



Article

<https://doi.org/10.1038/s41593-024-01668-6>

Flexible multitask computation in recurrent networks utilizes shared dynamical motifs

Received: 16 September 2022

Accepted: 26 April 2024

Published online: 9 July 2024

Check for updates

Laura N. Driscoll¹✉, Krishna Shenoy^{1,2,3,4,5,6,7} & David Sussillo^{1,5}

Flexible computation is a hallmark of intelligent behavior. However, little is known about how neural networks contextually reconfigure for different computations. In the present work, we identified an algorithmic neural substrate for modular computation through the study of multitasking artificial recurrent neural networks. Dynamical systems analyses revealed learned computational strategies mirroring the modular subtask structure of the training task set. Dynamical motifs, which are recurring patterns of neural activity that implement specific computations through dynamics, such as attractors, decision boundaries and rotations, were reused across tasks. For example, tasks requiring memory of a continuous circular variable repurposed the same ring attractor. We showed that dynamical motifs were implemented by clusters of units when the unit activation function was restricted to be positive. Cluster lesions caused modular performance deficits. Motifs were reconfigured for fast transfer learning after an initial phase of learning. This work establishes dynamical motifs as a fundamental unit of compositional computation, intermediate between neuron and network. As whole-brain studies simultaneously record activity from multiple specialized systems, the dynamical motif framework will guide questions about specialization and generalization.

Cognitive flexibility is a key feature of the human brain. Although artificial systems are capable of outperforming humans in specific tasks^{1–3}, they so far lack flexibility for rapid learning and task switching. A major open question in the fields of neuroscience and artificial intelligence is how the same circuit reconfigures to perform multiple tasks.

Conceptual models for cognitive flexibility propose a hierarchy of elementary processes that are reused across similar tasks^{4–6}. According to these models, the neural substrate for computation is modular such that combinations of previously learned subtasks may be reconfigured to perform unfamiliar tasks. This combination of subtasks is referred to as compositionality⁶. For example, a saccade task typically involves a cue that indicates in which direction to move the eyes. After learning a saccade task, a person could quickly learn an ‘anti’ version of the same task where the same cue now instructs a

saccade in the opposite direction. This new task may be quickly learned by combining a computational building block for the original task with a previously learned ‘anti’ building block. Although there is some experimental evidence that neural computation is compositional^{7,8}, a concrete model for its implementation hinges on identifying modular components with compositional potential. Although the time and effort required to train animals to perform many tasks has limited the exploration of multitask computation in biological networks, artificial neural networks now present an opportunity to explore the topic. The study of cognition through simulations in artificial networks has led to substantial advances in understanding neural computation in the past decade^{9–19}. However, researchers typically trained artificial neural networks to perform single tasks in isolation, with few exceptions^{20–25}, limiting the insights into biological neural circuits that perform many

¹Department of Electrical Engineering, Stanford University, Stanford, CA, USA. ²Department of Neurosurgery, Stanford University, Stanford, CA, USA.

³Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁴Department of Neurobiology, Stanford University, Stanford, CA, USA.

⁵Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA. ⁶Bio-X Institute, Stanford University, Stanford, CA, USA. ⁷Howard Hughes Medical Institute at Stanford University, Stanford, CA, USA. ✉e-mail: laura.driscoll@alleninstitute.org

tasks. One exception to this trend is the study by Yang et al.²¹, in which the authors trained a single network to perform 20 related tasks and, thereby, identified clustered representations in state space that supported task compositionality. In the present work, we identified the computational substrate that allowed for modular computation in these networks, which we call ‘dynamical motifs’. These motifs are distinctive features of the dynamics that reoccur across different implementations of similar computations.

We examined multitask networks through the lens of dynamical systems. This approach allowed us to explore the mechanisms underlying computation in a recurrently connected artificial network²⁶. We found that tasks that required the same computational elements (for example, memory, categorization and delayed response) were implemented by sharing and repurposing dynamical motifs (for example, attractors, decision boundaries and rotations).

Results

Network structure

We implemented a similar input–output structure and learning protocol as in previously examined multitasking recurrent neural networks (RNNs)²¹. These tasks included reaction-timed, delayed response and memory tasks with contextual integration, categorization, pro response and anti response components (see Extended Data Fig. 1, Supplementary Table 1 and ‘Tasks and performances’ in Methods for task definitions). For every task, the network received three noisy inputs: fixation (one-dimensional), stimulus (four-dimensional) and rule (15-dimensional) (Fig. 1a). The fixation input directed the network to either output zero or respond. The set of stimuli contained two separate two-dimensional vectors composed of $A\sin\theta$ and $A\cos\theta$, where each vector encoded a different one-dimensional circular variable (θ_1, θ_2) scaled by an amplitude (A_1, A_2). Depending on the rule, one stimulus vector may be contextually ignored. The rule input indicated the current task on any given trial, and this information was continuously available to the network throughout each trial. Rule input was encoded in a one-hot vector where the index associated with the current task was 1 and all other indices were 0.

The RNN is defined by

$$\tau \frac{dh}{dt} = -h(t) + \sigma(W_{rec}h(t) + W_{in}u(t) + b_{in} + \xi(t)) \quad (1)$$

$$z(t) = W_{out}h(t) + b_{out} \quad (2)$$

$$\sigma(h) = \ln(1 + \exp(h)) \quad (3)$$

All inputs, $u(t)$ (20×1), enter the system and induce a specific pattern of activity, $h(t)$ ($N_{rec} \times 1$), in the units of the RNN (equation (1)). We refer to this N_{rec} -dimensional vector, $h(t)$, as the state of the network at time t . There was noise in the inputs and independent noise in each unit, $\xi(t)$. The output, $z(t)$ (3×1), is a linear projection of the state (equation (2)). The output units indicate whether the network is responding in the first dimension and in which direction on a circle the RNN responds in the next two dimensions (for example, saccade direction) (Fig. 1a, right). For consistency, in most of this paper, we will focus on RNNs as described by equation (1), using diagonal initialization of W_{rec} , the softplus nonlinear activation function (equation (3)) and L2 activity and weight regularization. We identified shared dynamical motifs in all explored network designs and include comparisons to other parameter choices throughout. All network weights were trained to minimize the squared difference between the network output and a desired target using back propagation through time.

Our approach was to uncover the underlying learned dynamical systems in trained RNNs to mechanistically understand how networks implement computation. This approach utilizes fixed points

of equation (1) to provide an interpretable ‘skeleton’ of the complex high-dimensional dynamics^{26–28}. By studying how fixed points change as a function of the inputs, we may understand if and how fixed point structures are repurposed for different computations. In the absence of noise, inputs to the network were piecewise constant, where every change in the inputs marked the beginning of a new task period (Fig. 1a, vertical lines). Therefore, during each task period with unique inputs (for example, stimulus, context/memory and response), the network could be treated as a separate, autonomous dynamical system with a distinct set of fixed points from other task periods. Going forward, we use ‘dynamical motif’ to mean the high-dimensional nonlinear dynamics around a fixed point skeleton that implements computation for a specified input. See Methods for further details on network setup, training and fixed point analysis.

Single-task networks

We first trained individual networks to perform each task in isolation. For example, in a MemoryPro task, the network should respond in the same direction as the stimulus after a memory period. There were four periods (visually divided by vertical lines Fig. 1a). We quantified the fraction of state trajectory variance explained in the top 10 principal components (PCs) for a given task period and in the top two PCs defined by every other task period (Fig. 1b). This provides a reference for the fraction of variance captured in visualizations (Fig. 1c–h). We visualized the high-dimensional network state trajectories and fixed points in a low-dimensional PC space defined by performing principal component analysis (PCA) on state trajectories during the memory period of the MemoryPro task. In the first period (context), the rule input indicated which task the network performed for that trial. In a network trained to perform only the MemoryPro task, the context period inputs result in one fixed point at the center of a ring of fixed points (Fig. 1c). The central fixed point serves as an initial condition for performing the task computation during the ensuing stimulus period. Notice that the context period inputs are identical to the memory period (rule and fixation on, stimulus off) (Fig. 1a), so the fixed points are necessarily identical between these task periods. We show later that the additional ring of fixed points was relevant to the memory computation during the memory task period.

In the stimulus period, we examined the fixed point structure for each stimulus input separately. Stimulus period state trajectories for different stimuli diverged from the central initial condition toward stimulus-dependent fixed points, mapping out a stimulus representation that was orthogonal (null) to the response readout dimension (Fig. 1d). During the memory period, the state evolved toward a ring of fixed points that made up an approximate ring attractor (locally attracting structure in all dimensions except tangent to the ring, which is neither contracting nor expanding) (Fig. 1e). Although fixed points were identical in the context and memory periods, the network state interacted with different fixed points due to different initial conditions. Together, these fixed points stored the identity of the stimulus orientation based on the initial conditions of the state at the beginning of the memory period (end of the stimulus period). During the response period, the fixation input changed to zero, and a new ring attractor emerged (Fig. 1f). During the response period, the ring was oriented such that it had a non-zero dot product with the output weights (W_{out}) and was, therefore, output potent²⁹. The new ring caused the network to respond in the appropriate orientation based on the initial conditions of the response period (end of the memory period). Thus, the network responded with the appropriate orientation for this task ($\varphi_{response} = \theta_{stimulus}$; see Methods for definitions of all tasks).

What is the relationship between the ring of fixed points in the memory and response periods? To address this question, we traced locations of the fixed points during interpolation across memory and response period inputs, $(1 - \alpha)u_{memory} + \alpha u_{response}$, where α was incremented in 0.05 steps between 0 and 1. We identified fixed points for

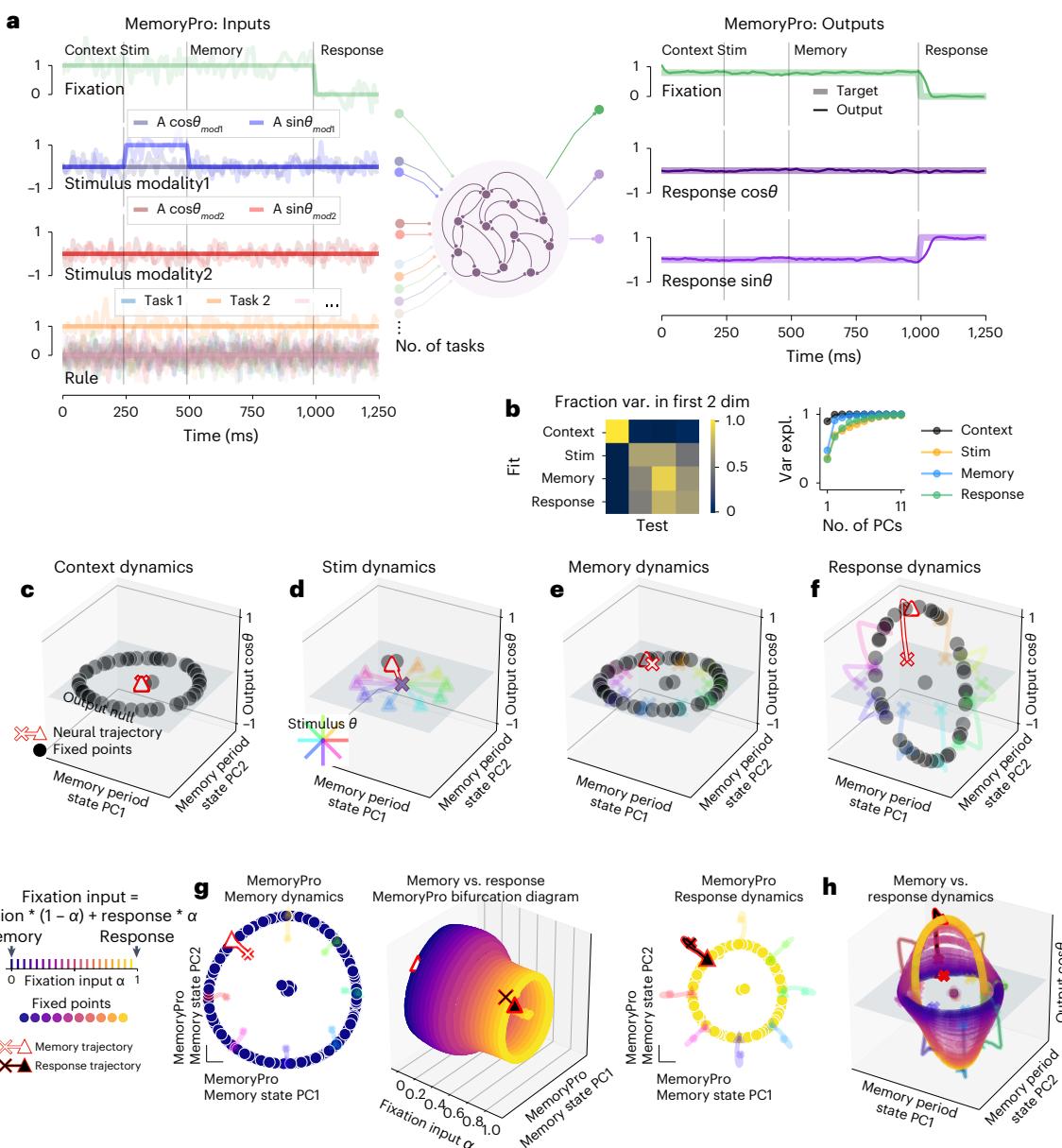


Fig. 1 | Single-task network shared fixed points across task periods. **a**, Left: noisy fixation, stimulus (modality 1 and modality 2) and rule input time series (overlaid without noise for clarity). Noise was used during training, and analyses were performed on running the network without noise. Vertical lines divide task periods: context, stimulus, memory and response. Right: targets (thick lines) overlaid with outputs of a trained network (thin lines). Stim, stimulus task period. **b**, Fraction variance explained in each task period by top two PCs of neural state trajectories for 1,024 stimulus conditions from every other task period (left). We performed PCA on task period Y and then calculated how much variance it explained on task period X. Right: top 11 PCs of neural state trajectories for 1,024 stimulus conditions from each task period. **c–f**, State space plots for single-task network performing MemoryPro during context (**c**), stimulus (**d**), memory (**e**) and response (**f**) task periods. State trajectories and fixed points projected onto the first two PCs defined by state trajectories during

the memory period on the x and y axes and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis. We visualized fixed point locations for $\theta_{stimulus} = 0$ (black dots) in all subpanels of Fig. 1 and additionally plotted state trajectories for other stimulus conditions (see 'Fixed points' in Methods for further details on fixed point identification). State trajectories (colored lines) are colored according to stimulus orientation with $\theta_{stimulus} = 0$ highlighted in red, starting from 'x' and ending with '▲'. **g**, Interpolation between inputs for memory ($\alpha = 0$) and response ($\alpha = 1$) periods. Middle: fixed points for 20 intermediate α values (x axis) projected into top two memory period PCs (as in **c–f**) (y and z axes) with memory $\alpha = 0$ (left) and response $\alpha = 1$ (right) fixed points and trajectories. **h**, Fixed points for input interpolation between memory (blue) and response (yellow) inputs. State trajectories are colored according to stimulus orientation. Same axes as **c–f**. All subpanels were generated from the same network with hyperparameters: 256 units, softplus activation, diagonal initialization.

each incremental input setting as a function of α ('Input interpolation' in Methods). By interpolating across input conditions for the memory and response periods, we traced how fixed points moved and changed stability as the dynamical system reconfigured.

For every intermediate input value throughout interpolation, an approximate ring attractor was present (Fig. 1g). The smooth transition

of this fixed point structure implies that each intermediate ring attractor was functionally the same ring attractor across input conditions. In this single-task network, the dynamical motif that performed memory and response computations was shared across task periods. The ring attractor rotated from output null space into output potent space when the fixation input changed to zero (Fig. 1h).

Two-task networks

We then trained networks to perform two tasks on interleaved batches. The MemoryPro and MemoryAnti tasks were both memory-guided response tasks that received identical stimulus inputs. The target outputs in the pro task were the same as the stimulus inputs ($\varphi_{\text{response}} = \theta_{\text{stimulus}}$), whereas, in the anti task, targets were in the opposite direction as the stimulus ($\varphi_{\text{response}} = \theta_{\text{stimulus}} + \pi$) (see Extended Data Fig. 1 and 'Tasks and performances' in Methods for full task definitions). We quantified the fraction of state trajectory variance explained in the top 10 PCs for a given task period and in the top two PCs defined by every other task period (Fig. 2a). This provided a reference for the fraction of variance captured in visualizations (Fig. 2b–i). We visualized and quantified variance of the network state in each task period for many additional subspaces (Extended Data Fig. 2).

Input interpolation across rule inputs for a network trained on the MemoryPro and MemoryAnti tasks revealed shared fixed points across tasks during the context/memory, stimulus and response periods (Fig. 2). Context period fixed points were similar to the single-task network throughout rule input interpolation, with one stable fixed point that was relevant to the context period and a ring of fixed points that was relevant to the memory period (Fig. 2b,c,f,g). Stimulus period rule input interpolation revealed two separate stable fixed points and an unstable fixed point between them for each intermediate input condition (Fig. 2d,e). The network state evolved away from the unstable fixed point, which smoothly moved in state space across interpolated input conditions, resulting in the network state evolving toward a different stable fixed point for each task. From that point onward, the state interacted with a shared ring attractor across both tasks (MemoryPro and MemoryAnti) and task periods (memory and response) according to the response direction (Fig. 2f–i). In summary, this network flexibly performed two related computations through small changes in fixed point locations. In addition to shared fixed points across different tasks and task periods, we could identify shared fixed points across different stimulus conditions for the same task period (Extended Data Fig. 3a–c).

One might expect that networks share fixed points due to the limited computational resources in small networks. We, therefore, trained networks that were nearly an order of magnitude larger and without noise in the inputs or recurrent units to determine whether supplying abundant computational resources might change this solution. To our surprise, we found that even large networks without noise still shared dynamical motifs (Extended Data Fig. 3d–l). We interpret these findings to mean that shared dynamical motifs are not a product of limited resources, and we explore possible explanations for shared motifs in 'Discussion'. Examples in Fig. 2 and Extended Data Fig. 3d–l provide a demonstration of what is and is not consistent across different networks. It is consistent that fixed points persist across different inputs and are often shared across tasks. On the other hand, the configuration of fixed points is not consistent across contexts (Fig. 2d and Extended Data Fig. 3g).

Identifying dynamical motifs in 15 task networks

To quantify shared structure across many tasks in a single network and to compare shared structure across multiple networks, we developed a modified version of the task variance metric described by Yang et al.²¹. We were motivated to study task periods because changes in the inputs reconfigure the RNN's dynamics across task periods. For example, when the stimulus input turns off in some tasks, the network goes from processing a stimulus to maintaining a memory of the stimulus. Task periods, therefore, provide the relevant granularity to identify the dynamical motifs that perform distinct computations.

We divided tasks into task periods and computed the variance across stimulus conditions for each unit, normalized across all task periods ('Task variance analysis' in Methods). The result was a matrix of each unit's normalized variance for each task period of every task (Fig. 3a), which we refer to as the variance matrix in subsequent

analyses. We sorted the rows and columns of this matrix based on similarity ('Clusters' in Methods). Clusters of units a–z were identified by performing hierarchical clustering on the columns and rows of the variance matrix and identifying a distance criterion to maximize the ratio of intercluster to intracluster distances (Fig. 3a, Extended Data Fig. 4 and 'Clusters' in Methods).

Sorting the rows and columns of the variance matrix revealed a blockwise structure, where groups of units had large variance for groups of task periods with similar computational requirements (Fig. 3a). Similar computations can be seen in the task period color labels (Stimulus1 and Stimulus2—for tasks with two sequential stimulus presentations, Memory and Response) and in the task names (Category, DecisionMaking, Memory, and so on) (Fig. 3a, left). For example, task period cluster 2 (Fig. 3a, right) corresponds to reaction-timed response task periods (see Extended Data Fig. 1 for definitions of all tasks). These tasks receive new stimulus information during a response period that must be incorporated into the computation immediately. Therefore, the network cannot prepare a response direction before the fixation cue disappears. On the other hand, in task period cluster 9, the network receives no new information during the response period and must, instead, use the memory of the stimulus to produce the correct output during the response period. These separate blocks in the variance matrix reveal two distinct clusters of units that contribute to response period dynamics: one for tasks with reaction-timed responses and another for tasks with memory-guided responses. Other unit clusters for stimulus (unit clusters k–o, task period clusters 4 and 5) and memory (unit clusters a and b, task period cluster 10) computations are apparent in the block-like structure aligned with task period type (Fig. 3a, task titles and task period color labels to the left of the variance matrix). Qualitative structure and quantitative variance of the fixed points for each task period within unit clusters demonstrate the relationship between dynamical motifs and unit clusters (Extended Data Fig. 5). Block structure in the variance matrix was robust to different network architectures and hyperparameters (examples in Extended Data Fig. 6). To quantify task period similarity across networks with different hyperparameters, we calculated the variance matrix for each trained network. We then sorted task periods according to similarity of rows in one reference network and computed the correlation matrix of the sorted rows in the variance matrix for each network ('Task variance analysis' in Methods). The correlation matrix for each trained network revealed the block-wise similarity of task periods (Fig. 3c). Higher correlation across trained networks compared to untrained networks confirmed that the block structure in the variance matrix emerged from learning the task computations rather than from network design choices or the structure of the inputs (Fig. 3b, right) (Pearson correlation coefficient between correlation matrices; 'Task variance analysis' in Methods).

The variance matrix quantified the extent to which the neural state evolved in similar subspaces across task periods, suggestive of which computations shared dynamical motifs. We highlight two different examples of shared memory dynamical motifs using rule input interpolation (Fig. 3d,e) and highlight their positions in the variance matrix (Fig. 3a, left of task period names, red and yellow squares). A pair of category memory task periods are within the same cluster in the variance matrix, suggesting that their computations are performed by a similar set of units (Fig. 3a, yellow square left: task period cluster 6, unit clusters t and u). Both category tasks used the same two point attractors for memory of the initial stimulus. Rather than store the identity of the initial continuous circular stimulus, the network stored which category it must respond to, regardless of task (Fig. 3d). In another example of a shared attractor across tasks, we found a ring attractor that was shared across several tasks (Fig. 3a; task period clusters 9 and 10, unit clusters a–d). All these tasks required memory of the initial continuous circular stimulus variable. To show that this ring attractor was shared across tasks, we interpolated across rule inputs for a pair of these tasks (IntegrationModality1 and IntegrationModality2)

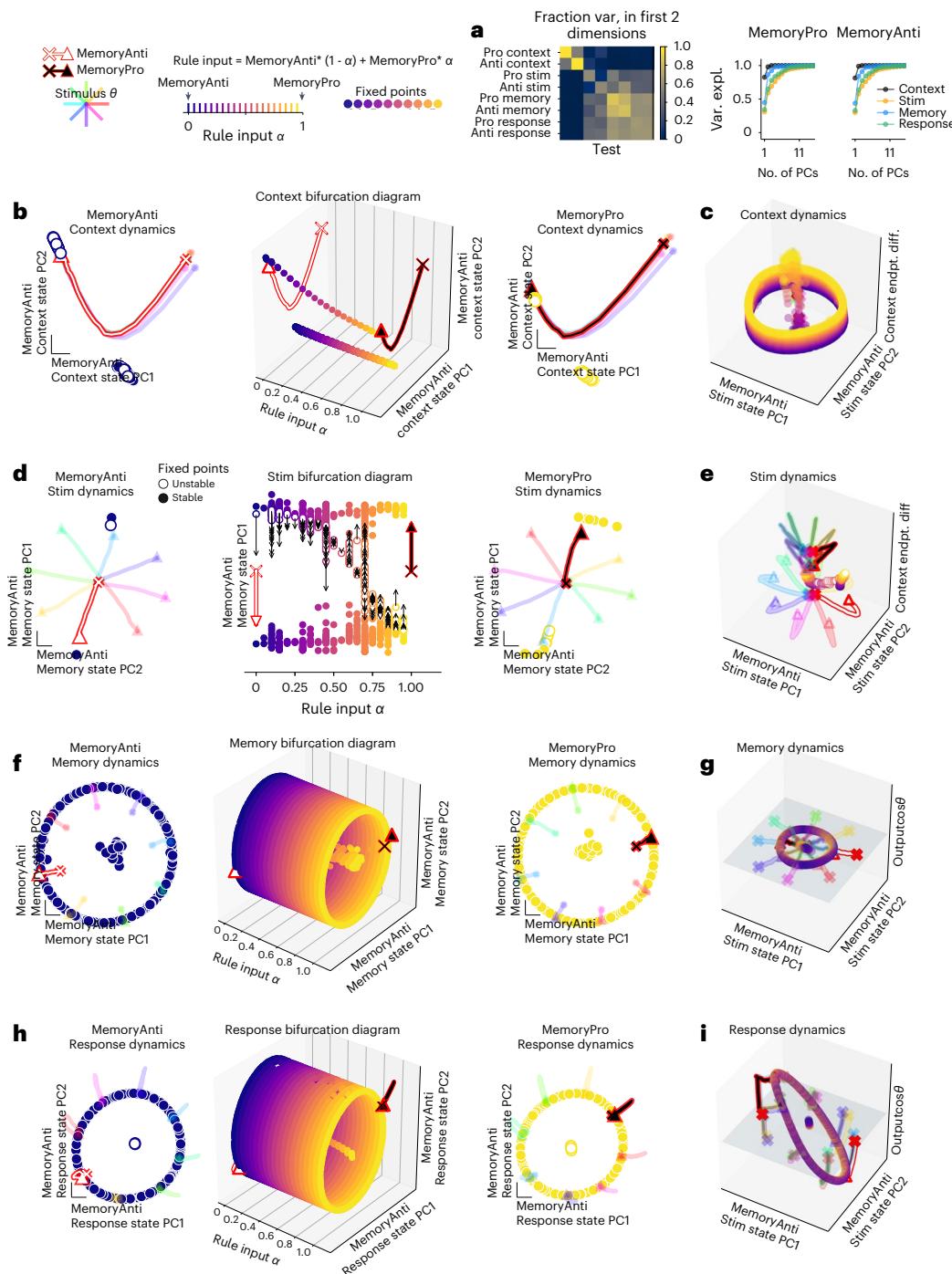


Fig. 2 | Two-task networks shared fixed points across related tasks. **a**, Fraction variance explained in each task period by top two PCs of neural state trajectories for 1,024 stimulus conditions from every other task period (left). We performed PCA on task period Y and then calculated how much variance it explained on task period X. Right: top 10 PCs of neural state trajectories for 1,024 stimulus conditions from the same task period. **b–i**, Fixed points for interpolation between inputs for MemoryAnti ($\alpha = 0$) and MemoryPro ($\alpha = 1$) tasks during context (**b,c**), stimulus (**d,e**), memory (**f,g**) and response (**h,i**) periods. **b**, Middle, fixed points for 20 intermediate α values (x axis) projected into top two PCs defined by state trajectories during the context period of the MemoryAnti task (y and z axes) with MemoryAnti $\alpha = 0$ (left) and MemoryPro $\alpha = 1$ (right) fixed points and trajectories. **c**, Rule input interpolation between tasks, MemoryAnti (blue fixed points, white state trajectory) and MemoryPro (yellow fixed points, black state trajectory) projected into the top two MemoryAnti stimulus period state PCs and the vector connecting context period state endpoints on the z axis. **d**, Same as **b** for stimulus period, with unstable (open) and stable (closed) fixed

points projected into top PC defined by the state trajectories during the memory period of the MemoryAnti task (y axis). Local linear dynamics around unstable fixed points are shown in black arrows: the state was initialized at the unstable fixed point and run forward for 10 steps. **e**, Same as **c** for stimulus period, projected into the top two MemoryAnti stimulus period state PCs and the vector connecting context period state endpoints on the z axis. **f**, Same as **b** for memory period, projected into top two PCs defined by the state trajectories during the memory period of the MemoryAnti task (y and z axes). **g**, Same as **c** for memory period, projected into the top two MemoryAnti stimulus period state PCs (x and y axes) and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis. **h**, Same as **b** for response period, projected into top two PCs defined by the state trajectories during the response period of the MemoryAnti task (y and z axes). **i**, Same as **g** for response period. All subpanels were generated from the same network with hyperparameters: 256 units, softplus activation, diagonal initialization. endpt. diff., endpoint difference; Var. expl., variance explained.

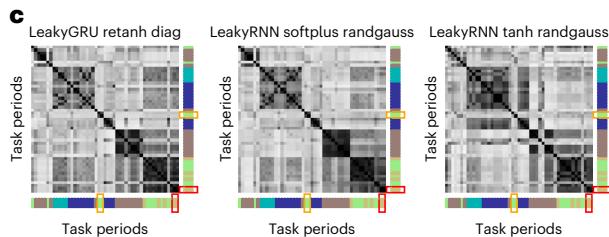
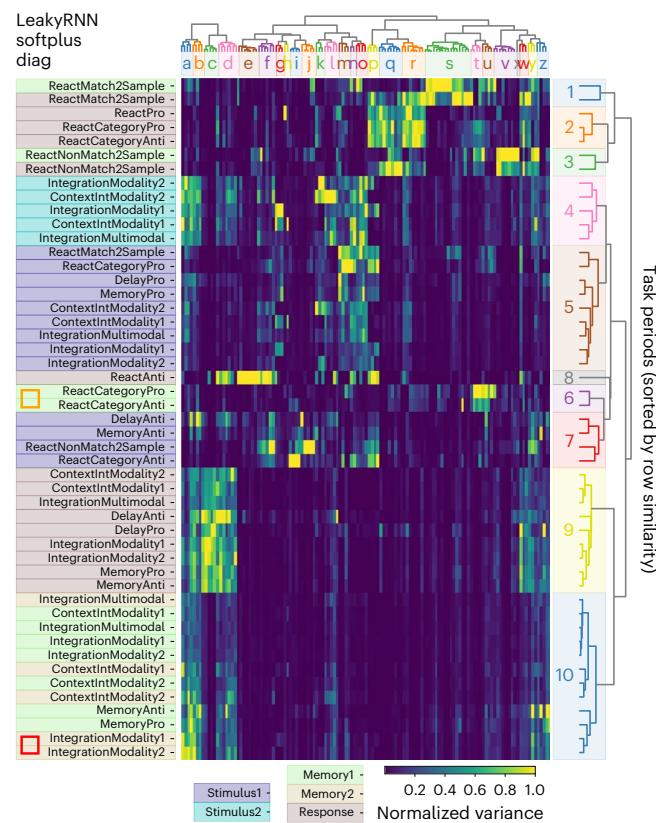
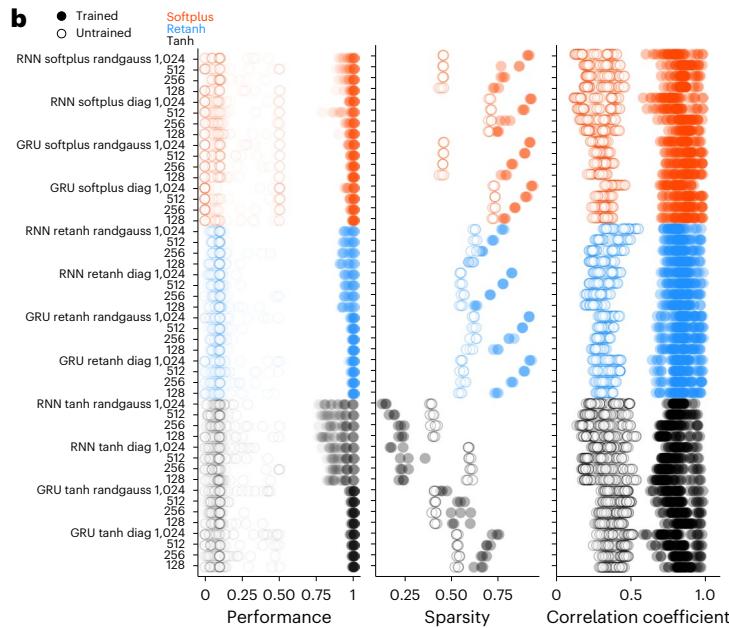
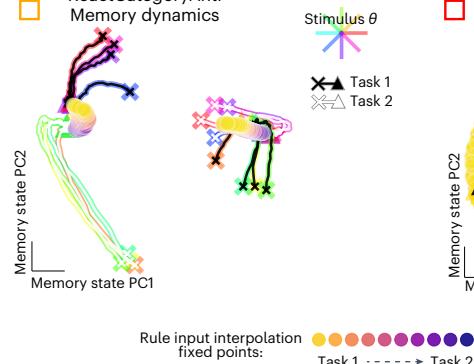
a Example network units (sorted by column similarity)

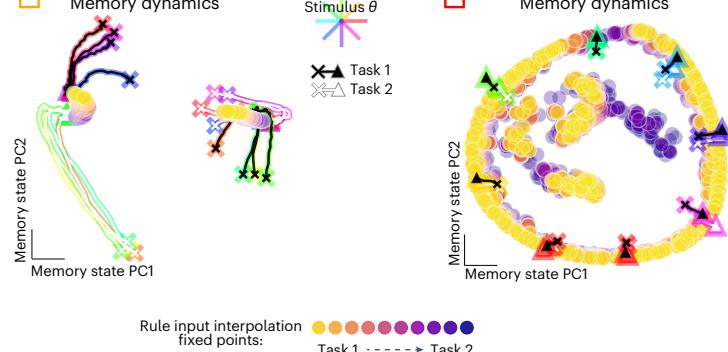
Fig. 3 | Modular organization in 15 task networks was not dependent on activation function, network initialization or size. **a**, Variance matrix: variance of unit activations across stimulus conditions, normalized across task periods (columns normalized by the maximum entry in each column). Rows and columns were sorted according to similarity ('Clusters' in Methods). Colors along the left side label each task period type (dark blue: initial stimulus; light blue: second stimulus; light green: initial memory; yellow: second memory; brown: response). Notice that blocks with high variance belong to same task period types (similar colors). **a,c–e**, Orange and red rectangles highlight category memory (ReactCategoryPro, ReactCategoryAnti) and continuous memory (IntegrationModality1, IntegrationModality2) tasks, respectively. **b**, Left: performance across all tasks for three networks of each of 48 hyperparameter settings. Middle: sparsity of the task variance measured as the fraction of entries <15% maximum unit variance. Right: task period correlation matrix (examples shown in **c**) for trained and untrained networks are sorted according to rows in **a** and correlated to trained networks for all other hyperparameter settings. **c**, Correlation matrix of rows in variance matrix (as in **a**) for three different



d ReactCategoryPro vs. ReactCategoryAnti
Memory dynamics



e IntegrationModality1 vs. IntegrationModality2
Memory dynamics



example networks; rows and columns sorted according to rows in **a**. Colors along the bottom and right axes label each task period type (dark blue: initial stimulus; light blue: second stimulus; light green: initial memory; yellow: second memory; brown: response). Notice that blocks with high correlation belong to the same task period types (similar colors). **d**, Shared point attractors for two category memory tasks as seen by input interpolation across tasks during memory period. State trajectories for eight stimulus conditions (colored by stimulus direction) starting from 'x' projected in PC space, defined by concatenating memory period state trajectories from both tasks for ReactCategoryPro (black) and ReactCategoryAnti (white) tasks. Rule input interpolation across tasks during memory period with fixed points for intermediate rule input conditions in filled circles. **e**, Same as **d** for two continuous circular variable memory tasks, highlighting shared ring attractors. State trajectories starting from 'x' projected in PC space, defined by concatenating memory period state trajectories from both tasks for IntegrationModality1 (black) and IntegrationModality2 (white) tasks. Panels **a**, **d** and **e** were generated from the same network with hyperparameters: 128 units, softplus activation, diagonal initialization, randgauss, random Gaussian.

and found a similar shared ring structure as in the two task networks (Fig. 2f,g). We highlight shared category and continuous memory dynamical motifs in networks with different activation functions in Extended Data Fig. 7.

In addition to clusters of task periods with similar variance, there were also some task periods that did not cluster with other task periods. For example, task period cluster 8 is dedicated to the ReactAnti task;

cluster 1 is dedicated to the ReactMatch2Sample task; and cluster 3 is ReactNonMatch2Sample (Fig. 3a). In these cases, the computation performed in the unique task period is so distinct from other computations, the dynamical motif is unlikely to be reused across tasks. These results are robust; the set of tasks that employed unique dynamical motifs was similar across hyperparameter settings (Fig. 3c and Extended Data Fig. 6).

Motif alignment to unit axes

One notable difference across network hyperparameters was sparsity in the variance matrix. We define sparsity to be the fraction of entries in the variance matrix below a threshold of 15% maximum unit variance. Networks with non-negative activation functions had sparse task variance matrices, whereas networks with the tanh activation function, which has a range of $(-1, 1)$, did not (Fig. 3b, middle). We understand sparsity to be a function of optimal network performance requiring potentially interfering dynamical motifs to be organized into orthogonal subspaces. In a network with activity regularization and where all units can take only positive values, this orthogonalization favors unit axes (Extended Data Fig. 8). We found clusters to be present in tanh networks, simply not aligned to unit axes and, therefore, non-identifiable using methods described in Yang et al.²¹. By examining the correlation matrix and the correlation coefficient across networks, we see that similar clusters are present in the tanh networks (Fig. 3b, right, and Fig. 3c, right).

Shared stimulus period dynamical motifs in 15 task networks

The variance matrix provides a useful overview of which task periods are implemented by similar clusters of units but falls short of addressing exactly how these subpopulations implement shared dynamical motifs. Shared motifs are implemented by organizing the state in the appropriate region of state space to evolve on the relevant shared dynamical landscape. To walk through this explanation in detail, we focus on stimulus period dynamics and highlight two examples, one in which dynamical motifs are shared and another where motifs are not shared.

Tasks with similar stimulus computations (integration, categorization, pro versus anti and reaction-timed versus delayed response) organized stimulus period initial conditions to be nearby in state space and evolve in a similar way after stimulus onset (see schematic in Fig. 4a). We visualized this organization in PC space defined by the final state of the context period across all tasks (Fig. 4b). To summarize the relationship between initial conditions and the ensuing stimulus dynamics for different tasks, we compared pairs of trials presented with the same stimulus across different tasks. We plotted the Euclidean distance between initial conditions against the angle between the state vector on the first timestep for pairs of trials (Fig. 4c). We observed that pairs of tasks with similar computations had initial conditions that were closer together and had smaller angles between state trajectories on the first timestep of the stimulus period compared to pairs of tasks with distinct computations. Similar initial conditions for stimulus onset resulted in shared context-dependent stimulus amplification in some networks (Extended Data Fig. 9). In these cases, the state update was scaled in magnitude according to whether the stimulus input was either Modality 1 or Modality 2, dependent on the position of that state at stimulus onset³⁰.

The relationship between context period states and stimulus period trajectory angles across tasks supports the idea that nearby initial conditions allowed tasks with similar stimulus computations to reuse the same dynamical landscape and, therefore, evolve in similar ways. We examine these features in two examples of comparisons between tasks that (1) share the same stimulus period dynamical motif and, then, tasks that (2) do not share the same dynamical motif (Fig. 4d–i).

In the case of two categorization tasks, ReactCategoryPro and ReactCategoryAnti, we found a shared stimulus motif (Fig. 4d–f). In the ReactCategoryPro task, the network was trained to respond if both sequential stimuli were less than or both greater than π ; whereas, in the ReactCategoryAnti task, the network was trained to respond if stimuli were on opposite sides of π . In either task, there was a decision boundary at $\theta_{\text{stimulus}} = \pi$. The initial conditions for these tasks were nearby, and trajectories during the stimulus period were aligned (Fig. 4c, ‘Category Motif’). We quantified stimulus response overlap by computing the fraction of variance explained for the state trajectory during one task by

the other task’s PCs (purple) compared to its own PCs (black) (Fig. 4d), revealing that both tasks were performed in an aligned subspace. Aligned stimulus responses for both category tasks were visualized in PC space defined by the stimulus period state trajectories of the ReactCategoryPro task (Fig. 4e, left and right). These analyses revealed that activity evolved in a qualitatively similar way for trials with the same stimulus conditions during both tasks, suggesting that the state trajectories could have occurred on a similar dynamical landscape.

To better understand the relationship between the dynamical landscapes across task contexts, we interpolated across rule inputs during the stimulus period for both category tasks with the same stimulus input. We found that similar stimulus responses were governed by shared stable and unstable fixed points, demonstrated by the smooth bridge of fixed points between both tasks (Fig. 4e, middle). We projected the unstable dimension of each unstable fixed point into PC space and found that this dimension was aligned with the direction of the state trajectories and roughly orthogonal to the decision boundary (Fig. 4e). We defined the most relevant fixed point to be the closest unstable fixed point to the state at the end of this task period. This simplification of one relevant fixed point was often necessary to tease apart how relevant dynamics are reconfigured across tasks while several to hundreds of other fixed points related to computations during other task periods also moved through state space. A continuous bridge mapped movement of the relevant fixed point, suggesting that rule inputs shifted a relevant shared fixed point that was reused across both tasks (Fig. 4e,f). Moreover, the stability of the local linear dynamics around this shared fixed point was consistent across all intermediate input conditions, as shown by the maximal real part of the eigenvalue of the linearized RNN state update around each interpolated fixed point location ('Analysis of fixed points for interpolated inputs' in Methods) (Fig. 4f). We interpret this result to mean that both category tasks reuse the unstable fixed point to move the state away from the category boundary.

The DelayAnti and ReactPro tasks were an example pair that did not share any dynamical motifs (Fig. 4g–i). The DelayAnti task began with the context period, followed by a stimulus presentation that signaled the opposite response direction ($\phi_{\text{response}} = \theta_{\text{stimulus}} + \pi$), followed by a ‘go’ cue that signaled when to initiate a delayed response (see Extended Data Fig. 1 for all task definitions). The ReactPro task began with the context period, followed by a stimulus presentation that signaled the same response direction and required an immediate response. During the context period, the network state evolved toward dissimilar locations for trials of either task, and trajectories during the stimulus period were not aligned (Fig. 4c, ‘Different Motifs’). We defined the subspace for the ReactPro task by performing PCA on the state trajectories during the stimulus period. We projected the DelayAnti task in the same subspace and found that little variance was captured by the other task PCs (Fig. 4g). We also visualized both tasks in a subspace defined by the first two PCs of the DelayAnti task and again found little overlap, suggesting that both tasks evolved in mostly non-overlapping subspaces (Fig. 4h). We interpolated across these two rule inputs during the stimulus period, revealing that there was a bifurcation where the relevant stimulus-dependent fixed point did not form a continuous bridge across interpolated rule inputs (Fig. 4h, middle). We quantified the distance between consecutive fixed points that were closest to the endpoint of the state trajectory for each interpolated input and identified a large discrete jump in the location of the relevant fixed point (Fig. 4i, right). We visualized the maximal real part of the eigenvalues of the linearized RNN state update around each consecutive fixed point, revealing qualitatively dissimilar local dynamics around fixed points for interpolated input conditions, indicated by crossing the stability threshold at one for this discrete dynamical system (Fig. 4i, left).

Taken together, these features suggest that shared dynamical motifs are implemented by evolving the state to the appropriate region

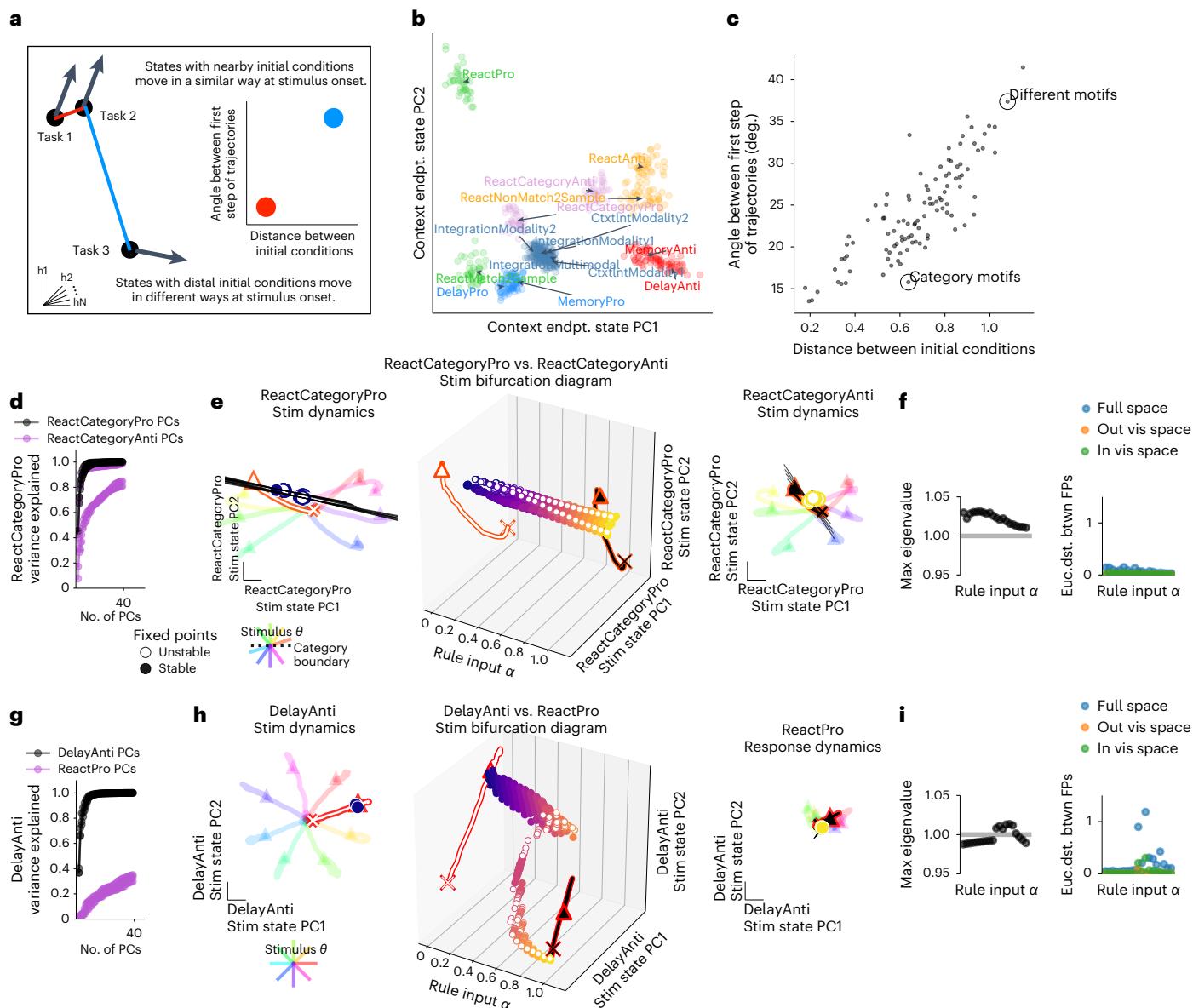


Fig. 4 | Tasks with similar stimulus computations were in nearby parts of state space and shared dynamical motifs. a, Schematic of analyses in **b** and **c**. **b**, The state for each trial (colored dot for each of 20 trials on each task) at the end of the context period (just before stimulus period) projected onto the top two PCs defined by the state at the end of the context period for all tasks. Trial-trial variation is due to input noise, independent noise and random initial conditions. Trials colored by similar stimulus computations as given by task definitions: pro delayed (light blue), anti delayed (red), pro integration (steel blue), categorization (pink), pro reaction (green) and anti reaction (orange). See Supplementary Table 1 for motif definitions. **c**, Euclidean distance between pairs of trials from different tasks at the end of the context period plotted against cosine angle between same pair after stimulus onset for a particular stimulus input, θ_{stimulus} , for one timestep and then averaged across stimulus angle inputs. Pairs of tasks in **d–i** are circled and labeled: ‘Category Motif’: ReactCategoryPro and ReactCategoryAnti (**d–f**) and ‘Different Motifs’: DelayAnti and ReactPro (**g–i**). **d**, Fraction of variance explained for ReactCategoryPro task by the ReactCategoryAnti task PCs (purple) compared to its own PCs (black) for five trained networks with different random seeds. **e**, Rule input interpolation across category tasks for one stimulus angle. Middle, unstable (open) and stable (closed) fixed points for 20 intermediate α values (x axis) projected onto top two PCs defined by state trajectories during the stimulus period of the ReactCategoryPro task (y and z axes) with ReactCategoryPro $\alpha = 0$ (left) and ReactCategoryAnti $\alpha = 1$ (right) fixed points and trajectories for eight different stimulus angles (rainbow colors). Expanding dimensions around unstable

fixed points are visualized as black lines. **f**, Left, maximum real eigenvalue for the linearization of the state update around each fixed point for the single unstable fixed point closest to the state at the end of the stimulus period for 20 consecutive α values between 0 and 1. Right, Euclidean distance between consecutive fixed points in the full state space (blue), outside of the visualized subspace in **e** (orange) and within the visualized space (green). We analyzed only one unstable fixed point that is most proximal to the end of the state trajectory for each input condition to highlight the task relevant fixed point (‘Analysis of fixed points for interpolated inputs’ in Methods). **g–i**, ‘Different Motifs’: DelayAnti and ReactPro. Same as **d–f** but for DelayAnti and ReactPro tasks. Note that a response is required at stimulus onset for the ReactPro task; therefore, the stimulus period is the same as the response period. **i**, Left, maximum real eigenvalue for the linearization of the state update around each fixed point for the single fixed point closest to the state at the end of the stimulus period for 20 consecutive α values between 0 and 1. Right, Euclidean distance between consecutive fixed points in the full state space (blue), outside of the visualized subspace in **h** (orange) and within the visualized space (green). We analyzed one fixed point that is most proximal to the end of the state trajectory for each input condition. Panels **b**, **c**, **e**, **f**, **h** and **i** were generated from the same network as in Fig. 3a with hyperparameters: LeakyRNN, 128 units, softplus activation, diagonal initialization. Other networks in **d** and **g** use the same hyperparameters. deg, degrees; endpt., endpoint; Euc.dist. btwn, Euclidean distance between; FP, fixed point; vis, visualized.

of state space such that it interacts with a shared fixed point across similar task computations. Category tasks shared both unstable and stable fixed points. On the other hand, stimulus period dynamics for the DelayAnti and ReactPro tasks evolved in separate subspaces and were governed by different stable fixed points. These analyses revealed that shared structure was not merely an artifact of all tasks within the same network; rather, only tasks with similar computations implemented shared dynamical motifs. Note, however, that we cannot rule out the possibility that an alternate path of interpolation between inputs might reveal shared fixed points. See 'Analysis of fixed points for interpolated inputs' in Methods and 'Discussion' for limitations of the interpolated inputs analysis.

Dynamical motifs result in modular lesion effects

Yang et al.²¹ found that network lesions affected sets of tasks that shared computational features. For example, if the output of a particular cluster of units was set to zero, then all tasks involving a particular computation decreased their performance, whereas other tasks were unaffected. Their work left open the major question of why lesion effects were modular. We identified the cause of these modular lesion effects to be related to the underlying modular dynamical motifs that perform computation.

We examined the impact of lesioning clusters of units described in the variance matrix in Fig. 3a. Many unit clusters had high variance for a set of task periods with similar computations. For example, unit clusters a and c had high variance for memory and response task period clusters 10 and 9, respectively (Fig. 3a). Other unit clusters had high variance for Modality 2 stimulus periods (unit cluster k) or anti stimulus periods (unit cluster f), etc. We lesioned a cluster of units by setting the output of all units within the cluster to be zero throughout a given trial.

In six example lesions, we demonstrate that some unit cluster lesions impacted only a subset of tasks that shared computational features where units had high variance. Here, we show that lesions either did or did not impact task-relevant computations depending on whether the relevant underlying dynamical motif was impacted (Fig. 5). A lesion to one cluster (Fig. 3a, unit cluster c) only impacted performance on tasks that included a delayed response (Fig. 5a–c). This motif

is responsible for rotating the ring attractor from output null space into output potent space²⁹. We visualized fixed points and state trajectories for task periods where the delayed response motif was either irrelevant (stimulus period) or relevant (response period) during MemoryPro and MemoryAnti tasks in lesioned (orange) and non-lesioned (blue) networks (Fig. 5b,c). Fixed points and, in turn, state trajectories were markedly impacted when the delayed response dynamical motif was lesioned during the response period of the MemoryPro and MemoryAnti tasks but not during the stimulus period of either task. In a second example, we lesioned a cluster of units with high variance during the performance of anti response tasks (Fig. 3a, unit cluster f) and found that there was minimal change in the fixed points and state trajectories of pro response tasks (Fig. 5d–f). We present additional examples for Modality1, Modality2, Category Memory and Continuous Memory motifs (Fig. 5g–r). Note that our clustering approach produces an arbitrary threshold to separate branches of the dendrogram into different groups (Fig. 3a and 'Clusters' in Methods). As a result, we sometimes combined different clusters that were on the same branch in our lesion studies. For category memory and continuous memory tasks, we lesioned two adjacent unit clusters (t and u for category memory and a and b for continuous memory). This highlights limitations in our approach to link clusters of units directly to dynamical motifs, which will be an important direction for future studies.

When tasks did not use the dynamical motif impacted by a unit cluster lesion, they were not impacted by the lesion. For example, MemoryPro stimulus period activity was not impacted by a lesion to the unit cluster which implemented the response period dynamical motif of the same task (Fig. 5b). The state evolved toward a stable fixed point at approximately the same location in both the lesioned (orange) and full (blue) network after this cluster was lesioned. However, this lesion resulted in minimal rotation into output potent space during the response period (Fig. 5c, middle and right). Conversely, there was little change in the fixed point structure after a lesion to the anti stimulus motif (cluster f) for all task periods during the MemoryPro task (Fig. 5e). Taken together, results from our lesion studies suggest that modular lesion effects are a result of modular fixed point structures that implement dynamical motifs.

Fig. 5 | Unit cluster lesions had modular effects on task period clusters that shared the same dynamical motif. **a–c**, Delayed response lesion.

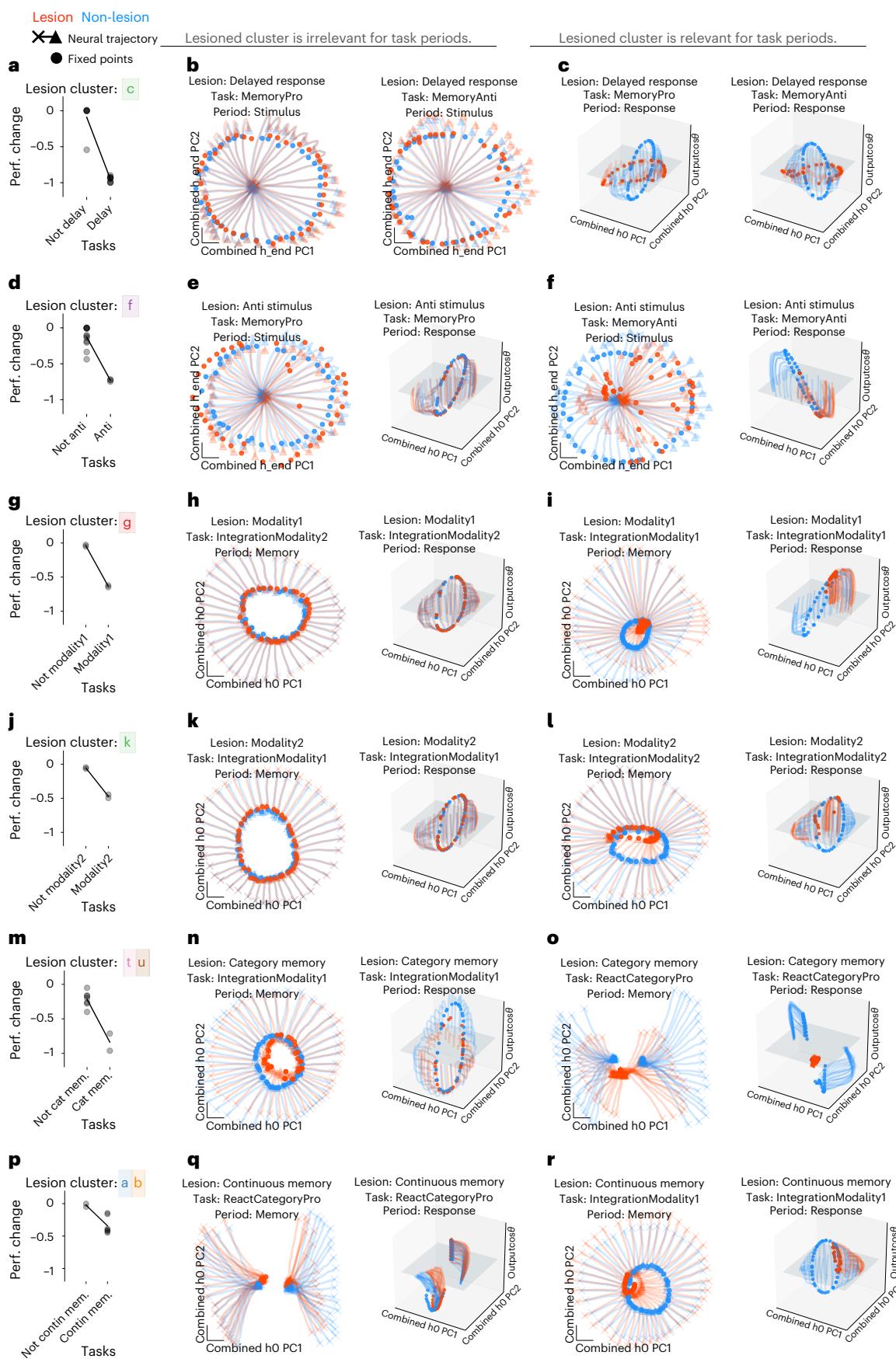
a, Fraction performance change (each point is one task) after delayed response (cluster c in Fig. 3a) lesion by setting unit output to zero for units within cluster with high variance during the response period of delayed response tasks (see Supplementary Table 1 for motif definitions). **b**, Fixed points (filled circles) and state trajectories (starting from 'x' and ending with '▲') during the performance of MemoryPro and MemoryAnti stimulus periods in lesioned (orange) and full (blue) network projected into the first two PCs (x and y axes) defined by the full and lesioned network state at the last timestep of the stimulus period. Stimulus period dynamics should not be relevant for the delayed response unit cluster lesion. As a result, fixed points and state trajectories diverge minimally from the full network. **c**, Same as **b** during the response periods in lesioned (orange) and full (blue) network projected into the first two PCs (x and y axes) defined by the full and lesioned network state at the first timestep of the response period and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis. Response period dynamics should be relevant for the delayed response unit cluster lesion. As a result, lesion prevents fixed points from rotating into output potent space and has a marked impact on network performance. **d–f**, Anti stimulus lesion. **d**, Same as **a** for anti stimulus (cluster f in Fig. 3a) during irrelevant MemoryPro stimulus and response task periods (**e**) and relevant MemoryAnti stimulus and response task periods (**f**). Fixed points and activity are projected onto the first two PCs defined by the network state at the last timestep of the stimulus period for both the lesioned and full network (x and y axes) (left) and the first timestep of the response period and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis (right). **g–i**, Modality1 Lesion. **g**, Same as **a** for Modality1 (cluster g in Fig. 3a). **h**, Irrelevant IntegrationModality2 stimulus and response task periods. **i**, Relevant IntegrationModality1 stimulus and response task periods. Fixed

points and activity are projected onto the first two PCs (x and y axes) defined by the network state at the first timestep of the memory task period for both the lesioned and full network and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis. **j–l**, Modality2 Lesion. **j**, Same as **a** for Modality2 (cluster k in Fig. 3a). **k**, Irrelevant IntegrationModality1 stimulus and response task periods. **l**, Relevant IntegrationModality2 stimulus and response task periods. Fixed points and activity are projected onto the first two PCs (x and y axes) defined by the network state at the first timestep of the memory task period for both the lesioned and full network (left) and the first timestep of the response task period for both the lesioned and full network and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis (right). **m–o**, Category memory lesion. **m**, Same as **a** for category memory (clusters t and u in Fig. 3a). **n**, Irrelevant IntegrationModality1 memory and response task periods. **o**, Relevant ReactCategoryPro memory and response task periods. Fixed points and activity are projected onto the first two PCs (x and y axes) defined by the network state at the first timestep of the memory task period for both the lesioned and full network (left) and the first timestep of the response task period for both the lesioned and full network and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis (right). **p–r**, Continuous memory lesion. **p**, Same as **a** for continuous memory (clusters a and b in Fig. 3a). **q**, Irrelevant ReactCategoryPro memory and response task periods. **r**, Relevant IntegrationModality1 memory and response task periods. Fixed points and activity are projected onto the first two PCs (x and y axes) defined by the network state at the first timestep of the memory task period for both the lesioned and full network and the output weight vector (from W_{out}) associated with $\cos\theta_{stimulus}$ on the z axis. All panels were generated from the same network as in Fig. 3a with hyperparameters: LeakyRNN, 128 units, softplus activation, diagonal initialization. Perf., performance. h_0 is the state at the beginning of the task period and h_{end} is the state at the end of the task period.

Fast learning of novel tasks by reusing dynamical motifs

Networks were able to rapidly learn new tasks sequentially by reconfiguring previously learned dynamical motifs. We first identified a task

where each task period shared a dynamical motif with at least one of the other 14 tasks: MemoryAnti. This task requires the anti stimulus motif (Fig. 2d) and the delayed response memory motif (Figs. 1e–h)



and 2f–i). We next trained a network to perform every task except the MemoryAnti task. After learning all other tasks, we trained only the N_{rec} -dimensional rule input vector for the MemoryAnti task (Fig. 6a). By training only one new set of rule input weights, we did not interfere with any previously learned dynamical motifs within the recurrent weight matrix, W_{rec} , enabling learning of new tasks without catastrophic forgetting. The network was able to learn the MemoryAnti task when previously trained on all other tasks (Fig. 6b, black). Similar results were observed for other tasks (Extended Data Fig. 10a–d).

To determine if the anti stimulus and delayed response dynamical motifs were sufficient, we pre-trained a network on tasks containing these motifs (DelayAnti and MemoryPro). Pre-training on this minimal set with relevant motifs achieved similar speed and proficiency as pre-training on all tasks (Fig. 6b, blue), suggesting that these motifs were sufficient. Conversely, pre-training on tasks lacking the anti motif (orange) or no pre-training (green) resulted in significantly slower learning or failure to learn MemoryAnti entirely (Fig. 6b).

Despite slow and variable learning, networks pre-trained on tasks lacking the anti motif were still able to learn the new dynamical motif by modifying the rule input vector (Fig. 6b, orange). This resulted in stimulus period state trajectories that were not highly overlapping with previously learned tasks (Fig. 6c). Linear interpolation between DelayPro and MemoryAnti rule inputs revealed a qualitative change in the dynamics across input conditions. This suggests that the MemoryAnti stimulus period relevant fixed point was distinct from the previously learned DelayPro stimulus period relevant fixed point (Fig. 6e,f). Despite learning a new stimulus period anti motif, the network was still able to reuse the previously learned memory motif (Fig. 6g–j). This result highlights the modularity of dynamical motifs.

Networks that were pre-trained with the relevant dynamical motifs reused the anti stimulus and memory dynamical motifs for fast learning of the novel MemoryAnti task. MemoryAnti state trajectories were in highly overlapping subspaces with the DelayAnti state trajectories during the stimulus period (Fig. 6k,l) and with the MemoryPro state trajectories during the memory period (Fig. 6o,p). Rule input interpolation between both anti tasks during the stimulus period (Fig. 6m) and memory tasks during the response period (Fig. 6q) provided strong evidence that the fixed point structures were shared.

In another example, we first trained on all tasks except ContextIntegrationModality2 and then used our transfer learning approach to learn the context inputs for the held-out task (Extended Data Fig. 10h–k). We found that the same stimulus fixed point was reused across ContextIntegrationModality1 and ContextIntegrationModality2 tasks (Extended Data Fig. 10h,i). This fixed point shifted in state

space corresponding to the relevant modality. The same plane attractor was reused to store the amplitude information of the stimulus throughout the memory period (Extended Data Fig. 10j,k). The requirement for these tasks to compare the amplitude of two consecutive stimuli resulted in a plane attractor rather than the ring attractor in other tasks, which has no amplitude information. See Extended Data Fig. 9 for detailed fixed point structures that implement integration tasks.

We wanted to better understand how transfer learning effectiveness relates to the uniqueness of a task's dynamical motifs. We trained networks on all but one task each, followed by training only the rule input for the held-out task. Performance was compared to single-task training (Extended Data Fig. 10). In Fig. 3, we calculated the correlation between rows in the variance matrix to identify task periods that might share dynamical motifs. In the following analysis, we used correlation between rows in the variance matrix to identify which tasks might benefit from pre-training on other tasks (see 'Transfer learning' in Methods for details).

Nearly all tasks shared similarities with others (high correlation) and benefited from pre-training early in training (faster performance gain) (Extended Data Fig. 10a–d and 10e–g, left). Tasks with unique dynamical motifs had at least one task period where the maximum correlation to other task periods was low (ReactAnti, ReactNonMatch2Sample and ReactMatch2Sample). These tasks could not be learned as well using this pre-training method compared to full network training, except for ReactAnti, which was easily learned by either training approach (Extended Data Fig. 10d and 10e–g). In summary, we found that rapid learning was not as successful in the context of novel tasks with unique dynamical motifs. These results provide support that rapid learning of novel tasks requires reconfiguration of relevant previously learned dynamical motifs.

Discussion

In this work, we examined how recurrently connected artificial networks flexibly repurposed their learned dynamics to perform multiple tasks. Our collection of commonly studied cognitive tasks could be broken down into an underlying set of subtasks (contextual integration, memory, categorization, anti-response, etc.; see Supplementary Table 1 for all motif definitions). We showed that networks learned this underlying subtask structure, which resulted in specialized computational building blocks that we call dynamical motifs, dedicated to each subtask. Using input interpolation and fixed point analyses, we showed how dynamical motifs were organized in relation to one another and often shared across tasks or task periods. Inputs reconfigured the dynamical

Fig. 6 | Dynamical motifs were reused for fast learning of novel tasks with familiar computational elements. a, Schematic of two-stage learning.

Networks were pre-trained on a set of tasks while all weights were plastic. The same network was then trained on a novel task by only learning weights for a single new one-hot rule input. **b**, Left: networks were pre-trained on two tasks that include pro stimulus and continuous memory motifs (orange), anti stimulus and continuous memory motifs (blue), all motifs (black) and no motifs (green). See Supplementary Table 1 for motif definitions. Right: performance during MemoryAnti task rule input weight training after pre-training on various sets of motifs for five different networks each. **c–j**, Network pre-trained on DelayPro and MemoryPro tasks (pro stimulus and continuous memory motifs) was then trained to perform MemoryAnti through weight changes to the MemoryAnti rule input ($N_{\text{rec}} \times 1$ vector). **c**, Fraction of MemoryAnti stimulus period variance explained by MemoryAnti stimulus period PCs (black) and DelayPro stimulus period PCs (purple) quantifies the extent to which both pro and anti tasks are in a similar subspace during the stimulus period. **d**, Stimulus period activity for DelayPro (white) and MemoryAnti (black) tasks for eight different stimulus angles (rainbow colors) projected into PC space (x and y axes) defined by state trajectories during MemoryAnti task and context period state endpoint difference between both tasks (z axis). **e**, Rule input interpolation across tasks for one stimulus angle. Middle: unstable (open) and stable (closed) fixed points

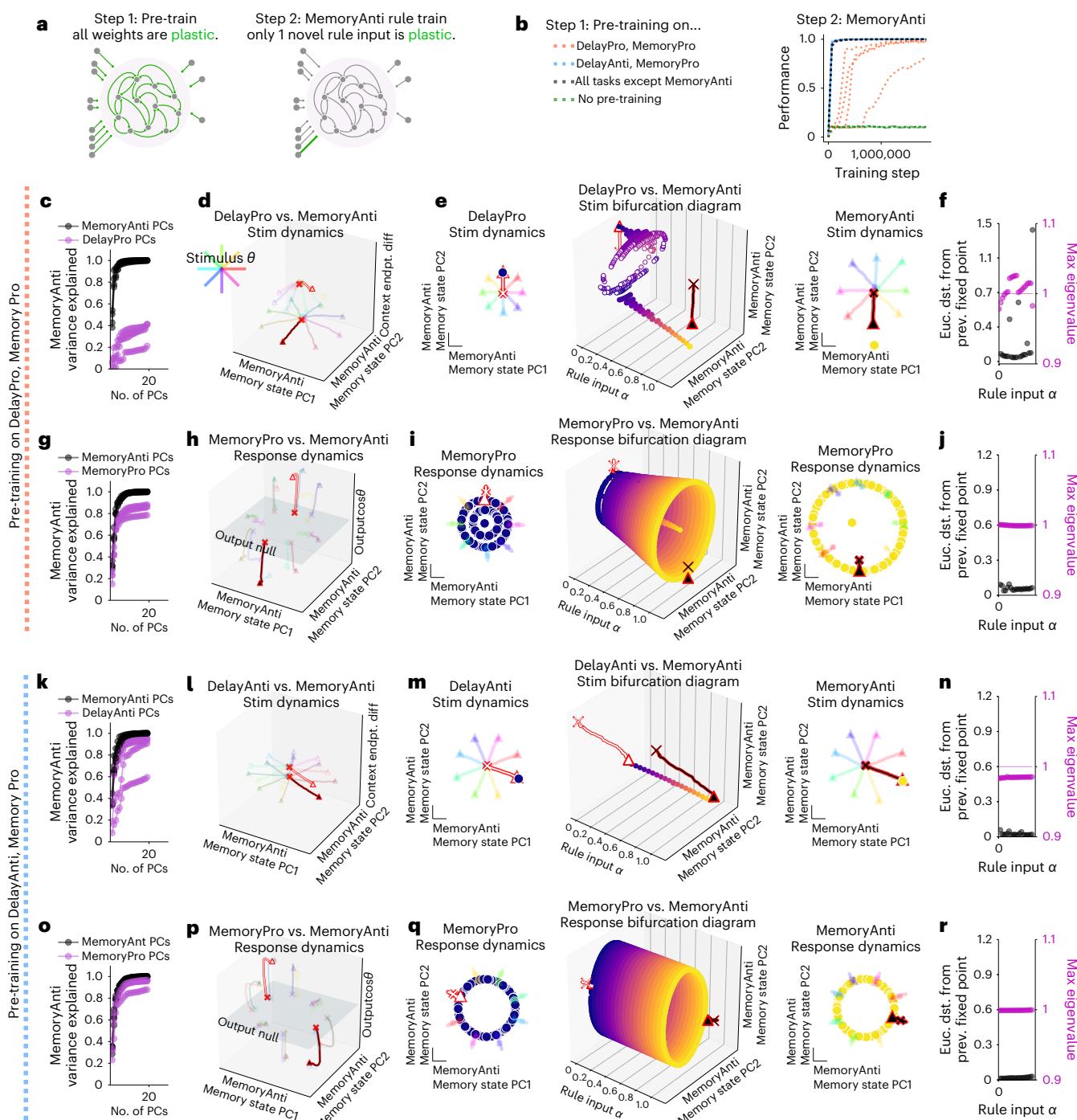
for 20 intermediate α values (x axis) projected onto top two PCs defined by state trajectories during the memory period of the MemoryAnti task (y and z axes) with DelayPro $\alpha = 0$ (left) and MemoryAnti $\alpha = 1$ (right) fixed points and activity for eight different stimulus angles (rainbow colors). **f**, Euclidean distance between fixed points (black) and maximum real eigenvalue for the linearization of the state update around each fixed point (purple) for the single fixed point closest to the state at the end of the stimulus period for 20 consecutive α values between 0 and 1. Analyzing only one fixed point that is most proximal to the end of the state trajectory for each input condition ('Analysis of fixed points for interpolated inputs' in Methods). **g–j**, Same as e–f for response period of MemoryPro and MemoryAnti tasks. **k–r**, Network pre-trained on DelayAnti and MemoryPro tasks (anti stimulus and continuous memory motifs) was then trained to perform MemoryAnti task through weight changes to the MemoryAnti rule input (length N_{rec} vector). **k–n**, Same as c–f for stimulus period of DelayAnti and MemoryAnti tasks with pre-training on DelayAnti and MemoryPro tasks. **o–r**, Same as c–f for response period of MemoryPro and MemoryAnti tasks with pre-training on DelayAnti and MemoryPro tasks. **d–f** and **h–j** were generated from one network, and **l–n** and **p–r** were generated from a different network, both with hyperparameters: LeakyRNN, 256 units, softplus activation, diagonal initialization. Euc. dst., Euclidian distance.

system in each task period, often resulting in smooth changes to the dynamical landscape underlying the performed computation. Motifs necessary to perform each subtask included different types of attractor structures, input amplifications, decision boundaries and rotations. The modular subtask structure in our set of tasks is analogous to the structure of language, mathematics and other natural behaviors in everyday life^{31,32}.

Our framework of examining subtask computation through the lens of dynamical motifs made it possible to explain lesions and learning results described previously^{19,21}. As in the study by Yang et al.²¹, we found that lesioning specific unit clusters resulted in specific deficits in sets of tasks that were related computationally. Units within a cluster had high variance during a set of task periods that shared a dynamical motif. When we lesioned a given unit cluster, the fixed points that

made up the associated dynamical motif were greatly impacted in terms of their locations and stability. A unit cluster associated with one dynamical motif could be lesioned with little impact to other computations that the network performed. This finding was surprising given the all-to-all connectivity possible in our networks as well as the fact that no regularizations or constraints to induce modularity were employed in the training of the RNNs. Recent work on subpopulation structure for the implementation of multiple tasks provides insight for these findings^{22,33}.

We demonstrated that networks equipped with relevant dynamical motifs could repurpose those motifs modularly for fast learning of novel tasks. The initial phase of learning novel dynamical motifs was a slow process. However, given a rich repertoire of previously learned motifs, a network could quickly repurpose motifs



to perform novel tasks by learning a single input weight vector. Our findings suggest that a useful lifelong learning strategy could include two stages of learning. Early in learning, it may be beneficial for highly plastic recurrent connections throughout the brain to learn novel subtasks (dynamical motifs). Late in learning, reduced plasticity in recurrent connections and new plastic layers that function as contextualizing inputs could repurpose previously learned subtasks. This hypothesis is interesting to consider in the context of critical periods³⁴ and re-aiming³⁵. We hypothesize that this two-stage process of slow and fast learning could provide some intuition for off-manifold and on-manifold brain–machine interface learning results in non-human primates^{36–38}. Reusable dynamical motifs may inform state-of-the-art models that require pre-training³⁹.

Our results are based on artificial systems, lacking the complexities of real brains. We used simplified networks without diverse cell types or prescribed architectures and only applied noisy static inputs. Although our learning rules are not biological, we hypothesize that optimized artificial neural networks and the principles that we uncover from them are informative about biological neural circuits based on principles of optimality and robustness⁴⁰. Although some constraints changed the way dynamical motifs were shared, our main finding that they are, in fact, shared was robust across all types of networks and hyperparameter choices that we tested, including large networks without noise (Extended Data Fig. 3e–l). These findings suggest that shared motifs are not a result of limited computational resources. We hypothesize that the modular organization of dynamical motifs was a result of the modular subtask structure of our tasks, but learning dynamics through gradient descent could play a role⁴¹. It will be of great interest to further explore the prevalence of dynamical motifs in other artificial⁴² and biological systems⁴³.

We examined a set of cognitive tasks commonly studied in humans and other animals^{21,22,25}. Networks learned a variety of fixed point structures capable of implementing these computations. Notably missing from this set of attractors were limit cycles, which have been shown to be important for computations that require specific timed responses. Additionally, tasks involving complex pattern generation or continuously changing inputs, such as real-time adaptation or tracking, may involve dynamics that are not well captured by our fixed point analysis. These limitations highlight the versatility of RNNs and the need for further exploration with a broader range of tasks. Fixed point structures often moved in different contexts rather than appeared or disappeared (Figs. 1g, 2f, h and 3d, e). This is consistent with previous work, where graded input results in a continuum of stable fixed points^{15,44}. Interpolations across contextual inputs, with intermediate values that the network was never exposed to during training, revealed an absence of bifurcations in the task-relevant dynamical motif. This finding suggested that motifs were reused across tasks and presents a conceptual advance for thinking about the relationship between different computations. Future research should explore the connection among high-dimensional parameter bifurcations, composition and computation^{45–48}.

Fixed points often persisted even when the inputs were not relevant to the current task (Fig. 2c). These irrelevant fixed points did not interfere with network computation because the network state was organized to be more proximal to task-relevant fixed points. This aligns with the concept of ‘sloppiness’ in flexible systems^{49–51}.

In this work, we focused on dynamical motifs that share fixed points across tasks because we could measure this feature through empirical bifurcation diagrams. However, tasks that do not share exactly the same fixed points could still evolve on a similar, nearby dynamical landscape. Notice that, in Fig. 4c, there are not two discrete clusters; rather, we found a continuous distribution of angle between trajectories and proximity of initial conditions. Recently developed analysis tools that quantify similarity of dynamical systems could

lay the groundwork for studying shared dynamical motifs without fixed points^{52,53}.

Our linear interpolation analysis presents limitations in identifying shared motifs. We assert that motifs are shared if interpolation reveals a smooth fixed point transition with no detected bifurcations. However, if we identify a bifurcation along one path, alternative interpolation paths might reveal continuous transitions. Examining all paths is impractical, so we cannot definitively claim that two motifs are not shared through this approach.

Compositional systems require modularity to recombine components in different contexts, fostering flexibility and generalization. We identified dynamical motifs as the underlying modular substrate that could support compositional computation. Previous work demonstrated that generalization to new tasks could be implemented without training by using a linear combination of task inputs²¹. We demonstrate that learning a single input is sufficient to link previously learned motifs within a fixed recurrent network. It will be of great interest to identify how and why a linear combination of inputs is able to recruit combinations of previously learned motifs. This will require a better understanding of how motifs are sequentially recruited and how inputs might reconfigure the space between and around fixed points that implement computation. We found that similar computations are proximal to each other in state space (Fig. 4b). This results in a hierarchical organization of tasks based on similarity (as shown in the variance matrix) and begins to address the question of compositionality. Although we think that this smoothness in state space^{14,17,54} is related to compositionality, it remains for future work to tackle head on why linear combinations of previously unseen task inputs should result in largely functioning systems.

Our findings offer several testable predictions. The method of studying unit variance across tasks could be readily performed on neural data (Fig. 3a). This analysis could be informative about perturbations to biological network activity that would most markedly impact performance on a computation of interest (Fig. 5). Results might apply across biological scales, from multiple motifs within a brain region to distinct computations in different cortical areas. Our approach for training networks sequentially by reusing previously learned dynamical motifs could be used to determine ideal curricula for training animals on complex tasks. For example, given a particular task of interest, one could train an artificial network to perform the task and inspect all relevant dynamical motifs. For example, the ‘anti’ and ‘memory’ motifs were the sufficient set of relevant motifs for the MemoryAnti task in Fig. 6b. Based on the task-relevant motifs, one could systematically design a set of tasks to learn a sufficient set of motifs rather than designing a curriculum through guesswork. Additionally, this work highlights the relevance of reporting training protocols as they may shape the dynamical motifs that implement computation^{55–57}. Beyond experimental predictions, our work provides intuition for why the brain exhibits functional specialization.

Through the lens of dynamical systems, we identified the underlying computational substrate for clustered representations described previously by Yang et al.²¹ and highlighted a new level of organization between the unit and the network: groups of units that implement dynamical motifs. More broadly, our findings highlight the relevance of dynamical systems as a framework to better understand the response properties of neurons in the brain. As researchers record more whole-brain activity, the framework of dynamical motifs will guide questions about specialization and generalization across brain regions.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-024-01668-6>.

References

1. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
2. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
3. Silver, D. et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**, 1140–1144 (2018).
4. Cole, M. W. et al. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat. Neurosci.* **16**, 1348–1355 (2013).
5. Yang, G. R., Cole, M. W. & Rajan, K. How to study the neural mechanisms of multiple tasks. *Curr. Opin. Behav. Sci.* **29**, 134–143 (2019).
6. Frankland, S. M. & Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annu. Rev. Psychol.* **71**, 273–303 (2020).
7. Willett, F. R. et al. Hand knob area of premotor cortex represents the whole body in a compositional way. *Cell* **181**, 396–409 (2020).
8. Reverberi, C., Görgen, K. & Haynes, J.-D. Compositionality of rule representations in human prefrontal cortex. *Cereb. Cortex* **22**, 1237–1246 (2012).
9. Laje, R. & Buonomano, D. V. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat. Neurosci.* **16**, 925–933 (2013).
10. Barak, O., Sussillo, D., Romo, R., Tsodyks, M. & Abbott, L. F. From fixed points to chaos: three models of delayed discrimination. *Prog. Neurobiol.* **103**, 214–222 (2013).
11. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
12. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
13. Carnevale, F., de Lafuente, V., Romo, R., Barak, O. & Parga, N. Dynamic control of response criterion in premotor cortex during perceptual detection under temporal uncertainty. *Neuron* **86**, 1067–1077 (2015).
14. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
15. Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron* **98**, 1005–1019 (2018).
16. Mastrogiosseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623 (2018).
17. Russo, A. A. et al. Motor cortex embeds muscle-like commands in an untangled population response. *Neuron* **97**, 953–966 (2018).
18. Pinto, L. et al. Task-dependent changes in the large-scale dynamics and necessity of cortical regions. *Neuron* **104**, 810–824 (2019).
19. Duncker, L., Driscoll, L., Shenoy, K. V., Sahani, M. & Sussillo, D. Organizing recurrent network dynamics by task-computation to enable continual learning. In Proc. 34th Conference on Neural Information Processing Systems <https://proceedings.neurips.cc/paper/2020/file/a576eafbce762079f7d1f77fca1c5cc2-Paper.pdf> (NeurIPS, 2020).
20. Masse, N. Y., Grant, G. D. & Freedman, D. J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proc. Natl Acad. Sci. USA* **115**, E10467–E10475 (2018).
21. 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).
22. Dubreuil, A., Valente, A., Beiran, M., Mastrogiosseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nat. Neurosci.* **25**, 783–794 (2022).
23. Riveland, R. & Pouget, A. Natural language instructions induce compositional generalization in networks of neurons. *Nat. Neurosci.* **27**, 988–999 (2024).
24. Yang, G. R., Ganichev, I., Wang, X.-J., Shlens, J. & Sussillo, D. A dataset and architecture for visual reasoning with a working memory. In 2018 Conference on Cognitive Computational Neuroscience <https://doi.org/10.32470/ccn.2018.1075-0> (Cognitive Computational Neuroscience, 2018).
25. Márton, C. D., Gagnon, L., Lajoie, G. & Rajan, K. Efficient and robust multi-task learning in the brain with modular latent primitives. Preprint at <https://arxiv.org/abs/2105.14108> (2021).
26. Sussillo, D. & Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**, 626–649 (2013).
27. Strogatz, S. H. & Strogatz, R. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* 2nd edn (Westview Press, 1994).
28. Golub, M. & Sussillo, D. FixedPointFinder: a Tensorflow toolbox for identifying and characterizing fixed points in recurrent neural networks. *J. Open Source Softw.* **3**, 1003 (2018).
29. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).
30. Pagan, M. et al. A new theoretical framework jointly explains behavioral and neural variability across subjects performing flexible decision-making. Preprint at bioRxiv <https://doi.org/10.1101/2022.11.28.518207> (2022).
31. 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).
32. Baroni, M. Linguistic generalization and compositionality in modern artificial neural networks. *Phil. Trans. R. Soc. Lond. B* **375**, 20190307 (2020).
33. Beiran, M., Dubreuil, A., Valente, A., Mastrogiosseppe, F. & Ostojic, S. Shaping dynamics with multiple populations in low-rank recurrent networks. *Neural Comput.* **33**, 1572–1615 (2021).
34. Hensch, T. K. Critical period regulation. *Annu. Rev. Neurosci.* **27**, 549–579 (2004).
35. Morehead, J. R., Qasim, S. E., Crossley, M. J. & Ivry, R. Savings upon re-aiming in visuomotor adaptation. *J. Neurosci.* **35**, 14386–14396 (2015).
36. Sadtler, P. T. et al. Neural constraints on learning. *Nature* **512**, 423–426 (2014).
37. Golub, M. D. et al. Learning by neural reassociation. *Nat. Neurosci.* **21**, 607–616 (2018).
38. Oby, E. R. et al. New neural activity patterns emerge with long-term learning. *Proc. Natl Acad. Sci. USA* **116**, 15210–15215 (2019).
39. Han, X. et al. Pre-trained models: past, present and future. *AI Open* **2**, 225–250 (2021).
40. Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S. & Sussillo, D. Universality and individuality in neural dynamics across large populations of recurrent networks. In Proc. 33rd Conference on Neural Information Processing Systems https://proceedings.neurips.cc/paper_files/paper/2019/file/07a4e20a7bbee7a736682b26b16ebe8-Paper.pdf (NeurIPS, 2019).
41. Turner, E. & Barak, O. The simplicity bias in multi-task RNNs: shared attractors, reuse of dynamics, and geometric representation. In Proc. 37th Conference on Neural Information Processing Systems https://proceedings.neurips.cc/paper_files/paper/2023/file/50d6dbc809b0dc96f7f1090810537acc-Paper-Conference.pdf (NeurIPS, 2023).
42. Goudar, V., Peysakhovich, B., Freedman, D. J., Buffalo, E. A. & Wang, X.-J. Schema formation in a neural population subspace underlies learning-to-learn in flexible sensorimotor problem-solving. *Nat. Neurosci.* **26**, 879–890 (2023).

43. Chen, G., Kang, B., Lindsey, J., Druckmann, S. & Li, N. Modularity and robustness of frontal cortical networks. *Cell* **184**, 3717–3730 (2021).
44. Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* **21**, 102–110 (2018).
45. Jaeger, H. in *Dynamics, Synergetics, Autonomous Agents* Vol. 8 (eds Tshacher, W. & Dauwalder, J.-P.) 29–48 (World Scientific, 1999).
46. Jordan, I. D., Sokół, P. A. & Park, I. M. Gated recurrent units viewed through the lens of continuous time dynamical systems. *Front. Comput. Neurosci.* **15**, 678158 (2021).
47. Beer, R. D. Codimension-2 parameter space structure of continuous-time recurrent neural networks. *Biol. Cybern.* **116**, 501–515 (2022).
48. Wang, X.-J. Theory of the multiregional neocortex: large-scale neural dynamics and distributed cognition. *Annu. Rev. Neurosci.* **45**, 533–560 (2022).
49. Marder, E. & Goaillard, J.-M. Variability, compensation and homeostasis in neuron and network function. *Nat. Rev. Neurosci.* **7**, 563–574 (2006).
50. Gutenkunst, R. N. et al. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, 1871–1878 (2007).
51. O’Leary, T., Sutton, A. C. & Marder, E. Computational models in the age of large datasets. *Curr. Opin. Neurobiol.* **32**, 87–94 (2015).
52. Ostrow, M., Eisen, A. J., Kozachkov, L. & Fiete, I. R. Beyond geometry: comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. In *Proc. 37th Conference on Neural Information Processing Systems* <https://openreview.net/forum?id=7blSUMwe7R> (OpenReview.net, 2023).
53. Galgali, A. R., Sahani, M. & Mante, V. Residual dynamics resolves recurrent contributions to neural computation. *Nat. Neurosci.* **26**, 326–338 (2023).
54. Saxena, S., Russo, A. A., Cunningham, J. & Churchland, M. M. Motor cortex activity across movement speeds is predicted by network-level strategies for generating muscle activity. *eLife* **11**, e67620 (2022).
55. Arlt, C. et al. Cognitive experience alters cortical involvement in goal-directed navigation. *eLife* **11**, e76051 (2022).
56. Latimer, K. W. & Freedman, D. J. Low-dimensional encoding of decisions in parietal cortex reflects long-term training history. *Nat. Commun.* **14**, 1010 (2023).
57. Molano-Mazón, M. et al. Recurrent networks endowed with structural priors explain suboptimal animal behavior. *Curr. Biol.* **33**, 622–638 (2023).

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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2024