

Privileged representational axes in biological and artificial neural networks

Corresponding Author: Professor Meenakshi Khosla

Version 0:

Decision Letter:

2nd June 2025

Dear Professor Khosla,

Thank you once again for your manuscript, entitled "Privileged representational axes in biological and artificial neural networks", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of considerable potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

It will be important to carry out the additional analyses requested by the reviewers, in order to more convincingly demonstrate the novel advance represented by these results.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

Nature Human Behaviour

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

This paper aims to demonstrate the existence of 'privileged axes' across brains, and models. To do so, they devise ways of aligning single unit responses across brains and show that representations that are rotated away from the native axes show worse alignment across brains (and models). The authors attribute the existence of these axes to the presence of nonlinear activation functions like ReLU. They also claim that the use of these axes confers computational benefits. The work is interesting, the paper is mostly clear, and it provides a useful additional way to compare neural systems.

I have two primary conceptual questions about this work and some technical clarifications needed.

First, this work uses single cell recordings as well as fMRI voxels and ECoG electrodes. How can these all be considered the native axes? Voxels are themselves some complex transformation of single cell activity. Should voxels be considered a rotation away from the native single cell basis? If so we shouldn't expect to see the same results using voxels vs single neurons.

Second, the authors note that the only rotation consistent with rectified activity is a permutation. I then wonder about the motivation for these statements:

"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 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"

"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)."

It seems like once we know that the only ReLU-consistent rotations are permutations, then we don't need to keep asking why we don't see other rotations. I assume I'm missing something, but it would be helpful if the authors could tell me what. In general, I feel like the paper somewhat conflates the questions of "why do these systems all have the same set of axes?" with "why are these the exact axes they all have?" and perhaps that is part of what is causing me confusion here.

Smaller technical questions:

Why are absolute correlation values reported only for the layer-wise analysis (and why are the normalized values not reported here)? It would be good to know the absolute correlation values throughout in order to contextualize how similar these privileged axes are.

"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)" - this means the reference model was retrained without the ReLU?

The DNNs aren't trained with a downstream wiring cost, so how are the results from these models used to claim that these axes stem from a need to reduce wiring cost? I understand that sparse readouts work on these axes, but the DNNs show that you can get these axes without any concern for wiring costs, right?

Can the authors say more about how the BNMF is axis dependent if it is based on latent components?

How is the second sentence in contrast to the first? It seems like the first explains how and the second explains why. "Our findings show that category selectivity can emerge merely from the statistics of the natural world in conjunction with neural architecture constraints, without specific priors for any of these categories (see also [56]). This is in contrast to the conventional view of selective responses, which attribute the development of selective responses in the cortex directly to computational advantages afforded by selective responses, e.g. augmenting storage capacity in associative memories, overcoming the superposition catastrophe, energy-efficiency, facilitating post-perceptual processing etc. [86, 89, 90]."

Though not required, it might be a good idea to bring a version of Figure S1 into the main text. I think it helps to explain important concepts and verify the methodology.

Where does the name 'ManyMonkeys' come from? It doesn't seem to be in the cited papers. If this name is new to this paper

it would be good to flag that.

Reviewer #2 (Remarks to the Author):

In this article Khosla et al. aim to assess the existence and extent of privileged representational axes in primate brains and artificial neural networks. The rationale is to provide a similarity metric (for comparisons across persons/species/models) that is less flexible than RSA, CCA and CKA or encoding models in that it is sensitive to the privileged tunings (or axes in multivariate space) of neuroimaging measurements. This manuscript is noteworthy in many aspects: it puts forward a new, multifaceted methodology, it evaluates artificial neural networks (ANNs) and biological neural networks, it uses several data sets across species and sensory modalities, argues that axis alignment metrics help to differentiate between models in model-brain comparison, provide feasibility demonstrations in models that privileged axes can reduce downstream wiring costs and improve few-shot generalization.

This is an impressive manuscript, amassing an unusually large and varied amount of analyses centered around the concept of privileged axes. At the same time, I am undecided what it amounts to in the end. My review focusses on two related issues:

A) To which degree is this groundbreaking work that provides a fresh, quantitative, biologically relevant and synthetic view on the human brain, and to which degree is it a sophisticated, but very complex way of showing what we know: that there are category selective regions and retinotopy in the human visual system, or tonotopy in the auditory system, across humans, and category selective units emerge there, too (no retinotopy is considered in ANNs, as activation across locations are averaged here).

B) Given the history of psychology and neuroscience over the last decades, and its reproducibility crises, I wonder how confident we should be in any effect size reported for the human brain that bordering at 100%. One might assume that such effect sizes either indicate a bias or bug, or relate to phenomena for which scientific protocols are so well advanced and sharpened (e.g. category selectivity), that rediscovering them is trivial. Or, the data is scaled in a way that suits the purpose of visualization well.

Major

- The authors claim that the alignment in privileged axes approaches the maximal alignment achievable through any axis rotation. But what does this mean? Is it that privileged axes are an interesting biological phenomenon, or is it that the data sets we currently have capture little interesting effects going beyond category selectivity and retinotopy, either because of dataset bias, or because of limitations of the recording methods?

- Is the method to derive the maximal alignment guaranteed to find the maximal alignment? Under what assumptions? It does find a maximal alignment in the simulations, but how certain can we be that it does so in the brain data?

- The relevance of axis alignment can be judged differently, depending on whether it is related to the maximal alignment, or minimal alignment after rotation. Generally, the alignment (in %max) drops very little with increasing rotations (e.g., Fig. 2D from 96% to 91%, approximating an asymptotic state. So in effect, even full rotation leads to an alignment of 91%. It seems this tells us that privileged axes account for 5% of a 100% alignment (and maybe 23% max in vision/NSD). This is a small number, not supporting the authors' claims about the importance of privileged axes. Admittedly, the number of units that shows drop in alignment is relatively large -but so is the number of neurons that show an increase.

- Echoing the last point, random ANNs have alignment scores of 85% (Fig. S7). So what does this tell us about the lower bound of the alignment measure and its sensitivity? Theoretically the lower bound is 0. This is in conflict with the claim that 'untrained networks do not show axis alignment with neural representations (Fig. S8).

- I wonder whether the null hypothesis, i.e. drawing rotations from $SO(n)$, is a reasonable assumption. For example, is it reasonable to assume that one human instance might show positive effects, where another person would show the exact same effects, just multiplied by -1 (a perfectly fine rotation). Relatedly, what if you constrain the null to contain rotation that result in rotated matrices of similar sparseness as the original matrix?

- S4 reports the brain-to-brain alignment values for the natural sounds data set, with larger impact of rotations when using the soft matching score than the score in Fit. 2. Intuitively, I would have expected the opposite. Can you clarify?

- The authors argue that axis alignment, associated with sparsity, decreases wiring costs (Fig. 5). This is an interesting point. However, as the authors also note, it is limited to the readout of categories. In the spirit of showing that the results here are about more than category preference, what about reading out other types of information that are category orthogonal (e.g. akin to Hong et al., 2016). Demonstrating this would strengthen the claims about the relevance of preferred axes.

- I agree that ANNs trained with supervision, particularly using categories related to categories for which category selective regions exist are likely to exhibit category selective tuning (p.14). The authors thus propose using self-supervised ANNs. I instead suggest either training supervised ANNs without categories related to category-selective regions, or removing such categories (and related ones in a wide sense) from the training material for self-supervised ANNs. This might better assess alignment beyond category selectivity.

- The authors provide an alignment analysis after excluding category selective voxels/electrodes units. I suggest a more axis-related approach for stronger inference. The X top ANN and neural components found using Bayesian Non-Negative Matrix Factorization correspond to categories for which category-selective regions exist. Exclude those components and then calculate alignment between brains.

- Interestingly, the authors show that their methodology yields significant results also beyond sensory cortices. However, only the first, most basic analyses are presented here without the sophistication of the additional analyses for sensory cortices. I clearly see the value of this, but at the same time, these analyses are highly incomplete, and one wishes for a more thorough treatment. At the same, the manuscript is already very long (to the point of being too long, losing its focus), and it seems more like two or three publications to related topics that were put into one manuscript for some reason. I invite the authors to condense the paper or split the manuscript in two or three to make it better readable.

Minor

- The intuitive explanation of the soft matching score (p. 4) is not intuitive, instead it refers to previous literature.

- L. 608: Why were not both units demonstrating selectivity for music and speech removed, but only one class? Few voxels might remain then, but that would be a clear point against the idea that there are privileged axes beyond category selectivity.

- The authors provide layer-specific alignment scores but averaged for layers. What is the pattern for each layer/block? How is the alignment score related to layers?

Reviewer #3 (Remarks to the Author):

I read the manuscript "Privileged representational axes in biological and artificial neural networks" with interest. The paper investigates whether representational axes matter in biological and artificial neural networks, challenging the view that information coding is best understood solely through population-code geometry, which is blind to native neural tunings. The authors developed methods to test for alignment of neural tuning across brains and DCNNs. Furthermore, the representational axes of DCNNs trained on natural inputs were aligned with those in perceptual cortices, and this axis-alignment metric proved better at differentiating competing models of biological sensory systems. The study also demonstrates that coding schemes prioritizing certain axes can improve task performance, reduce downstream wiring costs, and enhance generalization.

Overall, I found the question interesting and novel. However, I have some comments that I list here:

1) A common practice within research on population-code geometry involves evaluating proposed metrics or novel measures based on their test/retest reliability, often assessed across independent recording sessions for the same subject or across different random initializations of the same computational model (e.g., Walther et al., Neuroimage, 2016; Shahbazi et al., Neuroimage, 2021). These important analyses of consistency appear to be absent from the current paper, and their inclusion could additionally serve to provide a more principled basis for selecting between the pairwise and soft matching metrics introduced by the authors. Furthermore, while inter-subject brain alignment might be influenced by factors such as the quality of spatial normalization, a potential caveat that is not present when conducting within-subject brain analyses, evaluating within-subject reliability across sessions would help to address this by demonstrating the stability of the findings within individuals.

2) The authors employ a theoretically elegant approach to investigate privileged axes by applying geometry-preserving rotation operations to perturb the representational space of one population relative to another. This methodology forms a compelling foundation for testing the core hypothesis. As I understand it, the argument for privileged axes is primarily supported if the alignment is consistently greatest in the native, unrotated state ($\alpha=0$), and further strengthened if this native alignment also approaches the theoretical maximum achievable through any rotation. While the former condition—that the native alignment surpasses rotated alignments—appears to hold quite consistently across the presented results, the latter—that native alignment is near the estimated ceiling—does not seem to be universally met in all cases shown. Clarifying the relative weight or importance the authors place on these two criteria in drawing their conclusions, and perhaps discussing instances where the native alignment is not close to the theoretical maximum, could enhance the clarity and precision of the argument in the text.

3) While the work is undeniably interesting, a question that lingered is precisely what novel insights are uniquely provided by this line of thinking focused on representational form, insights that could not be gleaned from the more traditional approach of examining representational content. Figure 6 offers a hint by demonstrating that while content-based measures might suggest comparable similarity to the brain across different model architectures (apart from untrained networks), focusing on representational form reveals clearer distinctions. However, I am not entirely convinced that this differential alignment based on architectural differences constitutes a significant advantage in itself, and it prompts a deeper question about the interpretation: why should a structural difference between models necessitate or imply a specific functional segregation in their representation or their relationship to brain function in the manner observed? A more explicit discussion of the behavioral or functional consequences tied to these form-based differences in model-brain alignment could strengthen the argument for the significance of this novel analytical approach.

4) A relevant consideration in the context of this work is the extensive literature on hyper-alignment and related techniques aimed at aligning representational spaces across different individuals or datasets. While the differences between the methods employed in this paper and hyper-alignment approaches can be inferred, explicitly articulating these distinctions within the text would significantly enhance clarity for readers familiar with that body of work, helping to precisely position the contribution of the novel axis-dependent measures and the framework for assessing privileged axes.

I look forward to reading the revised manuscript

Hamed Nili

Version 1:

Decision Letter:

Our ref: NATHUMBEHAV-25041535A

3rd December 2025

Dear Dr. Khosla,

Thank you for submitting your revised manuscript "Privileged representational axes in biological and artificial neural networks" (NATHUMBEHAV-25041535A).

It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

[REDACTED]

[REDACTED]

Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

I thank the authors for their responses and am happy with their changes

Reviewer #2 (Remarks to the Author):

I thank the authors for their response. I do not disagree that the authors did a huge amount of work using a novel methodology, and in that the methodology is novel, the findings are novel, too. I am still less certain about the conceptual advance in understanding the brain, but that may be because the message is not particularly straight-forward or simple. However, the authors addressed the majority of my concerns, and the manuscript amasses a large amount of interesting work.

The authors have clarified the meaning of the effect sizes, and I appreciate their rebuttal versus replicability. The introduction of the rotation floor does provide a suitable context for the effects.

I thank the authors for their clarification of point 2.4. I do not see how reporting the raw values really reduces the probability of a misread as was ascribed to me. I invite the authors to make their thinking encapsulated in point 2.4 explicit to all readers, but I do not insist.

I appreciate the extensive clarification and additional analyses for my remaining points. I do not have further comments and the manuscript is ready for publication.

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for their revisions on the manuscript. I particularly appreciate the addition of the within-instance reliability analyses and the clarified theoretical framework regarding the rotation floor versus ceiling. I think the revised version is clearer and scientifically sound.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reply to the Reviewers

Re: Manuscript ID NATHUMBEHAV-25041535

“Privileged representational axes in biological and artificial neural networks”

Meenakshi Khosla, Sudhanshu Srivastava, Alex Williams, Josh McDermott and Nancy Kanwisher
Nature Human Behavior

Dear Editor,

We thank the reviewers for their careful reading of our manuscript and their helpful comments. We have updated our manuscript to address all of the reviewers’ comments. Below, we include the full text of each review along with our response to each comment, including all major changes made in the revised manuscript. We believe our response addresses all questions raised by the reviewers.

We thank the reviewers for their thoughtful and constructive feedback, which has greatly improved the clarity and rigor of our work. We have updated our manuscript to address all of the reviewers' comments. Below, we include the full text of each review along with our response to each comment, including all major changes made in the revised manuscript. We believe our response addresses all questions raised by the reviewers.

Reviewer Comments

Reviewer 1 (Remarks to the Author):

Reviewer #1, comment #1

This paper aims to demonstrate the existence of 'privileged axes' across brains, and models. To do so, they devise ways of aligning single unit responses across brains and show that representations that are rotated away from the native axes show worse alignment across brains (and models). The authors attribute the existence of these axes to the presence of nonlinear activation functions like ReLu. They also claim that the use of these axes confers computational benefits. The work is interesting, the paper is mostly clear, and it provides a useful additional way to compare neural systems. I have two primary conceptual questions about this work and some technical clarifications needed.

Our response #1.1

Thank you for the positive feedback. Below, we address the specific comments one by one.

Reviewer #1, comment #2

First, this work uses single cell recordings as well as fMRI voxels and ECoG electrodes. How can these all be considered the native axes? Voxels are themselves some complex transformation of single cell activity. Should voxels be considered a rotation away from the native single cell basis? If so we shouldn't expect to see the same results using voxels vs single neurons.

Our response #1.2

Thank you for raising this important point. We agree that fMRI voxels pool over many neurons, but we believe this does not undermine our conclusions. Voxel pooling acts as a local averaging of the neuronal responses within each voxel, not as an arbitrary orthogonal rotation of the underlying single-unit basis. Critically, this pooling respects the topographic organization of cortex: neurons with similar tuning properties are spatially clustered, so voxels inherit the selectivity of their constituent neurons.^{7,15} This averaging can weaken alignment scores, but it does not fundamentally reorient the representational axes—meaning that the privileged axes present at the single-unit level should still be detectable, albeit attenuated. Multiple studies have shown that in many cases voxel response profiles capture the same dominant tuning dimensions present at the neuronal level, so a voxel's tuning still reflects genuine single-cell tuning rather than random mixtures.¹⁵

We have added this in our results:

"Although fMRI voxels and 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 sensory cortex,^{7,15} allowing us to treat all modalities as reporting on similar privileged axes."

Reviewer #1, comment #3

Second, the authors note that the only rotation consistent with rectified activity is a permutation. I then wonder about the motivation for these statements: "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 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" "What causes brains and DCNNs to converge onto similar axes? Learning objectives by them selves should not result in

preferred axes (because axis rotation preserves task performance for linear readouts).” It seems like once we know that the only Relu-consistent rotations are permutations, then we don’t need to keep asking why we don’t see other rotations. I assume I’m missing something, but it would be helpful if the authors could tell me what. In general, I feel like the paper somewhat conflates the questions of “why do these systems all have the same set of axes?” with “why are these the exact axes they all have?” and perhaps that is part of what is causing me confusion here.

Our response #1.3

The fact that the only rotations that preserve ReLU activations are permutations does not, by itself, explain the strong *cross-instance* axis alignment we observe between independently trained networks—or between networks and brains. We note that the theoretical argument about ReLUs only specifies which transformations can be applied to a single network’s post-ReLU activations without changing the overall function computed by the network, and does not by itself predict alignment between two independent networks. In other words: theory (in “How ReLU Nonlinearities Disrupt Rotational Symmetry”) tells us that within one network, arbitrary $SO(N)$ rotations are ruled out: the only transformations that preserve the computation exactly are discrete permutations. It does not tell us why two different networks (or a network and a brain) happen to land on the same axes or what those axes will look like (i.e. what tuning functions will they emphasize). That empirical fact—that different networks trained on the same data and brains responding to the same stimuli all share essentially the same axis orientation—is what our rotation-sweep and cross-system alignment analyses reveal. The theoretical piece is a necessary first step - it shows that, once ReLUs are in place, continuous rotations of the activation axes cannot leave the network’s computations exactly unchanged, leaving only discrete permutations as true symmetries. But the core discovery is that training + data statistics + these nonlinearities drives all systems towards a common axis choice—and small continuous perturbations away from that axis immediately degrade alignment.

The reviewer is right that we address both why any axis-alignment should be found across instances, and also why the particular axes are chosen. With respect to the second point, our empirical results show that which orientation you pick really matters for performance: Aligning to native axes boosts performance; aligning to rotated axes hurts it. When we trained AlexNet variants with an extra loss term to force their late-layer axes toward a randomly rotated version of a fixed reference model (rather than toward its native axes), their ImageNet accuracy fell by about 3 points (45 % vs. 48%). Crucially, this drop only appears when a ReLU sits after that layer—removing the nonlinearity (and restoring full rotational symmetry) completely abolishes the performance gap.

That result tells us two things: (1) the network learns representations whose axes are tuned for maximal task performance under its nonlinear constraints; and (2) those very same axes are the ones that line up across independently trained networks (and brains). In other words, axis alignment is not just a theoretical artifact of ReLU breaking continuous symmetry—it actually confers a measurable performance advantage, and networks trained on natural images—and biological visual systems—consistently converge on exactly those advantageous axes.

Reviewer #1, comment #4

Smaller technical questions 1: Why are absolute correlation values reported only for the layer-wise analysis (and why are the normalized values not reported here)? It would be good to know the absolute correlation values throughout in order to contextualize how similar these privileged axes are.

Our response #1.4

We originally reported normalized alignment values in the axis-rotation plots because these emphasize alignment relative to the maximum attainable across all rotations, making it easier to compare across datasets with different overall scales. The corresponding raw alignment values were shown in Figure 6, alongside linear predictivity scores, which we felt was the most natural place to contextualize absolute values across datasets. However, we agree that providing absolute (unnormalized) alignment values for the rotation analyses adds important context. Following the reviewer’s suggestion (and a similar comment from Reviewer 2), we have revised all brain–brain, model–model, and model–brain alignment figures (Figs. 2–4 in the original manuscript) to report unnormalized alignment values, together with the corresponding maximum alignment attainable across all axis rotations (see Figs. A2,A3,A4 in response #2.5 below, showing unnormalized pairwise-matching values for each comparison and rotation level).

We have also updated Fig. 8 to present unnormalized alignment values for experiments excluding all category-selective units and voxels (see Fig. A5 below), and Fig. 9 in the revised manuscript to report unnormalized alignment between conspecific neural responses in macaque dorsomedial frontal cortex (DMFC) during a mental simulation task and human language cortex responses during sentence comprehension. Finally, all supplementary figures have been revised accordingly to show raw, unnormalized alignment values rather than normalized ones.

Reviewer #1, comment #5

Smaller technical questions 2: "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)" — this means the reference model was retrained without the relu?

Our response #1.5

Yes, we trained new reference models from scratch with the ReLU after *conv5* removed, and with no alignment loss term. We did this to keep everything other than the alignment loss consistent across the three conditions. We have clarified this in the manuscript. It now reads:

"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."

Reviewer #1, comment #6

Smaller technical questions 3: The DNNs aren't trained with a downstream wiring cost, so how are the results from these models used to claim that these axes stem from a need to reduce wiring cost? I understand that sparse readouts work on these axes, but the DNNs show that you can get these axes without any concern for wiring costs, right?

Our response #1.6

You're correct that our DNNs were never optimized with a wiring-cost objective. Rather than arguing that wiring constraints *cause* privileged axes to emerge, our sparseness/readout analysis shows a key *consequence* of those axes: they support much sparser linear decoders. In a biological system, that sparsity would translate directly into fewer axonal connections—and therefore lower wiring cost. In other words:

- Cause vs. consequence: Privileged axes in our models arise through standard factors (data statistics, nonlinearities, optimization dynamics), not through any built-in wiring penalty.
- Functional implication: Once those axes exist, downstream readouts become far sparser than they would along arbitrary directions. If a brain needed to economize on connections, aligning its representations to these axes would be an effective strategy.

We have clarified this in the manuscript when describing these results. It reads:

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

Reviewer #1, comment #7

Smaller technical questions 4: Can the authors say more about how the BNMF is axis dependent if it is based on latent components?

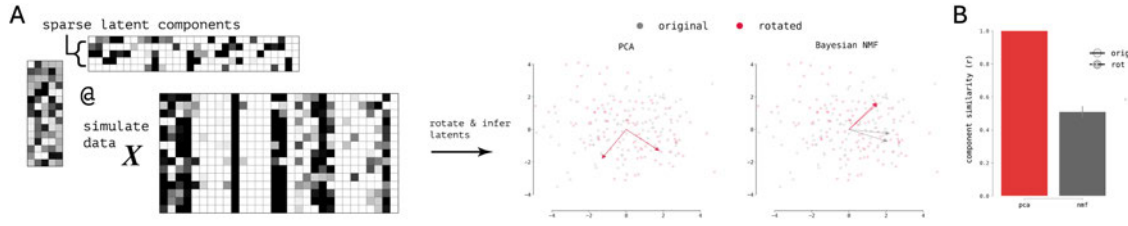


Figure A1: **Simulations of latent component recovery and rotation sensitivity** Different methods used to recover the latent components of simulated data **D**. BNMF but not PCA components are dissimilar after correcting for rotation, measured via Pearson’s r . (note: overlapping PCA components). Figure reproduced from Marvi et al. (ICLR 2025).

Our response #1.7

Bayesian NMF factorises the data matrix

$$\mathbf{D} \in \mathbb{R}^{N \times V} \approx \underbrace{\mathbf{R}}_{N \times C} \underbrace{\mathbf{V}}_{C \times V}, \quad \mathbf{R} \geq 0, \mathbf{V} \geq 0,$$

subject to **non-negativity** (and usually sparsity) priors on both factors. Those sign constraints break rotational symmetry. Let $Q \in \text{SO}(n)$ be any orthogonal rotation. After the change of basis $\mathbf{D}' = \mathbf{D}Q$, many entries that were non-negative in the native axes become negative. Because BNMF forbids negative loadings, the optimizer must find new factors $(\mathbf{R}', \mathbf{V}')$ to approximate $\mathbf{D}'$; the decomposition therefore depends on the chosen axes.

Principal-component methods, by contrast, minimise squared error with no sign restrictions, so the rotated solution is simply the original eigenvectors post-multiplied by Q ; PCA is axis-invariant.

In a companion study (Marvi et al., ICLR 2025),¹¹ we empirically demonstrated the axis-sensitivity of BNMF. Here, we generated simulated data as a mixture of sparse latent components and mixing coefficients. The data matrix $\mathbf{D} \in \mathbb{R}^{N,V}$ is constructed as the product of two sub-matrices: $\mathbf{R} \in \mathbb{R}^{N,C}$, which simulates the response of C components to N stimulus conditions, and $\mathbf{V} \in \mathbb{R}^{C,V}$, containing the mixing coefficients. The resulting equation, with added noise ϵ , is defined as: $\mathbf{D} = \mathbf{R}\mathbf{V} + \epsilon$. We then attempted to recover the latent components from $\mathbf{D}$ using PCA and Bayesian NMF after rotating the data. The results confirmed that PCA but not NMF components remain unchanged after rotation. These are shown in A1.

We have modified the Methods section with this clarification. It now reads:

“Due to the non-negativity constraints on both $\mathbf{R}$ and $\mathbf{V}$ (i.e., $R_{n,c} > 0$ and $V_{c,v} > 0$), the components identified by BNMF are axis-dependent. Specifically, when the data is rotated by an orthogonal matrix $Q \in \text{SO}(V)$ (i.e., $\mathbf{D}' = \mathbf{D}Q$), many entries that were non-negative in the native coordinate system become 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 like 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 Q . The axis-sensitivity of BNMF has also been empirically demonstrated in another recent paper with synthetic data.¹¹”

Reviewer #1, comment #8

Smaller technical questions 5: How is the second sentence in contrast to the first? It seems like the first explains how and the second explains why. “Our findings show that category selectivity can emerge merely from the statistics of the natural world in conjunction with neural architecture constraints, without specific priors for any of these categories (see also [56]). This is in contrast to the conventional view of selective responses, which attribute the development of selective responses in the cortex directly to computational advantages afforded by selective responses, e.g. augmenting storage capacity in associative memories, overcoming the superposition catastrophe, energy-efficiency, facilitating post-perceptual processing etc. [86, 89, 90].”

Our response #1.8

Our aim was to show that category selectivity can emerge even in the absence of narrowly tailored objectives—whether augmenting associative memory capacity, avoiding the superposition catastrophe, maximizing energy efficiency, or facilitating downstream processing. In fact, the only pressures required are quite broad: the statistical structure of natural inputs interacting with generic architectural constraints is itself sufficient to drive the spontaneous emergence of category-selective tuning. We have clarified this in the Discussion where we instead now write:

“These findings 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 also¹⁴) and without invoking any narrowly tailored objectives for associative-memory capacity, superposition avoidance, energy efficiency, or other downstream post-perceptual processing.^{3, 12, 13}”

Reviewer #1, comment #9

Smaller technical questions 6: Though not required, it might be a good idea to bring a version of Figure S1 into the main text. I think it helps to explain important concepts and verify the methodology.

Our response #1.9

We appreciate the suggestion. Given multiple reviewer requests to keep the already lengthy manuscript focused and readable, we opted to keep the validation simulations (Fig.S1) in the Supplement so the main text can focus on the empirical model- and brain-data results. We explicitly direct readers to Fig S1 for methodological verification.

Reviewer #1, comment #10

Smaller technical questions 7: Where does the name ‘ManyMonkeys’ come from? It doesn’t seem to be in the cited papers. If this name is new to this paper it would be good to flag that.

Our response #1.10

Thank you for catching that. The name “ManyMonkeys” was our own shorthand—borrowed from the informal label used on the dataset’s GitHub release (see <https://github.com/dapello/braintree/tree/master>). We have made that explicit by adding a brief note:

“To answer this question for the visual system, we leveraged the Natural Scenes Dataset (NSD, human fMRI)¹ and the macaque IT electrophysiology data of Dapello et al. (2022)⁵ (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 changes with axis rotation respectively.”

Reviewer 2 (Remarks to the Author):

Reviewer #2, comment #1

In this article Khosla et al. aim to assess the existence and extent of privileged representational axes in primate brains and artificial neural networks. The rationale is to provide a similarity metric (for comparisons across persons/species/models) that is less flexible than RSA, CCA and CKA or encoding models in that it is sensitive to the privileged tunings (or axes in multivariate space) of neuroimaging measurements. This manuscript is noteworthy in many aspects: it puts forward a new, multifaceted methodology, it evaluates artificial neural networks (ANNs) and biological neural networks, it uses several data sets across species and sensory modalities, argues that axis alignment metrics help to differentiate between models in model-brain comparison, provide feasibility demonstrations in models that privileged axes can reduce downstream wiring costs and improve few-shot generalization.

This is an impressive manuscript, amassing an unusually large and varied amount of analyses centered around the concept of privileged axes. At the same time, I am undecided what it amounts to in the end. My review focusses on two related issues: To which degree is this groundbreaking work that provides a fresh, quantitative, biologically relevant and synthetic view on the human brain, and to which degree is it a sophisticated, but very complex way of showing what we know: that there are category selective regions and retinotopy in the human visual system, or tonotopy in the auditory system, across humans, and category selective units emerge there, too (no retinotopy is considered in ANNs, as activation across locations are averaged here).

Our response #2.1

We thank the reviewer for their generous assessment of the breadth of analyses in our paper. The central idea of privileged axes is distinct from classic interpretable tunings such as category selectivity, retinotopy, or tonotopy. Privileged axes refer to systematic axis orientations that recur across independent instances of a system (brains, networks, or both). This is a different question from whether a single system contains interpretable features: it asks whether the same coordinate axes reappear across systems, and whether those axes confer computational advantages.

As we already show, our results go well beyond re-confirming that some voxels are category-selective or that early cortices are retino-/tonotopic. Our analyses in Fig. 8 (Section 7) indicate that the privileged axes identified in this study go beyond established category-specific neural responses (e.g., selectivity for faces, places, bodies, words, and food). These analyses further support our finding that the privileged axes are not limited to category-specific responses, suggesting instead a broader principle that transcends specific category selectivities.

We also want to emphasize that the evidence for privileged axes in biological networks was one of many results in our paper. First, we find axis alignment not only among humans and among monkeys but also among artificial neural networks—persisting throughout all layers, not just the early (Gabor-like) or late (category-selective) stages. Second, we show that networks and primates share these privileged axes (Figs. 3 and 4 in our manuscript) and we provide both theoretical and empirical reasoning behind why networks have privileged axes (Section 2 in Results; Fig. 3E). Third, we report several implications of privileged axes that extend across both vision and audition, including their ability to support improved generalization under sparsity constraints on the readout (Fig. 5). Fourth, we demonstrate the ability of axis-sensitive measures to distinguish competing models of biological systems (Fig. 6). Each of these findings is new and unexpected. Moreover, the existence of privileged axes in domains beyond vision, such as in audition and language, is also not well-acknowledged. For instance, a recent study has demonstrated the prevalence of music-selectivity in auditory networks¹⁰ but beyond this study, we are not familiar with any other study systematically analyzing category selectivity or privileged axes in auditory networks or their alignment with the auditory cortices.

Finally, we emphasize that the presence of category selectivity alone does not imply that the representational axes of these systems are privileged. It is the comparison against arbitrary (uniformly drawn) axes

that reveals their privileged nature. Arbitrary axes in a high-dimensional representational space could, in principle, also exhibit category-selective responses (and we show that some category-selectivity still persists in arbitrary axes). What makes an axis “privileged” is that native alignment outperforms a uniformly random basis (Haar) in a statistically principled test. To our knowledge that comparison (native vs Haar) has never been carried out before; it is the core methodological advance of the paper.

Reviewer #2, comment #2

Given the history of psychology and neuroscience over the last decades, and its reproducibility crises, I wonder how confident we should be in any effect size reported for the human brain that bordering at 100%. One might assume that such effect sizes either indicate a bias or bug, or relate to phenomena for which scientific protocols are so well advanced and sharpened (e.g. category selectivity), that rediscovering them is trivial. Or, the data is scaled in a way that suits the purpose of visualization well. The authors claim that the alignment in privileged axes approaches the maximal alignment achievable through any axis rotation. But what does this mean? Is it that privileged axes are an interesting biological phenomenon, or is it that the data sets we currently have capture little interesting effects going beyond category selectivity and retinotopy, either because of dataset bias, or because of limitations of the recording methods?

Our response #2.2

We understand the reviewer’s concern and have made extensive changes to address this point. Previously, we reported our axis-alignment as the ratio

$$\Delta_{\text{native}} = \frac{s_{\text{native}}}{s_{\text{ceiling}}} \quad \text{where} \quad s_{\text{ceiling}} = \max_{Q \in SO(N)} s(XQ, Y)$$

is the maximal alignment achievable by any rotation of the axes. This normalisation anchors the scale to the question we care about: *how much of the alignable structure is already present in the native axes?*. A ratio near one does not imply perfect correlation; it simply means that further rotations cannot improve the match by very much. In practice the ratios lie in the 0.80–0.98 range, not at 1.0 (with higher values for inter-model comparisons). To avoid the impression that this upper bound s_{ceiling} reflects a theoretical “ceiling”, we now refer to it as a *rotation ceiling*, since it represents the maximal alignment achievable given the restricted class of transformations we consider (rotations). More flexible linear or nonlinear mappings could in principle yield higher values, but our goal here was to quantify maximum achievable alignment under geometry-preserving transformations only.

Crucially, the pattern appears across three measurement modalities (fMRI, ECoG, single-unit recordings), two species, and several independently initialised CNN models—far more extensive replication than is typical in representational-similarity studies, making it highly unlikely that the effect is a dataset-specific or sampling fluke. This built-in replication across independent datasets and modalities directly addresses concerns raised by the reproducibility crisis, demonstrating that privileged axes are not idiosyncratic findings but robust, generalizable phenomena.

However, we agree with the reviewer that normalization can obscure the absolute effect size. Accordingly, in the revised manuscript we have removed the normalization and updated all figures to report the raw (unnormalized) pairwise-matching values for every comparison and rotation level (see Figs. A2,A3,A4 below), along with the corresponding maximum alignment value across all axis rotations (the *rotation ceiling*). In addition, Fig. 2 from the original manuscript has been expanded to include a new simulation panel (see Fig. A2B below) illustrating how alignment values at arbitrary axis rotations (what we now refer to as the *rotation floor*) depend on the intrinsic dimensionality of the underlying representation; see further discussion in our response #2.5 below. The native-versus-rotation floor drop is statistically significant in every species, model pair, and model-brain dataset combination we tested. Notably, when we look at the raw alignment values in the native axes against the maximum alignment achievable across all rotations, we see that pairwise alignment already accounts for a vast majority of what rotation-based methods (like hyperalignment or Procrustes) can achieve. For example, in the NSD fMRI data the native brain–brain alignment is 0.47 versus a rotation ceiling of 0.51; in the ManyMonkeys single-unit recordings it is 0.34 versus 0.35; and in human auditory cortex ECoG it is 0.57 versus 0.60.

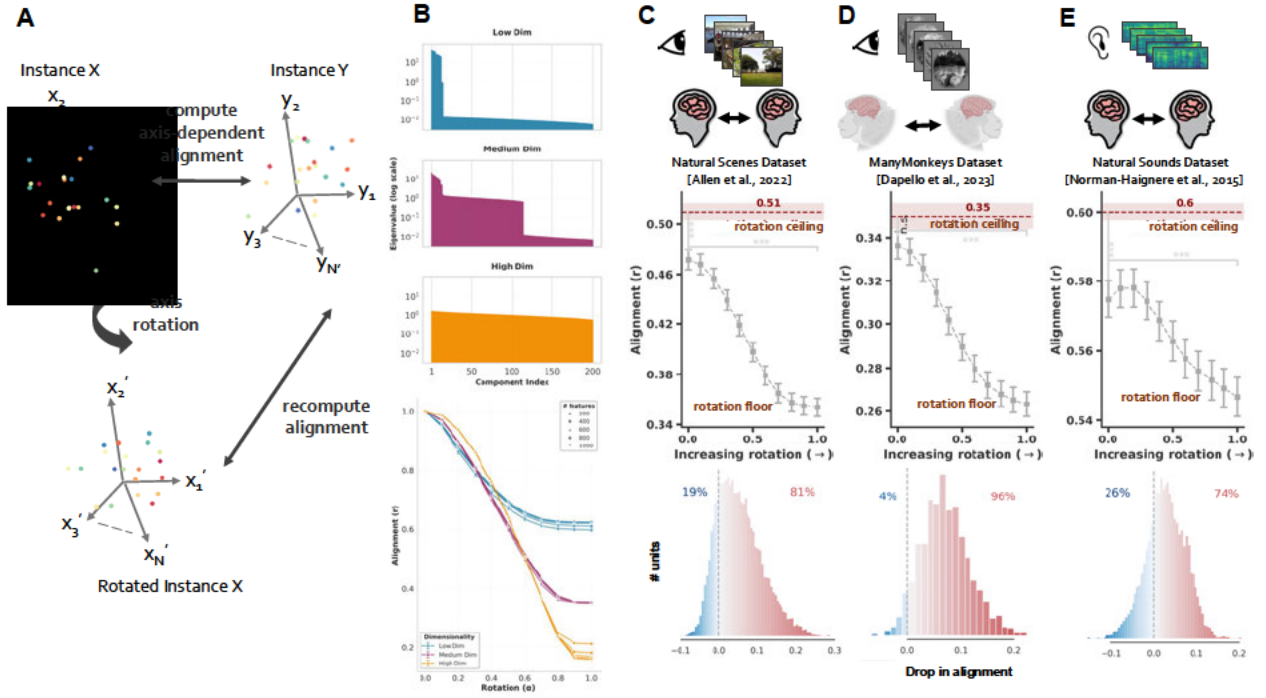


Figure A2: Brain-to-brain alignment (A) Cartoon schematic presenting the framework used to probe the prevalence of privileged axes in sensory cortical representations. Colored dots represent individual stimuli and axes $x_1, \dots, x_N$ and $y_1, \dots, y_N$ represent responses of individual voxels or 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. (B) Empirical “rotation floor” of pairwise matching alignment under three simulated data regimes ($M=3000$ stimuli; $N \in \{200, 400, 600, 800, 1000\}$; 5 random seeds). (*High-dim*) *Isotropic Gaussian data*: $X \in \mathbb{R}^{M \times N}$ drawn i.i.d. from $\mathcal{N}(0, 1)$, with $Y = X$. Even after applying random orthogonal transformations to X , alignment with Y remained at ~ 0.2 , indicating a nonzero rotation floor even for fully unstructured Gaussian data. (*Low-dim*) *Clustered latent modules*: X composed of $K = 6$ independent low-dimensional latent modules (module dim = 2; within-module noise = 0.1), $Y = X$. 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. (C), (D) and (E) depict alignment between conspecifics neural responses in human ventral visual cortex, macaque IT and human auditory cortex. This alignment is measured using the pairwise matching score and examined as a function of axis rotation (α) where $\alpha = 0$ indicates no rotation and $\alpha = 1$ indicates full rotation using a rotation matrix drawn uniformly from $SO(n)$. We also report the maximum possible value obtainable across all axis rotations (the “rotation ceiling”), determined using the ‘Optimal Axis Alignment’ procedure described in the text. Histograms depict the distribution of the difference in pairwise matching scores between the native and rotated basis across all neurons/voxels with the bars color-coded based on difference values. 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 ($\alpha = 1$), as well as with the optimal alignment values observed across all rotation angles. The shaded region indicates the SEM for the optimal alignment value across all rotation angles, computed from multiple data splits.

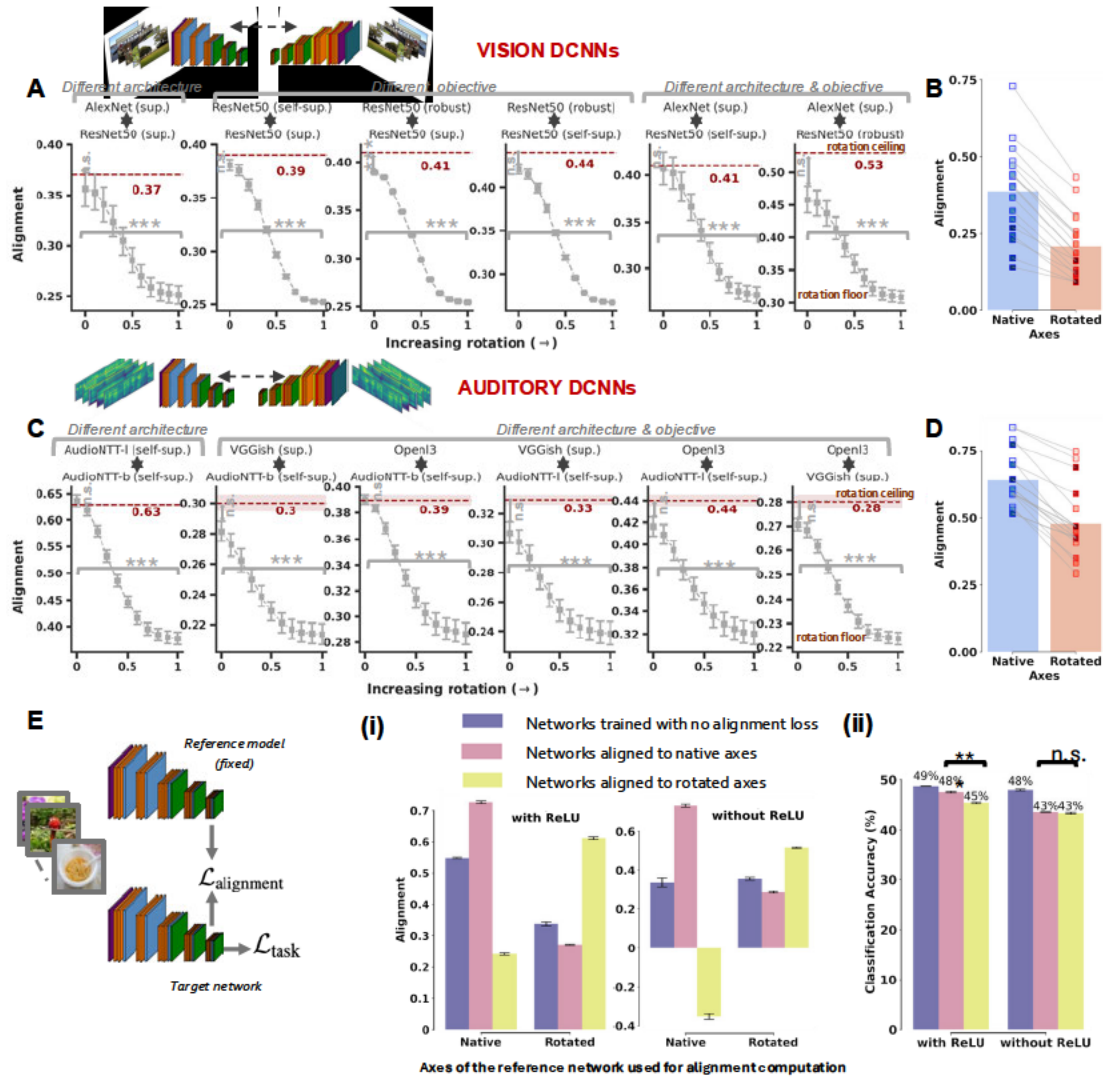


Figure A3: Model-to-Model alignment (A) Alignment between late-stage representations of different visual deep convolutional neural networks trained with different architectures and learning objectives. This alignment is measured using the pairwise matching score and examined as a function of axis rotation (α) where $\alpha = 0$ indicates no rotation and $\alpha = 1$ indicates full rotation using a rotation matrix drawn uniformly from $SO(n)$. We also report the maximum possible value obtainable across all axis rotations (the “rotation ceiling”), determined using the ‘Optimal Axis Alignment’ procedure described in the text. Alignment computed from activations to the 1000-image NSD stimulus set used for computing inter-subject alignment. 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 ($\alpha = 1$), as well as with the optimal alignment values observed across all rotation angles. The gray shaded region indicates the SEM 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 color-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-subject alignment analysis. In each case, we gradually rotate 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-l in the native and rotated axes corresponding to $\alpha = 1$. (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. Error bars represent the SEM across 10 networks with different initializations.

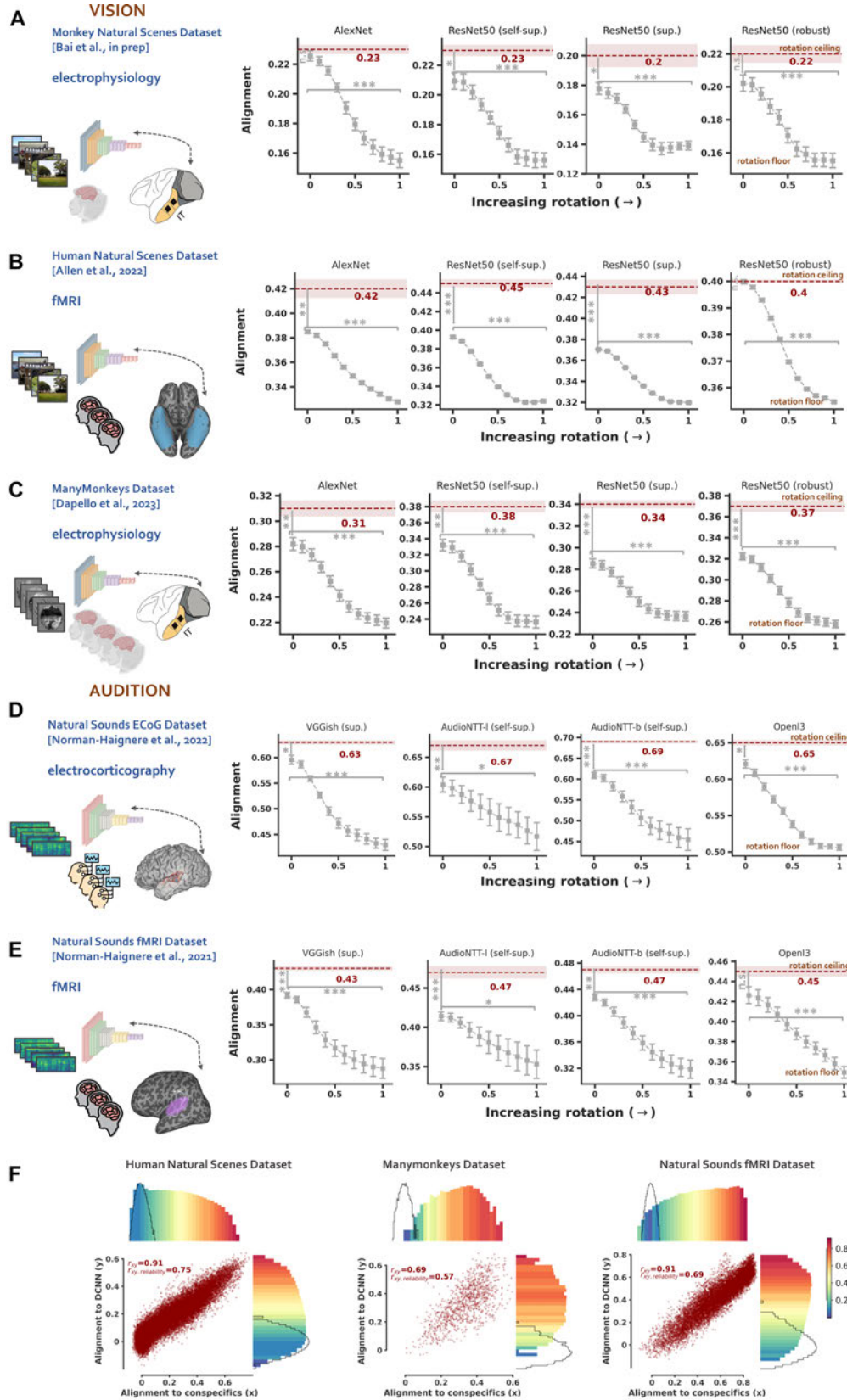


Figure A4:
Model-brain alignment.

(A–C) Alignment of visual DCNNs with neural responses in macaque IT and human ventral visual cortex, measured via the pairwise matching score as a function of rotation angle α ($\alpha=0$: native; $\alpha=1$: full random rotation from $SO(n)$). (D–E) Same for auditory DCNNs and human auditory cortex. The gray band denotes the SEM of the optimal “rotation ceiling” obtained via the Optimal Axis Alignment procedure; error bars are SEM across data splits. (F) One-to-one alignment of individual neurons/voxels across conspecifics versus their alignment with DCNN units (self-supervised visual/auditory models). Density plots show empirical (colored) and permutation-based null (black) distributions; color indicates voxel noise ceiling. Insets report raw and noise-ceiling-controlled partial correlations.

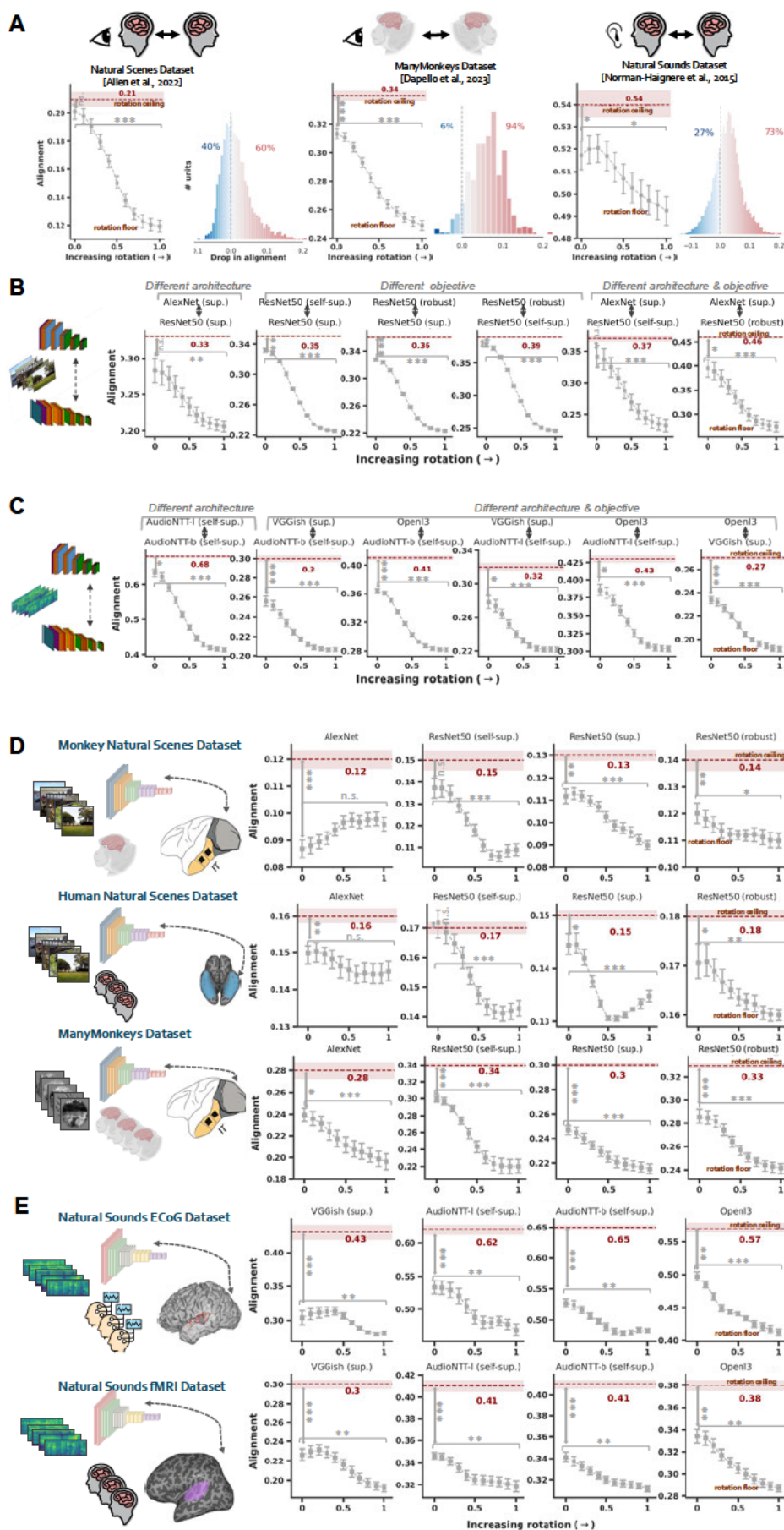


Figure A5: Axis-alignment is not limited to category selectivity Same analyses as Fig. A4 after removing all category-selective units/voxels. (A) **Brain-to-Brain:** Alignment between conspecifics neural responses in human ventral visual cortex, macaque IT and human auditory cortex after removing all 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 are SEM over multiple splits of the data. Significance tests compare the alignment values at no rotation ($\alpha = 0$) with those at full rotation ($\alpha = 1$), as well as with the optimal alignment values observed across all rotation angles. The gray shaded region indicates the SEM for the optimal alignment value across all rotation angles, computed from multiple data splits.

Moreover, privileged-axis alignment persists after we mask out all voxels and units that meet a liberal category-selectivity threshold, showing that the phenomenon cannot be reduced to a rediscovery of well-known category maps. We have now also updated Fig. 8 to report unnormalized values of pairwise alignment for these experiments (see Fig. A5 below). We have also updated Fig. 9 in the revised manuscript to report unnormalized alignment between conspecific neural responses in macaque dorsomedial frontal cortex (DMFC) during a mental simulation task and human language cortex responses during sentence comprehension. Finally, all supplementary figures have been revised accordingly to show raw, unnormalized alignment values rather than normalized ones.

We also emphasize that while establishing privileged axes in the brain was a critical first step, our study goes far beyond this. We go on to show, for the first time, that privileged axes also arise in artificial neural networks—and, critically, that these axes align between models and biological systems. We then explain why these axes emerge in models (and potentially also in brains), linking them to symmetry-breaking nonlinearities (e.g. ReLU-style activations) and demonstrate their concrete functional benefits, from sparser population codes to more efficient downstream read-outs and improved generalization. Altogether, our work offers a unified computational framework that explains—and extends—decades of isolated observations of category selectivity, retinotopy, and tonotopy by revealing the principled, mechanism-driven reasons neural systems converge on particular representational axes.

Note on normalization effects. Statistical values differ slightly between the normalized and unnormalized analyses because normalization rescales each subject-pair by its own baseline (“percent-of-optimum”) before averaging, whereas the unnormalized version averages in raw units. Since division and averaging do not commute, these two procedures yield different weighting across pairs and variance structures, leading to small shifts in the resulting t -statistics.

Reviewer #2, comment #3

Is the method to derive the maximal alignment guaranteed to find the maximal alignment? Under what assumptions? It does find a maximal alignment in the simulations, but how certain can we be that it does so in the brain data?

Our response #2.3

We acknowledged in our paper that 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). We optimize for Q^* by treating orthogonal matrices as a Riemannian manifold and performing gradient-based searches via the Cayley-transform trivialization in the geotorch library.⁴ Because the Cayley map is a smooth, locally diffeomorphic parametrization of $SO(n)$, this is equivalent to Riemannian gradient descent under a change of metric—hence standard results guarantee convergence to a critical point of our smooth alignment objective. We verified that for each optimization, random seeds converge essentially to the same value. Admittedly, gradient-based manifold optimization cannot offer a worst-case proof of global optimality in a nonconvex landscape; nonetheless, the compactness of $SO(n)$, the smoothness of our alignment metric, and the empirical consistency across initializations give us high confidence that the Q^* we report closely approximates the true maximal alignment.

Reviewer #2, comment #4

The relevance of axis alignment can be judged differently, depending on whether it is related to the maximal alignment, or minimal alignment after rotation. Generally, the alignment (in %max) drops very little with increasing rotations (e.g., Fig. 2D from 96% to 91%, approximating an asymptotic state. So in effect, even full rotation leads to an alignment of 91%. It seems this tells us that privileged axes account for 5% of a 100% alignment (and maybe 23% max in vision/NSD). This is a small number, not supporting the authors’ claims about the importance of privileged axes. Admittedly, the number of units that shows drop in alignment is relatively large—but so is the number of neurons that show an increase.

Our response #2.4

We understand and appreciate the reviewer’s concern which has prompted us to clarify our metrics and our interpretation of our findings. When measuring the effect size of axis-alignment, there are two reference points. The maximum alignment ceiling is included only to show how much additional structure could, in

principle, be captured by an optimal rotation (hence, we now call it the “rotation ceiling” instead); it is not the test of privileged axes.

The “5 %” is a *mis-read*: it is not correct to take the *raw* difference ($96 - 91 = 5$ percentage points) and treat it as “five percent of all possible alignment.” The relevant scale is not $0 \rightarrow 100$; a *randomly rotated* basis already attains 91% of the ceiling because rotations preserve all stimulus geometry. A Haar-random rotation scrambles axes, not information. Because the rotation is orthogonal it preserves every inter-stimulus distance, every linearly decodable feature, and the overall signal-to-noise ratio. Consequently, when we try to match two populations after such a rotation, we are still comparing two data matrices that encode similar information content—just in different coordinates. This will almost always give a non-zero baseline (see our response to the next comment). The correct figure of merit is therefore the *fraction of the rotation-invariant gap that native axes close*, i.e. the *alignable gap* above that rotation floor:

$$\Delta_{\text{native}} = \text{Share captured by native basis} = \frac{s_{\text{native}} - s_{\text{floor}}}{s_{\text{ceiling}} - s_{\text{floor}}} \quad 0 \leq \Delta_{\text{native}} \leq 1.$$

where s_{native} is the alignment in native axes, s_{ceiling} is the rotation ceiling and s_{floor} is the rotation floor (alignment under a uniformly random rotation drawn from $\text{SO}(n)$).

- $\Delta_{\text{native}} \approx 1$: the native basis is already as good as any rotation can make it.
- $0 \ll \Delta_{\text{native}} < 1$: the axes are still *privileged*—they close a large fraction of the possible gap, even though a fine-tuned rotation could squeeze out a little more.
- $\Delta_{\text{native}} \approx 0$: the axes behave like arbitrary coordinates.

In our case,

$$\Delta_{\text{native}} = \frac{96 - 91}{100 - 91} = 55\%.$$

Thus the native axes close **more than half** of the distance that *any* rotation could ever bridge—far from the “only five-percent” implied by subtracting ceilings from baselines.

Moreover, Fig. 5 demonstrates that the same axis advantage yields marked behavioral gains (sparser codes, cheaper read-outs, better few-shot generalisation).

Although we discuss these normalizations here for clarity in addressing the reviewer’s question, we have removed all normalization from the revised manuscript itself. All figures now report only the raw (un-normalized) alignment values in the native and rotated axes, along with the rotation ceiling, to avoid any ambiguity in interpreting effect sizes.

Reviewer #2, comment #5

Echoing the last point, random ANNs have alignment scores of 85% (Fig. S7). So what does this tell us about the lower bound of the alignment measure and its sensitivity? Theoretically the lower bound is 0. This is in conflict with the claim that ‘untrained networks do not show axis alignment with neural representations (Fig. S8).

Our response #2.5

The claim ‘untrained networks do not show axis alignment with neural representations’ is supported by the fact that random axes are indistinguishable from native axes of untrained models, in terms of alignment with neural representations (Figs. S7 and S8). Further, when we look at the raw alignment scores, they are indeed significantly lower for random ANNs than trained networks (see Fig 6).

The lower bound is never 0 practically because if two populations encode the same underlying stimulus variables—even after an arbitrary orthogonal transform—some positive similarity is inevitable. What determines the height of this floor is the latent structure of the data. To demonstrate why the alignment floor we observe—even for completely random rotations—never actually reaches zero, we ran two parallel simulations. In our first simulation (**high-dimensional** scenario in Fig A2,B), we generated an entirely isotropic population by drawing $\mathbf{X}$ from a standard normal distribution in N dimensions (with N varying over 200,

400, 600, 800, and 1,000) and then simply set $\mathbf{Y} = \mathbf{X}$ to ensure that the maximal possible alignment is 1, and any reduction in alignment must arise purely from our imposed rotations rather than from any intrinsic mismatch between $\mathbf{X}$ and $\mathbf{Y}$. When we applied a random orthogonal transform to $\mathbf{X}$ and re-computed the pairwise alignment with its unchanged copy $\mathbf{Y}$, we found that—even in this maximally structureless, full-rank case—the worst-case rotation still achieved an alignment of ~ 0.2 . In other words, even data that a priori fill their ambient N -dimensional space uniformly maintain a nonzero “floor” of similarity under arbitrary rotations.

In a second simulation (**low-dimensional** scenario in Fig A2,B) designed to mimic the clustered tuning often observed in real neural populations, we carved $\mathbf{X}$ up into a handful of tight clusters by letting small low-dimensional latent drivers feed into groups of neurons (again with $\mathbf{Y} = \mathbf{X}$ so that the true maximum alignment is pinned at 1). Here the same sweep through random rotations resulting in an alignment values ~ 0.6 , and this floor was again stable across all five choices of feature counts. This dramatic increase in the rotation “floor” simply reflects the fact that clustered—and more generally low-dimensional—data occupy only a small subspace of their full N -dimensional embedding. In a third simulation (**mid-dimensional** scenario in Fig A2,B), we created an intermediate case by combining both structured and random components: we generated $\mathbf{X}$ with partial clustered structure where half of the features formed loose clusters while others remained randomly distributed. This hybrid geometry yielded an alignment floor of ~ 0.37 under random rotations—between the extremes of fully isotropic (0.2) and highly clustered (0.6) representations. Together, these simulations demonstrate that the alignment floor is a direct reflection of the intrinsic dimensionality and geometric structure of the representations being compared.

Since both biological recordings and trained networks similarly reside on highly structured, low-dimensional manifolds rather than uniformly filling every axis, it is this latent geometry that enforces a practical lower bound well above zero.

Note that in these simulations we set $\mathbf{Y} = \mathbf{X}$, ensuring a rotation ceiling of 1. In empirical datasets, the corresponding ceiling is typically lower because the compared populations (e.g., brains or independently trained models) do not share identical latent dimensions and include measurement noise.

To illustrate an even more neuroplausible scenario, we simulated two populations of smooth, bell-shaped “receptive-field” neurons driven by the same low-dimensional stimulus $z \sim \text{Unif}([-1, 1]^2)$. Each unit is assigned a randomly chosen preferred stimulus $c_i \in [-1, 1]^2$ and exhibits a radial-basis tuning curve

$$f_i(z) = \exp(-\|z - c_i\|^2/b^2),$$

where b is a global bandwidth parameter. We used RBF tuning because it provides a simple response model that mimics generic “receptive-field” neurons while preserving analytic control over shared latent structure. A fraction f of the N units is *aligned*: the corresponding columns of the two response matrices $X, Y \in \mathbb{R}^{M \times N}$ are identical up to small i.i.d. Gaussian noise. The remaining $U = N - \lfloor fN \rfloor$ units share a common RBF basis $B \in \mathbb{R}^{M \times U}$ (constructed from U new centres) but are remixed independently in the two populations:

$$X_{\text{un}} = BW_X, \quad Y_{\text{un}} = BW_Y, \quad W_X, W_Y \sim \mathcal{N}(0, 1)^{U \times U}.$$

We concatenate the aligned and unaligned blocks and add low-amplitude i.i.d. Gaussian noise to obtain the final matrices $X = [X_{\text{aligned}} | X_{\text{un}}] + \eta_X$ and $Y = [Y_{\text{aligned}} | Y_{\text{un}}] + \eta_Y$. Empirically, we observe that even *after a complete axis rotation*, the alignment score never falls to zero; instead, **it remains in the 60 – 70% range for all values of f and N** (Fig. A6). This residual similarity is exactly what one would expect when both populations continue to encode the *same* latent stimulus variables—here the coordinates of z —even though the unit-wise correspondence has been obliterated by the rotation. The result thus confirms that the non-zero “floor” of the metric is a property of the shared representational content.

In summary, the alignment “floor” depends on the nature of the representation. For highly structured or low-dimensional representations—as in both brains and trained ANNs—the floor will inevitably be high.

This also helps contextualize the auditory cortex results. Prior work has shown that auditory cortex is especially low-dimensional—six components explain $> 80\%$ of voxel response variance norman2015distinct—and in our data this geometry naturally leads to a high rotation floor of $\sim 91\%$. Thus the auditory alignment numbers are not an inconsistency, but exactly what one would expect from the low-dimensional structure of that system.

We now highlight in the manuscript that the alignment floor depends on the latent structure of the representations. The critical evidence for privileged axes comes not from the absolute floor value but from the statistically principled test showing that native axes consistently outperform uniformly-drawn random rotations. It reads:

“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. 2B through a series of simulations systematically varying the intrinsic dimensionality and structure of the underlying representation. Each simulation involved two populations $X, Y \in \mathbb{R}^{M \times N}$ (with $M = 3000$ stimuli and $N \in \{200, 400, 600, 800, 1000\}$), and alignment was computed under repeated random rotations of X across five random seeds.

*In the **high-dimensional isotropic** regime, we drew X from an i.i.d. standard normal distribution and set $Y = X$ to ensure that the true maximal alignment equals 1. Even after applying random orthogonal transformations to X , alignment with Y remained at ~ 0.2 , indicating a nonzero 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, 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, i.e. the alignment “floor” 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.

Also, 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 (e.g., brains or independently trained models) do not share identical latent dimensions and include measurement noise.

”

Reviewer #2, comment #6

I wonder whether the null hypothesis, i.e. drawing rotations from $SO(n)$, is a reasonable assumption. For example, is it reasonable to assume that one human instance might show positive effects, where another person would show the exact same effects, just multiplied by -1 (a perfectly fine rotation). Relatedly, what if you constrain the null to contain rotations that result in rotated matrices of similar sparseness as the original matrix?

Our response #2.6

Thank you for this suggestion. We agree that considering constrained nulls is an important check. We sampled rotations uniformly from $SO(n)$ as this is statistically principled: the Haar measure is the unique probability measure on $SO(n)$ under which every orthogonal transformation is equiprobable. This property guarantees that any systematic advantage of the native axes over Haar-random axes must stem from structure in the data, not from bias in the sampling scheme.

Radial Basis Tuning Simulations

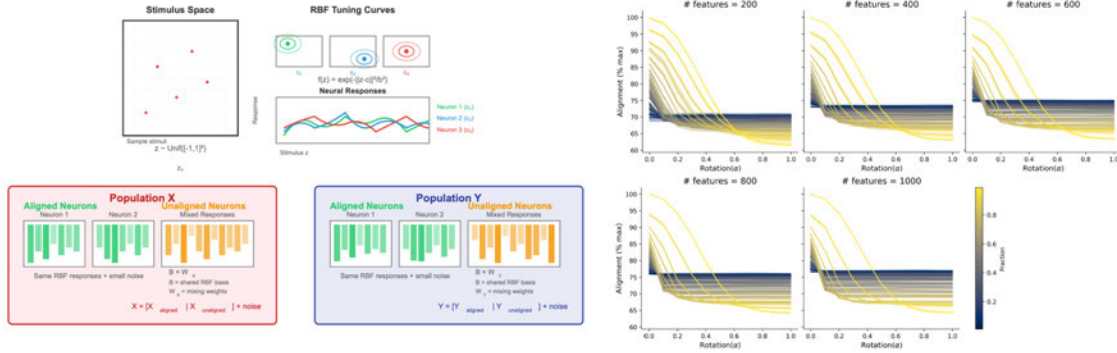


Figure A6: **Radial-basis receptive-field simulations:** A fraction f of units in X and Y are independently noisy RBF encodings of the same 2D stimulus (bandwidth= 0.3; alignment noise= 0.05), while the remaining units share only common RBF basis functions before independent random mixing (global noise= 0.1). Even after full rotation of X , alignment with Y never falls below 60–70%, reflecting the irreducible similarity imposed by the shared latent stimulus manifold. Curves show mean $\pm$ 1 SD across five seeds; shaded bands denote one standard deviation.

Nevertheless, to demonstrate that our conclusions are not an artefact of using Haar draws with unconstrained sparsity, we now implemented a second null in which every rotation is explicitly required to preserve the sparseness of each subject’s response. Specifically, with

$$S(x) = 1 - \frac{\langle |x| \rangle^2}{\langle x^2 \rangle}, \quad \bar{S}(X) = \frac{1}{n} \sum_{i=1}^n S(X_{*i}),$$

we solved

$$\min_{R \in \text{SO}(n)} |\bar{S}(XR) - \bar{S}(X)|.$$

We parameterised R as the weight of a square linear layer constrained to the Stiefel manifold via **GeoTorch** and minimised the absolute error with Adam (learning rate 10^{-4} , 200 iterations). For every subject in the *ManyMonkeys* dataset the optimiser produced a rotation whose rotated matrix matched the original mean sparsity to within $\leq 10^{-4}$ (fourth decimal place).

Even under this stricter sparsity-matched null, alignment in the native axes consistently exceeded that obtained from both Haar-random rotations and sparsity-preserving rotations (see Fig. A7 below, Fig. S12 in the revised manuscript). These findings demonstrate that neither lifetime sparsity nor the choice of the Haar measure drives our main result. For clarity and computational feasibility, we chose to perform the sparsity-matched control on the ManyMonkeys dataset: mean lifetime sparsity is a well-defined, single-unit property, whereas fMRI and ECoG signals—being spatially and temporally blurred aggregates—lack a clear per-voxel or per-electrode sparsity analogue.; extending the same control to every comparison (model–model, model–brain, brain–brain) across all datasets and metrics is in principle possible but would also require several orders of magnitude more compute and is therefore beyond the scope of the present study.

We have now incorporated these results into the revised manuscript with the following text:

“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 $R \in \text{SO}(n)$ that preserved the mean lifetime sparsity of each response matrix. Even under this stricter null, alignment in the native axes was consistently higher (Fig. S12). This confirms that our findings are not an artefact of the choice of null distribution, but reflect genuine axis structure in the data.”

Reviewer #2, comment #7

S4 reports the brain-to-brain alignment values for the natural sounds data set, with larger impact of rotations

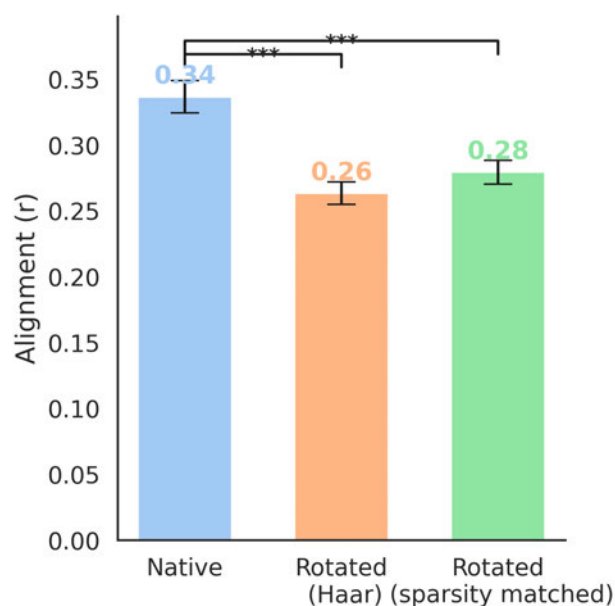


Figure A7: **Effect of axis rotations on pairwise alignment in single-unit recordings (ManyMonkeys dataset).** Alignment between pairs of monkeys is shown under three conditions: Native axes (no rotation), Haar-random rotation (axes rotated by a uniform draw from $SO(N)$), Sparsity-matched rotation (orthogonal 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.

when using the soft matching score than the score in Fig. 2. Intuitively, I would have expected the opposite. Can you clarify?

Our response #2.7

It’s hard to be certain without further experiments, but a likely explanation is simply that our two metrics respond very differently to any rotation of the axes. The pairwise score is fundamentally “local” – it lets each neuron pick its single best partner (even if many X’s map to the same Y), so even a full rotation need only disrupt a minority of top matches to leave the average largely intact. In contrast, the soft-matching score enforces a balanced, population-wide transport plan: every neuron’s tuning must be matched in concert, so a global reassignment caused by a rotation accumulates into a much larger overall drop. We now explicitly note this in the caption of Fig. S4, which reads:

“Note that the soft matching score shows a larger impact of rotations compared to pairwise matching in audition (cf. Fig. 2), likely because soft matching enforces a balanced, population-wide assignment where every unit must be matched in concert, whereas pairwise matching allows one-to-multiple assignments that may be more robust to rotations.”

Reviewer #2, comment #8

The authors argue that axis alignment, associated with sparsity, decreases wiring costs (Fig. 5). This is an interesting point. However, as the authors also note, it is limited to the readout of categories. In the spirit of showing that the results here are about more than category preference, what about reading out other types of information that are category orthogonal (e.g. akin to Hong et al., 2016)? Demonstrating this would strengthen the claims about the relevance of preferred axes.

Our response #2.8

We agree that it is critical to show that the wiring-cost benefits of privileged axes extend beyond just “obvious” category-selective cases (faces, places, etc.). In our original manuscript we already decode all 80 categories in our image set (not just the categories for which selective responses have been recorded in the brain), demonstrating that the axis-alignment advantage holds across the full gamut of object classes. In response to the reviewer’s suggestion, we have now also run linear readouts on six category-orthogonal object properties—object scale (s), two-dimensional translations (t_y , t_z) and three Euler rotations (r_{xy} , r_{xz} , r_{yz})—using the same neural recordings as Hong et al. (2016) (their “MajajHong” dataset).⁸ On

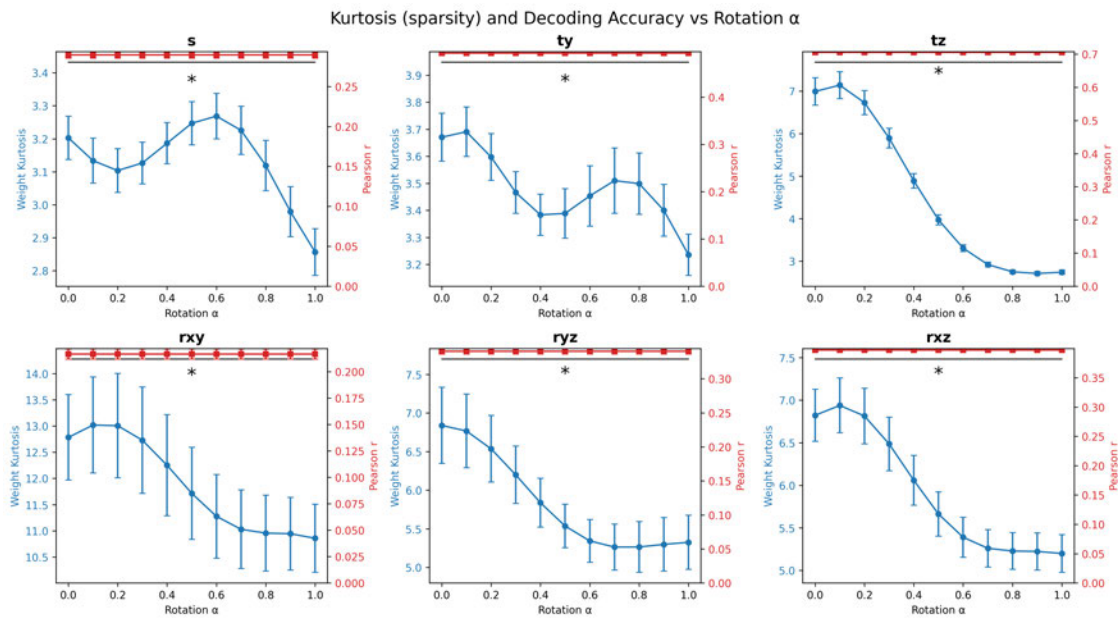


Figure A8: **Axis-rotation impacts readout sparsity for category-orthogonal properties in IT.** Each panel shows results for one of six continuous object properties: scale (s), horizontal translation (t_y), vertical translation (t_z), in-plane rotation (r_{xy}), elevation rotation (r_{xz}) and azimuth rotation (r_{yz}). On the x-axis is the axis-rotation parameter α (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 α . The blue curves (left ordinate) plot the corresponding readout sparsity, quantified by the kurtosis of the weight distribution: as α increases, sparsity systematically decreases (i.e. weight distributions become less peaked) for most properties, indicating higher inferred wiring cost when decoding from random axes.

each property we fit a separate linear readout (cross-validated ridge regression) to the penultimate-layer activations of self-supervised ResNet-50 (MoCo) as well as to the neural recordings. In each case we measured the *kurtosis* (κ) of the learned weight distribution, noting that higher κ corresponds to sparser weights and therefore lower implied wiring cost. As shown in Fig. A8 (Fig. S10 in the revised manuscript), decoding from the native (‘privileged’) axes of the IT data yields significantly larger κ than decoding from random axes, implying lower wiring cost. We observe the same pattern in Fig. A9 (Fig. S11 in the revised manuscript). Thus, the decrease in inferred wiring cost that privileged axes afford is not limited to category-labels but generalizes to the full suite of “nuisance” variables that Hong *et al.* showed are richly represented in IT cortex.

We now include these analyses in the Supplementary Materials (Figs. S10, S11) and briefly summarize them in the main paper with the following text:

“Beyond categorical selectivity, we also tested whether privileged axes confer wiring-cost benefits for category-orthogonal properties. Using the IT recordings of Hong et al. (2016), we fit linear readouts for object position, scale, and rotation and compared weight sparsity across native versus rotated axes and find that privileged axes reduce wiring cost for these very variables as well (Figs. S10, S11).”

Reviewer #2, comment #9

I agree that ANNs trained with supervision, particularly using categories related to categories for which category selective regions exist are likely to exhibit category selective tuning (p. 14). The authors thus propose using self-supervised ANNs. I instead suggest either training supervised ANNs without categories related to category-selective regions, or removing such categories (and related ones in a wide sense) from the training material for self-supervised ANNs. This might better assess alignment beyond category selectivity.

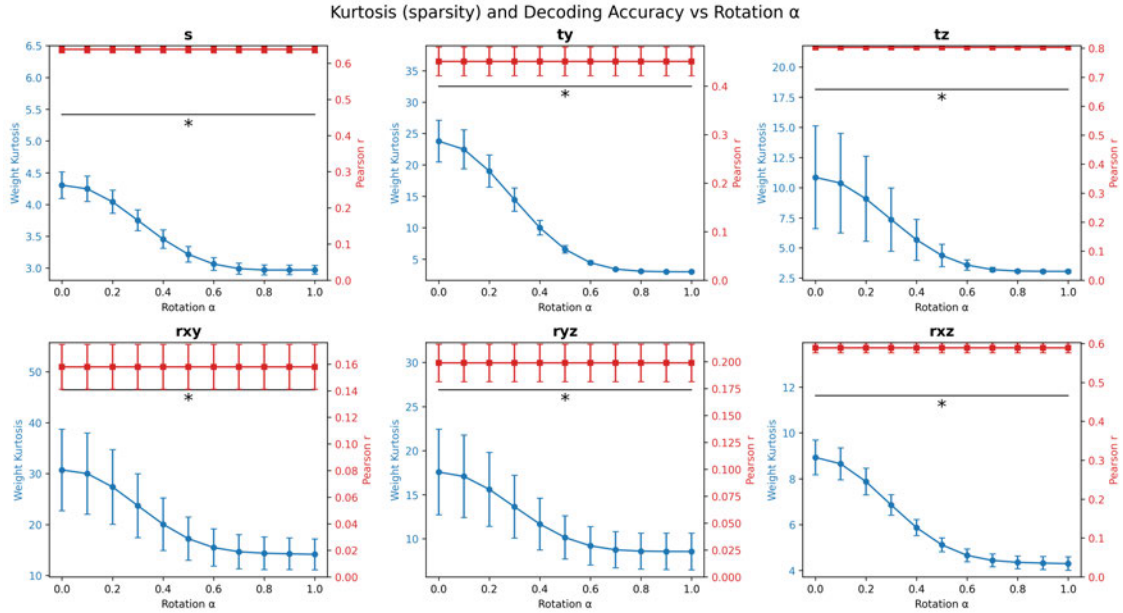


Figure A9: **Axis-rotation impacts readout sparsity for category-orthogonal properties in ANN activations.** Each panel shows results for one of six continuous object properties: scale (s), horizontal translation (t_y), vertical translation (t_z), in-plane rotation (r_{xy}), elevation rotation (r_{xz}) and azimuth rotation (r_{yz}). On the x-axis is the axis-rotation parameter α (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 α . The blue curves (left ordinate) plot the corresponding readout sparsity, quantified by the kurtosis of the weight distribution: as α increases, sparsity systematically decreases (i.e. weight distributions become less peaked) for all properties, indicating higher inferred wiring cost when decoding from random axes.

Our response #2.9

We appreciate the reviewer’s suggestion but this experiment is beyond the scope of the paper. Our objective in the present work is not to re-engineer the training diet, but to ask whether axis alignment persists once category-selective units are removed. We therefore adopt the simpler and more direct manipulation—masking category-selective voxels/units with a liberal threshold in both brains and models—because it cleanly answers that specific question and keeps the experimental space tractable.

Altering the entire training set would address a different (albeit important) issue: how training statistics shape the existence of category-selective axes in the first place. That enterprise is non-trivial: (i) completely purging a million-image corpus of every trace of face-, music-, or other category-selective content is practically infeasible and prone to leakage, and (ii) prior work shows that even when such images are withheld, selective units often still emerge spontaneously (e.g., Yamins et al, 2015; Baek et al., 2021; Kim et al., 2024).^{2,10,16}

Reviewer #2, comment #10

The authors provide an alignment analysis after excluding category selective voxels/electrodes units. I suggest a more axis-related approach for stronger inference. The X top ANN and neural components found using Bayesian Non-Negative Matrix Factorization correspond to categories for which category-selective regions exist. Exclude those components and then calculate alignment between brains.

Our response #2.10

Unfortunately, there is no straightforward way to remove NMF components. Components are mixtures, not pure voxels. Each component yields (i) a pattern of voxel weights and (ii) a response vector. The voxel weights are in arbitrary units and in general have modest but non-zero loadings on the vast majority of voxels. One could instead drop voxels whose weights exceed some ad-hoc threshold, but this adds an extra hyper-parameter and, in practice, duplicates what we already do: in Khosla et al., 2022 (Curr Biol.) we showed that quantitatively, the voxel weight maps demonstrated high correlations with the t statistics pertaining to the relevant categories from the functional localizer experiments (~ 0.8 correlation or higher for most categories). Thresholding those weight maps would therefore remove essentially the same voxels that our present localizer-based mask already excludes.

Reviewer #2, comment #11

Interestingly, the authors show that their methodology yields significant results also beyond sensory cortices. However, only the first, most basic analyses are presented here without the sophistication of the additional analyses for sensory cortices. I clearly see the value of this, but at the same time, these analyses are highly incomplete, and one wishes for a more thorough treatment. At the same, the manuscript is already very long (to the point of being too long, losing its focus), and it seems more like two or three publications to related topics that were put into one manuscript for some reason. I invite the authors to condense the paper or split the manuscript in two or three to make it better readable.

Our response #2.11

We opted to retain the language- and frontal-cortex analyses because they uniquely show that privileged axes are not confined to sensory systems—a conclusion that cannot be drawn from visual- and auditory-only results. To signal their exploratory status, we have:

- Retitled the section from “Privileged axes exist beyond the sensory cortices in higher-order neural systems” to “Preliminary evidence for privileged axes in higher-order cortex.”
- Inserted a two-sentence caveat at the start of that section:

“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.”

Reviewer #2, comment #12

Minor: The intuitive explanation of the soft matching score (p. 4) is not intuitive, instead it refers to previous literature.

Our response #2.12

Thank you for pointing this out. We have now improved the intuitive explanation of the soft-matching score, clarifying that it generalizes strict one-to-one matching to allow partial (“soft”) assignments between neurons in representations with different sizes.

“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 like deeming a representation $\mathbf{X}$ with many units as somewhat aligned to any representation that is mapped to it.⁹ 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 Williams et al. [30] 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 (see⁹ and Methods for more background on this similarity measure).”

Reviewer #2, comment #13

L. 608: Why were not both units demonstrating selectivity for music and speech removed, but only one class? Few voxels might remain then, but that would be a clear point against the idea that there are privileged axes beyond category selectivity.

Our response #2.13

We did exclude both classes: any unit whose t-value exceeded threshold for music OR speech was removed, so voxels selective for music, for speech, or for both were all discarded. The phrasing “music or speech” denotes this union. We have clarified this in the manuscript. It now reads:

“..while in auditory systems, we removed all units demonstrating significant selectivity for either music or speech (i.e., units selective for music, speech, or both were excluded). ”

Reviewer #2, comment #14

The authors provide layer-specific alignment scores but averaged for layers. What is the pattern for each layer/block? How is the alignment score related to layers?

Our response #2.14

We believe the requested information is already in the manuscript. Fig 3B (vision) and 3D (audition) show every individual layer (one dot per layer; colour saturation encodes depth). The native-axis pairwise matching score (left dot of each pair) is always higher than the rotated-axis score (right dot) and both scores are generally higher for early layers and decrease with network depth. Soft-matching versions appear in Suppl. Fig. S5B,D and show the same pattern.

Reviewer 3 (Remarks to the Author):

Reviewer #3, comment #1

I read the manuscript “Privileged representational axes in biological and artificial neural networks” with interest. The paper investigates whether representational axes matter in biological and artificial neural networks, challenging the view that information coding is best understood solely through population-code geometry, which is blind to native neural tunings. The authors developed methods to test for alignment of neural tuning across brains and DCNNs. Furthermore, the representational axes of DCNNs trained on natural inputs were aligned with those in perceptual cortices, and this axis-alignment metric proved better at differentiating competing models of biological sensory systems. The study also demonstrates that coding schemes prioritizing certain axes can improve task performance, reduce downstream wiring costs, and enhance generalization.

Overall, I found the question interesting and novel.

Our response #3.1

Thank you for your thoughtful summary and positive feedback.

Reviewer #3, comment #2

A common practice within research on population-code geometry involves evaluating proposed metrics or novel measures based on their test/retest reliability, often assessed across independent recording sessions for the same subject or across different random initializations of the same computational model (e.g. Walther et al., Neuroimage, 2016; Shahbazi et al., Neuroimage, 2021). These important analyses of consistency appear to be absent from the current paper, and their inclusion could additionally serve to provide a more principled basis for selecting between the pairwise and soft matching metrics introduced by the authors. Furthermore, while inter-subject brain alignment might be influenced by factors such as the quality of spatial normalization, a potential caveat that is not present when conducting within-subject brain analyses, evaluating within-subject reliability across sessions would help to address this by demonstrating the stability of the findings within individuals.

Our response #3.2

We agree that any new similarity metric should first meet the basic standard of test–retest reliability. In earlier drafts we focused primarily on between-subject and model–brain comparisons, but in the revised manuscript we have now added a systematic within-instance reliability analysis in the Supplementary (see Fig. A10). We performed this wherever the available data made it feasible—specifically, for brain datasets with a reasonably high number of sessions and good session-level data per subject/animal, and for models where we could train multiple independent seeds. We did not perform this analysis for the Audioset models, since the two architectures we report (Audioset-B and Audioset-L) already share the same training data and objective and highly similar architectures, making their between-model alignment effectively an upper bound on within-model reliability.

Across the datasets where we could perform this analysis, we find that within-brain alignment in the auditory fMRI dataset is relatively close to the corresponding between-brain alignment, whereas for the electrophysiology dataset the within–between gap is larger—likely reflecting the sparse, randomly sampled neuron populations in each animal. For models, within-model alignment is generally higher than between-model alignment, as expected, but the two are often in a similar range, indicating that the representational axis structure stable within a model/brain is also shared with other models/brains.

These new results address the reviewer’s point by demonstrating the stability of privileged-axis structure within individuals (or model instances).

We now include these analyses in the Supplementary (Fig. A10 below, Fig. S9 in the revised manuscript) and describe them in the main text as follows:

“To confirm that our axis-dependent metrics meet the basic standard of test–retest reliability, we also examined within-instance alignment wherever the data allowed—specifically, across independent recording sessions for the same subject or animal, and across different random initializations for the same model architecture. We found that tuning functions were reliably preserved within individuals and were often shared across individuals, indicating that the privileged-axis structure

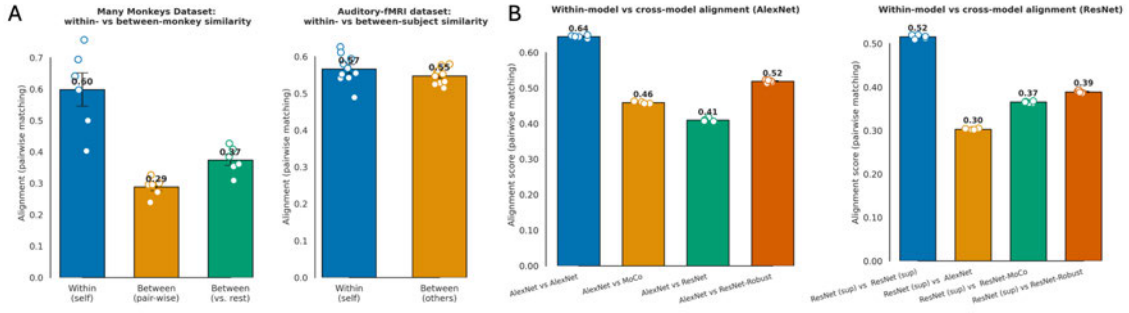


Figure A10: **Test–retest reliability of axis alignment within brains and within models.** (A) Within-brain alignment. Left: *ManyMonkeys* electrophysiology, estimated using split-half reliability across repeated stimulus presentations. Right: human auditory fMRI, where only three stimulus repetitions 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.

is both stable within a system and consistent across different instances (Supplementary Fig. S9).

Reviewer #3, comment #3

The authors employ a theoretically elegant approach to investigate privileged axes by applying geometry-preserving rotation operations to perturb the representational space of one population relative to another. This methodology forms a compelling foundation for testing the core hypothesis. As I understand it, the argument for privileged axes is primarily supported if the alignment is consistently greatest in the native, unrotated state ($=0$), and further strengthened if this native alignment also approaches the theoretical maximum achievable through any rotation. While the former condition—that the native alignment surpasses rotated alignments—appears to hold quite consistently across the presented results, the latter—that native alignment is near the estimated ceiling—does not seem to be universally met in all cases shown. Clarifying the relative weight or importance the authors place on these two criteria in drawing their conclusions, and perhaps discussing instances where the native alignment is not close to the theoretical maximum, could enhance the clarity and precision of the argument in the text.

Our response #3.3

Thank you for this insightful point. Our primary evidence for privileged axes comes from the consistent drop in alignment when moving from native axes ($\alpha = 0$) to fully random rotations ($\alpha = 1$), where those random axes are sampled uniformly from the Haar distribution over $SO(n)$. This comparison has a clear statistical grounding—every possible rotation is equally likely—so if the native basis were arbitrary, its alignment would be indistinguishable from this random baseline (what we call the “rotation floor”); the robust drop we observe after full rotation therefore suffices to show that the native axes are special.

The empirical ceiling—which we now refer to as the “rotation ceiling” to emphasize that it reflects the maximal alignment achievable under the specific transformation set considered (rotations)—serves a different purpose: it tells us how much alignment remains “on the table” if one could choose the best-possible rotation.

Our core claim of privileged axes does not hinge on that perfect maximality, but rather on their systematic advantage over arbitrary bases. Three situations can occur:

- **Native $\gg$ Rotation floor $\Rightarrow$** there is a privileged set of axes shared across systems.

- **Native $\approx$ Rotation ceiling** $\implies$ those axes already capture almost everything that can be aligned by *any* rotation.
- **Native $\gg$ Rotation floor** and **Native $\ll$ Rotation ceiling** $\implies$ 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-vs-rotation floor gap is the defining evidence for privileged axes; the native-vs-rotation ceiling comparison does not undermine the privileged-axis result, it simply tells us how much additional alignment could be squeezed out by the best-possible rotation. In our results, where the native alignment falls below the rotation ceiling, it still remains well above the rotation floor; thus, our core claims of privileged axes still hold. We have now added this explanation to the end of the section on “A formal framework for assessing privileged axes” to make these distinctions explicit for readers.

Interpreting the two benchmarks.

*For every pair of representations we therefore examine two reference points: (a) the **Rotation floor**, obtained by replacing X with XQ for Q drawn uniformly (Haar, $\alpha = 1$) from $SO(n)$; and (b) the **Rotation ceiling**, $s(XQ^*, Y)$, obtained by explicit optimization over all rotations. These two anchors let us distinguish three qualitatively different regimes: (i) **Native $\gg$ Rotation ceiling** $\implies$ axes are privileged: alignment cannot be explained by chance rotations; (ii) **Native $\approx$ Rotation ceiling** $\implies$ privileged axes already capture almost everything that can be aligned by *any* rotation.; (iii) **Native $\gg$ Rotation floor** and **Native $\ll$ Rotation ceiling** $\implies$ 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 vs. 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.*

Reviewer #3, comment #4

While the work is undeniably interesting, a question that lingered is precisely what novel insights are uniquely provided by this line of thinking focused on representational form, insights that could not be gleaned from the more traditional approach of examining representational content. Figure 6 offers a hint by demonstrating that while content-based measures might suggest comparable similarity to the brain across different model architectures (apart from untrained networks), focusing on representational form reveals clearer distinctions. However, I am not entirely convinced that this differential alignment based on architectural differences constitutes a significant advantage in itself, and it prompts a deeper question about the interpretation: why should a structural difference between models necessitate or imply a specific functional segregation in their representation or their relationship to brain function in the manner observed? A more explicit discussion of the behavioral or functional consequences tied to these form-based differences in model-brain alignment could strengthen the argument for the significance of this novel analytical approach.

Our response #3.4

Our axis-dependent metric is intentionally *stricter* than standard content metrics (linear predictivity, RSA, CKA etc.). Whereas those methods are blind to rotations, ours demands a one-to-one (or soft) match of tuning functions and therefore tests both **content** + **form** simultaneously. Two practical consequences follow.

Sharper model discrimination: Figure 6 shows one example: CNNs and ViTs are indistinguishable scores under linear predictivity, yet differ markedly under our form-based test. Thus the new metric resolves models that content-only tools treat as equivalent.

Mechanistic clues: Our theory and empirical results point to symmetry-breaking non-linearities as the driver of privileged axes. CNN layers end in a hard ReLU and training under this constraint often yields more monosemantic filters (cf. Bau et al., 2020) and, in our data, stronger axis-level alignment to cortex. In contrast, ViT architectures employ different nonlinearities (GeLU in MLP blocks and linear operations in attention layers), exhibit greater polysemanticity than CNNs (indicating fundamentally different tuning functions) and show poor form-alignment to cortex even though purely content-sensitive metrics rate CNNs and ViTs as equally “brain-like”. Thus the CNN $\gg$ ViT gap is not idle architectural plumbing but a concrete biophysical hint: cortical circuitry may employ ReLU-like operations (or any non-linearity that imposes a similar positivity constraint), a claim that content-only comparisons cannot address. We deliberately hedge: the present data cannot prove which non-linearity the brain uses, only that models with ReLU-style symmetry breaking align better in form. Nonetheless, this is the first empirical leverage we know of for adjudicating between architectures on the basis of their representational form.

We have modified the discussion section in the paper to reflect this point. It now reads:

“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 also⁶), 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.”

Finally, our paper links representational form to computational consequences that are invisible to content metrics: privileged axes observed in brains and networks lead to sparser codes, cheaper read-outs, and superior few-shot generalization (Fig. 5). These pay-offs motivate the functional relevance of privileged axes, and reinforce the practical value of analyzing form, not just content.

Reviewer #3, comment #5

A relevant consideration in the context of this work is the extensive literature on hyper-alignment and related techniques aimed at aligning representational spaces across different individuals or datasets. While the differences between the methods employed in this paper and hyper-alignment approaches can be inferred, explicitly articulating these distinctions within the text would significantly enhance clarity for readers familiar with that body of work, helping to precisely position the contribution of the novel axis-dependent measures and the framework for assessing privileged axes.

Our response #3.5

Thank you for pointing this out. The hyperalignment literature (e.g. Haxby et al. 2011, Guntupalli et al. 2016) focuses on finding the optimal rotation (or Procrustes) transform that brings multiple subjects’ response spaces into alignment with each other. By design, those methods seek the rotation that maximizes similarity and are therefore blind to which axes underlie that similarity. In contrast, our alignment methods forbid arbitrary rotations. Instead, we align representations using only voxel-level reassignments (or soft-permutations) and then systematically rotate each subject’s native axes to measure how quickly alignment breaks down. This allows us to reveal—and quantify—the extent to which particular representational axes are “privileged” (i.e. preserved across individuals and across biological vs. artificial networks). We now include this connection in the Discussion, which reads:

“Prior hyperalignment methods learn the optimal rotation to match different subjects’ data—thereby discounting all rotational degrees of freedom and essentially erasing any information carried by the original axes [hasb2011common](#), [guntupalli2016model](#). These approaches thus remain agnostic to how strongly the native axes themselves drive that match. 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 inter-subject correspondence, revealing the truly privileged nature of the brain’s representational axes.”

References

- [1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.
- [2] Seungdae Baek, Min Song, Jaeson Jang, Gwangsu Kim, and Se-Bum Paik. Face detection in untrained deep neural networks. *Nature communications*, 12(1):7328, 2021.
- [3] Jeffrey S Bowers, Ivan I Vankov, Markus F Damian, and Colin J Davis. Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological review*, 121(2):248, 2014.
- [4] Mario Lezcano Casado. Trivializations for gradient-based optimization on manifolds. In *Neural Information Processing Systems*, 2019.
- [5] Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D Cox, and James DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. *bioRxiv*, pages 2022–07, 2022.
- [6] Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel LK Yamins, and Kalanit Grill-Spector. A single computational objective drives specialization of streams in visual cortex. *bioRxiv*, pages 2023–12, 2023.
- [7] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.
- [8] Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622, 2016.
- [9] Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 326–341. PMLR, 2024.
- [10] Gwangsu Kim, Dong-Kyum Kim, and Hawoong Jeong. Spontaneous emergence of rudimentary music detectors in deep neural networks. *Nature Communications*, 15(1):148, 2024.
- [11] Ammar Marvi, Nancy Kanwisher, and Meenakshi Khosla. Sparse components distinguish visual pathways their alignment to neural networks. In *The Fourteenth International Conference on Learning Representations*, 2025.
- [12] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.
- [13] Mike Page. Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23(4):443–467, 2000.
- [14] Jacob S Prince, George A Alvarez, and Talia Konkle. A contrastive coding account of category selectivity in the ventral visual stream. *bioRxiv*, pages 2023–08, 2023.
- [15] Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.
- [16] Daniel Yamins, Michael Cohen, Ha Hong, Nancy Kanwisher, and James DiCarlo. The emergence of face-selective units in a model that has never seen a face. *Journal of vision*, 15(12):754–754, 2015.

We thank the reviewers for their careful reading of the revised manuscript and for their constructive feedback. We are grateful that Reviewers 1 and 3 found the revisions satisfactory, and that Reviewer 2 found the manuscript substantially improved and ready for publication.

No further scientific changes were requested by the reviewers. We have therefore not made additional substantive revisions in response to these comments, aside from editorial and formatting changes requested by the journal.