Reviewed Preprint
v1 • May 15, 2026
Not revised

For correspondence:
isabel-muzzio@uiowa.edu

^ Authors contributed equally

Competing interests: No competing interests declared

Funding: See page 31

Reviewing editor: Joshua Johansen, RIKEN Center for Brain Science, Japan

Learned and inferred valence arise from interactions between stable and dynamic subnetworks

Marc E Normandin^, Pedro M Ogallar^, Matthew R Lopez^, Isabel A Muzzio
Department of Psychological & Brain Sciences, University of Iowa, Iowa City, United States

eLife Assessment

This **valuable** study examines how the prelimbic cortex represents learned and generalized threat over time and identifies potentially distinct stable and dynamic subnetworks that may support these functions. The work is conceptually interesting and is strengthened by the longitudinal calcium imaging approach and the inclusion of key control groups. However, the evidence supporting the claims is **incomplete**, particularly because the interpretations regarding inference, time-dependent representational change, and the dissociation of neural activity from freezing behavior extend beyond what is currently established by the data.

https://doi.org/10.7554/eLife.111177.1.sa4

Abstract

Adaptive behavior requires assigning emotional value to sensory cues and inferring valence for novel stimuli to guide appropriate generalization. The prelimbic cortex (PL) is critical for threat expression and discrimination, yet its neuronal ensembles undergo pronounced turnover over time. How stable memory representations emerge from such network dynamism—and how they support inference to previously unexperienced stimuli—remains unresolved and central to debates on systems consolidation and the neural basis of generalization. Using longitudinal calcium imaging in freely moving mice, we tracked PL population activity for 30 days during two opposing versions of tone-discriminative fear learning and probed responses to conditioned and novel tones at recent and remote time points. Despite substantial ensemble reorganization, PL population dynamics reliably encoded graded emotional valence. Stimulus-evoked population similarity scaled precisely with behavioral generalization, and consistent population states emerged only for tones associated with shock or those eliciting strong generalized freezing, indicating that population-level similarity predicts inferred threat. Network analyses identified two functionally distinct subnetworks. Dynamic tone-selective ensembles encoded sensory features independent of learning and exhibited substantial turnover. In contrast, a valence-coding subnetwork whose neurons responded to all frequencies, integrated learned and inferred emotional value along a graded axis. Strikingly, only these graded valence neurons preserved cellular identity and response structure across time. These findings reveal that persistent valence-encoding subnetworks form a stable scaffold embedded within dynamic cortical ensembles. This architecture reconciles cortical turnover with long-term memory stability and provides a circuit-level mechanism for maintaining the emotional “gist” of experience while enabling flexible generalization.

Introduction

Adaptive behavior depends on evaluating sensory cues according to their emotional significance. Acoustic signals are particularly effective for threat detection because they provide rapid, omnidirectional information that often precedes danger. This broad accessibility reduces spatial

specificity and increases learning variability, making auditory associative learning a powerful model for studying how the brain extracts shared, robust features across experiences for proper generalization. The prelimbic cortex (PL) contributes to the expression (Burgos-Robles et al., 2009; Sierra-Mercado et al., 2011; Sotres-Bayon & Quirk, 2010) and the proper discrimination and generalization of threat memories (Rosas-Vidal et al., 2025; Stujenske et al., 2022). Accumulating evidence also points to a broader role for the PL in auditory processing, particularly through top-down attentional modulation along the auditory pathways (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006).However, it remains unclear whether the same or distinct PL neurons encode sensory features of sounds versus their learned or inferred emotional significance, and how these representations change with experience.

At the population level, episodic (DeNardo et al., 2019; Kitamura et al., 2017) and procedural (Do-Monte et al., 2015; Iqbal et al., 2026) memory ensembles undergo pronounced reorganization across days—a hallmark of systems consolidation—even as threat-associated behaviors remain stable over time. This dissociation between neural instability and behavioral persistence indicates that memory representations are not fixed at the level of individual neurons, but instead depend on population-level structure that is preserved across time (Deitch et al., 2021; Gallego et al., 2020). This idea raises several central questions: How can population-level representations remain functionally stable while their cellular constituents continually change? Are all neurons equally dynamic, or do distinct subpopulations exhibit differential stability? How do dynamic networks support appropriate generalization, ensuring that only potentially dangerous stimuli generate overlapping neural representations, amid ongoing ensemble turnover?

Recent studies demonstrate that PL neurons generalize emotional value by inferentially updating the valence of previously neutral stimuli after learning (Gu & Johansen, 2025). However, the neural mechanisms underlying this process remain unclear. Because generalization often extends to entirely novel stimuli that have never been experienced (for example, danger associated with a familiar siren generalizing to a new siren), it is unknown how subsets of neurons responding to these novel cues infer valence amid ongoing network dynamism. Here, our goal was to determine which network-level features preserve the essential components of a memory across time and how these elements support appropriate generalization to novel cues. We hypothesized that generalization to novel stimuli depends on stable subnetwork organization that enables comparisons between learned and inferred valence, as well as population-level features that reduce variability across related representations. To test this hypothesis, we combined longitudinal calcium imaging with computational analyses in freely moving mice to examine PL network dynamics during discriminative auditory learning and retrieval in the presence of both conditioned and novel tones. Our results show that stable cortical subnetworks integrate the emotional “gist” of memory and inferred valence for novel cues over time, despite ongoing ensemble reorganization, and that population-level firing rate similarity across stimulus presentations determines threat generalization.

Results

Logarithmic tone separation predicts freezing generalization

To assess memory and generalization, GRIN-lens–implanted and non-implanted mice were trained in a differential auditory fear-conditioning paradigm (Fig. 1a). One tone (CS^+; 80 dB) was paired with a mild foot shock (0.5 mA, 0.5 s), whereas a second tone (CS^-; 80 dB) was never paired with shock. In one group, the CS^+ was 15 kHz and the CS^- was 3 kHz (CS^+15; N = 27; 7 implanted, 20 non-implanted); in a second group, contingencies were reversed (CS^+3; N = 22; 5 implanted, 17 non-implanted). A no-shock control group was exposed to the same tones without shock (N = 13; 6 implanted, 7 non-implanted).

Memory retrieval was tested on days 1, 15, and 30 after conditioning to probe early, long-term, and remote memory (Bontempi et al., 1996). During retrieval, mice were tested in a novel context with the CS^+, CS^-, and two intermediate frequencies (7 and 11 kHz), presented in semi-random

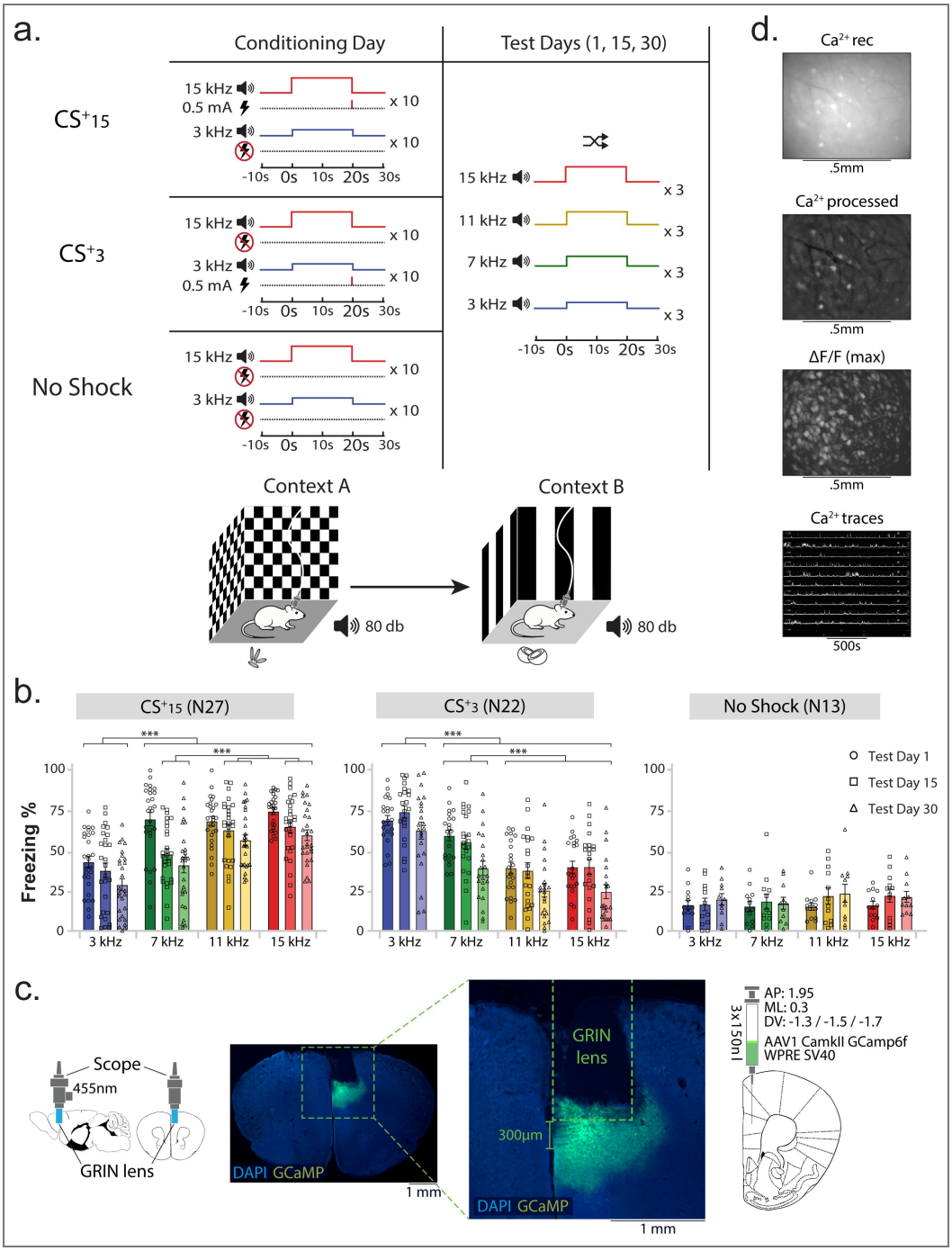

Figure 1. Experimental design, behavior, histology and calcium imaging.

a,Experimental design illustrating discriminative fear conditioning with either a 15 kHz CS$^+$ or a 3 kHz CS$^+$ (experimental groups), and no-shock controls tested at identical time points but never exposed to footshock. All groups were tested with 3 and 15 kHz tones and two intermediate frequencies (7 and 11 kHz) on days 1, 15 and 30 after conditioning. **b,** Behavioral responses across testing days (CS$^+$15 kHz: n = 27 mice; CS$^+$3 kHz: n = 22 mice; no-shock controls: n = 13 mice). Two-way repeated-measures ANOVA: CS$^+$15 : effect of day $F(2,52) = 13.28$, $p < 0.001$, frequency: $F(3,78) = 85.09$, $p < 0.001$, day × frequency interaction: $F(6,156) = 3.79$, $p = 0.002$; CS$^+$3: effects of day $F(2,42) = 14.42$, $p < 0.001$, frequency: $F(3,63) = 58.81$, $p < 0.001$, day × frequency interaction: $F(6,126) = 1.41$, $p = 0.217$; no-shock controls: effects of day: $F(2,20) = 1.38$, $p = 0.276$, frequency: $F(3,30) = 1.43$, $p = 0.253$; day × frequency interaction: $F(6,60) = 0.335$, $p = 0.916$. Significant Tukey multiple comparisons are denoted by asterisks, *** $p < 0.001$. **c,** Representative histological section showing GCaMP6f expression. Imaging depth did not exceed 300 µm, restricting recordings to PL. **d,** Example calcium imaging data showing raw fluorescence signals, deconvolved activity, thresholded events and corresponding activity traces.

order, with each tone repeated three times (Fig. 1a). Frequencies were spaced linearly to assess whether logarithmic frequency separation predicts similarity to the threat-associated tone. Although CS^+/CS^- separation was identical across groups (2.32 octaves), spacing between the CS^+ and adjacent frequencies differed (0.45 octaves for 11–15 kHz; 1.22 octaves for 3–7 kHz). Accordingly, if logarithmic separation governs threat generalization, freezing was expected to generalize to 11 kHz in CS^+15 mice, with weaker generalization between 3 and 7 kHz in CS^+3 mice.

In the CS^+15 group, mice consistently discriminated between the CS^+ and CS^-. Although generalization to intermediate frequencies was broad on day 1, it became progressively restricted to 11 kHz—the frequency adjacent to the CS^+—during long-term (day 15) and remote (day 30) retrieval ($p < 0.05$; Fig. 1b, left). CS^+3 mice also discriminated between the CS^+ and CS^-; however, these animals exhibited greater freezing to the CS^+ than to all intermediate frequencies across testing days ($p < 0.05$; Fig. 1b, center), indicating reduced generalization to the adjacent tone when the logarithmic separation was larger. No-shock control mice showed no significant differences in freezing across frequencies on any testing day ($p > 0.05$; Fig. 1b, right), confirming that freezing reflected associative learning. Together, these results indicate that logarithmic spacing, rather than simple proximity to the CS^+, was the primary determinant of generalization.

No sex differences were observed in CS^+15 mice (11 females, 16 males) or no-shock controls (6 females, 7 males; $p > 0.05$). In CS^+3 mice (11 females, 11 males), minor and inconsistent sex-related effects were detected on days 1 and 15, but these effects were absent by day 30 (Fig. S1). Accordingly, sex was not included as a factor in subsequent neural analyses.

Behavioral performance did not differ between implanted and non-implanted mice on days 1 and 15 ($p > 0.05$). On day 30, a modest main effect of group was observed in the CS^+3: group, reflecting higher overall freezing levels in implanted animals compared with non-implanted mice ($p < 0.03$). Importantly, both experimental groups exhibited robust frequency effects across days ($p < 0.001$), with no group × frequency interactions ($p > 0.05$), indicating preserved threat discrimination and generalization profiles. A similar elevation in overall freezing was observed in implanted no-shock controls ($p < 0.05$), suggesting that group differences reflect nonspecific changes in behavioral output rather than differences in associative learning.

PL sound-responsive networks exhibit dynamic properties over time

To examine the temporal evolution of PL neuronal responses, we performed longitudinal calcium imaging in CS^+15 (N = 7), CS^+3 (N = 5), and no-shock control mice (N = 6). Across groups, neurons were tracked during conditioning and all retrieval sessions. Figs. 1c and 1d show GCaMP6f expression in PL, representative calcium footprints, and activity traces. Sound-evoked responses were visualized using raster plots with calcium activity ranked across simultaneously recorded neurons (Fig. 2a). Across animals and sessions, we identified distinct neuronal populations showing positive modulation, negative modulation, mixed responses, or no consistent response to sound (Fig. 2b). Sound responder neurons were classified using a test that detected modulation based on magnitude relative to baseline variability, allowing reliable identification of both transient and sustained responses while remaining robust to noise. Fig. 2c summarizes the proportions of response types pooled across animals and sessions.

Given the ongoing debate over whether neocortical memory traces stabilize or remain dynamic over time (DeNardo et al., 2019; Kitamura et al., 2017; Kupke & Oliveira, 2025; Lopez et al., 2024; Mau et al., 2020; Rao-Ruiz et al., 2021; Refaeli et al., 2023; Terranova et al., 2023; Zaki & Cai, 2024), we next assessed the temporal stability of PL networks by tracking the cellular footprints of sound-responsive neurons across retrieval sessions. Network stability was visualized using Venn diagrams (Fig. 2c). A moderate proportion of neurons was present across all retrieval sessions, with no differences between groups ($p > 0.05$). Likewise, there were no group differences in the proportion of neurons overlapping across only two sessions ($p > 0.05$). For neurons present in only a single session, control mice exhibited a higher proportion on day 30

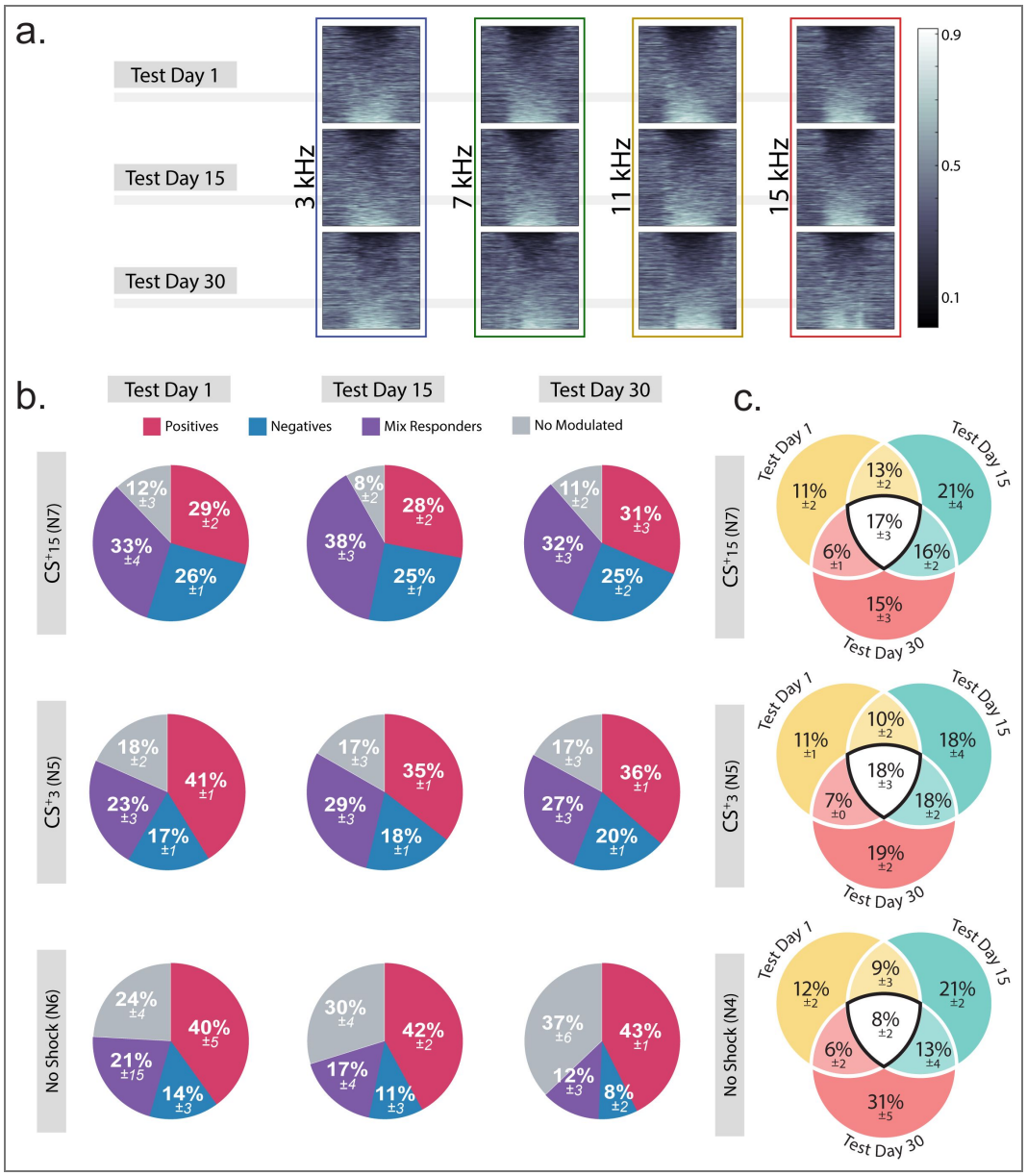

Figure 2. Neuronal activity distributions across days and stability of the active network.

a,Raster plots from an animal trained with a 15 kHz CS+, ordered by activity level. Raster plots illustrate positive sound-responsive neurons (bottom), negative sound-responsive neurons (top), and mixed or non-responsive neurons (middle). **b,** Proportions of neuronal response types across experimental groups and testing days. Number of cells: CS+15: conditioning: 3,834; day 1: 3,177; day 15: 4,186; day 30: 3,693, CS+3 conditioning: 1,628; day 1: 1,381; day 15: 1,881; day 30: 1,833, control: conditioning: 1,118; day 1: 986; day 15: 1,386; day 30: 563 (2 mice only yielded data up to day 15). **c,** Venn diagrams showing the proportions of consistently active neurons (active across all retrieval sessions; one-way ANOVA: $F(2,13) = 2.70$, $p = 0.104$), partially active neurons (active in two sessions; one-way ANOVA: days 1–15, $F(2,13) = 1.45$, $p = 0.271$; days 15–30, $F(2,13) = 1.02$, $p = 0.388$; days 1–30, $F(2,13) = 0.33$, $p = 0.723$), and transiently active neurons (active in a single session; day 1, Kruskal–Wallis: $H(2) = 1.90$, $p = 0.387$; day 15, one-way ANOVA: $F(2,13) = 0.06$, $p = 0.942$; day 30, Kruskal–Wallis: $H(2) = 7.52$, $p < 0.05$). Diagrams and analyses for control mice include only mice recorded for 30 days.

compared with CS15+ mice ($p < 0.05$), with no other differences observed. Overall, these data indicate that the majority of neurons overlapped across only a limited number of sessions or were transiently active, reflecting pronounced population-level dynamism over time.

Sound-modulated PL population responses encode learned and inferred valence

To assess population-level representations of learned and novel tones, calcium activity was averaged across three presentations for each frequency during a 10-s baseline, 20-s tone presentation, and 10-s post-stimulus interval. In CS$^+$15 mice, positively modulated sound-responsive neurons exhibited graded tone activity reflecting the contingency learned valence as well as the inferred valence of novel tones across testing days (Fig. 3a, top). Area-under-the-curve (AUC) analyses obtained by averaging recorded cells per animal revealed the highest responses at 15 and 11 kHz, intermediate responses at 7 kHz, and the lowest at 3 kHz; AUCs for 11 and 15 kHz exceeded those for 3 kHz ($p < 0.05$), with no difference between 11 and 15 kHz ($p > 0.05$) across testing days. This population level gradient mirrored behavioral generalization (Fig. 1b, left), with strong responses to 11 kHz and intermediate responses to 7 kHz. By contrast, negatively modulated neurons showed overlapping responses across frequencies with no significant differences at any time point ($p > 0.05$; Fig. 3a, bottom).

A complementary pattern was observed in CS$^+$3 mice. On day 1, the AUC for 3 kHz was greater than that for 11 kHz ($p < 0.05$), but only showed a trend relative to 15 kHz ($p = 0.68$; Fig. 3, middle panel); however, on days 15 and 30, it exceeded that for both 11 and 15 kHz ($p < 0.05$). Responses to 7 kHz reached significance relative to 3 kHz on days 1 and 15 ($p < 0.05$), but not on day 30. As in CS$^+$15 mice, negatively modulated neurons exhibited overlapping responses across frequencies and days ($p > 0.05$). In no-shock controls, although both positive and negative responses were present, population activity was not modulated by tone frequency or valence ($p > 0.05$; Fig. 3c, bottom panel), indicating that graded responses require associative learning. Together, these results show that despite substantial neuronal turnover, PL population responses encode graded sound-valence associations that reflect both learning and inference, closely matching behavioral generalization.

Consistently active neurons preserve valence representations as newly recruited neurons sharpen remote memory traces

Longitudinal tracking of individual neurons revealed how subpopulations with distinct stability profiles shape the evolution of cortical memory representations. We compared population responses across three stability types: consistently active neurons (active during conditioning and all retrieval sessions), emerging–retained neurons (recruited after conditioning and persisting through day 30), and transiently active neurons (only present during a single retrieval session). Consistently active, positively modulated neurons exhibited graded population responses reflecting both learned and inferred sound associations from day 1 onward in both experimental groups, which was quantified by calculating AUC per animal ($p < 0.05$; Fig. S2a–b, Table S1). In contrast, negatively responding neurons showed no graded tuning across days ($p > 0.05$), except for a day 1 difference between 15 and 3 kHz in CS$^+$15 mice ($p < 0.05$). In control mice, positively and negatively modulated neurons displayed variable and inconsistent activity across frequencies in all cell categories (consistently active, emerging–retained, and transiently active). As a result, these neurons were not included in further cell-type analyses.

Emerging-retained neurons recruited after conditioning exhibited graded population responses when recruitment occurred after day 1 (Fig. S3a–b, Table S1). This category encompassed neurons that emerged on day 1 and persisted through day 30, as well as neurons that emerged on day 15 and remained active through day 30. In both experimental groups, graded tuning was absent on day 1 ($p > 0.05$) but emerged on days 15 and 30, when population responses closely

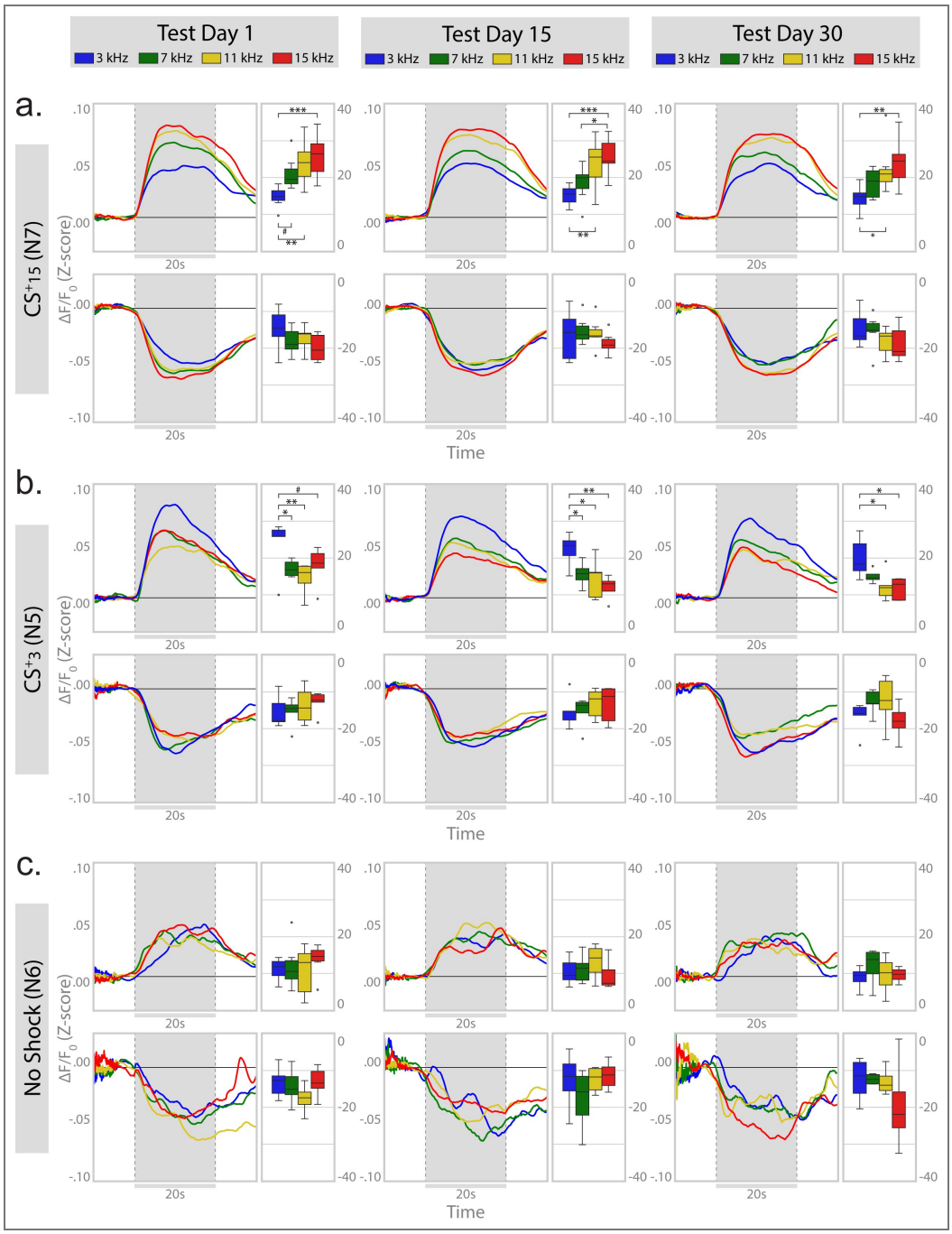

Figure 3. Population activity of positive sound responders shows emotional graded valence patterns of activity in response to tones.

a-c, Population responses in animals trained with a 15 kHz CS+ (a), a 3 kHz CS+ (b), and no-shock controls (c). In all groups, upper panels show positively responsive neurons and lower panels negatively responsive neurons. Boxplots show the median (center line), interquartile range (box), and whiskers extending to ±1.5× the interquartile range; points outside the whiskers represent individual observations beyond this range. CS+15: positively modulated: day 1: $F(3,18) = 9.963$, $p < 0.001$; day 15: $F(3,18) = 9.973$, $p < 0.001$; day 30: $F(3,18) = 6.627$, $p = 0.003$; negatively modulated: day 1: $F(3,18) = 2.483$, $p = 0.094$; day 15: $F(3,18) = 1.877$, $p = 0.178$; day 30: $F(3,18) = 2.753$, $p = 0.073$). CS+3: positively modulated: day 1: $F(3,12) = 6.899$, $p = 0.006$; day 15: $F(3,12) = 9.247$, $p = 0.002$; day 30: $F(3,12) = 6.123$, $p = 0.009$; negatively modulated: (day 1: $F(3,12) = 0.87$, $p = 0.484$; day 15: $F(3,12) = 0.512$, $p = 0.641$; Day 30: $F(3,12) = 1.448$, $p = 0.278$); no shock control: positively modulated: day 1: $F(3,15) = 0.527$, $p = 0.670$; day 15: $F(3,15) = 1.852$, $p = 0.181$; day 30: $F(3,9) = 1.046$, $p = 0.418$; negatively modulated: day 1: $F(3,13) = 1.205$, $p = 0.347$; day 15: $F(3,13) = 1.375$, $p = 0.294$; day 30: $F(3,9) = 0.95$, $p = 0.457$. Significant Tukey multiple comparisons are denoted by asterisks, *p < 0.05, **p < 0.01, ***p < 0.001.

mirrored behavioral generalization, with responses to 11 and 15 kHz becoming increasingly similar regardless of CS$^+$ identity ($p > 0.05$). By contrast, emerging–retained negatively responding neurons showed no significant emotional tuning across days.

Finally, we examined transiently active neurons, defined as cells present only on a single retrieval session. Transiently positive responders showed no valence tuning when active on day 1 but exhibited clear graded population responses when recruited on days 15 or 30 in both experimental groups (Fig. S4a–b, Table S1). In contrast, transiently negative responders showed no significant frequency tuning at any time point ($p > 0.05$, Table S1). Together, these results indicate that consistently active neurons maintain stable representations of learned and inferred sound associations across time, whereas neurons recruited after conditioning progressively acquire graded tuning at later retrieval stages. This dynamic refinement suggests that cortical memory representations become increasingly selective during systems consolidation, while a stable neuronal subpopulation preserves the core emotional content of the memory.

Population vector similarity at stimulus onset determines degree of generalization

Consistent neural responses across stimulus repetitions within sessions support stable population representations that enable recognition and generalization to similar cues (Hoshi et al., 2023). To assess population similarity within each retrieval session, we compared activity across repeated tone presentations (three per session, presented in semirandom order). Population activity was estimated using CASCADE, a validated spike-inference neural network (Rupprecht et al., 2021), and activity vectors were constructed from simultaneously recorded neurons. We quantified response similarity across tone pairs during stimulus presentation to assess temporal and rate consistency at the population level.

Rate-based population similarity analyses revealed reliable, temporally structured responses for 3/3 tone pairs in CS$^+$3 mice and for 15/15 and 15/11 tone pairs in CS$^+$15 mice that were absent in control mice for any frequency pair (Fig. 4a-d). Because population similarity peaked shortly after stimulus onset, we quantified similarity during the first 5 s after tone onset relative to the CS$^+$. In CS$^+$15 mice, population similarity was highest for 15/15 and 15/11 tone pairs ($p < 0.001$), with no difference between them ($p > 0.05$), and was significantly greater than for 15/3 comparisons ($p < 0.05$, Fig. 4e). In CS$^+$3 mice, similarity was highest for 3/3 tone pairs ($p < 0.004$) and significantly lower for 11/3 and 15/3 comparisons (Fig. 4f; $p < 0.05$). The 3/3 and 3/7 comparisons showed a non-significant trend ($p = 0.08$). No differences were observed among 7/15, 11/15, and 15/15 tone pairs in either group ($p > 0.05$). These findings indicate that population-level similarity at stimulus onset scales with behavioral threat generalization and is maximal for tones associated with robust threat responses.

Different subnetworks encode acoustic versus learned properties of sound association

Our previous analyses show that learned and inferred associations are represented at the population level. However, these results do not resolve whether graded responses arise from pooled activity of frequency-selective neurons or from subnetworks encoding integrated learned valence across tones. Because most neurons exhibit dynamic properties and only a subset remains stable over time, we hypothesized that the PL active ensemble segregates into functionally distinct subnetworks: one encoding tone-specific sensory features with dynamic characteristics, and another responding to all frequencies encoding stable core memory content and inferred emotional valence.

To test this hypothesis, we developed a clustering approach based on mutual information (MI), which captures both linear and nonlinear relationships between neuronal response profiles (Quian Quiroga & Panzeri, 2009). MI was used to group neurons with related activity patterns into functional subnetworks. Because MI does not preserve response polarity, neurons were subsequently classified by the sign of their pairwise correlations and re-clustered, consistently

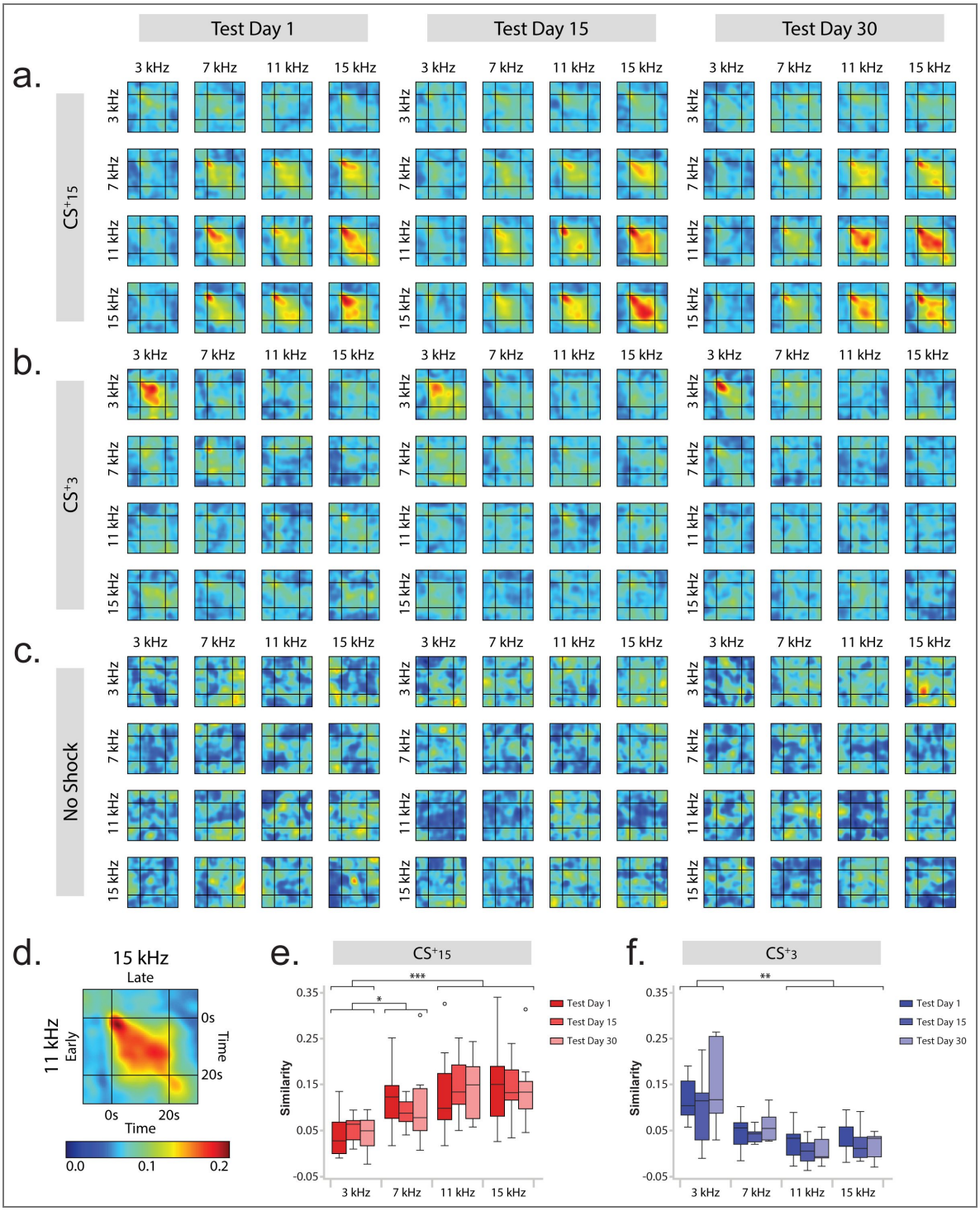

Figure 4. Population vector similarity across tones and time.

a-c, Population similarity maps for all tone pairs across the time course of sound presentation in the CS+15 (a), CS+3 (b), and no-shock control (c) groups. **d,** Schematic illustrating the color scale and map orientation; the y-axis denotes the earlier tone in the comparison, and the x-axis denotes the later tone. **e-f,** Box plots showing population similarity during the first 5 s following tone onset, quantified relative to the CS+ for the CS+15 (e, $F(3,36) = 12.025$, $p < 0.001$) and CS+3 (f, $F(3,24) = 7.435$, $p = 0.004$) groups. Significant Tukey multiple comparisons are denoted by asterisks, $p < 0.05$, $p < 0.01$, $p < 0.001$.

revealing two major response classes: positively and negatively modulated sound responders (Fig. 5a). Clustering was performed across all animals to derive global cluster identities (Fig. 5) and within each experimental and control group to identify learning-specific subnetworks (Figs. S5–S7).

Clustering quality was assessed by sorting signed MI values by global cluster identity and comparing within-versus across-cluster correlations (Fig. 5b). The resulting box-diagonal structure indicated robust clustering, with comparable quality indices across groups (CS^+15: 0.28; CS^+3: 0.27; controls: 0.30). In all cases, within-cluster correlations exceeded across-cluster correlations, supporting the reliability of the approach (CS^+15: 0.30 vs. −0.03; CS^+3: 0.27 vs. −0.02; controls: 0.28 vs. −0.04).

Across all groups, we identified positively and negatively modulated clusters selective for individual frequencies, indicating that frequency-specific representations are present in PL independent of learning (Fig. 5c–d; Figs. S5–S6), consistent with a role in auditory processing (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006). In contrast, clusters encoding graded learned associations—characterized by responses to all frequencies that scaled with learned and inferred emotional value—were observed exclusively in trained animals (Fig. 6; Figs. S5–S6). Responses peaking at 15 kHz were specific to the CS^+15 group (Fig. 6a–b; Fig. S5), whereas responses peaking at 3 kHz were unique to the CS^+3 group (Fig. 6d–e; Fig. S6).

If neurons encoding graded responses carry core mnemonic information, they should exhibit enhanced stability over time. To test this hypothesis, we quantified the proportion of registered neurons that retained their cluster identity across at least two retrieval sessions and compared these values to a shuffled null distribution (10,000 iterations), with multiple comparisons controlled using the Benjamini–Hochberg procedure. Only the positively modulated clusters encoding graded learned associations showed significant identity stability across all retrieval intervals (days 2–15, 15–30, and 2–30; Fig. 6a-c). In contrast, negatively modulated graded responders and a smaller positively modulated cluster in the CS^+3 group showed stability only between adjacent sessions (Fig. 6b, d, e, Table S2). We repeated this analysis focusing only on consistently active neurons finding the same results. These data demonstrate that graded valence clusters remain consistently active and retain stable cellular identity, enabling encoding of core memory components and providing a stable scaffold for inferred valence comparisons.

Graded clusters encode emotional valence but constitute only a fraction of the active population; yet valence coding at the population level remains accurate and precise. This indicates that neurons newly recruited into the population—likely frequency-selective and organized within learning-independent clusters—can be shaped by associative processes through modulation of firing activity. To test this hypothesis, we calculated the normalized baseline/stimulus mean firing rate ratio (BSR) of positively responding neurons within the identified clusters using CASCADE (Rupprecht et al., 2021).

Although clusters of neurons responding to individual tones emerged independently of learning (i.e., they were present in experimental and control mice), their activity was modulated by associative processes. Specifically, neurons in clusters selectively responding to 3 kHz exhibited higher BSR in both experimental groups compared with neurons from the same cluster in controls ($p < 0.05$), indicating elevated activity regardless of whether 3 kHz was associated with the CS^+ or CS^-. By contrast, the BSR of neurons selectively responding to 11 or 15 kHz in the CS^+15 group displayed significantly higher activity than corresponding neurons in the CS^+3 or control groups on days 15 and 30 ($p < 0.05$, Fig. S8 and Table S3). Importantly, in the CS+15 group—where robust threat generalization was observed to 11 kHz—clusters associated with 15 or 11 kHz exhibited similarly high BSR to both tones, indicating comparable firing-rate responses. As expected, the BSR of positively modulated graded-responder neurons mirrored freezing behavior across experimental groups, with higher BSR to the CS+ and to tones eliciting robust generalization, and lower BSR to tones that elicited discrimination ($p < 0.05$; Fig. 6h–i, Table S3).

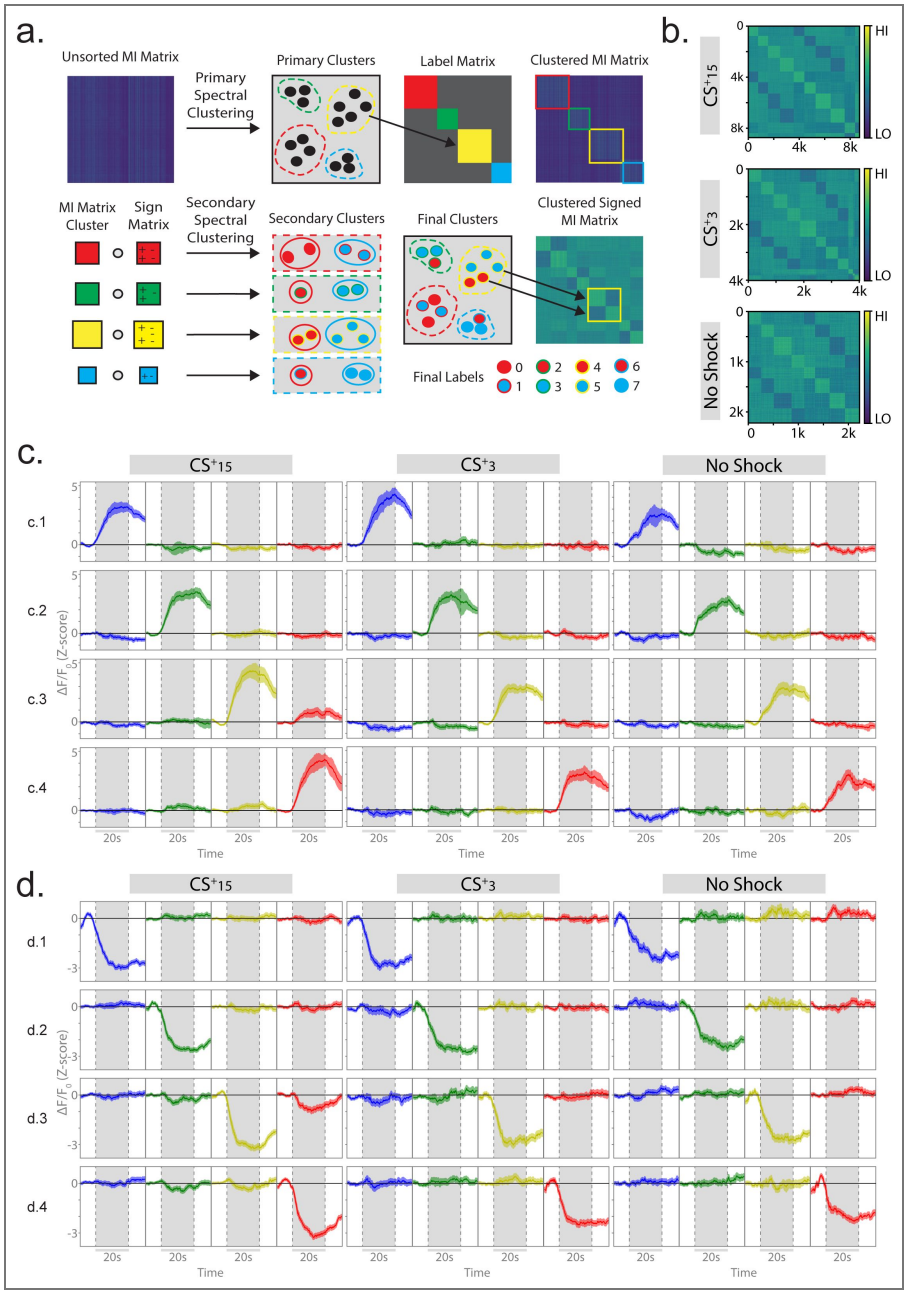

Figure 5. Clustering of PL subnetworks based on signed mutual information

a,Schematic of the mutual information (MI)–based clustering pipeline. An unsorted MI matrix computed from simultaneously recorded prelimbic (PL) neurons was first subjected to spectral clustering to identify primary clusters based on shared information structure, independent of response sign. The MI matrix was then reordered according to these primary cluster assignments. Within each primary cluster, MI values were combined element-wise with a corresponding sign matrix encoding the direction of correlation between cell pairs, yielding a signed MI matrix. Spectral clustering was subsequently applied independently within each primary cluster to identify secondary subclusters with distinct signed interaction patterns. Each subcluster was assigned a unique label, and all labels were combined to generate the final clustered signed MI matrix, enabling separation of positively and negatively modulated sound-responsive neurons. **b,** Final clustered signed MI matrices for each experimental and control groups. Matrices are sorted by cluster labels for the CS+15 group (top), CS+3 group (middle), and No Shock group (bottom). Color scale indicates signed MI strength (HI to LO). **c,** Average stimulus-aligned population responses for clusters showing strong positive modulation to individual tones. C.1–C.4 show primary responses to 3 kHz, 7 kHz, 11 kHz, and 15 kHz tones, respectively, across groups. **d,** Average stimulus-aligned population responses for clusters showing strong negative modulation to individual tones. D.1–D.4 show primary responses to 3 kHz, 7 kHz, 11 kHz, and 15 kHz tones, respectively, across groups.

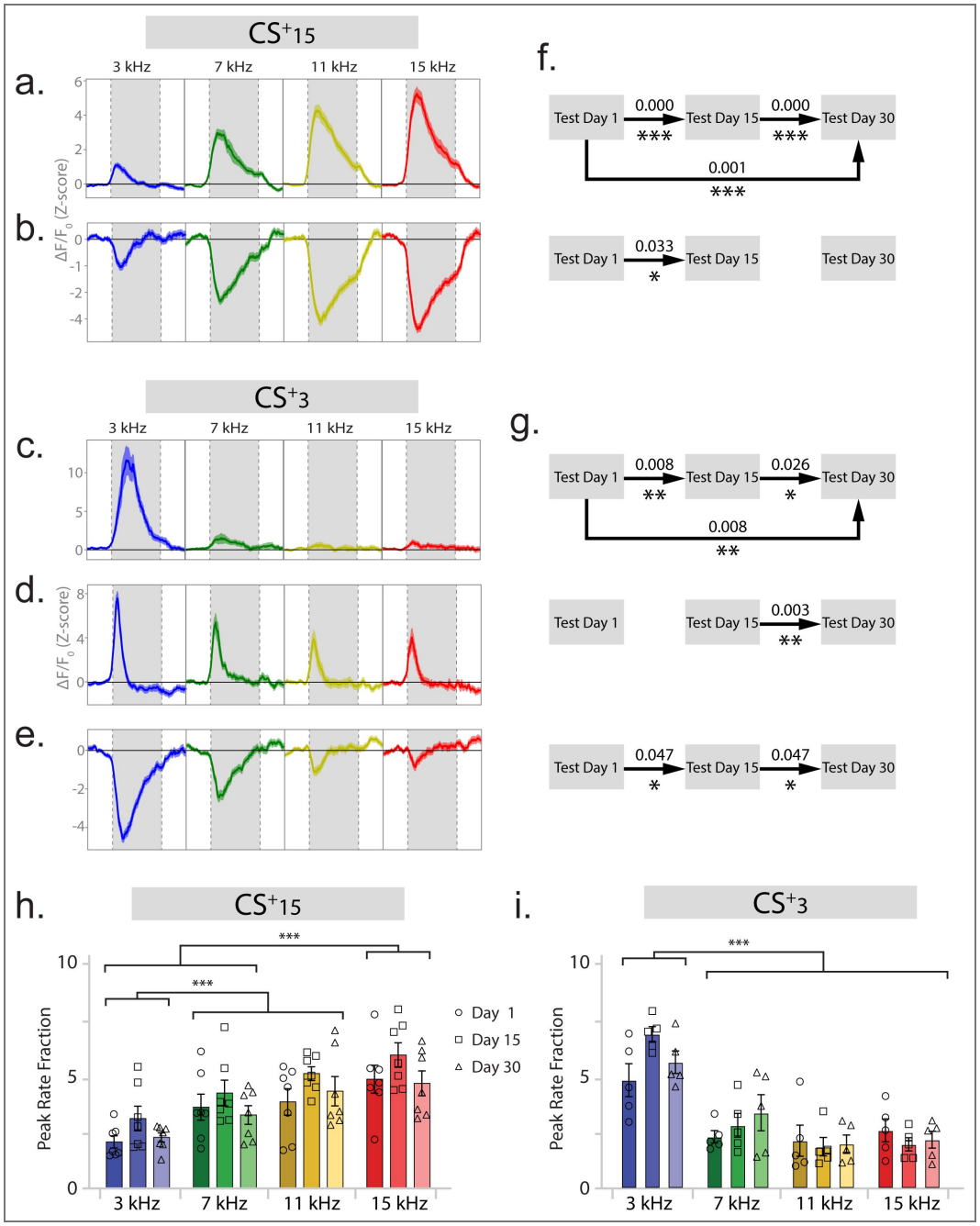

Figure 6. Graded emotional response clusters are present in experimental groups.

a-b, Average stimulus-aligned population responses for clusters showing graded emotional valence in animals trained with a 15 kHz CS+, showing positive (a) and negative (b) response patterns. **c-e,** Average stimulus-aligned population responses for clusters showing graded emotional valence in animals trained with a 3 kHz CS+, showing positive (c-d) and negative (e) response patterns. **f-g,** Analysis of stability of neuronal identity within graded valence clusters across retrieval sessions. Benjamini–Hochberg corrected significance denoted by asterisks. **h-i,** Baseline/stimulus firing rate ratio (BSR) showing changes in activity of graded clusters only present in the experimental CS+15 (h) and CS+3 (i) groups across days. Significant Tukey multiple comparisons are denoted by asterisks, *p < 0.05, **p < 0.01, ***p < 0.001.

Together, these results indicate that firing rates of stable graded clusters and dynamic tone selective ones encode both learned and inferred emotional valence.

Discussion

Our results show that PL populations encode learned emotional valence despite substantial turnover in active neurons over time. A subset of neurons remains consistently active across sessions, preserving core components of the memory trace and supporting inference of emotional valence for novel sounds, while neurons recruited after conditioning progressively acquire valence selectivity at remote time points. Population similarity at stimulus onset is observed for threat-associated cues and for stimuli that elicit robust generalization, consistent with the idea that generalization arises when similar inputs reactivate shared population states (Aschauer et al., 2022). Notably, populations responses emerge from coordinated activity across distinct subpopulations, including stable subnetworks encoding core memory content and inferred associations and dynamic sensory ensembles providing flexibility through learning-dependent modulation.

These findings bear directly on debates regarding systems consolidation and cortical memory stability (Frankland & Bontempi, 2005; Lopez et al., 2024). Classical systems consolidation models propose that memory formation initially depends on subcortical regions, followed by the gradual transfer and stabilization of memory traces in the neocortex (Moscovitch & Nadel, 1998; Nadel et al., 2000). In contrast, multiple-trace frameworks posit parallel encoding and continuous reorganization, with memory representations stabilizing only as they become increasingly schematic (Moscovitch & Nadel, 1998; Nadel et al., 2000; Tonegawa et al., 2018). Our data reconcile these views by demonstrating that cortical representations of emotional valence emerge rapidly after learning and persist within stable subnetworks, even as the broader population undergoes substantial turnover. This architecture preserves core mnemonic content while allowing flexibility in the surrounding ensemble.

At the circuit level, this organization aligns with principles established for memory engrams. A large body of work has shown that memories are encoded in sparse neuronal ensembles whose activity is both necessary and sufficient for memory expression (Josselyn & Tonegawa, 2020), and that these ensembles form stable functional units embedded within distributed circuits through strengthened synaptic connectivity (Tonegawa et al., 2018). Consistent with this framework, the graded valence subnetworks identified here exhibit hallmark properties of canonical engram populations, including stable cellular identity and persistent response profiles. Importantly, these subnetworks encode both learned contingencies and the inferred valence of novel stimuli along a graded representational axis, suggesting that strong recurrent connectivity provides a stable scaffold for emotional memory representations.

In the auditory cortex, neurons exhibiting sound-evoked suppression (“negative responders”) are thought to contribute to lateral inhibition, sharpening frequency tuning and improving signal-to-noise by suppressing activity in neurons tuned to similar or neighboring frequencies (Kato et al., 2017; Wehr & Zador, 2003). In our study, negative responders did not display consistent graded responses at the population level. However, we identified negatively graded clusters whose response profiles mirrored those of positively graded clusters, exhibiting opposite modulation across frequencies. This organization suggests that suppressive responses form structured subnetworks that may provide complementary inhibitory contrast to excitatory valence-encoding ensembles. Such opponent network dynamics could enhance discriminability at the population level without requiring individual negative responders to exhibit stable tuning across time.

Importantly, the precision of population graded emotional responses does not arise from learning-dependent recruitment of all participating neurons. Dynamic tone-selective responsive neurons emerge independently of learning, as they are present in both control and experimental mice, reflecting pre-existing PL sensory-driven properties (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006). However, the activity of these subnetworks is strongly modulated by associative learning, indicating that learning reshapes the gain of dynamic sensory ensembles. Conversely, graded valence emotional clusters only emerge following associative conditioning, preserving

elevated rate activity over time. This interaction between pre-existing sensory organization and learning-dependent modulation provides a mechanism by which PL circuits remain flexible while maintaining emotionally relevant information (Mau et al., 2020). Within this framework, PL supports threat generalization and inference by engaging overlapping and consistent population states for behaviorally generalizing cues. Together, this organization provides a circuit-level substrate through which PL compares novel sensory inputs with established emotional representations, enabling appropriate generalization or discrimination while preserving core emotional memory components despite ongoing ensemble reorganization.

Methods

Subjects

Female and male C57BL/6J mice (IMSR_JAX:000664; Jackson Laboratory, Bar Harbor, ME), aged 8–10 weeks, were housed on a 12 h light/dark cycle. All experiments were conducted during the light phase of the cycle. Animal housing and care were consistent with standards set by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC). All experimental procedures were approved by the University of Iowa Institutional Animal Care and Use Committee.

Discriminatory fear conditioning task

The training day, mice were placed in a square conditioning chamber (20 × 20 cm; Context A, Fig. 1a). The first 3 minutes served as a habituation period. Following habituation, experimental mice received presentations of a conditioned stimulus (CS^+) tone (15 or 3 kHz, 80 dB, 20.5 s) that co-terminated with a mild foot shock (unconditioned stimulus, US; 0.5 s, 0.5 mA). Between CS^+ presentations, animals were presented with a non-conditioned stimulus (CS^-) tone (3 or 15 kHz, 80 dB, 20.5 s) that was never paired with shock. Mice received 10 CS+ tone–shock pairings and 10 CS− tone presentations. Intertrial intervals between CS^+ and CS^- presentations alternated pseudo-randomly between 2, 4, and 6 minutes. Following the final CS^- presentation, mice remained in the chamber for an additional 60 s before being returned to their home cages. Control mice were exposed to the same auditory stimuli but did not receive foot shocks. On day 1, mice were placed in a novel context (Context B; 20 × 20 cm; see Fig. 1a) that differed from the training context in wall color, floor texture, and background scent. Animals were allowed to freely explore the chamber for 3 minutes, after which they were presented with auditory stimuli consisting of the conditioned tone (CS^+) the non-conditioned tone (CS^-), and two intermediate frequencies (7 and 11 kHz) spanning the same spectral range. These intermediate frequencies were spaced linearly rather than logarithmically. Tones were presented in a semi-random order, with three presentations per frequency and 3-min intertrial intervals. Following the final tone presentation, mice remained in the chamber for an additional 60 s period before being returned to their home cages.

Behavioral analysis

All behavioral sessions were video recorded using a mounted camera. Freezing behavior was defined as complete immobility except for respiratory movements and was quantified as a percentage of total trial time. Freezing was measured using FreezeFrame 5 software (Actimetrics, *RRID:SCR_014429*) for non-implanted mice and inertial measurement unit (IMU) data from the Inscopix miniscope system for implanted mice. This approach quantifies movement across three axes and circumvents visual occlusion introduced by the miniscope tether. To validate automated scoring, a subset of sessions from implanted and non-implanted mice were randomly selected and manually scored by an experimenter blind to experimental condition, following established criteria (Phillips & LeDoux, 1992; Wang et al., 2012). Automated and manual measurements were highly correlated (Pearson's r = 0.98 and 0.96), confirming strong inter-method reliability. Freezing responses were analyzed across three retrieval sessions (days 1, 15, and 30). Two animals in the non-shock control group (JAM061 and JAM062) were excluded from repeated-measures analyses due to missing data on day 30 but were retained for day-specific comparisons.

Viral construct stereotaxic surgery

Mice were anesthetized with isoflurane (3% in oxygen at 1 L/min) and placed in a stereotaxic apparatus (David Kopf Instruments, Tujunga, CA). During surgery, anesthesia was maintained at 1.5%. The head was positioned to ensure a flat skull in both the anterior–posterior (AP) and medial–lateral (ML) planes. Rimadyl (5 mg/kg, s.c.) was administered preoperatively and for two consecutive days postoperatively for analgesia. For calcium imaging experiments, three injections (150 nL each) of AAV1-CaMKIIα-GCaMP6f-WPRE-SV40 (Addgene cat. # 100834) were delivered into the PL at the following stereotaxic coordinates relative to bregma: AP +1.95 mm, ML ±0.3 mm, and DV −1.7, −1.5, and −1.3 mm. Injections were performed at a rate of 50 nL/min. After each injection, the syringe was left in place for 5 minutes to allow for viral diffusion and to minimize backflow. Mice were allowed to recover for four weeks prior to implantation of the GRIN lens and baseplate assembly for in vivo imaging.

Miniendoscope baseplate placement

Following AAV-GCaMP6f injections, a gradient-index GRIN lens (Inscopix, diameter: 1.0 mm, length: 4.0 mm, numerical aperture: 0.5, model: 1050-004637) and baseplate assembly (Inscopix, Mountain View, CA) were implanted above the PL. Stereotaxic coordinates relative to bregma were AP +1.94 mm, ML ±0.3 mm, and DV −1.5 mm. The cortical surface was identified and set as the zero reference point. To create a tract for lens placement, a 26-gauge syringe needle was lowered to approximately two-thirds of the final depth (DV −1.0 mm) and then slowly retracted. The GRIN lens was subsequently lowered to the final DV coordinate (−1.5 mm) and secured in place with a thin layer of cyanoacrylate adhesive applied around the lens perimeter and over the exposed skull. Dental cement (Lang Dental) was used to further stabilize the baseplate and anchor it to the skull. Mice were allowed to recover for at least two weeks before the onset of behavioral training.

Calcium imaging analysis

Calcium imaging data acquired using the Inscopix nVoke 2.0 miniscope system were processed using Inscopix Data Processing Software (IDPS; Inscopix, Mountain View, CA). Raw recordings were first preprocessed using the standard IDPS pipeline, which included spatial downsampling, background subtraction, spatial filtering, and rigid motion correction to compensate for movement artifacts. Fluorescence signals were then normalized to ΔF/F, defined as the change in fluorescence relative to a baseline fluorescence level computed for each pixel, to quantify activity-dependent calcium dynamics. Following preprocessing, neuronal signals were extracted using a constrained non-negative matrix factorization approach optimized for one-photon calcium imaging (CNMF-E implemented by IDPS). This method decomposes the imaging data into spatial components corresponding to putative neurons and their associated temporal activity traces while accounting for background and neuropil contamination. Extracted components were manually curated to exclude non-neuronal signals and artifacts based on established morphological and activity criteria (Fig. 1d).

Sound responder classification

To quantify stimulus-evoked modulation of neuronal activity, we implemented a window-based sound response test (SRT) that assesses changes in calcium activity relative to a pre-stimulus baseline while preserving temporal structure during the stimulus period.

Data alignment and normalization

For each neuron, calcium fluorescence traces ($\Delta F/F_0$) were aligned to sound onset and resampled onto a common relative time axis. Each aligned response comprised a pre-stimulus baseline period (time < 0 s) and a stimulus period of fixed duration (20 s). Analyses were first performed at the single-trial level and subsequently aggregated across trials. To normalize activity across trials, each trial was z-scored relative to its own baseline period. Specifically, the mean and standard deviation of baseline activity were computed separately for each trial, and all time points within that trial were normalized by subtracting the baseline mean and dividing by the baseline standard

deviation. A small constant was added to the denominator to prevent division by zero. This normalization ensured that trial-to-trial differences in baseline variance did not bias subsequent comparisons.

Sliding window construction

To capture temporally localized stimulus responses, normalized traces were segmented into overlapping sliding windows spanning the stimulus period. Windows were defined by a fixed duration (1 s) and stride (0.5 s), with window boundaries determined directly from the sampling interval of the data. Only windows fully contained within the stimulus epoch were included. This approach enabled detection of responses with variable onset latencies and durations without assuming a fixed response time.

Window-wise response quantification

For each window and trial, the median normalized activity within the window was computed to provide a robust estimate of window-level response magnitude. For each trial, the median baseline activity was computed using the same normalized data. Window responses were expressed as the difference between the window median and the corresponding trial baseline median. Baseline-subtracted window responses were then averaged across trials, yielding a mean response magnitude for each time window. An effect size was computed for each window as the mean baseline-subtracted response divided by the standard deviation of baseline medians across trials. This metric expresses response magnitude in units of baseline variability and provides a standardized measure of modulation strength. For single-trial cases, the raw normalized difference was used directly.

Statistical comparison to baseline

To determine whether activity within each window differed significantly from baseline, the distribution of normalized activity values within each window was compared to the pooled distribution of normalized baseline samples across trials. For each window, a two-sample Kolmogorov–Smirnov test was applied, yielding a window-specific p-value. This nonparametric test was selected to avoid assumptions about distributional shape and to remain sensitive to changes in both central tendency and distribution structure.

Identification of significant response epochs

A window was classified as significant if it met two criteria: (i) a p-value below a predefined threshold (typically $p < 0.05$) and (ii) an absolute effect size of at least one baseline standard deviation. Significant windows were further classified as positive (increased activity) or negative (decreased activity) based on the sign of the effect size. Contiguous significant windows of the same sign were grouped into response epochs. To account for brief interruptions due to noise, adjacent epochs separated by short gaps (≤1 window) were merged. For each neuron and response direction (positive or negative), the longest contiguous significant epoch was selected for reporting.

Summary response metrics

For each detected response epoch, several summary metrics were extracted, including: (1) the mean effect size across the epoch, (2) the peak response magnitude (reported both in normalized units and converted back to $\Delta F/F_0$ using the baseline standard deviation), (3) onset and offset times relative to stimulus onset, and (4) total response duration. Based on these metrics, responses were classified as positively modulated (normalized effect size ≥ 3 and significant for at least 1 s) or negatively modulated (normalized effect size ≤ −1.5 and significant for at least 1 s).

Mapping active memory ensembles across time

For longitudinal analyses, neuronal identity across imaging sessions was tracked using spatial footprint registration to identify putatively identical neurons across days. Spatial footprints of cells detected during conditioning were registered longitudinally using a probabilistic model implemented in CellReg (Inscopix (Sheintuch et al., 2017)). This algorithm aligns neurons based

on the similarity of their spatial footprints, enabling consistent tracking of neuronal identity across imaging sessions. Cells classified as *consistently active* were those reliably detected across all phases of conditioning and retrieval, including sessions on days 1, 15, and 30.

Average stimulus-aligned trace procedure

For each neuron and stimulus frequency, the session-long calcium fluorescence trace was used as the initial input and then z-scored. For each tone presentation, traces were temporally aligned to stimulus onset. A time window spanning from 10 s before tone onset to 10 s after tone offset was extracted to generate a stimulus-aligned trace (SAT). To ensure equal trace length and consistent temporal alignment across presentations, each SAT was linearly interpolated. Baseline activity was computed as the mean signal during the 10 s pre-stimulus period (−10 to 0 s) and subtracted from the corresponding SAT to normalize activity relative to baseline. Baseline-corrected SATs from all repetitions were then averaged to yield a single average stimulus-aligned trace (ASAT) for each neuron and tone frequency.

Generation of Population Sound-Response Curves

To generate population-level sound-response curves, an average trace was first computed for each combination of subject, session, tone frequency, and modulation classification by averaging across all ASATs from individual cells within that condition (cell-averaged trace). Subsequently, for each experimental group, session, frequency, and modulation classification these cell-averaged traces were averaged across animals to obtain the group-level average trace, representing the population response to each auditory stimulus.

Population similarity over time across tone pairs

To estimate neuronal firing rates from calcium activity, deconvolution was performed using CASCADE, a supervised neural network–based algorithm trained on ground-truth electrophysiological recordings (*RRID:SCR_005861*) (Rupprecht et al., 2021). This approach provides temporally precise estimates of spiking activity from calcium fluorescence traces. Estimated firing rate traces were subsequently Gaussian-smoothed across time using a 2-s kernel and z-scored within each session to allow comparisons across neurons and sessions. For population-level analyses, activity vectors were constructed by averaging estimated firing rates across 1-s time windows for each neuron (overlap 0.5 s). These population vectors were used to compute similarity metrics. To quantify the consistency and temporal structure of population responses, we computed population similarity matrices across time points by correlating population activity vectors across independent repetitions of the same tone (within-tone comparisons) as well as across different tone identities (cross-tone comparisons). For all tone comparisons, similarity matrices were not symmetrized, thereby preserving potential asymmetries arising from differences in the order of the stimuli across repetitions (e.g., 11 vs 3 kHz or 3 vs 11 kHz), tone identity, and/or stimulus-evoked population dynamics. All similarity matrices were indexed by time relative to stimulus onset (−10 to 30 s), with tone presentation occurring from 0 to 20 s. In all graphs we plotted the earlier tone presentation on the y axis, and the later one on the x axis. To analyze population similarity of tone pairs during stimulus onset, the population activity vector correlations occurring during the first 5 s were computed for each animal. This analysis focused on tone pairs relative to the CS+ in each experimental group.

Classification of subnetwork tone-response patterns

Stimulus response vectors

To identify subsets of neurons with shared activity patterns in response to auditory tones, fluorescence traces from individual cells were detrended using the 10th percentile value to remove slow baseline drifts and z-scored to normalize activity levels across cells. For each session (12 tone presentations), 40-s segments of the calcium trace were extracted around each tone period—comprising 10 s before tone onset (baseline), 20 s during tone presentation, and 10 s after tone offset. Segments were interpolated to a uniform set of timestamps to ensure consistent

temporal sampling across trials. For each tone frequency (3, 7, 11, and 15 kHz), baseline activity was subtracted and traces were averaged across repetitions. The four mean tone responses were concatenated in ascending frequency order to form a single composite response vector for each cell. All vectors from a session were compiled into a dataset for subsequent analyses.

Mutual information matrix

For each pair of cells, we computed the mutual information (MI) between their stimulus response vectors to quantify the statistical dependence between their responses. MI measures the reduction in uncertainty about one variable given knowledge of another. In this context, low MI values indicate that the stimulus response vectors of two cells are largely independent, whereas high MI values indicate shared information or coordinated response structure. The MI values for all cell pairs were assembled into a symmetric matrix (n_cells × n_cells), with each entry representing the information shared between a given pair of cells. Separate MI matrices were computed for each experimental group (non-shock, CS+3, CS+15), pooling cells across recording days within each group. In what follows, "MI matrix" refers to one of these matrices.

Adjacency matrix construction

To transform the MI matrix into a network representation, we applied a summed top-k adjacency procedure. For each cell (matrix row), connections to other cells were ranked by MI strength, and only the top-k values were retained. Each retained connection was weighted by the reciprocal of its rank (1 / rank), favoring the strongest associations. The resulting matrix was symmetrized to represent an undirected network of functional connectivity (if cell A is functionally connected to cell B then this enforces that B is also functionally connected to A).

Spectral clustering

The adjacency matrix was analyzed using spectral clustering to identify functional subnetworks. Spectral clustering converts the adjacency matrix into a graph Laplacian, computes its eigenvalues and eigenvectors, and projects the data into a low-dimensional space in which clusters are more separable. K-means clustering was then applied to the eigenvector representations to delineate groups of highly connected neurons. This approach is well suited for detecting non-linear or irregular structures in neuronal co-activity networks.

Hyperparameter selection

Two hyperparameters were determined: the top-k value and the number of clusters. We evaluated a range of top-k values (2 to 100) and, for each, computed the first 30 eigenvalues of the normalized graph Laplacian obtained during spectral clustering. Eigenvalues were averaged across top-k values to yield a single representative spectrum. The optimal cluster number was then determined using the eigengap method, defined as the largest gap between consecutive eigenvalues. The top-k values associated with this solution were averaged to obtain the final k.

Positively and negatively modulated subnetworks

Because MI is non-directional, it cannot distinguish between positively and negatively correlated responses. To separate these responses, Spearman's correlation was computed for all pairs of cells within the cluster, and only the sign of each correlation was retained, resulting in a sign matrix of the same dimensions as the MI matrix. A signed MI matrix was then created by element-wise multiplication of the sign matrix by the MI matrix. The signed MI matrix was then used as input for another round of clustering, as previously described. Briefly, this involved generating a top-k adjacency matrix, computing eigenvalues from the normalized Laplacian, determining the optimal number of clusters using the eigengap method, and assigning cluster labels. Based on this procedure, one cluster was separated into three sub-clusters, while all other clusters were separated into two sub-clusters.

Global labels

Since our clustering method begins with a MI matrix, and each group has its own MI matrix and associated subclusters and subcluster labels, we performed a final cluster labeling step that allowed for comparison across groups. The results are such that, if a set of cells in group A have been assigned cluster label L, then the set of cells in group B having similar response patterns will be assigned the same cluster label L, allowing for comparison across groups.

Cluster Stability

To assess the stability and reorganization of functional clusters across days, we tracked registered cells that were detected on at least two of the three imaging sessions (day 1, day 15, and day 30). For each pair of days (A, B) and each cluster label L on day A, we quantified label stability ("percent_same") — the percentage of cells that had label L on day A and retained the same label on day B:

$$\text{percent\_same}(L, A, B) = \frac{n_{\text{same}}}{n_L} \times 100$$

where n_L is the number of cells assigned to label L on day A and n_{same} is the number of those cells that were assigned the same label on day B. Higher percent_same values indicate that a larger fraction of cells maintained the same cluster identity across days, suggesting that the corresponding functional cell assemblies were stable over time. Conversely, lower percent_same values reflect cluster reorganization, where individual cells changed their cluster membership between sessions.

Statistics

Behavioral and neural data were analyzed using one- or two-way analyses of variance (ANOVAs) with repeated measures. For one-way ANOVAs, tone frequency was treated as the repeated factor; for two-way ANOVAs, a between-subjects group factor (e.g., experimental vs. control) was additionally included. When ANOVAs revealed significant main effects or interactions, post hoc comparisons were performed using Tukey's multiple-comparisons test. Data normality was assessed using the Shapiro–Wilk test and homogeneity of variances with the Brown–Forsythe test prior to statistical analysis. When assumptions were violated, Friedman tests were used for repeated-measures analyses, and Kruskal–Wallis tests were used for comparisons without repeated measures. Rank-based tests were followed by Dunn's post hoc multiple-comparison tests. Statistical significance was evaluated using a two-sided alpha level of 0.05. For analyses involving multiple cell-identity comparisons, p values were adjusted using the Benjamini–Hochberg false discovery rate procedure.

Histology

At the conclusion of experiments, mice were deeply anesthetized with isoflurane and transcardially perfused with phosphate-buffered saline (PBS), followed by 4% paraformaldehyde (PFA). Extracted brains were post-fixed in 4% PFA for 24 hours and subsequently transferred to a 20% sucrose solution containing sodium azide at 4 °C for an additional 24 hours. Frozen brains were coronally sectioned at 50 μm using a cryostat, and sections were mounted on Superfrost Plus microscope slides (Fisher Scientific). Histological verification was performed to confirm GCaMP expression and accurate lens placement within the PL.

Supplemental Figures

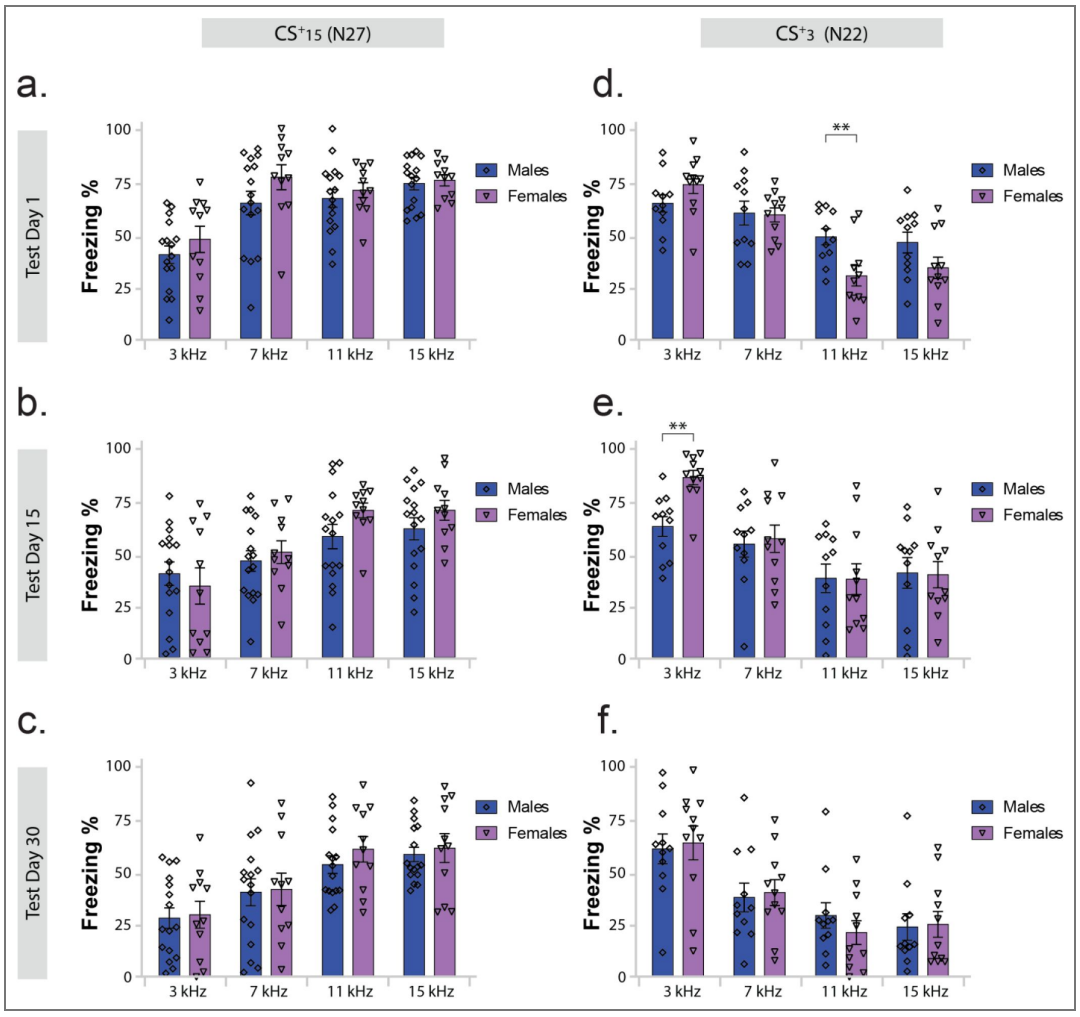

Figure S1. Sex differences in behavior. a-c, Sex differences in CS⁺15-trained mice. No sex differences were observed on any testing day (effect of sex: day 1, $F(1,25) = 1.52$, $p = 0.23$; day 15, $F(1,25) = 0.55$, $p = 0.467$; day 30, $F(1,25) = 0.17$, $p = 0.681$). On all days, there was a significant main effect of frequency, indicating successful learning in both males and females (day 1, $F(3,75) = 33.46$, $p < 0.001$; day 15, $F(3,75) = 26.32$, $p < 0.001$; day 30, $F(3,75) = 51.47$, $p < 0.001$). No sex × frequency interactions were detected on any testing day ($p > 0.05$). Post hoc comparisons showed that on day 1, mice discriminated 3 kHz from 7, 11 and 15 kHz ($p < 0.05$), but generalized among the higher frequencies ($p > 0.05$). On days 15 and 30, mice discriminated 3 kHz from 7, 11 and 15 kHz ($p < 0.05$) and 7 kHz from 11 and 15 kHz ($p < 0.05$), while generalization persisted between 11 and 15 kHz ($p > 0.05$). **d-f, Sex differences in CS⁺3-trained mice.** No main effect of sex was observed on any testing day (day 1, $F(1,20) = 1.24$, $p = 0.279$; day 15, $F(1,20) = 0.78$, $p = 0.388$; day 30, $F(1,20) = 0.003$, $p = 0.956$). A significant main effect of frequency was present across all days, indicating learning in both sexes (day 1, $F(3,60) = 40.65$, $p < 0.001$; day 15, $F(3,60) = 30.17$, $p < 0.001$; day 30, $F(3,60) = 35.19$, $p < 0.001$). A significant sex × frequency interaction was observed on day 1 ($F(3,60) = 6.91$, $p < 0.001$) and day 15 ($F(3,60) = 3.55$, $p < 0.02$), reflecting sex-specific differences at 11 kHz on day 1 and at 3 kHz on day 15. These effects were not consistent across frequencies or present on day 30 ($F(3,60) = 0.76$, $p = 0.522$), likely reflecting increased behavioral variability in females, rather than stable sex differences. Significant Tukey multiple comparisons are denoted by asterisks, ** $p < 0.01$.

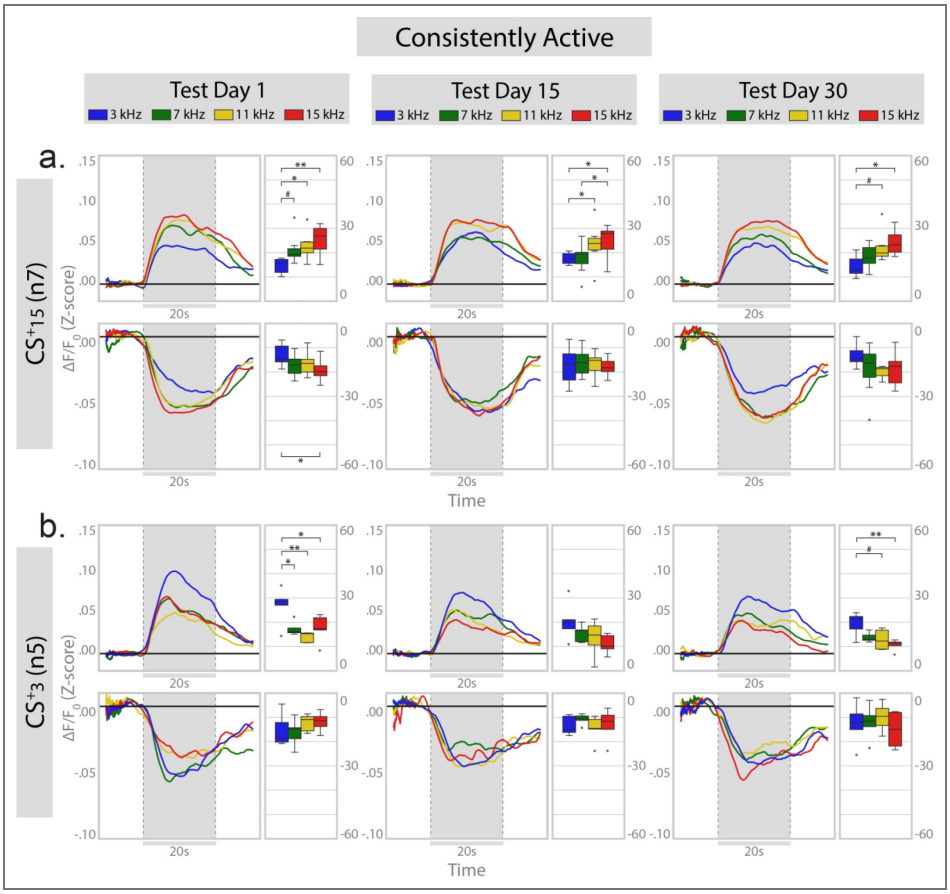

Figure S2. Population activity of consistently active neurons.

a-b, Population activity of consistently active neurons from animals trained with a 15 kHz CS+ (a) or a 3 kHz CS+ (b). Upper panels show positively tone-responsive neurons and lower panels show negatively tone-responsive neurons. Adjacent boxplots display areas under the population response curves. Consistently active positive responders exhibit graded responses across test days (day 1 to day 30). $CS^{+}15$: day 1, $F(3,18) = 6.072$, $p < 0.005$; day 15, $F(3,18) = 4.492$, $p = 0.016$; day 30, $F(3,18) = 4.599$, $p = 0.015$; $CS^{+}3$: Day 1, $F(3,12) = 9.304$, $p = 0.002$; Day 15, $F(3,12) = 3.022$, $p = 0.072$; day 30, $F(3,12) = 5.26$, $p = 0.015$; controls: day 1, $F(3,9) = 5.343$, $p < 0.02$; day 15, $F(3,8) = 0.399$, $p = 0.758$; day 30, $F(3,8) = 0.19$, $p = 0.90$, *p < 0.05; **p < 0.01.

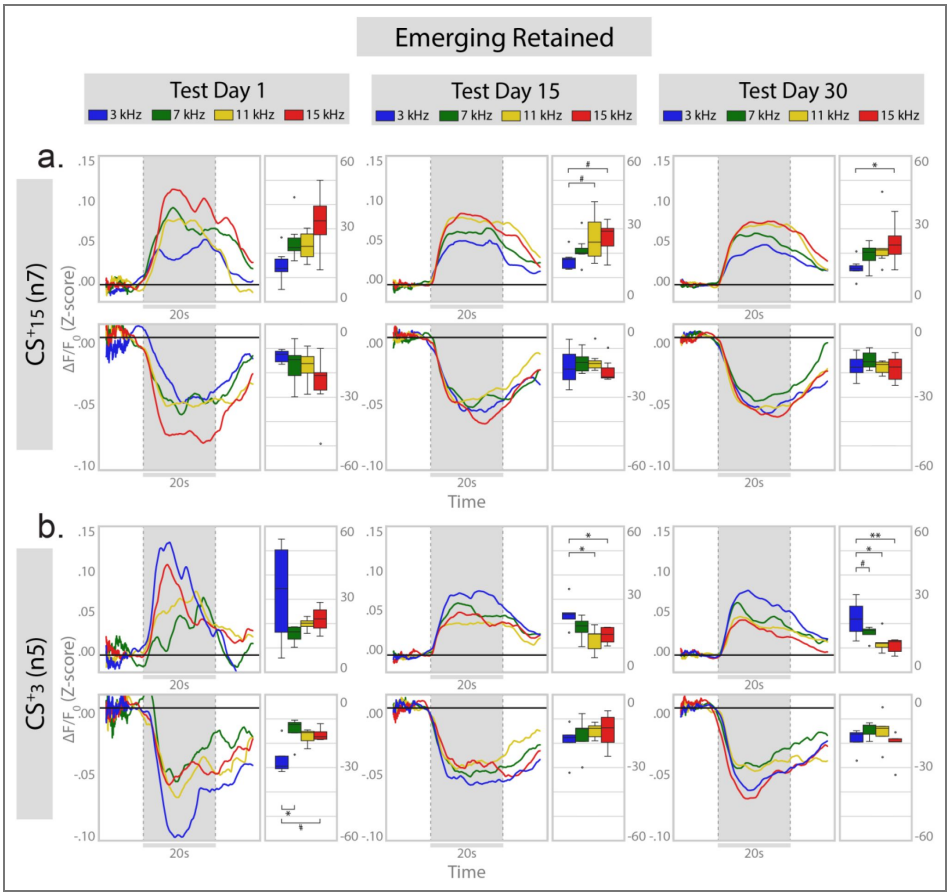

Figure S3. Population activity of emerging-retained neurons.

Emerging-retained neurons become part of the active ensemble after conditioning and are eded after that. **a-b,** Population activity of emerging-retained neurons from animals trained with a 15 kHz CS+ (a) or a 3 kHz CS+ (b). Upper panels show positively tone-responsive neurons and lower panels show negatively tone-responsive neurons. Adjacent boxplots display AUC. Emerging-retained neurons show greater variability on day 1 compared with later test days. Only positive responders show significant graded valence (CS^+15: Day 1, $F(3,16) = 4.587$, $p = 0.017$; day 15, $F(3,18) = 4.749$, $p = 0.013$; day 30, $F(3,18) = 4.748$, $p = 0.013$; CS^+3: day 1, $F(3,9) = 1.684$, $p = 0.239$; day 15, $F(3,12) = 7.728$, $p = 0.004$; day 30, $F(3,12) = 5.63$, $p = 0.012$). Significant post hoc multiple comparisons noted in the Figure with asterisks. *p < 0.05; **p < 0.01.

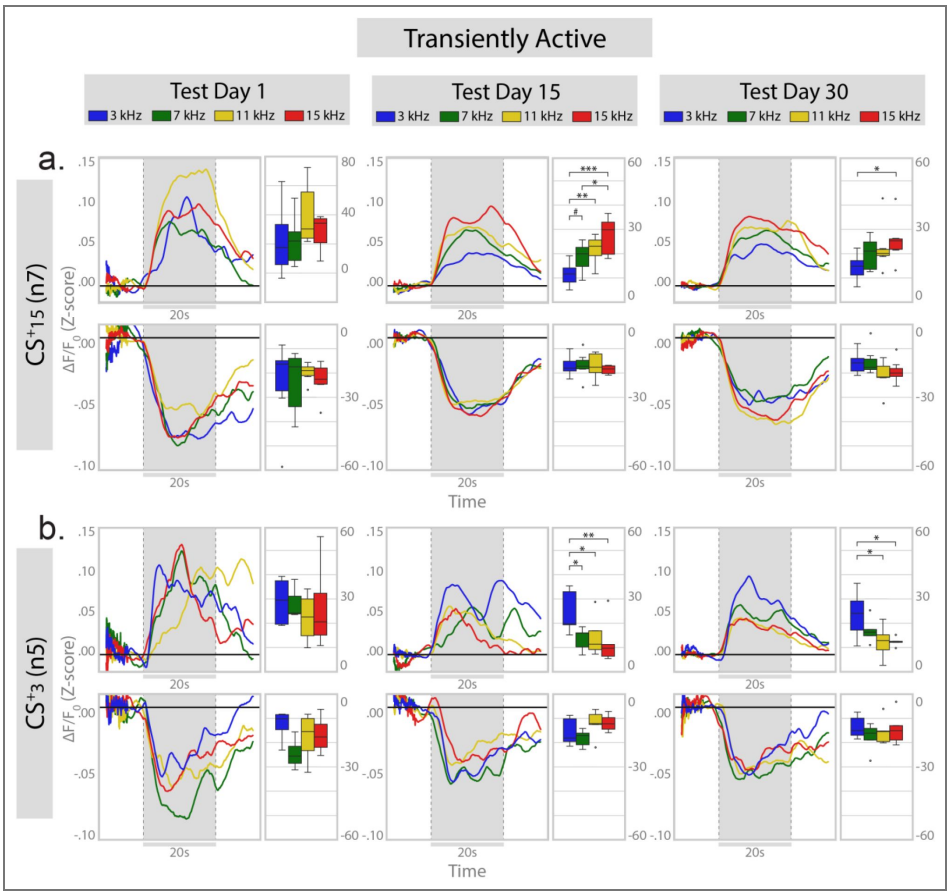

Figure S4. Population activity of transiently active neurons.

Transiently active neurons were active on only a single test day. **a-b,** Population activity of transiently active neurons from animals trained with a 15 kHz CS+ (a) or a 3 kHz CS+ (b). Upper panels show positively tone-responsive neurons and lower panels show negatively tone-responsive neurons. Adjacent boxplots display areas under the population response curves. Transiently active neurons did not show graded population responses on day 1; graded responses emerged by day 15 and were maintained through day 30 in positive sound responder neurons. *p < 0.05; **p < 0.01.

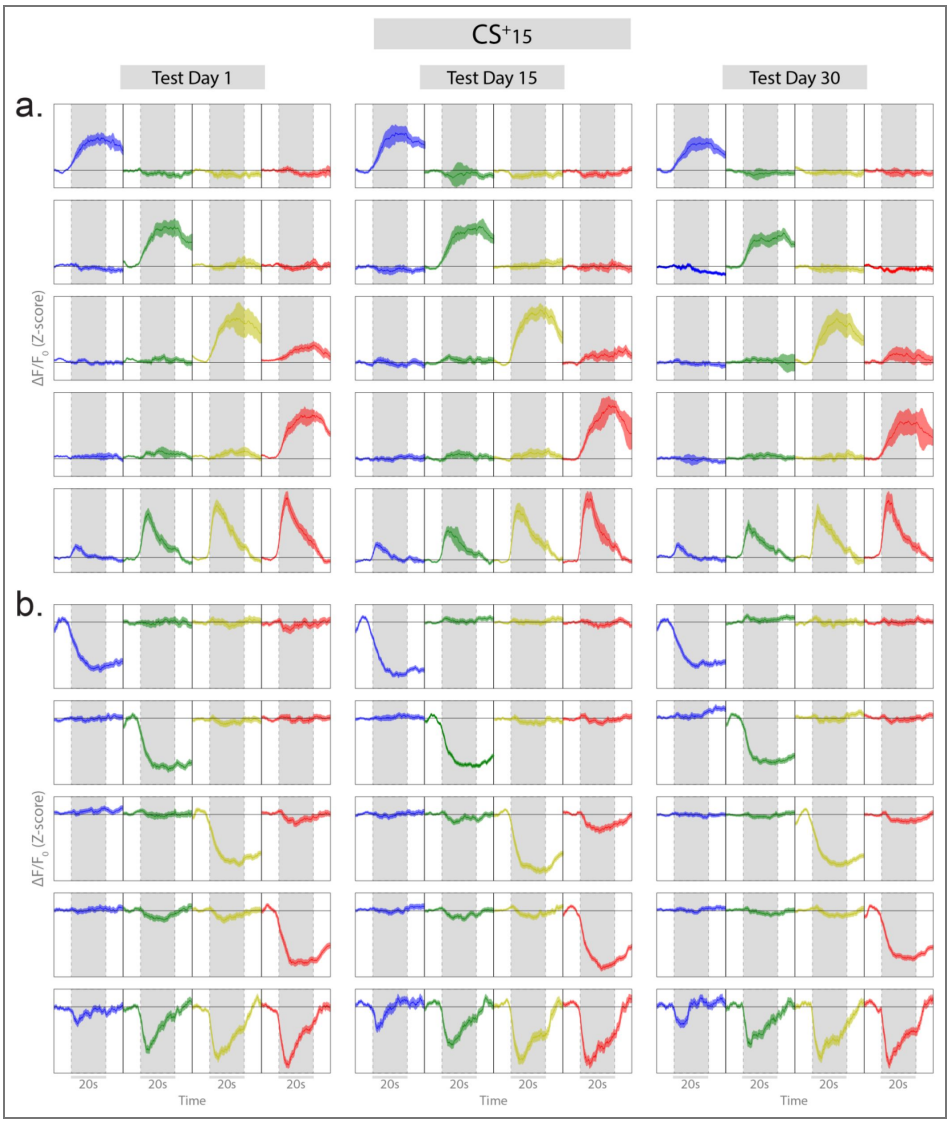

Figure S5. Clustering of PL subnetworks based on signed mutual information for the CS+ 15 kHz group per testing session.

a-b, Average stimulus-aligned population responses for clusters showing positive (a) or negative (b) modulation to individual tones (3, 7, 11, or 15 kHz, upper panels) or graded emotional tuning (a-b, bottom panels).

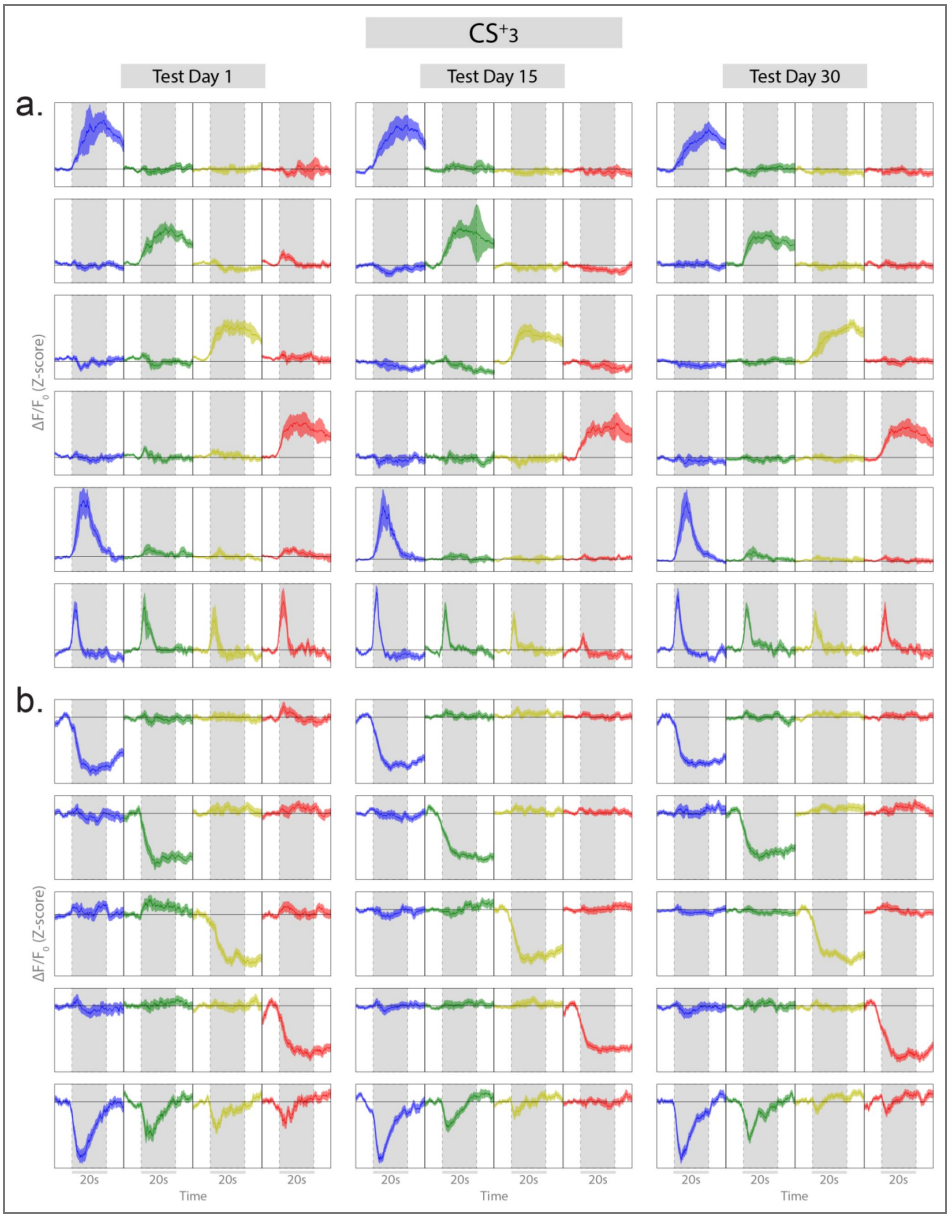

Figure S6. Clustering of PL subnetworks based on signed mutual information for the CS+ 3 kHz group per testing session.

a-b, Average stimulus-aligned population responses for clusters showing positive (a) or negative (b) modulation to individual tones (3, 7, 11, or 15 kHz, upper panels) or graded emotional tuning (a-b, bottom panels).

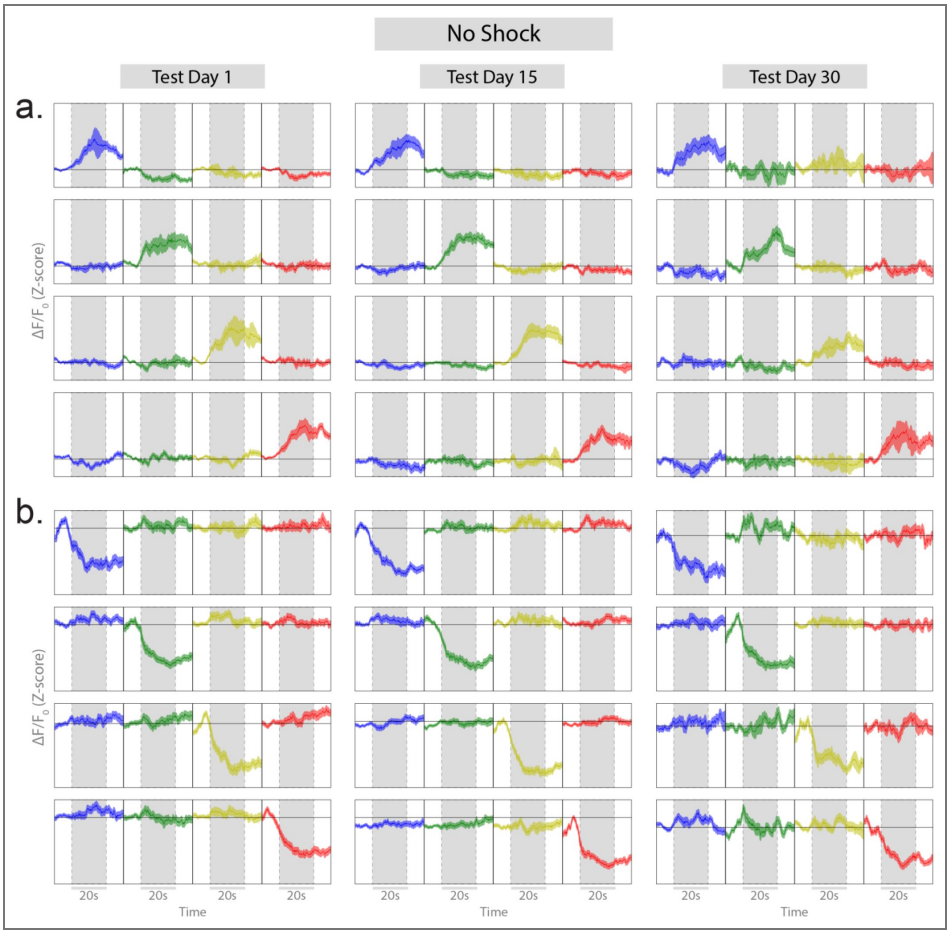

Figure S7. Clustering of PL subnetworks based on signed mutual information for the no shock control per testing session.

a-b, Average stimulus-aligned population responses for clusters showing positive (a) or negative (b) modulation to individual tones (3, 7, 11, or 15 kHz).

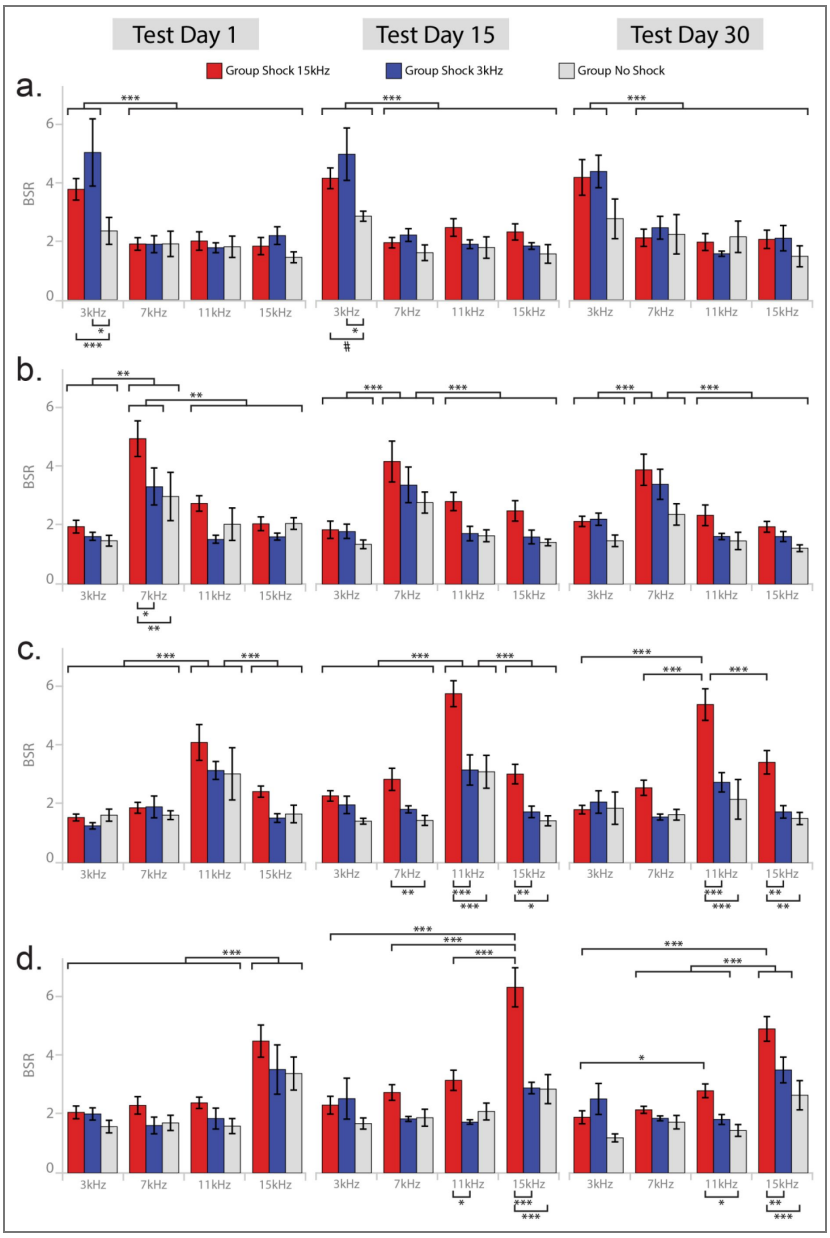

Figure S8.

a–d, Baseline-to-stimulus firing rate ratio (BSR) illustrating changes in positively responding neurons within tone-specific clusters shown in Fig. 5, for the CS+15 (red), CS+3 (blue), and control groups. **(a)** BSR of neurons in clusters primarily responsive to 3 kHz (Fig. 5c.1), **(b)** 7 kHz (Fig. 5c.2), **(c)** 11 kHz (Fig. 5c.3), and **(d)** 15 kHz (Fig. 5c.4). ANOVA results are reported in Table S3; Tukey's multiple-comparison tests indicate significance ($p < 0.05$; $p < 0.01$; $p < 0.001$).

Group	Type	Modulation	Test	Statistic	df	p
CS+15	Consistent	Positive	Test Day 1	F = 6.11	3, 18	0.005 **
			Test Day 15	F = 6.29	3, 18	0.004**
			Test Day 30	F = 4.43	3, 18	0.017*
		Negative	Test Day 1	F = 3.76	3, 18	0.030*
			Test Day 15	F = 0.31	3, 18	0.817
			Test Day 30	F = 1.80	3, 18	0.183
	Emerged Retained	Positive	Test Day 1	F = 3.18	3, 13	0.060
			Test Day 15	F = 3.24	3, 18	0.046 *
			Test Day 30	χ^2 = 8.66	3	0.0.4*
		Negative	Test Day 1	F = 1.89	3, 17	0.170
			Test Day 15	F = 1.45	3, 18	0.263
			Test Day 30	F = 2.63	3, 18	0.308
	Transiently Active	Positive	Test Day 1	F = 1.18	3, 17	0.332
			Test Day 15	F = 11.17	3, 18	<0.001 ***
			Test Day 30	F = 3.87	3, 18	0.030 *
		Negative	Test Day 1	F = 0.39	3, 17	0.764
			Test Day 15	F = 0.30	3, 18	0.827
			Test Day 30	F = 2.08	3, 18	0.139
CS+3	Consistent	Positive	Test Day 1	F = 9.32	3, 12	0.002 **
			Test Day 15	F = 3.18	3, 12	0.063
			Test Day 30	F = 5.39	3, 12	0.014 *
		Negative	Test Day 1	F = 1.28	3, 12	0.326
			Test Day 15	F = 0.53	3, 12	0.668
			Test Day 30	F = 0.47	3, 12	0.709
	Emerged Retained	Positive	Test Day 1	F = 1.49	3, 6	0.309
			Test Day 15	F = 6.18	3, 9	0.014 *
			Test Day 30	F = 9.51	3, 9	0.004 **
		Negative	Test Day 1	F = 4.75	3, 8	0.035
			Test Day 15	F = 0.65	3, 12	0.595
			Test Day 30	χ^2 = 1.39	3	0.356
	Transiently Active	Positive	Test Day 1	F = 0.28	3, 8	0.839
			Test Day 15	F = 6.45	3, 12	0.008 **
			Test Day 30	F = 5.49	3, 12	0.013 *
		Negative	Test Day 1	F = 1.12	3, 9	0.398
			Test Day 15	F = 1.57	3, 11	0.253
			Test Day 30	F = 0.426	3, 12	0.738

Table S1. Statistics corresponding to consistently active, emerging-retained, or transiently active neurons

Group	Cluster	% Stable D1-D15	p.(Test D1-TestD15)	% Stable D15-D30	p. (Test D15-TestD30)	% Stable D1-D30	p.(TestD1-TestD30)
CS+15	Common B1	13.5	0.391	5.0	1.000	9.3	0.870
	Common B2	3.4	1.000	7.1	1.000	8.2	1.000
	Common B3	11.1	0.576	13.4	0.391	19.2	0.124
	Common B4	15.5	0.391	15.1	0.391	20.9	0.121
	Common C1	19.3	0.079	11.9	0.391	12.9	0.576
	Common C2	10.4	0.765	10.0	0.661	10.3	0.576
	Common C3	11.9	0.697	20.2	0.334	17.8	0.697
	Common C4	20.2	0.011*	11.8	1.000	10.3	1.000
	Graded A	28.2	0.000***	22.8	0.000***	21.3	0.001***
	Graded B	11.1	0.034*	5.5	0.576	5.9	0.576
	Graded C	0.0	1.000	0.0	1.000	0.0	1.000
	Graded D	0.0	1.000	0.0	1.000	0.0	1.000
	Graded E	0.0	1.000	0.0	1.000	0.0	1.000
CS+3	Common B1	23.9	0.014*	16.5	0.460	19.0	0.476
	Common B2	13.4	0.708	6.5	0.982	11.9	0.792
	Common B3	10.8	0.852	9.8	0.939	21.1	0.185
	Common B4	10.4	0.756	7.7	0.982	19.6	0.521
	Common C1	20.9	0.230	20.6	0.074	13.0	0.708
	Common C2	11.5	0.714	8.0	0.579	8.6	0.592
	Common C3	9.7	0.739	5.3	0.982	10.3	0.846
	Common C4	8.9	0.872	7.6	0.972	12.8	0.592
	Graded A	0.0	1.000	0.0	1.000	0.0	1.000
	Graded B	0.0	1.000	0.0	1.000	0.0	1.000
	Graded C	21.2	0.008**	17.9	0.026*	25.0	0.008**
	Graded D	9.1	0.303	40.0	0.003**	14.3	0.303
	Graded E	12.9	0.047*	15.4	0.047*	8.3	0.589
No Shock	Common B1	7.8	1.000	8.3	1.000	22.2	1.000
	Common B2	11.5	1.000	10.0	1.000	0.0	1.000
	Common B3	21.6	1.000	16.7	1.000	40.0	1.000
	Common B4	11.5	1.000	5.9	1.000	0.0	1.000
	Common C1	10.0	1.000	7.1	1.000	0.0	1.000
	Common C2	5.5	1.000	21.4	1.000	21.4	1.000
	Common C3	12.0	1.000	4.0	1.000	33.3	1.000
	Common C4	6.4	1.000	22.2	1.000	0.0	1.000
	Graded A	0.0	1.000	0.0	1.000	0.0	1.000
	Graded B	0.0	1.000	0.0	1.000	0.0	1.000
	Graded C	0.0	1.000	0.0	1.000	0.0	1.000
	Graded D	0.0	1.000	0.0	1.000	0.0	1.000
	Graded E	0.0	1.000	0.0	1.000	0.0	1.000

Table S2. Statistics corresponding to cell identity across time for all neurons clustered using MI

Cluster	Groups	Test	Statistic	df	p
Statistics for clusters Responding to individual frequencies (Fig. 5c and Fig. S8)					
Fig. 5c, c.1 Maximal response at 3 kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =1.49	2, 15	0.258
			Frequency: F=30.51	3, 6	0.001***
			Interaction: 4.50	6, 45	0.001***
		Test Day 15	Group: F =4.59	2, 15	0.03*
			Frequency: F=31.99	3, 6	0.001***
			Interaction: 2.28	6,4 5	0.053
		Test Day 30	Group: F =0.36	3, 13	0.705
			Frequency: F=30.02	3, 6	0.001***
			Interaction: 2.78	6, 39	0.024*
Fig. 5c, c.2 Maximal response at 7kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =2.29	2,15	0.135
			Frequency: F=30.18	3, 6	0.001***
			Interaction: 2.62	6, 45	0.029*
		Test Day 15	Group: F =3.50	2, 15	0.057
			Frequency: F=27.46	3, 6	0.001***
			Interaction: 0.99	6, 45	0.441
		Test Day 30	Group: F =3.51	2,13	0.061
			Frequency: F=24.64	3, 6	0.001***
			Interaction: 1.057	6, 3 9	0.404
Fig. 5c, c.3 Maximal response at 11 kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =1.45	2, 15	0.266
			Frequency: F=20.44	3, 6	0.001***
			Interaction: 0.912	6, 4 5	0.495
		Test Day 15	Group: F =15.47	2, 15	0.001***
			Frequency: F=38.53	3, 6	0.001***
			Interaction: 3.40	6,4 5	0.007**
		Test Day 30	Group: F =11.46	2, 13	0.001***
			Frequency: F=14.20	3, 6	0.001***
			Interaction: 6.12	6, 3 9	0.001***
Fig. 5c, c.4 Maximal response at 15 kHz	CS+15 CS+3 No shock	Test Day 1	Group: F =2.71	3, 15	0.099
			Frequency: F=22.28	3, 6	0.001***
			Interaction: 0.38	6,4 5	0.89
		Test Day 15	Group: F =7.64	2,1 5	0.005**
			Frequency: F=33.86	3, 6	0.001***
			Interaction: 10.58	6, 45	0.001***
		Test Day 30	Group: F =8.01	2, 13	0.005**
			Frequency: F=30.05	3, 6	0.001***
			Interaction: 4.15	6, 39	0.003**
Statistics for clusters showing graded valence (Fig. 6)					
Graded cluster Fig. 6a	Experimental CS+15	Test Day 1, 15, and 30	Day: F =2.89	2, 12	0.10
			Frequency: F=29.28	3, 12	0.001***
			Interaction: 0.351	6,36	0.905
Graded cluster Fig. 6c	Experimental CS+3	Test Day 1	Day F =0.467	2, 8	0.643
			Frequency: F=53.37	3,12	0.001***
			Interaction: 2.63	6, 24	0.042*

Table S3. Statistics corresponding the baseline/stimulus firing rate changes (BSR) for tone-selective responders

Data availability

Datasets available at: *https://data.mendeley.com/datasets/9yyn63g346/1*.

Acknowledgements

This works has been funded by NSF (NSF/IOS 2303305 to IAM), NIH (R01 MH123260-01 to IAM; RISE GMO60655 to MRL).

Additional information

Author Contributions

MEN developed method to identify subnetworks, wrote code for analysis, and contributed to writing the manuscript, PMO collected data, conducted behavioral and in vivo recording experiments, and contributed to writing the manuscript, MRL contributed to experimental design, collected data, conducted behavioral and in vivo recording experiments. IAM supervised design, experiments, analysis, and writing of the manuscript.

Funding

Funder	Grant reference number	Author
National Science Foundation (NSF)	2303305	Isabel A Muzzio
National Institute of Mental Health and Neurosciences (NIMHANS)	MH123260-01	Isabel A Muzzio

Author ORCID iDs

Isabel A Muzzio: https://orcid.org/0000-0003-0156-0088

References

Aschauer D. F., Eppler J. B., Ewig L., Chambers A. R., Pokorny C., Kaschube M., Rumpel S (2022) Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization. *Cell Rep* **38**:110340 https://doi.org/10.1016/j.celrep.2022.110340 | PubMed

Bontempi B., Jaffard R., Destrade C (1996) Differential temporal evolution of post-training changes in regional brain glucose metabolism induced by repeated spatial discrimination training in mice: visualization of the memory consolidation process?. *Eur J Neurosci* **8**:2348-2360 https://doi.org/10.1111/j.1460-9568.1996.tb01198.x | PubMed

Burgos-Robles A., Vidal-Gonzalez I., Quirk G. J (2009) Sustained conditioned responses in prelimbic prefrontal neurons are correlated with fear expression and extinction failure. *J Neurosci* **29**:8474-8482 https://doi.org/10.1523/JNEUROSCI.0378-09.2009 | PubMed

Deitch D., Rubin A., Ziv Y (2021) Representational drift in the mouse visual cortex. *Curr Biol* **31**:4327-4339.e4326. https://doi.org/10.1016/j.cub.2021.07.062 | PubMed

DeNardo L. A., Liu C. D., Allen W. E., Adams E. L., Friedmann D., Fu L., Guenthner C. J., Tessier-Lavigne M., Luo L (2019) Temporal evolution of cortical ensembles promoting remote memory retrieval. *Nat Neurosci* **22**:460-469 https://doi.org/10.1038/s41593-018-0318-7 | PubMed

Do-Monte F. H., Quinones-Laracuente K., Quirk G. J (2015) A temporal shift in the circuits mediating retrieval of fear memory. *Nature* **519**:460-463 https://doi.org/10.1038/nature14030 | PubMed

Frankland P. W., Bontempi B (2005) The organization of recent and remote memories. *Nat Rev Neurosci* **6**:119-130 https://doi.org/10.1038/nrn1607 | PubMed

Gallego J. A., Perich M. G., Chowdhury R. H., Solla S. A., Miller L. E (2020) Long-term stability of cortical population dynamics underlying consistent behavior. *Nat Neurosci* **23**:260-270 https://doi.org/10.1038/s41593-019-0555-4 | PubMed

Gu X., Johansen J. P (2025) Prefrontal encoding of an internal model for emotional inference. *Nature* **643**:1044-1056 https://doi.org/10.1038/s41586-025-09001-2 | PubMed

Hockley A., Malmierca M. S (2024) Auditory processing control by the medial prefrontal cortex: A review of the rodent functional organisation. *Hear Res* **443**:108954 https://doi.org/10.1016/j.heares.2024.108954 | PubMed

Hoshi A., Hirayama Y., Saito F., Ishiguro T., Suetani H., Kitajo K (2023) Spatiotemporal consistency of neural responses to repeatedly presented video stimuli accounts for population preferences. *Sci Rep* **13**:5532 https://doi.org/10.1038/s41598-023-31751-0 | PubMed

Iqbal J., Kim S., Lawal S., Shah A., Punepalle L., Sanghvi H., Gallagher A., Wilson A., Xu B., Shrestha P (2026) A prelimbic molecular clock of protein synthesis for memory persistence. *bioRxiv* https://doi.org/10.64898/2026.01.02.697403 | PubMed

Josselyn S. A., Tonegawa S (2020) Memory engrams: Recalling the past and imagining the future. *Science* **367** https://doi.org/10.1126/science.aaw4325 | PubMed

Kato H. K., Asinof S. K., Isaacson J. S (2017) Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron* **95**:412-423.e414. https://doi.org/10.1016/j.neuron.2017.06.019 | PubMed

Kitamura T., Ogawa S. K., Roy D. S., Okuyama T., Morrissey M. D., Smith L. M., Redondo R. L., Tonegawa S (2017) Engrams and circuits crucial for systems consolidation of a memory. *Science* **356**:73-78 https://doi.org/10.1126/science.aam6808 | PubMed

Kupke J., Oliveira A. M. M (2025) The molecular and cellular basis of memory engrams: Mechanisms of synaptic and systems consolidation. *Neurobiol Learn Mem* **219**:108057 https://doi.org/10.1016/j.nlm.2025.108057 | PubMed

Lopez M. R., Wasberg S. M. H., Gagliardi C. M., Normandin M. E., Muzzio I. A (2024) Mystery of the memory engram: History, current knowledge, and unanswered questions. *Neurosci Biobehav Rev* **159**:105574 https://doi.org/10.1016/j.neubiorev.2024.105574 | PubMed

Mau W., Hasselmo M. E., Cai D. J (2020) The brain in motion: How ensemble fluidity drives memory-updating and flexibility. *eLife* **9** https://doi.org/10.7554/eLife.63550 | PubMed

Moscovitch M., Nadel L. (1998) Consolidation and the hippocampal complex revisited: in defense of the multiple-trace model. *Curr Opin Neurobiol* **8**:297-300 https://doi.org/10.1016/s0959-4388(98)80155-4 | PubMed

Nadel L., Samsonovich A., Ryan L., Moscovitch M (2000) Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**:352-368 https://doi.org/10.1002/1098-1063(2000)10:4<352::AID-HIPO2>3.0.CO;2-D | PubMed

Phillips R. G., LeDoux J. E (1992) Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behav Neurosci* **106**:274-285 https://doi.org/10.1037//0735-7044.106.2.274 | PubMed

R. Quian Quiroga, Panzeri S. (2009) Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci* **10**:173-185 https://doi.org/10.1038/nrn2578 | PubMed

Rao-Ruiz P., Visser E., Mitric M., Smit A. B., van den Oever M. C (2021) A Synaptic Framework for the Persistence of Memory Engrams. *Front Synaptic Neurosci* **13**:661476 https://doi.org/10.3389/fnsyn.2021.661476 | PubMed

Refaeli R., Kreisel T., Groysman M., Adamsky A., Goshen I (2023) Engram stability and maturation during systems consolidation. *Curr Biol* **33**:3942-3950.e3943. https://doi.org/10.1016/j.cub.2023.07.042 | PubMed

Rosas-Vidal L. E., Naskar S., Mayo L. M., Perini I., Masroor R., Altemus M., Ramos-Medina L., Zaidi S. D., Engelbrektsson H., Jagasia P., *et al.* (2025) Prefrontal correlates of fear generalization during endocannabinoid depletion. *J Clin Invest* **135** https://doi.org/10.1172/JCI179881 | PubMed

Rupprecht P., Carta S., Hoffmann A., Echizen M., Blot A., Kwan A. C., Dan Y., Hofer S. B., Kitamura K., Helmchen F., *et al.* (2021) A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nat Neurosci* **24**:1324-1337 https://doi.org/10.1038/s41593-021-00895-5 | PubMed

Sheintuch L., Rubin A., Brande-Eilat N., Geva N., Sadeh N., Pinchasof O., Ziv Y (2017) Tracking the Same Neurons across Multiple Days in Ca(2+) Imaging Data. *Cell Rep* **21**:1102-1115 https://doi.org/10.1016/j.celrep.2017.10.013 | PubMed

Sierra-Mercado D., Padilla-Coreano N., Quirk G. J (2011) Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear. *Neuropsychopharmacology* **36**:529-538 https://doi.org/10.1038/npp.2010.184 | PubMed

Sotres-Bayon F., Quirk G. J (2010) Prefrontal control of fear: more than just extinction. *Curr Opin Neurobiol* **20**:231-235 https://doi.org/10.1016/j.conb.2010.02.005 | PubMed

Stujenske J. M., O'Neill P. K., Fernandes-Henriques C., Nahmoud I., Goldburg S. R., Singh A., Diaz L., Labkovich M., Hardin W., Bolkan S. S., *et al.* (2022) Prelimbic cortex drives discrimination of non-aversion via amygdala somatostatin interneurons. *Neuron* **110**:2258-2267.e2211. https://doi.org/10.1016/j.neuron.2022.03.020 | PubMed

Terranova J. I., Yokose J., Osanai H., Ogawa S. K., Kitamura T (2023) Systems consolidation induces multiple memory engrams for a flexible recall strategy in observational fear memory in male mice. *Nat Commun* **14**:3976 https://doi.org/10.1038/s41467-023-39718-5 | PubMed

Tonegawa S., Morrissey M. D., Kitamura T (2018) The role of engram cells in the systems consolidation of memory. *Nat Rev Neurosci* **19**:485-498 https://doi.org/10.1038/s41583-018-0031-2 | PubMed

Wang M. E., Wann E. G., Yuan R. K., Ramos Alvarez M. M., Stead S. M., Muzzio I. A (2012) Long-term stabilization of place cell remapping produced by a fearful experience. *J Neurosci* **32**:15802-15814 https://doi.org/10.1523/JNEUROSCI.0480-12.2012 | PubMed

Wehr M., Zador A. M (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**:442-446 https://doi.org/10.1038/nature02116 | PubMed

Zaki Y., Cai D. J (2024) Memory engram stability and flexibility. *Neuropsychopharmacology* **50**:285-293 https://doi.org/10.1038/s41386-024-01979-z | PubMed

Zikopoulos B., Barbas H (2006) Prefrontal projections to the thalamic reticular nucleus form a unique circuit for attentional mechanisms. *J Neurosci* **26**:7348-7361 https://doi.org/10.1523/JNEUROSCI.5511-05.2006 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors combine discriminative auditory fear conditioning with longitudinal in vivo calcium imaging to ask how prelimbic (PL) representations of learned and generalized threat evolve across recent and remote memory time points. Using two different CS+ frequencies and a no-shock control group, they report that PL population activity tracks graded behavioral generalization, that population similarity is highest for tones eliciting strong threat responding, and that distinct subnetworks can be identified that appear to encode tone-specific sensory features versus learned threat-related response structure.

To my knowledge, this may be the first study to comprehensively examine neural encoding of fear generalization in prelimbic cortex (PL). The manuscript is ambitious and technically interesting, and several aspects are potentially important. In particular, the suggestion that neurons showing graded, learning-related response patterns become selectively stabilized over time is intriguing. The inclusion of two CS+ training conditions and a no-shock control also strengthens the case that at least some of the reported effects are related to associative learning rather than simple sensory differences. However, in its current form, the manuscript does not yet fully support the strength of the conceptual claims. Several issues limit confidence in the interpretation, including the possibility that repeated testing itself contributes to changes across days, uncertainty about the relationship between neural activity and freezing behavior, limited quantitative documentation of longitudinal cell registration, and a number of problems in figure clarity and statistical framing. Overall, the study contains promising observations, but the claims should be narrowed, and several analyses or controls would be needed to fully support the proposed framework.

Detailed Comments

(1) A general concern is that the repeated test procedure itself may contribute to extinction. Because the animals are exposed to multiple CS frequencies across multiple test days, and each tone is presented three times per session, some of the reported changes in behavior and neural activity across days could reflect extinction or repeated nonreinforced retrieval rather than the passage of time per se. This is especially relevant given that the manuscript makes claims about recent versus remote representations and representational drift over 30 days. At a minimum, the authors should discuss this limitation explicitly and temper claims about time-dependent changes. Ideally, they would include a control group in which animals are tested only once or twice (e.g., at an early and later time point with fewer CS frequencies), or a reduced-frequency testing design that minimizes extinction while still allowing evaluation of recent versus remote memory.

(2) More generally, some of the reported learning-related neural differences may be driven by behavioral differences, particularly freezing, rather than by learning or generalization per se. For example, animals that freeze more to certain frequencies may show corresponding neural response differences simply because freezing alters PL activity. The authors should examine this possibility more directly. Analyses testing whether recorded cells encode freezing behavior, or whether tone frequency-related neural differences remain robust when comparing high- and low-freezing epochs, would help determine whether the reported effects reflect learned stimulus value rather than behavioral state differences.

(3) A central feature of the manuscript is the analysis of neural response properties over an extended period of time, up to 30 days after learning. However, aside from a brief mention in the Methods that spatial registration was used, the manuscript provides very little quantitative information about this critical aspect of the study. The paper would be strengthened by including explicit metrics describing longitudinal cell tracking, such as the number and proportion of ROIs retained across all sessions, distributions of spatial-footprint correlations or centroid distances across days, and representative examples of matched imaging fields over time. Without this information, it is difficult to assess how strongly the longitudinal claims are supported.

(4) The text states that "Figs. 1c and 1d show GCaMP6f expression in PL, representative calcium footprints, and activity traces". However, the figure as presented does not clearly show all of these elements, at least not in a way that matches the description in the Results. The correspondence between text and figure should be corrected.

(5) The labeling of Figure 2a is insufficient for interpretation. The legend states that the panel shows raster plots of sound responsiveness, but the axes and scaling are not clearly defined. It is not clear from the figure what the x-axis represents, whether the y-axis corresponds to

individual neurons, where the CS period occurs, or what the activity scale at the right denotes. Also, the term 'rasters' implies that spikes were analyzed. It seems that the spike inference approach (CASCADE) was only used for later analyses. Perhaps 'heat-plot' would be more accurate here? Generally, this figure should be annotated more clearly so that the reader can understand it without referring back to the Methods.

(6) In relation to Figure 3, the analysis of population-averaged responses across tone frequencies is useful, but the manuscript would be stronger with additional statistical analyses across time and across groups. For example, if the authors want to argue that learning induces graded changes in neural responses and that these evolve across time, they should directly compare within-group responses across days and also compare matched frequencies between the conditioned groups and the no-shock controls. These analyses would help establish whether the observed differences are genuinely learning dependent and whether they change significantly over time.

(7) The inclusion of two different CS+ frequencies and a no-shock control is a strength of the study and substantially improves the interpretation that graded neural responses are related to learning and generalization rather than to simple sensory processing or passage of time. That said, I am not entirely comfortable with the use of the term "inference" throughout the manuscript. What is being measured here appears closer to sensory generalization than inference in a stronger cognitive sense. The current task does not clearly require that animals infer hidden structure or stimulus value through abstract reasoning; rather, the generalized stimulus may simply be treated as similar to the conditioned cue. The terminology should therefore be reconsidered or softened.

(8) I also found the use of the term "valence" somewhat problematic. The manuscript appears to use valence to refer to graded responding across tones with different aversive significance, but valence typically refers more broadly to distinctions between appetitive and aversive value. Here, terms such as "threat value," "aversive value," may be more precise. The authors should consider revising this language throughout.

https://doi.org/10.7554/eLife.111177.1.sa3

Reviewer #2 (Public review):

Summary:

The following points are those that occurred to me across readings of the paper. They are listed in what I take to be the order of their significance. Many of the points relate to the loose use of language and invocation of concepts that are not warranted, given the study design and results obtained.

Major Comments:

(1) The concept of ensemble turnover is interesting - the way it is introduced and discussed implies some type of spontaneous change in the neural underpinnings of fear discrimination and generalization in the PL. But, of course, every trial involves an opportunity to learn about the threat CS or the generalization test stimuli, and I am troubled by the thought that stability in the neural underpinnings of fear discrimination and generalization will actually reflect the level of defensive behaviours evoked on different trial types and/or the discrepancy between those behaviours and the outcome of a given trial in the generalization test. That is, stability in the neural underpinnings may be related to an animal's certainty or uncertainty in the contingency between a stimulus and danger; or, put another way, an animal's confidence that danger will or won't occur given the presence of some stimulus. This is not uninteresting. It is, however, not considered anywhere in the paper, which is overloaded with references to

inferred threat values and integration of information across different types of stimuli. The protocol is not one that requires inference about anything or integration across anything.

(2) I appreciate the link to Gu and Johansen in paragraph 3 of the Introduction, but the type of generalization under investigation here is not the same as the type of 'generalization' studied by Gu and Johansen [who used a sensory preconditioning protocol]. Nonetheless, the authors have forced the language used by Gu and Johansen into their paper, and this has created tension [at least for this reader] as the concepts introduced by Gu and Johansen [inference, integration] are simply not relevant given the generalization protocol used here. Here are a few examples of points where the tension might interfere with a reader's understanding:

a. 'We hypothesized that generalization to novel stimuli depends on stable subnetwork organization that enables comparisons between learned and inferred valence, as well as population-level features that reduce variability across related representations.'

I understand the words in the hypothesis, but can't form a representation of what is being said because of the reference to terms that stand in need of clarification [inferred valence, variability across related representations], but, ultimately, won't be clarified. This needs to be re-expressed so that the reader can appreciate what is being said.

b. 'Our results show that stable cortical subnetworks integrate the emotional "gist" of memory and inferred valence for novel cues over time, despite ongoing ensemble reorganization, and that population-level firing rate similarity across stimulus presentations determines threat generalization.'

Again, what does this mean? How is the gist of a memory integrated with inferred valence for novel cues over time? The statement simply doesn't make sense. This needs to be rewritten for clarity.

c. 'In $CS^{+}15$ mice, positively modulated sound-responsive neurons exhibited graded tone activity reflecting the contingency learned valence as well as the inferred valence of novel tones across testing days...'.

Can this be rewritten as 'In $CS^{+}15$ mice, positively modulated sound-responsive neurons exhibited graded activity to the tone CS and its variants that were used to assess generalization.'? The overloading of the text with references to 'contingency learned valence' and 'inferred valence' is unnecessary and makes it much harder to understand what has been shown in the results.

(3) Re the same passage of text as in 2c:

Is it the case that these neurons are simply tracking the expression of freezing to the various tones? The same question applies to the results obtained for the CS+3 mice. If this is the case, then why should the results be taken to support the banner statement that 'Sound-modulated PL population responses encode learned and inferred valence' - these analyses do not support that statement. And, as indicated, I don't believe that the language of learned and inferred valence is appropriate to such statements, given the nature of the protocol used and results obtained. It is a study looking at how populations of neurons in the PL respond during presentations of auditory stimuli that were subject to discriminative conditioning, and during tests of generalized freezing to other [intermediate] auditory stimuli.

(4) It is stated that:

'In no-shock controls, although both positive and negative responses were present, population activity was not modulated by tone frequency or valence'.

What does this mean? I can understand that population activity was not modulated by tone frequency. But what does it mean to say that it was not modulated by valence? Why should it

have been when none of the tones were conditioned in this group and, hence, mice were responding to all the tones equally? And given that this is true, I don't understand the use of 'valence' here, or the subsequent statements in this paragraph that 'graded responses require associative learning' and that 'PL population responses encode graded sound-valence associations that reflect both learning and inference, closely matching behavioral generalization.' The latter statement is particularly unwarranted and, again, highlights a major issue with the paper. It could and should be rewritten as 'PL population responses reflect behavioral generalization.' There is nothing in the additional language that adds to the reader's understanding of what has been shown. The reference to 'graded sound-valence associations that reflect both learning and inference' is completely unwarranted, given the nature of this study. It is anathema to the vast literature on stimulus generalization. If the authors wished to make statements of this sort, they should have taken a different approach, perhaps using protocols like those featured in Gu and Johansen.

(5) The section titled, 'Consistently active neurons preserve valence representations as newly recruited neurons sharpen remote memory traces' ends with the following summary:

'Together, these results indicate that consistently active neurons maintain stable representations of learned and inferred sound associations across time, whereas neurons recruited after conditioning progressively acquire graded tuning at later retrieval stages. This dynamic refinement suggests that cortical memory representations become increasingly selective during systems consolidation, while a stable neuronal subpopulation preserves the core emotional content of the memory.'

Once again, the summary is not in keeping with the results obtained. The 'dynamic refinement' of representations is far more likely to reflect the repeated testing across days 1, 15, and 30 rather than anything to do with systems consolidation - at the very least, it is the simplest interpretation of the results. The impact of repeated testing is evident in the sharpening of generalization gradients over time, which is contrary to what is otherwise observed in the literature - the incredibly well -documented broadening of generalization gradients with time. Given this impact of repeated testing, surely the changes in the neuronal population that underlie performance are more likely to reflect the learning that occurs on days 1, 15, and 30, which is reflected in reduced freezing to the non-conditioned tones. If this is a reasonable take on the results, then I don't see the basis for invoking systems consolidation at all, and I don't see the basis for inferring a stable neuronal subpopulation that preserves the emotional content of the memory. Rather, non-reinforced presentations of 'never-reinforced' tones result in recruitment of additional neurons that result in suppression of freezing responses to those stimuli.

(6) In the section titled, 'Population vector similarity at stimulus onset determines degree of generalization', it is stated that:

'Because population similarity peaked shortly after stimulus onset, we quantified similarity during the first 5 s after tone onset relative to the CS^{+}. In $CS^{+}15$ mice, population similarity was highest for 15/15 and 15/11 tone pairs with no differences between them.'

Isn't this consistent with the view that the population response in the PL simply reflects the level of freezing? Freezing to the 15-15 and 15-11 tones is most likely to be similar on their first presentation prior to the effects of extinction on the 11 Hz tone; hence the results obtained. That is, these results appear to clearly indicate that neuronal responses in the PL reflect the degree of stimulus generalization, as evidenced in freezing behavior. Given all that we know about the involvement of the PL in expressing fear responses, it is not appropriate to claim that 'population vector similarity at stimulus onset *determines* the degree of generalization. The PL responses simply reflect the varying levels of performance displayed to the different types of tones. What have I missed that could be taken to support additional statements?

Later in the same section, it is stated that 'population-level similarity at stimulus onset scales with behavioral threat generalization and is maximal for tones associated with robust threat responses.' For simplicity and, therefore, clarity, this should be rewritten as 'population-level similarity at stimulus onset reflects behavioral threat generalization.'

(7) In the section titled, 'Different subnetworks encode acoustic versus learned properties of sound association', it is stated that:

'Our previous analyses show that learned and inferred associations are represented at the population level. However, these results do not resolve whether graded responses arise from pooled activity of frequency-selective neurons or from subnetworks encoding integrated learned valence across tones.'

What does it mean to say 'integrated learned valence across tones'? As it presently stands, the meaning of the phrase is unclear. It only makes sense if one supposes that generalized freezing responses to the 11 and 7 kHZ tones reflect separate associations between those tones and the aversive foot shock US. This supposition is inconsistent with the rich literature on generalization of Pavlovian conditioned fear responses. Specifically, it is inconsistent with the many theories of fear generalization, which attribute the reduction in fear as one moves away from the specific conditioned stimulus to a decrement in the ability of the test stimulus to activate the trained CS-US association. My strong impression is that the authors would do well to ground their findings in theories of stimulus/fear generalization, of which there are many. This would better serve the results obtained [and the reader's appreciation of them] - at present, the unnecessary invocation of concepts does very little to enhance the reader's appreciation or understanding of what has been found in the study.

(8) Another example of what has been a common theme in this review :

'...we hypothesized that the PL active ensemble segregates into functionally distinct subnetworks: one encoding tone-specific sensory features with dynamic characteristics, and another responding to all frequencies encoding stable core memory content and inferred emotional valence.'

What does it mean to say 'all frequencies encoding stable core memory content and inferred emotional valence'? Do the authors mean to say '...and another that tracks freezing/defensive responses regardless of whether they were elicited by the trained CS or one of the generalization test stimuli'?

(9) It is stated that - 'Graded clusters encode emotional valence but constitute only a fraction of the active population; yet valence coding at the population level remains accurate and precise. This indicates that neurons newly recruited into the population-likely frequency-selective and organized within learning-independent clusters-can be shaped by associative processes through modulation of firing activity.'

What does this mean? Are the authors trying to say that - 'Some clusters of PL neurons track freezing responses. In spite of the fact that these are only a fraction of the total active neuronal population, the population-level response of PL neurons also tracks the levels of fear to the trained tone and its variants used in the test for generalization.' If this is what one wants to say, then the final statement in the reproduced section does not follow. That is, there is no indication that 'neurons newly recruited into the population-likely frequency-selective and organized within learning-independent clusters-can be shaped by associative processes through modulation of firing activity.' As noted, the characteristics of other ensembles that become active across the repeated tests on days 1, 15, and 30 are more likely to reflect learning from non-reinforcement that occurs within and across those sessions. Perhaps this is what is meant by the phrase, 'shaped by associative processes'? If so, it should be stated explicitly instead of left to the reader to work out.

(10) The following points all relate to the Discussion and reiterate many of the points above.

a. 'A subset of neurons remains consistently active across sessions, preserving core components of the memory trace and supporting inference of emotional valence for novel sounds, while neurons recruited after conditioning progressively acquire valence selectivity at remote time points.'

'Inference of emotional valence' is unclear and unwarranted for all of the reasons provided above regarding the use of language.

b. '...Our data reconcile these views by demonstrating that cortical representations of emotional valence emerge rapidly after learning and persist within stable subnetworks, even as the broader population undergoes substantial turnover. This architecture preserves core mnemonic content while allowing flexibility in the surrounding ensemble.'

These statements assume that the PL neuronal responses reflect something more than the levels of freezing behavior to the different stimuli; what are the grounds for this assumption?

c. 'Importantly, these subnetworks encode both learned contingencies and the inferred valence of novel stimuli along a graded representational axis, suggesting that strong recurrent connectivity provides a stable scaffold for emotional memory representations.'

What is a graded representational axis, and what part of the first statement suggests that 'strong recurrent connectivity provides a stable scaffold for emotional memory representations'? If the authors' goal was to make statements about emotional memory representations vis-à-vis emotional memory content, they should have used protocols that allowed them to probe such content. The auditory fear conditioning protocol used here [followed by tests for generalization to other auditory stimuli that differ in frequency from the conditioned tone] is not one that lends itself to analysis of emotional memory representations or content.

d. 'Dynamic tone-selective responsive neurons emerge independently of learning, as they are present in both control and experimental mice, reflecting pre-existing PL sensory-driven properties (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006).'

Maybe. They are also likely to have developed as a consequence of the repeated testing on days 1, 15, and 30, which involved intermixed exposures to the tones of different frequencies. That is, rather than 'pre-existing PL sensory-driven properties', the responses of these neurons might reflect the emergence of discrimination between the various tones across testing, and greater suppression of freezing to the non-trained tones compared to the trained tone across the various test intervals.

https://doi.org/10.7554/eLife.111177.1.sa2

Reviewer #3 (Public review):

Summary:

Normandin et al. explore the coding of stimuli predicting an aversive event in the prelimbic cortex. Stimuli could either be explicitly paired, explicitly unpaired, or novel but with an inferred association with the aversive event (generalization). Long-term tracking of GCaMP-positive neurons allowed them to examine how coding evolves out to a month following training. In general, they found two types of ensemble codes. One was ensembles coding for each stimulus independently, but with enhanced responding to the one eliciting a freezing response. The other was ensembles that responded to all stimuli in proportion to their similarity to the stimulus paired with the aversive event, either increasing or decreasing their

activation with the degree of freezing elicited by a stimulus. Importantly, this second set of ensembles was more stable across days, potentially providing a memory trace.

Strengths:

(1) The authors track ensembles in prelimbic cortex over long time scales, providing valuable information on the consolidation of neural codes.

(2) Neural coding of generalization is examined, which is under-examined in the field.

Weaknesses:

(1) Difficult to determine if responses treated as encoding stimulus valence are driven instead by the behavior that the stimulus elicits, freezing.

(2) The study implies that the identified ensembles are causally related to valence memory, but no experimental interventions are performed to justify this.

https://doi.org/10.7554/eLife.111177.1.sa1

Author response:

Public Reviews:

Reviewer #1 (Public review):

Summary:

The authors combine discriminative auditory fear conditioning with longitudinal in vivo calcium imaging to ask how prelimbic (PL) representations of learned and generalized threat evolve across recent and remote memory time points. Using two different CS+ frequencies and a no-shock control group, they report that PL population activity tracks graded behavioral generalization, that population similarity is highest for tones eliciting strong threat responding, and that distinct subnetworks can be identified that appear to encode tone-specific sensory features versus learned threat-related response structure.

To my knowledge, this may be the first study to comprehensively examine neural encoding of fear generalization in prelimbic cortex (PL). The manuscript is ambitious and technically interesting, and several aspects are potentially important. In particular, the suggestion that neurons showing graded, learning-related response patterns become selectively stabilized over time is intriguing. The inclusion of two CS+ training conditions and a no-shock control also strengthens the case that at least some of the reported effects are related to associative learning rather than simple sensory differences. However, in its current form, the manuscript does not yet fully support the strength of the conceptual claims. Several issues limit confidence in the interpretation, including the possibility that repeated testing itself contributes to changes across days, uncertainty about the relationship between neural activity and freezing behavior, limited quantitative documentation of longitudinal cell registration, and a number of problems in figure clarity and statistical framing. Overall, the study contains promising observations, but the claims should be narrowed, and several analyses or controls would be needed to fully support the proposed framework.

Detailed Comments

(1) A general concern is that the repeated test procedure itself may contribute to extinction. Because the animals are exposed to multiple CS frequencies across multiple test days, and each tone is presented three times per session, some of the reported changes in behavior and neural activity across days could reflect extinction or repeated

> *nonreinforced retrieval rather than the passage of time per se. This is especially relevant given that the manuscript makes claims about recent versus remote representations and representational drift over 30 days. At a minimum, the authors should discuss this limitation explicitly and temper claims about time-dependent changes. Ideally, they would include a control group in which animals are tested only once or twice (e.g., at an early and later time point with fewer CS frequencies), or a reduced-frequency testing design that minimizes extinction while still allowing evaluation of recent versus remote memory.*

We agree with the reviewer that repeated testing is an inherent limitation of longitudinal memory studies and may itself contribute to some neural changes across sessions. However, several aspects of our behavioral design and results argue against extinction or repeated nonreinforced retrieval as the primary drivers of the observed effects. Importantly, discrimination ratios remained stable or increased across time rather than progressively diminishing as would be expected under extinction (this new analysis will be added to the resubmission). Nevertheless, we will address this important point in the Discussion and explicitly acknowledge that repeated retrieval may contribute to some component of the observed representational changes.

> *(2) More generally, some of the reported learning-related neural differences may be driven by behavioral differences, particularly freezing, rather than by learning or generalization per se. For example, animals that freeze more to certain frequencies may show corresponding neural response differences simply because freezing alters PL activity. The authors should examine this possibility more directly. Analyses testing whether recorded cells encode freezing behavior, or whether tone frequency-related neural differences remain robust when comparing high- and low-freezing epochs, would help determine whether the reported effects reflect learned stimulus value rather than behavioral state differences.*

We thank the reviewer for raising this important point, which was also noted by the other reviewers. To address this issue, we will implement Reviewer 3's suggested Generalized Linear Model (GLM) analysis using inferred spiking activity derived from the Ca2+ signals, with both tone identity and freezing behavior included as predictors. Because freezing behavior varies across trials whereas stimulus identity is fixed, this approach will allow us to dissociate their respective contributions to neuronal activity. If, after accounting for freezing behavior, responsive neurons continue to exhibit graded coding consistent with inferred threat value, this would strengthen the interpretation that the identified ensembles reflect generalization gradients related to aversive value rather than freezing behavior alone. Otherwise, we will adjust the conclusions according to the interpretation that freezing itself drives the generalization gradients.

> *(3) A central feature of the manuscript is the analysis of neural response properties over an extended period of time, up to 30 days after learning. However, aside from a brief mention in the Methods that spatial registration was used, the manuscript provides very little quantitative information about this critical aspect of the study. The paper would be strengthened by including explicit metrics describing longitudinal cell tracking, such as the number and proportion of ROIs retained across all sessions, distributions of spatial-footprint correlations or centroid distances across days, and representative examples of matched imaging fields over time. Without this information, it is difficult to assess how strongly the longitudinal claims are supported.*

We thank the reviewer for this suggestion. We will include measures of registration quality in the resubmission.

> *(4) The text states that "Figs. 1c and 1d show GCaMP6f expression in PL, representative calcium footprints, and activity traces". However, the figure as presented does not clearly*

> *show all of these elements, at least not in a way that matches the description in the Results. The correspondence between text and figure should be corrected.*

We will correct correspondence between text and Figure.

> *(5) The labeling of Figure 2a is insufficient for interpretation. The legend states that the panel shows raster plots of sound responsiveness, but the axes and scaling are not clearly defined. It is not clear from the figure what the x-axis represents, whether the y-axis corresponds to individual neurons, where the CS period occurs, or what the activity scale at the right denotes. Also, the term 'rasters' implies that spikes were analyzed. It seems that the spike inference approach (CASCADE) was only used for later analyses. Perhaps 'heat-plot' would be more accurate here? Generally, this figure should be annotated more clearly so that the reader can understand it without referring back to the Methods.*

Thank you for this suggestion. We will clarify the labelling of the Figure 2a and call the graphs “activity-plots”.

> *(6) In relation to Figure 3, the analysis of population-averaged responses across tone frequencies is useful, but the manuscript would be stronger with additional statistical analyses across time and across groups. For example, if the authors want to argue that learning induces graded changes in neural responses and that these evolve across time, they should directly compare within-group responses across days and also compare matched frequencies between the conditioned groups and the no-shock controls. These analyses would help establish whether the observed differences are genuinely learning dependent and whether they change significantly over time.*

We will redo the Statistics of Figure 3 to take into account the following variables: group (CS15, CS3, no shocks), frequency (3, 7, 11, 15), and day of testing (2, 15, 30).

> *(7) The inclusion of two different CS+ frequencies and a no-shock control is a strength of the study and substantially improves the interpretation that graded neural responses are related to learning and generalization rather than to simple sensory processing or passage of time. That said, I am not entirely comfortable with the use of the term "inference" throughout the manuscript. What is being measured here appears closer to sensory generalization than inference in a stronger cognitive sense. The current task does not clearly require that animals infer hidden structure or stimulus value through abstract reasoning; rather, the generalized stimulus may simply be treated as similar to the conditioned cue. The terminology should therefore be reconsidered or softened.*

We thank the reviewer for appreciating the strengths of the experimental design and for this thoughtful suggestion regarding terminology. We agree that the term “inference” may overstate the cognitive processes engaged by the current task. Accordingly, we will revise the terminology throughout the manuscript to describe these effects as graded generalization of threat value across stimuli.

> *(8) I also found the use of the term "valence" somewhat problematic. The manuscript appears to use valence to refer to graded responding across tones with different aversive significance, but valence typically refers more broadly to distinctions between appetitive and aversive value. Here, terms such as "threat value," "aversive value," may be more precise. The authors should consider revising this language throughout.*

We will correct the language and use “threat value”.

> ***Reviewer #2 (Public review):***
>
> *Summary:*

> *The following points are those that occurred to me across readings of the paper. They are listed in what I take to be the order of their significance. Many of the points relate to the loose use of language and invocation of concepts that are not warranted, given the study design and results obtained.*
>
> *Major Comments:*
>
> *(1) The concept of ensemble turnover is interesting - the way it is introduced and discussed implies some type of spontaneous change in the neural underpinnings of fear discrimination and generalization in the PL. But, of course, every trial involves an opportunity to learn about the threat CS or the generalization test stimuli, and I am troubled by the thought that stability in the neural underpinnings of fear discrimination and generalization will actually reflect the level of defensive behaviours evoked on different trial types and/or the discrepancy between those behaviours and the outcome of a given trial in the generalization test. That is, stability in the neural underpinnings may be related to an animal's certainty or uncertainty in the contingency between a stimulus and danger; or, put another way, an animal's confidence that danger will or won't occur given the presence of some stimulus. This is not uninteresting. It is, however, not considered anywhere in the paper, which is overloaded with references to inferred threat values and integration of information across different types of stimuli. The protocol is not one that requires inference about anything or integration across anything.*

We thank the reviewer for these important points, which we address in further detail below.

Ongoing learning during test sessions: The reviewer correctly notes that unreinforced test presentations may constitute extinction-learning trials and that some neural changes across days could therefore reflect ongoing learning rather than spontaneous ensemble reorganization. However, new analyses indicate that extinction is unlikely to be the primary driver of our findings. Discrimination ratios do not decay over time; instead, they either sharpen or remain stable across sessions (new analyses to be included in the resubmission). These results argue against robust extinction as the primary source of the neural changes observed across sessions. This interpretation is also consistent with the strength of our conditioning protocol, which used 10 CS+ shock pairings and 10 CS− no-shock pairings specifically to minimize extinction across repeated testing sessions. Nevertheless, we acknowledge that the current design cannot fully dissociate time-dependent consolidation from retrieval-induced plasticity, and we will explicitly discuss this limitation in the revised Discussion.

Stability reflecting behavioral consistency: We agree this alternative cannot be fully excluded. However, the cluster stability analyses assess identity at the level of response profile across all four frequencies, not response magnitude alone. Tone-selective clusters, which also show consistent behavioral correlates (firing rate correlates with threat-value, Fig. S8), do not show equivalent profile stability, suggesting that the stability of graded clusters is not simply a consequence of behavioral consistency. This point will be added to the Discussion in the resubmission.

Language of "inference" and "integration": The reviewer is correct that responses to novel tones are consistent with graded stimulus generalization. We will substantially revise the manuscript to replace "inference" and "integration" with more precise language describing graded frequency generalization gradients.

> *(2) I appreciate the link to Gu and Johansen in paragraph 3 of the Introduction, but the type of generalization under investigation here is not the same as the type of 'generalization' studied by Gu and Johansen [who used a sensory preconditioning*

> *protocol]. Nonetheless, the authors have forced the language used by Gu and Johansen into their paper, and this has created tension [at least for this reader] as the concepts introduced by Gu and Johansen [inference, integration] are simply not relevant given the generalization protocol used here. Here are a few examples of points where the tension might interfere with a reader's understanding:*

We thank the reviewer for these specific and constructive criticisms. We will revise the manuscript throughout to remove or redefine terms like "inferred valence" and "integration," replacing them with clearer, more accurate descriptions of gradient generalization of threat value. Below we address each point raised by the reviewer regarding terminology clarifications.

> *(a) 'We hypothesized that generalization to novel stimuli depends on stable subnetwork organization that enables comparisons between learned and inferred valence, as well as population-level features that reduce variability across related representations.'*
>
> *I understand the words in the hypothesis, but can't form a representation of what is being said because of the reference to terms that stand in need of clarification [inferred valence, variability across related representations], but, ultimately, won't be clarified. This needs to be re-expressed so that the reader can appreciate what is being said.*

The hypothesis will be rewritten as: "We hypothesized that generalization to tones acoustically similar to the CS+ and CS− depends on the emergence of stable ensembles encoding threat value, and that population-level response similarity across stimuli would correlate with the degree of behavioral fear generalization, consistent with prior work in auditory cortex [1]."

> *(b) 'Our results show that stable cortical subnetworks integrate the emotional "gist" of memory and inferred valence for novel cues over time, despite ongoing ensemble reorganization, and that population-level firing rate similarity across stimulus presentations determines threat generalization.'*
>
> *Again, what does this mean? How is the gist of a memory integrated with inferred valence for novel cues over time? The statement simply doesn't make sense. This needs to be rewritten for clarity.*

The summary statement will be rewritten: "Our results show that stable cortical sub-ensembles preserve the emotional content of the fear memory over time, despite ongoing ensemble reorganization, and that population-level firing rate similarity in response to tones associated with threat correlates with the degree of behavioral threat generalization."

> *(c) 'In CS+15 mice, positively modulated sound-responsive neurons exhibited graded tone activity reflecting the contingency learned valence as well as the inferred valence of novel tones across testing days...'.*
>
> *Can this be rewritten as 'In CS+15 mice, positively modulated sound-responsive neurons exhibited graded activity to the tone CS and its variants that were used to assess generalization.'? The overloading of the text with references to 'contingency learned valence' and 'inferred valence' is unnecessary and makes it much harder to understand what has been shown in the results.*

We will adopt the reviewer's suggested rewording: "In CS+15 mice, positively modulated sound-responsive neurons exhibited graded activity to the tone CS and its variants that were used to assess generalization."

We will systematically review the entire manuscript to ensure consistency with this revised framing.

> *(3) Re the same passage of text as in 2c:*
>
> *Is it the case that these neurons are simply tracking the expression of freezing to the various tones? The same question applies to the results obtained for the CS+3 mice. If this is the case, then why should the results be taken to support the banner statement that 'Sound-modulated PL population responses encode learned and inferred valence' - these analyses do not support that statement. And, as indicated, I don't believe that the language of learned and inferred valence is appropriate to such statements, given the nature of the protocol used and results obtained. It is a study looking at how populations of neurons in the PL respond during presentations of auditory stimuli that were subject to discriminative conditioning, and during tests of generalized freezing to other [intermediate] auditory stimuli.*

The reviewer is correct that the graded population responses observed in PL could reflect freezing behavior across tone frequencies rather than encoding an abstract threat-value representation. This important concern was also raised by other reviewers. To address it directly, we will follow Reviewer 3's suggestion and implement a Generalized Linear Model (GLM) using inferred spiking activity derived from the Ca2+ signals, with both tone identity and freezing behavior included as predictors. This analysis will allow us to dissociate the respective contributions of tone frequency and freezing to the graded neural responses. Based on the outcome of this analysis, we will revise and appropriately adjust our conclusions.

In addition, we will revise the section heading and surrounding text to remove the terminology of "learned and inferred valence." Instead, the findings will be described more conservatively as: "PL population responses reflect behavioral generalization to auditory stimuli following discriminative fear conditioning."

> *(4) It is stated that:*
>
> *'In no-shock controls, although both positive and negative responses were present, population activity was not modulated by tone frequency or valence'.*
>
> *What does this mean? I can understand that population activity was not modulated by tone frequency. But what does it mean to say that it was not modulated by valence? Why should it have been when none of the tones were conditioned in this group and, hence, mice were responding to all the tones equally? And given that this is true, I don't understand the use of 'valence' here, or the subsequent statements in this paragraph that 'graded responses require associative learning' and that 'PL population responses encode graded sound-valence associations that reflect both learning and inference, closely matching behavioral generalization.' The latter statement is particularly unwarranted and, again, highlights a major issue with the paper. It could and should be rewritten as 'PL population responses reflect behavioral generalization.' There is nothing in the additional language that adds to the reader's understanding of what has been shown. The reference to 'graded sound-valence associations that reflect both learning and inference' is completely unwarranted, given the nature of this study. It is anathema to the vast literature on stimulus generalization. If the authors wished to make statements of this sort, they should have taken a different approach, perhaps using protocols like those featured in Gu and Johansen.*

The reviewer is correct that controls do not form threat associations; however, these animals still could respond differentially to distinct frequencies, something that is not reflected in the data. We will correct the section indicating that distinct neutral frequencies do not produce graded responses: "graded responses require associative learning" will be retained but reframed simply as: "graded frequency-dependent population responses were absent in

animals that did not receive fear conditioning." The concluding statement of the paragraph will be rewritten as: "PL population responses reflect behavioral generalization to acoustically similar stimuli following discriminative conditioning," in line with the reviewer's suggestion.

> *(5) The section titled, 'Consistently active neurons preserve valence representations as newly recruited neurons sharpen remote memory traces' ends with the following summary:*
>
> *'Together, these results indicate that consistently active neurons maintain stable representations of learned and inferred sound associations across time, whereas neurons recruited after conditioning progressively acquire graded tuning at later retrieval stages. This dynamic refinement suggests that cortical memory representations become increasingly selective during systems consolidation, while a stable neuronal subpopulation preserves the core emotional content of the memory.'*
>
> *Once again, the summary is not in keeping with the results obtained. The 'dynamic refinement' of representations is far more likely to reflect the repeated testing across days 1, 15, and 30 rather than anything to do with systems consolidation - at the very least, it is the simplest interpretation of the results. The impact of repeated testing is evident in the sharpening of generalization gradients over time, which is contrary to what is otherwise observed in the literature - the incredibly well -documented broadening of generalization gradients with time. Given this impact of repeated testing, surely the changes in the neuronal population that underlie performance are more likely to reflect the learning that occurs on days 1, 15, and 30, which is reflected in reduced freezing to the non-conditioned tones. If this is a reasonable take on the results, then I don't see the basis for invoking systems consolidation at all, and I don't see the basis for inferring a stable neuronal subpopulation that preserves the emotional content of the memory. Rather, non-reinforced presentations of 'never-reinforced' tones result in recruitment of additional neurons that result in suppression of freezing responses to those stimuli.*

We respectfully disagree with the reviewer's interpretation. While repeated testing cannot be entirely excluded as a contributing factor, several lines of evidence suggest that it cannot fully account for our observations.

Regarding extinction: discrimination ratios between CS+ and all other frequencies either remained stable or increased over time (new analysis included in resubmission), indicating that animals continued to discriminate threat value across the testing period rather than showing the progressive suppression expected under extinction — the opposite of what we observe.

Regarding the recruitment of new neurons: repeated non-reinforced tone exposure would be expected to produce stimulus-specific adaptation — characterized by reduced, less discriminative neural responsiveness and flatter tuning profiles [2]— not the progressive sharpening we observe. The same would be expected if these neurons represent or are associated with new extinction learning.

Finally, sharpening of generalization gradients during repeated within-subjects testing has been reported previously [3], suggesting that successive exposures may promote more precise discrimination in some cases. Consistent with this, discrimination learning has also been shown to narrow or sharpen fear generalization gradients rather than broaden them [4], supporting the idea that discriminative conditioning enhances stimulus specificity during testing. Although we cannot exclude the possibility that more extended training could eventually broaden the generalization gradient, under the training parameters and temporal window used in our study, the data support a progressive sharpening of the gradient over time. In the revised Discussion, we will present systems consolidation as the primary

interpretive framework and further elaborate on why repeated testing is unlikely to account for the full pattern of behavioral and neural findings reported here.

> *(6) In the section titled, 'Population vector similarity at stimulus onset determines degree of generalization', it is stated that:*
>
> *'Because population similarity peaked shortly after stimulus onset, we quantified similarity during the first 5 s after tone onset relative to the CS+. In CS+15 mice, population similarity was highest for 15/15 and 15/11 tone pairs with no differences between them.'*
>
> *Isn't this consistent with the view that the population response in the PL simply reflects the level of freezing? Freezing to the 15-15 and 15-11 tones is most likely to be similar on their first presentation prior to the effects of extinction on the 11 Hz tone; hence the results obtained. That is, these results appear to clearly indicate that neuronal responses in the PL reflect the degree of stimulus generalization, as evidenced in freezing behavior. Given all that we know about the involvement of the PL in expressing fear responses, it is not appropriate to claim that 'population vector similarity at stimulus onset *determines* the degree of generalization. The PL responses simply reflect the varying levels of performance displayed to the different types of tones. What have I missed that could be taken to support additional statements?*

The GLM analysis described in our response to reviewers 1 and 3 will directly address the contribution of freezing. We will report these results in the resubmission and revise the interpretive language in the manuscript accordingly.

However, regarding the analysis of population vector similarity, we need to clarify a point of confusion. The reviewer states “Freezing to the 15-15 and 15-11 tones is most likely to be similar on their first presentation prior to the effects of extinction on the 11 Hz tone; hence the results obtained”. The similarity vectors were calculated by correlating activity across all tone presentations within each testing day, not only the first two presentations. In Fig. 4, “Early” and “Late” refer to the order of a tone within a trial, which we will clarify more explicitly in the resubmission. Notably, repeated-measures analyses did not reveal any effect of the time variable (Fig. 4e,f), indicating that similarity across tone presentations remained high for tones associated with high threat value. Importantly, our data showed no evidence that responses to 11 kHz or 15 kHz in the CS15 group, or to 3 kHz in the CS3 group, exhibited extinction-like patterns at either the behavioral or neural level. Therefore, the persistence of high population similarity across time provides additional evidence against extinction as the primary explanation for our findings.

We will remove the word "determines" from the manuscript, as our data cannot conclusively establish a causal relationship.

> *Later in the same section, it is stated that 'population-level similarity at stimulus onset scales with behavioral threat generalization and is maximal for tones associated with robust threat responses.' For simplicity and, therefore, clarity, this should be rewritten as 'population-level similarity at stimulus onset reflects behavioral threat generalization.'*

We will make this correction.

> *(7) In the section titled, 'Different subnetworks encode acoustic versus learned properties of sound association', it is stated that:*
>
> *'Our previous analyses show that learned and inferred associations are represented at the population level. However, these results do not resolve whether graded responses arise from pooled activity of frequency-selective neurons or from subnetworks encoding integrated learned valence across tones.'*

> *What does it mean to say 'integrated learned valence across tones'? As it presently stands, the meaning of the phrase is unclear. It only makes sense if one supposes that generalized freezing responses to the 11 and 7 kHZ tones reflect separate associations between those tones and the aversive foot shock US. This supposition is inconsistent with the rich literature on generalization of Pavlovian conditioned fear responses. Specifically, it is inconsistent with the many theories of fear generalization, which attribute the reduction in fear as one moves away from the specific conditioned stimulus to a decrement in the ability of the test stimulus to activate the trained CS-US association. My strong impression is that the authors would do well to ground their findings in theories of stimulus/fear generalization, of which there are many. This would better serve the results obtained [and the reader's appreciation of them] - at present, the unnecessary invocation of concepts does very little to enhance the reader's appreciation or understanding of what has been found in the study.*

We thank the reviewer for raising this point. The phrase "integrated learned valence across tones" refers specifically to a subpopulation of neurons that respond to all four frequencies in a graded manner, with response magnitude scaling according to threat value. This is distinct from tone-selective neurons, which respond preferentially to a single frequency. The neurons responding to all tones in a graded manner are present only in conditioned animals and not in no-shock controls, demonstrating that their graded response profile is shaped by associative learning.

We agree, however, that the phrase "integrated learned valence" is unnecessarily opaque and we will replace it with more precise language: these neurons will be described as showing graded frequency-dependent responses whose magnitude scales with threat value. We believe this subpopulation represents a genuinely novel finding that complements the behavioral generalization literature by identifying a specific neural substrate for the generalization gradient within PL.

> *(8) Another example of what has been a common theme in this review:*
>
> *'...we hypothesized that the PL active ensemble segregates into functionally distinct subnetworks: one encoding tone-specific sensory features with dynamic characteristics, and another responding to all frequencies encoding stable core memory content and inferred emotional valence.'*
>
> *What does it mean to say 'all frequencies encoding stable core memory content and inferred emotional valence'? Do the authors mean to say '...and another that tracks freezing/defensive responses regardless of whether they were elicited by the trained CS or one of the generalization test stimuli'?*

As stated in our previous responses, in the resubmission we will determine the contribution of freezing. If we find that freezing predicts graded neural responses, we will adjust the language of the manuscript.

> *(9) It is stated that - 'Graded clusters encode emotional valence but constitute only a fraction of the active population; yet valence coding at the population level remains accurate and precise. This indicates that neurons newly recruited into the population-likely frequency-selective and organized within learning-independent clusters-can be shaped by associative processes through modulation of firing activity.'*
>
> *What does this mean? Are the authors trying to say that - 'Some clusters of PL neurons track freezing responses. In spite of the fact that these are only a fraction of the total active neuronal population, the population-level response of PL neurons also tracks the levels of fear to the trained tone and its variants used in the test for generalization.' If this is what one wants to say, then the final statement in the reproduced section does not*

> *follow. That is, there is no indication that 'neurons newly recruited into the population-likely frequency-selective and organized within learning-independent clusters-can be shaped by associative processes through modulation of firing activity.' As noted, the characteristics of other ensembles that become active across the repeated tests on days 1, 15, and 30 are more likely to reflect learning from non-reinforcement that occurs within and across those sessions. Perhaps this is what is meant by the phrase, 'shaped by associative processes'? If so, it should be stated explicitly instead of left to the reader to work out.*

We thank the reviewer for highlighting the lack of clarity in this passage and agree that the original phrasing was insufficiently precise. What we intended to convey is that only a subset of PL neurons displays graded tuning that tracks behavioral generalization across tones. Nevertheless, despite constituting only a fraction of the total active population, this graded coding is also reflected at the population level. Therefore, we suggest that neurons recruited into the active population after conditioning — likely frequency-selective neurons — contribute to the graded population responses through changes in their firing-rate activity, which is modulated by threat value (Fig. S8). We will rewrite this passage in the resubmission to make this interpretation explicit rather than leaving it to the reader to infer.

Regarding the reviewer's suggestion that the characteristics of newly recruited neurons more likely reflect learning from non-reinforced exposures during repeated test sessions, we respectfully maintain that this interpretation is difficult to reconcile with two aspects of our data. First, graded-response neurons are absent in no-shock controls that are exposed to nonreinforced repeated testing. Second, as detailed in our responses to previous points, the progressive sharpening of population responses over time is inconsistent with what would be expected from repeated non-reinforced exposure, which would more plausibly produce broader or flatter tuning profiles.

We agree that the phrase "shaped by associative processes" was ambiguous and will replace it with explicit language clarifying that we refer to fear conditioning as the associative process driving the emergence of graded responses, rather than any learning occurring during the test sessions themselves.

> *(10) The following points all relate to the Discussion and reiterate many of the points above.*
>
> *(a) 'A subset of neurons remains consistently active across sessions, preserving core components of the memory trace and supporting inference of emotional valence for novel sounds, while neurons recruited after conditioning progressively acquire valence selectivity at remote time points.'*
>
> *'Inference of emotional valence' is unclear and unwarranted for all of the reasons provided above regarding the use of language.*

We will modify the language as stated in the prior points.

> *(b) '...Our data reconcile these views by demonstrating that cortical representations of emotional valence emerge rapidly after learning and persist within stable subnetworks, even as the broader population undergoes substantial turnover. This architecture preserves core mnemonic content while allowing flexibility in the surrounding ensemble.'*
>
> *These statements assume that the PL neuronal responses reflect something more than the levels of freezing behavior to the different stimuli; what are the grounds for this assumption?*

We will incorporate new analysis (GLM) to better address this point and conclusions.

> *(c) 'Importantly, these subnetworks encode both learned contingencies and the inferred valence of novel stimuli along a graded representational axis, suggesting that strong recurrent connectivity provides a stable scaffold for emotional memory representations.'*
>
> *What is a graded representational axis, and what part of the first statement suggests that 'strong recurrent connectivity provides a stable scaffold for emotional memory representations'? If the authors' goal was to make statements about emotional memory representations vis-à-vis emotional memory content, they should have used protocols that allowed them to probe such content. The auditory fear conditioning protocol used here [followed by tests for generalization to other auditory stimuli that differ in frequency from the conditioned tone] is not one that lends itself to analysis of emotional memory representations or content.*

We thank the reviewer for this comment and agree that both phrases require clarification or revision.

By "graded representational axis" we intended to convey that PL population activity varies systematically as a function of stimulus similarity to the conditioned tone — that is, population responses are not categorical but scale continuously with spectral proximity to the CS+. We agree this was not clearly stated and will revise the manuscript accordingly.

Regarding recurrent connectivity, we agree with the reviewer that nothing in our data directly measures or manipulates connectivity between neurons. This statement was intended as a speculative interpretive hypothesis in the Discussion, motivated by the established literature linking strong recurrent connectivity in prefrontal circuits to stable population-level representations [5]. However, we acknowledge that invoking it in this context, without direct evidence, risks overstating our conclusions. We will revise this sentence to make its speculative nature explicit and ground it more carefully in the cited literature rather than presenting it as an inference from our own data.

In summary, we will ensure our conclusions will be restricted to population-level coding of learned threat value and its generalization across auditory frequencies. We will revise the relevant passages in the Discussion to ensure that speculative interpretations regarding emotional memory content are either removed or clearly flagged as speculative hypotheses.

> *(d) 'Dynamic tone-selective responsive neurons emerge independently of learning, as they are present in both control and experimental mice, reflecting pre-existing PL sensory-driven properties (Hockley & Malmierca, 2024; Zikopoulos & Barbas, 2006).'*
>
> *Maybe. They are also likely to have developed as a consequence of the repeated testing on days 1, 15, and 30, which involved intermixed exposures to the tones of different frequencies. That is, rather than 'pre-existing PL sensory-driven properties', the responses of these neurons might reflect the emergence of discrimination between the various tones across testing, and greater suppression of freezing to the non-trained tones compared to the trained tone across the various test intervals.*

We thank the reviewer for this point. Our interpretation that these neurons reflect pre-existing PL sensory-driven properties was based on the observation that tone-selective responses were present in control animals that never received conditioning, consistent with prior reports of sensory responsiveness in PL cortex ([6, 7]. Because these responses emerge from the first time we expose mice to the intermediate frequencies, they cannot be explained by repeated exposure. Moreover, we did not observe progressive refinement, emergence of discrimination-like changes, or suppression of responding to non-reinforced tones in control mice. This difference between conditioned and control animals indicates that repeated tone exposure alone is not sufficient to produce the observed dynamics — associative learning is necessary. We therefore maintain that the tone-selective responses of these neurons reflect

pre-existing sensory-driven properties of PL cortex that are present independently of conditioning history.

In summary, we thank the reviewer for suggesting clarifications to our interpretation, for raising the possibility that freezing behavior may contribute to graded neural responses, and for raising the question of whether repeated tone exposure may contribute to the properties of neurons recruited after conditioning. In the revised manuscript, we will include additional analyses to better dissociate the contributions of freezing behavior and tone identity, clarify passages that were insufficiently precise, and include a paragraph in the Discussion addressing potential alternative explanations alongside our own interpretation of the data.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *Normandin et al. explore the coding of stimuli predicting an aversive event in the prelimbic cortex. Stimuli could either be explicitly paired, explicitly unpaired, or novel but with an inferred association with the aversive event (generalization). Long-term tracking of GCaMP-positive neurons allowed them to examine how coding evolves out to a month following training. In general, they found two types of ensemble codes. One was ensembles coding for each stimulus independently, but with enhanced responding to the one eliciting a freezing response. The other was ensembles that responded to all stimuli in proportion to their similarity to the stimulus paired with the aversive event, either increasing or decreasing their activation with the degree of freezing elicited by a stimulus. Importantly, this second set of ensembles was more stable across days, potentially providing a memory trace.*
>
> *Strengths:*
>
> *(1) The authors track ensembles in prelimbic cortex over long time scales, providing valuable information on the consolidation of neural codes.*
>
> *(2) Neural coding of generalization is examined, which is under-examined in the field.*

We thank the reviewer for appreciating our design to track ensembles over time and the relevance of studying the neural substrates of generalization.

> *Weaknesses:*
>
> *(1) Difficult to determine if responses treated as encoding stimulus valence are driven instead by the behavior that the stimulus elicits, freezing.*

We thank the reviewer for this thoughtful and constructive comment. We agree that an alternative interpretation is that the graded-response ensembles may partially reflect freezing-related activity rather than mnemonic or salience-related representations of the conditioned stimuli themselves. In the revision, we will acknowledge that prior work has identified PL neurons that encode freezing independently of stimulus identity or associative content. Furthermore, we will implement the reviewer's suggested generalized linear model (GLM) approach using inferred spiking activity derived from the Ca2+ signals. Specifically, we will include both stimulus identity and freezing behavior as predictors. Because freezing varies across trials whereas stimulus presentation is fixed, this analysis will allow us to dissociate the relative contributions of stimulus-related versus freezing-related activity to the graded neuronal responses. We thank the reviewer for this excellent suggestion.

If graded stimulus coding remains significant after accounting for freezing behavior, this would strengthen the interpretation that these ensembles encode learned salience or associative properties of the stimuli rather than behavioral output alone. Conversely, if

freezing explains a substantial proportion of the variance, we will revise our interpretation accordingly.

> *(2) The study implies that the identified ensembles are causally related to valence memory, but no experimental interventions are performed to justify this.*

We appreciate the reviewer's point. We agree that our data are correlational in nature and that establishing a causal relationship between identified ensembles and valence memory would require experimental interventions such holographic two-photon manipulations, which are beyond the scope of the present study but represent an important direction for future work.

To provide an indirect link between ensemble organization and behavior within the constraints of the current dataset, we will examine inter-individual variability in the revised manuscript. Specifically, we will test whether the proportion of neurons participating in stable graded-response ensembles versus dynamic stimulus-specific ensembles predicts individual differences in freezing behavior and fear generalization across retrieval sessions. If animals with a higher proportion of stable graded-response neurons show stronger discrimination and less generalization to non-conditioned tones, this would strengthen the association between ensemble organization and behavioral outcome, while remaining correlational in interpretation.

We will modify the manuscript terminology accordingly, replacing causal language with phrasing that accurately reflects the associative nature of our conclusions.

References

(1) Aschauer, D.F., et al., Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization. Cell Rep, 2022. 38(6): p. 110340.

(2) Kato, H.K., S.N. Gillet, and J.S. Isaacson, Flexible Sensory Representations in Auditory Cortex Driven by Behavioral Relevance. Neuron, 2015. 88(5): p. 1027–1039.

(3) Vervliet, B., et al., Generalization gradients in human predictive learning: Effects of discrimination training and within-subjects testing. Learning and Motivation, 2011. 42(3): p. 210–220.

(4) Dunsmoor, J.E. and K.S. LaBar, Effects of discrimination training on fear generalization gradients and perceptual classification in humans. Behav Neurosci, 2013. 127(3): p. 350–6.

(5) Mante, V., et al., Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature, 2013. 503(7474): p. 78–84.

(6) Hockley, A. and M.S. Malmierca, Auditory processing control by the medial prefrontal cortex: A review of the rodent functional organisation. Hear Res, 2024. 443: p. 108954.

(7) Zikopoulos, B. and H. Barbas, Prefrontal projections to the thalamic reticular nucleus form a unique circuit for attentional mechanisms. J Neurosci, 2006. 26(28): p. 7348–61.

https://doi.org/10.7554/eLife.111177.1.sa0