Neural Mechanism of Compositional Generalization

Zilu Liang (zilu.liang@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Miriam Klein-Flugge (miriam.klein-flugge@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Abstract

Compositionality is a fundamental feature of cognition. Humans can break down learned knowledge into constituents and reassemble them flexibly to solve new problems. Using fMRI, we investigated the neural underpinnings of compositionality in a task where discrete features indicated locations on spatial axes. Participants were trained on a subset of stimuli with feedback and tested on the held-out without feedback. Successful generalization required decomposing the trained stimulus-location associations into rules and recombining the rules to solve the test. Different brain regions adopted distinct representation strategies: rules were represented in high-dimensional parallel manifolds in the hippocampus and composed by vmPFC to solve the test; the superior parietal gyrus put the test stimuli into a low-dimensional spatial reference frame; V1 represented both training and test stimuli as their locations on the ground-truth map.

Introduction

The brain's language of thought is compositional (Fodor & Pylyshyn, 1988; Frankland & Greene, 2020). Extensive literature has documented that biological agents learn simple building blocks that can be recomposed to solve novel problems efficiently (Harlow, 1949; Zhou et al., 2021; Samborska et al., 2022; Dekker et al., 2022; Luettgau et al., 2024). Across linguistic and non-linguistic domains, abundant evidence suggests that the representation of a compound (e.g. phrases) is related to its constituents (e.g. words)(Barron et al., 2013; Pylkkänen, 2019; Schwartenbeck et al., 2023). In computational linguistics and machine learning literature, vector addition is commonly used as the operation for binding the constituents into a compound, i.e., superimposing the patterns of constituents to form the representation of the compound (Mitchell & Lapata, 2008, 2010).

We investigated the neural representation supporting compositional generalization in a task (Dekker et al., 2022) that asked participants to predict spatial (treasure) locations based on nonspatial stimuli (Fig.1A). Two hidden rules mapped discrete stimulus features onto the continuous spatial axes (e.g. shapes onto an x-position and colors onto a y-position, Fig.1B). Participants were trained on a subset of stimuli-location associations and tested on the holdout set without feedback (Fig.1B). After pretraining, we scanned the participants whilst making decisions about all stimuli without feedback. Successful test performance required: 1) decomposing the trained associations into two rules; 2) recombining the rules to solve the test. We focused on participants who successfully solved the test, i.e., generalizers, and asked how their neural representation solved the decomposition and recomposition problem.

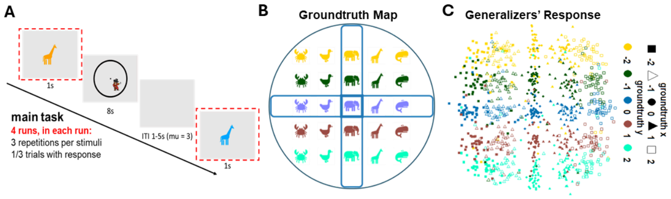

Figure 1. (A) The treasure hunt task performed in the scanner. (B) Groundtruth map of treasure locations. Stimuli highlighted in the blue box were training stimuli; others were test stimuli. (C) Response map of generalizers. Each dot indicated a generalizer's response to a stimulus averaged across runs.

Potential representation strategies

To motivate our analyses, we laid out the potential representation strategies. First, we considered the dimensionality of rule representation. Before training, the representation of discrete features was high-dimensional (highD), allowing for feature separation (Fig.2A left). After training, the representation could remain highD or compress to a low-dimensional axis (Fig.2A right, lowD). To address this, we estimated the dimensionality of rule representation with cross-validated singular value decomposition (Ahlheim & Love, 2018). Next, we considered the relationship between rules. Space could be used as a scaffold to represent the rules. If so, the rules would be represented orthogonally (Fig.2B left). Alternatively, a shared relational primitive, magnitude, could be used to describe the progression along the linear axis (left-ness and top-ness). If so, the rules would be represented in parallel (Fig.2B right, example of top of y aligned to left of x). We quantified axes alignment with a parallelism score (PS), the average cosine similarity between coding directions of different axes (e.g. crab-whale vs yellow-cyan).

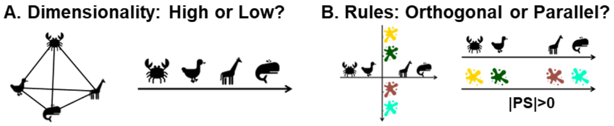
A. Dimensionality: High or Low?
B. Rules: Orthogonal or Parallel?
|PS|>0

Figure 2. Schematic illustration of the hypotheses.

Main Findings

Participants were classified into generalizers (N=41) and non-generalizers (N=15) following Dekker et al. (2022). The generalizers' responses in the scanner (Fig.1C) closely followed the groundtruth map (Fig.1B). Using searchlight representation similarity analysis (RSA), we found a lowD spatial map only for the test stimuli in the superior parietal gyrus (SPG) but for all stimuli in the early visual area (V1). We showed the V1 representation using multidimensional scaling (MDS) in Fig.3C.

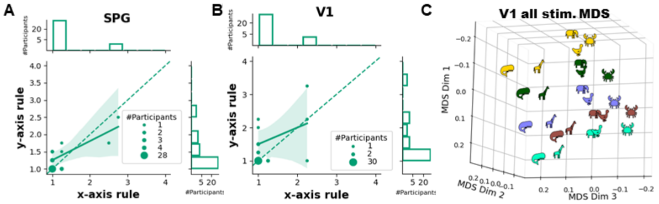

Figure 3. (A-B) SPG and V1 dimensionality. (C) V1 MDS; note how it resembles ground truth.

In stark contrast to V1 and SPG (Fig.3A-B), we found a higher-dimensional representation for both rules in the hippocampus (HPC) than in V1 and SPG (Fig.4A, both p<0.001). Moreover, in HPC, dimensionality was positively correlated with linear decodability of features after controlling for noise (r=0.44,p=0.002). Therefore, high dimensionality in HPC was not simply due to noise but helped maintain feature separation. HPC also showed abstraction of relational structure across the rules. Coding directions on different axes were significantly more parallel than expected by chance (absolute PS compared against null, p<0.001), with some generalizers being top-right aligned while some being top-left aligned (Fig.4B&D). This was further validated by a cross-validated RSA where we estimated the alignment in half of the data to generate a parallel model and let it compete with the orthogonal model to explain the neural RDM of the remaining half (Fig.4C).

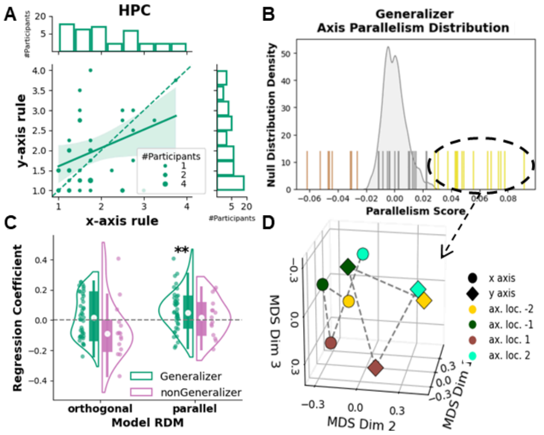

Figure 4. HPC representation. (A) Dimensionality. (B) PS for individual generalizers (sticks) and null distribution of PS (shaded area). Color indicates axis alignment classified by comparing each PS against the null. (C) Cross-validated RSA. (D) MDS of top-left parallel generalizers in (B): coding directions of the same colour pair were parallel between squares and diamonds.

In summary, we revealed the different representation strategies adopted by several regions in a compositional rule generalization task. HPC formed an abstract representation of the shared relational structure between rules, and the representation of each rule was high-dimensional. We further checked if test stimuli representation could be composed from the relevant training stimuli under the assumption of vector addition (cyan crab=cyan+crab). This analysis revealed that vmPFC but not HPC combined the rules in a highD space. Finally, the transformation of test stimuli representation into a lowD spatial map was found in SPG, whereas V1 represented all stimuli in a way that resembled the ground truth.

Acknowledgements

This research was generously funded by a Wellcome Trust Discovery Award (227928/Z/23/Z) to C.S., M.C.K.-F. is funded by a Wellcome Henry Dale Fellowship (223263/Z/21/Z) and a UKRI-underwritten ERC Starting Grant (EP/X021815/1).

References

Ahlheim, C., & Love, B. C. (2018). Estimating the functional dimensionality of neural representations. *NeuroImage*, *179*, 51–62. https://doi.org/10.1016/j.neuroimage.2018.06.015

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, *16*(10), 1492–1498. https://doi.org/10.1038/nn.3515

Dekker, R. B., Otto, F., & Summerfield, C. (2022). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences*, *119*(41), e2205582119. https://doi.org/10.1073/pnas.2205582119

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65. https://doi.org/10.1037/h0062474

Luettgau, L., Erdmann, T., Veselic, S., Stachenfeld, K. L., Kurth-Nelson, Z., Moran, R., & Dolan, R. J. (2024). Decomposing dynamical subprocesses for compositional generalization. *Proceedings of the National Academy of Sciences*, *121*(46), e2408134121. https://doi.org/10.1073/pnas.2408134121

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In J. D. Moore, S. Teufel, J. Allan, & S. Furui (Eds.), *Proceedings of ACL-08: HLT* (pp. 236–244). Association for Computational Linguistics. https://aclanthology.org/P08-1028/

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66. https://doi.org/10.1126/science.aax0050

Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J., & Akam, T. (2022). Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience*, *25*(10), 1314–1326. https://doi.org/10.1038/s41593-022-01149-8

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885-4897.e14. https://doi.org/10.1016/j.cell.2023.09.004

Zhou, J., Jia, C., Montesinos-Cartagena, M., Gardner, M. P. H., Zong, W., & Schoenbaum, G. (2021). Evolving schema representations in orbitofrontal ensembles during learning. *Nature*, *590*(7847), 606–611. https://doi.org/10.1038/s41586-020-03061-2