Introduction This project is based on a dataset collected during an object recognition task performed by non-human primates [1]. Each image corresponds to average firing rate of a set of neurons. We develop predictive models of IT neural activity using four approaches, with performance quantified by explained variance (EV).

1. Linear Models Predicting neural activity from image pixels using linear regression is challenging due to the high dimension of the input space. A linear model tends to overfit and is not able to capture complex neural dynamics. The use of dimensionality reduction using 1000 principal components and regularization using a 5-fold cross-validation over alpha using Ridge regression improves slightly the predictions, with an EV score reaching 9%. The best alpha value is extremely large (3×10^5), which emphasize the ill-conditioned problem and the need for more expressive models to capture the non-linearity of IT neural activity.

2. Task-Driven Models Task-driven models consist in modeling sensory cortex neural activity using hierarchical convolutional neural network trained on an object recognition task. Those models aim to mimic the encoding process by learning representations that match observed neural responses in the IT cortex [2]. We used activations from individual layers of a pretrained ResNet to predict IT neural activity and found that the EV increases with network depth. The highest EV was achieved from Layer 3 (40.7%). This suggests that deeper layers capture more relevant and brain-like features. The pretrained model significantly outperformed its randomly initialized counterpart showing that task-driven training is critical to develop predictive representations.

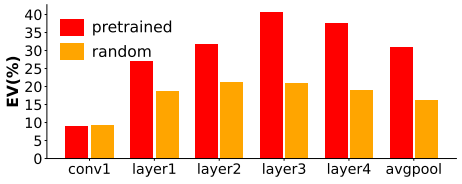

Figure 1: Explained variance scores (%) across layers for pretrained and randomly initialized task-driven models.

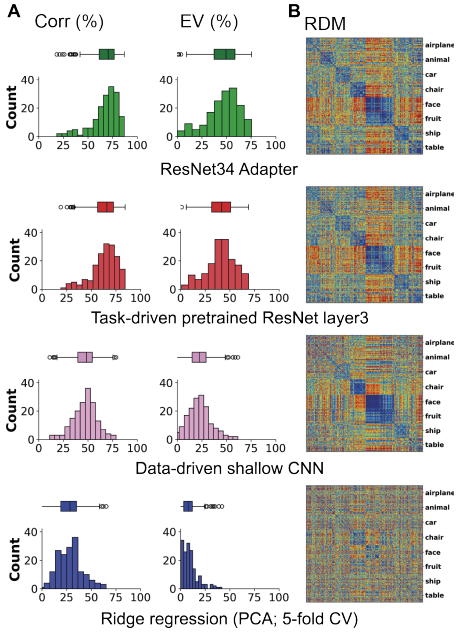

Figure 2: **A**. Histograms and boxplots of correlation (Corr) and explained variance (EV) scores across models. Higher values indicate stronger alignment with IT validation data. **B**. Object-level representation dissimilarity matrices (RDMs) computed from each model and rank-normalized to percentiles (blue = 0th, red = 100th) to visualize the relative representational distances among object categories. Figure adapted from [3].

3. Data-Driven Shallow CNN To assess how much of IT activity can be captured without any task-driven pre-training, we trained a compact four-block CNN. The network uses a 16, 32, 64, 128 filter progression before the global average pool collapses each feature map to a single scalar. No normalization, additional data processing or augmentation was used as all channels had close to 0 mean and 1 standard deviation already. Corrupted images (with noise or rotations) would still be paired to the original spike count, therefore skewing the targets. Additionally, there was too little parameters to add dropout. ShallowCNN explains 22.83% of the response variance, more than ridge, but below best layer of the task-driven pretrained ResNet. These results show that learning features directly from the limited IT dataset already captures a substantial fraction of neural variability, but the richer representations learned through large-scale object recognition remain more brain-like.

4. Finding the Best Model We screened common ImageNet backbones (AlexNet, VGG, ConvNeXt, ResNet, EfficientNet…) by attaching a small decoder that predicts firing rates. Across all tests, ImageNet pre-training and full fine-tuning outperformed random or frozen variants. Thus, subsequent work focused on the top-performing models VGG11_BN and ResNet, for which we conducted detailed comparisons of depth, decoder design, and training protocols.

VGG branch: We varied the number of VGG-style blocks from 1 to 7 and found that a 5-block variant, initialized from VGG11_BN weights and fully fine-tuned, achieved the highest EV. Further optimization of the two-layer classifier and training hyperparameters increased the EV to 45.44% (see Table 2 of the Appendix).

ResNet branch: Using a common decoder across all ResNet variants, we identified the best-performing architecture, ResNet34, by keeping the decoder fixed. We then further improved performance by optimizing the decoder structure and training schedule. The best-performing network is a 34-layer residual model initialized with ImageNet weights and fine-tuned end-to-end for the IT–prediction task. The complete training parameters and architecture specifications for this model can be found in Table 3 of the Appendix. We retain the full ResNet-34 feature trunk and replace its original 1000-class head with an adaptive-average-pool followed by a wide two-layer classifier. It achieved an explained variance of 46.54%. The result demonstrates that a mid-depth residual architecture, fine-tuned with a high-capacity classifier, offers the best trade-off between representational power and over-fitting on the limited IT dataset.

5. Discussion and conclusion Across increasingly expressive models we observe a clear monotonic rise in explained variance: linear-pixel regressors (9.33%), a shallow scratch-trained CNN (22.83%), task-driven ResNet features (40.70%) and, finally, a fine-tuned ResNet-34 adapter (46.54%). This progression highlights two principles: (i) hierarchical convolutional features are essential to match the nonlinear coding of IT, and (ii) starting from task-driven ImageNet representations and fine-tuning on neural data yields the most brain-like encoding.

Model	EV (%)	Corr (%)
Linear	-3.34	21.75
Linear + PCA	-8.81	21.42
Ridge Regression	9.33	28.68
Ridge Regression + PCA	9.24	28.51
Pretrained ResNet-50		
conv1	9.11	28.68
layer 1	27.05	51.83
layer 2	31.74	55.79
layer 3	40.70	63.24
layer 4	37.77	60.53
avgpool	31.10	54.89
Random ResNet-50		
conv1	9.21	28.44
layer 1	18.70	41.53
layer 2	21.23	44.25
layer 3	20.85	43.86
layer 4	19.07	41.77
avgpool	16.30	38.54
Shallow CNN	22.83	46.99
Resnet-34	**46.54**	**67.09**

Table 1: Model performance comparison with respect to Explained Variance (EV) and Pearson Correlation (Corr).

References

[1] Najib J. Majaj et al. "Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance". In: *The Journal of Neuroscience* 35.39 (2015). DOI: 10.1523/JNEUROSCI.5181-14.2015.

[2] Daniel L K Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". In: *Nature Neuroscience* 19.3 (Feb. 2016), pp. 356–365. ISSN: 1546-1726. DOI: 10.1038/nn.4244. URL: http://dx.doi.org/10.1038/nn.4244.

[3] Daniel L. K. Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences* 111.23 (May 2014), pp. 8619–8624. ISSN: 1091-6490. DOI: 10.1073/pnas.1403112111. URL: http://dx.doi.org/10.1073/pnas.1403112111.

Appendix

Reproducibility All experimental results can be reproduced using the code and data available at https://github.com/margaux-roulet/visionary/tree/main. For deterministic behavior, we fixed random seeds in both NumPy and PyTorch throughout all experiments.

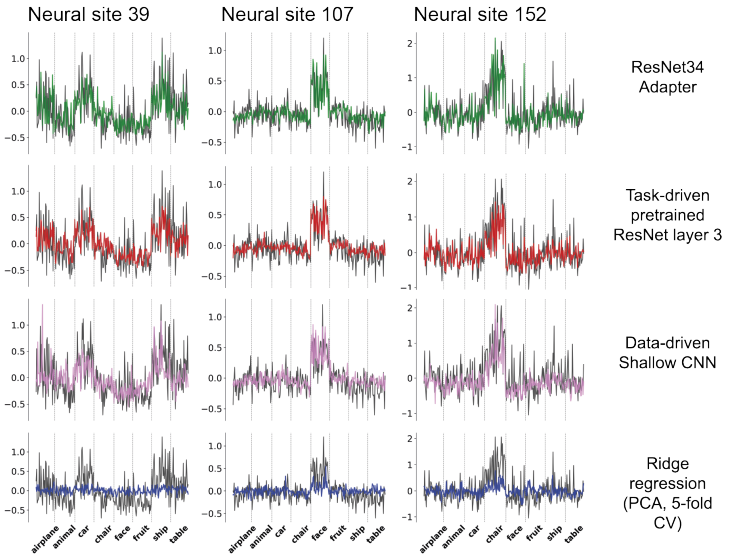

Figure 3: Neural activity responses of three neural sites to various stimulus categories. The responses show that individual neurons exhibit selective firing patterns. Figure adapted from [3]

Parameter	Value
Backbone	Pretrained VGG11_BN (unfrozen)
Classifier Hidden Layers	7000 → 7000 (ReLU, Dropout=0.47)
Optimizer	AdamW (LR=1e-4, Weight Decay=1e-5)
Scheduler	StepLR (step=15, γ=0.5)
Training	60 epochs (early stopping patience=7)

Table 2: Best VGG Model Configuration.

Parameter	Value
Backbone	Pretrained ResNet-34 (unfrozen)
Classifier Hidden Layers	7000 → 7000 (ReLU, Dropout=0.52)
Optimizer	Adam (LR=1e-4, Weight Decay=1e-5)
Scheduler	StepLR (step=15, γ=0.5)
Training	60 epochs (early stopping patience=10)

Table 3: Best ResNet-34 Model Configuration.