Mouse Visual Cortex Modelling

The impact of an ecological data diet on the neural predictivity of CNNs

Giacomo Amerio, Giovanni Lucarelli, Andrea Spinelli

github.com/giovanni-lucarelli/mice-representation

1 Introduction

Studies of the mouse visual system have identified a range of cortical areas that support diverse visual behaviors, including stimulus-reward associations, goal-directed navigation, and object-centered discrimination. Despite extensive investigation, a comprehensive understanding of the mouse visual cortex and its functional organization remains incomplete. Nayebi *et al.* [4] demonstrated that a shallow neural network architecture trained with a self-supervised objective and low-resolution visual inputs provides an optimal model of the mouse visual cortex. Their findings suggest that a lightweight, general-purpose visual system can effectively account for mouse visual representations. Acknowledging the inherently low visual acuity of mice, Nayebi *et al.* achieved improved neural predictivity by training their network on lower-resolution images. They used image resolution as a proxy for mouse visual acuity, rather than employing a biologically informed image preprocessing pipeline to approximate this property. They reported optimal neural predictivity using 64×64 pixel images and proposed that future work should incorporate more realistic, neurophysiologically informed preprocessing approaches.

The goal of this project is to extend Nayebi's analysis by developing and evaluating such preprocessing pipeline. Specifically, this study aims to investigate whether biologically informed visual transformations improve neural predictivity. Our findings indicate that the proposed ecological data diet yields higher neural predictivity compared to a traditional ImageNet-pretrained model. Moreover, a post-hoc analysis revealed that applying the data diet at inference time has a substantial impact on models trained without it. In addition, by disentangling the contributions of the individual components of the diet, we show that the diet attenuates ImageNet-specific features and reshapes the representational geometry of deeper layers to more closely resemble that of mouse visual cortex. Together, these results demonstrate that biologically motivated modifications to the input statistics can systematically steer standard deep networks toward more mouse-like representations.

2 Methods

2.1 Data Diet

A fair comparison between mouse visual perception and convolutional neural networks (CNNs) should respect the ecological limits of the mouse visual system, especially its limited spatial resolution. Mice have a low visual acuity; behaviorally measured constrast sensitivity functions (CSFs) show a band-pass profile that peaks at about 0.2 cycles per degree [5]. The CSF quantifies the inverse contrast required to detect sinusoidal gratigs at each spatial frequency, summarizing the behavior of the early visual pathway.

Following Muratore *et al.* [3], we treat the behavioral CSF as a functional approximation of the animal's visual system. We search for the combination of Gaussian blur and Gaussian noise that reproduces the mouse CSF measured behaviorally by Prusky. To calibrate these parameters, we simulate the standard grating detection task used in CSF experiments: blurred/noisy sinusoidal gratings are shown against noisy mid-gray images across a range of spatial frequencies and contrasts. The simulated display is divided into patches of 24 pixels; within each patch we compute the contrast as the standard deviation of pixel intensities, and concatenate these patch contrasts into a feature vector. We train a linear SVM to tell apart sinusoidal grating patches from a uniform mid-gray patch, using $n = 5000$ training examples for each class. For the grating class, we randomize orientation and phase. After training, we probe the SVM with gratings at different contrasts and spatial frequencies, and measure how often it correctly reports "grating" instead of "gray." For each spatial frequency, this accuracy-vs-contrast curve is treated as a psychometric function, from which we read off the contrast level that reaches threshold performance. We then collect those contrast thresholds across spatial frequencies (see fig. 1). To compare the resulting CSF to the mouse CSF, we perform a grid search over blur and noise variance to minimize the (absolute) error between the log contrast thresholds of our model and those measured in mice (see fig. 2). The optimal parameter combination identified by this simulated grating detection task is $\mathit{blur} = 0.176$ and $\mathit{noise} = 0.250$ (see fig. 3).

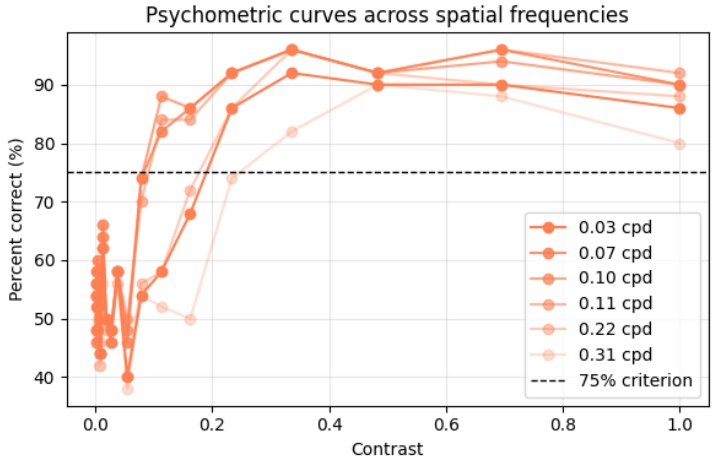

Figure 1: Psychometric curves are shown for different spatial frequencies. The intersection between the fitted psychometric curve and the 75% accuracy criterion is the threshold contrast chosen for the CSF.

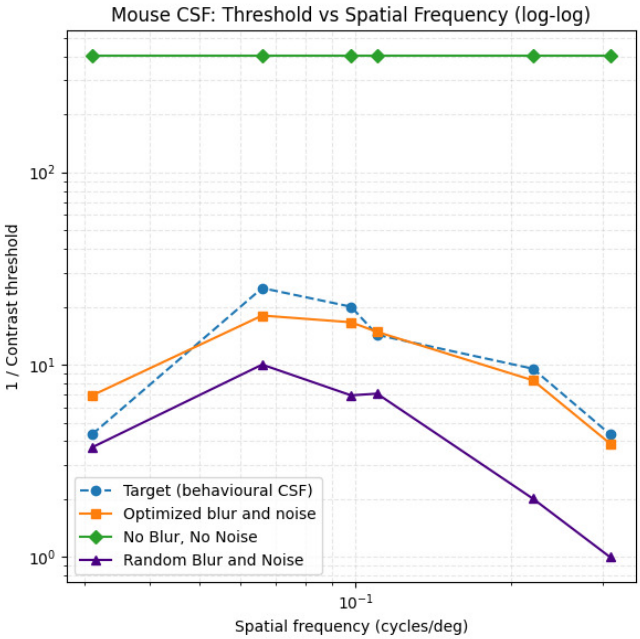

Figure 2: The contrast sensitivity function (CSF) measured for mice by Prusky *et al.* [5] (blue line) is shown along the CSFs obtained for a simulated observer, measuring the contrast of input images in small patches with different blur and noise parameters.

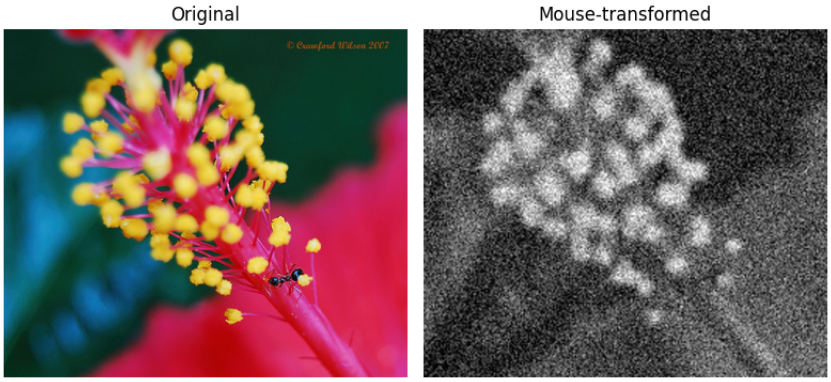

Figure 3: Illustration of an ImageNet sample after applying the optimized "artificial retina" preprocessing. The image has been transformed with a Gaussian blur ($\sigma = 0.176$) and additive Gaussian noise ($\sigma = 0.250$) to match the mouse contrast sensitivity function, as determined by the simulated psychophysical task (see fig. 2).

2.2 Representation Preprocessing

Allen Brain Neuropixels Visual Coding Neuropixels recordings from the Allen Brain Observatory Visual Coding dataset [1] were preprocessed following the methodology described by Nayebi *et al.* [4]. For each specimen and each visual area, we first computed the average temporal response across 10-ms time bins within the 0-250 ms post-stimulus window. This averaging was restricted to the largest contiguous time interval during which the median split-half reliability (computed across the population of recorded units within the specimen) exceeded a threshold of 0.3. Subsequently, for each visual area, we retained only those specimens with a number of responsive units greater than or equal to the 75th percentile across all specimens in that area. After this filtering procedure, the resulting neural response for each stimulus is represented as a real-valued vector, where each element corresponds to the mean response of an individual unit (neuron) to that specific stimulus. These vectors serve as the neural representations used in the subsequent neural mapping analyses.

AlexNet Activations We analyzed multiple variants of the AlexNet architecture, including: (1) an untrained network with Xavier initialization, (2) a pretrained model trained on the ImageNet dataset (PyTorch default), and (3) a pretrained model trained on ImageNet images preprocessed according to our ecological diet. In doing so, AlexNet is constrained to operate over the same spatial bandwidth as the mouse.

For each network, we computed activations in response to the same set of visual stimuli presented to the mice in the Allen Brain dataset. From each convolutional layer, activations were spatially averaged using global average pooling, resulting in a layer-specific feature vector whose dimensionality corresponds to the number of channels in that layer. These vectors constitute the model representations used for subsequent neural correspondence analyses.

	conv1	**conv2**	**conv3**	**conv4**	**conv5**
Channels	64	192	384	256	256

Table 1: Architecture of the convolutional layers in AlexNet. The number of channels in each layer corresponds to the dimensionality of the representation vector after global average pooling.

2.3 Neural mapping

The metric used in this work to map neuropixel responses to AlexNet activations is the Representation Similarity Analysis (RSA) [2].

Given two neural system X, Y, a set of n stimuli and let $\mathbf{z}_i^{(X)} \in \mathbb{R}^{|X|}, \mathbf{z}_i^{(Y)} \in \mathbb{R}^{|Y|}$ (where in general $|X| \neq |Y|$), be the responses of the two systems to each stimulus i.

We define RSA score

$$\text{RSA}^{(X,Y)} = \text{Corr}(\text{RDM}^{(X)}, \text{RDM}^{(Y)})$$

where

$$\text{RDM}_{i,j}^{(X)} = 1 - \text{Corr}(\mathbf{z}_i^{(X)}, \mathbf{z}_j^{(X)}), \quad \text{RDM}_{i,j}^{(Y)} = 1 - \text{Corr}(\mathbf{z}_i^{(Y)}, \mathbf{z}_j^{(Y)})$$

and RDM is the Representation Dissimilarity Matrix.

2.3.1 Interanimal consistency

One upper bound of the neural predictivity is defined by the inter-animal predictivity, *i.e.* the RSA score between the mice or equivalently the similarity score of one mouse and all the others. In our analysis a biological neural system is defined by the tuple (s, a) where s is the specimen and a is a visual area.

Corrected RSA The neuropixel data from the Allen Brain dataset are very noisy, hence to produce a robust estimate of the similarity score we used the corrected RSA proposed in Nayebi *et al.* [4], exploiting the presence of different trials for each stimulus.

Let $\mathbf{z}_{i,t}^{(X)} \in \mathbb{R}^{|X|}$ be the representation relative to the neural system $X = (s, a)$ for the stimulus i and the trial $t \in \{1, \dots, T\}$ (in the dataset for each stimulus we have $T = 50$). We define the X_1, X_2 two new neural systems such that $\mathbf{z}_i^{(X_1)} \in \mathbb{R}^{|X|}, \mathbf{z}_i^{(X_2)} \in \mathbb{R}^{|X|}$ are obtained by taking two random halves of trials and computing the average response, for each stimulus, over each subset of trials.

$$\text{RSA}_{\text{corrected}}^{(X,Y)} = \frac{\text{RSA}^{(X,Y)}}{\sqrt{\text{reliability}(X) \cdot \text{reliability}(Y)}}$$

and the reliability of a representation is computed as the similarity between two halves of the trials, using split-half correlation (Spearman-Brown corrected):

$$\text{reliability}(X) = \text{SB}(\text{RSA}^{(X_1,X_2)}), \quad \text{SB}(r) = \frac{2r}{1+r}.$$

The split has been performed 100 times by randomly assigning each trial to one of the two halves, and then averaging the similarity score over the splits.

Pooled interanimal consistency To obtain a more robust estimate of inter-animal similarity, we introduced a pooled version of the corrected RSA, where neural responses from multiple specimens of the same visual area are combined into a single, population-level representation. Specifically, let $\mathcal{S}_a = \{s_1, s_2, \dots, s_S\}$ denote the set of specimens for a fixed area a. For each specimen (s, a) we have trial responses $\mathbf{z}_{i,t}^{(s,a)} \in \mathbb{R}^{|(s,a)|}$ for stimulus i and trial t.

For each source specimen $(s, a) \in \mathcal{S}_a \setminus \{s^*\}$, we randomly split the available trials into two disjoint halves and compute the mean response over each half, obtaining $\mathbf{z}_i^{(s_1,a)}$ and $\mathbf{z}_i^{(s_2,a)}$. The pooled representations are then built by concatenating the averaged responses of all source specimens along the neural feature dimension:

$$\mathbf{z}_i^{(\text{pool}_1,a)} = \bigoplus_{s \in \mathcal{S}_a \setminus \{s^*\}} \mathbf{z}_i^{(s_1,a)}, \qquad \mathbf{z}_i^{(\text{pool}_2,a)} = \bigoplus_{s \in \mathcal{S}_a \setminus \{s^*\}} \mathbf{z}_i^{(s_2,a)}.$$

This "population pooling" procedure creates a new aggregate neural system combining information across multiple specimens from the same area.

At each bootstrap iteration, the pooled halves (pool_1, a) and (pool_2, a) are compared with two independent halves (s_1^*, a) and (s_2^*, a) of a held-out target specimen (s^*, a) using the corrected RSA metric defined above. To ensure symmetry, we compute the cross-half similarity:

$$\text{RSA}_{\text{num}} = \tfrac{1}{2}\Big(\text{RSA}^{(\text{pool}_1, a,\, s_2^*, a)} + \text{RSA}^{(\text{pool}_2, a,\, s_1^*, a)}\Big),$$

and normalize it by the geometric mean of the reliabilities of both the pooled and target systems:

$$\text{RSA}^{(\text{corr})}_{\text{pooled}}(a) = \frac{\text{RSA}_{\text{num}}}{\sqrt{\text{reliability}(\text{pool}, a) \cdot \text{reliability}(s^*, a)}}.$$

The procedure is repeated over 100 random trial partitions, and the final pooled inter-animal consistency score for area a is obtained as the mean and standard deviation of the resulting corrected RSA values.

2.3.2 Neural predictivity

The similarity between AlexNet activations and the neuropixel responses has been computes analogously to what introduced above. Specifically, for each layer l of the convolutional network and (s, a) specimen-are pair, we can define

$$\text{RSA}^{(s,a,l)} = \text{Corr}(\text{RDM}^{(s,a)}, \text{RDM}^{(\text{AlexNet}, l)})$$

and

$$\text{RSA}^{(s,a,l)}_{\text{corrected}} = \frac{\text{RSA}^{(s,a,l)}}{\sqrt{\text{reliability}(s, a)}}$$

where $\text{reliability}(\text{AlexNet}, l) = 1$ for all layers l, since the activations in a forward pass are deterministic.

3 Results

Table 2 reports the number of selected specimens and units after the Allen dataset preprocessing, and the pooled inter-animal consistency (RSA score) computed for each visual area. The mean RSA values, together with their corresponding standard errors of the mean (SEM), quantify the reliability of neural responses across animals within each region.

Area	Specimens	Units per Specimen	IC
AL	6	[85, 127, 184, 91, 166, 89]	0.648 ± 0.021
AM	7	[70, 94, 71, 72, 70, 74, 135]	0.570 ± 0.030
LM	6	[51, 53, 74, 56, 77, 58]	0.453 ± 0.031
PM	5	[65, 115, 62, 90, 54]	0.552 ± 0.020
RL	7	[76, 69, 67, 68, 79, 95, 111]	0.542 ± 0.027
V1	8	[110, 102, 91, 85, 88, 93, 94, 85]	0.592 ± 0.038

Table 2: Number of specimen IDs, total units, units per specimen ID, and interanimal consistency score (± standard error) for each visual area.

3.1 Data diet impact

In fig. 4 is shown the RSA score for each visual areas and for each convolutional layer of AlexNet. We can see that the RSA score is higher in the case of data diet both at training and inference time. At training time means that all the ImageNet dataset have been preprocessed with our "mouse-like" pipeline, whereas at inference time means that the pipeline has been applyed to the same Allen Brain stimuli that the mice saw.

Specifically, form fig. 5 where the untrained model (baseline) have been subtracted, we can see that, the diet affects negatively the first convolutional layer but leads to higer RSA score in the subsequent layers. Moreover we can observe that in some visual areas, namely AL, LM, RL, V1, for the last two layers the trained models provide a worse score with respect to the untrained one.

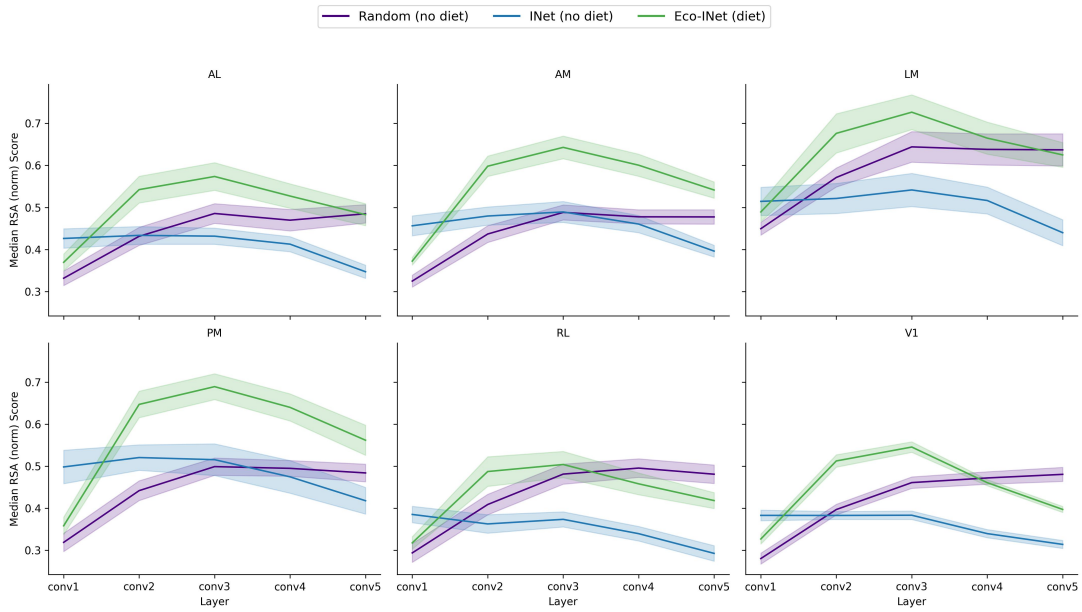

Figure 4: RSA score, normalized to the noise ceiling (pooled interanimal consistency), for all visual areas and for all AlexNet convolutional layers. In **violet** the untrained model, in **blue** the model trained without data diet (inference without diet), in **green** the one trained with data diet (inference with diet).

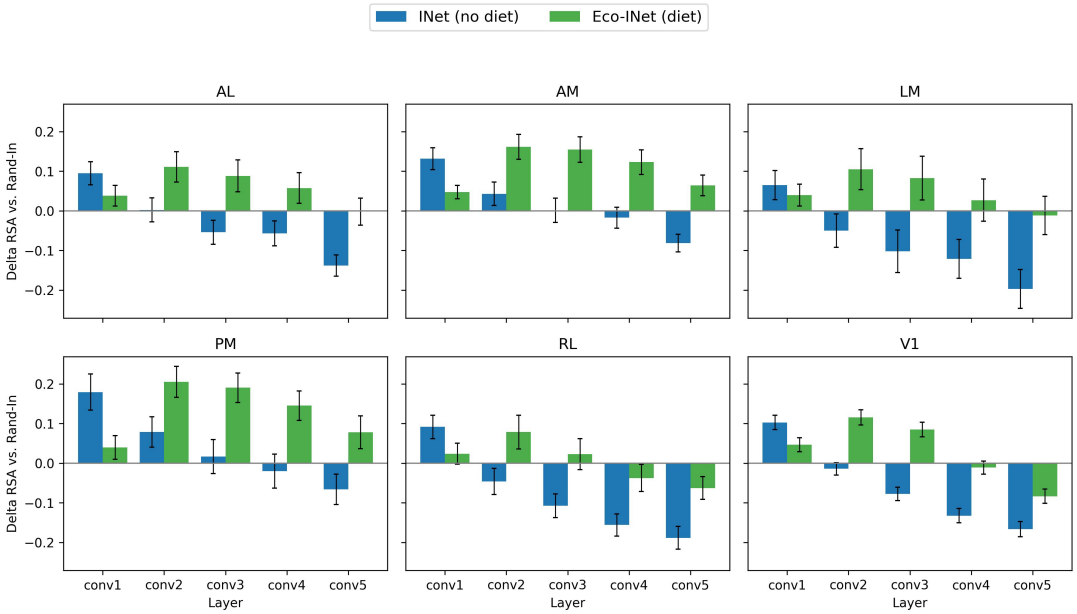

Figure 5: Delta RSA vs untrained model. In **blue** the model trained without data diet (inference without diet), in **green** the one trained with data diet (inference with diet).

Random diet fig. 6 illustrates the effect of a "random diet". Here, blur and noise variance are set to one for all images. For the untrained network, the highest RSA scores across areas and layers are obtained when *no* diet is applied at inference; all diets, including the random one, reduce predictivity. This contrasts with the trained models, which instead benefit from appropriately chosen diets.

Because the random diet uses a single random configuration of blur and noise, these results should be considered preliminary. A more systematic exploration of the random diet parameter space is required to draw robust conclusions.

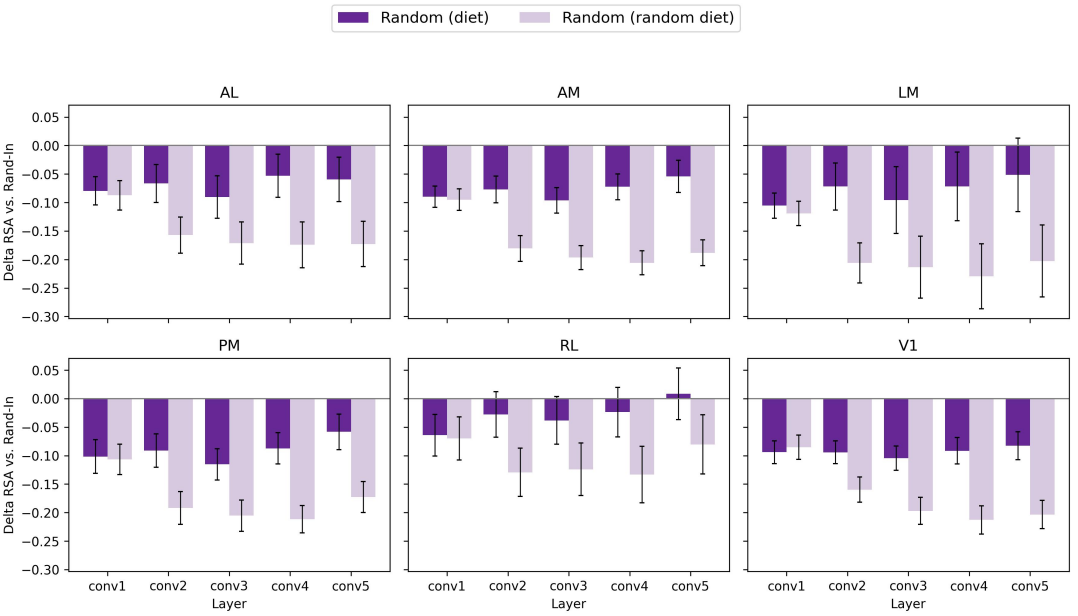

Figure 6: Delta RSA vs untrained model. In lighter shades of violet are shown different data diets (at inference time): our diet, Nayebi's diet and one random diet.

3.2 Decoupling training- and inference-time diets

We next disentangle the contributions of training- and inference-time preprocessing. When AlexNet is trained on the original (non-mouse-like) ImageNet dataset but the data diet is applied only at inference time to the Allen stimuli, we obtain qualitatively similar trends. As shown in fig. 7, applying the diet at inference consistently improves neural predictivity in all layers except `conv1`, for both the diet-trained and the standard ImageNet-trained models.

The effect is particularly pronounced for the network trained on unprocessed ImageNet: adding the diet only at inference brings its RSA scores close to those of the model trained directly on diet-processed ImageNet. This indicates that a substantial portion of the benefit can be achieved by reshaping the stimulus statistics at inference time, without necessarily re-training the model.

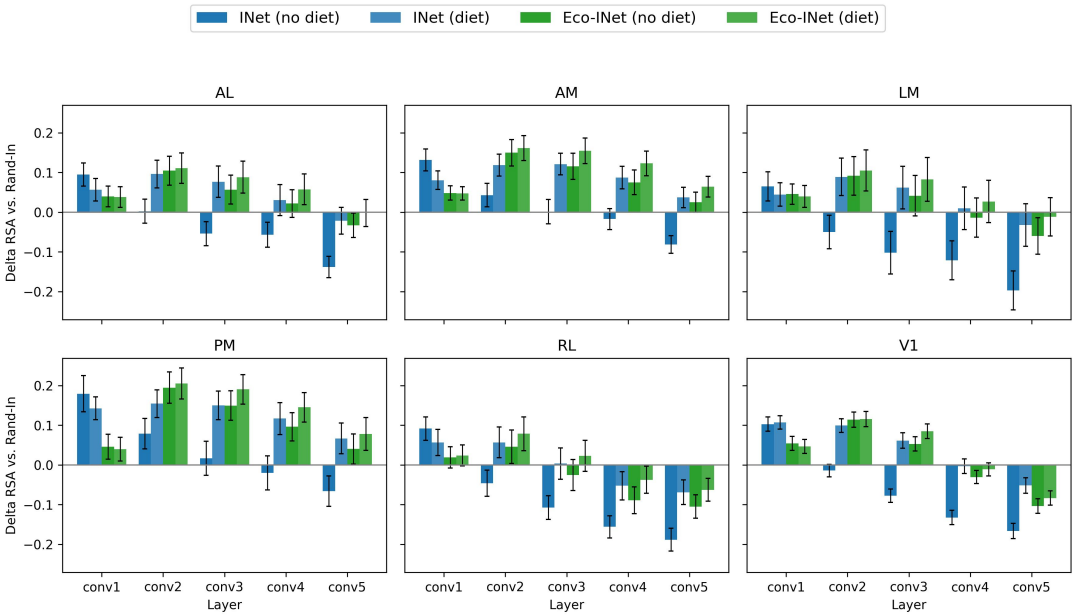

Figure 7: Delta RSA vs untrained model. In **blue** the model trained without data diet (inference without diet), in **light blue** the model trained withoud data diet but with diet at inference, in **green** the one trained with data diet (inference with diet).

3.3 Decoupling blur and noise

Figure 8 shows the separate contributions of Gaussian blur and Gaussian noise to the RSA score. In `conv1`, the best predictivity is obtained with a purely low-pass signal (blur only, no added noise) across all visual areas. However, this advantage decreases in deeper layers, where the full diet (blur + noise) yields higher scores.

The degradation in `conv1` when noise is added can be understood in terms of the approximate linearity of the first convolutional layer. Being the first stage of processing (with no additional pooling or non-linear transformations yet), `conv1` is well-approximated by a linear filter bank. Adding Gaussian noise to the input images is therefore effectively equivalent to adding noise directly on the layer's representations. Since this noise is independently sampled across images, increasing its variance reduces the correlation between feature patterns elicited by different stimuli. As a consequence, dissimilarities in the RDM become dominated by the noise term rather than by differences in stimulus statistics. In this way, the blur–acting

as a low-pass filter and already providing a good match to the cortical responses at conv1–is degraded when noise is added on top.

This interpretation is consistent with the behavior of the "random diet", defined by setting the variance of both the Gaussian blur and Gaussian noise to one. In this regime, the very large blur and noise produce highly corrupted representations, effectively destroying similarity structure both across images and across layers.

The key observation is that, in subsequent layers, adding noise to the input yields higher RSA scores. Noise that behaves as measurement noise at conv1 is gradually averaged out across the network hierarchy. The remaining effect of the diet acts as a form of "constructive interference" that brings the RDMs of AlexNet closer to the biological RDMs. In other words, the noise component behaves as a *biological regularizer* for the artificial network: it weakens AlexNet's ImageNet-specific ability to discriminate between images that are effectively similar from the mouse's perspective.

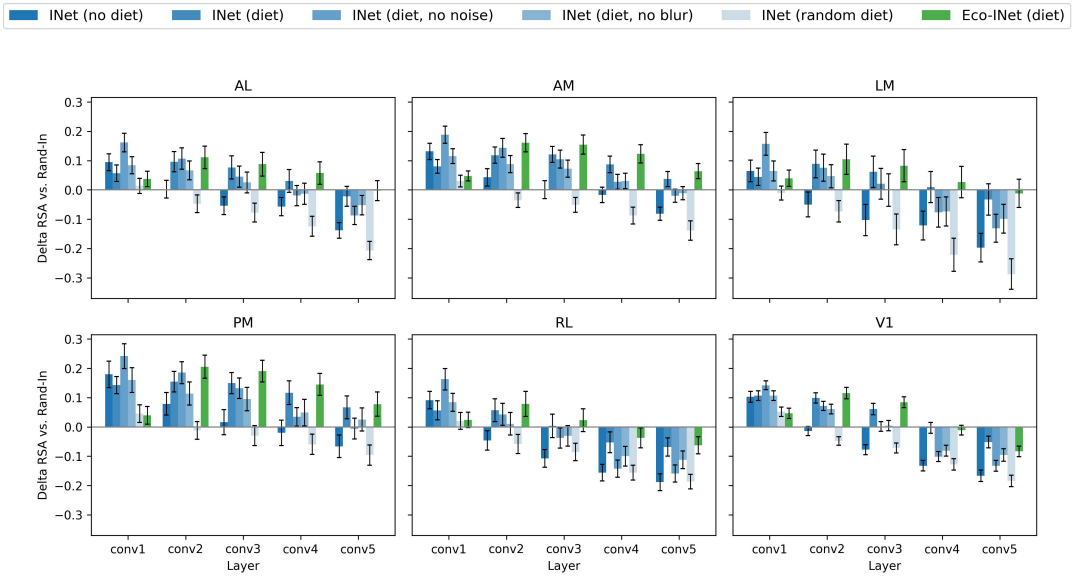

Figure 8

Nayebi diet To compare our results with Nayebi et al. [4], we defined a pseudo data diet analogous to their optimal configuration. Specifically, we rescale the stimuli to 64×64 pixels and then upsample them back to 224×224 to match AlexNet's input size. This pipeline implements a strong low-pass filter, harsher than our Gaussian blur, and can be regarded as a special case of our diet without the Gaussian noise.

As shown in fig. 9, the two diets exhibit very similar behavior across areas and layers, including conv1. Our blur-based diet achieves slightly better results in the first layer, likely because its kernel size was optimized to approximate Prusky's contrast sensitivity function (CSF) for mice.

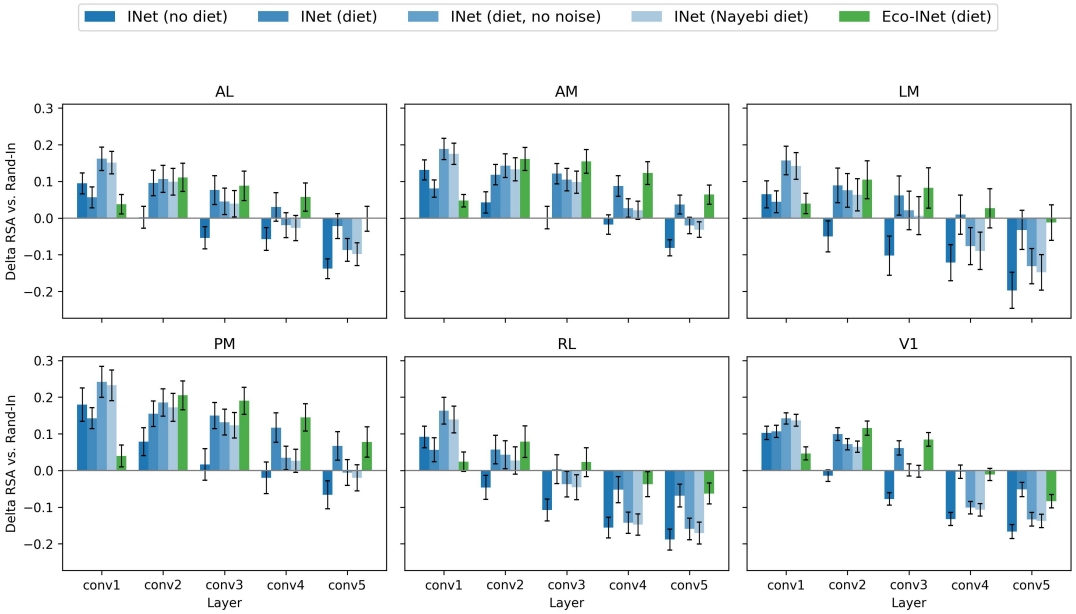

Figure 9: Delta RSA vs untrained model. In **blue** the model trained without data diet (inference without diet), in **light blue** the model trained withoud data diet but with diet at inference, in **lighter blue** the model trained withoud data diet but with Nayebi's diet at inference, in **green** the one trained with data diet (inference with diet).

4 Discussion

In this project, we developed a biologically-informed preprocessing pipeline to feed visual stimuli to artificial neural networks that more closely resemble those perceived by mice. We trained AlexNet on the transformed ImageNet dataset and evaluated its neural similarity to biological responses recorded in the Allen Neuropixels dataset.

Our results demonstrate that the ecological data diet, *i.e.*, the one obtained by our preprocessing strategy, provided both at training and inference time, provides the largest neural predictivity for all visual areas adn for all convolutional layers of AlexNet, except for the first one (see fig. 5). We note also that in all visual areas the peaks in neural predictivity are always in layers `conv2`, `conv3`. Therefore, we do not have evidence of a hierarchical structure in the mouse visual cortex, as already stated by Nayebi *et al.*

We also observed a clear decoupling effect, in the neural predictivity score, if the diet is provided at training or at inference time (see fig. 7). While all tested data diets improved the predictivity of pretrained models, they consistently decreased the predictivity of untrained networks with random weights (see fig. 6). This indicates that the diet engages meaningfully with learned feature represention. In the case of trained models (both with diet and without) the inference time diet always improves the score, but with a clear larger impact on the model trained on unprocessed images, leading to a score almost always close to the one obtained by the model trained on preprocessed images (see fig. 7).

By disentangling the contributions of Gaussian blur and Gaussian noise (fig. 8) and comparing them to the Nayebi pseudo-diet (fig. 9), we obtained additional insight into how ecological preprocessing shapes neural predictivity. In `conv1`, a purely low-pass signal (blur only, no noise) produced the highest RSA scores across all visual areas, while performance degraded when additive Gaussian noise was introduced. This is consistent with the approximate linearity of the first convolutional layer: `conv1` can be viewed as a linear filter acting directly on the input, so adding independent Gaussian noise to the images is effectively equivalent to adding noise to the layer's representations. Because this noise term is uncorrelated across images, it reduces correlations between activity patterns and causes dissimilarities in the RDM to be dominated by the noise rather than by stimulus-driven structure, thereby degrading a blur-only signal that already provides a good match to mouse spatial acuity. In deeper layers, however, the full ecological diet (blur + noise) consistently outperforms the blur-only condition: noise that behaves as "measurement noise" in `conv1` is progressively attenuated and averaged out by the hierarchical transformations (convolutions, nonlinearities, pooling), and what remains is a reshaping of the representational geometry that undermines AlexNet's ImageNet-specific reliance on high-frequency, texture-like cues and reduces distinctions between images that are effectively similar from the mouse's perspective. In this sense, blur and noise act as complementary components: the blur aligns early spatial statistics with the mouse contrast sensitivity function, while the noise acts as a *biological regularizer* that pushes deeper-layer representations toward a more mouse-like similarity structure. The Nayebi pseudo-diet can be interpreted as a particular instance of this blur-only regime with a stronger effective low-pass kernel (downscale to 64×64 followed by upscaling to 224×224), which explains its strong performance in `conv1`, whereas our calibrated blur combined with noise yields superior alignment in deeper layers. The behavior of the "random diet" further supports this view: when blur and noise variances are set to large, uncalibrated values, the representations become dominated by stochastic distortions and lose meaningful similarity structure at all depths (fig. 6).

Our overarching goal was to identify a model of the mouse visual system that incorporates the animal's natural visual experience. For this reason, we considered the model trained and evaluated with the ecological diet as our reference point, as it consistently processes mouse-like images. Surprisingly, however, most of the improvement in neural alignment originates at inference time: even a network trained on standard high-acuity images shows substantially increased predictivity once mouse-like transformations are applied to its inputs. This suggests that inference-time preprocessing may be the primary driver of alignment with biological neural activity. Consequently, an intriguing possibility emerges: perhaps modifying only the "artificial retina"—the transformations applied at inference—could be sufficient to repurpose a generic vision model such as AlexNet into a plausible "virtual brain" for different species.

Taken together, these findings highlight the central role of stimulus realism and input statistics in brain-model correspondence analyses. They also show that seemingly simple manipulations–such as calibrated blur and controlled noise–can substantially alter representational similarity, acting both as an approximation to species-specific optics and as an effective regularizer of artificial networks. Future work should explore more detailed retinal and optical models, extend these analyses to other architectures and species, and systematically characterize how different components of ecological preprocessing interact with learning dynamics. Such efforts will be crucial for disentangling the contributions of model architecture, training objective, and input statistics to the emergence of biologically plausible visual representations.

References

[1] Allen Institute for Brain Science. *Neuropixels Visual Coding — White Paper v1.0*. White Paper. Accessed: 2025-11-07. Allen Institute for Brain Science, 2024. URL: https://brainmapportal-live-4cc80a57cd6e400d854-f7fdcae.divio-media.net/filer_public/80/75/8075a100-ca64-429a-b39a-569121b612b2/neuropixels_visual_coding_-_white_paper_v10.pdf.

[2] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. "Representational similarity analysis - connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience* Volume 2 - 2008 (2008). ISSN: 1662-5137. DOI: 10.3389/neuro.06.004.2008. URL: https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008.

[3] Paolo Muratore, Alireza Alemi, and Davide Zoccolan. "Unraveling the complexity of rat object vision requires a full convolutional network and beyond". In: *Patterns* 6.2 (2025), p. 101149. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2024.101149. URL: https://www.sciencedirect.com/science/article/pii/S2666389924003210.

[4] Aran Nayebi et al. "Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation". In: *PLOS Computational Biology* 19.10 (Oct. 2023), pp. 1–36. DOI: 10.1371/journal.pcbi.1011506. URL: https://doi.org/10.1371/journal.pcbi.1011506.

[5] G.T. Prusky and R.M. Douglas. "Characterization of mouse cortical spatial vision". In: *Vision Research* 44.28 (2004). The Mouse Visual System: From Photoreceptors to Cortex, pp. 3411–3418. ISSN: 0042-6989. DOI: https://doi.org/10.1016/j.visres.2004.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0042698904004390.