An abstract relational map emerges in the human medial prefrontal cortex with consolidation

Alon Baram[1], Hamed Nili[1,2], Ines Barreiros[1], Veronika Samborska[1], Timothy E. J. Behrens[1,3] and Mona M. Garvert[1,4]

[1]Oxford Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, UK

[2]Department of Excellence for Neural Information Processing, Center for Molecular Neurobiology (ZMNH), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany

[3]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, UCL, London, UK

[4]Juniorprofessorship of Neuroscience, Faculty of Human Sciences, Julius-Maximilians-University of Würzburg, Germany

Corresponding authors: alon.baram@ndcn.ox.ac.uk, mona.garvert@uni-wuerzburg.de

Summary

Understanding the structure of a problem, such as the relationships between stimuli, supports fast learning and flexible reasoning. Rodent work has suggested that abstraction of structure away from sensory details occurs over the course of multiple days, in cortex. However, direct evidence of such explicit relational representations in humans is scarce, and it is unclear whether they emerge on similar timescales. Here, we combine a graph-learning paradigm with functional magnetic resonance imaging (fMRI) to look for such a relational map in the human brain. We first trained participants on two associative graphs with the same structure. We then scanned participants twice while they used this knowledge, with several days between scanning sessions. We found that the medial temporal lobe represented the specific associations between stimuli in each graph in both scanning sessions. In the medial prefrontal cortex (mPFC), on the other hand, fMRI repetition suppression and representational similarity analysis demonstrated the emergence of an abstract relational map over sessions. These results shed new light on how neural representations organising relational knowledge change with time.

Introduction

Humans and other animals constantly find themselves in situations they have never experienced before. How does our brain manage to choose adaptive behaviours given this diversity of experiences? Luckily, our world is replete with statistical structure (Shepard, 1987), which our brain represents as part of a “cognitive map” (Behrens et al., 2018; Tolman, 1948). One particularly effective way of harnessing this structure for generalisation is to separate (or “factorise”) the representation of the environment’s relational structure from its specific sensory details (Whittington et al., 2020). A potential locus for this abstraction is the medial prefrontal cortex (mPFC), which has been implicated in recognising commonalities (“schemas”) across experiences (Baldassano et al., 2018; El-Gaby et al., 2024; Gilboa and Marlatte, 2017; van Kesteren et al., 2013; Schapiro et al., 2013; Tse et al., 2007, 2011) and using these for inference (Zeithamova and Preston, 2010), as well as in concept learning (Kumaran et al., 2009; Mack et al., 2020), structural generalisation of reinforcement learning problems (Baram et al., 2021; Bein and Niv, 2025; Samborska et al., 2022) and generalisation of spatial tasks across different paths (Kaefer et al., 2020; Morrissey et al., 2017; Tang et al., 2023; Yu et al., 2018).

In addition, a recent line of research, inspired by rodent spatial studies, has shed new light on the relational properties of mPFC codes. Like the entorhinal cortex, the human mPFC contains grid cells (Jacobs et al., 2013), which are active on a regular triangular lattice during spatial navigation (Hafting et al., 2005). Importantly, grid cells maintain their cell-to-cell activity covariance structure across different spatial environments (Barry et al., 2012; Fyhn et al., 2007; Yoon et al., 2013), suggesting they encode the statistical relationships common to all spatial environments. Notably, a flurry of studies using a technique to indirectly index grid cell activity suggests that the same brain network that displays “grid-like” activity during spatial navigation (Doeller et al., 2010) also displays it during navigation to novel “locations” in a wide range of non-spatial and abstract 2D tasks (Bao et al., 2019; Bongioanni et al., 2021; Constantinescu et al., 2016; Park et al., 2021; Veselic et al., 2023). In most of these studies, the strongest grid-like signals emerge in the pregenual mPFC, suggesting this region encodes the 2D relational structure in these

tasks. While these studies focused on 2D environments, theory suggests the 2D grid representation is merely a special case of abstract structural encoding applicable to a much wider range of task structures (Whittington et al., 2020, 2022).

The mPFC has also been implicated in a related, yet distinct literature on abstraction via consolidation (i.e. on the time scale of days or weeks) across multiple memories (Kumaran et al., 2016; McClelland et al., 1995a). For example, mPFC activity in rodents initially reflects specific perceptual features of learnt associations, and only later encodes generalisable relational features that are shared between associations (Morrissey et al., 2017; Richards et al., 2014; Tse et al., 2011). However, the extent to which similar dynamics apply in humans remains unknown (Tompary and Davachi, 2017), and a more precise understanding of the form of such abstract representations has remained elusive.

Here, we hypothesised that the mPFC abstracts an explicit relational representation of the environment over the course of multiple days. We used fMRI repetition suppression (Barron et al., 2016; Grill-Spector and Malach, 2001) and representational similarity analysis (RSA) (Kriegeskorte et al., 2008; Nili et al., 2014) to assess whether such an abstract map emerges after training participants extensively on two graphs with different contexts, where the nodes are objects and the edges are associations between them. The objects and the graph structure were identical in both contexts, but pairwise distances between objects were decorrelated through their distribution on the two graphs. This allowed us to simultaneously test where the brain represents the task-relevant and irrelevant graphs in a task where relevance regularly switches. Crucially, we were also able to test for an abstracted graph representation which reflects the underlying common structure. To assess the role of consolidation in this abstraction, we examined these representations in two scanning sessions at least 24 hours apart.

We found that the medial temporal lobe represented the relevant and irrelevant graphs from the first scanning session, while a representation of the relational structure that was independent of the specific stimulus identities emerged in the mPFC over time. This was corroborated in two fully independent analyses, demonstrating that mPFC slowly

extracted an abstracted map that could presumably facilitate fast learning in new, but similarly structured situations.

Results

Subjects learn and flexibly switch between two graphs with the same structure

We designed a task to examine how the representation of abstract cognitive maps changes over the timescale of days. 23 participants were extensively trained on the relationships between a set of objects, each forming a graph (Figure 1A, B; training procedure outlined in the Methods section and Supplementary Figure 1). Importantly, the two graphs shared the same non-symmetric structure and differed only in terms of the distribution of objects across the graph´s nodes. Each object´s position on the blue graph thus corresponded to a unique object in the same position on the red graph. The objects were distributed such that pairwise distances between the objects were uncorrelated across the two graphs.

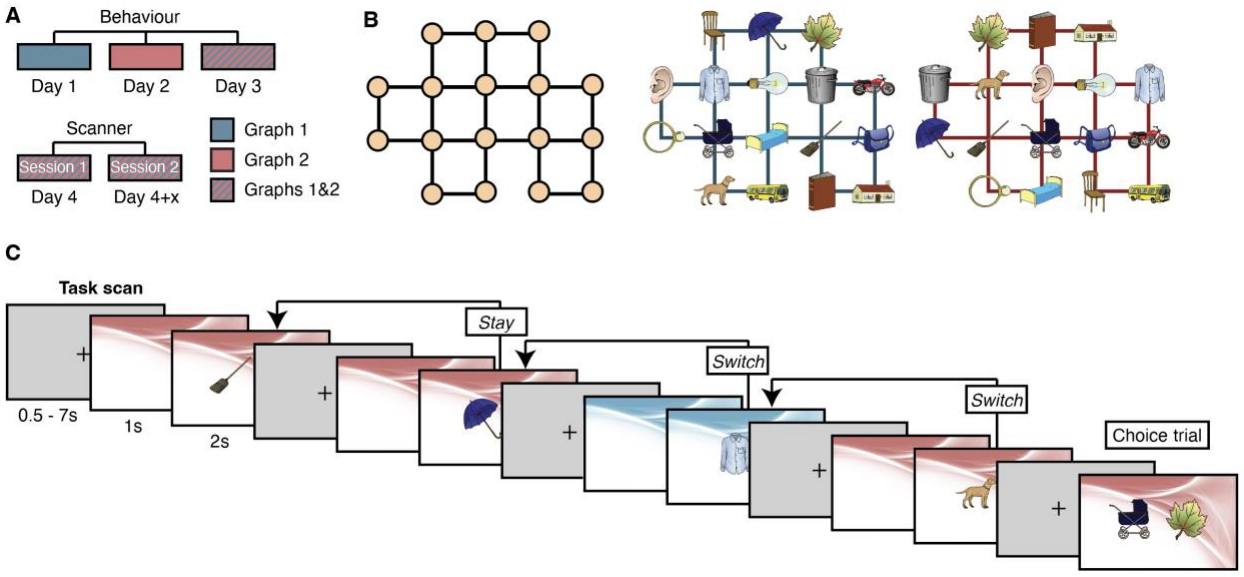

Figure 1. Experimental design. A Experimental design. Participants were trained online on the two graphs over three subsequent days (Supplementary Figure 1). On days 1 and 2, only one of the two graphs was trained, on day three both graphs were trained. Scan sessions took place on day 4 (session 1) and between one day and one week after the first session (session 2). **B** Graph structure (left) and example object distribution for the two graphs (middle and right). Across subjects, object distributions were randomised, and pairwise distances between objects on the two graphs were orthogonalised for every subject. **C** Task in the scanner. Participants were exposed to random sequences of objects after a brief context primer. Trials could be classified as *stay* trials (no change in context from trial to trial) and *switch* trials (change in context from one trial to the next). Occasionally, two objects were presented simultaneously. Participants were asked to indicate which of the two presented objects could be reached from the previously presented object in fewer steps in the relevant graph (*choice* trials).

Participants performed different tasks online, aimed at teaching them the pairwise associations between objects as well as sequences through the graph (Supplementary Figure 1). The underlying structure of the graphs was not made explicit. Each graph was learned separately on two subsequent days to minimise interference. Performance on the explicit structure-learning tasks did not differ between the two graphs learned on two separate training days (all $p > 0.15$, Supplementary Figure 1), suggesting that participants learned both graphs equally well. On the third day of training, participants repeated most of the tasks for both graphs to ensure that they learned to rapidly switch between them in a context-dependent manner. In all tasks, the context (i.e. the currently relevant graph) was indicated by a red or blue background image (order counterbalanced across subjects).

After three subsequent days of learning the object relationships, we presented subjects with sequences of objects in the scanner (Figure 1C). The first scanning session took place on the day after participants completed the training (session 1, day 4). The second scanning session took place between one day and one week after the first scanning session (session 2, day 4+x). The task in the scanner involved frequent context switches. The relevant context was always signaled by the same background color that subjects had learned to associate with a particular graph's object distribution during the training procedure. On 12.5% of trials (once after each object), participants were asked to indicate by a button press which one of two presented objects could be reached in fewer steps from a preceding object on the currently relevant graph (*choice* trials). The purpose of this cover task was to keep participants´ attention focused on the embedding of each presented object within its graph structure.

Participants learned the graph structures very well, as demonstrated in a subset of participants ($n = 19$), who were asked to arrange both sets of objects relative to one another in an "arena", after the last scanning session (Figure 2A-C). The pairwise distances in this arrangement correlated strongly with the 'true' link distances between pairs of objects (Figure 2D, $t_{18} = 6.64$, $p < 0.001$ for Graph 1, and $t_{18} = 6.56$, $p < 0.001$ for Graph 2, based on Fisher-transformed correlation coefficients). Moreover, the correlation strengths between the two graph structures correlated with each other ($r = 0.57$, $p = 0.01$,

based on Fisher-transformed correlation coefficients) and were not significantly different from each other (t_{18} = 0.25, p = 0.81, based on Fisher-transformed correlation coefficients). These results suggest that participants encoded both graph structures equally well. Indeed, we could recover both graph structures by applying multi-dimensional scaling to the distance matrix averaged across subjects (Figure 2B-C).

More importantly, participants also used their knowledge about the graph structures to solve the task in the scanner. Performance in the choice task was significantly above chance in both sessions (session 1: t_{22} = 16.49, p < 0.001, session 2: t_{21} = 17.35, p < 0.001), with an improvement in performance from scanning session 1 to 2 (Figure 2E, t_{21} = 2.94, p = 0.008). Across the two sessions performance in the choice task was highly correlated with performance in the post-scan arena task (Figure 2F, r = 0.65, p = 0.003), suggesting that participants who had learned the graph structures particularly well benefitted the most from this knowledge in the scanner.

In a logistic mixed-effects model, with subject included as a random effect, the likelihood of choosing the left or right option was significantly influenced by the difference in distances between the target object and the objects on either side. The difference in distances on the relevant graph between the target object and the chosen object had a strong positive effect on the likelihood of choosing a given option (Coef = 0.747, SE = 0.031, z = 24.12, p < 0.001, Figure 2G). This suggests that participants were more likely to choose the option on the side that was closer to the target object on the relevant graph. A smaller, yet significant, effect was also observed for the difference in distances on the irrelevant graph between the target object and the two options (Coef = 0.067, SE = 0.024, z = 2.79, p = 0.005). This indicates that the irrelevant graph also had a measurable, though smaller, influence on decision-making compared to the relevant graph

In a linear mixed-effects model with subject included as a random effect, the distance between the target object and the chosen object on the relevant graph had a strong positive effect on log response time (Coef = 0.363, SE = 0.010, z = 35.35, p < 0.001). Additionally, the distance between the target object and the unchosen object on the relevant graph also had a significant positive effect on response time (Coef = 0.075, SE

= 0.010, z = 6.87, p < 0.001). These results indicate that the length of mental simulation on the relevant graph, both from the chosen and unchosen options to the target, contributed significantly to longer response times. There was also a modest positive effect of the distance between the target object and the chosen object on the irrelevant graph (Coef = 0.037, SE = 0.010, z = 3.86, p < 0.001), suggesting some influence of irrelevant graph knowledge on response times. However, the distance between the target object and the unchosen object on the irrelevant graph did not significantly affect response times (Coef = -0.005, SE = 0.010, z = -0.49, p = 0.62).

Next, we examined whether these effects of graph distances on response times in the scanner correlated with explicit knowledge of the graphs. We ran a separate regression for each participant (predicting, again, log response times with the chosen and unchosen distances on the relevant and irrelevant graphs), and correlated across subjects the resulting parameter estimates with an index of performance in the arena task - the correlation coefficients between the arena distances and “true” distances. We found that both the distance to the chosen object and the distance to the unchosen object on the relevant graph had a stronger influence on response times for participants who performed better in the arena task (chosen: r = 0.65, p = 0.003; unchosen: r = 0.46, p = 0.045, Figure 2I, K). This suggests that participants with a better understanding of the graph structures used this knowledge more effectively to solve the choice task.

Notably, the effect of distance to the chosen object on the irrelevant graph was negatively correlated with arena task performance (r = -0.71, p = 0.001, Figure 2J), indicating that participants who knew the graph structures well were less distracted by irrelevant information. We found no significant correlation between performance in the arena task and the effect of unchosen distance on the irrelevant graph on response times, r = -0.31, p = 0.20, Figure 2L).

Together, these behavioural results suggest that participants had acquired and used structural knowledge about the relationships between objects to solve the cover task.

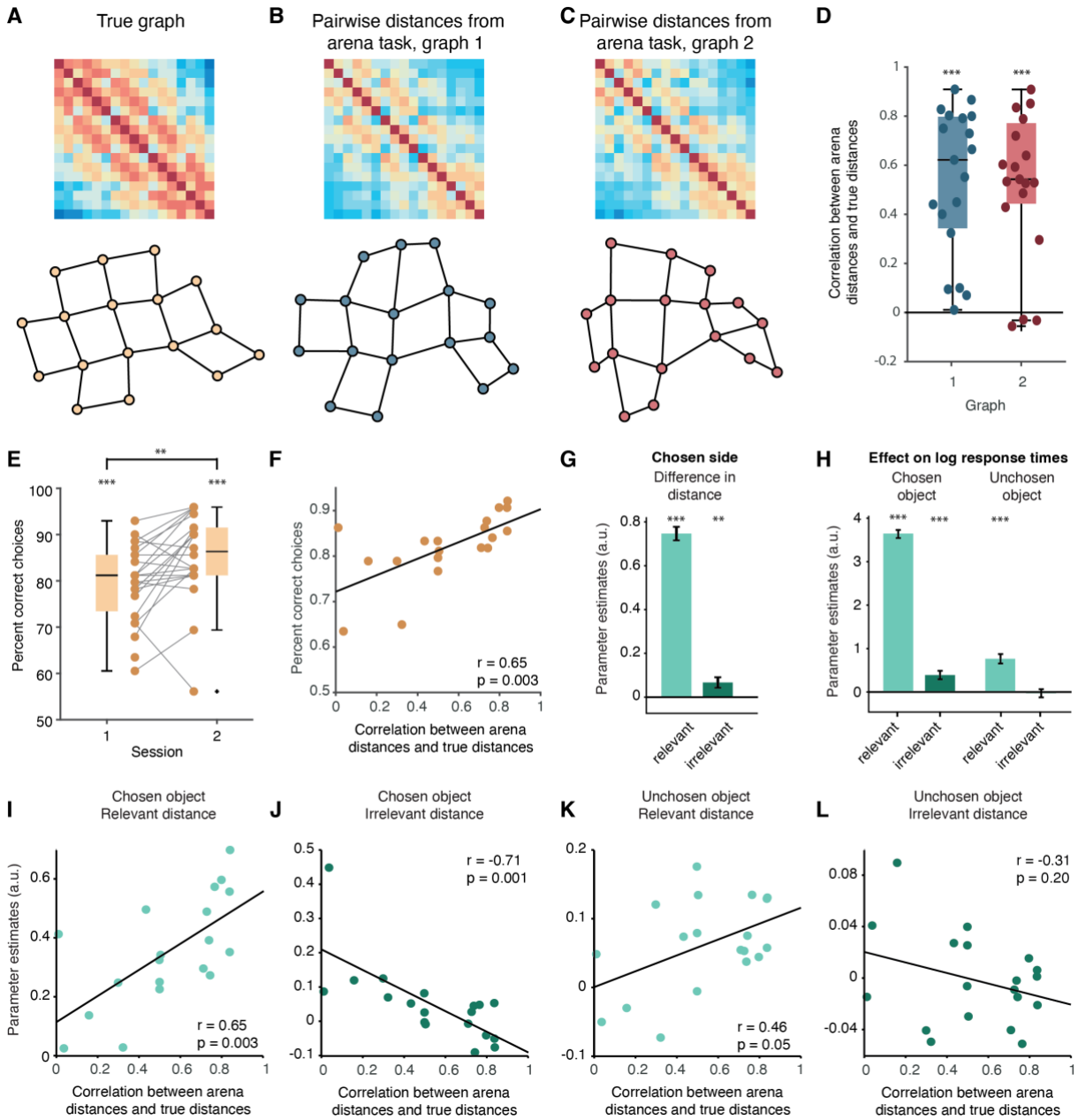

Figure 2. Behavioural performance. A Euclidian distances between object pairs (top) and reconstructed graph after projecting link distances into a two-dimensional space using multi-dimensional scaling (bottom). **B** Pairwise distances between objects of graph 1 as arranged by participants in an arena at the end of session 2, averaged across participants (top). Multi-dimensional scaling of the distance matrix was used to recover the graph (bottom). **C** Same as **B**, for graph 2. **D** Correlation between link distances in **A** and distances between positions in the arena for each graph (**B** and **C**). **E** Percent correct decisions in the choice task performed in the two scan sessions. Performance improves from session 1 to session 2. **F** Correlation between performance in the arena task and percent correct choices in the scanner, averaged across the two sessions. **G** Logistic mixed-effects regression coefficients reflecting the influence of the difference in distance to the chosen and unchosen objects on the relevant and irrelevant graph on chosen side. **H** Linear mixed effects regression coefficients reflecting the influence of relevant and irrelevant distances for chosen and unchosen objects on response times. In **G** and **H,** data from sessions 1 and 2 are pooled. **I** Correlation between arena performance and the influence of chosen/relevant distance on response times, averaged across the two sessions. **J** Correlation between arena performance and the influence of chosen/irrelevant distance on response times, averaged across the two sessions. **K** Correlation between arena performance and the influence of unchosen/relevant distance on response times, averaged across the two sessions. **L** Correlation between arena performance and the influence of unchosen/irrelevant distance on response times, averaged across the two sessions. *p < 0.05, **p < 0.01, ***p < 0.001

Representations of both the relevant and irrelevant graphs in the MTL

To solve this context-dependent task, the brain must represent the object-embedded graph that is currently relevant in each trial. Multiple studies have shown that the MTL represents such associative graphs (Mark et al., 2024; Park et al., 2020; Schapiro et al., 2016) when these are task-relevant. Importantly, we have previously shown that MTL also represents such graphs even when learnt implicitly, and when subjects were engaged in a task for which the graph structure was irrelevant (Garvert et al., 2017; Tacikowski et al., 2024). We thus hypothesised that the MTL might simultaneously represent both the currently relevant and irrelevant graphs.

To test this, we used functional magnetic resonance imaging (fMRI) repetition suppression (Grill-Spector and Malach, 2001), a technique based on the assumption that repeated stimuli evoke a suppressed response proportional to their distance in representational space (Barron et al., 2016). To investigate within-graphs effects, we focused first on *stay* trials (where the context remains stable relative to the previous trial). We reasoned that in regions encoding a map-like representation of the relevant graph, the degree of fMRI suppression should decrease as a function of distance on the relevant graph between the current and preceding object. (Figure 3A top). Meanwhile, in regions encoding the irrelevant graph, fMRI suppression *in the same trial* should scale with the distance *on the other* (currently irrelevant) graph to the (same) preceding object (Figure 3A bottom). Since pairwise object distances were orthogonalised across graphs 1 and 2, we were able to look at the representations of the relevant and irrelevant graphs simultaneously, in the same GLM.

A contrast averaging both relevant and irrelevant distances across both sessions identified a cluster in the left medial temporal lobe (Figure 3B-C). This cluster survived family wise error (FWE) small-volume correction (SVC) within a bilateral entorhinal cortex/subiculum mask (mask from Garvert et al., 2017, Supplementary Figure 2A, $p = 0.03$ corrected at peak level, peak $t_{22} = 4.39$, [-24, -22, -28]), and trended towards significance when SVC was performed in a larger MTL mask of bilateral hippocampus, parahippocampal cortex, and entorhinal cortex ($p = 0.053$ corrected at peak level, peak

t_{22} = 4.94, [-26, -22, -28], mask in Supplementary Figure 2B). In line with the comparable effects for graphs 1 and 2 across all our behavioural measures, we found no significant difference within this mask when comparing the suppression effects averaged across conditions for the two graphs ($t(21) = -1.41$, $p = 0.17$). A 2x2 ANOVA with the factors "Graph relevance" and "Session" did not reveal any significant main effects or interactions (Session: $F(1,20)=0.60$, $p=.446$; Relevance: $F(1,20)=0.64$, $p=.433$; Session × Relevance: $F(1,20)=1.72$, $p=.205$), indicating both the relevant and the irrelevant graphs were represented simultaneously in MTL, in both sessions.

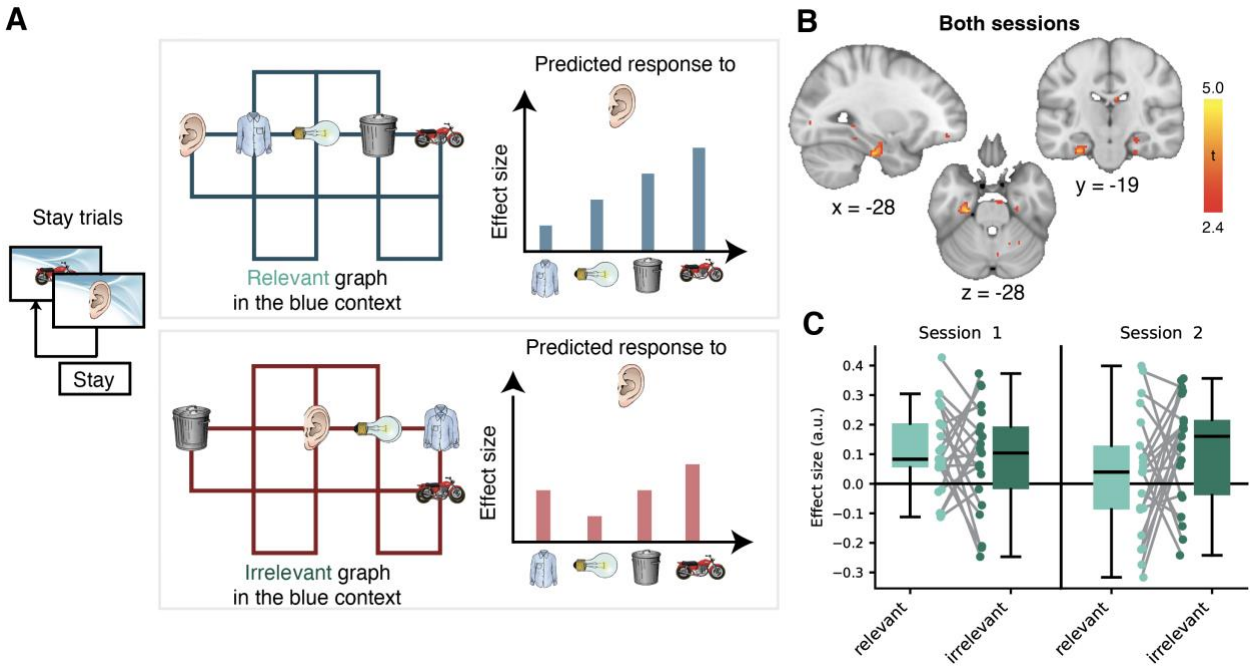

Figure 3. Relevant and irrelevant graph representations. A Analysis principle. On *stay* trials, distances between an object and its predecessor in the object sequence was defined on the relevant graph (here blue) and the irrelevant graph (here red). Visualised are distances from the Ear to four example objects. Since pairwise distances were orthogonalised for the two graphs, we were able to simultaneously assess the representation of relevant and irrelevant graphs. **B** MTL areas representing relevant and irrelevant distances across both sessions. **C** Visualisation of parameter estimates extracted from the left MTL cluster in B.

A generalisable abstract map in the mPFC

The medial prefrontal cortex is known to abstract the commonalities across experiences and represent schematic knowledge (Baldassano et al., 2018; El-Gaby et al., 2024; Gilboa and Marlatte, 2017; Samborska et al., 2022). In addition, in humans

mPFC contains grid cells (Jacobs et al., 2013) and grid-like signals (Bao et al., 2019; Constantinescu et al., 2016; Doeller et al., 2010; Park et al., 2021; non-human primates: Bongioanni et al., 2021; Veselic et al., 2023), suggested to encode statistical structure in 2D spaces in a way that reflects commonalities and ignores particularities across tasks and environments (Behrens et al., 2018; Whittington et al., 2020). Such abstract structure representations, grid-like or not, could be useful for transferring structural knowledge to new situations and thereby speed up learning.

In our experiment, such structure-coding would suggest the presence of an abstract map that is invariant to the specific context a stimulus was presented in. To test for such a cross-graph effects, we once again used repetition suppression, but this time on data from *switch* trials. We defined a cross-graph distance as the distance between the position of the currently presented object on its graph (Ring on Blue graph in Figure 4A) and the position of the preceding object on the other graph (Bed on Red graph). Such a cross-graph suppression effect would only be expected in brain areas that represent structure explicitly, and separately from the stimuli themselves.

A whole-brain analysis across both sessions identified a large significant cluster representing cross-graph distances in mPFC (Figure 4B, FWE-corrected on the cluster level, $p = 0.003$, peak MNI [2, 42, 8] $t_{21}=5.1$). This effect was driven by an emergence of a cross-graph suppression effect in session 2: within an ROI defined by the cross-graph effect across both sessions (threshold $p<0.01$), the difference between sessions was significant (Figure 4C-D, [session 2 > session 1] SVC FWE-corrected on cluster level: $p<0.05$, $t_{21}=3.66$, peak MNI [4,40,12] $t_{21}=3.66$). This was also true when tested on the peak of the both sessions cross-graph effect (MNI [2, 42, 8]: $p<0.005$, $t_{21}=3.1$). While these indirect repetition suppression effects do not allow us to make any inference about the particular types of cells involved in this representation, it is notable that these cross-graph suppression effects emerge across sessions in the same parts of mPFC showing grid-like coding of abstract relationships (mPFC ROI from Figure 2A in (Constantinescu et al., 2016), [session 2 > session 1]: SVC FWE-corrected on cluster level: $p = 0.03$, $t_{21} = 2.38$, Figure 4E).

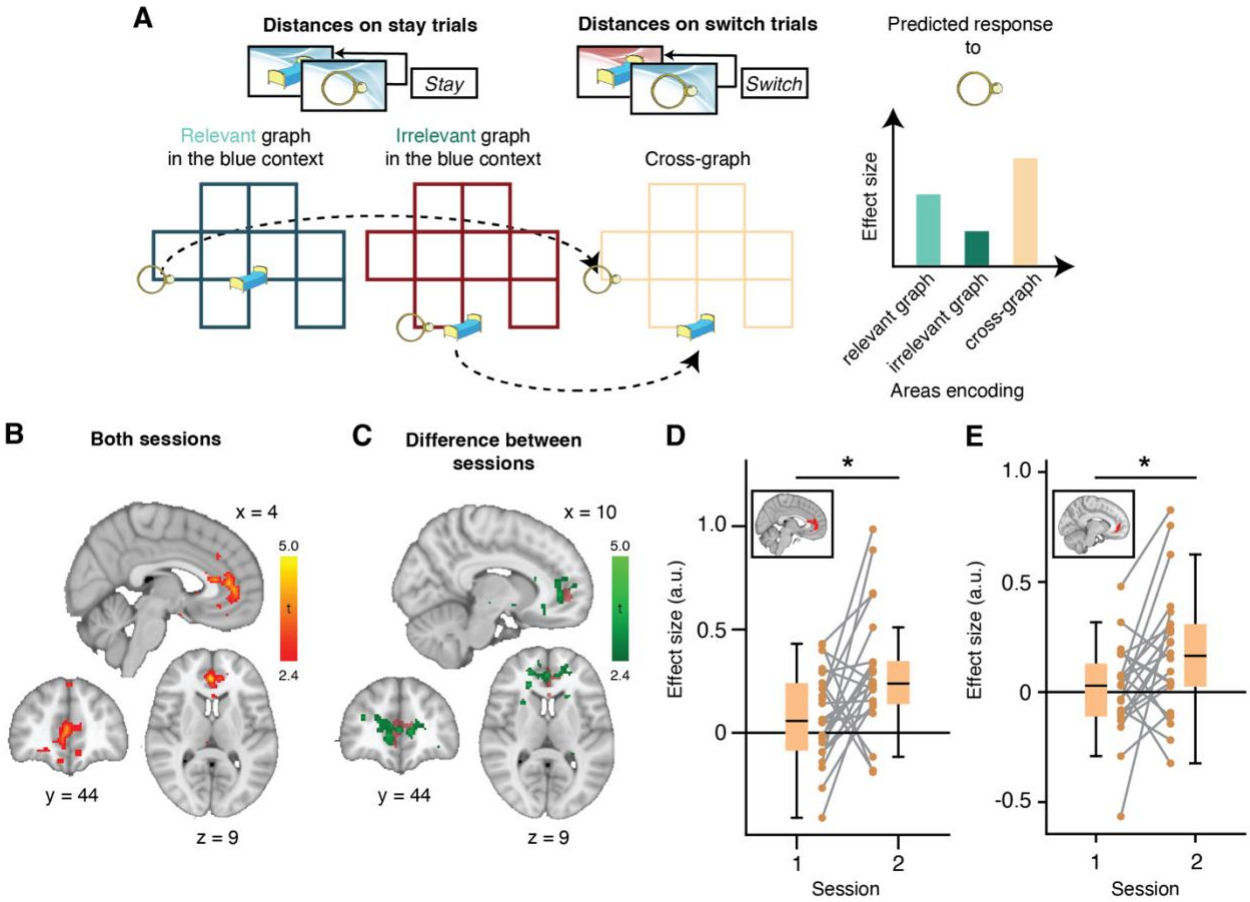

Figure 4. Abstract structure knowledge emerges in mPFC over time. A Logic of the cross-graph suppression analysis on *switch* trials. Cross-graph distances are defined as the distances between the position of the currently presented object on its graph and the position of the preceding object on the other graph. This measure corresponds to distances expected in an area that ignores the specific stimuli in a position, and instead reflects an abstract relational structure. **B** Across both sessions, cross-graph suppression is found in in the medial prefrontal cortex. **C** [session 2 > session 1] effect (green) overlaid on top of the mask extracted from **B** (red). **D** Visualisation of the both-sessions effect from **C**: the cross-graph suppression effect is significantly higher in session 2 compared to session 1 in the ROI from the both-sessions effect (inset). **E** A cross-graph suppression effect also emerges in an ROI defined based on grid-like coding of abstract relationships (Constantinescu et al., 2016). *p < 0.05

Next, we wanted to test for the emergence of an abstracted map in mPFC using representational similarity analysis (RSA, (Kriegeskorte et al., 2008). Like repetition suppression, RSA also tests for the modulation of representations by hypothesised distances between conditions, but multivariately – using the correlation distances between voxel patterns. Importantly, repetition suppression and RSA are fully orthogonal, providing independent evidence for the emergence of an abstract relational structure. Using a searchlight approach (Kriegeskorte et al., 2006), we calculated correlations between patterns of each pair of the 34 conditions (17 objects x 2 graphs, Figure 5A) resulting in a 34 x 34 symmetric matrix (termed “Data Representational Dissimilarity

Matrix", or "Data RDM") per searchlight. We performed this analysis on the cortical surface to ensure that only cortically adjacent voxels are included together in the same searchlight, and across temporally distant scanning runs to control for the effects of time (see Methods). We then constructed a hypothesis distance matrix (termed "Model RDM") by calculating the distance on the abstracted graph according to locations of the objects on their currently relevant graphs (Figure 5B). Note that for within-graph pairs (Figure 5C left) this simply corresponds to the distance on the currently relevant graph, while for across-graphs pairs (Figure 5D left) it corresponds to the cross-graph distance in Figure 4A. Importantly, following the repetition suppression results suggesting the abstract map emerged in mPFC between the two scanning sessions, we performed all RSA analyses on the [session 2 > session 1] contrasts.

We first tested whether RSA detects the emergence of a "full abstracted map" in mPFC – using both within and across-graphs elements of the RDM (Figure 5B left). We found a cluster in the right pregenual mPFC, in the same area as the repetition suppression [session 2 > session 1] cross-graph effect (P=0.01, FWE corrected using permutation test, SVC in a surface projection mask (green mask in Figure 5B right) of the [session 2 > session 1] cross-graph suppression effect from Figure 4C). To investigate this effect further, we tested separately for the within (Figure 5C left) and across (Figure 5D left) graphs parts of the RDM. Both of these analyses revealed small clusters in mPFC (Figure 5C-D, right), though in both cases these clusters did not reach significance (SVC FWE in an anatomical mask of mPFC: within-graph P=0.06; across-graphs P=0.11; SVC FWE in a surface projection mask of the [session 2 > session 1] across-graph effect from Figure 4D: within-graph P=0.08; across-graphs P=0.14). This is perhaps due to lack of statistical power, as these analyses utilise only half of the RDM elements. It is notable, however, that both these clusters were in the same part of mPFC: indeed, the across-graph effect was significant when corrected for multiple comparisons in an mPFC mask of the within-graph effect (p=0.03) and vice versa (p=0.01). This suggests that mPFC contains a shared representational substrate supporting both within- and across-graph structure, i.e. a common, context-general code.

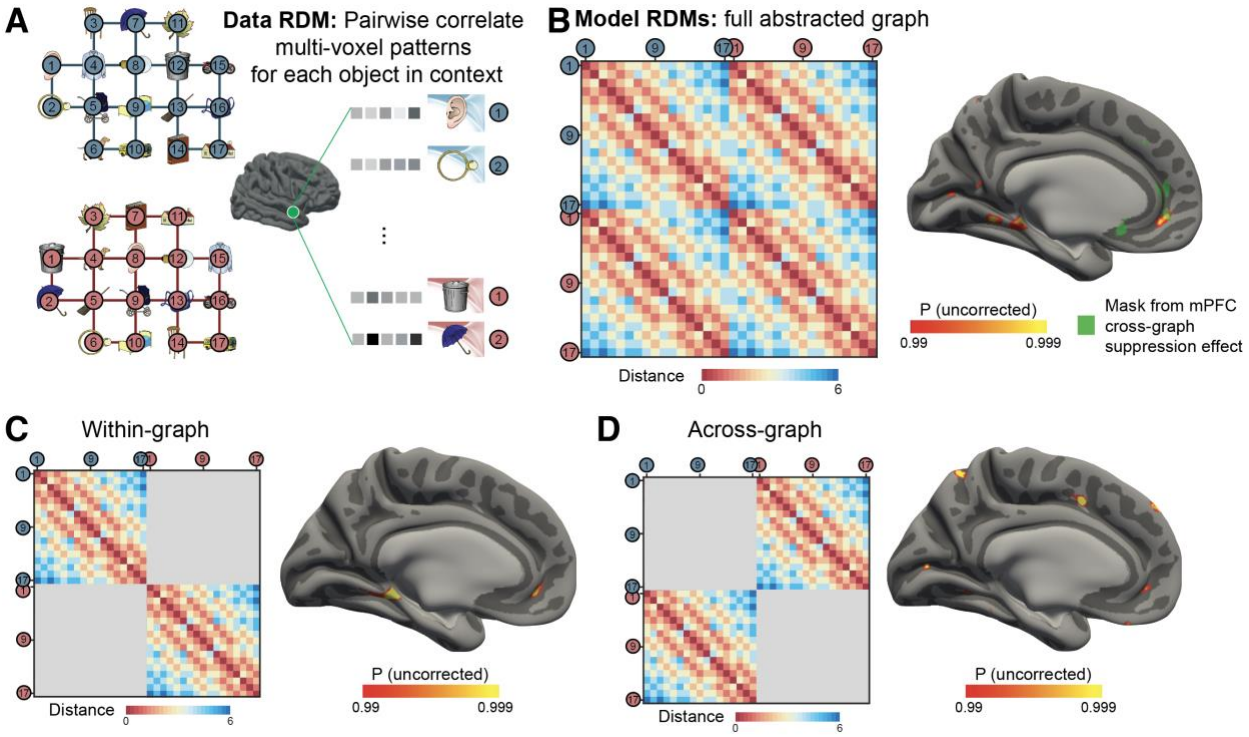

Figure 5. RSA results also suggest an abstracted map emerges in mPFC across sessions. A Representational similarity analysis (RSA) schematic. In each searchlight on the cortical surface, we extracted the dissimilarity of patterns across voxels separately for each of the 17 objects in each graph. **B.** Hypothesis matrix (Model RDM) for the full abstracted map (i.e. both within and across graphs, left) and the right medial surface results for its [session 2 > session 1] effect (right hemisphere, right). The mask from the [session 2 > session 1] cross-graph effect (Figure 4C) is displayed for reference in green. **C.** Same as **B**, but for the RDM elements corresponding only to within-graph pairs of objects (distances in the model RDM are the distances on the currently relevant graph). **D.** Same as **B** and **C**, but for the RDM elements corresponding only to across-graphs pairs of objects (distances in the model RDM are the cross-graph distances, illustrated in Figure 4A).

Discussion

Here, we used a graph-learning paradigm to investigate how abstract relational structures are represented in the brain and how these representations change over time. Participants learned two different graphs characterised by the same underlying associative structures, but a different distribution of stimuli. The retrieval of the graphs – both currently relevant and irrelevant - was mediated by the MTL, in both scanning sessions. Over the course of several days, an abstract, stimulus-independent structural representation emerged in the mPFC. This finding was supported by evidence from both repetition suppression and RSA.

While the mPFC has been implicated in many human studies probing schematic and structural encoding (Baldassano et al., 2018; Baram et al., 2021; Bein and Niv, 2025; Van Kesteren et al., 2010) as well as in the human cognitive maps literature (Park et al., 2020; Schapiro et al., 2013), to our knowledge this is the first time the changes in such an explicit abstract map following consolidation have been investigated. Separating the representation of task structure from stimuli is useful, because it facilitates the generalisation of knowledge across environments that share the same statistics (Behrens et al., 2018; Mark et al., 2024; Whittington et al., 2020). If task knowledge is organised in such a factorised form, then solutions to new tasks do not have to be learnt afresh, they can instead be inferred (Behrens et al., 2018; Mark et al., 2020).

Our results suggest that knowledge about the graph structure continues to be processed after the first scan, even in the absence of further training, and becomes more abstract over time. This is in line with theories of systems consolidation positing that memories change qualitatively after acquisition, from an encoding of specific episodic information to the abstraction of statistical structure that is shared between episodes (Richards et al., 2014). In this way, the neocortex gradually builds semantic knowledge of the world in a process that can take many weeks (McClelland et al., 1995b; O'Reilly and McClelland, 1994). In rodents, such consolidation effects in the mPFC manifest themselves in a gradual loss of selectivity for perceptual features characterizing specific stimuli and an increase in the representation of relational features that are shared

between stimuli (Morrissey et al., 2017). In humans, featural overlap between memories was represented in mPFC (as well as hippocampus) one week after, but not immediately after encoding (Tompary and Davachi, 2017).

Indeed, memory reactivation during sleep plays a critical role in extracting rules and identifying abstract patterns across experiences (Robertson, 2018; Stickgold and Walker, 2013). A potential mechanism that might be involved in this abstraction is neural replay (Foster, 2017), which in humans reorganises experiences according to pre-learnt structures (Liu et al., 2019). Replay also contains abstract structural codes in addition to the sensory ones (Liu et al., 2019), coincides with activation of the default mode network (including mPFC) (Higgins et al., 2020) and has been shown to be constrained by the structure of the task in a way that guides relational inference (Schwartenbeck et al., 2023). Recent theoretical work suggests that the amount of replay of specific experiences depends on the reliability versus generalisability of the environment, thus balancing the trade-off between memory and generalisation (Sun et al., 2023). Future work might explore a direct relationship between replay and the structural mPFC representations such as those we find here.

While not the focus of the current manuscript, it is worth noting the rapid context-switching task subjects experienced in the scanner can be related to other proposed functions of the prefrontal cortex, such as cognitive control (Badre, 2025; Miller and Cohen, 2001; Vaidya and Badre, 2022). Indeed, in such tasks the ventromedial PFC contain rich representations of relevant and irrelevant contexts (Moneta et al., 2023). It is noteworthy that the strongest effects we found in this study – the repetition suppression cross-graph effects – were found when considering only in the *switch* trials, when the context changed. One possibility is that the abstract mPFC map is only needed and used in these more demanding trials. It is possible this map is still represented in *stay* trials, but in a weaker fashion. Speculatively, our trending RSA effects, suggesting an abstract map representation exists in the same regions of mPFC both within and across graphs, might strengthen this hypothesis. Unfortunately, the corresponding RSA analysis to the strong repetition suppression effect – i.e. RSA considering only *switch* trials - was not feasible due to technical considerations of trial numbers for different conditions.

Another possible view on such context-switching tasks is through the lens of spatial cognition. In remapping experiments with rapid context switching ("teleporting"), the hippocampal representation rapidly flickers between the representations of both contexts, effectively implementing an inference process of the current context (Jezek et al., 2011). While our manuscript does not focus on the hippocampus, we note that our MTL effect (partly including hippocampus) of the relevant and irrelevant graphs is compatible with such a remapping effect, as a representation of a relevant graph (which is strong particularly in session 1, see Figure 3C) is indeed what is expected from hippocampal remapping.

In conclusion, our results suggest that the medial temporal lobe represents both relevant and irrelevant graph relationships immediately after learning, while an abstract relational map that captures the underlying structure, independent of specific stimuli, emerges in the mPFC over the course of days. This provides evidence for the abstraction of structural knowledge in the human prefrontal cortex through consolidation.

Methods

Participants

Twenty-three volunteers (aged 18 – 32 years, mean age ± standard deviation 23.2 ± 3.8 years, 14 female) participated in this study. One participant did not return to the lab for the second scan and only the first scan session of this participant was included in the analyses. All subjects reported normal or corrected-to-normal vision and no history of neurological or psychiatric disease. Participants gave written informed consent, and the study was approved by the Medical Sciences Inter-Divisional Research Ethics Committee (IDREC) of the University of Oxford (ref. number R49191/RE001). Participants were naïve to the purpose of the experiment. The scans took place at the Oxford Centre for Functional MRI of the Brain (FMRIB), Wellcome Centre for Integrative Neuroimaging.

Experimental structure

The experiment took place over three training and two scan sessions (Figure 1A). The three training sessions and the first scan session took place over four subsequent days. The second scanning session was scheduled between one and seven days after the first scan, based on participant and scanner availability. In the training sessions participants performed a set of tasks to learn the relational structures. Participants were free to do the training in their own time at home through the online testing platform Prolific (www.prolific.com). These tasks were run in custom-written JavaScript, CSS and HTML scripts. Because of problems with data storage on Prolific, for some subjects a few task blocks during the pre-training could not be recorded. However, all participants completed all tasks.

Stimuli

Thirty-one coloured and shaded object images that were similar in terms of familiarity and complexity were selected from the 'Snodgrass and Vanderwart 'Like' Objects' picture set (http://wiki.cnbc.cmu.edu/Objects, Rossion and Pourtois, 2004). For each subject, a subset of 17 objects was randomly chosen and assigned to the 17 nodes of the two graph structures (Figure 1B). The positions of the objects on the two graphs were randomised

such that pairwise distances between objects were orthogonal between graphs. The identity of the currently relevant graph was indicated by a blue or red background image. The colour associated with the graph learnt on the first day of training was randomly assigned and the alternative color was used for the second graph. Graphs one and two were initially learned separately on days one and two, and jointly on day three before the first scan.

Pre-training tasks

On each day, participants performed a set of computer-based tasks to learn the graph structures. In most of the tasks, participants could gain or lose coins depending on their performance. At the end of the experiment, the accumulated coins were converted to £ (every 100 coins = 1£ extra) and this money was paid in addition to the baseline payment of £7 per behavioural session. Visual feedback after each trial indicated whether a response was correct. On all days of the experiment, subjects were shown the number of correct trials and/or the number of coins collected at the end of each task.

To participate on subsequent days of the study, subjects were required to meet a minimum performance criterion of 400 coins collected on the first day of the experiment. This corresponds to about 70% of the total number of coins that could be obtained in the session. All trials and breaks between tasks were self-paced.

Tasks

The set of tasks was designed to teach participants the associations between pairs of objects and to familiarize them implicitly or explicitly with the object transitions in the graph (Supplementary Figure 1A).

Task 1 - Object pair learning

Subjects were presented twice with each one of the 24 object pairs (i.e. objects directly connected to each other by a link in the graph structure) and asked to memorize each pair (Supplementary Figure 1B).

Task 2 - Associative memory task

Subjects were presented twice with each object pair as well as with 48 pairs of objects they did not learn to associate with each other in a random order and asked whether the two objects were paired or not (Supplementary Figure 1C). Subjects gained or lost a coin for each correct or incorrect response, respectively.

Task 3 - Object orientation task

Subjects were exposed to object sequences generated from a random walk along the graph structure, such that each object would only be followed by a neighboring object on the graph structure (Garvert et al., 2017). Objects were presented in one of two orientations, which were mirror-images of each other (Supplementary Figure 1D). Subjects had to learn which orientation was associated with which button press. Thus, in this task the graphs were learnt in an implicit, rather than explicit manner. Subjects would gain or lose a coin for each correct or incorrect response, respectively. The task consisted of 100 trials.

Task 4 - Three-alternative forced choice task

Subjects were presented with one object at the top of the screen and had to pick the object it was paired with from three objects presented below (Supplementary Figure 1E). Each object was presented as many times as the number of objects it was linked to on the graph, such that each association was tested once. Subjects would gain or lose a coin for each correct or incorrect response, respectively.

Task 5 - Directed navigation task

In this task subjects navigated through the graph structure to find a given target object. The target object was displayed at the top of the screen. In the center of the screen the object corresponding to the current state were presented. Above this object, two cards were presented which were paired with this object. When choosing the next object, the object at the center was replaced by the selected object and the options were replaced by two new objects the chosen object was paired with (Supplementary Figure 1F). This means that by clicking through the objects, subjects moved from association to association until they reached the target. Below the objects, a progress bar and a number

of coins were presented. At the start of a trial, the number of coins corresponded to the distance on the graph between the starting and the target object. Selecting the object option closer to the target would move the bar by the distance corresponding to (the total length of the bar/minimum number of steps to reach target) to the right. Conversely, incorrect choices would move the progress bar to the left and a coin was subtracted from the amount that could be earnt on the trial. When subjects filled the progress bar, they got the number of coins displayed below. This means that the fewer steps it took a subject to reach the target, the more coins they gained. At the end of each trial, subjects were presented with the sequence of objects they had chosen to encourage learning. This task consisted of 30 trials.

Task 6 - Distance difference task

On each trial, participants were presented with a target object at the bottom of the screen and two goal objects presented on top of the target object. Participants were required to select the option that they would reach in fewer steps if they moved from association to association like in the directed navigation task from the target object (Supplementary Figure S1G). In other words, participants were required to select the option closer to the target object on the graph. Participants would gain or lose a coin for each correct or incorrect response, respectively. This task consisted of 30 trials.

fMRI experiment

Before the scanning session on day 4, participants performed a brief reminder training session consisting of one block of task 4 (three-alternative forced choice task) and one block of task 6 (goal navigation task). They were also familiarized with the task they would perform in the scanner through a brief pre-scan training session that was the same as the task performed in the scanner (15 trials).

In the scanner, subjects were presented with random object sequences. Each of the 17 objects was presented exactly four times per graph in each experimental block, resulting in a total number of 17 x 2 x 4 = 136 presented objects per block. There were 4 experimental blocks. 1 second before stimulus onset, the background colour changed

according to the relevant graph. Stimuli were then presented with the stimulus in context for 2 seconds, followed by a jittered inter-trial interval (ITI) generated from an exponential distribution with a mean of 1.7 seconds and truncated between 0.5 and 7.5 seconds (Figure 1C).

While observing the object sequences subjects performed a cover task of infrequently reporting by button press which one of two presented objects they would reach in fewer steps. This choice task occurred after random intervals exactly once after each object, i.e. in 12.5% of the total number of trials. Each correct button press was rewarded with £0.10 paid out in addition to the baseline fee to ensure that subjects attended to the stimuli. Subjects received a brief training on this task before they performed it in the scanner.

Arena task

At the end of the second scanning session, a subset of all participants (n=19) performed the Arena task separately for each context (Kriegeskorte and Mur, 2012). Here, all stimuli were initially displayed in random order outside a circular arena. Participants were instructed to arrange the objects inside the arena such that objects that were close to each other in the underlying graph structure were placed close together spatially. The item sets presented on each trial were adaptively selected using a “lift-the-weakest” heuristic, which optimizes subset construction based on previously collected judgments (Kriegeskorte and Mur, 2012). The resulting dissimilarity matrix, representing each participant’s perceived relationships between all items, was estimated by combining repeated judgments across pairs.

fMRI data acquisition and pre-processing

Visual stimuli were projected onto a screen via a computer monitor. Subjects indicated their choice using an MRI-compatible button box. MRI data were acquired using a 32-channel head coil on a 3 Tesla Magnetom Prisma scanner (Siemens, Erlangen, Germany). A T2*-weighted echo-planar imaging (EPI) sequence with a multi-band acceleration factor of 3 and within-plane acceleration factor (iPAT) of 2 was used to collect

72 transverse slices (interleaved order) with TR = 1.235 s, TE = 20 ms, flip angle = 65°, voxel resolution = 2x2x2 mm. Slices were tilted by 30˚ relative to the axial axis. The first five volumes of each block were discarded to allow for scanner equilibration. After two fMRI blocks, a T1-weighted anatomical scan with 1x1x1 mm resolution was acquired. In addition, a whole-brain field map with dual echo-time images (TE1 = 4.92 ms, TE2 = 7.38 ms, resolution 3x3x3 mm) was obtained to measure and later correct for geometric distortions due to susceptibility-induced field inhomogeneities. We performed slice time correction, corrected for signal bias, and realigned functional scans to the first volume in the sequence using a six-parameter rigid body transformation to correct for motion. Contrast images were later spatially normalized by warping subject-specific images to an MNI (Montreal Neurological Institute) reference brain and smoothed using a 6-mm full-width at half maximum Gaussian kernel. All preprocessing steps were performed with FSL (Jenkinson et al., 2012). Note that for the RSA analyses we used unsmoothed preprocessed data and later smoothed the resulting data RDMs (see below).

Behavioural data analysis

All data analysis of the training sessions, as well as of performance in the arena task after session 2, was performed using custom-written Matlab scripts (Mathworks, USA).

To obtain a metric for performance in the arena task, we first organized each subject's responses into pairwise distance matrices for each pair of objects. For visualisation purposes, we projected the matrices to a two-dimensional space using multi-dimensional scaling (MDS, Figure 2A-C bottom). Next, we correlated these distances with the "true" distance matrix, based on the number of shortest paths between object pairs (Figure 2D), as well as correlating the matrices of the two graphs with each other. To test for significance of these correlations across subjects we ran a one-sided t-test on the Fisher-transformed correlation coefficients.

Analysis of the behaviour in the choice trials during the scans was performed using MATLAB R2022b. We fitted a logistic mixed-effects model using the fitglme function, with subject included as a random effect. The dependent variable was the binary choice (left

vs. right option), and the independent variables were the differences in distance between the target object and the objects presented on each side (Figure 2G).

We then ran a linear mixed-effects model in MATLAB using the fitlme function, with subject included as a random effect. The dependent variable was the response times in the choice trials and the four independent variables were the distances between the target object and the 1) chosen object on the relevant graph; 2) unchosen object on the relevant graph; 3) chosen object on the irrelevant graph; 4) unchosen object on the irrelevant graph (Figure 2H). Next, we wanted to obtain subject-by-subject scores of the effects from this analysis, in order test for a correlation with the performance in the arena task (Figure 2I-L). We achieved this by running a separate linear regression for each subject, with the same dependent and independent variables as above.

fMRI data analysis: Repetition suppression

We modeled the onsets of objects separately for each graph and for *stay* and *switch* trials. Each regressor additionally contained two parametric modulators, one reflecting distance between the current object and the preceding object on the relevant graph and one reflecting the distance between the current object and the preceding object on the irrelevant graph. These two regressor were uncorrelated by design.

Contrasts were defined as [relevant] + [irrelevant] distances on *stay* trials (Figure 3) or as [relevant] distances on *switch* trials. The contrast images of all subjects from the first level were analyzed as a second-level random effects analysis. We report our results in the hippocampal–entorhinal formation, as this was our a priori ROI for the representations of the relevant and the irrelevant graphs, at a cluster-defining statistical threshold of $p<0.01$ uncorrected, combined with SVC for multiple comparisons (peak-level family-wise error [FWE] corrected at $p<0.05$). For the SVC procedure, we used two different anatomical masks. The first mask consisted of the entorhinal cortex and subiculum alone and was received with thanks from (Chadwick et al., 2015), (Supplemental Figure 2A). The second mask also contained other medial temporal lobe regions implicated in encoding physical space and comprised the hippocampus, entorhinal cortex, and parahippocampal cortex, as defined according to the maximum probability tissue labels

provided by Neuromorphometrics, Inc. (Supplemental Figure 2B). Activations in other brain regions were only considered significant at a cluster-defining threshold of $p<0.01$ uncorrected if they survived whole-brain FWE correction at the cluster level ($p<0.05$). While no areas survived this stringent correction for multiple comparisons, other regions are shown in Figures 3B, 4B and 4C at an uncorrected threshold of $p<0.01$ for completeness. While we used masks to correct for multiple comparisons in our ROI, all statistical parametric maps presented in the manuscript are unmasked. All the clusters used in the manuscripts for extraction of (independent) parameter estimates were obtained by thresholding at $t=2.5$.

To test whether suppression effects differed between the two graph structures, we extracted four condition estimates from the fMRI first-level models: graph 1 relevant, graph 2 relevant, graph 1 irrelevant, and graph 2 irrelevant. For each participant and session, the mean beta value within the MTL mask was obtained. Outliers exceeding ±3 SD within each session × condition were excluded. To ensure balanced data, only participants contributing all four conditions were retained. For each participant, we averaged the relevant and irrelevant conditions within each graph to obtain a single graph 1 and graph 2 suppression estimate. The resulting paired values were compared using a two-tailed paired t-test, testing the null hypothesis of no difference between the two graphs.

fMRI data analysis: Representational Similarity Analysis

Representational similarity analysis (Kriegeskorte et al., 2008) was performed on the cortical surface in subject space. We first ran a GLM where we modelled the presentation of each object in one of the graphs as a different condition (without any distinction between *stay* or *switch* trials). Note that here, unlike the repetition suppression analysis, we did not average the betas for each condition across experimental blocks, as we will use these beta maps to do RSA across blocks. We used Freesurfer on the T1 scans to create a grey matter surface map and then used a searchlight procedure (Kriegeskorte et al., 2006) to compute representational dissimilarities for each coordinate on the surface. The searchlight size was defined such that each coordinate reflected the

representational dissimilarity for up to 100 voxels within a 10 mm radius around it. Representational dissimilarity was computed as the correlation distances between voxel patterns after spatially whitening the data within the searchlight (i.e. (Walther et al., 2016)). To account for temporal autocorrelations and biases due to temporal proximity, we only correlated data from experimental blocks which are distant in time – i.e. corr([pattern from block 1], [patterns from block 3]), and corr([pattern from block 2], [patterns from block 4]), and averaged the two correlation coefficients to obtain a single 34 x 34 conditions RDM per subject per session. This data RDM was then related to each model RDM using Kendall's tau-a (Nili et al., 2014).

Each whole-brain map coordinate [x,y] thus contained the correlation of the brain's dissimilarity matrix with a model RDM when that coordinate is in the centre of the searchlight. To compute group statistics, we first smoothed the maps of each subject and session, and then moved them to the fsaverage space. Given our a-priori hypothesis (following the repetition suppression results regarding the emergence of the abstracted map across sessions), we then subtracted the map of session 1 from the map of session 2, and asked whether across participants the result was significantly larger than 0 using PALM (Permutation Analysis of Linear Models), a tool that allows inference on the surface through permutation methods (Winkler et al., 2014). We performed SVC FWE-correction (with cluster extent as the test statistic) within several masks derived by intersecting (orthogonal) effects of interest with an anatomical mPFC mask. The anatomical mPFC mask was the union of the Brodmann areas 24 and 32 masks from the PALS-B12 atlas (Van Essen, 2005)). The effects of interest masks were 1) the surface projection of the repetition suppression cross-graph effect (green mask in Figure 5B right); 2) the RSA effect of only the across-graphs RDM elements (Figure 5D); 3) the RSA effect of only the within-graph RDM elements (Figure 5C). Results are considered significant if they survive SVC FWE-correction at $p = 0.05$.

Acknowledgements

M.M.G. is supported by the Interdisciplinary Center for Clinical Research (IZKF) at the University of Würzburg (Project number S-533). T.B. is supported by a Wellcome Principal Research Fellowship (219525/Z/19/Z), a Wellcome Collaborator award (214314/Z/18/Z), a JS McDonnell Foundation award (JSMF220020372), and by the Jean Francois and Marie-Laure de Clermont Tonerre Foundation. The Wellcome Centre for Integrative Neuroimaging and Wellcome Centre for Human Neuroimaging are each supported by core funding from the Wellcome Trust (203139/Z/16/Z, 203147/Z/16/Z). The Sainsbury-Wellcome centre is supported by core funding from the Wellcome Trust (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755).

Author contributions

A.B., T.E.J.B. and M.M.G. conceived and designed the experiment and analyses. I.B., V.S. and M.M.G. collected the data. A.B. and M.M.G. conducted the analyses with support from H.N.. A.B. and M.M.G. wrote the manuscript with input from T.E.J.B. T.E.J.B. acquired the funding.

References

Badre, D. (2025). Cognitive Control. Annu. Rev. Psychol. *76*, 167–195.

Baldassano, C., Hasson, U., and Norman, K.A. (2018). Representation of Real-World Event Schemas during Narrative Perception. J. Neurosci. *38*, 9689–9699.

Bao, X., Gjorgieva, E., Shanahan, L.K., Howard, J.D., Kahnt, T., and Gottfried, J.A. (2019). Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. Neuron *102*, 1066-1075.e5.

Baram, A.B., Muller, T.H., Nili, H., Garvert, M.M., and Behrens, T.E.J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. Neuron *109*, 713-723.e7.

Barron, H.C., Garvert, M.M., and Behrens, T.E.J. (2016). Repetition suppression: A means to index neural representations using BOLD? Philos. Trans. R. Soc. B Biol. Sci. *371*, 20150355.

Barry, C., Ginzberg, L.L., O'Keefe, J., and Burgess, N. (2012). Grid cell firing patterns signal environmental novelty by expansion. Proc. Natl. Acad. Sci. *109*, 17687–17692.

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

Bein, O., and Niv, Y. (2025). Schemas, reinforcement learning and the medial prefrontal cortex. Nat. Rev. Neurosci. 2025 263 *26*, 141–157.

Bongioanni, A., Folloni, D., Verhagen, L., Sallet, J., Klein-Flügge, M.C., and Rushworth, M.F.S. (2021). Activation and disruption of a neural mechanism for novel choice in monkeys. Nat. 2021 5917849 *591*, 270–274.

Chadwick, M.J., Jolly, A.E.J., Amos, D.P., Hassabis, D., and Spiers, H.J. (2015). A goal direction signal in the human entorhinal/subicular region. Curr. Biol. *25*, 87–92.

Constantinescu, A.O., OReilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science (80-.). *352*, 1464–1468.

Doeller, C.F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. Nature *463*, 657–661.

El-Gaby, M., Harris, A.L., Whittington, J.C.R., Dorrell, W., Bhomick, A., Walton, M.E., Akam, T., and Behrens, T.E.J. (2024). A cellular basis for mapping behavioural structure. Nat. 2024 1–10.

Van Essen, D.C. (2005). A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex. Neuroimage *28*, 635–662.

Foster, D.J. (2017). Replay Comes of Age. Annu. Rev. Neurosci. *40*, 581–602.

Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., and Moser, E.I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. Nature *446*, 190–194.

Garvert, M.M., Dolan, R.J., and Behrens, T.E.J. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. Elife *6*, 1–20.

Gilboa, A., and Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. Trends Cogn. Sci. *21*, 618–631.

Grill-Spector, K., and Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychol. (Amst). *107*, 293–321.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E.I. (2005). Microstructure of a spatial map in the entorhinal cortex. Nature *436*, 801–806.

Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., and Woolrich, M. (2020). Replay bursts coincide with activation of the default mode and parietal alpha network. BioRxiv 1–20.

Jacobs, J., Weidemann, C.T., Miller, J.F., Solway, A., Burke, J.F., Wei, X.-X.X., Suthana, N., Sperling, M.R., Sharan, A.D., Fried, I., et al. (2013). Direct recordings of

grid-like neuronal activity in human spatial navigation. Nat. Neurosci. *16*, 1188–1190.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., and Smith, S.M. (2012). FSL. Neuroimage *62*, 782–790.

Jezek, K., Henriksen, E.J., Treves, A., Moser, E.I., and Moser, M.-B. (2011). Theta-paced flickering between place-cell maps in the hippocampus. Nature *478*, 246–249.

Kaefer, K., Nardin, M., Blahna, K., and Csicsvari, J. (2020). Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching. Neuron *106*, 154-165.e6.

van Kesteren, M.T.R., Beul, S.F., Takashima, A., Henson, R.N., Ruiter, D.J., and Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. Neuropsychologia *51*, 2352–2359.

Van Kesteren, M.T.R., Rijpkema, M., Ruiter, D.J., and Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. J. Neurosci. *30*, 15888–15894.

Kriegeskorte, N., and Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. Front. Psychol. *3*, 28167.

Kriegeskorte, N., Goebel, R., and Bandettini, P.A. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. *103*, 3863–3868.

Kriegeskorte, N., Mur, M., and Bandettini, P.A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

Kumaran, D., Summerfield, J.J., Hassabis, D., and Maguire, E.A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. Neuron *63*, 889–901.

Kumaran, D., Hassabis, D., and McClelland, J.L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. Trends Cogn. Sci. *20*, 512–534.

Liu, Y., Dolan, R.J., Kurth-Nelson, Z., and Behrens, T.E.J. (2019). Human Replay Spontaneously Reorganizes Experience. Cell.

Mack, M.L., Preston, A.R., and Love, B.C. (2020). Ventromedial prefrontal cortex compression during concept learning. Nat. Commun. 2020 111 *11*, 1–11.

Mark, S., Moran, R., Parr, T., Kennerley, S.W., and Behrens, T.E.J. (2020). Transferring structural knowledge across cognitive maps in humans and models. Nat. Commun. 2020 111 *11*, 1–12.

Mark, S., Schwartenbeck, P., Hahamy, A., Samborska, V., Baram, A.B., and Behrens, T.E. (2024). Flexible neural representations of abstract structural knowledge in the human Entorhinal Cortex. Elife *13*.

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

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

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Moneta, N., Garvert, M.M., Heekeren, H.R., and Schuck, N.W. (2023). Task state representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence. Nat. Commun. 2023 141 *14*, 1–21.

Morrissey, M.D., Insel, N., and Takehara-nishiuchi, K. (2017). Generalizable knowledge outweighs incidental details in prefrontal ensemble code over time. Elife *6*, 1–20.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. *10*, e1003553.

O'Reilly, R.C., and McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. Hippocampus *4*, 661–682.

Park, S.A., Miller, D.S., Nili, H., Ranganath, C., and Boorman, E.D. (2020). Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps. Neuron *107*, 1226-1238.e8.

Park, S.A., Miller, D.S., and Boorman, E.D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. Nat. Neurosci. 2021 249 *24*, 1292–1301.

Richards, B.A., Xia, F., Santoro, A., Husse, J., Woodin, M.A., Josselyn, S.A., and Frankland, P.W. (2014). Patterns across multiple memories are identified over time. Nat. Neurosci. *17*, 981–986.

Robertson, E.M. (2018). Memory instability as a gateway to generalization. PLOS Biol. *16*, e2004633.

Rossion, B., and Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. Perception *33*, 217–236.

Samborska, V., Butler, J.L., Walton, M.E., Behrens, T.E.J., and Akam, T. (2022). Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. Nat. Neurosci. 2022 2510 *25*, 1314–1326.

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., and Botvinick, M.M. (2013). Neural representations of events arise from temporal community structure. Nat. Neurosci. *16*, 486–492.

Schapiro, A.C., Turk-Browne, N.B., Norman, K.A., and Botvinick, M.M. (2016). Statistical learning of temporal community structure in the hippocampus. Hippocampus *26*, 3–8.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., and Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. Cell *186*, 4885-4897.e14.

Shepard, R.N. (1987). Toward a Universal Law of Generalization for Psychological Science. Science (80-.). *237*, 1317–1323.

Stickgold, R., and Walker, M.P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. Nat. Neurosci. 2013 162 *16*, 139–145.

Sun, W., Advani, M., Spruston, N., Saxe, A., and Fitzgerald, J.E. (2023). Organizing memories for generalization in complementary learning systems. Nat. Neurosci. *26*, 1438–1448.

Tacikowski, P., Kalender, G., Ciliberti, D., and Fried, I. (2024). Human hippocampal and entorhinal neurons encode the temporal structure of experience. Nat. 2024 1–8.

Tang, W., Shin, J.D., and Jadhav, S.P. (2023). Geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. Cell Rep. *42*.

Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208.

Tompary, A., and Davachi, L. (2017). Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex. Neuron *96*, 228-241.e5.

Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., and Morris, R.G.M. (2007). Schemas and Memory Consolidation. Science (80-.). *316*, 76–82.

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., and

Morris, R.G.M. (2011). Schema-Dependent Gene Activation and Memory Encoding in Neocortex. Science (80-.). *333*, 891–895.

Vaidya, A.R., and Badre, D. (2022). Abstract task representations for inference and control. Trends Cogn. Sci. *26*, 484–498.

Veselic, S., Muller, T.H., Gutierrez, E., Behrens, T.E.J., Hunt, L.T., Butler, J.L., and Kennerley, S.W. (2023). A cognitive map for value-guided choice in ventromedial prefrontal cortex. BioRxiv 2023.12.15.571895.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage *137*, 188–200.

Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. Cell *183*, 1249-1263.e23.

Whittington, J.C.R., McCaffary, D., Bakermans, J.J.W., and Behrens, T.E.J. (2022). How to build a cognitive map. Nat. Neurosci. 2022 2510 *25*, 1257–1272.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., and Nichols, T.E. (2014). Permutation inference for the general linear model. Neuroimage *92*, 381–397.

Yoon, K., Buice, M.A., Barry, C., Hayman, R., Burgess, N., and Fiete, I.R. (2013). Specific evidence of low-dimensional continuous attractor dynamics in grid cells. Nat. Neurosci. *16*, 1077–1084.

Yu, J.Y., Liu, D.F., Loback, A., Grossrubatscher, I., and Frank, L.M. (2018). Specific hippocampal representations are linked to generalized cortical representations in memory. Nat. Commun. 2018 91 *9*, 1–11.

Zeithamova, D., and Preston, A.R. (2010). Flexible Memories: Differential Roles for Medial Temporal Lobe and Prefrontal Cortex in Cross-Episode Binding. J. Neurosci. *30*,

14676–14684.