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

1 TWO PHOTON SCANS

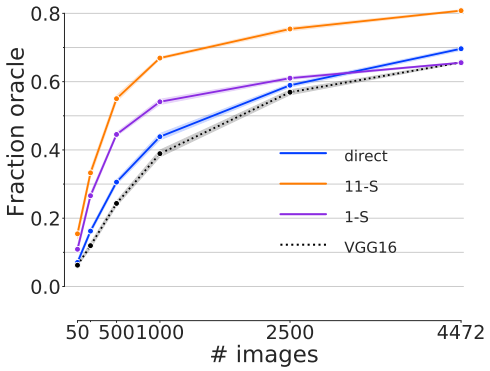
The following table lists details about the datasets used. A session marks a continuous experimental session that can comprise several scans and in which the mouse does not leave the scanner. A scan is a single continuous recording of neural activity. Spike inference from the two photon fluorescence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

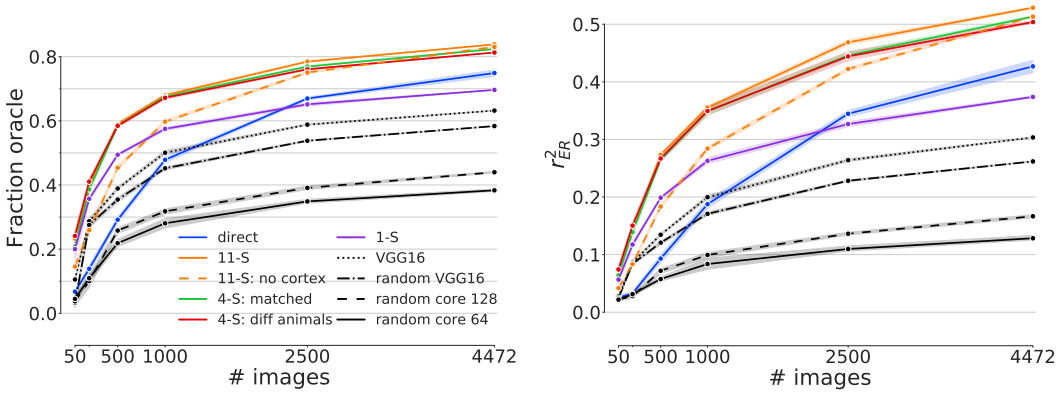


Figure 2: **Consistency across performance measures** (compare Fig 5 in the main paper). **Left:** *fraction oracle.* **Right:** *fraction of variance of the expected response.* Both measures qualitative show the same results.

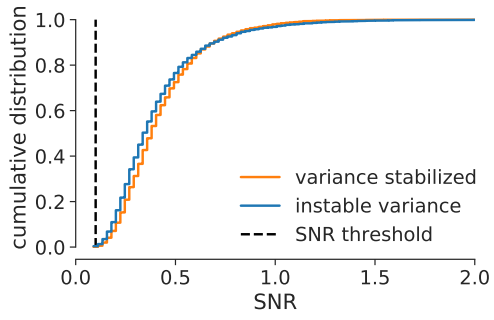


Figure 3: **Signal to noise ratio (SNR).** The SNR distribution across neurons of the evaluation dataset (blue dataset in Fig. 1 in the paper) both with (orange) and without (blue) variance stabilizing transform (Anscombe). The neurons do not fall below the threshold of 0.1 (black), justifying the use 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.

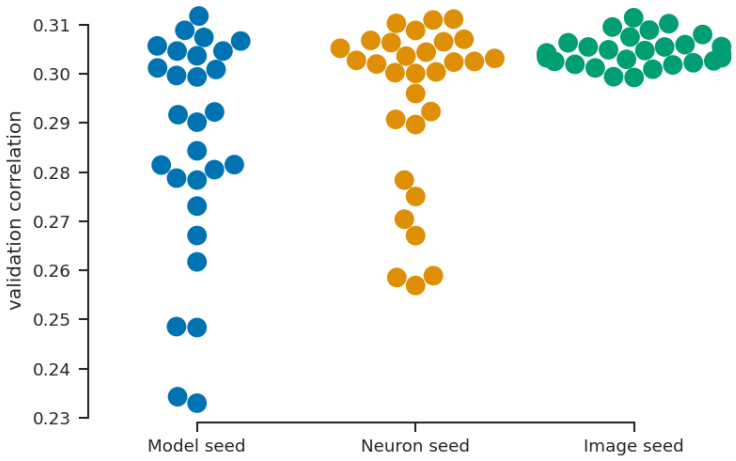


Figure 4: **Variation of model performance across seeds.** Several models (with Gaussian readout) were trained on 1000 neurons and 1000 images each, while varying the initialization of the model (blue), the specific set of 1000 neurons (yellow) and the specific set of 1000 images (green). Since the image 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.

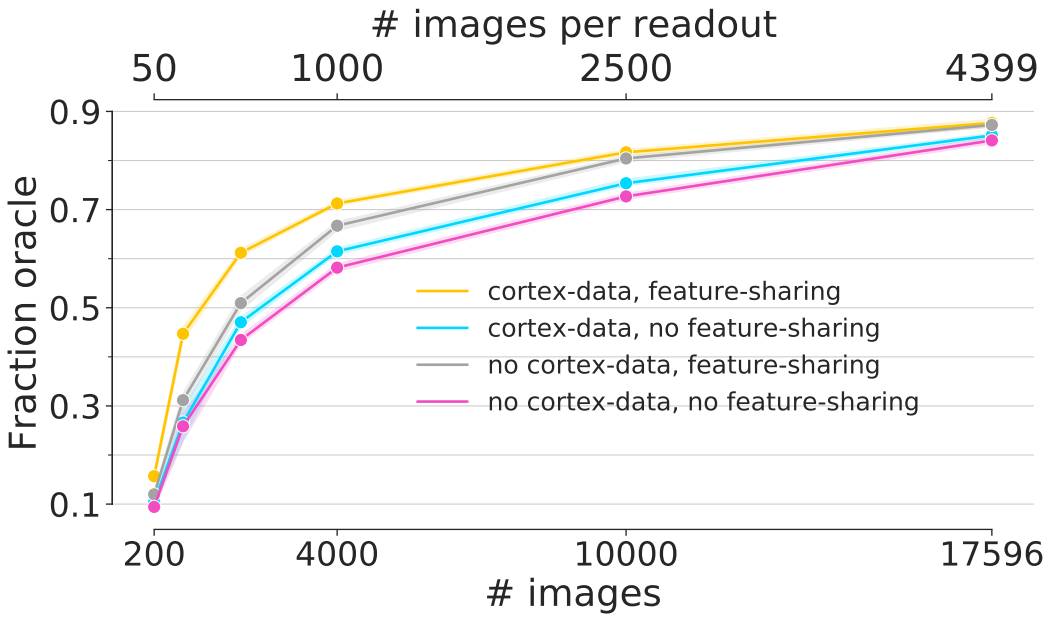


Figure 5: **Gaussian readout with and without feature sharing and position learning from anatomical data.** The training procedure is the same as in Fig. 3 in the paper. The *feature sharing* seems to be the main contributor to the better performance of the Gaussian readout compared to the factorized 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.

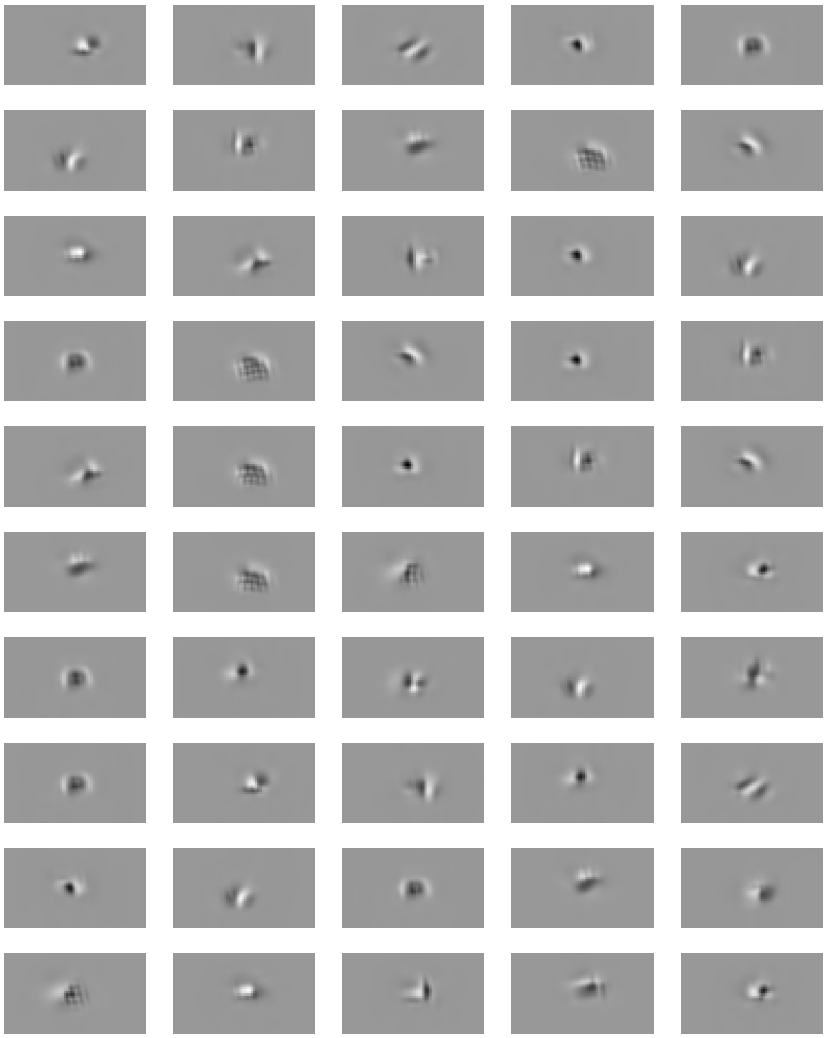


Figure 6: **Most exciting inputs (MEIs).** The image inputs that best drive the 50 best predicted neurons from the evaluation dataset, predicted with the best transfer core (Fig. 5 in the paper, orange line).