

<https://doi.org/10.1038/s44488-026-00013-z>

Partial recurrence can enable robust and efficient computation

Marcus Ghosh^{1,2,3} ✉ & Dan F. M. Goodman³

Neural circuits are sparse and bidirectional. Meaning that signals flow from early sensory areas to later regions and back. Yet, between connected areas there exist some but not all pathways. How does this structure, somewhere between feedforward and fully recurrent, shape circuit function? To address this question, we designed a recurrent neural network model in which a set of weight matrices (i.e. pathways) can be combined to generate every network structure between feedforward and fully recurrent. We term these architectures partially recurrent neural networks (pRNNs). We trained over 25,000 pRNNs on a novel set of reinforcement learning tasks, designed to mimic multisensory navigation, and compared their performance across multiple functional metrics. Our findings reveal three key insights. First, in dense-cue environments, most pRNN architectures match or exceed the task performance, learning speed or robustness of fully recurrent networks, despite using as few as one quarter the number of parameters; in sparse-cue environments, many match but a substantial fraction underperform. These results demonstrate that partial recurrence can enable energy efficient, yet performant solutions. Second, each pathway's functional impact is both task and circuit dependent. For instance, feedback connections enhance robustness to noise in some, but not all contexts. Third, different pRNN architectures learn solutions with distinct input sensitivities and memory dynamics, and these computational traits help to explain their functional capabilities. Overall, our results demonstrate that partial recurrence can enable robust and efficient computation - a finding that may help to explain *why* neural circuits are sparse and bidirectional, and shows *how* these principles can inform the design of artificial systems.

Across species, anatomical studies reveal a striking feature of neural circuits: they are sparse, yet bidirectional. At the level of single neurons, the *C. elegans* and *Drosophila* connectomes contain only $\approx 3\%$ ¹⁻³ and $< 1\%$ ⁴⁻⁶ of possible synapses, yet an abundance of feedback connections. Similarly, at the level of region-region pathways in the mouse^{7,8} and macaque neocortex⁹ only $\approx 50\%$ of areas connect, yet many of these connections are reciprocal.

This anatomical structure raises a fundamental question: *what advantages or disadvantages do sparse, bidirectional circuits confer*, compared to either: feedforward networks—in which information simply flows from early sensory areas to later areas, or fully recurrent networks—in which every component is bidirectionally connected? Determining this is challenging in vivo, but in silico, one can design networks with distinct architectures and compare different measures of function. Though, how should we explore the vast space of possible network structures?

Computational studies have primarily focussed on either feedforward¹⁰ or recurrent architectures^{11,12}. Though three lines of work have imbued

recurrent networks with properties closer to those found in real circuits. One approach is to impose a low-rank structure on the recurrent (hidden-hidden) weight matrix, effectively reducing its dimensionality by introducing structured correlations between units. This low-rank structure can be imposed prior to training¹³, or even analytically derived to implement specific task-relevant dynamics¹⁴. However, as networks with very distinct structures can share the same rank, and altering connection weights changes the rank, but not the topology, interpreting structure-function relations in low-rank networks remains challenging (Supplementary Section 1). Another approach, in task-trained models, is to add regularisation terms to the loss function, to encourage networks to learn recurrent connections with specific properties, such as modularity or small-worldness¹⁵. Though, as in the low-rank case¹³, these structural changes do not lead to changes in network function, defined as task performance¹⁵ (with the exception of¹⁶). Alternatively, one can train networks with sparse or modular recurrent weight matrices. One approach is to directly impose this structure^{17,18}.

¹Department of Life Sciences, Imperial, London, UK. ²I-X Centre for AI in Science, Imperial, London, UK. ³Department of Electrical and Electronic Engineering, Imperial, London, UK. ✉e-mail: m.ghosh@imperial.ac.uk

Another is to use learning algorithms which add and remove connections throughout training to achieve a given sparsity^{19,20}. Sparse networks are more energy efficient than dense networks, but this often comes with at least a small cost to performance in task-trained models^{17–20}.

Taken together, functional results from these three approaches suggest that either many structures are degenerate, in that they yield the same input-output mappings^{21,22}, or that unconstrained networks, with more parameters, would always be more performant. However, it is possible that more naturalistic tasks or alternative measures of function, for example, sample efficiency (i.e. learning speed) or robustness to noise, would reveal differences between these structures, and perhaps even scenarios in which sparser networks outperform dense networks.

Here, we explore the link between network structure and function in four steps. First, we introduce a set of multisensory maze tasks. In these tasks, we simulate networks as agents with noisy sensors, which provide clues about the shortest path through each maze. We focus on these tasks, as they involve translating the inputs from multiple sensory channels, $\pm$ internal states, to actions. As such, these tasks more closely resemble the kinds of challenges animals face than many machine learning tasks. Moreover, there are well-established theoretical models of multisensory integration^{23,24}, which we use to benchmark our tasks. Second, we define a recurrent neural network model in which signal propagation is bidirectional, yet sparse. We term these models partially recurrent neural networks (pRNNs). Third, we train all pRNN architectures on our maze tasks and compare their performance, sample efficiency and robustness to various perturbations. Finally, to link each architecture's structure to its function, we measure three computational traits from each network: their weights, input sensitivity and memory dynamics and quantify how predictive these traits are of network function. Our results demonstrate that partial recurrence can enable robust and efficient computation, and explain why networks with distinct structures function differently.

Results

General setup

We explore a range of scenarios in which agents must navigate multisensory mazes. All agents operate in discrete time, taking fixed-width steps on a grid. For implementation details, please see section 'Agents'.

In these tasks, we provide sensory cues in the form of positive spatial gradients leading towards a goal. While these gradients may seem abstract, they represent some stimuli well—such as chemical, thermal or luminance gradients. And, they provide a good approximation to the information provided by more complex cues. For example, tracks through the snow or mud which would appear crisper as one neared their source.

To perform these tasks, each agent implements a three-step sensation-action loop (Fig. 1A). First, they sense their local environment using sensors from two channels, such as vision and hearing. Often, we add independent Gaussian noise to each channel at each time step. Next, they implement a policy mapping their sensory inputs to potential actions. Finally, they choose and implement an action, either pausing or moving in a cardinal direction.

Optimal multisensory algorithms are brittle

To benchmark our tasks, we began by testing how well seven classical models of multisensory integration (Fig. 1B), interpreted as policies (section 'Multisensory Algorithms'), perform a simple maze task.

In this task (section 'Track maze'), agents start in the middle of a 1D track. Their aim is to reach a goal location at either the left or right end of the track within a fixed number of time steps. To determine which end to navigate to, agents must use the signals from their sensors, which are grouped into two channels (C_0 and C_1). Multisensory signals, from both channels (C_0 and C_1), lead towards the goal. While unisensory signals, from one channel (C_0 or C_1), lead away from the goal. Across trials, we vary both the goal location and which channel provides the unisensory signal. As the multisensory signal is twice as strong as the unisensory signal, the optimal

strategy is to sum the information across channels and head in the direction with the greatest total.

As a measure of performance, we define a fitness function from two metrics (section 'Functional Metrics'):

- **Speed**—the time taken to reach the goal. Normalised from 0 (not reached within a fixed number of time steps) to 1 (reached as quickly as possible).
- **Accuracy**—proximity to the goal after a fixed number of time steps. Normalised from 0 (ending the trial at the furthest point from the goal) to 1 (goal reached). We weight speed and accuracy equally, such that fitness ranges from 0 to 1 (always taking the shortest path from start to goal). Notably, for consistency with later tasks, agents sense in 2D and can try to move up or down—in which case they will collide with the wall and remain in place.

In principle, we could learn parameters for these algorithms. However, as each sensor is equally reliable, we simply treat them as non-learning policies, which we do not train or fit in any way. Intuitively, agents which ignore both channels ($-/-$), and so move randomly, perform poorly and are unaffected by increasing sensor noise (Fig. 1D, left). Agents which use one channel ($+/-$ and $-/+$) or consider both but base their decision on only one (2l-binary, 2l-max, Max-dv) perform reasonably well and decay with noise (Fig. 1D, left). Agents which fuse information across channels either linearly or nonlinearly perform optimally (Fig. 1D, left), demonstrating the potential for tasks like these to reveal behavioural differences between models. To explore this further, we tested the same seven algorithms on two variants of this task.

First, we modified the cues by matching the strength of the unisensory signal leading away from the goal to the multisensory signal leading towards it (Fig. 1C, middle) (section 'Track maze'). Consequently, at the first time step, a linear sum across channels is uninformative, as the total amount of signal on the agent's left and right is equal (in the absence of noise). Instead, the optimal strategy is to use a nonlinear function to weight the multisensory signal more highly than the unisensory signal. As expected, and akin to ref. 24, nonlinear fusion now significantly outperforms the other algorithms (Fig. 1D, middle). Strikingly, two algorithms (2l-max and Max-dv) even perform below chance ($-/-$), at low noise levels, as they base their decisions on only the strongest cue in either channel—which is misleading in this case.

Second, we introduced a gap in the sensory cues leading towards the goal (Fig. 1C, right). Now, all seven algorithms perform near chance ($-/-$), as just prior to the gap, they sense nothing ahead but something behind, so, fearing the void, turn back (Fig. 1D, right).

Together, these results highlight that these classic algorithms are brittle. In the sense that, while they are optimal in specific settings, relatively small changes can reduce them to chance performance or even mislead them entirely. This brittleness motivates our search for alternative, more robust multisensory models.

Partially recurrent neural networks

To explore alternative models, we began with a neural network with: a layer of input sensors (from two channels), a hidden layer and an output layer (with a node per action: move left, right, up, down or pause). Units in the input, hidden and output layers (respectively) use linear, rectified linear (ReLU) and softmax activation functions.

In this network, there are 9 possible connection pathways or weight matrices (Fig. 2A):

- **Feedforward** connections—from the input to the hidden layer (W_{ih}) and the hidden to output layer (W_{ho}).
- **Lateral** connections—connecting each unit to itself and the other units in its layer: input-input (W_{ii}), hidden-hidden (W_{hh}) and output-output (W_{oo}).
- **Skip** connections—connecting the inputs directly to the outputs (W_{io}) or vice versa (W_{oi}).

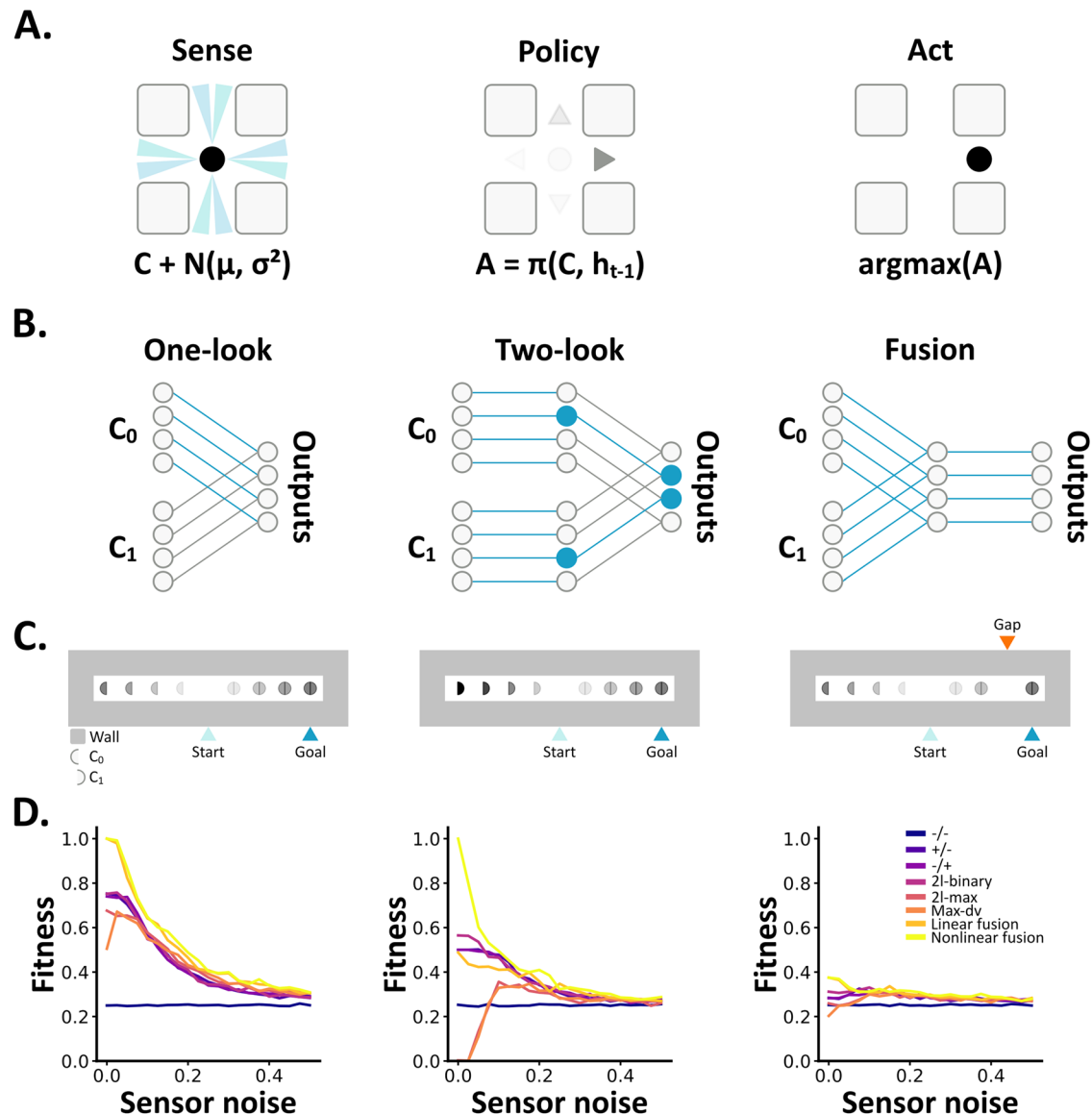


Fig. 1 | Multisensory algorithm behaviour. **A** In our tasks, agents must translate their noisy sensations ($C + N(\mu, \sigma^2)$), and in some cases hidden states (h_{t-1}), into actions (A) using algorithmic or learned policies (π). **B** Multisensory algorithms can solve these tasks by translating their sensory inputs (C) to actions (A) via different types of policies (section 'Multisensory Algorithms'). One-look algorithms (left) use one channel (blue connections) but not the other (grey connections). Two-look algorithms (middle) make independent decisions per channel (blue circles), then base their output on these. Fusion algorithms form their outputs by combining their inputs across sensory channels. **C** Left: an example track maze with a path shown in

white and walls in grey. Sensory cues are shown as circles, with cues from channel 0/1 shown on the left/right of each circle, such that semicircles denote unisensory cues. Light to dark fills indicate weak to strong signals. Middle: an example nonlinear track maze—the strength of the unisensory cues (leading away from the goal) matches the strength of the multisensory cues (leading towards). Right: an example track maze with a gap in the sensory cues. **D** Each algorithm's mean fitness as a function of increasing sensor noise on the: track maze (left), nonlinear track maze (middle) and track maze with a gap (right).

- *Backwards* connections—connecting the hidden units to the inputs (W_{hi}) or the output units to the hidden units (W_{oh}).

Assuming the feedforward connections (W_{ih} and W_{ho}) are always present, and each of the other 7 matrices can be present in any combination, yields $2^7 = 128$ distinct architectures. In this framework, excluding all 7 additional weight matrices yields a pure layer-layer feedforward network (Fig. 2B). While including all 7 generates a fully recurrent network—as every unit connects bidirectionally to every other unit (Fig. 2B). Between these two extremes are 126 architectures (e.g. Fig. 2B).

We term these architectures partially recurrent as information flow can be *bidirectional*—from early to later areas and back, yet *sparse*—in the sense that it cannot flow via all possible paths, as in a fully recurrent network. While several of these architectures have been studied individually, for

instance²⁵ and²⁶ define networks with output to hidden connections (W_{oh}) or lateral hidden connections (W_{hh}), respectively. Our contribution is to place these, and other architectures, into a taxonomy, which provides a common reference frame for comparing different architectures and highlighting under-explored structures.

All 128 circuits are shown with their sparse layer-layer pathways or sparse unit-unit matrices in (Supplementary Section 2). As we move through this space of network structures, how does function change? For instance, do networks with lateral hidden connections (W_{hh})—as in the standard recurrent neural network model, behave differently from those with lateral input (W_{ii}) or output (W_{oo}) connections? And, how would adding forward skip connections (W_{io}) to those architectures change their behaviour?

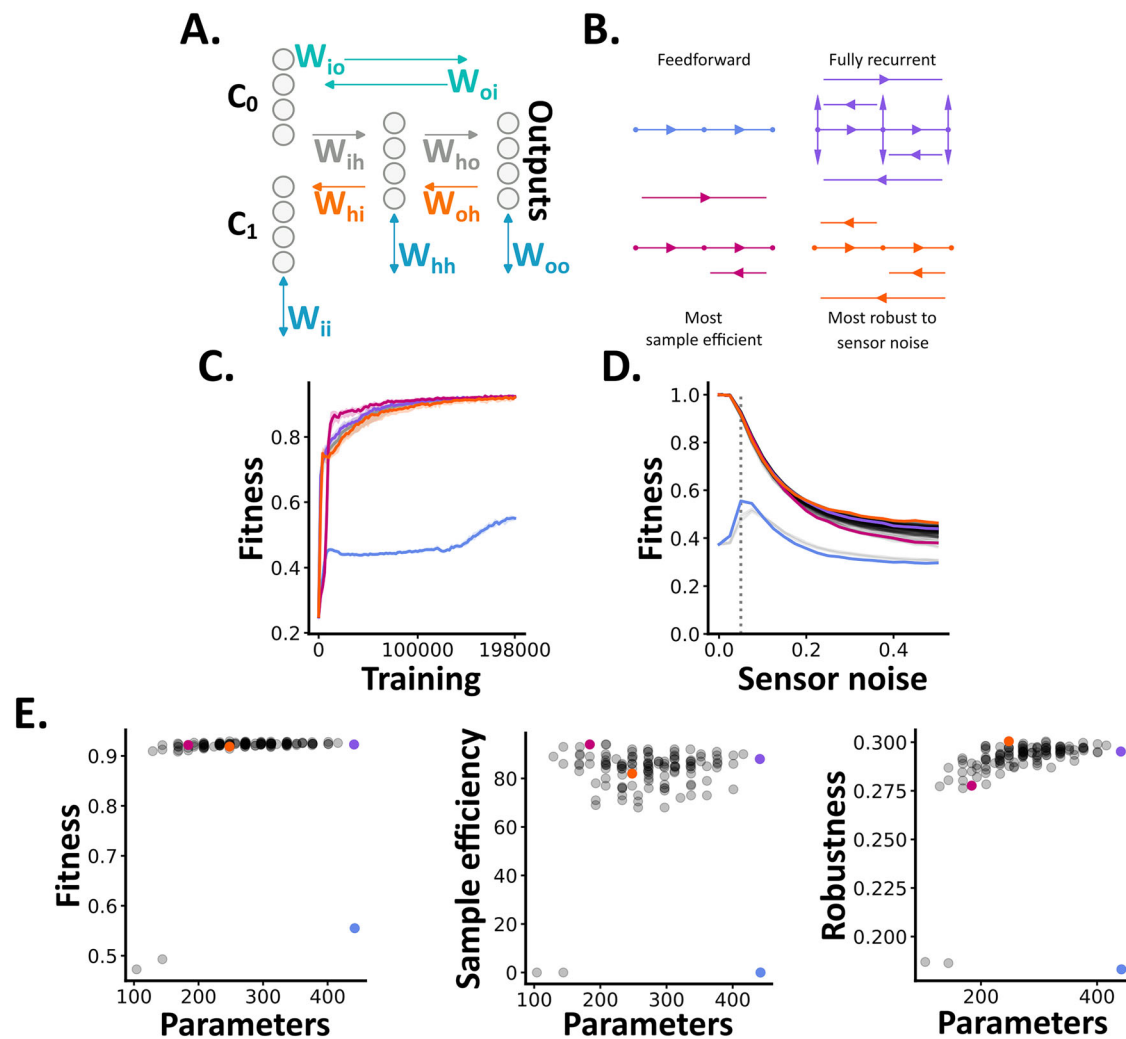


Fig. 2 | Partially recurrent neural networks. **A** Each architecture is composed of: input (*i*), hidden (*h*) and output units (*o*), two feedforward weight matrices W_{ih} and W_{ho} and any combination of the other 7 pathways. **B** 4 example architectures. **C** Test fitness across training on the track maze task, with a gap, for the 4 architectures shown in **(B)** (with matching colours), and all trained networks (grey). For each, we plot the median (solid line) and interquartile range (shaded surround). Note that here, and for subsequent panels, feedforward denotes the large-feedforward

architecture. **D** Each architecture's median fitness as a function of increasing sensor noise on the track maze task with a gap. All architectures are drawn in grey, with those from **(B)** highlighted. The dashed vertical line denotes the level of sensor noise during training. **E** Each architecture's number of trainable parameters vs median: fitness (left), sample efficiency (middle) and robustness to sensor noise (right). Each architecture is shown as a grey circle, with the architectures in **(B)** highlighted.

Partial recurrence enables efficient, yet performant solutions

We began by training all 128 architectures on the track maze task with a gap (Fig. 1C, right) using relatively small networks with 8 hidden units. To train each network's weights, we used reinforcement learning, specifically deep Q-learning^{27,28}. At each time step, pRNN agents act based on their current network output, without iterating to a steady state. To capture intra-architecture variability across random seeds²⁹, and to enable fair inter-architecture comparisons, we trained 50 networks per architecture, yielding 6400 trained networks.

Fully recurrent networks learned the task well, achieving a median fitness of 0.92 ± 0.01 std (Fig. 2C). By contrast, only two architectures were consistently unable to learn the task well: the pure layer-layer feedforward architecture (0.47 ± 0.02 std), and an architecture with additional forward skip connections W_{io} (0.49 ± 0.02 std). Notably, larger feedforward networks, with the same number of learnable parameters as the fully recurrent networks, were still unable to learn the task well (0.56 ± 0.01 std)(Fig. 2C). This is due to the fact that neither of these architectures retain any information from prior time steps, and so cannot solve tasks which require some form of memory. Strikingly, the remaining 125 architectures all achieved fitnesses close to the fully recurrent network (with median values between

0.91 and 0.93); despite having distinct pathways and far fewer learnable parameters—in some cases as low as $\frac{1}{4}$ of the fully recurrent network (Fig. 2E, left). While this may not be surprising, given the simplicity of the task, what is perhaps more surprising is that 16 architectures learned significantly faster than the fully recurrent network (Fig. 2C, E middle). And, 9 were significantly more robust to noise (Fig. 2D, E right). The fastest learning and most robust architectures are shown in Fig. 2B.

These results demonstrate that many architectures can implement robust behaviours; equivalent, or even better than a fully-connected network. Though, would this hold true in more complex tasks?

To test this, we designed a 2-parameter family of multisensory mazes. Specifically, we create maze structures (walls and paths) using the Aldous-Broder algorithm^{30,31} and fill these with sensory gradients leading towards the goal (Fig. 3A). This allows us to consider the more complex case where agents must navigate in four directions. Then, we introduce additional complexity by removing some fraction of walls—creating loops and open spaces (Fig. 3A), and some fraction of sensory cues—creating gaps as in the track maze (Fig. 3A). We focus on three scenarios: mazes with all walls and cues (M), mazes with 10% of walls removed (M_{sw}) and mazes with 10% of cues removed (M_{sc}). For each of these scenarios, we trained all 128

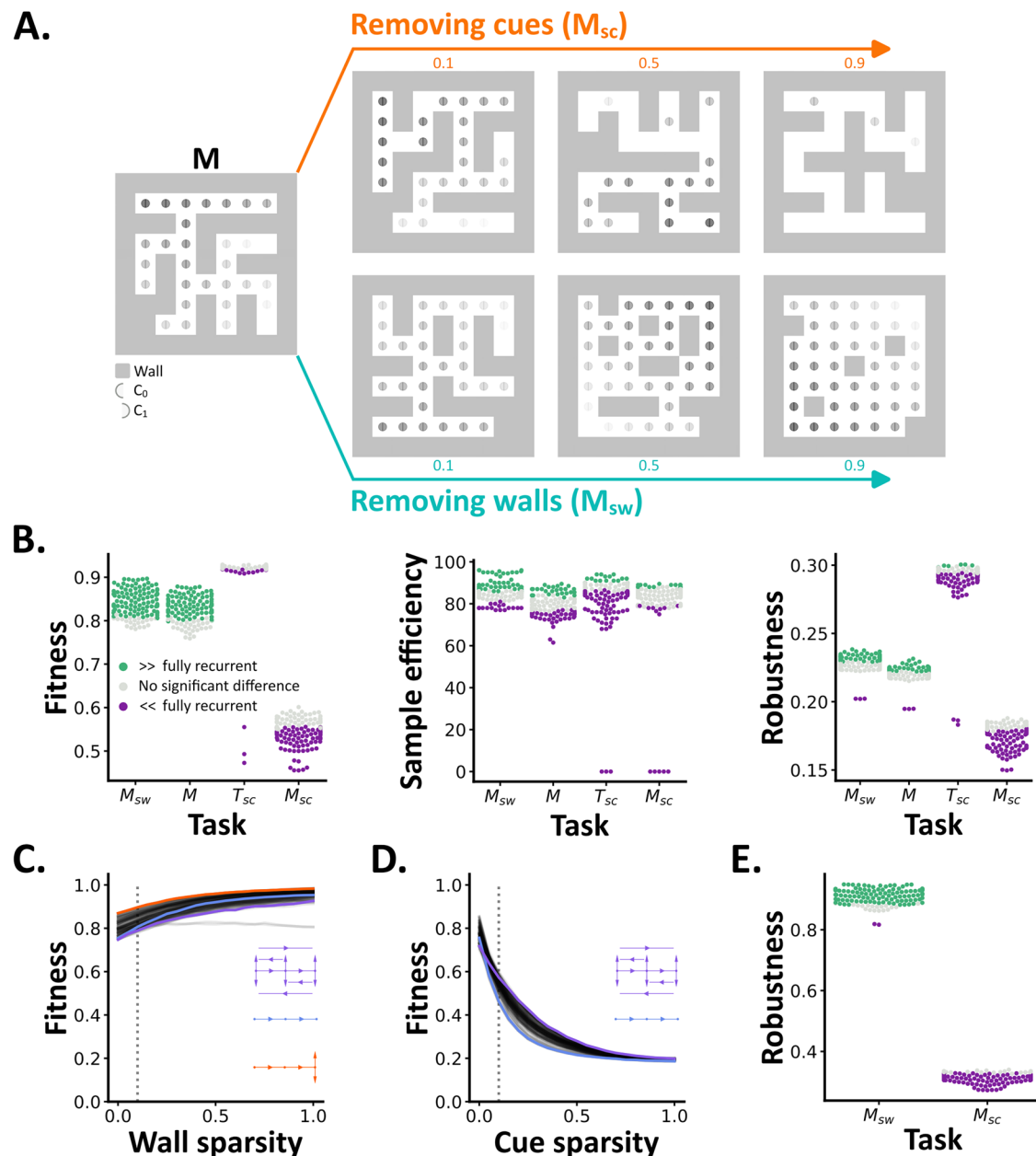


Fig. 3 | Multisensory mazes. **A** We create maze structures, shown with walls in grey and paths in white, and fill these with gradients of sensory cues (circles). We introduce further complexity by randomly removing some fraction of walls or cues. Seven maze structures are shown with various levels of wall and cue sparsity. **B** The median fitness (left), sample efficiency (middle) and robustness to sensor noise (right) across 50 trained networks, per task, for all 128 architectures (circles). Architectures which perform significantly better or worse ($p < 0.01$) than the fully

recurrent architecture are shown in green and purple, the rest are shown in grey. Each architecture's median fitness as a function of increasing wall (**C**) or cue sparsity (**D**). All architectures are drawn in grey, with select architectures highlighted in colour. Dashed vertical lines denote the level of sparsity during training. Note that here, feedforward denotes the large-feedforward architecture. **E** Statistical comparisons, as in (**B**), comparing each architecture's robustness to wall or cue sparsity to the fully recurrent architecture.

architectures, plus a large feedforward network, with 50 repeats per architecture. Combined with the networks from the track task with a gap (T_{sc}), this yields over 25,000 trained networks.

When navigating in environments with dense sensory cues (M_{sw} and M), every pRNN architecture achieved equivalent or significantly better fitness than the fully recurrent architecture (Fig. 3B, left); with the best achieving 11 and 9% higher fitness. Moreover, 30% of architectures were significantly more sample efficient (Fig. 3B, middle) and 47% were significantly more robust to noise (Fig. 3B, right). Conversely, when navigating in environments with sparse sensory cues (T_{sc} and M_{sc}), 41% of comparisons (across architectures and metrics) were significantly worse than the fully recurrent architecture, while the rest (54%) were mostly equivalent

(Fig. 3B). Additional experiments with homogenous activation functions per layer, and smaller and larger networks, support these findings (Supplementary Section 3).

Together, these results demonstrate that partially recurrent networks can outperform fully recurrent networks in environments with dense sensory cues, and perform equivalently in more complex scenarios. Though perhaps the main advantage of fully recurrent networks is that they are better able to adapt, or generalise, to novel situations?

To explore this, we tested how well networks trained at one level of wall (M_{sw}) or cue sparsity (M_{sc}) performed when tested across a wider range. In general, fitness increased with increasing wall sparsity (Fig. 3C)—as the average path length from start to goal decreased, but decreased with cue

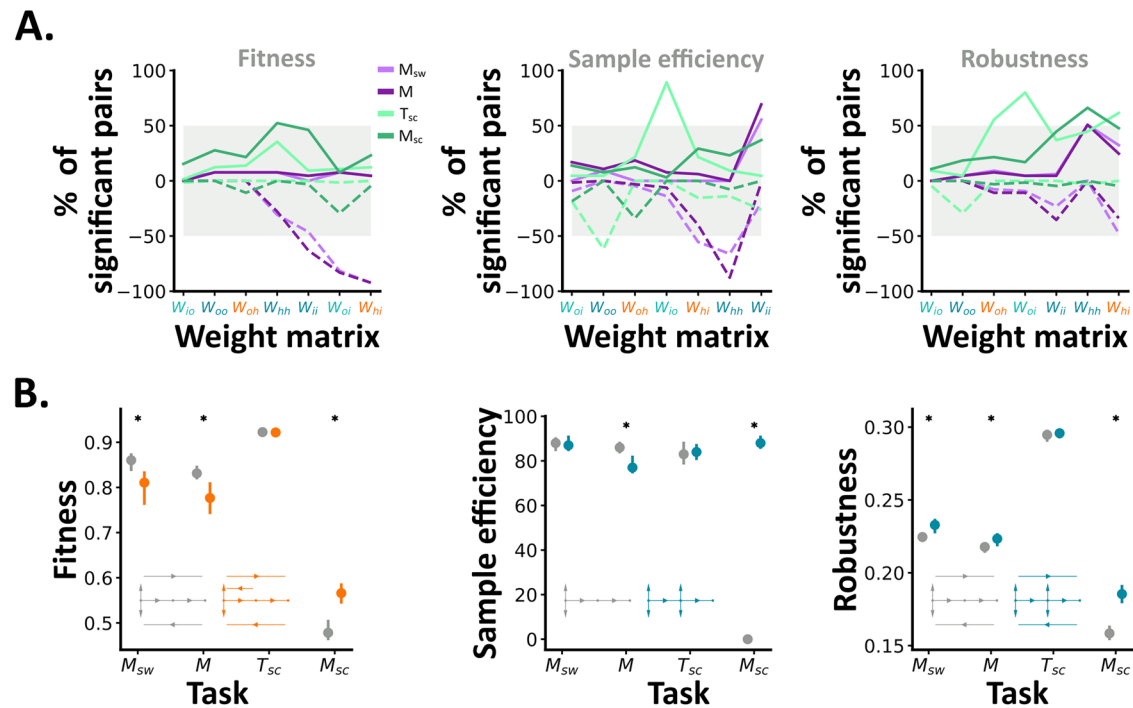


Fig. 4 | Counterfactual analysis. **A** The percentage of counterfactual pairs in which the architecture with an additional weight matrix has significantly greater (solid lines) or less (dashed lines): fitness (left), sample efficiency (middle) or robustness to sensor noise (right). Results from each task are shown in distinct colours. The grey underlay marks the region in which most differences are context- dependent.

B Example counterfactual pairs with task-dependent differences in fitness (left), sample efficiency (middle) and robustness to sensor noise (right). For each pair, we plot the two architectures, and the median and interquartile range for each task, across $n = 50$ independent networks. Statistically significant differences, according to a two-sided, asymptotic Mann–Whitney U test, are marked with a * ($p < 0.01$).

sparsity (Fig. 3D)—as the number of informative cues fell. Most pRNN architectures generalised significantly better to wall sparsity than the fully recurrent architecture, while the opposite was true for cue sparsity (Fig. 3E); illustrating that fully recurrent networks do not necessarily generalise better.

In sum, these results demonstrate that partial recurrence enables efficient, yet performant solutions. However, how does structure link to function across these architectures and tasks? For instance, in the track maze task the architecture which is most robust to noise, includes all three backwards pathways and the most sample efficient architecture includes forward skip connections (Fig. 2B). Do these examples highlight general structure-function principles?

Each pathway's functional impact is circuit and task-dependent

To explore the link between network structure and function, we used counterfactual analysis^{32,33} to ask: do pairs of architectures, which differ by only a single pathway, differ significantly in their function? The architectures in Fig. 4B (left) are one such pair, as they are identical, except for the backwards connection W_{hi} —present in one and absent in the other.

Specifically, for a given measure of function y and a pair of architectures which differ by one pathway i , we test if the distribution $y_{i=1}$ (across 50 networks) is stochastically less than or greater than $y_{i=0}$ (another set of 50 trained networks). From this approach, we draw two main conclusions.

First, each pathway's functional impact is *circuit dependent*. If a pathway had a consistent effect on function, for example, increasing robustness to sensor noise, then all or most networks with that pathway would differ significantly from their counterfactual pairs. Instead, for almost all pathways and functional metrics, we observe significant differences in less than 50% of pairs (Fig. 4A). For example, on average across tasks, only 42% of networks with backwards hidden-input (W_{hi}) connections are significantly more robust to sensor noise than their pairs without (Fig. 4A, right panel, rightmost x tick, solid lines).

Second, each pathways functional impact is *task dependent*. For instance, when trained and tested in mazes with dense sensory cues (M_{sw}

and M), 92% of networks with additional hidden-input connections (W_{hi}) were significantly less fit than their pair; conversely, in mazes with sparse cues (T_{sc} and M_{sc}), only 2% were less fit (Fig. 4A, left, purple vs green dashed lines). These task-dependent effects were even evident at the level of single circuit pairs. For example, the architecture with W_{hi} connections in Fig. 4B (left panel, orange) was significantly less fit than its pair (grey) in dense mazes, yet performed equivalently or better in sparse mazes. We observed similar effects for both sample efficiency and robustness to sensor noise (Fig. 4B, middle and right).

Overall, these data demonstrate that, depending on the circuit and task, small changes to network structure can lead to dramatic changes in function. Though, why do these changes alter network function?

From structure to function via computational traits

To explore why changes to network structure alter network function, we analysed our trained network's learned weights and input-output transformations. To do so, we measured three computational traits from every trained network: their pathway magnitudes, input-output sensitivity and memory dynamics. Then, we quantified how predictive these traits are of network function. For example, given a network's memory dynamics, how accurately can one predict its robustness to noise? These results lead us to propose a conceptual model, whereby: networks with different structures learn distinct computational traits, which determine their function.

One approach to interpreting neural networks is to analyse their learned weights. In this direction, we calculated the magnitude of each weight matrix ($\|W_i\|_F$) in every network. These values were surprisingly consistent across architectures and tasks: with large feedforward and forward skip magnitudes, and relatively low values otherwise (Fig. 5A).

Another approach is to analyse their learned input-output transformations; that is, how networks map their sensory inputs, plus any hidden states, to decisions/actions. Prior studies have investigated this by measuring how sensitive networks' outputs are to changes in their current input—using the norm of the input-output Jacobian³⁴. Or their memory for past inputs—

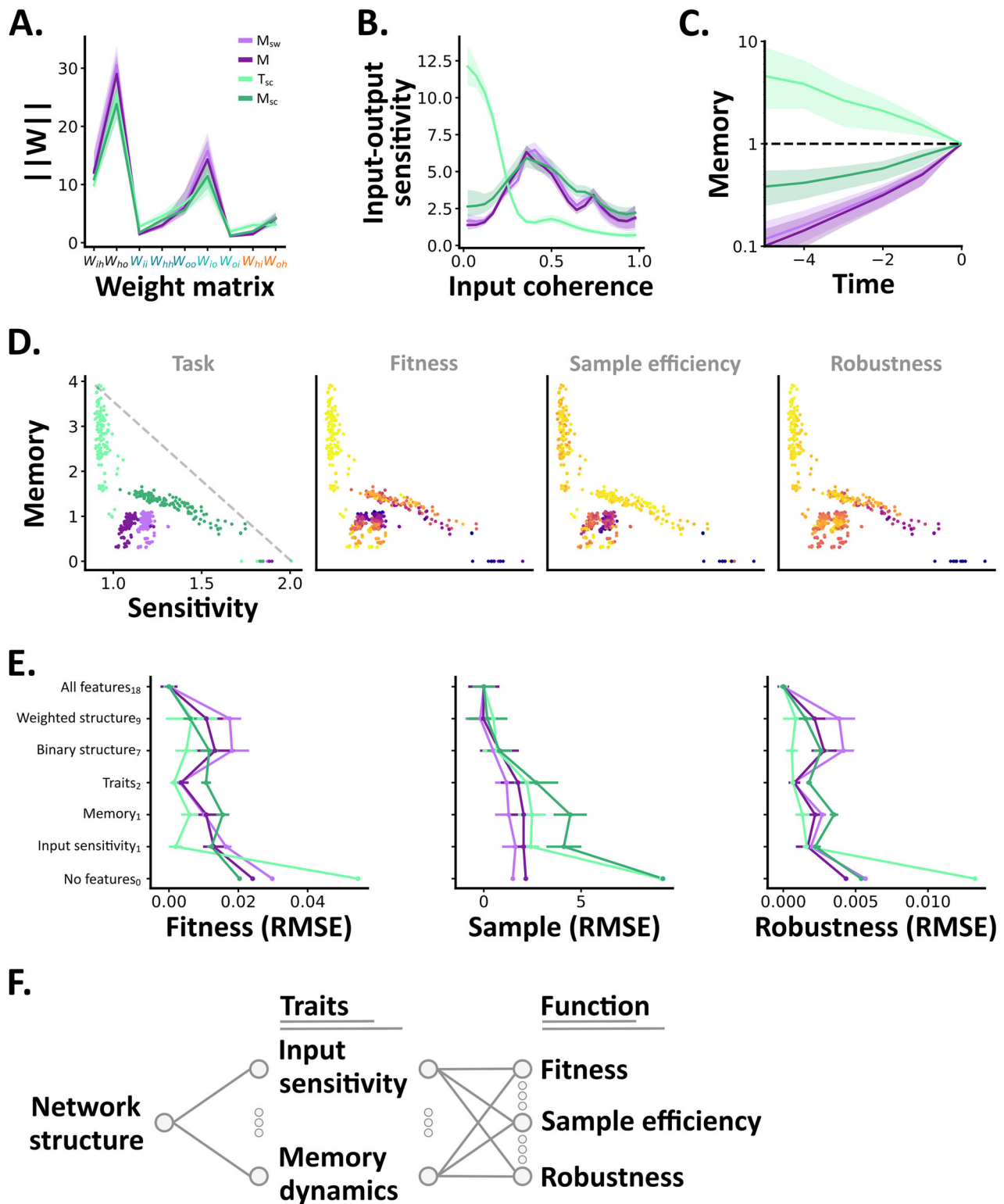


Fig. 5 | Structure-traits-function. **A** The learned magnitude of each weight matrix. Results from each task are shown as median (solid line) plus interquartile range (shaded surround) across all trained networks. **B** Input-output sensitivity as a function of input coherence plotted as in (A). **C** Memory as a function of time plotted as in (A). **D** Input sensitivity vs memory (log-log plots). Each circle represents the median of a single architecture from 50 trained networks per task. In the left panel, architectures are coloured by the task they are trained on. The dashed grey line illustrates a trade-off between input sensitivity and memory—simply plotted from: $(\min(x), \max(y))$ to $(\max(x), \min(y))$. In the other subplots, circles are coloured by

their fitness, sample efficiency and robustness from low (blue) to high (yellow) within each task. **E** Random forest models trained to predict fitness (left), sample efficiency (middle) and robustness to sensor noise (right) using different computational traits; subscripts note the number of input features per model. We plot the mean error plus standard deviation from $n = 10$ -fold cross-validation, with individual lines per task. **F** A conceptual model showing how networks with different structures learn solutions with distinct computational traits, which influence their function. The small circles, between labelled nodes, suggest additional computational traits or functional metrics.

by quantifying how well prior inputs can be reconstructed from the current network state³⁵ or how much information about past inputs is present in the current state³⁶, at different temporal lags. Building on these ideas, we measure how sensitive networks' outputs are to changes in their inputs at different timescales—by computing the norm of the input-output Jacobian across different temporal lags.

To do so, we collected data from 1000 test trials per network and defined a metric to describe how sensitive each network's outputs A , at time t , are to its inputs C at time $t - k$ ($k \geq 0$). Where the elements of C represent the inputs from each sensor, and the elements of A represent the network's output for each possible action.

$$S(t, k) = \left\| \frac{\partial A_t}{\partial C_{t-k}} \right\|_F \quad (1)$$

We define a network's *input-output sensitivity* as the mean $\langle \cdot \rangle$ of its sensitivity at time t to inputs at time t : $S = \langle S(t, 0) \rangle$. Calculating this metric from networks trained on the track maze task (T_{sc}), revealed an intuitive result: when the inputs to a network clearly indicate one direction (e.g. left > right), input sensitivity is low. And so, networks are relatively insensitive to changes in their inputs. By contrast, when inputs are ambiguous (i.e. left $\approx$ right), networks are highly sensitive to changes in their inputs (Fig. 5B, light green line). Networks trained on our other tasks showed a different sensitivity profile, with low sensitivities at both low and high levels of input coherence, and high values in between (Fig. 5B, dark green and purple lines). This may suggest that, in these tasks, networks learn to: filter out ambiguous inputs, detect subtle differences in semi-ambiguous inputs and simply follow clear inputs. Though this is just one possibility, and individual architectures may leverage different strategies.

We define memory at lag k to be the ratio of the sensitivity to past compared to current inputs, $M_k = \langle S(t, k) / S(t, 0) \rangle$. As such, as M_k increases, networks transition from being completely insensitive to their prior inputs (when $M_k = 0$) to being more sensitive to their prior than current inputs (when $M_k > 1$). In the track maze task (T_{sc}), we found that most networks' memory increased further back in time (Fig. 5C, light green line), meaning that their outputs at any time are more sensitive to changes in the past than the present. This approach works for this maze design, as, assuming an agent sets off in the correct direction, repeating its first action will take it to the goal.

However, in more complex scenarios (M_{sw} , M and M_{sc}), most networks' memories decayed with time (Fig. 5C, dark green and purple lines), showing that they are more sensitive to recent than past inputs; though the rate of this decay was task-dependent. Specifically, networks trained on the maze with sparse cues (M_{sc}) were more sensitive to their past inputs than networks trained on the tasks with dense cues (Fig. 5C dark green vs purple lines).

To explore the link between these computational traits and network function, we began by visualising the relation between each architecture's median input sensitivity, total memory and functional measures (Fig. 5D). This showed that different architectures learn distinct input-output sensitivities and memory dynamics and revealed a trade-off between the two, where improving one impairs the other. This is likely due to the fact that retaining hidden states improves memory, but combining those hidden states with current inputs interferes with the direct mapping of inputs to outputs.

Finally, to quantify the relation between these traits and our functional measures, we used random forest regression models—which can capture complex, nonlinear relations between different combinations of features and continuous values. As a lower bound, we calculated the error from simply predicting the mean of each functional measure. As an approximate upper bound, we fit random forest models to all 18 of our features of interest.

Random forest models trained to predict function from each network's weighted structure performed similarly to models trained on binary network structures (Fig. 5E) despite having an additional two parameters and using continuous instead of binary values. This demonstrates that knowing each pathways magnitude does not provide much additional information

over simply knowing if it is present or absent. By contrast, both input sensitivity and memory were highly predictive of function, achieving errors far lower than guessing the mean and close to or even better than models with many more features (Fig. 5E). For instance random forest models trained to predict robustness to sensor noise from memory achieved a mean normalised error of 0.002 across tasks; better than our lower bound of 0.007 and as good as models trained on each network's weighted structure (0.002)—despite using $\frac{1}{3}$ the number of features (Fig. 5E, right).

All together, these results lead us to propose that networks with different structures learn distinct input-output sensitivities and memory dynamics, which in turn influence their function (Fig. 5F). This model is primarily supported by two pieces of evidence. First, across architectures, we observe a diverse range of input-output sensitivities and memory dynamics (Fig. 5D). Second, knowing these traits allows us to predict network function with high accuracy (Fig. 5E). As these two traits describe how networks process and retain information, they are likely to be relevant to many tasks and would be interesting to approximate from real neural circuits. More broadly, we hope that providing a conceptual stepping stone between structure and function will help to bridge this gap.

Discussion

Many neural circuits sit somewhere between feedforward and fully recurrent. How does this structure shape their function? To explore this question, we define: a neural network model (pRNNs), a new set of tasks (multi-sensory mazes) and tools for linking network structure to function via computation. Below, we discuss our findings in the context of neuroscience and machine learning.

In biological circuits, connections cost both energy and space. So, a priori, partially recurrent networks should be more efficient than fully recurrent networks. However, does this efficiency come at the cost of performance?

While prior studies have found that sparsity comes with at least a small cost to performance in task-trained models^{15,17–20}, other work has shown that sparse architectures can support robust associative memory and efficient computation³⁷. Our findings reconcile these perspectives by demonstrating that partially recurrent networks—which are both sparse and bidirectional—can achieve high performance, sample efficiency and robustness, depending on their structure and the task. We attribute this difference to two aspects of our approach.

First, our choice of tasks. In general, prior studies have focussed on cognitive neuroscience or machine-learning style tasks in which networks must generate a single response to a sequence of inputs^{13,15,17,18}. By contrast, our maze tasks require networks to continuously translate the inputs from noisy sensors (and internal states), into actions; while being simple enough for us to focus on relatively small neural networks (with between 104 and 442 learnable parameters) and to train an exceptionally large number of networks (50 per architecture and task $\rightarrow$ over 25,000 trained networks). Working at this level of task complexity, resource constraints and statistical power allowed us to observe differences between classical models of multisensory integration (section 'Optimal Multisensory Algorithms are Brittle') and between our partially recurrent architectures (section 'Partial Recurrence Enables Efficient, yet Performant Solutions'). We therefore anticipate that these tasks will serve as a valuable foundation for future studies.

Second, our choice of model. Many prior studies have varied network structure by taking recurrent neural networks and constraining their hidden-hidden weight matrix (W_{hh}) to generate networks with distinct properties. However, within our framework, these networks are all variations on a single architecture, and so, by comparing every pRNN architecture we explored a far more diverse set of network structures.

That said, both of these choices come with their own caveats. In terms of our tasks, we use dense rewards to train our networks, which may influence the types of solutions networks learn as well as our inter-architecture results. In terms of our models, we only consider networks with a single hidden layer. Future work should aim to explore deeper models, closer to biological networks in both their structure and function. However,

the number of pRNN architectures scales super-exponentially with the number of layers (L): $2^{L^2-(L-1)}$. As such, complete searches of the space, as we present here, quickly become infeasible. One solution to this would be to develop search algorithms to navigate the space of pRNN architectures—akin to how³⁸ explored the space of symbolic cognitive models. However, our results demonstrate that even relatively small changes to network structure can lead to dramatic changes in function (section 'Each Pathway's Functional Impact is Circuit and Task Dependent'), suggesting that this may be challenging. Another solution would be to use neuroevolution algorithms^{39,40}. These algorithms optimise a given function (e.g. task performance) by iteratively evaluating and varying a population of neural network architectures. The advantage of this approach is that it enables the discovery of diverse solutions⁴¹. The downside is that it can be hard to interpret large collections of neural networks.

In this direction, we measured three computational traits from all of our trained networks. We found that different architectures learned distinct input-output sensitivities and memory dynamics, and observed a trade-off between the two (section 'From Structure to Function via Computational Traits'). As these metrics essentially quantify how networks process and retain information, they are likely to be relevant for many tasks in both neuroscience and machine learning, and could even be approximated from real circuits via connectome-based models^{42,43}. Though other computational traits may be more relevant for different tasks or architectures, developing these remains an exciting direction for future work.

Finally, we demonstrate that our computational traits are highly predictive of function. For instance, knowing a network's memory dynamics allows us to accurately predict its robustness to noise (section 'From Structure to Function via Computational Traits'). These results lead us to propose that networks with distinct architectures learn solutions with different computational traits which influence their function. This conceptual model suggests a shift in focus—from identifying which network structures yield different functions, to understanding why they do so, via their computational traits. And sparks numerous questions: to what extent do computational traits explain the functional differences between other neural network architectures? Could optimising networks for specific traits, rather than specific tasks, yield more flexible solutions? And could traits like these be key to understanding inter-circuit, inter-individual or even inter-species differences? Ultimately, our work provides a springboard for addressing fundamental questions, like these, related to how structure sculpts function in artificial and biological networks.

Methods

In brief, we compare how agents with different policies perform on a set of maze tasks (section 'Multisensory mazes'). These tasks are set in grid environments, with impassable walls, paths the agents can move on, and sensory cues which indicate the shortest path through each maze. To perform these tasks, each agent implements a three-step sensation-action loop (section 'Agents'). First, they sense their local environment. Often, we add independent Gaussian noise to each sensor at each time step. Next, they implement a policy—either an algorithm (section 'Multisensory Algorithms') or a forward pass through a neural network (section 'Partially Recurrent Neural Networks'). Finally, agents choose and implement an action from a discrete set; either pausing or moving in a cardinal direction. To compare models, we measure three functional metrics (section 'Functional Metrics') and three computational traits (section 'Computational Traits'). For statistical comparisons across models, we use Mann–Whitney U tests with corrections for multiple comparisons (section 'Analysis').

Multisensory mazes

All tasks are set on discrete grids, with tracks the agents can move on, and impassable walls. All tasks are run in discrete time, for a fixed number of time steps (see Table 1 for task hyperparameters).

Track maze. In this task, agents start in the middle of a track (Fig. 1C, left). Their aim is to reach a goal location at either the left or right end of

Table 1 | A table noting the hyperparameters we used for each task: track mazes with sensory gaps one step prior to the goal (T_{sc}), 2D mazes with dense walls and sensory cues (M), 2D mazes with 10% of walls removed (M_{sw}), and 2D mazes with 10% of sensory cues removed (M_{sc})

Task	Size ($n \times n$)	Trial length	# Training	# Testing	Wall sparsity	Cue sparsity
T_{sc}	11	6	200,000	1000	0.0	1 gap
M	9	20	400,000	1000	0.0	0.0
M_{sw}	9	20	400,000	1000	0.1	0.0
M_{sc}	9	20	400,000	1000	0.0	0.1

denotes the number of mazes used for training and testing.

the track within a fixed number of time steps. To do so, they must use sensory cues from two channels (C_0 and C_1). In each trial, one channel's cues are linearly spaced between 0 and 0.5, leading from start to goal. While the other channel's cues are linearly spaced between 0 and 0.5 from start to both ends of the track, and so are ambiguous. Across trials, we vary both the goal location and whether each channel is leading or ambiguous, yielding 4 configurations. Across configurations, the optimal strategy is to sum the information over both channels and head in the direction with the greatest total. For example, at the first time step, with the goal on the left and no sensor noise, an agent may sense:

$$C_{0L}(0.1) + C_{1L}(0.1) > C_{0R}(0.1) + C_{1R}(0.0)$$

We also consider two variants of this task.

Nonlinear track maze—we double the strength of the cues leading away from the goal (Fig. 1C, middle). This ensures that, at the first time step, there is an equal amount of information either side of the agent, and a linear strategy is uninformative. Following the example above, the agent would instead sense:

$$C_{0L}(0.1) + C_{1L}(0.1) = C_{0R}(0.2) + C_{1R}(0.0)$$

Now the optimal strategy is to use a nonlinear function to weight the multisensory signal (leading towards the goal) more strongly than the unisensory signal (leading away from the goal).

Track maze with a gap—we introduce a gap in the signals leading towards the goal, by setting both channels to zero m steps prior to the goal (Fig. 1C, right), i.e. when $m = 1$, the location prior to the goal has no sensory information, when $m = 2$, the two locations prior to the goal have no information. Consequently, the optimal strategy is to set off in the direction with the greatest total signal at the first time step and to persist in that direction.

General mazes. In these tasks, agents start in one corner of a maze (set on an $n \times n$ grid) and must navigate to a goal location in another corner (Fig. 3A). To generate each maze's structure (walls and paths), we use the Aldous–Broder algorithm^{30,31}. We then choose random corners for the start and goal locations, and fill sensory cues into each channel, based on the shortest-path distance from each location to the goal. These cues are linearly spaced between: 0, the furthest point from the goal (which is not necessarily the start position), to a maximum of $1/c$ at the goal location, where c is the number of sensory channels. Then, we introduce additional complexity by varying two parameters.

Wall sparsity—prior to filling in the sensory cues, we randomly remove some fraction of walls (excluding the maze's outer wall). Increasing this parameter from 0 to 1 interpolates between a perfect maze and an open field.

Cue sparsity—after filling in the sensory cues, we randomly set some fraction of cues (in both channels) to 0. Increasing this parameter from 0 to 1

interpolates between a maze with sensory cues at every location to a maze with no sensory cues.

Agents

Each agent occupies a location on a discrete grid, and implements a three-step sensation-action loop (Fig. 1A):

1. *Sense*—each of the agent's sensors reads one channel's signal from one adjacent square (e.g. C_{0L} , C_{1L} , C_{0R} , etc.). Independent Gaussian noise is optionally added to each sensor at each time step. If the agent is at (x_a, y_a) , then the reading from channel i in direction j is given by $C_{ij} = \text{Cue}(i, x_a + \Delta x_j, y_a + \Delta y_j) + \mathcal{N}(0, \sigma^2)$, where $\sigma = 0$ in the noise-free case. Sensor values are clipped to the range $[0, 1]$.
2. *Policy*—maps sensor inputs to an output vector with one value for each possible action. Policies are either multisensory algorithms (section 'Multisensory Algorithms') or neural networks (section 'Partially Recurrent Neural Networks'). Some policies are stateless, i.e. they map sensations to actions independently at every time step, but others combine their sensory inputs with internal (hidden) states.
3. *Act*—choose and implement an action: using $\text{argmax}(A)$ where A represents a vector of potential actions (either pausing or moving in a cardinal direction). When multiple actions are equally likely, agent's choose randomly between them. When an agent tries to move into a wall, it instead remains in place for that time step.

Multisensory algorithms

Throughout, we use the term *algorithm* to describe policies, which map a vector of sensory inputs (C) to a vector of potential actions (A), without using any learnable parameters or hidden states. C represents the agent's sensors (left, right, up and down), where C_{ij} denotes the values of sensory channel i in direction j . A represents possible actions (move left, right, up, down), where A_j is the value of the action in direction j . At each time step, agents choose $\text{argmax}(A)$ (with randomisation between equal values). We set $A_j = 0$ for all j at the start of every time step.

We explore the following 7 algorithms (Fig. 1B).

Random motion. Choose randomly between all actions:

$$A_j = 0 \quad \text{for all } j \quad (2)$$

Unisensory. Map each sensor's value, from channel i , directly to the corresponding output action:

$$A_j = C_{ij} \quad (3)$$

2-look binary. Within each channel, identify the sensor with the greatest activation $d(i)$. If they agree, choose the corresponding action. Otherwise, choose randomly between the suggested actions:

$$d(i) = \text{argmax}_j(C_{ij}) \quad (4)$$

For each i update:

$$A_{d(i)} \leftarrow 1 \quad (5)$$

2-look max. Within each channel, identify the sensor with the maximum activation $d(i)$. Add the activations from these sensors to their corresponding outputs.

For each i update:

$$A_{d(i)} \leftarrow A_{d(i)} + C_{i,d(i)} \quad (6)$$

Max-dv. Map the maximum value from each sensor pair (e.g. the left sensors) to the corresponding actions:

$$A_j = \max_i(C_{ij}) \quad (7)$$

Linear fusion. Sum the sensor signals across channels:

$$A_j = \sum_i C_{ij} \quad (8)$$

Nonlinear fusion. Sum the sensor signals across channels and add a nonlinear term. Here, we use the square root of their product, but any nonlinearity would suffice²⁴:

$$A_j = \sum_i C_{ij} + \sqrt{\prod_i C_{ij}} \quad (9)$$

Partially recurrent neural networks

Model. Given a neural network with: an input, a hidden and output layer, there are 9 possible weight matrices:

- *Feedforward* connections—from the input to the hidden layer (W_{ih}) and the hidden to the output layer (W_{ho}).
- *Lateral* connections—between the units within each layer: input-input (W_{ii}), hidden-hidden (W_{hh}) and output-output (W_{oo}). Note that these include self (unit _{b} -unit _{b}) connections.
- *Skip* connections—from the input to output layer (W_{io}) and output to input layer (W_{oi}).
- *Backwards* connections—from the hidden to input layer (W_{hi}) and output to hidden layer (W_{oh}).

Assuming the feedforward connections (W_{ih} and W_{ho}) are always present, and each of the other 7 matrices can be present in any combination, yields $2^7 = 128$ architectures. Relaxing this assumption would generate $2^9 = 512$ architectures, though some of these would have no path from input to output. When none of these 7 additional matrices are present, the model is purely feedforward. When all of the additional matrices are present, the model is fully recurrent as every unit connects to every other unit bidirectionally. We term the remaining 126 architectures, between these two extremes, *partially recurrent neural networks*.

Units in the input, hidden and output layers (respectively) use linear, rectified linear (ReLU) and softmax activation functions, with no bias terms.

In the feedforward case, we write the forward pass (or policy π) as:

$$A = g(W_{ho}f(W_{ih}C)) \quad (10)$$

Where A represents a vector of potential actions (move left, right, up, down or pause). C represents the agent's sensors (left, right, up, down from two channels). W_{ih} and W_{ho} are the two feedforward weight matrices, f is ReLU, and g a softmax across possible actions.

In the fully recurrent case, the forward pass (or policy π) is:

$$\begin{aligned} I &= C + W_{ii}I_{t-1} + W_{hi}H_{t-1} + W_{oi}A_{t-1} \\ H &= f(W_{ih}I + W_{hh}H_{t-1} + W_{oh}A_{t-1}) \\ A &= g(W_{ho}H + W_{oo}A_{t-1} + W_{io}I) \end{aligned} \quad (11)$$

Where I and H represent the activations of the input and hidden units at time t . And all hidden states are set to zero at the start of each trial.

In partially recurrent cases, each model's forward pass follows Eq. 11 with any absent weight matrices (determined by the model's structure) set to zero. See Table 2 for model hyperparameters.

Training. To optimise each model's weights, we used deep Q-learning^{27,28}. For each time step within each training episode, each agent senses its environment, chooses an action and acts. We use an epsilon-greedy action selection policy, where agents initially explore by choosing actions at random, but increasingly rely on their learned policy π as training progresses. Specifically, we define a vector ϵ with one value per training maze, which decreases step-wise from 0.95 to 0.25 over training.

Table 2 | A table listing the hyperparameters we used for our neural network models

Hyperparameters	Value
Input units (#)	8
Hidden units (#)	8
Output units (#)	5
Sensor noise scale	0.05
Learning rate*	0.001
Gamma*	0.9
Epsilon decay schedule*	See caption
Movement penalty*	0.1
Gradient clipping*	See caption

Hyperparameters only used during training are marked with a *. We used Epsilon-greedy action selection, with a linear, stepped decay schedule from 0.95 to 0.25. We used the clip_grad_norm function from PyTorch⁴⁴ to clip our gradients to a maximum value of 10.

At each time step, the agent selects an action a_t according to:

$$a_t = \begin{cases} \text{random action,} & \text{with probability } \epsilon \\ \text{argmax}(A), & \text{with probability } 1 - \epsilon \end{cases} \quad (12)$$

At each time step, agent's estimate the value of the action they took (termed the Q -value) using their policy (a forward pass through a neural network):

$$\hat{Q} = \pi(C_t, I_{t-1}, H_{t-1}, A_{t-1}) \quad (13)$$

We calculate the target Q -values:

$$Q = r + \gamma \max(Q_{t+1}) - 0.1 \quad (14)$$

Where the reward r is proportional to the reduction in shortest-path distance to the goal, scaled by a factor of 2. γ is a discount factor that determines how strongly the agent values future rewards. $\max(Q_{t+1})$ is the maximum predicted Q -value at the agent's new location, and the constant -0.1 introduces a small penalty for each time step to encourage faster solutions.

Next, we calculate the mean squared error between the two (the loss):

$$L = \frac{1}{2} (Q - \hat{Q})^2 \quad (15)$$

We accumulate these losses, then at the end of every episode, update the agent's weights using backpropagation. Episodes terminate after a fixed number of time steps or the goal being reached. As each model's forward pass is written in PyTorch⁴⁴, we rely on autodifferentiation for each model's backward pass.

Functional metrics

To describe each model's behaviour, we define three functional metrics.

Fitness. To measure performance in our maze tasks we define the following fitness function:

$$\begin{aligned} \text{fitness} &= \frac{1}{2} (\text{speed} + \text{accuracy}) \\ \text{speed} &= 1 - \frac{t - t_{\min}}{t_{\max} - t_{\min}} \\ \text{accuracy} &= \frac{d_{\max} - d_t}{d_{\max}} \end{aligned} \quad (16)$$

Where t is the time taken to reach the goal, $t_{\min}$ is the quickest time possible, $t_{\max}$ is the maximum number of steps in the trial, $d_{\max}$ is the furthest distance from the goal and d_t is the agent's distance from the goal at time t . Note that when the goal is not reached $t = t_{\max}$. This fitness function ranges from 0—ending the trial at the furthest point from the goal, to 1—taking the shortest path from start to goal.

We define each model's fitness by testing it on 1000 mazes. For our neural network models, we do this after training with all parameters set as they were during training (i.e. in distribution).

Sample efficiency. To measure a network's learning speed, we record it's test fitness 100 times, at equally spaced intervals, throughout training. This generates a vector, per network, F_t —the test fitness after t epochs. We define sample efficiency as 100 minus the point at which: F_t is first $\geq 0.9 \times \bar{F}$ where $\bar{F}$ denotes the median fitness across the entire set of networks trained on a given task. As such, higher values represent better sample efficiency. Networks, which do not reach this threshold, are assigned a value of 0.

Robustness. To test an agent's robustness, we repeatedly test it on a set of 1000 mazes as we: increase its sensor noise (e.g. Fig. 1D) or remove an increasing fraction of walls (Fig. 3C) and sensory cues (Fig. 3D) from each maze. We test: 21 values of noise (linearly spaced between 0 and 0.5) and 21 values of wall and cue sparsity (linearly spaced between 0 and 1). For our pRNNs, we test robustness after training, with all other parameters kept in distribution. To summarise each agent's robustness, we compute the area under each curve (fitness vs noise or sparsity) using the trapezoidal rule.

Computational traits

To explore why changes to network structure alter network function, we measure three computational traits from every network after training.

Pathway magnitudes. For each network, we calculate the Frobenius norm $\| \circ \|_F = \sqrt{\text{Tr}(A^T A)}$ of each weight matrix $\| W_i \|_F$ (Fig. 5A).

Input-output sensitivity. Following training, we test each network on a set of 1000 mazes, with all parameters set in distribution. At every time step during these trials, we collect each network's inputs (C), and the states of their input (I), hidden (H) and output (A) units. For each input state, we run a forward pass through the trained network. Then, following Eq. 1, we calculate the input-output Jacobian and take the Frobenius norm ($\| \circ \|_F$); using the `Jacobian` and `norm` functions from PyTorch⁴⁴.

When input-output sensitivity is high, small changes in the network's inputs will lead to large changes in its outputs. Conversely, when input-output sensitivity is low, changes to the network's inputs affect the outputs less. To obtain a single input-output sensitivity value per network, we calculate the mean norm across all states.

To plot each network's input sensitivity as a function of input coherence (c) as in Fig. 5B, we let:

- $\vec{d}_j = [d_{j,x}, d_{j,y}]$ be the unit direction vector for direction $j \in \{\text{left, right, up, down}\}$,
- $m_j = \sum_i C_{ij}$ be the total magnitude of sensory input in direction j across all channels i ,
- $\vec{v} = \sum_j m_j \vec{d}_j$ be the net input vector,
- $M = \sum_j m_j$ be the total input magnitude.

Then we define input coherence c as:

$$c = \begin{cases} \frac{\|\vec{v}\|}{M}, & \text{if } M > 0 \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where $\|\vec{v}\|$ denotes the Euclidean norm of the vector $\vec{v}$, which is always between 0 and 1. When c is close to 1, the input state clearly indicates the goal direction, while when c is low, the inputs are ambiguous.

Memory dynamics. Following training, we test each network on 1000 test trials, with all parameters held in distribution, and collect all of the network's states at every time step. Following Eq. 1, we define a network's memory at temporal lag k ($k \geq 0$) as: $M_k = \langle S(t, k)/S(t, 0) \rangle$. Where $S(t, k)$ represents the sensitivity of the network's outputs at time t w.r.t. its inputs at time k , $S(t, 0)$ is its sensitivity to its current inputs, and $\langle \cdot \rangle$ denotes the mean across all input states. As such, when $M_k = 0$ a change in the network's inputs at lag k , will have no effect at time t (as in a stateless model). When $M_k > 1$, networks are more sensitive to their prior than current inputs. And, when $0 < M_k < 1$, networks are sensitive to their prior inputs, but more sensitive to their current inputs. To obtain a single memory value per network, we sum each network's memory across all temporal lags. When drawing comparisons between networks trained on tasks with different trial lengths, as in Fig. 5D, we only sum lags up to the shortest trial length.

Analysis

Statistical comparisons. For a given metric, e.g. fitness, we compare each architecture (a distribution across 50 networks) to the fully recurrent architecture (a second distribution across 50 networks), using two-sided, asymptotic Mann–Whitney U tests. We then adjust the resulting (129) p -values, using the Benjamini–Hochberg procedure, to control the false discovery rate at 0.01. For instance, if 100 architectures are found to differ significantly from the fully recurrent architecture after correction, we would expect approximately 1 of these to be a false positive. Practically, we used the `mannwhitneyu` and `false_discovery_control` functions from SciPy⁴⁵.

Random forest models. To quantify the relation between our computational traits and our functional measures, we used random forest regression models. Starting from a binary matrix X with a row per trained network and a column per feature, and a vector y with a value per trained network. We fit random forest models (f), using different subsets of features (X_i): $\hat{y} = f(X_i)$. Practically, we used the `RandomForestRegressor` function from `scikit-learn`⁴⁶ with 10-fold cross-validation. To mitigate overfitting, we limited the depth of each tree to a maximum of 9 levels using the `max_depth` parameter.

Hyperparameters

Task hyperparameters. For our 4 tasks, we used the following hyperparameters.

Model hyperparameters. We used the following hyperparameters for all of our neural network models. Except for the large feedforward model, which used 34 hidden units—to match the fully recurrent architecture's number of trainable parameters. Results from networks with homogeneous activation functions per layer, and smaller and larger networks are shown in (Supplementary Section 3).

Data availability

All trained networks are available for download from <https://doi.org/10.5281/zenodo.20281927>. When loaded, each network is a class object with many attributes (its test results, learned weights, memory metrics etc) which can be analysed. These models can also be re-tested on new trials and tasks.

Code availability

All code was written in Python and is available as a fully documented Pip installable package, at https://github.com/ghoshm/Multimodal_mazes. All results were obtained from a single version of the code (Git tag: `multimodal_mazes-v1`).

Received: 6 February 2026; Accepted: 8 July 2026;

Published online: 07 August 2026

References

- White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*: the mind of a worm. *Philos. Trans. R. Soc. Lond.* **314**, 340 (1986).
- Cook, S. J. et al. Whole-animal connectomes of both *Caenorhabditis elegans* sexes. *Nature* **571**, 63–71 (2019).
- Witvliet, D. et al. Connectomes across development reveal principles of brain maturation. *Nature* **596**, 257–261 (2021).
- Winding, M. et al. The connectome of an insect brain. *Science* **379**, eadd9330 (2023).
- Schlegel, P. et al. Whole-brain annotation and multi-connectome cell typing of *Drosophila*. *Nature* **634**, 139–152 (2024).
- Lin, A. et al. Network statistics of the whole-brain connectome of *Drosophila*. *Nature* **634**, 153–165 (2024).
- Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
- Zingg, B. et al. Neural networks of the mouse neocortex. *Cell* **156**, 1096–1111 (2014).
- Markov, N. T. et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb. Cortex* **24**, 17–36 (2014).
- Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA*. **111**, 8619–8624 (2014).
- Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
- 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).
- Liu, Y. H. et al. How connectivity structure shapes rich and lazy learning in neural circuits. Preprint at <https://doi.org/10.48550/arXiv.2310.08513> (2024).
- Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623 (2018).
- Achterberg, J., Akarca, D., Strouse, D., Duncan, J. & Astle, D. E. Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nat. Mach. Intell.* **5**, 1369–1381 (2023).
- Zhang, X.-J., Moore, J. M., Gao, T.-T., Zhang, X. & Yan, G. Brain-inspired wiring economics for artificial neural networks. *PNAS Nexus* **4**, pgae580 (2025).
- Goulas, A., Damicelli, F. & Hilgetag, C. C. Bio-instantiated recurrent neural networks: Integrating neurobiology-based network topology in artificial networks. *Neural Netw.* **142**, 608–618 (2021).
- Béna, G. & Goodman, D. F. Dynamics of specialization in neural modules under resource constraints. *Nat. Commun.* **16**, 187 (2025).
- Bellec, G., Kappel, D., Maass, W. & Legenstein, R. Deep rewiring: training very sparse deep networks. Preprint at <https://doi.org/10.48550/arXiv.1711.05136> (2017).
- Jayakumar, S., Pascanu, R., Rae, J., Osindero, S. & Elsen, E. Top-kast: top-k always sparse training. *Adv. Neural Inf. Process. Syst.* **33**, 20744–20754 (2020).
- Edelman, G. M. & Gally, J. A. Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. USA*. **98**, 13763–13768 (2001).
- Marder, E. & Taylor, A. L. Multiple models to capture the variability in biological neurons and networks. *Nat. Neurosci.* **14**, 133–138 (2011).
- Jones, P. R. A tutorial on cue combination and signal detection theory: using changes in sensitivity to evaluate how observers integrate sensory information. *J. Math. Psychol.* **73**, 117–139 (2016).

24. Ghosh, M., Béna, G., Bormuth, V. & Goodman, D. F. Nonlinear fusion is optimal for a wide class of multisensory tasks. *PLOS Comput. Biol.* **20**, e1012246 (2024).
25. Jordan, M. I. *Serial Order: A Parallel Distributed Processing Approach*. Technical Report ICS Report 8604, (Institute for Cognitive Science, University of California, 1986).
26. Elman, J. L. Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
27. Mnih, V. et al. Playing Atari with deep reinforcement learning. Preprint at <https://doi.org/10.48550/arXiv.1312.5602> (2013).
28. Jensen, K. T. An introduction to reinforcement learning for neuroscience. *Neurons Behav. Data Anal. Theory.* 1–30 <https://doi.org/10.51628/001c.127771> (2024).
29. Patterson, A., Neumann, S., White, M. & White, A. Empirical design in reinforcement learning. *Journal of Machine Learning Research.* **25**, 1–63 (2024).
30. Aldous, D. J. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM J. Discret. Math.* **3**, 450–465 (1990).
31. Broder, A. Z. Generating random spanning trees. In *Proc. FOCS*, Vol. 89, 442–447 (IEEE Computer Society, 1989).
32. Verma, S., Dickerson, J. & Hines, K. Counterfactual explanations for machine learning: challenges revisited. Preprint at <https://doi.org/10.48550/arXiv.2106.07756> (2021).
33. Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.* **38**, 2770–2824 (2024).
34. Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J. & Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. Preprint at <https://doi.org/10.48550/arXiv.1802.08760> (2018).
35. Jaeger, H. *Short term Memory in Echo State Networks*. Technical report (GMD, 2001).
36. Ganguli, S., Huh, D. & Sompolinsky, H. Memory traces in dynamical systems. *Proc. Natl. Acad. Sci. USA.* **105**, 18970–18975 (2008).
37. Hoffmann, H. Sparse associative memory. *Neural Comput.* **31**, 998–1014 (2019).
38. Castro, P. S. et al. Discovering symbolic cognitive models from human and animal behavior. Preprint at <https://doi.org/10.1101/2025.02.05.636732> (2025).
39. Stanley, K. O. & Miikkulainen, R. Evolving neural networks through augmenting topologies. *Evolut. Comput.* **10**, 99–127 (2002).
40. Stanley, K. O., Clune, J., Lehman, J. & Miikkulainen, R. Designing neural networks through neuroevolution. *Nat. Mach. Intell.* **1**, 24–35 (2019).
41. Gaier, A. & Ha, D. Weight agnostic neural networks. *Adv. Neural Inf. Process. Syst.* **32**, 5364–5378 (2019).
42. Lappalainen, J. K. et al. Connectome-constrained networks predict neural activity across the fly visual system. *Nature* **634**, 1132–1140 (2024).
43. Shiu, P. K. et al. A *Drosophila* computational brain model reveals sensorimotor processing. *Nature* **634**, 210–219 (2024).
44. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Proc. Advances in Neural Information Processing Systems* **32**, 8024–8035 (2019).
45. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
46. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

We thank members of the Neural Reckoning group, Karim G. Habashy and Jeffery S. Bowers, for their constructive input.

Author contributions

M.G. and D.G. conceived these ideas and experiments. M.G. wrote code, analyzed data and produced figures. M.G. and D.G. wrote the manuscript.

Funding

M.G. is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences program.

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/s44488-026-00013-z>.

Correspondence and requests for materials should be addressed to Marcus Ghosh.

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 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) 2026