Open access • Posted Content • DOI:10.1101/2020.10.05.326256

# Generalization in data-driven models of primary visual cortex — Source link

Konstantin-Klemens Lurz, Mohammad Ali Bashiri, Konstantin F. Willeke, Akshay Kumar Jagadish ...+8 more authors

**Institutions:** University of Tübingen, Baylor College of Medicine, University of Göttingen

**Published on:** 07 Oct 2020 - bioRxiv (Cold Spring Harbor Laboratory)

**Topics:** Visual cortex, Retinotopy, Models of neural computation and Receptive field

# GENERALIZATION IN DATA-DRIVEN MODELS OF PRIMARY VISUAL CORTEX

**Konstantin-Klemens Lurz,**[1-2,*] **Mohammad Bashiri,**[1-2] **Konstantin Willeke,**[1-2]
**Akshay K. Jagadish,**[1] **Eric Wang,**[4-5] **Edgar Y. Walker,**[1,2,4-5]
**Santiago A. Cadena,**[2,3] **Taliah Muhammad,**[4-5] **Erick Cobos,**[4-5]
**Andreas S. Tolias,**[4-5] **Alexander S. Ecker,**[6] **Fabian H. Sinz**[1-5, **]

[1] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Germany
[2] International Max Planck Research School for Intelligent Systems, Tübingen, Germany
[3] Bernstein Center for Computational Neuroscience, University of Tübingen, Germany
[4] Department for Neuroscience, Baylor College of Medicine, Houston, TX, USA
[5] Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, USA
[6] Department of Computer Science / Campus Institute Data Science, University of Göttingen, Germany

[*]konstantin.lurz@uni-tuebingen.de, [**]fabian.sinz@uni-tuebingen.de

## ABSTRACT

Deep neural networks (DNN) have set new standards at predicting responses of neural populations to visual input. Most such DNNs consist of a convolutional network (core) shared across all neurons which learns a representation of neural computation in visual cortex and a neuron-specific readout that linearly combines the relevant features in this representation. The goal of this paper is to test whether such a representation is indeed generally characteristic for visual cortex, i.e. generalizes between animals of a species, and what factors contribute to obtaining such a generalizing core. To push all non-linear computations into the core where the generalizing cortical features should be learned, we devise a novel readout that reduces the number of parameters per neuron in the readout by up to two orders of magnitude compared to the previous state-of-the-art. It does so by taking advantage of retinotopy and learns a Gaussian distribution over the neuron's receptive field position. With this new readout we train our network on neural responses from mouse primary visual cortex (V1) and obtain a gain in performance of 7% compared to the previous state-of-the-art network. We then investigate whether the convolutional core indeed captures *general* cortical features by using the core in transfer learning to a different animal. When transferring a core trained on thousands of neurons from various animals and scans we exceed the performance of training directly on that animal by 12%, and outperform a commonly used VGG16 core pre-trained on imagenet by 33%. In addition, transfer learning with our data-driven core is more data-efficient than direct training, achieving the same performance with only 40% of the data. Our model with its novel readout thus sets a new state-of-the-art for neural response prediction in mouse visual cortex from natural images, generalizes between animals, and captures better characteristic cortical features than current task-driven pre-training approaches such as VGG16.

## 1 INTRODUCTION

A long lasting challenge in sensory neuroscience is to understand the computations of neurons in the visual system stimulated by natural images (Carandini et al., 2005). Important milestones towards this goal are general system identification models that can predict the response of large populations of neurons to arbitrary visual inputs. In recent years, deep neural networks have set new standards in predicting responses in the visual system (Yamins et al., 2014; Vintch et al., 2015; Antolík et al., 2016; Cadena et al., 2019a; Batty et al., 2016; Kindel et al., 2017; Klindt et al., 2017; Zhang et al., 2018; Ecker et al., 2018; Sinz et al., 2018) and the ability to yield novel response characterizations (Walker et al., 2019; Bashivan et al., 2019; Ponce et al., 2019; Kindel et al., 2019; Ukita et al., 2019).

Such a general system identification model is one way for neuroscientists to investigate the computations of the respective brain areas *in silico*. Such *in silico* experiments exhibit the possibility to study the system at a scale and level of detail that is impossible in real experiments which have to cope with limited experimental time and adaptation effects in neurons. Moreover, all parameters, connections and weights in an *in silico* model can be accessed directly, opening up the opportunity to manipulate the model or determine its detailed tuning properties using numerical optimization methods. In order for the results of such analyses performed on an *in silico* model to be reliable, however, one needs to make sure that the model does indeed replicate the responses of its biological counterpart faithfully. This work provides an important step towards obtaining such a generalizing model of mouse V1.

High performing predictive models need to account for the increasingly nonlinear response properties of neurons along the visual hierarchy. As many of the nonlinearities are currently unknown, one of the key challenges in neural system identification is to find a good set of characteristic nonlinear basis functions—so called *representations*. However, learning these complex nonlinearities from single neuron responses is difficult given limited experimental data. Two approaches have proven to be promising in the past: *Task-driven* system identification networks rely on transfer learning and use nonlinear representations pre-trained on large datasets for standard vision tasks, such as object recognition (Yamins & DiCarlo, 2016). Single neuron responses are predicted from a particular layer of a pre-trained network using a simple readout mechanism, usually an affine function followed by a static nonlinearity. *Data-driven* models share a common nonlinear representation among hundreds or thousands of neurons, and train the entire network end-to-end on stimulus response pairs from the experiment. Because the nonlinear representation is shared, it is trained via massive multi-task learning (one neuron–one task) and can be learned even from limited experimental data.

*Task-driven* networks are appealing because they only need to fit the readout mechanisms on top of a given representation and thus are data-efficient in terms of the number of stimulus-response pairs needed to achieve good predictive performance (Cadena et al., 2019a). Moreover, as their representations are obtained independently of the neural data, a good predictive performance suggests that the nonlinear features are characteristic for a particular brain area. This additionally offers the interesting normative perspective that the functional representations in deep networks and biological vision could be aligned by common computational goals (Yamins & DiCarlo, 2016; Kell et al., 2018; Kubilius et al., 2018; Nayebi et al., 2018; Sinz et al., 2019; Güçlü & van Gerven, 2014; Kriegeskorte, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann et al., 2019). In order to quantify the fit of the normative hypothesis, it is important to compare a given representation to other alternatives (Schrimpf et al., 2018; Cadena et al., 2019a). However, while representations pre-trained on ImageNet are the state-of-the-art for predicting visual cortex in primates (Cadena et al., 2019a; Yamins & DiCarlo, 2016), recent work has demonstrated that pre-training on object categorization (VGG16) yields no benefits over random initialization for mouse visual cortex (Cadena et al., 2019b). Since random representation should not be characteristic for a particular brain area and other tasks that might yield more meaningful representations have not been found yet, this raises the questions whether there are better ways to obtain a generalizing nonlinear representation for mouse visual cortex.

Here, we investigate whether such a generalizing representation can instead be obtained from *data-driven* networks. For this purpose, we develop a new data efficient readout which is designed to push non-linear computations into the core and test whether this core has learned general characteristic features of mouse visual cortex by applying the same criteria as for the task-driven approach: The ability to predict a population of unseen neurons in a new animal (transfer learning). Specifically, we make the following contributions: ❶ We introduce a novel readout mechanism that keeps the number of per-neuron parameters at a minimum and learns a bivariate Gaussian distribution for the readout position from anatomical data using retinotopy. With this readout alone, we surpass the previous state-of-the-art performance in direct training by 7%. ❷ We demonstrate that a representation pre-trained on thousands of neurons from various animals generalizes to neurons from an *unseen animal* (transfer learning). It exceeds the direct training condition by another 11%, setting the new state-of-the-art and outperforms a *task-driven* representation—trained on object recognition—by about 33%. ❸ We then show that this generalization can be attributed to the *representation* and not the readout mechanism, indicating that the data-driven core indeed captures generalizing features of cortex: A representation trained on a single experiment (4.5k examples) in combination with a readout trained on anatomically matched neurons from four experiments (17.5k examples) did not achieve this performance. ❹ Lastly, we find that transfer learning with our data-driven core is more data-efficient than direct training, achieving the same performance with only 40% of the data.

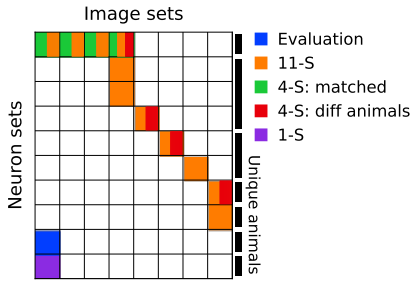



Figure 1: **Scans and training sets.** Overview of the datasets and how they are combined into different training sets. Each scan was performed on a specific set of neurons (rows) using a specific set of unique images (columns). Repeatedly presented test images were the same for all scans. Some scans were performed on the same neuron but with different image sets (first row). Colors indicate grouping of scans into training sets and match line colors in Fig. 5 to indicate which dataset a representation/core (not the readout) was trained on.

# 2 METHODS

## 2.1 DATA

**Functional data** The data used in our experiments consists of pairs of neural population responses and grayscale visual stimuli sampled and cropped from ImageNet, isotropically downsampled to $64 \times 36$ px, with a resolution of $0.53$ ppd (pixels per degree of visual angle). The neural responses were recorded from layer L2/3 of the primary visual cortex (area V1) of the mouse, using a wide field two photon microscope (Sofroniew et al., 2016). Activity was measured using the genetically encoded calcium indicator GCaMP6s. V1 was targeted based on anatomical location as verified by numerous previous experiments performing retinotopic mapping using intrinsic imaging. We selected cells based on a classifier for somata on the segmented cell masks and deconvolved their fluorescence traces (Pnevmatikakis et al., 2016). We did not filter cells according to visual responsiveness. The stimulation paradigm and data pre-processing followed the procedures described by Walker et al. (2019). A single scan contained the responses of approximately 5000–9000 neurons to up to 6000 images, of which 1000 images consist of 100 unique images which were presented 10 times each to allow for an estimate of the reliability of the neuron (see Appendix for a detailed description of the datasets). We used the repeated images for testing, and split the rest into 4500 training and 500 validation images. The neural data was preprocessed by normalizing the responses of the neurons by their standard deviation on the training set. To put the number of recorded neurons per scan into perspective, assuming that V1 has an area of about 4mm$^2$, that L2/3 is about 150-250$\mu$m thick and has a cell density of 80k excitatory cells per mm$^3$, entire V1 L2/3 should contain about 48k - 80k neurons (Garrett et al., 2014; Jurjut et al., 2017; Schüz & Palm, 1989), similar to the maximum number of neurons that we train a model on (72k neuron, `11-S` , Fig. 1, orange). Note, however, that this does not mean that these 72k neurons sample V1 or the visual field of a mouse evenly because of possible experimental biases in the choice of the recording location.

All together, we used 13 scans from a total of 7 animals (Fig. 1). Each scan is defined by the set of neurons it was performed on (rows/*neuron sets* in Fig. 1) and the set of images that were shown (columns/*image sets* in Fig. 1). Different image sets had non-overlapping training/validation images, but the same test images. Some of the scans were performed on the same neurons, but with different sets of natural images (first row in Fig. 1). These neurons were matched across scans by cross-correlating the structural scan planes against functionally recorded stacks (Walker et al., 2019). Stitching data from several scans in this way allowed us to increase the number of image presentations per neuron beyond what would be possible in a single scan. We combined these scans into different training sets (one color–one training set in Fig. 1) and named each one of them—*e.g.* `11-S` for a set with 11 **S**cans. The different sets are further explained in the respective experiments they are used in. All data from the seven mice used in this work has been recorded by trained personnel under a strict protocol according to the regulations of the local authorities at Balor College of Medicine.

## 2.2 NETWORKS AND TRAINING

The networks are split conceptually into two parts: a *core* and a *readout*. The core captures the nonlinear image representation and is shared among all neurons. The readout maps the features of the core into neural responses and contains all neuron specific parameters.

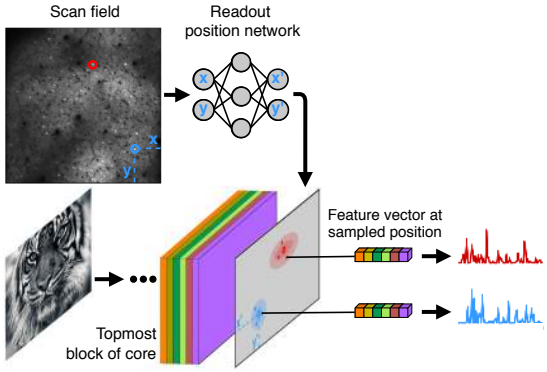



Figure 2: **Using retinotopy to learn the readout position from anatomical data.** The Gaussian readout for each neuron uses features from a single location on the final tensor of the core CNN (bottom). The position is drawn from a 2D Gaussian for every image during training. The parameters of the Gaussian for each neuron are learned during training. The means of the Gaussians are predicted from each neuron's coordinates on cortex by a *Readout Position Network* whose weights are shared across neurons and learned during training (top). During testing, the mean of the Gaussian is used as the neuron's position.

**Representation/Core** We model the core with a four-layer convolutional neural network (CNN), with 64 feature channels per layer. In each layer, the 2d-convolutional layer is followed by a batch normalization layer and an ELU nonlinearity (Ioffe & Szegedy, 2015; Clevert et al., 2015). All convolutional layers after the first one are depth-separable convolutions (Chollet, 2017) which we found to yield better results than standard convolutional layers in a search among different architecture choices.

**Readouts** We compared two different types of readouts to map the nonlinear features of the core to the response of each neuron. For each neuron, a tensor of $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ (**w**idth, **h**eight, **c**hannels) needs to be mapped to a single scalar, corresponding to the target neuron's response. All of our readouts assume that this function is affine with a linear weight tensor $\mathbf{w} \in \mathbb{R}^{w \times h \times c}$ , followed by an ELU offset by one (ELU+1), to keep the response positive. Furthermore, both readouts assume that in feature space the receptive field of each neuron does not change its position across features, but they differ in how this receptive field location is constrained and learned.

The *factorized readout* (Klindt et al., 2017) factorizes the 3d readout tensor into a lower-dimensional representation by using a spatial mask matrix $u_{ij}$ and a vector of feature weights $v_k$, i.e. $w_{ijk} = u_{ij} v_k$. The spatial mask $u_{ij}$ is restricted to be positive and encouraged to be sparse through an L1 regularizer.

Our novel *Gaussian readout* reduces the number of per-neuron parameters. It computes a linear combination of the feature activations at a single spatial position— parametrized as $(x, y)$ coordinates —via bilinear interpolation (Sinz et al., 2018). To facilitate gradient flow during training, we replace the spatial downsampling used in (Sinz et al., 2018) by a sampling step, which during training draws the readout position of each $n^{\text{th}}$ neuron from a bivariate Gaussian distribution $\mathcal{N}(\mu_n, \Sigma_n)$ for each image in a batch separately. This is the sampling version of (St-Yves & Naselaris, 2017) where the readout location is weighted spatially with a Gaussian profile. In our case, $\mu_n$ and $\Sigma_n$ are learned via the reparametrization trick (Kingma & Welling, 2014). Initializing $\Sigma_n$ large enough ensures that there is gradient information available to learn $\mu_n$ reliably. During training, $\Sigma_n$ shrinks as the estimate of the neuron position improves. During evaluation we always use the position defined by $\mu_n$, making the readout deterministic. This version of the Gaussian readout has $c + 7$ parameters per neuron (2 for $\mu$, 4 for $\Sigma$ because the linear mapping in the reparametrization trick is $2 \times 2$, and 1 for the scalar bias).

The second innovation of our Gaussian readout is to couple the location estimation of single neurons by exploiting the retinotopic organization of primary visual cortex (V1) and other areas. Since V1 preserves the topology of visual space, we estimate a neuron's receptive field location from its position $\mathbf{p}_n \in \mathbb{R}^2$ along the cortical surface available from the experiments. To that end, we learn a common function $\mu_n = f(\mathbf{p}_n)$ represented by a neural network that is shared across all neurons (Fig. 2). Since we work with neurons from local patches of V1, we model $f$ as a linear fully connected network. This approach turns the problem of estimating each neuron's receptive field location from limited data into estimating a single linear transformation shared by all neurons, and reduces the number of per-neuron parameters to $c + 5$. We initialized the *Readout Position Network* to a random orthonormal 2-2 matrix scaled by a factor which was optimized in hyper-parameter selection.

Finally, when training on several scans of anatomically matched neurons from the same mouse (see Data), we share the feature weights $v_k$ across scans. To account for differences in spike inference between scans, we introduced a scan-specific scale and bias for each neuron after the linear readout. We mention in the respective sections whether features are shared or not. The bias of each readout is initialized with the average response on the training set. The effects of both feature sharing and learning from cortical anatomy on the performance of the readout are shown in the Appendix.

**Training** The networks were trained to minimize Poisson loss $\frac{1}{m}\sum_{i=1}^{m}\left(\hat{r}^{(i)} - r^{(i)}\log\hat{r}^{(i)}\right)$ where $m$ denotes the number of neurons, $\hat{r}$ the predicted neuronal response and $r$ the observed response. We used early stopping on the correlation between predicted and measured neuronal responses on the validation set (Prechelt, 1998): if the correlation failed to increase during any 5 consecutive passes through the entire training set (epochs), we stopped the training and restored the model to the best performing model over the course of training. We found that this combination of Poisson objective and early stopping on correlation yielded the best results. After the first stop, we decreased the learning rate from $5 \times 10^{-3}$ twice by a decay factor of $0.3$, and resumed training until it was stopped again. Network parameters were iteratively optimized via stochastic gradient descent using the Adam optimizer (Kingma & Ba, 2015) with a batch size of 64. Once training completed, the trained network was evaluated on the validation set to yield the score used for hyper-parameter selection. The hyper-parameters were then selected with a Bayesian search (Snoek et al., 2012) of 100 trials and subsequently kept fixed throughout all experiments. Only the scale of the readout regularization was fine-tuned with additional Bayesian searches for the cases of different amounts of data independently. In transfer experiments, we froze all parameters of the core and trained a new readout only.

**Evaluation** We report performance as *fraction oracle* (see Walker et al., 2019), which is defined as the correlation of the predicted response and the observed single-trial test responses relative to the maximally achievable correlation measured from repeated presentations. We estimated the oracle correlation using a jackknife estimator (correlation of leave-one-out mean against single trial). Per data point, we trained 25 networks for all combinations of five different model initializations and five random partitions of the neurons into core and transfer sets. The image subsets were drawn randomly once and kept fixed across all experiments except in Fig. 5 where the full neuron set was used and 5 random partitions of image subsets were drawn instead. We selected the best performing models across initializations and calculated 95% confidence intervals over neuron- or image seeds.

## 3 Results

We investigated the conditions under which a *data-driven* core generalizes to new neurons in the same or different animals. We did this by pre-training a core on differently composed datasets (core sets, see 3.1) and testing that core in transfer learning to a new set of neurons (transfer set, see 3.1). Our main finding is that transferring a core trained on multiple scans (up to 35k unique images and 70k unique neurons) to a new set of 5335 unique neurons from a single scan (4.5k unique images) yields an improvement in performance of about 12% compared to a network directly trained on the single scan. This result was independent of whether the transferred core was trained on the same animal or not. By carefully choosing the Transfer learning conditions, we can attribute this boost in performance to the generalization of the core.

### 3.1 Transfer learning conditions

To test the generalization performance of the core, we used a separate set of 1000 neurons which we call *transfer set*. These neurons were not used to train the transferred core but only to fine-tune a new readout (note that Fig. 4 and 5 use different transfer sets, for intra- and inter- animal transfer respectively). A transfer set is a subset of neurons, not images. Thus, each transfer set also had a train, validation, and test split of its images. We compared the performance in transfer learning to that of a network directly trained end-to-end on the transfer set or subsets thereof (*direct* condition). Because the core of the *direct* condition was not transferred, it could adapt to the neurons at hand giving it a fair chance to outperform the transferred cores.

In the transfer learning conditions (all except *direct*), the core was always trained on a separate set of neurons (*core set*) and subsequently frozen, while the readout was always trained on top of

the frozen core using the transfer set. Again, a core set is a subset of the neurons, not images. In order to quantify the generalization of the core, we needed to decouple the amount of data it takes to train the core from the amount of data it takes to train the readout. We thus considered two transfer conditions: *diff-core/best-readout* and *best-core/diff-readout*. In the *best-core/diff-readout* condition, the core was trained on the core set using all images, while the readout was trained on the transfer set using different numbers of images. This condition tests the data efficiency of the readout. We expect that better cores lead to a higher data efficiency in the readout, *i.e.* require less data to achieve good performance. In the *diff-core/best-readout* condition, we trained the core on the core set using different numbers of images while the readout was trained on the transfer set using all images. Thereby we tested how the generalization of the core is affected by the amount of data used to train it.

### 3.2 Direct training and within animal/across neuron generalization

The following transfer experiments (Fig. 3 and 4) aim at investigating how the number of images and neurons provided to the core and the readout affects generalization to new neurons. To this end, we used a dataset that includes as many images as possible while still providing a reasonably large number of neurons. The `4-S:matched` dataset (first row in Fig. 1, green) provides 17596 images and 4597 neurons anatomically matched across 4 scans. Each scan was performed on the same neurons, but showing different sets of images (test set was identical). The dataset was concatenated along the image dimension and split into core and transfer sets of 3597 and 1000 neurons.

**Direct training** We used the `4-S:matched` core set to compare the performance of the Gaussian and factorized readout in the *direct* condition on the original core set before transfer. We tested both readouts with and without feature sharing (see Networks and training). The performance of the networks increased with the number of images (Fig. 3). It also increased with the number of neurons, but the number of images had a far stronger effect. While the performance saturated quickly with the number of neurons, saturation w.r.t. the number of images did not seem to be reached, even when using all 17.5k images. The Gaussian readout outperformed the factorized readout in predictive performance by 7% fraction oracle for the full `4-S:matched` dataset, reaching $0.886 \pm 0.005$ and $0.826 \pm 0.005$ fraction oracle respectively (mean$\pm$std). While the Gaussian readout profits from feature sharing, the factorized readout is hurt by it (Fig. 3, light vs. dark colors). This might be because the spatial masks in the factorized readout are less constrained in contrast to the Gaussian readout where the position network and the usage of only a single readout point exerts a stronger inductive bias. In all future experiments, we thus use feature sharing only for the Gaussian readout.

**Within animal/across neuron generalization** For both readouts, the generalization performance of the learned core, tested in the *diff-core/best-readout* condition, increased with the number of images used to train the core (Fig. 4, pink). The cores and readouts were trained on the core and transfer set of the `4-S:matched` dataset. As before, the Gaussian readout outperformed the factorized readout, exhibiting a stronger increase in performance with the number of images and a better final performance when the entire dataset was used to train the core. Even for a core trained on few data, a readout can yield good performance if it has access to enough images (pink line). Importantly,

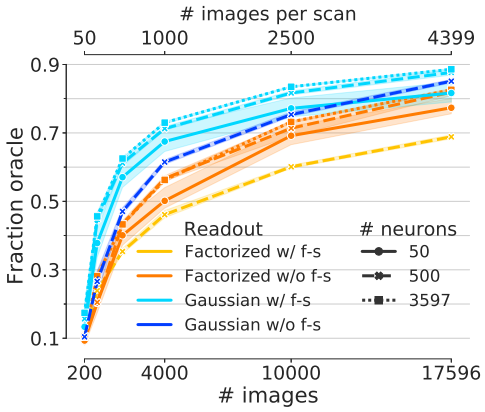



Figure 3: **Performance of end-to-end trained networks.** Performance for different subsets of neurons (linestyle) and number of training examples (x-axis). The same core architecture was trained for two different readouts with and without feature sharing (color) on the matched neurons of the `4-S:matched` core set (Fig. 1, green). Both networks show increasing performance with number of images. However, the network with the Gaussian readout achieves a higher final performance (light blue vs. orange). While the Gaussian readout profits from feature sharing (light vs. dark blue), the factorized readout is hurt by it (yellow vs. orange). Shaded areas depict 95% confidence intervals across random picks of the neuron subsets.

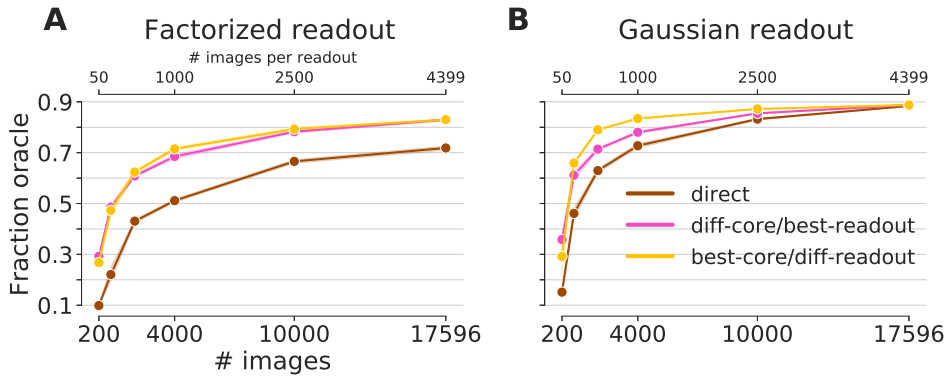



Figure 4: **Generalization to other neurons in the same animal.** A core trained on 3597 neurons and up to 17.5k images generalizes to new neurons (pink and yellow line). A core trained on the full data yields very good predictive performance even when the readout is trained on far less data (yellow). If the readout is trained with all data, even a core trained on few data can yield a good performance (pink). Both transfer conditions outperform a network directly trained end-to-end on the transfer dataset (brown). For the full dataset, all training conditions converge to the same performance. Except in the *best-core/diff-readout* condition for very few training data, the Gaussian readout (**B**) outperforms the factorized readout (**A**). The data for both the training and transfer comes from the `4-S:matched` dataset (Fig 1, green). Not that the different number of images can be from the core or transfer set, depending on the transfer condition.

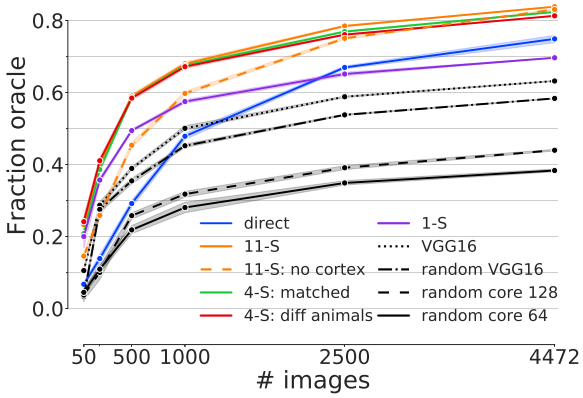



Figure 5: **Generalization across animals.** Prediction performance in fraction oracle correlation as a function of training examples in the transfer set for a Gaussian readout (x-axis) and different ways to obtain the core (colors). The transfer training was performed on the evaluation dataset (blue, Fig 1). Cores trained on several scans used in transfer learning outperform direct training on the transfer dataset (blue line; direct condition).

except for very low numbers of images, the fastest increase in performance occurred in the *best-core/diff-readout* condition, where a core trained on all images (but different neurons) was frozen and a new readout was trained on the transfer set for varying numbers of images (yellow lines). This result shows that a data-driven core provides general characteristic features for mouse V1, and these features generalize to new neurons. Importantly, for 4k images (about the size of a typical experiment), the performance of the *best-core/diff-readout* was approximately 7% better than the performance of the *diff-core/best-readout* condition ($0.834 \pm 0.003$ and $0.780 \pm 0.007$ fraction oracle respectively). This observation that a readout on all 17.5k images on a core from 4k images could not reach the performance of a readout on 4k images on a core from 17.5k images suggests that the better performance is due to the generalizing core and not the readout.

### 3.3 Generalization across animals

So far, we tested generalization performance of data-driven cores to different neurons in the same animal using the `4-S:matched` dataset. A stronger test for generalization is transfer learning across animals, where we may have to deal with inter-subject variability. To this end, we compared several cores derived from different core sets, random initialization, or ImageNet pre-training in terms of their generalization to neurons from a mouse which was not used to train the core. The transfer set consisted of 5335 neurons from a different mouse presented with images that were also in the

core set (Fig. 1, blue). Note that performance was still evaluated on a set of test images that were neither used to train core nor readout. Apart from the previously used core set with anatomically matched neurons (`4-S:matched` , Fig. 1, green), we also trained a core on a single scan from another animal (`1-S` , Fig. 1, purple), and on a set of four scans from different animals with different images (`4-S: diff animals` , Fig. 1, red). Finally we also trained a core on all datasets together, without any information on neuron matching or image sets (`11-S` , Fig. 1, orange). For completeness, we also compared to a task-driven VGG core (Simonyan & Zisserman, 2015), a randomly initialized VGG core (reading out from conv2-2 for both) (Cadena et al., 2019b), our core (64 features) with randomly initialized weights, and a scaled up version of our core with the same number of features as the VGG (128). As before, we compared the generalization performance to a model trained under the *direct* condition trained on the transfer set. Due to its better performance, we used the Gaussian readout for these experiments but a version with the factorized readout can be found in the Appendix.

We tested the generalization in the *best-core/diff-readout* condition on a single transfer set (Fig. 5), *i.e.* all performances are reported on the same transfer set (5335 neurons) (blue, Fig. 1 and 5), but differ in where the core was trained (colors, Fig. 5). The most striking finding is that representations trained on several scans not only reached better performances for fewer training images, but they also reached a better overall performance for the available data from a single scan compared to *direct* training (Fig. 5 orange, green, and red vs. blue). Interestingly, this improvement did not require matched neurons in the core set (Fig. 5) and any potential negative effects of inter-subject variability were outweighed by the benefits of using multiple scans to train the core. The two cores trained on a single scan reached about the same final performance (Fig. 5, blue vs. purple), with a slightly better performance for the directly trained model, as expected. However, the final performance of both single scan models was about 10% smaller than the transfer performance from the models pre-trained on four and eleven scans, respectively (Fig. 5, blue and purple vs. red, green, and orange). We provide visualizations of the receptive fields – produced via response maximization (Walker et al., 2019; Bashivan et al., 2019) – of some example neurons obtained from our best model (Fig. 5, orange at 4472 images) in the Appendix, Fig. 4.

Consistent with previous work (Cadena et al., 2019b), pre-trained and random VGG16 cores performed similarly (Fig. 5 gray, dotted vs. dash-dotted). Both VGG cores performed worse than a directly trained *data-driven* core (Fig. 5 blue vs. gray dotted and dash-dotted). Our core with random weights (64 features) performs worst, demonstrating that training on neural data extracts characteristic features. Scaling up this random core to VGG size (128) does not match its performance which could be due to the interaction of the initialization with architectural differences.

Lastly, we investigated the effect of constraining all neurons to share the transformation from cortical location to receptive field location by temporarily deactivating this feature (Fig. 5, `11-S: no cortex` dashed orange), and found that this constraint was particularly useful for small numbers of images. An equivalent figure for Fig. 5 for the factorized readout, a table with a detailed overview over the most important results in numerical form, as well as a comparison with other performance metrics can be found in the Appendix.

## 4 Related work

The idea of a common cortical feature representation is wide-spread in sensory and systems neuroscience, going back to the idea of V1 as a bank of Gabor filters or edge detectors (Jones et al., 1987; Olshausen & Field, 1996). A substantial body of recent work focuses on feature representations learned by training deep networks on vision tasks such as object recognition (Cadena et al., 2019a; Yamins & DiCarlo, 2016; Güçlü & van Gerven, 2014; Kriegeskorte, 2015; Khaligh-Razavi & Kriegeskorte, 2014). The brain-score[1] initiative compares different representations, resulting from pre-training on different tasks or different network architectures, with regards to performance of multiple neural prediction tasks (Schrimpf et al., 2018). In contrast to that, we focused on whether multi-task learning between thousands of neurons leads to a generalizing representation.

While *task-driven* representations perform comparably to *data-driven* representations in primates (Cadena et al., 2019a), Cadena et al. (2019b) recently demonstrated that they show almost no difference in predictive performance for mice. Our results corroborate this finding (Fig. 5) and show that a

[1] `https://www.brain-score.org/`

*data-driven* representation outperforms a *task-driven* representation by a substantial margin, even when tested on equal grounds with transfer learning.

Sharing a representation between neurons is commonly used to learn *data-driven* system identification networks (Cadena et al., 2019b;a; Batty et al., 2016; Sinz et al., 2018; Antolík et al., 2016; Klindt et al., 2017). Klindt et al. (2017) investigated the effect of the number of neurons and training examples onto the predictive performance of a data-driven network. However, this experiment was done on simulated data only and did not explore the generalization (transfer) performance of the learned representation. To the best of our knowledge, we are the first to systematically investigate the ability of data-driven representations to capture general characteristic features of visual cortex.

## 5 Discussion

Machine learning applications in biology are often faced with limited amount of data. Especially, for recent deep learning approaches this poses a challenging problem. One promising way to approach it is multi-task learning by training a shared nonlinear representation on multiple tasks or subjects. This increases the data volume and can help to extract inductive biases to achieve better generalization. Here, we investigated a particular instance of this problem: Modeling the responses of thousands of cortical neurons as a function of natural visual stimuli. We demonstrated that nonlinear *data-driven* representations, trained via massive multi-task learning through parameter sharing among thousands of neurons from mouse primary visual cortex, generalize to other neurons and mice, and significantly outperform common *task-driven* alternatives that are predictive for monkey V1 (Cadena et al., 2019b; Yamins & DiCarlo, 2016). As noted by Cadena et al. (2019b) already, this does not imply that all task-trained representations are necessarily suboptimal but rather indicates that we have not found the right task yet. But so far, our network sets a new state-of-the-art for neural response prediction in direct training *as well as* transfer learning.

Our transfer results strongly suggests that data trained cores can capture features that are characteristic of mouse primary visual cortex, and the fact that the Gaussian readout can predict novel neurons from this core with relatively few training example corroborates this idea. In addition, the good transfer learning performance indicates that inter-subject variability, which could affect the success of multi-task learning, does not seem to be a major problem for this application. For that reason, we believe that our *data-driven* core is the most characteristic representation to date to predict mouse primary visual cortex and that it could be a great tool in new experiments where data is scarce and/or training time is limited, such as the *inception loops* introduced by Walker et al. (2019) and Bashivan et al. (2019). To facilitate this, we share the weights of the trained representation together with its code online[2] to allow others to predict neural responses with it. Other possible applications include for example the analysis of the learned feature representations to investigate the operations in visual processing, or using the core to support vision tasks like image categorization (Li et al., 2019). Additionally, we also share the dataset that we evaluate our core on (Fig. 1, blue) so that other representations can be tested and compared with ours on the same data[3].

Our results are based on "deconvolved" calcium traces of neural activity integrated over 500 ms. Whether data driven cores can also provide generalizing features for shorter time scales, recordings with electrodes, to dynamic responses, or of higher areas remains to be seen in future studies.

### Acknowledgments

We thank A.K. Schalkamp, A. Nix, C. Blessing, S. Safarani, E. Froudarakis, and N. Patel for comments and discussions. KKL is funded by the German Federal Ministry of Education and Research through the Tübingen AI Center (FKZ: 01IS18039A). FHS is supported by the Carl-Zeiss-Stiftung and acknowledges the support of the DFG Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645. SAC and ASE are supported by the German Research Foundation (DFG grant EC-479/1-1). This work was supported by an AWS Machine Learning research award to FHS. MB was supported by the International Max Planck Research School for Intelligent Systems. Supported by the Intelligence Advanced Research Projects Activity via Department of Interior/Interior Business Center contract number D16PC00003.

[2] https://github.com/sinzlab/Lurz_2020_code
[3] https://gin.g-node.org/cajal/Lurz2020

## REFERENCES

J. Antolík, S. B. Hofer, J. A. Bednar, and T. D. Mrsic-flogel. Model Constrained by Visual Hierarchy Improves Prediction of Neural Responses to Natural Scenes. *PLoS Comput Biol*, pp. 1–22, 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004927.

P. Bashivan, K. Kar, and J. DiCarlo. Neural Population Control via Deep ANN Image Synthesis. pp. 1–33, 2019. doi: 10.32470/ccn.2018.1222-0.

E. Batty, J. Merel, N. Brackbill, A. Heitman, A. Sher, A. Litke, E. J. Chichilnisky, and L. Paninski. Multilayer network models of primate retinal ganglion cells. Number Nips, 2016.

S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput Biol*, pp. 201764, 2019a. doi: 10.1101/201764. URL https://www.biorxiv.org/content/early/2017/10/11/201764.article-info.

S. A. Cadena, F.H. Sinz, T. Muhammad, E. Froudarakis, E. Cobos, E. Y. Walker, J. Reimer, M. Bethge, and A. S. Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? In *NeurIPS 2019 Workshop Neuro AI*, 2019b.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do We Know What the Early Visual System Does? *J. Neurosci.*, 25(46):10577–10597, 11 2005. doi: 10.1523/JNEUROSCI.3726-05.2005.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.195.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). pp. 1–14, 2015. ISSN 09226389. doi: 10.3233/978-1-61499-672-9-1760. URL http://arxiv.org/abs/1511.07289.

A. S. Ecker, F. H. Sinz, E. Froudarakis, P. G. Fahey, S. A. Cadena, E. Y. Walker, E. Cobos, J. Reimer, A. S. Tolias, and M. Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex. pp. 5–7, 2018. doi: arXiv:1809.10504v1. URL http://arxiv.org/abs/1809.10504.

Marina E. Garrett, Ian Nauhaus, James H. Marshel, Edward M. Callaway, Marina E. Garrett, James H. Marshel, Ian Nauhaus, and Marina E. Garrett. Topography and areal organization of mouse visual cortex. *Journal of Neuroscience*, 34(37):12587–12600, sep 2014. ISSN 15292401. doi: 10.1523/JNEUROSCI.1124-14.2014. URL https://www.jneurosci.org/content/34/37/12587https://www.jneurosci.org/content/34/37/12587.abstract.

Umut Güçlü and Marcel A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway. Technical Report 27, 2014. URL http://arxiv.org/abs/1411.6422%0Ahttp://dx.doi.org/10.1523/JNEUROSCI.5023-14.2015http://arxiv.org/abs/1411.6422v1.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Technical report, 2015.

J. P. Jones, Palmer L A, and L. A. Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 1987. ISSN 0022-3077.

Ovidiu Jurjut, Petya Georgieva, Laura Busse, and Steffen Katzner. Learning enhances sensory processing in mouse V1 before improving behavior. *Journal of Neuroscience*, 37(27):6460–6474, jul 2017. ISSN 15292401. doi: 10.1523/JNEUROSCI.3485-16.2017. URL https://sites.google.com/a/nyu.

Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 2018. ISSN 10974199. doi: 10.1016/j.neuron.2018.03.044.

S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep Supervised , but Not Unsupervised , Models May Explain IT Cortical Representation. *PLoS computational biology*, 10(11), 2014. doi: 10.1371/journal.pcbi.1003915.

Tim C. Kietzmann, Courtney J. Spoerer, Lynn K.A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, 116(43):21854–21863, 10 2019. ISSN 10916490. doi: 10.1073/pnas.1905544116.

W. F. Kindel, Elijah D. Christensen, and Joel Zylberberg. Using deep learning to reveal the neural code for images in primary visual cortex. Technical report, 2017. URL http://arxiv.org/abs/1706.06208.

William F Kindel, Elijah D Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision*, 19(4):29, 10 2019. doi: 10.1167/19.4.29. URL https://app.dimensions.ai/details/publication/pub.1113752643https://jov.arvojournals.org/arvo/content_public/journal/jov/937940/i1534-7362-19-4-29.pdf.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

D. A. Klindt, A. S. Ecker, T. Euler, and M. Bethge. Neural system identification for large populations separating "what" and "where". In *Advances in Neural Information Processing Systems*, number Nips, pp. 4–6, 2017. ISBN 978-1-5108-6096-4. doi: 10.12751/nncn.bc2017.0132. URL http://arxiv.org/abs/1711.02653.

N. Kriegeskorte. Deep neural networks : a new framework for modelling biological vision and brain information processing. *Annual Reviews of Vision Science*, 2015. doi: 10.1101/029876.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, pp. 408385, 2018. doi: 10.1101/408385. URL https://www.biorxiv.org/content/10.1101/408385v1.

Zhe Li, Wieland Brendel, Edgar Y Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian H Sinz, Xaq Pitkow, and Andreas S Tolias. Learning From Brains How to Regularize Machines. 2019. URL http://arxiv.org/abs/1911.05072.

A. Nayebi, D. Bear, J. Kubilius, K. Kar, S. Ganguli, D. Sussillo, J. J. DiCarlo, and D. L. K. Yamins. Task-Driven Convolutional Recurrent Models of the Visual System. (NeurIPS):1–14, 2018. URL https://arxiv.org/pdf/1807.00053.pdfhttp://arxiv.org/abs/1807.00053.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 6 1996. doi: 10.1038/381607a0.

Eftychios A. Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A. Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, Misha Ahrens, Randy Bruno, Thomas M. Jessell, Darcy S. Peterka, Rafael Yuste, and Liam Paninski. Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron*, 2016. ISSN 10974199. doi: 10.1016/j.neuron.2015.11.037.

Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009, 2019. ISSN 10974172. doi: 10.1016/j.cell.2019.04.005. URL http://dx.doi.org/10.1016/j.cell.2019.04.005.

Lutz Prechelt. Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, 11(4):761–767, 6 1998. ISSN 08936080. doi: 10.1016/S0893-6080(98)00010-0.

M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, pp. 407007, 2018. doi: 10.1101/407007. URL https://www.biorxiv.org/content/early/2018/09/05/407007.

Almut Schüz and Günther Palm. Density of neurons and synapses in the cerebral cortex of the mouse. *Journal of Comparative Neurology*, 286(4):442–455, 1989. ISSN 10969861. doi: 10.1002/cne.902860404. URL https://pubmed.ncbi.nlm.nih.gov/2778101/.

Karen Simonyan and Andrew Zisserman. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. Technical report, 2015. URL http://www.robots.ox.ac.uk/.

F. Sinz, A. S. Ecker, P. Fahey, E. Walker, E Cobos, E. Froudarakis, D. Yatsenko, X. Pitkow, J. Reimer, and A. Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In *Advances in Neural Information Processing Systems 31*. 2018. doi: 10.1101/452672.

Fabian H. Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias. Engineering a Less Artificial Intelligence, 2019. ISSN 10974199.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. Technical report, 2012.

Nicholas James Sofroniew, Daniel Flickinger, Jonathan King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife*, 2016. ISSN 2050084X. doi: 10.7554/eLife.14472.

Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *The feature-weighted receptive field: an interpretable encoding model for complex feature spaces*, 2017. ISSN 1095-9572. doi: 10.1101/126318.

Jumpei Ukita, Takashi Yoshida, and Kenichi Ohki. Characterisation of nonlinear receptive fields of visual neurons by convolutional neural network. *Scientific Reports*, 9(1):3791, 2019. doi: 10.1038/s41598-019-40535-4.

B. Vintch, J. A. Movshon, and E. P. Simoncelli. A Convolutional Subunit Model for Neuronal Responses in Macaque V1. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(44):14829–41, 2015. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2815-13.2015. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2815-13.2015%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26538653%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4635132.

E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 2019. ISSN 15461726. doi: 10.1038/s41593-019-0517-x. URL http://dx.doi.org/10.1038/s41593-019-0517-x.

D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. ISSN 1097-6256. doi: 10.1038/nn.4244. URL http://www.nature.com/doifinder/10.1038/nn.4244.

D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24, 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4060707&tool=pmcentrez&rendertype=abstract.

Yimeng Zhang, Tai Sing, Lee Ming, Li Fang, Liu Shiming, T.-S. Lee, M. Li, F. Liu, and S. Tang. Convolutional neural network models of V1 responses to complex patterns. Technical report, 2018.

# GENERALIZATION IN DATA-DRIVEN MODELS OF PRIMARY VISUAL CORTEX (APPENDIX)

## 1 TWO PHOTON SCANS

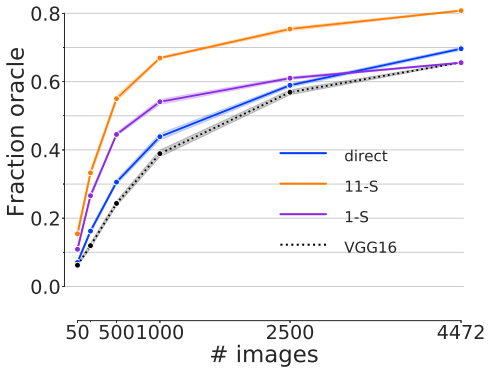
The following table lists details about the datasets used. A session marks a continuous experimental session that can comprise several scans and in which the mouse does not leave the scanner. A scan is a single continuous recording of neural activity. Spike inference from the two photon fluorescence signal is performed on the scan level.

The column **matched** indicates whether neurons were anatomically matched between scans. The four scans from mouse 22564 had 4625 matched neurons.

| animal_id | session | scan_idx | neurons | images | matched | in sets |
|---|---|---|---|---|---|---|
| 20457 | 5 | 9 | 5335 | 5993 | no | Evaluation |
| 20505 | 6 | 1 | 8367 | 5996 | no | 1-S |
| 22564 | 2 | 12 | 8115 | 5933 | yes | 4-S, 11-S |
| 22564 | 2 | 13 | 8199 | 5955 | yes | 4-S, 11-S |
| 22564 | 3 | 8 | 7916 | 5986 | yes | 4-S, 11-S |
| 22564 | 3 | 12 | 8182 | 5967 | yes | 4-S, 11-S |
| 22846 | 2 | 19 | 7700 | 5998 | no | 11-S |
| 22846 | 2 | 21 | 8044 | 5947 | no | 11-S |
| 22846 | 10 | 16 | 7344 | 5993 | no | 11-S |
| 23343 | 5 | 17 | 7334 | 5927 | no | 11-S |
| 23555 | 4 | 20 | 6848 | 5957 | no | 11-S |
| 23555 | 5 | 12 | 6559 | 5994 | no | 11-S |
| 23656 | 14 | 22 | 8107 | 5950 | no | 11-S |

## 2 GENERALIZATION ACROSS ANIMALS (EXTENSION)

We showed in Fig 4 in the main paper that the Gaussian readout outperforms the factorized readout in transfer-learning, especially in the low data regime. Consequently we conducted the main transfer experiment, the generalization across animals (Fig 5 in the main paper), with the Gaussian readout. For completeness, we here show the same experiment with the factorized readout for the relevant transfer cores *11-S*, *1-S* and *VGG16* (Fig. 1, left). The exact numeric values for this experiment with full data (5335 neurons, 4472 images) for both readouts can be found in Fig. 1 on the right. Consistent with the previous experiments, the Gaussian readout outperforms the factorized readout for direct training as well as transfer learning with data-driven cores. Interestingly however, the factorized readout scores higher than the Gaussian readout when compared on the task driven transfer core (VGG16), levelling its performance with the transfer core from one dataset (1-S). We hypothesize that this is caused by the factorized readout's less constrained spatial mask which can pool over more than one pixel in the final tensor and might thus enable it to compensate for the potentially suboptimal features in the VGG16 core.




| | Gaussian | | Factorized | |
|---|---|---|---|---|
| | mean | std | mean | std |
| 1-S | 0.696 | 0 | 0.656 | 0.002 |
| 11-S | 0.838 | 0.001 | 0.808 | 0.003 |
| 11-S: no cortex | 0.831 | 0.002 | - | - |
| 4-S: diff animals | 0.813 | 0.001 | - | - |
| 4-S: matched | 0.824 | 0.001 | - | - |
| VGG16 | 0.632 | 0.002 | 0.656 | 0.002 |
| direct | 0.749 | 0.013 | 0.696 | 0.012 |
| random VGG16 | 0.584 | 0.001 | - | - |
| random core 128 | 0.44 | 0.003 | - | - |
| random core 64 | 0.383 | 0.005 | - | - |

Figure 1: **Generalization across animals** (compare Fig 5 in the main paper). **Left:** Key experiments of Fig 5, conducted with the factorized readout. **Right:** Overview over the performances of the Gaussian and factorized readout models in the transfer-task across animals for full data (5335 neurons, 4472 images).

## 3 CONSISTENCY ACROSS OTHER PERFORMANCE MEASURES

Neural responses to (visual) stimuli suffer from trial-to-trial variability, even when keeping the input stimulus fixed. In order to get an unbiased estimate of the performance of a model that predicts such responses, the measure of performance needs to account for this statistical noise. Here we use the *fraction oracle* (Walker et al., 2019), see *Evaluation* in Section 2.2 *Networks and Training* in the main paper. However, there exists a variety of measures that attempt to tackle this issue and no standard measure has been established yet. Ideally, new findings should hold independently of the measure of performance and should be comparable across such measures. For this purpose we show the consistency of our main results (Fig. 5 in the paper), by comparing the *fraction oracle* to another measure, the *fraction of variance of the expected response* ($r^2_{ER}$) (Pospisil & Bair, 2020). The calculation of the $r^2_{ER}$ assumes that the variance over image repeats across unique images is constant. Note that this is not strictly true for our data, but to be able to compare the same model on equal ground, we chose to ignore the assumption for the sake of this comparison. Furthermore, the authors recommend that the signal-to-noise ratio of the data must be above a certain threshold (0.1 for data with 100 images and 10 repeats each, as in our case; see Fig. 14 in Pospisil & Bair (2020)). Our data meets this criterion (see Fig. 3). Figure 2 shows that both measures qualitatively yield the same results (same order of the curves).

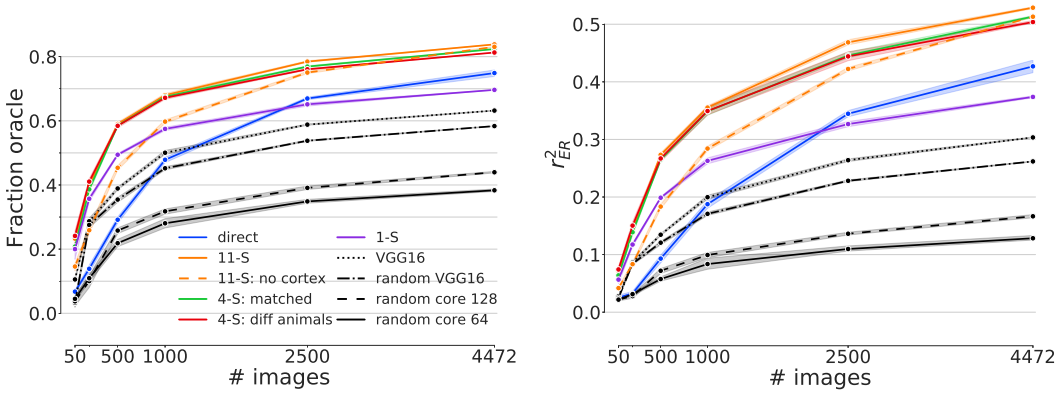



Figure 2: **Consistency across performance measures** (compare Fig 5 in the main paper). **Left:** *fraction oracle.* **Right:** *fraction of variance of the expected response.* Both measures qualitative show the same results.

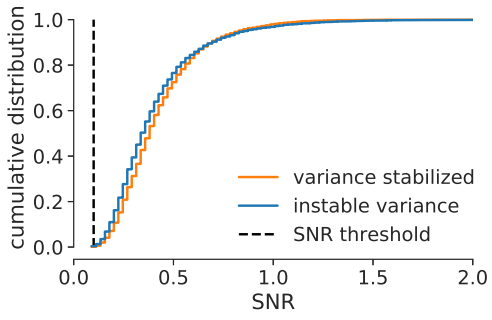



Figure 3: **Signal to noise ratio (SNR).** The SNR distribution across neurons of the evaluation dataset (blue dataset in Fig. 1 in the paper) both with (orange) and without (blue) variance stabilizing transform (Anscombe). The neurons do not fall below the threshold of 0.1 (black), justifying the use of the performance statistic $r^2_{ER}$ for our dataset (see Fig. 14 in Pospisil & Bair (2020)).

## 4 INFLUENCE OF SEEDS

The performance scores reported in our study are subject to three different sources of statistical uncertainty: The random initialization of the model weights, the specific set of images used to train the model and the specific set of neurons that we wanted to predict. In order to get an estimate of how much each of these factors contribute to the variance in the performance of our models, we trained a total of 90 models, 30 for each source of uncertainty, and varied the respective seeds. While the seed of one source was altered, the seeds of the remaining two sources were kept fixed. Since the impact of the neuron and image seed naturally increases with decreasing amounts of data, we conducted this experiment on a medium range data regime of 1000 images and 1000 neurons. The results can be seen in Fig. 4. While the main contributions to the variance in model performance seem to stem from the model initialization and the random subset of neurons, the image seed did not seem to have a major influence. We thus only used a single value for it in most experiments in the paper. Since we do not consider the variance caused by the random initialization of the model weights as relevant for the underlying scientific problem, we decided to pick the models which performed best on the validation set across 5 model initialization seeds. Finally, we computed 95% confidence intervals across 5 seeds of random neuron subsets. In the cases where all available neurons were used in an experiment, the statistics were computed across 5 image seeds instead (see section Data in the paper). The total number of trained models per data point was thus always 25.

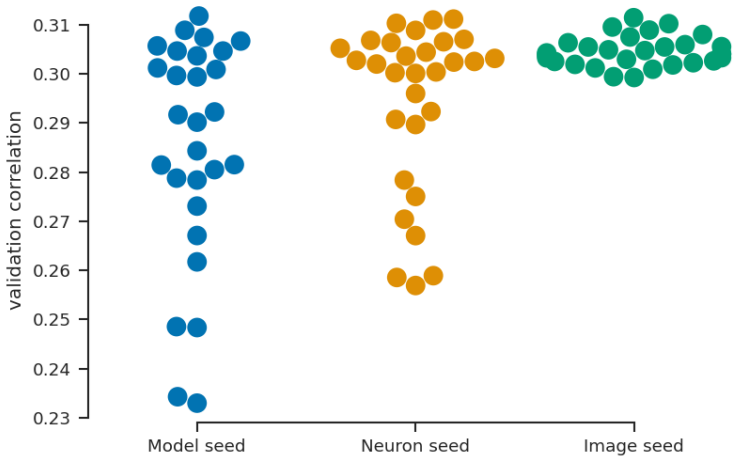



Figure 4: **Variation of model performance across seeds.** Several models (with Gaussian readout) were trained on 1000 neurons and 1000 images each, while varying the initialization of the model (blue), the specific set of 1000 neurons (yellow) and the specific set of 1000 images (green). Since the image seed did not have a major influence on model performance, we decided to only use a single seed value, select the best performing models across 5 model seeds and compute statistics across 5 neuron seeds throughout most of our experiments. In the cases where all available neurons were used in an experiment, the statistics were computed across 5 image seeds instead (see Chapter Data in the paper).

## 5 INFLUENCE OF CORTICAL DATA AND FEATURE SHARING ON THE GAUSSIAN READOUT

The models with Gaussian readout outperformed the ones with factorized readout, both in direct and transfer learning (see Fig. 3 and 4 in the paper). Here, we investigate which of its components this can be attributed to. To this end we trained models with Gaussian readout directly on the four matched datasets with 3625 neurons and varying number of images (Fig. 5, compare also Fig. 3 in the paper). We did this with and without using the components *feature sharing* and *cortex-data*: In the *feature sharing* condition, each neuron shared the same feature weight vector with its anatomical matches across the four datasets. The models with the *cortex-data* condition predicted the receptive field positions from anatomical cortical data via an affine transform. Both, *feature sharing* and *cortex-data* were switched on throughout the paper, and contributed to the good performance of the Gaussian readout. The better performance of the Gaussian readout compared to the factorized readout in Fig. 3 in the paper seems to be mainly due to *feature sharing*. The usage of cortical data to learn the receptive field positions is primarily advantageous for mid-range number of images.

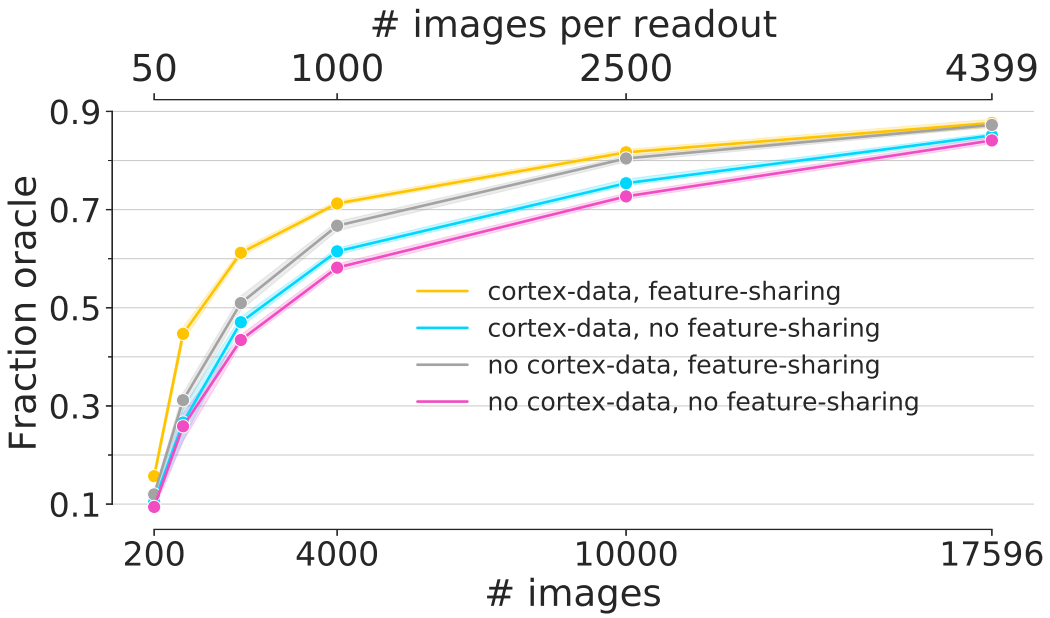



Figure 5: **Gaussian readout with and without feature sharing and position learning from anatomical data.** The training procedure is the same as in Fig. 3 in the paper. The *feature sharing* seems to be the main contributor to the better performance of the Gaussian readout compared to the factorized readout in Fig. 3 in the paper. The usage of cortical data to learn the receptive field positions seems to be primarily advantageous for low and mid range number of images.

## 6 MOST EXCITING INPUTS FOR MODELED NEURONS

One important application of general system identification models is the analysis of neural tuning, the relation that connects a neuron's response to the stimulus. Describing neural response properties by the stimuli that drive them best has a long tradition in neuroscience (such as Gabor filters and gratings in early visual cortex, or face-selective cells in higher layers). Walker et al. (2019) and Bashivan et al. (2019) introduced a method to obtain such most exciting inputs (MEIs) which we analogously generated for our model with the 11-S transfer core (Fig. 5 in the paper, orange line). Like Walker et al. (2019) we use an ensemble of networks to generate the MEI. In our case, we used an ensemble of five 11-S transfer cores from five seed initializations for which we each trained a readout with the evaluation dataset on top. In Fig. 6 we show these MEIs for the 50 neurons with the best test performance. Walker et al. (2019) have shown that many MEIs differ quit strongly from Gabor-like stimuli, which would be expected to be the best drivers for V1 neurons based on previous work in monkeys and cats. Our MEIs exhibit very similar characteristics to those presented by Walker et al. (2019), which were obtained from a directly trained network and experimentally verified, highlighting the generality of our transfer core.

## REFERENCES

P. Bashivan, K. Kar, and J. DiCarlo. Neural Population Control via Deep ANN Image Synthesis. pp. 1–33, 2019. doi: 10.32470/ccn.2018.1222-0.

Dean A Pospisil and Wyeth Bair. The unbiased estimation of the fraction of variance explained by a model. *bioRxiv*, pp. 2020.10.30.361253, nov 2020. doi: 10.1101/2020.10.30.361253. URL https://doi.org/10.1101/2020.10.30.361253.

E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 2019. ISSN 15461726. doi: 10.1038/s41593-019-0517-x. URL http://dx.doi.org/10.1038/s41593-019-0517-x.

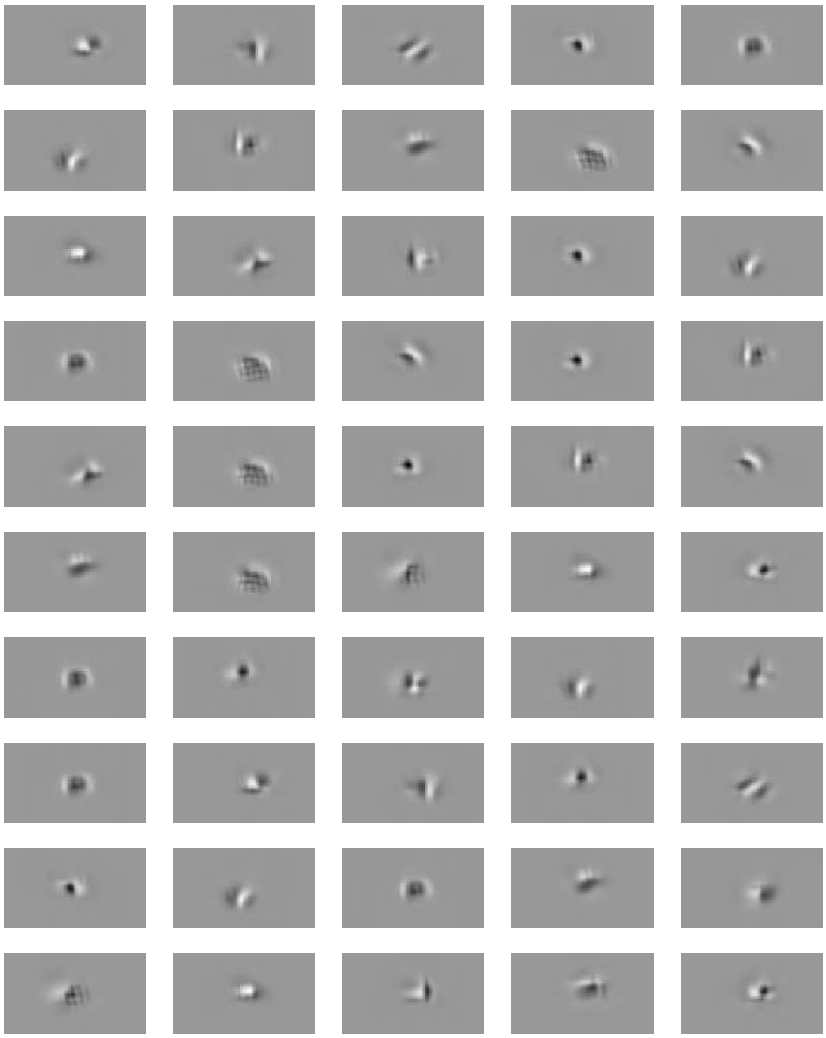


Figure 6: **Most exciting inputs (MEIs).** The image inputs that best drive the 50 best predicted neurons from the evaluation dataset, predicted with the best transfer core (Fig. 5 in the paper, orange line).