

Privileged representational axes in biological and artificial neural networks

In the format provided by the
authors and unedited

Supplementary Information

1.1 Simulations and Control Experiments

1.1.1 Proposed rotation-sensitive measures are sensitive to both representational content and form

We first present simulations demonstrating that the proposed pairwise matching and soft matching similarity scores effectively discriminate between representations that differ in the tuning of individual neurons. Consider one representation $\mathbf{X} \in \mathbb{R}^{N \times k}$ formed from sampling k linearly independent latent factors $\{v_1, \dots, v_k\}$ drawn from a normal distribution, $\mathbf{X} = [v_1 \ v_2 \ \dots \ v_k]$. $\mathbf{X}$ can be viewed as the activation space of k units for N stimuli. Now, consider a second representation $\mathbf{Y}$ derived as a linear combination of some of these factors using a mixing matrix $\mathbf{M} \in \mathbb{R}^{k \times l}$: $\mathbf{Y} = \mathbf{X}'\mathbf{M}$, where $\mathbf{X}' = [v_1 \ \dots \ v_i \ u_1 \ \dots \ u_{k-i}]$ and $i \in \{1, \dots, k-1\}$. Here, the remaining latent factors $\{u_1, \dots, u_{k-i}\}$ are independently drawn. The fraction i/k governs the alignment of latent factors between $\mathbf{X}$ and $\mathbf{Y}$: higher values result in increased alignment. Furthermore, the sparsity of $\mathbf{M}$, measured by the number of non-zero entries in $\mathbf{M}$, $\text{nnz}(\mathbf{M})$, controls the similarity between the single-neuron tuning functions of $\mathbf{X}$ and $\mathbf{Y}$: a sparse $\mathbf{M}$ leads to an increased alignment of tuning functions. We varied the alignment of latent factors (representational content) and the alignment of tuning functions (representational form) and studied the sensitivity of different measures of representational similarity to these two types of alignment. Our empirical simulations, shown in Extended Data Fig. 2A–D, confirm that the pairwise matching and the soft matching scores accurately capture stricter notions of representational similarity, extending beyond information content to encompass the alignment of single-neuron tuning functions as well. In contrast, conventional measures such as centered kernel alignment (CKA), representational similarity analysis (RSA) or linear predictivity are only sensitive to the alignment of latent factors (i.e. the information content), regardless of whether these factors are disentangled in one representation (here, $\mathbf{X}$) and entangled in the other ($\mathbf{Y}$).

1.1.2 Measuring alignment change with axis rotation reliably reveals the existence of privileged axes

We next present simulation results demonstrating that the proposed method of quantifying alignment change with axis rotations is diagnostic of the presence of a privileged axis within a representational system. We generated synthetic data varying in the presence of privileged axes, by mixing a set of factors using mixing coefficients that varied in sparsity. We first generated a matrix $\mathbf{F} \in \mathbb{R}^{m \times k}$ containing responses of k linearly independent factors to m stimulus conditions. Each column of $\mathbf{F}$ is sampled from a standard normal distribution, with the columns orthogonalized by the Gram-Schmidt procedure. We then generated two "neural" representations $\mathbf{Z}_1 \in \mathbb{R}^{m \times n}$ and $\mathbf{Z}_2 \in \mathbb{R}^{m \times n}$ as linear mixtures of these factors:

$$\begin{aligned} \mathbf{X}_1 &= [\mathbf{x}_{ij}^1 \sim \mathcal{N}(0, 1)], \mathbf{X}_2 = [\mathbf{x}_{ij}^2 \sim \mathcal{N}(0, 1)] \\ \mathbf{M}_1 &= \mathbf{S} \odot \mathbf{X}_1, \mathbf{M}_2 = \mathbf{S} \odot \mathbf{X}_2 \quad \text{where} \quad \mathbf{S} \in \{0, 1\}^{k \times n} \\ \mathbf{Z}_1 &= \mathbf{F}\mathbf{M}_1, \mathbf{Z}_2 = \mathbf{F}\mathbf{M}_2 \end{aligned}$$

This way of generating $\mathbf{Z}_1$ and $\mathbf{Z}_2$ causes individual neurons to be mixtures of a subset of the factors in $\mathbf{F}$. The level of sparsity in the mixing matrices is controlled by the number of non-zero entries in $\mathbf{S}$. When $\mathbf{S}$ is highly sparse, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ have privileged representational axes, with the "neurons" tending to be aligned with individual factors or small mixtures of factors in $\mathbf{F}$. When

$\mathbf{S}$ is dense, each "neuron" in $\mathbf{Z}_1$ and $\mathbf{Z}_2$ is a linear mixture of all independent factors. In this case, there are not privileged axes, i.e. the two representations are no more aligned in their native axis than an arbitrary axis. Here, we empirically examine how the alignment between $\mathbf{Z}_1$ and $\mathbf{Z}_2$, determined using the proposed rotation-sensitive measures, varies as we gradually rotate the representational axis in either $\mathbf{Z}_1$ or $\mathbf{Z}_2$. We emphasize that standard measures, like centered kernel alignment (CKA), representational similarity analysis (RSA) or linear predictivity would not show any change with axis rotations. As shown in Extended Data Fig. 2E-F, when $\mathbf{S}$ is sparse, i.e., when the representations have privileged axes, the alignment between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ drops sharply with axis rotation. Conversely, when $\mathbf{S}$ is dense and the representations lack distinct axes, this trend is not observed. This observation confirms the effectiveness of our method in identifying the prevalence of a distinct or 'privileged' basis.

1.1.3 Changes in alignment with axis rotations are not driven by the changing statistical structure of neural/unit responses

In our proposed framework, we assess changes in alignment through axis rotations to determine if a representational system has a privileged axis. If axis rotations (sampled uniformly from the $\text{SO}(n)$ group) consistently reduce the alignment between different instances of representations from a given system, it suggests the presence of a preferred axis. However, an alternative explanation could be that axis rotations alter the statistical distributions of unit/neuron activations, causing the observed alignment loss. For example, as demonstrated in Fig. 4A in the main text, the response distributions of units in their native axes was sparser (heavier-tailed and right-skewed) than that for units in an arbitrary basis. This pattern was consistent across both artificial and biological representations in vision and audition.

To test whether the differing response statistics can explain the alignment drop between native and arbitrary axes, we first conducted a simulation. We constructed two representations $\mathbf{X}$ and $\mathbf{Y}$ with unit responses drawn from a heavy-tailed right-skewed distribution, similar to the distributions observed for neural responses in their native axes (Fig. 4A in the main text). Specifically, the responses were sampled from a gamma distribution with a shape parameter (k) of 10 and a scale parameter (θ) of 1 (Extended Data Fig. 2H). Since the responses were drawn independently for the two representation instances ($\mathbf{X}$ and $\mathbf{Y}$), there should be no privileged axes shared between the two representations. If a heavy-tailed right-skewed distribution were sufficient to cause alignment to change with axis perturbations, one should expect the alignment between $\mathbf{X}$ and $\mathbf{Y}$, as measured using our axis-sensitive measures (pairwise matching and soft matching correlation scores), to drop with axis rotations (since axis rotations alter the response statistics). However, as demonstrated in Extended Data Fig. 2H, even full axis rotations ($\alpha = 0$) result in a non-significant change in the alignment values using either measure. This demonstrates that the changes in the distribution of unit/neuron responses after axis rotations does not explain the drop in alignment values. We note that for this experiment, when computing the pairwise matching score, we evaluated the correlation score on the same data that was used for finding the pairwise mappings (i.e. without cross-validation). Cross-validation would have resulted in scores hovering around zero due to the lack of a common stimulus dimension.

We conducted another control experiment to further demonstrate that the loss in alignment with increasing axis rotations can not be merely explained by the changing statistical structural of neural responses with axis rotations. We computed the alignment between neural responses in macaque IT (from the ManyMonkeys dataset) using the same brain-to-brain alignment procedure described in the main text. However, this time, we shuffled the stimulus ordering independently for each monkey. This shuffling preserves the response statistics of individual neurons while disrupting

any common stimulus-driven structure.

We again observed that even full axis rotations ($\alpha = 0$) resulted in no significant change in alignment between the shuffled neural representations in different monkeys, using both the pairwise matching and soft matching measures (Extended Data Fig. 2I). This finding reinforces the conclusion that changes in alignment values with axis rotations are not driven by changes in the statistical structure of neural responses alone. As in the previous experiment, we did not perform cross-validation when computing the pairwise matching score to avoid scores hovering around zero due to the lack of a common stimulus dimension.

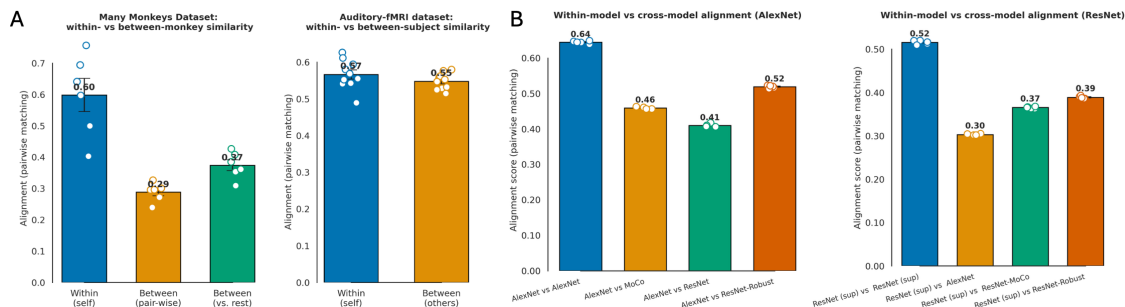
1.2 How ReLU Nonlinearities Disrupt Rotational Symmetry in Network Representations

Consider the representation of a hidden layer within a deep neural network that comprises an activation basis in $\mathbb{R}^n$. While $\mathbb{R}^n$ offers an infinite number of alternative bases, our objective is to identify those that could in principle be feasibly learned within the hidden layer representations. Mathematically, let $GL_n(\mathbb{R})$ represent the set of all invertible linear transformations in dimension n on a vector space over a field of real numbers $\mathbb{R}$. $GL_n(\mathbb{R})$ thus consists of all $n \times n$ matrices with real entries and non-zero determinants. Now, for ReLU nonlinearities, consider the subset $G_{ReLU} \subset GL_n(\mathbb{R})$ whose action on the network activations prior to ReLU has an equivalent transformation in $GL_n(\mathbb{R})$ after the pointwise application of this nonlinearity.

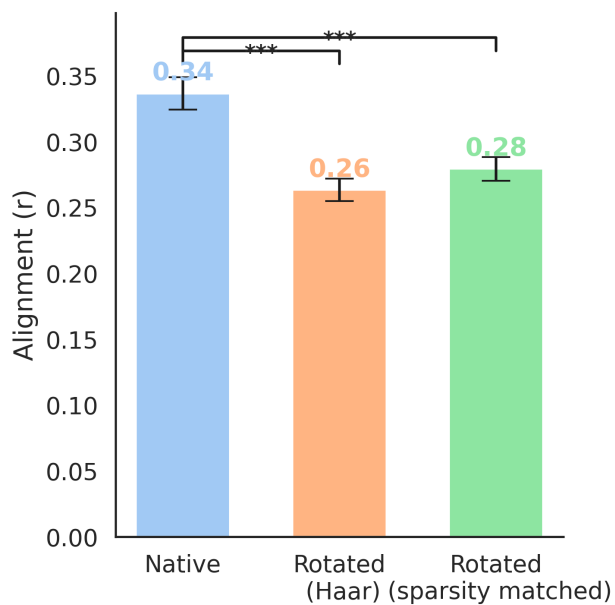
$$G_{ReLU} = \{T \in GL_n(\mathbb{R}) \mid \exists T' \in GL_n(\mathbb{R}) \text{ s.t. } ReLU \circ T = T' \circ ReLU\}$$

For linear transformations in deep network layers, the set of invertible linear transformations which when applied before the function allow for an equivalent transformation after their application, spans the entirety of $GL_n(\mathbb{R})$. Thus, a completely linear network should not have privileged axes. However, G_{ReLU} does not encompass $GL_n(\mathbb{R})$, but is substantially restricted to the set of matrices PD , where $P \in \mathcal{P}(N)$ is a permutation matrix and D is a diagonal matrix with positive entries. The integration of ReLUs into the network thus disrupts the existing symmetry and can endow the network with a preferred axis.

1.3 Additional Supplementary Figures

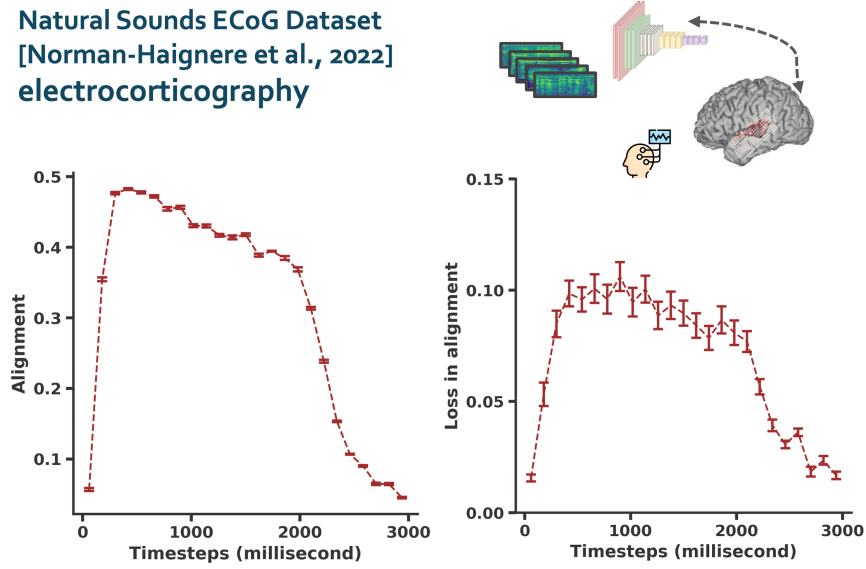


Supplementary Fig. S1: **Test-retest reliability of axis alignment within brains and within models.** (A) Within-brain alignment. Left: *ManyMonkeys* electrophysiology, estimated using split-half reliability across repeated stimulus presentations. Right: human auditory fMRI, where only three stimulus repetitions were available; here we used a 2-vs-1 split. To ensure comparability, between-subject alignment was recomputed using the same partitioning scheme. (B) Within-model alignment. Left: alignment between two independently trained seeds of AlexNet compared with AlexNet’s alignment to other models. Right: alignment between two supervised ResNet seeds compared with supervised ResNet’s alignment to other models. These results indicate that the representational axis structure stable within a model/brain is also shared with other models/brains. The larger within-between gap in the electrophysiology dataset likely reflects the sparse, randomly sampled neuron populations in each animal.

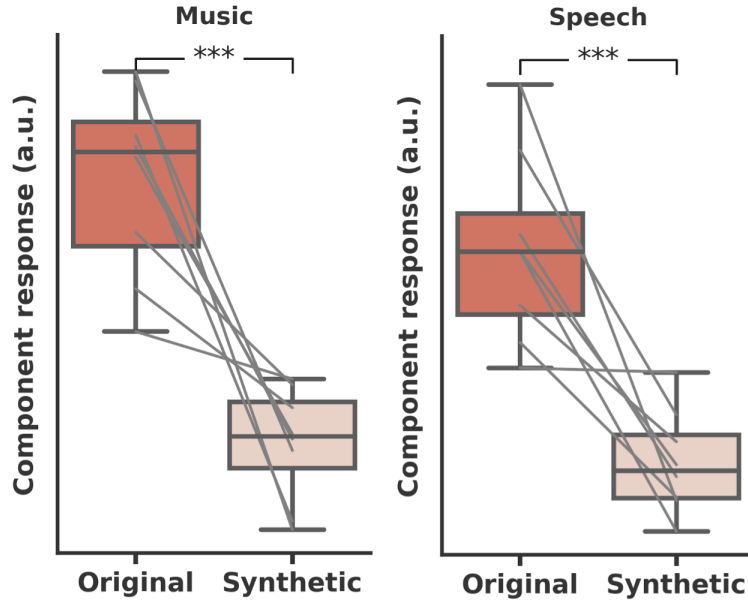


Supplementary Fig. S2: **Effect of axis rotations on pairwise alignment in single-unit recordings (ManyMonkeys dataset).** Alignment between pairs of monkeys is shown under three conditions: Native axes (no rotation), Haar-random rotation (axes rotated by a uniform draw from $SO(N)$), Sparsity-matched rotation (orthogonal transform optimized on the Stiefel manifold to preserve mean lifetime sparsity using GeoTorch). Error bars denote SEM over multiple random splits of stimuli. Both rotation controls (Haar-random and sparsity-matched) yield substantially lower alignment than the native axes, demonstrating that privileged axes cannot be explained by trivial sparsity-preserving coordinate changes.

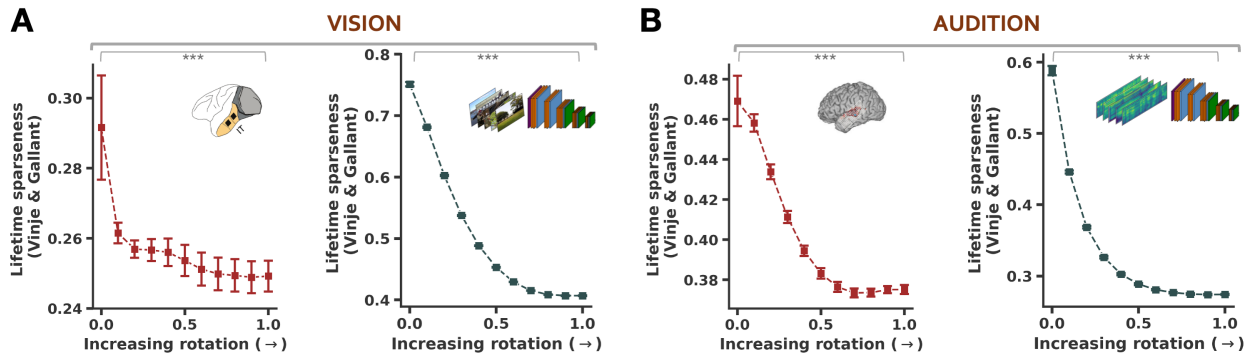
Natural Sounds ECoG Dataset
[Norman-Haignere et al., 2022]
electrocorticography



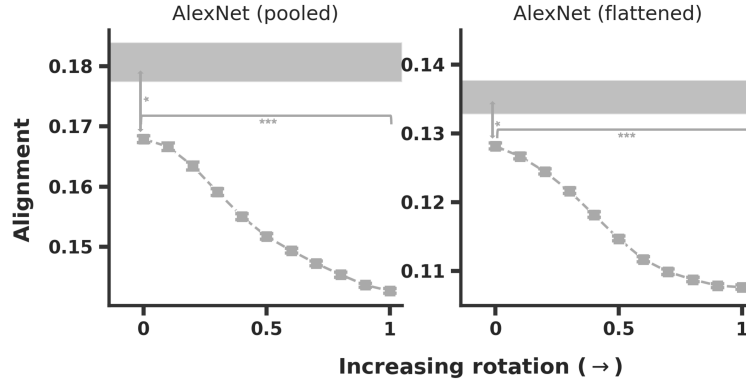
Supplementary Fig. S3: **Model-to-brain alignment as a function of time in the auditory domain** Left: Alignment of auditory deep convolutional neural network representations (self-supervised network with AudioNTT architecture) with biological neural representations in the auditory cortex (ECoG recordings) aggregated across different temporal windows following stimulus onset. Alignment is measured using the pairwise matching correlation score. Right: Difference in alignment of the auditory DCNN representation with neural representations in the human auditory cortex aggregated across different temporal windows at no rotation ($\alpha = 0$) and full rotation of the DCNN axis ($\alpha = 1$). Error bars are SEM over multiple splits of the data.



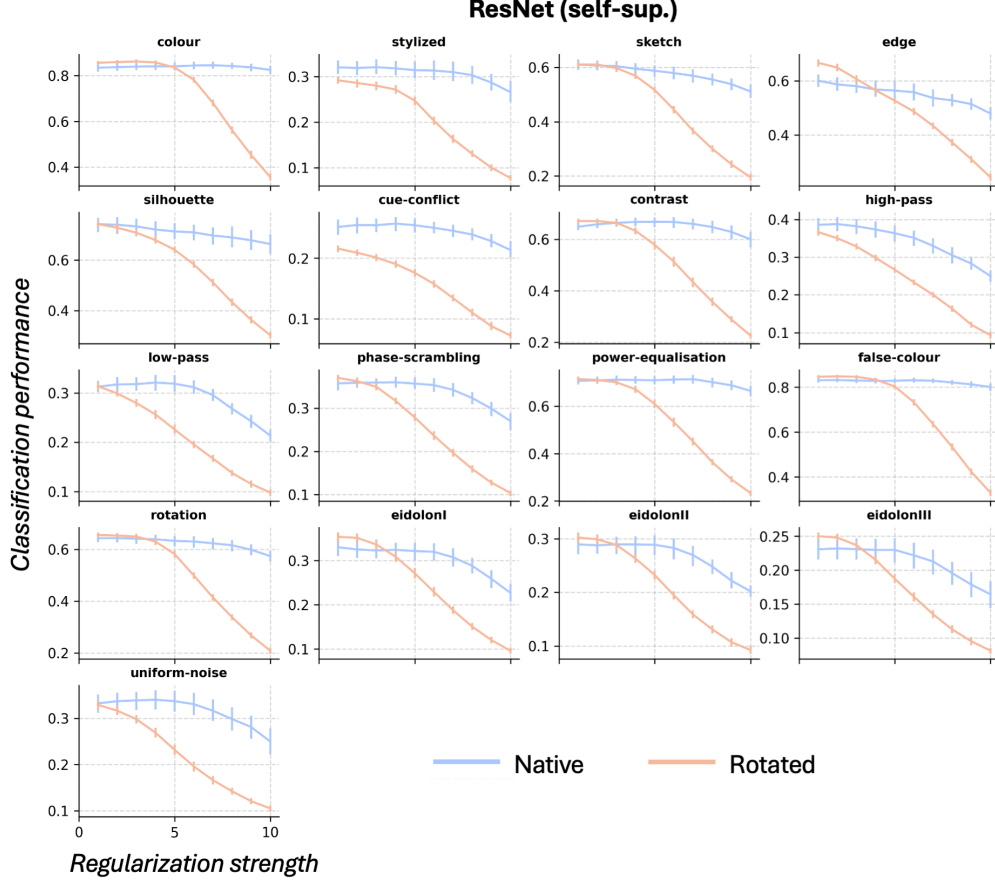
Supplementary Fig. S4: Responses of the music and speech-selective sparse components in the auditory self-supervised DCNN, as obtained by the NMF procedure, to a distinct set of 36 natural speech/music sounds and their corresponding modulation-matched controls. Signed t-test is used to evaluate whether the response to natural sounds was consistently greater than responses to corresponding modulation-matched sounds for these two categories. The modulation-matched synthetic sounds were generated to match the spectrotemporal modulation statistics of the corresponding natural music and speech sounds [1].



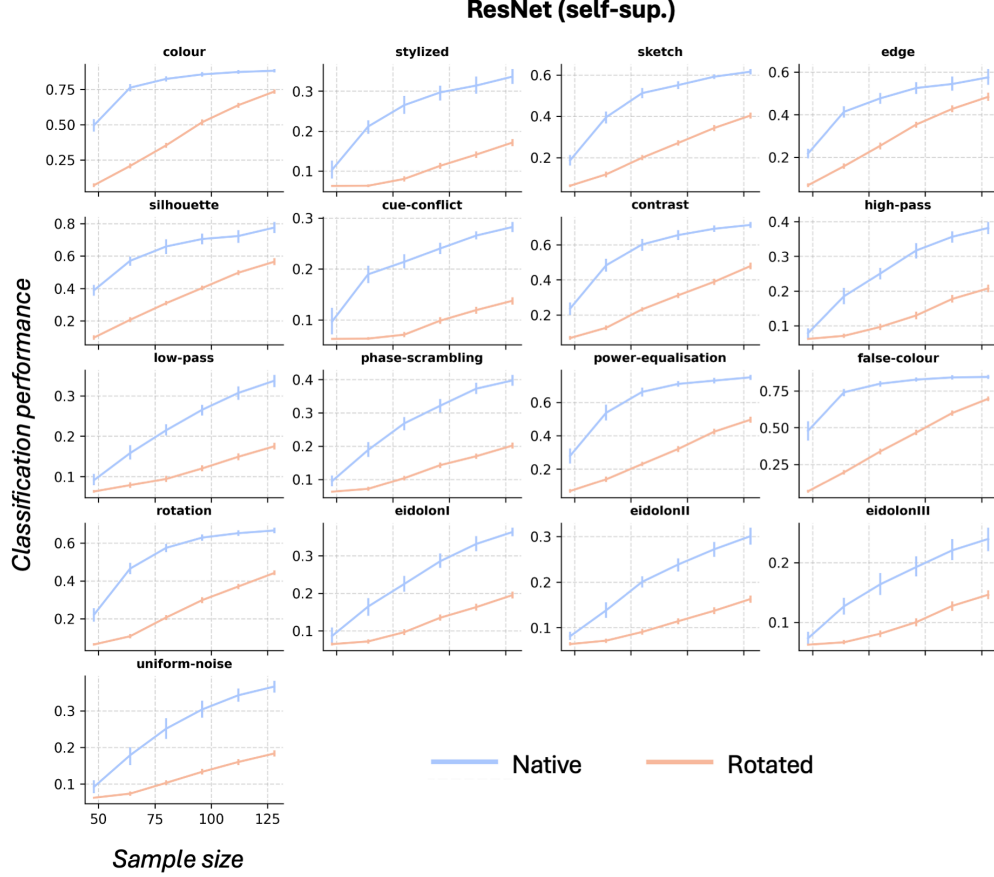
Supplementary Fig. S5: Lineplots show the mean lifetime sparseness, computed using the sparseness measure proposed in Vinje and Gallant (2000), across all units in the representation at different axis rotations. Values near 0 indicate low sparsity, and values near 1 indicate high sparsity. From left to right, the representations include the macaque IT neural responses (from the Macaque-NSD dataset), self-supervised visual DCNN activations, human auditory cortex responses (ECoG) and self-supervised auditory DCNN activations.



Supplementary Fig. S6: (Left) Alignment of the last convolutional layer of AlexNet with spatial pooling to neural representations in the high-level human ventral visual stream (Natural Scenes Dataset). This alignment is measured using the pairwise matching score and examined as a function of axis rotation (α) where $\alpha = 0$ indicates no rotation and $\alpha = 1$ indicates full rotation using a rotation matrix drawn uniformly from $SO(n)$. We also report the maximum possible value obtainable across all axis rotations (the “rotation ceiling”), determined using the ‘Optimal Axis Alignment’ procedure described in the text. (Right) Same as the left plot, but with flattened features instead of spatial pooling. Rotations are applied to the channels before flattening.



Supplementary Fig. S7: Few-shot generalization performance of penultimate-layer representations from a self-supervised ResNet (MoCo) under sparsity constraints. Sparse linear readouts (trained with l_1 -regularized linear regression) are evaluated on representations from a 16-class subset of the ImageNet dataset. The task is to predict entry-level category labels not only on the in-distribution (color) images but also on several out-of-distribution image types—including silhouettes, stylized images, and noise-distorted natural images—spanning a total of 17 datasets. All datasets (sourced from Geirhos et al. (2018) [2], Wang et al. (2019) [3], and Geirhos et al. (2021) [4]) use the coarser 16 labels from ImageNet, which include: Airplane, Bear, Bicycle, Bird, Boat, Bottle, Car, Cat, Chair, Clock, Dog, Elephant, Keyboard, Knife, Oven, and Truck. Each panel plots the classification accuracy (proportion correct) of the representation in both its native and rotated axes as a function of the sparsity penalty, with readouts trained on 5 images per category (80 images total).



Supplementary Fig. S8: Few-shot generalization performance of penultimate-layer representations from a self-supervised ResNet (MoCo) under sparsity constraints. Sparse linear readouts (trained with l_1 -regularized linear regression) are evaluated on representations from a 16-class subset of the ImageNet dataset. The task is to predict entry-level category labels not only on the in-distribution (color) images but also on several out-of-distribution image types—including silhouettes, stylized images, and noise-distorted natural images—spanning a total of 17 datasets. All datasets (sourced from Geirhos et al. (2018) [2], Wang et al. (2019) [3], and Geirhos et al. (2021) [4]) use the coarser 16 labels from ImageNet, which include: Airplane, Bear, Bicycle, Bird, Boat, Bottle, Car, Cat, Chair, Clock, Dog, Elephant, Keyboard, Knife, Oven, and Truck. Each panel plots the classification accuracy (proportion correct) of the representation in both its native and rotated axes as a function of increasing training sample size, with the sparsity penalty fixed at 10.

References

- [1] Sam V Norman-Haignere and Josh H McDermott. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS biology*, 16(12):e2005127, 2018.
- [2] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.
- [3] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.
- [4] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.