Neuron

Article

Tuned geometries of hippocampal representations meet the computational demands of social memory

Highlights

- CA2 represents novel conspecifics in a low-dimensional geometry, supporting abstraction
- Representations of familiar littermates are high-dimensional, supporting memory storage
- A shift in location of these representations enables abstract detection of familiarity
- The increase in dimensionality predicts behavioral detection of social familiarity

Authors

Lara M. Boyle, Lorenzo Posani, Sarah Irfan, Steven A. Siegelbaum, Stefano Fusi

Correspondence

sas8@cumc.columbia.edu (S.A.S.), sf2237@columbia.edu (S.F.)

In brief

Social memory consists of both familiarity detection and recollection of past social episodes. Whether and how the hippocampus fulfills these roles is unclear. Boyle, Posani, et al. find that the hippocampal CA2 region meets the distinct computational demands of these processes through tuning the geometry of structured neural activity.

Boyle et al., 2024, Neuron *112*, 1358–1371
April 17, 2024
https://doi.org/10.1016/j.neuron.2024.01.021

CellPress

CellPress

Neuron

Article

Tuned geometries of hippocampal representations meet the computational demands of social memory

Lara M. Boyle,[1,7] Lorenzo Posani,[2,3,7] Sarah Irfan,[4] Steven A. Siegelbaum,[1,3,5,6,8,*] and Stefano Fusi[1,2,3,6,8,9,*]

[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, USA

[3]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA

[4]Barnard College, New York, NY 10027, USA

[5]Department of Pharmacology, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY 10032, USA

[6]Kavli Institute for Brain Science, Columbia University, New York, NY 10027, USA

[7]These authors contributed equally

[8]Senior author

[9]Lead contact

*Correspondence: sas8@cumc.columbia.edu (S.A.S.), sf2237@columbia.edu (S.F.)

https://doi.org/10.1016/j.neuron.2024.01.021

SUMMARY

Social memory consists of two processes: the detection of familiar compared with novel conspecifics and the detailed recollection of past social episodes. We investigated the neural bases for these processes using calcium imaging of dorsal CA2 hippocampal pyramidal neurons, known to be important for social memory, during social/spatial encounters with novel conspecifics and familiar littermates. Whereas novel individuals were represented in a low-dimensional geometry that allows for generalization of social identity across different spatial locations and of location across different identities, littermates were represented in a higher-dimensional geometry that supports high-capacity memory storage. Moreover, familiarity was represented in an abstract format, independent of individual identity. The degree to which familiarity increased the dimensionality of CA2 representations for individual mice predicted their performance in a social novelty recognition memory test. Thus, by tuning the geometry of structured neural activity, CA2 is able to meet the demands of distinct social memory processes.

INTRODUCTION

Social memory, an animal's ability to recognize and remember experiences with other individuals of its species (conspecifics), consists of two distinct cognitive processes. As illustrated by the classic example of "the butcher on the bus,"[1] these include the ability to rapidly detect whether an individual is novel or familiar ("I know that person, but from where?") and the more effortful recollection of an individual's specific identity and the associated set of past experiences with that individual ("Ah, she's my butcher; I bought food from her last Tuesday"). These processes have conflicting demands and requirements. Familiarity must generalize to detect whether an individual is novel or familiar across different contexts. By contrast, memory of multi-dimensional social episodes requires distinct representations of past encounters with a given individual at different locations and events. How does the brain manage these conflicting memory requirements of familiarity detection, representation of social identity, and storage of social episodic memories?

Since the early studies of patient HM, it has been clear the hippocampus plays an important role in social memory.[2] However, whether the hippocampus is important for both familiarity and recollection remains controversial.[3,4] Studies in mice have found that the hippocampus,[5] and in particular the dorsal CA2 (dCA2)[6–8] and ventral CA1 (vCA1)[9] regions, are crucial for the storage, consolidation, and recall of social familiarity memory, acting through a dCA2 to vCA1 circuit.[10] Moreover, neurons in dCA2 change their firing to novel conspecifics[11,12] and can distinguish the identity of two novel individuals.[13] Neurons in vCA1 increase their firing to social stimuli[14] and preferentially fire around familiar individuals.[9] Although it is clear that these hippocampal regions encode social stimuli and are required for social memory, whether and how the hippocampus can support a generalized detection of familiarity while also storing a large number of detailed social episodic memories remains unknown. Moreover, because both dCA2[11–13,15,16] and vCA1[17] neurons also serve as place cells, encoding an animal's position in its environment,[18] it is uncertain how the hippocampus represents

and disambiguates social and spatial variables to meet the conflicting demands of social memory.

Here, we applied calcium imaging and computational approaches to identify how the population activity of dCA2 pyramidal neurons enables both a generalized readout of social familiarity compared with social novelty and the encoding of social/spatial episodic memories of familiar individuals. We found that CA2 accommodates the competing demands of familiarity and recollection by representing novel mice and familiar littermates in distinct geometric arrangements (i.e., the relationship between the neural population responses to distinct social/spatial stimuli in neural activity space). CA2 encodes novel animals in low-dimensional representations, which enables the identity of novel animals to be readily disentangled from their position. However, such representations have a low memory capacity and so cannot store the vast amount of information associated with past encounters with familiar animals. CA2 solves this problem by encoding littermates in higher-dimensional representations, increasing memory storage capacity at a modest cost to generalization across contexts. Moreover, by encoding familiarity along a direction in the neural space that is approximately orthogonal to the coding direction of identity, CA2 provides for the generalized or abstract decoding of social novelty. To our knowledge, these results provide the first evidence that transformations in the geometry of neural social representations enable a single neural population to discriminate social novelty from familiarity while supporting the recollection of experiences with highly familiar individuals.

RESULTS

Experimental approach and theoretical considerations of the geometry of social representations and its implications for social memory

Our goal was to characterize how social identity, social familiarity, and spatial location are represented by the activity of CA2 neurons. We used microendoscopic imaging of Ca^{2+} activity in dCA2 pyramidal neurons in freely moving mice as they explored two stimulus mice confined to wire cup cages during two 5-min trials, with the positions of the stimulus mice reversed in the two trials (Figures 1A–1C; Video S1). We then applied computational and theoretical approaches to probe the social and spatial information that was contained in CA2 neural representations.

Previous studies found that certain CA2 neurons act as place cells, responding to an animal's own spatial location,[11–13,15] whereas other CA2 neurons respond primarily to social stimuli.[12,13] Here, we focused on CA2 representations at the population level, using a linear classifier to decode the social and spatial information contained within CA2 activity. Of particular importance, linear classifiers provide insight into the geometry of CA2 social/spatial representations in neural activity space and, thus, can determine whether and how this geometry may differ in the encoding of social/spatial encounters with familiar littermates compared with novel animals.

We performed three recording sessions 1 week apart in which the same subject mouse explored a pair of novel stimulus mice, a pair of familiar littermates, and one novel and one littermate mouse in the different sessions. The subjects showed no preference for exploration of either of the two novel or two familiar individuals in the separate sessions (Figures 1D and S1). CA2 neurons were recorded from the same imaging field but were not aligned across the three sessions. We examined Ca^{2+} signals during periods when the subject mouse was actively exploring one of the two stimulus mice (Figures 1D and 1F; see STAR Methods). The Ca^{2+} signals for each neuron were then deconvolved to extract individual spikes grouped into 100-ms-long bins during interaction periods, labeled according to the identity (e.g., N1 and N2) and position (left, right) of the social encounters. We first describe the results in the sessions with a pair of novel mice and a pair of littermates. Later, we describe results in the session with one novel and one littermate mouse.

CA2 activity accurately encodes social identity and spatial location during exploration of either novel mice or littermates

We first asked whether CA2 activity contained sufficient information to allow for the linear decoding of the identity of the two explored mice, using the combined activity of neurons from the six subject mice as a single pseudo-population. To remove the influence of any spatial information, we grouped in one class the activity recorded around mouse N1 (or L1) in both trials of the task (where mouse 1 was in the left cup in trial 1 and right cup in trial 2) and in a second class the activity recorded around mouse N2 (or L2) in both trials of the task (where mouse 2 was in the right cup in trial 1 and left cup in trial 2), as shown in Figure 2A. Because we balanced the data by including neural activity during equal total lengths of time spent exploring the left and right cups containing the stimulus mice, the two classes differed only in the social identity of the mice. Using a cross-validated scheme, the linear classifier successfully decoded mouse identity during the trials with either the novel mice or the littermates (Figures 2C and 2D, novel animals; Figures 2F and 2G, littermates). Next, we asked whether CA2 activity contained sufficient information to decode position—whether a subject mouse was exploring the left versus right cup—irrespective of the identity of the mouse in the cup. In this case, we grouped activity data from the two trials around the left cup in one class and the right cup in a second class, balancing data so that the subject mouse spent equal time exploring the N1 and N2 (or L1 and L2) mice in the two classes. This removed social identity as a potential confound. Similar to the identity classifier, the position classifier successfully decoded position with high accuracy in both experiments (Figures 3C and 3D, novel animals; Figures 3F and 3G, littermates). The accurate decoding of identity and position is consistent with findings from electrophysiological recordings that CA2 neurons encode both social and spatial information.[9,12,13,15,16]

To explore whether decoding performance depended on CA2 neurons that were specialized for spatial or social information or relied on neurons with mixed social/spatial selectivity,[19] we examined the weights assigned to each neuron by the linear classifiers (Figures 3C and S2). Relatively few neurons had decoding weights specialized for position or identity in either novel or littermate experiments. Neurons that encoded position also typically encoded identity, and vice versa (Figures S2C–S2F; see STAR Methods). When we excluded the few specialized

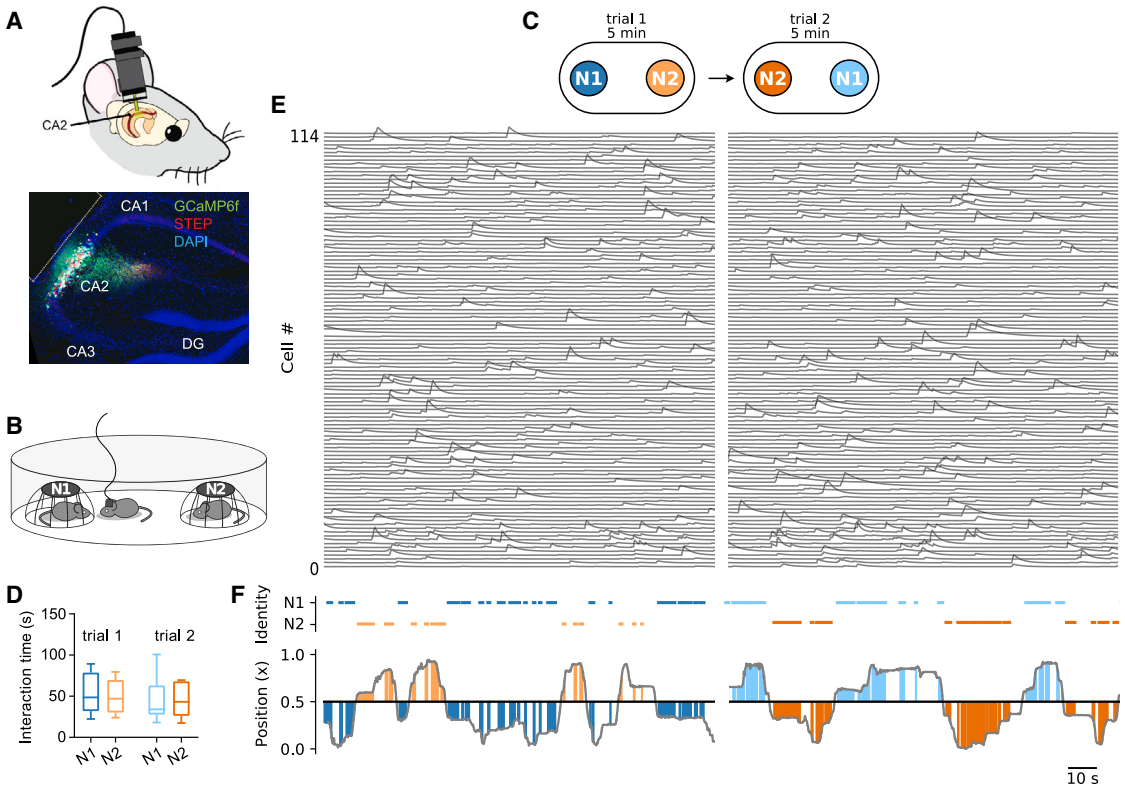

Figure 1. Experimental design

(A) Six Amigo2-Cre subject mice were injected with a Cre-dependent GCaMP6f AAV and implanted with a GRIN lens over dorsal CA2. We imaged a total of 439 and 595 CA2 pyramidal neurons in experiments in which the same subject explored two novel mice and two littermates, respectively, in separate sessions.

(B) Experimental protocol with novel stimulus mice (N1 and N2) under wire cup cages at opposite ends of an oval arena.

(C) Subjects explored the stimulus mice in two 5-min trials, with positions of stimulus mice swapped in each trial.

(D) Mean interaction time of subjects with stimulus mice in two trials. No significant difference was observed for exploration of N1 or N2 in either trial (two-way ANOVA for partner × trial $F(1,5) = 0.0530$, $p = 0.83$).

(E) Deconvolved calcium traces from 114 simultaneously recorded neurons across the two trials from a single subject.

(F) Subject position along axis defined by cup centers and interaction partner identity during trials. Colored lines on top and colored areas denote active interactions (sniffing) with stimulus mice. The colors correspond to the four combinations of spatial (left versus right cup) and social (mouse N1 versus N2) variables.

neurons from the decoding analysis, there was only a small decrease in decoding performance, similar to when we excluded an equal number of mixed-selectivity neurons (Figures S2G and S3H). Moreover, we could decode social identity and spatial location above chance from random subsamples of ~10% of the recorded CA2 population (Figures S2I and S2J), suggesting that CA2 encodes both spatial and social variables in a highly distributed and redundant code through neurons with mixed selectivity.[20]

Representational geometries and their computational implications for social memory

We next asked *how* social and spatial variables are encoded in CA2 population activity and whether the encoding of novel individuals differs from familiar littermates by analyzing the geometry of CA2 representations in neural activity space (Figures 3B–3D). We first consider the different possible geometries of neural representations under the four social/spatial conditions in the two trials of our experiments (mouse 1 and 2 in left and right cups). We then present theoretical results that illustrate the advantages and limitations of how these different geometric arrangements impact the encoding and recall of social/spatial information.

To provide a simplified example, we consider the responses of three hypothetical CA2 neurons in the four conditions across the two trials of an experiment. The positions of these responses in the neural activity space define their representational geometry (Figures 3B–3D). Because of noise, the neural responses to the four conditions form four point clouds. As our experiments have only four conditions, the maximum dimensionality (disregarding noise) is equal to three (3D), regardless of the number of neurons. We refer to such three-dimensional representations as high-dimensional, whereas one- or two-dimensional (1D or 2D) representations are low-dimensional.

Many different geometric arrangements are compatible with our findings that CA2 neurons encode identity and position.[13] However, these geometries have distinct computational properties that affect the ability of a binary linear classifier to read out social/spatial information. We use a binary linear classifier because it has two important advantages compared with nonlinear classifiers. First, as illustrated in Figure 3, it allows us to

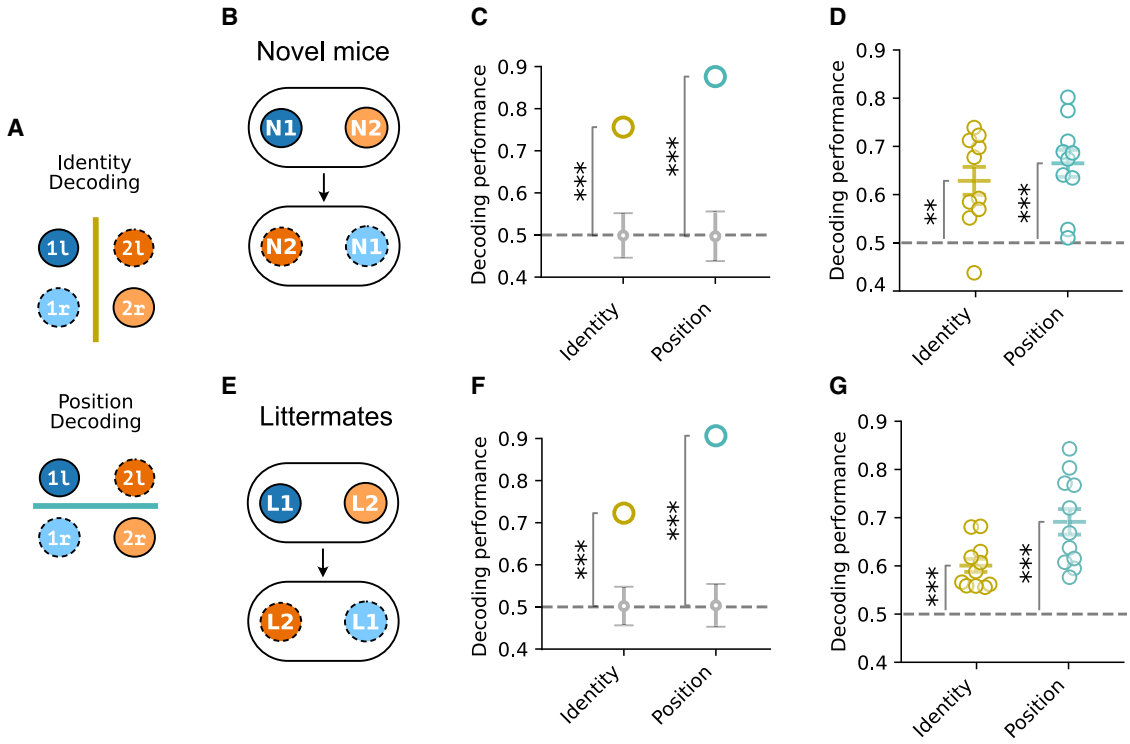

Figure 2. CA2 activity decodes identity and position of novel and littermate mice

(A) Scheme for decoding identity and position. Colored lines separate classes according to identity (1, 2) or position (l, r).

(B) Experiment with two novel mice.

(C) Decoding accuracy of novel mouse identity and position (colored circles) from a pseudo-population of 439 cells from 6 subjects was significantly greater than chance (null model; gray circles). Identity decoding = 0.76; null model = 0.50 ± 0.06. Position decoding = 0.88; null model = 0.50 ± 0.06.

(D) Decoding accuracy for individual subjects from a larger cohort (n = 10; circles). Mean identity decoding = 0.63 ± 0.03; mean position decoding = 0.67 ± 0.09.

(E) Experiment with two littermates.

(F) Littermate identity and position decoding accuracy from pseudo-population of 439 cells from 6 subjects were significantly greater than chance. Identity decoding = 0.72; null model = 0.50 ± 0.06. Position decoding = 0.91; null model = 0.50 ± 0.06.

(G) Decoding data with two littermates from individual subjects. Mean identity decoding = 0.60 ± 0.01; mean position decoding = 0.69 ± 0.03. Values for (D) and (G) are mean ± SEM (error bars). For null model distributions (C and F), values are mean ± SD, error bars show 2 SDs around the mean. p values are estimated from *Z* score of data compared with null models. p values in (D) and (G) are computed by a one-sample t test against a chance decoding accuracy of 50%. **p < 0.01, ***p < 0.001.

determine the geometric structure of the neural representations. Second, it has the neurobiological advantage of being readily implemented by a downstream neuron that linearly sums its inputs to reach a threshold for an action potential output.

Let us consider a simple geometry where CA2 neurons encode a single variable, social identity. In this case, the pairs of point clouds during exploration of the same individual at different locations overlap, defining a 1D object. A linear classifier trained to decode identity from neural activity identifies a plane optimally separating the clouds of points. In this case, the classifier will perform well (Figure 3E), as the cloud of population vectors determined for one identity (e.g., mouse 1, blue clouds in Figure 3B) is well separated from that determined for the other identity (e.g., mouse 2, orange clouds in Figure 3B). Importantly, a classifier trained to report identity when the animals are in the left cup (dark blue versus dark orange clouds in Figure 3B) will be able to generalize to accurately report identity when the animals are in the right cup (light blue versus light orange clouds in Figure 3B). This ability of a decoder trained on one set of conditions to decode a different set of related conditions is termed the cross-condition generalization performance (CCGP).[21] The simple, one-dimensional geometry shown in Figure 3B yields a high decoding performance and a high CCGP for identity (Figure 3E). Following Bernardi et al.,[21] we define this as an abstract representation of identity, as the neural responses to the different social identities do not depend on the position of the social interactions. However, such representations are of limited value for storing or encoding episodic memories, as they do not encode any other variable and, thus, are incompatible with our observations above (see also previous results[13]) that CA2 encodes identity and position.

More realistic population representations that encode both identity and position are shown in Figures 3C and 3D. In Figure 3C (left), identity and position are encoded by specialist neurons, each responsive to only one of the variables. The clouds of activity vectors are arranged in a planar, rectangular-like 2D shape along the two axes corresponding to the firing rates of the two classes of neurons. In this case, separate linear

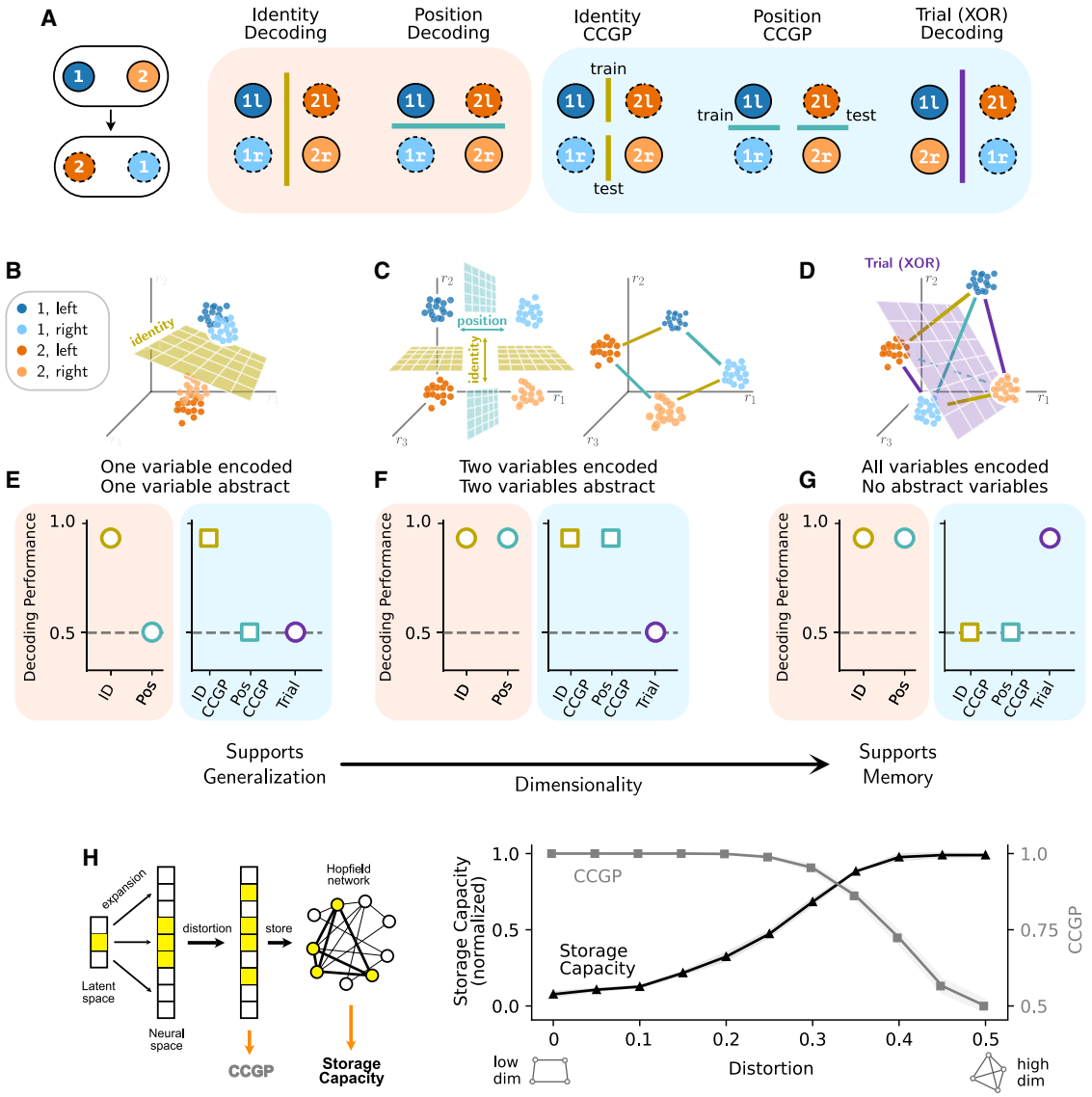

Figure 3. The effects of neural representation geometry on population decoding and memory storage

(A) Decoding and CCGP scheme (see main text and STAR Methods) for identity (ID), position (Pos), and trial (XOR). Mouse 1, blue; mouse 2, orange. Lighter shades, right cup (r); darker shades, left cup (l). Solid outline, trial 1 conditions; dashed outline, trial 2 conditions.

(B–D) Geometric arrangements for the coding of two variables (identity, ID, and position, Pos) with different dimensionality. Points plot firing rates of three neurons (r_1, r_2, and r_3) for a given color-coded condition. Noise results in point clouds. (B) One-dimensional arrangement in which ID but not Pos is encoded. A classifier plane (yellow) separating the point clouds decodes ID. (C) Example two-dimensional geometries in which ID and Pos are encoded and disentangled by specialized (left) or mixed-selective (right) neurons. For both geometries, the same decoding plane trained on one set of conditions decodes the other set of conditions (high CCGP). Pairs of two trials (XOR condition) lie at opposing vertices of the rectangles and cannot be linearly separated. (D) Neurons with a three-dimensional, tetrahedral geometric arrangement showing decoding of XOR (plane separates the two pairs of point clouds grouped by trials).

(E–G) Distinct fingerprints for decoding ID, Pos, and trial (XOR) and CCGP for ID and Pos for geometries shown above in (B)–(D).

(H) Left: scheme of computational model analyzing effects of dimensionality on memory storage capacity and CCGP. A Hopfield recurrent network of N neurons was trained to store and retrieve a set of patterns with geometrical dimensionality varying from L << N to N (see STAR Methods and Figure S3), starting with a low-dimensional representation (0 distortion) and adding increasing extents of random distortion by flipping the state of activation of each neuron in each pattern with a given probability ranging from 0 (L-dimensional) to 0.5 (N-dimensional), step size of 0.05. Right: results of the simulation (L = 10, N = 400). Curves and points show the average over n = 10 simulations. Storage capacity was normalized between 0 and 1.

classifiers can now decode both social identity and position (Figure 3F). Moreover, both variables are abstract with high CCGP (Figure 3F; Video S2). This type of representation is called "disentangled"[21–23] and is important for generalization and compositionality, the capacity to understand and produce a potentially infinite number of novel combinations from known components.

Although representations based on specialized neurons are not compatible with the finding that many neurons exhibit mixed

selectivity (Figure S2), these same computational properties (generalization of both variables) can be maintained with neurons with linear mixed selectivity, i.e., that respond linearly to both identity and position. The activity of such neurons still results in a planar, rectangular-like 2D geometric arrangement, but one that is rotated so that its edges are no longer aligned with the neural axes (Figure 3C, right). Despite the lack of alignment with neural axes, the coding direction for a given variable is still parallel to the coding direction for that same variable across conditions, yielding a high decoding accuracy and CCGP for both identity and position (Figure 3F). Low-dimensional mixed-selectivity representations (Figure 3C, right) have been observed in multiple brain areas.[21,23,24]

A drawback of low-dimensional representations is that they limit the number of variables that can be linearly decoded. For our experiments, the four social/spatial conditions can be grouped into three possible dichotomies according to social identity (mouse 1 in left and right cups versus mouse 2 in left and right cups), position (mouse 1 and 2 in left cup versus mouse 1 and 2 in right cup), and neither identity nor position (mouse 1 in left cup and mouse 2 in right cup versus mouse 1 in right cup and mouse 2 in left cup), which is termed the exclusive OR (XOR) dichotomy and reflects grouping by trial. Neither of the 2D representations in Figure 3 allows a linear classifier to separate the XOR (trial) pairs of point clouds.

In contrast to these low-dimensional (1D or 2D) geometries, the responses illustrated in Figure 3D show the highest-dimensional (3D) tetrahedral representation possible under the four conditions of our experiments. In this geometry, all three possible pairs of sub-groupings of the four conditions, including the trial (XOR) dichotomy, are linearly separable and individually decodable (Figure 3G; Video S3), which defines a representation with a high shattering dimensionality.[21] As both identity and position are decodable by 2D representations, the ability to decode the trial (XOR) dichotomy is diagnostic of higher-dimensional (in our case, 3D) geometries. A disadvantage of high-dimensional representations is a reduced capacity for generalization: the coding directions of each variable are no longer parallel across different conditions, yielding a low CCGP (Figure 3G; Video S4). In summary, whereas low-dimensional geometries provide for generalized decoding of identity/position and high CCGP, high-dimensional geometries allow the decoding of a greater number of variables (including XOR) at the cost of generalization.

The dimensionality of a representation also has important consequences for how a given brain region participates in memory storage, as it is directly related to the number of memories that can be stored in a recurrent neural network. For classical hippocampal-dependent episodic memory, the memory capacity is related to the number of distinct episodes that can be stored. As shown in Figure 3H and further developed with simulations and theoretical computations (see STAR Methods and Figure S3), a low-dimensional geometry severely limits memory storage capacity, where each memory is defined as a specific combination of variables (e.g., the encounter of an individual at a certain location). These limitations occur because low-dimensional representations have a greater correlation between the activity of different neurons compared with high-dimensional representations, effectively reducing the number of independent neurons available to participate in memory storage. Conversely, a high-dimensional geometry provides higher memory capacity at the price of a reduced generalization capacity (Figure 3E). Therefore, different geometries could satisfy different demands for social memory, with dimensionality controlling the tradeoff between generalization and memory storage capacity. As we will see below, the representations for novel animals are similar to the low-dimensional representation of Figure 3C, whereas the representations for littermates are more similar to the high-dimensional representation of Figure 3D.

Novel individuals are encoded in lower-dimensional representations than littermates

We first used principal-component analysis (PCA) to estimate the dimensionality of the CA2 pseudo-population activity during the exploration of novel animals or littermates based on the geometry of the centroids of the four social/spatial experimental conditions when projected in the space spanned by the first three principal components (see STAR Methods). For novel animals, we observed a rectangular-like planar 2D geometry: the coding directions for position were nearly parallel for the two mouse identities, as were the coding directions for identity in the two positions (Figures 4A and 4C–4E; Video S5). By contrast, for littermates, we observed a 3D geometry (Video S6). Although the coding directions of position retained some parallelism (Figures 4B, bottom center and 4C–4E), those of identity were orthogonal (Figure 4B, bottom left), allowing for linear decoding of the trial (XOR) dichotomy (Figure 4B, bottom right; see Figures 3D and 3G).

We next explored the geometry of CA2 social/spatial representations of the single pseudo-population in the original high-dimensional neural activity space using linear classifiers to examine CCGP for social identity and position. We also determined whether we could decode the XOR dichotomy (trial number) as an indirect measure of dimensionality, as discussed above. Identity CCGP was determined by training a linear classifier to decode social identity when the subject was exploring the two stimulus mice in one cup (e.g., left) and testing it on data when the subject was exploring the stimulus mice in the other cup (e.g., right), as shown in Figure 5A, and vice versa, averaging the performance values. Similarly, we measured position CCGP by training a classifier to decode right-left position when the cups were occupied by one mouse (e.g., N1 or L1) and tested on data recorded when the cups were occupied by the other mouse not used for training (e.g., N2 or L2). We found that CA2 neural activity supported generalized decoding of both identity and position when subjects explored novel mice, with a CCGP accuracy significantly greater than chance (Figures 5B and S4C). By contrast, when subjects explored littermates, identity CCGP was not significantly greater than chance. Although position CCGP with littermates was greater than chance, it was smaller than that seen with novel mice (Figures 5C and S4D). When we grouped neural data by trial to determine XOR decoding performance, we found a significant decoding performance during exploration of littermates (Figures 5C and S4B) but not during exploration of novel mice (Figure 5B and S4A). These differences in CCGP values and XOR decoding were statistically significant (Figure 5D) and consistent with the PCA results that littermates

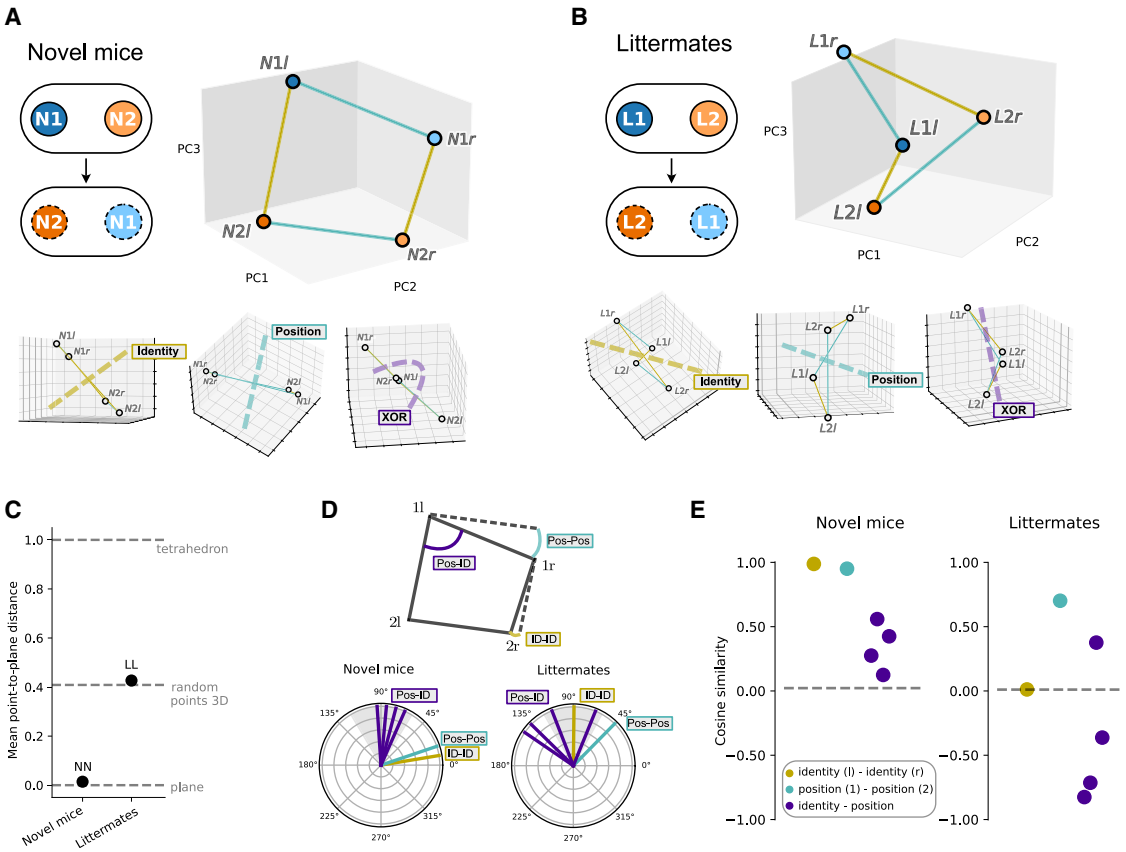

Figure 4. CA2 activity is higher dimensional for social/spatial representations of littermates compared with novel mice

(A) Top: left, experimental scheme with two novel mice. Right, PCA projections along the first 3 principal components of the four social/spatial conditions of the experiment—N1 or N2 in the left (l) or right (r) cups. Pseudo-population activity vectors from 6 mice. Bottom: PC projections from different viewpoints highlight the parallelism of coding directions and the consequent 2D geometry. Dashed lines show how linear planes separate conditions based on identity (left) and position (center) but not XOR (right).

(B) Same analysis as in (A) for the two-littermate experiment. The first three PC projections adopt a three-dimensional geometry that allows for decoding of identity, position, and XOR (bottom left, center, and right). Coding directions for identity are orthogonal (bottom left) so that a decoding plane for identity in the right cup will not necessarily decode identity in the left cup, resulting in a low CCGP for identity.

(C) Mean point-to-plane distance of individual conditions (e.g., N1-left) from the plane spanned by the remaining three conditions (e.g., N1-right, N2-left, and N2-right). The distance is normalized by the mean point-to-point distance between the three remaining conditions such that a perfect tetrahedron will give a mean point-to-plane distance equal to 1. A mean point-to-plane distance equal to 0 indicates a two-dimensional geometry. The geometry for two novel mice is nearly planar (mean point-to-plane distance = 0.01), while the geometry for littermates is compatible with a random sample of four points in three-dimensional space (mean point-to-plane distance = 0.43; random points = 0.41).

(D and E) Angular distance between coding directions of position and identity in the PCA space, visualized in angular degrees (D) and cosine similarity (E). Top panel of (D) shows how angles between coding directions are computed from the representational geometry of the four conditions. A cosine similarity of 1 is equivalent to an angle of 0 degrees, and a cosine similarity of 0 to a 90° angle.

are encoded in higher-dimensional social/spatial representations than novel mice.

To examine how the geometry of CA2 social/spatial representations varied among the six subject mice, we analyzed the decoding performance when linear classifiers were trained on data from each individual subject mouse. Although individual mouse decoding performances were reduced compared with the pseudo-population results, due to the lower number of neurons recorded from the individual mice, five out of six mice showed a significantly smaller identity CCGP ($p = 0.026$, Figure 5E) and a trend for an increased trial (XOR) decoding ($p = 0.055$, Figure 5E) from social/spatial representations of littermates compared with novel mice. By contrast, we did not observe a clear difference in position CCGP. Thus, our single animal and pseudo-population results provide a consistent picture that social/spatial representations during encounters with two novel animals are represented in a lower-dimensional geometry compared with the representations of littermates.

Familiarity is represented as an abstract variable that generalizes across identity and spatial position

To investigate whether CA2 provided an abstract representation of familiarity, we imaged CA2 activity (438 neurons from 5 subject mice) as subjects explored for 5 min an arena with a novel

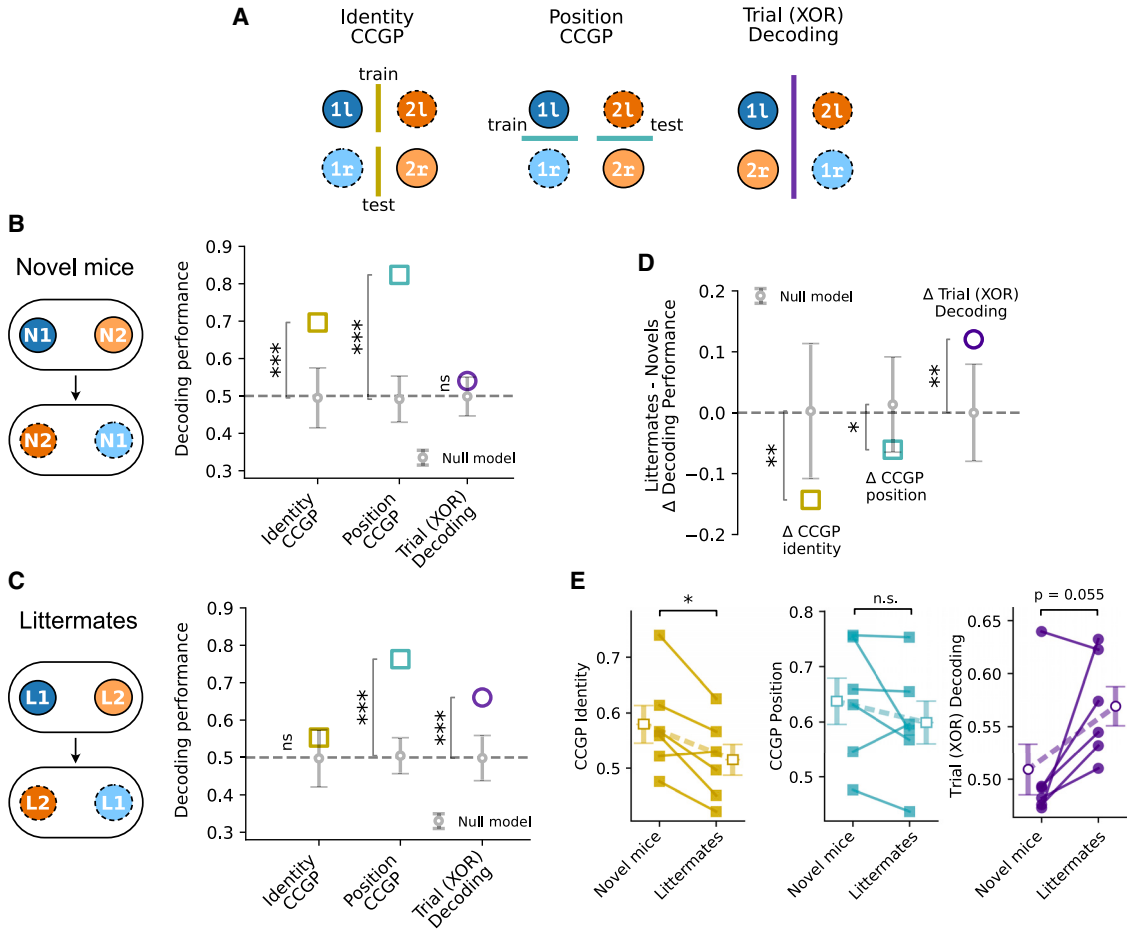

Figure 5. Greater generalization but reduced number of variables decoded for novel compared with littermate mice

(A) Scheme for decoding identity CCGP, position CCGP, and trial (XOR).

(B) Pseudo-population CCGP values for novel mouse identity (0.70) and position (0.82) are significantly greater than null model (0.49 ± 0.04 for both). Trial number (XOR) decoding (0.54) does not differ from null model (0.50 ± 0.06).

(C) For littermates, identity CCGP (0.55) is not significantly greater than null model (0.50 ± 0.05). Position CCGP (0.76) is greater than null model (0.50 ± 0.03). Trial number (XOR) decoding (0.66) is now greater than null model (0.50 ± 0.06).

(D) Difference in indicated decoding performance (Δ) with two littermates compared with two novel mice. Δ identity CCGP = −0.14; null = 0.00 ± 0.06; p = 0.0042. Δ position CCGP = −0.06; null = 0.00 ± 0.04; p = 0.029. Δ XOR = 0.12; null = 0.00 ± 0.04; p = 0.0012.

(E) Identity CCGP, position CCGP, and trial (XOR) decoding results from 6 individual subjects compared during exploration of novel mice and littermates. White-faced symbols and error bars connected by dashed lines show mean ± SE values from 6 mice. Identity CCGP was significantly greater for novel mice (0.58 ± 0.04) compared with littermates (0.51 ± 0.03; p = 0.017; Cohen's d = −1.57). There was no significant difference for position CCGP (novel mice: 0.64 ± 0.05; littermates: 0.60 ± 0.04; p = 0.26; Cohen's d = −0.56). There was a trend for a greater trial (XOR) decoding with littermates (0.57 ± 0.02) compared with novel mice (0.51 ± 0.03) that did not reach significance (p = 0.055; Cohen's d = 1.11). For null model distributions (B and C), values are mean ± SD, error bars show 2 SDs around mean; p values estimated from *Z* score of data compared with null model (B and C) or paired t test (E). *p < 0.05, **p < 0.01, ***p < 0.001.

conspecific and familiar littermate present under the two cups in trial 1 (Figure 6A). In trial 2, subjects explored for 5 min a different pair of novel and littermate mice, with the positions of the novel animal and littermate reversed from trial 1 (Figure 6A). Using a linear classifier, we found that we could decode both familiarity (grouping data around the two littermates in one class and the two novel mice in a second class) and position (grouping data around the left and right cups in the two classes; Figures S5B and S5C).

Next, we used CCGP to determine whether CA2 activity contained an abstract representation of familiarity (Figures 6A and 6B). We trained a classifier to discriminate between a pair of novel and littermate mice that were located in the same cup position in the two trials (e.g., littermate 1 versus novel 2 in the left cup) and tested whether this same classifier could discriminate the different pair of novel and littermate mice in the other cup (e.g., littermate 2 versus novel 1 in the right cup). We reasoned that if CA2 representations provided for an abstract coding of familiarity, we should observe a significant CCGP despite the difference in the specific identity of the novel animals and littermates and their different spatial locations. Remarkably, despite the many variables that changed between the training and testing conditions, CCGP for discrimination of littermates versus novel individuals was high, significantly greater than chance

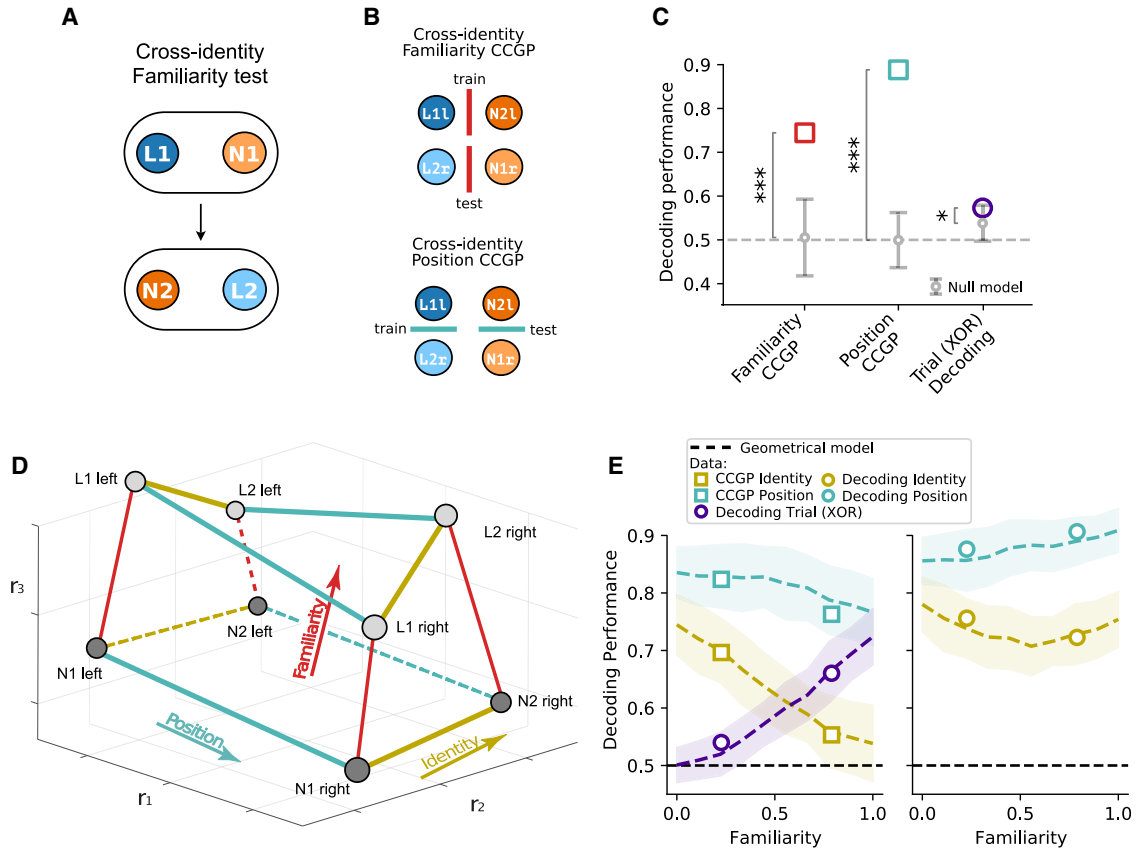

Figure 6. Decoding of familiarity based on tuned geometries of social/spatial representations

(A) Decoding familiarity. Subject mice (n = 5 mice, 438 cells) explored a pair of novel and littermate stimulus mice in a 5-min trial followed by a second 5-min trial with another pair of novel and littermate mice, with positions swapped from the first trial.

(B) Decoding schemes for familiarity and position CCGP.

(C) Familiarity and position CCGP (0.74) were significantly greater than their null models. (Familiarity: null = 0.51 ± 0.04; $p < 0.001$. Position: null = 0.50 ± 0.03; $p < 0.001$), as was XOR decoding (0.57; null model = 0.54 ± 0.02; $p < 0.05$).

(D) Geometrical model for how familiarity alters social/spatial representations, illustrated for three example neurons (firing rates r_1, r_2, and r_3). Dark and light gray circles represent firing rates to specific social and spatial variable combinations during interactions with novel and familiar animals, respectively.

(E) A best fit of the 6 parameters of the model based on the geometry depicted in (D) (see STAR Methods) reproduces our 10 experimental observations (circles and squares). Lines and shaded areas show mean ± SD values calculated from 100 model simulations. p values are estimated from the Z score of the data compared with null model. Null model error bars show 2 SDs around the mean. *$p < 0.05$, ***$p < 0.001$.

(Figures 6C and S5D). Thus, CA2 representations implement an abstract code for familiarity that generalizes across social identities and spatial locations. We also observed a high CCGP for position and a low trial (XOR) decoding performance (Figure 6C), suggesting that the social/spatial representations of the novel and littermate mice form a low-dimensional disentangled geometry.

The high CCGP was not associated with a global difference in CA2 neural activity around the novel compared with littermate mice under the conditions of our experiments (Figures S5E and S5F). Rather, we found that the difference in activity of individual neurons around the pair of novel and littermate mice in one cup was consistently and significantly correlated with the difference in neural activity around the other pair of novel and littermate mice in the other cup (Figure S5G). The fact that the coding direction for familiarity was conserved across the two conditions, despite the change in salient variables such as position and social identity, suggests that familiarity shifts the neural representation of different conspecifics in a common direction in neural space.

Reconstructing the full geometry of social/spatial representations

Based on our findings of CA2 social/spatial representations in the three experiments of Figures 4, 5, and 6, we hypothesized that the process of familiarization caused both a progressive, more-or-less parallel shift in these representations in a specific direction in neural activity space, explaining the abstract representation of familiarity versus novelty (Figure 6C), and a distortion or twisting of the planar representations of novel animals into an increasingly three-dimensional representation of familiar animals (Figure 6D), explaining the observed differences in CCGP and XOR decoding for novel compared with littermate mice (Figures 4 and 5). To test whether this geometric model

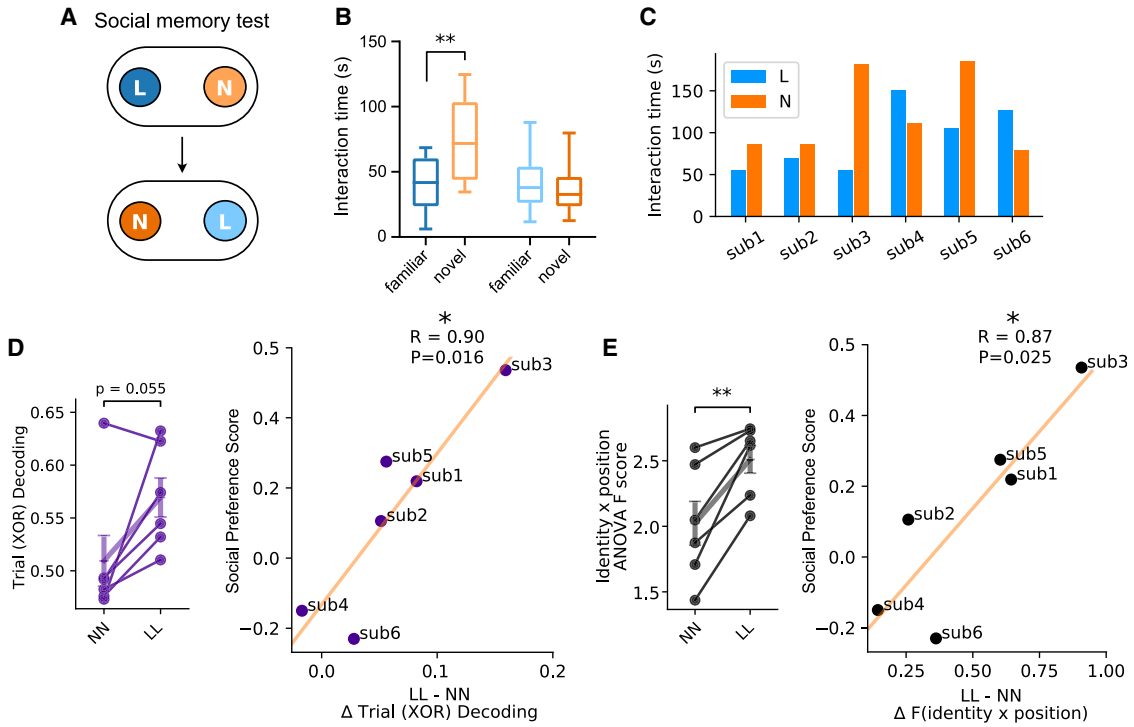

Figure 7. Increase in dimensionality of social/spatial representations of familiar compared with novel mice is correlated with social memory task performance

(A) Two-trial social novelty recognition memory task with novel and littermate mice.

(B) Box-whisker plots showing significantly greater exploration times of littermate and novel mouse in trial 1 (left) but not trial 2 (right). Two-way ANOVA: interaction partner × trial, $F(1,11) = 9.208$, $p = 0.011$. Šídák's multiple comparisons test: trial 1 $p = 0.0085$; trial 2 $p = 0.75$. n = 12 subjects, including the 6 used in Figures 3, 4, 5, and 6 and 6 used for other experiments.

(C) Interaction time with novel and littermate mice combined from the two trials for the 6 mice used for imaging.

(D) Left: data from Figure 5E replotted for comparison purposes showing XOR decoding for the 6 subjects during exploration of two novel (NN) and two littermate (LL) mice. Right: social novelty preference score (y axis, see STAR Methods) was strongly correlated with difference (Δ) between trial (XOR) decoding performance (x axis) during exploration of two littermates versus two novel mice ($r = 0.90$, $p = 0.016$).

(E) Left: single-neuron ANOVA non-linear interaction term (F score) was significantly greater during exploration of two littermates compared with two novel mice ($p = 0.0084$; Student's t test, n = 6 mice). Right: change in F score for interaction term was strongly correlated with social novelty preference ($r = 0.87$, $p = 0.025$). Error bars in (D) and (E) indicate mean and SEM. *$p < 0.05$, **$p < 0.01$.

could provide a quantitative description of our results, we calculated its predicted decoding performance using synthetic data as we increased the degree of familiarity, and thus the resultant representational shift and distortion, as a continuous variable. Despite the limited number of free parameters (6, see STAR Methods), the model provided a good fit to our findings at the two ends of the familiarity continuum we experimentally measured—fully novel compared with fully familiar mice (Figure 6E). This model also generated predictions for changes in activity that might occur with more intermediate degrees of familiarity than our present experiments explored.

The increase in representational dimensionality correlates with social memory behavioral performance

Based on our model in which familiarization produces a concerted shift in the location of social/spatial representations in neural activity space and increases the dimensionality of those representations, we predicted that the ability of a subject mouse to behaviorally discriminate a novel from a familiar mouse should be correlated with the increase in the dimensionality of the representations of familiar compared with novel mice. To test this idea, we ran a standard social novelty recognition memory test in which a subject mouse explored a novel and littermate mouse in cup cages at opposite ends of the oval arena for 5 min in two successive trials, with positions reversed in the trials (Figure 7A). Social novelty recognition memory was manifest as the increased exploration of the novel compared with the littermate mouse, which was more evident during the first presentation of the novel mouse in trial 1 (Figure 7B). We confirmed the importance of CA2 for social memory,[6,10] as its chemogenetic silencing using CA2-selective expression of the hM4Di inhibitory DREADD and systemic injection of the DREADD agonist clozapine N-oxide eliminated the behavioral preference for the novel animal (Figure S6).

As predicted, we found a strong correlation between the behavioral preference for social novelty and the increase in dimensionality of CA2 social/spatial representations as assessed by XOR decoding for littermates compared with novel

mice (R = 0.90, p = 0.016, n = 6; Figure 7D). The correlation was particularly noteworthy as the behavioral and imaging experiments were performed 2–3 weeks apart using distinct sets of novel and familiar mice. By contrast, neither the decoding performance for identity nor position was significantly related to behavior (Figure S7). As an independent measure of the change in dimensionality, we used a linear model ANOVA to determine single-neuron responses to identity, position, and their non-linear interaction (position × identity). We then used the F score for the interaction term, averaged from neurons for individual subjects, as a measure of dimensionality.[22] Consistent with the XOR results, the mean F score for the interaction term was significantly greater during interactions with littermates than novel animals (Figure 7E). Moreover, the increase in F score observed for individual subjects was also strongly correlated with their social novelty preference (R = 0.87, p = 0.025, Figure 7E), further supporting the relationship between the strength of social memory and representational dimensionality.

DISCUSSION

Despite previous findings that CA2 neurons respond to social interactions[12,13,25] and can distinguish novel from familiar animals,[12] it has been unclear whether and how CA2 representations support the discrimination of social novelty versus familiarity while enabling the storage and recall of detailed memories of prior encounters with a familiar individual. Moreover, as CA2 neurons also act as place cells,[11,12,15,16] it has been unclear as to how CA2 representations disentangle social and spatial information. Our results, based on large-scale calcium imaging of CA2 pyramidal neuron activity, indicate that CA2 simultaneously meets the demands of detecting social familiarity and discriminating individual social identity by representing novel animals and littermates in distinct geometries optimized, respectively, for generalization and high memory capacity.

The increase in dimensionality of littermate representations compared with those of novel mice is supported by four independent lines of evidence. First, our PCA results indicated that the (denoised) neural responses to the four conditions of our experiments required only two PC dimensions for novel animals, whereas littermates required three dimensions. Second, based on CCGP measurements, the social/spatial representations of novel individuals provided a greater degree of generalization or abstraction, a hallmark of low-dimensional representations, than those of littermates. Third, social/spatial representations of familiar individuals allowed for greater decoding accuracy of the XOR (or trial) dichotomy, an index of high-dimensional representations. Finally, a linear model ANOVA analysis showed that individual neurons exhibited a greater non-linear (high-dimensional) mixed interaction term (position × identity) for littermates than for novel animals.

Although, in principle, trial decoding could reflect the time difference between the two trials with littermates rather than the XOR dichotomy, a prior study found that time was not linearly decodable from CA2 activity when two littermates were presented in the same position in two trials.[12] Although our finding that a linear classifier failed to decode the XOR condition with novel animal representations provides strong support for a low-dimensional coding geometry, it does not necessarily mean that there is no information about this dichotomy in the recorded neural activity. Indeed, we can decode the XOR condition using a classifier with a non-linear (quadratic) kernel for experiments with both two novel and two littermate mice (Figure S7C). However, the accuracy of the non-linear XOR decoding with novel animals is considerably less than with littermates, qualitatively similar to our results with a linear decoder.

In principle, the global increase in mean CA2 firing rate during exploration of a novel mouse compared with a littermate reported in a prior *in vivo* electrophysiological study from our laboratories[12] could provide a neural mechanism for the generalized encoding of familiarity versus novelty. Although we found a similar global increase in calcium activity during the initial encounter with a novel mouse in our current experiments (data not shown), in the two trials of the full conditions of the experiment of Figure 6, there was no significant increase in global activity around the novel mouse (Figure S5). Thus, our finding of abstract decoding of familiarity cannot be explained by such a global difference in neural activity under the conditions of our experiments. The difference in results between the two studies may reflect differences in experimental protocol (number of presentations of the novel mouse) or differences in recording approach.

We were able to capture our key results in a geometric model in which the identity and spatial location of novel individuals were represented in a low-dimensional geometry and in which familiarization led to a concerted parallel shift in those representations along the familiarity decoding axis combined with a distortion into a higher-dimensional geometry. The parallel shift allows familiarity (compared with novelty) to be encoded in an abstract format, similar to what has recently been observed in the human medial temporal lobe for image recognition.[26] Whether and how these alterations occur at the cellular level remains unknown. The shift may be due to an external signal that encodes familiarity,[27] although plasticity within CA2 might help compute this signal. The increased dimensionality, which allows for increased memory capacity, could result from increased CA2 feedback inhibition (see Treves and Rolls[28]), which decorrelates neural activity. These two transformations were observed in multiple animals and were highly correlated. Of note, we found that the extent of the change in dimensionality of familiar versus novel representations was also highly correlated to a subject's behavioral performance in a social novelty memory test (Figure 7).

How do the different geometries of novel and familiar representations impact the ability of neurons downstream of CA2 to read out social/spatial information? As we noted, one of the attractions of using a linear classifier is that it has a ready neural implementation in downstream neurons integrating CA2 inputs according to a set of synaptic weights corresponding to the neural weights of the classifier. By grouping neural representations of novel and familiar individuals in distinct, linearly separable regions of neural activity space, downstream neurons can be programmed to read out whether a given individual is novel or familiar. Similarly, because novel animals are encoded in low-dimensional social/spatial representations, a downstream neuron can respond reliably to the identity of a novel individual,

independent of that individual's location. By contrast, the higher-dimensional representations of a littermate will ensure that distinct ensembles of downstream neurons differentially encode the memories of distinct social/spatial encounters with that given individual, a key component of episodic memory.

Interestingly, our results show similarities to recent descriptions of representational geometry for familiar and novel faces in the monkey inferotemporal (IT) cortex.[29] Similar to our findings, the representations for novel faces are low dimensional (see Chang et al.[24]). At short latencies, the dimensionality of familiar and unfamiliar face representations is similar, with the two geometries related by a simple translation. By contrast, familiar representations become distorted at longer latencies, though it is not clear whether their dimensionality changes.

To our knowledge, our findings provide the first indication that experience-dependent changes in representational geometry are associated with the behavioral discrimination of novel and familiar individuals. The balance of generalization and memory capacity achieved with these different geometries is likely an important feature that guides the encoding of complex social relationships to form a cognitive map of social space. Such coding may be both a product of and required for navigating complex social behaviors, such as pair bonding, social aggression, and the creation of social dominance hierarchies. Moreover, as abnormalities in social cognition are a hallmark of various psychiatric disorders, it will be of further interest to determine whether deficits in social memory in various mouse models of human genetic conditions linked to neuropsychiatric disease may be reflected in a loss of plasticity in social/spatial representational geometry.

Finally, we note the distortion that we observed for encounters with highly familiar littermates is likely to be an important universal component of efficient memory storage that goes beyond social memory. The idea that episodic memories should be "re-coded" to be stored more efficiently dates back to the studies of David Marr.[30] Re-coding has been the main idea behind the random non-linear transformations proposed in several studies.[28,31–33] These transformations were designed to generate well-separated representations of different memories, a function usually referred to as "pattern separation." Recent models proposed that pattern separation is a signature of a process of memory compression, used by the hippocampus to generate more efficient decorrelated representations.[34–38] In all of these cases, the transformations increase the dimensionality of the representations, similar to what we observed in CA2 activity during littermate interactions. Therefore, the increase of dimensionality we observed with familiarization could be the signature of efficient memory encoding, a mechanism that we predict should be seen in other situations involving different types of memories.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2024.01.021.

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; and 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 and to M.G. Posani for help with graphical illustrations. This work was supported by F30 MH120922-01 (L.M.B.) and grants R01-MH104602 and R01-MH116190 from NIH (PI, S.A.S.). L.P. and S.F. were also supported by NSF Neuronex, the Simons Foundation, the Gatsby Charitable Foundation, and the Swartz Foundation, and S.A.S. was also supported by a grant from the Zegar Family Foundation.

AUTHOR CONTRIBUTIONS

L.M.B. and S.A.S. conceived the project and designed the experiments with input from L.P. and S.F.; 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 were interpreted by L.M.B., L.P., S.A.S., and S.F., who wrote the paper with feedback from S.I.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 16, 2023
Revised: November 3, 2023
Accepted: January 19, 2024
Published: February 20, 2024

REFERENCES

1. Mandler, G. (1980). Recognizing: The judgment of previous occurrence. Psychol. Rev. *87*, 252–271.
2. Milner, B., Corkin, S., and Teuber, H.-L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of H.M. Neuropsychologia *6*, 215–234.
3. Slotnick, S.D. (2013). The nature of recollection in behavior and the brain. Neuroreport *24*, 663–670.
4. Wixted, J.T., and Squire, L.R. (2010). The role of the human hippocampus in familiarity-based and recollection-based recognition memory. Behav. Brain Res. *215*, 197–208.
5. Kogan, J.H., Frankland, P.W., and Silva, A.J. (2000). Long-term memory underlying hippocampus-dependent social recognition in mice. Hippocampus *10*, 47–56.
6. Hitti, F.L., and Siegelbaum, S.A. (2014). The hippocampal CA2 region is essential for social memory. Nature *508*, 88–92.
7. Stevenson, E.L., and Caldwell, H.K. (2014). Lesions to the CA2 region of the hippocampus impair social memory in mice. Eur. J. Neurosci. *40*, 3294–3301.
8. Smith, A.S., Williams Avram, S.K., Cymerblit-Sabba, A., Song, J., and Young, W.S. (2016). Targeted activation of the hippocampal CA2 area strongly enhances social memory. Mol. Psychiatry *21*, 1137–1144.
9. Okuyama, T., Kitamura, T., Roy, D.S., Itohara, S., and Tonegawa, S. (2016). Ventral CA1 neurons store social memory. Science *353*, 1536–1541.
10. Meira, T., Leroy, F., Buss, E.W., Oliva, A., Park, J., and Siegelbaum, S.A. (2018). A hippocampal circuit linking dorsal CA2 to ventral CA1 critical for social memory dynamics. Nat. Commun. *9*, 4163.
11. Alexander, G.M., Farris, S., Pirone, J.R., Zheng, C., Colgin, L.L., and Dudek, S.M. (2016). Social and novel contexts modify hippocampal CA2 representations of space. Nat. Commun. *7*, 10300.
12. Donegan, M.L., Stefanini, F., Meira, T., Gordon, J.A., Fusi, S., and Siegelbaum, S.A. (2020). 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.
13. Oliva, A., Fernández-Ruiz, A., Leroy, F., and Siegelbaum, S.A. (2020). Hippocampal CA2 sharp-wave ripples reactivate and promote social memory. Nature *587*, 264–269.
14. Rao, R.P., von Heimendahl, M., Bahr, V., and Brecht, M. (2019). Neuronal Responses to Conspecifics in the Ventral CA1. Cell Rep. *27*, 3460–3472.e3.
15. Mankin, E.A., Diehl, G.W., Sparks, F.T., Leutgeb, S., and Leutgeb, J.K. (2015). Hippocampal CA2 activity patterns change over time to a larger extent than between spatial contexts. Neuron *85*, 190–201.
16. Lu, L., Igarashi, K.M., Witter, M.P., Moser, E.I., and Moser, M.B. (2015). Topography of Place Maps along the CA3-to-CA2 Axis of the Hippocampus. Neuron *87*, 1078–1092.
17. Jung, M.W., Wiener, S.I., and McNaughton, B.L. (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. J. Neurosci. *14*, 7347–7356.
18. O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. Exp. Neurol. *51*, 78–109.
19. Stefanini, F., Kushnir, L., Jimenez, J.C., Jennings, J.H., Woods, N.I., Stuber, G.D., Kheirbek, M.A., Hen, R., and Fusi, S. (2020). A Distributed Neural Code in the Dentate Gyrus and in CA1. Neuron *107*, 703–716.e704.
20. Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.
21. Bernardi, S., Benna, M.K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C.D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. Cell *183*, 954–967.e21.
22. Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol. *37*, 66–74.
23. Nogueira, R., Rodgers, C.C., Bruno, R.M., and Fusi, S. (2023). The geometry of cortical representations of touch in rodents. Nat. Neurosci. *26* (2), 239–250.
24. Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013–1028.e14.
25. Cymerblit-Sabba, A., Stackmann, M.V.O.P., Williams, A., S., Granovetter, M., Cliz, N., Pereir, F., Smith, A., Song, J., Lee, H., and Young, W. (2020). Recognition memory via repetition suppression in mouse hippocampal dorsal CA2 pyramidal neurons expressing the vasopressin 1b receptor. https://doi.org/10.1101/2020.05.11.078915.
26. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A.N., and Rutishauser, U. (2020). Flexible recruitment of memory-based choice representations by the human medial frontal cortex. Science *368*, eaba3313.
27. Chen, S., He, L., Huang, A.J.Y., Boehringer, R., Robert, V., Wintzer, M.E., Polygalov, D., Weitemier, A.Z., Tao, Y., Gu, M., Middleton, S.J., Namiki, K., Hama, H., Therreau, L., Chevaleyre, V., Hioki, H., Miyawaki, A., Piskorowski, R.A., and McHugh, T.J. (2020). A hypothalmic novelty signal modulates hippocampal memory. Nat. *586*, 270–274.
28. Treves, A., and Rolls, E.T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. Hippocampus *2*, 189–199.
29. She, L., Benna, M., Shi, Y., Fusi, S., and Tsao, D. (2021). The neural code for face memory. Preprint at bioRxiv. https://www.biorxiv.org/content/10.1101/2021.03.12.435023v2.
30. Marr, D. (1971). Simple memory: a theory for archicortex. Philos. Trans. R. Soc. Lond. B Biol. Sci. *262*, 23–81.
31. McNaughton, B.L., and Morris, R.G.M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. Trends Neurosci. *10*, 408–415.
32. O'Reilly, R.C., and McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. Hippocampus *4*, 661–682.
33. McClelland, J.L., and Goddard, N.H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. Hippocampus *6*, 654–665.
34. Gluck, M.A., and Myers, C.E. (1993). Hippocampal mediation of stimulus representation: a computational theory. Hippocampus *3*, 491–516.
35. Benna, M.K., and Fusi, S. (2021). Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. Proc. Natl. Acad. Sci. USA *118*, e2018422118.
36. Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., and Norman, K.A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160049.
37. Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. Cell *183*, 1249–1263.e23.
38. Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., and Shea-Brown, E. (2021). Predictive learning as a network mechanism for extracting low-dimensional latent space representations. Nat. Commun. *12*, 1417.
39. Krashes, M.J., Koda, S., Ye, C., Rogan, S.C., Adams, A.C., Cusher, D.S., Maratos-Flier, E., Roth, B.L., and Lowell, B.B. (2011). Rapid, reversible activation of AgRP neurons drives feeding behavior in mice. J. Clin. Invest. *121*, 1424–1428.
40. Chen, T.W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature *499*, 295–300.

41. Giovannucci, A., Friedrich, J., Gunn, P., Kalfon, J., Brown, B.L., Koay, S.A., Taxidis, J., Najafi, F., Gauthier, J.L., Zhou, P., et al. (2019). CaImAn an open source tool for scalable calcium imaging data analysis. eLife *8*, e38173.

42. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. *21*, 1281–1289.

43. Zhou, P., Resendez, S.L., Rodriguez-Romaguera, J., Jimenez, J.C., Neufeld, S.Q., Giovannucci, A., Friedrich, J., Pnevmatikakis, E.A., Stuber, G.D., Hen, R., et al. (2018). Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. eLife *7*, e28728.

44. Rupprecht, P., Carta, S., Hoffmann, A., Echizen, M., Blot, A., Kwan, A.C., Dan, Y., Hofer, S.B., Kitamura, K., Helmchen, F., and Friedrich, R.W. (2021). A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. Nat. Neurosci. *24*, 1324–1337.

45. Pettit, N.L., Yuan, X.C., and Harvey, C.D. (2022). Hippocampal place codes are gated by behavioral engagement. Nat. Neurosci. *25*, 561–566.

46. Schuette, P.J., Reis, F.M.C.V., Maesta-Pereira, S., Chakerian, M., Torossian, A., Blair, G.J., Wang, W., Blair, H.T., Fanselow, M.S., Kao, J.C., and Adhikari, A. (2020). Long-Term Characterization of Hippocampal Remapping during Contextual Fear Acquisition and Extinction. J. Neurosci. *40*, 8329–8342.

47. Pedregosa, F., G., V., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

48. O'Neill, P.K., Posani, L., Meszaros, J., Warren, P., Schoonover, C.E., Fink, A.J.P., Fusi, S., and Salzman, C.D. (2023). The representational geometry of emotional states in basolateral amygdala. https://doi.org/10.1101/2023.09.23.558668.

49. Amit, D. (1989). Modeling Brain Function: the World of Attractor Neural Networks (Cambridge University Press).

50. Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558.

51. Okamoto, K., and Ikegaya, Y. (2019). Recurrent connections between CA2 pyramidal cells. Hippocampus *29*, 305–312.

52. Allegra, M., Posani, L., Gómez-Ocádiz, R., and Schmidt-Hieber, C. (2020). Differential Relation between Neuronal and Behavioral Discrimination during Hippocampal Memory Encoding. Neuron *108*, 1103–1112.e6.

53. Personnaz, L., Guyon, I., and Dreyfus, G. (1985). Information storage and retrieval in spin-glass like neural networks. J. Phyique. Lett. *46*, 359–365.

54. Kanter, I., and Sompolinsky, H. (1987). Associative recall of memory without errors. Phys. Rev. A Gen. Phys. *35*, 380–392.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit anti-PCP4	Sigma-Aldrich	#HPA005792; RRID:AB_1855086
Mouse anti-STEP (IgG1)	Cell Signaling Technology	#4396; RRID:AB_1904101
Bacterial and virus strains		
AAV2/5-hSyn-DIO-hM4D(Gi)-mCherry	Krashes et al.[39]	Addgene AAV2/5; 44362-AAV2/5
AAV2/1.syn.FLEX.GcaMP6f.WPRE.SV40	Chen et al.[40]	Addgene AAV1; 100837-AAV1
Experimental models: Organisms/strains		
Amigo2-Cre mice	The Jackson Laboratory	Cat# 030215; RRID:IMSR_JAX:030215
Software and algorithms		
MATLAB	Mathworks	https://www.mathworks.com/products/matlab.html
Inscopix Data Acquisition & Analysis Software	Inscopix	https://www.inscopix.com/
CaImAn	Giovannucci et al.[41]	https://github.com/flatironinstitute/CaImAn
ANY-maze	Stoelting Co.	https://www.any-maze.com/
DeepLabCut	Mathis et al.[42]	http://www.mousemotorlab.org/deeplabcut
Prism	Graphpad	https://www.graphpad.com/scientific-software/prism/
Decodanda (Decoding and Dimensionality analysis)	Github	https://github.com/lposani/decodanda
Custom Code (Calcium Imaging Preprocessing)	Github	https://github.com/Lboyle91/BoylePosani_Neuron
Python	Python	https://www.python.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Stefano Fusi (sf2237@columbia.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data reported in this paper will be shared by the lead contact with reasonable request.
- All original code has been deposited on GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Mice

Amigo2-Cre$^{+/-}$ and Cre$^{-/-}$ mice were housed with littermates, and kept on a 12-hour light-dark cycle in air-filtered, temperature- and humidity-controlled conditions with food and water available ad libitum.

METHOD DETAILS

Viral injection and GRIN lens implantation

Calcium imaging

A volume of 200 nL AAV2/1.syn.FLEX.GcaMP6f.WPRE.SV40 virus (titer: 6.5 x 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$^{+/-}$ male 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 (1.9x10^{12} pp/mL) was injected into dCA2 bilaterally using the following coordinates: anteroposterior (AP) -2.0mm, mediolateral (ML) +/-1.8mm, dorsoventral (DV) -1.7mm.

Immunofluorescent Labeling & Imaging

We perfused mice at the end of the experiments using saline followed by 4% PFA in ice-cold PBS. Brains were extracted and incubated in 4% PFA overnight. Brains were sliced in coronal orientation with thickness of 60 μm using a Leica VT1000S vibratome. Sections were permeabilized and blocked for 1 hour with 5% goat serum and 0.4% Triton-X in PBS at room temperature. Sections were incubated overnight with a CA2 marker primary antibody, either pcp4 (1:300, rabbit anti-pcp4 #HPA005792, Sigma-Aldrich) or STEP (1:1000, mouse anti-STEP # 4396, Cell Signaling Technology) at 4°C in 0.1% Triton-X in PBS plus 5% goat serum. The following day, slices were washed with PBS three times for 10 minutes in PBS and incubated with secondary antibodies (respectively: 1:500 goat anti-rabbit IgG, Life Technologies, or 1:500 goat anti-mouse IgG1, Life Technologies) for three hours. Slices were again washed three times in PBS for 10 minutes/wash. DAPI (ThermoFisher Scientific, #D1306) staining was applied at 1:1000 for 15 minutes in PBS at room temperature prior to mounting. Slices were mounted using Fluoromount (Sigma-Aldrich) and imaged using Zeiss LSM 700 confocal microscope.

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 software 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,[43] 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). OASIS is embedded into the CNMFe algorithm,[43] and has been validated and used in previous published reports for spike estimation in hippocampal pyramidal neurons.[44–46]

Behavior

Calcium recordings

We imaged dCA2 pyramidal neurons in a total of 15 Amigo2-Cre heterozygous male mice 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 reintroduced 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 (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 2x the cup radius (10 cm), 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.

Interaction with mice with similar degrees of novelty or familiarity

Six subject mice (n=439 neurons) 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 (Figure 1C). The same six subject mice (n=595 neurons) 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 (Figure 2E). Subsequent tests were run at least one week apart. The neurons for each subject were not registered across sessions.

Social Novelty Recognition Test

Five subject mice (neuron n=438) 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 (Figure 7A).

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 (Figure 6A). Of these twelve subject mice, six were additionally run in the two-novels and two-littermates tests as described above.

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.

Thirty-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 a memory recall trial: trial 3, one of the previously encountered mice in the learning trials (e.g. novel 1, now familiar 1) and a third previously unencountered novel mouse (novel 3). 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 to determine social memory above.

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavior Statistical Analysis

To determine whether there were significant differences in the interaction times of the subject mouse with different 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 between interaction partners. Statistical significance was defined as $p < 0.05$.

As a measure of preference for one interaction partner (B) against the other (A), in Figure 7D and 7E, we calculated the social preference score defined as:

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

Where t_A and t_B are the length of time the subject mouse interacted with mouse A and mouse B across both social interaction trials.

Single-neuron analysis

The response of each neuron to the experimental variables was quantified using a linear model approach combined with an ANOVA table to assess the statistical significance of adding individual terms to the linear model. Neural activity was binarized and fitted by a linear model comprised of a cell-specific intercept and three terms: identity (mouse 1 or mouse 2), position (left or right cup) and the interaction term (position x identity), along with a baseline intercept. The F-score of the ANOVA test for adding or removing each of these three terms was taken as a measure of individual response of the neuron to these variables. For analyses that compare the two-novels and two-littermates experiments, data was balanced across the two setups so that each of the four conditions has the same number of exploration time in the two experiments.

PCA analysis

To visualize the four conditions in a reduced-dimensionality space, we first performed a principal component analysis on the data resampled, as described below in the population decoding analysis, so that the pseudo population data contained the same number of samples for each of the four conditions. We then projected the resampled data in the space defined by the first three principal components. For each of the four conditions, we took the median of each position in the PC space as the position of the corresponding centroid. We then used the resulting four centroids to perform our geometrical quantifications as described in Figure 4.

Population decoding analysis

The decoding analysis was performed using a linear classifier based on a support vector machine (SVM) with custom-written Python scripts based on the scikit-learn SVC package.[47]

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 *#1* or *#2*). 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 [mouse 1 on left (#1-left), mouse 1 on right (#1,-right), mouse 2 on left (#2-left), mouse 2 on right (#2-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, *social stimulus identity* was decoded by grouping firing data around the familiar animal as class 0 (#1-left & #1-right) and grouping activity around the novel animal as class 1 (#2-left & #2-right). We decoded *stimulus position* by grouping activity around the left cup as class 0 (#1-left & #2-left) and grouping firing activity around the right cup as class 1 (#1-right & #2-right). For XOR decoding, we grouped together conditions that have no identity or position values in common, defining two classes that incidentally correspond to trial 1 and trial 2 of our experimental setup: (#1-left & #2-right) as class 0 and (#1-right & #2-left) as class 1.

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 was equally represented in the training and testing pseudo-simultaneous data set (e.g., for identity decoding: balancing #1-right and #1-left for class 0 and balancing #2-right and #2-left for class 1). 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. To allow for a meaningful comparison of decoding results across experiments, only subjects that explored the four conditions (#1-left, #1-right, #2-left, #2-right) for a minimum of 3 s each, divided into a minimum of 4 pseudo-trials, in all three experiments (two-novels, two-littermates, novel-littermate) were used in the decoding analysis.

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 ($<\mu_{null}>$) 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: $z = [\mu_{data} - <\mu_{null}>]/\sigma_{null}$.

Mixed selectivity analysis

We performed the following analysis to assess whether cells were specialized for one of two variables (hereby called variable *A* and variable *B*) or whether they encode the variables with mixed selectivity based on decoding weights. For a given variable *X* and each cell *i*, we identified its coding importance, defined as w_i^X, as the absolute value of its average decoding weight normalized by the standard deviation over *k* cross-validation folds[48]:

$$w_i^X := \left| \frac{\mathrm{E}\left[w_{i,n}^X\right]_n}{\sigma^2\left[w_{i,n}^X\right]_n} \right|$$

where $w_{i,n}^X$ is the SVM decoding weight of cell *i* in the *n*th cross-validation fold. We obtained a vector of all the values across the recorded population of cells:

$$\overrightarrow{W}^X = \left(w_1^X, w_2^X, \ldots, w_N^X\right)$$

We denoted these vectors as $\overrightarrow{W}^A$, for variable *A* and $\overrightarrow{W}^B$ for variable *B*. If the recorded population is specialized, neurons that encode *A* will not encode *B*, and vice-versa. Therefore, a population of specialized neurons will be characterized by anti-correlated values of $\overrightarrow{W}^A$ and $\overrightarrow{W}^B$ (Figure S2A). On the other hand, if neurons are not specialized (mixed selectivity), we expect no relationship between *A* and *B*coding, resulting in a null correlation between $\overrightarrow{W}^A$ and $\overrightarrow{W}^B$ (Figure S2B). A third possibility is that neurons are not specialized, but information is unevenly distributed across the population. In this case, neurons will typically encode neither or both variables, resulting in a positive correlation between $\overrightarrow{W}^A$ and $\overrightarrow{W}^B$ (Figure S2).

To assess whether the recorded population was specialized, we computed the Spearman correlation between the two coding importance vectors, denoted as $\rho(\overrightarrow{W}^A, \overrightarrow{W}^B)$. We then compared the value of $\rho(\overrightarrow{W}^A, \overrightarrow{W}^B)$ with those obtained by a null model where mixed selectivity is implemented by performing a solid random rotation of the coding weights of the two variables in the neural activity space. In this null model, the two coding vectors $\overrightarrow{W}^A$ and $\overrightarrow{W}^B$ have no relationship with each other. For each null model iteration *k* we sample a random rotation matrix R_k and use it to rotate the weights vectors of A and B decoding before taking the average across cross-validations. This procedure was repeated 100 times, each time with a new rotation matrix, to obtain a population of null values. The recorded population of cells was then classified as either mixed or selective depending on the significance of the correlation, computed using the z-score of the Spearman correlation compared to the null model:

- $\rho(\overrightarrow{W}^A, \overrightarrow{W}^B) <$ null model: selective population
- $\rho(\overrightarrow{W}^A, \overrightarrow{W}^B) \sim$ null model: mixed selectivity
- $\rho(\overrightarrow{W}^A, \overrightarrow{W}^B) >$ null model: mixed selectivity

Ablation analysis to assess decoding importance of specialized cells

To assess the relative importance of specialized and mixed neurons in the decoding performance for a given variable, we performed an "ablation" analysis where each class of cells (specialized and mixed) is selectively excluded from the data used for the decoding analysis. Say we are testing the decoding importance of two variables, *A* and *B*. First, we identified the n_A specialized cells for variable *A* and the n_B specialized cells for variable *B* as described in the multi-selectivity analysis methods. Those neurons that are not specialized for either A or B were labeled as "mixed." We then excluded all the $n_A + n_B$ neurons that were identified as specialized and performed the decoding analysis for variables *A* and *B* (as described in the decoding methods) to get the decoding performances DP_A^{sel} and DP_B^{sel}. We then randomly selected the same number of mixed neurons by choosing a random value of selectivity lower than the selectivity threshold used to identify specialized neurons, which is equivalent to a random angle in the W_A vs. W_B graph (see Figures S2C and S2E and multi-selectivity methods), and choosing the $n_A + n_B$ neurons that have the selectivity closest to this random value (orange dots in Figures S2C and S2E). We excluded these neurons and repeated the decoding analysis to obtain the decoding performances DP_A^{mix} and DP_B^{mix}. By repeating this random choice procedure 20 times, we obtained a population of decoding values for both variables when excluding mixed selectivity neurons from the decoding analysis (blue error bars in Figures S2G and S2H). Finally, the decoding performances DP_A^{sel} and DP_B^{sel} were compared to the population of DP_A^{mix} and DP_B^{mix} values to assess whether excluding specialized neurons affected decoding performance differently than excluding mixed-selective neurons. Statistical significance for each variable was assessed by computing the z-score of DP^{sel} compared to the corresponding DP^{mix} population.

Cross-condition generalization performance

Cross-condition generalization performance (CCGP) was computed as described in Bernardi et al.[21] 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 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.

We estimated the null model CCGP as described in Bernardi et al.[21] 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 littermates), 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.

Null model for decoding difference

To assess the significance of a difference between two decoding performances μ_A and μ_B, we first obtain a distribution of null model values for both performances as described above. We then created a null model distribution for the difference $\mu_A - \mu_B$ by taking all possible differences between null model values for μ_A and null model values for μ_B. The p value of the difference was then derived from the z-score of the performance difference $\mu_A - \mu_B$ compared to this distribution of differences. Note that the null model distribution of differences has a standard deviation that is approximately the sum of the two standard deviations of individual null model distributions.

Geometrical model

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 animals or two littermates (Figures 2 and 5).

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.

1. Reduces signal-to-noise ratio of the *identity* variable: $\mu_{id}(f) = \mu^0_{id} - \eta f$
2. Performs a global shift by vector length αf along a third coding direction orthogonal to identity and position axes
3. 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 points 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 and standard deviation for all decoding performance values (shown in Figure 4E). 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. We manually selected values of μ_{id}^0, μ_{pos}^0, γ, and η to reproduce the values of CCGP identity, CCGP position, and XOR decoding across the two-littermates and two-novels experiments at two optimized values of f_{NN} and f_{FF}, for a total of 6 fitted parameters to reproduce 10 experimental values: position decoding and CCGP, identity decoding and CCGP, and XOR decoding in the two-novels and two-littermates experiments. For the results shown in Figure 4, we used $N = 80$, $\mu_{pos}^0 = 0.8$, $\mu_{id}^0 = 0.6, \eta = 0.45, \alpha = 3.0$, $\gamma = 0.55$, $f_{NN} = 0.23$, and $f_{FF} = 0.78$.

Tradeoff between memory capacity and generalization in a Hopfield recurrent neural network

To study the trade-off between memory capacity and generalization capacity, we sampled patterns from geometries of different latent dimensionality, and measured (1) how many of these patterns a Hopfield recurrent neural network (RNN) can store and retrieve and (2) how well a linear classifier trained to decode a latent variable from these patterns is able to generalize across values of the other latent variables.

Sampling patterns with varying dimensionality

To obtain patterns of varying dimensionality, we first defined L and N as the minimum and maximum dimensionality, with $N \gg L$. A set of P binary L-dimensional latent patterns $\lambda^\mu, \mu \in [1, P]$ was then obtained by sampling L i.i.d. Bernoulli random variables P times:

$$\lambda^\mu = (\lambda_1^\mu, \lambda_2^\mu, \ldots, \lambda_L^\mu), \lambda_i^\mu \in \{0, 1\}$$

We then expanded each pattern λ^μ into an N-dimensional embedding by repeating each λ_i^μ value N/L times, so that the collection of patterns kept the original dimensionality (L) in the new (N-dimensional) space.

$$\xi^\mu = (\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu) = (\lambda_1^\mu, \ldots, \lambda_1^\mu, \lambda_2^\mu, \ldots, \lambda_2^\mu, \lambda_L^\mu, \ldots, \lambda_L^\mu)$$

We call the L-dimensional space the *latent* space, with λ_i being the ith latent variable, and the N-dimensional space the *embedding* space, which corresponds to the neural activity space of N neurons.

We then increased the dimensionality of the patterns by randomly flipping each value ξ_i^μ in each pattern with probability δ. Therefore, δ controls the dimensionality of the resulting set of patterns, which ranges from L (at $\delta = 0$) to N (at $\delta = 0.5$). We denote the collection of patterns obtained after applying this distortion as $\{\xi^\mu\}_\delta$.

Computing the memory capacity as a function of dimensionality

We then tested how δ affects the memory capacity of a Hopfield RNN of N neurons. Notably, when patterns are random and uncorrelated (in our case, δ=0.5) we expect a capacity that scales with the number of neurons N.[49,50] In the next section below, we show through a theoretical argument that, in the case of patterns that span a L-dimensional space in an N-dimensional embedding (δ=0), the critical capacity is reached at O(L) patterns. In the simulation shown in Figure 3H, we numerically computed the memory capacity for the intermediate cases by varying δ in 10 equally-spaced values from 0 to 0.5. For each value of δ, we constructed a set of P patterns $\{\xi^\mu\}_\delta$ as explained above. We then used these patterns to train a Hopfield model and tested its ability to retrieve each of the patterns used for training from a noisy version of the original (see the next section for more details). We then defined the maximum capacity of the model as the maximum value of P such that the fraction of retrieved patterns was larger than 95%. The resulting value of $P(\delta)$ for all 10 values of δ was normalized with $P(0.5)$ to be visualized in Figure 3H.

Computing the generalization capacity (CCGP) as a function of dimensionality

To compute how the generalization capacity of the neural code is affected by dimensionality, we used the same set of patterns $\{\xi^\mu\}_\delta$ generated at a given value of δ and tested how well a decoder trained to report one of the latent dimensions λ_l generalizes across values of a second latent dimension λ_m. For a given pair of latent dimensions (l,m), we divided the set of patterns $\{\xi^\mu\}_\delta$ into four classes depending on the value of the two latent variables. We then trained a linear SVM decoder to discriminate patterns in the $(\lambda_l = 0, \lambda_m = 0)$ class from patterns in the $(\lambda_l = 1, \lambda_m = 0)$ class, and tested it in the task of reporting patterns in the $(\lambda_l = 0, \lambda_m = 1)$ class from patterns in the $(\lambda_l = 1, \lambda_m = 1)$ class. This gave us a CCGP for the (l,m) pair, denoted as $CCGP_{lm}$. We then took the mean CCGP value over all the possible pairs $\frac{1}{L(L-1)}\sum_{l,m} CCGP_{lm}$ as a measure of generalization performance.

Theoretical derivation of the memory capacity of 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. Benna and Fusi[35]) or in recurrent neural network like the Hopfield network,[49,50] 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.[51] 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 Benna and Fusi[35]). This neural circuit could be implemented in the dentate gyrus, which is known to play an important role in pattern separation[28,32,52] (pattern separation is clearly a form of decorrelation).

Theoretical derivation

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[50] 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 Hopfield.[50] 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:

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

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):

$$I_k = \sum_{l=1}^{N_L} w_{kl} \xi_l^1 + \sum_{l=N_L+1}^{2N_L} w_{kl} \xi_l^1 + \ldots$$

The first sum contains neurons that encode only the first latent variable Λ_1, the second sum only the neurons that encode Λ_2 etc. and all the states of activation ξ_l^1 within each sum are the same: for example $\xi_l^1 = \Lambda_1^1$ for all $l = 1,...,N_L$. It is now convenient to switch to the indexes of the latent variables:

$$I_\nu = N_L w_{\nu 1} \Lambda_1^1 + N_L w_{\nu 2} \Lambda_2^1 + ... = \sum_{\lambda=1}^{L} w_{\nu\lambda} \Lambda_\lambda^1$$

Where ν is the index of the latent variable encoded by neuron k, the neuron whose state of activation has to be updated. $w_{\nu\lambda}$ is the value of the weight between a neuron encoding latent variable λ and a neuron encoding latent variable ν and it is given by:

$$w_{\nu\lambda} = \sum_{\mu=1}^{p} \Lambda_{\nu}^{\mu} \Lambda_{\lambda}^{\mu}$$

We then separate the sum over μ into two parts:

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

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, hence the max capacity p 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[49,53,54] it is clear that the memory capacity scales linearly with the dimensionality of the input patterns, which in our case is L.

Numerical simulations

We verified this theoretical result by numerical simulations. Given a latent dimensionality L and a neural dimensionality N, we constructed p patterns Λ^{μ} by expanding the latent space into correlated chunks of N_L neurons each in the neural space, as described above. We then used the p expanded patterns ξ^{μ} to train an Hopfield model and tested its ability to retrieve one of the patterns used for training from a noisy version of the original. To construct noisy versions, we randomly flipped 10% of the units. A successful retrieval was identified if the model converged to a pattern with less than 5% flipped neurons compared to the original pattern after one step of the Hopfield dynamics, hence getting closer to the original pattern.

We then computed the fraction of retrieved patterns for different values of L and N. As shown in Figures S3A and S3B, the fraction of retrieved pattern decreases when p increases, a sign of limited memory capacity. The decreasing profile varied with L (Figure S3B) but not with N (Figure S3A), as expected by the theory. We then defined the maximum capacity as the maximum value of p such that the fraction of retrieved patterns was larger than 95% (green dashed line in Figures S3A and S3B). As shown in Figure S3C, this capacity increases with N, but it saturates to a value that depends on the latent dimensionality N. Moreover, this value is much lower than the one obtained with the same number of neurons in a high-dimensional setup where patterns are random and uncorrelated (black dashed line in Figure S3C). Finally, we computed how the maximum storage capacity, i.e. the maximum value of memorized patterns when N is large enough, scales with the latent dimensionality L. We found a good linear scaling of the maximum storage capacity with L (Figure S3D), hence confirming the results obtained in the theoretical derivation above.

Limitations of the model

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.[49,54] 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 from the experimental observations that CA2 is not really dealing with these low dimensional representations because the representations of familiar animals are high dimensional. 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.