Towards a Foundation Model of the Mouse Visual Cortex

Eric Y. Wang[1,2], Paul G. Fahey[1,2], Kayla Ponder[1,2], Zhuokun Ding[1,2], Andersen Chang[1,2], Taliah Muhammad[1,2], Saumil Patel[1,2], Zhiwei Ding[1,2], Dat Tran[1,2], Jiakun Fu[1,2], Stelios Papadopoulos[1,2], Katrin Franke[1,2], Alexander S. Ecker[3,4], Jacob Reimer[1,2], Xaq Pitkow[1,2,5], Fabian H. Sinz[1,2,3,6], and Andreas S. Tolias[1,2,5] ✉

[1]Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, USA
[2]Department of Neuroscience, Baylor College of Medicine, Houston, USA
[3]Institute for Computer Science, University Göttingen, Göttingen, Germany
[4]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany
[5]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA
[6]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Germany

Understanding the brain's perception algorithm is a highly intricate problem, as the inherent complexity of sensory inputs and the brain's nonlinear processing make characterizing sensory representations difficult. Recent studies have shown that functional models capable of predicting large-scale neuronal activity in response to arbitrary sensory input can be powerful tools for characterizing neuronal representations by enabling unlimited *in silico* experiments. However, accurately modeling responses to dynamic and ecologically relevant inputs like videos remains challenging, particularly when generalizing to new stimulus domains outside the training distribution. Inspired by recent breakthroughs in artificial intelligence, where foundation models—trained on vast quantities of data—have demonstrated remarkable capabilities and generalization, we developed a "foundation model" of the mouse visual cortex: a deep neural network trained on large amounts of neuronal responses to ecological videos from multiple visual cortical areas and mice. The model accurately predicted neuronal responses not only to natural videos but also to various new stimulus domains, such as coherent moving dots and noise patterns, as verified *in vivo*, underscoring its generalization abilities. The foundation model could also be adapted to new mice with minimal natural movie training data. We applied the foundation model to the MICrONS dataset: a study of the brain that integrates structure with function at unprecedented scale, containing nanometer-scale morphology, connectivity with >500,000,000 synapses, and function of >70,000 neurons within a $\sim$1mm^3 volume spanning multiple areas of the mouse visual cortex. This accurate functional model of the MICrONS data opens the possibility for a systematic characterization of the relationship between circuit structure and function. By precisely capturing the response properties of the visual cortex and generalizing to new stimulus domains and mice, foundation models can pave the way for a deeper understanding of visual computation.

Visual cortex | Foundation model | Generalization | Artificial Intelligence

Correspondence: ***astolias@bcm.edu***

Introduction

A crucial step to decipher the brain's algorithm of perception is to build accurate functional models of neuronal activity that predict how the visual system responds to sensory stimuli and how activity is modulated by behavioral and internal brain states. Functional models of neuronal responses to visual inputs have a long history in neuroscience from simple linear-nonlinear (LN) models (Jones and Palmer, 1987; Heeger, 1992a,b), energy models (Adelson and Bergen, 1985), subunit/LN-LN models (Rust et al., 2005; Touryan et al., 2005; Vintch et al., 2015), to multi-layer neural network models (Zipser and Andersen, 1988; Lehky et al., 1992; Lau et al., 2002; Prenger et al., 2004).

Recently, deep artificial neural networks (ANNs) have become the new standard for modeling visual cortex (Yamins et al., 2014; Cadieu et al., 2014; Antolík et al., 2016; Batty et al., 2017; McIntosh et al., 2016; Klindt et al., 2017; Kindel et al., 2017; Cadena et al., 2019; Burg et al., 2021; Lurz et al., 2020; Bashiri et al., 2021; Christensen and Zylberberg, 2020; Cowley and Pillow, 2020; Ecker et al., 2018; Sinz et al., 2018; Bakhtiari et al., 2021; Nayebi et al., 2021; Willeke et al., 2022). Their ability to capture complex and highly nonlinear relationships have enabled them to accurately predict neuronal responses to arbitrary static natural images and even synthesize novel stimuli, such as the most exiting stimulus for individual neurons (Bashivan et al., 2019; Walker et al., 2019; Ponce et al., 2019; Franke et al., 2022; Höfling et al., 2022). These models offer the possibility to perform a nearly unlimited number of *in silico* experiments to systematically characterize neuronal representations, such as identifying what individual neurons are selective for (Walker et al., 2019; Bashivan et al., 2019; Franke et al., 2022; Fu et al., 2023), what they are invariant to (Ding et al., 2023b), and relating the geometry of the population activity to the sensory input and behavior. The resulting predictions can then be tested through *in vivo* closed loop experiments, such as the *inception loops* paradigm (Walker et al., 2019; Franke et al., 2022). This *in silico–in vivo* approach addresses inherent challenges of studying neuronal representations, including the high dimensionality of the input space, the non-linear nature of information processing in the brain, and the limited availability of time for conducting *in vivo* experiments.

However, although animals experience dynamic visual input analogous to videos, most models to date are designed for static images. Building models that accurately predict responses to video input is more challenging. Introducing the temporal component introduces an entirely new dimension of vision that is absent from static models. To deal with this increased complexity, dynamic models typically contain more parameters and require more data to train than their static

counterparts. Another challenge in neural network modeling is predicting on new stimulus domains outside the original training distribution (Hendrycks and Dietterich, 2019). For instance, when models are trained to generate responses to natural movies, they perform well at predicting unseen natural movies but exhibit a substantial decrease in prediction performance on other domains such as synthetic or parametric stimuli (Sinz et al., 2018). However, to build upon the long history of using parametric stimuli for visual psychophysics and neurophysiology (Britten et al., 1992; Salzman et al., 1990; Marshel et al., 2019) and to increase their usefulness for *in silico* experiments, it is crucial to develop functional models that generalize well to novel stimulus domains, such that tuning functions can be characterized *in silico* with parametric stimuli, for example.

Recently, so called *foundation models* (Bommasani et al., 2021) in artificial intelligence, characterized by their ability to train on massive amounts of data and build robust representations of their modeling domain, have demonstrated remarkable generalization and capabilities in downstream tasks (Brown et al., 2020; Radford et al., 2021). For example, foundation models of language are trained on vast quantities of text encompassing much of human knowledge. Trained to predict the next sub-word in text, these foundation models capture robust language and knowledge representations that can be transferred to new tasks with relatively little data. These tasks include answering unstructured questions and even passing medical licensing exams (Kung et al., 2023).

Inspired by these breakthroughs, we sought to develop a foundation model of the mouse visual cortex trained on extensive quantities of data to predict neuronal activity from dynamic video and behavior as inputs. We collected the responses to ecological video stimuli from ∼140,000 neurons across multiple areas of the visual cortex from 14 awake, behaving mice. Using a highly optimized recurrent deep neural network architecture trained on a subset of these data, we learned a common, data-driven dynamic "foundation core" that effectively captured the shared latent representations of all neurons we studied and accurately predicted neuronal responses across many mice and visual cortical areas. New models utilizing the foundation core demonstrated the ability to be rapidly and accurately fitted to new mice with minimal amounts of data, surpassing the performance of individualized models that were trained end-to-end for each mouse specifically. These models excelled not only in predicting neuronal responses to new natural movies (in-domain) but also generalized to accurately predict responses to various out-of-domain stimuli, including random moving dots, flashing dots, Gabor patches, coherent moving noise, and static natural images. Finally, using the foundation core, we produced accurate functional models for the MICrONS study (MICrONS Consortium et al., 2021): a publicly available dataset containing >70,000 neurons within a $\sim$1mm^3 cortical volume spanning multiple visual areas. In addition to neuronal function, the MICrONS dataset contains anatomical information about the morphology and connectivity of these neurons on the nanoscale-resolution, providing a comprehensive dataset for relating structure and function (functional connectomics, Ding et al. 2023a).

Results

State-of-the-art dynamic functional model of the mouse visual cortex. To model the dynamic neuronal responses of the mouse visual cortex, we developed an artificial neural network (ANN) that was comprised of four components: perspective, modulation, core, and readout (Fig. 1). The modular design enabled the ANN to accommodate diverse tasks and inputs. For instance, eye movements and different positioning of a mouse's head relative to the monitor can result in different perspectives of the same stimulus, despite best efforts to limit experimental variability. To account for this, the perspective component of our ANN uses ray tracing and eye tracking data to infer the perspective of the mouse from the presented stimulus on the monitor (Extended Data Fig. 1). To account for behavioral factors that modulate the activity of the visual cortex (Reimer et al., 2014), the modulation component transforms behavioral inputs (locomotion, pupil dilation) to produce dynamic representations of the mouse's behavioral and attentive state (Extended Data Fig. 2). The perspective and modulation components provide visual and behavioral inputs, respectively, to the core component of the ANN. Composed of convolutional feedforward and recurrent sub-networks, the core contains the majority of the ANN's modeling capacity and produces nonlinear representations of vision that are modulated by behavior. These representations are mapped onto the activity of individual neurons by the readout component, which performs a linear combination of the features generated by the core at one specific location, the neuron's receptive field. All four components of the ANN were trained end-to-end to predict time series of neuronal responses to natural movies (for details of model architecture and training, see Methods).

First, we evaluated the predictive accuracy of our ANN model architecture when trained on standard amounts of experimental data: i.e. on individual recording sessions lasting ∼1 hour. Predictive accuracy was measured by the correlation between the recorded and the predicted responses to a novel set of stimuli that were not included in model training. To account for *in vivo* noise, the correlation was normalized by an estimated upper bound on the performance that could be achieved by a perfect model (Schoppe et al., 2016). Using this normalized correlation coefficient (CC_{norm}) as the metric of predictive accuracy, we compared our model to the previous best-performing dynamic model of the mouse visual cortex (Sinz et al., 2018). Trained and tested on the same data from that study (dynamic V1 responses to natural movies), our model had a 25–46% increase in predictive accuracy on held-out test data across the three recording sessions used in Sinz et al. (2018) (Fig. 2a). This level of increase in performance is substantial for predictive models of the visual cortex. For comparison, in a recent competition to model neuronal responses to static images, the winning model out of 172 submissions from 26 teams provided an 18% improvement over the previous state-of-the-art static model (Willeke

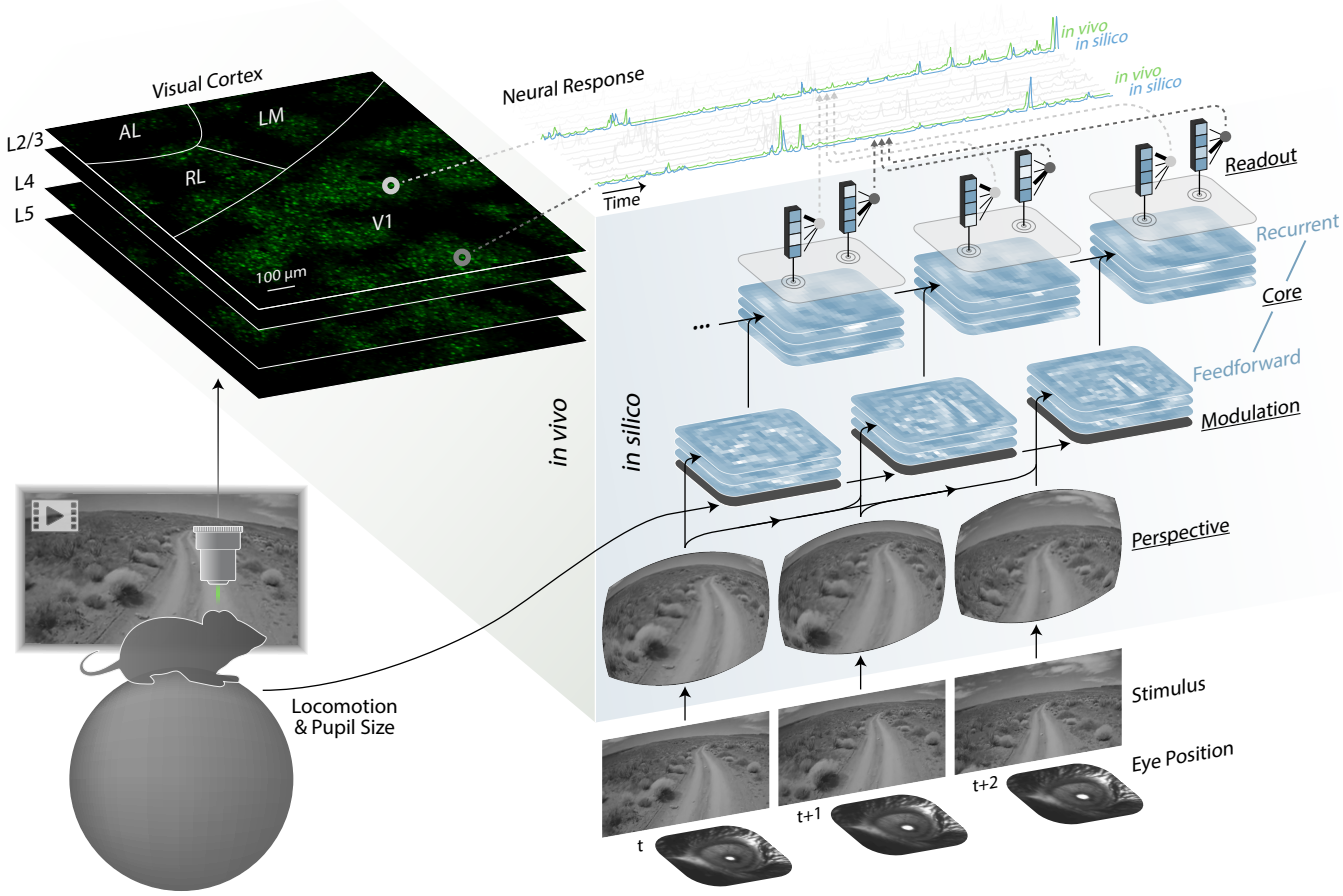


Fig. 1. ANN model of the visual cortex. The left panel (green) depicts an *in vivo* recording session of excitatory neurons from several areas (V1, LM, RL, AL) and layers (L2/3, L4, L5) of the mouse visual cortex. The right panel (blue) shows the architecture of the ANN model and the flow of information from inputs (visual stimulus, eye position, locomotion, and pupil size) to outputs (dynamic neuronal response). Underlined labels denote the four main components of the ANN: perspective, modulation, core, and readout. For the modulation and core, the stacked planes represent feature maps. For readout, the blue boxes represent the core's output features at the readout position of the neuron, and the fanning black lines represent readout feature weights. The top of the schematic displays the dynamic neuronal response for a sampled set of neurons. For two example neurons, *in vivo* and *in silico* responses are shown (green and blue, respectively).

et al., 2022, 2023). We also evaluated the predictive accuracy of our model on newly collected data that contained multiple visual areas (Fig. 2b). Interestingly, we found that the performance of our model for higher visual areas (LM, RL, AL) was similar to V1 (Fig. 2c), despite the increased complexity of neuronal tuning to more complex features exhibited by higher visual areas (Siegle et al., 2021; Goltstein et al., 2021).

Foundation models generalize to new subjects and stimulus domains. While our new ANN model architecture sets new standards for predicting dynamic neuronal responses of the visual cortex, the performance depends critically on the amount of data used for training, a property exhibited by ANNs in general (Fig. 2b). The remarkable performance of foundation models in other domains—e.g., natural language (Brown et al., 2020) and image generation (Radford et al., 2021)—originates from their vast quantities of training data. However, collecting large amounts of neuronal data from individual neurons and animals presents challenges. Individual recording sessions are limited in duration by experimental factors such as attentiveness and recording device stability. To overcome this limitation, we combined data from multiple recording sessions, totaling over 900 minutes of natural movie responses from 8 mice, 6 visual areas (V1, LM, AL, RL, AM, PM), and ∼67,000 neurons. This data was used to train a single, shared ANN core (Fig. 3a) with the goal of capturing common representations of vision that underlie the dynamic neuronal response of the visual cortex for a representative group of mice. This representation could then be used to fit models of new mice to improve their performance with limited data. Here we refer to the representative group of 8 mice as the "foundation cohort", the trained ANN component as the "foundation core", and ANNs derived from the foundation core as "foundation models".

To evaluate the representation of the visual cortex captured by the foundation core, we froze its parameters and transferred it to ANNs with new perspective, modulation, and readout components fitted to the new mice (Fig. 3a). Each new mouse was shown an assortment of stimuli, designated for either model training or testing. The training stimuli consisted of natural movies, and we used different portions of this, spanning from 4 to 76 minutes, to fit ANN components to the new mice. This approach aimed to examine the relationship between the models' performance and the amount of training data for each new mouse. The testing stimuli included natural

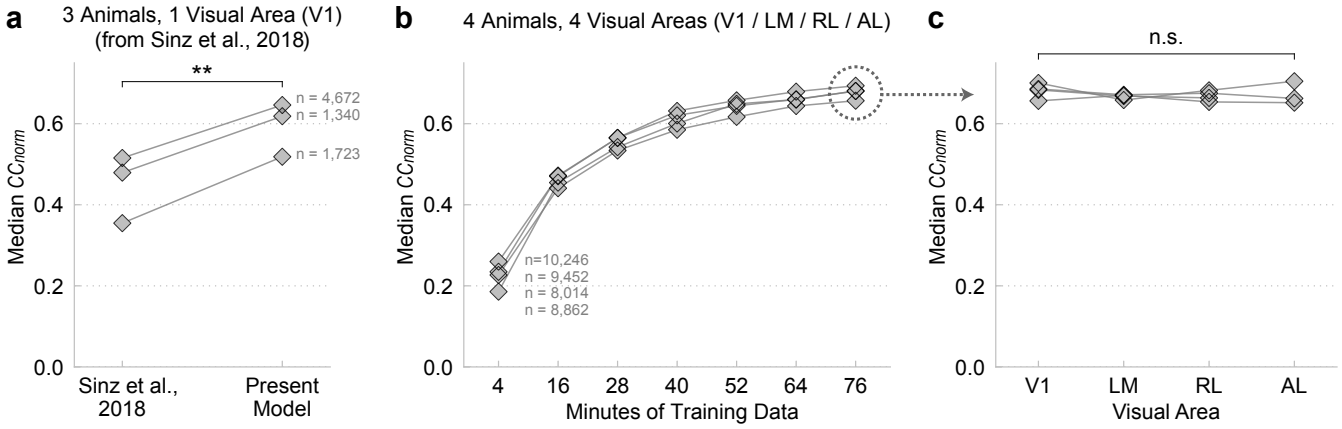


Fig. 2. Predictive accuracy of models trained on individual recording sessions. a, Predictive accuracy (median CC_{norm} across neurons, see Methods for details) of our model vs. the previous state-of-the-art dynamic model of the mouse visual cortex by Sinz et al. (2018). We trained and tested our model on the same set of data from Sinz et al. (2018): V1 neuronal responses to natural movies from 3 mice. Paired *t*-test (two-way): **, p < 0.01. n = number of neurons per mouse. **b,** Predictive accuracy of our models by the amount of data used for training for 4 new recording sessions and mice. For each recording session, training data was partitioned in to 7 fractions ranging from 4 to 76 minutes. Separate models (diamonds) were trained on the differing fractions of training data, but tested on the same held-out testing data. Models of the same mice are connected by lines. **c,** Predictive accuracy by visual area, from models that were trained on the full data. We did not find a statistically significant relationship between predictive accuracy and visual areas (linear mixed effects model (Lindstrom and Bates, 1988), Wald's test: n.s., p = 0.45).

movies that were not part of the training set (Fig. 3b'), and new stimulus domains like static natural images (Fig. 3c'), and 4 types of parametric stimuli (Fig. 3d'–g'), consisting of drifting Gabor filters, flashing Gaussian dots, directional pink noise, and random dot kinematograms. To test the role of the foundation core in prediction performance, we trained a set of control models that differed from the foundation models only by the core component. For these controls or "individual models", all four components—core, perspective, modulation, and readout—were trained end-to-end using training data from a single recording session. For the foundation models, training data from the new mice were only used to fit the perspective, modulation, and readout components, and the core was trained on the foundation cohort as described above and was frozen (Fig. 3a).

When tested on natural movies, foundation models outperformed individual models and required less training data from the new mice to achieve high levels of predictive accuracy (Fig. 3b). For instance, individual models required more than an hour of training data for to surpass a median CC_{norm} of 0.65 for all mice, whereas foundation models required less than half an hour (Fig. 3b). This performance gain was observed across all tested stimulus domains, including those that were in new stimulus domains (Fig. 3c'–g'), i.e., out-of-distribution (OOD) from the training domain of natural movies (Fig. 3b'). Importantly, no stimuli from the OOD domains were used to train any component of the models, including the foundation core. Nevertheless, foundation models were more accurate at predicting responses to new stimulus domains while requiring significantly less training data from the new mice (Fig. 3c–g). For example, when predicting drifting Gabor filters, the foundation models were able to achieve a performance of median $CC_{norm} > 0.55$ using only 16 minutes of natural movie training data. In contrast, the individual models required more than an hour of training data to reach the same performance level (Fig. 3d). This highlights the significant difference in the data efficiency of these models, i.e., the amount and sample complexity of training data required from new subjects to accurately fit their neuronal responses. Thus, training a foundation dynamic core on natural movie data pooled from multiple cortical layers, areas and animals produces a robust and transferable representation of the visual cortex that generalizes to new animals and improves model performance for not only natural movies but also novel stimulus domains.

Parametric tuning emerges from foundation models. By leveraging the foundation core and transfer learning, we were able to create accurate foundation models for individual mice (Fig. 3). These models enable essentially unlimited *in silico* experiments for studying representations, testing theories, generating novel hypotheses that can be verified *in vivo*. Here we assessed the precision with which classical tuning properties of the visual cortex could be replicated at the individual neuronal level in our foundation model. We presented mice—not part of the foundation cohort—with natural movie stimuli in order to train their ANN counterparts (Fig. 4a). Additionally, we presented parametric stimuli (Fig. 4b'–c') to measure the orientation, direction and spatial tuning of the recorded neurons. Subsequently, we presented the same parametric stimuli to the corresponding *in silico* neurons and measured their properties for comparison (Fig. 4b–c). This was done for 3 mice and ∼30,000 neurons from 4 visual areas (V1, LM, AL, RL).

To measure orientation and direction tuning, we presented directional pink noise (Fig. 4b'), which encoded coherent motion of different directions (0–360°) and orientations (0–180°). First, we computed the strength of orientation and direction tuning via selectivity indices for orientation (OSI) and direction (DSI). There was a significant correlation between *in vivo* and *in silico* estimates for both OSI (Fig. 4d) and DSI (Fig. 4f), which validated the foundation model's estimates

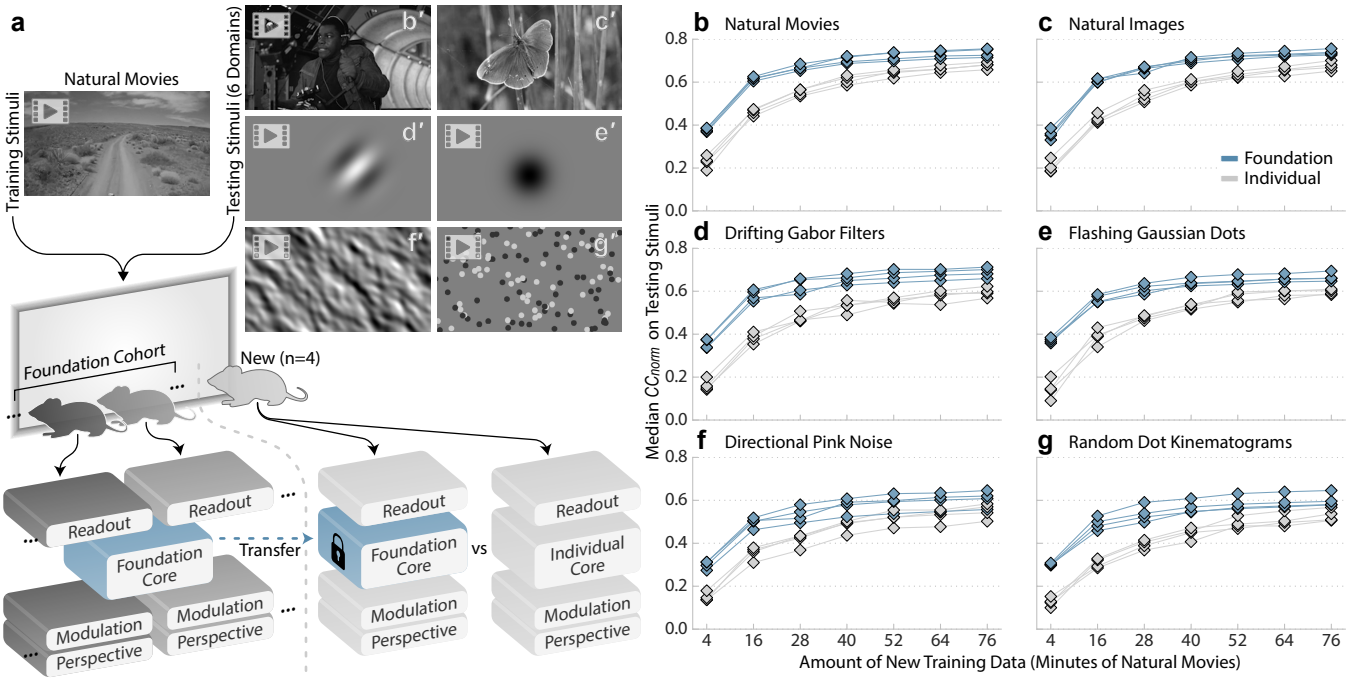


Fig. 3. Predictive accuracy of foundation models. a, Schematic of the training and testing paradigm. Natural movie data were used to train: 1) a combined model of the foundation cohort of mice with a single foundation core, and 2) foundation models vs. individual models of new mice. The models of the new mice were tested with stimuli from 6 different domains (**b'–g'**). **b–g**, Corresponding plots show the predictive accuracy (median CC_{norm} across neurons) as a function of the amount of training data for foundation models (blue) vs. individual models (gray) of the new mice. 4 mice $\times$ 7 partitions of training data $\times$ 2 types of models = 56 models (diamonds). Models of the same mouse and type (foundation / individual) are connected by lines. Number of neurons per mouse = 8,862 | 8,014 | 9,452 | 10,246.

of tuning strength for orientation and direction. Next, we estimated the preferred angles of orientation and direction of neurons by fitting a directional parametric model (mixture of von Mises distributions) to the responses. For strongly tuned neurons, the *in vivo* and *in silico* estimates of preferred angles of orientation and direction were closely matched (Fig. 4e,g). For example, for strongly orientation-tuned neurons with an *in silico* OSI > 0.5 (11% of neurons), the median difference between the *in vivo* and *in silico* estimates of preferred orientation was 4°, and with a lower OSI threshold of > 0.3 (43% of neurons), the median difference was 7°(Fig. 4e).

To measure spatial tuning, we presented flashing Gaussian dots (Fig. 4c') to the neurons described above. We computed a spike-triggered average (STA) of the stimulus, which was used to estimate: 1) the strength of spatial tuning for Gaussian dots (non-uniformity of the STA) via the spatial selectivity index (SSI); and 2) the preferred location (peak of the STA) via least squares fitting of the STA to a spatial parametric model (2D Gaussian distribution). As with orientation and direction tuning, we observed a significant correlation between *in vivo* and *in silico* estimates of spatial tuning strength, measured by SSI (Fig. 4h). For strongly tuned neurons with *in silico* SSI > 8 (1% of neurons), the median distance between the *in vivo* and *in silico* estimates of the preferred location was 0.02 of the monitor width (Fig. 4i), approximately 2° in visual space. Compared to directional pink noise, we observed that a much fewer proportion of the neurons we recorded were strongly tuned to Gaussian dots.

Together, these results demonstrate the accuracy of estimating tuning parameters for classical functional properties from our foundation model with no prior training on parametric stimuli. Therefore, rather than presenting parametric stimuli *in vivo*, parametric tuning can be performed *in silico* with an accurate and validated foundation model, freeing up valuable *in vivo* experimental time for other purposes.

Foundation model of the MICrONS mouse. Underlying the functional capabilities of the neocortex is an intricate circuitry of cellular and molecular structures. The MICrONS project was a landmark study in visual neuroscience that interrogated the relationship between structure and function of the mouse visual cortex at unprecedented scale and resolution. From a single mouse, the responses of >70,000 excitatory neurons presented with natural movies were measured through 14 sequential experiments, tiling a $\sim$1 mm^3 volume encompassing V1, LM, AL and RL. The volume subsequently underwent serial electron microscopy (EM) and dense morphological reconstruction (Fig. 5b), yielding the detailed structures of $\sim$60,000 excitatory neurons and $\sim$500 million synapses (MICrONS Consortium et al., 2021). The *in vivo* and EM data were co-registered to produce the largest integrated study of structure and function of the neocortex to date.

When combining functional studies of the brain with other modalities like anatomy, there is a finite amount of time for *in vivo* recordings before histological analysis renders the tissue unusable. While traditionally this would limit the number of functional studies that can be performed *in vivo*, predictive models allow essentially unlimited experiments to be performed *in silico*. To enable this for the MICrONS project,

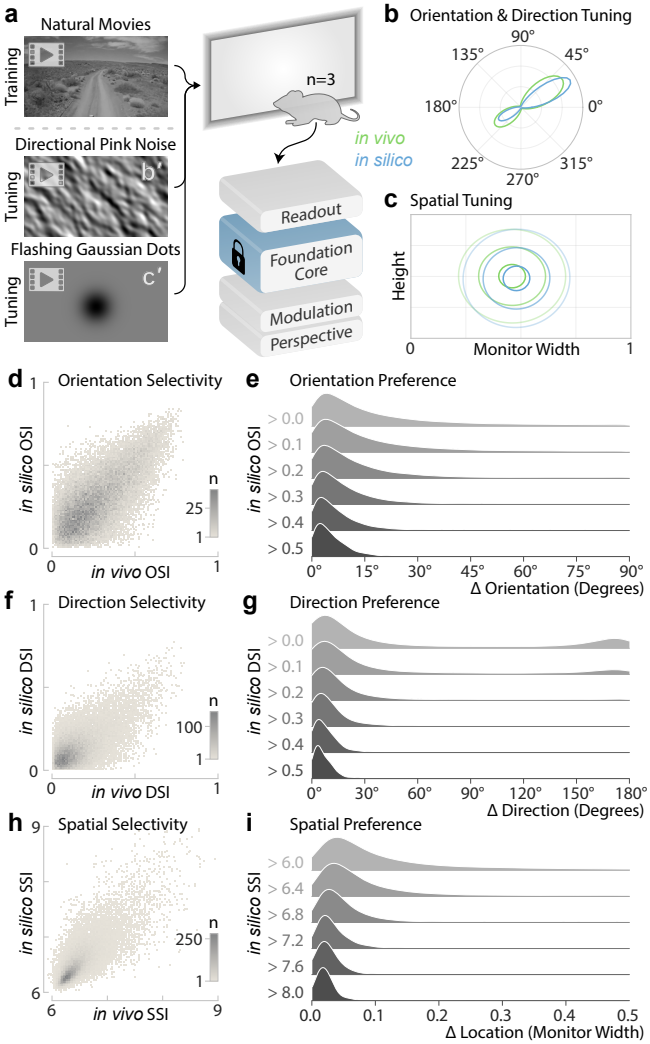


Fig. 4. Emergence of parametric tuning from the foundation model. a, Schematic of the experimental paradigm: foundation models of new mice (n=3) were trained with natural movies, and estimates of parametric tuning were computed from *in vivo* and *in silico* responses to synthetic stimuli (**b',** directional pink noise; **c',** flashing Gaussian dots). **b,c,** *In vivo* and *in silico* estimates of an example neuron's parametric tuning to orientation/direction (**b**) and spatial location (**c**). **d,f,h,** Binned scatter plots of *in vivo* and *in silico* estimates of selectivity indices (SI) for orientation (**d,** OSI), direction (**f,** DSI), and spatial (**h,** SSI). The color indicates the number of neurons (n) in each bin. **e,g,i,** Density histograms of differences between *in vivo* and *in silico* estimates of preferred orientation (**e**), direction (**g**), and spatial location (**i**). In each panel, histograms containing increasingly selective groups of neurons, thresholded by *in silico* OSI (**e**) / DSI (**g**) / SSI (**i**), are stacked from top to bottom. The density histograms were produced via kernel density estimation using Scott's bandwidth.

responses to natural movies were collected for the purpose of model training. Due to the challenge of completing all 14 experiments in the same animal in as short a period as possible, the amount of training data collected from each experiment (mean 42 minutes, range 33–53 minutes, depending on optical quality and animal behavioral profile) was less than the other recording sessions in this study. With the available amount of data, individual models, with all components trained from scratch on a single experiment, achieved a median CC_{norm} of 0.48–0.65, when tested on a held-out set of natural movies. By applying our foundation modeling paradigm—transferring the foundation core and fitting only the perspective, modulation, and readout components—the median CC_{norm} increased to 0.58–0.76 (Extended Data Fig. 3). This highlights the advantage of the foundation modeling approach when there is a limited amount of data available for training.

Fundamental properties of cortical organization—retinotopy and functional specialization into visual areas—emerged from the readout parameters of our foundation model of the MICrONS mouse. To map the core's representations onto the activity of individual neurons, the readout component of the ANN performs a linear combination of the core's features at a single position, the neuron's receptive field (Fig. 1). This means that the readout parameters for each neuron are factorized into two elements: position and feature weights (Fig. 5a). The position encodes the center of the neuron's spatial receptive field, and the feature weights encode the relative importance of the core's features at that position to the neuron. These parameters were initialized and trained without explicit information pertaining to cortical anatomy or visual areas. Nevertheless, by training the model to predict neuronal responses to natural scenes, retinotopy and area-specific properties of neuronal function emerged implicitly from the readout positions and feature weights, respectively (Fig. 5c–d).

One of the main organizing principles and anatomical features of the visual cortex is retinotopy: the topographic mapping between the cortical position of neurons and the location of their receptive fields. In mice, there is a nasotemporal (azimuthal) progression of receptive fields from lateral to medial regions of V1, and an inferosuperior (altitudinal) progression from rostral to caudal regions. Higher visual areas also exhibit characteristic patterns of retinotopy. These patterns were recapitulated by our model's readout positions: the x coordinate corresponded with azimuthal retinotopy, and y corresponded with altitude (Fig. 5c). In contrast to previous retinotopic methods, which either rely on special stimuli designed to elicit spatial tuning (Garrett et al., 2014; Zhuang et al., 2017) or anatomical coordinates to inform models (Bashiri et al., 2021), our model enables retinotopic characterization of the visual cortex solely from neuronal responses to naturalistic stimuli.

The visual cortex is functionally specialized, with different inter-connectivity and functional properties exhibited by lower and higher visual areas. As visual information propagates from lower to higher visual areas, the complexity of neuronal encoding increases (Siegle et al., 2021), leading to a more explicit representation of higher-level features like more linearly decodable information about objects (Froudarakis et al.; Goltstein et al., 2021). Since the model's readout feature weights encode differential tuning to visual features, we hypothesized that they would reveal functional differences between neurons in lower (V1) versus higher (LM, RL, AL) visual areas of the MICrONS volume. To test this hypothesis, we visualize the readout feature weights by performing nonlinear dimensionality reduction via uniform manifold approximation and projection (UMAP, McInnes et al. 2018). We observed that neurons were orga-

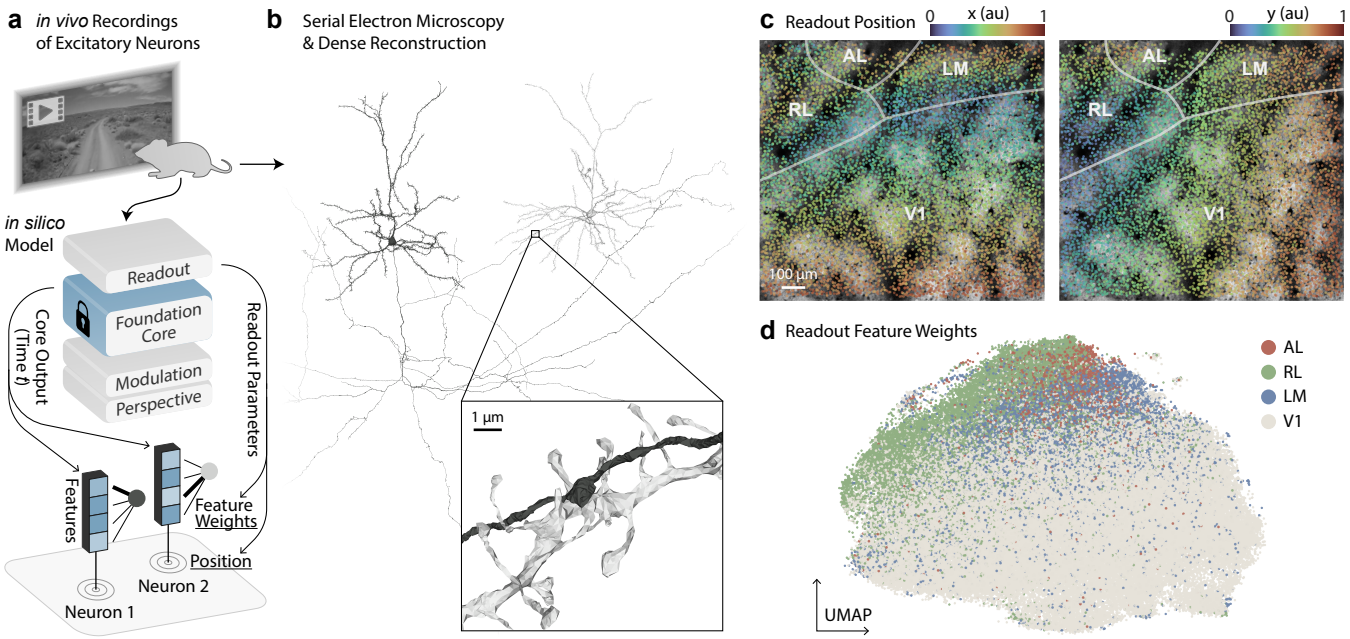


Fig. 5. The foundation model of the MICrONS volume reveals the functional organization of the visual cortex. a, Schematic of a foundation model of the MICrONS mouse, trained on excitatory neuronal responses to natural movies. At the bottom, the readout at a single time point is depicted, showing the readout positions and feature weights for two example neurons. **b,** Meshes of two example neurons, reconstructed from serial electron microscopy. The zoom-in cutout shows a synapse between these two neurons, with the pre-synaptic axon in black and post-synaptic dendrite in silver. **c,** Colored scatter plots of readout positions of all neurons from a recording session of the MICrONS mouse, overlayed on top-down a view of the recording window with annotated visual areas (V1, LM, RL, AL) and boundaries. The left and right plots are colored by the x and y coordinates of the readout positions, respectively. **d,** Visualization of the readout feature weights of all neurons in the MICrONS volume, projected onto a 2-dimensional embedding via UMAP.

nized according to their visual area, and V1 neurons occupied mostly different regions of the UMAP than higher visual areas (Fig. 5d). This suggests that properties underlying the functional organization of the visual cortex are inherently captured by the parameters of our foundation model. This makes it a valuable tool for studying computations across visual areas of the mouse cortex including analysis of the functional connectivity of the MICrONS volume.

Discussion

We introduce a major step towards a foundation model for the mouse visual cortex that achieves state-of-the-art performance at predicting dynamic neuronal responses across multiple visual areas. Beyond excelling in the natural movie domain on which it was trained, it accurately predicted responses to new stimulus domains, including coherent random moving dots, dynamic Gabor patches, flashing dots, directional pink noise, and natural static images. The model's generalization performance on new stimulus domains highlights its ability to capture non-linear transformations from image space to neuronal activity in the mouse visual cortex. The foundation core enabled accurate models of new mice to be fitted with limited training data, outperforming models with cores that were individually trained for each mouse.

Our work was inspired by recent breakthroughs in artificial intelligence, where foundation models (Bommasani et al., 2021), trained on massive data volumes, have demonstrated remarkable generalization in many downstream tasks. For example, training on next word prediction (Brown et al., 2020) can be transferred to downstream tasks—e.g., conversing naturally with humans, or passing professional licensing exams (Kung et al., 2023)—with relatively small amounts of new data. Applied to neuroscience, the foundation modeling paradigm overcomes a major limitation of previous common approaches where models are individually trained using data from a single experiment. The limited amount of data hinders the accuracy of models as they try to learn from scratch the complex non-linearities of the brain, even though there is a great deal of similarity in how visual neurons respond. By contrast, foundation models combine data from multiple experiments and subjects, giving them access to a much larger and richer set of data; only the specific idiosyncrasies of each individual mouse and its neurons must be learned separately. In other words, the similarities between neurons and subjects can be leveraged to identify common features of the brain, producing a more unified and accurate model of the brain that is informed by multiple subjects rather than one.

In neuroscience, previous work (Lurz et al., 2021) has shown that static models of the visual cortex benefit from pretraining on large amounts of data pooled from multiple subjects. In this study, we demonstrate that data pooling and transfer learning can extend to a more universal model of the visual cortex that predicts dynamic neuronal responses to moving stimuli for both lower and higher visual cortical areas. Importantly, our foundation models also predicted responses to new stimulus domains even without any fine tuning. Indeed, this accurate extrapolation underscores the potential of foundation models to study complex biological sys-

tems such as the brain.

Here, we used the foundation modeling approach to fit the valuable MICrONS dataset, enabling the exhaustive study of functional connectomics within a large volume of the mouse visual cortex. Building a digital twin for datasets of this type effectively "immortalizes" the functional properties of the recorded neurons in the study. If the foundation core can demonstrate generalization to novel arbitrary stimulus domains of interest, the digitally twinned neurons can be characterized using these new stimuli that were not presented at the time of the *in vivo* data collection. Naturally, conducting new animal validation experiments, similar to those presented in this study, will be necessary to confirm that the foundation core indeed generalizes to these newly introduced stimulus domains of interest. In large projects like MICrONS, where the longevity of a particular dataset is especially desirable, the good generalization performance of foundation models has clear benefits because we don't want the models' value to depend on previous strategic decisions about allocating experimental time to specific experimental questions.

This *in silico* representation also offers some unique additional advantages in the types of analyses that can be performed. For instance, the architecture of the foundation core allows each modeled neuron's predicted responses to be represented by a tuning function, which can be separated into two components: a spatial component (indicating the position of the neuron's receptive field) and a feature component (describing what the neuron responds to). This factorization of the tuning function was utilized by another study—derived from a version of our model—to analyze the relationship between the functional properties of neurons and synaptic connectivity (Ding et al., 2023a). The researchers discovered that the feature component, but not the spatial component, predicted which neurons were connected at a fine synaptic scale.

Our present foundation model merely scratches the surface, as it only models parts of the mouse visual system under passive viewing conditions. By expanding this approach to encompass complex, natural behaviors in freely-moving subjects, incorporating additional brain regions, diverse cell types, and creating foundation models for other species could be a paradigm shift in neuroscience. Foundation models can be employed to study vision, cognition and motor control during intricate, unconstrained natural behaviors in which identical conditions rarely occur twice. For instance, we can conduct comprehensive *in silico* experiments to explore relationships between the high dimensional neuronal activity and behavioral spaces to generate hypotheses and to design simpler experiments to run *in vivo*, such as inception loops (Walker et al., 2019; Franke et al., 2022).

Moreover, by considerably reducing the neuron-hours required to model new individuals and behaviors, foundational models facilitate more efficient and cost-effective neuroscience experiments. For instance, we can establish high-throughput research platforms that, with minimal new data, generate predictions of individual subjects' neuronal activity and behavior. When causal manipulations are incorporated in the foundation model, such as pharmacological interventions, we could then swiftly screen drugs tailored for a desired phenotypic neuronal or behavioral outcome. Ultimately, the development of multimodal foundational neuroscience models offers a powerful new approach to deciphering the algorithms underpinning natural intelligence.

ACKNOWLEDGEMENTS

The authors thank David Markowitz, the IARPA MICrONS Program Manager, who coordinated this work during all three phases of the MICrONS program. We thank IARPA program managers Jacob Vogelstein and David Markowitz for co-developing the MICrONS program. We thank Jennifer Wang, IARPA SETA for her assistance.

The work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DoI/IBC) contract numbers D16PC00003, D16PC00004, and D16PC0005. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. AT acknowledge support from NSF NeuroNex grant 1707400. AT acknowledges support from National Institute of Mental Health and National Institute of Neurological Disorders And Stroke under Award Number U19MH114830 and National Eye Institute/National Institutes of Health Core Grant for Vision Research T32-EY-002520-37. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

ASE received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101041669) as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 432680300 (SFB 1456, project B05)

We also would like to thank Matthias Bethge, Mackenzie Mathis, Blake Richards, Anthony Zador and Joel Zylberberg for many stimulating discussions regarding building foundation models for the brain.

Methods

Neurophysiological experiments MICrONS data in Fig. 5 was collected as described in MICrONS Consortium et al. (2021), and data in Fig. 2a was collected as described in Sinz et al. (2018). Data collection for all other figures is described below.

All procedures were approved by the Institutional Animal Care and Use Committee of Baylor College of Medicine. Fourteen mice (Mus musculus, 6 females, 8 males, age 2.2-4 months) expressing GCaMP6s in excitatory neurons via Slc17a7-Cre and Ai162 transgenic lines (recommended and generously shared by Hongkui Zeng at Allen Institute for Brain Science; JAX stock 023527 and 031562, respectively) were anesthetized and a 4 mm craniotomy was made over the visual cortex of the right hemisphere as described previously (Reimer et al., 2014; Froudarakis et al., 2014). Animals were allowed at least 5 days to recover before experimental scans.

Mice were head-mounted above a cylindrical treadmill and calcium imaging was performed using Chameleon Ti-Sapphire laser (Coherent) tuned to 920 nm and a large field of view mesoscope (Sofroniew et al., 2016) equipped with a custom objective (excitation NA 0.6, collection NA 1.0, 21 mm focal length). Laser power after the objective was increased exponentially as a function of depth from the surface according to: $P = P_0 \times e^{(z/L_z)}$, where P is the laser power used at target depth z, P0 is the power used at the surface (not exceeding 20 mW), and Lz is the depth constant (220 μm). The greatest laser output of 100 mW was used at approximately 420 μm from the surface.

The craniotomy window was leveled with regards to the ob-

jective with six degrees of freedom. Pixel-wise responses from an ROI spanning the cortical window (>2400 x 2400 μm, 2-5 μm/px, between 100-220 μm from surface, >2.47 Hz) to drifting bar stimuli were used to generate a sign map for delineating visual areas (Garrett et al., 2014). Area boundaries on the sign map were manually annotated.

For eleven out of fifteen scans (including four of the foundation cohort scans), our target imaging site was a 1200 × 1100 μm^2 area spanning L2-L5 at the conjunction of lateral primary visual cortex (V1) and three lateral higher visual areas: anterolateral (AL), lateromedial (LM), and rostrolateral (RL). This resulted in an imaging volume that was roughly 50% V1 and 50% higher visual area. This target was chosen in order to mimic the area membership and functional property distribution in the MICrONS animal (MICrONS Consortium et al., 2021) Each scan was performed at 6.3 Hz, collecting eight 620 × 1100 μm^2 fields per frame at 2.5 μm/px xy resolution to tile a 1200-1220 × 1100 μm^2 FOV at four depths (two planes per depth, 20-40 μm overlap between coplanar fields. The four imaging planes were distributed across layers with at least 45 μm spacing, with two planes in L2/3 (depths: 170-200 μm and 215-250 μm), one in L4 (300-325 μm), and one in L5 (390-420 μm).

For the remaining 4 foundation cohort scans, our target imaging site was a single plane in L2/3 (depths 210-220 μm), spanning all visual cortex visible in the cortical window (typically including V1, LM, AL, RL, PM, and AM). Each scan was performed at 6.8-6.9 Hz, collecting four 630 μm width adjacent fields (spanning 2430 μm ROI, with 90 μm total overlap). Each field was a custom height (2010-3000 μm) in order to encapsulate visual cortex within that field. Imaging was performed at 3 μm/px.

Movie of the animal's eye and face was captured throughout the experiment. A hot mirror (Thorlabs FM02) positioned between the animal's left eye and the stimulus monitor was used to reflect an IR image onto a camera (Genie Nano C1920M, Teledyne Dalsa) without obscuring the visual stimulus. The position of the mirror and camera were manually calibrated per session and focused on the pupil. Field of view was manually cropped for each session. The field of view contained the left eye in its entirety, and was captured at ∼20 Hz. Frame times were time stamped in the behavioral clock for alignment to the stimulus and scan frame times. Video was compressed using Labview's MJPEG codec with quality constant of 600 and stored the frames in AVI file.

Light diffusing from the laser during scanning through the pupil was used to capture pupil diameter and eye movements. A DeepLabCut model (Mathis et al., 2018) was trained on 17 manually labeled samples from 11 animals to label each frame of the compressed eye video (intraframe only H.264 compression, CRF:17) with 8 eyelid points and 8 pupil points at cardinal and intercardinal positions. Pupil points with likelihood >0.9 (all 8 in 72-99% of frames per scan) were fit with the smallest enclosing circle, and the radius and center of this circle was extracted. Frames with < 3 pupil points with likelihood >0.9 (<1.2% frames per scan), or producing a circle fit with outlier > 5.5 standard deviations from the mean in any of the three parameters (center x, center y, radius, <0.2% frames per scan) were discarded (total <1.2% frames per scan). Gaps of <= 10 discarded frames were replaced by linear interpolation. Trials affected by remaining gaps were discarded (<18 trials per scan, <0.015%).

The mouse was head-restrained during imaging but could walk on a treadmill. Rostro-caudal treadmill movement was measured using a rotary optical encoder (Accu-Coder 15T-01SF-2000NV1ROC-F03-S1) with a resolution of 8000 pulses per revolution, and was recorded at ∼100 Hz in order to extract locomotion velocity.

Monitor Positioning and Calibration Visual stimuli were presented with Psychtoolbox in MATLAB to the left eye with a 31.0 x 55.2 cm (height x width) monitor (ASUS PB258Q) with a resolution of 1080 x 1920 pixels positioned 15 cm away from the eye. When the monitor is centered on and perpendicular to the surface of the eye at the closest point, this corresponds to a visual angle of 3.8 °/cm at the nearest point and 0.7 °/cm at the most remote corner of the monitor. As the craniotomy coverslip placement during surgery and the resulting mouse positioning relative to the objective is optimized for imaging quality and stability, uncontrolled variance in animal skull position relative to the washer used for head-mounting was compensated with tailored monitor positioning on a six dimensional monitor arm. The pitch of the monitor was kept in the vertical position for all animals, while the roll was visually matched to the roll of the animal's head beneath the headbar by the experimenter. In order to optimize the translational monitor position for centered visual cortex stimulation with respect to the imaging field of view, we used a dot stimulus with a bright background (maximum pixel intensity) and a single dark square dot (minimum pixel intensity). Randomly ordered dot locations drawn from either a 5 x 8 grid tiling the screen (20 repeats) or a 10 x 10 grid tiling a central square (approx 90 degrees width and height, 10 repeats), with each dot presentation lasting 200 ms. For five scans (four foundational cohort scans, 1 scan from Fig. 4), this dot-mapping scan targeted the V1/RL/AL/LM conjunction, and the final monitor position for each animal was chosen in order to maximize inclusion of the population receptive field peak response in cortical locations spanning the scan FOV. In the remaining scans, the procedure was the same, but the scan FOV spanned all of V1 and some adjacent higher visual areas, and thus the final monitor position for each animal was chosen in order to maximize inclusion of the population receptive field peak response in cortical locations corresponding to the extremes of the retinotopic map. In both cases, the yaw of the monitor visually matched to be perpendicular to and 15 cm from the nearest surface of the eye at that position.

A photodiode (TAOS TSL253) was sealed to the top left corner of the monitor, and the voltage was recorded at 10 KHz and timestamped with a 10 MHz behavior clock. Simultaneous measurement with a luminance meter (LS-100 Konica Minolta) perpendicular to and targeting the center of the monitor was used to generate a lookup table for linear inter-

polation between photodiode voltage and monitor luminance in cd/m^2 for 16 equidistant values from 0-255, and one baseline value with the monitor unpowered.

At the beginning of each experimental session, we collected photodiode voltage for 52 full-screen pixel values from 0 to 255 for one second trials. The mean photodiode voltage for each trial was collected with an 800 ms boxcar window with 200 ms offset. The voltage was converted to luminance using previously measured relationship between photodiode voltage and luminance and the resulting luminance vs. voltage curve was fit with the function $L = B + A \cdot P^{\gamma}$ where L is the measured luminance for pixel value P, and the median γ of the monitor was fit as 1.73 (range 1.58 - 1.74). All stimuli were shown without linearizing the monitor (i.e. with monitor in normal gamma mode).

During the stimulus presentation, display frame sequence information was encoded in a 3 level signal, derived from the photodiode, according to the binary encoding of the display frame (flip) number assigned in-order. This signal underwent a sine convolution, allowing for local peak detection to recover the binary signal together with its behavioral time stamps. The encoded binary signal was reconstructed for $>$96% of the flips. Each flip was time stamped by a stimulus clock (MasterClock PCIe-OSC-HSO-2 card). A linear fit was applied to the flip timestamps in the behavioral and stimulus clocks, and the parameters of that fit were used to align stimulus display frames with scanner and camera frames. The mean photodiode voltage of the sequence encoding signal at pixel values 0 and 255 was used to estimate the luminance range of the monitor during the stimulus, with minimum values of approximately 0.005 - 1 cd/m^2 and maximum values of approximately 8.0 - 11.5 cd/m^2.

Stimulus Composition Dynamic stimuli libraries of natural movies, global directional pink noise ("Monet"), and local directional parametric stimuli ("Trippy"), was as described in MICrONS Consortium et al. (2021). Static natural image library was as described in Walker et al. (2019).

Dynamic Gabor filters were generated as described in Petkov and Subramanian (2007). We used a spatial envelope that had a standard deviation of $\sim$10° in the center of the monitor. A 10-second trial consisted of 10 Gabor filters (each lasting 1 second) with randomly sampled spatial positions, directions of motion, phases, spatial and temporal frequencies.

Random dot kinematograms were generated as described in Morrone et al. (2000). The diameter of the dots was $\sim$2° in the center of the monitor. Each 10-second trial contained 5 patterns of optical flow, each lasting 2 seconds. The patterns were randomly sampled in terms of type of optical flow (translation: up/down/right/left, radial: in/out, rotation: clockwise/anticlockwise), and coherence of random dots (50%, 100%).

Four of the foundational cohort scans (V1,LM,AL,RL) contained 192 minutes of trials, randomly intermixed, as follows:

- **Oracle Natural Movies:** 6 natural movie clips, 2 from each category. 10 seconds each, 20 repeats per scan, 20 minutes total. Conserved across all 4 scans.
- **Unique Natural Movies:** 144 natural movies, 48 from each category. 10 seconds each, 1 repeat per scan, 24 minutes total. Conserved across all 4 scans.
- **2x Repeat Natural Movies:** 144 natural movies, 48 from each category. 10 seconds each, 2 repeat per scan, 48 minutes total. Conserved across all 4 scans.
- **4x Repeat Natural Movies:** 90 natural movies, 30 from each category. 10 seconds each, 4 repeats, 60 minutes total. Conserved across all 4 scans.
- **Local Directional Parametric Stimulus ("Trippy"):** 20 seeds, 15 seconds each, 4 repeats, 20 minutes total. Conserved across all 4 scans.
- **Global Directional Parametric Stimulus ("Monet"):** 20 seeds, 15 seconds each, 4 repeats, 20 minutes total. Conserved across all 4 scans.

The remaining four foundational cohort scans (V1,LM,AL,RL,AM,PM) contained 111-116 minutes of trials, randomly intermixed, as follows:

- **Oracle Natural Movies:** 6 natural movie clips, 2 from each category. 10 seconds each, 15-20 repeats per scan, 15-20 minutes total. Conserved across 4 all scans.
- **Unique Natural Movies:** 576 natural movies, 192 from each category. 10 seconds each, 1 repeat per scan, 96 minutes total. Unique to each scan.

The four scans in Fig. 3 contained 141 minutes of trials, randomly intermixed, as follows:

- **Oracle Natural Movies:** 6 natural movie clips, 2 from each category. 10 seconds each, 10 repeats per scan, 10 minutes total. Conserved across all 4 scans.
- **Unique Natural Movies:** 480 natural movies, approx 160 from each category. 10 seconds each, 1 repeat per scan, 80 minutes total. Conserved across all 4 scans.
- **Global Directional Parametric Stimulus ("Monet"):** 6 seeds, 10.8 seconds each, 10 repeats, 10.8 minutes total. Conserved across all 4 scans.
- **Gabor Movies:** 6 sequences, 10 seconds each, 10 repeats, 10 minutes total. Conserved across all 4 scans.
- **Random Dot Kinematogram Movies:** 6 sequences, 10 seconds each, 10 repeats, 10 minutes total. Conserved across all 4 scans.
- **Static Natural Image:** 40 images, 1.5 seconds each, 10 repeats, 10 minutes total. Conserved across all 4 scans.
- **Gaussian Dot Images:** 6 sequences, 10.2 seconds each, 10 repeats, 10.2 minutes total. Conserved across all 4 scans.

The four scans in Fig. 4 contained 130-150 minutes of trials, randomly intermixed, as follows:

- **Oracle Natural Movies:** 6 natural movie clips, 2 from each category. 10 seconds each, 10-30 repeats per scan, 10-30 minutes total. Conserved across all 3 scans.

- **Unique Natural Movies:** 360 natural movies, approx 120 from each category. 10 seconds each, 1 repeat per scan, 60 minutes total. Conserved across all 3 scans.
- **Global Directional Parametric Stimulus ("Monet"):** 120 seeds, 15 seconds each, 1 repeat, 30 minutes total. Conserved across all 3 scans.
- **Gaussian Dot Images:** 120 sequences, 15 seconds each, 1 repeat, 30 minutes total. Conserved across all 3 scans.

Neural network architecture Our model of the visual cortex is an artificial neural network composed of four components: perspective, behavior, core, and readout. These components are described in detail in the following sections.

Perspective network The perspective network uses ray tracing to infer the perspective or retinal activation of a mouse at discrete time points from two input variables: stimulus (movie frame) and eye position (estimated center of pupil, extracted from the eye tracking camera). To perform ray tracing, we modeled the following physical entities: 1) topography and light ray trajectories of the retina; 2) rotation of the retina; 3) position of the monitor relative to the retina; 4) intersection of the light rays of the retina and the monitor.

1) The topography of the retina was modeled as a uniform 2D grid consisting of 108×72 (width by height) points. Let i denote the horizontal displacement from the center of the grid, and let j denote the veritcal displacement. The trajectory of light for point ij is defined by the vector $\mathbf{l}_{ij} \in \mathbb{R}^3$:

$$\theta_{ij} = \sqrt{i^2 + j^2} \ ,$$
$$z = \frac{\theta_{ij}}{\tan \theta_{ij}} \ ,$$
$$\mathbf{l}_{ij} = \begin{bmatrix} i \\ j \\ z \end{bmatrix} \ ,$$

where θ_{ij} is the visual angle (radians) of point ij of the retina.

2) We used pupil tracking data to infer the rotation of the occular globe and the retina. At each time point t, a multilayer perceptron (MLP with 3 layers and 8 hidden units per layer) was used to map the pupil position onto the 3 ocular angles of rotation:

$$\text{MLP} : \begin{bmatrix} p_{xt} \\ p_{yt} \end{bmatrix} \mapsto \begin{bmatrix} \widehat{\theta}_{xt} \\ \widehat{\theta}_{yt} \\ \widehat{\theta}_{zt} \end{bmatrix} \ .$$

The vector $[p_{xt}, p_{yt}]$ is the center of the pupil in the frame of the tracking camera at time t. $\widehat{\theta}_{xt}, \widehat{\theta}_{yt}, \widehat{\theta}_{zt}$ are the estimated angles of rotation of about the x (adduction/abduction), y (elevation/depression), z (intorsion/extorsion) axes of the occular globe at time t. Let $R_x, R_y, R_z \in \mathbb{R}^{3\times 3}$ denote rotation matrices about x, y, z axes. Light rays for each point $\mathbf{l}_{ij}$ of the retinal grid were rotated according to:

$$\widehat{\mathbf{l}}_{ijt} = R_z(\widehat{\theta}_{zt}) R_y(\widehat{\theta}_{yt}) R_x(\widehat{\theta}_{xj}) \mathbf{l}_{ij} \ ,$$

where $\widehat{\mathbf{l}}_{ijt} \in \mathbb{R}^3$ represents the ray of light—accounting for occular globe rotation—that hits the retina at point ij and time t.

3) The monitor was modeled as a plane with 6 degrees of freedom: 3 for translation and 3 for rotation. Translation of the monitor plane relative to the retina was parameterized by $\mathbf{m}_0 \in \mathbb{R}^3$. Rotation was parameterized by angles $\bar{\theta}_x, \bar{\theta}_y, \bar{\theta}_z$:

$$\begin{bmatrix} \mathbf{m}_x & \mathbf{m}_y & \mathbf{m}_z \end{bmatrix}^\intercal = R_z(\bar{\theta}_z) R_y(\bar{\theta}_y) R_x(\bar{\theta}_x) \ ,$$

where $\mathbf{m}_x, \mathbf{m}_y, \mathbf{m}_z \in \mathbb{R}^3$ are the horizontal, vertical, and normal unit vectors of the monitor, respectively.

4) We computed the line-plane intersection between the monitor plane and $\widehat{\mathbf{l}}_{ijt}$, the gaze-corrected trajectory of light for point ij of the retina at time t:

$$\mathbf{m}_{ijt} = \frac{\mathbf{m}_0 \cdot \mathbf{m}_z}{\widehat{\mathbf{s}}_{ijt} \cdot \mathbf{m}_z} \widehat{\mathbf{l}}_{ijt},$$

where $\mathbf{m}_{ijt}$ is the point of intersection between the monitor plane and the light ray $\widehat{\mathbf{l}}_{ijt}$. This was projected onto the monitor's horizontal and vertical unit vectors:

$$m^x_{ijt} = (\mathbf{m}_{ijt} - \mathbf{m}_0) \cdot \mathbf{m}_x \ ,$$
$$m^y_{ijt} = (\mathbf{m}_{ijt} - \mathbf{m}_0) \cdot \mathbf{m}_y \ ,$$

yielding m^x_{ijt}, m^y_{ijt}, the horizontal and vertical displacements from the center of the monitor/stimulus. To produce inferred activation of the retina at point ij at time t, we performed bilinear interpolation of the stimulus at the four pixels surrounding the line-plane intersection at m^x_{ijt}, m^y_{ijt}.

Modulation network The modulation network is a small LSTM network with 6 hidden units. It transforms behavioral variables, i.e., locomotion and pupil size, and previous states of the network, to produce dynamic representations of the behavioral state and arousal of the mouse.

$$\text{LSTM} : \begin{bmatrix} r_t \\ p_t \\ p'_t \end{bmatrix}, \mathbf{h}^M_{t-1}, \mathbf{c}^M_{t-1} \mapsto \mathbf{h}^M_t, \mathbf{c}^M_t \ ,$$

where r_t is the running/treadmill speed, p_t is the pupil diameter, p'_t is the instantaneous change in pupil diameter, and $\mathbf{h}^M_t, \mathbf{c}^M_t \in \mathbb{R}^6$ are the "hidden" and "cell" state vectors of the modulation LSTM network at time t. The hidden state vector formed the output of the modulation network, and its 6 dimensions were expanded spatially to produce 6 uniform feature maps:

$$\mathbf{h}^M_t \in \mathbb{R}^c \to \mathbf{H}^M_t \in \mathbb{R}^{c \times h \times w} \ ,$$

where $c = 6$ is the number of feature channels, and $h = 12$ and $w = 18$ are the height and width of the feature maps, respectively. The modulation feature maps, $\mathbf{H}^M_t$, become the modulatory inputs to the recurrent portion of the core.

Core network The core network transforms inputs from the perspective and modulation networks to produce feature representations of vision modulated by behavior. This occurs in

two stages: 1) a feedforward network transforms the output of the perspective network to produce an intermediate feature representation, and 2) a recurrent network transforms this intermediate representation, along with modulation inputs and previous hidden states, to produce the output of the core network.

1) The feedforward network is a 3-layer 3D convolutional neural network transforms the visual input provided by the perspective network. Using $k^{t\times h\times w}, s^{t\times h\times w}, \mathbb{R}^{c\times t\times h\times w}$ to denote the spatiotemporal dimensions of the convolutional kernel (k), stride (s), and visual input/feature maps ($\mathbb{R}$), the following describes the shape of the layers L:

$$
\begin{aligned}
&L1, k^{3\times3\times3}, s^{1\times3\times3} : \mathbb{R}^{1\times t\times 72\times 108} \mapsto \mathbb{R}^{32\times t\times 24\times 36},\\
&L2, k^{3\times3\times3}, s^{1\times1\times1} : \mathbb{R}^{32\times t\times 24\times 36} \mapsto \mathbb{R}^{64\times t\times 24\times 36},\\
&L3, k^{2\times2\times2}, s^{1\times2\times2} : \mathbb{R}^{64\times t\times 24\times 36} \mapsto \mathbb{R}^{128\times t\times 12\times 18},
\end{aligned}
$$

where t is the number of time points/movie frames. The convolutional output of each layer was transformed with a GeLU nonlinearity (Hendrycks and Gimpel, 2020) with a residual connection (He et al., 2015). To enforce causality, the 3D convolutions were shifted along the temporal dimension such that no inputs from future time points contributed to any outputs.

2) The recurrent network produces dynamic representations of visual and behavioral information from the perspective/feedforward and modulation networks, respectively. It consists of 4 "cells", or cannonical layers that perform the same nonlinear operation but with different inputs and parameters. The cells are interconnected, and each cell receives inputs from every other cell, in addition to the feedforward and modulation networks. Let $*$ denote a 2D spatial convolution. The following describes the input to a cell C at time t:

$$
\mathbf{X}_t^C = \mathbf{F}_t^C + \mathbf{W}^{CM} * \mathbf{H}_t^M + \sum_{C'} \mathbf{W}^{CC'} * \mathbf{H}_{t-1}^{C'} ,
$$

where $\mathbf{X}_t^C \in \mathbb{R}^{32\times12\times18}$ (channel, height, width) is the input to cell C at time t. $\mathbf{F}_t^C \in \mathbb{R}^{32\times12\times18}$ is a portion of the feedforward network's output at time t: 32 out of the 128 feature maps. The other 96 feature maps serve as inputs for the other 3 cells $C' \neq C$. From those cells, $\mathbf{H}_{t-1}^{C'} \in \mathbb{R}^{32\times12\times18}$ is the hidden state at time $t-1$. Finally, $\mathbf{H}_t^M \in \mathbb{R}^{6\times12\times18}$ is the output of the modulation network at time t. For the 2D convolutions, we used a kernel size of $k^{1\times1}$.

For each cell C, the input $\mathbf{X}_t^C$ was transformed with a convolutional LSTM (Conv-LSTM, SHI et al. 2015):

$$
\text{Conv-LSTM}^C : \mathbf{X}_t^C, \mathbf{H}_{t-1}^C, \mathbf{C}_{t-1}^C \mapsto \mathbf{H}_t^C, \mathbf{C}_t^C ,
$$

where $\mathbf{H}_{t-1}^C, \mathbf{H}_t^C \in \mathbb{R}^{32\times12\times18}$ are "hidden" states (previous and currrent), and $\mathbf{C}_{t-1}^C, \mathbf{C}_t^C \in \mathbb{R}^{32\times12\times18}$ are "cell" states of the LSTM network. We used a kernel size of $k^{3\times3}$ for all convolutions of the Conv-LSTM.

For each time point t, the output of each cell was $\mathbf{H}_t^C \in \mathbb{R}^{32\times12\times18}$ (32 feature maps). The outputs from the 4 cells were concatenated, and then a $k^{1\times1}$ convolution was used to reduce the number of concatenated feature maps by half. This yielded $\mathbf{R}_t \in \mathbb{R}^{64\times12\times18}$, the combined output of the recurrent network at time t (4 cells $\times$ 32 / 2 = 64 feature maps).

We performed multi-head recurrent processing, where multiple recurrent networks were used to transform the feedforward and modulation inputs in parallel. We used 4 recurrent heads, and their outputs were concatenated to produce $\mathbf{C}_t \in \mathbb{R}^{256\times12\times18}$, the output of the core network at time t (4 heads $\times$ 64 = 256 feature maps).

Readout network The readout network maps the core's outputs onto the activity of individual neurons. For each neuron, the readout parameters are factorized into two components: spatial position and feature weights. Let $\mathbf{p}^n \in \mathbb{R}^2$ denote the spatial position (x, y), and let $\mathbf{w}^n \in \mathbb{R}^{256}$ denote the feature weights for a neuron n. To produce the response of that neuron at time t, the following readout operation is performed:

$$
\begin{aligned}
\mathbf{c}_t^n &= \text{Interpolate}(\mathbf{C}_t, \mathbf{p}^n) ,\\
r_t^n &= \exp(\mathbf{c}_t^n \cdot \mathbf{w}^n + b^n) ,
\end{aligned}
$$

where $\mathbf{c}_t^n \in \mathbb{R}^{256}$ is a feature vector that is produced via bilinear interpolation of the core network's output feature maps $\mathbf{C}_t \in \mathbb{R}^{256\times12\times18}$ at time t, interpolated at the spatial position $\mathbf{p}^n$. The feature vector $\mathbf{c}_t^n$ is then combined with the feature weights $\mathbf{w}^n$ and a scalar bias b^n to produce the response r_t^n of neuron n at time t.

Due to the bilinear interpolation at a single position, each neuron only reads out from the core's output feature maps within a 2×2 spatial window. While this adheres to the functional property of spatial selectivity exhibited by neurons in the visual cortex, the narrow window limits exploration of the full spatial extent of features during model training. To facilitate the spatial exploration of the core's feature maps during training, for each neuron n, we sampled the readout position from a 2D Gaussian distribution: $\mathbf{p}^n \sim \mathcal{N}(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n)$. The parameters of the distribution $\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n$ (mean, covariance) were learned via the reparameterization trick (Kingma and Welling, 2013). We observed empirically that the covariance $\boldsymbol{\Sigma}^n$ naturally decreased to small values by the end of training, meaning that the readout converged on a specific spatial position. After training, and for all testing purposes, we used the mean of the learned distribution $\boldsymbol{\mu}^n$ as the single readout position $\mathbf{p}^n$ for neuron n.

Model training The perspective, behavior, core, and readout networks were assembled together to form an ANN that was trained to match the recorded dynamic neuronal responses from the training dataset. Let y_t^i be the recorded *in vivo* response, and let r_t^i be the predicted *in silico* response of neuron i at time t. The ANN was trained to minimize the Poisson negative log likelihood loss, $\sum_{it} r_t^i - y_t^i \log(r_t^i)$, via stochastic gradient descent with Nesterov momentum (Sutskever et al., 2013). The ANN was trained for 200 epochs with a learning rate schedule that consisted of a linear warm-up in the first 10 epochs, cosine decay (Loshchilov and Hut-

ter, 2016) for 90 epochs, followed by a warm restart and cosine decay for the remaining 100 epochs. Each epoch consisted of 512 training iterations / gradient descent steps. We used a batch size of 5, and each sample of the batch consisted of 70 frames (2.33 seconds) of stimulus, response, and behavioral data.

Model testing We generated model predictions of responses to stimuli that were included in the experimental recordings but excluded from model training. To evaluate the accuracy of model predictions, for each neuron we computed the correlation between the mean *in silico* and *in vivo* responses, averaged over stimulus repeats. The average *in vivo* response aims to estimate the true expected response of the neuron. However, when the *in vivo* response is highly variable and there are a limited number of repeats, this estimate becomes noisy. To account for this, we normalized the correlation by an upper bound proposed by Schoppe et al. (2016). Using $\overline{\cdot}$ to denote average over trials/stimulus repeats, the normalized correlation CC_{norm} is defined as follows:

$$CC_{\text{norm}} = \frac{CC_{\text{abs}}}{CC_{\text{max}}} \,,$$
$$CC_{\text{abs}} = \frac{\text{Cov}(\overline{r}, \overline{y})}{\sqrt{\text{Var}(\overline{r})\text{Var}(\overline{y})}} \,,$$
$$CC_{\text{max}} = \sqrt{\frac{N\text{Var}(\overline{y}) - \overline{\text{Var}(y)}}{(N-1)\text{Var}(\overline{y})}} \,,$$

where r is the *in silico* response, y is the *in vivo* response, and N is the number of trials. CC_{abs} is the Pearson correlation coefficient between the average *in silico* and *in vivo* responses. CC_{max} is the upper bound of achievable performance given the the *in vivo* variability of the neuron and the number of trials.

Parametric tuning To estimate parametric tuning, we presented parametric stimuli to the mice and the models. Specifically, we used directional pink noise parameterized by direction/orientation and flashing Gaussian blobs parameterized by spatial location. Orientation, direction, and spatial tuning were computed from the recorded responses from the mice and the predicted responses from the models. This resulted in analogous *in vivo* and *in silico* estimates of parametric tuning for each neuron. The methods for measuring the tuning to orientation, direction, and spatial location are explained in the following sections.

Orientation and Direction tuning We presented 16 angles of directional pink noise, uniformly distributed between $[0, 2\pi)$. Let $\overline{r}_\theta$ be the mean response of a neuron to the angle θ, averaged over repeated presentations of the angle. The orientation and direction selectivity indices (OSI and DSI) were computed as

$$\text{OSI} = \frac{|\sum_\theta \overline{r}_\theta \, e^{i2\theta}|}{\sum_\theta \overline{r}_\theta} \,,$$
$$\text{DSI} = \frac{|\sum_\theta \overline{r}_\theta \, e^{i\theta}|}{\sum_\theta \overline{r}_\theta} \,,$$

i.e., the normalized magnitude of the first and second Fourier components.

To determine the parameters for orientation and direction tuning, we used the following parametric model:

$$f(\theta \,|\, \mu, \kappa, \alpha, \beta, \gamma) = \alpha e^{\kappa \cos(\theta - \mu)} + \beta e^{\kappa \cos(\theta - \mu + \pi)} + \gamma \,,$$

which is a mixture of two von Mises functions with amplitudes α and β, preferred directions μ and $\mu + \pi$, and dispersion κ, plus a baseline offset of γ. The preferred orientation is the angle that is orthogonal to μ between $[0, \pi]$, i.e., $(\mu + \pi/2) \bmod \pi$. To estimate the parameters $\mu, \kappa\, \alpha, \beta, \gamma$ that best fit the neuronal response, we performed least squares optimization, minimizing $\sum_\theta (f(\theta \,|\, \mu, \kappa\, \alpha, \beta, \gamma) - r_\theta)^2$.

Parameters were estimated via least square optimization for both the *in vivo* and *in silico* responses. Let $\widehat{\mu}, \overline{\mu}$ be the angles of preferred directions estimated from *in vivo*, *in silico* responses, respectively. The angular distances between the *in vivo* and *in silico* estimates of preferred direction (Fig. 4g) and orientation (Fig. 4e) were computed as follows:

$$\Delta\text{Direction} = \arccos\left(\cos\left(\widehat{\mu} - \overline{\mu}\right)\right) \,,$$
$$\Delta\text{Orientation} = \arccos\left(\cos\left(2\widehat{\mu} - 2\overline{\mu}\right)\right)/2 \,.$$

Spatial tuning To measure spatial tuning, we presented "on" and "off" (white and black), flashing (300 ms) Gaussian dots. The dots were isotropically shaped, with a standard deviation of approximately 8 visual degrees in the center of the monitor. The position of each dot was randomly sampled from a 17×29 grid tiling the height and width monitor. We observed a stronger neuronal response for "off" compared to "on", and therefore we used only the "off" Gaussian dots to perform spatial tuning from the *in vivo* and *in silico* responses.

To measure spatial tuning, we first computed the spike triggered average (STA) of the stimulus. Let $\mathbf{x} \in \mathbb{R}^2$ denote the spatial location (height and width) in pixels. The value of the STA at location $\mathbf{x}$ was computed as follows:

$$\overline{s}_{\mathbf{x}} = \frac{\sum_t |s_{\mathbf{x}t} - s_0| r_t}{\sum_t r_t} \,,$$

where r_t is the response of the neuron, $s_{\mathbf{x}t}$ is the value of the stimulus at location $\mathbf{x}$ and time t, and s_0 is the blank or gray value of the monitor.

To measure the spatial selectivity of a neuron, we computed the covariance matrix or dispersion of the STA. Again using $\mathbf{x} \in \mathbb{R}^2$ denote the spatial location (height and width) in pix-

els:

$$z = \sum_{\mathbf{x}} \overline{s}_{\mathbf{x}} \ ,$$
$$\overline{\mathbf{x}} = \sum_{\mathbf{x}} \overline{s}_{\mathbf{x}} \mathbf{x} \,/\, z \ ,$$
$$\Sigma_{\mathrm{STA}} = \sum_{\mathbf{x}} \overline{s}_{\mathbf{x}} (\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\intercal} \,/\, z \ .$$

The spatial selectivity index, or strength of spatial tuning, was defined as the negative log determinant of the covariance matrix:

$$\mathrm{SSI} = -\log |\Sigma_{\mathrm{STA}}| \ .$$

To determine the parameters of spatial tuning, we used least squares to fit the STA to the following parametric model:

$$f(\mathbf{x} \,|\, \boldsymbol{\mu}, \Sigma, \alpha, \gamma) = \alpha \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\intercal} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) + \gamma \ ,$$

which is a 2D Gaussian component with amplitude α, mean $\boldsymbol{\mu}$, and covariance Σ, plus a baseline offset of γ.

From the *in vivo* and *in silico* responses, we estimated two sets of spatial tuning parameters. Let $\widehat{\boldsymbol{\mu}}, \overline{\boldsymbol{\mu}}$ be the means (preferred spatial locations) estimated from *in vivo* and *in silico* responses. To measure the difference between the preferred locations (Fig. 4i), we computed the Euclidean distance:

$$\Delta\mathrm{Location} = \|\widehat{\boldsymbol{\mu}} - \overline{\boldsymbol{\mu}}\| \ .$$

Data availability. All MICrONS data have already been released on BossDB (https://bossdb.org/project/microns-minnie, please also see https://www.microns-explorer.org/cortical-mm3 for details). Additional data including foundation model architecture and learned weights will be released upon publication.

Code availability. All code will be released on github upon publication.

Bibliography

E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.*, 2(2):284–299, Feb. 1985.

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.*, pages 1–22, 2016.

S. Bakhtiari, P. Mineault, T. Lillicrap, C. Pack, and B. Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34:25164–25178, 2021.

M. Bashiri, E. Walker, K.-K. Lurz, A. Jagadish, T. Muhammad, Z. Ding, Z. Ding, A. Tolias, and F. Sinz. A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, 34, 2021.

P. Bashivan, K. Kar, and J. J. DiCarlo. Neural population control via deep image synthesis. *Science (New York, N.Y.)*, 364(6439), 2019. ISSN 1095-9203. doi: 10.1126/science.aav9436.

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. In *Proceedings of the International Conference for Learning Representations (ICLR)*, 2017.

R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12): 4745–4765, 1992.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

M. F. Burg, S. A. Cadena, G. H. Denfield, E. Y. Walker, A. S. Tolias, M. Bethge, and A. S. Ecker. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6): e1009028, July 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009028.

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 Computational Biology*, 15(4):e1006897, Apr. 2019. doi: 10.1371/journal.pcbi.1006897.

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

E. Christensen and J. Zylberberg. Models of primate ventral stream that categorize and visualize images. *bioRxiv*, pages 2020–02, 2020.

B. Cowley and J. Pillow. High-contrast "gaudy" images improve the training of deep neural network models of visual cortex. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33*, pages 21591–21603. Curran Associates, Inc., 2020.

Z. Ding, P. G. Fahey, S. Papadopoulos, E. Wang, B. Celii, C. Papadopoulos, A. Kunin, A. Chang, J. Fu, Z. Ding, S. Patel, K. Ponder, J. Alexander Bae, A. L. Bodor, D. Brittain, J. Buchanan, D. J. Bumbarger, M. A. Castro, E. Cobos, S. Dorkenwald, L. Elabbady, A. Halageri, Z. Jia, C. Jordan, D. Kapner, N. Kemnitz, S. Kinn, K. Lee, K. Li, R. Lu, T. Macrina, G. Mahalingam, E. Mitchell, S. S. Mondal, S. Mu, B. Nehoran, S. Popovych, C. M. Schneider-Mizell, W. Silversmith, M. Takeno, R. Torres, N. L. Turner, W. Wong, J. Wu, W. Yin, S.-C. Yu, E. Froudarakis, F. H. Sinz, H. Sebastian Seung, F. Collman, N. M. da Costa, R. Clay Reid, E. Y. Walker, X. Pitkow, J. Reimer, and A. S. Tolias. Functional connectomics reveals general wiring rule in mouse visual cortex. Mar. 2023a.

Z. Ding, D. T. Tran, K. Ponder, E. Cobos, Z. Ding, P. G. Fahey, E. Y. Wang, T. Muhammad, J. Fu, S. A. Cadena, S. Papadopoulos, S. Patel, K. Franke, J. Reimer, F. H. Sinz, A. S. Ecker, X. Pitkow, and A. S. Tolias. Bipartite invariance in mouse primary visual cortex. Submitted to bioRxiv, 2023b.

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, 2018. URL https://arxiv.org/abs/1809.10504.

K. Franke, K. F. Willeke, K. Ponder, M. Galdamez, N. Zhou, T. Muhammad, S. Patel, E. Froudarakis, J. Reimer, F. H. Sinz, and A. S. Tolias. State-dependent pupil dilation rapidly shifts visual feature selectivity. 610(7930):128–134, 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-05270-3.

E. Froudarakis, U. Cohen, M. Diamantaki, E. Y. Walker, J. Reimer, P. Berens, H. Sompolinsky, and A. S. Tolias. Object manifold geometry across the mouse cortical visual hierarchy. URL http://biorxiv.org/lookup/doi/10.1101/2020.08.20.258798.

E. Froudarakis, P. Berens, A. S. Ecker, R. J. Cotton, F. H. Sinz, D. Yatsenko, P. Saggau, M. Bethge, and A. S. Tolias. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nat. Neurosci.*, 17(6):851–857, June 2014.

J. Fu, S. Shrinivasan, K. Ponder, T. Muhammad, Z. Ding, E. Wang, Z. Ding, D. T. Tran, P. G. Fahey, S. Papadopoulos, et al. Pattern completion and disruption characterize contextual modulation in mouse visual cortex. *bioRxiv*, pages 2023–03, 2023.

M. E. Garrett, I. Nauhaus, J. H. Marshel, and E. M. Callaway. Topography and areal organization of mouse visual cortex. *J. Neurosci.*, 34(37):12587–12600, Sept. 2014.

P. M. Goltstein, S. Reinert, T. Bonhoeffer, and M. Hübener. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. 2021. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-00914-5.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL http://arxiv.org/abs/1512.03385.

D. J. Heeger. Half-squaring in responses of cat striate cells. *Vis. Neurosci.*, 9(5):427–443, 1992a.

D. J. Heeger. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.*, 9(2):181–197, 1992b.

D. Hendrycks and T. G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019. URL http://arxiv.org/abs/1903.12261.

D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs), 2020. URL http://arxiv.org/abs/1606.08415.

L. Höfling, K. P. Szatko, C. Behrens, Y. Qiu, D. A. Klindt, Z. Jessen, G. W. Schwartz, M. Bethge, P. Berens, K. Franke, A. S. Ecker, and T. Euler. A chromatic feature detector in the retina signals visual context changes. Dec. 2022.

J. P. Jones and L. A. Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1187–1211, Dec. 1987.

W. F. Kindel, E. D. Christensen, and J. Zylberberg. Using deep learning to reveal the neural code for images in primary visual cortex. 2017.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

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*, pages 4–6, 2017.

T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. 2(2):e0000198, 2023. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000198.

B. Lau, G. B. Stanley, and Y. Dan. Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences*, 99(13):8974–

8979, 2002.

S. Lehky, T. Sejnowski, and R. Desimone. Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *The Journal of Neuroscience*, 12(9):3568–3581, Sept. 1992. doi: 10.1523/jneurosci.12-09-03568.1992.

M. J. Lindstrom and D. M. Bates. Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data. 83(404):1014, 1988. ISSN 01621459. doi: 10.2307/2290128.

I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. URL http://arxiv.org/abs/1608.03983.

K.-K. Lurz, M. Bashiri, K. Willeke, A. K. Jagadish, E. Wang, E. Y. Walker, S. A. Cadena, T. Muhammad, E. Cobos, A. S. Tolias, A. S. Ecker, and F. H. Sinz. Generalization in data-driven models of primary visual cortex. In *Proceedings of the International Conference for Learning Representations (ICLR)*, page 2020.10.05.326256, Oct. 2020.

K.-K. Lurz, M. Bashiri, K. Willeke, A. Jagadish, E. Wang, E. Y. Walker, S. A. Cadena, T. Muhammad, E. Cobos, A. S. Tolias, A. S. Ecker, and F. H. Sinz. Generalization in data-driven models of primary visual cortex. In *International Conference on Learning Representations*, 2021.

J. H. Marshel, Y. S. Kim, T. A. Machado, S. Quirin, B. Benson, J. Kadmon, C. Raja, A. Chibukhchyan, C. Ramakrishnan, M. Inoue, et al. Cortical layer–specific critical dynamics triggering perception. *Science*, 365(6453):eaaw5202, 2019.

A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.*, 21(9):1281–1289, Aug. 2018.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

L. T. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. A. Baccus. Deep learning models of the retinal response to natural scenes. *Adv. Neural Inf. Process. Syst.*, 29(Nips): 1369–1377, 2016.

MICrONS Consortium, J. Alexander Bae, M. Baptiste, A. L. Bodor, D. Brittain, J. Buchanan, D. J. Bumbarger, M. A. Castro, B. Celii, E. Cobos, F. Collman, N. M. da Costa, S. Dorkenwald, L. Elabbady, P. G. Fahey, T. Fliss, E. Froudakis, J. Gager, C. Gamlin, A. Halageri, J. Hebditch, Z. Jia, C. Jordan, D. Kapner, N. Kemnitz, S. Kinn, S. Koolman, K. Kuehner, K. Lee, K. Li, R. Lu, T. Macrina, G. Mahalingam, S. McReynolds, E. Miranda, E. Mitchell, S. S. Mondal, M. Moore, S. Mu, T. Muhammad, B. Nehoran, O. Ogedengbe, C. Papadopoulos, S. Papadopoulos, S. Patel, X. Pitkow, S. Popovych, A. Ramos, R. Clay Reid, J. Reimer, C. M. Schneider-Mizell, H. Sebastian Seung, B. Silverman, W. Silversmith, A. Sterling, F. H. Sinz, C. L. Smith, S. Suckow, Z. H. Tan, A. S. Tolias, R. Torres, N. L. Turner, E. Y. Walker, T. Wang, G. Williams, S. Williams, K. Willie, R. Willie, W. Wong, J. Wu, C. Xu, R. Yang, D. Yatsenko, F. Ye, W. Yin, and S.-C. Yu. Functional connectomics spanning multiple areas of mouse visual cortex. July 2021.

M. C. Morrone, M. Tosetti, D. Montanaro, A. Fiorentini, G. Cioni, and D. C. Burr. A cortical area that responds specifically to optic flow, revealed by fMRI. *Nature Neuroscience*, 3(12):1322–1328, 2000. ISSN 1097-6256, 1546-1726. doi: 10.1038/81860.

A. Nayebi, N. C. Kong, C. Zhuang, J. L. Gardner, A. M. Norcia, and D. L. Yamins. Shallow unsupervised models best predict neural responses in mouse visual cortex. *bioRxiv*, pages 2021–06, 2021.

N. Petkov and E. Subramanian. Motion detection, noise reduction, texture suppression and contour enhancement by spatiotemporal gabor filters with surround inhibition. *Biological Cybernetics*, 97(5-6):423–439, 2007. doi: 10.1007/s00422-007-0182-0.

C. R. Ponce, W. Xiao, P. F. Schade, T. S. Hartmann, G. Kreiman, and M. S. Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009.e10, 2019.

R. Prenger, M. C.-K. Wu, S. V. David, and J. L. Gallant. Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Netw.*, 17(5-6):663–679, 2004.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

J. Reimer, E. Froudarakis, C. R. Cadwell, D. Yatsenko, G. H. Denfield, and A. S. Tolias. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*, 84(2): 355–362, Oct. 2014.

N. C. Rust, O. Schwartz, J. A. Movshon, and E. P. Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

C. D. Salzman, K. H. Britten, and W. T. Newsome. Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346(6280):174–177, 1990.

O. Schoppe, N. S. Harper, B. D. B. Willmore, A. J. King, and J. W. H. Schnupp. Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, 10, Feb. 2016. doi: 10.3389/fncom.2016.00010.

X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

J. H. Siegle, X. Jia, S. Durand, S. Gale, C. Bennett, N. Graddis, G. Heller, T. K. Ramirez, H. Choi, J. A. Luviano, P. A. Groblewski, R. Ahmed, A. Arkhipov, A. Bernard, Y. N. Billeh, D. Brown, M. A. Buice, N. Cain, S. Caldejon, L. Casal, A. Cho, M. Chvilicek, T. C. Cox, K. Dai, D. J. Denman, S. E. J. de Vries, R. Dietzman, L. Esposito, C. Farrell, D. Feng, J. Galbraith, M. Garrett, E. C. Gelfand, N. Hancock, J. A. Harris, R. Howard, B. Hu, R. Hytnen, R. Iyer, E. Jessett, K. Johnson, I. Kato, J. Kiggins, S. Lambert, J. Lecoq, P. Ledochowitsch, J. H. Lee, A. Leon, Y. Li, E. Liang, F. Long, K. Mace, J. Melchior, D. Millman, T. Mollenkopf, C. Nayan, L. Ng, K. Ngo, T. Nguyen, P. R. Nicovich, K. North, G. K. Ocker, D. Ollerenshaw, M. Oliver, M. Pachitariu, J. Perkins, M. Reding, D. Reid, M. Robertson, K. Ronellenfitch, S. Seid, C. Slaughterbeck, M. Stoecklin, D. Sullivan, B. Sutton, J. Swapp, C. Thompson, K. Turner, W. Wakeman, J. D. Whitesell, D. Williams, A. Williford, R. Young, H. Zeng, S. Naylor, J. W. Phillips, R. C. Reid, S. Mihalas, S. R. Olsen, and C. Koch. Survey of spiking in the mouse visual system reveals functional hierarchy. 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-03171-x.

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.

N. J. Sofroniew, D. Flickinger, J. King, and K. Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife*, 5:e14472, June 2016.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

J. Touryan, G. Felsen, and Y. Dan. Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, 45(5):781–791, 2005.

B. Vintch, J. A. Movshon, and E. P. Simoncelli. A convolutional subunit model for neuronal responses in macaque V1. *J. Neurosci.*, 35(44):14829–14841, 2015.

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. *Nat. Neurosci.*, 22(12):2060–2065, Dec. 2019.

K. F. Willeke, P. G. Fahey, M. Bashiri, L. Pede, M. F. Burg, C. Blessing, S. A. Cadena, Z. Ding, K.-K. Lurz, K. Ponder, T. Muhammad, S. S. Patel, A. S. Ecker, A. S. Tolias, and F. H. Sinz. The sensorium competition on predicting large-scale mouse primary visual cortex activity, 2022. URL http://arxiv.org/abs/2206.08666.

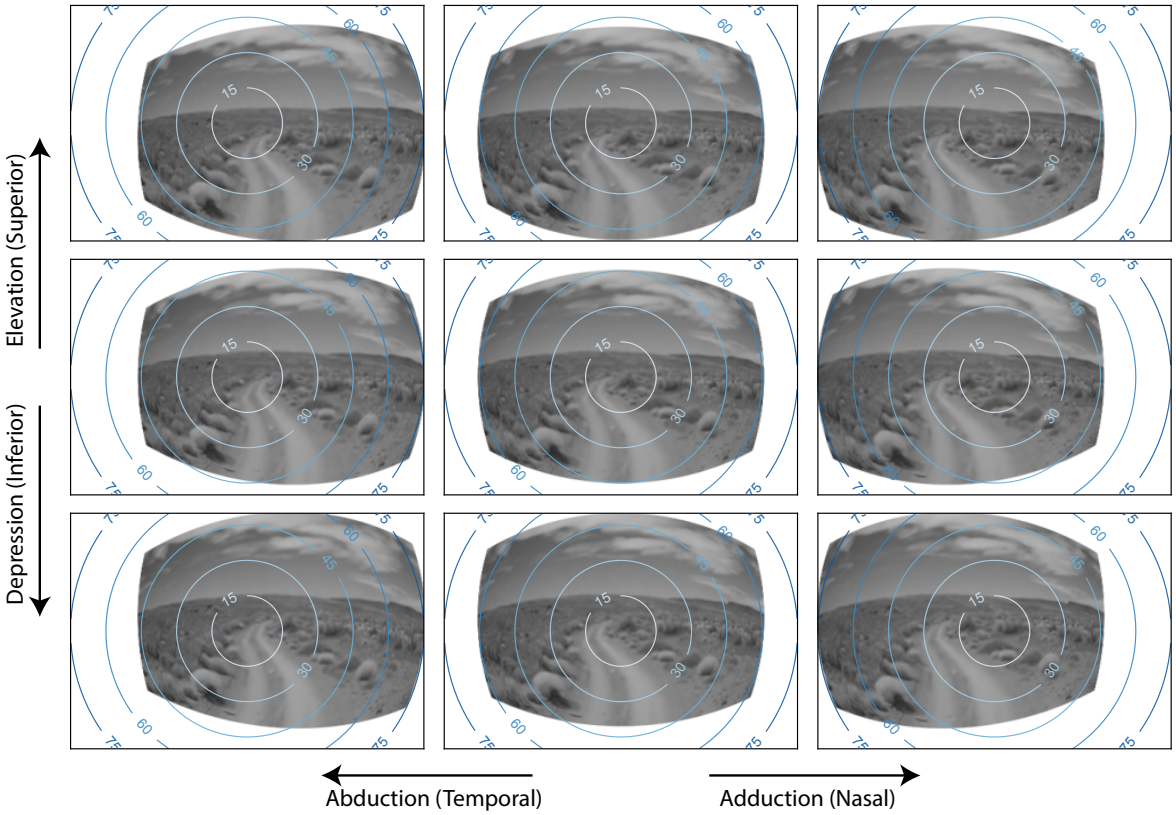
K. F. Willeke, P. G. Fahey, M. Bashiri, L. Pede, M. F. Burg, C. Blessing, S. A. Cadena, Z. Ding, K.-K. Lurz, K. Ponder, T. Muhammad, S. S. Patel, A. S. Ecker, A. S. Tolias, and F. H. Sinz. The sensorium competition webpage, 2023. URL https://sensorium2022.net/home.

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*, May 2014.

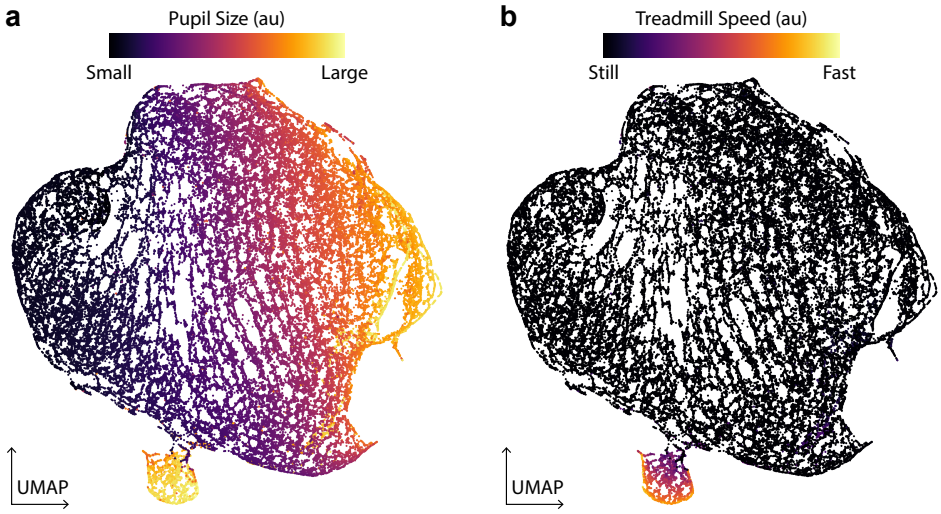
J. Zhuang, L. Ng, D. Williams, M. Valley, Y. Li, M. Garrett, and J. Waters. An extended retinotopic map of mouse cortex. *eLife*, 6:e18372, jan 2017. ISSN 2050-084X.

D. Zipser and R. A. Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331(6158):679–684, 1988.

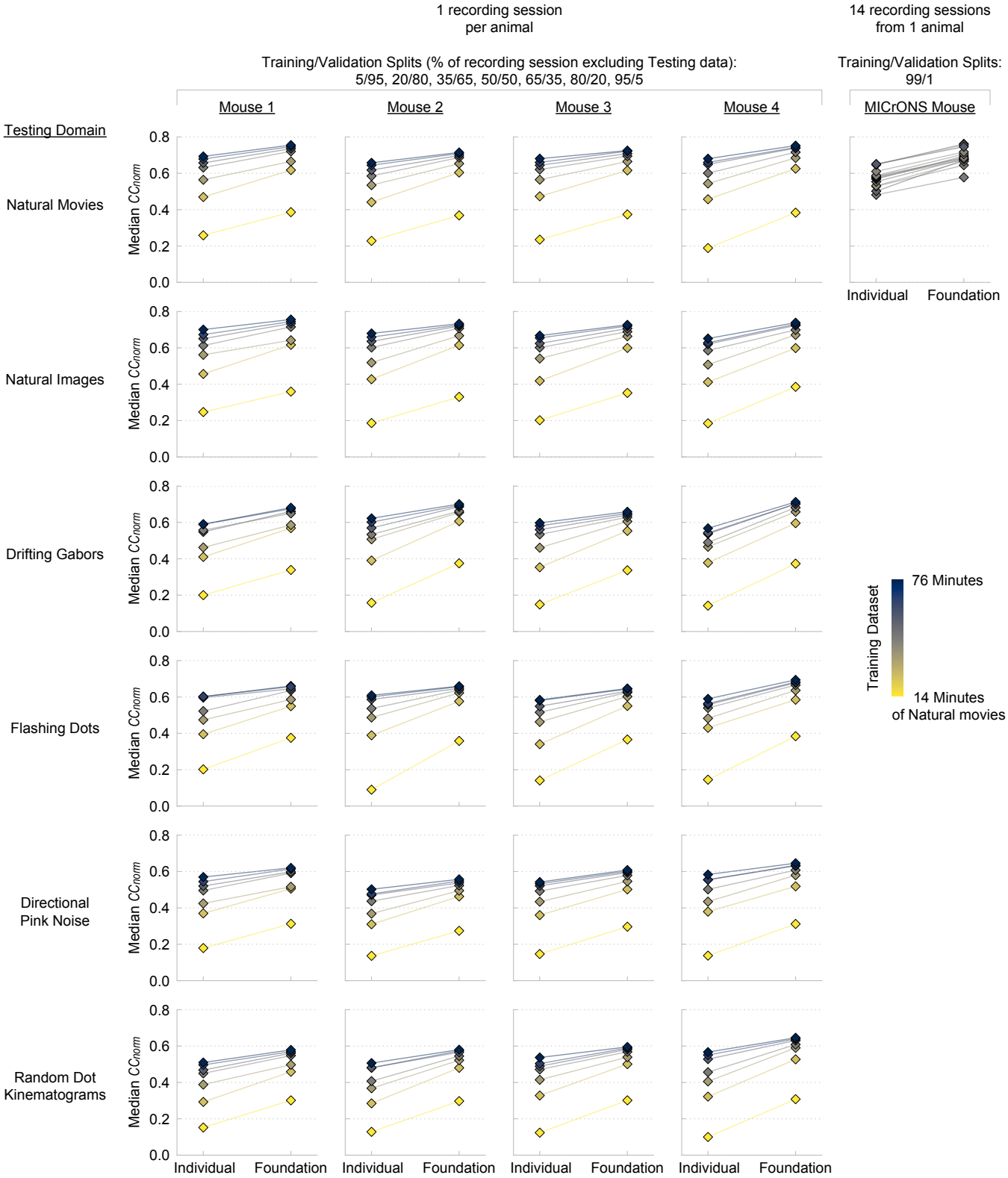
Extended data



Extended Data Fig. 1. ANN perspective. Inferred perspective of an example stimulus frame shown to the left eye, produced by the perspective network of the ANN. Perspectives for 9 rotational positions of the occular globe are shown: 3 for abduction/adduction $\times$ 3 for elevation/depression. The concentric circles denote visual angles in degrees. (See Methods for details on the perspective network.)



Extended Data Fig. 2. ANN modulation. Visualization of the modulation network's output, projected onto 2 dimensions via UMAP. **a**, **b** show the same data from an example recording session and modulation network. Each point on the plot indicates a point in time from the recording session. The colors indicate measurements of pupil size (**a**) and treadmill speed (**b**) at the respective points in time. (See Methods for details on the modulation network.)



Extended Data Fig. 3. ANN performance: Individual vs. Foundation Predictive accuracy (median CC_{norm} across neurons) of foundation models (with the foundation core) vs. individual models (with cores trained on individual recording sessions). For the 4 mice in the 4 left columns, 1 recording session was performed, and that data was partitioned into 7 training/validation splits, which were used to train separate individual/foundation models. The predictive accuracy of those models (diamonds) is reported for 6 testing stimulus domains (rows). For the MICrONS mouse, 14 recording sessions were performed, for each recording session, a model was trained using nearly all (99%) of the data available for training/validation. The MICrONS models were only tested on the natural movies, due to the lack of the other stimuli in the recording sessions. All models were trained only using natural movies.