Large-scale brain-wide neural recording in nonhuman primates

Eric M. Trautmann [1-3], Janis K. Hesse [4], Gabriel M. Stine [5,6], Ruobing Xia [7,8], Shude Zhu [7,8], Daniel J. O'Shea [9-11], Bill Karsh [12], Jennifer Colonell [12], Frank F. Lanfranchi [13], Saurabh Vyas [2], Andrew Zimnik [1,2], Natalie A. Steinmann [1,2], Daniel A. Wagenaar [4], Alexandru Andrei [14], Carolina Mora Lopez [14], John O'Callaghan [14], Jan Putzeys [14], Bogdan C. Raducanu [14], Marleen Welkenhuysen [14], Mark Churchland [1-3,15], *Tirin Moore [8,17], *Michael Shadlen [1,2,15,16], *Krishna Shenoy [8-11,17-19], *Doris Tsao [4,20], †Barundeb Dutta [14], †Timothy Harris [12,21]

* These authors contributed equally to this work
† These authors contributed equally to this work

Author Affiliations

1. Department of Neuroscience, Columbia University Medical Center, New York, NY, USA.
2. Zuckerman Institute, Columbia University, New York, NY, USA
3. Grossman Center for the Statistics of Mind, Columbia University, New York, NY, USA.
4. Department of Molecular and Cell Biology, University of California, Berkeley
5. McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA
6. Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA
7. Department of Neurobiology, Stanford University, Stanford, CA
8. Howard Hughes Medical Institute, Stanford University, Stanford, CA
9. Department of Electrical Engineering, Stanford University, Stanford, CA
10. Wu Tsai Neuroscience Institute, Stanford University, Stanford, CA
11. Bio-X Institute, Stanford University, Stanford, CA
12. Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA
13. Computation & Neural Systems, California Institute of Technology, Pasadena, CA
14. IMEC, Leuven, Belgium
15. Kavli Institute for Brain Sciences, Columbia University, New York, NY
16. Howard Hughes Medical Institute, Columbia University, New York, NY
17. Neuroscience Graduate Program, Stanford University, Stanford, CA
18. Department of Bioengineering, Stanford University, Stanford, CA
19. Department of Neurosurgery, School of Medicine, Stanford University, Stanford, CA
20. Howard Hughes Medical Institute, Berkeley, CA
21. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

We dedicate this manuscript to Krishna Shenoy (1968-2023), whose visionary leadership set this work in motion. His passion and dedication inspired a generation of neuroscientists and engineers, and his presence will continue to resonate within our field and community.

Abstract

High-density, integrated silicon electrodes have begun to transform systems neuroscience, by enabling large-scale neural population recordings with single cell resolution. Existing technologies, however, have provided limited functionality in nonhuman primate species such as macaques, which offer close models of human cognition and behavior. Here, we report the design, fabrication, and performance of Neuropixels 1.0-NHP, a high channel count linear electrode array designed to enable large-scale simultaneous recording in superficial and deep structures within the macaque or other large animal brain. These devices were fabricated in two versions: 4416 electrodes along a 45 mm shank, and 2496 along a 25 mm shank. For both versions, users can programmably select 384 channels, enabling simultaneous multi-area recording with a single probe. We demonstrate recording from over 3000 single neurons within a session, and simultaneous recordings from over 1000 neurons using multiple probes. This technology represents a significant increase in recording access and scalability relative to existing technologies, and enables new classes of experiments involving fine-grained electrophysiological characterization of brain areas, functional connectivity between cells, and simultaneous brain-wide recording at scale.

Introduction

High-channel count electrophysiological recording devices such as Neuropixels probes[1,2] are transforming neuroscience with rodent models by enabling recording from large populations of neurons anywhere in the rodent brain[3–9]. The capabilities provided by these approaches have led to important new discoveries, such as a continuous attractor network of grid cells in the rat hippocampus[9], establishing a causal topographical mapping between activity in the cortex and striatum in mice[4], and revealing how thirst modulates brain-wide neural population dynamics in mice[7]. The Neuropixels 1.0 probe has also been used to record neurons acutely in nonhuman primates (NHPs) like macaques[10–13] and in humans[14,15], but their short length (10 mm) restricts access to all but the most superficial targets, and the thinness of the shanks (24 μm) renders them fragile and difficult to insert through primate dura mater.

In a number of discussions with a community of primate researchers, scientists articulated the need for a probe more suitable for use with rhesus macaques. In particular, this community desired a technology that is capable of easily accessing as many neurons in as much of the brain as possible, and a technology which is compatible with conventional acute (single-session) recording techniques. These needs collectively suggested a probe design with a long linear array of dense recording sites, similar to the rodent probes, but with mechanical properties that would enable transdural insertion in macaques.

Current commercially-available technologies do not fully satisfy these needs. Linear array technologies, such as V or S-probes (Plexon Inc.), provide access to the whole brain, but are limited to 64 channels and have relatively large diameters (e.g., 380 μm) that increase as recording channels are added. Surface arrays like the Utah array[16] or floating microwire

arrays[17], allow for recording from up to 256 channels simultaneously, but are limited to recording at pre-specified depths in the superficial cortex, require opening the dura for placement, and cannot be moved after implantation. This latter constraint adds considerable risk to an experimental workflow, as such surgeries may not yield successful recording quality in animals trained for years on an experimental task. Alternative approaches using individually-driven single electrodes have achieved recordings in deep brain regions from dozens or hundreds of neurons by a few dedicated labs[18–20], though these approaches do not allow for dense sampling within a specific target region.

Alternatively, two photon (2P) calcium imaging, allows for large-scale population recordings at single cell resolution, but with limited temporal resolution (due to both calcium signaling kinetics and imaging scan rates) and limited recording depth (0–500 μm from the pial surface). In addition, the genetic modification required for 2P imaging remains challenging to implement reliably[21]. While approaches like microendoscopic imaging have been developed to address the imaging depth limitation, these require insertion of large (1mm) imaging lenses and provide access to a comparably small number of cells without moving the implanted lens[22].

We developed a high-density integrated silicon electrode array, optimized for recording in NHPs, and designed to enable flexible and configurable recording from large populations of neurons throughout the entire brain with single-neuron and single-spike resolution. While the probe's design and electronic specifications are based on the Neuropixels 1.0 probe (http://Neuropixels.org), fabricating these probes with the desired combination of mechanical and electrical properties using a standard CMOS photolithography process required additional advanced engineering. Each probe is larger than a photolithographic reticle, requiring "stitching" of electrical traces across the boundaries between multiple reticles in order to achieve the required probe geometry[23]. Relative to Neuropixels 1.0, the two variants of Neuropixels-NHP feature a longer, wider, and thicker shank (45 mm long, 125 μm wide, and 90 μm thick and 25 mm long, 125 μm wide, and 60 μm thick; Fig. 1a). For each probe variant, the full length of the shank is populated with recording sites with a density of 2 sites every 20 micrometers (Fig. 1b). A switch under every site allows flexible selection of the 384 simultaneous recording channels across these 4416 or 2496 sites (11 or 6 banks of 384 channels respectively plus one half-sized bank at the shank-base junction; Fig. 1c, Extended Data Fig. 1).

Here, we describe the Neuropixels 1.0-NHP, the engineering challenges surmounted in fabrication, and illustrate the ability to address a range of novel experimental use cases for neuroscientific data collection in large animal models. In particular, we focus on recording large populations of neurons, recording deep in the brain, and recording neurons from multiple brain regions simultaneously using one or multiple probes. We illustrate these use cases with example experiments designed to address specific questions in sensory, motor, and cognitive neuroscience using macaques. The scale and access provided by Neuropixels 1.0-NHP enable a wide range of new experimental paradigms, while streamlining neural data collection at a fraction of the cost of existing alternatives.

Results

Technology

The Neuropixels 1.0-NHP probe uses the same signal-conditioning circuits as the Neuropixels 1.0 probe[1], integrating 384 low-noise readout channels with programmable gain and 10-bit resolution, in a 130nm Silicon-on-Insulator CMOS platform. The probe shank consists of an array of shift-register elements and a switch matrix to enable the selection of electrode groups. The probe is fabricated as a monolithic piece of silicon, which includes the shank and base electronics, and the total probe lengths are 54 mm and 34 mm for the longer and shorter versions respectively, both larger than the optical reticle. Methods, known as *stitching*, were developed to overcome this limitation[24], which involves aligning features from different exposures between adjacent reticles.

The typical maximum size of a 130 nm CMOS chip is limited by the maximum reticle size (ie., 22 mm x 24 mm in this case) that can be exposed in a single lithography step. To build larger chips, the CMOS chip is divided into sub-blocks smaller than the reticle field, which are later used to re-compose (i.e. stitch together) the complete chip by performing multiple reticle exposures per layer. The stitched boundary regions of the small design blocks are overlapped (i.e., double exposed) to ensure uninterrupted metal traces from one reticle to the next, which is accounted for in the design. Since multiple levels of metal wires are needed to realize and interconnect the large number of electrodes at high density, the conducting wires in the shank must be aligned between adjacent reticles for four of the six stacked aluminum metal layers. To enable stitching and to simplify photo-mask design and reuse, the probe is designed with three elements as shown in Fig. 1d: a 5 mm tip (Fig. 1e), either one or two 20 mm middle shank segments (stitch segments 2 and 3 in Fig. 1d) for the 25 mm and 45 mm versions respectively, and the 10 mm base segment. Thus either three reticles (25 mm shanks) or four reticles (45 mm shanks) are required for each probe. The probe base is 6 mm wide, allowing four probes to be written across these three (25 mm shank) or four (45 mm shank) reticles.

Four design strategies were developed to mitigate the expected degradation of signal with increased shank length: 1) the 384 metal wires connecting the electrodes to the recording circuits in the base were made wider than the Neuropixels 1.0 probe shank to keep their resistance and thermal noise contribution low; 2) the spacing between the metal wires was increased to limit the signal coupling and crosstalk along the shank; 3) the power-supply wires connected to shank circuits were made wider to minimize voltage attenuation and fluctuations along the shank; and 4) the size of the decoupling capacitors is increased. The additional width and spacing of the metal lines also mitigated the impact of anticipated reticle misalignments, magnification, and rotational errors in the overlap regions during stitching. Fig. 1e shows details of the shank electrode placement and the dense CMOS layout in the shank and Fig. 1f shows a scanning-electron-microscope (SEM) image of the Al metal wires running along one of the 0.24-μm stitching regions in the shank. Due to the double exposure of the masking photoresist

at the stitching region, the wires become 24%-54% narrower in this region, but remain continuous circuits without interruption.

One of the major design changes to this device, compared to the Neuropixels 1.0 (rodent) probe[1], was to strengthen and thicken the probe shank, which was necessary both to support the longer 45 mm shank length and to allow the probe to penetrate primate dura. To achieve this, we increased the thickness of the shank by 3.75x, from 24um to 90μm. In addition, since the increased thickness altered the bending profile of the shank, we added stress compensation layers, which help to reduce intrinsic stresses within the material in the shank, and which enabled us to keep the bending within the same range as the rodent probes, despite the 450% increase in length.

For the data reported here, we used the Neuropixels 1.0-NHP probe version with a 125 μm wide, 45 mm long shank, 4416 selectable electrodes or pixels, and a 48 mm^2 base. To facilitate insertion of the shank into the brain and to minimize dimpling and tissue damage[1], the 20° top-plane chisel tapered shanks are mechanically ground to a 25° bevel angle on the side plane using a modified pipette micro grinder (Narishige EG-402). This procedure results in a tip which is sharp along both axes (Fig. 1g), allowing insertion through the dura in many conditions with reduced penetration forces. Additional discussion of insertion mechanics, methods, and hardware are provided in methods and documented in a hardware and methods wiki.

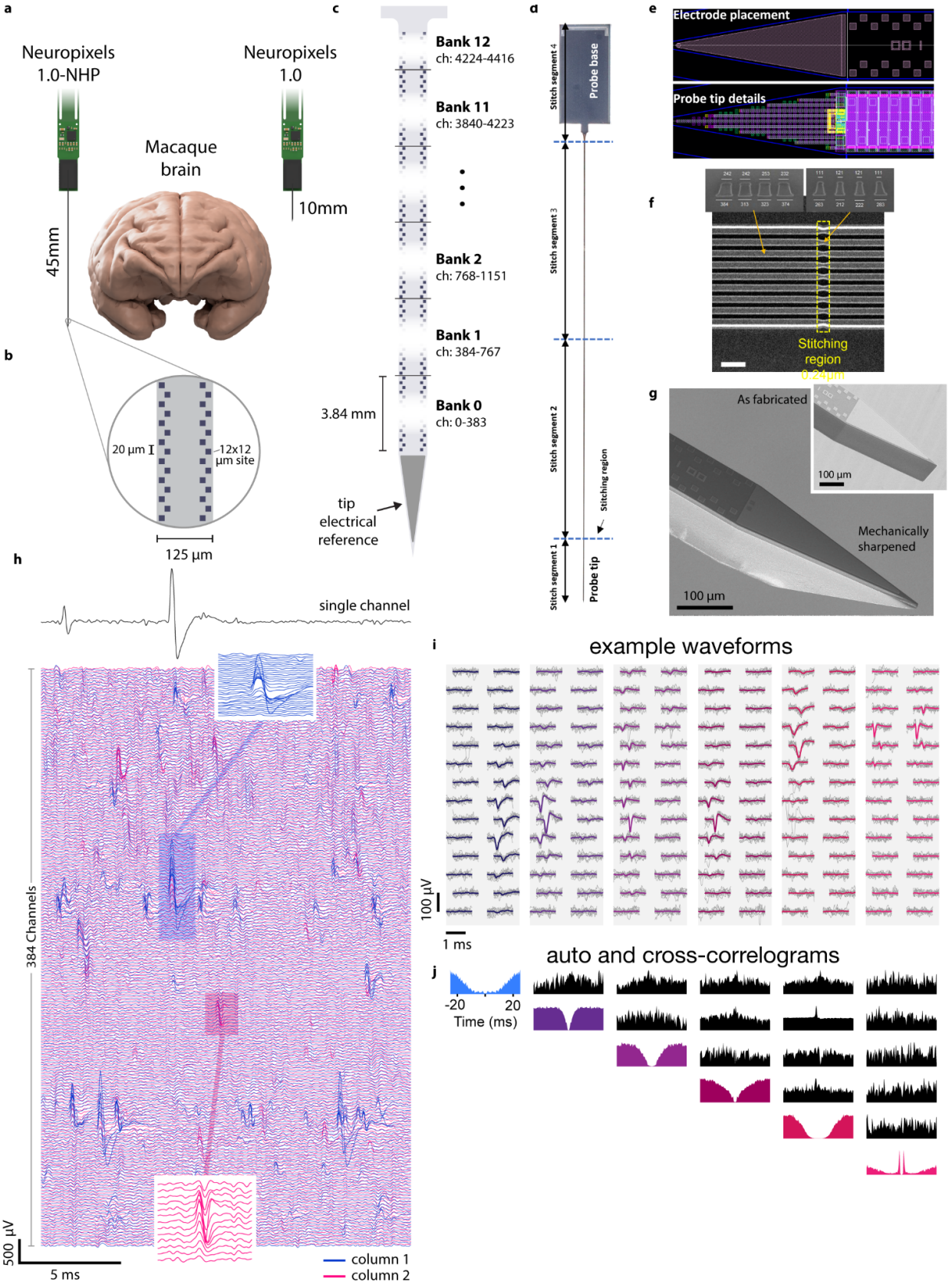


Figure 1 - Neuropixels 1.0-NHP probe characterization and engineering. **a,** comparison of probe geometry with macaque brain and Neuropixels 1.0 probe. **b,** electrode site layout. **c,** 4416 recording sites cover the full length of shank, grouped in 11.5 banks of 384 channels. **d,** Neuropixels 1.0-NHP probe photograph indicating the four segments or subblocks used for the sub-field stitching fabrication process. **e,** Details of the shank tip electrode layout (top) and CMOS circuit layout (bottom). **f,** Top-down scanning-electron-microscope (SEM) image of one of the metal layers in the shank when crossing the stitching region (scale bar: 1 μm); top-left: cross-section taken outside the stitching overlap region; top-right: cross-section at the narrowest point; the metal wires are narrower due to the double resist exposure. **g,** SEM picture of shank tip mechanically ground out-of-plane to 25°; inset: probe tip geometry as fabricated, pre-grinding. **h,** raw electrical recordings from 384 simultaneous channels in the motor cortex of a rhesus macaque monkey. Insets: expanded view of single spike events from single neurons. **i,** Example waveforms from isolated single neurons (gray) and median waveform (colored). **j,** auto and cross correlograms for neurons shown in d).

Scientific applications

The dense, high-channel count, programmable sites of Neuropixels 1.0-NHP provide a number of advantages relative to existing neural recording technologies appropriate for primates and other large animal models. First, the large number of simultaneously-recordable channels represents a transformative capability in itself. Large-scale recordings: A) permit rapid surveys and mapping of recorded brain regions, B) enable analyses which infer the neural state on single trials from large population recordings, and C) make it practical to infer functional connectivity using correlational analyses of spike timing. Second, the high density of recording sites enables high quality, automated spike sorting[25] when recording from one or both columns within a 3.84 mm bank of electrodes (Fig. 1c,h–j). Users can programmatically select to record with full density from one column in each of two banks, for 7.68 mm total length of high-quality single unit recordings (Extended Data Fig. 1). In addition, this high density enables continuous tracking of neurons in the event of drifting motion between the probe and tissue[14,15]. Third, the ability to programmatically select recording sites allows experimenters to decouple the process of optimizing a recording location from positioning the probe. This pragmatic yet important detail allows experimenters to leave the probe in place to settle prior to an experiment, in order to improve positional stability during the recording and to reduce the impact of any transient tissue response to insertion on subsequent recordings. Finally, users can leverage programmable site selection to perform a survey of activity along the entire length of the probe, without moving the probe, to map the relative position of electrophysiological features. This procedure can reduce or eliminate ambiguity of the probe's recording depth and location with respect to a target brain area. We illustrate these collective advantages using example recordings in different macaque brain structures in pursuit of diverse topics, including: 1) retinotopic organization of extrastriate visual cortex, 2) neural dynamics throughout the motor system, 3) face recognition in face patches of inferotemporal (IT) cortex, and 4) neural signals underlying decision making in posterior parietal cortex.

Dense recordings throughout primate visual cortex

More than half of the macaque neocortex is visual in function[26], and a multitude of visual areas, containing varied retinotopic organization and neurons with distinct feature-selective properties (e.g. motion, color, etc.), lie beyond the primary visual cortex (V1). Most of these visual areas, however, are distributed throughout the brain and located deep within the convolutions of the occipital, temporal and parietal lobes (Fig. 2a). As a consequence of this, and the limitations of prior single-neuron recording technologies, the majority of electrophysiological studies in visual neuroscience have focused on only a subset of visual areas. Fewer than half of the identified visual areas have been well-studied (e.g. areas V4, MT), while most (e.g. DP, V3A, FST, PO) have only been sparsely investigated, which is surprising given the clear similarities between the macaque and human visual systems[27,28]. A more thorough and systematic investigation of neural representation across the numerous contributing regions is only practical with technologies that enable large-scale surveys via simultaneous population recordings, with the capability of accessing both superficial and deep structures. Our initial tests with the Neuropixels 1.0-NHP probe demonstrated that it is well-suited for that purpose.

During individual experimental sessions, the activity of thousands of single neurons across multiple visual cortical areas could be recorded using a single NHP probe. Figure 2b shows the spike waveforms of 202 neurons recorded from one bank of electrodes spanning 6–10 mm below the cortical surface. As anticipated, the location of the neurons' visual receptive fields (RFs) varied as a function of the position along the Neuropixels probe. The RFs shifted in an orderly way for stretches of approximately 1 mm, consistent with a topographic representation, and then shifted abruptly, reflecting likely transitions between different retinotopic visual areas (e.g., refs. [29–31]; Fig. 2c, Extended Data Fig. 2a). Across the full depth, RFs tiled much of the contralateral hemifield, and included some of ipsilateral visual space as well (Fig. 2d, Extended Data Fig. 2b).

In other sessions, we recorded from up to five probe banks spanning 0–19 mm beneath the pial surface (Fig. 2e). In one example session, 3,029 single neurons were recorded, of which 2,729 neurons were visually responsive, exhibiting short-latency responses after stimulus onset (Fig. 2f). As with the single-bank recordings (Fig. 2c,d), neuronal RFs shifted gradually for contiguous stretches, punctuated by abrupt changes at specific depths. In the example shown (Fig. 2g), RFs at more superficial sites (0-3 mm) were located at more eccentric locations of the visual field, and then abruptly shifted towards the center and closer to the lower vertical meridian (LVM; ~3 mm). At the same location, neurons became more selective to the direction of motion (Extended Data Fig. 2d), suggesting a transition from area V4 to areas MT or MST. After that, RFs were located more centrally at the lower contralateral visual field and were observed across several mm. At deeper sites (~6–7 mm), smaller RFs clustered near the horizontal meridian (HM) for more than 1 mm, then quickly shifted toward the upper vertical meridian (UVM; ~8 mm). Finally, at the deepest sites (>10 mm), RFs generally became larger and much less well defined. These data illustrate how the Neuropixels 1.0-NHP's dense

sampling and single-unit resolution facilitates large-scale and unbiased mapping of the response properties of neurons across multiple visual areas in the primate brain.

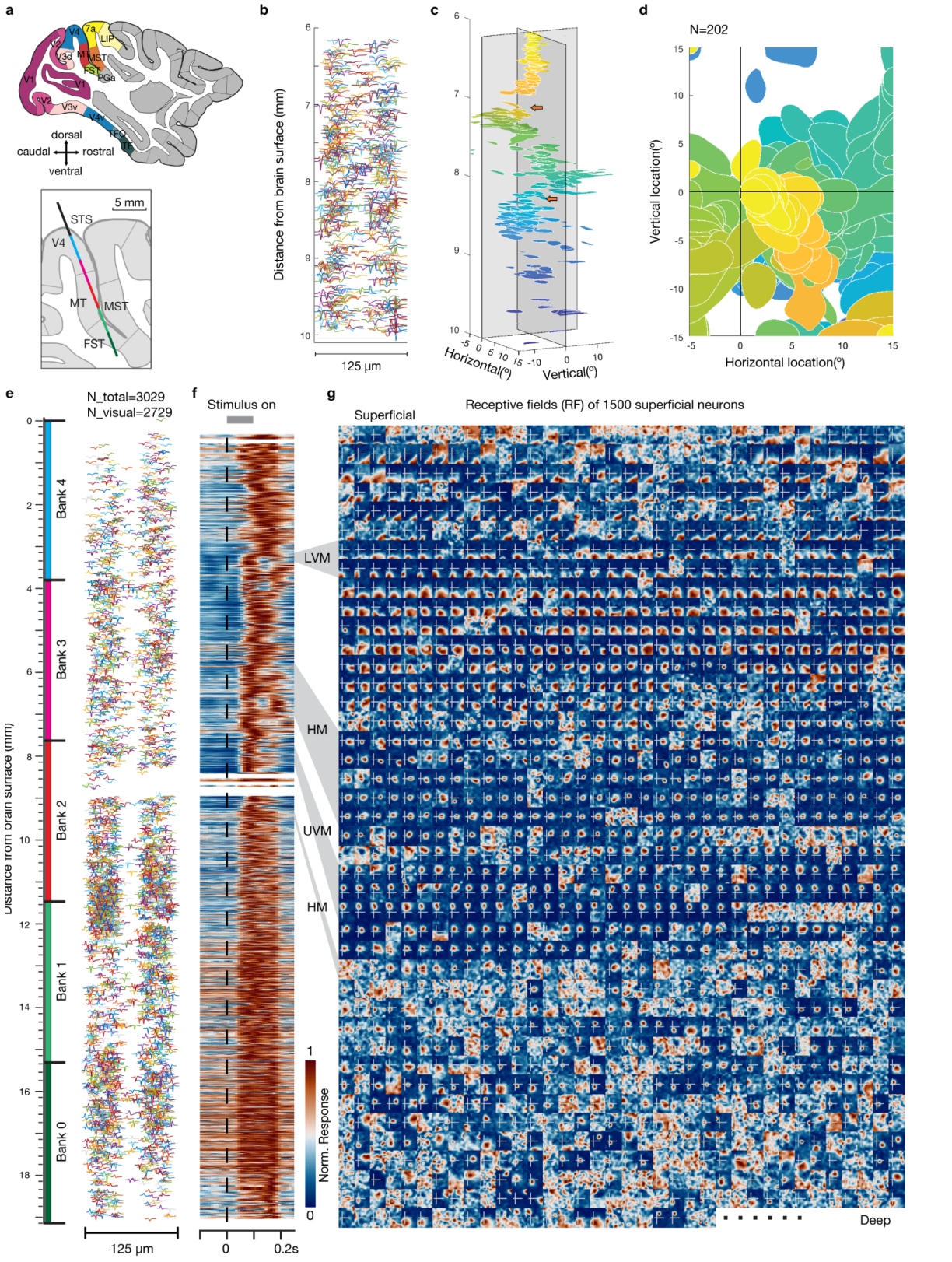


Figure 2 - Single and Multi-bank recordings across multiple visual cortical areas. **a,** Visual areas within macaque neocortex shown in a sagittal section. The inset shows the estimated probe trajectory of one multi-bank recording. **b,** Spike waveforms of single neurons recorded across a single bank of (384) electrode contacts (3.84 mm) shown at their measured location on the probe surface. **c,** Distribution of receptive fields (RFs) of 202 visually responsive neurons across cortical depth in a single-bank recording. Arrows denote abrupt changes in RF progressions and putative visual area boundaries. **d,** Top-view of **c** illustrating the coverage of RFs across the contralateral visual field. **e,** Spike waveforms of 3,029 single neurons recorded across 5 banks of electrode contacts (~19 mm) shown at their measured location on the probe surface. **f,** Heat map of stimulus evoked responses for all 2729 visually responsive neurons. Each neuron is plotted at its corresponding cortical depth. Dashed black line denotes stimulus onset; gray line at top denotes (0.1 sec) duration. Gray shading on the right denotes depths where RFs fell on the lower vertical meridian (LVM), horizontal meridian (HM), or upper vertical meridian (UVM). **g,** RF heat maps for 1500 of the superficial most neurons, indicating visual field locations where stimuli evoked responses for each single neuron. White crosshair in each map denotes the estimated horizontal and vertical meridians. RFs are arranged in a 42(rows) by 36 (columns) array.

Large-scale recording throughout the motor system

Next, we demonstrate the utility of this technology for studying the multiple brain areas involved in movement control. Primary motor cortex (M1) is situated at the rostral bank of the central sulcus and extends along the precentral gyrus. Sulcal M1 contains the densest projections of descending corticomotoneuronal cells and corticospinal neurons, which collectively are understood to convey the primary efferent signals from the brain to the periphery[32,33]. Constraints of existing technology have led to two broad limitations in studies of the motor system. First, motor electrophysiologists have been forced to choose between simultaneous recording from populations of superficial neurons in gyral motor cortex (PMd and rostral M1) using utah arrays, or alternatively, recording fewer neurons in sulcal M1 using single-wire electrodes or passive arrays of 16–32 contacts (e.g., Plexon S-probes or Microprobes Floating Microwire Arrays[17]). Recording from large populations of neurons in sulcal M1 has not been feasible.

Second, the motor cortex is only one part of an extensive network of cortical and subcortical structures involved in generating movement[34]. Many investigations of the motor system focus on the primary motor cortex, and comparably fewer experiments have systematically investigated neural responses from the numerous additional brain structures involved in planning and controlling movements, in part due to the challenge of obtaining large-scale datasets in subcortical structures in primates. Areas such as the supplementary motor area (SMA) and the basal ganglia (BG) are understood to be important for planning and controlling movements, but systematic investigation of the functional roles and interactions between these regions is hampered by the challenge of simultaneously recording from multiple areas.

We developed an insertion system capable of simultaneous insertion of multiple Neuropixels 1.0-NHP probes to superficial and deep structures of rhesus macaques. We tested this approach using a motor behavioral task, in which a monkey used isometric forces to track the position of a scrolling path of dots (Fig. 3a, task described in ref. [35]). During this task, we

recorded from the primary motor and premotor cortex (M1 and PMd, Fig. 3b) while the monkey generated repeated forces across multiple trials (Fig 3d). Single neurons exhibit a diversity of temporal dynamics throughout the motor behavior (Fig. 3e–f), and the highest-variance principal components illustrate population dynamics throughout the motor behavior (Fig 3g). Predicting the arm force via linear regression from the neural population reveals improved model performance as more neurons are included, and performance does not saturate when including all 360 recorded neurons for an example session (Fig. 3h). Despite the apparent simplicity of a one-dimensional force tracking task, it is necessary to sample from many hundreds of neurons in order to capture a complete portrait of the population-level neural dynamics.

In many experimental situations, it is desirable to be able to precisely localize the probe within the anatomy of the target region. Current source density (CSD) is often used to infer the recording depth using consistent patterns of sources and sinks when the recording trajectory is normal to the cortical lamina[11,36,37]. The high-density of recording sites on the Neuropixels 1.0-NHP allows for averaging of CSD within local regions and high-quality CSD calculations, revealing spatial and temporal structure throughout the force-tracking task (Fig. 3i, Extended Data Fig. 3). This approach may be used to gain confidence in the recording location, and to map the response properties of individual neurons to specific locations within the cortex or deep brain structures.

In order to simplify trajectory planning for one or multiple probes, we developed an insertion system which allows for precise computer-aided design (CAD) modeling of anatomy and probe insertion geometry (Fig. 3j). A 3D-printed chamber adapter is used to precisely orient multiple probes at angles that can be skew to the recording chamber's principle axis (Fig. 3k). Using this approach, it is possible to record from multiple probes within a compact target region. Here, we demonstrate recordings of 673 neurons in gyral PMd using three probes (Fig. 3l), and up to 819 neurons using six probes in a separate session.

This approach is also well optimized for recording with precise targeting in deep structures, to study other brain areas which are understood to contribute to motor control. The Neuropixels 1.0-NHP probe simplifies survey experiments by enabling recording from multiple disparate regions using a single probe via site programmability (e.g. basal ganglia and motor cortex, Fig 3m), and via recording from hundreds of neurons on multiple probes in disparate targets simultaneously (Fig 3n–o). Using three probes, we demonstrate simultaneous recording from M1, the internal globus pallidus (GPi) of the basal ganglia, and the supplementary motor area (SMA). Neurons in all three regions displayed modulated trial-averaged activity patterns during a sequential multi-target reaching task.

Collectively, the capabilities enabled by the Neuropixels 1.0-NHP probe make new classes of experiments feasible, in addition to dramatically reducing the challenge of recording neural data. This is particularly true for a range of deeper brain regions known to be important for motor control, but which have historically received less attention due to the relative challenge of accessing populations of neurons in these areas.

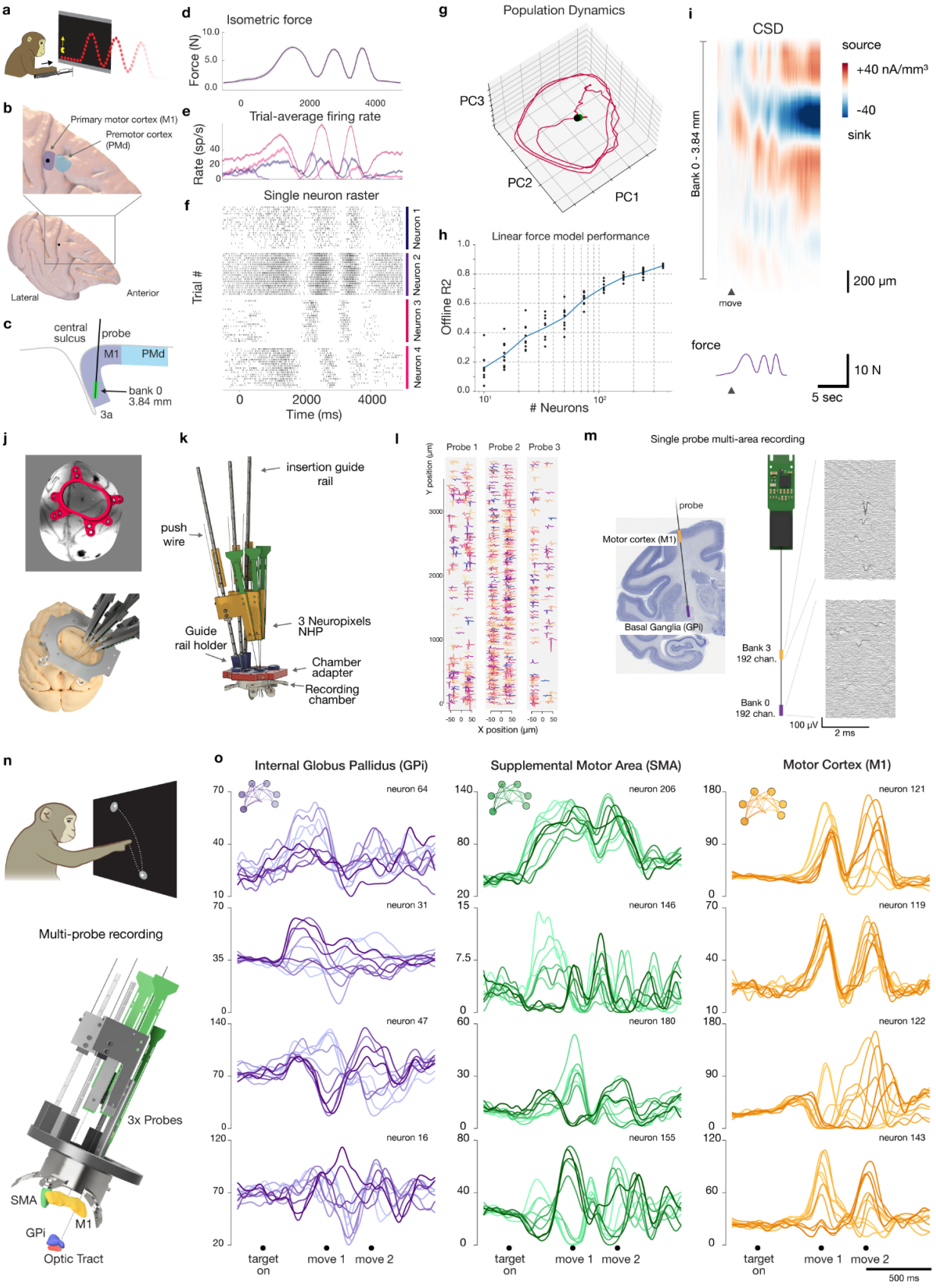


Figure 3 - Large scale recording throughout the rhesus motor system. **a,** Isometric force path-tracking (pacman) behavioral task. **b,** Probe insertion point in motor cortex visualized on macaque brain rendering. **c,** Schematic recording target in sulcal M1, sagittal section. **d,** Force generated by a monkey's arm during the pacman task. **e,** Trial-averaged firing rate of four example neurons. **f,** Single-trial spiking rasters of four example neurons. **g,** Low-dimensional PCA trajectories of the population neural activity. **h,** Linear model force prediction accuracy as a function of the number of neurons included in the analysis. **i,** Current source density (CSD) calculated using LFP-band signals. **j,** Application 2: multi probe recording within a single brain region. (top) Curvilinear brain from MRI imaging with recording chamber model. (bottom) 3D model of insertion setup for three probes with trajectories targeting close-packing arrangement in M1 while minimizing mechanical interference. **k,** 3D model of insertion hardware, illustrating use of linear rail guides for precise probe positioning. **l,** Waveforms from roughly 600 simultaneously-recorded units in motor cortex. **m,** Application 3: simultaneous dual region recording using a single probe in M1 and GPi. **n,** Application 4: simultaneous multi-region recording using multiple probes. (top) Sequential target reaching task. (bottom) 3D model of recording targets (GPi, M1, and SMA - reconstructed from MRI imaging) and insertion hardware. **o,** PSTHs from four neurons from each of GPi, SMA, and M1.

Face recognition in IT cortex

Next, we demonstrate use of the probes in the inferotemporal (IT) cortex, a brain region that is challenging to access due to its depth, spanning the lower gyrus and ventral surface of the temporal lobe. IT cortex is a critical brain region supporting high-level object recognition, and has been shown to harbor several discrete networks[38], each specialized for a specific class of objects. The network that was discovered first and has been most well-studied in nonhuman primates is the face patch system. This system consists of six discrete patches in each hemisphere[39], which are anatomically and functionally connected. Each patch contains a large concentration of cells that respond more strongly to images of faces than to images of other objects. Studying the face patch system has yielded many insights that have transferred to other networks in IT cortex, including increasing view invariance going from posterior to anterior patches and a simple, linear encoding scheme[38]. As such, this system represents an approachable model for studying high-level object recognition[40]. The code for facial identity in these patches is understood well enough that images of presented faces can be accurately reconstructed from neural activity of just a few hundred neurons[41].

A major remaining puzzle is how different nodes of the face patch hierarchy interact to generate object percepts. To answer this question, it is imperative to record from large populations of neurons in multiple face patches simultaneously to observe the varying dynamics of face patch interactions on a single-trial basis. This is essential since the same image can often invoke different object percepts on different trials[42]. Here, we recorded with one probe in each of two face patches, middle lateral (ML) and anterior fundus (AF), simultaneously (Fig. 4a). The Neuropixels 1.0-NHP probes recorded responses of 1,127 single- and multi-units across both face patches during a single session (Fig. 4a, right). Changing the visual stimulus to a monkey face yielded a clear visual response across both face patch populations. We measured responses of visually responsive cells to 96 different stimuli containing faces and non-face objects (Fig. 4b–c). A majority of cells in the two patches showed clear face-selectivity. With single-wire tungsten electrodes, this dataset would take

about two years to collect, but is now possible in a single two-hour experimental session[41]. In addition to the gain in efficiency created by this technology, simultaneous recordings of multiple cells and multiple areas allows for investigation of how populations encode object identity in case of uncertain or ambiguous stimuli, where the interpretation of the stimulus may vary from trial to trial but is nevertheless highly coherent on each trial. The anatomical depth of face patches puts them far out of reach for shorter high-density probes. For example, face patch AM sits roughly ~42 mm from the craniotomy (Fig. 4d) along a conventional insertion trajectory.

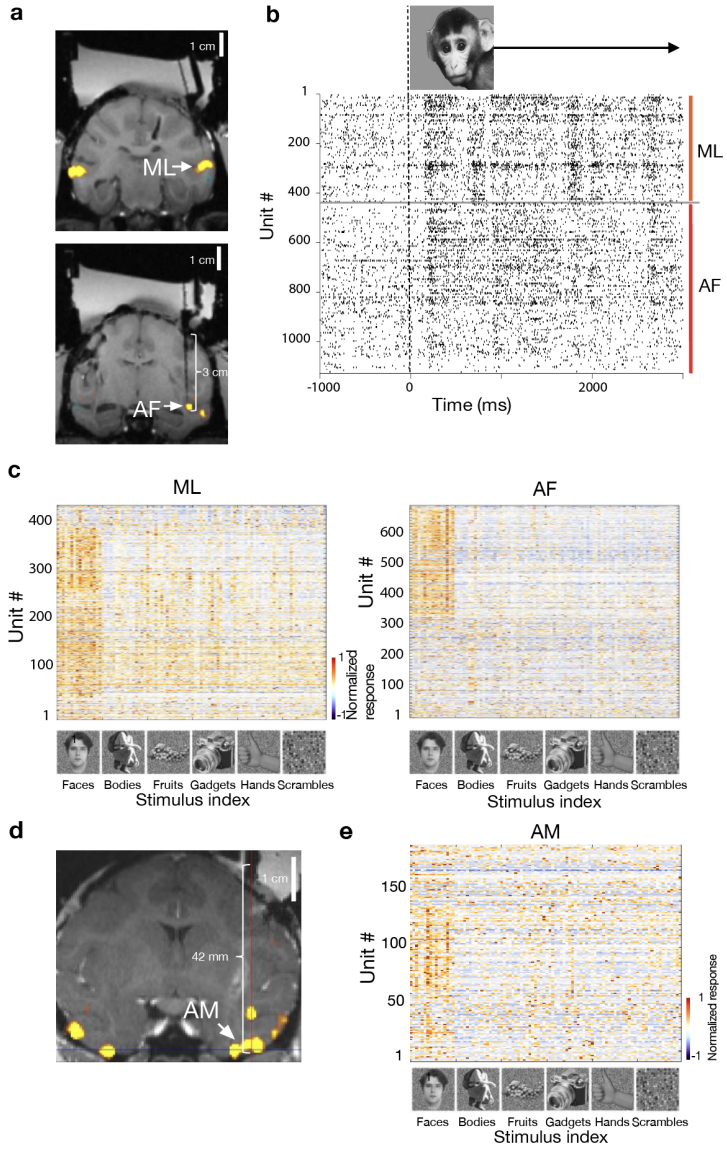


Figure 4 - Deep, simultaneous recordings from two face patches in IT cortex. a, Left: Simultaneous targeting of two face patches. Coronal slices from magnetic resonance imaging scan show inserted tungsten electrodes used to verify targeting accuracy for subsequent recordings using Neuropixels 1.0-NHP (top: face patch ML, bottom: face patch AF). Color overlays (yellow) illustrate functional magnetic resonance imaging contrast in response to faces vs. objects. **b,** Response rasters for a single stimulus presentation of simultaneously recorded neurons in ML and AF to a monkey face, presented at t=0. Each line in the raster corresponds to a spike

from a single neuron or multi-unit cluster (including both well isolated single units and multi-unit clusters. **c,** Neuropixels 1.0-NHP enables recordings from many face cells simultaneously. These plots show average responses (baseline-subtracted and normalized) of visually responsive cells (rows) to 96 stimuli (columns) from six categories, including faces and other objects. Bottom panel shows exemplar stimuli from each category. The plots include 438 cells or multiunit clusters in ML (left) and 689 in AF (right), out of which a large proportion responds selectively to faces. Units are sorted by channel, revealing that face cells are spatially clustered across the probe. **d,** Same as a, but for the deepest IT face patch AM (recording performed in a different session from data in a–c). **e,** Same as c, for face patch AM.

Single-trial correlates of decision making in LIP

For many cognitive functions, the processes that give rise to behavior vary across repetitions of a task. For example, certain perceptual decisions are thought to arise through the accumulation of noisy evidence to a stopping criterion, such that their evolution is unique on each trial. This process is widely observed and known as drift-diffusion[43,44]. Its neural correlate has been observed in the lateral intraparietal area (LIP). Neurons in LIP have spatial response fields (RFs; Fig. 5c, top) and represent the accumulated evidence for directing the gaze toward the response field. Limitations in recording technology prevented previous studies from recording many LIP neurons simultaneously, requiring that neural activity be averaged across similar decisions. Such averaging highlights the shared features of activity across decisions (i.e., the “drift” component) but discards the unique dynamics that give rise to each individual decision (i.e., the diffusion component).

Neuropixels recording in LIP reveals the neural correlates of a single decision. These recordings yield 50–250 simultaneously recorded neurons, of which 10–35 share a RF that overlaps one of the contralateral choice targets used by the monkey to report its decision in a dynamic motion discrimination task (Fig. 5a). The average activity of these target-in-RF neurons, on a single trial, tracks the monkey’s accumulated evidence as it contemplates its options. The signal explains much of the variability in the monkey’s choices and reaction times[45] and conforms to drift-diffusion dynamics (Fig. 5d–e, top).

Neuropixels technology also enables multi-area recordings from ensembles of neurons that share common features. For example, neurons in the superior colliculus (SC) receive input from LIP and, like LIP neurons, also have spatial RFs and decision-related activity. An ideal experiment to understand how the two areas interact is to record simultaneously from populations of neurons in LIP and SC that share the same RF. This experiment is nearly impossible with previous recording technology because of the anatomical organization in LIP and SC. SC is retinotopically organized such that nearby neurons share the same response field. This level of organization is absent in LIP. Therefore, it is improbable that a given LIP neuron will have a response field that overlaps with those of a cluster of SC neurons. With patience, one can find one LIP neuron with the desired RF, but the likelihood of encountering more than two is vanishingly small. This challenge is overcome by the large number of neurons yielded by Neuropixels 1.0-NHP probe recording, allowing for post-hoc identification of neurons with overlapping RFs. The lower panel of Figure 5c depicts the response field of an

example SC neuron, recorded with a 32-channel V-probe (Plexon). The response field overlaps the left-choice target as well as 15 simultaneously recorded neurons in the SC. Importantly, it also overlaps the RF of 17 of the 203 neurons recorded with the Neuropixels probe in area LIP. (e.g., Figure 5c, top). Unlike LIP, single trial analysis of the SC population reveals dynamics that are not consistent with drift-diffusion (Figure 5d-e, bottom). Instead, SC exhibits bursting dynamics, which were found to be related to the implementation of a threshold computation[46]. The distinct dynamics in LIP and SC during decision making were only observable through single-trial analyses, the resolution of which are greatly improved with the high yield of the Neuropixels 1.0-NHP recording.

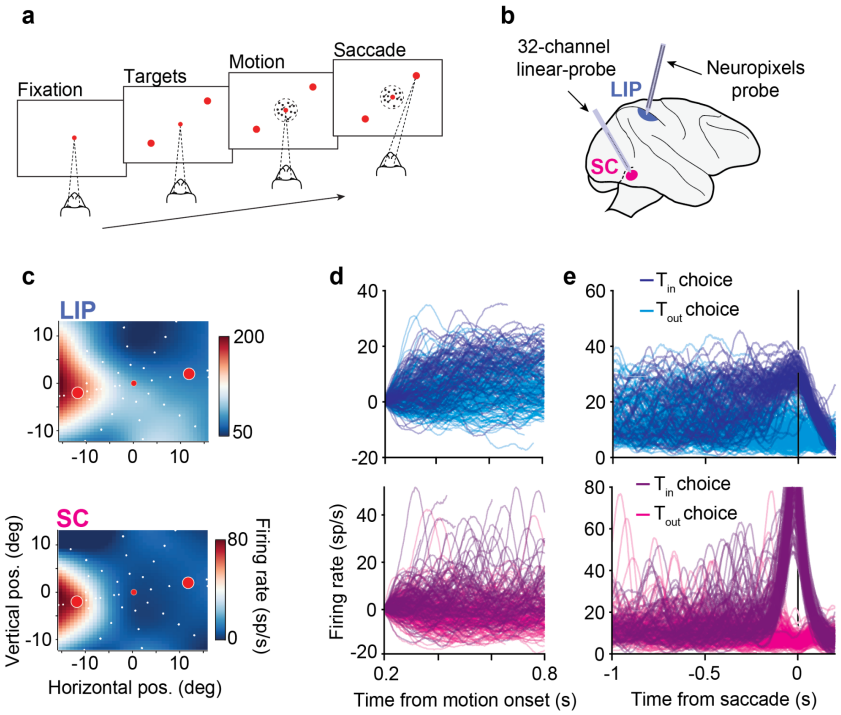


Figure 5 - Single-trial dynamics of a decision process in multiple brain regions. a, Dynamic motion discrimination task using random dot motion (RDM) stimulus. Upon fixation and onset of the choice targets, a random-dot-motion stimulus appears in the center of the display. Monkeys discriminate the direction of motion and, when ready, indicate their choice with a saccade to one of two choice-targets. **b**, Simultaneous recordings in LIP and SC. Populations of neurons were recorded in LIP with a Neuropixels 1.0-NHP probe and in SC with a multi-channel V-probe (Plexon). **c,** The response field (RF) of an example neuron in LIP (top) and SC (bottom). The colormap depicts the mean firing rate, interpolated across target locations (white circles), during the delay epoch of an oculomotor delayed-saccade task. Red circles depict the location of the choice targets in the RDM task in this session. **d**, Single trial activity in LIP (top, n = 17) and SC (bottom, n = 15) from an example session. Each trace depicts the smoothed firing rate average (Gaussian kernel, σ = 25 ms) of the neuronal population on a single trial, aligned to the onset of the motion stimulus. Rates are offset by the mean firing rate 0.18–0.2 s after motion onset to force traces to begin at zero. The line color indicates the animal's choice on that trial. **e**, The same trials as in **d** aligned to saccade initiation, without baseline offset.

Inferring functional connectivity with high density recordings

Understanding how the anatomical structure of specific neural circuits enables unique neural computations remains an important but elusive goal of systems neuroscience. One step towards connecting disparate levels of experimental inquiry is mapping functional connections, or inferred connectivity between neurons using correlative measures of spike timing[6,11,47]. This is often impractical or extremely challenging when only recording several neurons, as the likelihood of recording from a connected pair of neurons is quite low (e.g., 0.96% for two neurons recorded within the same bank in mice[6]), but the probability of successfully recording from two functionally-connected neurons increases by the square of the number of neurons recorded.

The Neuropixels 1.0-NHP probe typically yields 200–450 (and sometimes more) neurons when recording with 384 channels in cortical tissue. Applying the same methodology established in ref.[6] to 13 sessions from rhesus PMd and M1 yielded 111 ± 89 putative connected pairs per session, and a connection probability of 0.73% ± 0.61%. Fig. 6a shows six example jitter-corrected cross correlogram plots between pairs of neurons with significant peaks in the CCG. In many examples the CCG peak is lagged between one neuron relative to the other, consistent with a 1–2 ms synaptic delay. For other neuron pairs, the CCG peak is synchronous between the two neurons, suggesting that they may receive common input (e.g., Fig 6a, top right panel). Fig 6b illustrates the distribution of spike timing delays for 479 neurons from an example recording session. Nearby neurons are more likely to exhibit functional connections than neurons located further apart (Fig. 6c). Using this approach, we can map the full set of putative connections for a given recording across cortical lamina (Fig. 6d).

In addition to the cortical distance, we further assessed the dependence of functional connectivity on the tuning similarity between neuronal pairs. For example, neuron populations with diverse receptive fields (RFs) obtained from multiple visual areas (Fig. 2) allow us to quantitatively determine the extent to which functional connections between neurons depend on their RF overlapping. Using signal correlation (r_{sig}) as a measure of such tuning similarity in visual fields[48], we show that functionally connected neuron pairs exhibit higher r_{sig} compared with non-significant pairs (Fig. 6e). Specifically, synchronous pairs (putatively receiving common inputs) tend to share highly overlapping RFs (r_{sig} mean=0.60, $p<10^{-4}$ compared with non-significant pairs), while asynchronous pairs (putatively exhibiting synaptical connections) more likely to share moderately overlapping RFs (r_{sig} mean=0.43, $p<10^{-4}$ compared with non-significant pairs, and $p<10^{-4}$ compared with synchronous pairs). Moreover, the amplitude of the significant CCG peaks is positively correlated with the r_{sig} (Fig. 6f), indicating that neurons with similar RFs tend to have stronger functional connections, which is consistent with previous studies (e.g., [49,50]).

This same methodology can be applied to assess functional connectivity between multiple simultaneously recorded regions recording using separate probes to assess feedforward and feedback connectivity between two regions. Fig. 6g (left) shows the jitter-corrected CCG for a putatively connected pair of neurons where one neuron is located in face patch ML and the

other in face patch AF in IT cortex. Fig 6g (right) shows a single-trial spike raster of the putative post-synaptic neuron in area AF, triggered on spikes of the neuron in ML, illustrating a putative feedforward connection, while Fig 6h illustrates a putative feedback connection with opposite timing response from the cell pair shown in Fig. 6g. Remarkably, pairs of face cells were over 10 times as likely to be connected (1.6%) as other pairs of non-face responsive cells (0.13%) between ML and AF (Fig. 6i).

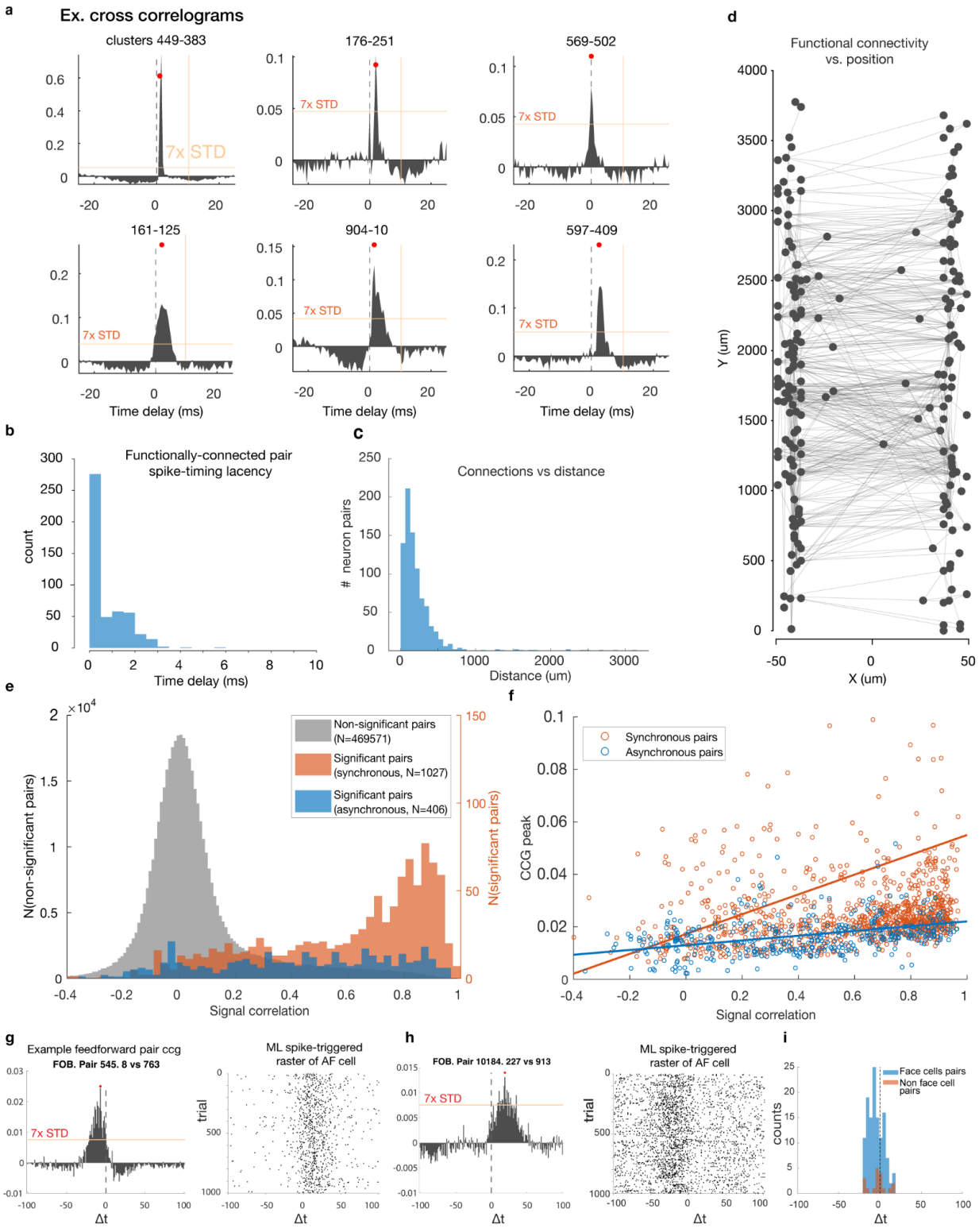


Figure 6 - Inferring functional connectivity from unit cross-correlation and high-density recording. **a,** Jitter-corrected cross correlograms of four example pairs of neurons exhibiting significant correlations in spike timing. **b,** Distribution of spike timing latency for 479 putative connected pairs from one

recording session. **c,** Histogram of connections as a function of distance between two cells. **d,** Neuronal connectivity matrix inferred from CCG analysis, with neurons ordered by depth along the probe. **e,** Distribution of signal correlation for pairs of neurons with different CCG types in the visual cortex. **f,** relationship between signal correlation and peak value of CCG. **g,** Example putative connected cell pair identified using two probes in area ML and AF with putative feedforward connection. **h,** Example putative connected cell pair identified using two probes in area ML and AF with putative feedback connection. **i,** The population of functionally connected cells between AL and MF regions is dominated by cells that respond to faces.

Discussion

We have presented a new recording technology and suite of techniques to enable electrophysiological recordings using high-density integrated silicon electrodes in nonhuman primates like rhesus. This technology enables large-scale recordings from populations of hundreds of neurons from deep structures in brain areas that are inaccessible using alternative technologies. The key methodological advance in the Neuropixels 1.0-NHP probe is the longer recording shank, which required developing techniques to adapt photolithographic silicon manufacturing methods to allow for “stitching” across multiple reticles. This advance allows for the manufacturing of monolithic silicon devices that span across multiple reticles to achieve larger sizes than could otherwise be manufactured. Creating a long and thin probe shank also required developing approaches for reducing bending due to internal stresses within the shank.

This technology combines the advantages of multiple approaches - recording with single-neuron spatial resolution and single-spike temporal resolution, while providing recording access to the majority of the macaque brain. Programmable site selection enables recording from multiple brain structures using a single probe, as well as surveying multiple recording sites along the shank without moving the probe. The combination of the compact form factor, commercially-available and turn-key recording hardware, and comparably modest cost enable straightforward scaling in the size of simultaneously recorded neural populations. This capabilities may be essential for achieving accurate estimates of neural dynamics on single trials, or for estimating the value of small-variance neural signals embedded in the neural population response.

The high spatial resolution offers a number of advantages over sparser sampling, including high quality single-unit isolation, automated drift correction (e.g., refs. [25,51], Extended Data Fig. 4), and localizing the position and depth of the recording electrodes within a brain structure (e.g., inferring probe depth with respect to cortical lamina) using current source density or other features of the recording. The high density also offers additional advantages not described here, such as identifying putative neuron subclasses using extracellular waveforms[5,14,52].

The probe will be commercially available and integrates seamlessly with the existing set of community-supported hardware and software tools for Neuropixels probes. The Neuropixels 1.0-NHP recording system is straightforward to set up and to integrate with other experimental hardware, like behavior or stimulus control computers. Combining the low total system cost

(roughly $7–15k) and large-scale recordings enables a dramatic reduction in the recording cost per neuron acquired relative to existing technologies.

While highly capable, the Neuropixels 1.0-NHP probes are limited in several ways. First, this technology is not optimized for simultaneous, dense sampling across a wide swath of cortex. For applications requiring horizontal sampling, planar recording technologies like Utah arrays or two-photon calcium imaging may be more appropriate. Second, in contrast with many passive electrodes, it is not currently possible to use the Neuropixels probe to deliver intracortical microstimulation (ICMS), though future versions of the probe may add this functionality. The Neuropixels technology is, however, capable of recording while stimulating through external electrodes, as recently demonstrated by O'Shea, Duncker, et al.[13]. Third, the Neuropixels 1.0-NHP design is not explicitly optimized for chronic implantation. While it is likely possible to leave the probe in place over multiple days or sessions, this theoretical capability remains untested and requires new implant designs. The probe base contains active electronics, and is not designed for implantation under dura. As such, a chronic implant design may require mounting the probe in a manner that allows it to mechanically "float" with the brain, to prevent relative motion between the probe and the tissue as the brain moves. As such, the current probe is most appropriate for acute recordings, though it conceivably might be implanted for sub-chronic recordings with appropriate insertion methods and hardware. Lastly, we note that while it is theoretically possible to insert the entire 45 mm long shank into the brain, inserting a probe this deep introduces additional practical challenges to overcome—primarily a requirement for precise alignment of the probe's insertion axis with the insertion location. A thorough discussion of these considerations is presented in a user's wiki https://github.com/etrautmann/Neuropixels-NHP-hardware/wiki.

Taken together, these methodological advances enable new classes of neuroscientific experiments in large animal models and provide a viable scaling path towards recording throughout the whole brain.

Materials and Methods Summary

Probe design and recording system

The Neuropixels 1.0-NHP probe consists of an integrated base and "shank" fabricated as a monolithic piece of silicon using a 130 nm CMOS lithography process. The 6 mm x 9 mm base is mounted to a 7.2 x 23 mm PCB, which is attached to a 7.2 x 40 mm long flexible PCB. This flex PCB plugs into a ZIF connector on a headstage (15 x 16 mm, 900 mg), which is connected to a PXIe controller mounted in an PXIe chassis using a 5 m twisted-wire cable. The base electronics, headstage, cable, PXIe system, and software are identical to the Neuropixels 1.0 probe. Data collection was performed using SpikeGLX software[53] and the system is fully compatible with OpenEphys software.

The Neuropixels 1.0-NHP probe is manufactured in two variants: 1) 45 mm long x 125 μm wide x 90 μm thick, featuring 4416 electrodes comprising 11.5 banks of 384 channels each; and 2) 25 mm long, 125 μm wide, and 60 μm thick, featuring 2496 electrodes comprising 6.5 banks of 384 channels. We note here that the commercial release of these probes will feature electrodes distributed in two aligned vertical columns, in contrast to the "zig-zag" columns described in this manuscript, in order to optimize data collection for automated drift correction to enhance automated spike sorting. We also note that a third variant, identical to Neuropixels 1.0 but with a thicker shank (100 μm vs. 25 μm), is also planned for commercial release.

Recording sites are 12 μm x 12 μm, made of titanium nitride , and have an impedance of 150 kΩ at 1kHz. The tips of the probes were mechanically beveled to a 25° angle using the Narishige EG-402 micropipette beveler. During recordings, electrical measurements were referenced to either: 1) the large electrical reference point on the tip of the electrode, 2) an external electrical reference wire placed within the recording chamber, or 3) a stainless steel guide tube cannula. Electrical signals are digitized and recorded separately for the action potential (AP) band (10 bits, 30 kHz, 5.7 μV mean input-referred noise) and local field potential (LFP) band (10 bits, 2.5 kHz).

Recording sites are programmatically selectable with some constraints on site-selection. See Extended Data Fig. 1 for a description of site selection rules and common configurations. Spike sorting was performed using Kilosort 2.5 and Kilosort 3.0, and results were curated using Phy. Analysis was performed using custom scripts written in Matlab and Python, leveraging the open source software package neuropixels-utils: (https://github.com/djoshea/neuropixel-utils).

Probe insertion

Several distinct methods were used to mount and insert probes, guided by the unique constraints of inserting probes to different depths, and depending on the recording chambers and mechanical access available for different primates used in these studies, as well as the

existing hardware used by each of four distinct research groups. For single probe insertions, probes were mounted using custom adapters to a commercially available electrode drive (e.g., Narishige corp.), and inserted through a blunt guide tube for superficial recordings, and a sharp penetrating guide tube for deeper recordings. When using a non-penetrating guide tube, the dura was typically penetrated with a tungsten electrode prior to using a Neuropixels probe to create a small perforation in the dura to ease insertion. When inserting electrodes to deep targets (>20 mm), the alignment between the drive axis and the probe shank is essential for enabling safe insertion, as misalignment can cause the probe to break. For this application, we developed several approaches to maintain precise alignment of the probe and drive axis. First, we employed a linear rail bearing (IKO International) and custom 3D printed fixture to maintain precise alignment of the insertion trajectory. This approach is discussed in detail in the accompanying Neuropixels-NHP wiki (https://github.com/cortex-lab/neuropixels/wiki).

For the experiments shown in figure 4, we developed a dovetail rail system that maintains precise alignment between a penetrating guide tube and the Neuropixels probe. The choice of appropriate insertion method depends on the mechanical constraints introduced by the recording chamber design, the depth of recording targets, number of simultaneous probes required, and choice of penetrating or non-penetrating guide tube. The interaction of these constraints and a more thorough discussion of insertion approaches is provided on the Neuropixels users wiki . Open-source designs for mechanical mounting components for Neuropixels-1.0-NHP to drives from Narishige, NAN, and other systems are available in a public repository: https://github.com/etrautmann/Neuropixels-NHP-hardware.

Visual cortex recordings and analysis

Two male adult rhesus monkeys (*Macaca Mulatta*, 11 and 16 kg), monkey T and monkey H, served as experimental subjects. Each animal was surgically implanted with a titanium head post, and a cylindrical titanium recording chamber (30 mm diameter). In each animal, the placement of the recording chamber was centered at ~17 mm from the midline and ~7 mm behind ear-bar-zero, and craniotomy was performed, allowing access to multiple visual areas in the superior temporal sulcus (STS). All surgeries were conducted using aseptic techniques under general anesthesia and analgesics were provided during post-surgical recovery.

We measured visual RFs by randomly presenting a single-probe stimulus out of either a 7 (H) * 11 (V) probe grid extending 18 (H) * 30 (V) degree of visual angles (dva)(monkey T), or a 14 (H) *17 (V) probe grid extending 26 (H) * 32 (V) dva (monkey H). The probes consisted of a circular drifting Gabor gratings (2 deg in diameter, 0.5 cycle/deg in spatial frequency, 4 deg/sec in speed, 100% Michelson contrast) and was presented for a duration of 0.1 sec. Monkeys were recorded with a drop of juice if they maintained fixating at the fixation spot throughout the trial.

For a given probe location, we obtained the neuronal activity by counting all the spikes during the stimulus presentation period, accounted for by a time delay of 50 ms. Neuronal RFs were defined as the probe locations that elicited more than 90% of the peak visual responses.

Motor cortex recordings and analysis

Details of the pacman behavioral task and experimental hardware presented in Fig. 3 are described in [54]. Three monkeys (*Macaca Mulatta)* served as experimental subjects. In each, a head post and recording chamber were implanted over premotor and primary motor cortex using aseptic surgical procedures and general anesthesia. Placement of the chambers were guided using structural MRI. Recordings in Monkey C were performed using a standard 19 mm plastic recording chamber (Christ Inc.), while Monkey I and J were implanted with custom low-profile footed titanium chambers (Rogue Research).

We conducted 13 sessions in Monkey C, and 10 sessions in Monkey I, targeting sulcal and gyral M1 and PMd. On a subset of 15 sessions in Monkey C, we also targeted GPi in the basal ganglia. In Monkey J, we report data from one session while simultaneously recording in GPi, SMA, and M1.

For monkey C, the Neuropixels 1.0-NHP probe was held using a standard 0.25” dovetail mount rod with a custom adapter to mount it to a hydraulic drive (Narishige Inc.). A 21 gauge blunt guide tube, 25 mm in length, was held using a custom fixture and placed over the desired recording location. The dura was then penetrated with a tungsten electrode (FHC, size E), which was bent at 27 mm to prevent the tip from inserting further than 2 mm past the end of the guide tube. This electrode was inserted manually via forceps, once or several times, as necessary, which also provided feedback on the depth and difficulty of penetrating the dura. The Neuropixels 1.0-NHP probe was then aligned using the Narishige tower XY stage, lowered into the guide tube, and carefully monitored to ensure that the tip of the probe was aligned with the dural penetration. This procedure sometimes took several attempts to find the correct insertion point, but generally was successful in less than a few minutes.

For Monkeys I and J, the Neuropixels 1.0-NHP probe was held using a custom fixture mounted to a linear rail bearing (IKO Inc.). This apparatus was designed to enable close packing of many probes, and to solve the challenge of precisely targeting structures deep in the brain without trial and error. The linear rail is mounted in a custom 3D printed base, which mounts directly to the recording chamber. The geometry of the 3D printed base component determines the insertion trajectories and prevents mechanical interference between probes and the chamber. This base also provides support for either sharp or blunt guide tubes, as required. In general, blunt guide tubes were preferred, but if necessary sharp guide tubes were sometimes used when the dura had become thicker and difficult to penetrate. The linear bearing was connected to a commercial drive system (NAN Inc.) via a ~50 mm long, .508 mm stainless steel wire, which provided rigid connection between the Nan drive electrode mount and the Neuropixels probe mounted on the rail bearing, while allowing a small amount of

misalignment between the drive axis and the insertion axis. This apparatus greatly simplifies the procedure of using many probes in a small space, while not relying on commercial drives to provide the mechanical rigidity required to safely insert a delicate probe. Additional details on the custom hardware is provided in the Neuropixels 1.0-NHP user wiki:

https://github.com/etrautmann/Neuropixels-NHP-hardware.

Spike sorting was performed using Kilosort 2.5 and manually curated using Phy. PCA trajectories were calculated after smoothing spikes with a 25 ms gaussian kernel and averaging across successful trials. Offline force model prediction performance was computed using a 50 ms time lag between arm force and neural activity. Neurons were randomly sub-selected, and 80% of trials from six target conditions were used to train a linear regression model in Python using scikit-learn, while the remaining 20% of trials were used to calculate model performance. Ten iterations were performed for each level of neurons retained. Details of the current source density analysis are presented further below.

LIP recordings and analysis

Details of the collection and analysis of the data presented in Figure 5 are described in[46]. Two monkeys (*Macaca Mulatta*, 8–11 kg) served as experimental subjects. In each, a head post and two recording chambers were implanted using aseptic surgical procedures and general anesthesia. Placement of the LIP chamber was guided by structural MRI. The SC chamber was placed on the midline and angled back 38° from vertical in the anterior-posterior axis.

We conducted eight recording sessions in which activity in LIP and SC was recorded simultaneously. In LIP we used a single Neuropixels 1.0-NHP probe, yielding 54–203 single units per session. In SC, we used 16-, 24-, and 32-channel V-probes (Plexon) with 50–100 µm electrode spacing, yielding 13–36 single units per session. In each session, we first lowered the SC probe and approximated the response fields (RFs) of SC neurons using a few dozen trials of a delayed saccade task. Because our penetrations were approximately normal to the retinotopic map in SC, the RFs of the SC neurons were highly similar within a session. We proceeded only if the center of the RFs were at least 7° eccentric in order to ensure minimal overlap with the motion stimulus.

If the RF locations in SC were suitable, we then lowered the Neuropixels probe into LIP through a dura-penetrating, stainless steel guide tube (23G) at 5 µm/s using a MEM microdrive (Thomas Recording) that was attached to a chamber-mounted, three-axis micromanipulator. Custom-designed adapters were used for mounting the Neuropixels probe onto the drive (see wiki). Once the target depth was reached (~10 mm below the dura), we allowed for 15–30 minutes of settling time to facilitate recording stability. In order to precisely measure RF locations in both areas, the monkeys performed 100–500 trials of the delayed saccade task and LIP neurons with RFs that overlapped those of the SC neurons were identified post-hoc. Finally, the monkeys performed a reaction-time RDM motion discrimination task until satiated (typically 1,500–3,000 trials).

Neurons in both areas were sorted using Kilosort 2.0 and manually curated in Phy. We restricted our analysis of the LIP data to neurons with RFs that overlapped those of the simultaneously recorded SC neurons. (164 of 1,084 total LIP neurons). Spike trains were discretized into 1 ms bins and convolved with a Gaussian kernel (σ = 25 ms) to produce the single-trial activity traces depicted in Figure 5d and e.

Face patch recordings and analysis

Three monkeys (*macaca mulatta*) served as experimental subjects. Each animal was surgically implanted with an MRI-compatible Ultem head post, and a large rectangular recording chamber (61 x 46 mm diameter, 65 x 50 mm diameter, and 58 x 61 mm diameter respectively), covering most of the animal's acrylic implant. Monkeys were trained to passively fixate on a spot for juice reward while visual stimuli of 5° size, such as images of faces or objects, were presented on an LCD screen (Acer). We targeted face patches ML and AF (monkey 1), face patches MF and AL (monkey 2), and face patches MF and AM (monkey 3) in the IT cortex for electrophysiological recordings. Face patches were identified using fMRI. Monkeys were scanned in a 3T scanner (Siemens), as described previously[55]. MION contrast agent was injected to increase signal-to-noise ratio. During fMRI, monkeys passively viewed blocks of faces and blocks of other objects to identify face-selective patches in the brain.

Before targeting fMRI-identified face patches with Neuropixels probes, we performed scout recordings with tungsten electrodes with 1 MΩ impedance (FHC) using grids designed with the software Planner[56]. While inserting tungsten electrodes, we performed structural MRI scans to confirm correct targeting (Fig. 4a). Subsequently, we performed a total of 72 Neuropixels insertions. In order to perform very deep recordings (e.g., 42 mm from the craniotomy, Fig. 4d), we lowered a cannula holder to touch or gently push the dura. The cannula holder contained a short cannula to penetrate the dura. A probe holder, that held the probe, was slid through the cannula holder through matching dovetails. This dovetail mechanism was designed to ensure that the direction of probe movement matched the direction of the cannula, as even small differences in angles would risk breakage of the probe when inserted deeply into the cannula. The probe holder was advanced using an oil hydraulic micromanipulator (Narishige), but importantly, the precise direction of probe movement was constrained by the dovetail between probe holder and cannula holder rather than the micromanipulator.

Neuropixels data were recorded using SpikeGLX and OpenEphys, and spikes were sorted using Kilosort 3.0. To compute responses for Fig. 4c,e average spike rates from 50 ms to 250 ms after trial onset were computed, and baselines, averaged from 0 ms to 50 ms after trial onset, were subtracted.

Functional connectivity via cross-correlation analysis

Functional interactions between pairs of neurons were measured with an established cross-correlation method. Cross-correlogram (CCG) were calculated using spike trains from pairs of simultaneously recorded neurons, either during the whole stimulus presentation period or during the inter-trial-intervals. The CCG is defined as:

$$CCG(\tau) = \frac{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N} x_1^i(t)x_2^i(t+\tau)}{\theta(\tau)\sqrt{\lambda_1\lambda_2}}$$

Where M is the number of trials, N is the number of time bins within a trial, x_1^i and x_2^i are the spike trains of neuron 1 and 2 from trial i, τ is the time lag relative to the reference spikes, and λ_1 and λ_2 are the mean firing rate of the two neurons, respectively. $\theta(\tau)$ is a triangular function calculated as $\theta(\tau) = N - |\tau|$ that corrects for the overlapping time bins at different time lags. A jitter corrected method was used to remove correlations caused by stimulus-locking or slow fluctuations.

$$CCG_{jitter\ corrected} = CCG_{original} - CCG_{jittered}$$

Where $CCG_{original}$ and $CCG_{jittered}$ are CCGs calculated with the above equation using original dataset and dataset with spike timing randomly perturbed (jittered) within the jitter window, respectively. The correction term ($CCG_{jittered}$) captured slow correlation longer than the jitter window (caused by common stimulation or slow fluctuation in the population response), thus once it's subtracted, only the fine temporal correlation is preserved. A 25-ms jitter window was chosen based on previous studies[6].

As with previous studies[6,11], here a CCG is classified as significant if the peak of jitter-corrected CCG occurred within 10ms of zero time lag and if this peak is more than 7 standard deviations above the mean of the noise distribution (CCG flank).

Current source density analysis and session alignment

For each session, we first flagged bad LFP channels by following the approach described in the International Brain Lab (IBL) spike sorting pipeline[57]. Briefly, we identify dead channels, with unusually low similarity to the common average reference (CAR) signal, and noisy channels with high spectral power above 0.8 of the Nyquist frequency or low similarity with the high pass filtered (CAR). We low-pass filtered the LFP using a zero-phase, fifth order Butterworth filter with 25 Hz corner frequency. We also corrected in the frequency domain for the relative sampling offsets among the channels sharing a common ADC on the probe, as also described in the IBL pipeline. We then extracted LFP signals aligned to the onset of force production for each trial. We infilled the bad channels using kriging interpolation, using weights decreasing

with distance d as $w(d) \propto e^{-(d/d_0)^p}$ where d_0 = 20 μm and p = 1.3. We computed the event related potential (ERP) as the average of the aligned LFP signals across trials within each force profile condition and the average over all trials as well. We then resampled the ERP spatially along a vertical column running parallel with the center of the probe, sampled every 10 μm, again using kriging. To compute the current source density (CSD) for each condition individually and across all trials, we computed the second spatial derivative of the ERP using a smoothing, differentiating Savitzy Golay filter of second order, with a 450 μm frame size. To align the CSDs across sessions, we mean squared error as the loss function, i.e. normalized by the number of overlapping rows in the CSD. For each pair of the sessions, the loss function used the set of shared conditions collected in both experimental sessions and concatenated the CSDs for these shared conditions in time. If no conditions were present in both sessions, the grand average CSDs were used. We then found the optimal alignment jointly across all sessions by performing constrained optimization using a genetic algorithm.

References

1. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).
2. Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, (2021).
3. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).
4. Peters, A. J., Fabre, J. M. J., Steinmetz, N. A., Harris, K. D. & Carandini, M. Striatal activity topographically reflects cortical activity. *Nature* **591**, 420–425 (2021).
5. Jia, X. *et al.* High-density extracellular probes reveal dendritic backpropagation and facilitate neuron classification. *J. Neurophysiol.* **121**, 1831–1847 (2019).
6. Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).
7. Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* **364**, 253 (2019).
8. Vesuna, S. *et al.* Deep posteromedial cortical rhythm in dissociation. *Nature* **586**, 87–94 (2020).
9. Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).
10. Trautmann, E. M. *et al.* Accurate Estimation of Neural Population Dynamics without Spike Sorting. *Neuron* **103**, 292–308.e4 (2019).
11. Trepka, E. B., Zhu, S., Xia, R., Chen, X. & Moore, T. Functional interactions among neurons within single columns of macaque V1. *Elife* **11**, (2022).

12. Sun, X. *et al.* Cortical preparatory activity indexes learned motor memories. *Nature* **602**, 274–279 (2022).
13. O'Shea, D. J. *et al.* Direct neural perturbations reveal a dynamical mechanism for robust computation. *bioRxiv* 2022.12.16.520768 (2022) doi:10.1101/2022.12.16.520768.
14. Paulk, A. C. *et al.* Large-scale neural recordings with single neuron resolution using Neuropixels probes in human cortex. *Nat. Neurosci.* **25**, 252–263 (2022).
15. Chung, J. E. *et al.* High-density single-unit human cortical recordings using the Neuropixels probe. *Neuron* (2022) doi:10.1016/j.neuron.2022.05.007.
16. Maynard, E. M., Nordhausen, C. T. & Normann, R. A. The Utah intracortical Electrode Array: a recording structure for potential brain-computer interfaces. *Electroencephalogr. Clin. Neurophysiol.* **102**, 228–239 (1997).
17. Musallam, S., Bak, M. J., Troyk, P. R. & Andersen, R. A. A floating metal microelectrode array for chronic implantation. *J. Neurosci. Methods* **160**, 122–127 (2007).
18. Schwarz, D. A. *et al.* Chronic, wireless recordings of large-scale brain activity in freely moving rhesus monkeys. *Nat. Methods* **11**, 670–676 (2014).
19. Dotson, N. M., Hoffman, S. J., Goodell, B. & Gray, C. M. A Large-Scale Semi-Chronic Microdrive Recording System for Non-Human Primates. *Neuron* **96**, 769–782.e2 (2017).
20. Mao, D. *et al.* Spatial modulation of hippocampal activity in freely moving macaques. *Neuron* **109**, 3521–3534.e6 (2021).
21. Trautmann, E. M. *et al.* Dendritic calcium signals in rhesus macaque motor cortex drive an optical brain-computer interface. *Nat. Commun.* **12**, 3689 (2021).
22. Bollimunta, A. *et al.* Head-mounted microendoscopic calcium imaging in dorsal premotor cortex of behaving rhesus macaque. *Cell Rep.* **35**, 109239 (2021).

23. Rominger, J. P. & Lin, B. J. Seamless Stitching For Large Area Integrated Circuit Manufacturing. *SPIE Proceedings* Preprint at https://doi.org/10.1117/12.968412 (1988).
24. Lin, B. J. The Paths To Subhalf-Micrometer Optical Lithography. in *Optical/Laser Microlithography* vol. 0922 256–269 (SPIE, 1988).
25. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Kilosort, H. K. D. realtime spike-sorting for extracellular electrophysiology with hundreds of channels. bioRxiv. Preprint at (2016).
26. Felleman, D. J. & Van Essen, D. C. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex* vol. 1 1–47 Preprint at https://doi.org/10.1093/cercor/1.1.1 (1991).
27. Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B. & Tootell, R. B. H. Faces and objects in macaque cerebral cortex. *Nat. Neurosci.* **6**, 989–995 (2003).
28. Orban, G. A., Van Essen, D. & Vanduffel, W. Comparative mapping of higher visual areas in monkeys and humans. *Trends Cogn. Sci.* **8**, 315–324 (2004).
29. Gattass, R. & Gross, C. G. Visual topography of striate projection zone (MT) in posterior superior temporal sulcus of the macaque. *J. Neurophysiol.* **46**, 621–638 (1981).
30. Maguire, W. M. & Baizer, J. S. Visuotopic organization of the prelunate gyrus in rhesus monkey. *J. Neurosci.* **4**, 1690–1704 (1984).
31. Gattass, R., Sousa, A. P. & Gross, C. G. Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of neuroscience* **8**, 1831–1845 (1988).
32. Dum, R. P. & Strick, P. L. Motor areas in the frontal lobe of the primate. *Physiol. Behav.* **77**, 677–682 (2002).
33. Rathelot, J.-A. & Strick, P. L. Subdivisions of primary motor cortex based on

cortico-motoneuronal cells. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 918–923 (2009).

34. Scott, S. H. The computational and neural basis of voluntary motor control and planning. *Trends Cogn. Sci.* **16**, 541–549 (2012).
35. Marshall, N. J. *et al.* Flexible neural control of motor units. *bioRxiv* 2021.05.05.442653 (2022) doi:10.1101/2021.05.05.442653.
36. Nicholson, C. & Freeman, J. A. Theory of current source-density analysis and determination of conductivity tensor for anuran cerebellum. *J. Neurophysiol.* **38**, 356–368 (1975).
37. Mitzdorf, U. Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. *Physiol. Rev.* **65**, 37–100 (1985).
38. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
39. Tsao, D. Y., Moeller, S. & Freiwald, W. A. Comparing face patch systems in macaques and humans. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19514–19519 (2008).
40. Hesse, J. K. & Tsao, D. Y. The macaque face patch system: a turtle's underbelly for the brain. *Nat. Rev. Neurosci.* **21**, 695–716 (2020).
41. Chang, L. & Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013–1028.e14 (2017).
42. Hesse, J. K. & Tsao, D. Y. A new no-report paradigm reveals that face cells encode both consciously perceived and suppressed stimuli. *Elife* **9**, (2020).
43. Shadlen, M. N. & Kiani, R. Decision making as a window on cognition. *Neuron* **80**, 791–806 (2013).
44. Ratcliff, R. & McKoon, G. The diffusion decision model: theory and data for two-choice

decision tasks. *Neural Comput.* **20**, 873–922 (2008).

45. Steinemann, N. A. *et al.* Direct observation of the neural computations underlying a single decision. *bioRxiv* 2022.05.02.490321 (2022) doi:10.1101/2022.05.02.490321.
46. Stine, G. M., Trautmann, E. M., Jeurissen, D. & Shadlen, M. N. A neural mechanism for terminating decisions. *bioRxiv* 2022.05.02.490327 (2022) doi:10.1101/2022.05.02.490327.
47. Jia, X. *et al.* Multi-regional module-based signal transmission in mouse visual cortex. *Neuron* **110**, 1585–1598.e9 (2022).
48. Kohn, A. & Smith, M. A. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J. Neurosci.* **25**, 3661–3673 (2005).
49. DeAngelis, G. C., Ghose, G. M., Ohzawa, I. & Freeman, R. D. Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons. *J. Neurosci.* **19**, 4046–4064 (1999).
50. Ts'o, D. Y., Gilbert, C. D. & Wiesel, T. N. Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *J. Neurosci.* **6**, 1160–1170 (1986).
51. Windolf, C. *et al.* Robust Online Multiband Drift Estimation in Electrophysiology Data. *bioRxiv* 2022.12.04.519043 (2022) doi:10.1101/2022.12.04.519043.
52. Lee, E. K. *et al.* Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex. *eLife* vol. 10 Preprint at https://doi.org/10.7554/elife.67490 (2021).
53. Karsh, B. & Others. SpikeGLX. Preprint at (2016).
54. Marshall, N. J. *et al.* Flexible neural control of motor units. *Nat. Neurosci.* **25**, 1492–1504 (2022).

55. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).

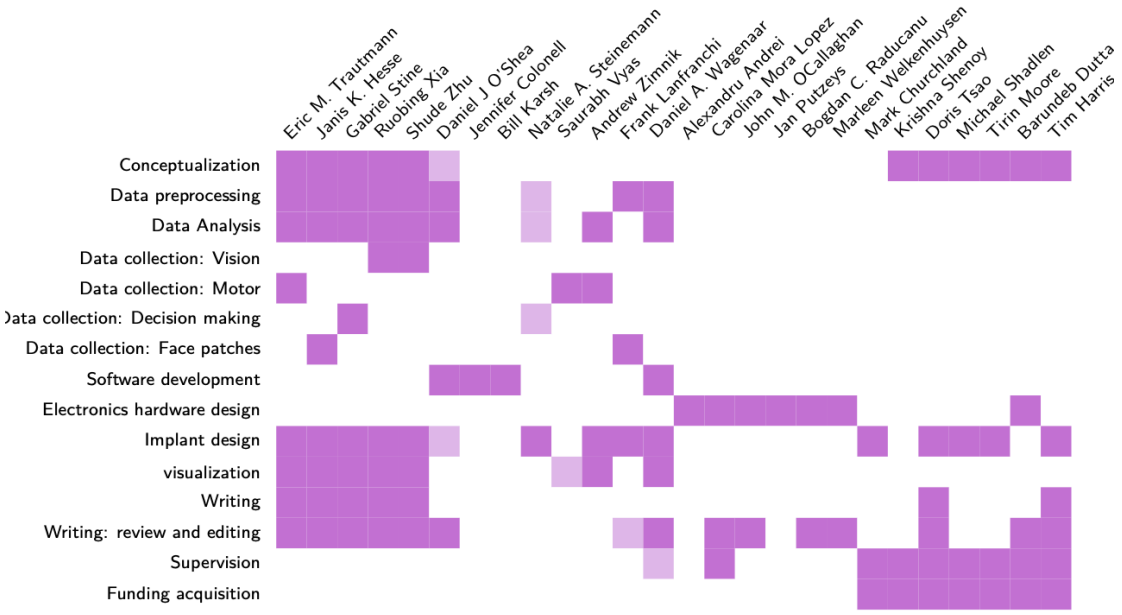
56. Ohayon, S. & Tsao, D. Y. MR-guided stereotactic navigation. *J. Neurosci. Methods* **204**, 389–397 (2012).

57. International Brain Laboratory *et al. Spike sorting pipeline for the International Brain Laboratory*. https://figshare.com/articles/online_resource/Spike_sorting_pipeline_for_the_International_Brain_Laboratory/19705522 (2022) doi:10.6084/m9.figshare.19705522.v1.

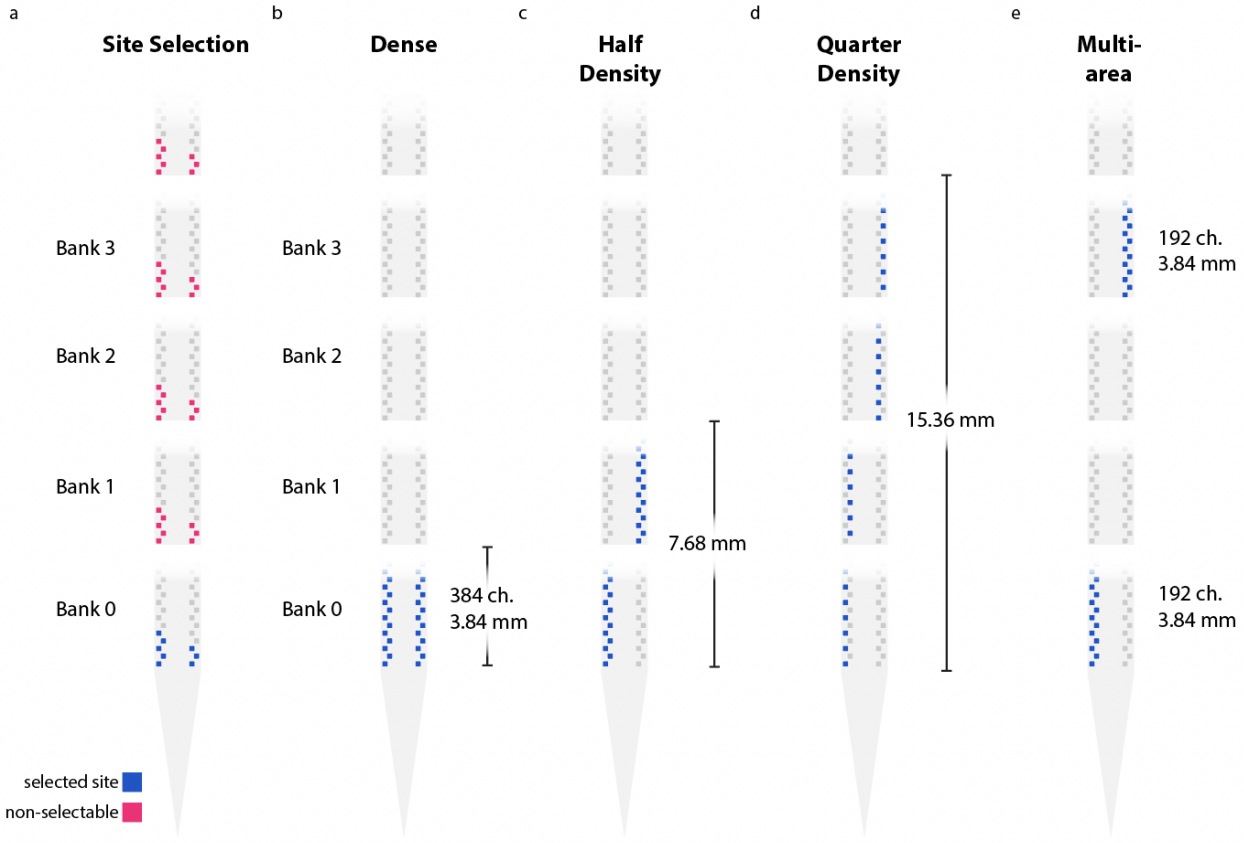
Acknowledgements

We thank the support of the Howard Hughes Medical Institute, who funded the development of the probe. We thank Yanina Pavlova, Danielle Abreu Lopes, Stephen Cital, Cornel Duhaney, Brian Madeira, Mackenzie Risch, and Michelle Wechsler for surgical assistance and expert veterinary care. for their assistance in the planning and execution of surgeries, animal training and general support, and Stephen Ryu for surgical expertise. We thank Bob Schneeveis and Tanya Tabachnik for engineering assistance. In addition, we thank Columbia University's ICM for the quality of care they provide for our animals, especially during the pandemic and lockdown. We thank Wei-lung Sun, for probe testing and software development (HHMI Janelia). E.M.T. is supported by the Grossman center and the Brain and Behavior Research Foundation. S.V is supported by NIH NRSA NINDS F32, N.A.S is supported by NIH Brain Initiative (MR01NS113113). T.M. is supported by EY014924, NS116623. A.Z. is supported by the American Parkinson Disease Post-Doctoral Fellowship. D.J.O is supported by SCGB (543045).

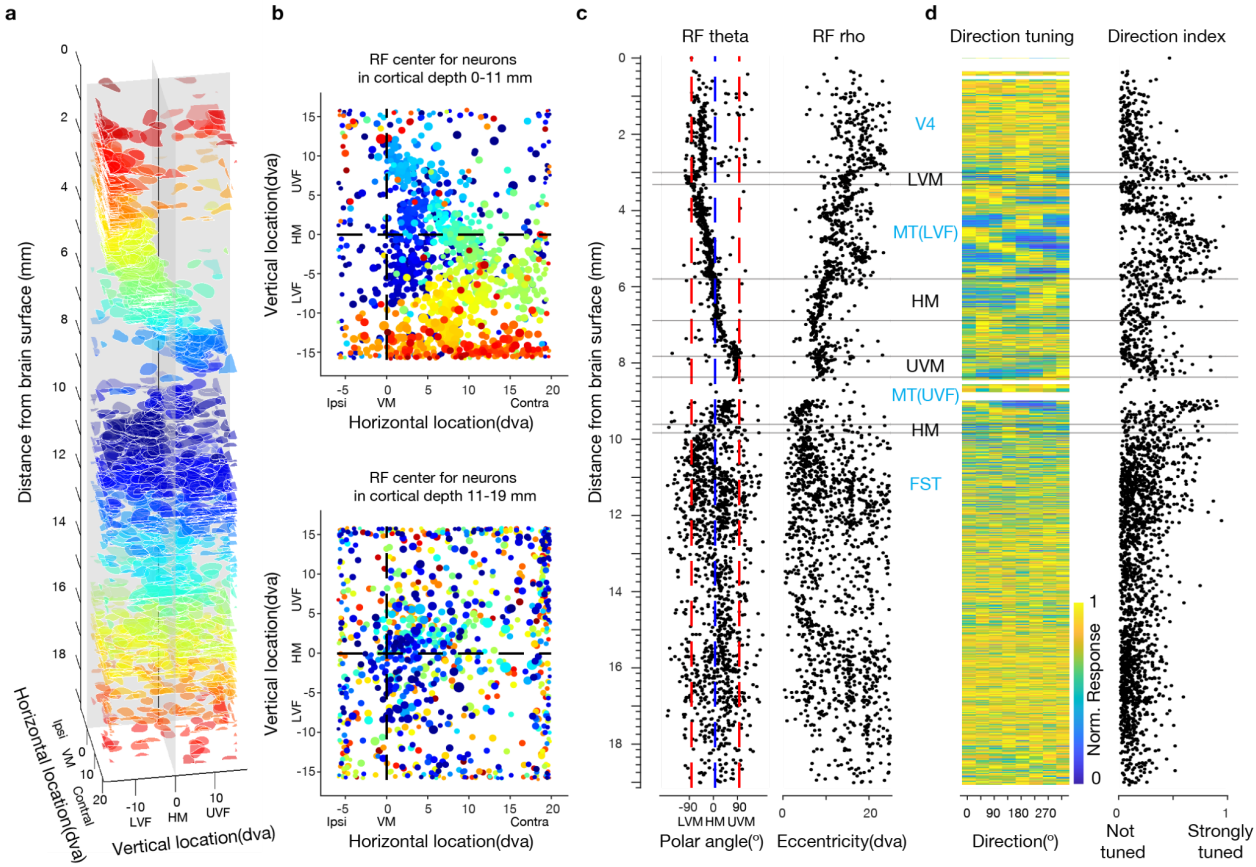
Author Contributions



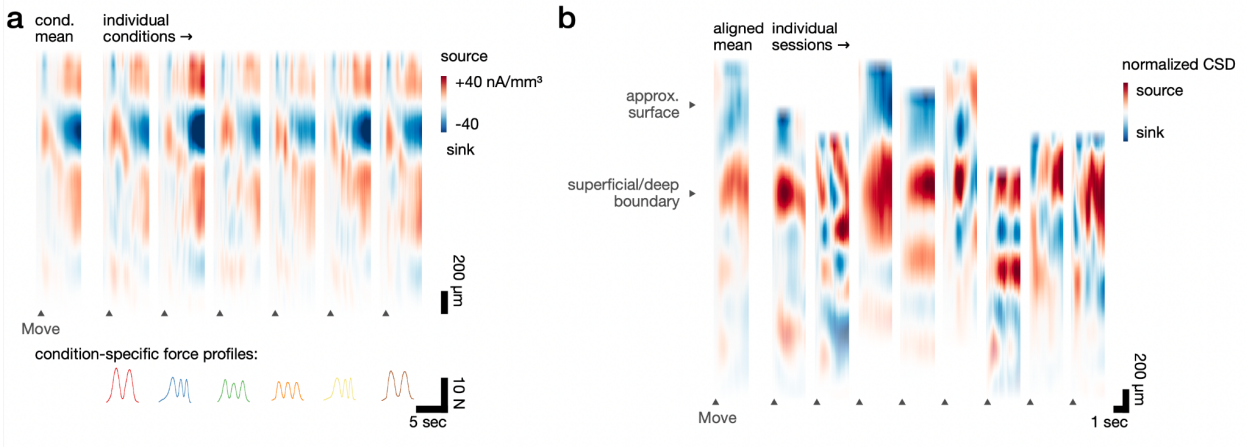
Extended data



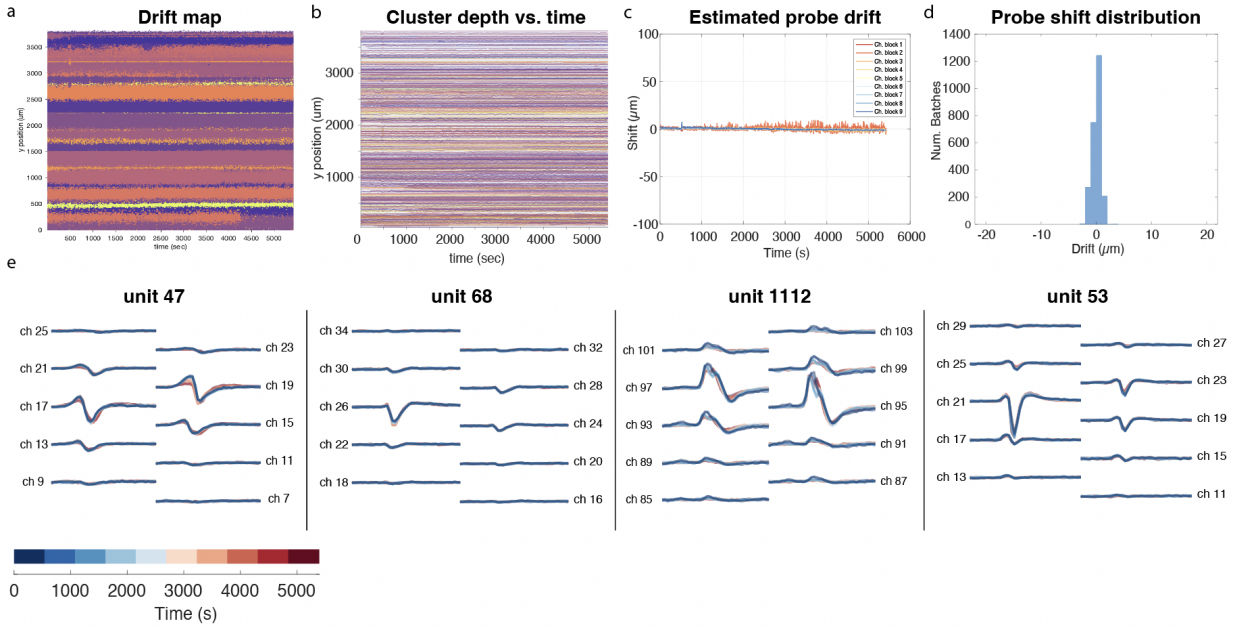
Extended Data Figure 1. Site selection rules and common configurations. a, Site selection rule - any electrode selected on any bank is unavailable for selection in other banks. Note that multiple electrodes can be connected to a single readout channel, as explored by [2], but this impacts the signal to noise and detection for small units, and is a specialized method and not currently in widespread use. **b,** Dense recording from Bank 0, places 384 channels spanning 3.84 mm. **c,** half density configuration covering Banks 0 and 1, spanning 7.68mm. **d,** Quarter density, covering banks 0–3, spanning 15.36 mm. **e,** multi-area recording.



Extended Data Figure 2. Retinotopic organization and functional properties of single neurons across multiple visual areas. a, Distribution of receptive fields (RFs) of 2729 visually responsive neurons across cortical depth from the 5-banks recording across multiple visual areas. Color scale represents cortical depth. **b,** Top-view of **a,** illustrating the progression of RFs across visual fields. RFs from the superficial and deeper part of the brain are demonstrated separately for clarity. **c**, Polar angle (theta) and Eccentricity (rho) of each RF's geometric center across cortical depth. **d**, Left, heat map of evoked responses across drift directions of grating (vertical thickness is greater for less dense neuronal population). Color scale represents the magnitude of evoked responses. Right, direction index as quantified by the differences of responses to the preferred and its opposite direction divided by the sum of the two. In **c** and **d**, each neuron is plotted at its corresponding cortical depth. Horizontal lines denote the section of cortex where the center of RFs falls on Lower vertical meridians (LVM), horizontal meridian (HM), upper vertical meridian (UVM), and horizontal meridian (HM), respectively, superficial to deep. Putative visual areas are identified and labeled. LVF: lower visual field; UVF: upper visual field; FST: fundus of the superior temporal (FST) area.



Extended Data Figure 3. Aligning recordings to cortical lamina using current source density. a, Single session current source density (CSD) plot during motor behavior. (top left) Mean CSD plot over all behavioral conditions, time locked to movement onset. (top right panels) Single-condition CSD plots illustrating spatial and temporal dynamics of CSD during behavior. (bottom row) Trial-averaged force behavior during pacman task for each condition. **b,** Probe locations from Individual recording sessions aligned using CSD to infer recording depth relative to other sessions.



Extended Data Figure 4. Acute recording stability a, Drift map of individual spikes, colored according to clusters identified via Kilosort 2.5. **b,** Depth estimate of each cluster over time within a recording session. **c,** Tissue drift relative to the probe estimated using Kilosort 2.5. Different traces indicate different blocks of channels on the probe. **d,** Distribution of probe drift estimates across all alignment batches. **e,** Neuron waveforms for four units identified using Kilosort 2.5. Individual traces represent median waveform calculated using 100 randomly spikes during one of 10 blocks during the session, with block indicated by line color. Data for panels a-e collected from Monkey I in the motor cortex during a motor behavioral experiment.