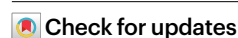


# Privileged representational axes in biological and artificial neural networks

Received: 3 April 2025

Accepted: 3 June 2026

Published online: 23 October 2026



Meenakshi Khosla<sup>1,2</sup>✉, Sudhanshu Srivastava<sup>1</sup>, Alex H. Williams<sup>3,4</sup>,  
Josh McDermott<sup>5,6</sup> & Nancy Kanwisher<sup>5,6</sup>

How do neurons code information? Recent work emphasizes the geometry and information content of population codes, using measures that are blind to the native tunings (or ‘axes’) of neural responses. But might these representational axes matter, with some privileged systematically over others? We developed methods to test for alignment of neural tuning across brains and deep convolutional neural networks (DCNNs). Across both vision and audition, both brains and DCNNs consistently favoured certain axes for representing the natural world. Moreover, the representational axes of DCNNs trained on natural inputs were aligned to those in perceptual cortices, such that axis-sensitive model–brain similarity metrics better differentiated competing models of biological sensory systems. We further show that coding schemes that privilege certain axes can facilitate task performance, reduce downstream wiring costs and improve generalization. These results motivate a new framework for understanding neural tuning in biological and artificial networks and its computational benefits.

The quest to unravel the neural basis of perception is guided by two opposing paradigms. One view holds that neural representations are best understood at the level of an entire population of neurons<sup>1–6</sup>. Under this view, the focus shifts to the geometry of the population representation, and the role of individual neurons is considered less central. Consequently, the coordinate axes of the representation are often treated as arbitrary, such that a representational system, denoted as A, is viewed as equivalent to another system, B, derived by rotating the axes of the original representation, A. This view is supported by the argument that changing the coordinate axes preserves both linearly decodable information and representational geometry (the relative distances among patterns of response in representational space). An alternate view contends that the actual tunings of neurons (that is, representational axes) matters<sup>1,7–9</sup>, and System A and System B could differ importantly from each other (Fig. 1a).

How can one resolve the apparent conflict between these two viewpoints and decide which framework is more suitable for understanding the neural code? Two questions must be answered. One is empirical:

Do brain regions have a preferred way of implementing representations at the level of the tunings of individual neurons? The second question is theoretical: Why might the implementation of a representation at the neuron level matter, that is, what are the consequences of choosing a particular representational axis?

Decades of experimental investigations have provided loose evidence bearing on the empirical question. Individual neurons, and sometimes whole brain regions, have been reported to encode complex, meaningful and identifiable features of our external world such as the presence of a navigational boundary, face or speech sound<sup>10–13</sup>. These and many other selective response profiles are consistently found across individuals and sometimes even across species<sup>11,14–19</sup>, raising the possibility that some brain systems might favour certain representational coordinate systems over others. However, this body of evidence remains predominantly qualitative and leaves open the possibility that beyond a few specific tuning functions, axes could mostly differ across brains. Thus, an important question remains unresolved: Are neural representational axes arbitrary, or do neural populations in

<sup>1</sup>Department of Cognitive Science, University of California San Diego, San Diego, CA, USA. <sup>2</sup>Department of Computer Science and Engineering, University of California San Diego, San Diego, CA, USA. <sup>3</sup>Center for Neural Science, New York University, New York, NY, USA. <sup>4</sup>Center for Computational Neuroscience, Flatiron Institute, New York, NY, USA. <sup>5</sup>McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>6</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: [mkhosla@ucsd.edu](mailto:mkhosla@ucsd.edu)

fact code for information in more systematic ways, such that responses of individual neurons might be aligned across brains?

Many data analysis practices in modern neuroscience implicitly assume that representational axes are irrelevant and are therefore poorly suited to address this question. For example, linear decoding models are insensitive to arbitrary rotations of the neural representational axes: imposing a rotation on the representations will simply transform the optimal decoding weight matrix and lead to no change in predictive accuracy. Similar assumptions are built into the most popular methods for testing the alignment of representations across brains and artificial networks, including representational similarity analysis (RSA)<sup>20</sup>, canonical correlation analysis (CCA)<sup>21</sup>, centred kernel alignment (CKA)<sup>22,23</sup>, Procrustes shape distance<sup>23</sup> and linear predictivity scores<sup>24</sup>. Thus, while such comparative analyses of biological and artificial neural networks have been instrumental in explaining the ‘content’ of neural codes (that is, what information is linearly accessible from a representation and its representational geometry), they fail to explain the ‘form’ of neural codes (that is, how individual neurons encode information) due to their invariance to rotations of response patterns. Moreover, these analytical tools implicitly assume a downstream readout that has access to the entire neural population and is capable of computing linear transforms. However, the accuracy of this assumption is yet to be determined.

Here we show that the view of neural and artificial coding that disregards the importance of representational axes is incomplete, as it cannot account for a striking empirical fact: certain axes of representation are ‘privileged’, that is, systematically aligned across brains and across DCNNs trained to process natural visual and auditory inputs. To assess this alignment, we introduce measures for quantifying alignment between representations that overcome the constraints imposed by existing rotation-insensitive comparison metrics. These axis-dependent alignment measures reveal alignment across conspecifics’ perceptual cortices, across different DCNNs and between perceptual cortices and respective DCNNs in their native axes. Remarkably, in most cases, this alignment approaches the maximal alignment achievable through any axis rotation. These findings held in both vision and audition, and our preliminary analyses suggest that they also extend to more cognitive representations including language. We also provide theoretical and empirical insights into the emergence of privileged axes in networks, demonstrating that certain axes enable superior task performance in architectures with pointwise nonlinearities such as rectified linear units (ReLUs). Further, our alignment measures, which map individual neurons or voxels to individual network units, establish a more stringent match between brains and artificial networks, and

better discriminates model fit to the brain compared with currently standard methods such as regression-based model prediction. Finally, we demonstrate that the axes privileged in these coding schemes convey computational benefits in terms of downstream wiring costs, sparse coding and the capacity for few-shot generalization under biologically realistic constraints.

## Results

### A formal framework for assessing privileged axes

Consider a representation  $\mathbf{X} \in \mathbb{R}^{M \times N}$  formed by collecting the responses of  $N$  neurons to  $M$  stimuli. The activation space corresponds to  $\mathbb{R}^N$ . The native basis  $\{e_i\}_{i=1}^N$  (with  $e_i$  defined such that its  $i$ th entry equals 1 and all others equal 0) is but one of an infinite number of possible bases for  $\mathbb{R}^N$ . Any geometry-preserving change of basis can be constructed by sampling a rotation matrix (that is, a special orthogonal matrix)  $\mathbf{Q} \in \mathbb{R}^{N \times N}$  and applying the corresponding linear transformation to  $\mathbf{X}$ , yielding the transformed representation  $\mathbf{XQ}$ . Critically, this change of basis preserves the information ‘content’ of a representation (that is, what can be linearly decoded) as well its ‘geometry’ (that is, distances between points). This raises the question: are the native basis and the axes it defines arbitrary?

Here we address this question using a formal framework for assessing ‘whether’ and to what extent a representational system (for example, a brain area or DCNNs) has privileged representational axes. We test for such axes by examining multiple instances of a given representational system (for example, different individuals of a species, or different instantiations of artificial neural networks) and asking whether there is any consistent relationship between their axes.

In what follows, we first introduce axis-dependent measures of representational alignment. We then use them to measure the alignment between instances of brains or artificial neural networks and assess how the alignment changes under axis perturbations. If the pair of representation instances are consistently more aligned in their native basis than in the perturbed bases, this provides evidence that the representational scheme has privileged axes. We then contrast these values with the maximal alignment achievable under all possible bases. If the alignment in the native basis approaches the maximal alignment under all possible bases, that provides further support for privileged axes.

We note that by ‘privileged axis’, we refer exclusively to a consistent inclination towards certain bases as evident across different system instances. Privileged axes need not be interpretable, or have other distinguishing properties. In later sections, however, we consider the properties of the privileged axes in brains and models and whether those might explain why they are privileged.

**Fig. 1 | Sensory cortices exhibit privileged representational axes. a**, Cartoon schematic presenting two representational frameworks in three dimensions. Each stimulus elicits a unique response pattern in the three units. In System A, the cardinal axes correspond to individual neurons, where neurons 1, 2 and 3 exhibit selective responses to faces, scenes and bodies, respectively. In System B, we depict the representation achieved by arbitrarily rotating the coordinate axes of System A. In this configuration, the units do not exhibit clear tuning to any of the categories. **b**, Cartoon schematic presenting the framework used to probe the prevalence of privileged axes in sensory cortical representations. Coloured dots represent individual stimuli and axes  $x_1, \dots, x_N$  and  $y_1, \dots, y_N$  represent responses of individual voxels or neurons. Neural representations in different conspecifics are compared with axis-alignment measures sensitive to single-neuron tuning. For each representation pair, we change the axes of one representation along a smooth  $SO(n)$  manifold and quantify how the pairwise matching alignment changes with the amount of axis rotation. **c**, Empirical ‘rotation floor’ of pairwise matching alignment under three simulated data regimes ( $M = 3,000$  stimuli;  $N \in \{200, 400, 600, 800, 1,000\}$ ; 5 random seeds). High-dim, isotropic Gaussian data:  $\mathbf{X} \in \mathbb{R}^{M \times N}$  drawn i.i.d. from  $\mathcal{N}(0, 1)$ , with  $\mathbf{Y} = \mathbf{X}$ . Even after applying random orthogonal transformations to  $\mathbf{X}$ , alignment with  $\mathbf{Y}$  remained at  $-0.2$ , indicating a non-zero random baseline even for fully unstructured Gaussian data. Low-dim, clustered latent modules:  $\mathbf{X}$  composed of  $K = 6$  independent low-dimensional

latent modules (module dim = 2; within-module noise = 0.1),  $\mathbf{Y} = \mathbf{X}$ . Random rotations yielded an alignment value close to 0.6 across all tested feature counts  $N$ . Mid-dim, regime in which half of the features participated in clustered latent modules and the remaining half were isotropic. This yielded an intermediate rotation floor of  $-0.37$ , confirming that alignment values vary smoothly with the intrinsic dimensionality and structure of the data. **d–f**, Alignment between conspecifics neural responses in human ventral visual cortex (**d**), macaque IT (**e**) and human auditory cortex (**f**). This alignment was measured using the pairwise matching score and examined as a function of axis rotation ( $\alpha$ ) 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. Histograms depict the distribution of the difference in pairwise matching scores between the native and rotated basis across all neurons/voxels, with the bars colour coded on the basis of difference values. Significance tests were two-sided paired  $t$ -tests comparing alignment values at no rotation ( $\alpha = 0$ ) with those at full rotation ( $\alpha = 1$ ), as well as with the optimal alignment values observed across all rotation angles. Error bars and shaded regions indicate s.e.m. across  $n = 10$  stimulus train/test splits per comparison. \*\*\* $p < 0.0001$  (maximum annotated  $p = 1.8 \times 10^{-6}$ ); NS, not significant.

**Quantifying representational alignment.** We introduce two complementary measures of axis alignment (Extended Data Fig. 1). The first measure is an intuitive quantification of direct pairwise correspondence between tuning functions in different representations. The second measure is symmetric and benefits from the desirable properties of metric spaces in its theoretical framework.

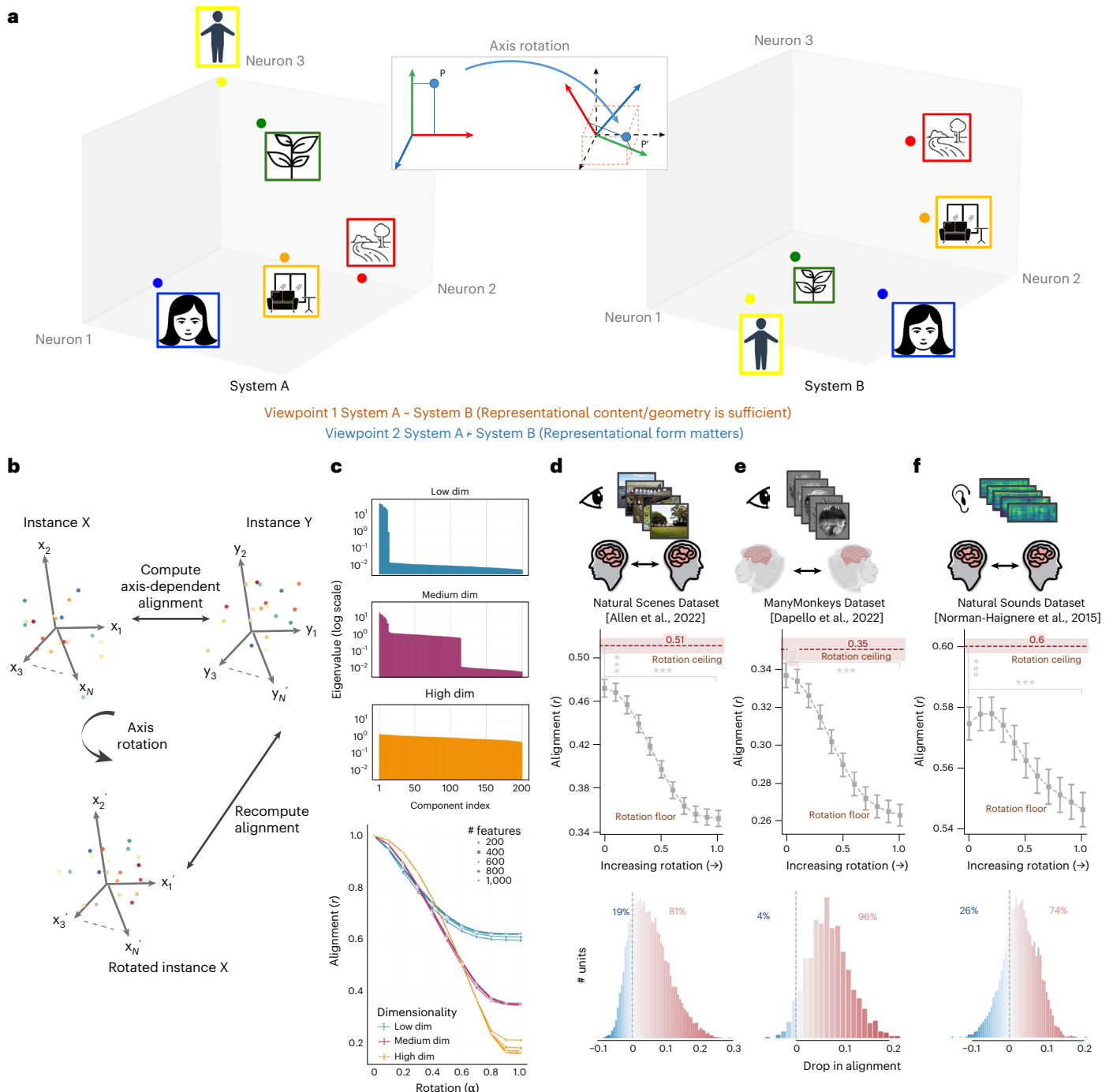
**Pairwise matching score.** Our first measure is a direct measure of similarity between units in different representations. Here, a unit in a biological representation corresponds to the smallest unit of measurement in brain recordings (a voxel in functional magnetic resonance imaging (fMRI) data, a single neuron in neurophysiology data), whereas a unit in artificial networks corresponds to a single ‘filter’ in the convolutional layer of a DCNN with its response averaged over all spatial locations.

Consider two matrices  $\mathbf{X} \in \mathbb{R}^{M \times N_x}$  and  $\mathbf{Y} \in \mathbb{R}^{M \times N_y}$ , which represent the activations of  $N_x$  and  $N_y$  units, respectively, for  $M$  inputs (for example, stimuli). The proposed measure proceeds in two stages:

- In the first stage, we find the optimal assignment matrix  $\mathbf{A}^* \in \mathbb{R}^{N_x \times N_y}$  with entries  $a_{ij}^*$  that maps each unit in  $\mathbf{X}$  to a single unit in  $\mathbf{Y}$  by optimizing a correlation-based objective. The optimization problem is defined as follows:

$$\{a_{ij}^*\}_{i=1..N_x, j=1..N_y} = \arg \max_{a_{ij}} \sum_i \sum_j \rho(\mathbf{X}_i^{\text{train}}, \mathbf{Y}_j^{\text{train}}) a_{ij},$$

$$\text{s.t.} \quad a_{ij} \in \{0, 1\}, \sum_j a_{ij} = 1$$
(1)



Intuitively, the optimization procedure aims to find, for each unit in  $\mathbf{X}$ , the unit in  $\mathbf{Y}$  that exhibits the maximum response correlation with that unit in  $\mathbf{X}$  over a training stimulus set. The constraints on  $a_{ij}$  ensure that each unit in  $\mathbf{X}$  is matched to exactly one unit in  $\mathbf{Y}$  (although each unit in  $\mathbf{Y}$  can have multiple units in  $\mathbf{X}$  mapped to it). As a result, some units in  $\mathbf{Y}$  may remain unmatched after the assignment.

- In the second stage, using the unit matches obtained in the first stage, we compute the correlation between each unit in  $\mathbf{X}$  and its corresponding matched unit in  $\mathbf{Y}$  on held-out data and then average these values across all matched pairs to derive the final alignment metric

$$s_{\text{pairwise}}(\mathbf{X}^{\text{test}}, \mathbf{Y}^{\text{test}}, a_{ij}^*) = \frac{1}{N_X} \sum_i \sum_j \rho(\mathbf{X}_i^{\text{test}}, \mathbf{Y}_j^{\text{test}}) a_{ij}^* \quad (2)$$

Note that this measure is asymmetric by design. The choice to allow repetitions with a bipartite semi-matching assignment (that is, allowing a single unit in  $\mathbf{Y}$  to be matched to multiple units in  $\mathbf{X}$ )<sup>25</sup> instead of a strict bipartite matching is motivated by potential variations in the number of neurons/units across different systems.

**Soft-matching score.** The pairwise matching score just described is intuitive but is inherently asymmetric: we can obtain different alignment values when aligning units from  $\mathbf{X}$  to  $\mathbf{Y}$ , compared to the reverse. Due to the asymmetry, it could also yield counter-intuitive outcomes such as deeming a representation  $\mathbf{X}$  with many units as somewhat aligned to any representation that is mapped to it<sup>26</sup>.

Motivated by these limitations, we employ another novel axis-dependent and symmetric representational similarity measure to quantify the similarity in the tuning curves of units across two representations: the soft-matching score. Instead of forcing a strict one-to-one mapping of units across representations, soft matching finds an optimal ‘soft’ assignment, allowing each neuron from one representation to partially match multiple neurons from the other. This matching is formulated using optimal transport theory, providing a principled way to handle comparisons between representations with different numbers of units. Soft matching generalizes the permutation-based similarity from ref. 23 by relaxing exact neuron-to-neuron matching into a weighted matching (‘soft-permutation’) framework. Intuitively, one can think of soft matching as spreading each neuron’s ‘similarity’ across several neurons in the other representation, rather than restricting it to a single best match (see ref. 26 and Methods for more background on this similarity measure). Through empirical simulations, we have shown that both the pairwise matching and soft-matching alignment measures effectively capture similarity in the ‘form’ of representations, above and beyond ‘content’ comparisons (see Extended Data Fig. 2 and Supplementary Section 1.1).

Our main findings in this paper hold for both the pairwise matching and soft-matching alignment metrics. In the main text, we show results derived using the pairwise matching score, while results based on the soft-matching score are provided in Extended Data Figs. 3 and 4.

To confirm that our axis-dependent metrics meet the basic standard of test–retest reliability, we examined within-instance alignment wherever the data allowed, specifically, across independent recording sessions for the same participant or animal, and across different random initializations for the same model architecture. We found that tuning functions and their associated axes were reliably preserved within individuals and across independent model initializations, confirming that the privileged-axis structure is stable within a given system (Supplementary Fig. 1).

**Probing alignment changes with axis perturbations.** To test whether brains and models have privileged axes, we ask whether the alignment between instances is greater than would be expected from arbitrary

axes. We draw a random rotation matrix  $\mathbf{Q}$  uniformly from the special orthogonal group  $\text{SO}(n)$ . We also consider smaller rotations (closer to the identity matrix  $\mathbf{I}$ ) along the  $\text{SO}(n)$  manifold by computing fractional powers of the sampled rotation matrix,  $\mathbf{Q}^\alpha$  ( $0 < \alpha < 1$ ). For each pair of representation instances, we rotate the axes of one instance along this  $\text{SO}(n)$  manifold and quantify how the alignment between instances changes with the amount of rotation ( $\alpha$ ). We note that each rotated representation has exactly the same linearly decodable information and population geometry and differs only in the tuning of its coordinate axes. The alignment observed at the maximal rotation ( $\alpha = 1$ ), corresponding to a fully random orthogonal transform, serves as the ‘rotation floor’—a geometry-preserving control that quantifies the level of alignment expected by chance.

**Deriving maximal alignment across all axis rotations.** The aforementioned methodology tests whether the alignment between axes in different instances of the system is higher in their native basis compared with a uniformly drawn basis. But it leaves open the possibility that some rotation matrix  $\mathbf{Q}^*$  in the  $\text{SO}(n)$  manifold yields a higher alignment than the native basis. We use the Lie exponential and perform gradient-based optimization on the  $\text{SO}(n)$  matrix manifold<sup>27</sup> to find  $\mathbf{Q}^*$  and compute this maximal alignment for both alignment measures  $s \in \{s_{\text{softmax}}, s_{\text{pairwise}}\}$ .

$$\mathbf{Q}^* = \arg \max_{\mathbf{Q} \in \text{SO}(n)} s(\mathbf{XQ}, \mathbf{Y}) \quad (3)$$

See Methods for further details. Henceforth, we refer to this maximal alignment across all axis rotations as the ‘rotation ceiling’.

We validate the measures of alignment with empirical simulations that show the framework to be effective in detecting the presence or absence of privileged axes in various representational systems (Extended Data Fig. 2).

**Interpreting the two benchmarks.** For every pair of representations, we therefore examine ‘two’ reference points: (1) the ‘rotation floor’, obtained by replacing  $\mathbf{X}$  with  $\mathbf{XQ}$  for  $\mathbf{Q}$  drawn uniformly (Haar,  $\alpha = 1$ ) from  $\text{SO}(n)$ ; and (2) the ‘rotation ceiling’,  $s(\mathbf{XQ}^*, \mathbf{Y})$ , obtained by explicit optimization over all rotations. These two anchors let us distinguish three qualitatively different regimes: (1) native  $\gg$  rotation floor  $\Rightarrow$  axes are privileged: alignment cannot be explained by chance rotations; (2) native  $\approx$  rotation ceiling  $\Rightarrow$  privileged axes already capture almost everything that can be aligned by any rotation; (3) native  $\gg$  rotation floor and native  $\ll$  rotation ceiling  $\Rightarrow$  coarse privileged axes exist, but extra orthogonal rotations could still tighten the match (permutation-only style mapping leaves some alignment ‘on the table’). Thus the ‘native versus rotation floor’ comparison is the statistical test for privileged axes, while the gap to the rotation ceiling quantifies how much additional alignment could be squeezed out by the best-possible rotation.

**Empirical characterization of the ‘rotation floor’.** It is important to note that the rotation floor is not expected to reach zero. Populations that encode the same latent variables inevitably retain some similarity, and the exact baseline depends on the representational structure: high-dimensional isotropic responses yield a low rotation floor, whereas structured, low-dimensional responses, as in biological populations, yield a substantially higher one.

We illustrate this dependence explicitly in Fig. 1c through a series of simulations systematically varying the intrinsic dimensionality and structure of the underlying representation. Each simulation involved two populations  $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{M \times N}$  (with  $M = 3,000$  stimuli and  $N \in \{200, 400, 600, 800, 1000\}$ ), and alignment was computed under repeated random rotations of  $\mathbf{X}$  across five random seeds.

In the high-dimensional isotropic regime, the entries of  $\mathbf{X}$  were independent and identically distributed (i.i.d.), drawn from a standard normal distribution and set  $\mathbf{Y} = \mathbf{X}$  to ensure that the true maximal

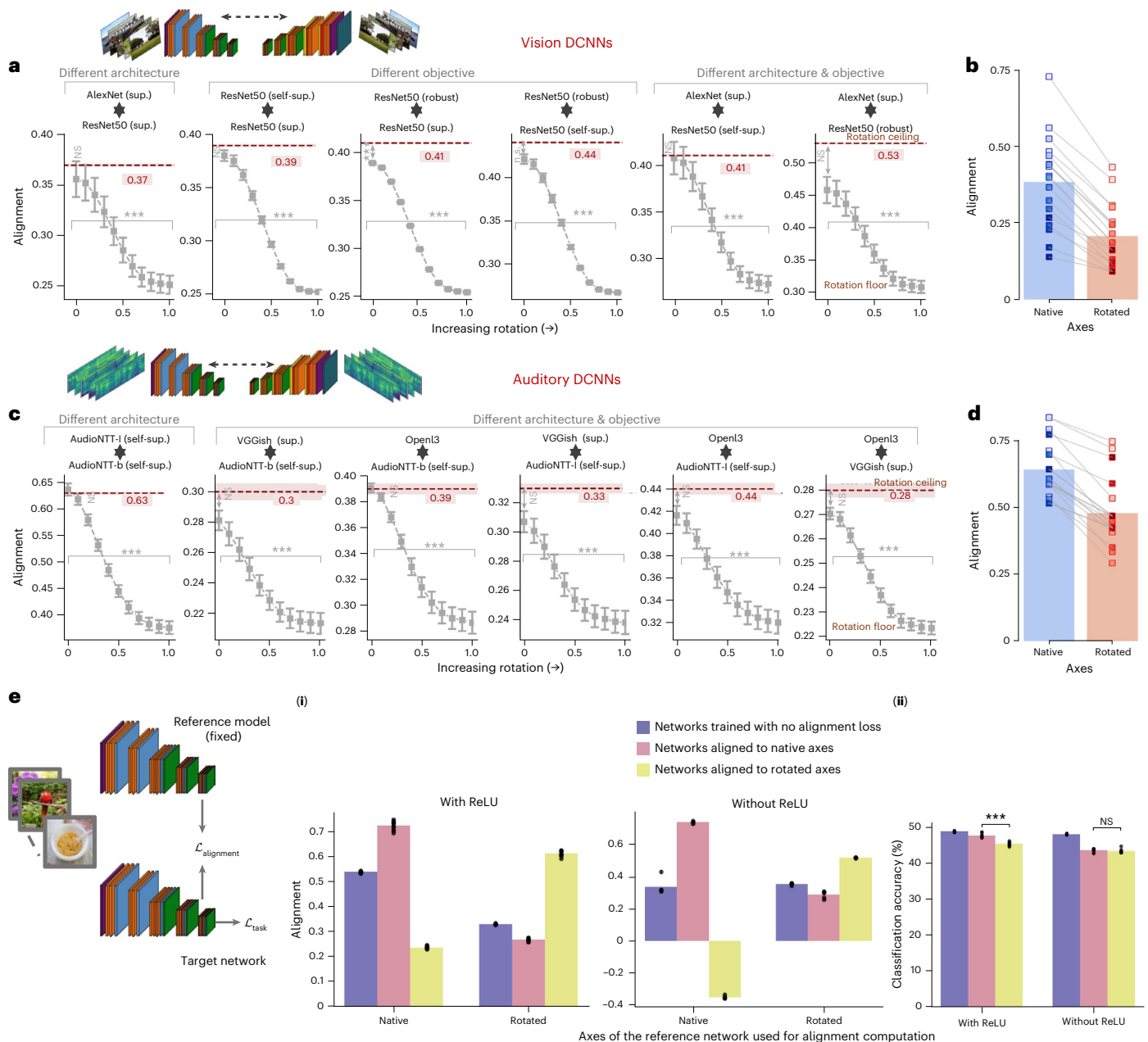


**Fig. 2 | Model-to-model alignment.** **a**, Alignment between late-stage representations of different visual deep convolutional neural networks trained with different architectures and learning objectives. This alignment is measured using the pairwise matching score and examined as a function of axis rotation ( $\alpha$ ), 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. Alignment computed from activations to the 1,000-image NSD stimulus set used for computing inter-participant alignment. Error bars are s.e.m. over multiple splits of the data. Significance tests compare the alignment values at no rotation ( $\alpha = 0$ ) with those at full rotation ( $\alpha = 1$ ), as well as with the optimal alignment values observed across all rotation angles. The red shaded region indicates the s.e.m. for the optimal alignment value across all rotation angles, computed from multiple data splits. **b**, Raw alignment scores, computed as correlation of response profiles in held-out data after pairwise matching, between every pair of corresponding layers of ResNet (sup.) and ResNet (self-sup.) in the native and rotated axes corresponding to  $\alpha = 1$ . Each line connecting two dots represents a single layer. The dots are colour coded to indicate layer depth, with darker shades signifying greater depth. **c**, Same as **a**, but for auditory neural networks

and computed from activations to the 165 natural sounds used for auditory inter-participant alignment analysis. In each case, we gradually rotated the axis in one representation while computing its alignment with the other representation in its native basis. **d**, Same as **b** but for auditory networks, specifically between every pair of corresponding layers of AudioNTT-b and AudioNTT-I in the native and rotated axes corresponding to  $\alpha = 1$ . For **a–d**, significance annotations indicate two-sided paired  $t$ -tests computed across  $n = 10$  stimulus train/test splits for each model–model pair comparison: \*\*\* $p < 0.0001$ . **e**, Left: schematic illustrating the experimental setup for manipulating axis alignment in neural networks. (i) Alignment values of target networks with the native and rotated axes of the reference network, measured using pairwise matching scores. The left plot shows networks with ReLU in the last convolutional layer, while the right plot shows networks without ReLU in that layer. (ii) Classification performance of the target networks with and without ReLU in the last convolutional layer across three alignment conditions: native axes alignment, rotated axes alignment and no alignment loss. Bars indicate the mean across networks, error bars indicate s.e.m., and overlaid points indicate individual independently initialized networks ( $n = 10$  networks per condition). Significance annotations indicate two-sided  $t$ -tests unless otherwise stated: \*\*\* $p < 0.0001$ .

alignment equals 1. Even after applying random orthogonal transformations to  $X$ , alignment with  $Y$  remained at  $-0.2$ , indicating a non-zero rotation floor even for fully unstructured Gaussian data.

In the ‘low-dimensional clustered’ regime,  $X$  was constructed from six independent latent modules, each two-dimensional (2D), with within-module noise of 0.1 and independent weights mapping latent drivers to neuron groups. Again,  $Y = X$ . Here, random rotations yielded an alignment value close to 0.6, a substantial increase relative to the isotropic case. This arises because clustered (and hence low-rank) data occupy only a small subspace of their ambient  $N$ -dimensional embedding and therefore share a large proportion of variance even under random rotations.

To bridge these extremes, we simulated a ‘mid-dimensional hybrid’ regime in which half of the features participated in clustered latent modules and the remaining half were isotropic. This yielded an intermediate rotation floor of  $-0.37$ , confirming that alignment values vary smoothly with the intrinsic dimensionality and structure of the data.

Together, these analyses show that the rotation floor, that is the alignment baseline under random rotations, depends on the intrinsic geometry and dimensionality of the representations being compared. Thus, the meaningful test of privileged axes is the systematic advantage of the native basis over uniformly random rotations, rather than the absolute baseline itself.

Moreover, it is important to note that in these simulations we set  $Y = X$ , ensuring a rotation ceiling of 1. In empirical datasets, the corresponding ceiling is typically lower because the compared populations (for example, brains or independently trained models) do not share identical latent dimensions and include measurement noise.

We apply the method outlined above for assessing whether and to what extent a representational system has privileged axes, addressing the following questions: (1) Do sensory cortices (both visual and auditory) have privileged representational axes?, (2) Do deep convolutional neural network models trained on natural visual and auditory input also have privileged axes?, (3) Do biological and artificial neural networks share the same representational axes?, (4) Do the particular axes privileged in brains and networks confer computational advantages over other possible axes?, (5) Does brain–model axis alignment improve our ability to choose between competing models of biological computations (for example, DCNNs versus transformers)?, (6) Are these aligned basis directions interpretable?, (7) Are there other non-interpretable tuning functions that consistently characterize these aligned axes? and (8) Are privileged axes also found in higher-order cognitive domains?

### Sensory cortices have privileged axes

We first tested whether perceptual cortices exhibit privileged axes for the representation of natural stimuli. To answer this question for the visual system, we leveraged the Natural Scenes Dataset (NSD, human fMRI)<sup>28</sup> and the macaque IT electrophysiology data of ref. 29 (hereafter referred to as the ‘ManyMonkeys’ dataset) to assess how axis-dependent alignment between conspecifics’ neural responses in the human ventral visual cortex and macaque IT, respectively, changes with axis rotation. We found that visual representations (both human fMRI and monkey neurophysiology) exhibit near-maximal alignment between individuals (values approaching the rotation ceiling in Fig. 1d,e), and this alignment decreases gradually with increasing axis rotation of the representation of one participant/animal away from its native axes, indicating a non-random relationship between the two bases (Fig. 1d,e).

To address the same question in the auditory system, we analysed a dataset of fMRI responses to natural sounds (measured in the auditory cortex of multiple participants<sup>30</sup>). Here again we found a similar pattern of decreasing alignment with axis rotations, providing evidence that the auditory cortex also has privileged axes for representing auditory stimuli (Fig. 1f). While the auditory cortex analysis reveals a slight increase in alignment with small rotations, this effect is not statistically significant ( $p = 0.69$  using  $t$ -test across 5 folds) and may represent

random variability. Importantly, the claim of privileged axes is based on comparisons between native axes and fully random arbitrary axes (for example,  $\alpha = 1$ ), where arbitrary rotations are sampled from the Haar distribution, ensuring uniformity across  $SO(n)$ . However, the overall effect size (that is, the reduction in alignment upon axis rotation) is smaller in the auditory cortex than in the visual cortex. This may be due to the lower dimensionality of the auditory data, where a modest number of components account for most of the variance<sup>30</sup>. In such spaces, rotations may cause less divergence between native and rotated axes, reducing the observed effect.

Thus, across three datasets spanning visual and auditory cortices, we find that the majority of neurons/voxels lose this direct alignment with neurons/voxels in conspecifics upon axis rotation (Fig. 1d–f, bottom). Although fMRI voxels and electrocorticography (ECoG) electrodes each pool activity over many neurons, a wealth of evidence shows that these coarser signals nonetheless preserve the same dominant tuning dimensions present at the single-unit level due to the topographic clustering of functionally similar neurons in the sensory cortex<sup>11,31</sup>, allowing us to treat all modalities as reporting on similar privileged axes.

We also conducted this alignment analysis by sampling rotation matrices through a different method, specifically, as a composition of plane rotations. This approach provides finer control over the degree of rotation and the number of plane (2D) rotations involved. Our findings indicate that even minor perturbations to the native axes reduce alignment between conspecifics (see Methods and Extended Data Fig. 9). Our results thus show that these biological representational systems have a privileged set of axes, challenging widespread assumptions about their arbitrariness.

As a further control, we asked whether our conclusions might depend on the use of Haar-random rotations, which do not preserve the sparsity of unit responses. We therefore implemented a sparsity-matched null, constructing orthogonal matrices  $R \in SO(n)$  that preserved the mean lifetime sparsity of each response matrix. Even under this stricter null, alignment in the native axes was consistently higher (Supplementary Fig. 2). This confirms that our findings are not an artefact of the choice of null distribution, but reflect genuine axis structure in the data.

### Deep convolutional networks have privileged axes

To further understand the privileged axes of biological sensory systems, we next asked whether and when artificial neural networks trained on natural stimuli also develop privileged axes.

In the domain of vision, we selected four standard deep convolutional networks, each trained on ImageNet<sup>32</sup>, that span different architectures and learning objectives. These include: (1) ResNet50 with supervised training, henceforth called ResNet50 (sup.)<sup>33</sup>, (2) AlexNet with supervised training, denoted as AlexNet (sup.)<sup>34</sup>, (3) ResNet50 trained in a self-supervised fashion with a contrastive learning objective, specifically momentum contrast<sup>35</sup>, denoted as ResNet50 (self-sup.) and (4) ResNet50 with adversarial training, henceforth ResNet50 (robust)<sup>36</sup>. All these models perform competitively on large-scale object categorization. We first directly compared later layers of DCNNs as these are known to best model the high-level visual cortex.

Even when comparing models with different architectures (for example, ResNet versus AlexNet) and learning objectives (for example, supervised versus self-supervised), we find that DCNNs are almost maximally aligned with each other in their native basis (Fig. 2a), indicating that they contain privileged axes. Importantly, untrained networks do not exhibit this phenomenon (see Extended Data Fig. 5).

We performed similar comparisons in the auditory domain, employing four different models, all trained on the AudioSet corpus<sup>37</sup>. These include: (1) a VGG-inspired architecture with supervised training, henceforth called VGGish (sup.)<sup>38</sup>, (2) a convolutional architecture with

a high-dimensional representation embedding (2,048), trained with contrastive learning, labelled here as AudioNTT-I (self-sup.)<sup>39</sup>, (3) a more compact network akin to (2) but with a reduced representation embedding dimensionality (512), denoted as AudioNTT-b (self-sup.)<sup>39</sup>, and (4) the  $L^3$ -net audio model, trained with self-supervision on an audio-visual correspondence pretext task, denoted as OpenI3 (ref. 40). Irrespective of the architecture and learning objectives, auditory DCNNs after training again converge onto privileged axes (Fig. 2c).

**Do all layers have privileged axes?** Our results thus far highlight the prevalence of privileged axes in late-stage network representations. But do they exist in earlier layers as well? To explore this, we perform layer-by-layer comparisons in both the vision and auditory domains. In the vision domain, we compare each pair of layers in the supervised and self-supervised (MoCo) versions of ResNet, on the grounds that using the same architecture allows for a direct comparison of layers. For the auditory domain, we compare the AudioNTT-b and AudioNTT-I networks, which differ in the number of units in some layers but otherwise share the same architecture, enabling a similar direct comparison. In the case of convolutional layers, each ‘channel’ corresponds to a distinct representational unit. We sample random positions for all stimuli within the convolutional layer to find the optimal matching between representational units and compute the final alignment on held-out data.

Our findings reveal that, for both vision and auditory networks, alignment is greater in the native axes compared with rotated axes for all model layers (Fig. 2b,d). Specifically, in the vision networks, alignment along the native axes was between 43% and 159% higher than along the rotated axes for each layer. Similar trends were observed in the auditory networks, with alignment ranging from 12% to 85% higher in the native axes.

**Why do different networks have similar bases?** The dominant view from previous work is that the basis is an arbitrary and inconsequential aspect of network representations. This idea stems from the notion that any representation within a neural network is equivalent up to a full-rank matrix multiplication because the successive weight matrix can counterbalance any such multiplication, rendering the initial choice of basis irrelevant<sup>22,41</sup>. However, this degeneracy is altered by the introduction of nonlinearities within deep network architectures. Specifically, consider the set of invertible linear transformations that commute with the nonlinear operators used in DCNNs, such as ReLUs. This set does not encompass the entire space of invertible matrices, but rather is constrained to a limited subset within that space. Nonlinearities might thus act as symmetry-breaking mechanisms, potentially predisposing networks to favour certain activation bases.

To understand this constraint, consider a representation of post-ReLU activations  $x$ . The ReLU operation ensures that all activations in  $x$  are non-negative. If we apply a rotation to these activations, we get  $y = Ux$ , where  $U$  is a rotation matrix. For both  $x$  and  $y$  to be valid post-ReLU activations, they must remain non-negative after the transformation. For this, the matrix  $U$  must be a non-negative matrix. But this means that  $U$  must be a permutation matrix, because every orthogonal matrix with non-negative entries is necessarily a permutation matrix. This means that  $U$  can only permute (or shuffle) the activation units, rather than performing arbitrary rotations. Thus, the nonlinearity introduced by ReLU disrupts the rotational symmetry of the activation space, potentially explaining why different networks converge to similar representational axes. (see Supplementary Section 1.2 ‘How ReLU nonlinearities disrupt rotational symmetry in network representations’ for additional technical details).

We next investigate the empirical validity of the theoretical claim that privileged axes can enhance task performance under structural constraints, such as nonlinear activation functions that disrupt rotational symmetry in neural networks. Specifically, we trained neural

networks for ImageNet classification while explicitly maximizing or minimizing alignment with the native representational axes of a reference network (also trained on ImageNet classification).

We used a custom novel loss function involving two parts: a cross-entropy loss ( $\mathcal{L}_{\text{task}}$ ) for optimizing task performance (classification), and an alignment loss ( $\mathcal{L}_{\text{alignment}}$ ) to maximize axis alignment with a fixed reference model:  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{alignment}}$ . We tested three conditions: (1) networks trained to align with the native axes of the reference model ( $\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{alignment}}^{\text{native}}$ ), (2) networks trained to align with rotated axes ( $\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{alignment}}^{\text{rotated}}$ ) and (3) networks trained without any alignment loss, using only the task loss ( $\mathcal{L}_{\text{task}}$ ). For this experiment, we used AlexNet as the reference model, using the last convolutional layer for the alignment loss. We then trained new networks (each with the AlexNet architecture) to test the different alignment conditions.

Our results confirm that the custom loss function effectively achieves the intended alignment: networks trained with the alignment loss for the native axes show improved alignment compared to those trained without the alignment loss, while networks trained to align with rotated axes exhibit better alignment with the rotated axes and reduced alignment with the native axes relative to the no-alignment-loss condition (Fig. 2e(i)). Importantly, our results show that training networks to align with rotated axes consistently leads to a significant drop in task performance compared to networks trained to align with the native axes of the reference model ( $p < 10^{-6}$ , based on independent  $t$ -tests across 10 networks; Fig. 2e(ii)). Networks trained to align with the native axes achieve a classification accuracy of ~48%, consistently outperforming those trained to align with rotated axes, which achieve ~45%. To test the role of non-negativity in this result, we then removed the ReLU activation after the last convolutional layer (both in the reference model and the new networks), making the layer rotationally symmetric. For this comparison, new reference models were trained from scratch without the post-conv5 ReLU, and new models with the custom loss function described above were trained as well, ensuring a direct comparison in this condition. In this case there was no significant performance difference ( $t_{(16)} = 0.90$ ,  $P = 0.38$ ) between networks aligned to native versus rotated axes (Fig. 2e(ii)). These findings support the hypothesis that privileged axes facilitate better task performance under structural constraints such as nonlinear activation functions which break the rotational symmetry in networks. The role of ReLUs in establishing a privileged axis in networks is further supported by the finding that, in networks trained without ReLU in the last convolutional layer and without alignment loss, the alignment with the native axes is similar to that with the rotated axes. In contrast, networks trained with ReLU (in the absence of alignment loss) exhibit a marked difference in alignment between the native and rotated axes (Fig. 2e(i)).

### DCNNs and sensory cortices have aligned privileged axes

The presence of privileged axes in perceptual cortices and DCNNs raises the question of whether sensory areas and DCNNs have the same privileged axes. To address this question, we measured the axis-dependent alignment between brain representations and DCNNs, and examined how it changes with increased axis rotations. For the visual domain, we incorporated data not only from the ManyMonkeys dataset and the Human NSD but also from a novel dataset capturing responses in macaque IT to images from the human NSD. We conducted analogous experiments with auditory models and human auditory cortex responses, measured using both fMRI and intracranial recordings (ECOG)<sup>42</sup>.

Across all brain–model pairs in both vision and audition, alignment was high without rotation (often, approaching the maximal value) and decreased by as much as ~20–35% as DCNN axes were rotated (Fig. 3a–e). This result suggests that sensory cortices and DCNNs tend to exhibit the same axes.

Are the tuning functions that are most aligned between conspecifics also the ones most aligned between brains and DCNNs? To find out,

we asked whether the capacity of a neuron's or voxel's tuning function to be accurately represented by that of a single neuron or voxel in a conspecific can predict its peak alignment with the tuning function of a single unit in a DCNN (Fig. 3f). Across neurons/voxels, we found a significant correlation between the two (partial correlation:  $r = 0.75$ ,  $p < 10^{-100}$ , 95% CI [0.74, 0.76]), even after adjusting for the varying reliability of response measurements in different neurons/voxels. This result indicates that the neurons or voxels that exhibit high one-to-one alignment across conspecifics also tend to be more one-to-one aligned with DCNNs.

We emphasize that standard approaches for comparing artificial and biological neural representations (linear encoding, RSA or CKA) conceal this similarity in 'representational form' between DCNNs and sensory regions. Assessing similarity in form requires measures of axis alignment such as those introduced here.

Why do biological and machine systems converge on similar representational axes? Importantly, untrained networks do not show axis alignment with neural representations (see Extended Data Fig. 5). This result suggests that optimization constrained by natural input statistics and the architectural primitives in DCNNs (for example, ReLUs) might help explain why brains favour certain coordinate systems over others.

### Computational implications of privileged axes

Given that axis rotations maintain the information content and population geometry of a representation, why do high-level sensory cortical and DCNN representations exhibit specific, non-random axes? We hypothesized that the representational form employed by the brain could affect efficiency, downstream wiring costs and the capacity for few-shot generalization under anatomical and physiological constraints.

**Efficiency.** We operationalized efficiency as economy in the number of neurons active in response to a stimulus. To test whether the native axes result in more efficient representations, we compared the lifetime sparseness of individual units and population sparseness in DCNN and cortical representations using the native basis, and compared them to those for uniformly drawn random bases.

In both vision and audition, both brain and model native axes produced representations with higher lifetime and population sparseness than rotated axes, as measured with kurtosis (Fig. 4a) as well as an alternative measure of lifetime sparseness proposed in ref. 43 (Supplementary Fig. 5). This effect is also evident in the response distribution of units in the native axes, which had a more pronounced heavy-tailed appearance compared with units in an arbitrary basis (Fig. 4a left, top versus bottom).

**Downstream wiring costs.** We next hypothesized that the axes in the brain are those that reduce wiring costs. Due to space constraints on the number of possible connections between neurons in two regions, a neuron in one region likely cannot connect to every neuron in another

region, such that its readout from that region might need to be driven by a small subset of neurons. We asked whether the axes found in the brain are those that facilitate a sparse readout of behaviourally relevant information (for example, object category labels) from patterns of neural activity.

We trained linear decoders on representations from DCNNs and sensory cortices, aiming to predict the category labels of different stimuli from their evoked representations. For visual representations, we used NSD images for extracting both model (here, the self-supervised visual DCNN) and brain representations (macaque IT and human high-level ventral visual cortex). The decoder classes were the 80 object categories in the MS-COCO dataset. Since an image in the NSD can have more than one object category, we set this up as a multilabel classification task and quantified the classification performance for each category label with the area under the receiver operating characteristic (AUC-ROC) measure. As shown in Fig. 4b, all model and sensory cortical representations perform well on this decoding task (AUC-ROC  $\approx 0.69$ – $0.93$ ). This decoding accuracy is not affected by the change of basis, since arbitrary rotations preserve the information that can be linearly accessed in a representation. We then assessed how sparsely the category information is 'read out' in each basis by estimating the kurtosis of the readout weight distribution of each category over all units. Results revealed that the native basis yields a much sparser readout than an arbitrary basis, and even small deviations from the native basis lead to notable reductions in readout sparsity. This reduction is not dominated by a few select categories, but almost every category is read out more sparsely in the native basis.

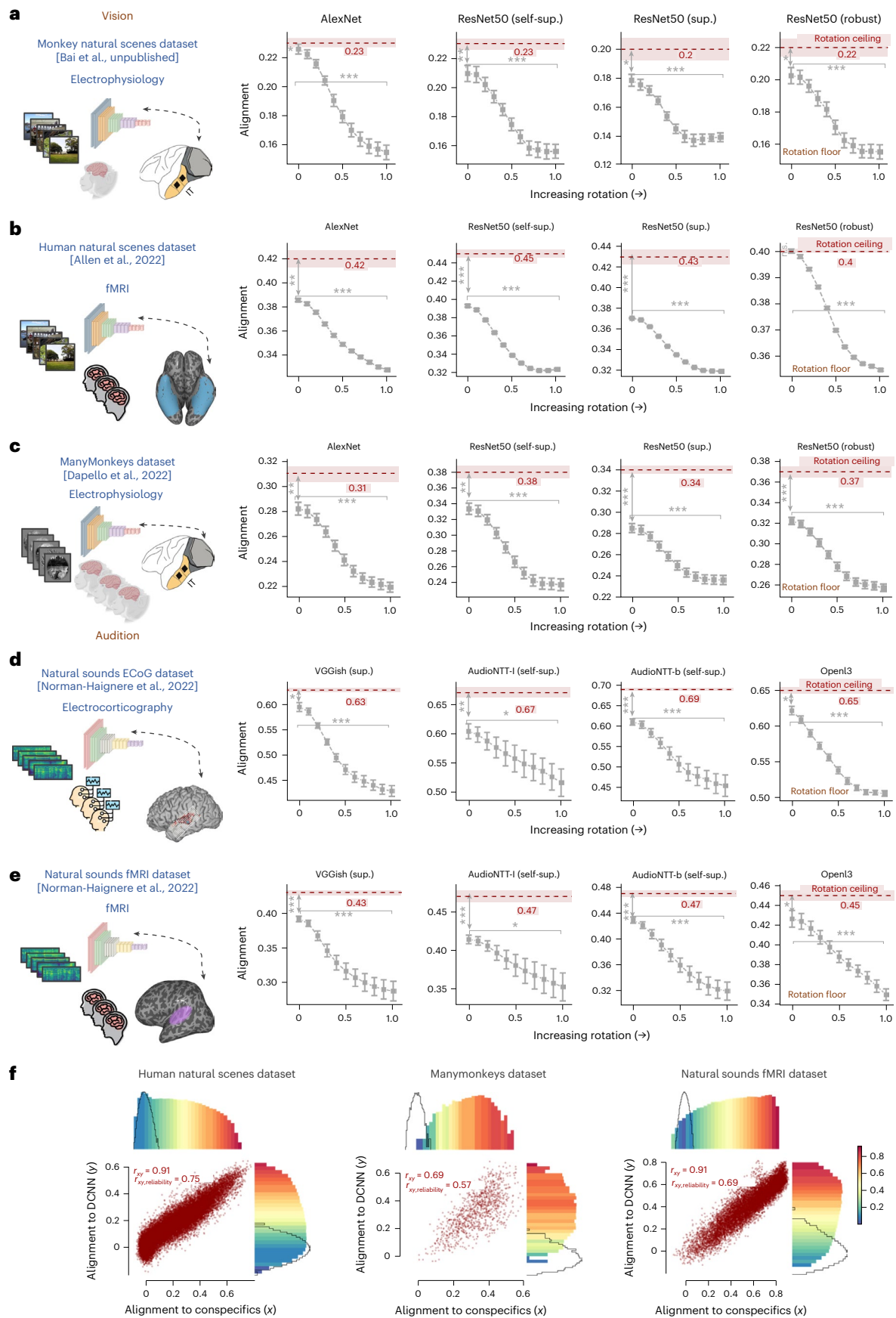
Analogous experiments in the auditory domain yielded similar results. Here we trained linear classifiers on intracranial cortical responses and self-supervised DCNN activations to natural auditory sounds. The classifier decoded the category labels of the sounds, which spanned 11 human-interpretable categories (for example, environmental sounds, human vocalizations, song and so on). As in the visual domain, a change of basis preserved the decoding accuracies but dramatically decreased the sparsity of the classifier weights. These results indicate that object category decoding can be implemented with lower wiring cost using the axes found in the brain. Whether the same conclusions hold for the within-category information presumably contained within many of these neural populations is an important question for future research.

Beyond categorical selectivity, we also tested whether privileged axes confer wiring-cost benefits for category-orthogonal properties. Using the IT recordings in ref. 44 as well as the corresponding DCNN activations for these stimuli, we fit linear readouts for object position, scale and rotation, compared weight sparsity across native versus rotated axes, and find that privileged axes reduce wiring cost for these very variables as well (Extended Data Figs. 6 and 7).

Importantly, our results do not imply that privileged axes arise from wiring optimization. The DCNNs analysed here were not trained with any explicit wiring-cost penalty. Rather, the sparsity

**Fig. 3 | Model-to-brain alignment. a–c,** Alignment of various visual deep convolutional neural networks with biological neural representations in the macaque IT and human ventral visual stream. This alignment was measured using the pairwise matching score and examined as a function of axis rotation ( $\alpha$ ), 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. **d,e,** Alignment of auditory DCNNs with neural representations in the human auditory cortex, again assessed as a function of DCNN axis rotation. Across **a–e**, significance tests were two-sided paired  $t$ -tests comparing alignment at no rotation ( $\alpha = 0$ ) with full rotation ( $\alpha = 1$ ), as well as with the optimal alignment observed across rotations. Error bars and shaded regions indicate s.e.m. across  $n = 5$  stimulus train/test splits per model–brain comparison; \* $p < 0.01$ , \*\*\* $p < 0.0001$ . The largest

annotated  $p$  value was 0.009. **f,** The one-to-one alignment of each neuron/voxel with neurons/voxels in other conspecifics as a function of its alignment with units in a DCNN (self-supervised vision/auditory networks). The raw correlation between these variables and the partial correlation after excluding the effect of different noise ceilings of various neurons/voxels are reported. Coloured density plots depict the distribution of alignment values (always calculated on stimuli held out from those used to find matching voxels/units; horizontal: alignment with conspecifics, vertical: alignment with DCNNs), with colours indicating the mean noise ceiling of voxels within each bin, while black density plots represent the null distribution generated using the permutation method, where the stimulus order in one representation of each pair was randomly shuffled before computing the alignments. The alignment values seen in brains and models are frequently much higher than would be expected by chance.



and wiring-efficiency results highlight a functional consequence of the learned axes: once such privileged orientations emerge through standard factors—data statistics, nonlinearities and optimization dynamics—they naturally support much sparser linear readouts. In biological systems, this sparsity would directly translate into fewer axonal connections and thus lower wiring cost.

#### Capacity for few-shot generalization under wiring constraints.

We also hypothesized that privileged neural axes might have consequences for generalization, particularly impacting sample efficiency during learning of downstream tasks under biological constraints that might limit the number of neurons accessible for read-out. Specifically, we posited that the privileged axes in the brain could enhance the generalization capabilities of downstream linear readouts, especially during constrained learning scenarios, such as few-shot learning and enforced sparsity on the readout. To test this, we initially focused on categories corresponding to the dominant components encoded in the ventral visual pathway and high-level auditory cortex. Previous decomposition analyses of brain data have identified faces, places, bodies, text and food as key features represented in the ventral visual pathway<sup>17</sup>, with music and speech emerging as canonical components in responses to natural sounds<sup>30</sup>. We trained sparse linear decoders on top of neural representations in both native and arbitrary bases. A sparsity constraint, meant to mimic anatomical constraints on the number of inter-region connections, was enforced on the decoder through an  $l_1$  penalty. The  $l_1$  penalty also makes the generalization performance of the decoder axis-dependent. A few-shot learning scenario was instantiated by training the linear decoders using only a very small fraction (10%) of labelled stimulus–category pairs. In the visual domain, separate sparse linear regressors were trained to predict the presence of each of the five categories in an image (quantified using salience ratings) as listed above: faces, places, bodies, text and food. For auditory representations, two distinct sparse linear classifiers were utilized to discern the presence of music or speech in a sound waveform. For each candidate representation and basis (native versus arbitrary), we then computed the generalization performance of the trained sparse linear decoders on held-out stimuli. Across most categories, the generalization performance of a representation in its native basis exceeded its performance in an arbitrary basis (Fig. 4c(i)). This pattern was consistent in both neural network and sensory cortical representations, and was observed across both visual and auditory domains.

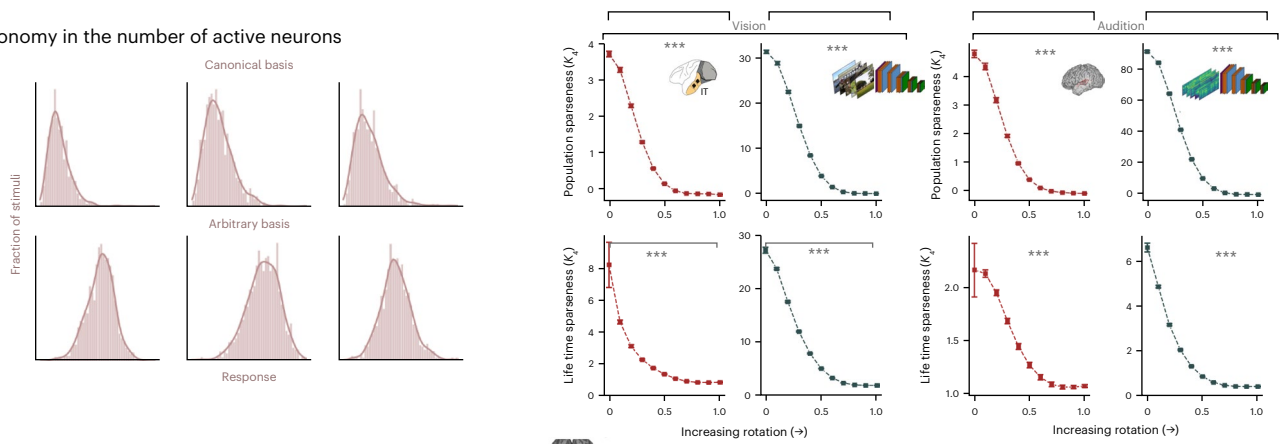
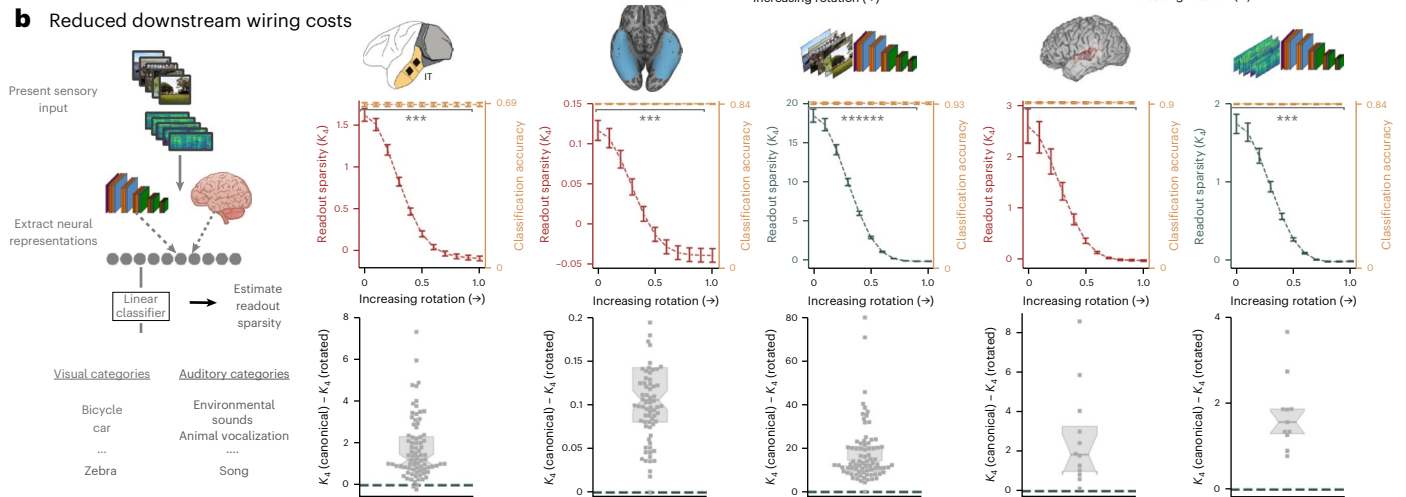
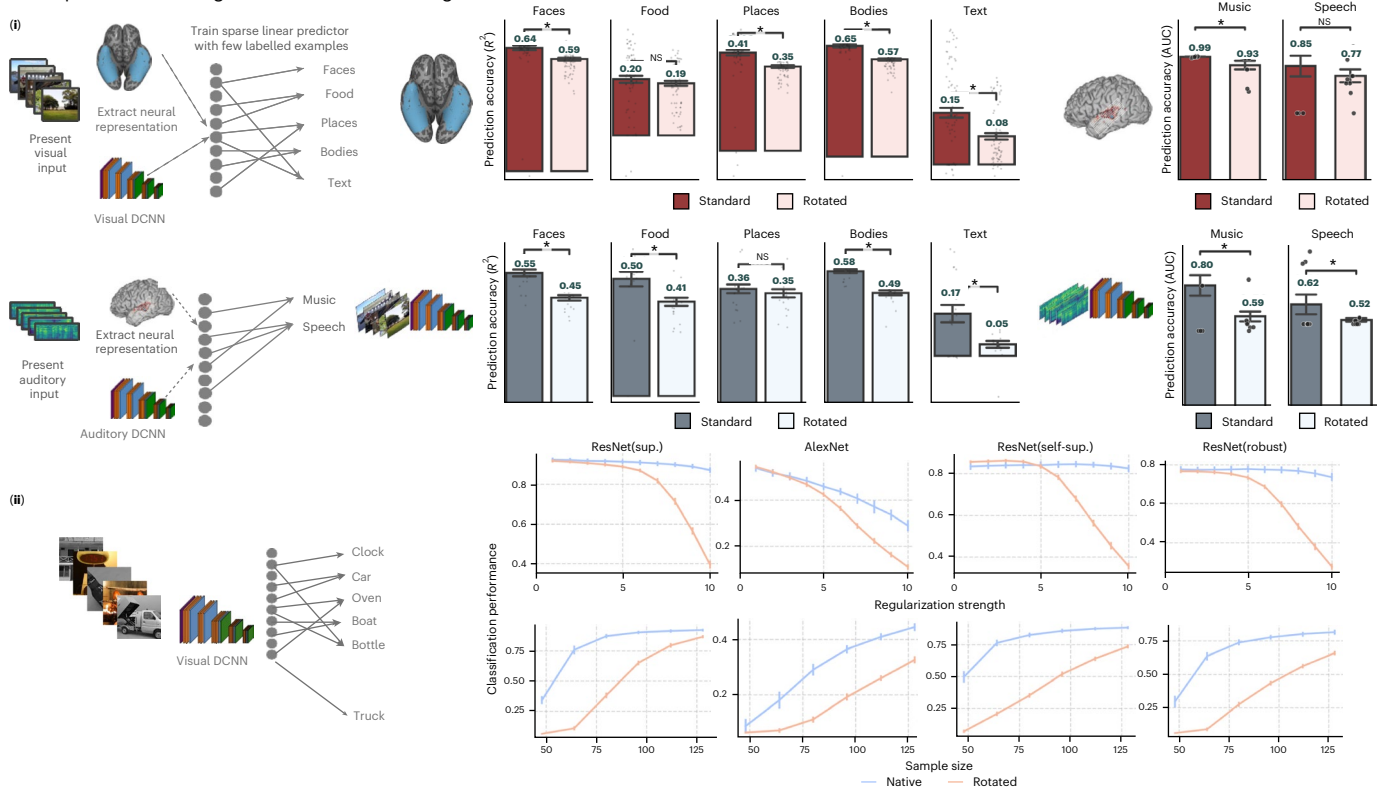
Previous findings show that native axes enhance performance in predicting the presence of categories corresponding to canonical components of neural responses to natural images or sounds. However, could this reflect a broader phenomenon, with native axes supporting better generalization under sparsity constraints in more complex tasks? To explore this question, we investigated the generalization of visual network representations in an object classification task comprising 16 ImageNet categories from ref. 45. These include airplane, bicycle, truck and several other categories not traditionally linked to selectivity in biological vision, providing a broader test of the influence of privileged axes.

To test generalization under biologically motivated constraints, we trained sparse linear readouts on different neural network representations to classify images into these 16 categories. By varying the  $l_1$  sparsity penalty on the readouts, we observed striking differences between representations in their native versus rotated axes. Notably, representations in the native axes outperformed those in rotated axes under high sparsity penalties across all four vision networks studied (Fig. 4c(ii), top). This gap remained across a wide range of training sample sizes for the linear readouts, although it generally diminished as more data became available in most networks (Fig. 4c(ii), bottom). Similar results regarding the superior performance of native axes in comparison to arbitrary axes in few-shot generalization under read-out sparsity were obtained across 16 other datasets featuring various out-of-distribution image types (for example, silhouettes, stylized images and noise-distorted natural images), underscoring the robustness of this finding (Supplementary Figs. 7 and 8). Note that the networks themselves lacked any explicit  $l_1$  regularization in their training phase. These results suggest that privileged axes confer substantial functional advantages, particularly under biologically motivated constraints on the readout, such as sparsity, offering new insights into their relevance in artificial and biological systems.

Why do privileged axes facilitate few-shot learning? Intuitively, generalization error in a decoding model comprises a bias component, stemming from the decoding model's simplifying assumptions, and a variance component, arising from fluctuations in the decoder weights based on the training set. Imposing a sparsity constraint will reduce the variance component across all axes by narrowing the hypothesis space. However, the bias component is then strongly influenced by specific axes. For many axes, the sparsity penalty might exclude the actual solution from the hypothesis space, increasing bias and, in turn, the generalization error. See Methods for a more complete description of this theoretical proposition.

**Fig. 4 | Computational implications of privileged axes.** **a**, Left: distribution of responses to natural stimuli for 3 neurons in the native basis (top) and 3 in the rotated basis (bottom) using the Macaque-NSD dataset. Right: line plots showing the mean population sparseness (above) and lifetime sparseness (below), computed as kurtosis, at different axis rotations. 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. **b**, Left: schematic illustrating the procedure for estimating downstream readout sparsity. Linear decoders are trained on top of neural representations to predict stimulus categories. The sparsity of the learned readout weights are calculated as kurtosis. Right, top: line plots showing the mean readout sparsity, computed as kurtosis, across all categories in the representation at different axis rotations. Right, bottom: histograms depicting the distribution of the difference in sparsity between the native and rotated axes across all categories. The centre line indicates the median, box bounds indicate the first and third quartiles, and overlaid points indicate individual categories. From left to right, the representations include the macaque IT neural responses, human ventral visual stream responses, self-supervised visual DCNN activations, human auditory cortex responses and self-supervised auditory DCNN activations. **c**, (i) Left schematic shows the procedure for estimating the few-shot generalization performance of each representation under sparsity constraints. We train sparse linear readouts (using  $l_1$ -regularized linear regression) on top

of neural representations to predict the salience/presence of each category in the stimulus using only few labelled examples. Left: few-shot generalization performance of the human ventral visual cortex representation (top) and the self-supervised visual DCNN representation (bottom) for the regression task (that is, predicting category salience), estimated using  $R^2$  for each of the 5 visual categories. Right: performance of the human auditory cortex representation (top) and the self-supervised auditory DCNN representation (bottom) for the classification task (that is, predicting whether the category is present or not), estimated using AUC-ROC for each of the 2 auditory categories. (ii) Left: procedure for estimating the few-shot generalization performance of each representation under sparsity constraints. We train sparse linear readouts (using  $l_1$ -regularized linear regression) on top of network representations of the 16-class ImageNet dataset to predict the classification label of the image among the 16 entry-level categories of ImageNet. Right, top: the classification accuracy (proportion correct) of penultimate convolutional layer representations from various networks, both in their native and rotated axes, as a function of the sparsity penalty, with readouts trained on 5 images per category (80 images total). Right, bottom: the classification accuracy as the training sample size increases, with the sparsity penalty fixed at 10. Significance annotations indicate two-sided  $t$ -tests unless otherwise stated: \* $p < 0.01$ , \*\*\* $p < 0.0001$ . For **b**, all annotated comparisons had  $p < 10^{-10}$  and  $n = 5$  across splits; for **c**, the largest annotated  $p$  value was 0.007 ( $n = 20$  splits for vision and  $n = 10$  splits for audition). Data in line plots and bar plots are presented as mean  $\pm$  s.e.m.

**a** Economy in the number of active neurons**b** Reduced downstream wiring costs**c** Improved few-shot generalization under wiring constraints

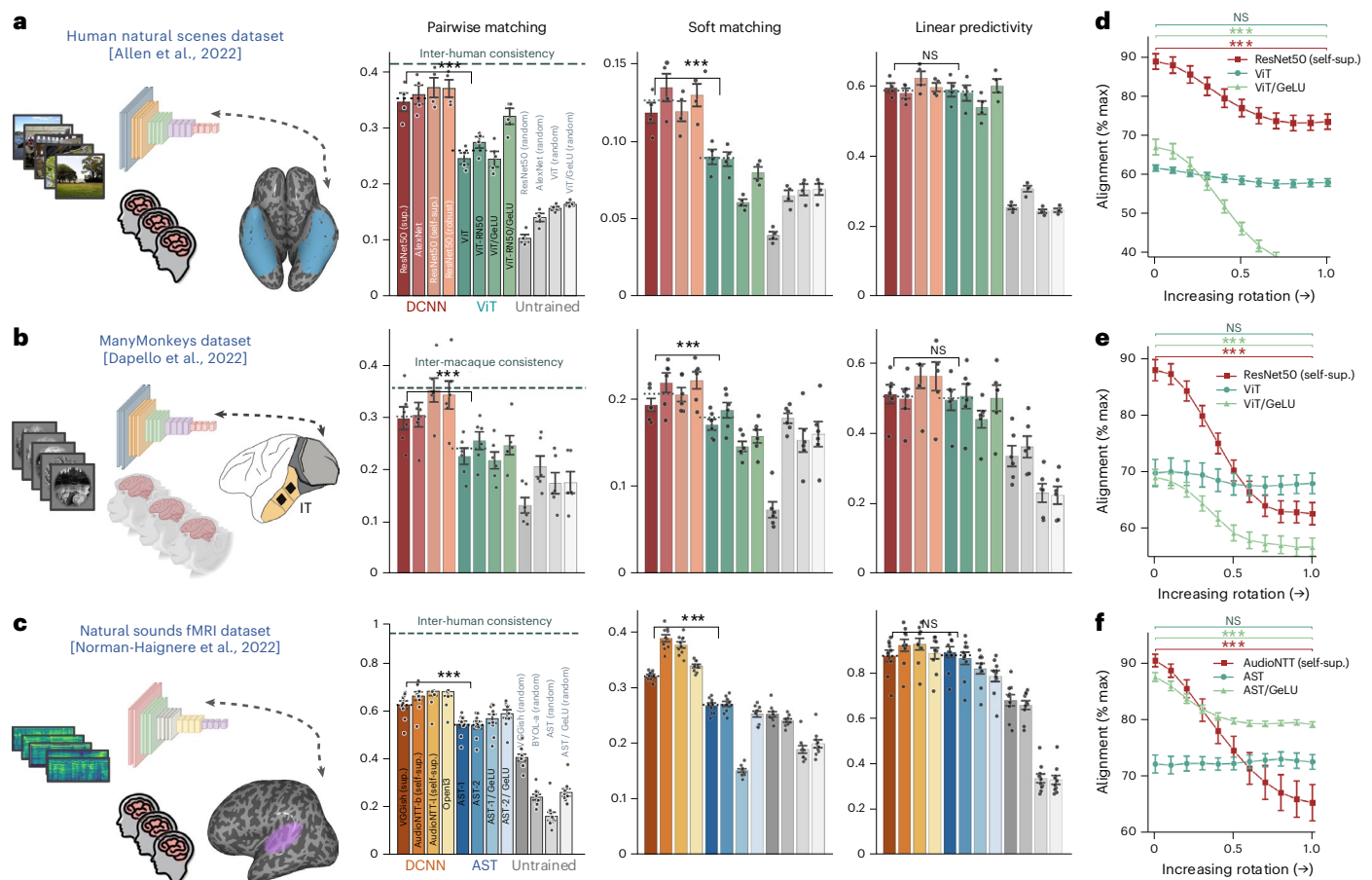


**Fig. 5 | Axis-dependent alignment methods better discriminate fit of computational models to the brain.** **a, b**, Alignment of various visual deep convolutional neural networks with neural representations in the human ventral visual stream (**a**) and macaque IT (**b**), examined using two measures of axis alignment (the pairwise matching correlation (left) and the soft-matching correlation score (middle)) and linear predictivity. Here and in **c**, left plots display inter-individual consistency for the pairwise matching correlation. **c**, Alignment of auditory DCNNs with neural representations in the human auditory cortex, assessed using the same three metrics. Bars indicate the mean across animals/participants and error bars indicate s.e.m. Overlaid points indicate individual animals/participants:  $n = 6$  monkeys for macaque IT,  $n = 4$  participants for human NSD and  $n = 10$  participants for auditory cortex.

Significance annotations indicate two-sided paired  $t$ -tests: \*\*\* $p < 0.0001$ ; maximum  $p = 3 \times 10^{-8}$ . **d–f**, Pairwise matching correlation between artificial neural network representations and neural responses in human ventral visual cortex (NSD (**d**)), macaque inferotemporal cortex (ManyMonkeys dataset (**e**)) and human auditory cortex (natural sounds fMRI dataset (**f**)) as a function of model representation axis rotation ( $\alpha$ ), expressed as a proportion of the maximum possible value obtainable across all axis rotations, determined using the ‘optimal axis alignment’ procedure described in the text. Here, the model representations include penultimate layers of self-supervised DCNNs (architectures ResNet50 and AudioNTT), transformers (ViT and AST) and the final GeLU layers of transformers (ViT/GeLU and AST/GeLU).

### Axis alignment better distinguishes model fit to the brain

To investigate whether privileged axes depend on the type of architecture underlying a representation, we extend the previous analyses to the transformer model architecture. The transformer has recently emerged as another promising model class that can successfully capture neural activity and human behaviour across vision and audition<sup>46–48</sup>. Transformers do not contain convolutions but instead leverage parallel self-attention mechanisms to effectively capture contextual information by allowing every position in an input sequence (for example, a patch in an image) to attend to every other position. Each transformer layer further contains a feed-forward neural network, also known as the multi-layer perceptron (MLP) block, that operates independently on each position and introduces nonlinearity to the model through its activation functions (often, a softened non-negativity constraint such as Gaussian error linear unit (GeLU)). This complements the self-attention mechanism by enabling more complex transformations, equipping the transformer to model a rich set of functions.

Recent studies have revealed that transformer architectures achieve comparable performance to deep convolutional neural network architectures in predicting neural responses throughout the

sensory cortices<sup>49,50</sup>. Here we ask whether transformers also exhibit axis alignment with the brain. We extracted visual representations from two vision transformers trained on ImageNet with category labels: (1) a standard ViT architecture (ViT)<sup>47</sup> and (2) a hybrid model where the input sequence to the ViT is formed from intermediate feature maps of a regular ResNet50, henceforth denoted as ViT-RN50 (ref. 47). We also extracted sound representations from two Auditory Spectrogram Transformers (AST) trained on the AudioSet corpus with category labels and initialized with different random seeds<sup>48</sup>, henceforth denoted as AST-1 and AST-2. For each of these networks, we extracted the penultimate representation (output of the final attention block) as well as the representation in the last feed-forward layer with a non-linear GeLU activation function, labelled ‘GeLU’. We then compared each of these DCNN/transformer representations to the corresponding sensory areas using the most widely used brain–model similarity metric: linear predictivity after training a linear model that maps DNN activations to brain activity. We contrast this standard metric with our measures of axis alignment.

Although DCNNs and transformers exhibited comparable linear predictivity (as expected given previous results), they differed

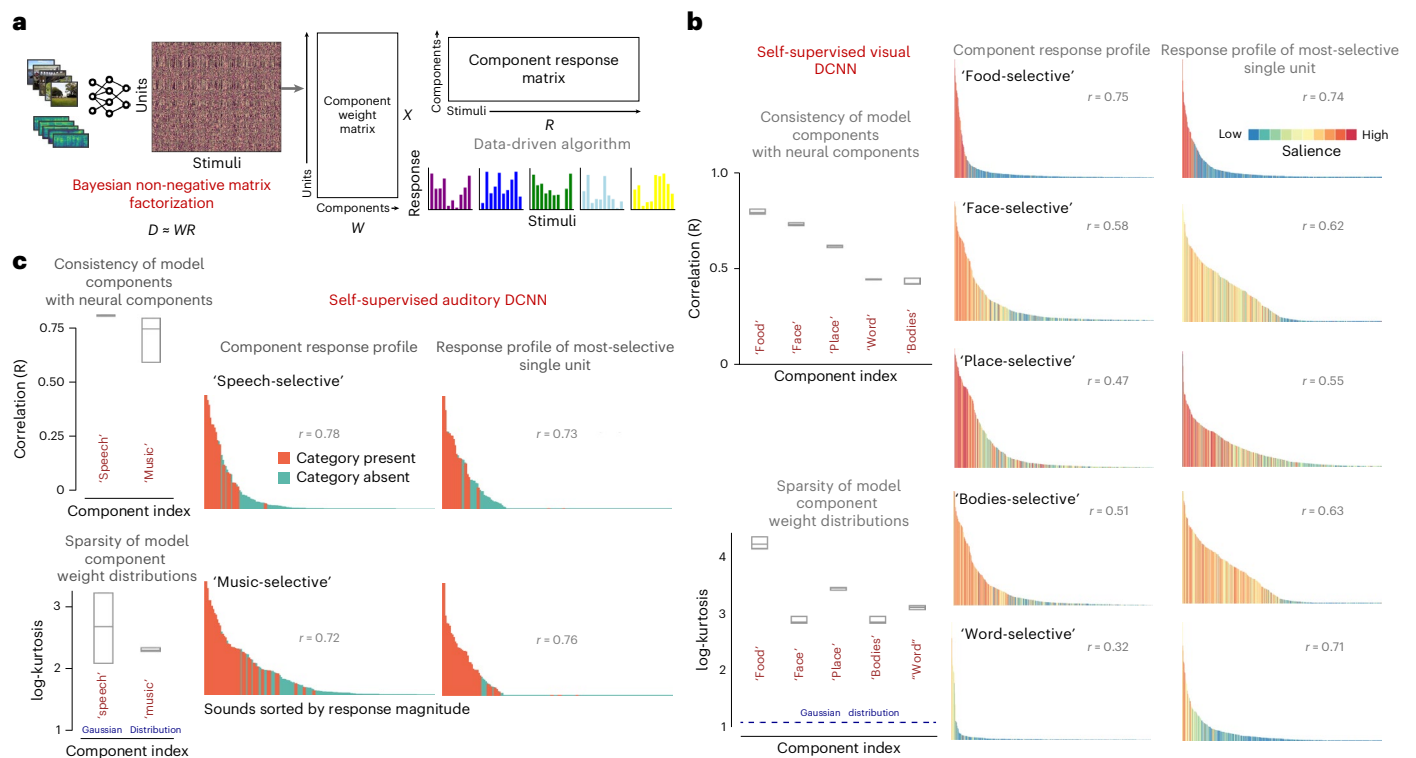


**Fig. 6 | What are the preferred axes in brains and networks? a**, Overview of the factorization approach applied to derive the dominant components of DCNN representations. Bayesian non-negative matrix factorization was used to decompose DCNN activations as a product of two lower-dimensional matrices: (1) a response profile matrix characterizing the response of each component to all stimuli and (2) a component-by-unit weight matrix that expresses the contribution of each unit to every component. **b**, Left: correlations between the top 5 visual DCNN components, sorted by their consistency with neural components, and their corresponding neural counterparts (top); weight distribution sparsity ( $\kappa_k$ ) of these top 5 DCNN components across units (bottom). Middle: response profile of each component across the 515 images viewed by all

NSD participants. Each bar corresponds to an image, with the colour denoting the behavioural salience rating for the preferred category of the component (for example, the salience of faces for Component 2). Correlations (Pearson's  $r$ ) between the magnitude of component responses and salience ratings for the preferred category of each component across all 515 images are reported. Right: response profile of the single unit that is most selective for the respective category. **c**, The same analysis as in **b**, but for auditory DCNN components. Correlations (Pearson's  $r$ ) of component responses to all 165 sounds with a binary vector indicating whether the preferred category of the component (music/speech) was present/absent in the stimuli are reported. Boxplots reflect variability over multiple iterations of the NMF.

markedly in the extent to which their axes were aligned to those of the brain (Fig. 5a–c, right). The representational axes of DCNNs were much better aligned to those of the brain than were those of transformers. This trend held true both for macaque IT recordings and for human ventral visual stream fMRI measurements, and for human auditory cortical fMRI measurements. Further, in the visual domain, the DCNN–brain axis alignment was close to the estimated noise ceiling (as measured by inter-participant axis alignment).

We next asked whether the representational axes of transformers exhibit a non-random relationship with sensory cortical representations despite being worse than that of DCNNs. To address this, we again computed axis-dependent alignment as a function of axis rotation. While the basis of the penultimate layer was nearly indistinguishable from an arbitrary basis by its alignment with neural representations in both audition and vision, the GeLU-layer axes were more aligned than an arbitrary basis. This result is consistent with the idea that nonlinearities are important for generating brain-like privileged axes (because the penultimate layer lacks an ensuing nonlinearity).

Altogether, these results show that by adopting a more stringent measure of similarity, our axis-alignment metric better discriminates which artificial network models best fit the brain.

### Privileged axes arise partially from category-selective tunings

What are the actual tunings of the privileged representational axes in brains and networks? In neuroscience, a wealth of evidence has revealed the presence of neurons tuned to meaningful features of our

external environment. For example, in the macaque inferotemporal cortex, many single neurons selectively respond to images of faces<sup>11</sup>. In humans, both neuroimaging studies and intracranial recordings within the visual cortex have identified neural populations tuned to images of faces, scenes, bodies, visually presented letter sequences or words, and food<sup>12,14–17</sup>. Distinct neural populations within the auditory cortex analogously show selectivity for music, speech and song<sup>30,42</sup>. It has remained unclear why such category-selective tuning is so consistently observed across neural recordings in distinct individuals. The findings in the previous section demonstrating aligned axes between DCNNs and sensory cortices raise the possibility that DCNNs possess similar tuning functions.

To explore this issue, we employed a data-driven, hypothesis-agnostic factorization approach to discover the dominant components underlying DCNN representations. The rationale was to reduce the representational dimensionality to a tractable number for interpretability analyses while minimizing biases towards predetermined categories. Models trained with supervision might plausibly acquire category selectivity via training labels that correspond to such categories. Accordingly, we instead analysed self-supervised models which do not receive any explicit supervision in the form of semantic labels, as these seemed the strongest test of the presence of selectivity related to particular categories.

We first aggregated the responses of all  $n$  units in the self-supervised visual DCNN (trained with momentum contrast) to  $m$  natural stimuli from the NSD, resulting in a data matrix  $D \in \mathbb{R}^{m \times n}$ .

We employed the Bayesian non-negative matrix factorization method previously applied to neural data<sup>17</sup> to factorize the DCNN responses to natural stimuli into two lower-dimensional matrices  $D \approx RW$ : (1) a response profile matrix  $R \in \mathbb{R}^{m \times k}$  that characterizes the response of each of  $k \ll n$  components to all  $m$  stimuli and (2) a component-by-unit weight matrix  $W \in \mathbb{R}^{k \times n}$  that expresses the contribution of each unit to every component's response. This approach is intrinsically axis dependent. We performed the same factorization approach on high-dimensional responses in the human ventral visual pathway in the NSD. For every DCNN component, we then computed its correlation with the top five neural components, which we had previously shown to be selective to faces, scenes, bodies, text and food, respectively<sup>17</sup>.

For every neural component, we found a highly correlated ( $-0.44$ – $0.80$ ) DCNN component. We visualized the response profiles of each of these matched DCNN components across all stimuli and correlated the component responses with the salience ratings for their corresponding category in each of these stimuli (Fig. 6). These salience ratings were collected in an independent behavioural experiment to quantify the salience of each of the five visual categories within individual stimuli<sup>17</sup>.

This analysis revealed that the DCNN component responses were each indeed highly selective to a different category. Are these category-selective component responses indicative of distributed responses among all units, or are they driven by a few specific units? To answer this, we next turned our attention to the component-by-unit weight matrix and computed the kurtosis of the unit weight distributions for each component. As shown in Fig. 6b (left, bottom), unit weight distributions were highly kurtotic, relative to a Gaussian, indicating a peakier, heavy-tailed distribution skewed towards higher values, suggesting that the components were indeed sparse and received contributions from only a few units, similar to the neural components. We also evaluated whether there were single units in the DCNN selective for each of these five categories. This analysis revealed single units whose responses were highly correlated ( $R \approx 0.44$ – $0.80$ ) with each of the categories, respectively (Fig. 6b, right). These findings are consistent with previous reports of category selectivity in DCNNs<sup>51,52</sup>, but extend previous findings by showing that category selectivity emerges in a hypothesis-neutral fashion as the dominant tuning in DCNNs, and in showing that these dominant responses are matched to those in the brain.

In the auditory self-supervised DCNN, the same factorization approach yielded an analogous result: sparse components highly selective for music and speech (Fig. 6c), again reminiscent of the category-selective tuning previously reported in the auditory cortices<sup>30</sup>. Analysis of single DCNN unit responses further reveals that some single units themselves exhibit selectivity for these complex high-level categories (Fig. 6c, right), consistent with one recent report of music selectivity in audio-trained DCNNs<sup>53</sup>. Control experiments showed that the observed selectivity for high-level abstract categories such as music or speech was not explained by generic acoustic features, such as spectrotemporal modulations (Supplementary Fig. 4).

Are these aforementioned categories more prominent in the tuning functions of the native axes than in arbitrary axes? To find out, we computed the number of units selective for each of these categories in both the native and arbitrary axes. As in previous analyses, we generated a sequence of rotation matrices that interpolate between a uniformly drawn random rotation matrix and the identity matrix

(no rotation). We determine the count of category-selective units in each of these representational bases for both sensory cortical representations and self-supervised DCNN representations. To compute selectivity, we contrast the response of each unit to stimuli containing the category in question against all other stimuli with a  $t$ -statistic (following the standard approach of functional localizers in neuroimaging experiments), comparing it to a fixed  $t$ -value threshold to identify the number of selective units. We used an image set encompassing text, bodies, faces and places<sup>54</sup>, and the 165 sounds from the auditory dataset<sup>30</sup>.

As shown in Extended Data Fig. 8, category-selective tuning is most pronounced in the native axes, with deviations from these axes consistently reducing single-unit selectivity. This result provides evidence that category-selective tuning is a substantial contributor for the privileged axes found in sensory systems and models. We note that the alignment of representational axes in penultimate DCNN layers with human-interpretable categories emerges independently of DCNN task performance, given that arbitrary rotations of the axes preserve downstream task performance.

The emergence of a sparse category-selective tuning structure in neural networks sheds light on their origins in the brain. While prevailing theories often attribute category selectivity in brains to downstream behavioural pressures and the ecological importance of these categories, our findings show that category selectivity can arise without any of these explicit pressures. Its presence in self-supervised models that are not optimized for a behavioural task indicates that category selectivity can arise simply from the statistics of our natural input and optimization for contrastive objectives<sup>51,52</sup>, potentially in conjunction with particular architectural features such as nonlinearities.

### Axis alignment is not limited to category selectivity

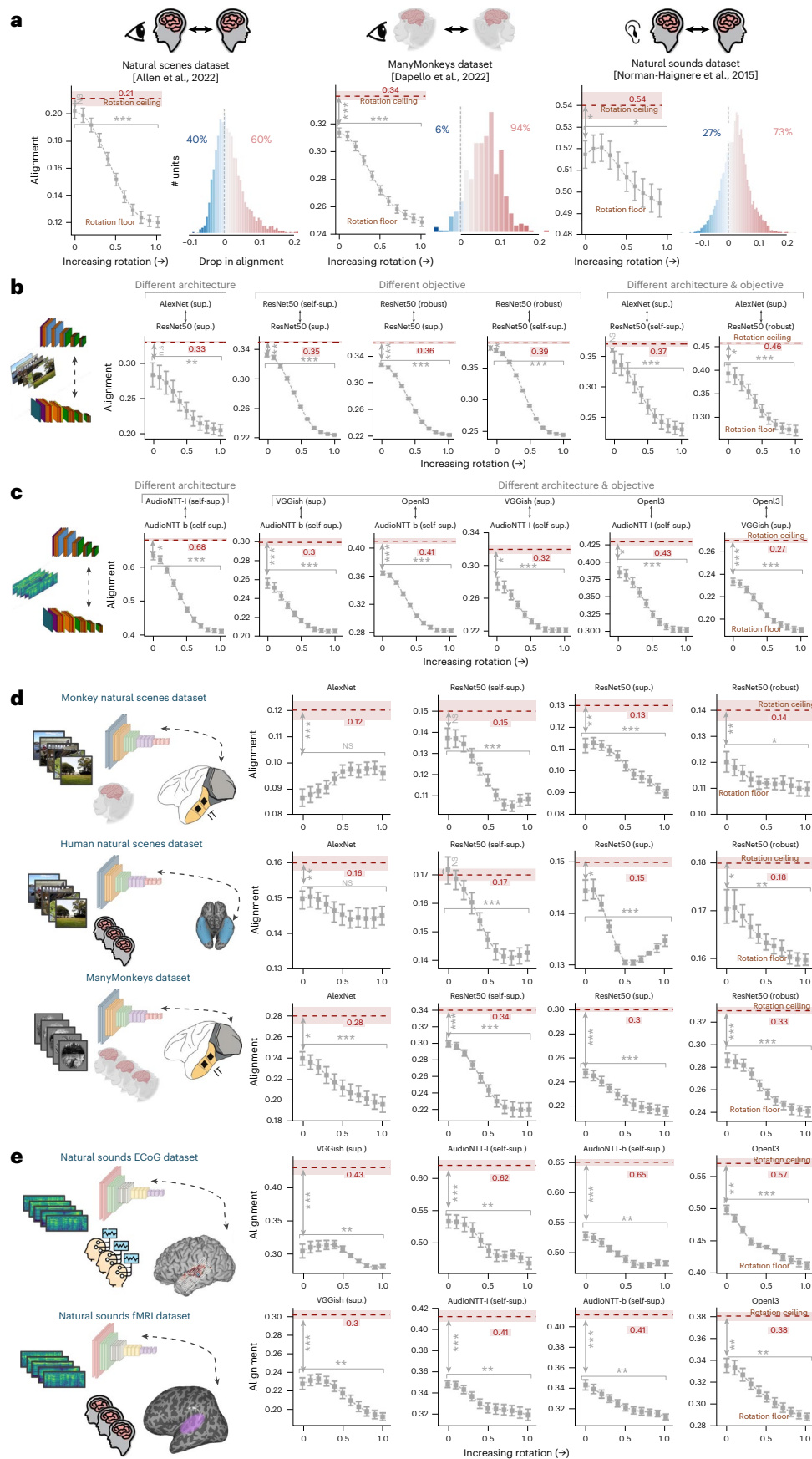
Previous analyses have shown that category-selective tuning functions—such as those for faces, bodies, places, visually presented words, and food in the visual domain, or music and speech in the auditory domain—are dominant components of neural representations in both biological and artificial neural networks. Do the privileged axes in these systems, or the aligned axes between them, reflect only these category-selective responses? Or do they represent a broader principle that transcends specific category selectivities?

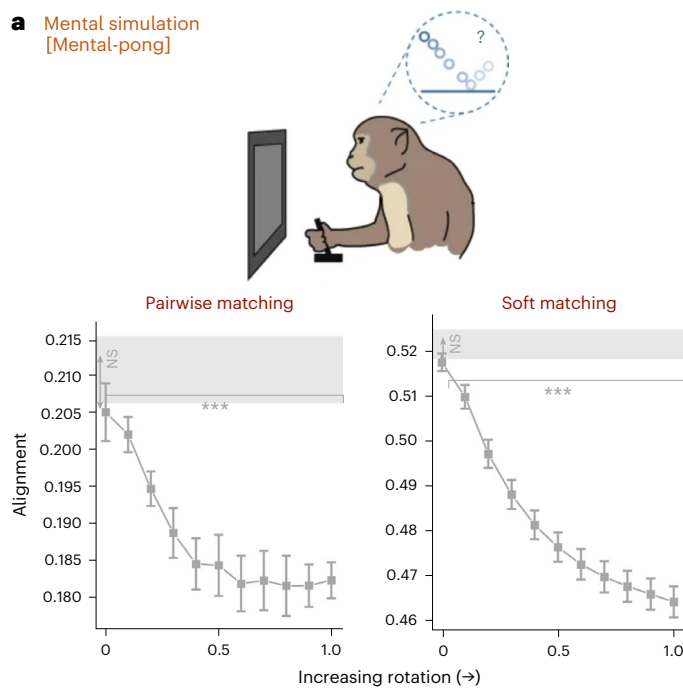
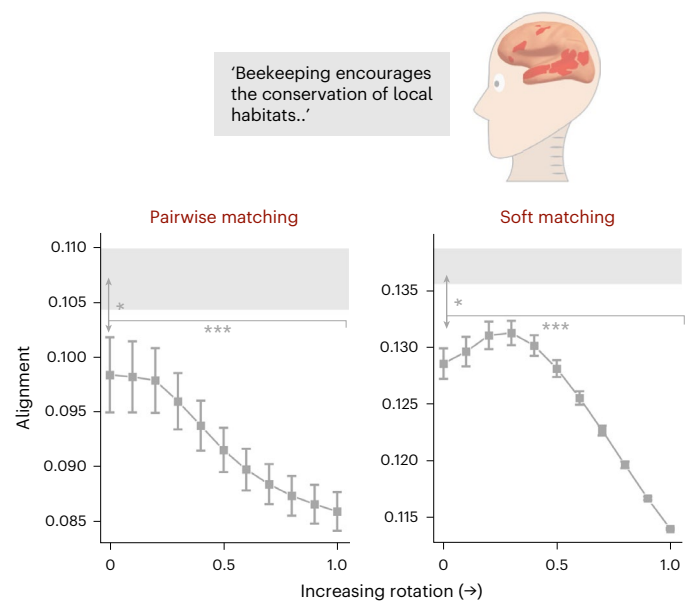
To address this question, we reanalysed privileged axes across brains, networks and their alignment while excluding category-selective neurons, voxels and units for each category with known functional specialization in the brain. A liberal category-selectivity criterion was used to exclude visually category-selective units for the five major visual categories mentioned above and auditory units selective for music or speech (that is, units selective for music, speech or both were excluded).

These exclusions were designed to test for alignment effects over and above category-selective responses. As shown in Fig. 7, the finding of axis alignment remains even after excluding all category-selective units, voxels and electrodes: axis alignment is preserved across brains, networks and their alignment with each other. Note however that particularly for the visual fMRI data, a smaller percentage of non-category-selective voxels lost alignment with rotation (60%, Fig. 7a) than for the full set of voxels (81% in Fig. 1d), indicating that a large proportion of axis-aligned voxels were category selective in this dataset.

**Fig. 7 | Axis alignment is not limited to category selectivity.** Same analyses as for Fig. 3 after removing all category-selective units/voxels. **a**, Brain-to-brain: Alignment between conspecifics neural responses in human ventral visual cortex, macaque IT and human auditory cortex after removing all category-selective voxels/electrodes. **b,c**, Model-to-model: Alignment between different visual (**b**) and between different auditory (**c**) deep convolutional neural networks after removing all category-selective units. **d,e**, Model-to-brain: Alignment of different visual (**d**) and auditory (**e**) deep convolutional neural networks with

neural representations in the primate visual and auditory cortex, respectively, after removing all category-selective units in networks and category-selective voxels/neurons/electrodes in brain recordings. Error bars and grey shaded regions across all panels indicate s.e.m. across  $n = 5$  stimulus train/test splits per comparison. Significance tests were two-sided paired  $t$ -tests comparing alignment values at no rotation ( $\alpha = 0$ ) with full rotation ( $\alpha = 1$ ), as well as with the optimal alignment values observed across all rotation angles. \* $p < 0.01$ , \*\* $p < 0.001$ , \*\*\* $p < 0.0001$ . The largest annotated  $p$  value was 0.002.



**a** Mental simulation  
[Mental-pong]**b** Language  
[Pereira, 2018]**Fig. 8 | Privileged axes extend to higher-order cortical representations.**

**a**, Alignment between conspecifics neural responses in macaque dorsomedial frontal cortex during a mental simulation task, measured using the pairwise matching (left) and soft-matching correlation (right) scores and assessed as a function of axis rotation ( $\alpha$ ) in one animal's brain. **b**, Alignment between pairs of conspecifics neural responses in the human language cortex during sentence comprehension, measured using the pairwise matching (left) and soft-matching correlation (right) scores and examined as a function of axis rotation ( $\alpha$ ) in

one participant's representation. 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. For **a** and **b**, error bars and grey shaded regions indicate s.e.m. across  $n = 5$  stimulus train/test splits per comparison. Significance tests were two-sided paired  $t$ -tests comparing alignment values at no rotation ( $\alpha = 0$ ) with full rotation ( $\alpha = 1$ ), as well as with the optimal alignment values observed across all rotation angles. \* $p < 0.01$ , \*\*\* $p < 0.0001$ . The largest annotated  $p$  value was 0.0032.

In a few instances of model-to-brain alignment, the difference between native and rotated axes was less clear after excluding category-selective units. This outcome could be attributed either to the small number of remaining units, neurons or voxels after applying the exclusion criterion or to an intriguing possibility that axis alignment in these cases is largely driven by category-selective tuning. The latter interpretation could offer a promising avenue for distinguishing between models. Regardless, it is clear that in the majority of cases, there is substantial axis alignment even without category-selective units.

**Preliminary evidence for privileged axes in higher-order cortex**

Are privileged axes exclusive to sensory domains, or could neural systems supporting higher-level cognition also privilege certain coordinate systems over others? Here we focus on two such high-level brain regions: the human language cortex, associated with language comprehension and production, and the macaque dorsomedial frontal cortex (DMFC). The following analyses are a first-pass extension of our framework to language and frontal areas. They demonstrate generality but a full mechanistic account is beyond the scope of the present paper and will be pursued in future work.

To extract language cortex representations, we analysed fMRI activations measured in 9 participants as they read 384 sentences, each representing a unique stimulus condition, from short passages covering a wide variety of topics<sup>55</sup>. For the DMFC, we examined neural activity in a large population of neurons in 2 macaques during a ball interception task (dubbed 'mental-pong')<sup>56</sup>. Here, the start position of the ball was randomized in each trial, resulting in 79 unique conditions.

We calculated representational alignment for each pair of human participants/animals within these datasets, both in the native axes and across a sequence of arbitrary axes derived from axis rotations.

Even in the language and frontal areas, alignment was much higher for the native axes than for any arbitrary axes (Fig. 8). These results raise the possibility that the phenomenon of privileged axes may be ubiquitous throughout the cortex. We leave a thorough characterization of these privileged axes as well as the parallel analysis of computational models of these high-level cognitive abilities for future work.

**Discussion**

How is information coded in the brain? To answer this question, neuroscientists have increasingly adopted methods that analyse the geometry of the population code that are blind to the actual tuning of neurons. But are neural tunings in fact arbitrary and irrelevant, or might they matter, privileging some representational axes over others? We developed methods to probe for privileged representational axes in biological and artificial neural networks, and applied them to multiple types of neural data from diverse brain systems, and to DCNNs trained on natural sensory stimuli. We found that representational axes were consistent between individuals, between artificial neural networks varying in architecture and learning objectives, and between brain systems and artificial neural networks trained on the same modality. These results indicate that representational axes in neural systems are not arbitrary and can arise in artificial systems with none of the priors and meanings assigned to these axes by humans. We further provide evidence linking the prevalence of privileged axes in networks to nonlinear activation functions such as ReLU, and show that the privileged axes in both the brain and DCNNs are those that yield sparse representations, reduce downstream wiring costs and that aid generalization by downstream classifiers under constrained learning scenarios. Finally, our metric of axis alignment also distinguishes the fit of models to the brain that are not well discriminated by standard metrics.

A key contribution of the present work is the development and application of quantitative measures to formally test whether the representational axes in different animals/participants are more similar than would be expected from arbitrary bases, and on the other hand whether they approach the maximum alignment possible. Our findings provide, to our knowledge, presumably the first quantitative evidence demonstrating that high-level sensory cortices possess preferred axes for encoding natural stimuli, building on earlier qualitative observations of interpretable tuning functions in individual neurons. Our findings do not dismiss the utility of population-level analyses in understanding the neural code, but rather propose that such analyses should not ignore neural tuning functions. We found that among the large set of possible population codes with identical geometry and content, only some are empirically observed. We later discuss potential mechanistic reasons for why certain assignments of representations to individual neurons are favoured, along with their computational implications.

This axis-dependent question differs from that addressed by previous hyperalignment methods. Such methods learn the optimal rotation to match different participants' data, thereby discounting all rotational degrees of freedom and effectively erasing any information carried by the original axes<sup>37</sup>. These approaches thus remain agnostic to how strongly the native axes themselves drive inter-participant correspondence. By contrast, our axis-dependent framework restricts alignments to voxel reassignments or soft permutations and then systematically perturbs the axes themselves, directly measuring how much the native coordinate system contributes to shared representational structure. This distinction allows us to reveal, and quantify, the truly privileged nature of the brain's representational axes.

A parallel question arises for artificial neural networks. Previous research has offered hints of privileged axes in artificial neural networks, such as the emergence of edge detector units or units selective for specific semantic concepts across a variety of networks<sup>34,58–61</sup>. Such observations suggest a possible alignment of unit tuning curves beyond just the alignment of representational content. However, these qualitative observations of tuning similarity leave unclear whether the alignment of tuning functions across networks exceeds what would be expected by chance. Our approach allows this comparison and reveals that the axis alignment across different models far exceeds that observed in arbitrary bases, indeed approaching the maximal possible alignment achievable through any axis rotation. While the existence of privileged axes in artificial neural networks does not guarantee that these axes will be interpretable to humans, our experiments suggest that in many cases they are. At the very least, our findings provide a clear rationale for focusing on the native axes, rendering interpretability analyses of these bases not just convenient, but also well founded.

The finding of privileged axes in both brains and models also reframes longstanding claims about representational convergence. A large body of work supports the idea that different computational architectures often end up learning similar representations (known as the universality, or 'convergent learning' conjecture in machine learning)<sup>61–63</sup>. Over the past decade, it has been suggested that this universality also extends to comparisons between artificial and biological networks, with task-optimized neural networks argued to reproduce aspects of biological representations in multiple domains, including vision<sup>64,65</sup>, audition<sup>50,66</sup> and language<sup>67–71</sup>. This representational convergence across brains and models is taken to reflect the strength of behavioural constraints and the statistical regularities of our natural world in narrowing down the solution set for a nervous system. However, the results supporting this convergence primarily rely on measures of representational content that are invariant to axis rotations, most commonly neural predictivity under a linear mapping model and representational similarity analysis. Our results show that the convergence of biological and artificial neural systems goes beyond representational content to encompass representational form. This conclusion is consistent with several recent papers that also note

alignment between the representational units of the visual cortex and DCNNs<sup>31,72,73</sup>. The comparison against arbitrary bases in our work underscores the non-trivial nature of this similarity.

What causes brains and DCNNs to converge onto similar axes? Learning objectives by themselves should not result in preferred axes (because axis rotation preserves task performance for linear readouts). Rather, they must interact with the neural architecture to yield consistent axes. In artificial neural networks, particularly those employing ReLU nonlinearity, we can understand the representational invariances and the mechanisms leading to the emergence of privileged axes through theoretical constructs such as the set of operations that commute with the nonlinearity. Although it is theoretically possible for neural networks to learn hidden layer representations that can be inverted after ReLU operations, our empirical results demonstrate that this is not typically the case: training networks to align with arbitrary axes results in a significant performance drop. This finding aligns with recent work<sup>74</sup>, which explored whether orthonormal transformations of a representation can be inverted by subsequent layers. Their experiments, particularly with Myrtle CNNs trained on CIFAR-10, show that such inversions do not always restore task performance. Taken together, these results suggest that the non-invertibility of activation functions in neural networks limits the set of axes that can contribute to effective task performance. We speculate that the privileged axes arising in the brain might be similarly causally linked to the nonlinear activation functions of neurons, and that such nonlinearities could explain the axis alignment between biological and artificial neural networks.

These findings raise the question of why particular axes might matter computationally. Representations are best understood in the context of the computations that transform them for use in downstream tasks<sup>75</sup>. While there have been many studies comparing the downstream impact of different kinds of representation, for instance, analysing the effect of the dimensionality of a representation or how representations with mixed selectivity compare to representations with 'pure' selectivity<sup>76</sup>, little attention has been paid to the impact of representational axes. We proposed several respects in which representational axes could impact computation, all related in some way to sparsity. First, axes affect the efficiency of the neural code. We found that axes in both biological and DCNNs are characterized by heavy-tailed (kurtotic) response distributions when processing natural stimuli. High kurtosis is one measure of lifetime sparseness and is associated with efficient (non-redundant) representations<sup>77</sup>. This finding is consistent with previous reports of high lifetime sparseness in early and high-level visual areas<sup>43,78,79</sup>. Second, the axes affect downstream wiring costs. Readouts in the brain are likely to have to be restricted to small subsets of all possible neurons to avoid prohibitive volumes of brain tissue. We found that the axes of both biological networks and DCNNs facilitate sparser readouts of category information compared with arbitrary bases. This finding may relate to a recent study of neural responses in primate prefrontal cortex during a working memory task, revealing how neural implementations of representational geometries influence wiring costs<sup>80</sup>. Third, axes affect the capacity for generalization. If connections are constrained to be sparse, the generalization of any readout of a representation is dependent not only on the representation's geometry but also on the tuning of representational units. We found that the axes observed in the brain support better few-shot generalization under sparse linear readouts compared with arbitrary bases. Our theoretical analysis suggests that any task that relies on contributions from a small subset of neurons will show better few-shot generalization if the representational axes are those in which natural stimuli are sparsely represented. A similar relationship between disentanglement and improved generalization in the context of multitask learning was also noted in recent theoretical work<sup>81</sup>.

These computational consequences also highlight why axis-dependent alignment measures are useful in comparative analyses. Given our finding that systematic neural tuning is a basic fact

of brains, it requires explanation, and metrics that are blind to this phenomenon are incomplete at best. Our test of axis alignment fills this gap and enables better discrimination between models in terms of their fit to neural data. For instance, DCNNs exhibit greater similarity to biological sensory systems than transformers when evaluated via axis alignment, despite being indistinguishable under axis-independent measures. This divergence reflects fundamental differences in neural coding: CNNs employ hard ReLU nonlinearities that break rotational symmetry, yielding stronger form-alignment to cortex, whereas transformers use softer nonlinearities (GeLU, attention operations) that may be responsible for the weaker alignment. The CNN–transformer gap may thus point to a concrete biophysical hypothesis: single-neuron tuning functions in cortical circuits may be preferentially shaped by ReLU-like symmetry-breaking operations. Further, our finding of one-to-one mappings between units in networks and units in brains constitutes evidence of a more literal and stringent kind of similarity between networks and brains (see also ref. 52), more strongly supporting the use of these networks as models of the brain. These results indicate that measures of axis alignment, along with other recent model–brain comparison metrics, should be added to commonly employed batteries of model–brain comparisons, not merely as a stricter test but as a tool for revealing mechanistic principles invisible to rotation-invariant metrics.

Given that brains and ANNs share privileged representational axes, what are they? To address this question in a data-driven rather than ad hoc fashion, we used a sparse factorization algorithm to pick out the dominant tuning functions in both deep neural network representations and the corresponding neural systems. In both cases, the predominant tuning functions were category selective for the categories previously identified in previous studies of neural tuning (faces, places, bodies and words for visual networks and brain regions, and speech and music for auditory networks and brain regions), despite not being constrained in any way to be interpretable or to conform to presupposed hypotheses. Two recent papers also described category-selective tuning in networks trained on visual self-supervised learning objectives<sup>51,72</sup>.

Our study further shows that brain-like category selectivity in individual network units (1) is not confined to visual processing, but also extends to auditory models, (2) is a dominant feature of high-level DCNN representations, as the specialized categories emerge spontaneously from a data-driven decomposition of network activations and (3) is not an inevitable result of any high-level representational space, since rotating the axes of the representation preserves the content but eliminates selective responses. These together suggest that category selectivity can arise purely from broad pressures—the statistical structure of natural inputs in conjunction with generic architectural constraints, without specific priors for any of these categories (see also ref. 51) and without invoking any narrowly tailored objectives for associative-memory capacity, superposition avoidance, energy efficiency or other downstream post-perceptual processing<sup>79,82,83</sup>.

Importantly, however, we find axis alignment across brains and models even after removing category-selective neurons, voxels or units, indicating that axis alignment reflects a more general principle of coding.

It is also important to distinguish privileged axes from related ideas about distributed coding and mixed selectivity. Several recent lines of work address the broad question tackled here of whether neural codes are best understood by considering the actual tunings of neurons versus the geometry of neural population codes. However, these other lines of work differ substantively from the concept of privileged axes introduced here. First, the idea of mixed selectivity (or superposition in neural networks<sup>84</sup>) refers to the tendency of neurons to code not for single interpretable dimensions of the stimulus or task<sup>85</sup>, but instead for mixtures of these dimensions<sup>86,87</sup>. Our concept of privileged axes is orthogonal to this distinction, since privileged axes need not be

interpretable. Instead, privileged axes arise from systematicity, specifically, the consistency observed across instances of a representational system. If neurons across different instances are systematically modulated by the same mixture of stimulus variables, this would still qualify as evidence for privileged axes. A related question<sup>88</sup> is whether neurons cluster into functional categories with similar response profiles<sup>80</sup>, consistent with classic findings, or whether the distribution of neural tunings reveals no such categorical structure<sup>89</sup>. (The idea that neurons may cluster in their response profiles is conceptually separable from mixed selectivity, since one could in principle find clusters of neurons with the same mixed selectivity<sup>90</sup>.) Similar to our concept of privileged axes, clustering of neurons into functional categories with similar selectivity may suggest that some neural tunings are favoured over others. Yet, a mere clustering of neural selectivities lacks the critical evidence presented here that those selectivities are systematically aligned across individuals, and that such alignment between two individuals is non-random in that it is lost when the representational space of one individual is rotated.

Our work leaves many questions unanswered. First, assessing the prevalence of privileged axes in the brain is limited by existing neural activity data, which are currently limited in either coverage (electrophysiology) or spatial and temporal resolution (fMRI). Thus, it is possible that other representational axes may exist that are less consistent across participants or between brains and ANNs. Second, our reliance on gradient-based optimization procedures to derive the optimal alignment could present limitations in certain situations and for certain models (by working less well due to local optima). Third, while our work suggests that nonlinearities could play a role in determining tuning functions, it remains unknown which nonlinearities produce brain-like privileged axes in computational models of neurobiological systems. Fourth, our theoretical framework and empirical results explain why similar activation bases are consistently preferred in network representations, but they do not provide a clear answer to what these activation bases should be, particularly in terms of neural tuning. For example, why do we observe category selectivity in the privileged axes? This is a complex and difficult theoretical question. Recent work has provided theoretical evidence that in the presence of biological constraints on neurons such as non-negativity (as routinely employed in modern DCNNs via ReLU operations) and sparsity, single neurons become aligned to axes of data generative factors<sup>91</sup>. However, whether this happens for realistic natural sensory input is an open question. Studies of how experience with varying sensory diets shapes the tuning functions in in-silico models could provide insight. Some studies have offered theoretical insights into the emergence of certain kinds of ‘privileged axes’ in deep networks, but these often rely on simplifying assumptions, even in relatively simple models such as single-layer networks<sup>92</sup>. Providing a fully developed theoretical framework for the privileged axes observed in our study is an important question for future research. Finally, the concept of axis alignment in the brain is scale dependent. We focused here on a meso-scale level of analysis, considering representations within modality-specific cortices and networks and finding privileged axes in all cases we examined. But this work leaves open the question of whether there is a scale of analysis at which representations transition from privileged axes to arbitrary axes. Our preliminary analyses of axis alignment within category-selective regions of the high-level visual cortex suggest that at this finer scale, voxel responses in conspecifics are no more aligned in the native basis than in an arbitrary basis, providing hints that at this scale, representations may not have privileged axes (see Extended Data Fig. 10). However, these conclusions are limited by the coarse resolution of fMRI. To validate these initial observations, further research employing higher spatial resolution within category-selective patches is essential.

Perhaps most fundamentally, our work does not answer why biological and artificial networks have representational dimensions selective to some categories of input (for example, faces, places, speech)

but apparently not others. One speculative hypothesis is that the data generating causes in the natural world consist of sparsely interacting parts, and DNNs that perform (generalize) well learn to represent these parts in individual neurons in the presence of certain nonlinearities. This idea resembles the ‘Platonic representation hypothesis’<sup>93</sup>, but takes it one step further to argue that not only do networks (and brains) tend to converge on the causal structure of the world, but also that many of the key dimensions of the world will be reflected in neurons selective for that dimension.

## Methods

### Brain activity datasets

**Natural scenes dataset.** A detailed description of the Natural Scenes Dataset (NSD; <http://naturalscenesdataset.org>) is provided elsewhere<sup>28</sup>. Here we briefly summarize the data acquisition and preprocessing steps. The NSD dataset contains measurements of fMRI responses from 8 participants who each viewed 9,000–10,000 distinct colour natural scenes (22,000–30,000 trials) over the course of 30–40 scan sessions. Scanning was conducted at 7 T using whole-brain echo-planar imaging at 1.8-mm resolution and 1.6-s repetition time. Images were taken from the Microsoft Common Objects in Context (COCO) database<sup>94</sup>, square cropped and presented at a size of 8.4° × 8.4°. A special set of 1,000 images were shared across participants; the remaining images were mutually exclusive across participants. Images were presented for 3 s with 1-s gaps in between images. Participants fixated centrally and performed a recognition task in which they were instructed to indicate whether they have seen each presented image at any point in the past. Informed consent was obtained from the participants and the study was approved by the University of Minnesota Institutional Review Board.

The fMRI data were preprocessed by performing one temporal interpolation (to correct for slice time differences) and one spatial interpolation (to correct for head motion). A general linear model was then used to estimate single-trial beta weights, with the haemodynamic response function estimated for each voxel and the GLMdenoise technique used for denoising<sup>95,96</sup>. In this work, we used the 1.8-mm volume ‘nativesurface’ preparation of the NSD data and version 3 of the NSD single-trial betas (‘beta\_fithrf\_GLMdenoise\_RR’). Every stimulus considered in this study had 3 repetitions, and we only analysed data in the four NSD participants who had full 3 repetitions for each of their 10,000 images. To derive voxel responses to each stimulus, we averaged single-trial betas after z-scoring every voxel within each scan session.

We selected ventral visual stream voxels by using the ‘streams’ atlas provided in the native space of each participant in NSD<sup>28</sup>. Briefly, this region of interest (ROI) collection reflects large-scale divisions of the visual cortex into primary visual cortex and intermediate and high-level ventral, lateral and dorsal visual areas. These were manually drawn for each participant by NSD curators and were based on voxel-level reliability metrics. For this study, we extracted the ROI mask corresponding to the ‘higher-level ventral stream’ label. This ROI was drawn to follow the anterior lingual sulcus (ALS), including the anterior lingual gyrus (ALG) on its inferior border, and to follow the inferior lip of the inferior temporal sulcus (ITS) on its superior border. The anterior border was drawn on the basis of the midpoint of the occipital temporal sulcus (OTS). This parcel is very broad and results in 7,000–9,000 voxels per participant.

**ManyMonkeys dataset.** The ManyMonkeys Dataset, as described in refs. 97,29, comprises neural recordings from the inferotemporal (IT) cortex of 6 rhesus monkeys, captured during a passive viewing task. This dataset features a collection of neural responses to a set of 640 high-definition images. Each image represents a 2D rendering of a 3D object model superimposed on a randomly generated background, with variations in the object’s position, size and orientation. The diverse stimulus array includes images of bears, elephants, human faces, apples, cars, dogs, chairs, airplanes, birds and zebras.

Neural activities were monitored through the implantation of 2–3 micro-electrode arrays in the IT cortex of the monkeys, accessing ~58–280 neural sites in each monkey. These arrays were strategically placed to sample a breadth of neural sites across the posterior–anterior axis of the IT cortex. Spike events were identified in the voltage traces on the basis of a threshold-crossing event, specifically, when the falling edge of the voltage signal dropped below 3 standard deviations of the baseline voltage fluctuations.

For each image presented, the response of a neural site was quantified as the average firing rate within a 70–170-ms window following the image’s appearance.

**Monkey-NSD dataset.** The Monkey-NSD Dataset, collected by Yoon Bai in Jim DiCarlo’s laboratory at MIT, comprises neural recordings from the IT cortex of one monkey to the shared 1,000 images viewed by all participants of the NSD. It encompasses a total of 116 neural recording sites, with 49 located in the posterior IT region and 67 in the anterior IT. For this particular dataset, an advanced spike-sorting procedure was employed to identify spike events, offering an enhancement in the reliability of the recorded data compared to traditional spike-thresholding methods.

**Auditory datasets. fMRI.** The fMRI data analysed here were originally collected and analysed in ref. 30. We summarize the data collection procedure with slight modifications from the original study for brevity.

Ten individuals participated in the study, each undergoing 2–3 fMRI sessions, each lasting 1.5 h. They were scanned while listening to a set of 165 natural sounds, each lasting 2 s, designed to broadly sample common auditory experience. The experimental design utilized a block structure, with each 17-s block comprising 5 presentations of one sound clip separated by scanner acquisitions. Within each block, one sound was presented at a 7-dB lower intensity to enable a sound intensity discrimination task. Participants were instructed to respond to this quieter sound by pressing a button. Following each sound, a corresponding fMRI volume was acquired using sparse scanning techniques that avoid temporal overlap between auditory stimulus presentation and scanner acquisition sounds. This resulted in a total of 1 run per participant, each including 15 sound blocks interspersed with 4 silence blocks to facilitate baseline response measurement. The functional scans were performed on a 3-T scanner equipped with a 32-channel head coil, with a repetition time of 3.4 s and an in-plane resolution of 2.1 × 2.1 mm, leading to a voxel size of 2.1 × 2.1 × 4 mm.

The acquired fMRI volumes underwent several preprocessing steps using FMRIB Software Library and custom MATLAB scripts. Motion and slice timing corrections were applied, non-brain tissue was removed, and the voxel time series were linearly detrended. Each functional run was co-registered to the participant’s anatomical volume. Subsequently, these volumes were mapped to vertices on the cortical surface reconstructed via FreeSurfer, and spatial smoothing was applied using a 2D Gaussian kernel with a full width at half maximum (FWHM) of 3 mm.

The average response of a voxel to each sound was computed by analysing the acquisitions from the second to the fifth post-stimulus onset, discounting the first to account for haemodynamic delay. These averaged responses were expressed as percent signal change (PSC) relative to the response during silence blocks. The PSC data were then interpolated onto a 2-mm isotropic grid on the cortical sheet.

Adopting the criteria from ref. 30, we selected voxels within a defined anatomical region that included the superior temporal and posterior parietal cortex, whose response reliability for the auditory stimuli exceeded a criterion. This resulted in a dataset of 11,065 voxels across all participants, with an average of 1,106.5 voxels per participant (range 799–1,498).

**Electrocorticography.** The human intracranial ECoG recordings were initially acquired for a separate study<sup>42</sup>. The stimulus set

for this dataset comprised the same set of 165 natural sounds used in the fMRI study described above<sup>30</sup>. The study involved 15 individuals diagnosed with epilepsy and implanted with subdural electrode grids for clinical purposes. Superior temporal gyrus (STG) exposure was notably greater in one hemisphere for each participant (8 with right hemisphere placement and 7 with left hemisphere placement). The predominant electrode configuration featured a 2.3-mm-diameter electrode with an inter-electrode distance of 6 mm on temporal lobe arrays (with a 10-mm spacing on arrays positioned in the frontal, parietal and occipital lobes, although these generally did not yield consistent sound-evoked responses). Two patients received implants with a denser grid arrangement (1-mm exposed diameter, 3-mm inter-electrode spacing). In addition, one individual was fitted with stereotactic electrodes in lieu of ECoG grids. Participants completed a variable number of experimental runs, ranging from 3 to 7.

The selection of electrodes for the final analysis was guided by the reliability of the broadband gamma activity (70–140 Hz) elicited by the acoustic stimuli, specifically the split-half correlation between responses to odd and even presentations, which had to exceed 0.2. This criterion resulted in inclusion of 272 electrodes across all participants. The signals from these electrodes were subjected to a common-average referencing process, calibrated against the mean across all electrodes within each participant. To ensure the purity of this referencing, electrodes with abnormal 60 Hz power levels were omitted. Subsequent to this referencing, a notch filter was applied to attenuate 60 Hz electrical noise and its harmonics. Broadband gamma power was quantified by the signal's envelope within the 70–140 Hz filtered range. The response of each electrode during a 3-s window locked to sound onset was measured, with each auditory stimulus lasting 2 s. These responses were normalized to percent signal change from baseline (the response during silence) by calculating the deviation from, and then dividing by, the average activity in the 300-ms pre-stimulus interval. For comparative analysis with neural network models, we averaged the responses within the 3-s post-onset window for each electrode to each stimulus, treating each electrode as a representational unit. We also broke down the analysis of model-to-brain alignment in different temporal windows for this ECoG dataset and found that the raw alignment scores as well the difference in alignment of the auditory DCNN representation with auditory cortex representations at no rotation and full rotation of the DCNN axis peak at ~300 ms following stimulus onset (Supplementary Fig. 3).

### Model selection

We compiled a set of models spanning different architectures and learning objectives that we could compare to brain data. We analysed late model stages within each candidate model on the grounds that they provide a hypothesis of a neural implementation of high-level auditory or visual processing. Specifically, we analysed the last convolutional layer for CNN models and the last attention layer for transformer models. For CNN models, we averaged the activations across the spatial dimensions to derive a single response value per unit (channel/filter). We note that our analysis is equally applicable to flattened feature representations (that is, that are not spatially pooled). In Supplementary Fig. 6, we demonstrate that our results hold equally well in this alternative setting. Specifically, when examining the alignment of AlexNet representations with the high-level ventral visual stream in humans (natural scenes dataset), flattening the feature representations (with rotations applied along the channel dimension before flattening) yields both qualitatively and quantitatively similar outcomes to those with spatially pooled activations. For transformer models, we averaged across tokens.

In the visual domain, we investigated a set of  $n = 6$  candidate models, spanning deep convolutional and vision transformer architectures, all trained on ImageNet<sup>32</sup>. The candidate models were:

- **ResNet50 (supervised)**: a convolutional architecture with residual connections for image classification. We used the `avgpool` layer before the final fully connected layer, yielding 2,048 units.
- **AlexNet (supervised)**: a convolutional architecture for image classification. We used the `avgpool` layer before the fully connected layers, yielding 256 units.
- **ResNet50 (self-supervised)**: a ResNet50 architecture trained using contrastive learning with momentum contrast. We used the `avgpool` layer before the final fully connected layer, yielding 2,048 units.
- **ResNet50 (robust)**: an adversarially trained ResNet50 architecture for image classification. We used the `avgpool` layer before the final fully connected layer, yielding 2,048 units.
- **ViT**: a standard vision transformer architecture for image classification. We analysed either the `norm` layer after `blocks.11` and before the classification head, yielding 768 units, or the `blocks.11.mlp.act` GeLU activations, yielding 3,072 units.
- **ViT-RN50**: a hybrid architecture in which the input sequence to the ViT is formed from intermediate feature maps of a ResNet50. As with ViT, we analysed either the `norm` layer after `blocks.11` and before the classification head, yielding 768 units, or the `blocks.11.mlp.act` GeLU activations, yielding 3,072 units.

In the auditory domain, we investigated a set of  $n = 6$  candidate models, spanning deep convolutional and transformer architectures, all trained on AudioSet<sup>37</sup>. The candidate models were:

- **VGGish (supervised)**: a VGG-inspired architecture for audio classification. We used the `pool4` layer, yielding 2,048 units.
- **AudioNTT-I (self-supervised)**: a convolutional architecture with a high-dimensional embedding trained using Bootstrap Your Own Latent for Audio (BYOL-a). We used the `features` layer, yielding 2,048 units.
- **AudioNTT-b (self-supervised)**: a more compact variant of AudioNTT-I. We used the `features` layer, yielding 512 units.
- **L<sup>3</sup>-Net**: a convolutional architecture trained with self-supervision on an audio-visual correspondence pretext task. We used the `embedding` layer, yielding 512 units.
- **AST-1**: an audio spectrogram transformer architecture for audio classification. We analysed either the `norm` layer after `blocks.11` and before the classification head, yielding 768 units, or the `blocks.11.mlp.act` GeLU activations, yielding 3,072 units.
- **AST-2**: the same architecture as AST-1, trained with a different random seed. We analysed the same two layers: `norm`, yielding 768 units, or `blocks.11.mlp.act`, yielding 3,072 units.

### Methodological framework for probing axis alignment

We propose a formal framework for assessing ‘whether’ and ‘to what extent’ a representational system (that is, a brain region or DCNN) has a privileged representational axis.

Our framework entails three main steps. First, we propose two complementary measures of axis alignment. We use each to measure the alignment between two brains or two artificial neural networks or between a brain and a network. Second, we measure how this axis alignment changes when one of the two instances undergoes axis perturbations. Third, we compare the axis alignment under native and perturbed axes to the maximum alignment achievable under any axis rotations to assess the extent to which tuning functions expressed in the native axes are privileged.

All analyses of alignment were performed using training, validation and test splits of data. When measuring alignment between two brains, two models, or a model and a brain, we used the training split to obtain the correspondence between representations, and the test

split to measure the alignment between representations with independent data. When estimating the maximum possible alignment, we performed gradient descent using a training set, then used a validation set to determine the training epoch that gave the best alignment, and then used the test split to measure the alignment value.

**Measures of axis alignment.** Existing approaches for comparing representations focus on population-level similarity measures based on linear predictivity (as assessed using  $L_2$ -regularized linear regression models) or RSA. These measures are insensitive to the axes (for instance, RSA is invariant to axis rotations) and hence cannot be employed to study axis alignment. To overcome this limitation, we propose two new measures for comparing the similarity between tuning functions in different representations.

**Pairwise matching score.** Most analyses in the main text use a direct axis-dependent measure of pairwise similarity between units in different representational systems (the only exceptions are Figs. 5 and 8, which also present the results for a ‘soft matching’ score described below). Here, a unit in a biological representation corresponds to the smallest unit of brain measurement (for example, a voxel in fMRI data, a single neuron in neurophysiology data), whereas a unit in an artificial neural network corresponds to a single ‘filter’ in the convolutional layer of a DCNN or an activation unit in the MLP layer of a transformer model.

Consider matrices  $\mathbf{X} \in \mathbb{R}^{M \times N_x}$  and  $\mathbf{Y} \in \mathbb{R}^{M \times N_y}$ , which represent the activations of  $N_x$  and  $N_y$  units, respectively, for  $M$  instances. The ‘pairwise matching’ score employs a bipartite semi-matching assignment algorithm<sup>25</sup> to find the alignment between these two representations and proceeds in two stages:

- In the first stage, we find the optimal assignment matrix  $\mathbf{A}^* \in \mathbb{R}^{N_x \times N_y}$  that maps each unit in  $\mathbf{X}$  to a single unit in  $\mathbf{Y}$  by optimizing a correlation-based objective. The optimization problem is defined as follows:

$$\begin{aligned} \{a_{ij}^*\}_{i=1 \dots N_x, j=1 \dots N_y} &= \arg \max_{a_{ij}} \sum_i \sum_j \rho(\mathbf{x}_i^{\text{train}}, \mathbf{y}_j^{\text{train}}) a_{ij}, \\ \text{s.t.} \quad a_{ij} &\in \{0, 1\} \\ \sum_j a_{ij} &= 1 \end{aligned} \quad (4)$$

Intuitively, the optimization process aims to find the unit in  $\mathbf{Y}$  that exhibits the maximum response correlation with each unit in  $\mathbf{X}$  based on a training stimulus set. The constraint ensures that each unit in  $\mathbf{X}$  is matched to exactly one unit in  $\mathbf{Y}$  as shown in Extended Data Fig. 1a. As a result, some units in  $\mathbf{Y}$  may remain unmatched after the assignment.

- In the second stage, using the optimal assignment matrix  $\mathbf{A}^*$  obtained from the first stage, we compute the correlation-based objective on held-out data. This entails calculating the correlation between each unit and its corresponding matched unit on the held-out data and then averaging these values across all matched pairs to derive the final alignment metric:

$$S_{\text{pairwise}}(\mathbf{X}^{\text{test}}, \mathbf{Y}^{\text{test}}, \mathbf{A}^*) = \sum_i \sum_j \rho(\mathbf{x}_i^{\text{test}}, \mathbf{y}_j^{\text{test}}) a_{ij}^* \quad (5)$$

Note that this measure is asymmetric by design. The choice of allowing repetitions with a bipartite semi-matching assignment (that is, allowing a single unit in  $\mathbf{Y}$  to be matched to multiple units in  $\mathbf{X}$ ) instead of a strict bipartite matching is motivated by potential variations in the number of neurons/units across different systems such as different artificial or biological networks, which may lead to varying levels of redundancy.

**Soft-matching score.** We also used a second similarity measure designed to address some limitations of the pairwise matching score, specifically, that it is asymmetric (it yields different values depending on which of two representations is used as the ‘reference’) and does not satisfy the triangle inequality. This alternative measure satisfies all the requirements of a proper metric space.

Consider again two matrices  $\mathbf{X} \in \mathbb{R}^{M \times N_x}$  and  $\mathbf{Y} \in \mathbb{R}^{M \times N_y}$ , which represent the activations of  $N_x$  and  $N_y$  units, respectively, for  $M$  instances. If  $N_x = N_y = N$ , instead of finding the optimal bipartite semi-matching assignment as in the direct pairwise matching score, we can attempt to find the optimal assignment between units without replacement, also known as a linear assignment problem<sup>98</sup> and as bipartite matching in graph theory<sup>25</sup>. This yields a ‘one-to-one’ matching distance between representations:

$$d_{\mathcal{P}}(\mathbf{X}, \mathbf{Y}) = \min_{\mathbf{P} \in \mathcal{P}(N)} \|\mathbf{X} - \mathbf{Y}\mathbf{P}\|_F \quad (6)$$

where the optimization is performed over the group of  $N$ -dimensional permutation matrices,  $\mathcal{P}(N)$ . The number of possible permutations is factorial in  $N$ , making brute force search impossible. However, an efficient algorithm can be achieved by replacing the space of permutations by the space of doubly stochastic matrices<sup>99,100</sup>: square, non-negative matrices whose rows and columns sum to one, which form the ‘Birkoff polytope’  $\mathcal{B}(N)$ . The resulting optimization problem is convex, and its solution can be shown to coincide with the optimal permutation.

Experimental datasets and neural network models typically have different numbers of neurons/units, meaning that no one-to-one matching of units is possible. Thus, we employ a generalization of the one-to-one matching as proposed in recent work<sup>26</sup>. To do this, we relax the set of permutations to ‘soft permutations’ which result from extending the Birkoff polytope to rectangular matrices. Specifically, consider a non-negative matrix  $\mathbf{P} \in \mathbb{R}^{N_x \times N_y}$  whose rows each sum to  $1/N_x$  and whose columns each sum to  $1/N_y$ . The set of all such matrices defines a ‘transportation polytope’<sup>101</sup>, denoted as  $\mathcal{T}(N_x, N_y)$ . Optimizing over this set of rectangular matrices results in a ‘soft matching’ or ‘soft permutation’ of neuron labels in the sense that every row and column of  $\mathbf{P}$  may have more than one non-zero element as shown in Extended Data Fig. 1b.

The expression for the soft-matching score is:

$$S_{\text{softmatch}}(\mathbf{X}, \mathbf{Y}) = \max_{\mathbf{P} \in \mathcal{T}(N_x, N_y)} \sum_{i,j} \mathbf{P}_{ij} \rho(\mathbf{x}_i, \mathbf{y}_j) \quad (7)$$

where  $\rho(\mathbf{x}_i, \mathbf{y}_j)$  is the Pearson correlation between neuron  $i$  in network  $\mathbf{X}$ , and neuron  $j$  in network  $\mathbf{Y}$ . The score ranges between 0 and 1, providing a normalized alternative to  $d_{\mathcal{P}}$ . This measure has the appealing theoretical properties of optimal transport distances. Intuitively, the optimal mapping matrix  $\mathbf{P}^*$  obtained by solving the optimization can be interpreted as a soft permutation or soft matching of neurons across the two representations.

Computing the above distance measures involved the resolution of a linear programme, for which we employed efficient implementations of the simplex algorithm as in the Python Optimal Transport Library<sup>102</sup>. The calculation of the soft-matching score was computationally expensive and was prohibitive for the inter-participant comparisons of visual fMRI data, which had many voxels per participant. As a result, the soft-matching score was omitted from analyses that involve inter-participant comparisons for visual fMRI data. We included it in all other analyses to show that our conclusions generalize across two measures of axis alignment.

**Quantifying alignment with a change of basis.** We drew a random rotation matrix  $\mathbf{Q}$  from the Haar distribution (the only uniform distribution on the special orthogonal group  $\text{SO}(n)$ ) using an efficient

algorithm based on QR decomposition implemented via Householder transformations<sup>103</sup>. The Haar measure ensures that each special orthogonal matrix is equally likely to be chosen.

We also considered smaller rotations (closer to the identity matrix  $I$ ) along the  $SO(n)$  manifold by computing fractional powers of the sampled rotation matrix,  $Q^\alpha$  ( $0 < \alpha < 1$ ). These fractional powers were computed by (1) triangulating  $Q$  into an upper quasi-triangular matrix  $T$  using the Schur decomposition ( $Q = WTW^*$ ), (2) computing  $T^\alpha$  by estimating the fractional powers of the individual  $2 \times 2$  blocks comprising  $T$  and (3) reconstructing the fractional power of the original rotation matrix  $Q$  as  $Q^\alpha = WT^\alpha W^*$ . For each representation pair  $(X, Y)$ , we changed the basis of one representation, say  $X$ , along this smooth  $SO(n)$  manifold to yield a transformed representation,  $X' = XQ^\alpha$ , and quantified how our axis-dependent alignment metric  $s$  (either  $s_{\text{pairwise}}$  or  $s_{\text{softmax}}$ ) changes with the amount of axis rotation ( $\alpha$ ) by computing  $s(XQ^\alpha, Y)$ . If increasing rotations (from  $I$  to  $Q$ ) consistently reduce the alignment between representations, this provides empirical evidence that the system has a privileged basis. We note that each rotated representation has exactly the same linearly decodable information and population geometry and differs only in the tuning of its coordinate axes. In all experiments studying the influence of axis rotations on representational alignment, we stochastically sampled  $Q$  10 times and computed the alignment values across all sampled rotations for 10 values of  $\alpha$  (spaced equally within the range  $[0.1, 1]$ ).

**Sampling rotation matrices as a composition of plane (2D) rotations.** In additional analyses, we investigated the effect of small axis perturbations on alignment between neural representations by sampling the special orthogonal (rotation) matrices using a different technique. Specifically, we constructed rotations in  $n$  dimensions as the product of  $n(n-1)/2$  plane rotations<sup>104–106</sup>:

$$f(\theta_{1,2}, \theta_{1,3}, \dots, \theta_{n-1,n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^n R_{i,j}(\theta_{i,j}) \quad (8)$$

where  $R_{i,j}(\theta_{i,j})$  denotes the rotation in the  $i, j$  plane embedded in the  $n$ -dimensional representation with angle  $\theta_{i,j}$ . Each 2D rotation affects only two coordinates at a time, leaving the others unchanged. For example, rotating a 3D representation can be achieved by combining individual rotations in each of the 2D planes within the 3D space. Parameterizing these single rotations with angles  $(\theta_{1,2}, \theta_{1,3}, \theta_{2,3})$ , we can, for instance, write  $R_{1,3}(\theta_{1,3})$  as

$$R_{1,3}(\theta_{1,3}) = \begin{pmatrix} \cos \theta_{1,3} & 0 & \sin \theta_{1,3} \\ 0 & 1 & 0 \\ -\sin \theta_{1,3} & 0 & \cos \theta_{1,3} \end{pmatrix} \quad (9)$$

The combined rotation matrix can then be obtained by multiplying these matrices:

$$R = R_{1,2}(\theta_{1,2})R_{1,3}(\theta_{1,3})R_{2,3}(\theta_{2,3}) \quad (10)$$

By sampling rotation matrices in this manner, we controlled  $\theta_{i,j}$  to set the amount of each 2D rotation. We set  $\theta_{i,j}$  to be a constant across all  $i, j$ , choosing from  $\{\pi/4, \pi/16, \pi/32, \pi/64\}$ . For each of these angles, we constructed the final rotation matrices by composing a varying number ( $c$ ) of 2D rotations, drawing  $c$  from 10 linearly spaced values between 0 and  $n(n-1)/2$ . Here,  $c = n(n-1)/2$  corresponds to the rotation of all possible planes in an  $n$ -dimensional representational space.

We then applied these rotations to one representation within each pair of neural representations and studied how the angle of 2D rotation ( $\theta_{i,j} \in \{\pi/4, \pi/16, \pi/32, \pi/64\}$ ) and the number of 2D rotations ( $c \in [0, n(n-1)/2]$ ) affected alignment, as measured using pairwise matching. We performed this sensitivity analyses on brain-to-brain alignment as a proof of concept, and found that higher rotation angles ( $\theta_{i,j}$ ) and

higher number of 2D rotations ( $c$ ) more strongly reduced alignment. These results are presented in Extended Data Fig. 9.

**Finding the optimal axis alignment.** The procedure described in the previous section can reveal whether the alignment between axes in different instances of the system is higher in their native basis compared with a uniformly drawn basis. But such a result would leave open the possibility that the alignment is not as high as it could be, that is, that there is a rotation matrix  $Q^*$  in the  $SO(n)$  manifold that yields a higher alignment than the standard basis. To assess whether the standard basis yields close to the maximum possible alignment, we used the Lie exponential and performed gradient-based optimization on the  $SO(n)$  matrix manifold<sup>27</sup> to find  $Q^*$  and computed this maximal alignment. The difference between the maximal basis alignment and natural basis alignment reveals the extent to which a representational system has a privileged basis.

More formally, we sought to find the optimal axis alignment between any pair of representations  $(X, Y)$ . For an alignment metric  $s$  (either  $s_{\text{pairwise}}$  or  $s_{\text{softmax}}$ ), we computed the optimal alignment  $s^*$ :

$$s^*(X, Y) = \max_{Q \in SO(n)} s(XQ, Y) \quad (11)$$

The above optimization is a constrained optimization problem. However, with a reparametrization, this constraint optimization can be turned into an unconstrained optimization, relying on the fact that every rotation matrix ( $Q$ ) can be expressed as a matrix exponential of a skew-symmetric matrix ( $S$ ):

$$\underbrace{\max_{Q \in SO(n)} s(XQ, Y)}_{\text{Constrained-optimization}} \equiv \underbrace{\max_{S \in \text{Skew}(n)} s(X \exp(S), Y)}_{\text{Unconstrained-optimization}} \quad (12)$$

We found the solution for this unconstrained optimization problem using a regular gradient descent optimizer (Adam) by making use of the Geotorch library<sup>27</sup>. We used a learning rate of 0.001 and ran the optimization algorithm for 40 epochs in all experiments. Training curves were monitored for each experiment to ensure convergence.

### Experimental setup for manipulating axis alignment in neural networks

We empirically tested the hypothesis that privileged axes improve task performance under structural constraints such as nonlinear activation functions. Neural networks were trained for ImageNet classification using a custom loss function, combining task performance optimization (cross-entropy loss  $\mathcal{L}_{\text{task}}$ ) and axis alignment with a reference model. We evaluated three conditions: (1) alignment with native axes ( $\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{alignment}}^{\text{native}}$ ), (2) alignment with rotated axes ( $\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{alignment}}^{\text{rotated}}$ ) and (3) no alignment loss ( $\mathcal{L}_{\text{task}}$ ).

Given internal representations  $Y_{\text{ref}}$  of the fixed reference model and  $Y_{\text{model}}$  of the model to be trained, and a rotation matrix  $R$ , the alignment loss for aligning to the native axes was:

$$\mathcal{L}_{\text{alignment}}^{\text{native}} = s_{\text{pairwise}}(Y_{\text{ref}}, Y_{\text{model}}) - s_{\text{pairwise}}(Y_{\text{ref}}R, Y_{\text{model}}) \quad (13)$$

where  $s_{\text{pairwise}}$  is the pairwise matching score used elsewhere in the paper, that is,

$$s_{\text{pairwise}}(Y^1, Y^2) = \max_{a_{ij}} \sum_i \sum_j \rho(Y_i^1, Y_j^2) a_{ij}, \quad (14)$$

$$\text{s.t.} \quad a_{ij} \in \{0, 1\}, \sum_j a_{ij} = 1$$

Likewise, the alignment loss for aligning to the rotated axes was:

$$\mathcal{L}_{\text{alignment}}^{\text{rotated}} = s_{\text{pairwise}}(Y_{\text{ref}}R, Y_{\text{model}}) - s_{\text{pairwise}}(Y_{\text{ref}}, Y_{\text{model}}) \quad (15)$$

For this experiment, we used AlexNet as the reference model, focusing on the last convolutional layer containing 256 channels for the alignment loss. We then trained new networks on the basis of the AlexNet architecture to test the different alignment conditions.

The networks were trained with the SGD optimizer and a learning rate of 0.005. The rotation matrix for the alignment loss was drawn uniformly from  $SO(n)$ . We set the batch size to 256 and trained the models for 120 epochs, after which no improvements in classification performance or alignment were observed. The alignment loss was scaled by a factor  $\alpha = 8$ . We trained 10 networks for each of the 3 conditions mentioned above.

### Analyses of the dominant components underlying DCNN representations

**Bayesian non-negative matrix factorization algorithm.** We identified the dominant components underlying DCNN representations using Bayesian non-negative matrix factorization (BNMF) in which the data matrix (units  $\times$  stimuli) is modelled as the product of two lower rank matrices, both of which are constrained to be non-negative. The first matrix (henceforth referred to as the response profile matrix) encodes the response profiles of each of a set of components to all stimuli, and the second matrix (henceforth referred to as the component-by-unit weight matrix) specifies the relative contribution of all units to each component.

Mathematically, the Bayesian NMF algorithm models the data matrix  $\mathbf{D}$  as

$$\mathbf{D} = \mathbf{R}\mathbf{V} + \mathbf{E} \quad (16)$$

where  $\mathbf{D}$  is the stimuli  $\times$  units data matrix for a DCNN representation,  $\mathbf{R}$  is the stimuli  $\times$  components ( $N \times C$ ) response profile matrix,  $\mathbf{V}$  is the components  $\times$  units ( $C \times V$ ) weight matrix and  $\mathbf{E}$  is a stimuli  $\times$  units ( $N \times V$ ) residual matrix. In the Bayesian approach to NMF,  $\mathbf{R}$ ,  $\mathbf{V}$  and  $\mathbf{E}$  are assumed to follow prior densities chosen for plausibility and mathematical convenience. For efficient inference, following ref. 107, we chose a zero-mean normal residual matrix  $\mathbf{E}$  with variance  $\sigma^2$ , yielding a normal data likelihood:

$$p(\mathbf{D}|\mathbf{R}, \mathbf{V}, \mathbf{E}) \sim \prod_{n=1, \dots, N; v=1, \dots, V} \mathcal{N}(\mathbf{D}_{n,v}; (\mathbf{R}\mathbf{V})_{n,v}, \sigma^2) \quad (17)$$

We further assumed that  $\mathbf{R}$  and  $\mathbf{V}$  are independently exponentially distributed with scales  $\rho_{n,c}$  and  $\gamma_{c,v}$ :

$$p(\mathbf{R}) \approx \prod_{n=1, \dots, N; c=1, \dots, C} \rho_{n,c} \exp(-\rho_{n,c} R_{n,c}) \mathbb{1}(R_{n,c} > 0) \quad (18)$$

and  $p(\mathbf{V}) \sim \prod_{c=1, \dots, C; v=1, \dots, V} \gamma_{c,v} \exp(-\gamma_{c,v} V_{c,v}) \mathbb{1}(V_{c,v} > 0)$

The conditional probabilities of  $\mathbf{R}$  (that is,  $P(\mathbf{R}|\mathbf{D}, \mathbf{V}, \sigma^2)$ ) and  $\mathbf{V}$  (that is,  $P(\mathbf{V}|\mathbf{D}, \mathbf{R}, \sigma^2)$ ) thus have a rectified Gaussian distribution. Following ref. 107, the prior for the variance in  $\mathbf{E}$  was assumed to have an inverse gamma distribution, resulting in an inverse-gamma conditional probability, that is,  $P(\sigma^2|\mathbf{D}, \mathbf{R}, \mathbf{V})$ . Parameters for  $\mathbf{R}$ ,  $\mathbf{V}$  and  $\mathbf{E}$  were optimized by sequentially drawing samples from these conditional densities using the Bayesian Markov Chain Monte Carlo (MCMC) sampling method derived in ref. 107. Due to the non-negativity constraints on both  $\mathbf{R}$  and  $\mathbf{V}$  (that is,  $R_{n,c} > 0$  and  $V_{c,v} > 0$ ), the components identified by BNMF are axis dependent. Specifically, when the data were rotated by an orthogonal matrix  $\mathbf{Q} \in SO(V)$  (that is,  $\mathbf{D}' = \mathbf{D}\mathbf{Q}$ ), many entries that were non-negative in the native coordinate system became negative. Because BNMF forbids negative values in  $\mathbf{R}$  and  $\mathbf{V}$ , the optimizer must find entirely new factors ( $\mathbf{R}', \mathbf{V}'$ ) to approximate  $\mathbf{D}'$ , making the decomposition dependent on the chosen axes. In contrast, methods such as principal components analysis (PCA) minimize squared error without sign restrictions and are therefore axis invariant: the rotated solution is simply the original eigenvectors post-multiplied by  $\mathbf{Q}$ . The axis sensitivity of BNMF has also been empirically demonstrated in another recent paper with synthetic data<sup>108</sup>.

**Bayesian NMF analysis of visual representations.** We fixed the number of components for representations in the self-supervised visual DCNN ('MoCo') to 20, the same as the optimal number of components estimated using the Bayesian information criterion for the high-dimensional responses in the human ventral visual pathway in the NSD<sup>17</sup>. For obtaining the NMF components, we used the stimulus set shown to each of the 4 NSD participants who had full 3 repetitions for each of their 10,000 images. For the 10,000 images corresponding to each NSD participant's data, we extracted the DCNN components separately. Further, since Bayesian NMF is a stochastic algorithm sensitive to initialization, we ran the algorithm  $N = 10$  times on the data matrix, each time obtaining  $C = 20$  components. We next compared these DCNN components with the 20 components obtained by applying the same factorization approach to high-level ventral visual stream responses in the 4 NSD participants. For each of these  $4 \times 10$  different iterations of NMF yielding 20 components each, we identified the optimal assignment of those components to the 5 ventral visual stream components that were most consistent across NSD participants (see details in ref. 17). These components were previously shown to be selective to faces, scenes, bodies, text and food, respectively<sup>17</sup>. Here we first computed the mean response across participants for each of these 5 neural components to the 515 images viewed by all NSD participants. We next computed the negative pairwise correlation of the response of each of these 5 neural components with the response of the 20 DCNN components and applied the linear sum assignment algorithm<sup>109</sup> on the resulting correlation matrix to match each of the DCNN components to the most similar neural component. We report the mean correlations between the resulting 5 visual DCNN components and their corresponding neural counterparts across all 40 iterations normalized by the noise ceiling of each neural component (Fig. 6b, left panel, top). The noise ceiling of each neural component was calculated by splitting the 4 NSD participants into 2 groups, computing the correlation between the mean response profiles of the component within each group, averaging the resulting correlation across all possible splits and applying the Spearman–Brown correction.

For each of the DCNN components, we also report the log-kurtosis (a measure of sparsity) to reveal the extent to which component responses are driven by a few specific units as opposed to distributed responses among all units (Fig. 6b, left panel, bottom). We also computed the correlation of each of the DCNN component responses with the salience ratings for their preferred category as obtained in an independent behavioural experiment<sup>17</sup> (reported in Fig. 6b, middle).

**Bayesian NMF analysis of auditory representations.** We determined the optimal number of components for representations in the self-supervised auditory DCNN ('BYOL-a') to be 10. This was achieved by systematically varying the component count from 5 to 25 in intervals of 5, and determining the number that minimized the Bayesian information criterion. The NMF components were derived using model activations for the 165 sounds from the auditory fMRI dataset<sup>30</sup>. We again ran the Bayesian NMF algorithm  $N = 10$  times on the data matrix, each time obtaining  $C = 10$  components. For each of these 10 different iterations, we identified the optimal assignment of the 10 components to the 2 auditory fMRI components previously shown to be selective to speech and music<sup>30,110</sup> using a linear assignment procedure analogous to that used for the visual NMF analysis. We report the mean correlations between the resulting 2 auditory DCNN components and their corresponding neural counterparts across all 10 iterations (Fig. 6c, left, top). We also computed the correlation of each of these DCNN component responses with a binary vector encoding the presence/absence of their preferred category of stimuli, that is, 'music'/'speech' (reported in Fig. 6c, middle). We conducted additional control analyses to determine whether speech and music selectivity in DCNNs could be attributed to generic acoustic features, such as spectrotemporal modulations. For this purpose, we analysed the component responses

to a different set of 36 natural sounds and their modulation-matched counterparts, as used in ref. 110. These component responses, derived from sounds different from the original 165 sounds in the study, were estimated by multiplying the DCNN activations by the pseudoinverse of the component weights.

### Analyses of category selectivity in individual units of DCNNs and single voxels of fMRI data

The Bayesian NMF decomposition analysis reveals the dominant components underlying DCNN representations but does not tell us whether category-selective tuning is more prominent in the native basis of these representations than in arbitrary bases. To assess category selectivity across different axes of visual DCNN and ventral visual stream representations, we used an image set encompassing characters, bodies, faces and places<sup>54</sup> that is commonly used to localize category-selective regions in fMRI experiments (stimuli available at <http://vpnl.stanford.edu/fLoc/>). There were ten categories in this stimulus set, grouped into five stimulus domains: characters (word and number), bodies (body and limb), faces (adult and child), places (corridor and house) and objects (car and instrument). Stimuli consisted of an object from the category superimposed on a scrambled background (different backgrounds for different stimuli).

The same 10-category image set was also presented to NSD participants in separate functional localizer (fLoc) experimental runs with six repetitions of each stimulus. We used the GLM betas from the fLoc runs as the ventral visual stream representation in each NSD participant.

For each of the DCNN (MoCo) and cortical representations, we computed  $t$ -values to quantify the selectivity of individual units (single filters in DCNNs, single voxels in fMRI fLoc data) for different image categories. For instance, selectivity for faces was quantified as a  $t$ -value contrasting adult and child faces versus all other categories. We then computed the number of units selective for each category using a  $t$ -value threshold of 10. We also assessed how the number of selective units changed under a shift of basis for each representation. As in our methodological framework for computing representational alignment, we drew a random rotation matrix  $Q$  from a uniform distribution on the special orthogonal group  $SO(n)$  to change the basis and considered smaller rotations along the  $SO(n)$  manifold by computing fractional powers of the sampled rotation matrix,  $Q^\alpha$ , where  $\alpha$  ranged across 10 equally spaced intervals within [0.1, 1]. We stochastically sampled rotation matrices 10 times and computed the mean number of selective units across all sampled rotations for each  $\alpha$ .

In the auditory domain, we used the 165 sounds used to obtain the fMRI dataset to identify category-selective units (we did not have data from any independent functional localizer runs in the auditory domain). For both the auditory cortex and auditory DCNN (BYOL-a) representations, we computed the number of units selective for music and speech separately using a  $t$ -value threshold of 10. Here,  $t$ -values were computed by contrasting the response to each category of sound (music or speech) versus all other sounds. We followed the same procedure as in the visual domain for assessing how the number of selective units changes with an altered basis.

### Analysing the impact of excluding category-selective units on privileged axes

To study whether privileged axes in biological and artificial neural networks and aligned axes between these systems simply reflect known category-selective tuning functions, we re-ran our entire analysis of privileged axes across brains, networks and the alignment between them after excluding category-selective neurons, voxels and units for each category with known functional specialization in the brain. In the visual domain, this includes categories such as faces, bodies, places, visually presented words, and food; in the auditory domain, it includes music and speech.

We applied a liberal exclusion criterion. For visual networks and the visual cortex, we removed all units with a  $t$ -statistic greater than 1 for category-present versus category-absent comparisons across any of the five visual categories. In the auditory domain, we followed the method of ref. 30, excluding the top 10% of units with the highest selectivity for music or speech based on the  $t$ -statistic for category presence. Specifically, we used the 165-sound set and labels for ‘music’ and ‘speech’ provided by refs. 30,42 to identify music- and speech-selective units. Responses to music were contrasted with responses to all other sounds, and similarly for speech.

For the visual datasets and networks, the specific criteria were as follows:

- **Natural scenes dataset:** We removed all voxels selective for either faces, places, text, bodies or food. For faces, places, text and bodies, we used responses to the fLoc stimuli provided with NSD. Since the fLoc experiment lacked a food localizer, we used the behavioural salience ratings associated with the 1,000 shared NSD images to identify food-containing images (salience  $\geq 7$ ) and non-food images (salience = 0). Brain responses to these images were compared to compute the  $t$ -statistic for food selectivity.
- **ManyMonkeys dataset:** The dataset primarily contained images from different categories such as bears, elephants, human faces, apples, cars, dogs, chairs, airplanes, birds and zebras. Face-selective units were identified by contrasting responses to human faces against all other categories. Body-selective units were identified by comparing responses to images containing animals against those containing cars, chairs or airplanes.
- **Macaque natural scenes dataset:** Category-selective units for faces, places, text, bodies or food were identified using behavioural salience ratings<sup>17</sup>. Images with salience ratings equal to or greater than 7 were categorized as containing the specific category, while images with salience of 0 were categorized as absent. To compute  $t$ -statistics, brain responses to category-present images were compared against responses to category-absent images, as defined by these salience thresholds.
- **Vision networks:** Using the same stimuli as described for NSD, we extracted network activations and applied identical criteria to identify units selective for faces, places, text, bodies or food.

### Comparison of DCNNs and transformer architectures with different metrics

We compared the representation of each DCNN/Transformer model to that of the corresponding sensory areas of the human brain using the two axis-dependent alignment measures described above (pairwise matching and soft-matching scores) as well as the arguably most widely used brain–model similarity metric (linear predictivity). For computing linear predictivity scores, we followed the standard procedure for fitting task-optimized models to brain responses based on a full linear readout. We extracted features from penultimate pretrained layers of each network and fitted  $l_2$ -regularized linear regression models separately for each network. The regularization parameter was selected by first fitting a regression model for each parameter among 8 log-spaced values in  $[1 \times 10^{-4}, 1 \times 10^4]$ , using a training split of data, and then selecting the parameter value that gave the best accuracy on a validation split. This selection was performed independently for each animal/participant and each neural network model. We quantified ‘linear predictive accuracy’ on the test stimuli as the Pearson correlation coefficient between the predicted and measured response at each voxel or neuron. Here we use the same train–validation–test split as used for the pairwise matching correlation scores and described under the respective section for each dataset.

Because the human visual fMRI data were derived from an event-related design, the voxel responses were relatively noisy, hence we noise corrected the linear predictivity estimates to give a better sense of the quality of the model predictions (Fig. 5a). We performed the correction using the noise ceiling for each voxel, which was computed using the standard procedure followed in ref. 28 by considering the variability in voxel responses across repeat scans. We also noise corrected the pairwise matching score, as this was natural given that the score is based on a correlation. Because there was no natural way to noise correct the soft-matching scores, we report the raw scores for this measure. The auditory fMRI data were derived from a block design in which each sound was presented multiple times in a row, and were substantially more reliable than the visual data, hence were not noise corrected.

Following the idea that a good model should be similar to the brain in the same sense that conspecifics are similar to each other<sup>31</sup>, we also compared the values of our axis-alignment measure (pairwise matching score) to an ‘inter-animal’ or ‘inter-human’ consistency score derived for each dataset. This consistency score was derived using an identical train–val–test split as that described above. The consistency score was averaged across all pairs of humans/animals.

### Analyses of computational implications of privileged axes

**Efficiency.** We operationalized efficiency in terms of how sparsely a unit/neuron responds to all presented natural stimuli (‘lifetime sparseness’). We quantified lifetime sparseness as the Fisher’s kurtosis ( $\kappa_4$ ) of the response distributions of each unit/neuron in response to all presented stimuli (a normal distribution of such responses across stimuli would yield  $\kappa_4 = 0$ ). This lifetime sparseness measure was calculated for individual units/voxels/neurons and then averaged across all units/voxels/neurons. Similarly, population sparseness was computed as the kurtosis of the responses of the entire population of neurons to a single image, and we report the mean population kurtosis across all images.

$$\kappa = \frac{1}{N} \sum_{i=1}^N \left( \frac{r_i - \mu}{\sigma} \right)^4 - 3 \quad (19)$$

where  $r_i$  is the response of the  $i$ th neuron,  $N$  is the total number of neurons,  $\mu$  is the mean response, and  $\sigma$  is the standard deviation of responses. We assessed how sparseness changes with axis perturbations by again drawing a random rotation matrix  $\mathbf{Q}$  from a uniform distribution on the special orthogonal group  $\text{SO}(n)$  to change the basis and considered smaller rotations along the  $\text{SO}(n)$  manifold by computing fractional powers of the sampled rotation matrix,  $\mathbf{Q}^\alpha$ , where  $\alpha$  ranged across 10 equally spaced intervals within [0.1, 1]. We stochastically sampled these rotation matrices 10 times and computed the mean sparseness values across all sampled rotations for each  $\alpha$ . We applied this same procedure to single-neuron recordings in macaque IT (as in the Macaque-NSD dataset), intracranial recordings in human auditory cortex<sup>42</sup>, and penultimate layer activations in self-supervised visual (‘MoCo’) and auditory (‘BYOL-a’) networks.

Since kurtosis could distort lifetime sparseness comparisons due to the non-negative response distributions in the native axes, we also employed another measure of lifetime sparseness previously proposed by ref. 43. Here, sparseness ( $S$ ) for a set of responses  $\{r_i\}_{i=1, \dots, n}$ , where  $r_i$  denotes the response of the neuron/unit to the  $i$ th stimulus and  $n$  denotes the number of stimuli, is quantified as

$$S = \frac{1 - (\sum_i |r_i|/n)^2 / \sum_i (r_i^2/n)}{1 - 1/n} \quad (20)$$

Results obtained using this metric are reported in Supplementary Fig. 5.

**Downstream wiring costs.** We also tested whether the axes found in the brain are those that facilitate a sparse readout of behaviourally

relevant information (for example, object category labels) from patterns of neural activity. To explore this hypothesis, we employed linear decoders trained on neural representations from both computational models and actual brain data, tasked with predicting stimulus categories of interest. In the visual domain, we used the shared 1,000 NSD images for extracting both model (here, the self-supervised visual DCNN trained using momentum contrast) and brain representations, which included human ventral visual stream responses in the NSD and macaque IT responses in the Macaque-NSD dataset. The decoder classes were the 80 object categories in the MS-COCO dataset from which the NSD stimuli were drawn. Since an image in the NSD can have more than one object category, we set this up as a multilabel classification task. We trained separate ridge classification models on each visual representation. The model was regularized with an  $l_2$  penalty, with the aim of ensuring equivalent performance across all representational axes (since changing the axes does not alter the  $l_2$ -norm). The regularization parameter was selected by first fitting a regression model for each parameter among 8 log-spaced values in  $[1 \times 10^{-4}, 1 \times 10^4]$ , using a training split of data and then selecting the parameter value that gave the best accuracy on a validation split. This selection was performed independently for each brain and model representation. We quantified the classification performance for each category label separately with the area under the receiver operating characteristic (AUC-ROC) measure and computed the mean across all categories to derive the aggregate prediction performance. We estimated the final prediction performance by averaging the results across 5 splits of the data for each representation, holding out one-third of the data for testing in each split. We computed the readout sparsity by estimating the sparseness of the readout weight distribution for each category over all units/neurons/voxels using Fisher’s kurtosis ( $\kappa_4$ ). Aggregate readout sparsity was computed by taking the mean across all categories (Fig. 4b, top right panels).

We again repeated the procedure for different axis perturbations by drawing random rotation matrices  $\mathbf{Q}$  from a uniform distribution on the special orthogonal group  $\text{SO}(n)$  to change the basis and considered smaller rotations along the  $\text{SO}(n)$  manifold by computing fractional powers of the sampled rotation matrices,  $\mathbf{Q}^\alpha$ , where  $\alpha$  ranged across 10 equally spaced intervals within [0.1, 1]. We stochastically sampled these rotation matrices 10 times and computed the mean classification performance and readout sparsity across all sampled rotations for each  $\alpha$ . We also visualized the difference between the readout sparsity in the native axes and mean sparsity over all fully rotated axes ( $\alpha = 1$ ) for each of the 80 categories as a histogram (Fig. 4b, bottom right panels).

In the auditory domain, we followed a similar procedure as in the visual domain, training ridge classification models on intracranial cortical responses and self-supervised DCNN (‘BYOL-a’) activations to natural auditory sounds. The classifier decoded the category labels of the sounds, which spanned 11 human-interpretable categories (environmental sounds, English speech, foreign speech, animal vocalizations, animal non-vocalizations, human vocalizations, human non-vocalizations, mechanical sounds, music, nature sounds and song). The regularization strength parameter was optimized independently for each representation by validating among 9 log-spaced values in  $[1 \times 10^{-4}, 1 \times 10^4]$ . We evaluated the classification performance by computing the mean AUC-ROC measure (one versus rest) using the predicted probability of each stimulus category. We estimated the classification performance by averaging results across 5 splits of the data, holding out one-third of the data for testing in each split, as in the visual domain. We computed the readout sparsity by estimating the sparseness of the readout weight distribution of each category over all units/electrodes using Fisher’s kurtosis ( $\kappa_4$ ). We followed the same approach for comparison against arbitrary uniformly drawn bases as described for the visual representations.

**Few-shot generalization.** We also tested the hypothesis that privileged axes could improve generalization for certain tasks under constrained learning scenarios. Specifically, we hypothesized that privileged axes would result in improved generalization capabilities for predicting categories whose selective responses are manifested in these axes, and that these benefits would be most evident during constrained learning scenarios such as few-shot learning or learning with enforced sparsity on the downstream linear readout. This hypothesis is based on a theoretical assertion described in the next section. To test this hypothesis empirically, we trained sparse linear decoders on top of neural representations in both native and arbitrary bases. A sparsity constraint, meant to mimic anatomical constraints on the number of inter-region connections, was enforced on the decoder through an  $l_1$  penalty. The  $l_1$  penalty also makes the generalization performance of the decoder axis-dependent. A few-shot learning scenario was instantiated by training the linear decoders using only a small fraction (10%) of labelled stimulus–category pairs. In the visual domain, separate sparse linear regressors (Lasso) were trained to predict the presence of each of five categories in an image: faces, places, bodies, text and food. The presence of each category was quantified using human salience ratings. These salience ratings were collected in an independent behavioural experiment<sup>17</sup> by asking participants how noticeable each of the five categories was in every image. For auditory representations, two distinct sparse linear classifiers (logistic regression with  $l_1$  penalty) were trained to detect the presence of music or speech in a sound waveform. The regularization parameter for the  $l_1$  penalty was selected by first fitting a regression model for each parameter among 8 log-spaced values in  $[1 \times 10^{-4}, 1 \times 10^4]$ , using a training split of data and then selecting the parameter value that gave the best accuracy on a validation split. This selection was performed independently for each task and for each model and brain representation. We considered representations from the human ventral visual stream and the self-supervised visual DCNN ('MoCo') in response to all images for which salience ratings were available (515 in total). In the auditory domain, we considered representations from the human auditory cortex as measured using intracranial recordings and the self-supervised auditory DCNN ('BYOL-a') in response to all 165 sounds. We performed comparisons against arbitrary axes by again drawing random rotation matrices  $\mathbf{Q}$  from a uniform distribution on the special orthogonal group  $\text{SO}(n)$  to change the basis. We stochastically sampled these rotation matrices 10 times and computed the mean value of the classification metrics across all sampled rotations.

For each candidate representation and basis (native versus arbitrary), we computed the generalization performance of the trained sparse linear decoders on held-out stimuli. For the visual tasks of predicting category salience, we quantified the generalization performance by computing the coefficient of determination  $R^2$  between the predicted and measured responses. For the auditory tasks of predicting category presence versus absence, we quantified generalization performance by computing the AUC-ROC using the predicted probabilities for each class. We aggregated the performance metrics across 5 splits of the data using a 10–90 train–test split corresponding to the few-shot scenario.

Why do privileged axes facilitate few-shot learning?

Here we provide a mathematical explanation for why certain axes facilitate few-shot learning when using sparsity constraints on the downstream readout. Consider the representation  $\mathbf{X} \in \mathbb{R}^{M \times N}$  as a matrix encoding the activations of  $N$  neurons for  $M$  stimuli. Let  $\mathbf{y} \in \mathbb{R}^M$  be a target variable that represents a meaningful factor of variation in our external world. We assume that it is continuous valued, such as the salience of a behaviourally relevant category in a stimulus. The readout neuron learns to encode this factor from these  $M$  instances of stimuli. Here we focus on the learning problem where the representation is fixed and only the readout weights  $\mathbf{w} \in \mathbb{R}^N$  are learned. Assuming a linear readout, the weights can be learned within a linear estimation framework:

$$\tilde{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 \quad (21)$$

In many such situations in biological nervous systems, the readout neuron may not have access to all the neurons of a representation due to wiring constraints. Thus,  $\|\mathbf{w}\|_0 = N'$  for some  $N' < N$ . During learning, this assumption can be instantiated as a sparsity constraint on the learned weights  $\|\mathbf{w}\|_0$ . Under this assumption, the learning problem can be expressed as follows:

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_0 \quad (22)$$

where  $\lambda$  is a regularization parameter that controls the strength of the sparsity penalty. The generalization error of the above estimation problem can be decomposed into bias and variance terms. The expected error can be expressed in terms of the ideal weights  $\mathbf{w}^*$  as:

$$E\left[\left(\hat{\mathbf{w}}^T \mathbf{x} - \mathbf{w}^{*T} \mathbf{x}\right)^2\right] = \underbrace{E\left[\left(\hat{\mathbf{w}}^T \mathbf{x}\right)^2\right] - E\left[\mathbf{w}^{*T} \mathbf{x}\right]^2}_{\text{Variance}} + \underbrace{\left(E\left[\mathbf{w}^T \mathbf{x}\right] - \mathbf{w}^{*T} \mathbf{x}\right)^2}_{\text{Bias}^2} \quad (23)$$

Although there is no closed-form solution for the bias and variance of the estimator subject to a sparsity constraint (equation (22)), the bias–variance trade-off suggests that reducing the size of the hypothesis class using sparsity-based regularization should reduce the variance of the estimator<sup>12</sup>. This would hold true for any representation  $\mathbf{X}$ , irrespective of the representational form it uses to encode the target variable  $\mathbf{y}$ . Consider a transformed representation  $\mathbf{X}' = \mathbf{X}\mathbf{R}$  with equivalent representational content as  $\mathbf{X}$  obtained by randomly rotating the axes of the representation  $\mathbf{X}$  through multiplication with a rotation matrix  $\mathbf{R}$ . The solution to the unconstrained optimization problem  $\mathbf{w}' = \underset{\mathbf{w} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}'\mathbf{w}'\|_2^2$  can be expressed in terms of the solution  $\tilde{\mathbf{w}}$  to equation (21) as it simply reflects a reparametrization,

$$\mathbf{w}' = \mathbf{R}^{-1} \tilde{\mathbf{w}} \quad (24)$$

Now, let us assume that the standard basis representation  $\mathbf{X}$  employs a sparse code to encode the target variable  $\mathbf{y}$ , that is,  $\|\tilde{\mathbf{w}}\|_0 < N$ . In this case, the sparsity constraint in equation (22) does not exclude the estimator  $\tilde{\mathbf{w}}$  from the hypothesis class, leaving the bias unaffected. On the other hand, multiplication with a random  $\mathbf{R}^{-1}$  will generally destroy the sparsity of  $\tilde{\mathbf{w}}$ , hence  $\mathbf{w}'$  may no longer be included in the hypothesis space of equation (22). This would cause an increase in the bias term for the rotated representation  $\mathbf{X}'$ , increasing the overall generalization error of the estimator. Thus, generally, while the sparsity constraint would reduce the variance of the estimator for both  $\mathbf{X}$  and  $\mathbf{X}'$ , it can increase the bias term for  $\mathbf{X}'$ , resulting in worse generalization in comparison to  $\mathbf{X}$ . In practice, for all our experiments, we considered a convex relaxation of the learning problem described above (equation (22)) using the L1 norm instead of L0 norm since the convex relaxation can be easily solved using coordinate descent.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

This study only uses publicly available models and datasets.

## Code availability

The code used in this study is available in OSF at <https://osf.io/xdc5y/> (ref. 113).

## References

- Barack, D. L. & Krakauer, J. W. Two views on the cognitive brain. *Nat. Rev. Neurosci.* **22**, 359–371 (2021).
- Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
- Gallego, J. A. et al. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* **9**, 4233 (2018).
- Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
- Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
- Eichenbaum, H. Barlow versus Hebb: when is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neurosci. Lett.* **680**, 88–93 (2018).
- Barlow, H. B. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1**, 371–394 (1972).
- Parker, A. J. & Newsome, W. T. Sense and the single neuron: probing the physiology of perception. *Annu. Rev. Neurosci.* **21**, 227–277 (1998).
- Endres, D. & Foldiak, P. Interpreting the neural code with formal concept analysis. *Adv. Neural Inf. Process. Syst.* **21**, 425–432 (2008).
- Barry, C. et al. The boundary vector cell model of place cell firing and spatial memory. *Rev. Neurosci.* **17**, 71–98 (2006).
- Tsao, D. Y., Freiwald, W. A., Tootell, R. B. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).
- Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
- Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
- Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
- Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).
- Polk, T. A. & Farah, M. J. The neural development and organization of letter recognition: evidence from functional neuroimaging, computational modeling, and behavioral studies. *Proc. Natl Acad. Sci. USA* **95**, 847–852 (1998).
- Khosla, M., Murty, N. A. R. & Kanwisher, N. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Curr. Biol.* **32**, 4159–4171 (2022).
- O'keefe, J. & Nadel, L. Précis of O'Keefe & Nadel's the hippocampus as a cognitive map. *Behav. Brain Sci.* **2**, 487–494 (1979).
- Lever, C., Burton, S., Jeewajee, A., O'Keefe, J. & Burgess, N. Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29**, 9771–9777 (2009).
- Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
- Morcos, A., Raghu, M. & Bengio, S. Insights on representational similarity in neural networks with canonical correlation. *Adv. Neural Inf. Process. Syst.* **31**, 5727–5736 (2018).
- Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. *PMLR* **97**, 3519–3529 (2019).
- Williams, A. H., Kunz, E., Kornblith, S. & Linderman, S. W. Generalized shape metrics on neural representations. *Adv. Neural Inf. Process. Syst.* **34**, 4738–4750 (2021).
- Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
- Lawler, E. L. *Combinatorial Optimization: Networks and Matroids* (Dover, 2001).
- Khosla, M. & Williams, A. H. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *PMLR* **243**, 326–341 (2024).
- Lezcano-Casado, M. Trivializations for gradient-based optimization on manifolds. *Adv. Neural Inf. Process. Syst.* **32**, 9157–9168 (2019).
- Allen, E. J. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
- Dapello, J. et al. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.07.01.498495> (2022).
- Norman-Haignere, S., Kanwisher, N. G. & McDermott, J. H. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88**, 1281–1296 (2015).
- Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).
- Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 109–1105 (2012).
- He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9726–9735 (IEEE, 2020).
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. Preprint at <https://doi.org/10.48550/arXiv.1706.06083> (2017).
- Gemmeke, J. F. et al. Audio set: an ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* 776–780 (IEEE, 2017).
- Hershey, S. et al. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* 131–135 (IEEE, 2017).
- Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N. & Kashino, K. BYOL for audio: self-supervised learning for general-purpose audio representation. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **31**, 137–151 (IEEE, 2021).
- Arandjelovic, R. & Zisserman, A. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision* 609–617 (IEEE, 2017).
- Szegedy, C. et al. Intriguing properties of neural networks. Preprint at <https://doi.org/10.48550/arXiv.1312.6199> (2013).
- Norman-Haignere, S. V. et al. A neural population selective for song in human auditory cortex. *Curr. Biol.* **32**, 1470–1484 (2022).
- Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
- Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).

45. Geirhos, R. et al. Generalisation in humans and deep neural networks. *Adv. Neural Inf. Process. Syst.* **31**, 7538–7550 (2018).
46. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
47. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at <https://doi.org/10.48550/arXiv.2010.11929> (2020).
48. Gong, Y., Chung, Y.-A. & Glass, J. Ast: audio spectrogram transformer. Preprint at <https://doi.org/10.48550/arXiv.2104.01778> (2021).
49. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383 (2024).
50. Tuckute, G., Feather, J., Boebinger, D. & McDermott, J. H. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLoS Biol.* **21**, e3002366 (2023).
51. Prince, J. S., Alvarez, G. A. & Konkle, T. A contrastive coding account of category selectivity in the ventral visual stream. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.08.04.551888> (2023).
52. Finzi, D., Margalit, E., Kay, K., Yamins, D. L. K. & Grill-Spector, K. A single computational objective can produce specialization of streams in visual cortex. *Nat. Commun.* **17**, 9762 (2026).
53. Kim, G., Kim, D.-K. & Jeong, H. Spontaneous emergence of rudimentary music detectors in deep neural networks. *Nat. Commun.* **15**, 148 (2024).
54. Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal processing capacity in high-level visual cortex is domain specific. *J. Neurosci.* **35**, 12412–12424 (2015).
55. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
56. Rajalingham, R., Sohn, H. & Jazayeri, M. Dynamic tracking of objects in the macaque dorsomedial frontal cortex. *Nat. Commun.* **16**, 346 (2025).
57. Haxby, J. V. et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
58. Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization. *Distill* <https://distill.pub/2017/feature-visualization> (2017).
59. Olah, C. et al. The building blocks of interpretability. *Distill* <https://distill.pub/2018/building-blocks> (2018).
60. Bau, D. et al. Understanding the role of individual units in a deep neural network. *Proc. Natl Acad. Sci. USA* **117**, 30071–30078 (2020).
61. Li, Y., Yosinski, J., Clune, J., Lipson, H. & Hopcroft, J. Convergent learning: Do different neural networks learn the same representations? Preprint at <https://doi.org/10.48550/arXiv.1511.07543> (2015).
62. Bansal, Y., Nakkiran, P. & Barak, B. Revisiting model stitching to compare neural representations. *Adv. Neural Inf. Process. Syst.* **34**, 225–236 (2021).
63. Geirhos, R. et al. On the surprising similarities between supervised and self-supervised models. Preprint at <https://doi.org/10.48550/arXiv.2010.08377> (2020).
64. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
65. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
66. Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644 (2018).
67. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
68. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
69. Schrimpf, M. et al. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
70. Jain, S. & Huth, A. Incorporating context into language encoding models for fMRI. *Adv. Neural Inf. Process. Syst.* **31**, 6628–6637 (2018).
71. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Adv. Neural Inf. Process. Syst.* **32**, 14954–14964 (2019).
72. Margalit, E. et al. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron* **112**, 2435–2451.e7 (2024).
73. Arend, L. et al. *Single Units in a Deep Neural Network Functionally Correspond with Neurons in the Brain: Preliminary Results* Technical Report (Center for Brains, Minds and Machines (CBMM), 2018).
74. Brown, D., Vyas, N. & Bansal, Y. On privileged and convergent bases in neural network representations. Preprint at <https://doi.org/10.48550/arXiv.2307.12941> (2023).
75. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 2010).
76. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
77. Willmore, B. & Tolhurst, D. J. Characterizing the sparseness of neural codes. *Network* **12**, 255–270 (2001).
78. Weliky, M., Fiser, J., Hunt, R. H. & Wagner, D. N. Coding of natural scenes in primary visual cortex. *Neuron* **37**, 703–718 (2003).
79. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14**, 481–487 (2004).
80. Lin, X.-X., Nieder, A. & Jacob, S. N. The neuronal implementation of representational geometry in primate prefrontal cortex. *Sci. Adv.* **9**, eadh8685 (2023).
81. Lachapelle, S. et al. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. *PMLR* **202**, 18171–18206 (2023).
82. Bowers, J. S., Vankov, I. I., Damian, M. F. & Davis, C. J. Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychol. Rev.* **121**, 248–261 (2014).
83. Page, M. Connectionist modelling in psychology: a localist manifesto. *Behav. Brain Sci.* **23**, 443–467 (2000).
84. Elhage, N. et al. Toy models of superposition. Preprint at <https://doi.org/10.48550/arXiv.2209.10652> (2022).
85. Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).
86. Tye, K. M. et al. Mixed selectivity: cellular computations for complexity. *Neuron* **112**, 2289–2303 (2024).
87. Klindt, D., Sanborn, S., Acosta, F., Poitevin, F. & Miolane, N. Identifying interpretable visual features in artificial and biological neural systems. Preprint at <https://doi.org/10.48550/arXiv.2310.11431> (2023).
88. Dubreuil, A., Valente, A., Beiran, M., Mastrogioseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nat. Neurosci.* **25**, 783–794 (2022).

89. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784–1792 (2014).
90. Ostojic, S. & Fusi, S. Computational role of structure in neural activity and connectivity. *Trends Cogn. Sci.* **28**, 677–690 (2024).
91. Whittington, J. C., Dorrell, W., Ganguli, S. & Behrens, T. Disentanglement with biological constraints: a theory of functional cell types. Preprint at <https://doi.org/10.48550/arXiv.2210.01768> (2022).
92. Lufkin, L., Saxe, A. M. & Grant, E. Nonlinear dynamics of localization in neural receptive fields. Preprint at <https://doi.org/10.48550/arXiv.2501.17284> (2025).
93. Huh, M., Cheung, B., Wang, T. & Isola, P. The platonic representation hypothesis. Preprint at <https://doi.org/10.48550/arXiv.2405.07987> (2024).
94. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *Computer Vision – ECCV 2014* (eds Fleet, D. et al.) 740–755 (Springer, 2014).
95. Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. A. GlmDenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7**, 247 (2013).
96. Prince, J. S. et al. Improving the accuracy of single-trial fMRI response estimates using GLMsingl. *eLife* **11**, e77599 (2022).
97. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
98. Burkard, R., Dell’Amico, M. & Martello, S. *Assignment Problems* (Society for Industrial and Applied Mathematics, 2012).
99. Ziegler, G. M. *Lectures on Polytopes* (Springer, 2012).
100. Birkhoff, G. Three observations on linear algebra. *Univ. Nac. Tucuman. Rev. A* **5**, 147–151 (1946).
101. De Loera, J. A. & Kim, E. D. Combinatorics and geometry of transportation polytopes: an update. *Discrete Geom. Algebr. Comb.* **625**, 37–76 (2013).
102. Flamary, R. et al. POT: Python optimal transport. *J. Mach. Learn. Res.* **22**, 1–8 (2021).
103. Ozols, M. A. *How to Generate a Random Unitary Matrix* (Latvijas Universitāte, 2009).
104. Pinchon, D. & Siohan, P. Angular parametrization of rectangular paraunitary matrices. *HAL* <https://hal.science/hal-01289570v1> (2016).
105. Clements, W. R., Humphreys, P. C., Metcalf, B. J., Kolthammer, W. S. & Walmsley, I. A. Optimal design for universal multiport interferometers. *Optica* **3**, 1460–1465 (2016).
106. Quessard, R., Barrett, T. & Clements, W. Learning disentangled representations and group structure of dynamical environments. *Adv. Neural Inf. Process. Syst.* **33**, 19727–19737 (2020).
107. Schmidt, M. N., Winther, O. & Hansen, L. K. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation. ICA 2009* (eds Adali, T. et al.) 540–547 (Springer, 2009).
108. Marvi, A., Kanwisher, N. & Khosla, M. Sparse components distinguish visual pathways & their alignment to neural networks. *J. Vis.* **24**, 759 (2025).
109. Crouse, D. F. On implementing 2D rectangular assignment algorithms. *IEEE Trans. Aerosp. Electron. Syst.* **52**, 1679–1696 (2016).
110. Norman-Haignere, S. V. & McDermott, J. H. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biol.* **16**, e2005127 (2018).
111. Cao, R. & Yamins, D. Explanatory models in neuroscience, part 1: taking mechanistic abstraction seriously. *Cogn. Syst. Res.* **87**, 101244 (2024).
112. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
113. Srivastava, S. Privileged representational axes in biological and artificial neural networks. OSF <https://osf.io/xdc5y/> (2026).

## Acknowledgements

This work was presented on 25 August 2023 at the Conference on Cognitive Computational Neuroscience in Oxford, UK. We thank S. Fusi, E. Fedorenko, D. Yamins and N. Kriegeskorte for helpful discussions.

## Author contributions

M.K. and N.K. conceptualized the project. M.K., N.K., S.S., J.M. and A.H.W. designed the methodology. M.K. and S.S. conducted formal analysis. M.K. performed visualization. M.K., N.K., J.M., A.H.W. and S.S. reviewed and edited the paper. N.K. administered the project.

## Funding

This work was supported by the ICoN Fellowship and Open Philanthropy, now Coefficient Giving, Grant to M.K.; the National Institutes of Health grant DP1HD091947 to N.K.; and the Center for Brains, Minds, and Machines, funded by NSF STC award CCF-1231216.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41562-026-02515-3>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41562-026-02515-3>.

**Correspondence and requests for materials** should be addressed to Meenakshi Khosla.

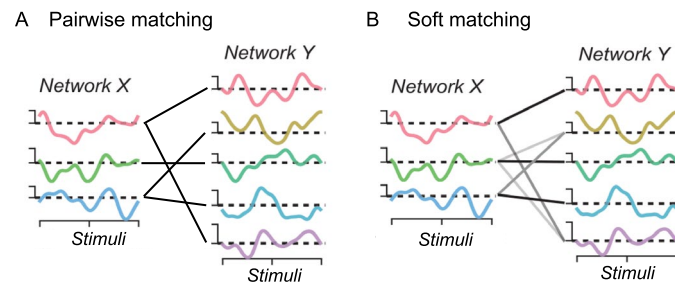
**Peer review information** *Nature Human Behaviour* thanks Hamed Nili and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

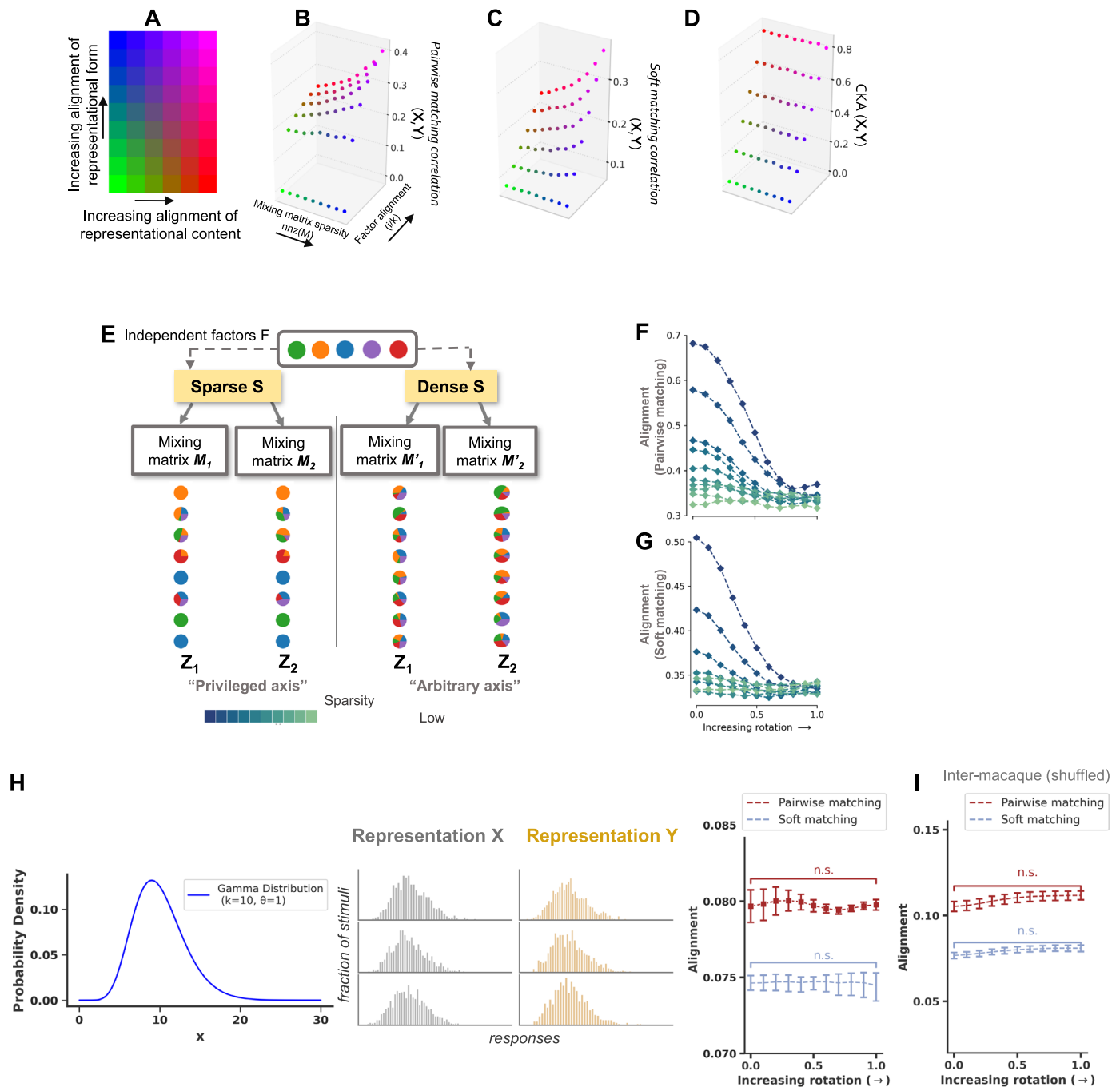
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2026



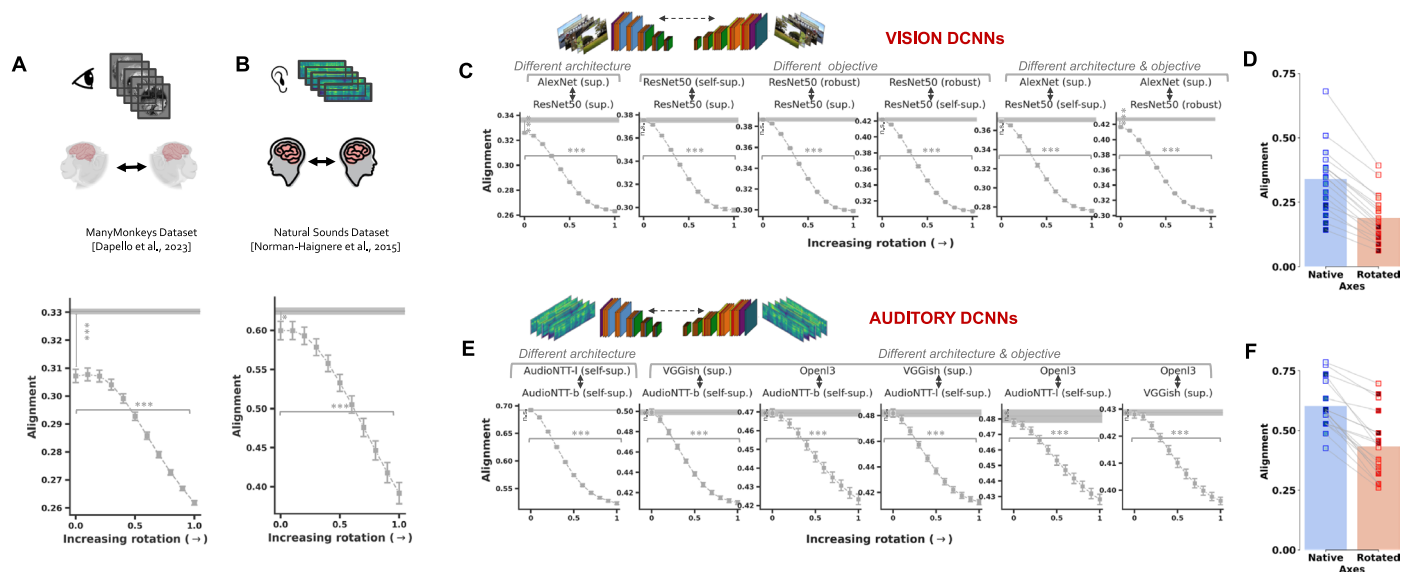
**Extended Data Fig. 1 | Schematic comparison of pairwise and soft matching alignment measures.** **A** schematic illustration of pairwise matching using example tuning curves from neurons in two representations over a 1D stimulus space. Lines show matched units across the two representations. **B** Schematic

illustration of soft matching. Grayscale lines show matched similar tuning curves across two representations. The darkness of the line indicates the strength of the match.



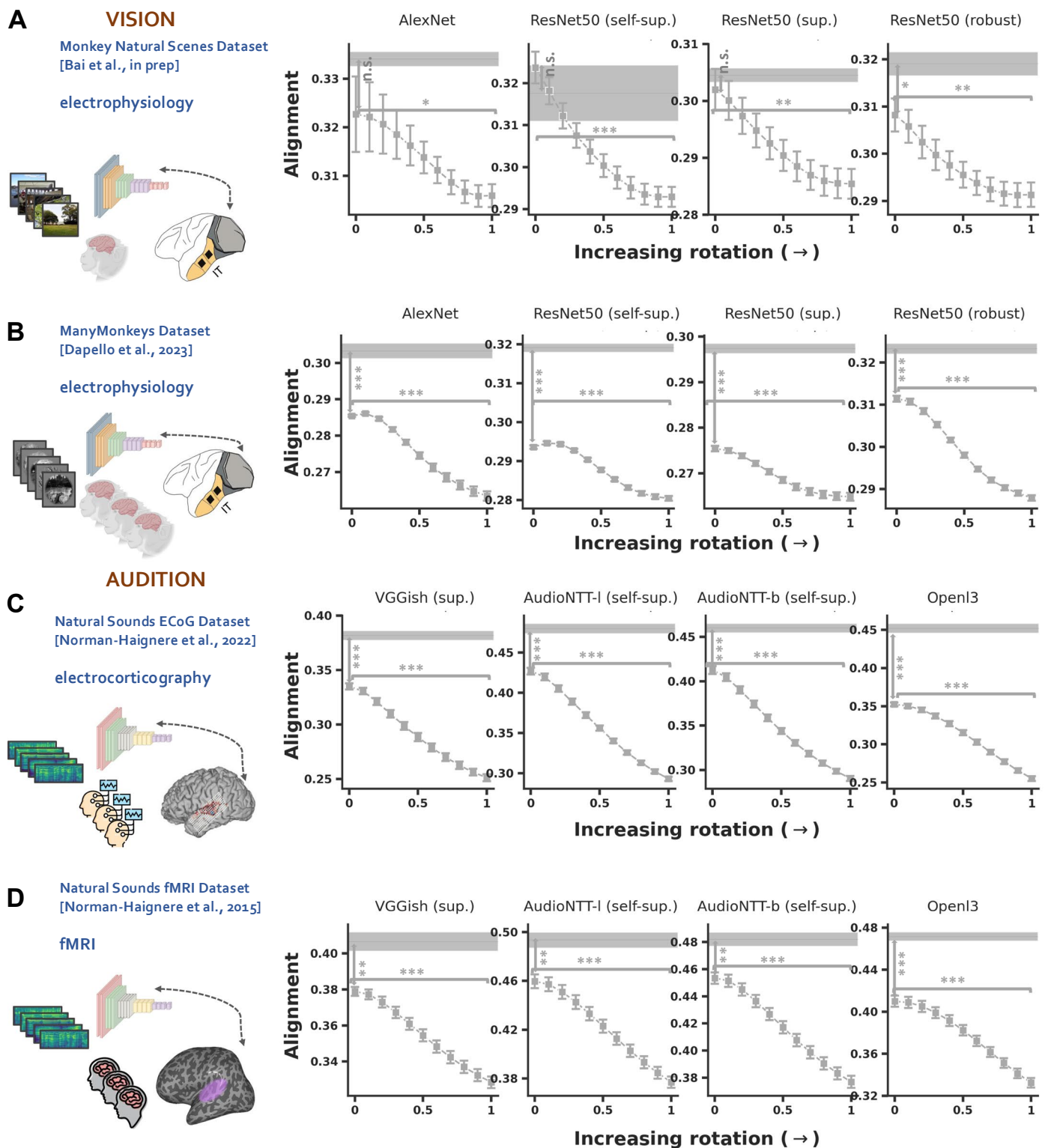
**Extended Data Fig. 2 | Simulations validating axis-dependent alignment measures.** Proposed measures effectively capture neural tuning alignment. We sample two representations,  $X$  and  $Y$ , varying the extent of alignment in their representational content and form, with the color scheme illustrated in (A). (B), (C) and (D) visualize the changes in alignment between  $X$  and  $Y$  with respect to these manipulated variables, as assessed using our pairwise matching score ( $s_{\text{pairwise}}$ ) soft matching score ( $s_{\text{softmatch}}$ ) and centered kernel alignment (CKA), respectively. CKA distinguishes representations that differ in content but is blind to differences in representational form, whereas our two new measures are sensitive to differences in both content and form. In this simulation,  $N = 4000$ ,  $k = 100$  and  $l = 1000$ . (E) Illustrated schematic featuring two instances ( $Z_1$  and  $Z_2$ ) of a representational system where each unit is a linear mixture of several independent latent factors. Colors depict the independent factors that each unit encodes. The sparsity of the mixing matrix controls the distinctiveness of the coordinate axis: high sparsity reflects the privileged axis scenario with a direct alignment of neural tuning functions, whereas low sparsity reflects the arbitrary axis scenario. (F) and (G) depict the change in alignment as a function of axis rotation ( $\alpha$ ), as measured using the pairwise matching score ( $s_{\text{pairwise}}$ )

and the soft matching correlation score ( $s_{\text{softmatch}}$ ), respectively. The dark blue curves correspond to the privileged axis scenarios, consistently displaying a gradual alignment decrease with axis rotations, in contrast to the light blue curves representing the arbitrary axis scenarios. While our illustration shows privileged axes aligning with independent data generative factors, the concept of privileged axes in our study is broader, and includes any consistent tuning function alignment across different system instances regardless of what the tuning functions encode. In this simulation,  $m = 1000$ ,  $k = 50$  and  $n = 100$ . (H) Left to right: Density function of a gamma distribution with a shape parameter of 10 and a scale parameter of 1; response distributions of sample units in simulated representations  $X$  and  $Y$  sampled from the previous gamma distribution; Alignment between  $X$  and  $Y$  plotted against axis rotation, assessed using pairwise matching and soft matching correlation scores. (I) Alignment between conspecific neural responses in macaque IT after shuffling of the stimulus order (preserving the response distributions but eliminating any shared representational axes), evaluated with pairwise matching and soft matching correlation scores, and examined as a function of axis rotation ( $\alpha$ ).



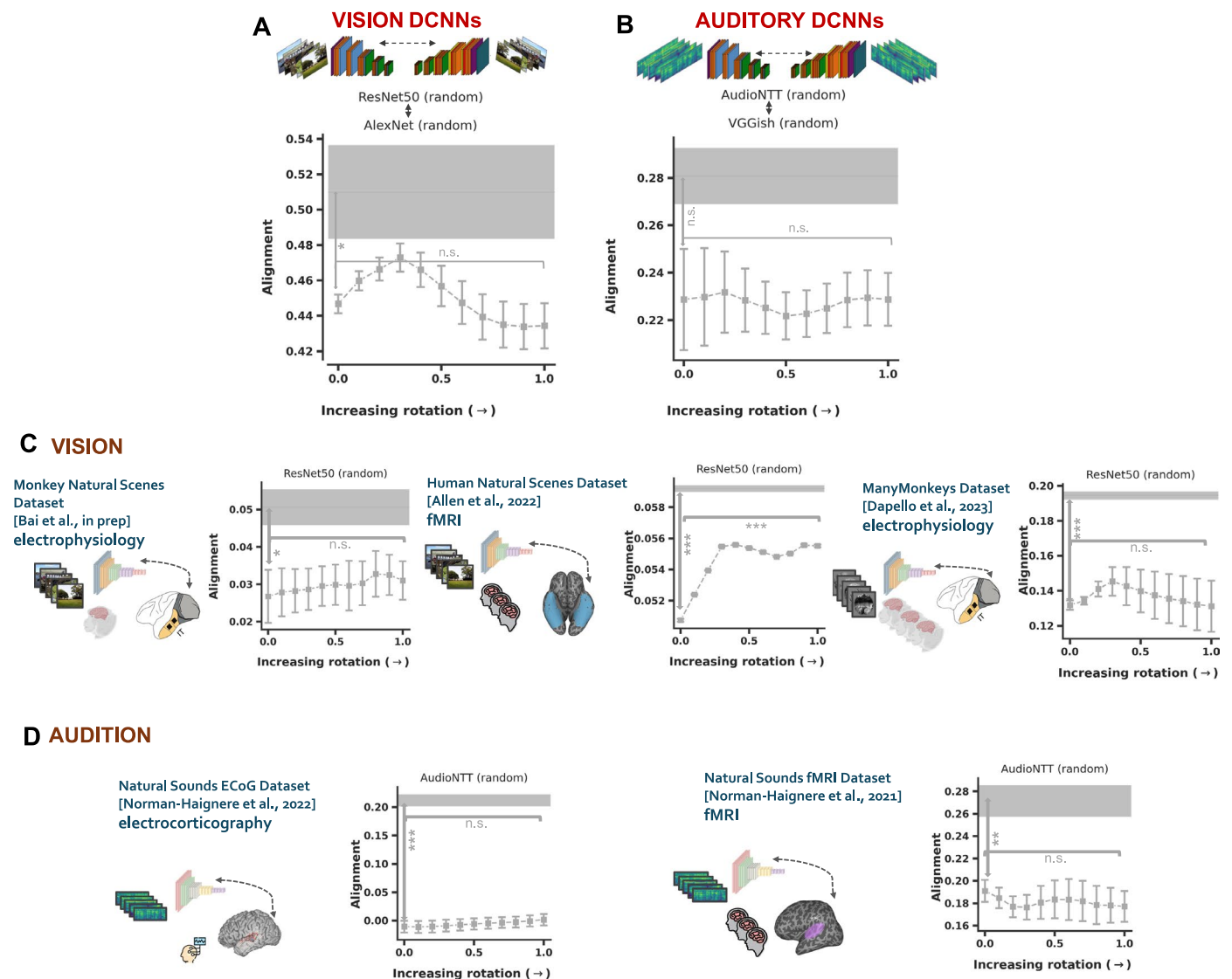
**Extended Data Fig. 3 | Soft matching reveals privileged axes in brains and DCNNs. (A–B)** Alignment between conspecific neural responses in macaque IT and human auditory cortex, measured using the soft matching score as a function of axis rotation ( $\alpha$ ). **(C)** Alignment between late-stage representations of visual DCNNs trained with different architectures and learning objectives, measured using the soft matching score. Alignment was computed from activations to the 1000-image NSD stimulus set used for the inter-subject alignment analysis. **(D)** Layerwise soft matching alignment between corresponding layers of supervised ResNet50 and self-supervised ResNet50 in the native axes and after full rotation ( $\alpha = 1$ ). Each line connects the native and rotated alignment values for a single layer; dots are color-coded by layer depth, with darker colors indicating later layers. **(E)** Same as **(C)**, but for auditory neural networks, computed from activations to the 165 natural sounds used for auditory

inter-subject alignment analysis. **(F)** Same as **(D)**, but for corresponding layers of AudioNTT-b and AudioNTT-l. For panels **(A–C, E)**,  $\alpha = 0$  indicates the native axes and  $\alpha = 1$  indicates full rotation by a matrix drawn uniformly from  $SO(n)$ . The rotation ceiling denotes the maximum alignment obtainable across all axis rotations, estimated using the Optimal Axis Alignment procedure described in the text. Data are presented as mean  $\pm$  SEM across  $n = 5$  stimulus train/test splits. Gray shaded regions indicate SEM for the rotation ceiling across splits. Significance tests were two-sided paired t-tests comparing alignment at  $\alpha = 0$  with alignment at  $\alpha = 1$ , as well as with the rotation ceiling; \*\*\* $p < 0.0001$ , n.s., not significant. In addition, soft matching shows a larger effect of rotation than pairwise matching, likely because soft matching enforces a balanced population-wide assignment, whereas pairwise matching allows one-to-many assignments that may be more robust to rotation.



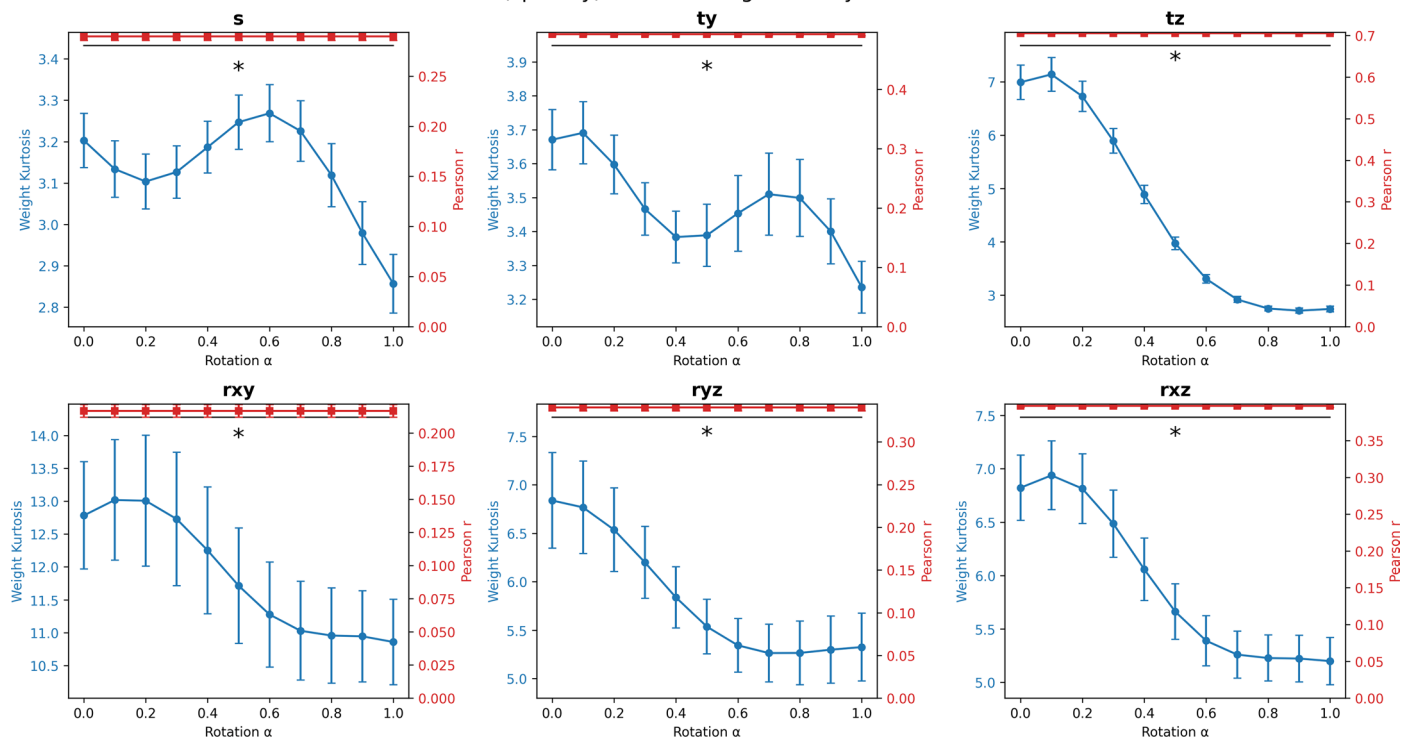
**Extended Data Fig. 4 | Model-to-brain alignment using the soft matching score.** (A–B) Alignment of various visual deep convolutional neural networks with biological neural representations in the macaque IT. This alignment is measured using the soft matching correlation score and examined as a function of axis rotation ( $\alpha$ ) 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.

(C–D) Alignment of auditory DCNNs with neural representations in the human auditory cortex, again assessed as a function of DCNN axis rotation. Error bars are SEM over 5 splits of the data. Significance tests were two-sided paired t-tests comparing alignment at  $\alpha = 0$  with alignment at  $\alpha = 1$ , as well as with the rotation ceiling; \*\*\* $p < 0.0001$ , \*\* $p < 0.001$ , \* $p < 0.01$ , n.s., not significant. The gray shaded region indicates the SEM for the optimal alignment value across all rotation angles, computed from 5 data splits.



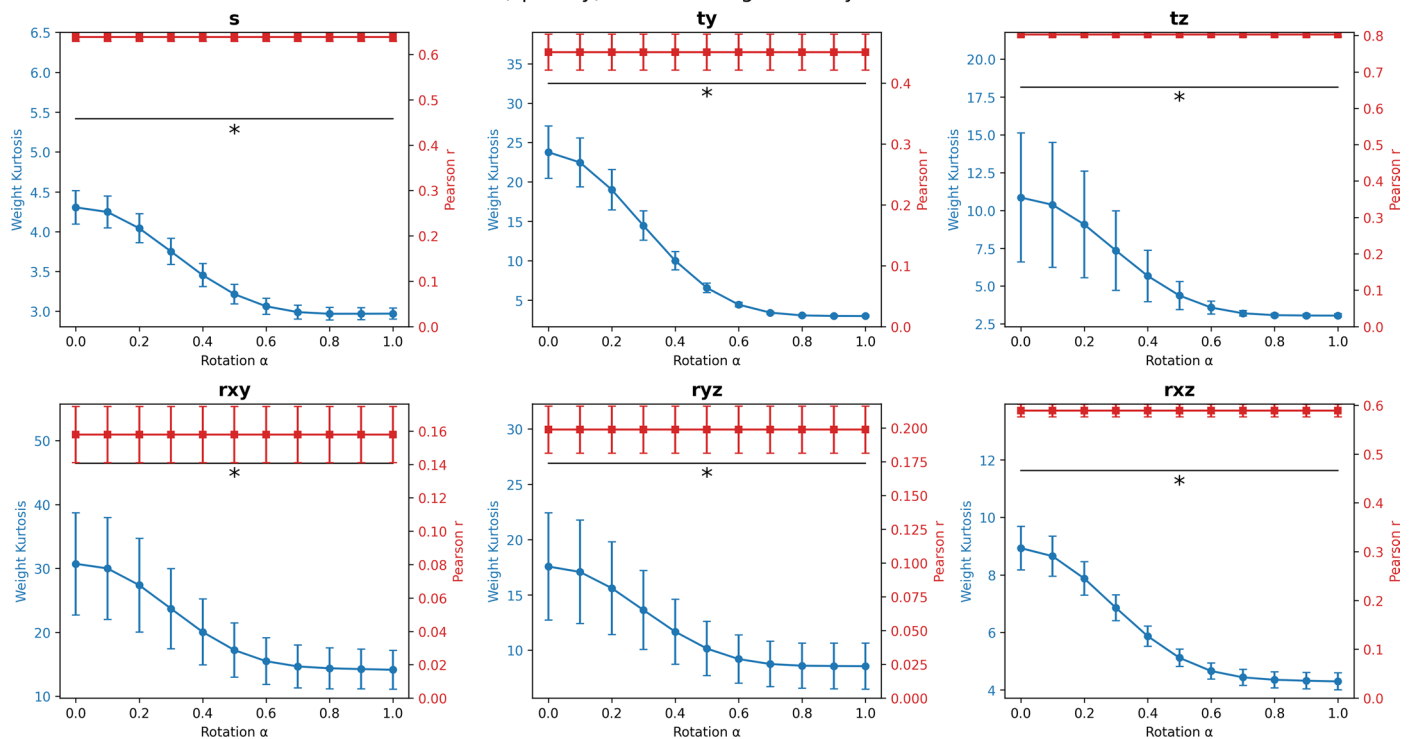
**Extended Data Fig. 5 | Untrained DCNNs do not exhibit privileged axis alignment with each other or with the brain. (A)** Pairwise matching alignment between untrained visual DCNNs with different architectures, ResNet50 and AlexNet, computed from activations to the 1000-image NSD stimulus set. **(B)** Same as **(A)**, but for untrained auditory DCNNs with different architectures, AudioNTT and VGGish, computed from activations to the 165 natural sounds used for auditory inter-participant alignment. **(C)** Pairwise matching alignment between untrained visual DCNNs and biological neural representations in macaque IT and human ventral visual cortex. **(D)** Pairwise matching alignment between untrained auditory DCNNs and neural representations in human

auditory cortex. For all panels, alignment was assessed as a function of axis rotation ( $\alpha$ ), where  $\alpha = 0$  indicates the native axes and  $\alpha = 1$  indicates full rotation by a matrix drawn uniformly from  $SO(n)$ . The rotation ceiling denotes the maximum alignment obtainable across all axis rotations, estimated using the Optimal Axis Alignment procedure described in the text. Data are presented as mean  $\pm$  SEM across  $n = 5$  stimulus train/test splits. Gray shaded regions indicate SEM for the rotation ceiling across splits. Significance tests were two-sided paired t-tests comparing alignment at  $\alpha = 0$  with alignment at  $\alpha = 1$ , as well as with the rotation ceiling; \*\*\* $p < 0.0001$ , n.s., not significant.

Kurtosis (sparsity) and Decoding Accuracy vs Rotation  $\alpha$ 

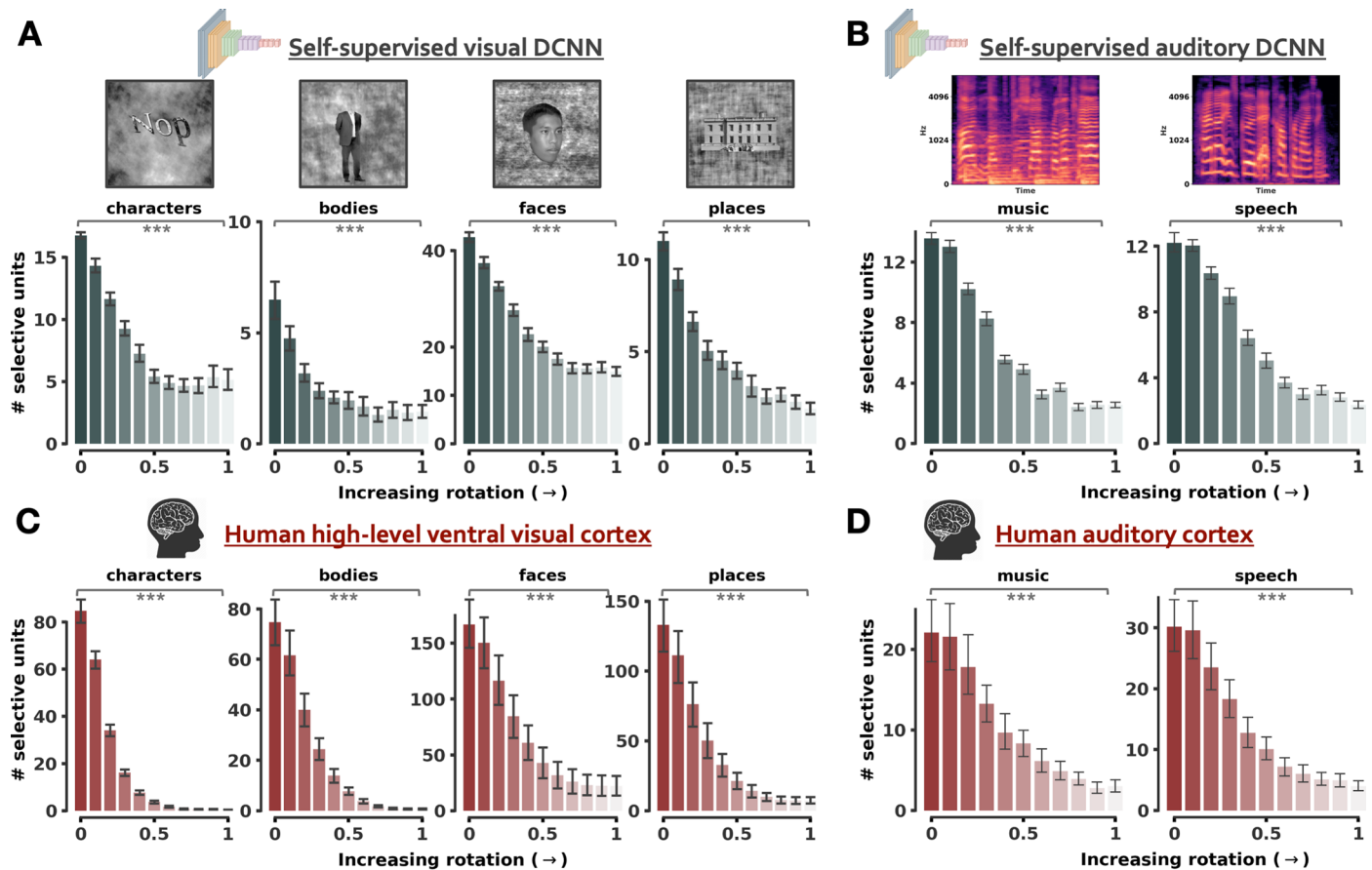
**Extended Data Fig. 6 | Axis-rotation impacts readout sparsity for category-orthogonal properties in IT.** Each panel shows results for one of six continuous object properties: scale (s), horizontal translation ( $t_x$ ), vertical translation ( $t_z$ ), in-plane rotation ( $r_{xy}$ ), elevation rotation ( $r_{xz}$ ) and azimuth rotation ( $r_{yz}$ ). On the x-axis is the axis-rotation parameter  $\alpha$  (0=no rotation, 1=full random rotation). The red curves (right ordinate) plot linear readout decoding accuracy (Pearson  $r$ ) of each property as a function of axis rotation: accuracy remains constant across

all  $\alpha$ . The blue curves (left ordinate) plot the corresponding readout sparsity, quantified by the kurtosis of the weight distribution: as  $\alpha$  increases, sparsity systematically decreases (that is weight distributions become less peaked) for most properties, indicating higher inferred wiring cost when decoding from random axes. Significance was assessed using two-sided paired t-tests comparing readout sparsity at no rotation and full rotation across  $n=5$  data splits; \* $p < 0.01$ .

Kurtosis (sparsity) and Decoding Accuracy vs Rotation  $\alpha$ 

**Extended Data Fig. 7 | Axis-rotation impacts readout sparsity for category-orthogonal properties in ANN activations.** Each panel shows results for one of six continuous object properties: scale (s), horizontal translation ( $t_y$ ), vertical translation ( $t_z$ ), in-plane rotation ( $r_{xy}$ ), elevation rotation ( $r_{yz}$ ) and azimuth rotation ( $r_{xz}$ ). On the x-axis is the axis-rotation parameter  $\alpha$  (0=no rotation, 1=full random rotation). The red curves (right ordinate) plot linear-readout decoding accuracy (Pearson  $r$ ) of each property from the penultimate layer activations of

ResNet50 (MoCo) as a function of axis rotation: accuracy remains constant across all  $\alpha$ . The blue curves (left ordinate) plot the corresponding readout sparsity, quantified by the kurtosis of the weight distribution: as  $\alpha$  increases, sparsity systematically decreases (that is weight distributions become less peaked) for all properties, indicating higher inferred wiring cost when decoding from random axes. Significance was assessed using two-sided paired t-tests comparing readout sparsity at no rotation and full rotation across  $n=5$  data splits; \* $p < 0.01$ .



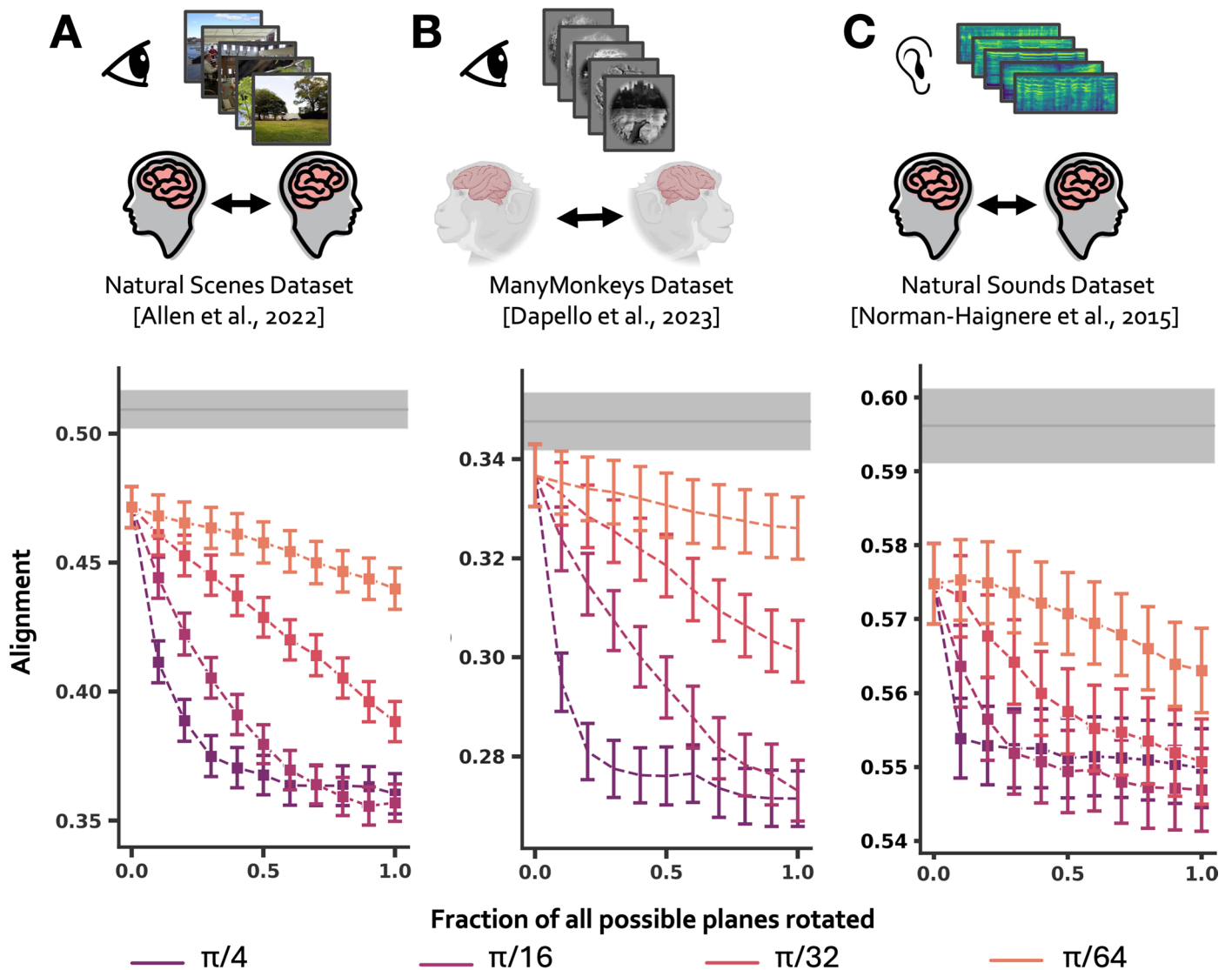
**Extended Data Fig. 8 | Category-selectivity is an axis-dependent phenomenon.**

(A) Number of single units in a self-supervised visual DCNN selective for each of four categories: characters, bodies, faces, places. Here, and in B, selectivity is assessed via t-statistics, contrasting responses to stimuli from the chosen category against others, mirroring standard functional localizer procedures.

(B) Number of single units in a self-supervised auditory DCNN selective for music

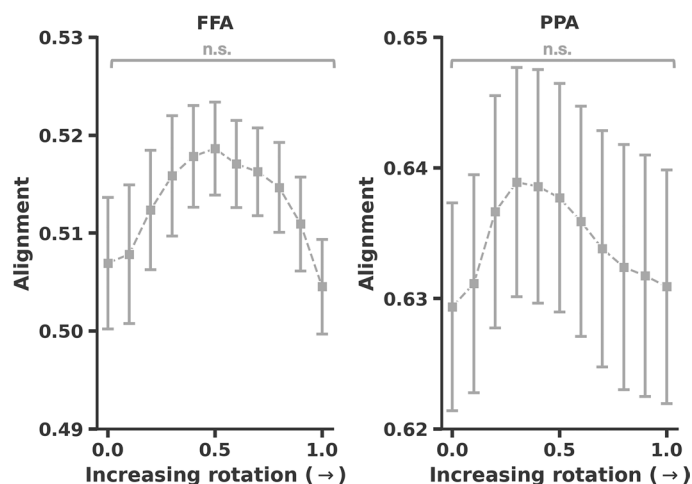
(left) and speech (right). Spectrograms of example stimuli for these categories are presented above. (C) and (D) show analogous plots for neural representations in the human high-level ventral visual cortex and auditory cortex, respectively. Here, the unit of analyses are single voxels. Statistical tests compare # selective units at no rotation and full rotation using paired t-test across 10 data splits;

\*\*\* $p < 0.0001$ .



**Extended Data Fig. 9 | Brain-to-brain alignment under smaller axis perturbations.** (A), (B) and (C) depict alignment between conspecifics neural responses in human ventral visual cortex, macaque IT and human auditory cortex, measured using the pairwise matching score. Alignment is examined as a function of the number of plane (2D) rotations (c) normalized by the total

number of planes (x-axis) and the angle of each 2D rotation (drawn from  $\{\pi/4, \pi/16, \pi/32, \pi/64\}$ ). Error bars are SEM over 5 splits of the data. The results show that higher angles of plane rotation and a greater number of plane rotations more strongly reduce alignment.



**Extended Data Fig. 10 | Alignment between conspecific neural responses in human fusiform face area (FFA, left panel) and the parahippocampal place area (PPA, right panel) as a function of axis rotation, measured using the pairwise matching score.** Alignment computed from fMRI activations to the 1000-image NSD stimulus set. Error bars are SEM over multiple splits of the data.

Significance tests compare the alignment values at no rotation ( $\alpha = 0$ ) with those at full rotation ( $\alpha = 1$ ) using paired t-test; n.s.: not significant. In each subject, the FFA and PPA were defined as the top 100 most face-selective and place-selective voxels, respectively, based on functional localizer experiments and anatomical boundaries provided with NSD.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- |                                     |                                     |                                                                                                                                                                                                                                                            |
|-------------------------------------|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement                                                                                                                                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly                                                                                                                                    |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>                                                               |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A description of all covariates tested                                                                                                                                                                                                                     |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons                                                                                                                                        |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings                                                                                                                                                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes                                                                                                                                     |
| <input type="checkbox"/>            | <input type="checkbox"/>            | Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated                                                                                                                                                         |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection We used publicly available data for this study

Data analysis Analysis code is provided here: <https://osf.io/xdc5y/>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

This study uses publicly available datasets.

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender N/A

Reporting on race, ethnicity, or other socially relevant groupings N/A

Population characteristics N/A

Recruitment N/A

Ethics oversight N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size N/A

Data exclusions For NSD, we only used data from all 4 four participants who completed full 3 repetitions of all images

Replication N/A

Randomization N/A

Blinding N/A

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a                                 | Involved in the study                                  |
|-------------------------------------|--------------------------------------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants                        |

### Methods

| n/a                                 | Involved in the study                           |
|-------------------------------------|-------------------------------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

|                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Seed stocks           | Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.                                                                                                                                                                                                                                                                                          |
| Novel plant genotypes | Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied. |
| Authentication        | Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.                                                                                                                                                                                                                                       |

---

# 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 ( $\theta$ ) 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.

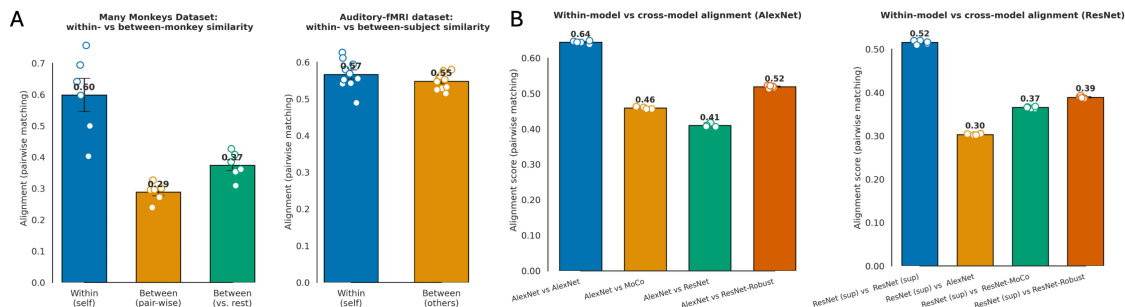
## 1.2 How ReLU Nonlinearities Disrupt Rotational Symmetry in Network Representations

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.

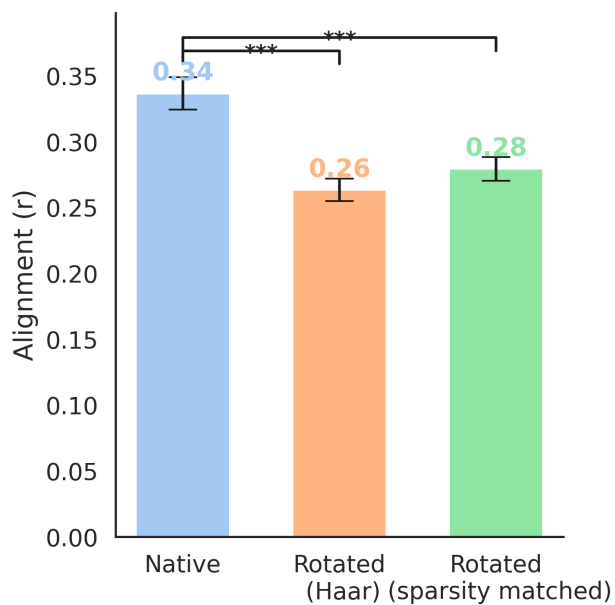
$$G_{ReLU} = \{T \in GL_n(\mathbb{R}) \mid \exists T' \in GL_n(\mathbb{R}) \text{ s.t. } ReLU \circ T = T' \circ ReLU\}$$

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

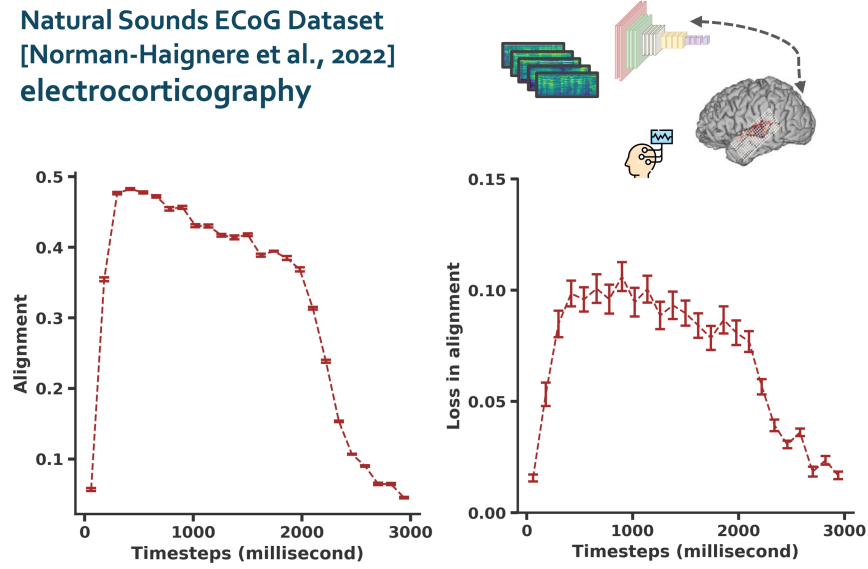


Supplementary Fig. S1: **Test-retest reliability of axis alignment within brains and within models.** (A) Within-brain alignment. Left: *ManyMonkeys* electrophysiology, estimated using split-half reliability across repeated stimulus presentations. Right: human auditory fMRI, where only three stimulus 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.

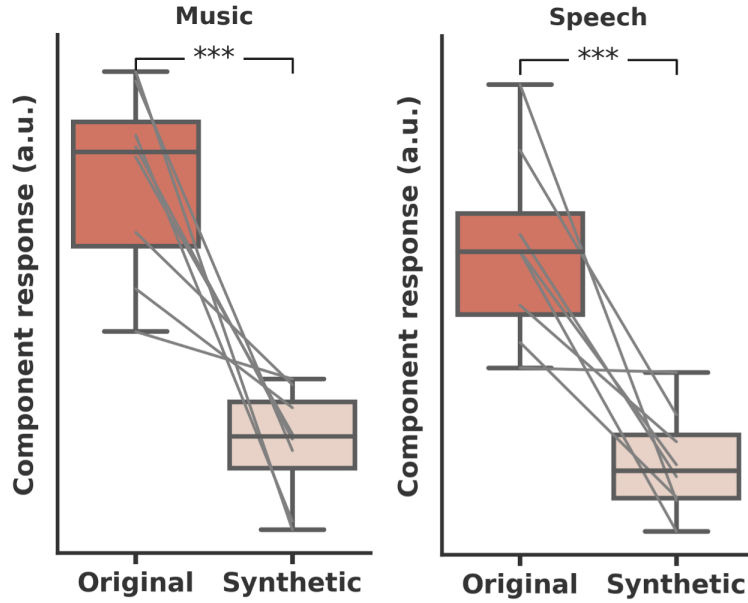


Supplementary Fig. S2: **Effect of axis rotations on pairwise alignment in single-unit recordings (*ManyMonkeys* dataset).** Alignment between pairs of monkeys is shown under three conditions: Native axes (no rotation), Haar-random rotation (axes rotated by a uniform draw from  $SO(N)$ ), Sparsity-matched 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.

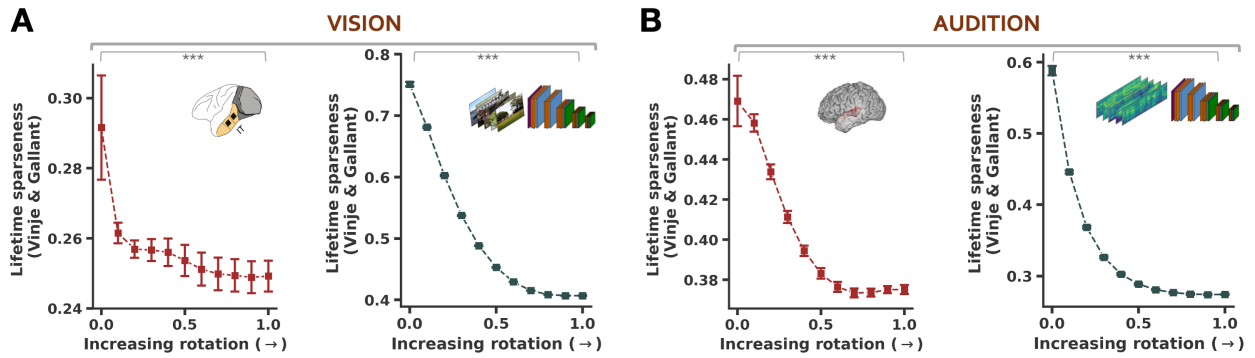
**Natural Sounds ECoG Dataset**  
**[Norman-Haignere et al., 2022]**  
**electrocorticography**



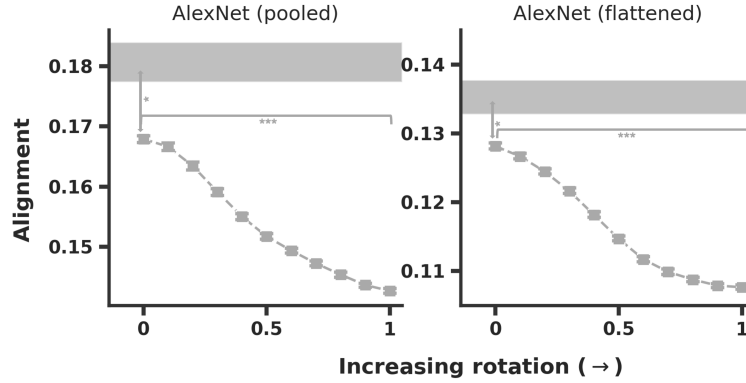
Supplementary Fig. S3: **Model-to-brain alignment as a function of time in the auditory domain** Left: Alignment of auditory deep convolutional neural network representations (self-supervised network with AudioNTT architecture) with biological neural representations in the auditory cortex (ECoG recordings) 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.



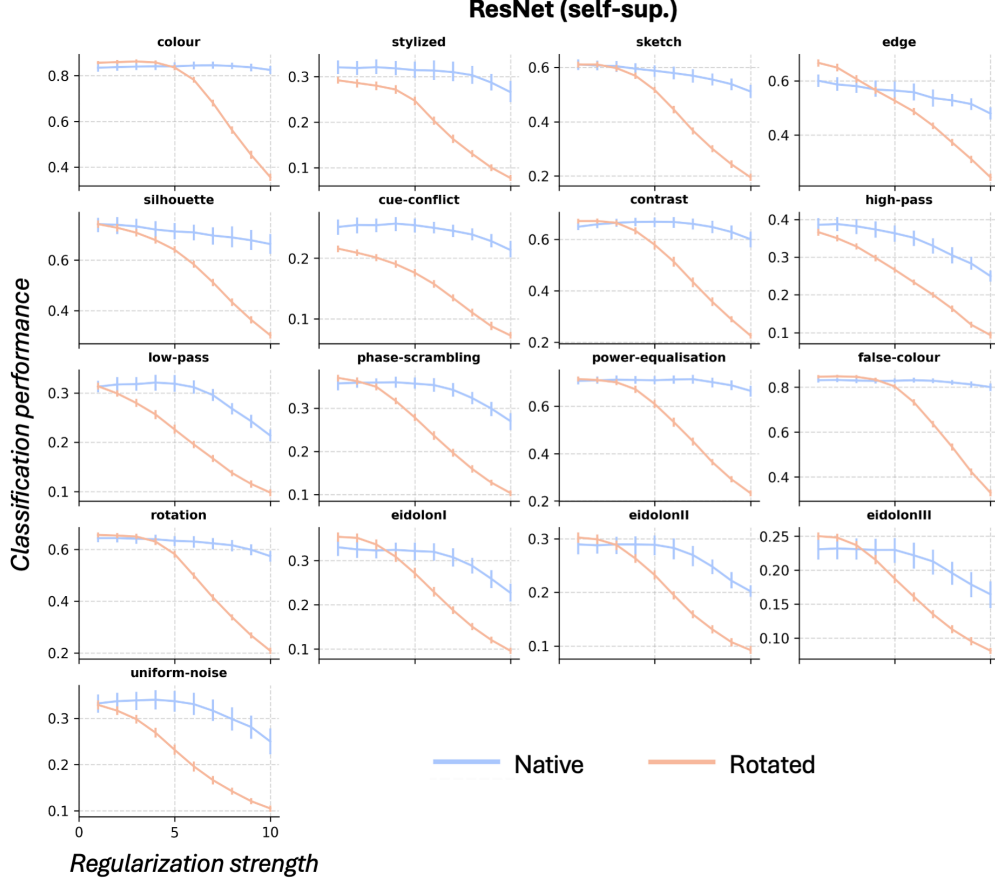
Supplementary Fig. S4: Responses of the music and speech-selective sparse components in the auditory self-supervised DCNN, as obtained by the NMF procedure, to a distinct set of 36 natural speech/music sounds and their corresponding modulation-matched controls. Signed t-test is used to evaluate whether 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].



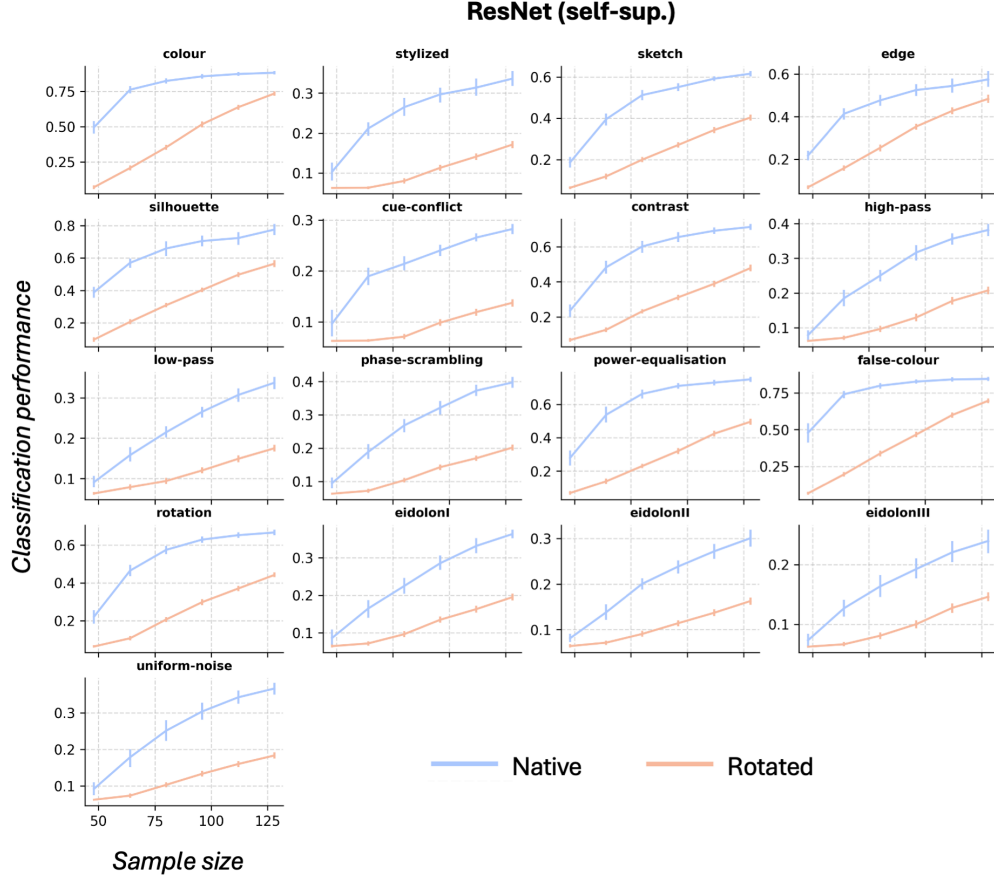
Supplementary Fig. S5: Lineplots show the mean lifetime sparseness, computed using the sparseness measure proposed in Vinje and Gallant (2000), across all units in the representation at different axis rotations. Values near 0 indicate low sparsity, and values near 1 indicate high sparsity. From left 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 ( $\alpha$ ) 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.