Tackling the 2021 Algonauts Challenge with Semi-Supervised Networks & Bayesian Optimization

Robert Tjarko Lange *
Technical University Berlin

1 Introduction

Deep neural networks have been widely adopted as state-of-the-art models of the visual ventral stream (e.g. Cadieu et al., 2014; Yamins and DiCarlo, 2016; Cichy et al., 2016). Most studies have focused on predicting neural responses to static visual stimuli. But arguably the majority of human visual processing involves movement and dynamic scenes. The 2021 Algonauts challenge (Cichy et al., 2021) attempts to address this forefront of current research and focuses on explaining human neural responses to short video stimuli. Participants were tasked to predict fMRI voxel activity for 10 subjects who viewed short video clips of moving objects and humans. The challenge was divided into two tracks: A *mini-track* focusing on a set of 9 regions of interest (ROI) and a *full-track* emphasizing whole brain activity. In the following report we outline our 5th place solution approach to the mini-track and share insights gathered through exploration and hypothesis testing.

Solution approach. Our solution is based on a self-supervised pre-trained ResNet-50 network (SimCLR-v2; Chen et al., 2020b), which was fine-tuned on supervised ImageNet labels. After extracting the layer-wise features for the video frames, we temporally mean-aggregate the individual activations and reduce their dimensionality using principal component analysis (PCA) to 50 features. Afterwards, we run a Bayesian Optimization (BO; Snoek et al., 2012) loop to tune the number of components in a partial least squares (PLS) regression encoding model. The BO procedure is layer-, subject- and ROI-specific and optimizes a 10-fold cross-validated (CV) correlation metric. We select the combination of best performing network layer and encoder hyperparameter and retrain a final PLS encoding model on all datapoints. The general pipeline is depicted in figure 1.

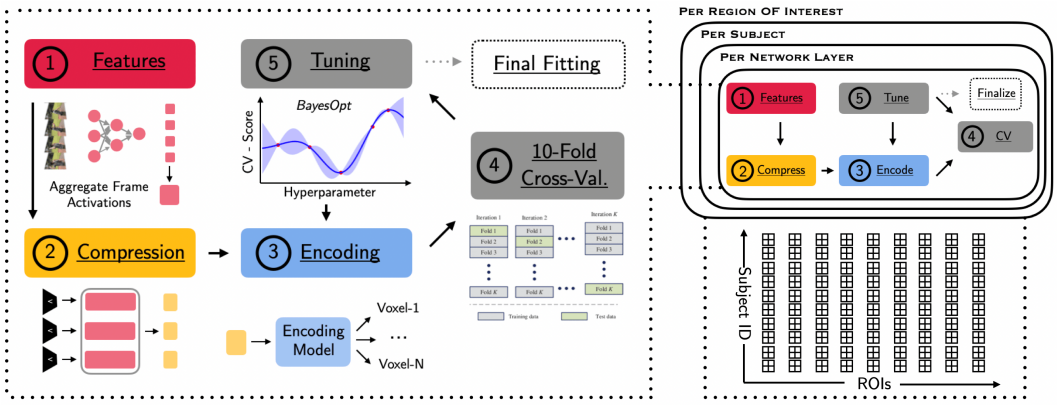

Figure 1: Proposed general solution pipeline consisting of 6 steps: We extract neural network features by computing layer-wise activations for a set of video frames. The frame activations are temporally aggregated and dimensionality reduced. Based on the preprocessed features we train encoding models to predict normalized voxel activity. The hyperparameters are tuned using CV and BO. Finally, the best layer activations and encoder hyperparameters are used to fit a final encoding model, which predicts the test set activity. Steps 1 and 2 can be computed once and amortized, while steps 4, 5 and 6 are computed in parallel for all subjects and ROIs.

*Code to replicate the results can be found at `https://github.com/RobertTLange/algonauts-2021`.

Compute resources. We precompute all layer-specific network features, their temporal aggregation and the compression on a GPU accelerator. Depending on the chosen video sampling rate, compression technique, network architecture and number of network layers, this pre-processing takes between 2 and 5 hours. Afterwards, we parallelize the encoding model optimization routine across ROIs and subjects using the `MLE-Toolbox` (Lange, 2021) on a CPU cluster. Each individual BO loop requires between 2 and 4 CPU cores and lasts between 2 and 4 hours depending on the number of feature dimensions, encoding model and fitted voxels. Given a pre-trained feature extractor network the entire pipeline takes between 4 and 9 hours.

2 Initial Exploration of the Solution Space

The space of possible solution attempts is vast and impossible to search through exhaustively. We therefore initially focused on exploring the impact of three main ingredients: The neural network architecture used to extract frame-dependent features, the dimensionality reduction technique used to compress the layer-wise activations and the encoding model used to predict ROI- and subject-specific voxel activations. In order to compare the different ingredient configurations across all subjects, we computed subject-aggregated ROI-fit scores by calculating a weighted average of subject-specific correlation scores. The subject-ROI weight is computed as the ratio of subject-specific ROI voxels and total ROI voxels across all individuals (see SI B). Our insights are summarized as follows:

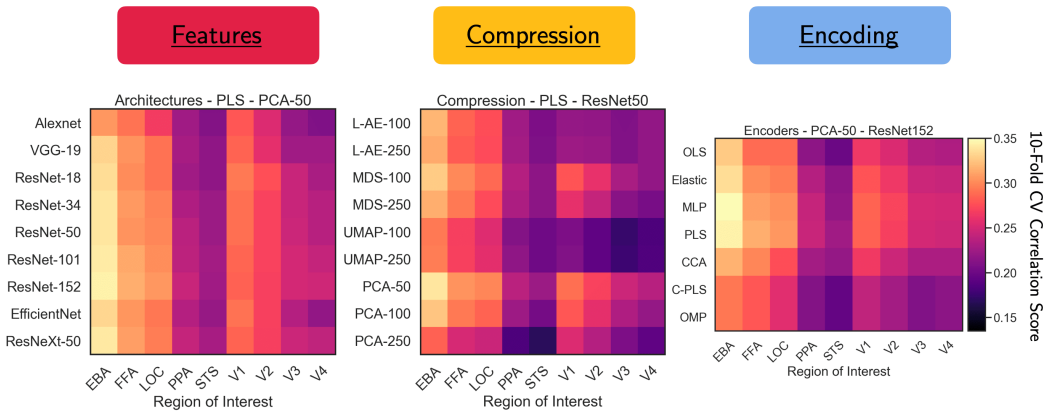

Figure 2: Exploration of different dimensions of the solution space. **Left**: Neural Network Architectures. **Middle**: Dimensionality Reduction Techniques. **Right**: Encoding Models.

Neural Network Features. We evaluated a set of common convolutional neural network architectures including VGG-19, residual networks of different sizes and several others (figure 2 left). To do so, we fixed the compression technique (PCA-50) and encoding model (PLS). We observed that all models tend to perform qualitatively similar across different ROIs and that larger ResNet architectures generalized better. Only the AlexNet baseline performed fairly poor, indicating the importance of top-down task performance (i.e. classification accuracy on ImageNet (Deng et al., 2009)).

Feature Compression Techniques. Next, we were interested in whether linear dimensionality reduction is sufficient to capture the relevant variation in network activity. We compared CV correlation scores for a simple multi-layer perceptron (MLP) auto-encoder (AE), multi-dimensional scaling (MDS; McInnes et al., 2018), UMAP (Cox and Cox, 2008) and PCA with different compression dimensionality. Figure 2 (middle) reveals that both PCA-50 and MDS-100 provide comparably strong performance for ResNet-50 features with downstream PLS encoding. Using more feature dimensions, on the other hand, quickly hinders generalization, highlighting the risk of overfitting.

Encoding Models. Finally, we compared several encoding approaches and chose to fix ResNet-50 features and PCA-50 compression. We compared multi-output linear regression, elastic net (Zou and Hastie, 2005), MLP, PLS, canonical correlation analysis (CCA Thompson, 1984), canonical PLS and orthogonal matching pursuit (OMP) encoders. Both PLS regression and MLP encoding resulted in the best encoding correlation scores across all ROIs (figure 2 right). Due to computational and data amount considerations we decided to further focus primarily on PLS regression.

3 Proposed Solution: SimCLR-v2, PCA-50 & PLS Regression

We wondered whether network features resulting from pure supervised training were suitable for predicting neural responses to moving stimuli. Active vision involves partially noisy data measurements resulting from saccades and top-down attention processes. We therefore hypothesized that a self-supervised pre-training procedure, which relies on random data augmentations may yield more robust and predictive features. Furthermore, unsupervised models were previously successfully explored as models of the visual cortex (Nayebi et al., 2021; Zhuang et al., 2021). We focused our attention on a recently proposed self-supervised training paradigm: SimCLR (Chen et al., 2020a,b)[2]. SimCLR relies on a set of random data augmentations and a contrastive learning objective (see figure 3 left), which promotes similar representations to result from augmented versions of the same image (positive pair) and dissimilar representations for different base images (negative pair). SimCLR-v2 (Chen et al., 2020b) additionally proposes to afterwards fine-tune on a subset labeled images and to perform student-teacher distillation on unlabelled examples. Due to the observed importance of task-specific performance when fitting neural responses (see section 2, we used SimCLR-v2 ResNet networks, which were fine-tuned on 100 percent of ImageNet labels.

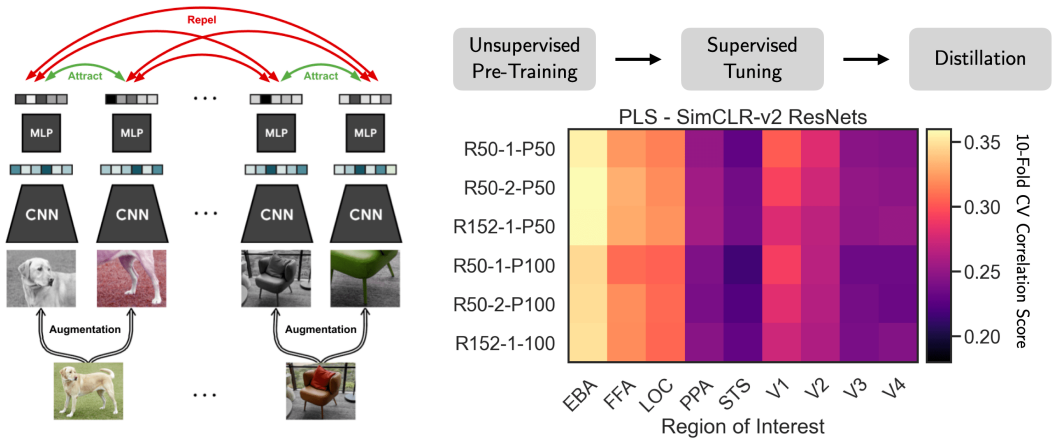

Figure 3: Self-Supervised SimCLR-v2 solution. **Left**: Illustration of contrastive learning paradigm. Figure taken from Chen et al. (2020a). **Right**: Performance of different SimCLR-v2 pre-trained models for different ROIs. ResNet-50 features with PCA-50 and without increased channel width perform best.

Figure 3 (right) depicts the weighted voxel-prediction correlation for different sizes of SimCLR-v2 ResNets (with/without selective kernels and increased channel width). The ResNet-50 architecture without increased channel width performs best across all ROIs. This model configuration was used in our final submission. We provide more information about the layer-wise performance and BO procedure in figure 5.

4 Additional Experiments: Temporal Aggregation, VOneNets & Auto-ML

This section summarizes further explorations which did not yield significant improvements. Nonetheless, we believe that there may be valuable insights to be gained:

Temporal Filtering. The raw fMRI is processed using adaptive temporal filters in order to extract an estimate of the hemodynamic response. We wondered whether a similar filtering of layer activations over time (video frames) might improve their predictivity. We created convolution kernels using a double gamma formulation:

$$\mathcal{K}(t; \texttt{peak}, \texttt{under}, \beta_{\texttt{under}}) = \texttt{Gamma}(t, \texttt{peak}) - \beta_{\texttt{under}}\texttt{Gamma}(t, \texttt{under}),$$

where `Gamma`, `under`, `peak` denote the Gamma distribution density and shape parameters. We convolve the unit activity recorded from network layers across video frames and mean the signal

[2]We rely upon the official pre-trained checkpoints which were converted to be PyTorch (Paszke et al., 2019) compatible. The checkpoints may be found here: `https://github.com/google-research/simclr`. Furthermore, the conversion code is available at `https://github.com/Separius/SimCLRv2-Pytorch`.

(figure 4, top). When comparing different values for `peak`, `under`, $\beta_{\texttt{under}}$, mean and median temporal aggregation we surprisingly did not find a significant differences. We also validated several levels of temporal downsampling of video frames and did not find a strong effect (figure 4, bottom left).

VOneNetworks. As of writing the top-4 neural and behavioral predictivity models on the BrainScore benchmark (Schrimpf et al., 2020) are provided by different VOneNet architectures (Dapello et al., 2020). It was a natural question whether these architectures, which exchange the first layer with a stochastic biology-informed Gabor Filter bank, would perform well in predicting BOLD responses to video snippets. We found that Cornet-S (Kubilius et al., 2019) provide the best base architecture for the VOne frontend. Neither adversarial training nor eliminating of stochasticty improved the VOne-Resnet-50 score. Furthermore, the VOne-Cornet-S test score obtained on the holdout dataset did not exceed the SimCLR-v2 submission. This raises the question whether low-level physiology-inspired models generalize from fitting invasive macaque recordings to the challenge's fMRI setting.

Auto-ML. We briefly experimented with replacing the BO loop with an Auto-ML procedure, *Auto-Sklearn v.1.0* (Feurer et al., 2019). Is the focus on a single compression technique and encoding model class across all subjects and ROIs too restrictive? This appears not to be the case (results not shown): While the best Auto-ML activity predictions did significantly better on the training data, they did not generalize well to the unseen test data. We note that this may be due to a fairly restrictive search resulting from heavy disk space requirements or insufficient tuning of the Auto-ML parameters.

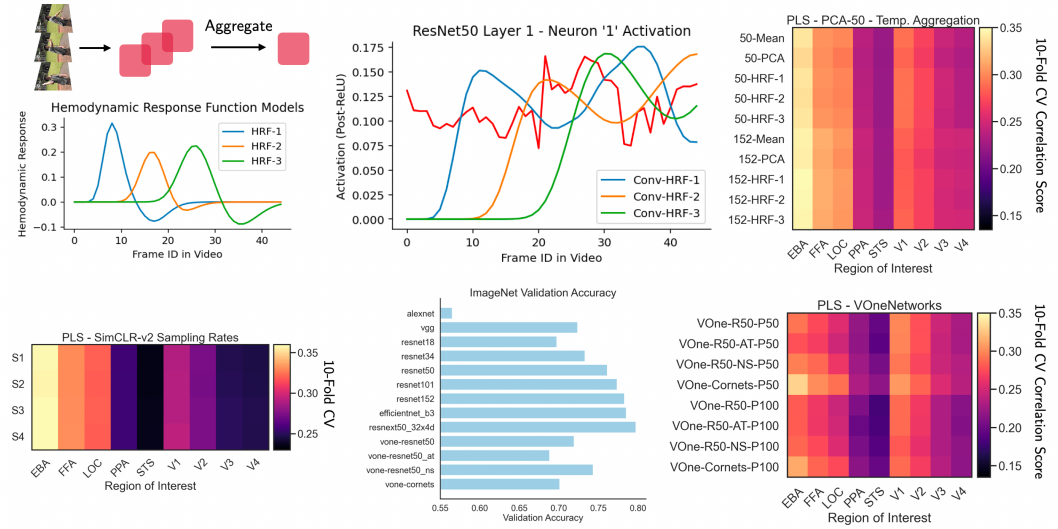

Figure 4: Additional investigations. **Top**: Illustration and evaluation of temporal feature aggregation approaches. **Bottom**: Sampling rate comparison & VOneNets (Dapello et al., 2020) evaluation for different base architectures.

5 Conclusion

We introduced a general BO-based pipeline to cross-validate different ingredients to the encoding procedure. This pipeline was leveraged to compare the capabilities of model architectures, compression and encoding methods to fit neural activity resulting from small video snippet stimulation. We found contrastive pre-training and supervised fine-tuning yield a powerful feature extractor, which fits neural variation well. If there had been more time we would have liked to further investigate the following questions:

- *Prediction ensemble*. How to combine predictions from several encoding models?
- *Cross-subject/ROI encoding*. How to maximally leverage the sparse training signal?
- *Transformer-CNN hybrids* (e.g. d'Ascoli et al., 2021). What is the right amount of inductive bias?

Finally, we thank the organizers for putting together this exciting challenge and the open source community for public network checkpoints and software infrastructure.

References

CADIEU, C. F., H. HONG, D. L. YAMINS, N. PINTO, D. ARDILA, E. A. SOLOMON, N. J. MAJAJ, AND J. J. DICARLO (2014): "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS computational biology*, 10, e1003963.

CHEN, T., S. KORNBLITH, M. NOROUZI, AND G. HINTON (2020a): "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 1597–1607.

CHEN, T., S. KORNBLITH, K. SWERSKY, M. NOROUZI, AND G. HINTON (2020b): "Big self-supervised models are strong semi-supervised learners," *arXiv preprint arXiv:2006.10029*.

CICHY, R. M., K. DWIVEDI, B. LAHNER, A. LASCELLES, P. IAMSHCHININA, M. GRAUMANN, A. ANDONIAN, N. MURTY, K. KAY, G. ROIG, ET AL. (2021): "The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion," *arXiv preprint arXiv:2104.13714*.

CICHY, R. M., A. KHOSLA, D. PANTAZIS, A. TORRALBA, AND A. OLIVA (2016): "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Scientific reports*, 6, 1–13.

COX, M. A. AND T. F. COX (2008): "Multidimensional scaling," in *Handbook of data visualization*, Springer, 315–347.

DAPELLO, J., T. MARQUES, M. SCHRIMPF, F. GEIGER, D. D. COX, AND J. J. DICARLO (2020): "Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations," *BioRxiv*.

D'ASCOLI, S., L. SAGUN, G. BIROLI, AND A. MORCOS (2021): "Transformed CNNs: recasting pre-trained convolutional layers with self-attention," *arXiv preprint arXiv:2106.05795*.

DENG, J., W. DONG, R. SOCHER, L.-J. LI, K. LI, AND L. FEI-FEI (2009): "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 248–255.

FEURER, M., A. KLEIN, K. EGGENSPERGER, J. T. SPRINGENBERG, M. BLUM, AND F. HUTTER (2019): "Auto-sklearn: efficient and robust automated machine learning," in *Automated Machine Learning*, Springer, Cham, 113–134.

KUBILIUS, J., M. SCHRIMPF, K. KAR, H. HONG, N. J. MAJAJ, R. RAJALINGHAM, E. B. ISSA, P. BASHIVAN, J. PRESCOTT-ROY, K. SCHMIDT, ET AL. (2019): "Brain-like object recognition with high-performing shallow recurrent ANNs," *arXiv preprint arXiv:1909.06161*.

LANGE, R. T. (2021): "MLE-Toolbox: A Reproducible Workflow for Distributed Machine Learning Experiments," Available at `https://pypi.org/project/mle-toolbox/`, version 0.3.0.

MCINNES, L., J. HEALY, AND J. MELVILLE (2018): "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*.

NAYEBI, A., N. C. KONG, C. ZHUANG, J. L. GARDNER, A. M. NORCIA, AND D. L. YAMINS (2021): "Unsupervised Models of Mouse Visual Cortex," *bioRxiv*.

PASZKE, A., S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA, ET AL. (2019): "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, 32, 8026–8037.

SCHRIMPF, M., J. KUBILIUS, H. HONG, N. J. MAJAJ, R. RAJALINGHAM, E. B. ISSA, K. KAR, P. BASHIVAN, J. PRESCOTT-ROY, F. GEIGER, ET AL. (2020): "Brain-score: Which artificial neural network for object recognition is most brain-like?" *BioRxiv*, 407007.

SNOEK, J., H. LAROCHELLE, AND R. P. ADAMS (2012): "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, 25.

THOMPSON, B. (1984): *Canonical correlation analysis: Uses and interpretation*, 47, Sage.

YAMINS, D. L. AND J. J. DICARLO (2016): "Using goal-driven deep learning models to understand sensory cortex," *Nature neuroscience*, 19, 356–365.

ZHUANG, C., S. YAN, A. NAYEBI, M. SCHRIMPF, M. C. FRANK, J. J. DICARLO, AND D. L. YAMINS (2021): "Unsupervised neural network models of the ventral visual stream," *Proceedings of the National Academy of Sciences*, 118.

ZOU, H. AND T. HASTIE (2005): "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67, 301–320.

A Bayesian Optimization Hyperparameter Ranges

Encoding Model	Parameter	Min	Max	Variable Type	Prior
OLS	L2 λ	1e−07	1e−01	Real	$\mathcal{U}$
Elastic Net	α	0.01	0.25	Real	$\mathcal{U}$
	L1 Ratio	0.01	1	Real	$\mathcal{U}$
MLP	No. hidden layers	$\{1, 2, 3, 4\}$		Categorical	$\mathcal{U}$
	No. hidden units	$\{64, 128, 256, 512\}$		Categorical	$\mathcal{U}$
	Activation fct.	{ReLU, PReLU, leaky ReLU, ELU}		Categorical	$\mathcal{U}$
	Optimizer	{Adam, RMSProp}		Categorical	$\mathcal{U}$
	Learning rate	{1e−05, 5e−05, 1e−04, 5e−04, 1e−03}		Categorical	$\mathcal{U}$
	Weight decay	{0.0, 1e−07, 1e−06, 1e−05}		Categorical	$\mathcal{U}$
	Dropout	$\{0.0, 0.1, 0.2\}$		Categorical	$\mathcal{U}$
PLS	No. components	1	100	Integer	$\mathcal{U}$
CCA	No. components	1	50	Integer	$\mathcal{U}$
C-PLS	No. components	1	50	Integer	$\mathcal{U}$
OMP	Non-zero Coeffs.	1	50	Integer	$\mathcal{U}$

Table 1: Encoding hyperparameter ranges to tune with Bayesian optimisation (25 iters. per feature layer).

B Final Submission Evaluation Details

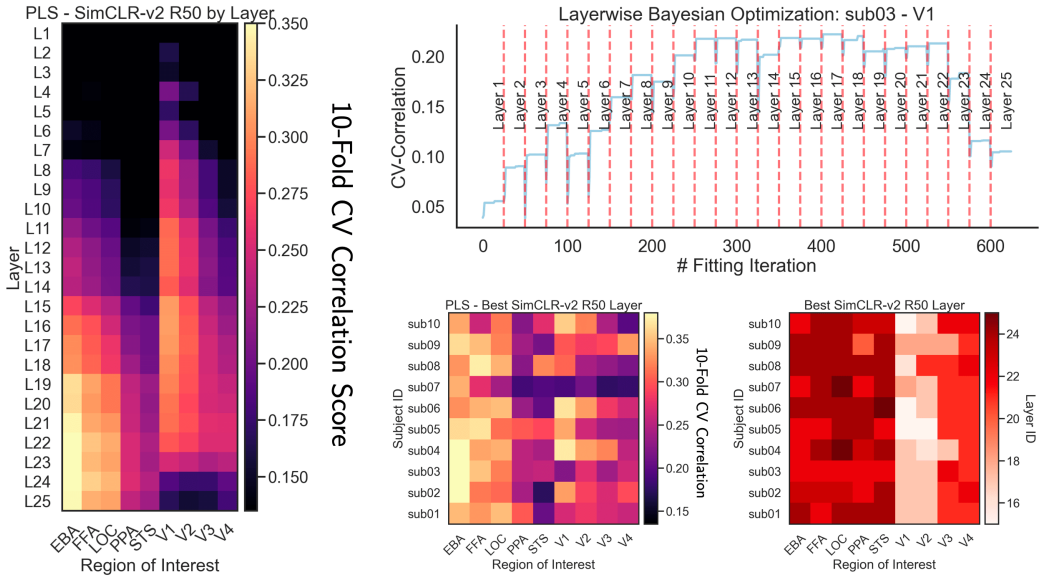

Figure 5: Additional results for SimCLR-v2 encoding. **Left**: Layer-wise voxel-weighted CV correlation scores. Later layers fit EBA, FFA, LOC, PPA and STS well, while earlier layers excel at fitting the ventral stream (V1-V4). **Right, Top**: Bayesian optimization best score dynamics across all SimCLR-v2 layers for subject 3 and V1. **Right, Bottom**: Best CV score and layer id for each subject and region of interest.