Structural abstraction of emotion knowledge in hippocampal-prefrontal systems

Yumeng Ma (yumeng.ma@emory.edu)
Department of Psychology, Emory University
Atlanta, GA 30032 USA

Philip A. Kragel (pkragel@emory.edu)
Department of Psychology, Department of Psychiatry and Behavioral Sciences, Emory University
Atlanta, GA 30032 USA

Abstract

Hippocampal-prefrontal systems organize knowledge in a map-like way across multiple domains, from physical space to abstract concepts. It is well known that knowledge about emotional events is organized in a low-dimensional space, raising the question of whether the brain uses cognitive maps to represent emotion concepts. Using functional magnetic resonance imaging while participants viewed emotionally evocative film clips, we decoded patterns of neural activity in hippocampal-prefrontal systems to predict representations of emotion concepts in a computational model of relational memory inspired by the hippocampal formation. Our findings demonstrate that hippocampal-prefrontal systems contain map-like representations of emotion concepts at multiple levels of granularity. These findings provide new insight into how the brain organizes knowledge about emotional events to react adaptively in a complex environment.

Keywords: emotion concepts; cognitive map; hippocampal formation; prefrontal cortex

Introduction

The brain builds mental models of relations between entities in a map-like way to support flexible behavior. A growing body of research suggests that the hippocampal-prefrontal system encodes cognitive maps not only for physical spaces but also for conceptual spaces (Behrens et al., 2018). Prior studies have primarily studied cognitive maps using artificially constructed concepts in controlled laboratory experiments in which abstract relationships are explicitly taught and reinforced (Constantinescu et al., 2016). Despite evidence supporting the flexible use of cognitive maps to guide behavior, it remains unknown whether similar mechanisms are employed to organize conceptual knowledge that naturally develops from our rich and dynamic experiences in daily life, such as knowledge about emotional episodes.

Decades of behavioral research have demonstrated that when humans communicate about emotional experiences and recognize emotional expressions, they do so using an affective space organized by dimensions of valence and arousal (Remmington et al., 2000). Given this evidence, we hypothesize that emotion concepts are mapped within hippocampal-prefrontal systems using mechanisms similar to those used for mapping physical and other conceptual spaces (Constantinescu et al., 2016; Park et al., 2021). This account predicts that individual emotion concepts are encoded in the activity of distinct hippocampal populations, and that a relational network linking emotions to one another is represented in population activity in the entorhinal cortex and ventromedial prefrontal cortex (vmPFC). Investigating whether cognitive mapping mechanisms extend to knowledge about emotions provides insight into how the brain models complex, naturally occurring experiences.

Methods & Results

We tested whether emotion concepts could be represented using cognitive maps by analyzing behavioral and fMRI measures from the Emo-FilM dataset (Morgenroth et al., 2025). In this study, participants (N = 29) watched a series of cinematic videos that robustly engaged multiple emotion concepts. An independent group of subjects provided self-report ratings that reflect normative emotional experience. To explicitly model hippocampal-prefrontal systems, we used the Tolman-Eichenbaum Machine (TEM; Whittington et al., 2020) to simulate interconnected neural populations that learn to represent emotion concepts (**p** cells) and structural relations among them (**g** cells) in order to predict upcoming sensory experiences.

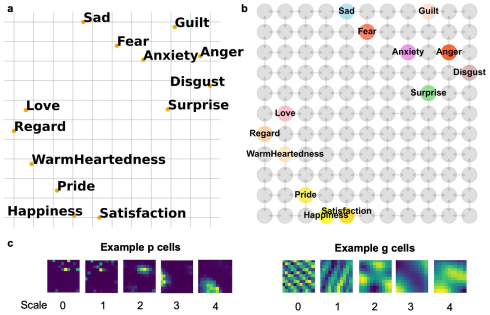

Figure 1. Learning emotion concepts with TEM. a) Multidimensional scaling on self-reported ratings of emotion categories across all film stimuli. b) The 11-by-11 discrete environment derived from emotion self-report. c) Rate maps of example **p** and **g** cells obtained by averaging the activity of each cell at each node after training. Brighter colors indicate greater activity.

To simulate learning, we created an environment based on the normative film ratings, such that agents exploring

the environment experienced emotion transitions that approximate human ratings. We used multidimensional scaling to map emotion categories into a two-dimensional space, which was discretized into a square environment (Fig. 1). TEM agents repeatedly explored this environment, learning to map locations of emotion concepts in the two-dimensional space. To model human brain activity as participants watched emotional videos, we averaged the activations of **p** and **g** cells weighted by the emotion ratings at each time point of the film stimuli, approximating trajectories in the conceptual space that took place during film viewing.

Because TEM includes artificial neurons at multiple levels of abstraction, it can represent both fine and coarse-grained aspects of experience (e.g., whether an event might be described as 'horrifying', 'scary', or more generally 'bad'). Following Whittington et al. (2020), we specified five levels of abstraction. Because representations in **p** and **g** differ the most at fine-grained levels/small scales (see Fig. 1c), we predicted that hippocampal decoding of **p** would be more accurate than **g** for units with narrow firing fields, as they better capture specific emotions rather than large distances in the emotion space.

We specified two fMRI decoding models using partial least squares regression to predict activity of **p** and **g** from patterns of hippocampal BOLD response (using within subject modeling and leave-one-film-out cross-validation). Model performance was quantified as the Fisher-transformed correlation between time series of the decoded and actual activity in TEM. We found that information about emotion concepts (i.e., the average response of cells in **p**) could be decoded from the hippocampus (z = 0.0543, 95% bootstrap *CI* [0.0463, 0.0623]), and that **p** could be decoded more accurately than **g** (Δz = 0.0017, 95% bootstrap *CI* [0.0001, 0.0034], p = .0145). Comparisons across representational scales revealed that fine-grained information was read out more accurately for **p** than **g** (Δz = 0.0352, 95% bootstrap *CI* [0.0254, 0.0448], p < .0001; Fig. 2, left), suggesting that the hippocampus exhibits sparse concept representations more than relational representations at small scales.

Given better decoding of small-scale **p** activity in the hippocampus, we next grouped TEM activity of emotion concepts by representational scale and examined whether the granularity of representations varied across the long axis of the hippocampus. Examining differences between small (0-1) and large (2-4) scale representations of space, we found that more information about small-scale **p** cell activity was present in posterior compared to anterior hippocampus (Δz = 0.0052, 95% bootstrap *CI* [0.0033, 0.0071], p = .0019).

Next, because the entorhinal cortex and vmPFC are thought to represent conceptual spaces in a relational manner as opposed to locations in the relational graph, we tested whether more information about **g** was present these two regions compared to the hippocampus. An analysis of variance testing this hypothesis revealed that decoding performance varied as a function of brain region, TEM component (**g** and **p**), and representational scale (i.e., a three-way interaction; $F(8, 616) = 12.44$, p < .0001; Fig. 2).

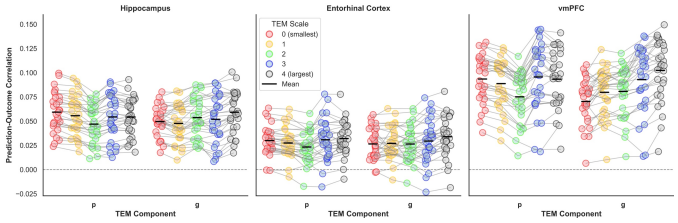

Figure 2. Performance of decoding models trained to predict TEM **p** and **g** activity from BOLD activity in hippocampus, entorhinal cortex, and vmPFC.

Post hoc tests revealed decoding **g** was more accurate than **p** at larger scales in the vmPFC (Δz = 0.0712, 95% bootstrap *CI* [0.0614, 0.0809], p < .0001; Fig. 2, right), and to a lesser degree in the entorhinal cortex (Δz = 0.0099, 95% bootstrap *CI* [0.0001, 0.0199], p = .0461). Decoding **g** was more accurate at larger scales in the vmPFC than in the hippocampus (Δz = 0.0542, 95% bootstrap *CI* [0.0447, 0.0640], p < .0001) and entorhinal cortex (Δz = 0.0604, 95% bootstrap *CI* [0.0506, 0.0700], p < .0001).

We additionally observed better decoding of **p** at small scales in the hippocampus compared to both entorhinal cortex (Δz = 0.0191, 95% bootstrap *CI* [0.0093, 0.0289], p = .0003) and vmPFC (Δz = 0.0182, 95% bootstrap *CI* [0.0086, 0.0279], p = .0031). Together, these results suggest that the vmPFC better represents large-scale structural abstractions (e.g., that one portion of a video was more pleasant than another), whereas the hippocampus contributes more to small-scale, fine-grained representations of individual concepts.

In sum, these findings, together with evidence that hippocampal-prefrontal systems represent self-reported emotions (Ma & Kragel, 2025), suggest that emotion concepts are encoded in a map-like way at multiple levels of granularity.

References

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468.

Ma, Y., & Kragel, P. A. (2025). The representation of emotion knowledge in hippocampal-prefrontal systems. *bioRxiv*. https://doi.org/10.1101/2025.06.03.657722

Morgenroth, E., Moia, S., Vilaclara, L., Fournier, R., Muszynski, M., Ploumitsakou, M., Almató-Bellavista, M., Vuilleumier, P., & Van De Ville, D. (2025). Emo-FilM: A multimodal dataset for affective neuroscience using naturalistic stimuli. *Scientific Data*, *12*(1), 684.

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature Neuroscience*, *24*(9), 1292–1301.

Remmington, N. A., Fabrigar, L. R., & Visser, P. S. (2000). Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology*, *79*(2), 286–300.

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249-1263.e23.