# Generalization of Covariance Structure in Human and Neural Network

**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:**

**The ability to identify structure in sensory data and flexibly reuse this knowledge in new settings is key to human cognition. Here, we report neural network simulations that studied the nature of the neural representations that support transfer learning. We designed a paradigm in which agents learn the mapping from stimuli onto one label set, and tested whether they detected the covariance structure among feature dimensions (e.g. colour or shape) and generalise to 1) new stimuli from unseen combinations of features, 2) an entirely new set of labels. We train a multilayer perceptron (MLP) with one hidden layer on the same task. We derive a model of representation that supports structure transfer in our task, allowing us to make predictions for a future human study.**

**Keywords: generalisation; abstraction; representational geometry**

## Introduction

A key hallmark of intelligence is the ability to detect regularities and patterns in the world and repurpose this knowledge to make inferences in new settings (structure transfer). However, a mechanistic account of how structure transfer can be achieved is not yet established.

Here, we formalise structure learning and transfer as a problem of learning covariance structure among variables and generalising it to (1) new combinations of features and to (2) a new response space. We ask how the representational geometry formed in the hidden layer of the network relates to its ability to perform the task. Previous literature suggests that the geometry of neural representation is important for effective knowledge generalisation (Bernardi et al., 2020; Ito et al., 2022; Johnston & Fusi, 2022). We developed a label prediction task and analysed the hidden layer representation of an MLP performing the task. The results allow us to make predictions for human participants performing the same task in a future study.

## Methods

The task uses 64 stimuli from the factorial combinations of 4 colours, 4 shapes, and 4 textures (Fig. 1A). The task involves learning to predict labels (e.g., pseudowords for a potential human cohort: Horst and Hout (2016)) assigned to the stimuli according to a hidden rule (Fig. 1B). Shape, colour and texture are pseudo-randomly assigned to be a constraining, relevant and irrelevant dimension for each participant. Stimulus labels depend on both the constraining and the relevant dimensions, but the structure (shared labels for 2 of the 4 features) only depends on the relevant dimension. There are two symmetric pairs of correlated relevant features (e.g. blue-yellow and red-green; see Fig. 1C).

An MLP with 20 hidden units was trained to perform the task in 30 runs with different initializations and stimulation sequences. We start from small initial weights (the "rich" regime, see Flesch et al. (2022)). In each run, the model goes through 5000 epochs with each being a training-transfer cycle. During training, the network only learns half the associations (Fig. 1C, top). This partial training setup allows us to examine the agent's generalisation of learned colour-pairing rules to new shape instances (new-instance transfer, Fig. 1C, bottom). Whilst all training trials have feedback, only a subset of transfer trials do. We choose these "anchor" trials (Fig. 1C bottom, grey trials) such that an agent who grasps the structure could transfer (zero-shot) to all other "test" trials (Fig. 1C bottom, white trials). In addition to this "full" training procedure, two control procedures disrupting structural learning are included: 1) the "anchor only" procedure that removes the partial training on half of the associations, and 2) the "train random" procedure in which the associations in partial training are randomised.

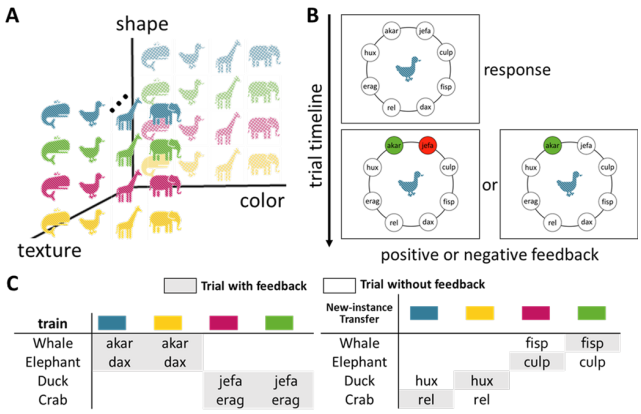


Figure 1: (A) Schematic illustration of the feature space and example stimuli. (B) The timeline for training and anchor trials. (C) Schematic illustration of stimuli-label associations in training and transfer.

## Results

The MLP successfully learns and generalises in the full training procedure (Fig. 2, top panels), but not in control training procedures (Fig. 2, middle and bottom panels). Under full training, the network reaches 100% accuracy in test trials in 28 out of 30 runs while performance in training and anchor trials converges in all runs.

To understand how the MLP might be solving the task, we study the neural geometry formed as it learns. The task has 8 labels. The high-dimensional representation that does not support generalisation is a regular 7-simplex formed by 8 equidistant points in seven-dimensional space (Fig. 3A, left). To generalise, coding directions of shape should be parallel between two colour pairs and vice versa (Fig. 3B, left). This yields a parallel tetrahedron representation (Fig. 3B, right) that encodes both abstract variables: 1) the shape of the stimulus, and 2) the pair of correlated colours that the stimulus belongs to.

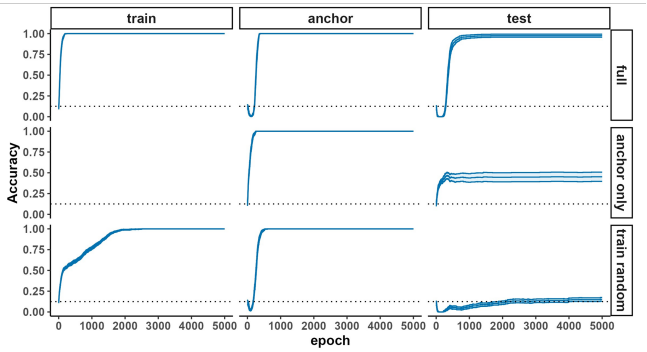


Figure 2: Performance curve of different trial types (train, anchor, test) in MLP trained under different procedures. Dashed black lines: chance levels. Ribbons indicate mean± standard error.

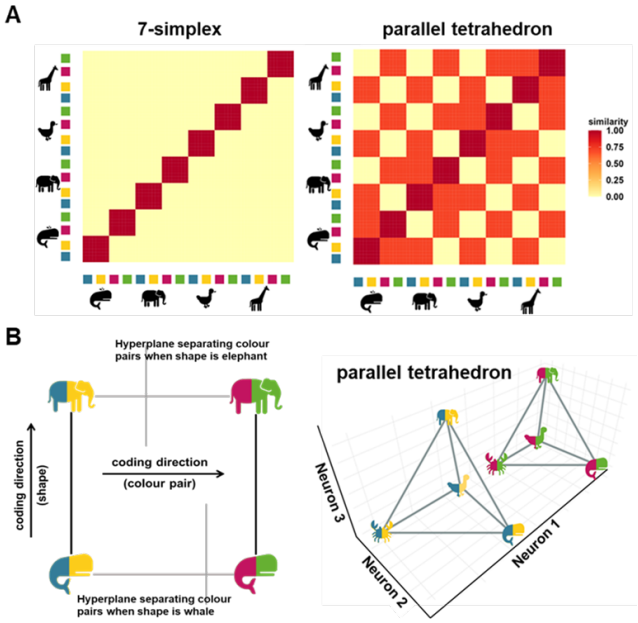


Figure 3: (A) The representational dissimilarity matrix of the 7-simplex model (left) and the parallel tetrahedron model (right). (B) Perfect generalisation requires coding directions of the same variable to be parallel.

To test our hypothesis, we quantify the level of abstraction in the hidden layer of the MLP with two measures. The first one is the correlation between the representational dissimilarity matrix (RDM) of the hidden layer and RDMs of different models (Fig. 3A). The second one is the parallelism score (PS) for each abstract variable (Bernardi et al., 2020; Ito et al., 2022). This is the cosine similarity between the coding directions of a given abstract variable. A PS of 1 indicates that the coding directions are perfectly parallel. Conversely, a PS of 0 indicates orthogonal coding directions.

Consistent with our prediction, after full training, hidden layer RDM is more similar to the parallel tetrahedron model than the 7-simplex model (Fig. 4A), and PS for both abstract variables is higher than initialization (Fig. 4B). Control training procedures lower the correlation with the parallel tetrahedron model, as well as PS (Fig. 4A-B).

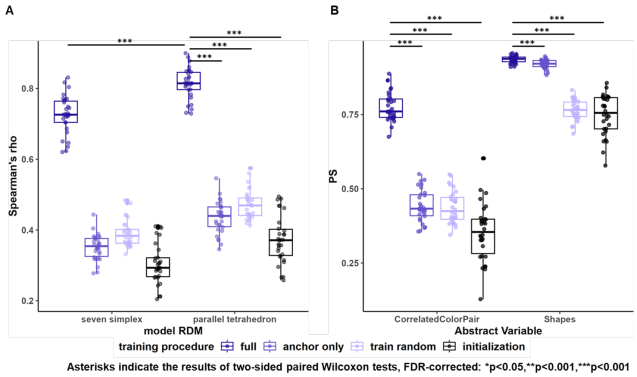


Figure 4: (A-B) Level of abstraction in observed representational geometry of MLP quantified by (A) correlation with model RDM and (B) parallelism scores.

## Discussion

Our simulations demonstrate the neural geometry that supports zero-shot generalisation of relational structure to new compositions of stimuli features. It represents the stimuli in a low-dimensional space spanned by a set of latent variables. This echoes previous work that demonstrated the formation of low-dimensional neural codes that represent the common structure of problems with learning in the orbitofrontal cortex (Zhou et al., 2021) in biological agents as well as in the hidden layers of neural networks (Johnston & Fusi, 2022). As a next step, we will collect behavioural data from human participants and compare it to that of neural networks.

## Acknowledgments

This work was supported by the European Research Council (award REP-725937 to C.S.) and Wellcome Henry Dale fellowship (223263/Z/21/Z to M.K.F.).

## References


Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967 e921. https://doi.org/10.1016/j.cell.2020.09.031

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270 e1211. https://doi.org/10.1016/j.neuron.2022.01.005

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behav Res Methods*, *48*(4), 1393-1409. https://doi.org/10.3758/s13428-015-0647-3

Ito, T., Klinger, T., Schultz, D. H., Murray, J. D., Cole, M. W., & Rigotti, M. (2022). Compositional generalization through abstract representations in human and artificial neural networks. arXiv:2209.07431. Retrieved September 01, 2022, from https://ui.adsabs.harvard.edu/abs/2022arXiv220907431I

Johnston, W. J., & Fusi, S. (2022). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *bioRxiv*, 2021.2010.2020.465187. https://doi.org/10.1101/2021.10.20.465187

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