Parameter symmetries determine representational geometry in overparameterized nonlinear networks

Marvin Theiss ✉
University of Tübingen
& IMPRS-IS

Lukas Braun ✉
Allen Institute
for Neural Dynamics

Andrew M. Saxe ✉
Gatsby Unit & SWC,
UCL

Erin Grant ✉
University of Alberta
& Amii

Abstract

Representations are routinely used in machine learning, psychology, and neuroscience to probe the computations of biological and artificial systems. Yet it remains unclear to what extent computation constrains representation in artificial neural networks. One key obstacle is that these networks admit *parameter symmetries*: transformations of the parameters that preserve function exactly while reshaping representational geometry. Here we show that known parameter symmetries act on representations through just three primitives: *addition*, *duplication*, and *scaling*. This yields a closed-form descriptor of representational geometry as a sum of task-linked features and symmetry-induced noise, extending recent dissociation results from linear to nonlinear networks. Overall, our results delineate when representation can support inferences about computation, and when it cannot.

Computation and representation dissociate

A central question in representational alignment (Sucholutsky et al., 2025) is whether solving the same task ensures systems develop similar representations. One critical challenge is that even in linear networks, computation and representation can dissociate: networks can implement the same input-output function while expressing different representational geometries, and similarity measures quantifying their alignment can become misleading unless additional constraints hold at the level of implementation (Braun et al., 2025). The same issue already arises in very simple networks with nonlinear activation functions. First, one-hidden-layer networks trained on a nonlinear task can arrive at different zero-loss solutions with different representational geometries, demonstrating that task optimization alone does not constrain representation (Fig. 1B). Second, when networks are *overparameterized*, i.e., endowed with more neurons than are necessary to implement a given function, parameter symmetries (Fig. 1A) allow representational geometry to vary with a fixed input-output function. For example, a single neuron can be rescaled or split across several copies, and groups of neurons can cancel out. In this regime, representational geometries of overparameterizations of the same solution can be nearly uncorrelated, while those of solutions implementing different functions can be perfectly correlated (Fig. 1B). The resulting picture is a double dissociation: the same function can support very different representations, and representations of networks implementing different functions can look deceptively similar.

From symmetries to feature transforms

Despite the variety of parameter symmetries in overparameterized networks (Martinelli et al., 2024; Şimşek et al., 2021), their effects on representational geometry are simpler than their parameter-level descriptions suggest. At the level of hidden activations, all known symmetries reduce to combinations of three primitive transformations: *addition*, *duplication*, and *scaling* (Fig. 2A). Viewing the representational similarity matrix (RSM) as a sum of rank-one contributions from hidden features, i.e.,

$$\mathrm{RSM} = \sum_{j=1}^{N_h} \mathbf{z}_j \mathbf{z}_j^\top, \quad \mathbf{z}_j^\top = \sigma(\mathbf{w}_j^\top \mathbf{X} + b_j \mathbf{1}^\top) \in \mathbb{R}^{1\times P}, \quad (1)$$

where N_h is the layer width, $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function, $\mathbf{w}_j$ and b_j are the network's weights and biases, respectively, and $\mathbf{X}$ is the input matrix used to probe the network, these three feature transforms reveal a decomposition into a *task-linked* core and *symmetry-induced* nuisance structure (Fig. 2B). Duplication and scaling reweight similarity structure already present in the task-linked core, whereas addition contributes low-rank terms previously absent from the RSM. This decomposition into task-linked and nuisance components makes explicit how the same function can support very different geometries through a constrained set of transformations.

Discussion

Taken together, these results move toward a more principled account of when representational comparisons are well-posed. They also suggest a route to characterizing when links between representation and computation may be partially restored in nonlinear networks, extending the analysis of linear networks by Braun et al. (2025) to the nonlinear setting. More broadly, these results sharpen a central lesson: without accounting for degeneracy, representational similarity alone need not constitute evidence of shared computation.

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

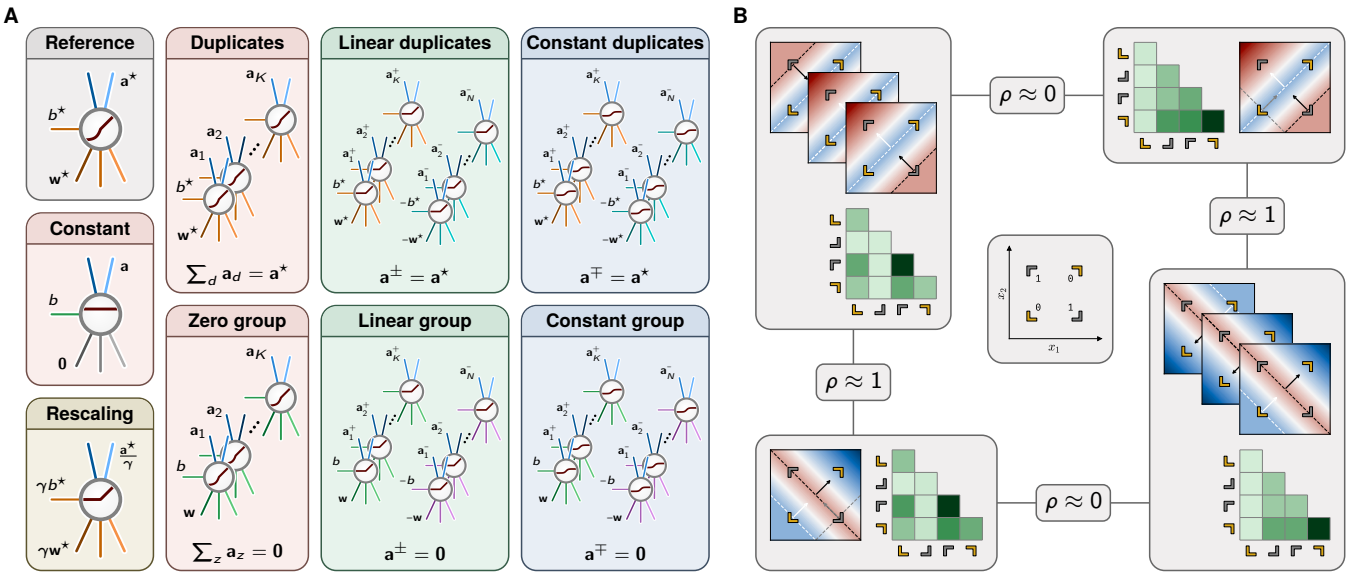

Figure 1: **Parameter symmetries can doubly dissociate function and representation**. **(A)** Parameter symmetries let a network implement the same input-output function through different parameterizations. Rescaling symmetries rescale *individual* neurons; other symmetries exploit redundancy among *groups* of neurons, either by distributing one neuron's computation across multiple copies or by coupling groups through algebraic structure in the activation function. **(B)** Two-neuron, one-hidden-layer ReLU networks trained on XOR (center) from small initialization yield six solutions, shown as heatmaps over input space with neuron decision boundaries overlaid. These form two families (top left and bottom right) solving the same task with different functions, which exhibit nearly uncorrelated RSMs. Moreover, overparameterizations of the *same* function can have uncorrelated RSMs (top and bottom), while *different* functions can have perfectly correlated RSMs (left and right): a double dissociation of function and representation.

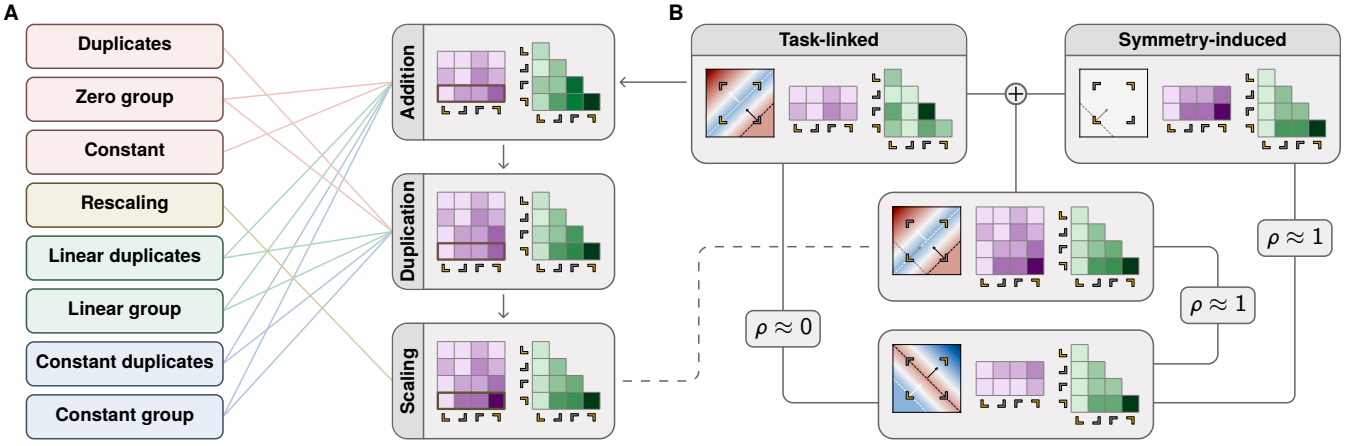

Figure 2: **Primitive transforms unify symmetries and isolate task-linked structure**. **(A)** Parameter symmetries decompose into three primitive feature transforms: *addition*, *duplication*, and *scaling*. Shown are their effects on hidden activations and RSMs when a zero group is introduced to a solution from panel B of Figure 1 and one neuron is then rescaled. **(B)** Decomposing the full RSM of the resulting overparameterized network (center) into *task-linked* and *symmetry-induced* components reveals the source of the dissociation observed in Figure 1. Whereas the task-linked component is nearly uncorrelated with the RSM of a solution from the opposite family, the symmetry-induced component is nearly perfectly correlated with it.

Acknowledgments

We thank Flavio Martinelli for helpful discussions that clarified aspects of the parameter symmetries cataloged in Martinelli et al. (2024). MT thanks Felix A. Wichmann for his continued guidance and support throughout MT's doctoral studies.

MT was supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). LB was supported by the Allen Institute. AMS was supported by a Schmidt Science Polymath Award, a Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z), and the Gatsby Charitable Foundation (GAT3850). EG was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2026-07018 and DGECR-2026-00191). AMS is a CIFAR Fellow in Learning in Machines & Brains, and EG is a CIFAR Azrieli Global Scholar in Learning in Machines & Brains and a Canada CIFAR AI Chair.

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.

Martinelli, F., Şimşek, B., Gerstner, W., & Brea, J. (2024). Expand-and-Cluster: Parameter recovery of neural networks. *Proceedings of the 41st International Conference on Machine Learning*, *235*, 34895–34919.

Şimşek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., & Brea, J. (2021). Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 9722–9732.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Cueva, C. J., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K., Oktar, K., Greff, K., Hebart, M. N., Cloos, N., Kriegeskorte, N., … Griffiths, T. L. (2025). Getting aligned on representational alignment. *Transactions on Machine Learning Research*.