Generalization of Covariance Structure in Human and Neural Network

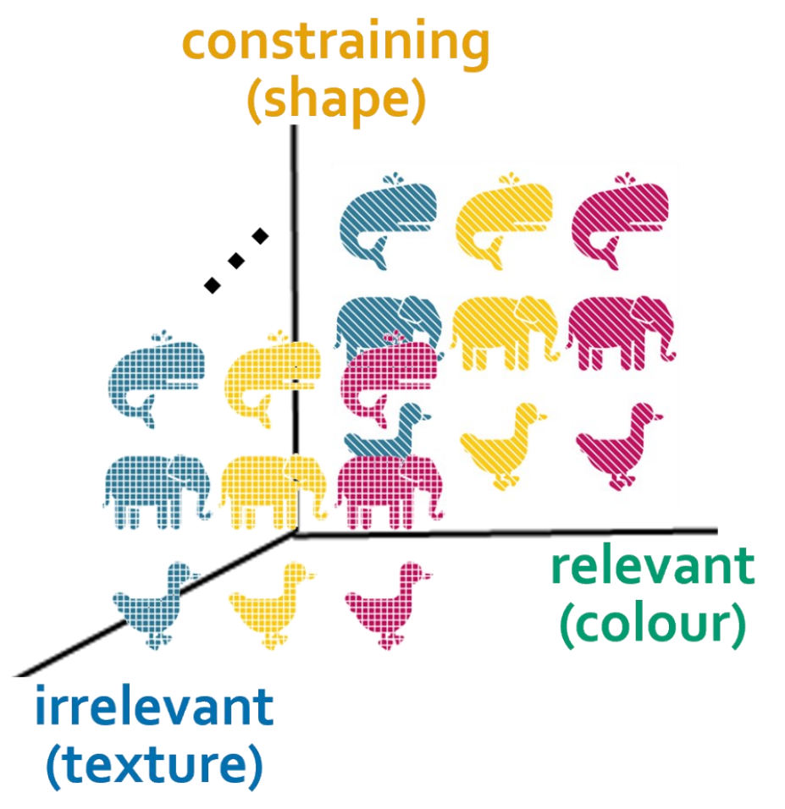
Zilu Liang, Miriam Klein-Flugge, Christopher Summerfield
Department of Experimental Psychology, University of Oxford

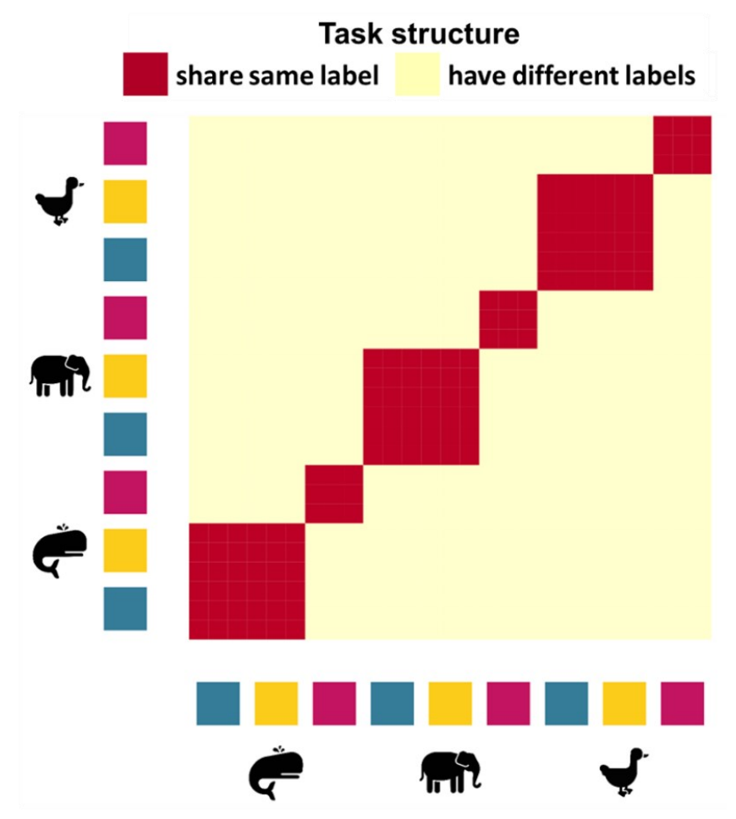
Introduction

- We designed a task to study the learning of covariance structure among feature variables and structure transfer to: 1) a new response space, and 2) new combinations of learned features.
- Using this task, we:
    - Investigate effective curriculum for structural learning and transfer in humans.
    - Test a theoretical model of representation that supports structure transfer in the task with neural network

The label prediction task (9-grid)

- constructed 27 stimuli from the factorial combination of 3 shapes, 3 colours, and 3 textures
- Participants learn to predict labels of the stimuli based on these features.
- **Task rule - One pair of correlated colours:** *blue and yellow stimuli of the same shape always have the same label.*





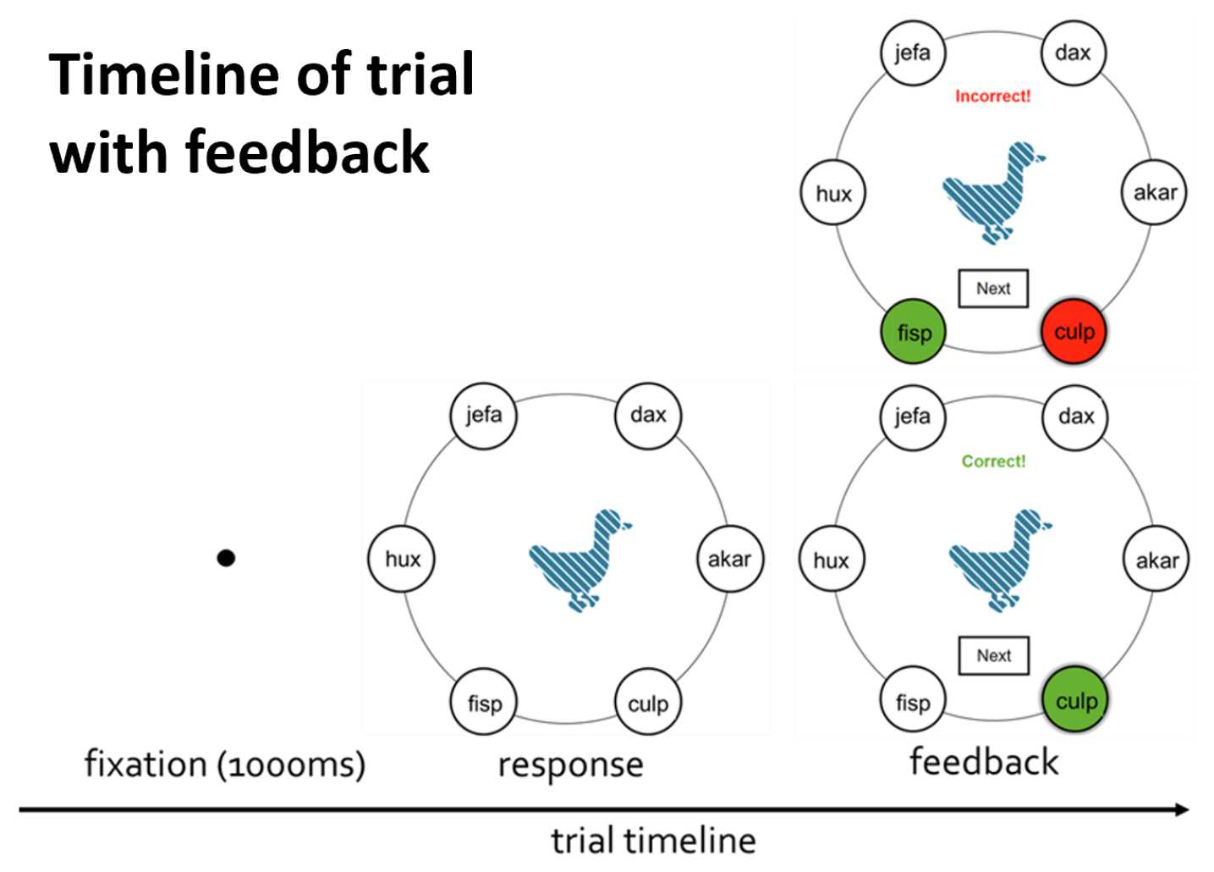
train	🟦	🟨	🟥
Whale	akar	akar	fisp
Elephant	dax	dax	culp
Duck	hux	hux	jefa

transfer	🟦	🟨	🟥
Whale	rel	rel	erag
Elephant	kern	kern	gip
Duck	lep	lep	blap

trial type
training | anchor — with feedback
test — no feedback

- Training: mapping stimuli to a set of 6 labels.
- Transfer: mapping stimuli to a new set of 6 labels.
- 10 training-transfer cycles (Interleaved train/transfer blocks)

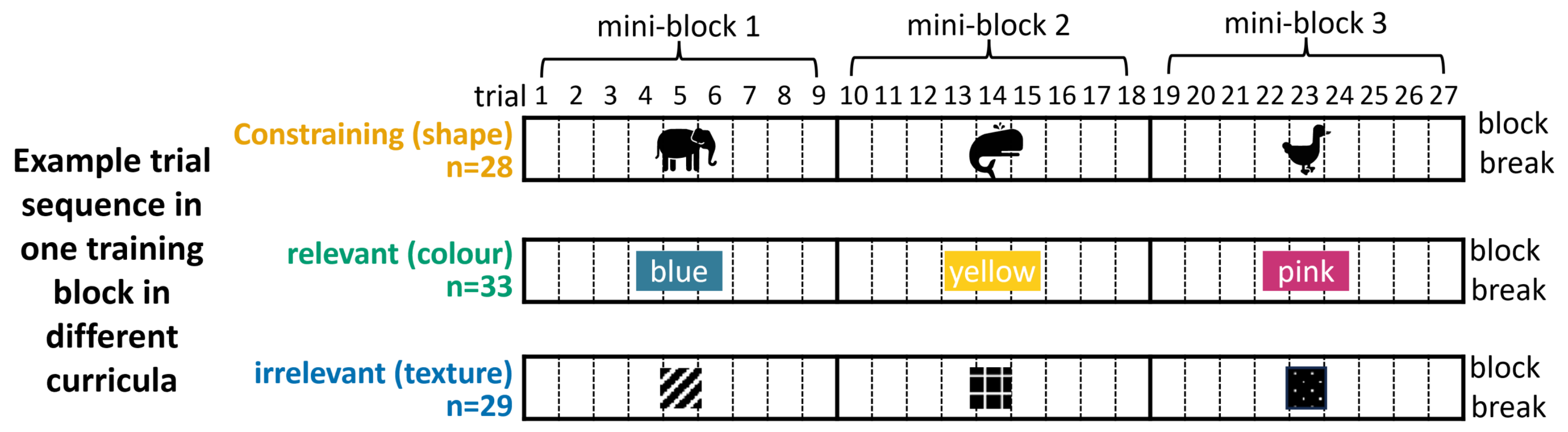
Timeline of trial with feedback



Human participants performance under different training curricula

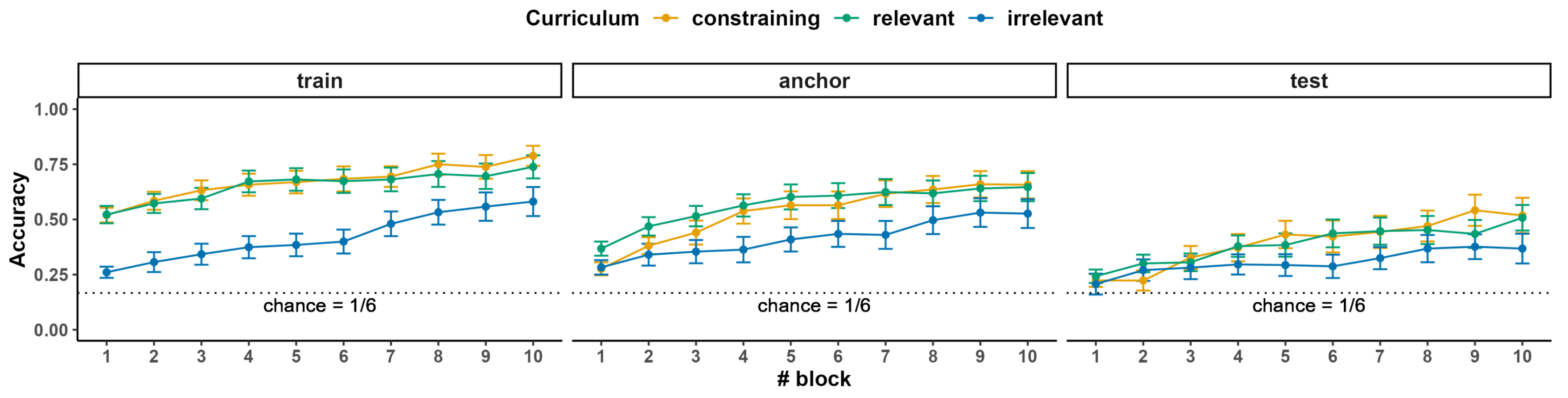
Training curricula

- Curricula that help to "start small" and form factorized representations (Flesch et al., 2018; Dekker et al., 2022): training curricula that introduce temporal autocorrelation (blocking) to task-relevant dimensions
- Compared to a training curriculum that blocks task-irrelevant dimension.


mini-block 1
mini-block 2
mini-block 3
trial 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
Example trial sequence in one training block in different curricula
Constraining (shape)
n=28
relevant (colour)
n=33
irrelevant (texture)
n=29
blue
yellow
pink
block break

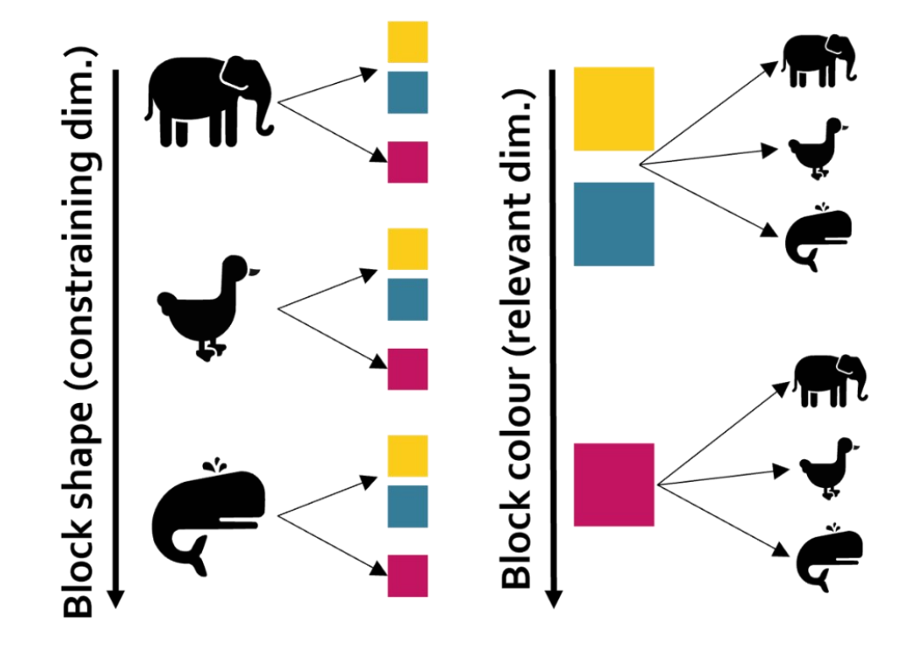
Result

- Participants' learning and transfer benefit from training curricula that blocks the task-relevant dimensions.

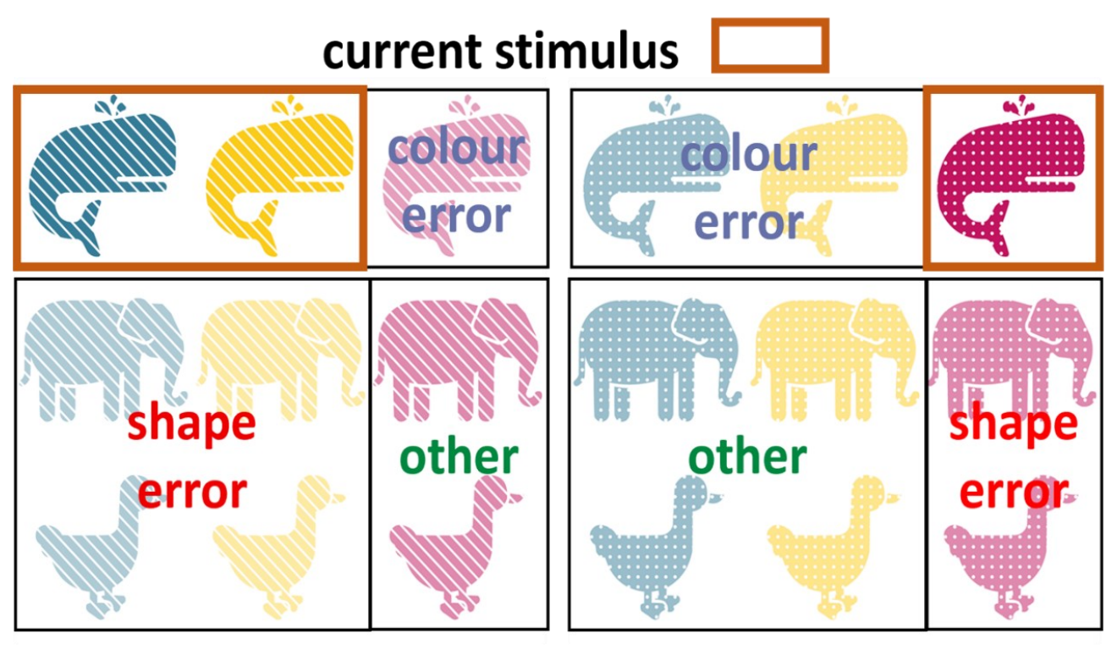


- The slow varying feature in the training stimuli contextualize stimuli-label associations
- Participants are less likely to be confused by associations from the same context than to be confused by associations from different contexts (Collins & Koechlin, 2013).

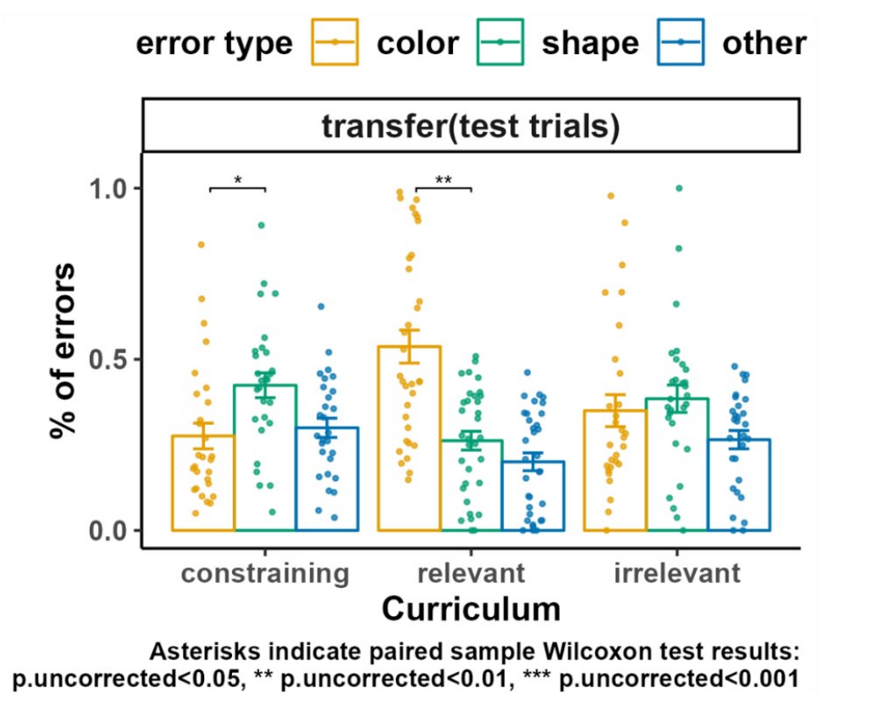
Slow-varying feature contextualize stimuli-label associations



Error classification



Error Distribution



Asterisks indicate paired sample Wilcoxon test results: p.uncorrected<0.05, ** p.uncorrected<0.01, *** p.uncorrected<0.001

If you have any questions, feel free to contact zilu.liang@psy.ox.ac.uk

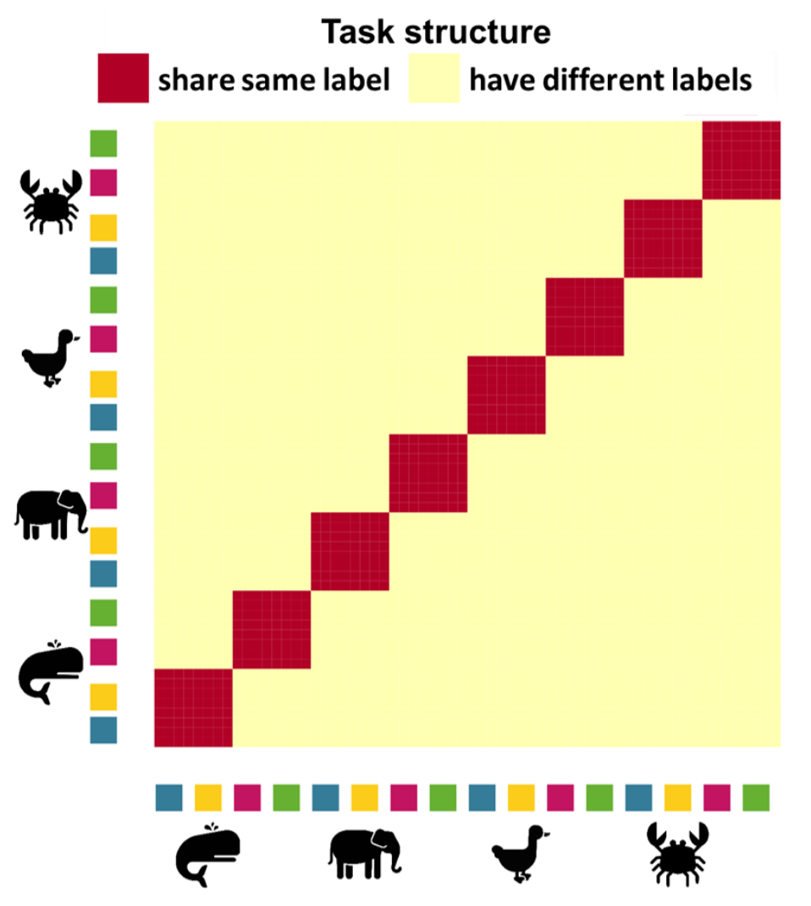
You can also find the online version of this poster and the abstract by scanning the QR codes:

Abstract | Poster | Task Demo

Extension to 16-grid label prediction task and neural network simulation

Extension to 16-grid task

- 64 stimuli generated from 4 shapes, 4 colours and 4 textures
- **two pairs of correlated colours:** *blue-yellow and red-green*



trial type: training | anchor | test

train	🟦	🟨	🟥	🟩
Whale	akar	akar		
Elephant	dax	dax		
Duck			jefa	jefa
Crab			erag	erag

New-instance Transfer	🟦	🟨	🟥	🟩
Whale			fisp	fisp
Elephant			culp	culp
Duck	hux	hux		
Crab	rel	rel		

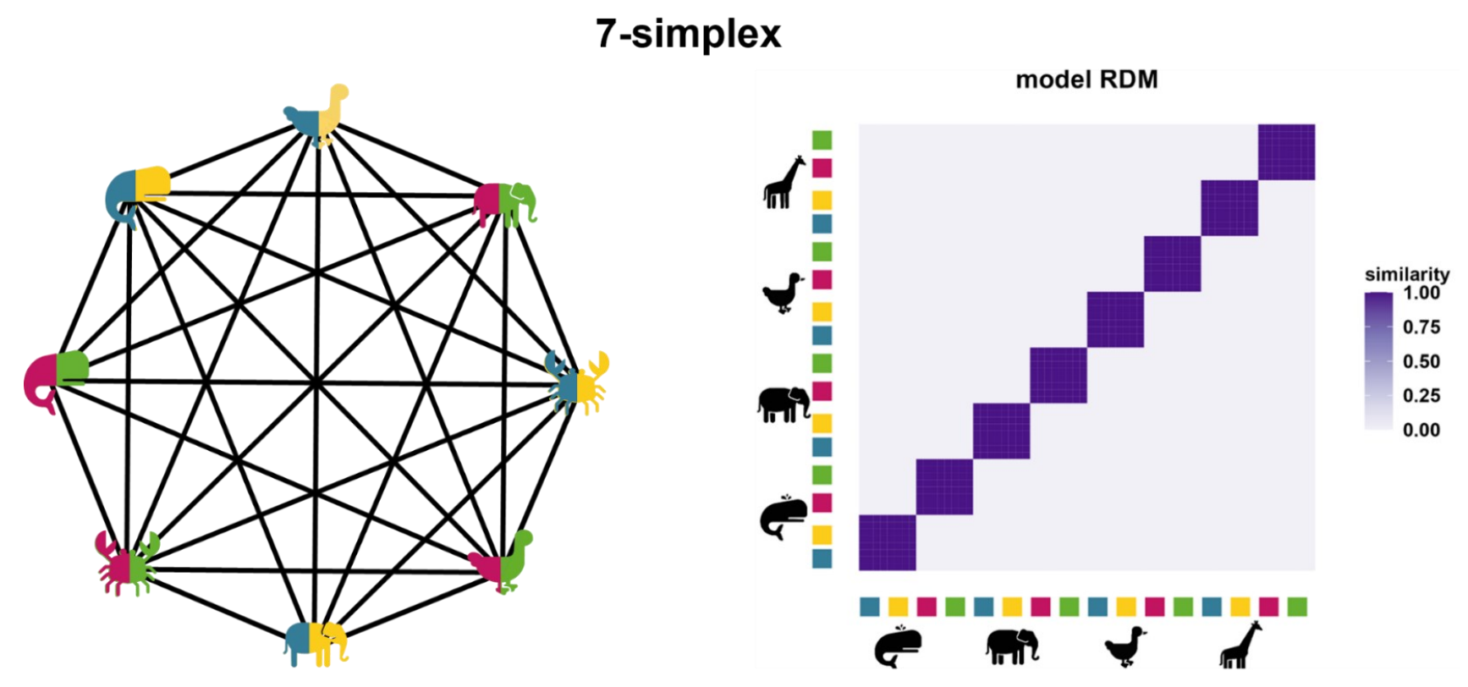
Neural Network Simulation

- Architecture: MLP with one hidden layer trained with SGD update
- A full training procedure same as behavioural study and two control training procedures disrupting learning

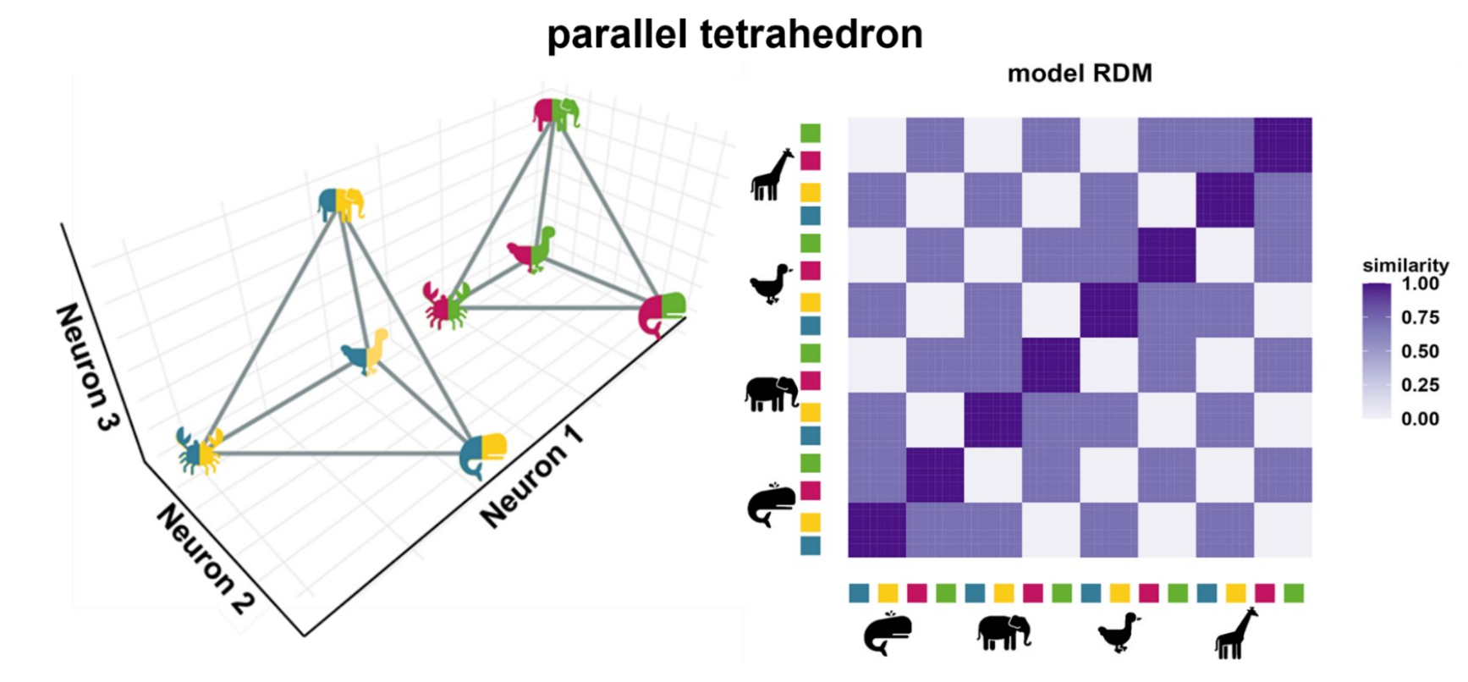
	trial type		
training procedures	training	anchor	Test
full	√	√	√
control 1: anchor only	X	√	√
control 2: train random	Train on random stimuli-label mapping	√	√

Hidden layer representation that supports structure transfer in neural network

Theoretical models of representation geometry



- **7-simplex model** : A high dimensional representation hinders the generalization of the relational structure.



- **Parallel tetrahedron model**: A low dimensional abstract representation that reproduces the latent structure supports knowledge generalization (Bernardi et al., 2020; Johnston & Fusi, 2022).

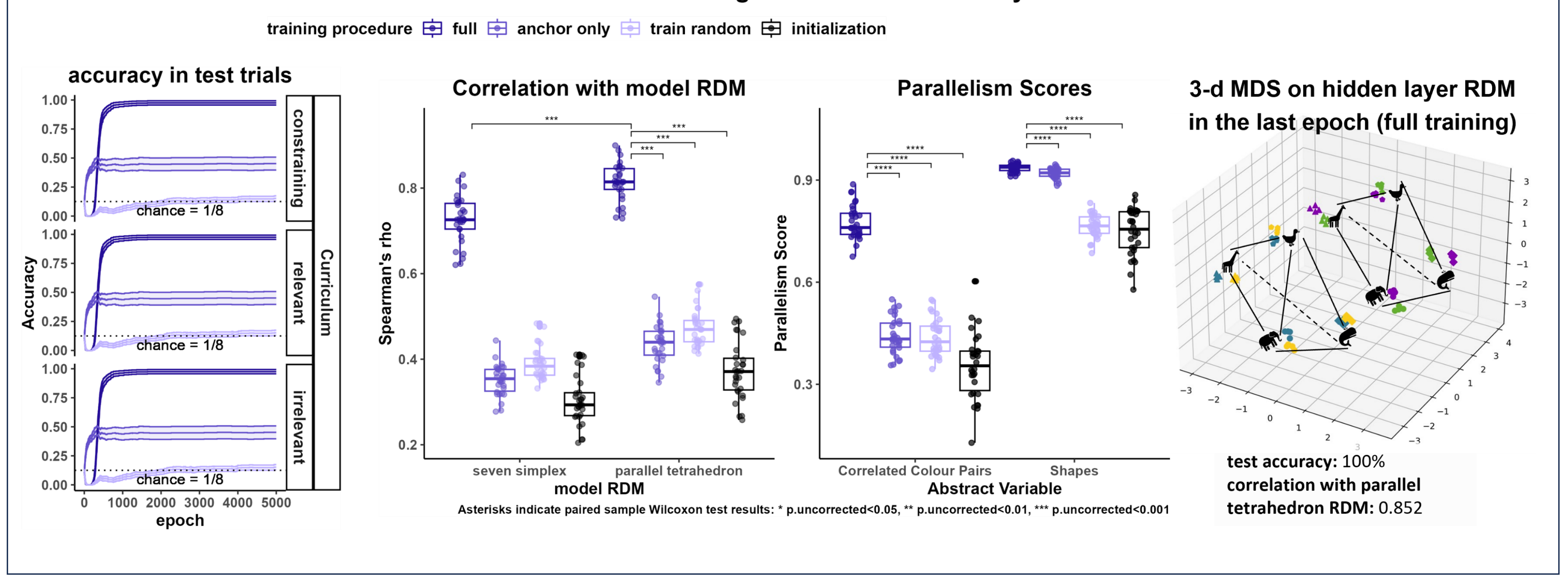
Quantifying abstraction in hidden layer representation

- Correlation between hidden layer RDM and model RDM
- Parallelism Score (Bernardi et al., 2020; Ito et al., 2022) of abstract variables: correlated colour pairs and shape

Result

- MLP with one hidden layer trained with SGD update is not sensitive to training curriculum
- After full training , the network learns the structure and can transfer
- After full training, hidden layer representation resembles the parallel tetrahedron model

16-grid task: one hidden layer MLP



Asterisks indicate paired sample Wilcoxon test results: * p.uncorrected<0.05, ** p.uncorrected<0.01, *** p.uncorrected<0.001

test accuracy: 100%
correlation with parallel tetrahedron RDM: 0.852

Discussion and Future directions

Summary

- **Structure learning and transfer in human is sensitive to training curricula.**
- **Neural geometry that represents the stimuli in a low-dimensional space spanned by a set of latent variables supports zero-shot generalization of relational structure to new compositions of stimuli features**

Next steps

- Modelling curriculum effect
- Behavioral study of the 16-grid task

Acknowledgements

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.).