Deep Convolutional Neural Network Models of Mouse Visual Cortex: Does Self-Supervised Learning Produce More Brain-Like Representations?

Author: Callum Messiter
Student Number: 24183058
Supervisors: Ali Haydaroglu and Prof Kenneth D. Harris
Department: Department of Neuromuscular Diseases

August 2025

Abstract: 249 words
Introduction: 747 words
Results: 1737 words
Methods: 2010 words
Discussion: 1480 words
Limitations of Methods: 749 words
Total: 6971 words

I acknowledge the use of ChatGPT 4.0 (Open AI, https://chat.openai.com) to list background sources, engage in a socratic dialogue, and check my final draft for grammatical errors and language issues.

1 Abstract

Deep convolutional neural networks (DCNNs) have become powerful models of visual cortex, but their success in primates has not always translated to rodents. Mouse vision is tuned to ethological demands—motion, predator detection, navigation—rather than fine-grained object categorisation. This raises the question: does self-supervised learning (SSL) provide a more biologically faithful account of mouse vision.

Here, we compared a self-supervised SimCLR model with a supervised VGG-19, using calcium imaging data from mouse V1 and higher visual areas (HVAs). Neural responses to 1,500 natural images were predicted via ridge regression, and model–brain alignment was assessed across multiple dimensions: neural predictivity (fraction of explainable variance explained, FEV), representational geometry (RSA), dimensionality -eigenspectrum decay rate (α) and PC90 - semantic clustering, and qualitative feature visualisation.

SimCLR consistently outperformed VGG-19 in predicting neural responses, with its intermediate layers achieving the highest FEV. These layers also aligned most closely with mouse visual cortex in representational geometry, matched the long-tailed eigenspectrum decay ($\alpha \equiv 1$) characteristic of biological vision, and showed low semantic clustering comparable to neural data. Feature visualisation revealed that SimCLR intermediate layers emphasise mid-level features—textures, spatial contrasts, and orientations—mirroring the tuning of mouse visual cortex.

Regression analysis indicated that no single representational property explained neural predictivity. Rather, predictive layers shared a characteristic profile: high neural–geometric alignment, moderate dimensionality, and low semanticity.

These findings support SSL as a biologically plausible learning framework for modelling mouse vision, and highlight representational isomorphism—beyond mere predictivity—as a critical benchmark for assessing model–brain alignment.

2 Acknowledgements

Thank you to Ali Haydaroğlu for his time, and his kindness in supervision. Thank you to Prof Kenneth Harris and Prof Matteo Carandini for the opportunity to do science. Thank you to Prof Kenneth Harris, Prof Matteo Carandini, Dr Andrew Landau, Dr Sophie Skriabine and Ali Haydaroğlu for their valuable suggestions for analyses.

Thank you to Dr Andrew Batchelor—and especially to Prof Daniel Bendor—for giving me a chance. I am deeply appreciative of Dan's consistent support throughout the programme. I thank Dr Hana Roš for her warm support, Dr Sadra Sadeh for his time and advice, and Dr Max Garagnani, whose generosity helped me find my way to the world of brains, minds and machines.

I am also grateful to Oliver Millar, Mridul Pandey and Sailaja Kuruvada, for their friendship.

Above all, I thank Inès for her steadfast love and support. We share this degree.

Finally, I dedicate this work to Lorraine.

3 Introduction

To what extent is machine intelligence equivalent to biological intelligence? This problem lies at the intersection of computational neuroscience and computer vision. Supervised deep convolutional neural networks (DCNNs) outperform humans on object recognition tasks (Kheradpisheh et al., 2016). In this sense the model passes a domain-constrained Turing test: it is functionally equivalent to the human visual cortex.

Framed in Marr's terms, both systems solve the same computational problem (Marr, 1982): invariant object recognition. Yet the same problem can be solved by many algorithms, and the same algorithm realised on many substrates. This is Putnam's "multiple realisability" (Putnam, 1967). What, then, is going on inside these systems? How similar are their internal mechanisms?

If we are to know what the visual system does, our model must predict neural responses to arbitrary stimuli (Carandini, 2005). Functional equivalence is not enough: we must probe the internal mechanisms, at Marr's algorithmic level, decomposed into (i) the representations and (ii) the transformations that operate on them.

Yamins & DiCarlo (2016) specify the task: to discover the nonlinear *transformations* across the visual hierarchy—input–output mappings that convert raw sensory inputs into high-level representations. We argue that this is a *neural* functionalism: such a model captures the right cascade of transformations, but not necessarily the representational *content*. Are the model's representations **isomorphic** to those of cortex, such that stimulus relationships can be mapped across spaces without loss?

Crucially, high neural-response predictivity does not guarantee this isomorphism. Cadena et al. (2019b) illustrate this: an untrained DCNN predicts neural responses in mouse visual cortex as accurately as a trained network, and hierarchical correspondence between trained model layers and brain areas is missing. Thus, we should assess model representations both in terms of predictivity *and* biological plausibility.

Good progress has been made on predictivity. In humans (Güçlü & van Gerven, 2015) and macaques (Yamins et al., 2014; Yamins & DiCarlo, 2016; Nayebi et al., 2018; Cadena et al., 2019), DCNNs surpass classical models in predicting neural responses to natural images. Their success reflects three factors: (i) neurally inspired computations like filtering, pooling, and normalisation (Carandini, 2005; Serre, 2007; Carandini & Heeger, 2012); (ii) architectural constraints like local receptive fields (RF), retinotopy, and RF expansion (Malach, 2002); and (iii) their learning objective - the major focus of the present study. The task of the primate ventral stream is invariant object categorisation. Accordingly, DCNNs trained on this task predict neural responses in V4 and IT with unrivalled accuracy (Yamins et al., 2014; Yamins & DiCarlo, 2016; Cadena et al., 2019).

In contrast, DCNNs have been less successful in mice (Cadena et al., 2019; de Vries et al., 2020). Mouse visual cortex is shallower and more parallel (Harris et al., 2019; Siegle et al., 2021); indeed, even a two-layer CNN can predict mouse V1 well (Du et al., 2025). More critically, mice, not benefiting from category labels during development, nor the high visual acuity of primates (Prusky et al., 2000; Acar, 2019), may not primarily rely on fine-grained, invariant object categorisation. Instead, their visually guided behaviours are more general-purpose (Nayebi et al., 2023). A classification objective thus forces model representations to align with category boundaries that the mouse visual system may not have evolved to encode.

Instead, self-supervised objectives - learning invariances directly from natural statistics, without labels - may be better suited to mice. SimCLR, for example, learns by aligning augmented views of the same image (Chen et al., 2020). This instance-discrimination objective encourages representations that capture features stable across common transformations in natural vision (e.g., translation, rotation, scale), without relying on human-defined categories. Such models outperform supervised DCNNs in mice (Bakhtiari et al., 2021; Nayebi et al., 2023) and also perform strongly in primates (Zhuang et al., 2020).

Despite promising studies, modelling mouse vision with self-supervised DCNNs is relatively

underexplored. This study thus has two aims. First, we tested how the learning objective affects neural predictivity in mouse visual cortex, comparing supervised VGG-19 with self-supervised SimCLR. We expected SimCLR to outperform VGG-19, since its contrastive objective better matches the ethological demands of mice. Second, we went beyond predictivity, to assess representational isomorphism. Quantitatively, we built a multidimensional profile of each model layer—geometry, dimensionality, and semanticity—and related these metrics to neural predictivity. Qualitatively, we visualised representations from the most predictive DCNN layers, aligned with the dominant axis of neural population variance, to assess whether model and brain emphasise similar visual features. In sum, this study attempts to determine systematically if self-supervised learning produces DCNN model representations more aligned with mouse visual cortex.

4 Methods & Materials

All analyses described in this thesis were performed solely by the author. The author did not participate in neuronal data collection, for which the full methodological details can be found in Schmutz & Haydaroğlu et al. (2025) and Haydaroğlu et al. (2025).

Unless specified otherwise, the code for these analyses was written by the author. It is available at: https://github.com/cjdm94/self-supervised-dcnn-models-of-mouse-visual-cortex.

4.1 Neuronal data collection

Neural data were obtained from calcium imaging of excitatory neurons in mouse visual cortex (Schmutz & Haydaroğlu et al., 2025; Haydaroğlu et al., 2025). Recordings spanned V1 and higher visual areas, yielding on average 19k neurons across three mice. Mice viewed 1,866 natural-image mosaics (Stringer et al., 2019), each presented twice in random order. Preprocessing (Suite3D; Haydaroğlu et al., 2025) produced one scalar response per neuron per stimulus presentation. Full experimental details are reported in the original publications.

4.2 Neuronal data filtering

For each of the three mice, neuronal datasets were obtained comprising deconvolved calcium activity (standard-deviation normalised) from 19,223 ± 2,948 neurons across two presentations of 1,573 natural images. Neurons eliciting reliable stimulus responses were identified using signal-related variance (SRV), defined as the cross-trial covariance between repeats, normalised by the average variance (Stringer et al., 2019).

Statistical significance was assessed by generating null SRV distributions from 100 shuffles of stimulus–response pairings, with neurons considered reliable if their SRV exceeded the 99th percentile of the null. Reliable neurons were ranked by SRV, and the top 500 were retained for analysis. For each retained neuron, trial-averaged responses were computed by averaging across repeats.

4.3 Preparing natural images for DCNN feature extraction

Each stimulus mosaic (375 × 1500 px; see Section 4.2) contained three natural images. The left-most image (375 × 500 px) was extracted for analysis and preprocessed according to each DCNN's training pipeline. For SimCLR (Chen et al., 2020), trained on STL-10 (Coates et al., 2011), images were resized to a 96-px shorter side, center-cropped to 96×96 px, rescaled to $[0, 1]$, and normalised to zero mean and unit variance using mean = [0.5, 0.5, 0.5] and std = [0.5, 0.5, 0.5]. For VGG-19 (Simonyan & Zisserman, 2015), trained on ImageNet (Deng et al., 2009), images were resized to a 224-px shorter side, center-cropped to 224 × 224 px, rescaled to $[0, 1]$, and normalised using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

4.4 SimCLR feature extraction

We used a ResNet-18–based SimCLR DCNN pretrained on STL-10 (Chen et al., 2020), with the projection head removed to retain the encoder backbone. Features were extracted from intermediate layers (layer1–layer4) and the final fully connected (fc) layer using forward hooks. We selected layer1–layer4 and fc to span the representational hierarchy of ResNet-18, from low-level spatial features to high-level embeddings. This allowed me to probe a range of feature complexities and compare them with hierarchical processing stages in mouse visual cortex. Feature extraction was performed using PyTorch in evaluation mode. Stimuli were passed through the DCNN in batches of 64 via a DataLoader. Feature maps were retained in their original spatial format, for both intermediate layers (C × H × W) and fc (C x D). Extraction was deterministic due to fixed random

seeds and run either on CPU or GPU, depending on availability. The resulting features, for each mouse and its corresponding subset of natural images, were saved for downstream analysis.

4.5 VGG-19 feature extraction

We extracted features from a pre-trained VGG-19 DCNN using PyTorch's torchvision library. Five layers were selected—conv2_2, conv3_4, conv4_4, conv5_4, and fc2—to span the network hierarchy from early to deep stages. This selection was designed to align approximately with the representational depth of SimCLR layers (1–4 and fc), enabling fair cross-network comparisons. Convolutional layer outputs were spatially pooled using adaptive average pooling to 1×1 per channel, then flattened. Fully connected outputs were flattened directly. Activations were captured via forward hooks during inference, and features were computed in batches. This yielded a fixed-length vector per image per layer, suitable for representational comparison with SimCLR and neural data.

4.6 Gabor filter bank (GFB) feature extraction

To extract low-level visual features, we applied a custom Gabor filter bank to each greyscale image. Filters were generated using OpenCV's getGaborKernel function with a kernel size of 31 pixels. We used 8 orientations ($\theta \in [0, \pi)$), 3 spatial wavelengths ($\lambda = 5, 10, 20$), and 2 aspect ratios ($\gamma = 0.5, 1.0$) to approximate the range of spatial frequency and orientation selectivity observed in mouse V1 neurons. Filters were applied using 2D convolution in PyTorch, and responses were spatially mean-pooled to yield one feature per filter. This produced a compact vector of Gabor responses for each image, capturing edge and texture information across scales and orientations.

4.7 Dimensionality reduction and standardisation of representations

To standardise dimensionality across analyses, we reduced feature vectors to a fixed number of PCs using principal component analysis (PCA). For SimCLR, VGG-19, and neural data, we retained 100 PCs, balancing compression with representational richness. This enabled consistent comparisons across DCNNs and neural responses for multiple metrics (e.g. FEV, RSA, semanticity) across multiple mice (n=3). For Gabor features, which are lower dimensional and used only for FEV prediction, we retained 10 PCs. We selected these values based on performance saturation and to ensure comparability across layers, networks, and biological data.

4.8 Predicting neuronal activity from DCNN activations

A regularised linear model (RidgeCV) was fit to map layerwise DCNN activations onto trial-averaged neural responses.

Images were partitioned into training (80%) and test (20%) sets using fixed random indices. Predictions were generated from approximately 1500 natural images per mouse (1573, 1585, and 1380 for mice 1–3, respectively) to a subpopulation of 500 neurons.

$$\hat{W}_\lambda = \arg\min_W |Y - XW|_F^2 + \lambda |W|_F^2 = (X^\top X + \lambda I)^{-1} X^\top Y. \tag{1}$$

The weight matrix $\hat{W}_\lambda$ was estimated using ridge regression by minimising the objective function $\|Y - XW\|_F^2 + \lambda \|W\|_F^2$, with the solution expressed as $\hat{W}_\lambda = (X^\top X + \lambda I)^{-1} X^\top Y$, where λ represents the regularisation parameter.

For each DCNN layer, flattened features were optionally reduced using PCA before regression (see Section 4.7). Two pipelines were implemented per layer per mouse: one using the full neural population (no PCA) and one using only the first principal component (PC1) of neural responses.

Ridge regression was trained on the training set using five-fold cross-validation to select the regularisation strength, and evaluated on the test set. All variance estimates were computed with unbiased estimators ($ddof = 1$).

1500 stimuli provided sufficient samples for cross-validation, while 500 neurons captured population responses at a manageable scale. Dimensionality reduction preserved variance while avoiding instability from fitting tens of thousands of correlated features. Ridge regression was used because features far exceeded stimuli; the L2 penalty regularised weights, stabilised solutions, and improved generalisation.

4.9 Evaluating predictive performance

The performance of regression models was evaluated with the fraction of explainable variance explained (FEV). FEV was computed as

$$\mathrm{FEV}_j = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(y_{i,j} - \hat{y}_{i,j})^2 - \sigma^2_{\mathrm{noise},j}}{\mathrm{Var}[y_j] - \sigma^2_{\mathrm{noise},j}} \tag{2}$$

where $y_{i,j}$ is the observed trial-averaged response of neuron j to test image i, and $\hat{y}_{i,j}$ is the model's predicted response. $\sigma^2_{\mathrm{noise},j}$ denotes the trial-to-trial variance for neuron j, estimated by averaging the variance across repeated presentations of each test image. $\mathrm{Var}[y_j]$ is the variance of the responses averaged by the neuron j' in the trial across the test images.

This formulation removes the contribution of noise from both the numerator and denominator, isolating the variance that is in principle explainable. A perfect model yields $\mathrm{FEV}_j = 1$, while a model no better than noise yields $\mathrm{FEV}_j = 0$.

Since the FEV is computed per neuron, mean FEV is taken as a summary of regression performance.

4.10 Assessing representational similarity between DCNN activations and neural responses

Representational geometry alignment between DCNN activations and neural responses was assessed using Representational Similarity Analysis (RSA). For each DCNN layer, representational dissimilarity matrices (RDMs) were computed across image stimuli using pairwise correlation distances. Neural RDMs were derived from either the full population or a specified PC. DCNN activations were optionally reduced via PCA (10–500 PCs) prior to RDM computation. DCNN and neural RDMs were vectorised (upper triangle) and compared using Spearman correlation. For single-PC analyses, a stable Spearman implementation was applied to avoid undefined values. A fixed random seed ensured reproducibility. This procedure yielded layer-wise estimates of representational similarity across multiple feature dimensionalities.

4.11 Dimensionality analysis of DCNN and neural representations

To estimate the effective dimensionality of visual representations, PCA was applied to DCNN activations, and neural responses. For each layer and the neural data, cumulative explained variance (EV) was computed, along with the proportion of components required to explain 68% EV (chosen as this was the lowest maximum cumulative variance reached by any DCNN layer - SimCLR layer1 - when using all available components).

Dimensionality under a fixed 128-dimensional bottleneck was assessed by drawing 20 random subsamples per layer, performing PCA, and averaging the number of PCs needed to reach 90% EV.

Analyses were performed separately for each mouse and averaged across animals. Features were flattened (samples × features) and not standardised, as scaling was comparable by design.

4.12 Estimating the eigenspetrum decay rate

α was quantified as the decay rate of the eigenspectrum of stimulus responses, following Stringer et al. (2019). For mouse visual cortex, trial-averaged responses were split into two repeats; cvPCA provided unbiased eigenvalues by projecting PCs from one repeat onto the other. For DCNNs, standard PCA was applied to stimulus-by-unit response matrices. In both cases, eigenvalues were fit with a power law $\lambda_k \propto k^{-\alpha}$ over ranks $k = 11$–500 using log–log OLS regression. Model fits used linearly spaced indices; neural fits used log-spaced indices. Eigenvalues below five times the estimated noise floor (median tail) were excluded.

4.13 Assessing semantic clustering compactness

Semantic structure was assessed using t-distributed stochastic neighbor embedding (t-SNE) and silhouette scores for DCNN activations and neural responses. t-SNE was used instead of UMAP due to its stronger emphasis on local structure.

For STL-10, DCNN activations from the test set were labeled with ground-truth classes. Activations from each layer were reduced by PCA (10–500 PCs), followed by 2D t-SNE (perplexity = 30). Silhouette scores were computed on the embeddings using class labels.

For mouse stimuli, 468 manually annotated images (78 per class across 6 categories) were sampled to balance class sizes. t-SNE embeddings were computed for DCNN activations using the same procedure. Trial-averaged neural responses (500 neurons) were reduced to 100 PCs before t-SNE. Semanticity was quantified as silhouette score using the manual labels.

All analyses used a fixed random seed. For cross-layer comparisons, silhouette scores were taken from activations standardised to 100 PCs.

4.14 Visualising DCNN layerwise representational profiles

The joint profile of each DCNN layer was plotted (FEV, RSA, α, and semanticity). Radar plots were generated for SimCLR and VGG-19, each evaluated on the full neural population and PC1. For each case, mean metric values across mice (n=3) were computed per layer and min–max normalised across layers, preserving profile shape while removing scale differences.

4.15 Regression analysis: representational properties underlying neural predictivity

Ordinary least squares (OLS) regression was used to test which representational properties best explained neural predictivity. Analyses were run separately for SimCLR and VGG-19, each evaluated on (i) the full neural population and (ii) PC1. For each case, per-mouse, per-layer datasets (5 layers $\times$ 3 mice = 15 datapoints) were compiled with three candidate predictors: Spearman correlation, α (the eigenspectrum decay rate), and silhouette score on ecologically relevant stimuli. Mean FEV served as the dependent variable. Eight models were fit per case, covering all predictor subsets plus a null model. Predictor collinearity was assessed via pairwise correlations, revealing moderate correlation between α and semanticity. Regressions were implemented in Python (statsmodels) and evaluated using adjusted R^2 and Akaike Information Criterion (AIC).

4.16 Visualising features encoded by DCNN layers

A ridge regression model was fit for each layer to map DCNN activations to PC1 scores of mouse visual cortex activity. Synthetic images were then optimised to maximise predicted PC1 scores. Starting from a random grayscale image, the Adam optimiser (200 iterations) updated only the image, with DCNN weights frozen. For SimCLR, inputs were 96×96, repeated across three channels (learning rate 0.05, weight decay 1e6); for VGG-19, inputs were 224×224 with global average pooling (learning rate 0.05, no weight decay). At each step, features were extracted via a forward

hook, scored by matrix multiplication with regression weights, and a loss combining the negative score and L2 pixel penalty (=0.001) was backpropagated. Final images were rescaled to [0,1], upsampled to 192×192, and saved as 8-bit grayscale PNGs.

5 Results

5.1 SimCLR predicts mouse visual cortex neural responses better than VGG-19

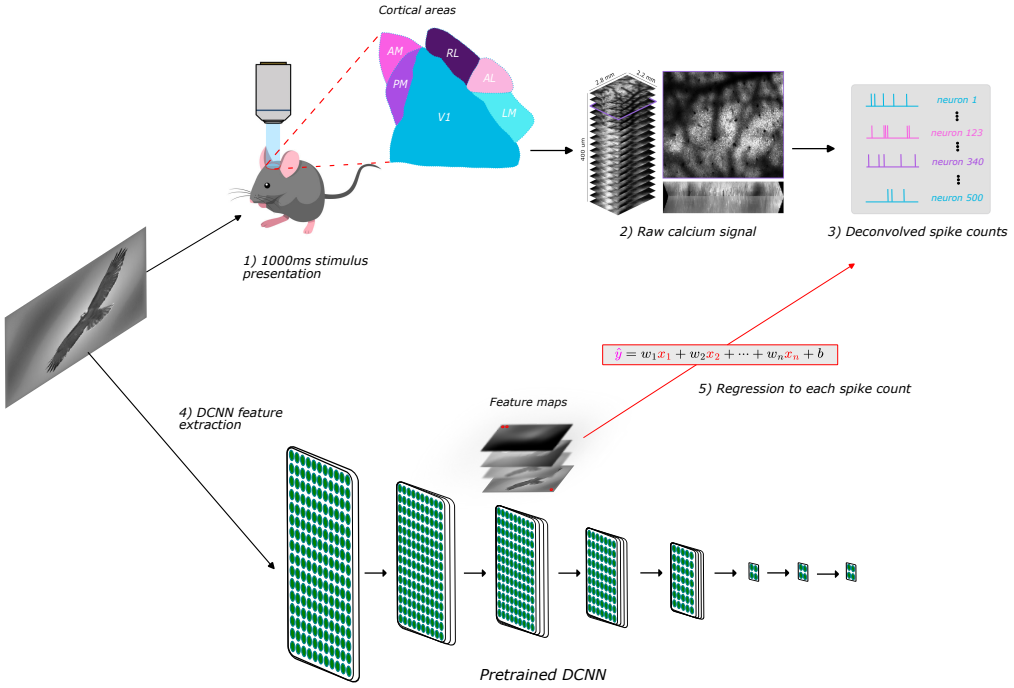
We first evaluated how well DCNN layers predicted neural responses to natural images in mouse visual cortex, using ridge regression to map layer activations to responses. Predictive accuracy was measured as fraction of explainable variance explained (FEV; see Methods 4.9). In parallel, we assessed prediction of the *first principal component (PC1)* of neural responses, which captures the dominant axis of population variation.

Both DCNNs far exceeded Gabor baselines. *SimCLR consistently outperformed VGG-19* across mice, supporting our primary hypothesis. For full-population responses, SimCLR layer2 reached 0.178 ± 0.009 FEV versus 0.121 ± 0.009 for VGG-19 conv4_4, a median paired gain of 0.074 (95% CI: 0.068–0.080). For PC1, SimCLR layer3 achieved 0.53 ± 0.04 compared to 0.39 ± 0.04 for VGG-19 conv2_2, an improvement of 0.127 (95% CI: 0.108–0.152). Although the small sample ($n = 3$) limited formal inference (minimum two-sided $p = 0.25$), the effect was consistent across animals.

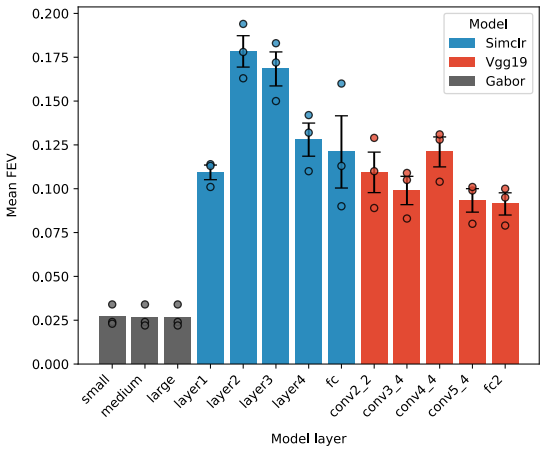
For both networks, *intermediate layers were most predictive*, consistent with our secondary hypothesis. For population responses, SimCLR layers 2–3 and VGG-19 conv4_4 outperformed shallow and deep layers. For PC1 (Figure 1c), SimCLR layers 2–3 and VGG-19 conv2_2, conv3_4, and conv4_4 were best.

Overall, these results show that SimCLR representations—especially at intermediate layers—better align with mouse visual cortex than those of category-supervised VGG-19, supporting the view that self-supervised objectives may be more biologically plausible for mouse vision.

(a) Neural encoding framework



(b) Prediction: DCNN → **Neural**



(c) Prediction: DCNN → **Neural PC1**

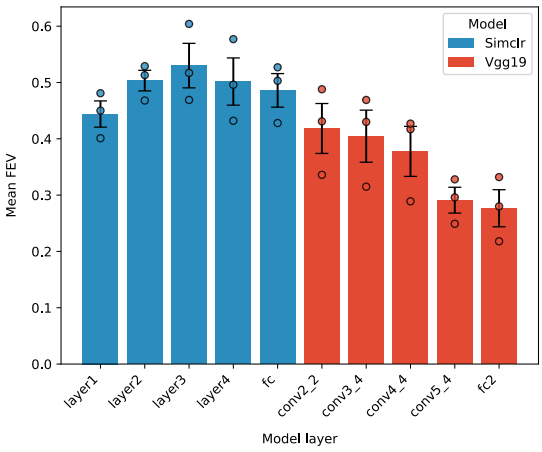


Figure 1: DCNN performance in predicting neural responses. Each bar represents the mean FEV across mice for the specific DCNN layer (SimCLR in blue, VGG-19 in red, Gabor baseline in grey). Error bars show the standard error of the mean (SEM), and dots represent individual mice (n = 3). (a) Neural encoding framework. (b) Mean FEV for predicting full neural population responses. (c) Mean FEV for predicting PC1 of neural responses.

5.2 SimCLR intermediate layers show weak semantic clustering most similar to mouse visual cortex

Having established which DCNN layers best predict neural responses, we next asked how their internal representations align with mouse visual cortex. To do this, we built a representational profile of each layer spanning several properties. We began with semanticity, defined as the degree to which a representational space clusters stimuli by semantic category (e.g. "cat," "car," "airplane"). This measure is particularly relevant here, since supervised DCNNs are explicitly trained to separate categories, whereas self-supervised models—and mouse visual cortex—may not exhibit such clustering.

To assess semantic clustering, we first visualised representations with t-SNE. For STL-10 images, VGG-19 fc2 produced clear category-separated clusters (Figure 2a), whereas SimCLR fc showed noisier but still discernible separation (Figure 2b). To quantify these trends, we then computed silhouette scores, which measure clustering quality by comparing within-cluster cohesion to between-cluster separation (range –1 to 1). Scores increased with depth in both models, peaking at VGG-19 fc2 (Figure 2e), confirming that semantic clustering emerges hierarchically and is stronger in the supervised network.

We next assessed ecologically relevant images—those shown to mice—in both DCNN layers and neural data. Clustering was far weaker here, for mouse visual cortex (Figure 2d) as well as DCNN representations (example layer, Figure 2c). Unlike STL-10, silhouette scores were negative across all sources (Figure 2f), indicating poor alignment with human-defined semantic categories. Neural scores were closest to SimCLR layers 2–4, suggesting that mid-level self-supervised representations best approximate the weak clustering present in mouse visual cortex.

Overall, semantic clustering emerges in deeper layers of both supervised and self-supervised models for standard datasets but is largely absent for naturalistic stimuli. Mid-level SimCLR layers show the closest correspondence to neural data, though this similarity should be interpreted cautiously given the domain mismatch between training images (colour, labelled) and test images (greyscale, unlabelled, ecologically relevant).

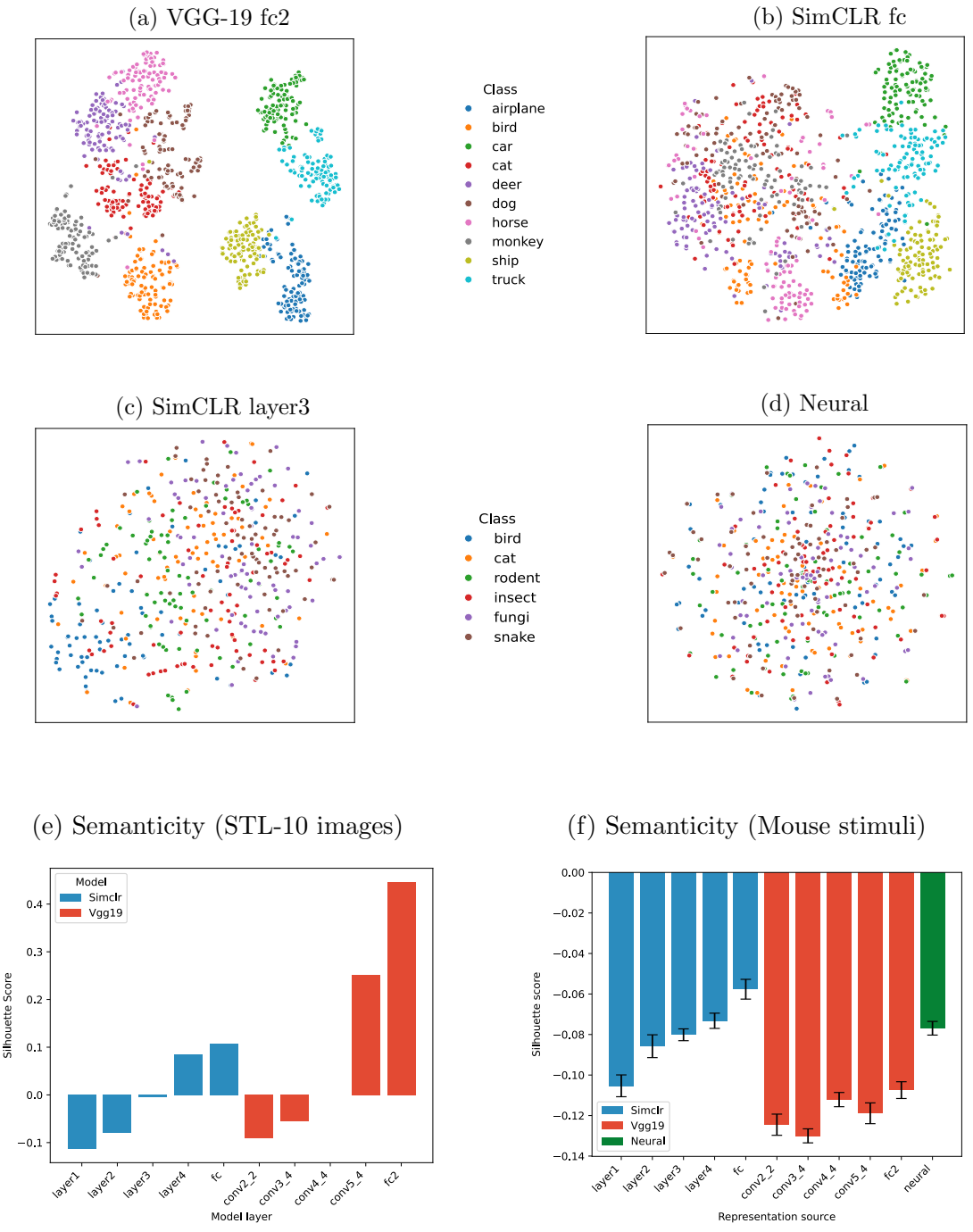


Figure 2: Semantic clustering of images by DCNN layers and the neural population. (a) t-SNE plot showing clustering of STL-10 images (not seen by mice) by VGG-19 fc2. Each dot represents a single image. (b) t-SNE plot: clustering of STL-10 images by SimCLR fc. (c) t-SNE plot showing clustering of ecologically relevant images (shown to mice), by SimCLR layer3. (d) t-SNE plot showing clustering of ecologically relevant images (shown to mice), by the neural population (Mouse 1). (e) Silhouette scores for layer activations evoked by the STL-10 image dataset (Coates et al., 2011). Each bar represents the value for a specific model layer (SimCLR in blue, VGG-19 in red). (f) Silhouette score for layer activations evoked by ecologically relevant images shown to mice (Stringer et al., 2019). Each bar represents the mean value across mice (n = 3) for that model layer (SimCLR in blue, VGG-19 in red, neural population in green). Error bars show the standard error of the mean (SEM).

5.3 SimCLR intermediate layers most closely approximate the dimensionality of mouse visual cortex

Because semantic clustering depends on the underlying representational capacity, we next examined dimensionality, i.e. how many axes of variation the code affords. Supervised DCNNs trained for object recognition are thought to collapse variation into a low-dimensional space aligned with category axes. Biological visual codes, however, may retain higher-dimensional structure. We therefore examined dimensionality directly.

We used two complementary metrics. First, effective dimensionality (ED). We measured this as the of principal components (PCs) needed to explain 90% of variance (PC90) - reflecting the dominance of components at the head of the eigenspectrum. Second, the power-law exponent (α), reflecting the contribution of the eigenspectrum tail.

SimCLR layers consistently required more PCs than VGG-19 (Figure 3a), indicating higher ED. Cumulative variance plots confirmed this compression: SimCLR saturated faster than neural data (Figure 3c), and VGG-19 faster still (Figure 3d).

The neural eigenspectrum had a mean α of 0.85 ± 0.05 SEM (Figure 3e), close to the $\alpha \approx 1$ reported previously in mouse V1 (Stringer et al., 2019). In SimCLR, α increased with depth, consistent with the observations of Elmoznino & Bonner (2024), with layer3 - the most predictive layer (Figure 1c) - showing an exponent (0.96 ± 0.002) closest to the neural mean (Figure 3f).

This value also lies near the theoretical critical boundary ($\alpha \approx 1$) at which codes are simultaneously expressive (capturing fine stimulus distinctions), smooth (similar stimuli evoke similar responses), and compressed (information concentrated into fewer dimensions) (Stringer et al., 2019).

In summary, both DCNNs largely underestimate the full dimensional complexity of mouse visual cortex. However, these results indicate that SimCLR representations are more distributed than the highly compressed features of VGG-19. Strikingly, SimCLR layer3 closely approximates both the ED (PC90) and α of mouse V1.

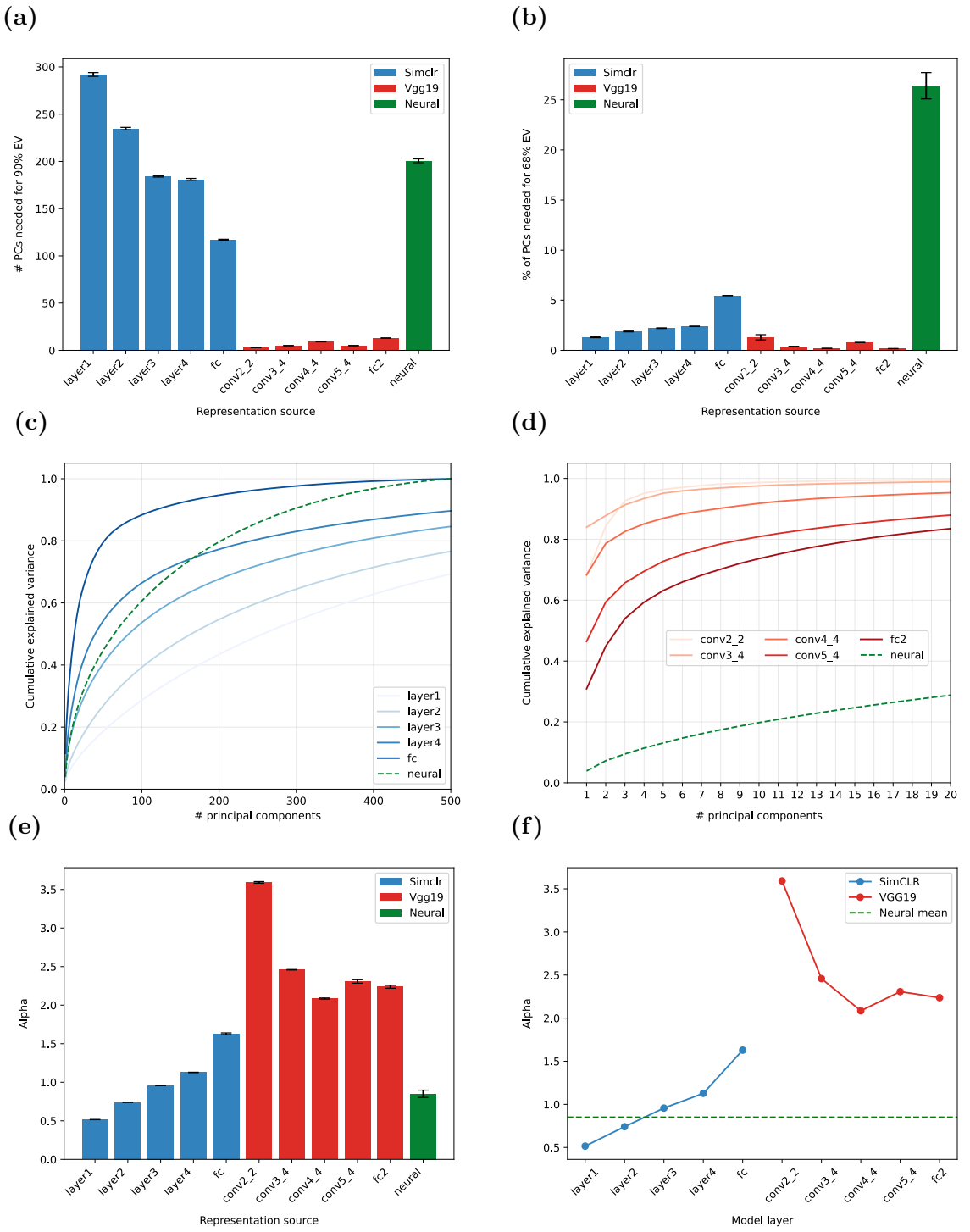


Figure 3: Comparison of representational dimensionality: DCNN layers vs. mouse visual cortex. Each bar represents the mean value across mice (n = 3) for a specific DCNN layer (SimCLR in blue, VGG-19 in red, neural population in green). Error bars show the standard error of the mean (SEM). The dashed green line shows the mean neural α. (a) PC90: Percentage of principal components (PCs) needed to explain 90% of the variance in a 128-dimensional subsample of each layer's features, averaged across mice. (d) PC68: Percentage of total PCs required to explain 68% of the variance using the full dimensionality of each layer's representations. We use 68% as the comparison threshold because this is the lowest maximum cumulative variance reached by any DCNN layer (SimCLR layer1), even when using all available components. (c) Cumulative explained variance as a function of the number of PCs for SimCLR layers and neural data, using full dimensionality. (d) Cumulative explained variance for VGG19 layers and neural data, using the full dimensionality (up to 20 PCs shown for comparability). (e) Power-law exponent (α) for each representation source, quantifying the decay rate of the eigenspectrum. (f) Same α values as in (e), with pointed lines plotted per DCNN (each point represents a layer of the network) to show trends across SimCLR (blue) and VGG-19 (red) hierarchies.

5.4 SimCLR intermediate layers align more closely with the representational geometry of mouse visual cortex

Dimensionality and semanticity probe specific aspects of representational geometry: one captures variance structure, the other category structure. We next examined the global arrangement of stimuli in DCNN latent spaces using representational similarity analysis (RSA), which measures how well pairwise distances between model features align with neural responses.

SimCLR consistently matched neural geometry more closely than VGG-19 (Figure 2b). Its best layer (layer3) reached 0.213 ± 0.061, compared with 0.146 ± 0.026 for VGG-19 (conv3_4). SimCLR showed stable alignment across layers, whereas VGG-19 declined in deeper layers. This perhaps reflects its shift toward category-specific abstraction, as we showed in Figure 2e.

For PC1 (Figure 2c), the dominant axis of variance, SimCLR layer3 achieved 0.141 ± 0.021 versus 0.139 ± 0.047 for VGG-19 conv3_4. Because PC1 is one-dimensional, its RDM captures distances only along a "line". Thus many independent axes of variation (orientation, contrast, spatial frequency, etc.) are discarded, leading to higher variance, and limiting confidence in the comparison.

In sum, SimCLR, in particular its intermediate layers, not only predicts neural responses more accurately but also better captures their representational geometry. This is most clear for the full-population response. This suggests that self-supervised learning yields feature spaces more aligned with mouse visual cortex.

(a) RSA: CNN ↔ **Neural** **(b)** RSA: CNN ↔ **Neural PC1**

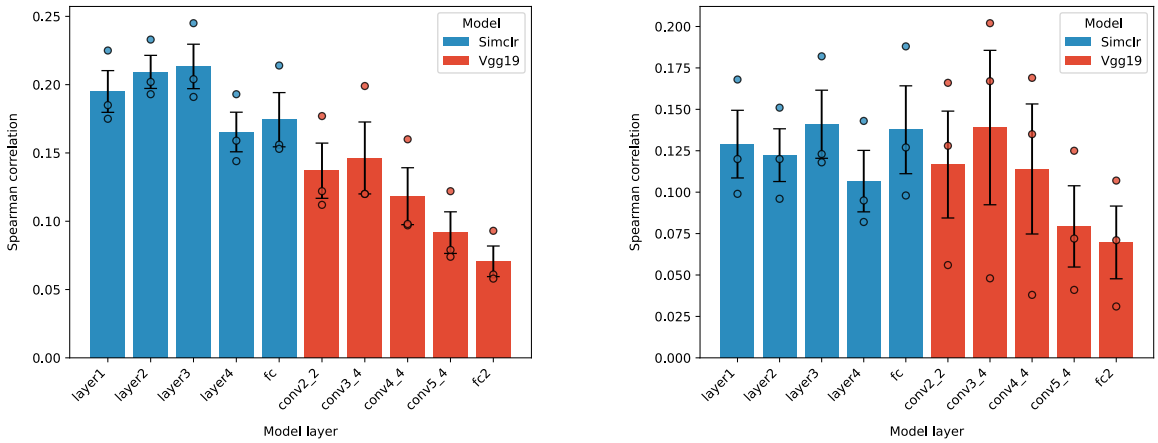


Figure 4: Model alignment with representational geometry of neural responses. Each bar represents the mean Spearman correlation across mice for a specific DCNN layer (SimCLR in blue, VGG-19 in red). Error bars show the standard error of the mean (SEM), and dots represent individual mice (n = 3). (a) Representational similarity analysis schematic. (b) Spearmann correlation between DCNN layer activations and full neural population responses. (c) Spearmann correlation between layer activations and PC1 of neural responses.

5.5 Individual metrics fail to explain neural predictivity

Neural predictivity implies a functional isomorphism, at the level of nonlinear transformations, between a DCNN layer and the neural population. This suggests that predictive layers capture key aspects of the brain's coding strategy. Previous results established that SimCLR intermediate layers were the best neural predictors (Figure 1b,c) and most closely matched mouse visual cortex in semanticity (Figure 2f) and dimensionality (Figure 3a,f). We therefore asked which representational properties confer this predictivity. Specifically, I tested whether any single metric—representational similarity (RSA), power-law exponent (α), or semanticity (silhouette score)—could account for neural predictivity (FEV) of the full population and PC1.

Scatter plots (Figure 5) revealed weak and inconsistent associations between FEV and these metrics (ρ = –0.3 to 0.6, all n=5, p¿0.2). RSA (Fig 5a,b) showed the strongest positive trends (up to ρ=0.6), while α and semanticity hovered near zero or negative. None reached significance, reflecting limited sample size.

These findings suggest that neural predictivity is not explained by a single property, but rather reflects a multidimensional representational profile.

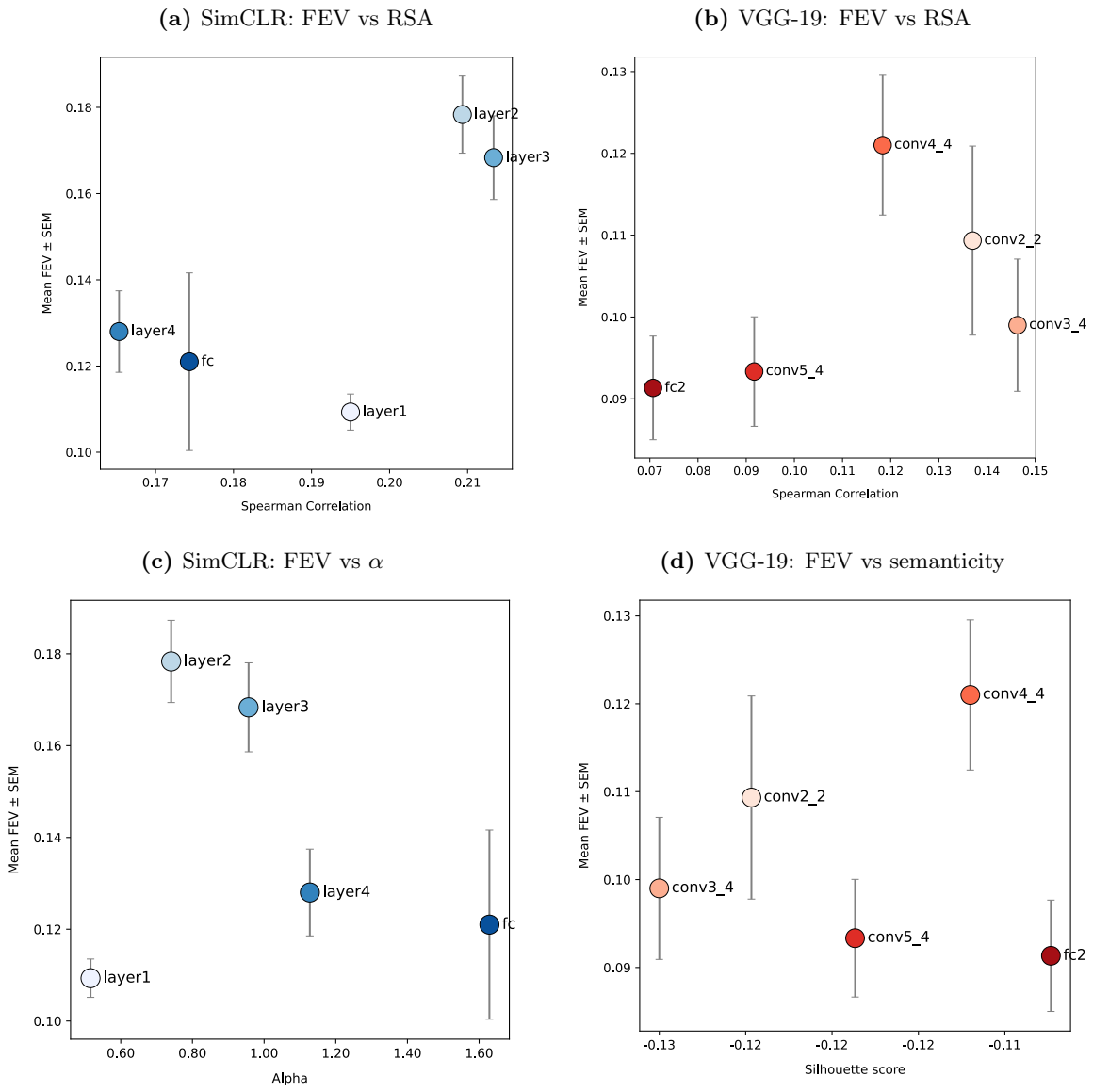


Figure 5: Neural predictivity does not correlate with individual representational metrics. Example scatter plots. (a) FEV vs. RSA (SimCLR). (b) FEV vs. RSA (VGG-19) (c) FEV vs. α (SimCLR).. (d) FEV vs semanticity (VGG-19; ecologically relevant stimuli). All metrics measured against the full neural population; relationships were similar when considering the first neural principal component (PC1), and are omitted here for clarity. Each point represents the mean FEV and metric value for a specific DCNN layer across mice ($n = 3$). Error bars indicate SEM for FEV only. While intermediate layers of SimCLR achieve the highest FEV, no single metric reliably explains layerwise predictivity.

5.6 DCNN layers most predictive of mouse visual cortex exhibit a balance of geometric similarity, moderate dimensionality, and low semanticity

First, we assessed the neural predictivity of each DCNN layer. Then, we went beyond prediction, gathering metrics to build a comprehensive representational profile of each layer. These qualitative profiles (Figure 6a,b) suggest that the most predictive layers exhibit a balance of high neural–geometric similarity, low semanticity, and moderate dimensionality—a pattern most evident in SimCLR intermediate layers.

We next asked whether a combination of representational properties could explain the predictivity of a given DCNN layer. To test this, we performed ordinary least squares (OLS) regressions using RSA, α, and semanticity as predictors of layerwise FEV. For VGG-19, the best model combined RSA and semanticity (Adj. $\mathbf{R^2} = 0.39$), consistent with its category-driven training objective; RSA alone explained less variance (0.23), and α or semanticity alone performed worse than null. For SimCLR, RSA alone provided the strongest fit (Adj. $\mathbf{R^2} = 0.37$), with α and semanticity adding no value due to their high collinearity ($r = 0.89$). For VGG-19, correlations were more moderate (e.g. RSA–semanticity $r = –0.46$). Thus, VGG-19 predictivity reflects both geometric alignment with mouse visual cortex, and semantic clustering. SimCLR appears to be driven almost entirely by geometric alignment. These trends should be interpreted with caution given the limited sample size and predictor collinearity, which constrain statistical power.

Nonetheless, they provide a qualitatively useful description. Across models, the DCNN layers that best predicted mouse visual cortex exhibit a characteristic profile: high neural-geometric alignment, low semanticity, and moderate dimensionality. This co-occurrence hints at a signature of biologically aligned representations and motivates a deeper question: what visual features define this geometry?

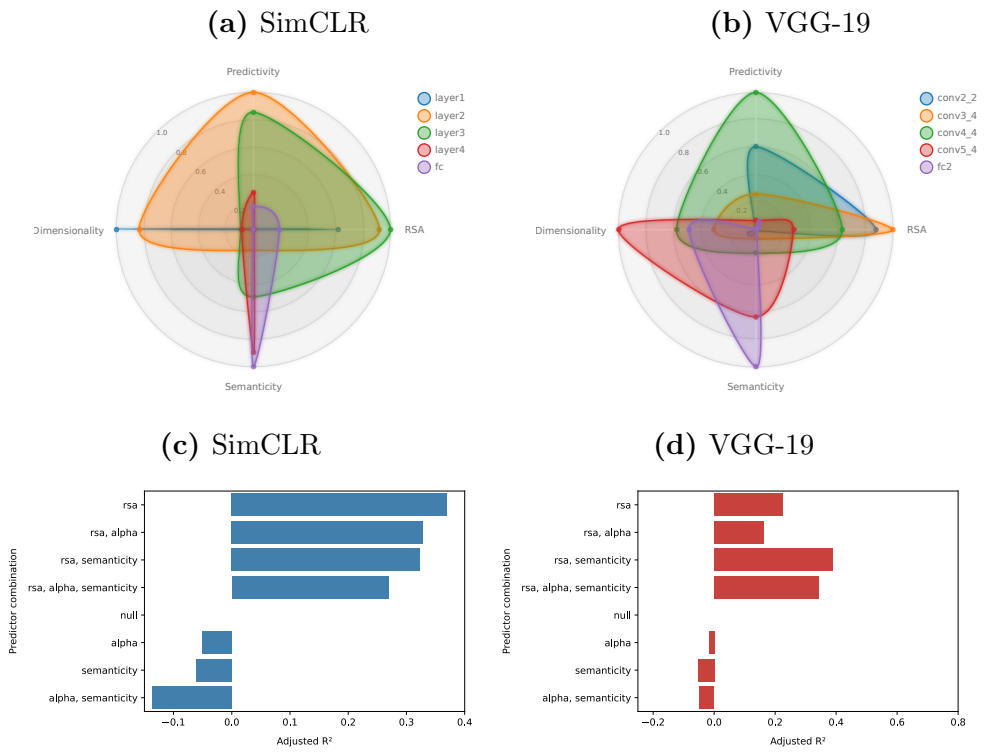


Figure 6: The representational profiles of DCNN layers reveal properties underlying neural predictivity. For each DCNN, we computed four metrics per layer: neural predictivity (FEV), geometric similarity (RSA), dimensionality (α), and semanticity. Radar plots (a-d) show each layer's profile across these metrics, min–max normalised within model to visualise relative differences. Regression models (c&d) used ordinary least squares (OLS) on per-mouse, per-layer values, treating each mouse-layer combination, on a given metric (e.g. RSA for Mouse1 vs. SimCLR layer3), as a datapoint. Each bar represents a different combination of predictors.

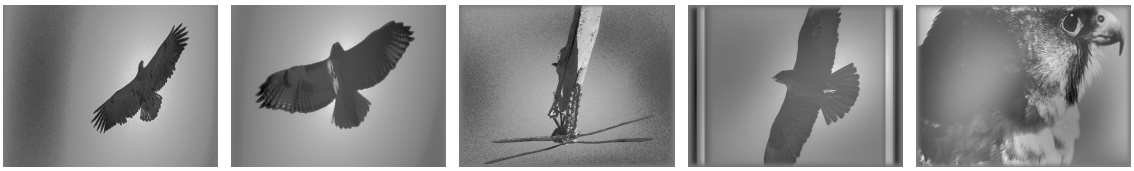
5.7 SimCLR intermediate layers capture mid-level features aligned with mouse visual cortex

We first identified the model layers most predictive of mouse visual cortex and then characterised their representational profiles in terms of semanticity, dimensionality, and neural-geometric alignment. Across these measures, SimCLR—particularly its intermediate layers—showed closer biological alignment than VGG-19. SimCLR layer3 predicted neural PC1 responses better than any other layer (Figure 1c). It also displayed effective dimensionality comparable to mouse cortex (Figure 3a), an eigenspectrum decay rate (α) closely matching neural data (Figure 4f), and, together with layer4, the strongest similarity in semantic clustering. Consistently, its representational geometry was also the most similar to mouse visual cortex (Figure 4b).

What visual features does such a DCNN layer encode? To help here, we utilised feature visualisation, which optimises an image to maximise predicted neural (PC1) responses through a given model layer. This provides an intuitive window into what that layer "thinks" is most salient to mouse visual cortex. Our analysis confirmed that SimCLR layer3 produces synthetic images resembling natural PC1-driving stimuli, capturing textures, spatial contrasts, and orientation structure (Figure 8d). Earlier layers lacked interpretable features (Figure 8b), while deeper layers (Figure 8e,f) produced more abstract patterns. VGG-19 showed a similar pattern of hierarchical abstraction, but its synthetic images were less naturalistic (Figure 8g-k). This reinforces that mouse visual cortex emphasises mid-level visual features rather than high-level semantic categories.

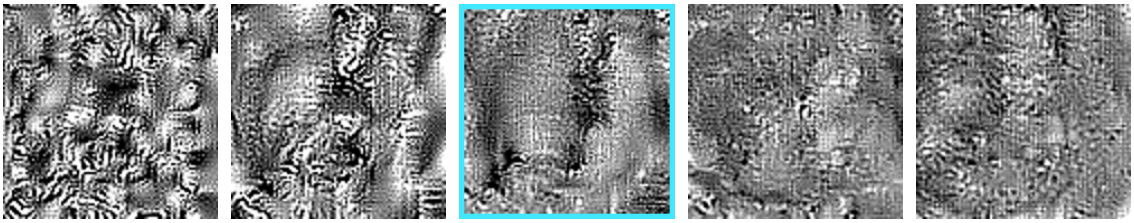
Taken together, this qualitative evidence complements our quantitative results, suggesting that SimCLR layer3—and by extension mouse visual cortex—may indeed be tuned to mid-level visual features such as textures, spatial contrasts, and orientation structure. While this conclusion should be drawn cautiously, the convergence of predictive accuracy, representational alignment, and feature visualisation strengthens the case for mid-level feature selectivity as a signature of rodent vision.

(a) Mouse 1: Top PC1-activating natural stimuli

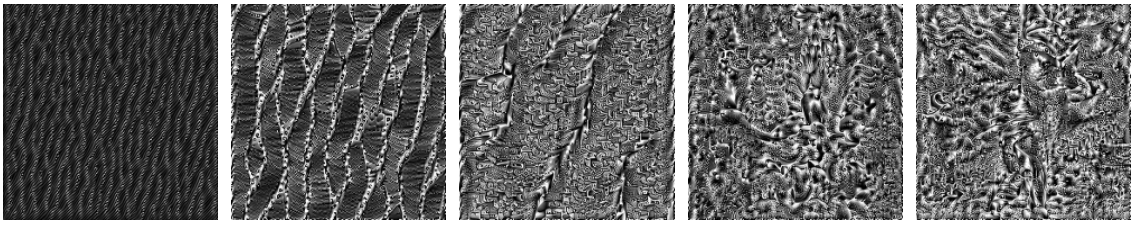


Mouse 1: Synthetic images maximising PC1 activation

(b) layer1 **(c)** layer2 **(d)** *layer3* **(e)** layer4 **(f)** fc



(g) conv2_2 **(h)** conv3_4 **(i)** conv4_4 **(j)** conv5_4 **(k)** fc2



SimCLR layer3: natural and synthetic images activating PC1

(l) M2: natural **(m)** M2: synthetic **(n)** M3: natural **(o)** M3: synthetic

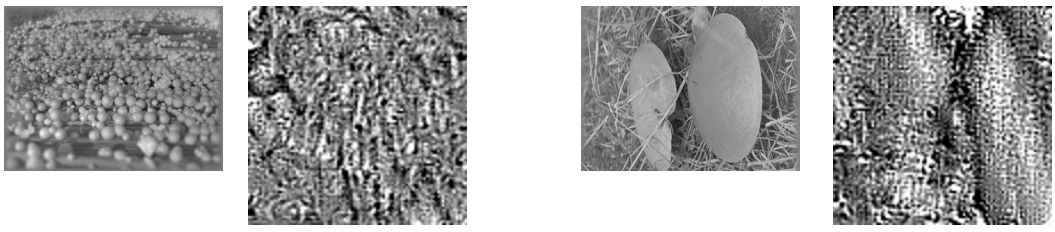


Figure 7: SimCLR layer3 synthetic images resemble natural PC1 drivers. Synthetic images maximising predicted PC1 responses differ by layer because optimisation is constrained by each layer's representational basis. They should be interpreted not as direct readouts of neural selectivity, but as model-based hypotheses of PC1 feature preferences filtered through that basis. (a) Top PC1-activating natural images for Mouse 1. (b–f) Synthetic images maximising PC1 activation from SimCLR layers 1–fc (Mouse 1). (g–k) Synthetic images maximising PC1 activation from VGG-19 layers conv2_2–fc2 (Mouse 1). (l) Top PC1-activating natural image (mouse 2) (m) SimCLR layer3 synthetic image maximally activating PC1 (mouse 2) (n) Top PC1-activating natural image (mouse 3) (o) SimCLR layer3 synthetic image maximally activating PC1 (mouse 3)

6 Discussion

6.1 From primate object recognition to mouse visual processing: rethinking the learning objective

Self-supervised learning is a promising framework for modelling visual cortices, yet its fidelity beyond primates has not been extensively studied. Here, we tested the hypothesis that SimCLR (with a contrastive objective) would outperform a category-supervised VGG-19 in predicting mouse visual cortex responses, and that intermediate layers would be most predictive. Our results confirmed both: SimCLR achieved higher prediction accuracy, and intermediate layers performed best. This adds weight to emerging evidence that self supervised DCNNs outperform supervised counterparts. It also suggests that models must match mouse biology in architecture and ethological demands in objective.

Table 1: Summary of studies using computational models to predict neural responses in visual cortex, grouped by species and training objective. Direct comparison of predictive accuracy across studies is not straightforward due to differences in recording modality, brain region coverage, stimulus type, dataset size, regression method, and evalutation metric. Instead the focus here is on the relative predictive performance afforded by self-supervised vs. supervised training methods

Study	Training	Testing	Result
Mouse			
Nayebi et al. (2023)	Self-supervised	Learning objective	Contrastive training superior to supervised and other self-supervised
Bakhtiari et al. (2021)	Self-supervised	Learning objective	Contrastive Predictive Coding superior to supervised ResNet (ventral areas only; natural videos)
Wang et al. (2025)	Supervised	Visual diet	Much larger, pooled dataset yielded 25–46% higher predictive accuracy vs. state-of-the-art (natural videos)
Cadena et al. (2019)	Supervised	Learning objective	A VGG16 CNN trained on ImageNet outperformed a classical Gabor filter bank model in predicting mouse visual cortex responses
Macaque			
Zhuang et al. (2020)	Self-supervised	Learning objective	Models with a self supervised (contrastive) learning objective predict macaque V1, V4 and IT as well as supervised ResNet-18
Lotter et al. (2020)	Self-supervised	Architecture	Computing explicit error signals at each layer improves predictivity of model trained for prediction-error minimisation (natural videos)
Higgins et al. (2021)	Self-supervised	Learning objective	A disentangling objective yielded the best predictive accuracy for macaque IT, vs supervised and other self-supervised objectives

Study	Training	Stimulus	Result
Yamis et al. (2014)	Supervised	Learning objective	Supervised CNNs optimised for invariant object recognition predicted macaque IT cortex responses better than baseline (V1- and V2-) models
Yamins & DiCarlo (2016)	Supervised	Learning objective	HCNNs trained for object categorization predict neural responses in the ventral stream without direct neural fitting, outperforming baseline models including V1- and V2-like feature models
Cadena et al. (2019)	Supervised	Learning objective	Supervised VGG-19 predicts macaque neural responses substantially better than traditional models
Nayebi et al. (2018)	Supervised	Anatomy	Adding recurrent cells and selective feedback to CNNs improves prediction of neural responses in macaque ventral stream
Human			
Konkle et al. (2022)	Self-supervised	Learning objective	CNNs with a contrastive learning objective predict neural responses in the human ventral visual stream on par with supervised baselines
Güçlü & van Gerven (2015)	Supervised	Learning objective	A supervised CNN predicts fMRI responses to natural images across V1–LO better than a traditional (Gabor) model
St-Yves et al. (2023)	Supervised	Learning objective	A supervised CNN trained on image-neural response pairs predicts fMRI responses to natural images across V1–V4 better than a standard category-supervised CNN
Cichy et al. (2016)	Supervised	Learning objective	DNN (AlexNet) trained for object categorisation predicts human visual cortex responses in space and time in a hierarchical fashion, better than untrained baseline models
Mouse & Macaque			
Du et al. (2025)	Supervised	Architecture	A simple two-layer CNN can predict V1 responses in mouse and monkey V1 nearly as well as deep models
Human & Macaque			
Khaligh-Razavi & Kriegeskorte (2014)	Supervised	Learning objective	A deep CNN trained with 1.2 M labeled images (ImageNet) explained IT representations better than any other model tested (including HMAX, VisNet, SIFT, GIST, and others)

6.2 Intermediate abstraction in DCNNs and mouse visual cortex

Mouse higher areas appear to move beyond edge detection toward mid-level, combinatorial representations, like curvature, texture, boundaries, depth/motion cues (Marshel et al., 2011). Neurons tuned to boundaries or silhouettes may provide rudimentary, distributed precursors of IT-style shape selectivity (Han et al., 2022). Overall, these findings suggest that mouse visual cortex supports mid-level feature representations—richer than the edge detectors of V1, yet short of the high-level object and category selectivity characteristic of the primate ventral stream.

Interpretability studies of hierarchical DCNNs reinforce this framework: early layers encode simple edges and orientations, intermediate layers encode textures and shapes, and late layers category-specific semantic features, closely paralleling the functional hierarchy of the primate ventral stream. We therefore expected that the mid-level layers of SimCLR and VGG-19 would provide the best match to the level of abstraction in mouse visual cortex (V1 and higher visual areas), and correspondingly would yield the strongest neural predictions. This expectation was borne out (Figures 1a&b).

6.3 On the relationship between predictivity and representational isomorphism

Why care about a model's ability to predict neural responses? Models are aligned to biology by tuning architecture, learning objective, and visual diet, then assessed by behavioural performance and neural predictivity. If neural responses can be explained by a linear transformation of model features, the representations may share functional properties with the brain. Yet this logic has limits.

High predictivity does not imply convergent computations. In mouse visual cortex, Cadena et al. (2019) found that a VGG-16 trained on object recognition did not outperform an untrained network, whose convolutional filters already generated a feature space linearly mappable to neural responses. Crucially, unlike in primates, no hierarchical correspondence between layers and areas was observed. Predictivity can therefore arise from architectural priors rather than biological alignment.

Training objectives further complicate interpretation. As Zhuang et al. (2021) note, self-supervised methods often learn high-level features through proxy tasks, which may capture relevant scene variables but also embed task-specific confounds. Different regimes can yield overlapping feature sets that blur inference.

As Yamins & DiCarlo (2016) emphasise, the goal is to uncover the nonlinear transformations across the visual hierarchy. To establish representational isomorphism, however, we must probe geometry, dimensionality, and semantic structure. This study therefore complements neural predictivity with a systematic assessment of representational profiles.

6.4 Beyond neural predictivity: signatures of representational alignment with mouse visual cortex

6.4.1 Semantcity: weak category clustering

In primates, invariant object recognition depends on inferotemporal (IT) cortex, which encodes robust category clusters (Hung et al., 2005; McKee et al., 2014). These correspond to the deepest, fully connected layers of CNNs trained on ImageNet (DiCarlo et al., 2012), and late VGG-19 layers likewise cluster stimuli by semantic category (Figures 2a&e).

Rodents show a more limited form of invariance, generalising across transformations such as size and rotation (Zoccolan et al., 2009; Tafazoli et al., 2017; Vinken et al., 2021). This relies on extrastriate cortex rather than V1, which is tuned to low-level features (Girman & Lund, 1999; Niell & Stryker, 2008; Ringach et al., 2016) and unlikely to support categorical coding (Zoccolan

et al., 2009). Accordingly, rodent neural data are best captured by low–mid CNN layers (Vinken & Op de Beeck, 2021), reflecting a less abstract, non-semantic representational regime.

Consistent with this, I observed low semantic clustering in mouse V1 and extrastriate areas (Figure 2d). The most predictive SimCLR layers (2–4) matched this low semanticity (Figure 2b), whereas layers with strongest clustering (VGG-19 conv3_4, fc2) poorly predicted neural activity (Figures 6a&b). Although semanticity alone did not explain predictivity, its consistent reduction in predictive layers suggests it forms part of the characteristic profile of mouse-aligned representations: mid-level, moderately dimensional, and non-categorical (Figure 6a&b).

6.4.2 Dimensionality: long-tailed, high-dimensional eigenspectra

Neural population responses can be described along orthogonal axes of variation, spanning a latent representational space. In mouse V1, for instance, one axis may encode orientation, another spatial frequency, and another contrast. Recent work has focused on eigenspectrum-derived metrics—particularly the power-law exponent (α) and the number of components needed to explain 90% of variance (PC90). A steep eigenspectrum ($\uparrow \alpha$) concentrates variance in a few dominant dimensions (low PC90), while a shallow spectrum ($\downarrow \alpha$) distributes variance broadly (high PC90). These measures quantify representational dimensionality and reveal signatures of biological vision, such as the long-tailed spectra reported in mouse V1 (Stringer et al., 2019).

CNN studies highlight pitfalls of using predictivity alone. Classification-trained networks compress variance into task-relevant axes, producing low-dimensional, semantically clustered codes (Yamins et al., 2014). By contrast, primate IT and human LO align with high-dimensional eigenspectra (Elmoznino Bonner, 2024), approximating the theoretical $1/i$ decay (Stringer et al., 2019).

Our results suggest mouse vision follows a similar principle. SimCLR intermediate layers—those most predictive of neural responses—matched neural PC90 and α more closely than VGG-19, approximating the $1/i$ decay that defines smooth, maximally expressive representations. Still, despite α-semanticity colinearity (Figure 6c), the most predictive layers combined high RSA with moderate α and low semanticity (Figures 6a&b), indicating dimensionality shapes, but does not fully explain, neural predictivity.

6.4.3 Representational geometry

Of the representational metrics examined, RSA has been most widely applied in mice. Studies consistently show that mid-level layers best align with mouse visual cortex, while deeper layers decline. This holds across diverse architectures—classical CNNs (AlexNet, VGG, ResNet), mouse-specific CNNs (MouseNet), and even Transformers (Nayebi et al., 2023). Conwell et al. (2021) further reported a correspondence between cortical hierarchy and model depth: early layers map to V1, and intermediate layers to higher areas.

Our results follow this pattern. Intermediate layers of both networks aligned best, with SimCLR layer3 showing the highest similarity (Figures 4a&b). VGG-19's deep layers declined in RSA and neural predictivity (Figures 1a&b), consistent with their association with primate IT, which encodes category-based geometries (Hung et al., 2005; Kiani et al., 2007; McKee et al., 2014). By contrast, rodent cortex remains tuned to low- and mid-level features under natural conditions (Marshel et al., 2011; Han et al., 2022), though object recognition can emerge with training (Zoccolan et al., 2009; Vermaercke et al., 2014; Goltstein et al., 2021).

RSA also emerged as the most consistent predictor of neural predictivity (Figures 6c&d), though given the limited statistical power this data should be taken as preliminary.

6.4.4 Feature visualisation

Classical experiments revealed the tuning of single neurons—for example, orientation-selective cells in V1 (Hubel & Wiesel, 1962). In mice, V1 acts as an initial filter for orientations, motion,

spatial scales, contrast, and to a lesser extent color (Niell & Stryker, 2008). More recently, large-scale optical imaging and dense electrophysiology have enabled recordings from tens of thousands of neurons simultaneously, allowing population-level representations to be probed. At this scale, tuning is less obvious: rather than encoding simple variables, neural populations vary along many overlapping axes, reflecting the structure of high-dimensional natural images.

Feature visualisation offers a way forward. Originally developed for individual units and validated by showing synthetic images can drive stronger responses than natural stimuli (Walker et al., 2019), we extend this approach to entire populations. A linear readout from model features to neural PC1 was trained, and gradient ascent generated images maximising predicted responses.

SimCLR layer3 produced synthetic images most similar to natural PC1-driving stimuli (Figure 7d), consistent across mice (Figures 7l-o) and aligned with its superior predictivity (FEV) and RSA. These results suggest mouse cortex prioritises textures, contrasts, and orientation structure—mid-level features tuned to ethological demands rather than semantic categories.

6.5 Implications and future directions

This study shows that self-supervised learning better models mouse vision than category supervision, supporting the view that mouse visual cortex is tuned to ethological demands rather than fine-grained object categorisation. Intermediate SimCLR layers were the most predictive and representationally aligned, suggesting that mid-level features best capture the abstraction of mouse vision. Progress will therefore require training objectives and visual diets tailored to rodent ethology—datasets in the spirit of Stringer et al. (2019), but scaled for modern self-supervised CNNs.

More broadly, predictive accuracy alone is insufficient: models can map onto neural responses without converging on biological representations. Biologically aligned layers showed a consistent signature—high neural-geometric similarity, moderate dimensionality, and low semantic clustering—highlighting the value of representational isomorphism as a criterion for model–brain alignment. Future work should therefore combine geometric, dimensional, and semantic analyses, and experimentally validate population-level feature visualisation. Walker et al. (2017) showed that synthetic images could drive stronger responses than natural stimuli at the single-neuron level; similar validation at the population level would strengthen interpretability.

Finally, this programme extends beyond sensory neuroscience. As multimodal AI systems integrate theories of attention, metacognition, and access consciousness, representational isomorphism offers a principled framework for assessing machine–human equivalence and reverse-engineering higher cognition.

7 Limitations of Methods

This study has several limitations, which are considered in turn below.

7.1 SimCLR and VGG-19 architectures are mismatched

Alignment between artificial and biological vision depends on (at least) three factors: architecture, learning objective, and visual diet. Here, I compare SimCLR (a **ResNet-18 backbone** trained with a *contrastive self-supervised* objective) with VGG-19 (a **deeper feedforward architecture** trained with *category supervision*). As a result, differences in neural predictivity, and other measures of representational alignment, cannot be attributed solely to the learning objective.

Indeed, prior work has shown that architectural priors alone can shape neural predictivity. For example, a simple two-layer CNN predicts V1 responses nearly as well as deep models (Du et al., 2025), and a randomly initialised (untrained) VGG rivals the performance of its trained counterpart in mouse visual cortex (Cadena et al., 2019).

Thus, to isolate the effect of self-supervised versus supervised objectives, future work should train both regimes on matched architectures and diets, enabling a cleaner assessment of learning objective.

Nevertheless, my initial finding—that SimCLR predicts neural responses more accurately than VGG-19—is consistent with prior work. When architecture and visual diet are controlled, contrastive and other self-supervised methods reliably outperform supervised models in neural predictivity, in mouse (Bakhtiari et al., 2021; Nayebi et al., 2023), macaque (Zhuang et al., 2020), and human visual cortex (Konkle et al., 2022). Taken together, these results strengthen the conclusion that the self-supervised learning objective itself confers higher neural predictive accuracy.

7.2 Test stimuli are out-of-distribution

The models analysed here were trained on large, human-labelled datasets: ImageNet (Deng et al., 2009) for VGG-19, and STL-10 (Coates et al., 2011) for SimCLR. STL-10 is a reduced subset of ImageNet, containing far fewer categories (10 vs. 1,000) and images, but like ImageNet it consists of colour photographs of human-relevant object classes. The neural recordings, however, were collected with a different stimulus set, from Stringer et al. (2019). These ethologically relevant images were greyscale, contrast-normalised, and drawn from categories such as birds, cats, rodents, insects, snakes, and food items, with a mix of low- and high-spatial frequencies.

This mismatch creates a challenge for interpretation. On STL-10, both models—especially in their deepest layers—showed clear semantic clustering. On the Stringer stimuli, however, both the models and the mouse visual cortex showed weak clustering. This could reflect a genuine similarity, since mouse visual cortex may not primarily encode the same human-semantic category boundaries. But it could equally reflect poor out-of-distribution generalisation in the models, which fail to transfer category structure from their training sets to these stimuli.

This visual diet mismatch may also subtly affect measures such as FEV and RSA. Models trained on human-centric datasets may emphasise features that are underrepresented in the ethologically relevant stimuli used for mouse recordings. Thus, closer alignment of training and test datasets will be needed to disentangle these explanations.

7.3 Explaining neural predictivity from representational metrics is underpowered

The regression analysis linking neural predictivity (FEV) to representational metrics was limited by both statistical power and conceptual scope. With only five layers per model across three mice (n=15 datapoints per analysis), even moderate associations were unlikely to reach significance. This small sample also inflated uncertainty in the estimated contributions of RSA, α, and semanticity, making it difficult to disentangle their effects, particularly given collinearity between

α and semanticity. Pairwise correlations further showed that no single metric could reliably explain layerwise neural predictivity. While OLS regressions incorporated multiple predictors, they remained underpowered and restricted to a narrow set of representational properties. Moreover, adding more intermediate layers from SimCLR and VGG-19 would likely not meaningfully resolve this limitation. This is because redundancy between adjacent layers could limit any real gain in explanatory power. Larger datasets, additional models, and multivariate approaches will be required to more robustly explain why certain layers predict neural responses more accurately than others.

7.4 Number of mice is limited (n = 3)

This study relied on neural recordings from three mice, which constrains the robustness of the reported findings. I did observe consistent patterns across individuals. Quantitatively, for example, intermediate SimCLR layers aligned most closely with neural responses. Qualitatively, feature visualisation of SimCLR layer3 produced synthetic images that resembled the dominant neural response axis in all three mice. However, the small sample size limits confidence in generalising these results. With more animals, it would be possible to assess inter-individual variability in representational geometry and dimensionality, and to test whether the observed alignment between SimCLR mid-level layers and mouse visual cortex is stable across subjects.

References

Bakhtiari, S., Mineault, P.J., Lillicrap, T., Pack, C.C. Richards, B.A. (2021) The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. Advances in Neural Information Processing Systems (NeurIPS 2021), Spotlight

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput Biol 15(4): e1006897. https://doi.org/10.1371/journal.pcbi.1006897

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A. S., Ecker, A. S. (2019, December). How well do deep neural networks trained on object recognition characterize the mouse visual system? In Real Neurons Hidden Units: Future directions at the intersection of neuroscience and AI (Contributed talk). Workshop at the Neural Information Processing Systems (NeurIPS) 2019, Vancouver, Canada.

Carandini, M., Heeger, D. Normalization as a canonical neural computation. Nat Rev Neurosci 13, 51–62 (2012). https://doi.org/10.1038/nrn3136

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., Rust, N. C. (2005). Do we know what the early visual system does?. The Journal of neuroscience : the official journal of the Society for Neuroscience, 25(46), 10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709.

Cichy, R., Khosla, A., Pantazis, D. et al. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6, 27755 (2016). https://doi.org/10.1038/srep27755

Adam Coates, Honglak Lee, Andrew Y. Ng An Analysis of Single Layer Networks in Unsupervised Feature Learning AISTATS, 2011.

Conwell, C., Prince, J.S., Kay, K.N. et al. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. Nat Commun 15, 9383 (2024). https://doi.org/10.1038/s41467-024-53147-y

J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Du, F., Angel Núñez-Ochoa, M., Pachitariu, M. et al. A simplified minimodel of visual cortical neurons. Nat Commun 16, 5724 (2025). https://doi.org/10.1038/s41467-025-61171-9

Du, F., Angel Núñez-Ochoa, M., Pachitariu, M. et al. A simplified minimodel of visual cortical neurons. Nat Commun 16, 5724 (2025). https://doi.org/10.1038/s41467-025-61171-9

Elmoznino, E. and Bonner, M.F., 2024. High-performing neural network models of visual cortex benefit from high latent dimensionality. PLOS Computational Biology, 20(1), p.e1011792. https://doi.org/10.1371/journal.pcbi.1011792

Girman, S. V., Sauvé, Y., Lund, R. D. (1999). Receptive field properties of single neurons in rat primary visual cortex. Journal of neurophysiology, 82(1), 301–311. https://doi.org/10.1152/jn.1999.82.1.301

Goltstein, P.M., Reinert, S., Bonhoeffer, T. et al. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. Nat Neurosci 24, 1441–1451 (2021). https://doi.org/10.1038/s41593-021-00914-5

Güçlü, U., van Gerven, M. A. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. The Journal of neuroscience : the official journal of the Society for Neuroscience, 35(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Han, X., Vermaercke, B. Bonin, V. Diversity of spatiotemporal coding reveals special-

ized visual processing streams in the mouse cortex. Nat Commun 13, 3249 (2022). https://doi.org/10.1038/s41467-022-29656-z

Harris, J.A., Mihalas, S., Hirokawa, K.E. et al. Hierarchical organization of cortical and thalamic connectivity. Nature 575, 195–202 (2019). https://doi.org/10.1038/s41586-019-1716-z

Haydaroğlu, A., Chang, T., Landau, A., Krumin, M., Dodgson, S., Baruchin, L. J., Cozan, M., Guo, J., Meyer, D., Reddy, C. B., Zhong, J., Ji, N., Schröder, S., Harris, K. D., Vaziri, A., Carandini, M. (2025). Suite3D: Volumetric cell detection for two-photon microscopy. bioRxiv : the preprint server for biology, 2025.03.26.645628. https://doi.org/10.1101/2025.03.26.645628

Higgins, I., Chang, L., Langston, V. et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. Nat Commun 12, 6456 (2021). https://doi.org/10.1038/s41467-021-26751-5

Hung, C. P., Kreiman, G., Poggio, T., DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science (New York, N.Y.), 310(5749), 863–866. https://doi.org/10.1126/science.1117593

Khaligh-Razavi, S.M. and Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLOS Computational Biology, 10(11), p.e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kiani, R., Esteky, H., Mirpour, K., Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of neurophysiology, 97(6), 4296–4309. https://doi.org/10.1152/jn.00024.2007

Konkle, T., Alvarez, G.A. A self-supervised domain-general learning framework for human ventral stream representation. Nat Commun 13, 491 (2022). https://doi.org/10.1038/s41467-022-28091-4

Lotter, W., Kreiman, G. Cox, D. A neural network trained for prediction mimics diverse features of biological neurons and perception. Nat Mach Intell 2, 210–219 (2020). https://doi.org/10.1038/s42256-020-0170-9

Malach, R., Levy, I., Hasson, U. (2002). The topography of high-order human object areas. Trends in cognitive sciences, 6(4), 176–184. https://doi.org/10.1016/s1364-6613(02)01870-3

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco: W.H. Freeman.

Marshel, J. H., Garrett, M. E., Nauhaus, I., Callaway, E. M. (2011). Functional specialization of seven mouse visual cortical areas. Neuron, 72(6), 1040–1054.

McKee, J. L., Riesenhuber, M., Miller, E. K., Freedman, D. J. (2014). Task dependence of visual and category representations in prefrontal and inferior temporal cortices. The Journal of neuroscience : the official journal of the Society for Neuroscience, 34(48), 16065–16075. https://doi.org/10.1523/JNEUROSCI.1660-14.2014

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J.J. and Yamins, D.L.K., 2018. Task-driven convolutional recurrent models of the visual system. arXiv preprint arXiv:1807.00053.

Nayebi, A., Kong, N.C.L., Zhuang, C., Gardner, J.L., Norcia, A.M. and Yamins, D.L.K., 2023. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. PLOS Computational Biology, 19(10), p.e1011506.

Niell, C. M., Stryker, M. P. (2008). Highly selective receptive fields in mouse visual cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience, 28(30), 7520–7536. https://doi.org/10.1523/JNEUROSCI.0623-08.2008

Prusky, G. T., West, P. W., Douglas, R. M. (2000). Behavioral assessment of visual acuity in mice and rats. Vision research, 40(16), 2201–2209. https://doi.org/10.1016/s0042-6989(00)00081-x

Putnam, H., 1967. Psychological predicates. In: W.H. Capitan and D.D. Merrill, eds. Art, Mind, and Religion. Pittsburgh: University of Pittsburgh Press, pp.37–48.

Ringach, D. L., Shapley, R. M., Hawken, M. J. (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. The Journal of neuroscience : the official journal of the Society for Neuroscience, 22(13), 5639–5651. https://doi.org/10.1523/JNEUROSCI.22-13-

05639.2002

Schmutz, V., Haydaroglu, A., Wang, S., Feng, Y., Carandini, M., Harris, K. D. (2025). High-dimensional neuronal activity from low-dimensional latent dynamics: a solvable model. bioRxiv : the preprint server for biology, 2025.06.03.657632.

Serre, T., Oliva, A., Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America, 104(15), 6424–6429. https://doi.org/10.1073/pnas.0700622104

Siegle, J.H., Jia, X., Durand, S. et al. Survey of spiking in the mouse visual system reveals functional hierarchy. Nature 592, 86–92 (2021). https://doi.org/10.1038/s41586-020-03171-x

Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

St-Yves, G., Allen, E.J., Wu, Y. et al. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. Nat Commun 14, 3329 (2023). https://doi.org/10.1038/s41467-023-38674-4

Stringer, C., Pachitariu, M., Steinmetz, N. et al. High-dimensional geometry of population responses in visual cortex. Nature 571, 361–365 (2019). https://doi.org/10.1038/s41586-019-1346-5

Sina Tafazoli, Houman Safaai, Gioia De Franceschi, Federica Bianca Rosselli, Walter Vanzella, Margherita Riggi, Federica Buffolo, Stefano Panzeri, Davide Zoccolan (2017) Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex eLife 6:e22794

Vermaercke, B., Cop, E., Willems, S., D'Hooge, R., Op de Beeck, H. P. (2014). More complex brains are not always better: rats outperform humans in implicit category-based generalization by implementing a similarity-based strategy. Psychonomic bulletin review, 21(4), 1080–1086. https://doi.org/10.3758/s13423-013-0579-9

Vinken, K. and Op de Beeck, H., 2021. Using deep neural networks to evaluate object vision tasks in rats. PLOS Computational Biology, 17(10), p.e1009453.

Vinken, K., Vermaercke, B. and Op de Beeck, H.P., 2014. Visual categorization of natural movies by rats. Journal of Neuroscience, 34(32), pp.10645–10658.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. Nature neuroscience, 22(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

Wang, E.Y., Fahey, P.G., Ding, Z. et al. Foundation model of neural activity predicts response to new stimulus types. Nature 640, 470–477 (2025). https://doi.org/10.1038/s41586-025-08829-y

Yamins, D., DiCarlo, J. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19, 356–365 (2016). https://doi.org/10.1038/nn.4244

D.L.K. Yamins,H. Hong,C.F. Cadieu,E.A. Solomon,D. Seibert, J.J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex, Proc. Natl. Acad. Sci. U.S.A. 111 (23) 8619-8624, https://doi.org/10.1073/pnas.1403112111 (2014).

C. Zhuang,S. Yan,A. Nayebi,M. Schrimpf,M.C. Frank,J.J. DiCarlo, D.L.K. Yamins, Unsupervised neural network models of the ventral visual stream, Proc. Natl. Acad. Sci. U.S.A. 118 (3) e2014196118, https://doi.org/10.1073/pnas.2014196118 (2021).

D. Zoccolan,N. Oertelt,J.J. DiCarlo, D.D. Cox, A rodent model for the study of invariant visual object recognition, Proc. Natl. Acad. Sci. U.S.A. 106 (21) 8748-8753, https://doi.org/10.1073/pnas.0811583106 (2009).

de Vries, S.E.J., Lecoq, J.A., Buice, M.A. et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nat Neurosci 23, 138–151 (2020). https://doi.org/10.1038/s41593-019-0550-9

Kheradpisheh, S., Ghodrati, M., Ganjtabesh, M. et al. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. Sci Rep 6, 32672 (2016). https://doi.org/10.1038/srep32672