Neural responses in early, but not late, visual cortex are well predicted by random-weight CNNs with sufficient model complexity

Amr Farahat[1,2,*] **and Martin Vinck**[1,2]

[1]Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany
[2]Donders Centre for Neuroscience, Department of Neurophysics, Radboud University, Nijmegen, the Netherlands
[*]amr.farahat@esi-frankfurt.de

ABSTRACT

Convolutional neural networks (CNNs) were inspired by the organization of the primate visual system, and in turn have become effective models of the visual cortex, allowing for accurate predictions of neural stimulus responses. While training CNNs on brain-relevant object-recognition tasks may be an important pre-requisite to predict brain activity, the CNN's brain-like architecture alone may already allow for accurate prediction of neural activity. Here, we evaluated the performance of both task-optimized and brain-optimized convolutional neural networks (CNNs) in predicting neural responses across visual cortex, and performed systematic architectural manipulations and comparisons between trained and untrained feature extractors to reveal key structural components influencing model performance. For human and monkey area V1, random-weight CNNs employing the ReLU activation function, combined with either average or max pooling, significantly outperformed other activation functions. Random-weight CNNs matched their trained counterparts in predicting V1 responses. The extent to which V1 responses can be predicted correlated strongly with the neural network's complexity, which reflects the non-linearity of neural activation functions and pooling operations. However, this correlation between encoding performance and complexity was significantly weaker for higher visual areas that are classically associated with object recognition, such as monkey IT. To test whether this difference between visual areas reflects functional differences, we trained neural network models on both texture discrimination and object recognition tasks. Consistent with our hypothesis, model complexity correlated more strongly with performance on texture discrimination than object recognition. Our findings indicate that random-weight CNNs with sufficient model complexity allow for comparable prediction of V1 activity as trained CNNs, while higher visual areas require precise weight configurations acquired through training via gradient descent.

Introduction

The development of convolutional neural networks (CNNs) was originally inspired by features of the primate visual system, such as its hierarchical organization [Felleman and Van Essen, 1991, Vezoli et al., 2021] and localized receptive fields (RFs) with repeated feature kernels across space [Hubel et al., 1959, Fukushima et al., 198 LeCun et al., 1989]. CNNs, and deep neural networks (DNNs) in general, have in turn become effective models of the primate visual ventral stream, allowing for relatively accurate prediction of neural responses to novel, natural stimuli [Cadena et al., 2019, Güçlü and Van Gerven, 2015, Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014, Zhuang et al., 2021]. The efficacy of task-optimized CNNs in predicting a given brain area's neural responses (henceforth referred to as "encoding performance") is thought to depend on several factors, such as the network architecture, the objective function, training dataset and learning rules used

for training[Richards et al., 2019, Cichy and Kaiser, 2019, Doerig et al., 2023].

Although earlier work postulated that CNNs can effectively predict neural activity because they are trained on ecologically relevant object recognition tasks [Cadieu et al., 2014, Mehrer et al., 2021], the effect of training may differ substantially between hierarchical levels of the primate ventral stream. It stands to reason that training networks on a diet of natural images with an objective function probing for invariant image classification (e.g. supervised object recognition [Yamins et al., 2014] or contrastive self-supervised learning [Zhuang et al., 2021]) is important for predicting neural activity in higher regions of the primate ventral stream, given their high degree of functional specialization [Cadieu et al., 2014, Rust and DiCarlo, 2010, DiCarlo et al., 2012, Hung et al., 2005]. Indeed, several studies highlight the importance of training objectives and datasets in predicting the activity in higher primate ventral stream regions, while suggesting that differences in architecture between trained CNNs have a smaller influence on explaining responses throughout the primate ventral stream [Storrs et al., 2021, Conwell et al., 2024, Zhuang et al., 2021]. Yet, it is less clear to what extent training CNNs is essential for predicting activity in early visual areas of the primate cortex, which show much less functional specialization and may be involved in a wider range of functions beyond object recognition including scene segmentation [Self et al., 2013], motion processing [Gur and Snodderly, 2007], and salience detection [Li, 2002]. While CNNs trained for object recognition outperform traditional models like linear-nonlinear Poisson models and Gabor filters in predicting macaque V1 responses to natural images [Simoncelli et al., 2004, Willmore et al., 2008, Cadena et al., 2019], this superior prediction may be either due to the CNN's architecture or the used training objective. A recent study showed that the RF size of neurons in object-recognition trained CNNs was an important determinant of encoding performance, suggesting that the CNN's architecture does play an important role for predicting V1 activity [Miao and Tong, 2024].

Furthermore, it is plausible that differences between trained and untrained CNNs in predicting neural activity depend strongly on the initial architecture of the CNN with random weights. Recent studies showed that the generalization capacity of DNNs can be attributed to their loss landscapes upon initialization dictated by their architectural design [Chiang et al., 2022, Ramasinghe et al., 2022]. In particular, random-weight DNNs can show strong differences in the complexity of input/output functions dependent on e.g. the non-linear activation function used in the network [Teney et al., 2024]. It is possible that training CNNs steers them towards a certain non-linear complexity matching neural complexity, thereby masking initial differences in architecture, but that a random-weight CNN with an appropriate RF size and non-linear complexity may already allow for accurate prediction of brain activity.

Here, we systematically test whether certain architectural components contribute to CNNs' ability to encode neural data of early and high visual areas in primates' brains. Specifically, we constructed CNN models with a linear readout to predict neural data and investigated when training the convolutional filters is necessary for good encoding performance versus only training the linear readouts for an otherwise random-weight CNN.

Results

Neural encoding performance of VGG16 model

We analyzed three neural datasets: (1) Firing rates of 166 V1 neurons from two macaques, recorded while the animals passively viewed natural and texture images [Cadena et al., 2019]; (2) activity from 168 multi-unit sites in the IT cortex of two macaque monkeys, passively viewing 3200 grayscale images [Cadieu et al., 2014]; and (3) the fMRI Natural Scenes Dataset (NSD) [Allen et al., 2022], comprising fMRI responses from 8 human subjects viewing thousands of color natural images. Neural activity was predicted by linearly transforming the three-dimensional activation maps of each convolutional layer

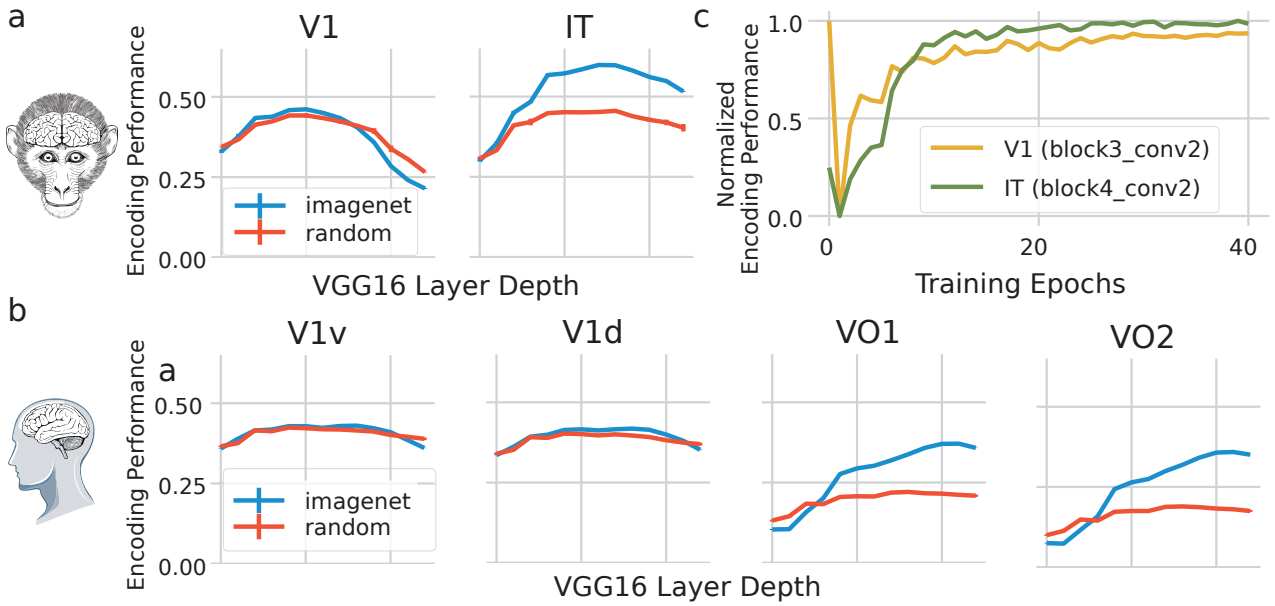

Figure 1. Encoding performance of VGG16 model. **(a)** Encoding performance of a linear readout optimized on top of the representations of the convolutional layers of VGG16 model either upon random initialization (in red) or pretraiend on ImageNet dataset for object recognition (in blue) for two neural datasets recorded from the early visual cortex V1 (left) or the higher visual area IT (right) in macaques. **(b)** same as **a** but for two early visual cortex ROIs (V1v and V1d) or higher visual areas (VO1 and VO2) in huamn fMRI data. **(c)** The normalized encoding performance of the best performing layer for V1 and IT brain areas in macaques tracked over training time on ImageNet dataset from random initialization till full convergence. Each line is the average of 3 training iterations.

in the VGG16 model into a one-dimensional vector representing neural activity (either firing rates for macaque datasets or voxel activations for the human fMRI dataset). See supplementary Fig. S1 For a visual illustration of the models. The weights of this linear transformation were fitted on 80% of the dataset and tested on the remaining held-out test set. The "encoding performance" for each VGG16 layer was quantified as the median Pearson correlation between the predicted and actual neural responses across all neuronal sites. Consistent with previous findings, the layers of the ImageNet-trained VGG16 model could predict a substantial amount of variance in neural responses to unseen (i.e. test-set) stimuli in both early and higher visual areas of the macaque and human brain (Fig. 1).

Next, we investigated the influence of object recognition training on the encoding performance. When the VGG16 model was initialized with random weights, thereby omitting task training, its ability to predict primate V1 responses showed only minor differences compared to the ImageNet-trained VGG16 model (Fig. 1; difference trained vs. random-weight CNNs: 0.009 for macaque V1, 0.008 for human V1v, 0.012 for human V1d). By contrast, a substantial decrease in encoding performance was observed when predicting responses in higher visual areas using random-weight CNNs (inferotemporal cortex (IT) in macaque and ventral occipital areas (VO1 and VO2) in humans (Fig. 1; difference trained vs. random-weight CNNs: 0.132 for macaque IT, 0.138 for human VO1, and 0.176 for human VO2). The loss in encoding performance for IT was statistically much larger than for V1 (Mann-Whitney U rank test $p \ll 0.001$). We found qualitatively similar results using other popular convolutional models such as Resnet50 [He et al., 2016], Inception [Szegedy et al., 2015], and DenseNet [Huang et al., 2016] (see supplementary Fig. S3). We furthermore observed only minor differences in V1 encoding performance

comparing trained and random-weight CNNs for other measurements of V1 brain activity such as the amplitude of gamma band in local field potentials in an independent data set [Uran et al., 2022] (see supplementary Fig. S4).

To examine the impact of ImageNet training on the encoding performance of VGG16, we quantified the encoding performance across each training epoch, starting with randomly initialized weights and progressing to full convergence (Fig. 1c). Although the network's performance on object recognition improved monotonically with training, the encoding performance showed a markedly different profile: Starting from the first epoch, the encoding performance for V1 declined notably after the initial training epoch compared to the randomly initialized weights. Hence, training on the object recognition initially decreases the encoding performance, i.e. the ability to predict V1 activity. The encoding performance for V1 only recovered upon reaching full convergence. For IT, however, the encoding performance mostly showed a monotonic increase from the first towards the last training epochs.

Together, these findings indicate that training a CNN architecture (VGG16) on object recognition is not essential for predicting primate V1 activity, as random-weight CNNs demonstrate comparable performance to trained CNNs.

Simple convolutional models for encoding early and higher visual areas

To further investigate the efficacy of randomly initialized networks in predicting neural responses, we constructed simpler CNN models, systematically changed their architecture and training, and then evaluated their neural encoding performance across various brain regions in macaques and humans. We varied network depth between shallow (2 layers) and deeper (4 layers) architectures, while adjusting convolutional kernel sizes to maintain consistent receptive field sizes across models (see Methods and supplementary Fig. S2). Additionally, we evaluated average and maximum pooling operations and four distinct activation functions (ReLU, ELU, Tanh, and Linear). In this case, we optimized the neural network weights and a linear readout to directly predict neural activity, rather than training on an object recognition task (see Methods), similar to a previous study that developed a shallow neural-network model to predict V1 activity [Du et al., 2024].

For the trained models, we observed that, with the exception of linear networks, all models achieved comparable performance for the V1 area in both macaques and humans (Fig. 2a). Specifically, linear networks employing average pooling exhibited poor encoding performance, while different non-linear activation functions or linear networks utilizing max pooling operations yielded higher and comparable V1 encoding performance. The V1 encoding performance was comparable between shallow and deep CNNs. The performance of the models in predicting responses from higher visual areas (IT, VO1, and VO2) was comparable across non-linear activation functions but improved for deep compared to shallow networks. In sum, when convolutional filters are optimized, the architectural bias of the models is subtle, i.e., the performance difference between different activation functions and pooling mechanisms is almost negligible (except for linear networks with average pooling).

We then evaluated the performance of the networks when only the linear readout was optimized, while the convolutional filters were frozen at their randomly initialized weights. For these and subsequent analyses, unless otherwise specified, we focused on shallow 2-layer networks for the early visual cortex and the deeper 4-layer networks for higher visual areas. The comparison of the encoding performance of models with trained convolutional layers to those with randomly initialized weights showed that random ReLU networks approached the performance of their trained counterparts in predicting V1 responses for both macaques and humans (Fig. 2b). The differences between random and trained ReLU networks were 0.045 for the average pooling models and 0.068 for the maximum pooling models. Compared to V1, there was a much greater difference in encoding performance in higher visual areas (IT, VO1, VO2; Fig. 2b).

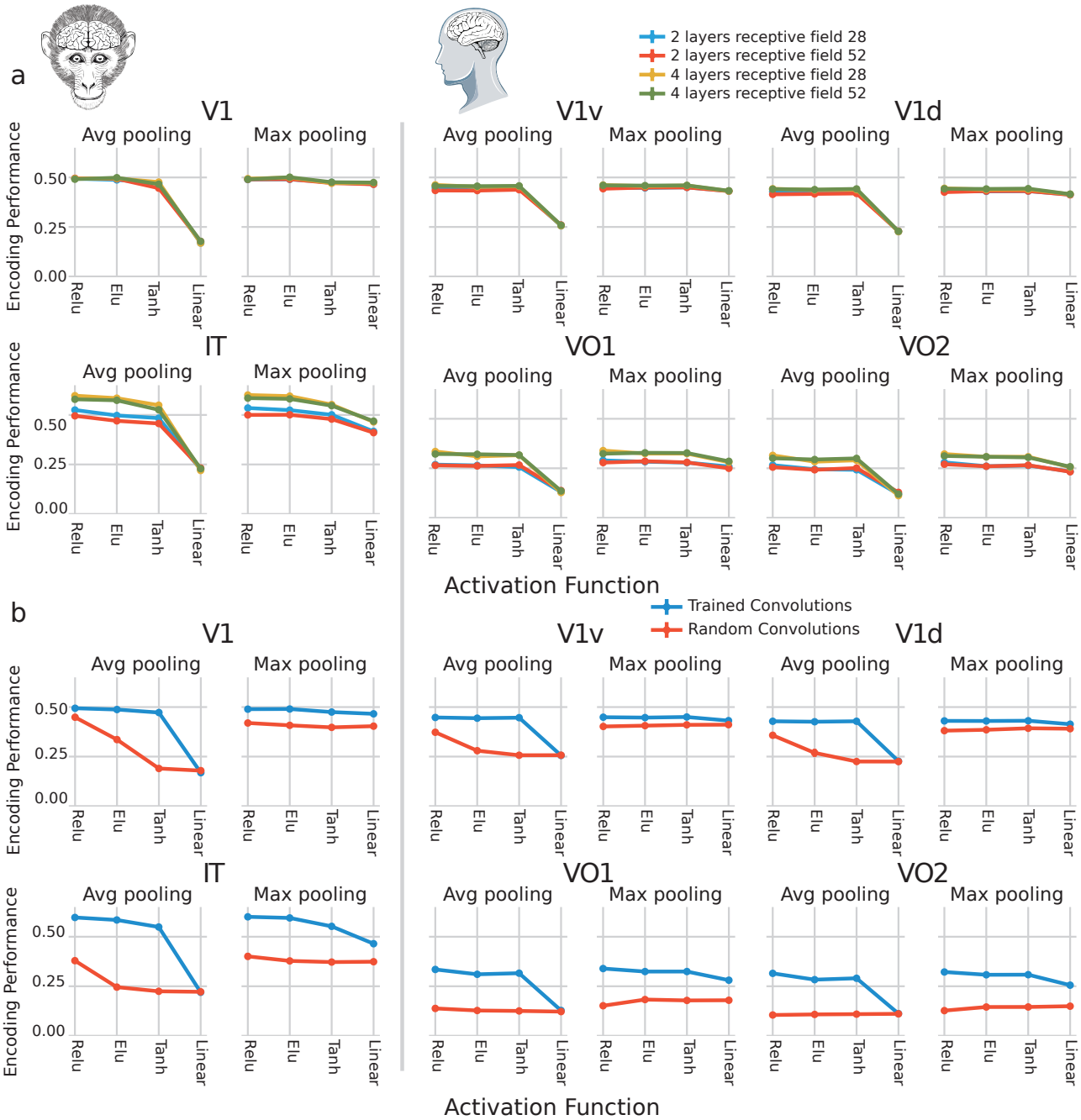

Figure 2. Encoding performance of simple convolutional models on early and higher visual areas of macaques and human brains. (a) Encoding performance of shallow (2 layers) and deeper (4 layers) convolutional models each manipulated to have effective receptive field of either 28 or 52 pixels^2 fully trained on neural data from early (V1, V1v, and V1d) and higher (IT, VO1, and VO2) visual brain areas of macaques and humans. Each model has 8 variants: 2 different pooling strategies (average pooling and maximum pooling) and 4 different activation functions (ReLU, ELU, Tanh, and Linear). **(b)** Encoding performance of the best performing fully trained models from **a** (in blue) – the 2-layer 28×28 models for V1, V1v, and V1d brain areas and the 28×28 4-layer models for IT, VO1, and VO2 brain areas – and their randomly initialized counterparts (in red). In the latter case, only the linear readout was trained on top of the randomly initialized weights. Each line is the average of 5 training iterations.

The differences between random and trained ReLU networks were 0.214 for the average pooling models and 0.196 for the maximum pooling models. In contrast to fully trained networks, networks with randomly initialized convolutional weights showed substantial differences in V1 encoding performance across the activation functions, especially those using average pooling operations (Fig. 2b; $s = 0.106$ for random nonlinear models and $s = 0.009$ for trained nonlinear models).

In summary, random ReLU networks achieved significantly higher encoding performance than other random networks for both early and higher visual cortices. Furthermore, random networks with max pooling operations exhibited substantially higher performance than their counterparts with average pooling, except for ReLU networks, which achieved comparable encoding performance in both scenarios. However, the difference between the best performing random architecture and its trained counterpart was substantially smaller for early visual cortex in comparison to higher visual areas. In conclusion, we identified the ReLU activation function and maximum pooling as key architectural components that significantly contribute to the V1 encoding performance of CNNs. This is evidenced by the comparable performance achieved by randomly initialized models that incorporate these components as compared to their fully trained counterparts.

Complexity of deep neural networks explains their V1 encoding performance

We showed that random-weight CNNs varied substantially in their encoding performance of brain responses as a function of the non-linear activation function and pooling strategies, especially for early visual cortex (Fig. 2b). However, after having trained the CNNs to predict neural activity, they showed comparable performance across the nonlinear activation functions. Recent studies have shown that random-weight neural networks represent functions of different complexities depending on their architectural components such as their activation functions [Teney et al., 2024, Chiang et al., 2022, Ramasinghe et al., 2022]. We thus wondered whether complexity may explain the encoding performance of random-weight CNNs.

We approximated the function implemented by each output node in each model using a set of Chebyshev polynomials (see Methods). Complexity was then quantified as the average of the polynomial orders, weighted by their coefficients [Teney et al., 2024]. Fig. 3 shows the complexity distribution of output nodes in our models of V1 and IT responses, with both random-weight and trained convolutional kernels (analyses of fMRI models, see supplementary Fig. S5).

The random-weight CNNs with a low V1 and IT encoding performance (linear, ELU, Tanh with average pooling) generally showed low complexities as compared to their trained counterparts. By contrast, the random-weight CNNs with high V1 and IT encoding performance, such as ReLU and models with max pooling, had higher complexities than the random-weight CNNs with a low encoding performance. Finally, there was a greater overlap among the complexity distributions of models with trained convolutions than those with random convolutions, which is a consequence of the fact that the networks were optimized for the same target function. That is, training CNNs with a different architecture makes their complexity more homogeneous. For V1, random-weight CNNs with a comparable complexity to the trained counterparts also have similar encoding performance, suggesting that model complexity is a main driver of encoding performance. By contrast, for IT, there are major differences in encoding performance between random-weight and trained models despite similar complexity, suggesting that the specific configuration of weights is an additional important factor for IT.

We explored the relationship between model complexities and their encoding performance by plotting the median of the complexity distribution of each model's output nodes against its encoding performance (Fig. 4). To quantify the relationship between complexity and encoding performance, we fitted a quadratic function. We found a systematic relationship between the median complexity of the models and their encoding performance for V1 in both humans and macaques (Explained variance was 86%, 81%, and 84%

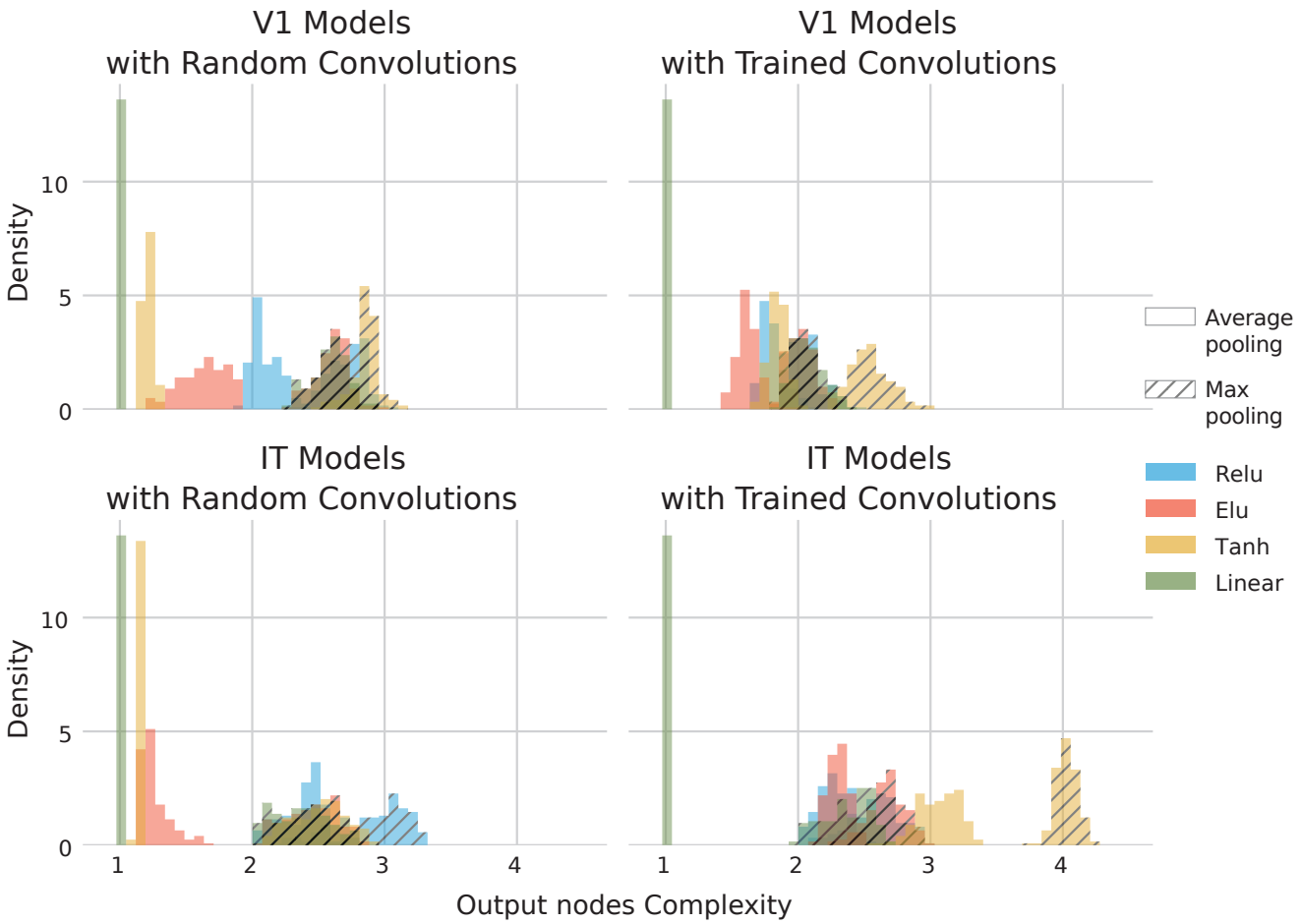

Figure 3. Distributions of encoding models' complexity across neurons. Each distribution across neurons represent the complexity (see methods) of a certain model configuration (with respect to the pooling strategy and activation function) trained on V1 (upper row) and IT (lower row) data. Only linear readout was trained on top of random convolutional features (left column) or the model was fully trained (right column). Each distribution is the average of 5 training iterations. The corresponding distributions for the human fMRI models are in supplementary Figure S2.

for the areas V1, V1v, and V1d respectively). For higher visual areas, the relationship was substantially weaker (Explained variance was 63%, 57%, and 55% for the areas IT, VO1, and VO2 respectively). Specifically, while random-weight models with similar model complexity as trained counterparts had comparable encoding performance for V1, there was major increase in encoding performance for IT. Together, these results suggests that for V1, complexity alone explains encoding performance, while for higher visual areas, the precise configurations of connection weights discovered through gradient descent are crucial for strong encoding performance. To further test the dependence of encoding performance on the precise configuration of weights, we shuffled the weights of all convolutional kernels of all layers of the trained models across all dimensions (space, input channels and output channels), freezing the weights, and subsequently retraining the linear readout. For a fair comparison, the deeper models (4 convolutional layers) were used for both early and higher visual areas. As anticipated, a much stronger decrease in encoding performance was observed for the shuffled models in higher visual areas compared to area V1 for both macaque (Fig. 6a) and human (Fig. 6b) brains.

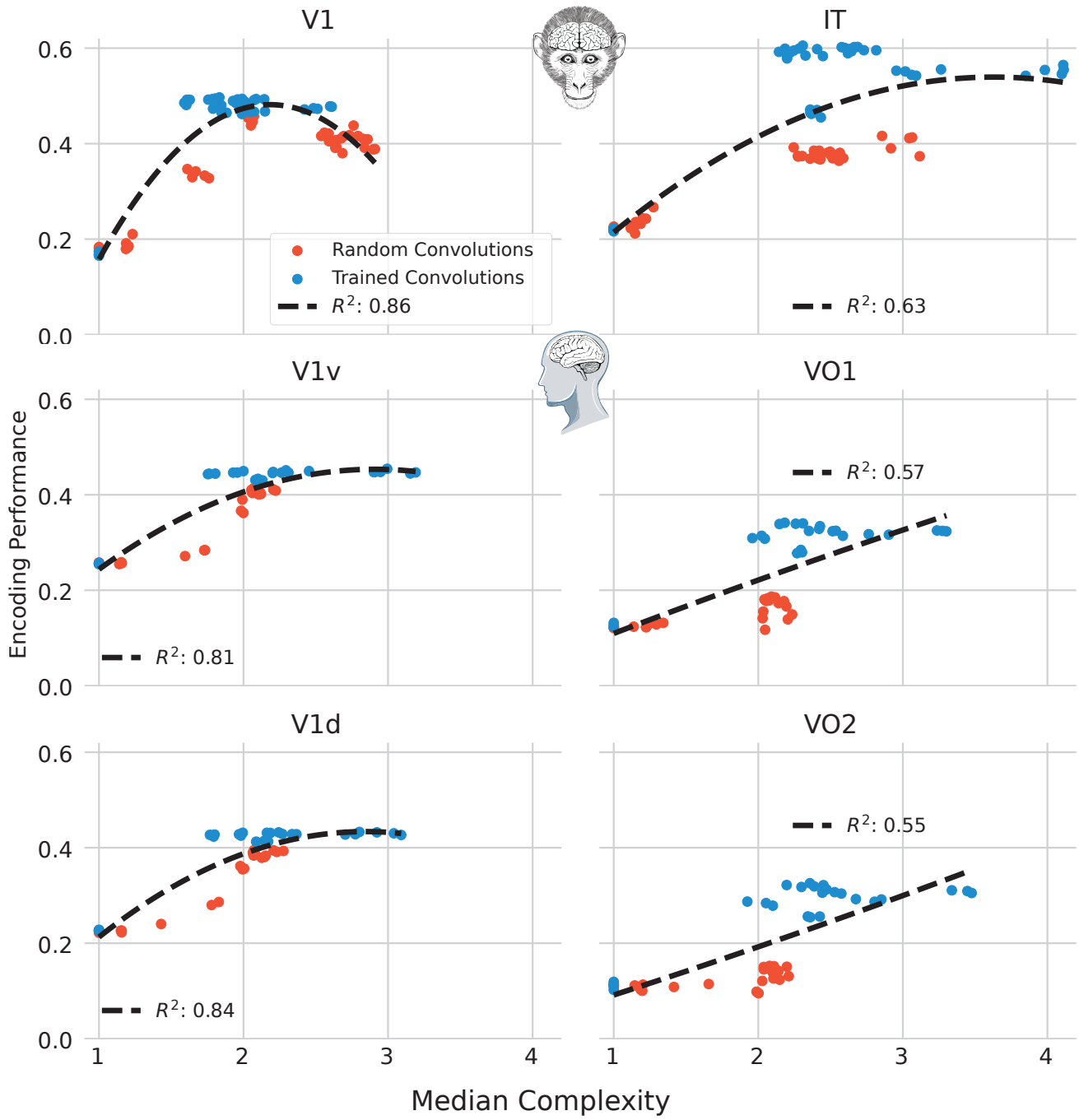

Figure 4. The relationship between the median complexity of the models and their encoding performance. Each panel shows the relationship between the median complexity of the models and their encoding performance for a certain brain area in macaques and humans. Models vary in their activation functions (4), pooling strategy (2) and wether only the readout was trained (in red) or the full model was trained (in blue). Each model configuration is represented 5 times that vary in their random initializations. The dashed black curve represent the fitted quadratic function to the data whose goodness of fit is quantified through R^2 printed in the legend.

Precise configuration of convolutional weights is critical for object recognition but not texture discrimination

Next, we investigated what kind of visual computations / tasks can be performed by random-weight CNNs, and which tasks are strongly dependent on training. To this end, we created a Texture-MNIST dataset for which two different tasks can be defined (Fig. 5b). Texture-MNIST is a dataset in which every sample is an MNIST digit filled with a texture batch (see Methods). Texture batches are randomly sampled from 10 high-quality texture images. We trained the 4-layer models to predict either the object (digit) identity or the texture patch identity. Similar to the neural data, we either trained only the readout, leaving the convolutional layers frozen at their randomly initialized state, or we trained both the readout and the convolutional layers. We observed that random-weight ReLU networks, with either average or maximum pooling, outperformed all other random-weight networks in predicting the correct identity of the texture class and the digit class (Fig. 5a). However, random-weight ReLU networks achieved almost the same performance as the trained networks on the texture discrimination task, while there was a major difference in performance for digit recognition task.

Similar to the neural data, we also investigated the dependence of task performance on the complexity of the models. We found that the complexity of the models showed a very strong relationship with texture discrimination accuracy (explained variance 95%) but not for digit recognition accuracy (explained variance 32%). These findings demonstrate that object recognition performance requires a precise optimization of convolutional kernels, while texture discrimination can already be subserved by random-weight neural networks. We further confirmed this observation by shuffling the weights of the optimized convolutional kernels then retraining the readout. Object recognition performance showed a stronger decrease than the texture discrimination performance (Fig. 6c).

Trained networks develop similar orientation selectivity to V1

We demonstrated that the representations of random-weight CNNs, with an architectural bias that entails sufficient model complexity, suffice for encoding the responses of early visual cortex in macaques and humans. However, it is well-established that V1 neurons exhibit selectivity for certain features, such as the orientation of a bar or grating stimulus [Hubel et al., 1959]. We sought to determine whether random-weight CNNs also possess such feature tuning. To test this, we generated Gabor patches of varying orientations and phases and presented them to the neural networks to assess their orientation selectivity, and then compared the selectivity distributions to experimental V1 data (Fig. 7a). We analyzed the central neurons in the last convolutional layer (i.e. those with a receptive field in the center of the image) of the random-weight neural networks and the networks trained to predict V1 activity. To compute the activation for each orientation, we averaged across all the different phases of the Gabor stimuli. In Fig. 7b, we show the four most orientation-selective neurons in each trained model (for random models, see supplementary Fig. S6a). We then quantified the neurons' orientation selectivity by calculating the circular variance of their tuning curves. Circular variance is a measure used to quantify the dispersion of data points around a circle, with a lower value indicating more concentrated responses, and thus, higher selectivity [Mazurek et al., 2014] (see Methods). We compared the circular variance distribution of the artificial neurons in the models (Fig. 7c for trained models and supplementary Fig. S6b for random-weight models) with the circular variance distribution of V1 neurons recorded from alert macaque monkeys [Gur et al., 2005]. To this end, we quantified the difference between distributions using the Wasserstein (i.e. Earth Mover) distance, which we term the "V1 deviation score" (lower scores indicate more similarity). Random-weight ReLU networks exhibited the lowest median circular variance among random-weight models (Fig. 7d), i.e. they were the most orientation-selective. Moreover, random-weight ReLU models also demonstrated the lowest V1 deviation score among all random-weight networks (Fig.

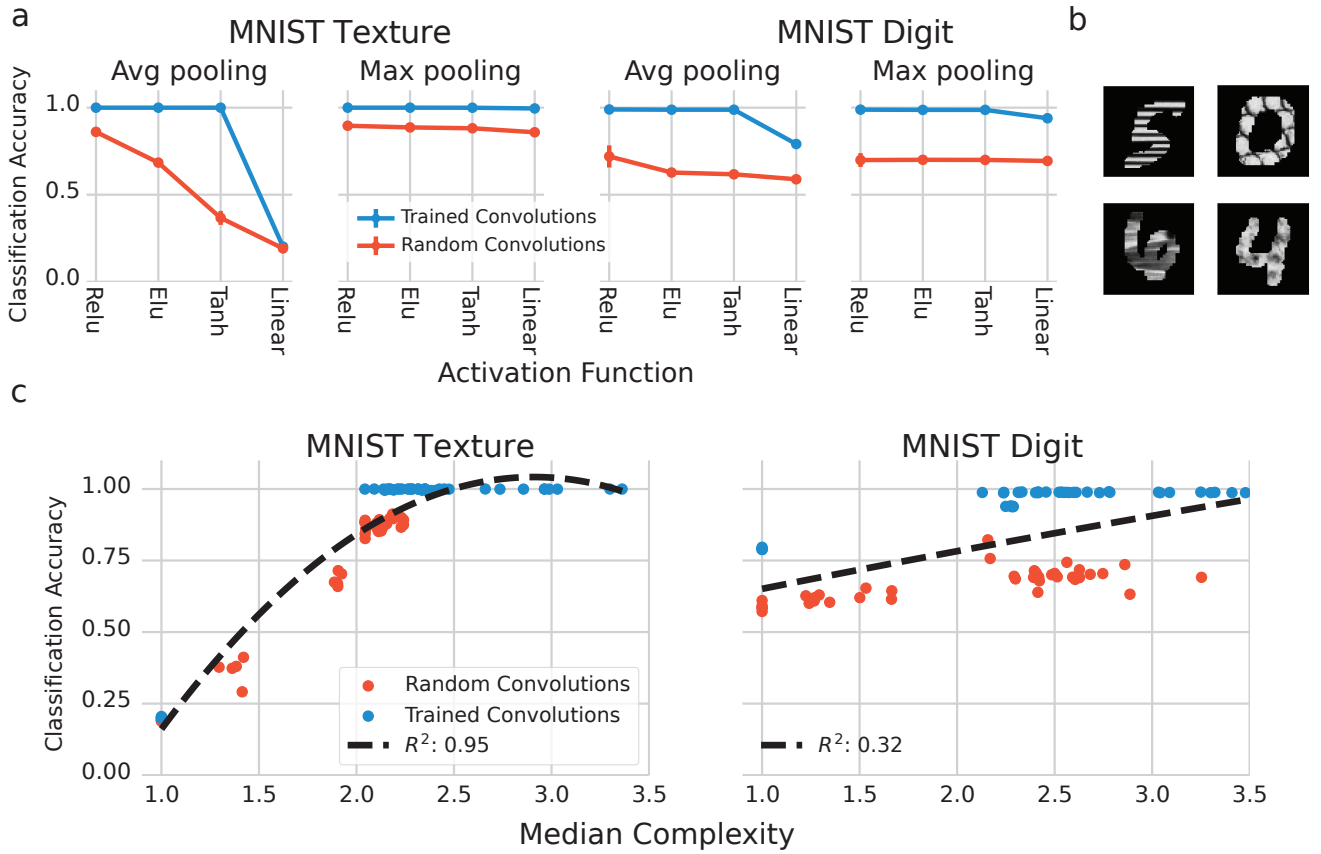

Figure 5. Texture discrimination and digit recognition performance on the Texture-MNIST dataset. **(a)** Performance of our 4-layer models trained on the Texture-MNIST dataset to either predict the texture identity (MNIST Texture) or the digit identity (MNIST Digit). Only the linear classification layer was trained (in red) or the full convolutional model was trained (in blue). Each line is the average of 5 training iterations. **(b)** Examples of the Texture-MNIST dataset. **(c)** The relationship between the median complexity of the mdoels and their texture discrimination accuracy (left) and digit recogntion accuracy (right).

7e). Furthermore, training the models on V1 data led to stronger orientation selectivity (i.e. a lower median circular variance) for all the models except the linear ones (Fig. 7d). Moreover, training the models on V1 data also led to lower V1 deviation scores for all of the models except the linear ones (Fig. 7e), with trained ReLU models being the most similar to the V1 orientation selectivity distribution. To examine whether the V1 deviation scores decreased because the networks were specifically trained to predict V1, we also tested the orientation selectivity of the models trained on IT data. We found that IT trained models were less orientation-selective than V1 trained models (supplementary Fig. S7a), and had higher V1 deviation scores (supplementary Fig. S7b).

To examine if networks with more similar orientation tuning to V1 also better predicted V1 activity (on another dataset), we examined the relationship between both variables across random and trained models (Fig. 7f). We found a strong negative monotonic relationship (Spearman correlation = -0.87) between the V1 deviation score and V1 encoding performance, suggesting that the more similar the orientation selectivity of the models is to that of V1 neurons, the better the models are at predicting V1 responses (on another dataset). However, it is also worth noting that although trained networks demonstrated small

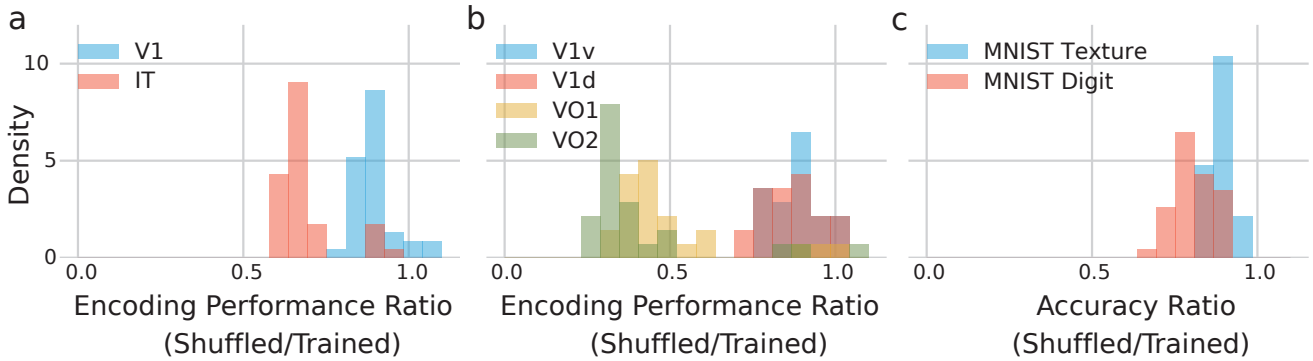

Figure 6. Effect of shuffling the convolutional weights on the neural encoding performance and classification accuracy. For each fully trained model, we shuffled the convolutional weights in the model and retrained the linear readout to predict the neural responses and to classify the texture/digit identity. Then we calculated the ratio between the performance of the model after shuffling and its performance before shuffling. **(a)** The distribution of the encoding performance ratio for all models across different configurations for the macaque brain areas. **(b)** same as **(a)** but for the fMRI human brain areas. **(c)** same as **(a)** but for the models trained for texture discrimination and digit recognition.

variation in their V1 encoding performance, they showed high variability in their V1 deviation score (Fig. 7f), suggesting that having similar orientation selecting to V1 neurons is not a strong requirement for the model to be able to effectively encode V1 responses.

Discussion

We investigated the factors contributing to the success of CNNs in predicting responses within the visual cortex in both macaques and humans. We demonstrate that, unlike higher visual areas, accurate prediction of early visual cortex responses does not necessarily depend on the optimization of convolutional kernels via training. Instead, architectural components, such as pooling strategies and activation functions, play a crucial role. Specifically, we found that the ReLU activation function and maximum pooling were the most critical components for achieving high encoding performance for early visual cortex, even in the absence of optimization of the convolutional kernels based on a task or neural data. Furthermore, we observed that random-weight CNNs exhibited substantial variability in the complexity of the functions they represented, depending on their architectural components. Notably, the CNN's complexity explained significantly more variance in the encoding performance in early visual cortex compared to higher visual areas. These findings held true across electrophysiological data from macaques and fMRI data from humans. Additionally, model complexity explained significantly more variance in performance on a texture discrimination task than on a digit (object) recognition task using the same dataset. This suggests that precise configuration of convolutional kernels is more essential for object recognition and, consequently, more critical for predicting responses in higher visual areas. Importantly, our results indicate that training the full convolutional models masked the effect of architectural bias, as the variability in encoding performance across different model configurations decreased significantly after training. However, when employing an alternative metric for aligning models with experimental V1 data – namely, the circular variance distribution of the model's artificial neurons as a proxy for their orientation selectivity – a different picture emerged: Trained models exhibited significant variability in their alignment with V1 orientation selectivity, despite displaying low variability in their V1 encoding performance. Overall, our findings

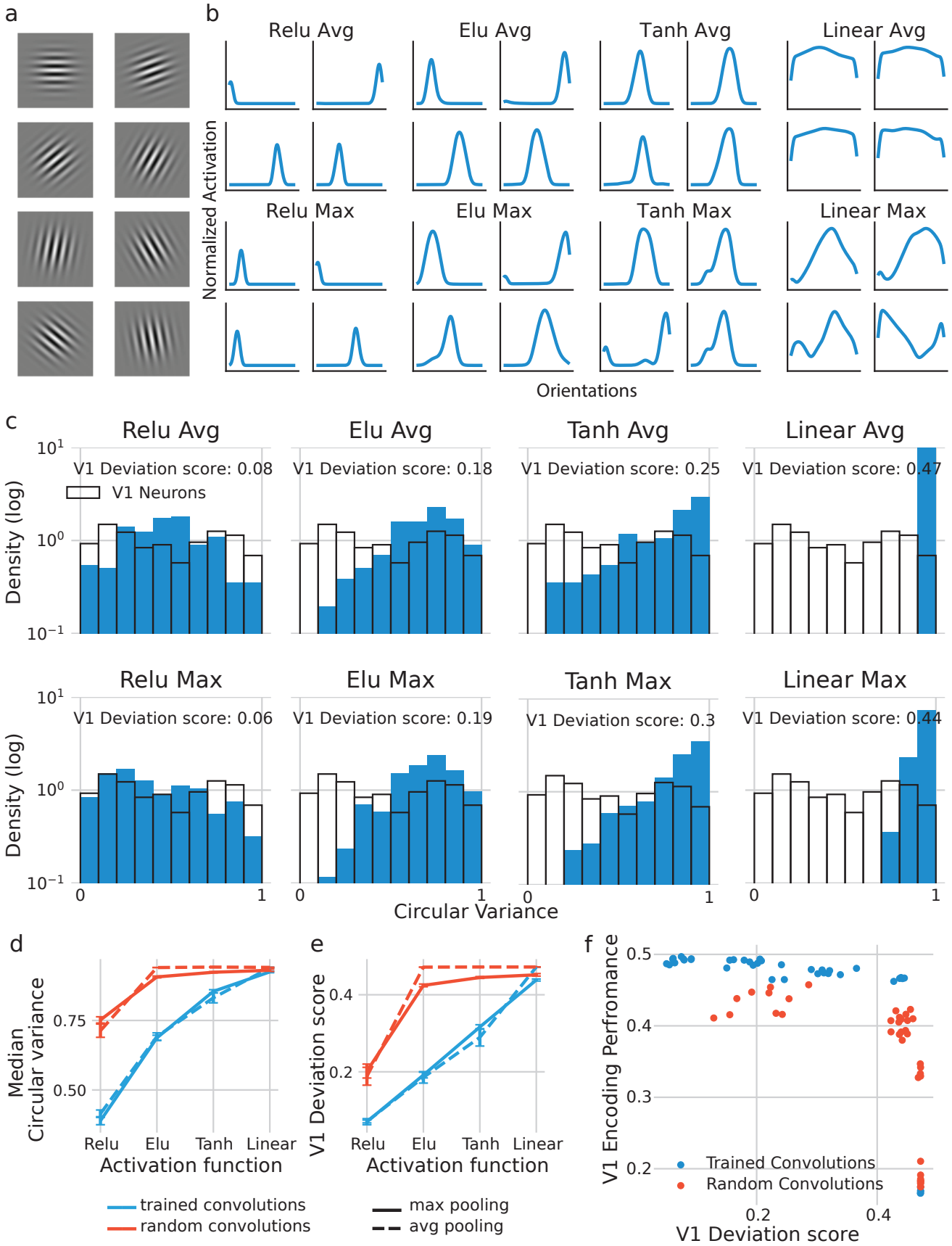
a
b
Relu Avg
Elu Avg
Tanh Avg
Linear Avg
Relu Max
Elu Max
Tanh Max
Linear Max
Normalized Activation
Orientations
c
Relu Avg
V1 Deviation score: 0.08
V1 Neurons
Elu Avg
V1 Deviation score: 0.18
Tanh Avg
V1 Deviation score: 0.25
Linear Avg
V1 Deviation score: 0.47
Relu Max
V1 Deviation score: 0.06
Elu Max
V1 Deviation score: 0.19
Tanh Max
V1 Deviation score: 0.3
Linear Max
V1 Deviation score: 0.44
Density (log)
Circular Variance
d
Median Circular variance
Relu
Elu
Tanh
Linear
Activation function
trained convolutions
random convolutions
e
V1 Deviation score
Activation function
max pooling
avg pooling
f
V1 Encoding Perfromance
Trained Convolutions
Random Convolutions
V1 Deviation score

Figure 7. Orientation selectivity of random and trained models **(a)** Examples of the Gabor patches that were represented to the models. **(b)** Tuning curves of the most orientation selective artificial neuron in the last convolutional layer of each model configuration trained on the V1 data. For the randomly initialized models see supplementary figure S2a. **(c)** Distributions of the circular variance of the artificial neurons of the last convolutional layer of each model configuration trained on the V1 data. For the randomly initialized models see supplementary figure S2b. **(d)** Median circular variance of the artificial neurons of the last convolutional layer of random and V1 trained models of different configurations averaged across 5 iterations. Error bars are the standard deviation. **(e)** V1 deviation scores of the random and V1 trained models averaged across 5 iterations. Error bars are the standard deviation. **(f)** The relationship between V1 deviation score and the V1 encoding performance for all the models that varied in their activation functions, pooling strategies, and their random initializations. Spearman correlation $= -0.87$, $p \ll 0.001$.

indicate that random-weight CNNs with sufficient model complexity allow for comparable prediction of V1 activity as trained CNNs, while higher visual areas require precise weight configurations acquired through training via gradient descent.

Our work highlights the importance of random-weight controls to reveal architectural bias. According to the deep learning framework for neuroscience [Richards et al., 2019], various deep neural network models, trained under specific constraints regarding their architecture, objective function, and training data, can serve as tests for specific hypotheses about brain function [Zhuang et al., 2021, Dwivedi et al., 2021, Doerig et al., 2023, Cichy and Kaiser, 2019]. The degree to which these models' representations of complex natural stimuli align with brain responses can provide evidence for certain hypotheses, such as the necessity of recurrent connections for cortical information processing [Kietzmann et al., 2019, Kubilius et al., 2019]. However, it has been shown that multiple models with different architectures, trained similarly, or multiple models with the same architecture, trained differently, can achieve similar performance in predicting neural data [Storrs et al., 2021, Conwell et al., 2024]. Further investigations revealed that task optimization increased the effective dimensionality [Del Giudice, 2021] of CNNs representations. This increase in dimensionality correlated with their encoding performance of higher visual brain areas [Elmoznino and Bonner, 2024]. A recent study showed that substantially scaling the dimensionality of random CNNs, but not transformers or fully connected networks, led to a comparable encoding performance of higher visual areas to ImageNet-trained AlexNet model suggesting a strong contribution of architecture bias to DNN encoding performance of neural data [Kazemian et al., 2024]. In the present work we showed that scaling the complexity of random CNNs, in the low-dimensional regime, leads to a comparable encoding performance for lower, but not higher visual areas. Moreover, it has been shown that layers of task-optimized VGG19 and Alexnet models with matching receptive field sizes yielded similar V1 encoding performance despite having different depths i.e. different numbers of nonlinear transformations suggesting the significance of receptive field size in predicting V1 data [Miao and Tong, 2024]. These findings underscore the necessity of implementing appropriate controls when employing task-optimized or brain-optimized DNNs for predicting neural activity. Specifically, the use of random-weight models with variable architectural components is crucial to reveal the architectural biases of neural encoding models.

Beyond the hypothesized neural factors influencing the neural encoding performance of DNNs, their success in predicting brain data could originate from non-neural, biologically implausible design choices implemented by researchers. For instance, the emergence of grid-like representations in DNNs optimized for path integration has been shown to depend on specifically designed readout mechanisms [Schaeffer et al., 2022]. Moreover, numerous studies have demonstrated that the mea-

sured similarity between models and the brain can be highly dependent on the chosen similarity metric [Soni et al., 2024, Davari et al., 2022]. Specifically, neural predictability scores based on linear regression can be heavily influenced by the inductive biases of linear regression, the dimensionality of the model representations, or the ratio of the number of stimuli in the benchmarking dataset to the dimensionality of the model representations [Canatar et al., 2024, Schaeffer et al., 2024, Bowers et al., 2023, Elmoznino and Bonner, 2024]. When we assessed the brain alignment of V1 models using two independent scores: encoding performance on one dataset and V1 deviation score on another dataset, we found large difference in the variability of both scores despite correlating with each other. Therefore, consistent with the existing literature, our findings highlight the importance of moving beyond single metrics of model-brain alignment. Instead, our findings highlight the importance of a multidimensional model assessment approach that enables the dissection of the similarities and differences between computational models and the brain [Wichmann and Geirhos, 2023, Biscione et al., 2024, Jacob et al., 2021, Rajesh et al., 2024]. This holds true in particular considering that many of these similarity measures are correlational [Bowers et al., 2023], and models with high prediction scores can still operate in qualitatively different ways than the brain [Geirhos et al., 2019, Geirhos et al., 2018, Baker et al., 2018, Farahat et al., 2023, Wichmann and Geirhos, 2023]. This multidimensional approach will facilitate more targeted model improvement and informed hypothesis generation in future research.

By systematically manipulating the architectural components of the models and comparing the performance of randomly initialized models with their fully trained counterparts, we identified the essential components that underpin neural encoding, raising the question to what extent they mimic the architecture of visual cortex. The ReLU activation function emerged as a key factor in generating visual representations that supported the most efficient encoding performance, considering the number of trainable parameters. The ReLU activation function was introduced to DNNs as a more biologically plausible alternative to Sigmoid and Tanh functions, given its one-sided nature (outputting zero for negative inputs) and its promotion of sparse representations [Glorot et al., 2011, Attwell and Laughlin, 2001, Douglas et al., 1995]. Indeed, ReLU networks, even with random convolutions, not only exhibited the best encoding performance for V1 but also displayed the smallest distance to the orientation selectivity distribution of V1 neurons. Importantly, while models fully fitted to predict V1 data, with ReLU, ELU, or Tanh activation functions, exhibited similar encoding performance, they still displayed substantial variability in their similarity to V1 orientation selectivity, with ReLU networks being the most V1-like. This result demonstrates that the combined application of multiple model assessment metrics and systematic architectural manipulations enables the identification of key, potentially biologically plausible architectural components that contribute significantly to neural encoding performance. Furthermore, the fact that non-linearities with sufficient model complexity are a major factor in predicting neural activity fits with the general idea that non-linearities are a central component of cortical inter-areal interactions beyond mere linear information transmission [Vinck et al., 2023, DiCarlo et al., 2012].

Beyond examining the encoding performance and tuning properties of the models' representations, it is imperative to understand the computational advantages of models' representations in supporting visual tasks. We demonstrated that random-weight CNN representations, which were sufficient for predicting early visual cortex responses, performed well in discriminating between texture families compared to their trained counterparts. Conversely, these random-weight representations were considerably worse than fully trained models in invariantly classifying the identity of digits within images. Studies have shown that V1 activity, in particular superficial cortex, exhibits selectivity for texture statistics, albeit less pronounced than in a higher visual area, LM [Bolaños et al., 2024, Ziemba et al., 2019]. In humans, texture discrimination task-learning has been shown to induce local changes within the early visual cortex without requiring the recruitment of higher visual areas [Schwartz et al., 2002]. Additionally, a decoder

trained on macaque V1 population activity elicited by texture samples could discriminate between 15 different texture families [Ziemba et al., 2016]. In contrast, several studies have demonstrated that IT neurons possess the tolerance to identity-preserving transformations that is essential for object recognition [Rust and DiCarlo, 2010, Hung et al., 2005].

The efficacy of random features in machine learning has been well-documented, often rivaling hand-crafted or even learned features across various learning tasks [Rahimi and Recht, 2008a, Gallicchio and Scardapane, 20 Scardapane and Wang, 2017]. Random-weight CNNs were shown to be frequency-selective and translation-invariant which explains their superior performance over random non-convolutional networks on image classification tasks [Saxe et al., 2011]. Moreover, only training a small fraction of the convolutional weights or only training the batch normalization layers in random-weight CNNs led to object recognition performance competitive with their trained counterparts [Rosenfeld and Tsotsos, 2019, Frankle et al., 2021]. Furthermore, the structure of a random generator CNN can capture significant low-level image statistics even without any learning. This inherent structure acts as a prior, making random-weight CNNs useful for various image processing tasks such as image restoration, denoising, inpainting and super-resolution [Ulyanov et al., 2018]. These findings emphasize the significant contribution of convolution and pooling operations, independent of learning, in visual processing tasks. Consequently, it is plausible that random features with the right convolutional architectural bias could effectively model the representations found in V1, considering the diverse range of visual tasks that V1 supports. One hypothesis is that V1 comprises an array of neurons representing high-dimensional, non-linear random basis functions, capable of supporting a diverse set of downstream functions [Rahimi and Recht, 2007, Rahimi and Recht, 2008a, Rahimi and Recht, 2008b]. In addition to our results that showed that random-weight ReLU networks exhibit orientation-selective neurons, previous research on biologically plausible recurrent models of mouse V1 demonstrated the emergence of orientation selectivity even when both feedforward and recurrent connections are randomly initialized [Hansel and van Vreeswijk, 2012, Pattadkal et al., 2018].

Our findings contribute to the growing body of literature that emphasizes the importance of conducting controlled experiments to systematically investigate the architectural and training components that contribute to the neural encoding performance of computational models. Moreover, our results underscore the necessity of developing comprehensive batteries of neural and perceptual metrics to facilitate more informed conclusions about the similarities between computational models and the brain [Biscione et al., 2024, Jacob et al., 2021]. Finally, considering the computational benefits of the models' representations that support the prediction of brain responses is valuable, as it helps formulate hypotheses regarding the functional roles of different brain areas [Dwivedi et al., 2021, Cichy and Kaiser, 2019].

Methods

Datasets

V1 monkey dataset

We used a public dataset that consists of neural activity recordings from 166 neurons across different layers of V1 brain area in two monkeys [Cadena et al., 2019]. The monkeys were shown 7,250 images, each presented 1-4 times for a duration of 60 milliseconds. Each image was displayed within a circular window spanning 2 degrees of visual angle, with the edges gradually fading out to blend with the surroundings.

IT monkey dataset

We used a publicly available IT monkey dataset which consists of neural recordings from 168 multiunit sites within the inferotemporal (IT) cortex of two macaque monkeys [Cadieu et al., 2014]. The monkeys

were presented with 3,200 unique grayscale images, each showing one of 64 objects from eight categories. These images were designed to mimic real-world visual scenes by placing the cropped object images onto various natural image backgrounds at different positions, orientations, and sizes.

fMRI human dataset

The Natural Scenes Dataset (NSD) is a publicly available fMRI dataset that captures the brain activity of eight human participants as they viewed thousands of natural images (9,000–10,000 distinct color natural images for each subject repeated up to 3 trials) [Allen et al., 2022]. The images were taken from the Microsoft Common Objects in Context (COCO) database square-cropped and presented at a size of 8.4° x 8.4°. We used the regions of interest (ROI) V1v and V1d manually drawn based on the results of a population receptive field (pRF) experiment, and the higher-order ROIs VO1 and VO2, defined using a visual probabilistic atlas.

Texture-MNIST dataset

We created the Texture-MNIST dataset to probe different models' texture and shape discrimination abilities. We created binary masks from the MNIST dataset, resized and overlaid them over 64×64 patches of texture randomly copped from a high-resolution texture image unique for each digit class. Using this dataset, we can train our models to either predict the class of the object (digit) or the class of the texture of each image.

Models

We used simple DNN models consisting of a convolutional block and a linear readout. The convolutional block included two and four convolutional layers for the shallow and deeper models respectively. Each convolutional layer is followed by a batch normalization layer and an activation function. To maintain an efficient number of trainable parameters we used depthwise separable convolutions in all convolutional layers except for the first one [Du et al., 2024]. Furthermore, shallow models had 16 and 256 feature maps in their 2 convolutional layers, whereas deeper models had feature maps that progressively increased from 16 to 32, 64, and finally 256 across the network depth. For the shallow models, we had a pooling layer after each activation function and for the deeper models, we had the pooling layer after every other activation layer. For the shallow models, convolutional layers had either filter size of 9×9 or 17×17 pixels, leading to an effective receptive field of the models of 28×28 or 52×52 respectively. For the deeper models, we had filter sizes of 5×5 or 9×9 pixels leading to the same effective receptive fields as the sallow models (see supplementary Fig. S1 for an illustration of the detailed architecture of the shallow and deep 28×28 models). We tested a variety of activation functions including ReLU, ELU, Tanh, and Linear. Moreover, we considered average and maximum pooling operations, each with a pooling window of 2×2 pixels.

The three-dimensional activation maps of the convolutional block were transformed to the neural responses through a linear readout factorized using three one-dimensional weight vectors $\mathbf{w}_c$, $\mathbf{w}_x$, and $\mathbf{w}_y$ for the channels, and two spatial dimensions respectively. For the image classification tasks, global average pooling was applied to the three-dimensional activation maps to obtain the feature vector used for classification.

Complexity measurement

To calculate the complexity of the function represented by a neural network, we evaluate the network on a regularly sampled grid in its input space [Teney et al., 2024]. Our networks were trained with input normalized to the range from -1 to 1. Therefore, we sampled 100 corners in the hypercube $[-1,1]^d$, where d is the input dimension of the network. We sampled 50 points regularly on each of the 100 lines connecting each corner with its succeeding corner and evaluated the network at each sample input point.

Let x be the regularly sampled line in the range $[-1,1]$, and y be the activation of a certain output node evaluated at the data points lying on that line in the input space hypercube. We compute the coefficients $\mathbf{c}$ of Chebyshev polynomials that fit the data (x,y) by minimizing the least square error:

$$E = \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{d} c_k T_k(x_i) \right)^2$$

where k is the polynomial order and $T_k(x_i)$ is the $k-$th Chebyshev polynomial evaluated at the i-th input x_i. The Chebyshev polynomials $T_k(x)$ are recursively defined as:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$$

We define the complexity metric as the average of the polynomial orders weighted by their corresponding Chebyshev coefficients:

$$\text{Complexity}_{Chebychev} = \frac{\sum_{k=0}^{d} c_k k}{\sum_{k=0}^{d} c_k}$$

For each output node, we average the 100 complexity metrics from the 100 lines to obtain one complexity measurement per one output node.

Comparison to experimental V1 data

To go beyond regression, we tested the alignment of the artificial neurons in our models with experimental data recorded from the V1 area by comparing their orientation selectivity distributions. We presented the models with Gabor patches of different orientations and phases [Kong et al., 2022]. Orientations were sampled at 10° steps in the range from 0° to 180° and 20 phases were sampled evenly from 0° to 360°. The spatial frequency of the Gabor patches was chosen to allow the receptive field of the model neurons to contain two cycles i.e. for the receptive field of 28×28 pixels and input size of 128×128 pixels, we used a spatial frequency of 9.142 cycles/image. Specifically, we generated the Gabor patches according to the formula:

$$f(x,y;\sigma,\lambda,\psi,\theta,\gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(\frac{2\pi x'}{\lambda} + \psi\right)$$

and

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

where $\sigma = 20$ pixels is the standard deviation of the Gaussian envelope which controls the size of the Gabor patch, λ is the wavelength of the sinusoidal factor in pixels, ψ is the phase offset of the sinusoidal factor, θ is the orientation of the Gabor patch in radians and $\gamma = 1$ is the spatial aspect ratio, specifying the ellipticity of the support of the Gabor function.

We calculated the orientation tuning curves of the center location of every artificial neuron at the last convolutional layer of our models. The tuning curves were scaled to the range from 0 to 1 to avoid negative responses in activation functions such as Tanh, Elu, and Linear. To establish a baseline for artificial neuron responses, the minimum response value across the V1 test set was determined. This minimum value was then used in the scaling of the tuning curves. Then we followed the analysis steps mentioned in [Gur et al., 2005] to calculate the orientation selectivity distribution of the V1 neurons recorded from awake monkeys. Briefly, we linearly interpolated the tuning curves with 1° steps, then smoothed them with a Hanning filter with a 7° half-width at half-height. We then quantified the orientation selectivity

of neurons from the tuning curves by calculating the circular variance (CV) [Mazurek et al., 2014]. The circular variance was calculated from the smoothed tuning curves resampled at regular 15° intervals according to the equation:

$$CV = 1 - \frac{|\sum_k r_k e^{i2\theta_k}|}{\sum_k r_k}$$

where θ_k is the orientation in radians and r_k is the corresponding response.

For each model, we simulated 100 in-silico electrophysiology experiments by randomly sampling with replacement 339 neurons from its last convolutional layer. We calculated their CV as described. From each experiment, we compared the distribution of CV with the corresponding distribution obtained from 339 V1 neurons recorded from alert macaques (Figure 3 in [Gur et al., 2005]) by calculating a V1 deviation score and then averaging the scores over the 100 experiments to obtain one score per model. The V1 deviation score was computed as the Wasserstein distance between the distribution of circular variance of the model's neurons and the corresponding distribution of experimental V1 neurons.

Acknowledgments

This project was financed by the BMF (Bundesministerium fuer Bildung und Forschung), Computational Life Sciences, project BINDA (031L0167); an ERC starting grant (850861) SPATEMP; DFG VI Grants (908/5-1 and 908/7-1); an NWO VIDI Grant; and the Dutch Brain Interface Initiative (DBI2).

References

Allen et al., 2022. Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Attwell and Laughlin, 2001. Attwell, D. and Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21(10):1133–1145.

Baker et al., 2018. Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613.

Biscione et al., 2024. Biscione, V., Yin, D., Malhotra, G., Dujmovic, M., Montero, M. L., Puebla, G., Adolfi, F., Heaton, R. F., Hummel, J. E., Evans, B. D., et al. (2024). Mindset: Vision. a toolbox for testing dnns on key psychological experiments. *arXiv preprint arXiv:2404.05290*.

Bolaños et al., 2024. Bolaños, F., Orlandi, J. G., Aoki, R., Jagadeesh, A. V., Gardner, J. L., and Benucci, A. (2024). Efficient coding of natural images in the mouse visual cortex. *Nature Communications*, 15(1):2466.

Bowers et al., 2023. Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385.

Cadena et al., 2019. Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897.

Cadieu et al., 2014. Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963.

Canatar et al., 2024. Canatar, A., Feather, J., Wakhloo, A., and Chung, S. (2024). A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, 36.

Chiang et al., 2022. Chiang, P.-y., Ni, R., Miller, D. Y., Bansal, A., Geiping, J., Goldblum, M., and Goldstein, T. (2022). Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*.

Cichy and Kaiser, 2019. Cichy, R. M. and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317.

Conwell et al., 2024. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383.

Davari et al., 2022. Davari, M., Horoi, S., Natik, A., Lajoie, G., Wolf, G., and Belilovsky, E. (2022). Reliability of cka as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*.

Del Giudice, 2021. Del Giudice, M. (2021). Effective dimensionality: A tutorial. *Multivariate behavioral research*, 56(3):527–542.

DiCarlo et al., 2012. DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Doerig et al., 2023. Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., Van Gerven, M. A., Kriegeskorte, N., et al. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450.

Douglas et al., 1995. Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., and Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, 269(5226):981–985.

Du et al., 2024. Du, F., Núñez-Ochoa, M. A., Pachitariu, M., and Stringer, C. (2024). Towards a simplified model of primary visual cortex. *bioRxiv*, pages 2024–06.

Dwivedi et al., 2021. Dwivedi, K., Bonner, M. F., Cichy, R. M., and Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS computational biology*, 17(8):e1009267.

Elmoznino and Bonner, 2024. 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):e1011792.

Farahat et al., 2023. Farahat, A., Effenberger, F., and Vinck, M. (2023). A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations. *Neural networks*, 167:400–414.

Felleman and Van Essen, 1991. Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47.

Frankle et al., 2021. Frankle, J., Schwab, D. J., and Morcos, A. S. (2021). Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *International Conference on Learning Representations*.

Fukushima et al., 1983. Fukushima, K., Miyake, S., and Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834.

Gallicchio and Scardapane, 2020. Gallicchio, C. and Scardapane, S. (2020). Deep randomized neural networks. In *Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*, pages 43–68. Springer.

Geirhos et al., 2019. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Geirhos et al., 2018. Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

Gieselmann and Thiele, 2008. Gieselmann, M. and Thiele, A. (2008). Comparison of spatial integration and surround suppression characteristics in spiking activity and the local field potential in macaque v1. *European Journal of Neuroscience*, 28(3):447–459.

Glorot et al., 2011. Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Güçlü and Van Gerven, 2015. Güçlü, U. and Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Gur et al., 2005. Gur, M., Kagan, I., and Snodderly, D. M. (2005). Orientation and direction selectivity of neurons in v1 of alert monkeys: functional relationships and laminar distributions. *Cerebral Cortex*, 15(8):1207–1221.

Gur and Snodderly, 2007. Gur, M. and Snodderly, D. M. (2007). Direction selectivity in v1 of alert monkeys: evidence for parallel pathways for motion processing. *The Journal of physiology*, 585(2):383–400.

Hansel and van Vreeswijk, 2012. Hansel, D. and van Vreeswijk, C. (2012). The mechanism of orientation selectivity in primary visual cortex without a functional map. *Journal of Neuroscience*, 32(12):4049–4064.

He et al., 2016. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer.

Huang et al., 2016. Huang, G., Liu, Z., and Weinberger, K. Q. (2016). Densely connected convolutional networks. corr. *arXiv preprint arXiv:1608.06993*.

Hubel et al., 1959. Hubel, D. H., Wiesel, T. N., et al. (1959). Receptive fields of single neurones in the cat's striate cortex. *J physiol*, 148(3):574–591.

Hung et al., 2005. Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866.

Jacob et al., 2021. Jacob, G., Pramod, R., Katti, H., and Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, 12(1):1872.

Kazemian et al., 2024. Kazemian, A., Elmoznino, E., and Bonner, M. F. (2024). Convolutional architectures are cortex-aligned de novo. *bioRxiv*, pages 2024–05.

Khaligh-Razavi and Kriegeskorte, 2014. Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915.

Kietzmann et al., 2019. Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863.

Kong et al., 2022. Kong, N. C., Margalit, E., Gardner, J. L., and Norcia, A. M. (2022). Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739.

Kubilius et al., 2019. Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32.

LeCun et al., 1989. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Li, 2002. Li, Z. (2002). A saliency map in primary visual cortex. *Trends in cognitive sciences*, 6(1):9–16.

Mazurek et al., 2014. Mazurek, M., Kager, M., and Van Hooser, S. D. (2014). Robust quantification of orientation selectivity and direction selectivity. *Frontiers in neural circuits*, 8:92.

Mehrer et al., 2021. Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118.

Miao and Tong, 2024. Miao, H.-Y. and Tong, F. (2024). Convolutional neural network models applied to neuronal responses in macaque v1 reveal limited nonlinear processing. *Journal of Vision*, 24(6):1–1.

Pattadkal et al., 2018. Pattadkal, J. J., Mato, G., van Vreeswijk, C., Priebe, N. J., and Hansel, D. (2018). Emergent orientation selectivity from random networks in mouse visual cortex. *Cell reports*, 24(8):2042–2050.

Peter et al., 2019. Peter, A., Uran, C., Klon-Lipok, J., Roese, R., van Stijn, S., Barnes, W., Dowdall, J. R., Singer, W., Fries, P., and Vinck, M. (2019). Surface color and predictability determine contextual modulation of v1 firing and gamma oscillations. *elife*, 8:e42101.

Rahimi and Recht, 2007. Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Rahimi and Recht, 2008a. Rahimi, A. and Recht, B. (2008a). Uniform approximation of functions with random bases. In *2008 46th annual allerton conference on communication, control, and computing*, pages 555–561. IEEE.

Rahimi and Recht, 2008b. Rahimi, A. and Recht, B. (2008b). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21.

Rajesh et al., 2024. Rajesh, N., Jacob, G., and Arun, S. (2024). Brain-like emergent properties in deep networks: impact of network architecture, datasets and training. *arXiv preprint arXiv:2411.16326*.

Ramasinghe et al., 2022. Ramasinghe, S., MacDonald, L., Farazi, M., Saratchandran, H., and Lucey, S. (2022). How you start matters for generalization. *arXiv preprint arXiv:2206.08558*.

Richards et al., 2019. Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770.

Rosenfeld and Tsotsos, 2019. Rosenfeld, A. and Tsotsos, J. K. (2019). Intriguing properties of randomly weighted networks: Generalizing while learning next to nothing. In *2019 16th conference on computer and robot vision (CRV)*, pages 9–16. IEEE.

Rust and DiCarlo, 2010. Rust, N. C. and DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995.

Saxe et al., 2011. Saxe, A. M., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., and Ng, A. Y. (2011). On random weights and unsupervised feature learning. In *Icml*, volume 2, page 6.

Scardapane and Wang, 2017. Scardapane, S. and Wang, D. (2017). Randomness in neural networks: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2):e1200.

Schaeffer et al., 2024. Schaeffer, R., Khona, M., Chandra, S., Ostrow, M., Miranda, B., and Koyejo, S. (2024). Position: Maximizing neural regression scores may not identify good models of the brain. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Schaeffer et al., 2022. Schaeffer, R., Khona, M., and Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, 35:16052–16067.

Schwartz et al., 2002. Schwartz, S., Maquet, P., and Frith, C. (2002). Neural correlates of perceptual learning: a functional mri study of visual texture discrimination. *Proceedings of the National Academy of Sciences*, 99(26):17137–17142.

Self et al., 2013. Self, M. W., van Kerkoerle, T., Supèr, H., and Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area v1 in figure-ground segregation. *Current biology*, 23(21):2121–2129.

Simoncelli et al., 2004. Simoncelli, E. P., Paninski, L., Pillow, J., Schwartz, O., et al. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, 3(327-338):1.

Soni et al., 2024. Soni, A., Srivastava, S., Khosla, M., and Kording, K. P. (2024). Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, pages 2024–08.

Storrs et al., 2021. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., and Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064.

Szegedy et al., 2015. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. 2015. *arXiv preprint arXiv:1512.00567*.

Teney et al., 2024. Teney, D., Nicolicioiu, A. M., Hartmann, V., and Abbasnejad, E. (2024). Neural redshift: Random networks are not random functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4786–4796.

Ulyanov et al., 2018. Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454.

Uran et al., 2022. Uran, C., Peter, A., Lazar, A., Barnes, W., Klon-Lipok, J., Shapcott, K. A., Roese, R., Fries, P., Singer, W., and Vinck, M. (2022). Predictive coding of natural images by v1 firing rates and rhythmic synchronization. *Neuron*, 110(7):1240–1257.

Vezoli et al., 2021. Vezoli, J., Magrou, L., Goebel, R., Wang, X.-J., Knoblauch, K., Vinck, M., and Kennedy, H. (2021). Cortical hierarchy, dual counterstream architecture and the importance of top-down generative networks. *Neuroimage*, 225:117479.

Vinck and Bosman, 2016. Vinck, M. and Bosman, C. A. (2016). More gamma more predictions: gamma-synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. *Frontiers in systems neuroscience*, 10:35.

Vinck et al., 2023. Vinck, M., Uran, C., Spyropoulos, G., Onorato, I., Broggini, A. C., Schneider, M., and Canales-Johnson, A. (2023). Principles of large-scale neural interactions. *Neuron*, 111(7):987–1002.

Wichmann and Geirhos, 2023. Wichmann, F. A. and Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9(1):501–524.

Willmore et al., 2008. Willmore, B., Prenger, R. J., Wu, M. C.-K., and Gallant, J. L. (2008). The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform. *Neural computation*, 20(6):1537–1564.

Yamins et al., 2014. Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Zhuang et al., 2021. Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118.

Ziemba et al., 2016. Ziemba, C. M., Freeman, J., Movshon, J. A., and Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, 113(22):E3140–E3149.

Ziemba et al., 2019. Ziemba, C. M., Perez, R. K., Pai, J., Kelly, J. G., Hallum, L. E., Shooner, C., and Movshon, J. A. (2019). Laminar differences in responses to naturalistic texture in macaque v1 and v2. *Journal of Neuroscience*, 39(49):9748–9756.

Supplementary Figures

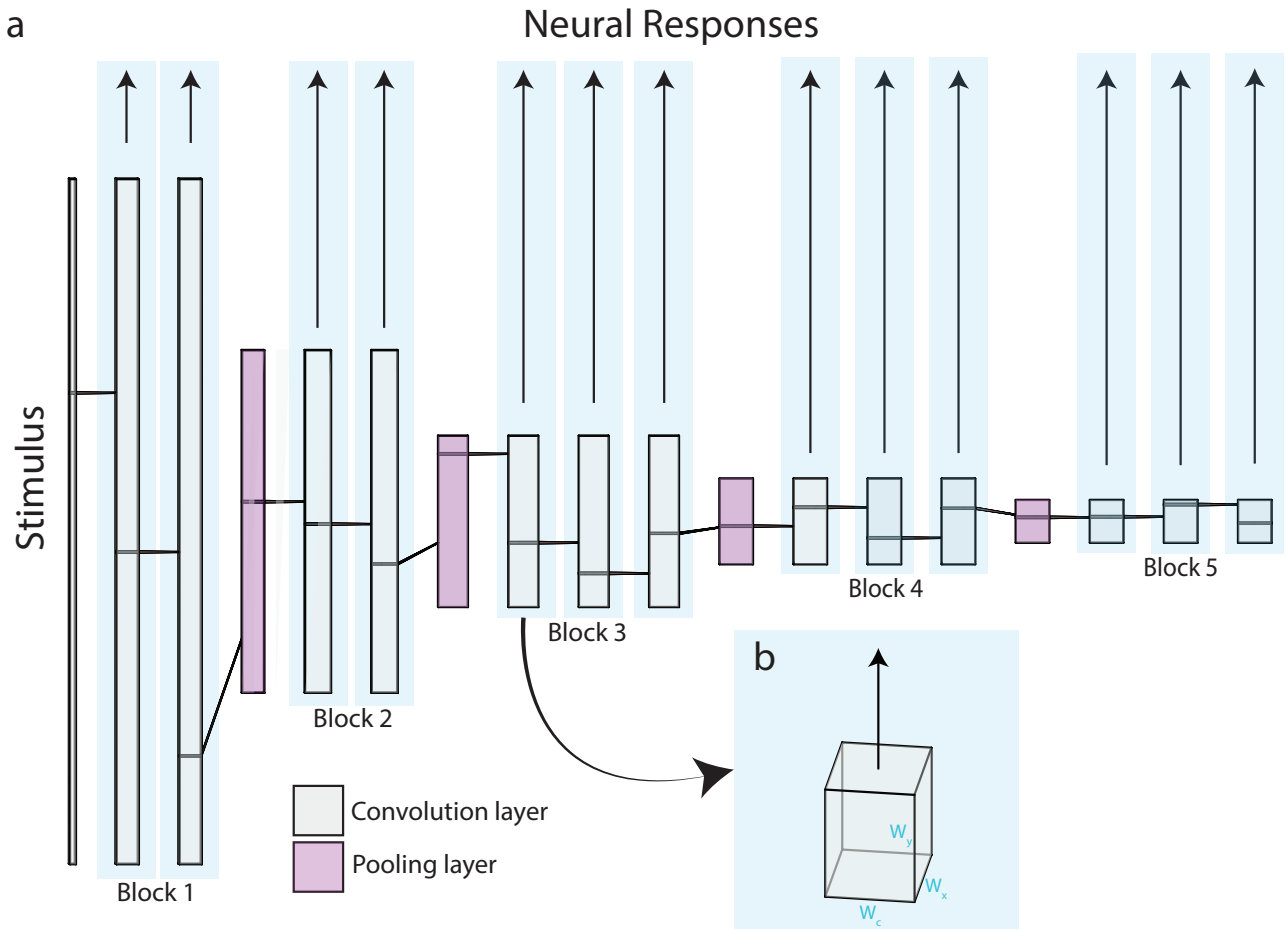

Figure S1. VGG16 encoding models. (a) On top of the three-dimensional feature maps of each convolutional layer (gray rectangles), we trained a linear readout (arrows) to predict the neural responses of a certain brain area. We used ImageNet-trained VGG16 model and randomly-initialized variants. **(b)** Linear readout was factorized into 3 one-dimensional weight vectors $\mathbf{w}_c$, $\mathbf{w}_x$, and $\mathbf{w}_y$ for the channels and the two spatial dimensions respectively.

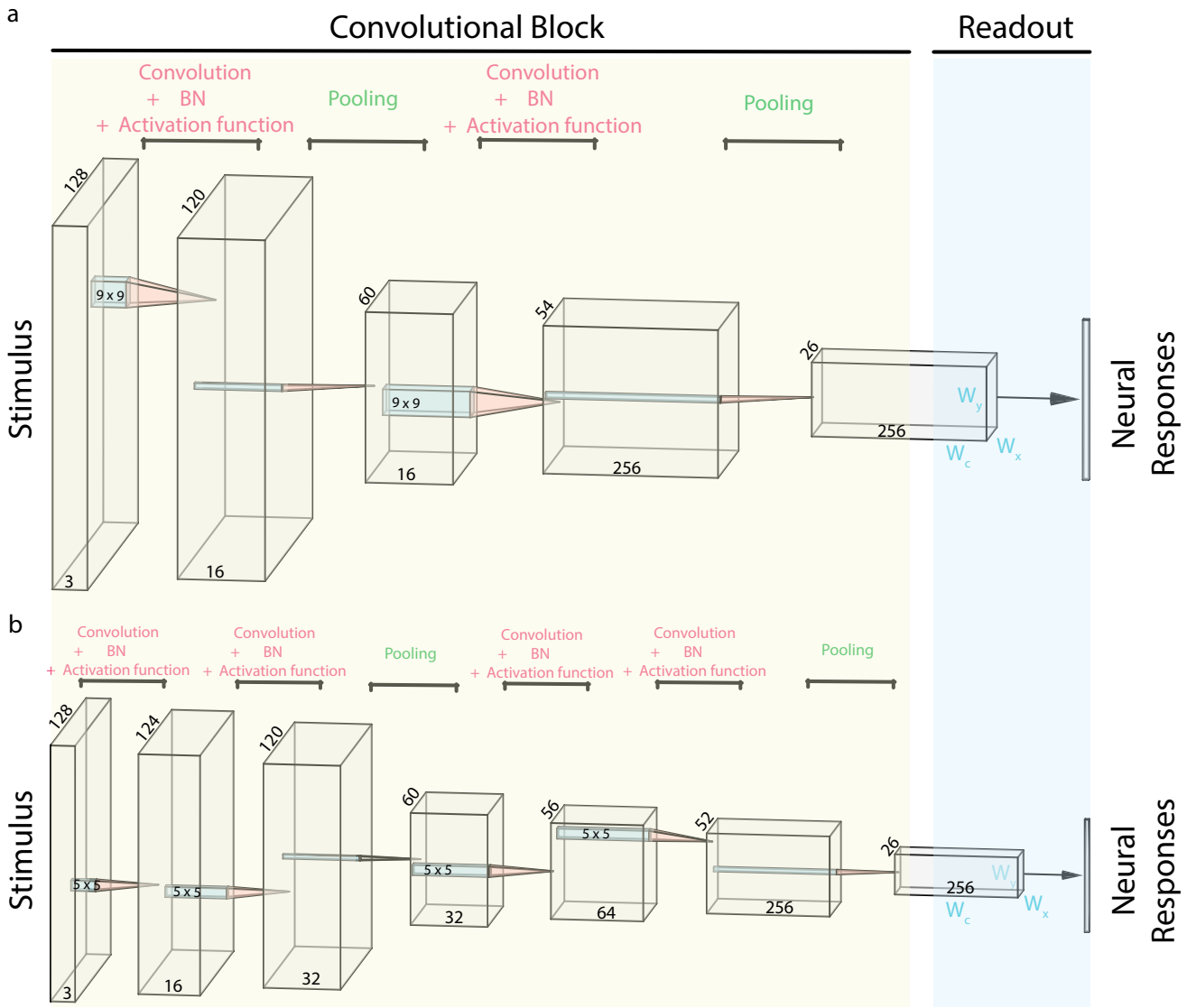

Figure S2. Architecture of CNNs used for encoding neural responses. All models are formed of a convolutional block and a linear readout. Readout is factorized into three one-dimensional vectors ($\mathbf{w}_c$, $\mathbf{w}_x$, and $\mathbf{w}_y$) that transform the channels and the two spatial dimensions of the feature maps of the last convolutional layer into neural activity. **(a)** The convolutional block of shallow models used for encoding early visual cortex activity is formed of two convolutional layers with 9×9 filter sizes. Each layer is followed by a batch normalization (BN) layer, an activation function (ReLU, ELU, Tanh, or linear), and a pooling layer of 2×2 window size and stride = 2 (maximum or average pooling). **(b)** The convolutional block of the deeper models used for encoding higher visual areas is formed of 4 convolutional layers with 5×5 filter sizes. Each layer is followed by a batch normalization (BN) layer and an activation function (ReLU, ELU, Tanh, or linear). Every other layer is followed by a pooling layer of 2×2 window size and stride = 2 (maximum or average pooling). The spatial resolution of feature maps is printed on top of each block of feature maps. Number of feature maps (channels) is printed at the bottom. All convolutional layers are depthwise separable except for the first one. The effective receptive field of neurons in the feature maps of the last convolutional layer of both models is 28×28 pixels.

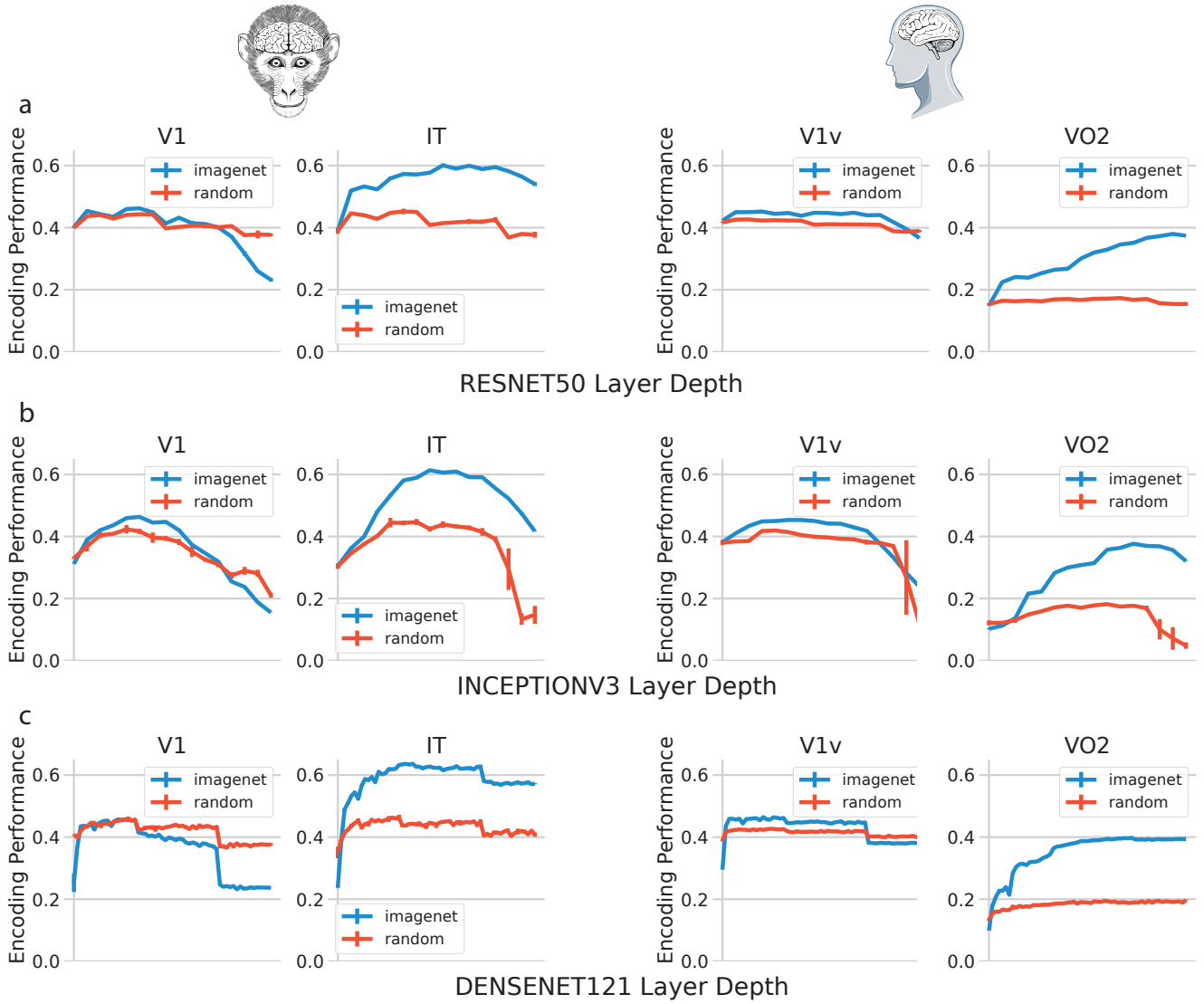

Figure S3. Encoding performance of popular convolutional architectures. **(a)** Encoding performance of a linear readout optimized on top of the representations of the convolutional layers of RESNET50 model either upon random initialization (in red) or pretraiend on ImageNet dataset for object recognition (in blue) for four neural datasets: Two electrophysiological datasets recorded from macaques: the early visual cortex V1 and the higher visual area IT and two fMRI datasets recorded from humans: the early visual cortex V1v and the higher visual area VO2. **(b)** same as **a** but for INCEPTIONV3 model. **(c)** same as **a** but for DENSENET121 model.

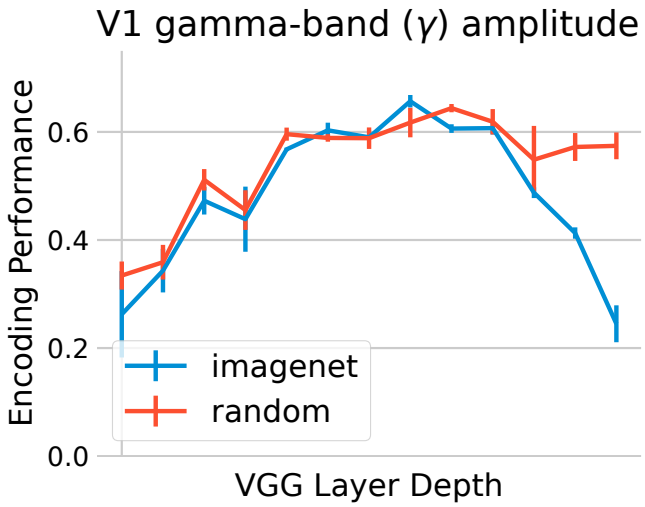

Figure S4. Data from [Uran et al., 2022]. In brief, recordings were made from one awake monkey passively viewing 11 degrees black/white natural images using multi-site recordings. Gamma-band local field potential (LFP) activity, which indexes rhythmic, synchronized activity in the local network, was measured by detrending the power spectrum and measuring the magnitude of the local spectral peak in the LFP. Note that gamma-band responses reflect surround modulations beyond the classic receptive field, which may explain the enhanced reliance on deeper layer representations for predicting V1 gamma-band activity [Uran et al., 2022, Peter et al., 2019, Vinck and Bosman, 2016, Gieselmann and Thiele, 2008].

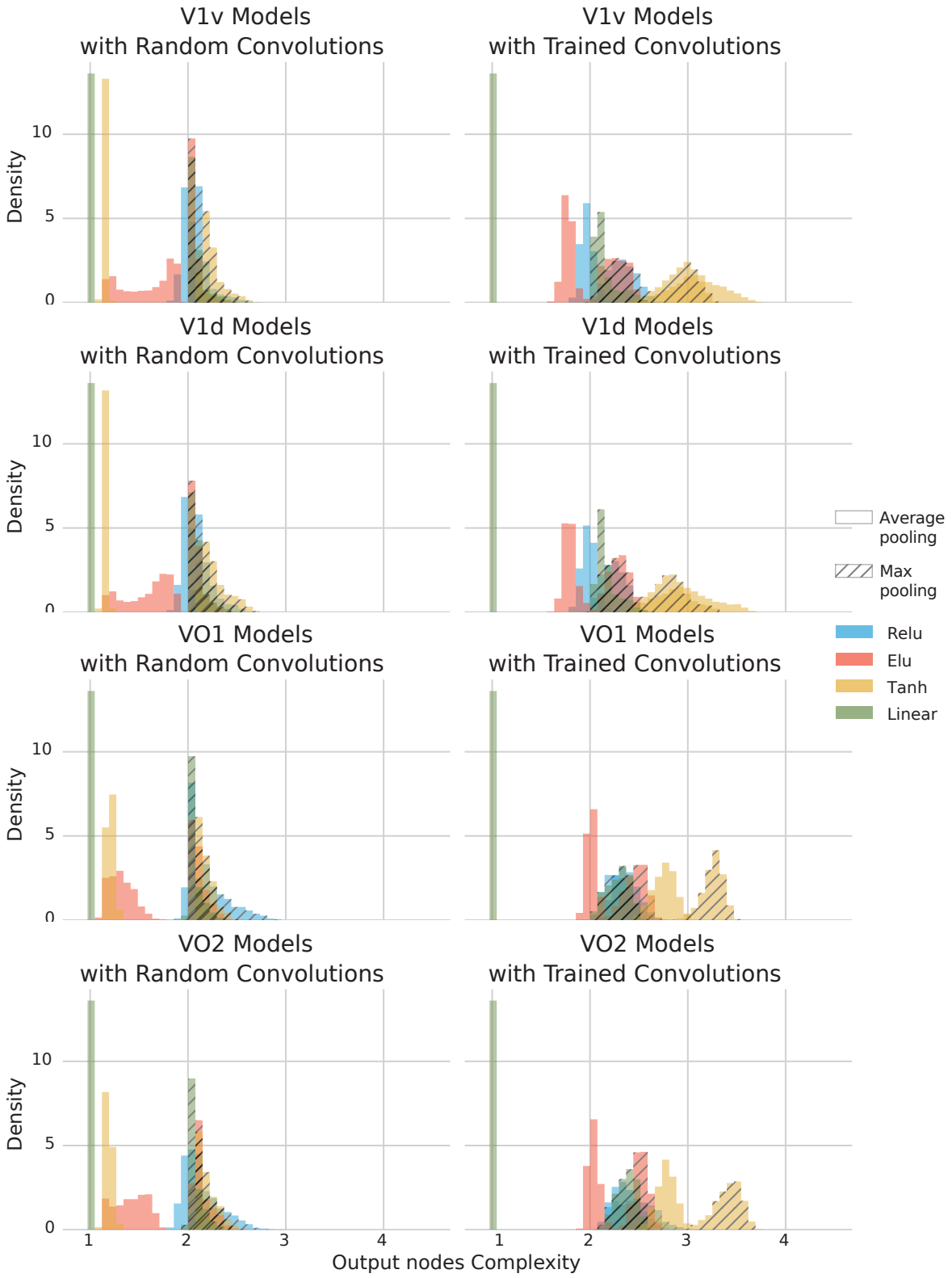

Figure S5. Distributions of encoding models' output nodes complexity. Each distribution represent the complexity of output nodes (see methods) of a certain model configuration (with respect to the pooling strategy and activation function) trained on human V1v, V1d (upper two rows), VO1 and VO2 (lower two rows) data. Only linear readout was trained on top of random convolutional features (left column) or the model was fully trained (right column). Each distribution is the average of 3 training iterations.

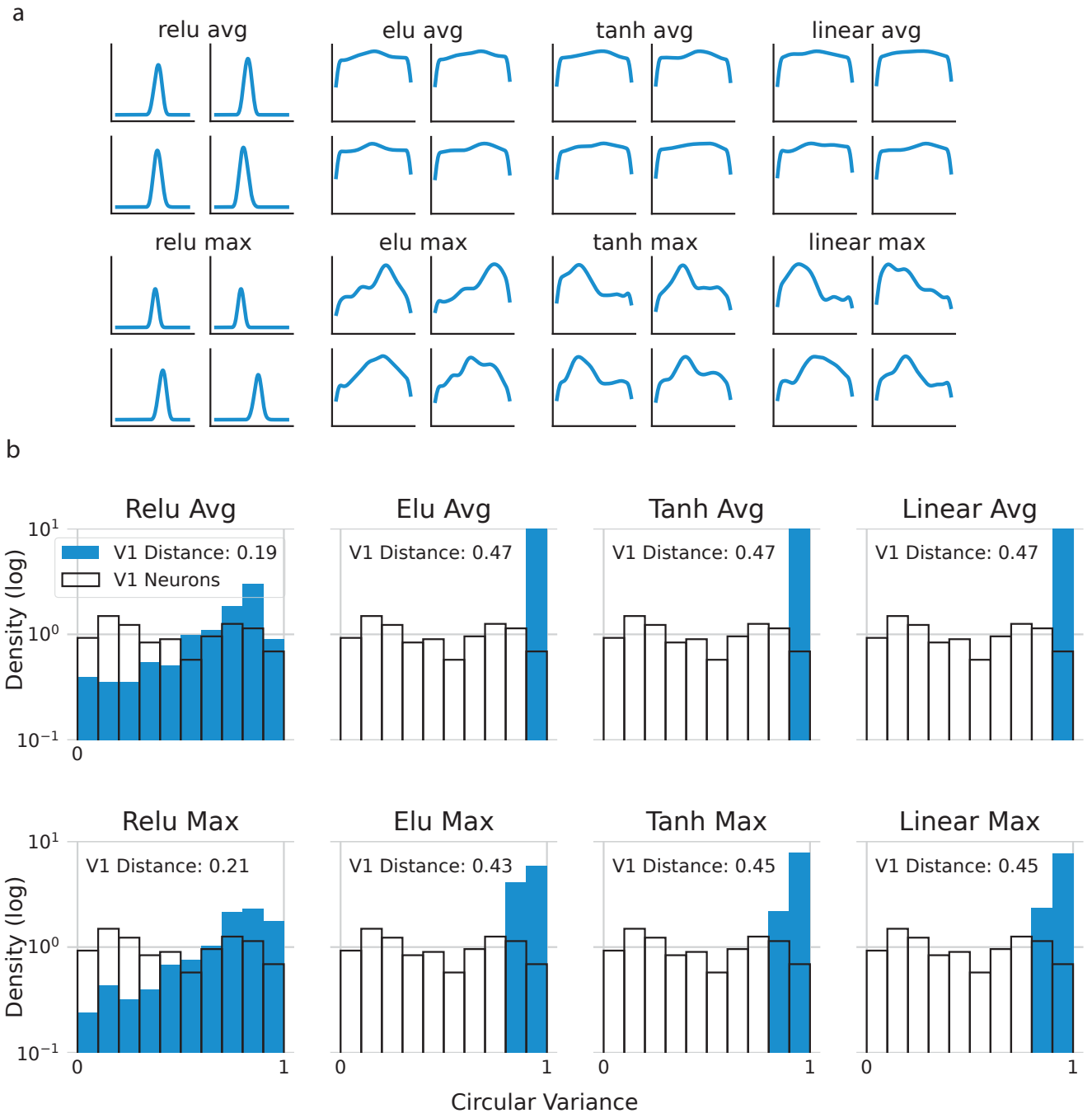

Figure S6. Orientation selectivity of random models. **(a)** Tuning curves of the most orientation-selective artificial neuron in the last convolutional layer of each model configuration upon random initialization. **(b)** Distributions of the circular variance of the artificial neurons of the last convolutional layer of each model configuration upon random initialization.

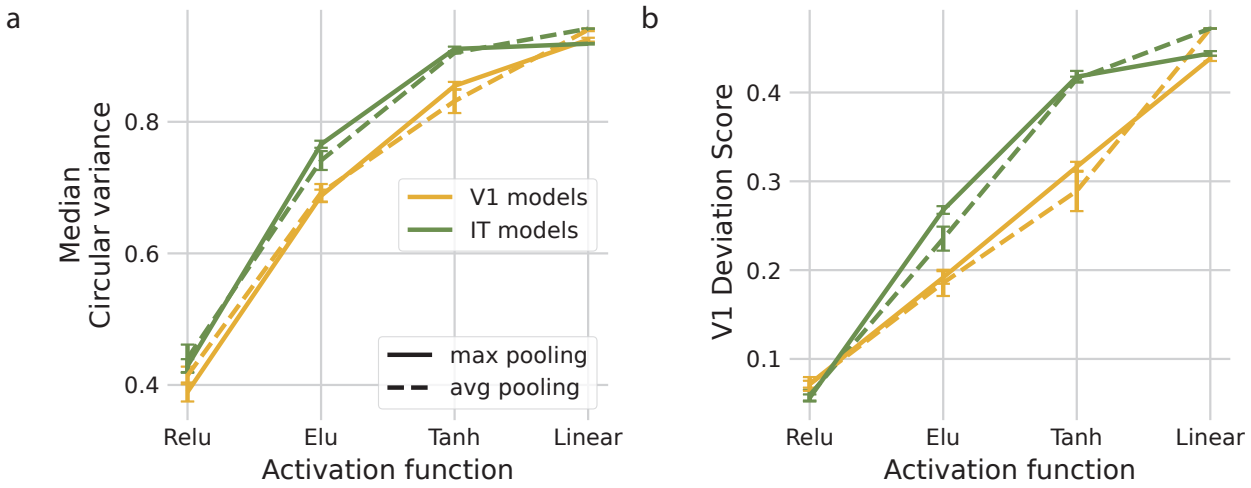

Figure S7. **(a)** Median circular variance of the artificial neurons of the last convolutional layer of V1 trained and IT trained models of different configurations averaged across 5 iterations. Error bars are the standard deviation. **(b)** V1 deviation scores of V1 trained and IT trained models averaged across 5 iterations. Error bars are the standard deviation.