Brain-Like Pathways Form in Models With Heterogeneous Experts

Jack Cook[1] Danyal Akarca[2] Rui Ponte Costa[1,*] Jascha Achterberg[1,*]
[1]Centre for Neural Circuits and Behaviour, University of Oxford
[2]Department of Electrical and Electronic Engineering, Imperial College London
[*]Joint senior authors

This work is also described in Cook et al. (2025)

Abstract

The brain is made up of a vast set of heterogeneous brain regions that organize themselves into sub-networks and processing pathways to respond to task demands. Examples of such pathways can be seen in the ventral and dorsal visual streams, the Multiple-Demand Network during task execution, and the interaction between cortical and subcortical networks during learning. In this work we ask how do these processing pathways develop from a set of heterogeneous brain regions. Do regions automatically group into systems or are additional priors required? We study this by using neural networks, specifically by extending the Mixture-of-Expert architecture. We show that heterogeneous regions do not automatically form processing pathways by themselves. Training with a processing-complexity routing cost, when scaled based on task performance, results in the development of replicable processing pathways. When comparing our model to the brain, we observe that these pathways match how the brain utilizes different systems to learn and execute tasks of varying task complexities. Our findings establish specific biases that may underlie the formation of processing pathways observed in the brain. At the same time, our model allows us to conduct fine-grained analyses of how sets of pathways interact during problem solving across domains of neuroscience.

Keywords: Processing pathways; Cortical; Subcortical; Mixture-of-Experts; Modularity; Neural Network

Introduction

The brain is made up of heterogeneous regions that achieve complex functions by working together as pathways or systems. This organizational principle manifests across sensory systems (Grill-Spector & Malach, 2004), cognitive networks (Duncan, 2025), and emotion-related circuits (Etkin, Egner, & Kalisch, 2011). Using a static and vision-specific neural networks, Finzi, Margalit, Kay, Yamins, and Grill-Spector (2023) found spatial constraints as a cause of pathways in visual systems. Here we ask the larger question of when do pathways develop from heterogeneous regions and how does a network learn to combine its pathways, in a domain-general network architecture allowing for dynamic recombination of modules.

Model and task setup

Our model setup builds on the idea of a Heterogeneous Mixture-of-Experts model (HMoE) proposed for large-scale AI (Jawahar et al., 2022). In HMoE, a network is built based on layers containing multiple expert networks with heterogeneous characteristics like varying expert sizes. A layer-specific routing network decides which experts of a given layer should process the information of a given timestep. We introduce a crucial change to HMoE: Where experts are usually feedforward networks in standard MoEs, we substitute them with Gated Recurrent Units (GRUs), to have recurrent processing in the experts. In our network (Fig.1A) each layer has three experts of different complexities: complex (32 unit GRU), simple (16 unit GRU), skip (no processing / identify function; similar to Raposo et al. (2024)). We train models in a supervised fashion to solve 82 cognitive tasks of the ModCog task set with >90% accuracy (Khona*, Chandra*, Ma, & Fiete, 2022) (extension of NeuroGym tasks). The set includes tasks such as multi-stimuli integration and GoNogo tasks (Fig.1B).

Do pathways develop?

First we study whether pathways develop in a HMoE architecture by themselves. For brevity we focus on testing for the existence of pathways by using the single criteria of pathway-replicability: if pathways develop according to task demands, then networks from different training runs should rely on roughly similar pathways across the same set of tasks.

Here we test for this by measuring how 'complex' the pathways are that are used to solve different tasks. We then correlate this vector of pathway complexities (length = number of tasks) across training runs. Pathway complexity is measured as the sum of squared expert networks sizes used across layers (complex expert = 32; simple expert = 16; skip = 0). As the router of each layer outputs a weighting of the three experts of each layer, we also weight this cost by the routing weight. Fig.1B shows that the baseline model does not form replicable pathways developing, as shown by varying correlations across training runs. We now introduce modifications to the model optimization, namely (1) a routing cost to incentivize the router to prioritize less complex experts and (2) a loss normalization to normalize the routing cost from (1) by the current overall model performance. Fig.1B shows that these modifications result in replicable pathways across training runs. Following results are averaged across these training runs.

How do pathways match the brain?

Now that we have a model which develops replicable pathways, we want to compare how these pathways compare to processing pathways that we observe in the brain. The first comparison we want to focus on is to the **Multiple-Demand (MD) Network**, a large network of cortical regions that acti-

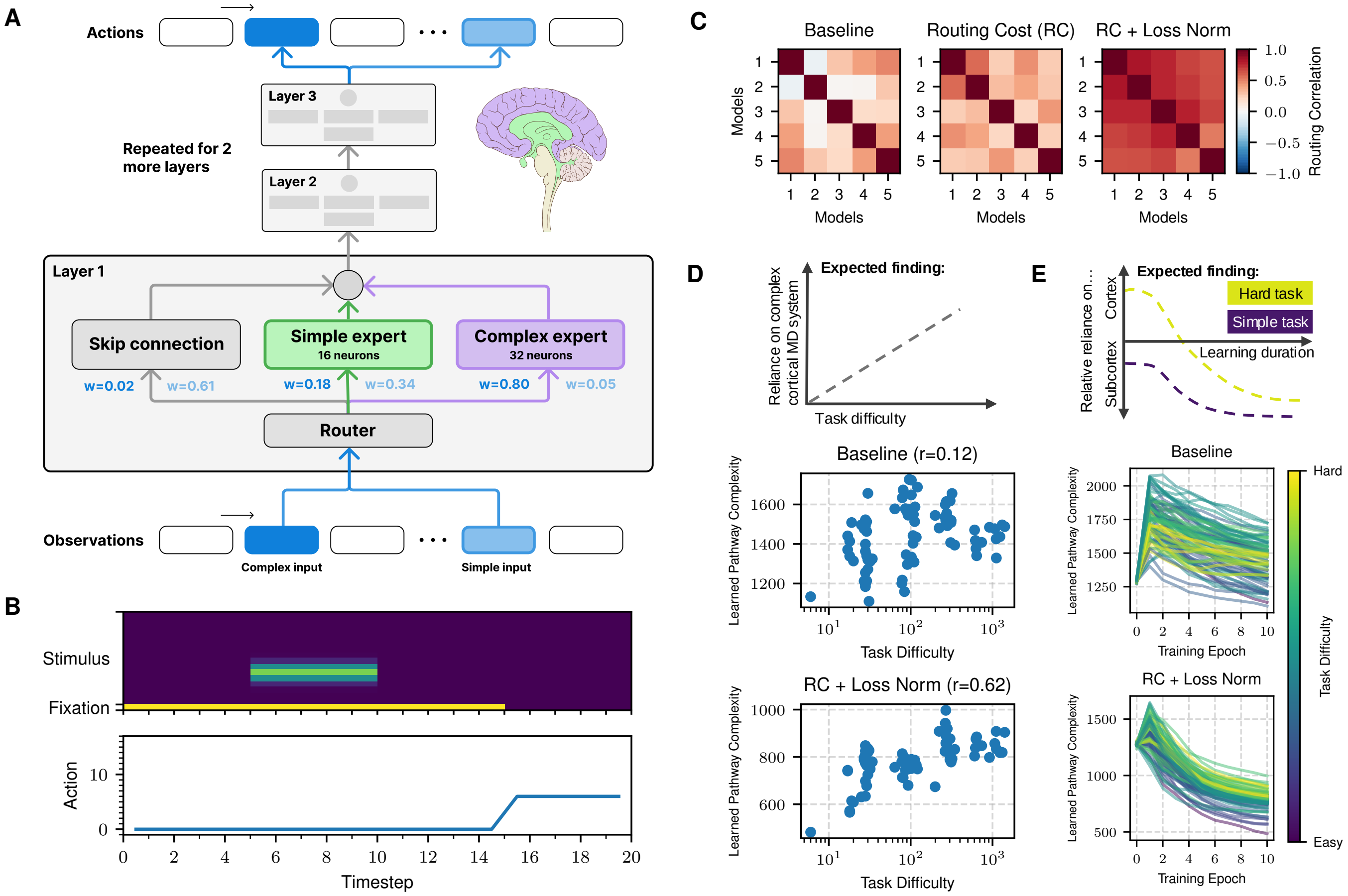

Figure 1: **A** Baseline model architecture. **B** Go trial of the Go-NoGo task. Model observes time series and if a stimuli appears the model has to respond with ’Go’ during response period. **C** Stability of pathways across training runs (within matrix) and across model modification (across matrices). **D** Model shows MD-system-like processing of task of differing task complexity (each dot is one task). Results averaged across 5 training runs. **E** Model mirrors cortical-subcortical trade-off when learning tasks of different complexities (each line is one task, colored by complexity). Results averaged across 5 training runs.

vate in response to task complexity (Duncan, 2025). We want to test whether our model also develops distinct pathways related to task complexity. For this we test whether the pathway complexity of the pathway used to solve a given task is predicted by the task’s complexity (Fig.1D top). The latter is measured by how many learning steps a separate GRU network needs to learn the task. We find that the baseline network does only develop a weak MD-like pattern (Fig.1D middle), whereas the network with the additional modifications does show the expected pattern (Fig.1D bottom).

The second finding that we want to compare to can be observed **during learning across cortical and subcortical pathways** (Hong, Lacefield, Rodgers, & Bruno, 2018; Peters, Fabre, Steinmetz, Harris, & Carandini, 2021; Wolff, Ko, & Ölveczky, 2022). Here we observe that for simple tasks complex cortical pathways are not required but simpler subcortical pathways are sufficient to learn and execute a skill. In more complex task settings however, we see that cortical pathways are active during learning and that skills can then passed down to simpler cortical pathways over time for execution. We test whether our model shows the same behavior by plotting the task complexity used to solve each cognitive task over learning (Fig.1D). We find that our model behaves similarly to the brain, where complex tasks are initially pushed upwards to complex pathways during learning, before then being increasingly passed downwards to simpler pathways. Simple tasks on the other hand, are continuously passed down and do not rely on complex pathways for learning. This behavior is not present in the baseline model.

New possible investigations

While priors and conditions causing modularity in networks have been studied (Achterberg, Akarca, Strouse, Duncan, & Astle, 2023; Yang, Joglekar, Song, Newsome, & Wang, 2019), prior network architectures did not allow us to study how heterogeneous modules dynamically organize into sub-networks to perform functions. The dynamic and time-step dependent routing of our model allows us to observe how different pathways of partially overlapping regions interact to solve tasks. For example, we can observe how networks gradually rely on more complex pathways over the task duration as (and if) more complex stimuli appear (not depicted in this report). As a result, the architecture presented here gives us a new way of understanding how a large distributed network combines its sub-networks to achieve cognition.

Acknowledgments

This research is supported by the BBSRC (BB/X013340/1), the EPSRC (EP/X029336/1), and an ERC-UKRI Frontier Research Guarantee Starting Grant (EP/Y027841/1) awarded to R.P.C. J.C.'s work is supported by a Rhodes Scholarship. We additionally thank Modal for compute credits.

References

Achterberg, J., Akarca, D., Strouse, D., Duncan, J., & Astle, D. E. (2023). Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nature Machine Intelligence*, *5*(12), 1369–1381.

Cook, J., Akarca, D., Ponte Costa, R., & Achterberg, J. (2025). *Brain-like processing pathways form in models with heterogeneous experts.* Retrieved from `https://arxiv.org/abs/2506.02813`

Duncan, J. (2025). Construction and use of mental models: Organizing principles for the science of brain and mind. *Neuropsychologia*, *207*, 109062.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in cognitive sciences*, *15*(2), 85–93.

Finzi, D., Margalit, E., Kay, K., Yamins, D. L., & Grill-Spector, K. (2023). A single computational objective drives specialization of streams in visual cortex. *bioRxiv*, 2023–12.

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.*, *27*(1), 649–677.

Hong, Y. K., Lacefield, C. O., Rodgers, C. C., & Bruno, R. M. (2018). Sensation, movement and learning in the absence of barrel cortex. *Nature*, *561*(7724), 542–546.

Jawahar, G., Mukherjee, S., Liu, X., Kim, Y. J., Abdul-Mageed, M., Lakshmanan, L. V., . . . Gao, J. (2022). Automoe: Heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. *arXiv preprint arXiv:2210.07535*.

Khona*, M., Chandra*, S., Ma, J. J., & Fiete, I. (2022). Winning the lottery with neurobiology: faster learning on many cognitive tasks with fixed sparse rnns. *arXiv*. Retrieved from `[https://arxiv.org/abs/2207.03523]`

Peters, A. J., Fabre, J. M., Steinmetz, N. A., Harris, K. D., & Carandini, M. (2021). Striatal activity topographically reflects cortical activity. *Nature*, *591*(7850), 420–425.

Raposo, D., Ritter, S., Richards, B., Lillicrap, T., Humphreys, P. C., & Santoro, A. (2024). Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*.

Wolff, S. B., Ko, R., & Ölveczky, B. P. (2022). Distinct roles for motor cortical and thalamic inputs to striatum during motor skill learning and execution. *Science advances*, *8*(8), eabk0231.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, *22*(2), 297–306.