The hippocampus represents abstract graph structure at multiple scales along its longitudinal axis

Valerio Rubino[1] **Anna-Lea Beyer**[2,3] **Charley M. Wu**[2,3] **Manuela Piazza**[1]

[1] Center for Mind/Brain Science, University of Trento, 38068 Italy
[2] Center for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Darmstadt, Germany
[3] Hessian.AI, Darmstadt, Germany

valerio.rubino@unitn.it

Abstract

Humans extract latent structure from their environment, and growing evidence suggests that the hippocampus encodes such structure through predictive representations. In spatial domains, these representations are organized along the hippocampal longitudinal axis, with posterior regions encoding finer-grained, local structure and anterior regions supporting more global, coarse-grained structure. To test whether this principle extends to abstract, non-spatial relational structures, we reanalyzed fMRI data from participants viewing sequences drawn from a previously learned graph. We found that the left hippocampus represents abstract graph structure at multiple scales, with predictive scale increasing from posterior to anterior regions. Our findings suggest that hippocampal multi-scale predictive representations may constitute a domain-general mechanism for relational processing.

Introduction

Humans readily extract latent structure from their surroundings. For instance, by observing interactions within a group, we can infer its underlying social network and, at a larger scale, how groups relate to one another. These multi-scale relational structures can be naturally formalized as graphs, in which nodes correspond to entities and edges to their relationships (Kemp & Tenenbaum, 2008; Mark et al., 2020).

There is growing evidence that the hippocampal formation supports structural learning through predictive representations (Behrens et al., 2018; Garvert et al., 2017; Mark et al., 2026; Schapiro et al., 2015), which resemble the Successor Representation (SR; Stachenfeld et al., 2017). The SR was originally introduced in reinforcement learning (Dayan, 1993), and encodes each state in terms of its expected discounted future occupancy, governed by a parameter γ (Gershman, 2018). By varying γ, the SR naturally captures structure at multiple scales (Momennejad, 2020).

In spatial domains, such multi-scale organization of SR-like maps is reflected along the hippocampal longitudinal axis (Brunec & Momennejad, 2021). While behavioral and computational evidence suggests that SR-like representations also support planning (Momennejad et al., 2017), social reasoning (Son et al., 2024), and conceptual knowledge (Haga et al., 2023), it remains unknown whether the hippocampus encodes non-spatial relational structure at multiple scales. Here, we test this possibility by reanalyzing fMRI data from participants who previously learned an abstract graph structure (Garvert et al., 2017).

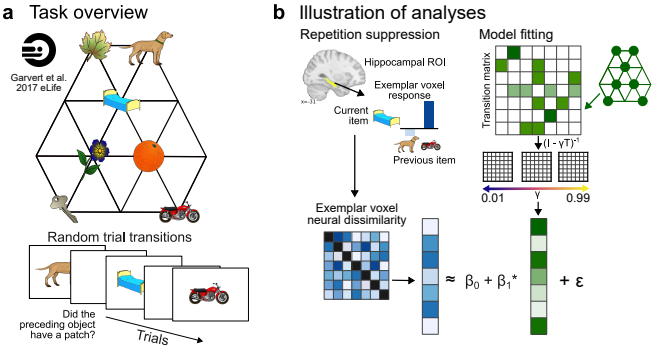

Figure 1: (**a**) During fMRI scanning, participants viewed pseudorandom sequences of seven items drawn from a previously learned graph. (**b**) We estimated voxel-wise neural dissimilarity and modeled it using the SR.

Results

In the original Garvert et al. (2017) study, participants learned the structure of a 12-node graph (Fig. 1a). On the following day, they underwent fMRI scanning across three runs while viewing pseudorandom sequences drawn from a subset of seven nodes, with transition probabilities reflecting the learned graph structure. We used the preprocessing pipeline from Garvert et al. (2017). Using a repetition suppression approach, we modeled each trial as one of the 42 possible transitions between node pairs. For each voxel, we estimated one β coefficient per transition, with larger responses reflecting reduced repetition suppression between items (and thus greater dissimilarity). Averaging β estimates across the three runs yielded a dissimilarity matrix for each voxel (Fig. 1b, left).

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026.

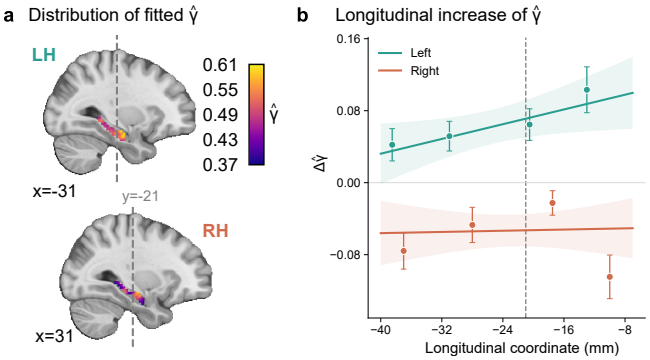

Figure 2: **(a)** Spatial distribution of fitted $\hat{\gamma}$ values across the hippocampus. **(b)** Longitudinal gradient of $\Delta\hat{\gamma}$ (demeaned within participant) along the hippocampal longitudinal axis, shown separately for each hemisphere. The dashed line marks the anterior–posterior boundary at $y = -21$ mm (Poppenk et al., 2013).

All functional images were normalized to Montreal Neurological Institute (MNI) space. We defined a hippocampal region of interest (ROI) using the probabilistic Julich Brain Atlas (Amunts et al., 2020), including cornu ammonis and dentate gyrus, thresholded at $p \geq 0.80$. We modeled voxel-wise neural dissimilarities within bilateral hippocampal ROIs using parametric representational similarity analysis (Fan et al., 2024; Kriegeskorte, 2008; Fig. 1b). From the graph learned on the previous day, we computed the SR (Dayan, 1993; Gershman, 2018; Stachenfeld et al., 2017; Fig. 1b, right), as:

$$\mathbf{M} = (\mathbf{I} - \gamma\mathbf{T})^{-1}, \tag{1}$$

where $\mathbf{T}$ is the transition probability matrix and $\gamma \in (0, 1)$. We predicted voxel-wise neural dissimilarity as:

$$\beta_0 + \beta_1\, \mathrm{M}(\gamma), \tag{2}$$

and fit β_0, β_1, and γ to minimize the mean squared error.

We averaged voxel-wise R^2 within each hemisphere for each participant, and found them significantly higher than zero (left: $t(22) = 3.93$, $p < .001$; right: $t(22) = 4.78$, $p < .001$), with no statistically significant difference between hemispheres (paired $t(22) = 2.03$, $p = .055$).

Next, we tested whether $\hat{\gamma}$ varied along the hippocampal longitudinal axis (Fig. 2a), using mixed-effects models predicting voxel-wise $\hat{\gamma}$, including participants as random intercepts. We found that $\hat{\gamma}$ increased with the y-coordinate in MNI space ($b = 0.03$, $[0.02, 0.04]$, $p < .001$; Fig. 2b), with a significant negative interaction with the right hemisphere ($b = -0.03$, $[-0.05, -0.01]$, $p = .002$): the effect was significant in the left hippocampus ($b = 0.03$, $[0.02, 0.04]$, $p < .001$), but not in the right ($b = 0.00$, $[-0.01, 0.01]$, $p = .948$).

To rule out that the gradient was driven by poorly fitted voxels, we restricted analyses to voxels exceeding progressively stricter goodness-of-fit thresholds, and found consistent results ($R^2_{\min} \in [0, .01, .025, .05]$). Furthermore, averaging $\hat{\gamma}$ across voxels at each y-coordinate still yielded a significant effect ($b = 0.04$, $[0.01, 0.07]$, $p = .009$). The effect remained significant ($b = 0.03$, $[0.02, 0.05]$, $p < .001$) when controlling for the effects of estimated slope $\hat{\beta}_1$ ($b = -0.06$, $[-0.24, 0.13]$, $p = .548$) and R^2 ($b = 1.70$, $[1.28, 2.12]$, $p < .001$). Therefore, the gradient was not driven by collinearity between β_1 and $\hat{\gamma}$ or by differences in model fit. Results qualitatively replicated when predicting $\hat{\gamma}$ on the logit scale, given its bounded range $(0, 1)$.

Collectively, these results suggest that the left hippocampus represents abstract graph structure at multiple scales along its longitudinal axis.

Discussion

Our environment contains structure at multiple scales (Rubino et al., 2026). Poppenk et al. (2013) proposed that the hippocampus concurrently represents them along its longitudinal axis, with posterior regions encoding finer, more detailed representations and anterior regions encoding coarser ones. Accordingly, Brunec and Momennejad (2021) showed that spatial representations become progressively coarser along this axis. This gradient was captured by the SR, with γ increasing from the posterior to the anterior hippocampus.

Replicating and extending these findings, we found that the hippocampus represents multi-scale SR-like maps of a non-spatial, abstract graph structure, with γ increasing along its longitudinal axis. One key difference is that this effect was restricted to the left hippocampus, possibly reflecting a bias toward representing discrete structure rather than continuous space (Jordan, 2019).

These findings suggest that the multi-scale SR provides a domain-general model for organizing relational knowledge, potentially underlying other hippocampal coarse-graining gradients observed in non-spatial domains such as narrative processing (Collin et al., 2015) and conceptual learning (Viganò & Piazza, 2021). Indeed, Stoewer et al. (2022) showed that abstract categories, such as *animals* or *vehicles*, emerge from SR-based semantic models with higher γ values. Thus, similar representations may support semantic abstraction, suggesting a possible mechanistic account of hippocampal contributions to semantic knowledge acquisition (Elward & Vargha-Khadem, 2018).

Representing structure at multiple scales may enable flexible, task-dependent switching between local and global representations. Kahn and Daw (2025) showed that humans adaptively arbitrate between detailed and temporally abstract world models. Our results suggest that such flexibility may be supported by specialization along the hippocampal longitudinal axis, with anterior regions supporting coarser, longer-range predictions and posterior regions encoding finer, more local structure.

Acknowledgments

We used large language models as general-purpose tools for language editing, and for debugging and refining analysis code. All analyses and results were verified by the authors. MP and VR are supported by the Italian Ministry of Education, University and Research under the PRIN 2022 programme (Grant Agreement No. 2022EBC78W), and by a EU Next Generation PhD fellowship (PNRR funds, DM 118/2023). CMW and ALB are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (C^4: 101164709), by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223), and the Excellence Cluster "Reasonable AI" by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy – EXC-3057.

References

Amunts, K., Mohlberg, H., Bludau, S., & Zilles, K. (2020). Julich-brain: A 3d probabilistic atlas of the human brain's cytoarchitecture. *Science*, *369*(6506), 988–992. https://doi.org/10.1126/science.abb4588.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organising knowledge for flexible behaviour. https://doi.org/10.1101/365593.

Brunec, I. K., & Momennejad, I. (2021). Predictive representations in hippocampal and prefrontal hierarchies. *The Journal of Neuroscience*, *42*(2), 299–312. https://doi.org/10.1523/jneurosci.1327-21.2021.

Collin, S. H. P., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience*, *18*(11), 1562–1564. https://doi.org/10.1038/nn.4138.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613.

Elward, R. L., & Vargha-Khadem, F. (2018). Semantic memory in developmental amnesia. *Neuroscience letters*, *680*, 23–30.

Fan, Y., Wang, M., Fang, F., Ding, N., & Luo, H. (2024). Two-dimensional neural geometry underpins hierarchical organization of sequence in human working memory. *Nature Human Behaviour*, *9*(2), 360–375. https://doi.org/10.1038/s41562-024-02047-8.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*. https://doi.org/10.7554/elife.17086.

Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *The Journal of Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/jneurosci.0151-18.2018.

Haga, T., Oseki, Y., & Fukai, T. (2023). A unified neural representation model for spatial and semantic computations. https://doi.org/10.1101/2023.05.11.540307.

Jordan, J. T. (2019). The rodent hippocampus as a bilateral structure: A review of hemispheric lateralization. *Hippocampus*, *30*(3), 278–292. https://doi.org/10.1002/hipo.23188.

Kahn, A. E., & Daw, N. D. (2025). Humans rationally balance detailed and temporally abstract world models. *Communications psychology*, *3*(1), 1.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008.

Mark, S., Moran, R., Parr, T., Kennerley, S. W., & Behrens, T. E. J. (2020). Transferring structural knowledge across cognitive maps in humans and models. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18254-6.

Mark, S., Schwartenbeck, P., Hahamy, A., Samborska, V., Baram, A. B., & Behrens, T. E. (2026). Flexible neural representations of abstract structural knowledge in the human entorhinal cortex. *eLife*, *13*. https://doi.org/10.7554/elife.101134.3.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8.

Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. https://doi.org/10.31234/osf.io/b6sr8.

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, *17*(5), 230–240. https://doi.org/10.1016/j.tics.2013.03.005.

Rubino, V., Dayan, P., & Wu, C. M. (2026). Simplicity guides the discovery and use of compositionality. *PsyArXiv*. https://doi.org/10.31234/osf.io/25pha_v1.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2015). Statistical learning of temporal community structure in the hippocampus: Statistical learning of temporal community structure.

Hippocampus, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523.

Son, J.-Y., Vives, M.-L., Bhandari, A., & FeldmanHall, O. (2024). Replay shapes abstract cognitive maps for efficient social navigation. *Nature Human Behaviour*, *8*(11), 2156–2167. https://doi.org/10.1038/s41562-024-01990-w.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650.

Stoewer, P., Schlieker, C., Schilling, A., Metzner, C., Maier, A., & Krauss, P. (2022). Neural network based successor representations to form cognitive maps of space and language. *Scientific Reports*, *12*(1), 11233.

Viganò, S., & Piazza, M. (2021). The hippocampal-entorhinal system represents nested hierarchical relations between words during concept learning. *Hippocampus*, *31*(6), 557–568. https://doi.org/10.1002/hipo.23320.