Abstract representations emerge in human hippocampal neurons during inference behavior

Hristos S. Courellis[1,2,*], Juri Mixha[1,2,*], Araceli R. Cardenas[3], Daniel Kimmel[4], Chrystal M. Reed[5], Taufik A. Valiante[3], C. Daniel Salzman[4], Adam N. Mamelak[1], Ralph Adolphs[2], Stefano Fusi[4,#], Ueli Rutishauser[1,2,5,6,#]

[1] Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[2] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA

[3] Krembil Research Institute and Division of Neurosurgery, University Health Network (UHN), University of Toronto, Toronto, ON, Canada

[4] Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, NY, USA

[5] Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[6] Center for Neural Science and Medicine, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA

*=Joint first authors, #=joint senior authors.

Abstract:

Humans have the remarkable cognitive capacity to rapidly adapt to changing environments. Central to this capacity is the ability to form high-level, abstract representations that take advantage of regularities in the world to support generalization[1,2]. However, how these representations are encoded in populations of neurons, how they emerge through learning, and how they relate to behavior remains unknown[3]. Here we characterized the representational geometry of populations of neurons (single-units) recorded in the hippocampus, amygdala, medial frontal cortex, and ventral temporal cortex of neurosurgical patients who are performing an inferential reasoning task. We find that only the neural representations formed in the hippocampus simultaneously encode multiple task variables in an abstract, or disentangled, format. This representational geometry is uniquely observed after patients learn to perform inference, and consisted of disentangled directly observable and discovered latent task variables. Interestingly, learning to perform inference by trial and error or through verbal instructions led to the formation of hippocampal representations with similar geometric properties. The observed relation between representational format and inference behavior suggests that abstract/disentangled representational geometries are important for complex cognition.

Introduction:

Humans have a remarkable capacity to make inferences about hidden states that describe their environment[1,4,5] and use this information to adjust their behavior. One core cognitive function that enables us to perform inference is the construction of abstract representations of the environment. Abstraction is a process through which relevant shared structure in the environment is compressed and summarized, while superfluous details are discarded or represented so that they do not interfere with the relevant ones[6]. This process often leads to the discovery of latent variables that parsimoniously describe the environment. By performing inference on the value of these variables, frequently from partial information, the appropriate actions for a given context can rapidly be deployed[7–9], thereby generalizing from past experience to novel situations. For example, a latent variable specifying being in a left- or right-driving country can be used by a pedestrian to infer which way to look for oncoming traffic when crossing a road, even in the absence of a sensory cue such as traffic moving in that pedestrian's field of view, and when crossing roads they have never before encountered. Through abstraction, the common, underlying structure of the world is represented in a way that facilitates adaptive behavior.

How are latent variables encoded in the human brain? How are existing representations modified to accommodate them? Does their format matter for behavior? At the level of neuronal activity, the answers to such questions have remained elusive. Prior research has implicated the hippocampus in the implementation of abstraction and inference-related computations, both through neuroimaging in humans during tasks that require abstraction and generalization[1,7,10,11], and through neurophysiology in rodents and non-human primates engaged in tasks with abstract spatial and non-spatial components[3,12–15]. To date, relatively few studies have explored the role that the format, or geometry, of task variable representations plays in shaping computation in the human brain at the single neuron level[16,17], and no study has, to our knowledge, reported the emergence and manipulation of this geometry on the short timescales that would be required for rapid inference in humans.

We recorded the activity of large populations of neurons in the brains of awake, behaving epilepsy patients to study the emergence of abstract representations. Patients performed a serial reversal learning task with two latent contexts distinguished by different stimulus-response-outcome maps. We find that as patients learned to perform inference on the latent context, a representation of context emerged in an abstract format in the hippocampus, being disentangled from simultaneously represented stimulus information (i.e. represented in an approximately orthogonal subspace of the neural activity space). The emergence of the abstract context variable was correlated with an individual's ability to rapidly perform inference on the state of the latent variable context and was absent during error trials. Furthermore, we found that this abstract hippocampal context representation could emerge in two ways: by learning through experience and through verbal instructions informing patients about the latent structure of the task.

Results:

Humans perform inference in a context dependent task.

Patients completed a deterministic trial-based serial reversal learning task[3] with two blocked alternating latent contexts (Fig. 1A). Each of the two latent contexts is defined by a unique stimulus-response-outcome map (SRO map, Fig.1B,C), associating each of the four stimuli with a correct response (Left/Right button press on a button box) and a monetary reward level (outcomes) that is received deterministically after a correct response. The stimuli used are four different images (Fig. 1C, left). The two SRO maps are related since all stimulus-response pairings are inverted between the two (see Fig. 1C). With this design, an individual performing inference can detect a change in latent context after receiving feedback that their response was incorrect in a single trial and immediately update their stimulus-response associations for the remaining stimuli even though they have not yet been encountered in the new context. We refer to the trials in which a given stimulus is encountered for the first time following a covert context switch as inference trials and to the remaining trials as non-inference trials.

Patients (17 total, 3 male) completed 42 sessions (180-320 trials/session, 10-16 blocks/session) of the task, typically in pairs of two back-to-back sessions on the same recording day (mean = 2.4 sessions per day, min = 2, max = 4, see Table 1). Novel stimuli were used in every session, thus requiring patients to re-learn the SRO maps through trial and error at the start of every session. Of the 42 sessions, 6 were excluded from analysis due to at-chance performance in non-inference trials (Binomial Test, $p > 0.05$), indicating that patients did not learn the SR maps (Fig. 1B,C). Performance on non-inference trials was well above chance for the remaining 36 sessions (Fig. 1D). Each of the 36 included sessions was classified as either "Inference Present" (IP) or "Inference Absent" (IA) sessions depending on whether the patient performed significantly above chance on the first of the three possible inference trials occurring after context switches (Fig. 1E). Our task is designed such that by performing inference, patients can respond correctly the first time they see an image in a new context following the initial error trial (seen as significantly below chance performance in Fig. 1B at time point 1). This can be achieved by patients flexibly updating the currently active SRO map immediately after encountering an error, thereby allowing them to perform accurately for the remaining three stimuli that had not yet been seen in the new context. We took accuracy on the first of these three opportunities (the first inference trial) following a context switch as the behavioral signature of successful behavioral inference (timepoint 2 in Fig. 1E, Binomial Test, $p < 0.05$). Block-wise estimates of task performance for IA (Fig. 1F) and IP (Fig. 1G) sessions reveal that during IA sessions, patients exhibit poor inference performance after every context switch throughout the task, although the performance at the end of every block is high. In contrast, during IP sessions, inference performance rapidly rose over the first few blocks and remained high throughout the duration of the session. Note that within a given session, the two latent contexts had identical stimuli, responses, and outcomes; the only difference was the respective SRO associations. Correspondingly, subject-level accuracies (Fig. S1B) and reaction times (Fig. S1C) for the two contexts (arbitrarily labeled 1 and 2 across sessions) were not significantly different, indicating that there was no systematic performance bias for one of the two contexts.

When first performing the task, patients were told that they needed to learn arbitrary stimulus-response associations that would change over time, but were not informed about the latent contexts and their related structure. Sessions were recorded in back-to-back pairs with verbal instructions (see Methods) detailing the latent structure of the task provided during the inter-session period (mean length of break = 241 s, range 102-524s), which was considerably shorter than the sessions themselves (mean = 1154 s, range 898-1900s). We next considered whether patients could discover the latent task structure before receiving instructions, and if not, whether verbal instructions successfully shaped behavior. Patients were split into three groups: An "instruction successful" (IS) group, which is composed of patients who did not perform inference (IA) during pre-instruction sessions and who did perform inference (IP) during post-instruction sessions (5 patients, 10 sessions, Fig. S1D); An "instruction unsuccessful" (IU) group, which were patients who did not perform inference (IA) during both the pre-and post-instruction sessions (4 patients, 8 sessions, Fig. S1E); and a "session one interference" (S1) group, which were patients who exhibited inference behavior (IP) during both pre- and post-instruction sessions (3 patients, 6 sessions, Fig. S1F). Only patients who performed accurately in non-inference trials in both pre-and post instruction sessions were included in one of these three groups (Fig. S1D-F, "last"; 5 patients excluded, see Table 1). Thus, patients exhibited a variety of inference behaviors, with some exhibiting inference after receiving instructions (IS), some failing to exhibit inference after the instructions (IA), and some exhibiting inference even before receiving the instructions (S1). Below, we contrast the neural representations between these groups of patients to examine how instructions shape neural representations on short timescales.

Electrophysiology and analysis approach

Neural data recorded over the 36 (of 42) included sessions yielded 2694 (of 3124) well isolated single-units, henceforth neurons, distributed across the hippocampus (HPC, 494 neurons), amygdala (AMY, 889 neurons), pre-supplementary motor area (preSMA, 269 neurons), dorsal anterior cingulate cortex (dACC, 310 neurons), ventromedial prefrontal cortex (vmPFC, 463 neurons), and ventral temporal cortex (VTC, 269 neurons) (Fig. 1H,K). Only well isolated neurons as assessed by rigorous spike sorting quality metrics were included (see methods). Action potentials discharged by these neurons were counted during two 1s long trial epochs: during the baseline period (base, -1s to 0s prior to stimulus onset), and during the stimulus period (stim, 0.2s to 1.2s after stimulus onset). For the stimulus period, since patients would sometimes respond before 1.2s (reaction time = 1.08 ± 0.04s over sessions), we determined that 75.15% of all spikes occurred before a response was provided across all recorded neurons, indicating that analyses performed with these spike counts predominantly, but not exclusively, reflect pre-decision processing. Single neuron responses during these periods were heterogeneous. During the stimulus period, some neurons exhibited selectivity to one or several of the four variables stimulus identity, response, (predicted) outcome, and context that we examined (Fig. 1I-K, example PSTHs of stimulus identity, response, and context coding neurons). Other neurons were modulated by combinations of these variables (Fig. 1L, PSTH of a neuron with conjunctive stimulus and context coding). Across all brain areas, 54% of units (1447/2694) exhibited some form of task modulation, with 26% of units (706/2694) exhibiting only interaction effects, 17% (449/2694) exhibiting only main effects, and 11% (292/2694)

exhibiting both when fitting a 3-Way ANOVA for Response, Context, and Outcome (Fig. 1N, RCO column, chance = 135/2694 units, factor significance at $p < 0.05$). When neurons were separated into those recorded from IA and IP sessions, we found that anywhere between 5% and 15% of neurons in IA and IP sessions are significant for each of the main and interaction effects of the 3-Way ANOVA, with significant reductions in the proportion of neurons tuned to Outcome, Response x Context, and Response x Context x Outcome (Fig. S1M, $p_O = 0.0007$, $p_{RxC} = 0.0395$, $p_{RxCxO} = 0.0048$, two sample z-test) in IP sessions compared to IA sessions. Similar analyses conducted on separate 2-Way ANOVA for Stimulus Identity and Context (Fig. 1N, SC column, Fig. S1N, $p_S = 0.0165$, two sample z-test comparing IA with IP sessions), and for Stimulus Identity and Response (Fig. 1N, SR column, Fig. S1O, $p_S = 0.0287$), revealed a significant decrease in the fraction of neurons tuned to Stimulus Identity in IP compared to IA, again with no significant changes in the proportion of neurons coding for Context, Response, or interactions. These findings indicate that, while there is diverse tuning to many task variables simultaneously across all brain regions, there is no obvious increase in univariate tuning at the level of single neurons pooled across all areas that correlates with the ability to perform inference.

Measures of Neural Population Geometry

Given the heterogenous nature of the response pattern at the single neuron level (also see Fig. S1G-L), we adopted a population-level approach. This approach allows us to assess which variables are encoded in the distributed neural patterns, considering also the correlations of the neural responses across multiple conditions. Most importantly, it enables us to say significantly more about how these variables are represented, and, in particular, to study the geometry of the neural representation, which has important computational implications. The geometry of a representation is defined by the arrangement of the eight points in the activity space that represent the experimental conditions. Low dimensional geometries (e.g., when the eight points define a cube) are typically the result of a process of abstraction and confer to a linear readout the ability to generalize. For example, consider a simplified situation in which there are only two stimuli in two contexts, each associated with a different response (Fig. 2A-C). Imagine that the 4 points (2 per context) are arranged on a relatively low dimensional square (the maximal dimensionality for 4 points is 3), with context encoded along one side and the response along one of the two orthogonal sides (Fig. 2A). Then, a linear decoder for context, trained only on the conditions for which the response is Right, can readily generalize to the Left response conditions (Fig. 2B). This ability to generalize is due to the particular arrangement of the points, which make the coding direction of context for the Left responses parallel to the context coding direction for the Right responses (Fig. 2C). Also, in this case, context and response are represented in orthogonal subspaces, and hence, they are called disentangled variables[18,19]. In the square example, the ability of a linear decoder to generalize across conditions (cross condition generalization) also applies to the variable response (i.e. a response decoder trained on context 1 conditions will generalize to context 2 conditions). In previous work[3], we used the cross-condition generalization performance (CCGP) as the defining characteristic of an abstract representation of a variable. Indeed, when context CCGP is high, as in the square example, there is a subspace (the 1D space defined by the context coding direction) in which the representation

of context does not depend on the response, or in other words, it is disassociated from the specific instances of the response (Left or Right). In this case we would say that context is represented in an abstract format, or more synthetically that the representation of context is abstract.

Notice that if the 4 points of the example are at random locations in the activity space defining a tetrahedron, the representation is "unstructured" and does not have any of the generalization properties we described. On the other hand, these high dimensional representations allow a linear readout to separate (or shatter) the points in any arbitrary way, and hence confer to the readout the flexibility of implementing any possible task. The number of ways the points can be separated into two groups by a linear decoder (dichotomies) is called shattering dimensionality (SD)[3,20]. Recorded neural representations can have both the generalization properties of the abstract representations, and the flexibility of the high dimensional representations[3].

We constructed neural pseudopopulations of all recorded neurons in each brain area, and then we systematically analyzed and compared the representational geometry in IA and IP sessions. We computed for each relevant variable the decoding accuracy, which tells us whether a variable is encoded at all or not, and the CCGP, which informs us about the specific representational geometry. The CCGP is a cross-validated quantity estimated training and testing decoders on single trials, and sometimes, the available data is not sufficient to reveal a low dimensional structure. So, we complemented this analysis with the parallelism score (PS), which measures the parallelism of the coding directions of a specific variable. The coding directions are estimated using the average activity for each condition. A high PS strongly indicates that the variable is represented in an abstract format. The PS is also a more direct geometrical measure that focuses on the structure of the representation (the CCGP depends on the geometry but also on the noise and its shape).

To perform the analysis in an unbiased manner, we did not consider only the variables defining the task, but all the variables that correspond to the 35 possible dichotomies of the 8 conditions (see Fig. S2 for illustration of the five dichotomies that correspond to specific task variables, Table 2).

<u>Hippocampal neural geometry correlates with inference behavior</u>.

We first examined the the decodability of each balanced dichotomy in different brain areas for sessions where inference was present (IP) and sessions where inference was absent (IA). Following stimulus onset, the hippocampal neural population exhibited a significant increase in average decodability across all balanced dichotomies in IP sessions relative to IA sessions, indicative of an increase in shattering dimensionality (Fig. 2D, IA vs IP $p_{RS} = 2.7x10^{-3}$, Ranksum over dichotomies). Latent context (Fig. 2D, red, IA vs. IP $p_{RS} = 2.9x10^{-27}$, $p_{IA} = 0.12$, $p_{IP} = 5.1x10^{-5}$; p_{IA} and p_{IP} are significance tests vs. chance and p_{RS} a pairwise comparison between IA and IP) and stim pair (Fig. 2D, purple, IA vs. IP, $p_{RS} = 5.0x10^{-27}$, $p_{IA} = 0.015$, $p_{IP} = 7.9x10^{-7}$) emerged as the most strongly decodable named dichotomies in the IP sessions. The stim pair dichotomy corresponds to a grouping of stimulus

identities that elicit the same response, a relationship that remains static across contexts by construction of the SRO maps (Fig. 1C, S2). This difference in representation between IA and IP sessions is unique to HPC, as no other recorded region exhibited a significant change in decodability of the variable context (Fig. 2F, red). Rather, in vmPFC stim pair (Fig. 2F, purple, IA vs. IP, $p_{RS} = 6.9x10^{-20},\ p_{IA} = 0.21, p_{IP} = 0.0097$) and in preSMA response (Fig. 2F, green, IA vs. IP, $p_{RS} = 5.2x10^{-13}, p_{IA} = 0.091, p_{IP} = 0.0057$) increased significantly in decodability in IP compared to IA sessions, which is as expected (see discussion). The stim pair variable was also significantly decodable in AMY both during IA and IP, with no significant difference between the two (Fig. 2F, purple, IA vs. IP, $p_{RS} = 0.88, p_{IA} = 0.0016, p_{IP} = 0.0018$). The global increase (that is, averaged across all dichotomies) in shattering dimensionality of the HPC representation can also be quantified by measures that probe for non-linear interactions of variables in the population. The decodability of the parity dichotomy is one such metric, since decoding parity requires a high degree of non-linear encoding of variables in a neural population (see methods). We observed that in the HPC but not in other brain areas, parity decodability increased significantly in IP relative to the IA sessions (Fig. 2D, orange, IA vs. IP, $p_{RS} = 1.5x10^{-21}, p_{IA} = 0.27, p_{IP} = 0.0055$). Generalizing this finding, dividing different dichotomies into increasing levels of "difficulty", with more difficult dichotomies requiring stronger non-linear interactions of task variables to shatter, reveals that average SD is highest for the most difficult dichotomies in the HPC (Fig. S7). Together, these findings suggest that non-linearities in the HPC population response in the IP relative to the IA group are increasing the number of dichotomies that decodable by a simple linear decoder, conferring flexibility to a read out (a linear readout can implement a large number of input-output function without the need to modify (re-learn) the representation). Notably, no other brain region exhibited significant parity decodability (Fig. 2F, orange, all $p > 0.05$), no other region featured a significant increase in SD, and no other region exhibited the emergence of multiple, simultaneously decodable dichotomies between IA and IP sessions.

We next examined the format of the decodable named dichotomies (context, stim pair, parity) using geometric measures of abstraction. During the stim period, CCGP (Fig. 2E, S3D) was significantly elevated for both the context (Fig. 2B, red, IA vs. IP, $p_{RS} = 2.0x10^{-2}\ ,\ p_{IA} = 0.51, p_{IP} = 0.02$) and stim pair (Fig. 2E, purple, IA vs. IP, $p_{RS} = 2.0x10^{-28}, p_{IA} = 0.17, p_{IP} = 0.0011$) variables in IP but not in IA sessions. In addition, the PS was significantly larger than expected by chance for the variables context and stim pair in the IP but not IA sessions (Fig. S3E, red, $p_{IA} = 0.55, p_{IP} = 1.4x10^{-15}\ and$ Fig. S3E, purple, $p_{IA} = 0.17, p_{IP} = 1.7x10^{-8}$). CCGP larger than expected by chance was also observed for the stim pair variable in vmPFC during IP and not in IA (Fig. S3A, purple, $p_{IA} = 0.45, p_{IP} = 0.014$), the response variable in preSMA (Fig. S3A, green, $p_{IA} = 0.050, p_{IP} = 0.0010$), and the stim pair variable in AMY during both IA and IP (Fig. S3A, purple, $p_{IA} = 0.050, p_{IP} = 0.039$). Taken together, the increased CCGP and PS values in IP relative to IA sessions indicate that the context and stim pair variables are both simultaneously represented in an abstract format in the HPC in sessions where inference behavior was observed (IP sessions). This means that for example a decoder for context trained only on the conditions in which the subject responded Left, can readily generalize to all the conditions in which the subject responded Right. These two variables were not

represented in an abstract format when inference behavior is absent (IA sessions). While other task variables were also represented in an abstract format in other brain regions, only the hippocampus simultaneously represented these two variables in an abstract format (Fig. S3A,B). These two disentangled variables are represented in approximately orthogonal subspaces.

We also conducted a parallel analysis during the pre-stimulus baseline (Fig. 2, inset), analyzing the geometry of persistent representations of the previous trial. We found that context alone was encoded in an abstract format in the HPC only in sessions where inference was present (Supplement S.1), indicating that the hippocampal context representation was persistently maintained in an abstract format across trial epochs.

Together, these findings indicate the emergence of an abstract representation of latent context in the HPC that is correlated with patient's ability to perform inference about the current state of the task. This context variable is encoded in the hippocampal representation alongside: (i) an abstract, behaviorally-relevant stimulus category variable (stim pair, but see below), and (ii) an increased in parity dichotomy decodability and increased SD from IA to IP sessions. Together, these findings indicate that the ability to perform inference correlates with the hippocampus explicitly encoding a context variable, whose state is not directly observable from any low-level sensory feature of the task, in a representation that simultaneously exhibits features of high dimensionality (parity decodability) and low dimensionality (two variables in an abstract format).

<u>Hippocampal representation of context is absent in incorrect trials</u>

We next asked whether the presence of context as an abstract variable were associated with trial-by-trial performance. To examine this question, we compared the decodability and geometry of all dichotomies between correct and incorrect (error) trials in IP sessions, in which patients on average performed inference correctly. For this analysis, the first trial of every block was excluded due to being necessarily incorrect by design. Contrasting correct with incorrect trials, we found that dimensionality was significantly higher in correct trials (Fig. 3A, IP-c vs. IP-I, average SD 0.59 vs 0.54, $p_{RS} = 0.0048$), as was decodability of the parity dichotomy (Fig. 3A, IP-c vs. IP-i, orange, $p_{RS} = 0.029$). Furthermore, the dichotomies context and stimulus pairing were significantly more decodable in correct compared to incorrect trials (Fig. 3A, IP-c vs. IP-i, red, $p_{RS} = 1.2x10^{-20}$, purple, $p_{RS} = 1.0x10^{-9}$). Furthermore, PS for context but not other variables was significantly elevated in correct trials but not in incorrect trials (Fig. 3B, red, IP-c vs. IP-i, $p_{IP-c} = 1.1x10^{-16}, p_{IP-i} = 0.083$) in IP sessions. Similarly, the baseline representation of context also showed this effect, being decodable during correct trials but not during incorrect trials (Fig. 3C, red, IP-c vs. IP-i, $p_{RS} = 3.5x10^{-12}, p_{IP-c} = 0.012, p_{IP-i} = 0.47$). Context was present in an abstract format only in correct trials based on the PS for context being significantly larger than chance during correct but not incorrect trials (Fig. 3D, red, IP-c vs. IP-i $p_{IP-c} = 0, p_{IP-i} = 0.94$). The lack of decodability and parallelism of context in the baseline immediately prior to an incorrect trial indicates that the geometry of the representation at baseline is correlated with correct behavior in the upcoming trial. Together, these findings

demonstrate that both the content and format of the hippocampal neural representation is correlated with behavior on a trial-by-trial basis. This effect is present during the stim period in incorrect trials, where shattering dimensionality, context decodability, and context PS significantly decreased. This effect is also present during the baseline period, where reduction in the decodability and parallelism of the context variable is associated with an error on an upcoming trial.

Ablation analysis for neurons with univariate tuning and in seizure-onset zones

To determine whether the geometric findings presented here arose as a consequence of strong classical (univariate) tuning, two separate ablation analyses were conducted which entailed re-computing neural population geometry after removing all neurons with significant main effect tuning in (i) a 2x2x2 ANOVA for Response, Context, and Outcome, and (ii) a 4x2 ANOVA for Stimulus Identity and Context (Supplement S.2). In both analyses, context was still significantly decodable and encoded in an abstract format, though in (ii) the stimulus pair representation was no longer decodable. These control analyses indicate that the hippocampal context representation studied here does not arise only due to the emergence of classically context tuned neurons, and is instead a consequence of broadly distributed context tuning at the level of the population.

Furthermore, to ensure that the reported hippocampal representational geometry was not influenced by pathology, hippocampal neurons recorded from seizure onset zones were removed and the geometry was re-analysed (Supplement S.3). All findings were recapitulated, suggesting that hippocampal pathology did not contribute to these results.

Abstraction of stimulus coding across contexts uniquely increases in the hippocampus

Individual HPC neurons in humans prominently encode the identity of visual stimuli[21]. Visually tuned neurons, whose firing rate is strongly modulated by the identity of presented images, are a prominent example of such encoding[22]. We therefore next asked how the variable context, which we show above becomes encoded in the HPC as patients learn to perform inference, interacts with the stimulus identity variable. To examine this question, we conducted geometric analyses over pairs of stimuli (e.g. stimulus A vs B) in the two contexts. This approach is different from the analysis shown so far, in which we considered all 35 possible balanced dichotomies of task conditions. In each balanced dichotomy, different stimulus identities are grouped together (e.g. stimuli A&B vs stimuli C&D), thereby not allowing an examination of the representational geometry for individual stimuli. To examine how stimulus identity coding relates to context encoding, we instead consider all possible stimulus pairs individually (A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D) in both contexts. As a comparison, we concurrently performed this analysis on neurons recorded in ventral temporal cortex (VTC), which are also strongly modulated by stimulus identity[23–25], but in which context was not decodable at the level of the balanced dichotomy analysis (basline period: Fig. S8A, red; compare with Fig. 2G; stimulus period: Fig. S8B, red, compare with Fig. 2D).

First, we verified that neurons in both areas encoded the identity of the four stimuli presented. This was the case in both HPC and VTC: 109/494 (22%) of neurons in HPC (Fig. 4A

shows an example) and 195/269 (73%) of neurons in VTC (Fig. 4F shows an example) were significantly modulated by stimulus identity following stimulus onset (1x4 ANOVA, $p < 0.05$). Similarly to HPC, at the population level, VTC neurons encoded stimulus identity-related balanced dichotomies in an abstract format (Fig. S8B-D, purple, brown, pink, $p_{IA/IP} < 10^{-10}$). Furthermore, error trial analysis revealed that stimulus-related dichotomies were also decodable during errors in VTC (Fig. S8E, purple, brown, pink, $p_{IP-i} < 10^{-10}$). This finding contrasts with HPC (compare to Fig. 3A, stim pair AC vs. BD dichotomoy) and is consistent with the idea that neurons in VTC veridically represented the stimulus viewed on the screen by the patient during both correct and error trials). These analyses demonstrate that VTC and HPC neurons encode stimulus information, but only HPC does so simultaneously with context. This allowed us to next examine how context interacted with stimulus identity coding in the HPC.

We next conducted a geometric stimulus-pair analysis to study the interaction of stimulus identity and context coding in the same neural population. The stimulus-pair analysis was designed to detect the presence of simultaneous abstract coding of stimulus identity across contexts (see Fig. S9A-C for illustration) and abstract coding of context across stimuli (see Fig. S9D-F for illustration).

The average stimulus decoding accuracy across all stimulus pairs in the HPC did not differ significantly between IA and IP sessions (0.73 vs. 0.76; Fig. S10C, $p_{RS} = 0.13$, RankSum over stimulus pairs), indicating that the decodability of stimulus information was not different when patients could perform inference vs. when they could not (see also Fig. S10A). In contrast, the geometry of the stimulus representation became disentangled from context: both the stimulus CCGP (Fig. 4B, $p_{RS} = 0.041$) and stimulus PS (Fig. 4C, $p_{RS} = 0.040$) were significantly increased in IP compared to IA sessions. This means that a decoder trained to differentiate between stimulus A and B in one context generalized better to the other context in IP compared to IA sessions (and vice-versa for all other possible pairs), and suggests that the stimulus responses reorganized with respect to the emerging context variable. Note that while context was not decodable in IA sessions as a balanced dichotomy, the fact that stimulus decoders did not generalize well between the two contexts indicates that context did modulate stimulus representations in the HPC, but in a way that was entangled with stimulus identity (see below). The reorganization effect was specific to the HPC: in VTC, the neural population geometry was unchanged, as indicated by no significant differences in stimulus decodability (Fig. S10D, $p_{RS} = 0.15$), stimulus CCGP (Fig. 4D, $p_{RS} = 0.15$) and stimulus PS (Fig. 4E, $p_{RS} = 0.39$) between IA and IP sessions. These analyses demonstrate that the representational geometry for stimulus identity in HPC becomes significantly more disentangled from context in IP sessions compared to IA sessions in a manner that reflects an abstract format.

Having explored the generalization of stimulus encoding across contexts, we next turned our attention to study the generalization of the context code across stimuli. The presence of abstract coding for one variable (stimulus identity) does not necessarily imply that the other variable is also present in an abstract format, though we do have evidence that this is the case in hippocampus from the CCGP and PS analysis over balanced dichotomies (Fig. 2B, S4E). Context decoding analysis conducted over stimuli (e.g. considering only trials with stimuli A&B

shown, decode context) in HPC revealed that context was decodable for many stimuli both during IA and IP sessions, without significant difference between the two (0.63 vs. 0.67; Fig. S11A, $p_{RS} = 0.065$). However, despite being decodable, context encoding during IA sessions was not in an abstract format for stimulus pairs as indicated by low context CCGP (Fig. 4G, $p_{IA} > 0.10$ for all stim pairs) and low context PS (Fig. 4H, $p_{IA} > 0.17$ for all stim pairs except for AB, where $p_{IA} = 0.033$) values that were not significantly greater than chance. In contrast, in IP sessions, both context CCGP (Fig. 4G, $p_{RS} = 0.012$) and context PS (Fig. 4H, $p_{RS} = 0.015$) increased significantly relative to the IA group. This finding indicates that context emerged as an abstract variable at the level of individual stimulus pairs in the HPC.

We next contrast these findings with VTC. While context was decodable from some stimulus pairs during both IA and IP sessions (Fig. S11B, $p_{IA} \in (0.013, 0.074)$, $p_{IP} \in (0.020, 0.081)$ for all stim pairs), there was no significant change in context decodability between IA to IP sessions (Fig. S11B, $p_{RS} = 0.18$). Rather, there was a significant decrease in context CCGP (Fig. S11C, $p_{RS} = 0.026$) and no significant difference in context PS (Fig. S11D, $p_{RS} = 0.39$) from IA to IP. Together, these findings indicate that in the HPC, the context variable in the IP sessions is in an abstract format because context coding directions become aligned (i.e. parallel) across stimuli. For VTC, on the other hand, the lack of an abstract context representation across stimulus identities in both IA and IP sessions suggests that context, though decodable for individual stimulus pairs, is not organized in an abstract format and does not meaningfully correlate with inference behavior.

In summary, these findings indicate that the emergence of context as an abstract variable in the hippocampus when patients can perform inference is coupled with the reorganization of stimulus representations so they are also more disentangled, thereby forming a jointly abstracted code for stimuli and context. This transformation of the representation is visible directly in the data when projecting the neural representations in 3 dimensions using Multidimensional Scaling (Fig. 4I, S12). This reorganization occurs without the encoding of additional stimulus information since individual stimuli are equally decodable in the presence or absence of inference behavior. Critically, this change in geometry relies on the encoding of context in an abstract format and is unique to hippocampus, with systematic reorganization of stimulus representations with inference being absent from ventral temporal cortex.

<u>The tuning changes of hippocampal neurons that reflect the geometry changes</u>

We next examined what aspects of neuronal activity changed in the hippocampus to give rise to the abstract neural representations that we observed (i.e. the representations with elevated CCGP). We considered the following non-mutually exclusive possibilities (Fig. 5A-D). (i) The distances between conditions in state space could increase (Fig. 5A vs. Fig. 5B), either as a result of increased firing rate of variable-coding neurons, an increase in the fraction of tuned neurons, or an increase in the depth of tuning of these neurons. (ii) The variance of the population response projected along the coding direction could decrease (Fig. 5C). (iii) Parallelism could increase due to increases in the consistency of firing rate modulation in response to one variable over values of another (Fig. 5D).

We first examined whether mean firing rates differed between IA and IP sessions in the HPC. The firing rate across conditions decreased from 3.37±0.13 to 1.36±0.03 Hz, a 60% reduction on average during the stimulus period (Fig. 5E, $p_{RS} = 8.3x10^{-5}$). Firing rates were also reduced during the baseline period (3.29±0.09 to 1.38±0.02 Hz, 58% reduction, Fig. S14G). This firing rate reduction was unique to the HPC, with every other recorded region exhibiting no significant differences or increases in firing rate between IA and IP during the stimulus (Fig. S13C) and baseline periods (Fig. S14H).

The firing rate reduction led to a decrease in the average distance between class centroids (separation) across all dichotomies in IP (5.77 ± 0.22 to 4.17 ± 0.07 Hz, $p_{RS} = 2.9x10^{-8}$, Fig. 5G). However, the centroid distance for a single dichotomy, the context dichotomy, increased from IA to IP (4.3 vs 5.0 Hz, $p_{IA} = 0.87, p_{IP} = 0.076$, $p_{\Delta Dist} = 0.040$, Fig. 5G,H,S13G). In fact, context was the dichotomy with the largest change in distance in firing rate space when comparing the IP and IA conditions (Fig. 5H). This isolated significant rise in context separability was not seen in any of the other recorded areas during the stimulus period (Fig. S13A,B). Similarly, during the baseline period, the centroids for context moved together the fewest in the HPC (5.6 vs 5.0 Hz, $p_{IA} = 0.68, p_{IP} = 0.0007$, $p_{\Delta Di} = 0.027$, Fig. 5K,L) despite the significant decrease in distance over all dichotomies that was also observed here due to the firing rate reduction (5.85±0.08 to 4.25±0.04 Hz, $p_{RS} = 6.5x10^{-13}$, Fig. 5K).

Next, we assessed changes in the variability of the population response along the coding direction of each dichotomy. The variance along the coding direction of neuronal responses in the HPC decreased for all dichotomies in IP when compared to IA sessions during both the stimulus period (2.51 ± 0.16 vs. 1.53 ± 0.06, $p_{RS} = 6.5x10^{-13}$, Fig. 5I) and the baseline period (2.49 ± 0.09 vs. 1.58 ± 0.02, $p_{RS} = 6.5x10^{-13}$, Fig. S14A,B). However, this decrease could be a simple consequence of the reduction in firing rates under the assumption of Poisson statistics. We conducted a condition-wise Fano-factor analysis to assess whether the variance reduction was beyond that expected for a the reduction in firing rates. This analysis revealed no significant differences in Fano factors between IA and IP sessions during the stimulus period (1.39±0.22 vs 1.36±0.14 , $p_{RS} = 0.99$, Fig. 5F) and the baseline period (1.61±0.26 vs 1.45±0.11, $p_{RS} = 0.19$). Together, these two findings suggest that the decrease in variance along dichotomy coding directions is explained by the decreases in firing rate.

Though the increase in distances between dichotomy centroids for context appears to be a distributed, population-level phenomenon (see Fig. S5A-F), we sought to determine if a signature of this increase could be detected in the tuning of individual neurons (Supplement S.4). Indeed, the proportion of neurons exhibiting univariate context tuning did increase from IA to IP sessions, thus partially explaining the increased representational distance, but it should be noted that their exclusion does not qualitatively alter the hippocampal population geometry (Fig. S5).

Finally, to explain the increases in parallelism for stimuli across contexts for single neurons (see Fig. S11E-H for examples), we conduct an analysis of the consistency with which stimulus representations are modulated (firing rate increased or decreased) across contexts. This context modulation consistency is computed for each neuron, and can take on values between 0

and 4, with 0 indicating no consistency in modulation and 4 indicating all stimuli exhibit the same firing rate modulation between contexts (see Methods for details). We find a significant increase in the consistency of context modulation in HPC from IA to IP sessions (Fig. S5J,11I, 2.6±0.3 vs 1.6±0.2, $p_{RS} = 0.0039$). This effect was specific to HPC: in VTC, this metric decrease significantly (Fig. S11I, 1.8±0.2 vs 2.9±0.3, $p_{RS} = 0.0049$). These findings indicate that, for HPC, stimulus parallelism arises from shifts in the stimulus representations across contexts for individual neurons that significantly increase in consistency during IP sessions.

The changes in neural state space responses for HPC are summarized in Fig. 5M, and feature aspects of our previous hypotheses: (i) condition averages for context increase in separation despite relaxing towards the origin (decrease in firing rate), (ii) are accompanied by decreases in variance along the coding direction, and (iii) neurons become increasingly consistent (parallel) in their modulation across stimulus and context dimensions. Together, these changes explain the implementation of the context coding dimension in the hippocampal representation and how it is able to emerge as a simultaneous, linearly encoded variable alongside stimulus identity.

Context representations outside of the Hippocampus.

A representation of latent context that correlated with inference behavior was also observed during the baseline period in the dACC (Supplement S.5). Interestingly, context emerged in the task representation through a different implementation strategy, namely an increase in firing rates rather than a decrease, as was observed in HPC.

Verbal instruction induces the representation of context in an abstract format.

In all previous analyses, we compared sessions where patients could perform inference (IP) vs. those in which patients were not (IA), without regard to how patients transitioned from the IA to the IP group. We provided verbal instructions detailing the latent task structure following the first session (Fig. 6, inset) to all patients, allowing us to examine whether instructions lead to changes in neural representations. As shown above, patients were divided based on behavior into three types: those who exhibited inference behavior in the very first session (S1 group), those who did so after being given verbal instructions (IS group), and those who did not perform inference even after being provided with verbal instructions (IA group). We next compared the neural representation of context between these three groups of patients.

The instructions provided to these three groups were identical, and all patients acknowledged receipt of the instructions. All included patients exhibited high performance on non-inference trials both before and after being given instructions, indicating that they understood the task and learned the SRO maps. The principal difference between the IS (Fig. 6A, S15A,B) and IU (Fig. 6D, S15C,D) groups is therefore, their ability to perform inference following the verbal instructions, with both groups performing the task accurately otherwise. The S1 group, on the other hand, exhibited above-chance inference performance during both pre and post-instruction sessions (Fig. 6F, S15E,F).

In the IS group, context was decodable in the HPC during the stimulus period on correct trials in the session following the verbal instructions (Fig. 6B, $p_{pre-c} = 0.17, p_{post-c} = 0.016, p_{RS} = 3.1x10^{-19}$). This representation of context was in an abstract format, as indicated by significant increases in both CCGP (Fig. S16A; $p_{pre} = 0.28, p_{post} = 0.047, p_{RS} = 8.4x10^{-16}$) and PS (Fig. S16B; $p_{pre-} = 0.023, p_{post-} = 1.2x10^{-6}$). Successful performance in the task was associated with context being represented abstractly in HPC, as both the decodability (Fig. 6B, $p_{post-} = 0.99$, post-c vs. post-i, $p_{RS} = 4.3x10^{-20}$) and PS (Fig. S16B, $p_{post-} = 1.1x10^{-4}$) of context decreased significantly on incorrect trials in the post-instruction session. Context was also encoded in an abstract format during the baseline period in the same performance dependent manner as context in the stimulus period (Fig. S16C-E). At the single neuron level, this effect can be appreciated by an increase in the proportion of neurons that are significantly linearly tuned to context ($p < 0.05$, one-way ANOVA for context) during both the stimulus (8% (6/75 neurons) vs 18% (17/93 neurons), $p = 0.027$) and baseline (7% (5/75 neurons) vs 16% (15/93 neurons), $p = 0.029$) periods in the post-instruction session compared to the pre-instruction session (Fig. 6C shows an example). Thus, the ability of IS group patients to perform inference following instructions was associated with the rapid emergence of an abstract context variable in their hippocampus.

In contrast to the IS group, in patients in the IU group, context was not encoded by HPC neurons neither during the stimulus (Fig. 6E, S16F,G, all $p_{pre/post} > 0.05$) nor the baseline (Fig. S16H-J all $p_{pre/post} > 0.05$) periods in the post-instruction session. Furthermore, there was no significant change in tuning to context at the single-neuron level in the hippocampus of patients in the IU group both during the stimulus period (6% pre vs 6% post, $p = 0.41$) and the baseline period (8% pre vs 5% post, $p = 0.27$). These data indicate that receiving verbal instructions describing the latent context is insufficient to generate an abstract context representation in the hippocampus, as is the case for the IU group, but rather that the context representation is a correlate of an individual's inference behavior, which depends on that variable.

For the S1 patient group, context was decodable already during the pre-instruction session (Fig. 6G, $p_{pre} = 0.014$), and dropped slightly below significance ($p_{post} = 0.17$) during the post-session, likely due to the small number of patients and neurons present in this analysis. The CCGP was not significance (Fig. S16K, $p_{pre} = 0.21$, $p_{post} = 0.31$), but the PS in both the pre and post-sessions was highly significant and near the top of the dichotomy rank order in both cases (Fig. S16L, $p_{pre} = 1.5x10^{-9}$, $p_{post} = 1.7x10^{-6}$). This finding suggests that the context variable these patients learned experientially during the pre-instruction session was in an abstract format as assessed by PS but not CCGP, and that receiving the instructions did not significantly alter the behavior or the neural representation of context. A similar trend was observed with the baseline context representation for these patients (Fig. S16M-O). Note that such discrepancies between CCGP and PS are not unexpected since, when less data is available (fewer neurons, decreased firing rates), single trial measures (Decoding, CCGP) become less sensitive to representational structure than measures that operate on condition averages (PS).

We also examined firing rate changes of HPC neurons separately for the IS, IU, and S1 groups (Fig. S13D). This analysis revealed significant reductions in firing rate across all conditions from pre- to post-sessions for the IS patients alone (-0.39 ± 0.15 Hz, $p_{IS} = 1.4x10^{-4}$), confirming that reductions in hippocampal firing rate in the same neurons recorded across adjacent sessions were associated with increases in inference performance, and not simply a consequence of the passage of time or loss of recording signal. HPC neurons from IU patients exhibited an increase in firing rate (0.10 ± 0.04 Hz, $p_{IU} = 1.2x10^{-4}$) and S1 firing rates did not significantly change (0.06 ± 0.08 Hz, $p_{S1} = 0.08$).

Lastly, we compared the geometry of the context representations formed by each of these patient groups using the PS. We recomputed the PS for context with balanced neuron counts for IS/IU/S1 x pre/post. Only the IS group significantly increased in PS for context from chance levels similar to IU pre-instructions ($p_{IS,pre} = 0.20, p_{IU,pre} = 0.58$, Fig. 6H) to a level comparable to S1 post-instructions ($p_{IS,post} = 0.0028, p_{S1,post} = 0.0035$, Fig. 6H). All these analyses suggest that HPC neurons in the S1 group carried an abstract representation of context before receiving explicit instructions, and retained that geometry after receiving instructions. The IS group, on the other hand, could not construct an abstract representation of context and perform inference without instructions. However, upon receiving instructions, they could perform inference, and neurons in their HPC carried a task representation whose geometry resembled that of the S1 group, indicating that a surprisingly similar representational geometry can be constructed through either experience or instruction. Lastly, the IU group could not leverage the information provided in the instructions to perform inference, and accordingly, their hippocampi never carried an abstract context representation.

Discussion

The ability to perform inference in this task was associated with the hippocampus forming an abstract representation of the environment. This representation encoded latent context and stimulus identity in approximately orthogonal subspaces, was behaviorally relevant on the level of individual trials, and emerged with learning. To implement this representation, the context coding directions for different visual stimuli became more parallel, and the distance between contexts in neural state space slightly increased, even though hippocampal neurons overall reduced their firing rates, thereby reducing variance. This representation could emerge quickly, with some patients spontaneously learning the latent task structure during their first session and others exhibiting abstract representations within minutes of receiving verbal instructions explaining the task structure despite having no previous experience before the data shown here was recorded. Abstract representations of context and stimulus identity following stimulus onset were only present in the hippocampus and not in the other brain areas we examined. Together, this data reveals that hippocampal population codes can be restructured by learning and verbal instructions within minutes to support inference in a new task.

How can a neural or biological network efficiently encode multiple variables simultaneously[3,26]? One solution is to encode variables in an abstract format so they can be re-used in novel situations to facilitate generalization and compositionality[2,18,27–30]. Here, we show that in the human brain, such a disentangled representation emerged as a function of learning to perform inference in our task. The format by which latent context and stimulus identity was represented was predictive of the ability to perform behavioral generalization that relies on contextual inference. Crucially, patients performed well on non-inference trials in all sessions included in the analysis, indicating that they understood the task and successfully learned the SRO associations in both contexts. Therefore, the difference between the inference present and absent sessions was only in whether they performed inference following the covert context switch. For those sessions where patients did not perform inference, there was no systematic relationship between context coding vectors across stimuli. For sessions where patients performed inference, there was alignment of the context coding direction across stimuli (making them parallel), indicating that the context variable had been disentangled from the stimulus identity variable in the hippocampi of these patients. As a result, the two variables became disentangled, thereby allowing for generalization.

Inferential reasoning is thought to rely on cognitive maps, which have been observed in the hippocampus and other parts of the brain[12,31–35] . Cognitive maps are thought to underlie inferential reasoning in various complex cognitive and spatial domains[7,31,32,36–38]. However, little is known about how maps for cognitive spaces emerge at the cellular level in the human brain as a function of learning. Here, we show that a cognitive map that organizes stimulus identity and latent context in an ordered manner emerges in the hippocampus. The cognitive map emerges because task states in one context, indexed by stimulus identity, become systematically related to the corresponding task states in the other context through a dedicated context coding direction that is disentangled from stimulus identity. Furthermore, the relational codes between task states (stimuli) in each context are preserved across contexts.

However, hippocampal cognitive maps observed in other studies are often different from those that we observed. Indeed, the encoded variables are often observed to non-linearly interact, which is a signature of high dimensional representations. These representations are believed to be the result of a decorrelation of the neural representations (recoding) that is aimed at maximizing the memory capacity[39–41]. This form of pre-processing leads to widely observed response properties, like those of place cells[42]. However, there is some evidence of hippocampal neurons that encode one task variable independently of others[14,43–49]. In these studies, no correspondence was shown between different representational geometries in the hippocampus and differences in behavior. Here, the task representations generated when patients cannot perform inference (but can still perform the task) are systematically different from the abstract hippocampal representations of context and stimulus identity that correlate with inference behavior[3]. Finally, it is important to stress that we also observed an increase in the shattering dimensionality (SD). High SD can be compatible with the low dimensionality of disentangled representations[3,49].

Interestingly, we observed that there are other brain regions in which the representations strongly encode stimulus identity but do not undergo any major reorganization. This is the case in the ventral temporal cortex, a region analogous to macaque IT cortex, in which neurons construct a high-level representation of visual stimuli[50–52]. Some studies conducting unit recordings in this general region in humans have verified that neurons exhibit strong tuning to stimulus identity [24,25,53]. VTC does what a purely visual area is supposed to do: VTC responses to stimuli were not modulated by latent context in a systematic way. Despite being decodable for individual stimulus pairs, context was not represented in an abstract format. Rather, in VTC, context was only weakly decodable for a subset of the stimuli, context decodability did not change between IA and IP sessions (Fig. S11B,C), and stimulus identity geometry was not reorganized relative to context in IP sessions (Fig. 4D,E). Our study, therefore shows that disentangled context-stimulus representations emerged in the hippocampus, but not in the upstream visually responsive region VTC.

The hippocampus was not unique in its representation of context. A weaker representation of context was also found in the dACC, only during the baseline period, and trivially abstract by virtue of being the only encoded variable. This finding is in line with work implicating the dACC in the representation of task rules and task sets [54–59]. Following stimulus onset, however, dACC did not contain a representation of latent context (Fig. 2). Hippocampus being the sole region to encode context during stimulus processing also runs contrary to previous observations that neurons in the medial frontal cortex (dACC and preSMA) are tuned to task context paradigms with explicitly cued context switches[16,17,60,61]. In contrast, in our present task, context switches were uncued and had to be inferred from outcomes. It remains an open question to examine whether cued vs. inferred context switches engage different mechanisms of switching between contexts and/or different context encoding schemes in the hippocampus.

The focus of our study was to examine how representations of context, stimulus identity, response, and predicted outcome change as a function of learning in the human brain. In a prior study in macaques[3], the representation of the same variables in very well trained animals was

examined in HPC, dlPFC, and ACC. Several notable differences exist between the two studies. First, context was encoded in an abstract format at baseline and was decodable after stimulus onset in all three brain areas examined in the macaques. In contrast, in humans, context is only strongly decodable in the HPC. We hypothesize that the wide-spread encoding of context in the macaque study was due to the extensive training the animals received before recordings commenced, unlike our patients who had no task experience. It is possible that early on during learning, latent context representations are present only in the hippocampus, and are propagated to cortex (dACC) with extensive task experience. This hypothesis is supported by prominent direct and indirect projections from the hippocampus to dACC in primates[62–64], and flexible, context-dependent interactions between medial frontal cortical neurons and hippocampal outputs[16,65]. Second, the human hippocampus exhibited abstract stimulus representations, unlike the abstract response or "choice" representation in the macaques. Notably, the abstract stimulus pair and response dichotomies are constructed such that high CCGP for one will necessarily lead to below-chance CCGP for the other, which was indeed the case for both our study (high stim pair, low response) and the primate study (low stim pair, high response). One potential reason for these differences is a species difference: human HPC neurons are strongly modulated by the identity and semantic category of presented images[16,21,66–68], making it natural to organize representations of context relative to this existing representation. Similarly, representations of choices are not prominent in the human HPC[16]. Another potential reason is a difference in task construction: our task employed semantically identifiable images, whereas the prior experiment with macaques used fractals. Third, unlike macaque HPC, human HPC did not encode predicted outcome. We note that in our task, outcome prediction was not necessary to perform the task because context switches were signaled by the accuracy of the response (correct or incorrect), which was independent of predicted outcome received for a correct response. Furthermore, all possible task-states were uniquely indexable using stimulus identity and context, rendering outcome prediction representation unnecessary for unambiguously defining the current task state. Finally, another possibility is that the reward for the macaques (juice volume) was more motivationally salient than the small monetary reward (25¢ or 5¢) patients received. We hypothesize that these reasons obviated the need for a predictive representation of outcome to complete the task in our patients. It remains an important open question whether representations similar to those seen in macaques emerge in the other brain areas we examined following extensive training with tens of thousands of trials. Our data indicates that on short experiential timescales, the human hippocampus generates a representation that encodes the minimum set of variables required to solve the task at hand.

In our study, verbal instructions resulted in changes in hippocampal task representations that correlated with behavioral changes. Providing a language-based description of latent task structure can result in reformatting hippocampal representations to reflect that structure. The emergence of this representation in the session immediately following the instructions in the IS group is correlated with their newfound ability to perform inference and suggests that hippocampal representations can be modified on the timescale of minutes through verbal instructions in line with behavioral changes. This change in representation is qualitatively different from the standard neurophysiological approach of studying the emergence of a "learning set", wherein a low-dimensional representation of abstract task structure emerges

slowly over days through trial-and-error learning and correlates with increased learning rate for novel problems[44,69,70]. Importantly, the instructions provided here were themselves abstract, detailing the existence of stimulus pairs and contexts, and their application to specific stimuli was left as an exercise to the patient. The neurophysiological effect of providing instructions with varying degrees of abstraction as opposed to concrete instructions that specify stimulus-response associations is an area of ongoing research, and the role of the hippocampus in the construction of abstraction hierarchies remains to be elucidated[71,72]. Our finding of similar representational structure in the hippocampus in subjects who learned spontaneously and those who only learned after receiving verbal instructions suggests that both ways of learning can potentially lead to the same solution in terms of the underlying neural representation. In complex, high-dimensional environments, learning abstract representations through trial and error becomes exponentially costly (the curse of dimensionality), and instructions can be used to steer attention towards previously undiscovered latent structure that can be explicitly represented and utilized for behavior. The process of instruction-dependent restructuring of hippocampal representations is likely cortical-dependent, given the role of the cortex in language comprehension, but the exact mechanism by which this process occurs remains to be explored[27,73]. Our findings suggest that when high-level instructions successfully alter behavior, the underlying neural representation is rapidly modified to resemble one learned through experience.

Acknowledgements

We thank the members of the labs of R. Adolphs, U. Rutishauser, and M. Meister for discussion. We thank all subjects and their families for their participation and the staff of the Cedars-Sinai and Toronto Western Epilepsy Monitoring Units for their support.

Funding: This work was supported by the BRAIN Initiative through the NIH Office of the Director (U01NS117839 to U.R.), NIMH (R01MH110831 to U.R.), the Caltech NIMH Conte Center (P50MH094258 to R.A. and U.R.), the Simons Foundation Collaboration on the Global Brain (to R.A., S.F., and U.R.), the Moonshot R&D JPMJMS2294 (to R.A.), and by a merit scholarship from the Josephine De Karman Fellowship Trust (to H.S.C.).

Author contributions: Conceptualization: J.M., U.R., R.A., C.D.S., S.F., Task design: J.M. and D.L.K., Data collection: J.M., H.S.C, and A.R.C., Data analysis: H.S.C. and J.M., Writing: H.S.C., U.R., and S.F., Clinical care and experiment facilitation: C.M.R. Surgeries: A.N.M. and T.A.V., Supervision: U.R., R.A., and S.F.

Competing interests: None.

Methods:

Participants: The study participants were 17 adult patients who were implanted with depth electrodes for seizure monitoring as part of an evaluation for treatment for drug -resistant epilepsy (see Table 1). 14 were monitored at Cedars-Sinai Medical Center (CSMC) and the other 3 were monitored at Toronto Western Hospital (TWH). All patients provided informed consent and volunteered to participate in this study. All research protocols were approved by the institutional review boards of CSMC, TWH, and the California Institute of Technology.

Psychophysical Task and Behavior: Participants performed a serial reversal learning task. There were two possible static stimulus-response-outcome (SRO) maps, each of which was active in one of the two possible contexts. Context was latent and switches between context were uncued. Each recording session consisted of 280-320 trials grouped into 10-16 blocks of variable size (15-32 trials/block) with block transitions corresponding to a change in the latent context. Each trial consisted of a blank baseline screen, stimulus presentation, speeded response from the participant, followed by feedback after a brief delay (Fig. 1A). Responses were either "left" or "right" in every trial. In each session, stimuli were four unique images, each chosen from a different semantic category (human, macaque, fruit, car). If a patient performed multiple sessions, new images not seen before by the patient were chosen for each session. The task was implemented in MATLAB (The Mathworks, Inc., Natick, MA) using PsychToolbox-3[74]. Images were presented on a laptop positioned in front of the patient and subtended approximately 10 degrees of visual arc (300 px^2, 1024x768 screen resolution, 15.6 inch (40 cm) monitor, 50 cm viewing distance). Patients provided responses using a binary response box (RB-844, Cedrus Inc.).

Receipt of reward in a given trial was contingent on the accuracy of the response provided. In each trial, either a high or low reward (25¢ or 5¢) was given if the response was correct, and no reward (0¢) if incorrect. Whether a given trial resulted in high or low reward if the response was correct was determined by the fixed SRO map (see Fig. 1C). Stimulus-response associations were constructed such that two out of four images (randomly selected) were assigned one response and the other two images were assigned the other (e.g. human and fruit = left, macaque and car = right). Thus, in each context, each stimulus was uniquely specified by a combination of its correct response (left/right) and reward value (high/low). Crucially, the SRO maps of the two possible contexts were constructed so that they were the opposite of each other from the point of view of the associated response (Fig. 1C). To fully orthogonalize also associated reward, half of the reward values stayed the same and the others switched. This structured relationship of stimuli across contexts led to the full orthogonalization of the response, context, and reward variables (Fig. 1B-C). Crucially, the stimulus-response map inversion across contexts provided the opportunity for patients to perform inferential reasoning about the current state of the SRO map, and therefore the latent context.

Since rewards were provided deterministically, participants could switch context upon receiving a single error. Therefore, if patients performed inference, they should be able to respond correctly after receiving a single error. The behavioral signature of inferential reasoning was thus the accuracy in the trials that occurred immediately after the first error trial.

785 Specifically, we took a participant's performance on the first instance of each of the three
786 remaining stimuli in the new context is to measure a participants inference capabilities.
787 Patients completed multiple sessions of the task, in each of which new stimuli were
788 chosen. After completion of the first session, the experimenter provided a standardized
789 description of the latent contexts and SRO reversal to the patient (see below). These instructions
790 were given regardless of how well the patient performed in the immediately preceding session.
791 After this brief interlude, the participants completed the task again with a novel set of four
792 stimuli.
793

794 Instructions given to patients:

795 ---------- Instruction set 1 (before first session) ----------

796 In this task, we will show you a series of images, 4 of them in total. Your objective is to learn the
797 correct response for each image (either left or right). In the beginning, you will not know what
798 the correct answer is, so take a guess. The correct answer for an image may occasionally change,
799 so pay close attention. For every correct answer you will receive a reward of either 25 or 5 cents.
800 For an incorrect answer you will receive 0 cents. This is real money that you will receive before
801 you leave the hospital in the form of a gift card to your favorite place (ex. Starbucks). You will
802 have the opportunity to take a break halfway through.

803 ---------- Instruction set 2 (before second session) ----------

804 You may have noticed that some images have the same correct response and some images have
805 the same reward. Even when the correct response changes, they usually change together. In this
806 experiment, we are going to try a different strategy. Pay attention to which images go together
807 (i.e. have the same correct response and similar reward). This should make it a lot easier to
808 perform the task. To make the task a little more difficult, now the correct response for each
809 image will change a little more frequently.

810 **Behavioral Control:** We administered a control version of the task identical to the 'first session'
811 described above to n=49 participants recruited on Amazon Mechanical Turk (MTurk). We then
812 used this data to calibrate the difficulty of the task. A majority (~75%) of the control subjects
813 demonstrated proper inference performance, and the remaining 25% demonstrating slow
814 updating of SROs after a context switch, consistent with a behavioral strategy where each
815 stimulus is updated independently (see Fig. S1A).

816 **Electrophysiology:** Electrode Placement and Recording: Extracellular electrophysiological
817 recordings were conducted using microwires embedded within hybrid depth-electrodes (AdTech
818 Medical Inc.). The patients we recruited for this study had electrodes implanted in at least the
819 hippocampus, as well as in addition subsets of amygdala, dACC, pre-SMA, vmPFC, and VTC as
820 determined by clinical needs (see Table 1). Implant locations were often bilateral but some
821 patients only had unilateral implants as indicated by clinical needs. Broadband potentials (0.1Hz
822 – 9kHz) were recorded continuously from every microwire at a sampling rate of 32kHz (ATLAS

system, Neuralynx Inc.). All patients included in the study had well isolated single neuron(s) in at least one of the brain areas of interest.

Electrode Localization: Electrode localization was conducted using a combination of pre-operative MRI and post-operative CT using standard alignment procedures as previously described[16,60]. Electrode locations were co-registered to the to the MNI152-aligned CIT168 probabilistic atlas[75] for standardized location reporting and visualization. Placement of electrodes in gray matter was confirmed through visual inspection of subject-specific CT/MRI alignment, and not through visualization on the atlas.

Spike Detection and Sorting: Raw electric potentials were filtered with a zero-phase lag filter with a 300Hz-3kHz passband. Spikes were detected and sorted using the OSort software package[76]. All spike sorting outcomes were manually inspected and putative single-units were isolated and used in all subsequent analyses. We evaluated the quality of isolated neurons quantitatively using our standard set of metrics[66,77,78] including proportion of inter-spike interval violations < 3ms, signal-to-noise ratio of the waveform, projection distance between pairs of isolated clusters, and isolation distance of each cluster relative to all other detected spikes.

Selection of Neurons, Trials, and Analysis Periods: Activity of neurons was considered during two epochs throughout each trial: the baseline period (base), defined as -1s to 0s preceding stimulus onset on each trial, and the stimulus period (stim), defined as 0.2s to 1.2s following stimulus onset on each trial. Spikes were counted for every neuron on every trial during each of these two analysis periods. The resulting firing rate vectors were used for all encoding and decoding analyses. Tests of single-neuron selectivity were conducted using N-way ANOVAs with significance at $P < 0.05$, where N was either 2 for models of stim id (A, B, C, D) and context (1, 2), or 3 for models including outcome (High, Low), response (Left, Right), and context (1, 2). All variables were categorical, and all models were fit with all available interaction terms included.

Population analysis – decoding: Single-trial population decoding analysis was performed on pseudo-populations of neurons assembled across all neurons recorded across all patients. We pooled across sessions within each anatomically specified recording area as described previously[16,17]. Decoding was conducted using support vector machines (SVM) with a linear kernel and L2 regularization as implement in matlab's fitcsvm function. No hyperparameter optimization was performed. All decoding accuracies are reported for decoding accuracy for individual trials. Decoding accuracy is estimated out-of-sample using 5-fold cross-validation unless otherwise specified (e.g. cross-condition generalization). Many of the decoding analyses in this work consist of grouping sets of distinct task conditions into classes, then training an SVM to discriminate between those two groups of conditions. Neurons included in the analysis were required to have at least K correct trials of every unique condition in order to be included in the analysis (K = 15 trials unless otherwise stated). To construct the pseudopopulation, we then randomly sampled K trials from every unique condition and divided those trials into the groups required for the current decoding analysis for every neuron independently. Randomly sampling correct trials in this way allowed us to destroy noise-correlations that might create locally correlated sub-spaces from neurons recorded in the same area and session[16].

To account for the variance in decoding performance that arose from this random sub-sampling procedure, all reported decoding accuracies are the average resulting from 1000 iterations of sub-sampling and decoder evaluation. A similar trial balancing and sub-sampling procedure was conducted for all analyses that report decoding accuracy on incorrect trials, but with K = 1 trial/condition required as incorrect for the neuron to be included in analysis. Various other analyses conducted throughout this work, including representation geometry measures, centroid distances, and coding direction variances, all rely on this procedure of balanced correct and incorrect trial sub-sampling, and averaging over 1000 iterations of the computed metric to study the relationships between task conditions in an unbiased manner. All reported values have been computed with this approach unless otherwise stated.

Construction of Balanced Dichotomies: Our task has 8 possible states (Fig. 1B). We characterized how neurons represented this task space by assessing how a decoder could differentate between all possible "balanced dichotomies" of these 8 task conditions (Fig. 1B). The set of all possible balanced dichotomies is defined by all possible ways by which the 8 unique conditions can be split into two groups containing 4 of the conditions each (e.g. 4 points in context 1 vs 4 points in context 2 is the context dichotomy). There are 35 possible balanced dichotomies (nchoosek(8,4)/2). Some of the possible balanced dichotomies are easily interpretable because they correspond to variables that were manipulated in the task. We refer to these balanced dichotomies as the "named dichotomies", which are: context, response, outcome, stimulus pair (stim pair), and parity. These dichotomies are shown individually in Fig. S2. The stim pair dichotomy corresponds to the grouping of stimuli for which the response is the same in either context (A&C vs. D&B; see Fig. S2). The parity dichotomy is the balanced dichotomy with the maximal non-linear interaction between the task variables (Fig. S2).

Defining decoding difficulty of dichotomies: We quantify the relative degree of non-linear variable interactions needed by a neural population to classify a given dichotomy using a difficulty metric that rates dichotomies that require proximal task conditions to be placed on opposite sides of the decision boundary as more difficult. Note that proximity of task conditions in task space here is defined with respect to the variables that were manipulated to construct the task space. The conditions corresponding to (Response L, Outcome Low, Context 1) and (Response L, Outcome Low, Context 2) are proximal since their task specifications differ by a single variable (hamming distance 1) whereas (Response L, Outcome Low, Context 1) and (Response R, Outcome High, Context 2) are distal since their task specifications differ by all three variables (hamming distance 3). With this perspective, we can systematically grade the degree of non-linearity required to decode a given dichotomy with high accuracy as a function of the number of adjacent task conditions that are on opposite sides of the classification boundary for that dichotomy. For a set of 8 conditions specified by 3 binary variables, this corresponds to the number of adjacent vertices on the cube defined by the variables that are in opposing classes (See Fig. S7A). We define this number as the "difficulty" for a given dichotomy, and can compute it directly for every one of the 35 balanced dichotomies. The smallest realizable dichotomy difficulty is 4, and corresponds only to named dichotomies that align with the axis of one of the three binary variables used to specify the task space. The largest realizable dichotomy is 12, and this corresponds to the parity dichotomy since the dichotomy difficulty (number of

adjacent conditions with opposing class membership) is maximized in this dichotomy by definition. All remaining dichotomies lie between these two extremes in difficulty, and computing average decoding accuracy over dichotomies of increasing difficulty gives a sensitive readout of the degree of non-linear task variable interaction present in a neural population.

Geometric Analysis of Balanced Dichotomies: We used three measures to quantify the geometric structure of the neural representation[3]: shattering dimensionality (SD), cross-condition generalization performance (CCGP), and parallelism score (PS).

SD is defined as the average decoding accuracy across all balanced dichotomies. It is an index of the expressiveness of a representation, as representations with higher SD allow more dichotomies to be decoded. The content of a representation is assessed by considering which balanced dichotomies are individually decodable better than expected by chance.

CCGP assesses the extent to which training a decoder on one set of conditions generalized to decoding a separate set of conditions. Note that to compute CCGP, all trials from a set of conditions are held out from the training data, which is different from the "leave-one-out" type decoding used to estimate SD. The remaining held-in conditions are used to train the decoder, and performance is then evaluated on the held-out conditions (trial-by-trial performance). The CCGP for a given balanced dichotomy is the average over all possible 16 combinations of held-out conditions on either side of the dichotomy boundary. One of the 4 conditions on each side of the dichotomy are used for testing, whereas the remaining three on each side of the dichotomy are used for training. For each of the 16 possible train/test splits, the decoder is trained on all correct trials from the remaining six conditions, and performance is evaluated on the two held-out conditions.

PS assesses how coding directions for one variable are related to each other across values of other variables in a decoder agnostic manner. The PS is defined for every balanced dichotomy as the cosine of the angle between two coding vectors pointing from conditions in one class to conditions in the other for a given dichotomy. These vectors are computed by selecting four conditions (two on either side of the dichotomy), computing the normalized vector difference between the mean population response for each of the two pairs, then computing the cosine between said coding vectors.This procedure is repeated for all possible pairs of coding vectors, and the average over all cosines is reported. Since the correct way of "pairing" conditions on either side of the dichotomy is not known a-priori, we compute the cosine average for all possible configurations of pairing conditions on either side of the dichotomy, then report the PS as the maximum average cosine value over configurations.

Null distribution for geometric measures: We used two approaches to construct null distributions for significance testing of the geometric measures SD, CCGP, and PS.

For the SD and decoding accuracy of individual dichotomies, the null distribution was constructed by shuffling trial labels between the two classes on either side of each dichotomy prior to training and testing the decoder. After shuffling the order of the trial labels, the identical procedures for training and testing were employed. This way of constructing the null distribution destroys the information content of the neural population while preserving single-neuron properties such as mean firing rate and variance.

For the CCGP and PS, we employed a geometric null distribution[3]. Prior to training, we randomly swapped the responses of pairs of neurons within a given condition. For example, for one task condition, all of neuron 1's responses are assigned to neuron 2 and all of neuron 2's responses are assigned to neuron 1, for another task condition, all of neuron 1's responses are assigned to neuron 3, etc…). This way of randomly shuffling entire condition responses leads to the situation where neural population response statistics by-condition are held constant, but the systematic cross-condition relationships that exist for a given neuron are destroyed. This way of shuffling creates a maximally high dimensional representation, thereby establishing a conservative null distribution for the geometric measures CCGP and PS.

Analysis of Incorrect Trials: For determining decoding accuracy for trials in which subjects provided an incorrect response ("error trials"), decoders were trained and evaluated out of sample on all correct trials in IA and IP sessions (denoted as "IA-c" and "IP-c" trials respectively). The accuracy of the decoder was then evaluated on the left out incorrect trials in the IP sessions (denoted as "IP-i"" trials) that were balanced by task condition. Neurons from sessions without at least one incorrect trial for each of the 8 conditions were excluded. We did not estimate CCGP separately for correct and incorrect trials. The PS was estimated using only correct trials for IP-c and IA-c. For IP-i, parallelism was computed using one coding vector (difference between two conditions) from correct trials and one coding vector from incorrect trials. All other aspects of the PS calculation remained as described earlier. The very first trial after a context switch was excluded from analysis (it was incorrect but by design, as the subject cannot know when a context switch occurred).

Stimulus Identity Geometry Analysis (Fig. 4): We repeated the geometric analysis described above for subsets of trials to examine specifically how the two variables context and stimulus interact with each other. To do so, we considered each possible pair of stimuli (AB, AC, AD, BC, BD, CD) separately. For each stimulus pair, we then examine the ability to decode and the structure of the underlying representation for two variables: stimulus identity (see Table 3) and context (see Table 4).

For stimulus identity, what is decoded is whether the stimulus identity is the first or second possible identity in each pair (i.e. "A vs. B" for the AB pair). Stimulus CCGP (Fig. 4B,D) is calculated by training a decoder to decide "A vs. B" in context 1 and testing the decoder in context 2 and vice-versa (the CCGP is the average between these two decoders). Stimulus PS (Fig. 4C,E) is the angle between the two coding vectors "A vs. B" in context 1 and 2.

For context, decoding accuracy is estimated by training two decoders to decide "Context 1 vs. Context 2" for each of the two stimuli in a stimulus pair. The reported decoding accuracy is the average between these two decoders (Fig. S11A,B). For example, for the stimulus pair AB, one such decoder each is trained for all "A" trials and all "B" trials. Context CCGP (Fig. 4G, S11C) is calculated by training a decoder to differentiate between Context 1 and 2 based on the trials in the first identity of the pair, and tested in the second pair and vice-versa. The reported Context CCGP value for a given stimulus pair is the average between the two. Similarly, context PS (Fig. 4H, S11D) is the angle between the two coding vectors Context 1 vs. Context 2 estimated separately for the first and second stimulus in a pair.

Distance/Variance Analysis (Fig. 5): We computed a series of metrics to quantify aspects of the population response that changed between IA and IP sessions. We used (i) the firing rate, (ii) distance in neural state space between classes for balanced dichotomies and stimulus dichotomies (dichotomy distance), (iii) variance of neural spiking projected along the coding directions for those dichotomies (coding direction variance), and (iv) the condition-wise fano factor.

Firing rate (Fig. 5E) was the mean firing rate averaged across all neurons during the stimulus period, reported separately for correct trials of every unique task condition. Values reported during the baseline (Fig. S14G,H) are computed with an identical procedure using firing rates from before 1s prior to stimulus onset.

Dichotomy distance (Fig. 5G,H,K,L) was defined as the Euclidean distance in neural state space between the centroids of the two classes on either side of the decision boundary for that dichotomy. Centroids were computed by constructing the average response vector for each class using a balanced number of correct trials from every condition included in each class through a resampling procedure (described below). Null distributions reported for dichotomy distances are geometric null distributions.

Coding direction variance (Fig. 5I) was computed for a given balanced dichotomy by projecting individual held-out trials onto the coding vector of the decoder trained to differentiate between the two groups of the balanced dichotomy being evaluated. The coding direction was estimated by training a linear decoder on all trials except eight (one from each condition either side of the dichotomy). The vector of weights estimated by the decoder (one for each neuron) was normalized to unit magnitude to estimate the coding vector. The projection of the left out trial onto this coding vector was than calculated using the dot product. This process was repeated 1000 times, generating a distribution of single trial projections onto the coding vector for each dichotomy. The variance of the distribution of 1000 projected data point was then computed and reported as the variance for a given balanced dichotomy (Fig. 5I).

The condition-wise Fano factor (Fig. 5F) was computed separately for each neuron. We used all correct trials for a given balanced dichotomy to estimate the mean firing rate and standard deviation and then took the ratio between the two to calculate the Fano factor for each neuron. Reported fano factors are the average of all fano factors across all neurons from that area/behavioral condition. Fano factors are computed by-condition since grouping trials across conditions could lead to task variable coding (signal) contaminating the fano-factor measurement, which should ideally only reflect trial-by-trial variation around the mean for approximately poisson-distributed firing rates.

The context-modulation consistency (Fig. 5J) was also computed separately for each neuron. Context modulation consistency is the tendency for a neuron's firing rate to shift consistently (increase or decrease) to encode context across stimuli. For each neuron, it was computed by deteriming the sign of the difference (+/-) between the mean firing rate for a given stimulus between the two contexts, and summing the number of stimuli that exhibit the same modulation (either increase or decrease) across the two contexts. This consistency can take on values between 0 (increase in firing rate to encode context for half of the stimuli, decrease in firing rate for the other half) and 4 (either increase or decrease in firing rate for all four stimuli).

Bootstrap Re-sampled Estimation of Measures and Null Distributions: All the measures described in the preceding sections were estimated using a trial and neuron-based re-sampling wmethod. This resampling strategy was used to assure that every measure reported is comparable between a set of conditions by assuring that the same number of neurons and data points are used to train and test classifiers. Metrics were re-computed 1000 times with resampling and all null distributions were computed with 1000 iterations of shuffling and re-computing. Plotted boundaries of null distributions correspond to the 5th and 95th percentiles as sestimated from the 1000 repetitions.

A single iteration of the re-sampling estimation procedure proceeds as follows. For all analyses that involved a comparison of a metric between two behavioral conditions (IA vs IP or pre vs. post instruction), the same number of neurons was included in both conditions by on a region by region basis. For a neuron to be included, at least 15 correct trials for each of the 8 unique task conditions had to exist (120 correct trials total). Across patients, the number of correct trials per condition varied: min = 10.9 ± 1.3 trials/condition, mean = 25.0 ± 0.6 trials/condition, max = 39.6 ± 1.2 trials/condition (mean ± s.e.m.). After identifying the neurons that met this inclusion criteria, an equal number were randomly sampled from both behavioral conditions. The number of considered neurons was set to the number of neurons available in the smallest group.

When constructing feature matrices for decoding, 15 trials were randomly selected from each unique condition that was included in the given analysis. Trial order was shuffled independently for every neuron within condition to destroy potential noise correlations between neurons that were simultaneously recorded. For decoding and SD, out-of-sample accuracy was estimated with 5-fold cross validation. For generalization analyses (CCGP), all trials were used in training since performance is evaluated on entirely held-out conditions. For vector-based measures (dichotomy distance, variance, PS), all trials in relevant conditions were used to compute condition centroids. In the case of variance estimation, all trials except one on either side of the dichotomy boundary were used to learn the coding axis, then the held-out trials were projected onto the coding axis. As previously stated, these procedures were repeated 1000 times with independent random seeds to ensure independent random sampling of neurons and trials across iterations.

Bibliography

1. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology* **20**, 251–256 (2010).

2. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**, e253 (2017).

3. Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S. & Salzman, C. D. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).

4. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* **331**, 1279–1285 (2011).

5. Kemp, C. & Tenenbaum, J. B. Structured statistical models of inductive reasoning. *Psychological Review* **116**, 20–58 (2009).

6. Ho, M. K., Abel, D., Griffiths, T. L. & Littman, M. L. The value of abstraction. *Current Opinion in Behavioral Sciences* **29**, 111–116 (2019).

7. Vaidya, A. R., Jones, H. M., Castillo, J. & Badre, D. Neural representation of abstract task structure during generalization. *eLife* **10**, e63226 (2021).

8. Radulescu, A., Shin, Y. S. & Niv, Y. Human Representation Learning. *Annual Review of Neuroscience* **44**, 253–273 (2021).

9. Niv, Y. Learning task-state representations. *Nat Neurosci* **22**, 1544–1553 (2019).

10. Robert, S., Arno, V. & J.d, M. M. Distinct hippocampal and cortical contributions in the representation of hierarchies. *eLife* **12**, (2023).

11. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).

12. Knudsen, E. B. & Wallis, J. D. Hippocampal neurons construct a map of an abstract value space. *Cell* **184**, 4640-4650.e10 (2021).

13. Aronov, D., Nevers, R. & Tank, D. W. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* **543**, 719–722 (2017).

14. Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D. & Tank, D. W. Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).

15. Courellis, H. S., Nummela, S. U., Metke, M., Diehl, G. W., Bussell, R., Cauwenberghs, G. & Miller, C. T. Spatial encoding in primate hippocampus during free navigation. *PLOS Biology* **17**, e3000546 (2019).

16. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by human medial-frontal cortex. *Science* **368**, eaba3313 (2020).

17. Fu, Z., Beam, D., Chung, J. M., Reed, C. M., Mamelak, A. N., Adolphs, R. & Rutishauser, U. The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* **376**, eabm9922 (2022).

18. Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D. & Botvinick, M. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat Commun* **12**, 6456 (2021).

19. Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D. & Lerchner, A. Towards a Definition of Disentangled Representations. Preprint at https://doi.org/10.48550/arXiv.1812.02230 (2018).

20. Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K. & Fusi, S. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

21. Fried, I., Rutishauser, U., Cerf, M. & Kreiman, G. *Single Neuron Studies of the Human Brain: Probing Cognition*. (MIT Press, 2014).

22. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* **3**, 946–953 (2000).

23. Gross, C. G. Single neuron studies of inferior temporal cortex. *Neuropsychologia* **46**, 841–852 (2008).

24. Khuvis, S., Yeagle, E. M., Norman, Y., Grossman, S., Malach, R. & Mehta, A. D. Face-Selective Units in Human Ventral Temporal Cortex Reactivate during Free Recall. *J. Neurosci.* **41**, 3386–3399 (2021).

25. Decramer, T., Premereur, E., Zhu, Q., Van Paesschen, W., van Loon, J., Vanduffel, W., Taubert, J., Janssen, P. & Theys, T. Single-Unit Recordings Reveal the Selectivity of a Human Face Area. *J Neurosci* **41**, 9340–9349 (2021).

26. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]* (2014).

27. Ito, T., Klinger, T., Schultz, D. H., Murray, J. D., Cole, M. W. & Rigotti, M. Compositional generalization through abstract representations in human and artificial neural networks. Preprint at https://doi.org/10.48550/arXiv.2209.07431 (2022).

28. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci* **22**, 297–306 (2019).

29. Johnston, W. J. & Fusi, S. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nat Commun* **14**, 1040 (2023).

30. Muhle-Karbe, P. S., Sheahan, H., Pezzulo, G., Spiers, H. J., Chien, S., Schuck, N. W. & Summerfield, C. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron* (2023) doi:10.1016/j.neuron.2023.08.021.

31. Epstein, R. A., Patai, E. Z., Julian, J. B. & Spiers, H. J. The cognitive map in humans: spatial navigation and beyond. *Nat Neurosci* **20**, 1504–1513 (2017).

32. Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L. & Kurth-Nelson, Z. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).

1143 33. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map*. *Oxford University Press: Oxford,*
1144 *UK. (1978)* (Oxford University Press, 1978).

1145 34. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal Cortex as a Cognitive
1146 Map of Task Space. *Neuron* **81**, 267–279 (2014).

1147 35. Igarashi, K. M., Lee, J. Y. & Jun, H. Reconciling neuronal representations of schema, abstract
1148 task structure, and categorization under cognitive maps in the entorhinal-hippocampal-frontal circuits.
1149 *Current Opinion in Neurobiology* **77**, 102641 (2022).

1150 36. Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W. & Behrens, T. E. J. How to build a
1151 cognitive map. *Nat Neurosci* **25**, 1257–1272 (2022).

1152 37. Vaidya, A. R. & Badre, D. Abstract task representations for inference and control. *Trends in*
1153 *Cognitive Sciences* **26**, 484–498 (2022).

1154 38. Morton, N. W., Schlichting, M. L. & Preston, A. R. Representations of common event structure
1155 in medial temporal lobe and frontoparietal cortex support efficient inference. *Proceedings of the National*
1156 *Academy of Sciences* **117**, 29338–29345 (2020).

1157 39. Marr, D. Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci* **262**, 23–
1158 81 (1971).

1159 40. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning
1160 systems in the hippocampus and neocortex: insights from the successes and failures of connectionist
1161 models of learning and memory. *Psychol Rev* **102**, 419–457 (1995).

1162 41. Gluck, M. A. & Myers, C. E. Hippocampal mediation of stimulus representation: A
1163 computational theory. *Hippocampus* **3**, 491–516 (1993).

1164 42. Benna, M. K. & Fusi, S. Place cells may simply be memory cells: Memory compression leads to
1165 spatial tuning and history dependence. *Proceedings of the National Academy of Sciences* **118**,
1166 e2018422118 (2021).

1167 43. Tang, W., Shin, J. D. & Jadhav, S. P. Geometric transformation of cognitive maps for
1168 generalization across hippocampal-prefrontal circuits. *Cell Reports* **42**, 112246 (2023).

1169 44. Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J. & Akam, T. Complementary task
1170 representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nat*
1171 *Neurosci* **25**, 1314–1326 (2022).

1172 45. Wood, E. R., Dudchenko, P. A., Robitsek, R. J. & Eichenbaum, H. Hippocampal neurons encode
1173 information about different types of memory episodes occurring in the same location. *Neuron* **27**, 623–
1174 633 (2000).

1175 46. Grieves, R. M., Wood, E. R. & Dudchenko, P. A. Place cells on a maze encode routes rather than
1176 destinations. *eLife* **5**, e15986 (2016).

1177 47. Frank, L. M., Brown, E. N. & Wilson, M. Trajectory Encoding in the Hippocampus and
1178 Entorhinal Cortex. *Neuron* **27**, 169–178 (2000).

1179 48. Sun, C., Yang, W., Martin, J. & Tonegawa, S. Hippocampal neurons represent events as
1180 transferable units of experience. *Nat Neurosci* **23**, 651–663 (2020).

1181 49. Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A. & Fusi, S. Tuned geometries of
1182 hippocampal representations meet the demands of social memory. 2022.01.24.477361 Preprint at
1183 https://doi.org/10.1101/2022.01.24.477361 (2023).

1184 50. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal
1185 cortex. *Nature* **583**, 103–108 (2020).

1186 51. Hesse, J. K. & Tsao, D. Y. The macaque face patch system: a turtle's underbelly for the brain.
1187 *Nat Rev Neurosci* **21**, 695–716 (2020).

1188 52. Tanaka, K. Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience* **19**, 109–
1189 139 (1996).

1190 53. Axelrod, V., Rozier, C., Malkinson, T. S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V. &
1191 Naccache, L. Face-selective neurons in the vicinity of the human fusiform face area. *Neurology* **92**, 197–
1192 198 (2019).

1193 54. Duncan, J. & Owen, A. M. Common regions of the human frontal lobe recruited by diverse
1194 cognitive demands. *Trends Neurosci* **23**, 475–483 (2000).

1195 55. Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C.,
1196 Burgund, E. D., Grimes, A. L., Schlaggar, B. L. & Petersen, S. E. A Core System for the Implementation
1197 of Task Sets. *Neuron* **50**, 799–812 (2006).

1198 56. Ma, L., Chan, J. L., Johnston, K., Lomber, S. G. & Everling, S. Macaque anterior cingulate cortex
1199 deactivation impairs performance and alters lateral prefrontal oscillatory activities in a rule-switching
1200 task. *PLOS Biology* **17**, e3000045 (2019).

1201 57. Monosov, I. E., Haber, S. N., Leuthardt, E. C. & Jezzini, A. Anterior cingulate cortex and the
1202 control of dynamic behavior in primates. *Curr Biol* **30**, R1442–R1454 (2020).

1203 58. Economides, M., Guitart-Masip, M., Kurth-Nelson, Z. & Dolan, R. J. Anterior Cingulate Cortex
1204 Instigates Adaptive Switches in Choice by Integrating Immediate and Delayed Components of Value in
1205 Ventromedial Prefrontal Cortex. *J Neurosci* **34**, 3340–3349 (2014).

1206 59. Cohen, Y., Schneidman, E. & Paz, R. The geometry of neuronal representations during rule
1207 learning reveals complementary roles of cingulate cortex and putamen. *Neuron* **109**, 839-851.e9 (2021).

1208 60. Fu, Z., Wu, D.-A. J., Ross, I., Chung, J. M., Mamelak, A. N., Adolphs, R. & Rutishauser, U.
1209 Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal
1210 Cortex. *Neuron* **101**, 165-177.e5 (2019).

1211 61. Wang, S., Mamelak, A. N., Adolphs, R. & Rutishauser, U. Abstract goal representation in visual
1212 search by neurons in the human pre-supplementary motor area. *Brain* **142**, 3530–3549 (2019).

1213 62. Insausti, R. & Muñoz, M. Cortical projections of the non-entorhinal hippocampal formation in the
1214 cynomolgus monkey (Macaca fascicularis). *Eur J Neurosci* **14**, 435–451 (2001).

1215 63. Aggleton, J. P., Wright, N. F., Rosene, D. L. & Saunders, R. C. Complementary Patterns of
1216 Direct Amygdala and Hippocampal Projections to the Macaque Prefrontal Cortex. *Cereb Cortex* **25**,
1217 4351–4373 (2015).

64. Wang, J., John, Y. & Barbas, H. Pathways for Contextual Memory: The Primate Hippocampal Pathway to Anterior Cingulate Cortex. *Cereb Cortex* **31**, 1807–1826 (2020).

65. Spellman, T., Rigotti, M., Ahmari, S. E., Fusi, S., Gogos, J. A. & Gordon, J. A. Hippocampal–prefrontal input supports spatial encoding in working memory. *Nature* **522**, 309–314 (2015).

66. Kamiński, J., Sullivan, S., Chung, J. M., Ross, I. B., Mamelak, A. N. & Rutishauser, U. Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat Neurosci* **20**, 590–601 (2017).

67. Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. *Proceedings of the National Academy of Sciences* **105**, 329–334 (2008).

68. Rutishauser, U., Ye, S., Koroma, M., Tudusciuc, O., Ross, I. B., Chung, J. M. & Mamelak, A. N. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat Neurosci* **18**, 1041–1050 (2015).

69. Zhou, J., Jia, C., Montesinos-Cartagena, M., Gardner, M. P. H., Zong, W. & Schoenbaum, G. Evolving schema representations in orbitofrontal ensembles during learning. *Nature* **590**, 606–611 (2021).

70. Zhou, J., Montesinos-Cartagena, M., Wikenheiser, A. M., Gardner, M. P. H., Niv, Y. & Schoenbaum, G. Complementary Task Structure Representations in Hippocampus and Orbitofrontal Cortex during an Odor Sequence Task. *Current Biology* **29**, 3402-3409.e3 (2019).

71. Badre, D. & D'Esposito, M. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat Rev Neurosci* **10**, 659–669 (2009).

72. Badre, D. & Nee, D. E. Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences* **22**, 170–188 (2018).

73. Cole, M. W., Laurent, P. & Stocco, A. Rapid instructed task learning: a new window into the human brain's unique capacity for flexible cognitive control. *Cogn Affect Behav Neurosci* **13**, 1–22 (2013).

74. Brainard, D. H. The Psychophysics Toolbox. *Spatial Vision* **10**, 433–436 (1997).

75. Tyszka, J. M. & Pauli, W. M. In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. *Human Brain Mapping* **37**, 3979–3998 (2016).

76. Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods* **154**, 204–224 (2006).

77. Hill, D. N., Mehta, S. B. & Kleinfeld, D. Quality Metrics to Accompany Spike Sorting of Extracellular Signals. *J. Neurosci.* **31**, 8699–8705 (2011).

78. Courellis, H., Nummela, S., Miller, C. & Cauwenberghs, G. A computational framework for effective isolation of single-unit activity from in-vivo electrophysiological recording. in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)* 1–4 (2017). doi:10.1109/BIOCAS.2017.8325164.

Figure 1

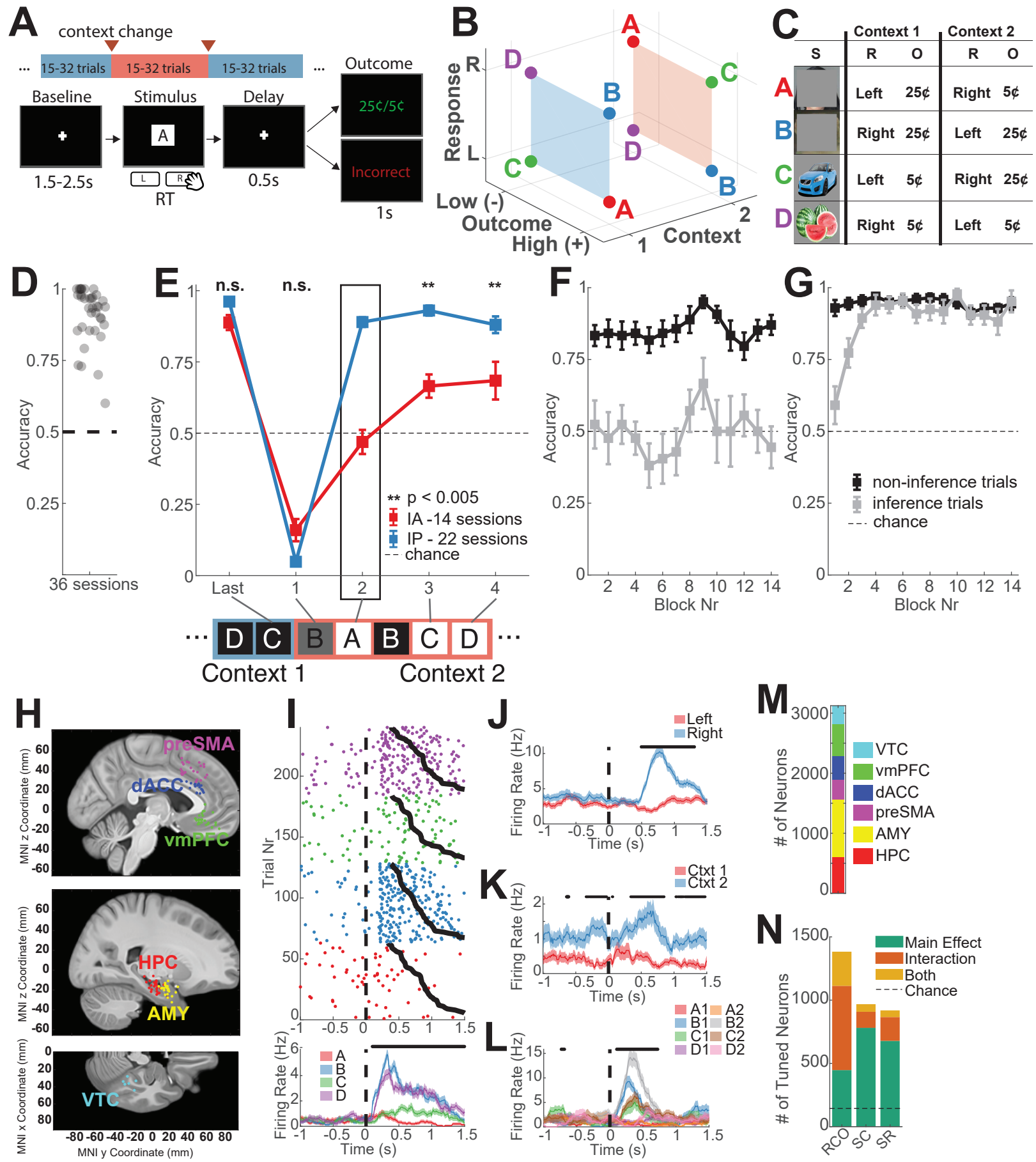

A
context change
15-32 trials
15-32 trials
15-32 trials
Baseline
Stimulus
Delay
Outcome
25¢/5¢
Incorrect
1.5-2.5s
RT
0.5s
1s
B
Response
R
L
Low (-)
Outcome
High (+)
Context
1
2
C
Context 1
Context 2
S
R
O
R
O
Left 25¢ Right 5¢
Right 25¢ Left 25¢
Left 5¢ Right 25¢
Right 5¢ Left 5¢
D
Accuracy
36 sessions
E
n.s.
n.s.
**
**
** p < 0.005
IA -14 sessions
IP - 22 sessions
chance
Last
Context 1
Context 2
F
Block Nr
G
non-inference trials
inference trials
chance
H
preSMA
dACC
vmPFC
HPC
AMY
VTC
MNI z Coordinate (mm)
MNI x Coordinate (mm)
MNI y Coordinate (mm)
I
Trial Nr
Firing Rate (Hz)
Time (s)
J
Left
Right
K
Ctxt 1
Ctxt 2
L
A1 A2
B1 B2
C1 C2
D1 D2
M
# of Neurons
VTC
vmPFC
dACC
preSMA
AMY
HPC
N
# of Tuned Neurons
Main Effect
Interaction
Both
Chance
RCO
SC
SR

Figure 2

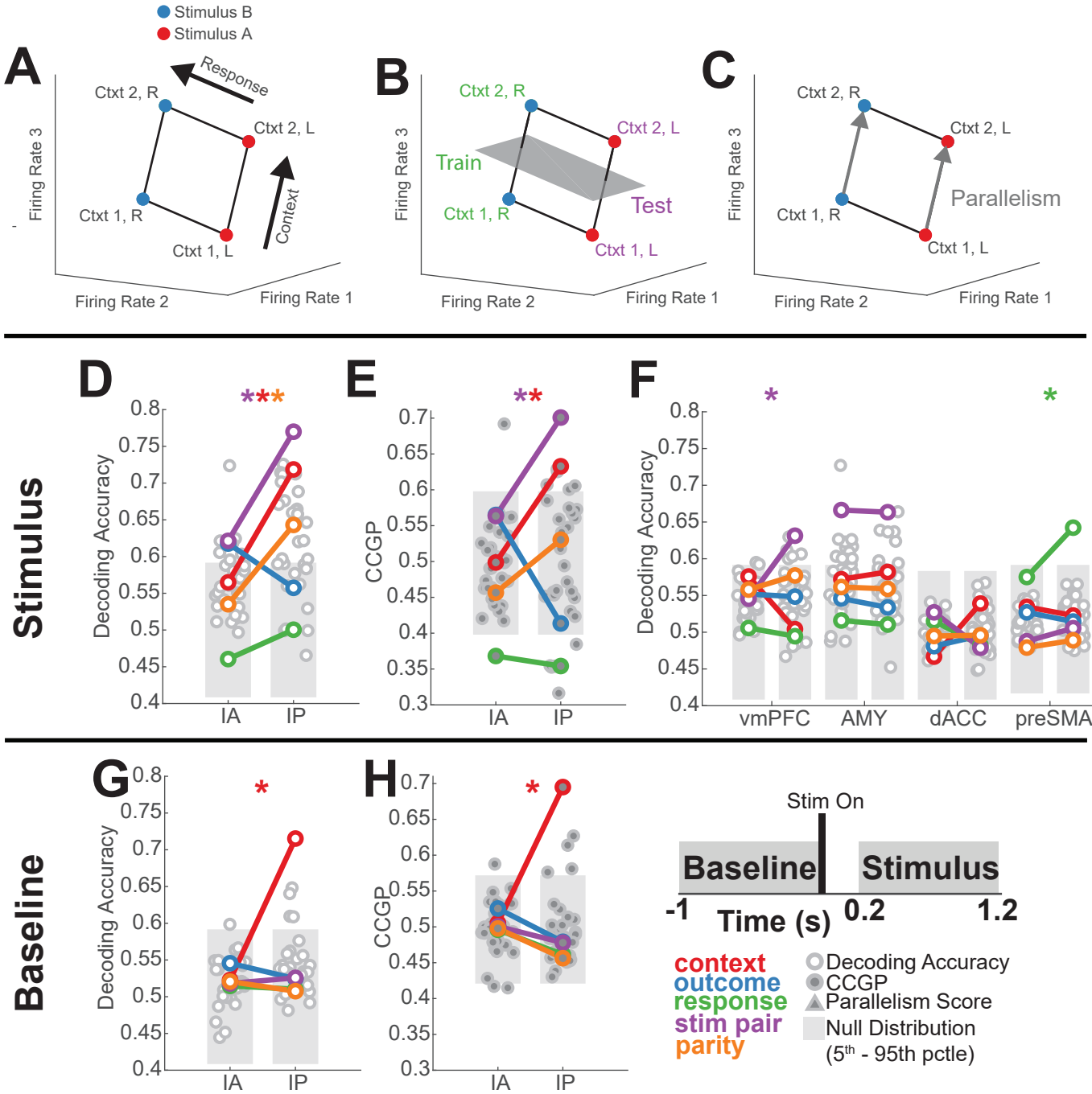

Stimulus B
Stimulus A
A
Response
Context
Ctxt 2, R
Ctxt 2, L
Ctxt 1, R
Ctxt 1, L
Firing Rate 3
Firing Rate 2
Firing Rate 1
B
Train
Test
C
Parallelism
Stimulus
Baseline
D
E
F
G
H
Decoding Accuracy
CCGP
IA
IP
vmPFC
AMY
dACC
preSMA
Stim On
Baseline
Stimulus
-1
0.2
1.2
Time (s)
context
outcome
response
stim pair
parity
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

Figure 3

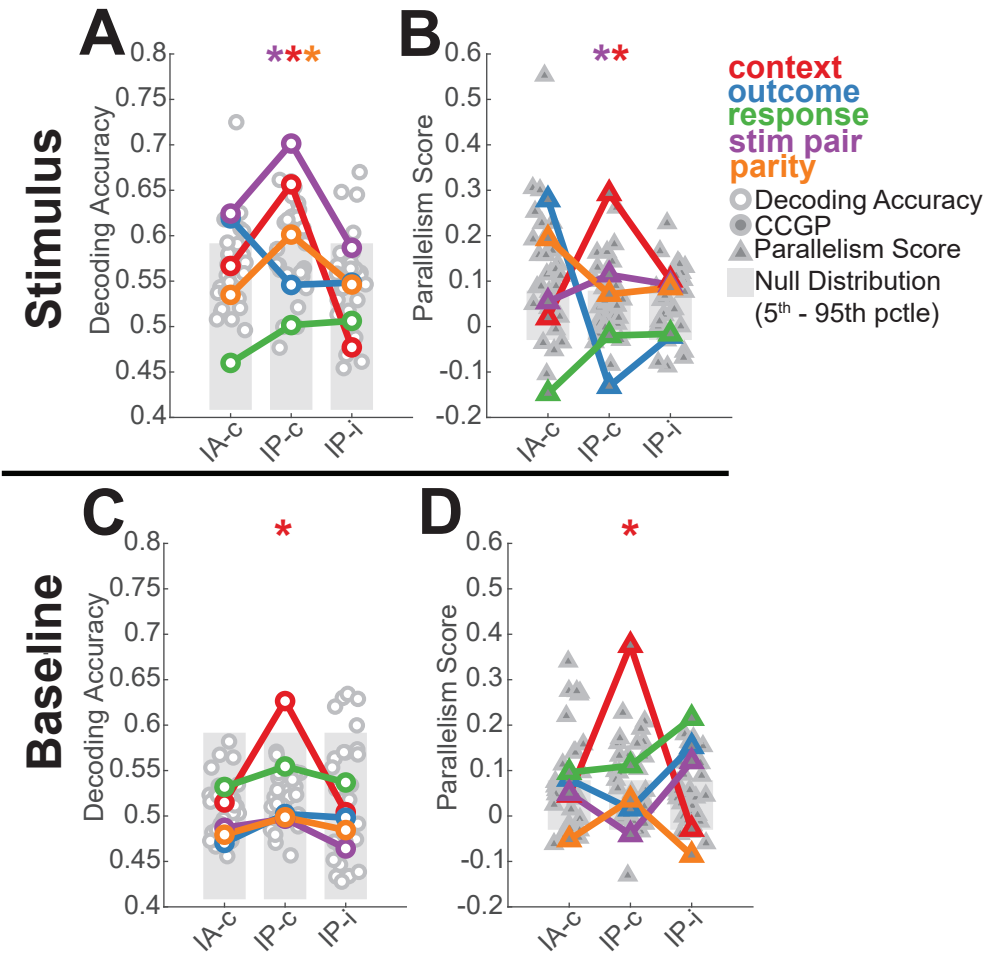

A
B
C
D
Stimulus
Baseline
Decoding Accuracy
Parallelism Score
IA-c
IP-c
IP-i
context
outcome
response
stim pair
parity
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

Figure 4

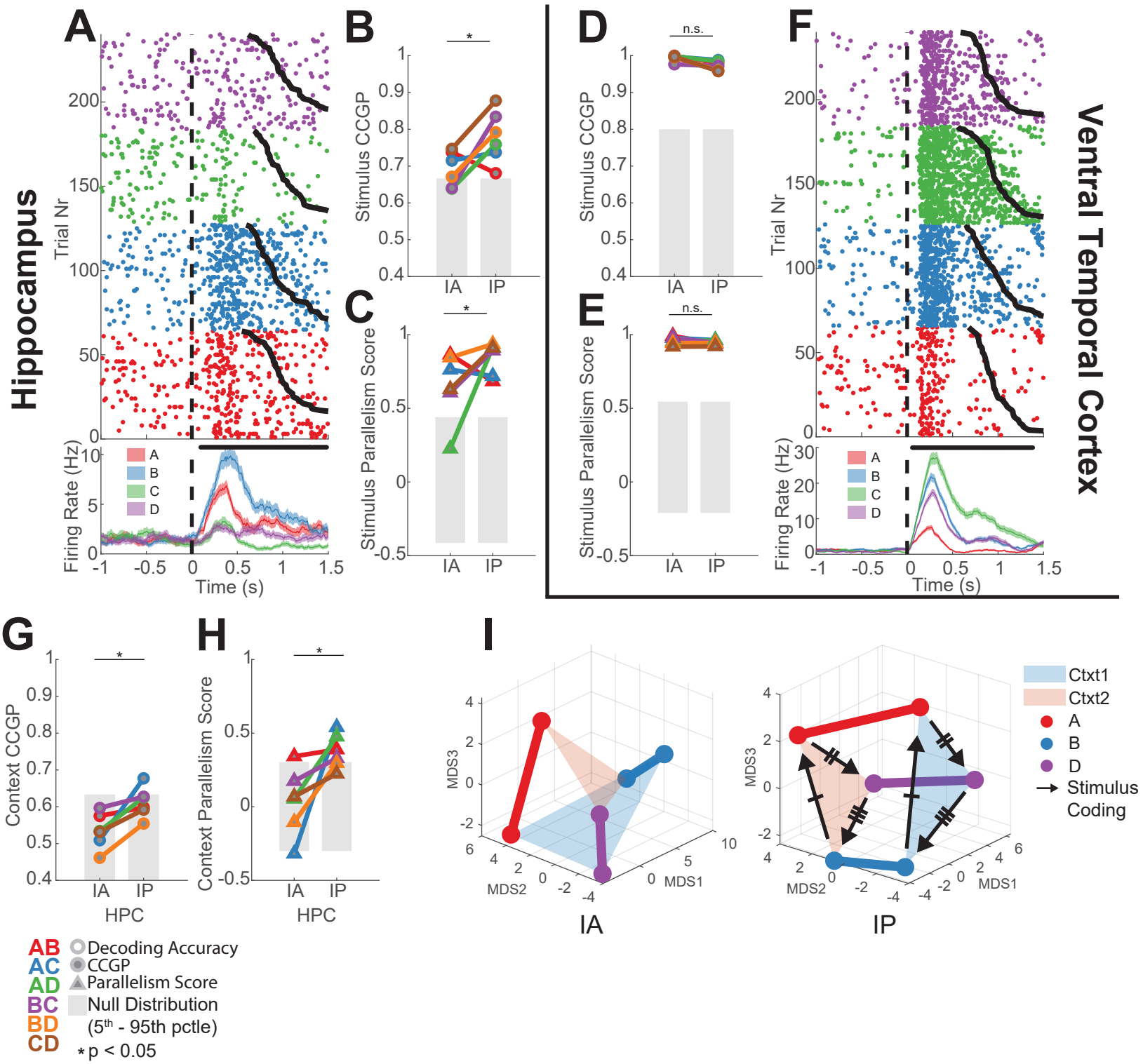

A
B
C
D
E
F
G
H
I
Hippocampus
Ventral Temporal Cortex
Trial Nr
Firing Rate (Hz)
Time (s)
A
B
C
D
Stimulus CCGP
Stimulus Parallelism Score
IA
IP
n.s.
*
Context CCGP
Context Parallelism Score
HPC
MDS1
MDS2
MDS3
Ctxt1
Ctxt2
Stimulus Coding
AB
AC
AD
BC
BD
CD
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)
* p < 0.05

Figure 5

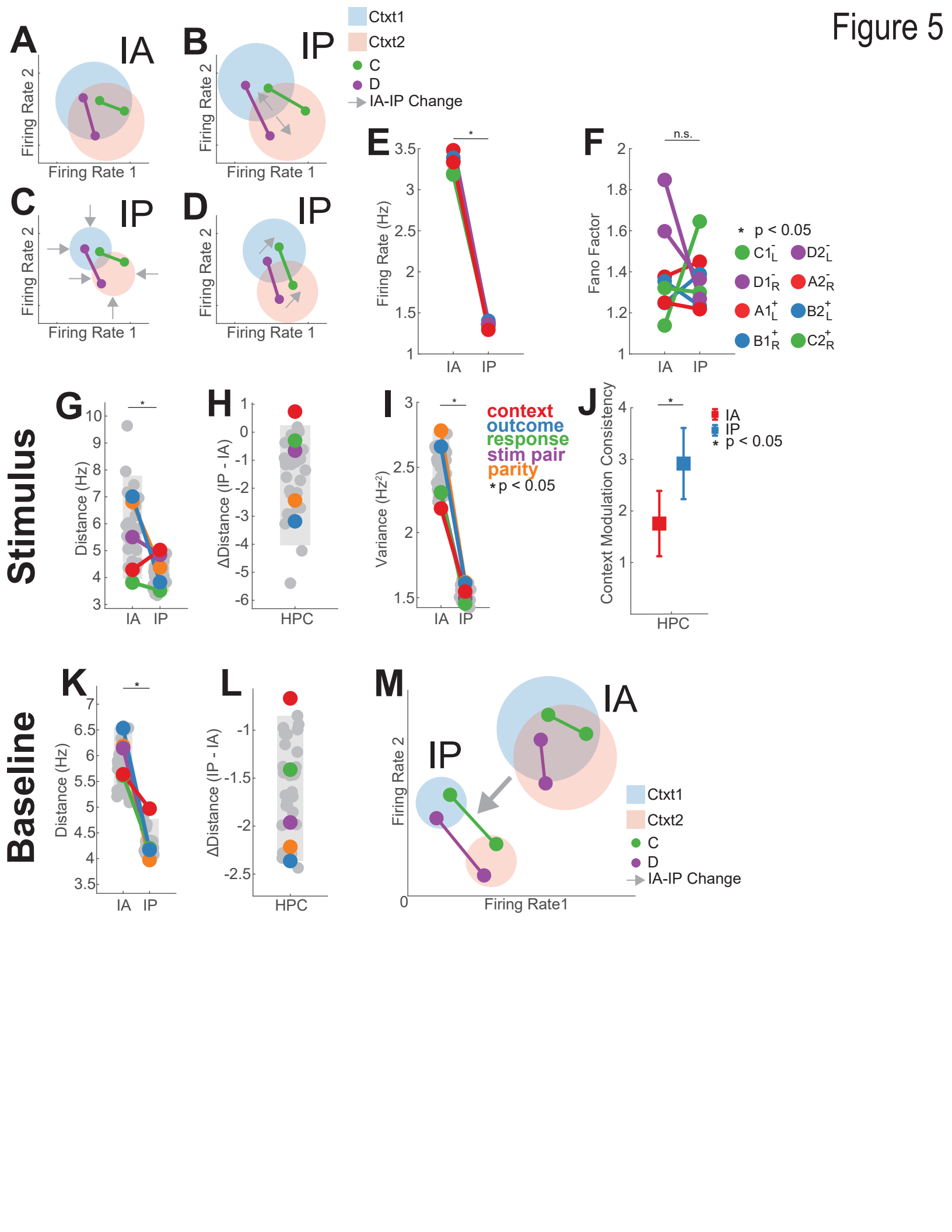

A
IA
B
IP
C
IP
D
IP
Ctxt1
Ctxt2
C
D
IA-IP Change
Firing Rate 2
Firing Rate 1
E
Firing Rate (Hz)
IA
IP
F
Fano Factor
n.s.
* p < 0.05
Stimulus
G
Distance (Hz)
H
ΔDistance (IP - IA)
HPC
I
Variance (Hz²)
context
outcome
response
stim pair
parity
* p < 0.05
J
Context Modulation Consistency
IA
IP
* p < 0.05
Baseline
K
Distance (Hz)
L
ΔDistance (IP - IA)
M
Firing Rate2
Firing Rate1

Figure 6

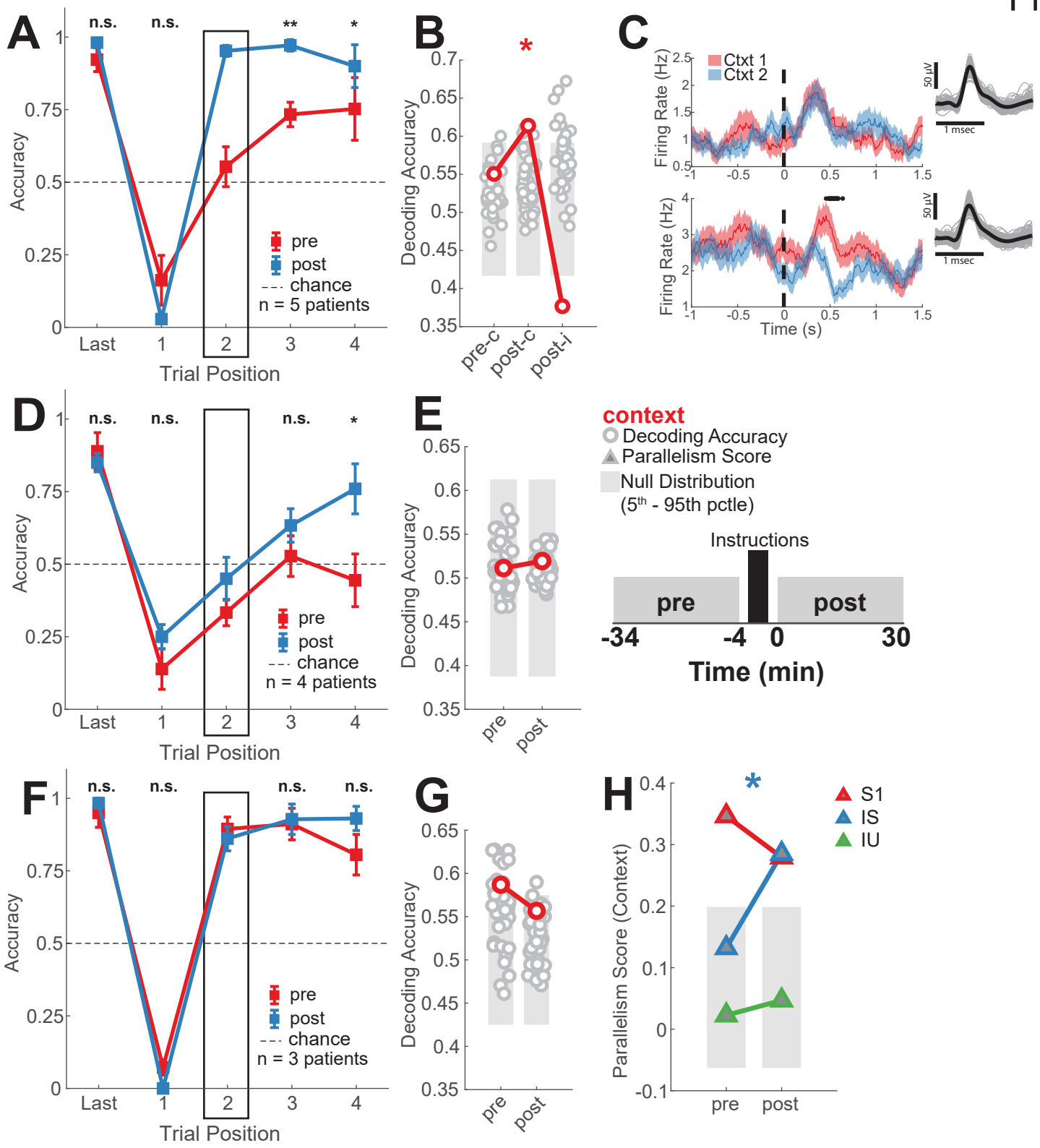

A
n.s.
n.s.
**
*
Accuracy
pre
post
chance
n = 5 patients
Last
1
2
3
4
Trial Position
B
*
Decoding Accuracy
pre-c
post-c
post-i
C
Ctxt 1
Ctxt 2
Firing Rate (Hz)
50 µV
1 msec
Time (s)
D
n.s.
n.s.
n.s.
*
Accuracy
pre
post
chance
n = 4 patients
Trial Position
E
Decoding Accuracy
pre
post
context
Decoding Accuracy
Parallelism Score
Null Distribution
(5th - 95th pctle)
Instructions
pre
post
-34
-4
0
30
Time (min)
F
n.s.
n.s.
n.s.
n.s.
Accuracy
pre
post
chance
n = 3 patients
Trial Position
G
Decoding Accuracy
pre
post
H
*
S1
IS
IU
Parallelism Score (Context)
pre
post

Figure 1. Task, behavior, recording locations, and single-neuron tuning. (A) The task consisted of variable-length blocks (15-32 trials) that alternated between two latent contexts (red and blue). Context changes (red arrows) were covert. Trials consisted of a pre-stimulus baseline followed by stimulus presentation during which patients executed the associated response (left or right button press) in a speeded manner. After button press, the stimulus was replaced with a fixation cross, followed by the outcome (either high/low reward or incorrect) was presented after a fixed 0.5s delay. **(B-C)** Illustration of the task structure. Each stimulus (A-D) is associated with a single correct response and results in either a high or low reward if the correct response is given. All stimulus-response-reward relationships are inverted between context 1 (blue) and 2 (orange). This visualization is reflective of the disentangled structure of the task variables, and does not necessarily reflect how neurons will organize their responses in neural state-space to each of these conditions. **(C)** Example images (left) associated with the stimuli A-D. Note: stimuli A and B are masked due to copyright. These associations are randomized for every session. **(D)** Patients exhibited high accuracy on non-inference (baseline) trials. Each dot corresponds to the average non-inference trial performance over a single session. Black dashed line indicates chance. Only sessions where patients exhibited above-chance accuracy on non-inference trials are shown (36/42 sessions, $p < 0.05$, Binomial Test on all non-inference trials). **(E)** Task performance split by whether inference was present (IP) or absent (IA) on the first inference trial following context switches in a given session. Sessions where inference performance was significantly above chance (22 sessions, $p < 0.05$, Binomial Test on inference trial 1) were deemed "inference present" (IP, blue), and those where inference performance was not above chance (14 sessions, $p > 0.05$, Binomial Test on inference trial 1) were considered "inference absent" (IA, red). Plot shows performance on the last trial before the context switch, the first trial after the context switch, and for the remaining three inference trials averaged over all trials in each session (mean ± s.e.m. across sessions). Dashed line marks chance. Black box indicates inference trial 1. **(F-G)** Task performance as a function of time in the task for the **(F)** inference absence and **(G)** inference present groups. Shown is the accuracy for the last non-inference trial before a switch (black) and the first inference trial after a switch (gray). Accuracy is shown block-by-block averaged over a 3-block window (mean ± s.e.m. across sessions). **(H)** Electrode locations. Each dot corresponds to a single microwire-bundle. Locations are shown on the same hemisphere (right) for visualization purposes only. Shown are pre-Supplementary Motor Area (preSMA, purple), dorsal Anterior Cingulate Cortex (dACC, blue), ventromedial Prefrontal Cortex (vmPFC, green), Hippocampus (HPC, red), Amygdala (AMY, yellow), and Ventral Temporal Cortex (VTC, teal). **(I-L)** Example neurons tuned to task variables. **(I)** Hippocampal neuron that encodes stimulus identity. Raster trials are reordered based on stimulus identity, and sorted by reaction time therein (black curves). Stimulus onset occurs at time 0. Black points above PSTH indicate times where 1-way ANOVA over the plotted task variables was significant ($p < 0.05$). **(J-L)** PSTH of three other example neurons that encode response **(J)**, context **(K)**, and mixtures of stimulus id and context **(L)**. **(M)** Number of single units recorded across all brain areas (3124 neurons recorded in total). **(N)** Number of single units across all brain areas exhibiting significant Main effects or interaction effects (n-way ANOVA with interactions, $p < 0.05$, see methods) to at least one of the principal task variables (R = Response, C = Context, O = Outcome, S = Stimulus ID) or to combinations of variables. A unit is linearly tuned if it has at least one significant main effect, and non-linearly tuned if it has at least one significant interaction term in the ANOVA model.

Figure 2. Emergence of multiple abstract variables in Hippocampus supports inference.

(A) Simplified example of a neural state space where each axis is the firing rate of one neuron. Points correspond to the response of the neurons to different task states, i.e. two stimuli (red and blue) that elicit two responses (R and L) in two contexts. Note the coding vectors for response and context (black arrows) are not aligned with the axes as each neuron might respond to mixtures of variables. The axes of this state space differ from those shown in Fig. 1B, with the latter being defined by experimenter-selected variables rather than neural firing rates.

(B) Example of cross-condition generalization. A decoder is trained to classify context only on response "R" conditions (green) and is evaluated on its ability to decode context on response "L" conditions (purple). If context is represented in an abstract format (i.e. disentangled from response), then the decoder should generalize to the held-out response condition, yielding a high cross-condition generalization performance (CCGP) for context.

(C) Example of context parallelism. Coding vectors for context (gray arrows) are parallel, indicating that the coding direction for context is identical for different responses, and thus that context and response are disentangled. Details for computing the parallelism score (PS) of a balanced dichotomy with 8 conditions are provided in the methods.

(D-F) During the stimulus presentation period, context (red) and stimulus pair (purple) become decodable in the IP group in the HPC. Context is encoded in an abstract format. Decoding accuracy **(D)** and CCGP **(E)** are shown for all 35 balanced dichotomies during the stimulus period (0.2 to 1.2 s following stimulus onset, see inset). A subset of the dichotomies are named (color code) because they represent task specific variables (see Fig. S2). Swarm plots for decoding accuracy and CCGP are light circles and dark circles respectively. Gray bars denote the 5th-95th percentile of the shuffle-null distribution for decoding accuracy and geometric null distribution for CCGP. Stars denote named dichotomies that are above chance in IP and are significantly different from their corresponding IA value ($p_{RS} < 0.05/35$, Ranksum Test, Bonferroni corrected for multiple comparisons across all dichotomies).

(F) Identical analysis to **(D)** showing decodability of balanced dichotomies from neurons recorded in other brain regions (except VTC, which is shown in Fig. 4 and S8).

(G-H) Same as for **(D-F)**, but for spikes counted during the baseline period prior to stimulus onset. Context (red) becomes decodable in IP sessions and is in an abstract format. Trials are labeled according to the current trial. Decoding accuracy **(G)** and CCGP **(H)** computed in HPC for all balanced dichotomies with spikes counted during the pre-stimulus baseline period (-1 to 0s prior to stimulus onset, see inset). All plotting conventions identical to those in **(D-F)**, except Baseline analysis is conducted with task variables from previous trial.

Note: all reported geometric measures, decoding accuracies, or angles are the average of 1000 runs with condition-wise trial resampling as described in the methods. All null distributions are constructed from 1000 iterations of shuffled trial-resampling using either trial-label shuffling (shuffle null) or random rotations designed to destroy low-dimensional structure (geometric null). Also, neuron counts are balanced between IA and IP for every brain area to ensure that dimensionally-sensitive values (e.g. vector angles, decoding accuracies, etc..) are directly comparable. See methods for details.

Figure 3. Geometry of hippocampal context representations is compromised during error trials.

(A-B) Following stimulus onset, context (red) is not decodable (A) and not in an abstract format (B) in incorrect trials occurring during IP sessions. Decoding accuracy and parallelism scores are estimated separately during correct and incorrect trials in IP sessions (IP-c and IP-i, respectively) and in correct trials only during IA sessions (IA-c). Stars denote named dichotomies that are above chance in the IP-c trials and are significantly different from their corresponding IA-c value ($p < 0.05/35$, Ranksum Test, Bonferroni multiple comparison corrected across all dichotomies). **(C-D)** Same as **(A-B)**, but estimated for spikes counted during the baseline period and task states from previous trial. Note: IP-c values here are slightly lower than the IP values reported in Fig. 2 because additional neurons were removed by subsampling to equalize the number of neurons used for both correct and incorrect trials.

Figure 4. Stimulus representations become structured around context with inference in HPC but not VTC.

(A-C) Responses in HPC following stimulus onset carry information about stimulus identity. **(A)** Example neuron in the HPC. Trials are reordered by stimulus identity, and are sorted by reaction time therein (black line).

(B,C) Stimulus geometry across contexts, with geometric analysis conducted over pairs of stimuli in each context. Data points shown correspond to different stimulus pairs (color coded, see lower right for legend). Significance of differences is tested using RankSum comparing IA and IP over all stimulus pairs (* indicates $p < 0.05$, n.s. otherwise). All other conventions identical to those in Fig. 2,3.

(B) CCGP ($p_{RS} = 0.041$) and **(C)** parallelism score ($p_{RS} = 0.040$) for stimulus coding across contexts significantly increased in IP compared to IA.

(D-F) Same as **(A-C)**, but for VTC.

(D) CCGP ($p_{RS} = 0.15$) and **(E)** parallelism score ($p_{RS} = 0.39$) for stimulus coding across contexts does not differ significantly between IA and IP.

(G-H) Context encoding across stimulus pairs for HPC. Plotting conventions are identical to those in panels (B-E). **(G)** CCGP for context across stimuli ($p_{RS} = 0.012$) and **(H)** parallelism score for context coding vectors between pairs of stimuli ($p_{RS} = 0.015$) both significantly increase from IA to IP.

(I) Changes in neural geometry in HPC. MDS of condition-averaged responses of all recorded HPC neurons shown for IA (left) and IP (right). Colored points are average population vector responses to a stimulus (point color) in each context (plane color). Stimuli of the same identity in either context are connected by a line of the same color. Black arrows represent the coding direction between two different stimuli in each context. Comparing left with right, note that context emerges as a decodable variable (red and blue planes separate), and both stimulus coding (black arrows) and context coding (colored lines) are parallel reflecting the abstract format of both variables.

Note: only three stimuli are shown here since the full stimulus ID space is 3D (4 unique, unrelated stimuli) and context is encoded in its own independent dimension, thus requiring 4 dimensions to fully visualize the neural geometry of the hippocampus during inference. All 2D projections of stimulus pairs are shown in Fig. S12.

Figure 5. Firing rate properties underlying the observed changes of population-level hippocampal neural geometry.

(A-D) Illustration of different hypothesized firing rate pattern changes that could give rise to the observed population level geometry changes. The four different hypotheses are illustrated with two stimuli (C and D), each present in two different contexts (blue and red). Condition responses are defined by the firing rates of two hypothetical neurons. The solid-colored points represent the condition average for each stimulus and the larger shaded circles represent the trial-by-trial variation of the two neurons for each context. Gray arrows signify changes that have occurred in IP plots **(B-D)** relative to the IA plot **(A)**.

These response of the neurons to the stimuli during IA **(A)** can be shaped by **(B)** increasing the distance between the context centroids, **(C)** decreasing the variance along the coding direction in the absence of changes in distance, or **(D)** straightening the neural responses without changing distances/variances so that the geometry becomes more orthogonalized.

(E) Changes in hippocampal firing rate from IA to IP. Points correspond to the average firing rate over neurons in the HPC for each of 8 unique task conditions, and are colored according to stimulus identity in that condition (e.g. task condition $C1_L^-$ describes: stimulus C, context 1, outcome $-$, response L). Neuronal firing rates were lower during IP compared to IA sessions ($p_{RS} = 8.3x10^{-5}$, RankSum over conditions).

(F) Same as **(E)**, but for condition-wise fano factors. Fano factor (FF) here is computed as the ratio of the condition-wise variance and the condition-averaged firing rate, computed by neuron and averaged over neurons. Points correspond to average FF over all hippocampal neurons. There was no significant difference (RankSum over conditions between IA and IP ($p_{RS} = 0.99$)).

(G) Population distances between centroids for all 35 balanced dichotomies. The colored connected points represent distances for the named dichotomies indicated in the legend to the right. Gray bars indicate the 5th-95th percentile of the geometric null distribution. Across all dichotomies, distances decreases from IA to IP ($p_{RS} = 2.9x10^{-8}$, RankSum over dichotomies).

(H) Context alone is the only dichotomy whose distance significantly increases from IA to IP **(**red, $p_{\Delta Dist} = 0.040$**)**. The null distribution shown H is the distribution of differences between the IP and IA null distributions shown in.

(I) Average variance projected along the coding direction decreased on average between IA and IP sessions ($p_{RS} = 6.5x10^{-13}$). Variance was computed in a cross-validated manner (see methods) resulting in a distribution of trial-by-trial population activity along each dichotomy coding direction from which the coding variance was computed. The 5th-95th percentile of the geometric null distribution was also used here for the null distribution.

(J) Change in the consistency of context-modulation for stimuli averaged over all neurons in HPC. Greater context modulation consistency for individual neurons results in greater parallelism score for context at the population level. HPC neurons on average exhibit a significant increase in context modulation consistency between IA and IP sessions ($p_{RS} = 0.0039$) during the stimulus period.

(K,L) Same as **(G,H)**, but for spike counts during the baseline period and grouping trials by task state of the previous trial. Distance was significantly reduced across all dichotomies (**K**, $p_{RS} = 6.4x10^{-13}$, RankSum over dichotomies) and context alone exhibits a distance reduction that is smaller than would be expected by chance (**L,** red, $p_{\Delta Dist} = 0.027$)

(M) Illustration of implementational changes to neural state space using the conventions introduced in **(A-D)**. We find that, when comparing IA with IP sessions, that (i) context dichotomy distance increased (indicated by the increased distance between the red and blue shaded circles), (ii) variance decreased due to a reduction in firing rate (indicated by decreased shaded circle radius and movement towards the origin of state space), and (iii) an increase in the consistency of stimulus modulation across contexts (indicated by lines becoming parallel).

Figure 6. Abstract hippocampal representation of context is present following successful verbal instructions about latent context.

(A,D,F) Behavioral performance shown separately for pre-instruction and post-instruction sessions (pre and post). **(A)** Inference performance for the instruction-successful (IS) patient group (5 patients, 10 sessions). **(C)** Example hippocampal neuron with univariate context encoding in the post (bottom) but not the pre instruction (top) session. (one-way ANOVA, $p_{pre} = 0.40, p_{post} = 0.010$). **(D)** Inference performance for the instruction-unsuccessful (IU) patient group (4 patients, 8 sessions). **(F)** Inference performance for the session one inference (S1) patient group (3 patients, 6 sessions).

(B,E,G) Encoding of context in the stimulus period in pre-and post-instruction sessions during pre-instruction correct trials (pre-c), post-instruction correct trials (post-c) and post-instruction incorrect trials (post-i). The first trial following a switch is excluded from this analysis. **(E, inset)** Schematic of the recording procedure, showing a pre- and post-instruction session shaded in gray (30 min duration), with a 4 minute inter-session break (mean duration = 241 s, range 102-524s) during which instructions detailing task structure were provided. **(B)** Context emerges as significantly decodable post-instructions but not pre-instructions in the IS group in a task-performance-dependent manner ($p_{pre-c} = 0.17$, $p_{post-} = 0.016, p_{RS} = 3.1x10^{-19}, p_{post-i} = 0.99$). **(E)** Context is not significantly decodable in the IU group neither pre- nor post-instructions ($p_{pre} = 0.44$, $p_{post} = 0.42$). **(G)** Context is decodable both pre- and post-instructions in the S1 group ($p_{pre} = 0.014, p_{post} = 0.17$).

(H) Summary of changes in PS for context for all three subject groups (S1 red, IS blue, IU green). Neuron counts are sub-sampled to match across all groups so that PS values are directly comparable. Significant increase in context PS from pre to post is indicated for the IS group ($p_{IS,pre} = 0.20, p_{IS,post} = 0.0028$), but not for the IU ($p_{IU,pre/post} < 0.5$) and S1 ($p_{S1,pre/post} < 0.005$) groups.

Figure S1

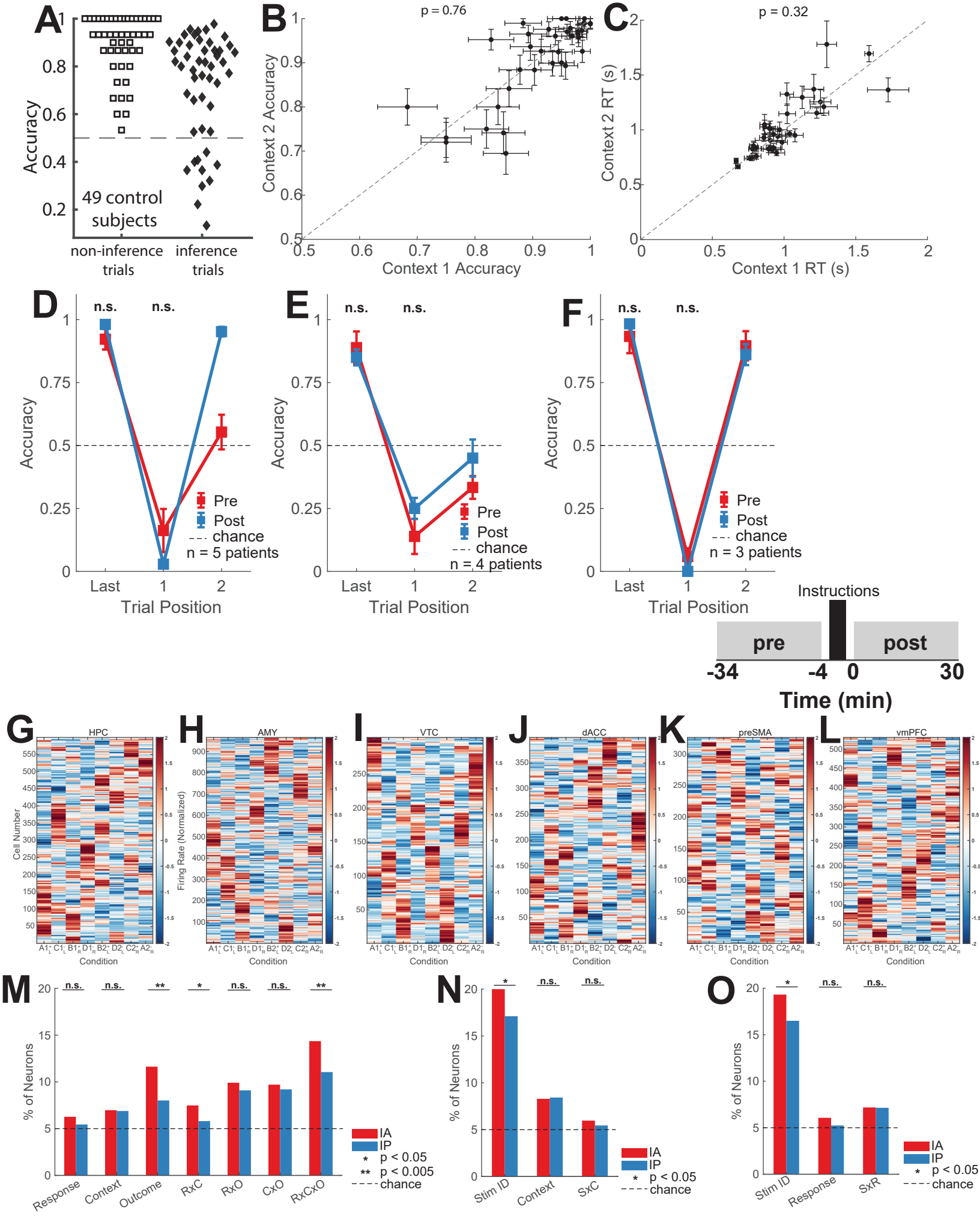

A
Accuracy
49 control subjects
non-inference trials
inference trials
B
p = 0.76
Context 2 Accuracy
Context 1 Accuracy
C
p = 0.32
Context 2 RT (s)
Context 1 RT (s)
D
n.s.
n.s.
Accuracy
Pre
Post
chance
n = 5 patients
Last
1
2
Trial Position
E
n.s.
n.s.
Accuracy
Pre
Post
chance
n = 4 patients
Trial Position
F
n.s.
n.s.
Accuracy
Pre
Post
chance
n = 3 patients
Trial Position
Instructions
pre
post
-34
-4
0
30
Time (min)
G
HPC
Cell Number
Condition
H
AMY
Firing Rate (Normalized)
Condition
I
VTC
Condition
J
dACC
Condition
K
preSMA
Condition
L
vmPFC
Condition
M
n.s.
n.s.
**
*
n.s.
n.s.
**
% of Neurons
Response
Context
Outcome
RxC
RxO
CxO
RxCxO
IA
IP
* p < 0.05
** p < 0.005
chance
N
*
n.s.
n.s.
% of Neurons
Stim ID
Context
SxC
IA
IP
* p < 0.05
chance
O
*
n.s.
n.s.
% of Neurons
Stim ID
Response
SxR
IA
IP
* p < 0.05
chance

Figure S2

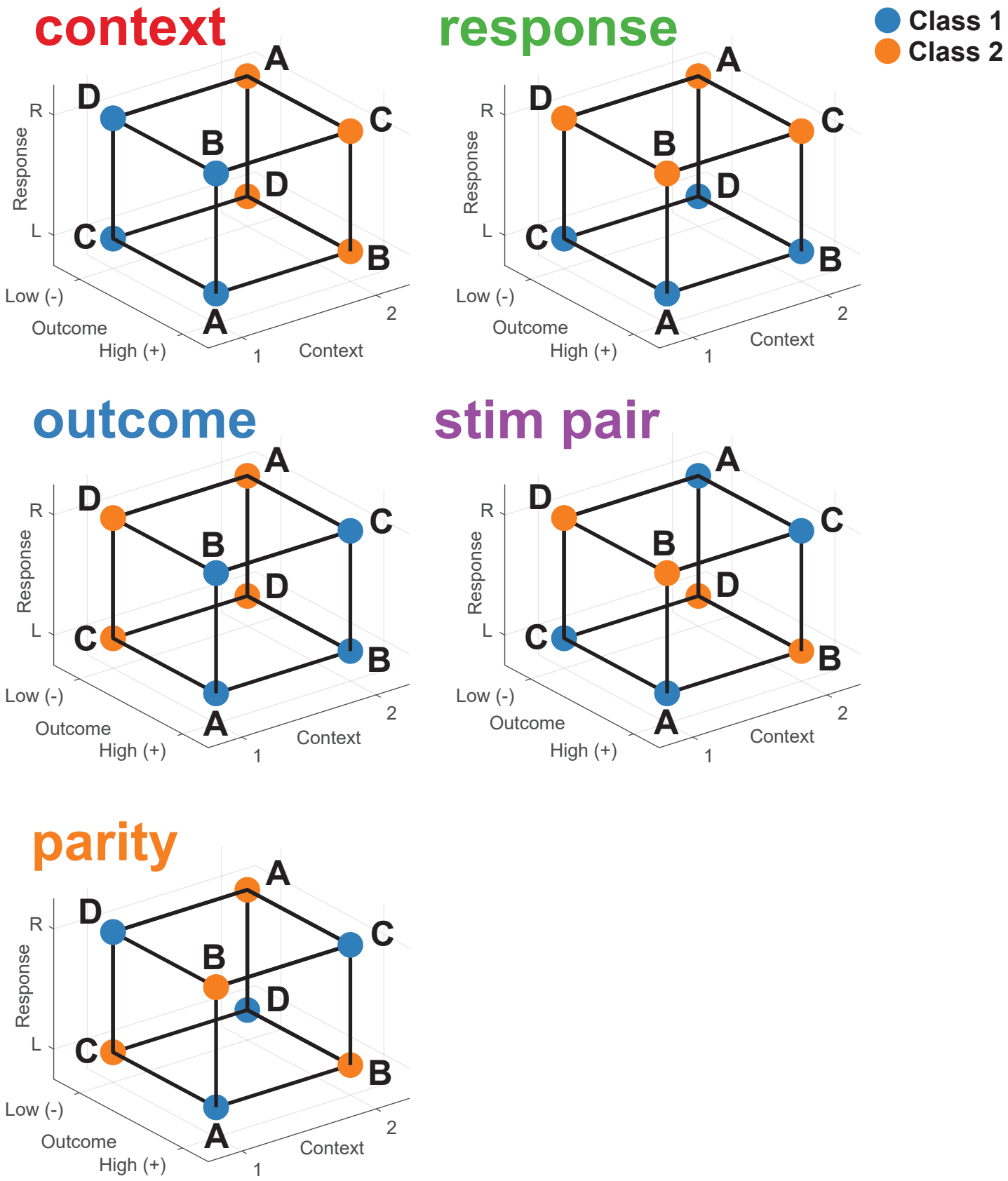

context
response
outcome
stim pair
parity
Class 1
Class 2
Response
R
L
Low (-)
Outcome
High (+)
Context
1
2
A
B
C
D

Figure S3

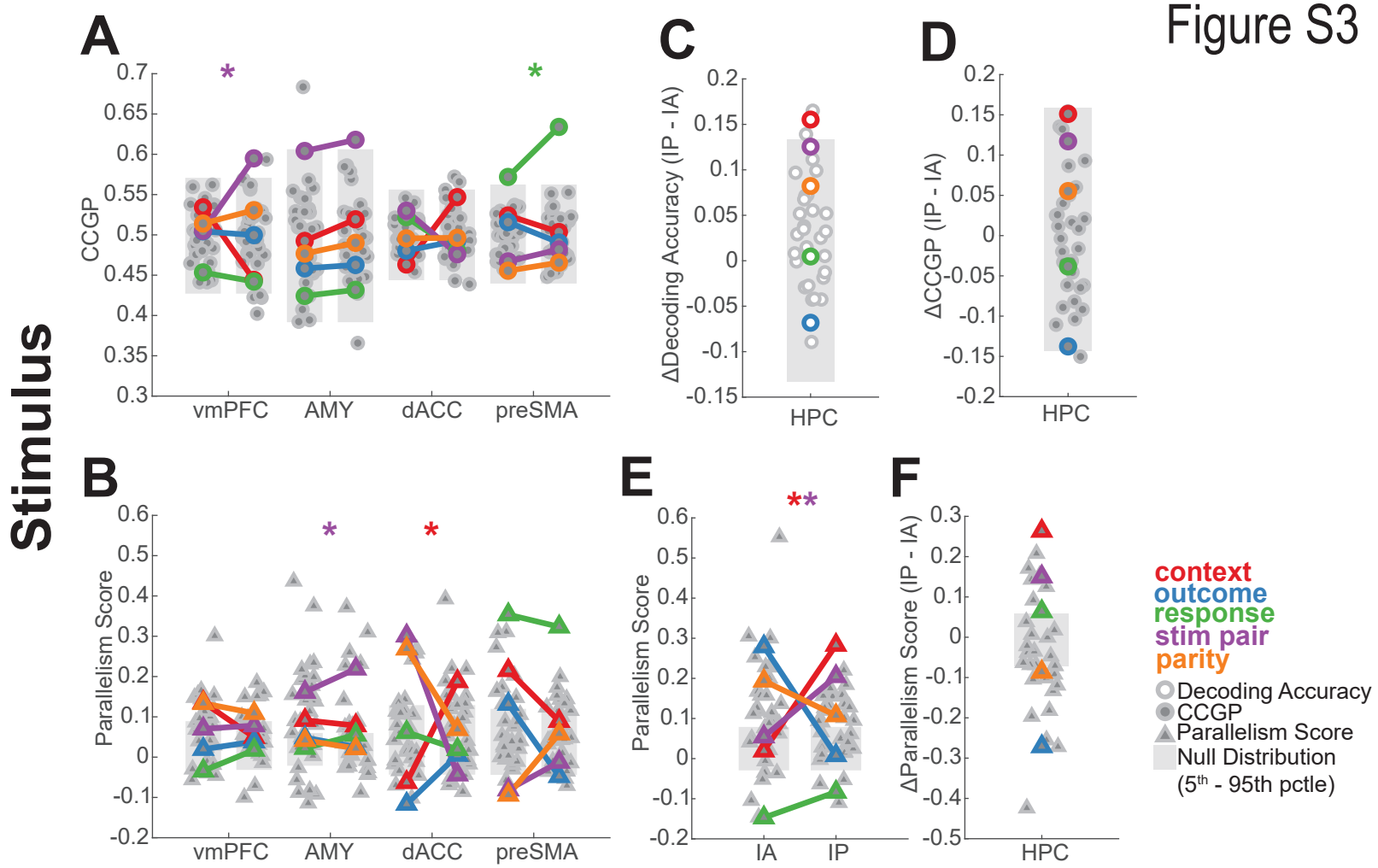

Stimulus
A
CCGP
vmPFC
AMY
dACC
preSMA
B
Parallelism Score
C
ΔDecoding Accuracy (IP - IA)
HPC
D
ΔCCGP (IP - IA)
E
IA
IP
F
ΔParallelism Score (IP - IA)
context
outcome
response
stim pair
parity
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

Figure S4

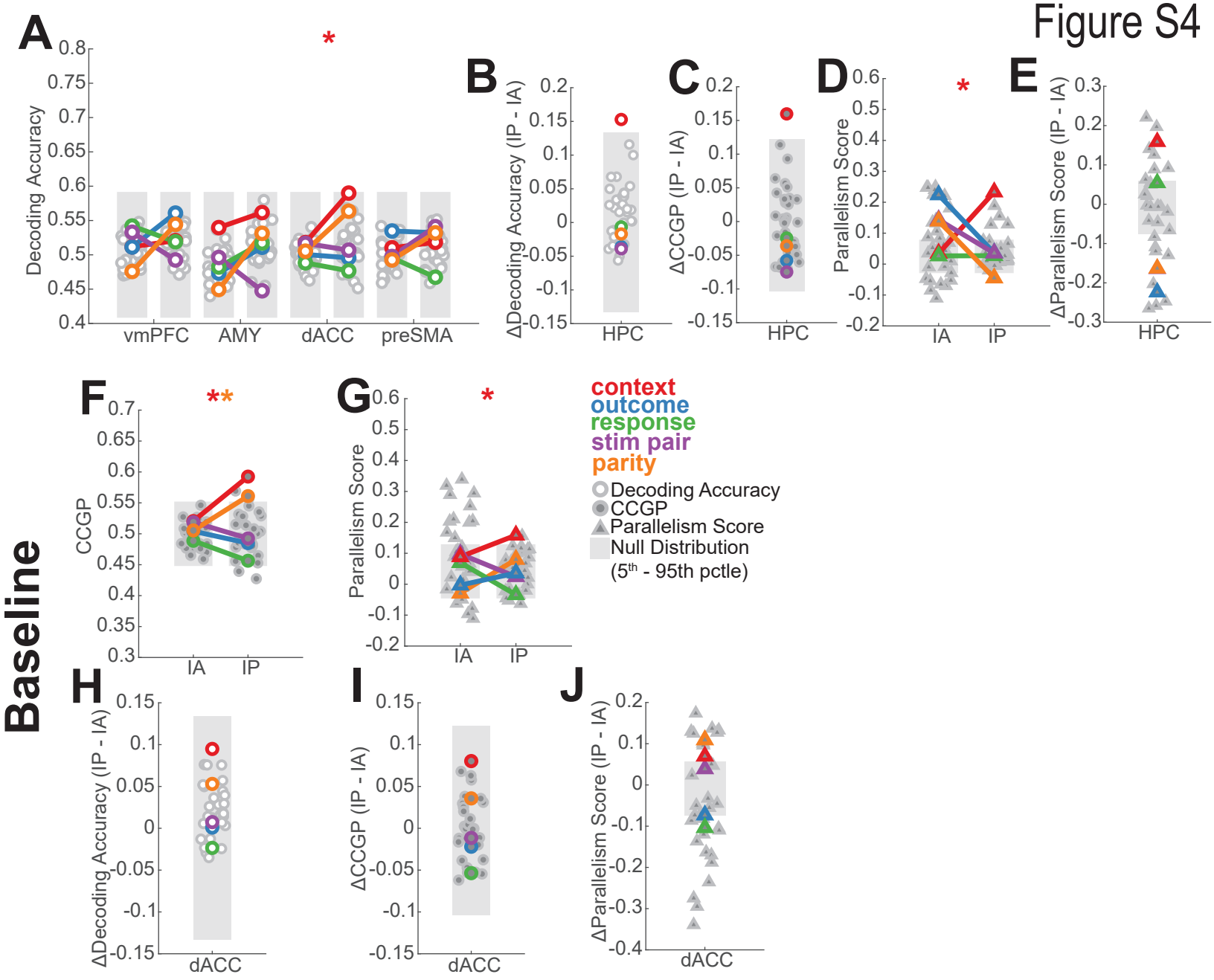

A
B
C
D
E
F
G
H
I
J
Baseline
Decoding Accuracy
vmPFC
AMY
dACC
preSMA
ΔDecoding Accuracy (IP - IA)
HPC
ΔCCGP (IP - IA)
Parallelism Score
IA
IP
ΔParallelism Score (IP - IA)
CCGP
context
outcome
response
stim pair
parity
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

Figure S5

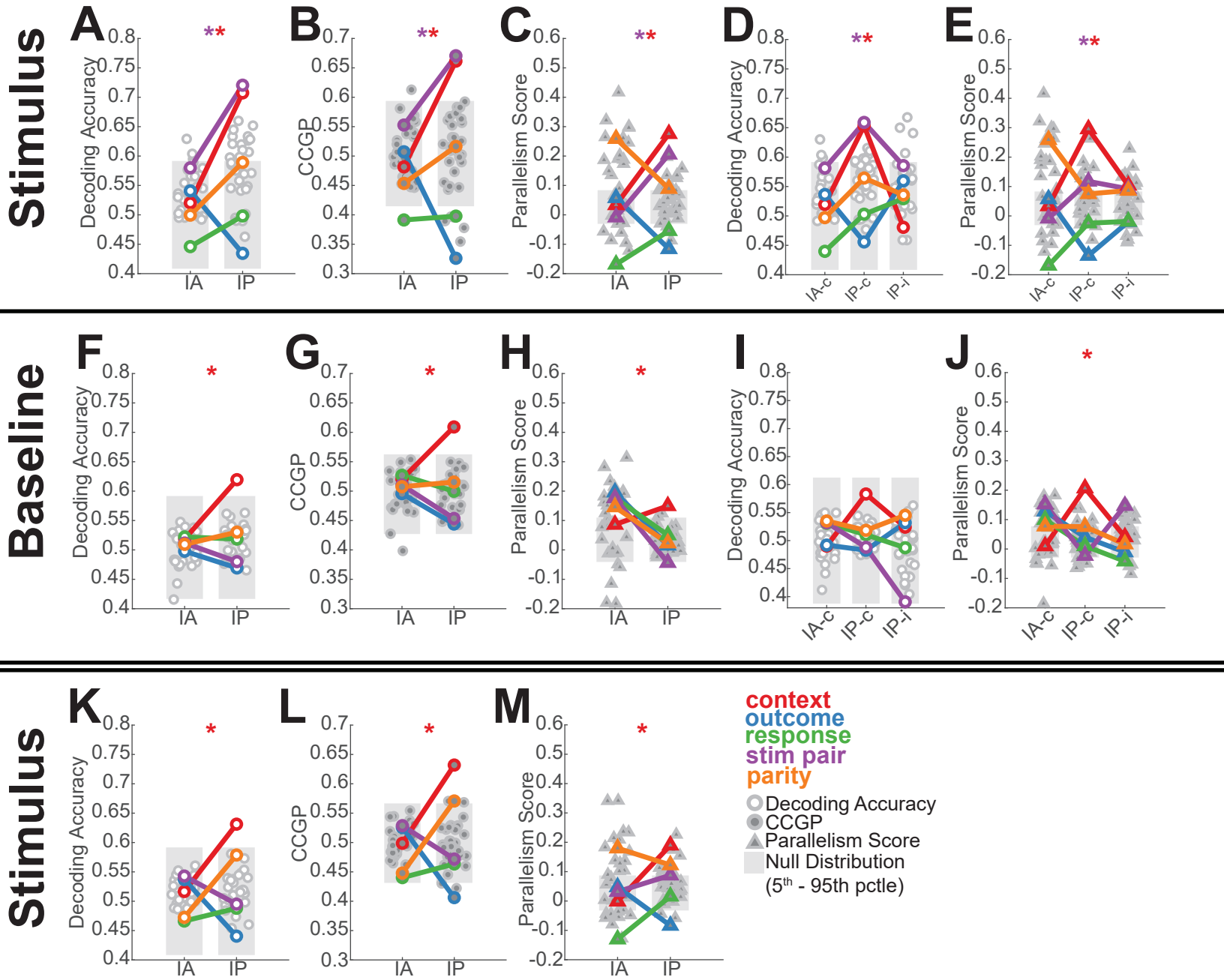

Stimulus
Baseline
Stimulus
A
B
C
D
E
F
G
H
I
J
K
L
M
Decoding Accuracy
CCGP
Parallelism Score
IA
IP
IA-c
IP-c
IP-i
context
outcome
response
stim pair
parity
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

Figure S6

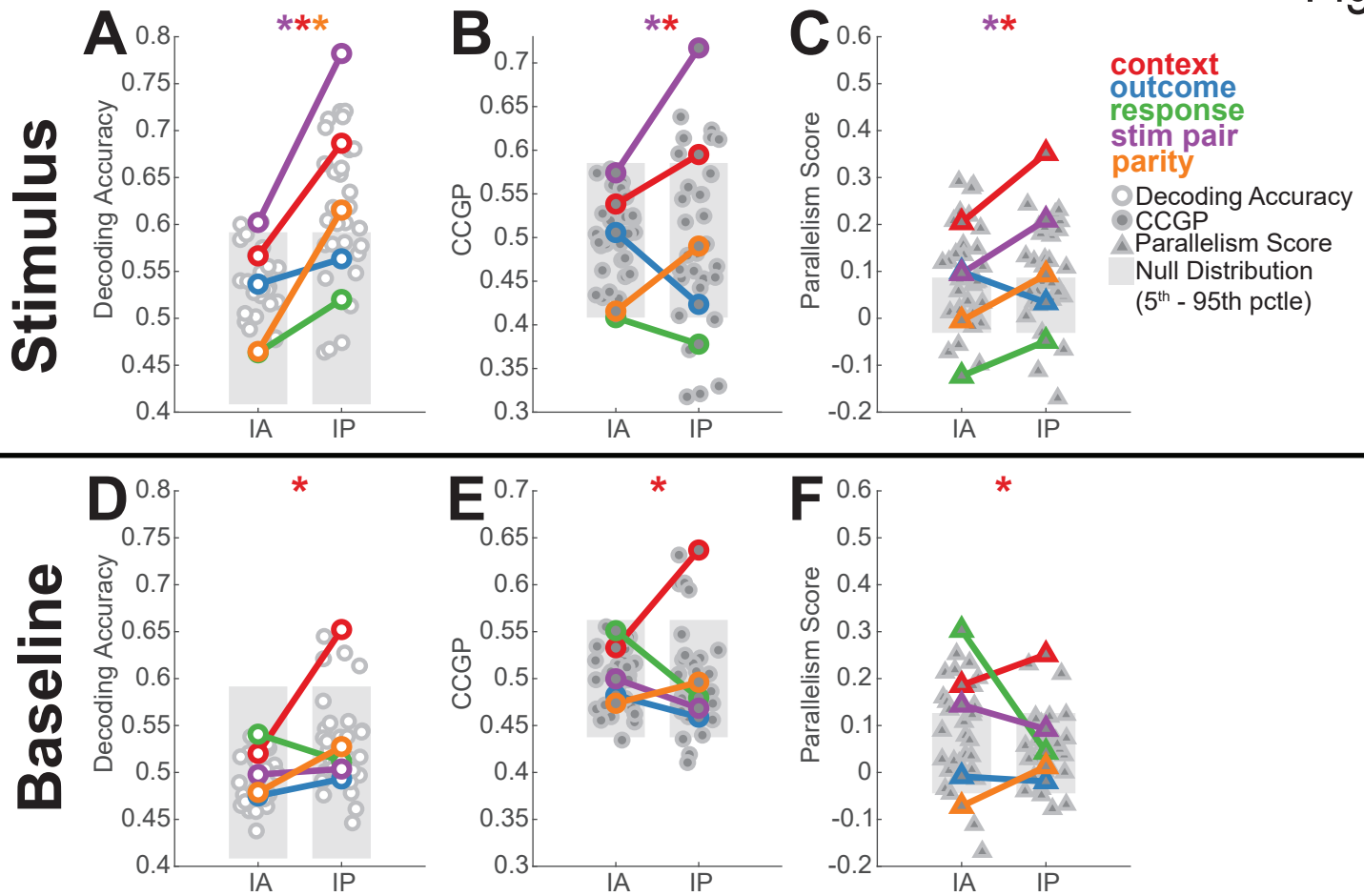

Stimulus
Baseline
A
B
C
D
E
F
Decoding Accuracy
CCGP
Parallelism Score
IA
IP
context
outcome
response
stim pair
parity
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

Figure S7

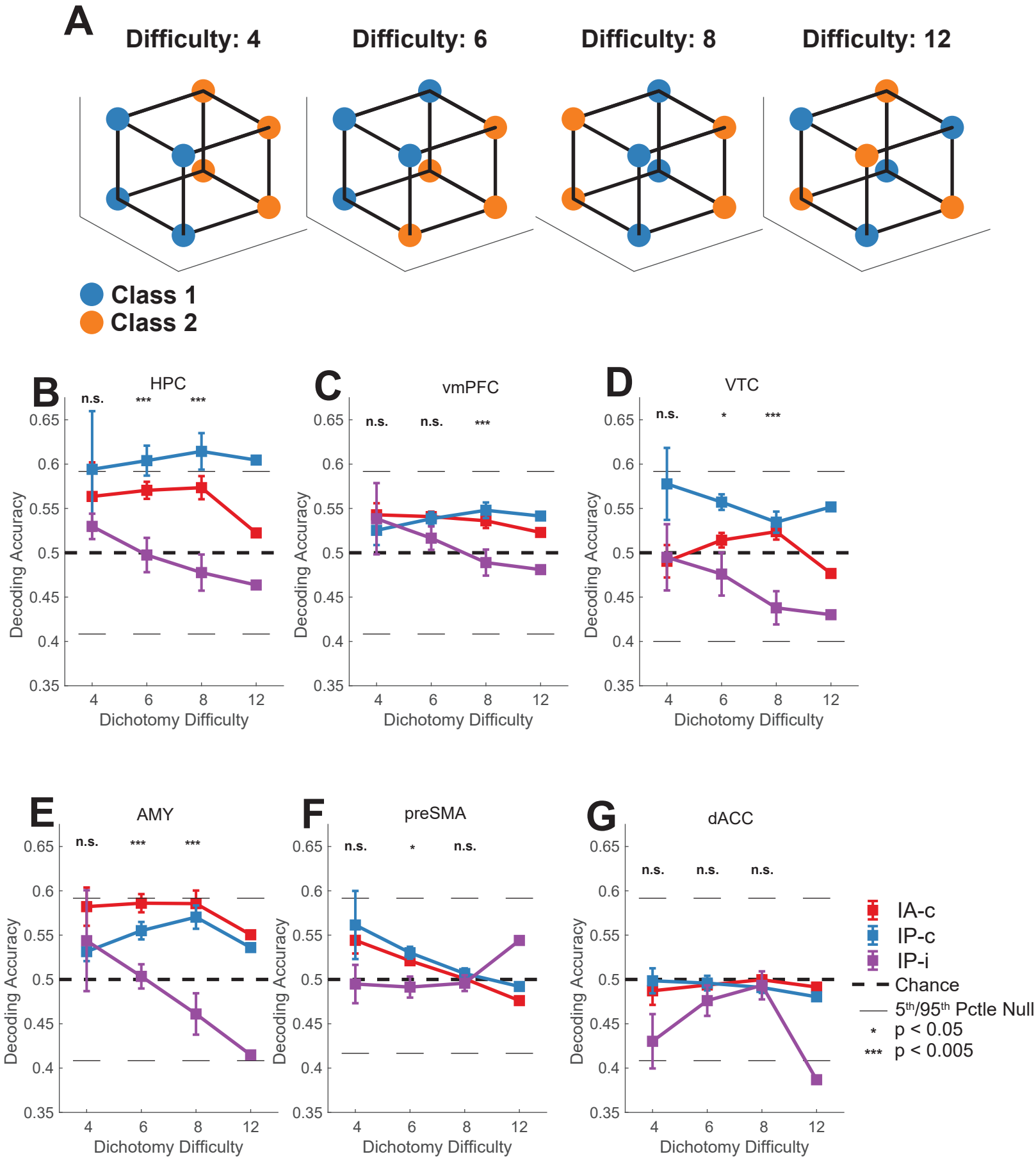
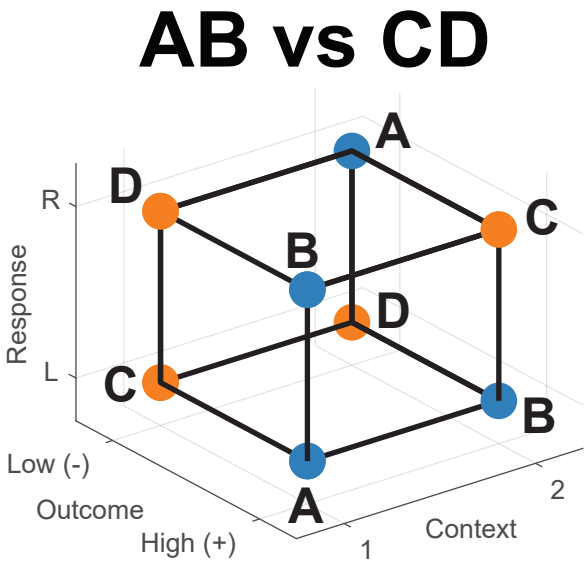
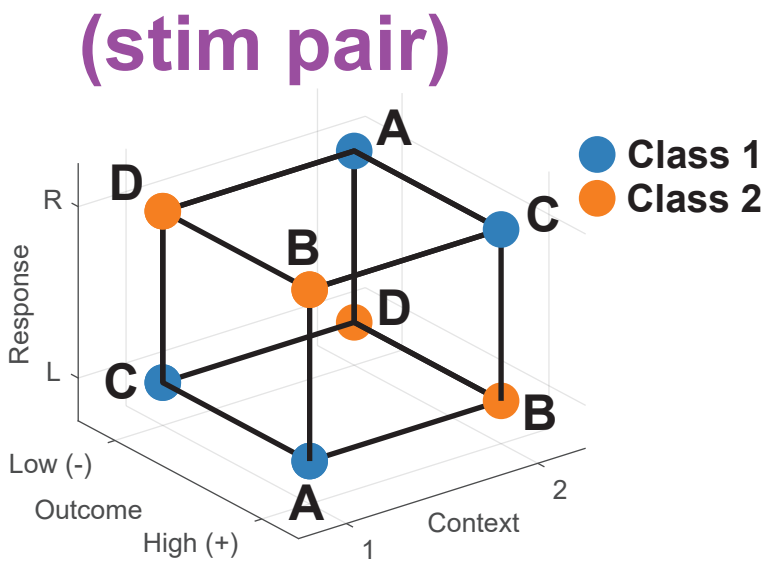
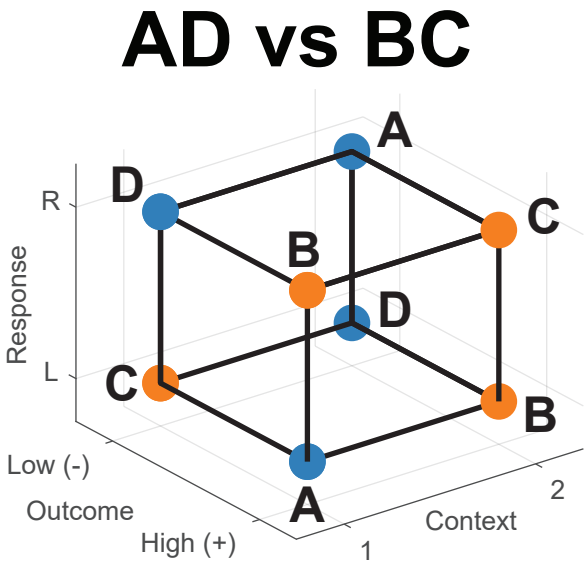
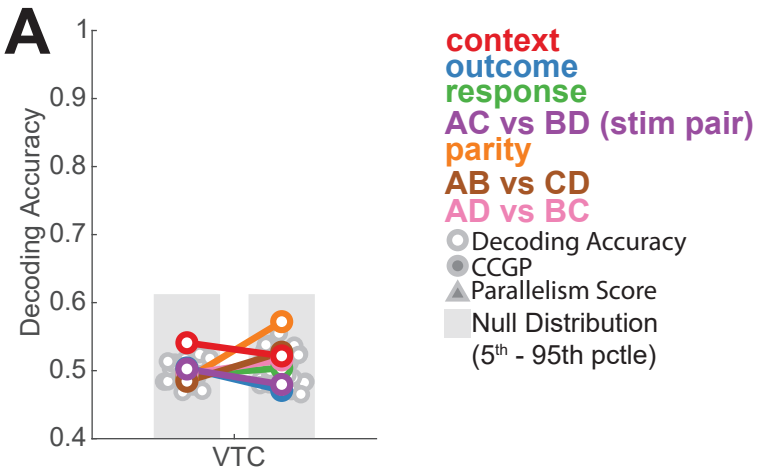

A
Difficulty: 4
Difficulty: 6
Difficulty: 8
Difficulty: 12
Class 1
Class 2
B
HPC
C
vmPFC
D
VTC
E
AMY
F
preSMA
G
dACC
Decoding Accuracy
Dichotomy Difficulty
n.s.
*
***
IA-c
IP-c
IP-i
Chance
5th/95th Pctle Null
p < 0.05
p < 0.005

Figure S8


AB vs CD
A
B
C
D
R
L
Response
Low (-)
High (+)
Outcome
1
2
Context


AC vs BD
(stim pair)
Class 1
Class 2
A
B
C
D
R
L
Response
Low (-)
High (+)
Outcome
1
2
Context


AD vs BC
A
B
C
D
R
L
Response
Low (-)
High (+)
Outcome
1
2
Context


A
Decoding Accuracy
VTC
context
outcome
response
AC vs BD (stim pair)
parity
AB vs CD
AD vs BC
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

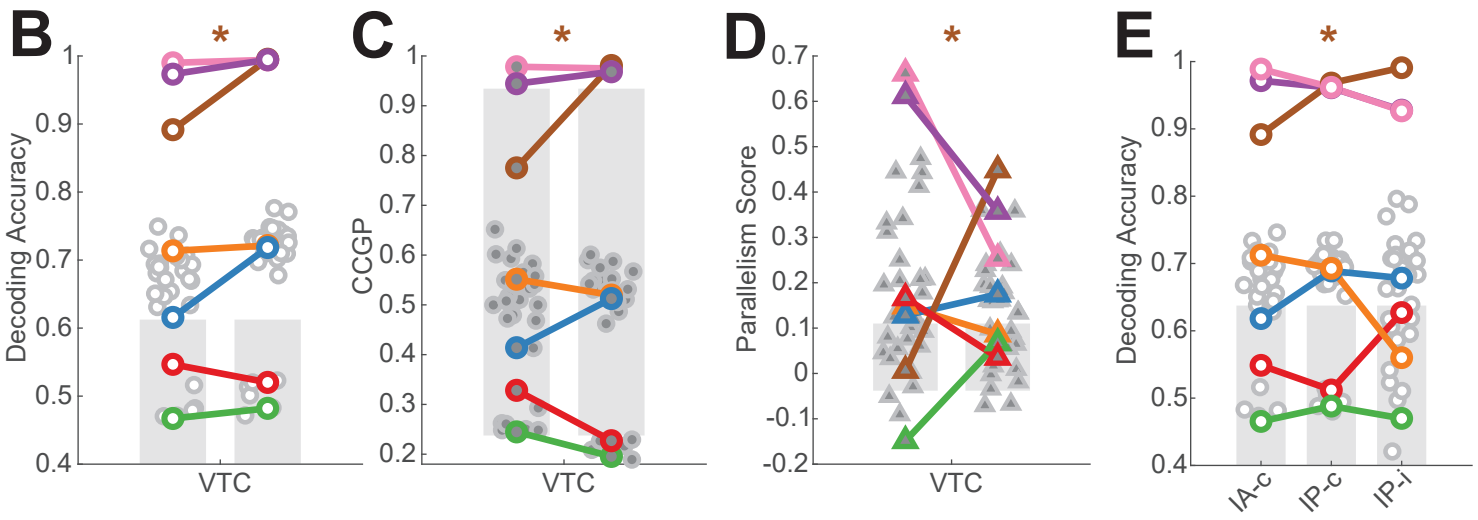
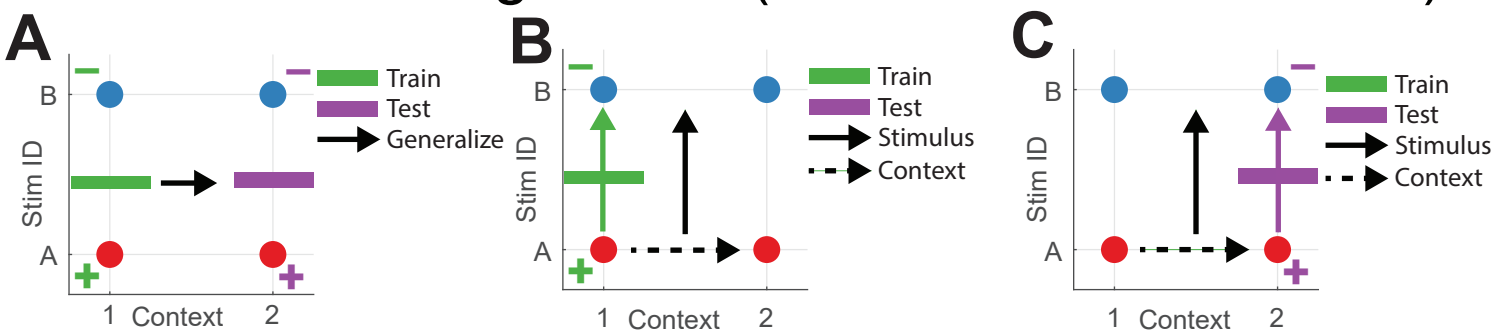

B
Decoding Accuracy
VTC
C
CCGP
VTC
D
Parallelism Score
VTC
E
Decoding Accuracy
IA-c
IP-c
IP-i

Figure S9

Stimulus Decoding/CCGP (train/test within context)



Context Decoding/CCGP (train/test within stimulus)

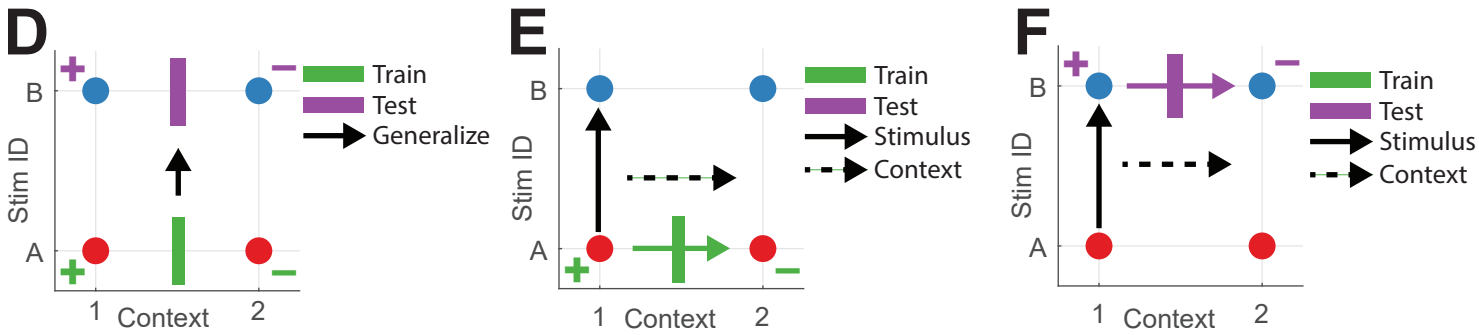
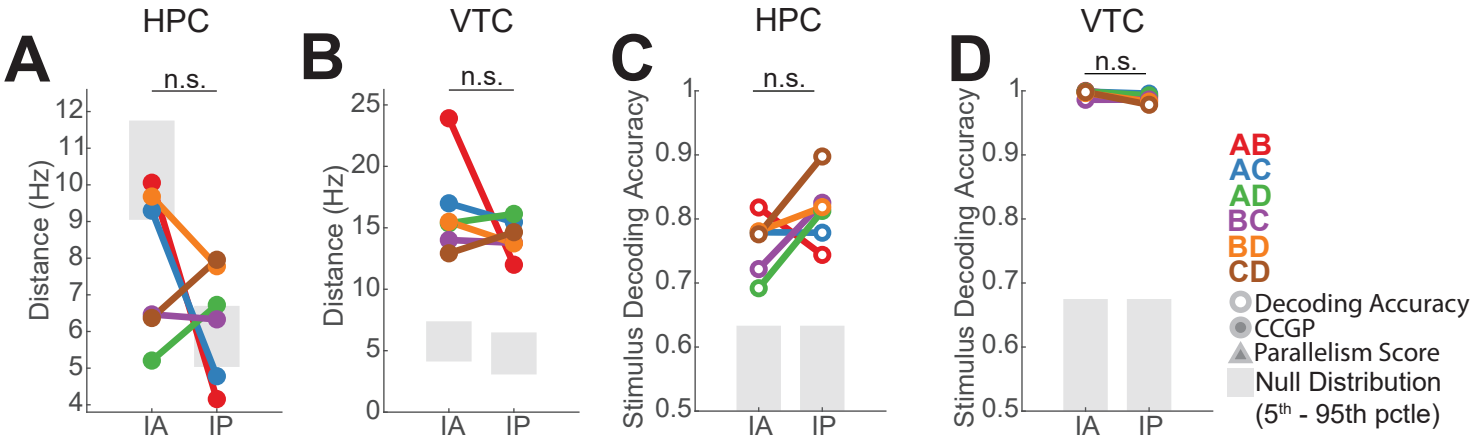


Figure S10


A
HPC
n.s.
Distance (Hz)
12
11
10
9
8
7
6
5
4
IA
IP
B
VTC
n.s.
Distance (Hz)
25
20
15
10
5
0
IA
IP
C
HPC
n.s.
Stimulus Decoding Accuracy
1
0.9
0.8
0.7
0.6
0.5
IA
IP
D
VTC
n.s.
Stimulus Decoding Accuracy
1
0.9
0.8
0.7
0.6
0.5
IA
IP
AB
AC
AD
BC
BD
CD
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)

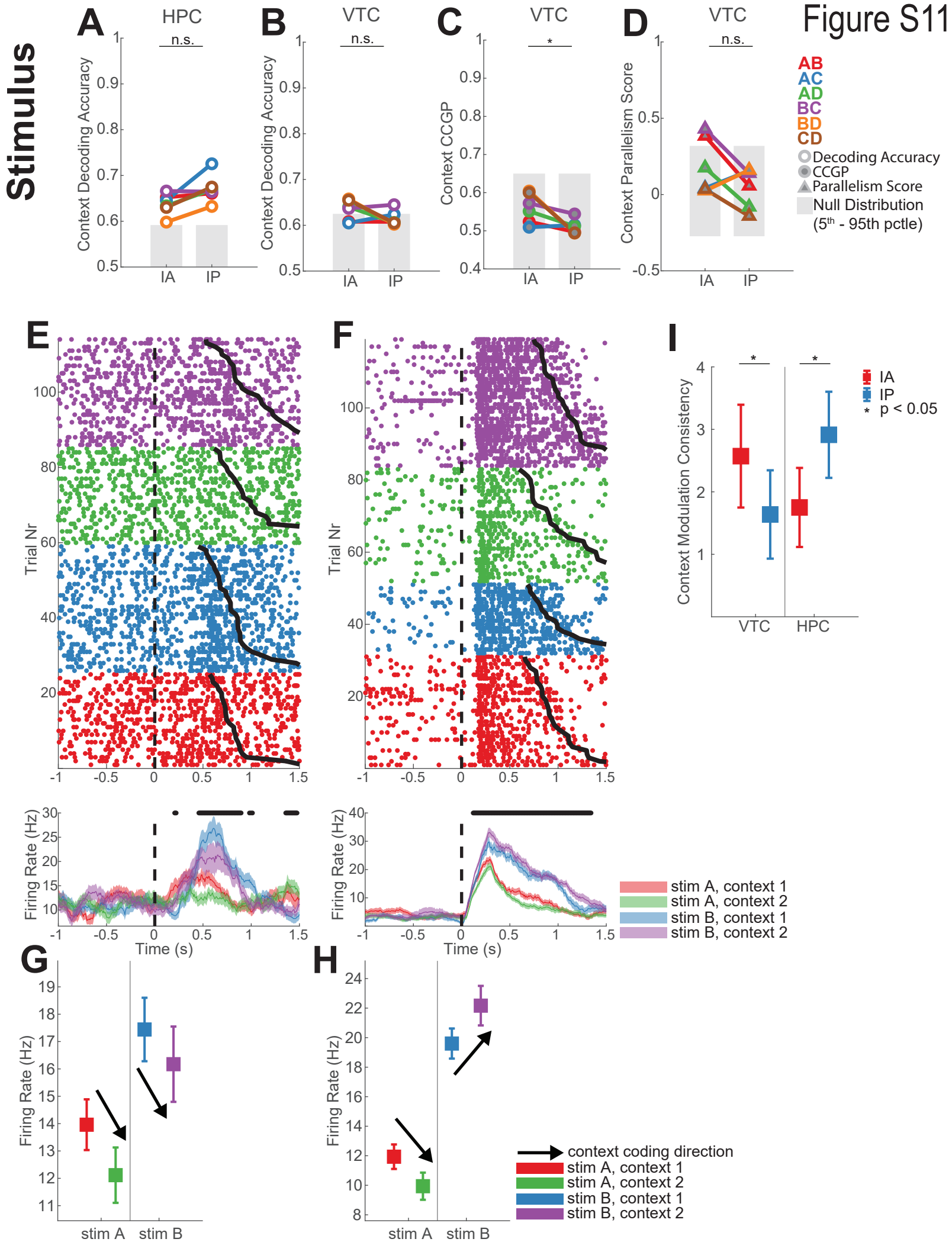


Figure S11
Stimulus
A
HPC
n.s.
Context Decoding Accuracy
IA
IP
B
VTC
n.s.
Context Decoding Accuracy
C
VTC
*
Context CCGP
D
VTC
n.s.
Context Parallelism Score
AB
AC
AD
BC
BD
CD
Decoding Accuracy
CCGP
Parallelism Score
Null Distribution
(5th - 95th pctle)
E
F
Trial Nr
Firing Rate (Hz)
Time (s)
stim A, context 1
stim A, context 2
stim B, context 1
stim B, context 2
I
Context Modulation Consistency
VTC
HPC
IA
IP
* p < 0.05
G
H
Firing Rate (Hz)
stim A
stim B
context coding direction
stim A, context 1
stim A, context 2
stim B, context 1
stim B, context 2

Figure S12

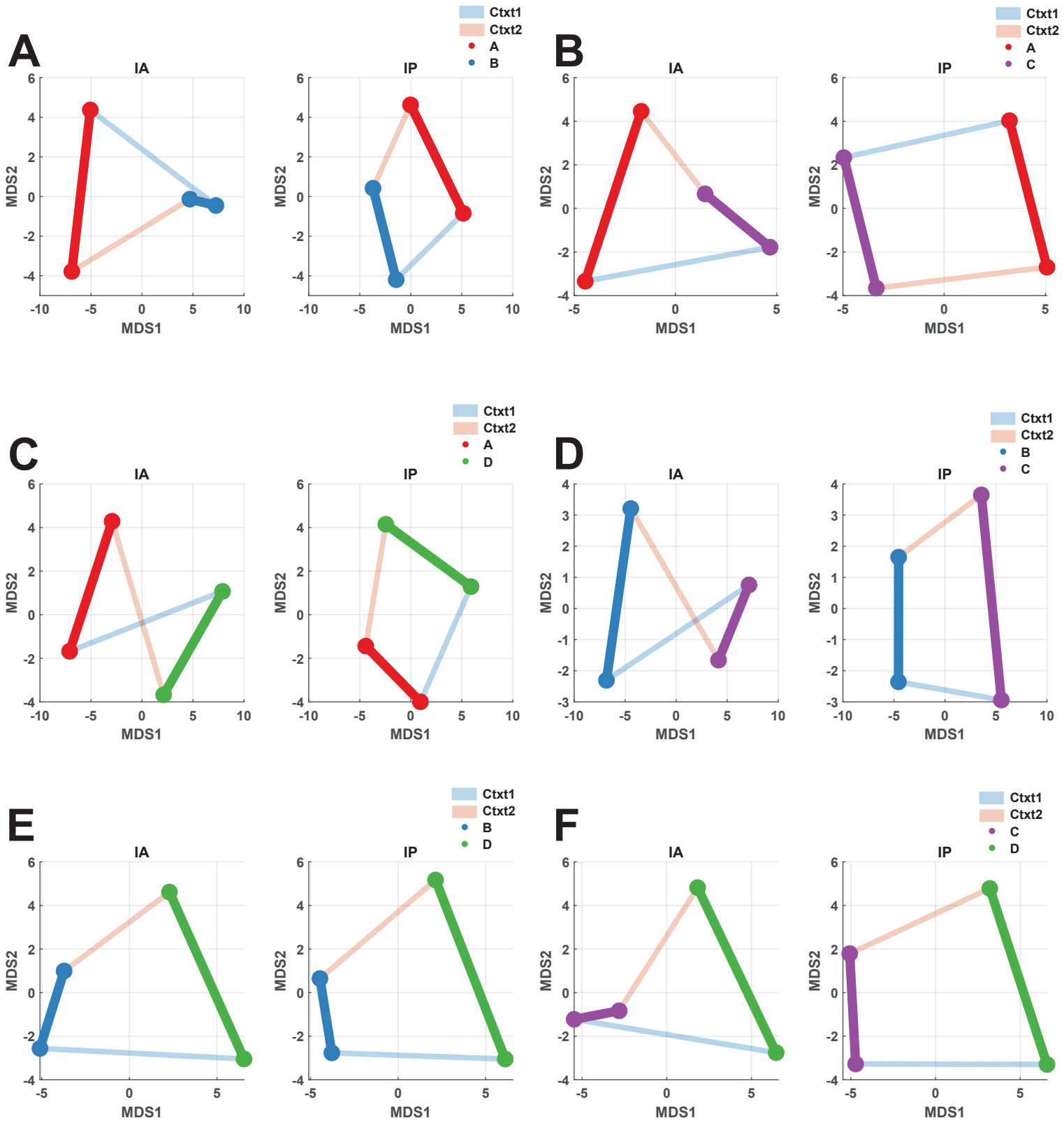

A
IA
IP
Ctxt1
Ctxt2
A
B
B
IA
IP
Ctxt1
Ctxt2
A
C
C
IA
IP
Ctxt1
Ctxt2
A
D
D
IA
IP
Ctxt1
Ctxt2
B
C
E
IA
IP
Ctxt1
Ctxt2
B
D
F
IA
IP
Ctxt1
Ctxt2
C
D
MDS1
MDS2

Figure S13

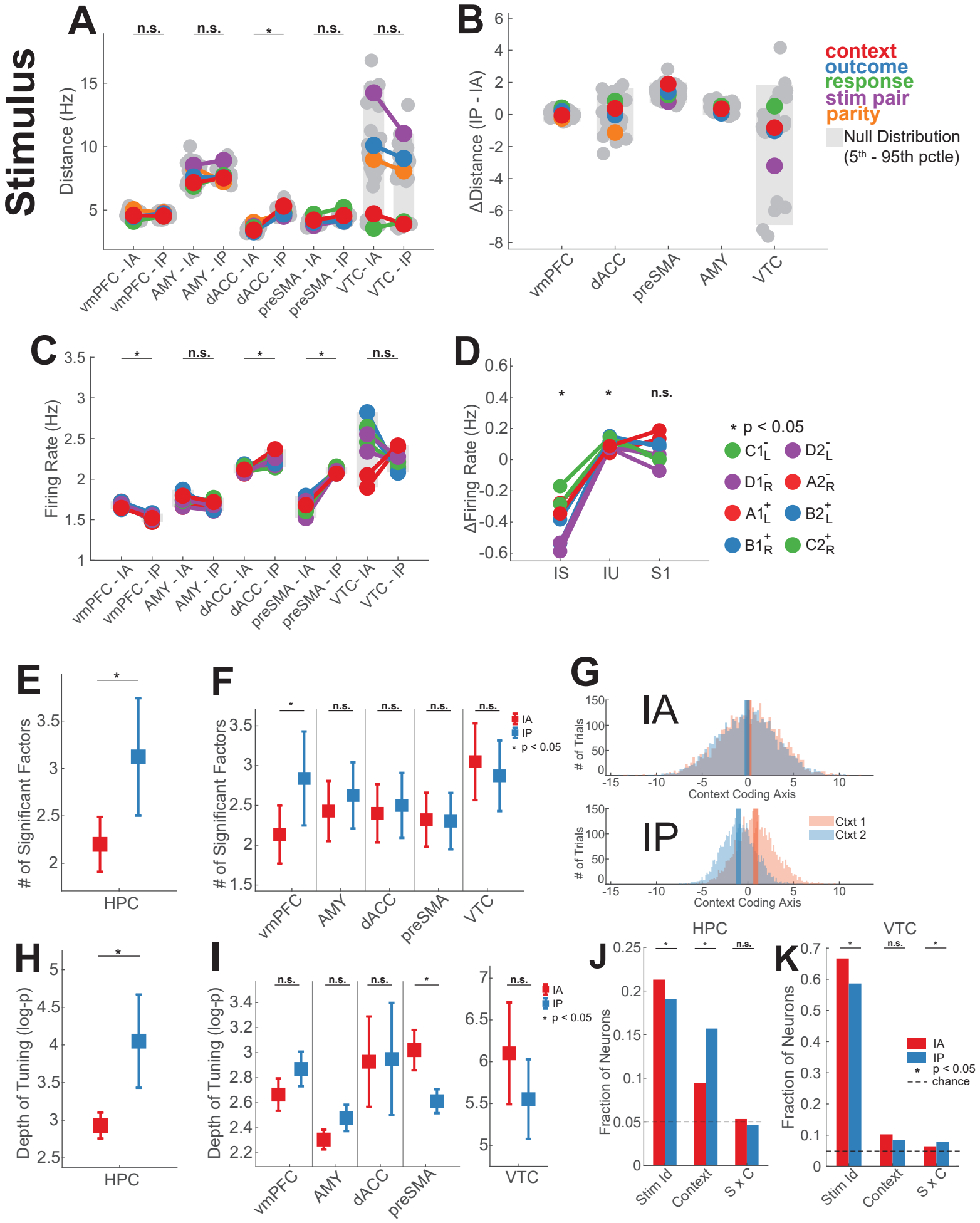

Stimulus
A
Distance (Hz)
n.s.
*
vmPFC - IA
vmPFC - IP
AMY - IA
AMY - IP
dACC - IA
dACC - IP
preSMA - IA
preSMA - IP
VTC- IA
VTC - IP
B
ΔDistance (IP - IA)
vmPFC
dACC
preSMA
AMY
VTC
context
outcome
response
stim pair
parity
Null Distribution
(5th - 95th pctle)
C
Firing Rate (Hz)
D
ΔFiring Rate (Hz)
IS
IU
S1
* p < 0.05
C1-L
D2-L
D1-R
A2-R
A1+L
B2+L
B1+R
C2+R
E
# of Significant Factors
HPC
F
vmPFC
AMY
dACC
preSMA
VTC
IA
IP
G
IA
IP
# of Trials
Context Coding Axis
Ctxt 1
Ctxt 2
H
Depth of Tuning (log-p)
I
J
HPC
Fraction of Neurons
Stim Id
Context
S x C
K
VTC
chance

Figure S14

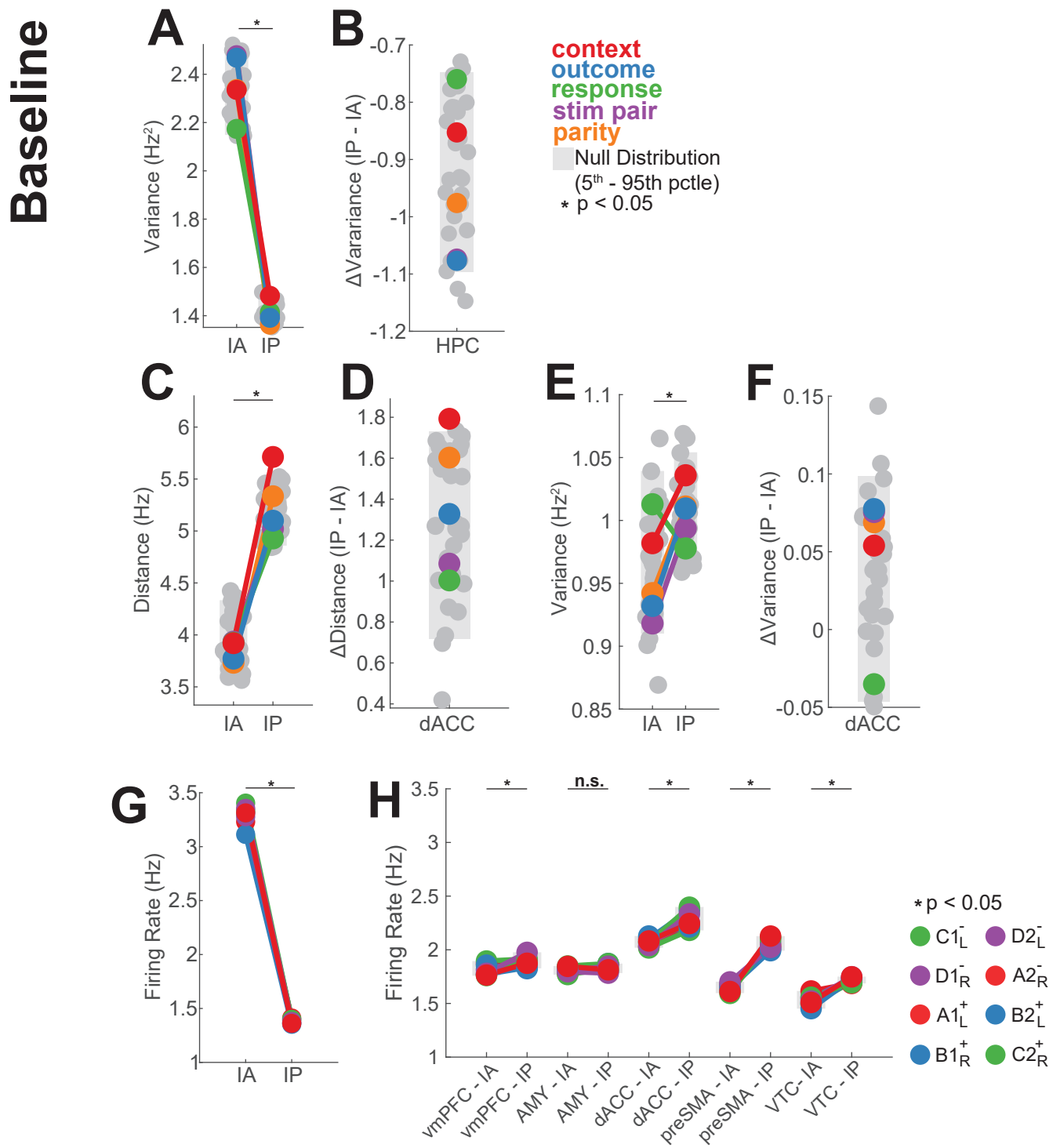
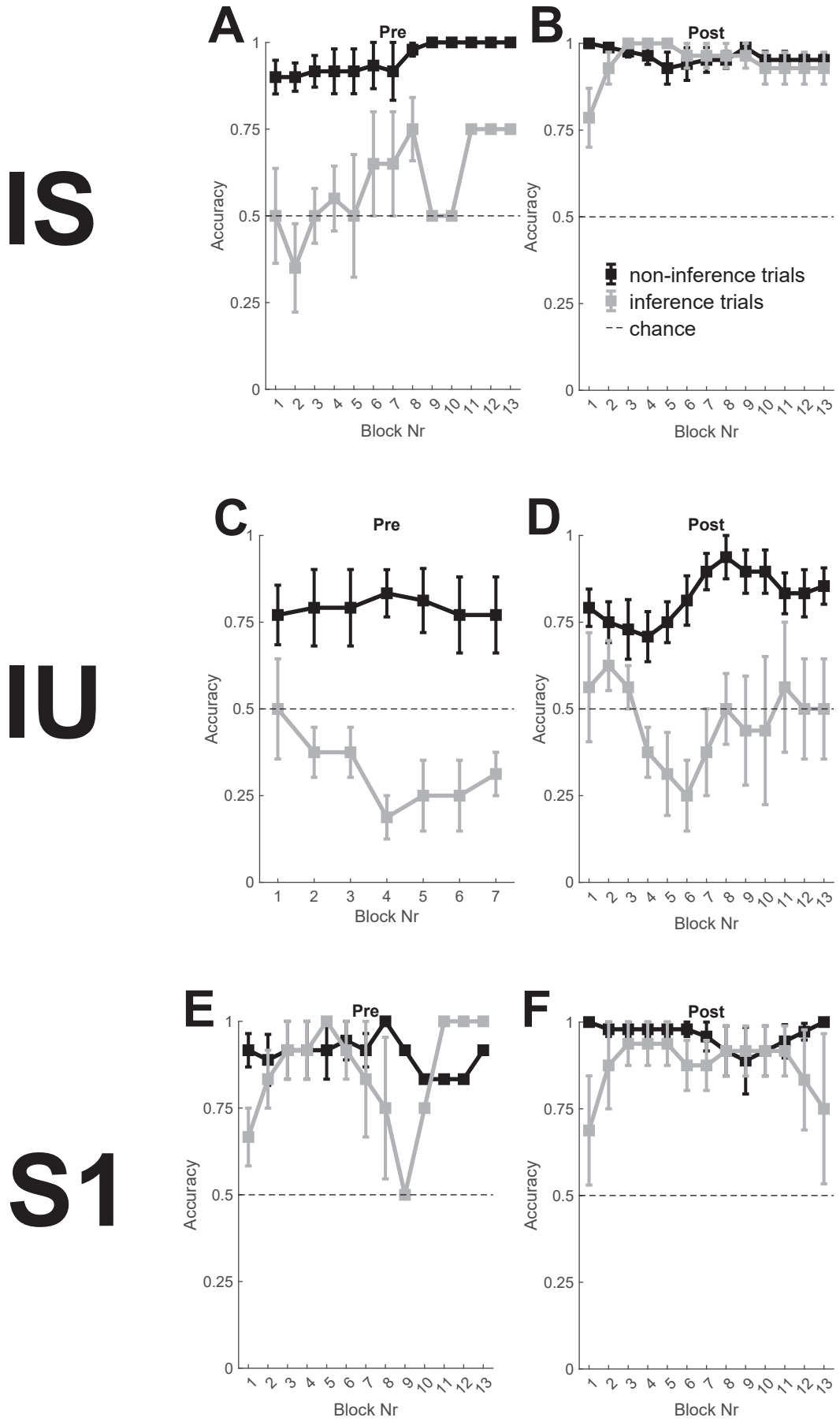

Baseline
A
B
C
D
E
F
G
H
context
outcome
response
stim pair
parity
Null Distribution
(5th - 95th pctle)
* p < 0.05
Variance (Hz²)
ΔVarariance (IP - IA)
Distance (Hz)
ΔDistance (IP - IA)
Variance (Hz²)
ΔVariance (IP - IA)
Firing Rate (Hz)
IA
IP
HPC
dACC
n.s.
vmPFC - IA
vmPFC - IP
AMY - IA
AMY - IP
dACC - IA
dACC - IP
preSMA - IA
preSMA - IP
VTC- IA
VTC - IP
* p < 0.05
C1-L
D2-L
D1-R
A2-R
A1+L
B2+L
B1+R
C2+R

Figure S15


IS
IU
S1
A
Pre
B
Post
C
Pre
D
Post
E
Pre
F
Post
Accuracy
Block Nr
non-inference trials
inference trials
chance

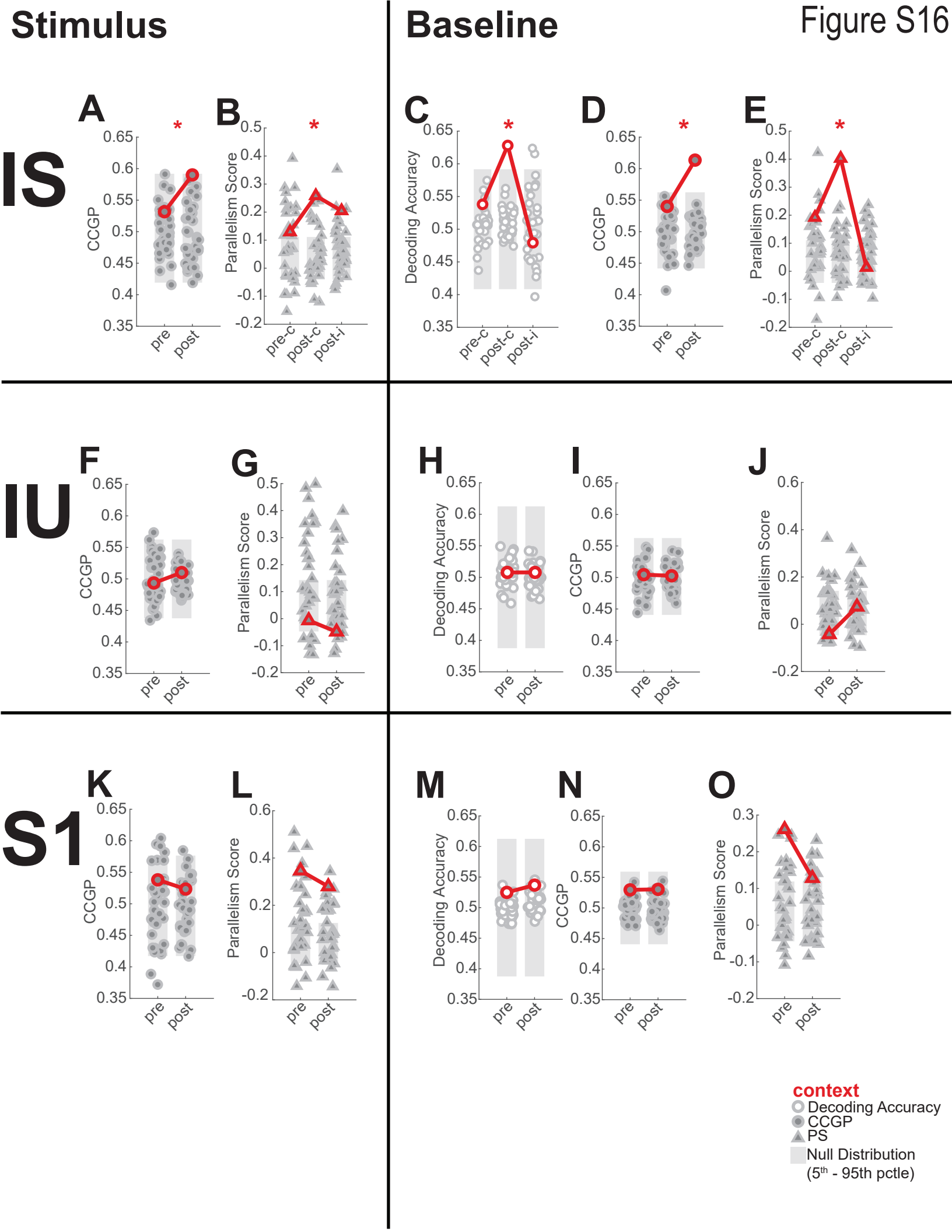

Stimulus
Baseline
Figure S16
IS
IU
S1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
CCGP
Decoding Accuracy
Parallelism Score
pre
post
pre-c
post-c
post-i
context
Decoding Accuracy
CCGP
PS
Null Distribution
(5th - 95th pctle)

Figure S1. Task behavior and single-neuron responses across all recorded regions.

(A) Task performance on individual sessions from 49 control subjects recruited through an online platform (Amazon MTurk). Accuracy is reported as an average for each subject over all non-inference trials (left) and inference trials (right). The horizontal gray dashed line corresponds to chance (50%). This task variant is equivalent to the first session of the task encountered by patients where they were given general instructions about learning stimulus-response mappings, but were not informed of the latent task structure. Subjects exhibited a variety of behaviors, with 46/49 subjects performing above chance on non-inference trials, indicating that the SRO maps were generally learnable despite the wide variation in performance on inference trials.

(B) Non-inference performance for context 1 is plotted against context 2 for each of the 36 sessions included in the analysis. Error bars correspond to SEM computed over blocks. The diagonal gray dashed line indicates identical block performance (y=x). The reported p-value is computed by paired t-test between the mean accuracies for Context 1 and Context 2 across all sessions.

(C) Same as (B), but with reaction time (RT), computed as time from stimulus onset to button press for every trial. Mean RT's are also computed by block.

(D) Truncated behavioral performance plot similar to Fig 1E. Plot shows performance on the last trial before the context switch, the first trial after the context switch, and for the first inference trial (Trial 2) averaged over all trials in each session (mean ± s.e.m. across sessions). Dashed line marks chance. Red and blue lines correspond to session performance before and after instructions detailing latent context are provided. This plot shows performance for the "instruction-successful" (IS) patient group – did not exhibit significant inference performance pre-instruction, but did exhibit inference performance post-instruction. Trial 2 pre/post performance was used to classify patients, so difference significance is not computed. All trials where pre/post performance difference was insignificant ($p > 0.05$) are shown with n.s.

(E) Same as (D), but for the "instruction-unsuccessful" (IU) patient group – did not exhibit significant inference performance during either pre-instruction or post-instruction sessions.

(F) Same as (D), but for the "session one inference" (S1) patient group – exhibited significant inference performance during both pre-instruction and post-instruction sessions.

(G) Normalized activity for all neurons recorded from the hippocampus is plotted as a heat map by region after computing the trial-averaged response to each unique condition (8 total, specified by unique Response-Context-Outcome combinations). Z-scored firing rates are computed from 0.2s to 1.2s after stimulus onset for every trial. Each row of the heat map corresponds to the activity of a single neuron, and columns correspond to each of the 8 conditions. Neurons are ordered such that adjacent rows (neurons) are maximally correlated in 8-dimensional condition response space. This approach would allow for modular tuning to visibly emerge in the heat map if groups of neurons were clustered in their response profiles. Clearly, the responses here are very diverse.

(H) Same as (G), but for amygdala.

(I) Same as (G), but for ventral temporal cortex.

(J) Same as (G), but for dorsal anterior cingulate cortex.

(K) Same as (G), but for pre-supplementary motor area.

(L) Same as (G), but for ventromedial prefrontal cortex.

(M) Percentage of neurons across all areas that exhibit tuning to each of the three binary variables manipulated in the experiment. Tuning was assessed by fitting either a 2x2x2 (Response-Context-Outcome) ANOVA for every individual neuron's firing rate during a 1s window during the stimulus presentation period. Significant neurons were counted as $p < 0.05$ for main effects (Linear) or interaction effects (Nonlinear) involving the stated variables (RxC = Response x Context interaction, RxO = Response x Outcome interaction, CxO = Context x Outcome interaction, RxCxO = Response x Context x Outcome interaction). Significance in the change of percentage of neurons exhibiting modulation to each factor is determined via z-test, where "*" indicates $p < 0.05$, "***" indicates $p < 0.005$, and "n.s." indicates "not significant".

(N) Same analysis as **(M)**, but for a 4x2 ANOVA for stimulus identity and context.

(O) Same analysis as **(M)**, but for a 4x2 ANOVA for stimulus identity and response.

Note: the corresponding analysis between stimulus identity and outcome cannot be conducted since those variables are correlated by task construction.

Figure S2. Visual representation of all named balanced dichotomies.

Named balanced dichotomies correspond to condition splits that have clearly interpretable meaning with respect to the construction of the task when evaluating the decodability or disentanglement (Cross Condition Generalization Performance, Parallelism Score) for that dichotomy. For example, the context dichotomy (top left), arises from assigning all conditions for context = 1 to one class and all conditions for which context = 2 to the other class. Performing binary classification on neural responses with trial labels arranged in this way corresponds to decoding context from the neural population. The specific assignment of class labels 1 and 2 is arbitrary, and inverting the labels still corresponds to the same meaning for the dichotomy. All named dichotomies shown here are color coded to reflect their value in all Shattering Dimensionality, CCGP, and PS plots, and this color code remains consistent throughout the paper whenever balanced dichotomies are considered.

The "stim pair" dichotomy corresponds to the special split of stimulus identities where the stimuli that have the same response in each context are grouped together (e.g. Stimuli A and C have Response L in Context 1 and Response R in Context 2, v.v. for Stimuli B and D). There are more balanced dichotomies that correspond to splits of stimulus identity (shown in Fig. S7). High decodability of any one of these balanced dichotomies reflects stimulus-id coding in the neural population.

The "parity" dichotomy is another special dichotomy that corresponds to the most difficult, or non-linear, dichotomy that can be constructed, and high decodability of this dichotomy is a signature of a high-dimensional representation. The term parity is in reference to the fact that, if task states were represented as 3-bit binary words where each bit corresponds to the value of the response, context, and outcome variable that describes the state (e.g. 000 for Left, Ctxt 1, Low, 111 for Right, Ctxt 2, High), then one class of the parity dichotomy corresponds to all states with an even number of ones, and the other class corresponds to all states with an odd number of ones. Note, for this dichotomy, no node of a given class shares an edge with another node of the same class. If one views the faces of the cube, one can see that the standard 2D XOR dichotomy between class 1 and 2 is present on every face. The notion of parity, can be generalized to arbitrarily many dimensions.

Figure S3. Additional geometric analysis for hippocampus during stimulus processing.

(A) Cross-condition generalization performance (CCGP) reported for other regions over balanced dichotomies. Each dot corresponds to the CCGP for a single dichotomy. The reported values are averages over 1000 repetitions of resampling of trials and neurons as described in the methods. For every region, the left column corresponds to inference absent (IA) sessions and the right column to inference present (IP sessions). Colored lines are drawn to connect IA and IP values for named dichotomies using the standard color-coding scheme. The gray background bars indicate the 5th (bottom) to 95th (top) percentile of the null distribution. Note that the null distribution differs by area due to the different number of neurons present in each area. Significant named dichotomies are marked when the dichotomies are: above 95th pctle of null in IP (i.e. significantly above chance during inference), significantly different between IA and IP (RankSum $p < 0.01/35$, Bonferroni corrected for balanced dichotomies). Significant increases from IA to IP were observed in vmPFC for stim pair (purple, $p_{IA} = 0.45, p_{IP} = 0.014$) and preSMA for response (green, $p_{IA} = 0.045, p_{IP} = 0.0010$). Note that stim pair CCGP in AMY was above chance for both IA and IP sessions (purple, $p_{IA} = 0.050, p_{IP} = 0.039$).

(B) Same as (A), but for Parallelism Score (PS). Significant increases in PS were present for stim pair in amygdala (purple, $p_{IA} = 1.3x10^{-4}, p_{IP} = 9.0x10^{-8}$) and context in the dorsal anterior cingulate (red, $p_{IA} = 0.99, p_{IP} = 3.8x10^{-12}$).

(C) Change in decoding accuracy computed as IP – IA for every balanced dichotomy. The gray shaded bar again indicates 5th-95th pctle of null, which is populated by computing IP – IA for 1000 random pairs of dichotomies in the null distributions of IP and IA computed separately. In general, the null distribution for any "IP-IA" difference plot is computed by drawing such samples from the IA and IP null distributions of the associated area and metric (e.g. the null distribution here is computed using the IP and IA nulls in Fig. 2C, the null distribution for Fig. S3D is computed from Fig. 2D, the null distribution for Fig. S3F is computed from Fig. S3E, etc…).

(D) Same as (C), but for changes in CCGP from IA to IP.

(E) Parallelism score plot for hippocampus in the IA and IP condition. Coloring, plotting, and significance conventions are identical to those used for Decoding Accuracy and CCGP plots (e.g. Fig. 2C,D). Null distribution here (gray background bars) is computed using the geometric null, the same procedure as CCGP. We note that the 5th and 95th pctle of null is quite narrow since the rotation procedure used to generate the null distribution produces approximately orthogonal coding vectors in high dimensional spaces. Here, context was significantly elevated in IP compared to IA (red, $p_{IA} = 0.55, p_{IP} = 1.4x10^{-1}$), as was stim pair (purple, $p_{IA} = 0.17, p_{IP} = 1.7x10^{-8}$).

(F) Same as (C), but for changes in PS from IA to IP.

Figure S4. Additional geometric analysis for hippocampus and anterior cingulate cortex during the baseline period.

(A) Decoding accuracy reported for regions other than hippocampus, analogous to Fig. 2E, but for the baseline period instead of the stimulus period. Additionally, trial labels here correspond to the conditions (Response, Context, Outcome) encountered during the previous trial. Details regarding the reported decoding accuracies and null distributions are identical to those described in Fig. 2A. Decoding accuracy for all balanced dichotomies is reported for the inference absent (IA, Left) and inference present (IP, right) conditions, with color-coded lines for named dichotomies connecting IA and IP for each region. Significant increase from IA to IP was observed in dACC for context (red, $p_{IA} = 0.37, p_{IP} = 0.049$).

(B) Change in decoding accuracy for all balanced dichotomies in hippocampus associated with the presence of inference (i.e. IP – IA), again computed during the baseline period. Procedure for all reported accuracies and null distribution construction identical to that described in Fig. S3C, except that the analysis used baseline firing rates and condition labels from the previous trial instead of the current trial. Here, context is the only balanced dichotomy whose increase in decodability is significantly above null.

(C) Change in cross-condition generalization performance (CCGP) for all balanced dichotomies in hippocampus. See Fig. S3C for plotting details. Context is also the only dichotomy for which the CCGP rises significantly more than the 95th pctle of the geometric null distribution.

(D) Parallelism score (PS) for hippocampus in the IA and IP condition. Plot is analogous to Fig. S3E, but computed during the baseline with previous trial labels instead of during the stimulus with current trial labels. Context is the only named dichotomy for which the PS increased to significance above the geometric null in the IP condition. (red, $p_{IA} = 0.37, p_{IP} = 1.2x10^{-10}$)

(E) Same as **(B)** and **(C)**, but for parallelism score. Context is the only named dichotomy to increase significantly in IP, and the other un-named dichotomies that also significantly rise are correlated with context.

(F) CCGP for balanced dichotomies in the dorsal anterior cingulate cortex (dACC). Associated decoding accuracy for balanced dichotomies in IA and IP shown in (A). Here in the dACC, context (red, $p_{IA} = 0.26, p_{IP} = 0.018$) is also found to be in an abstract format.

(G) Parallelism score (PS) for balanced dichotomies in the acc. See Fig. S3E for plotting details. Here, context (red, $p_{IA} = 0.18, p_{IP} = 0.013$)) emerges as significant in the IP condition.

(H) Change in decoding accuracy for balanced dichotomies in acc with inference (IP – IA). The associated plot for IA and IP is in (A). Though context is the dichotomy with the greatest increase in decoding accuracy, it is still below the null 95th pctle in this case.

(I) Same as (H), but for CCGP. Here, context is also the dichotomy with the greatest increase, but is still below null 95th pctle.

(J) Same as (H), but for PS. Increase in PS for parity is notably significant ($p_{\Delta} = 0.0016$). Context also significantly increases ($p_{\Delta} = 0.026$).

(K) Scatter plot of context decoder weights during baseline and stimulus periods. Each point in the scatter plot corresponds to the weights assigned to a single neuron. Weights are averages over 1000 repetitions of decoding with trial re-sampling as described in the methods. Weights are z-scored across neurons for every repetition. Circled neurons are those with weights in the top 25th pctle, separately identified for

baseline and stimulus decoding, with neurons in the top 25th pctle for baseline context decoding alone circled red, top 25th pctle for stimulus context decoding alone circled blue, and those in the top 25th pctle of both stimulus and baseline context decoding circled green.

(L) Angle distribution for circled neurons shown in (K). Angles are computed as the angle generated by the vector $<\beta_{baseline}, \beta_{stimulus}>$ with the vector $<1,0>$ for each neuron. Thus, angle 0 corresponds to a neuron with $\beta_{stimulus} = 0$ and $\beta_{baseline} > 0$, angle 90 corresponds to a neuron with $\beta_{stimulus} > 0$ and $\beta_{baseline} = 0$, and a neuron with an intermediate angle contributes to both.

Figure S5. Hippocampal representation geometry unchanged by removal of linearly tuned neurons and neurons recorded inside seizure onset zones.

Identical analysis to the main geometric analysis shown in Fig. 2, except that hippocampal neurons are excluded from the analysis with the following criteria: in **(A-J)**, neurons with significant linear tuning for Context, Response, or Outcome (2x2x2 ANOVA, Any Main Effect $p < 0.01$), and in **(K-M)**, neurons with significant linear tuning for Stimulus Identity or Context (4x2 ANOVA, Any Main Effect $p < 0.01$).

Using the 3-Way ANOVA applied neuron-by-neuron, 455/494 neurons were retained for the stimulus period analysis **(A-C)** and 458/494 neurons were retained for the baseline period analysis **(D-F)**. All primary results for changes in hippocampal geometry were recapitulated apart from decodability of the parity dichotomy during the stimulus period **(A)**.

(A) Context decodability (red, $p_{IA} = 0.36, p_{IP} = 0.0001, p_{RS} = 1.6x10^{-31}$). Stim pair decodability (purple, $p_{IA} = 0.078, p_{IP} = 4.2x10^{-5}, p_{RS} = 6.6x10^{-31}$) during the stimulus presentation.

(B) Context CCGP (red, $p_{IA} = 0.63, p_{IP} = 0.0016, p_{RS} = 5.2x10^{-3}$). Stim pair CCGP (purple, $p_{IA} = 0.17, p_{IP} = 0.00095, p_{RS} = 5.3x10^{-34}$) during the stimulus presentation.

(C) Context PS (red, $p_{IA} = 0.40, p_{IP} = 3.7x10^{-13}$). Stim pair PS (purple, $p_{IA} = 0.83, p_{IP} = 1.2x10^{-7}$) during the stimulus presentation.

(D) Context decodability (red, $p_{IA-c} = 0.36, p_{IP-c} = 0.0029, p_{IP-i} = 0.64, p_{RS} = 1.5x10^{-2}$). Stim pair decodability (purple, $p_{IA-c} = 0.071, p_{IP-c} = 0.0021, p_{IP-i} = 0.062, p_{RS} = 2.0x10^{-5}$) during the stimulus presentation.

(E) Context PS (red, $p_{IA-c} = 0.40, p_{IP-c} = 4.6x10^{-15}, p_{IP-i} = 0.012$) during the stimulus presentation.

(F) Context decodability (red, $p_{IA} = 0.37, p_{IP} = 0.013, p_{RS} = 2.2x10^{-26}$) during the baseline.

(G) Context CCGP (red, $p_{IA} = 0.31, p_{IP} = 0.0044, p_{RS} = 1.9x10^{-3}$) during the baseline.

(H) Context PS (red, $p_{IA} = 0.12, p_{IP}$ =0.0055) during the baseline.

(I) Context decodability (red, $p_{IA-c} = 0.55, p_{IP-c} = 0.12, p_{IP-i} = 0.37$) during the baseline.

(J) Context PS (red, $p_{IA-c} = 0.66, p_{IP-c} = 8.5x10^{-9}, p_{IP-i} = 0.30$) during the baseline.

Using the 2-Way ANOVA applied neuron-by-neuron, 412/494 neurons were retained for the stimulus period analysis **(A-C)**. The stim-pair dichotomy is no longer decodable after removal of all stimulus-identity tuned neurons, but context is still present in an abstract format.

(K) Context decodability (red, $p_{IA} = 0.38, p_{IP} = 0.0088, p_{RS} = 4.1x10^{-28}$) during the stimulus presentation.

(L) Context CCGP (red, $p_{IA} = 0.51, p_{IP} = 6.0x10^{-4}, p_{RS} = 2.5x10^{-34}$) during the stimulus presentation.

(M) Context PS (red, $p_{IA} = 0.77, p_{IP} = 2.3x10^{-6}$) during the stimulus presentation.

Figure S6. Hippocampal representation geometry unchanged by removal of neurons recorded inside seizure onset zones.

Identical analysis to the main geometric analysis shown in Fig. 2, except that hippocampal neurons neurons recorded in seizure onset zones (SOZs, post-hoc identification) were removed. 410/494 neurons were retained for analysis. The exclusion neurons recorded from SOZ hippocampi led to the full hippocampal geometric analysis being effectively identical to that reported in Fig. 2, with every significant named dichotomy increase during stimulus **(A-C)** and baseline **(D-F)** periods being recapitulated in the absence of SOZ hippocampal neurons.

(A) Context decodability (red, $p_{IA} = 0.12, p_{IP} = 0.00044, p_{RS} = 1.0x10^{-26}$). Stim pair decodability (purple, $p_{IA} = 0.034, p_{IP} = 2.0x10^{-7}, p_{RS} = 3.5x10^{-32}$). Parity decodability (purple, $p_{IA} = 0.74, p_{IP} = 0.019, p_{RS} = 2.4x10^{-30}$).

(B) Context CCGP (red, $p_{IA} = 0.74, p_{IP} = 0.019, p_{RS} = 1.1x10^{-3}$). Stim pair CCGP (purple, $p_{IA} = 0.084, p_{IP} = 2.7x10^{-5}, p_{RS} = 1.2x10^{-33}$).

(C) Context PS (red, $p_{IA} = 3.5x10^{-7}, p_{IP} = 0$). Stim pair PS (purple, $p_{IA} = 0.027, p_{IP} = 1.6x10^{-7}$)

(D) Context decodability (red, $p_{IA} = 0.35, p_{IP} = 0.0025, p_{RS} = 1.1x10^{-30}$).

(E) Context CCGP (red, $p_{IA} = 0.20, p_{IP} = 0.00018, p_{RS} = 2.5x10^{-34}$).

(F) Context PS (red, $p_{IA} = 0.0022, p_{IP} = 2.0x10^{-5}$).

Figure S7. Effect of inference and errors on shattering dimensionality as a function of dichotomy difficulty.

The signature for a high-dimensional representation is a greater degree of non-linear mixing of task variables. "Dichotomy difficulty" is a systematic measure that quantifies the relative amount of non-linear interaction of task variables needed in a population of neurons to make a given dichotomy decodable (see methods for detailed description). **(A)** Example schematics of dichotomies of increasing difficulty. The cubes here represent different unique task conditions realized by three binary variables, and node coloring represents membership of a condition to one of two arbitrary classes assigned for the purposes of dichotomy decoding (identical to Fig. S2). Note: the difficulty 4 dichotomy corresponds to context and difficulty 12 dichotomy corresponds to parity (Fig. S2). **(B-G)** Decoding accuracy as a function of dichotomy difficulty for different regions. Reported values (mean +/- SEM) are computed over dichotomy decoding accuracies, where the average decoding accuracy for each dichotomy is computed with 1000 repetitions of re-sampled estimation (see methods). The blue, red, and green curves correspond to correct IA trials, correct IP trials, and incorrect IP trials respectively. Black dashed lines indicate chance level (50% for binary decoding), horizontal black lines indicate the 5th and 95th pctle of the null distribution. P-values are computed by conducting a one-way ANOVA over dichotomies independently for every dichotomy difficulty (Bonferroni MCC). This value is not meaningfully computable for difficulty 12, which contains a single dichotomy (the parity dichotomy), and is therefore not reported. Hippocampus alone **(B)** exhibits an increase in decoding accuracy from IA to IP sessions, with more difficulty dichotomies rising above 95th pctle null in IP. A collapse in representational dimensionality on incorrect trials (purple curves) is present in the hippocampus **(B)**, and is also present in other areas, most prominently in the ventral temporal cortex **(D)** and the amygdala **(E)**.

Figure S8. Balanced dichotomy analysis of neural geometry in ventral temporal cortex.

Ventral temporal cortex (VTC) strongly encodes high-level features of visual stimuli, necessitating the introduction of two new dichotomies that capture stimulus identity, while not directly corresponding to any of the principal manipulated variables in the task (Response, Context, Outcome). The AB vs CD and AD vs BC dichotomies are "stimulus" dichotomies in that they represent systematic differences in coding between unrelated stimuli arbitrarily paired together, unlike the AC vs BD dichotomy which pairs stimulus identities for which responses are identical across the two contexts. That is, A/C response is L in Context 1 and R in Context 2, and v.v. B/D response is R in Context 1 and L in Context 2. These are the images whose correct responses "switch together" across contexts. Note that the new stimulus dichotomies are correlated with other named dichotomies: AB vs CD is correlated with outcome and AD vs BC is correlated with the parity dichotomy. Thus, high AB vs CD decodability will lead to increased outcome decodability. However, CCGP is robust to these dichotomy correlations, and will be low for correlated dichotomies even if decodability is increased. These three dichotomies (AB vs CD, AC vs BD, and AD vs BC) are particularly relevant to vtc given its strong stimulus representations.

(A) Dichotomy decodability during pre-stimulus baseline. None of the balanced dichotomies are decodable during IA or IP ($p > 0.05$ for all dichotomies).

(B) Dichotomy decodability during the stimulus presentation period. All three named stimulus dichotomies are highly decodable both during IA and IP. Correlated dichotomies also demonstrated above-chance decodability. Dichotomies: purple, $p_{IA} = 6.8x10^{-1}$, $p_{IP} = 6.6x10^{-14}$, brown, $p_{IA} = 2.2x10^{-9}, p_{IP} = 6.0x10^{-14}$, pink, $p_{IA} = 1.1x10^{-1}$, $p_{IP} = 6.7x10^{-14}$. Notably, context is not significantly decodable in either IA or IP (red, $p_{IA} = 0.24, p_{IP} = 0.38$).

(C) Dichotomy CCGP for VTC during the stimulus presentation period. Two stimulus dichotomies are in an abstract format in IA and all three are in an abstract format in IP (purple, $p_{IA} = 0.0054, p_{IP} = 0.0036$, brown, $p_{IA} = 0.057, p_{IP} = 0.0029$, pink, $p_{IA} = 0.0030, p_{IP} = 0.0032$). All remaining dichotomy CCGP values are at chance apart from response and context, which are significantly below chance in IP (green, $p_{IA} = 0.93, p_{IP} = 0.96$, red, $p_{IA} = 0.84, p_{IP} = 0.94$).

(D) Dichotomy PS for VTC during the stimulus presentation period. Again, two stimulus dichotomies are in an abstract format in IA, and all three are in an abstract format in IP (purple, $p_{IA} = 0, p_{IP} = 4.3x10^{-1}$, brown, $p_{IA} = 0.73, p_{IP} = 0$, pink, $p_{IA} = 0, p_{IP} = 5.9x10^{-7}$).

(E) Dichotomy decodability analysis for incorrect trials during the stimulus presentation period. Decoders are trained on correct trials and evaluated on incorrect trials (balanced by condition) in IP sessions. Plotting conventions are identical to those described in Fig. 3,S6. Note that all three stimulus identity-related dichotomies are still highly significantly decodable during incorrect trials in IP sessions (purple, $p_{IP-i} = 7.8x10^{-11}$, brown, $p_{IP-i} = 1.1x10^{-13}$, pink, $p_{IP-i} = 8.7x10^{-11}$).

Figure S9. Cross-condition generalization performance for stimulus identity and context defined over stimulus pairs.

To fully disentangle and study the interaction between stimulus coding and context, geometric analysis of balanced dichotomies must be replaced by new analyses that are defined for pairs of individual stimuli, thus allowing for the study of stimulus coding un-ambiguously without arbitrarily grouping together stimuli as is necessary in the balanced dichotomy approach. When considering a pair of stimuli (e.g. A and B) across two contexts (e.g. 1 and 2), there are four possible task conditions (A1, B1, A2, B2). On these points, stimulus (A1A2 vs B1B2) and context (A1B1 vs A2B2) can be decoded in a straightforward manner, but is not informative about the format in which stimulus and context are encoded. The CCGP for stimulus across contexts **(A-C)** and for context across stimuli **(D-F)** provide information about the structure of the two variables and how they interact.

Consider within-context training/testing. The procedure is summarized in **(A)**, which shows a linear decoder (green bar) trained between stimuli A and B in context 1 (green + and – correspond to class labels for training). The decoder is then generalized to context 2, where stimulus identity is decoded (purple bar, + and – for class labels). This procedure is broken down step-by-step for training in **(B)** and testing in **(C)**. In addition, arrows showing persistent stimulus and context coding vectors (black/dashed arrows) have been drawn alongside the vector orthogonal to the hyperplane learned during train/test (colored arrow passing through the bar). Note that, for this formulation of Stimulus CCGP, the stimulus coding vector and the normal vector to the hyperplane are parallel in **(B)** and **(C)**. Thus, in cases with high within-context train/test Stimulus CCGP, stimulus information is present in an abstract format across contexts.

The same procedures for computing CCGP can be applied for studying the format of context organizing across pairs of stimuli **(D-F)**, with schematic details identical to those described above for Stimulus CCGP. Here, high Context CCGP indicates that context is encoded abstractly across the different stimuli.

Figure S10. Additional geometric stimulus-pair analysis for hippocampus and ventral temporal cortex.

(A) Average distances between stimulus representations in hippocampus (HPC) in IA and IP. All plotted points correspond to named, interpretable groups of conditions defined by pairs of stimuli presented in both contexts. For example, the green dot in IA indicates the average distance (~5.1Hz) between the condition centroids for stimulus A and stimulus D (averaged over contexts). Distance is computed as Euclidean distance between the stimulus centroids, each of which is an N (# of neurons) dimensional vector of average firing rates during stimulus presentation. Neuron counts are balanced between IA and IP to allow for direct distance comparisons. Null distributions here are geometric nulls, and are identical to those used for CCGP and PS. Significance of the difference between IA and IP inter-stimulus distances is established by RankSum test computed over stimulus pairs, and n.s. indicates $p > 0.05$.

(B) Same as **(A)**, but for ventral temporal cortex (VTC).

(C) Decodability of stimuli also did not significantly change between IA and IP for HPC. Here, decoding accuracies are reported for each unique pair of stimuli with 1000 repetitions of trial sub-sampling. Null distributions are constructed with trial-label shuffling, and the gray bars correspond to the boundary of the 5th to 95th pctle of the null. Significance of the difference between IA and IP decodability is also established by Ranksum test over average decoding accuracies and n.s. indicates $p > 0.05$.

(D) Same as **(C)**, but for VTC.

Figure S11. Additional context CCGP analysis over stimulus pairs for hippocampus and ventral temporal cortex (stimulus period).

Change in context decoding accuracy from IA to IP evaluated over individual stimulus pairs is shown for the hippocampus **(A)** and ventral temporal cortex **(B)**. Individual points correspond to context decoding accuracy averaged over 1000 repetitions of decoding/CCGP/Parallelism Score estimation with trial re-sampling. All plotting conventions for geometric plots are identical to those used in Fig. 4,S10.

(C) and **(D)** show Context CCGP and Context Parallelism Score over stimuli for VTC. Analogous plots for HPC are Fig. 4G,H.

Two exemplar neurons are shown, one from HPC **(E,G)** and one from VTC **(F,H)** that feature both stimulus tuning and context modulation. Plotting conventions are identical to previous raster/psth plots apart from the colors and conditions plotted, which here are two stimuli (A and B) in the two contexts. Responses to task conditions for the two neurons are summarized in **(G,H)**, which show mean ± s.e.m. firing rates by condition for spikes counted on individual trials during the stimulus period (0.2s to 1.2s after stimulus onset). The same trials used to compute **(G)** and **(H)** are shown in **(E)** and **(F)** respectively, and condition colors are matched between the two sets of plots. Black arrows indicate the direction in which the firing rate for a stimulus is modulated by a shift in context. The HPC neuron **(G)** shows consistent modulation by context since both arrows point downward, whereas the VTC neuron **(H)** shows inconsistent modulation by context since one arrow points downwards and the other points upward.

(I) Change in the consistency of context-modulation for stimuli averaged over all neurons in VTC and HPC. Context modulation consistency is the tendency for a neuron's firing rate to shift consistently (increase or decrease) to encode context across stimuli. This consistency can take on values between 0 (increase in firing rate to encode context for half of the stimuli, decrease in firing rate for the other half) and 4 (either increase or decrease in firing rate for all four stimuli). An interaction effect is observed between context modulation consistency for HPC neurons and VTC neurons in IA and IP in the absence of main effects (2 x 2 ANOVA, $p_{Area} = 0.36$, $p_{IA/IP} = 0.64$, $p_x = 4.5x10^{-5}$), revealing significant increases in context modulation consistency in HPC from IA to IP with concurrent decreases in VTC.

Figure S12. Hippocampal MDS plots showing increases in context and stimulus parallelism for all stimulus pairs.

MDS plots analogous to that shown in Fig. 4K, but plotted in 2D for individual stimulus pairs. Colored points represent the mean condition response of all HPC neurons during IA or IP to a given stimulus in a given context. Stimuli are color coded according to identity (e.g. in A, red points are condition responses to stimulus A and blue points are condition responses to stimulus B), and are connected by a line of the same color to reflect the context coding direction for that stimulus. Stimuli in the same context are connected by shaded lines that are blue for context 1 and red for context 2. Since MDS is conducted independently for IA and IP, individual MDS axes are not directly comparable between IA and IP, but the relative distances are comparable since the number of neurons is matched between IA and IP, and both are dimensionally reduced to the same number of MDS dimensions (N_{dim} = 2). Condition averages are computed using only correct trials. Evidence of disentangling of context and stimulus identity is evident across most stimulus pairs, with the notable exception of the B/D stimulus pair **(E)**, which is perfectly correlated with outcome and therefore cannot be dissociated from outcome using CCGP. The emergence of quadrilaterals with approximately parallel sides for all other stimulus pairs **(A-D, F)** is a signature of disentangling of stimulus identity and context, and clearly demonstrates the abstract format of both variables.

Figure S13. Implementation of geometric changes in hippocampal representation – Stimulus period.

(A) Distances between centroids for balanced dichotomies shown for all regions other than HPC. Plotting conventions are identical to those used in Fig. 5E. Note: neuron counts were only balanced across IA/IP within-region, so distances in different regions are computed in spaces with different dimensionality and are therefore not meaningfully comparable. Significant change in average dichotomy separation determined through Bonferroni MCC RankSum where * indicates $p < 0.05/35$, and n.s. otherwise.

(B) Changes in inter-centroid distance for balanced dichotomies. Points in these plots are the differences between IP and IA distances shown in **(A)**. Null distributions are computed No distances for named dichotomies increased or decreased more than would be expected by chance (outside 5th-95th pctle null).

(C) Firing rates for individual task conditions (8 total) for all regions other than HPC. Plotting conventions identical to those used in Fig. 5I. Task conditions are color coded based on the identity of the presented stimulus (same as Fig. 1B. 5I,J). Significant change in average dichotomy separation determined through Bonferroni MCC RankSum where * indicates $p < 0.05$, and n.s. otherwise.

(D) Changes in hippocampal firing rates for 3 different sub-groups of session pairs: instruction-successful sessions (IS, n = 5 patients, 10 sessions), instruction-unsuccessful sessions (IU, n = 4 patients, 8 sessions), and instances of first-session inference (S1, n = 3 patients, 6 sessions). These session groups were constructed to compare the effect of inference acquisition with the passage of time (see Methods for details). Firing rate changes here are computed during the stimulus presentation period (0.2s to 1.2s after stim onset) from consecutive pre-instruction and post-instruction sessions. Points are average changes in condition-averaged firing rates (8 unique conditions). Changes in firing rate that significantly differed from zero (t-test, $p < 0.05/3$) are indicated with a "*". IS group alone exhibited significant decrease in firing rate. IU group exhibited an increase in firing rate.

Changes in hippocampal single-neuron tuning quantified by 3-way ANOVA (Response, Context, Outcome) with interactions. Significant factors ($p < 0.05$) were identified for every neuron and averages of both the number of factors per neuron **(E)** and the depth of tuning of those factors quantified through $-\log_{10}(p_{ANOVA})$ **(H)** reported (mean ± s.e.m. across neurons) for the IA (red) and IP (blue) groups. Significance of difference between IA and IP for both the number of factors (**E**, $p_{RS} = 0.041$) and the tuning strength (**H**, $p_{RS} = 0.027$) was assessed by RankSum test over neurons between the two groups, and "*" indicates $p_{RS} < 0.05$.

(F) same as **(E),** but for all regions other than HPC.

(I) same as **(H),** but for all regions other than HPC.

(G) The change in the distribution of trials projected along the coding direction for context was visualized during IA (above) and IP (below). The red and blue histograms are the distribution of projected trials from context 1 and 2 respectively, with the red and blue vertical lines indicating the mean of each distribution. Positive and negative values for projection were arbitrarily established by computing the coding vector as (ctxt1 – ctxt2).

(J) Plot showing the fraction of hippocampal neurons that exhibit task selectivity for IA (red) and IP (blue) sessions. Selectivity is determined independently for every neuron using a 4x2 ANOVA (Stimulus Identity, Context), with a per-factor significance threshold of $p < 0.05$. Significant differences in tuned fractions between IA and IP assessed with z-test. **(J)** Plot showing the fraction of hippocampal neurons that exhibit task selectivity for IA (red) and IP (blue) sessions. Selectivity is determined independently for

every neuron using a 4x2 ANOVA (Stimulus Identity, Context), with a per-factor significance threshold of $p < 0.05$. Significant differences in tuned fractions between IA and IP assessed with z-test.

(K) same as **(J)**, but for VTC.

Figure S14. Implementation of geometric changes in hippocampal representation – Baseline period.

(A) Average variance of individual trials projected onto the coding direction for every dichotomy. Plotting conventions identical to those in Fig. 5I. Average variance along coding directions decreased significantly between IA and IP sessions ($p_{RS} = 6.5x10^{-1}$, RankSum over dichotomies).

(B) Changes in variance between IP and IA for all dichotomies shown in **(A)**. No named dichotomies fall outside the 5th-95th pctle of the null distribution.

(C) Population distance between dichotomy centroids for dACC at baseline. All plotting conventions identical to those used in Fig. 5G. Average distance between dichotomy centroids increased when comparing IA to IP sessions ($p_{RS} = 2.9x10^{-8}$, RankSum over dichotomies). Notably, context centroids emerged as significantly separated in IP than expected by chance ($p_{IA} = 0.48$, $p_{IP} = 0.0065$)

(D) Changes in distance between IP and IA sessions for all dichotomies shown in **(C)**. Context alone (red, $p_{\Delta} = 0.047$) exhibited a greater increase in distance than expected by chance.

(E) Same as **(A)**, but for projective variance in dACC during the baseline. Average variance along coding directions increased significantly between IA and IP sessions ($p_{RS} = 6.0x10^{-3}$, RankSum over dichotomies).

(F) Same as **(B)**, but for differences in dACC variance during the baseline computed using **(E)**.

(G) Baseline firing rate averaged by condition (8 total) for the hippocampus. Plotting conventions are identical to those in Fig. 5I, S13C. Reduction from IA to IP is significant ($p < 0.05$, RankSum over conditions).

(H) Baseline firing rates averaged by condition for all regions other than hippocampus. Significance of change in firing rate also assessed by RankSum over conditions ("*" indicates $p < 0.05$, n.s. otherwise). Note: most regions (apart from AMY) exhibit slight, but significant increases in baseline firing rate during in IP compared with IA.

Figure S15. Additional analysis of the effect of instructions on behavior.

Task performance shown as a moving average over blocks for the instruction-successful (IS) subject group **(A,B)**, the instruction unsuccessful (IU) subject group **(C,D)**, and the session one inference (S1) subject group **(E,F)**. Plots are similar to those shown in Fig. 1F,G, except Pre and Post plots show performance over time for sessions recorded immediately preceding and immediately following verbal instructions describing latent task structure. For example, for the IS group, pre-instruction performance is plotted in **(A)** for non-inference trials (black) and inference trials (gray). Average performance is computed as a moving average with a 3-block window on the last three trials before a context switch (non-inference) and on the first inference trial after a switch (inference). Error bars are standard errors computed over subjects. Chance performance is indicated with the dashed line at $y = 0.5$.

Figure S16. Additional analysis of the effect of instructions on hippocampal neural geometry.

Geometric metrics shown are computed over balanced dichotomies, and are plotted using the same conventions as discussed previously (see Fig. 2, methods for details), except for the left and right columns of each plot correspond to pre-instruction and post-instruction geometry respectively. Plots in the figure are segregated according to subject group (instructions successful - IS, instructions unsuccessful IU, session one inference - S1) and analysis period (Stimulus, Baseline). Only context is plotted as a named dichotomy for visual clarity.

(A) CCGP (context, red, $p_{pre} = 0.27, p_{post} = 0.046, p_{RS} = 1.4x10^{-31}$) and **(B)** PS (context, red, $p_{pre-c} = 0.029,\ p_{post-c} = 3.5x10^{-6}, p_{post-i} = 0.0028$) for the IS group during the stimulus period.

(C) Decoding accuracy (context, red, $p_{pre-c} = 0.35, p_{post-c} = 0.0014, p_{post-i} = 0.55, p_{RS} = 1.4x10^{-20}$), **(D)** CCGP (context, red, $p_{pre} = 0.33, p_{post} = 0.0037, p_{RS} = 3.0x10^{-34}$), and **(E)** PS (context, red, $p_{pre-c} = 0.017, p_{post-c} = 7.5x10^{-8}, p_{post-i} = 0.40$) for the IS group during the baseline period.

(F) CCGP (context, red, $p_{pre} = 0.56, p_{post} = 0.39, p_{RS} = 0.004$) and **(G)** PS (context, red, $p_{pre} = 0.81, p_{post} = 0.95$) for the IU group during the stimulus period.

(H) Decoding accuracy (context, red, $p_{pre} = 0.45, p_{post} = 0.45, p_{RS} = 0.68$), **(I)** CCGP (context, red, $p_{pre} = 0.45, p_{post} = 0.47, p_{RS} = 0.15$), and **(J)** PS (context, red, $p_{pre} = 0.93, p_{post} = 0.30$) for the IU group during the baseline period.

(K) CCGP (context, red, $p_{pre} = 0.23, p_{post} = 0.19, p_{RS} = 0.0045$) and **(L)** PS (context, red, $p_{pre} = 6.3x10^{-8}, p_{post} = 4.5x10^{-7}$,) for the S1 group during the stimulus period.

(M) Decoding accuracy (context, red, $p_{pre} = 0.37, p_{post} = 0.47, p_{RS} = 0.036$), **(N)** CCGP (context, red, $p_{pre} = 0.30, p_{post} = 0.50, p_{RS} = 5.9x10^{-7}$), and **(O)** PS (context, red, $p_{pre} = 1.7x10^{-5}, p_{post} = 0.029$) for the S1 group during the baseline period.

Supplementary Sections

S.1 Hippocampus encodes context abstractly at baseline during inference sessions.

We analyzed the baseline period preceding stimulus onset (Fig. 2, inset) to study the geometry of task representations that might persist from the previous trial (all labels were defined by the prior just completed trial for this analysis). Context was encoded as an abstract variable in the HPC in IP and not in IA sessions. Unlike the stim epoch, context was the only named dichotomy to emerge as significantly decodable in the HPC for IP sessions (Fig. 2D, red, IA vs. IP, $p_{RS} = 1.1x10^{-33}, p_{IA} = 0.35, p_{IP} = 6.4x10^{-5}$, S4B), indicating that any task condition information from the previous trial other than the context (i.e. outcome, response) was not significantly decodable from the population during this epoch. Context during baseline was also encoded abstractly as shown by the increase in CCGP (Fig. 2E, S4C, red, IA vs. IP, $p_{RS} = 2.4x10^{-34}$, $p_{IA} = 0.45, p_{IP} = 7.7x10^{-6}$) and PS (Fig. S4D,E, red, $p_{IA} = 0.37, p_{IP} = 1.2x10^{-1}$). These results indicate that an abstract representation of context that persisted from the previous trial also emerged in the hippocampus in sessions where patients performed inference.

S.2 Hippocampal context representation does not rely on classically tuned neurons

To determine whether the geometric findings presented here arose as a consequence of strong classical (univariate linear) tuning, two ablation analyses were conducted in which all neurons that had significant univariate linear tuning to at least one task variable were excluded and all measures were re-computed. In the first analysis, neurons tuned to one or several of response, reward, or context (3-Way ANOVA, any main effect $p < 0.01$) were excluded. Of the 494 neurons in HPC, 41 were excluded for exhibiting classic tuning prior to recomputing all measures. Qualitatively, findings remained largely unchanged by the exclusion of this population of specialized neurons (Fig. S5A-E). The Parity dichotomy was no longer decodable above chance in IP sessions (Fig. S5A, orange, $p_{IP-c} = 0.056$), but the rise in SD (average over dichotomies) was still present (Fig. S5A, $p_{RS} = 0.0012$, Ranksum over dichotomies). A similar analysis was also conducted for the baseline representation of context, and again the findings qualitatively remained unchanged (Fig. S5F-J). Note that, although the decodability of context at baseline during IP sessions did decrease in significance when conducting the error trial analysis (Fig. S5I, red, $p_{IP-c} = 0.12$), it was still significantly increased compared to IA sessions ($p_{RS} = 5.7x10^{-19}$) and was significantly reduced in error trials ($p_{IP-} = 0.37$, IP-c vs IP-i, $p_{RS} = 0.0002$). In the second analysis, neurons tuned to stimulus identity or context (2-Way ANOVA, any main effect $p < 0.01$) were excluded. In this analysis, 82 neurons were excluded, leading to a loss of decodability of stimulus dichotomies due to the removal of visually-selective neurons as expected (Fig. S5K-M). Context, however, remained decodable and present in an abstract format. Together, these analyses indicate that the abstract context representation in the hippocampus is a highly distributed variable whose geometry is not a simple consequence of strong univariate tuning of a small population of "context" neurons.

S.3 Hippocampal context and stimulus representations are not driven by SOZ neurons

To ensure that the reported findings were not influenced by pathology, in a further control analysis we excluded all HPC neurons (86/494) recorded from electrodes that were clinically identified to reside in medial temporal seizure onset zones (SOZs). Repeating our analyses on the neurons that remain revealed that results were qualitatively unchanged (Fig. S6), though the parallelism for context was now significantly above chance during IA sessions for both the stimulus (Fig. S6C, red, $p_{IA} = 3.5x10^{-7}$) and baseline (Fig. S6F, red, $p_{IA} = 0.0022$) periods. These findings suggests that neurons residing in hippocampal tissue with high disease burden do not meaningfully contribute to the task representation[17], thereby leading to marginally increased representation strength once these neurons are removed (note that the decodability and CCGP for context in IA are still not significantly different from chance, confirming our finding).

S.4 Changes in hippocampal single-neuron tuning explain changes in representation with inference

To determine if changes in univariate tuning could partially explain the observed changes in the hippocampal representation from IA to IP sessions, average tuning properties of single neurons were computed. Both the number of significant factors (main effects or interactions) per neuron (3-way ANOVA – Response, Context, Outcome, $p < 0.05$) and the tuning strength of neurons for those significant factors indexed by the log-p-value of those ANOVA factors were significantly elevated in IP compared to IA (Fig. S13E, $p_{RS} = 0.010$, S13H, $p_{RS} = 0.0089$). These effects were only observed in HPC, with either no change or a decrease in tuning observed in the other areas except for a significant increase in vmPFC (Fig. S13F,I). We also separately considered the linear and non-linear (interaction) terms for the 4 (Stimulus) x 2 (Context) ANOVA and found that, while the fraction of hippocampal neurons exhibiting significant ($p < 0.05$, Main Effect) stimulus identity tuning significantly decreased from IA to IP (Fig. S13J, 21.3% vs. 19.1% of neurons, $p = 0.002$), the fraction of neurons exhibiting significant context tuning increased from IA to IP (Fig. S13J, 9.5% vs. 15.7% of neurons, $p = 2.2x10^{-6}$), which could also partially explain the increase in centroid distance and resultant context decodability. Such "context neurons", however, cannot account for this effect fully as their removal from the pseudopopulation does not qualitatively alter the hippocampal population geometry (Fig. S5). As a control, this analysis was also performed for VTC neurons, in which we did not find a significant difference in the percentage of neurons with univariate context tuning (Fig. S13K).

S.5 Context representations outside of the hippocampus

The dorsal Anterior Cingulate Cortex (dACC) was the only other region to exhibit significant changes in its latent context representation as a function of patients' ability to perform inference. However, this context representation was limited to the baseline period, since geometric analysis for the stimulus period revealed an absence of significant context decodability and CCGP in IA and IP sessions (Fig. 2C, S3A). However, the PS for context in the dACC did increase significantly from IA to IP sessions (Fig. S3B, red, IA vs. IP, $p_{RS} = 2.4x10^{-19}, p_{IA} = 0.99, p_{IP} = 3.8x10^{-12}$), reflecting an increase in context parallelism in condition averages that was not detectable with the other metrics (which are based on single trial decoding).

During the baseline, however, context also emerged as the only decodable dichotomy in IP sessions in the dACC (Fig. S4A,H, red, $p_{IA} = 0.37, p_{IP} = 0.049$). We found this context variable was also represented in an abstract format, emerging as the dichotomy with the highest CCGP (Fig. S4F,I, red, $p_{IA} = 0.26, p_{IP} = 0.018$) and PS (Fig. S4G,J, red, $p_{IA} = 0.18, p_{IP} = 0.013$) in IP sessions, while being at chance for both metrics in IA sessions. As in HPC, the lack of decodability for previous trial outcome and response suggests that any variables encoded at earlier epochs during the previous trial (e.g. post-reply delay or outcome) were extinguished from the population by the onset time of the current baseline epoch.

Analysis of dichotomy centroid distances during the baseline period revealed context also emerged as the dichotomy with the greatest separation in the dACC in IP trials (3.9 vs. 5.7 Hz, $p_{IA} = 0.48, p_{IP} = 0.0065, p_{\Delta Dist} = 0.046$, Fig. S14C,D). However, this implementation of the context representation was achieved through significant increases in condition-wise firing rates (Fig. S14H, dACC IA vs IP, $p_{RS} = 0.049$) as opposed to the firing rate decrease observed in the hippocampus. Together, these analyses indicate that a weak, but nonetheless significant, representation of latent context encoded in an abstract format also emerged in the dACC, and that this variable was accommodated in the representation through a different implementational strategy.