

Dynamically shifting from compositional to conjunctive brain representations supports cognitive task learning

Received: 18 March 2025

Accepted: 6 October 2025

Published online: 18 November 2025

 Check for updatesRavi D. Mill  & Michael W. Cole 

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 cognitive task learning. The strengthening of conjunctive representations hence serves as a computational signature of learning, reflecting cortical-subcortical dynamics that optimize task representations in the human brain.

Learning cognitive tasks is a common feature of everyday life, yet the neural basis for this essential human skill remains unknown. Studies have consistently reported behavioral improvement as tasks are performed from initial novel presentation through extensive practice, which has been linked to a progression from controlled to automatic processing^{1,2}. fMRI has alluded to 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^{3–7}. The activation decreases are most prominent in cognitive control-related regions^{8–10}. However, a precise computational account for how neural representations change over learning remains elusive.

In the present report, we assimilated concepts from the language, long-term memory, context-dependent decision-making and cognitive control neuroscientific literatures to develop hypotheses about the neural basis of task learning. These hypotheses are elaborated below, and summarized here: i) novel task performance is reliant on task-general compositional brain representations, ii) practiced task

performance is reliant on task-specialized conjunctive brain representations, iii) these representational dynamics are differentially evoked in cortex versus subcortex (constituting a division of labor between these systems), and iv) progressive strengthening of conjunctions specifically in cortex associates with signatures of effective task practice.

Our first hypothesis was informed by early language theories^{11,12} and ongoing research involving artificial neural networks^{13–15}, which position compositionality as an important representational principle. Compositions provide a format for representing abstract information that can be retrieved and recombined based on task context, theoretically allowing for flexible thought and behavior^{16–20}. Interest in compositionality has been accelerated by observed shortcomings of artificial neural network models in performing rapid instructed task learning²¹ - a human ability allowing for competent task performance with little to no practice. Longstanding artificial neural network deficits in transferring prior learning in “rapid” task contexts (e.g. zero-shot transfer)^{22,23} have recently been ameliorated by models embedded

with compositional principles^{15,24}. Extending these findings to human cognitive task learning, we predicted that compositional representations in the brain are critical in processing novel tasks. The degree to which compositional representations are evoked in empirical brain data (versus simulated data used for artificial neural network modeling), and how this might vary over the course of extended task learning, remains uncertain.

Whilst compositional representations theoretically allow for task competency in novel contexts, compositionality alone is unlikely to account for established effects of practice in improving behavioral performance towards true task proficiency. This informed our second hypothesis, which frames a complementary role for conjunctive representations as the neural substrate of task practice. In long-term memory research, nonlinear conjunctions of sensory features have long been theorized as a key computation, under varying terminology: chunking²⁵, binding^{26–28}, relational representation²⁹, configural (versus elemental) association^{30,31} and object-based (versus feature-based) representation³². Extending these ideas to the domain of cognitive task learning, we predicted that conjunctions of task rule elements would strengthen over extended learning, and uniquely associate with signatures of effective practice (reduced representational interference and improved behavior). This would offer a contrasting trajectory to compositional representations, which would be more strongly evoked early in learning, and then progressively weaken.

Finally, we also predicted that these representational dynamics would be differentially evoked across cortical and subcortical systems. This was founded on seminal neuropsychological studies of long-term memory^{29,33–36} demonstrating that regions including but not limited to the medial temporal lobe (MTL) subserve learning and memory. This was formalized in many multiple memory systems frameworks^{31,37–42}.

Critical amongst these was complementary learning systems theory^{40,43–45}, which posited differentiable roles for MTL and cortical regions in learning. MTL was theorized to perform fast learning (via fast conjunction formation) whereas cortex performs slow learning (via slow conjunction formation). These distinct roles in the service of a common learning function was characterized as a “division of labor” between cortical and subcortical systems. Embedding this division of labor in artificial neural network models accurately simulated memory deficits from animal lesioning studies⁴⁰. Extending to cognitive task learning, we predicted that repeated task practice is underpinned by strengthening of conjunctive representations specifically in cortex. This follows from the alignment of the theorized slow learning function of cortex with the slow timescale of practice-related behavioral improvement. Critically, to our knowledge, the predictions of complementary learning systems theory have not been tested in empirical human brain data - in the domain of task learning, but also generally.

Furthermore, complementary learning systems theory places predominant emphasis on the role of conjunctive representations in extended learning, thereby neglecting the reciprocal importance of compositions under task novelty. This contrasts with the inverse pattern of emphasis in recent artificial neural network models - promoting the importance of compositions (as a substrate of abstraction) and neglecting conjunctions (as a substrate of specialization). We offer an integrative theoretical account in Fig. 1A, in which cortical-subcortical dynamics underpin a shift from relying on compositional representations when tasks are novel, to relying on conjunctive representations with task practice. Unlike complementary learning systems theory, we conceptualize subcortical involvement in learning as not exclusive to the MTL, but also strongly involving the cerebellum, given its prior linkages with task learning^{46,47}, behavioral automaticity^{3,48,49}, and flexibility generally^{50–52}. Figure 1B schematizes how cortical conjunctive strengthening might drive signatures of effective task practice - reduced representational interference and improved behavior.

In devising an analytic framework to test the schematized learning principles in empirical human brain data, we were inspired by recent

innovations in the context-dependent decision-making literature. Machine learning approaches have been applied to functional brain data to quantify the degree to which neural representations evoked during decision-making are context-specific (conjunctive) or context-general (compositional)^{53–62}. Whereas these representational geometry approaches have been primarily applied at the within-trial timescale, here we applied similar methods over the longer timescale of extended task learning.

We tested our hypotheses using fMRI recordings in healthy adult human participants as they performed multiple complex cognitive tasks from initial novel presentation through repeated practice. We developed an updated version of the concrete permuted rule operations paradigm (C-PRO2), which preserved the compositional (each task is created by permuting sensory, logic and motor rule types) and multi-task (allowing for more generalizable inferences than single-task approaches) nature of previous versions^{9,63,64}, but now with more complex, naturalistic and diverse features (see Methods and Fig. 2A). This combination of theoretical, analytic and task design 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 C-PRO2 paradigm

We first verified that task learning had taken place as expected in our updated C-PRO2 paradigm (see Methods and Fig. 2A). Figure 2B plots the block-averaged accuracy across key conditions: practiced tasks presented in the first Practice session (4 tasks repeated 36 blocks each), practiced tasks in the second Test session (same 4 tasks repeated for a further 15 blocks each), and novel tasks in the Test session (60 tasks presented once). Subjects were reliably above chance accuracy (cross-task average=77.6%) in all three conditions (one-sample t-tests against 50%, all $p < 0.001$, all Cohen's $d > 1.60$), confirming that participants understood the paradigm.

The cross-condition profile demonstrated that practice improved behavior: practiced tasks in both the Practice (mean = 79.0%) and Test (80.3%) sessions were more accurately performed than novel (73.6%) tasks (practiced task in Practice session > novel task in Test session, $t(43) = 2.90$, $p = 0.006$, Cohen's $d = 0.44$; practiced task in Test session > novel task in Test session, $t(43) = 5.14$, $p < 0.001$, Cohen's $d = 0.77$). Accuracy for practiced tasks in the Test session was numerically albeit non-significantly higher than practiced tasks in the Practice session ($t(43) = 1.10$, $p = 0.276$, Cohen's $d = 0.17$), and there were no reliable between-condition differences in RT (paired t-tests, all $p > 0.450$).

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) with repeated practice (paired t-test for last block > first block performance, accuracy $t(41) = 6.27$, $p < 0.001$, Cohen's $d = 0.97$; RT $t(39) = 2.47$, $p = 0.018$, Cohen's $d = 0.39$). These results confirmed that the key task learning manipulation was successful.

Evidence of task pattern separation: neural task similarity decreases with practice to reduce interference

Our fMRI analyses generally adopted multivariate pattern analysis techniques. We first used a classification approach (4-way decoding of practiced task identity) as a sanity check to verify that task information was reliably decodable from multivariate neural activations in our paradigm (Fig. S1).

Subsequent analyses targeted the critical question of how the geometry of neural representations changed over learning. We tested whether representational similarity in the brain decreased with repeated task practice, consistent with pattern separation processes that

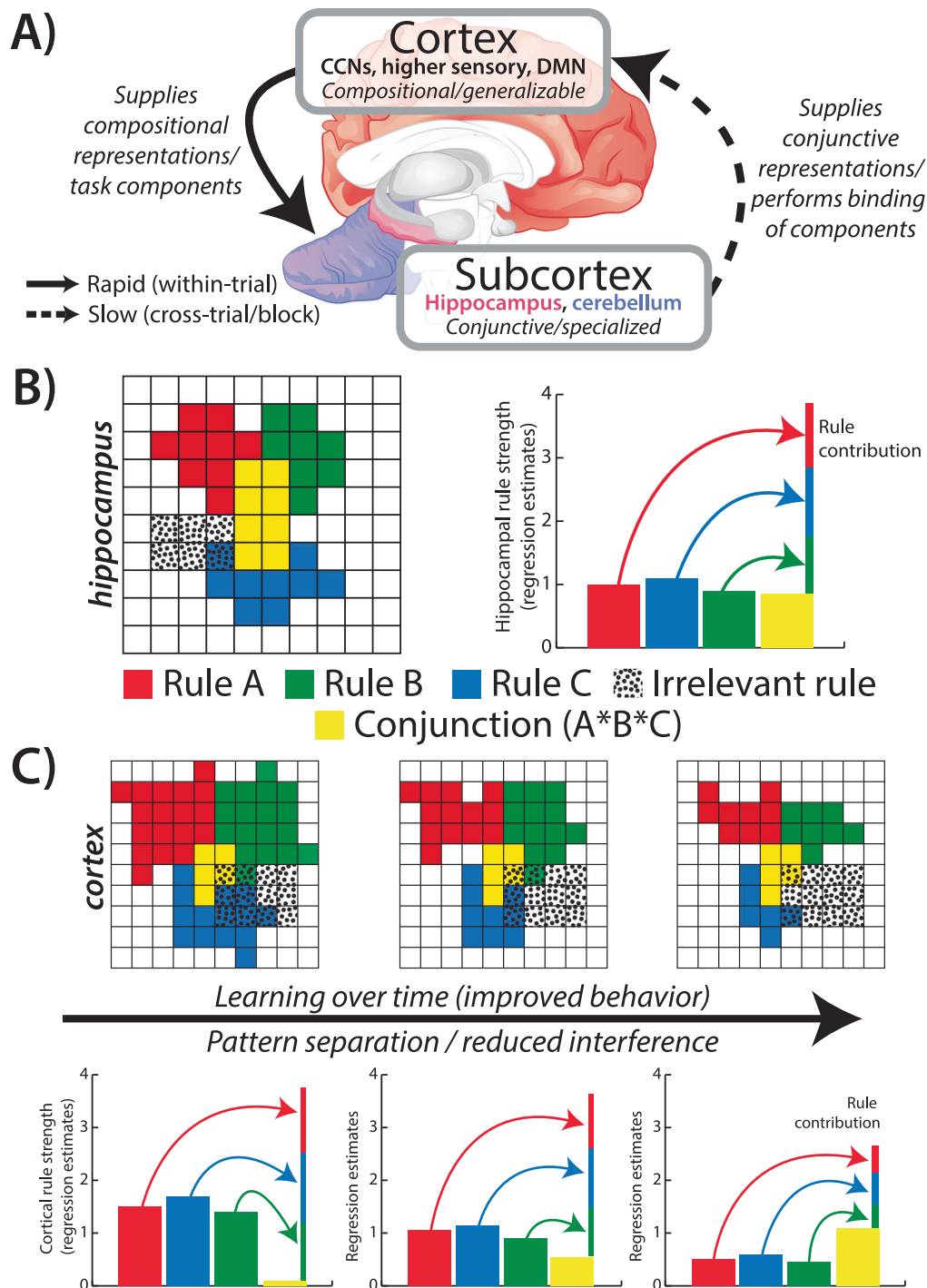


Fig. 1 | Hypothesized cortical-subcortical dynamics underlying changes to neural representational geometry over task learning. **A** Depiction in neural space. Early in learning, compositional representations of task rules are retrieved and reinstated in cortical networks (including cognitive control networks, CCNs, higher sensory networks and the default mode network, DMN), facilitating transfer of relevant pre-experiment knowledge to the current context (here, performance of C-PRO2 paradigm tasks; see Fig. 2A for design). These compositional representations are then rapidly bound by subcortex (hippocampus and cerebellum prominently) into more specialized conjunctive representations, and over practice propagated back to cortex, where they gradually strengthen relative to the compositional representations to facilitate behavioral improvement. **B** Depiction in representational space: hypothesized spatial activity pattern evoked in hippocampus early in learning. Complementary depictions of the same representational profile are provided: the left panel presents the hypothetical within-region activation pattern (each square representing activation of within-region voxels), and the right panel presents hypothetical results from applying our machine learning

approach to quantify rule representation strength in this region (regression weights from fitting linear compositional and non-linear conjunctive rule templates to the evoked activation pattern). Early in learning, hippocampus binds compositional task rule representations into non-linear conjunctions, which promotes encoding of a more distinct, specialized representation of the current task. One computational benefit of such conjunctive coding is to allow for a pattern-separated representation of the task, which is more resistant to interference from irrelevant rules (depicted as reduced overlap of task-relevant rules with an irrelevant rule's dotted pattern). **C** Equivalent activation (top panel) and regression space (bottom panel) depictions for a cortical region over learning. As tasks are repeatedly practiced, conjunctive codes increasingly dominate in cortical regions, with reliance on compositional codes reduced in parallel. The computational benefit of these representation dynamics is to promote task pattern separation, reducing interference with neural encoding of other tasks, and with a particular impact (compared to subcortex) in improving behavioral performance with extended practice.

reduce interference. This was hypothesized as a key signature of practice (Fig. 1), which is later linked to more precise representational dynamics. Here, our goal was to preliminarily establish pattern separation in the domain of cognitive task learning, versus more typically studied spatial memory and object recognition domains.

Our method of estimating neural task similarity is depicted in Fig. 3A. We analyzed the resulting regional similarity timecourses in two complementary ways, which, coupled with our rigorous fMRI preprocessing steps, ruled out potential contamination by temporal artifacts (see Methods). Firstly, we estimated the dynamic trajectory (regression slope over time) of the practiced task similarity timecourses in the first Practice session. Secondly, we contrasted the block-averaged similarity of practiced versus novel tasks presented in the second Test session. Consistent with our hypothesis that task practice reduces neural task similarity (increases pattern separation), we predicted that i) regional similarity trajectories would generally decrease over time for practiced tasks in the Practice session, ii) similarity would be lower for practiced versus novel tasks in the Test session. The results find support for both predictions (see Supplementary Data 1 for a list of significant regions and accompanying statistics), with Fig. 3B depicting the conjunction of these two analyses to isolate regions most strongly linked to task pattern separation. These included medial temporal lobe regions typically associated with pattern separation (bilateral hippocampus), but also cortical and subcortical regions (bilateral cerebellum, bilateral caudate nucleus and right amygdala) not as conventionally linked.

The Practice session regional trajectories were averaged into large-scale functional networks, based on the established Cole-Anticevic Brain Network Partition (CAB-NP⁶⁵; Fig. 3C). Whilst all networks showed a significant negative trajectory (FDR $p < 0.05$), this was numerically strongest in networks linked to higher cognitive functions: default mode network (DMN, trajectory $t(43) = -11.83$, $p < 0.001$, Cohen's $d = 1.78$), cognitive control networks (CON, $t(43) = -9.62$, $p < 0.001$, Cohen's $d = 1.45$; DAN, $t(43) = -7.84$, $p < 0.001$, Cohen's $d = 1.18$; FPN, $t(43) = -10.75$, $p < 0.001$, Cohen's $d = 1.62$) and the higher-order sensory networks (Language, $t(43) = -9.77$, $p < 0.001$, Cohen's $d = 1.47$; Visual 2, $t(43) = -7.71$, $p < 0.001$, Cohen's $d = 1.16$). The Supplement (Fig. S2) provides results at the whole-brain level, which again revealed a negative trajectory, and further highlighted that this held generally across all 12 task rules (3 rule types * 4 levels each = 12 rules total; see Methods).

Reduction in neural task similarity over practice was also reliably associated with behavioral improvement, for both accuracy (Fig. S3A) and RT (Fig. S3B). Overall, the results reveal changes to neural representational geometry consistent with task pattern separation, which reduces interference with practice.

Quantifying compositional and conjunctive rule representations during task learning yields a dissociation between subcortex and cortex

Preceding analyses revealed that neural task representations became more distinct from each other over learning, but left unclear what specifically was changing to produce this. We hypothesized that the reduction in cross-task neural similarity over learning was driven by an increasing reliance on conjunctive (versus compositional) representations of the task rules (Fig. 1).

To test this, we developed a machine learning approach to quantify the strength of compositional and conjunctive representations on a block-to-block basis (see Fig. 4A and Methods for details). Templates for single rule types (sensory, logic and motor) and their conjunction were estimated using held-out Test session data, and then fit simultaneously via robust regression to the regional practiced task activations, over Practice session blocks. The resulting beta coefficients quantified the strength of compositional versus conjunctive rule representations over time.

The overall (cross-block average) strength of each rule type is presented in Fig. 4B. Recovery of positive region peaks for the compositional rule types, i.e. visual and auditory regions for sensory rules, lateral prefrontal regions for logic rules, and motor cortex for motor rules (FDR $p < 0.05$; see Supplementary Data 1 for accompanying statistics), was consistent with prior literature^{66,67} and served as analysis sanity checks. 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 between subcortex and cortex in fulfilling distinct learning computations.

Rule representation dynamics reveal a neurocomputational signature of task practice: progressive replacement of compositional with conjunctive representations in cortex

To further examine dissociable roles for cortex and subcortex in task learning, we conducted dynamic analyses of rule representation strength. We hypothesized that with task practice, compositional representations in cortex would be slowly replaced with conjunctive representations (Fig. 1). To ensure sufficient power, these analyses used network summaries of cortical regions (regional rule beta coefficients averaged into affiliated CAB-NP networks), and 5 subcortical regions-of-interest that showed strong pattern separation and conjunctive representation effects (see Table S1 and Fig. S4).

The results are provided in Fig. 5, with left columns depicting rule strength timecourses, and right columns formalizing the timecourse trajectories. Firstly, we observed significant subcortex > cortex conjunctive strength, and significant cortex > subcortex compositional strength across all blocks (all paired t-tests significant in the respective directions, for all rule types, FDR $p < 0.05$), confirming the separable sensitivities suggested by the block-averaged results (Fig. 4B).

Further dissociation was revealed in the dynamic trajectories. Whereas conjunctive codes progressively strengthened in cortex (cortical average, trajectory $t(43) = 2.70$, $p = 0.009$, Cohen's $d = 0.41$), they progressively weakened in subcortex (subcortical average, trajectory $t(43) = -2.77$, $p = 0.008$, Cohen's $d = 0.42$), and this difference was significant (paired t-test cortex > subcortex, $t(43) = 3.52$, $p = 0.001$, Cohen's $d = 0.53$). Contrastingly, compositional codes progressively weakened across cortex (mean across 3 rules, cortical network average, $t(43) = -5.68$, $p < 0.001$, Cohen's $d = 0.86$) and subcortex (mean across 3 rules, subcortical region average, $t(43) = -4.62$, $p < 0.001$, Cohen's $d = 0.70$). This profile is consistent with the hypothesized importance of strengthening conjunctive representations in cortical networks specifically over repeated task practice (Fig. 1).

The Supplement (Fig. S5) provides an expanded statistical explanation of the Fig. 5 results, with a focus on correctly interpreting the early negative cortical conjunction estimates (Fig. 5A) as arising from representational interference between the compositions. This references established statistical understanding of the interaction term in regression models, complemented by simulations with a known ground truth, to substantiate that the brain truly exhibits an increasing reliance on cortical conjunctions with practice.

The Supplement (Fig. S6) also reports equivalent representational dynamics over learning when using an alternative, established measure of estimating compositional representations (cross-condition generalization performance⁵⁴). As detailed in the Supplement, we were precluded from applying an alternative method of estimating conjunctive representations⁵⁹ by inherent restrictions in our task design, specifically the lack of novel task repetitions.

Strengthening of cortical conjunctions is uniquely associated with behavioral improvement over practice

We hypothesized that the dynamic shift from compositional to conjunctive representations over task practice would be associated with behavioral improvement. We computed associations between the rule

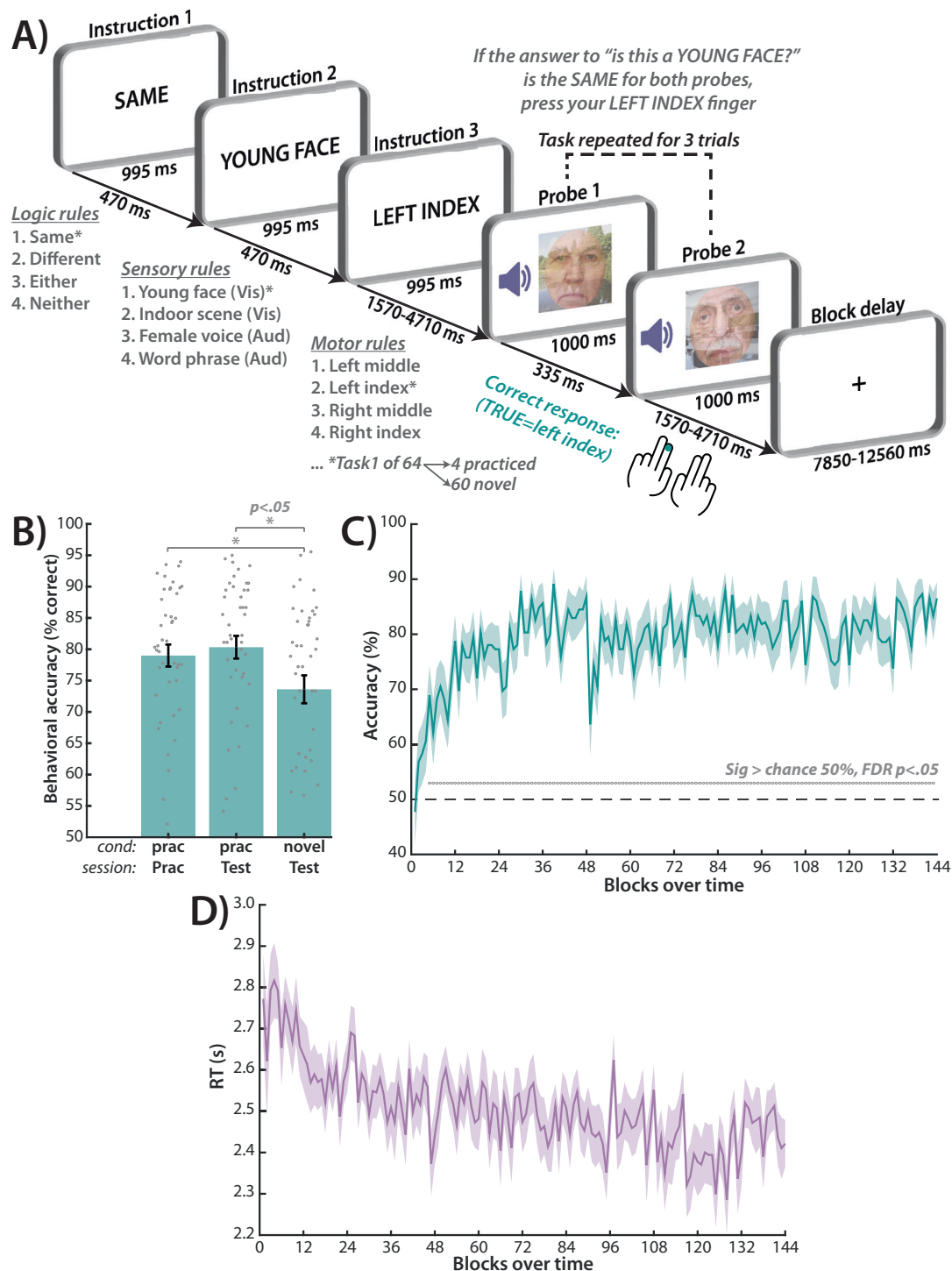


Fig. 2 | C-PRO2 task design and observed behavioral learning effects. **A** Format for one block of the concrete permuted rule operations paradigm, version 2 (C-PRO2; see Methods for details), depicting an example trial with accompanying correct response. This block format was preserved for all 64 tasks, across practiced/novel task conditions and Practice/Test sessions. A task block began with three instruction screens (one each for logic, sensory and motor rule types), which conveyed the rules to be applied for the ensuing set of three trials. Each trial concurrently presented composite visual (dimensions: young or old face, indoor or outdoor scene) and composite auditory (dimensions: female or male voice, word or non-word content) stimuli over two sequential probes. Subjects were required to integrate the task rule information to process the incoming stimulus information across the probes, and make their true/false categorization responses. Depicted composite face¹⁰² and scene¹⁰¹ databases (see Methods for details). **B** Session-

averaged accuracy across key task conditions: practiced tasks in the first Practice session (prac Prac), practiced tasks in the second Test session (prac Test), novel tasks in the Test session (novel Test). Subjects were reliably above chance performance across conditions (all $p < 0.001$), and practiced tasks were better performed than novel ones (prac Prac > novel Test $p = 0.006$, prac Test > novel Test $p < 0.001$). Error bars reflect standard error of the mean, with subject datapoints overlaid. Asterisks denote significance of paired t-tests at $p < 0.05$. **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-tests conducted at each block, multiple comparison-corrected at FDR $p < 0.05$; denoted by the gray dots) that plateaued around 30 blocks. **D** Reaction time (RT) over task blocks in the Practice phase, also showing improvement in performance over learning (quickening of RT). Patches in **C** and **D** reflect standard error of the mean. For panels **B–D**, sample $n = 44$, with two-sided t-tests (one-sample and paired) conducted.

strength timecourses and block-to-block variation in reaction time (RT). The results in Fig. 6 demonstrate that strengthened cortical conjunctions were linked to behavioral improvement (quicker RT) with practice, particularly in the FPN ($t(43) = -4.11$, $p < 0.001$, Cohen's $d = 0.62$). Conversely, associations between compositional strength and behavior over this extended timescale were generally in the harmful direction (slowing RT), for both cortex and subcortex, and especially for the sensory rule type (FPN sensory $t(43) = 4.03$, $p < 0.001$, Cohen's $d = 0.61$). The findings are consistent with our hypothesis that cortical conjunctive strengthening is a key computational signature of behavioral optimization with practice.

This pattern of brain-behavior association held in the FPN even after regressing out linear temporal effects prior to computing the associations via partial correlation (FPN conjunctive-RT association: $t(43) = -2.16$, $p = 0.036$, Cohen's $d = 0.33$; FPN sensory-RT association: $t(43) = 2.99$, $p = 0.005$, Cohen's $d = 0.45$). This was an exhaustive control for temporal confounds/artifacts, conducted in addition to removing linear trends and expanded noise regressors during pre-processing (see Methods). No significant associations were observed with behavioral accuracy (all accuracy/RT analyses were FDR-corrected for the full set of multiple comparisons across both measures; see Methods).

Our use of methods of brain-behavior association that were sensitive to both linear and non-linear relationships (i.e. Spearman's correlation) was validated by stepwise regression modeling of the behavioral RT timecourses (see Methods). This revealed significant behavioral prediction accuracy when entering the time predictor linearly ($\beta = 0.40$; $R^2 = 0.21$, $p < 0.001$ via two-sided Wilcoxon signed-rank test versus 0, $Z(43) = 4.91$) and non-linearly (i.e. power law; $\beta = 0.39$; $R^2 = 0.15$, $p < 0.001$, $Z(43) = 5.06$). Contrasting R^2 for linear > non-linear models via two-sided paired Wilcoxon: R^2 difference = 0.056, $p < 0.001$, $Z(43) = 3.44$. Critically, the combined model in which linear and non-linear predictors were entered simultaneously (combined model $R^2 = 0.21$, $p < 0.001$, $Z(43) = 4.41$) yielded numerically greater behavioral prediction accuracy than the linear only model (R^2 difference = 0.007, $p = 0.065$, paired $Z(43) = 1.84$), and significantly greater accuracy than the non-linear only model (R^2 difference = 0.062, $p = 0.004$, paired $Z(43) = 2.89$). These results suggest that both linear and non-linear temporal effects were present in behavior, which justified our use of Spearman's correlation in the brain-behavior association analyses.

Strengthening of cortical conjunctions over practice is associated with stronger pattern separation

We hypothesized that one advantage of strengthening conjunctive representations with practice is to increase pattern separation, and consequently reduce cross-task interference. We therefore interrogated whether pattern separation dynamics (reduced neural task similarity over learning) were associated with rule representation strength. For each region, we computed the Spearman's correlation between the separate measures of neural task similarity (estimated per Fig. 3A) and compositional/conjunctive rule strength (estimated per Fig. 4A), across Practice session blocks. The results in Fig. 7 demonstrate that compositional weakening over this extended timescale was associated with improved pattern separation (reduced neural task similarity). This held across cortex/subcortex, and across the 3 single rule types (cross-rule cortex average, $t(43) = 39.74$, $p < 0.001$, Cohen's $d = 5.99$; cross-rule subcortex average, $t(43) = 14.16$, $p < 0.001$, Cohen's $d = 2.14$). Weakening of subcortical conjunctions was also associated with stronger pattern separation (subcortex average, $t(43) = 4.62$, $p < 0.001$, Cohen's $d = 0.70$). Critically, conjunctions in cortex uniquely showed the opposite, beneficial association: stronger cortical conjunctions related to stronger pattern separation (cortex average, $t(43) = -24.26$, $p < 0.001$, Cohen's $d = 3.66$).

The results are again consistent with the hypothesized differentiation in function between cortex and subcortex over task learning. Critically, the evidenced beneficial relationship between cortical conjunctions and pattern separation was predicted by our theoretical framework (Fig. 1), as a signature of effective practice that reduces representational interference.

Cortical conjunctions continue to strengthen across a second learning session

If cortical conjunctions serve as a neural signature of the extent/robustness of task learning (per our hypotheses), then they should continue to strengthen with additional practice. Whereas preceding analyses focused on the first Practice session, here we estimated the strength of conjunctive representations for practiced tasks presented in the second Test session (separated from the first session by 1-7 days; see Methods).

Consistent with our predictions, the timecourses in Fig. 8A revealed further strengthening of cortical conjunctions as practice progressed in the Test session. This visual impression was formalized via contrasts of rule strength over temporal epochs (averages over 6 blocks at the beginning and end of each session; see Methods). This revealed significantly stronger (more positive) cortical conjunctions for Test session epochs compared to Practice session epochs (early Prac < early Test, $t(43) = 3.61$, $p < 0.001$, Cohen's $d = 0.54$; early Prac < late Test, $t(43) = 4.06$, $p < 0.001$, Cohen's $d = 0.61$; late Prac < early Test, $t(43) = 2.99$, $p = 0.005$, Cohen's $d = 0.45$; late Prac < late Test, $t(43) = 3.15$, $p = 0.003$, Cohen's $d = 0.48$; all significant at FDR $p < 0.05$). The remaining contrasts were in the same direction of stronger cortical conjunctions with increasing practice, albeit non-significant (early Prac < late Prac, early Test < late Test, both $p > 0.510$, both Cohen's $d < 0.10$).

Subcortical conjunctions in the Test session showed the opposite trajectory, progressively weakening as practice continued (Fig. 8A). This was formalized in Fig. 8B as significantly weaker subcortical conjunctions in Test compared to Practice epochs (early Prac > early Test, $t(43) = 3.11$, $p = 0.003$, Cohen's $d = 0.47$; early Prac > late Test, $t(43) = 3.64$, $p < 0.001$, Cohen's $d = 0.55$; late Prac > early Test, $t(43) = 2.27$, $p = 0.028$, Cohen's $d = 0.34$; late Prac > late Test, $t(43) = 2.62$, $p = 0.012$, Cohen's $d = 0.39$; all FDR $p < 0.05$). The remaining contrasts were numerically in the same direction but non-significant (early Prac > late Prac, $t(43) = 1.65$, $p = 0.107$, Cohen's $d = 0.25$; early Test > late Test, $t(43) = 1.00$, $p = 0.322$, Cohen's $d = 0.15$). These findings reveal that the conjunctive dynamics observed in the first Practice session were extended with additional practice in the second Test session. The opposing trajectories between cortex and subcortex further support our hypothesized division of labor between these systems in task learning. Critically, the results substantiate cortical conjunctive strength as a neural index of the extent of task practice, even as it extends across separated learning episodes.

For completeness, the Supplement presents compositional rule strength estimated in the Test session (Fig. S7). This revealed transient compositional re-engagement at the start of the Test session, as a response to increased cognitive demand following the lapse in task performance, which was reflected in behavior (Fig. S8). The differing dynamic profile of compositions and cortical conjunctions is again consistent with our hypothesis that they underpin separable neural computations, albeit suggesting that compositional engagement occurs under general increase in cognitive demand, rather than exclusively under novelty.

Behavioral cost of task switching is uniquely indexed by cortical conjunction strength

We leveraged another C-PRO2 manipulation to substantiate the link between cortical conjunction strength and the robustness of practiced task representations. Task switching demands were varied across the 4

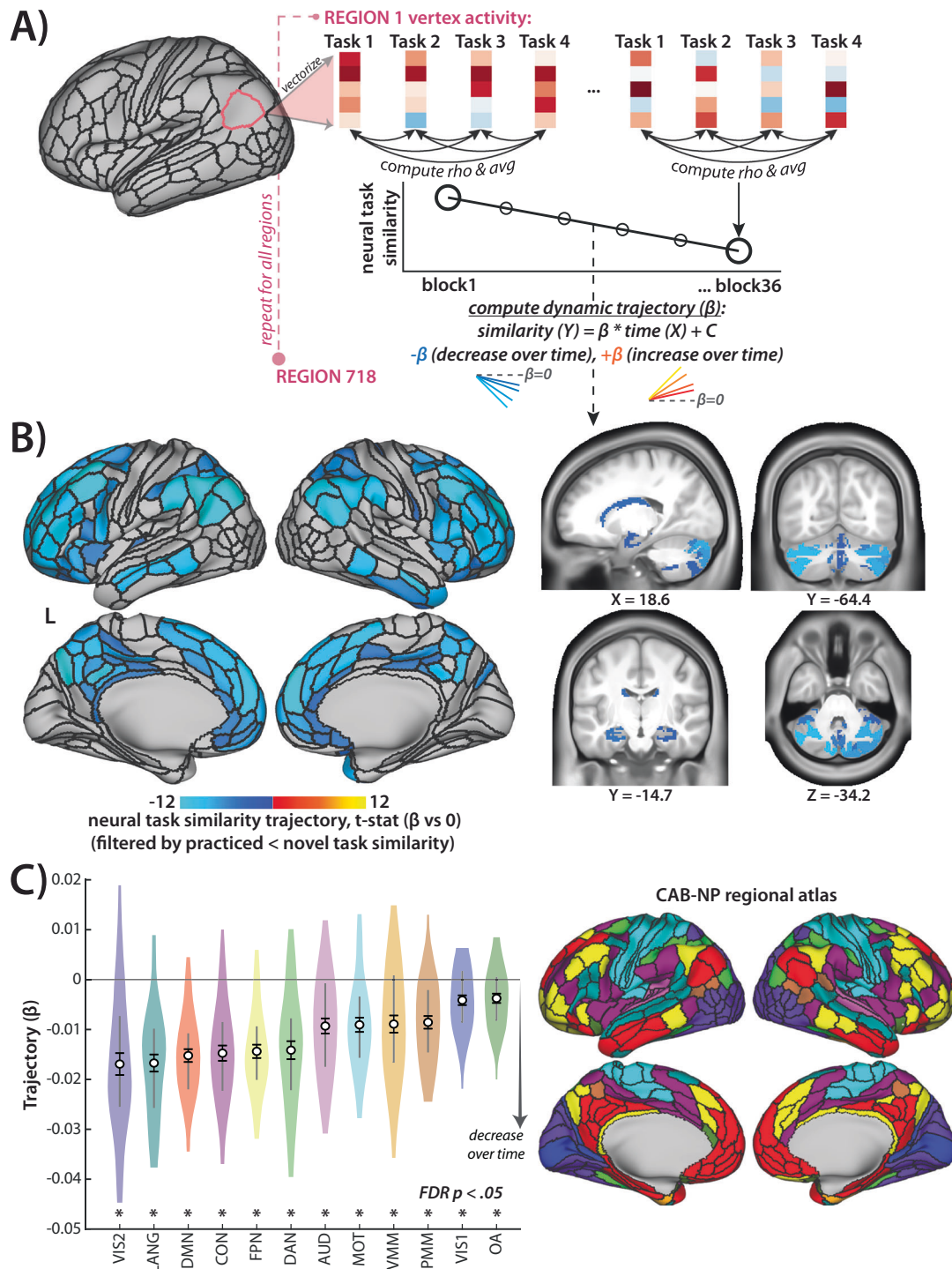


Fig. 3 | Task practice reduces neural task similarity, consistent with pattern separation processes that reduce interference. **A** Schematic of analysis approach. For each region of interest, the similarity between task activation features (within-region vertices for cortex, and within-region voxels 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 process repeated per region. **B** Strongest pattern separation regions, i.e. those showing a significant dynamic trajectory for practiced < novel task similarity difference in the Test session, both at FDR $p < 0.05$. Significant cortical regions are shown in the left panel, and significant subcortical regions are shown in the right panel. Neural task similarity decreased (i.e. pattern separation increased) with task practice. **C** Regional similarity trajectories in the Practice session averaged into functional networks from the Cole-Anticevic

Brainwide Network Partition atlas (CAB-NP; depicted in the right panel, with the network color index used here and in subsequent figures). Network abbreviations: 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). The violins in the figure represent the subject distribution, circles the mean, horizontal notches the standard error of the mean, and vertical lines the interquartile range. Asterisks denote significance via one-sample t -test against 0, after multiple comparison correction at FDR $p < 0.05$. Networks linked to higher cognitive functions showed the strongest practice-induced increases in pattern separation (VIS2, LANG, DMN, CON, FPN, DAN). For panels **B–C**, sample $n = 44$, with two-sided one-sample t -tests conducted. Brain underlays were generated using Connectome Workbench¹³⁸.

practiced tasks per subject: 2 tasks were presented under high switching (Switch condition), and the remaining 2 tasks under reduced switching (NoSwitch condition; see Methods). We predicted that the Switch condition would interfere with effective practice, leading to a behavioral cost relative to the NoSwitch condition, which would be uniquely indexed by the strength of cortical conjunctions.

We first verified that high task switching yielded the anticipated behavioral cost. Behavioral timecourses revealed slower RT for Switch compared to NoSwitch tasks early in the Practice session (epoch average over blocks 1–6; Fig. 8C); Switch early epoch mean=2691.2 ms, NoSwitch early epoch mean=2609.6 ms, paired $t(43) = 2.40$, $p = 0.021$, Cohen's $d = 0.36$. Accuracy also worsened for Switch tasks during this period (Fig. S9A); Switch early epoch mean=66.85%, NoSwitch early epoch mean=73.61%, paired $t(43) = 2.24$, $p = 0.030$, Cohen's $d = 0.34$. Late epoch Switch vs NoSwitch contrasts were non-significant (both $p > 0.240$, Cohen's $d < 0.18$). This profile is consistent with increased task switching disrupting effective practice.

We then assessed whether the strength of cortical conjunctions was sensitive to this switching cost. Conjunctive representations were estimated separately for Switch and NoSwitch conditions. The resulting timecourses (Fig. S9B) revealed similar dynamic trajectories across Switch and NoSwitch, verifying that prior analyses that collapsed across these conditions did not occlude dramatic differences. However, the intercept visibly differed between switching conditions. We formalized these effects via a 2 (brain system: cortex, subcortex) $\times$ 2 (switch condition: Switch, NoSwitch) repeated measures ANOVA on the session-level (block-averaged) conjunctive representation strength. This yielded a significant main effect of system (Subcortex $>$ Cortex, $F(1,43) = 77.0$, $p < 0.001$, $\eta^2_g = 0.413$) and a non-significant main effect of condition (Switch $>$ NoSwitch, $F(1,43) = 2.67$, $p = 0.109$, $\eta^2_g = 0.004$). Critically, the system-by-switch condition interaction was significant ($F(1,43) = 5.97$, $p = 0.019$, $\eta^2_g = 0.009$), reflecting a reliable reversal in the sign of the Switch vs NoSwitch effect between cortical and subcortical conjunctions. Planned paired t -tests revealed that cortical conjunctions were numerically stronger for NoSwitch tasks (Switch estimate(t)=−2.59, NoSwitch estimate(t)=−1.72; Switch $>$ NoSwitch paired $t(43) = -0.77$, $p = 0.449$, Cohen's $d = 0.12$), whereas subcortical conjunctions were significantly stronger for Switch tasks (Switch estimate(t)=8.07, NoSwitch estimate(t)=7.22; Switch $>$ NoSwitch paired $t(43) = 2.48$, $p = 0.017$, Cohen's $d = 0.37$). This dissociable profile again supports the division of labor between cortical and subcortical systems in task learning.

We then examined whether block-to-block variation in cortical conjunction strength was associated with block-to-block variation in behavioral switch costs. Figure 8D plots correlations between conjunctive rule strength differences (NoSwitch minus Switch) and behavioral RT differences (NoSwitch minus Switch) across Practice session blocks. The results in Fig. 8D reveal that conjunctive strength in the FPN appropriately and uniquely indexed the behavioral costs of task switching ($t(43) = -3.99$, $p < 0.001$, Cohen's $d = 0.60$; only network that survived at FDR $p < 0.05$). No reliable associations were observed with behavioral accuracy (all brain-behavior associations were corrected for the full set of multiple comparisons across RT and accuracy; see Methods).

Follow-up analyses revealed FPN's early conjunctive strength to be significantly greater for NoSwitch versus Switch tasks (NoSwitch estimate(t)=0.34, Switch estimate(t)=−2.09; paired $t(43) = 2.14$, $p = 0.038$, Cohen's $d = 0.32$), demonstrating sensitivity to the period of maximal behavioral cost. FPN's brain-behavior association index also held when confined to this early epoch of maximal cost (Spearman's correlation confined to blocks 1–6; $t(43) = -2.29$, $p = 0.027$, Cohen's $d = 0.35$). This FPN index was evidenced even after regressing out the linear temporal effect before computing the index, to rule out potential temporal confounds/artifacts (partial $t(43) = -3.45$, $p = 0.001$,

Cohen's $d = 0.52$). This aligns with earlier brain-behavior associations in the FPN that were also robust to this control analysis.

For completeness, the Supplement provides equivalent results for compositional rule strength (Fig. S10), revealing overall modest differences across switching conditions, and no statistically reliable brain-behavior association index. This further substantiates the unique link between cortical FPN conjunctions and practice-related behavioral improvement. The behavioral index captured more robust strengthening of conjunctions in the FPN under reduced task switching demands, which associated appropriately with behavioral improvement in that condition.

Cross-region association of rule representation dynamics reveals subcortical binding of compositions into conjunctions

Preceding analyses interrogated rule representation dynamics separately within each region and rule representation type (compositions and conjunctions). In ensuing analyses, we examined profiles of cross-region and cross-rule associations, to reveal insight into network interactions underlying the observed task learning dynamics.

We used a regularized multivariate measure of cross-region association, inspired by recent advancements in functional connectivity estimation (see Methods)⁶⁸. This measure was computed for all pairwise combinations of rule strength timecourses (i.e. compositional-compositional, conjunctive-conjunctive, and compositional-conjunctive associations), across Practice session blocks. We hypothesized that compositional-compositional associations would primarily capture how the same kind of rule information was propagated across the brain, and that this would prominently involve cortical networks. Conversely, compositional-conjunctive associations would capture the transformation of different rule representation formats, and prominently involve subcortical regions. These predictions follow from the differential sensitivities of cortex/subcortex to compositions/conjunctions in prior analyses, and multiple memory systems theories linking subcortex to binding operations^{26,28}. We did not have strong predictions regarding the cross-region conjunctive-conjunctive associations, but the highly specialized nature of these representations raised potential difficulties in registering them at the spatial scale of fMRI (versus, e.g. the scale of neural populations).

The group-averaged matrix for compositional-compositional cross-region associations is provided in Fig. 9A. This revealed generally positive associations, which were more prominent in cortex than subcortex. The visual impressions were formalized by averaging the distal (off-diagonal) associations for each cortical network/subcortical region (Fig. 9B), which revealed significantly greater associations for the cortical average ($t(43) = 40.4$, $p < 0.001$, Cohen's $d = 6.09$) compared to the subcortical average (subcortical $t(43) = 16.0$, $p < 0.001$, Cohen's $d = 2.41$; cortical $>$ subcortical, paired $t(43) = 23.9$, $p < 0.001$, Cohen's $d = 3.61$). The conjunctive-conjunctive results (Fig. 9C, D) revealed no significant cross-region association, suggesting that these were primarily local, or that distal associations were at too fine a scale to be registered with fMRI.

Figure 9E plots the critical compositional-conjunctive cross-region associations, which were visibly more positive for subcortex versus cortex. This impression was formalized by averaging the off-diagonal elements, which revealed significantly positive associations (at FDR $p < 0.05$) involving a cerebellar region (cVIS2; $t(43) = 3.07$, $p = 0.004$, Cohen's $d = 0.46$) and the subcortical average ($t(43) = 3.08$, $p = 0.004$, Cohen's $d = 0.47$; hippocampus was significant at uncorrected $p < 0.05$, $t(43) = 2.63$, $p = 0.012$, Cohen's $d = 0.40$). The subcortical average was also significantly greater than the cortical average (paired $t(43) = 3.67$, $p < 0.001$, Cohen's $d = 0.55$). Note that the compositional-conjunctive analysis was the only one of the three to estimate meaningful within-region associations (i.e. a non-zero matrix diagonal; Fig. 9E), due to associating two different rule representation

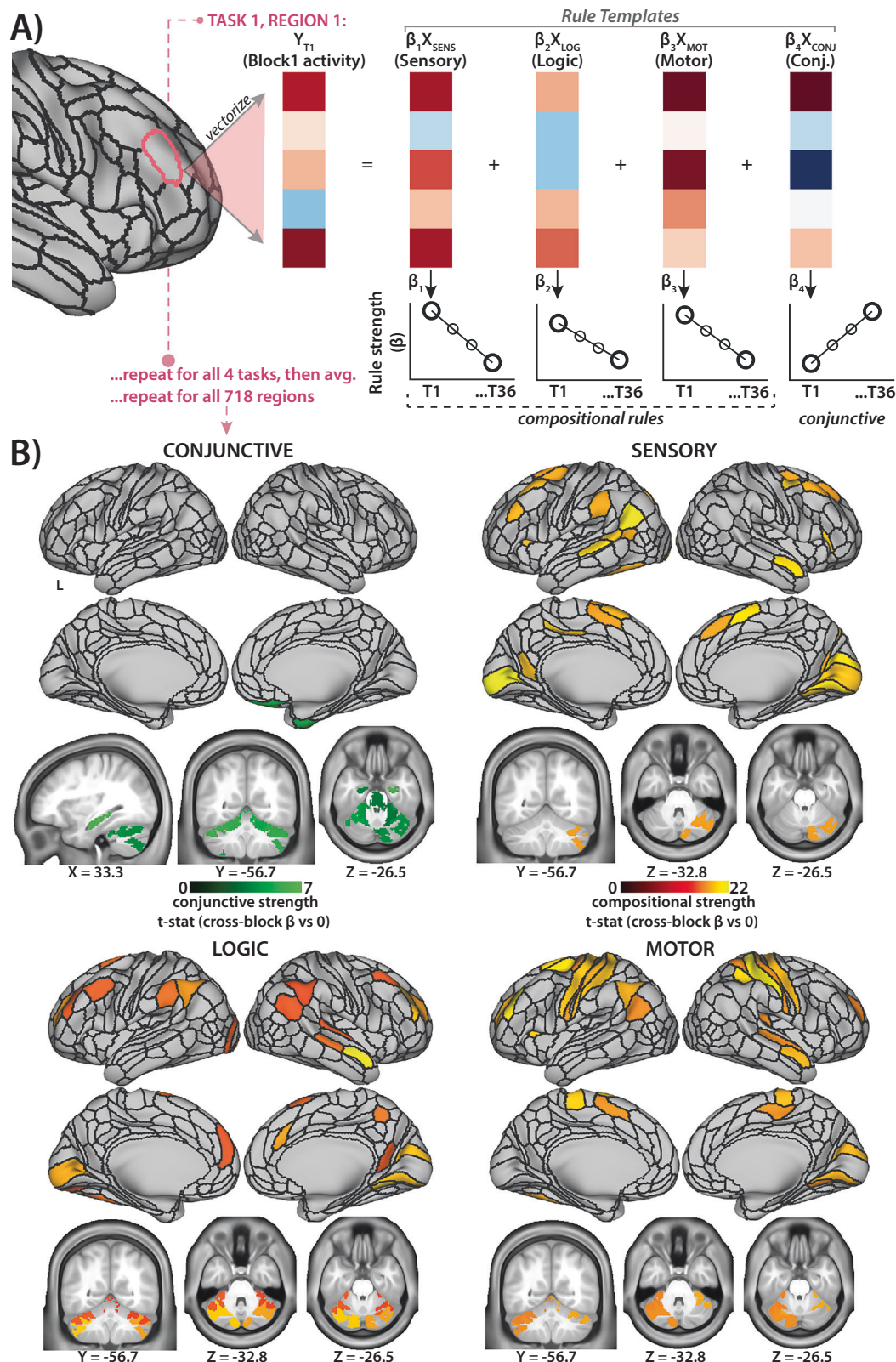


Fig. 4 | Cross-block rule strength analysis reveals differential sensitivity of subcortex and cortex to conjunctive and compositional codes. **A** Schematic of our machine learning approach to detecting rule representation strengths over learning (see text for details). **B** Applying the approach to quantify overall (cross-block averaged) conjunctive (top left) and compositional (clockwise from top right) rule strengths, for cortex and subcortex. To ease visualization of positive region peaks, the compositional brain plots are thresholded as the top 5% t-stats, as many regions survived FDR $p < 0.05$ correction. The conjunctive plot yielded more sparse significance and is hence depicted at the standard FDR $p < 0.05$ threshold.

Conjunctive codes were strongest in subcortical regions (bilateral hippocampus, bilateral cerebellum, bilateral diencephalon, bilateral brainstem, left thalamus), with compositional codes detectable in cortex as well as subcortex. Within subcortex, bilateral cerebellar regions were most prominent in representing compositions (surviving the top 5% thresholding here; see Supplementary Data 1 for regional statistics). Sensible peaks were obtained for each single rule type (e.g. bilateral motor cortex regions for the motor rule), validating the analytic approach. For panel B, sample $n = 44$, with two-sided one-sample t-tests conducted. Brain underlays were generated using Connectome Workbench³⁸.

strength timeseries (compositional and conjunctive; see Methods). This revealed general anticorrelation between the two representation types (on-diagonal blue values in the Fig. 9E matrix), recapitulating the opposing within-region compositional versus conjunctive trajectories seen in Fig. 5.

The pattern of findings was highly similar when using an alternative association method (cross-validated PC regression^{66,69}), with cortex again significantly more prominent in propagating compositional-compositional information (cortical average $t(43) = 23.4$, $p < 0.001$, Cohen's $d = 3.53$; subcortical average $t(43) = 13.9$, $p < 0.001$, Cohen's $d = 2.10$; paired $t(43) = 8.86$, $p < 0.001$, Cohen's $d = 1.34$), and subcortex significantly more prominent in transforming compositions into conjunctions (cortical average $t(43) = -1.45$, $p = 0.156$, Cohen's $d = 0.22$; subcortical average $t(43) = 0.94$, $p = 0.354$, Cohen's $d = 0.14$; paired $t(43) = 2.47$, $p = 0.018$, Cohen's $d = 0.37$). The findings are therefore robust to analytic variation, and substantiate our prediction that subcortical regions are critical to cross-region binding of task-general compositions into task-specialized conjunctions. This again highlights a division of labor between cortex (propagating the same kind of compositional information across the brain) and subcortex (binding compositions into conjunctions).

Discussion

We assimilated concepts from multiple separate neuroscientific subfields to test a neurocomputational framework of task learning (Fig. 1). This formalized a dynamic shift from compositional to conjunctive representations over practice. We developed methods to quantify this shift in the functioning human brain as multi-task learning progressed (Fig. 2A). The findings support our hypotheses: compositional representations are strongly engaged early in learning to facilitate transfer of prior knowledge to novel contexts, while progressive strengthening of cortical conjunctions drives effective practice (behavioral optimization and reduced interference). Shifting from task-general compositional representations to task-specific conjunctive representations hence serves as a key neural underpinning of successful practice.

We firstly demonstrated decreases in neural representational similarity with task practice, extending pattern separation computations beyond the spatial memory and visual object recognition domains of typical study^{30,70,71} to cognitive task learning. We then evidenced a precise computational basis for such changes: increasing nonlinear conjunction (binding) of each task's rule components. Conjunctions have long been considered an important neurocomputational function of long-term memory^{25,27,29–32}, and are central to complementary learning systems theory⁴⁰. Here we used machine learning to dynamically quantify the strength of conjunctive and compositional codes in empirical data, taking inspiration from recent analytic innovations in context-dependent decision-making^{54,55,59,62}. Whilst this literature has demonstrated the importance of conjunctions in within-trial dynamics underlying behavior, here we highlight their role in optimizing human behavior at the longer timescale (minutes/hours) of task learning.

Conversely, compositional representations, previously highlighted as a basis for flexible knowledge transfer^{13–15,18,19,21,54}, were prominent early in learning when tasks were novel, and weakened thereafter. The Supplement (Fig. S7 and S8) also provides evidence for a more general link between compositional representations and high demand conditions (novelty and lapses in task performance). Our findings add to emerging evidence across subfields of a fundamental trade-off between encoding and retrieving generalities of experience (compositional coding), and encoding and retrieving specifics of experience (conjunctive coding^{17,32,40,72,73}). This trade-off provides a neurocomputational underpinning to the transition from controlled to automatic processing^{1,3,9,10}.

We consistently observed differential involvement of cortex and subcortex in the interplay between compositional and conjunctive

codes. In the present timeframe (minutes/hours, relatively early in learning), cortex more strongly represented compositions, whereas subcortex more strongly represented conjunctions. This substantiates the division of labor posited by complementary learning systems theory^{40,44,45}, whilst adding computational detail through distinguishing conjunctive and compositional codes (the latter being neglected in the original theory). Elucidating representational dynamics was key in specifying this division of labor: compositions were engaged early in both cortex and subcortex and then weakened; conjunctions were engaged early in subcortex and then marginally weakened, whilst conjunctions in cortex slowly strengthened. The slow strengthening of cortical conjunctions was uniquely associated with behavioral improvement and pattern separation, highlighting it as a key signature of effective practice. Importantly, this strengthening continued into a second task learning episode (Fig. 8A–B), and was distinct from dynamics of compositional representations in this period (Fig. S7). 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.

Our findings also linked rule representation dynamics to behavioral costs of task switching^{74–78}. Increased task switching during practice worsened behavior (Fig. 8C, Fig. S9A), and these behavioral costs were appropriately and uniquely indexed by FPN conjunctive strength (Fig. 8D, Fig. S10). Our findings suggest that task switching disrupts the time-intensive (and behaviorally critical) process of strengthening cortical conjunctions over practice, leading to increased representational interference and behavioral costs. This contrasts with learning under reduced task switching conditions, which allows more uninterrupted time for cortical conjunctive strengthening. These findings again portray cortical conjunctive strengthening as a key neurocomputational signature of effective learning, with relevance in explaining performance limitations arising from task switching.

Beyond the clear differentiation between cortex and subcortex over learning, we also observed differentiable large-scale network involvement. Network-averaged results consistently highlighted the higher-order cortical networks (DMN, CCNs and higher-sensory networks) as more prominently showing the representational shift over learning, with relatively weaker effects (more stable representations) in primary sensory-motor networks (VIS1, AUD, MOT). Whilst the memory function of the DMN is increasingly well-established⁷⁹, a role for CCNs and higher sensory networks in learning computations (e.g. pattern separation) has recently been acknowledged⁸⁰, building on earlier work linking these networks to general cognitive control functions^{63,66,81}, and control operations impacting long-term memory^{42,82–84}. Future work will be necessary to elucidate the precise nature of interactions involving the higher networks (i.e. within them and extending to the primary networks and subcortex), which give rise to the representational dynamics we observed^{85,86}. The centrality of CCNs to these interactions is supported by FPN conjunctions showing the strongest associations with behavioral learning effects.

Here, our main aim was to capture the distinct representational geometries of compositional and conjunctive representations, and their differential engagement over task practice. However, although these two formats are representationally distinct, they are computationally linked through the process by which compositional elements are nonlinearly bound to construct specialized conjunctions (Fig. 1A). We provided preliminary evidence for this compositional-conjunctive binding, by examining cross-region synchronization of representational dynamics (Fig. 9). Whilst cortical networks were prominent in propagating compositional rule information, subcortical regions were prominent in transforming/binding compositions into conjunctions. The Supplement (Fig. S5) presents simulations consistent with a network mechanism wherein early task-general compositional engagement is progressively sparsified around task-specific sub-circuits, thereby boosting efficiency of information processing from stimulus

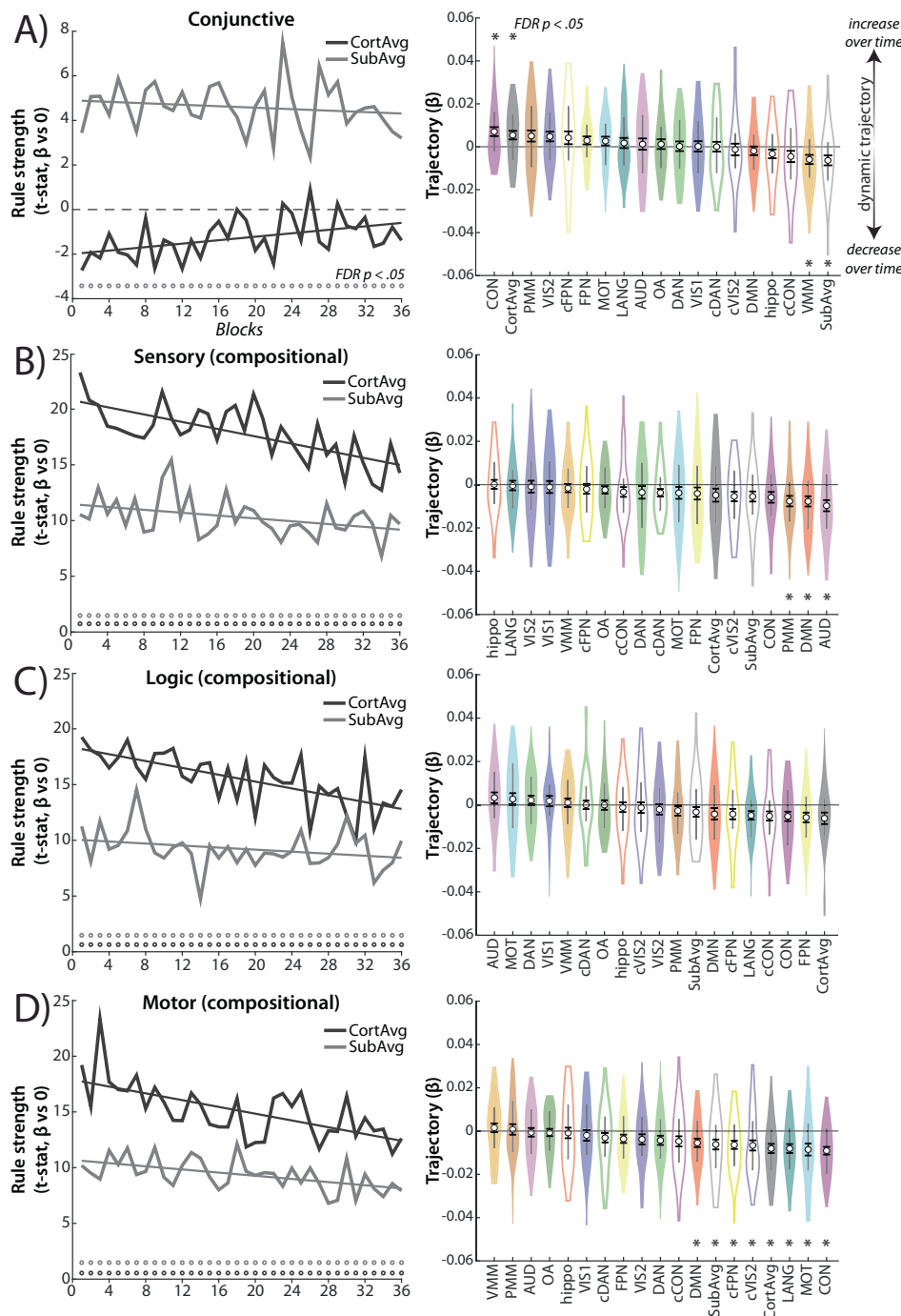


Fig. 5 | Replacement of compositional with conjunctive codes in cortex is a key computational signature of task learning. Left columns plot rule strength timecourses (estimated per Fig. 4A) for the different rule types. To simplify visualization, depicted timecourses are the average across higher cortical networks (black line; see Methods) and the average across subcortical regions (gray line), with linear trend lines overlaid. Dots denote significant rule strength (via two-sided one-sample t -tests against 0 at each block, multiple comparison-corrected at FDR $p < 0.05$). Right columns provide the dynamic trajectory analysis results for each cortical network and subcortical region (and their averages), which formalize the timecourse trends. Violin plots follow previous conventions (violin=subject datapoint distribution, white circle=mean, horizontal notches=standard error of the mean, vertical line=interquartile range). Cortical networks are presented with solid violins and subcortical regions with hollow violins. Asterisks denote significant trajectories (via two-sided one-sample t -tests versus 0 for each region/network, at FDR $p < 0.05$). **A** Dynamics of conjunctive rule representations. Whereas conjunctions were detected early in subcortex and reliably weakened over time, they were not present early in cortex but reliably strengthened over time. The dynamic trajectory analysis revealed generally

increasing conjunctive strength in the cortical networks: 9/12 cortical networks were numerically positive, with the CON ($p = 0.002$) and the cortical network average ($p = 0.009$) reaching significance at FDR $p < 0.05$. 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 subcortex, see Methods) and the subcortical average significantly decreasing over time (both $p = 0.008$). See Figure S5 for a simulated demonstration that the observed early negative cortical conjunctive estimates indicate dominance of compositions over conjunctions, with increases in conjunctive estimates over time truly indicating progressive strengthening of the conjunctions. Panels **B–D** provide the compositional (single rule) results. Compositional codes were reliably present in cortex and subcortex, and emerged early. However, compositional codes were significantly stronger in cortex > subcortex (via two-sided paired t -tests conducted at each block, FDR $p < 0.05$). The dynamic trajectory analyses revealed a generally negative trend, with compositional codes weakening over time in cortex and subcortex (both $p < 0.001$). Sample $n = 44$ for all panels.

to response over practice. Many questions remain over precisely how transformation of compositional to conjunctive representations occurs in the brain, over what timescales, and to the extent it relies on local/distal and cortical/subcortical network interactions. Future work seeking answers to these questions may benefit from methods that formally target network interactions, such as recent empirical neural network modeling approaches^{58,66,87,88}.

Whilst we believe our approach to fMRI preprocessing and expanded nuisance regression was rigorous enough to rule out major artifacts (see Methods), future research seeking convergence of our findings across different neural modalities (e.g. source EEG/MEG) would capitalize on the unique artifacts in these data to increase confidence. The higher temporal resolution of source EEG/MEG may also provide insight into learning-induced changes to representational geometry that differ across task instruction and execution phases^{89,90}.

It will also be important to study representational dynamics beyond the limited window of task learning studied here (minutes/hours of practice), and towards a more extended timescale (hours/days of practice). This is especially as we may hone our efficiency in performing critical tasks over a lengthy timescale (days/months/years) in daily life. Whereas the current limited timescale was appropriate in capturing early retrieval of compositions (promoting minimal competency in specific tasks) and later encoding of new cortical conjunctions (promoting efficiency in specific tasks), a more extended timescale may reveal how new compositions are encoded (providing new building blocks for task-general transfer). Whether the conjunctive coding mechanism presented here (nonlinear sparsification of pre-existing compositional representations around task-relevant sub-circuits; Fig. 1C, Figure S5) is sufficient to give rise to this novel flexibility, or whether long-term learning drives emergence of more materially “new” representations/circuits within subcortex⁹¹ or across cortex-subcortex⁹², remains unclear.

Although the present work provides important empirical substantiation of the long-theorized division of labor between cortex and subcortex, future work seeking to separate roles amongst subcortical regions would help advance beyond the relatively coarse “cortex versus subcortex” distinction. Whilst the involvement of the hippocampus in learning has been widely studied, the cerebellum has received less interest, despite being linked to relevant functions of procedural motor skill learning^{46,47} and behavioral automaticity^{48,49}. Our results suggest a precise computational basis for cerebellar contributions to learning, via early encoding of rule compositions and conjunctions. These computations are supported by cerebellum’s dense bidirectional connectivity with MTL and cortical regions^{93–95}. Future work that seeks to differentiate between subcortical regions would 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 over task learning. We believe the neurocomputational signature identified here – progressive cortical conjunctive strengthening that drives pattern separation and behavioral improvement – is a general learning mechanism, with relevance to multiple neuroscientific subfields. 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; FWA number: 00003913, IORG number: IORG0000357).

C-PRO2 stimulus creation and behavioral validation study

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 stimuli⁹ or concrete multi-sensory stimuli (C-PRO; visual lines and auditory tones)⁶⁴, 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^{96–99}.

Visual stimuli were sampled from established image databases, and varied on the following dimensions: young versus old faces (CAL/PAL face database)¹⁰⁰, and indoor versus outdoor scenes (SUN database)¹⁰¹. The young (18–32 years of age) and old (64+ years) faces were taken from the Ebner¹⁰² 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 standardize 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 database¹⁰³, 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 software¹⁰⁴. 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 example: 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

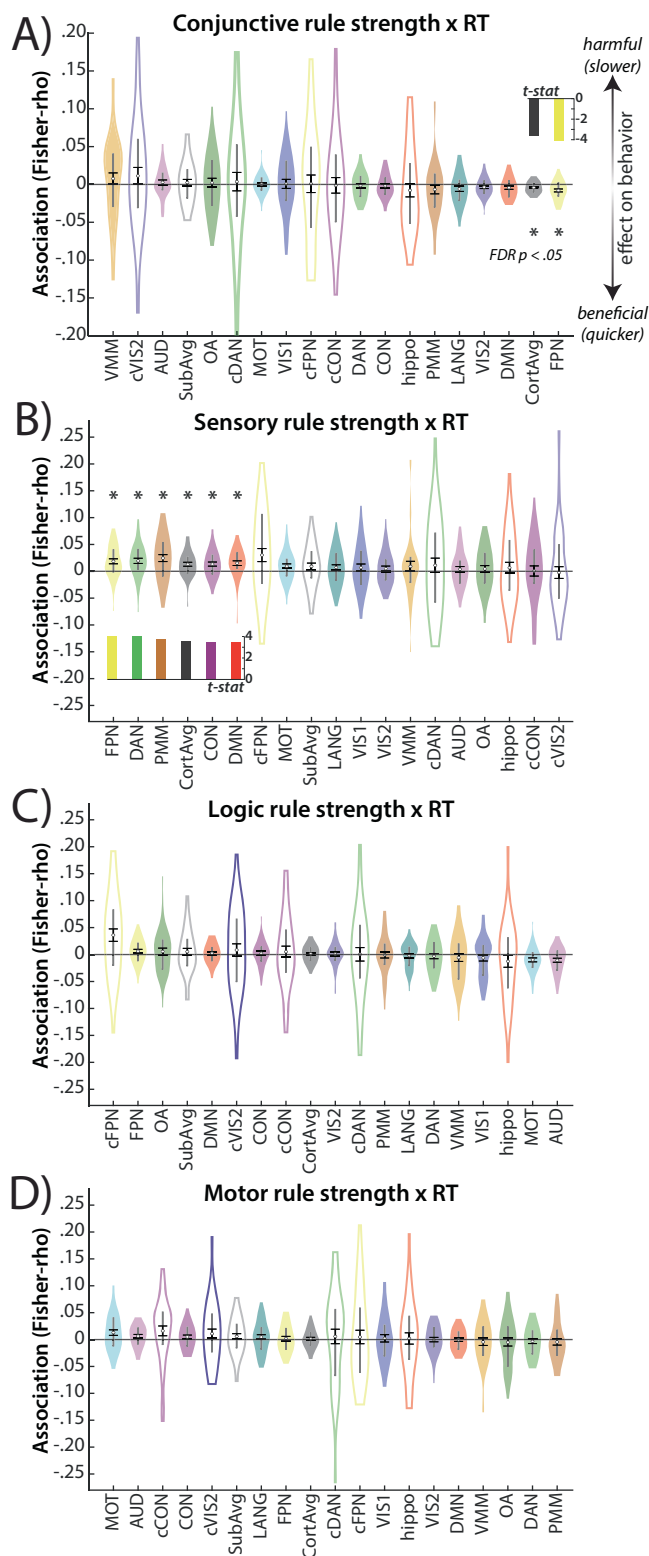


Fig. 6 | Strengthening of cortical conjunctions over learning is associated with behavioral improvement (quickening of RT). Panels display Fisher-transformed Spearman correlations (Fisher-rho) between rule representation strength and behavioral RT, with t-statistics for significant associations inlaid, along with asterisks denoting multiple comparison-corrected significance at FDR $p < 0.05$. Violin plots follow conventions in Fig. 5 (solid violins=cortical regions, hollow violins=subcortical regions). **A** Conjunctive strength was generally numerically associated with beneficial effects on behavior, reaching significance in the FPN cortical network and the cortical network average (both $p < 0.001$, surviving at FDR $p < 0.05$). **B–D** Conversely, maintaining stronger rule compositions over this timescale of extended practice generally yielded harmful behavioral associations (slowing of RT), with this effect most pronounced for sensory rules in the cognitive control networks (CON, DAN, FPN; $p < 0.001$ for all 3 networks), DMN ($p = 0.001$) and PMM ($p < 0.001$). For all panels, sample $n = 44$, with two-sided one-sample t-tests conducted.

arising due to congruence between the male/female visual face stimuli and the male/female auditory voice stimuli.

Stimuli for the main study were created by permuting across these 78×4 possible stimulus combinations, pseudo-randomly at the group level to maximize stimulus variety. This constitutes a highly complex and variable stimulus set that balances the greater naturalism of recent movie-watching studies¹⁰⁵ with greater experimental control. Increasing stimulus variability increased the likelihood of results from this paradigm generalizing to wider contexts. These validated stimuli were included in the full C-PRO2 task (described below).

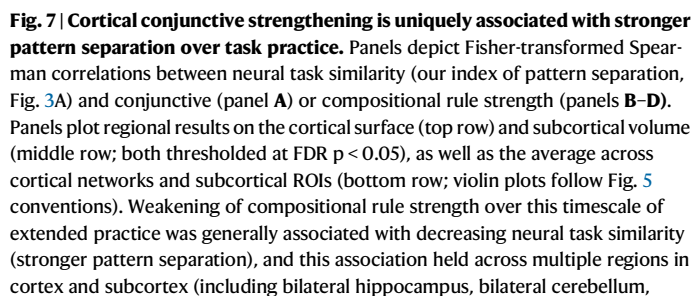
C-PRO2 task design

The C-PRO2 paradigm presents 64 unique tasks per subject, based on permuting 3 rule types (logic, sensory and motor). Each rule type has 4 levels each, meaning that there are 12 rules total (see Fig. 2A for the list of rules). Hence, each of the 64 tasks is defined by its unique combination of 3 rules, one each sampled from the rule types (one sensory, one logic and one motor rule per task). 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 ($16 \times 4 = 64$), 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. Each block began with instruction cue screens that conveyed the rules for the ensuing task, followed by three trials in which subjects applied those rules. Each trial presented two multi-sensory stimulus probes, comprising visual (young/old faces superimposed on indoor/outdoor scenes) and auditory (vocalizations varying in lexical/non-lexical content and normative female/male speakers) stimuli, presented concurrently. Subjects attended to the relevant stimulus dimension based on the cued sensory rule, formed their decision based on the cued logic rule, and responded based on the cued motor rule. How subjects needed to integrate across the various rules to perform an example task trial is detailed in the text above the Probe 1 and Probe 2 events in Fig. 2A. To clarify the response format, for a task with the motor rule “LEFT INDEX”, subjects would make a “TRUE” decision by pressing their left index finger, and a “FALSE” decision by pressing their left middle finger. That is, subjects were instructed to make TRUE/FALSE decisions for a given task by using opposing fingers of the same hand (as cued by the motor rule for that task).

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

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 $p > 0.16$). An additional constraint for the final set of face stimuli was to balance the number of normative male and female faces in each category (totaling 39 male and 39 female face stimuli). This controlled for potential confounds



14

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 (4 practiced tasks, each repeated for 36 blocks). The fMRI Test session involved 10 runs of 12 blocks, amounting to 120 blocks total (each practiced task repeated for $15 \times 4 = 60$ 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 to derive templates for the rule representation strength 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, the order of presentation of the 4 practiced tasks was controlled, 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 “NoSwitch” condition (no switching within a 12-block run). To simplify the majority of our analyses, we collapsed across switch/no switch conditions when analyzing the practiced tasks. We also conducted analyses focused on this task switching manipulation (see later Methods section, Fig. 8C, D, Fig. S9, Fig. S10). 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, given the surplus of novel compared to practiced tasks).

fMRI acquisition

Data were acquired by the 3T Siemens Trio MRI scanner 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 the scanner 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 1 mm 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 Practice/Test sessions for a given subject. 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^{106,107}. 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)¹⁰⁸.

The HCP pipeline outputs preprocessed fMRI timeseries data in a combined CIFTI grayordinate file format, comprising ~60 k 2-D cortical surface vertices and ~30 k 3-D subcortical volume voxels, both at 2 mm resolution, in standard MNI space. The motivation behind HCP's separate treatment of cortex and subcortex has been described in detail previously¹⁰⁷, but in brief: the cortical gray matter is characterized by convoluted folding patterns with minimal gray matter thickness variation between functionally specialized regions, favoring a 2-D surface representation (with units of vertices). In contrast, the subcortical gray matter has less convoluted folding and more prominent variation in gray matter thickness between anatomical regions, favoring a 3-D volumetric representation (with units of voxels). This separate treatment of 2-D cortex and 3-D subcortex has been previously demonstrated to mitigate mixing of signals (e.g. in cortex, 3-D voxels will mix signals over opposing sulcal banks, which is avoided with 2-D vertices)^{109,110}, and allow for system-appropriate cross-subject alignment approaches that improve group-level analysis sensitivity (e.g. the multi-modal MSM-All algorithm for cortex, and nonlinear volumetric warping for subcortex)^{108,111,112}.

The preprocessed CIFTI grayordinate data for each subject (dimensions: ~90k grayordinates x timepoints) were then submitted to general linear modeling (GLM), to clean residual noise from the data and estimate task activations. Separate GLMs were estimated for the fMRI Practice and fMRI Test sessions. We adapted recent best practice recommendations for fMRI nuisance regression of resting-state data¹¹³ to optimally remove artifacts from our task data. Referencing the resting-state literature was motivated by prior reports highlighting greater propensity for artifact contamination in these data (versus time-locking to manipulated task events)^{114,115}, leading to development of more rigorous artifact controls than standard task fMRI approaches. Nuisance regressors included in the task 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)¹¹⁶. By sensitively estimating signals from predominantly non-neuronal areas, the aCompCor regressors have been shown to be effective in reducing a host of physiological (e.g. cardiac, respiratory) and technical confounds in fMRI, when compared to regressing out the averaged white matter/ventricular signals, or regressing out externally recorded physiological signals^{116,117}. Derivatives and quadratics for the 6 motion and 10 physiological regressors were also included, meaning that a total of 64 nuisance regressors were modeled. This constituted a more rigorous and expanded approach to task fMRI nuisance regression than is typically conducted, which aimed to maximally reduce physiological, technical and motion artifacts in our analyses. Critically, each acquired fMRI run was linearly detrended before being entered into the GLM, which removed trend/drift artifacts. We also conducted an exhaustive partial correlation control analysis to rule out any residual temporal

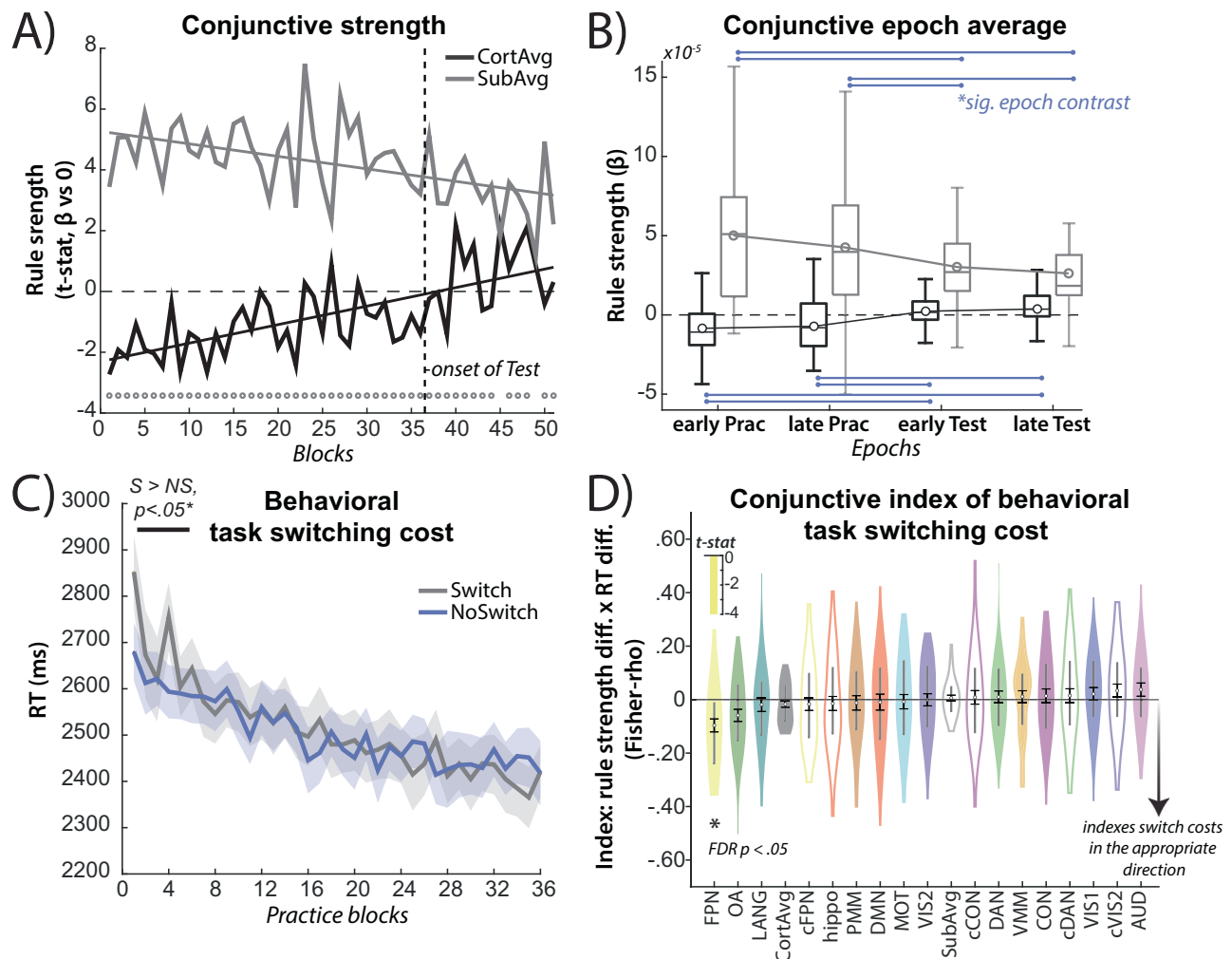


Fig. 8 | Cortical conjunctions strengthen into the second Test session, and index behavioral costs of task switching. **A** Test session conjunction analysis: dynamics of conjunctive rule strength across the first Practice and second Test sessions (onset of the latter marked with a dotted line). Estimation of conjunctive strength and norms of plotting are the same as in Fig. 5. Cortical conjunctions (black line) continue to strengthen and subcortical conjunctions (gray line) continue to weaken with practice. **B** Test session conjunction analysis: epoch averages of the timecourse results in panel A (see Methods for details). Horizontal lines in each box represent the median, circles the mean, edges of each box the inter-quartile range, and whiskers the data range. Blue lines indicate significant paired t-test contrasts between epochs (multiple comparison-corrected at FDR $p < 0.05$; see text for details). These plots formalized the trends in panel A: cortical

conjunctions were significantly stronger in Test versus Practice session epochs; subcortical conjunctions were significantly weaker in Test versus Practice epochs. **C** Task switching analysis: behavioral cost of increased task switching. Reaction time was significantly slower for Switch compared to NoSwitch practiced tasks in the early Practice session epoch ($p = 0.021$). **D** Task switching analysis: brain-behavior association. Asterisks denote significant associations via one-sample t -tests against 0 for each network/region, at multiple comparison-corrected FDR $p < 0.05$. Differences in FPN conjunctive strength across conditions (NoSwitch > Switch) reliably and appropriately indexed the behavioral cost of task switching (NoSwitch RT < Switch RT; $p < 0.001$, surviving at FDR $p < 0.05$). For all panels, sample $n = 44$, with two-sided t -tests conducted (paired and one-sample).

confounds in our brain-behavior association results (see later Methods section).

For both Practice and Test sessions, separate task regressors for each presented block were modeled, as boxcars convolved with the canonical hemodynamic response function, spanning the onset of the first instruction screen to the offset of the last trial for that block. Our decision to model fMRI block activations spanning instruction and trial phases was driven by a desire to capture task set/rule representation processes that extend across both phases, following prior evidence^{118–121}. We also wanted to capitalize on established benefits of modeling extended blocks (versus shorter events) in improving fMRI activation detection sensitivity^{122,123}. 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

Generally, the units of our fMRI analyses were task activations estimated for cortical vertices and subcortical voxels, both at 2 mm scale, and spanning the whole brain. We used within-region vertex/voxel activations as features in our multivariate analyses. Region boundaries were identified from established functional atlases: 360 cortical regions of interest from the Glasser atlas¹⁰⁸, and 358 subcortical regions from the related Cole-Anticevic Brain-wide Network Partition (CAB-NP)⁶⁵. The derivation of each atlas was described in detail in the referenced papers, but in brief: the Glasser cortical regions were defined by aggregating estimates of regional boundaries across multiple MRI-estimated brainmaps (myelin content, cortical thickness, task fMRI activations, resting-state functional connectivity fingerprints), using machine learning methods that were verified manually by human expert anatomists and via cross-validation in held-out data. This multi-modal approach increases confidence in the region

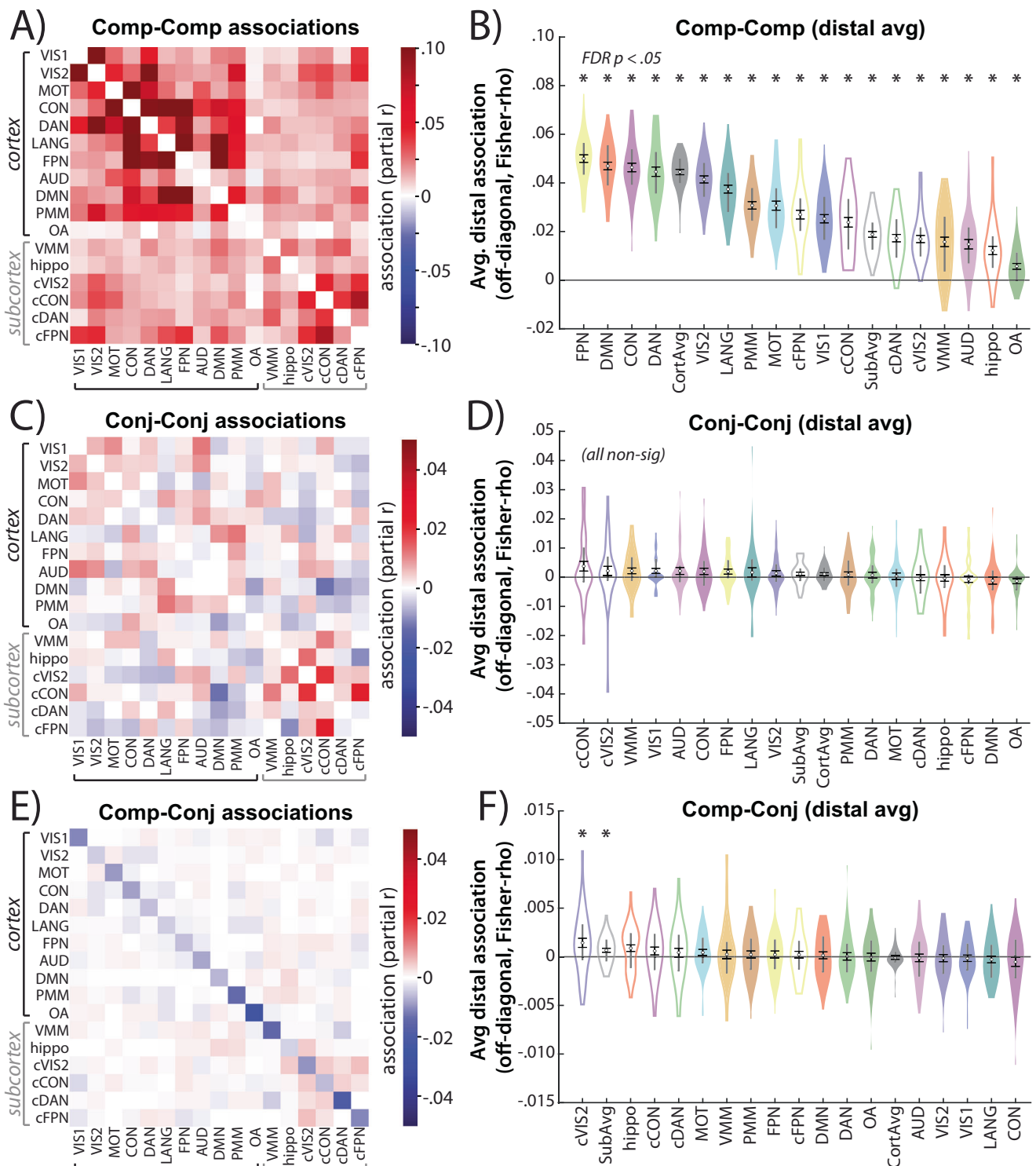


Fig. 9 | Subcortex is critical for cross-region binding of rule information. Cross-region associations are plotted for the various rule-to-rule combinations: compositional-compositional (Comp-Comp, **A, B**) conjunctive-conjunctive (Conj-Conj, **C, D**) and compositional-conjunctive (Comp-Conj, **E, F**). Left panel matrices plot group-averaged associations between cortical networks and subcortical regions. Right panel violin plots depict the averaged distal associations for each network/region, i.e. the average of the off-diagonal elements for each matrix row. These violin plots follow previous conventions (see Fig. 5 and Methods), with

asterisks denoting significance of two-sided one-sample t-tests against 0, multiple comparison-corrected at FDR $p < 0.05$. The results revealed stronger compositional-compositional cross-region associations (red matrix values and positive distal averages) for cortical networks, and stronger compositional-conjunctive associations for the subcortical networks (cVIS2 $p = 0.004$, subcortical average $p = 0.004$, both surviving at FDR $p < 0.05$). The latter results highlight the more prominent role of subcortex in binding compositions into conjunctions. Sample $n = 44$ for all panels.

definitions versus relying on any single modality. The CAB-NP atlas extended the Glasser atlas, firstly by linking each cortical region to 12 large-scale functional networks, which was achieved by hierarchical clustering of region-to-region resting-state functional connectivity. The network clustering was tuned towards clear separation of the primary sensory and motor networks, referencing established anatomical knowledge, and task fMRI activation data in the same subjects. The network definitions were then used to link the ~30k subcortical voxels to larger functional regions by first computing the resting-state functional connectivity between each subcortical voxel and the 360 cortical regions, and assigning each voxel to the network with which it shared the highest average connectivity. The voxel-level results were then spatially clustered into coherent larger subcortical regions. The CAB-NP network assignments for all 360 cortical and 358 subcortical regions were verified via cross-validation in held-out data.

The use of these established regional atlases allowed for a principled means to define the functional region boundaries needed for our machine learning analyses, with greater generalizability of these boundaries (and reduced overfitting) due to their definition in previous, held-out data. These held-out data also spanned much larger samples than the present report, further increasing confidence. These specific atlases also fulfilled our aims of including parcellations of both cortex and subcortex in our analyses, which addressed general trends in memory and decision-making subfields of excluding cortex and subcortex respectively. The CAB-NP's 12 functional network affiliations (see Fig. 3C) were also used to summarize the regional effects at the network level and aid functional interpretations. 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^{124,125}) 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. Similarity timecourses were estimated for practiced tasks in the first (Practice) and second (Test) sessions. The Practice session timecourses were used to quantify dynamic trajectories over time (Fig. 3B, C), as well as to examine associations with task behavior (Fig. S3) and compositional/conjunctive rule strength (Fig. 7, detailed below).

The Test session timecourses were averaged over blocks, and contrasted between practiced and novel tasks (Fig. 3B). For the Test session practiced tasks, this involved averaging neural task similarity over all pairwise practice-to-practice task similarity values, over all 15 block presentations. For the Test session novel tasks (which were only presented once), this involved averaging over all pairwise novel-to-novel task similarity values, after excluding any novel-to-novel task pairs that shared a rule. The latter step was important to ensure that similarity was not spuriously inflated by rule overlap between the novel tasks, thereby maintaining parity with the practiced tasks that by design did not share rules. For example, the similarity estimate between the novel task “YOUNG FACE-SAME-LEFT INDEX” and another novel task “YOUNG FACE-BOTH-LEFT MIDDLE” was excluded from the overall novel similarity average, due to sharing the “YOUNG FACE” sensory rule. The resulting averaged practiced task similarity and

averaged novel task similarity for each region was then contrasted via paired t-test. Note that the number of data points for practiced and novel tasks entering this analysis was balanced, given that both conditions were presented 60 times each at Test.

For the results in Fig. 3B, an analytic conjunction approach was used, in which the practiced <novel task similarity contrast (i.e. regions showing reliably stronger pattern separation for practiced versus novel tasks) was used to filter the Practice session trajectory analysis (i.e. any region showing a reliable positive or negative neural similarity trend over time). This analytic conjunction approach helped isolate regions that showed the strongest pattern separation effects of task learning, whilst further ruling out any contamination by temporal artifacts.

Estimation of compositional and conjunctive rule strength

To interrogate precise changes in neural representational geometry that occurred during task learning, we developed a machine learning approach to quantify the relative contributions of compositional (single-rule) versus conjunctive rule representations over time (Fig. 4A). This took inspiration from other multivariate approaches developed in context-dependent decision-making^{54,59}. For a to-be-analyzed practice task (in the Practice session), we created compositional representation templates for each constituent rule (one each from sensory, logic and motor rule types) by averaging over relevant novel task activations in the held-out Test session. Note that availability of the novel tasks in the Test session was critical in enabling estimation of the compositional rule templates. For example, estimating the compositional representation for the “YOUNG FACE” rule (Fig. 2A) with only access to the practiced tasks would result in a template that confounded the rule representation with the task representation, given that this rule was only presented in one practiced task context in the first Practice session. With access to the novel tasks in the second Test session, which presented “YOUNG FACE” in multiple distinct novel tasks, a compositional template for this rule could be created by averaging over all presentations of that rule in the novel tasks. This allowed estimation of compositional rule templates that were not confounded by the practiced task representations. Finally, a conjunctive template was created by multiplying the 3 compositional 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:

$$Y_t = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 * X_2 * X_3 + C \quad (1)$$

where Y_t is the activation vector for the practiced task on that block, $X_{1,3}$ 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 β_{1-4} therefore quantify the strength of each rule type in the activity representation evoked in that block. Note that we excluded practiced tasks presented in the Test session from data that were averaged to create each rule template, so as to prevent any unintended biases (e.g. inflating detection of the conjunctions, which likely were more strongly evoked in those practiced tasks).

We used robust regression to estimate the beta coefficients, given its improved resilience against outliers compared to regular least squares regression¹²⁶. 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 strength of each rule representation during each block. Repeating over blocks and practiced tasks yielded timecourses of rule representation strength for each subject. The timecourses were averaged across the 4 practiced tasks, with the process repeated for each region to derive subject-level rule strength estimates.

Note that by estimating rule representation templates in held-out data, which are then fit to first-level regional activation patterns, our approach could be conceptualized as estimating/fitting spatial encoding models^{86,124}. Encoding models likely provide stronger tests of representational hypotheses than decoding models or representational similarity analysis, given the specification of particular spatial activation patterns as hypotheses rather than more abstract decoding hyperplanes or representational similarity trends. Another beneficial feature of our approach is that we estimated the representation templates using fMRI data from an entirely different session, thereby going beyond previous calls for “cross-run” multivariate analyses, towards implementing a “cross-session” approach. This addresses temporal proximity confounds that have been previously highlighted in representational geometry analyses^{127,128}.

For the overall rule strength results in Fig. 4B, we averaged the subject-level 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 (Figs. 5–9). Subcortical ROI selection was identified a priori as an important analysis step, following previous evidence of reduced fMRI signal-to-noise in subcortex versus cortex^{129,130}, with this imbalance ameliorated by an ROI approach¹³¹. Selected ROIs were subcortical regions showing the strongest pattern separation and conjunctive strength effects, with regional significance >0 in both the neural task similarity (Fig. 3B) and overall conjunctive strength (Fig. 4B) analyses (both FDR $p < 0.05$). The Supplement (Table S1 and Fig. S4) details the 5 subcortical ROIs selected, with accompanying network information from the CAB-NP atlas.

In another attempt to ensure sufficient signal-to-noise, the analyses in Figs. 5–9 also averaged cortical region results into each of the 12 CAB-NP networks. These cortical network averages excluded all subcortical regions affiliated with each network, to ensure clean separation between cortical and subcortical contributions. The averages across cortical networks (CortAvg) and subcortical regions (SubAvg) were also included in Figs. 5–9 to aid results visualization (e.g. rule strength timecourses for the cortical and subcortical averages were presented in Fig. 5, rather than presenting all 17 timecourses in a single plot). The ventromedial multimodal (VMM) network was included in the subcortical average, given that the bilateral perirhinal regions within this network are intimately linked to the function of subcortical MTL regions^{45,132}. The cortical network average focused on the 6 networks linked to higher cognitive functions - the cognitive control networks (FPN, CON, DAN), default mode network (DMN) and the higher sensory networks (VIS2, LANG) - following prior work identifying cognitive control-related representational geometry in these networks^{58,63}. Multiple comparison correction of all relevant statistics included these cortical and subcortical averages in the adjustments.

To interrogate the association between pattern separation (neural task similarity) and rule strength (Fig. 7), Spearman's correlation was computed between the block-wise neural task similarity timecourse and the rule strength timecourses (3 compositional and 1 conjunctive, 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.

We also extended estimation of the compositional and conjunctive representations into practice task blocks presented in the second Test session. Each of the 4 practiced tasks were repeated a further 15 times in this session. The rule templates were once again calculated by averaging over relevant held-out novel tasks in the Test session, to prevent circularity. Visual impressions of the resulting rule strength timecourses were formalized by averaging the estimates over 6-block epochs spanning early Practice (blocks 1-6), late Practice (blocks 31-36), early Test (blocks 37-42) and late Test (blocks 46-51).

We followed similar analytic conventions when estimating compositional and conjunctive rule strength separately for the task switching sub-conditions of the practiced tasks (2 Switch tasks versus 2 NoSwitch tasks, see earlier Methods for design details, see Fig. 8C, D for results). Compositional and conjunctive rule strength was estimated across Practice session blocks per the approach in Fig. 4A, separately for each of the 4 tasks, with the resulting timecourses averaged into Switch and NoSwitch conditions before subsequent analysis. These timecourses were presented visually in Fig. S9B, and also summarized in the text as the epoch-averaged rule strength (early i.e. block 1-6, or late i.e. block 31-36), as above. Construction of the brain-behavior index linking cortical conjunctions to task switching behavioral costs is detailed in the later “Behavioral analyses” section.

Dynamic trajectory analysis

For both the neural task similarity and rule strength analyses, trends over time in the observed block-to-block timecourses were quantified using a simple linear regression approach⁸⁴:

$$Y = \beta X + C \quad (2)$$

where Y is the neural measure (neural task similarity or compositional/conjunctive rule strength), 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 timecourses¹²⁶. Dynamic trajectories were estimated for each individual subject's neural timecourses, and then submitted to group random effects statistics. This quantitative approach provided a more statistically formalized means of assessing representational trends over time, versus relying on visual inspection alone.

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. We plotted behavioral timecourses across chronological blocks for accuracy (Fig. 2C) and RT (Fig. 2D) to verify recovery of standard learning effects (improved performance with practice). 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 brain-behavior associations computed Spearman's rank correlation between the relevant blockwise imaging measure (neural task similarity or rule representation strength), separately with behavioral accuracy and behavioral RT. For the neural task similarity measure (Fig. S3), 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 strength (Fig. 6), 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 prior to random effects group statistics.

For the task switching analyses, we firstly calculated behavioral costs of switching, by contrasting RT (Fig. 8C) and accuracy (Fig. S9A) between Switch versus NoSwitch tasks. Trends in the behavioral timecourses were summarized via epoch averages that followed previous conventions (early: block 1-6, late: block 31-36). We also computed a brain-behavior association index separately for each cortical network and subcortical region (Fig. 8D). This linked blockwise differences in conjunctive rule strength between task switching conditions to blockwise differences in behavior (RT) between the same

conditions. For each subject, Spearman's correlation was computed between rule strength (NoSwitch minus Switch) and behavioral RT (NoSwitch minus Switch), such that negative rho values captured an index of behavioral switch costs in the appropriate direction (i.e. quicker RT for NoSwitch versus Switch). These subject-level rho values were again analyzed via a random effects approach at the group level.

We selected Spearman's correlation as our method of brain-behavioral association given its established sensitivity to monotonic relationships that may be linear or non-linear (e.g. exponential/power law relationships). To validate this choice, we conducted stepwise regression modeling of behavioral RT to examine whether linear and/or non-linear temporal relationships were observable in the behavioral data. This analysis predicted subject-level RT timecourses from linear and non-linear (power law) variables, fit in separate models, and simultaneously in a combined model. The coefficient of determination (R^2) was used to compare model accuracy across these variations (linear only, non-linear only and combined). The exponent of the non-linear predictor was estimated flexibly for each subject by first fitting a linear regression to log-transformed linear time (X) and RT (Y) variables.

We also conducted a control analysis involving partial Spearman's correlation, which regressed out the linear time variable prior to computing the various brain-behavioral associations (accompanying Figure S3, Fig. 6 and Fig. 8D). Considering that we had already removed the linear trend and various technical and physiological artifacts via expanded nuisance regression, this additional conservative step was an attempt to exhaustively rule out any residual temporal artifacts in the brain-behavior associations.

Cross-region association of rule representation dynamics

For each subject, we computed cross-region associations between rule representation timeseries estimated over Practice session blocks (Fig. 4A). Our chosen association measure was graphical LASSO (GLASSO) - a multivariate method based on partial correlation that can identify direct associations from indirect and confounded associations, unlike alternative methods (e.g. Pearson correlation). It also uses L1 regularization to prevent overfitting to noise (unlike alternative methods, e.g. unregularized partial correlation), by adding a cost function that penalizes extreme estimates. These features were important to our goal of estimating direct associations amongst multiple variables (rule representation timeseries for 12 cortical networks plus 5 subcortical regions), whilst accounting for indirect associations between those variables and minimizing noise. The benefits of estimating direct (versus indirect/confounded) cross-region neural associations have been expounded previously¹³³. The specific GLASSO method has been extensively used in the machine learning field¹³⁴, and was recently extended to human neural functional connectivity⁶⁸, yielding improved inter-session reliability and validity compared to other methods. Whilst that study computed associations amongst "first-level" fMRI BOLD timeseries, we anticipated similar benefits when computing associations amongst "second-level" rule representation timecourses. In this way, our approach can be considered a variation of previous informational connectivity approaches¹³⁵.

Publicly released toolbox code was used to estimate cross-region associations via GLASSO (<https://colelab.github.io/ActflowToolbox/>, which adapted the Python package GGLasso¹³⁶). Each GLASSO model fit rule strength timeseries for all cortical networks and subcortical regions simultaneously. The lambda1 hyperparameter determining the degree of regularization was estimated by gridsearch over a default range of values (0.001–0.316, log-spaced) for each subject, with 6-fold crossvalidation (ensuring balanced train/test sets across folds). This selected the lambda1 value yielding the highest accuracy in predicting held-out rule representation strength timepoints for target networks/regions (Y), from fitting the representation timeseries of source network/regions (X).

Separate GLASSO models were run for the various combinations of rule types: compositional-compositional, compositional-conjunctive and conjunctive-conjunctive. Each yielded a 17-by-17 adjacency matrix of cross-region association estimates (partial correlation r values), which was symmetric across the diagonals. Off-diagonal matrix elements captured cross-region associations of rule representation dynamics, between cortical networks and subcortical regions (defined in Fig. S4). For the compositional-compositional analysis, we estimated the associations separately for each rule type (sensory-sensory, logic-logic and motor-motor) and then averaged them before computing group-level statistics. Similarly, for the compositional-conjunctive analysis, we estimated the associations separately for each compositional rule type (sensory-conjunction, logic-conjunction, motor-conjunction) and then averaged them. Averaging over compositional rule types was supported in both cases by the results being very similar for the separate rules.

The resulting adjacency matrices were visualized after group averaging (Fig. 9 left panels), with the cross-region (distal) associations summarized by taking the average of off-diagonal elements for each cortical network/subcortical region (i.e. each row in the subject's matrix, excluding the diagonal; Fig. 9 right panels). Note that whilst the compositional-compositional and conjunctive-conjunctive analyses targeted propagation of the same type of compositional rule representation across regions, the compositional-conjunctive analysis targeted transformation (binding) of rule representations across regions, through associating two different rule representation types (compositions and conjunctions). The compositional-conjunctive analysis was also unique in having a meaningful diagonal, which captured within-region associations between the two distinct rule types. The matrix diagonal for the other two analyses (compositional-compositional and conjunctive-conjunctive) captured associations of the same rule representation type within the same network/region, and was therefore all ones (which were manually set to zero to aid visualization).

Statistical analysis

All presented statistics – neural task similarity rho values, rule representation strength betas, behavioral association rho values, and rho values for associations of neural task similarity and rule strength – used 2-sided one-sample t -tests against 0 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. Statistics were corrected for multiple comparisons across regions, networks or block timepoints (where appropriate) using the false discovery rate procedure ($FDR\ p < 0.05$)¹³⁷. Note that all brain-behavior association p -values were corrected for the full set of multiple comparisons across both accuracy and RT, to prevent inflation of experimenter degrees of freedom.

Supplementary Data 1 contains a spreadsheet listing 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>). The Connectome Workbench software¹³⁸ was used for visualization of the regional results. Violin plots were used to visualize network- and region-averaged effects, in which the colored violin was a kernel density estimation of the subject datapoint distribution, the white circle was the mean, the horizontal notches the standard error of the mean and the gray vertical line the interquartile range.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Preprocessed fMRI data and behavioral data used in this report are accessible via the Open Science Foundation (<https://osf.io/fjba4/>). For figures plotting results at the brain region level, Supplementary Data 1 provides a list of significant regions and accompanying statistics. Source data are provided with this paper.

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

- Schneider, W. & Chein, J. M. Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cogn. Sci.* **27**, 525–559 (2003).
- Schneider, W. & Shiffrin, R. M. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* **84**, 1–66 (1977).
- Chein, J. M. & Schneider, W. Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. *Cogn. Brain Res.* **25**, 607–623 (2005).
- Hampshire, A. et al. Probing cortical and sub-cortical contributions to instruction-based learning: Regional specialisation and global network dynamics. *NeuroImage* **192**, 88–100 (2019).
- Kami, A. et al. Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature* **377**, 155–158 (1995).
- Poldrack, R. The neural basis of visual skill learning: an fMRI study of mirror reading. *Cereb. Cortex* **8**, 1–10 (1998).
- Ruge, H. & Wolfensteller, U. Rapid Formation of Pragmatic Rule Representations in the Human Brain during Instruction-Based Learning. *Cereb. Cortex* **20**, 1656–1667 (2010).
- Bassett, D. S., Yang, M., Wymbs, N. F. & Grafton, S. T. Learning-induced autonomy of sensorimotor systems. *Nat. Neurosci.* **18**, 744–751 (2015).
- Cole, M. W., Bagic, A., Kass, R. & Schneider, W. Prefrontal Dynamics Underlying Rapid Instructed Task Learning Reverse with Practice. *J. Neurosci.* **30**, 14245–14254 (2010).
- Mohr, H. et al. Integration and segregation of large-scale brain networks during short-term task automatization. *Nat. Commun.* **7**, 13217 (2016).
- Chomsky, N. *Syntactic Structures*. 116 (Mouton, Oxford, England, 1957).
- Montague, R. Universal grammar. *Theoria* **36**, 373–398 (1970).
- Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
- McClelland, J. L. & Rumelhart, D. E. Distributed memory and the representation of general and specific information. *J. Exp. Psychol. Gen.* **114**, 159–188 (1985).
- Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
- Baron, S. G. & Osherson, D. Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage* **55**, 1847–1852 (2011).
- Behrens, T. E. J. et al. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).
- Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W. & Braver, T. S. Rapid Transfer of Abstract Rules to Novel Contexts in Human Lateral Prefrontal Cortex. *Front. Hum. Neurosci.* **5**, (2011).
- Collins, A. G. E. & Frank, M. J. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).
- Reverberi, C., Görgen, K. & Haynes, J.-D. Compositionality of Rule Representations in Human Prefrontal Cortex. *Cereb. Cortex* **22**, 1237–1246 (2012).
- Cole, M. W., Laurent, P. & Stocco, A. Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cogn. Affect. Behav. Neurosci.* **13**, 1–22 (2013).
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
- Palatucci, M., Pomerleau, D., Hinton, G. E. & Mitchell, T. M. Zero-shot Learning with Semantic Output Codes. in *Advances in Neural Information Processing Systems* (eds. Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. & Culotta, A.) vol. 22 (Curran Associates, Inc., 2009).
- Ito, T. et al. Compositional generalization through abstract representations in human and artificial neural networks. in *Advances in Neural Information Processing Systems* (eds. Koyejo, S. et al.) vol. 35 32225–32239 (Curran Associates, Inc., 2022).
- Wickelgren, W. A. 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. *Psychol. Rev.* **86**, 44–60 (1979).
- Ranganath, C. Binding Items and Contexts: The Cognitive Neuroscience of Episodic Memory. *Curr. Dir. Psychol. Sci.* **19**, 131–137 (2010).
- Shimamura, A. P. & Wickens, T. D. Superadditive memory strength for item and source recognition: The role of hierarchical relational binding in the medial temporal lobe. *Psychol. Rev.* **116**, 1–19 (2009).
- Yonelinas, A. P., Ranganath, C., Ekstrom, A. D. & Wiltgen, B. J. A contextual binding theory of episodic memory: systems consolidation reconsidered. *Nat. Rev. Neurosci.* **20**, 364–375 (2019).
- Eichenbaum, H. What H.M. Taught Us. *J. Cogn. Neurosci.* **25**, 14–21 (2013).
- Marr, D. Simple memory: a theory for archicortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **262**, 23–81 (1971).
- Sutherland, R. J. & Rudy, J. W. Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology* **17**, 129–144 (1989).
- Radulescu, A., Shin, Y. S. & Niv, Y. Human Representation Learning. *Annu. Rev. Neurosci.* **44**, 253–273 (2021).
- Brooks, D. & Baddeley, A. What can amnesic patients learn?. *Neuropsychologia* **14**, 111–122 (1976).
- Corkin, S. Lasting Consequences of Bilateral Medial Temporal Lobectomy: Clinical Course and Experimental Findings in H.M. *Semin. Neurol.* **4**, 249–259 (1984).
- Squire, L. R. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* **99**, 195–231 (1992).
- Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21 (1957).
- Hirsh, R. The hippocampus, conditional operations, and cognition. *Physiol. Psychol.* **8**, 175–182 (1980).
- Manns, J. R. & Eichenbaum, H. Evolution of declarative memory. *Hippocampus* **16**, 795–808 (2006).
- Nadel, L., Samsonovich, A., Ryan, L. & Moscovitch, M. Multiple trace theory of human memory: Computational, neuroimaging,

- and neuropsychological results. *Hippocampus* **10**, 352–368 (2000).
40. O'Reilly, R. C. & Rudy, J. W. Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychol. Rev.* **108**, 311–345 (2001).
 41. Sherry, D. F. & Schacter, D. L. The evolution of multiple memory systems. *Psychol. Rev.* **94**, 439–454 (1987).
 42. Simons, J. S. & Spiers, H. J. Prefrontal and medial temporal lobe interactions in long-term memory. *Nat. Rev. Neurosci.* **4**, 637–648 (2003).
 43. Kumaran, D., Hassabis, D. & McClelland, J. L. What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
 44. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
 45. O'Reilly, R. C., Bhattacharyya, R., Howard, M. D. & Ketz, N. Complementary Learning Systems. *Cogn. Sci.* **38**, 1229–1248 (2014).
 46. Beukema, P. & Verstynen, T. Predicting and binding: interacting algorithms supporting the consolidation of sequential motor skills. *Curr. Opin. Behav. Sci.* **20**, 98–103 (2018).
 47. Ferrari, C. et al. TMS Over the Cerebellum Interferes with Short-term Memory of Visual Sequences. *Sci. Rep.* **8**, 6722 (2018).
 48. Lang, C. E. & Bastian, A. J. Cerebellar Damage Impairs Automaticity of a Recently Practiced Movement. *J. Neurophysiol.* **87**, 1336–1347 (2002).
 49. Nixon, P. D. & Passingham, R. E. The cerebellum and cognition: cerebellar lesions impair sequence learning but not conditional visuomotor learning in monkeys. *Neuropsychologia* **38**, 1054–1072 (2000).
 50. Gao, J.-H. et al. Cerebellum Implicated in Sensory Acquisition and Discrimination Rather Than Motor Control. *Science* **272**, 545–547 (1996).
 51. King, M., Hernandez-Castillo, C. R., Poldrack, R. A., Ivry, R. B. & Diedrichsen, J. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nat. Neurosci.* **22**, 1371–1378 (2019).
 52. Nakai, T. & Nishimoto, S. Representations and decodability of diverse cognitive functions are preserved across the human cortex, cerebellum, and subcortex. *Commun. Biol.* **5**, 1245 (2022).
 53. Badre, D., Bhandari, A., Keglovits, H. & Kikimoto, A. The dimensionality of neural representations for control. *Curr. Opin. Behav. Sci.* **38**, 20–28 (2021).
 54. Bernardi, S. et al. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954–967.e21 (2020).
 55. Dang, W., Jaffe, R. J., Qi, X.-L. & Constantinidis, C. Emergence of Non-Linear Mixed Selectivity in Prefrontal Cortex after Training. *J. Neurosci.* JN-RM-2814-20 (2021) <https://doi.org/10.1523/JNEUROSCI.2814-20.2021>.
 56. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258–1270.e11 (2022).
 57. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
 58. Ito, T., Yang, G. R., Laurent, P., Schultz, D. H. & Cole, M. W. Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. *Nat. Commun.* **13**, 673 (2022).
 59. Kikimoto, A., Mayr, U. Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proc. Natl Acad. Sci.* **117**, 10603–10608 (2020).
 60. Mack, M. L., Preston, A. R. & Love, B. C. Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* **11**, 46 (2020).
 61. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
 62. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
 63. Cole, M. W. et al. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat. Neurosci.* **16**, 1348–1355 (2013).
 64. Ito, T. et al. Cognitive task information is transferred between brain regions via resting-state network topology. *Nat. Commun.* **8**, 1027 (2017).
 65. Ji, J. L. et al. Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage* **185**, 35–57 (2019).
 66. Mill, R. D. et al. Network modeling of dynamic brain interactions predicts emergence of neural information that supports human cognitive behavior. *PLOS Biol.* **20**, e3001686 (2022).
 67. Schultz, D. H., Ito, T. & Cole, M. W. Global connectivity fingerprints predict the domain generality of multiple-demand regions. *Cereb. Cortex* **32**, 4464–4479 (2022).
 68. Peterson, K. L., Sanchez-Romero, R., Mill, R. D. & Cole, M. W. Regularized partial correlation provides reliable functional connectivity estimates while correcting for widespread confounding. *Imaging Neurosci.* **3**, IMAG.a.162 (2025).
 69. Mill, R. D., Gordon, B. A., Balota, D. A. & Cole, M. W. Predicting dysfunctional age-related task activations from resting-state network alterations. *NeuroImage* **221**, 117167 (2020).
 70. Bakker, A., Kirwan, C. B., Miller, M. & Stark, C. E. L. Pattern Separation in the Human Hippocampal CA3 and Dentate Gyrus. *Science* **319**, 1640–1642 (2008).
 71. Brunec, I. K., Robin, J., Olsen, R. K., Moscovitch, M. & Barense, M. D. Integration and differentiation of hippocampal memory traces. *Neurosci. Biobehav. Rev.* **118**, 196–208 (2020).
 72. Reagh, Z. M. & Ranganath, C. Flexible reuse of cortico-hippocampal representations during encoding and recall of naturalistic events. *Nat. Commun.* **14**, 1279 (2023).
 73. Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W. & Behrens, T. E. J. How to build a cognitive map. *Nat. Neurosci.* **25**, 1257–1272 (2022).
 74. Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci.* **115**, E10313–E10322 (2018).
 75. Musslick, S. & Cohen, J. D. Rationalizing constraints on the capacity for cognitive control. *Trends Cogn. Sci.* **25**, 757–775 (2021).
 76. Sakai, K. Task Set and Prefrontal Cortex. *Annu. Rev. Neurosci.* **31**, 219–245 (2008).
 77. Xu, S., Simoens, J., Verguts, T. & Braem, S. Learning where to be flexible: Using environmental cues to regulate cognitive control. *J. Exp. Psychol. Gen.* **153**, 328–338 (2024).
 78. Yeung, N., Nystrom, L. E., Aronson, J. A. & Cohen, J. D. Between-Task Competition and Cognitive Control in Task Switching. *J. Neurosci.* **26**, 1429–1438 (2006).
 79. Kaefer, K., Stella, F., McNaughton, B. L. & Battaglia, F. P. Replay, the default mode network and the cascaded memory systems model. *Nat. Rev. Neurosci.* **23**, 628–640 (2022).
 80. Amer, T. & Davachi, L. Extra-hippocampal contributions to pattern separation. *eLife* **12**, e82250 (2023).
 81. Cocuzza, C. V., Ito, T., Schultz, D., Bassett, D. S. & Cole, M. W. Flexible Coordinator and Switcher Hubs for Adaptive Task Control. *J. Neurosci.* **40**, 6949–6968 (2020).
 82. Dobbins, I. G., Foley, H., Schacter, D. L. & Wagner, A. D. Executive Control during Episodic Retrieval. *Neuron* **35**, 989–996 (2002).
 83. Mill, R. D., Cavin, I. & O'Connor, A. R. Differentiating the functional contributions of resting connectivity networks to memory

- decision-making: fMRI support for multistage control processes. *J. Cogn. Neurosci.* **27**, 1617–1632 (2015).
84. Mill, R. D., O'Connor, A. R. & Dobbins, I. G. Pupil dilation during recognition memory: Isolating unexpected recognition from judgment uncertainty. *Cognition* **154**, 81–94 (2016).
 85. de-Wit, L., Alexander, D., Ekroll, V. & Wagemans, J. Is neuroimaging measuring information in the brain? *Psychon. Bull. Rev.* **23**, 1415–1428 (2016).
 86. Ito, T., Hearne, L., Mill, R., Cocuzza, C. & Cole, M. W. Discovering the Computational Relevance of Brain Network Organization. *Trends Cogn. Sci.* **24**, 25–38 (2020).
 87. Cole, M. W., Ito, T., Bassett, D. S. & Schultz, D. H. Activity flow over resting-state networks shapes cognitive task activations. *Nat. Neurosci.* **19**, 1718–1726 (2016).
 88. Hearne, L. J. et al. Activity flow underlying abnormalities in brain activations and cognition in schizophrenia. *Sci. Adv.* **7**, eabf2513 (2021).
 89. Baggio, G. et al. Multiple neural representations of elementary logical connectives. *NeuroImage* **135**, 300–310 (2016).
 90. Stokes, M. G. et al. Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron* **78**, 364–375 (2013).
 91. Kempermann, G. et al. Human Adult Neurogenesis: Evidence and Remaining Questions. *Cell Stem Cell* **23**, 25–30 (2018).
 92. Sun, W., Advani, M., Spruston, N., Saxe, A. & Fitzgerald, J. E. Organizing memories for generalization in complementary learning systems. *Nat. Neurosci.* **26**, 1438–1448 (2023).
 93. Buckner, R. L. The Cerebellum and Cognitive Function: 25 Years of Insight from Anatomy and Neuroimaging. *Neuron* **80**, 807–815 (2013).
 94. Gao, Z. et al. A cortico-cerebellar loop for motor planning. *Nature* **563**, 113–116 (2018).
 95. Strick, P. L., Dum, R. P. & Fiez, J. A. Cerebellum and Nonmotor Function. *Annu. Rev. Neurosci.* **32**, 413–434 (2009).
 96. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
 97. Formisano, E., De Martino, F., Bonte, M. & Goebel, R. 'Who' Is Saying 'What'? Brain-Based Decoding of Human Voice and Speech. *Science* **322**, 970–973 (2008).
 98. Kanwisher, N., McDermott, J. & Chun, M. M. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *J. Neurosci.* **17**, 4302–4311 (1997).
 99. Lafer-Sousa, R., Conway, B. R. & Kanwisher, N. G. Color-Biased Regions of the Ventral Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in Macaques. *J. Neurosci.* **36**, 1682–1697 (2016).
 100. Minear, M. & Park, D. C. A lifespan database of adult facial stimuli. *Behav. Res. Methods Instrum. Comput.* **36**, 630–633 (2004).
 101. Xiao, J., Hays, J., Ehinger, K. A., Oliva, A. & Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 3485–3492 (IEEE, San Francisco, CA, USA, 2010). <https://doi.org/10.1109/CVPR.2010.5539970>.
 102. Ebner, N. C. Age of face matters: Age-group differences in ratings of young and old faces. *Behav. Res. Methods* **40**, 130–136 (2008).
 103. Weinberger, S. *Speech Accent Archive*. (2016).
 104. Boersma, P. & Weenik, D. *Praat: Doing Phonetics by Computer*. (2023).
 105. Sonkusare, S., Breakspear, M. & Guo, C. Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends Cogn. Sci.* **23**, 699–714 (2019).
 106. Elam, J. S. et al. The Human Connectome Project: A retrospective. *NeuroImage* **244**, 118543 (2021).
 107. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
 108. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
 109. Anticevic, A. et al. Comparing surface-based and volume-based analyses of functional neuroimaging data in patients with schizophrenia. *NeuroImage* **41**, 835–848 (2008).
 110. Tucholka, A., Fritsch, V., Poline, J.-B. & Thirion, B. An empirical comparison of surface-based and volume-based group studies in neuroimaging. *NeuroImage* **63**, 1443–1453 (2012).
 111. Fischl, B. et al. Whole Brain Segmentation. *Neuron* **33**, 341–355 (2002).
 112. Klein, A. et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage* **46**, 786–802 (2009).
 113. Ciric, R. et al. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* **154**, 174–187 (2017).
 114. Huijbers, W., Van Dijk, K. R. A., Boenniger, M. M., Stirnberg, R. & Bretelet, M. M. B. Less head motion during MRI under task than resting-state conditions. *NeuroImage* **147**, 111–120 (2017).
 115. Laumann, T. O. et al. On the Stability of BOLD fMRI Correlations. *Cereb. Cortex* (2016) <https://doi.org/10.1093/cercor/bhw265>.
 116. Behzadi, Y., Restom, K., Liao, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
 117. Caballero-Gaudes, C. & Reynolds, R. C. Methods for cleaning the BOLD fMRI signal. *NeuroImage* **154**, 128–149 (2017).
 118. Cookson, S. L. & Schumacher, E. H. Dissociating the Neural Correlates of Planning and Executing Tasks with Nested Task Sets. *J. Cogn. Neurosci.* **34**, 877–896 (2022).
 119. Dosenbach, N. U. F. et al. A Core System for the Implementation of Task Sets. *Neuron* **50**, 799–812 (2006).
 120. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
 121. Wilk, H. A., Ezekiel, F. & Morton, J. B. Brain regions associated with moment-to-moment adjustments in control and stable task-set maintenance. *NeuroImage* **59**, 1960–1967 (2012).
 122. Liu, T. T., Frank, L. R., Wong, E. C. & Buxton, R. B. Detection Power, Estimation Efficiency, and Predictability in Event-Related fMRI. *NeuroImage* **13**, 759–773 (2001).
 123. Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N. A. & Dale, A. M. Stochastic Designs in Event-Related fMRI. *NeuroImage* **10**, 607–619 (1999).
 124. Diedrichsen, J. & Kriegeskorte, N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Comput. Biol.* **13**, e1005508 (2017).
 125. Haynes, J.-D. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron* **87**, 257–270 (2015).
 126. Holland, P. W. & Welsch, R. E. Robust regression using iteratively reweighted least-squares. *Commun. Stat. - Theory Methods* **6**, 813–827 (1977).
 127. Alink, A., et al. Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. Preprint at <https://doi.org/10.1101/032391> (2015).
 128. Cai, M. B., Schuck, N. W., Pillow, J. W. & Niv, Y. Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLOS Comput. Biol.* **15**, e1006299 (2019).
 129. De Hollander, G., Keuken, M. C., Van Der Zwaag, W., Forstmann, B. U. & Trampel, R. Comparing functional MRI protocols for small, iron-rich basal ganglia nuclei such as the subthalamic nucleus at 7 T and 3 T. *Hum. Brain Mapp.* **38**, 3226–3248 (2017).
 130. Keuken, M. C., Isaacs, B. R., Trampel, R., Van Der Zwaag, W. & Forstmann, B. U. Visualizing the Human Subcortex Using Ultra-

- high Field Magnetic Resonance Imaging. *Brain Topogr.* **31**, 513–545 (2018).
131. Isherwood, S. J. S., Keuken, M. C., Bazin, P. L. & Forstmann, B. U. Cortical and subcortical contributions to interference resolution and inhibition – An fMRI ALE meta-analysis. *Neurosci. Biobehav. Rev.* **129**, 245–260 (2021).
 132. Tomás Pereira, I., Agster, K. L. & Burwell, R. D. Subcortical connections of the perirhinal, postrhinal, and entorhinal cortices of the rat. I. afferents. *Hippocampus* **26**, 1189–1212 (2016).
 133. Reid, A. T. et al. Advancing functional connectivity research from association to causation. *Nat. Neurosci.* **22**, 1751–1760 (2019).
 134. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
 135. Coutanche, M. N. & Thompson-Schill, S. L. Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Front. Hum. Neurosci.* **7**, (2013).
 136. Schaipp, F., Müller, C. L. & Vlasovets, O. GGLasso - a Python package for General Graphical Lasso computation. Preprint at <https://doi.org/10.48550/ARXIV.2110.10521> (2021).
 137. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
 138. Marcus, D. S. et al. Informatics and Data Mining Tools and Strategies for the Human Connectome Project. *Front. Neuroinformatics* **5**, (2011).

Acknowledgements

The authors acknowledge support by the US National Institutes of Health under award R01-MH109520 to M.W.C. and award R56-MH138448 to M.W.C., and support by the US National Science Foundation under award 2219323 to R.D.M. and M.W.C. 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.

Author contributions

R.D.M. and M.W.C. conceptualized the study design. R.D.M. collected the dataset and analyzed it, under the supervision of M.W.C. R.D.M. and M.W.C. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-65041-2>.

Correspondence and requests for materials should be addressed to Ravi D. Mill.

Peer review information *Nature Communications* thanks Benjamin Rangel, Carlo Reverberi, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025