



Connecting Concept Convexity and Human-Machine Alignment in Deep Neural Networks

Teresa Dorszewski¹, Lenka Tětková¹, Lorenz Linhardt^{2,3}, Lars Kai Hansen¹

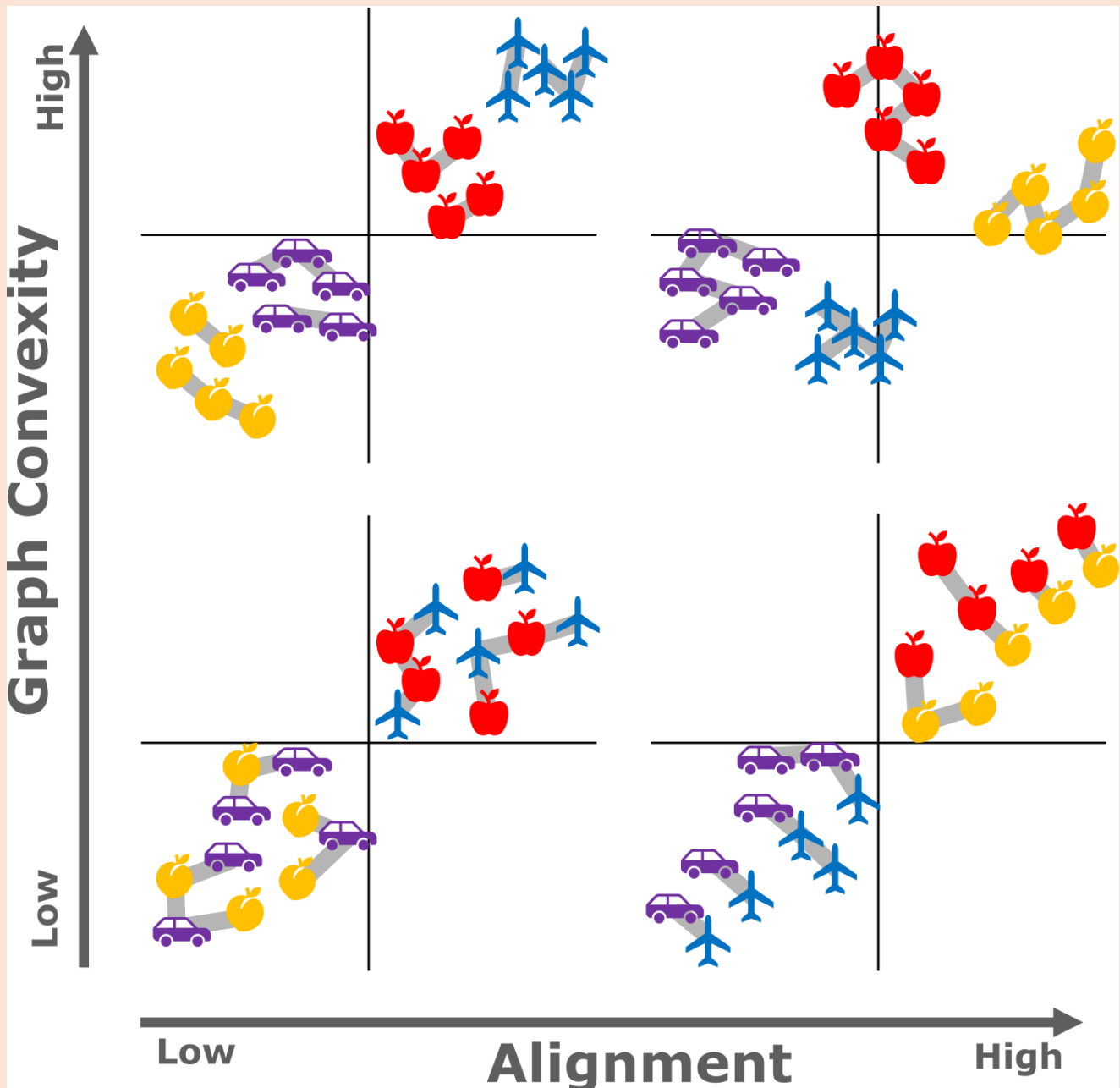
¹Technical University of Denmark, ²Technische Universität Berlin, ³Berlin Institute for the Foundations of Learning and Data – BIFOLD

Motivation

Measures of human-likeness of neural network representations are often derived from cognitive science and quantify a similarity in the representational structure of humans and neural networks. In recent years, a variety of measures has been proposed. In this work, we ask what the relationship between two of these measures, **graph convexity** and **odd-one-out accuracy**, is.

Any relationship of those measures is conceivable. What is their relationship in real DNNs?

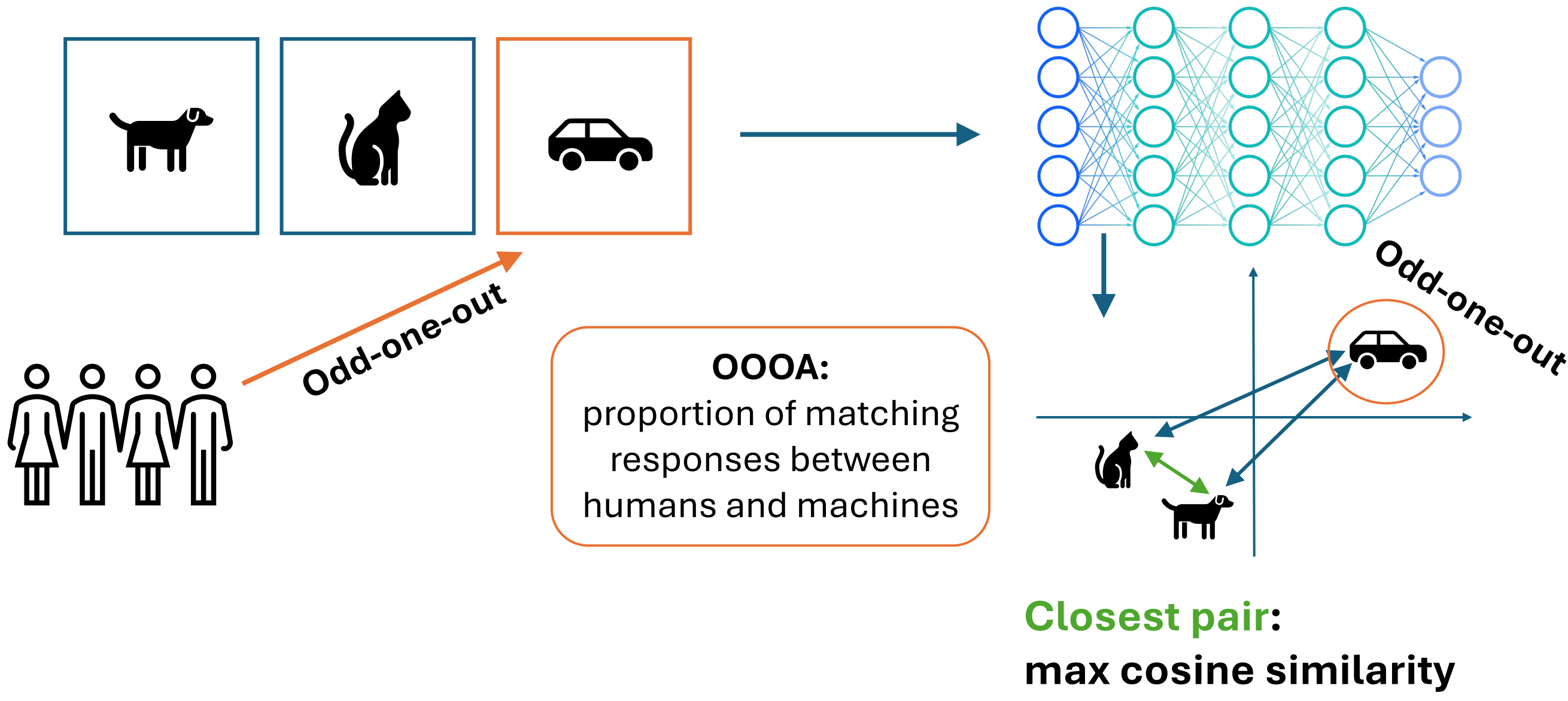
- High correlation may indicate that the same underlying property is measured.
- Low correlation opens the question of what aspects of human likeness each measure quantifies.



Methods

Human-Machine Alignment

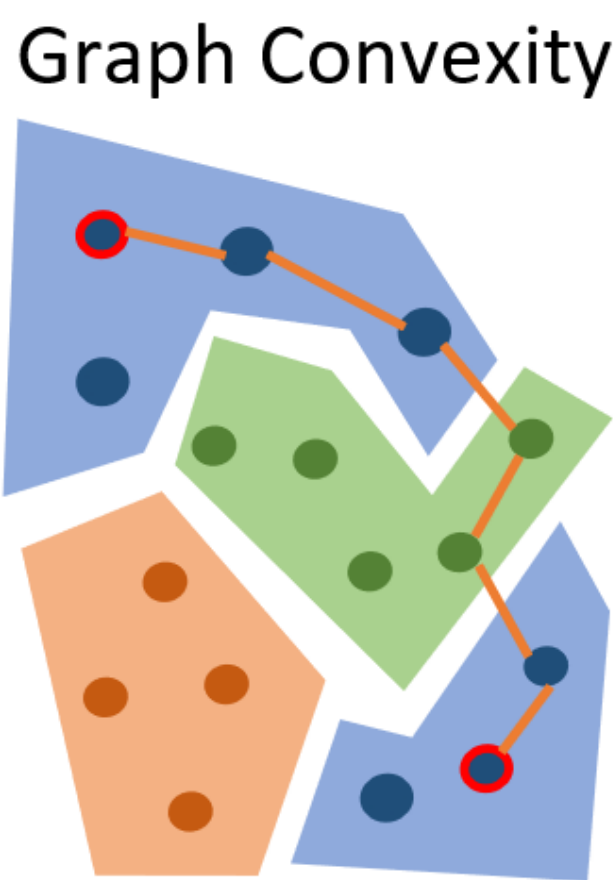
Odd-one-out accuracy (OOOA) [1]



Convexity

Graph convexity score [2]:

1. Extract representations of data
2. Build nearest-neighbors graph
3. For all pairs in one class, find shortest path
4. Graph convexity score: avg. proportion of points in path belonging to same class



Results

High Correlation (Fig c)

- High correlation of convexity and OOOA in early layers
- Different trends for pretrained and finetuned models

Observation on OOOA (Fig b)

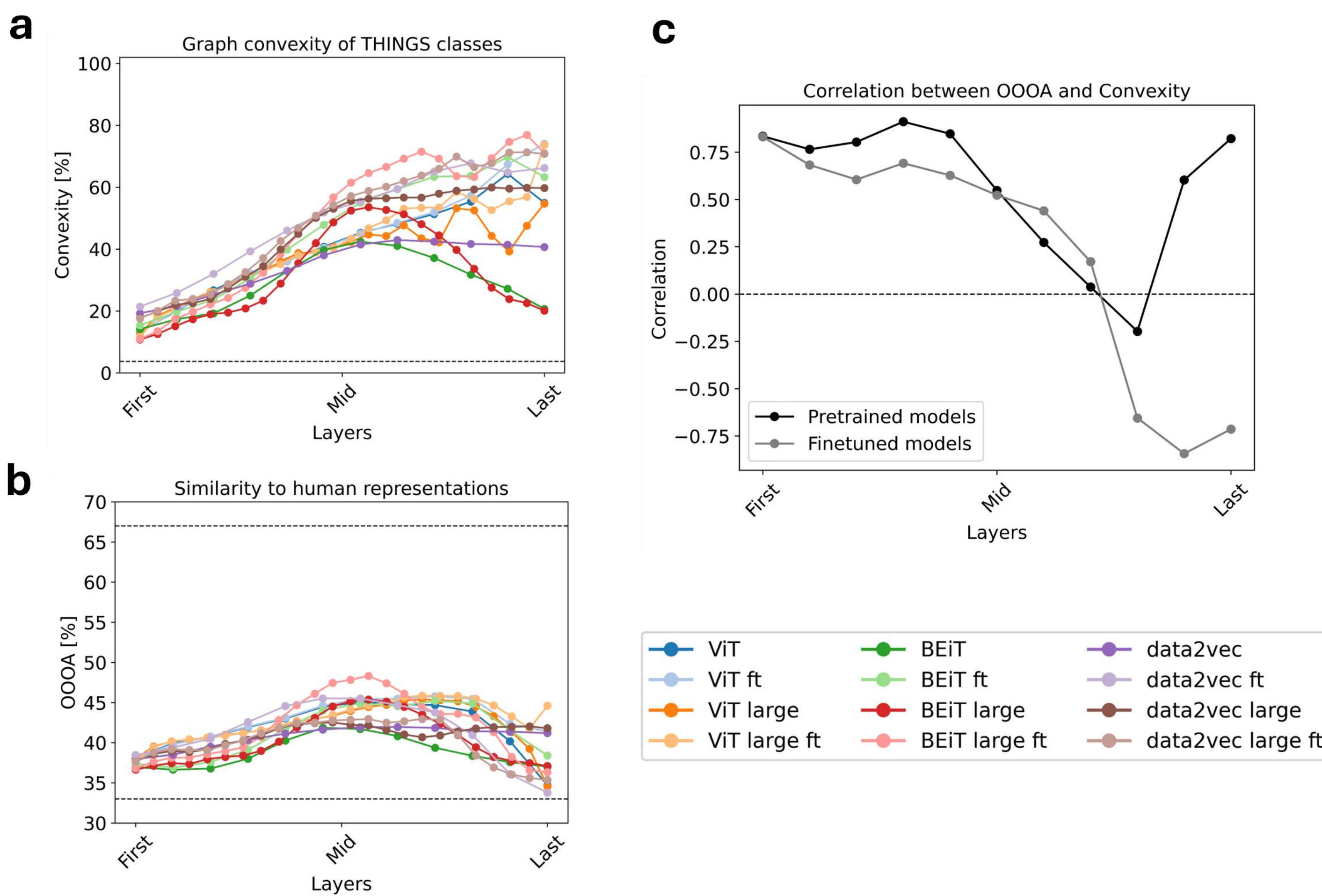
- Highest OOOA in the middle layers
- ! current research focuses on late layers

Impact of Increasing Human Alignment

- Increasing human-alignment using a latent space transform [3] generally increases convexity in pretrained models

Models + Layers	Change in OOOA	Change in Convexity
Pretrained First	+ 4.9 ± 1.4	+ 1.7 ± 1.4
Pretrained Middle	+ 9.7 ± 3.2	+ 0.4 ± 2.6
Pretrained Last	+ 13.7 ± 1.0	+ 3.1 ± 2.5
Finetuned First	+ 5.0 ± 1.4	+ 1.5 ± 1.4
Finetuned Middle	+ 9.1 ± 1.4	- 2.2 ± 1.6
Finetuned Last	+ 14.1 ± 4.4	- 2.5 ± 3.3

- Increasing convexity via finetuning does not necessarily lead to higher alignment



Conclusion and Outlook

Complex relationship between latent space geometries (convexity) and human-machine alignment (OOOA)

Future research directions:

- Enhancement of convexity and human-machine alignment → more aligned and generalizing models
- Investigation of factors influencing convexity and OOOA
 - (pre)training strategy
 - model architecture
 - training data