Title:

Neural representation dynamics reveal computational principles of cognitive task learning

Submission type: Article

Abbreviated title: Neural representation dynamics of learning

Authors: Ravi D. Mill[1], Michael W. Cole[1]
[1] Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ, 07102, USA

Corresponding author:
Ravi D. Mill
Center for Molecular and Behavioral Neuroscience, Rutgers University
197 University Avenue, Newark, NJ 07102
rdm146@rutgers.edu

Author contributions: RDM and MWC conceptualized the study design. RDM collected the dataset and analyzed it, under the supervision of MWC. RDM and MWC wrote the manuscript.

Conflict of interest: The authors declare no competing financial interests.

Acknowledgements: The authors acknowledge support by the US National Institutes of Health under award R01 MH109520, and support by the US National Science Foundation under award 2219323. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey, for providing access to the Amarel cluster and associated research computing resources that have contributed to the reported results. We would also like to thank Julia Hamilton, Emily Winfield, Doug Schultz, Richard Chen, Nicole Lalta, Kirsten Petersen and Kaustubh Kulkarni for assistance with piloting and data collection.

Abstract

During cognitive task learning, neural representations must be rapidly constructed for novel task performance, then optimized for robust practiced task performance. How the geometry of neural representations changes to enable this transition from novel to practiced performance remains unknown. We hypothesized that practice involves a shift from compositional representations (task-general activity patterns that can be flexibly reused across tasks) to conjunctive representations (task-specific activity patterns specialized for the current task). Functional MRI during learning of multiple complex tasks substantiated this dynamic shift from compositional to conjunctive representations, which was associated with reduced cross-task interference (via pattern separation) and behavioral improvement. Further, we found that conjunctions originated in subcortex (hippocampus and cerebellum) and slowly spread to cortex, extending multiple memory systems theories to encompass task representation learning. The formation of conjunctive representations hence serves as a computational signature of learning, reflecting cortical-subcortical dynamics that optimize task representations in the human brain.

Introduction

How are tasks represented in the human brain over the course of learning? Studies of cognitive task learning consistently report improvements in behavior as tasks are performed from initial novel presentation through extensive practice, which has been linked to a progression from controlled to automatic processing (Schneider & Chein, 2003; Schneider & Shiffrin, 1977). fMRI has revealed neural changes accompanying this progression, in the form of widespread decreases in the magnitude and extent of regional activations over practice, with some spatially limited increases (Chein & Schneider, 2005; Hampshire et al., 2019; Kami et al., 1995; Poldrack, 1998; Ruge & Wolfensteller, 2010). The activation decreases are most prominent in cognitive control-related regions, which also shift their patterns of functional connectivity as learning progresses (Bassett et al., 2015; Cole et al., 2010; Mohr et al., 2016). However, a precise computational basis for such task practice effects, accounting for how practice changes neural representations over learning, remains lacking. In the present report, we assimilated concepts from the long-term memory and context-dependent decision-making literatures to develop novel hypotheses regarding the neurocomputational mechanisms of task learning. These formalized roles for both subcortex and cortex (correcting over-emphasis on the latter in the task learning literature) in fulfilling the critical learning computations of pattern separation and the dynamic trade-off between compositional and conjunctive representations. We hypothesized that progressive replacement of compositional with conjunctive representations in cortex drives behavioral improvement over repeated task practice.

The foundations of these learning computations can be traced back to neuropsychological research describing deleterious effects of medial temporal lobe (MTL) damage on long-term memory function (Corkin, 1984; Eichenbaum, 2013; Scoville & Milner, 1957). Critically, preservation of certain types of memory following MTL damage, such as procedural motor skill learning (Brooks & Baddeley, 1976; Corkin, 1984; Squire, 1992) and recognition memory (Murray & Mishkin, 1986; Stark et al., 2019; Vargha-Khadem et al., 1997), highlighted that regions including but not limited to MTL subserve memory. This has sparked much theoretical interest through a broad class of multiple memory systems frameworks (Hirsh, 1980; Manns & Eichenbaum, 2006; Nadel et al., 2000; O'Reilly & Rudy, 2001; Sherry & Schacter, 1987; Simons & Spiers, 2003; Sutherland & Rudy, 1989). Much work remains in linking these theories to empirical evidence.

Key to this endeavor is formalizing the neural representations and computations that implement learning, and developing methods to quantify their operation in the functioning brain. Progress has been made in elucidating core neural representations of spatial memory, including place cells encoding an animal's position in their environment (Ekstrom et al., 2003; O'Keefe & Nadel, 1978), and grid cells mapping the environment with a grid-like representation (Doeller et al., 2010; Hafting et al., 2005). Non-spatial variants of these representational principles have also been theorized (McClelland et al., 1995; Wood et al., 1999). One influential idea has been the cognitive map (Tolman, 1948), which was popularized as a theoretical basis for representing spatial distances (O'Keefe & Nadel, 1978), and recently broadened to encode relationships/similarity amongst memorized concepts generally (Behrens et al., 2018; Eichenbaum & Cohen, 2014). In empirically grounding this abstract concept, pattern separation has been proposed as a key computational basis for constructing cognitive maps. This captures separation or reduced similarity of neural representations over learning to prevent interference between encoded information, and facilitate

recall from partial cues (Marr, 1971; O'Reilly & Mcclelland, 1994). MTL pattern separation has been supported empirically in animals and humans (Bakker et al., 2008; Berens et al., 2018; Brunec et al., 2020; Favila et al., 2016; Leutgeb et al., 2007). Despite theoretical compatibility, it remains unclear whether pattern separation serves as a general neurocomputation of memory that also influences cognitive task learning. Other uncertain aspects pertain to its domain-generality, given limited focus of prior work on spatial memory and visual recognition tasks. It is also unknown whether pattern separation emerges widely in the brain, given limited focus on MTL in animal lesioning studies, and restricted field-of-view fMRI studies seeking improved MTL spatial resolution at the expense of whole-brain coverage.

Certain multiple memory systems frameworks have incorporated contributions from cortical regions to pattern separation and related learning computations. Complementary learning systems theory (Kumaran et al., 2016; McClelland et al., 1995; O'Reilly et al., 2014; O'Reilly & Rudy, 2001) linked increasing pattern separation to another key memory computation – increasing nonlinear conjunctions of sensory features. A role for conjunctive representations in memory has long been conjectured with varying terminology: chunking (Wickelgren, 1979), configural associations (Marr, 1971; Sutherland & Rudy, 1989), binding (Ranganath, 2010; Shimamura & Wickens, 2009; Yonelinas et al., 2019), and relational representation (Eichenbaum, 2013). Whereas the majority of this work focused on conjunction formation in MTL, evidence of MTL-lesioned rodents being able to perform tasks explicitly requiring stimulus conjunctions (Whishaw & Tomie, 1991) prompted complementary learning systems proponents to concretize contributions of non-MTL regions. Subsequent updates to the theory (O'Reilly & Rudy, 2001) posited a division of labor, with MTL performing fast learning (via fast pattern separation and conjunctive formation) biased towards extracting specifics of experience, whereas cortex performs slow learning (slow pattern separation and conjunction formation) biased towards extracting generalities. Formalization of these core learning computations and the division of labor between subcortex/cortex in artificial neural network models accurately simulated behavioral findings from animal lesioning work (O'Reilly & Rudy, 2001). Critically, to our knowledge, the predictions of complementary learning systems theory have not been tested in empirical human brain data. This is largely a result of difficulty in developing analytic approaches that can quantify the engendered computations dynamically in the functioning brain as learning progresses.

Insight towards this goal can be gained from recent analytic innovations in the context-dependent decision-making literature, which has interrogated neural representational geometry underlying behavioral flexibility. Approaches inspired by machine learning and multivariate pattern analysis (e.g. dimensionality estimation and representational similarity analysis) have been applied to functional brain data to quantify the degree to which neural representations evoked during decision-making are context-specific (conjunctive) and/or context-general (abstract or compositional) (Badre et al., 2021; Bernardi et al., 2020; Flesch et al., 2022; Fusi et al., 2016; Ito et al., 2020; Kikumoto & Mayr, 2020; Mack et al., 2020; Mante et al., 2013; Rigotti et al., 2013). Compositional representations have been proposed to facilitate retrieval of memorized information and flexible recombination based on task context (Cole et al., 2011; Collins & Frank, 2013; Lake et al., 2015; McClelland & Rumelhart, 1985; Yang et al., 2019). This raises parallels with the distinction made in complementary learning systems theory between memory extracting specifics versus generalities of experience.

Whereas a role for compositionality in allowing for flexible transfer of knowledge to new contexts seems likely, nascent evidence linking within-trial emergence of conjunctive representations to accurate behavior (Bernardi et al., 2020; Kikumoto & Mayr, 2020; Rigotti et al., 2013) suggests a complementary role for conjunctions in subserving behavioral efficiency in familiar contexts. While these results illustrate the importance of conjunctive codes in guiding moment-to-moment decision-making, empirical study of the dynamics of these representations over the longer timescale of task learning, and how these trade off with compositional representations, has not been conducted. Methods capable of dynamically quantifying the relative contributions of conjunctive versus compositional representations also may illuminate how cortical-subcortical representational dynamics evolve over learning.

To this end, we hypothesized a task learning framework in which cortical-subcortical dynamics drive emergence of conjunctive versus compositional brain representations, yielding positive impacts on behavior over practice (Fig 1). This was tested empirically in the present report via fMRI recordings in healthy adult human participants as they practiced multiple complex cognitive tasks.

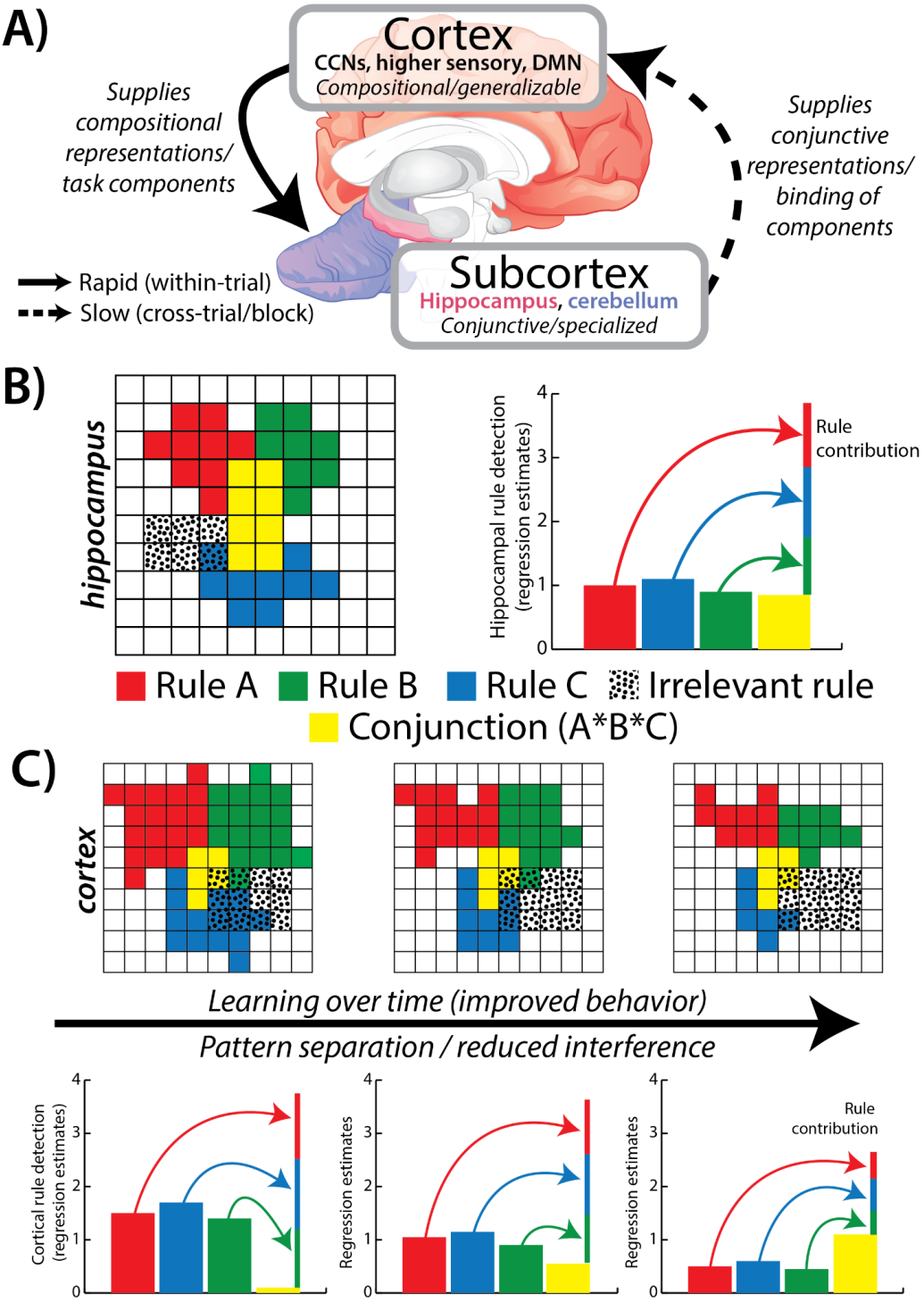

Fig 1. Hypothesized cortical-subcortical dynamics underlying changes to neural representational geometry over task learning, which drive behavioral optimization. A) Depiction in neural space. Early in learning, compositional representations of task rules are retrieved and reinstated in cortical networks (including cognitive control networks, CCNs, higher sensory networks and the default mode network, DMN), facilitating transfer of relevant pre-experiment knowledge to the current context (e.g. performance of C-PRO2 paradigm tasks, see Fig 2A for design). These compositional representations are rapidly bound by subcortex (hippocampus and cerebellum prominently) into more specialized conjunctive representations, and over practice propagated back to cortex where they gradually replace the earlier compositional representations to facilitate behavioral improvement. **B)** Depiction of hippocampal learning dynamics in representational space. Early in learning, the hippocampus binds compositional task rule representations into non-linear conjunctions,

which promotes encoding of a more distinct, specialized representation of the current task. The computational benefit of forming such conjunctive representations is to allow for a pattern separated representation of the task, which is more resistant to interference from irrelevant rules (depicted as the dotted pattern in the Figure). The left panel depicts this process as within-region activation patterns, and the right panel depicts this with reference to the machine learning approach we adopted to empirically quantify each rule type representation over learning (i.e. relative regression weights from fitting template rule types to a region's blockwise activation pattern; see Methods and Fig 5A). **C)** Equivalent activation (top panel) and regression space (bottom panel) depictions for cortical region learning dynamics. As tasks are repeatedly practiced, conjunctive codes increasingly dominate in cortical regions, and in parallel reliance on compositional codes is reduced. The computational benefit of these representational dynamics is to promote pattern separation of task codes in the brain, reducing interference with neural encoding of other tasks, and with a particular impact (compared to subcortex) in improving behavioral efficiency over extended learning.

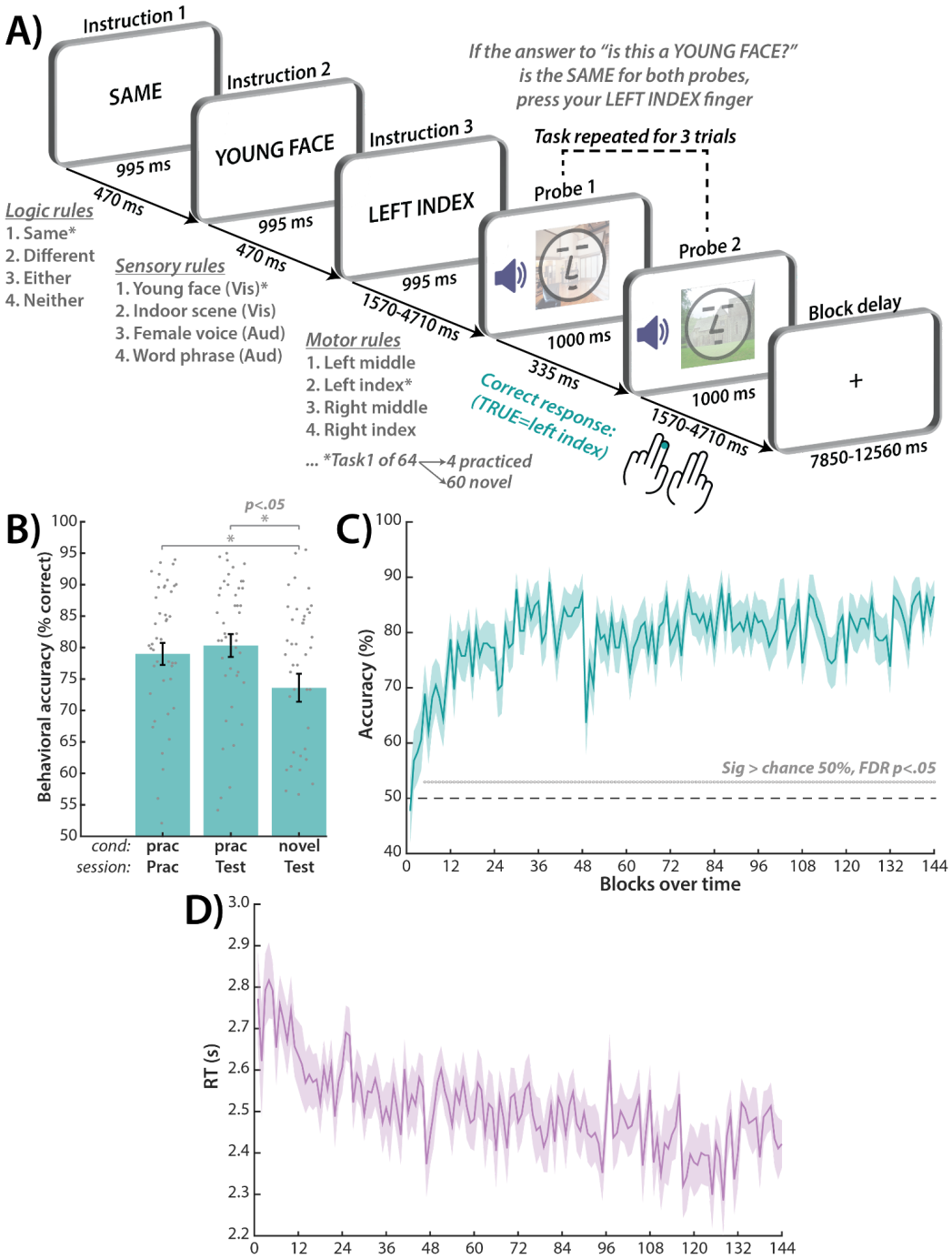

Fig 2. C-PRO2 task design and behavioral learning effects. A) Format for one block of the C-PRO2 task paradigm (see Methods for details), with an example trial and accompanying correct response depicted. This block format was preserved for all 64 tasks, across practiced/novel conditions and Practice/Test sessions. A task block began with three instruction screens (one for each of the logic, sensory and motor rule dimensions), which conveyed the task rules to be applied for the ensuing set of three trials. Each trial concurrently presented composite visual (varying across dimensions: young versus old face, and indoor versus outdoor scene) and composite auditory (varying across dimensions: female versus male voice, and word versus non-word content) stimuli over two sequential probe events. Subjects were required to integrate task instruction information and stimulus information across the probes to make their true/false categorization responses. Note that cartoon faces are used in the figure to conform with biorxiv's policy prohibiting photographs of real human faces in preprints. The actual experiment presented photographs of real human faces from an established database (Minear & Park, 2004; Ebner, 2008), which varied along the old-young age

dimension (see Methods for details). **B)** Session-averaged accuracy across key task conditions (practiced tasks presented in the Practice session, practiced tasks in the Test session, novel tasks in the Test session) revealed that subjects were reliably above chance performance, and that practiced tasks were better performed than novel ones (significance denoted by asterisks for paired t-tests, $p<.05$). Error bars represent standard error of the mean, with individual subject datapoints overlaid. **C)** Accuracy over task blocks in the Practice phase, showing expected learning effects: rise in accuracy over time (significantly above-chance by block 5 via one-sample t-test, significance denoted by gray dots at FDR $p<.05$) reaching asymptotic performance around 30 blocks. **D)** Reaction time (RT) over task blocks in the Practice phase, showing equivalent improvement in performance over learning (i.e. quickening of RT towards an asymptote). Patches in C) and D) reflect standard error of the mean.

Whilst the depicted framework takes considerable inspiration from prior multiple memory systems theories (particularly complementary learning systems theory), some differences between our framework and those theories are worth highlighting. Firstly, we go beyond prior theories' emphasis on conjunctions of sensory features to examine whether conjunctions also impact on learning of higher-order task rule information. To this end, we developed an updated version of the concrete permuted rule operations (C-PRO2) paradigm, which preserved the compositional (each task is created by permuting sensory, logic and motor rules) and multi-task (allowing for more generalizable inferences than single-task approaches) nature of previous versions (Cole et al., 2010; Cole, Reynolds, et al., 2013; Ito et al., 2017), but now with more complex, naturalistic and variable task features (see Methods and Fig 2A). Secondly, we conceptualize subcortical involvement in learning as not exclusive to the MTL, but rather also strongly involving the cerebellum, given its prior linkages with task learning (Beukema & Verstynen, 2018; Ferrari et al., 2018), development of behavioral automaticity (Chein & Schneider, 2005; Lang & Bastian, 2002; Nixon & Passingham, 2000), and flexible context-dependent processing generally (J.-H. Gao et al., 1996; King et al., 2019; Nakai & Nishimoto, 2022). A more comprehensive examination of cortical and subcortical involvement in task learning was enabled by our whole-brain fMRI approach, rather than a priori focusing on sub-selected regions-of-interest (e.g. MTL). Collectively, this combination of theoretical, task design and analytic innovations allowed us to empirically substantiate a neurocomputational signature of task learning (Fig 1), in which cortical-subcortical dynamics underpin a transition from compositional to conjunctive representations with practice.

Results

Behavioral evidence of task learning in novel C-PRO2 paradigm

We first verified that task learning had taken place as expected in our newly developed C-PRO2 paradigm (see Methods and Fig 2A for task details). Fig 2B plots the block-averaged accuracy across key conditions: practiced tasks presented in the Practice session (4 tasks repeated for 36 blocks each), practiced tasks in the Test session (same 4 tasks repeated for a further 15 blocks each), and novel tasks in the Test session (60 tasks presented just once). Subjects were reliably above chance accuracy (cross-task average=77.6%) in all three conditions (via one-sample t-tests against 50% chance, $p<.05$), which served as a sanity check that participants understood the paradigm. The cross-condition profile confirmed the influence of repeated practice on behavior: practiced tasks in

both the Practice (mean=79.0%) and Test (80.3%) sessions were more accurately performed than novel (73.6%) tasks, tested via paired t-tests (practiced task in Practice session > novel task in Test session, t=2.90, p=.006; practiced task in Test session > novel task in Test session, t=5.14, p<.001). Accuracy for practiced tasks in the Test session was numerically albeit non-significantly higher than practiced tasks in the Practice session (t=1.10, p=.276), and there were no reliable between-condition differences in RT (paired t-tests, all p>.450). Further evidence of learning came from timecourse analyses of performance accuracy (Fig 2C) and reaction time (RT; Fig 2D) over Practice session blocks. In both cases, behavior improved (accuracy increased and RT quickened) towards asymptotic levels over repeated practice (paired t-test for last block > first block improvement, accuracy t=6.27, p<.001; RT t=2.47, p=0.018). These results confirmed that the key task learning manipulation was successful, which validated the ensuing neural analyses.

Evidence for distributed pattern separation: neural task similarity decreases over the course of learning

We began by testing our hypothesis that pattern separation increases over task practice, consistent with a key computational role in task learning (Fig 1). This involved analyses of fMRI data examining the geometry of multivariate activity representations. As reported in the supplement (Fig S1), we first used a classification approach (4-way decoding of practiced task identity) to verify that task information was reliably decodable in our paradigm. We found that decodable information was distributed across multiple regions, albeit graded such that regions in cognitive control, motor and higher sensory networks were most prominent (consistent with previous reports; Cocuzza et al., 2020; Ito et al., 2017; Mill et al., 2022). This “distributed-but-graded” information profile has been observed previously across multiple neural data modalities (Kauvar et al., 2020; Mill et al., 2022; Siegel et al., 2015).

After verifying that the collected fMRI data contained decodable task information, subsequent analyses targeted the critical question of how the geometry of these representations changed over learning. We hypothesized that neural task representations would become increasingly distinct (i.e. dissimilar) from each other over practice, consistent with pattern separation encoding processes operating in the task learning domain (Fig 1). To test this, we calculated neural task similarity (pairwise Spearman’s correlation) between the 4 practiced tasks’ neural activations (Fig 3A), across Practice session blocks and across regions. Trends in the resulting regional timecourses were formalized by estimating the dynamic trajectory (regression slope over time, see Methods). The empirical trajectories in Fig 3B were predominantly negative, capturing a reduction in cross-task neural similarity over time, consistent with pattern separation (see Supplementary Table for a list of significant regions and accompanying statistics). Critically, this trajectory was highly general, encompassing medial temporal lobe regions typically linked to pattern separation (e.g. bilateral hippocampus), but also many cortical and subcortical regions (bilateral cerebellum, bilateral dorsal striatum and right amygdala) not as conventionally highlighted in the memory literature.

The regional trajectories were averaged into large-scale functional networks, based on region-to-network affiliations in the CAB-NP network parcellation (Fig 3C). Whilst all networks showed a significant negative trajectory (one-sample t-test, FDR p<.05), this was numerically strongest in the default mode network (DMN, trajectory t=-11.83, p<.001), cognitive control

networks (CON, t=-9.62, p<.001; DAN, t=-7.84, p<.001; FPN=-10.75, p<.001) and the higher-order sensory networks (Language, t=-9.77, p<.001; Visual 2, t=-7.71, p<.001;). The supplement (Fig S2) provides results of this approach at the whole-brain level (treating all 718 cortical/subcortical regions as features), which again revealed a negative similarity trajectory, and further highlighted that this held generally across the 12 task rule types.

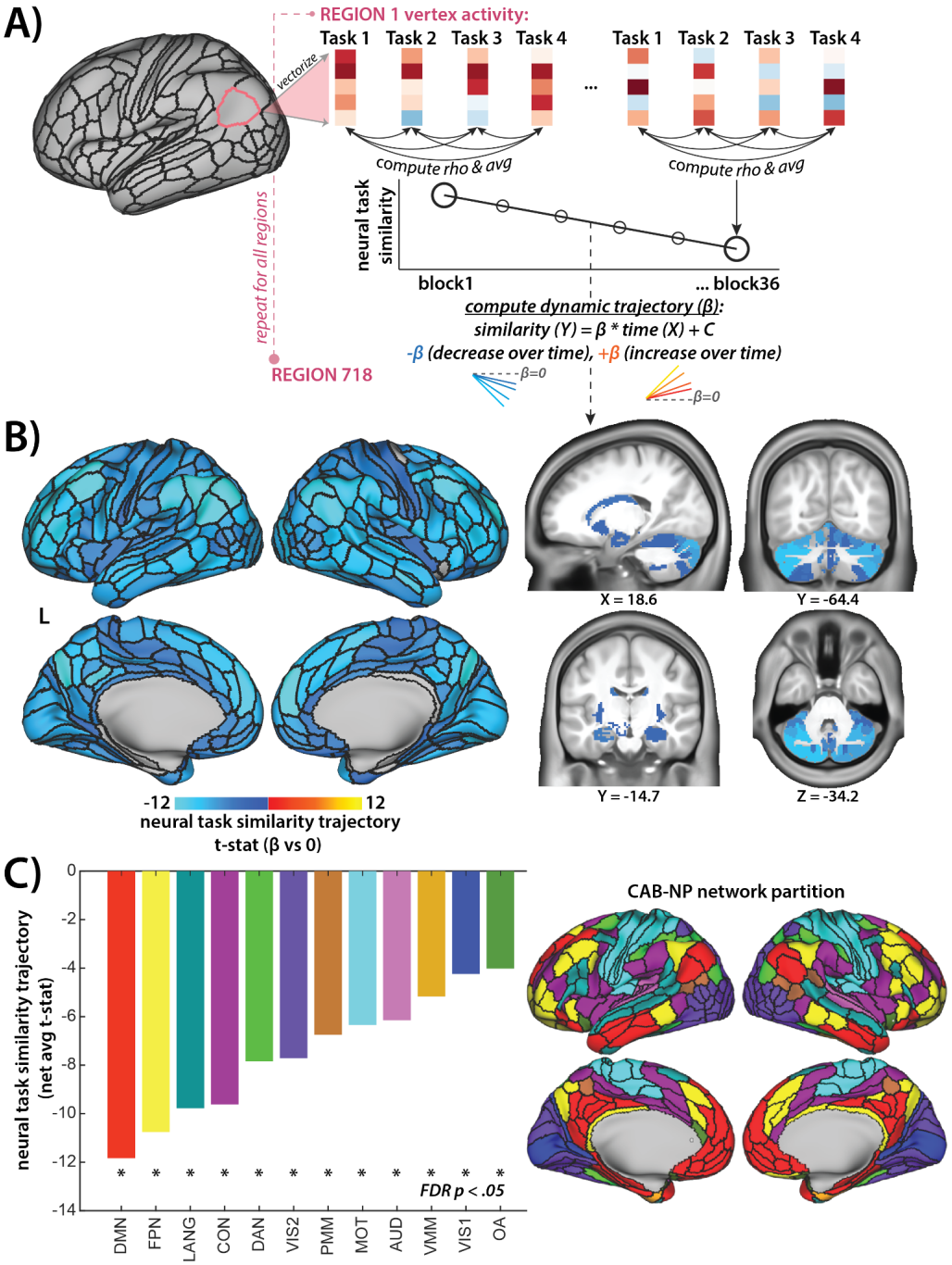

Fig 3. Learning leads to general reductions in neural task similarity over time, consistent with distributed pattern separation encoding processes. A) Schematic of analysis approach. For each region of interest, the similarity between task activation features (within-region vertices for cortex, and within-region voxels for subcortex) was computed using Spearman's correlation. The average across pairwise task similarities was

computed for each block, yielding a timecourse of neural task similarity over repeated practice. A regression approach then formalized the positive/negative trajectory over time, with this entire process repeated for all regions. **B)** Empirical dynamic trajectories (one-sample t-statistic for beta coefficients vs 0) plotted for each brain region, for cortex (left panel) and subcortex (right panel), thresholded at FDR p<.05. Multiple regions showed a significant negative trajectory, including regions typically linked to pattern separation (e.g. bilateral hippocampus), as well as multiple cortical/subcortical regions not as typically linked. **C)** Region-wise results in panel B were averaged into functional networks based on the CAB-NP network partition (depicted in the right panel). Whilst all networks showed a reliable negative trajectory (one-sample t-test, FDR p<.05), this was most prominent in the default mode network, cognitive control networks (DAN, FPN, CON) and higher-order sensory networks (Visual 2, Language).

The reduction in neural task similarity over practice was also associated with behavioral improvement, for both accuracy (reduced task similarity linked to greater accuracy, Fig 4A) and RT (reduced task similarity linked to reduced i.e. quicker RT, Fig 4B). The Supplementary Table lists all significant regions for this analysis and accompanying statistics. The network-averaged behavioral associations (left panels of Fig 4) converged across accuracy/RT on the CCNs as the numerically strongest network drivers of pattern separation processes that optimized behavior over learning (for accuracy: CON association t-stat=-5.19, DAN t=-3.73, FPN t=-4.11; for RT: CON t=7.38, DAN t=6.21, FPN t=6.86; all p<.001). The DMN (accuracy t=-3.28, p=.002; RT t=6.22, p<.001) and higher sensory networks (accuracy: Language t=-2.74, p=.009; Visual 2 t=-2.74, p=.005; RT: Language t=5.27, p<.001; Visual 2 t=5.66, p<.001) were also prominent.

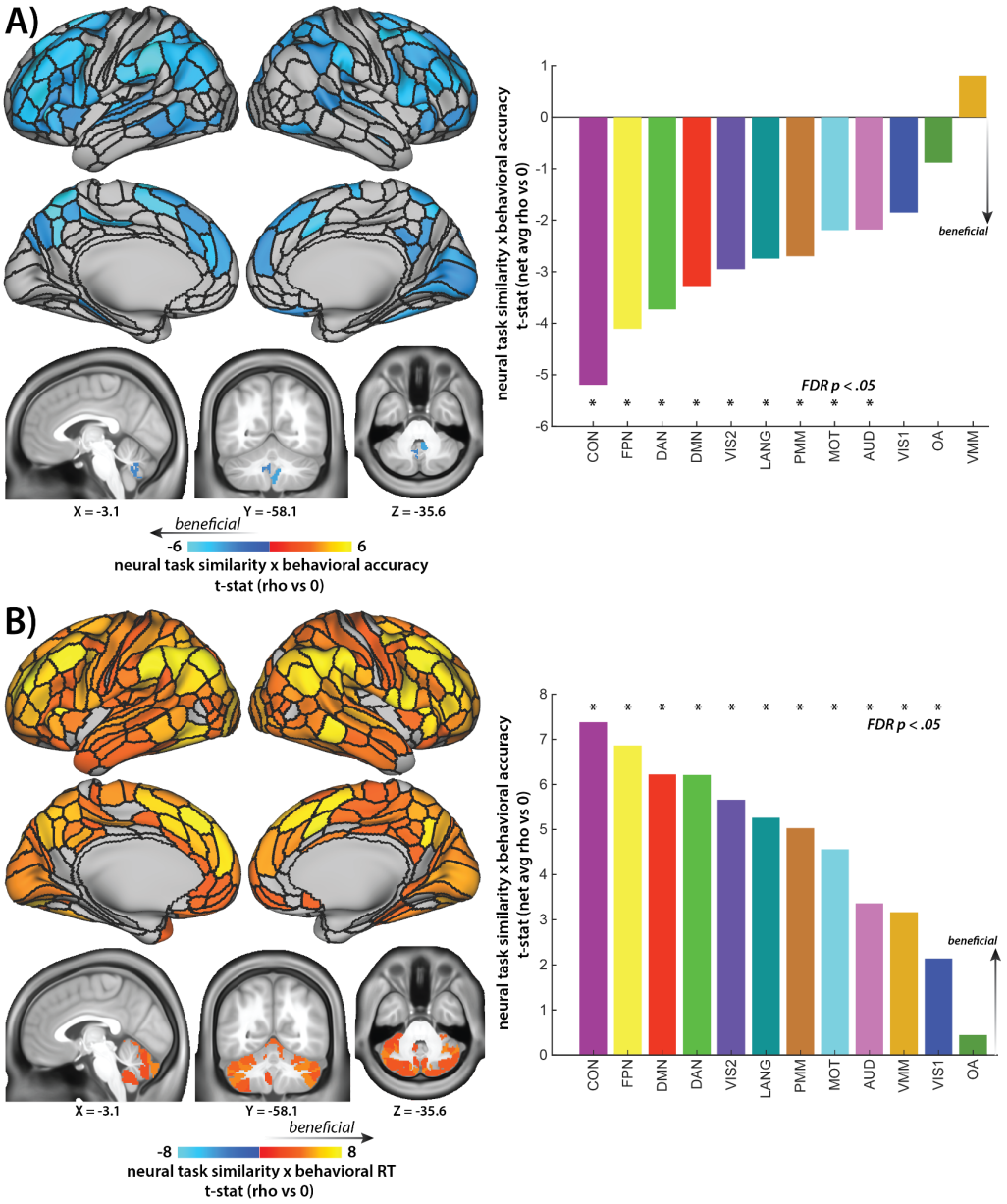

Fig 4. Degree of neural task similarity decrease (pattern separation) correlates with behavioral improvement over learning. A) Associations between regional neural task similarity and behavioral accuracy (% correct), computed via Spearman's correlation (one-sample t-stat for Fisher-transformed rho vs 0, at FDR p<.05). Negative values reflect beneficial associations i.e. as neural task similarity decreased over learning, behavioral accuracy increased. This was observed in multiple cortical regions and two subcortical regions in bilateral cerebellum. **B)** Equivalent analysis for RT, with positive values reflecting beneficial associations i.e. as neural task similarity decreased over learning, RT decreased. Multiple cortical regions reached significance, as did multiple subcortical regions (all in bilateral cerebellum). Network averages of the regional associations are presented to the right in both panels, revealing prominent loci for these behavioral influences in the CCNs, DMN and higher sensory networks.

Overall, the results reveal changes to neural representational geometry consistent with pattern separation encoding processes. These are shown here to operate at the higher-order level of cognitive task learning (departing from the focus on sensory pattern separation in the long-term memory literature), and to be more widely distributed across the brain than previously anticipated.

Pattern separation of neural task representations was also found to associate positively with behavioral task improvement, highlighting the functional relevance of these dynamics.

Quantifying conjunctive versus compositional rule representations during learning reveals a dissociation between subcortex and cortex

Preceding analyses revealed that neural task representations became more distinct over learning, but left unclear which specific aspect of the representations were changing to produce this. We hypothesized that the reduction in cross-task neural similarity over learning was driven by an increasing reliance on non-linear rule conjunctions over compositional single-rule representations (Fig 1).

To test this hypothesis, we developed a machine learning approach to quantify the relative strength of compositional and conjunctive representations on a dynamic block-to-block basis (see Fig 5A and Methods for details). Templates for the single rule types (sensory, logic and motor) and their non-linear conjunction were estimated using held-out Test session data, and then fit simultaneously via robust multiple linear regression to the regional practiced task activations, separately for each Practice session block. Test session data – which recombined rules into 64 unique task sets (4 practiced and 60 novel) – was essential in allowing us to separately estimate representation templates for each specific rule. The resulting beta coefficients for each practice session block quantified the relative contributions of compositional versus conjunctive rule types over time.

The overall (cross-block average) detectability of each rule type is presented in Fig 5B. Recovery of region peaks for the single rule types, i.e. visual and auditory regions for sensory rules, lateral prefrontal regions for logic rules, and motor cortex for motor rules (FDR $p<.05$; see Supplementary Table for a list of significant regions and accompanying statistics), was consistent with prior literature (Mill et al., 2022; Schultz et al., 2022) and served as sanity checks for the method. The results also revealed stronger representation of conjunctions in subcortex versus cortex, with compositions strongly represented across both. This dissociable spatial profile supports our hypothesized division of labor (Fig 1) between subcortex and cortex in respectively performing conjunctive and compositional neural computations during task learning.

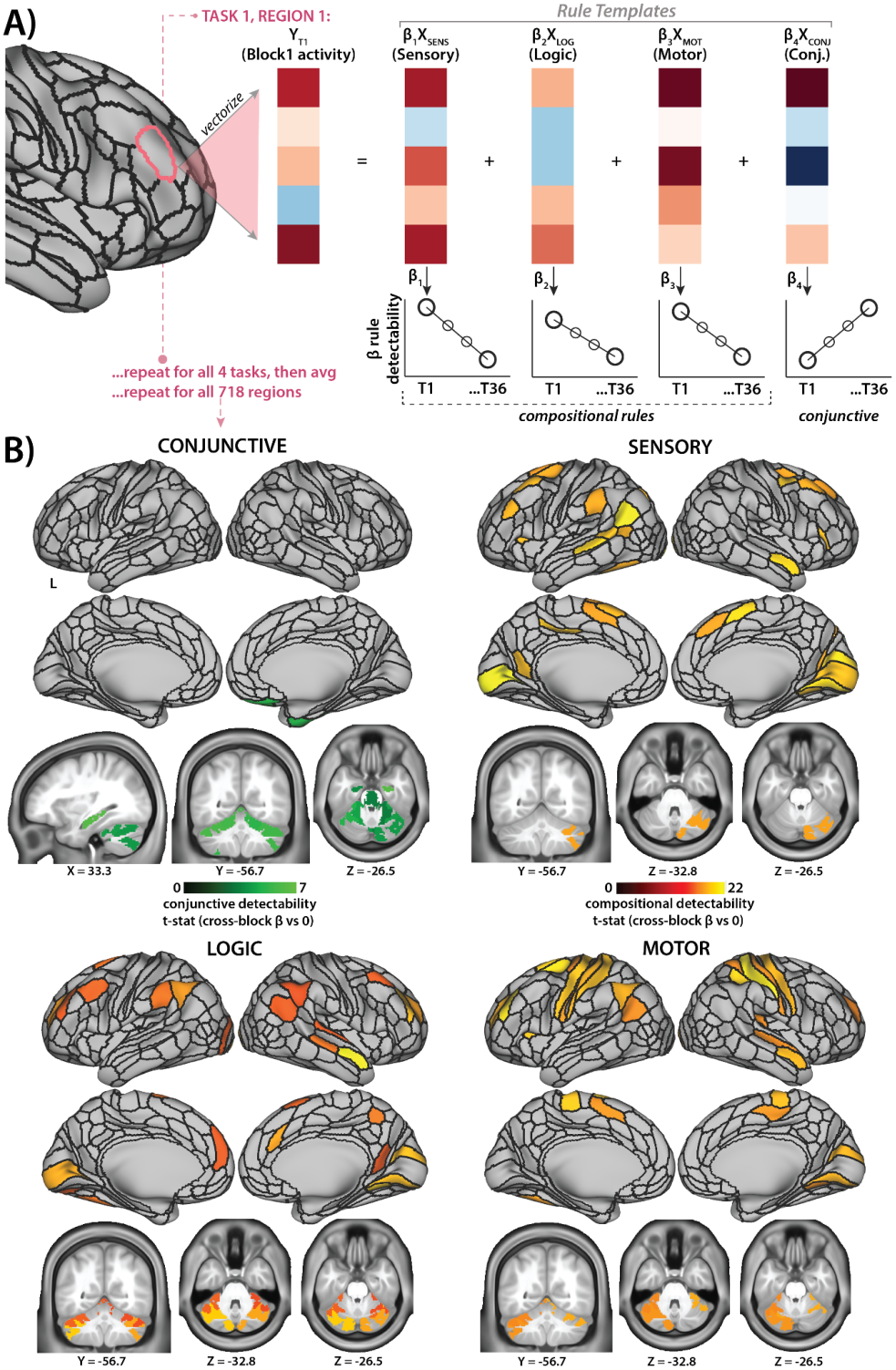

Fig 5. Rule type detectability analysis reveals differential sensitivity of subcortex and cortex to conjunctive and compositional codes. A) Schematic of our machine learning-inspired approach to detecting rule representation strengths over learning (see text for details). **B)** Results of applying the approach to quantify overall (cross-block) conjunctive (top left) and compositional (clockwise from top right) rule strengths, for cortex and subcortex. To ease visualization of region peaks, the compositional brain plots are thresholded as the top 5% t-stats, as many regions survived FDR p<.05 correction. The conjunctive plot yielded more sparse significance and is hence depicted at the standard FDR p<.05 threshold. Conjunctive codes were maximally detectable in subcortical regions (bilateral hippocampus, bilateral cerebellum, bilateral diencephalon, bilateral brainstem, left thalamus), whereas compositional codes were detectable in cortex as well as subcortex.

Bilateral cerebellar regions were most prominent in representing compositions (surviving the top 5% thresholding here), but multiple subcortical regions survived at FDR p<.05 (see supplementary information, Fig S3). Sensible peaks were obtained for each single rule type (e.g. bilateral motor cortex regions for the motor rule), validating the analytic approach.

Representational dynamics reveal progressive replacement of compositional with conjunctive representations in cortex as a neurocomputational signature of task learning

To further test dissociable roles for cortex and subcortex in task learning, we conducted dynamic analyses focusing on block-to-block variation of rule type detectability. We hypothesized that compositional representations in cortex would be slowly replaced with conjunctive representations (already present early in subcortex) with learning (Fig 1). To ensure sufficient power, these analyses were applied to network summaries of cortical regions (i.e. regional rule beta coefficients averaged into their affiliated CAB-NP networks), and 5 subcortical regions-of-interest that showed strong pattern encoding and conjunctive rule representation effects (see supplementary information for details, Table S1 and Fig S4).

The results are provided in Fig 6, with left columns depicting detectability timecourses across rule types, and right columns formalizing dynamic trajectories of the timecourse trends. The results support the expected cortex/subcortex dissociation based on their relative strength in representing compositional/conjunctive codes over task learning. Firstly, we observed significant subcortex > cortex conjunctive detectability, and significant cortex > subcortex compositional detectability across all blocks (all paired t-tests significant in the respective directions, for all rule types, FDR p<.05), confirming the separable representational sensitivities observed in the block-averaged results (Fig 5B). The dynamic trajectory analyses revealed further dissociation in the trajectories of cortical versus subcortical representational geometries over time. Whereas conjunctive codes progressively increased in cortex (cortical network average, trajectory t=2.70, p=.009), they progressively decreased in subcortex (subcortical region average, trajectory t=-2.77, p=.008). This difference in conjunctive code trajectories between cortex and cortex was significant (paired t-test cortex > subcortex trajectories, t=3.52, p=.001). Contrastingly, compositional codes progressively decreased across both cortex (mean across the 3 rule types, cortical network average, t=-4.22, p<.001) and subcortex (cross-rule subcortical region average, t=-3.78, p<.001). This dynamic profile is consistent with the hypothesized importance of strengthening conjunctive (versus compositional) representations over repeated task practice, derived from a multiple memory systems perspective of learning (Fig 1).

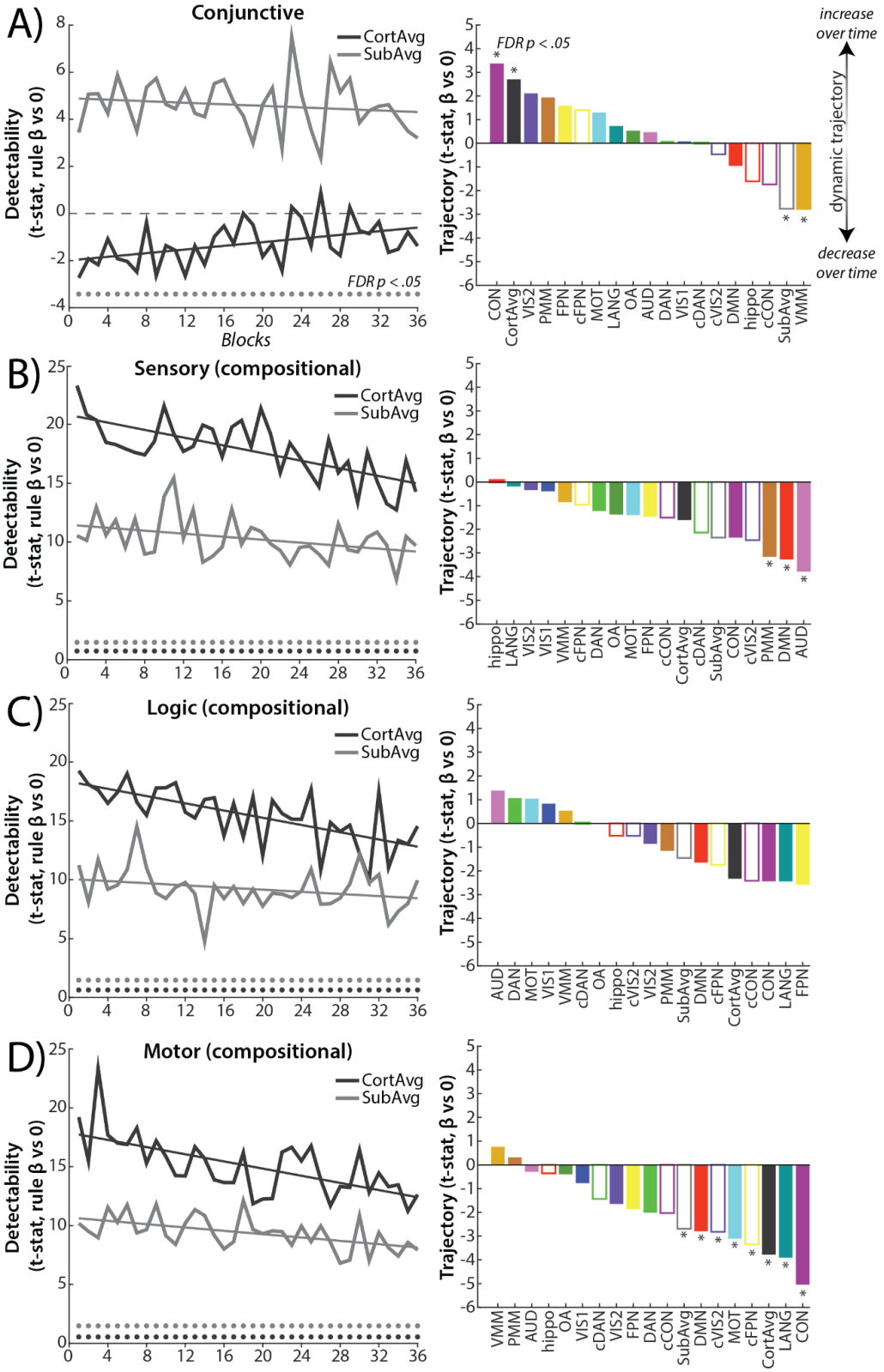

Fig 6. Replacement of compositional with conjunctive codes in cortex is a key representational signature of task learning. Left columns depict detectability timecourses (one-sample t-stats for rule beta coefficients vs 0, estimated per Fig 5A) for the different rule types. To simplify presentation, depicted timecourses are the average across cortical networks (black line) and the average across subcortical ROIs (gray line), with linear trend lines overlaid. Dots denote significant rule type detectability via one-sample t-test versus 0, FDR p<.05 (black=cortical, gray=subcortical). Right columns provide the dynamic trajectory analysis results for each

cortical network and subcortical region (and their averages), which formalize the timecourse trends over time. Cortical networks are presented as solid bars, and subcortical regions as outlined bars. Asterisks denote significant trajectories via one-sample t-test versus 0, FDR p<.05. **A)** Dynamics of conjunctive rule representations. Whereas conjunctive codes were reliably detected early in subcortex and decreased over time, such codes were not reliably detected in cortex but increased over time. The dynamic trajectory analysis revealed generally increasing conjunctive detectability in the cortical networks: 9/12 cortical networks were numerically positive, with the CON and the cortical network average reaching significance at FDR p<.05. Conversely, the trajectory was in the opposite direction for subcortex (4/5 regions were negative), with VMM (comprising perirhinal regions whose functions are intimately linked to MTL/subcortex, see Methods) and the subcortical region average showing a significant decrease over time. Panels **B-D)** provide the compositional (single rule) detectability results. Compositional codes were reliably present in cortex and subcortex, and emerged early (see significance dots in left panels). However, compositional codes were significantly stronger in cortex > subcortex; as reported in the main text, paired t-tests yielded significance in this direction for all single rule types, across all blocks (FDR p<.05). The dynamic trajectory analyses revealed a generally negative trend, with compositional codes decreasing over time in both cortex and subcortex. In interpreting the left panels, note that colored lines reflect one-sample t-tests contrasting the block-to-block rule beta estimates against 0, separately for cortex and subcortex (with significance denoted by colored dots). These are different from the paired t-tests reported in the main text, which contrasted the subcortex versus cortex rule beta estimates at each block.

We hypothesized that this dynamic shift from compositional to conjunctive representations contributes to the behavioral improvement that emerges with task practice. Accordingly, we probed the functional relevance of these dynamics, computing behavioral associations between the rule type detectability timecourses and block-to-block variation in reaction time. The results in Fig 7 demonstrate that greater conjunctive rule detectability in cortical networks (particularly the FPN; association t-stat=-4.11, p<.001) was linked to behavioral improvement (quickening RT) over task learning. Conversely, associations between compositional rule detectability and RT were generally in the harmful direction (slowing), for both cortex and subcortex, and especially for the sensory rule type. This pattern of findings is consistent with our hypothesis that conjunctive strengthening in cortex is the key computational signature of behavioral optimization over learning.

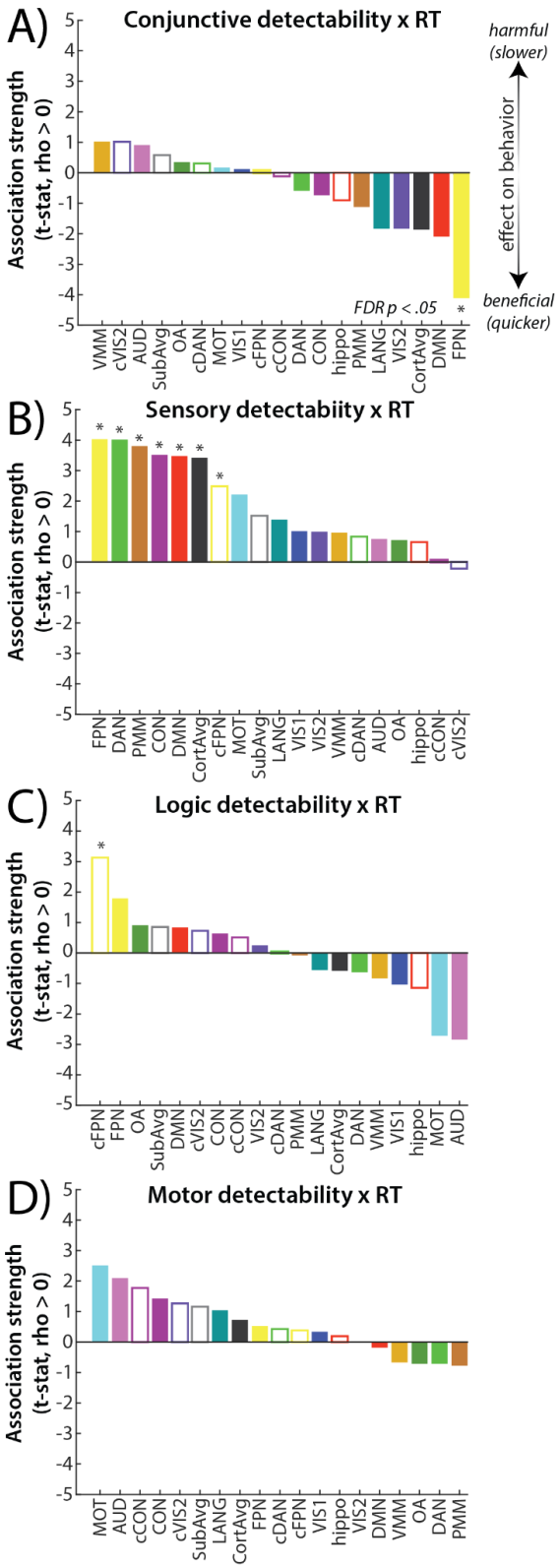

Fig 7. Strengthening of conjunctive codes in cortex over learning is associated with behavioral improvement (quickening of RT). Panels display the Spearman rank correlation between rule type detectability and behavioral RT. Conjunctive rule detectability was generally numerically associated with beneficial effects on behavior, reaching significance in the FPN cortical network (FDR p<.05). Conversely, greater compositional rule detectability over time generally yielded harmful behavioral associations (slowing of RT), with this effect most pronounced for sensory rules in the cognitive control networks (CCNs) and DMN.

Stronger conjunctive representations in cortex are associated with improved pattern separation
We hypothesized that one advantage of strengthening conjunctive (rather than compositional) representations with practice is to increase representational pattern separation, and thereby reduce cross-task interference. To test this computational hypothesis more directly, we interrogated whether pattern separation processes (quantified as neural task similarity decrease over learning; Fig 3-4) were associated with the detectability of conjunctive or compositional rule representations (Fig 5-7). The results in Fig 8 demonstrate that weakening compositional rule detectability over time was associated with improved pattern separation (reduced neural task similarity). This held across cortex/subcortex, and across the 3 single rule types (cross-rule and cross-cortex average, $t=36.86$; cross-rule and cross-subcortex average, $t=14.16$; both $p<.001$; see Supplementary Table for a list of significant regions and accompanying statistics). Decreasing conjunctive rule detectability in subcortex was also associated with improved pattern separation (cross-subcortex average, $t=4.62$, $p<.001$). However, conjunctions in cortex showed the opposite, beneficial association: increasing conjunction formation in cortex related to improved pattern separation (reduced neural task similarity; cross-cortex average, $t=-19.14$, $p<.001$). The overall spatial profile is again consistent with the hypothesized differentiation in function between cortex and subcortex, arising from their differential sensitivities to compositions and conjunctions as task learning progresses. Critically, the evidenced beneficial relationship between conjunctive coding and pattern separation was predicted by our multiple memory systems framework (Fig 1), and further supports conjunctive strengthening in cortex as a key neural signature of successful task learning.

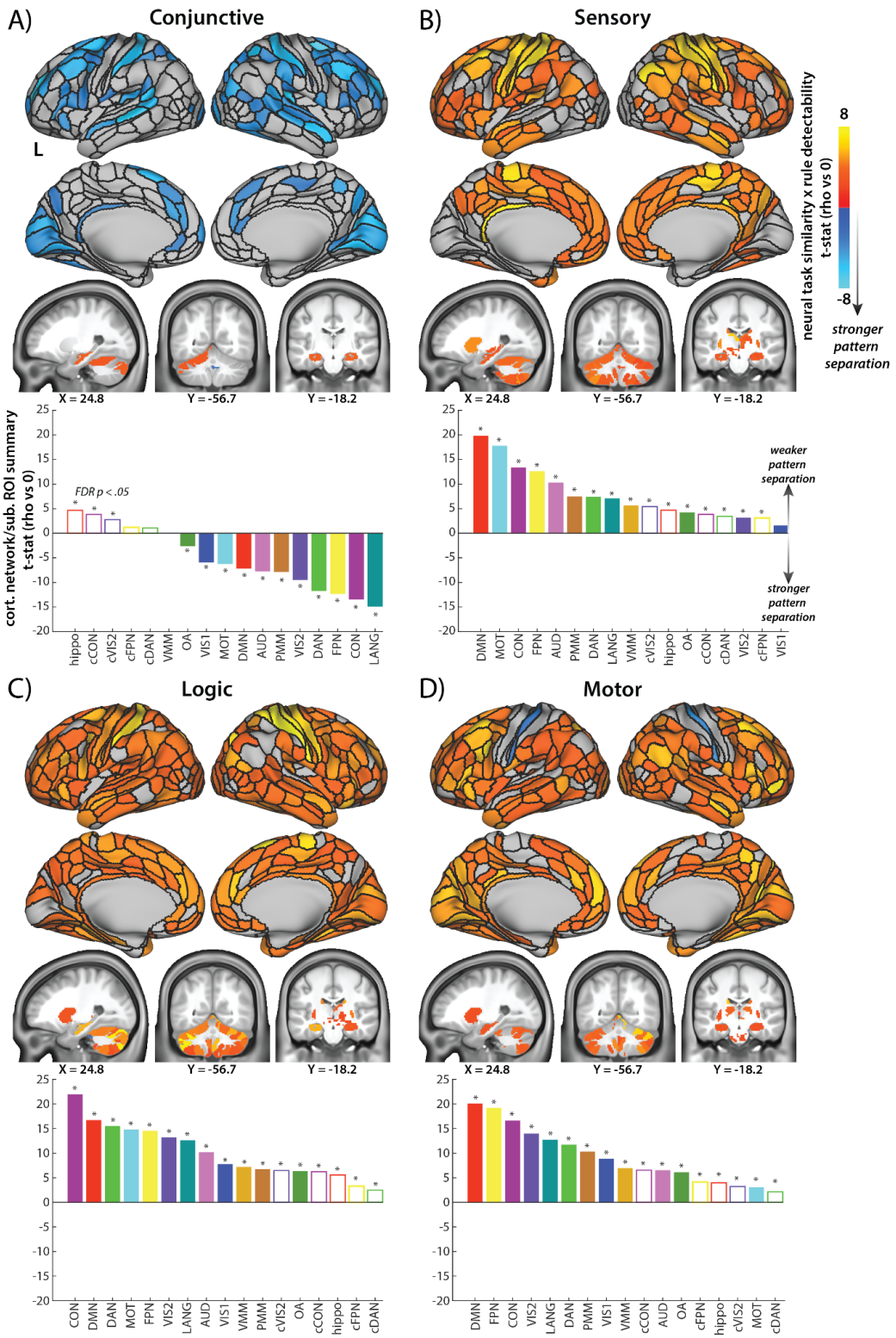

Fig 8. Stronger rule conjunctions in cortex are associated with improved pattern separation over task learning. Panels depict Spearman correlations between neural task similarity (our index of pattern separation, Fig 3A) and conjunctive (panel **A**) or compositional rule detectability (panels **B-D)**, thresholded at FDR p<.05. Panels plot regional results on the cortical surface (top row) and subcortical volume (middle row), as well as the average across cortical networks and subcortical ROIs (bottom row). Decreasing compositional rule detectability over time was associated with decreasing neural task similarity (improved pattern separation), and this association held across multiple regions in cortex and subcortex (including bilateral hippocampus, bilateral cerebellum, bilateral dorsal striatum, bilateral thalamus, right amygdala and bilateral brainstem). The

direction of association between pattern separation and conjunctive detectability varied for cortex versus subcortex. For subcortex, decreasing conjunctive representations were related to decreasing neural task similarity (improved pattern separation). For cortex, increasing conjunctive representations were beneficial in relating to decreasing neural task similarity (improved pattern separation). Significance in the network/regional average plots across panels (denoted by asterisks) highlighted that the effects were spatially distributed, albeit with overall greatest prominence in the CCNs, DMN and higher sensory networks.

Discussion

In the present report, we assimilated concepts from the separate neuroscientific subfields of long-term memory and context-dependent decision-making to test a framework in which cortical-subcortical representational dynamics drive task learning (Fig 1). This formalized key learning computations: pattern separation and the trade-off between compositional and conjunctive representations. We developed analytic approaches to quantify these computations in the functioning human brain as learning progressed in a novel paradigm, involving multi-task and multi-rule processing of complex stimuli (Fig 2A). The findings provide clear support for the framework's tenets: compositional representations form early in learning to facilitate transfer of prior (pre-experiment) knowledge to novel contexts, while progressive strengthening of cortical conjunctive codes drives increased pattern separation and behavioral optimization with practice (Fig 1). Replacement of context-independent compositional with context-specific conjunctive representations in cortex hence serves as a key representational transformation signifying successful task learning in the brain.

We hypothesized that as complex tasks are practiced, their neural representation undergoes pattern separation to become more distinct from other tasks. Supporting this hypothesis, we used a regression-based method to quantify the negative trajectory of cross-task representational similarity in brain regions, finding that the degree to which tasks become more dissimilar over time correlated with behavioral improvement. This supports the importance of pattern separation encoding processes previously theorized (Marr, 1971; O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001) and empirically studied in the long-term memory literature (Bakker et al., 2008; Brunec et al., 2020; Favila et al., 2016; Leutgeb et al., 2007). Critically, pattern separation emerged in distributed brain regions over learning, including but not limited to traditionally studied MTL regions. This aligns with emerging fMRI evidence from visual object recognition tasks suggesting that pattern separation processes are also performed by non-MTL regions (Amer & Davachi, 2023; Bowman et al., 2019; Liang et al., 2020; Nash et al., 2021; Pidgeon & Morcom, 2016). We extend the applicability of pattern separation not just spatially across the brain, but also from typically studied domains (spatial memory and visual object recognition) to the higher-order domain of task learning. Further, these computations are likely task-general, given the multi-task and multi-rule format of our C-PRO2 paradigm (i.e. pattern separation trajectories presented across variation in sensory, logic and motor rules).

Beyond showing that task representations become more dissimilar from each other over time, our findings also evidenced a more precise computational basis for such changes: increasing non-linear conjunction (binding) of each task's rule components. Conjunctions have long been considered an important neurocomputation of long-term memory (Eichenbaum, 2013; Marr, 1971; Shimamura & Wickens, 2009; Sutherland & Rudy, 1989; Wickelgren, 1979), with complementary learning systems theory explicitly linking conjunctions to pattern separation (O'Reilly & Rudy, 2001). Here we developed a machine learning-inspired approach to quantify the strength of conjunctive and compositional codes on a dynamic block-to-block basis, taking inspiration from recent analytic innovations in the context-dependent decision-making literature (Bernardi et al., 2020; Kikumoto & Mayr, 2020; Rigotti et al., 2013). Whilst this literature has demonstrated the important role of conjunctions in within-trial representational dynamics driving behavior, here we show their

importance in optimizing behavior at the longer timescale (minutes/hours) of block-to-block task learning. The present results emphasize conjunctions as a computational basis for behavioral efficiency with repeated task practice.

Conversely, compositional representations, previously highlighted as a basis for transferring knowledge flexibly to novel contexts (Bernardi et al., 2020; Cole et al., 2011; Cole, Laurent, et al., 2013; Collins & Frank, 2013; Lake et al., 2015; McClelland & Rumelhart, 1985; Yang et al., 2019), were prominent early in learning (when the C-PRO2 task paradigm was novel) and declined thereafter. Our findings add to emerging evidence across memory and decision-making subfields of a fundamental trade-off in neurocognitive function between encoding and retrieving generalities of experience (compositional coding), and encoding and retrieving specifics of experience (conjunctive coding) (Behrens et al., 2018; O'Reilly & Rudy, 2001; Reagh & Ranganath, 2023; Whittington et al., 2022). This trade-off provides a neurocomputational underpinning to the transition from controlled/novel to automatic/practiced processing, as studied behaviorally and neurally in the dedicated task learning literature (Chein & Schneider, 2005; Cole et al., 2010; Mohr et al., 2016; Schneider & Chein, 2003). The research avenues opened in the present report therefore have the potential for integrating memory and decision-making processes (typically studied in separate fields) under a unified neurocomputational model of learning.

Our findings also illustrated differential involvement of cortex and subcortex in the dynamic interplay between compositional and conjunctive codes over learning. In the presently studied timeframe (minutes/hours, relatively early in learning), cortex was found to more strongly represent compositions, whereas subcortex more strongly represented conjunctions. This substantiates the division of labor and associated computational roles posited by complementary learning systems theory (McClelland et al., 1995; O'Reilly et al., 2014; O'Reilly & Rudy, 2001): subcortical regions underpin fast learning with a bias towards extracting specifics, and cortical regions underpin slow learning with a bias towards extracting generalities. Elucidating representational dynamics was key in specifying the mechanisms underlying this division of labor: compositions decreased over learning in both cortex and subcortex, whereas conjunctions decreased in subcortex and increased in cortex. The increase in cortical conjunctive coding was also uniquely associated with behavioral improvement and greater pattern separation, highlighting it as a signature of successful task learning. This provides empirical support for the computational tenets of multiple memory systems theories, which were extended here to characterize learning in the higher-order task rule domain (versus prior application to spatial memory and visual object recognition), and to encompass a broader view of subcortex beyond MTL (particularly cerebellum, discussed in detail later).

Future work quantifying these computations in different domains of memory (e.g. episodic memory) and studying them over longer timescales (e.g. days/weeks of practice) will be necessary to clarify whether replacement of compositions with conjunctions in cortex is truly a general signature of learning. The time-intensive nature of cortical conjunction formation also suggests a computational recharacterization of an established pattern of memory deficits following MTL lesions (Corkin, 1984; Scoville & Milner, 1957). MTL damage does not cause retrograde amnesia, defined as an inability to retrieve old memories (Eichenbaum, 2013; McKenzie & Eichenbaum, 2011). Our findings suggest this arises from conjunctive representations (the putative computational basis for "old memories") having had sufficient time to migrate from MTL and connected subcortical regions to cortex, and

undergo extensive consolidation. This results in retrieval of old memories not requiring the MTL, and hence the lack of retrograde deficits following MTL damage. Compatibly, MTL damage does cause anterograde amnesia (inability to form new memories), reflecting the necessary involvement of MTL regions in forming early conjunctions when memories are newly encoded, which precedes the later process of conjunctions migrating to (and being consolidated in) cortex.

These computational dynamics might also have a role in established multi-tasking deficits (Musslick & Cohen, 2021). Beyond more widely appreciated "switch costs" of having to retrieve and apply task sets for multiple distinct tasks (resulting in interference; Sakai, 2008; Yeung et al., 2006), our findings also make the prediction that increased task switching disrupts the time-intensive (and behaviorally critical) process of learning cortical conjunctions. Per the framework delineated in Fig 1, failure to adequately strengthen cortical conjunctions under high task switching demands should also make tasks susceptible to representational interference. This contrasts learning under reduced task switching conditions, such as with sustained blocked task presentation formats, which allow enough uninterrupted time for cortical conjunctions to strengthen. This explanation aligns well with recent reports highlighting the benefits of blocked versus interleaved (rapid switching) task learning formats (Flesch et al., 2018). Such applications - recharacterizing established deficits following MTL damage, and as a computational underpinning for multi-tasking limitations - highlight the potential clinical and real-world relevance of cortical conjunctive replacement as a general neural signature of effective learning.

Beyond the clear differentiation between cortex and subcortex in representing compositions and conjunctions over learning, we also observed differentiable large-scale network involvement. Although pattern separation effects were distributed, the network-averaged results revealed a spatial gradation, with regions within the DMN, CCNs and higher-sensory networks showing the strongest similarity trajectories. Whilst the memory function of the DMN is increasingly well-established (Kaefer et al., 2022), a role for CCNs and sensory networks in pattern separation has recently been acknowledged (Amer & Davachi, 2023), building on earlier work linking these networks to general cognitive control functions (Cocuzza et al., 2020; Cole, Reynolds, et al., 2013; Mill et al., 2022), and control operations impacting long-term memory (Dobbins et al., 2002; Mill et al., 2015, 2016; Simons & Spiers, 2003). Future work will be necessary to elucidate the precise nature of interactions between the DMN, CCNs and sensory networks that give rise to the representational dynamics we observed (de-Wit et al., 2016; Ito et al., 2019). One possibility is that CCNs perform a goal-directed filtering of sensory information (via interactions with Visual2/Language networks) upstream of MTL and cerebellum (Amer & Davachi, 2023), to guide conjunction formation and pattern separation over learning. The importance of CCN interactions is supported by CCNs (and the FPN particularly) showing the strongest associations between conjunctive coding, pattern separation and behavioral improvement with practice. Separating the sub-functions of FPN and CON (a CCN also shown to be prominent in our findings) will be another important goal.

Towards targeting network interactions more formally, recent network modeling approaches that integrate empirical estimates of functional connectivity into generative models of function have been proposed (Cole et al., 2016; Hearne et al., 2021; Ito et al., 2022; Mill et al., 2022). Such methods allow for adjudication between candidate network models of task learning, which could illuminate the sequence of information flow across brain regions underlying the representational dynamics

observed here. This generative network modeling approach also has the potential to illuminate the mechanism of interaction between cortex and subcortex, and the degree to which they operate independently or dependently (synergistically), which has been the subject of much speculation (O'Reilly et al., 2014; Schapiro et al., 2017; Whittington et al., 2022).

Future extensions of the present study should also seek to differentiate the roles of MTL and cerebellum in task learning. Whereas hippocampus has been widely scrutinized in multiple memory systems frameworks, cerebellum has received less interest, despite being linked to relevant functions of procedural motor skill learning (Beukema & Verstynen, 2018; Ferrari et al., 2018) and behavioral automaticity (Lang & Bastian, 2002; Nixon & Passingham, 2000). Importantly, our results elucidate a precise computational basis for cerebellar contributions to learning, via early encoding of compositional and conjunctive codes. These computations are supported by cerebellum's dense bidirectional connectivity with MTL and cortical regions (Buckner, 2013; Z. Gao et al., 2018; Strick et al., 2009), and its recently highlighted role in multi-task cognition (J.-H. Gao et al., 1996; King et al., 2019; Nakai & Nishimoto, 2022). Our results suggest this multi-task involvement likely reflects a role in task learning computations. In considering differential roles within subcortex, one possibility is that MTL subserves the initial formation of conjunctive codes by referencing a cognitive map (Whittington et al., 2022), which then requires translation by cerebellum into a "forward model" (Hull, 2020) representational format that is more centered on action. This integrates the retrieved information from memory with current reward information to formulate an action plan that is propagated to cortex. Future work that manipulates variables hypothesized to influence forward model updating processes, such as reward and expectation, will be useful in probing differentiation between cerebellum and MTL. Separating between and within subcortical regions (e.g. hippocampal subfields) would also generally benefit from the improved spatial resolution of high field fMRI.

To conclude, we elucidated specific cortical and subcortical computations underlying changes to neural representational geometry that emerge over task learning. We believe the neurocomputational signature identified here – increased conjunction formation in cortex that drives increased pattern separation and improved behavior – serves as a general learning mechanism, synergizing insights from the separate subfields of context-dependent decision-making, long-term memory, and task learning. Future examination of the network underpinnings and generality of this mechanism has potential to advance fundamental understanding of learning in multiple contexts.

Methods

Subjects

Usable fMRI data were collected from 44 subjects (age mean=22.25 years, range=18-36; 23 self-reported female), out of a total of 46 (one subject was excluded due to a computer malfunction, and another for failing to use the response pad correctly). The sample was recruited from Rutgers University-Newark and neighboring communities, with the inclusion criteria of right-handedness, absence of current or past psychiatric or neurological illness, and absence of standard MRI contraindications. Prior to MRI data collection, a separate sample of 34 subjects (age mean=25.74 years, range 19-40; 18 self-reported female) was included in a behavioral study that validated the stimuli created for the task (see next section for details). All subjects gave informed written consent prior to participating, following ethical guidelines from the Rutgers University Institutional Review Board (IRB).

C-PRO2 stimulus creation and task design

Our aim of elucidating general neural principles of cognitive task learning informed substantial updating of the permuted rule operations (PRO) paradigm. Whereas previous versions used abstract word stimuli (Cole et al., 2010) or concrete multi-sensory stimuli (C-PRO; visual lines and auditory tones; Ito et al., 2017), the present concrete version (C-PRO2) used more complex, naturalistic and variable multi-sensory stimuli. The sensory dimensions of these stimuli (described below) were chosen based on prior literature suggesting they would elicit strong and spatially distinct neural representations (Epstein & Kanwisher, 1998; Formisano et al., 2008; Kanwisher et al., 1997; Lafer-Sousa et al., 2016).

Visual stimuli were sampled from established image databases, and varied on the following dimensions: young versus old faces (CAL/PAL face database; Minear & Park, 2004), and indoor versus outdoor scenes (SUN database; Xiao et al., 2010). The young (18-32 years of age) and old (64+ years) faces were taken from the Ebner (Ebner, 2008) update of the original CAL/PAL database, which sub-sampled to emotionally neutral expressions and color images, and additionally processed the images to remove eye-catching non-facial features (e.g. jewelry) and standardized low-level visual features (luminance and color, across background and clothing). To further reduce low-level feature variability, we used Adobe Photoshop to standardize image size and opacity (50%), for both faces and scenes. Faces were also cropped to remove hair, which could serve as an unwanted low-level visual cue for the young versus old face category judgment. The final composite visual stimuli consisted of faces overlaid centrally on scenes (Fig 2A), capitalizing on the reduced opacity of the constituent images.

Auditory stimuli were sampled from the ACCENT database (Weinberger, 2016), which contains audio recordings of the same 30-second short story read by multiple human speakers of varied gender, nationalities and accents. Our sampling was confined to US American accents to match the location of the study, with an equal proportion of normatively masculine (male) and normatively feminine (female) voices. The continuous speech recordings were then manually segmented into non-overlapping 3-word phrases using the PRAAT software (Boersma & Weenik, 2023). Further processing to control for low-level auditory feature variability included adding a 50 ms fade in/out

period at the start/end of each sound (to correct for acoustic clicks), equating the duration of all sound files to 800 ms (without affecting pitch, using PRAAT's PSOLA algorithm) and equating the intensity of all files to 75 dB. Reversed versions of each segmented audio file were also created to serve as the alternate response category. Hence, the final set of auditory stimuli varied across two dimensions: male versus female speaker, and word (lexical) vs non-word (non-lexical i.e. reversed) content.

To ensure that stimuli were reliably discriminable across all 4 sensory dimensions, a behavioral validation study was conducted on a preliminary set of 100 faces, 100 scenes, 100 word phrases and 100 non-word phrases (each with an equal proportion of relevant sub-categories e.g. 50 young and 50 old for the face condition). This study presented composite visual or composite auditory stimuli on individual trials, with subjects cued to categorize one dimension (e.g. "is the face in the image old or young?"), with the secondary dimension (in this case: scenes) randomly selected. Each stimulus was presented twice per subject, with categorization accuracy averaged across presentations and subjects. These data were first used to eliminate non-discriminable stimuli (any item yielding accuracy below 50% chance), and extreme high/low accuracy items were then iteratively eliminated from each of the 4 tasks to reduce accuracy differences between categories for each dimension. This procedure resulted in 78 stimuli for each of the 4 dimensions, which were reliably discriminable relative to chance (average accuracies across all stimulus categories were > 87%), and did not differ between categories (e.g. male voices were not significantly more discriminable than female voices; all between-category 2-sample t-tests non-significant at $p > 0.16$). Stimuli for the main study were created by permuting across these 78^4 possible stimulus compositions (pseudo-randomly at the group level to maximize stimulus variety), thereby constituting a highly complex and variable stimulus set that balances the greater naturalism of recent movie-watching studies (Sonkusare et al., 2019) with greater experimental control of simpler tasks. The goal of increasing stimulus variability was to increase the likelihood of results from this paradigm generalizing to a wide variety of contexts.

These stimuli were included as the updated sensory dimension in the full C-PRO2 task (Fig 2A). Following previous versions, C-PRO2 presents 64 unique tasks per subject, based on permuting logic, sensory and motor rule dimensions (each with 4 levels). For each subject, 4 tasks were repeatedly presented (practiced task condition), with the remaining 60 tasks presented just once (novel task condition). Each of the 12 rules were featured once across the 4 practiced tasks assigned to a given subject, to ensure that individual rules were equally practiced overall. Practiced tasks were counterbalanced across subjects, such that all 64 tasks were included as practiced tasks across the group. There were 16 practiced task combinations total (16 * 4 = 64), meaning practiced tasks were reused every 16 subjects. Pre-scan task instructions were kept minimal to preserve learning effects to be tracked in the scanner. The format of an example task block is depicted in Fig 2A, in which rules for the present task were cued via early instruction screens, followed by three trials in which subjects' categorization of the composite auditory-visual stimuli was determined by the task rule instructions. Feedback on performance accuracy was provided at the end of each 12-block run. The presentation order of the task rule instruction screens was counterbalanced across subjects, with a consistent order maintained for a given subject across Practice/Test sessions. Permuting this general task format across practiced/novel conditions, and across the logic/sensory/motor rule dimensions, thereby constituted a flexible multi-task paradigm. This firstly provided the necessary compositional task rule structure to test our hypotheses (probing the tradeoff between compositional and

conjunctive codes over learning), and also allowed for more generalizable inferences on the neural basis of task learning than standard single-task paradigms.

Experimental procedure

Subjects participated in three sessions, each separated by 1-7 days: fMRI Practice (task fMRI during repeated presentation of the 4 practiced C-PRO2 tasks, intermixed blockwise), fMRI Test (task fMRI during presentation of the 4 practiced tasks and the 60 novel tasks, intermixed blockwise), and EEG Test (EEG acquired during an identical repeat of the fMRI Test session). The third session involved dense-array EEG data recordings (using the EGI HydroCel Geodesic Sensory Net; 1000 Hz sampling rate, 256 sensors, Cz reference electrode), which were not analyzed in the present report. 15 minutes of resting-state data were also acquired in each session (pre-task for the fMRI sessions and post-task for the EEG session). The fMRI Practice session involved 12 runs of 12 blocks, amounting to 144 blocks total (each practiced task repeated for 36*4 blocks). The fMRI Test session involved 10 runs of 12 blocks, amounting to 120 blocks total (each practiced task repeated for 15*4 = 60 practiced blocks, and each novel task presented once = 60 novel blocks). Analyses in the present report primarily focused on the fMRI Practice session. The fMRI Test session was used here to derive templates for the rule type quantification analysis (see later Methods section). More extensive analysis of the fMRI Test and EEG Test sessions is planned for future studies. All task sessions were programmed and presented using E-Prime 2.

Note that in the Practice session, presentation of the 4 practiced tasks was intermixed blockwise, with 2 tasks assigned to a “switch” condition (~50% likelihood of switching between them within a 12-block run) and 2 tasks assigned to a “no switch” condition (each task presented alone within a 12-block run). To simplify the present analyses, we collapsed across switch/no switch when analyzing the practiced tasks, with deeper scrutiny of these conditions planned for future follow-ups. For the Test session, the first 12-block run always presented the 4 practiced tasks (with the intention of reactivating the learned representations from the Practice session), with novel tasks randomly intermixed with the practiced tasks in subsequent runs, and the last run always presenting novel tasks only (to ensure all were presented despite the surplus of novel compared to practiced tasks).

fMRI acquisition

Data were acquired by the 3T Siemens Trio MRI scanner housed at the Rutgers University Brain Imaging Center (RUBIC). Whilst all subjects underwent whole-brain multiband echo-planar imaging (EPI) sequences with a 32-channel head coil, enforced replacement of scanner hardware (specifically, the head coil) that occurred during data collection forced us to change the fMRI sequence parameters. Subjects 1-18 had the following sequence: TR = 785 ms, TE = 34.8 ms, flip angle = 55°, Bandwidth 1924/Hz/Px, in-plane FoV read = 208 mm, 72 slices in axial orientation, 2 mm isotropic voxels, multiband factor = 8. Subjects 19-20 had the following sequence: TR = 785 ms, TE = 37.2 ms, flip angle = 55°, Bandwidth 1924/Hz/Px, in-plane FoV read = 208 mm, 56 slices in axial orientation, 2 mm isotropic voxels with a 1mm slice gap, multiband factor = 8. Subjects 21-46 had the following sequence: TR = 785 ms, TE = 37.2 ms, flip angle = 55°, Bandwidth 1924/Hz/Px, in-plane FoV read = 224 mm, 64 slices in axial orientation, 2.2 mm isotropic voxels, multiband factor = 8. The same fMRI sequence was maintained across all rest/task acquisitions and across Practice/Test sessions for

a given subject. Throughout, the same anatomical T1 and T2 MRI scans were collected during the Test session (both with 0.8 mm isotropic voxels). Spin echo fieldmaps were acquired at the beginning of Practice and Test sessions to enable correction of distortions in the functional and structural images. The acquisition order for the Practice session was: fieldmaps, resting-state (15 mins), 12 runs of C-PRO2 (practiced tasks only, ~7 mins each). The order for the Test session was: fieldmaps, T1, T2, resting-state (15 mins), 10 runs of C-PRO2 (practiced and novel tasks, ~7 mins each).

fMRI preprocessing

fMRI data were preprocessed using the Human Connectome Project (HCP) minimal preprocessing pipeline, which has been described in detail previously (Glasser et al., 2013). Briefly, the pipeline performs reconstruction and segmentation of the anatomical images, followed by functional image distortion correction (using the acquired fieldmaps), motion realignment, volume-to-surface mapping, spatial normalization to a standard MNI surface template and cross-subject surface alignment using multimodal structural/functional MRI information (MSM-All; Glasser et al., 2016).

The HCP pipeline outputs preprocessed CIFTI grayordinate fMRI timeseries (comprising ~60k 2-D cortical surface vertices, and ~30k 3-D subcortical voxels) in standard MNI space, which were then submitted to a general linear model (GLM). Separate GLMs were estimated for the fMRI Practice and fMRI Test sessions. Following best practice recommendations for fMRI artifact removal (Ciric et al., 2017), nuisance regressors included in the GLM were the 6 standard motion parameters, and physiological regressors specified per the aCompCor procedure (first 5 principal components estimated within white matter and ventricular masks; Behzadi et al., 2007). Derivatives and quadratics for the 6 motion and 10 physiological regressors were also included, meaning that a total of 64 nuisance regressors were modeled. For both sessions, separate task regressors for each presented block were modeled, as boxcars convolved with the SPM canonical hemodynamic response function, spanning the onset of the first instruction screen to the offset of the last trial for that block. For each vertex/voxel, this yielded 144 task beta activation estimates spanning presentation of the practiced tasks across the Practice session, and 120 task beta activation estimates for presentation of the practiced and novel tasks in the Test session.

Parcellation into functional regions and networks

Our general analysis approach required that the vertex-/voxel-wise data be linked to larger regions of interest, so as to apply multivariate approaches that treated within-region vertices/voxels as features. This was intended to approximate application of multivariate decoding approaches to neural populations in animal multi-unit electrophysiology research (Saxena & Cunningham, 2019; Siegel et al., 2015), to the extent feasible with the coarser spatial scale of human fMRI. To enable this, we linked the cortical vertex activations to 360 regions of interest defined by the Glasser surface atlas (Glasser et al., 2016). The subcortical voxel activations were linked to 358 subcortical regions of interest using the CAB-NP functional network partition (Ji et al., 2019). The use of these established regional atlases allowed for a principled means to define functionally relevant regions, with likely better generalizability given that these definitions were established in previous held-out data. This also fulfilled the need to include parcellations of both cortex and subcortex in our analyses, consistent with our goal of examining cortical-subcortical dynamics of task learning, and contrary to

general trends in memory and context-dependent decision-making subfields of excluding cortex and subcortex respectively. The CAB-NP also provides 12 functional network affiliations for all 718 regions (see Fig 3), which enabled further inferences of how task learning representational dynamics emerged at the level of large-scale networks. The expanded names of network labels provided in figures are as follows: AUD (primary auditory), CON (cingulo-opercular network), DAN (dorsal attention network), DMN (default mode network), FPN (fronto-parietal network), LANG (language i.e. higher auditory network), MOT (motor), OA (orbito-affective), PMM (posterior multi-modal), VIS1 (primary visual), VIS2 (higher visual), VMM (ventral multi-modal).

Estimation of neural task similarity (pattern separation measure)

To test our pattern separation hypotheses, we used an approach inspired by representational similarity analysis (RSA; Diedrichsen & Kriegeskorte, 2017; Haynes, 2015) which we term neural task similarity. The approach is schematized in Fig 3A. For a given region, the Spearman rank correlation was computed between the vectorized activations of its within-region features (vertices for cortex, voxels for subcortex), across all 4 practiced tasks. This yielded 6 pairwise rho values for each block, which were averaged to provide a summary similarity value for that block. Repeating over blocks populated the neural task similarity timecourse for that region, which was used to estimate a formal dynamic trajectory over time (see below), and was also used to examine associations with task behavior (see below).

Quantification of compositional versus conjunctive rule type detectability

To interrogate more precise changes in neural representational geometry that occurred during task learning, we developed a machine learning-inspired approach to quantify the relative contributions of compositional (single rule) versus conjunctive rule types over time (Fig 5A). This took inspiration from other regression-based approaches developed in the field of context-dependent decision-making (Kikumoto & Mayr, 2020). For a to-be-analyzed practice task (in the Practice session), we created representational templates for each constituent rule type (sensory, logic, motor) by averaging over relevant blockwise activations in the held-out Test session. For example, if the practiced task contained the “YOUNG FACE” rule (Fig 2A), a sensory rule template for that task would be created by averaging over all instances of that rule across novel tasks presented in the Test session. Note that availability of the Test session was critical in creating these rule templates, as each rule was assigned uniquely across the 4 practiced tasks, which prevented decomposition of practiced tasks into component rule representations. A conjunctive template was created by multiplying the 3 single rule templates together, akin to a non-linear interaction term. We then fit a multiple linear regression model with all 4 template terms (spatial regressors) as predictors to the regional task activation vector for a given block:

$$Y_t = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 * X_2 * X_3 + C$$

where Y_t is the activation vector for the practiced task on that block, $X_{1\text{-}3}$ are the single rule templates for sensory, logic and motor rules, and $X_1*X_2*X_3$ is the rule template for the conjunctive rule (non-linear interaction term). Coefficients $\beta_{1\text{-}4}$ therefore quantify the strength of each rule type in the activity representation evoked in that block. We used robust regression to estimate the beta

coefficients, given its improved resilience against outliers compared to regular least squares regression (Holland & Welsch, 1977). This was implemented via Matlab's "robustfit" function, with default parameters (bisquare weight function, tuning constant=4.69). The resulting beta coefficients captured the unique detectability of each rule type during each block. Repeating over blocks and practiced tasks yielded timecourses of rule type detectability for each subject. The timecourses were averaged across the 4 practiced tasks, with the process repeated for each region to derive subject-level rule type detectability estimates.

Note that for the overall rule type decodability results in Fig 5B, we averaged the subject-level rule estimates across time (blocks), before conducting group-level random effects statistics. These results informed selection of subcortical regions-of-interest (ROIs) for ensuing analyses that preserved the block-to-block dynamics (Fig 6-8). Selected ROIs were subcortical regions showing the strongest pattern separation and conjunctive decodability effects, with regional significance > 0 in both the neural task similarity (Fig 3B) and overall conjunctive decodability (Fig 5B) analyses (both FDR $p<.05$). Note that we incorporated an a priori theoretical focus on hippocampus and cerebellar regions, along with effects in the independent pattern separation analysis, to offset any potential circularity introduced by selecting ROIs on the basis of the block-average conjunctive analysis. The supplement (Table S1 and Fig S4) details the 5 subcortical ROIs selected, with accompanying region and network information taken from the CAB-NP brain parcellation. Identifying subcortical ROIs ensured sufficient signal-to-noise and statistical power for the analyses of representational dynamics. These criteria also informed the decision to confine cortical analyses in Fig 6 and 7 to the average for each of the 12 CAB-NP networks. Note that the cortical network averages excluded all subcortical regions affiliated for each network, to ensure clean separation between cortical and subcortical contributions. The averages across cortical networks and subcortical regions were also included to aid timecourse visualization (versus presenting 19 timecourses in a single plot) in Fig 6. Note that the ventromedial multimodal (VMM) network was included in the subcortical average, given that regions within this network (i.e. bilateral perirhinal cortex) are intimately linked to the function of MTL regions in subcortex (O'Reilly et al., 2014; Tomás Pereira et al., 2016).

For the final analysis (Fig 8) interrogating the association between pattern separation (neural task similarity) and rule type detectability, Spearman's correlation was computed between the block-wise neural task similarity timecourse and the 4 rule detectability timecourses (each averaged across the 4 practiced tasks). The resulting subject-level rho values were submitted to group random effects statistics, at both the regional and network-averaged levels.

Dynamic trajectory analysis

For both the neural task similarity and rule type detectability analyses, trends over time in the observed block-to-block timecourses were quantified using an approach we term dynamic trajectory analysis (inspired by prior work; Mill et al., 2016). Rather than trusting visual impressions of timecourse trends, this method provided a means to statistically formalize them, with the added benefit of providing an easily presentable regional summary (versus presenting timecourses for all 718 regions). As depicted in Fig 3A, this involved fitting a simple linear regression:

$$Y = \beta X + C$$

where Y is the neural measure (neural task similarity or compositional/conjunctive rule type detectability), X is time (block 1-36), C is a constant, and β is a coefficient capturing the linear trajectory over time (positive=increasing, negative=decreasing). Robust regression was again used to ensure improved resilience of the estimated coefficients to noise/outliers in the timecourses (Holland & Welsch, 1977). Dynamic trajectories were estimated for each individual subjects' neural timecourses, and then submitted to group random effects statistics to assess significance.

Behavioral analyses

Standalone analyses of behavior focused on performance accuracy (% correct) and reaction time (RT), averaged across the 3 trials presented within a task block. To validate our practiced versus novel task manipulation, cross-block-averaged accuracy was contrasted pairwise between 3 task conditions (practiced tasks in the Practice session, practiced tasks in the Test session and novel tasks in the Test session) via paired sample t-tests (see Fig 2B). We also plotted behavioral timecourses across chronologically presented blocks for accuracy (Fig 2C) and RT (Fig 2D) to verify recovery of standard learning effects (rise in performance towards an asymptote). For analyses that contrasted behavior against chance performance accuracy (Fig 2B and Fig 2C), even though chance was technically 25% (given that 4 button responses were available, 2 for each hand), we considered 50% (2 button options on a given hand) as chance under the assumption that subjects usually correctly identified which hand to use for a given task.

Analyses of the fMRI data that tested for behavioral associations computed Spearman's rank correlation between the relevant blockwise imaging measure (neural task similarity or rule type detectability) and behavioral RT. For the neural task similarity measure (Fig 4), this was computed between the neural task similarity timecourse (one timecourse computed across all tasks, see above) and the cross-task averaged RT timecourse. For rule type detectability (Fig 7), this was computed separately between a given rule type's timecourse (sensory, logic, motor or conjunctive) and RT for a given task. The resulting Spearman's timecourses were then averaged across the 4 practiced tasks, with the resulting cross-task average for each rule type submitted to group random effects statistics.

Statistical analysis

All presented statistics – neural task similarity rho values, rule type regression betas, behavioral association rho values, and rho values for the association of neural task similarity and rule type detectability – used one-sample t-tests against 0 (2-sided) to assess significance, treating subject-level estimates as a random effect. Where specified in the Results, paired sample t-tests (2-sided) were used for contrasts across conditions. Spearman rho values underwent the Fisher-z transformation to improve normality prior to running the t-tests. T-statistics resulting from these contrasts are depicted in all figures, thresholded using the false discovery rate (FDR $p<.05$; Benjamini & Hochberg, 1995) procedure to correct for multiple comparisons across regions, networks or block timepoints (where appropriate). The Supplementary Table provides lists of significant regions for all regional analyses, with accompanying t and p values. This includes the CAB-NP region number (1-360 are cortical, and 361-718 are subcortical), allowing for extraction of anatomical information with reference to CAB-NP documentation (https://github.com/ColeLab/ColeAnticevicNetPartition).

Data availability

Preprocessed fMRI data for each subject will be uploaded to an Open Science Foundation repository (with the access link provided) upon publication.

Code availability

Custom MATLAB code used to empirically quantify the strength of rule type (i.e. compositional and conjunctive) representations on a block-to-block basis in the practiced task data is provided here https://github.com/ColeLab/CPRO2_learning_release.

References

Amer, T., & Davachi, L. (2023). Extra-hippocampal contributions to pattern separation. *ELife*, *12*, e82250. https://doi.org/10.7554/eLife.82250

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20–28. https://doi.org/10.1016/j.cobeha.2020.07.002

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. L. (2008). Pattern Separation in the Human Hippocampal CA3 and Dentate Gyrus. *Science*, *319*(5870), 1640–1642. https://doi.org/10.1126/science.1152882

Bassett, D. S., Yang, M., Wymbs, N. F., & Grafton, S. T. (2015). Learning-induced autonomy of sensorimotor systems. *Nature Neuroscience*, *18*(5), 744–751. https://doi.org/10.1038/nn.3993

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Berens, S. C., Horst, J. S., & Bird, C. M. (2018). Cross-Situational Learning Is Supported by Propose-but-Verify Hypothesis Testing. *Current Biology*, *28*(7), 1132-1136.e5. https://doi.org/10.1016/j.cub.2018.02.042

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4),

954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Beukema, P., & Verstynen, T. (2018). Predicting and binding: Interacting algorithms supporting the consolidation of sequential motor skills. *Current Opinion in Behavioral Sciences*, *20*, 98–103. https://doi.org/10.1016/j.cobeha.2017.11.014

Boersma, P., & Weenik, D. (2023). *Praat: Doing phonetics by computer*. http://www. praat. org/

Bowman, C. R., Chamberlain, J. D., & Dennis, N. A. (2019). Sensory Representations Supporting Memory Specificity: Age Effects on Behavioral and Neural Discriminability. *The Journal of Neuroscience*, *39*(12), 2265–2275. https://doi.org/10.1523/JNEUROSCI.2022-18.2019

Brooks, D., & Baddeley, A. (1976). What can amnesic patients learn? *Neuropsychologia*, *14*(1), 111–122. https://doi.org/10.1016/0028-3932(76)90012-9

Brunec, I. K., Robin, J., Olsen, R. K., Moscovitch, M., & Barense, M. D. (2020). Integration and differentiation of hippocampal memory traces. *Neuroscience & Biobehavioral Reviews*, *118*, 196–208. https://doi.org/10.1016/j.neubiorev.2020.07.024

Buckner, R. L. (2013). The Cerebellum and Cognitive Function: 25 Years of Insight from Anatomy and Neuroimaging. *Neuron*, *80*(3), 807–815. https://doi.org/10.1016/j.neuron.2013.10.044

Chein, J. M., & Schneider, W. (2005). Neuroimaging studies of practice-related change: FMRI and meta-analytic evidence of a domain-general control network for learning. *Cognitive Brain Research*, *25*(3), 607–623. https://doi.org/10.1016/j.cogbrainres.2005.08.013

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage*, *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Cocuzza, C. V., Ito, T., Schultz, D., Bassett, D. S., & Cole, M. W. (2020). Flexible Coordinator

and Switcher Hubs for Adaptive Task Control. *The Journal of Neuroscience*, *40*(36), 6949–6968. https://doi.org/10.1523/JNEUROSCI.2559-19.2020

Cole, M. W., Bagic, A., Kass, R., & Schneider, W. (2010). Prefrontal Dynamics Underlying Rapid Instructed Task Learning Reverse with Practice. *Journal of Neuroscience*, *30*(42), 14245–14254. https://doi.org/10.1523/JNEUROSCI.1662-10.2010

Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W., & Braver, T. S. (2011). Rapid Transfer of Abstract Rules to Novel Contexts in Human Lateral Prefrontal Cortex. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00142

Cole, M. W., Ito, T., Bassett, D. S., & Schultz, D. H. (2016). Activity flow over resting-state networks shapes cognitive task activations. *Nature Neuroscience*, *19*(12), Article 12. https://doi.org/10.1038/nn.4406

Cole, M. W., Laurent, P., & Stocco, A. (2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(1), 1–22. https://doi.org/10.3758/s13415-012-0125-7

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, *16*(9), 1348–1355. https://doi.org/10.1038/nn.3470

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Corkin, S. (1984). Lasting Consequences of Bilateral Medial Temporal Lobectomy: Clinical Course and Experimental Findings in H.M. *Seminars in Neurology*, *4*(02), 249–259. https://doi.org/10.1055/s-2008-1041556

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*(5), 1415–1428. https://doi.org/10.3758/s13423-016-1002-0

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework

for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*(4), e1005508. https://doi.org/10.1371/journal.pcbi.1005508

Dobbins, I. G., Foley, H., Schacter, D. L., & Wagner, A. D. (2002). Executive Control during Episodic Retrieval. *Neuron*, *35*(5), 989–996. https://doi.org/10.1016/S0896-6273(02)00858-9

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661. https://doi.org/10.1038/nature08704

Ebner, N. C. (2008). Age of face matters: Age-group differences in ratings of young and old faces. *Behavior Research Methods*, *40*(1), 130–136. https://doi.org/10.3758/BRM.40.1.130

Eichenbaum, H. (2013). What H.M. Taught Us. *Journal of Cognitive Neuroscience*, *25*(1), 14–21. https://doi.org/10.1162/jocn_a_00285

Eichenbaum, H., & Cohen, N. J. (2014). Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? *Neuron*, *83*(4), 764–770. https://doi.org/10.1016/j.neuron.2014.07.032

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184–188. https://doi.org/10.1038/nature01964

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, *7*(1), 11066. https://doi.org/10.1038/ncomms11066

Ferrari, C., Cattaneo, Z., Oldrati, V., Casiraghi, L., Castelli, F., D'Angelo, E., & Vecchi, T. (2018). TMS Over the Cerebellum Interferes with Short-term Memory of Visual Sequences. *Scientific Reports*, *8*(1), 6722. https://doi.org/10.1038/s41598-018-25151-y

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44). https://doi.org/10.1073/pnas.1800755115

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). “Who” Is Saying “What”? Brain-Based Decoding of Human Voice and Speech. *Science*, *322*(5903), 970–973. https://doi.org/10.1126/science.1164318

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Gao, J.-H., Parsons, L. M., Bower, J. M., Xiong, J., Li, J., & Fox, P. T. (1996). Cerebellum Implicated in Sensory Acquisition and Discrimination Rather Than Motor Control. *Science*, *272*(5261), 545–547. https://doi.org/10.1126/science.272.5261.545

Gao, Z., Davis, C., Thomas, A. M., Economo, M. N., Abrego, A. M., Svoboda, K., De Zeeuw, C. I., & Li, N. (2018). A cortico-cerebellar loop for motor planning. *Nature*, *563*(7729), 113–116. https://doi.org/10.1038/s41586-018-0633-x

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Hampshire, A., Daws, R. E., Neves, I. D., Soreq, E., Sandrone, S., & Violante, I. R. (2019). Probing cortical and sub-cortical contributions to instruction-based learning: Regional specialisation and global network dynamics. *NeuroImage*, *192*, 88–100. https://doi.org/10.1016/j.neuroimage.2019.03.002

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Hearne, L. J., Mill, R. D., Keane, B. P., Repovš, G., Anticevic, A., & Cole, M. W. (2021). Activity flow underlying abnormalities in brain activations and cognition in schizophrenia. *Science Advances*, *7*(29), eabf2513. https://doi.org/10.1126/sciadv.abf2513

Hirsh, R. (1980). The hippocampus, conditional operations, and cognition. *Physiological Psychology*, *8*(2), 175–182. https://doi.org/10.3758/BF03332848

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, *6*(9), 813–827. https://doi.org/10.1080/03610927708827533

Hull, C. (2020). Prediction signals in the cerebellum: Beyond supervised motor learning. *ELife*, *9*, e54073. https://doi.org/10.7554/eLife.54073

Ito, T., Hearne, L., Mill, R., Cocuzza, C., & Cole, M. W. (2019). Discovering the Computational Relevance of Brain Network Organization. *Trends in Cognitive Sciences*, S1364661319302402. https://doi.org/10.1016/j.tics.2019.10.005

Ito, T., Kulkarni, K. R., Schultz, D. H., Mill, R. D., Chen, R. H., Solomyak, L. I., & Cole, M. W. (2017). Cognitive task information is transferred between brain regions via resting-state network topology. *Nat. Commun.*, *8*(1), 1027. https://doi.org/10.1038/s41467-017-01000-w

Ito, T., Yang, G. R., Laurent, P., Schultz, D. H., & Cole, M. W. (2020). *Constructing neural network models from brain data reveals representational transformations underlying adaptive behavior* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.12.24.424353

Ito, T., Yang, G. R., Laurent, P., Schultz, D. H., & Cole, M. W. (2022). Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. *Nature Communications*, *13*(1), 673. https://doi.org/10.1038/s41467-022-28323-7

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., & Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57. https://doi.org/10.1016/j.neuroimage.2018.10.006

Kaefer, K., Stella, F., McNaughton, B. L., & Battaglia, F. P. (2022). Replay, the default mode network and the cascaded memory systems model. *Nature Reviews Neuroscience*, *23*(10), 628–640. https://doi.org/10.1038/s41583-022-00620-6

Kami, A., Meyer, G., Jezzard, P., Adams, M. M., Turner, R., & Ungerleider, L. G. (1995). Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature*, *377*(6545), 155–158. https://doi.org/10.1038/377155a0

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kauvar, I. V., Machado, T. A., Yuen, E., Kochalka, J., Choi, M., Allen, W. E., Wetzstein, G., & Deisseroth, K. (2020). Cortical Observation by Synchronous Multifocal Optical Sampling Reveals Widespread Population Encoding of Actions. *Neuron*, *107*(2), 351-367.e19. https://doi.org/10.1016/j.neuron.2020.04.023

Kikumoto, A., & Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences*, *117*(19), 10603–10608.

https://doi.org/10.1073/pnas.1922166117

King, M., Hernandez-Castillo, C. R., Poldrack, R. A., Ivry, R. B., & Diedrichsen, J. (2019). Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nature Neuroscience*, *22*(8), 1371–1378. https://doi.org/10.1038/s41593-019-0436-x

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Lafer-Sousa, R., Conway, B. R., & Kanwisher, N. G. (2016). Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *The Journal of Neuroscience*, *36*(5), 1682–1697. https://doi.org/10.1523/JNEUROSCI.3164-15.2016

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lang, C. E., & Bastian, A. J. (2002). Cerebellar Damage Impairs Automaticity of a Recently Practiced Movement. *Journal of Neurophysiology*, *87*(3), 1336–1347. https://doi.org/10.1152/jn.00368.2001

Leutgeb, J. K., Leutgeb, S., Moser, M.-B., & Moser, E. I. (2007). Pattern Separation in the Dentate Gyrus and CA3 of the Hippocampus. *Science*, *315*(5814), 961–966. https://doi.org/10.1126/science.1135801

Liang, J. C., Erez, J., Zhang, F., Cusack, R., & Barense, M. D. (2020). Experience Transforms Conjunctive Object Representations: Neural Evidence for Unitization After Visual Expertise. *Cerebral Cortex*, *30*(5), 2721–2739. https://doi.org/10.1093/cercor/bhz250

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 46. https://doi.org/10.1038/s41467-019-13930-8

Manns, J. R., & Eichenbaum, H. (2006). Evolution of declarative memory. *Hippocampus*, *16*(9), 795–808. https://doi.org/10.1002/hipo.20205

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *262*(841), 23–81. https://doi.org/10.1098/rstb.1971.0078

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159–188. https://doi.org/10.1037/0096-3445.114.2.159

McKenzie, S., & Eichenbaum, H. (2011). Consolidation and Reconsolidation: Two Lives of Memories? *Neuron*, *71*(2), 224–233. https://doi.org/10.1016/j.neuron.2011.06.037

Mill, R. D., Cavin, I., & O'Connor, A. R. (2015). Differentiating the functional contributions of resting connectivity networks to memory decision-making: FMRI support for multistage control processes. *Journal of Cognitive Neuroscience*, *27*(8), 1617–1632. https://doi.org/10.1162/jocn_a_00808

Mill, R. D., Hamilton, J. L., Winfield, E. C., Lalta, N., Chen, R. H., & Cole, M. W. (2022). Network modeling of dynamic brain interactions predicts emergence of neural information that supports human cognitive behavior. *PLOS Biology*, *20*(8), e3001686. https://doi.org/10.1371/journal.pbio.3001686

Mill, R. D., O'Connor, A. R., & Dobbins, I. G. (2016). Pupil dilation during recognition memory: Isolating unexpected recognition from judgment uncertainty. *Cognition*, *154*, 81–94. https://doi.org/10.1016/j.cognition.2016.05.018

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 630–633. https://doi.org/10.3758/BF03206543

Mohr, H., Wolfensteller, U., Betzel, R. F., Mišić, B., Sporns, O., Richiardi, J., & Ruge, H. (2016). Integration and segregation of large-scale brain networks during short-term task automatization. *Nature Communications*, *7*(1), 13217. https://doi.org/10.1038/ncomms13217

Murray, E. A., & Mishkin, M. (1986). Visual recognition in monkeys following rhinal cortical ablations combined with either amygdalectomy or hippocampectomy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *6*(7), 1991–2003. https://doi.org/10.1523/JNEUROSCI.06-07-01991.1986

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775. https://doi.org/10.1016/j.tics.2021.06.001

Nadel, L., Samsonovich, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*(4), 352–368. https://doi.org/10.1002/1098-1063(2000)10:4<352::AID-HIPO2>3.0.CO;2-D

Nakai, T., & Nishimoto, S. (2022). Representations and decodability of diverse cognitive functions are preserved across the human cortex, cerebellum, and subcortex. *Communications Biology*, *5*(1), 1245. https://doi.org/10.1038/s42003-022-04221-y

Nash, M. I., Hodges, C. B., Muncy, N. M., & Kirwan, C. B. (2021). Pattern separation beyond the hippocampus: A HIGH-RESOLUTION WHOLE-BRAIN investigation of mnemonic discrimination in healthy adults. *Hippocampus*, *31*(4), 408–421. https://doi.org/10.1002/hipo.23299

Nixon, P. D., & Passingham, R. E. (2000). The cerebellum and cognition: Cerebellar lesions impair sequence learning but not conditional visuomotor learning in monkeys. *Neuropsychologia*, *38*(7), 1054–1072.

https://doi.org/10.1016/S0028-3932(99)00138-4

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press ; Oxford University Press.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary Learning Systems. *Cognitive Science*, *38*(6), 1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682. https://doi.org/10.1002/hipo.450040605

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311–345. https://doi.org/10.1037/0033-295X.108.2.311

Pidgeon, L. M., & Morcom, A. M. (2016). Cortical pattern separation and item-specific memory encoding. *Neuropsychologia*, *85*, 256–271. https://doi.org/10.1016/j.neuropsychologia.2016.03.026

Poldrack, R. (1998). The neural basis of visual skill learning: An fMRI study of mirror reading. *Cerebral Cortex*, *8*(1), 1–10. https://doi.org/10.1093/cercor/8.1.1

Ranganath, C. (2010). Binding Items and Contexts: The Cognitive Neuroscience of Episodic Memory. *Current Directions in Psychological Science*, *19*(3), 131–137. https://doi.org/10.1177/0963721410368805

Reagh, Z. M., & Ranganath, C. (2023). Flexible reuse of cortico-hippocampal representations during encoding and recall of naturalistic events. *Nature Communications*, *14*(1), 1279. https://doi.org/10.1038/s41467-023-36805-5

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Ruge, H., & Wolfensteller, U. (2010). Rapid Formation of Pragmatic Rule Representations in the Human Brain during Instruction-Based Learning. *Cerebral Cortex*, *20*(7), Article 7.

https://doi.org/10.1093/cercor/bhp228

Sakai, K. (2008). Task Set and Prefrontal Cortex. *Annual Review of Neuroscience*, *31*(1), 219–245. https://doi.org/10.1146/annurev.neuro.31.060407.125642

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, *55*, 103–111. https://doi.org/10.1016/j.conb.2019.02.002

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science*, *27*(3), Article 3. https://doi.org/10.1207/s15516709cog2703_8

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schultz, D. H., Ito, T., & Cole, M. W. (2022). Global connectivity fingerprints predict the domain generality of multiple-demand regions. *Cerebral Cortex*, *32*(20), 4464–4479. https://doi.org/10.1093/cercor/bhab495

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439–454. https://doi.org/10.1037/0033-295X.94.4.439

Shimamura, A. P., & Wickens, T. D. (2009). Superadditive memory strength for item and source recognition: The role of hierarchical relational binding in the medial temporal lobe. *Psychological Review*, *116*(1), 1–19. https://doi.org/10.1037/a0014500

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible

sensorimotor decisions. *Science*, *348*(6241), 1352–1355. https://doi.org/10.1126/science.aab0551

Simons, J. S., & Spiers, H. J. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews Neuroscience*, *4*(8), 637–648. https://doi.org/10.1038/nrn1178

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. https://doi.org/10.1037/0033-295X.99.2.195

Stark, S. M., Kirwan, C. B., & Stark, C. E. L. (2019). Mnemonic Similarity Task: A Tool for Assessing Hippocampal Integrity. *Trends in Cognitive Sciences*, *23*(11), 938–951. https://doi.org/10.1016/j.tics.2019.08.003

Strick, P. L., Dum, R. P., & Fiez, J. A. (2009). Cerebellum and Nonmotor Function. *Annual Review of Neuroscience*, *32*(1), 413–434. https://doi.org/10.1146/annurev.neuro.31.060407.125606

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*(2), 129–144. https://doi.org/10.3758/BF03337828

Tomás Pereira, I., Agster, K. L., & Burwell, R. D. (2016). Subcortical connections of the perirhinal, postrhinal, and entorhinal cortices of the rat. I. afferents. *Hippocampus*, *26*(9), 1189–1212. https://doi.org/10.1002/hipo.22603

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential Effects of Early Hippocampal Pathology on Episodic and Semantic Memory. *Science*, *277*(5324), 376–380. https://doi.org/10.1126/science.277.5324.376

Weinberger, S. (2016). *Speech Accent Archive*. http://accent.gmu.edu

Whishaw, I. Q., & Tomie, J.-A. (1991). Acquisition and retention by hippocampal rats of simple, conditional, and configural tasks using tactile and olfactory cues: Implications for hippocampal function. *Behavioral Neuroscience*, *105*(6), 787–797. https://doi.org/10.1037/0735-7044.105.6.787

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). How to build a cognitive map. *Nature Neuroscience*, *25*(10), 1257–1272. https://doi.org/10.1038/s41593-022-01153-y

Wickelgren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review*, *86*(1), 44–60. https://doi.org/10.1037/0033-295X.86.1.44

Wood, E. R., Dudchenko, P. A., & Eichenbaum, H. (1999). The global record of memory in hippocampal neuronal activity. *Nature*, *397*(6720), 613–616. https://doi.org/10.1038/17605

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492. https://doi.org/10.1109/CVPR.2010.5539970

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Yeung, N., Nystrom, L. E., Aronson, J. A., & Cohen, J. D. (2006). Between-Task Competition and Cognitive Control in Task Switching. *The Journal of Neuroscience*, *26*(5), 1429–1438. https://doi.org/10.1523/JNEUROSCI.3109-05.2006

Yonelinas, A. P., Ranganath, C., Ekstrom, A. D., & Wiltgen, B. J. (2019). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews Neuroscience*, *20*(6), 364–375. https://doi.org/10.1038/s41583-019-0150-4