Deep Neural Network Models as Digital Twins for Functional Characterization of Visual Cortex

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt
von

Konstantin Friedrich Willeke
aus Jena

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	27.01.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Fabian Sinz
2. Berichterstatter:	Prof. Dr. Felix Wichmann

Erklärung / Declaration

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel

“Deep neural network models as digital twins for functional characterization of visual cortex”

selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled

"Deep neural network models as digital twins for functional characterization of visual cortex"

submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Ort/Place, Datum/Date

Unterschrift/Signature

Acknowledgements

Throughout my PhD, I've been tremendously lucky.

I was lucky to have found a mentor and advisor in Fabian who is uniquely kind, supportive, and knowledgeable. Working under your direction was truly a privilege and I'll forever be indebted that you took the gamble and gave me the chance to learn from you.

I also want to thank my inspiring co-advisors Andreas and Alex. The research atmosphere and community within and across all of your labs has been incredible. You both were mentors that I will always look up to. I'm also thankful for the support of Georg Martius and for the help and advice of Felix Wichmann.

To all of my amazing colleagues at the Sinzlab, I'd like to say that I'm immensely grateful. Your help, collegiality, and good spirits mean more than I could ever express. Having you as my colleagues and friends on this journey has been without a doubt the most fun part of my PhD. I'd like to thank Edgar, Christoph, Konstantin, Mohammad, Arne, Shahd, Pawel, and Akshay for the great times we've had in the early days of the lab and for developing many of the tools together that allowed us to do our research. The times we spent together at the kicker table, discussing our ideas, and struggling together to make datajoint and nnfabrik work will forever make me smile. Suhas, Pavi, Dominik, Caio, Sarah, you've made my time in the lab truly special and rewarding. I'm also indebted to my colleagues from the labs of Alex and Andreas. Thank you so much for your guidance and help Santiago, Ivan, Max, Mara, Polly, Laura, Michaela, Saumil, Kelli, Tori, Jiakun, Paul, Nikos, Maria, Zhiwei, Zhuokun, Taliah, and Kayla. And a special thanks to Silke, Veronika, Camila, and Cate for your unbelievable help and patience.

I want to thank the vibrant and inspiring neuroscience and AI research community in Tübingen. I was lucky to be a part of such a genuinely friendly community of teachers, researchers, students, and friends at the AI Center, CIN, Hertie, IMPRS-IS, and especially the GTC. I also want to extend a heartfelt thank you to Ziad Hafed and Antimo Buonocore for your guidance.

To my parents, Joachim and Angelika, and my brother and sister, Simon and Lea, I want to say that your support and love mean the world to me. I feel blessed to have such a great and caring family. To Katrin, it's been the greatest luck of all that during this journey, I've met you. Thank you for more than everything.

Contents

List of publications included in this thesis

Peer reviewed publications

- [1]: Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H Sinz, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, 2022

- [2]: Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F. Burg, Santiago A. Cadena, Zhiwei Ding, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Kaiwen Deng, Yuanfang Guan, Yiqin Zhu, Kaiwen Xiao, Xiao Han, Simone Azeglio, Ulisse Ferrari, Peter Neri, Olivier Marre, Adrian Hoffmann, Kirill Fedyanin, Kirill Vishniakov, Maxim Panov, Subash Prakash, Kishan Naik, Kantharaju Narayanappa, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. Retrospective on the sensorium 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 314–333. PMLR, 28 Nov–09 Dec 2022. URL `https://proceedings.mlr.press/v220/willeke22a.html`

Submitted publications

- [3]: Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. May 2023. doi: 10.1101/2023.05.12.540591

Statement of contributions according to § 5 (2)

Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H Sinz, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, 2022

Abstract
To increase computational flexibility, sensory processing changes with behavioral context. In the visual system, active behavioral states characterized by motor activity and pupil dilation [1, 2] enhance sensory responses but typically leave the preferred stimuli of neurons unchanged. Here we find that behavioral state also modulates stimulus selectivity in mouse visual cortex in the context of colored natural scenes. Using population imaging in behaving mice, pharmacology, and deep neural network modeling, we identified a rapid shift of color selectivity towards ultraviolet stimuli during an active behavioral state. This was exclusively caused by pupil dilation, resulting in a dynamic switch from rod to cone photoreceptors, thereby extending their role beyond night and day vision. The change in tuning facilitated the decoding of ethological stimuli, such as aerial predators against the twilight sky. In contrast to previous studies that have used pupil dilation as an indirect measure of brain state, our results suggest that statedependent pupil dilation itself differentially recruits rods and cones on fast timescales to tune visual representations to behavioral demands.

Contribution

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Konstantin F. Willeke	1*	20 %	20 %	50 %	30 %
Status in publication process:		*Published*			

*Shared co-first authorship with Katrin Franke

Statement of contributions according to § 5 (2)

Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. May 2023. doi: 10.1101/2023.05.12.540591

Abstract
Deciphering the brain's structure-function relationship is key to understanding the neuronal mechanisms underlying perception and cognition. The cortical column, a vertical organization of neurons with similar functions, is a classic example of primate neocortex structure-function organization. While columns have been identified in primary sensory areas using parametric stimuli, their prevalence across higher-level cortex is debated. A key hurdle in identifying columns is the difficulty of characterizing complex nonlinear neuronal tuning, especially with high-dimensional sensory inputs. Here, we asked whether area V4, a mid-level area of the macaque visual system, is organized into columns. We combined large-scale linear probe recordings with deep learning methods to systematically characterize the tuning of >1,200 V4 neurons using in silico synthesis of most exciting images (MEIs), followed by in vivo verification. We found that the MEIs of single V4 neurons exhibited complex features like textures, shapes, or even high-level attributes such as eye-like structures. Neurons recorded on the same silicon probe, inserted orthogonal to the cortical surface, were selective to similar spatial features, as expected from a columnar organization. We quantified this finding using human psychophysics and by measuring MEI similarity in a non-linear embedding space, learned with a contrastive loss. Moreover, the selectivity of the neuronal population was clustered, suggesting that V4 neurons form distinct functional groups of shared feature selectivity, reminiscent of cell types. These functional groups closely mirrored the feature maps of units in artificial vision systems, hinting at shared encoding principles between biological and artificial vision. Our findings provide evidence that columns and functional cell types may constitute universal organizing principles of the primate neocortex, simplifying the cortex's complexity into simpler circuit motifs which perform canonical computations.

Contribution

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Konstantin F. Willeke	1*	50 %	10 %	80 %	30 %
Status in publication process:		*Under review*			

*Shared co-first authorship with Kelli Restivo

Statement of contributions according to § 5 (2)

Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F. Burg, Santiago A. Cadena, Zhiwei Ding, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Kaiwen Deng, Yuanfang Guan, Yiqin Zhu, Kaiwen Xiao, Xiao Han, Simone Azeglio, Ulisse Ferrari, Peter Neri, Olivier Marre, Adrian Hoffmann, Kirill Fedyanin, Kirill Vishniakov, Maxim Panov, Subash Prakash, Kishan Naik, Kantharaju Narayanappa, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. Retrospective on the sensorium 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 314–333. PMLR, 28 Nov–09 Dec 2022. URL `https://proceedings.mlr.press/v220/willeke22a.html`

Abstract

The neural underpinning of the biological visual system is challenging to study experimentally, in particular as neuronal activity becomes increasingly nonlinear with respect to visual input. Artificial neural networks (ANNs) can serve a variety of goals for improving our understanding of this complex system, not only serving as predictive digital twins of sensory cortex for novel hypothesis generation in silico, but also incorporating bio-inspired architectural motifs to progressively bridge the gap between biological and machine vision. The mouse has recently emerged as a popular model system to study visual information processing, but no standardized large-scale benchmark to identify state-of-the-art models of the mouse visual system has been established. To fill this gap, we proposed the SENSORIUM benchmark competition. We collected a large-scale dataset from mouse primary visual cortex containing the responses of more than 28,000 neurons across seven mice stimulated with thousands of natural images, together with simultaneous behavioral measurements that include running speed, pupil dilation, and eye movements. The benchmark challenge ranked models based on predictive performance for neuronal responses on a held-out test set, and included two tracks for model input limited to either stimulus only (SENSORIUM) or stimulus plus behavior (SENSORIUM+). As a part of the NeurIPS 2022 competition track, we received 172 model submissions from 26 teams, with the winning teams improving our previous state-of-the-art model by more than 15%. Dataset access and infrastructure for evaluation of model predictions will remain online as an ongoing benchmark. We would like to see this as a starting point for regular challenges and data releases, and as a standard tool for measuring progress in large-scale neural system identification models of the mouse visual system and beyond.

Contribution

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Konstantin F. Willeke	1*	50 %	10 %	90 %	50 %
Status in publication process:		*Published*			

*Shared co-first authorship with Paul Fahey

Abstract

Understanding how the brain processes visual information remains a fundamental challenge in neuroscience. Recent advances in deep learning have revolutionized our ability to model visual processing, enabling the development of "digital twins" - deep neural networks that accurately predict how individual neurons respond to arbitrary visual stimuli. This thesis explores how these models can advance our understanding of visual processing across different scales and species, through three projects combining computational prediction with biological validation. In the first project, we investigate the dynamic relationship between behavioral state and visual processing in mouse primary visual cortex. By extending digital twin models to incorporate both color processing and behavioral variables, we predict how neural responses change with behavioral state. Through systematic generation of maximally exciting inputs (MEIs) conditioned on behavioral measurements and subsequent closed-loop validation experiments, we uncover a novel mechanism where mice rapidly modulate their color processing based on pupil size to enhance detection of behaviorally relevant stimuli. The second project addresses a fundamental question in visual neuroscience: how is functional selectivity organized in higher visual areas? Using digital twins to characterize neuronal responses in macaque area V4, we identify distinct functional groups of neurons sharing preferences for complex visual features. Our key finding demonstrates that neurons within the same cortical column exhibit similar response preferences, providing evidence that columnar organization, previously established in early visual areas, extends to higher-order visual processing as a general principle of cortical computation. The third project establishes a standardized benchmark platform for evaluating digital twin models of mouse V1, addressing the growing need for systematic comparison of neural prediction models. We provide a comprehensive dataset of thousands of neurons responding to naturalistic stimuli, coupled with evaluation tools and metrics for rigorous model comparison. This framework promotes collaborative advancement through competitive model development and establishes clear criteria for measuring progress in neural response prediction. Taken together, this work establishes digital twins as powerful tools for investigating neural computation, enabling systematic exploration of hypotheses through a combination of computational modeling and targeted biological experiments. The development of standardized benchmarks ensures that progress in this field remains rigorous and reproducible. These advances provide a framework for investigating neural computation across scales, from single neurons to population dynamics, accelerating our understanding of visual processing in the brain.

Zusammenfassung

Das Verständnis der visuellen Informationsverarbeitung im Gehirn bleibt eine fundamentale Herausforderung in den Neurowissenschaften. Jüngste Fortschritte im Deep Learning haben unsere Fähigkeit zur Modellierung visueller Verarbeitung revolutioniert und die Entwicklung von "digitalen Zwillingen" ermöglicht - neuronale Netzwerke, die präzise vorhersagen, wie einzelne Neuronen auf beliebige visuelle Reize reagieren. Diese Dissertation untersucht, wie diese Modelle unser Verständnis der visuellen Verarbeitung über verschiedene Skalen und Spezies hinweg erweitern können, durch drei verbundene Projekte, die spezifische Vorhersagen mit biologischer Validierung kombinieren. Im ersten Projekt untersuchen wir die dynamische Beziehung zwischen Verhaltenszustand und visueller Verarbeitung im primären visuellen Kortex (V1) der Maus. Durch die Erweiterung der digitalen Zwillingsmodelle um Farbverarbeitung und Verhaltensparameter sagen wir vorher, wie sich neuronale Antworten mit dem Verhaltenszustand ändern. Durch systematische Generierung maximal erregender Inputs basierend auf Verhaltensmessungen und nachfolgende Closed-Loop-Validierungsexperimente entdecken wir einen neuartigen Mechanismus, bei dem Mäuse ihre Farbverarbeitung basierend auf der Pupillengröße modulieren, um die Erkennung verhaltensrelevanter Reize zu verbessern. Das zweite Projekt befasst sich mit einer grundlegenden Frage der visuellen Neurowissenschaft: Wie ist die funktionelle Selektivität in höheren visuellen Arealen organisiert? Mithilfe digitaler Zwillinge zur Charakterisierung neuronaler Antworten im Makaken-Areal V4 identifizieren wir verschiedene funktionelle Neuronengruppen mit gemeinsamen Präferenzen für komplexe visuelle Merkmale. Unser zentrales Ergebnis zeigt, dass Neuronen innerhalb derselben kortikalen Säule ähnliche Antwortpräferenzen aufweisen, was belegt, dass sich die kolumnäre Organisation, die bereits in frühen visuellen Arealen nachgewiesen wurde, als allgemeines Prinzip der kortikalen Verarbeitung auf die höhere visuelle Verarbeitung erstreckt. Das dritte Projekt etabliert eine standardisierte Benchmark-Plattform zur Evaluierung digitaler Zwillingsmodelle des Maus-V1 und adressiert damit den wachsenden Bedarf an systematischen Vergleichen neuronaler Vorhersagemodelle. Wir stellen einen umfassenden Datensatz von tausenden Neuronen zur Verfügung, die auf naturalistische Stimuli reagieren, gekoppelt mit Evaluierungswerkzeugen und Metriken für rigorose Modellvergleiche. Dieses Framework fördert kollaborative Fortschritte durch kompetitive Modellentwicklung und etabliert klare Kriterien zur Messung des Fortschritts in der neuronalen Antwortvorhersage. Zusammengenommen etabliert diese Arbeit digitale Zwillinge als leistungsfähige Werkzeuge zur Untersuchung neuronaler Verarbeitung und ermöglicht die systematische Erforschung

von Hypothesen durch die Kombination von computergestützter Modellierung und gezielten biologischen Experimenten. Die Entwicklung standardisierter Benchmarks stellt sicher, dass der Fortschritt in diesem Bereich rigoros und reproduzierbar bleibt. Diese Fortschritte bieten einen Rahmen für die Untersuchung neuronaler Verarbeitungsprozesse über verschiedene Skalen hinweg, von einzelnen Neuronen bis hin zu Populationsdynamiken.

1 Introduction

1.1 The computational view of sensory processing

The neuron doctrine posits that individual neurons are the fundamental units of the nervous system's structure. This theory originated from the groundbreaking work of Santiago Ramón y Cajal [4]. Cajal's meticulous observations demonstrated the presence of physical discontinuities between neuronal processes, directly contradicting the prevailing the reticular theory, advocated by Camillo Golgi, which incorrectly proposed that neurons formed a continuous syncytium [5, 6]. Alongside the consolidation of the single neuron as the fundamental anatomical unit of the brain, the functional view of the neuron doctrine developed. Sherrington [7] first argued that individual neurons are the basic functional units of the nervous system. Sherrington also proposed the concept of the receptive field, as the area of skin that, when stimulated, elicits a scratch reflex [7]. With the development of techniques for recording activity from single axons, the receptive field concept gained robust support across other sensory domains, including the visual system [8]. In this view, each sensory neuron possesses its own unique receptive field, a specific feature of the sensory world that drives its activity and thus determines its function [9, 10]. A prime example that helped to shape the exploration of the visual system was the so-called bug detector neuron in the frog retina - single neurons with small, motion-sensitive receptive fields that seem to be optimally designed for detecting moving flies [9, 11, 12]. Subsequently, to understand the brain's function, single neurons, and their receptive fields became the focus of electrophysiology research, especially after David Hubel introduced the tungsten microelectrode [13]. This opened up the still ongoing tradition of single-cell recordings and the careful mapping of receptive fields throughout the visual system, from the retina to higher cortical areas. This led to breakthrough findings such as the topographically organized receptive fields in cortical columns by Mountcastle [14], as well as Hubel and Wiesel's findings about the retinotopic organization of V1, orientation, and ocular dominance columns [15–19].

These groundbreaking results firmly established the single neuron as the brain's anatomical, functional, and perceptual unit. The hierarchical organization of the visual system, with simple and complex cells in V1 feeding into higher cortical areas [17], suggested that grandmother cells responsible for the perception of highly specific features such as persons might exist at the top of this hierarchy [10]. The identification of face cells in the temporal cortex of monkeys and humans, which respond selectively to images of particular individuals, provided support for this notion [20–24].

However, the feature detector view of single neurons has encountered significant challenges, especially when applied to higher cortical areas. Neurons in regions such as the inferotemporal cortex, prefrontal cortex, and hippocampus exhibit a breadth and flexibility of tuning that is difficult to reconcile with the notion of simple feature detection. Neurons in these areas often respond to complex combinations of features and show remarkable adaptability in their responses [25–29]. Despite these challenges to the single neuron doctrine, it provided inspiration to David Marr and his ideas on best identifying" the levels of understanding of the brain [30]. Marr proposed that a comprehensive understanding of an information processing system, such as the brain, requires analysis at three levels: computational, algorithmic, and implementation. The computational level specifies the goal or problem to be solved by the system, the constraints it must operate under, and the resources available. In the context of the visual system, the computational level would define the specific tasks that the system is designed to perform, such as detecting edges, computing motion, or recognizing objects, or more generally, what transformation occurs between input and output [30].

At the level of single neurons, the computational level would specify which feature or transformation a neuron or group of neurons is designed to detect or compute. For example, in V1, simple cells are optimized for detecting oriented edges at specific spatial locations, while complex cells are tuned to oriented edges with greater spatial invariance [15–19]. The computational level thus provides a framework for interpreting the response properties of neurons in terms of their computational goals. According to Marr [30], the computations performed by single neurons and circuits should be the first fundamental step in understanding neuronal processing in general, which then guides the inquiry into the biological implementation and exact algorithm carried out. Similar views on the primacy of the computational level were already being expressed by Barlow, for example, in his efficient coding hypothesis [31], and in subsequent works [32, 33].

1.2 Models of single neurons of the visual cortex

Along with the early characterization of visual response functions of neurons, computational modeling has played a crucial role in understanding and formalizing the neuron's computations. By representing the receptive field as a mathematical function, models have provided a quantitative framework for describing the stimulus selectivity of neurons at various stages of the visual hierarchy. The concept of a receptive field is typically represented in a model that begins with a linear filter. Filtering involves computing the weighted sum of the intensities at each local region of an image, that is, the value of each pixel intensity, using the values of a filter. A linear filter then describes the stimulus selectivity for a neuron: images that closely resemble the filter elicit strong responses, whereas images that have little resemblance to the filter produce minimal responses. The first simple mathematical models of neurons at the early visual processing stages, namely the retina, LGN, and

V1 simple cells, were composed of a single linear filter, a simple Gabor filter [34–36]. In contrast, neurons in higher processing stages (V1 complex cells and beyond) require two or more filters [35, 37–39]. After the linear filtering is applied, the next step is to convert the filtered outputs into a positive scalar firing rate response of the neuron of interest. This is done by passing the output into a static nonlinearity, such as half-wave rectification [40–42]. A similar concept was introduced in the context of artificial neural networks by Fukushima [43] as the so called ramp function, more commonly known as the rectified linear unit (ReLU). Early models considered the output of the static-nonlinearity as the neuron's spiking activity and were called LN-models, for their combination of a linear filter with a non-linear response function. Later models added a Poisson process for generating precise spike times from the firing rate, and were thus termed LNP, Linear-Nonlinear-Poisson models [44–46]. This approach of finding the optimal parameters of the filter is also referred to as system identification, a quantification of the function from input stimuli to neuronal response.

The LNP model, the most basic encoding model, performs well despite its simplicity in predicting spike rates to basic stimuli when considering responses of photoreceptors or ganglion cells in the retina, as well as LGN and simple cell V1 responses [47–49]. The existing standard model for V1 neurons was largely derived from experiments using a narrow class of test stimuli, often designed for characterizing linear systems, such as spots, white noise, or sine-wave gratings. These stimuli were found in painstaking experiments and often by sheer luck in finding the right stimulus properties to make the neurons of interest fire [17, 46, 49]. While the hope is that insights gained from these reduced stimuli will generalize to more complex situations [50], such as natural scenes, this assumption does not hold for nonlinear systems. In a nonlinear system, the response to a reduced set of stimuli won't be able to predict the response to an arbitrary combination of those stimuli. For example, changing even one stimulus parameter, such as the contrast, requires a re-learning of the linear filter [51, 52]. It is, therefore, nearly impossible to map the response to a large enough set of simple stimuli. Consequently, the use of natural ecologically relevant stimuli, such as natural images and video, or other stimuli, generated with natural image statistics became more relevant. [33, 53–56].

To capture the complex nonlinear response properties of neurons in the visual cortex to natural stimuli, so-called LN-LN cascade models, which incorporate additional nonlinear stages to better account for the hierarchical visual processing, were proposed (reviewed in [57]). These models either learn convolutional subunits [58–60] or use handcrafted wavelet representations [61] to capture more complex response properties. While LN-LN cascade models outperform simple LN models in predicting V1 responses to natural stimuli, they still face challenges in capturing the full range of nonlinearities observed, for example, in V1 neurons [59]. Although deeper, multi-layer networks like HMAX [62] showed some success in predicting neuronal responses in higher-level visual areas, particularly in the macaque's inferior temporal (IT) cortex, these models achieved only limited predictive accuracy overall [63–65]. A likely reason for the failures of these deeper cascade models was

the limitation in data availability to train these networks, as well as the limited sophistication of network architectures that could be fit to neural data.

1.3 Deep neural network models

Since the advent of deep learning in 2012 [66, 67], a new approach for computational modeling of visual sensory neurons emerged. Deep convolutional neural networks (CNNs) have proven themselves to be powerful tools for understanding the computational principles and architectural motifs underlying object recognition. The development of CNNs has been driven by the convergence of two lines of research: computational modeling of biological vision using artificial neural networks (ANNs) [43], as described in the previous chapter, and the development of object recognition algorithms in computer vision [67, 68]. A breakthrough in CNN performance on object recognition tasks [66, 67] highlighted the importance of architectural motifs that are shared with biological image processing systems, such as hierarchical processing by stacking of linear-nonlinear layers, parallel processing streams via parallel convolutional filters, and lateral inhibition through response normalization. Furthermore, beyond the inspirations drawn from neuroscience, the ground-breaking success of CNNs in computer vision and object recognition has been, in large part, driven by advancements in computational resources. Large-scale datasets [69] and the utilization of optimized GPUs for parallel computation [67] made it possible to train networks with millions, and even billions, of parameters through error backpropagation [67, 70–72]. Together, these advancements led to big leaps in computer vision, where previous efforts with limited data, such as recognition of handwritten digits with CNNs [73], were successful but not powerful enough for more challenging applications.

Along with the state-of-the-art performance on object recognition benchmarks, ANNs have since been widely used to study the visual system, particularly for insights into similarities in representations between features of deep neural networks and biological neurons across the visual stream. Two especially fruitful approaches that make use of ANNs are representational similarity analysis (RSA) and single-neuron encoding models. RSA quantifies the dissimilarities between population activity patterns elicited by different experimental conditions, using typically a large number of visual stimuli, and summarizes them in representational dissimilarity matrices (RDMs) [74, 75]. If the response geometries of the internal representations of ANNs and the brain agree, they are considered similar. By focusing on population-level representational geometries, RSA circumvents the need for an explicit mapping between ANN units and individual neurons. RDMs can be computed using population activity patterns from various levels of granularity, including whole networks, network layers, feature maps, or individual units [76]. These ANN RDMs can then be directly compared with brain RDMs from neural populations in any regions of interest, or an additional data fitting step can be employed to optimize the agreement between the ANN and the brain [77–79].

Furthermore, ANNs have proven to be highly effective as encoding models of the brain by learning latent feature representations that capture similar nonlinear transformations carried out by the brain. For example, it has been shown that neuronal activity can be predicted in monkeys and humans at various stages of the visual hierarchy without being explicitly trained on neural data [78, 80–87]. In this approach, the activity of each neuron (or voxel in case of fMRI studies) is predicted by a weighted sum of the network's hidden unit activations of a particular layer. In essence, the linear filter of an LNP model is simply replaced by a large number of ANN features, which turns the neuronal prediction into a large-scale multivariate approach to fit each neuron independently. Using these pre-trained features of an ANN to solve a different task, that is, predicting the activity of single neurons, is also known as transfer learning and has been widely successful in many fields [88, 89] of machine learning. Interestingly, it has been shown a promising correspondence exists between deeper layers of ANNs and which stage of the ventral visual stream these layers are able to fit best [81, 84, 90]. Moreover, it has also been shown that better task performance, i.e., classification performance on ImageNet, led to better features for neuronal predictions [91, 92]. In a similar vein, using features of ANNs that were trained to be robust against image distortions also improved the fit to neuronal data [91, 93]. However, more recent experiments with a large variety of ANN architectures were not able to confirm these results and showed a reversal of alignment between neuronal data and ANN features [91, 94–96].

Another approach to using ANNs as encoding models is to fit all parameters end-to-end directly to neuronal data. This data-driven modeling, as compared to task-driven modeling outlined above, became feasible using the same advances as in training ANNs for image classification: weight-sharing through convolutional architectures, batch normalization, and better model training procedures [60, 85, 86, 97–101]. Compared to LN-LN cascade models, these data-driven ANNs rival or exceed the performance of task-driven models [102]. One key advantage of data-driven ANNs is their ability to facilitate straightforward comparisons between different architectures and their inductive biases. This enables researchers to investigate which computational functions are crucial for predicting visual neuron activity, while also exploring how well established neuroscientific models can be incorporated into ANN frameworks [87, 103–106].

Taken together, both task and data-driven models based on ANNs are the most accurate models to date for predicting the activity of visual neurons in response to arbitrary images. By varying the statistics of the dataset, changing the model architectures, or the training objectives, specific hypotheses can be tested regarding which computations are performed by the neurons of interest [107, 108]. Due to of the high prediction accuracy of these models, they are also referred to as *digital twins* of biological neurons, as they reliably capture the response properties of individual neurons remarkably well. This enables systematic experimentation for testing specific hypotheses about neural computation using the digital twins.

1.4 Interpretability methods for deep neural models of visual processing

While ANN models have demonstrated superior performance in capturing the complex input-output relationships of visual neurons compared to traditional Linear-Nonlinear-Poisson (LNP) or cascaded LN-LN models, this improved predictive power comes at the cost of reduced interpretability of the digital twins. Although the simpler LNP models offer at least a degree of interpretability by either hand-designing or inspecting each component, this approach is no longer feasible for ANN models. This is why ANN models can be considered *black box models*, with their inner function equally unknown as the biological counterpart they are trying to model [109], which would defeat the purpose of leveraging these models to investigate the computational function of visual neurons, i.e. their properties. Nevertheless, there are emerging techniques in model interpretability for deep-learning-based computer vision models that try to address the black box character of these models. By investigating the computational properties and learned algorithms of ANNs solving computer vision tasks, the problems and solutions become similar to those in trying to understand the functions of individual visual neurons and populations of neurons in visual cortex [107, 110–113].

Early artificial neural networks of the 1980s featured transparent architectures with clear relationships between weights and functions, allowing straightforward interpretation through manual inspection [114]. The introduction of backpropagation then marked a shift toward more sophisticated visualization techniques, including feature visualization, activation maximization, and attribution techniques that map network decisions to input features [115–117]. Interpretability research for ANNs can be summarized into four categories: behavioral, attributional, concept-based, and mechanistic [110, 111]. *Behavioral* analysis is concerned only with the input-output relations and treats the model as a black box. It is used to gain insights into the behavior of the model, such as a model's robustness against image perturbations [118–120]. *Attributional* methods similarly aim to explain the entire model's behavior by showing input feature influence on the output without the need of understanding the internal structure. This is done by visualizing the input features that most highly drive the output model activation through analyzing the gradients with respect to individual outputs [121–125].

Concept-based methods quantify the interpretability of a model by measuring the degree of alignment between hidden unit activations and ground-truth labels in a pre-defined dictionary of concepts. An example approach is called network dissection [126–128], which is able to identify the unit activations within the model that correspond to semantically meaningful concepts by matching the activation level of these units to a large input dataset with pre-segmented images. The alignment between the units and concepts in a model trained for image classification can then be tested by selectively removing, that is, dropping out, these units and observing the reduction in accuracy for classes related to these concepts [128]. Lastly, *mechanistic interpretability* is a collection of bottom-up approaches that aim to study the

individual model components through analyses of hidden units, model layers, and their connections. These methods seek to identify specific circuits or components responsible for driving particular model behaviors [129–131].

Both neuroscience and mechanistic interpretability aim to understand information processing systems by identifying specific computational circuits and their functions [107, 113, 132]. This approach is exemplified by research on curve detection in convolutional neural networks [133]. Using techniques like gradient ascent [132] and image patch analysis, researchers identified neurons in CNNs that are selectively tuned to curves at different orientations [133], which mirrors classic neuroscience studies that mapped similar curvature circuits in primate visual areas V2 and V4 [134–136]. In particular, feature visualization techniques are well suited for analyzing deep neural networks trained to predict neural responses in the visual system, as identifying optimal stimuli has been fundamental to understanding visual processing. Optimization-based methods use gradient descent to generate synthetic inputs from random noise that maximize the responses of a unit in an artificial neural network, revealing their preferred stimuli [132, 137–142]. This process parallels the serach for the stimulus that elicits the highest response in visual neurons used in neurophysiology, where researchers map neural response properties through controlled stimulus manipulation.

Using digital twins, optimal stimuli for individual neurons can thus be found similarly, by using gradient ascent on the pixels of an image. The activity of a single neuron of the digital twin then acts as the objective to be maximized, by optimizing the parameters of the image iteratively to increase the firing rate of the target neuron. The resulting image is called a *MEI*, maximally exciting input. This approach has its downsides, because without regularization, the resulting image often contains high frequency artifacts due to either overfitting or because the optimization can get stuck in local maxima [137, 143]. The other major downside of this approach is that without *in-vivo* verification, it is unclear whether the obtained image is indeed driving the response of a single neuron highly and is thus informative about the role of the neuron in visual processing.

Recently, several research groups have overcome the substantial experimental and computational obstacles to verify MEIs in-vivo by running closed loop experiments [144–146]. In these experiments, first a large set of natural images and neuronal responses are collected such that a deep-learning based digital twin of the recorded neurons can be trained (Fig. 1.1a). Then MEIs can be generated using the trained digital twin for a selection of neurons (Fig. 2.1b). These MEIs can then be shown back to the animal while still recording from the same neurons, which can be achieved by recording a neuron across multiple days with calcium-imaging [144] or chronically implanted multi-electrode arrays [145]. In these experiments, it was convincingly shown that the MEIs indeed highly and selectively excite the target neurons [144, 145]. Using this approach, Walker et al. [144] and colleagues demonstrated that the feature selectivity of neurons in primary visual cortex of mice is more complex than previously thought, as the MEIs revealed more complex

visual features compared to classical Gabor stimuli, which were less exciting stimuli for the majority of neurons compared to the MEIs. Another elegant study by Bashivan et al. [145] demonstrated successful closed-loop experiments in macaque area V4, which revealed that activity of single units as well as groups of units can be selectively excited or inhibited by *in-silico* optimized stimuli. Subsequent studies using closed-loop approaches were able to extensively characterize visual selectivity in mouse visual cortex by studying center-surround interactions [147], spatial-frequency selectivity and its possible role in object segmentation [148], and the hierarchical visual feature complexity in higher visual areas [149]. Digital twins and MEIs were also used to study the retina in mice and primates [150–152]. Finally, MEIs can also be combined with extensive in-silico characterization to compare the activity elicited by MEIs to classical stimuli such as oriented gratings [153, 154].

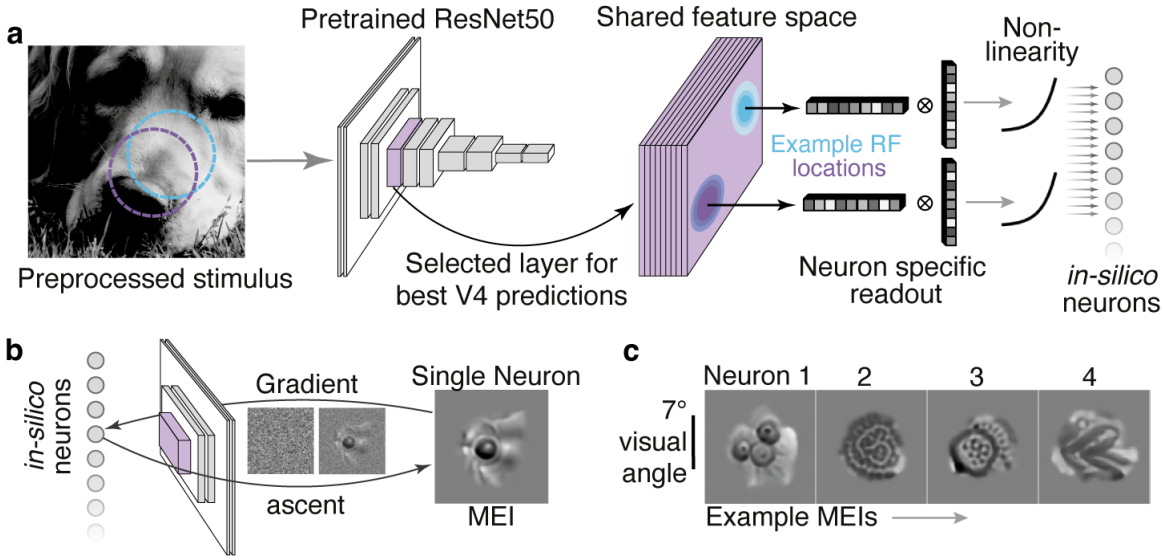


Figure 1.1: **Overview of the digital twin approach to study the functional characterization of neurons in visual cortex using maximally exciting inputs (MEIs)**. **a,** First, a deep neural network model is trained on pairs of images and the neuronal responses. The model is then able to predict responses of individual *in-silico* neurons, to arbitrary images. This neural predictive model is called a *digital twin*. **b,** The activity of a single neuron of the digital twin becomes an objective to be maximized. Starting from a random image, namely white noise, each pixel of the image is iteratively optimized to increase the firing rate of the target neuron via regularized gradient ascent. **c,** Example MEIs obtained for single neurons in macaque area V1.

1.5 Work presented in this thesis

This thesis builds upon the recent work of building digital twins of the visual cortex using deep neural network models [60, 86, 97–100, 102, 153, 155–163], in particular in the context of closed-loop experiments for in-vivo verification of the predictions of digital twins [144, 147, 148, 150–152, 154].

The first project, presented in chapter 2.1, investigates the effect of brain states on visual processing in mice by extending the digital twin and MEI approach in the

color as well as the behavioral domain. Many studies have shown the influence of behavior and internal brain states on visual responses along the entire processing hierarchy [164–168]. In this work, we generated MEIs conditioned on behavioral variables such as pupil size and locomotion speed. The results of extensive closed-loop experiments demonstrated that there is a powerful link between changes in the behavioral state and neuronal tuning. Through subsequent in-silico and in-vivo experiments, we uncovered a novel mechanism in mice to shift color-selectivity in area V1 to better detect behaviorally relevant stimuli.

The second project, summarized in chapter 2.2, examines the functional organization of macaque area V4 by systematically mapping feature selectivity of single neurons. While cortical columns have been established as a fundamental organizing principle in early visual areas, characterizing such structural-functional relationships in higher visual areas remains challenging due to the increasing complexity of neuronal feature preferences. The digital twin approach revealed distinct functional groups of neurons that share preferences for complex features such as curves, textures, and eye-like patterns. We found that neurons recorded along the same cortical column exhibited similar feature selectivity, providing evidence for a columnar organization of visual processing in area V4.

Finally, in the third project, detailed in chapter 2.3, we developed a standardized benchmark platform for digital twins of mouse primary visual cortex, facilitating systematic and standardized comparison of neural predictive models. We established a comprehensive dataset comprising thousands of neurons responding to naturalistic stimuli, coupled with evaluation tools and metrics to enable rigorous model comparisons. The benchmark platform serves as an open resource for tracking progress in neural response prediction, promoting collaborative advancement through competitive model development. Through standardized evaluation metrics, the platform enables direct comparison of different modeling approaches, accelerating progress in understanding visual processing mechanisms in the mouse brain.

Additionally, I have contributed to numerous other projects which are related, but not a part of this thesis. In these other works, we demonstrated the generalization capabilities of digital twins across different mice [169], enhanced model robustness through co-training on macaque V1 activity and object classification tasks [170], and improved MEI naturalism using diffusion models in macaque V4 [171]. We further evaluated how deep neural networks trained on diverse computer vision tasks predict responses in macaque V1 and V4 [172], extended neural prediction benchmarks to natural videos [173], and leveraged digital twins for functional cell type clustering [174]. Finally, we also investigated the fine-scale organization of orientation-selective subregions within individual mouse V1 neurons [175].

Publications included in this thesis

- [1]: Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob

Reimer, Fabian H Sinz, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, 2022

- [2]: Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F. Burg, Santiago A. Cadena, Zhiwei Ding, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Kaiwen Deng, Yuanfang Guan, Yiqin Zhu, Kaiwen Xiao, Xiao Han, Simone Azeglio, Ulisse Ferrari, Peter Neri, Olivier Marre, Adrian Hoffmann, Kirill Fedyanin, Kirill Vishniakov, Maxim Panov, Subash Prakash, Kishan Naik, Kantharaju Narayanappa, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. Retrospective on the sensorium 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 314–333. PMLR, 28 Nov–09 Dec 2022. URL `https://proceedings.mlr.press/v220/willeke22a.html`

- [3]: Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. May 2023. doi: 10.1101/2023.05.12.540591

Other publications

- [169]: Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. In *International Conference on Learning Representations*, 2021

- [171]: Paweł A. Pierzchlewicz, Konstantin F. Willeke, Arne F. Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, Andreas S. Tolias, and Fabian H. Sinz. Energy guided diffusion for generating neurally exciting images. May 2023. doi: 10.1101/2023.05.18.541176. URL `https://doi.org/10.1101/2023.05.18.541176`

- [175]: Jiakun Fu, Konstantin F. Willeke, Pawel A. Pierzchlewicz, Taliah Muhammad, George H. Denfield, Fabian Hubert Sinz, and Andreas S. Tolias. Heterogeneous orientation tuning across sub-regions of receptive fields of v1 neurons in mice. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4029075. URL `https://doi.org/10.2139/ssrn.4029075`

- [172]: Santiago A Cadena, Konstantin F Willeke, Kelli Restivo, George Denfield, Fabian H Sinz, Matthias Bethge, Andreas S Tolias, and Alexander S Ecker. Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. *bioRxiv*, page 2022.05.18.492503, May 2022

- [173]: Polina Turishcheva, Paul G. Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F. Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S. Ecker. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. 2023. doi: 10.48550/ARXIV.2305.19654. URL https://arxiv.org/abs/2305.19654

- [170]: Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 739–751. Curran Associates, Inc., 2021

- [174]: Max F Burg, Thomas Zenkel, Michaela Vystrčilová, Jonathan Oesterle, Larissa Höfling, Konstantin Friedrich Willeke, Jan Lause, Sarah Müller, Paul G. Fahey, Zhiwei Ding, Kelli Restivo, Shashwat Sridhar, Tim Gollisch, Philipp Berens, Andreas S. Tolias, Thomas Euler, Matthias Bethge, and Alexander S Ecker. Maximally discriminative stimuli for functional cell type identification. 2024. URL https://openreview.net/forum?id=9W6KaAcYlr

2 Results

2.1 State-dependent pupil dilation rapidly shifts visual feature selectivity

This chapter is based on the following publication:

- Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H Sinz, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, 2022

2.1.1 Motivation

An animal's behavior and internal state powerfully shape how the brain processes information, allowing for flexibility in cortical processing depending on the behavioral context the animal is in [164–168]. During active behavioral states, marked by changes in pupil size [176] and movement [177], brain activity is both elevated and decorrelated, which improves signal-to-noise ratio and makes it easier to distinguish behaviorally-relevant stimuli [176–182]. Although the response strength changes with the behavioral state, the basic tuning of visual neurons (e.g., preferred orientation) usually remains stable across quiet and active states [167, 177, 180]. In this work, we examine how behavioral states change the tuning of mouse cortical neurons to colored natural images, moving beyond simplistic parametric stimuli. We included colored stimuli, as the color domain of visual input is crucial for natural behaviors of mice such as predator detection, prey hunting, and analyzing self-movement [183–185].

We investigated the link between the behavioral state and changes in neural tuning by combining calcium imaging of the mouse primary visual cortex with deep neural network modeling. Our model extends recent work [85, 144, 169, 186], by accurately predicting how populations of neurons respond to visual stimuli by including behavioral variables in our deep neural network models, which allowed us to perform *in-silico* experiments to understand neural tuning in the context of the animal's behavioral state. We then tested the *in-silico* model predictions by performing closed-loop experiments [144, 145] and by creating naturalistic-inspired stimuli and validating them in an independent cohort of mice that had no prior exposure to the experimental paradigm.

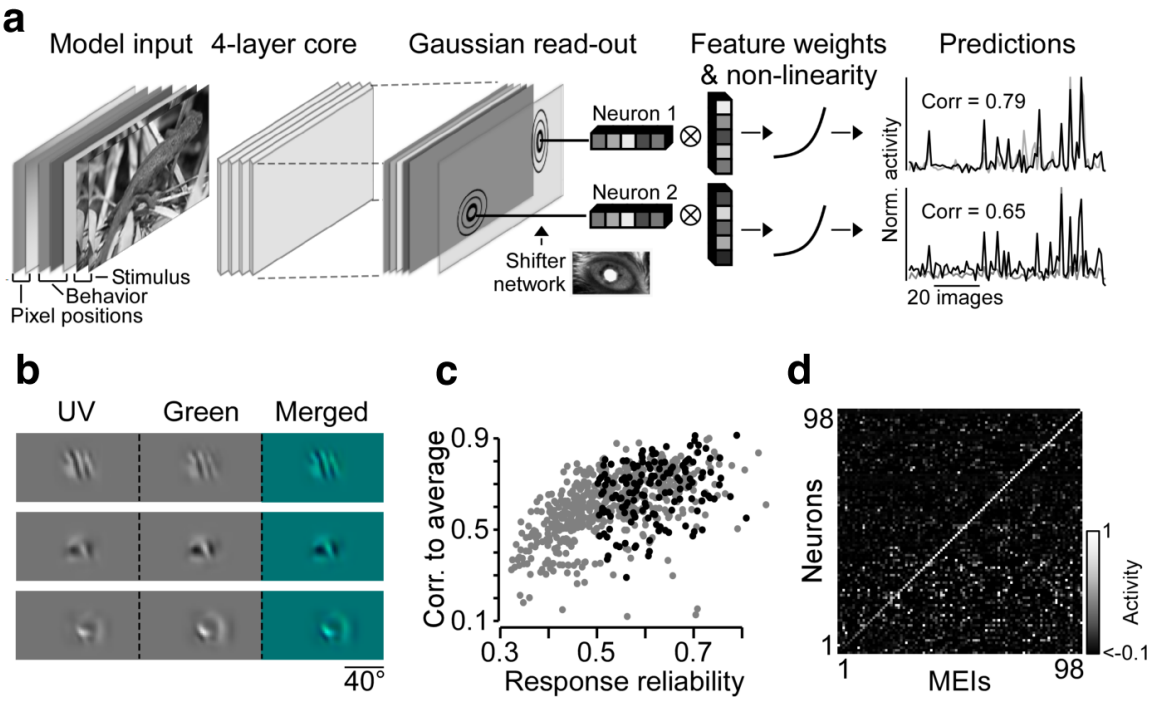


Figure 2.1: **Deep neural network model captures color tuning properties of mouse V1. a,** Our model was trained to predict neuronal responses of neurons in primary visual cortex of mice using visual input (images consisted of two color channels: UV and green) and behavior (pupil size, change of pupil size, movement speed) as inputs. We trained the model end-to-end following a core-readout architecture, whereby general features are learned by the core, which a neuron-specific readout transforms into a scalar response for a given stimulus. The right-hand graph shows how the model's predictions (black) accurately match the average responses of two example neurons (gray) to a set of 90 held-out images. **b,** Maximally exciting images (MEIs) of three example neurons generated from the predictive model. MEIs were optimized jointly for the two color channels. **c,** We plotted each neuron's response consistency to natural images against how well our model predicted its average response. Neurons chosen for further experimental testing (i.e. closed-loop experiments) are marked in black. **d,** Results of an closed-loop experiments, showing the confusion matrix for the neuronal responses of all neurons to all MEIs. The diagonal shows the neurons' response to their own respective MEI, indicating that most neurons had their strongest response to their own MEI.

2.1.2 Results and Synopsis

We discovered that color tuning in mouse V1 neurons shifts towards greater UV sensitivity during active states, especially for neurons representing the upper visual field. Using drugs to manipulate pupil size, we were able to show that pupil dilation alone is both necessary and sufficient to cause this change in color selectivity. We also found that pupil dilation increases light intake enough to switch visual processing from rod-dominated (color-blind) to cone-dominated (color-sensitive), even at constant light levels. Finally, we showed that an increased UV sensitivity during active states could help mice better detect overhead predators against the UV-rich sky.

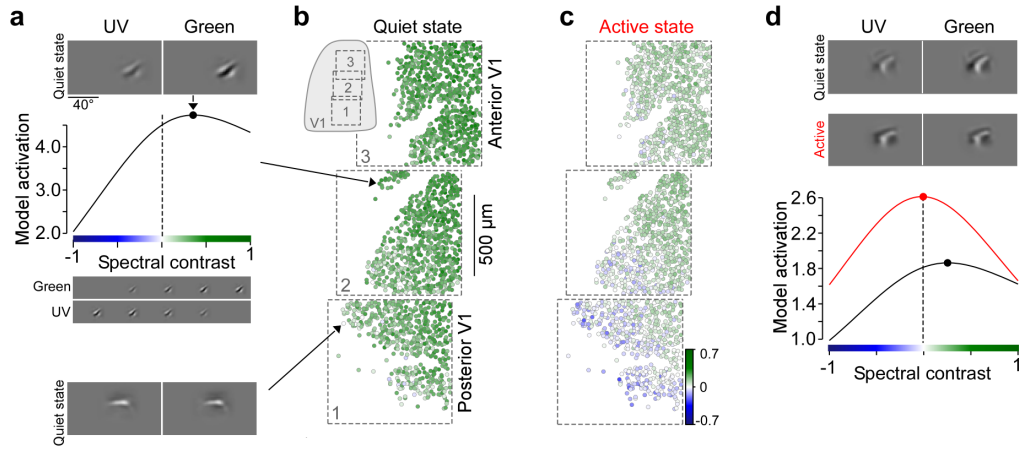


Figure 2.2: **The animal's behavioral state changes the color tuning of optimal colored images. a,** Example MEI of a neuron in mouse primary visual cortex, optimized for a quiet behavioral state (characterized by a small pupil size and no locomotion speed). MEI stimuli for the same neuron with varying spectral contrasts are plotted below - from all contrast in the UV channel (left) to all contrast in green image channel (right), with the predicted activity of the target neuron shown for all contrast levels. The dotted line shows an MEI with no color preference (spectral contrast of 0). The neuron's preferred color contrast is marked by the arrow at the top of the curve. The very bottom shows the MEI of a different neuron and its corresponding anatomical location within primary visual cortex in (b). **b,** We recorded 1,759 neurons along the back-to-front (posterior-anterior) axis of mouse V1 (3 scans, 1 mouse). The neuron color shows the spectral contrast of their optimal image (MEI) in a quiet state. The top inset shows the scan locations within V1. **c,** Same as (b), but using MEIs optimized for an active state (both pupil size and locomotion at their highest levels). We observed a large change in color contrast across the entire V1 population from the quiet state (b) towards more UV sensitivity. **d,** Example MEIs are shown for quiet (black) and active (red) states. The graph below plots the corresponding color tuning curves, indicating an overall increased activity for this state, as well as a shift of color selectivity towards UV.

Closed-loop experiments for colored stimuli in mouse V1. We built a deep neural network to model the recorded neurons (Fig. 2.1a). This model learned to predict neuron responses based on the input of natural images and the animal's behavior. We generated colored images (MEIs) that maximally activated individual neurons and used these to confirm that our model accurately captures how color images affect mouse V1 neurons (Fig. 2.1b). For most neurons, the UV and green channels of the MEI were similar, suggesting that true color opponency is rare with our type of images (Fig. 2.1b). To directly test how well our model predicted real neural responses, we selected MEIs from a subset of neurons (Fig. 2.1c) and presented these images to the mice the following day. These MEIs strongly activated their corresponding neurons, further validating our modeling approach (Fig. 2.1d).

Behavioral state changes color tuning. To analyze how behavior changes color tuning, we used our trained CNN model to perform detailed in-silico experiments. Focusing on two common states – quiet (small pupil, no movement) and active (large pupil, movement) – we optimized MEIs to maximally activate each neuron for both behavioral states. We then systematically varied the color contrast of this image

to create a color tuning curve (Fig. 2.2a). In both states, neurons' preferred spectral contrast changed systematically along the posterior-to-anterior axis of V1 (Fig. 2.2b). UV sensitivity increased towards posterior V1, aligning with the distribution of color-sensitive cells in the mouse eye. However, in the quiet state, nearly all neurons preferred a green-biased image, even at the posterior end of V1 where the eye is most UV-sensitive. During active periods, neurons' color tuning consistently shifted towards higher UV sensitivity (Fig. 2.2b-d), while model-predicted activity levels also increased overall (Fig. 2.2d). Importantly, the overall image structure that excited neurons remained similar across states (see example in Fig. 2.2d).

Pupil dilation causes shift in color selectivity. We then investigated the mechanism behind the observed color tuning shifts in mouse V1. The behavioral state affects brain activity via neuromodulators released throughout the visual system, but it also changes the amount of light entering the eye by altering pupil size. Could either mechanism explain our results? To isolate the effect of pupil size, we used eye drops to pharmacologically dilate or constrict the pupil. We recorded V1 responses from the same animal to natural images in these altered states and trained separate CNN models on these data. Remarkably, dilating the pupil with atropine was enough to shift color tuning towards higher UV sensitivity, even if the animal remained quiet and still. This shift was evident in MEIs optimized for the dilated versus normal pupil size for a quiet brain state.

Next, we temporarily constricted the pupil with carbachol drops to see if pupil dilation is necessary for the color shift normally seen in active states. We found that the typical increase in V1 activity during active states was still present, suggesting neuromodulation was unaffected. However, with the pupil constricted, neurons had a stronger preference for green stimuli during quiet periods and, crucially, the UV shift during active periods was absent. This means that while neuromodulation may play a role, it is not enough to cause the tuning shift. We then investigated if the amount of light reaching the retina can be solely responsible for the shift in color tuning. We estimated that the pupil size changes we observed would cause a large increase in light reaching the photoreceptors, which changes light levels enough to shift the mouse visual system from rod-dominated (green sensitive) to cone-dominated (UV sensitive). To test this directly, we dimmed our stimulus by 1.5 orders of magnitude while keeping the pupil dilated with atropine, mimicking a rod-dominated state. As expected, V1 neurons in this low-light condition preferred greener images. Taken together, our results provide strong evidence that pupil dilation during active states causes a dynamic shift from rod-dominated to cone-dominated vision, which in turn causes the observed changes in color tuning.

Shift in selectivity is behaviorally relevant. Finally, we investigated whether the shift in color tuning during active states improves how the mouse brain processes natural images. We tested this directly by training mice to recognize objects displayed in either UV or green light (Fig. 2.3b). As they viewed these, we recorded activity in the posterior part of V1. We then trained a non-linear SVM (Fig. 2.3a)

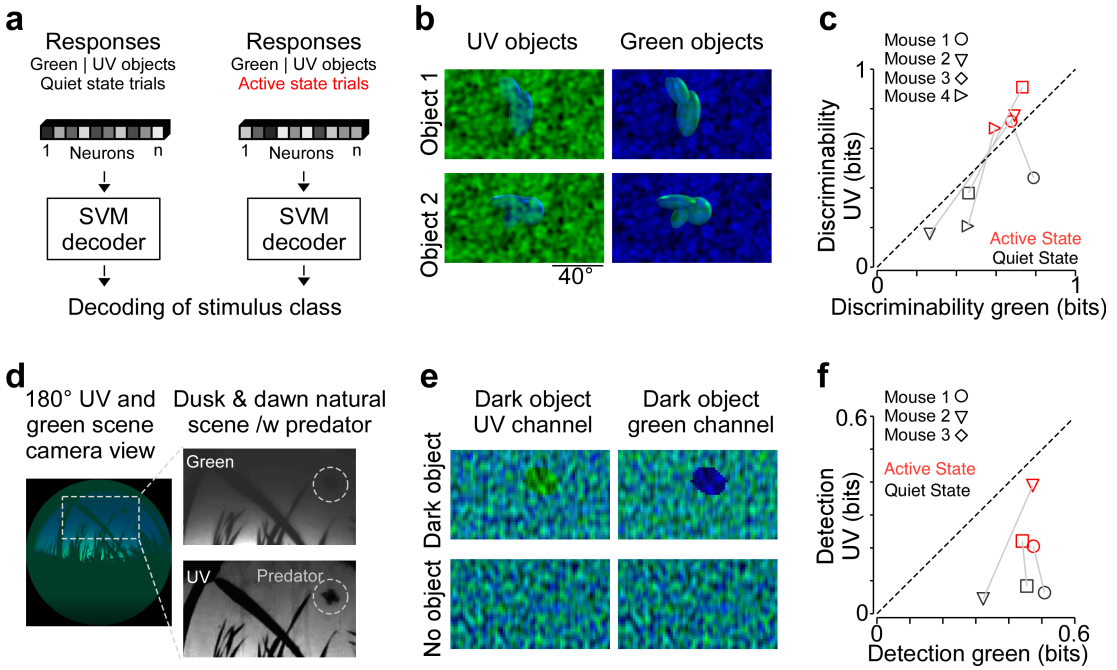


Figure 2.3: **Behavioral relevance of shift of color preference with a change of behavioral state. a,** Overview of the decoding paradigm. We trained a non-linear support vector machine (SVM) to decode whether a mouse was seeing a green or UV object, using brain activity recorded during either quiet (small pupil) or active (large pupil) behavioral states. **b,** Stimuli for the object discrimination task: Example 5-second video stimuli showing UV and green stimuli with added noise. **c,** Decoding results of discrimination task show the model discriminability of green vs. UV objects using the activity of 200 randomly chosen neurons for each mouse. UV vs. green discrimination was better during active trials (red) compared to quiet (gray). **d,** Natural scene photographed with a mouse-eye camera, showing a drone (mimicking a predator). Right-side images show just the UV or green part of the left image; the drone is circled, highlighting the easier detection of the drone in the UV image channel. **e,** We designed simplified stimuli inspired by (d) for an object detection task: a dark object in either the UV or green channel on top of noise. We also presented images with no object present. We then again trained an SVM to predict whether an object was present or absent. **f,** Similar to (c), but showing how well the model could detect the dark UV vs. green objects from (e). Detection performance increased for UV objects when the animals were in the active state.

to decode the object identity from the activity of the entire neuronal population. Consistent with previous studies, decoding improved when the mouse was active, which can be readily explained by increased neuronal firing rates and thus a higher signal-to-noise ratio. But more importantly, the improvement was larger for UV objects than green objects, matching our observation that neurons become more UV-sensitive when the animal is active (Fig. 2.3c). Why might this UV sensitivity boost be useful? Mice are most at risk from aerial predators at dawn and dusk, when the sky is much brighter in UV light than in green. A heightened UV sensitivity during active states could therefore help mice more readily spot predators against the background of the UV-rich sky (Fig. 2.3d). We tested this idea with simplified stimuli: by presenting either visual noise or noise plus a dark object in either the UV or green part of the image (Fig. 2.3e). We found that, as predicted, the brain's ability to detect a dark object against the UV background significantly increased when the

mouse was active (Fig. 2.3f). Detection of a dark object against a green background also improved, but less so overall (Fig. 2.3f). This suggests that the population-level shift towards UV sensitivity could be a survival adaptation, helping mice detect predators against a bright UV sky.

2.1.3 Discussion and Outlook

Our work uncovers a novel mechanism for mice to dynamically tune their visual system: state-dependent pupil dilation that optimizes visual selectivity of the mouse visual system for behaviorally relevant tasks. The impact of behavior and internal state on how the brain processes sensory input has been studied for decades across many species, showing that response modulation of visual neurons can lead to improved performance in behavioral tasks [167, 177, 179, 187–191]. In some reported cases, however, the tuning of sensory neurons themselves also shifts via behavioral modulation [165, 166, 192, 193]. In this work, we show that mouse visual neurons change their color tuning depending on the animal's behavioral state. Our results suggest that the shift toward higher UV sensitivity during active states is indeed behaviorally relevant by helping mice detect predators against the UV-rich sky. These results are in line with prior work, which implicated that UV vision evolved as an adaptation for predator and prey detection in different natural environments [194, 195].

Previous studies linked behavior-dependent visual changes to brain chemicals like acetylcholine and norepinephrine, which are elevated in active behavioral states [196, 197]. In contrast, our work emphasizes that the related dynamic changes in pupil size also play a key role. We propose that changes in pupil size shift the balance between rod and cone photoreceptors, which, based on their different color sensitivities, are able to rapidly change the spectral sensitivity in visual cortex. Pupil changes linked to behavioral states exist across most vertebrates [198], which suggests that pupil-mediated tuning may be a general mechanism for quickly optimizing the visual system for different tasks, utilizing a rapid shift between rod- and cone-driven regimes and thereby changing the visual feature selectivity. We believe that our results provide the first mechanistic explanation for the longstanding question of why pupil dilation with behavior and the internal state is such a ubiquitous phenomenon.

2.2 Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization

This chapter is based on the following publication:

- Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. May 2023. doi: 10.1101/2023.05.12.540591

2.2.1 Motivation

Understanding how the brain works means understanding how its structure relates to its function. A key discovery in neuroscience was the concept of 'cortical columns': groups of neurons that all share similar response characteristics [14] stacked in vertical columns. This organization has been proposed as a fundamental building block of the cortex [199, 200], with ample evidence for columnar organization for visual features such as ocular dominance, spatial frequency, orientation selectivity, motion direction, and motion disparity [201–203]. The advantages of such an organization seem obvious: through minimizing synaptic distances, computations within a layer are facilitated [204–206]. However, it remains a major challenge to characterize the neuronal response selectivity in higher visual areas as features become increasingly complex, with response preferences found in natural images such as 2D and 3D shape, texture, objects, and faces [134, 207–216]. This high-dimensional space is difficult to explore systematically, making it difficult to link a neuron's preferences to its cortical location. If neurons with similar preferences are indeed grouped together topographically, this would point towards a general organizing principle in the cortex. Using deep learning-driven models of visual neurons, it is now becoming possible to much more accurately predict the visually evoked activity of neurons along the cortical hierarchy in response to arbitrary images [85, 159, 169, 211, 217]. This lets us perform unlimited virtual experiments, identifying a neuron's preferred stimulus [1, 146, 149, 211, 217–219] or mapping other characteristics of the visual tuning functions such as invariances and contextual modulations [147, 148, 153]. As described earlier, the model predictions of these highly predictive models (*digital-twins*) can then be verified in the animal with *closed-loop* experiments - showing the model-generated stimuli back to the animal in an experimental closed-loop setting. Here, we adapted the closed-loop technique to study visual area V4 in macaque monkeys to systematically characterize the visual selectivity of single neurons in order to relate the neurons' functions to their anatomical location in search of a columnar organization of response types within higher-level visual cortex.

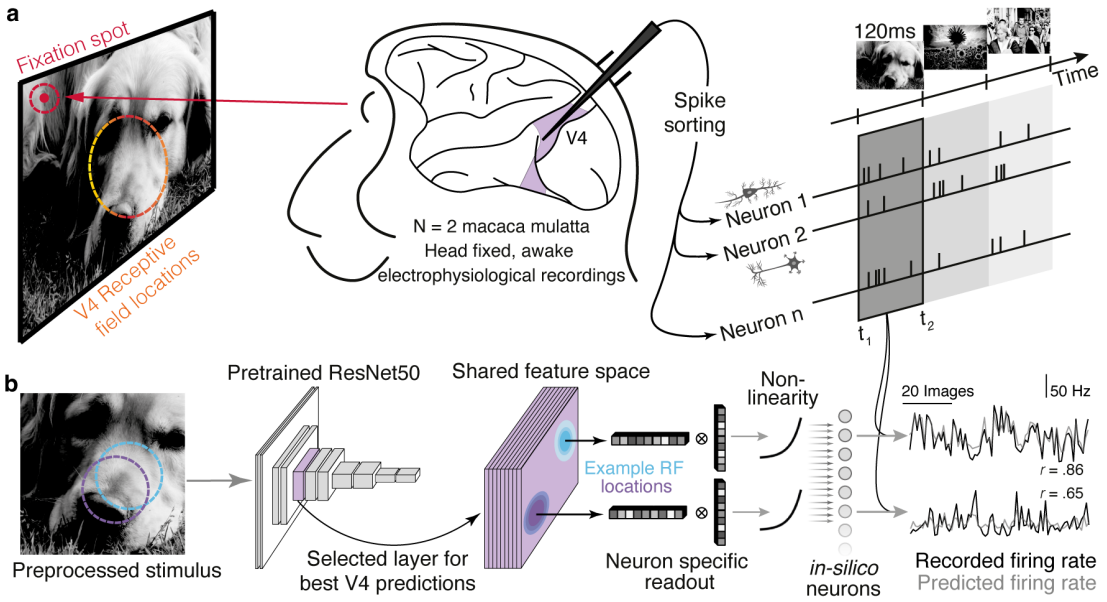


Figure 2.4: **Deep neural networks model the tuning properties of single neurons in primate area V4. a,** Experimental setup: Two awake, head-fixed macaque monkeys were presented with static natural images. While the animals were fixating such that the neurons' receptive fields were in the center of the monitor, 15 images were shown back-to-back for 120ms per image in each trial, for about 1,000 trials per recording. Neuronal activity was recorded using 32-channel laminar silicon probes. After analyzing the recordings, we isolated the activity of individual V4 neurons. **b,** We trained a neuron-specific model using pre-processed images (100x100 pixel crops) and recorded neuron responses. This model builds on a ResNet50 network pre-trained for image classification. We selected the ResNet50 layer that best predicted V4 activity and calculated neuron responses using Gaussian readout followed by a non-linear function. The right-side graph compares the averaged responses of two sample neurons (gray) with model predictions (black) for 75 test images, showing a high correlation.

2.2.2 Results and Synopsis

Closed-loop experiments verify optimal stimuli for single neurons in V4 To understand how monkey V4 neurons respond to natural images, we combined large-scale recordings with deep neural network modeling. We showed thousands of natural images to awake monkeys while recording the spiking activity of V4 neurons using 32-channel probes that spanned the depth of the cortex (Fig. 2.4a). The monkeys were trained to maintain their gaze on a small spot offset from center, ensuring the images were centered over the neurons' receptive fields. Through post-hoc analysis, we isolated the activity of 1,224 individual V4 neurons in response to more than 10,000 images. To predict and analyze V4 responses, we used a deep convolutional neural network pre-trained for image classification [71, 93] to extract complex image features relevant for V4 (Fig. 2.4b). We then trained a readout [169] to predict the response of each neuron to a presented image. Using the CNN as a 'digital twin' of the V4 population, we generated images (MEIs) designed to excite individual neurons maximally (Fig. 2.5a). These experiments revealed

highly diverse preferences across neurons (see examples in Fig. 2.5b), including textures, curves, and edges, resembling other stimuli that were used in previous studies of V4 with carefully hand-designed stimuli [134, 135, 220]. To confirm that our generated MEIs strongly excite V4 neurons, we developed a closed-loop experimental paradigm to verify that the MEIs indeed elicit high activations in the biological neurons (Fig. 2.5c). In a 'generation' session, we first recorded V4 neurons while showing the monkey natural images and subsequently trained the model on pairs of neuronal responses and natural images. Based on the model's predictions, we chose six neurons with the highest performance on a held-out test set and generated MEIs for these neurons. After generating the MEIs, in a 'verification' experimental session, we showed these MEIs, along with matching control images, to the monkey while continuously recording from the same neurons. Control images were the most activating natural image patches selected from a large set of natural images for the target neurons. Despite their qualitative similarity to natural image patches, the MEIs indeed consistently produced a stronger response than control images, and they were more exciting than the MEIs or control images for other neurons, demonstrating the validity of the closed-loop paradigm for primate area V4 (Fig. 2.5d).

MEIs reveal columnar organization in primate area V4 We noticed that MEIs from the same recording session seemed remarkably similar, compared to MEIs from other sessions (Fig. 2.5e). While we found MEIs representing diverse features like lines, curves, and grids, within a given recording session, there was usually less variation. For example, most neurons in one session preferred a fur-like texture pattern (Fig. 2.5e). Unlike V1 neurons, which align along well-defined dimensions like orientation or frequency, V4 neurons' intricate MEI patterns complicate direct comparisons of their tuning similarities. To tackle this, we implemented an unsupervised deep learning approach [221–223] to create a two-dimensional embedding space of MEI images, emphasizing their image feature similarities. This approach trains a model to bring closer the embeddings of variations of the same image while separating those of visually distinct images. Through data augmentation, including random rotations, shifts, and scaling, the model learns to identify images as similar if they only differ through these transformations. These augmentations retain critical attributes of images, ensuring, for example, that a rotated or scaled feature like an edge or a texture is still recognized as the same feature when rotated or scaled. To train this self-supervised model, we generated 50 MEIs for each neuron and used each MEI as a positive example such that for a single MEI, we obtain two randomly augmented MEIs, which should then have a small distance in embedding space, and a large distance compared to all other MEIs. After training the model on the entire population of MEIs, we obtained a two-dimensional embedding space with similar MEIs positioned nearby, and more dissimilar MEIs having a large distance in the embedding space. After analyzing the structure of the embedding space, we found that nearby MEIs in the embedding space were qualitatively similar (Fig. 2.6) and that MEIs generated from the same neurons were closely clustering together. Most

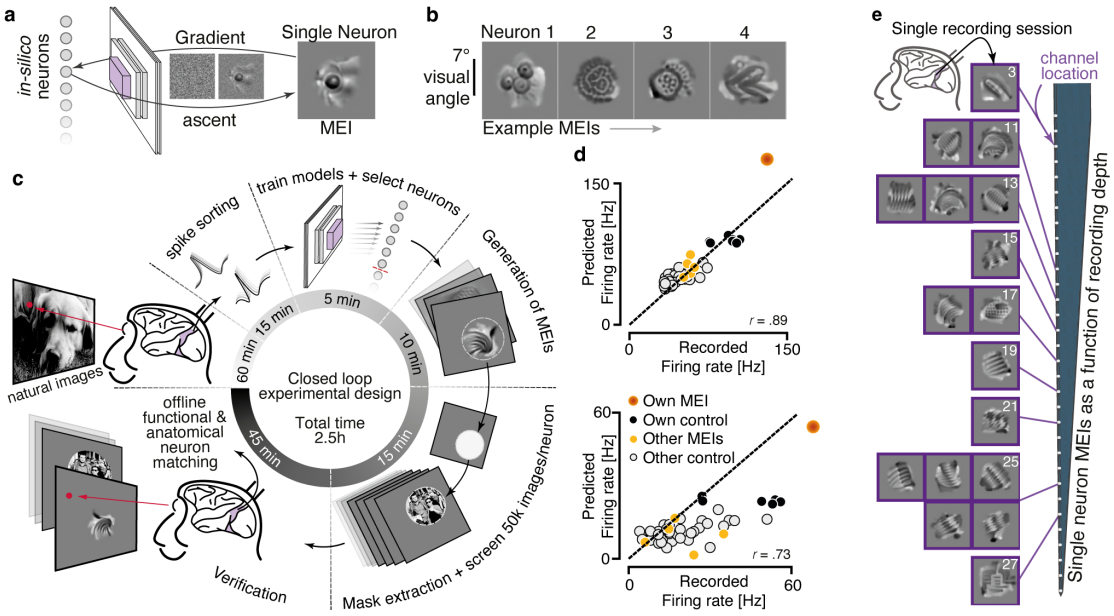


Figure 2.5: **A closed-loop experiment verified the model's ability to generate optimal stimuli for single V4 neurons. a,** We optimized the most exciting image (MEI) for each V4 neuron using gradient ascent. **b,** MEIs of four example neurons. **c,** Schematic illustrating the closed-loop experimental paradigm for recordings in primate V4. First, we recorded brain activity in a "generation session." Then, we built a model, chose six individual neurons, and generated MEIs for these selected units. We also searched for the highest activating natural images to act as control stimuli. Finally, we showed MEIs and control images to the monkey in a "verification session" while still recording from the same neurons. We carefully checked in offline analyses that we were recording from the same neurons across the entire duration of the closed loop experiment. **d,** Comparison of predicted vs real recorded neuronal activity from two sample neurons in response to their own MEI and control images, as well as MEIs/controls of other neurons. **e,** MEIs of 17 neurons recorded in a single experimental session, arranged such that the recording location of each neuron in depth along the recording electrode can be observed. Number insets in the MEIs indicate the channel (along a 32-channel depth probe), with higher channel numbers meaning deeper recording depth within area V4.

importantly, we also found that neurons from the same recording session indeed shared similar features as measured by their smaller mean pairwise distances in the embedding space against a shuffled control. These findings collectively support the hypothesis that V4 neurons in monkeys are organized in a columnar manner, where neurons with identical tuning align vertically.

MEIs reveal novel clustering of response modes Inspired by the structure of the learned embedding space, we asked whether V4 neurons cluster into functional groups based on their preferred features. We used a clustering algorithm [224] to group MEIs based on similarity, resulting in 17 distinct groups (Fig. 2.6). MEIs within a group were strikingly similar (e.g., features resembling 'eyes' in group 11). MEIs from different groups, especially those far apart in the embedding space, were quite distinct (e.g., texture- or grid-like in clusters 8 and 4). To ensure these groups weren't an artifact of our embedding method, we also computed the similarity

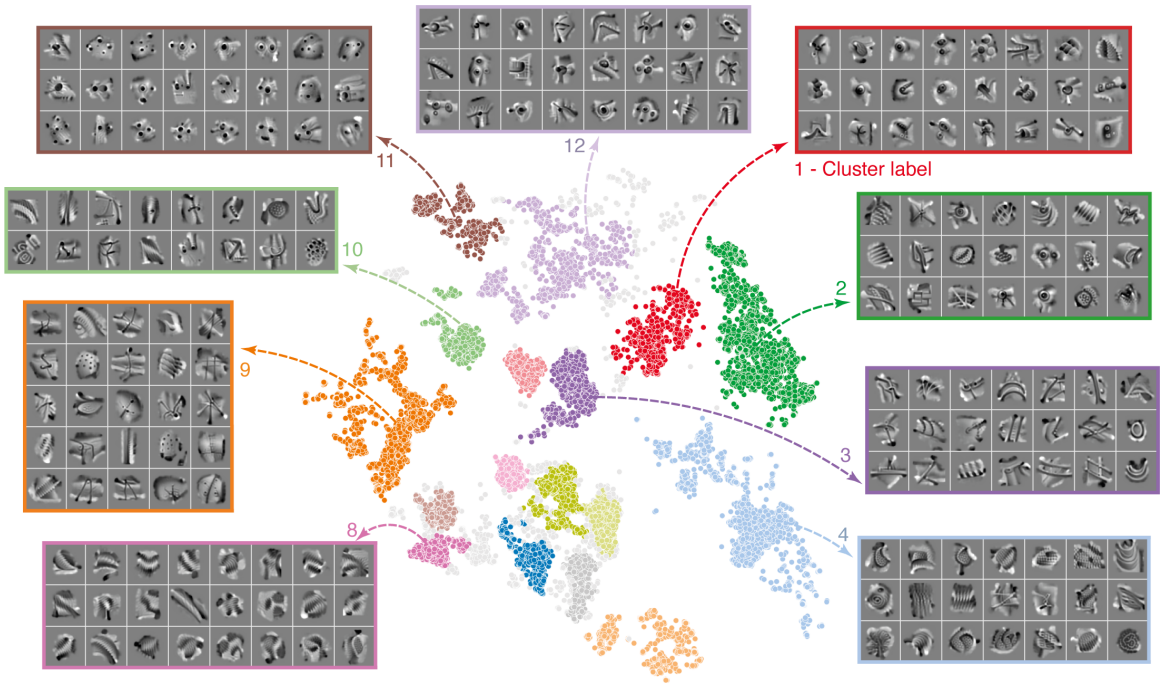


Figure 2.6: **Clustering of V4 neurons into distinct response modes.** We embedded the positions of 19,688 highly activating MEIs (from 889 neurons) in a 2D similarity space. MEIs were color-coded based on their cluster assignment using a hierarchical clustering algorithm. We found 17 distinct clusters, which we refer to as response modes. For nine of the 17 clusters, we display a random selection of neurons belonging to that cluster, highlighting the similarities of MEIs within the clusters.

of MEIs by comparing how the model's predicted population response would differ. MEIs within a group produced more similar model responses than MEIs across groups, confirming that our clusters represent real functional distinctions. Remarkably, these V4 groups closely resemble the feature visualizations of mid-level layers of deep neural networks trained on object recognition [130], which is a promising framework to develop testable hypotheses about V4 neuronal properties using features of artificial vision systems as a starting point.

2.2.3 Discussion and Outlook

Here, we adapted closed-loop experiments [1, 144, 145] to study cortical columns in visual area V4 in monkeys. We built a deep neural network model of over 1,200 V4 neurons and used this model to perform closed-loop experiments. We found that neurons recorded along the depth of a single electrode shared similar feature selectivities. Remarkably, we uncovered groups of V4 neurons, akin to functional cell types, that share preferences for complex features such as eyes, fur patterns, grids, or curves. These feature selectivity groups in V4 neurons can also be found in deep neural networks trained solely trained on object recognition [129, 130, 143], which is compelling evidence for shared computational principles in biological and artificial vision.

Previous work has used similar 'digital twin' approaches to study the mouse visual

system [1, 148, 149, 175, 217], as well as macaque V4 [145]. Our MEIs include the textures seen in these studies as well as shape-like features like corners, curves, and even eye-like patterns which weren't shown in prior V4 studies [145]. This difference likely arises from the fact that we were able to study single neurons, rather than multi-unit activity that may have resulted in mixed preferences and more texture-like MEIs. Our results are consistent with classic V4 studies that used simpler parametric stimuli demonstrating selectivity for shapes and especially curvature [135, 207, 208, 212, 214–216, 225, 226]. Because we recorded across cortical layers using a laminar depth probe, we were able to study the vertical organization of V4. The existence of columns in V4 has been debated, particularly for features that are challenging to describe parametrically such as orientation [209, 227]. Many studies have shown using imaging techniques that V4 is organized into areas such as 2D versus 3D shape [228], curvature [229], spatial frequency [230], and even motion direction [231]. These studies found clustering of neurons with selectivities within certain categories, and how these selectivities change across the cortical surface. A recent study [232] used Neuropixel probes to study columns in area V4 and found no evidence of similar neuronal selectivities along the cortical depth and therefore concluded that there is no columnar organization in area V4. However, this approach used a classical and thus restricted set of stimuli such as a fixed set of shapes and textures. It is therefore possible that the tuning characterization with this set of stimuli is not able to capture the feature tuning of V4 neurons to a sufficient degree to reveal columnar organization. Our approach overcomes this by directly identifying the most exciting image for each neuron as a proxy for feature selectivity. Our 'digital twin' approach thus allowed us to compare MEIs without relying on hand-designed features and to quantify the similarities of MEIs using features of the MEIs directly. We found that neurons recorded simultaneously across layers had more similar MEIs than neurons chosen at random. However, this effect was not equally strong across all recordings. This variability could arise from imperfect alignment of our recording probes or from a more complex arrangement of V4 columns, similar to the pinwheel structure of orientation tuning in V1.

This work highlights the power of deep learning for uncovering principles of neuronal computation in the brain. Our results suggest that cortical columns are not limited to primary sensory areas like V1 but also exist in mid-level visual areas like V4. Since neurons within a column are more likely to be connected, this organization likely facilitates the development of more complex tuning properties through local interactions. While we focused on MEIs to demonstrate columnar similarity, this is an incomplete picture. To understand computation within a column, it becomes a future task to characterize the full tuning function of its neurons along the entire range of excitability for each neuron.

2.3 Retrospective on the SENSORIUM 2022 competition

This chapter is based on the following publication:

- Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F. Burg, Santiago A. Cadena, Zhiwei Ding, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Kaiwen Deng, Yuanfang Guan, Yiqin Zhu, Kaiwen Xiao, Xiao Han, Simone Azeglio, Ulisse Ferrari, Peter Neri, Olivier Marre, Adrian Hoffmann, Kirill Fedyanin, Kirill Vishniakov, Maxim Panov, Subash Prakash, Kishan Naik, Kantharaju Narayanappa, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. Retrospective on the sensorium 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 314–333. PMLR, 28 Nov–09 Dec 2022. URL `https://proceedings.mlr.press/v220/willeke22a.html`

2.3.1 Motivation

In computer vision, benchmarks like the ImageNet Challenge helped launch the deep learning revolution [69, 233]. With the advent of deep neural network models for neural system identification [60, 86, 97–100, 102, 144, 155–160, 234], neuroscience can similarly benefit from large-scale benchmarks to compare, develop, and refine new models. Neural system identification methods can reveal previously undetected nonlinear processes in sensory processing by developing models that better account for stimulus-response relationships. Higher model performance, quantified by explained variance in neural responses to stimuli, suggests the capture of computational principles that simpler, parametric models missed [46, 107, 235]. Through iterative model refinement and validation with in-vivo experiments, these models can help uncover the computational mechanisms underlying sensory processing in the brain. Standardized benchmarks can provide a controlled framework for comparing different models' performance using consistent metrics and shared datasets. This systematic evaluation process drives competitive model development, where sequential refinements by multiple research groups collectively advance the field's capabilities. In this work, we present a benchmark with the explicit goals of (1) providing a large-scale dataset of thousands of neurons in mouse primary visual cortex in response to naturalistic stimuli to be used as a reference dataset to measure progress, (2) creating an open benchmark platform to track progress, and (3) providing evaluation tools and metrics to allow for straightforward model comparison.

However, benchmarks in neuroscience are not novel. For example, Brain-Score [91, 92] compares how well artificial neural networks trained on other tasks fit various primate brain areas. The Algonauts challenge [236] focuses on predicting human fMRI responses to images and videos. The Allen Institute's mouse visual

cortex dataset [237] provides a similar large-scale reference dataset, which however is not set up as a benchmark and lacks tools to measure performance. Finally, the Neural Latents benchmark [238] is a competition centered around neural forecasting in the motor cortex, with the goal of predicting either motor behavior or neuronal responses given certain behaviors. Because these benchmarks do not satisfy the need for a benchmark for measuring progress in neural system identification models, we created the Sensorium competition to focus on building the best possible predictive model of the mouse visual cortex. We collected a massive dataset of over 28,000 neurons in seven mice, showing them thousands of natural images while monitoring behavior like locomotion speed, eye movements, and pupil dilation. Our benchmark ranked all models based on how accurately they predict responses of the visual neurons to unseen images. The Sensorium competition has two tracks: models that use image data only, and models that can also use behavioral data. As part of the 2022 NeurIPS conference, our competition ran from May to October 2022. Overall, 26 teams submitted a total of 172 models, with the winners substantially improving the previous state-of-the-art. Here, I describe the competition in detail, the winning models, and the lessons learned for future benchmarks in neuroscience.

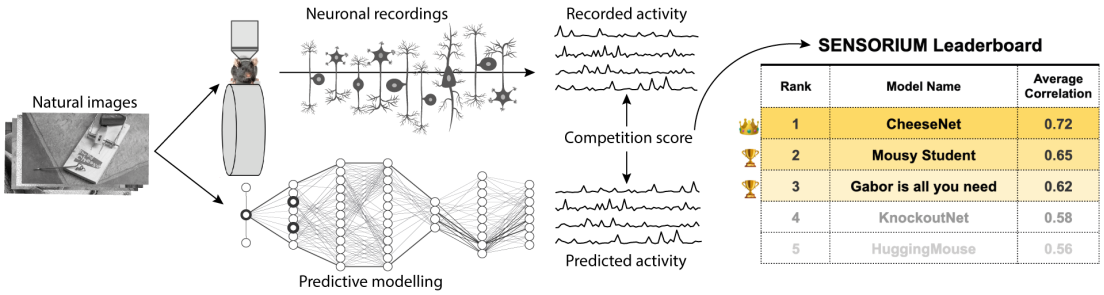


Figure 2.7: **Overview of the Sensorium competition.** We recorded and made available a large-scale dataset of neuronal activity in the primary visual cortex of mice. We invited participants to the competition to train models on pairs of natural image stimuli and their matched recorded neuronal activity of individual neurons. Based on the submitted predictions, we automatically ranked the best models and selected three winners in two competition tracks.

2.3.2 Results and Synopsis

The Sensorium 2022 competition (and ongoing benchmark) aims to find the best models for predicting how neurons respond to arbitrary natural images (Fig. 2.7). When the competition began, we released a large-scale training dataset where both images and responses were presented. Additionally, we created a set of competition test images that were shown to two animals, for which we did not make the neural responses public. These test images were split into 'live' and 'final test' sets. The 'live test' set was used to rank models on a public leaderboard during the contest. The 'final test' set was used only at the very end to prevent teams from overfitting their models on the publicly available test data. This split let us give feedback during the competition without compromising the final results.

The competition has two tracks, SENSORIUM and SENSORIUM+, that use the exact same image sets but have different goals and datasets: Participants predict the average activity of 7,776 neurons in one mouse across 10 repeats of 200 unique test images. For the SENSORIUMtrack, the only input for predicting the responses are the images themselves. This forces models to focus purely on the image-driven response, which is the standard approach for most neural modeling work. In the SENSORIUM+track, participants predict the responses of 7,538 neurons to the same 200 images, but given both the image and full behavioral data. This track aims for more accurate models, as a significant part of neural variability comes from the animal's behavior and internal brain state, as shown in the first work of this thesis.

Our dataset includes recordings from seven mice, with a total of over 28,000 neurons responding to 25,200 images. Each recording contains 6,000-7,000 image presentations. Note that due to our imaging method, the same neuron may appear at multiple depths, inflating the total count. We divided the recordings into three categories: 'pre-training', 'live test', and 'final test'. Five recordings are designated as 'pre-training'. These are provided solely for model development. They include responses to 5,000 unique images, plus 100 images shown 10 times each (the 'public test' set). The other two 'competition' recordings contain the same 5,000 training images along with an additional 100 repeated 'live test' images. We initially hid the neural responses to the 'live test' set to provide iterative feedback. These recordings also contain 10 repetitions of 100 unique 'final test' images. Responses to these were permanently hidden, used only to determine the final winner. All five 'pre-training' recordings include the behavioral data, as does the SENSORIUM+ track. We also released the anatomical location of all recorded neurons. To set a starting point, we trained two types of baseline models for each competition track: a simple linear-nonlinear model (LN) and a state-of-the-art convolutional neural network (CNN) as used in the first work presented in this thesis.

During the four-month competition, 26 teams submitted a total of 172 models. We were pleased to see that the state-of-the-art CNN baseline was outperformed by more than 15% in both tracks. The winning entries to the competition follow the overall model architecture of using a CNN-based core as a backbone and the Gaussian readout developed by [169], as provided in the baseline models. The improvements in predictive performance were largely due to changes in core architecture, training scheme, and including data augmentations of the input image data. The largest performance improvement was obtained by all winning teams through pre-training the model on all available pre-training data and carefully fine-tuning the results to the competition data. Furthermore, the winning team used distillation and large model ensembles. These changes led to substantial gains – over 15% in both tracks. It's impressive to see such improvement in just four months while the competition was active, and we are especially excited about the potential for entirely new modeling approaches. A recent publication by Li et al. [239] used the benchmark data in developing a novel model based on the novel vision Transformer [240] architecture.

2.3.3 Discussion and Outlook

This benchmark competition aims to unite computational neuroscience and machine learning communities to advance our understanding of sensory processing in the brain. Our long-term goal is to steadily expand the benchmark with new datasets, challenges, and metrics. While the initial 2022 Sensorium benchmark focuses specifically on predicting V1 layer 2/3 responses to static images, future iterations using our benchmarking framework now let us explore many possible directions, including: Moving beyond V1 to higher visual areas as well as other cortical areas, simulating the broader color range that mice can see, adding other sensory modalities, using different recording techniques such as electrophysiology, and studying different animal models, including non-human primates. Most importantly, going forward, modeling progress should not be measured purely by predictive performance, as this is not a goal in itself. Rather, interpretability techniques and other methods for scientific discoveries need to be developed and turned into metrics so that progress can be measured along this axis. Building on the success of our initial benchmark, our second iteration of the Sensorium competition [241] implemented key advances in experimental design. By transitioning to natural movies as stimuli and incorporating out-of-distribution parametric videos, we significantly expanded the complexity and ecological validity of the neural prediction challenge. This evolution in stimulus complexity provides a more stringent test of model generalization while maintaining strong community engagement. The systematic evaluation and improvement of predictive models through standardized benchmarks can accelerate progress in computational neuroscience. We encourage broad participation from the research community through participation in existing benchmarks, contributing new datasets, and developing novel evaluation metrics.

3 Discussion and conclusion

The emergence of deep learning in 2012 marked a transformative moment in computational neuroscience, particularly in modeling visual processing. Deep neural networks have become a powerful tool in understanding neural computation, establishing a productive co-development between machine learning and visual neuroscience. These models now achieve state-of-the-art performance in predicting neural responses to arbitrary stimuli and serve as *digital twins* of biological systems. This approach enables researchers to conduct extensive *in-silico* experiments that would be impractical *in-vivo*, effectively narrowing down the large experimental search space to select the most promising hypotheses to test. The digital twin framework has proven particularly valuable for investigating complex neural computations across different brain regions and states, allowing for systematic characterization of neuronal feature selectivity. This thesis builds upon these advances in neural network modeling of the visual cortex, advancing our understanding of visual processing through systematic investigation of neuronal response properties, their biological organization, and the development of standardized benchmarks for evaluating and comparing different modeling approaches.

The role of behavioral states in neural processing represents a key focus in modern neuroscience research. The first project in this thesis demonstrates how digital twin models can be used to study state-dependent visual processing in mice. While previous work established that behavioral states - such as changes in pupil size and movement - modulate neural responses across the visual system, most studies relied on simple artificial stimuli that poorly reflect natural vision. Our study extended this framework by examining how behavioral states influence the processing of colored natural images, which are crucial for survival behaviors such as predator detection. We extended digital twin models to incorporate both color processing and behavioral variables, creating predictions of neural responses and feature selectivity across different states. By generating maximally exciting inputs (MEIs) conditioned on behavioral measurements, we could systematically probe how neural tuning changes with behavioral states. These model-driven predictions were then validated through closed-loop experiments, where we presented the computed optimal stimuli while recording from the same neurons. This approach revealed that mice actively and rapidly modulate their color processing in primary visual cortex (V1) based on changes in pupil size, specifically to enhance the detection of relevant stimuli. This work extends the understanding of behavioral modulation of visual processing, showing how the visual system dynamically adapts its processing based on behavioral context.

The relationship between neural structure and function represents a fundamental question in neuroscience. The second project in this thesis addresses this relationship by investigating the functional organization of macaque visual area V4, building upon the foundational concept of cortical columns - a key organizational principle where neurons with similar response properties are arranged in vertical columns. While columnar organization is well-documented in early visual areas for properties like orientation selectivity, ocular dominance, and motion direction, establishing this principle in higher visual areas has remained challenging because of the increased complexity of neural responses. Our study leveraged digital twin models and closed-loop experiments to overcome these limitations. Our approach revealed distinct functional groups of neurons sharing preferences for specific complex features, reminiscent of functional cell types. Crucially, we found that neurons within the same cortical column exhibited similar response preferences, providing strong evidence for columnar organization in area V4. The presence of columnar structure in macaque V4 suggests that this organizational principle could extend beyond early visual areas, potentially representing a general strategy for efficient neural computation.

With the rapid development of digital twin models for various brain areas, standardized benchmarks to compare the best predictive models are becoming increasingly necessary. The third project in this thesis establishes a benchmark platform for evaluating digital twin models of mouse primary visual cortex. Our platform provides three key components: a comprehensive dataset of thousands of neurons and their responses to a large and diverse set of naturalistic stimuli, an open framework for tracking modeling progress, and standardized evaluation metrics for direct model comparison. This approach enables rigorous evaluation of different modeling strategies, from traditional linear-nonlinear models to state-of-the-art deep learning architectures. We hope that this benchmark framework drives innovation in digital twins for neuronal response prediction, similar to the advancements in object recognition in computer vision through the ImageNet [233, 242] challenge. By establishing clear metrics and providing tools for straightforward comparison, we have created an ecosystem that accelerates progress in understanding visual computation through the use of digital twin models.

Together, these projects as well as a growing literature of digital twin models of visual cortex [60, 86, 97–100, 102, 153, 155–163] demonstrate the success in predicting individual neuronal responses and extracting insight into visual processing by conducting experiments in-silico.

These results in this thesis however remain limited due to their focus on single neuron feature selectivity as a way of quantifying the computation performed by a neuron. A neuron's computational role extends far beyond simple stimulus-triggered responses - it encompasses precise spike timing and synchronization with local or population-wide activity [243]. The complexity of neural computation is evident even in primary sensory areas, where neurons in awake animals show considerable response variability to identical stimuli [244, 245], suggesting more

sophisticated computational principles than previously recognized. A particularly compelling challenge to single-neuron models comes from widespread organized spontaneous activity across brain regions, especially in humans [168, 246–249]. The development of large-scale recording techniques, such as multi-electrode arrays and calcium imaging, has revealed that neural populations carry information not detectable when examining single neurons in isolation [250–252]. These methods have uncovered complex activity patterns that go well beyond individual receptive field properties [168, 186, 252–255]. The traditional interpretation of specialized neurons, such as "face cells" [24, 256], illustrates the limitations of single-neuron approaches. The probability of finding individual neurons that encode specific faces seems implausibly low when recording from single neurons within cortical areas containing hundreds of thousands of neurons. More likely, facial recognition emerges from distributed activity across neural populations [257, 258]. Similar population-level encoding has been demonstrated for spatial information in hippocampal place cells [259]. These findings suggest that it is necessary to reconceptualize receptive fields as individual manifestations of distributed circuit computations - specific patterns of activity across many neurons responding to particular stimuli or locations. This perspective calls for broadening the scope from single neurons to understanding how neuronal groups collectively process information [243, 260] to incorporate the computational goals of larger circuits [30, 261–263], the nonlinear and adaptive properties of neurons [46, 50], and the dynamics of neural populations [250] to address the distributed nature of information processing in the brain [264, 265].

Recent developments in understanding artificial neural networks parallel the drive in neuroscience from single-neuron to population-level analysis. Early mechanistic interpretability research focused on understanding individual neurons and circuits within deep neural networks. However, this approach revealed limitations: convolutional networks for image classification also showed polysemantic units that respond to multiple concepts rather than specific features [266]. This puts a limit to how far individual neurons can be understood as dedicated detectors for specific features, either high- or low-level. Discoveries such as these align with earlier neural network theory, which emphasized distributed representations as the foundation of information processing [70, 267–269]. Supporting this view, computational neuroscience research demonstrated that mixed representations are not only common in biological brains but also provide advantages in coding efficiency [168, 270]. Thus, following the growing literature of studying neural populations and manifolds, mechanistic interpretability has evolved similarly. Most recently, the emergence of sparse autoencoders can be seen as a bridge between single neuron and population interpretability analysis. This technique decomposes network activations using sparse autoencoders [271–273], similar to sparse component analysis methods in neuroscience [274]. These decomposed representations prove more interpretable than raw network activations of hidden units, as validated by both human evaluation and automated analysis using language models [275, 276]. These insights have important implications for digital twins of the visual system. While current models excel at predicting individual neuron responses, they often maintain a traditional

focus on single-unit properties. Incorporating modern interpretability techniques focused on distributed representations and population dynamics could lead to more sophisticated models that better capture the complexity of biological visual processing as in the work from Bashiri et al. [157]. Taken together, this represents a crucial step toward understanding how populations of neurons collectively process visual information and get insights into the algorithms carried by the neuronal populations of different brain areas.

Overall, the studies presented in this thesis demonstrate the power of digital twins as tools for understanding visual processing across different species and scales. Looking ahead, the combination of computational modeling with targeted biological experiments promises to accelerate our understanding of neural computation, while the development of standardized benchmarks can ensure that progress in this field can be meaningfully measured and compared. The integration of digital twins into computational neuroscience establishes a powerful paradigm for testing theories of neural computation and, ultimately, brain function.

4 References

[1] Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H Sinz, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, 2022.

[2] Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Hansel, Christoph Blessing, Konstantin-Klemens Lurz, Max F. Burg, Santiago A. Cadena, Zhiwei Ding, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Kaiwen Deng, Yuanfang Guan, Yiqin Zhu, Kaiwen Xiao, Xiao Han, Simone Azeglio, Ulisse Ferrari, Peter Neri, Olivier Marre, Adrian Hoffmann, Kirill Fedyanin, Kirill Vishniakov, Maxim Panov, Subash Prakash, Kishan Naik, Kantharaju Narayanappa, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. Retrospective on the sensorium 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 314–333. PMLR, 28 Nov–09 Dec 2022. URL `https://proceedings.mlr.press/v220/willeke22a.html`.

[3] Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. May 2023. doi: 10.1101/2023.05.12.540591.

[4] Santiago Ramón y Cajal. *Estructura de los centros nerviosos de las aves*. 1888.

[5] Camillo Golgi. Sulla sostanza grigia del cervello. *Gazetta Medica Italiana*, 33: 244–246, 1873.

[6] Gordon M Shepherd. *Foundations of the neuron doctrine*. Oxford University Press, 2015.

[7] Charles Scott Sherrington. Observations on the scratch-reflex in the spinal dog. *The Journal of physiology*, 34(1-2):1, 1906.

[8] Haldan Keffer Hartline. The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2):400–415, 1938.

[9] Jerome Y Lettvin, Humberto R Maturana, Warren S McCulloch, and Walter H Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11): 1940–1951, 1959.

[10] Horace B Barlow. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394, 1972.

[11] Humberto R Maturana, Jerome Y Lettvin, Warren S McCulloch, and Walter H Pitts. Anatomy and physiology of vision in the frog (rana pipiens). *The Journal of general physiology*, 43(6):129, 1960.

[12] Howard Eichenbaum. Barlow versus hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neuroscience Letters*, 680:88–93, July 2018. ISSN 0304-3940. doi: 10.1016/j.neulet.2017.04.006. URL http://dx.doi.org/10.1016/j.neulet.2017.04.006.

[13] David H Hubel. Tungsten microelectrode for recording from single units. *Science*, 125(3247):549–550, 1957.

[14] V B Mountcastle. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.*, 20(4):408–434, July 1957.

[15] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

[16] David H Hubel and Torsten N Wiesel. Integrative action in the cat's lateral geniculate body. *The Journal of physiology*, 155(2):385, 1961.

[17] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[18] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, 28(2):229–289, 1965.

[19] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[20] Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984.

[21] M Tanaka, H Weber, and OD Creutzfeldt. Visual properties and spatial distribution of neurones in the visual association area on the prelunate gyrus of the awake monkey. *Experimental Brain Research*, 65:11–37, 1986.

[22] Keiji Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139, 1996.

[23] Gabriel Kreiman, Christof Koch, and Itzhak Fried. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9):946–953, September 2000. ISSN 1546-1726. doi: 10.1038/78868. URL http://dx.doi.org/10.1038/78868.

[24] R. Quian Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435 (7045):1102–1107, June 2005. ISSN 1476-4687. doi: 10.1038/nature03687. URL http://dx.doi.org/10.1038/nature03687.

[25] Charles G Gross. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5): 512–518, 2002.

[26] Howard Eichenbaum, Paul Dudchenko, Emma Wood, Matthew Shapiro, and Heikki Tanila. The hippocampus, memory, and place cells: is it spatial memory or a memory space? *Neuron*, 23(2):209–226, 1999.

[27] Tom Hartley, Colin Lever, Neil Burgess, and John O'Keefe. Space in the brain: how the hippocampal formation supports spatial cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):20120510, 2014.

[28] Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

[29] Rajeev V Rikhye and Mriganka Sur. Spatial correlations in natural scenes modulate response reliability in mouse visual cortex. *Journal of Neuroscience*, 35 (43):14661–14680, 2015.

[30] David C. Marr and Tomaso A. Poggio. From understanding computation to understanding neural circuitry. 1976. URL https://api.semanticscholar.org/CorpusID:61076330.

[31] H. B. Barlow. *Possible Principles Underlying the Transformations of Sensory Messages*, page 216–234. The MIT Press, September 2012. doi: 10.7551/mitpress/9780262518420.003.0013. URL http://dx.doi.org/10.7551/mitpress/9780262518420.003.0013.

[32] David J Field. What is the goal of sensory coding? *Neural computation*, 6(4): 559–601, 1994.

[33] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[34] Christina Enroth-Cugell and John G Robson. The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of physiology*, 187(3):517–552, 1966.

[35] J Anthony Movshon, Ian D Thompson, and David J Tolhurst. Receptive field organization of complex cells in the cat's striate cortex. *The Journal of physiology*, 283(1):79–99, 1978.

[36] C Enroth-Cugell and P Lennie. The control of retinal ganglion cell discharge by receptive field surrounds. *J. Physiol.*, 247(3):551–578, June 1975.

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

[38] Jon Touryan, Brian Lau, and Yang Dan. Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience*, 22(24): 10811–10818, 2002.

[39] Jon Touryan, Gidon Felsen, and Yang Dan. Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, 45(5):781–791, 2005.

[40] Edward H Adelson and James R Bergen. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299, 1985.

[41] David J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197, August 1992. ISSN 1469-8714, 0952-5238. doi: 10.1017/S0952523800009640.

[42] 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, December 1987.

[43] Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5 (4):322–333, 1969. ISSN 0536-1567. doi: 10.1109/tssc.1969.300225. URL `http://dx.doi.org/10.1109/TSSC.1969.300225`.

[44] Liam Paninski, Eero Simoncelli, and Jonathan Pillow. Maximum likelihood estimation of a stochastic integrate-and-fire neural model. *Advances in Neural Information Processing Systems*, 16, 2003.

[45] Eero P Simoncelli, Liam Paninski, Jonathan Pillow, Odelia Schwartz, et al. Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, 3(327-338):1, 2004.

[46] Matteo Carandini, Jonathan B. Demb, Valerio Mante, David J. Tolhurst, Yang Dan, Bruno A. Olshausen, Jack L. Gallant, and Nicole C. Rust. Do we know what the early visual system does? *J. Neurosci.*, 25(46):10577–10597, November 2005. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3726-05.2005.

[47] EJ Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: computation in neural systems*, 12(2):199, 2001.

[48] Kareem A Zaghloul, Kwabena Boahen, and Jonathan B Demb. Different circuits for on and off retinal ganglion cells cause different contrast sensitivities. *Journal of Neuroscience*, 23(7):2645–2654, 2003.

[49] Bruno A Olshausen and David J Field. How close are we to understanding v1? *Neural computation*, 17(8):1665–1699, 2005.

[50] Nicole C Rust, Odelia Schwartz, J Anthony Movshon, and Eero P Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

[51] Ethan A Benardete and Ehud Kaplan. The dynamics of primate m retinal ganglion cells. *Visual neuroscience*, 16(2):355–368, 1999.

[52] Kerry J Kim and Fred Rieke. Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. *Journal of Neuroscience*, 21 (1):287–299, 2001.

[53] Valerio Mante, Robert A Frazor, Vincent Bonin, Wilson S Geisler, and Matteo Carandini. Independence of luminance and contrast in natural scenes and in the early visual system. *Nature neuroscience*, 8(12):1690–1697, 2005.

[54] Javier Portilla. *International Journal of Computer Vision*, 40(1):49–70, 2000. ISSN 0920-5691. doi: 10.1023/a:1026553619983. URL `http://dx.doi.org/10.1023/A:1026553619983`.

[55] Jon Touryan, Gidon Felsen, and Yang Dan. Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, 45(5):781–791, 2005.

[56] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

[57] R Christopher Decharms and Anthony Zador. Neural representation and the cortical code. *Annual review of neuroscience*, 23(1):613–647, 2000.

[58] Nicole C Rust, Odelia Schwartz, J Anthony Movshon, and Eero P Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–956, 2005.

[59] 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.

[60] 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.

[61] Ben Willmore, Ryan J Prenger, Michael C-K Wu, and Jack L Gallant. The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform. *Neural computation*, 20(6):1537–1564, 2008.

[62] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[63] Nikolaus Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3:879, 2009.

[64] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

[65] Davide Zoccolan, Minjoon Kouh, Tomaso Poggio, and James J DiCarlo. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27(45):12292–12307, 2007.

[66] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[67] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[68] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 0018-9219. doi: 10.1109/5.726791. URL `http://dx.doi.org/10.1109/5.726791`.

[69] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[70] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[72] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[73] Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[74] Nikolaus Kriegeskorte. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008. doi: 10.3389/neuro.06.004.2008. URL `https://doi.org/10.3389/neuro.06.004.2008`.

[75] Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021.

[76] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep

learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

[77] Philipp Kaniuth and Martin N Hebart. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, 257:119294, 2022.

[78] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[79] Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064, 2021.

[80] 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*, 111(23):8619–8624, May 2014. doi: 10.1073/pnas.1403112111. URL `https://doi.org/10.1073/pnas.1403112111`.

[81] Charles F Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.*, 10(12):e1003963, 2014.

[82] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[83] Nikolaus Kriegeskorte and Jörn Diedrichsen. Peeling the onion of brain representations. *Annual Review of Neuroscience*, 42(1):407–432, July 2019. ISSN 1545-4126. doi: 10.1146/annurev-neuro-080317-061906. URL `http://dx.doi.org/10.1146/annurev-neuro-080317-061906`.

[84] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[85] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[86] Yimeng Zhang, T-S Tai Sing Lee, Ming Li, Fang Liu, Shiming Tang, Tai Sing, Lee Ming, Li Fang, Liu Shiming, T-S Tai Sing Lee, Ming Li, Fang Liu, and Shiming Tang. Convolutional neural network models of V1 responses to complex patterns. *J. Comput. Neurosci.*, pages 1–22, 2018.

[87] Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L.K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, May 2024. ISSN 0896-6273. doi: 10.1016/j.neuron.2024.04.018. URL http://dx.doi.org/10.1016/j.neuron.2024.04.018.

[88] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2013. URL https://arxiv.org/abs/1310.1531.

[89] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[90] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[91] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[92] Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020.

[93] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020.

[94] Soma Nonaka, Kei Majima, Shuntaro C Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9), 2021.

[95] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

[96] Yifei Ren and Pouya Bashivan. How well do models of visual cortex generalize to out of distribution samples? *PLOS Computational Biology*, 20(5):e1011145, May 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011145. URL http://dx.doi.org/10.1371/journal.pcbi.1011145.

[97] 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.

[98] Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan Litke, EJ Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion cell responses. In *International Conference on Learning Representations*, 2017.

[99] Lane T McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A Baccus. Deep learning models of the retinal response to natural scenes. *Adv. Neural Inf. Process. Syst.*, 29(Nips):1369–1377, 2016.

[100] William F Kindel, Elijah D Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. *Journal of vision*, 19(4):29–29, 2019.

[101] William F. Kindel, Elijah D. Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision*, 19(4):29, April 2019. doi: 10.1167/19.4.29. URL https://doi.org/10.1167/19.4.29.

[102] Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, April 2019. doi: 10.1371/journal.pcbi.1006897.

[103] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

[104] Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019.

[105] Umut Güçlü and Marcel AJ Van Gerven. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7, 2017.

[106] Kohitij Kar and James J DiCarlo. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1):164–176, 2021.

[107] Fabian H. Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, September 2019. doi: 10.1016/j.neuron.2019.08.034. URL https://doi.org/10.1016/j.neuron.2019.08.034.

[108] James J. DiCarlo, Daniel L. K. Yamins, Michael E. Ferguson, Evelina Fedorenko, Matthias Bethge, Tyler Bonnen, and Martin Schrimpf. Let's move forward: Image-computable models and a common model evaluation scheme

are prerequisites for a scientific understanding of human vision. *Behavioral and Brain Sciences*, 46, 2023. ISSN 1469-1825. doi: 10.1017/s0140525x23001607. URL http://dx.doi.org/10.1017/S0140525X23001607.

[109] Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1(1):417–446, November 2015. ISSN 2374-4650. doi: 10.1146/annurev-vision-082114-035447. URL http://dx.doi.org/10.1146/annurev-vision-082114-035447.

[110] Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, October 2021. ISSN 2471-285X. doi: 10.1109/tetci.2021.3100641. URL http://dx.doi.org/10.1109/TETCI.2021.3100641.

[111] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. URL https://arxiv.org/abs/2404.14082.

[112] Kohitij Kar, Simon Kornblith, and Evelina Fedorenko. Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nature Machine Intelligence*, 4(12):1065–1067, December 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00592-3. URL http://dx.doi.org/10.1038/s42256-022-00592-3.

[113] Grace W. Lindsay and David Bau. Testing methods of neural systems understanding. *Cognitive Systems Research*, 82:101156, December 2023. ISSN 1389-0417. doi: 10.1016/j.cogsys.2023.101156. URL http://dx.doi.org/10.1016/j.cogsys.2023.101156.

[114] David E. Rumelhart and James L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press, 1986. ISBN 9780262291408. doi: 10.7551/mitpress/5236.001.0001. URL http://dx.doi.org/10.7551/mitpress/5236.001.0001.

[115] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, Proceedings of the 26th International Conference On Machine Learning, ICML 2009, pages 609–616, 2009. ISBN 9781605585161. 26th International Conference On Machine Learning, ICML 2009 ; Conference date: 14-06-2009 Through 18-06-2009.

[116] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006. ISSN 1530-888X. doi: 10.1162/neco.2006.18.7.1527. URL http://dx.doi.org/10.1162/neco.2006.18.7.1527.

[117] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[118] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 655–670. Springer, 2019.

[119] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[120] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209): 1–90, 2021.

[121] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[122] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[123] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[124] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[125] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[126] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[127] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018.

[128] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, September 2020. ISSN 1091-6490. doi: 10.1073/pnas.1907375117. URL `http://dx.doi.org/10.1073/pnas.1907375117`.

[129] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

[130] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 2020. doi: 10.23915/distill.00024.002. https://distill.pub/2020/circuits/early-vision.

[131] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, 2022. URL `https://arxiv.org/abs/2207.13243`.

[132] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[133] Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 2021. doi: 10.23915/distill.00024.006. https://distill.pub/2020/circuits/curve-circuits.

[134] A Pasupathy and C E Connor. Responses to contour features in macaque area V4. *J. Neurophysiol.*, 82(5):2490–2502, November 1999.

[135] Anitha Pasupathy and Charles E Connor. Population coding of shape in area V4. *Nat. Neurosci.*, 5(12):1332–1338, November 2002.

[136] Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981, May 2013. ISSN 1546-1726. doi: 10.1038/nn.3402. URL `http://dx.doi.org/10.1038/nn.3402`.

[137] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[138] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[139] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[140] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[141] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[142] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016.

[143] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. URL `https://distill.pub/2017/feature-visualization`.

[144] Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.*, 22(12):2060–2065, December 2019.

[145] Pouya Bashivan, Kohitij Kar, and James 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.

[146] Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

[147] Jiakun Fu, Suhas Shrinivasan, Kayla Ponder, Taliah Muhammad, Zhuokun Ding, Eric Wang, Zhiwei Ding, Dat T Tran, Paul G Fahey, Stelios Papadopoulos, Saumil Patel, Jacob Reimer, Alexander S Ecker, Xaq Pitkow, Ralf M Haefner, Fabian H Sinz, Katrin Franke, and Andreas S Tolias. Pattern completion and disruption characterize contextual modulation in mouse visual cortex. March 2023.

[148] Zhiwei Ding, Dat T Tran, Kayla Ponder, Erick Cobos, Zhuokun Ding, Paul G Fahey, Eric Wang, Taliah Muhammad, Jiakun Fu, Santiago A Cadena, Stelios Papadopoulos, Saumil Patel, Katrin Franke, Jacob Reimer, Fabian H Sinz, Alexander S Ecker, Xaq Pitkow, and Andreas S Tolias. Bipartite invariance in mouse primary visual cortex. March 2023.

[149] Rudi Tong, Ronan da Silva, Dongyan Lin, Arna Ghosh, James Wilsenach, Erica Cianfarano, Pouya Bashivan, Blake Richards, and Stuart Trenholm. The feature landscape of visual cortex. November 2023. doi: 10.1101/2023.11.03.565500. URL `http://dx.doi.org/10.1101/2023.11.03.565500`.

[150] Larissa Hoefling, Klaudia P Szatko, Christian Behrens, Yongrong Qiu, David Alexander Klindt, Zachary Jessen, Gregory S Schwartz, Matthias Bethge, Philipp Berens, Katrin Franke, Alexander S Ecker, and Thomas Euler. A chromatic feature detector in the retina signals visual context changes.

December 2022. doi: 10.1101/2022.11.30.518492. URL http://dx.doi.org/10.1101/2022.11.30.518492.

[151] Matías A. Goldin, Baptiste Lefebvre, Samuele Virgili, Mathieu Kim Pham Van Cang, Alexander Ecker, Thierry Mora, Ulisse Ferrari, and Olivier Marre. Context-dependent selectivity to natural images in the retina. *Nature Communications*, 13(1), September 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33242-8. URL http://dx.doi.org/10.1038/s41467-022-33242-8.

[152] Michaela Vystrčilová, Shashwat Sridhar, Max F. Burg, Tim Gollisch, and Alexander S. Ecker. Convolutional neural network models of the primate retina reveal adaptation to natural stimulus statistics. March 2024. doi: 10.1101/2024.03.06.583740. URL http://dx.doi.org/10.1101/2024.03.06.583740.

[153] Ivan Ustyuzhaninov, Max F. Burg, Santiago A. Cadena, Jiakun Fu, Taliah Muhammad, Kayla Ponder, Emmanouil Froudarakis, Zhiwei Ding, Matthias Bethge, Andreas S. Tolias, and Alexander S. Ecker. Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*, 2022. doi: 10.1101/2022.02.10.479884. URL https://www.biorxiv.org/content/early/2022/02/10/2022.02.10.479884.

[154] Eric Y Wang, Paul G Fahey, Kayla Ponder, Zhuokun Ding, Andersen Change, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat T Tran, Jiakun Fu, et al. Towards a foundation model of the mouse visual cortex. *bioRxiv*, pages 2023–03, 2023.

[155] Max F. Burg, Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Andreas S. Tolias, Matthias Bethge, and Alexander 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.

[156] Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay K Jagadish, Eric Wang, Edgar Y Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S Tolias, Alexander S Ecker, and Fabian 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, October 2021.

[157] Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, and Fabian Sinz. A flow-based latent state generative model of neural population responses to natural images. *Adv. Neural Inf. Process. Syst.*, 34, December 2021.

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

[159] Alexander S. Ecker, Fabian H. Sinz, Emmanouil Froudarakis, Paul G. Fahey, Santiago A. Cadena, Edgar Y. Walker, Erick Cobos, Jacob Reimer, Andreas S. Tolias, and Matthias Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex, 2018.

[160] 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.

[161] Erick Cobos, Taliah Muhammad, Paul G. Fahey, Zhiwei Ding, Zhuokun Ding, Jacob Reimer, Fabian H. Sinz, and Andreas S. Tolias. It takes neurons to understand neurons: Digital twins of visual cortex synthesize neural metamers. December 2022. doi: 10.1101/2022.12.09.519708. URL `http://dx.doi.org/10.1101/2022.12.09.519708`.

[162] Hui-Yuan Miao and Frank Tong. Convolutional neural network models of neuronal responses in macaque v1 reveal limited non-linear processing. August 2023. doi: 10.1101/2023.08.26.554952. URL `http://dx.doi.org/10.1101/2023.08.26.554952`.

[163] Fengtong Du, Miguel Angel Núñez-Ochoa, Marius Pachitariu, and Carsen Stringer. Towards a simplified model of primary visual cortex. July 2024. doi: 10.1101/2024.06.30.601394. URL `http://dx.doi.org/10.1101/2024.06.30.601394`.

[164] C H Rowell. Variable responsiveness of a visual interneurone in the Free-Moving locust, and its relation to behaviour and arousal. *Journal of Experimental Biology*, 1971.

[165] M Eugenia Chiappe, Johannes D Seelig, Michael B Reiser, and Vivek Jayaraman. Walking modulates speed sensitivity in drosophila motion vision. *Curr. Biol.*, 20(16):1470–1475, August 2010.

[166] S Treue and J H Maunsell. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, 382(6591):539–541, August 1996.

[167] C J McAdams and J H Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.*, 19(1): 431–441, January 1999.

[168] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437), 2019.

[169] Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. In *International Conference on Learning Representations*, 2021.

[170] Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 739–751. Curran Associates, Inc., 2021.

[171] Paweł A. Pierzchlewicz, Konstantin F. Willeke, Arne F. Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, Andreas S. Tolias, and Fabian H. Sinz. Energy guided diffusion for generating neurally exciting images. May 2023. doi: 10.1101/2023.05.18.541176. URL https://doi.org/10.1101/2023.05.18.541176.

[172] Santiago A Cadena, Konstantin F Willeke, Kelli Restivo, George Denfield, Fabian H Sinz, Matthias Bethge, Andreas S Tolias, and Alexander S Ecker. Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. *bioRxiv*, page 2022.05.18.492503, May 2022.

[173] Polina Turishcheva, Paul G. Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F. Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S. Ecker. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. 2023. doi: 10.48550/ARXIV.2305.19654. URL https://arxiv.org/abs/2305.19654.

[174] Max F Burg, Thomas Zenkel, Michaela Vystrčilová, Jonathan Oesterle, Larissa Höfling, Konstantin Friedrich Willeke, Jan Lause, Sarah Müller, Paul G. Fahey, Zhiwei Ding, Kelli Restivo, Shashwat Sridhar, Tim Gollisch, Philipp Berens, Andreas S. Tolias, Thomas Euler, Matthias Bethge, and Alexander S Ecker. Maximally discriminative stimuli for functional cell type identification. 2024. URL https://openreview.net/forum?id=9W6KaAcYlr.

[175] Jiakun Fu, Konstantin F. Willeke, Pawel A. Pierzchlewicz, Taliah Muhammad, George H. Denfield, Fabian Hubert Sinz, and Andreas S. Tolias. Heterogeneous orientation tuning across sub-regions of receptive fields of v1 neurons in mice. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4029075. URL https://doi.org/10.2139/ssrn.4029075.

[176] Jacob Reimer, Emmanouil Froudarakis, Cathryn R Cadwell, Dimitri Yatsenko, George H Denfield, and Andreas S Tolias. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*, 84(2):355–362, October 2014.

[177] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, February 2010.

[178] Martin Vinck, Renata Batista-Brito, Ulf Knoblich, and Jessica A Cardin. Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron*, 86(3):740–754, May 2015.

[179] Corbett Bennett, Sergio Arroyo, and Shaul Hestrin. Subthreshold mechanisms underlying state-dependent modulation of visual responses. *Neuron*, 80(2): 350–357, October 2013.

[180] Sinem Erisken, Agne Vaiceliunaite, Ovidiu Jurjut, Matilde Fiorini, Steffen Katzner, and Laura Busse. Effects of locomotion extend throughout the mouse early visual system. *Curr. Biol.*, 24(24):2899–2907, December 2014.

[181] Liang Liang, Alex Fratzl, Jasmine D S Reggiani, Omar El Mansour, Chinfei Chen, and Mark L Andermann. Retinal inputs to the thalamus are selectively gated by arousal. *Curr. Biol.*, 30(20):3923–3934.e9, October 2020.

[182] Alexander S Ecker, Philipp Berens, R James Cotton, Manivannan Subramaniyan, George H Denfield, Cathryn R Cadwell, Stelios M Smirnakis, Matthias Bethge, and Andreas S Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.

[183] Ellen J Gerl and Molly R Morris. The causes and consequences of color vision. *Evolution: Education and Outreach*, 1(4):476–486, October 2008.

[184] A Szél, P Röhlich, A R Gaffé, B Juliusson, G v Aguirre, and T Van Veen. Unique topographic separation of two spectral classes of cones in the mouse retina. *J. Comp. Neurol.*, 325(3):327–342, 1992.

[185] Tom Baden, Timm Schubert, Le Chang, Tao Wei, Mariana Zaichuk, Bernd Wissinger, and Thomas Euler. A tale of two retinal domains: near-optimal sampling of achromatic contrasts in natural scenes through asymmetric photoreceptor distribution. *Neuron*, 80(5):1206–1217, December 2013.

[186] Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, and Fabian Sinz. A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, 34, 2021.

[187] Emmanuel Eggermann, Yves Kremer, Sylvain Crochet, and Carl C H Petersen. Cholinergic signals in mouse barrel cortex during active whisker sensing. *Cell Rep.*, 9(5):1654–1660, December 2014.

[188] Sylvia Schröder, Nicholas A Steinmetz, Michael Krumin, Marius Pachitariu, Matteo Rizzi, Leon Lagnado, Kenneth D Harris, and Matteo Carandini. Arousal modulates retinal output. *Neuron*, 107(3):487–495.e9, August 2020.

[189] H Spitzer, R Desimone, and J Moran. Increased attention enhances both behavioral and neuronal performance. *Science*, 240(4850):338–340, April 1988.

[190] Maria C Dadarlat and Michael P Stryker. Locomotion enhances neural encoding of visual stimuli in mouse V1. *J. Neurosci.*, 37(14):3764–3775, April 2017.

[191] J W de Gee, Z Mridha, M Hudson, Y Shi, H Ramsaywak, S Smith, N Karediya, M Thompson, K Jaspe, W Zhang, and M J McGinley. Mice regulate their attentional intensity and arousal to exploit increases in task utility. May 2022.

[192] Tatiana Bezdudnaya, Monica Cano, Yulia Bereshpolova, Carl R Stoelzel, Jose-Manuel Alonso, and Harvey A Swadlow. Thalamic burst mode and inattention in the awake LGNd. *Neuron*, 49(3):421–432, February 2006.

[193] Mark L Andermann, Aaron M Kerlin, Demetris K Roumis, Lindsey L Glickfeld, and R Clay Reid. Functional specialization of mouse higher visual cortical areas. *Neuron*, 72(6):1025–1039, December 2011.

[194] Thomas W Cronin and Michael J Bok. Photoreception and vision in the ultraviolet. *J. Exp. Biol.*, 219(Pt 18):2790–2801, September 2016.

[195] Yongrong Qiu, Zhijian Zhao, David Klindt, Magdalena Kautzky, Klaudia P Szatko, Frank Schaeffel, Katharina Rifai, Katrin Franke, Laura Busse, and Thomas Euler. Natural environment statistics in the upper and lower visual field are reflected in mouse retinal specializations. *Curr. Biol.*, June 2021.

[196] Laura Busse. The influence of locomotion on sensory processing and its underlying neuronal circuits. *eNeuroforum*, 24(1):A41–A51, February 2018.

[197] David M Schneider. Reflections of action in sensory cortex. *Curr. Opin. Neurobiol.*, 64:53–59, October 2020.

[198] Rylan S Larsen and Jack Waters. Neuromodulatory correlates of pupil dilation. *Front. Neural Circuits*, 12:21, March 2018.

[199] Jonathan C Horton and Daniel L Adams. The cortical column: a structure without a function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 360(1456):837–862, April 2005.

[200] D Purves, DR Riddle, and AS LaMantia. Iterated patterns of brain circuitry (or how the cortex gets its spots). *Trends in neurosciences*, 15(10):362–368, 1992.

[201] Thomas D Albright, Robert Desimone, and Charles G Gross. Columnar organization of directionally selective cells in visual area mt of the macaque. *Journal of neurophysiology*, 51(1):16–31, 1984.

[202] Gregory C DeAngelis and William T Newsome. Organization of disparity-selective neurons in macaque area mt. *Journal of Neuroscience*, 19(4):1398–1415, 1999.

[203] Kenichi Ohki and R Clay Reid. In vivo two-photon calcium imaging in the visual system. *Cold Spring Harbor Protocols*, 2014(4):pdb–prot081455, 2014.

[204] Xiaolong Jiang, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel, and Andreas S Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350(6264):aac9462, 2015.

[205] Cathryn R Cadwell, Federico Scala, Paul G Fahey, Dmitry Kobak, Shalaka Mulherkar, Fabian H Sinz, Stelios Papadopoulos, Zheng H Tan, Per Johnsson, Leonard Hartmanis, et al. Cell type composition and circuit organization of clonally related excitatory neurons in the juvenile mouse neocortex. *Elife*, 9: e52951, 2020.

[206] Luke Campagnola, Stephanie C Seeman, Thomas Chartrand, Lisa Kim, Alex Hoggarth, Clare Gamlin, Shinya Ito, Jessica Trinh, Pasha Davoudian, Cristina Radaelli, et al. Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, 375(6585):eabj5861, 2022.

[207] Anitha Pasupathy and Charles E Connor. Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, 86(5):2505–2519, 2001.

[208] Anitha Pasupathy, Dina V Popovkina, and Taekjun Kim. Visual functions of primate area v4. *Annual Review of Vision Science*, 6:363–385, 2020.

[209] Rendong Tang, Qianling Song, Ying Li, Rui Zhang, Xingya Cai, and Haidong D Lu. Curvature-processing domains in primate V4. *Elife*, 9, November 2020.

[210] Doris Y Tsao, Winrich A Freiwald, Tamara A Knutsen, Joseph B Mandeville, and Roger BH Tootell. Faces and objects in macaque cerebral cortex. *Nature neuroscience*, 6(9):989–995, 2003.

[211] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), 2019.

[212] Taekjun Kim, Wyeth Bair, and Anitha Pasupathy. Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, 39(24):4760–4774, 2019.

[213] Ramanujan Srinath, Alexandriya Emonds, Qingyang Wang, Augusto A Lempel, Erika Dunn-Weiss, Charles E Connor, and Kristina J Nielsen. Early emergence of solid shape coding in natural and deep network vision. *Curr. Biol.*, 31(1):51–65.e5, January 2021.

[214] Robert Desimone and Stanley J Schein. Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of neurophysiology*, 57(3): 835–868, 1987.

[215] Jack L Gallant, Jochen Braun, and David C Van Essen. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091): 100–103, 1993.

[216] E Kobatake and K Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, 71(3):856–867, March 1994.

[217] 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.*, 2019.

[218] William F Kindel, Elijah D Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. *Journal of vision*, 19(4):29–29, 2019.

[219] Larissa Höfling, Klaudia P Szatko, Christian Behrens, Yongrong Qiu, David A Klindt, Zachary Jessen, Gregory W Schwartz, Matthias Bethge, Philipp Berens, Katrin Franke, Alexander S Ecker, and Thomas Euler. A chromatic feature detector in the retina signals visual context changes. December 2022.

[220] Anitha Pasupathy, Dina V Popovkina, and Taekjun Kim. Visual functions of primate area V4. *Annu Rev Vis Sci*, 6:363–385, September 2020.

[221] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[222] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. February 2020.

[223] Niklas Böhm, Philipp Berens, and Dmitry Kobak. Unsupervised visualization of image datasets using contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=nI2HmVA0hvt`.

[224] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205. URL `https://doi.org/10.21105/joss.00205`.

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

[226] Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6: 414–417, 1983.

[227] Rundong Jiang, Ian Max Andolina, Ming Li, and Shiming Tang. Clustered functional domains for curves and corners in cortical area v4. *Elife*, 10:e63798, 2021.

[228] Ramanujan Srinath, Alexandriya Emonds, Qingyang Wang, Augusto A Lempel, Erika Dunn-Weiss, Charles E Connor, and Kristina J Nielsen. Early emergence of solid shape coding in natural and deep network vision. *Current Biology*, 31(1):51–65, 2021.

[229] Jia Ming Hu, Xue Mei Song, Qiannan Wang, and Anna Wang Roe. Curvature domains in v4 of macaque monkey. *Elife*, 9:e57261, 2020.

[230] Yiliang Lu, Jiapeng Yin, Zheyuan Chen, Hongliang Gong, Ye Liu, Liling Qian, Xiaohong Li, Rui Liu, Ian Max Andolina, and Wei Wang. Revealing detail along the visual hierarchy: neural clustering preserves acuity from v1 to v4. *Neuron*, 98(2):417–428, 2018.

[231] Peichao Li, Shude Zhu, Ming Chen, Chao Han, Haoran Xu, Jiaming Hu, Yang Fang, and Haidong D Lu. A motion direction preference map in monkey v4. *Neuron*, 78(2):376–388, 2013.

[232] Tomoyuki Namima, Erin Kempkes, Polina Zamarashkina, Natalia Owen, and Anitha Pasupathy. High-density recording reveals sparse clusters (but not columns) for shape and texture encoding in macaque v4. October 2023. doi: 10.1101/2023.10.15.562424. URL http://dx.doi.org/10.1101/2023.10.15.562424.

[233] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *arXiv*, 2014.

[234] Katrin Franke, Konstantin F. Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H. Sinz, and Andreas S. Tolias. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, September 2022. doi: 10.1038/s41586-022-05270-3. URL https://doi.org/10.1038/s41586-022-05270-3.

[235] Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 33(10): 2017–2031, September 2021. ISSN 1530-8898. doi: 10.1162/jocn_a_01544. URL http://dx.doi.org/10.1162/jocn_a_01544.

[236] R. M. Cichy, K. Dwivedi, B. Lahner, A. Lascelles, P. Iamshchinina, M. Graumann, A. Andonian, N. A. R. Murty, K. Kay, G. Roig, and A. Oliva. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion, 2021. URL https://arxiv.org/abs/2104.13714.

[237] Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, Peter A. Groblewski, Gabriel K. Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, Kate Roll, Marina Garrett, Tom Keenan, Leonard Kuan, Stefan Mihalas, Shawn Olsen, Carol Thompson, Wayne Wakeman, Jack Waters, Derric Williams, Chris Barber, Nathan Berbesque, Brandon Blanchard, Nicholas Bowles, Shiella D. Caldejon, Linzy Casal, Andrew Cho, Sissy Cross, Chinh Dang, Tim Dolbeare, Melise Edwards, John Galbraith, Nathalie Gaudreault, Terri L. Gilbert, Fiona Griffin, Perry Hargrave,

Robert Howard, Lawrence Huang, Sean Jewell, Nika Keller, Ulf Knoblich, Josh D. Larkin, Rachael Larsen, Chris Lau, Eric Lee, Felix Lee, Arielle Leon, Lu Li, Fuhui Long, Jennifer Luviano, Kyla Mace, Thuyanh Nguyen, Jed Perkins, Miranda Robertson, Sam Seid, Eric Shea-Brown, Jianghong Shi, Nathan Sjoquist, Cliff Slaughterbeck, David Sullivan, Ryan Valenza, Casey White, Ali Williford, Daniela M. Witten, Jun Zhuang, Hongkui Zeng, Colin Farrell, Lydia Ng, Amy Bernard, John W. Phillips, R. Clay Reid, and Christof Koch. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, December 2019. doi: 10.1038/s41593-019-0550-9. URL https://doi.org/10.1038/s41593-019-0550-9.

[238] Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity, 2021. URL https://arxiv.org/abs/2109.04463.

[239] Bryan M. Li, Isabel M. Cornacchia, Nathalie L. Rochefort, and Arno Onken. V1t: large-scale mouse v1 response prediction using a vision transformer, 2023. URL https://arxiv.org/abs/2302.03023.

[240] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[241] Polina Turishcheva, Paul G Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, et al. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *arXiv preprint arXiv:2305.19654*, 2023.

[242] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[243] Rafael Yuste. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8):487–497, July 2015. ISSN 1471-0048. doi: 10.1038/nrn3962. URL http://dx.doi.org/10.1038/nrn3962.

[244] Bilal Haider, Michael Häusser, and Matteo Carandini. Inhibition dominates sensory responses in the awake cortex. *Nature*, 493(7430):97–100, November 2012. ISSN 1476-4687. doi: 10.1038/nature11665. URL http://dx.doi.org/10.1038/nature11665.

[245] Alfonso Renart and Christian K Machens. Variability in neural activity and behavior. *Current Opinion in Neurobiology*, 25:211–220, April 2014. ISSN 0959-4388. doi: 10.1016/j.conb.2014.02.013. URL http://dx.doi.org/10.1016/j.conb.2014.02.013.

[246] Michael D. Fox and Marcus E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9):700–711, September 2007. ISSN 1471-0048. doi: 10.1038/nrn2201. URL http://dx.doi.org/10.1038/nrn2201.

[247] Kenneth D Harris and Alexander Thiele. Cortical state and attention. *Nat. Rev. Neurosci.*, 12(9):509–523, August 2011.

[248] Dario L Ringach. Spontaneous and driven cortical activity: implications for computation. *Current Opinion in Neurobiology*, 19(4):439–444, 2009. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2009.07.005. URL https://www.sciencedirect.com/science/article/pii/S0959438809000786. Sensory systems.

[249] Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, January 2011. ISSN 1095-9203. doi: 10.1126/science.1195870. URL http://dx.doi.org/10.1126/science.1195870.

[250] Kenneth D. Harris. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, 6(5):399–407, May 2005. ISSN 1471-0048. doi: 10.1038/nrn1669. URL http://dx.doi.org/10.1038/nrn1669.

[251] Stefano Panzeri, Jakob H. Macke, Joachim Gross, and Christoph Kayser. Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, 19(3):162–172, March 2015. ISSN 1364-6613. doi: 10.1016/j.tics.2015.01.002. URL http://dx.doi.org/10.1016/j.tics.2015.01.002.

[252] Bruno B. Averbeck, Peter E. Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, May 2006. ISSN 1471-0048. doi: 10.1038/nrn1888. URL http://dx.doi.org/10.1038/nrn1888.

[253] Nicholas James Sofroniew, Daniel Flickinger, Jonathan King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife*, 5:e14472, June 2016.

[254] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, March 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2434. URL http://dx.doi.org/10.1038/nmeth.2434.

[255] Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, June 2012. ISSN 1476-4687. doi: 10.1038/nature11129. URL http://dx.doi.org/10.1038/nature11129.

[256] R Quian Quiroga, Leila Reddy, Christof Koch, and Itzhak Fried. Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of neurophysiology*, 98(4):1997–2007, 2007.

[257] R. Quian Quiroga, G. Kreiman, C. Koch, and I. Fried. Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, 12(3):87–91, March 2008. ISSN 1364-6613. doi: 10.1016/j.tics.2007.12.003. URL http://dx.doi.org/10.1016/j.tics.2007.12.003.

[258] Le Chang and Doris Y. Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028.e14, June 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.05.011. URL http://dx.doi.org/10.1016/j.cell.2017.05.011.

[259] Edvard I. Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31 (1):69–89, July 2008. ISSN 1545-4126. doi: 10.1146/annurev.neuro.31.061307.090723. URL http://dx.doi.org/10.1146/annurev.neuro.31.061307.090723.

[260] R.G.M Morris. D.o. hebb: The organization of behavior, wiley: New york; 1949. *Brain Research Bulletin*, 50(5–6):437, November 1999. ISSN 0361-9230. doi: 10.1016/s0361-9230(99)00182-3. URL http://dx.doi.org/10.1016/s0361-9230(99)00182-3.

[261] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, January 2012. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn3136.

[262] B Olshausen and D Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, August 2004. ISSN 0959-4388. doi: 10.1016/j.conb.2004.07.007. URL http://dx.doi.org/10.1016/j.conb.2004.07.007.

[263] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, December 1997. ISSN 0042-6989. doi: 10.1016/s0042-6989(97)00169-7. URL http://dx.doi.org/10.1016/S0042-6989(97)00169-7.

[264] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema,

João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, October 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0520-2. URL http://dx.doi.org/10.1038/s41593-019-0520-2.

[265] David L. Barack and John W. Krakauer. Two views on the cognitive brain. *Nature Reviews Neuroscience*, 22(6):359–371, April 2021. ISSN 1471-0048. doi: 10.1038/s41583-021-00448-6. URL http://dx.doi.org/10.1038/s41583-021-00448-6.

[266] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3), March 2018. ISSN 2476-0757. doi: 10.23915/distill.00010. URL http://dx.doi.org/10.23915/distill.00010.

[267] Simon J. Thorpe. Local vs. distributed coding. *Intellectica*, 8:3–40, 1989. URL https://api.semanticscholar.org/CorpusID:70175501.

[268] Geoffrey E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1–3):185–234, September 1989. ISSN 0004-3702. doi: 10.1016/0004-3702(89)90049-0. URL http://dx.doi.org/10.1016/0004-3702(89)90049-0.

[269] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982. ISSN 1091-6490. doi: 10.1073/pnas.79.8.2554. URL http://dx.doi.org/10.1073/pnas.79.8.2554.

[270] Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, May 2013. ISSN 1476-4687. doi: 10.1038/nature12160. URL http://dx.doi.org/10.1038/nature12160.

[271] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[272] Trenton Bricken, Rylan Schaeffer, Bruno Olshausen, and Gabriel Kreiman. Emergence of sparse representations from noise. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3148–3191.

PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bricken23a.html.

[273] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

[274] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381 (6583):607–609, June 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL http://dx.doi.org/10.1038/381607a0.

[275] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[276] Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features, 2024. URL https://arxiv.org/abs/2411.02193.

5 Appendix

This chapter contains the publications that are summarized in the result section.

5.1 State-dependent pupil dilation rapidly shifts visual feature selectivity

Article

State-dependent pupil dilation rapidly shifts visual feature selectivity

https://doi.org/10.1038/s41586-022-05270-3

Received: 5 December 2021

Accepted: 23 August 2022

Published online: 28 September 2022

Check for updates

Katrin Franke[1,2,3,4,9]✉, Konstantin F. Willeke[5,6,9], Kayla Ponder[3,4], Mario Galdamez[3,4], Na Zhou[3,4], Taliah Muhammad[3,4], Saumil Patel[3,4], Emmanouil Froudarakis[3,4,7], Jacob Reimer[3,4], Fabian H. Sinz[3,4,5,6,10] & Andreas S. Tolias[3,4,8,10]

To increase computational flexibility, the processing of sensory inputs changes with behavioural context. In the visual system, active behavioural states characterized by motor activity and pupil dilation[1,2] enhance sensory responses, but typically leave the preferred stimuli of neurons unchanged[2–9]. Here we find that behavioural state also modulates stimulus selectivity in the mouse visual cortex in the context of coloured natural scenes. Using population imaging in behaving mice, pharmacology and deep neural network modelling, we identified a rapid shift in colour selectivity towards ultraviolet stimuli during an active behavioural state. This was exclusively caused by state-dependent pupil dilation, which resulted in a dynamic switch from rod to cone photoreceptors, thereby extending their role beyond night and day vision. The change in tuning facilitated the decoding of ethological stimuli, such as aerial predators against the twilight sky[10]. For decades, studies in neuroscience and cognitive science have used pupil dilation as an indirect measure of brain state. Our data suggest that, in addition, state-dependent pupil dilation itself tunes visual representations to behavioural demands by differentially recruiting rods and cones on fast timescales.

Neuronal responses in animals are modulated by their behavioural and internal states to flexibly adjust information processing to different behavioural contexts. This phenomenon has been well described across animal species, from invertebrates[11,12] to primates[4,9]. In the mammalian visual cortex, neuronal activity is desynchronized and sensory responses are enhanced during an active behavioural state[1–3,5,7,8], which is characterized by pupil dilation[1] and locomotion activity[2]. Mechanistically, these effects have been linked to neuromodulators such as acetylcholine and noradrenaline (reviewed in refs. [13,14]). Other than changes in response gain, the tuning of visual neurons, such as orientation selectivity, typically does not change across quiet and active states[2,3,5,7,8]. So far, however, this has largely been studied in non-ecological settings using simple synthetic stimuli.

In this work, we study how behavioural state modulates cortical visual tuning in mice in the context of naturalistic scenes. Crucially, these scenes include the colour domain of the visual input due to its ethological relevance across species (reviewed in ref. [15]). Mice, like most mammals, are dichromatic and have two types of cone photoreceptor that express ultraviolet (UV)-sensitive and green-sensitive short-wavelength and medium-wavelength opsins (S-opsin and M-opsin, respectively)[16]. These UV-sensitive and green-sensitive cone photoreceptors predominantly sample the upper and the lower visual field, respectively, through uneven distributions across the retina[16,17].

To systematically study the relationship between neuronal tuning and behavioural state in the context of naturalistic scenes, we combined in vivo population calcium imaging of the primary visual cortex (V1) in awake, head-fixed mice with deep convolutional neural network (CNN) modelling. We extended a recently described model[18,19] to predict neuronal responses on the basis of both the visual input and the behaviour of the animal jointly. This enabled us to characterize the relationship between neuronal tuning and behaviour in extensive in silico experiments without the need to experimentally control the behaviour. Finally, we experimentally confirmed in vivo the in silico model predictions[18,20].

Using this approach, we demonstrate that colour tuning of mouse V1 neurons rapidly shifts towards higher UV sensitivity during an active behavioural state. By pharmacologically manipulating the pupil, we show that this is solely caused by pupil dilation. Dilation during active behavioural states sufficiently increases the amount of light entering the eye to cause a dynamic switch between rod-dominated and cone-dominated vision, even for constant ambient light levels. Finally, we show that the increased UV sensitivity during active periods may tune the mouse visual system to improved detection of predators against the UV background of the sky. Our results identify a new functional role of state-dependent pupil dilation: to rapidly tune visual feature representations to changing behavioural requirements in a bottom-up manner.

[1]Institute for Ophthalmic Research, Tübingen University, Tübingen, Germany. [2]Center for Integrative Neuroscience, Tübingen University, Tübingen, Germany. [3]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA. [4]Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, USA. [5]Institute for Bioinformatics and Medical Informatics, Tübingen University, Tübingen, Germany. [6]Department of Computer Science, Göttingen University, Göttingen, Germany. [7]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, Heraklion, Greece. [8]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. [9]These authors contributed equally: Katrin Franke and Konstantin F. Willeke. [10]These authors jointly supervised this work: Fabian H. Sinz and Andreas S. Tolias. ✉e-mail: katrin.franke@uni-tuebingen.de

CNNs identify optimal coloured stimuli

Here we studied the relationship between neuronal tuning in mouse V1 and the behaviour of the animal, specifically focusing on colour processing because of its behavioural relevance (reviewed in ref. [15]). We presented coloured naturalistic images (Extended Data Fig. 1) to awake, head-fixed mice positioned on a treadmill (Fig. 1a) while recording the calcium activity of L2/3 neurons in V1 using two-photon imaging (Fig. 1c,d). We simultaneously recorded locomotion activity, pupil size and instantaneous changes in pupil size, which have all been associated with distinct behavioural states[1,2]. Visual stimuli were presented using a projector with UV and green light-emitting diodes (LEDs)[21] (Fig. 1b), which enabled the differential activation of UV-sensitive and green-sensitive mouse photoreceptors. We recorded neuronal responses along the posterior–anterior axis of V1 (Fig. 1c), sampling from various vertical positions across the visual field. This choice was motivated by the gradient of spectral sensitivity of mouse cone photoreceptors across the retina[16,17].

We used a deep CNN to learn an in silico model of the recorded neuron population as a function of the visual input and the behaviour of the animal[18] (Fig. 1e). The CNN had the following input channels: (1) UV and green channels of the visual stimulus; (2) three channels set to the recorded behavioural parameters (that is, pupil size, change in pupil size and locomotion); and (3) two channels that were shared across all inputs encoding the x and y pixel positions of the stimulus image. The third criterion was previously shown to improve CNN model performance in cases for which feature representations depend on image position[22], similar to the gradient in mouse colour sensitivity across visual space. Our neural predictive models also included a shifter network[18] that spatially shifted the receptive fields of model neurons according to the recorded pupil position traces. For each dataset, we trained an ensemble of four-layer CNN models end-to-end[19] to predict the neuronal responses to individual images and behavioural parameters. The prediction performance of the resulting ensemble model (Extended Data Fig. 2) was comparable to state-of-the-art predictive models of mouse V1 (ref. [19]).

Using our CNN ensemble model as a 'digital twin' of the visual cortex, we synthesized maximally exciting inputs (MEIs) for individual neurons (Fig. 1f and Extended Data Fig. 3a). To this end, we optimized the UV and green colour channels of a contrast-constrained image to produce the highest activation in the given model neuron using regularized gradient ascent[18,20]. For most of the neurons, MEI colour channels were positively correlated, which indicated that colour opponency is rare given our stimulus paradigm (Extended Data Figs. 3 and 4). Inception loop experiments[18] confirmed that the computed MEIs strongly drive the recorded neurons. For these experiments, we randomly selected MEIs of 150 neurons above a response reliability threshold for presentation on the next day (Fig. 1g). For most neurons, the MEIs were indeed the most exciting stimuli: responses of neurons to their own MEI were significantly larger than to other MEIs (Fig. 1h; for statistics, see figure legends and Supplementary Methods). Together, these findings demonstrate that our modelling approach accurately captures the tuning properties of mouse V1 neurons in the context of coloured naturalistic scenes.

V1 colour tuning changes with behaviour

To study how cortical colour tuning changes with behavioural state, we performed detailed in silico characterizations using the above-described trained CNN model. To that end, we focused on two well described and spontaneously occurring behavioural states[1,2]: (1) a quiet state with no locomotion and a small pupil (3rd percentile of locomotion and pupil size across all trials) and (2) an active state indicated by locomotion and a larger pupil (97th percentile). For each neuron and distinct behavioural state, we optimized a MEI and then generated a

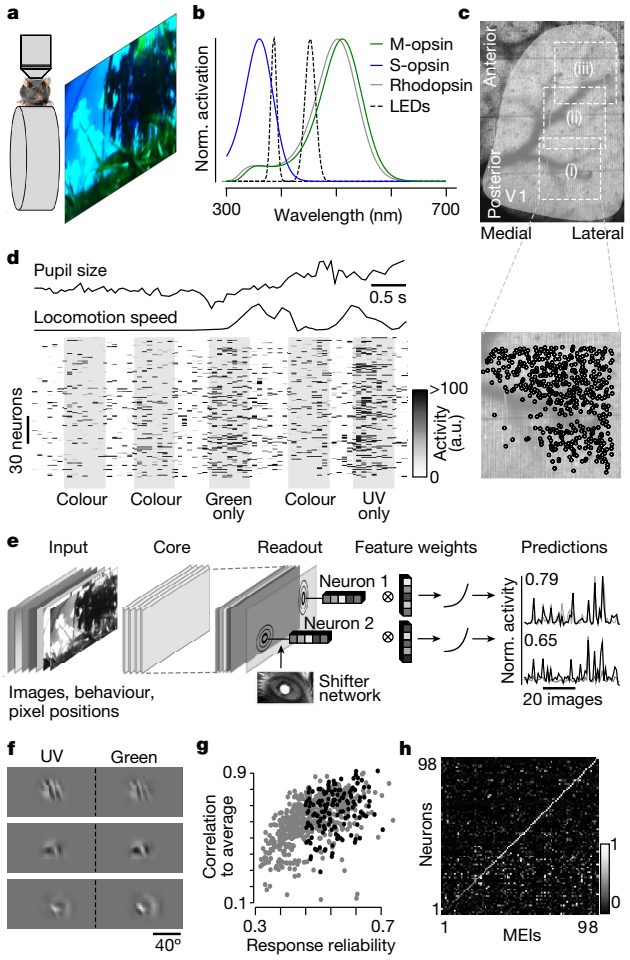


Fig. 1 | Deep neural networks capture mouse V1 tuning properties in the context of coloured naturalistic scenes. a, Schematic of the experimental setup. Awake, head-fixed mice on a treadmill were presented with UV-coloured and green-coloured naturalistic scenes (Extended Data Fig. 1). **b**, Normalized (Norm.) sensitivity spectra of mouse S-opsin and M-opsin expressed by cones and rhodopsin expressed by rods, with LED spectra for visual stimulation. **c**, Cortical surface of a transgenic mouse expressing GCaMP6s, with positions of three scan fields ((i)–(iii), 650 × 650 µm each). The bottom image shows cells (n = 478) selected for further analysis. **d**, Neuronal activity (shown in arbitrary units (a.u.); n = 150 cells) in response to coloured naturalistic scenes and simultaneously recorded behavioural data (pupil size and locomotion speed). **e**, Schematic of the model architecture. The model input consists of two image channels, three behaviour channels and two position channels that encode the x and y pixel position of the input images[22]. A four-layer convolutional core is followed by a Gaussian readout and a nonlinearity[19]. Readout positions were adjusted using a shifter network[18]. Traces on the right show average responses (grey) to test images of two example neurons and corresponding model predictions (black). **f**, MEI images of three example neurons (from n = 658). See also Extended Data Fig. 3. **g**, Response reliability to natural images plotted against model prediction performance of all cells of one scan. Neurons selected for experimental verification (inception loop) are indicated in black. **h**, Confusion matrix of the inception loop experiment[18] depicting the activity of each selected neuron to presented MEIs. Neurons are ordered on the basis of the response to their own MEI (>65% showed the strongest response to their own MEI). Responses of neurons to their own MEI (along the diagonal) were significantly larger than to other MEIs (P = 0 for a one-sided permutation test, n = 10,000 permutations).

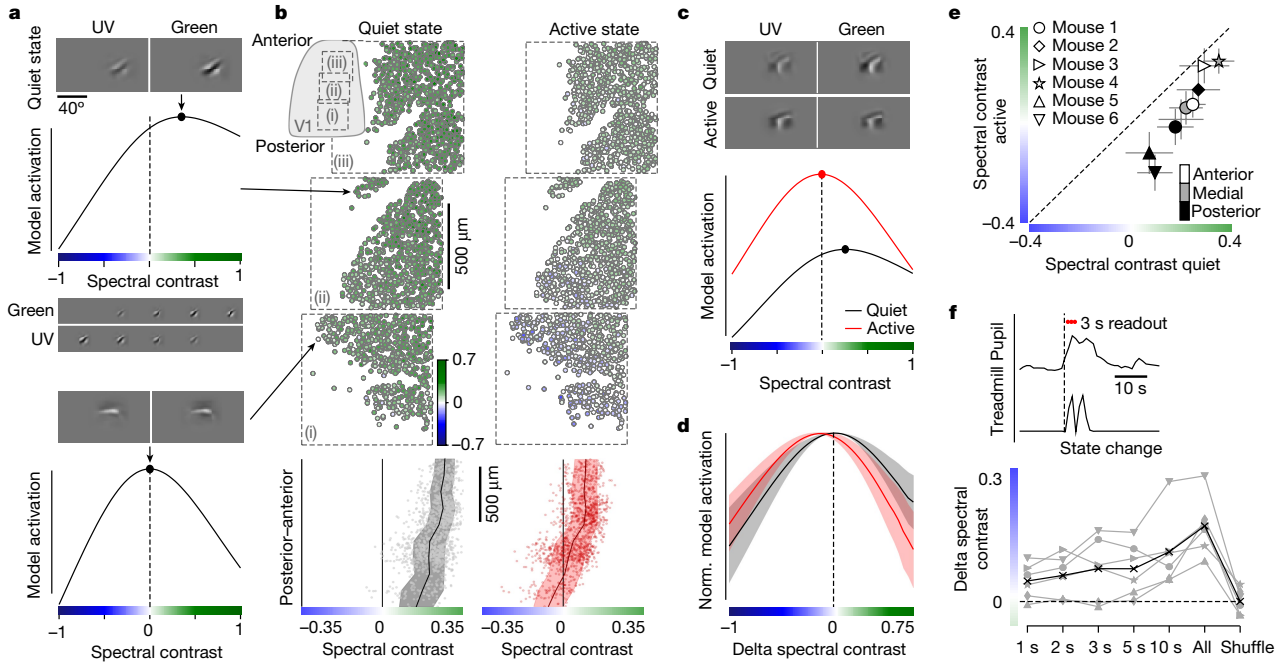


Fig. 2 | V1 colour-tuning changes with the behavioural state. a, MEIs optimized for a quiet state (3rd percentile of pupil and locomotion) and model activations for varying MEI spectral contrasts ($n = 50$) of two example neurons (from $n = 1{,}759$). Example stimuli are shown below. Arrows indicate the cortical position of neurons. **b**, Neurons ($n = 1{,}759$ neurons, $n = 3$ scans, $n = 1$ mouse) along the posterior–anterior V1, colour-coded on the basis of the spectral contrast of quiet and active state (97th percentile) MEIs. Inset shows the scan positions within V1. Bottom shows MEI spectral contrasts of neurons from the top, with binned average and s.d. shading. The spectral contrast significantly varied across the anterior–posterior V1 axis ($P = 10^{-16}$ for the smooth term on the cortical position of the generalized additive model (GAM); see Supplementary Methods for more details). **c**, MEIs of an example neuron optimized for a quiet and an active state, with colour-tuning curves shown below. **d**, Population mean with s.d. shading of peak-normalized colour-tuning curves from **b** and **c** aligned with respect to the peak of the tuning curves from the quiet state. The optimal spectral contrast shifted significantly towards higher UV sensitivity during active periods ($P = 10^{-16}$ for the behavioural state coefficient of the GAM). **e**, Mean MEI spectral contrast of quiet and active states across animals ($n = 478$ (mouse 1, posterior), 623 (mouse 1, medial), 658 (mouse 1, anterior), 843 (mouse 2), 711 (mouse 3), 822 (mouse 4), 769 (mouse 5), 706 (mouse 6) cells, $n = 8$ scans, $n = 6$ animals). Error bars indicate the s.d. across neurons. Wilcoxon signed-rank test (two-sided): $P = 10^{-78}$ (mouse 1, posterior), 10^{-103} (mouse 1, medial), 10^{-109} (mouse 1, anterior), 10^{-139} (mouse 2), 10^{-50} (mouse 3), 10^{-136} (mouse 4), 10^{-127} (mouse 5), 10^{-111} (mouse 6). **f**, Pupil size and treadmill velocity over time. Dashed line indicates the state change from quiet to active. Red dots indicate active trials used for analyses for a 3-s readout period. Bottom, change in mean MEI spectral contrast ($n = 6$ animals) between quiet and active states for different readout lengths after the state change, with mean across animals (black). All, all trials; Shuffle, shuffled behaviour relative to responses. One-sample t-test across animals (two-sided): $P = 0.038$ (1 s), $P = 0.029$ (2 s), $P = 0.053$ (3 s), $P = 0.03$ (5 s), $P = 0.021$ (10 s), $P = 0.001$ (All), $P = 0.92$ (Shuffled).

colour-tuning curve by predicting the activity of the neuron to varying colour contrasts of this MEI (Fig. 2a and Extended Data Fig. 5).

For both behavioural states, the optimal spectral contrast of neurons systematically varied along the anterior–posterior axis of V1 (Fig. 2b). The UV sensitivity significantly increased from anterior to posterior V1, which is in line with the distribution of cone opsins across the retina[16,17] and with previous studies of V1 (ref. [23]) and the dorsal lateral geniculate nucleus[24]. Nevertheless, for quiet behavioural periods, nearly all neurons preferred a green-biased stimulus (Fig. 2b, left), even the ones positioned in the posterior V1, which receives input from the ventral retina, where cones are largely sensitive to UV light[17]. This distribution of V1 colour preferences indicates that visual responses during quiet states are largely driven by rod photoreceptors that are sensitive to green light[25].

By contrast, during active periods, the colour tuning of neurons systematically shifted towards higher UV sensitivity (Fig. 2b–d). This was accompanied by an overall increase in neuronal activation predicted by the model (Fig. 2c and Extended Data Fig. 6a,d), which is in agreemnt with previous results[2,5]. The shift in colour selectivity was observed across animals for both the posterior and anterior V1 (Fig. 2e). As a result, neurons in the posterior V1 exhibited UV-biased MEIs, whereas neurons in the anterior V1 largely maintained their preference for green-biased stimuli. This is consistent with a cortical distribution of colour tuning expected from a shift from rod-dominated to cone-dominated visual responses[25]. Notably, the spatial structure of the MEIs was largely unchanged across behavioural states (Fig. 2c and Extended Data Fig. 5).

The shift in colour selectivity with behavioural state was fast, operating on the timescale of seconds (Fig. 2f). To test the temporal dynamics of the shift in tuning, we identified state changes from quiet to active periods by detecting rapid increases in pupil size after a prolonged quiet period. Then we sampled active trials within different time bins after the state change, trained CNN models on this subselection of active trials and all quiet trials and optimized MEIs as described above. The shift in colour selectivity with behavioural state was evident for a 10-s readout window for all animals tested. Notably, for the majority of animals ($n = 4$ out of 6), the shift was already present when training a model based on active trials that sampled just 1 s after the state change.

We wanted to confirm the above prediction from our in silico analysis that mouse V1 colour tuning rapidly shifts towards higher UV sensitivity during active periods. To that end, we used a well-established sparse noise paradigm for mapping the receptive fields of visual neurons

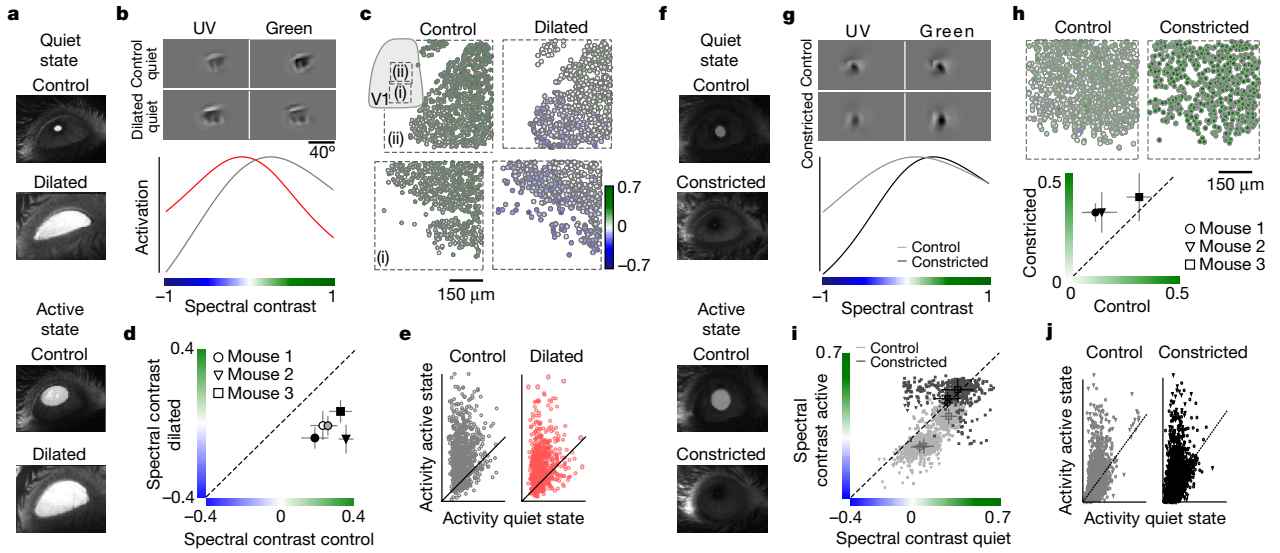


Fig. 3 | Pupil dilation causes the state-dependent shift in V1 colour selectivity. a, Example images from the eye camera during a quiet and an active state and for control and dilated conditions (atropine). **b**, MEIs of an example neuron (from $n = 478$) optimized for a quiet state for the control (black) and dilated (red) conditions (top) and peak-normalized colour-tuning curves (bottom). Neurons were matched anatomically across recordings. **c**, Neurons ($n = 1{,}101$) recorded in two experiments for the control (from Fig. 2) and the dilated condition, colour coded on the basis of the spectral contrast of the quiet state MEI. The spectral contrast significantly varied across the anterior–posterior V1 axis for the dilated condition ($n = 1{,}859$, $P = 10^{-16}$ for the smooth term on the cortical position of the GAM; see Supplementary Methods for more details). **d**, Mean spectral contrasts of quiet state MEIs in the control compared with the dilated condition ($n = 478$ (mouse 1, posterior, control), 623 (mouse 1, medial, control), 658 (mouse 1, anterior, control), 711 (mouse 2, control), 1,109 (mouse 3, drug), 464 (mouse 1, posterior, drug), 689 (mouse 1, medial, drug), 706 (mouse 1, anterior, drug), 723 (mouse 2, drug), 1,090 (mouse 3, drug) cells, $n = 10$ scans, $n = 3$ animals). Error bars indicate the s.d. across neurons. Two-sample t-test (two-sided): $P = 0$ for all scans. **e**, Mean activity of neurons from **c** during the quiet and active behavioural periods in the control and dilated conditions. **f**,**g**, Same as **a** (**f**) and **b** (**g**), but for pupil constriction with carbachol. **h**, Neurons recorded in posterior V1 ($n = 751$ (control) and 518 (constricted)), colour coded on the basis of the spectral contrast of a quiet state MEI. Bottom shows the mean spectral contrast of quiet state MEIs in control compared with the constricted condition ($n = 822$ (mouse 1, control), 769 (mouse 2, control), 1,109 (mouse 3, control), 751 (mouse 1, drug), 1,037 (mouse 2, drug), 1,028 (mouse 3, drug) cells, $n = 6$ scans, $n = 3$ mice). Error bars indicate the s.d. across neurons. Two-sample t-test (two-sided): $P = 0$ (mouse 1), 0 (mouse 2), 10^{-38} (mouse 3). **i**, Spectral contrast of quiet state MEIs compared with the spectral contrast of active state MEIs ($n = 778$ neurons, $n = 6$ scans, $n = 3$ mice), for the control (grey) and the constricted conditions (black). Only neurons with a test correlation value of >0.3 are shown. Wilcoxon signed-rank test (two-sided): $P = 10^{-134}$ (mouse 1, control), 10^{-127} (mouse 2, control), 10^{-170} (mouse 3, control), $P = 0.98$ (mouse 1, constricted), 0.0003 (mouse 2, constricted), 10^{-6} (mouse 3, constricted). **j**, Same as **e**, but for neurons from **h** in the control and the constricted conditions.

(Extended Data Fig. 7a). Trials were separated into quiet (<50th percentile) and active periods (>75th percentile) using the simultaneously recorded pupil size trace. For each neuron and behavioural state, we estimated a spike-triggered average (STA) that represented the preferred stimulus of the neuron in the context of the sparse noise input (Extended Data Fig. 7b). Consistent with the in silico analysis, we observed that most V1 neurons preferred a green-biased stimulus during the quiet behavioural state (Extended Data Fig. 7c). Moreover, neurons in the posterior and medial V1 showed increased UV sensitivity during active periods (Extended Data Fig. 7c,d). The UV shift was also present in the anterior V1, but only for more extreme pupil size thresholds (20th and 85th percentiles; Extended Data Fig. 7e). Finally, we confirmed that V1 colour preference shifted within a few seconds after onset of an active behavioural state (Extended Data Fig. 7e). Together, these results confirm the prediction of the CNN model that mouse V1 colour tuning rapidly changes with behavioural state, particularly for neurons that sample the upper visual field.

Pupil dilation shifts neuronal tuning

Next, we investigated the mechanism underlying the observed behaviour-related changes in colour tuning of mouse V1 neurons. On the one hand, the behavioural state of the animal affects neuronal activity through neuromodulation that acts on multiple stages of the visual system[6,8,26–28]. On the other hand, state-dependent pupil dilation results in higher light intensities at the level of the retina that might also affect visual processing[29,30].

To experimentally test the relative contribution of these two mechanisms, we dissociated state-dependent neuromodulatory effects from changes in pupil size by pharmacologically dilating and constricting the pupil with atropine and carbachol eye drops, respectively (Fig. 3a,f). We recorded visual responses to naturalistic scenes during control and pharmacology conditions and trained separate CNN models (Extended Data Fig. 2c).

Pupil dilation with atropine eye drops was sufficient to shift the colour tuning of neurons towards higher UV sensitivity, whereas locomotion activity was not necessary. During a quiet state with no locomotion, MEI colour tuning systematically shifted towards higher UV sensitivity for the dilated pupil compared with the control condition (Fig. 3b–d). We confirmed the role of pupil size in modulating colour tuning of mouse V1 neurons by also recording visual responses to the sparse noise stimulus after dilating the pupil with atropine (Extended Data Fig. 8).

To test whether pupil dilation is not only sufficient but also necessary for the behavioural shift in colour tuning, we dissociated pupil dilation from neuromodulation during active periods by temporarily constricting the pupil with carbachol eye drops (Extended Data Fig. 2f). The gain increase of neuronal responses with locomotion persisted under these pharmacological manipulations of the pupil[6,26,28]

(Fig. 3e,j), which indicated that this well-known effect of neuromodulation was unaffected. For quiet periods, pupil constriction resulted in a systematic shift towards higher green sensitivity compared with the control condition (Fig. 3g,h). Notably, we did not observe a significant shift towards higher UV sensitivity during active periods for the constricted condition, whereas the shift was evident in the control condition (Fig. 3i). This suggests that neuromodulation or other internal state-dependent mechanisms during active behavioural periods are not sufficient to drive the shift in colour tuning with behaviour, whereas state-dependent pupil dilation is necessary for the effect.

Tuning shift is caused by photoreceptors

Previous studies have shown that in mice, pupil size regulates retinal illuminance levels by more than one-order of magnitude[31]. This affects the relative activation levels of the green-sensitive rods and UV-sensitive and green-sensitive cones, thereby changing cortical colour preferences in anaesthetized mice[25]. To test whether our data could be explained by a shift from rod to cone photoreceptors during active behavioural periods because of a larger pupil (Fig. 4a), we estimated activation levels of mouse photoreceptors as a function of pupil size[10]. For our experiments, we observed up to a tenfold increase in pupil area and an equal increase in the estimated photoisomerization rate for an active compared with a quiet behavioural state (Fig. 4a, bottom). Therefore, the change in retinal light level due to pupil dilation during an active state is probably sufficient to dynamically shift the mouse visual system from a rod-dominated to a cone-dominated operating regimen.

If this was true, we would expect that the shift in colour selectivity can be reproduced for constant pupil sizes by changing ambient light levels. We experimentally confirmed this prediction by reducing the light intensity of the visual stimulus by 1.5-orders of magnitude while keeping the pupil size constant across recordings through pharmacological dilation with atropine (Fig. 4b). The low-light-intensity condition was expected to predominantly activate rod photoreceptors, which are green sensitive. Indeed, V1 neurons exhibited more green-biased MEIs for the low compared with the high light condition. Together with our pupil dilation and constriction experiments, this result strongly suggests that pupil dilation during active states results in a dynamic shift from rod-driven to cone-driven visual responses and a corresponding shift in spectral sensitivity.

Tuning shift affects population decoding

Next, we tested whether the shift in colour tuning during an active state might increase visual performance at the level of large populations of neurons in response to naturalistic stimuli. First, we applied a contrast-constrained image reconstruction paradigm[32] using the above-described trained CNN model (Extended Data Fig. 9a). Stimulus reconstruction from neuronal activity has previously been used to infer the most relevant visual features encoded by the neuron population[33], such as the colour sensitivity of neurons. Most reconstructed images for a quiet behavioural state exhibited higher contrast in the green channel, whereas the contrast was shifted towards the UV channel during active states (Extended Data Fig. 9b,c). This indicated that the increase in UV sensitivity during active periods observed at the single-cell level might contribute to specific visual tasks such as stimulus discrimination performed by populations of neurons in mouse V1.

We experimentally confirmed this prediction by showing that the decoding of UV objects selectively improved during active periods. To that end, we modified a recent object-decoding paradigm[34]. Mice passively viewed movie clips with two different objects presented in either the UV or green image channel (Fig. 5b) while recording the population calcium activity in the posterior V1 as described above. We estimated

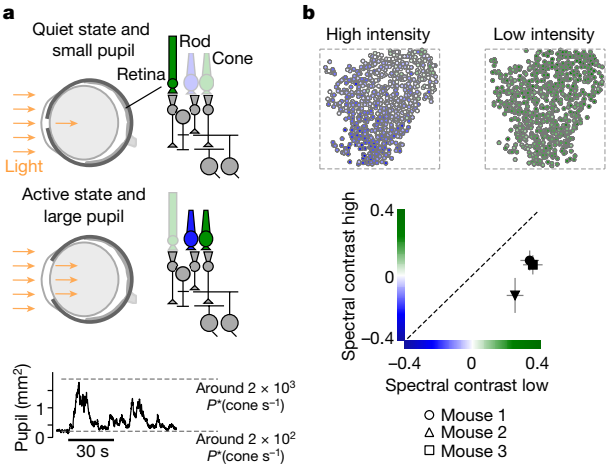


Fig. 4 | Pupil dilation during an active behavioural state differentially recruits rod and cone photoreceptors. a, Schematic of the mouse eye for a quiet behavioural state with a small pupil (top) and an active state with a large pupil (middle). Right, simplified circuit diagrams of the vertebrate retina. Activation of rod and cone photoreceptors are indicated by the degree of transparency. Arrows indicate the amount of light entering the eye through the pupil. Photoreceptors are coloured on the basis of their peak wavelength sensitivity. Bottom, pupil area recorded during functional imaging, with estimated photoisomerization rates (P^* (cone s^{-1})) for a small pupil and a large pupil. **b**, Top, neurons recorded in the posterior V1, colour coded on the basis of the spectral contrast of their quiet state MEI under the dilated condition for a high monitor intensity (n = 1,125 cells) and a 1.5-order of magnitude lower monitor intensity (n = 1,059 cells). Bottom, the mean spectral contrast of quiet state MEIs in low compared with a high monitor intensity condition (n = 1,125 (mouse 1, low), 651 (mouse 2, low), 1,090 (mouse 3, low), 1,059 (mouse 1, high), 627 (mouse 2, high), 1,068 (mouse 3, high) cells, n = 6 scans, n = 3 mice). Error bars indicate the s.d. across neurons. Two-sample t-test (two-sided): P = 0 for all scans.

the discriminability of object identity of UV and green objects from the recorded neuronal responses using a nonlinear support vector machine (SVM) decoder (Fig. 5a). Consistent with previous reports[1,35,36], decoding discriminability was higher during active compared with quiet behavioural periods (Fig. 5c). However, the increase in decoding discriminability of UV objects was larger than for green objects, which is consistent with an increase in UV sensitivity during active behavioural periods. This result was statistically significant compared with the result of a permutation test that shuffled quiet and active trials. The selective increase in decoding discriminability of UV objects was also present for a subset of recordings with modified stimuli, such as with reduced object contrast or different object polarity (Extended Data Fig. 10).

We then considered the behavioural relevance of this increase in UV sensitivity during an active state for mice. It has recently been shown that during dusk and dawn, aerial predators in the natural environment of mice are more visible in the UV than the green wavelength range[10] (Fig. 5d). Therefore, an increase in UV sensitivity of mouse visual neurons for an alert behavioural state might facilitate the detection of predators visible as dark silhouettes in the sky. To investigate this hypothesis on the level of populations of neurons, we presented parametric stimuli inspired by these natural scenes, which contained either only noise or an additional dark object in the green or UV image channel, to passively viewing mice (Fig. 5e). This experiment revealed that decoding detection of the behaviourally relevant stimulus—corresponding to the dark object being presented in the UV channel—was substantially increased for an active behavioural state. Decoding detection of the green objects did not increase to a similar extent (Fig. 5f). This result

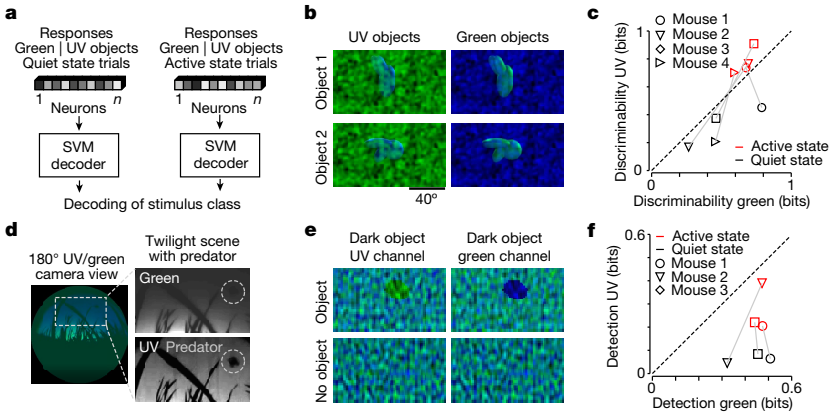


Fig. 5 | Shift in colour preference during an active state facilitates decoding of behaviourally relevant stimuli. a, Schematic illustrating the decoding paradigm. Neuronal responses for either quiet or active trials to green or UV objects were used to train a nonlinear SVM decoder to predict stimulus classes. **b**, Example stimulus frames of green and UV objects on top of noise. Stimulus conditions were presented as 5-s movie clips in random order. **c**, Scatter plot of the decoding discriminability of green compared with UV objects for quiet and active trials ($n = 4$ animals) for a SVM decoder trained on all neurons of each scan ($n = 1{,}090$ (mouse 1), 971 (mouse 2), 841 (mouse 3), 918 (mouse 4) cells). Grey lines connect the quiet and the active state performance of the same animal, with slopes larger than one indicating a larger increase in decoding performance for UV versus green objects. *P* values obtained from one-sided permutation test: $P < 0.002$ (mouse 1), $P < 0.044$ (mouse 2), $P < 0.024$ (mouse 3), $P < 0.01$ (mouse 4). **d**, Natural scene recorded at sunrise with a custom camera adjusted to the spectral sensitivity of mice[10], with a drone mimicking an aerial predator. Right images show single colour channels of a cropped image from the left, with the mock predator highlighted by a white dashed open circle. **e**, Parametric stimuli inspired by natural scene in **d** showing a dark object in either the UV or the green image channel (top) or noise only (bottom), with the object present or absent as the decoding objective. Stimuli were shown for 0.5 s with 0.3–0.5-s periods of grey screen in between. **f**, Similar to **c**, but for decoding detection of green versus UV dark objects from **e** ($n = 773$ (mouse 1), 1,049 (mouse 2), 1,094 (mouse 3) cells). *P* values were obtained from one-sided permutation test (see Methods for detail): $P < 0.008$ (mouse 1), $P < 0.009$ (mouse 2), $P < 0.008$ (mouse 3).

suggests that on the population level, the shift towards higher UV sensitivity might be behaviourally relevant as it selectively improves the decoding detection of dark objects in the UV channel, analogous to a predatory bird flying in a UV-bright sky.

Discussion

Our work identified a new mechanism by which state-dependent pupil dilation dynamically tunes the feature selectivity of the mouse visual system to behaviourally relevant stimuli.

The fact that sensory responses are modulated by the motor activity and the internal state of the animal was first demonstrated in elegant studies of invertebrates many decades ago[11,37]. Since then, modulation of sensory responses as a function of behavioural and internal states, such as attention, has been described in many animals[2,4,38,39]. Across animal species, state-dependent modulation predominantly affects neuronal responsiveness[2,9,27,28], which results in better behavioural performance[7,35,36,40]. In a few cases, however, the tuning properties of sensory circuits are also affected by this modulation. In the visual system, this has been reported, for instance, for temporal tuning in *Drosophila*[12], rabbits[39] and mice[41], as well as for direction selectivity in primates[4]. In these cases, the visual system might bias processing towards visual features relevant for current behavioural goals, such as higher temporal frequencies during periods of walking, running or flying.

Here, we demonstrated a shift in neuronal tuning with behavioural state in mice, focusing on the colour domain, which has rarely been studied in the context of behavioural modulation. Our results suggested that the shift towards higher UV sensitivity during active behavioural periods may help support ethological tasks, such as the detection of predators in the sky. In particular, UV vision has been implicated in predator and prey detection in several animal species as an adaptation to living in different natural environments (reviewed in ref. [42]). This is related to the stronger scattering of short wavelength light in general as well as ozone absorption[43] in the sky, which probably facilitate the detection of objects as dark silhouettes against a UV-bright background in the sky[10], underwater and against the snow[42]. However, it will be important to directly test the behavioural relevance of the described shift in colour tuning during an active state for mouse predator detection. For example, combining an overhead detection task of a looming stimulus presented in UV or green light conditions[44] with pharmacological pupil manipulations or careful tracking of pupil dynamics[45] will reveal whether pupil dilation results in better behavioural detection of UV stimuli, as suggested by our results.

Mechanistically, state-dependent modulation of visual responses has been linked to neuromodulators such as acetylcholine and noradrenaline (reviewed in refs. [13,14]), which are released with active behavioural states and alert internal states. Our results demonstrated that in addition to internal brain state mechanisms, dynamic changes in pupil size are both sufficient and necessary to affect cortical tuning (see also Supplementary Discussion). We propose that this mechanism changes colour sensitivity through differential rod versus cone activation, which is reminiscent of the Purkinje shift described in humans[46], although acting on faster timescales. A recent neurophysiological study[25] that used anaesthetized mice demonstrated that pharmacological pupil dilation at constant ambient light levels is sufficient to induce a shift from rod-driven to cone-driven visual responses in V1. Our data indicated that a switch between the rod and cone system can also happen dynamically at the timescale of seconds in behaving mice as a consequence of changes in pupil size across distinct behavioural states. As rod and cone photoreceptors differ with respect to spatial distribution, temporal resolution and degree of nonlinearity (discussed in ref. [47]), dynamically adjusting their relative activation might influence the sensory representation of the visual scene far beyond the colour domain of the visual input.

Changes in pupil size driven by behavioural and internal states of the animal are common features shared across most vertebrate species studied so far (reviewed in ref. [48]), including amphibians, birds and mammals (see also Supplementary Discussion). Notably, pupil

dilation is probably under voluntary control for some animals such as birds and reptiles (discussed in ref. [49]), and potentially even for some humans[50]. We propose that state-dependent pupil size changes might act as a general mechanism across species to rapidly switch between the rod-driven and cone-driven operating regimen, thereby tuning the visual system to different features, as suggested here for predator detection in mice during dusk and dawn. Our findings provide a functional explanation to the long-standing debate of why pupil size is modulated with internal and behavioural states.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05270-3.

1. Reimer, J. et al. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron* **84**, 355–362 (2014).
2. Niell, C. M. & Stryker, M. P. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65**, 472–479 (2010).
3. Vinck, M., Batista-Brito, R., Knoblich, U. & Cardin, J. A. Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron* **86**, 740–754 (2015).
4. Treue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).
5. Erisken, S. et al. Effects of locomotion extend throughout the mouse early visual system. *Curr. Biol.* **24**, 2899–2907 (2014).
6. Reimer, J. et al. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat. Commun.* **7**, 13289 (2016).
7. Bennett, C., Arroyo, S. & Hestrin, S. Subthreshold mechanisms underlying state-dependent modulation of visual responses. *Neuron* **80**, 350–357 (2013).
8. Liang, L. et al. Retinal inputs to the thalamus are selectively gated by arousal. *Curr. Biol.* **30**, 3923–3934.e9 (2020).
9. McAdams, C. J. & Maunsell, J. H. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.* **19**, 431–441 (1999).
10. Qiu, Y. et al. Natural environment statistics in the upper and lower visual field are reflected in mouse retinal specializations. *Curr. Biol.* **31**, 3233–3247.e6 (2021).
11. Rowell, C. H. Variable responsiveness of a visual interneurone in the free-moving locust, and its relation to behaviour and arousal. *J. Exp. Biol.* **55**, 727–747 (1971).
12. Chiappe, M. E., Seelig, J. D., Reiser, M. B. & Jayaraman, V. Walking modulates speed sensitivity in *Drosophila* motion vision. *Curr. Biol.* **20**, 1470–1475 (2010).
13. Busse, L. The influence of locomotion on sensory processing and its underlying neuronal circuits. *eNeuroforum* **24**, A41–A51 (2018).
14. Schneider, D. M. Reflections of action in sensory cortex. *Curr. Opin. Neurobiol.* **64**, 53–59 (2020).
15. Gerl, E. J. & Morris, M. R. The causes and consequences of color vision. *Evol. Educ. Outreach* **1**, 476–486 (2008).
16. Szél, A. et al. Unique topographic separation of two spectral classes of cones in the mouse retina. *J. Comp. Neurol.* **325**, 327–342 (1992).
17. Baden, T. et al. A tale of two retinal domains: near-optimal sampling of achromatic contrasts in natural scenes through asymmetric photoreceptor distribution. *Neuron* **80**, 1206–1217 (2013).
18. Walker, E. Y. et al. Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.* **22**, 2060–2065 (2019).
19. Lurz, K.-K. et al. Generalization in data-driven models of primary visual cortex. In *Proc. International Conference on Learning Representations* (2021).
20. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
21. Franke, K. et al. An arbitrary-spectrum spatial visual stimulator for vision research. *eLife* **8**, e48779 (2019).
22. Liu, R. et al. An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems* (2018).
23. Rhim, I., Coello-Reyes, G., Ko, H.-K. & Nauhaus, I. Maps of cone opsin input to mouse V1 and higher visual areas. *J. Neurophysiol.* **117**, 1674–1682 (2017).
24. Denman, D. J., Siegle, J. H., Koch, C., Reid, R. C. & Blanche, T. J. Spatial organization of chromatic pathways in the mouse dorsal lateral geniculate nucleus. *J. Neurosci.* **37**, 1102–1116 (2017).
25. Rhim, I., Coello-Reyes, G. & Nauhaus, I. Variations in photoreceptor throughput to mouse visual cortex and the unique effects on tuning. *Sci. Rep.* **11**, 11937 (2021).
26. Fu, Y. et al. A cortical circuit for gain control by behavioral state. *Cell* **156**, 1139–1152 (2014).
27. Schröder, S. et al. Arousal modulates retinal output. *Neuron* **107**, 487–495.e9 (2020).
28. Eggermann, E., Kremer, Y., Crochet, S. & Petersen, C. C. H. Cholinergic signals in mouse barrel cortex during active whisker sensing. *Cell Rep.* **9**, 1654–1660 (2014).
29. Tikidji-Hamburyan, A. et al. Retinal output changes qualitatively with every change in ambient illuminance. *Nat. Neurosci.* **18**, 66–74 (2015).
30. Grimes, W. N., Schwartz, G. W. & Rieke, F. The synaptic and circuit mechanisms underlying a change in spatial encoding in the retina. *Neuron* **82**, 460–473 (2014).
31. Pennesi, M. E., Lyubarsky, A. L. & Jr. Pugh, E. N. Extreme responsiveness of the pupil of the dark-adapted mouse to steady retinal illumination. *Invest. Ophthalmol. Vis. Sci.* **39**, 2148–2156 (1998).
32. Safarani, S. et al. Towards robust vision by multi-task learning on monkey visual cortex. In *Advances in Neural Information Processing Systems* (2021).
33. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252**, 1854–1857 (1991).
34. Froudarakis, E. et al. Object manifold geometry across the mouse cortical visual hierarchy. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.20.258798 (2020).
35. Dadarlat, M. C. & Stryker, M. P. Locomotion enhances neural encoding of visual stimuli in mouse V1. *J. Neurosci.* **37**, 3764–3775 (2017).
36. Spitzer, H., Desimone, R. & Moran, J. Increased attention enhances both behavioral and neuronal performance. *Science* **240**, 338–340 (1988).
37. Wiersma, C. A. & Oberjat, T. The selective responsiveness of various crayfish oculomotor fibers to sensory stimuli. *Comp. Biochem. Physiol.* **26**, 1–16 (1968).
38. Maimon, G., Straw, A. D. & Dickinson, M. H. Active flight increases the gain of visual motion processing in *Drosophila*. *Nat. Neurosci.* **13**, 393–399 (2010).
39. Bezdudnaya, T. et al. Thalamic burst mode and inattention in the awake LGNd. *Neuron* **49**, 421–432 (2006).
40. de Gee, J. W. et al. Mice regulate their attentional intensity and arousal to exploit increases in task utility. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.04.482962 (2022).
41. Andermann, M. L., Kerlin, A. M., Roumis, D. K., Glickfeld, L. L. & Reid, R. C. Functional specialization of mouse higher visual cortical areas. *Neuron* **72**, 1025–1039 (2011).
42. Cronin, T. W. & Bok, M. J. Photoreception and vision in the ultraviolet. *J. Exp. Biol.* **219**, 2790–2801 (2016).
43. Hulburt, E. O. Explanation of the brightness and color of the sky, particularly the twilight sky. *J. Opt. Soc. Am.* **43**, 113–118 (1953).
44. Storchi, R. et al. Measuring vision using innate behaviours in mice with intact and impaired retina function. *Sci. Rep.* **9**, 10396 (2019).
45. Meyer, A. F., Poort, J., O'Keefe, J., Sahani, M. & Linden, J. F. A head-mounted camera system integrates detailed behavioral monitoring with multichannel electrophysiology in freely moving mice. *Neuron* **100**, 46–60.e7 (2018).
46. Wald, G. Human vision and the spectrum. *Science* **101**, 653–658 (1945).
47. Lamb, T. D. Why rods and cones? *Eye* **30**, 179–185 (2016).
48. Larsen, R. S. & Waters, J. Neuromodulatory correlates of pupil dilation. *Front. Neural Circuits* **12**, 21 (2018).
49. Douglas, R. H. The pupillary light responses of animals; a review of their distribution, dynamics, mechanisms and functions. *Prog. Retin. Eye Res.* **66**, 17–48 (2018).
50. Eberhardt, L. V., Grön, G., Ulrich, M., Huckauf, A. & Strauch, C. Direct voluntary control of pupil constriction and dilation: exploratory evidence from pupillometry, optometry, skin conductance, perception, and functional MRI. *Int. J. Psychophysiol.* **168**, 33–42 (2021).

Methods

Neurophysiological experiments

All procedures were approved by the Institutional Animal Care and Use Committee of Baylor College of Medicine. Owing to the explanatory nature of our study, we did not use randomization or blinding. No statistical methods were used to predetermine sample sizes.

Mice of either sex (*Mus musculus*, $n = 13$; 6 weeks to 5 months of age) expressing GCaMP6s in excitatory neurons through the Slc17a7-Cre and the Ai162 transgenic lines (stock numbers 023527 and 031562, respectively, The Jackson Laboratory) were anaesthetized, and a 4-mm craniotomy was made over the visual cortex of the right hemisphere as previously described[1,51]. For functional recordings, awake mice were head-mounted above a cylindrical treadmill, and calcium imaging was performed using a Ti-Sapphire laser tuned to 920 nm and a two-photon microscope equipped with resonant scanners (Thorlabs) and a ×25 objective (MRD77220, Nikon). The laser power after the objective was kept below 60 mW. The rostral–caudal treadmill movement was measured using a rotary optical encoder with a resolution of 8,000 pulses per revolution. We used light diffusing from the laser through the pupil to capture eye movements and pupil size. Images of the pupil were reflected through a hot mirror and captured with a GigE CMOS camera (Genie Nano C1920M; Teledyne Dalsa) at 20 fps at 1,920 × 1,200 pixel resolution. The contour of the pupil for each frame was extracted using DeepLabCut[52], and the centre and major radius of a fitted ellipse were used as the position and dilation, respectively, of the pupil.

For image acquisition, we used ScanImage. To identify V1 boundaries, we used pixelwise responses to drifting bar stimuli of a 2,400 × 2,400 μm scan at 200 μm depth from the cortical surface[53], recorded using a large-field-of-view mesoscope[54] not used for other functional recordings. In V1, imaging was performed using 512 × 512 pixel scans (650 × 650 μm) recorded at approximately 15 Hz and positioned within L2/3 at around 200 μm from the surface of the cortex. Imaging data were motion-corrected, automatically segmented and deconvolved using the CNMF algorithm[55]; cells were further selected by a classifier trained to detect somata based on the segmented masks. In addition, we excluded cells with low stimulus correlation. For this, we computed the first principal component (PC) of the response matrix of the size number of neurons × number of trials. For each neuron, we then estimated the linear correlation of its responses to the first PC, as the first PC captured unrelated background activity. We excluded neurons with a correlation lower or higher than −0.25 or 0.25, respectively. This resulted in 450–1,100 selected soma masks per scan depending on the response quality and the blood vessel pattern. A structural stack encompassing the scan plane and imaged at 1.6 × 1.6 × 1 μm *xyz* resolution with 20 repeats per plane was used to register functional scans of the same neurons into a shared *xyz* frame of reference. Cells registered to the same 3D stack were then anatomically matched for distances of <10 μm. For inception loop experiments, we confirmed the anatomical matching with a functional matching procedure, using the responses of cells to the same set of test images (see also ref. [18]) and only included anatomically matched neurons with a response correlation of >0.5 for further analysis. To bring different recordings of the same animal across the posterior–anterior axis of V1 into the same frame of reference, we manually aligned the mean image of each functional recording to the mean image of the 2,400 × 2,400 μm scan acquired at the mesoscope (see above) using the blood vessel pattern. Then, each cell within the functional scan was assigned a new *xy* coordinate (in μm) in the common frame of reference. To illustrate coarse differences across visual space, scan fields were manually assigned into three broad location categories within V1 (posterior, medial and anterior) depending on their position relative to V1 boundaries.

Visual stimulation

Visual stimuli were presented to the left eye of the mouse on a 42 × 26 cm light-transmitting Teflon screen (McMaster-Carr) positioned 12 cm from the animal, covering approximately 120 × 90° visual angle. Light was back-projected onto the screen by a DLP-based projector (EKB Technologies)[21] with UV (395 nm) and green (460 nm) LEDs that differentially activated mouse S-opsin and M-opsin. LEDs were synchronized with the scan retrace of the microscope. Note that the UV LED not only drives UV-sensitive S-opsin but also slightly activates green-sensitive M-opsin and rhodopsin because of their sensitivity tail for shorter wavelengths (β-band). This cross-activation could be addressed by using a silent substitution protocol, whereby one type of photoreceptor is selectively stimulated by presenting a steady excitation to all other photoreceptor types using a counteracting stimulus. However, this comes at the cost of overall contrast. We considered that our imperfect spectral separation of photoreceptor types was suitable to investigate most questions concerning chromatic processing in the visual system (discussed in ref. [21]), especially as photoreceptor-type-isolating stimulation in natural scenes is rare.

Light intensity (measured as the estimated photoisomerization rate, P^*(cone s^{-1})) was calibrated using a spectrometer (USB2000+, Ocean Optics) to result in equal activation rates for mouse M-opsin and S-opsin (for details see ref. [21]). In brief, the spectrometer output was divided by the integration time to obtain counts per s and then converted into electrical power (in nW) using the calibration data (in μJ per count) provided by Ocean Optics. The intensity (in μW) of the entire screen set to maximal intensity (255 pixel values) was approximately 1.28 and 1.39 for green and UV LEDs, respectively. To obtain the estimated photoisomerization rate per photoreceptor type, we first converted electrical power into energy flux (in eV s^{-1}) and then calculated the photon flux (in photons s^{-1}) using the photon energy (in eV). The photon flux density (in photons s^{-1} μm^{-2}) was then computed and converted into the photoisomerization rate using the effective activation of mouse cone photoreceptors by the LEDs and the light collection area of cone outer segments. In addition, we considered both the wavelength-specific transmission of the mouse optical apparatus[56] and the ratio between pupil size and retinal area[57]. See the calibration iPython notebook provided online (https://github.com/katrinfranke/open-visual-stimulator) for further details. For a pupil area of 0.2 mm^2 during quiet trials and maximal stimulus intensities (255 pixel values), this resulted in 400 P^*(cone s^{-1}) corresponding to the mesopic range. During active periods, the pupil area increased to 1.9 mm^2, resulting in 4,000 P^*(cone s^{-1}) corresponding to the low photopic regimen.

Before functional recordings, the screen was positioned such that the population receptive field across all neurons, estimated using an achromatic sparse noise paradigm, was within the centre of the screen. The screen position was fixed and kept constant across recordings of the same neurons. We used Psychtoolbox in MatLab for stimulus presentation and showed the following light stimuli.

Natural images. We presented naturalistic scenes from the available ImageNet online database[58]. We selected images on the basis of two criteria (Extended Data Fig. 1). First, to avoid an intensity bias in the stimulus, we selected images with no significant difference in the mean intensity of the blue and green image channels across all images. Second, we selected images with high pixelwise mean squared error (MSE > 85) across colour channels to increase chromatic contrast, resulting in a lower pixel-wise correlation across colour channels compared with a random selection. Then, we presented the blue and green image channels using the UV and green LEDs of the projector, respectively. For a single scan, we presented 4,500 unique coloured and 750 monochromatic images in UV and green, respectively. We added monochromatic images to the stimulus to include images without correlations across colour channels, thereby diversifying the input to the model. As the test set, we used 100 coloured and 2 × 25 monochromatic images that were repeated 10 times uniformly spread throughout the recording. Each image was presented for 500 ms, followed by a grey screen (UV and green LEDs at 127 pixel value) for 300–500 ms, sampled

uniformly from that range. The mean intensity of presented natural images across the green and UV colour channels varied between 5 and 204 (8-bit, gamma-corrected). For a small pupil during quiet states, this corresponded to approximately 8 and 320 photoisomerizations (P^*) per cone and second (P^*(cone s^{-1})). Each natural image was preceded by a grey blank period (all pixel values set to 127), which reduced the range of monitor intensities to approximately 57.2–213 P^*(cone s^{-1}) when integrating over 1 s, spanning less than one-order of magnitude. For the light intensities we were using, previous studies have found that the pupil size is relatively constant for changes in ambient light intensities below one-order of magnitude[31,59]. Indeed, we found that ambient monitor intensity does not contribute strongly to the recorded changes in pupil size (Extended Data Fig. 1).

Sparse noise. To map the receptive fields of V1 neurons, we used a sparse noise paradigm. UV and green bright (pixel value of 255) and dark (pixel value of 0) dots of approximately 10° visual angle were presented on a grey background (pixel value of 127) in a randomized order. Dots were presented for eight and five positions along the horizontal and vertical axis of the screen, respectively, excluding screen margins. Each presentation lasted 200 ms and each condition (for example, UV bright dot at position $x = 1$ and $y = 1$) was repeated 50 times. For a subset of recordings ($n = 2$ animals, $n = 3$ scan fields; compare with Extended Data Fig. 7e), each condition was repeated 150 times to increase the number of trials for more extreme behavioural states.

Full-field binary white noise. We used a binary full-field noise stimulus of UV and green LEDs to estimate temporal kernels of V1 neurons. The intensity of UV and green LEDs was determined independently by a balanced 15-min random sequence updated at 10 Hz. A similar stimulus was recently used in recordings of mouse[60] and zebrafish retina[61].

Coloured objects. To test for object discrimination, we used two synthesized objects rendered in Blender (https://www.blender.org) as previously described[34]. In brief, we smoothly varied object position, size, tilt and axial rotation. For bright objects, we also varied either the location or energy of four light sources. Stimuli were rendered as bright objects on a black screen and Gaussian noise in the other colour channel (condition 1), bright and dark objects on a grey screen and Gaussian noise in the other colour channel (conditions 2 and 3) or as bright objects on a black screen without Gaussian noise (condition 4). Per object and condition, we rendered movies of 875 s, which we then divided into 175 5-s clips. We presented the clips with different conditions and objects in a random order.

Images with dark objects. For the object detection task, we generated images with independent Perlin noise[62] in each colour channel using the perlin-noise package for Python (https://pypi.org/project/perlin-noise/). For all images except the noise images, we added a dark ellipse (pixel value of 0) of varying size, position and angle to one of the colour channels. We adjusted the contrast of all images with a dark object to match the contrast of noise images, such that the distribution of image contrasts did not differ between noise and object images. We presented 2,000 unique noise images and 2,000 unique images with a dark object in the UV and green image channels, respectively. Each image was presented for 500 ms, followed by a grey screen (UV and green LEDs at 127 pixel value) for 300–500 ms, sampled uniformly from that range.

For the presentation of naturalistic scenes and object movies and images, we applied a gamma function of 1.9 to the 8-bit pixel values of the monitor.

Preprocessing of neuronal responses and behavioural data

Neuronal responses were first deconvolved using constrained non-negative calcium deconvolution[55]. For all stimulus paradigms except the full-field binary white noise stimulus, we subsequently extracted the accumulated activity of each neuron between 50 ms after stimulus onset and offset using a Hamming window. For the presentation of objects, we segmented the 5-s clips into 9 bins of 500 ms, starting 250 ms after stimulus onset. Behavioural traces were extracted using the same temporal offset and integration window as deconvolved calcium traces. To train our models, we isotropically downsampled stimuli images to 64 × 36 pixels. Input images, the target neuronal activities, behavioural traces and pupil positions were normalized across the training set during training.

Pharmacological manipulations

To pharmacologically dilate and constrict the pupil, we applied 1–3% atropine and carbachol eye drops, respectively, to the left eye of the animal facing the screen for visual stimulation. Functional recordings started after the pupil was dilated or constricted. Pharmacological pupil dilation lasted >2 h, enabling the use of all the data for further analysis. By contrast, carbachol eye drops constricted the pupil for approximately 30 min and were re-applied once during the scan. For analysis, we only selected trials with constricted pupils and we matched data analysed in the control scans to the same trial numbers.

Sparse noise spatial receptive field mapping

We estimated spatial STAs of V1 neurons in response to the sparse noise stimulus by multiplying the stimulus matrix with the response matrix of each neuron[63] separately for each stimulus colour and polarity as well as behavioural state. For the behavioural state, we separated trials into small (<50th percentile) and large pupil trials (>75th percentile). We used different pupil size thresholds for the two behavioural states compared to the model owing to the shorter recording time. For recordings with pupil dilation, we used locomotion speed instead of pupil size to separate trials into two behavioural states. For each behavioural state, STAs computed on the basis of on and off dots were averaged to produce one STA per cell and stimulus colour. Green and UV STAs of the same behavioural state were peak-normalized to the same maximum. To assess STA quality, we generated response predictions by multiplying the flattened STA of each neuron with the flattened stimulus frames and compared the predictions to the recorded responses by estimating the linear correlation coefficient. For analysis, we only included cells for which the correlation was >0.2 for at least one of the stimulus conditions.

In contrast to the modelling results, the STA spectral contrast for a quiet state varied only slightly across the anterior–posterior axis of the V1. This was probably due to the different pupil size thresholds for quiet and active state used in the STA paradigm compared to the model. To verify this, we used the data in response to natural images (Fig. 2) to train a separate model without behaviour as input channels on trials with small pupil (<50th percentile) and subsequently optimized MEIs, which is a procedure more similar to the STA paradigm. When looking at the spectral contrast of the resulting MEIs, we observed a smaller variation of colour preference across the anterior–posterior axis of V1, thereby confirming our prediction (data not shown).

To confirm that the shift in colour preference with behaviour in response to the sparse noise was not dependent on the specific pupil size thresholds we used, we presented 150 instead of 50 repeats per stimulus condition in a subset of experiments. The larger number of trials for more extreme behavioural states allowed us to compute STAs for behavioural states more similar to the model (<20th versus >85th percentile). This resulted in a stronger shift in colour preference during active periods compared with the lower thresholds of pupil sizes (data not shown), which indicated that we had probably underestimated the effect for the shorter recordings shown in Extended Data Fig. 7a–c.

Full-field binary noise temporal receptive field mapping

We used the responses to the 10 Hz full-field binary noise stimulus of UV and green LEDs to compute temporal STAs of V1 neurons. Specifically,

we upsampled both stimulus and responses to 60 Hz and then multiplied the stimulus matrix with the response matrix of each neuron. Per cell, this resulted in a temporal STA in response to UV and green flickers, respectively. The kernel quality was measured by comparing the variance of each temporal STA with the variance of the baseline, defined as the first 100 ms of the STA. Only cells with at least five times more variance of the kernel compared with baseline were considered for further analysis.

Simulated data using Gabor neurons

We simulated neurons with Gabor receptive fields with varying Gabor parameters across the two colour channels. We normalized each Gabor receptive field to have a background of 0 and an amplitude range between −1 and 1. To generate responses of simulated neurons, we used the same set of training images presented during functional recordings. First, we subtracted the mean across all images from the training set, multiplied each Gabor receptive field with each training image and computed the sum of each multiplication across the two colour channels c. We then passed the resulting scalar response per neuron through a rectified linear unit (ReLU) to obtain the simulated response r, such that

$$r = \mathrm{ReLU}\left(\sum_{c,x,y} \mathrm{image}_{c,x,y}\mathrm{Gabor}_{c,x,y}\right),$$

where

$$\mathrm{Gabor}_{c,x,y} = \alpha_c \exp\left(-\frac{x'^2 + \gamma_c^2 y'^2}{2\sigma_c^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \psi_c\right)$$

with $x' = x\cos(\theta_c) + y\sin(\theta_c)$ and $y' = -x\sin(\theta_c) + y\cos(\theta_c)$. We varied orientation θ, size σ, spatial aspect ratio γ, phase ψ and colour preference α independently for each colour channel and neuron, while keeping spatial frequency λ constant across all neurons. Finally, we passed the simulated responses r through a Poisson process and normalized the responses by the respective standard deviation of the responses across all images. We used the responses of the simulated Gabor neurons together with the natural images to train the model (see below). Our model recovered both the colour opponency and the colour preference of simulated neurons. Only extreme colour preferences were slightly underestimated by our model, which is probably due to high correlations across the colour channels of natural scenes.

In silico tuning characterization

It has been our main interest to investigate the change in tuning properties with the behavioural state of animals. Ideally, this includes manipulating the behaviour of the animal and investigating the resulting effect on different visual tuning properties. Although this is experimentally challenging and time-consuming, it is straightforward with a deep-learning-based neuronal predictive model that emulates the biological circuit. This allowed us to selectively study how tuning to colour or spatial features changes with behaviour. To perform our in in silico tuning characterization, we created a CNN model, which was split into two parts: the core and the readout. The core computed latent features from the inputs, which were shared among all neurons. The readout was learned per neuron and mapped the output features of the core onto the neuronal responses through regularized regression.

Core of the CNN model. We based our model on the work from ref. [19], as it was demonstrated to set the state of the art for predicting the responses of a population of mouse V1 neurons. In brief, we modelled the core as a 4-layer CNN, with 64 feature channels per layer. Each layer consisted of a 2D convolutional layer followed by a batch-normalization layer and ELU nonlinearity[64,65]. Except for the first layer, all convolutional layers were depth-separable convolutions[66], which led to better performance while reducing the number of core parameters. Each depth-separable layer consisted of a 1 × 1 pointwise convolution followed by a 7 × 7 depth-wise convolution, again followed by a 1 × 1 pointwise convolution. Without stacking the outputs of the core, the output tensor of the last layer was passed on to the readout.

Readout of the CNN model. To obtain the scalar firing rate for each neuron, we computed a linear regression between the core output tensor of dimensions $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ (w, width; h, height; c, channels) and the linear weight tensor $\mathbf{w} \in \mathbb{R}^{c \times w \times h}$, followed by an exponential linear unit (ELU) offset by one (ELU+1) to keep the response positive. We made use of the recently proposed Gaussian readout[19], which considerably simplifies the regression problem. Our Gaussian readout learned the parameters of a 2D Gaussian distribution $\mathcal{N}(\mu_n, \Sigma_n)$ and sampled a location of height and width in the core output tensor in each training step for every image and neuron. Given a large enough initial Σ_n to ensure gradient flow, Σ_n, that is, the uncertainty about the readout location, decreased during training for more reliable estimates of the mean location μ_n, which represented the centre of the receptive field of a neuron. At inference time (that is, when evaluating our model), we set the readout to be deterministic and to use the fixed position μ_n. We therefore learned a position of a single point in core feature space for each neuron. In parallel to learning the position, we learned the weights of the weight tensor of the linear regression of size c per neuron. Furthermore, we made use of the retinotopic organization of V1 by coupling the recorded cortical 2D coordinates $\mathbf{p}_n \in \mathbb{R}^2$ of each neuron with the estimation of the receptive field position μ_n of the readout. We achieved this by learning a common function $\mu_n = f(\mathbf{p}_n)$, which is shared by all neurons. We set f to be a randomly initialized linear fully connected network of size 2-2 followed by tanh nonlinearity.

Shifter network. Because we used a free viewing paradigm when presenting the visual stimuli to the head-fixed mice, the receptive field positions of the neurons with respect to the presented images had considerable trial-to-trial variability due to eye movements. To inform our model of the trial-dependent shift of the receptive fields of neurons, we shifted μ_n, the receptive field centre of the model neuron, using the estimated pupil centre (see the section 'Neurophysiological experiments'). We accomplished this by passing the pupil centre through a small shifter network, a three-layer fully connected network with $n = 5$ hidden features, again followed by a tanh nonlinearity, that calculates the shift Δx and Δy per trial. The shift was then added to μ_n of each model neuron.

Input of behaviour and image position encoding. In addition to the green and UV channels of the visual stimulus, we appended five extra channels to each input to the model. We added three channels of the recorded behavioural parameters in each given trial (pupil size, instantaneous change of pupil size and locomotion speed), such that each channel simply consisted of the scalar for the respective behavioural parameter, transformed into the stimulus dimensions. This enabled the model to predict neuronal responses as a function of both visual input and behaviour and therefore to learn the relationship between behavioural states and neuronal activity. This modification enabled us to investigate the effect of behaviour by selecting different inputs in the behavioural channels while optimizing the image channels. Furthermore, we added a positional encoding to the inputs, which consisted of two channels that encoded the horizontal and vertical pixel positions of the visual stimulus. These encodings can be thought of as simple greyscale gradients in either direction, with values from [−1, ..., 1]. Appending position encodings of this kind has been shown to improve the ability of CNNs to learn spatial relationships between pixel positions of the input image and high level feature representations[22]. We found that including

the position embedding increased the performance of our model (Extended Data Fig. 2b). We also observed a smoother gradient of colour tuning across the different scan fields (Fig. 2b and Extended Data Fig. 6b) when adding the position encoding. This indicated that the model learned the well-described colour sensitivity tuning of mouse cone photoreceptors across visual space.

Model training and evaluation

We first split the unique training images into the training and validation set, using a split of 90% to 10%, respectively. Then we trained our networks with the training set by minimizing the Poisson loss $\frac{1}{m}\sum_{i=1}^{m}(\hat{r}^{(i)} - r^{(i)}\log\hat{r}^{(i)})$, where m denotes the number of neurons, $\hat{r}$ the predicted neuronal response and r the observed response. After each full pass through the training set (that is, epoch), we calculated the correlation between the predicted and the measured responses across all neurons on the validation set: if the correlation failed to increase during a fixed number of epochs, we stopped the training and restored the model to its state after the best performing epoch. After each stopping, we either decreased the learning rate or stopped training altogether if the number of learning-rate decay steps was reached. We optimized the parameters of the network through stochastic gradient descent using the Adam optimizer[67]. Furthermore, we performed an exhaustive hyperparameter selection using a Bayesian search on a held-out dataset. All parameters and hyperparameters can be found in our GitHub repository (see the Code availability section). When evaluating our models on the test set (Extended Data Fig. 2a–c), we used two different types of correlation. First, referred to as test correlation, we computed the correlation between the prediction by the model and neuronal responses across single trials, including the trial-by-trial variability across repeats. Second, we computed the correlation of the predicted responses with the average responses across repeats and refer to it here as the correlation to average. We also computed the fraction of variance explained, using $\hat{r}^2_{\mathrm{ER}}$ proposed in ref. [68], which provides an unbiased estimate of the variance explained based on the expected neuronal response across image repetitions. However, our model computed different predictions for each repetition of a given test set image because we also fed the behavioural parameters of each trial into the model. We therefore simply averaged the model responses across repetitions and calculated the $\hat{r}^2_{\mathrm{ER}}$ accordingly. When evaluating the model performance for the pharmacology conditions (Extended Data Fig. 2c), we found that they led to a lower model performance compared with the control condition. This could be due to the fact that for the dilated condition, we did not incorporate pupil-related behavioural parameters into the model owing to difficulties in pupil tracking for this pharmacological condition. For the drug condition with carbachol, we selected a subset of trials in which the pupil was constricted (see the 'Pharmacological manipulations' section), which led to fewer trials to train the models with. Finally, for some of our datasets that had either a low number of trials or a low yield of neurons, we trained a single model on multiple datasets[19], such that the convolutional core of the model was trained with more examples. The training of the per-neuron readout was unaffected by this joint training of datasets. We assigned a model identifier to each trained model (which can be found in Supplementary Table 1) such that datasets that were trained together in one model could be easily identified.

Ensemble models

For all analyses and for the generation of MEIs, we used an ensemble of models rather than individual models. Instead of training just one model for each dataset, we trained ten individual models that were initialized with different random seeds. We then selected the five best models as measured by their performance on the validation set to be part of a model ensemble. The inputs to the ensemble model were passed to each member, and the resulting predictions were averaged to obtain the final model prediction.

Generation of MEIs

We used a variant of regularized gradient ascent on our trained deep neural network models to obtain a MEI image for each neuron, given by $\mathbf{x} \in \mathbb{R}^{h\times w\times c}$. Because of our particular model inputs (see the section 'Input of behavioural parameters and image position encoding'), each MEI, like the natural images used for training, had seven channels of which we optimized only the first two: the green and UV colour channels. To obtain MEIs, we initialized a starting image with Gaussian white noise. We set the behavioural channels of the starting image to the desired behavioural values (usually <3rd and >97th percentile for quiet and active states, respectively). In addition, we set the position channels to the default position encoding. Then, in each iteration of our gradient ascent method, we presented the image to the model and computed the gradients of the first two image channels (green and UV) with respect to the model activation of a single neuron. During gradient descent optimization, we smoothed the gradient by applying Gaussian blur with a σ of 1 pixel. To constrain the contrast of the image, we calculated the Euclidean (L2) norm of the resulting MEI

$$\|\mathrm{MEI}\|_2 := \sqrt{\sum_{i=1}^{c}\sum_{j=1}^{w}\sum_{k=1}^{h}\mathrm{MEI}_{ijk}^2}$$

across all pixels MEI_{ijk} of the two colour channels c and compared the L2 norm to a fixed norm budget b, which we set to 10. The norm budget can be effectively thought of as a contrast constraint. An L2 norm of 10, calculated across all pixel intensities of the image, proved to be optimal such that the resulting MEI had minimal and maximal values similar to those found in our training natural image distribution. If the image exceeded the norm budget during optimization, we divided the entire image by factor f_{norm} with $f_{\mathrm{norm}} = \|\mathrm{MEI}\|_2/b$. Additionally, we made sure that the MEI could not contain values outside the 8-bit pixel range by clipping the MEI outside these bounds, corresponding to 0 or 255 pixel intensity. As an optimizer, we used stochastic gradient descent with a learning rate of 3. We ran each optimization for 1,000 iterations, without an option for early stopping. Our analyses showed that the resulting MEIs were spatially highly correlated across behavioural states (Extended Data Fig. 5a–c). To validate this finding, we performed a control experiment using two separate models exclusively trained on trials from active or quiet states. We again split the trials into quiet and active periods using pupil size (quiet, <50th percentile; active, >75th percentile). When inspecting the MEIs generated from these two models, we found that the MEIs were again highly correlated across colour channels, albeit less than for the model that was trained on the entire data. This can partially be explained by the limited amount of data for the model trained with trials from the active state that occurred less frequently in our data. Furthermore, we found that the spatial structure of MEIs of anatomically matched neurons across the control and pharmacology conditions was highly similar, which suggested that the two models trained separately both converged on the same tuning properties, despite differences in the prediction performance (Extended Data Fig. 2)

Spectral contrast

For estimating the chromatic preference of the recorded neurons, we used spectral contrast (SC). It is estimated as a Michelson contrast ranging from −1 to 1 for a neuron responding solely to UV and green image contrast, respectively. We decided to quantify the spectral sensitivity in relative terms for each behavioural state because visual responses to both green and UV stimuli are gain modulated in an active state. Therefore, interpretation of absolute response amplitudes to UV and green stimuli across behavioural states can be challenging. See Extended Data Fig. 6a,d for an illustration of how responses to stimuli of diverse spectral contrasts are gain modulated during an active state. We define SC as

$$SC = \frac{r_{green} - r_{UV}}{r_{green} + r_{UV}}$$

where r_{green} and r_{UV} correspond to the following criteria: (1) the norm of the green and UV MEI channels to estimate the chromatic preference of neurons in the context of naturalistic scenes; (2) the amplitude (mean of all pixels >90th percentile) of the UV and green spatial STAs to estimate the chromatic preference of neurons in the context of the sparse noise paradigm; (3) the norm of the green and UV channels of reconstructed images to quantify chromatic preference at a populational level; and (4) the norm of the green and UV channels of simulated Gabor receptive fields to obtain each simulated chromatic preference of neurons.

In silico colour-tuning curves

To generate in silico colour-tuning curves for recorded V1 neurons, we systematically varied the L2-norm of the green and UV MEI channels while keeping the overall norm across colour channels constant (with norm = 10). We used $n = 50$ spectral contrast levels, ranging from all contrast in the UV channel to all contrast in the green channel. We then presented the modified MEIs to the model and plotted the predicted responses across all $n = 50$ spectral contrast levels. Modified MEIs were either presented to the model for a quiet or active state (see also above).

Temporal dynamics of shift in colour tuning with behaviour

To investigate the timescale of the shift in colour selectivity with behaviour, we tested how fast we could observe the shift after a transition from a quiet to an active behavioural state. To achieve this, we identified state changes from quiet to active periods by detecting rapid increases in pupil size above a certain threshold (>95th percentile of differentiated pupil size trace) after a prolonged quiet state period (>5 s below the 50th percentile of pupil size). Results were consistent across varying thresholds (data not shown). We then sampled active trials with pupil sizes >75th percentile of pupil size for varying readout windows (1, 2, 3, 5 and 10 s) after that state change. Model training was performed on all quiet trials (<50th percentile of pupil size) and the selection of active trials. MEIs and STAs were then estimated as described above.

Reconstruction analysis

We visualized which image features the population of model neurons are sensitive to by using a new resource-constrained image reconstruction method based on the responses of a population of model neurons[32]. The reasoning behind the resource-constrained reconstruction is to recreate the responses of a population of neurons when presented with a target image by optimizing a new image and matching the responses of neurons given that new image as close as possible to the responses of the target image. By limiting the image contrast of the reconstructed image during the optimization, the reconstructions will only contain the image features that are most relevant to recreate the population responses, thereby visualizing the sensitivities and invariances of the population of neurons. As target images for our reconstruction, we chose natural images from our test set. For each reconstruction, we first calculated the responses $f(\mathbf{x}_0)$ of all model neurons when presented with target image $\mathbf{x}_0$. We then initialized an image ($\mathbf{x}$) with Gaussian white noise as the basis for reconstruction of the target image by minimizing the squared loss between the target responses and the responses from the reconstructed image $\ell(\mathbf{x}_0, \mathbf{x}) = \|f(\mathbf{x}) - f(\mathbf{x}_0)\|^2$ subject to a norm constraint. In this work, we set the contrast (that is, L2-norm, see section 'Generation of MEIs' for details) of the reconstructions to 40, which corresponds to about 60% of the average norm of our natural image stimuli. We chose this value to be high enough to still allow for qualitative resemblance between the reconstructed image and the target while keeping the constraint tight enough to avoid an uninformative trivial solution; that is, the identical reconstruction of the target. We improved the quality of the reconstructions by using an augmented version of our model, which reads out each neuronal response not from the actual receptive field position μ of the model neuron (see 'Readout' for details), but from all height × width positions in feature space, except the $n = 10$ pixels around each border to avoid padding artefacts. This yielded 18 × 46 = 828 copies per neuron, and with the $N = 478$ original model neurons of mouse 1 in Extended Data Fig. 9c, this resulted in overall $n = 395{,}784$ augmented neurons for our reconstruction analyses. A stochastic gradient descent with a learning rate of 1,000 produced the qualitatively best reconstructions, resulting in images with the least amount of noise. We always optimized for 5,000 steps per image, without the early stopping step of the optimization process.

Decoding analysis

We used a SVM classifier with a radial basis function kernel to estimate the decoding accuracy between the neuronal representations of two stimulus classes: either object 1 and object 2 (object discrimination) or dark object and no object (object detection). We used all neurons recorded within one scan and built four separate decoders for UV and green stimuli and small and large pupil trials, respectively. Then we trained each decoder with randomly selected training trials (usually 176 trials, but only 60–126 trials for $n = 3$ scans owing to the lower number of trials with locomotion activity), tested its accuracy with randomly selected test trials (15% of train trials) and computed the mean accuracy across $n = 10$ different training–test trial splits. Finally, we converted the decoding accuracy into discriminability, the mutual information (MI) between the true class and its estimate using

$$MI(c, \hat{c}) = \sum_i \sum_j P_{ij} \log_2 \frac{P_{ij}}{P_{i\cdot} P_{\cdot j}}$$

where P_{ij} is the probability of observing the true class i and predicted class j and $P_{i\cdot}$ and $P_{\cdot j}$ denote the respective marginal probabilities.

To quantify the significance for each animal, we compared the observed shift in decoding performance of UV versus green objects across behavioural states per animal with a distribution of shifts ($n = 500$) obtained when shuffling the labels of quiet and active trials using bootstrapping. Specifically, we sampled half of the training data and test data from quiet trials, and the other half from active trials at random. We then trained SVMs to compute the decoding accuracy based on this particular shuffling. We repeated this $n = 500$ times and obtained a P value by computing the upper quantile of the real shift given the distribution of shifts obtained when shuffling the behavioural states.

Response reliability

We calculated the signal-to-noise ratio (SNR)[68] as our measure for response reliability. It is defined as follows:

$$SNR = \frac{\frac{1}{m} \sum_{i=1}^{m} (\mu_i - \bar{\mu})^2}{\hat{\sigma}^2}$$

The SNR expresses the ratio of the variance in the expected responses against trial-by-trial variability across repeats. Here, μ_i corresponds to the expected response to the ith stimulus, with the average expected response given as

$$\bar{\mu} = \frac{1}{m} \sum_{i=1}^{m} \mu_i$$

The trial-by-trial variance $\hat{\sigma}^2$ was computed by averaging the variance across repeats over all stimuli. We assumed that $\hat{\sigma}^2$ is constant across all responses to different stimuli. This is achieved by a variance stabilizing transform of the responses r, for which we used the Anscombe transformation. We therefore obtained the transformed responses $\hat{r}$ as follows:

$$\hat{r} = 2\sqrt{r + \frac{3}{8}}$$

The SNR is a reliable estimate of data quality for neuronal responses across diverse recording modalities and brain regions[68].

Statistical analysis

We used generalized additive models (GAMs) to analyse the relationship of MEI spectral contrast, cortical position and behavioural state (see Supplementary Methods for details). GAMs extend the generalized linear model by allowing the linear predictors to depend on arbitrary smooth functions of the underlying variables[69]. In practice, we used the mgcv-package for R to implement GAMs and perform statistical testing. For all other statistical tests, we used Wilcoxon signed-rank test and two-sampled or one-sampled *t*-test.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The stimulus images and neuronal data used in this paper are stored at https://gin.g-node.org/cajal/Franke_Willeke_2022.

Code availability

Our coding framework uses general tools such as PyTorch, Numpy, scikit-image, matplotlib, seaborn, DataJoint[70], Jupyter and Docker. We also used the following custom libraries and code: neuralpredictors (https://github.com/sinzlab/neuralpredictors) for torch-based custom functions for model implementation; nnfabrik (https://github.com/sinzlab/nnfabrik) for automatic model training pipelines using DataJoint; nndichromacy for utilities, (https://github.com/sinzlab/nndichromacy); and mei (https://github.com/sinzlab/mei) for stimulus optimization.

51. Froudarakis, E. et al. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nat. Neurosci.* **17**, 851–857 (2014).
52. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
53. Garrett, M. E., Nauhaus, I., Marshel, J. H. & Callaway, E. M. Topography and areal organization of mouse visual cortex. *J. Neurosci.* **34**, 12587–12600 (2014).
54. Sofroniew, N. J., Flickinger, D., King, J. & Svoboda, K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife* **5**, e14472 (2016).
55. Pnevmatikakis, E. A. et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89**, 285–299 (2016).
56. Henriksson, J. T., Bergmanson, J. P. G. & Walsh, J. E. Ultraviolet radiation transmittance of the mouse eye and its individual media components. *Exp. Eye Res.* **90**, 382–387 (2010).
57. Schmucker, C. & Schaeffel, F. A paraxial schematic eye model for the growing C57BL/6 mouse. *Vision Res.* **44**, 1857–1867 (2004).
58. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
59. Grozdanic, S. et al. Characterization of the pupil light reflex, electroretinogram and tonometric parameters in healthy mouse eyes. *Curr. Eye Res.* **26**, 371–378 (2003).
60. Szatko, K. P. et al. Neural circuits in the mouse retina support color vision in the upper visual field. *Nat. Commun.* **11**, 3481 (2020).
61. Yoshimatsu, T., Schröder, C., Nevala, N. E., Berens, P. & Baden, T. Fovea-like photoreceptor specializations underlie single UV cone driven prey–capture behavior in zebrafish. *Neuron* **107**, 320–337.e6 (2020).
62. Perlin, K. An image synthesizer. *SIGGRAPH Comput. Graph.* **19**, 287–296 (1985).
63. Schwartz, O., Pillow, J. W., Rust, N. C. & Simoncelli, E. P. Spike-triggered neural characterization. *J. Vis.* **6**, 484–507 (2006).
64. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd International Conference on Machine Learning* (2015).
65. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). In *Proc. International Conference on Learning Representations* (2016).
66. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition* (2017).
67. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations* (2015).
68. Pospisil, D. A. & Bair, W. The unbiased estimation of the fraction of variance explained by a model. *PLoS Comput. Biol.* **17**, e1009212 (2021).
69. Wood, S. N. *Generalized Additive Models: An Introduction with R* (Chapman and Hall/CRC, 2006).
70. Yatsenko, D. et al. DataJoint: managing big scientific data using MATLAB or Python. Preprint at *bioRxiv* https://doi.org/10.1101/031658 (2015).
71. Tan, Z., Sun, W., Chen, T.-W., Kim, D. & Ji, N. Neuronal representation of ultraviolet visual stimuli in mouse primary visual cortex. *Sci. Rep.* **5**, 12597 (2015).
72. Mouland, J. W. et al. Extensive cone-dependent spectral opponency within a discrete zone of the lateral geniculate nucleus supporting mouse color vision. *Curr. Biol.* **31**, 3391–3400.e4 (2021).

Acknowledgements We thank G. Horwitz, T. Euler, M. Mathis, T. Baden, L. Höfling and Y. Qiu for feedback on the manuscript and D. Kim, D. Sitonic, D. Tran, Z. Ding, K. Lurz, M. Bashiri, C. Blessing and E. Walker for technical support and helpful discussions. We also thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Konstantin F. Willeke. This work was supported by the Carl-Zeiss-Stiftung (to F.H.S.), the DFG Cluster of Excellence 'Machine Learning—New Perspectives for Science' (to F.H.S.; EXC 2064/1, project number 390727645), an AWS Machine Learning research award (to F.H.S.), the Intelligence Advanced Research Projects Activity (IARPA) through the Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003 (to A.S.T.), grant R01 EY026927 (to A.S.T.), grant U01 UF1NS126566 (to A.T.), a NEI/NIH Core Grant for Vision Research (P30EY002520) and an NSF NeuroNex grant 1707400 (to A.S.T.). The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. 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 US Government.

Author contributions K.F.: conceptualization, methodology, validation, software, formal analysis, investigation, writing (original draft), visualization, supervision and project administration. K.F.W.: conceptualization, methodology, validation, software, formal analysis, investigation, writing (original draft), visualization and data curation. K.P.: investigation, validation and writing (reviewing and editing). M.G.: investigation and validation. N.Z.: investigation and methodology. T.M.: investigation. S.P.: methodology, software, validation and writing (reviewing and editing). E.F.: methodology, investigation and writing (reviewing and editing). J.R.: validation and writing (reviewing and editing). F.H.S.: conceptualization, writing (reviewing and editing), methodology, software, data curation, supervision and funding acquisition. A.S.T.: conceptualization, experimental and analysis design, supervision, funding acquisition and writing (reviewing and editing).

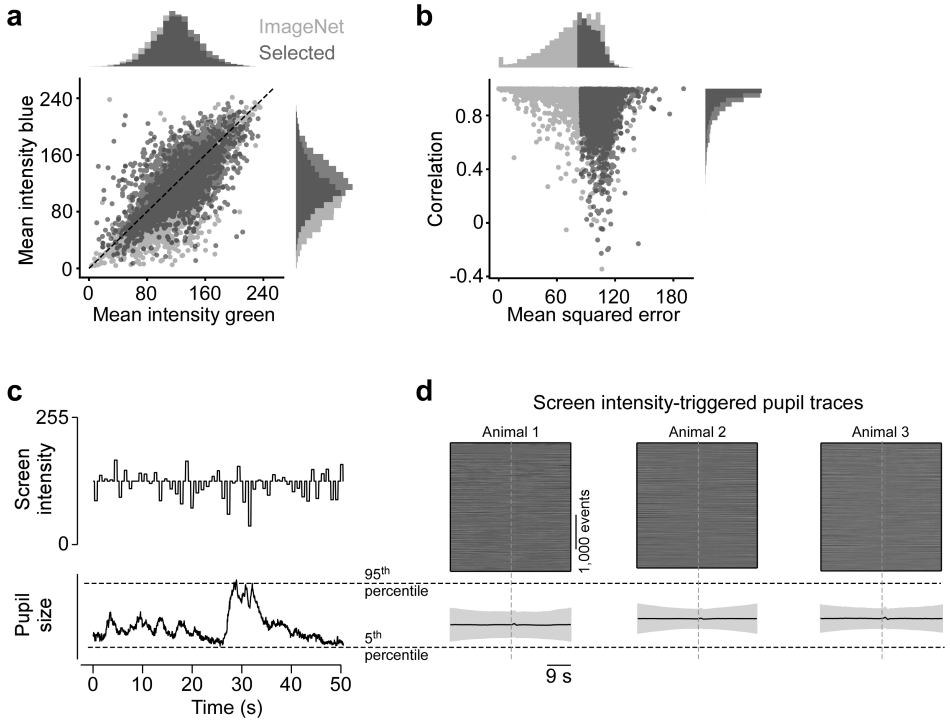
Competing interests The authors declare no competing interests.

Additional information

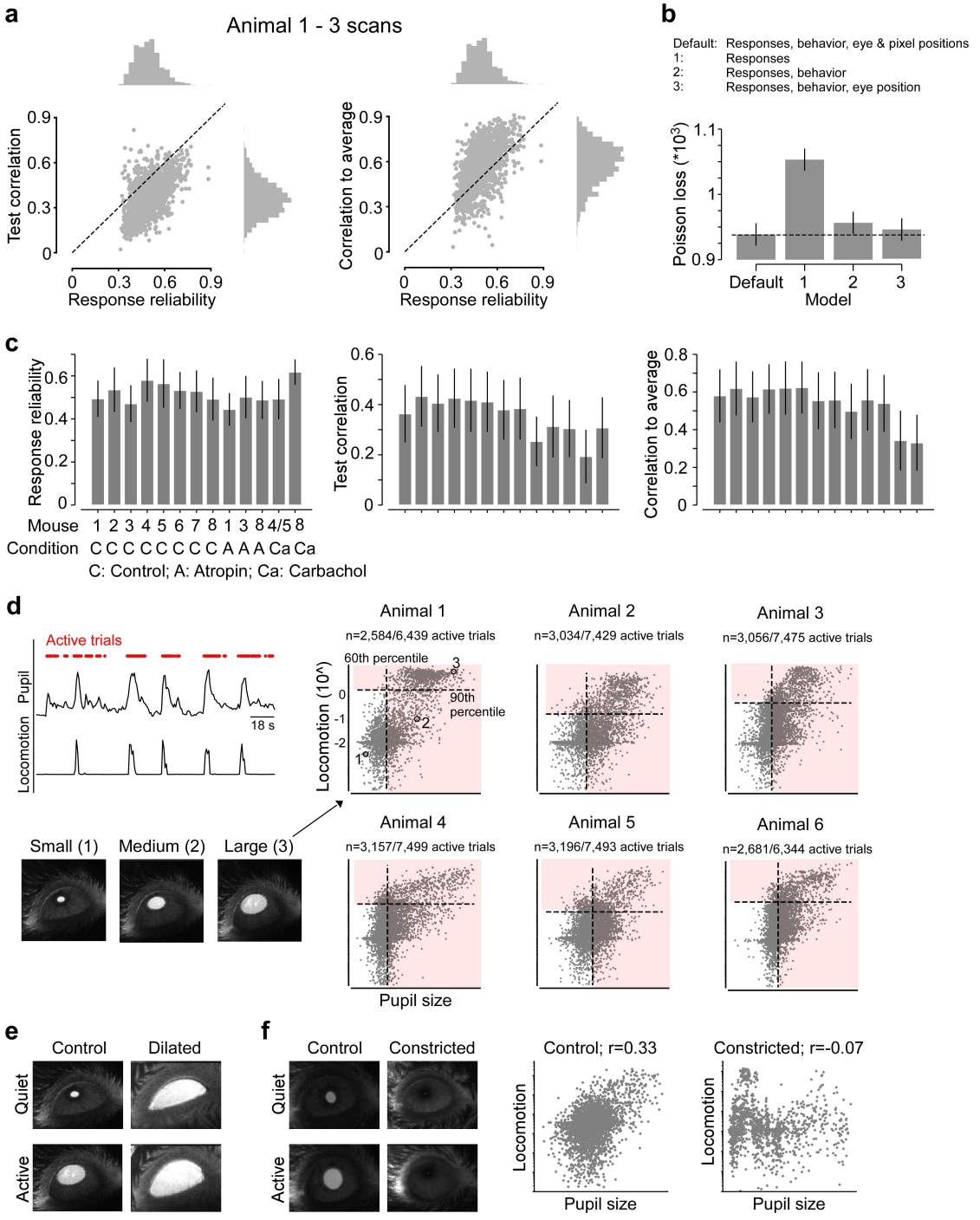
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05270-3.

Correspondence and requests for materials should be addressed to Katrin Franke.

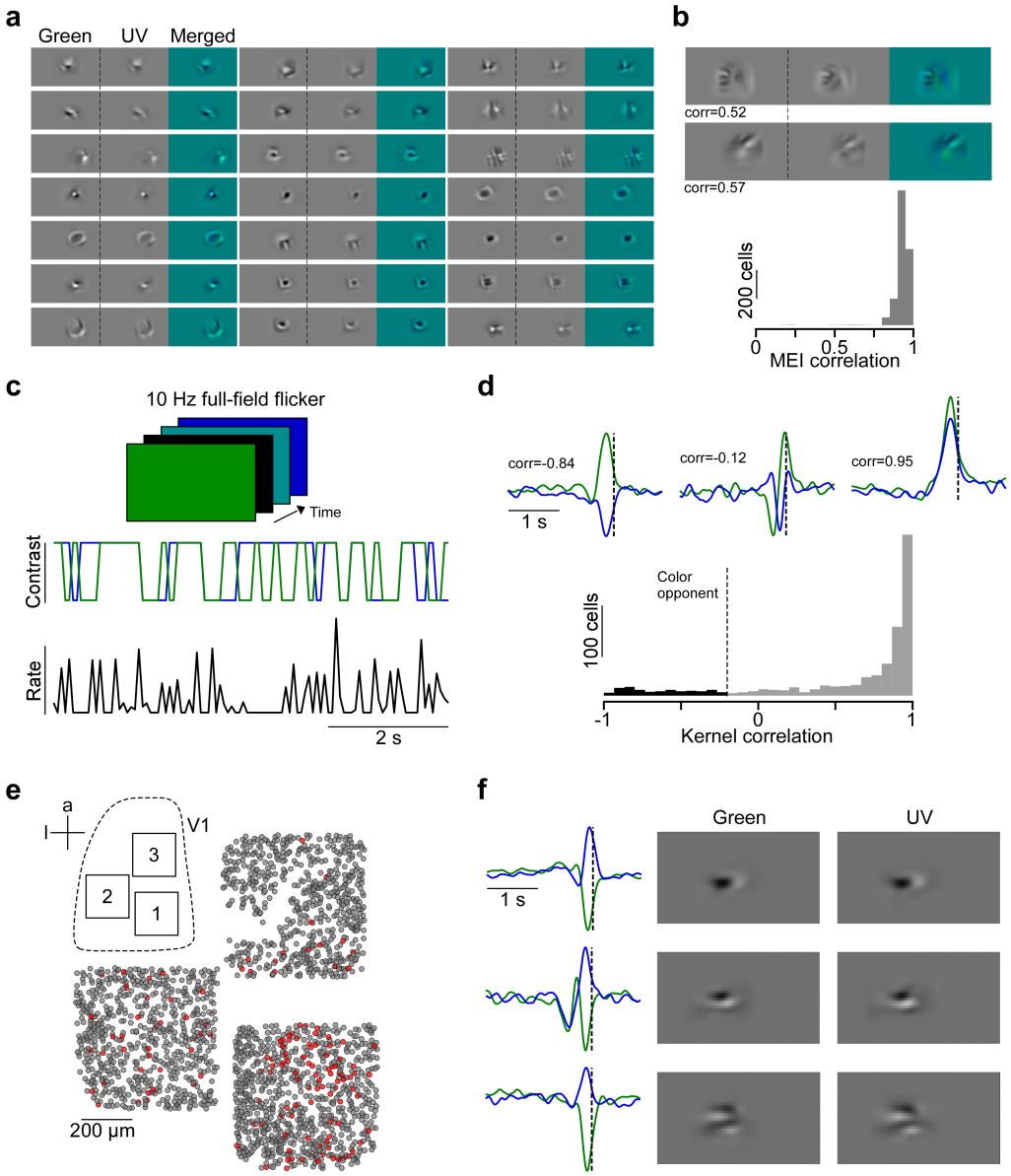
Peer review information *Nature* thanks Najib Majaj, Nathalie Rochefort and Aman Saleem for their contribution to the peer review of this work. Peer reviewer reports are available.



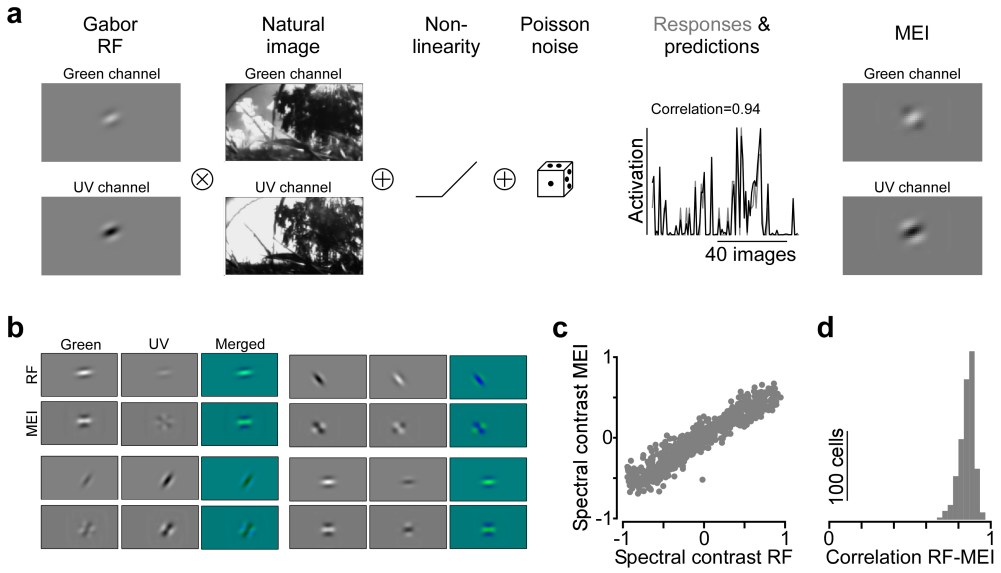
Extended Data Fig. 1 | Selection of coloured naturalistic scenes and pupil changes with monitor intensity. a, Mean intensity in 8-bit pixel space of green and blue channel of randomly sampled ImageNet images (light gray; n=6.000) and selected images (dark gray; n=6.000). Images were selected such that the distribution of mean intensities of blue and green image channels were not significantly different. Selected images can be downloaded from the online repository (see Data Availability in Methods section). **b**, Distribution of correlation and mean squared error (MSE) across green and blue image channels. To increase chromatic content, only images with MSE > 85 were selected for visual stimulation. **c**, Mean screen intensity (top) and pupil size changes (bottom) for n=50 trials. Dotted lines in the bottom indicate 5th and 95th percentile, respectively. **d**, Screen-intensity triggered pupil traces (top) for n=3 scans performed in different animals. Vertical dotted line indicates time point of screen intensity increase. Bottom shows mean change in pupil size (black; s.d. shading in gray) upon increase in screen intensity. Compared to pupil dilation induced by the behavioural state, the changes in monitor intensity over time only elicited minor changes in pupil size.



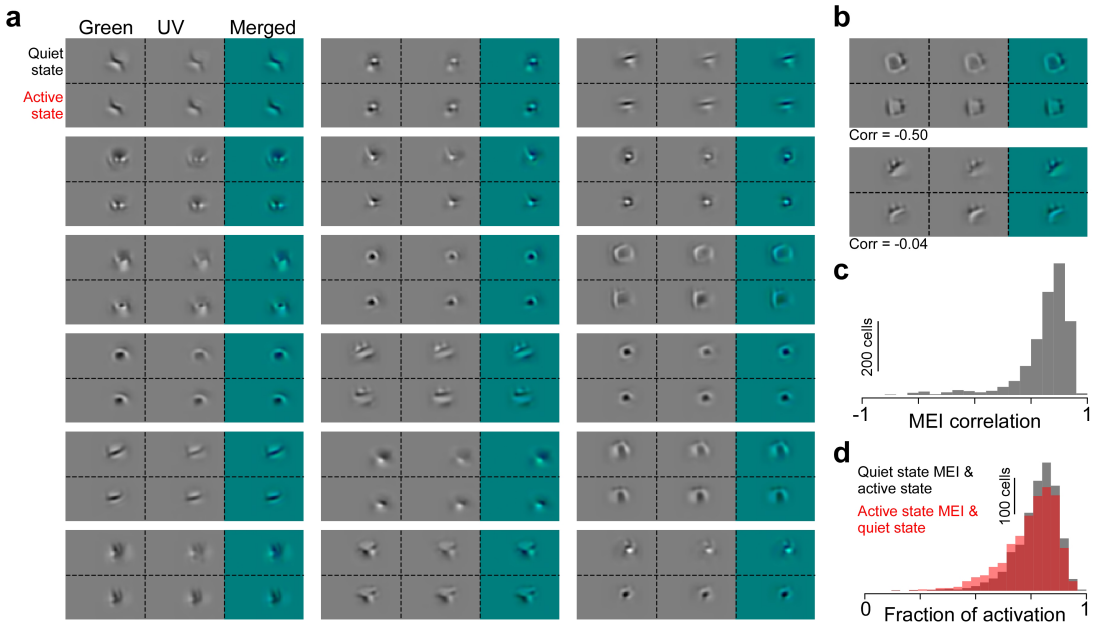
Extended Data Fig. 2 | Model performance and descriptive analysis of behaviour. a, Response reliability plotted versus test correlation (left) and correlation to average (right) for data shown in Fig. 2 (n=1.759 cells, n=3 scans, n=1 mouse). **b**, Mean Poisson loss (lower is better) for different models trained on the dataset from (a). The default model is used for all analysis, while models 1-3 are shown for comparison. Dotted line marks mean Poisson loss of default model. The default model had significantly lower Poisson loss values compared to all three alternative models (Wilcoxon signed rank test (two-sided), n=1,759: $p < 10^{-288}$ (model 1), 10^{-200} (model 2), 10^{-18} (model 3)). Error bars show 95% confidence interval. **c**, Mean response reliability, test correlation and correlation to average across neurons (error bars: s.d. across neurons; n=478 to n=1,160 neurons per recording) for n=10 models, with control and drug condition indicated below. **d**, Pupil size and locomotion speed trace of example animal, with active trials indicated by red dots. Trials were considered active if pupil size > 60th percentile and/or locomotion speed > 90th percentile. Plots on the right show mean pupil size across trials versus mean locomotion speed across trials. Dotted lines indicate 60th and 90th percentile of pupil size and locomotion speed, respectively. **e**, Example frames of eye camera for a quiet and active behavioural period for control and dilated condition. For the dilated condition, the eye was often squinted during quiet periods. **f**, Same as (e), but for control and constricted condition. Right plots show pupil size versus locomotion speed of trials used for model training for control and constricted condition.



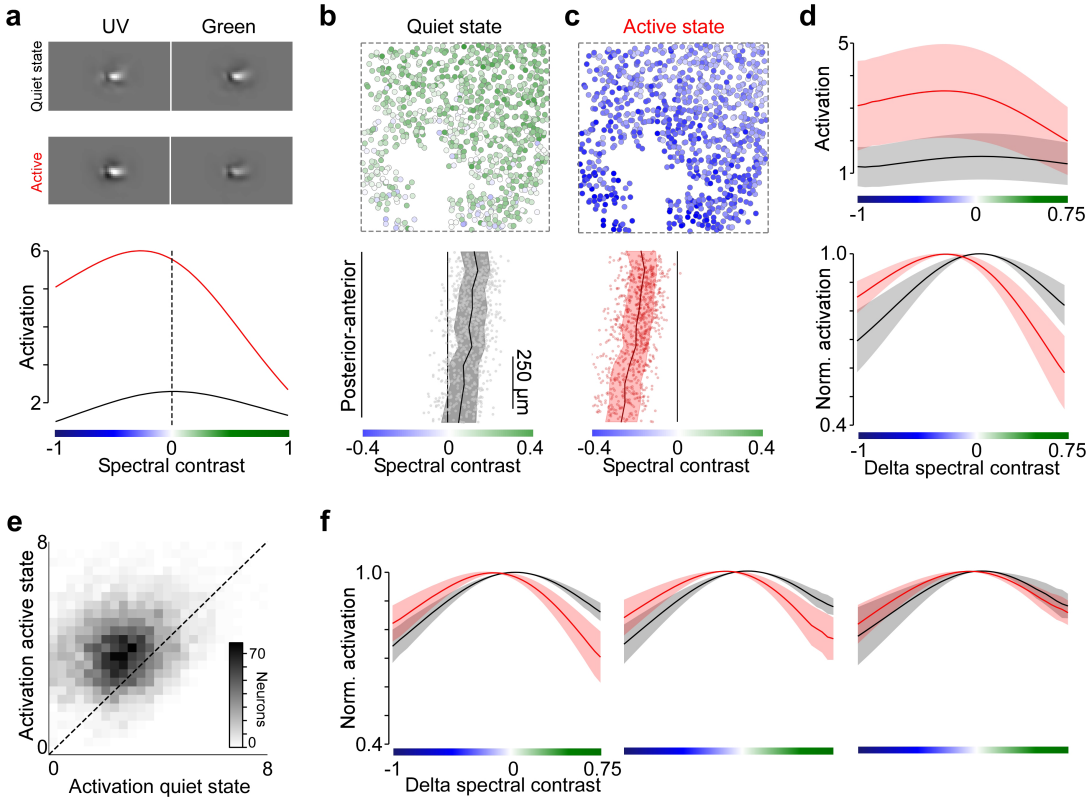
Extended Data Fig. 3 | Spatial and temporal colour opponency of mouse V1 neurons. a, MEIs of 21 exemplary neurons illustrate structural similarity across colour channels. **b**, Distribution of correlation across colour channels for dataset shown in Fig. 2. MEIs on top show example cells with relatively low correlation across colour channels. **c**, Schematic illustrating paradigm of 10 Hz full-field binary white noise stimulus and corresponding response of exemplary neuron. **d**, Temporal kernels estimated from responses to full-field noise stimulus from (c) of three exemplary neurons and distribution of kernel correlations (n=924 neurons, n=1 scan, n=1 mouse; scan 1 from (e)). Dotted line indicates correlation threshold of -0.25 – cells with a kernel correlation lower than this threshold were considered colour-opponent. A fraction of neurons (<5%) exhibited colour-opponent temporal receptive fields (see also [71]) in response to this full-field binary noise stimulus – in line with recent retinal work [60]. **e**, Neurons recorded in 3 consecutive scans at different positions within V1, colour-coded based on colour-opponency (red: opponent). **f**, Temporal kernels in response to full-field coloured noise stimulus of three exemplary neurons (left) and MEIs of the same neurons. Neurons were anatomically matched across recordings by alignment to the same 3D stack. This indicates that colour-opponency of mouse V1 neurons depends on stimulus condition, similar to neurons in mouse dLGN [72], which might be due to e.g. differences in activation of the neuron's surround or static versus dynamic stimuli.



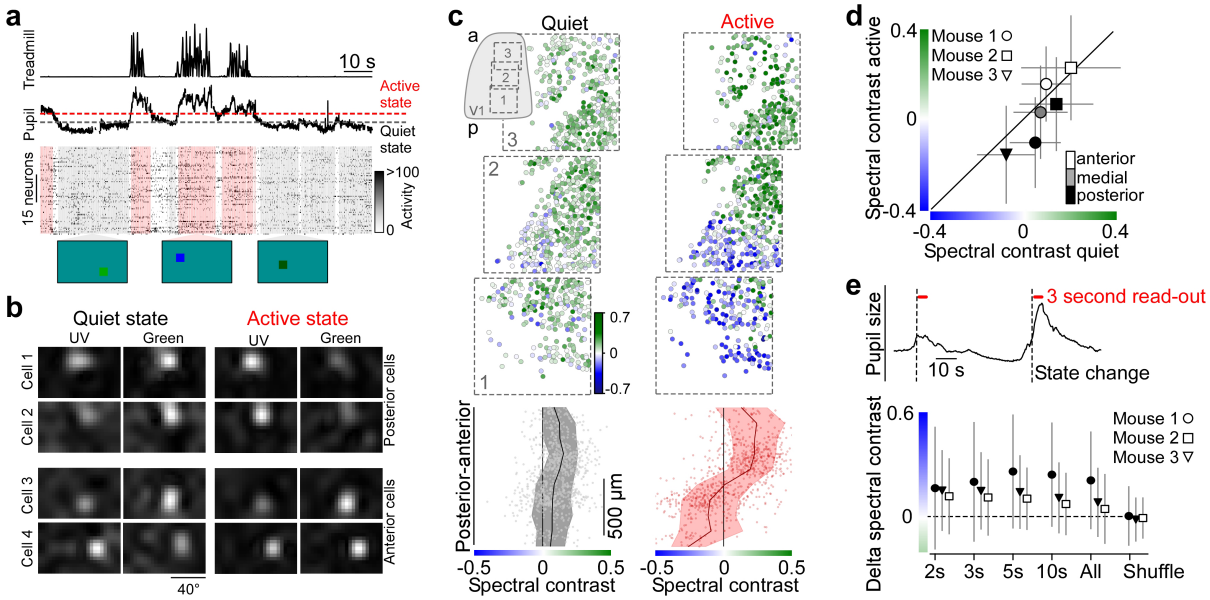
Extended Data Fig. 4 | Model recovers colour opponency and colour preference of simulated neurons. a, We simulated neurons with Gabor receptive fields (RFs) of varying size, orientation, spectral contrast and colour-opponency (correlation across colour channels). Then, responses of simulated neurons with Gabor RFs were generated by multiplication of the RFs with the natural images also used during experiments. Corresponding responses were passed through a non-linearity and a poisson process before model training. Model predictions and optimized MEIs closely matched the simulated responses and Gabor RFs, respectively. **b**, Gabor RFs and corresponding MEIs of four example neurons, some of them with colour-opponent RFs and MEIs. **c**, Spectral contrast of Gabor RFs plotted versus spectral contrast of computed MEIs. The model faithfully recovered the simulated neurons' colour preference. Only extreme colour preferences were slightly underestimated by our model, which is likely due to correlations across colour channels of natural scenes. This also suggests that it is unlikely that the low number of colour-opponent MEIs (Extended Data Fig. 3) is due to an artifact of modelling. **d**, Correlation of the MEI with the ground truth gabor RF.



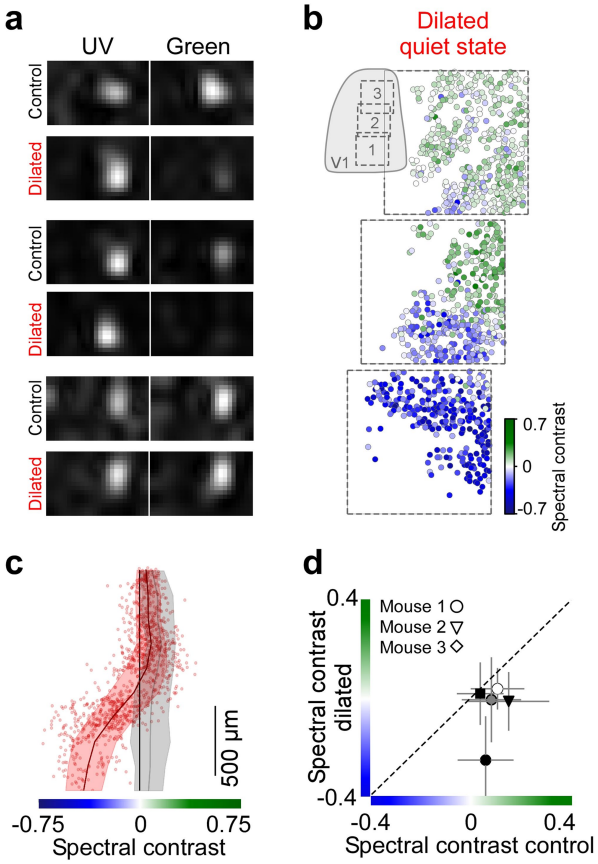
Extended Data Fig. 5 | MEI structure is consistent across quiet and active states. a, MEIs optimized for a quiet (top row of each sub-panel) and active (bottom row) behavioural state of 18 example neurons illustrate structural similarity of MEIs across states. **b**, MEIs of two exemplary neurons with low correlation across behavioural states. **c**, Distribution of MEI correlation across states (n=1,759 neurons, n=3 scans, n=1 mouse). **d**, MEI activation for incongruent behavioural state (n=1,759 neurons, n=3 scans, n=1 mouse). Gray: Model activation of MEI optimized for a quiet state presented to the model for active state relative to model activation of MEI optimized and presented for active state (activation=1). Red: Model activation of MEI optimized for active state presented to the model for quiet state relative to model activation of MEI optimized and presented for quiet state (activation=1). This suggests that MEIs optimized for different behavioural states lead to similar activations in the model and thus share similar tuning properties for the majority of neurons.



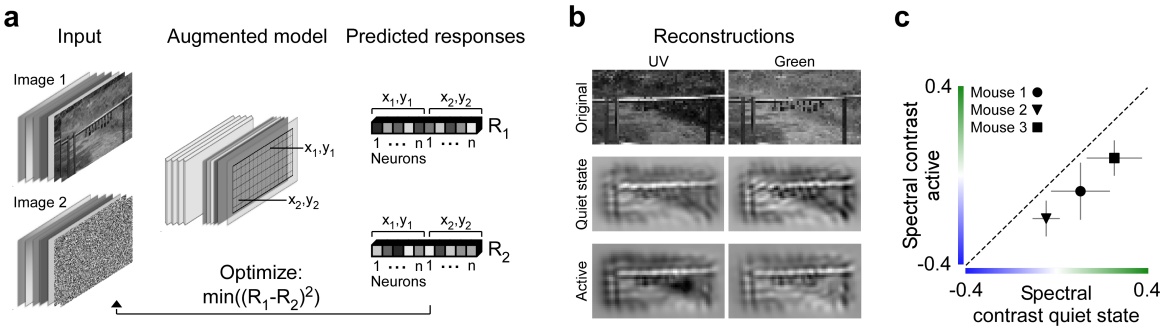
Extended Data Fig. 6 | Behavioural modulation of colour tuning of mouse V1 neurons - additional data. a, MEIs optimized for quiet and active state of exemplary neuron and corresponding colour tuning curves. **b**, Neurons recorded in posterior V1 colour coded based on spectral contrast of their quiet state MEI (top) and distribution of spectral contrast along posterior-anterior axis of V1 in an additional example animal. Black line corresponds to binned average (n=10 bins), with s.d. shading in gray. **c**, Like (b), but for active state. **d**, Mean of colour tuning curves of neurons from (b, c), aligned with respect to peak position of quiet state tuning curves. Shading: s.d. across neurons from this scan. Top shows higher model activation for active state tuning curves, in line with gain modulation of visual responses. Bottom shows peak-normalized tuning curves, illustrating (i) a shift towards lower spectral contrast values for the peak response, (ii) lower activation relative to peak for green-biased stimuli for an active state and (iii) stronger activation relative to peak for UV-biased stimuli for an active state. This suggests that during an active state, the increase in UV-sensitivity is accompanied by a decrease in green-sensitivity. **e**, Density plot of model activation in response to MEIs optimized for a quiet versus an active behavioural state, for n=6,770 neurons from n=7 mice. **f**, Mean of peak-normalized colour tuning curves of quiet (black) and active state (red), aligned with respect to peak position of quiet state tuning curves for n=3 scans from n=3 mice. Shading: s.d. across neurons.



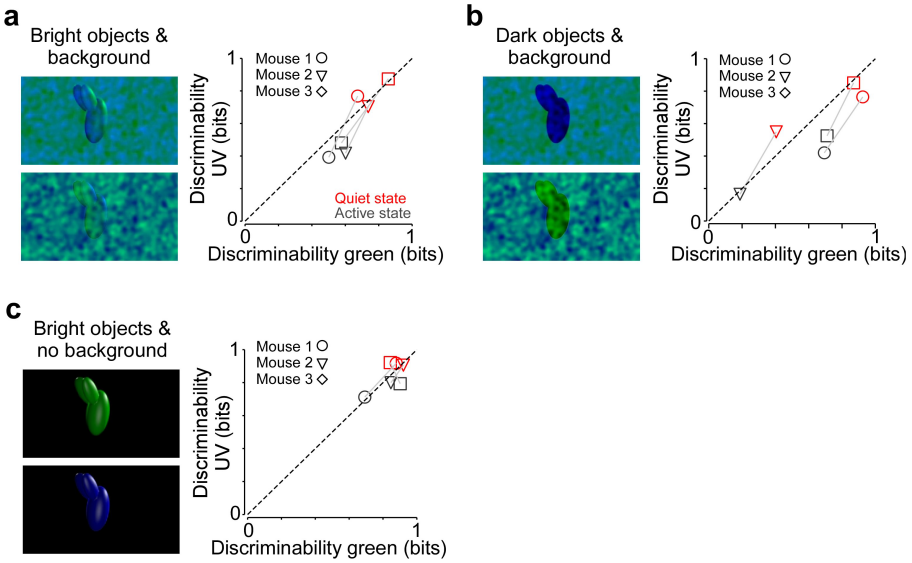
Extended Data Fig. 7 | Behavioural shift of colour preference of mouse V1 neurons in the context of a coloured sparse noise paradigm. a, Activity of n=50 exemplary V1 neurons in response to UV and green On and Off dots (10° visual angle) flashed for 0.2 seconds and simultaneously recorded locomotion speed and pupil size. Horizontal dashed lines indicate thresholds for quiet (black; < 50th percentile of pupil size) and active trials (red, > 75th percentile of pupil size). We adjusted the definition of quiet and active state compared to our *in-silico* analysis to ensure a sufficient number of trials in each state despite the shorter recording time (25 minutes for sparse noise versus 120 minutes for naturalistic images). Shading below in red and gray highlights trials above or below these thresholds. Bottom images show single stimulus frames. **b**, Spike-triggered average (STA) of 4 example neurons estimated from quiet and active trials, separated by posterior and anterior recording position. STAs estimated based on On and Off stimuli were combined to yield one STA per cell and pupil size. **c**, Neurons recorded in three consecutive experiments along the posterior-anterior axis of V1 (n=981 neurons, n=3 scans, n=1 mouse), colour coded based on spectral contrast of their STA estimated for quiet (left) and active trials (right). Bottom shows spectral contrast along the posterior-anterior axis of V1 of cells from (c, top), with binned average (black, n=10 bins) and s.d. shading (gray). Spectral contrast varied only slightly, but significantly along the anterior-posterior axis of V1 for quiet periods (n=981, p=10^{-7} for smooth term on cortical position of Generalized Additive Model (GAM); see Supplementary Methods). The small change in spectral contrast across the anterior-posterior axis of V1 is likely due to the fact that we pooled data from a wider range of pupil sizes. For an active state, optimal spectral contrast also changed with behavioural state (n=981, p=10^{-16} for behavioural state coefficient of GAM), with a significant interaction between cortical position and behavioural state modulation (p=10^{-7}; see Supplementary Methods). **d**, Mean STA spectral contrast of quiet versus active state for n=6 scans from n=3 mice. Error bars: s.d. across neurons recorded in one scan that passed quality threshold. Marker shape and filling indicate mouse ID and cortical position along the posterior-anterior axis, respectively. STA spectral contrast was significantly shifted (p=10^{-101}/3.68*10^{-51}/10^{-59}/10^{-303}, Wilcoxon signed rank test (two-sided)) towards UV for posterior and medial scan fields. The shift was not evident in anterior V1. This was likely due to the different definitions of quiet and active state in the model compared to the sparse noise recordings: For pupil size thresholds more similar to the ones used in the model (20th and 85th percentile), we observed a stronger UV-shift in STA colour preference with behaviour, also for anterior V1. **e**, Top: pupil size trace with state changes from quiet to active indicated by vertical dashed lines. Red dots show selected trials using a 3 second read-out window. Bottom: difference in STA spectral contrast of quiet versus active state for different read-out times after state change. All: all trials with quiet and active trials defined as < 20th and > 85th percentile of pupil size. Shuffle: all trials with shuffled behaviour parameters relative to neuronal responses. Dashed horizontal line indicates delta spectral contrast=0. Data shows mean and s.d. across neurons (n=996/702/964 cells, n=3 scans, n=3 animals).



Extended Data Fig. 8 | Pharmacological pupil dilation replicates shift in colour selectivity with sparse noise stimulus. a, STAs of three example neurons, estimated for quiet trials in control condition (black) and dilated condition (red). **b**, Neurons recorded in three consecutive experiments across the posterior-anterior axis of V1 (n=1,079 neurons, n=3 scans, n=1 mouse), colour coded based on STA estimated for quiet trials in the dilated condition. See Extended Data Fig. 7 for STAs estimated for the control condition of the same animal. **c**, Spectral contrast of STAs of neurons from (b) along the posterior-anterior axis of V1 (red dots), with binned average (n=10 bins; red line) and s.d. shading. Black line and gray shading corresponds to binned average and s.d. of neurons recorded at the same cortical positions in control condition (cf. Extended Data Fig. 7). Spectral contrast significantly varied across anterior-posterior axis of V1 for the dilated condition (n=1,079, $p=10^{-16}$ for smooth term on cortical position of GAM). Optimal spectral contrast changed with pupil dilation (n=1,079 (dilated) and n=943 (control), $p=10^{-16}$ for condition coefficient of GAM), with a significant interaction between cortical position and behavioural state modulation (see Supplementary Methods). **d**, Mean spectral contrast of quiet state STAs in control condition versus spectral contrast of quiet state STAs in dilated condition (n=10 scans, n=3 mice). Error bars: s.d. across neurons. Two-sample t-test (two-sided): $p=10^{-135}/10^{-20}/10^{-29}/10^{-194}/0.0006$.



Extended Data Fig. 9 | Reconstructions of coloured naturalistic scenes predict colour tuning shift for a population of neurons. a, Schematic illustrating reconstruction paradigm. As the receptive fields of neurons recorded within one of our scans only covered a fraction of the screen, we used an augmented version of our CNN model for image reconstruction where the receptive field of each model neuron was copied to each pixel position of the image except the image margins. For a given target input image (image 1), this results in a predicted response vector (R1) of length number of neurons times number of pixels. During image reconstruction, a novel image (image 2) is optimized such that its corresponding response vector (R2) matches the response vector of the target image as closely as possible. **b**, Green and UV image channels of exemplary test image (top) and reconstructions of this image for a quiet (middle) and active state (bottom). For reconstructions, neurons from scan 1 in Fig. 2 were used. **c**, Spectral contrasts of reconstructed test images (n=100) in quiet state versus active state for n=3 models trained on scans from n=3 animals. Wilcoxon signed rank test (two-sided): $p=10^{-18}/10^{-18}/10^{-18}$.



Extended Data Fig. 10 | Additional data and stimulus conditions for decoding paradigm. a, Exemplary frames of stimulus condition with lower object contrast than in Fig. 5c due to gray background in the object colour channel. Right: Scatter plot of decoding discriminability of green versus UV objects for quiet (gray) and active (red) trials for n=3 animals. Each marker represents the decoding performance of the SVM decoder trained on all neurons of the respective scan. The decoding performance for the two behavioural states are connected with gray lines, with slopes larger than one for all animals, corresponding to a larger increase in decoding performance for UV versus green objects. P-values obtained from a one-sided permutation test: < 0.012 (Mouse 1), < 0.032 (Mouse 2), < 0.112 (Mouse 3). **b**, Like (a), but for stimulus condition with objects as dark silhouettes and noise in the other colour channel. P-values obtained from a one-sided permutation test: < 0.02 (Mouse 1), < 0.1 (Mouse 2), < 0.038 (Mouse 3). **c**, Like (a), but for stimulus condition with high contrast objects and no noise in the other colour channel. P-values obtained from a one-sided permutation test (see Methods for detail): 0.44 (Mouse 1), 0.404 (Mouse 2), 0.024 (Mouse 3). The observed variability in (a) and (b) across animals might be related to different recording positions along the anterior-posterior axis of V1 and differences in the animal's behaviour, i.e. the time spent in a quiet versus active behavioural state. For the stimulus condition in (c), we might also observe a ceiling effect caused by the fact that these stimuli are relatively easy to discriminate, as indicated by high object discriminability even during quiet behavioural periods.

5.2 Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization

Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization

Konstantin F. Willeke[1-2,*], Kelli Restivo[3-4,*], Katrin Franke[3-4], Arne F. Nix[1-2], Santiago A. Cadena[3], Tori Shinn[3-4], Cate Nealley[3-4], Gabrielle Rodriguez[3-4], Saumil Patel[3-4], Alexander S. Ecker[2,5], Fabian H. Sinz[1-4,†,✉], and Andreas S. Tolias[3-4,6†,✉]

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

Deciphering the brain's structure-function relationship is key to understanding the neuronal mechanisms underlying perception and cognition. The cortical column, a vertical organization of neurons with similar functions, is a classic example of primate neocortex structure-function organization. While columns have been identified in primary sensory areas using parametric stimuli, their prevalence across higher-level cortex is debated. A key hurdle in identifying columns is the difficulty of characterizing complex nonlinear neuronal tuning, especially with high-dimensional sensory inputs. Here, we asked whether area V4, a mid-level area of the macaque visual system, is organized into columns. We combined large-scale linear probe recordings with deep learning methods to systematically characterize the tuning of >1,200 V4 neurons using *in silico* synthesis of most exciting images (MEIs), followed by *in vivo* verification. We found that the MEIs of single V4 neurons exhibited complex features like textures, shapes, or even high-level attributes such as eye-like structures. Neurons recorded on the same silicon probe, inserted orthogonal to the cortical surface, were selective to similar spatial features, as expected from a columnar organization. We quantified this finding using human psychophysics and by measuring MEI similarity in a non-linear embedding space, learned with a contrastive loss. Moreover, the selectivity of the neuronal population was clustered, suggesting that V4 neurons form distinct functional groups of shared feature selectivity, reminiscent of cell types. These functional groups closely mirrored the feature maps of units in artificial vision systems, hinting at shared encoding principles between biological and artificial vision. Our findings provide evidence that columns and functional cell types may constitute universal organizing principles of the primate neocortex, simplifying the cortex's complexity into simpler circuit motifs which perform canonical computations.

Correspondence: ***sinz@uni-goettingen.de; astolias@bcm.edu***

Introduction

From the intricate layering of neurons with diverse functions in the retina (e.g. Masland, 2001) to the topographic 1937), for decades neuroscientists have been pursing the quest to discover general organizing principles that relate the structure (anatomy) and function (physiology) of the brain. For instance, the concept of functional cortical columns, discovered by Mountcastle (1957) in the somatosensory cortex and later in the primary visual cortex (V1) by Hubel & Wiesel (1968), has been hypothesized to represent a fundamental computational circuit motif repeated throughout the primate neocortex (discussed in Horton & Adams, 2005). In this arrangement, neurons with similar function are vertically organized across cortical layers. Considering that connections within the cortex are locally dense and span cortical layers, this configuration enables neurons with similar response functions to synaptically interact, thereby facilitating computations to transform information within and across the layers (e.g. Cadwell et al., 2020; Campagnola et al., 2022; Jiang et al., 2015)

Obtaining a comprehensive understanding of the relationship between anatomy and function requires a thorough characterization of neuronal stimulus selectivity or tuning. The selectivity of neurons in monkey and cat V1 for simple visual features, such as orientation, phase, or spatial frequency (Issa et al., 2000; Victor et al., 1994), enables the characterization of these neurons' tuning properties using well-defined parametric stimuli like gratings. This has greatly facilitated the identification of general organizing principles of neuronal function in early visual areas of the cortical hierarchy (e.g. Ohki & Reid, 2014). However, neurons in higher visual areas prefer more complex visual features found in natural scenes, such as shapes, textures, objects, and faces (e.g. Bashivan et al., 2019; Kim et al., 2019a; Tang et al., 2020; Tsao et al., 2003), which are not easily described using parametric stimuli. The immense diversity and high dimensionality of the natural image space make it challenging to systematically characterize more complex visual function and link it to an organizing structure. Therefore it remains unknown whether neuronal tuning to complex spatial patterns like shapes

in the macaque visual cortex (Bashivan et al., 2019; Gallant et al., 1993; Kim et al., 2019a; Pasupathy & Connor, 2002; Tang et al., 2020), is also organized topologically across cortical layers, and whether cortical columns represent a universal principle reflecting the organization of the primate cortex. Addressing this question requires a flexible method to comprehensively characterize neuronal function without making strong assumptions about the underlying neuronal tuning model.

Advancements in deep learning promise to overcome these challenges. Specifically, recent deep learning functional models of the brain can accurately predict responses to arbitrary stimuli (Bashivan et al., 2019; Cadena et al., 2019; Walker et al., 2019), enabling essentially unlimited *in silico* experiments including ones that are virtually impossible in the real brain. This can be used for a comprehensive characterization of neuronal tuning function, such as identifying the neurons' optimal stimuli (Bashivan et al., 2019; Franke et al., 2022; Höfling et al., 2022; Walker et al., 2019), map their invariances (Ding et al., 2023b), characterize contextual modulation (Fu et al., 2023) or characterize how multiple distinct tuning properties or nonlinear contextual effects relate to each other (Ustyuzhaninov et al., 2022). The predictions derived from these *in silico* analyses can then be verified through *in vivo* closed-loop experiments, known as inception loops, which have been successfully applied to single neurons in mice (Ding et al., 2023b; Franke et al., 2022; Fu et al., 2023; Höfling et al., 2022; Walker et al., 2019) and populations of neurons in macaque visual cortex (Bashivan et al., 2019).

In this study, we adapted the inception loop paradigm for macaque electrophysiological single-unit recordings to systematically map stimulus selectivity and analyze the structure-function organization for neurons in visual area V4. We used deep neural networks to build an accurate model of >1,200 recorded V4 neurons, capable of predicting responses to arbitrary images and used it to synthesize the most exciting image (MEI) for individual neurons, which we subsequently verified *in vivo*. We found that neurons recorded on the same silicon probe orthogonal to the cortical surface appeared to have similar spatial features compared to MEIs of neurons recorded across silicon probes inserted in different locations, and verified this impression with human psychophysics and a non-linear embedding space based on image similarity. Furthermore, the MEIs formed isolated clusters in the non-linear embedding space, indicating that V4 neurons separate into distinct functional groups that are selective for specific complex visual features such as eye-like structures, oriented fur patterns, grid-like motifs, or curvatures. Interestingly, these functional groups closely resemble the feature maps of early- to mid-level units in deep neural networks trained on image classification (Olah et al., 2020), suggesting that computational principles are shared among biological and artificial visual systems. Our findings provide evidence that functional cortical columns may be a generalizable canon-sensory areas like V1 and S1.

Results

Deep neural network approach captures tuning properties of individual monkey V4 neurons To systematically study the neuronal tuning properties of monkey V4 neurons in the context of natural scenes, we combined large-scale neuronal recordings with deep neural network modeling. To this end, we presented natural images to awake, head-fixed macaque monkeys and monitored the spiking population activity of V4 neurons using acute electrophysiological recordings with 32-channel linear arrays spanning 1,920 µm in depth, covering the majority of the 2 mm cortical depth (Fig. 1a; Denfield et al., 2018). In each recording session, we displayed 9,000–12,075 gray-scale images from the ImageNet database (Deng et al., 2009) organized in a trial structure, where each trial consisted of 15 images, each presented for 120 ms, followed by a gray screen 1,200 ms inter-trial period. During image presentation, the monkey was trained to maintain fixation on a fixation spot offset from the center of the monitor (Fig. 1a). The spot's exact location was selected prior to each recording session such that the neurons' population receptive field (RF), determined by using a sparse random dot stimulus, was centered on the monitor. Post-hoc spike sorting of the neuronal activity recorded across 100 sessions from two monkeys isolated the single-unit visual activity of 1,224 individual V4 neurons (Fig. 1c), resulting in a large dataset of well-isolated single-unit activity in monkey V4.

To predict the responses of the recorded neurons and characterize the neurons' tuning properties, we used a deep convolutional neural network (CNN) model (Fig. 1b). Based on previous work in monkey (Bashivan et al., 2019; Cadena et al., 2019, 2022), we used a pre-trained goal-directed neural network as a non-linear feature space shared across all neurons and fitted only a simple linear-nonlinear neuron-specific readout (Lurz et al., 2020). Specifically, we chose a robust and high-performing ResNet-50 (Salman et al., 2020) as goal-directed neural network, trained on an image classification task. We selected one of its intermediate layers (layer 3.0) as non-linear feature space because it resulted in the best response predictions of the recorded V4 neurons. This yielded a correlation between response predictions and mean neuronal responses across repetitions of 0.43 (Fig. 1d). Together, these results show that our modeling approach accurately captures tuning properties of monkey V4 neurons in the context of naturalistic scenes.

Treating our CNN model as a functional digital twin of the population of V4 neurons, we synthesized maximally exciting images (MEIs) (Bashivan et al., 2019; Walker et al., 2019) for individual V4 neurons *in silico* (Fig. 1e). To this end, we optimized a contrast-constrained image to produce the highest activation in the model neuron using regularized gradient ascent. The resulting MEI corresponds to the optimal stimulus of a neuron according to

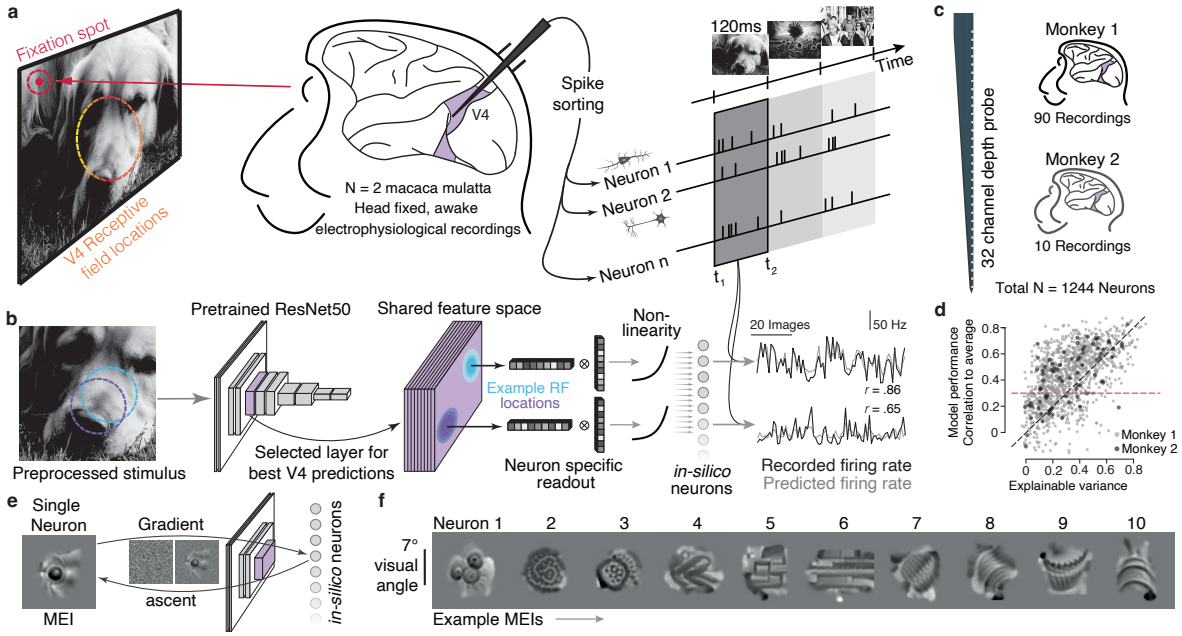


Fig. 1. Deep neural network approach captures tuning properties of individual monkey V4 neurons a, Schematic illustrating experimental setup: Awake, head-fixed macaque monkeys were presented with static natural images after fixating for 300 ms (120 ms presentation time per image, 15 images per trial, 1200 ms inter-trial period), while recording the neuronal activity in V4 using 32-channel probes. Animals were fixating on a fixation spot such that the recorded neurons' population receptive field was centered on the monitor. Post-hoc spike sorting resulted in single-unit activity of individual V4 neurons. **b,** Schematic illustrating model architecture: The pre-processed stimuli (100 x 100 pixels crop) and neuronal responses were used to train a neuron-specific read-out of a ResNet50 pre-trained on an image classification task. Specifically, the ResNet50 layer with the best V4 predictions was selected to represent a shared feature space across neurons and we computed the neuronal responses by passing the neuron-specific feature activations to a Gaussian readout and a subsequent non-linearity. Traces on the right show average responses (gray) to 75 test images of two example neurons and corresponding model predictions (black). **c,** Schematic illustrating 32-channels along the probe used for electrophysiological recordings and number of recording sessions per monkey. In total, we recorded the single-unit activity of n=1,244 neurons. **d,** Explainable variance as a measure of response reliability to natural images plotted versus model prediction performance (correlation between prediction and average neural response to repeated presentations) of all cells. Dotted red line indicates a prediction performance of 0.3 used in subsequent analyses (explainable variance mean $\pm$ s.d. $= 0.33 \pm 0.19$, correlation to average mean $\pm$ s.d. $= 0.43 \pm 0.21$) **e,** Schematic illustrating optimization of most exciting images (MEIs). For each *in silico* neuron, we optimized its MEI using gradient ascent over n=100 iterations. **f,** MEIs of ten example neurons. The whole gray box (full extent) is 14.82° degrees visual angle in width and height.

that MEIs strongly differ across neurons, indicating selectivity for distinct stimulus features like texture, curvature and edges (Fig. 1f), which resemble the features found in the MEIs of V4 multi-unit activity (Bashivan et al., 2019). Our MEIs were also consistent with tuning properties of macaque V4, such as shape, curvature, and texture selectivity, previously, identified using parametric stimuli (Kim et al., 2019b; Pasupathy & Connor, 2001; Pasupathy et al., 2020). However, in contrast to these previous studies, our data-driven approach uncovers tuning properties of single V4 neurons without making any parametric assumptions about the neurons' stimulus selectivity or the need for preselecting an ensemble of visual stimuli.

Closed-loop paradigm verifies model-derived optimal stimuli of single V4 neurons To demonstrate the model's accuracy and that the computed MEIs indeed strongly drive the recorded neurons, we developed a closed-loop paradigm for acute electrophysiological recordings of single neurons (Fig. 2a). Specifically, after fitting the readout for single-unit responses recorded in a "generation" session where natural images were shown, we selected the ification, generated their MEIs and presented them back to the animal on the same day while recording from the same neurons in a "verification" session. Single units were matched across the generation and verification session using spike waveform similarity and functional consistency of responses to the same natural images (Suppl. Fig. 1). As a control stimulus for each selected unit, we presented the seven most exciting natural image crops identified by the model by screening 5,000 natural images not used during model training. Each crop was matched to the size, position, and contrast of the MEI of a particular neuron. These control stimuli perceptually resembled the MEI (Fig. 2b), demonstrating that visual features of model-synthesized MEIs are representative for elements of natural scenes.

Overall, our model faithfully predicted responses of V4 neurons to full-field natural images and synthesized stimuli, and reliably synthesized strongly exciting images. Despite the high structural similarity of MEI and control images, the MEI consistently elicited higher neuronal responses than the control images, as well as MEIs and control images of other neurons (Fig. 2c,d,f), suggesting that

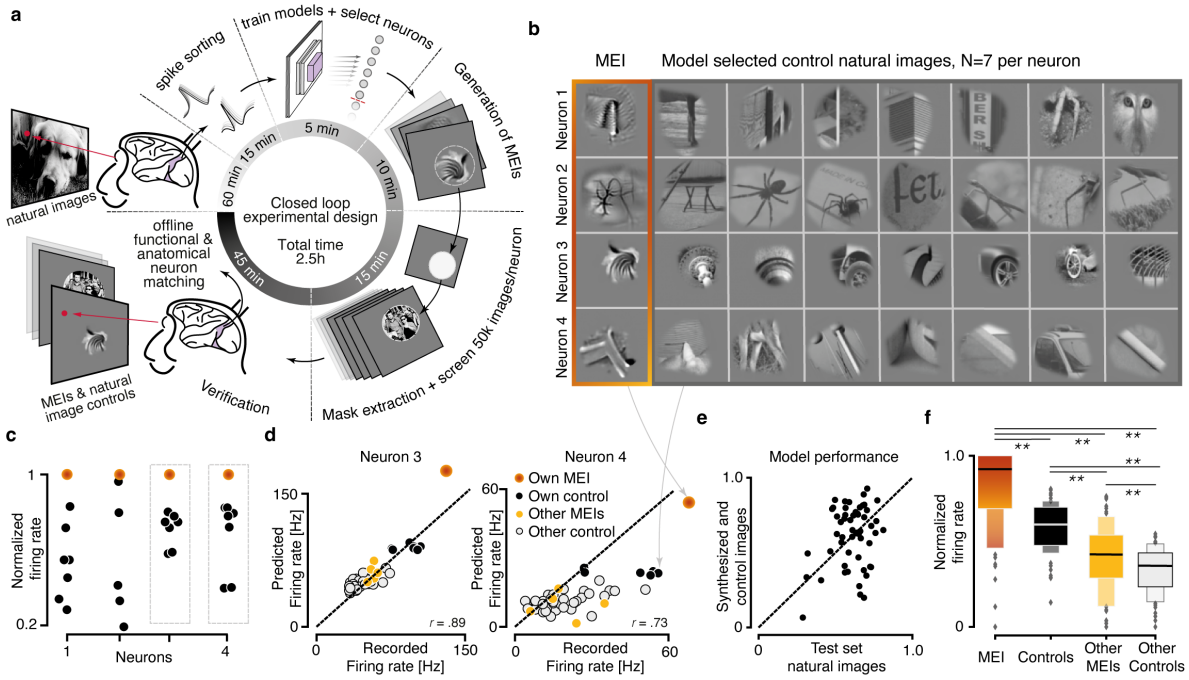


Fig. 2. Closed-loop paradigm verifies model-derived optimal stimuli of single V4 neurons a, Schematic illustrating closed-loop experimental paradigm for acute recordings in monkey V4. In brief, after recording and spike sorting of the "generation session", we train a model, select neurons for experimental confirmation, generate MEIs and identify exciting natural image control stimuli, and present both MEIs and controls back to the animal while recording from the same neurons in the "verification session". Functional and waveform matching of units across recordings is performed offline. **b,** MEI and the seven most exciting natural image crops, selected from 50k natural images, for four example neurons. Natural images were matched in size, location and contrast to the MEI. **c,** Peak-normalized recorded responses of the neurons in (b) to their MEI (orange) and control images (black; mean across n=20 repeats). **d,** Recorded versus predicted neuronal activity of two example neurons to their MEI and control stimuli, as well as to MEIs and control stimuli of other neurons. **e,** Scatter plot of model performance on the test set of natural images and the closed-loop stimuli (as shown in d, but for all neurons). Correlation to average: mean $\pm$ s.d. $= 0.61 \pm 0.11$; Synthesized and selected stimuli: mean $\pm$ s.d. 0.61 ± 0.20, $n = 55$ neurons. A paired t-test showed no significant difference $p = .89$. **f,** Distribution of peak-normalized mean responses to each neuron's MEI and control stimuli, as well as MEIs and control stimuli of other neurons for all closed-loop neurons ($n = 55$ neurons, $n = 24$ sessions, $n = 1$ monkey). P-values for a paired t-test are: MEI-Control, 3.22e-08; MEI-OtherMEIs, 2.57e-14; MEI-OtherControls, 3.06e-19; Control-OtherMEIs, 2.86e-07; Control-OtherControls, 1.62e-19; OtherMEIs-OtherControls, 2.99e-05. P-values were corrected for multiple comparisons with Bonferroni correction.

addition, the model accurately predicted the closed-loop neurons' average responses to their own MEIs, to control stimuli, and to the MEIs and control stimuli of other neurons of the same session (two example neurons in Fig. 2d). The absolute scale of the firing rate predictions did not always perfectly match the recorded firing rate, likely due to slow drifts in overall firing rates of some neurons (e. g. Fig. 2d, right). Nevertheless, across neurons, the model trained on the generation session made accurate predictions for the verification session, with no significant difference in prediction performance between full-field natural images and synthesized MEIs ($\rho = 0.61$, Fig. 2e). Moreover, the amplitude of neuronal responses to control stimuli and MEIs of other neurons only slightly differed, suggesting that there is little difference between MEIs and contrast- and size-matched natural images (Fig. 2f).

Columnar organization of optimal stimuli in macaque V4

Studying how visual selectivity is organized in a particular brain area has revealed key principles of vision, including the pinwheel of orientation columns in primary visual cortex (Bonhoeffer & Grinvald, 1991). In monkey V4, shape and texture (Kim et al., 2019b; Pasupathy & Connor, 2001; Pasupathy et al., 2020; Srinath et al., 2020), but it remains unclear whether V4 tuning properties are organized in a columnar manner (Ghose & Ts'o, 1997; Hatanaka et al., 2022; Tang et al., 2020). Therefore, we next asked whether our data-driven approach reveals an organizing principle of stimulus selectivity in V4.

We noticed that MEIs from individual sessions exhibited strong mutual perceptual similarity compared to MEIs from other sessions (Fig. 3a,b). While the range of preferred stimuli we found spanned a large variety from oriented and comb-like patterns, to grid-like motifs and patterns that resembled eye-like structures, the perceived variability within many sessions was much smaller than across sessions. For example, most neurons in one example session preferred oriented and comb-like patterns (Fig. 3a), while neurons from other example sessions preferred curved edges (session 2 in Fig. 3b) and grid-like patterns (session 3 in Fig. 3b).

To quantify the perceptual similarity of MEIs within a session, we performed a simple psychophysics experiment, where human observers were presented with two sets of

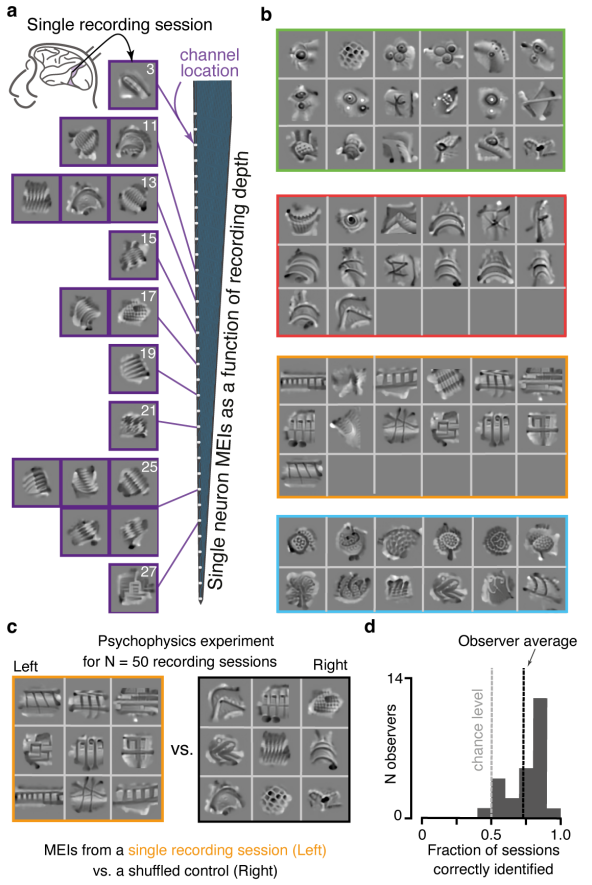


Fig. 3. Columnar organization of optimal stimuli of V4 neurons a, MEIs of 17 neurons recorded in a single experimental session, arranged according to each neuron's channel location along the recording probe. Numbers indicate channel, with higher channel numbers meaning greater recording depth. **b,** MEIs of varying numbers of neurons for four different sessions (indicated by different colors). **c,** Schematic illustrating paradigm of simple psychophysics experiment. In one trial, subjects were presented with MEIs of 9 neurons recorded within one session (left) or randomly sampled from all neurons except the target session (right), and reported the location (left or right) of the set of MEIs that looked more consistent (i.e. shared the same image features). The experiment included n=50 trials/sessions. **d,** Distribution of fraction of sessions correctly identified across n=25 observers, with change level and observer average indicated by dotted lines. Mean across subjects, $= 0.73$; subject-variability in s.d., $= 0.13$; session-variability in s.d., $= 0.21$.

session and (2) a set of nine neurons randomly sampled across sessions. The two sets were presented side-by-side (as shown in Fig. 3c, but without the colored frames), with each set being shown on the left or right at random. The observers then had to report in a two-alternative forced-choice paradigm which set of MEIs looked perceptually more similar. On average, the observers classified MEIs of the same session as being more consistent than a random set of MEIs from different sessions for 73% of the sessions (Fig. 3d), suggesting that within-session MEIs indeed share similar image features. Since neurons from a single session are arranged roughly vertically across cortical layers, this result suggests that the preferred stimuli of V4 neurons may be organized in a columnar manner.

To further quantify whether tuning properties of V4 neurons are reflected in the MEI similarity in a similar fashion, we performed nonlinear dimensionality reduction to embed the MEIs in a two-dimensional space based on image feature similarity (but not based on session information). In contrast to V1 neurons, whose tuning properties can be compared along clearly defined axes such as orientation or spatial frequency, the complex MEI structure of V4 neurons makes it challenging to quantify the tuning similarity between neurons. To resolve this problem, we used an unsupervised deep learning technique that learns a two-dimensional image-embedding based on mutual similarity of images (Böhm et al., 2023). The model is trained by forcing the two-dimensional representations (embeddings) of different variations of the same image to be close to each other, while pushing the embeddings of different images apart (Fig. 4a). By choosing data augmentations used to create different image variations, we inform the unsupervised learning model what we consider to be similar images. We used random rotations, shifts, and scaling, meaning that MEIs that differ only by those transformations should end up in the same neighborhood in the embedding space (Fig. 4a). We chose these augmentations because they generally preserve the identity of many mid- and high-level image features: for instance, a corner or an eye remain a corner or an eye even after rotating, shifting or scaling them. Moreover, we observed that multiple MEIs of the same neuron, generated by starting from different initial noise images during optimization, often exhibited variations in some or all of these dimensions (Suppl. Fig. 2). Since the training of the model is exclusively based on image identity, it does not provide any information about which recording session a particular MEI originated from. Thus, any clustering of MEIs according to recording sessions in the learned embedding space is purely based on MEI similarity.

The resulting embedding space placed neurons with similar MEIs close to each other (neurons 1 and 3 in Fig. 4b) and neurons with different MEI features far away (neurons 1 and 2). Similarly, multiple MEIs of the same neuron, generated by starting from different initial noise images during optimization, were placed nearby in the embedding space as well (groups of the same color in Fig. 4b and Fig. 4c,d). These observations suggest that the model indeed learned to embed MEIs based on image similarity. We next quantified whether neurons recorded within one session share tuning properties as suggested by the observed MEI similarity within sessions (cf. Fig. 3). To this end, we computed the mean pairwise distance in the embedding space across neurons from one session and compared it to a null distribution of distances obtained by computing the mean pairwise distance across randomly picked neurons from different sessions (Fig. 4c,e). While the within-session distance varied across sessions (Fig. 4f), on a population level, it was significantly smaller than the across-session distance (Fig. 4g). The percentage of sessions that showed a significantly smaller distance in MEI similarity than the null distribution increased

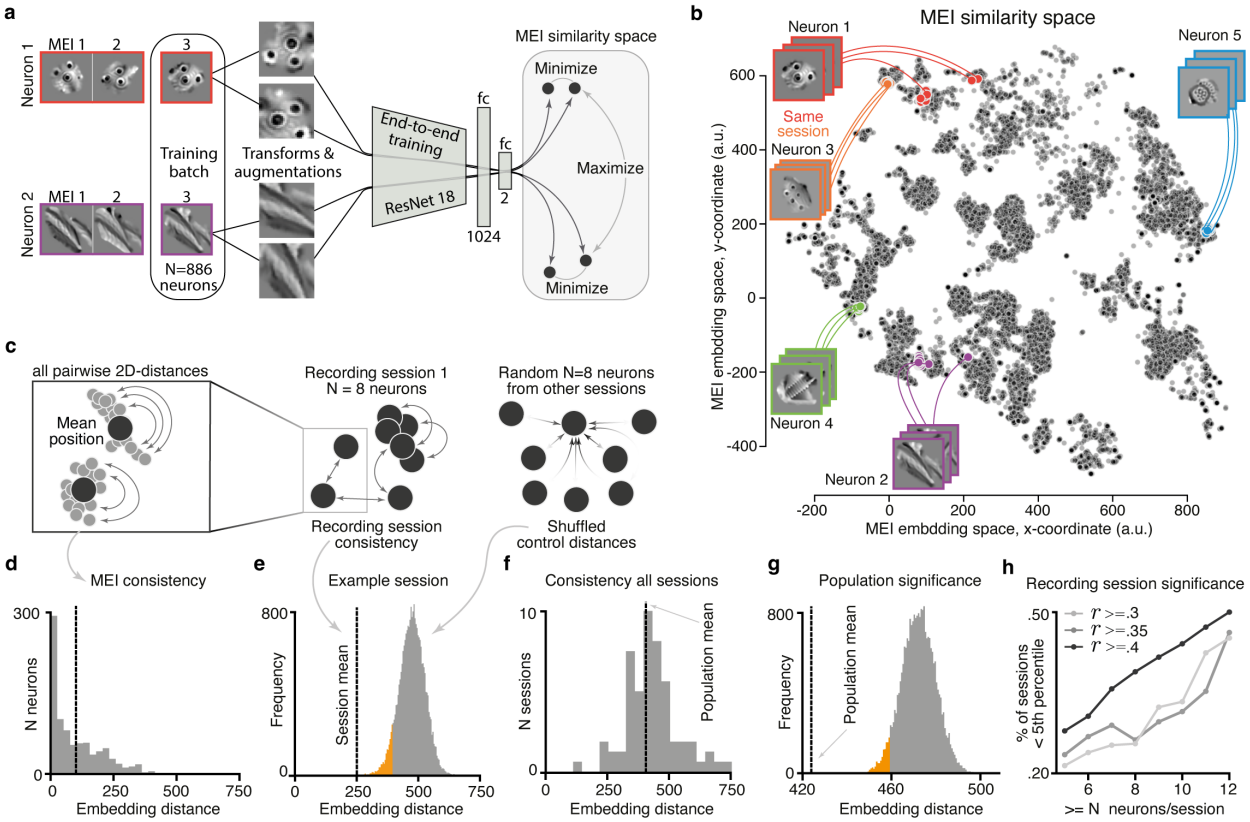


Fig. 4. Contrastive clustering of MEIs confirms columnar-like organization of V4 visual tuning selectivity a, Schematic illustrating contrastive learning approach to quantify MEI similarity. Per neuron, we optimize n=50 MEIs initialized with different random seeds, then select highly activating MEIs, and use one MEI per neuron (n=889) as a training batch. Each MEI is transformed and augmented twice and the model's objective then is to minimize the distance in a 2D MEI similarity space between different transforms of the same MEI, while maximizing the distance to MEI transforms of other neurons. **b,** Position of all highly activating MEIs (n=19688) of n=889 neurons in a 2D MEI similarity space, with MEIs of five example neurons indicated in different colors. Dots of the same color indicate MEIs optimized from different random seeds of the same neuron. **c,** Schematic illustrating analyses performed on the 2D MEI similarity space. We computed the pairwise 2D distances across all MEIs of one neuron to estimate MEI consistency (left), and all pairwise distances across MEIs of the same recording session to estimate recording session consistency (right). For the latter, we used the distances across a random selection of neurons from other sessions as control. **d,** Distribution of distances across MEIs of the same neuron. Vertical dotted line indicates mean of the distribution. **e,** Mean distance across neurons from one example session (vertical line), with a null distribution generated by bootstrapping distances across the same number of neurons randomly sampled from all other sessions. Orange shading indicates values <5% percentile. Note that the null distribution depends on how many neurons were recorded in each session, as it estimates the standard error of the mean for each session. **f,** Histogram of session means like in e), but for all sessions. Grand mean across all sessions is indicated by the vertical line. Mean = 423.97 $\pm$ 105.20 s.d. **g,** Mean within-session distance across all sessions from f) along with the mean null distribution across sessions in gray. The population mean significantly deviates from the null distribution ($p < 4 \times 10^{-5}$; 25,000 bootstrap samples). Orange shading indicates values <5% percentile. **h,** Percentage of sessions with the within-session distance <5% percentile of the null distribution for different numbers of neurons per session (x-axis) and different model predictions thresholds (shades of gray). The percentiles obtained from the embedding space (including all neurons above a prediction threshold of 0.3) were significantly correlated with the observer agreement (percent correct) of the psychophysics experiment ($\rho = -0.33$, p=.019, n=50 sessions).

with higher prediction performance of the model (Fig. 4h). For more than 12 neurons per session and with the highest performance threshold (correlation to average >0.4), half of the sessions displayed a significantly smaller within- than across-session distance, indicative for high similarity of MEIs. Importantly, the within-session distances estimated based on the embedding space significantly correlated with the observer agreement (percent of observers who correctly classified a specific session) from the psychophysics results ($\rho = -.32$, p=.025), suggesting that MEI distance in the embedding space is informative about MEI perceptual similarity. We additionally confirmed that MEIs of neurons recorded within one session are more similar to each other than to MEIs of neurons recorded namely the representational similarity of MEIs in neuronal response space, which closely mimics perceptual similarity (Kriegeskorte, 2008) (Suppl. Fig. 3). Taken together, the above analyses strongly suggest that neuronal tuning properties of monkey V4 neurons are organized in a columnar structure, with neurons sharing the same tuning aligned within a vertical column.

V4 neurons cluster into distinct functional groups that resemble feature maps of artificial vision systems At the level of retinal ganglion cells, neurons cluster into specific functional groups or output channels (Baden et al., 2016; Goetz et al., 2022). Whether a similar functional clustering persists for cortical neurons is still an open question. For

tween simple and complex cells represents two ends of a continuous spectrum or two discrete categories (Mechler & Ringach, 2002). Previous results in mouse primary visual cortex suggest that neurons cluster according to function, but not in an entirely discrete manner (Ustyuzhaninov et al., 2019). Motivated by the structure of the MEI embedding space that exhibited isolated "islands" of MEIs (cf. Fig. 4b), we asked whether tuning properties of V4 neurons fell into similar functional groups, characterized by their preferred stimulus. To address that question, we clustered the two-dimensional embedding vectors of the MEIs using hierarchical clustering (DBSCAN; McInnes et al., 2017), resulting in 17 different functional groups (Fig. 5). Different MEIs within the same group showed high perceptual similarity and exhibited similar stimulus preferences, such as for eye-like structures in group 11. In contrast, neurons assigned to different groups substantially differed with respect to their preferred visual feature, especially for groups located on opposite sides of the embedding space. For example, while MEIs in group 11 were characterized by eye-like structure, MEIs of group 3 and 8 displayed grid-like and comb-like patterns, respectively.

To ensure that the group structure in the embedding space is not an artifact of the contrastive neighborhood embedding method, we additionally used an independent method to compare within-group to across-group similarities. To that end, we computed the representational similarity of the MEIs using the predicted neuronal responses from our model. Specifically, we centered the receptive fields of all neurons in the model and compared the similarity of two MEIs via the cosine of the predicted population response vectors. Importantly, this similarity metric is unrelated to the embedding space used for clustering and, therefore, provides an independent verification of the identified functional groups. We found that MEIs of neurons assigned to the same group were significantly closer to each other in the neuronal response space than MEIs of neurons assigned to different groups (Suppl. Fig. 3): The within-group similarity in neuronal response space was significantly higher than the across-group similarity for all of the 17 functional groups.

Interestingly, the MEIs of the functional groups of monkey V4 neurons we identified closely resemble feature visualizations of single units found in modern deep neural networks trained on image recognition. For example, a similar preference for specific complex features can be found in the early layers, specifically layer `mixed3a`, of the InceptionV1 deep network (Olah et al., 2020; Szegedy et al., 2015). This alignment between V4 and deep network features is in line with previous results that found boundary selective units in the AlexNet deep network (Krizhevsky et al., 2012; Pospisil et al., 2018), similar to boundary neurons in monkey V4 (Pasupathy & Connor, 2002). Olah et al. (2020) manually grouped the feature visualizations into different categories like "Oriented fur," "Eyes/circles/loops" and "Divots/boundaries" (Fig. 5b). tional groups of V4 neurons we identified using hierarchical clustering, suggesting that there are general encoding principles that are shared among the primate visual system and artificial vision system. The resemblance between V4 neuronal and deep artificial neural network feature selectivity can be used to generate specific hypotheses about visual tuning properties of primate V4 neurons beyond spatial patterns. For example, as the artificial vision systems were trained on color images, the resulting feature visualizations also characterize model units' color tuning and can be used to derive predictions about color boundary encoding in monkey V4 functional groups (Fig. 5c), which could subsequently be verified in *in vivo* experiments.

Discussion

Our work provides evidence for a columnar organization of tuning to spatial patterns in visual area V4, based on detailed single neuron tuning properties. This finding was enabled by employing an iterative image synthesis method based on a deep neural network model of neuronal activity. This digital twin of V4 facilitated a characterization of spatial pattern tuning that was not biased by any prior parametric assumption about the type of selectivity, and identified preferred stimuli (MEIs) that evoked stronger responses than competitive natural stimuli selected from a large set of previously unseen images. By learning a nonlinear image similarity embedding space and using human psychophysics, we demonstrated that the MEIs from a single recording penetration perpendicular to the cortical surface were more similar to each other than from a random selection of neurons, and that the MEIs of neurons in V4 clustered into separate functional groups. Our findings suggest that, despite the complex stimulus preferences of V4 neurons, their preferred selectivity is organized in a columnar manner.

Are our MEIs consistent with previous findings? The idea to use neuronal encoding models to synthesize optimal stimuli for the brain is not new (Lehky et al., 1992) and has already successfully been used in mouse primary visual cortex (Franke et al., 2022; Walker et al., 2019), mouse retina (Höfling et al., 2022), and also macaque V4 (Bashivan et al., 2019). Many of our MEIs display features, such as different types of textures, that qualitatively resembled those found by Bashivan et al. (2019). However, some of our MEIs exhibited shape-like features, including curved strokes, corners, and even higher-level attributes such as individual eye-like stimuli (cluster 7 in Fig. 5a), which have not been reported before.
One distinction between our study and that of Bashivan et al. (2019) is that we used silicon probes, enabling us to separate spikes from individual neurons, while they employed chronically implanted Utah arrays and optimized stimuli for single 'sites' (multi-units) that likely comprise a mix of multiple neurons. Consequently, it remains unclear

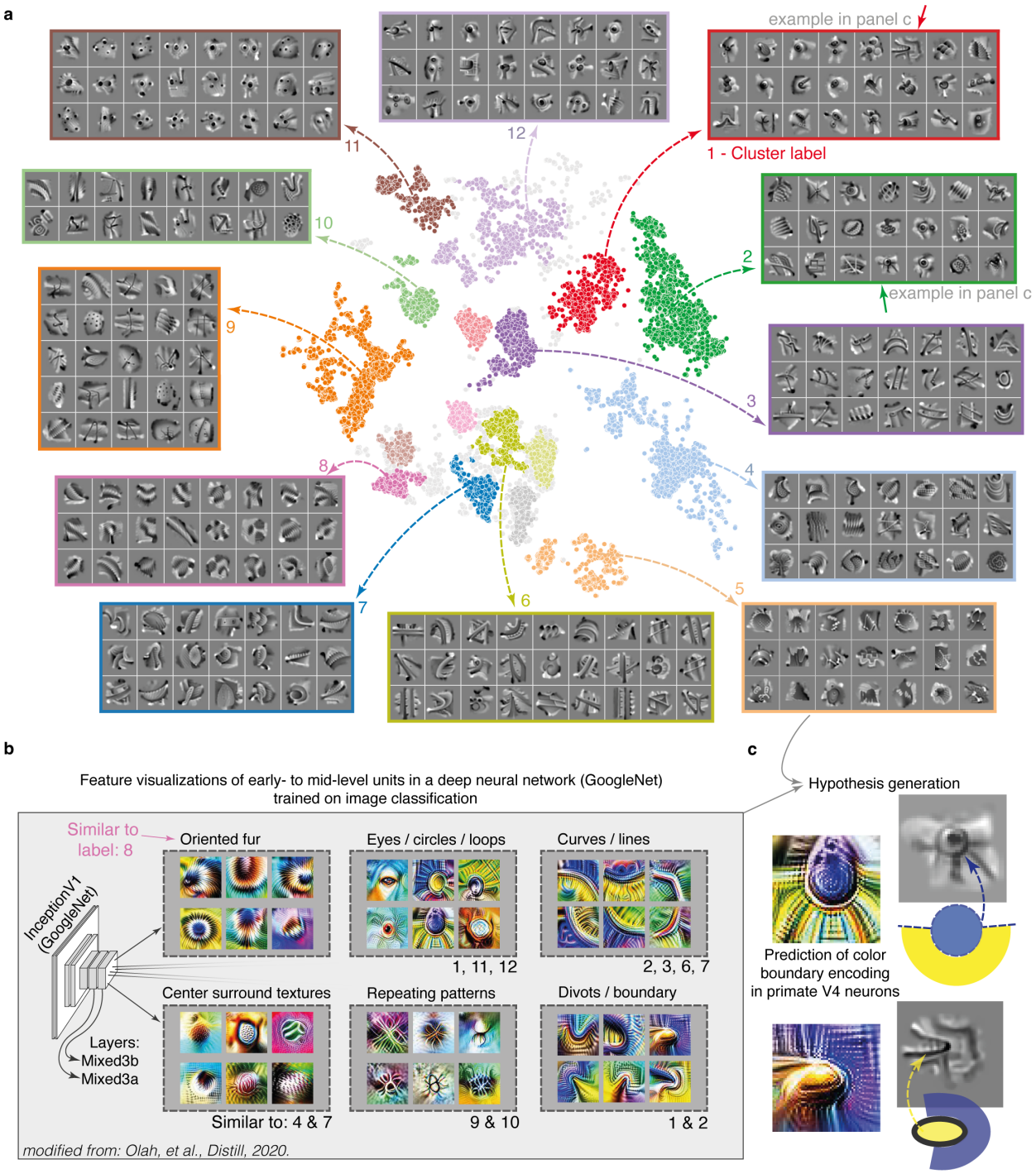


Fig. 5. V4 neurons cluster into distinct response modes that resemble feature maps of artificial vision systems a, Position of all highly activating MEIs (n=19,688) of n=889 neurons in the 2D MEI similarity space, color coded based on cluster assignment obtained from the hierarchical clustering algorithm HDBSCAN. For n=12 clusters, we show a random selection of MEIs of different neurons assigned to this cluster. For examples of the other clusters, see Suppl. Fig. 4 and for independent verification of the clusters, see Suppl. Fig. 3. Light gray dots indicate MEIs that could not be assigned to any of the clusters with high probability. **b,** Feature visualizations of eary- to mid-level units in the deep neural network InceptionV1 trained in an image classification task (Olah et al., 2020). Units are grouped into distinct categories based on (Olah et al., 2020), with clusters from (a) resembling these categories indicated below. **c,** Example units of the neural network trained on image classification compared with example MEIs exhibiting similar spatial patterns. The resemblance between the two can be used to generate hypotheses, such as to predict color boundary encoding in primate V4 neurons, that can be subsequently tested experimentally.

from the mixing of spikes from multiple neurons or if single neurons already prefer such intricate spatial patterns as displayed by the MEIs. Additionally, MEIs for multi-unit activity could average out specific features like shapes. Our results provide experimentally validated evidence that the complexity of MEIs is already present at the level of single neurons in macaque V4.

Our MEIs are also consistent with previously described tuning properties of V4 neurons. V4 is part of the ventral pathway which plays a major role in object and shape recognition (Felleman & Van Essen, 1991; Mishkin et al., 1983). Previously, V4 cells have been reported to be selective to complex shapes (Kobatake & Tanaka, 1994), and be tuned to convex and concave shapes of object boundaries at specific locations in the visual field (Gallant et al., 1993; Pasupathy & Connor, 2001). In addition to shape, V4 neurons are also known to be selective to texture (Kobatake & Tanaka, 1994). Kim et al. (2019a) found that tuning of single V4 neurons can be placed along a continuum from strong tuning for boundary curvature of shapes to strong tuning for perceptual dimensions of texture. From visual inspection, our MEIs reproduce these tuning properties. On the one hand, we find MEIs that clearly exhibit curvature elements (e.g. cluster 7 in Fig. 5a) or 'eye'-like elements (cluster 1). On the other hand, many MEIs have a texture component such as 'fur' (cluster 8), dots (cluster 11), or grid-like elements (cluster 6). Interestingly, generating multiple MEIs from different starting points – known as Diverse Exciting Images (DEIs; Cadena et al., 2018; Ding et al., 2023b) – resulted in multiple MEIs that had similar shape and texture features (Suppl. Fig. 2), indicating that single cells in V4 are neither texture- nor shape-invariant. Together, our MEIs are consistent with previous results but paint a more detailed picture of the complex and diverse tuning properties of single cells in V4.

Topological organization of MEIs in V4 One advantage of our recordings is that we can record simultaneously across layers using silicon probes. This enabled us to characterize the vertical organization of tuning selectivity to complex spatial patterns in area V4, and investigate evidence for a columnar organization. The presence of columns in V4 has been a matter of debate and controversy, predominantly studied in the domain of color (Kotake et al., 2009) or orientation (Ghose & Ts'o, 1997) because stimuli for these domains are more accessible to low dimensional parametrization. For instance, in the color domain, one study has reported weak columnar organization (Kotake et al., 2009), while another found no evidence (Tanaka et al., 1986). Beyond that, other studies using natural images or parametric stimuli for curvature found that neurons within a layer with similar tuning are locally clustered (Hatanaka et al., 2022; Tang et al., 2020). However, since these studies only focused on neurons from superficial layers, it remains unknown whether there is a columnar organization for these features that spans all cortical layers.

lectivity for complex spatial features that extend beyond those characterized by parametric stimuli defined by orientation, or parameterized curvature (Tang et al., 2020). Consequently, in the absence of a detailed identification of the optimal stimuli, assessing the hypothesis of functional columns proves to be challenging. For example, if neurons are tuned to similar grid-like textures but with different orientations, using grating stimuli (Ghose & Ts'o, 1997) or predetermined parameterized curvature stimuli (Tang et al., 2020) will obscure the true underlying organization. Our deep learning based image synthesis method avoids these challenges by identifying most exciting stimuli in the high-dimensional pixel space.

To compare MEIs regardless of their spatial complexity, we used human psychophysics and deep learning techniques to assess the similarity between them, specifically employing contrastive learning to create a non-linear embedding based on similarity among MEIs (Böhm et al., 2023). We discovered that neurons across layers recorded in orthogonal penetrations to the cortical surface had MEIs that were perceptually more similar than those from randomly sampled neurons. However, the strength of the columnar effect varied and was not equally evident in all recording sessions. This variability in effect size could have been caused by several factors. First, despite best efforts, our electrode penetrations may not have been perfectly orthogonal to the cortical surface, because the electrodes were aligned relative to the recording chamber. The recording chamber was implanted such that its center was orthogonal to the surface of the cortex. Since the cortex is curved, penetrations further away from the center may not have been perfectly orthogonal to the cortical surface. The brain may also have moved or have been slightly compressed during insertion, potentially resulting in slightly angled penetrations. Second, if there is a topographical organization in V4, it may feature both homogeneous zones and regions with more mixed selectivity, analogous to pinwheels in orientation maps in V1. Thus, we expect a certain fraction of penetrations to be close to such heterogeneous zones. Because extracellular electrodes record the activity of neurons in a roughly cylindrical region around the electrode, we expect a fraction of penetrations to exhibit mixed tuning even if the penetrations were perfectly vertical. Estimating what fraction of penetrations should exhibit consistent tuning is difficult, because the size of the columns and the relationships between them are not yet understood. However, judging from our psychophysical results and the power analysis in Fig. 4h, we consider it likely that more than 50% of the penetrations actually contain a significant bias towards certain types of selectivity. To more accurately delineate and map the topological organization of spatial form tuning in V4, future studies need to combine recording techniques like two-photon functional imaging with deep learning and inception loop approaches.

Similarity to tuning in deep networks The MEIs for V4

cial neural networks trained on image recognition tasks. While there are differences between biological and artificial vision (reviewed in Sinz et al., 2019), deep networks trained on large-scale vision tasks are the closest human engineered system to biological visual system we know. Several previous works have found similarities between the primate visual system and deep network representations of visual stimuli (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016; Yamins et al., 2014). On a single neuron level, previous work has pointed out similarities between tuning in early vision and selectivities of single units in deep networks (Krizhevsky et al., 2012; Olah et al., 2020; Zeiler & Fergus, 2014) and a recent investigation has shown that single units in deep networks can exhibit similar object boundary tuning as V4 neurons (Pospisil et al., 2018). The striking similarity between our single cell MEIs of V4 neurons and single units in the InceptionV1 architecture (Olah et al., 2020) provide an even stronger case for similarities in tuning in primate early vision and deep networks. What's more, by using this similarity we can derive predictions about the color selectivity of V4 neurons despite having shown only gray-scale images in the experiment and during model training (Fig. 5b).

Future directions Our research highlights the power of applying deep learning to comprehensively characterize neuronal representations and helping to elucidate the relations of structure and function of the brain. The concept of cortical columns is an attractive model of the cortex, as it breaks down cortical computation into smaller building blocks of information processing. Our results support the hypothesis that distinct functional groups of neurons organized into cortical columns are not only a feature of primary sensory areas like V1 and S1, but also mid-level visual cortical area V4. Given that neuronal connections within a cortical area tend to be locally dense Gilbert & Wiesel (1989), there's a significantly higher likelihood of connection between neighboring neurons within a column, both within and across layers. Therefore, the purpose of a columnar architecture is not to set precise spatial boundaries between neurons. Instead, it serves to enhance synaptic opportunities between neurons that share similar response properties, facilitating circuit computations such as the development of new types of invariances or the creation of more complex feature selectivity. In this study, our primary focus was on MEIs, with an emphasis on the tuning similarities among neurons recorded from the same probe. However, it is crucial to emphasize that MEIs merely represent the tip of the iceberg. To decipher columnar computations, we need to characterize the full tuning function, including neuronal invariances, as well as examine the differences between neurons within a column. These differences may include variations in their invariances and distinctions in the aspects of the features they encode, which would be consistent with an increasing complexity of feature selectivity across formations occurring within columns and ultimately, when combined with synaptic resolution connectomics, will reveal the underlying mechanisms by which neurons interact to compute Bock et al. (2011); Consortium et al. (2021); Ding et al. (2023a); Reid (2012).

ACKNOWLEDGEMENTS The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Konstantin Willeke and Arne Nix. The authors would also like to thank Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak for their guidance with using their recently developed t-simCNE model. The authors also thank Edgar Y. Walker, George Denfield, Christoph Blessing, Mohammad Bashiri, Konstantin-Klemens Lurz, Max Burg, Shanqian Ma, Robert Petrovic, Elena Offenberg and Paul Fahey for technical support and helpful discussions. The research was funded by the Carl-Zeiss-Stiftung (KW, FHS), the Cyber Valley Research Fund (AN, FHS). FHS is further supported by the German Federal Ministry of Education and Research (BMBF) via the Collaborative Research in Computational Neuroscience (CRCNS) (FKZ 01GQ2107), as well as the Collaborative Research Center (SFB 1233, Robust Vision) and the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1, project number 390727645). ASE received funding for this project from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101041669) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 432680300 (SFB 1456, project B05). The work was also 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. We also acknowledge support from the National Institute of Mental Health and National Institute of Neurological Disorders And Stroke under Award Number U19MH114830 and National Eye Institute award numbers R01 EY026927 and Core Grant for Vision Research T32-EY-002520-37 as well as the National Science Foundation Collaborative Research in Computational Neuroscience IIS-2113173. 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.

Reference

Baden, T., Berens, P., Franke, K., Román Rosón, M., Bethge, M., & Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. *Nature*, *529*(7586), 345–350.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439).

Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., Yurgenson, S., Soucy, E. R., Kim, H. S., & Reid, R. C. (2011). Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, *471*(7337), 177–182.

Böhm, N., Berens, P., & Kobak, D. (2023). Unsupervised visualization of image datasets using contrastive learning. In *The Eleventh International Conference on Learning Representations*. URL https://openreview.net/forum?id=nI2HmVA0hvt

Bonhoeffer, T., & Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature*, *353*(6343), 429–431.

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

Cadena, S. A., Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, (pp. 217–232).

Cadena, S. A., Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2022). Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. *bioRxiv*, (p. 2022.05.18.492503).

Cadwell, C. R., Scala, F., Fahey, P. G., Kobak, D., Mulherkar, S., Sinz, F. H., Papadopoulos, S., Tan, Z. H., Johnsson, P., Hartmanis, L., et al. (2020). Cell type composition and circuit organization of clonally related excitatory neurons in the juvenile mouse neocortex. *Elife*, *9*, e52951.

Calabrese, A., & Paninski, L. (2011). Kalman filter mixture model for spike sorting of non-stationary data. *Journal of neuroscience methods*, *196*(1), 159–169.

Campagnola, L., Seeman, S. C., Chartrand, T., Kim, L., Hoggarth, A., Gamlin, C., Ito, S., Trinh, J., Davoudian, P., Radaelli, C., et al. (2022). Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, *375*(6585), eabj5861.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. URL https://arxiv.org/abs/2002.05709

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Consortium, M., Bae, J. A., Baptiste, M., Bishop, C. A., Bodor, A. L., Brittain, D., Buchanan, J., Bumbarger, D. J., Castro, M. A., Celii, B., et al. (2021). Functional connectomics spanning multiple areas of mouse visual cortex. *BioRxiv*, (pp. 2021–07).

Denfield, G. H., Ecker, A. S., Shinn, T. J., Bethge, M., & Tolias, A. S. (2018). Attentional fluctuations induce shared variability in macaque primary visual cortex. *Nature communications*, *9*(1), 1–14.

hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, (pp. 248–255). Ieee.

Ding, Z., Fahey, P. G., Papadopoulos, S., Wang, E., Celii, B., Papadopoulos, C., Kunin, A., Chang, A., Fu, J., Ding, Z., et al. (2023a). Functional connectomics reveals general wiring rule in mouse visual cortex. *bioRxiv*, (pp. 2023–03).

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

Ecker, A. S., Berens, P., Cotton, R. J., Subramaniyan, M., Denfield, G. H., Cadwell, C. R., Smirnakis, S. M., Bethge, M., & Tolias, A. S. (2014). State dependence of noise correlations in macaque primary visual cortex. *Neuron*, *82*(1), 235–248.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., & Tsipras, D. (2019a). Robustness (python library).
URL https://github.com/MadryLab/robustness

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., & Madry, A. (2019b). Adversarial robustness as a prior for learned representations.
URL https://arxiv.org/abs/1906.00945

Feather, J., Leclerc, G., Mądry, A., & McDermott, J. H. (2022). Model metamers illuminate divergences between biological and artificial neural networks.
URL https://doi.org/10.1101/2022.05.19.492678

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

Franke, K., Willeke, K. F., Ponder, K., Galdamez, M., Zhou, N., Muhammad, T., Patel, S., Froudarakis, E., Reimer, J., Sinz, F. H., et al. (2022). State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, *610*(7930), 128–134.

Fu, J., Shrinivasan, S., Ponder, K., Muhammad, T., Ding, Z., Wang, E., Ding, Z., Tran, D. T., Fahey, P. G., Papadopoulos, S., Patel, S., Reimer, J., Ecker, A. S., Pitkow, X., Haefner, R. M., Sinz, F. H., Franke, K., & Tolias, A. S. (2023). Pattern completion and disruption characterize contextual modulation in mouse visual cortex.

Gallant, J. L., Braun, J., & Van Essen, D. C. (1993). Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, *259*(5091), 100–103.

Ghose, G. M., & Ts'o, D. Y. (1997). Form processing modules in primate area V4. *J. Neurophysiol.*, *77*(4), 2191–2196.

Gilbert, C. D., & Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, *9*(7), 2432–2442.

Goetz, J., Jessen, Z. F., Jacobi, A., Mani, A., Cooler, S., Greer, D., Kadri, S., Segal, J., Shekhar, K., Sanes, J. R., & Schwartz, G. W. (2022). Unified classification of mouse retinal ganglion cells using function, morphology, and gene expression. *Cell Rep.*, *40*(2), 111040.

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

Guo, C., Lee, M. J., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & DiCarlo, J. J. (2022). Adversarially trained neural representations may already be as robust as corresponding biological neural representations.
URL https://arxiv.org/abs/2206.11228

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, *12*(10), 993–1001.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.
URL https://doi.org/10.1038/s41586-020-2649-2

Hatanaka, G., Inagaki, M., Takeuchi, R. F., Nishimoto, S., Ikezoe, K., & Fujita, I. (2022). Processing of visual statistics of naturalistic videos in macaque visual areas V1 and V4. *Brain Struct. Funct.*, *227*(4), 1385–1403.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

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

Horton, J. C., & Adams, D. L. (2005). The cortical column: a structure without a function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *360*(1456), 837–862.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, *195*(1), 215–243.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, (pp. 448–456). PMLR.

Issa, N. P., Trepel, C., & Stryker, M. P. (2000). Spatial frequency maps in cat visual cortex. *Journal of Neuroscience*, *20*(22), 8504–8514.

Jiang, X., Shen, S., Cadwell, C. R., Berens, P., Sinz, F., Ecker, A. S., Patel, S., & Tolias, A. S. (2015). Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, *350*(6264), aac9462.

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

Kim, T., Bair, W., & Pasupathy, A. (2019a). Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, *39*(24), 4760–4774.

Kim, T., Bair, W., & Pasupathy, A. (2019b). Neural coding for shape and texture in macaque area V4. *J. Neurosci.*, *39*(24), 4760–4774.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., notebooks – a publishing format for reproducible computational workflows. In F. Loizides, & B. Schmidt (Eds.) *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, (pp. 87 – 90). IOS Press.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, *71*(3), 856–867.

Kotake, Y., Morimoto, H., Okazaki, Y., Fujita, I., & Tamura, H. (2009). Organization of color-selective neurons in macaque visual area V4. *J. Neurophysiol.*, *102*(1), 15–27.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
URL https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Lehky, S. R. R. S. R. R., Sejnowski, T. J. J., & Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *J. Neurosci.*, *12*(9), 3568–3581.

Li, Z., Brendel, W., Walker, E. Y., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F. H., Pitkow, X., & Tolias, A. S. (2019). Learning from brains how to regularize machines.
URL https://arxiv.org/abs/1911.05072

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts.
URL https://arxiv.org/abs/1608.03983

Lurz, K.-K., Bashiri, M., Willeke, K. F., Jagadish, A. K., Wang, E., Walker, E. Y., Cadena, S., Muhammad, T., Cobos, E., Tolias, A., et al. (2020). Generalization in data-driven models of primary visual cortex. *bioRxiv*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks.
URL https://arxiv.org/abs/1706.06083

Masland, R. H. (2001). The fundamental plan of the retina. *Nat. Neurosci.*, *4*(9), 877–886.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205.
URL https://doi.org/10.21105/joss.00205

Mechler, F., & Ringach, D. L. (2002). On the classification of simple and complex cells. *Vision research*, *42*(8), 1017–1033.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, *2014*(239), 2.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, *6*, 414–417.

Mordvintsev, A., Pezzotti, N., Schubert, L., & Olah, C. (2018). Differentiable image parameterizations. *Distill*. Https://distill.pub/2018/differentiable-parameterizations.

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.*, *20*(4), 408–434.

Ohki, K., & Reid, R. C. (2014). In vivo two-photon calcium imaging in the visual system. *Cold Spring Harbor Protocols*, *2014*(4), pdb–prot081455.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). An overview of early vision in inceptionv1. *Distill*. Https://distill.pub/2020/circuits/early-vision.

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, *86*(5), 2505–2519.

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.*, *5*(12), 1332–1338.

Pasupathy, A., Popovkina, D. V., & Kim, T. (2020). Visual functions of primate area V4. *Annu Rev Vis Sci*, *6*, 363–385.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 32*, (pp. 8024–8035). Curran Associates, Inc.

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, *60*, 389–443.

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'artiphysiology' reveals v4-like shape tuning in a deep network trained for image classification. *Elife*, *7*.

Reid, R. C. (2012). From functional architecture to functional connectomics. *Neuron*, *75*(2), 209–217.

Safarani, S., Nix, A., Willeke, K., Cadena, S. A., Restivo, K., Denfield, G., Tolias, A. S., & Sinz, F. H. (2021). Towards robust vision by multi-task learning on monkey visual cortex.
URL https://arxiv.org/abs/2107.14344

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., & Madry, A. (2020). Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*.

Shan, K. Q., Lubenov, E. V., & Siapas, A. G. (2017). Model-based spike sorting with a mixture of drifting t-distributions. *Journal of neuroscience methods*, *288*, 82–98.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967–979.

Srinath, R., Emonds, A., Wang, Q., Lempel, A. A., Dunn-Weiss, E., Connor, C. E., & Nielsen, K. J. (2020). Early emergence of solid shape coding in natural and deep network vision. *Curr. Biol.*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1–9).

Tanaka, M., Weber, H., & Creutzfeldt, O. (1986). Visual properties and spatial distribution of neurones in the visual association area on the prelunate gyrus of the awake monkey. *Experimental Brain Research*, *65*, 11–37.

Tang, R., Song, Q., Li, Y., Zhang, R., Cai, X., & Lu, H. D. (2020). Curvature-processing domains in primate V4. *Elife*, *9*.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nature neuroscience*, *6*(9), 989–995.

Ustyuzhaninov, I., Burg, M. F., Cadena, S. A., Fu, J., Muhammad, T., Ponder, K., Froudarakis, E.,

code of non-linear computations in the mouse primary visual cortex. *bioRxiv*.
URL https://www.biorxiv.org/content/early/2022/02/10/2022.02.10.479884

Ustyuzhaninov, I., Cadena, S. A., Froudarakis, E., Fahey, P. G., Walker, E. Y., Cobos, E., Reimer, J., Sinz, F. H., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Rotation-invariant clustering of neuronal responses in primary visual cortex.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579–2605.
URL http://jmlr.org/papers/v9/vandermaaten08a.html

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: image processing in python. *PeerJ*, *2*, e453.

Victor, J. D., Purpura, K., Katz, E., & Mao, B. (1994). Population encoding of spatial frequency, orientation, and color in macaque v1. *Journal of neurophysiology*, *72*(5), 2151–2166.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.*.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., & Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (september 2017).
URL https://doi.org/10.5281/zenodo.883859

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

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

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

Yatsenko, D., Reimer, J., Ecker, A. S., Walker, E. Y., Sinz, F., Berens, P., Hoenselaar, A., Cotton, R. J., Siapas, A. S., & Tolias, A. S. (2015). Datajoint: managing big scientific data using matlab or python. *BioRxiv*, (p. 031658).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, vol. 8689, (pp. 818–833).

AUTHOR CONTRIBUTIONS
KFW Conceptualization, Methodology, Validation, Software, Formal Analysis, Investigation, Writing - Original Draft, Visualization **KR** Conceptualization, Methodology, Validation, Software, Formal Analysis, Investigation, Data acquisition, Data curation, Writing - Original Draft, Visualization; **KF** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Original Draft, Writing - Review & Editing, Supervision; **AFN** Methodology, Software, Writing - Review & Editing; **SAC** Conceptualization, Methodology, Software; **TS,CN,GR,SP** Data acquisition, Data curation, Methodology; **ASE** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Review & Editing, Supervision; **FHS** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding Acquisition, Project administration; **AST** Conceptualization, Methodology, Investigation, Experimental and analysis design, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding Acquisition, Project administration;

Materials and Methods

Ethics statement Electrophysiological data were gathered from a pair of healthy male rhesus macaque monkeys (*Macaca mulatta*), aged 17 and 19 years, and weighing 16.4 and 10.5 kg, respectively, at the time of the study. The research adhered to NIH guidelines and received approval from the Institutional Animal Care and Use Committee at Baylor College of Medicine (permit number: AN-4367). The monkeys were individually housed in a spacious room near the training facility, in the company of approximately ten other monkeys, allowing for abundant visual, olfactory, and auditory interactions. They were maintained on a 12-hour light/dark cycle.

The Center for Comparative Medicine at Baylor College of Medicine ensured that the monkeys received regular veterinary check-ups, balanced nutrition, and environmental enrichment. Surgical procedures involving the monkeys were performed using general anesthesia and adhering to standard aseptic techniques. Postoperative pain was man-

Electrophysiological recordings Non-chronic recordings were conducted using a 32-channel linear silicon probe (NeuroNexus V1x32-Edge-10mm-60-177), with surgical methods and recording protocols previously outlined (Denfield et al., 2018). In summary, custom titanium recording chambers and head posts were implanted under complete anesthesia and sterile conditions. Initially, the bone remained unaltered, and only before recordings were small trephinations (2 mm) made over lateral V4, with eccentricities spanning from 1.7 to 18.3 degrees of visual angle. Recordings took place after a week of each trephination. A Narishige Microdrive (MO-97) and guide tube were used to carefully lower the probes, penetrating the dura.

Data acquisition and spike sorting Electrophysiological data were continuously collected as a broadband signal (0.5Hz–16kHz), digitized at 24 bits. The spike sorting methods employed in this study resemble those used in (Cadena et al., 2019; Denfield et al., 2018), with the code accessible at https://github.com/aecker/moksm. The linear array of 32 channels was divided into 14 groups, each containing six neighboring channels (with a stride of two), which were treated as virtual electrodes for spike detection and sorting. Spikes were identified when channel signals exceeded a threshold equal to five times the standard deviation of the noise.

Following spike alignment, the first three principal components of each channel were extracted, resulting in an 18-dimensional feature space utilized for spike sorting. A Kalman filter mixture model was fitted to monitor waveform drift, which is common in non-chronic recordings (Calabrese & Paninski, 2011; Shan et al., 2017). Each cluster's shape was modeled using a multivariate t-distribution (degrees of freedom = 5) with a ridge-regularized covariance matrix. The cluster count was determined based on a penalized average likelihood with a constant cost for each additional cluster (Ecker et al., 2014).

Lastly, a custom graphical user interface was used to manually confirm single-unit isolation by evaluating the units' stability (based on drifts and cell health throughout the session), identifying a refractory period, and inspecting scatter plots of channel principal component pairs.

Visual stimulation and eye tracking Visual stimuli were generated by a specialized graphics workstation and presented on a 16:9 HD widescreen LCD monitor (23.8") with a 100 Hz refresh rate and a resolution of 1920×1080 pixels, positioned at a 100 cm viewing distance (yielding approximately $\sim 63px/^\circ$). The monitor underwent gamma correction to ensure a linear luminance response profile. A custom-made, camera-based eye tracking system confirmed that monkeys kept their gaze within roughly $\sim 0.95^\circ$ around a $\sim 0.15^\circ$-sized red fixation target. Offline evaluations revealed that the monkeys typically fixated with greater accuracy.

Upon maintaining fixation for 300 ms, a visual stimulus was displayed. If the monkeys sustained their gaze throughout

drop of juice at the end of the trial.

Receptive field mapping and stimulus placing At the start of each session, we determined receptive fields in relation to a fixation target using a sparse random dot stimulus. A solitary dot, spanning 1° of the visual field, was displayed on a uniform gray background, with its location and polarity (black or white) randomly changing every 30 ms. Each fixation trial persisted for two seconds. Multi-unit receptive field profiles for each channel were obtained through reverse correlation. The population receptive field location was estimated by fitting a 2D Gaussian to the spike-triggered average across channels at the time lag that optimized the signal-to-noise ratio.

The natural image stimulus occupied the entire screen. The fixation spot was adjusted so that the mean of the population receptive field was as close to the screen's center as feasible. Due to the recording sites' location in both monkeys, this positioning involved placing the fixation spot near the screen's upper border, shifted to the left.

Natural image stimuli We selected a collection of 24,075 images from 964 categories (~ 25 images per category) from ImageNet (Deng et al., 2009), transformed them to grayscale, and cropped the central 420×420 pixels. For images that were smaller than 420×420, a central crop was taken and the resulting image was re-scaled to 420×420. Each image had an 8-bit intensity resolution (values ranging from 0 to 255). From this set, we randomly chose 75 images as our *test-set*. Out of the remaining 24,000 images, we designated 20% as *validation-set* at random, leaving 19,200 images in the *train-set*. Natural images were displayed during the standalone generation recordings of 1,244 units and during the generation phase of closed-loop recordings for 82 units. Specifically, ~ 12k unique *train-set* images were displayed during the standalone generation recordings, and ~ 7.5k unique *train-set* images were displayed during the generation phase of closed-loop recordings. Across sessions, train images were randomly sampled from the *train-set* such that the full set was exhausted before cycling back through the same images. The 75 *test-set* images were displayed in every recording session. Note that selecting images from ImageNet means that the pre-trained convolutional network (see below) has likely seen our natural stimuli (but not the neuronal responses) during training on the classification task.

During standalone generation recording sessions, ~ 1000 successful trials (~ 12k train images and 75 repeated test images) were recorded, whereas 600 successful trials (~ 7.5k train images and 75 repeated test images) were recorded during the generation phase of closed-loop experiments. In both instances, each trial involved continuous fixation for 2.4 seconds, which includes 300 ms of a gray screen (intensity 128) at the beginning and end of the trial, as well as 15 consecutive images displayed for 120 ms each without any gaps. Trials contained either train-images, which were repeated during the experiment. Throughout the recording session, all trails were randomly interleaved, with test images being repeated 40-50 times during standalone generation recordings and 20 times during the generation phase of closed-loop recordings. Training and validation images were sampled without replacement, so each image was effectively displayed once or not at all. Images were upscaled using bicubic interpolation to match the screen width (1920 pixels) while maintaining their aspect ratio. The upper and lower 420-pixel bands were cropped out to cover the entire screen, effectively stimulating both classical and beyond classical receptive fields of V4 neurons. After sorting the neurons, spikes associated with each image presentation were counted within a 70-160 ms time window after stimulus onset.

Image preprocessing for model training. Starting out from an original image size of 420×420 with a resolution of $14\text{px}/^\circ$ we cropped the upper and lower bands to fit the full screen for presentation for a resulting image size of 420×236. We then cropped the images so that only the bottom center 200×200 pixels remained. Then, we downsampled the images to either 80×80 or 100×100 pixels ($5.8\text{px}/^\circ$ or $7\text{px}/^\circ$), for the closed-loop model training and non closed-loop model training, respectively.

Model architecture Our neural predictive model of primate V4 consisted of two main parts: A pretrained *core* that computes nonlinear features of input images, and a *Gaussian readout* (Lurz et al., 2020) that maps these features to the neuronal responses of the single neurons.

As the core of our model, we used a ResNet50 (He et al., 2016) which was adversarially trained on ImageNet (Deng et al., 2009) to have robust visual representations (Salman et al., 2020), which yields improved transfer-learning performance (Engstrom et al., 2019a,b; Madry et al., 2017). Interestingly, it has been previously shown that robust features not only allow for better transfer-learning but they appear to be more similar to biological networks and also improve neural predictivity (Feather et al., 2022; Guo et al., 2022; Li et al., 2019; Safarani et al., 2021). Building on previous work (Cadena et al., 2022), we selected the first residual block of layer 3 of the ResNet, `layer3.0`, to read out from, and found that the adversarially robust training with $\epsilon = 0.1$ yielded the highest predictive performance, compared to all other ResNet models and layers. The corresponding size of the output feature map at layer `layer3.0` was 1024.

The input images $\mathbf{x}$ were forwarded through all layers up to a selected layer, to output a tensor of feature maps. Importantly, the parameters of the pretrained network were always kept fixed. We then applied batch-normalization (Ioffe & Szegedy, 2015). Lastly, we rectified the resulting tensor with a ReLU unit to obtain the final nonlinear feature space $\Phi(\mathbf{x}) \in \mathbb{R}^{w \times h \times c}$ (**w**idth, **h**eight, **c**hannels) shared by all neurons.

To predict the response of a single neuron from the $\Phi(\mathbf{x}) \in$

For each neuron n, this readout learns the coordinates $(x^{(n)}, y^{(n)})$ of the position of the receptive field on the output tensor and extracts a feature vector $\Phi_{x^{(n)},y^{(n)},:} \in \mathbb{R}^c$ at this location from Φ. To this end, the Gaussian readout learns the parameters of a 2D Gaussian distribution $\mathcal{N}(\mu_n, \Sigma_n)$ and samples a location in feature space $\Phi(\mathbf{x})$ during each training step for every neuron n. Σ_n is initialized large enough to ensure that the entire visual field can be covered, and then decreases in size during training to have a more reliable estimate of the mean location μ_n. At inference time (i.e. when evaluating our model), the readout is deterministic and uses the fixed position μ_n. Although this framework allows for rotated and elongated Gaussian functions, we found that for our data, an isotropic formulation of the covariance – parametrized by a single scalar σ_n^2 – was performing equally well as compared to a fully parametrized Gaussian. Taken together, total number of parameters per neuron of the readout were $c+4$ (number of channels, bivariate mean, variance, and bias). This extracted feature vector $\Phi_{x^{(n)},y^{(n)},:}$ is then used in a linear-nonlinear model to predict the neuronal response. To this end, an affine function of the resulting feature vector at the chosen location was computed, followed by a rectifying nonlinearity f, chosen to be an ELU (Clevert et al., 2015) offset by one ($\mathrm{ELU}+1$) to make responses positive (Eq. 1). The weight vector $\mathbf{w}_n \in \mathbb{R}^c$ was L_1 regularized during training.

$$\hat{r}_n(\mathbf{x}) = f\left(\sum_k \Phi_{x^{(n)},y^{(n)},k}(\mathbf{x}) \cdot w_{n,k} + b_n\right) \quad \textbf{(1)}$$

Model training We trained our model to minimize the summed Poisson loss across N neurons between observed spike counts r and the models' predicted spike counts $\hat{r}$ in addition to the L_1 regularization of the readout parameters.

$$\mathcal{L} = \sum_{i=1}^{N} (\hat{r}_n - r_n \log \hat{r}_n) + \lambda \sum_{n,k} |w_{nk}| \quad \textbf{(2)}$$

We trained the models either on the full dataset of n=100 recording sessions with n=1244 neurons and an image size of 100 by 100 pixels, or on an individual sessions during a closed loop recording with a reduced image size of 80 by 80 pixels in order to save time with model training and stimulus generation. For training on the full dataset, an epoch consisted of the cycling through the whole training set of 19200 images. However, because only ≈9,000 - 13,000 images were shown in each session, if a particular image was not shown in a session, we simply zeroed out the gradients of all neurons in the sessions that did not contain the image in question. When training models on a single closed loop recording, we cycled through all ≈9,000 training images that were shown in that session. In all cases, we used a batch size of 64 and after each batch, we updated the weights using the Adam optimizer (Kingma & Ba, 2014). The initial learning rate was $3 \cdot 10^{-4}$ After each epoch, we computed the Poisson loss on the entire validation set. Similar to the training set, not all images were shown in all sessions, so that we again zeroed out the loss for the sessions in question. We then used early stopping to decide whether to decay the learning rate or stop the training altogether. We scaled the learning rate by a factor of 0.3 once the validation loss did not improve for five consecutive epochs. Before decaying the learning rate, we restored the weights to the best ones based on the poisson loss on the validation set. After four early stopping steps were completed, we stopped the training. On average, this resulted in ≈ 50 training epochs, for a training time of 2 minutes for a closed loop session, and 15 minutes for the entire dataset on a NVIDIA 2080ti GPU.

Ensemble models. Instead of using a single trained model, we used a model ensemble for all of our analyses and for MEI generation. To predict the neuronal responses to individual images, we trained readout weights for each member of an ensemble of five models initialized with different random seeds and used the average prediction across the ensemble for further analyses (Hansen & Salamon, 1990). We always trained ten individual models with a different random seed, which determined the model initialization as well as the drawing of training batches. Then, we selected the five models with the highest performance on the validation set to form a model ensemble. The inputs to the ensemble model were passed to each member, and the resulting predictions were averaged to obtain the final model prediction.

Explainable variance As a measure of response reliability, we estimated the fraction of the stimulus-driven variability as compared to the overall response variability. More specifically, we computed the ratio between each neurons' total variance minus the variance of the observation noise, over the total variance (Eq. 3).To estimate the variance of the observation noise, we averaged the variance of responses across image repeats for all of the 75 repeated full-field natural image test stimuli $\sigma^2_{noise} = \mathrm{E}_j[\mathrm{Var}_t[r_t|x_j]]$ where t corresponds to the repeats and x_j represents a unique image:

$$\mathrm{EV} = \frac{\mathrm{Var}[r] - \sigma^2_{noise}}{\mathrm{Var[r]}}. \quad \textbf{(3)}$$

Model performance measures To measure the predictive performance of our models, we calculated the correlation to average (Cadena et al., 2022; Franke et al., 2022; Willeke et al., 2022) on the held out test set images. Given a neuron's response r_{ij} to image i and repeat j and the model predictions o_i, the correlation is computed between the predicted responses and the average neuronal response $\bar{r}_i$ to the i^{th} test image (averaged across repeated presentations of the same stimulus):

$$\mathrm{corr}(\bar{r}, o) = \frac{\sum_i (\bar{r}_i - \bar{r})(o_i - \bar{o})}{\quad} \quad \textbf{(4)}$$

where $\bar{r}_i = \frac{1}{J}\sum_{j=1}^{J} r_{ij}$ is the average response across J repeats, $\bar{r}$ is the average response across all repeats and images, and $\bar{o}$ is the average prediction across all repeats and images.

Generation of MEIs We used the trained model to synthesize maximally exciting input images (MEIs) for each neuron using regularized gradient ascent (Bashivan et al., 2019; Franke et al., 2022; Walker et al., 2019). Starting out with a randomly initialized Gaussian white noise image given by $\mathbf{x} \in \mathbb{R}^{h \times w}$, with height h and width w, we showed the image to the model and computed the gradients of a single target neuron w.r.t. the image. To avoid high frequency artifacts, after each iteration we applied Gaussian blur with a σ of 1 pixel to smoothen the image. Additionally, we constrained the entire image to have a fixed energy budget, which we implemented as a maximum L2 norm of the standardized image, calculated across all pixel intensities. We chose an L2 norm of 25 for all neurons such that the resulting MEIs had minimal and maximal values similar to those found in our training natural image distribution. If the MEI exceeded the allowed norm budget after any iteration, we divided the MEI by factor f_{norm} with $f_{norm} = \|\text{MEI}\|_2 / b$. Additionally, enforced that the MEI could not contain values outside of the 8-bit pixel range by clipping the MEI outside of the bounds that correspond to 0 or 255 pixel-intensity. We used the stochastic gradient descent (SGD) optimized with learning rate of 10. We ran each optimization for 1000 iterations, without early stopping.

MEIs with transparency masks Furthermore, we employed a novel technique of synthesizing MEIs using a transparency channel based on the idea of Mordvintsev et al. (2018). Given an MEI optimized with a method as described in the section above, it is difficult to distinguish which of the MEI features are important, i.e. strongly activate the neuron, and which ones are not. Through our inbuilt L2 energy constraint, there only is finite contrast that the model is able to distribute across the image. However, this is still uninformative regarding which features of the MEI are most important. Thus, we adopted a transparency method as a differentiable parametrization (Mordvintsev et al., 2018), which jointly optimizes the MEI and a transparency mask, with the objective of making the MEI itself as transparent as possible (i.e. uninformative parts of the MEI), while still retaining the high neuronal activation of the resulting image. More specifically, we optimized an image $\mathbf{x} \in \mathbb{R}^{c \times h \times w}$, with channels c, height h and width w. We set the channels $c = 2$, treated the first channel as the MEI $\mathbf{x}_{mei}$, and the second channel as a transparent mask $\mathbf{x}_\alpha$, which we optimized jointly as follows: (1) In each iteration, we drew a random image $\mathbf{x}_{bg}$ from the training set as a background. (2) We clip the second MEI channel, i.e. the transparent mask, between the values 0–1, with 1 meaning that the MEI is fully opaque, and 0 meaning the the MEI is not visible at all. (3) Then, we blend the background and the MEI with the transparent mask $\mathbf{x}_\alpha$ to get the combined image $\mathbf{x}_{combined}$ with

$$\mathbf{x}_{combined} = x_{bg} \times (1 - x_\alpha) + x_{mei} \times x_\alpha \quad \textbf{(5)}$$

We then showed the combined image to the model to compute the neuronal response of the target neuron. With no additional constraints, the model did in fact learn to set all values of the transparent mask to 1, so that it is necessary to change the MEI objective function $\mathbf{L}_{old}$ from simply maximizing the neuronal response r to penalize alpha values of zero, as suggested by Mordvintsev et al. (2018):

$$\mathbf{L}_{new} = r \times (1 - mean(x_\alpha)) \quad \textbf{(6)}$$

We used the same learning rate and optimizer as for our default gradient descent MEIs. We also applied the same L2 contrast constraint of 25, but only to the MEI image channel x_{mei}. For our closed loop experiments, we ran the initial 16 sessions with the default gradient ascent MEIs, and the final 8 sessions with the transparency method, observing similar success for the *in vivo* verification. Because the transparency method seemed to generate MEIs that were less spread out and had features that were perceptually easier to identify, we used exclusively this method to generate the MEIs for the full dataset. For all the 889 out of 1244 neurons that passed our threshold of correlation to average larger than 0.3, we generated 50 MEIs per neuron for a total of 44,450 MEIs. Each MEI took on average 2 minutes on a NVIDIA 2080ti GPU, resulting in a total compute time of ≈60 GPU weeks.

Centering of MEIs In this work, we do not further analyze the transparent MEI masks. However, we use the transparent masks to center the MEIs for the similarity quantification. After MEI optimization, the MEI is located at the approximate RF location of each neuron. To center the MEIs, we find the center of mass of the transparent mask, and move the MEI such that the center of mass is at the center. All MEIs shown throughout this manuscript are shown at the center location for ease of comparison.

Psychophysics In our psychophysics experiment,the objective was to assess the perceptual similarity between MEIs originating from either a single recording session as compared to a random selection of neurons from different recording sessions. For this purpose, we devised a two-alternative forced choice task in which we showed two sets of MEIs, arranged in 3x3 grids as the basis for the similarity evaluation task. Participants were instructed to simply report which of the MEIs in the two grids looked more similar to each other. No other information apart from the displayed MEIs was provided.
To obtain the images for the task, we randomly chose 50 out of 52 sessions that had a minimum of 9 well-predicted neurons, using the threshold of correlation to average larger than 0.3. From these selected sessions, the top 9 well-predicted neurons were identified, resulting in a total pool of 450 neurons. For each session, participants were

and a randomly selected set of 9 additional neurons from the pool, sampled without replacement. Specifically, for each neuron, we presented five MEIs, generated by starting from a different initial noise image during optimization, as images and compiled in a GIF. Participants were shown the two example sessions which were not included in the study, together with the correct solution. The subjects had unlimited time to make their choice.

Subject recruitment involved lab members that were familiar with the concept of an MEI from all research groups contributing to this article. In total, we recorded the responses of N=25 observers.

Closed-loop procedure Our closed loop experiments were composed of a generation and verification session. In the generation session, we recorded the electrophysiological responses of V4 neurons to full-field gray-scale natural images for 600 trials, corresponding to 9,000 unique images. After showing of all of these trials, we stopped and promptly restarted the neurophysiological acquisition system to continue recording neuronal activity while we processed the data from the generation session. During this analysis period, we isolated single units and used these units to train several models to predict neuron responses to natural images. We produced an ensemble of the best five models, used this ensemble to select the six highest predicted units and generate ten MEIs per unit. We computed the mask of each MEI by calculating the z-score of each pixel value and setting a threshold ($z = 0.35$) to isolate the area within the mask. We then created a convex hull to close any holes and smoothed the edges with a Gaussian filter ($\sigma = 2$). We used this masking procedure for our natural image control selection procedure. Specifically, we screened a distinct set of 5,000 natural images, different from our training, validation, and test set images. This set was selected randomly from a database of 100k images, and we screened this same image set for all neurons across all closed-loop experiments. Each image was masked with the ten MEI masks of each unit and re-normalized to the contrast of the MEIs, resulting in 50,000 total screened images for each unit. We selected the seven highest activating masked natural images of each unit to use as controls for the verification recording.

Closed-loop stimulation paradigm During the verification session of the closed-loop experiment, we displayed one MEI and seven controls for each of the six units. These stimuli were centered on the RF and repeated 10 times throughout the experiment. We also showed the same full-field test set natural images 20 times each, to enable functional response matching between the generation and verification units. Each trial contained 15 images, composed of either (i) randomly interleaved generated stimuli (i.e. MEIs and masked controls) or (ii) full-field test set natural images, as described previously. Trials were randomly interleaved. In total, 100 trials of full-field images and 180

Closed-loop unit matching To assess the stability of our closed-loop neurons, we developed a procedure for spike waveform matching single units between the generation and verification session recordings. To do this, we performed a *post hoc* spike sorting of the full recording session, which was produced by stitching together the raw data files of the generation and verification recordings. The verification recording includes the analysis period immediately after the generation recording, during which we isolate single units and generate their MEIs. This continuity of recording allows us to more easily track single units throughout the experiment. Furthermore, spike sorting the full session allows us to use a drift correction of the spike sorting model based on a kalman-filter across the entire experiment, which aided the accuracy of our single unit tracking.

We then executed a two-step matching procedure using the spike sorting results of both the generation and full sessions. To determine potentially matched units, we first took the spikes of the generation session units and assigned them to full session units. We then calculated the proportion of spikes that were assigned exclusively to one unit in both the generation and verification sessions. If a unit in the generation session had at least 95% of its spikes assigned to only one single unit in the full session, this was considered a possible match. To confirm true matches, we assigned the full session spikes to the generation session units. If a unit in the full session had at least 95% of its spikes assigned to the potentially matched single unit in the generation session, these units were verified to be a match, and thus certified as stable. We additionally assessed neuron stability by computing the functional consistency of each unit. To compute this measure, we used the test set of 75 full-field natural images that were shown in both generation and verification session and calculated the Pearson correlation to the repeat-averaged responses in both sessions. We set a minimum functional consistency threshold of 0.5. Taken together, we were able to waveform match 82 neurons from 24 closed-loop sessions, with 27 of these failing to meet the functional consistency criterion, resulting in n=55 neurons for the analysis of the closed-loop paradigm.

MEI similarity quantification using contrastive learning We used a recently proposed method of contrastive learning (Böhm et al., 2023) to quantify the perceptual similarity of our neurons' MEIs. The method is based on SimCLR (Chen et al., 2020), a method of self-supervised contrastive learning of visual representations in which images are embedded in a high-dimensional space. In this space, augmentations of the same image are learned to have small distances while simultaneously training the model to increase the distance to all other images. Böhm et al. (2023) adapted this procedure for a two-dimensional space by combining the ideas of contrastive learning with 2-d neighbor embeddings as used in t-SNE (van der Maaten & Hinton, 2008). Their new method, called t-

between image embeddings from cosine-similarity used in SimCLR, which would constrain the embeddings to the unit circle, to Euclidean distance $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|$ between the embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$ of two augmentations i and j of the same image. We use this method to train a model end-to-end to embed our MEIs using the loss function as proposed by Böhm et al. (2023):

$$\ell_{t-\mathrm{SimCNE}}(i,j) = -\log \frac{1/\left(1+d_{ij}^2\right)}{\sum_{k\neq i}^{2b} 1/\left(1+d_{ik}^2\right)}. \quad \textbf{(7)}$$

In our adaptation, i and j correspond to two data augmentations of the same MEI, and z denotes the 2-d embedding, which is the model output. For a given batch size b, the resulting batch size is $2b$ because each MEI is augmented twice. As pointed out by the authors of SimCLR (Chen et al., 2020), this approach requires careful selection of data augmentations, as well as choosing the largest batch size possible.

Taken together, our training procedure consists of these steps: We first select all 889 neurons above the model performance threshold of correlation to average larger than 0.3, as described earlier. For each neuron, from the 50 MEIs that we optimized per neuron, we select only the MEIs that elicited at least 90% of the predicted firing rate of the highest-acivating MEI, which resulted in anywhere from 3 to 42 MEIs per neuron, with a total of n=19,688 MEIs used in this and subsequent analyses. This selection was important as there can be failures during optimizing MEIs, resulting in noisy images which the contrastive learning algorithm was very sensitive to. Next, from all the MEIs that we selected, we assembled a batch of size n=889, by randomly drawing one MEI for each neuron. Subsequently, we preprocessed the MEIs and applied the data augmentations in the following order: (1) centering the MEIs as described above, (2) center cropping to resize the MEIs from 100x100 to 75x75 pixels, (3) random rotation between 0 and 30 degrees, (4) random rescaling between 32x32 and 75x75 pixels, (5) random cropping from the resulting image to a final size of 32x32 pixels. Steps (2) to (4) were applied twice to each MEI to obtain two augmentations per MEI. The resulting training batch thus consisted of a randomly drawn MEI per neuron, augmented twice, for a total batch size of 889*2=1,778 images, with the model's objective being to minimize the distance between each pair, and to maximize the distance from all-to-all images which are not pairs. It is important to point out that with this training scheme, the trained model was given no information about recording sessions or which MEIs belong to which neuron. Once the model was fully trained, we obtained the full 2-d embedding of all MEIs from all neurons by applying transformations (1) and (2), i.e. centering and center cropping the MEIs, as well as rescaling the MEIs to 32x32 pixels, and subsequently showing the resulting images to the model. Finally, We obtained a 2-d location of each neuron by taking the mean location over

Contrastive learning: model architecture and training We closely followed the t-simCNE authors (Böhm et al., 2023) in our choice of model architecture and training paradigm. As a model backbone, we employed a randomly initialized ResNet18 (He et al., 2016) with an output size of 512 and a reduced kernel size of the first convolutional layer from 7×7 to 3×3. Following the authors, we also added one hidden ReLU layer with n=1,024 units followed by a linear output layer of n=128 units, which was reduced to n=2 during the final stage of training. The training consisted of three stages: First, the model was trained for 3,000 epochs with the output layer size of 128. In the second stage, the output layer was disregarded and replaced with a linear output layer of n=2, followed by a training of only this layer for 200 epochs while the rest of the model was frozen. Lastly, in the third stage, the whole model was fine-tuned for another 1,000 epochs with a reduction of the learning rate by a factor of 1,000. In each of the three stages, we used the initial learning rate of $0.03 \cdot b/256 \approx 0.1$ with b=889, preceded by linear warm-up for ten epochs, followed by cosine annealing (Loshchilov & Hutter, 2016) with a final learning rate of 0. In each epoch, we accumulated the loss from 10 batches and optimized the model using SGD with a momentum of 0.9.

Contrastive learning: within-neuron MEI distances As mentioned above, it is crucial that our contrastive learning training scheme did not provide the model with any information about either the recording sessions or the neurons' identity of each MEI. Therefore, an important sanity-check for the trained model is to analyze the 2-d embedding distances of all MEIs that we optimized for a single neuron. We performed this analysis by simply calculating all pairwise distances and taking the average of all MEIs per neuron.

Contrastive learning: Session distances We first averaged all pairwise distances across neurons recorded within one session to obtain a within-session distance in the 2-d embedding space. We then compared this within-session distance to a shuffled control. For this control, we used bootstrapping to sample at random the same number of neurons as in a given session, with the constraint that they cannot be from the session in question and that all randomly drawn neurons originate from different sessions. We then computed the mean pairwise distance for these shuffled neurons, and repeated this process 25,000 times for each session to get the null distribution. We then calculated the percentile of the true within-session distance against each sessions' null distribution to obtain a p-value.

Contrastive learning: HDBSCAN cluster cutting After we obtained the 2-d embedding space of MEIs, we clustered this entire two-dimensional space using hierarchical density-based spatial clustering of applications with noise (HDBSCAN McInnes et al., 2017), the same clustering that was employed by the t-simCNE authors (Böhm et al., 2023). We searched over the parameter grid min-

samples, which indicates the minimum number of samples to be considered non-noise, $\in \{5, 15, \ldots, 145\}$ with the constraint min-samples $\leq$ min-cluster-size. The resulting clusterings will create one clustered group labelled -1, for MEIs that were unable to be clustered. We selected the parameters min-cluster-size = 200 and min-samples = 10 which resulted in the lowest number of unassigned MEIs. Using this approach, the MEIs of each neuron get assigned a label, with the possibility that not all MEIs of single neuron were assigned to the same cluster. We assigned a cluster ID to a neuron if more than half of its MEIs had the same cluster ID. Furthermore, we excluded neurons for which the more than half of the MEIs were in the unassigned category -1 of the HDBSCAN algorithm. Based on these two criteria, 828 out of 889 neurons were given a valid clustering ID.

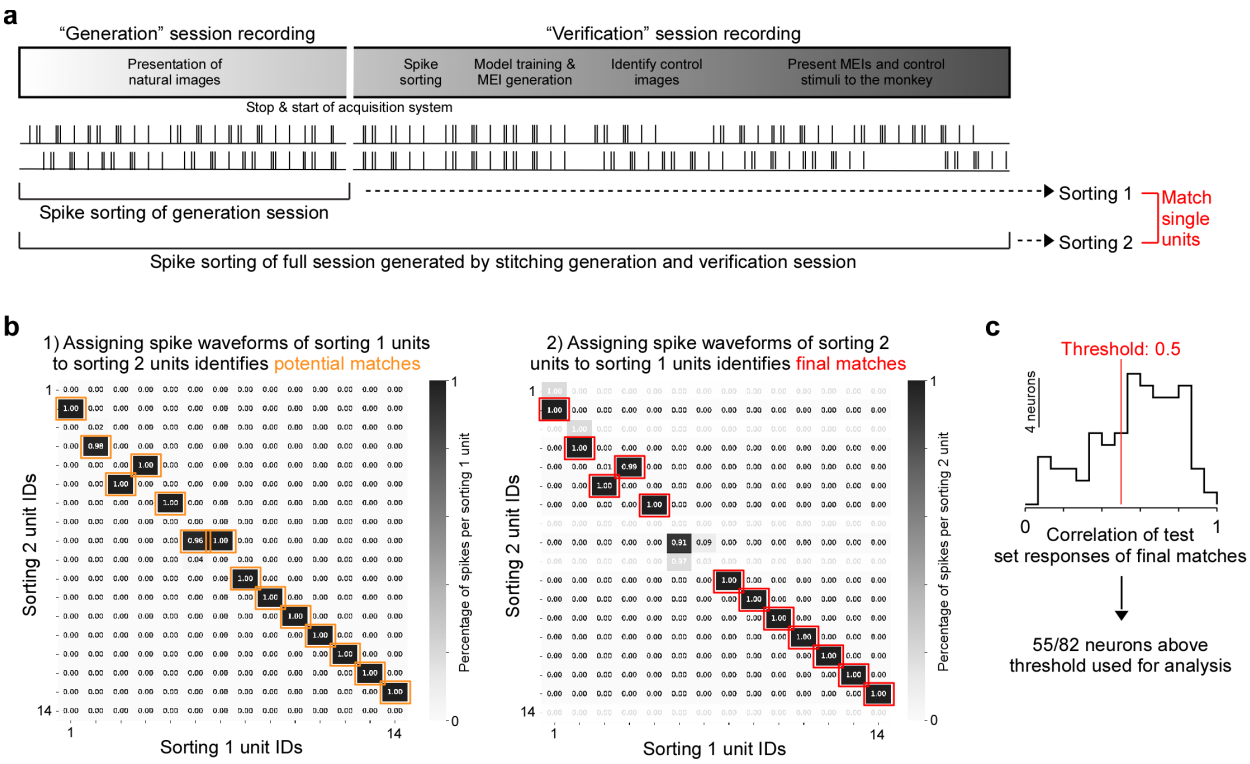
Representational similarity of MEIs As an independent way to verify the similarity of MEIs, we computed their representational similarity. For this purpose, we used the centered MEIs, and computed the responses of all 889 well-predicted neurons from our neural predictive model to each MEI. We controlled for the different RF positions of each *in silico* neuron by artificially centering the *in silico* neuron's RF to the center by setting the readout location of the Gaussian readout to the center of the core output. We used this approach for each model ensemble member individually and again took the model ensemble average as the predicted neuronal activity. Consequently, for each MEI, we obtained a response vector of length 889, which we then used for pairwise comparisons using cosine similarity. We then compared the average cosine similarity either within clusters or within recording sessions to the average similarity across sessions/clusters. To obtain a p-value for the similarity within vs. across clusters, we computed the similarity matrix 100 times. In each of the 100 runs, per neuron we drew at random one of the neurons MEIs and computed the population response vector for that MEI. We thus ended up with 100 similarity matrices, for which the pairwise comparisons were thus computed based on different MEIs. For each cluster, we then computed the average similarity across the 100 runs, and compared this similarity. We then took (1) the average cosine similarity within each cluster across the 100 runs, and (2) for each of the 100 runs, for each cluster we average the similarity to all other clusters. We report the p-value as percentile of the within-session similarity to the distribution of across-session similarity.

Data and code availability

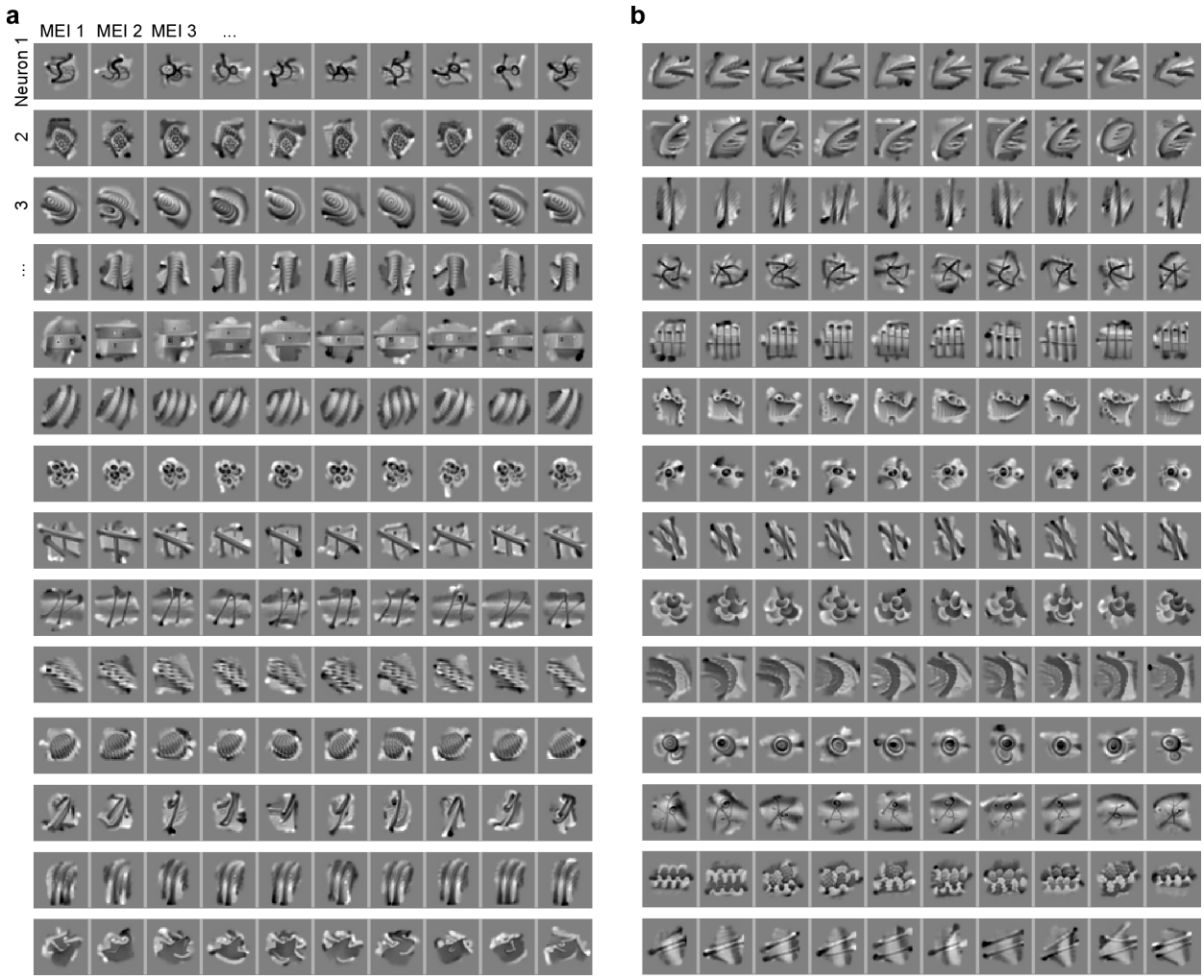
The analysis code will be publicly available in an online repository latest upon journal publication. Our coding framework uses Pytorch (Paszke et al., 2019), Numpy (Harris et al., 2020), scikit-image (Van der Walt et al., 2014), matplotlib (Hunter, 2007), seaborn (Waskom et al., 2017), DataJoint (Yatsenko et al., 2015). We also used the following custom libraries and code: `neuralpredictors` (https://github.com/sinzlab/neuralpredictors) for torch-based custom functions for model implementation, `nnfabrik` (https://github.com/sinzlab/nnfabrik) for automatic model training pipelines using DataJoint, `nnvision` (https://github.com/sinzlab/nnvision) for specific model definitions, analysis, and figures.

Supplementary Information

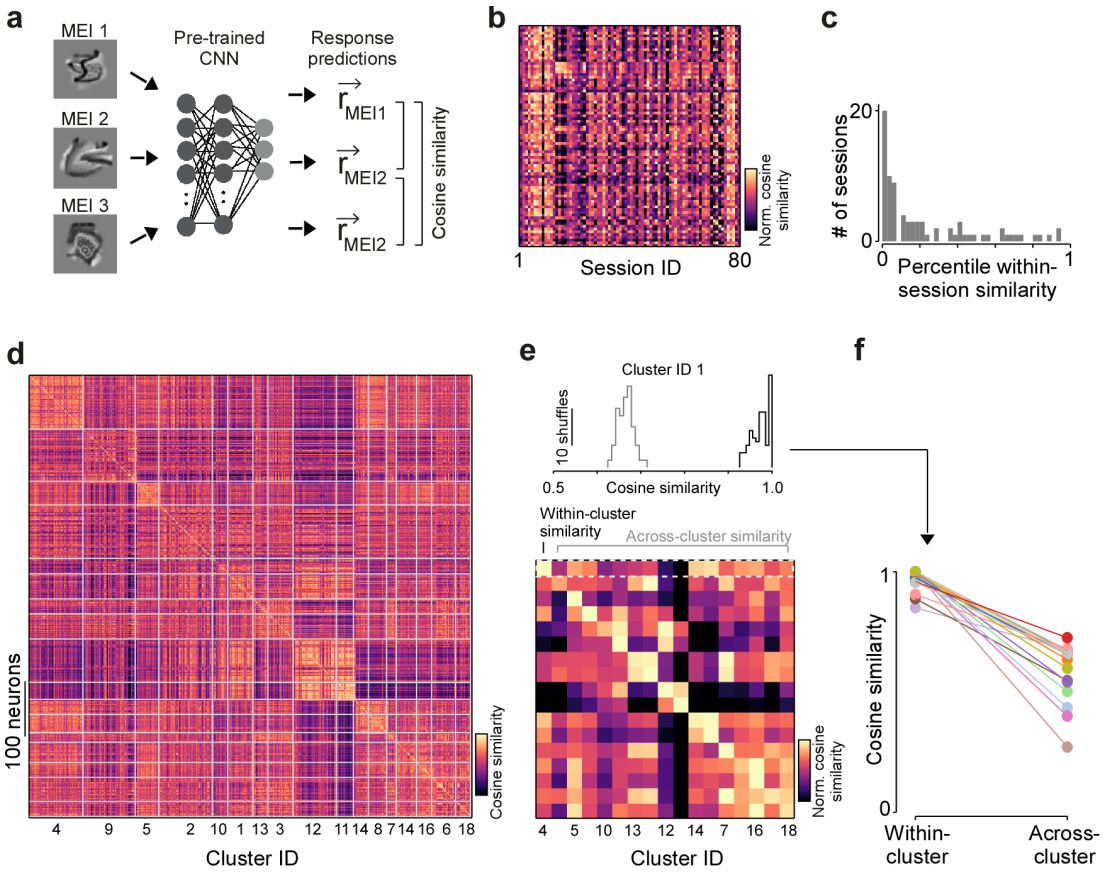
Supplemental Fig. 1 - Waveform and functional matching of single units across recordings
Supplemental Fig. 2 - Diverse model-derived stimuli of individual monkey V4 neurons
Supplemental Fig. 3 - Similarity of optimal stimuli in neuronal response space
Supplemental Fig. 4 - Overview of optimal stimuli of V4 response modes



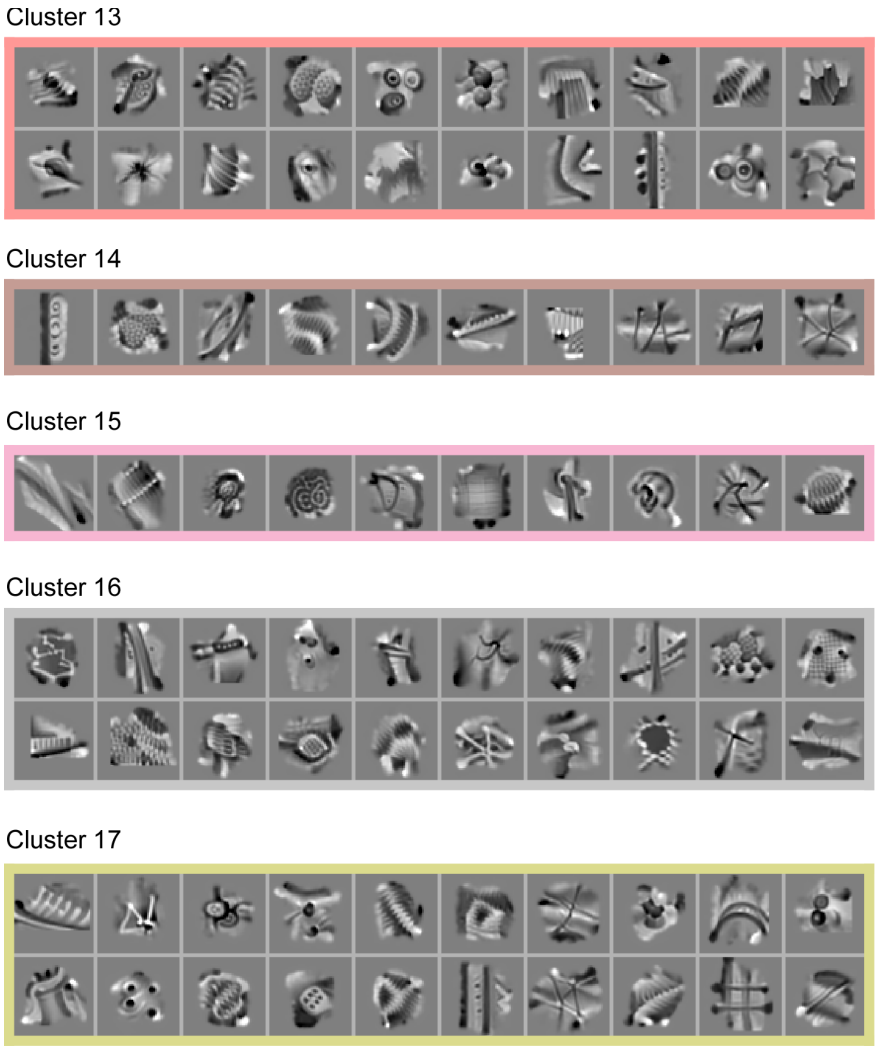
Supplemental Fig. 1. Spike waveform and functional matching of single units across recordings. a, Schematic illustrating spike sorting of the closed-loop experimental paradigm. The “generation session” is spike sorted directly after the recording, resulting in “Sorting 1”. This data is then used for model training and optimization of MEIs, which are presented back to the animal during the “verification session”. The verification session recording starts immediately after the generation session recording ends, to ensure a continuous monitoring of the recorded units over time. After the experiment, the generation and verification session recordings are concatenated (“full session”) and spike sorted, resulting in “Sorting 2”. **b,** Unit matching based on spike waveforms across Sorting 1 (generation session) and Sorting 2 (full session) for an example session. The left plot shows the percentage of spikes of the Sorting 1 units assigned to the units of Sorting 2. Units were assigned by passing the principal components of each spike, extracted using the Sorting 1 Gaussian Mixture model (GMM), to the Sorting 2 model. For a potential match (orange), at least 95% of the spikes of a single unit of Sorting 1 had to be assigned to an individual unit of Sorting 2. The right plot shows the percentage of spikes of Sorting 2 units assigned to the units of Sorting 1. For a final match (red), at least 95% of the spikes of a single Sorting 2 unit had to be assigned to the potential match of Sorting 1. **c,** Distribution of correlations of mean test set responses for all final matches. We only included matched units into the analysis, if their functional correlation was at least 0.5.



Supplemental Fig. 2. Diverse model-derived stimuli of individual monkey V4 neurons. a, For a set of 14 example neurons, we show 10 MEIs per neuron, generated from different starting points (random seeds) during the MEI optimization. The different MEIs exhibit the same visual feature but somewhat differ with respect to orientation, scale and position. **b,** Same as (a), for another set of 14 neurons.



Supplemental Fig. 3. Similarity of optimal stimuli in neuronal response space. a, MEIs of three example neurons. Right shows a schematic illustrating how we compared the similarity of MEIs using representational similarity. In brief, each MEI was presented to the trained CNN model that was used to produce the MEIs to obtain a response vector. The response vectors were then compared using cosine similarity. **b,** Mean cosine similarity of MEIs within a single recording session (diagonal) and across recording sessions for n=88 sessions, peak-normalized for each row. **c,** Distribution of percentiles of within-session similarity. For example, a percentile of 0.05 means that the MEI similarity within the session was larger than the MEI similarity to 95% of the other sessions. For n=30/55 sessions, the percentile was <0.05. **d,** Cosine similarity of MEIs of n=889 neurons, sorted based on cluster assignment (cf. Fig. 5). **e,** Mean cosine similarity of MEIs within a cluster (diagonal) and across clusters, peak normalized per row. The matrix depicts the mean across n=100 similarity matrices, each generated based on a different random selection of MEIs per neuron. Top shows distribution of cosine similarity within an example cluster (black) and the mean similarity to all other clusters (gray). **f,** Mean within-cluster and across-cluster similarity for all clusters.



Supplemental Fig. 4. Overview of optimal stimuli of V4 response modes. a, MEIs of example neurons for five response modes not shown in Fig. 5.

5.3 Retrospective on the SENSORIUM 2022 competition

Retrospective on the SENSORIUM 2022 competition

Konstantin F. Willeke[*,1,3], Paul G. Fahey[*,2], Mohammad Bashiri[1,3], Laura Hansel[3], Christoph Blessing[3], Konstantin-Klemens Lurz[1,3], Max F. Burg[1,3], Santiago A. Cadena[1,3], Zhiwei Ding[2], Kayla Ponder[2], Taliah Muhammad[2], Saumil S. Patel[2], Kaiwen Deng[4], Yuanfang Guan[4], Yiqin Zhu[5], Kaiwen Xiao[5], Xiao Han[5], Simone Azeglio[6,7], Ulisse Ferrari[6], Peter Neri[7], Olivier Marre[6], Adrian Roggenbach[8], Kirill Fedyanin[9], Kirill Vishniakov[10], Maxim Panov[9], Subash Prakash[1], Kishan Naik[1], Kantharaju Narayanappa[1], Alexander S. Ecker[1,3], Andreas S. Tolias[2], Fabian H. Sinz[1,3]

[1] University of Tübingen; [2] Baylor College of Medicine, Houston [3] University of Göttingen; [4] University of Michigan; [5] Tencent AI Lab; [6] Sorbonne University; [7] Ecole Normale Superieure; [8] University of Zurich; [9] Technology Innovation Institute; [10] Mohamed bin Zayed University of Artificial Intelligence

[*] equal contribution

KONSTANTIN-FRIEDRICH.WILLEKE@UNI-TUEBINGEN.DE, PAUL.FAHEY@BCM.EDU, SINZ@CS.UNI-GOETTINGEN.DE

Editors: Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

Abstract

The neural underpinning of the biological visual system is challenging to study experimentally, in particular as neuronal activity becomes increasingly nonlinear with respect to visual input. Artificial neural networks (ANNs) can serve a variety of goals for improving our understanding of this complex system, not only serving as predictive digital twins of sensory cortex for novel hypothesis generation *in silico*, but also incorporating bio-inspired architectural motifs to progressively bridge the gap between biological and machine vision. The mouse has recently emerged as a popular model system to study visual information processing, but no standardized large-scale benchmark to identify state-of-the-art models of the mouse visual system has been established. To fill this gap, we proposed the SENSORIUM benchmark competition. We collected a large-scale dataset from mouse primary visual cortex containing the responses of more than 28,000 neurons across seven mice stimulated with thousands of natural images, together with simultaneous behavioral measurements that include running speed, pupil dilation, and eye movements. The benchmark challenge ranked models based on predictive performance for neuronal responses on a held-out test set, and included two tracks for model input limited to either stimulus only (SENSORIUM) or stimulus plus behavior (SENSORIUM+). As a part of the NeurIPS 2022 competition track, we received 172 model submissions from 26 teams, with the winning teams improving our previous state-of-the-art model by more than 15%. Dataset access and infrastructure for evaluation of model predictions will remain online as an ongoing benchmark. We would like to see this as a starting point for regular challenges and data releases, and as a standard tool for measuring progress in large-scale neural system identification models of the mouse visual system and beyond.

Keywords

mouse visual cortex, system identification, neural prediction, natural images

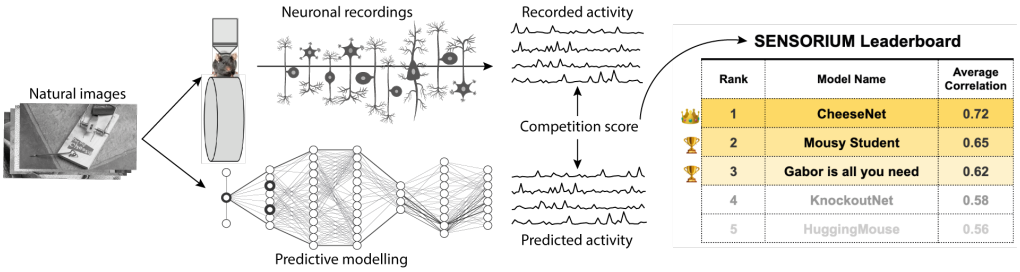


Figure 1: **A schematic illustration of the SENSORIUM competition.** We provide large-scale datasets of neuronal activity in the primary visual cortex of mice. Participants of the competition trained models on pairs of natural image stimuli and recorded neuronal activity.

Introduction

Understanding how the visual system processes visual information is a long standing goal in neuroscience. Neural system identification approaches this problem in a quantitative, testable, and reproducible way by building accurate predictive models of neural population activity in response to arbitrary input. If successful, these models can serve as functional digital twins for the visual cortex, allowing computational neuroscientists to derive new hypotheses about biological vision *in silico*, and enabling systems neuroscientists to test them *in vivo* (Walker et al., 2019; Ponce et al., 2019; Bashivan et al., 2019; Franke et al., 2022). In addition, highly predictive models are also relevant to machine learning researchers who use them to bridge the gap between biological and machine vision (Li et al., 2019; Safarani et al., 2021; Li et al., 2022; Sinz et al., 2019).

The work on predictive models of neural responses to visual inputs has a long history that includes simple linear-nonlinear (LN) models (Jones and Palmer, 1987; Heeger, 1992a,b), energy models (Adelson and Bergen, 1985), more general subunit/LN-LN models (Rust et al., 2005; Touryan et al., 2005; Schwartz et al., 2006; Vintch et al., 2015), and multi-layer neural network models (Zipser and Andersen, 1988; Lehky et al., 1992; Lau et al., 2002; Prenger et al., 2004). The deep learning revolution set new standards in prediction performance by leveraging task-optimized deep convolutional neural networks (CNNs) (Yamins et al., 2014; Cadieu et al., 2014; Cadena et al., 2019) and CNN-based architectures incorporating a shared encoding learned end-to-end for thousands of neurons (Antolík et al., 2016; Batty et al., 2017; McIntosh et al., 2016; Klindt et al., 2017; Kindel et al., 2019; Cadena et al., 2019; Burg et al., 2021; Lurz et al., 2021; Bashiri et al., 2021; Zhang et al., 2018; Cowley and Pillow, 2020; Ecker et al., 2018; Sinz et al., 2018; Walker et al., 2019; Franke et al., 2022).

The core idea of a neural system identification approach to improve our understanding of an underlying sensory area is that models that explain more of the stimulus-driven variability may capture nonlinearities that previous low-parametric models have missed (Carandini et al., 2005). Subsequent analysis of high performing models, paired with ongoing *in vivo* verification, can eventually yield more complete principles of brain computation. This motivates continually improving our models to explain as much as possible of the stimulus-driven variability and analyze these models to decipher principles of brain computations.

Standardized large-scale benchmarks are one approach to stimulate constructive competition between models compared on equal ground, leading to numerous incremental improvements that accumulate to substantial progress. In machine learning and computer vision, benchmarks have been an important driver of innovation in the last ten years. For instance, benchmarks such as the ImageNetChallenge (Russakovsky et al., 2015) helped jump start the revolution in artificial intelligence through deep learning. Similarly, neuroscience can benefit from more large-scale benchmarks to drive innovation and identify state-of-the-art models. This is especially true in the mouse visual cortex, which has recently emerged as a popular model system to study visual information processing, due to the wide range of available genetic and light imaging techniques for interrogating large-scale neural activity.

Existing neuroscience benchmarks vary substantially in the type of data, model organism, or goals of the contest (Schrimpf et al., 2018; Cichy et al., 2021; de Vries et al., 2019; Pei et al., 2021). For example, the **Brain-Score** benchmark (Schrimpf et al., 2018) ranks *task*-pretrained models that best match areas across primate visual ventral stream and other behavioral data, but do not provide neuronal training data. Instead, participants design objectives, learning procedures, network architectures, and input data that result in representations that are predictive of the withheld neural data. The **Algonauts** challenge (Cichy et al., 2021) competition ranks neural predictive models of human brain fMRI visual cortex activity in response to natural images and videos. Additionally, large data releases such as the mouse visual cortex dataset from **Allen Institute for Brain Science** (de Vries et al., 2019) are often not designed for a machine learning competition (consisting of only 118 natural images in addition to parametric stimuli and natural movies), and lack benchmark infrastructure for measuring predictive performance against a withheld test set. Lastly, the **Neural Latents** benchmark (Pei et al., 2021) also targets neuronal response prediction, but for cognitive, somatosensory, and motor areas with a focus on latent variable models.

To fill this gap, we created the `SENSORIUM` benchmark competition to facilitate the search for the best predictive model for mouse visual cortex. We collected a large-scale dataset from mouse primary visual cortex containing the responses of more than 28,000 neurons across seven mice stimulated with thousands of natural images, together with simultaneous behavioral measurements that include running speed, pupil dilation, and eye movements. Benchmark metrics will rank models based on predictive performance for neuronal responses on a held-out test set, and includes two tracks for model input limited to either stimulus only (`SENSORIUM`) or stimulus plus behavior (`SENSORIUM+`).

Our competition was part of the NeurIPS 2022 competition track, receiving 172 model submissions from 26 teams between May 20 and Oct 15, 2022. The winning teams substantially improved our previous state-of-the-art model in both competition tracks (`SENSORIUM`: +13.6%; `SENSORIUM+`: +18%). In this retrospective, we first describe the competition in detail, followed by the results of the competition, with descriptions from the winning teams outlining their approach. Finally, we reflect on the competition results as well as our lessons learned for future iterations.

The `SENSORIUM` Competition

The goal of the `SENSORIUM` 2022 competition and ongoing benchmark is to identify the best models for predicting sensory neural responses to arbitrary natural stimuli. At the start of

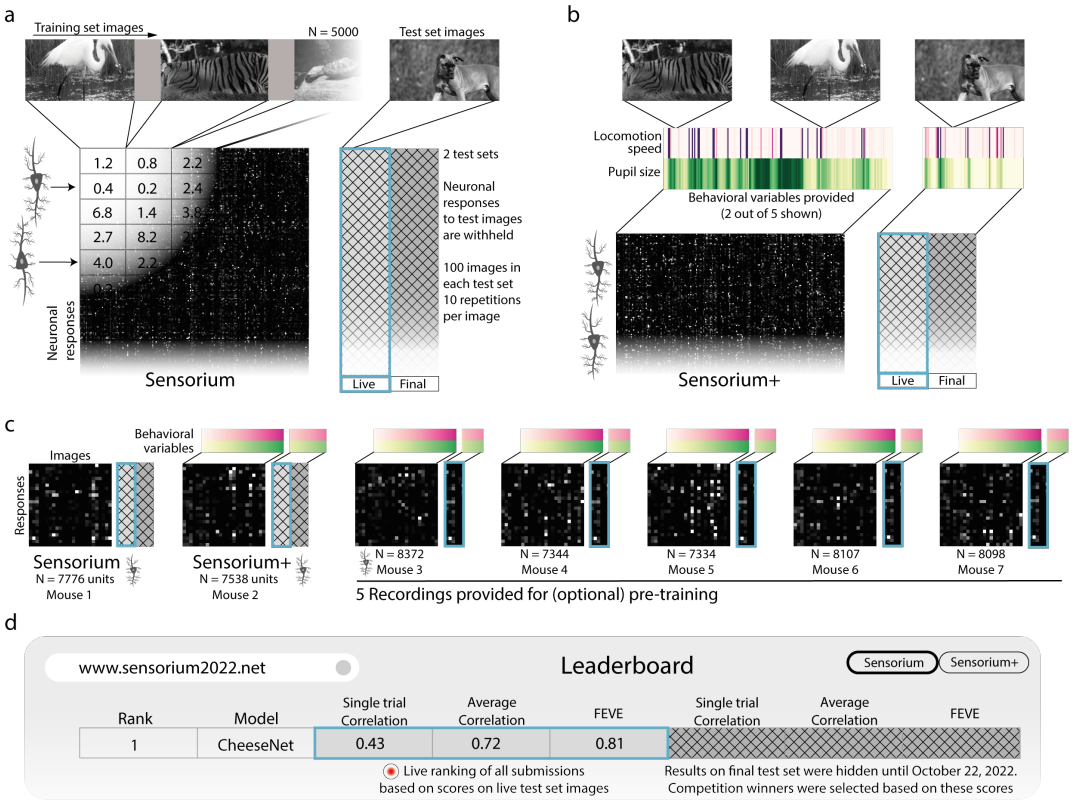


Figure 2: **Overview of the data and the competition structure.** **a**, Single recording from the `SENSORIUM` track. For ≈ 5000 training images, the neuronal activity of each neuron is provided. For 100 *live* and 100 *final* test set images shown 10 times each, neuronal responses are withheld. **b**, `SENSORIUM+` track is the same as (**a**) but behavioral variables are available. **c**, Overview of seven dataset recordings. Five *pre-training* recordings are not part of the competition evaluation, but can be used to improve model performance. In the *public test* set, 100 images are shared with the *live test* set (blue frame), but neuronal responses are provided. **d**, *Live test* scores are displayed on the live leaderboard, while *final test* set scores were only revealed after the submissions closed.

the competition, a training dataset for refining model performance was publicly released (Fig. 2). For two animals, the neuronal responses to a set of competition test set images were permanently withheld. The competition test set images are divided into two exclusive groups: *live* and *final test*. Performance metrics computed on the *live test* images are used to maintain a public leaderboard on our website, while the performance metrics on the *final test* images were only used to score entries after the submission period has ended (Fig. 2d). By separating the *live test* and *final test* set performance metrics, we were able to provide feedback on *live test* set performance to participants wishing to submit updated predictions

(up to one submission per day), while protecting the *final test* set from overfitting over multiple submissions.

The competition has two tracks, `SENSORIUM` and `SENSORIUM+` , predicting two datasets with the same stimuli, but from two different animals and with differing model inputs.

`SENSORIUM`. In the first challenge, participants predict the average neuronal activity of 7,776 neurons in response to 10 repetitions of 200 unique natural images of our competition *live test* and *final test* image sets. The data provided for the test set includes the natural image stimuli but not the behavioral variables (Fig. 2a). Thus, this challenge focuses on stimulus-driven responses, treating other correlates of neural variability, such as behavioral state, as noise. This track resembles most of the current efforts in the community (Schrimpf et al., 2018) to identify stimulus-response functions without additional information about the brain and behavioral state.

`SENSORIUM+` In the second challenge, participants predict the single-trial neuronal activity of 7,538 neurons in response to 200 unique natural images of our competition *live test* and *final test* image sets. In this case, both the natural image stimuli and the accompanying behavioral variables are provided (Fig. 2b). As a significant part of response variability correlates with the animal's behavior and internal brain state (Niell and Stryker, 2010; Reimer et al., 2014; Stringer et al., 2019), their inclusion can result in models that capture single trial neural responses more accurately (Bashiri et al., 2021; Franke et al., 2022).

Data

The competition dataset was designed to compare neural predictive models that capture neuronal responses $\mathbf{r} \in \mathbb{R}^n$ of n neurons as a function $\mathbf{f}_\theta(\mathbf{x})$ of either only natural image stimuli $\mathbf{x} \in \mathbb{R}^{h \times w}$ (image **h**eight,**w**idth), or as a function $\mathbf{f}_\theta(\mathbf{x}, \mathbf{b})$ of both natural image stimuli and behavioral variables $\mathbf{b} \in \mathbb{R}^k$. We provide $k = 5$ variables: locomotion speed, pupil size, instantaneous change of pupil size (second order central difference), and horizontal and vertical eye position. See Fig. 2 and (Willeke et al., 2022) for an overview of the dataset.

Natural images. Natural images from ImageNet (Russakovsky et al., 2015) were cropped to fit a monitor with 16:9 aspect ratio, converted to gray scale, and presented to mice for 500 ms, preceded by a blank screen period between 300 and 500 ms (Fig. 2a,b)

Neuronal responses. We recorded the response of excitatory neurons in layer 2/3 of the right primary visual cortex in awake, head-fixed, behaving mice using calcium imaging. Activity was extracted and accumulated 50 - 550 ms after each stimulus onset.

Behavioral variables. During imaging, mice were head-mounted over a cylindrical treadmill, and an eye camera captured changes in the pupil position and dilation. Behavioral variables were similarly extracted and accumulated 50 - 550 ms after each stimulus onset.

Dataset. Our complete data corpus comprises seven recordings in seven animals (Fig. 2c), including the neuronal activity of over 28,000 neurons to 25,200 unique images, with 6,000–7,000 image presentations per recording (see (Willeke et al., 2022) for details). We report a conservative neuron count estimate, due to multiple segmented units from single neurons appearing in multiple, densely placed calcium imaging recording planes. Fig. 2c shows the uncorrected number of 54,569 units.

Five of the seven recordings, which we refer to as *pre-training recordings* (Fig. 2c, right), are provided solely for training and model generalization, and are not included in the

competition performance metrics. They contain 5,000 single presentations of natural images, randomly intermixed with 10 repetitions of 100 natural images, and all corresponding neuronal responses. The 100 repeated images and responses serve as a *public test* set.

In the two remaining *competition recordings* (Fig. 2c, left), the mice were also presented with 5,000 single presentations of training stimuli as well as the *public test* images. However, during the contest, we withheld the responses to the *public test* images and use them for the live leaderboard. We thus refer to these images as *live test* set. Furthermore, the competition recordings contain 10 repetitions of 100 additional natural *test* images that were randomly intermixed during the experiment. These *test* images are only present in the two competition recordings. The responses to these images were also withheld and used to determine the winner of the competition after submissions are closed (Fig. 2a,b). We refer to these images as our *final test* set. By providing both *live* and *final test* scoring, participants receive the benefit of iterative feedback while avoiding overfitting on the final scoring metrics.

In our first competition track (`SENSORIUM`, Fig. 2a), we withhold the behavioral variables, such that only the natural images can be used to predict the neuronal responses. For the other competition track (`SENSORIUM+`, Fig. 2b), as well as the five pre-training recordings, we are releasing all the behavioral variables. Lastly, we released the anatomical locations of the recorded neurons for all datasets. The complete corpus of data is available to download at https://sinzlab.org/sensorium2022.html.

Performance Metrics

Across the two benchmark tracks, three metrics of predictive accuracy are automatically and independently computed for the 100 *live test* set images and 100 *final test* set images, for which ground-truth neuronal responses are withheld (see (Willeke et al., 2022) for details)

Correlation to Average We calculate the *correlation to average* of 100 model predictions to the withheld, observed mean neural response across 10 repeated presentations of the same stimulus. This metric is computed for both the `SENSORIUM` and `SENSORIUM+` tracks to facilitate comparison. Correlation to average on the *final test* set served as the ultimate ranking score in the `SENSORIUM` track to determine competition winners.

Fraction of Explainable Variance Explained (FEVE) While correlation to average is a common metric, it is insensitive to affine transformations of either the neuronal response or predictions. This metric computes the ratio between the variance explained by the model and the explainable variance in the neural responses (Cadena et al., 2019). The explainable variance accounts for only the stimulus-driven variance and ignores the trial-to-trial variability in responses. This metric is computed for `SENSORIUM` but not `SENSORIUM+` , due to the lack of repeated trials with identical behavior fluctuations necessary to estimate explainable variance. For numerical stability, we compute the FEVE only for neurons with an explainable variance larger than 15% (N=4319 for `SENSORIUM` and N=4548 for `SENSORIUM+`).

Single Trial Correlation Lastly, to measure how well models account for trial-to-trial variations we compute the *single trial correlation* between predictions and single trial neuronal responses, without averaging across repeats. This metric is computed for both the `SENSORIUM` and `SENSORIUM+` tracks to facilitate comparison. Single trial correlation on the *final test* set served as the ultimate ranking score in the `SENSORIUM+` track to determine competition winners.

	Competition results					
	Live test set			Final test set		
	Single trial Correlation	Average Correlation	FEVE	Single trial Correlation	Average Correlation	FEVE
SENSORIUM						
LN Baseline	0.197	**0.363**	0.222	0.207	**0.377 (-28.6%)**	0.232
CNN Baseline	0.274	**0.513**	0.433	0.287	**0.528 (0%)**	0.439
#3: Azeglio et al.	0.307	**0.580**	0.549	0.319	**0.587 (+11.1%)**	0.516
#2: Zhu et al.	0.314	**0.589**	0.512	0.325	**0.598 (+13.2%)**	0.503
#1: Deng et al.	0.316	**0.594**	0.576	0.325	**0.600 (+13.6%)**	0.559
SENSORIUM+						
LN Baseline	**0.257**	0.373	-	**0.266 (-30.7%)**	0.385	-
CNN Baseline	**0.374**	0.571	-	**0.384 (0%)**	0.578	-
#3: Fedyanin et al.	**0.397**	0.605	-	**0.410 (+6.7%)**	0.618	-
#2: Deng et al.	**0.428**	0.643	-	**0.437 (+13.8%)**	0.650	-
#1: Roggenbach	**0.444**	0.625	-	**0.453 (+18.0%)**	0.632	-

Table 1: **Performance of the competition winners of both tracks.** For the final test set, the improvement over the CNN baseline model is shown in percentages.

Baseline

To establish baselines, we trained a simple linear-nonlinear model (LN Baseline) as well as a state-of-the-art convolutional neural network (CNN baseline, Lurz et al., 2021) model for both competitions. For each baseline, we trained a single model (based on one random seed) on only the training data from each competition track (for details, see Willeke et al. (2022)).

Results and Participation

During the four month submission period, 26 teams submitted a total of 172 models (SENSORIUM: 124, SENSORIUM+: 78). To our delight, our state-of-the-art baseline models were outperformed in both tracks by more than 15% (Table 1). We invited the winning teams of each track to describe both their successful and fruitless approaches.

SENSORIUM Rank 1: Deng & Guan

Our winning submission in the SENSORIUM and the SENSORIUM+ track only had minor changes to the SOTA model from the CNN Baseline (Lurz et al., 2021). In the core, we added a convolutional layer whose kernel size and strides were 4 before the first layer to replace the scaling operation in the original model. We set the channel number as 32 for the scaling

layer and increased the filtering numbers of the following 4 layers to 128, 256, 256, and 256. We did not change the readout and the other hyperparameters.

The key features of our winning solution were: for the SENSORIUM track, 1) we added the positions of the objects in the images as additional channels to the inputs; 2) we trained multiple models by using different train-validation splits and averaged the predictions of these models; 3) for the SENSORIUM+ track, we utilize the pretraining datasets in an ensemble way. We trained multiple models using different pretraining datasets and then averaged the predictions. The idea of adding object positions came from the contribution of "pupil behavior" in the SENSORIUM+ track. We hypothesized that the objects would catch the attention and their positions would reflect the pupil's behaviors. The object detection model was trained on the ILSVRC 2017 dataset. We sampled 250 from each category, converted them to gray-scale images, and fine-tuned the PyTorch YOLOv5 large model (Jocher et al., 2022). To label the objects in the competition data, we set the NMS confidence threshold as 0.05 and the IOU threshold as 0.5. The parameter image size was 256. The bounding boxes were merged into a larger one for each image, and the position was a vector (x, y, width, height) in the YOLO format. We gave the vector (0.5, 0.5, 1, 1) for the images without bounding boxes. Finally, there were 6 channels in our SENSORIUM model inputs: the image normalized by the provided mean and standard deviation, the centered first-channel image, and the object positions. For SENSORIUM+ , we removed the object positions. Our ablation studies as well as a description of unsuccessful attempts can be found in the appendix.

SENSORIUM Rank 2: Zhu, Xiao, & Han

Architecture. Our model was based on the CNN Baseline (Lurz et al., 2021) and is initialized in the same way. We used the shared core approach so that we can pre-train our model with data from all seven mice. The feature map in first layer looked like the gradient map of input gray image. Therefore, we use sobel operator to pre-extract the x, y-axis gradient map of the image, and input it into the model together with the image.

Training. We only used the data provided by the competition. To make full use of them, our model was trained in a two-stage manner. First, we update all parameters with all data until the validation score no longer improves. Second, we fine-tuned the core and readout networks of the target mice with smaller learning rate. We utilize self-distillation to generate more robust model. Specifically, we utilize the training set data to train a teacher model. Then, we use teacher model to predict neural responses for all data, and mix new image-response pairs data with real data. The student model with better performance can be obtained by being trained with mixed data. Our optimization object is to minimize the joint loss $Loss = L_{poisson} + \lambda L_{corr}$, where $L_{possion}$ donate Poisson loss, $L_{corr} = 1 - Corr(r_{av}, o_{av})$ is correlate loss, λ is a balance factor. We fix $\lambda = 1000$ in all experiments.

Inference. Model ensembling always works at inference time. However, directly ensembling multiple models will have large redundancy. Therefore, we designed a greedy ensemble strategy to achieve their best outcome. We trained more than 100 models with different seeds and number of core convolution layers. All of them were used to form our ensemble model. We then test each model one by one. If the validation score improves when we block any model, then this model will be removed. We repeat this process 3 times in total.

`SENSORIUM` **Rank 3: Azeglio, Ferrari, Neri, & Marre**

Our approach entailed building upon the baseline CNN model developed by the organizers (Lurz et al., 2021). This model comprises of a nonlinear core and a readout layer informed by biological retinotopy. Our focus was on the front-end modules situated between the input and the core, specifically two components. The first component was Scattering Networks, as introduced by Bruna and Mallat (2013), which enforce geometric constraints. These networks have two notable features: 1) their representations are both translation invariant and robust to minor deformations, and 2) deeper models can be achieved without the need for additional parameters as all parameters are fixed. The second component was VOneBlock, developed by Dapello and collaborators (Dapello et al., 2020), which implements biologically grounded constraints through a linear-nonlinear Poisson model composed of a Gabor filter bank with fixed weights, simple and complex cell nonlinearities, and neuronal stochasticity (independent Gaussian noise).

1, 2, 3... Ensembling! Based on the front-ends previously discussed, we decided to implement four distinct models: 1) Scattering front-end, baseline CNN core, and Gaussian readout; 2) VOneBlock front-end, baseline CNN core, and Gaussian readout; 3) Scattering front-end, Squeeze and Excitation CNN core (as described by Hu et al. (2018)) and Gaussian readout; and 4) VOneBlock front-end, Squeeze and Excitation CNN core, and Gaussian readout. Following training and evaluation of the various models, we combined them into an ensemble model by taking an average of their predictions. The performance of individual cores can be found in Table 4 in the appendix, along with a discussion of unsuccessful approaches. Code is available at `https://github.com/sazio/sensorium`.

`SENSORIUM+` **Rank 1: A. Roggenbach**

In addition to the visual input, neural activity in the visual cortex also depends on the ongoing neural activity and the behavioral state (Stringer et al., 2019; Arieli et al., 1996; Syeda et al., 2022). The key addition of the model is to account for these non-sensory effects based on past neural activity. This is implemented by combining the output of the provided baseline model with a modulator network which consists of three parts.

First, a ten-dimensional network state of the activity of all neurons in the last known timestep is extracted. This low-dimensional projection is calculated by passing the neural activity through a reduced rank auto-regression network for the next time step (to remove the stimulus information which is not predictive for the next time step) and calculating a non-negative matrix factorization on this output. These features are linearly combined for each neuron in the modulator network. Second, the activity history of each neuron is passed through a filter bank with varying temporal kernels, resulting in five history features per neuron which are linearly combined. Third, the output of the provided core model is added to the previously described network state and history output, passed through a ReLU+1 non-linearity and multiplied with a scalar gain. This learned gain regressor is encouraged to be smooth by reading out the regressor with a half-normal distribution kernel and by applying a L2-penalty on the temporal difference.

Additionally, the pupil and running regressors for the core module are normalized between 0 and 1 and hyperparameters are slightly adjusted. Ensembling of five models with different

seeds and train/validation splits further increased the performance. Trained models and code are available at https://github.com/AdrianRoggenbach/adrian_sensorium.

SENSORIUM+ Rank 2: Deng & Guan

The model is identical to the rank 1 model of SENSORIUM.

SENSORIUM+ Rank 3: Fedyanin, Vishniakov, & Panov

We explored several directions to improve the CNN baseline model with varying success. Namely, we tried improving the feature extractor (core), simplifying a readout layer, incorporating geometric and color data augmentations, self-supervised pretraining, and model ensembling. In this section, we report validation set results for recording 27204-5-13.

Core Design Improvement. Initially, we discovered that full-resolution images gave worse results than images that were downscaled by the factor of 0.25. In our investigation, we found that size of the feature map produced by the core has much more pronounced effect on the final result than the input image size. We used this information to design a deeper encoder based on ResNet (He et al., 2016). We made the following changes to baseline ResNet-18 model: changing the number of layers from 18 to 9, changing stride of several layers from 2 to 1, replacing ReLU with ELU, and adding Dropout layers.

Table 2: Model performance with altered core module. Performance reported for a single recording.

	SENSORIUM	SENSORIUM+
Baseline (Lurz et al., 2021)	0.296	0.378
ResNet-18 (He et al., 2016)	0.236	0.310
ResNet-18, stride=1	0.288	0.284
ResNet-9	0.286	0.379
+ELU	0.300	0.400
+Dropout	0.311	0.409

Ensembling. We trained 20 basic models on both SENSORIUM and SENSORIUM+, using the different weight initializations, which gave improvement of +**2.4%** on SENSORIUM and +**3.2%** on SENSORIUM+ test sets.

Reflections

Competition results

It is noteworthy that the majority of the winning teams relied heavily on our CNN baseline model architecture, which remained mostly unchanged. Common successful strategies included:

- using additional transformations of the input data or temporal dependencies
- pre-training the core on the extra datasets, with fine-tuning on the respective competition dataset
- creating large model ensembles, together with improvements in model training
- adjustments of the core-architecture

These changes led to substantial gains in model performance, larger than 15% in both competition tracks. While we consider the improvements in model accuracy with these strategies as impressive (especially given the limited four month competition period), we look

forward to modeling approaches that differ more substantially from our previous state-of-the-art model. On that note, a recent publication by Li et al. (2023) utilized the benchmark data while describing an entirely novel modeling approach based on the Vision Transformer (Dosovitskiy et al., 2020).

Lessons learned for future iterations

We hope that this benchmark infrastructure serves as a catalyst for both computational neuroscientists and machine learning practitioners to advance the field of neuro-predictive modeling. Our broader goal is to continue to populate the underlying benchmark infrastructure with future iterations of dataset releases, new challenges, and additional metrics.

As is the case for benchmarks in general, by converging in this first iteration on a specific dataset, task, and evaluation metric in order to facilitate constructive comparison, `SENSORIUM` 2022 also becomes limited to the scope of those choices. In particular, we opted for simplicity for the first competition hosted on our platform in order to appeal to a broader audience across the computational neuroscience and machine learning communities. *A priori*, it is not clear how well the best performing models of this competition would transfer to a broader or more naturalistic setting where stimuli could be out of domain for the models. Having established our benchmarking framework, possible directions to extend in future challenges are:

- including cortical layers beyond L2/3 and areas in mouse visual cortex beyond V1
- replacing static image stimuli with dynamic movie stimuli in order to better capture the temporal evolution of representation and/or simulation
- replacing grayscale stimuli with coverage of UV- and green-sensitive cone photoreceptors
- increasing the number of animals and recordings in the test set beyond one per track to emphasize generalization across animals and brain states
- moving beyond passive stimulus viewing by incorporating a decision making paradigm
- including different or multiple sensory domains (e.g., auditory, olfactory, somatosensory, etc) and motor areas
- recording neural responses with different techniques (e.g., electrophysiology) that emphasize different population sizes and spatiotemporal resolution
- recording neural responses in different animal models, such as non-human primates.
- inverting model architecture to reconstruct visual input from neural responses.

We believe that predictive models have become an important tool for neuroscience research. In our view, systematically benchmarking and improving these models along with the development of accurate metrics will be of great benefit to neuroscience as a whole. We therefore invite the research community to join the benchmarking effort by continuing to participate in the benchmark, and by contributing new datasets and metrics to our benchmarking system. We would like to cast the challenge of understanding information processing in the brain as a joint endeavor in which we engage together as a whole community, iteratively re-defining what is the state-of-the-art in predicting neural activity and leveraging models to pursue the question of how the brain makes sense of the world.

Acknowledgments

KKL is funded by the German Federal Ministry of Education and Research through the Tübingen AI Center (FKZ: 01IS18039A). This work was supported by an AWS Machine Learning research award to FHS. MB and SAC were supported by the International Max Planck Research School for Intelligent Systems. This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101041669).

This research was supported by National Institutes of Health (NIH) via National Eye Institute (NEI) grant RO1-EY026927, NEI grant T32-EY002520, National Institute of Mental Health (NIMH) and National Institute of Neurological Disorders and Stroke (NINDS) grant U19-MH114830, and NINDS grant U01-NS113294. This research was also supported by National Science Foundation (NSF) NeuroNex grant 1707400. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NEI, NIMH, NINDS, or NSF.

This research was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract no. D16PC00003, and with funding from the Defense Advanced Research Projects Agency (DARPA), Contract No. N66001-19-C-4020. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. 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, DARPA, or the US Government.

Deng & Guan report that this study was also supported by NIH Grants: NIH/NIGMS R35GM133346.

AR acknowledges the use of Fenix Infrastructure resources, which are partially funded from the European Union's Horizon 2020 research and innovation programme through the ICEI project under the grant agreement No. 800858.

References

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

Amos Arieli, Alexander Sterkin, Amiram Grinvald, and Ad Aertsen. Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. Science, 273 (5283):1868–1871, sep 1996. doi: 10.1126/science.273.5283.1868. URL https://doi.org/10.1126%2Fscience.273.5283.1868.

Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, and Fabian Sinz. A flow-based latent state generative model of neural population responses to natural images. Adv. Neural Inf. Process. Syst., 34, December 2021.

Pouya Bashivan, Kohitij Kar, and James 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.

Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan Litke, EJ Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion cell responses. In International Conference on Learning Representations, 2017.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. IEEE transactions on pattern analysis and machine intelligence, 35(8):1872–1886, 2013.

Max F. Burg, Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Andreas S. Tolias, Matthias Bethge, and Alexander 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.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. PLOS Computational Biology, 15(4):e1006897, April 2019. doi: 10.1371/journal.pcbi.1006897.

Charles F Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput. Biol., 10(12): e1003963, 2014.

Matteo Carandini, Jonathan B. Demb, Valerio Mante, David J. Tolhurst, Yang Dan, Bruno A. Olshausen, Jack L. Gallant, and Nicole C. Rust. Do we know what the early visual system does? J. Neurosci., 25(46):10577–10597, November 2005. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3726-05.2005.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

R. M. Cichy, K. Dwivedi, B. Lahner, A. Lascelles, P. Iamshchinina, M. Graumann, A. Andonian, N. A. R. Murty, K. Kay, G. Roig, and A. Oliva. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion, 2021. URL https://arxiv.org/abs/2104.13714.

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

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. Advances in Neural Information Processing Systems, 33:13073–13087, 2020.

Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, Peter A. Groblewski, Gabriel K. Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, Kate Roll, Marina Garrett, Tom Keenan, Leonard Kuan, Stefan Mihalas, Shawn Olsen, Carol Thompson, Wayne Wakeman, Jack Waters, Derric Williams, Chris Barber, Nathan Berbesque, Brandon Blanchard, Nicholas Bowles, Shiella D. Caldejon, Linzy Casal, Andrew Cho, Sissy Cross, Chinh Dang, Tim Dolbeare, Melise Edwards, John Galbraith, Nathalie Gaudreault, Terri L. Gilbert, Fiona Griffin, Perry Hargrave, Robert Howard, Lawrence Huang, Sean Jewell, Nika Keller, Ulf Knoblich, Josh D. Larkin, Rachael Larsen, Chris Lau, Eric Lee, Felix Lee, Arielle Leon, Lu Li, Fuhui Long, Jennifer Luviano, Kyla Mace, Thuyanh Nguyen, Jed Perkins, Miranda Robertson, Sam Seid, Eric Shea-Brown, Jianghong Shi, Nathan Sjoquist, Cliff Slaughterbeck, David Sullivan, Ryan Valenza, Casey White, Ali Williford, Daniela M. Witten, Jun Zhuang, Hongkui Zeng, Colin Farrell, Lydia Ng, Amy Bernard, John W. Phillips, R. Clay Reid, and Christof Koch. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nature Neuroscience, 23(1):138–151, December 2019. doi: 10.1038/s41593-019-0550-9. URL https://doi.org/10.1038/s41593-019-0550-9.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

Alexander S. Ecker, Fabian H. Sinz, Emmanouil Froudarakis, Paul G. Fahey, Santiago A. Cadena, Edgar Y. Walker, Erick Cobos, Jacob Reimer, Andreas S. Tolias, and Matthias Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex, 2018.

Katrin Franke, Konstantin F. Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H. Sinz, and Andreas S. Tolias. State-dependent pupil dilation rapidly shifts visual feature selectivity. Nature, 610(7930):128–134, September 2022. doi: 10.1038/s41586-022-05270-3. URL https://doi.org/10.1038/s41586-022-05270-3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

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.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, Imyhxy, , Lorna, Zeng Yifu, Colin Wong, Abhiram

V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, Tkianai, YxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation, 2022. URL https://zenodo.org/record/7347926.

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, December 1987.

William F Kindel, Elijah D Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. Journal of vision, 19(4):29–29, 2019.

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.

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.

SR Lehky, TJ Sejnowski, and R Desimone. Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. The Journal of Neuroscience, 12(9):3568–3581, September 1992. doi: 10.1523/jneurosci.12-09-03568.1992. URL https://doi.org/10.1523/jneurosci.12-09-03568.1992.

Bryan M. Li, Isabel M. Cornacchia, Nathalie L. Rochefort, and Arno Onken. V1t: large-scale mouse v1 response prediction using a vision transformer, 2023. URL https://arxiv.org/abs/2302.03023.

Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. Advances in neural information processing systems, 32, 2019.

Zhe Li, Josue Ortega Caro, Evgenia Rusak, Wieland Brendel, Matthias Bethge, Fabio Anselmi, Ankit B Patel, Andreas S Tolias, and Xaq Pitkow. Robust deep learning object recognition models rely on low frequency information in natural images. bioRxiv, page 2022.01.31.478509, February 2022.

Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay K Jagadish, Eric Wang, Edgar Y Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S Tolias, Alexander S Ecker, and Fabian 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, October 2021.

Lane T McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A Baccus. Deep learning models of the retinal response to natural scenes. Adv. Neural Inf. Process. Syst., 29(Nips):1369–1377, 2016.

Rafael Navarro, Pablo Artal, and David R Williams. Modulation transfer of the human eye as a function of retinal eccentricity. JOSA A, 10(2):201–212, 1993.

Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. Neuron, 65(4):472–479, 2010.

D Pamplona, J Triesch, and CA Rothkopf. Power spectra of the natural input to the visual system. Vision research, 83:66–75, 2013.

Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity, 2021. URL https://arxiv.org/abs/2109.04463.

Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell, 177(4):999–1009.e10, 2019.

Ryan Prenger, Michael C-K Wu, Stephen V David, and Jack L Gallant. Nonlinear V1 responses to natural scenes revealed by neural network analysis. Neural Netw., 17(5-6): 663–679, 2004.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

Jacob Reimer, Emmanouil Froudarakis, Cathryn R Cadwell, Dimitri Yatsenko, George H Denfield, and Andreas S Tolias. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. Neuron, 84(2):355–362, October 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis., 115(3): 211–252, December 2015.

Nicole C Rust, Odelia Schwartz, J Anthony Movshon, and Eero P Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. Neuron, 46(6):945–956, 2005.

Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. Adv. Neural Inf. Process. Syst., 34:739–751, December 2021.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? BioRxiv, page 407007, 2018.

Odelia Schwartz, Jonathan W Pillow, Nicole C Rust, and Eero P Simoncelli. Spike-triggered neural characterization. J. Vis., 6(4):484–507, July 2006.

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.

Fabian H. Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias. Engineering a less artificial intelligence. Neuron, 103(6):967–979, September 2019. doi: 10.1016/j.neuron.2019.08.034. URL https://doi.org/10.1016/j.neuron.2019.08.034.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. Science, 364(6437), 2019.

Atika Syeda, Lin Zhong, Renee Tung, Will Long, Marius Pachitariu, and Carsen Stringer. Facemap: a framework for modeling neural activity based on orofacial tracking. bioRxiv, pages 2022–11, 2022.

Jon Touryan, Gidon Felsen, and Yang 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.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. Nat. Neurosci., 22(12):2060–2065, December 2019.

Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Pede, Max F. Burg, Christoph Blessing, Santiago A. Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The sensorium competition on predicting large-scale mouse primary visual cortex activity, 2022. URL https://arxiv.org/abs/2206.08666.

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, 111(23):8619–8624, May 2014. doi: 10.1073/pnas.1403112111. URL https://doi.org/10.1073/pnas.1403112111.

Yimeng Zhang, T-S Tai Sing Lee, Ming Li, Fang Liu, Shiming Tang, Tai Sing, Lee Ming, Li Fang, Liu Shiming, T-S Tai Sing Lee, Ming Li, Fang Liu, and Shiming Tang. Convolutional neural network models of V1 responses to complex patterns. J. Comput. Neurosci., pages 1–22, 2018.

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

Appendix A. Additional material from the winning teams

SENSORIUM Rank 1: Deng & Guan

The ablation study results on the SENSORIUM model list in Table 3. They are the evaluations of the 5 datasets for pre-training. All of our modifications contributed to the performance improvement of the baseline model. We also found some unexpected contributions in these methods when comparing the ensemble model and the single model. The object positions significantly improved the performances in the single model, but only had minor effects in the ensemble model.

Model ablation study		
Method	Model type	Score
Best model	ensemble	0.6254±0.0235
Remove object bounding boxes	ensemble	0.6234±0.0239
Fewer filters	ensemble	0.6068±0.0236
Replace Conv-scale	ensemble	0.6007±0.0225
Remove centered image	ensemble	0.5964±0.0227
Best model	single	0.5895±0.0238
Remove object bounding boxes	single	0.5794±0.0241
Fewer filters	single	0.5745±0.0242
Replace Conv-scale	single	0.5706±0.0248
Remove centered image	single	0.5646±0.0225

Table 3: Ablation study for the core used in the SENSORIUM track.

We also tried several other methods to utilize the information of the objects but failed. These methods included constructing a matrix with the same shape as the image and assigning 0 or 1, or different weights between 0 and 1, for the elements outside and inside the bounding box. We tried adding this matrix as an additional channel and multiplying it on the original images to clip the image, but none of these experiments could outperform the baselines. One limitation of our current model is that it is not an end-to-end solution, but needs a separate model or extra manual effort to provide the object information. In future work, we may explore the model's ability to determine the regions of interest for Gaussian readout sampling by itself.

SENSORIUM Rank 2: Zhu, Xiao, & Han

Unsuccessful approaches We try to improve the performance via pre-training CNN core. The first idea come to our mind is training in CLIP (Radford et al., 2021) way. CLIP improves performance of backbone network by minimizing the similarity of two different modalities. We consider images and neural responses in mouse visual cortex as two intrinsically related modalities. Therefore, we feed the features extracted by the core into a fully connected

layer to match dimensions of neural responses, thereby minimizing their similarity. This pre-training method enables our model to converge faster, but unfortunately, it does not bring performance improvements.

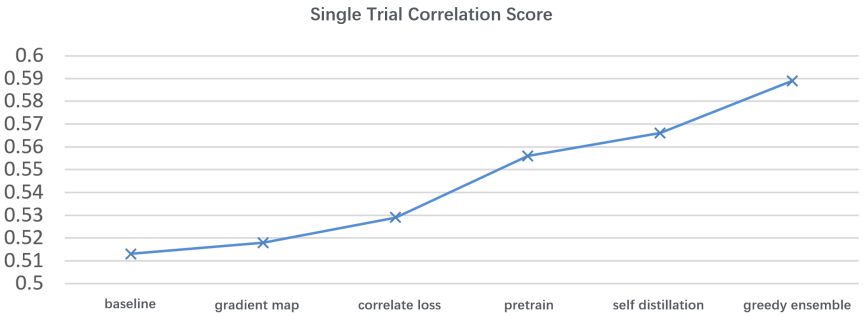


Figure 3: Ablation study for different training strategies in the SENSORIUM track.

SENSORIUM Rank 3: Azeglio, Ferrari, Neri, & Marre

Tried and Failed. Because the thin lens model does not take into account the spherical nature of the eye, it is not applicable to certain transformations of the input image. Thin lens approximations characterize the geometry of planar image projection, while spherical transformations map images onto a spherical surface. We incorporated spherical projection into our processing pipeline by sampling local power spectra across the visual field, and applying it as a preprocessing step to input images. We relied on the human Modulation Transfer Function (Navarro et al., 1993) estimated by Pamplona et al. (2013), which we appropriately scaled for application to mice. Compared with the baseline model, our results showed improvements in the third significant digit. Further attempts to extend our methods to SENSORIUM+ were unfruitful, possibly because this dataset lacks information about pupil position and dilation. Despite the above limitations, we remain interested in exploring the potential of this approach in future work.

Model	Single Trial Correlation	Correlation to Average	FEVE
Baseline CNN	0.29	0.543	0.482
Scattering CNN	0.31	0.56	0.495
Scattering SE[1]-CNN	0.31	0.559	0.492
VOneBlock CNN	0.30	0.557	0.487
VOneBlock SE-CNN	0.30	0.556	0.486
Ensemble	0.324	0.587	0.549

Table 4: Performance on the live test set of individual cores in the model ensemble.

Simplification of Readout Layer. We investigated how changing the shape of Σ in the Gaussian readout affects the final result. Our initial thought was that having a 2×2 covariance matrix for each of the neurons could be redundant. In our experiments in Table 5, we found that the shape of the Σ has almost no effect on the final result, and instead of having a 2×2 matrix for each neuron, fixing sigma to a single number seems to be sufficient.

Σ shape	$n \times 2 \times 2$	$n \times 2 \times 1$	$n \times 1 \times 1$	2×2	1×2	1×1
# params	31104	15552	7776	4	2	1
Acc	**0.311**	0.307	0.308	0.305	0.306	0.306

Table 5: Performance as a function of number of readout parameters. Different shapes of Σ with ResNet-9 encoder, where n represents the number of neurons.

Data augmentation. We tried augmenting the data, but with no further benefit (see Table 6). We guess one needs to match the readout shift for the geometrical augmentation precisely, but we didn't validate the hypothesis.

Aug	No Aug	Blur	ColorJitter	RRC2	RRC5	HFlip
Acc	**0.311**	0.305	0.296	0.103	0.167	0.258

Table 6: Model performance for various data augmentations. No Aug: ResNet-9 with no augmentations. RRC2: RandomResizedCrop with 20% crop lower bound. RRC5: RandomResizedCrop with 50% crop lower bound. Blur: Gaussian Blur. HFlip: Horizontal Flip. For each augmentation, the probability of application was set to 20%.

Self-supervised pre-training. For self-supervised pre-training we chose MoCo v2 (Chen et al., 2020) as our base framework. We perform pre-training only on the Core part of the network without Readout and Shifter. To do this, we add a pooling and linear layer, which produces the embedding of size 128. We tried different augmentations, but the pre-trained model didn't improve the final results.