Neural Geometric Representations of Social Memory for Multi-individuals in Medial Prefrontal Cortex

Tianyu Li[1], Xingzhi He[1], Xiang Gu[1], Xiaohui Zhang[1*]

[1] STATE KEY LABORATORY OF COGNITIVE NEUROSCIENCE AND LEARNING, IDG/MCGOVERN INSTITUTE FOR BRAIN RESEARCH, BEIJING NORMAL UNIVERSITY, BEIJING 100875, CHINA

* Correspondence: X-h Zhang (xhzhang@bnu.edu.cn)

Abstract

Social memory, an ability to recognize and recall acquainted individual and related experience, is critically dependent on the medial prefrontal cortex (mPFC). However, how the mouse mPFC neurons encode the memory for multiple social-individuals remain unknown. We, here, report that low-dimensional, tuned geometries of population neural activity in the mPFC stably represent individual social identities and animals maintain social discriminative memory over two weeks, using a novel multi-social discrimination task in which a mouse encountered four conspecifics paired with different condition stimulus. By *in vivo* recording from the mPFC in test mouse during multi-social discrimination task, only few neurons exhibit exclusive spiking responses to specific demonstrators. Instead, demonstrator identities were stably represented by distinct population-level subspaces in the mPFC across days. Furthermore, pairing the test mouse's social investigations to individual demonstrators with rewarding, aversive or neutral stimulus, respectively, substantially increased geometric distances and angles among these identity neural-subspaces. This enhanced representational separation is attributed to progressive recruitments of neurons encoding both social identity and associated valence in conditioned social experience. Thus, these findings elucidate that tuned neural activity subspaces in the mPFC stably encode acquired multiple social individuals over experience and provide a computational framework for understanding complex social cognition.

Introduction

Social recognition is evolutionarily conserved and fundamental to survival and reproduction across species. Studies using non-invasive techniques have demonstrated that human and other primates rely on conserved sensory cues (particularly visual) and integrated neural networks to recognize conspecifics[1-4]. Rodents, which are generally social animals, engage in behaviours such as communal nesting and cooperative care[5,6]. Conventional research on social memory has largely relied on the three-chamber assay, designed to assess binary discrimination between familiar and unfamiliar individuals or different emotional states[7-13]. However, this assay often conflates novelty-driven exploration with genuine social recognition and fail to disentangle neural representations of social identity from confounding factors such as spatial location[8,9,14,15]. Furthermore, its two-choice design fails to adequately capture natural social environments, where individuals must distinguish among multiple conspecifics. Thus, whether mice can discriminate between three or more individuals, and how such capacity is neural implemented, remains poorly understood.

Neural circuits underlying the regulation of social memory involve several brain regions, including the hippocampus[10,13,16-21], nucleus accumbens (NAc)[8,21], and medial prefrontal cortex (mPFC)[6,8,9,21-28]. In particular, the mPFC has been shown to play a key role in social information processing, such as storing social memories (engrams)[28], exhibiting dynamic activity during interactions[9], and supporting normal social behaviour[25]. Yet, functional spiking responses of mPFC neurons exhibited mixed selectivity to multiple behavioural or conditioning cues[14,15,23,29,30], which has complicated the interpretation of results from theses conventional binary social test paradigms. Furthermore, it remains unclear how mPFC neurons represent individual social identities include the mechanisms for stable population-level how the representations evolve with social experience, and how social memory guides individual-specific social preferences.

While previous studies have explored multi-mouse social interaction and recognition[16,24], the neural dynamics representation of social identity remains poorly described. In the present study, we developed a novel one-*versus*-four social discrimination task that extends behavioural and neural analyses beyond binary (one-versus-two) comparisons. Using chronic multi-wire array *in vivo* electrophysiological recordings in free-moving mice, we identified distinct functional classes of mPFC neurons, including those conjunctively encoding social and spatial features. By establishing a cross-session alignment approach based on regularized regression[31,32] and joint dimensionality reduction[33], we revealed stable, low-dimensional neural subspaces that uniquely encode individual demonstrators. These subspaces underwent experience-dependent reorganization during associative

social learning, correlating with shifts in social preference. Thus, our findings reveal a mechanistic basis for the function of social discrimination among multiple social individuals.

Result

mPFC Neurons Exhibit Diversely Selective Responses in the Multi-individual Social Test

The widely-used three-chamber test paradigms have generally failed to account for the modulation of neuronal activity by spatial variables and mixed selectivity [8,9,14,15,29,30], despite the fact that a substantial proportion of mPFC neurons encode a combination of social and location information[8]. To exclude spatial influence and determine whether mice can discriminate among multi-social demonstrators at the single-neuron level, we performed in vivo electrophysiological recordings in the mPFC using a 32-channel micro-wire array (Fig. 1a-d, Extended Data Fig. 1) in new designed paradigm, where mice freely investigated to four distinct novel sex-match and same age demonstrators from different cage or litter, each placed in a separate corner of a square open-field arena in two consecutive test with an interval of 30–60 minutes. In each test, to dissociate social information from spatial coding, each mouse underwent two tests, the positions of all demonstrators were changed (Fig. 1a). This design controlled for location-specific responses while minimizing effects of temporal memory and motivational decline. During these sessions, mice spent around 60% of the time dwell in social investigation zone (Fig. 1b), and among these, around 53% involved social behaviours (Fig. 1b). We recorded a total of 1800~ single neurons, and over 70% of mice contributing at least 20 well-isolated neurons per day (Fig. 1c-d).

To characterize neuronal selectivity based on their tuning to social and spatial features, we fit a regularized Lasso regression model to the firing rate of each unit, using social identity and corner position as predictors[31,32,34] (Fig. 1e). Model significance and generalizability were assessed via nested 10-fold cross-validation. Neurons were categorized into five functional groups (Fig. 1e-h): (1) **Pure social neurons** are significantly modulated by one or more specific social demonstrators across both sessions (3.2%, shuffle =0%);(2) **Pure location neurons** show significantly responsive to one or more arena corners across both sessions (4.5%, shuffle =0%); (3) **Conjunctive neurons** express significantly tuned to both social and spatial features across sessions (24.4%, shuffle =0%); (4) **Combinative neurons**: responded exclusively to a specific demonstrator in a specific location

(35.7%, shuffle =1.3%); (5) **Non-responsive neurons** are no significant tuning to any variable (32.3%, shuffle =98.7%). Consistent with earlier studies, these findings demonstrate the multi-selectivity properties of mPFC neurons, a large fraction of which act as combinative or conjunctive cells.

Among the 512 neurons classified as pure social or conjunctive, we further quantified social specificity and multi-selectivity. Approximately 46.5% neurons exhibited selective firing to only one demonstrator, while the remaining 53.5% neurons responded significantly to multiple individuals (Fig. 1i). The mixed tuning properties of mPFC neurons indicate that social memories are organized in a distributed code similar to working memory, a strategy that likely serves the functional purpose of maximizing the storage capacity for complex social information [14,30]. The distribution of different neuron types showed no significant differences or trends across days (Fig. 1h-i).

Together, these results indicate that only a little subset of mPFC neurons encode social identity in a highly specific manner. The majority of responsive neurons exhibit multi-selectivity or conjunctive tuning, underscoring the role of the mPFC in integrating diverse behavioural variables and confirming its highly mix-selective functional characteristics[8,14].

Neuronal Ensembles Encode Multiple Social Identities

Given the integrative and hub-like functional region of the mPFC in processing diverse information on social recognition[2,6,8,9,12,22,23,25,28,35], we next asked whether population activity in this region could support discrimination among four social individuals. We trained support vector machines (SVMs) using either linear or radial basis function (rbf) kernel, a kind of non-linear kernel, with 5-fold cross-validation (Fig. 2a). Both classifiers performed significantly above chance levels (shuffled accuracy: 25.0 ± 2.5%), confirming that population activity in the mPFC robustly encodes social identity (Fig. 2b-c). However, the rbf kernel consistently yielded higher F1-scores than the linear kernel (Fig. 2c), indicating that social representations are inherently nonlinear. This nonlinearity can be attributed to the prevalence of combinative and conjunctive neurons, which constituted the majority of responsive cells and are known to perform nonlinear integration of social and spatial variables. The F1-score showed no significant differences or trends across multiple days of recording (Fig. 2d). Together, mice have the ability to discriminate multi-individual conspecifics in population neuronal level.

Moreover, to evaluate the relative contribution of functionally distinct neurons to decoding performance, we compared social-related neurons (including social-responsive, conjunctive, and combinative neurons) with non-social neurons (location-tuned and non-responsive neurons). The feature importance of social-related neurons is significantly higher than non-social neurons. To compare population contributions, we performed random subsampling of equal numbers of neurons from each group. This revealed significantly higher decoding accuracy from social-related neurons (Fig. 2e, top). Conversely, randomly ablating an equal number of neurons from each group caused a more pronounced accuracy drop when social-related neurons were removed (Fig. 2e, bottom). These complementary results collectively confirm the dominant role of social-related neurons in supporting social discrimination. Thus, the social-related neurons contain more information of social conspecifics identity. Although social-related neurons encode more social information, it is an oversimplification to regard non-social neurons as entirely devoid of functional relevance. In fact, social memory is a highly complex cognitive process facilitated by the collaborative engagement of multiple neural principal components[36] (Extended Data Fig. 2).

In summary, these results demonstrate that the mPFC population enables discrimination of multiple individuals through nonlinear mixed selectivity, with socially tuned ensembles making the strongest contributions to social recognition.

Stable Neural Activity Subspaces for Identity Representation Across Days

Having established the ability to distinguish among social individuals at the population level, we next asked how the mPFC encodes and maintains representations of social identity across multiple days. Comparing the neurons activity across days in the mPFC is challenging due to: (1) the mix encoding of multiple behavioural variables[29], introducing noise during dimensionality reduction, and (2) the inability to track the same subset of neuron ensemble across sessions using micro-wire array electrode, because of electrode tips drifting , confusing neuron sorting and complicating cross-day alignment. To address these issues, we adopt neural geometry perspective to analyse neural activity, because many studies have identified behaviorally relevant organization within the full neural state space, whose organizational structure persists over long time periods performing the same task when the recorded neurons change and even across different individual[37]. We developed a novel method combining regularized regression and joint dimensionality reduction. First, we employed nonlinear dimensionality reduction (Cebra[38], multi-session mode) to align neural data from different days into

a common embedding space (Fig. 3a, Extended Data Fig. 3c-f). This analysis revealed that the neural embeddings exhibited a highly similar structure across days. Then, we used a Lasso-regularization (L1 regularization) linear model to isolate firing rates related only to social identity and location, effectively denoising the neural responses (see Methods). We then concatenated regression coefficient matrices across days into a unified design matrix and performed joint singular value decomposition (SVD) to align all sessions into a low-dimensional subspace (Fig. 3b). The similarity of dimensions is high, so all recording data from different day in same mouse can embed in low dimensional subspace further is comparable across day (Fig. 3c).

Within the neural geometry framework, distinct stimuli engage specific low-dimensional subspaces in population activity, forming stable stimulus-specific representations[37]. We next asked whether each social demonstrator is stably represented within a specific neural subspace. For example, for Demo 1, we defined its putative subspace using activity from both of its occupied locations across the two tests (Fig. 3d). After joint SVD, these activity patterns formed an orthogonal basis defining a demonstrator-specific subspace. Within-day subspace similarity was moderate (Variance Accounted For, VAF ≈ 60%) (Fig. 3e), but across days, the same demonstrator's subspace was highly stable, with VAF≈ 93% (Fig. 3f). This suggests a stable neural subspace for each social individual across day. After permutation test, we found 83% of pairwise comparison, was significantly stable (Fig 3g, Extended Data Fig 3). Moreover, we get similar result when we use non-linear dimension reduction (Extended Data Fig. 3c-f).

Furthermore, we found that subspaces derived from social-related neurons were significantly more stable across days than those from non-social neurons (Extended Data Fig 4). This reveal that the stable neural subspace for each social individual is statistically significant, and social-related neurons contribute more information to relative subspace. The inhibition of behavioural paradigm-related neurons using cFos-TRAP and chemo-genetics disrupted the neural embedding. Consequently, the dataset from the chemo-inhibition failed to align with those from the untreated group and control. Thus, our results establish a key role for mPFC behavioural paradigm-related neurons in organizing the neural subspaces (Extended Data Fig 5).

Together, these results demonstrate that social identity is encoded in stable, low-dimensional neural subspace. Thus, our data establish a principle of stable subspace encoding for social identity, further supported by the observation that interactions with the same individual engage similarly stable geometries.

Conditioning Social Learning Enhances Social Discrimination to Multi-individuals

Having established neural discrimination of social individuals at single neurons, population, and geometry levels, it is necessary to determine whether this discrimination is evident at the behavioural level. Social memory, social experiment and social history can flexibly influence valence assignment to social stimuli[1]. Based on these, we developed a "multi-individual social learning" paradigm, consisting of three phases (Fig. 4a): (1) In the pre-training phase, test mice underwent repeated social tests to determine baseline preferences for four unfamiliar conspecifics after habituation for 4 days (Fig. 4 b, c), avoiding regression-to-the-mean effects. The most preferred demonstrator was designated the Demonstrator Aversive (to avoid ceiling effects), and the least preferred became the Demonstrator Reward (to avoid floor effects). The remaining two served as neutral controls (Demonstrator Neutral 1 and Demonstrator Neutral 2). (2) During the 8 days conditioning phase, test mice received air puffs when interacting with the aversive mouse and milk rewards with the rewarded mouse. No treatment was applied to the neutral mice. To control for order and location biases, demonstrator positions and trial sequences were randomized across day. (3) Post-training tests, same procedure as pre-conditioning test, revealed a significant increase in preference for the rewarded mouse and a decrease for the aversive mouse (Fig. 4c), while neutral individuals elicited no change.

As human and other primates discriminate conspecific most rely on visual sensory input, we identify sensory modalities supporting this discrimination by tested mice in complete darkness and found no impairment in preference modulation (Fig 4.f-g). However, in mice lacking hormone-sensing receptors (*Trpc2*-null)[6,39,40], the preference changes were abolished (Fig. 4e), implicating chemosensory cues as necessary for individual discrimination. So, it demonstrated that hormone perception pathway plays the key role in multi-individual recognition. We next assessed the duration of this social memory. The observed preference shifts persisted for at least two weeks (Fig. 4f), indicating the formation of long-term social memory. This demonstrates that mice can discriminate between multiple individuals and that this memory endures over a long-term period, which is distinct from the short-term memory (lasting approximately 12 hours) reported in previous study[13]. And we also test the TRAPed behavioural paradigm-relative mPFC neurons plays the key role in this memory[28,41] (Extend data fig. 6).

Together, these results demonstrate that mice can behaviourally discriminate among at least three

individuals and form robust, long-term valence-specific memories mediated primarily by olfactory signals.

Reorganization of Neural Geometric Representations in Social Memory

An animal's social history (for example, experience of air puff or milk reward during social interaction) can alter the attractive or repulsive value of social stimuli and, consequently, bi-directionally modulate the motivation to seek social contact[1,35]. We next tested how associative conditioning altered neural representations in the mPFC. At the single-unit level, we observed a decrease in the proportion of non-responsive neurons (Fig. 5a, b). This indicates the recruitment of more neurons in behavioural paradigm.

Population decoding further revealed enhanced discriminability after training. SVM classifiers achieved significantly higher accuracy in distinguishing among all individuals (Fig. 5 c), with the most pronounced improvement for the Reward and Aversive demonstrator (Fig. 5 d-g). This suggests that associative conditioning improves population-level representations, particularly for individuals associated with valence outcomes.

Together, these findings indicate that preference shifts following learning enhanced population-level discriminability—especially for reward-associated and aversive-associated social targets.

Post-training reorganization of social coding subspaces

We next asked whether associative conditioning induces reorganization within the neural geometries encoding social information. Although the overall neural geometry remained highly similarity before and after conditioning (Fig. 5h), demonstrating that population activity can be embedded and compared in a consistent low-dimensional space across days, we identified specific and meaningful changes within demonstrator-specific subspaces (Fig. 5i-m). The neuronal ensemble encoding a specific stimulus demonstrated remarkable stability. We observed a 70% overlap between the original c-Fos-TRAPed population and the c-Fos-positive cells activated during the recall of the same demonstrator in the match group. This overlap was significantly greater than the 40% observed

in the mismatch control group, which is recalled by different demonstrator (Extend data fig. 7).

Notably, the angle and distance between the Reward and Aversive subspaces significantly increased at post-conditioning phase, indicating greater neural separation of these valence-opposed individuals (Fig. 5 i-j). This separation was further supported by a decrease in VAF, confirming increased distinctiveness between these neural representations (Fig. 5 m). In addition, the modulation index, reflecting the population-level response magnitude to a given stimulus, was markedly elevated in the Reward subspace, consistent with improved population-level decoding accuracy (Fig. 5 d) and indicating stronger engagement of neurons representing the rewarded individual (Fig. 5 k). Furthermore, significant angular separation and reduced shared variance (VAF) were also observed between the Aversive and Neutral subspaces, suggesting that discriminability among these representations was likewise enhanced.

Together, these results indicate that learning drives selective reorganization within social respective subspaces, enhancing separation of behaviourally salient individuals while refining representations among others, thus facilitating more precise social discrimination.

Evolution of Conditioning Responsive Neurons to Social Neurons during Learning

At the learning timescale, neural geometry properties likely change to help improve the ability to perform certain behaviours through learning[37]. To understand how conditioning stimulus reshapes social representations during conditioning phase, we analysed neural activity throughout the training period. We used Elastic-Net regression (L1 + L2 regularization) to capture complex temporal responses to discrete events (e.g., air puff, milk reward) while performing feature selection[29] (Extend data fig. 8 a). Spatial effect is controlled because all demonstrators occurred in a fixed location. Neurons were classified into four types (Fig. 6 a-d): (1) **Social-response neurons**: significantly modulated during social interaction but unresponsive to conditioning stimuli; (2) **Conditioning-response neurons**: responsive only to reward or aversive stimuli, not to social demonstrators; (3) **Joint-response neurons**: significantly tuned to both a specific demonstrator and its associated conditioning stimulus; (4) **Non-response neurons**: unmodulated throughout. The proportion of Joint-response neurons increased gradually across days (Fig. 6 e), while other classes remained no significant change (Fig. 6 e), indicating a learning-dependent shift toward integrated encoding of

social and valence-events.

Modulation indices across conditions revealed a hierarchical representational structure: conditioning stimuli (milk and air puff) showed the highest modulation (Fig. 6 h), followed by valence-relative demonstrators (Reward and Aversive) (Fig. 6 i), with two neutral demonstrators exhibiting the lowest activity (Fig. 6 j). Across sessions, modulation increased for milk reward and the Reward demo but decreased for air puff, reflecting a strategic shift from avoiding punishment to pursuing reward, consistent with improved population-level decoding accuracy (Fig. 5 d) and subspace changing after conditioning (Fig. 5 k).

Subspace analysis revealed several learning-related dynamics (Fig. 6 k-n): (1) The angle between Reward and Aversive subspaces remained stable and near-orthogonal throughout training, indicating immediate and sustained discrimination of these salient individuals (Fig. 6 k); (2) Neutral demonstrators (N1 and N2) showed higher shared variance and less distinct separation, consistent with their behavioural irrelevance (Fig. 6 l); (3) The Reward and milk reward subspaces were orthogonal throughout, indicating stable unrelative between social and reward representations (Fig. 6 m); (4) In contrast, the Aversive and air-puff subspaces converged toward orthogonality over time, suggesting that mice gradually learned to dissociate the aversive stimulus from the social context (Fig. 6 n). The demonstrator Aversive shift separative to other demonstrator subspace during conditioning phase (Extend data fig.8).

Together, these results demonstrate that learning drives increased integration of social and valence information at the single-unit level, while subspace dynamics reflect a strategic reorientation and change the preference of respective demonstrator.

Discussion

Social interaction is fundamental to survival and reproductive success across species, facilitating cooperative foraging, predator defence, mate selection, parental care, and environmental adaptation[42]. Rodents display various social behaviours[1,2,42], and many different behavioural paradigms have been used for investigating their neural circuit bases. Social behaviour is highly vulnerable to brain network dysfunction caused by neurological and neuropsychiatric conditions

such as autism spectrum disorders (ASDs)[21,25,43]. Central to complex social behaviours is the ability to recognize and distinguish among multiple individuals. Here, we developed a novel behavioural paradigm for multi-individual social discrimination and combined it with *in vivo* electrophysiology to reveal stable, distinct neural subspaces for the representation of different social individuals. Furthermore, we show that valence-conditioning induces representational drift within these subspaces, enhancing discriminability among individuals.

Previous work has highlighted the functional complexity of mPFC neurons—particularly projection neurons—which often exhibit mixed selectivity to diverse behavioural variables[9,14,15,23,30]. This likely reflects the convergent inputs from and divergent outputs to numerous brain regions. It has been proposed that different subpopulations within the mPFC may encode social valence via distinct projection targets. Our results not only confirm this complexity but also provide the direct evidence for a high-dimensional subspace organization of social representations in the mouse mPFC.

Conventional social memory studies often rely on binary (one-*versus*-two) discrimination tasks[7-13]. However, as suggested by earlier reports, such designs are prone to false positives due to the multi-variable encoding properties of mPFC neurons. Moreover, cortex-wide activity was dominated by movements, especially uninstructed movements not required for the task, which contribute to noise in the electrophysiology recording dataset. Our approach—using multiple tests with counterbalanced demonstrator positions—allowed us to dissociate pure social neurons, pure spatial neurons, conjunctive neurons and combinative neurons, offering a more refined dissection of social-spatial integration.

Using a new multi-individual conditioning paradigm, we induced long-lasting shifts in social preference that persisted for at least two weeks. This extended durability contrasts with some previous reports suggesting social memories last only hours[10], though it aligns with findings from head-fixed paradigms that also show longer stability[16]. These differences may arise from variations in behavioural design, salience of conditioning stimuli, or neural recording techniques.

The sensory basis of social discrimination is another key aspect. Social recognition in mice is inherently multimodal: auditory cues convey emotional and gender information, while chemosensory signals communicate kinship, health, and danger[20]. Even in humans with reduced olfactory sensitivity, pheromonal cues continue to operate subconsciously[20]. Our study indicates that visual

input is dispensable for long-term social preference shifts, whereas hormonal/chemosensory signal, mediated by both the main and accessory olfactory systems, is essential. Consistent with this, *Trpc2*-knockout mice, which lack functional vomeronasal transduction[40], exhibit profound deficits in social and sexual discrimination.

From a methodological standpoint, analysing mPFC electrophysiology across days presents major challenges: electric-wire drift and heterogeneous tuning obscuring population codes. We provided a new analytical framework combining regularized regression and joint dimensionality reduction to overcome these obstacles. This approach revealed stable social subspaces that undergo systematic angular shifts during learning, refining social representations.

Social valence is a dynamically computed assessment of a social stimulus as positive or negative, which is flexibly determined by the integration of multiple factors including social alignment, attributes, memory, rank, and isolation history[1]. This multifactorial process, implemented by overlapping neural circuits governing social cognition and valence, enables adaptive behavioral responses such as approach or avoidance and is crucial for survival and often dysregulated in psychiatric disorders[20]. Our results show that the positive or aversive experience triggers a shift in neural geometry, which underlies the change in social preference for the demonstrator.

The brain is a highly complex, self-organizing system, not merely a passive machine that reacts to external stimuli. Furthermore, the dynamic processes within neuronal populations are crucial for researchers to understand the representation of memory. The emergence of stimulus-specific, cross-day stable subspaces, and their experience-dependent separation, aligns powerfully with attractor dynamics in manifold frameworks[37]. In such frameworks, memories are stored as low-energy attractor states shaped by recurrent connectivity and Hebbian plasticity. The progressive divergence of social representations with learning suggests repulsion between attractors to minimize interference, thereby increasing pattern separation. This continuous refinement of neural geometry via adaptive plasticity offers a more dynamic account of social memory than static engram theories, highlighting how learning sculpts neural manifolds for computational efficiency. Our result support these perspectives, that social memory shape a stable geometry.

Concluding Remarks

In summary, our study advances our understanding of social memory on multiple fronts: we

introduce a robust behavioural paradigm for multi-individual discrimination, identify distinct and stable neural subspaces for social representations, demonstrate experience-dependent reorganization of these subspaces, and resolve the critical role of chemo-sensation—rather than vision—in supporting long-term social recognition. Our findings argue for a dynamic, distributed and subspace-based neural encoding strategy within the mPFC that supports complex social behaviour over extended timescales. These results not only clarify previously contradictory findings in social memory duration but also provide a new analytical framework for studying population coding across days in highly heterogeneous brain regions. The interplay between sensory input, learning-induced plasticity, and attractor-like dynamics in neural manifolds opens new avenues for investigating how brains represent and remember social worlds.

Limitation

While our study provides mechanistic insights into social representation, several limitations should be considered. First, our cross-day alignment approach, while innovative, relies on linear dimensionality reduction and may not fully capture all nonlinear neural dynamics across sessions. Future work employing large-scale calcium imaging could more directly track the same neurons over time to validate our approach. Second, all experiments were conducted in adult male mice. Whether similar subspace coding principles govern social learning in females or across different ages remains an open and important question for future investigation. Finally, our behavioural paradigm is that the interactions between the demonstrator and test mice were not completely free and unrestricted, and thus may not fully recapitulate the ethologically rich and dynamic social encounters that occur in naturalistic settings.

Acknowledgements We would thank Dr. Qiang Li (SJTU) for providing the *Trpc2*-null mice and Dr. Haitao Wu (BIBMS) for technical help on whole-brain section imaging. This work was supported by grants from the Scientific & Technological Innovation (STI) 2030-Major Project (2022ZD0204900 to X-h.Z.) and the National Natural Science Foundation of China (32130043 to X-h.Z.)

Author contributions X.Z. conceived the project and supervised the work. T-y.L. designed and conducted all experiments and data analysis. X-z.H. did viral injection experiment. X.Z and T.L. wrote the manuscript with comments from all other authors.

1 Padilla-Coreano, N., Tye, K. M. & Zelikowsky, M. Dynamic influences on the neural encoding of social valence. *Nature Reviews Neuroscience* **23**, 535-550, doi:10.1038/s41583-022-00609-1 (2022).

2 Kanterman, A. & Shamay-Tsoory, S. From social effort to social behavior: An integrated neural model for social motivation. *Neuroscience & Biobehavioral Reviews* **173**, doi:10.1016/j.neubiorev.2025.106170 (2025).

3 Atzil, S., Gao, W., Fradkin, I. & Barrett, L. F. Growing a social brain. *Nature Human Behaviour* **2**, 624-636, doi:10.1038/s41562-018-0384-6 (2018).

4 Shamay-Tsoory, S. G., Saporta, N., Marton-Alper, I. Z. & Gvirts, H. Z. Herding Brains: A Core Neural Mechanism for Social Alignment. *Trends in Cognitive Sciences* **23**, 174-186, doi:10.1016/j.tics.2019.01.002 (2019).

5 Sun, F., Wu, Y. E. & Hong, W. A neural basis for prosocial behavior toward unresponsive individuals. *Science* **387**, doi:10.1126/science.adq2679 (2025).

6 Chen, P. & Hong, W. Neural Circuit Mechanisms of Social Behavior. *Neuron* **98**, 16-30, doi:10.1016/j.neuron.2018.02.026 (2018).

7 Qi, C. *et al.* Anterior cingulate cortex parvalbumin and somatostatin interneurons shape social behavior in male mice. *Nature Communications* **16**, doi:10.1038/s41467-025-59473-z (2025).

8 Murugan, M. *et al.* Combined Social and Spatial Coding in a Descending Projection from the Prefrontal Cortex. *Cell* **171**, 1663-1677.e1616, doi:10.1016/j.cell.2017.11.002 (2017).

9 Liang, B. *et al.* Distinct and Dynamic ON and OFF Neural Ensembles in the Prefrontal Cortex Code Social Exploration. *Neuron* **100**, 700-714.e709, doi:10.1016/j.neuron.2018.08.043 (2018).

10 Hitti, F. L. & Siegelbaum, S. A. The hippocampal CA2 region is essential for social memory. *Nature* **508**, 88-92, doi:10.1038/nature13028 (2014).

11 Fujima, S., Sato, M., Nakai, N. & Takumi, T. Parvalbumin interneurons in the insular cortex control social familiarity and emotion recognition. *Cell Reports*, doi:10.1016/j.celrep.2025.116085 (2025).

12 Chen, Z. *et al.* A prefrontal-thalamic circuit encodes social information for social recognition. *Nature Communications* **15**, doi:10.1038/s41467-024-45376-y (2024).

13 Okuyama, T., Kitamura, T., Roy, D. S., Itohara, S. & Tonegawa, S. Ventral CA1 neurons store social memory. *science* (2016).

14 Pinto, L. & Dan, Y. Cell-Type-Specific Activity in Prefrontal Cortex during Goal-Directed Behavior. *Neuron* **87**, 437-450, doi:10.1016/j.neuron.2015.06.021 (2015).

15 Khilkevich, A. *et al.* Brain-wide dynamics linking sensation to action during decision-making. *Nature* **634**, 890-900, doi:10.1038/s41586-024-07908-w (2024).

16 Kong, E., Lee, K.-H., Do, J., Kim, P. & Lee, D. Dynamic and stable hippocampal representations of social identity and reward expectation support associative social memory in male mice. *Nature Communications* **14**, doi:10.1038/s41467-023-38338-3 (2023).

17 Ray, S. *et al.* Hippocampal coding of identity, sex, hierarchy, and affiliation in a social group of wild fruit bats. *Science* **387**, doi:10.1126/science.adk9385 (2025).

18 Villegas, A. & Siegelbaum, S. A. Modulation of aggression by social novelty recognition memory in the hippocampal CA2 region. *Cell Reports* **44**, doi:10.1016/j.celrep.2025.115992

(2025).
19 Watarai, A., Tao, K. & Okuyama, T. Representation of sex-specific social memory in ventral CA1 neurons. *Science* **389**, doi:10.1126/science.adp3814 (2025).
20 Zorab, J. M. *et al.* Serotonin and neurotensin inputs in the vCA1 dictate opposing social valence. *Nature* **642**, 154-164, doi:10.1038/s41586-025-08809-2 (2025).
21 Sato, M., Nakai, N., Fujima, S., Choe, K. Y. & Takumi, T. Social circuits and their dysfunction in autism spectrum disorder. *Molecular Psychiatry* **28**, 3194-3206, doi:10.1038/s41380-023-02201-0 (2023).
22 Wittmann, M. K. *et al.* Basis functions for complex social decisions in dorsomedial frontal cortex. *Nature* **641**, 707-717, doi:10.1038/s41586-025-08705-9 (2025).
23 Levy, D. R. *et al.* Dynamics of social representation in the mouse prefrontal cortex. *Nature Neuroscience* **22**, 2013-2022, doi:10.1038/s41593-019-0531-z (2019).
24 Li, S. W. *et al.* Frontal neurons driving competitive behaviour and ecology of social groups. *Nature* **603**, 661-666, doi:10.1038/s41586-021-04000-5 (2022).
25 Mediane, D. H., Basu, S., Cahill, E. N. & Anastasiades, P. G. Medial prefrontal cortex circuitry and social behaviour in autism. *Neuropharmacology* **260**, doi:10.1016/j.neuropharm.2024.110101 (2024).
26 Ko, J. Neuroanatomical Substrates of Rodent Social Behavior: The Medial Prefrontal Cortex and Its Projection Patterns. *Frontiers in Neural Circuits* **11**, doi:10.3389/fncir.2017.00041 (2017).
27 Liu, Y.-c. *et al.* Oxytocin modulates inhibitory balance in the prelimbic cortex to support social memory consolidation during REM sleep. *Theranostics* **15**, 3257-3274, doi:10.7150/thno.109104 (2025).
28 Xing, B. *et al.* A Subpopulation of Prefrontal Cortical Neurons Is Required for Social Memory. *Biological Psychiatry* **89**, 521-531, doi:10.1016/j.biopsych.2020.08.023 (2021).
29 Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience* **22**, 1677-1686, doi:10.1038/s41593-019-0502-4 (2019).
30 Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585-590, doi:10.1038/nature12160 (2013).
31 Yang Xie, P. H., Junru Li, Jingwen Chen, Weibin Song, Xiao-Jing Wang, Tianming Yang, Stanislas Dehaene, Shiming Tang, Bin Min, Liping Wang. Geometry of sequence working memory in macaque prefrontal cortex. *Science*, doi:10.1126/science.abm0204 (2022).
32 Golub, Å. B. a. G. H. Numerical Methods for Computing Angles Between Linear Subspaces. *MATHEMATICS OF COMPUTATION* **27** (1973).
33 Ehret, B. *et al.* Population-level coding of avoidance learning in medial prefrontal cortex. *Nature Neuroscience* **27**, 1805-1815, doi:10.1038/s41593-024-01704-5 (2024).
34 Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* **58**, 267-288 (1996).
35 Olsson, A., Knapska, E. & Lindström, B. The neural and computational systems of social learning. *Nature Reviews Neuroscience* **21**, 197-212, doi:10.1038/s41583-020-0276-4 (2020).
36 Voitov, I. & Mrsic-Flogel, T. D. Cortical feedback loops bind distributed representations of

working memory. *Nature* **608**, 381-389, doi:10.1038/s41586-022-05014-3 (2022).

37 Perich, M. G., Narain, D. & Gallego, J. A. A neural manifold view of the brain. *Nature Neuroscience* **28**, 1582-1597, doi:10.1038/s41593-025-02031-z (2025).

38 Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360-368, doi:10.1038/s41586-023-06031-6 (2023).

39 Leypold, B. G. *et al.* Altered sexual and social behaviors in trp2 mutant mice. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6376-6381 (2002).

40 Stower, L., Holy, T. E., Meister, M., Dulac, C. & Koentges, G. Loss of Sex Discrimination and Male-Male Aggression in Mice Deficient for TRPC2. *Science* **295** (2002).

41 Guenthner, Casey J., Miyamichi, K., Yang, Helen H., Heller, H. C. & Luo, L. Permanent Genetic Access to Transiently Active Neurons via TRAP: Targeted Recombination in Active Populations. *Neuron* **78**, 773-784, doi:10.1016/j.neuron.2013.03.025 (2013).

42 Li, H., Zhao, Z., Jiang, S. & Wu, H. Brain circuits that regulate social behavior. *Molecular Psychiatry* **30**, 3240-3256, doi:10.1038/s41380-025-03037-6 (2025).

43 Pruitt, A., Gupta, A. R. & Hoffman, E. J. Molecular and Genetic Mechanisms in Autism Spectrum Disorder. *Annals of Neurology*, doi:10.1002/ana.70013 (2025).

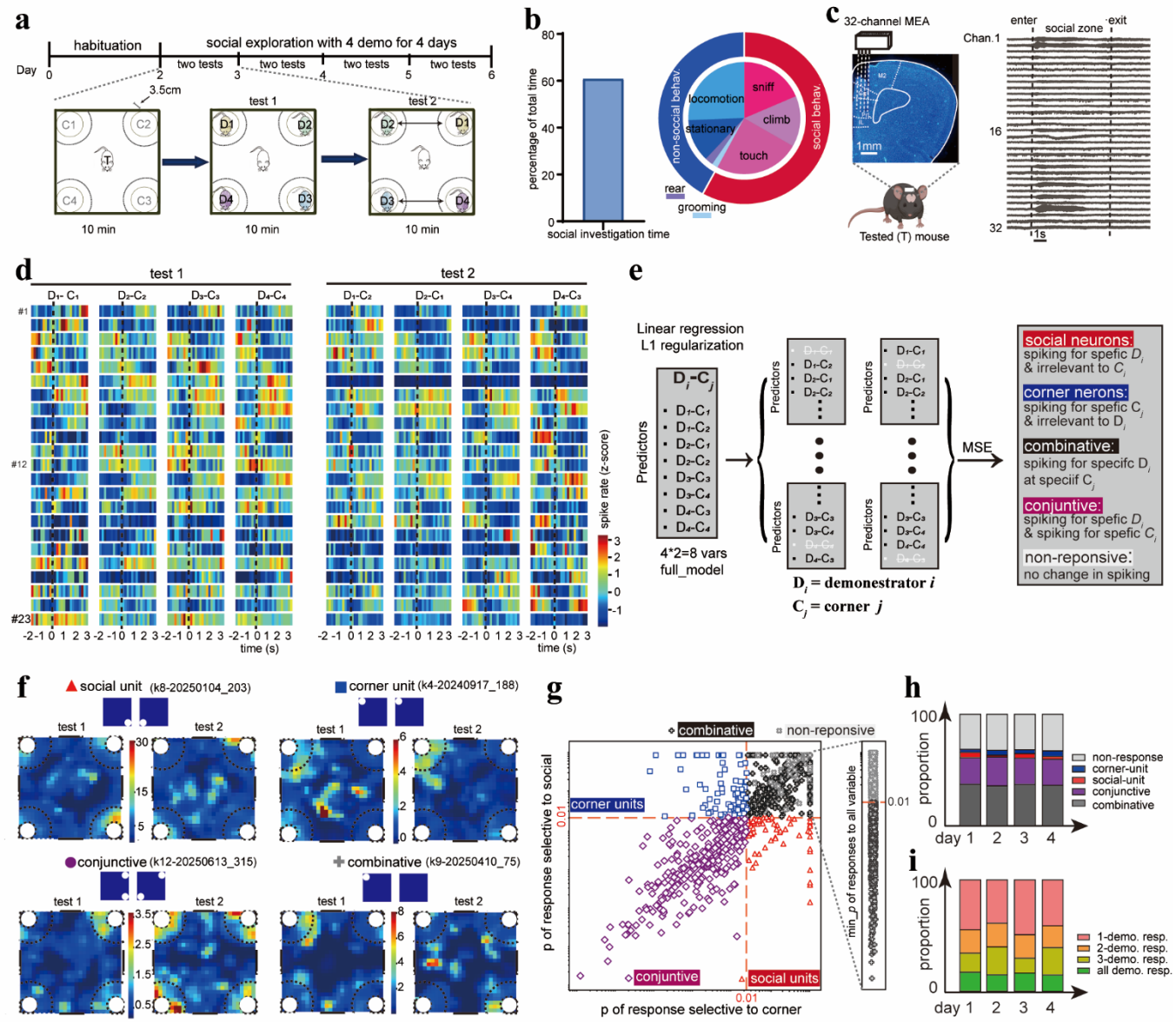


Fig. 1. Neural response features of individual mPFC neurons in the multi-social test. **a,** Schematic of the four-demonstrator social exploration test. **b,** The proportions of a test animal's time spent in the social zone (left) and related social exploration behaviours (right, red) as well non-social explorations (blue) in the 10 min test time. **c,** *Left*: Schematic illustration of implantation of a 32-channel micro-wire array (MEA) in the mPFC. *Right:* Raw traces of *in vivo* recorded neural activities from individual 32 channels, covering a complete time duration spent in the social zone (dotted line: time points for entering and exiting, respectively). **d,** Changes of spiking rates (*z*-scores) of representative mPFC single-units in the social zone in the consecutive social-test 1 and 2 sesssions, respectively. Dotted gray lines indicate the time of entering the social zone. **e,** A diagram illustrating the procedure of utilizing the Lasso Regression to classify neuronal units encoding distinct spatial or social features. D*i* and C_j depict the demonstrator mice *i* and the corner *j*, respectively. **f,** Representative pure social or corner neuron, conjunctive neuron and combinative neuron, respectively from the analysis in the **e**, indicated by distinct heat-maps of their mean spiking rates at different arena locations between the social-test 1 and 2 sessions. **g,** Distributions of *p*-

values for clarifying the featured social neurons (red triangles), corner neurons (blue squares), conjunctive neurons (purple circles), combinative neurons (dark grey), and non-responsive units (light grey), based on the criterion of $p < 0.01$. **h,** Changes of the proportions of distinct featured units across the 4 days of social test. **i,** Changes of the proportions of social units encoding demonstrator 1, 2, 3 and 4, respectively, across 4 test days.

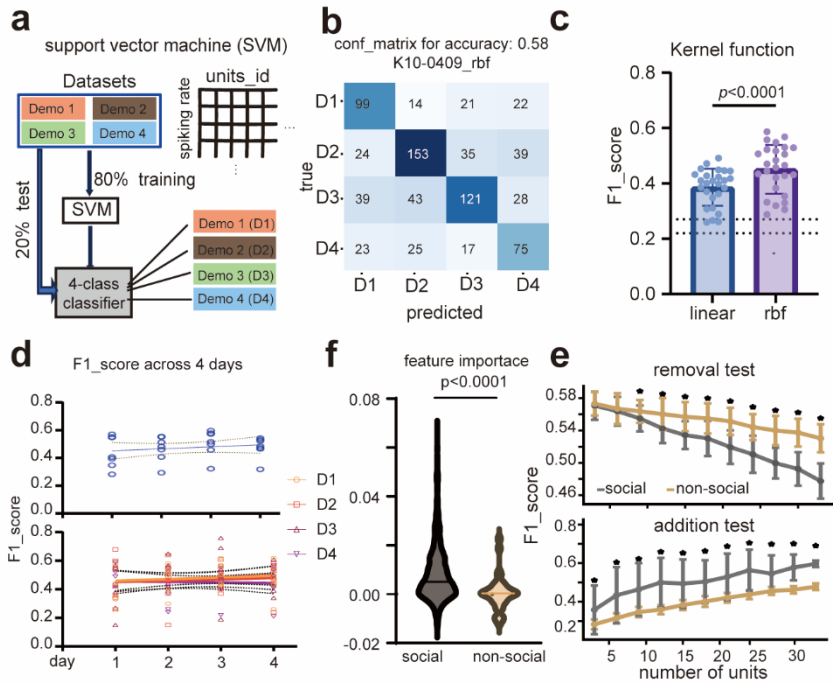


Fig. 2. Decoding multiple individuals based on mPFC population neuronal activities.

a, Schematic diagram illustrating the use of support vector machine (SVM) classifier to test the accuracy in discriminating multi-individuals (Demo *i*) on the basis of neuronal spiking activities recorded during the test-mouse dwelling within the social zone. Spiking trains of all neurons were binned with a 250-ms sliding window, and 80% and 20% were used for the classifier training and test, respectively, in which all class were balanced. **b,** An example of the confusion matrix calculated from the SVM model with RBF kernel (test session). **c,** Comparison of calculated F1 scores from the SVM classifiers with a linear or a radial basis function (rbf) kernel (*t*-test). Dotted lines: the shuffled F1 score (0.25±0.04). **d,** Changes of the averaged F1_score (*upper*) or the individual F1_scores (*bottom*) of the SVM for classifying distinct demonstrators across days. Dots represent individual tested mice. **f,** Comparison of the feature importance values of these social-related neurons (including the pure social, conjunctive, and combinative neurons) *versus* those non-social neurons (including the corner and non-responsive neurons). **e,** Changes of the classification accuracy in discriminating distinct demonstrators when the SVM model were trained with continuous ablations (*upper*) or additions (*bottom*) of social-related neurons (yellow) and non-social neurons, respectively. *: $p < 0.01$, *t*-test.

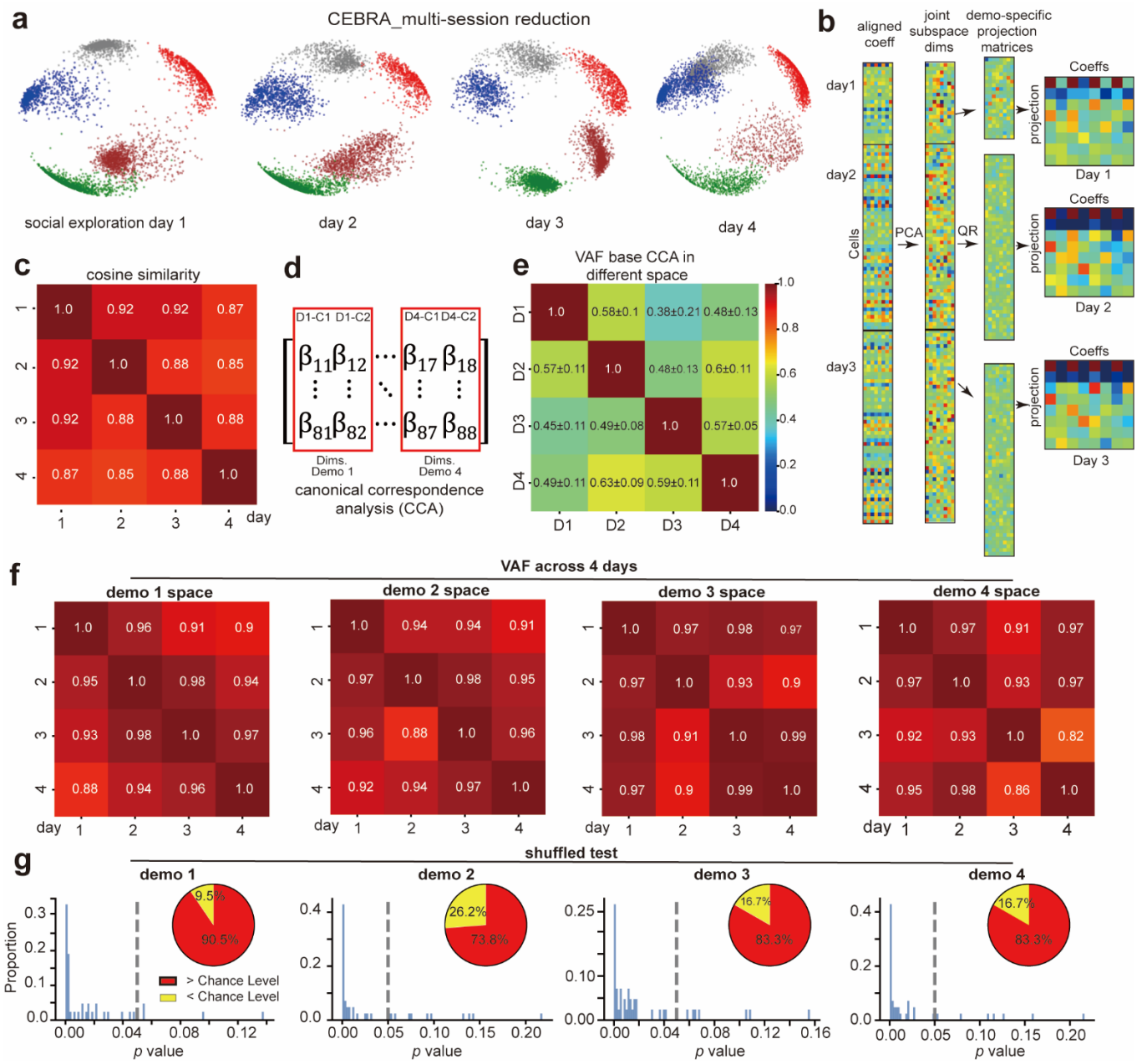


Fig. 3. Stable neural representations of individual identity in the mPFC.

a, Stable cluster structures of neuronal activities in the mPFC across the 4 test days, revealed by the CEBRA multi-session reduction. Clusters in colours represent units that spiked within the social zone and encoded distinct 4 demonstrators, while clusters in gray are ones that spiked outside the social zone. **b,** A schematic diagram illustrating a unified dimensionality reduction framework based on the joint Singular Value Decomposition (SVD). Matrices from different days were concatenated along the aligned "demonstrator × location" variable dimension and projected into a unified low-dimensional space, in which the horizontal and vertical axis represent the coefficients and the projections as 'super neuron' capturing the overall encoding patterns of neuronal population. **c,** Cosine similarity between neural representations across days, calculated from the **b**. **d,** A diagram

illustrating the Schematic of Canonical Correlation Analysis (CCA) to analyse cross-day stability of subspaces, constructed by the dimensionality reduction of the two projection arrays, 'D*i*- C_1' and 'D*i*-C_2'. D: demonstrator, C: corner. **e,** The mean Variance Accounted For (VAF) between subspaces for distinct social identities. **f,** The VAF for each social subspace across 4 days. **g,** Distributions of *p*-values for statistic comparison between the true VAF and the shuffled VAF (1000 times). Inserts: the proportions of shuffled data points above (red) and below (yellow) the chance level. Dashed lines indicate the α value at 0.05. See the Extended Table 1 for detailed statistic results.

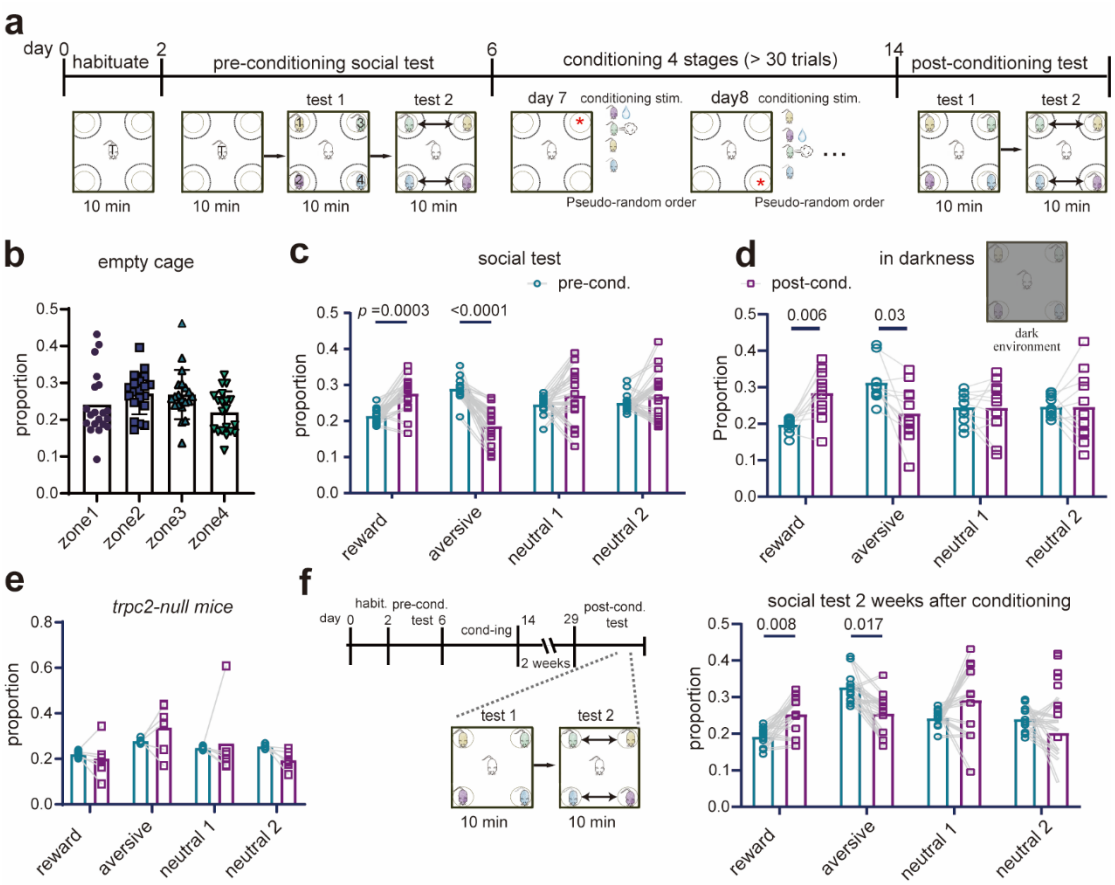


Fig. 4. Recognition and memory of multi-social individuals tested by a social conditioning learning.

a, The social conditioning learning for multi-social individuals. The social conditioning is to associate milk reward (R), aversive air puff (A), and blank consequence with social investigations to 4 individual demonstrators in the test mice. The social conditionings with individual demonstrators were conducted at pseudo-random orders and at randomly selected corners across 8 days, followed by a post-condition multi-social test to evaluate the change in the rank of social preferences to individual demonstrators. **b,** No spatial preference to individual locations of empty houses in the test mice, indicated by the proportions of exploration time spent on each corner. **c,** Changes of the proportions of time for social investigations to 4 individual demonstrators before (pre-cond.) and after the social conditionings shown in the **a**. **d,** Same as the **c** except the post-cond. social test was conducted in darkness. **e,** Same as the **c** except use of the *Trpc2-null* test mice. **f,** The protocol (*left*) and results (*right*) of conditioned social discrimination memory tested 2 weeks after the social conditioning in the wild type mice. The *p* values were calculated by the Two-way ANOVA with Bonferroni's multiple comparisons test (see detailed information in the Extended Table 1).

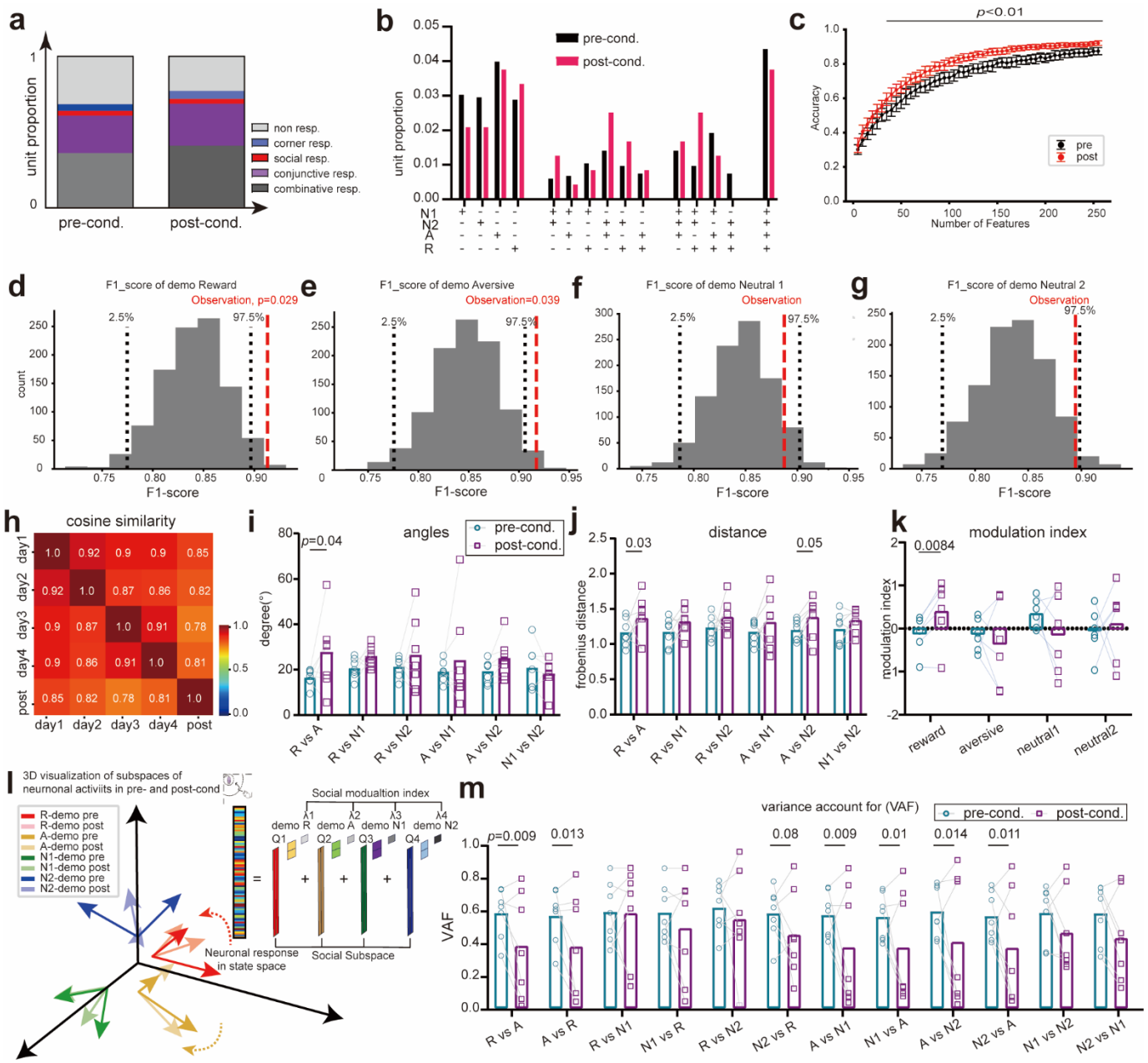


Fig. 5. Social conditioning-induced selective enhancement of neural-activity representations of the reward demonstrator and Aversive demonstrator.

a, Changes of the proportions of different types of recorded neurons showing distinct selectivity in the multi-social tests before and after the social conditioning. Resp.: responses. Note that the proportion of non-responsive neurons significantly decreased (chi-square test, $\chi^2 = 10.30$, $p = 0.036$; $Z_{non\text{-}resp.} = 3.168$, $p_{non\text{-}resp.} = 0.0076$). **b,** No significant changes in the proportions of social-resp. units encoding distinct demonstrators before and after the social conditioning (chi-square test, $\chi^2 = 12.12$, $p = 0.67$). **c,** Consistently higher decoding accuracies in the SVM using the post-conditioning units than that using the pre-conditioning units, with increasing numbers of units (*: $p < 0.01$, t-test). **d–g,** F1 scores for population unit activity decoding of distinct Reward, Aversive,

Neutral 1 and 2 demonstrator before and after conditioning, indicating selective increases in the accuracy of decoding the Reward and Aversive demonstrators. **h,** Stable population activity representations of social individuals across pre- and post-cond. test days, indicated the cosine similarity in the joint SVD. **i–j,** Change of the angles and distances between distinct neural subspaces encoding Reward, Aversive, Neutral 1 and 2 demonstrators, respectively, before and after the conditioning. Note the significantly-increased angle and distance between the Reward and Aversive demonstrator subspaces (Two-way ANOVA with Tukey's multiple comparisons test). **k,** Changes of the modulation index of neural subspace representations for individual demonstrators before and after the conditioning, indicating a selective enhanced intensity in the Reward-demonstrator representation (Two-way ANOVA with Tukey's test). **l,** A schematic illustration of changes in neural subspace geometries for encoding distinct demonstrators based on angle and distance metrics (as shown panels **i**–*j*). **m,** VAF for distinct demonstrator subspaces before and after the conditioning, showing specific reduction of VAF for Reward and Aversive demonstrator, respectively ((Two-way ANOVA with Tukey's test, see detailed information in the Extended Table 1).

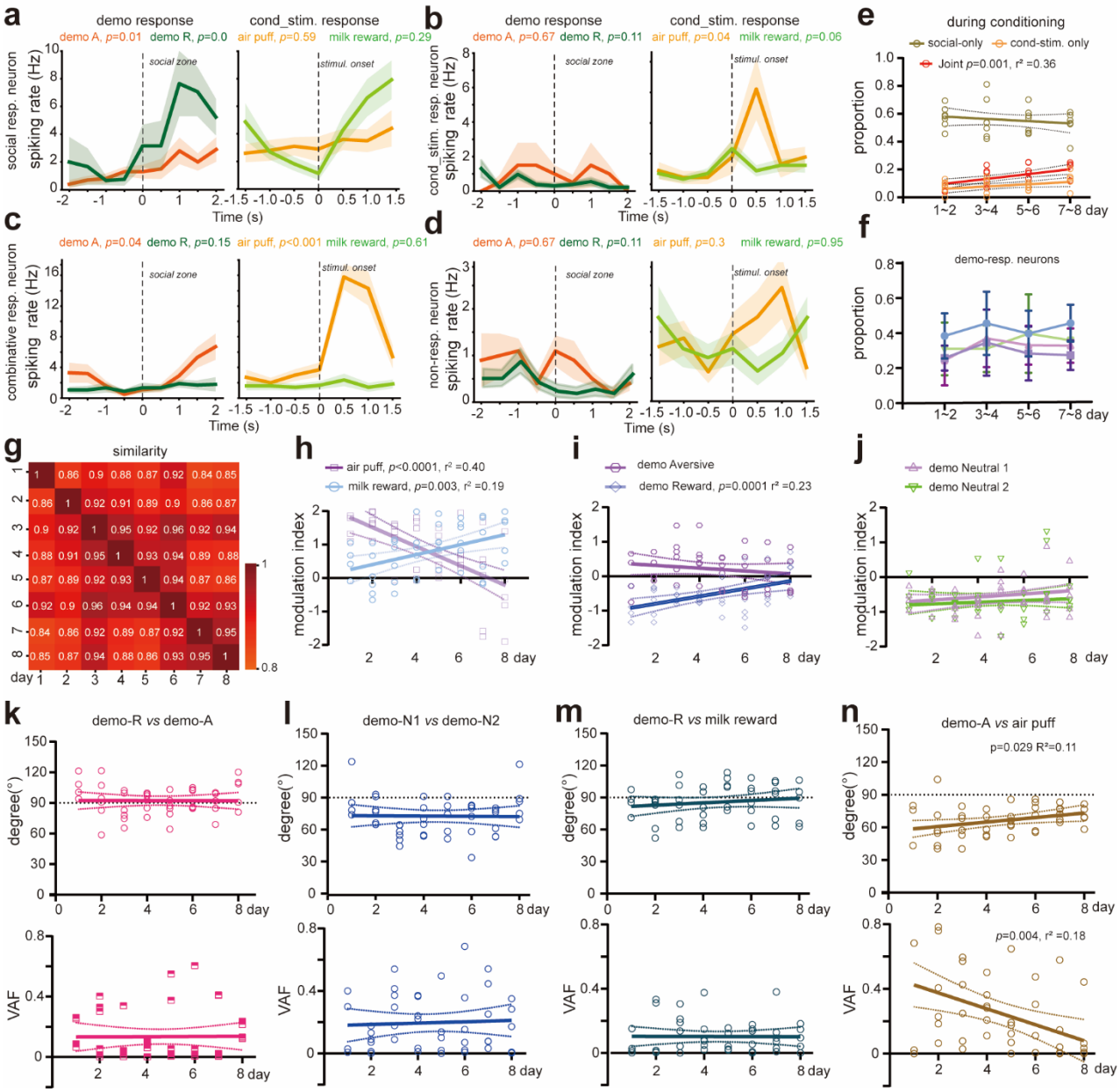


Fig. 6. Preferential shifts of neural representations toward the Reward demonstrator during social conditioning.

a–d, Representative social-response unit (**a**), conditioning stimuli-response unit (**b**), conjuctive (social or conditioning stimulus) response unit (**c**) and non-response unit (**d**), showing differential changes in spiking rates (binned with 500 ms) before and after the onset of entering the social zone (*left*) or receiving conditioning stimuli (*right*), respectively, in the social conditioning. **e,** Changes of the proportions of these three types of responsive units across the 8 conditioning days. Note a trend of increasing the proportion of conjuctive response units with the conditioning days (simple linear regression). **f,** Similar as **e** except proportions of demonstrator-specific encoding units were analysed. **g,** Cosine similarity of population activity representations in the joint SVD among the 8 days of social conditioning. **h–j,** Opposite changes of the modulation index of neural subspace

representations for reward and aversive condition stimulus (**h**), or Reward and Aversive demonstrators (**i**), while there is no significant change for Neutral 1 and 2 (**j**) during the 8 days of social conditioning. **k–n,** Comparisons of dynamic change of the angles (*top*) and VAF (*bottom*) between paired individual neural subspaces for distinct demonstrators or conditioning stimulus. Note that subspaces for the Aversive demonstrator and the air-puff stimuli become more separated by increasing the angle and reducing VAF simultaneously during the 8 days of social conditioning

Method

Animals

Adult wild type (WT) C57BL/6J mice (at postnatal 3-6 months) were used in the most of animal behavioral tests and *in vivo* recording experiments. In several sets of experiments, two additional transgenic lines, the *Trpc2* knock-out mice (021208; RRID:IMSR_JAX:021208) and the *cFos-CreER* (Jax. No. 021882) were used as the test mice as well. The *cFos-CreER* line was crossed with the tdTomato reporter line *Ai9* (Jax No.007905) to fluorescently identify experience-related, cFos-expressing neuron (presumably active neurons)[1]. Only male mice were used in this study.

In the behavioral tasks, one WT or transgenic mouse was used as the test mouse, while other four mice, at matching ages, were selected as the social stimulus or demonstrator mice in the multi-individual social test from other different cages. During the conditioning phase, the test mouse was regularly housed with its littermates, and after being implanted with a recording chamber, it was then housed alone in order to protect the implanted electrode. All mice were raised on a revered 12 h light (18:00–6:00)/dark (6:00-18:00) cycle, under an environmental condition with 23~25 °C ambient temperature and 40-60 % relative humidity. The protocols of animal surgery and experiments were approved by the Animal Care and Use Committee (IACUC) of State Key Laboratory of Cognitive Neuroscience & Learning at Beijing Normal University (reference NO.: IACUC-BNUNKLCNL-2021-17).

Animal Surgeries, Electrode-array Implantation and Viral Injection

Mice were anesthetized using isoflurane (4% for induction, 0.5–1.5% for maintenance) and tightly positioned on a stereotaxic apparatus (RMD, China). A custom-built micro-drive equipped with a 32-channel micro-wire array (KD-MWA, Kedou Brain-Computer Technology, China) was implanted in the medial prefrontal cortex (mPFC) at one hemisphere. Stereotaxic coordinates (relative to the Bregma) were: +1.9 mm AP, + 0.25 mm ML, and –0.8 to –1.0 mm DV from the cortical surface, to make sure all wire in PL or IL area. Then the micro-drive assembly was affixed to the skull using three small stainless-steel screws (M1, 1mm), cyanoacrylate adhesive and dental acrylic. After the implantation, animals were individually housed to recover on a heating pad until fully ambulatory.

For viral injection, the *cFos-CreER* mice (at postnatal 8-9 weeks) were fixed on a stereotaxic device under the same anesthesia condition. The skull was exposed under antiseptic conditions and a small craniotomy was made with a thin drill over the prefrontal cortex. (typical coordinates: 1.9 mm anterior to Bregma; ± 0.25mm lateral to the midline). An electronic injector (RWD, China) attached to the stereotaxic devise was used to infuse ~300 nL viral solution (*pAAV-EF1α-DIO-hM4D(Gi)-mcherry-WPRE*, 1.12 x 10^{13} particles per ml, Obio Technology, Shanghai) into the target cortical

region through a borosilicate glass micropipette (tip size, 10 μm), at a delivery rate of 20 nL/min. At the end of injection, the micropipette was held in the place for about 10 min before retraction. The mice were then placed in a stable-temperature incubator to recover from anesthesia, and after a full recovery they were returned to their home cages.

Multi-individual Social Test

Apparatus of Multi-individual Social Test

A custom-built wooden square arena, 50 cm (length) × 50 cm (width) × 50 cm (height), was used for the behavioral social tasks, in which a test mouse made social interaction and recognition with other four demonstrator counterparts (Figure 1A). The demonstrator mice were housed in a wire pencil cup (8 cm diameter and 10 cm heigh) and positioned at four corners individually in the arena. A semi-circle area with 3.4 cm away from the cup's circle edge was defined as the social investigation zone at each corner[2], and its outer edge was installed with six infrared photoelectric sensor to detect the time of entering and leaving the zone precisely. The arena was lit unevenly during the test, and animal's behaviors in the arena were recorded by a digital camera (IRAYPLE A7300MU90; 60 fps, 2048×1536 pixels) mounted 80 cm above the center of the arena. In some experiments, two additional cameras were used to record from the arena sides. Recorded videos were analyzed by the DeepLabCut software[3] (version 2.3.9, https://github.com/DeepLabCut/DeepLabCut?tab=readme-ov-file), in which the test mouse's positions were tracked and its defined social investigation actions (e.g. sniffing and touching the demonstrator mice in the social zone) as well as other non-social behaviors (locomotion and stationary manners) were accordingly labelled and categorized for further analyzed.

Both test and demonstrator mice were separately habituated to the arena for 2–3 days prior to experimentation. Our social behavioral paradigm comprised three phases: the social preference rank test (pre-conditioning phase), the social conditioning, and the post-conditioning social preference rank test. In above phases, all controls to video-recording, social investigation detection, conditioning stimuli delivery and time alignments of electrophysiology recording were conducted by an Arduino-based finite-state machine (BPod r0.5, Sanworks) and our own custom MATLAB scripts (MathWorks, version R2023b).

Social Preference Rank Test (pre-conditioning phase)

After 2- or 3-day habituation to the testing arena, the test mice encountered four housed demonstrator mice at individual corners for the first time. The test mouse was allowed to freely explore the arena and make social investigations to individual demonstrator mice for 10 min. Its time spent within individual demonstrator's social zones were calculated. To minimalize potential biases to specific

corners or demonstrator in the test mouse, we set two 10-min test sessions, in which the corners of housing two demonstrators were swapped each other (as shown by the schematic diagram in the Figure 1A) in each day. For each test mice, its proportions of social investigation time spent on four individuals were averaged from the two sessions across 4 days to obtain the social preference rank.

Social Conditioning Protocol

Based on the initial social preference rank order for a test mouse, we then set the milk reward and aversive air puff stimuli paired with social interaction when the test mice were socially investigating these two demonstrators that resized the first and the last positions, respectively. Thus, the latter two demonstrators were designating as the *Aversive* demonstrator and *Reward* demonstrator, respectively. Meanwhile, social investigation of the test mice to the rest two demonstrators yielded no any outcome, and similarly these two demonstrators were designating as the *Neutral 1* and *Neutral 2*.

Each test mouse was conditioned during its social interactions with *Reward, Aversive, Neutral 1* or *Neutral 2* demonstrators separately in the same arena 8 days (Figure 4A). Details of social conditioning stimuli were as follows:

When the test mouse made a social investigation to individual housed demonstrator ≥ 3 seconds in the social zone, a specific outcome was triggered based on the designated demonstrator identity:

- ***Reward* demonstrator**: 2 drops of sweet milk delivered outside the social zone to promote continued social exploration (Extended Data video 5).
- ***Aversive* demonstrator**: An air puff was administered to discourage interaction (Extended Data video 6).
- ***Neutral* demonstrators**: No stimulus was delivered.

A daily conditioning session comprised ≥ 30 trials (intervalled by 30 sec) or lasted up to 15 min for individual demonstrators, respectively. Corner positions and conditioning trial sequences for different demonstrators were daily randomized to avoid possible preference to specific corner or conditioning order in the test mice (Figure 4A).

Post-conditioning Multi-individual Social Test

After conditioning, mice underwent two consecutive social tests wherein all demonstrator positions were changed between sessions to control for spatial and novelty effects. Changes in social preference were evaluated using Two-way ANOVA (Extended Data Table 1).

***in vivo* Electrophysiology in Behaving Mice**

A 4×8 array of KD-MWA micro wires (Nitinol, 25μm, impedance 100kΩ-300kΩ) was implanted

into the mPFC of the test mouse. A wire was inserted above V1/2 area as a common reference. The microwire array was connected to a 32-channel head-stage, and electric signals were recorded by the CerePlex-Direct data acquisition system (Blackrock Microsystems, USA). Neuronal spiking signals from extracellular recording were digitized at 40 kHz, bandpass-filtered (250 Hz - 8 kHz) and acquired by a computer for further analysis. The *in vivo* recording was made while mice freely explored the environments and housed demonstrator mice in arena for 10 minutes. A commutator was used to reduce the cable torques on the headstage and minimalize any possible restraint on mouse's free movements. Time of recording neural signals was synchronized with that of video-recording behaviors.

We used the KiloSort 2.5 (https://github.com/MouseLand/Kilosort) to sort raw spike waveforms into single units (presumptive neurons)[4], using the following parameters: Nblock = 0 (optimized for micro-wire arrays), detection threshold = 9, lambda = 10, and AUC threshold = 0.7. Putative units identified by the KiloSort as "good" or "mua" (multi-unit activity) were subsequently manually curated in the Phy 2.0 (https://github.com/kwikteam/phy), based on waveform. Units with firing rates <0.3 Hz were omitted. Final unit inclusion was based on the following quality metrics, computed via the SpikeInterface (https://github.com/SpikeInterface/spikeinterface): inter-spike interval violation ratio < 0.5, presence ratio > 0.95 per session, amplitude cutoff < 0.1, and signal-to-noise ratio (SNR) > 2.

Linear regression model

To identify task variables modulating neural activity during pre- and post-conditioning social tests, we employed a regularized linear regression model. Eight one-hot encoded task variables $S_{demo,loc}$ were defined, corresponding to each combination of demonstrator identity—Neutral 1, Neutral 2, Aversive, or Reward—and its spatial location across the two test sessions (4 demonstrators×2 sessions = 8 combinative variable).

The firing rate $y_i$of the ***i***th neuron was modeled as a linear combination of these task variables:

$$y_i = \sum_{demo=1}^{4} \sum_{loc=1}^{2} \beta_i(demo, loc) S_{demo,loc} + \varepsilon_i$$

Where $\beta_i(demo, loc)$ denotes the regression coefficient for each demonstrator-location predictor, and ε_i represents the residual error. To mitigate overfitting and perform feature selection, we applied L_1 (Lasso) regularization. The regularization strength was optimized via 10-fold cross-validation[5].

To identify neurons significantly encoding a specific predictor, we compared full and reduced models

using a cross-validated variance-explained criterion. For each neuron, we iteratively refit the model on 90% of the data (10-fold cross-validation), each time omitting one predictor of interest. A significant increase in the mean squared error (MSE) of the residuals upon predictor removal indicated that the neuron encoded that variable.

Neuronal classification was based on spiking response characteristics during the pre- or post-conditioning social test:

- **Social-response neurons**: significantly responsive to the same demonstrator across different locations.
- **Corner-response neurons**: significantly responsive to the same location across different demonstrators.
- **Combinative neurons**: significantly responsive to a specific demonstrator-location combination.
- **Conjunctive neurons**: met criteria for both social- and corner-response units.

All social-responsive, conjunctive, and combinative neurons were classified as **social-informative neurons**; the remainder were categorized as **non-social informative neurons**. During the conditioning phase, where all demonstrators occupied the same location, spatial confounds were controlled, and encoding was assessed solely based on the significance of residual MSE increase.

To assess how task variables influenced neural activity during the conditioning phase, we employed a multivariable linear regression model similar to that used in the social tests, but with an expanded set of predictors. Task variables included demonstrator identity (*Neutral* 1, *Neutral* 2, *Aversive*, or *Reward*) and conditioning type (Milk reward or air puff). Since reward and aversive stimuli occurred as discrete events but could elicit sustained neural responses, we utilized Elastic Net regularization—a hybrid approach combining L_1 (Lasso) and L_2 (Ridge) penalties. This method simultaneously performs feature selection (via Lasso) and models temporal persistence in neural activity (via Ridge, Extended data fig. 8a)[6].

Neuronal classification was based on spiking response characteristics during the conditioning phase:

- **Social-response neurons**: significantly responsive to the social demonstrator.
- **Condition-response neurons**: significantly responsive to Milk reward or Air puff.
- **Combinative neurons**: significantly responsive to Milk reward with demonstrator *Reward*, or to Air puff with demonstrator *Aversive*.

Population decoding

To determine whether population activity patterns in the mPFC encode discriminable social information and to characterize the linearity of such representations, we performed population decoding using the support vector machines (SVMs). We trained the multi-class SVMs with both linear and radial basis function (RBF, non-linear) kernels to classify neural population activity into one of the four demonstrator identities.

For each recording session, the SVMs were trained and evaluated independently at each social interaction time point using 5-fold cross-validation. Model performance was quantified both by decoding accuracy (proportion of correct classifications) and the F_1 score. The F_1 score, the harmonic mean of precision and recall, provided a robust measure of classifier performance, especially under class imbalance, and offered a more comprehensive evaluation of the model's ability to discriminate neural activity patterns associated with different social stimuli.

The F1 score was calculated as:

$$F_1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

where **Precision**: The proportion of predicted positive samples that are actually positive.

- Precision = TP / (TP + FP)
- TP (True Positive): Correctly predicted positive samples
- FP (False Positive): Incorrectly predicted positive samples

- **Recall**: The proportion of actual positive samples that are correctly identified as positive.
  - Recall = TP / (TP + FN)
  - FN (False Negative): Incorrectly predicted negative samples

To assess the statistical significance of the decoder performance, we performed a non-parametric cluster-based permutation test (1000 times).

CEBRA for Cross-Session Neural Embedding

To uncover latent neural dynamics and enable alignment of population activity across multiple days, we applied the CEBRA (Consistent EmBeddings of High-dimensional Recordings using Auxiliary variables; https://cebra.ai), a machine learning method designed for nonlinear dimensionality reduction of time-series data[7]. CEBRA learns a mapping from high-dimensional neural activity into a low-dimensional embedding space that preserves temporal, behavioral, or neural correlation structure.[7]

Neuronal spike data were binned into 250-ms intervals and combined across sessions. We used CEBRA in discrete, multi-session mode with the following configuration: temperature_mode = "auto", learning_rate = 3e−4, max_iterations = 10000, time_offsets = 10, output_dimension = 3, distance = 'cosine', and conditional = 'time_delta'. This set of parameters was chosen to prioritize temporal consistency across sessions while capturing discriminative features related to social stimuli.

The resulted embeddings were then used to quantify representational stability and similarity across days. We computed session-pairwise similarity using the adaptive Hausdorff distance[8], which provides a robust metric for comparing the geometry between two sets of neural trajectories in the embedded space. This approach allowed us to assess whether social representations occupied consistent subspaces across days and conditioning phases.

Joint SVD Projection Method for Cross-Dataset Neural Response Analysis

To enable consistent comparison of neural representations across day, we developed a unified dimensionality reduction framework based on joint Singular Value Decomposition (SVD). This method identifies shared neural response patterns while preserving dataset-specific variance structures through subspace orthogonalization[9].

Let $\{D_1 \in \mathrm{R}^{\mathrm{N}_1 \times S}, D_2 \in \mathrm{R}^{\mathrm{N}_2 \times \mathrm{S}}, \ldots, D_{\mathrm{k}} \in \mathrm{R}^{\mathrm{N}_{\mathrm{k}} \times \mathrm{S}}\}$ denote ***k*** neural response datasets, where $N_i$represents the number of neurons in dataset ***i*** and ***S*** denotes the number of stimulus conditions (i.e., combinations of demonstrator and location, the total amount = 8). Each entry in D_{i} corresponds to a centered regression coefficient $\beta_i(demo, loc)$ for a given neuron and stimulus.

1. ***Global Matrix Construction***

All datasets are vertically concatenated into a combined matrix:

$$\beta_{\text{combined}} = [D_1^\top D_2^\top \cdots D_k^\top]^\top \in R^{N \times S} \quad \text{where} \quad N = \sum_{i=1}^{k} N_i$$

2. ***Joint SVD Decomposition***

The centered matrix is factorized via singular value decomposition:

$$\beta_{\text{combined}} = U \Sigma V^\top$$

Where the matrix ***U*** represents coordinated patterns of neural activity within the population, while ***V*** encodes linear combinations of the eight stimulus conditions that are associated with specific neural response profiles.

3. ***Dataset-Specific Projection***

For each dataset D_i:

(1) Extract the top-$\boldsymbol{n}$ ($\boldsymbol{n}$ is neuron numbers of corresponding dataset) left singular vectors: $\boldsymbol{U_i} \in R^{N_i \times n}$

(2) Perform QR decomposition to orthonormalize the basis:

$$\boldsymbol{Q}_i \boldsymbol{R_i} = \mathbf{qr}(\boldsymbol{U_i}),$$

with sign alignment via correlation matching to ensure consistency.

(3) Project the centered data into the local subspace:

$$P_i = D_i{}^T Q_i \in R^{S \times n}$$

This framework facilitates the comparison of neural population dynamics across experiments, sessions, or individuals, while respecting the intrinsic variance structure of each dataset. It is particularly useful for identifying stable neural subspaces and tracking their reorganization over learning or time.

During conditioning phase, the framework is similar but the $\boldsymbol{S} = \mathbf{6}$ (4 social demonstrator variable + 2 conditioning stimulus).

Computation of Subspaces for Social Identities

To identify low-dimensional subspaces capturing the majority of social and location-related variance before and after conditioning, we utilized the regression coefficients, $\beta_i(demo, loc)$, obtained from joint SVD projection. Each social-location combination was represented at the population level by an 8-dimensional vector

$$\beta_i(demo, loc) = [\beta_1(demo, loc), \ \cdots, \ \beta_8(demo, loc)]^\top.$$

A demonstrator-specific subspace was defined as the set of vectors corresponding to the same demonstrator across different locations (ei. $\beta_i(demo\ Reward, loc1)$ and $\beta_i(demo\ Reward, loc2)$ form the demonstrator-Reward subspace). For each such group $\{\beta(demo, loc)\}_{loc=1,2}$, we performed principal component analysis (PCA) to identify the first two orthogonal axes capturing the greatest response variance. This allowed the approximation:

$$\beta(demo, loc) \approx V_{demo} \boldsymbol{\kappa}(demo, loc)$$

where $V_{\text{demo}} = [v^1_{\text{demo}}, v^2_{\text{demo}}] \in R^{8\times 2}$ forms an orthogonal basis for the demonstrator-specific subspace, with $[|v^1_{demo}|_2 = |v^2_{demo}|_2 = \sqrt{8}]$, and $\boldsymbol{\kappa}(demo, loc)$ denotes the projection coefficients of $\beta(demo, loc)$ onto this subspace. For subsequent analysis, we used a normalized version of the basis, denoted $V_{demo}/\sqrt{N}$, referred hereafter simply as $\hat{V}_{demo}$. To facilitate comparison across different social subspaces, we used orthonormal axes q^1_{demo} and q^2_{demo} rather than the original v^1_{demo} and v^2_{demo} to define each social subspace[5].

Principal Angles Between Subspaces

To quantify the relative orientation of different social demonstrator subspaces in the neural state space during both pre- and post-conditioning phases, we computed the principal (or canonical) angles between subspaces—a geometric measure of alignment between two linear manifolds.

For any two demonstrator subspaces a*a* and b*b*, with corresponding orthonormal bases $\hat{V}_a$ and $\hat{V}_b$, we performed singular value decomposition (SVD) on their inner product matrix:

$$\hat{V}_a^T \hat{V}_b = P_a C P_b^T,$$

Where both P_a and P_b are 2×2 orthogonal matrices and C is a 2×2 diagonal matrix whose elements are the demo cosines of the principal angles θ_1 and θ_2:

$$C = \text{diag}(\cos(\theta_1), \cos(\theta_2))$$

The principal angles $\theta_1, \theta_2 \in [0, \pi/2]$ provide a measure of subspace similarity, with smaller angles indicating greater alignment[5].

During the conditioning phase, in which all demonstrators were presented at the same location within a single day, a simplified similarity measure was sufficient. Thus, we computed the cosine similarity and angle directly between subspace basis vectors without full principal angle analysis.

Variance Accounted For (VAF) Ratio

To further quantify the alignment between different social demonstrator subspaces, we introduced a variance-accounted-for (VAF) ratio. For a given demonstrator and location, let

$$\boldsymbol{g}_{demo}(loc) = V_{demo}\boldsymbol{\kappa}(demo, loc)$$

The VAF ratio between two demonstrator subspaces $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as:

$$\mathrm{VAF_{ab}} = \frac{\mathrm{Var}\left(\widehat{\mathrm{V}}_{\mathrm{b}}\widehat{\mathrm{V}}_{\mathrm{b}}^{\mathrm{T}}\mathrm{G_a}\right)}{\mathrm{Var}(\mathrm{G_a})}$$

where

$$G_a = [g_a(location\ test\ 1), g_a(location\ test\ 2)]$$

contains the reconstructed activity patterns for demonstrator a*a* across both test locations. Note that $\mathrm{VAF_{ab}}$ is asymmetric; it measures how well subspace ***b*** accounts for variance in the neural representation of demonstrator a*a*. The ratio ranges from 0 (orthogonal subspaces) to 1 (complete overlap)[5].

To compute the Variance Accounted For (VAF) between neural representations during the conditioning phase, we applied the following procedure:

1. ***Data Centering***:
Each dataset containing coefficients β_i, obtained from L1+L2 linear regression about demonstrator variable and conditioning variable, was centered by subtracting its mean:

$$a_{\mathrm{centered}} = a - \mu_a,\ b_{\mathrm{centered}} = b - \mu_b$$

where μ_a and μ_b denote the means of datasets ***a*** and ***b***, respectively.

2. ***Covariance Calculation***:
The covariance between the two centered datasets was computed as:

$$\mathrm{Cov}(a, b) = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(a_{\mathrm{centered,i}} \times b_{\mathrm{centered,i}}\right)$$

where ***N***=6, corresponding to the four demonstrators and two conditioning stimuli.

3. ***Variance Estimation***:
The variances of the centered datasets were estimated using Bessel's correction (ddof = 1) to account for sample bias:

$$\mathrm{Var(A)} = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{A_{centered,i}} - \mu_{\mathrm{A_{centered}}}\right)^2, \mathrm{Var(B)} = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{B_{centered,i}} - \mu_{\mathrm{B_{centered}}}\right)^2$$

These quantities were subsequently used to compute the VAF ratio, quantifying the shared variance between neural activity patterns under different stimulus conditions.

4. ***VAF computation***:

The Variance Accounted For (VAF) was then calculated as:

$$\mathrm{VAF} = \frac{[\mathrm{Cov}(A,B)]^2}{\mathrm{Var}(A)\cdot\mathrm{Var}(B)}$$

This metric quantifies the proportion of variance in one dataset that is explained by the other, yielding values between 0 (indicating no shared variance) and 1 (representing complete shared variance).

Gain-Modulation Approximation of Collective Variables

To assess whether a simplified gain-modulation mechanism could account for neural population responses, we approximated the full linear model at the level of collective variables using a gain-field modulation framework. Specifically, we modeled the latent variable $\boldsymbol{\kappa}(demo, loc)$ as:

$$\kappa(demo, loc) = \lambda_{demo} O_{demo} f_{loc} + \zeta(demo, loc)$$

where λ_{demo} is a demonstrator-specific gain factor, f_{loc} (of length 2) is a spatial tuning vector independent of demonstrator identity, O_{demo} (of size 2 ×2) is an orthogonal matrix aligning demonstrator and spatial features, and $\boldsymbol{\zeta}(demo, loc)$ denotes the approximation error.

Following model fitting, gain factors were rescaled by the mean $\lambda\lambda$ value across all conditions, and spatial vectors were adjusted accordingly to preserve relative modulation strength.

During the conditioning phase, where spatial variation was absent, the modulation index was defined as the Euclidean (L_2) norm of the demonstrator-specific gain vector.

All modulation indices were z-scored across sessions prior to cross-day comparison to account for baseline shifts in response magnitude.

Data availability

The preprocessed data of this paper (https://github.com/Tianyulied/Neural-geometric-representations-of-social-memory-for-multi-individuals-in-medial-prefrontal-cortex) is available on GitHub. The raw data (electrophysiological recordings and video files) are not publicly available due to their substantial size. They are available from the corresponding author upon reasonable request.

Code availability

Code for this paper (https://github.com/Tianyulied/Neural-geometric-representations-of-social-memory-for-multi-individuals-in-medial-prefrontal-cortex) is available on GitHub.

1 Guenthner, Casey J., Miyamichi, K., Yang, Helen H., Heller, H. C. & Luo, L. Permanent Genetic Access to Transiently Active Neurons via TRAP: Targeted Recombination in Active Populations. *Neuron* **78**, 773-784, doi:10.1016/j.neuron.2013.03.025 (2013).

2 Rein, B., Ma, K. & Yan, Z. A standardized social preference protocol for measuring social deficits in mouse models of autism. *Nat Protoc* **15**, 3464-3477, doi:10.1038/s41596-020-0382-9 (2020).

3 Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* **14**, 2152-2176, doi:10.1038/s41596-019-0176-0 (2019).

4 Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, doi:10.1126/science.abf4588 (2021).

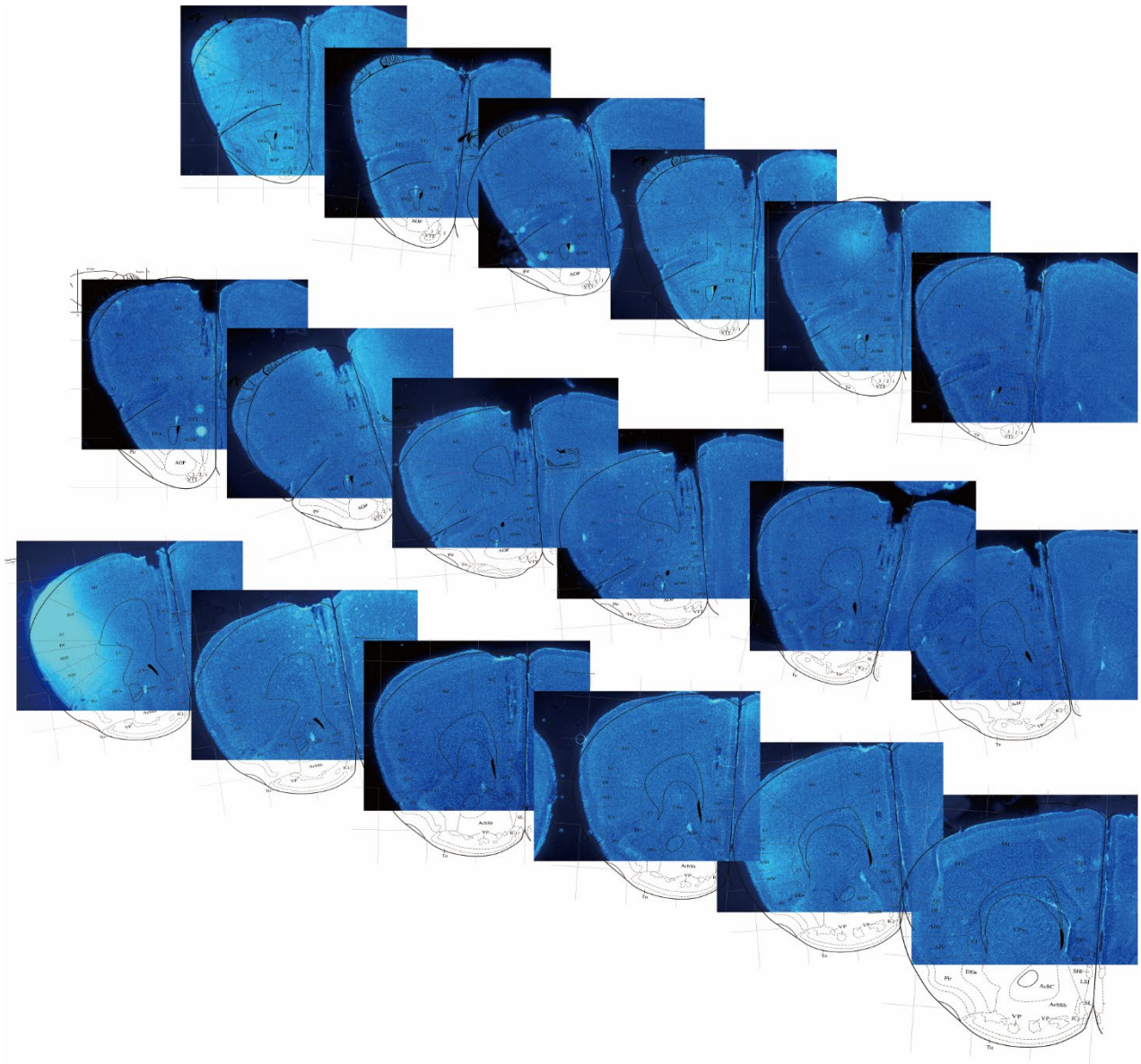
5 Yang Xie, P. H., Junru Li, Jingwen Chen, Weibin Song, Xiao-Jing Wang, Tianming Yang, Stanislas Dehaene, Shiming Tang, Bin Min, Liping Wang. Geometry of sequence working memory in macaque prefrontal cortex. *Science*, doi:10.1126/science.abm0204 (2022).

6 Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience* **22**, 1677-1686, doi:10.1038/s41593-019-0502-4 (2019).

7 Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360-368, doi:10.1038/s41586-023-06031-6 (2023).

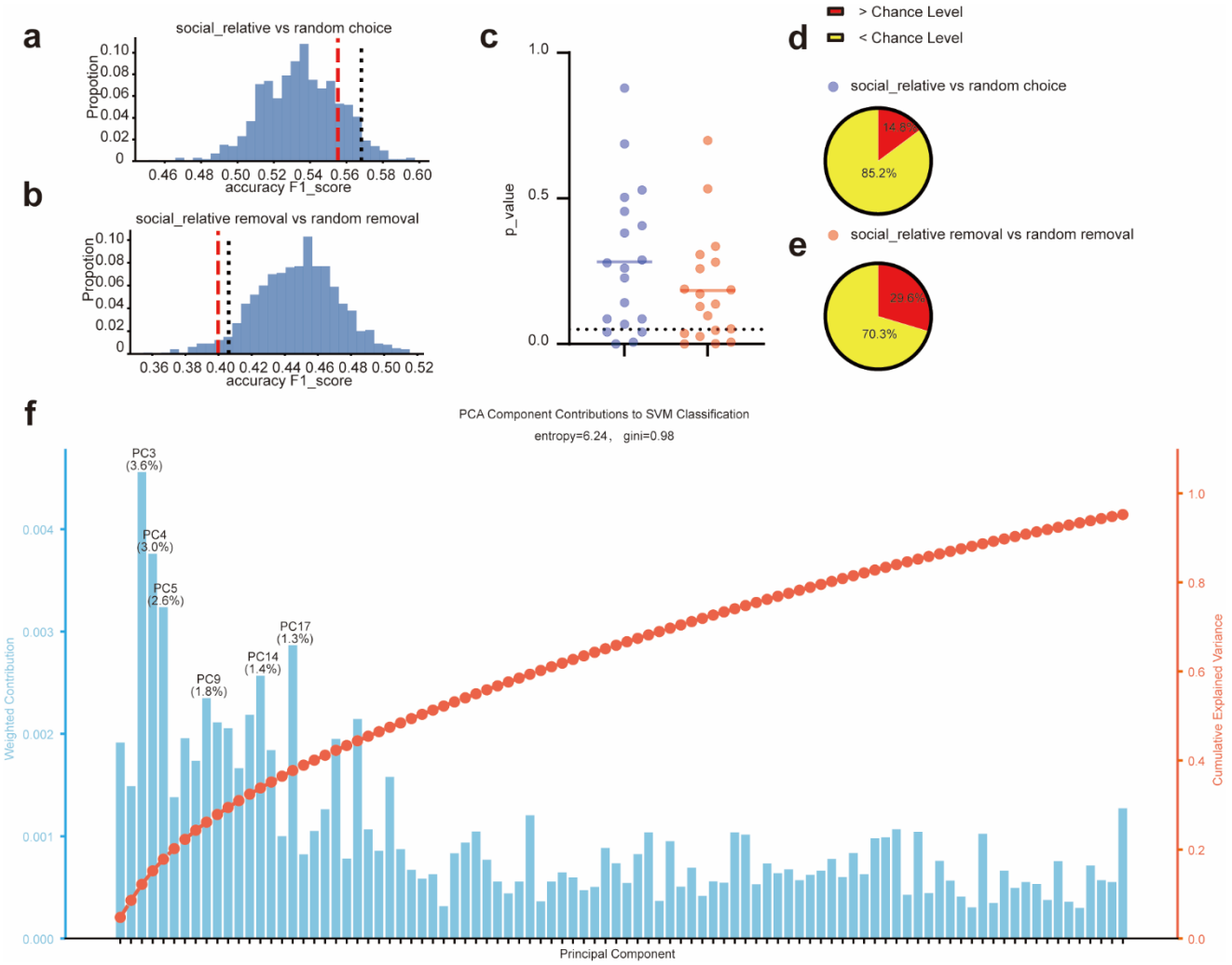
8 Lange, K. Computation of the Hausdorff Distance between Two Compact Convex Sets. *Algorithms* **16**, doi:10.3390/a16100471 (2023).

9 Ehret, B. *et al.* Population-level coding of avoidance learning in medial prefrontal cortex. *Nature Neuroscience* **27**, 1805-1815, doi:10.1038/s41593-024-01704-5 (2024).



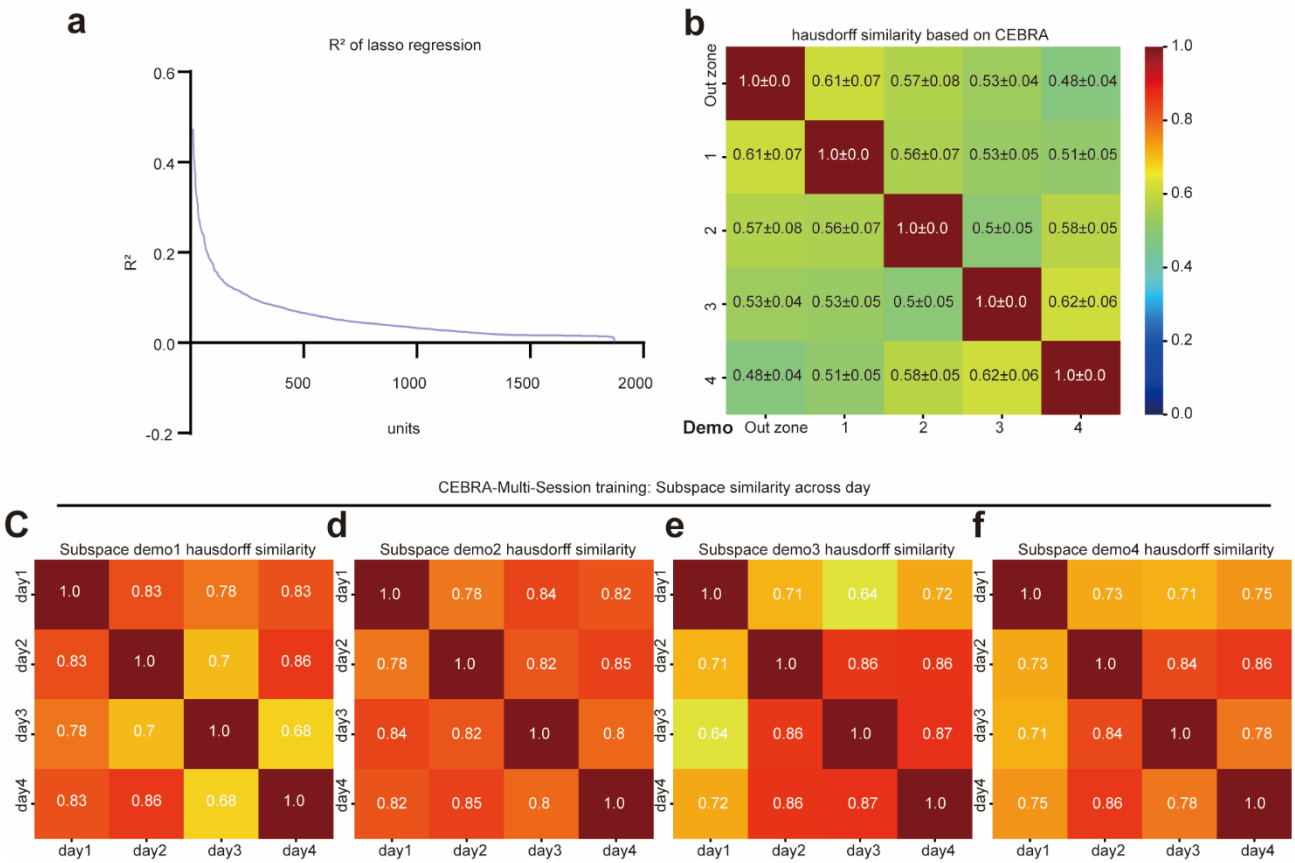
Extend Data Fig. 1 Histological verification of microelectrode tracks in mouse brain sections.

Serial sections of mouse brain (spaced at 80 μm intervals). Image obtained from the Allen Mouse Brain Atlas (https://mouse.brain-map.org).



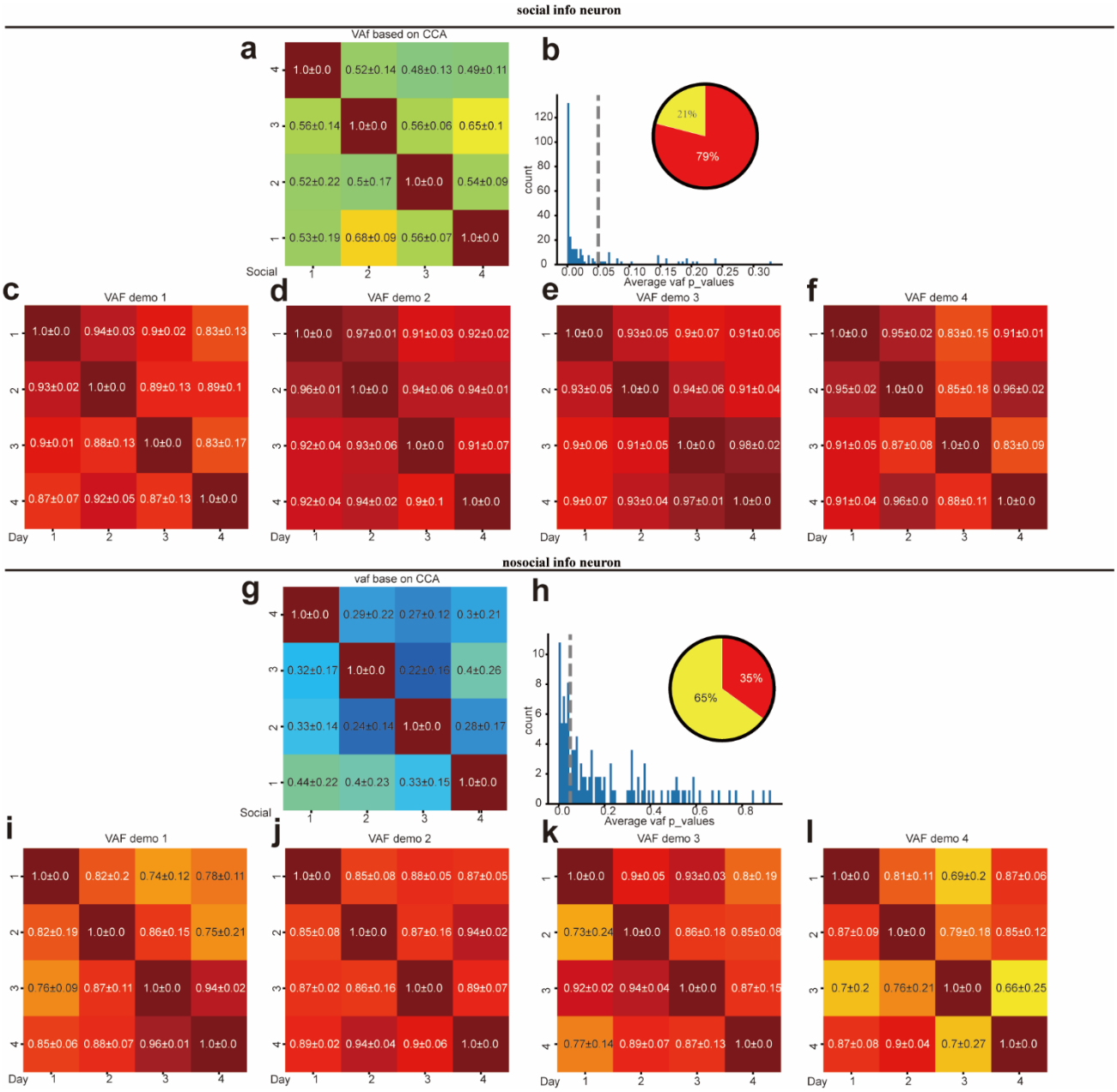
Extend Data fig. 2 Most cells carrying social memory information that is distributed in uncorrelated modes of activity.

a-b, Performance of control classifiers. F1 scores of support vector machine (SVM) classifiers trained on randomly selected (a) or randomly omitted (b) non-social neurons (with neuron numbers matched to social-related neurons; 1,000 iterations). The black dashed line indicates the 95% chance level threshold. The red line shows the true F1 score of social-related neurons. **c,** Statistical significance of control tests. Distribution of p-values from the random selection test and the random omission test. **d-e,** Proportion of above-chance classifications. The ratio of iterations yielding above-chance performance in the random selection (**d**) and random omission (**e**) tests. **f,** Component analysis of population activity. Explained variance ratio and component weights after principal component analysis (PCA). The distribution of weights across many components, as indicated by Gini and entropy metrics, demonstrates that social discrimination is supported by a complex recognition process involving multiple, uncorrelated activity modes. Detailed statistics are provided in Extended Data Table 1.



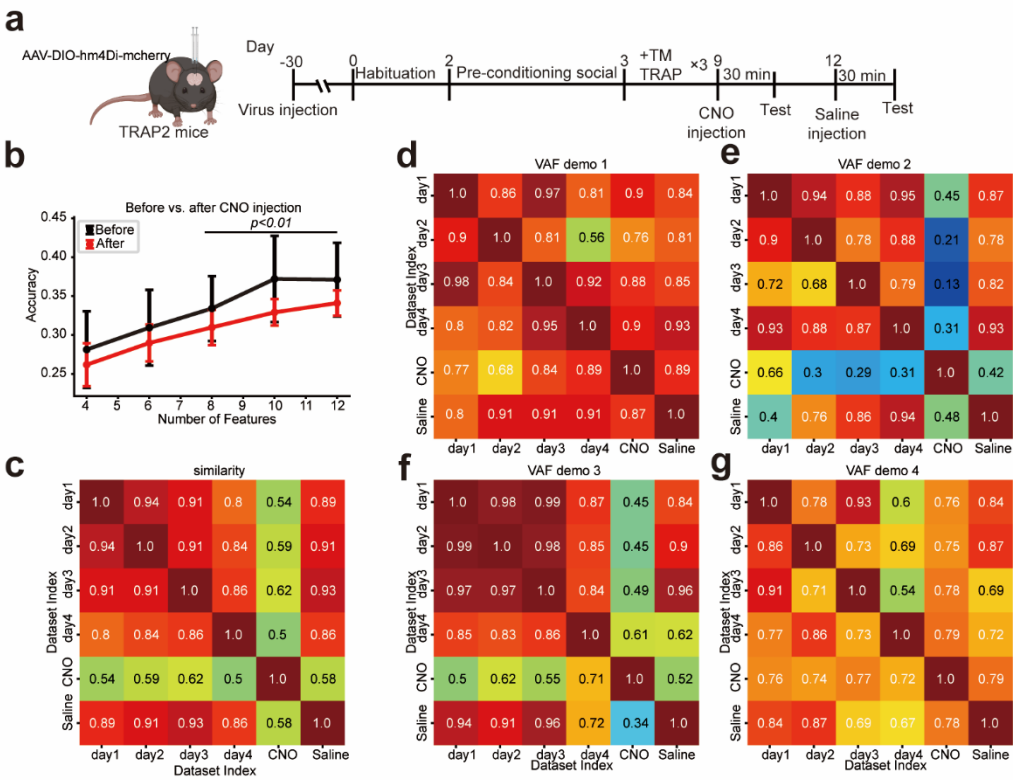
Extend Data fig. 3 Stability of neural representations across days assessed via nonlinear dimensionality reduction and similarity analysis.

a, The ratio of variance in social discrimination explained by neural activity features selected via Lasso regression. **b,** Similarity of neural representations across days, quantified using the Hausdorff similarity metric between neural embeddings generated by CEBRA (multi-session mode) for each pair of days. **c-f,** Consistency of individual social subspaces. Similarity matrices (Hausdorff similarity) for each of the four identified social subspaces across the 4 recording days. Detailed statistics are provided in Extended Data Table 1.



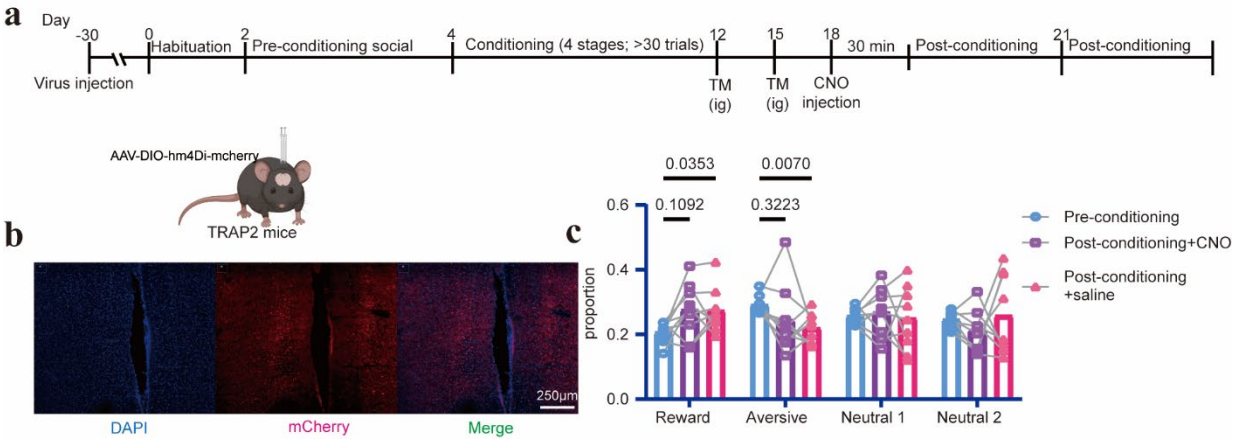
Extend data fig.4 Social-related neurons exhibit higher cross-day representational stability than non-social neurons.

a, VAF values were calculated between neural representations on different days using the dataset containing only social-related neurons. **b,** Statistical significance of cross-day stability. Distribution of p-values from the comparison between the true VAF and VAF from shuffled data (1,000 iterations). Insets show the proportion of shuffled data points above (red) and below (yellow) the chance level. The dashed line indicates the significance threshold ($\alpha = 0.05$). **c-f,** Stability of individual social subspaces. VAF values for each of the four social subspaces across the 4-day period (using social-related neurons). **g-l,** Parallel analyses were consistent with **a–f**, but using the dataset containing only non-social neurons. Detailed statistics are provided in Extended Table 1.



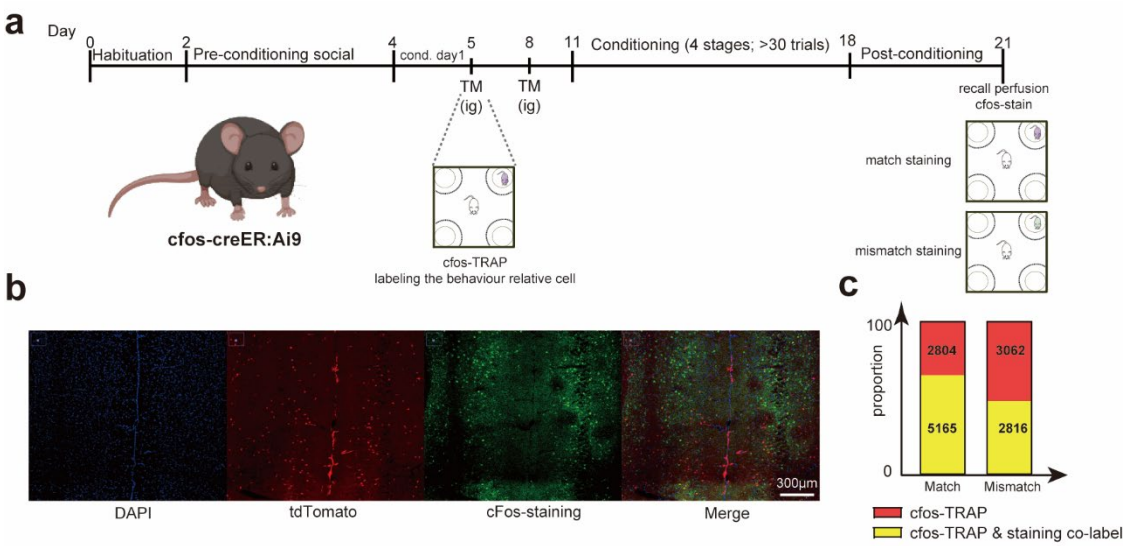
Extend data fig 5, Chemogenetic inhibition of behaviour-encoding neurons in mPFC disrupts population representations.

a, Schematic of viral injection, cFos-TRAP labelling, and chemogenetic inhibition (via CNO) of neurons activated during the behavioural paradigm. **b,** Population decoding accuracy for social discrimination was significantly reduced after CNO injection to inhibit the TRAPed neuronal ensemble, compared to control conditions. **c,** Neural activity from the chemogenetic inhibition session failed to align within the same low-dimensional subspace as the embeddings from the untreated and control sessions, demonstrating a breakdown in representational structure. **d-g,** Variance accounted for (VAF) values for each of the four social subspaces across days following chemogenetic inhibition, showing disrupted stability. Detailed statistics are provided in Extended Table 1.



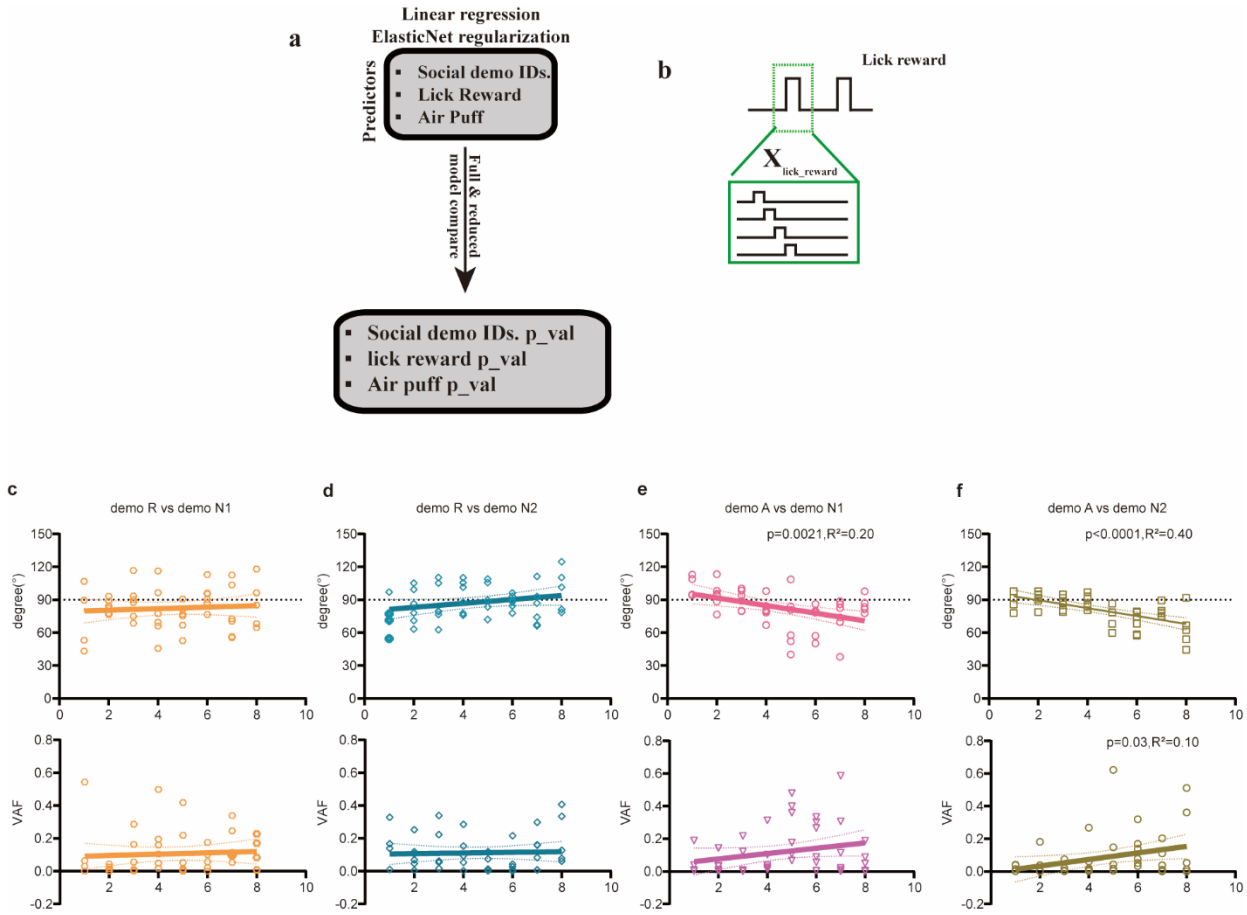
Extend data fig 6, Chemogenetic inhibition of behaviour-encoding neurons in the mPFC impairs social discrimination memory.

a, Schematic of the cFos-TRAP and chemogenetic inhibition approach used to target and manipulate neurons activated during the behavioural paradigm. **b,** Representative image showing TRAPed neurons (expressing mCherry, red) in the medial prefrontal cortex (mPFC). **c,** Changes in the proportion of time spent investigating four individual social demonstrators across three key phases: before conditioning (Pre-conditioning), after CNO injection prior to the social discrimination test (Post-conditioning + CNO), and after saline injection prior to the test (Post-conditioning + Saline). Inhibition of the TRAPed neuronal ensemble significantly disrupted memory-dependent social investigation. Detailed statistics are provided in Extended Table 1.



Extend data fig 7, Stability of the neuronal ensemble encoding a specific social stimulus across days.

a, Schematic of the cFos-TRAP and subsequent cFos immunostaining protocol to identify neurons co-activated during the initial encoding and during memory recall in the mPFC. **b,** Representative image of co-labelling. TRAPed neurons (red, mCherry) and recall-activated neurons (green, cFos+) in the mPFC. White arrows indicate examples of co-labelled neurons. **c,** Specificity of ensemble reactivation. The proportion of co-labelled cells (overlap between TRAPed and recall-activated neurons) was significantly higher when recall was triggered by the same demonstrator (Match group) compared to a different demonstrator (Mismatch control group) (χ^2 test, $p < 0.001$; N_match = 7969, N_mismatch = 5878). This demonstrates stable reactivation of the stimulus-specific ensemble. Detailed statistics are provided in Extended Data Table 1.



Extended Data fig.8 Dynamic reconfiguration of neural encoding toward the reward-paired demonstrator during conditioning.

a, Illustration of the linear regression framework with combined L1 (Lasso) and L2 (Ridge) regularization (Elastic Net) used to identify predictive neural features. **b,** The temporally shifted design matrix was used to compute regression weights (kernels) for discrete behavioural events, isolating the neural components associated with the air puff (aversive, $\beta_{\text{air puff}}$) and lick reward (milk reward, $\beta_{\text{lick reward}}$) **c–d,** Angles and variance accounted for (VAF) between selected subspaces.