Cross-Modal Representation of Identity in Primate Hippocampus

Timothy J Tyree[1,3+], Michael Metke[1,2+] & Cory T Miller [1,2*]

[1] Cortical Systems and Behavior Laboratory
[2] Neurosciences Graduate Program
[3] Department of Physics
University of California San Diego

+equal contributions
*corresponding author: corymiller@ucsd.edu

Faces and voices are the dominant social signals used to recognize individuals amongst human and nonhuman primates [1–5]. Yet, evidence that information across these signals can be integrated into a modality-independent representation of individual identity in the primate brain has been reported only in human patients [6–9]. Here we show that, like humans, single neurons in the marmoset monkey hippocampus exhibit invariant neural responses when presented with the faces or voices of specific individuals. However, we also identified a population of single neurons in hippocampus that were responsive to the cross-modal identity of multiple conspecifics, not only a single individual. An identity network model revealed population-level, cross-modal representations of individuals in hippocampus, underscoring the broader contributions of many neurons to encode identity. This pattern was further evidenced by manifold projections of population activity which likewise showed separability of individuals, as well as clustering for family members, suggesting that multiple learned social categories are encoded as related dimensions of identity in hippocampus. The constellation of findings presented here reveal a novel perspective on the neural basis of identity representations in primate hippocampus as being both invariant to modality and comprising multiple levels of acquired social knowledge.

Effectively navigating the sophisticated societies that typify primates relies on the ability to rapidly recognize each individual encountered and to infer relationships between those conspecifics to make appropriate social decisions [10–13]. An abstract, modality-independent representation of identity would confer a significant advantage for this purpose as prior experiences are bound together rather than separated by sensory input [14]. Evidence of such mechanisms, however, has been limited to 'concept cells', a sparse population of highly-selective neurons in the human hippocampus responsive to a single individual across different views and modalities [6,7]. These neurons are significant for several reasons including their putative role in declarative memory functions [8] and their potential uniqueness to humans [9]. By contrast, evidence of identity coding in the nonhuman primate brain has been limited to unimodal faces or voices[2,4,5,15–18]. Here we tested whether cross-modal representations of identity are evident in common marmoset monkeys (*Callithrix jacchus*) - a highly social New World primate - by recording the activity of single hippocampal neurons in response to the faces and/or voices of familiar conspecifics. We presented subjects with multiple exemplars of individual marmoset faces - from different viewpoints (Figure 1A) - and voices as unimodal stimuli, consistent with previous studies[6], as well as concurrently. These cross-modal stimuli involved presentations of faces and voices from the same (identity match) or different individuals (identity mismatch; Figure 1A). Overall, we observed no difference in visual behavior between the identity match and identity mismatch trials, but marmosets fixated significantly longer and made significantly fewer saccades in the voice-only trials relative to the face-only trials (Figure S1) indicating that these distinct social signals differentially affected marmoset visual behavior.

To test whether— like humans— neurons in the marmoset hippocampus exhibit invariant identity representations, we performed analyses identical to those described previously[6,19] and identified a small subpopulation of neurons in the hippocampus of marmoset monkeys (2.9%; N=67) that met the same criteria of modality invariant concept cells established in humans. This class highly-selective neurons, therefore, is not limited to our own species. Figure 1B plots the responses of an exemplar neuron to the face or voice of the preferred individual and two other conspecifics, while Figure 1C shows the Receiver Operator Characteristic (ROC) curve from the same recording session while the area under the ROC curve (AUC) quantifies the ability to predict the face or voice of the preferred individual. Although the vast majority of putative concept cells in marmosets (91.3%) were selective to a single individual (Figure 1D), the magnitude of selectivity (Figure 1E) was not as pronounced as in humans (AUC≈1)[6,19] and these neurons were widely distributed in all subfields recorded in the marmoset hippocampus (Figure 1F). A potentially important feature in our study that distinguishes it from prior human research is the familiarity of the individuals presented in the stimuli and the subjects. Whereas previous studies in humans have typically presented stimuli from well-known celebrities with whom subjects did not likely have a personal relationship, or single family members [6,19], here we presented subjects with a large set of familiar individuals in our colony that included both family members and unrelated individuals ($N_{individuals}$=12). Given that both human and nonhuman primates live in stable societies characterized by intricate, long-term relationships with known individuals, the pattern of responses here may reflect how the identity of individuals within one's social network are coded in this brain structure.

A potential limitation on these highly-selective neurons being a cornerstone for various memory functions [8] is the inherent disparity between the near infinite amount of information in the world and the limited number of neurons in the brain. A contrasting neural coding strategy would be for individual neurons to contribute to a range of different functions that collectively support population-level computations for complex behaviors [20–22]. In the case of social identity, this could be achieved by single neurons being sensitive to the cross-modal identity of multiple conspecifics, rather than only a single individual. To test this, we broadcast the faces and voices of familiar conspecifics concurrently while manipulating whether these signals were from the same (identity match) or different (identity mismatch) individuals (Figure 2A). We conjectured that if hippocampal neurons exhibited a significant change in activity between the identity match and mismatch trials, it would indicate that the cross-modal identities of multiple conspecifics are encoded in single hippocampal neurons. Analyses revealed another subpopulation of individual neurons (Figure 2B) that exhibited a significant difference in median firing rate over the stimulus duration between identity match and mismatch trials (N=217; 9.2%; Figure 2C), thereby confirming the existence of a mechanism distinct from putative concept cells for representing individual identity across modalities within single primate hippocampal neurons.

Quantifying firing rate differences over the entire stimulus, as in the previous analysis, however, may fail to capture more selective neural coding features relevant to cross-modal representations of identity, particularly for the long stimuli used here (>3s). To test this, we identified intervals of time during which individual neurons exhibited a significantly different median firing rate between identity match and mismatch trials, which we labeled as predictive time bins (Figures 2D,S2). This analysis revealed 732 hippocampal neurons with at least one predictive time bin, with most responding more during identity match trials (Figure 2E). Most of these predictive neurons only exhibited a single predictive time bin (Figure 2F). The duration of our predictive time bins ranged from 0.20-3.60s (median: 0.40s, IQR: 0.20-0.80s), with a median significantly less than that of uniformly sampled time bins (Figure 2G) according to a Wilcoxon–Mann–Whitney test ($p<0.001$, N=2×968).

Using an ensemble gradient-boosted decision tree model as our population-level decoder [23], we next analyzed whether identity match *versus* mismatch (MvMM) trials could be accurately classified using only those neurons with predictive time bins. The predictive population yielded almost perfect classification of MvMM trials, while the non-predictive population performed at chance (Figure 2H). The ROC curve for the predictive population further illustrates this success (Figure 2I), suggesting that this coding feature was highly reliable for distinguishing cross modal identity from hippocampal activity. Remarkably, only a relatively small number of randomly chosen predictive time bins (N=50) were necessary for the typical decoding testing accuracy to exceed 80% (Figure 2J), suggesting accurate cross-modal identity information was encoded by many neurons at the population-level. When holding the number of time bins fixed and equal to the total number of predictive time bins, our population-level decoder predicted MvMM with a testing accuracy exceeding 80% *despite* having 50% of its predictive time bins replaced by random time bins. When it considered the entire predictive population with an equal number of random time bins, our population-level decoder predicted MvMM with a testing accuracy exceeding 85% (Figure 2K). Although this decoder is limited in the sense that it is completely agnostic to individual identity, these results overwhelmingly support robust encoding of cross-modal identity at the population-level because distinguishing between identity match and mismatch trials relies on information about the faces and voices from multiple individual animals in each monkey's social network.

To explicitly test whether the primate hippocampus represents the identity of multiple marmosets within the same population, we developed an identity network model (INM) that integrated two different neural decoding approaches. We analyzed data only from the three observers with at least three known relationships with the individuals. The first approach was identical to the neural decoder described above, classifying MvMM trials blind to individual identity. The second approach resulted in identity-specific decoders for each individual to detect the face or voice in any trial. Our INM combined these two approaches to achieve cross-modal decoding of individual identity (Figure 3A). This combination was critical because while the MvMM decoder was highly accurate at classifying MvMM trials, it performed poorly on its own classifying the specific individuals (Figure 3B). By contrast, the identity-specific decoder performed almost perfect classification for individuals' faces or voices, but not when the face and voice were presented together (Figure 3C). Notably, performance of this latter decoder parallels results from putative concept cells, in which individuals are identified based on independent unimodal stimulus presentations [6,19], suggesting that recognition of an individual from their face or voice alone may be distinct from the mechanisms that integrate multiple modalities simultaneously into a representation of identity. When combined across individuals, the INM successfully distinguished ten individuals, supporting the identity of the individual being represented across both modalities (Figure 3D). Furthermore, decoding performance was at least 5× above chance when distinguishing ten individuals (Figures 3E,S3). Together, these results demonstrate for the first time that cross-modal representations of *individual* identity are evident at the population-level in the primate hippocampus.

Because putative concept cells were included in these analyses, we investigated whether their contribution to cross-modal representations of identity were disproportionate to their sparse distribution. To test this conjecture, we compared INM performance when these neurons were included and removed from the analysis. As an important control, we considered only individuals for whom at least one concept cell exhibited a preference in this analysis. The ROC curve averaged over all subjects (Figure 3F), as well as over each individual subject (Figures 3G,S4), was not appreciably affected by eliminating putative concept cells. This suggests that putative concept cells are no more significant for decoding the identity of individuals than other neurons in the population, at least for familiar conspecifics in the social network. Rather these neurons may function as a bellwether, indicating the presence of information pertinent to an identity concept within the broader hippocampal population.

The success of the INM provided compelling evidence that an individual within a marmoset's social network can be decoded from their face and/or voice, but an individual's identity is also coupled to their social relationships, such as their family. To test whether hippocampus encodes categorical attributes of social identity, we applied nonlinear dimensionality reduction techniques shown to be powerful tools for revealing elements of brain functions by projecting features of neural activity into low-dimensional manifolds [24], including in studies of hippocampus [25,26]. As a first step to this end, we tested whether the spatial clustering of manifold projects would replicate the findings of the INM.

Using the same individual-specific predictive time bins for marmoset faces and voices as in the INM (i.e. Figure 3C), manifold projections revealed a similar pattern of separability between individuals (Figure 4A), effectively replicating our previous result using an entirely different analysis. Notably, further analyses showed that removal of predictive time bins significantly decreased separability of individuals in this analysis (Figure S5). To investigate whether the representation of identity can be described by the relative timing of spike events, we complemented this rate-coded representation with an analysis of *signed connection rate* as a parameterless event-coded measure, which considered only the minimum of the hindsight delay and the foresight delay (Figure 4B, left). We evaluated our signed connection rate at the spike times of each neuron to describe how one neuron "connects" with another, revealing statistical distributions specific to any given pair of neurons (Figure 4B, right). Results using this event coded measure again revealed excellent separability for individual faces and voices (Figure 4C), effectively replicating the effect observed for the rate coded features and INM using a facet of neural activity that is entirely independent of time bins, further illustrating that cross-modal representations of identity are robustly encoded in the population activity of marmoset hippocampal neurons.

Having shown that manifold projections can represent cross-modal identity in event-coded hippocampal activity, we tested whether other social categories pertinent to an individual's broader identity may likewise be represented in the same manifold projections. We calculated the mean squared range (MSR) of each individual from the mean over all individuals. We then investigated the MSR of two social categories: cagemates (i.e. family members) and non-cagemates. In both subjects with at least three individuals presented as stimuli in each social category, we observed a significantly larger median MSR for family members compared to non-cagemates (Figure 4D), suggesting hippocampal dynamics represent both individual identity and the relatedness of individuals. Notably, while the initial individual identity analysis was supervised, the clustering that emerged based on respective social relatedness was unsupervised. Figure 4E shows the bundled edges from a two-dimensional manifold projection to visualize the relationships between conspecifics in hippocampal activity, which supports connections existing between identities in the neural representation. We next computed an unsupervised latent firing rate as the manifold projection of the absolute value of our signed connection rate. Though it did not demonstrate separation of individual identity, it did exhibit trajectories that appeared stable in time and comparable between trials (Figure 4F). Averaging over N=12 recording sessions, we found the motion of mean latent firing rate was significantly larger at multiple time points while one observer, Baloo, was observing cagemates compared to non-cagemates (Figure 4G). The other observer demonstrated similar motion for individuals from both social categories. While idiosyncrasies in the face and voice directly convey individual identity, information about familial relationships must be abstracted based on experience observing conspecifics interacting with the other individuals around them [13,27]. These results suggest that neural representations of social identity in primate hippocampus are not only invariant to the sensory modality but reflect the rich corpus of acquired knowledge about the relationships between familiar marmosets in the social network.

Here we showed that the cross-modal identity of multiple conspecifics are represented in the marmoset hippocampus. Although we identified putative concept cells similarly to humans [6,8], we discovered that this population of highly selective neurons is not the only mechanism for representing concepts of individuals. Rather, both single neurons and the broader population in hippocampus encode the cross-modal identity of multiple conspecifics. Furthermore, analyses revealed that a population-level code represents not only the identity of multiple familiar individuals invariant to the modality of the social signal, but information pertinent to social categories, as well. Similar to the role of hippocampus in other contexts [28], these representations likely reflect learned social schema for identity that facilitates the rapid recognition of individuals necessary for guiding decision-making during the complex interactions characteristic of complex societies. The presence of unimodal representations of identity in the primate face and voice patches [17,18] with the close anatomical connectivity between the temporal cortex and medial temporal lobe [29,30] supports an integrative social recognition circuit in which substrates in the broader network plays distinct but complementary roles that together govern natural primate social brain functions [31–33]. Elucidating how such a schema confers an advantage for primates, however, will likely require experiments in which freely-moving monkeys leverage these neural representations to navigate their social landscape [22].

Acknowledgements.
This work was supported by grants from the NIH to CTM (R01 NS109294, R01 DC012087).

Methods

Subjects.

Four adult marmosets (2 male, 2 female) served as subjects in these experiments. All animals are socially housed with 2-8 conspecifics in the Cortical Systems and Behavior Laboratory at the University of California San Diego (UCSD). The UCSD marmoset colony in the Miller Lab houses ~70 animals in 15 family groups. All procedures were approved by the Institutional Animal Care and Use Committee at the University of California San Diego and follow National Institutes of Health guidelines. A total of 47 recording sessions were performed with these subjects over the course of the experiment and analyzed here.

Experiment Design.

Neurophysiological recordings were performed while subjects were head and body restrained in our standard marmoset chair [34,35]. Visual stimuli were presented on a LED screen from a BenQ monitor 1080 positioned 24 cm in front of the animal. Acoustic stimuli were presented at 70-80 dB SPL. All behavior was collected in an anechoic chamber illuminated only by the screen, which had a dynamic range from 0.5 to 230 cd/m^2, with luminance linearity verified by photometer. Stimulus presentation was controlled using custom software and eye position was monitored by infrared camera tracking of the pupil. For hardware, calibration, and validation, Ref. [36].

Subjects initiated trials by holding fixation of gaze for 100ms at a center fixation dot on the screen, at which point stimulus presentation was initiated. The 150ms period immediately post-stimulus was discarded to account for the time for visual signals to propagate from the retina to the hippocampus. This latency has been measured to be in the range 100-200ms in macaques [36–39] and relatively recent data suggests it is comparable in marmosets [40]. This biophysical argument supports our estimate of the stimulus onset t=0 occurring 150ms after stimulus was presented. Unless otherwise specified, baseline firing rates were estimated from the 500ms preceding t=0. Stimulus responses were initially measured by comparing the peristimulus baseline firing rate to firing rates averaged from t=0 to 3s.

Stimuli were divided amongst unimodal – face-only and voice-only – and cross-modal – identity match and identity mismatch on a trial-by-trial basis. Up to twelve conspecifics were represented per stimulus set (min 9, max 14). Face stimuli comprised multiple examples of each individual marmoset from different head orientation.

Each individual marmoset was represented in multiple distinct stimuli (N=36.0+/-15.3) for each individual in each recording session across each of the four stimulus modalities. Monkeys with fewer than 10 presentations per individual in a recording session were not considered in any analysis. The stimulus duration of trials involving vocalizations (i.e. voice-only and cross-modal) necessarily varied because each “phee” call differed in duration (mean: 3.02+/-0.74s). The face stimulus duration was 3.5s. Stimuli were presented in 10-trial blocks, with an inter-block active forage trial with juice reward to maintain attention. Each recording set was composed of 400 face and/or voice stimuli, split into 2 subsets.

All stimuli were composed of faces and/or voices of conspecific monkeys in our colony familiar to each subject. A total of 16 individual monkeys were represented overall (9 male, 7 female). Test subjects were not included in their own stimulus sets. Because our goal was to test for representations of individual identity rather than cross-modal perceptual integration of face/voice biomechanical movements (i.e. McGurk Effect [41]), we presented subjects with static face stimuli so as not to introduce confounds that may emerge due to temporal misalignments of the face and vocalizations during identity mismatch trials.

All face stimuli were photographs of monkeys from our colony taken while animals were in our standard marmoset chair [35,42] with a light background behind them. The animals are trained to sit comfortably while a neck guard restricted their mobility. While seated, subjects could freely change head direction. Photographs of each subject were visually inspected and selected based on image quality and suitable representation of multiple head orientations(Figure 1A,2A). Photos used as stimuli were cropped to only show the neck guard and the face/head, so as to eliminate views of the rest of the body and chair.

All voice stimuli were marmoset “phee” calls comprising two pulses, the species-typical long-distance contact calls [43]. Previous work has shown that that marmosets are able to recognize the caller’s identity when hearing “phee” calls [44]. Recordings were made at 44.1 kHz sampling rate while a monkey engaged in natural vocal interactions with a visually occluded conspecific in a soundproof chamber and hand-selected using custom code. Only examples with high SNR and minimal background noise were selected for stimuli.

Surgical and neural recording details.

The surgical procedure employed here has been described previously [45]. Briefly, we performed an initial surgery to affix a post to the skull on each animal to restrain subjects’ head during experiment preparation. Following recovery, a second procedure was performed to embed the drive housing and the electrode array for stable chronic electrophysiological recording. We implanted a 64-channel microwire brush array (MBA, Microprobes [46]) either unilaterally or bilaterally into the hippocampus using preoperative MRI stereotaxic coordinates. Electrode locations were confirmed by postoperative MRI and histology. All surgeries were performed under sterile and anesthetized conditions. The implants were inserted 7-13 degrees of angle off the vertical using the medial sulcus as reference before the operation has taken place.

Neural recordings were performed with an Intan 512ch Recording Controller system via an RHD2164 64-channel amplifier chip, sampled at 30kHz [47]. Neurophysiology data was analyzed using Spyking Circus [48] yielding across all recording sessions 2,358 isolated single-units, referred to as neurons in the main text and in the remainder of Methods. Standard procedures were employed to remove obvious recording errors, which resulted in less than 1% of trials being removed from the analysis *a priori*.

Determining statistically significant differences in firing rates.

Unless otherwise specified, the Wilcoxon-Mann-Whitney test was used to test for statistically significant differences between two samples of firing rates. This includes the determination of preferential responses to the MvMM task averaged over the 3 seconds following stimulus onset (Fig. 2C) and during predictive time bins (Fig. 2E). The Wilcoxon-Mann-Whitney test compares median values without making any assumption of normality [49].

Identifying concept cells.

Hippocampal neurons were tested for invariant response to individuals in the face-only and voice-only trials using an ROC analysis identical to that described in human hippocampus [6]. For each isolated single neuron we performed the analysis for all identities where at least 3 unimodal stimuli (either face or voice but not both) were presented for either mode: face-only and voice-only. Above-threshold firing-rate responses to any stimulus of the preferred subject was considered a positive trial. Significance of an ROC for a given subject was determined by comparison to 99 surrogate ROC curves, which resulted from using the same labels with randomly selected trials. Curves that surpassed all surrogates were considered significant ($p<0.01$). Neurons that met or exceeded these thresholds were determined to be putative concept cells for individual identity in marmosets.

Identifying predictive time bins.

Hippocampal neurons were analyzed in terms of their firing rate response during time bins that we identified as candidate. For each neuron, our procedure consisted of three stages. The first stage was to generate a large list of time bins of varying duration using an extension of the sliding window approach. The second stage identified a subset of time bins as having a general ability to distinguish trials. We required this subset to be mutually disjoint. Candidate time bins resulted from the third stage, which varied each time bin independently according to our refining procedure.

The first stage extended the sliding window approach by using 200ms time bins evenly distributed between 0 and 3.6 sec, the maximum stimulus duration (Figure S2A). Time bins of duration greater than 200ms were constructed by joining adjacent time bins, leading to a maximum allowed time bin duration of 3.6 seconds (Figure S2B). A general ability to weakly distinguish trials was determined by splitting the training trials according to three-fold stratified cross-validation and then computing the training AUC of each fold (Figure S2C). Training AUC was initially computed from the ROC curve that resulted from an above-threshold firing rate response determining a positive trial. Separately, training AUC was computed from a below-threshold firing rate response as determining a positive trial. In either case, if the training AUC was greater than chance (AUC>0.5) for all folds, then the time bin was retained for stage two (Figure S3D). The same convention for *above* versus *below* firing rate response as determining a positive trial was used for stage two and for stage three. All population-level decoders were blind to this convention of sign.

The second stage selected a disjoint set of candidate time bins, optimizing for their ability to distinguish trials by maximizing the mean AUC averaged over the same cross-validation. To achieve this, time bins were selected in decreasing order of their mean AUC and included if doing so maintained the disjointness of time bins retained (Figure S2D).

The third stage refined the resulting disjoint set by considering a number of random perturbations of each remaining candidate time bin and keeping only the optimal perturbation. The random perturbations shifted the start times and the end times independently by a random amount identically sampled from the normal distribution with zero mean and standard deviation equal to the duration of the unperturbed time bin. We generated a sample of N=100 perturbed time bins and removed those with a duration <10ms. Perturbations were additionally removed if they exhibited a start time before stimulus onset t=0 or if they exhibited an end time after t=3.6 seconds. A worsening AUC in any of the folds resulted in rejection of the given candidate time bin.

If any of the resulting training AUC values were smaller than that of the unperturbed time bin, there that perturbation was removed from consideration. The overall training AUC was computed for each perturbation using all training trials together. The perturbed time bin with the largest overall training AUC was kept instead of the unperturbed time bin. Perturbed time bins were allowed to overlap with other remaining time bins, thereby relaxing the condition of disjointness for the sake of parallelizability. A flowchart summarizes the procedure (Figure S2E).

If no perturbations remained under consideration, then the unperturbed time bin was kept from stage two. Approximately 31.4% (N=304/968) of the predictive time bins in Figure 2 were unaffected by the refinement procedure of the third stage. Any remaining candidate time bins were considered predictive only if they presented a statistically significant difference in median firing rate for the true (e.g. identity match) training trials compared to the false (e.g. identity mismatch) training trials (Figure S2F). Significance was determined according to p<0.05, where p was the statistic computed as the mean p-value resulting from a Wilcoxon–Mann–Whitney test conducted over the training trials averaged over five stratified cross-validation folds over training, which was a sufficient statistic in the sense that all time bins with p<0.05 also exhibited a statistically significant difference in median value at the same level of significance according to a Wilcoxon-Mann-Whitney test conducted over all MvMM trials. This procedure provided the sufficient features used in our population-level decoders. Data and code are made available to the reader (see Author Contributions).

Generating random time bins from the non-predictive population.

Time bins were randomly selected for neurons uniformly drawn from the population that exhibited zero predictive time bins (the 'non-predictive population'). Time bins possessed start and end times drawn from a uniform random sample from t=0 to 3.5 seconds, the latter of which was the median stimulus offset time. Time bins with a duration briefer than 0.2 seconds were immediately removed from consideration.

The number of random time bins involved in Figure 2G,I was equal to the number of predictive time bins involved. To include the entire aggregated predictive population as only ~1% of the mixed population in Figure 2K, we considered a pool of random time bins that was at least 100X larger than the aggregated pool of predictive time bins. This is a consequence of all random sampling from any time bin or neuron population was performed without replacement. Furthermore, uniform random sampling was conducted for all pools of time bins in analyses. During the training of the population-level neural decoders, the condition of uniform random sampling of time bins was relaxed.

Training the population-level neural decoders.

Population-level decoders were trained on the training trials before computing predictions for the separate testing trials. Decoders were trained and tested on a Quadro RTX 5000 GPU typically in less than five seconds of runtime.

The population-level decoders trained using firing rates directly as inputs. Neither translating nor scaling of the firing rates was performed, as the decoders were both location and scale invariant [23]. The prediction was estimated by the weighted average of values returned by an ensemble of decision trees (Figure S2G) relative to a default value of one half (controlled by base_score in Table 1). For each training epoch, at least 25 decision trees were trained (controlled by num_parallel_tree). While a unique solution exists for a given decision tree, a heuristic algorithm was used to approximate the unique solution using the quantile method [50].

Decision trees were trained to minimize the binary cross-entropy loss function (equivalently, to maximize likelihood) at the ensemble-level by considering only a fraction of the training trials (controlled by subsample). Decision node rules considered only a fraction of the input firing rates (controlled by colsample_bynode) to determine placement of its weight. The weight of a node was limited to a certain amount (controlled by max_delta_step). The complexity of the decision node rules was further limited using linear and quadratic regularization (controlled by reg_alpha and reg_lambda in Table 1, respectively).

Each decision tree was gradient boosted in the sense that nodes were recursively added in accordance to an estimate of the gradient of a training loss computed at the ensemble-level. If inserting a decision node failed to improve the loss by a sufficiently large amount (controlled by gamma), then that decision node was removed from the tree. To further limit structural complexity, the maximum tree depth was set to no more than five decisions (controlled by max_depth). The weight for a new decision tree was scaled down by a factor (controlled by learning_rate). Training terminated for a given decision tree when the total weight for the next decision node was smaller than a certain amount (controlled by min_child_weight). After all decision trees terminated training, the training epoch was complete. After a fixed, predetermined number of training epochs, the ensemble terminated training. Then, predictions were computed for the testing trials. Predictions were used to evaluate the predictive ability of a given set of one or more predictive time bins in terms of AUC (Figure S2H).

Determining hyperparameter settings for the population-level neural decoders.

The parameter settings for our population-level decoders resulted from a series of coarse grid searches each conducted over a wide range of settings for one pair of hyperparameters at a time. Each parameter setting considered five-fold stratified cross-validation involving the training trials only with the goal of maximizing mean testing AUC. Early stopping was used during this tuning procedure, which supported a minimum 60 training epochs for the match vs mismatch

(MvMM) predictive population and a minimum 67 training epochs for the identity-specific predictive population as sufficient according to early stopping. By increasing the number of epochs, stability of performance became immediately apparent for up to 500 epochs for both MvMM and identity-specific decoders. We made no use of early stopping anywhere else apart from the hyperparameter tuning procedure described here.

This hyperparameter tuning procedure was conducted only on the training trials for Archie observing Waylon in recording session #8. Archie (male) and Waylon (female) were never cagemates – though they may have known eachother in the colony. These training trials (from session #8) were complementary to testing trials from no more than one of the multiple recording sessions summarized in Figure 3. The hyperparameter settings that resulted are reported in Table 1.

Hyperparameter	MvMM	identity-specific
base_score	0.5	0.5
num_parallel_tree	25	50
subsample	0.2	0.2
colsample_bynode	0.1	0.1
max_delta_step	0.5	1
reg_alpha	0.4	0.3
reg_lambda	0.4	0.3
gamma	0.1	5
max_depth	5	2
learning_rate	0.9	0.6
min_child_weight	0.5	1

Table 1 Table of hyperparameters for our population-level decoders. Numerical values were passed as arguments to the constructor of xgboost.XGBClassifier instances [20]. Columns correspond to the two types of predictive populations reported in the main text.

Aggregating predictive populations from multiple recording sessions.

For the aggregated MvMM decoder reported in Figure 2, the training procedure was performed on 120 training trials randomly selected from 150 MvMM trials aggregated from N=14 recording sessions. Trials chosen to be aggregated together shared the same label for each recording session involved. Training consisted of 200 training epochs using the hyperparameter settings listed in Table 1. The results of all except for the first ten training epochs were considered when computing predictions with the remaining 30 testing trials. We controlled for moderate unbalanced sampling of training trials by scaling the positive weights by a factor of 5. Moderate unbalanced sampling was a necessary consequence of the trial stimulus selection being randomized. The train and test procedure was repeated for each fold involved in five-fold stratified cross-validation of the aggregated trials. Predictive time bins were identified separately for each fold.

If a recording session did not possess at least 150 cross-modal trials, then it was removed from consideration in the aggregated predictive population reported in Figure 2. A number of recording sessions (N=10/24) were removed from the aggregation procedure for exhibiting fewer than 15 predictive time bins in any of the folds. Predictions for the testing trials of all folds were summarized in Figure 2H,I by concatenating predictions and ground truth labels across folds before computing the ROC curve and the associated testing AUC. The appearance of almost perfect predictions of MvMM can be attributed to predictive populations being aggregated from multiple recording sessions (N=14). Testing performance was lower for recording sessions considered individually, as is reported in Figure 3. This supports a larger number of predictive time bins corresponding to more predictive population-level decoders. Interaction between firing rates from multiple recording sessions was not permitted anywhere except for quantifying the effect of restricting the abundance of predictive time bins.

Quantifying of the effect of restricting the abundance of predictive time bins. To systematically vary the relative abundance of predictive time bins, we randomly sampled time bins from the non-predictive population. Their firing rates were concatenated with those of all available predictive time bins. We took a random sample of predictive time bins in addition to a statistically independent random sample of time bins from the non-predictive population. The relative sizes of the samples were chosen to reflect a given relative abundance of predictive time bins. Testing accuracy was computed at the same relative abundance over many statistically independent samples (N=100) in order to estimate the mean testing accuracy conditioned on the relative abundance of predictive time bins considered by the decoder. Uncertainty in mean testing accuracy was estimated by bootstrapping that same sample of testing accuracies, resulting in 95% confidence intervals less than 1% for both traces reported in Figure 2K. Many random time bins ($N>10^5$) were independently generated for this analysis in order to estimate the mean testing accuracy at the 1% minimum relative abundance reported in the main text while simultaneously involving the entire aggregated predictive population.

The fold with the median testing performance (AUC=0.9911) provided the predictive time bins (N=347) and the aggregated trials (N=150) that were used to quantify the effect of restricting the number of predictive time bins in Figure 2J. The fold with the lowest testing performance (AUC=0.9111) provided the predictive time bins (N=335) and the aggregated trials (N=150) that were used to quantify the effect of restricting the relative abundance of the predictive time bin in Figure 2K.

Summarizing testing performance from multiple predictors.

Population-level decoders were trained as MvMM or identity-specific predictors for each individual identity in each recording session involved in Figure 3. To account for variations in prediction magnitude between decoders, predictions were scaled linearly to a maximum value of unity before combining ROC traces the multiple recording sessions summarized in Figure 3B-D,F-G. No such scaling was involved with the multiclass predictions reported in Figure 3E.

Sampling trials for multiple predictive populations from the same recording session.

For a given recording session, the following criteria were respected while partitioning testing trials from training trials involving the identity network model (INM) discussed in the main text. Testing trials for the INM were also testing trials for both the MvMM decoder and the identity-specific decoders. Because stimuli involving individuals were sampled uniformly, the frequency of a given individual could be small for a given recording session. To account for this, individuals were considered only if they exhibited at least forty appearances in a given recording session.

Because of the uniform nature of our uniform sampling over the larger space of cross-modal stimuli, each recording session had relatively few trials involving both the face and the voice of a particular individual. This resulted in far more negative trials being presented to the observer relative to the number of true trials for the INM. This was also the case for both the MvMM decoders and the identity-specific decoders reported in Figure 3. All three binary classification tasks had balanced samples randomly selected, which were then randomly shuffled before 30% were randomly selected to be the testing trials. The remaining 70% of trials were considered for training. Unbalanced sampling in the training set was accounted for by scaling the positive weights by a factor of 5 for the MvMM decoders and 100 for the identity-specific decoders. Decoders involved in Figure 3 used 200 training epochs, all of which were used in testing decoder performance except the first training epoch. The only exception was the identity-specific decoders involved in evaluating the INM for the winner-take-all model in Figure 3E, which considered all 500 training epochs.

Decoding multiple identities using a winner-take-all model.

We used the winner-take-all model to predict the identities of multiple individuals shown during identity match and face-only trials. The ten individuals summarized (Figure 3E) have their detailed testing performance reported supplementarily (Figure S3). The winner-take-all model predicted the correct identity with an overall testing accuracy of 89% (N_{trials}=198). For a given recording session, the following procedure was performed to generate the predictions for the winner-take-all model. First, we identified all identities involved in a sufficient number of identity match trials (N_{trials}≥12). All identity match trials involving the identities identified were shuffled and 30% were randomly selected as testing trials to be withheld from training with the remaining 70% of trials.

We considered predictions of our INM to approximate a predicted probability that a given trial from the testing set involved the given identity. The presence of the individual was modeled using the decoder outputs in the winner-take-all model if the INM had the sufficient number of predictive time bins available. After repeating this procedure for all individuals in the recording session, the predicted identity of the winner-take-all model corresponded to that of the maximum predicted value.

Quantifying relative contribution of concept cells in decoders of their preferred identities.

To investigate the possibility of concept cells exhibiting any clearly observable significance in the INM at the population-level, we removed all concept cells from consideration and recomputed the testing predictions of Figure 3D for each individual that was statistically preferred by a concept cell ($N_{neurons}$=29). After recording the testing AUC, we repeated a comparable procedure as a control that randomly removed an equivalent number of predictive time bins from any neuron that was not found to be a concept cell. This control procedure was repeated many times (N=200) and then averaged to estimate the mean control testing AUC, which was not significantly different from a normal distribution according to the omnibus test for normality (p>0.05, N=200). The aforementioned control and test procedures were conducted using independent randomized samples.

ROC curves were computed with above-threshold values indicating a positive trial for the three observers with at least three cagemates amongst the identities presented. The INM appeared successful despite the removal of concepts cells independently for multiple observers: Archie (Figure S4A), Baloo (Figure S4B), and Hades (Figure S4C). Removing concept cells from the INM for all recording sessions involving one observer resulted in a mean testing AUC that was not significantly smaller than that of the control according to a one-tailed paired student's t-test. We independently replicated this same statistical insignificance of concept cells at the population-level for multiple observer subjects (p>0.05, N=3). This insignificance was consistent with a comparable analysis that made no assumption of normality, which suggested the median testing AUC was also not significantly smaller when all concept cells were removed relative to the control (p>0.05, N=3). It is uncertain whether this insignificance can be attributed to these concept cells being observed in nonhuman primates, as no comparable predictive time bin analysis has ever been performed in humans to the knowledge of the authors.

Computing signed connection rate.

Our event coded representation relied on our signed connection rate measure, which we computed using our two primitive event measures. The first we referred to as the hindsight delay, $\tau_- > 0$, which is the amount of time since a given neuron has spiked. The second we refer to as the foresight delay, $\tau_+ > 0$, which is the amount of time until a given neuron will spike. A schematic illustrating the computation of the hindsight delay is shown (Figure 4B, left). A similar computation is found for the foresight delay by time inversion. If the given neuron has not yet spiked, then we take the hindsight delay to approach infinity. Similarly, if the given neuron was not observed to spike again, then we take the foresight delay to approach infinity. Note that our primitive event measures do not evaluate to non-positive real numbers.

The magnitude of our signed connection rate is the multiplicative inverse of the minimum of the hindsight delay and the foresight delay. Finally, we set the sign of our signed connection rate to be negative if the hindsight delay was used. Using the standard conventions of real analysis, our signed connection rate is now well-defined at all times for all neurons that exhibited at least two spikes. Equivalently, our signed connection rate was computed according to a real function of two variables

$$c(\tau_+,\tau_-) = \frac{\Theta(\tau_- - \tau_+)}{\tau_+} - \frac{\Theta(\tau_+ - \tau_-)}{\tau_-},$$

where Θ(x)=1 if x is nonnegative, otherwise, Θ(x)=0.

We evaluated our signed connection rate for every neuron at the spike times of each neuron. This was our attempt to measure how a single neuron "connects" with any other neuron. In doing this, we observed statistical distributions that appeared specific to a given neuron pair (Figure 4B, right). We considered a given neuron to have an approximately symmetric signed connection rate if it exhibited no more than twice as many negative values as positive values in these statistical distributions.

Estimating manifold projections.

We used uniform manifold approximation and projection (UMAP) to compute our manifold projections in Figure 4 of the main text, which presents descriptive manifold projections computed from predictive firing rate features and separately from our signed connection rate measure of spiking events. The same parameter settings on the same optimization algorithm was used for both rate and event coded manifold projections. We used predictive firing rates from the MvMM population in Figure S5C,E. We used firing rates concatenated from the identity-specific predictive populations otherwise with the exception of our analysis of the apparent Euclidean distance in the rate coded representation (Figure S5A,B). The rate coded manifold projections considered neuron spikes from t=0 to 2 seconds after the stimulus onset. Similarly, the event coded manifold projections considered neuron spikes from t=0 to 2 seconds after the stimulus onset. The average predictive time bin from the MvMM predictive population reported in Figure 2 was centered from t=0 to 2 seconds after the stimulus onset, with 45.5% of predictive time bins ending earlier (N=440/968), which supports 2 seconds as a reasonable choice for the max time considered by the rate and event coded manifold projections.

The UMAP algorithm was composed of two steps that can fruitfully be described as graph construction and graph projection [24]. The graph was constructed from a given set of comparable observations. The graph was projected to a low-dimensional space of real numbers. In the optimization procedure, five negative samples were selected for each positive sample. The minimum distance between two observations was set to 0.1 Hz. The number of nearest neighbors was initialized to 50 for rate-coded representations and 100 for our event-coded representations. Repulsion strength was initialized to unity. Local connectivity was set to 1 Hz in estimating probability distances. We trained for 500 epochs at a learning rate initialized to unity for all observations. The resulting function was equipped with a learned graph of the observations, which projected to the manifolds visualized in Figure 4,S5. An example of connections from such a learned graph were visualized (Figure 4E).

For our rate-coded manifold projections, the inclusion of predictive time bins (p<0.05) appeared sufficient for the separation of individuals (Figure S5A), which was supported by computing the minimum distance between the centroid of any individual and then comparing across multiple recording sessions. Minimum distances that were computed from predictive time bins exhibited a significantly smaller median value when compared to candidate time bins that were not predictive (p>0.85) according to a Wilcoxon-Mann-Whitney test (p<0.001, N=29), suggesting predictive activity leads to better separation of individuals in comparable rate coded representations (Figure S5B). Shown are examples of rate-coded manifold projections that used predictive firing rates as trial-by-trial observations. Event-coded manifold projections used signed connection rates as spike-by-spike observations for Hades (Figure S5C,D) and for Baloo (Figure S5E,F).

Estimating latent firing rate.

Our latent firing rate was computed using unsupervised nonlinear dimensionality reduction of the absolute value of the signed connection rate for all neurons that had no less than one third of its computed signed connection rate values as positive (i.e. approximately symmetric). In computing the latent firing rate, we used a method of nonlinear dimensionality reduction that made no assumption of uniformity [51]. The output metric and the input metric were both Euclidean (flat), which supports the output having the same units as the input. The output was embedded in six-dimensional real space, and the first three dimensions were plotted for an exemplar recording session (Figure S5G). After this output was computed at the spike times of all neurons involved, it was analyzed as a time series by time ordering according to the evaluation time.

By considering latent firing rates evaluated at the times t=0 to 4 seconds after a stimulus onset, we observed relatively stable trajectories for multiple recording sessions conducted over multiple observers. We performed a median filter with a sliding window of 50 neuron spikes before plotting our estimates of the latent firing rates. Shown are three exemplar identity match trials, where Baloo observed the face and voice of her mother, her father, and her sister (Figure S5H). Our rationale for choosing 6 dimensions to embed the latent firing rate considered an experimentally observed latent space of neuronal networks having between four and six real valued dimensions [22]. This interval includes 5.25695…≈5.2, which approximates the dimension that maximizes the volume of the unit hypersphere. Since 6 is the smallest integer that is greater than 5.2, we considered 6 dimensions in estimating the manifold projection time series we referred to as latent firing rate.

Determining anatomical positions of implants.

All implants were followed by at least one postoperative MRI (Figure S6). The scans were aligned to anatomical features with RadiAnt Dicom viewer [52] and the position along the anterior-posterior axis was determined by measurement from the center of the array to the ear canal. Because implants were stereotactically performed coronally, all recordings for a given array were assigned the same anterior-posterior (AP) position.

Because of the 1mm spread of the microwire brush arrays, it was difficult to precisely estimate the position of any given electrode, or indeed the entire bundle on a particular day. We used the position of the tip of the electrode from each MRI and extrapolated the trajectory by estimating position along the drive axis by cross-referencing with contemporaneous notes made of the date and distance of every movement of the drive. Based on a centroid at each estimated position, we chose particular sessions for we had the greatest confidence that the majority of the array was located predominantly in one or two hippocampal fields. Because the relative positioning of individual electrodes was not clearly observable, all reported analyses were developed to be agnostic to neuron location.

Confirming implant location by MRI.

MRI was performed at the UCSD Center for Functional Magnetic Resonance Imaging in a 7.0T Bruker 20cm small animal imaging system using Advance II software [35]. Preoperative images were analyzed in Osirix DICOM Viewer [53] and stereotactic coordinates were established using a pair of saline-filled barrels affixed above the putative posterior end of temporal sulcus (marked on the skull during headcap surgery). Array positioning and tract trajectory was verified by post-operative MRI. Follow-up scans were performed occasionally to update array position.

Determination of anatomical positioning was performed using RadiAnt DICOM Viewer (Medixant, n.d.). Stereotactic alignment was performed using a number of clearly defined and readily identifiable anatomical landmarks. 2D coronal slices were made vertical by rotating to align the medial longitudinal fissure with a vertical line. Yaw was corrected by re-slicing the coronal plane to align both interaural canals. Pitch correction was performed by re-slicing MRI so that the 4th ventricle was aligned vertically with the isthmus of the corpus callosum.

Position on the anterior-posterior axis was calculated relative to the interaural canal. Measurement was taken from the coronal slice at which the array first entered the hippocampal complex (Figure 1F, left). Arrays were implanted with as little pitch as possible, so AP position variability is negligible along the electrode trajectory.

Electrode positions are not precisely determinable with our brush arrays, as microwires are not visible at the resolution of the scans and individual tips are not individually distinguishable by any practical means available. Electrode splay of the 64-ch MBA in tissue has been measured at approximately 1mm [46,54], so we approximated electrode position by use of a 1mm spherical voxel centered at the tip of the array.

We used a Microdrive with a 500µm thread pitch that could reliably make controlled movements with a precision of 30-40µm. An array tip was identified for every MRI in each subject and position was extrapolated based on contemporaneous notes regarding electrode movement. Once putative array centroids have been hand-tagged they were assigned to one of the hippocampal subfields. Centroids were deemed to be in a hippocampal subfield if more than 70% of their volume fell within that area, as assessed by hand-traced MRI. Recording sessions where the centroid fell significantly between two subregions were not counted in anatomical analyses. CA2 and CA3 were combined due to insufficient granularity in this methodology and resolution in our scans to effectively differentiate them. Figure S6 shows the estimated position of each electrode array in the hippocampus for all subjects.

Citations.

1. Miller, C. T. *et al.* Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219–233 (2016).
2. Tsao, D. Y. & Livingstone, M. S. Neural mechanisms for face perception. *Annu. Rev. Neurosci.* **31**, 411–438 (2008).
3. Leopold, D. A. & Rhodes, G. A comparative view of face perception. *J. Comp. Psychol.* **124**, 233–251 (2010).
4. Belin, P., Bodin, C. & Aglieri, V. A "voice patch" system in the primate brain for processing vocal information? *Hear. Res.* **366**, 65–74 (2018).
5. Freiwald, W., Duchaine, B. & Yovel, G. Face Processing Systems: From Neurons to Real-World Social Perception. *Annu. Rev. Neurosci.* **39**, 325–346 (2016).
6. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
7. Quian Quiroga, R., Kraskov, A., Koch, C. & Fried, I. Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain. *Curr. Biol.* **19**, 1308–1313 (2009).
8. Quiroga, R. Q. Concept cells: The building blocks of declarative memory. *Nat. Rev. Neurosci.* **13**, 587–597 (2012).
9. Sliwa, J., Planté, A., Duhamel, J.-R. & Wirth, S. Independent Neuronal Representation of Facial and Vocal Identity in the Monkey Hippocampus and Inferotemporal Cortex. *Cereb. Cortex* **26**, 950–966 (2016).
10. Chang, S. W. C. *et al.* Neuroethology of primate social behavior. *Proc. Natl. Acad. Sci. U. S. A.* **110 Suppl 2**, 10387–10394 (2013).
11. Tremblay, S., Sharika, K. M. & Platt, M. L. Social Decision-Making and the Brain: A Comparative Perspective. *Trends Cogn. Sci.* **21**, 265–276 (2017).
12. Platt, M. L., Seyfarth, R. M. & Cheney, D. L. Adaptations for social cognition in the primate brain. *Philosophical Transactions of the Royal Society of London* **371**, 20150096 (2016).
13. Bergman, T. J., Beehner, J. C., Cheney, D. L. & Seyfarth, R. M. Hierarchical classification by rank and kinship in baboons. *Science* **302**, 1234–1236 (2003).
14. Sliwa, J., Duhamel, J.-R., Pascalis, O. & Wirth, S. Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1735–1740 (2011).
15. Gothard, K. M., Battaglia, F. P., Erickson, C. A., Spitler, K. M. & Amaral, D. G. Neural responses to facial expression and face identity in the monkey amygdala. *J. Neurophysiol.* **97**, 1671–1683 (2007).
16. Landi, S. M. & Freiwald, W. A. Two areas for familiar face recognition in the primate brain. *Science* **357**, 591 (2017).
17. Chang, L. & Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013-1028.e14 (2017).
18. Perrodin, C., Kayser, C., Logothetis, N. K. & Petkov, C. Voice cells in primate temporal lobe. *Curr. Biol.* **21**, 1408–1415 (2011).
19. Quiroga, R. Q., Kreiman, G., Koch, C. & Fried, I. Sparse but not "grandmother-cell" coding in the medial temporal lobe. *Trends Cogn. Sci.* **12**, 87–91 (2008).
20. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
21. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
22. Miller, C. T. *et al.* Natural behavior is the language of the brain. *Curr. Biol.* **32**, R482–R493 (2022).
23. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]* (2016) doi:10.1145/2939672.2939785.

24. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
25. Nieh, E. H. *et al.* Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).
26. Beyeler, M., Rounds, E. L., Carlson, K. D., Dutt, N. & Krichmar, J. L. Neural correlates of sparse coding and dimensionality reduction. *PLoS Comput. Biol.* **15**, e1006908 (2019).
27. Seyfarth, R. M. & Cheney, D. L. Knowledge of Social Relations. in *The Evolution of Primate Societies* (eds. Mitani, J., Call, J., Kappeler, P. M., Palombit, R. & Silk, J. B.) 629–642 (University of Chicago Press, 2012).
28. Baraduc, P., Duhamel, J.-R. & Wirth, S. Schema cells in the macaque hippocampus. *Science* **363**, 635–639 (2019).
29. Landi, S. M., Viswanathan, P., Serene, S. & Freiwald, W. A. A fast link between face perception and memory in the temporal pole. *Science* **373**, 581–585 (2021).
30. Rolls, E. T. & Wirth, S. Spatial representations in the primate hippocampus, and their functions in memory and navigation. *Prog. Neurobiol.* **171**, 90–113 (2018).
31. Freiwald, W. A. Social interaction networks in the primate brain. *Curr. Opin. Neurobiol.* **65**, 49–58 (2020).
32. Sliwa, J. & Freiwald, W. A. A dedicated network for social interaction processing in the primate brain. *Science* **356**, 745–749 (2017).
33. Cléry, J. C., Hori, Y., Schaeffer, D. J., Menon, R. S. & Everling, S. Neural network of social interaction observation in marmosets. *Elife* **10**, e65012 (2021).
34. Mitchell, J. F., Reynolds, J. H. & Miller, C. T. Active vision in marmosets: a model system for visual neuroscience. *J. Neurosci.* **34**, 1183–1194 (2014).
35. Nummela, S. U., Jutras, M. J., Wixted, J. T., Buffalo, E. A. & Miller, C. T. Recognition Memory in Marmoset and Macaque Monkeys: A Comparison of Active Vision. *J. Cogn. Neurosci.* **31**, 1318–1328 (2019).
36. Erickson, R. G. & Dow, B. M. Foveal tracking cells in the superior temporal sulcus of the macaque monkey. *Exp. Brain Res.* **78**, 113–131 (1989).
37. Jutras, M. J. & Buffalo, E. A. Recognition memory signals in the macaque hippocampus. *Proceedings of the National Academy of Sciences* **107**, 401–406 (2010).
38. Rolls, E. T., Judge, S. J. & Sanghera, M. K. Activity of neurones in the inferotemporal cortex of the alert monkey. *Brain Res.* **130**, 229–238 (1977).
39. Rolls, E. T. *et al.* Hippocampal neurons in the monkey with activity related to the place in which a stimulus is shown. *J. Neurosci.* **9**, 1835–1845 (1989).
40. Solomon, S. S. *et al.* Visual motion integration by neurons in the middle temporal area of a New World monkey, the marmoset. *J. Physiol.* **589**, 5741–5758 (2011).
41. Ghazanfar, A. A. & Logothetis, N. K. Facial expressions linked to monkey calls. *Nature* **423**, 937–938 (2003).
42. Mitchell, J. F., Priebe, N. J. & Miller, C. T. Motion dependence of smooth pursuit eye movements in the marmoset. *J. Neurophysiol.* **113**, 3954–3960 (2015).
43. Miller, C. T., Mandel, K. & Wang, X. The communicative content of the common marmoset phee call during antiphonal calling. *Am. J. Primatol.* **72**, 974–980 (2010).
44. Miller, C. T. & Thomas, A. W. Individual recognition during bouts of antiphonal calling in common marmosets. *Journal of Comparative Physiology A* **198**, 337–346 (2012).
45. Courellis, H. S. *et al.* Spatial encoding in primate hippocampus during free navigation. (2019) doi:10.5061/dryad.kk63d49.
46. McMahon, D. B., Bondar, I. V., Afuwape, O. A., Ide, D. C. & Leopold, D. A. One month in the life of a neuron: longitudinal single-unit electrophysiology in the monkey visual system. *J. Neurophysiol.* **112**, 1748–1762 (2014).
47. Orbán, G. Intan RHD 2000 file of electrophysiological recordings. (2019) doi:10.17632/W767NNK5WH.

48. Yger, P. *et al.* A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *Elife* **7**, (2018).
49. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *ann. math. stat.* **18**, 50–60 (1947).
50. Ke, G. *et al.* LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
51. Narayan, A., Berger, B. & Cho, H. Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability. *bioRxiv* 2020.05.12.077776 (2020) doi:10.1101/2020.05.12.077776.
52. Medixant. *RadiAnt DICOM Viewer*. (2021).
53. Rosset, A., Spadola, L. & Ratib, O. OsiriX: an open-source software for navigating in multidimensional DICOM images. *J. Digit. Imaging* **17**, 205–216 (2004).
54. McMahon, D. B., Russ, B. E., Elnaiem, H. D., Kurnikova, A. I. & Leopold, D. A. Single-Unit Activity during Natural Vision: Diversity, Consistency and Spatial Sensitivity among AF Face Patch Neurons. *Journal of Neuroscience* **35**, 5537–5548 (2015).

Author Contributions.

T.T. performed the primary data analyses and wrote the paper. M.M. collected the data and assisted in analyzing some data. C.T.M. managed the project, designed the experiment and wrote the paper.

The authors declare no competing interests.

Materials and Correspondence should be directed to Cory Miller – corymiller@ucsd.edu

Data and Code are available at the Dryad repository: https://doi.org/10.5061/dryad.qnk98sfkv
The repository with downloadable files for Editors and Reviewers upon submission can be found here:
https://datadryad.org/stash/share/p3UkRDi05pvkMK4uEutemVsN8WZDfCbOk2L3bKTdfPg

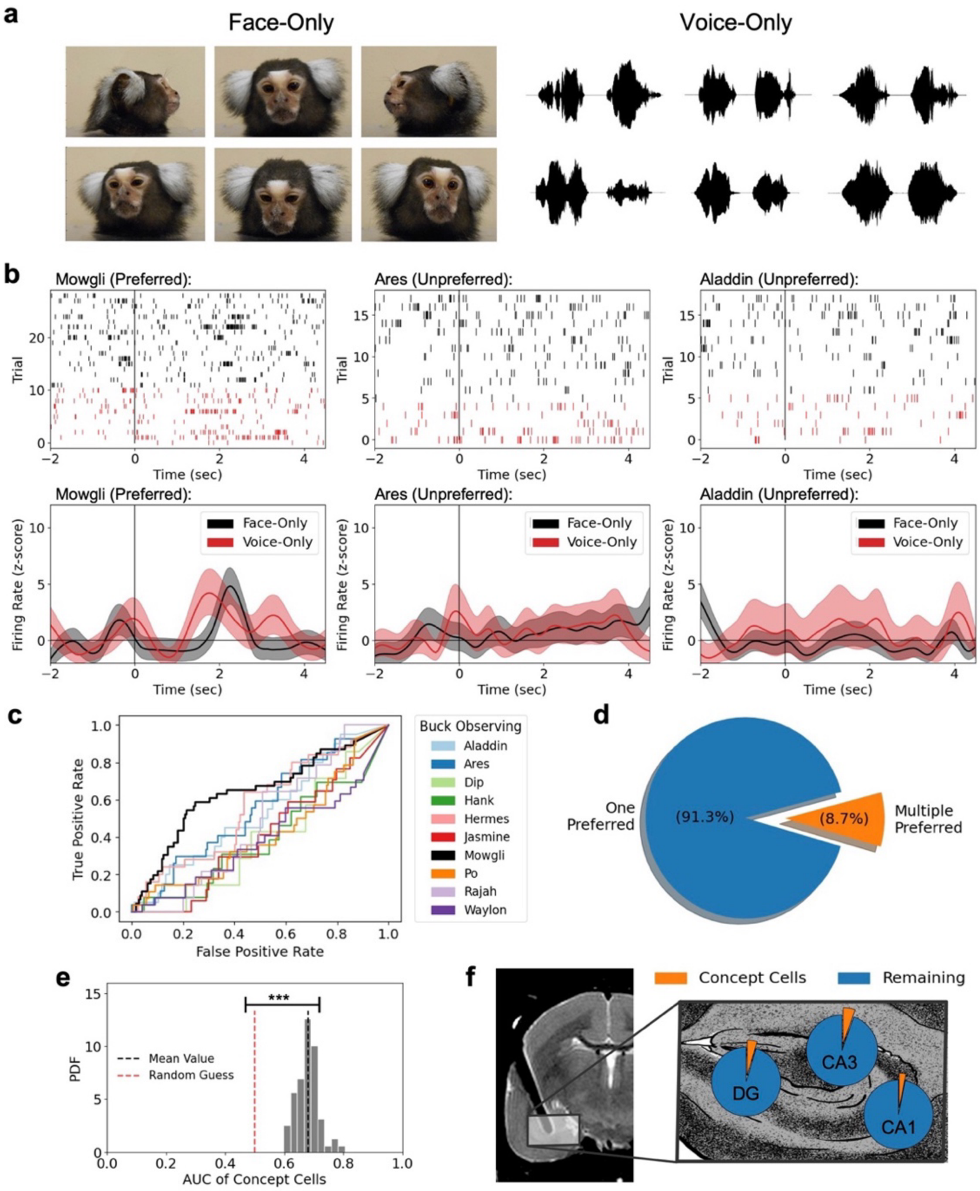


Figure 1. Putative concept cells in marmoset hippocampus. [a] Unimodal Face-Only and Voice-Only stimuli from a single marmoset included multiple pictures of (left) faces from different orientations and (right) multiple two pulsed phee calls. **[b]** Exemplar neuron responding selectively to face and voice of Mowgli (left) but not for (middle) Ares or (right) Aladdin during (black) face-only and (red) voice-only trials. Shown are (top) rasters of spike times and (bottom) mean change in firing rate relative to the 500ms prestimulus baseline firing rates. Indicated is (shaded) 95% confidence intervals of the mean estimated by bootstrap. **[c]** Receiver-Operator-Characteristic (ROC) curve of time-averaged firing rate of the same concept cell in response to multiple individuals. The preferred identity is (black) Mowgli, who is the sister of the observer, Baloo (AUC=0.643). **[d]** Pie chart of the proportion of putative concept cells that preferred (blue) one individual versus (orange) more than one individual. **[e]** Histogram of areas under the ROC curve (AUC) for individuals preferred by all putative concept cells in marmoset hippocampus. The (black dotted) mean value (AUC=0.680+/-0.009) was significantly larger than (red dotted) random chance according to a student's t-test ($p<0.001$, $N_{neurons}=67$). Unless otherwise stated,

uncertainty indicates 95% confidence intervals. **[f]** The anatomical distribution of (orange) putative concept cells in hippocampal subfields relative to (blue) all other neurons recorded.

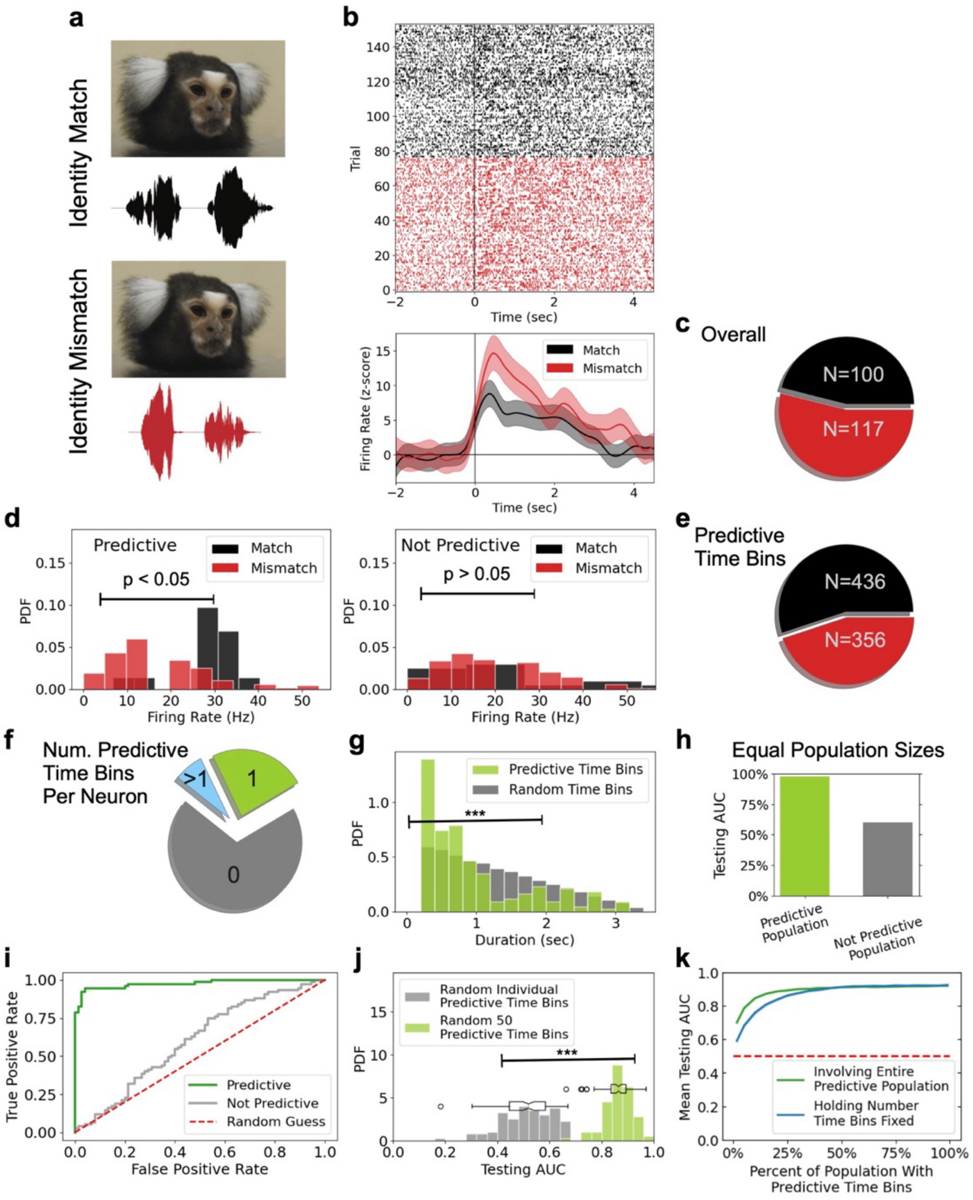


Figure 2. Single neurons and population activity in hippocampus represent multiple individuals. [a] Cross-modal stimuli for identity match (top) - face and voice stimuli are from a single monkey - and mismatch (bottom) - face and voice stimuli are from different monkeys- trials. Red waveform indicates voice stimulus from a different animal than one depicted in the face stimulus and black waveform voice stimulus above **[b]** Exemplar neuron showing increased firing rate during (red) identity mismatch trials than during (black) identity match trials. Shown is (left) the raster of spike times and (right) the mean change in firing rate relative to 500ms prestimulus baseline firing rates. **[c]** Pie chart showing the proportion of neurons with a significantly larger median firing rate for (black) identity match trials or for (red) mismatch trials during the 3 seconds immediately following onset of cross-modal stimulus ($p<0.05$, $N_{trials}\geq81$). Indicated are total counts out of all neurons recorded. **[d]** Histograms showing an exemplar neuron responding more for (black) identity match trials than for (red) identity mismatch trials during (left) a predictive time bin but not during (right) not a predictive time bin. **[e]** Pie chart showing the proportion of neurons with a significantly larger median firing rate for (black) match trials or for (red) mismatch trials during the 3 seconds immediately following onset of cross-modal stimulus

($p<0.05$, $N_{trials} \geq 81$). Indicated are total counts out of all neurons recorded. Neurons exhibiting multiple predictive time bins were permitted in both categories. **[f]** Pie chart showing the proportion of neurons exhibiting one or more predictive time bins (31.0%, N=732/2358). The majority of these predictive neurons exhibited (green) only one predictive time bin (76.8%, N=562/732). **[g]** Histograms of time bin durations are shown for (green) the predictive time bins and for (gray) uniformly distributed random time bins. Predictive time bins favored shorter durations according to a Wilcoxon-Mann-Whitney test ($p<0.001$, N=968). **[h]** Bar plot of testing AUC resulting from (green) our population-level decoder (AUC=0.977) and from (gray) an equal number of time bins randomly sampled from the non-predictive population (AUC=0.600). Time bins were aggregated over 14 recording sessions recorded from four different observers. **[i]** ROC curve resulting from the same (green) population-level decoder and from (gray) an equal number of time bins randomly sampled from the same non-predictive population. Indicated is (red dotted) random chance. **[j]** Histograms showing testing AUC for (gray) individual predictive time bins (median AUC=0.519, IQR: 0.445-0.584) was significantly smaller than (green) 50 predictive time bins (median AUC=0.864, IQR: 0.835-0.893) according to a Wilcoxon–Mann–Whitney test ($p<0.001$, $N_{samples}=100$). Independent random samples were drawn uniformly from the same predictive population without replacement. **[k]** Mean testing AUC versus relative abundance of predictive time bins averaged over independent random samples of time bins drawn from the same predictive and non-predictive populations without replacement ($N_{samples}=100$). Traces are shown for random samples with (green) all available predictive time bins (N=335) and for random samples with (blue) fixed total number of time bins (N=335). Uncertainty in the mean was less than 1% for both traces.

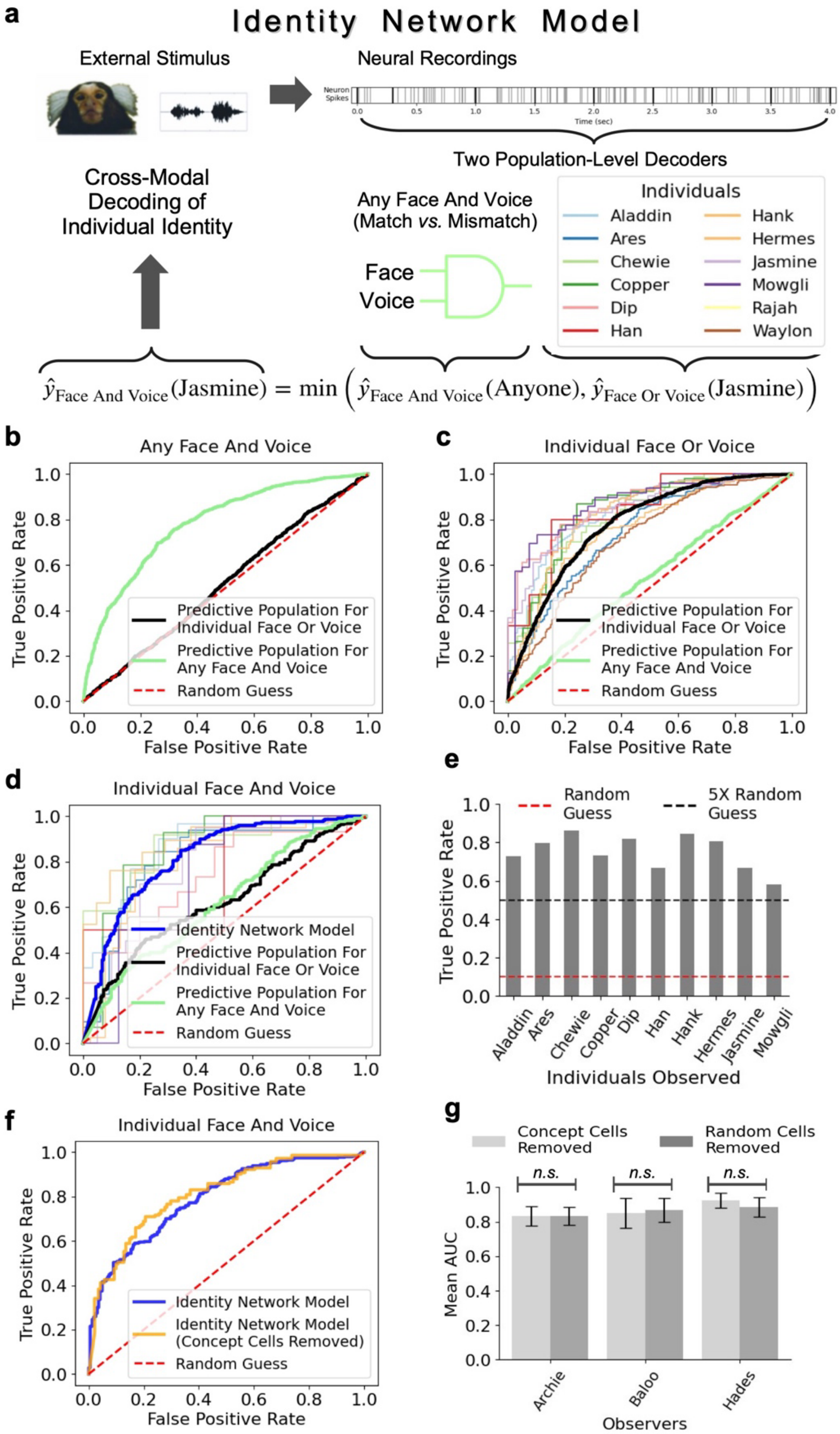


Figure 3 Cross-modal decoding of individual identity. [a] Schematic showing our identity network model (INM) predicting the presence of the face and voice of Jasmine as the minimum prediction returned by either the MvMM predictive population or the identity-specific predictive population. Identity-specific predictive populations were identified for each individual observed (N=12). Legend indicates colors corresponding to individuals.

[b] ROC curves for the detection of identity match trials. Firing rates were considered from (green) MvMM predictive time bins selected for their ability to detect any face and voice matching (AUC=0.782) and from (black) identity-specific predictive time bins selected for the detection of the face or voice of an individual (AUC=0.516). **[c]** ROC curves for the detection of face or voice of individuals. Firing rates were considered from (green) MvMM predictive time bins (AUC=0.536) and from (black) identity-specific predictive time (AUC=0.779), both averaged over individuals. Colored lines indicate results for individual identity-specific predictive populations. **[d]** ROC curves for the detection of both face and voice of individuals. Firing rates were considered from (green) MvMM predictive time bins (AUC=0.615), from (black) identity-specific predictive time bins (AUC=0.622), and from (blue) INM (AUC=0.818), similarly averaged over individuals. Colored lines indicate results of the INM for individuals. **[e]** Bar plot showing true positive rates predicted by the winner-take-all model, which considered predictions from the INM specific to ten individuals. Indicated is (red dotted) the true positive rate resulting from random chance. True positive rates were computed over testing identity match trials from 23 recording sessions conducted over three observers (N_{trials}=198). **[f]** ROC curves for the INM predicting the face and voice of individuals that were preferred by at least one concept cell. Shown is (orange) the INM with concept cells removed from consideration (AUC=0.810) and (blue) the INM with an equal number of predictive time bins randomly removed from both MvMM and identity-specific populations (AUC=0.800). **[g]** Bar plot showing mean testing AUC predicted by the INM for the same three observers. The mean AUC for (light gray) the INM with concept cells removed was not significantly different from that of (dark gray) the INM with an equal number of predictive time bins removed from decoding according to a paired student's t-test conducted over multiple successful sessions recorded from the same three observers: Archie (p=0.97, N=13), Baloo (p=0.18, N=3), and Hades (p=0.19, N=13). The median AUC similarly exhibited no significant difference according to comparable Wilcoxon-Mann-Whitney tests conducted on the same three observers: Archie (p=0.44, N=13), Baloo (p=0.49, N=3), and Hades (p=0.32, N=13). Predictions were generated independently for each recording session. Unless otherwise specified, ROC curves summarize the same 19 recording sessions conducted over the same three observers.

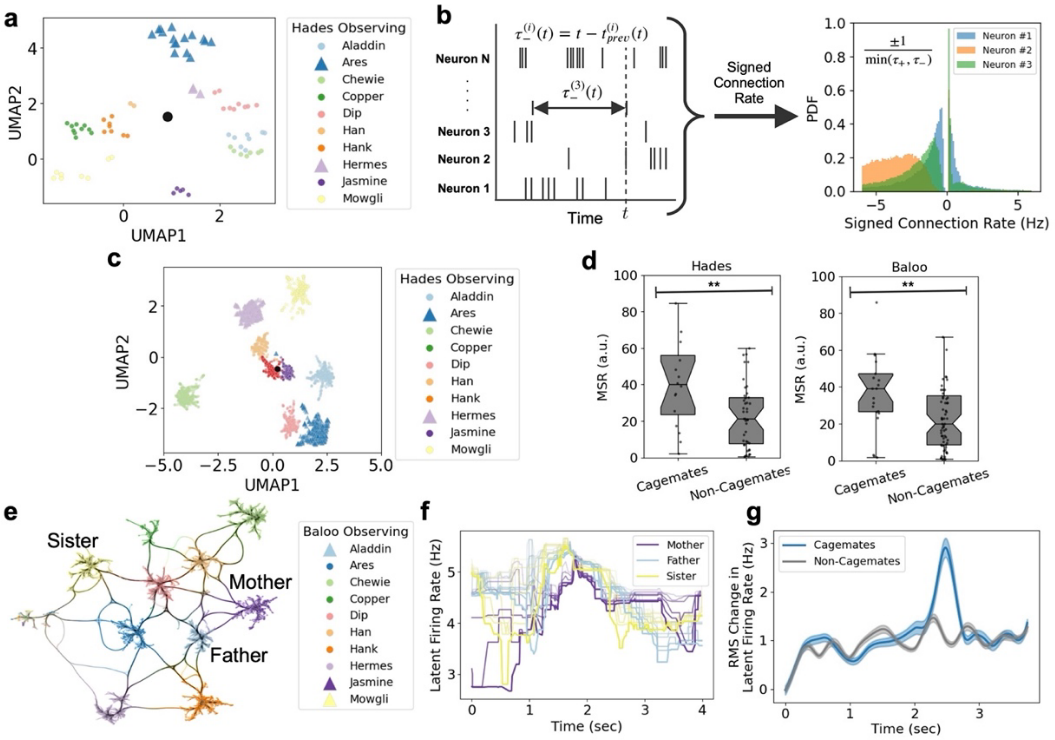


Figure 4 Cross-modal representation of identity using rate and event coded measures. [a] Two-dimensional manifold projection of our rate coded representation of individual identity computed from the firing rates of identity-specific predictive time bins. Time bins were aggregated from all individuals in one exemplar recording session. Manifold projections were estimated using nonlinear dimensionality reduction by UMAP. One dot represents one identity match trial. Indicated is (black) the mean. Colors correspond to individuals in the legend. Triangles indicate the individual was a cagemate with the observer, Hades. **[b]** Schematic illustrating how spike times were used to compute (left) the hindsight delay for a given neuron. The foresight delay was computed similarly by time inversion. The signed connection rate considered only the minimum of the foresight delay and the hindsight delay and took a negative value only if the hindsight delay was smaller than the foresight delay. While well-defined for all neurons at all times, the signed connection rate evaluated for a given neuron at the spike times of another neuron resulted in statistical distributions that appear to describe a relationship between the two neurons. Different morphologies of these statistical distributions were observed for different pairs of neurons, as is shown by (right) signed connection rate histograms of four example neurons in connection with the same reference neuron. **[c]** Two-dimensional manifold projection of our event coded representation of individual identity computed as the manifold projection of signed connection rates of all neurons identified in the same exemplar recording session. One dot represents one spike from one neuron. Indicated is (black) the mean. Colors correspond to individuals in the legend. Triangles indicate the individual was a cagemate with the observer, Hades. **[d]** Boxplot of MSR from the mean of our event coded representations. Median MSR was significantly larger when (left) Hades was observing the face and voice of cagemates versus non-cagemates (p=0.0025, N=60). Similarly, a significantly larger median MSR was observed when (right) Baloo was observing the face and voice of cagemates versus non-cagemates according to a Wilcoxon–Mann–Whitney test (p=0.0063, N=93). **[e]** Connectivity graph of the same event coded representation of individual identity for one recording session exemplar. Colors correspond to individuals in the legend. Indicated are family members of the observer who were also (triangles) cagemates of the observer, Baloo. Connections were bundled by dividing them into smaller edges and allowing those smaller edges to attract. **[f]** Latent firing rate trajectory is

shown each for Baloo's (yellow) sister Mowgli, (blue) father Aladdin, and (purple) mother Jasmine for the same recording session exemplar. For a given color, each line corresponds to one of the six dimensions of the latent firing rate trajectory. **[g]** Root mean square change in our latent firing rate was averaged over (blue) all cagemates and (gray) all non-cagemates presented to Baloo. Uncertainty indicates 95% confidence intervals of the mean as estimated by bootstrap. Results were averaged over the identity match trials from 12 recording sessions conducted on the observer, Baloo.

Supplementary Materials.

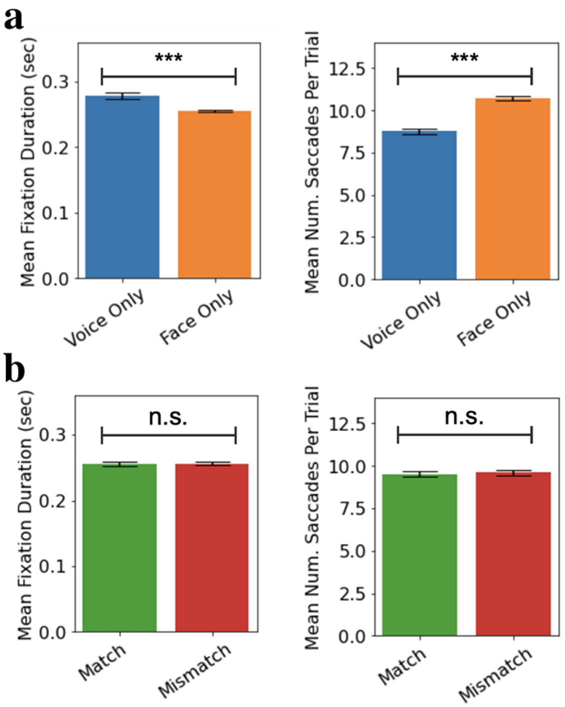
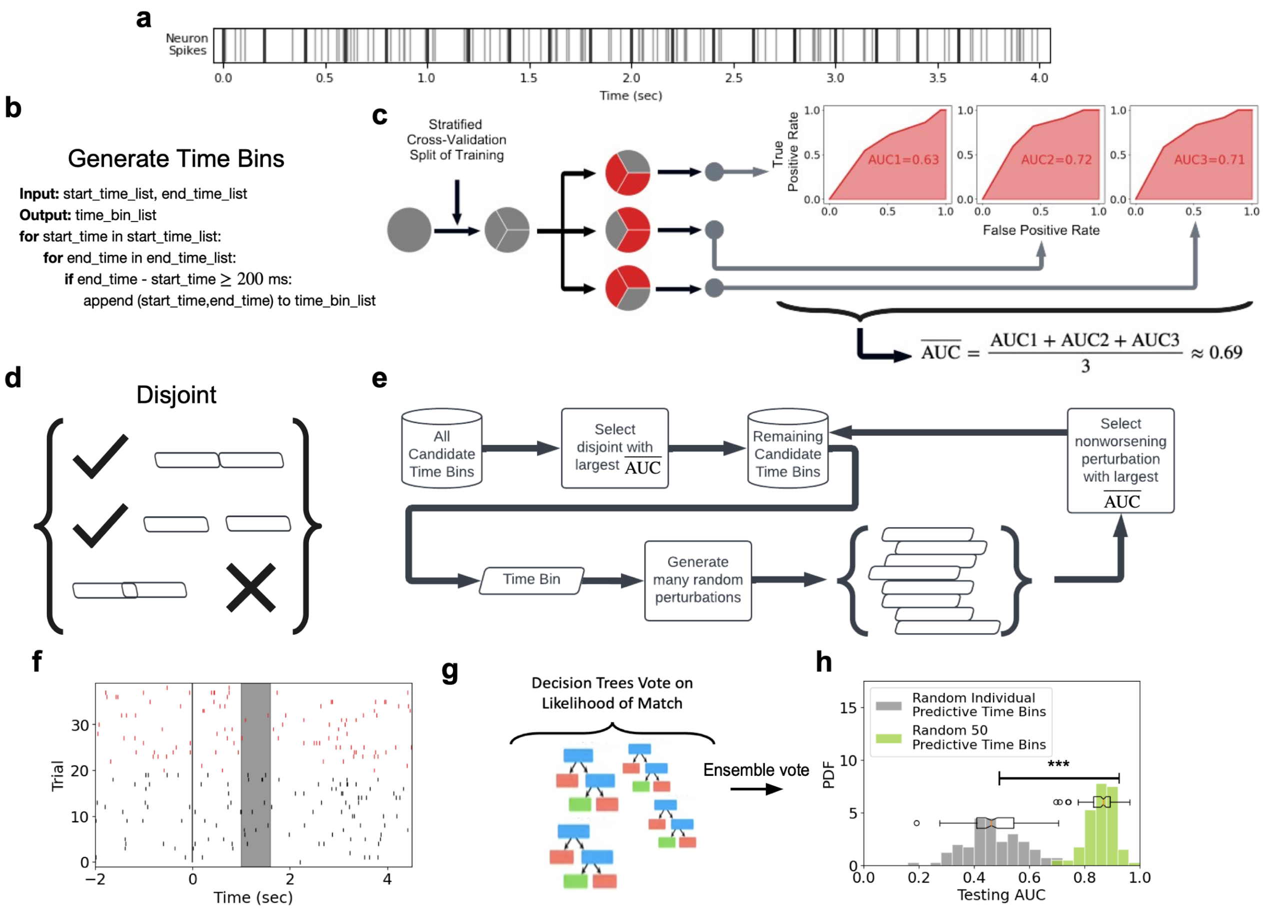


Figure S1. Visual behavior during unimodal and cross-modal stimulus presentations. [a] Barplots showing (left) the mean fixation duration per trial and (right) the mean number of saccades per trial. Indicated are (blue) voice-only trials and (orange) face-only trials. For voice-only trials, the mean fixation duration was 0.279+/-0.005 seconds, which was not significantly different from a normal distribution according to the omnibus test for normality (p=0.98, N=21,673). For face-only trials, the mean fixation duration was 0.256+/-0.002 seconds, which was not significantly different from a normal distribution according to an omnibus test (p=0.39, N=70,182). The voice-only trials exhibited a statistically significantly larger mean fixation duration compared to the face-only trials according to a one-sided student's t-test (p<0.001, N≥21,673). For the voice-only trials, the mean number of saccades per trial was 8.76+/-0.17, which was not significantly different from a normal distribution according to an omnibus test (p=0.22, N=2,333). For the face-only trials, the mean number of saccades per trial was 10.70+/-0.10, which was not significantly different from a normal distribution according to an omnibus test (p=0.24, N=6,226). The voice-only trials exhibited a statistically significantly larger mean number of saccades per trial compared to the face-only trials according to a one-sided student's t-test (p<0.001, N≥21,673). **[b]** Bar plots showing (left) the mean fixation duration per trial and (right) the mean number of saccades per trial. Indicated are (black) identity match trials and (red) identity mismatch trials. For identity match trials, the mean fixation duration was 0.256+/-0.003 seconds, which was not significantly different from a normal distribution according to an omnibus test (p=0.05, N=33,955). For identity mismatch trials, the mean fixation duration was 0.256+/-0.003 seconds, which was not significantly different from a normal distribution according to an omnibus test (p=0.70, N=35,475). The identity match trials exhibited no statistically significant different mean fixation duration compared to the identity mismatch trials according to a student's t-test (p=0.85, N≥33,955). For identity match trials, the mean number of saccades per trial was 9.51+/-0.15, which was not significantly different from a normal distribution according to an omnibus test (p=0.92, N=3,369). For identity mismatch trials, the mean number of saccades per trial was 9.59+/-0.14, which was not significantly different from a normal distribution according to an omnibus test (p=0.09, N=3,496). The identity match trials exhibited no statistically significant different mean number of saccades per trial compared to the identity mismatch trials according to a student's t-test (p=0.44, N≥3,369). Unless otherwise stated, uncertainty indicates 95% confidence intervals of the average approximated via bootstrap


a
Neuron Spikes
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Time (sec)
b
Generate Time Bins
c
Stratified Cross-Validation Split of Training
True Positive Rate
1.0
0.5
0.0
AUC1=0.63
AUC2=0.72
AUC3=0.71
0.0
0.5
1.0
False Positive Rate
d
Disjoint
e
All Candidate Time Bins
Select disjoint with largest $\overline{\text{AUC}}$
Remaining Candidate Time Bins
Select nonworsening perturbation with largest $\overline{\text{AUC}}$
Time Bin
Generate many random perturbations
f
Trial
30
20
10
0
−2
0
2
4
Time (sec)
g
Decision Trees Vote on Likelihood of Match
Ensemble vote
h
Random Individual Predictive Time Bins
Random 50 Predictive Time Bins
***
PDF
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
Testing AUC

```
Input: start_time_list, end_time_list
Output: time_bin_list
for start_time in start_time_list:
    for end_time in end_time_list:
        if end_time - start_time ≥ 200 ms:
            append (start_time,end_time) to time_bin_list
```

$$\overline{\text{AUC}} = \frac{\text{AUC1} + \text{AUC2} + \text{AUC3}}{3} \approx 0.69$$

Figure S2. Identification of predictive time bins. [a] Schematic showing the spike times of an example neuron firing versus time after the stimulus onset at t=0. Indicated are (black) start and end times. **[b]** Pseudocode showing the method of generating time bins of variable duration. **[c]** Flow chart showing training trials being split by stratified cross-validation to result in multiple receiver operator characteristic (ROC) traces. Each training fold resulted in an area under the curve (AUC), which were then averaged to produce the mean training AUC as an estimator of the general ability of a time bin to distinguish true trials from false trials. Time bins satisfying a list of properties were considered as candidate time bins (described in Methods). **[d]** Schematic showing our definition of disjoint time bins. Time bins are disjoint if and only if they share no time interval in common. **[e]** Flow chart showing the procedure that resulted in all predictive time bins (described in Methods). **[f]** Spike raster for an exemplar neuron showing a response to (red) identity mismatch and (black) identity mismatch trials. Indicated is (shaded) an exemplar predictive time bin. **[g]** Schematic showing decision trees voting on the likelihood of an identity match trial, resulting in the predictions of our population-level decoders. **[h]** Histograms of testing AUC values are shown for (gray) random individual predictive time bins and (green) 50 randomly selected predictive time bins, exhibiting a statistically significant difference of median value according to a Wilcoxon-Mann-Whitney test (p<0.001, N=100). Predictive time bins for the MvMM binary classification task were randomly considered from multiple recording sessions (N=14).

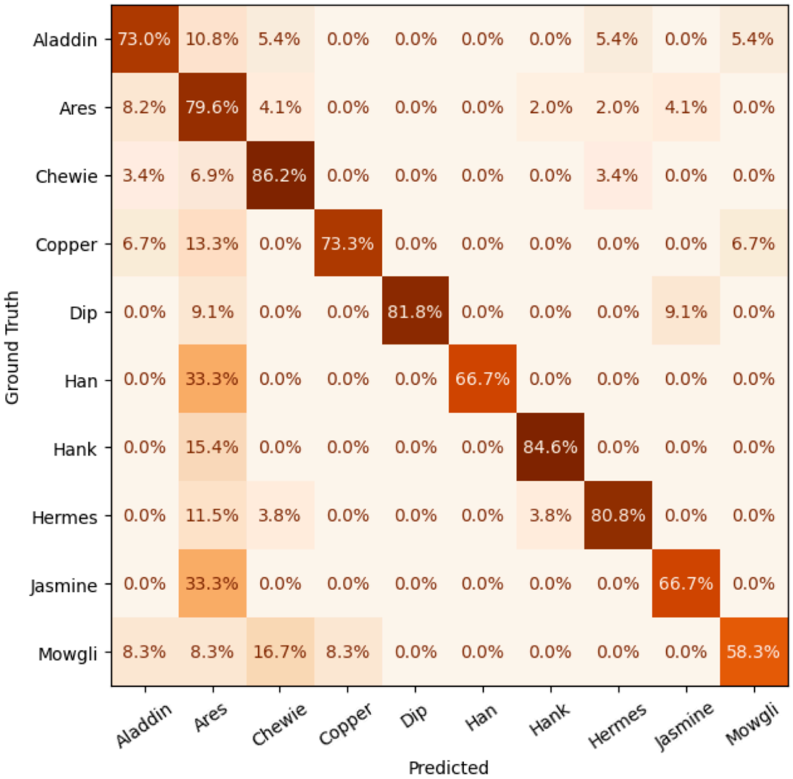


Figure S3. Multiple individuals classified by winner-take-all model. Confusion matrix reporting the winner-take-all predictions of the INM on ten individuals shown to three observers over 23 recording sessions (testing accuracy=0.89, sensitivity=0.91, specificity=0.87, precision=0.91, negative predictive value=0.87, N=198 identity match trials). The winner-take-all model is described in Methods.

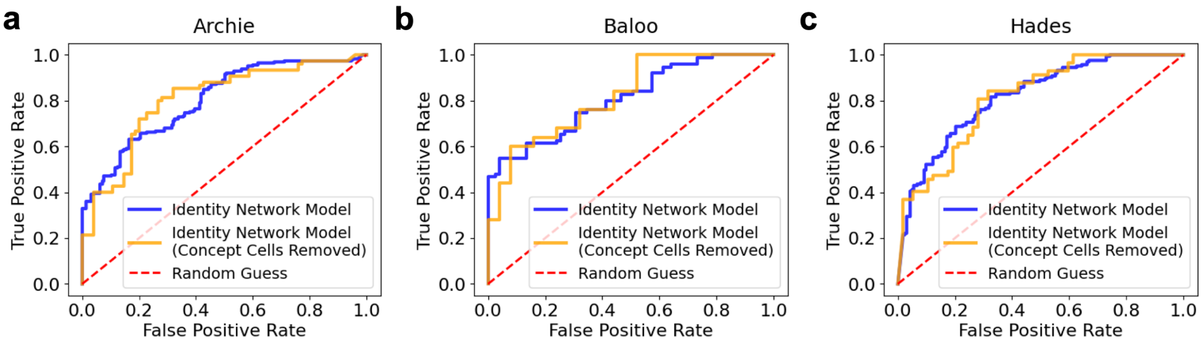


Figure S4. Identity network model for each individual subject. ROC curves were computed by averaging over all recording sessions for each of three observers **[a]** Archie, **[b]** Baloo, and **[c]** Hades. These ROC curves demonstrate the predictive power of our INM both with (blue) all cells considered and with (orange) all concept cells removed. Individual identities were averaged over if they were preferred by at least one concept cell. We controlled for network size by removing the same number of cells from both ROC curves. We did this for both the MvMM predictive population and the identity-specific predictive population in evaluating the INM.

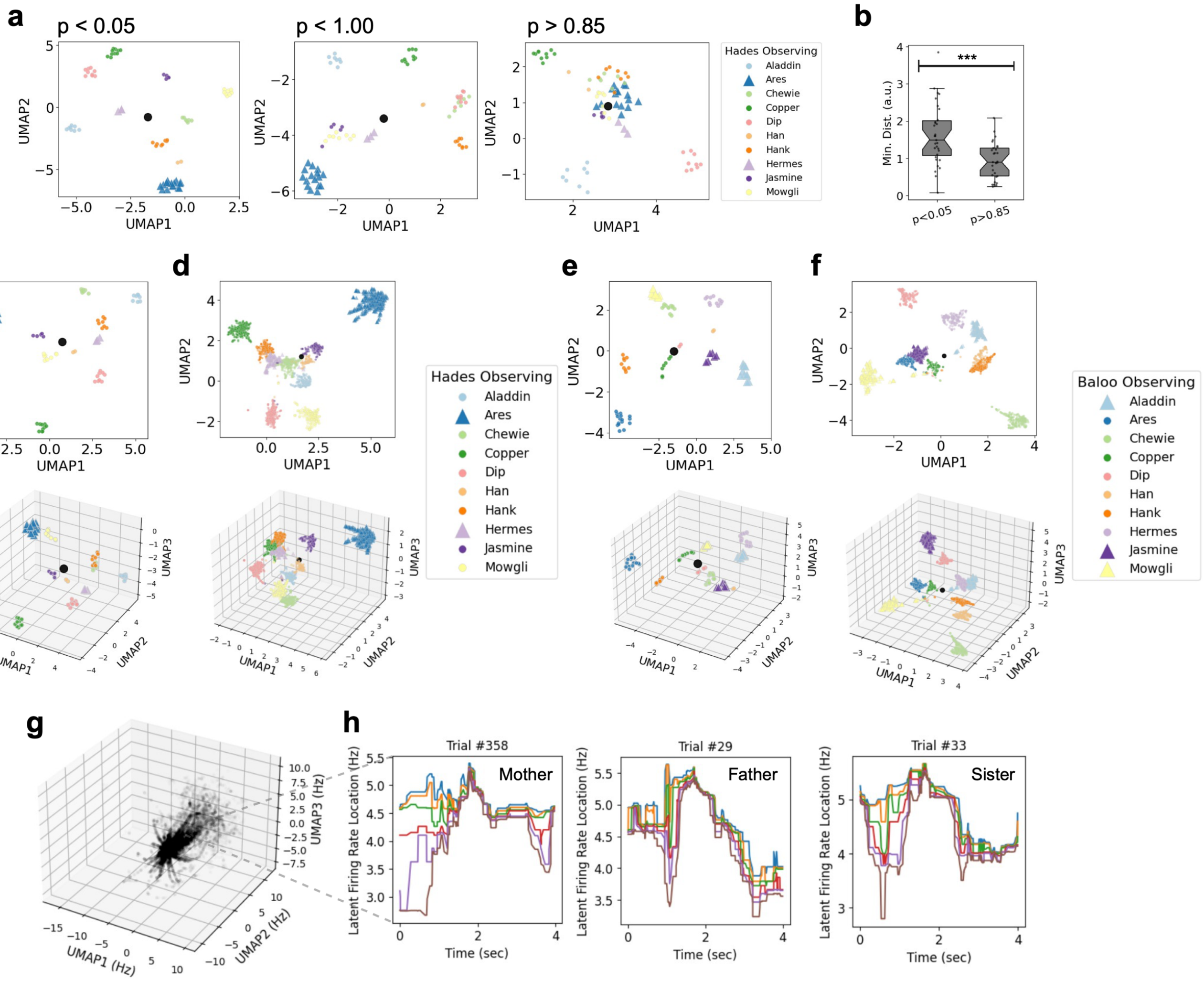


Figure S5. Low-dimensional projections of our rate code and event code.
[a] Scatter plot showing an exemplar recording session as two-dimensional rate-coded representations of individual identity, where the firing rates were computed from all candidate time bins exhibiting (left) p<0.05, (middle) p<1.00, and (right) p>0.85. **[b]** Box-and-whisker plots showing the minimum distance between any individual in our rate-coded representation of individual identity. The median minimum distance of (left) p<0.05 was significantly smaller than the median minimum distance of (right) p>0.85 according to a Wilcoxon-Mann-Whitney test (p<0.001, N_{sessions}=29). **[c-f]** Shown are the (top) first two axes and (bottom) first three axes of our representations of individual identity for two distinct observers: **[c,d]** Hades and **[e,f]** Baloo. **[c,e]** Shown are manifold projections of our predictive time bins and **[d,f]** our signed connection rate. Colors indicate individuals, and triangles indicate family members. The signed connection rate was evaluated no more than two seconds after stimulus onset, which was evaluated whenever the neuron with the largest overall spike count fired. **[g]** Shown are the first three axes of our six dimensional latent firing rate, which was an unsupervised manifold projection of the absolute value of the signed connection rate from the same neuron with the largest overall spike count – the reference neuron – to all neurons that appeared approximately symmetric (defined in Methods). **[h]** Shown are time traces of our latent firing rate for an exemplary trial from each of three family members of Baloo. Each color represents one dimension. The order of dimensions is consistent between panels.

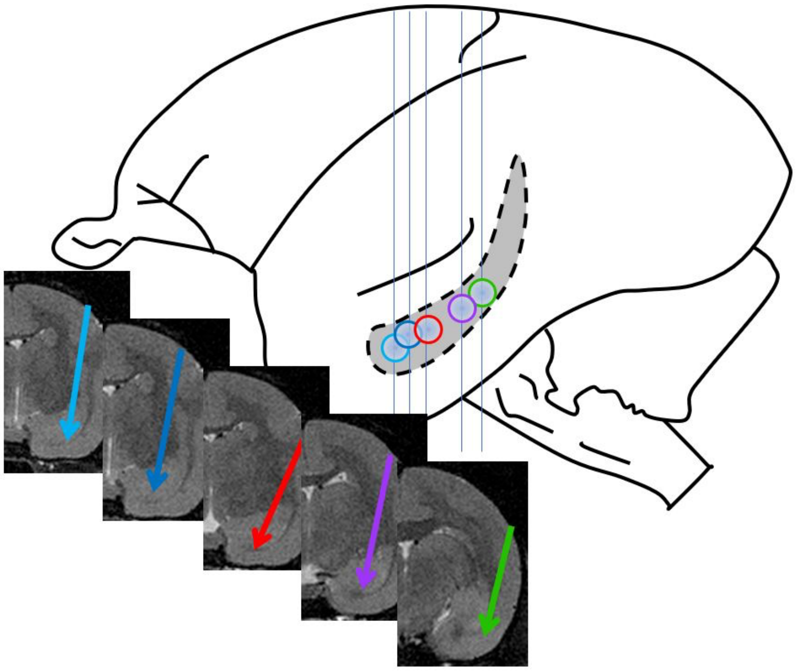


Figure S6. Anatomical locations of microwire bundles across animals. Arrows on MRI indicate trajectory of each MBA in marmoset hippocampus. Each color indicates a different animal's array.   Circles correspond to AP position.