# Growing Brains in Recurrent Neural Networks for Multiple Cognitive Tasks

**Ziming Liu**[12*], **Mikail Khona**[1*], **Ila R. Fiete**[1], **Max Tegmark**[12]
[1] **MIT** [2] **IAIFI**
{zmliu, mikail, fiete, tegmark}@mit.edu

## Abstract

Recurrent neural networks (RNNs) trained on a diverse ensemble of cognitive tasks, as described by Yang et al. (2019); Khona et al. (2023), have been shown to exhibit functional modularity, where neurons organize into discrete functional clusters, each specialized for specific shared computational subtasks. However, these RNNs do not demonstrate *anatomical modularity*, where these functionally specialized clusters also have a distinct spatial organization. This contrasts with the human brain which has both functional and anatomical modularity. Is there a way to train RNNs to make them more like brains in this regard? We apply a recent machine learning method, brain-inspired modular training (BIMT), to encourage neural connectivity to be local in space. Consequently, hidden neuron organization of the RNN forms spatial structures reminiscent of those of the brain: spatial clusters which correspond to functional clusters. Compared to standard $L_1$ regularization and absence of regularization, BIMT exhibits superior performance by optimally balancing between task performance and sparsity. This balance is quantified both in terms of the number of active neurons and the cumulative wiring length. In addition to achieving brain-like organization in RNNs, our findings also suggest that BIMT holds promise for applications in neuromorphic computing and enhancing the interpretability of neural network architectures.

**Keywords:** Cognitive Neuroscience, Recurrent Neural Networks, Interpretability, Brain-Inspired Machine Learning

## 1. Introduction

Brain-inspired modular training (BIMT) is a recent method for making artificial neural networks modular and interpretable (Liu et al., 2023). The key idea of BIMT is to encourage local neural connections via two optimization terms: distance-dependent weight regularization and discrete neuron swapping. Here we ask whether we can use BIMT, which was inspired by the brain, to also answer a fundamental question about neuroscience: Can spatial constraints and wiring costs (Chen et al., 2006; Chklovskii et al., 2002; Chklovskii and Koulakov, 2004) lead to the emergence of anatomical modules which are also functionally distinct? The brain is modular with strong indications of gross localization of function (Ferrier, 1874), for example, visual processing corresponding to core object recognition is localized to the brain regions of the ventral visual stream while voluntary motor control is confined to a few motor and premotor cortical regions.

We study how BIMT can lead to the emergence of spatial modules in a multitask learning setting relevant to cognitive systems neuroscience with a commonly used set of

[*] Equal contribution

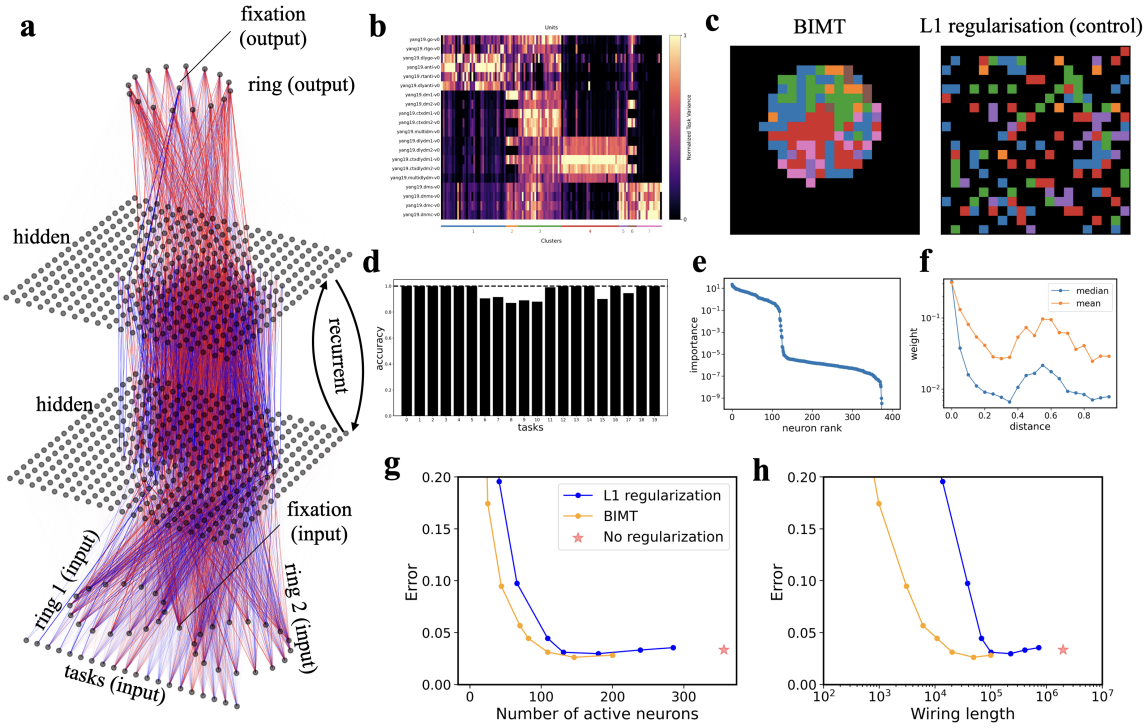


Figure 1: Training an RNN with BIMT on cognitive tasks. **a**: Visualization of the network. Each line represents a weight; blue/red means positive/negative weights; thickness corresponds to magnitudes. **b**: The hidden neurons are clustered into functional modules. **c**: These functional modules (distinguished by colors) are also clustered in space, visually resembling a brain. By contrast, $L_1$ regularization leads to no anatomical modules. The network has (**d**) good performance, (**e**) high sparsity and good locality (**f**). Trade-off between performance (error) and sparsity. (**g**): using the number of active neurons as the sparsity measure. (**h**): using the wiring length as the sparsity measure. In both cases, the Pareto frontier of BIMT is better than that of $L_1$ regularization and no regularization.

tasks, `20-Cog-tasks` (Yang et al., 2019). We trained recurrent neural networks (RNNs) with BIMT in the supervised setup. We observe brain-like organization emerging in the hidden layer of RNN, i.e., neurons that are functionally similar are also localized in space (see Figure 1c). Such locality and sparsity are gained with no sacrifice in performance, or even sometimes slightly improve the performance. We introduce methods in Section 2 and include results in Section 3.

## 2. Method

**RNN architecture** We take a simple recurrent neural network (RNN) relevant to systems neuroscience, which is defined by

$$\begin{aligned} \mathbf{h}_{t+1} &= \phi(\mathbf{W}\mathbf{h}_t + \mathbf{W}_{\text{in}}\mathbf{u}_t + \mathbf{b}^h), \\ \mathbf{o}_{t+1} &= \mathbf{W}_{\text{out}}\mathbf{h}_{t+1} + \mathbf{b}^o, \end{aligned} \tag{1}$$

where $\mathbf{u}_t \in \mathbb{R}^{n_u}$, $\mathbf{h}_t \in \mathbb{R}^{n_h \times n_h}$, $\mathbf{o}_t \in \mathbb{R}^{n_o}$, $\mathbf{W} \in \mathbb{R}^{n_h \times n_h} \times \mathbb{R}^{n_h \times n_h}$. We place hidden neurons uniformly on a 2D grid $[0,1]^2$ (see Figure 1a), so the $ij$ neuron (the neuron in the $i^{\text{th}}$ row and the $j^{\text{th}}$ column) is located at $(i/n_h, j/n_h)$. The distance between the $ij$ neuron and the $mn$ neuron is thus $\mathbf{D}_{ij,mn} \equiv (|i-m| + |j-n|)/n_h$. We define the RNN's connection cost as

$$\ell_{cc} = \underbrace{\|\mathbf{W}\|_1 + \|\mathbf{W}_{\text{in}}\|_1 + \|\mathbf{W}_{\text{out}}\|_1 + |\mathbf{b}^h|_1 + |\mathbf{b}^o|_1}_{\text{vanilla } L_1 \text{ regularization}} + \underbrace{A\|\mathbf{D} \odot \mathbf{W}\|_1}_{\text{distance-aware regularization}}, \tag{2}$$

where $\|\mathbf{M}\|_1 \equiv \sum_{ij} |M_{ij}|$ and $|\mathbf{v}|_1 \equiv \sum_i |v_i|$ are matrix and vector $L_1$-norm, respectively. $A$ is a hyper-parameter controling the strength of locality constraint.

**Cognitive tasks** The `20-Cog-tasks` are a set of simple cognitive tasks inspired by experiments with rodents and non-human primates performed by systems neuroscientists Yang et al. (2019). These tasks are designed to fall into families where each family is defined by a set of computations drawn from a common pool of computational primitives. Thus, the tasks have shared subtasks and an optimal solution is to form clusters of neurons specialized to these subtasks and share them across tasks, illustrated in Figure 2a. The prediction loss $\ell_{\text{pred}}$ is the cross-entropy between the ground truth and the predicted reaction.

**BIMT** loss simply combines the prediction loss and the connection cost, i.e., the total loss function is

$$\ell = \ell_{\text{pred}} + \lambda \ell_{cc}, \tag{3}$$

where $\lambda \geq 0$ is the strength of penalizing connection costs. When $\lambda = 0$, it boils down to train a fully-connected RNN without sparsity constraint; when $\lambda > 0$ but $A = 0$, it boils down to train with vanilla $L_1$ regularization. Besides adding connection costs as regularization, BIMT allows swapping neurons to further reduce $\ell_{cc}$ by avoiding local minima, e.g., if a neural network is initialized to be performing well but has non-local connections, without swapping, the network would not change much and maintain those non-local connections.

## 3. Results

**BIMT learns a 2D brain that solves all `20-Cog-tasks`**

We show results for $A = 1.0$ and $\lambda = 10^{-5}$ in Figure 1. **a** shows the connectivity graph of the RNN. All the weights are plotted as lines whose thicknesses are proportional to their magnitudes [1], and blue/red means positive/negative weights. BIMT learns to prune away peripheral neurons and concentrate important neurons only in the middle. Following Yang et al. (2019), we cluster neurons into functional modules based on their normalized task variance (shown in **b**). These functional modules, colored by different colors, are shown to be also anatomically modular, i.e., spatially local (shown in **c**), visually resembling a brain. By contrast, $L_1$ regularization does not induce any anatomical module. **d** shows that the network performs reasonably well. The network is sparse; **e** shows that it only contains around 100 important neurons (measured by sum of task variances). The connections in the hidden layer are mostly local, as we hoped; **f** shows that weights decay fast as distance increases, which is similar to fixed local-masked RNN (Khona et al., 2023). However intriguingly, there is a second peak of (relatively) strong connections around distances being 0.6, which is probably

1. Some weights are visually vanishing, because their magnitudes are too small, although we indeed plot them.

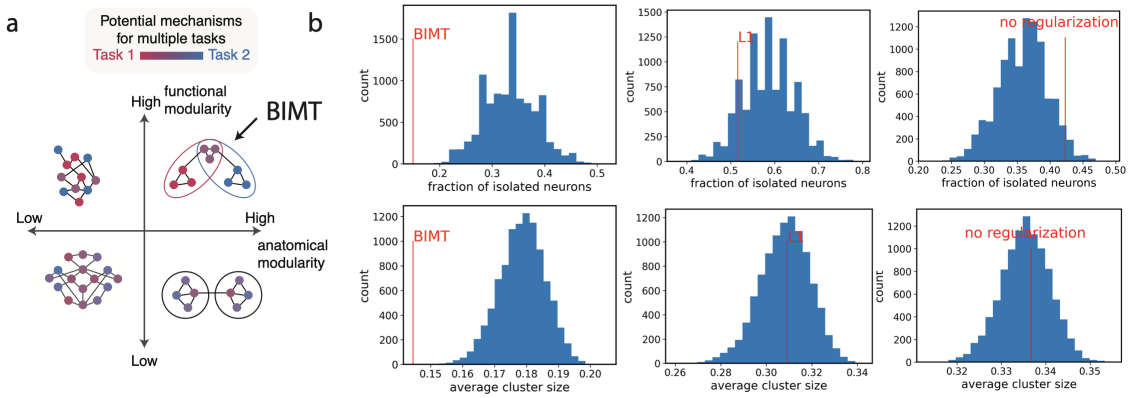


Figure 2: (a) The myriad ways anatomical and functional modularity can present itself in trained RNNs. (b) We test anatomical modularity for neural networks with BIMT (left), L1 regularization (middle) or no regularization (right). We propose two metrics, fraction of isolated neurons (top) and average (functional) cluster size (bottom) to measure anatomical modularity. For both metrics, smaller is better. We compare the trained network with networks whose useful hidden neurons are randomly shuffled. No regularization and L1 regularization are in the distributions of randomly shuffled networks, while BIMT is significantly out of distribution (smaller) than random ones, indicating anatomical modularity.

attributed to inter-module connections.

**Sparsity vs Accuracy Tradeoff** There is a Pareto frontier showing the trade-off between sparsity and accuracy, shown in Figure 1g,h. We also compute the trade-off for networks with vanilla $L_1$ regularization. We use two sparsity measures: the number of active neurons, and the wiring length [2]. Under both measures, BIMT is superior than $L_1$ regularization in terms of having better Pareto frontier.

**Anatomical and functional modularity correspondance** Anatomical and functional modularity correspondance means that neurons with similar functions are placed close to each other in space. Because neurons of fully-connected layers have permutation symmetries, there is no incentive for them to develop anatomical modularity. By contrast, since BIMT penalizes connection costs, BIMT networks potentially have anatomical modularity. In Figure 1**c**, each neuron's functional cluster is marked and there are clear spatial clusters in which all neurons belong to the same functional cluster. Quantitatively, we propose two metrics: (1) the fraction of isolated neurons. A neuron is isolated if none of its (eight) neighbors belongs to the same functional cluster. (2) the average size of functional clusters. For both metrics, the smaller the better. For baselines, we randomly shuffle important hidden neurons [3]. We compute the two metrics for networks trained with BIMT, $L_1$ regularization, or no regularization in Figure 2. Only BIMT networks are seen to be significantly out-of-distribution from baselines, implying anatomical modularity. In future work, we hope to explore improvements of BIMT to further increase the functional modularity of the "brain" seen in Figure 1**c**.

2. sum of lengths of all active connections; a connection is active if its weight magnitude is larger than $10^{-2}$
3. A neuron is important if the sum of its task variances is above $10^{-3}$.

## References

Beth L Chen, David H Hall, and Dmitri B Chklovskii. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences*, 103 (12):4723–4728, 2006.

Dmitri B Chklovskii and Alexei A Koulakov. Maps in the brain: what can we learn from them? *Annu. Rev. Neurosci.*, 27:369–392, 2004.

Dmitri B Chklovskii, Thomas Schikorski, and Charles F Stevens. Wiring optimization in cortical circuits. *Neuron*, 34(3):341–347, 2002.

David Ferrier. The localization of function in the brain. *Proceedings of the Royal Society of London*, 22(148-155):228–232, 1874.

Mikail Khona, Sarthak Chandra, Joy J Ma, and Ila R Fiete. Winning the lottery with neural connectivity constraints: Faster learning across cognitive tasks with spatially constrained sparse rnns. *Neural Computation*, pages 1–20, 2023.

Ziming Liu, Eric Gan, and Max Tegmark. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *arXiv preprint arXiv:2305.08746*, 2023.

Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306, 2019.