Simulating hippocampal-mPFC interactions with a mixture-of-experts model

Dhairyya Singh and Anna C. Schapiro
Department of Psychology, University of Pennsylvania
{dsin,aschapir}@sas.upenn.edu

Abstract

The hippocampus supports both learning fine-grained details of individual experiences and extracting shared structure across them, despite their competing computational demands. Prior work suggests a division of labor within the hippocampus: the trisynaptic pathway (TSP) builds pattern-separated representations for episodic learning, whereas the monosynaptic pathway (MSP) builds overlapping representations for structure learning. But how does the brain deploy the right pathway for a given task? We introduce a mixture-of-experts framework with MSP- and TSP-like neural network experts and an mPFC-inspired gating network that computes trial-wise weightings of their outputs. Trained end-to-end, the gating network learns to preferentially recruit MSP for categorization and TSP for exemplar learning. We show that the model reproduces findings from rodent contextual fear conditioning experiments, where mPFC disruption impairs fear-memory specificity without disrupting fear learning. Adaptive gating shifts recruitment toward TSP when a novel shock context must be encoded, reducing representational overlap with similar contexts. These results offer a mechanistic theory of how mPFC may dynamically engage hippocampal pathways to balance competing memory demands.

Introduction

The hippocampus supports both episodic and structure learning (Dickerson & Eichenbaum, 2010; Mack et al., 2018)—functions that impose contrasting computational demands. C-HORSE, a biologically-grounded model of the hippocampus, assigns these roles to separate pathways: the TSP builds sparse pattern-separated representations that orthogonalize episodes, while the MSP builds overlapping representations that extract structure (Schapiro et al., 2017). A critical puzzle remains: how does the brain adaptively deploy the right pathway for a given task? Prior work implicates mPFC–hippocampal interactions as a candidate control mechanism (Eichenbaum, 2017; Preston & Eichenbaum, 2013), and evidence from rodents suggests that the mPFC-Re-HPC circuit specifically regulates memory specificity in fear conditioning (Xu & Südhof, 2013). We explore learning mechanisms that could allow the PFC to exert this kind of task-specific control over hippocampal representations.

We propose a heterogenous Mixture-of-Experts system (MoE; Jacobs et al., 1991) comprising MSP- and TSP-like experts, imbued with properties of the biological subcircuits, and an mPFC-inspired gating network. In prior work we showed that the experts developed contrasting representational profiles: MSP learned shared structure across items, whereas TSP separated individual items (Singh, Williams, & Schapiro, 2025). The mPFC-like gate learned to adaptively recruit hippocampal subcircuits according to task demands: at the aggregate level, TSP was preferentially recruited for exemplar recognition and MSP for categorization, while trial-wise TSP recruitment during categorization increased for more novel category exceptions. Here, we extend this framework to contextual fear conditioning, asking whether adaptive gating can account for how mPFC disruption selectively impairs fear-memory specificity.

Results

We used a shared-latent mixture-of-experts architecture set up as an autoencoder in which MSP- and TSP-like experts encoded synthetic environmental context vectors concatenated with a freeze/no-freeze outcome label. Trial-wise mPFC gating weights (α) combined the experts' hidden states into a common latent representation, which was decoded to reconstruct the full context–label pattern **(Fig. 1A)**. This allowed us to examine how control over pathway recruitment shapes hippocampal representations and fear generalization.

The simulation was modeled on the Xu and Südhof (2013) paradigm, in which animals experience tone-shock pairings in a specific context for three trials and were later tested in: the training context, an altered but similar context, and in response to the tone alone in the altered context **(Fig 1B)**. Empirically, PFC disruption selectively reduced memory specificity: lesioned animals showed intact freezing in the training context and tone test, and overgeneralized the freezing behavior to the altered context **(Fig 1C)**.

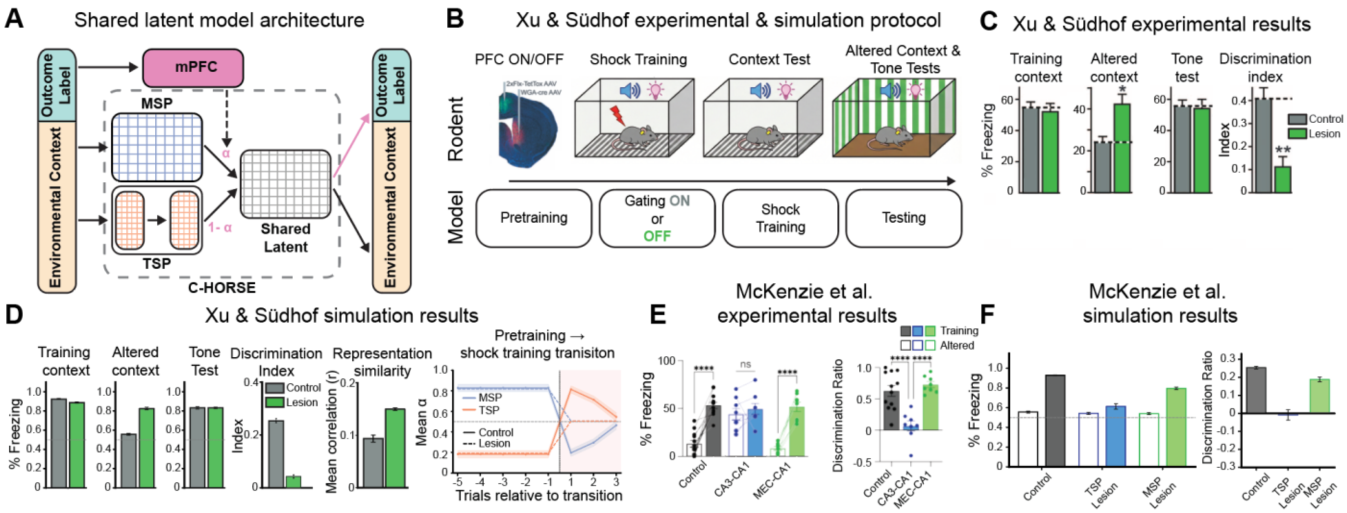

Figure 1: Model architecture, task structure and results. **(A)** MoE architecture. Trial-wise gating weights were applied to the MSP and TSP hidden states before integration into a shared latent representation, from which the final outcome-label and environmental context outputs were reconstructed. **(B)** Xu and Südhof (2013) experimental and simulation protocol. **(C)** Xu and Südhof (2013) experimental results: freezing across conditions and discrimination index for control and lesion groups. **(D)** Xu and Südhof (2013) simulation results: model freezing across conditions, discrimination index, representational similarity between the shock-training and altered contexts, and MSP/TSP recruitment across the pretraining to shock-training transition. **(E)** McKenzie et al. (2026) experimental results: freezing and discrimination ratio across control and hippocampal lesion conditions. **(F)** McKenzie et al. (2026) simulation results.

The model was first pretrained on 100 similar context vectors paired with a no-freeze outcome label, approximating a homogeneous safe home cage experience. It was then trained on a single novel shock-context vector paired with a freeze outcome label for three trials and tested on vectors corresponding to the shock context, a less similar altered context, and a tone-test context with high similarity to the shock context **(Fig. 1B)**. Freezing was measured from the reconstructed outcome label, with overgeneralization indexed as an indiscriminate freezing response to the altered context. mPFC lesions were implemented by clamping gating weights to equal recruitment of both experts (α = 0.5), inactivating learned mPFC control. The simulation reproduced the empirical pattern: the intact model preserved discrimination, while the lesioned model showed impaired discrimination **(Fig. 1D)**.

This behavioral effect was accompanied by a sharp shift in pathway recruitment: During pretraining, learned gating favored MSP, consistent with similarity among safe contexts. Upon shock training, control models showed a marked spike in TSP recruitment, whereas this shift was fully attenuated under lesion, given the clamped gating **(Fig. 1D)**. Consequently, lesioned models showed greater representational overlap between shock-training and altered contexts, explaining the behavioral overgeneralization as a failure of pattern separation.

Our account predicts that TSP supports fear-memory specificity by separating the shock context from similar safe contexts. We therefore tested whether the model could capture a pathway-specific dissociation predicted by this mechanism. McKenzie et al. (2026) found that disrupting TSP but not MSP impaired contextual discrimination, consistent with our account **(Fig. 1E)**. Our simulation of their paradigm reproduced this dissociation: TSP lesions impaired discrimination, while MSP lesions preserved it **(Fig. 1F)**.

Conclusion

Our model offers a novel mechanistic account of how mPFC could learn to flexibly coordinate complementary hippocampal pathways to support task-appropriate memory use. In contextual fear conditioning, adaptive gating promotes memory specificity by engaging TSP-like pattern separation during encoding of novel, behaviorally significant contexts. This work provides a solution to how the brain may resolve competing demands on memory through learned, dynamic control over hippocampal representations.

Acknowledgements & Disclosures

This work was funded by National Institutes of Health grants R01 MH129436 and RF1 NS127128. Generative AI was used for minor language editing.

References

Dickerson, B. C., & Eichenbaum, H. (2010). The episodic memory system: Neurocircuitry and disorders. Neuropsychopharmacology, 35(1), 86–104. https://doi.org/10.1038/npp.2009.126

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. Nature Reviews Neuroscience, 18(9), 547–558. https://doi.org/10.1038/nrn.2017.74

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1), 79–87. https://doi.org/10.1162/neco.1991.3.1.79

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. Neuroscience Letters, 680, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

McKenzie, C., Ghosh, A., Ramsaran, A. I., Wu, B., Srinivasan, S., Liu, T., Josselyn, S. A., Richards, B., & Frankland, P. W. (2026). Distinct CA1 inputs support shifts in neural dimensionality and memory resolution. bioRxiv. https://doi.org/10.64898/2026.01.14.699532

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. Current Biology, 23(17), R764–R773. https://doi.org/10.1016/j.cub.2013.05.041

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Singh, D., Williams, A. B., & Schapiro, A. C. (2025). Resolving the tension between exemplar and structure learning through a mixture-of-experts model of hippocampal-mPFC interaction [Abstract]. Cognitive Computational Neuroscience, Amsterdam, The Netherlands.

Xu, W., & Südhof, T. C. (2013). A Neural Circuit for Memory Specificity and Generalization. Science, 339(6125), 1290–1295. https://doi.org/10.1126/science.1229534