IMPROVING GEOMETRIC CONSISTENCY OF NEURONAL FUNCTIONAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks provide a powerful framework for learning functional embeddings of neurons from their responses to visual stimuli. These embeddings can reveal structure in neuronal function, but biological interpretations require their structure to be reproducible across model fits. We identify three sources of inconsistency in data-driven core-readout predictive models for mouse visual cortex. First, the functional basis of the embeddings spanned by the core is not orthonormal, leading to distorted Euclidean distances when computed naively. Second, learnt receptive-field locations vary across model fits. Third, the image encoders (cores) learn different functional subspaces across different model initializations. Here we show that the first issue can be resolved by whitening the functional basis. Whitening corrects the geometric distortion, improving the consistency of local neighborhoods and global structure across models trained with different random weight initializations. The other two issues are mitigated by introducing a new Retinotopic Factorized Linear readout (ReFL). ReFL improves the consistency of readout locations and makes core functional decomposition more similar. Combining whitening with the ReFL readout yields the most consistent functional embeddings. These results highlight the importance of both the orthonormality of the functional basis and the readout mechanism for the reproducibility of learned neuronal embeddings.

1 INTRODUCTION

A central question in neuroscience is whether neurons can be grouped into distinct cell types or whether they exist along a continuum (Zeng, 2022). One approach to this question examines neurons' anatomical and electrophysiological properties, with recent evidence suggesting a largely continuous organization in the mouse brain (Scala et al., 2019; Weis et al., 2025). An alternative approach focuses on neuronal function, characterizing how neurons map high-dimensional visual stimuli to their responses (Klindt et al., 2017; Ustyuzhaninov et al., 2022). Traditionally, neuronal function has been described using a limited set of classical functional parameters, such as orientation selectivity, spatial frequency preference, and phase invariance (Awang, 1962; DeAngelis et al., 1994). However, recent work shows that even in early visual areas, such parameters are insufficient to fully characterize neuronal function (Walker et al., 2019; Fu et al., 2024; Li et al., 2025).

Over the last decade, deep neural networks have emerged as powerful models for predicting neuronal responses in the visual cortex to arbitrary visual stimuli, providing various frameworks for studying the computations underlying biological vision. Several model classes learn per-neuron functional vectors either as model weights (Klindt et al., 2017; Mi et al., 2023; Willeke et al., 2022; 2026; Azabou et al., 2023; Zhang et al., 2025) or as input-dependent embeddings (Arora et al., 2025; Yu et al., 2025; Wu et al., 2025). These representations have been shown to encode information about brain areas (Azabou et al., 2025; Zhang et al., 2025; Arora et al., 2025) and known cell types (Mi et al., 2023; Yu et al., 2025; Arora et al., 2025; Wu et al., 2025; Nellen et al., 2025; Akhmetali et al., 2026). Additionally they could potentially support data-driven discovery of new cell subtypes (Nellen et al., 2025; Akhmetali et al., 2026) as well as the analysis of neuronal functional similarity (Ustyuzhaninov et al., 2022).

The reproducibility of these embeddings across model fits remains unclear, but reproducible embeddings are essential for making biological claims. Few works have studied the stability of learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

neuronal representations (Turishcheva et al., 2024a; Nellen et al., 2025; Akhmetali et al., 2026), focusing on clustering consistency across models fitted with different seeds. These studies use the core–readout framework to approximate neuronal input–output functions (Fig. 3). The core computes nonlinear basis functions from visual input, while the per-neuron readout selects a location and uses the neuron's weight vector to linearly combine the basis functions at that position to predict individual neural activity (Klindt et al., 2017). The readout weight vector represents coordinates in the function space spanned by the core outputs. Hence, we can interpret it as a neuron's functional embedding. Prior reproducibility analyses only focused on the functional embeddings. We extend this work by analyzing both functional embedding consistency and readout location stability.

We show that current state-of-the-art Gaussian readout (Lurz et al., 2021) learn differently shifted readout locations across seeds, despite the underlying receptive fields remaining unchanged. Motivated by this, we introduce ReFL, a Retinotopic Factorized Linear readout that learns centered receptive fields without explicit guidance to do so. Additionally, we show that the core's functional decomposition of models trained with the Gaussian readout varies across seeds, reducing functional embedding consistency. ReFL improves this consistency as well. We further find that functional readout vectors are represented in a non-orthonormal basis, whose basis functions are correlated and have different scales, which distorts distances and local neighborhood relations. To address this, we estimate the covariance matrix of the core channels across encoded stimuli, whiten the space, and transform the functional vectors into the resulting orthonormal coordinate frame, enabling correct distance-based comparisons.

To summarize, our contributions are:

- We evaluate both global and local embedding consistency across model fits, motivated by evidence that distinct clusters may not exist (Nellen et al., 2025; Weis et al., 2025) and that clustering can be unstable (Turishcheva et al., 2024a).
- We identify that neuronal functional embeddings lie in a non-orthonormal space, motivating whitening before computing distances which improves embedding consistency.
- We introduce Retinotopic Factorized Linear readout architecture – ReFL – which improves functional embedding consistency globally and locally over the Gaussian readout.
- We show that the core functional decomposition is more consistent across seeds with ReFL than with Gaussian readouts, which leads to improved readout embedding consistency.

2 Background and Related work

Predictive models for visual cortex. Data-driven population models pioneered by Antolík et al. (2016); Batty et al. (2016); McIntosh et al. (2016) introduced the core-readout framework, in which a shared nonlinear core extracts features from visual stimuli, while each neuron's readout consists of a receptive-field location and a weight vector that combines the core features at that location to predict its activity. The core therefore defines a shared functional basis by mapping visual stimuli to nonlinear basis functions, while the readout weights provide the coordinates of each neuron's function in this basis. The resulting readout vector can therefore be interpreted as a dense representation of a neuron's function. Early models were extended by including behavioral modulation (Sinz et al., 2018; Franke et al., 2022), accounting for the eye perspective transformation (Wang et al., 2025), extended to video data (Sinz et al., 2018; Wang et al., 2025; Turishcheva et al., 2024b; Höfling et al., 2024) and conditioning on other responses (Bashiri et al., 2021; Schmidt et al., 2024). While several stimuli-driven deep predictive networks now try to improve model results via scaling (Wang et al., 2025; Willeke et al., 2026), another perspective is to make models as compact as possible for better interpretability while preserving state-of-the-art performance (Du et al., 2025; Cowley et al., 2026).

Several recent works (Ustyuzhaninov et al., 2022; Turishcheva et al., 2024a; Nellen et al., 2025; Akhmetali et al., 2026) tried to use the readout functional embeddings to define cell types, however, most of them leaned towards the lack of discrete functional cell types among excitatory cells in mouse V1. Other work tried to cluster optimal stimuli derived from these models on monkey V4 (Willeke et al., 2023) and mouse V1 (Tong et al., 2023) ultimately reaching the same conclusion.

Features orthonormality and whitening. Whitening is a classical linear-algebraic transformation that removes second-order dependencies and normalizes feature scales (Hyvärinen & Oja, 2000). It is often used to obtain more compact and well-conditioned representations, for example as a

preprocessing step for independent component analysis (Hyvärinen & Oja, 2000). Given centered features $\mathbf{x}$ with covariance $\mathbf{\Sigma}$, whitening applies $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{x}$, yielding $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}$. The resulting features are therefore uncorrelated and unit-variance, forming an orthogonal representation.

Recently, whitening neural activations has been used in contrastive self-supervised learning to prevent representation collapse and improve optimization and generalization (Zbontar et al., 2021; Ermolov et al., 2021; Bardes et al., 2022) and in supervised learning (Cogswell et al., 2016; Laakom et al., 2023; Zhu et al., 2025; Desjardins et al., 2015; Luo, 2017; Huang et al., 2018; 2019; 2021; Qiao et al., 2019; Belmekki et al., 2024) to improve generalization as well. For a more detailed overview please see App. D. Note that most prior works whiten *activations during training* at the batch level whereas we whiten the *weights post-hoc* using the whole dataset, since our preliminary experiments with during training whitening were not successful.

Consistency metrics. To make biological claims from neural network embeddings, we must ensure that the learned information is robust to stochasticity, a known issue in deep neural networks (Summers & Dinneen, 2021). We assess this by comparing neuronal representations across model fits differing only in random seed, while accounting for possible differences in coordinate frames across seeds (Moschella et al., 2023). Prior work (Turishcheva et al., 2024a; Nellen et al., 2025; Akhmetali et al., 2026) focused on clustering and therefore used Adjusted Rand Index (ARI) (Hubert & Arabie, 1985), which measures the agreement between two clusterings by calculating how often pairs of points are assigned to the same cluster—or separated into different clusters—in both groupings, adjusted by chance. This analysis implies discrete functional types, inconsistent with growing evidence for continuous neuronal organization (Weis et al., 2025; Turishcheva et al., 2024a; Nellen et al., 2025; Tong et al., 2023). We therefore additionally use standard representational alignment metrics: Centered Kernel Alignment (CKA) (Kornblith et al., 2019) and Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). Linear CKA is the normalized Frobenius inner product between the centered Gram matrices of two representation spaces, while RSA is the Spearman rank correlation between pairwise distances from different spaces. Both are invariant to orthogonal rotations, reflections, translation, and uniform scaling; RSA is additionally invariant to monotonic distance warps. However, CKA and RSA describe global geometric similarity as a single number and may miss fine-grained local changes. Recent work therefore uses local neighborhood metrics, such as mutual-kNN consistency (Huh et al., 2024; Gröger et al., 2026), which assess local alignment without imposing discrete boundaries between groups. In this paper, we use chance-adjusted mutual-kNN consistency unless stated otherwise. For formal definitions see App. T.2.

Ignoring the coordinate basis distorts local geometry. Before analyzing model embeddings, we use a toy example to show why computing distances in embedding space requires care. Pairwise Euclidean distances are used in the clustering algorithms and k-nearest-neighbor analyses, which quantify the consistency of neuronal embeddings across model seeds. However, the standard Euclidean distance formula is only valid in an orthonormal basis. When embeddings are represented in a non-orthonormal basis, computing Euclidean distance naively distorts the computed distances and, consequently, local neighborhood relationships. Importantly, the underlying distances between the points themselves do not change.

To illustrate the local neighborhood distortion, consider a toy example of 400 points randomly sampled from a two-dimensional isotropic Gaussian distribution (Fig. 1A), represented either in a non-orthonormal basis B (red) or a rotated orthonormal basis Q (blue). In this example, each point would be a neuronal embedding and B, Q would be the sets of basis functions given by the output of two cores. The orthonormal transformation only rotates the space and preserves pair-

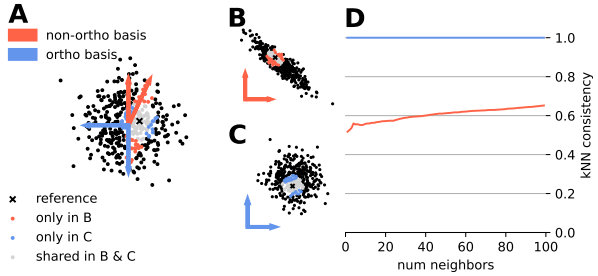


Figure 1: Non-orthonormal basis distorts nearest neighbors (NN), when distance computation does not account for it. **A:** Original data. 400 data points and two chosen bases. Colored points: 100 NN using standard Euclidean distances in the respective basis (**B** and **C**). **B–C:** Same data represented in both bases. **D:** kNN consistency of original and transformed data.

wise distances and thus fully keeping the k nearest neighbors of all points (100% k-NN consistency for all k; Fig. 1D, blue). In contrast, the non-orthonormal transformation shears and stretches the space (Fig. 1B), distorting distances and local neighborhoods and causing a severe drop in k-NN consistency between points in original space and points in B's embedding space (Fig. 1D, red). Thus, distance-based metrics require knowing the coordinate basis.

3 METHOD

Whitening the functional basis is needed to preserve local structure of the space. Following earlier work (Klindt et al., 2017; Lurz et al., 2021), we use a predictive model that maps stimuli to neuronal responses using the core–readout architecture. The core implemented as a CNN maps each image patch $x \in \mathcal{X} \subset \mathbb{R}^{P\times P}$ to a vector of C channels (Fig. 3). At a fixed readout location, the core output for each channel defines a function $f_i : \mathcal{X} \to \mathbb{R}, \quad i = 1, \ldots, C$, mapping an image patch $x \in \mathcal{X}$ to a scalar. Combining the core channel functions yields the feature vector $\mathbf{f}(x) = (f_1(x), \ldots, f_C(x))^\top$. To map an image patch x to the n'th neurons' pre-activation response $\hat{y}_n(x)$ the readout combines these feature values linearly using a learned weight vector $\mathbf{r}_n$: $\hat{y}_n(x) = \mathbf{f}(x)^\top \mathbf{r}_n$. Thus, each neuron's function is a linear combination of the core's feature channels. In other words, the core channels span a functional space and neurons learn coordinates in this shared basis to describe their response function. The readout weight vectors are these coordinates. Note that the actual neuronal responses have to be positive, so in the model the pre-activation response $\hat{y}_n(x)$ is passed through a non-linearity, specifically $ELU(\hat{y}_n(x)+b_n)+1$, where b_n is a per-neuron learnt bias.

To calculate meaningful Euclidean distances between readout vectors we need to ensure the orthonormality of their basis (Fig 1). Here the basis functions are the core channels $\{f_i\}_{i=1}^C$. To determine whether these functions are orthogonal, we define their inner product with respect to the natural image patch distribution as

$$\begin{aligned}\langle f_i, f_j\rangle_p &= \int_{\mathcal{X}} f_i(x)\, f_j(x)\, p(x)\, dx \\ &= \mathbb{E}_{x\sim p}[f_i(x)\, f_j(x)],\end{aligned}$$

where $p(x) : \mathcal{X} \to \mathbb{R}$ is the probability density function of the natural image patch distribution.

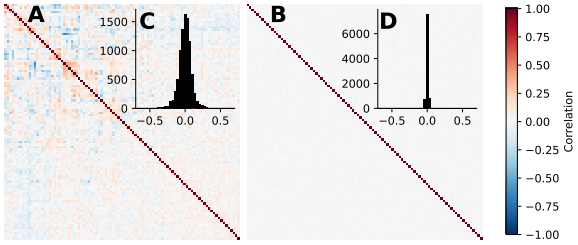


Figure 2: Core channel correlation matrix of a fully trained baseline model using the Gaussian readout (**A**) before and (**B**) after whitening. Whitening transform was computed on training data and visualization is on validation data. **C-D:** Histogram of off diagonal correlation values.

To estimate if the basis is indeed orthogonal, we mean-center the channel functions $\tilde{f}_i(x) = f_i(x) - \mathbb{E}_{x\sim p}[f_i(x)]$, $\tilde{\mathbf{f}}(x) = (\tilde{f}_1(x), \ldots, \tilde{f}_C(x))^\top$ such that the inner products form a covariance matrix $\Sigma_{ij} = \langle \tilde{f}_i, \tilde{f}_j\rangle_p$. If $\Sigma = I$, the channel functions are pairwise orthogonal (uncorrelated) and have unit norm (unit variance). Orthogonality yields linear independence and our core functions actually form a basis.

Since the image distribution $p(x)$ is unknown and thus the underlying covariance matrix of the core functions, we estimate it empirically from B image patches sampled from the training set $X = (x_1, \ldots, x_B)^\top \in \mathcal{X}^B$. (For details see App. T.3.1.) We encode the image patches with the core functions f_i and mean-center the resulting channels to obtain the corresponding feature matrix $\tilde{\mathbf{F}}(X) \in \mathbb{R}^{B\times C}$ which yields the sample covariance $\hat{\Sigma} = \frac{\tilde{\mathbf{F}}(X)^\top \tilde{\mathbf{F}}(X)}{B-1}$. As shown in Fig. 2A,C, $\hat{\Sigma}$ contains substantial off-diagonal correlations, demonstrating that the learned core channels are not orthogonal. We therefore whiten the centered features by applying the linear whitening transformation $\tilde{\mathbf{f}}_w(x) = \tilde{\mathbf{f}}(x)W$, where $W \in \mathbb{R}^{C\times C}$ is chosen by Cholesky decomposition such that $W^\top \hat{\Sigma} W = I_C$. This whitening transformation is just a change of basis of the core functions. To preserve the response predictions, the readout vectors must be transformed by the corresponding inverse basis transformation. Since whitening transformation is defined by using the covariance of the core features, we first show that mean-centering (needed for covariance) leaves the response

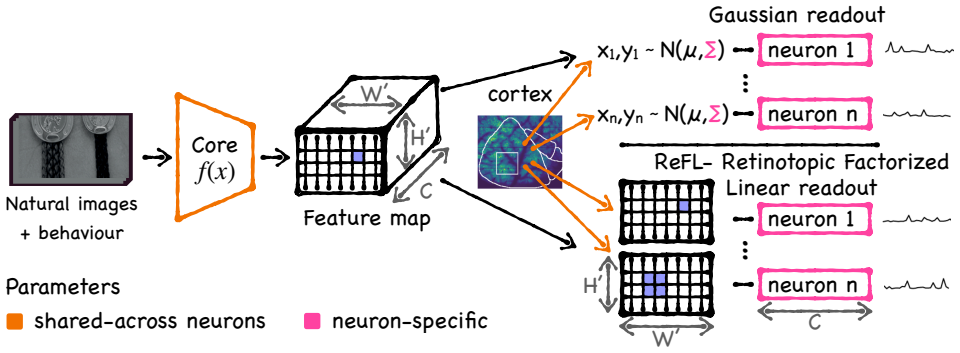


Figure 3: **Model architecture**: The *Core* $f(x)$ transforms image x into feature map of shape (H', W', C). The per-neuron readout converts the feature map into a scalar neuron response via two steps: spatial and functional. The spatial step collapses the feature map into a feature vector, which the functional step maps to a scalar via a dot product with a learned readout vector, interpreted as the neuron's functional embedding. The spatial step differs between the two readout architectures: The current state-of-the-art *Gaussian Readout* learns a 2d Gaussian distribution over the feature map whose mean specifies the neuron's spatial location. The *ReFL Readout* learns a spatial mask and does weighted pooling to collapse the spatial dimensions. Both use a retinotopic mapping meaning they learn shared-across-neurons functions to predict the spatial step.

predictions unchanged by absorbing the mean shift into the per-neuron bias. Let $\boldsymbol{\mu} = \mathbb{E}_{x\sim p}[\mathbf{f}(x)]$:

$$\hat{y}_n(x) + b_n = (\mathbf{f}(x) - \boldsymbol{\mu})^\top \mathbf{r}_n + \boldsymbol{\mu}^\top \mathbf{r}_n + b_n = \tilde{\mathbf{f}}(x)^\top \mathbf{r}_n + \underbrace{\boldsymbol{\mu}^\top \mathbf{r}_n + b_n}_{\tilde{b}_n}. \tag{1}$$

Second, whitening projects the core functions into a new basis, so the readout weights are transformed by the inverse basis transformation to preserve the response predictions

$$\hat{y}_n(x) = \tilde{\mathbf{f}}(x)^\top \mathbf{r}_n = \tilde{\mathbf{f}}(x)^\top W W^{-1} \mathbf{r}_n = \tilde{\mathbf{f}}_w(x)^\top \underbrace{W^{-1}\mathbf{r}_n}_{\tilde{\mathbf{r}}_n}. \tag{2}$$

For more details see App. T.3.

ReFL readout architecture. ReFL readout predicts a spatial mask of shape $H' \times W'$ for each neuron, similarly to the original Klindt et al. (2017) factorized readout. However, the original factorized readout learns a mask for every neuron separately which leads to many neuron-specific parameters. ReFL readout learns a shared mapping from a neurons' planar cortex location to its spatial mask. This means every parameter needed for the spatial step is shared across all neurons. This mapping is implemented as a linear layer ($\mathbb{R}^2 \to \mathbb{R}^{H'\times W'}$) followed by a Gaussian smoothing kernel and a softmax. The Gaussian smoothing prevents overfitting by limiting the model to smooth masks (for details see App. M). The softmax acts as normalization such that masks have only positive entries, which sum to one. ReFL readout also incorporates eye-tracking data through a shifter network that maps pupil position in screen coordinates as a shift of the readout location on the feature map. Unlike the original shifter (Sinz et al., 2018), which shifts a learned 2D position as in the Gaussian readout Lurz et al. (2021), ReFL readout shifts the feature map instead of every neuron's mask.

4 Results

ReFL readout and whitening improve the consistency of neuronal functional embeddings. The core maps visual input into a space of basis functions, while the readout learns a receptive field location and a functional vector (Fig. 3) for each neuron. To make biological claims about these functional vectors, we need them to be consistent across model fits. We start by asking whether whitening and our new readout improve the consistency of functional embeddings. The first measure of consistency we consider is the adjusted rand index (ARI) (Hubert & Arabie, 1985), which

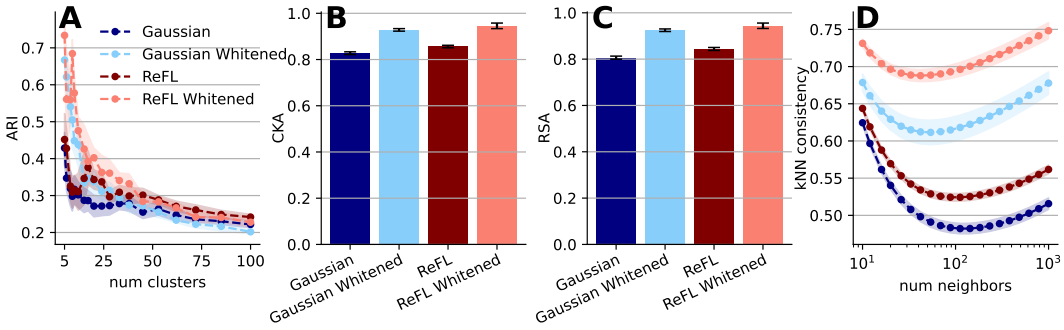


Figure 4: Readout vector consistency across seeds measured as (**A**) ARI, (**B**) Linear CKA, (**C**) RSA and (**D**) kNN consistency. **A**: Shaded areas: SD of three pairwise model comparisons, averaged across five k-means initializations. **B–D**: Shaded areas and error bars: SD of three pairwise model comparisons.

quantifies to what extent clustering the functional embeddings of different model fits leads to consistent cluster assignments (Turishcheva et al., 2024a; Nellen et al., 2025; Akhmetali et al., 2026). Following Turishcheva et al. (2024a), we use k-means clustering (Arthur et al., 2007). In agreement with prior studies (Turishcheva et al., 2024a; Nellen et al., 2025), ARI decreases monotonically with the number of clusters (Fig. 4A), indicating greater stability of coarse than fine-grained structure. Whitening improves ARI on a reasonable amount of clusters (≤ 40), while it has little effect at higher numbers. This is true for both readouts. Similarly, the ReFL readout improves over the Gaussian readout, irrespective of whether we whiten the basis or not. Combining both leads to the most consistent clustering up to 40 clusters across model refits, matching the Gaussian readout baseline afterwards.

Since ARI cannot distinguish clustering issues (lack of clusters, non-isotropic clusters) from global space inconsistencies, we tested two additional global stability metrics: Centered Kernel Alignment (CKA) (Kornblith et al., 2019) and Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). Whitening significantly increased both CKA (Fig. 4B) and RSA (Fig. 4C) indicating better global stability. Again, combining whitening and ReFL readout achieved the best results.

Both CKA and RSA describe global geometric similarity as a single number but may miss fine-grained local changes. To evaluate local relationships, we used chance-adjusted mutual kNN consistency (Huh et al., 2024; Gröger et al., 2026) (Fig. 4D), which measures what fraction of the k nearest neighbors (in embedding space) are, on average across neurons, shared between two model fits. The kNN consistency decreased from ten to hundred neighbors before increasing at larger scales. The early decrease in kNN consistency may be caused by neurons that were recorded multiple times across recording planes (Willeke et al., 2022). These 'duplicates' were treated as independent neurons with separate receptive fields and functional embeddings. Whitening improved kNN consistency for both Gaussian and ReFL readout, with the whitened ReFL readout significantly outperforming the whitened Gaussian.

Possible hypothesis for improved readout consistencies: The improved functional-embedding consistency with ReFL readout suggests that the two readouts learn representations differently despite similar performance. We consider two possible non-mutually exclusive explanations for it. First, the Gaussian readout could learn more variable readout locations for the same neurons across different model initial-

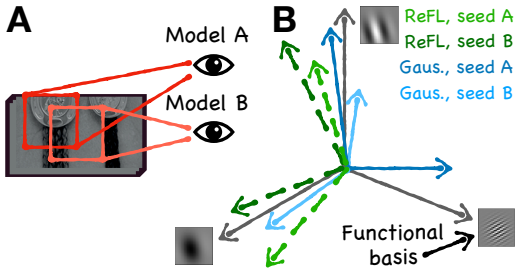


Figure 5: Two non-mutually exclusive explanations why ReFL functional embeddings are more consistent: **(A)** Gaussian readout could learn more different receptive fields for the same neurons across different model initializations; **(B)** the models with ReFL readout could learn more consistent functional decomposition in the core.

izations. If so, the same neurons will process different input patches across fits, leading to different learned functions (Fig. 5A). Second, ReFL's readout might change the learning dynamics and lead to more consistent core representations across fits, improving functional-embedding consistency (Fig. 5B). We test both of these hypothesis below.

Gaussian readout locations shift more across seeds than ReFL, but the core compensates by shifting its kernels, keeping neurons' receptive fields consistent across seeds. First we check how consistent the receptive field positions are. To do so, we analyze the readout locations learned by both the Gaussian and ReFL readout models. For the Gaussian readout, this is simply the (x, y) coordinate learned by the readout. The ReFL readout predicts a 2D heatmap, which we map to a single coordinate using the mask's center of mass $(\bar{x}, \bar{y})$ (App. I). We find that the Gaussian readout learns substantially shifted readout locations between model initializations 6A), whereas ReFL produces much more consistent readout locations (Fig. 6B). Additionally, the ReFL readout positions are centered (Fig. 6B), agreeing with the real-world population receptive field. However, consistent readout locations alone do not guarantee consistent receptive fields: the core can compensate for shifts in the readout by shifting the filters within its convolutional kernels. To test if this actually happens we compare RFs after correcting for core shifts. For each neuron we computed the gradient of its predicted response with respect to a mean (zero-valued) input image via backpropagation, yielding a linear approximation of the neuron's spatial receptive field. We then take the center-of-mass of this gradient to compute the neurons' RF, corrected for the core shifts (white crosses in Fig. 6C, D). Qualitatively, Fig. 6C shows a systematic shift in the learnt readout location (red dot and bounding boxes) between the seeds for the Gaussian readout, while the learnt and core-shift-corrected RFs align much better for the ReFL readout (Fig. 6D). However, the core-shift-corrected RFs (white crosses) from different seeds averaged across neurons (Fig. 6E) are equally consistent between both model types, so it seems like the core successfully compensates for the shifts in the Gaussian readout.

ReFL readout makes the core learn more similar functions. Next, we test how similar the functional spaces are across model initializations. To do so, we compare two models A and B with the same readout type but trained with different seeds. For each training batch, we randomly sample a hundred neurons and extract the C-dimensional core output at each neuron's center-of-mass position (Fig. 8A, step ①). This results in a set of C-dimensional core-feature vectors for model A and B for identical image-neuron pairs, while accounting for differences in the readout

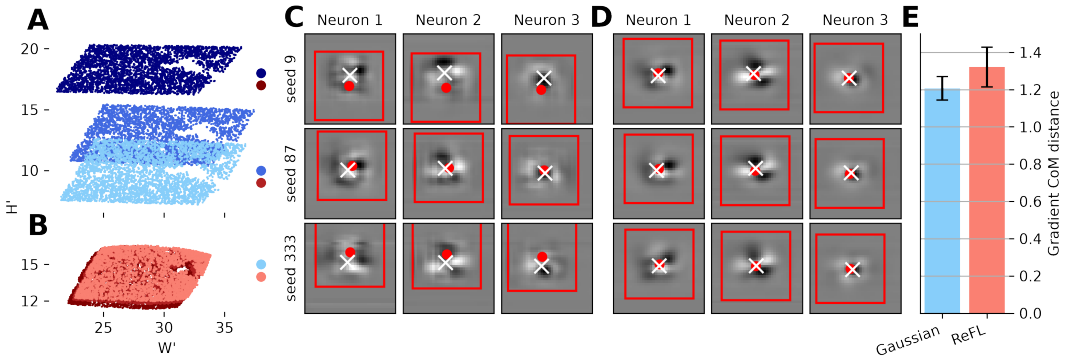


Figure 6: Gaussian readout locations shift more across seeds than ReFL, but the core compensates by shifting its kernels, keeping image receptive fields consistent across seeds. **A:** Mean Gaussian readout locations on the feature map. **B:** Center-of-mass (CoM) of masks learned by the ReFL readout. **A–B**: One point is one neuron. **C–D:** Per-neuron gradients with respect to zero-valued images that indicate the direction of input changes that maximize each neuron's predicted response for models using Gaussian (**C**) and ReFL (**D**) readouts. *White crosses*: Gradient's CoM shows the neurons' RF positions after correcting for core shifts. *Red box*: Image patch shown at the readout's location. *Red dot*: Learnt readout location. For ReFL readout we use the mask's CoM as the readout location. **E:** Average distance in pixels between the shift-corrected readout locations (white crosses) of the same neurons across seeds, (mean ± SD) across three pairwise seed comparisons.

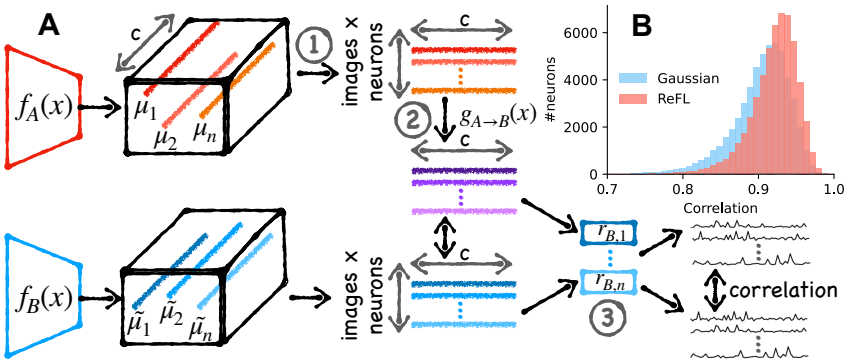


Figure 7: Core functional decomposition is more consistent across seeds with ReFL than with Gaussian readouts. **A**. To test whether the core functions of seeds A and B are similar, we test whether core A can substitute for core B. ① We extract core outputs at the center-of-mass (CoM) positions of the sampled neurons, using the same neurons across seeds and each model's own CoM positions. ② We train a linear regression $g_{A\to B}$ on the training set to map features from space A to space B. ③ On the validation set, we extract features at each neuron's readout location, map them from A to B via $g_{A\to B}$, and do the dot product with the functional embedding from B to obtain responses $\hat{y}_{A\to B}$. Without feature substitution, we obtain $\hat{y}_{B\to B}$. For both Gaussian and ReFL readouts, we use readout locations / masks from seed A to account for the core shifts for the models with the Gaussian readout. **B**. Histogram of correlations between original model predictions $\hat{y}_{B\to B}$ and core-replaced predictions $\hat{y}_{A\to B}$.

locations between the two models. On these pairs, we fit a linear regression model from A's feature space to B's, using the training split only (Fig. 8A, step ②).

On the validation split, we then replace the core features from model B with the mapped features from core A, while keeping B's readout fixed, yielding neuronal response predictions $y_{A\to B}$ (Fig. 8A, step ③). We quantify functional similarity as the per-neuron correlation between $y_{A\to B}$ and the predictions $y_{B\to B}$ of the original model B (Fig. 8B). Swapped cores trained with ReFL readout produce more correlated responses, suggesting more consistent core basis functions.

Note that this analysis is limited to linear mappings and may miss nonlinear correspondences between core A and B.

Functional readout vectors become less similar as cortical distance increases. To evaluate functional organization across the cortex, we compute the cosine similarity between the functional embeddings (readout vectors) of every pair of neurons and relate it to their physical cortical distance by averaging similarities within cortical distance bins. As shown in Fig. 8, functional similarity decays monotonically with increasing cortical distance, starting at a mean cosine similarity of ~ 0.23 for nearby neurons (< 50 μm) and dropping to slightly negative values at larger distances (> 700 μm).

Crucially, these weak functional-vs-cortical correlations are not simply induced by the use of retinotopy data (App. E for a control experiment) but may reflect biological organization, con-

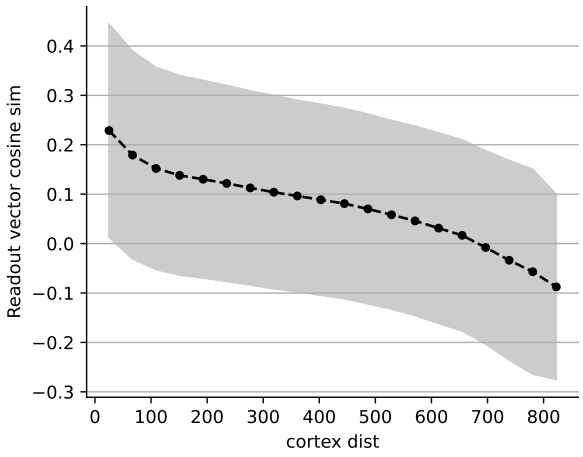


Figure 8: Cosine similarity between functional readout vectors vs the distance in the cortex for pairs of neurons from the same animal, averaged across seven animals. We only consider neurons from the top recording plane. Shaded area: SD across neurons.

sistent with the subtle functional clustering of nearby neurons in mouse V1 revealed by fine-scale imaging (Kondo et al., 2016; Ringach et al., 2016; Fahey et al., 2019) despite the overall "salt-and-pepper" organization for rodents (Ohki et al., 2005; Ohki & Reid, 2007).

5 DISCUSSION

In this work, we demonstrate that evaluating distances in non-orthonormal spaces distorts local neighborhood relationships leading to incorrect assessments of representational consistency. To resolve this, we estimate the functional basis covariance matrix from core output channels and perform post-hoc whitening to project functional vectors into an orthonormal coordinate frame. Furthermore, we introduce ReFL readout, a novel readout architecture that predicts per-neuron spatial masks directly from cortical coordinates via a shared retinotopic mapping. Evaluating both global and local consistency we show that ReFL readout together with post-hoc whitening substantially improves the consistency of functional neuronal embeddings. Additionally it stabilizes readout locations and improves reproducible learning of core functions.

Limitations and future work. First, we apply whitening post hoc rather than learning neuronal embeddings directly in an orthonormal basis the spectrum of the the core changes substantially during training making during training whitening challenging. We speculate that this reflects *simplicity bias* and *spectral bias*: SGD first learns simple, low-frequency features and progressively adds more complex components (Kalimeris et al., 2019; Zhang et al., 2026; Rahaman et al., 2019). Thus, early covariance estimates may be far from their final form or rank-deficient, making online whitening potentially harmful to representation consistency. Resolving this is important for combining whitening with DECEMber (Nellen et al., 2025), which ideally performs clustering in the correct orthonormal space. Second, Rahaman et al. (2019) showed that higher frequencies can become easier to learn as manifold complexity increases, raising the question of the optimal dimensionality of the functional space, which we fix here. While we focus on readout consistency, we show that some inconsistency comes from the core learning different functional decompositions across seeds. Guiding the core toward more similar decompositions could further improve neuronal embedding consistency. This may be particularly relevant because natural images contain relatively little high-frequency information (Burton & Moorhead, 1987). Future work may therefore investigate the core's functional decomposition and its consistency across seeds. Third, our whitening approach does not directly extend to models such as POYO (Azabou et al., 2023) or OmniMouse (Willeke et al., 2026), where neuronal embeddings are not clearly constrained by visual space. Fourth, while ReFL readout produces more consistent masks than the Gaussian readout, the underlying mechanism remains unclear. We speculate that this is due to easier optimization. With a Gaussian readout, shifting the center from the top-left to the bottom-right requires gradually moving the mass, as there are no gradients from the target position. In contrast, the ReFL readout can shift faster because it receives gradients from multiple spatial positions. Another possibility is that neurons benefit from an explicit bias to pool information across image patches directly in the mask readout, instead of learning it as a part of the core. The mask's size and shape can then encode functional properties. Although it is unclear why this inductive bias helps and it comes with violating the conceptual separation of 'what' and 'where'.

REFERENCES

Mels Akhmetali, Lisa Schmors, Na Zhou, Jacob Reimer, Katrin Franke, and Fabian H. Sinz. Pick your channel: Ultra-sparse readouts for recovering functional cell types, 2026. URL https://openreview.net/forum?id=eybUA13VG0.

Ján Antolík, Sonja B. Hofer, James A. Bednar, and Thomas D. Mrsic-Flogel. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLOS Computational Biology*, 12(6):1–22, 06 2016. doi: 10.1371/journal.pcbi.1004927. URL https://doi.org/10.1371/journal.pcbi.1004927.

Vinam Arora, Divyansha Lachi, Ian Knight, Mehdi Azabou, Blake Richards, Cole Hurwitz, Joshua H Siegle, and Eva L Dyer. Know thyself by knowing others: Learning neuron identity from population context. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=zt3RKc6VBp.

David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pp. 1027–1035, 2007.

Adriawan Awang. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 1962.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh B Nachimuthu, Michael Jacob Mendelson, Blake Aaron Richards, Matthew G Perich, Guillaume Lajoie, and Eva L Dyer. A unified, scalable framework for neural population decoding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sw2Y0sirtM.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IuU0wcO0mo.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, and Fabian Sinz. A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, 34:15801–15815, 2021.

Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan M. Litke, E. J. Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion cell responses. In *International Conference on Learning Representations*, 2016. URL https://api.semanticscholar.org/CorpusID:39002941.

Zakariae Belmekki, Jun Li, Patrick Reuter, David Antonio Gómez Jáuregui, and Karl Jenkins. Feature map similarity reduction in convolutional neural networks. *arXiv preprint arXiv:2411.03226*, 2024.

Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied optics*, 26(1):157–170, 1987.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations (ICLR)*, 2016.

Benjamin R Cowley, Patricia L Stan, Jonathan W Pillow, and Matthew A Smith. Compact deep neural network models of the visual cortex. *Nature*, 652(8111):947–954, 2026.

Adrián Csiszárik, Péter Kőrösi-Szabó, Akos Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. *Advances in Neural Information Processing Systems*, 34:5656–5668, 2021.

Gregory C DeAngelis, RALPH D Freeman, and IZUMI Ohzawa. Length and width tuning of neurons in the cat's primary visual cortex. *Journal of neurophysiology*, 71(1):347–374, 1994.

Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, et al. Natural neural networks. *Advances in neural information processing systems*, 28, 2015.

Fengtong Du, Miguel Angel Núñez-Ochoa, Marius Pachitariu, and Carsen Stringer. A simplified minimodel of visual cortical neurons. *Nature Communications*, 16(1):5724, 2025.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pp. 3015–3024. PMLR, 2021.

Paul G Fahey, Taliah Muhammad, Cameron Smith, Emmanouil Froudarakis, Erick Cobos, Jiakun Fu, Edgar Y Walker, Dimitri Yatsenko, Fabian H Sinz, Jacob Reimer, et al. A global map of orientation tuning in mouse visual cortex. *BioRXiv*, pp. 745323, 2019.

Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H Sinz, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, 2022.

Jiakun Fu, Paweł A Pierzchlewicz, Konstantin F Willeke, Mohammad Bashiri, Taliah Muhammad, Maria Diamantaki, Emmanouil Froudarakis, Kelli Restivo, Kayla Ponder, George H Denfield, et al. Heterogeneous orientation tuning in the primary visual cortex of mice diverges from gabor-like receptive fields in primates. *Cell reports*, 43(8), 2024.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uz0gAAYydl.

Yena Han, Tomaso A Poggio, and Brian Cheung. System identification of neural systems: If we got it right, would we know? In *International conference on machine learning*, pp. 12430–12444. PMLR, 2023.

Larissa Höfling, Klaudia P Szatko, Christian Behrens, Yuyao Deng, Yongrong Qiu, David Alexander Klindt, Zachary Jessen, Gregory W Schwartz, Matthias Bethge, Philipp Berens, et al. A chromatic feature detector in the retina signals visual context changes. *Elife*, 13:e86860, 2024.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 791–800, 2018.

Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4869–4878. IEEE, 2019.

Lei Huang, Yi Zhou, Li Liu, Fan Zhu, and Ling Shao. Group whitening: Balancing learning efficiency and representational capacity. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9507–9516. IEEE, 2021.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where". *Advances in neural information processing systems*, 30, 2017.

Satoru Kondo, Takashi Yoshida, and Kenichi Ohki. Mixed functional microarchitectures for orientation selectivity in the mouse primary visual cortex. *Nature communications*, 7(1):13210, 2016.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28648–28662, 2021.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, and Moncef Gabbouj. Wld-reg: A data-dependent within-layer diversity regularizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8421–8429, 2023.

Bryan M Li, Wolf De Wulf, Danai Katsanevaki, Arno Onken, and Nathalie L Rochefort. Movie-trained transformer reveals novel response properties to dynamic stimuli in mouse visual cortex. *bioRxiv*, pp. 2025–09, 2025.

Ping Luo. Learning deep architectures via generalized whitened neural networks. In *International Conference on Machine Learning*, pp. 2238–2246. PMLR, 2017.

Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A. Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S. Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021. URL https://openreview.net/forum?id=Tp7kI90Htd.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14334–14343, 2024.

Lane T. McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A. Baccus. Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1369–1377. Curran Associates, Inc., 2016.

Lu Mi, Trung Le, Tianxing He, Eli Shlizerman, and Uygar Sümbül. Learning time-invariant representations for individual neurons from population dynamics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EcN3l6Xmnx.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Nina S. Nellen, Polina Turishcheva, Michaela Vystrčilová, Shashwat Sridhar, Tim Gollisch, Andreas S. Tolias, and Alexander S. Ecker. Learning to cluster neuronal function. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Eufm2Jmjod.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KJNcAkY8tY4.

Kenichi Ohki and R Clay Reid. Specificity and randomness in the visual cortex. *Current opinion in neurobiology*, 17(4):401–407, 2007.

Kenichi Ohki, Sooyoung Chung, Yeang H Ch'ng, Prakash Kara, and R Clay Reid. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433 (7026):597–603, 2005.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Dario L Ringach, Patrick J Mineault, Elaine Tring, Nicholas D Olivas, Pablo Garcia-Junco-Clemente, and Joshua T Trachtenberg. Spatial clustering of tuning in mouse primary visual cortex. *Nature communications*, 7(1):12270, 2016.

Pau Rodríguez, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016.

Aninda Saha, Alina Bialkowski, and Sara Khalifa. Distilling representational similarity using centered kernel alignment (cka). In *BMVC 2022-33rd British Machine Vision Conference Proceedings*, pp. 1–12. British Machine Vision Association (BMVA), 2022.

Federico Scala, Dmitry Kobak, Shen Shan, Yves Bernaerts, Sophie Laturnus, Cathryn Rene Cadwell, Leonard Hartmanis, Emmanouil Froudarakis, Jesus Ramon Castro, Zheng Huan Tan, et al. Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. *Nature communications*, 10(1):4174, 2019.

Finn Schmidt, Suhas Shrinivasan, Polina Turishcheva, and Fabian Sinz. Modeling dynamic neural activity by combining naturalistic video stimuli and stimulus-independent latent factors. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. URL https://neurips.cc/virtual/2024/101471.

Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, 31, 2018.

Cecilia Summers and Michael J Dinneen. Nondeterminism and instability in neural network optimization. In *International Conference on Machine Learning*, pp. 9913–9922. PMLR, 2021.

Rudi Tong, Ronan da Silva, Dongyan Lin, Arna Ghosh, James Wilsenach, Erica Cianfarano, Pouya Bashivan, Blake Richards, and Stuart Trenholm. The feature landscape of visual cortex. *bioRxiv*, 2023. doi: 10.1101/2023.11.03.565500. URL https://www.biorxiv.org/content/early/2023/11/05/2023.11.03.565500.

Polina Turishcheva, Max Burg, Fabian H. Sinz, and Alexander Ecker. Reproducibility of predictive networks for mouse visual cortex, 2024a. URL https://arxiv.org/abs/2406.12625.

Polina Turishcheva, Paul G Fahey, Michaela Vystrčilová, Laura Hansel, Rachel Froebe, Kayla Ponder, Yongrong Qiu, Konstantin F Willeke, Mohammad Bashiri, Eric Wang, et al. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *ArXiv*, pp. arXiv–2305, 2024b.

Ivan Ustyuzhaninov, Max F Burg, Santiago A Cadena, Jiakun Fu, Taliah Muhammad, Kayla Ponder, Emmanouil Froudarakis, Zhiwei Ding, Matthias Bethge, Andreas S Tolias, et al. Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*, pp. 2022–02, 2022.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12): 2060–2065, 2019.

Eric Y Wang, Paul G Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, et al. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058):470–477, 2025.

Marissa A Weis, Stelios Papadopoulos, Laura Hansel, Timo Lüddecke, Brendan Celii, Paul G Fahey, Eric Y Wang, J Alexander Bae, Agnes L Bodor, Derrick Brittain, et al. An unsupervised map of excitatory neuron dendritic morphology in the mouse visual cortex. *Nature Communications*, 16 (1):3361, 2025.

KF Willeke, K Restivo, K Franke, AF Nix, SA Cadena, T Shinn, et al. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. biorxiv. 2023. *doi. org/10.1101/2023.05*, 12, 2023.

Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Pede, Max F. Burg, Christoph Blessing, Santiago A. Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The sensorium competition on predicting large-scale mouse primary visual cortex activity, 2022. URL https://arxiv.org/abs/2206.08666.

Konstantin Friedrich Willeke, Polina Turishcheva, Alex Gilbert, Goirik Chakrabarty, Hasan Atakan Bedel, Paul G. Fahey, Yongrong Qiu, Marissa A. Weis, Michaela Vystrčilová, Taliah Muhammad, Lydia Ntanavara, Rachel E Froebe, Kayla Ponder, Zheng Huan Tan, Emin Orhan, Erick Cobos, Sophia Sanborn, Katrin Franke, Fabian H. Sinz, Alexander S. Ecker, and Andreas S. Tolias. Omnimouse: Scaling properties of multi-modal, multi-task brain models on 150b neural tokens. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mEw4lhAn0F.

Wei Wu, Can Liao, zizhen Deng, Zhengrui Guo, and Jinzhuo Wang. Neuron platonic intrinsic representation from dynamics using contrastive learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vFanHFE4Qv.

Han Yu, Hanrui Lyu, YiXun Xu, Charlie Windolf, Eric Kenji Lee, Fan Yang, Andrew M Shelton, Olivier Winter, International Brain Laboratory, Eva L Dyer, Chandramouli Chandrasekaran, Nicholas A. Steinmetz, Liam Paninski, and Cole Lincoln Hurwitz. In vivo cell-type and brain region classification via multimodal contrastive learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=10JOlFIPjt.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022.

Yedi Zhang, Andrew Saxe, and Peter Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. In *International Conference on Learning Representations*, volume 2026, pp. 57507–57551, 2026.

Yizi Zhang, Yanchen Wang, Mehdi Azabou, Alexandre Andre, Zixuan Wang, Hanrui Lyu, International Brain Laboratory, Eva L Dyer, Liam Paninski, and Cole Lincoln Hurwitz. Neural encoding and decoding at scale. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vOdz3zhSCj.

Jiachen Zhu, Katrina Evtimova, Yubei Chen, Ravid Shwartz-Ziv, and Yann LeCun. Variance-covariance regularization improves representation learning, 2025. URL https://openreview.net/forum?id=jqff3wzkLT.

A REPRODUCIBILITY STATEMENT

We used open source and openly available SENSORIUM 2022 dataset (Willeke et al., 2022). We also publish all our code, environment setting and models hyperparameters upon publication. The models in this work require less than 10Gb of GPU RAM memory and less than 5 hours of training per model. Hence we believe all our work is reproducible via public resources, such as Google Colab.

B ETHICS STATEMENT

Our work does not include any human data, which could lead to potential ethics concerns. Our work does not involve any funding conflicts. We believe that our work focuses on fundamental scientific questions and does not pose more ethics concerns as any other AI research project.

C AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, edit a research paper to improve readability, specifically rewriting text drafts to make it more clear and concise and summarizing or analysing existing literature, specifically to identify more relevant literature and sourcing/searching for information. We also used generative AI to create or edit software code and implement parts of the methods, specifically generate small patches of code, write scripts, clean up the codebase. LLM generated code and text was always verified by humans. We have not used generative AI tools for independently creating research ideas, proposing or refining hypotheses, support qualitative and thematic data analysis and interpreting the results. We also have not used generative AI for help with develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs Additionally, we have not used generative AI tools for create or modify scientific figures or images, suggest experimental parameters, creation of artifacts, draft parts of a research paper, transcribe recordings of research material, discover research topics or identify gaps, brainstorming, format references, suggest a structure for a research paper, propose a title or keywords for a research paper. We have reviewed all AI-assisted work.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Not applicable for our work: generate synthetic data sets, clean and reformat dataset, assist with translation, formulate questions for surveys or interviews.

D RELATED WORKS - DURING TRAINING WHITENING

Recently, whitening neural activations has been used in contrastive SSL to prevent representation collapse and improve optimization and generalization. Barlow Twins (Zbontar et al., 2021) encourages the cross-correlation matrix to approach the identity, reducing correlations between feature dimensions, while W-MSE (Ermolov et al., 2021) uses Cholesky decomposition to whiten batch representations before computing MSE between positive pairs. Later, VICReg (Bardes et al., 2022) similarly reduces feature correlations by regularizing off-diagonal covariance terms, but enforces only a minimum variance rather than unit variance.

In supervised learning, the focus was more on feature independence (orthogonality) but not on unit norm. Cogswell et al. (2016) introduced DeCov regularization loss, which penalized of-diagonal covariance but have been criticized not to be scale-invariant. To address it Laakom et al. (2023) introduced WLD-Reg, a loss penalizing similarity across within layer activations without imposing unit-variance on axis. While DeCov captures only pairwise relationships, determinant-based WLD-Reg captures the global multidimensional structure of activations, encouraging greater diversity. Recently VCReg (Zhu et al., 2025) adapted VICReg's variance-covariance regularization (Bardes et al., 2022) and outperformed both DeCov and WLD-Reg, but still enforcing orthogonality rather than orthonormality. Alternatively, Desjardins et al. (2015); Luo (2017); Huang et al. (2018; 2019; 2021) all worked on extending BatchNorm to produce standardized and decorrelated features

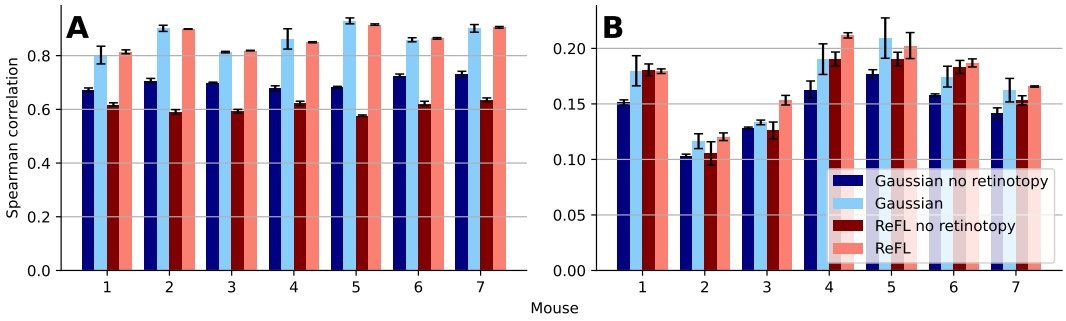


Figure 9: **Spearman correlation** between cortical distance and – **(A)** within seed distance between receptive-field center-of-masses or **(B)** within seed functional embeddings distances. We ensure that lower Spearman correlations without retinotopy are not due to worse fits by checking average test-set correlations across 5 of 7 mice with test data: Gaussian no retinotopy 0.387 ± 0.003, Gaussian 0.392 ± 0.002, ReFL no retinotopy 0.381 ± 0.001, ReFL 0.393 ± 0.001.

focusing on convergence speed rather that basis space for representations. Additionally, weight standardization (Qiao et al., 2019) standardizes weight vectors in convolutional layers instead of activations and OrthoReg (Rodríguez et al., 2016) regularizes CNN kernels toward orthogonality, though Belmekki et al. (2024) later showed that kernel orthogonality does not imply feature-map orthogonality in CNNs.

Note that most prior works whiten *activations during training* at the batch level whereas we whiten the *weights post-hoc* using core output activations to compute the Gram matrix.

E Coordinates interaction.

The improved functional consistency could result from conditioning the masks on cortical (x, y) positions. We therefore asked whether the functional embeddings inherit retinotopic structure by construction. To test this we computed Spearman correlations ρ between cortical distances and distances between receptive-field CoMs (Fig. 9A) or functional embeddings (Fig. 9B). We did it separately for each animal since cortical coordinate frames differ across animals. Cortical conditioning strongly anchors where neurons look in visual space ($\rho \approx 0.95$–1.0, Fig. 9A), but functional properties correlate only weakly with cortical proximity ($\rho \approx 0.12$–0.22, Fig. 9B), with no consistent difference between architectures. As a control, we repeated the analysis for Gaussian and ReFL readouts without retinotopy. Removing cortical positions, as expected, substantially reduced the correlation between learned RF and cortical distances, but it had little effect on the weak correlation between cortical distances and embedding distances. Thus, these weak functional-vs-cortical correlations are not simply induced by the use of retinotopy data but may reflect biological organization, consistent with the subtle functional clustering of nearby neurons in mouse V1 revealed by fine-scale imaging (Kondo et al., 2016; Ringach et al., 2016; Fahey et al., 2019) despite the overall "salt-and-pepper" organization (Ohki et al., 2005; Ohki & Reid, 2007).

F Retinotopic mapping visualization

The ReFL readout maps the 2d cortex position of a neuron to its spatial mask of shape $(x' \times y')$ via a linear layer $(2 \to x' \cdot y')$ followed by a gaussian smoothing kernel and a softmax over x' and y'. This means every pixel has a separate affine transformation consisting of 3 learned parameters: A bias and a weight for both cortex positions. To visualize these parameters we show the bias and both weights on a grid of shape $(x' \times y')$ so their positions correspond to the pixel they map to. We show the parameters after they were smoothed by the the models gaussian kernel.

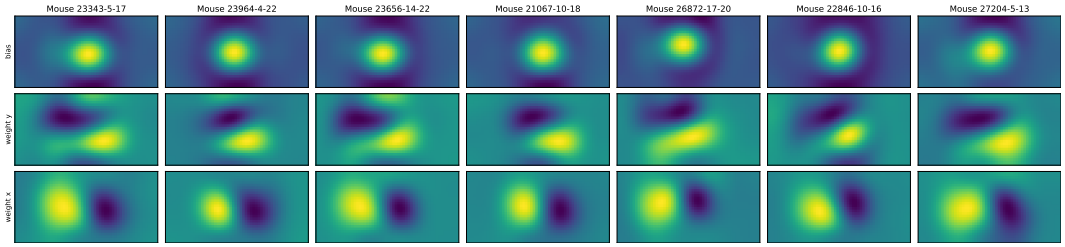


Figure 10: Visualization of the retinotopic mappings parameters. For all 7 animals we show the bias and both weights in a grid of shape $(x' \times y')$ corresponding to the feature map dimensions. The bias learns a "blob" while the weights learn "ramps" which move the bias blob based on cortex position.

G READOUT MASK EXAMPLES

We visualized 32 randomly selected RF masks learned by the ReFL readout along with their validation correlation values measuring how well the model predicts this neurons activity.

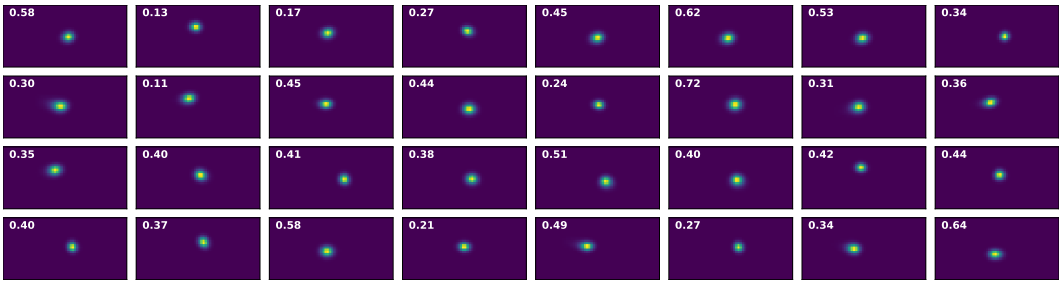


Figure 11: 32 randomly selected RF masks learned by the ReFL readout. White text is this neurons validation correlation.

H VISUALIZING LEARNING

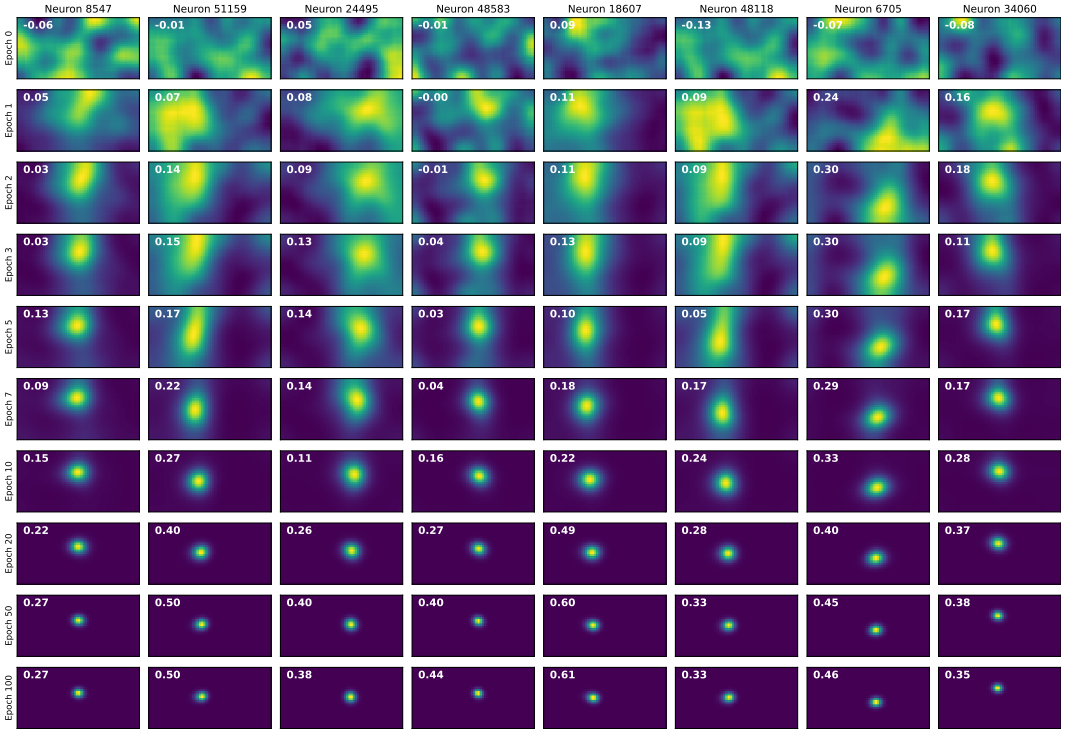


Figure 12: For 8 randomly selected neurons (columns) we visualize the RF masks over the course of training the ReFL readout where rows are epochs. The white text correspond to the validation correlation of the neurons at this epoch.

I CENTER-OF-MASS (COM) DISTANCE AS METRIC FOR RECEPTIVE FIELD CONSISTENCY.

We want to compare the consistency of predicted receptive fields across seeds and readout architectures. Since the Gaussian readout predicts a single readout location while ReFL readout predicts a 2D heatmap, we map ReFL readout to a single coordinate using the spatial center of mass $(\bar{x}_n, \bar{y}_n)$. Specifically, for each unit n we take a normalized 2D spatial heatmap $S_n \in \mathbb{R}^{H' \times W'}$ (where $\sum_{i,j} S_n(i,j) = 1$) and compute:

$$\bar{x}_n = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} j \cdot S_n(i,j), \quad \bar{y}_k = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} i \cdot S_n(i,j) \tag{3}$$

where i and j denote row (height) and column (width) indices, respectively.

J 7 MICE READOUT COMPARISON

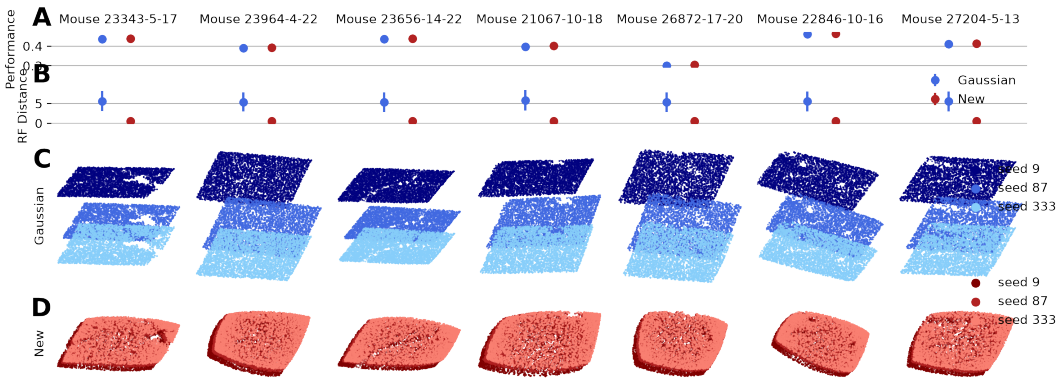


Figure 13: Gaussian readout and ReFL readout comparison on all 7 mice with 3 seeds each. **A:** Model performance. **B:** Average distance in pixels between seeds. **A-B:** Error bars mark the standard deviation across seeds. **C:** Mean locations on the feature map learned by the Gaussian readout. (One point is one neuron.) **D:** Center-of-mass of masks learned by the ReFL readout. (One point is one neuron)

C-D: Grids were mean centered to account for possible shifts in core functions.

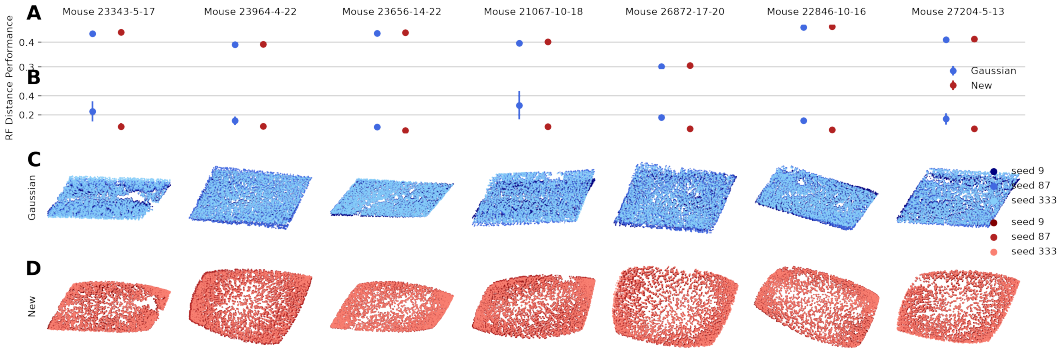


Figure 14: Same as plot above but grids were mean centered to account for possible core filter shifts.

K SHIFTER CONSISTENCY

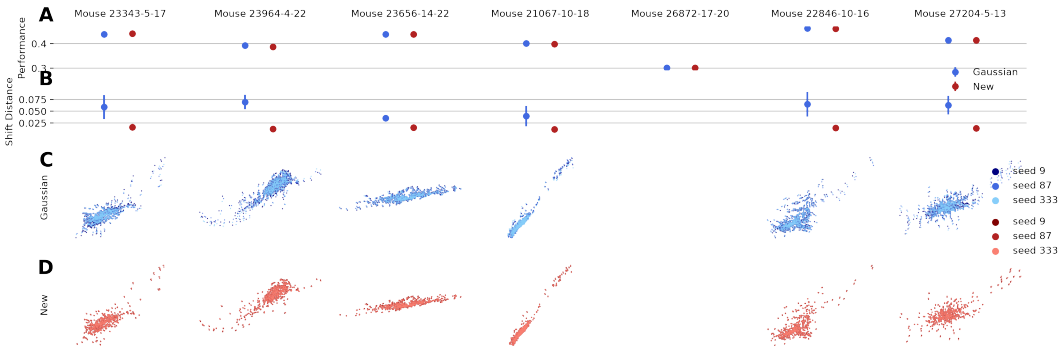


Figure 15: Shifter comparison for Gaussian readout and ReFL readout on all 7 mice across 3 seeds each. Mouse Nr. 5 has no behavioral data so also no shifter. **A:** Model performance. **B:** Average distance in pixels between shifts of different seeds. Error bars mark the standard deviation across seeds. **C-D:** Shift values (one per image) on the feature map learned by shifters of models with the Gaussian readout (**C**) and the ReFL readout (**D**). One point corresponds to one pupil position recorded while showing one image.

L FROZEN CORE EXPERIMENTS

To test the consistency of readouts given the same functional basis we took a pretrained model, froze the core and reinitialized the readout and shifter multiple times.

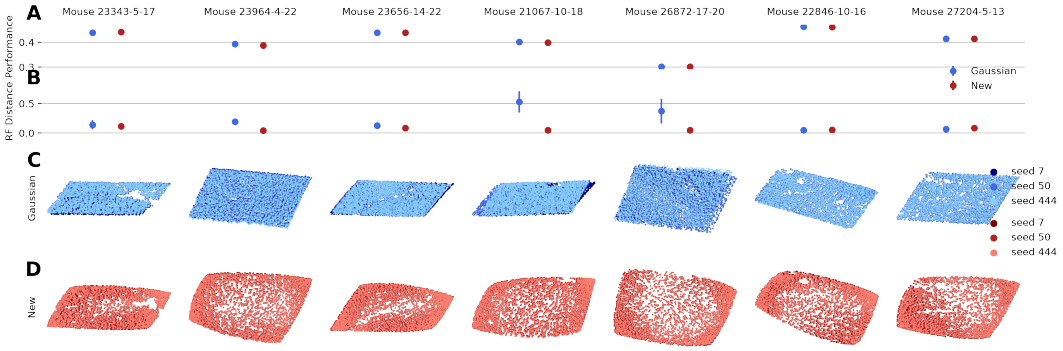


Figure 16: Gaussian readout and ReFL readout comparison on all 7 mice with 3 seeds each. Both the Gaussian and the ReFL readout models were initialized with a frozen core (one core for Gaussian and one for ReFL readout) and only the readout and shifter were retrained. **A:** Model performance. **B:** Average distance in pixels between seeds. **A-B:** Error bars mark the standard deviation across seeds. **C:** Mean locations on the feature map learned by the Gaussian readout. (One point is one neuron.) **D:** Center-of-mass of masks learned by the ReFL readout. (One point is one neuron). **C-D**: Grids were mean centered to account for possible shifts in core functions.

M SMOOTHNESS PARAMETER FOR REFL READOUT

We did a sweep for the smoothness parameter of the ReFL readout.

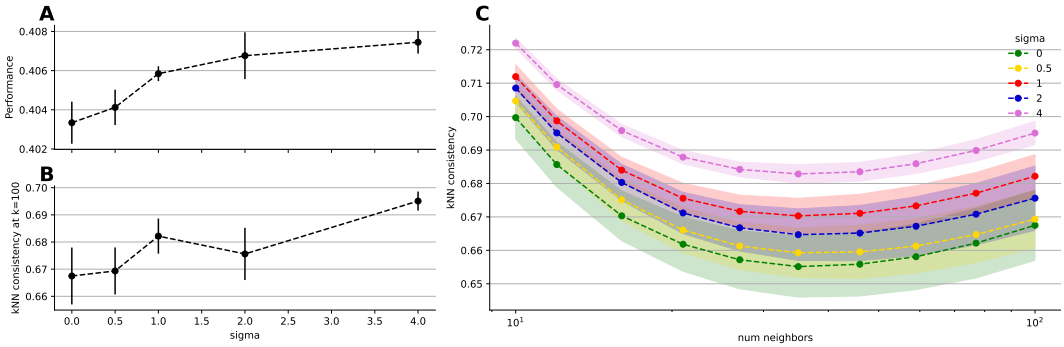


Figure 17: Comparison of different smoothness values (sigma) for the ReFL readout on predictive performance (**A**) and kNN consistency at k=100 (**B**). Errorbars are standard deviation across 3 seeds. **C**: Shows the the full kNN consistency curves with shaded areas showing the standard deviation across 3 seeds.

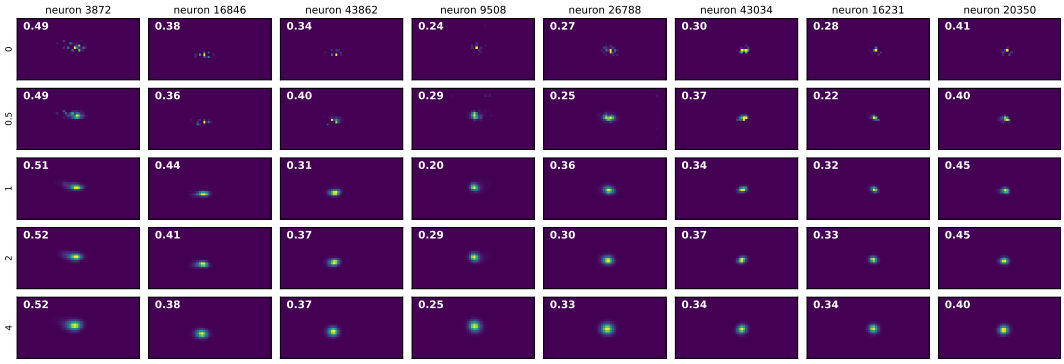


Figure 18: RF masks of 8 randomly selected neurons for different smoothness parameters. For increasing smoothness values the masks get visibly smoother.

N WHITENING DURING TRAINING DOES NOT YIELD BETTER RESULTS THAN POSTHOC WHITENING

Since we only apply the whitening procedure post hoc and we use adaptive L1 penalty on the readout vectors Turishcheva et al. (2024a) during training, we tried to incorporate the whitening into the training to regularize the readout vectors as if they live in the whitened space. For this we kept an exponential moving average (EMA) of the cores mean vector and covariance matrix. This was used to transform the readout vectors into the whitened space for computing the L1 penalty. We compare this to the post-hoc whitening approach where the L1 penalty is computed in the non-whitened space. Here we compare both L1 approaches with different regularization strengths and show that during-training whitening does not improve consistency or performance. All models used the ReFL readout.

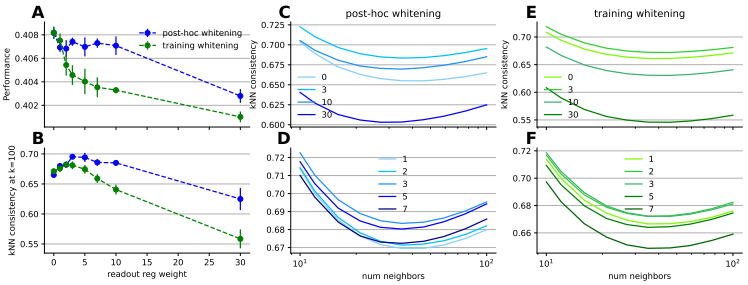


Figure 19: Comparison of different readout regularization strengths for the ReFL readout on predictive performance (**A**) and kNN consistency at k=100 (**B**). **C-F** show the full kNN curves for post-hoc whitening and training whitening.

O NORMALIZING READOUT VECTORS FOR COMPUTING METRICS

Normalizing readout vectors to unit euclidean norm leads to worse RSA and improves kNN consistency at low k.

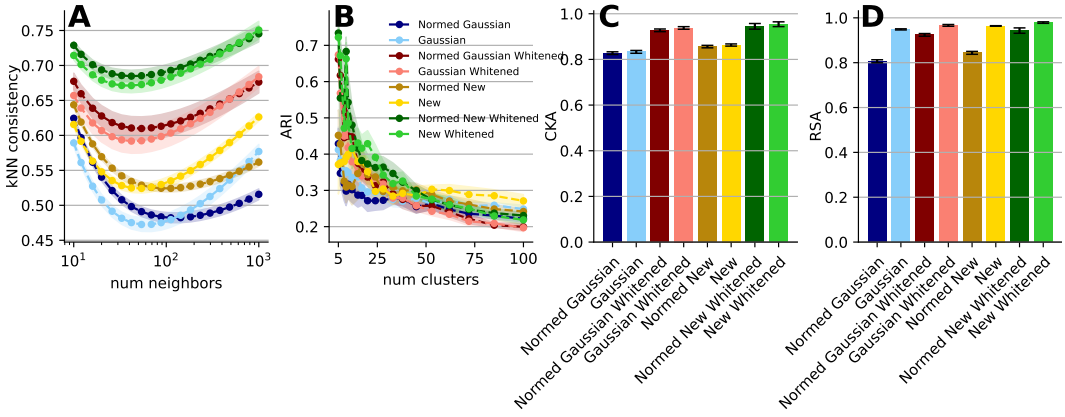


Figure 20: Comparison of normalizing readout vectors to unit euclidean norm vs not normalizing them. Readout vector consistency across seeds measured as kNN consistency (**A**), ARI (**B**), Linear CKA (**C**) and RSA (**D**).

P CHANCE ADJUSTING KNN CONSISTENCY

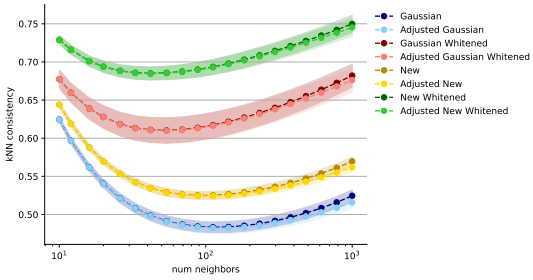


Figure 21: Comparison of chance adjusting kNN consistency (‘adjusted‘) vs no chance-adjustment. It only has a small effect for large number of neighbors.

Q ARI BASELINE

ARI measures clustering consistency. Its scores can be lower when the data lacks clearly separated clusters even if the datapoints itself are perfectly consistent, especially when using a non-deterministic algorithm such as k-means, as k-means would return slightly different partition depending on the initialization. To assess how much ARI scores are limited by the structure of our embeddings and k-means non-determinism, we compute ARI on the same embeddings but with 5 different KMeans initializations. We compare this to computing ARI across different embedding spaces (different trained models) to see if ARI scores are limited by the data structure or by changes in the structure across trained models.

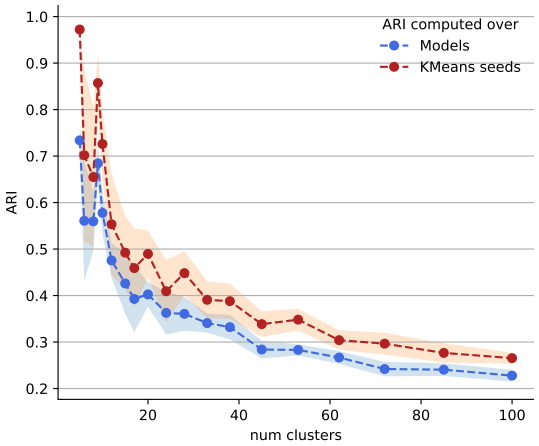


Figure 22: We have $3 \cdot 5$ sets of labels form 3 model embeddings combined with 5 different ARI initializations. We compare 2 approaches: 1. Computing ARI between labels of embeddings from different models and averaging over KMeans seeds, 2. Computing ARI between labels of different KMeans seeds and averaging over embeddings from different models. 1. acts as the upper bound of how far we can get because it isolates how high ARI gets if the embedding space is the same. 2. Shows how close we are to this upper bound when computing ARI between embeddings from different model initializations.

R RETINOTOPY IS NECESSARY

To test if the shared retinotopic mapping is necessary we compare to a model that replaces this mapping with neuron-specific mask parameters. The masks are still smoothed using the gaussian kernel and softmaxed exactly like the retinotopy model.

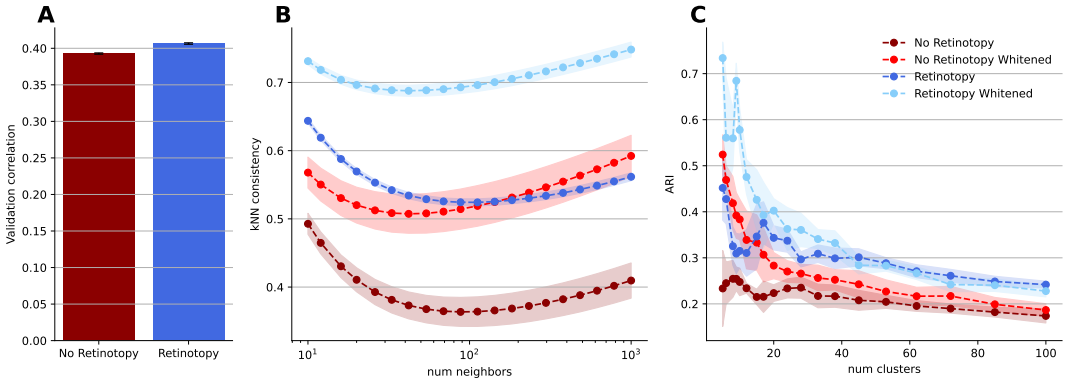


Figure 23: Comparison using a shared-across-neurons retinotopic mapping vs neuron-specific mask parameters on predictive performance (**A**), kNN consistency (**B**), and ARI (**C**).

S IMPROVING THE BASELINE MODEL

We compare the ReFL readout to a slightly modified version of the sensorium2022 baseline model Willeke et al. (2022) which uses the Gaussian readout. We replace two MLPS with linear layers and show that this leads to improved performance and higher consistency. The first one is the shared-across-neurons retinotopic mapping which is used to map from the neurons cortex location to the feature map location. It uses one hidden layer with 30 hidden units ($2 \rightarrow 30 \rightarrow 2$), an ELU in the hidden layer and a tanh in the final layer to restrict the output to between -1 and 1. We replace this with a linear layer ($2 \rightarrow 2$) followed by a tanh. The second component we simplify is the shifter which is responsible of mapping the recorded pupil position of the animal to a shift on the feature map which is applied to all neurons RF positions. It consist of 2 hidden layers with 5 hidden units each ($2 \rightarrow 5 \rightarrow 5 \rightarrow 2$) and tanh transfer functions in between. We replace this with a linear layer ($2 \rightarrow 2$) without a bias term, followed by a tanh. The reason this doesn't need a bias term is that it would create an ambiguity with the retinotopic mapping as the bias would be equally applied to all neurons. Please note that we could only reduce the grid mean predictor to a linear layer as Sensorium 2022 dataset (Willeke et al., 2022) contains only of V1 neurons. Hence, there is no 'sign flip' from higher visual areas.[1] Otherwise, we would need to stick to MLP, as sign flip requires the non-linear interaction between coordinates. The MLP grid mean predictor has been shown to capture the properties of retinotopy sign flips (Bashiri et al., 2021).

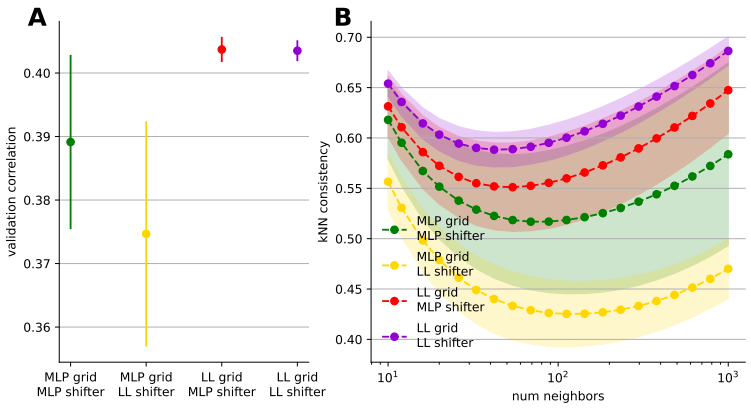


Figure 24: Comparison of baseline (Gaussain readout) models covering all combinations of MLP vs linear layer (LL) in both modules: retinotopy mapping (grid) and shifter. We trained on 3 seeds per architecture. **A**: Predictive performance. Errorbars show the standard deviation across seeds. A linear retinotopy mapping (grid) reduces variance and improves performance. **B**: kNN consistency measuring neighborhood consistency in readout vectors across models. Shaded areas show the standard deviation across comparisons.

[1] A sign flip is a directional reversal in how visual space is mapped onto the brain surface when moving from one cortical visual area to the next, e.g. if you move a light spot across the screen from left to right: the neural activity travels left to right across V1 but the moment it hits the border of a higher visual area, for example LM, the neural activity turns around and travels right to left across LM.

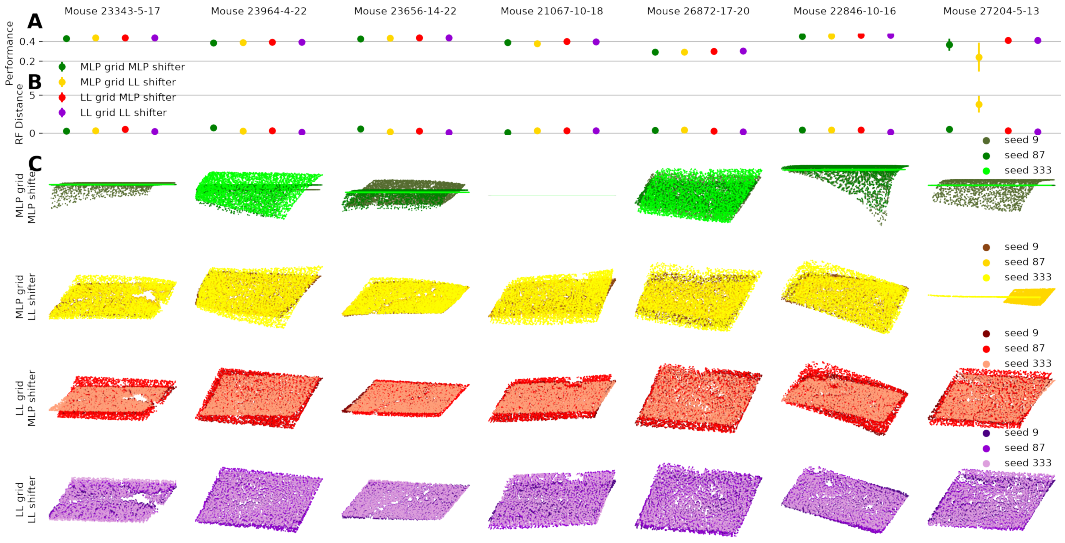


Figure 25: Comparison of baseline models covering all combinations of MLP vs linear layer (LL) in both modules: retinotopy mapping (grid) and shifter. **A**: Predictive performance on all 7 mice. **B**: Average distance in RFs across seeds in pixels. **C**: Mean locations on the feature map learned by the Gaussian readout. One point is one neuron. Grids were mean centered to account for possible shifts in core functions.

T EXPERIMENTAL DETAILS

T.1 DATA

We use data from the NeurIPS 2022 Sensorium Competition Willeke et al. (2022). The dataset comprises responses of primary visual cortex (V1) neurons in seven mice to grayscale natural images, recorded using two-photon calcium imaging. In addition to neuronal activity, behavioral variables were collected for all but one mouse, including running speed, pupil center position, pupil dilation, and the temporal derivative of pupil dilation.

T.2 EVALUATION.

Models functional performance. Consistent with prior work Willeke et al. (2022); Turishcheva et al. (2024a); Nellen et al. (2025), we assess predictive performance by computing the Pearson correlation between measured and predicted neural responses across images in the test set for 5 out of 7 mice , averaged across neurons. 2 mice were used for the competition purposes, hence, their test set is not publicly available.

Local functional embeddings consistency. We aim to quantify the consistency of functional embeddings locally without relying on an assumed cluster structure and to evaluate this consistency across multiple scales, capturing both local and more global organization.

Let $Y, Y' \in \mathbb{R}^{n \times d}$ be two representations of n neurons. For each neuron $i \in \{1, \ldots, n\}$, define:

$$N_Y(i) = \{j_1, \ldots, j_k\} \subset \{1, \ldots, n\} \setminus \{i\}$$

as the indices of the k nearest neighbors of neuron i in Y (similarly for $N_{Y'}(i)$).

$$\text{kNN consistency}(Y, Y') = \frac{1}{nk} \sum_{i=1}^{n} |N_Y(i) \cap N_{Y'}(i)|$$

where $|\cdot|$ denotes set cardinality.

kNN consistency measures, on average across all neurons, the fraction of k-nearest neighbors in Y that also appear among the k-nearest neighbors in Y'. In the paper we use kNN consistency and mutual kNN consistency interchangeably.

kNN consistency chance adjustment We want to adjust kNN consistency for chance such that zero means as good as chance under random neighbor assignment. This means we need to find the expected kNN consistency under random neighbor assignments for n objects and a given k.

Given two sets of random neighborhood assignments, treat the first set as fixed, then draw the neighbors in the second set. This is equivalent to drawing k objects from n objects without replacement, where k of the n objects are successes (they are in the k nearest neighbors in set 1). This is modeled by a hypergeometric distribution with population size $N = n - 1$, number of success states $K = k$, and number of draws $m = k$. The expected value of a hypergeometric distribution, and therefore number of shared neighbors is

$$\mathbb{E}[X] = \frac{mK}{N} = \frac{k^2}{n-1}.$$

Since kNN consistency is the number of shared neighbors normalized by k, the expected consistency under random assignments is $\mathbb{E}[X]/k = k/n$. We therefore chance-adjust the observed kNN consistency by subtracting this expected value and rescaling so that a perfect match remains 1:

$$\text{Adjusted kNN consistency} = \frac{\text{kNN consistency} - k/(n-1)}{1 - k/(n-1)}$$

Thus, random agreement corresponds to 0, while perfect agreement remains 1.

Global functional embeddings consistency - clustering. Clustering consistency is commonly used to assess structure in neural representations by measuring the agreement between different clustering solutions (Turishcheva et al., 2024a; Nellen et al., 2025). A standard metric for this purpose is the adjusted Rand index (ARI) (Hubert & Arabie, 1985), which quantifies the similarity between two cluster assignments while correcting for chance. Let $\mathbf{z}, \mathbf{z}' \in \{1, \dots, K\}^N$ denote two clusterings of N neurons into K clusters. The ARI is defined as

$$\text{ARI}(\mathbf{z}, \mathbf{z}') = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}}, \tag{4}$$

where n_{ij} denotes the number of neurons assigned to cluster i in $\mathbf{z}$ and cluster j in $\mathbf{z}'$, and a_i and b_j are the corresponding marginal cluster sizes. ARI equals 1 for identical clusterings and has expected value 0 under random assignments, negative values indicate less agreement than expected by chance.

Although ARI-based consistency measures are widely used to evaluate cluster structure, they are less informative when the underlying structure is continuous. Poor ARI scores can then be interpreted as either an effect of poor consistency or general lack of clustering structure.

Global spatial consistency without clustering - CKA and RSA.
Centered Kernel Alignment (CKA) (Kornblith et al., 2019) is a similarity metric used to compare representations (typically feature activations) extracted from different neural network layers, architectures, or training runs (Nguyen et al., 2021; Kornblith et al., 2021; Han et al., 2023; Maniparambil et al., 2024; Saha et al., 2022). Given two activation matrices $\mathbf{X} \in \mathbb{R}^{n \times p_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p_2}$ for n paired input samples, CKA computes the normalized scalar similarity between their centered kernel matrices $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^\top$ (assuming linear kernels):

$$\text{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\text{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\text{HSIC}(\mathbf{K}, \mathbf{K}) \cdot \text{HSIC}(\mathbf{L}, \mathbf{L})}}$$

where $\text{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2}\text{Tr}(\mathbf{KHLH})$ is the Hilbert-Schmidt Independence Criterion, Tr is matrix trace, and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. Intuitively, CKA can be viewed as a

multivariate correlation coefficient: it measures the normalized covariance between two centered Gram matrices. The numerator is their Frobenius inner product, i.e. the dot product of the matrices after flattening them into vectors.

Properties:

- CKA is symmetric, meaning $\text{CKA}(X, Y) = \text{CKA}(Y, X)$
- CKA is invariant to orthogonal transformations (rotations/reflections) and isotropic scaling.
- CKA bounded to a score in $[0, 1]$, where 1 means identical representational structure up to an orthogonal transformation.
- CKA only measure geometric alignment and not the functional similarity (Csiszárik et al., 2021)

Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) compares activation patterns across modalities, models, or seeds by transforming high-dimensional feature spaces into relational distance structures. For a set of n neuron embeddings within a given seed, an $n \times n$ Representational Dissimilarity Matrix (RDM) $\mathbf{D}$ is constructed using pairwise Euclidean distances, where $D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. To evaluate the representational alignment between seed A and seed B, the strictly lower triangular entries of their respective RDMs $\mathbf{D}_A$ and $\mathbf{D}_B$ are flattened into 1D vectors $\mathbf{d}_A, \mathbf{d}_B \in \mathbb{R}^{\frac{n(n-1)}{2}}$. These distances are then replaced by their ordinal positions to form rank vectors $\mathbf{r}_A$ and $\mathbf{r}_B$. The RSA similarity score is then computed using Spearman's rank correlation ρ between $\mathbf{r}_A$ and $\mathbf{r}_B$, making RSA invariant to monotonic non-linear transformations.

T.3 Whitening core channel activations with invariant response predictions

We take a trained baseline model and apply a linear whitening transformation to the cores channel activations and to the readout vectors. This transformation achieves identity channel covariance and identical response predictions by construction.

In Section 3 we assumed the core channel functions $f_i, \quad i = 1, \ldots, C$ to be elements of $L^2(\mu)$, i.e.,

$$\int_X f_i(x)^2 d\mu(x) < \infty,$$

meaning the integral of the squared channel functions is finite. This means the inner product over core channels is defined, and the covariance matrix is equivalent to the gram matrix. This approach therefore serves as a baseline for spatial consistency.

T.3.1 Selecting features for the whitening

In our setting, $f(X)$ consists of image embeddings sampled according to the learned receptive field of each neuron. For the Gaussian readout we just sample at the mean positions for each neuron, using bilinear interpolation as the learnt mean receptive field positions are not discrete. For ReFL readout, for each image, the CNN features are weighted and pooled using the neuron's learned spatial mask, yielding one pooled representation per neuron – and we do it for every neuron. This allows us to estimate the patch distribution $p(x)$ at the spatial locations relevant to each neuron.

T.3.2 Core activations and whitening transform.

Let the core feature maps for all B training images be

$$X \in \mathbb{R}^{B \times H \times W \times C},$$

where C is the number of channels. We flatten the first three dimensions (B, H, W) into a single index $N = BHW$ and obtain

$$\tilde{X} \in \mathbb{R}^{N \times C}.$$

We center the activations by subtracting the channel-wise mean

$$\mu = \frac{1}{N} \sum_{n=1}^{N} \tilde{X}_n \in \mathbb{R}^C, \qquad X_c = \tilde{X} - \mathbf{1}\mu^\top,$$

where $\tilde{X}_n$ represents the n-th row of $\tilde{X}$ and $\mathbf{1}$ represents a vector of ones used to match dimensions. The empirical covariance is

$$\Sigma = \frac{1}{N-1} X_c^\top X_c \in \mathbb{R}^{C \times C}.$$

Using any suitable matrix factorization (we used Cholesky), we compute a whitening transform $W \in \mathbb{R}^{C \times C}$ such that

$$W^\top \Sigma W = I.$$

The whitened activations are then

$$X' = X_c W,$$

which have identity covariance across channels by construction.

We now have a linear transformation to get core activations which have an identity covariance matrix. Next we change the readout vectors to preserve the same response predictions using the whitened core activations.

T.3.3 Readout transformation that preserves predictions.

Consider a single neuron with readout vector $r \in \mathbb{R}^C$ and scalar bias $b \in \mathbb{R}$. Let

$$F \in \mathbb{R}^{B \times C}$$

denote the per-image feature vectors extracted from B images via the model's Gaussian readout. The original predicted responses are

$$\hat{y} = Fr + b\mathbf{1} \in \mathbb{R}^B.$$

Define the whitened features

$$F' = (F - \mathbf{1}\mu^\top)W,$$

using the same whitening matrix W and mean μ. We seek r' and b' which predict responses based on whitened features F' such that predictions are unchanged:

$$F'r' + b'\mathbf{1} = Fr + b\mathbf{1}.$$

Choosing

$$r' = W^{-1}r$$

yields

$$\begin{aligned} F'r' + b'\mathbf{1} &= (F - \mathbf{1}\mu^\top)WW^{-1}r + b'\mathbf{1} \\ &= Fr - \mathbf{1}(\mu^\top r) + b'\mathbf{1}. \end{aligned}$$

The term $\mathbf{1}(\mu^\top r)$ is constant across images and can be absorbed into the bias by setting

$$b' = b + \mu^\top r.$$

Thus, (r', b') together with whitened core activations F' produce identical response predictions, while the core channels are decorrelated with unit variance. Importantly, this procedure can be applied post hoc, requiring neither architectural changes nor additional loss functions.