An In Silico Exploration of the Effect of Surprising Information on Hippocampal Representations

Emily M. Heffernan, Michael L. Mack

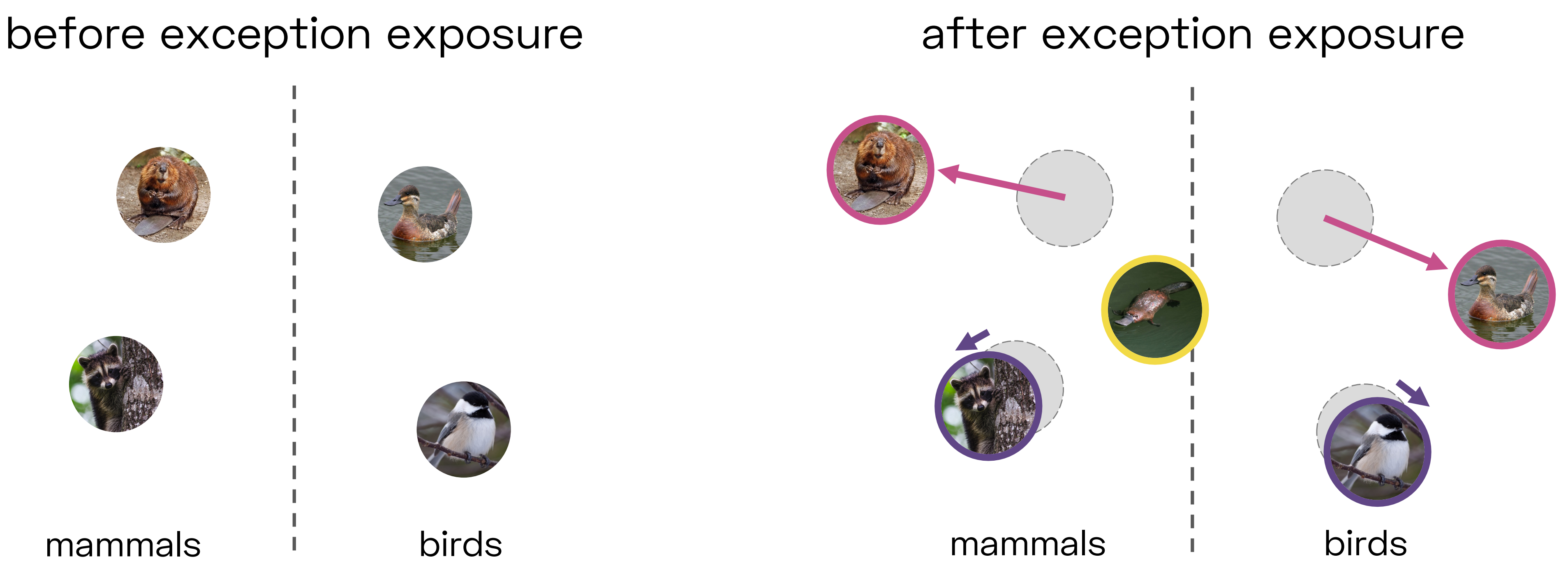
Department of Psychology, University of Toronto

Introduction

In rule-plus-exception category learning, the learner must both generalize to understand the category rule and rapidly form distinct representations of exceptions.

Though exceptions are thought to be uniquely represented in memory,[1] they may affect the learning of rule-followers.[2,3] Here we used modelling simulations to explore how exception introduction impacts existing representations of rule-following items.

We predicted that introducing exceptions would impact existing category representations, and that rule-followers that were **similar to exceptions** would be impacted more than rule-followers that were **dissimilar to exceptions.**



Do category exceptions impact the nature of rule-following items? And, is this impact a function of similarity?

Modelling Simulations

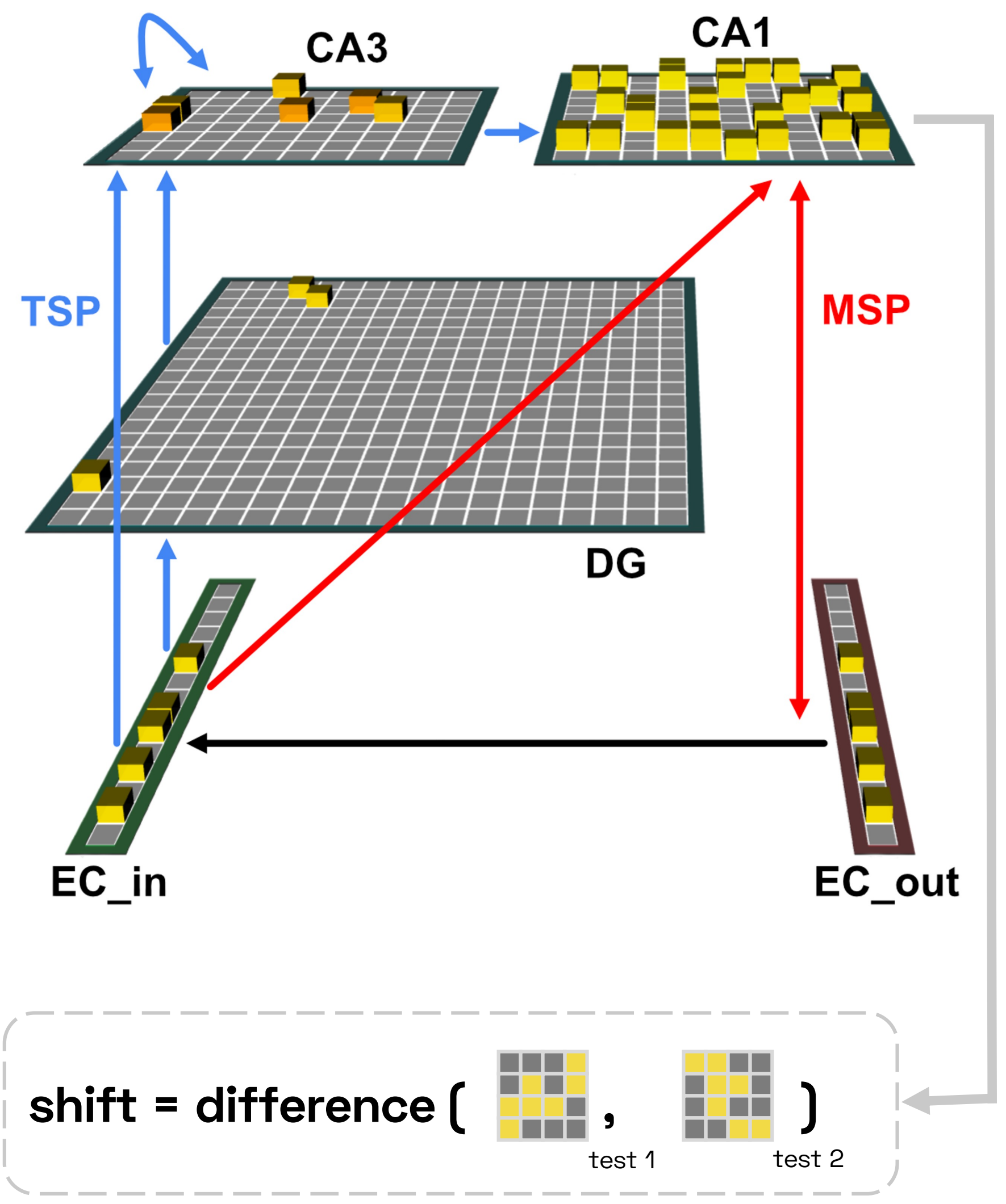
We used the existing C-HORSE model of hippocampus, which captures hippocampal subfields and white matter connections.[4,5]

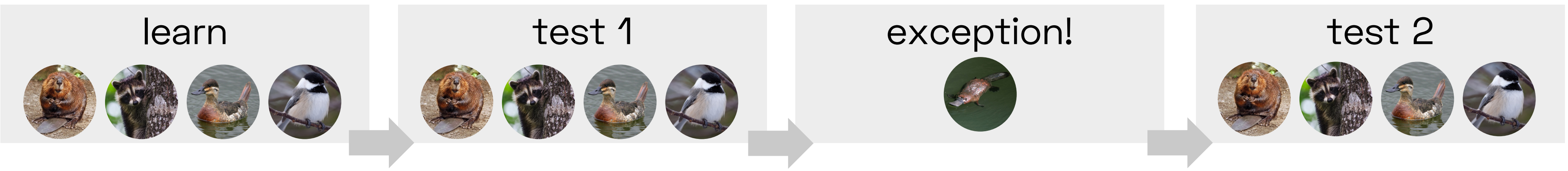
Why hippocampus? Past work has highlighted its role in rule-plus-exception learning,[e.g.,1,6,7] and the present model accounts for behavioural category learning results.[3]

The model was exposed to stimuli (coded as binary-valued vectors) that varied across two category-relevant and one category irrelevant dimension.[3,8]

Subfield representations of rule-followers was recorded before (test 1) and after (test 2) exception introduction.

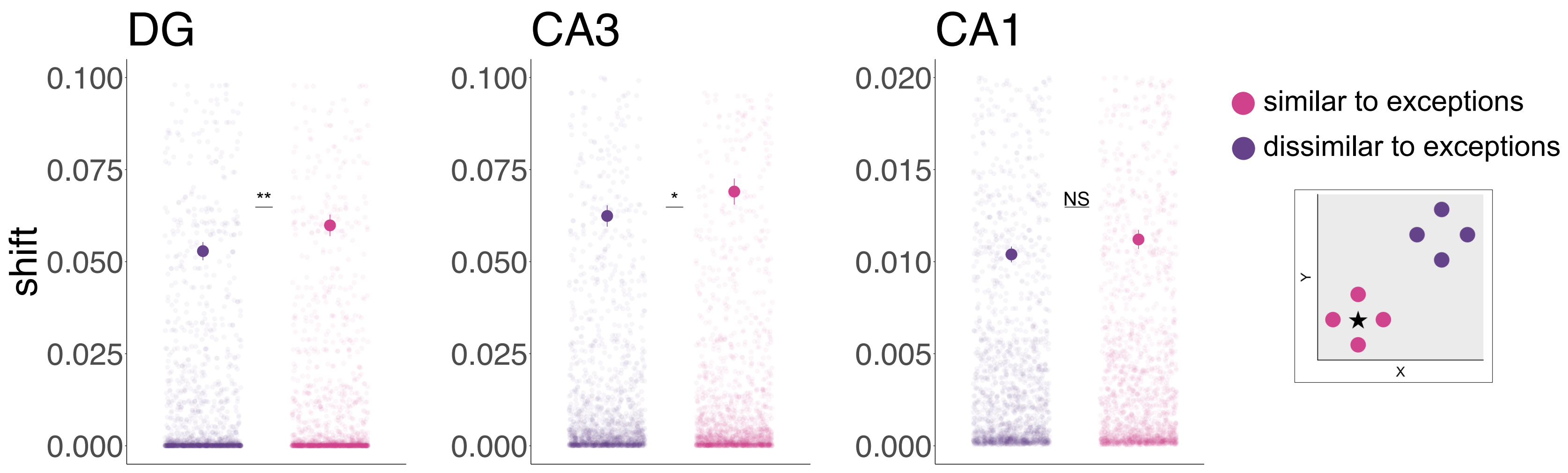
Shift was formalized as the difference between a stimulus's representation before and after exception exposure, calculated using cosine similarity.





Exceptions shift rule-follower representations in a similarity-dependent manner

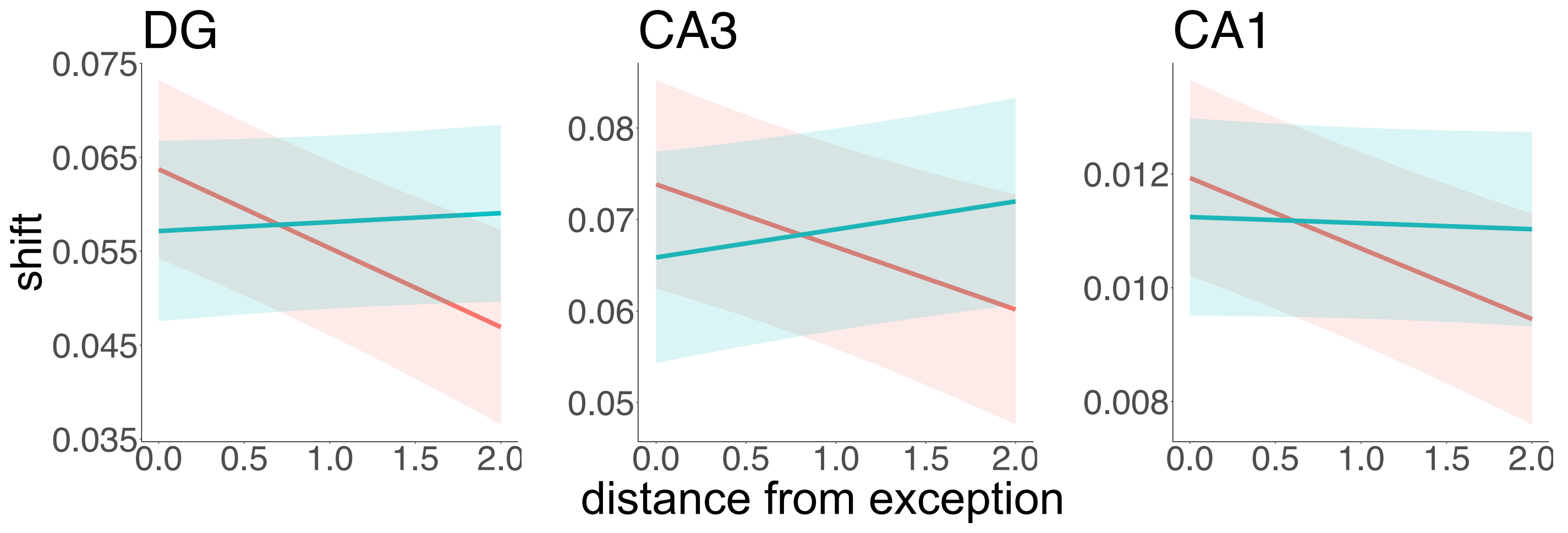
In CA3 and DG subfields, shift was significantly greater for items **similar** to exceptions compared to items **dissimilar** to exceptions. The same trend was apparent in CA1.



Category-relevant similarity predicts shift better than feature-based similarity

As past work has indicated that hippocampus maps conceptually relevant, but not irrelevant, features,[6,9] we explored how well concept- and feature-based similarity predicted shift.

AIC comparisons of model fits indicated that similarity computed using **category-relevant** features was a better predictor of shift than distance that also included **category-irrelevant** features.



Conclusions

Modelling results indicate that the introduction of exceptions impacts rule-follower representation. Rule-followers similar to exceptions experienced greater shift than those dissimilar to exceptions, suggesting that the influence of exceptions on category structure is coded in a task-specific manner.

Similarity along conceptually relevant dimensions better predicted shift than similarity that included conceptually irrelevant dimensions, providing further computational evidence of hippocampal encoding of category information.

These findings shed light on how we may accommodate new information at the potential expense of what we have already learned and provides a foundation for future neuroimaging studies.

References

1. Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. Cerebral Cortex, 22(2), 260–273.
2. Castro, L., Yang, S., Savic, O., Sloutsky, V., & Wasserman, E. (2021). Not all exceptions are created equal: Learning of exceptions in pigeons' categorization. Psychonomic Bulletin & Review, 28(4), 1344–1353.
3. Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. Scientific Reports, 11(1), 21429.
4. Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus. PLoS Computational Biology, 9(6), e1003067.
5. 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).
6. Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. Proceedings of the National Academy of Sciences of the United States of America, 113(46), 13203–13208.
7. Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. Journal of Neuroscience, 38(10).
8. Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs: General and Applied, 75, 1–42.
9. Theves, S., Fernández, G., & Doeller, C. F. (2020). The Hippocampus Maps Concept Space, Not Feature Space. Journal of Neuroscience, 40(38), 7318–7325.

Acknowledgements

Thanks to Meg Schlichting, Morgan Barense, and members of Budding Minds and Mack Lab for their helpful feedback and suggestions. This research is supported by the NSERC Discovery Grant to MLM (RGPIN-2017-06753), the Canada Foundation for Innovation and Ontario Research Fund to MLM (36601), the Brain Canada award to MLM, the CIHR (PJT-178337) grant to MLM, and the NSERC CGS-D award to EMH.