NETWORK SCIENCE

# Emergence and reconfiguration of modular structure for artificial neural networks during continual familiarity detection

**Shi Gu[1,2]*, Marcelo G. Mattar[3], Huajin Tang[4,5], Gang Pan[4,5]**

Copyright © 2024 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

**Advances in artificial intelligence enable neural networks to learn a wide variety of tasks, yet our understanding of the learning dynamics of these networks remains limited. Here, we study the temporal dynamics during learning of Hebbian feedforward neural networks in tasks of continual familiarity detection. Drawing inspiration from network neuroscience, we examine the network's dynamic reconfiguration, focusing on how network modules evolve throughout learning. Through a comprehensive assessment involving metrics like network accuracy, modular flexibility, and distribution entropy across diverse learning modes, our approach reveals various previously unknown patterns of network reconfiguration. We find that the emergence of network modularity is a salient predictor of performance and that modularization strengthens with increasing flexibility throughout learning. These insights not only elucidate the nuanced interplay of network modularity, accuracy, and learning dynamics but also bridge our understanding of learning in artificial and biological agents.**

## INTRODUCTION

Biological and artificial neural networks (ANNs) are powerful models capable of learning and adapting to new information. Such models are used routinely in neuroscience and artificial intelligence (AI) for various applications related to learning. Neuroscience has traditionally focused on understanding the organizational mechanisms of biological neural networks and how they support cognitive process (*1*), although these insights can sometimes inform the development of AI models (*2*–*4*). AI, on the other hand, leverages ANNs in a variety of applications, from computer vision and natural language processing, to more complex tasks that involve decision-making and prediction (*5*), with these applications offering insights into the brain's intricate mysteries (*6*, *7*). While ANNs are often viewed as a digital manifestation of the brain's workings, a holistic framework that perceives an ANN as a dynamic neural system is conspicuously absent.

Current applications of AI in neuroscience can be broadly classified into two categories. The first category applies ANNs as prediction tools to strengthen the power of identifying associations, e.g., using sophisticated models to map brain connectome to labels and developing encoding models based on ANN features (*6*, *8*). The second category builds recurrent neural network (RNN) models to execute cognitive tasks, with the goal of understanding the relationship among tasks and the principles of cognition through manipulating the trained ANNs (*9*, *10*). The methodological philosophy behind these two categories of approaches is that the similarity in performance may suggest similarity in the structure and representation. Yet, the dynamic nature of learning, intrinsic to both artificial and biological entities, remains largely uncharted, underscoring the chasms that persist in our comprehension of ANN methodologies and neural dynamics.

A substantial amount of research has been devoted to understanding the efficacy of ANNs, based primarily on concepts from computational optimization and statistical learning theory. Such perspectives, however, are insufficient to capture the dynamic, nonlinear nature of learning in biological neural networks (*11*, *12*). While we are now able to construct networks capable of impressive feats (*13*), grasping their underlying learning dynamics is still an emerging frontier. This calls for versatile analytical instruments, capable of dissecting the temporal intricacies of ANNs, reflecting the persistent adaptability evident in biological neural networks during learning phases. We argue that a perspective grounded in computational neuroscience is indispensable, drawing from a set of techniques we label Artificial Network Neuroscience.

Here, we illustrate these ideas by studying the learning dynamics of synaptic networks—specifically, Hebbian feedforward (HebbFF) neural networks—during a continual familiarity detection task (*14*). We choose this task and model for two reasons. First, memory-related tasks have been widely studied in network neuroscience especially for the network reconfiguration across the learning procedure (*15*–*18*). Second, the HebbFF model endowed with synaptic plasticity reproduces experimental results in the memory domain (*14*, *19*, *20*), making it an appropriate model for dynamic examination. We hypothesize that leveraging analytical techniques from network neuroscience will facilitate the interpretation of dynamic shifts underlying neural network reconfigurations.

To study the learning dynamics of HebbFF, we developed a multipronged approach examining the dynamic reconfiguration of the networks from different perspectives. We begin with an analysis of the modularity over temporal scales and its relationship to variations in task accuracy and distribution entropy across diverse learning paradigms. We then explored the synchronicity of states during the training phase and its subsequent correlation with accuracy. We show that network modularization increases with learning and that network flexibility serves as a robust metric encapsulating model performance, in line with results from neuroscience in biological organisms. We hope that the analytical approach described here will

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. [2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China. [3]Department of Psychology, New York University, New York, NY 10003, USA. [4]College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [5]State Key Laboratory of Brain Machine Intelligence, Zhejiang University, Hangzhou, China.
*Corresponding author. Email: gus@uestc.edu.cn

Downloaded from https://www.science.org on July 27, 2024

shed light on the interplay between network modularity, accuracy, and learning dynamics, and ultimately advance our understanding of ANNs and their biological counterparts.

## RESULTS

We explored the dynamic reconfiguration of ANNs, focusing on a class of recurrent (RNNs) called HebbFF network. This network determines the familiarity of a stimulus $\mathbf{x}(t)$ based on whether it matches a stimulus encountered previously (*14*). The network input is an $N$-dimensional vector $\mathbf{x}(t)$, and the output $y(t)$ indicates whether $\mathbf{x}(t)$ equals $\mathbf{x}(t-R)$ (Fig. 1A). Note that this task setting resembles classical working memory paradigms, which have been investigated widely in network neuroscience terms of dynamic network reconfiguration (*15*, *17*). The parameters of this continual familiarity detection task include a repeat interval length, $R$, and a vector length, $N$, which we set to $R = 5$ and $N = 100$. We use a neural network with $M = 120$ units in the hidden layer to provide sufficient representation power for encoding the input (Fig. 1B). A more detailed discussion on the memory capacity can be found at (*14*).

We trained 120 network instances representing different subjects. Each network was trained on a training dataset of 2500 samples for 1000 epochs. After every 10 epochs, we evaluated the partially trained model on a testing dataset comprised of a separate 2500 samples, examining the activation of the hidden layer units to investigate the emergence of modularization. For each model instance, thus, we created a $120 \times 2500 \times 100$ tensor with 120 regions, 2500 test time points, and 100 selected training epochs. As the weights of HebbFF are updated through the task, the activation vectors contain information about short memory, which supports the combination of neighboring cases into time windows. The functional connectivity between each pair of units is computed as the correlation between their activations over the 2500 test samples.

In cognitive neuroscience, network modularization is recognized as an important mechanism that enhances computational efficiency and flexibility by facilitating localized information processing (*21*). To understand the role of modularization in HebbFF, we analyzed the evolution of the network's community structure throughout the learning trajectory. We found a pronounced decline in the number of communities early in the learning trajectory, concomitant with an escalation in overall network modularity (Fig. 2, A and B). Such trends are indicative of a scenario wherein the quantitative reduction in module count is counterbalanced by an amplification in the representational capability of the extant modules. A deeper probe into the modular allegiance matrix across varied learning epochs reveals a nascent modular structure as early as epoch 10 (Fig. 2C). By epoch 200, this structure began to delineate with more pronounced modules. By epochs 390 and 580, we observed a further crystallization of these modules, with distinct community structure surfacing. By epochs 770 and 960, the matrix displayed pronounced and contrastive modules, underscoring a heightened modular allegiance. Such dynamics underscore the iterative refinement and consolidation of a community structure during the learning process, echoing the behavior of biological neural networks. These results highlight the emergence of modular, localized processing in fine-tuning learning dynamics, raising the important question of whether such structure is related to task performance.

To analyze more closely the increase of modularity during learning, we examined the activations of the hidden layer. We found that, on average, the hidden layer's activation initially increased in very early stages but declined as training matured toward its latter phases (Fig. 3A).

Downloaded from https://www.science.org on July 27, 2024

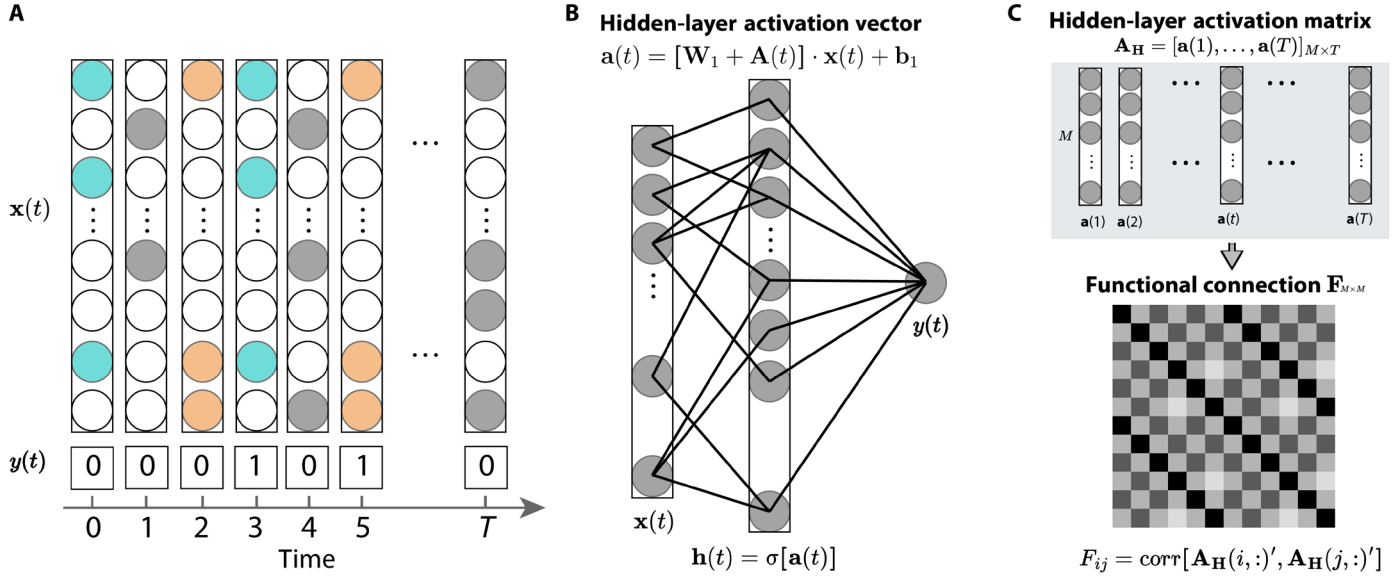


**Fig. 1. Functional connection for HebbFF network.** (**A**) An illustration of the continual familiarity detection task. The network output indicates whether a given stimulus $\mathbf{x}(t)$ matches the stimulus presented previously. (**B**) The HebbFF network, using a hidden layer, encodes the input stimulus and predicts its familiarity through a linear classifier. Here, $M$ is the number of neurons, and $T$ is the number of testing samples. (**C**) Analogous to the brain network, hidden layer activations are extracted in response to a task sequence, consisting of a hidden-layer activation matrix of dimension $M \times T$. The function connection between the neurons $i$ and $j$ is defined as their Pearson's correlation over the testing samples.

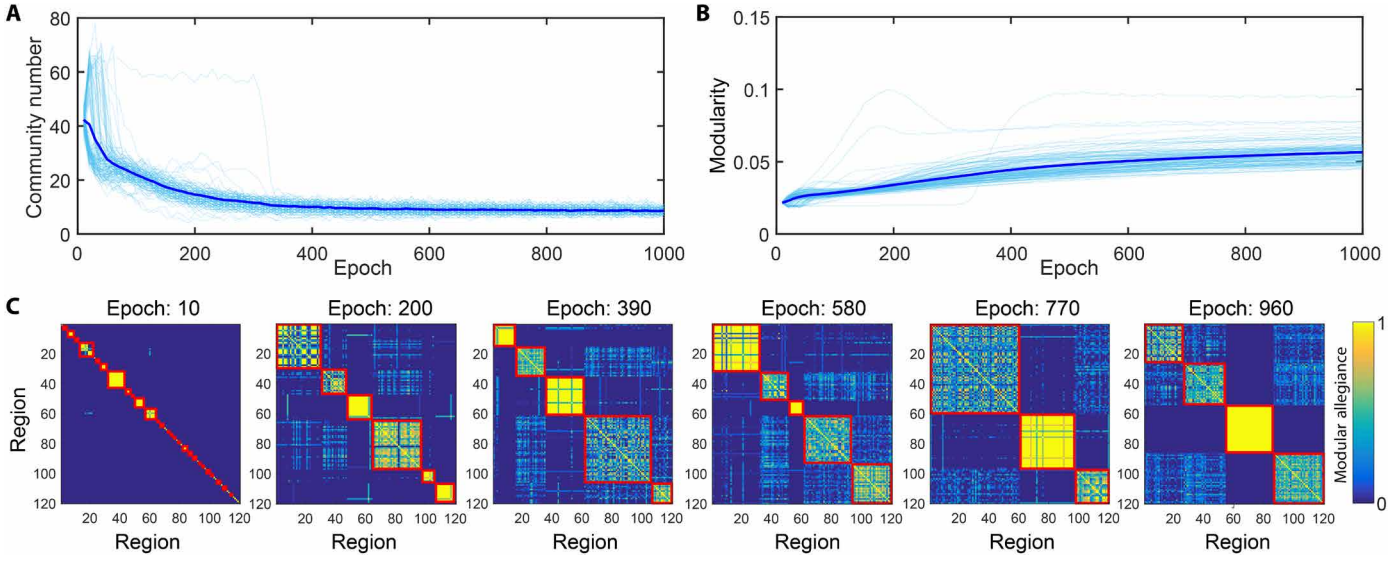


**Fig. 2. Modularization of hidden layer activation over training.** (**A**) The number of communities decreases, and (**B**) the modularity increases through the training. (**C**) The modularization becomes more significant through trainings with less modules and more contrastive modules.

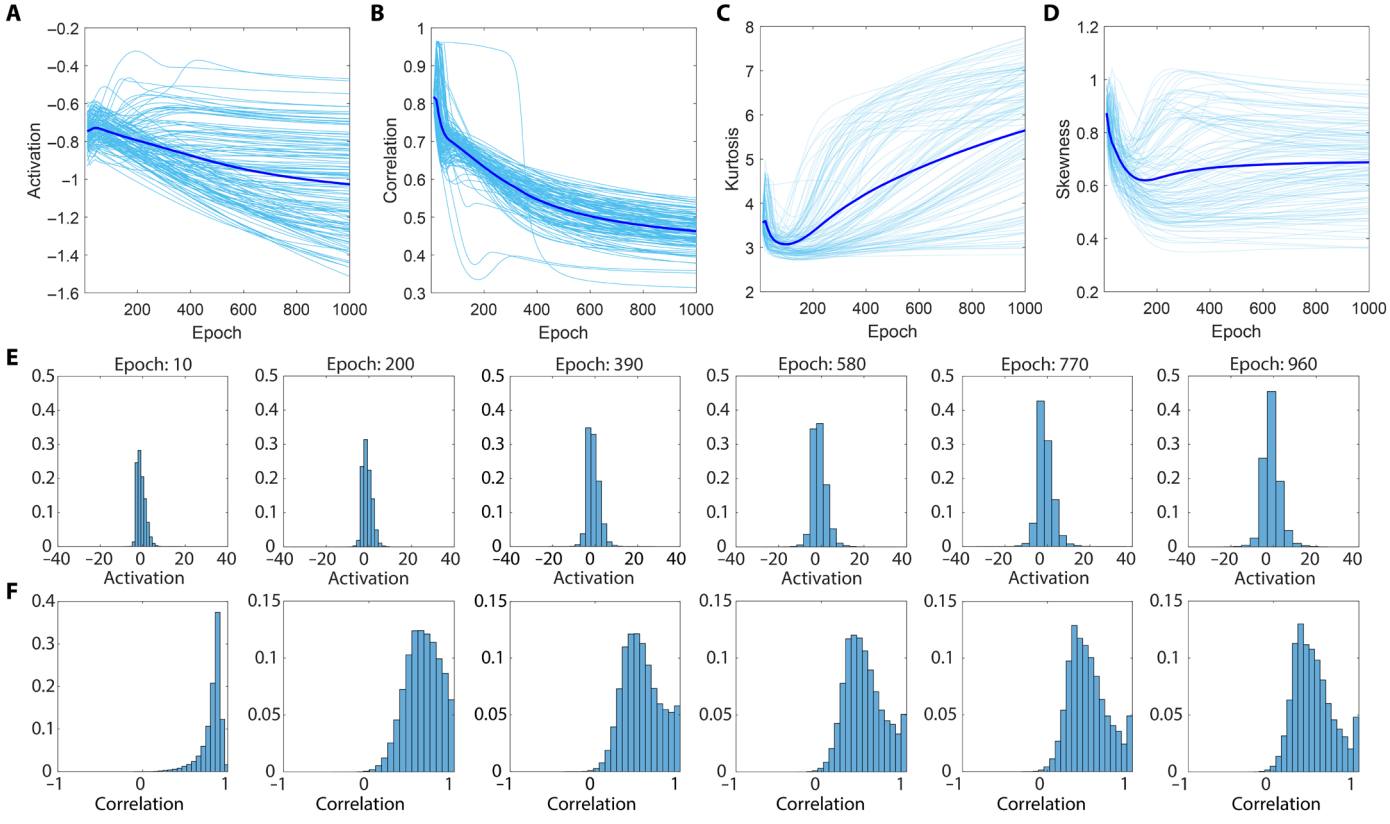


**Fig. 3. Evolution of activation patterns in the hidden layer over training periods.** (**A**) Tracing the trajectory of average activation in the hidden layer, an initial moderate increase is evident in the early training epochs, succeeded by a subsequent decline persisting until the latter epochs. (**B**) An analysis of the activation values' correlation among neurons in the hidden layer exhibits a consistent decrease as training progresses. (**C**) A study of the kurtosis of the activation values across epochs shows an initial decrease, followed by a resurgence in the later stages of training. (**D**) The skewness of activation values delineates an intriguing pattern: an initial decline, transitioning into a bifurcation in later epochs, manifesting both an ascendant trajectory and an alternate trajectory with an initial surge followed by a decline. (**E**) Epoch-specific histograms depict the activation distribution across select epochs, illustrating a growing concentration of values around zero and a reduction in distributional asymmetry as training advances. (**F**) Histograms across epochs, analogous to those in (E), elucidate the shifts in correlation values. Early training predominantly showcases correlations proximate to unity. As training matures, the distribution gravitates toward zero, concluding in a mildly zero-skewed distribution punctuated by an isolated peak at one.

Downloaded from https://www.science.org on July 27, 2024

Such a reversal could result from the enhancement of the distribution's negative spectrum or a possible attenuation of its positive counterpart. Additional insights into this behavior can be extracted from observing the activation distributions across model instances across select epochs, from early to late stages (Fig. 3E). We found that the probability of near-zero activations increased throughout learning, consistent with the increased modularity described previously (Fig. 2C).

We also found interesting patterns in the kurtosis and skewness of the activation distributions. We observed a slight contraction in the kurtosis of hidden layer neural activations early in learning, followed by a marked expansion in the later stages (Fig. 3C). This trend corroborates our observations in Fig. 3E, which demonstrate an accumulation of activation values in the vicinity of zero. The skewness, meanwhile, displayed an intriguing pattern. Early in learning, the distribution of neural activations showed a consistent decline in skewness, representing an increase in the symmetry of the distribution. Yet, as training enters its intermediate and late phases, the skewness bifurcated into dual trajectories. While one set of networks followed a steady increase in skewness, the other showed a rapid increase before a slower decrease. This raises the question of whether these distinct trajectories represent distinct modes of learning, and whether they are relevant for behavior.

To gain additional insights, we examined the correlations within the hidden layer. We found that correlations tended to decrease throughout learning (Fig. 3B). In early stages, hidden layer activations were highly correlated (Fig. 3F, left). As training unfolded, however, the distribution of correlation shifted toward zero, culminating in a slightly zero-skewed distribution punctuated by an isolated peak at one (Fig. 3F, right). Such behavior suggests a desynchronization within the network, suggesting an augmentation in representational power that may underlie an increase in model performance.

In network neuroscience, a modular structure holds dual signification: first, as a descriptor of learning phase differentiation; second, as an indicator of interindividual variability during cognitive assessments. To understand the interplay between modularity and performance within HebbFF, we explored the correlation dynamics between modularity and accuracy, both within specified epoch windows and across disparate model instances (Fig. 4). Looking across windows of 100 epochs, we found a consistently positive correlation between modularity and accuracy, suggesting that more modular

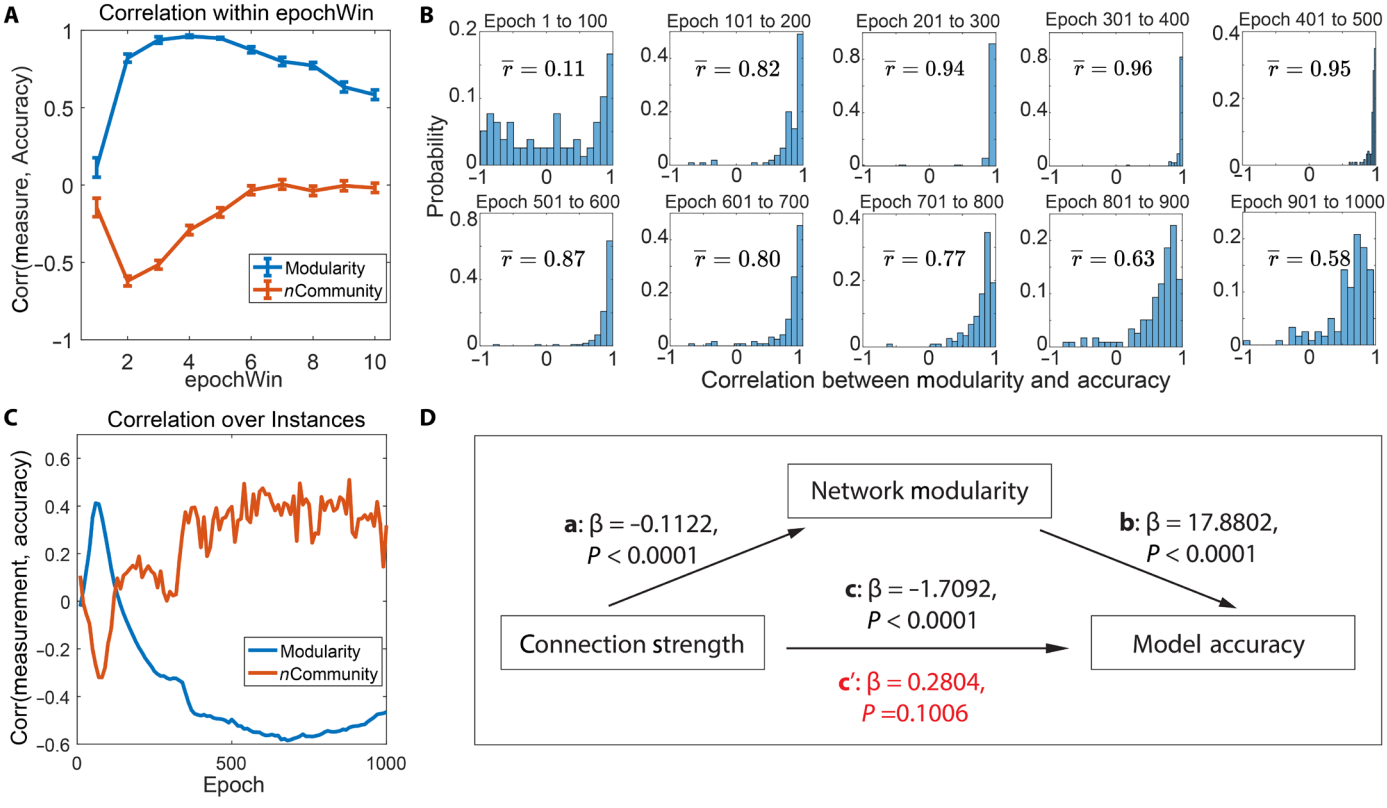


**Fig. 4. Correlation dynamics between network modularity and performance accuracy.** (**A**) Correlation progression between modularity, community number, and accuracy across distinct epoch windows (epochWin), each time point comprises 10 selected epochs corresponding to 100 original epochs in the training procedure. For the modularity and model accuracy, initial stages showcase a surge in correlation nearing unity, which subsequently declines to ~0.6 by the concluding stages. For the number of communities and model accuracy, the correlations display an opposite trend and goes to almost zeros in the late stages. (**B**) Depiction of correlation distributions corresponding to distinct epoch windows, capturing the variability within individual error bars of (A). (**C**) Overall correlation trajectory between modularity, number of communities, and accuracy, calculated across diverse model instances. For the modularity and model accuracy, a marked positive correlation during the formative training stages inversely transitions to a negative domain during intermediate and advanced stages. For the number of communities and model accuracy, the correlations display an opposite trend and remain positive in the late stages. (**D**) Mediation analysis for the connection strength, network modularity and model accuracy. The network modularity fully mediates the effect of the connection strength on the model accuracy.

Downloaded from https://www.science.org on July 27, 2024

networks tend to perform better. This correlation approached one during the median training stages, subsequently retracting to ~0.6 in late stages (Fig. 4A). This trend is congruent with the overarching rise in modularity as training progresses depicted previously (Fig. 2B) and associated with an opposite trend for the change of the number of communities. Given the near-monotonic increase in modularity (Fig. 2B) and in task accuracy (Fig. 4A), the trends in correlations suggest that the heightened modularity facilitates, in median training stages, the memorization of states in the form of strengthening the separation among different states.

Holding the epoch constant and examining correlations across model instances, we found a distinct pattern. The modularity-accuracy correlation demonstrates an initial upswing during the formative training epochs, which then descends to negative magnitudes during the intermediate and concluding phases. The community-number-accuracy correlation displays an overall opposite trend and remains positive in the concluding phases (Fig. 4C). Examining these patterns across epoch windows and model instances reveals a particularly intriguing trend. While a single model's learning curve exhibits a positive correlation between accuracy and modularity, this association switches to negative when examined across model instances. When we calculate the correlation across instances, different instances vary in the number of communities, with a larger number of communities resulting in improved model capacity and better model performances. However, more communities unnecessarily lead to higher modularity values. For a single model, the community structure remains almost unchanged, especially in the late stage of training. Thus, the modularity value is mainly affected by the strength where higher modularity suggests better separation of different modes and, consequently, higher model accuracy (see fig. S2 and eqs. S1 to S8 for a more detailed mechanistic discussion based on idealized models). These findings suggest that, within models subjected to extended training, a robust modular structure, manifesting as markedly modularized activation states, could potentially compromise representational capacity, consequently attenuating overall performance.

As stated above, the change in connection strength may also be predictive of the model's accuracy. Thus, to examine whether the effect of modularity on the model accuracy is caused by changes in connection strength, we performed a mediation analysis where model accuracy is the dependent variable, connection strength is the independent variable, and network modularity is the mediator variable. We used a mixed-effects regression model. First, the connection strength significantly predicts the network modularity (effect **a**: $\beta = -0.1122$, $P_F < 0.0001$, note that $P_F$ denotes $P$ values associated with the $F$ statistic) and model accuracy (effect **c:** $\beta = -1.7092$, $P_F < 0.0001$). In addition, network modularity predicts model accuracy (effect **b**: $\beta = 17.8802$, $P_F < 0.0001$). However, when we execute the regression of model accuracy on both network modularity and connection strength, the effect of connection strength becomes insignificant (**c**′: $\beta = 0.2804$, $P_F = 0.1006$). Based on these results, we can conclude that network modularity fully mediates the effect of the connection strength on the model accuracy. The causal link between network modularity and model accuracy is also supported by a Granger's causality analysis (see fig. S1).

The results thus far illustrate the relevance of network modularity, a static metric of mesoscale structure, for task performance. To complement these findings, we examined also the network's modular flexibility, which quantifies the dynamic reconfiguration rate of the networks' modular structure. In biological networks, modular flexibility has been empirically linked to task execution proficiency in neuroscience studies, displaying distinct patterns across learning phases. Here, we examined modular flexibility to assess its connection with model performance across learning stages and across model instances (Fig. 5).

To examine the dynamics of modular flexibility, the sequence of 2500 time points was divided into 50 time windows, each with length 50. The dynamic functional connectivity matrices (*22*) were then constructed as the Pearson's matrix within each of the 50 time windows (Fig. 5A). We applied the multi-slice Louvain algorithm (*23*) to the 50-layer networks to obtain the temporally varying community association (Fig. 5A). We then calculated the network flexibility (*15*) for each model instance. Initially, we segmented the entire learning procedure (1000 epochs, recorded every 10 epochs) into 10 epoch windows. Within each of these windows, we calculated the correlation between flexibility and accuracy and subsequently aggregated these correlations to produce a composite histogram and error bars. We found a remarkable parallel between increase in modular flexibility and accuracy over the learning process (Fig. 5B). Similarly to the trend found for modularity, correlations between flexibility and accuracy (across model instances) increased in the early learning stages before declining in the subsequent stages (Fig. 5C). Within trained models, we found that lower flexibility, which we associated with increased representational stability, may predict higher memory performance. We found several correlation peaks in epochs 101 to 200 and 201 to 300, resonating with the latter phases of the warm-up and the onset of the progressive stage, respectively. Epoch-specific correlations for the 120 model instances display an intriguing pattern (Fig. 5D). Although flexibility and modularity follow similar trajectories, the peak of the correlation between flexibility and accuracy precedes the peak of the correlation between modularity and accuracy. This suggests that dynamic metrics (e.g., flexibility) might be precursors to the emergent modular representations throughout training. Furthermore, while flexibility contributes to learning, optimal performance may require stable representations.

The patterns observed in the distributions of hidden unit activations (Fig. 3, A to D), when viewed alongside the evolution in modular flexibility (Fig. 5A), suggest varying learning trajectories across different model instances. To discern whether prominent learning patterns exist, we analyzed the performance of HebbFF on the test set throughout the learning process (Fig. 6). This analysis revealed two distinct learning modes. In the first mode (learning mode I), accuracy remained nearly constant for the first 200 epochs (Fig. 6, A and B). In the second mode (learning mode II), accuracy increased rapidly in the first few epochs before decreasing again around epoch 100 (Fig. 6, E and F).

In brain networks, network flexibility represents the community adaptation rate throughout time, with higher flexibility often correlating with heightened cognitive task performance. We wished to determine whether this relationship between network flexibility and performance is also found in HebbFF networks. Both learning modes, despite differing warm-up behaviors, consistently exhibited a monotonic increase in modular flexibility, possibly to increase the network's representation capability and accuracy (Fig. 6, C and G). This raises the possibility that modular flexibility is a potentially broader network characteristic, transcending its traditional heuristic function.

Downloaded from https://www.science.org on July 27, 2024

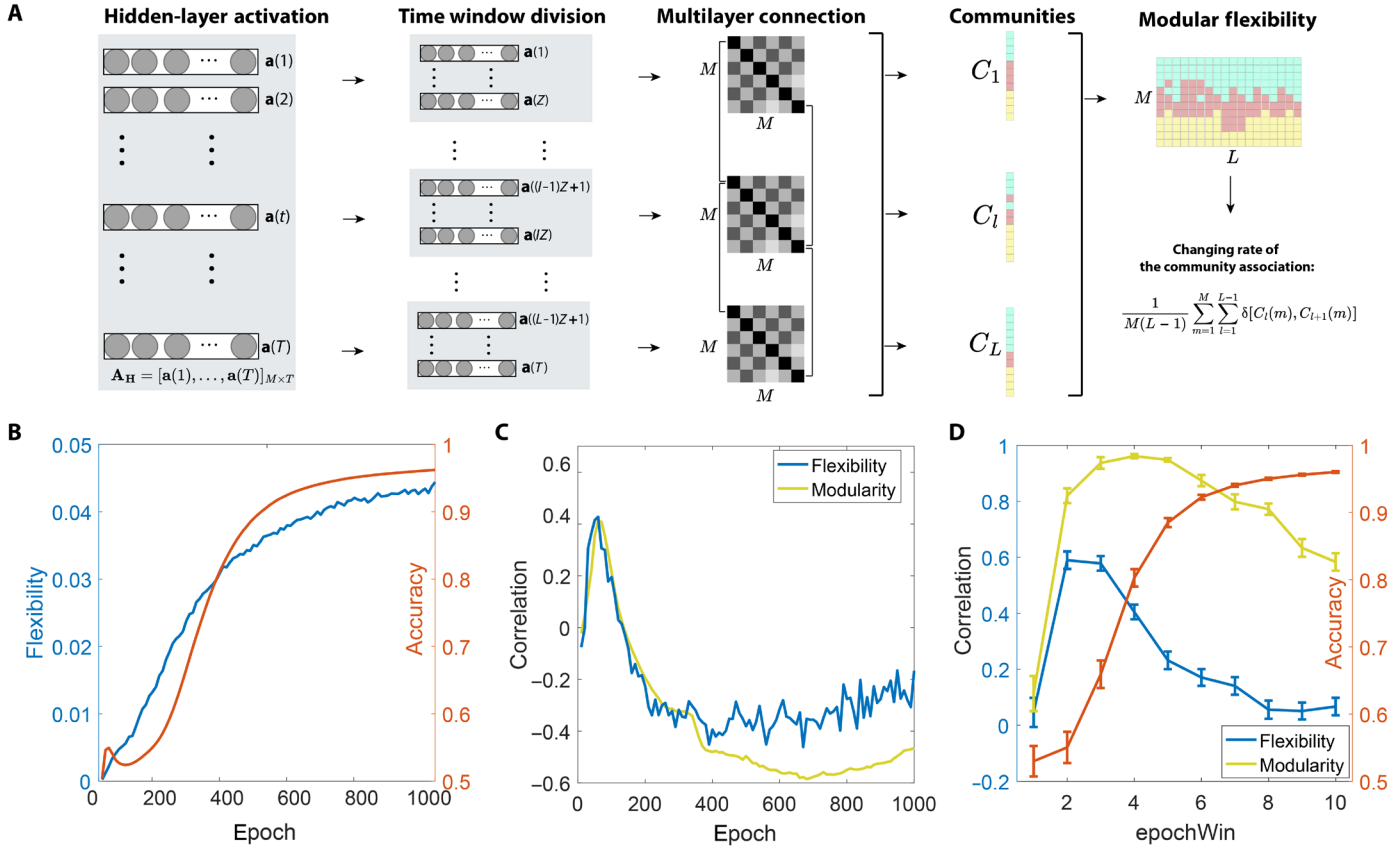


**Fig. 5. Evaluating the interplay between accuracy and modular flexibility during training.** (**A**) Illustration of the construction of multilayer functional networks through correlation within time windows. Here, *M* is the number of nodes, and *L* is the number of time windows. We apply the multilayer community detection algorithm to obtain the community structure for the hidden layer in each epoch window and compute the modular flexibility as the changing rate of the community association across all epoch windows. (**B**) Over the training progression, modular flexibility consistently elevates in tandem with accuracy. Initially, there is a distinctive oscillation in the accuracy curve, mirroring a "warm-up" phase, while modular flexibility follows a clear upward trend. (**C**) The epoch-wise correlation between modular flexibility and accuracy (spanning individual model instances) first ascends, recording positive values, before descending into negative territory. (**D**) Analyzing correlations within designated epoch windows, we discern that while both modularity and flexibility exhibit an upward trajectory, flexibility achieves its correlation zenith earlier, subsequently declining to near-zero levels in the concluding stages.

To further examine the representational capacity of the hidden layers, we gauged the distribution entropy of hidden layer activation, interpreting it as the average informational content. For mode I, despite the averaged entropy curve increasing before the 100-epoch mark, substantial fluctuations were noted between 100 and 300 epochs (Fig. 6D). In contrast, mode II experienced a nearly monotonic rise in entropy after a transient in the first few epochs, suggesting consistent informational augmentation parallel to increasing modular flexibility. Intriguingly, both modes demonstrated entropy (*24*) growth in latter stages, implying that global optima optimization correlates with enhanced representation.

Last, to analyze the generalizability of the link between network modularity and model accuracy, we examined the correlation between the hidden layer modularity and the model accuracy in a dynamic vision sensor image recognition task (*25*). This task aims to categorize an image sequence, a task very different from the familiarity detection task (Fig. 7A). The input sequence first goes through a pretrained encoder to transfer into the hidden states. The hidden states are then updated in the form of $\mathbf{h}(t+1) = \mathbf{W} \cdot \mathbf{h}(t)$. Here, the updating weight matrix $\mathbf{W}$ consists of the elementwise product of two matrix $\mathbf{S}$ that can be either fixed or trained via the Hebbian rule, and $\mathbf{F}$ is trained via the backpropagation. If we fix $\mathbf{S}$ as a constant matrix, then the model reduces to a typical RNN model. The hidden layer has 200 dimensions. We ran the model with 120 random seeds to obtain multiple model instances and averaged the curves to achieve stable results (see the Supplementary Materials for a more detailed description of the model setup). First, functional connections, defined as the covariance of $\mathbf{h}$, displayed an increasing modularity (Fig. 7B) along with the increase of model performance (Fig. 7C). This was true regardless of whether we updated $\mathbf{W}$ with the Hebbian rule for $\mathbf{S}$ or not. Using the Hebbian rule, however, led to higher modularity and better performance as well as fast convergence and earlier overfitting. Overall, these results support our conclusion that ANN modularization is a characteristic of the training procedure.

## DISCUSSION

We studied the learning dynamics of the HebbFF neural network and identified patterns of modularization over the training process.

Downloaded from https://www.science.org on July 27, 2024

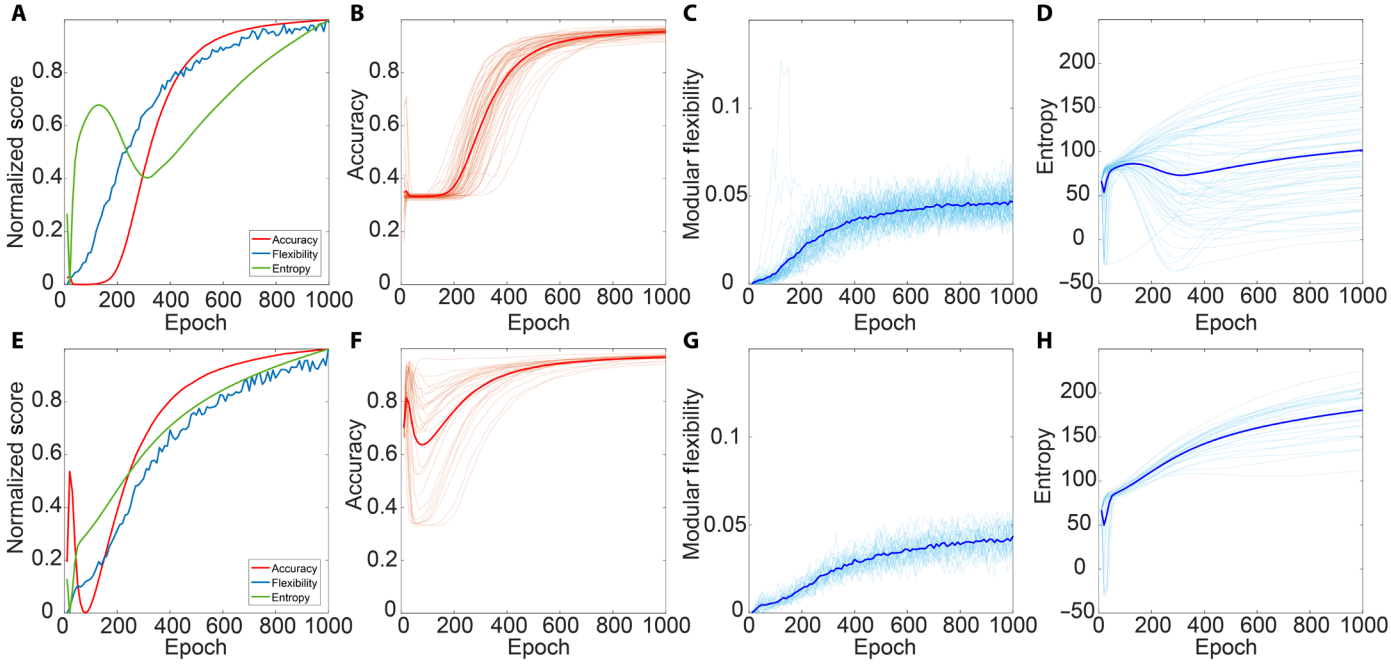


**Fig. 6. Divergent modes of learning in HebbFF networks.** (**A**) Learning mode I showcases different evolutionary trends for accuracy, modular flexibility, and distribution entropy. The accuracy curve (**B**) exhibits a warm-up period of ~20 epochs before ascending to its upper limit. Simultaneously, the modular flexibility (**C**) consistently grows, reaching a plateau, while the distribution entropy of the hidden layer (**D**) peaks during the warm-up phase and then diminishes during the early accuracy surge before experiencing another incline as training concludes. In contrast, learning mode II (**E**) reveals distinct patterns in accuracy, modular flexibility, and distribution entropy compared to mode I. Specifically, the accuracy (**F**) swiftly climbs to a suboptimum and then recedes, facilitating further exploration within the loss landscape to pinpoint global optima. Despite the fluctuating accuracy, both the modular flexibility (**G**) and distribution entropy (**H**) perpetually ascend. For clarity, the normalized score represents the original measurement adjusted to fit within the range of 0 to 1.

By examining the evolution of modularity, modular flexibility, and entropy throughout the learning process, we obtained insights into the learning behavior of the HebbFF neural network that the modularization of the activation pattern of hidden layer neurons predicts the model performance both along training and across model instances. The relationship between modularity and task performance highlights the potential of network neuroscience to characterize the learning behaviors of ANNs.

The relationship between network architecture and learnability has long been an important topic for network science and computational cognitive neuroscience (*26*). Modularization of network architecture has been shown to support both sustained activity (*27*, *28*) and the adaption to varied goals (*29*) with high executive efficiency (*30*). Leveraging RNNs, recent work highlights that spatial constraints (*31*) and specialized information processing (*32*) may also demand a certain level of modularization. In brain networks, prior works also demonstrate a correlation between network segregation and integration with the execution of cognitive tasks (*33*–*35*).

In the specific case of memory-related tasks, a positive correlation between modularity and performance was found to help an individual learn new skills without forgetting old skills (*36*). Typically, these results are interpreted in terms of challenging tasks requiring integration between modules, with the need for integration decreasing through learning (*33*). In line with this prior work, we found that the modularity of HebbFF networks increased during the learning process. These changes reflect the enhancement of modularity values associated with learning (*37*) and neurodevelopment (*38*). The fluctuations in flexibility throughout the learning process also mimic prior findings (*15*, *16*), underscoring the role of increased flexibility in supporting task learning for given subjects. In contrast, in late learning stages, we observed a negative association between modularity and accuracy, as well as between flexibility and accuracy. Such negative association is also observed in reservoirs model, where both overly weak and strong modularization can harm the model performance (*39*). For brain networks, the negative association aligns with the decrease of modularity (*40*) and flexibility (*41*) through the adolescent neurodevelopment. In particular, the Simpson's paradox (*42*) in Fig. 4 (A and C) may stem from a complex interplay between local cohesion and global connectivity (*39*), potentially reflecting developmental trend and individual difference in neurodevelopment.

In our investigation of HebbFF networks, we also observed that the processes of segregation and integration occurred not only among different functional modules but also within the representation of diverse features. This suggests that network modularization, as it pertains to feature representation, may hold substantial relevance for understanding the functionality of the brain. In the brain, different systems handle different aspects of the input and decision-making processes, implying that features could be encoded across multiple interconnected systems. This holistic approach to understanding neural networks, by considering the role of feature representation in network modularization, may provide unique insights into the complex dynamics of both artificial and biological learning systems. The consistency of the results on HebbFF and brain

Downloaded from https://www.science.org on July 27, 2024

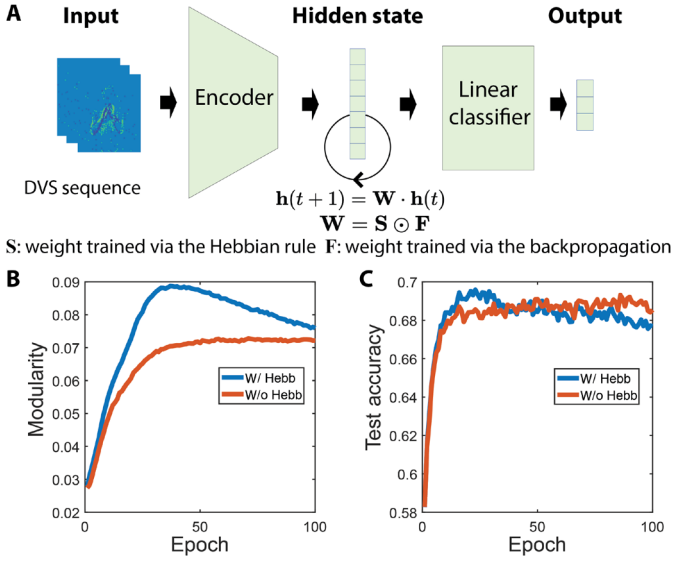


**Fig. 7. Generalization on the spiking neural network with DVS input.** To explore the generalizability of the observation of modularity over the training, we examine the modularity change over training in an image recognition task with a spiking neural network. (**A**) The model takes an image sequence acquired with the dynamic vision sensor (DVS) camera and outputs the category label. The hidden layer states are updated in a recurrent form with mixed mechanism of Hebbian rules and backpropagation. (**B**) Network modularity in hidden layer increased regardless of whether a Hebbian learning was used. The network trained with the Hebbian rule had higher modularity values than the network trained without the Hebbian rule. (**C**) Test accuracy for networks trained with and without Hebbian learning.

networks support the link between the network characteristics and the cognitive execution (*43*). In addition, the role of modularization in the representation is also supported in previous work varying from neuroscience (*44*) to machine learning (*45*). Thus, beyond offering a nuanced interpretation of network methodologies, our study points to a potentially unifying understanding of dynamic reconfiguration roles in both human cognition and machine learning.

Existing network neuroscience analyses of functional brain networks typically model the brain as a dynamic system that follows either a random walk (*46*) or a given geometric flow (*47*). These analyses are based on the implicit assumption that the variation of the information segregation and integration would cause corresponding difference in the signal space. Here, because functional connectivity is constructed directly from patterns of activation in the hidden layer units, we provide a more directive examination on the relationship between the learning behavior and the representation topology. This approach offers numerous insights into the performance of ANNs, including how the topology of the feature representation varies in early stages of learning, even when accuracy does not increase. While previous work (*12*) investigates how network topology affects the performance, our work adds to this literature by suggesting that, in addition to the network structure and its associated static measurement, the networked dynamics may provide additional evidence on how the model performance is related to the structure. This aligns with neuroscience findings that the dynamics of functional connectivity provides better predictions of cognitive scores than static measures (*48*). Accordingly, our work suggests a previously unidentified family of measurements and can facilitate the design of brain-inspired neural networks.

More broadly, our work contributes to the field of AI for neuroscience by providing a better understanding of the dynamical behavior of ANNs through a neuroscience-inspired lens. The endeavor to understand the human brain and develop efficient AI systems has often been a mutually beneficial process. However, a major challenge in AI for neuroscience has been bridging the gap between the static, linear analysis commonly used in machine learning and the dynamic, nonlinear characteristics that are intrinsic to the biological brain. Traditional methods for analyzing neural networks, such as studying the weight and bias parameters, may not capture the complete picture of how an ANN learns, adapts, and evolves over time. In this context, our study provides previously unidentified tools and methodologies to better understand these dynamics. By constructing brain-like networks from HebbFF and applying techniques like community detection and modular analysis, we mirror the modular structure and dynamic reconfiguration seen in the biological brain. In essence, we provide an avenue to study the temporal evolution and adaptation of ANNs during training, similar to the continual reconfiguration observed in the brain during learning. This approach enhances our understanding of the similarities and differences between ANNs and the human brain, opening new avenues for improving the design and training of ANNs. Our work also takes a step toward closing the gap between the simplistic activation and loss landscapes usually used in AI research and the complex, high-dimensional, and dynamic landscapes that are likely in the brain.

In our study, we have primarily concentrated on memory tasks and the emergence of modular structures identified through modularity maximization (*49*). We note, however, that various other methodologies can identify modular structures, and their significance quantified through different means, such as block-based models (*50*) or by comparing the relative strengths of inter-block and intra-block connections (*51*). Moreover, while flexibility in brain networks is commonly interpreted as a marker of neural plasticity at the network level (*26*), in ANNs, it relates to the system's capacity to adapt and represent multiple states. This adaptability is influenced not only by the architecture of the network but also by the chosen hyperparameters.

In sum, our discoveries contribute to the broader aim of intersection between AI and neuroscience, using AI not only to replicate but also to understand and learn from the intricate workings of the brain. The tools and methods that we developed present previously unexplored opportunities to study learning dynamics in both artificial and biological neural networks. Such cross-fertilization of ideas can potentially lead to more efficient, adaptable, and robust AI systems while providing insights into the neuroscience of learning and memory. Future work should expand our approach to other cognitive tasks like multimodal matching, value decision, and perception tasks, as well as to other types of ANNs including RNN and deep feedforward networks. We hope that multimodal continual tasks learned through complex networks could serve as a digital analog of the brain in terms of cognitive execution and may provide novel insights into how functional modules reconfigure to support complex tasks.

## MATERIALS AND METHODS

### HebbFF network architecture

We adopted the same HebbFF network for a continual familiarity detection task in (*14*). In this task, the Hebbian network takes a stream

Downloaded from https://www.science.org on July 27, 2024

of stimuli in the form of randomly generated $N \times 1$–dimensional input vector $\mathbf{x}(t)$ and returns a label $y(t)$ that indicates whether $\mathbf{x}(t)$ has appeared previously. For both the training and testing datasets, the $\mathbf{x}(t)$ is generated as a $\pm 1$ binary vector that equals to $\boldsymbol{x}(t-R)$ with probability $P$. Here, $R$ is the repeat interval length. The HebbFF network consists of three layers: the input, output, and hidden layers. The hidden layer consists of $M$ neurons. We use an $M \times 1$ vector $\mathbf{a}(t)$ to denote the hidden state and $\mathbf{h}(t) = \sigma[\mathbf{a}(t)]$ to denote the activation state after an activation function $\sigma(\cdot)$, where $\sigma(x) = \frac{1}{1+\exp(-x)}$ takes the form of the sigmoid function. The hidden state $\mathbf{a}(t)$ is obtained through a linear transformation on the input $\mathbf{x}(t)$, which is given by

$$\mathbf{a}(t) = \left[\mathbf{W}_1 + \mathbf{A}(t)\right]\mathbf{x}(t) + \mathbf{b}_1 \tag{1}$$

where $\mathbf{W}_1$ is an $M \times N$ matrix denoting the affine transformation and $\mathbf{b}_1$ is an $M \times 1$ vector denoting the bias. The matrix $\mathbf{A}(t)$ is the plasticity matrix that is updated at every time step to take care of the memory. It is calculated as

$$\mathbf{A}(t+1) = \lambda \cdot \mathbf{A}(t) + \eta \cdot \mathbf{h}(t)\mathbf{x}(t)^T \tag{2}$$

where $\lambda$ is the learnable decay parameter and $\eta$ is the learnable familiarity learning rate. When $\eta > 0$, it is called the Hebbian learning; when $\eta < 0$, it is called the anti-Hebbian learning. As demonstrated in (*14*) and supported by our own experiments, we adopt the anti-Hebbian learning rules for the familiarity detection. Further, the readout is given by

$$\hat{y}(t) = \sigma\left[\mathbf{W}_2\mathbf{h}(t) + b_2\right] \tag{3}$$

where $\mathbf{W}_2$ of dimension $1 \times M$ and $b_2$ of dimension $1 \times 1$ are learnable transformation weight and bias. The training loss is then set as

$$\mathscr{L} = \frac{1}{T}\sum_{t=1}^{T} y(t)\log\left[\hat{y}(t)\right] + \left[1-y(t)\right]\log\left[1-\hat{y}(t)\right] \tag{4}$$

which represents the cross-entropy between the label $y$ and prediction $\hat{y}$. The activation values $\mathbf{a}(t)$ and $\mathbf{A}(t)$ are updated through the training and thus can be used to analyze the network dynamics. The network structure is shown in Fig. 1 (A and B).

### Construction of brain-like networks from the HebbFF activation

For an instance of HebbFF, after being training for $t$ epochs, we evaluate the model on the testing dataset $\mathbf{X}_{\text{test}}$ of dimension $N \times T$ and collect the hidden states into a matrix $\mathbf{A}_{\text{H}} = [\mathbf{a}(1), \ldots, \mathbf{a}(T)]$ of dimension $M \times T$, which can be taken as the recorded activation sequence of the HebbFF for continuously executing $T$ tasks. We then divide the full $\mathbf{A}_{\text{H}}$ into $L$ time windows each of the length $Z = T/L$. As the length parameter $T$ is fully controllable, we assume that $T$ can be divided up by $Z$ for simplicity. We obtain a multilayer network $\{F^k{}_{ijl}\}$ where $F^k{}_{ijl}$ is the Pearson's correlation value of the $i$th and $j$th rows in $\mathbf{A}_{\text{H}}$ within the $l$th time window after the $k$th epoch's training. On the basis of these $\{F^k{}_{ijl}\}$ along the full $K$ epochs' training, we can perform network-based analysis to characterize the learning behavior of HebbFF through training.

### Modularization

Community detection is a method that decomposes a system into subsystems (*23*). For a given multilayer network $\{F_{ijl}\}$ where $i$ and $j$ denote the regions and $l$ denote the layers. The multilayer modularity function is given as

$$Q = \frac{1}{2\mu}\sum_{ijlr}\left(F_{ijl} - \gamma_l\frac{k_{il}k_{jl}}{2m_l}\delta_{lr} + \delta_{ij}C_{jlr}\right)\delta(g_{il}, g_{jr}) \tag{5}$$

where the adjacency matrix of layer $l$ has component $F_{ijl}$; $\gamma_l$ is the resolution parameter of layer $l$; $g_{il}$ and $g_{jr}$ give the community assignments of nodes $i$ and $j$ in layers $l$ and $r$, respectively; and $k_{il}$ is the strength of node $i$ in the layer $l$ with $2\mu = \sum_{jr}\kappa_{jr}$, $\kappa_{jl} = k_{jl} + c_{jl}$, and $c_{jl} = \sum_r C_{jlr}$. When $F_{ijl}$ is signed, one can split the positive and negative part and construct the modularity function similarly as shown in (*52*). Through maximizing $Q$, we can get the community structure $g_{il}$ of each node in each layer, which allows us to further investigate the networked dynamics of modules. Further, the module-allegiance matrix $\mathbf{P}$ is defined as the matrix whose element $P_{ij}$ denotes the frequency of nodes $i$ and $j$ in the same partition.

### Dynamic reconfiguration of HebbFF networks

On the basis of the constructed network series $\{A_{ijl}\}$ and their associated community structure $g_{il}$, we can then define the modular flexibility $f_i$ as the changing frequency of the community association over time (*15*), i.e., $f_i = \frac{1}{L-1}\sum_{l=1}^{L-1}\left[1 - \delta\left(g_{i,l}, g_{i,l+1}\right)\right]$. Further, we can define the modular flexibility of the system as $f = \frac{1}{M}\sum f_i$. If the system has a high flexibility, then it indicates that the module structure of the system changes fast, thus suggesting a high flexibility in the representation of the learned features.

### Information entropy of activation variables

For a HebbFF network, the modular structure of the activation-induced correlation network is supported by the similarity of neuron's activation patterns. To quantify the strength of different region's participation in support the temporal dynamics, we adopt the entropy of a random variable quantifies the average information contained in the outcome (*24*). It is defined as the expectation of the log of the density for a continuous distribution. Mathematically, it is denoted as $H(x) = E_{\mathbf{x}}[-\log P(\mathbf{x})]$. For each HebbFF, when it processed $K$ samples, we can collect $K$ activation vector $\mathbf{a}_1, \ldots, \mathbf{a}_K$. On the basis of these $\mathbf{a}_1, \ldots, \mathbf{a}_K$, we can estimate the $H(\mathbf{a})$ as the information contained by the hidden layer. Here, as our purpose is not accurately defining the information quantity, thus, for simplicity, we assume that this $\mathbf{a}_i$ follows a multivariate Gaussian distribution.

### Mixed-effects model

We use the mixed effects model for the regression in examining the relationship between modularity, connection strength, and model accuracy. Suppose the dependent variable is $Y_{ij}$ for the $i$th model and $j$th sample and the independent variable is $X_j$. Then, the mixed effects model is given as $Y_{ij} = \mu + U_i + V_i \cdot X_j + \beta \cdot X_j + \epsilon_{ij}$, where $U_i$ and $V_i$ are the random effects for the $i$th model instance and $\beta$ and $\mu$ are the fixed effects. We then apply the fixed effects to construct the mediation analysis.

## Supplementary Materials

**This PDF file includes:**

Supplementary Results
Figs. S1 and S2
References

Downloaded from https://www.science.org on July 27, 2024

## REFERENCES AND NOTES


1. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X. J. Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
2. J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie, L. Shi, Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019).
3. C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, B. Kay, Opportunities for neuromorphic computing algorithms and applications. *Nat. Comput. Sci.* **2**, 10–19 (2022).
4. S. Ullman, Using neuroscience to develop artificial intelligence. *Science* **363**, 692–693 (2019).
5. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
6. Y. Takagi, S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023), pp. 14453–14463.
7. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
8. C. Du, K. Fu, J. Li, H. He, Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10760–10777 (2023).
9. T. Ito, G. R. Yang, P. Laurent, D. H. Schultz, M. W. Cole, Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. *Nat. Commun.* **13**, 673 (2022).
10. G. R. Yang, M. Molano-Mazón, Towards the next generation of recurrent network models for cognitive neuroscience. *Curr. Opin. Neurobiol.* **70**, 182–192 (2021).
11. X.-J. Wang, Theory of the multiregional neocortex: Large-scale neural dynamics and distributed cognition. *Annu. Rev. Neurosci.* **45**, 533–560 (2022).
12. S. Xie, A. Kirillov, R. Girshick, K. He, "Exploring randomly wired neural networks for image recognition" in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019), pp. 1284–1293.
13. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712 (2023).
14. D. Tyulmankov, G. R. Yang, L. F. Abbott, Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron* **110**, 544–557.e8 (2022).
15. D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, S. T. Grafton, Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7641–7646 (2011).
16. U. Braun, A. Schäfer, H. Walter, S. Erk, N. Romanczuk-Seiferth, L. Haddad, J. I. Schweiger, O. Grimm, A. Heinz, H. Tost, A. Meyer-Lindenberg, D. S. Bassett, Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11678–11683 (2015).
17. U. Braun, A. Harneit, G. Pergola, T. Menara, A. Schäfer, R. F. Betzel, Z. Zang, J. I. Schweiger, X. Zhang, K. Schwarz, J. Chen, G. Blasi, A. Bertolino, D. Durstewitz, F. Pasqualetti, E. Schwarz, A. Meyer-Lindenberg, D. S. Bassett, H. Tost, Brain network dynamics during working memory are modulated by dopamine and diminished in schizophrenia. *Nat. Commun.* **12**, 3478 (2021).
18. Y. He, X. Liang, M. Chen, T. Tian, Y. Zeng, J. Liu, L. Hao, J. Xu, R. Chen, Y. Wang, J.-H. Gao, S. Tan, J. Taghia, Y. He, S. Tao, Q. Dong, S. Qin, Development of brain state dynamics involved in working memory. *Cereb. Cortex* **33**, 7076–7087 (2023).
19. S. Qin, S. Farashahi, D. Lipshutz, A. M. Sengupta, D. B. Chklovskii, C. Pehlevan, Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nat. Neurosci.* **26**, 339–349 (2023).
20. H. G. Rodriguez, Q. Guo, T. Moraitis, "Short-term plasticity neurons learning to learn and forget" in *Proceedings of the 39th International Conference on Machine Learning* (ACM, 2022), pp. 18704–18722.
21. O. Sporns, R. F. Betzel, Modular brain networks. *Annu. Rev. Psychol.* **67**, 613–640 (2016).
22. M. G. Preti, T. A. Bolton, D. Van De Ville, The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage* **160**, 41–54 (2017).
23. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328**, 876–878 (2010).
24. D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge Univ. Press, 2003).
25. A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, D. Modha, "A low power, fully event-based gesture recognition system" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), pp. 7243–7252.
26. P. Zurn, D. S. Bassett, Network architectures supporting learnability. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190323 (2020).
27. S.-J. Wang, C. Hilgetag, C. Zhou, Sustained activity in hierarchical modular neural networks: Self-organized criticality and oscillations. *Front. Comput. Neurosci.* **5**, 30 (2011).
28. M. Kaiser, C. Hilgetag, Optimal hierarchical modular topologies for producing limited sustained activation of neural networks. *Front. Neuroinform.* **4**, 8 (2010).
29. N. Kashtan, U. Alon, Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci.* **102**, 13773–13778 (2005).
30. J. Clune, J.-B. Mouret, H. Lipson, The evolutionary origins of modularity. *Proc. Biol. Sci.* **280**, 20122863 (2013).
31. J. Achterberg, D. Akarca, D. J. Strouse, J. Duncan, D. E. Astle, Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nat. Mach. Intell.* **8**, 1369–1381 (2023).
32. J. Tanner, S. Mansour L, L. Coletta, A. Gozzi, R. F. Betzel, Functional connectivity modules in recurrent neural networks: Function, origin and dynamics. arXiv:2310.20601 (2023).
33. D. S. Bassett, M. Yang, N. F. Wymbs, S. T. Grafton, Learning-induced autonomy of sensorimotor systems. *Nat. Neurosci.* **18**, 744–751 (2015).
34. G. Tononi, O. Sporns, G. M. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci.* **91**, 5033–5037 (1994).
35. R. Wang, M. Liu, X. Cheng, Y. Wu, A. Hildebrandt, C. Zhou, Segregation, integration, and balance of large-scale resting brain networks configure different cognitive abilities. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022288118 (2021).
36. K. O. Ellefsen, J.-B. Mouret, J. Clune, Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS Comput. Biol.* **11**, e1004128 (2015).
37. K. Finc, K. Bonna, X. He, D. M. Lydon-Staley, S. Kühn, W. Duch, D. S. Bassett, Dynamic reconfiguration of functional brain networks during working memory training. *Nat. Commun.* **11**, 2435 (2020).
38. G. L. Baum, R. Ciric, D. R. Roalf, R. F. Betzel, T. M. Moore, R. T. Shinohara, A. E. Kahn, S. N. Vandekar, P. E. Rupert, M. Quarmley, P. A. Cook, M. A. Elliott, K. Ruparel, R. E. Gur, R. C. Gur, D. S. Bassett, T. D. Satterthwaite, Modular segregation of structural brain networks supports the development of executive function in youth. *Curr. Biol.* **27**, 1561–1572.e8 (2017).
39. N. Rodriguez, E. Izquierdo, Y.-Y. Ahn, Optimal modularity and memory capacity of neural reservoirs. *Netw. Neurosci.* **3**, 551–566 (2019).
40. J. R. Cohen, M. D'Esposito, The segregation and integration of distinct brain networks and their relationship to cognition. *J. Neurosci.* **36**, 12083–12094 (2016).
41. S. Gu, P. Fotiadis, L. Parkes, C. H. Xia, R. C. Gur, R. E. Gur, D. R. Roalf, T. D. Satterthwaite, D. S. Bassett, Network controllability mediates the relationship between rigid structure and flexible dynamics. *Netw. Neurosci.* **6**, 275–297 (2022).
42. E. H. Simpson, The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Series B* **13**, 238–241 (1951).
43. J. D. Medaglia, M.-E. Lynall, D. S. Bassett, Cognitive network neuroscience. *J. Cogn. Neurosci.* **27**, 1471–1491 (2015).
44. J. M. Shine, Neuromodulatory influences on integration and segregation in the brain. *Trends Cogn. Sci.* **23**, 572–583 (2019).
45. Z. Liu, O. Kitouni, N. S. Nolte, E. Michaud, M. Tegmark, M. Williams, Towards understanding grokking: An effective theory of representation learning. *Adv. Neural Inf. Process. Syst.* **35**, 34651–34663 (2022).
46. E. Bullmore, O. Sporns, Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198 (2009).
47. G. Deco, M. L. Kringelbach, Great expectations: Using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron* **84**, 892–905 (2014).
48. R. Liégeois, J. Li, R. Kong, C. Orban, D. Van De Ville, T. Ge, M. R. Sabuncu, B. T. T. Yeo, Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nat. Commun.* **10**, 2317 (2019).
49. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
50. C. Aicher, A. Z. Jacobs, A. Clauset, Learning latent block structure in weighted networks. *J. Complex Netw.* **3**, 221–248 (2015).
51. A. Nematzadeh, E. Ferrara, A. Flammini, Y.-Y. Ahn, Optimal network modularity for information diffusion. *Phys. Rev. Lett.* **113**, 088701 (2014).
52. S. Gu, C. H. Xia, R. Ciric, T. M. Moore, R. C. Gur, R. E. Gur, T. D. Satterthwaite, D. S. Bassett, Unifying the notions of modularity and core–periphery structure in functional brain networks during youth. *Cereb. Cortex* **30**, 1087–1102 (2020).
53. J. Geweke, Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.* **77**, 304–313 (1982).


Downloaded from https://www.science.org on July 27, 2024

54. W. Fang, Y. Chen, J. Ding, Z. Yu, T. Masquelier, D. Chen, L. Huang, H. Zhou, G. Li, Y. Tian, SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Sci. Adv.* **9**, eadi1480 (2023).

**Acknowledgments:** We thank S. Deng in UESTC for helping implement the SNN experiments. **Funding:** S.G. is supported by NSFC Key Program, 62236009; Shenzhen Fundamental Research Program (General Program), JCYJ 20210324140807019; NSFC General Program, 61876032; and Key Laboratory of Data Intelligence and Cognitive Computing, Longhua District, Shenzhen.
**Author contributions:** Conceptualization, methodology, investigation, and visualization: S.G. Writing—original draft: S.G. and M.G.M. Writing—review and editing: S.G., M.G.M., H.T., and G.P.
**Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All original data in this work was generated by program provided at https://github.com/dtyulman/hebbff with some homemade adaption for storage and parallel execution. The code for data processing and analysis can be found at Zenodo (https://doi.org/10.5281/zenodo.12207489).

Submitted 8 November 2023
Accepted 21 June 2024
Published 26 July 2024
10.1126/sciadv.adm8430

Downloaded from https://www.science.org on July 27, 2024

Science Advances
AAAS

# Supplementary Materials for

## Emergence and reconfiguration of modular structure for artificial neural networks during continual familiarity detection

Shi Gu *et al.*

Corresponding author: Shi Gu, gus@uestc.edu.cn

*Sci. Adv.* **10**, eadm8430 (2024)
DOI: 10.1126/sciadv.adm8430

**This PDF file includes:**

**Supplementary Results**

**Granger's causality between Modularity and Accuracy.** For the causal relationship between the modularity and accuracy, we analyze the relationship between modularity and accuracy through the granger's causality with $F$-test (*53*). We set the $\text{maxLag} = 2$ and repeated the tests on all the 120 instances that we used in the manuscript. The first degree of freedom ($df_1$) is then 2 and the second degree of freedom ($df_2$) is 94. In Fig. S1, we can see that for most of the cases, the causal relation is statistically significant. Further, as we have run multiple model instances, we consider the random effect of different instances by adopting the mixed effect model in calculating the regression. Then we have $df_1 = 360$ , $df_2 = 11040$ and $F = 12.6558$, which corresponds to $p < 1e^{-14}$.

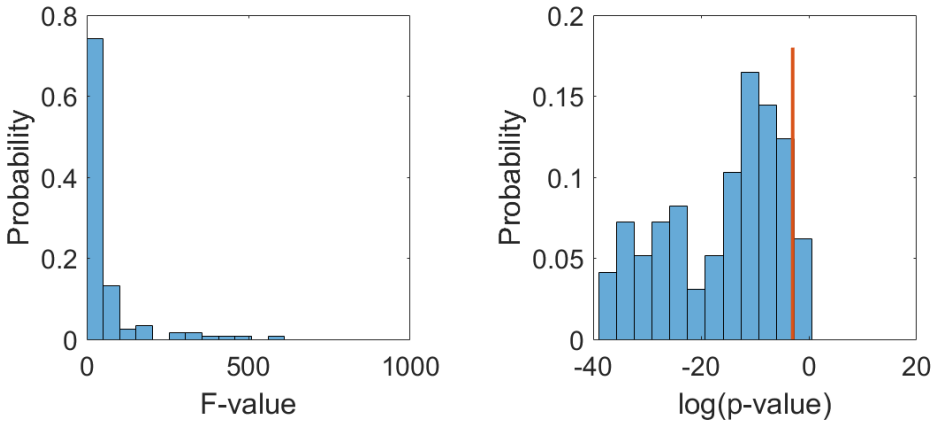


**Figure S1. Granger's Causality of Network Modularity on Model Accuracy.** The max lag of causality is set to 2. The degrees of freedom for F is $df_1$ = 2 and $df_2$ = 94. The left panel showed the distribution of F-values. The right panel showed the distribution of the log of $p$-values. The red line corresponds to $p$ = 0.05.

**Simpson's Paradox of the opposite trend for correlations over instances and within epochs.** To explain why the correlations between modularity and model accuracy are opposite when calculated over different model instances and within the epoch window for each model instance, we examine the change of community numbers. We find that the community number remains to display a positive correlation in the late training states (Fig. S2a) and goes to almost zero when the correlation is calculated along the training procedure. Thus we infer that when we calculated the correlation over instances, different instances varied in the number of communities where more communities indicated improved model capacity (*like* (*14*)) thus better model performances. However, more communities may not lead to higher modularity values (Fig. S2c, $Q_1 > Q_2$). Respectively, for the same model, especially in the late stage of training, the community structure almost remained unchanged thus the modularity value is mainly affected by the strength (Figure S2c, $Q_2 < Q_3$) where higher modularity suggests better separation of

different modes thus higher model accuracy. Such differences may affect cause the different trend of correlations when calculated over instances and training epochs.

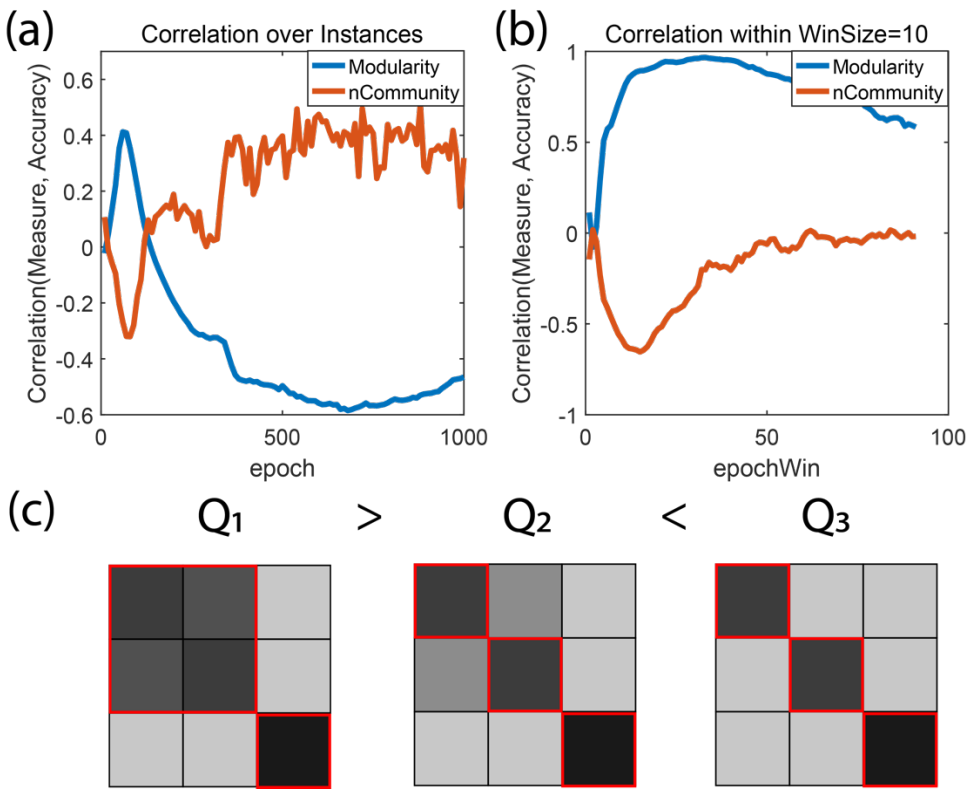


**Figure S2. Dynamics of Network Modularity and Community Structure During Training Across Instances.** (a) Modularity (Q) and the number of communities (nCommunity) are plotted over the course of training. The correlation between these network metrics and accuracy was computed over 120 instances to discern patterns. (b) For each individual model instance, a sliding window of 10 epochs was used to calculate the evolving correlation between network measurements (modularity and number of communities) and model performance, capturing the temporal dynamics within the training process. (c) Illustrative examples depict how variations in modularity (Q) relate to changes in the number of communities and the strength of connections within the network. This toy model aims to clarify potential fluctuations in modularity with respect to network structure and training progression.

**Model-based Explanation on the Relationship between Network Modularity and Model Accuracy.** The main observation comes in two parts: 1) for a given network, the modularity increases with the training; 2) for multiple model instances, the modularity negatively correlates with the accuracy. The first observation can be taken as an immediate corollary of the strengthened modularization in the representation space through model training. Specially, the distribution of the hidden layer activation gradually shifts from the initial random distribution to a multi-modal distribution, which was already fully demonstrated in *Tyulmankov et al.2022* (*1*). Thus, the modularity positively correlates with the accuracy as both increases. For the second observation, the relationship is a bit complex. Through the model training, the model needs to both separate different modes into different hidden neurons and have one stand out for the

correct label. These two goals determine the final model accuracy and are mainly affected by the relative strength of the transformation and bias. Thus, identifying the dependence of modularity and accuracy on the model parameter may provide mechanistic explanation on their connection. We disentangle the explanation in four major steps below.

**Step 1:** Derive the form of covariance matrix of the hidden layer activation which we use to construct the Function Connection (FC) matrix.

**Step 2:** Derive the dependence of model accuracy on the trained parameter with approximation when $t$ is large and the model adopts an idealized setting as *Tyulmankov et al.2022* (*1*).

**Step 3:** Discuss how the common parameter affect i) the modularity through its impact on the FC and ii) the model performance through the approximation of the probability function.

**Step 4:** Identify the negative trend of impact from the parameters thus explains for the negative correlation between the network modularity and model accuracy.

Here, we provide a derivation of how the similarity matrix of the hidden layer's activation depends on the addressing function that is updated over the training (*14*). Following the same setup of the idealized model in, the firing rate of the hidden layer is given by $\boldsymbol{h}(t) = \Theta(\boldsymbol{W}_1\boldsymbol{x}_{\boldsymbol{W}}(t) + \boldsymbol{A}(t)\boldsymbol{x}_{\boldsymbol{A}}(t) + \boldsymbol{b}_1)$, where the input is a d vector split into two separate parts $x(t) = [\boldsymbol{x}_{\boldsymbol{W}}(t); \boldsymbol{x}_{\boldsymbol{A}}(t)]$, of dimension $n$ and $D$ respectively. The similarity matrix we used to compute network-based measurements is then defined as the covariance of $\boldsymbol{h}(t)$ w.r.t $\boldsymbol{x}(t)$ with large $t = T$. For the simplicity of derivation, we assumed that the mean of $h$ is 0 and ignore the form of $\Theta$ for simplification. Then the covariance of h is approximately calculated as

$$COV\big(\boldsymbol{h}(t)\big) \approx \mathbb{E}_{\boldsymbol{x}}((\boldsymbol{W}_1\boldsymbol{x}_{\boldsymbol{W}}(t) + \boldsymbol{A}(t)\boldsymbol{x}_{\boldsymbol{A}}(t) + \boldsymbol{b}_1)(\boldsymbol{W}_1\boldsymbol{x}_{\boldsymbol{W}}(t) + \boldsymbol{A}(t)\boldsymbol{x}_{\boldsymbol{A}}(t) + \boldsymbol{b}_1)^T). \quad (S1)$$

Noticing that $x_A$ and $x_W$ are independent with each other with mean 0, we can then ignore the cross term and have

$$COV\big(\boldsymbol{h}(\boldsymbol{t})\big) \approx \mathbb{E}_x(\boldsymbol{W_1} \cdot (\boldsymbol{x_w}(t)\boldsymbol{x_W}(t)^T) \cdot \boldsymbol{W_1^T}) + \boldsymbol{A}(t) \cdot (\boldsymbol{x_A}(t)\boldsymbol{x_A}(t)^T) \cdot \boldsymbol{A}(t)^T + \boldsymbol{b}_1 \cdot \boldsymbol{b}_1^T). \quad (S2)$$

As $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ are independent with $\boldsymbol{x}$, and $\mathbb{E}_x((\boldsymbol{x_w}(t)\boldsymbol{x_W}(t)^T) = \boldsymbol{I}_{n\times n}$, we further have

$$COV\big(\boldsymbol{h}(\boldsymbol{t})\big) \approx \boldsymbol{W_1}\boldsymbol{W_1^T} + \boldsymbol{b_1}\boldsymbol{b_1^T} + \mathbb{E}_x\,(\boldsymbol{A}(t) \cdot (\boldsymbol{x_A}(t)\boldsymbol{x_A}(t)^T) \cdot \boldsymbol{A}(t)). \quad (S3)$$

Noticing that $\boldsymbol{A}(t) = \lambda\boldsymbol{A}(t-1) - \eta\boldsymbol{h}(t-1)\boldsymbol{x_A}(t-1)^T$, thus $\boldsymbol{A}(t)$ is also independent of $\boldsymbol{x}(t)$. We can then write the formula as

$$COV\big(\boldsymbol{h}(t)\big) = \boldsymbol{W}_1\boldsymbol{W}_1^T + \boldsymbol{b}_1\boldsymbol{b}_1^T + \mathbb{E}_x(\boldsymbol{A}(t) \cdot \boldsymbol{A}(t)^T), \quad (S4)$$

and further decompose the remaining term as

$\mathbb{E}_x\,(\boldsymbol{A}(t) \cdot \boldsymbol{A}(t)^T) = \mathbb{E}_x\,(\lambda^2\boldsymbol{A}(t-1)\boldsymbol{A}(t-1)^T + \eta^2\boldsymbol{h}(t-1)\boldsymbol{x_A}(t-1)^{\boldsymbol{T}}\boldsymbol{x_A}(t-1)\boldsymbol{h}(t-1)^T$.

The cross-terms are cancelled due to the independence between $x(t)$ and $x(t-1)$. Further, as $\boldsymbol{x_A}(t-1)^T\boldsymbol{x_A}(t-1) = D$, we actually have a form of

$$\mathbb{E}_x\,(\boldsymbol{A}(t) \cdot \boldsymbol{A}(t)^T) = \lambda^2\mathbb{E}_x(\boldsymbol{A}(t-1) \cdot \boldsymbol{A}(t-1)^T) + D\eta^2 COV\big(\boldsymbol{h}(t-1)\big). \quad (S5)$$

When the size of test sample is very large, we may assume that $COV\big(\boldsymbol{h}(t)\big) \approx COV\big(\boldsymbol{h}(t-1)\big)$ and $\mathbb{E}_x\,(\boldsymbol{A}(t) \cdot \boldsymbol{A}(t)^T) \approx \mathbb{E}_x(\boldsymbol{A}(t-1) \cdot \boldsymbol{A}(t-1)^T)$. Combined with Eqn. (S4) and (S5), we can then obtain

$$COV(\boldsymbol{h}) = \frac{1-\lambda^2}{1-\lambda^2-D\eta^2}(\boldsymbol{W_1}\boldsymbol{W_1^T} + \boldsymbol{b_1}\boldsymbol{b_1^T}). \quad (S6)$$

Thus, as the training goes, the community structure is mainly determined by $(\boldsymbol{W_1}\boldsymbol{W_1^T} + \boldsymbol{b_1}\boldsymbol{b_1^T})$. When the addressing function $\boldsymbol{W_1}$displays modularized representation, the covariance of the hidden layer activation would also display modularized structure. On the other hand, as shown in the (*14*) , the capacity of the addressing function $\boldsymbol{W_1}$ can be reflected by its modular structure as well, which connects the model accuracy and the modularized structure of the similarity matrix.

If we write $\boldsymbol{b_1}$ as a same scalar in *Tyulmankov et al.2022* with $\boldsymbol{b_1} = \beta D \cdot \boldsymbol{1}$, we will have

$$COV(\boldsymbol{h}) = \frac{1-\lambda^2}{1-\lambda^2-D\eta^2}(\boldsymbol{W_1}\boldsymbol{W_1^T} + \beta^2 D^2 \boldsymbol{1_{N\times N}}) \quad (S7)$$

On the other hand, mathematically the accuracy is given by

$$Accuracy = P_{TP} + P_{TN}$$

where the TP and FP can be estimated following the similar derivation on page 21 in *Tyulmankov et al.2022.* Specially, notice the opposite definition in the setup between *Tyulmankov et al.2022* and ours on the "true" case where we define the new sample as 0 and existing case as 1 while *Tyulmankov et al.2022* defines in the opposite way in the derivation. We can approximately have

$$Accuracy \approx 1 + erfc\left(\frac{n+\beta}{\alpha_\lambda\sqrt{2}}\right) - erfc\left(\frac{n+\beta-\lambda^{R-1}}{\alpha_\lambda\sqrt{2}}\right) \approx 1 - \frac{2\lambda^{R-1}}{\sqrt{\pi}}\exp\left(-\frac{(n+\beta)^2}{2\alpha_\lambda^2}\right). \quad (S8)$$

where $n, \beta$ are defined the same as above and $\alpha_\lambda$ follows $\sqrt{\frac{f_{eff}D}{N(1-\lambda^2)}} = \alpha_\lambda D$ and $f_{eff}$ denotes the faction of stimuli reported as novel by the network.

Based on Equation (S7), when $\boldsymbol{W_1}\boldsymbol{W_1^T}$ is fixed with a modular structure, if $\beta$ increases, the $\beta^2 D^2 \boldsymbol{1_{N\times N}}$ will increase and the modularity will decrease as $\beta^2 D^2 \boldsymbol{1_{N\times N}}$ uniformly add weight to each element. Based on Equation (S8), when $\beta$ increases, $\frac{2\lambda^{R-1}}{\sqrt{\pi}}\exp\left(-\frac{(n+\beta)^2}{2\alpha_\lambda^2}\right)$ decreases thus $-\frac{2\lambda^{R-1}}{\sqrt{\pi}}\exp\left(-\frac{(n+\beta)^2}{2\alpha_\lambda^2}\right)$ increases, leading to an increased Accuracy. Therefore, for different trained models, the variation of $\beta$ can lead to opposite trend of changes in the modularity and accuracy. This explains for the negative correlation of between network modularity and model performances when the correlation is calculated over samples.

It also worth noting that $\lambda$ and $\eta$ may also change thus the overall weight may not vary monotonically with $\beta$, which may be the reason why the modularity mediates the effect of overall weight on the performance.

**Model setups for the Spiking Neural Network with DVS input.** In this experiment, we aim to investigate the changes in modularity of the weight matrix of recurrent connections in the hidden layer induced by training DVS classification tasks in spiking neural networks (SNNs). We employ a streamlined network structure comprising an encoder module, a hidden state module, and a linear classification head. The encoder module projects the image onto the hidden state dimension through fully connected layers, generating the spiking trains that enter the hidden state module. The hidden state module consists of 200 neurons, each maintaining a membrane potential state $u(t)$ and receiving both the current forward spike vector $a(t)$ and the previous

time step's output spike vector $s(t)$. The forward process in the hidden state module can be mathematically described as

$$\boldsymbol{u}(t) = \tau\, \boldsymbol{u}(t-1) + \boldsymbol{W}_f \cdot \boldsymbol{h}(t) + \boldsymbol{W} \cdot \boldsymbol{s}(t-1), \tag{S9}$$

$$\boldsymbol{s}(t) = \Theta(u(t) - V_{\text{th}}), \tag{S10}$$

where $\tau = 0.5$ is the leaky factor for SNN neurons, $\boldsymbol{W}_f$ is the forward weight matrix, $\boldsymbol{W}$ is the recurrent weight matrix, $\Theta(x)$ is a step function that outputs 1 when x exceeds 0, otherwise outputs 0, and $V_{\text{th}} = 1.0$ is the default threshold for neurons. Specially, the recurrent weight matrix $\boldsymbol{W} = \boldsymbol{S} \odot \boldsymbol{F}$, where $\boldsymbol{S}$ denotes the structural weight matrix (like dendritic strength), $\boldsymbol{F}$ is the functional weight matrix that is updated by the backpropagation algorithm, and $\odot$ means Hadamard product. The $\boldsymbol{S}$ matrix is updated via the Hebbian rule as

$$S_{ij} = \eta(f_i - \phi_i)\big(f_j - \phi_j\big), \tag{S11}$$

$$S_{ij} = \text{clip}\big(S_{ij}, 0.0, 1.0\big), \tag{S12}$$

where $\eta$ is the learning rate of the Hebbian rule, $f_i$ and $f_j$ are the spiking rates of neurons $i$ and $j$. Respectively, $\phi_i$ and $\phi_j$ are the threshold for the Hebbian rule that are equal to 0.25, and the clip function is to constrain $S_{ij}$ between 0 and 1. Finally, the output spike matrix $\boldsymbol{S}(t)$ passes through the classification head. The experiments were conducted using the DVS-Gesture dataset (*25*), which consists of 11 distinct actions performed by 29 subjects captured by an event camera. The dataset comprises a total of 1342 samples with a resolution of $128 \times 128$. We utilized Spikingjelly (*54*) for data preprocessing. Spikingjelly converts the event dataset to 10 spike image frames of size $2 \times 128 \times 128$ and automatically splits the dataset into 1176 training samples and 288 testing samples. The frames are reshaped into a 32768-dimensional feature representation and fed into the SNN model. We employ the Adam optimizer with a learning rate of 1e-3 to optimize the entire network. For the Hebbian rule, we use the learning rate of 1e-4. The training is conducted for 100 epochs with a batch size of 128. We compared the variations in modularity of the functional weight matrix $F$ under two conditions: updating the $S$ by the Hebbian rule or using a fixed $S$. We perform this comparison across 120 different random seed settings to ensure the robustness of the experimental results.

**REFERENCES AND NOTES**


1. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X. J. Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).

2. J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie, L. Shi, Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019).

3. C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, B. Kay, Opportunities for neuromorphic computing algorithms and applications. *Nat. Comput. Sci.* **2**, 10–19 (2022).

4. S. Ullman, Using neuroscience to develop artificial intelligence. *Science* **363**, 692–693 (2019).

5. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).

6. Y. Takagi, S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023), pp. 14453–14463.

7. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

8. C. Du, K. Fu, J. Li, H. He, Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10760–10777 (2023).

9. T. Ito, G. R. Yang, P. Laurent, D. H. Schultz, M. W. Cole, Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. *Nat. Commun.* **13**, 673 (2022).

10. G. R. Yang, M. Molano-Mazón, Towards the next generation of recurrent network models for cognitive neuroscience. *Curr. Opin. Neurobiol.* **70**, 182–192 (2021).

11. X.-J. Wang, Theory of the multiregional neocortex: Large-scale neural dynamics and distributed cognition. *Annu. Rev. Neurosci.* **45**, 533–560 (2022).

12. S. Xie, A. Kirillov, R. Girshick, K. He, "Exploring randomly wired neural networks for image recognition" in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019), pp. 1284–1293.

13. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712 (2023).

14. D. Tyulmankov, G. R. Yang, L. F. Abbott, Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron* **110**, 544–557.e8 (2022).

15. D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, S. T. Grafton, Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7641–7646 (2011).

16. U. Braun, A. Schäfer, H. Walter, S. Erk, N. Romanczuk-Seiferth, L. Haddad, J. I. Schweiger, O. Grimm, A. Heinz, H. Tost, A. Meyer-Lindenberg, D. S. Bassett, Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11678–11683 (2015).

17. U. Braun, A. Harneit, G. Pergola, T. Menara, A. Schäfer, R. F. Betzel, Z. Zang, J. I. Schweiger, X. Zhang, K. Schwarz, J. Chen, G. Blasi, A. Bertolino, D. Durstewitz, F. Pasqualetti, E. Schwarz, A. Meyer-Lindenberg, D. S. Bassett, H. Tost, Brain network dynamics during working memory are modulated by dopamine and diminished in schizophrenia. *Nat. Commun.* **12**, 3478 (2021).

18. Y. He, X. Liang, M. Chen, T. Tian, Y. Zeng, J. Liu, L. Hao, J. Xu, R. Chen, Y. Wang, J.-H. Gao, S. Tan, J. Taghia, Y. He, S. Tao, Q. Dong, S. Qin, Development of brain state dynamics involved in working memory. *Cereb. Cortex* **33**, 7076–7087 (2023).

19. S. Qin, S. Farashahi, D. Lipshutz, A. M. Sengupta, D. B. Chklovskii, C. Pehlevan, Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nat. Neurosci.* **26**, 339–349 (2023).

20. H. G. Rodriguez, Q. Guo, T. Moraitis, "Short-term plasticity neurons learning to learn and forget" in *Proceedings of the 39th International Conference on Machine Learning* (ACM, 2022), pp. 18704–18722.

21. O. Sporns, R. F. Betzel, Modular brain networks. *Annu. Rev. Psychol.* **67**, 613–640 (2016).

22. M. G. Preti, T. A. Bolton, D. Van De Ville, The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage* **160**, 41–54 (2017).

23. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328**, 876–878 (2010).

24. D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge Univ. Press, 2003).

25. A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, D. Modha, "A low power, fully event-based gesture recognition system" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), pp. 7243–7252.

26. P. Zurn, D. S. Bassett, Network architectures supporting learnability. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190323 (2020).

27. S.-J. Wang, C. Hilgetag, C. Zhou, Sustained activity in hierarchical modular neural networks: Self-organized criticality and oscillations. *Front. Comput. Neurosci.* **5**, 30 (2011).

28. M. Kaiser, C. Hilgetag, Optimal hierarchical modular topologies for producing limited sustained activation of neural networks. *Front. Neuroinform.* **4**, 8 (2010).

29. N. Kashtan, U. Alon, Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci.* **102**, 13773–13778 (2005).

30. J. Clune, J.-B. Mouret, H. Lipson, The evolutionary origins of modularity. *Proc. Biol. Sci.* **280**, 20122863 (2013).

31. J. Achterberg, D. Akarca, D. J. Strouse, J. Duncan, D. E. Astle, Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nat. Mach. Intell.* **8**, 1369–1381 (2023).

32. J. Tanner, S. Mansour L, L. Coletta, A. Gozzi, R. F. Betzel, Functional connectivity modules in recurrent neural networks: Function, origin and dynamics. arXiv:2310.20601 (2023).

33. D. S. Bassett, M. Yang, N. F. Wymbs, S. T. Grafton, Learning-induced autonomy of sensorimotor systems. *Nat. Neurosci.* **18**, 744–751 (2015).

34. G. Tononi, O. Sporns, G. M. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci.* **91**, 5033–5037 (1994).

35. R. Wang, M. Liu, X. Cheng, Y. Wu, A. Hildebrandt, C. Zhou, Segregation, integration, and balance of large-scale resting brain networks configure different cognitive abilities. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022288118 (2021).

36. K. O. Ellefsen, J.-B. Mouret, J. Clune, Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS Comput. Biol.* **11**, e1004128 (2015).

37. K. Finc, K. Bonna, X. He, D. M. Lydon-Staley, S. Kühn, W. Duch, D. S. Bassett, Dynamic reconfiguration of functional brain networks during working memory training. *Nat. Commun.* **11**, 2435 (2020).

38. G. L. Baum, R. Ciric, D. R. Roalf, R. F. Betzel, T. M. Moore, R. T. Shinohara, A. E. Kahn, S. N. Vandekar, P. E. Rupert, M. Quarmley, P. A. Cook, M. A. Elliott, K. Ruparel, R. E. Gur, R. C. Gur, D. S. Bassett, T. D. Satterthwaite, Modular segregation of structural brain networks supports the development of executive function in youth. *Curr. Biol.* **27**, 1561–1572.e8 (2017).

39. N. Rodriguez, E. Izquierdo, Y.-Y. Ahn, Optimal modularity and memory capacity of neural reservoirs. *Netw. Neurosci.* **3**, 551–566 (2019).

40. J. R. Cohen, M. D'Esposito, The segregation and integration of distinct brain networks and their relationship to cognition. *J. Neurosci.* **36**, 12083–12094 (2016).

41. S. Gu, P. Fotiadis, L. Parkes, C. H. Xia, R. C. Gur, R. E. Gur, D. R. Roalf, T. D. Satterthwaite, D. S. Bassett, Network controllability mediates the relationship between rigid structure and flexible dynamics. *Netw. Neurosci.* **6**, 275–297 (2022).

42. E. H. Simpson, The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Series B.* **13**, 238–241 (1951).

43. J. D. Medaglia, M.-E. Lynall, D. S. Bassett, Cognitive network neuroscience. *J. Cogn. Neurosci.* **27**, 1471–1491 (2015).

44. J. M. Shine, Neuromodulatory influences on integration and segregation in the brain. *Trends Cogn. Sci.* **23**, 572–583 (2019).

45. Z. Liu, O. Kitouni, N. S. Nolte, E. Michaud, M. Tegmark, M. Williams, Towards understanding grokking: An effective theory of representation learning. *Adv. Neural Inf. Process. Syst.* **35**, 34651–34663 (2022).

46. E. Bullmore, O. Sporns, Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198 (2009).

47. G. Deco, M. L. Kringelbach, Great expectations: Using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron* **84**, 892–905 (2014).

48. R. Liégeois, J. Li, R. Kong, C. Orban, D. Van De Ville, T. Ge, M. R. Sabuncu, B. T. T. Yeo, Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nat. Commun.* **10**, 2317 (2019).

49. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

50. C. Aicher, A. Z. Jacobs, A. Clauset, Learning latent block structure in weighted networks. *J. Complex Netw.* **3**, 221–248 (2015).

51. A. Nematzadeh, E. Ferrara, A. Flammini, Y.-Y. Ahn, Optimal network modularity for information diffusion. *Phys. Rev. Lett.* **113**, 088701 (2014).

52. S. Gu, C. H. Xia, R. Ciric, T. M. Moore, R. C. Gur, R. E. Gur, T. D. Satterthwaite, D. S. Bassett, Unifying the notions of modularity and core–periphery structure in functional brain networks during youth. *Cereb. Cortex* **30**, 1087–1102 (2020).

53. J. Geweke, Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.* **77**, 304–313 (1982).

54. Wei Fang, Y. Chen, J. Ding, Z. Yu, T. Masquelier, D. Chen, L. Huang, H. Zhou, G. Li, Y. Tian SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Sci. Adv.* **9**, eadi1480 (2023).