The geometry of hippocampal CA2 representations enables abstract coding of social familiarity and identity

Lara Boyle[1*], **Lorenzo Posani**[2,3*], **Sarah Irfan**[4], **Steven A Siegelbaum**[1,3,5,6†], **and Stefano Fusi**[1,2,3,6†]

[1]Department of Neuroscience, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY 10027 USA; [2]Center for Theoretical Neuroscience, Columbia University, New York, NY 10027; [3]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027 USA; [4]Barnard College, New York, NY 10027; [5]Department of Pharmacology, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY 10032; [6]Kavli Institute for Brain Science, Columbia University, New York, NY 10027

Abstract

Social recognition memory encompasses two distinct processes: familiarity - the ability to rapidly distinguish a novel from familiar individual - and recollection, the recall of detailed episodic memories of prior encounters with familiar individuals (1). Although it is clear that the hippocampus is important for different forms of episodic memory (2), including spatial memory (3) and social recognition memory (4–7), whether and how neural activity in this single brain region may be able to encode both social familiarity and social recollection remains unclear (8–13). We addressed such questions using microendoscopic calcium imaging from pyramidal neurons in the dorsal CA2 region of the hippocampus (dCA2), an area crucial for social recognition memory (6, 14, 15) that encodes social and spatial information (16–18), as mice explored novel and familiar conspecifics. Here we demonstrate that the geometry of dCA2 representations in neural activity space enables social familiarity, social identity, and spatial information to be readily disentangled. Importantly, highly familiar littermates were encoded in higher-dimensional neural representations compared to novel individuals. As a result of this coding strategy, dCA2 neural activity was able to both provide an abstract, low-dimensional representation of social familiarity that could readily distinguish a novel from familiar individual and encode detailed episodic memories associated with familiar individuals.

Keywords: social memory; social recognition; recollection; familiarity; identity; novelty; geometry; abstraction; CA2; hippocampus; calcium imaging; one-photon; miniscope; microendoscope

[*] **Lara Boyle** and **Lorenzo Posani** contributed equally to this work

[†]Corresponding and co-senior authors: **Steven A. Siegelbaum**, sas8@cumc.columbia.edu; **Stefano Fusi**, sf2237@columbia.edu

The example of the "butcher on the bus" provides a classic illustration of the distinction between familiarity and recollection processes of social recognition memory (1). Encountering a known individual in a novel context may evoke an immediate sense of familiarity, which then requires conscious effort to recollect the details of episodic memories associated with that individual (19). To explore the neural mechanisms of familiarity and recollection, we injected a Cre-dependent virus into dCA2 of Amigo2-Cre mice to express the genetically encoding calcium indicator GCaMP6f selectively in dCA2 pyramidal neurons. Following lens implantation, we measured calcium events in a large number of dCA2 pyramidal neurons in awake, behaving animals (Fig. 1a-c, Suppl Movie 1) as they interacted with novel mice and familiar littermates.

We first imaged dCA2 activity as a mouse performed a test of social novelty recognition (Fig. 1d). After habituation to an oval arena containing two empty wire pencil cups at its opposite ends (left and right sides), one novel mouse and one familiar littermate were placed under each cup, and the subject mouse was allowed to explore the stimulus mice for five minutes (trial 1). To investigate the relation between social and spatial responses, we exchanged the positions of the stimulus mice and allowed the subject mouse to explore them for an additional 5 min (trial 2). As previously reported, subject mice showed a robust preference for the novel over the familiar individual in the initial presentation of the conspecifics in trial 1 (Fig. 1e-g). Pharmacogenetic silencing of dCA2 confirmed that dCA2 pyramidal neurons were necessary for social recognition in this arena (Suppl Fig. 1).

dCA2 encodes both social and spatial features

To determine whether the population activity of dCA2 neurons contained information about social or spatial features of experiences, we used a linear classifier to decode spatial and social variables from dCA2 firing activity as the subject mouse explored the novel individual and familiar littermate in the two trials (Fig 2b). A linear classifier is simple enough that it can be readily implemented by an individual neuron, implying that if the classifier can successfully extract information from dCA2 activity then so could, in principle, a single neuron. This neuron could be within dCA2, recurrently connected to the other dCA2 neurons (20), or it could be a neuron downstream from dCA2 (such as ventral CA1 (21)). We first trained the classifier to distinguish mouse interactions with the novel versus familiar animal by combining firing data around a given mouse in both trials. Next, we trained the classifier to distinguish whether the subject mouse was exploring around the left versus right cup, combining firing data around a given cup in both trials. To maximize decoding performance, we trained the classifier on the pseudo-simultaneous population of dCA2 neurons that we recorded (1096 neurons, n=11 mice).

The linear classifier was able to decode significantly above chance levels with which of the stimulus mice the subject was interacting and the spatial position where the interaction occurred (Fig. 2c). Additionally, the classifier could also decode whether the subject mouse was participating in trial 1 versus trial 2 (Fig. 2c). However, trial decoding performance was much lower than the decoding of position or social interactions (Fig. 2c). This distinction was further evidenced when we trained the linear classifier on the population of dCA2 neurons recorded in each individual mouse. The smaller populations of dCA2 neurons from each subject provided sufficient information to decode both social and spatial variables (Fig. 2d), whereas trial decoding accuracy was not greater than chance levels (Fig. 2d).

To what extent is the accuracy of dCA2 social decoding relevant to the behavioral performance of the mouse in distinguishing novel and familiar individuals? Although previous studies have found that dCA2 is crucial for social recognition memory and that dCA2 neuron firing responds to the presence of a conspecific (16–18) and contains information about social novelty (17, 22), it is not known whether the precision of social encoding in dCA2 is related to an animal's behavioral ability to discriminate a novel from a familiar individual. We therefore compared the behavioral performance of each individual mouse in discriminating a novel from familiar individual with the accuracy of dCA2 activity-based decoding of interaction partner in the first trial. This comparison revealed a strong correlation between behavioral and neural discrimination (Fig. 2e), suggesting the behavioral relevance of dCA2 decoding accuracy. In contrast, there was no correlation between behavioral preference and decoding accuracy when the subject mouse explored two stimulus mice with similar degrees of novelty, such as when the cups contained two novel mice or two familiar littermates (Suppl. Fig. 2b and c, respectively).

Does social and spatial decoding rely on separate subpopulations of highly selective dCA2 cells that respond mainly to social or spatial cues? We addressed this question by examining the weights assigned to each neuron by the linear classifier during decoding of social or spatial information. Most dCA2 neurons contributed to the decoding of both social and spatial information, and the fraction of selective cells that contributed to either social or spatial decoding was comparable to chance levels (see Suppl. Fig. 3; a cell exhibiting mixed selectivity across the trials can be observed in Suppl. Movie 2&3). The few highly selective cells did not contain essential social or spatial information as there was little decrease in decoder performance when these neurons (ranked by their normalized difference in social and spatial decoding weights) were omitted from the input to the classifier (Suppl. Fig. 4). In contrast, there was a larger decrease in decoding performance when we omitted neurons with high decoding information content for both variables (ranked by the sum of their social and spatial decoding weight; Suppl. Fig. 4). A comparison between the performance decays upon exclusion of selective or informative neurons revealed that the decoding performance relied more on high-information neurons than on high-selectivity neurons (Suppl. Fig. 4).

Social and spatial features are represented in different subspaces of the neural activity space

Our results so far indicate that, although individual dCA2 neurons responded to different combinations of social and spatial variables, dCA2 activity contained sufficient information to decode social and spatial features during exploration of a novel and familiar animal.

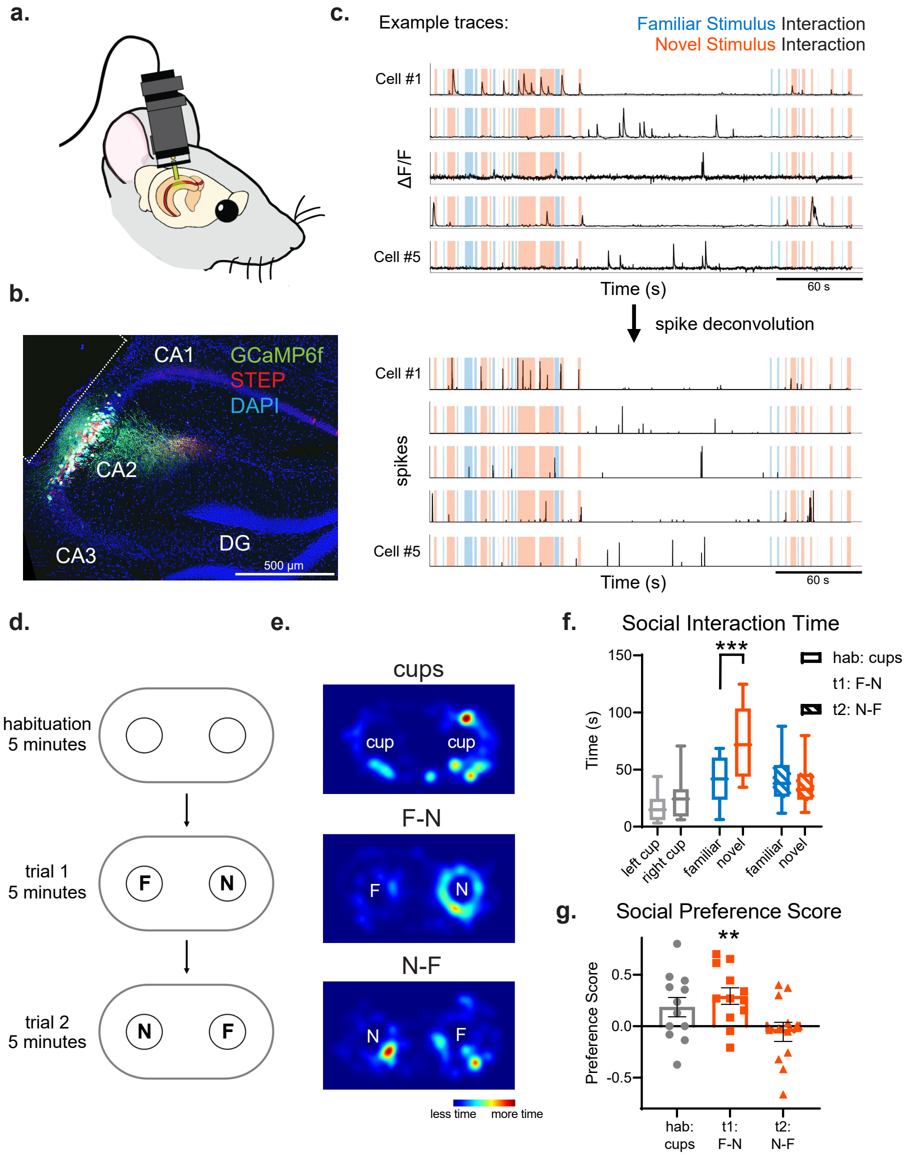


Fig. 1. Single dCA2 pyramidal neurons show distinct Ca2+ responses during social and spatial exploration. a) Amigo2-Cre mice were injected in dCA2 with Cre-dependent virus expressing GCaMP6f. dCA2 pyramidal neuron calcium levels were imaged via microendoscopy. b) Image of lens path over dCA2 showing co-expression (white cells) of GCaMP6f and STEP, a dCA2 marker protein. c) GCaMP6f fluorescent signals (top) and deconvolved spike traces (bottom) from five example dCA2 cells during trial 1. Periods of interaction with novel or familiar mice are color-coded. d) Experimental paradigm. A subject mouse was habituated to an arena containing two empty wire cup cages. In trial 1 the cups contained one novel (N) and one familiar mouse. In trial 2, the positions of the mice were swapped. e) Heatmap of subject mouse position in the three trials. f) Time spent actively interacting with empty cups or mice. Subjects interacted significantly more with the novel versus familiar mouse in trial 1, but not trial 2 (two-way ANOVA: Interaction Partner x Trial, F(2, 22) = 7.652, p<0.01. Šídák's multiple comparisons test: habituation trial (left versus right cup), p>0.05; trial 1 (N versus F), p<0.001; trial 2 (F versus N), p>0.05). g) Interaction discrimination index = [(time exploring left cup) – (time exploring right cup)]/[time spent exploring both cups] for three trials. A significant preference is observed in trial 1 only. One-sample t-test against zero: habituation trial, t=1.980, p=0.073; trial 1, t=3.644, p<0.01; trial 2, t=0.59, p>0.05. Bars show mean ± SEM. ** p<0.01, *** p<0.001.

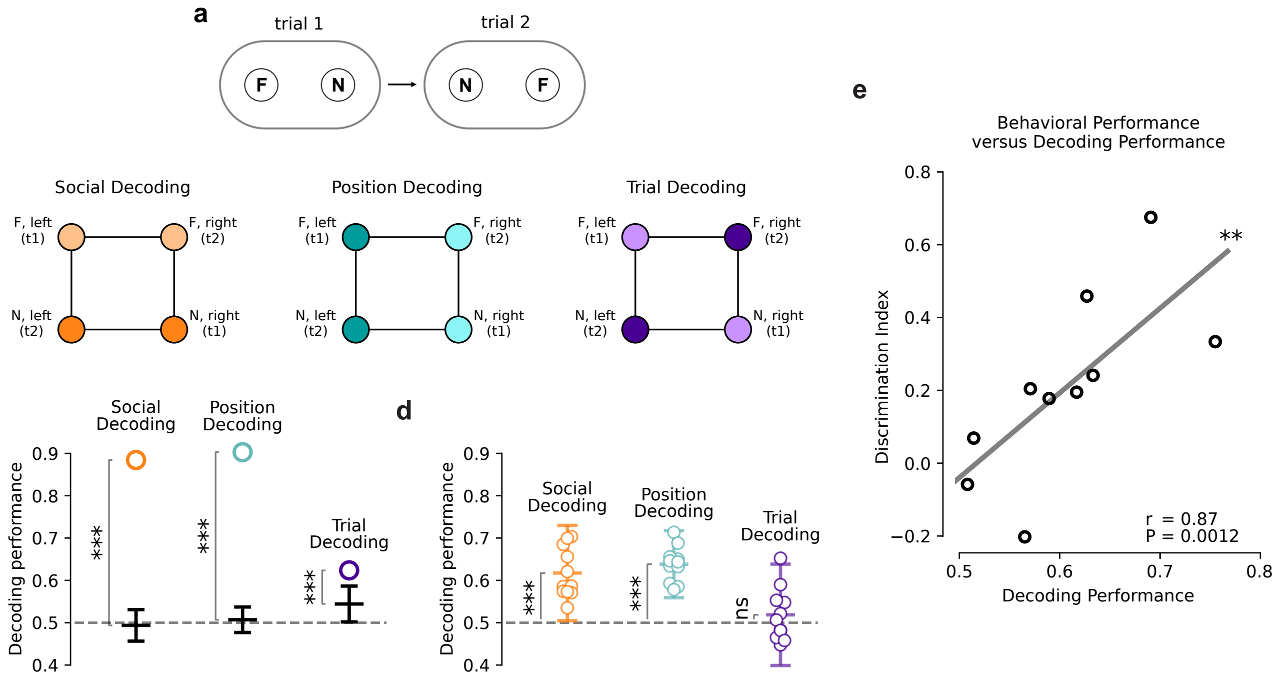


Fig. 2. Decoding social and position information from dCA2 population activity. a) Schema for experiment with a novel and familiar mouse as in Figure 1. b) Linear classifiers (SVMs) were trained to decode social interaction partner (N versus F), position (left versus right) and trial (trial 1 versus trial 2). Calcium spike data from both trials were grouped as indicated by colors. The decoded dichotomy is indicated by light vs. dark colors. c) Decoding of interaction partner (social), position, and trial from pseudo-simultaneous data from 1096 cells recorded from 11 mice. Open circles show average decoding performance from 20 cross-validations. Horizontal line and error bars show mean ± 2SD of distribution of chance values from shuffled data (see Methods). Social decoding performance = 0.88; chance = 0.49 ± 0.02 (mean ± SD, throughout figure); p<0.001. Spatial decoding performance = 0.90; chance = 0.51 ± 0.02; p<0.001. Trial decoding performance = 0.62; chance = 0.54 ± 0.02; p<0.001. In all cases statistical significance determined by z-score relative to chance distribution. d) Decoding performance from individual animals (open circles). Horizontal lines and error bars show mean ± SD SD determined from individual decoding values from 11 animals. Social decoding performance = 0.62 ± 0.06 (mean ± SD); p<0.001, paired t-test comparing decoding performance with chance value (0.5) calculated for each individual animal (t=6.25, n=11). Spatial decoding performance = 0.64 ± 0.04, p<0.001, paired t-test (t=11.12, n=11). Trial decoding performance = 0.52 ± 0.06; p>0.05, paired t-test (t=0.96, n=11). e) Decoding performance from individual animals during trial 1 was strongly correlated with the subject mouse's ability to discriminate the novel from familiar animal (Spearman's r=0.87, p<0.01, n=10). See Methods for a detailed discussion on chance level estimations and p values. ns = non-significant, ** p<0.01, *** p<0.001.

Next, we explored how these social and spatial variables interacted with each other in the population code. If the responses of individual neurons depend non-linearly on the two variables, the neural representations will be typically high dimensional (23). In this case, for example, a linear classifier trained to discriminate the location of social interactions with a specific individual will not be able to decode the position of interactions with a different conspecific. Conversely, neural activity could depend linearly on the two variables (linear mixed selectivity (24–26)), in which case the changes in neural response to alteration of one variable will be invariant to changes in the other variable. In this case a linear classifier trained to discriminate between two positions will generalize, allowing it to decode position when the identity of the animal at each position is changed. Analogously, a linear decoder trained to report the identity of a pair of animals will generalize when the position of the animals is altered. Such generalized representations can be considered to be abstract (27), as the classifiers report a given variable independent of changes in one or more other variables. These abstract representations — which are widely studied in the machine learning community, where they are called disentangled representations — have been observed in several brain areas (27–29).

To quantify the ability of dCA2 representations to generalize, we asked whether a linear classifier trained on one set of social and spatial conditions could accurately decode social and spatial information recorded in a different set of conditions not used for training. This measure is termed the cross-condition generalization performance (CCGP) (27). As illustrated in Figure 3b, high CCGP values for both social and spatial features require a low dimensional geometry of these representations in the dCA2 population firing space, with the social and spatial features represented in approximately orthogonal subspaces. CCGP for social novelty was determined first by training a linear classifier to distinguish the novel from the familiar mouse using the subset of dCA2 firing data recorded around only one of the two cups in each of the two trials, so that the spatial location was fixed. We then tested whether the linear classifier trained for social decoding around one cup (e.g., the left cup) could be used for the successful decoding of the novel and familiar animals when they were located in the other cup (e.g., the right cup). Similarly, we examined the spatial CCGP by first training the classifier based on dCA2 firing when the subject mouse explored a given mouse (e.g., the novel mouse) in the left and right cups across the two trials and testing whether that classifier could distinguish left from right using activity around the same two cups when they contained the other animal (e.g., the familiar mouse).

We found that dCA2 activity enabled a high CCGP using pseudo-population data for both social and spatial information (Figure 3c top). The same findings were corroborated by analyzing CCGP for each individual separately (Fig 3c bottom). Together, these results suggest that the representational geometry of social and spatial representations in dCA2 is relatively low dimensional, hence allowing simple linear decoders, implementable by single downstream neurons, to extract social and spatial information independently.

dCA2 provides a low-dimensional abstract representation of novelty versus familiarity

Although the above experiments showed that dCA2 activity can distinguish a novel from familiar mouse, the data so far did not allow us to determine whether dCA2 did indeed encode familiarity versus novelty or whether it encoded the two social identities of the distinct individuals present in these trials, independent of their degree of familiarity. To address this issue, we exposed subject mice (438 cells from 5 mice) to one pair of novel and familiar mice in trial 1 and then to a different pair of novel and familiar mice in trial 2, with the positions of the novel and familiar mice swapped between the two trials (Fig. 4a). On average the subjects explored the novel mice to a greater extent than the familiar mice on both trials (Suppl. Fig. 5). A linear classifier trained on dCA2 activity during interactions with the two novel versus the two familiar mice accurately decoded whether the subject mouse was interacting with a novel or familiar mouse (Suppl. Fig. 6). Similarly, a classifier trained on dCA2 activity when a mouse was exploring the left or right cup accurately reported the left-right position of the subject mouse (Suppl. Fig. 6).

To understand whether dCA2 activity provided an abstract representation of social novelty versus familiarity, we determined the CCGP for familiarity across the two trials with the two distinct pairs of mice (Fig. 4b). Remarkably, a classifier trained to distinguish interactions with one novel and familiar mouse located in the same cup across the two trials accurately decoded interactions with a distinct novel and familiar mouse located in the other cup across the two trials. We observed a significant CCGP for decoding familiarity versus novelty when the classifier was trained on either pseudo-population data or data from individual subjects (Fig. 4c,d). Thus, the representation of novelty versus familiarity in dCA2 was generalizable, or abstract, with respect to individual identity and subject position. Abstraction of familiarity was not simply the result of an invariance of the representations with respect to the position or the identity of individuals. Indeed, position could still be decoded (Suppl. Fig. 6), and it was also represented in an abstract format as we observed a highly significant CCGP value (Fig. 4c,d). We thus conclude that the hippocampal social representations, and specifically those in dCA2, have a sufficiently low-dimensional geometry to encode the abstract concept of social familiarity.

dCA2 encodes social identity for novel and familiar individuals

Given that familiarity is represented in an abstract format, it was natural to ask whether dCA2 also contained sufficiently detailed information to discriminate between individual identities, separately from their degree of familiarity. We approached this question using two additional social interaction tests in a paradigm similar to that used in Figure 1. In one test, the subject mouse explored a pair of novel mice placed under the two cups, whose positions were reversed in trials 1 and 2 (732 cells from 10 mice; Fig 5b). In the second test we placed a pair of familiar mice under the cups in the two trials of the test (1083 cells from 11 mice; Fig 5c). Thus, in both cases the two stimulus mice present during a given test had similar degrees of novelty or familiarity.

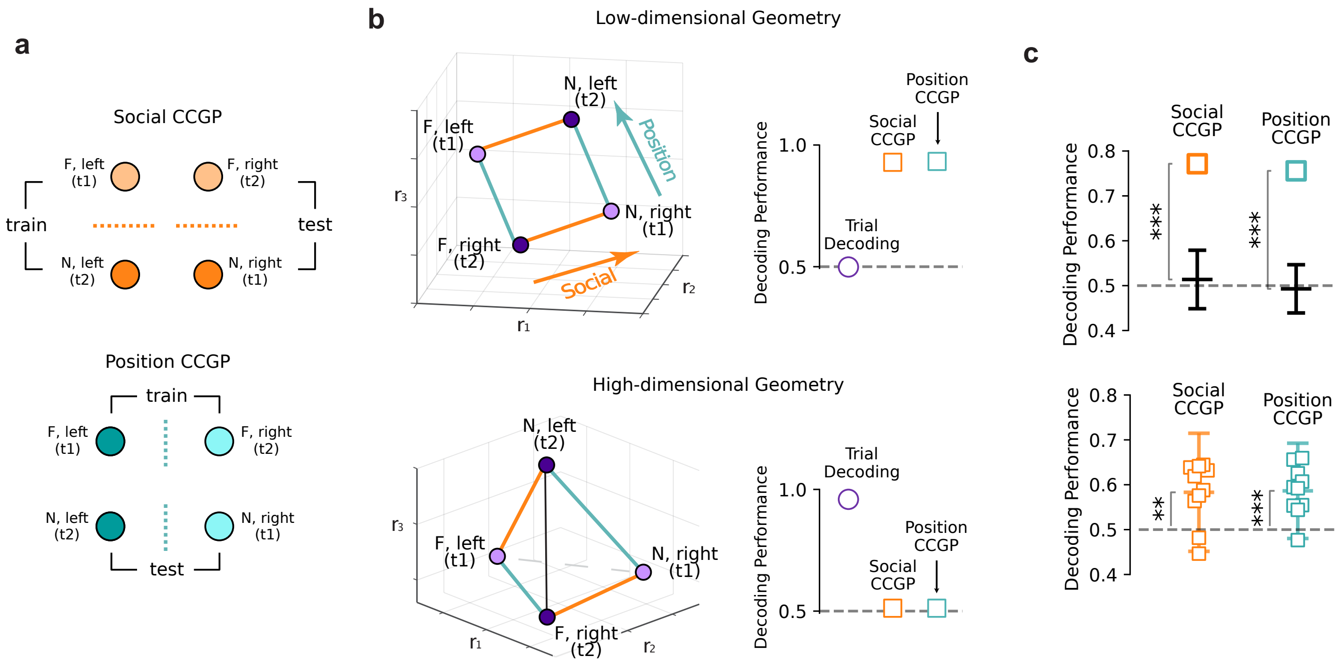


Fig. 3. Cross-condition generalization performance (CCGP) for decoding social and position information. a) Protocol for measuring CCGP. Top, social CCGP obtained by training classifier to distinguish N from F mouse when in one cup across trials (e.g., left) and testing that classifier to identity same mice when in the other cup (e.g., right), and vice versa (train on left, test on right). CCGP values averaged from left, right pairs. Bottom, spatial CCGP determined by training classifier to distinguish left versus right cup when they contain same animal (e.g., F) and testing by decoding left from right when cups contain the other animal (e.g., N), and vice versa. CCGP obtained from average of two spatial decoding results. b) Graphical representation showing how low dimensional (near planar) representation of four conditions of experiment in neural firing space provides for a higher CCGP compared to a higher-dimensional neural representation of same four conditions (tetrahedral, 3-dimensional geometry). c) CCGP values for social and position decoding. Top, CCGP determined for pseudo-simultaneous data. Social CCGP = 0.77, chance level = 0.51 ± 0.03 SD (mean ± SD throughout figure); p<0.001, z-score relative to chance distribution. Spatial CCGP = 0.76; chance level = 0.49 ± 0.03, p<0.001, z-score relative to chance distribution. Horizontal lines and error bars show mean ± 2 STDs of null model values. Bottom, CCGP values determined for individual animals (symbols). Social CCGP = 0.58 ± 0.07; p<0.01, paired t-test against chance levels (t=3.75, n=10). Spatial CCGP = 0.59 ± 0.05; p<0.001 (t=4.76, n=10). Horizontal lines and error bars show mean ± SD. ** p<0.01, *** p<0.001.

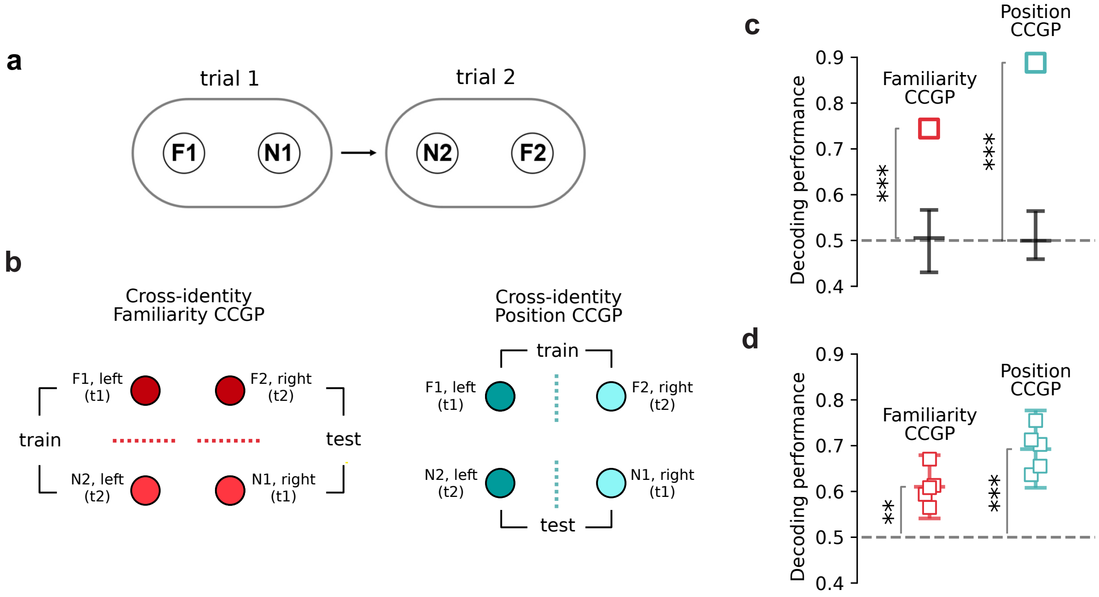


Fig. 4. dCA2 provides a generalized, abstract representation of novelty-familiarity. a) Schema of experiment. A subject mouse interacted for 5 min with one novel (N1) and one familiar mouse (F1) in trial 1 and interacted for 5 min with a different novel (N2) and familiar mouse in trial 2. Positions of novel and familiar mice were swapped in the two trials. b) Decoding scheme for calculation of CCGP for social novelty versus familiarity (left) and for right versus left position (right). In both cases, training and testing conditions were swapped and decoding results averaged to obtain CCGP. c) CCGP values calculated from pseudo-simultaneous population data. Familiarity (versus novelty) CCGP = 0.74; chance level = 0.51 ± 0.04 (mean ± SD, throughout figure); p<0.001, determined from z-score value relative to chance distribution. Position CCGP = 0.89; chance level = 0.50 ± 0.03; p<0.001 (z-score comparison). Horizontal lines and error bars show mean ± 2 SDs for distribution of chance values. d) CCGP values determined for individual subjects. Familiarity (vs novelty) CCGP = 0.61 ± 0.04; p<0.01, paired t-test against individual chance levels (t=5.79, n=5). Position CCGP = 0.69 ± 0.04; p<0.001, paired t-test (t=9.26, n=5). Horizontal lines and error bars show mean ± SD calculated from individual decoding values. ** p<0.01, *** p<0.001.

The subject mice showed no significant behavioral preference for one novel mouse over the other or one familiar mouse over the other (Suppl. Fig. 7). Despite the lack of behavioral preference, the linear decoder successfully classified interactions with one novel mouse versus the other using pseudo-simultaneous population data (Fig. 5a,d). The classifier similarly decoded the individual identity of familiar mice (Fig. 5e). Thus, dCA2 population activity contained sufficiently detailed information to distinguish two mice based on identity, and not only by differences in degree of novelty. In addition to decoding social identity, we were also able to decode left versus right position at a high accuracy when the cups contained either two novel animals or two familiar mice (Fig 5d,e). In contrast, the accuracy of trial decoding from the pseudo-population data was much lower than that for the social identity or position decoding, although significantly higher than chance, for both novel and familiar mice (Fig 5d,e). When we analyzed neural data from individual subjects, the decoding performance for social identity and position were significantly higher than chance (Suppl. Fig. 8). However, trial decoding was higher than chance levels only when the subject mice interacted with two familiar mice; trial decoding did not differ from chance levels when the subject mice interacted with two novel mice (Suppl. Fig. 8).

Distinct representational geometries of novel and familiar conspecifics

How does dCA2 meet the distinct demands of representing both the identities of novel and familiar animals required for recollection while also providing a generalized, abstract representation of novelty, required for familiarity detection? To approach this question, we investigated the geometries of novel and familiar representations, first by measuring the CCGP values for spatial position and social identity in the protocol described above using two novel and two familiar mice.

Our analysis of pseudo-simultaneous data yielded a significant CCGP value for social identity of two novel animals across different positions (Fig. 5f). The CCGP for social identity of two familiar animals was greater than chance levels, but lower than that obtained with the two novel animals (Fig. 5g). When we analyzed data from individual mice, we found a significant CCGP for social identity during interactions with the two novel animals (Suppl. Fig. 8), but not for the two familiar animals (Suppl. Fig. 8). In contrast, CCGP values for position were similar and significantly greater than chance for the two sets of conditions, both when determined for the pseudo-population data (Fig. 5f,g) and for individual mice (Suppl. Fig. 8).

To enable a more direct comparison between the two experiments, we calculated CCGP values for the subset of mice (630 cells from 7 mice) that was run in both the two-novel and two-familiar animal sessions. We found that the CCGP for social identity calculated with pseudo-population decreased markedly when the subject was exposed to two familiar mice compared to two novel mice (Fig. 6a). A smaller, yet significant, decrease was also observed for position CCGP (Fig. 6a). When we compared values calculated for each individual subject, we found that mean CCGP for social identity decreased

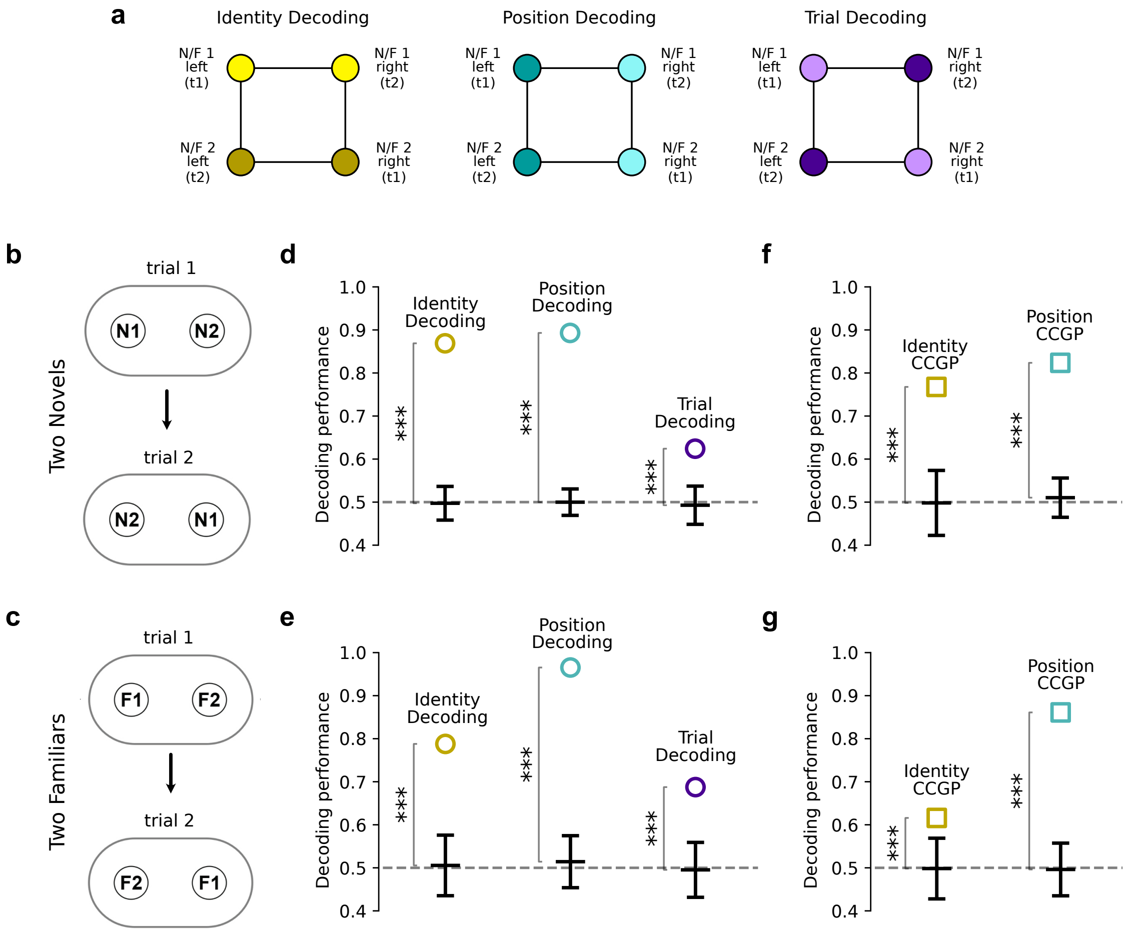


Fig. 5. Encoding of mouse identity, position and trial during interactions with novel or familiar mice. a) Scheme of decoding analysis. b,c) A subject mouse explored two novel (b) or familiar (c) animals in trial 1; the same two animals were present in trial 2 with positions swapped. d) Decoding of identity, position and trial with two novel mice using pseudo-simultaneous data. Novel identity decoding = 0.87; chance = 0.50 ± 0.02 (mean ± SD, throughout figure); p<0.001 (z-score comparison relative to chance). Position decoding = 0.89; chance = 0.50 ± 0.015; p<0.001 (z-score comparison). Trial decoding = 0.62; chance = 0.49 ± 0.02 p<0.001 (z-score comparison). e) Decoding of identity, position, and trial with two familiar mice using pseudo-simultaneous data. Familiar identity decoding = 0.79; chance = 0.51 ± 0.04; p<0.001 (z-score comparison). Position decoding = 0.97; chance = 0.51 ± 0.03; p<0.001 (z-score comparison). Trial decoding = 0.69; chance = 0.50 ± 0.032; p<0.001 (z-score comparison). f) CCGP for identity and position with two novel mice using pseudo-simultaneous data. Novel identity CCGP = 0.77; chance = 0.50 ± 0.04; p<0.001 (z-score comparison). Position CCGP = 0.82; chance = 0.51 ± 0.02; p<0.001 (z-score comparison). g) CCGP with two familiar mice. Familiar identity CCGP = 0.62; chance = 0.50 ± 0.04; p<0.001 (z-score comparison). Position CCGP = 0.86; chance = 0.50 ± 0.03; p<0.001 (z-score comparison). Horizontal lines and errors bars show mean ± 2 SD of chance distribution. *** p<0.001.

significantly during exploration of familiar compared to novel mice. Moreover, we observed a significant decrease in identity CCGP during trials with familiar compared to novel mice in 6 out of 7 subjects, with a significant increase in identity CCGP in only 1 out of 7 subjects (Suppl. Fig. 9). In contrast, we did not observe a consistent effect of familiarity on position CCGP (Suppl. Fig. 9).

The differences in CCGP values for social identity indicate that dCA2 encodes novel animals in lower dimensional representations compared to familiar individuals. Support for this conclusion comes from an analysis of the ability of a linear classifier to decode the trial variable. The low-dimensional planar, rectangle-like representation of social identity and position with novel animals (Fig 6b, dark grey points) implies that the pairs of conditions in the two trials correspond to the opposite corners of the rectangle. In this case a linear classifier will not be able to separate one trial from the next as the classification problem is equivalent to solving the exclusive OR (XOR) problem (where the two conditions of each trial that are grouped together by the classification task share no common social or spatial variable). In contrast, a higher dimensional tetrahedral representation (Fig. 6b, light gray points) would allow the points in opposite corners to be separated by a linear plane, thus enabling trial decoding. As the trial variable is the only dichotomy of the four conditions that is non-linearly separable, a higher than chance decoding accuracy for trial (provided the other two variables, identity and position, are also decodable), indicates that the representation is high dimensional and the points can be divided into two groups (shattered) in all possible ways by a linear decoder (high shattering dimensionality (23, 27)).

As noted above in discussing the results of Figure 5, trial decoding with data from the entire pseudo-population was greater during interactions with two familiar animals than with two novel animals (see also Suppl. Fig. 8). To determine whether this difference was statistically significant, we compared trial decoding in the two-novel or two-familiar animal sessions when calculated for those individual subject mice run in both tasks, as done for the CCGP analysis. We found that trial decoding from the pseudo-simultaneous population of cells recorded from this subset of animals was significantly greater during interactions with two familiar mice compared to two novel mice (Fig 6a). We similarly found a consistently greater accuracy of trial decoding during interactions with familiar mice compared to novel mice when analyzing data for individual subjects (significant increase in 5 out of 7 subjects; no subjects showed a significant decrease, Suppl. Fig. 9). The trial decoding data therefore provides additional support that familiar animal representations have a higher dimensionality compared to novel animal representations.

To provide a comprehensive overview of our conclusions, we devised a geometric model that was able to capture our key findings. We first assumed that novel individuals are represented in a low dimensional geometry, illustrated in this example by having the activity of a sample population of three neurons confined to a two-dimensional plane (Fig. 6b). We then posited that increasing levels of familiarity progressively shift the neural plane away from the novel animal representations, accounting for the abstract decoding of familiarity seen in Fig. 4 (the coding directions of familiarity, red in the figure, are approximately parallel, meaning that a decoder trained on one animal in one position would generalize to other animals and positions). To explain the decreased CCGP associated with trials containing two familiar mice compared to two novel mice, as well as the increased ability to decode the trial variable when interacting with familiar mice (Fig. 6a), we posited a relatively small random displacement of the neural firing points for familiar animals away from the shifted planar representation, providing an increased dimensionality. Finally, because decoding performance was higher for decoding the identities of two novel animals compared to two familiar animals (Fig. 5d,e, Suppl Fig. 10) we surmised that the distance between representations on the identity axis was reduced with increasing familiarity. This geometric description was sufficient to recapitulate all our major findings, across tests (Fig. 6c, Suppl Fig. 10).

Discussion

The classic "butcher on the bus" scenario posits that social recognition memory has at least two components: familiarity—the sense of whether one has previously encountered an individual—and recollection—the detailed recall of previous encounters with specific individuals1. Up to now, the neural mechanisms underlying social familiarity and recollection have been obscure, including uncertainty as to whether a given brain region can participate in both processes (8–13, 30). Our experiments and analyses, based on large-scale calcium imaging of hippocampal dCA2 pyramidal neuron from mice engaged in social/spatial interactions, demonstrate that dCA2 encodes both social familiarity and social identities of individuals with similar degrees of novelty or familiarity, the latter being a key requisite of recollection. Moreover, we find that familiar individuals are encoded with a higher dimensional geometry than novel individuals, one consequence of which may be to enable the richer memory store associated with familiar individuals compared to novel ones.

We focused on dCA2 because of its prominent role in the encoding, consolidation and recall of social memory, as assessed by the ability of a mouse to distinguish between a novel and familiar conspecific (6, 14, 21). Although previous studies using in vivo recordings found that dCA2 neuron firing responds during social interactions (16–18) and can distinguish between a novel and familiar animal (17), dCA2 neurons also act as place cells, firing as an animal explores specific locations (16–18, 31, 32). Thus, it has to date been unclear as to whether and how dCA2 social/spatial firing can be disentangled by downstream neurons to decode social novelty versus familiarity and spatial location. Moreover, prior studies did not fully address whether dCA2 represents the identity of individual conspecifics with identical degrees of social novelty or familiarity nor how the representations of novel and familiar conspecifics may differ.

Using a linear classifier to decode social and spatial interactions, we found that dCA2 activity encodes both social and spatial information. Although most dCA2 neurons respond to both social and spatial cues, the geometry of these social/spatial representations in dCA2 firing space provides a

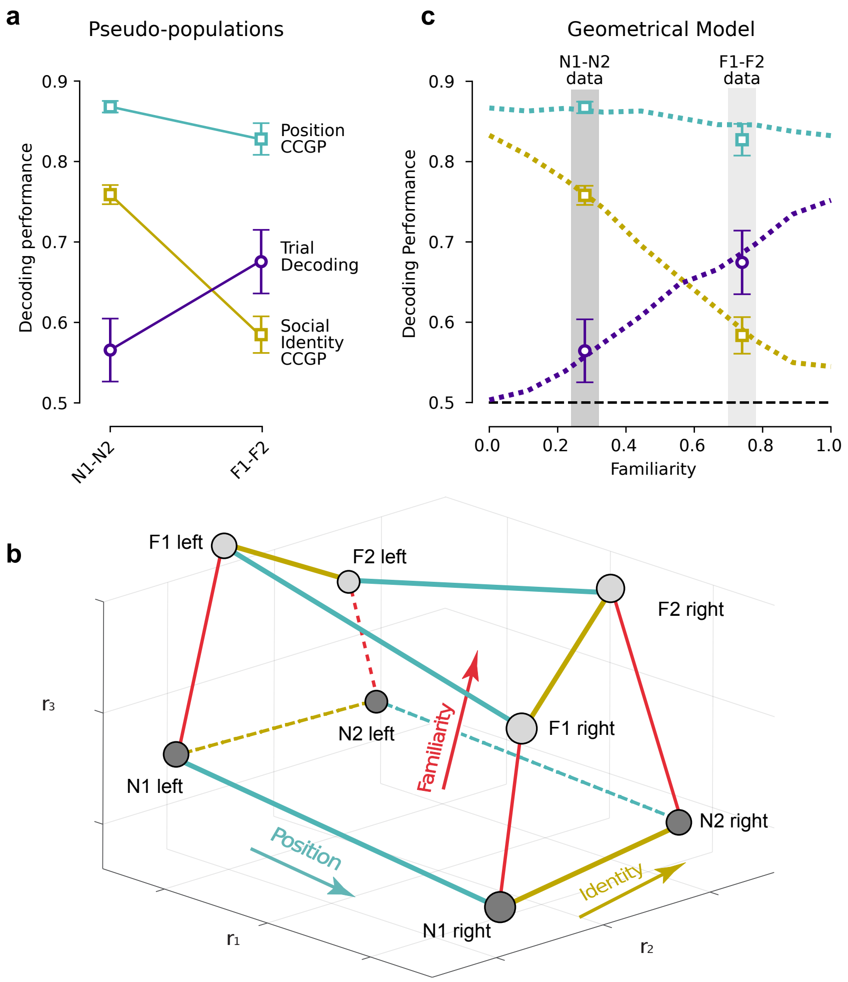


Fig. 6. Familiar conspecifics are represented in higher dimensions than novel conspecifics in dCA2. a) Comparison of CCGP values during interactions with two novel and two familiar mice, computed on pseudo-simultaneous data. Social identity CCGP during interactions with novel animals (0.76 ± 0.01; mean ± SD, throughout figure) was significantly greater than social identity CCGP with familiar animals (0.58 ± 0.02; $p<0.001$, Mann-Whitney U test). Position CCGP during interactions with novel animals (0.87 ± 0.01) was greater than during interaction with familiar animals (0.83 ± 0.02; $p<0.001$, Mann-Whitney U test). Trial decoding during interactions with two novel animals (0.59 ± 0.04) was less than with two-familiar animals (0.68 ± 0.04; $p<0.001$, Mann-Whitney U test). b) Graphical representation of geometrical model for encoding of social and spatial information by three example neurons (firing rates r1, r2, r3). Dark and light gray circles represent firing rates during specific combinations of social and spatial variables during interactions with novel and familiar animals, respectively. Increasing familiarity both shifts and distorts in neural firing space the planar, low-dimensional social-spatial representations of novel animals. c) Model qualitatively reproduces experimental observations (data from panel a is overlaid on predictions from model). Symbols and error bars in a,b show mean ± SD, calculated using 20 cross validations for Trial decoding, and over 20 pseudo-population re-sampling for CCGP (see Methods). Lines in c show mean values over 200 simulations (see Methods).

low-dimensional representation that allows a linear classifier to decode social information in a way that generalizes to novel locations (i.e., locations not used to train the decoder). Moreover, a linear classifier trained to decode position can readily generalize to a situation in which novel animals are placed at the same locations. We also found that the ability of dCA2 social representations from a given mouse to decode a novel from familiar animal was highly correlated with the behavioral ability of that mouse to distinguish a novel from familiar animal, supporting the view that dCA2 representations may be important for encoding social memory.

In addition to detecting novelty versus familiarity, we found that dCA2 also encodes the identities of individual animals with identical degrees of novelty or familiarity, enabling a linear classifier trained on dCA2 firing to distinguish one novel mouse from another or one familiar littermate from another. Of particular importance, the representations of the novel and familiar animals adopt distinct geometries. This enables a simple, generalized read-out of novelty versus familiarity, which may contribute to the ability of both mice and humans to rapidly distinguish a familiar individual from a stranger.

A geometrical model captured the key elements of our findings, in which increasing familiarity causes a roughly parallel shift in the low dimensional manifold of novel animal representations, enabling the generalized decoding of novel from familiar animals. At the same time, familiarization distorts the low-dimensional representation seen with novel animals with non-linear perturbations that are different for every combination of mouse identity and position. These perturbations make the representations higher dimensional.

What are the consequences of having familiar identities represented in a higher dimensional space than novel identities? A number of studies have reported that the dimensionality of neural activity is linked to functional or behavioral complexity (23, 27, 33–35). Low-dimensional representations, and in particular disentangled representations (29), are generally robust to noise and allow for generalization not possible with high dimensional representations. These disentangled representations also allow dCA2 to represent a large number of different situations (e.g. different animals encountered at different locations in different contexts), as large as all the novel situations that an animal can potentially encounter. If L is the number of disentangled variables (latent variables) that are represented and each variable has only two values, then the number of representable situations scales as 2^L, which can be much larger than the number of neurons N. Some of these situations are actually experienced by the animal and so could be stored in memory. Recollecting a memory can be modeled as a process in which the pattern of activity that represents a particular experience is fully reconstructed at a later time (36–39), when a memory cue is presented. In the case of disentangled representations, the number of memories that can be stored and recollected is relatively low, scaling only linearly with L (see Supplemental Information), and is exceeded by the memory capacity of high-dimensional representations, which scales linearly with N (see e.g. (36, 37)). Indeed, disentangled representations that have similar geometrical properties as those that we observed can be constructed by combining together L separate populations of neurons, each encoding a single latent variable. The correlations between all the neurons within each population make the representations low dimensional. Each population of neurons can be regarded as a single effective neuron, and hence the memory capacity will scale with the number of independent effective neurons L, and not with the total number of neurons N.

We suggest that when a novel situation is actually experienced by an animal its low-dimensional representation is transformed into a higher-dimensional representation suited for storage in episodic memory, likely through the process of synaptic plasticity, thereby greatly enhancing the number of memories that can be stored. Interestingly, this transformation still allows familiarity to be represented as an abstract variable. In the simple geometrical models described in Figure 6, we showed that the geometry of familiar social/spatial representations can be obtained by transforming the geometry of the social/spatial representations of novel animals in two steps. The first is a rigid translation in the activity space. This allows familiarity to be encoded in an abstract format. The second is a relatively small shift of each different social/spatial condition (i.e., a combination of position and identity) in different directions in neural activity space, making the representations of familiar animals higher dimensional. This transformation allows familiarity to be represented in an abstract format without sacrificing the elevated memory capacity of high dimensional representations.

Recent findings on the representational geometry for both familiar and novel faces in monkey inferotemporal (IT) cortex (26) show certain similarities and differences with our results. Similar to our findings, the representations for novel faces are low dimensional (see also (25)). At short latencies, the dimensionality of familiar and unfamiliar face representations is similar, with the two geometries related by a simple translation. In contrast at longer latencies, the geometry of familiar representations becomes distorted. It is unclear whether this distortion is non-linear, as in our data, or whether it could be explained with a linear transformation, which would preserve the dimensionality of the representations. One clear difference is that the distortion observed in IT enables improved discrimination of familiar faces whereas in our case the ability of the decoder to discriminate between two animals remains the same or slightly decreases for familiar compared to novel animals. This could be due to a difference between neocortex and hippocampus, a difference between primates and rodents, and/or a difference in the nature of social familiarity for two-dimensional pictures of faces compared to interactions with living conspecifics.

Our results provide the first demonstration, to our knowledge, that experience-dependent changes in representational geometry can contribute to the differential cognitive processing of novel and familiar individuals. The balance of generalization and flexibility may be an important feature that guides the encoding of complex social relationships to form a cognitive map of social space. Such coding may be required for navigating complex social behaviors, such as pair bonding, social aggression, and the creation of social dominance hierarchies.

Funding. : The work from this publication was supported by National Institute of Mental Health of the National Institutes of Health under award number F30 MH120922-01A1 (L.M.B.) and grants R01-MH104602 and R01-MH116190 from NIH (PI, S.A.S.). L.P. and S.F. were supported also by NSF Neuronex, the Simons Foundation, the Gatsby Charitable Foundation and the Swartz Foundation.

Author Contributions. : L.M.B. and S.A.S conceived the project and designed the experiments, L.M.B. and S.I. collected the data, and L.M.B. and L.P. analyzed the data with guidance from S.F. The data was interpreted by L.M.B., L.P., S.A.S, and S.F., who wrote the paper with feedback from S.I.

Competing interests. : Authors have no competing interests to report.

ACKNOWLEDGMENTS. We thank T. Tabachnik for help designing and obtaining the oval arena; R. Sahai, K. Lewis, and A. Fisher for assistance in obtaining immunofluorescent images and implementing DeepLabCut; D. Salzman, R. Hen, S. Hassan, P Kassraian-Fard, and A Villegas for critical discussions and comments on the manuscript. We are grateful to R Nogueira, V Fascianelli, S Muscinelli, and J Minxha for many valuable and knowledgeable discussions.

References.

1. G Mandler, Recognizing: The judgment of previous occurrence. *Psychol. review* **87**, 252 (1980).
2. LR Squire, Memory systems of the brain: a brief history and current perspective. *Neurobiol. learning memory* **82**, 171–177 (2004).
3. T Hartley, C Lever, N Burgess, J O'Keefe, Space in the brain: how the hippocampal formation supports spatial cognition. *Philos. Transactions Royal Soc. B: Biol. Sci.* **369**, 20120510 (2014).
4. S Steinvorth, B Levine, S Corkin, Medial temporal lobe structures are needed to re-experience remote autobiographical memories: evidence from hm and wr. *Neuropsychologia* **43**, 479–496 (2005).
5. JH Kogan, PW Franklandand, AJ Silva, Long-term memory underlying hippocampus-dependent social recognition in mice. *Hippocampus* **10**, 47–56 (2000).
6. FL Hitti, SA Siegelbaum, The hippocampal ca2 region is essential for social memory. *Nature* **508**, 88–92 (2014).
7. T Okuyama, T Kitamura, DS Roy, S Itohara, S Tonegawa, Ventral ca1 neurons store social memory. *Science* **353**, 1536–1541 (2016).
8. JT Wixted, LR Squire, The role of the human hippocampus in familiarity-based and recollection-based recognition memory. *Behav. brain research* **215**, 197–208 (2010).
9. A Kafkas, D Montaldi, How do memory systems detect and respond to novelty? *Neurosci. letters* **680**, 60–68 (2018).
10. E Atucha, A Karew, T Kitsukawa, MM Sauvage, Recognition memory: Cellular evidence of a massive contribution of the lec to familiarity and a lack of involvement of the hippocampal subfields ca1 and ca3. *Hippocampus* **27**, 1083–1092 (2017).
11. MB Merkow, JF Burke, MJ Kahana, The human hippocampus contributes to both the recollection and familiarity components of recognition memory. *Proc. Natl. Acad. Sci.* **112**, 14378–14383 (2015).
12. CB Kirwan, JT Wixted, LR Squire, A demonstration that the hippocampus supports both recollection and familiarity. *Proc. Natl. Acad. Sci.* **107**, 344–348 (2010).
13. MM Sauvage, NJ Fortin, CB Owens, AP Yonelinas, H Eichenbaum, Recognition memory: opposite effects of hippocampal damage on recollection and familiarity. *Nat. neuroscience* **11**, 16–18 (2008).
14. EL Stevenson, HK Caldwell, Lesions to the ca 2 region of the hippocampus impair social memory in mice. *Eur. J. Neurosci.* **40**, 3294–3301 (2014).
15. AS Smith, SW Avram, A Cymerblit-Sabba, J Song, WS Young, Targeted activation of the hippocampal ca2 area strongly enhances social memory. *Mol. psychiatry* **21**, 1137–1144 (2016).
16. GM Alexander, et al., Social and novel contexts modify hippocampal ca2 representations of space. *Nat. communications* **7**, 1–14 (2016).
17. ML Donegan, et al., Coding of social novelty in the hippocampal ca2 region and its disruption and rescue in a 22q11. 2 microdeletion mouse model. *Nat. Neurosci.* **23**, 1365–1375 (2020).
18. A Oliva, A Fernández-Ruiz, F Leroy, SA Siegelbaum, Hippocampal ca2 sharp-wave ripples reactivate and promote social memory. *Nature* **587**, 264–269 (2020).
19. CM MacLeod, The butcher on the bus: A note on familiarity without recollection. *Hist. Psychol.* **23**, 383 (2020).
20. K Okamoto, Y Ikegaya, Recurrent connections between ca2 pyramidal cells. *Hippocampus* **29**, 305–312 (2019).
21. T Meira, et al., A hippocampal circuit linking dorsal ca2 to ventral ca1 critical for social memory dynamics. *Nat. communications* **9**, 1–14 (2018).
22. SI Hassan, S Bigler, SA Siegelbaum, Coding of social odors in the hippocampal ca2 region as a substrate for social memory. *bioRxiv* (2021).
23. M Rigotti, et al., The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
24. D Raposo, MT Kaufman, AK Churchland, A category-free neural population supports evolving demands during decision-making. *Nat. neuroscience* **17**, 1784–1792 (2014).
25. L Chang, DY Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).
26. L She, MK Benna, Y Shi, S Fusi, DY Tsao, The neural code for face memory. *bioRxiv* (2021).
27. S Bernardi, et al., The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
28. R Nogueira, CC Rodgers, RM Bruno, S Fusi, The geometry of cortical representations of touch in rodents. *bioRxiv* pp. 2021–02 (2021).
29. I Higgins, et al., Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. communications* **12**, 1–14 (2021).
30. A Kafkas, D Montaldi, Two separate, but interacting, neural systems for familiarity and novelty detection: A dual-route mechanism. *Hippocampus* **24**, 516–527 (2014).
31. L Lu, KM Igarashi, MP Witter, EI Moser, MB Moser, Topography of place maps along the ca3-to-ca2 axis of the hippocampus. *Neuron* **87**, 1078–1092 (2015).
32. EA Mankin, GW Diehl, FT Sparks, S Leutgeb, JK Leutgeb, Hippocampal ca2 activity patterns change over time to a larger extent than between spatial contexts. *Neuron* **85**, 190–201 (2015).
33. R Nogueira, et al., Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nat. Commun.* **8**, 1–13 (2017).
34. CC Rodgers, et al., Sensorimotor strategies and neuronal representations for shape discrimination. *Neuron* (2021).
35. S Fusi, EK Miller, M Rigotti, Why neurons mix: high dimensionality for higher cognition. *Curr. opinion neurobiology* **37**, 66–74 (2016).
36. JJ Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. national academy sciences* **79**, 2554–2558 (1982).
37. DJ Amit, *Modeling brain function: The world of attractor neural networks.* (Cambridge university press), (1989).
38. MA Gluck, CE Myers, Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus* **3**, 491–516 (1993).
39. MK Benna, S Fusi, Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. *Proc. Natl. Acad. Sci.* **118** (2021).
40. P Zhou, et al., Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife* **7**, e28728 (2018).
41. F Pedregosa, et al., Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).
42. A Treves, ET Rolls, Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus* **2**, 189–199 (1992).
43. RC O'reilly, JL McClelland, Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* **4**, 661–682 (1994).
44. M Allegra, L Posani, R Gómez-Ocádiz, C Schmidt-Hieber, Differential relation between neuronal and behavioral discrimination during hippocampal memory encoding. *Neuron* **108**, 1103–1112 (2020).
45. L Personnaz, I Guyon, G Dreyfus, Information storage and retrieval in spin-glass like neural networks. *J. de Physique Lettres* **46**, 359–365 (1985).
46. I Kanter, H Sompolinsky, Associative recall of memory without errors. *Phys. Rev. A* **35**, 380 (1987).

Materials and Methods

Viral injection and GRIN lens implantation.

Calcium imaging. A 200 nL volume of AAV2/1.syn.FLEX.GCaMP6f. WPRE.SV40 virus (titer: 6.5×10^{11} pp/mL, Penn Vector Core) was injected at a rate of 150 nL/min into the right hemisphere above dorsal hippocampal CA2 using stereotactic coordinates: AP -2.0 mm, ML +1.8 mm, DV -1.7 mm from bregma of 3-6 month-old male heterozygous Amigo2-Cre (Cre$^{+/-}$) mice. Three weeks following injection, a 1.2 mm diameter circular craniotomy was centered at the following coordinates: AP -2.0 mm, ML +2.5 mm. We inserted a GRIN lens (Inscopix, 1.0 mm diameter, 4.0 mm length) into the craniotomy at a depth of -1.4 to -1.5 mm relative to bregma at a 10° angle from the midline, so that the lens was parallel to the CA2 cell body layer. The Inscopix Proview system imaged cells during implantation to adjust the position of the lens to optimize visible fluorescence. Kwik-sil was placed around the craniotomy and the lens secured in place using Metabond dental cement. The top of the Proview lens cuff was filled with Kwik-cast to protect the lens. Mice were housed with littermates for one week before a plastic baseplate was placed over the lens and secured with Metabond dental cement. The baseplate and microscope were placed over the lens and the position was adjusted until cells were maximally in focus.

Pharmacogenetic silencing of CA2. We injected 8 Amigo2-Cre$^{-/-}$ (controls) and 12 Amigo2-Cre$^{+/-}$ mice in dCA2 with a Cre-dependent virus expressing the inhibitory hM4Di designer receptor exclusively activated by designer drugs (iDREADD), AAV2/8 hSyn.DIO.hM4D (Gi)-mCherry. 200 nL of virus (1.9×10^{12} pp/mL) was injected into dCA2 bilaterally using the following coordinates: anteroposterior (AP) -2.0mm, mediolateral (ML) +/-1.8mm, dorsoventral (DV) -1.7mm.

Extraction of Calcium Signals.

Data Acquisition, Preprocessing and Motion-correction. On the day of the experiment, mice were moved to the behavior room and subject mice and littermates were separated into holding cages. Mice were allowed to acclimate to the environment for 30 minutes. An nVista 3.0 Inscopix miniaturized microscope was inserted into the baseplate and used to record calcium fluorescence from dCA2 pyramidal neurons during social and non-social behavior using Inscopix data acquisition software (20 frames per second, 50-ms exposure, 0.2-0.3 mW/mm^2 EX-LED). The working distance between the microscope objective and the lens was adjusted to maximize cell focus, and this distance was maintained between trials and from session to session. To align behavior and calcium videos, a 5V TTL pulse from an AMi-2 Optogenetic interface triggered calcium recordings through Anymaze softward at the start of each trial along with a behavior video recording. Behavior recordings were collected at a rate of 20 Hz. The raw videos from separate sessions were concatenated and then run through Inscopix Data Analysis software. Videos were preprocessed to correct defective pixels and 4x spatially down-sampled. Background fluorescence was removed using a spatial band-pass filter and fluorescence videos were motion-corrected using the Inscopix motion correction algorithm. The preprocessed and motion corrected tiff files were then exported for cell identification and signal deconvolution.

Segmentation and ROI Identification. Cell regions-of-interest (ROIs) were identified using the Python CaImAn package for large-scale calcium imaging data. The spatial footprints and deconvolved signal for the active sources (ROIs) were extracted using CNMFe (40), and then the scaled raw traces and spatial footprints were exported to Matlab. We used a custom GUI to evaluate individual ROIs and spatial footprints, and those with non-spherical or non-oval shapes caused by motion artifacts were excluded from analysis. We detrended the raw traces over a window of 50 s using custom scripts. Finally, the computed traces, separated by session, were deconvolved using the OASIS algorithm for nonnegative signal deconvolution (baseline = trace median, noise = trace MAD, spike thresholds = 2x MAD).

Behavior.

Calcium recordings. We imaged dCA2 pyramidal neurons in a total of fifteen Amigo2-Cre heterozygous mice (with one excluded due to non-specific expression in dCA1) in multiple tests probing social recognition and memory. Prior to the first test, mice were handled and habituated for three days on the following schedule: Handling (day 1), handling, exposure to oval arena for 15 minutes (day 2), handling, exposure to holding cage for 30 minutes, scruffing/insertion of the microscope, and to the oval arena for 15 minutes with microscope inserted (day 3). Mice were additionally habituated in the oval arena to empty cups for 10 minutes. No changes in subject mouse behavior, including during social interaction, were observed compared to wild-type controls.

In each test, subject mice were placed into an oval arena that consisted of two half-circles with radius 15 cm connected to a central square area with length of 30 cm (total dimensions: length 60 cm, width 30 cm, height 45 cm). Wire pencil cups (radius 5 cm) were placed 10 cm from the two ends of the arena along the midline and will hereafter be referred to as left cup and right cup. Stimulus mice were placed underneath the cups as described for each test. Between consecutive trials, subject mice were removed to a holding cage to which they had been previously habituated for approximately 2 minutes while the oval arena was cleaned with 70% alcohol wipes to remove any olfactory cues, wiped with paper towels, cleaned with water, and then wiped with paper towels until dry. The cups with or without stimulus mice were re-introduced to the arena, and finally the subject mouse was re-introduced into the arena and the trial initiated in ANY-maze. The position of the two stimulus mice were randomized to the left or right cups in the first trial, and the positions then swapped in the second trial. Stimulus mice were age- and sex-matched to subject mice (males 3-6 months old).

In each trial, the subject mouse was free to explore the arena. Periods of interaction with cups or conspecifics in the arena, defined as times when the subject's head was oriented towards the center of the cup within a zone equal to 1.5x the cup radius (7.5cm), while the subject was actively sniffing, were manually scored. In a minority of tests and trials, the subject mouse climbed on top of the wire pencil cups. In these cases, the period atop the cup was excluded from analysis. The behavior videos were run through a deep neural network trained using DeepLabCut to recognize the position of the mouse head and body, as well as location of the objects placed in the arena. Errors in the DeepLabCut output were corrected using an automated custom Matlab script.

Familiar versus novel mouse recognition test. Twelve subject mice underwent the following three 5-min trials: habituation trial, two empty cups (left and right); trial 1, novel mouse and familiar littermate in the two cups; trial 2, same novel mouse and familiar littermate with positions swapped (Fig. 1c).

Social Novelty Recognition Test. Five subject mice underwent the following three 5-min trials: habituation trial, two empty cups; trial 1, novel mouse 1 and familiar littermate 1; trial 2, novel mouse 2 and familiar littermate 2, with novel/familiar animal positions swapped relative to trial 1 (Fig. 4a).

Interaction with mice with similar degrees of novelty or familiarity. Twelve subject mice were exposed to two novel mice using three 5-min trials: habituation trial, two empty cups; trial 1, two novel mice; trial 2, the same two novel mice with positions swapped (Fig. 5b). Twelve subject mice were exposed to two familiar littermates using three 5-min trials: habituation trial, two empty cups; trial 1, two familiar littermates in the cups; trial 2, the same familiar littermates with positions swapped (Fig. 5c).

Subject mice were run through one or more of the above social memory tests. If a subject was run through more than one test, subsequent tests were run one week apart. As an exception, subjects run first through the test for individual preference (familiar) then through the test for familiar recognition were run on the same day with a thirty-minute delay period between tests. The breakdown of tests per subject is as follows (in test order): 7/15 subjects were given the two-novel mice interaction test followed by the two-familiar mice interaction test followed by a test for familiar recognition (stimulus identities same between trials). 3/15 two-novel mice interaction test only, 2/15 familiar recognition test (stimulus identities changed between trials), familiar recognition test (stimulus identities same between trials), two-novel mice interaction test, two-familiar interaction test. 3/15 familiar recognition test (stimulus identities changed between trials), familiar recognition test (stimulus identities same between trials), two-familiar interaction test.

Behavior Statistical Analysis. To determine whether there were significant differences in the interaction times of the subject mouse with different social and non-social stimuli, we ran a two-way ANOVA of trial and interaction partner with repeated measures for both factors using Graphpad Prism software (version 9.0.1). Šidák's multiple comparisons test was used post-hoc to determine significant differences across trials or between interaction partners. Statistical significance was defined as $p < 0.05$. In addition, for each trial we calculated a preference score for interacting with each partner through the following equation:

$$\text{Preference Score}\,(B:A) = \frac{t_B - t_A}{t_B + t_A}$$

Where t_A is the length of time the subject mouse interacted with one mouse and t_B is the length of time the subject mouse interacted with the other mouse. We used the one-sample t-test to determine whether the preference score was significantly different than zero.

Effect of CA2 silencing on social memory. Three weeks after iDREADD viral injection, Amigo2-Cre heterozygous mice (n=12) and wild-type littermates (n=8) were habituated to IP injection for four days. On the third and fourth day, mice were additionally habituated to the same oval arena used in calcium recording experiments for 5 minutes and to an individual holding cage for 30 minutes. On the fifth day, mice were moved to the experimental room and allowed to acclimate to the environment for 30 minutes in their individual holding cages. Mice were then injected intraperitoneally 30 minutes prior to testing with 10 mg/kg clozapine-n-oxide (CNO), the ligand for the iDREADD receptors, to reduce CA2 activity.

30-minutes post-injection, subject mice were run through two 5-minute learning trials in the oval arena: trial 1, novel mouse 1 and novel mouse 2 in the two cups; trial 2, the same two mice with positions swapped. In between each trial, the subject mouse was returned to the holding cage for approximately 2 minutes. Following trial 2, the subject mouse was returned to its holding cage. After a two-hour interval, the subject mouse was returned to the arena for two memory recall trials: trial 3, one of the previously encountered mice in the learning trials (e.g. novel 1) and a third previously unencountered novel mouse (novel 3); trial 4, the same two mice with positions swapped. The behavior videos were manually scored for interactions, defined by the same criteria as those applied during calcium imaging behavior, by an investigator blinded to the identities of the subject mice and the individuals under the cups. Memory recall was assessed by the greater interaction time with novel 3 compared to the previously encountered mouse, using the same statistical analysis described above.

Population decoding analysis.

Linear classifier. The decoding analysis was performed using a linear classifier based on a support vector machine with custom-written Python scripts based on the scikit-learn SVC package (41).

Data labeling. For each subject and session, we selected neural data corresponding to periods in which the subject was actively interacting with one of the two cups. We then divided the neural recordings into 100 ms time bins and labeled them according to whether the subject was interacting with the left or right cup and to the identity of the animal under the cup (labeled as F, as in familiar, or N, as in novel, in the familiar versus novel recognition test, and #1 or #2 for novel-novel and familiar-familiar interaction tests). In each test there were always two trials, with the positions of animals swapped in trials 1 and 2. Thus, for each test there were a total of 4 social/spatial conditions [e.g., for the familiar-novel test: familiar on left (F, left), familiar on right (F, right), novel on left (N, left), novel on right (N, right)]. We then divided the four conditions into binary dichotomies (class 0 and class 1) according to the variable we wished to decode. For example, in the novel-familiar test, social stimulus identity was decoded by grouping firing data around the familiar animal as class 0 [(F, left) and (F, right) conditions] and grouping firing activity around the novel animal as class 1 [(N, left) and (N, right) conditions). We decoded stimulus position by grouping firing activity around the left cup as class 0 [(F, left) and (N, left) conditions] and grouping firing activity around the right cup as class 1 [(F, right) and (N, right) conditions]. For trial decoding (also referred to as XOR), we grouped firing activity around the two cups of trial 1 as class 0 [(F, left) and (N, right)] and trial 2 as class 1 [(F, right) and (N, left)].

Cross-validation and pseudo-simultaneous population activity. For each subject and session, we divided data from each class of conditions (0 and 1) into training and test pseudo-trials, which each trial defined by a bout of interaction, with bout duration lasting from the beginning to end of a given interaction. Bout durations lasting longer than 1 s were split into multiple 1-s-long pseudo-trials. We randomly selected 75% of pseudo-trials for training a classifier and the remaining 25% were used for testing decoding performance. We next constructed a set of pseudo-population activity vectors from the training and testing datasets from a given animal by dividing each pseudo-trial into 100-ms bins, with each bin having its associated population activity vector containing the mean event rate observed during that time bin for each neuron recorded. We then randomly sampled q population vectors (where $q = 5$ unless otherwise noted) from the training data set of each subject and concatenated them to form a single qn-long vector, where n is the total number of recorded neurons in a given subject. This procedure was repeated $T = 2qn$ times to create a training data set of pseudo-population firing rate vectors. We then followed the same procedure to build the pseudo-population testing data vectors, by sampling population vectors from the testing data set of each subject. In some cases we performed decoding analysis on data from all N neurons from all animals tested in a given behavioral task. In this case, we randomly sampled q population vectors from the training data set for each individual animal. Next we concatenated those extended population vectors into one pseudo-simultaneous qN-long vector. We repeated this process sampling successive sets of random population vectors for a total of $T = 2qN$ pseudo-simultaneous training set vectors. We then repeated this process to obtain the testing data set vectors.

To disentangle the selectivity to position and stimulus identity, which are correlated variables, the sampling procedure described above was performed in a balanced way so that each condition within each class (e.g., (F, right) & (F, left) for class 0 and (N, right) & (N, left) for class 1 in decoding stimulus identity) for each subject was equally represented in the training and testing pseudo-simultaneous data set. Only subjects that explored all conditions for a minimum of 3 s each, divided into a minimum of 4 pseudo-trials, were included in the analysis.

The pseudo-simultaneous training data set was then used to train a SVM linear classifier, which was tested on the pseudo-simultaneous testing data set to assess the decoding performance as the fraction of correctly classified pseudo simultaneous vectors. The whole procedure, from training-testing division to performance assessment, was repeated for $k = 20$ times to implement a k-fold cross-validation scheme, taking the mean score (μ_{data}) as the estimated performance value of the decoding procedure.

Null model and p-value. We tested the decoding performance obtained by the cross-validated procedure described above against a null model where the labels (0 and 1 as defined above) of pseudo-trials were randomly shuffled. After each shuffling, the same cross-validation procedure was repeated, obtaining a null-model value for decoding performance. We repeated the shuffling n_{null} times to obtain a distribution of null model performance values, yielding a mean null decoding performance $\langle\mu_{null}\rangle$ and standard deviation of the null distribution σ_{null}. The p value was then derived from the z-score of the performance computed on data compared to the distribution of n_{null} null-model values: $z = \frac{\mu_{data} - \langle\mu_{null}\rangle}{\sigma_{null}}$.

Correlation between decoding performance and behavior. We compared decoding performance versus behavioral preference in the familiar versus novel social recognition test. It was assessed by comparing the two quantities for left (familiar) vs. right (novel) cup in the first half of the littermate recognition test. To compute decoding performance, we used the decoding scheme described above, with the difference that training and testing data was not sampled and concatenated across different subjects to build pseudo-simultaneous population vectors but was analyzed for each animal individually, with $q = 1$. To account for the greater variance of decoding performance across the random training-testing assignment of trials in individuals, we used a 120-fold cross validation scheme. Behavioral performance was computed as the Preference Score computed in the first trial of the littermate recognition test. When specified, the absolute value of the Preference Score was also computed. To exclude animals with a strong left-right preference (independently on the identity of the stimulus animal in the corresponding cup), we computed a left-right preference score as the absolute value of the normalized difference in exploration time for the left vs. the right cup across the whole session: $PS_{lr} = \frac{|t_l - t_r|}{t_l + t_r}$. Subjects with a strong left-right preference, defined as $PS_{lr} > 0.5$, were excluded from the analysis.

Multi-selectivity analysis. We performed the following analysis to assess whether decoding of social and spatial information was primarily driven by cells specialized for one of the two variables (social identity or spatial location of the stimulus). First, for each variable we identified its coding direction as the normalized average decoding weights vector over k cross validations. We denoted these vectors as $\vec{W}^{\text{position}}$, for position decoding, and $\vec{W}^{\text{social}}$ for social familiarity/identity decoding. For each cell i, we then computed a specialization index defined as the absolute value of the normalized difference between the two decoding weights:

$$\sigma_i := \frac{|W_i^{\text{position}} - W_i^{\text{social}}|}{W_i^{\text{position}} + W_i^{\text{social}}}$$

Given a specialization threshold θ_σ, we then computed a population specialization index as the fraction of cells whose specialization was larger than θ_σ. Finally, for each value of θ_σ, we compared the specialization index with a null model that assigned to each cell a random positive value of the two weights by keeping their quadratic sum $\left(W_i^{\text{position}}\right)^2 + \left(W_i^{\text{social}}\right)^2$ conserved – equivalent to a random rotation of the weight vector in the corresponding two-dimensional plane.

Cross-condition generalization performance. Cross-condition generalization performance (CCGP) was computed as described in (27). We first constructed pseudo-simultaneous activity vectors as described above, except we did not group data from pairs of conditions with the same decoding variable. Rather pseudo-trials used for training a given classification came from one of the pairs of conditions that both contained the decoding dichotomy for a given classification while sharing the same non-decoding variable. The corresponding testing set consisted of data from the other pair of conditions that shared the other non-decoding variable. For example, when decoding social identity, one training set consisted of data during interactions with mouse 1 versus mouse 2, when both were in the left cup, and the testing set consisted of data with mouse 1 and mouse 2 in the right cup. The decoding for a given dichotomy was then repeated, swapping the classes of pseudo-trials used for the training and testing data (e.g., training with data obtained with mouse 1 and mouse 2 in the right cup and testing on data with mouse 1 and mouse 2 in the left cup). CCGP was obtained from the mean decoding performance from the two pairs of training and testing conditions. Only animals that showed a decoding performance greater than chance levels for both spatial and social variables (with the threshold p<0.05 computed as described above) were used in the by-subject analysis of CCGP values.

Null model for CCGP. We estimated the null model CCGP as described in (27). To obtain a meaningful null model for generalization performance, it is important to maintain the level of decodability observed experimentally while selectively randomizing generalization between different pairs of conditions. To achieve this, we performed a solid rotation-translation of the pseudo-population vectors sampled from each condition in the neural activity space (using $q = 5$ as described for the decoding analysis) by random shuffling of the neuron index. After the four independent rotations, we computed the CCGP as described above to obtain a null model CCGP value, and repeated this to obtain 20 null model CCGP values. As described in the decoding section, the significance of the CCGP value for the experimental data was computed from its z-score with respect to the population of null model CCGP values.

Comparing decoding performance and CCGP across experiments. To compare the decoding performance or CCGP of the same subject in different experimental paradigms (for example, interacting with the two novel or the two familiar animals), we balanced the subject's behavior so that each of the four conditions had the same interaction time (the minimum) between the two paradigms. If the two sessions had a different number of recorded neurons, say n_{min} and n_{max}, we randomly sub-sampled the session with a larger number of neurons to match the smaller one. The random choice of n_{min} out n_{max} neurons was repeated for each cross-validation (for decoding) or each pseudo-simultaneous data sampling (for CCGP) when decoding the n_{max} session.

Exclusion analysis. To assess whether simple decoding performance or CCGP relies on a set of specialized cells, we ran an exclusion analysis by progressively excluding neurons from the linear classifier based on their ranking through two different metrics:

- Selectivity, defined above as $\sigma_i := \frac{|W_i^{\text{position}} - W_i^{\text{social}}|}{W_i^{\text{position}} + W_i^{\text{social}}}$
- Information, defined as $I_i := W_i^{\text{position}} + W_i^{\text{social}}$

We then ranked the cells according to their Selectivity (σ) or Information (I) scores, and measured CCGP after excluding the top $p\%$ of ranked neurons from the classifier. We denoted these two measures as CCGP(p, σ) and CCGP(p, I), respectively. For each value of p, we then assessed the relative importance of Information and Selectivity by computing the difference between the two scores:

$$\Delta_{I,\sigma}(p) := \text{CCGP}(p, I) - \text{CCGP}(p, \sigma)$$

A negative value of $\Delta_{I,\sigma}(p)$ indicates that, for the purpose of generalization performance, Information is a more relevant feature than Selectivity, as CCGP(p, I) decreases more than CCGP(p, σ) when the top $p\%$ of cells are excluded. Vice versa, a positive value of $\Delta_{I,\sigma}(p)$ indicates that selective neurons are more important, for the purpose of generalization performance, than informative cells. To obtain a single value for each session and individual, we computed the area between the two curves as AUC$(I, \sigma) := \sum_{p=1}^{100} \Delta_{I,\sigma}(p)$. The population of AUC values for each experimental setup that was then tested against a chance level of AUC=0 using a one-sample t-test. The same analysis was also performed for decoding performance.

Geometrical model. In order to test our geometrical interpretation of the experimental data, we developed a statistical model in which increasing degrees of familiarity led to a progressive and continuous change in the geometry of social/spatial representations. The model is composed of a population of N neurons whose firing rate is described by two binary latent variables, corresponding to position and stimulus identity of animals with the same degree of familiarity, reproducing the data from the interaction test with two novel or two familiar animals (Fig.s 5, 6).

In the absence of noise, each of the four conditions of an experiment would be associated with a point in N-dimensional neural firing space. To introduce response variability to the same stimulus, the population firing probability for each condition was described by an isotropic Gaussian distribution with unit variance centered around a condition-specific centroid in the neural firing space.

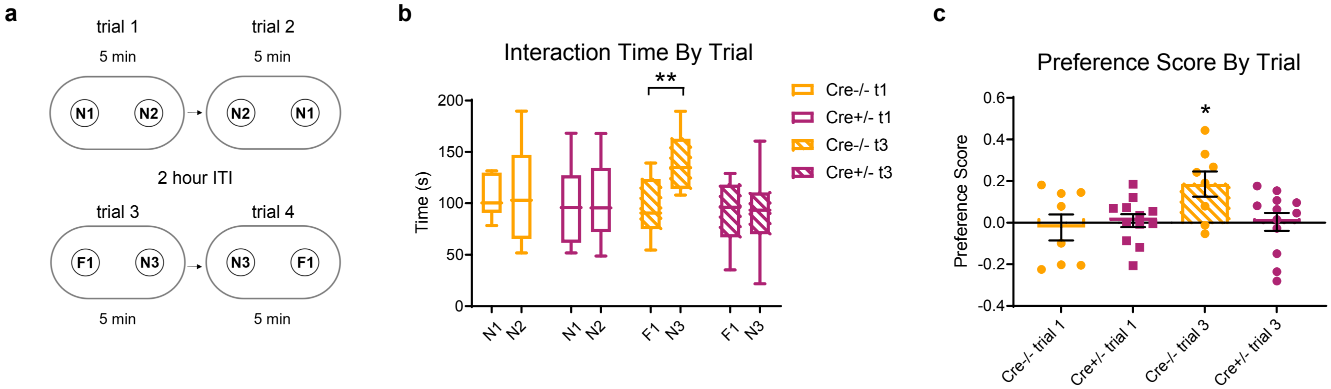
To account for our results during interactions with two novel animals, the means of the four gaussian distributions were arranged so that the two coding directions for the variables were orthogonal – reproducing a low-dimensional, or abstract, representational geometry in the firing space approximated by a two-dimensional rectangle. The length of two arms of the rectangle, denoted as μ^0_{pos} and μ^0_{id}, correspond to the signal-to-noise ratio in the representations of position and social identity variables, respectively, which in turn are reflected in decoding performance.

We accounted for the changes we observed in decoding of familiar compared to novel animals by introducing a familiarity latent variable, denoted as f, in which increasing degrees of familiarity modify the planar, rectangular representation of novel animals as follows.

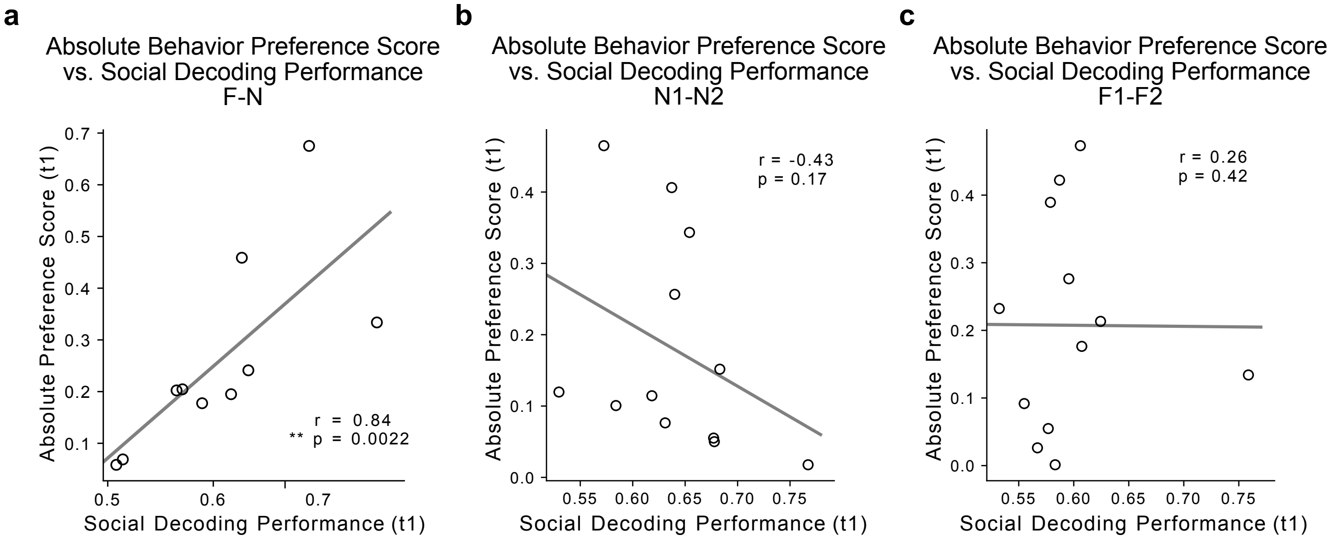
- Reduces signal-to-noise ratio of the identity variable
$$\mu_{\text{id}}(f) = \mu^0_{\text{id}} - \eta f$$
- Performs a global shift by vector length αf along a third coding direction orthogonal to identity and position axes
- Increases the representational dimensionality of the two variables by shifting each of the four condition centroids by a vector of length γf along a random direction for each condition

Using this model, we created simulated data for the activity of N neurons during a set of simulated sessions as a mouse is allowed to interact with two individuals of the same degree of familiarity, f, in left and right cups, with positions swapped in two trials. For each given condition (given mouse in a given cup), we randomly sampled $T = 5000$ N-dimensional point from the distribution in neural activity space for that condition. We then analyzed the simulated data using the same linear decoding and CCGP procedures we used for the experimental data analysis. For each value of f, we repeated the sampling and analysis for $n = 200$ simulated sessions and took the mean for all decoding performance values. We carried out this analysis for a set of values of f ranging from 0 (fully novel) to 1 (completely familiar) at increments of 0.1. For the present analysis we used $N\ = 80,\ \mu_{\mathrm{pos}} = 0.7,\ \mu^0_{\mathrm{id}} = 0.6,\ \eta = 0.5,\ \alpha = 3.0,\ \gamma = 0.06$.

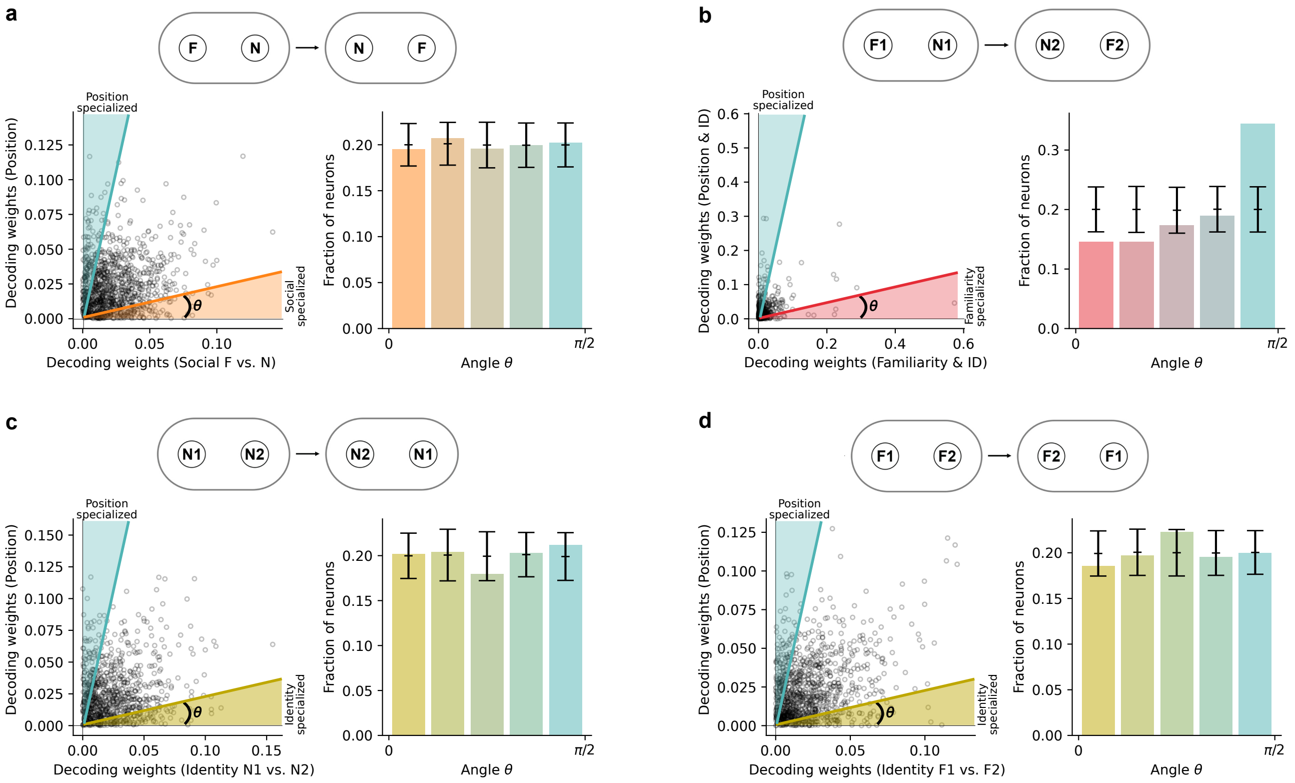
Supplementary Information - Supplementary Figures



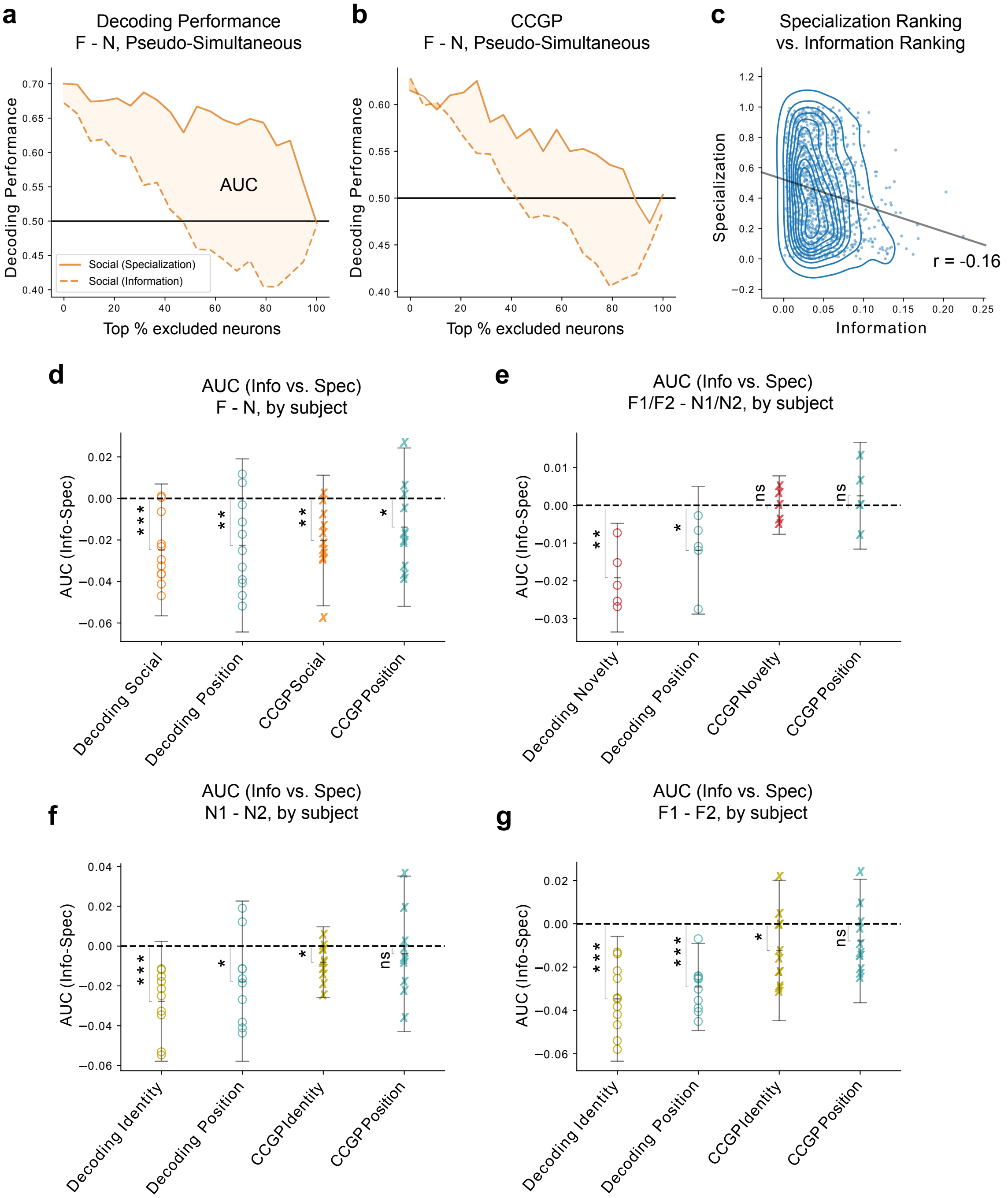
Supplementary Figure 1. dCA2 silencing impairs social memory recognition in a two-choice test in an oval chamber. a) Experimental setup: Amigo2-Cre$^{-/-}$ (control) and Amigo2- Cre$^{+/-}$ mice were injected with Cre-dependent virus to express iDREADD in dCA2. After viral expression both groups were systemically injected with CNO 30-minutes prior to a social memory test, which consisted of two learning trials (Trials 1 and 2) and one recall trial (Trial 3). Trial 1, A subject mouse explored for 5 min two novel stimulus mice (N1 and N2) placed in pencil cup cages at opposite ends of an oval chamber. Trial 2, The positions of the two novel mice were swapped and the subject mouse explored the stimulus mice for an additional 5 min. Trial 3, After a two-hour intertrial interval, one of the two now familiar novel mice (eg N2) was exchanged for a third novel mouse (N3). Memory recall was assessed by the increased time spent exploring the third novel mouse compared to the now-familiar mouse from the previous trials (F1). b) Cre$^{-/-}$ and Cre$^{+/-}$ mice showed similar interactions with N1 and N2 in the two learning trials (only trial 1 data showed here). Cre$^{-/-}$ mice showed expected increased interaction time with the novel compared to familiar individual in trial 3. Cre$^{+/-}$ mice, in which dCA2 was inhibited, did not show a preference for novel over familiar mouse. Two-way repeated-measures ANOVA: Genotype x Interaction Partner F(3,36)=3.624, p=0.022. Šídák's multiple comparisons test. Trial 1 (N1 versus N2): Cre$^{-/-}$ mice, p>0.99; Cre$^{+/-}$ mice p>0.99. Trial 3 (F1 versus N3): Cre$^{-/-}$ mice, p=0.0014; Cre$^{+/-}$ mice, p=0.99. c) Memory performance in indicated trials assessed by preference score: [(time exploring N3) – (time exploring F1)]/[(time exploring N3) + (time exploring F1)]. One-sample t-test against zero. Trial 1: Cre$^{-/-}$ mice, t=0.3680, df=7, p=0.72; Cre$^{+/-}$ mice t=0.3065, df=11, p=0.76. Trial 3: Cre$^{-/-}$ mice, t=3.080, df=7, p=0.018; Cre$^{+/-}$ mice, t=0.09393, df=11, p=0.93. Bars show mean ± SEM. * p<0.05, ** p<0.01.



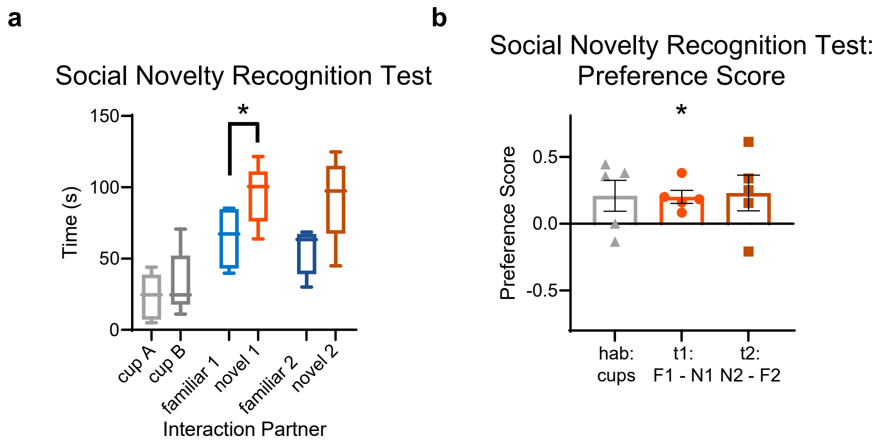
Supplementary Figure 2. Plots of decoding performance compared with subject mouse social memory preference score. a) There is a significant correlation between the absolute value of the preference score and social decoding accuracy. Experiment from Figures 1, 2, 3 (Spearman's correlation r=0.84, p<0.01). b) No significant correlation is observed between absolute preference score and social decoding performance during interactions with two novel mice (Spearman's correlation r =-0.43, p>0.05). c) There was no significant correlation between absolute preference score and social decoding performance during interactions with two familiar littermates (Spearman's correlation r=0.26, p>0.05). ** p<0.01.



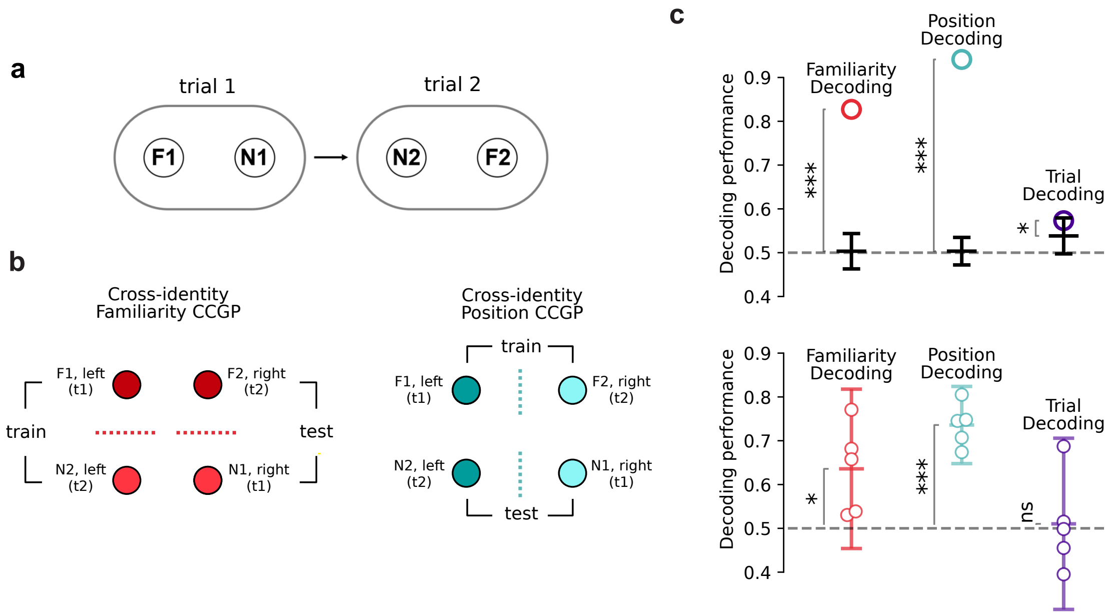
Supplementary Figure 3. Most dCA2 cells have mixed selectivity, encoding information for position, novelty, and identity. a) left: Position decoding versus social decoding weights for each dCA2 neuron (F-N experiments from Figures 1 and 2). Selective neurons lie in a triangular wedge near each axis, defined by an arbitrary angle θ. Right: Histogram of angles representing the population of cells divided into 5 classes of selectivity (bars are color graded corresponding to the degree, or lack of thereof, of selectivity for a given variable). Black bars show mean and two STDs of a null model obtained by assigning a random angle to each point in the left plot. b, c, d) Same analysis of a for the experiment of, respectively: Figure 4 with two different pairs of Novel/Familiar animals in trials 1 and 2; Figure 5, with two novel animals in each trial; Figure 5, with two littermates in each trial. A selectivity distribution consistent with the null model is observed for experiments in a, c, and d. The null model assumes that the two variables are equally decodable; therefore, a one-sided deviation from chance levels as the one observed in b is expected when one variable is more decodable than the other (position in this case, see Suppl. Fig. 6)



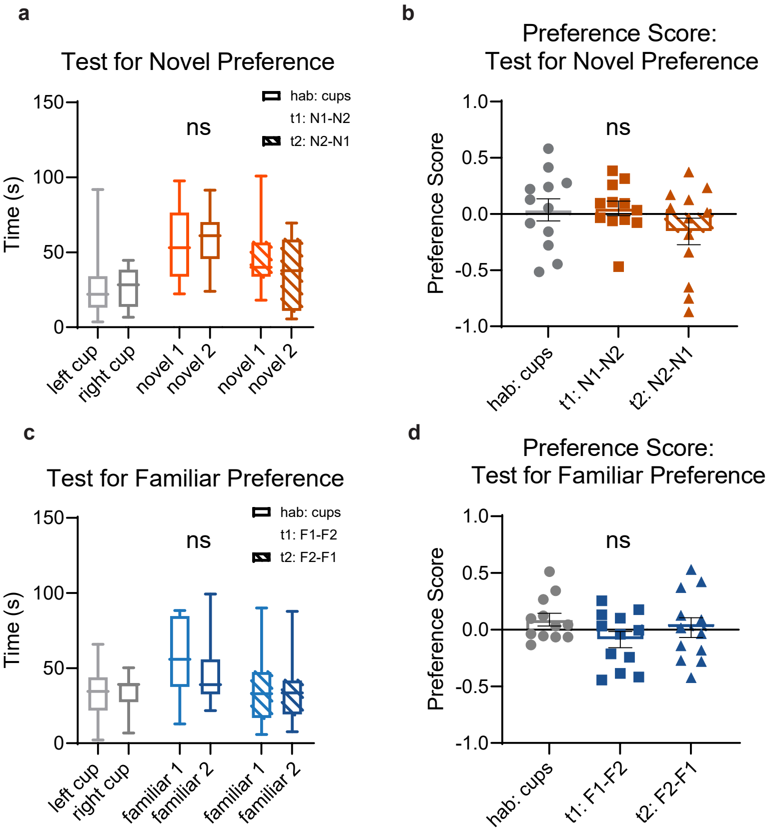
Supplementary Figure 4. Decoding performance relies on cells with high information content rather than high spatial/social selectivity. a) Effect on decoding performance of omitting top fraction of rank-ordered cells from linear classifier. Cells were ranked either by information content (sum of social and spatial decoding weights) or by selectivity (normalized difference in social and spatial decoding weights). The area between the curves (AUC) reveals the greater importance of information content compared to selectivity. Data shown for F-N experiment from Figures 1-2. b) The same analysis for CCGP from Figure 3. c) Selectivity and information rankings from the F-N experiment are moderately negatively correlated, indicating that cells differentially contribute to the two measures (Spearman r=-0.16, p<0.001). d-g) AUC values with significance defined by a one-sample t-test against zero. d) AUC values by subject, for social (social decoding AUC = -0.025 ± 0.016, t=-4.94, p<0.001) and spatial (spatial decoding AUC = -0.023 ± 0.021, t=03.43, p=0.0064) decoding performance and CCGP (social CCGP AUC = -0.020 ± 0.016, t=-4.08, p=0.0022; spatial CCGP AUC = -0.014 ± 0.019, t=-2.29, p=0.045) in F-N experiment of Figures 1-3. e) AUC values for experiment of Figure 4 with a different pair of novel and familiar mice in trials 1 and 2. Statistics: Social AUC (decoding): -0.019 ± 0.007, t=-5.33, p=0.0060; spatial AUC (decoding): -0.012 ± 0.008, t=-2.83, p=0.048; social AUC (CCGP): 0.00 ± 0.004, t=0.04, p=0.97; spatial AUC (CCGP): 0.003 ± 0.007, t=0.72, p=0.51. f) AUC values for experiment of Figure 5 with two novel animals. Statistics: Social AUC (decoding): -0.028 ± 0.015, t=-5.54, p<0.001; spatial AUC (decoding): -0.018± 0.020, t=-2.63, p=0.027; social AUC (CCGP): -0.008 ± 0.009, t=-2.73, p=0.023; spatial AUC (CCGP): -0.004 ± 0.020, t=-0.60, p=0.56. g) AUC values for experiment in Figure 5 with two familiar animals. Statistics: Social AUC (decoding): -0.035 ± 0.014, t=-7.61, p<0.001; spatial AUC (decoding): -0.029 ± 0.010, t=-9.17, p<0.001; social AUC (CCGP): -0.012 ± 0.016, t=-2.41, p=0.037; spatial AUC (CCGP): -0.008 ± 0.014, t=-1.76, p=0.11. ns = non-significant, * p<0.05, ** p<0.01, *** p<0.001.



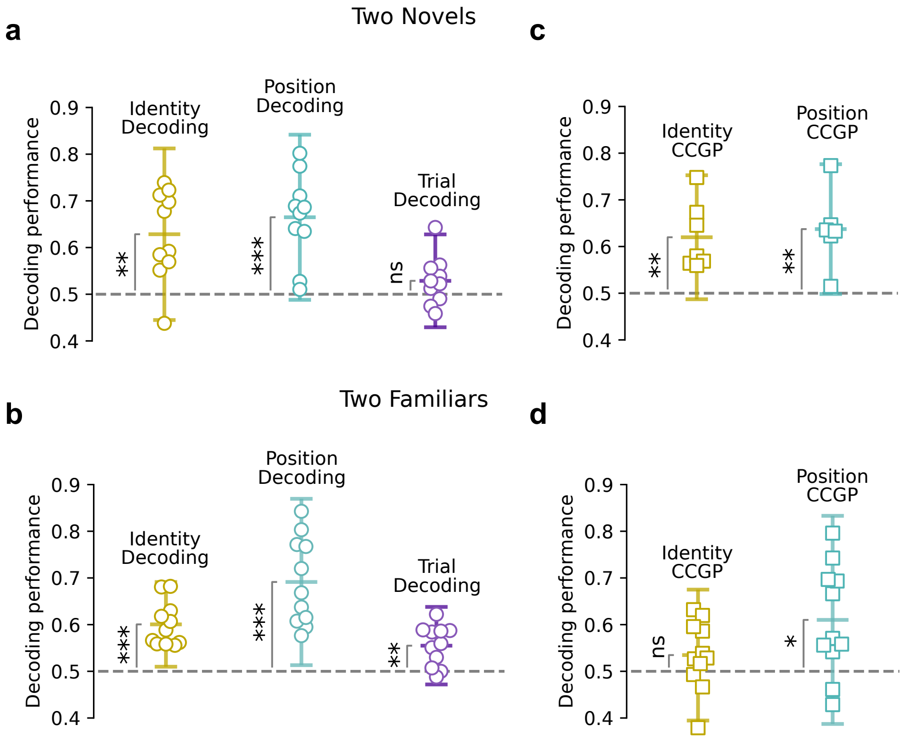
Supplementary Figure 5. Behavior during social novelty recognition test of Figure 4. a) Time spent exploring empty cups in habituation trial and during presentation of two pairs of novel/familiar animals in trials 1 and 2. Two-way repeated-measures ANOVA. Interaction Partner, F(1,4)=22.19, p<0.01; Interaction Partner x Trial, F(1.212, 4.849)=1.100, p>0.05. Šídák's multiple comparisons test: habituation trial (cup A versus cup B), p>0.05; trial 1 (N1 versus F1), p<0.05, trial 2 (N2 versus F2), p>0.05. b) Preference score in three trials. One-sample t-test against zero: habituation trial, t=1.802, df=4, p=0.15; trial 1, t=4.097, df=4, p<0.05; trial 2, t=1.716, df=4, p>0.05. Bars show mean ± SEM. * p<0.05.



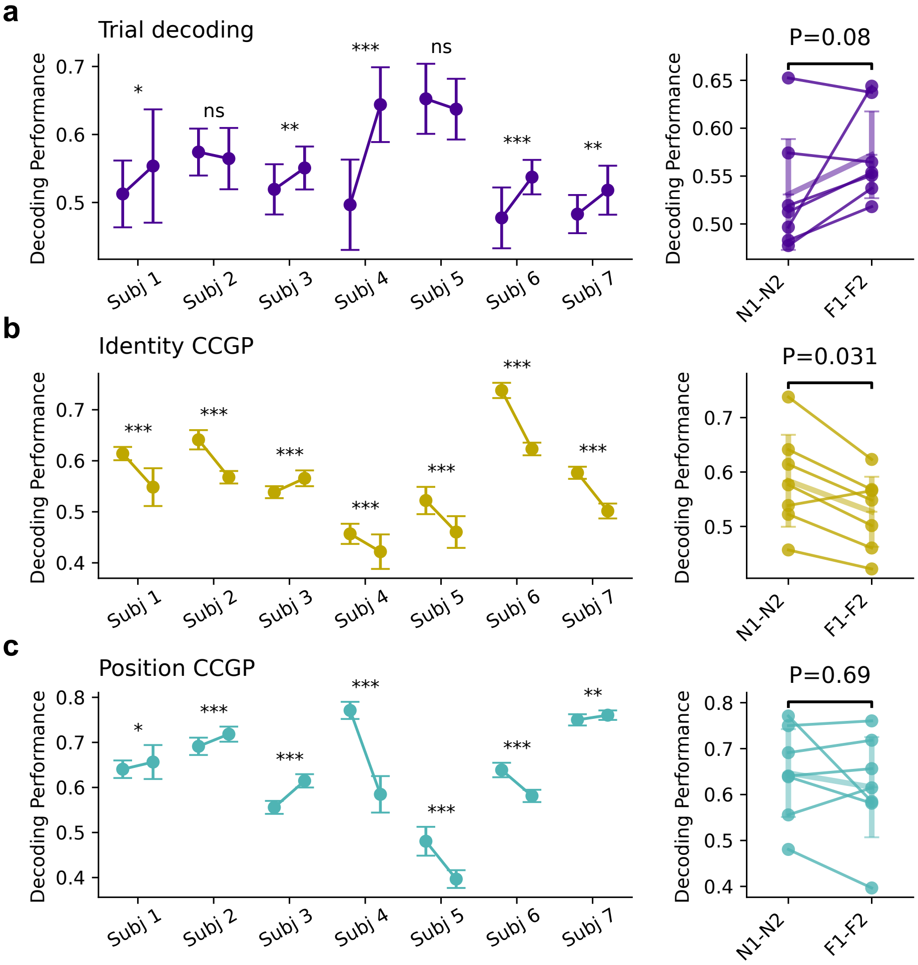
Supplementary Figure 6. Decoding of social novelty/familiarity, position, and trial during experiment of Figure 4. a) Schema of experiment using two pairs of novel/familiar animals in trials 1 and 2. b) Grouping of conditions for decoding social novelty, position, and trial. c, top) Performance of linear classifier for decoding social familiarity (familiarity decoding performance = 0.83, null = 0.50 ± 0.02 [mean ± SD, throughout figure], p<0.001), position (position decoding performance = 0.94, null = 0.50 ± 0.02, p<0.001), and trial (trial decoding performance = 0.57, null = 0.54 ± 0.02, p<0.05) for pseudo-simultaneous dCA2 data. Open circles show average decoding performance from 20 cross-validations. Horizontal line and error bars show mean ± 2SD of distribution of chance values from shuffled data. c, bottom) Performance of linear classifier for same conditions for single subjects (mean familiarity decoding performance = 0.64 ± 0.09, paired t-test against individual chance levels t=2.99, n=5, p<0.05; mean position decoding performance = 0.74 ± 0.04 STD, t=10.71, n=5, p<0.001; trial decoding performance = 0.51 ± 0.10, t=0.21, n=5, p>0.05). Open circles indicate performance of individual subjects. Horizontal lines and error bars show mean ± SD across individual animals. ns = non-significant, * p<0.05, *** p<0.001.



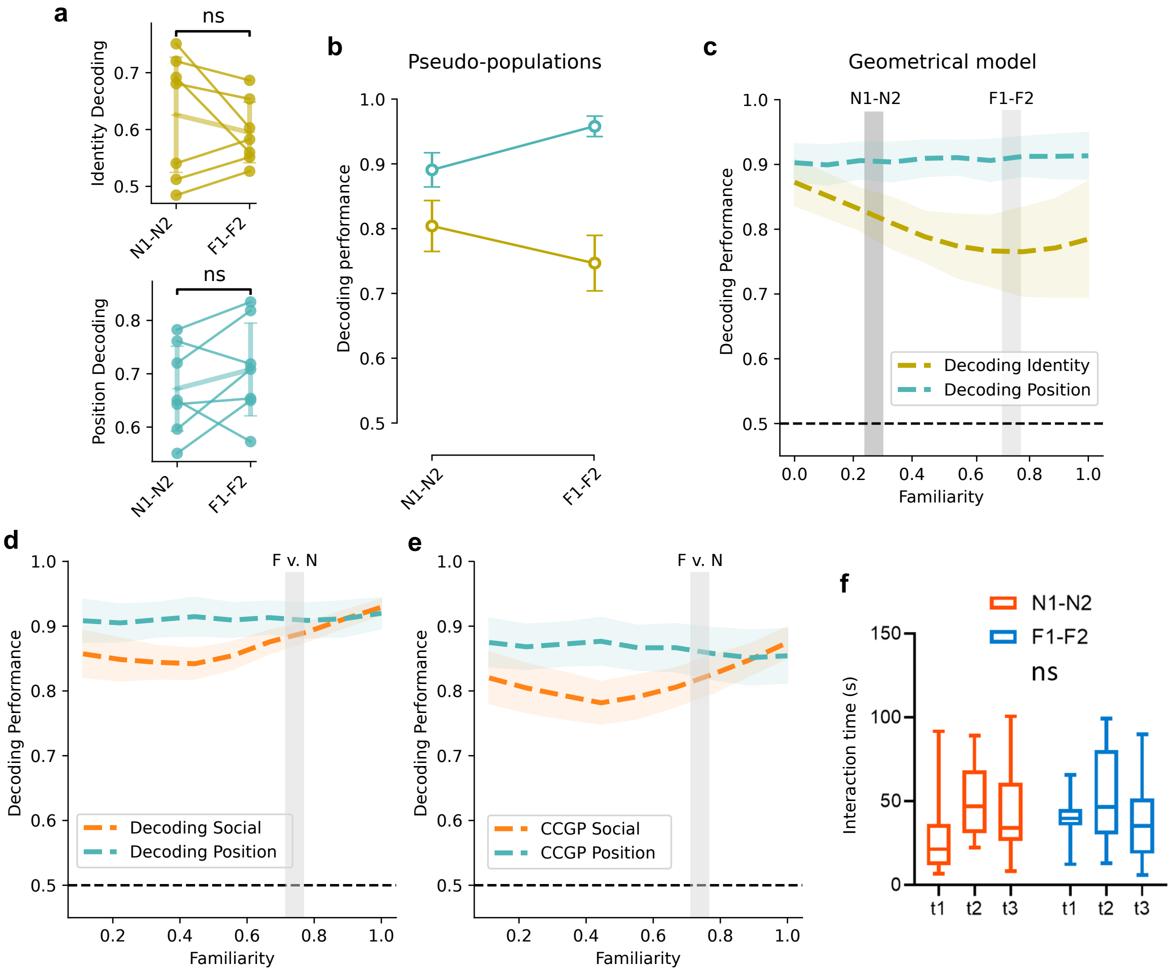
Supplementary Figure 7. Behavioral data associated with tests in Figure 5. a) Interaction durations with empty cups (habituation trial) or cups containing two novel mice in trials 1 or 2 (data for Fig. 5b). No significant difference was observed for any trial: Two-way ANOVA for Partner x Trial F(2,22) = 0.8633, p=0.44. b) Preference scores in three trials are not significantly different from zero. One-sample t-test against zero: habituation trial, t=0.3668, df=11, p=0.72; trial 1 (N1-N2), t=0.7748, df=11, p=0.45; trial 2 (N2-N1), t=1.317, df=11, p=0.21. c) Interaction durations with empty cups or cups containing two littermates (data for Fig. 5c). No significant difference was observed for any trial: Two-way ANOVA for Interaction Partner x Trial F(2,22)=1.079, p=0.36. d) Preference scores in three trials were not significantly different from zero. One-sample t-test against zero, habituation trial, t=1.569, df=11, p=0.15; trial 1 (F1-F2), t=1.249, df=11, p=0.24; trial 2 (F2-F1), t=0.2079, df=11, p=0.84. Bars show mean ± SEM.



Supplementary Figure 8. Decoding performance and CCGP for individual subjects in the two-novel and two-familiar setup (Fig. 5). Throughout figure, open circles (decoding performance) or squares (CCGP values) show performance for individual subjects. Horizontal line and error bars show mean across subjects ± SD. a) Two-novel decoding performance for individual subjects. Decoding of social identity (decoding performance = 0.63 ± 0.09 [mean ± SD, throughout figure], paired t-test against individual chance levels t=4.43, n=10, p<0.01) and position (decoding performance = 0.67 ± 0.09, t=5.65, n=10, p<0.001) was significant. Trial decoding (decoding performance = 0.53 ± 0.05, t=1.65, n=10, p>0.05) was at chance levels. b) Two-familiar decoding performance of identity, position, and trial for individual subjects (identity decoding performance = 0.60 ± 0.05, paired t-test against individual chance levels, t=7.16, n= 11, p<0.001; position decoding performance = 0.69 ± 0.09, t=6.84, n=11, p<0.001; trial decoding performance = 0.56 ± 0.04, t=4.33, n=11, p<0.01). c) Cross-condition generalization performance is significantly greater than chance in classifying two novel identities and left-right position for individuals' data (two-novel identity CCGP = 0.62 ± 0.07, paired t-test against individual chance levels, t=4.67, n=7, p<0.01; position CCGP=0.64 ± 0.07, t=4.61, n=7, p<0.01). d) Two-familiar decoding performance of identity, position, and trial for individual subjects (identity decoding performance = 0.60 ± 0.05, paired t-test against individual chance levels, t=7.16, n= 11, p<0.001; position decoding performance = 0.69 ± 0.09, t=6.84, n=11, p<0.001; trial decoding performance = 0.56 ± 0.04, t=4.33, n=11, p<0.01). d Two-familiar test CCGP values for individual subject data show that identity CCGP is non-significantly different from chance, while position CCGP is significantly higher than chance (two-familiar identity CCGP = 0.54 ± 0.07, paired t-test against individual chance levels t=1.38, n=11, p>0.05; position CCGP=0.61 ± 0.11, t=3.13, n=11, p<0.05). ns, non-significant; *, p<0.05; **, p<0.01; ***, p<0.001.



Supplementary Figure 9. Comparisons of trial decoding performance, identity CCGP, and position CCGP in tests with two novel mice or two familiar mice from Figures 5,6. a) left: by-subject paired comparison of trial decoding performance during interaction with two novel mice (left side of each subject plot) and two familiar mice (right side of each subject plot). Significance levels are computed with a Mann-Whitney U test on n=20 cross validation repetitions (for decoding performance) or n=20 pseudo-population re-sampling (for CCGP). right: summary of left panel showing only mean performance values. Circles indicate single subject performances. Horizontal lines and error bars show mean ± SD. Significance is computed via Wilcoxon signed-rank test across individual subjects. b, c) same analysis of a done for identity and position CCGP, respectively. ns, non-significant; *, p<0.05; **, p<0.01; ***, p<0.001.



Supplementary Figure 10. Comparisons of decoding performances of geometrical model, experimental data and behavior in tests with two novel mice or two familiar mice from Figures 5,6. a) Experimental results for social and spatial decoding performance from dCA2 populations in individual subject mice, same analysis of Suppl. Fig. 9. Shown statistics refer to a Wilcoxon signed-rank test. Circles indicate single subject performances. Horizontal lines and error bars show mean ± SD. b) Experimental results for social and spatial decoding performance from pseudo-simultaneous population data [mean ± SD from 20 cross-validation folds] in tests with two novel (N1-N2) or two familiar (F1-F2) mice. c) Geometrical model results for spatial and social decoding performance of two mice as function of increasing degrees of familiarity (0, completely novel mice; 1.0, familiar littermates). d) Geometrical model results for social and spatial decoding in the test with one familiar and one novel mouse (F-N) from Figures 1, 2, 3. e) Geometrical model results for social and spatial CCGP in the F-N test from Figures 1, 2, 3. Note that the model correctly reproduces that social decoding and position decoding, as well as social and position CCGP, have comparable values in the F-N experiment. f) Interaction times with empty cups in habituation trial and two novel mice (N1-N2) or two familiar mice (F1-F2) in trials 1 and 2 for the experiment shown in Figure 5. No significant difference in interaction was observed between the two tests: two-way repeated-measures ANOVA of Test F(1,13)=0.3093, p=0.59, or Test x Trial F(2,26)=1.663, p=0.21. Shaded areas in c,d,e show SD over 200 simulations. Bars in (f) show mean ± SEM. ns = non-significant.

Supplementary Video 1. $\Delta F/F$ calcium imaging recording of CA2 pyramidal neurons. Recording is 4x speed.

Supplementary Video 2. Animation of data gathered from example subject mouse in the 5-minute experiment corresponding to Figure 1, trial 1, demonstrating subject exploring familiar littermate in left cup and novel conspecific in right cup. Mouse head and body represented by small and large thin black circles around the blue and magenta dots, respectively. Active interaction with conspecifics is demonstrated by black fill within the head and body circles. Left and right cup are represented by wide black circles. Ring around body represents the $\Delta F/F$ activity of a single sample neuron, with a scale from black (no activity) to red (high activity), normalized to trial maximum $\Delta F/F$ value. Grey imprints of the head and body represent the head and body positions at the time of maximum event $\Delta F/F$. Yellow line indicates the path of the subject in the arena. The sample cell demonstrates a high selectivity for calcium events in the presence of the novel conspecific in the right cup over the familiar littermate in the left cup.

Supplementary Video 3. Animation of data gathered from example subject mouse in the 5-minute experiment corresponding to Figure 1, trial 2, with stimulus conspecific positions swapped from Supplementary Movie 2 such that the familiar littermate is in the right cup and novel conspecific in the left cup. Mouse head and body represented by small and large thin black circles around the blue and magenta dots, respectively. Active interaction with conspecifics is demonstrated by black fill within the head and body circles. Left and right cup are represented by wide black circles. Ring around body represents the $\Delta F/F$ activity of a single sample neuron, with a scale from black (no activity) to red (high activity), normalized to trial maximum $\Delta F/F$ value. Grey imprints of the head and body represent the head and body positions at the time of maximum event $\Delta F/F$. Yellow line indicates the path of the subject in the arena. With the positions of the conspecifics reversed, the sample cell now demonstrates little to no selectivity for the novel conspecific in the left cup or familiar conspecific in the right cup. The high selectivity in trial 1 and low selectivity in trial 2 (when positions are reversed) is consistent with a neuron demonstrating mixed selectivity.

Supplementary Information - Memory capacity for disentangled representations

Our goal is to compare memory storage capacity of low- and high-dimensional representations. We assume that a memory of an experience is recollected when the neural circuit is presented with a cue and it can reconstruct the patterns of activity corresponding to the experience stored in memory. This can be implemented with a feed-forward network that essentially implements an autoencoder (see e.g. (39)) or in recurrent neural network like the Hopfield network (36, 37), in which each attractor of the neural dynamics represents one memory (this scenario would be compatible with the anatomy of dCA2, which is known to have recurrent excitatory connections (20)). In both cases, the synaptic weights are chosen in a way that the recollected memory is reconstructed: for the autoencoder the memory is simply reconstructed in the output layer, and for a recurrent network it is reconstructed after relaxation in an attractor. Also, in both cases a partial cue (e.g. a pattern that has a limited overlap with the one stored in memory) will lead to the reconstruction of the full stored memory.

In order to estimate the memory capacity we need to make assumptions about the nature of the memories. For random uncorrelated patterns the memory capacity of the Hopfield model is $p \sim N$: the number of attractors p scales linearly with the number N of neurons. Random patterns are high dimensional, as long as p is not too large (i.e. when $p < N$) and N is large enough, so this is one illustrative and highly representative case of memories that are represented with high dimensional geometries. Real world memories are not random and uncorrelated but it is not unreasonable to consider the random representations if one assume that the brain has a neural circuit that decorrelates the representations (recoding), at least so some extent, before storing them in memory (see for example (39)). This neural circuit could be implemented in the dentate gyrus, which is known to play an important role in pattern separation (42–44) (pattern separation is clearly a form of decorrelation).

In order to estimate the memory capacity we start by considering one possible way of constructing disentangled representations. The representations we now define are not the only possible type of disentangled representations, but they are a representative and illustrative example. Moreover, they have a geometry that is compatible with the observed low dimensional representations. Each pattern is obtained by concatenating L vectors of N_L neurons, each encoding one latent variable Λ_λ, with $\lambda = 1, ..., L$ (e.g. we could assume that $L = 2$ and the first N_L neurons of the full vector encode the position of the animal, and the second N_L neurons encode the identity). For simplicity we assume that each latent variable is encoded by the same number of neurons. All the neurons within each group of N_L neurons have the same activation state, which equal to the value of the latent variable Λ_λ that they encode, and hence they are perfectly correlated. Following (36) we assume that there are only two activation states ±1 for each neuron.

The patterns to be memorized are ξ_i^μ where μ is the memory index, i is the index of the neuron ($i = 1, ..., N$). As discussed above, the patterns are obtained by concatenating vectors that encode different latent variables. Hence $\xi_i^\mu = \Lambda_\lambda^\mu$ for $i = (\lambda-1)N_L+1, ..., (\lambda-1)N_L+N_L$, where Λ_λ^μ is value of the latent variable indexed by λ for memory μ. For example if $L = 2$, the memory μ would have the following form:

$$\xi^\mu = \overbrace{\xi_1^\mu, \xi_2^\mu, ...\xi_{N_L}^\mu}^{N_L}, \overbrace{\xi_{N_L+1}^\mu, \xi_{N_L+2}^\mu, ...\xi_N^\mu}^{N_L} = \overbrace{\Lambda_1^\mu, \Lambda_1^\mu, ...\Lambda_1^\mu}^{N_L}, \overbrace{\Lambda_2^\mu, \Lambda_2^\mu, ...\Lambda_2^\mu}^{N_L}$$

We assume that $\Lambda_\lambda^\mu = \pm 1$ with equal probability. In other words the patterns Λ_λ^μ are random and uncorrelated. This implies that each memory is constructed by choosing randomly each latent variable. This could correspond to a particular episode in which, for example, a certain animal is encountered at a particular location. The identity of the animal and the location are assumed to be random. These representations are low dimensional as their dimensionality is L and L is assumed to be much smaller than N.

We now estimate the memory capacity using a simple signal to noise analysis, as in (36). If the initial state is set by the input, and it is $s_l(t)$, then the state of activation at time $t+1$ of neuron s_k is given by the following expression:

$$s_k(t+1) = \text{sign}\left(\sum_{l=1}^{N} w_{kl} s_l(t)\right)$$

where $k, l = 1, ...N$ and $N = LN_L$ and w_{kl} is the synaptic weight connecting neuron l to neuron k. The argument of the sign function is total synaptic current to neuron k and we call it I_k. We assume that w_{kl} is computed using the Hopfield prescription:

$$w_{kl} = \sum_{\mu=1}^{p} \xi_k^\mu \xi_l^\mu$$

We now focus on the total incoming synaptic current to neuron k:

$$I_k = \sum_{l=1}^{N} w_{kl} s_l(t) = \sum_{l=1}^{N} \sum_{\mu=1}^{p} \xi_k^\mu \xi_l^\mu s_l(t)$$

We consider the case in which a generic pattern is presented, for example memory 1: $s(t) = \xi^1$. In the sum over l, we can now group together all the neurons that encode the same latent variable (they all have the same state of activation) and express the total synaptic current as a function of the Λ variables, which are independent by construction (both with respect to λ and to μ):

$$I_\nu = N_L \left(\Lambda_\nu^1 \sum_{\lambda=1}^{L} \Lambda_\lambda^1 \Lambda_\lambda^1 + \sum_{\mu>1} \sum_{\lambda=1}^{L} \Lambda_\nu^\mu \Lambda_\lambda^\mu \Lambda_\lambda^1 \right)$$

Where Λ is the index of the latent variable encoded by neuron k, the neuron whose state of activation has to be updated. We separated the sum over μ into two parts: the first term reproduces the stored memory (Λ_ν^1) that has to be recollected and hence is usually called (memory) signal. The second accounts for the interference from the other memories, and under the assumption that the values of the latent variables are random and uncorrelated, it is basically just noise. As $\Lambda_\lambda^1 \Lambda_\lambda^1 = 1$, the signal scales like $N_L L$ and the noise term has a variance of approximately $N_L^2 pL$ (there are pL independent terms in the noise). So the signal to noise ratio (SNR) is $L/\sqrt{pL} = \sqrt{L/p}$. This means that the SNR of the memory to be recollected remains large enough, even in the presence of other memories, as long as $p < L$. Hence the maximum number of memories that can be recollected scales as L, the number of latent variables. Notice that N_L cancels out and the max capacity p hence scales as L (it does not depend on the total number of neurons but only on the number of latent variables).

This result is not surprising and it holds also for other learning rules. For example for the pseudo-inverse approach (37, 45, 46) it is clear that the memory capacity scales linearly with the dimensionality of the input patterns, which in our case is L.

Notice that we had to assume that the weights between neurons encoding the same latent variable are all set to zero. Otherwise we have a problem similar to the presence of autapses in the Hopfield model (synapses that connect a neuron with itself): the autapses greatly enhance the stability of the input cue, at the expense of the ability to recall the stored memory (37, 46). By setting all the synapses between neurons encoding the same latent variable to zero, we ensure that the network recollects the memory stored in the synaptic weights and it does not simply reproduce the cue. We neglected the corrections due to these zero weights in the formulae above because they do not change the scaling properties we are interested in when L, N_L and N are large enough.

The simple calculations reported here have only the purpose to illustrate some properties of memory systems storing disentangled representations. It has several limitations: 1) the disentangled representations we considered are not the only possible low dimensional representations, and in particular we should consider representations that are rotated, which would be more similar to those observed in the experiment. In the simple case considered above each neuron encodes only one disentangled variable. 2) it will be interesting to consider representations that are not fully disentangled and have a dimensionality that is intermediate 3) the learning rule is very simple and it is biologically plausible but it doesn't consider the problem of autapses (how does the system set to zero the connections between neurons representing the same latent variable?). On the other hand it seems to be clear that CA2 is not really dealing with these low dimensional representations. The only purpose of the calculations reported here is to show that there is a problem of memory capacity with low dimensional representations and that is probably the reason why they are not used in CA2 to represent familiar animals.