# Representation can dissociate from function and behaviour in task-optimised vision networks

**Marvin Theiss** ✉
University of Tübingen & IMPRS-IS

**Ben Fausten** ✉
University of Tübingen

**Felix A. Wichmann** ✉
University of Tübingen

Code: github.com/wichmann-lab/ccn-2026-dissociation

## Abstract

Building models that explain how biological brains process information is a central goal of computational neuroscience. A common way to evaluate such models is to ask how well they align with neural data. This alignment is routinely quantified by comparing model and brain representations using measures such as RSA and CKA. In doing so, high representational similarity is often taken as evidence of similar computation and behaviour. Here, we show that representational similarity can dissociate from both function and behaviour in task-optimised vision networks. Across architectures and datasets, including comparisons to macaque IT cortex, networks can match a target model's outputs, task performance, and trial-by-trial behaviour while exhibiting a wide range of representational similarity scores. Conversely, networks can closely match task performance and representational similarity, yet differ substantially in trial-by-trial behaviour. These findings show that representational similarity alone is insufficient evidence for shared computation and motivate stricter criteria for brain-model alignment.

## Introduction

Artificial neural networks (ANNs) are widely regarded as the most powerful models of brain computation in computational neuroscience. This view is articulated explicitly in the neuroconnectionist research programme (Doerig et al., 2023), which argues that ANNs occupy a "Goldilocks zone" of biological abstraction: abstract enough to remain analytically useful, yet close enough to biology to support meaningful comparison with brains. Within this framework, research alternates between model creation and evaluation. A key part of evaluation is alignment with neural data, often quantified by comparing model and brain representations using similarity measures such as Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) and Centred Kernel Alignment (CKA) (Kornblith et al., 2019).

Due to their central role in assessing alignment, representational similarity measures have been scrutinised in several ways, including sensitivity to confounds in stimulus structure and training data (Dujmović et al., 2024), dependence on their statistical properties (Gröger et al., 2026), and the fact that high representational similarity can coexist with poor behavioural alignment (Wichmann & Geirhos, 2023), to name but a few. Beyond concerns about confounds and metric properties, recent theory shows that function and representation can dissociate in deep networks: similar representations can arise from different functions, and identical functions from different representations (Braun et al., 2025), at least for *linear* ANNs. Here we show that an analogous dissociation also arises in task-optimised *nonlinear* vision networks widely used in computational neuroscience; worryingly, representation can dissociate from both function and behaviour. Across architectures and datasets, including comparisons to macaque IT cortex, networks can match outputs, task performance, and trial-by-trial behaviour while exhibiting a wide range of representational similarity scores. Conversely, networks closely matched in task performance and representational similarity can differ substantially in trial-by-trial behaviour. These findings challenge the interpretation of high representational similarity as evidence of shared computation and motivate stricter criteria for brain-model alignment.

## Methods

**Experimental paradigm.** We considered two experiment families. In all cases, networks were trained for image classification. In the first, networks were trained to match a fixed target network's outputs while realising a broad range of representational similarity scores, evaluated against the target model or neural data, in same- and cross-architecture settings. In the second, networks were trained to match the task performance and representational similarity of independently trained task-optimised baselines while differing in trial-by-trial behaviour. In all experiments, the optimised network shared a frozen feature extractor with the target. By fixing early visual features while allowing downstream representations to vary, we explicitly test how tightly function and behaviour constrain downstream representational geometry. Training used a stochastic augmented Lagrangian formulation (Dener et al., 2020, Algorithm 1).

**Quantifying similarity.** We distinguish representation, function, and behaviour. Representational similarity was quantified at the penultimate layer using distance-based RSA (Spearman or Pearson correlation of scaled squared

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026. 

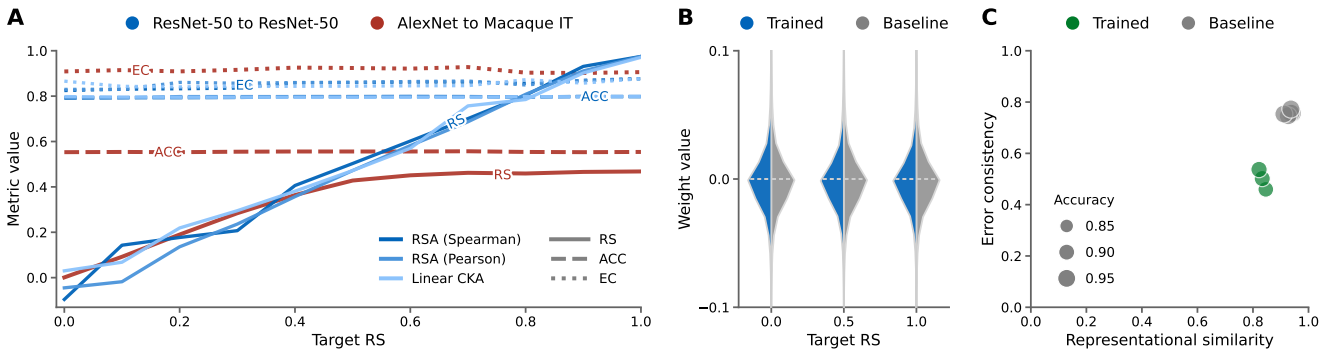


Figure 1: **Representation can dissociate from function and behaviour**. **(A)** Sweeping target similarity from 0 to 1 produces ResNet-50 networks spanning the full range of representational similarity scores (RS) to the target model, while top-1 accuracy (ACC) and error consistency (EC) remain at baseline levels (blue). A similar sweep against macaque IT responses yields a broad range of ANN–IT similarity scores saturating around 0.47 (red). **(B)** Weight distributions of networks from (A) closely match those of independent baselines. **(C)** Networks (green) can closely match representational similarity and task performance of independent baselines at substantially different EC levels.

Euclidean distances) and linear CKA. Functional similarity was quantified by mean squared error (MSE) between logits. Behavioural similarity was quantified by trial-by-trial error consistency (Geirhos et al., 2020).

**Networks, datasets, neural data.** Experiments used LeNet-5 (LeCun et al., 1998), AlexNet (Krizhevsky et al., 2012), VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), and ViT (Dosovitskiy et al., 2021) trained on Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky, 2009), and ImageNet (Deng et al., 2009). Brain comparisons relied on direct recordings from macaque IT cortex (Cadieu et al., 2014).

## Results

**Dissociating representation from function.** In our main experiment (ResNet-50 trained on ImageNet), we sweep target representational similarity scores from 0 to 1 in steps of 0.1 and recover networks at each target (Fig. 1A). Across this range, top-1 accuracy, logit MSE, and error consistency remain within the range observed across independently trained baselines. Thus, representational similarity can vary from near-zero to near-perfect alignment without any noticeable changes in function or behaviour. The same dissociation recurs across *same*-architecture comparisons spanning networks and datasets of different scales (LeNet-5 $\times$ Fashion-MNIST, AlexNet $\times$ CIFAR-10, VGG / ResNet / ViT $\times$ ImageNet). It also persists in *cross*-architecture comparisons between AlexNet and ResNet-50 trained on ImageNet.

The same dissociation emerges in comparisons to neural data. Using ImageNet-trained AlexNet, we obtain similarity scores to macaque IT data from Cadieu et al. (2014) ranging from 0.00 to 0.47 (Fig. 1A). Across this range, task performance and behavioural similarity remain within the range of independently trained baselines.

Despite their divergent internal representations, these networks are not pathological outliers. They remain indistinguishable from standard trained networks in task performance, logits, trial-by-trial behaviour, and weight variation, with learned weights staying within the range observed across independently trained baselines (Fig. 1B).

**Dissociating behaviour from representation.** The converse also holds. In AlexNet $\times$ CIFAR-10 experiments, we closely match representational similarity and top-1 accuracy to independent baselines while nearly halving error consistency (Fig. 1C). Thus, representational alignment need not imply similar behaviour either.

## Discussion

Our results demonstrate that representation is substantially underdetermined by function and behaviour in task-optimised vision networks. Across architectures, datasets, and comparisons with macaque IT cortex, networks can closely match outputs, task performance, and trial-by-trial behaviour, while exhibiting widely varying representational similarity to a fixed reference. This suggests that high representational alignment, on its own, is weaker evidence for shared computation than is often assumed. The converse result reinforces the same conclusion: even when task performance and representational similarity are matched, trial-by-trial behaviour can still vary substantially. Together, these findings argue against treating representational similarity as a sufficient criterion for brain-model alignment. Stronger alignment claims should instead rest on convergent evidence across representational, functional, and behavioural levels.

## Acknowledgments

This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Marvin Theiss. Felix A. Wichmann acknowledges funding from VIS4NN, Programa Fundamentos 2022, Fundación BBVA. Further, Felix A. Wichmann is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/2 – Project number 390727645.

## References

Braun, L., Grant, E., & Saxe, A. M. (2025). Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. *Proceedings of the 42nd International Conference on Machine Learning*, *267*, 5355–5382.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology*, *10*(12), e1003963.

Dener, A., Miller, M. A., Churchill, R. M., Munson, T., & Chang, C.-S. (2020, September 15). *Training neural networks under physical constraints using a stochastic augmented Lagrangian approach*. arXiv preprint.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., Van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *9th International Conference on Learning Representations*.

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2024). Inferring DNN-brain alignment using representational similarity analyses can be problematic. *ICLR 2024 Workshop on Representational Alignment (Re-Align)*.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

Gröger, F., Wen, S., & Brbić, M. (2026, February 16). *Revisiting the Platonic representation hypothesis: An Aristotelian view*. arXiv preprint.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. E. (2019). Similarity of neural network representations revisited. *Proceedings of the 36th International Conference on Machine Learning*, *97*, 3519–3529.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Krizhevsky, A. (2009, April 8). *Learning multiple layers of features from tiny images* (tech. rep.). University of Toronto.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations*.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, *9*, 501–524.

Xiao, H., Rasul, K., & Vollgraf, R. (2017, August 25). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*. arXiv preprint.