The Helmholtz Hippocampus: A biologically plausible generative model of Hippocampus

Summary Generative models have recently revolutionised machine learning and long been thought fundamental to biological intelligence. In animals, data suggests the hippocampal formation learns and uses a generative model to support its role in spatial and non-spatial memory. Here we introduce a biologically plausible model of the hippocampal formation tantamount to a Helmholtz machine that we apply to a continuous stream of inputs. Fast theta-band oscillations (5-10 Hz) gate the direction of information flow through the network, training it akin to a high-frequency wake-sleep algorithm. Our model can accurately infer the latent state from sensory stimuli and generate realistic sensory predictions offline. Trained on a navigation task it learns to path integrate by developing a ring attractor and can flexibly transfer this structure between environments matching previous theoretical but biologically implausible proposals. Whereas many models trade-off biological plausibility with generality, our model captures a variety of hippocampal cognitive functions under one simple and local learning rule.

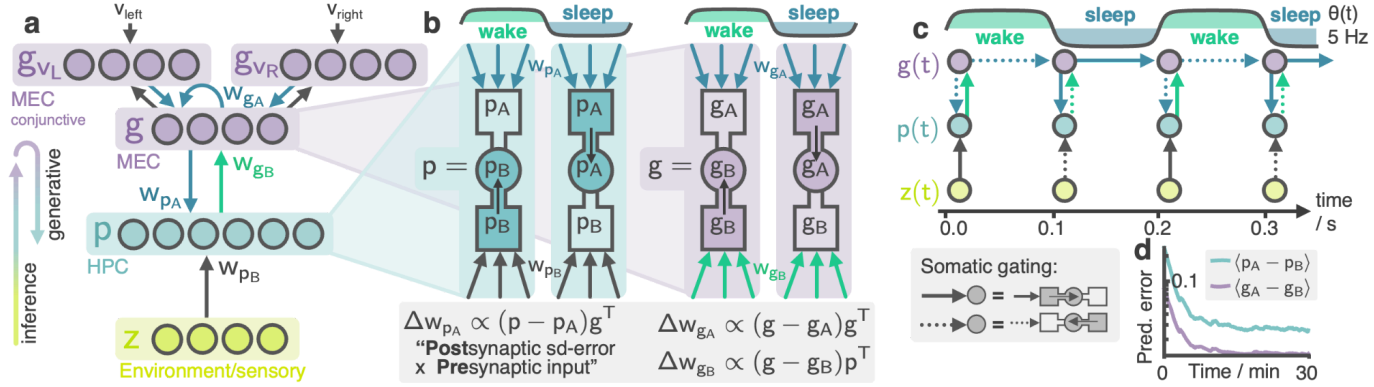

Figure 1: *The Helmholtz Hippocampus* ***(a)*** *During inference, sensory information z(t) flows up through HPC to MEC. Concurrently the generative model combines velocity signals with the hidden MEC state and returns these predictions to HPC for comparison to real sensory data (during training) or use in planning/consolidation.* ***(b)*** *HPC and MEC neurons are multicompartmental. Theta phase (5 Hz) controls which compartment (basal, B, or apical, A) drives the soma while plasticity continuously adjusts synaptic weights to minimise the local somatic-dendritic discrepancy (sd-error). Neurons are rate-based but learning rules naturally extend to spikes[6].* ***(c)*** *During early theta phases basal somatic drive means the network is in a "wake" state where the inference model is sampled and the generative model is trained. During late theta the network is in a "sleep" state (generative sampled, inference trained).* ***(d)*** *Dendritic discrepancies decrease over time indicating successful training of the inference and generative models.*

Model details Theory and experiment indicate the Hippocampal formation concurrently *infers* self-location[1] and *generates* trajectories offline for planning or consolidation[2]. Models accounting for these dual capabilities, where existing[3], lean heavily on unrealistic assumptions e.g. non-local learning rules, leaving unclear what mechanisms support these functions in the brain.

To resolve this we propose a novel framework: The Helmholtz Hippocampus (Fig. 1a). A hierarchical network receives a continuous stream of sensory information from the environment, z(t). Inference occurs as the high-dimensional and/or noisy stimuli pass through hippocampus (HPC, p(t)) to medial entorhinal cortex (MEC, g(t)) atop the hierarchy. Matching physiology[4] MEC also receives inputs tuned to the agent's velocity. Being recurrent, MEC can function independently and integrate velocity signals into the hidden state "generating" predictions which are sent down to HPC. In both HPC and MEC, bottom-up (inference) and top-down (generative) signals arrive at distinct basal and apical dendritic compartments respectively (Fig. 1b).

A 5 Hz theta oscillation controls whether the model is in "inference" or "generative" mode (Fig. 1b) by gating somatic drive[5]. In the first half of each theta cycle, basal dendrites drive the soma; inference occurs as information flows from the environment to MEC. In the second half of theta, flow reverses and the generative mode takes over as MEC integrates velocity inputs and sends predictions to HPC. Learning is simple: randomly initialised synaptic strengths are continuously adjusted by a local Hebbian-style rule minimising the somatodendritic voltage discrepancy. This causes dendrites to "predict" the soma[6] and hence, in alternating phases, generative predictions approach sensory inferences and vice versa. Learning converges after ~10 mins (Fig. 1d).

Mathematically we show this is equivalent to a well studied class of generative models called Helmholtz machines[7] (HMs). HMs are trained by switching between "wake" and "sleep" phases which alternately sample the inference and generative models. A powerful intuition emerges: early/late theta phases are analogous to wake/sleep cycles as understood in machine learning. This normative framework, our results show, accounts for many aspects of hippocampal function.

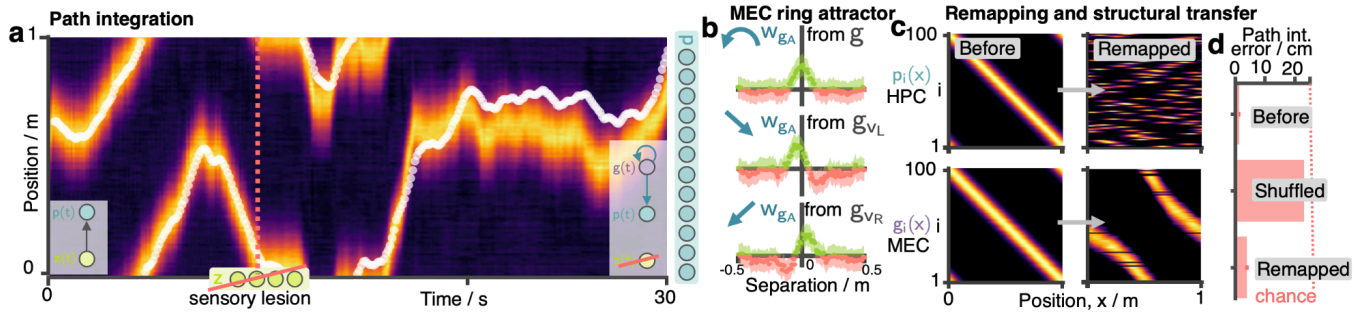

Figure 2: *Path integration and transfer learning.* ***(a)*** *A lesion experiment demonstrates learned path integration: After training, sensory input was removed. HPC activity, driven only by top down MEC input, continued to track the agent's position.* ***(b)*** *MEC contains a ring attractor: Plotting the average synaptic weight (based on the distance between MEC receptive fields) reveals a ring attractor connectivity structure with symmetric centre-surround recurrent weights and oppositely asymmetric weights from the left/right velocity inputs.* ***(c)*** *A sensory shuffle demonstrates structure transfer between environments and realistic remapping. After shuffling sensory inputs the system was retrained with* w_{gA} *fixed. MEC receptive fields reorganised with a constant phase shift reminiscent of biological remapping.* ***(d)*** *Path integration abilities recovered despite no learning to* w_{gA} *demonstrating that MEC transfers structure between environments.*

Experimental results We applied our model to a navigation task where an agent moved around a 1D loop track and Gaussian position-tuned inputs were fed to HPC. After learning, in a lesion experiment, sensory inputs were removed and the model was placed in "generative" mode. Despite the lesion HPC activity was minimally perturbed and a persistent neural signal continued to track the agent's position for a further 20 seconds or more (Fig. 2a). We hypothesised that a self-stabilising velocity-integrating attractor in MEC drove persistent activity in post-lesion HPC. Calibrated path integration, a core theoretical proposal for MEC-HPC function[8], is shown here to be learned in a biologically plausible model without non-local rules, a first to our knowledge.

To understand this further we calculated the average interneuron synaptic strength in MEC as a function of how far apart the receptive fields were (Fig. 2b). MEC cells with nearby receptive fields were strongly recurrently connected, those far-apart weakly inhibited one another. Synapses from the velocity inputs to MEC were asymmetric. This connectivity can be readily interpreted as that of a quasi-continuous ring attractor[9] where top-down velocity inputs are tuned to push a self-stabilised bump along a manifold proportional to the agent's velocity. Note that neither path integration nor a ring attractor – though predicted by previous theoretical models – was the target of a global learning rule (e.g. unlike [8]) but emerged assumption-free from our local learning rule.

Some hypotheses suggest that MEC stores environmental structure while HPC acts as a retrainable 'scratchboard' for translating between sensorially different but structurally similar environments[3]. Intuitively this makes sense; the laws of motion don't change, and needn't be relearned, when moving from the park to the street. We tested if this was true in our model. After initial training (in a simplified regime $w_{gB} = \delta$) we shuffled the sensory inputs to simulate a new environment, temporarily destroying path integration abilities (Fig. 2d). The system was then retrained with MEC weights (w_{gA}) frozen. Despite no MEC plasticity, path integration abilities almost perfectly recovered (Fig. 2d). MEC receptive fields reorganised with a constant phase shift along the track, while HPC cells (by construction) independently remapped. This matched biological remapping[10] and supported the hypothesis that MEC enables plasticity-free transfer of structural knowledge (here, a ring attractor tuned for path integration) between environments.

In conclusion, we reimagine the Hippocampal formation through the lens of Helmholtz machines. This framework unifies disparate aspects of its function – its ability to encode self location, path integrate and perform offline planning – and generates testable predictions, based on normative principles, regarding the role of LFP oscillations such as theta. It is interpretable, biologically plausible and extends naturally to other brain regions such as PFC which use generative models.

References **[1] Sanders, H et al. (2020)** Hippocampal remapping as hidden state inference. **[2] Carr, M et al. (2011)** Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. **[3] Whittington, J et al. (2020)** The Tolman-Eichenbaum machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. **[4] Sargolini, F et al. (2006)** Conjunctive representation of position, direction, and velocity in entorhinal cortex. **[5] Hasselmo, M et al. (2002)** A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. **[6] Urbanczik, R et al. (2013)** Learning by the Dendritic Prediction of Somatic Spiking. **[7] Dayan, P et al. (1995)** The Helmholtz Machine. **[8] Sorscher, B et al. (2023)** A unified theory for the computational and mechanistic origins of grid cells. **[9] Zhang, K (1996)** Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble. **[10] Fyhn, T et al. (2007)** Hippocampal remapping and grid realignment in entorhinal cortex.