Can Systems Neuroscientists Understand an Artificial Neural Network?

James M. Shine[1,2], Mike Li[1,2,6], Oluwasanmi Koyejo[3,4], Ben Fulcher[1,5] and Joseph T. Lizier[1,6]

Affiliations

1 Centre for Complex Systems, The University of Sydney, NSW 2006, Australia

2 Brain and Mind Centre, The University of Sydney, NSW 2050, Australia

3 Beckman Institute for Advanced Science and Technology, University of Illinois Champaign

4 Department of Computer Science, University of Illinois at Urbana-Champaign

5 School of Physics, The University of Sydney, NSW 2006, Australia

6 Complex Systems Research Group, Faculty of Engineering, The University of Sydney, NSW 2006, Australia

Corresponding author:

* James M. Shine – mac.shine@sydney.edu.au

Abstract

Network neuroscience has catalysed crucial insights into the systems-level organisation of the brain, however the lack of a 'ground truth' inherently limits direct interpretation. In parallel, deep learning approaches have advanced our algorithmic understanding of intelligence, however the principles that govern learning-induced modifications to network structure remain relatively opaque. Here, we combine the benefits of these two approaches to overcome each of their inherent weaknesses. Specifically, we train a shallow, feedforward neural network to classify handwritten digits and then used a combination of systems neuroscience and information theoretic tools to perform 'virtual brain analytics'[1] on the resultant edge weights and nodal activity patterns. We identified three distinct stages: early in learning, training aligned network edges with information-rich regions of the nodes in up-stream layers of the network, and did so in separate stages for inputs to each layer; whereas later in learning, network activity patterns reconfigured so as to maximize digit category separation in a low-dimensional state space. Our results offer a systems-level perspective of how artificial neural networks function – in terms of multi-stage reorganization of edge weights and activity patterns so as to most effectively exploit the information content of input data during edge-weight training – while simultaneously enriching our understanding of the methods used by systems neuroscience.

In the human brain, capacities such as cognition, attention, and awareness are thought to emerge from the coordinated activity of billions of neurons[2]. Traditional measures typically used to map these functions from neuroimaging data were designed to identify 'activated', localized regions of the brain that characterize a particular cognitive context[3]. This historical focus on localization has led to a number of key insights about neural function, however it has also made it more challenging to create links between systems-level neural organization and psychological capacities.

A potential means for mapping psychological functions to neural circuitry involves the analysis of neuroimaging data from a systems-level perspective[4–6]. By treating neuroimaging datasets as if the data arise from networks of interacting parts, systems neuroscientists are able to characterize high-dimensional datasets in ways that help to understand how brain networks process information[7,8]. Across multiple spatial[9] and temporal[10] scales, these approaches have revealed a number of systems-level principles of brain function. A salient example is the measurement of network modularity, which quantifies the extent to which a network is comprised of a relatively weakly inter-connected set of tight-knit sub-modules.

Systems-level organization has demonstrable computational advantages[11,12] and has been shown to effectively map onto higher-level cognitive functions[13,14]. However, systems-level approaches in human neuroimaging are inherently indirect, as we don't yet have access to *'ground truth'* neuroimaging datasets that directly link structure with function, let alone over the course of interesting behavioural changes, such as the learning of mappings between stimulus and response. For this reason, although we have ready access to high-quality neuroimaging datasets[15–17], it remains relatively challenging to infer precisely which aspects of brain system organization are revealed in neuroimaging data are integral for facilitating behaviour using traditional approaches[18].

Linking adaptation of network structure to enhanced task performance is also a central issue in the field of machine learning. Although some of the details of implementation differ[19], neuroscience and machine learning share some remarkable similarities. For example, the original neural-network algorithms were in part inspired by the anatomy of the cerebral cortex[19–21], and in the case of deep, layered neural networks, both systems share a common property of distributed computation facilitated by complex topological wiring between large numbers of (relatively) simple computational units. Over the last few decades, neural networks[20] have been trained to outperform world experts at complex strategy

games, such as Chess and Go[22]. Although the algorithms that are used to train neural network weights are well understood, the manner in which neural networks reconfigure in order to facilitate high levels of classification accuracy remains relatively opaque[3,20,21]. It is this process of adapting a complex network of interacting components to perform a useful task that has as yet escaped a detailed analysis using the established tools of network neuroscience, which themselves have been used to quantify structure–function relationships in the brain for over a decade.

Whilst the question of how network reconfiguration supports learning is mirrored in machine learning and network neuroscience, the different contexts of these fields provides a timely opportunity to bring them together synergistically to investigate the problem[1]. First, we can observe that the process of adapting a complex network of interacting components to perform a useful task is more simply captured and observed in the training of neural networks. Studying this process in a machine learning setting offers fine time-scale, full-system observations of network structure and activity that are not currently possible in neuroscience. In this way, our approach allows us to potentially identify deeper synergies between the two fields[1,21]. For instance, macroscopic human brain networks constructed from multi-region interactions in neuroimaging data

demonstrate substantial reconfiguration as a function of task performance: early in the course of learning, the brain is relatively inter-connected and integrated, but this pattern typically gives way to a more refined, segregated architecture as a simple motor skill becomes second-nature[6,23]. Whether this same topological change also occurs in ML network remains a critical open question.

In addition, the synthetic nature of ML networks means that we can directly interrogate the functional signature of specific elements within ML networks as they train. While direct access to neuronal interconnections is not permitted in contemporary neuroimaging approaches, we can leverage the nature of neural networks to directly observe changes in the distributed patterns of connectivity inherent to ML neural networks as they change over the course of learning. In this manner, several studies have investigated the extent to which trained neural networks attain a modular structure[24,25], though have not yet looked at the manner in which this develops during the training process, nor using the well-defined measures of modularity derived from systems neuroscience. Using this vantage point, we can test the hypothesis that the functional capacities of neural networks are distributed across the different nodes and connections that define their architecture, which is an idea that is inherently challenging to study in biological brains. Importantly, the established tools of network neuroscience, which have

been used to quantify structure–function relationships in the brain for over a decade, are perfectly placed for such analysis[26–29].

In this study, we use a network neuroscience approach to understand how network reconfiguration supports training within a machine learning setting. This combined approach is used to provide a general understanding of the process of adapting a complex network of interacting components to perform a useful task, which is of paramount theoretical importance to both fields. Specifically, we use the tools of systems neuroscience and information theory to analyse a feedforward neural network as it learns to classify a set of binary digits (the classic MNIST data set). While this approach does not in any way test the boundaries of machine learning performance, it does afford a unique opportunity to better interpret the outcomes of systems-level analytic approaches on how network reconfiguration supports learning.

By tracking the topology of the network over the course of training, we identify three distinct phases of topological reconfiguration. Early in learning, training reconfigured the edges of the network so that they are strongly aligned with information-rich regions of the nodes in up-stream layers of the network, and did so in separate stages for inputs to each layer. Later in learning, network activity

patterns reconfigured so as to maximize digit category separation in a low-dimensional state space. These results provide important insights into how network reconfiguration supports learning in feed-forward neural networks, contributing to the cause of "explainable AI"[19–21]. This simpler setting also enriches our understanding of these methods themselves and aids interpretation of their results in a neuroscience setting. Through this approach, we hope to provide a clear interpretation of network activity over the course of learning that simultaneously informs our understanding of both systems neuroscience and machine learning.

Results

Feed Forward Neural Network Construction and Training

We applied systems neuroscience and information theoretic methods to analyze the structure of a feedforward neural network as it was trained (across 100,000 epochs with stochastic gradient descent) to rapidly classify a set of ten hand-written digits (Modified National Institute of Standards and Technology [MNIST] dataset[30]). Although a neural network with a single hidden layer is theoretically sufficient for high performance on MNIST[30], neural networks with more hidden layers provide benefits of both computational and parameter efficiency[31]. For the

sake of simplicity, we chose a relatively basic network in which edge weights and nodal activity patterns could be directly related to performance.

With these constraints in mind, we constructed a feedforward network with two hidden layers — a 100-node hidden layer (HL1) that received the 28 x 28 input (pixel intensities from the MNIST dataset) and a 100-node hidden layer (HL2) that received input from HL1 — and a 10-node output layer (Fig. 1A). The edges between these layers were given unique labels: the edges connecting the input nodes to the first hidden layer were labelled as α edges (dark blue in Fig. 1A); the edges connecting the two hidden layers were labelled as β edges (orange in Fig. 1A); and the edges connecting the second hidden layer to the readout layer were labelled as γ edges (dark green in Fig. 1A). The absolute value of edge weights from all three groups increased non-linearly over the course of training.

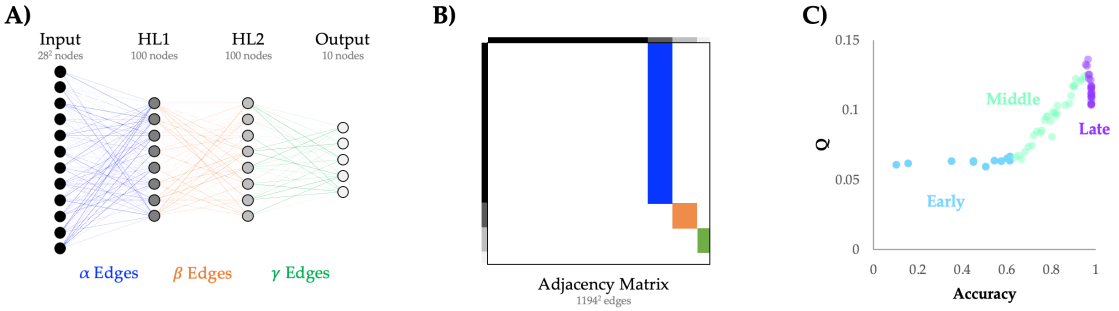

Figure 1. A feed-forward neural network exhibits three topologically distinct phases of reconfiguration throughout learning the MNIST dataset. A) A large (60,000 item) corpus of hand-drawn digits (28 x 28 pixel array with 256 intensity values per pixel) were vectorized and entered

into a generic feed-forward neural network with two hidden layers – a 100-node layer (HL1) that received the 28 x 28 input and a 100-node layer (HL2) that received the input from HL1 – and a 10-node output layer (argmax); B) the edges connecting the input → HL1 (dark blue; α), HL1 → HL2 (orange; β) and HL2 → output (dark green; γ) were embedded within an asymmetric weighted and signed connectivity matrix; C) classification accuracy showed a non-linear relationship with Q (calculated across the whole network): there is an initial learning phase that was independent of network modularity (light blue), after which there is a positive linear relationship between accuracy and Q (Pearson's r = 0.981; light green), and finally a sustained drop in Q, as accuracy saturates in the later stages of learning (light purple).

The Topological Signature of Feed Forward Neural Network During Training

Inspired by results from systems neuroscience[6,23] and complex systems[27,29] linking network topology and function, we hypothesized that the topological structure of the neural network should reconfigure so as to maximally extract the relevant information from the input dataset, and that this reconfiguration should relate to the improved performance of the network across the training phrase. To test this prediction, we needed a means for translating the edges of the neural network into a format that was amenable to network science approaches (i.e., a weighted and directed adjacency matrix). To achieve this aim, we created a sparse node x node matrix, and then mapped the α (Input-HL1), β (HL1-HL2) and γ (HL2-output) edges accordingly, yielding the adjacency matrix shown in Fig. 1B.

With the network edge weights framed as a graph, we were next able to apply topological analyses from the systems neuroscience literature to the edge weights of the feed forward neural network as it was trained to classify the MNIST dataset. In particular, we were interested in whether the topology of the neural network over the course of learning mirrored patterns observed in the analysis of fMRI networks in human participants[12]. By tracking functional networks derived from fMRI data over the course of 10 sessions in which participants learned to map visual stimuli to motor responses, it was observed that effective learning was associated with an increase in network modularity[12], Q, which quantifies the extent with which the network can be clustered into tight-knit communities with relatively sparse connections between them and is thought to be a key property of complex networks[12]. From this work, we hypothesized that the neural network should show a similar shift towards heightened modularity over the course of learning the MNIST dataset.

To test this hypothesis, we applied used the Louvain algorithm to estimate Q from the neural network graphs at each training epoch. Our results provided partially supportive evidence for our hypothesis (Fig. 1C), allowing us to confirm our hypothesis in the intermediate stages of learning, but reject the hypothesis for early or late stages. Interestingly, we also observed a nonlinear relationship

between Q and classification accuracy (Fig. 1C). From the shape of this curve, we identified three phases of topological adjustment through the training process (labelled as 'Early', 'Middle', and 'Late' in Fig. 1C). Early in the course of training, there was a substantial improvement in accuracy without a noticeable change in Q (light blue in Fig. 1C). In the Middle phase, we observed an abrupt increase in Q (light green in Fig. 1C) that tracked linearly with performance accuracy (r = 0.981, $p_{PERM} < 0.0001$, permutation test). Finally, the level of Q began to drop in the Late training stage (Fig. 1C; light purple). These results demonstrate that the modularity of the neural network varies over the course of training in a way that is tightly associated with the classification performance of the network.

Edge weight alterations are concentrated on informative inputs

The fact that Q didn't change early in training, despite substantial improvements in accuracy, was somewhat surprising. This result was made even more compelling by that that that we observed substantial edge-weight alteration during the initial phase (Fig. 2A), however with no alteration in the overall topology. To better understand this effect, we calculated the variance of changes in edge strength across all outgoing edges from input pixels (σ Edge Δ) in the Input→HL1 sub-network (α edge; blue in Fig. 1A/B) over the course of the Early phase. We found that the α edge weights that changed the most over this phase

were located along the main stroke lines in the middle of the image (e.g., the outside circle and a diagonal line; Fig. 2A). Similar to the manner in which an eye saccades to a salient target[32], we hypothesized that the feedforward network was reconfiguring early in training so as to align with the most informative regions of the input space.

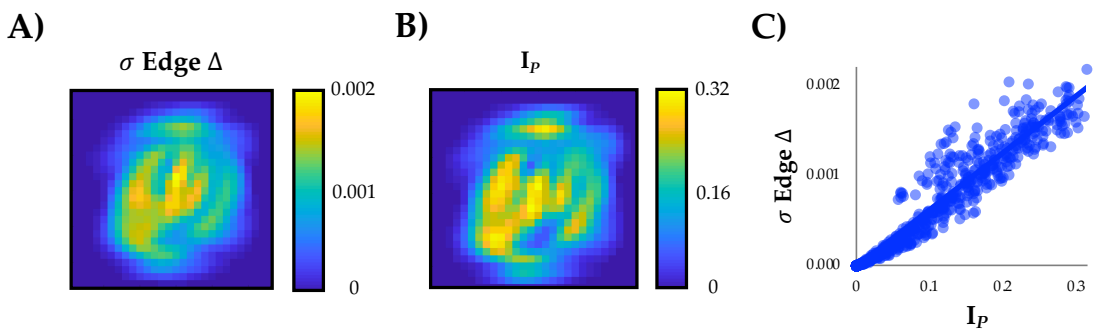

Figure 2. Topologically silent alterations in network edges during the Early phase of training. A) although network modularity was static in the Early phase, the standard deviation of changes in edge strength, σ Edge Δ, in the first hidden layer of the network did change substantially over the course of the Early training phase (first 10 epochs; cf. Fig. 1C); B) Pixel information, I_P = MI(pixel,class); C) We observed a strong positive correlation between σ Edge Δ and I_P: $r = 0.965$.

To test this hypothesis, we binarized the pixel activity across the 60,000 items from the training set, with a threshold that varied across each pixel so as to maximize the mutual information (MI) that the binarized pixel provides about the class (i.e., the digit), and then calculated the information held by each pixel (I_P:

MI(pixel,class); Fig. 2B). We observed a clear, linear correspondence between I_P and the edges that reconfigured the most during the Early phase (Fig. 2C; $r = 0.965$, $p_{PERM} < 0.0001$). This result supported our hypothesis that the network was adjusting to concentrate sensitivity to class-discriminative areas of input space, which we demonstrate occurs via the reconfiguration of edge weights relating to the most class-discriminative areas of the input space.

Topological Segregation During the Middle Phase of Learning

Outside of the initial phase of learning, we observed a substantial increase in network Q that scaled linearly with improvements in classification accuracy (Middle Phase II; Fig. 1C, green). To better understand how node-level network elements reconfigured during the Middle phase, we computed two metrics for each node that quantify how its connections are distributed across network modules: (i) module-degree z-score (MZ); and (ii) participation coefficient (PC)[33]. MZ and PC have together been used characterize the cartographic profile of complex networks: MZ measures within-module connectivity, and PC measures between-module connectivity and thus captures the amount of inter-regional integration within the network (see Methods for details; Fig. 2A)[33]. These measures have been previously used in combination with whole-brain human fMRI data to demonstrate a relationship between heightened network integration and cognitive

function[14,34], however the algorithmic utility of integrative topological organization is less well understood. Importantly, the calculation of both MZ and PC relies on the community assignment estimated from the Louvain algorithm, and hence affords a sensitivity to changes in network topology over the course of training.

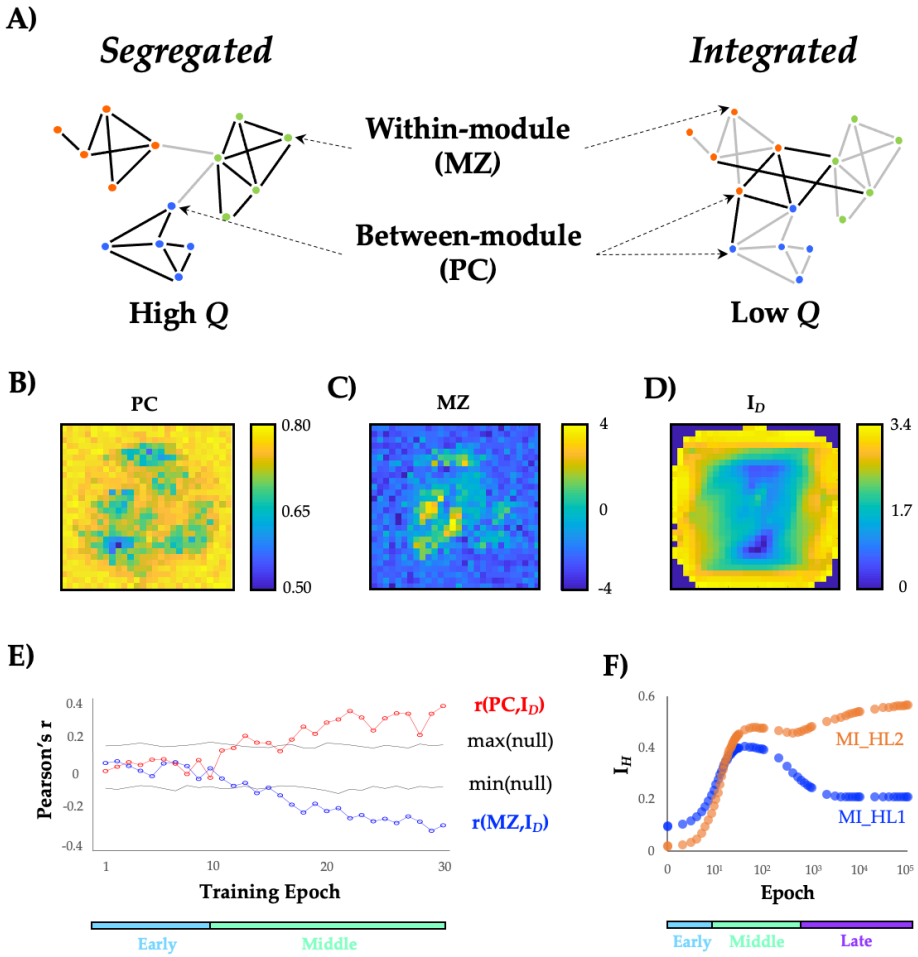

Figure 3 – Topological changes in the Middle epoch. A) a cartoon depiction of two topological extremes: on the left is a segregated network, with tight-knit communities that are weakly-interconnected – this network would be characterized by high *Q*, and would have more nodes with

high module degree *z*-score (MZ) than nodes with high Participation Coefficient (PC); on the right is an integrated network, which has stronger connections between nodes in different communities, and hence a lower *Q*, and more nodes with high PC than nodes with high MZ; B) participation coefficient (PC) of Input layer nodes at training epoch 30; C) module degree *z*-score (MZ) of Input layer at training epoch 30; D) Digit information, I_D = MI($pixel_{On}$,class); E) Pearson's correlation, *r*, between I_D and PC (red) and MZ (blue) across first 30 training epochs. Black lines represent the upper and lower bounds (95th and 5th percentiles) of a permuted null dataset (10,000 iterations) and coloured bars represent learning phases; F) I_H = MI(node,class) for HL1 (blue) and HL2 (orange) nodes – note that both subnetworks increase I_H during the Middle phase, but that the Late phase dissociates the two layers.

Using this cartographic approach[33], we were able to translate the edge weights in the network into values of PC and MZ for each node of the network for each epoch of training. Figures 2B and 2C show the distribution of PC and MZ values for the nodes in the Input layer at training epoch 30, which was indicative of the patterns in the Middle phase. We first noted that PC and MZ mapped onto different locations in the input space, and hence were somewhat inversely correlated across all epochs ($r = -0.107$; $p = 3 \times 10^{-89}$). PC was associated with a relatively 'patchy' appearance around the main stroke areas, suggestive of a distributed topological coverage of the input space, as well as high values on the edges of the input space (Fig. 2C). In contrast, high MZ values (which are indicative of local hubs within network communities) were located along the main stroke lines, so as to align

network hubs with detection or absence of higher order patterns in the input stream (Fig. 2D). We hypothesized that these changes were indicative of a topological reconfiguration that reorganized the neural network so as to align network hubs with key aspects of the input stream.

To test this hypothesis, we related the PC and MZ for each node of the network across all epochs of training to the amount of information available in each pixel of the input space (Fig. 3D). In order to more precisely relate changes in MZ and PC to the manner in which each pixel held information about the digit class, we calculated the partial information (i.e., I_D: MI($pixel_{On}$,class); as well as its inverse, MI($pixel_{Off}$,class)) carried by each pixel about the class. In contrast to I_P, I_D for example quantifies how informative each pixel is for tracking multiple different digit classes, but only when the pixel is active ($pixel_{On}$). High values of I_D imply that the recruitment of the particular pixel is associated with a reduction in uncertainty (i.e., an increase in information) about the digit. As detailed in the Methods, I_P is inversely correlated to I_D and dominated by samples when the pixel is inactive; hence we focus on information carried when the pixel is active I_D (Fig. 3D).

When relating I_D to PC and MZ across training, we observed a significant positive correlation between I_D and MZ that emerged towards the end of the Middle phase (Fig. 2E). Specifically, we observed a double dissociation that emerged over the course of learning in the input layer (Fig. 2F), wherein during the Middle phase I_D was positively correlated with nodal participation coefficient (max $r = 0.396$, $p_{PERM} < 0.0001$), but negatively correlated with mean module degree z-score (max $r = -0.352$, $p_{PERM} < 0.0001$). In other words, the topology of the neural network reconfigured so as to align highly informative active pixels with topologically integrated hubs (nodes with higher PC). These pixels are comparatively rarely active, but highly informative when this occurs, and the result suggests that this requires the network to send information about such events to many downstream modules. By contrast, more segregated hubs (nodes with higher MZ) were likely to be associated with highly informative inactive pixels, which are also more informative on average of digit class. This may indicate that the network is reconfiguring so as to organize sets of informative nodes into modules in a way that supports the creation of higher order 'features' in the next layer. In neuroscience, nodes within the same module are typically presumed to process similar patterns of information[12], suggesting that the topology of the neural network studied here may be aligning to detect the presence or absence of low-dimensional features within the input space.

Inter-layer correspondence

Given that the same gradient descent algorithm used to train the network was applied consistently across all layers of the network, we predicted that the same principles identified in the input layer should propagate through the network, albeit to the abstracted 'features' captures by each previous layer. Similar to the manner in which a locksmith sequentially opens a bank vault, we hypothesized that each layer of the neural network should align with the most informative dimensions of its input in turn, such that the information could only be extracted from an insulated layer once a more superficial layer was appropriately aligned with the most informative aspects of its input stream. To test this hypothesis, we investigated how the mutual information MI(node,class) in each node about the digit class evolved across training epochs. As shown in Fig. 2F, mean MI within both hidden layers 1 (MI_{HL1}) and 2 (MI_{HL2}) increased during the first two epochs, but then diverged at the point in learning coinciding with the global decrease in modularity, Q (cf. Fig. 1D). Crucially, despite the decrease in MI_{HL1} there was still an increase in MI_{HL2}, suggesting that the Layer 2 nodes are improving their ability to combine information available in separate individual Layer 1 nodes to become

more informative about the class. This suggests that Layer 1 nodes specialise (and therefore hold less information overall, lower MI_{HL1}) in order to support the integration of information in deeper layers of the neural network (increased MI_{HL2}).

Validation with the eMNIST dataset

Thus far, we have demonstrated that neural network edges reconfigured in three distinct phases to augment the relationship between input and hidden layer nodes and the class-relevant information in the data being fed into the network. To determine whether these training principles identified on the MNIST dataset were generalizable to other datasets, we trained a new feed-forward neural network (identical in architecture to the original network) on the eMNIST dataset[52], which is similar to MNIST, but uses hand-written letters, as opposed to numbers. Although learning was more protracted in the eMNIST dataset (likely due to the increased complexity of the alphabet, relative to the set of digits), we were able to replicate the changes in network structure across training the MNIST dataset: (i) the network shifted from integration to segregation; layers reconfigured in serial; and nodal roles (with respect to inferred network modules) were similarly related to class-relevant information in individual pixels (Fig. S1). These results suggest that the insights obtained from the MNIST analysis may represent general learning principles of multilayered neural networks.

Late phases were associated with low-dimensional pattern separation

Next, we investigated whether the manner in which the extent to which the nodal topology of the networks trained on the two datasets differed (i.e., whether different regions of the input space had higher or lower PC and MZ) was proportional to the most informative locations of the inputs space in each dataset (ΔI_D). Specifically, the difference in the pattern of a node's edges across inferred network modules between the eMNIST and MNIST datasets (ΔPC) was correlated with the difference in image input characteristics between the two datasets (ΔI_D vs. ΔPC: $r = 0.301$, $p_{PERM} < 0.0001$; ΔI_D vs. ΔMZ: $r = -0.247$, $p_{PERM} < 0.0001$). This result provides further confirmation that neural networks learn by reorganizing their nodal topology into a set of phases that act so as to align network edges and activity patterns with the most informative pixels within the training set.

We found that pixels demonstrated unique roles across learning with respect to the emerging modular architecture of the training neural network, and that these roles were shaped by their class-relevant information. As the edge weights were reconfigured across training, we observed that outgoing edge strength increases for highly informative inputs. As these weights change, they alter the activity of each of the nodes in the hidden layers, which ultimately pool their activity via

modules to affect the class predictions, which are read out based on the activity of the final output layer. So how do the changes in edge weight translate into nodal activity? Based on recent empirical electrophysiological[35] and fMRI[36] studies, we hypothesized that the activity patterns would be distributed across the neural network in a low-dimensional fashion. Specifically, by way of analogy to the notion of manifold untangling in the ventral visual system[37], we predicted that across training, the high-dimensional initial state of the system (i.e., the random weights) would become more low-dimensional as pixel-pixel redundancies were discovered through the learning process.

To test this hypothesis, we used dimensionality-reduction[38] to analyze the 'activity' of all of the nodes within the neural network, across the different training epochs. The primary intuition behind these approaches is that, given the highly inter-connected nature of biological systems, reducing the dimensionality of data using statistical techniques, such as principal component analysis (PCA), can allow investigators to essentially ignore small details in order to track the representative activity of the system over time, often revealing key features of its organization[39]. We applied PCA to the nodal activity across all four layers of the feedforward network – i.e., the Input, HL1, HL2 and Output nodes – which were first standardized and then either concatenated (to calculate the dimensionality of

the entire process) or analyzed on an epoch-to-epoch basis (to calculate the effect of training; see Methods for details). The concatenated state-space embedding was low-dimensional (120/994 components, or 12.2%, explained ~80% of the variance) and the pixel-wise loading of each of the top eigenvalues (λs) for the Input layer (Fig. 4A) was correlated with both information theoretic measures used in the prior analyses ($I_P - \lambda_1$: $r = 0.218$, $p < 0.0001$; λ_2: $r = 0.189$, $p < 0.0001$; $\lambda_3 = 0.158$, $p < 0.0001$; and $I_D - \lambda_1$: $r = 0.338$, $p < 0.0001$; λ_2: $r = 0.123$, $p < 0.0001$; λ_3: $r = 0.062$, $p = 0.080$), suggesting a direct correspondence between class-relevant information in the input space and the low-dimensional embedding. Crucially, test trials that were incorrectly classified (Epoch 10,000, though results were consistent for other epochs) were associated with lower absolute loadings on the ten most explanatory EVs ($EV_{1\text{-}10}$; Figure S3; FDR $p < 0.05$). This suggests that alignment with the low-dimensional space was related to accurate performance. These results are tangentially related to recent empirical neuroscientific studies that employed dimensionality reduction on electrophysiological[35] and fMRI data[36] to show that learning and cognitive task performance are typically more effective when constrained to a low-dimensional embedding space.

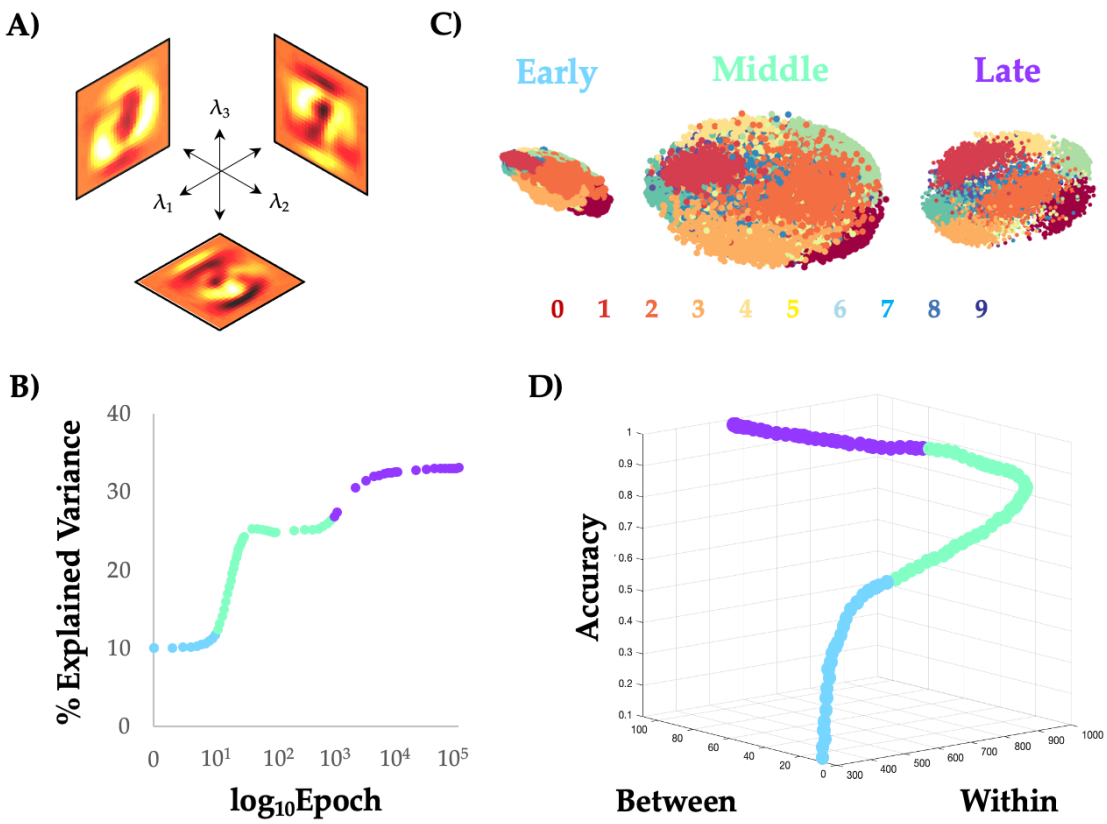

Figure 4 – Unravelling the manifold: low-dimensional projections of feed-forward neural network activity during MNIST training reveal category-specific untangling. A) The first three principal components (eigenvalues 1-3: λ_1/ λ_2/ λ_3) of the Input nodes; B) The percentage of variance explained by EV_1, when the PCA was fit on data from each training epoch separately; C) 3D scatter plot of the items from the training set during three different stages: during the Early phase (Epochs 1–10), the topological embedding of the different digits showed substantial overlap, which is reflected in the low between-category distance (i.e., distance between mean of each digit); in the Middle phase (Epochs 11–300), the embedding showed a relative expansion in the low-dimensional space; and during the Late phase (Epochs 300+), the distance within each category dropped dramatically; D) 3D scatter plot of between-category and within-category distance, along with training accuracy – note that maximal accuracy is associated with increases in both within- and between-category distance.

By conducting a PCA on each epoch in turn, we found that training was associated with a nonlinear alteration in network dimensionality that aligned with the topologically identified phases (Fig. 4B and Movie S1). The network began in a relatively high-dimensional configuration, consistent with the random initiation of nodal activity. During the Early phase (light blue in Fig. 4B), as the edge weights reconfigured to align with I_P (Fig. 3D), the dimensionality of the system remained relatively high. During the Middle phase (light green in Fig. 3B), there was a sharp reduction in dimensionality, however the dimensionality collapse was diminished mid-way through the phase. The Late phase (purple in Fig. 3B) was associated with a final, albeit mild, reduction in dimensionality. Interestingly, heightened levels of training led to a tighter correspondence between nodal topological signatures (PC/MZ, calculated at each epoch) and the principal component loadings of nodes in the Input layer (Fig. S2), suggesting that the topology of the neural network reconfigured over training to better map onto the low-dimensional manifold that concentrates class-relevant information in the training dataset.

Previous theoretical work in systems neuroscience has argued that a primary computational benefit of the nervous system may be to 'untangle' low-dimensional relationships between different information streams[37]. Interestingly,

the same concept has been used to explain the function of both the visual system[37] and effective decision making[40], suggesting that the capacity may reflect a relatively domain general property of nervous system organization. There is also evidence to suggest that the same concept may underpin the functionality of convolutional neural networks trained on naturalistic images[26]. Based on these results, we interrogated our data for the presence of manifold untangling, and found that the increase in topologically rich, low-dimensionality was associated with a relative 'untangling' of the low-dimensional manifold (Fig. 4C): the Middle phase was associated with a general expansion in the low-dimensional embedding distance Within categories (light green in Fig. 4D), which then allowed the system to both expand Between categories and contract within Categories during the Late phases of learning (purple in Fig. 4D). This ultimately had the effect of augmenting classification accuracy. Indeed, the contraction of the within category embedding distance – which takes place first – co-occurs with the drop of MI_{HL1}, with the following expansion of Between category distance co-occurring with the increase in MI_{HL2}. At the sub-network level, the activity on nodes in HL2 was substantially more low-dimensional than HL1 (Fig. S4), further expanding on the notion that different computational constraints are imposed on neural networks, depending on the depth of network layers.

Discussion

This work demonstrates the benefits of translating neural networks into the language of systems neuroscience. We applied topological metrics to the weighted, directed adjacency matrix of a simple feedforward neural network during its training to quantify how connections are distributed across the network, and thus directly tracked changes in network edge strength (Fig. 2), topology (Fig. 3) and activity (Fig. 4) that coincided with distinct phases of accuracy-related change in the configuration of the network. The results of our study both help to validate the study of network topology in systems neuroscience, while also improving our understanding of how neural networks alter their structure so as to better align the topology of the network with the available streams of information being fed to the network.

The approaches used in this study were designed to aid in the interpretation of the inner workings of the "black box" of the brain, which is a similar issue to that recognised today in seeking explanations for the inner workings of deep neural networks[21]. An important open question is whether other distinct network architectures, such as recurrent[41], convolutional[42], echo state[43] or generative adversarial networks[44], will share similar or distinct mechanisms[3,20,21]. It is also well known that the brain contains numerous non-linear mechanisms, including gain

modulation[45] and circuit-based mechanisms[46], that may in turn form the basis of distinct topological motifs. Regardless, our work provides evidence that network science can provide intuitive explanations for the computational benefits of neural network learning, and helps to highlight the growing intersection between artificial and biological network analyses[47]. Taking inspiration from biology, other authors have previously suggested that the concept of modularity may be explicitly employed to improve the design of deep neural network architecture in various ways[48,49]. Our findings add to this perspective by demonstrating that the training of such networks may be implicitly interpretable using these same concepts.

What can network neuroscience learn from the results of this experiment? For one, our observations provide evidence that the tools of systems neuroscience and engineering can indeed be used to understand the function of a complex, high-dimensional system[5,38,39]. In addition, there are a number of benefits to analysing neural networks that are not readily apparent in neurobiological analyses. For instance, in the case of feed-forward neural networks, we know the direct mapping between inputs and nodes, whereas in neuroscience, this mapping is challenging (e.g., the location of the animal's gaze can alter the information entering the brain). In addition, standard network approaches in neuroscience require a noisy

estimation of network edges, whereas the process here allows us to observe the edge weights directly. It is our hope that these benefits can be used to improve our understanding of the computational benefits of particular systems-level organizing principles, which in turn can be brought back to neuroscience to accelerate progress in our understanding of the systems-level mechanisms that comprise effective neurological function.

In conclusion, we used a systems-level perspective to demonstrate a series of three serial phases over the course of network training that relate to manner in which the topological configuration of the neural network is aligned with the information content provided by the data fed into the network.

Methods

Feedforward Neural Network

A prototypical, four-layer, feed-forward neural network with no non-linearities was created with randomized weights (edge strengths: -1 to 1). The input layer was designed to take 784 inputs, which themselves were flattened from a 28x28 greyscale pixel array from the MNIST dataset[50]. The input layer was fully connected to a hidden layer of 100 nodes (HL1), which in turn was fully connected to a second hidden layer of 100 nodes (HL2). The second hidden layer was then fully connected to a 10-node output layer. The activation function was a standard sigmoid for the incoming connections at each hidden layer (exponent = 1), and a soft max at the output layer. The maximum value across the nodes of the output layer was taken to reflect the 'response' of the network. Each result in our study was also replicated in a separate eMNIST dataset, which was identical to MNIST, but had 26 hand-written letters, as opposed to 10 hand-written digits[51].

Training Approach

The network was trained with backpropagation using a Stochastic Gradient Descent optimiser. To aid interpretation, the learnt bias at each neuron was kept to zero and no regularisation was used. The weights and activities were saved as

training progressed over the course of a number of epochs (SGD: 100,000). Accuracy was defined as the percentage of trials in a held-out, 10,000 trial testing set in which the maximum value of the output layer was matched with the test category.

Network Construction

The weighted and signed edges from each asymmetric layer of the neural network were concatenated together to create an asymmetric connectivity matrix. Each connectivity profile was placed in the upper triangle of the matrix (see Fig. 2). To ensure that this step did not adversely affect the topological estimates, each experiment was conducted separately on: a) each layer in turn; b) only the upper triangle of the connectivity matrix. Similar patterns were observed when we re-ran each network separately, suggesting that the embedding did not adversely affect topological interpretation.

Modularity Maximization

The Louvain modularity algorithm from the Brain Connectivity Toolbox (BCT[73]) was used on the neural network edge weights to estimate community structure. The Louvain algorithm iteratively maximizes the modularity statistic, Q, for different community assignments until the maximum possible score of Q has been

obtained (see Equation 1). The modularity of a given network is therefore a quantification of the extent to which the network may be subdivided into communities with stronger within-module than between-module connections.

$$Q_T = \frac{1}{v^+}\sum_{ij}\left(w_{ij}^+ - e_{ij}^+\right)\delta_{M_iM_j} - \frac{1}{v^+ + v^-}\sum_{ij}\left(w_{ij}^- - e_{ij}^-\right)\delta_{M_iM_j} \qquad [1]$$

where v is the total weight of the network (sum of all negative and positive connections), w_{ij} is the weighted and signed connection between nodes i and j, e_{ij} is the strength of a connection divided by the total weight of the network, and δ_{MiMj} is set to 1 when nodes are in the same community and 0 otherwise. '+' and '–' superscripts denote all positive and negative connections, respectively.

For each epoch, we assessed the community assignment for each region 500 times and a consensus partition was identified using a fine-tuning algorithm from the BCT. We calculated all graph theoretical measures on un-thresholded, weighted and signed undirected, asymmetric connectivity matrices[73]. The stability of the γ parameter (which defines the resolution of the community detection algorithm) was estimated by iteratively calculating the modularity across a range of γ values (0.5-2.5; mean Pearson's r = 0.859 +- 0.01) on the time-averaged connectivity matrix for each subject – across iterations and subjects, a γ value of 1.0 was found to be

the least variable, and hence was used for the resultant topological analyses. A consensus clustering partition was defined across all epochs using *consensus_und.m* from the BCT. The resultant solution contained 10 clusters that each contained nodes that were distributed across multiple layers (i.e., Input, HL_1, HL_2 and Output).

Cartographic Profiling

Based on time-resolved community assignments, we estimated within-module connectivity by calculating the time-resolved module-degree Z-score (MZ; within module strength) for each region in our analysis (Equation 2)[74], where κ_{iT} is the strength of the connections of node *i* to other nodes in its module s_i at time *T*, $\bar{\kappa}_{s_{iT}}$ is the average of κ over all the regions in s_i at time *T*, and $\sigma_{\kappa_{s_{iT}}}$ is the standard deviation of κ in s_i at time *T*.

$$\mathrm{MZ} = \frac{\kappa_{iT} - \bar{\kappa}_{s_{iT}}}{\sigma_{\kappa_{s_{iT}}}} \qquad [2]$$

The participation coefficient, *PC*, quantifies the extent to which a node connects across all modules (i.e. between-module strength) and has previously been used to successfully characterize hubs within brain networks (e.g. see [75]). The PC for each node was calculated within each temporal window using Equation 3, where

κ_{isT} is the strength of the positive connections of node *i* to nodes in module *s* at time *T*, and κ_{iT} is the sum of strengths of all positive connections of nodes *i* at time *T*. Consistent with previous approaches in neuroscience[14,52], negative connections were removed prior to calculation. The participation coefficient of a region is therefore close to 1 if its connections are uniformly distributed among all the modules and 0 if all of its links are within its own module.

$$\mathrm{PC} = 1 - \sum_{s=1}^{n_M} \left(\frac{\kappa_{isT}}{\kappa_{iT}}\right)^2 \qquad [3]$$

Mutual Information

We calculated three separate Information measures. To calculate the Information content within each pixel (I_P), we binarized the pixel activity across the 60,000 items from the training set, with a threshold that varied across each pixel so as to maximize the mutual information (MI) that the binarized pixel provides about the class, and then calculated the information within each pixel: MI(pixel,class). To calculate the Information content within each pixel when the pixel was active (after thresholding), we averaged the pointwise MI for each training item, $I_D = \log_2 \frac{p(\text{digit}|\text{pixel})}{p(\text{digit})}$, only over the items where the pixel was on (pixel_{On}). Note that I_P and I_D were inversely correlated across the 28x28 input dimension (r = -0.560, p_{PERM}

< 0.0001), suggesting that the total information from the pixel is dominated by samples when the pixel is inactive. To calculate the Information content within each hidden layer node (I_H), we calculated the mutual information for each node (binarized at activity = 0.5) with the digit class. All MI values were computed using the open source JIDT software[53].

Principal Components Analysis

Activity values from the test trials from the input, HL1 and HL2 layers from each epoch were concatenated to form a multi-epoch time series. The data were normalized and then a spatial PCA was performed on the resultant data[38]. The top 3 eigenvectors were used to track the data within a low-dimensional embedding space (Fig. 3), and the percentage explained variance was tracked across all learning epochs. The eigenvectors from the concatenated data were then used to estimate the leading eigenvalues across all training epochs. The analysis was also re-run with activity patterns in HL1 and HL2 separately (i.e., independent of the input layer; Fig. S4). The average value for each exemplar was then used to create two distance measures: Between-category distance, which was defined as the average between-category Euclidean distance at each epoch; and Within-category distance, which was defined as the average within-category Euclidean distance within each epoch.

Permutation Testing

We used non-parametric testing to determine statistical significance of the relationships identified across our study[54]. A distribution of 10,000 Pearson's correlations was calculated for each comparison, against which the original correlation was compared. Using this approach, the *p*-value was calculated as the proportion of the null distribution that was less extreme than the original correlation value. In many instances, the effects we observed were more extreme than the null distribution, in which case the *p*-value was designated as $p_{PERM} < 0.0001$.

References

1. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **95**, 245–258 (2017).
2. Shine, J. M. The thalamus integrates the macrosystems of the brain to facilitate complex, adaptive brain network dynamics. *Progress in Neurobiology* 101951 (2020) doi:10.1016/j.pneurobio.2020.101951.
3. Hasson, U., Nastase, S. A. & Goldstein, A. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron* **105**, 416–434 (2020).
4. Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12574–12579 (2016).
5. Shine, J. M. *et al.* Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nat Neurosci* **22**, 289–296 (2019).
6. Bassett, D. S., Yang, M., Wymbs, N. F. & Grafton, S. T. Learning-induced autonomy of sensorimotor systems. *Nature Neuroscience* **18**, 744–751 (2015).
7. Wibral, M., Lizier, J. T. & Priesemann, V. Bits from brains for biologically inspired computing. *Frontiers in Robotics and AI* **2**, 5 (2015).
8. Hamrick, J. & Mohamed, S. Levels of Analysis for Machine Learning. *arXiv:2004.05107 [cs, stat]* (2020).
9. Favre-Bulle, I. A., Vanwalleghem, G., Taylor, M. A., Rubinsztein-Dunlop, H. & Scott, E. K. Cellular-Resolution Imaging of Vestibular Processing across the Larval Zebrafish Brain. *Current biology : CB* **28**, 3711-3722.e3 (2018).

10. Kitzbichler, M. G., Smith, M. L., Christensen, S. R. & Bullmore, E. Broadband Criticality of Human Brain Network Synchronization. *PLoS Comput Biol* **5**, e1000314 (2009).
11. Ellefsen, K. O., Mouret, J.-B. & Clune, J. Neural Modularity Helps Organisms Evolve to Learn New Skills without Forgetting Old Skills. *PLoS Comput Biol* **11**, e1004128 (2015).
12. Sporns, O. & Betzel, R. F. Modular Brain Networks. *Annual review of psychology* **67**, annurev-psych-122414-033634 (2015).
13. Betzel, R. F., Fukushima, M., He, Y., Zuo, X. N. & Sporns, O. Dynamic fluctuations coincide with periods of high and low modularity in resting-state functional brain networks. *NeuroImage* **127**, 287–297 (2016).
14. Shine, J. M. *et al.* The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance. *Neuron* **92**, 544–554 (2016).
15. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
16. Markram, H. *et al.* Reconstruction and Simulation of Neocortical Microcircuitry. *Cell* **163**, 456–492 (2015).
17. Phillips, J. W. *et al.* A repeated molecular architecture across thalamic pathways. *Nature Neuroscience* **22**, 1925–1935 (2019).
18. Jonas, E. & Kording, K. P. Could a Neuroscientist Understand a Microprocessor? *PLoS Comput Biol* **13**, e1005268 (2017).

19. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat Rev Neurosci* (2020) doi:10.1038/s41583-020-0277-3.
20. Sejnowski, T. J. The unreasonable effectiveness of deep learning in artificial intelligence. *Proc Natl Acad Sci USA* 201907373 (2020) doi:10.1073/pnas.1907373117.
21. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat Neurosci* **22**, 1761–1770 (2019).
22. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
23. Mohr, H. *et al.* Integration and segregation of large-scale brain networks during short-term task automatization. *Nature communications* **7**, 13217 (2016).
24. Csordás, R., van Steenkiste, S. & Schmidhuber, J. Are Neural Nets Modular? Inspecting Functional Modularity Through Differentiable Weight Masks. *arXiv:2010.02066 [cs]* (2021).
25. Ballard DH. Modular learning in neural networks. *AAAI-87 Proceedings* **1**, 279–284 (1987).
26. Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat Commun* **11**, 746 (2020).
27. Shwartz-Ziv, R. & Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv:1703.00810 [cs]* (2017).

28. Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proc Natl Acad Sci USA* **115**, E10313–E10322 (2018).
29. Sussillo, D. Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology* **25**, 156–163 (2014).
30. Leshno, M., Lin, V. Ya., Pinkus, A. & Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks* **6**, 861–867 (1993).
31. Mhaskar, H., Liao, Q. & Poggio, T. Learning Functions: When Is Deep Better Than Shallow. *arXiv:1603.00988 [cs]* (2016).
32. Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. Perceptions as Hypotheses: Saccades as Experiments. *Front. Psychology* **3**, (2012).
33. Guimerà, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
34. Bertolero, M. A., Yeo, B. T. T. & D'Esposito, M. The modular and integrative functional architecture of the human brain. *Proceedings of the National Academy of Sciences of the United States of America* 201510619 (2015).
35. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).
36. Shine, J. M. *et al.* The Low-Dimensional Neural Architecture of Cognitive Complexity Is Related to Activity in Medial Thalamic Nuclei. *Neuron* **104**, 849-855.e3 (2019).
37. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends in Cognitive Sciences* **11**, 333–341 (2007).

38. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* **17**, 1500–1509 (2014).

39. Kato, S. *et al.* Global brain dynamics embed the motor command sequence of Caenorhabditis elegans. *Cell* **163**, 656–669 (2015).

40. Yoo, S. B. M. & Hayden, B. Y. Economic Choice as an Untangling of Options into Actions. *Neuron* **99**, 434–447 (2018).

41. Mastrogiuseppe, F. & Ostojic, S. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron* **99**, 609-623.e29 (2018).

42. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 8619–8624 (2014).

43. Gallicchio, C. & Scardapane, S. Deep Randomized Neural Networks. *arXiv:2002.12287 [cs, stat]* (2020).

44. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]* (2014).

45. Salinas, E. & Sejnowski, T. J. Book Review: Gain Modulation in the Central Nervous System: Where Behavior, Neurophysiology, and Computation Meet. *Neuroscientist* **7**, 430–440 (2001).

46. Freeman, W. J. Nonlinear gain mediating cortical stimulus-response relations. *Biological cybernetics* **33**, 237–247 (1979).

47. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun* **10**, 3770 (2019).

48. Jo, J., Verma, V. & Bengio, Y. Modularity Matters: Learning Invariant Relational Reasoning Tasks. *arXiv:1806.06765 [cs, q-bio, stat]* (2018).

49. Kirsch, L., Kunze, J. & Barber, D. Modular Networks: Learning to Decompose Neural Computation. *arXiv:1811.05249 [cs, stat]* (2018).

50. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).

51. Cohen, G., Afshar, S., Tapson, J. & van Schaik, A. EMNIST: an extension of MNIST to handwritten letters. *arXiv:1702.05373 [cs]* (2017).

52. Bertolero, M. A., Yeo, B. T. T. & D'Esposito, M. The diverse club. *Nature communications* **8**, 1277 (2017).

53. Lizier, J. T. JIDT: An Information-Theoretic Toolkit for Studying the Dynamics of Complex Systems. *Front. Robot. AI* **1**, (2014).

54. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping* **15**, 1–25 (2002).