Development of Extracellular Electrophysiology Methods for Scalable Neural Recording

by

Jacob G. Bernstein

S.B., Massachusetts Institute of Technology (2007)
S.M., Massachusetts Institute of Technology (2009)

Submitted to the Program in Media Arts and Sciences
School of Architecture + Planning
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

© Massachusetts Institute of Technology 2016. All rights reserved.
DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

Pages contain inkjet marks/speckeling
Development of Extracellular Electrophysiology Methods for Scalable Neural Recording

by Jacob G. Bernstein

Submitted to the Program in Media Arts and Sciences
School of Architecture + Planning
on September 1, 2016, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences

Abstract

In order to map the dynamics of neural circuits in mammalian brains, there is a need for tools that can record activity over large volumes of tissue and correctly attribute the recorded signals to the individual neurons that generated them. High-resolution neural activity maps will be critical for the discovery of new principles of neural coding and neural computation, and to test computational models of neural circuits. Extracellular electrophysiology is a neural recording method that has been developed to record from large populations of neurons, but well-known problems with signal attribution pose an existential threat to the viability of further system scaling, as analyses of network function become more sensitive to errors in attribution. A key insight is that blind-source separation algorithms such as Independent Component Analysis may ameliorate problems with signal attribution. These algorithms require recording signals at much finer spatial resolutions than existing probes have accomplished, which places demands on recording system bandwidth. We present several advances to technologies in neural recording systems, and a complete neural recording system designed to investigate the challenges of scaling electrophysiology to whole brain recording. We have developed close-packed microelectrode arrays with the highest density of recording sites yet achieved, for which we built our own data acquisition hardware, developed with a computational architecture specifically designed to scale to over several orders of magnitude. We also present results from validation experiments using colocalized patch clamp recording to obtain ground-truth activity data. This dataset provides immediate insight into the nature of electrophysiological signals and the interpretation of data collected from any electrophysiology recording system. This data is also essential in order to optimize probe development and data analysis algorithms which will one day enable whole-brain activity mapping.

Thesis Supervisor: Dr. Edward S. Boyden, III
Title: Associate Professor of Media Arts and Sciences
Associate Professor of Brain and Cognitive Sciences
Associate Professor of Biological Engineering
This following served on the examination committee for this dissertation:

Signature redacted

Reader...

Dr. Nancy Kopell
Professor
Co-Director, Center for BioDynamics
Department of Mathematics
Boston University

Signature redacted

Reader....

Dr. Joseph Paradiso
Associate Professor
Co-Director, Things That Think Consortium
MIT Media Lab
Massachusetts Institute of Technology
# Contents

1 **Introduction** ............................................. 13
   1.1 Problem Statement ..................................... 13
   1.2 Introduction ........................................... 14
   1.3 Background ............................................. 16
      1.3.1 Spike Sorting - Clustering and ICA .............. 17
   1.4 Outline of Chapters ................................... 18

2 **Hardware for Scalable Electrophysiology** ........... 23
   2.1 Introduction ........................................... 23
   2.2 Results ................................................ 24
      2.2.1 System Design ..................................... 25
         Probes ................................................. 25
         Headstages .......................................... 27
         Data Acquisition ................................... 29
   2.3 Discussion .............................................. 32
   2.4 Methods ................................................ 36
      2.4.1 Probes ............................................. 36
      2.4.2 Headstages ....................................... 38
      2.4.3 Data Acquisition ................................ 38
   2.5 Contribution .......................................... 42

3 **Spike Sorting with Close-Packed Electrode Arrays** .... 45
   3.1 Introduction ........................................... 45
3.2 Results

3.2.1 ICA on Ground-Truth Recordings

3.2.2 Analysis of Multi-Unit Independent Components

3.3 Discussion

3.4 Methods

3.4.1 Surgical Procedures

3.4.2 Colocalized Electrophysiology Recordings

3.4.3 Data Acquisition and Analysis Pipeline

3.4.4 Pipette Position Estimation

3.4.5 Spike Identification in Patch Signal and Burst Exclusion

3.4.6 Filtering and Artifact Subtraction in MEA Signal

3.4.7 ICA Implementation

3.4.8 Matching Independent Components to Patched Neuron

3.4.9 Feature Extraction and Discriminant Analysis

3.5 Contribution

4 Devices for 3D Neuromodulation

4.1 Introduction

4.2 Results

4.2.1 Fiber array design, fabrication, and operation

4.2.2 Implantation, utilization, and validation of fiber arrays in behaving mice

4.2.3 Integration of electrophysiological recording and fiber array illumination

4.3 Discussion

4.4 Methods

4.4.1 LED and Fiber-coupled LED Array Fabrication: Design and Preparation of Key Structural Components

4.4.2 Fiber Array Fabrication: Preparation of Other Components

4.4.3 Fiber Array Fabrication: Assembly

4.4.4 LED Driver Circuit Design and Operation

4.4.5 Fiber Array Testing: Light Power
4.4.6 Fiber Array Testing: Thermal Testing In Vivo ......................... 93
4.4.7 Fiber Array Fabrication: Optimization for Electrophysiology ...... 93
4.4.8 Electrophysiological Recording and Data Analysis .................... 94
4.4.9 Fiber Array Fabrication: Accessories for Surgery Practice and Surgery
   Facilitation ................................................................. 95
4.4.10 Fiber array implantation ............................................. 96
List of Figures

2-1 Photograph of 1000-channel data acquisition system .................. 25
2-2 Close-packed electrodes manufactured on a thin silicon substrate .... 26
2-3 First generation 512-Channel signal-conditioning headstage ............. 28
2-4 Implementation of direct communication design of data acquisition (DAQ) module .................................................................................. 31
2-5 Methods for scaling and integrating probes and headstages ............... 34
2-6 Schematic for using multiple data acquisition systems in parallel ....... 35
2-7 Block diagram of the FPGA circuitry ..................................................... 39
3-1 Colocalized patch-clamp and multielectrode array recording ................ 48
3-2 Multi-unit component containing spikes from patched neuron .......... 51
3-3 Comparison of clustering on electrode and IC waveforms ............... 52
3-4 Data Analysis Flow Chart .................................................................... 59
3-5 Estimate of pipette position by curve fit to $\frac{1}{r}$ falloff ..................... 61
3-6 Probe track across complete neuron hunt ........................................... 62
3-7 Comparison of intracellular and extracellular waveform during a burst . 64
3-8 Impact of ISI on burst spike amplitude .............................................. 65
3-9 Effect of filter parameters on extracellular spike waveform ............ 66
3-10 Motion artifact subtraction ................................................................. 67
3-11 Comparison of median TP and FP waveforms on MEA for neuron shown in Figures 3-2 and 3-3 ................................................................. 69
3-12 Comparison of error rates vs. number of predictors for linear discriminants across three sets of features ......................................................... 70
3-13 QQ-plots of neuron clusters in IC space and MEA feature space ............... 71

4-1 Design and fabrication of optical fiber arrays ............................................. 79
4-2 Accessory devices, and modifications, for implantation and utilization of fiber
arrays ........................................................................................................... 82
4-3 A custom coaxial cable design enables decoupled LED activation and recording 83
4-4 Thermal characteristics of fiber array use ..................................................... 85
Chapter 1

Introduction

1.1 Problem Statement

There is great interest among neuroscientists in recording the activity of as many neurons as possible in living mammalian brains[1], in order to advance research in neural coding and computation[2–4], as well as improve assistive technologies such as brain machine interfaces[5, 6]. In order to map the dynamics of neural circuits in mammalian brains, there is a need for tools that can record activity over large volumes of tissue and correctly attribute the recorded signals to the individual neurons that generated them. Extracellular electrophysiology, a method for sensing the electrical currents generated in the brain by neural activity, has been neuroscience’s primary method of observing the activity of large populations of neurons, yet after decades of tool development, modern tools fall short of the demands of neural circuit mapping by several orders of magnitude[7]. Additionally, well-known problems with signal attribution pose an existential threat to the viability of further system scaling, as analyses of network function become more sensitive to errors in attribution[8], and alternative technologies, such as optical sensing, receive greater attention in the tool development community. However, electrophysiology maintains certain advantages over alternative methods, especially as it is currently the only method approved and suited for human brain-machine interfaces. As these tools are currently being used for restoration of function in quadriplegia[9] and treatment of neural disorders such as Parkinson’s disease[10], epilepsy[11], and severe depression[12, 13], and these therapies have much to benefit from
more complete mapping of neural circuit activity, it is of utmost importance to discover a path for scalable electrophysiology.

One of the primary challenges to developing electrophysiology tools is that neural recording systems are comprised of modules utilizing a wide range of engineering disciplines, including MEMS, microelectronics, electrochemistry, embedded electronics, software engineering, and data analysis, and tool development has traditionally been pursued within groups focused on individual modules within their domain of expertise. Unfortunately, interdependencies between system modules have limited the effectiveness of these approaches, and we believe tool development must be understood and directed at the level of complete recording systems in order to unlock the potential of modern engineering capabilities. A key insight is that blind-source separation algorithms such as Independent Component Analysis may ameliorate problems with signal attribution. These algorithms are dependent on recording signals at much finer spatial resolutions than existing probes have accomplished, and these probes place demands on the size and bandwidth of downstream signal conditioning and data acquisition modules. Fortunately, all of the modules in modern neural recording systems are built with electronics microfabrication methods which are the subject of continuous research and development by the electronics industry, so there is a great opportunity to utilize modern tools to achieve rapid increases in recording system scale, and there exists a dependable roadmap for the development of more advanced tools into the future.

1.2 Introduction

Electrophysiology tools have been instrumental in neuroscience research and clinical advancements for decades, including the discovery of fundamental physiological, computational, and coding properties of neurons\[14, 15\], and the treatment of injuries\[16\] and diseases\[13, 17\] of the nervous system. Electrophysiological recording systems have been consistently improved to meet the demands of the field, and measured by the number of neurons that can be recorded simultaneously, the capacity of advanced recording system has doubled roughly every seven years\[7\]. In the near future, clinicians would like to deploy advanced neural recording systems to enable therapeutic applications of brain-machine interfaces with
closed-loop control, such as decoding complex motor plans in the brains of persons with paralysis in order to control robotic limbs [5], and mapping, predicting, and disrupting the evolution of seizures in epileptic patients[6]. Many neuroscientists would even like to record entire mammalian brains one day[1-4], but at the current rate of development (extrapolating from [7]), it would take more than 100 years to record the activity of the approximately 100,000,000 neurons in the brain of an adult mouse[18]. Thus, understanding the obstacles and limitations to scaling neural recording systems has important implications for the development of neuroscience.

In electrophysiology recording systems, recording capacity is primarily determined by the number of electrodes that can be fabricated, implanted, and recorded, which is constrained by limitations on the size of probes that can be used in and around the brain. Recording the signal from an electrode also requires special amplifiers and data acquisition systems, which have the potential to impose bottlenecks in the signal chain. Technological approaches to these challenges have converged on a common solution: all of the hardware in the electrophysiology signal chain can be manufactured on silicon substrates, using fabrication systems developed for the microelectronics industry. Yet while the capabilities of microelectronics systems have double every two years according to Moore's law, neural recording system capacity has not kept pace. Thus, the scalability of neural recording systems must not currently be gated by the fabrication constraints of individual system components.

The scalability of neural recording systems is also determined by the data analysis processes used. The algorithms conventionally used for the first step in data analysis, spike sorting, are error prone, which leads to two major problems (see Section 1.3 for an overview of spike sorting challenges, and Section 1.3.1 for more details). First, dealing with the errors is a waste of effort. While automated algorithms may make fewer errors than human operators[19], most electrophysiologists use spike sorting algorithms that requires human interaction. This makes analysis time a limiting resource in neural recording experiments of many neurons. Second, the nature of conventional spike-sorting errors (in both automated and supervised systems) also leads to systematic biases in measures of neural activity, which compound exponentially in downstream analyses of correlated activity of multiple neurons [8, 20]. This negates much of the value of increasing neural recording capacity[21].
The art of scalability in the context of neural recording systems is deriving a strategy that increases both number and reliability of neurons recorded. The uniqueness of our approach to recording system development is that we have approached the problem of spike sorting from first principles, analyzing the assumptions incorporated into conventional spike sorting techniques, and identified an algorithm, Independent Components Analysis (ICA), that is designed to address the problems of conventional spike sorting. ICA places demands on the design of probes, amplifiers, and data acquisition systems, which we have analyzed to derive the optimal development path for neural recording system hardware.

1.3 Background

Extracellular electrophysiology is the method of inserting electrodes in neural tissue to sense the voltages generated by the activity of nearby neurons. The first step to analyzing data from electrophysiology recordings is to separate the signals measured by the electrodes into contributions from different neurons. Extracellular electrodes recording nearby neurons are like microphones recording voices in a crowded room. Current algorithms use features of spike waveforms to distinguish spikes coming from different neurons, the way a listener might distinguish multiple participants in a conversation by the sounds of their voices. Unfortunately, waveforms can be corrupted by co-occurring activity of other neurons, which lead to unavoidable failure modes with conventional spike sorting. For this reason, there is great interest in neural recording technologies that do not have this spike sorting problem.

Alternative methods of recording brain activity do exist, such as optical imaging of exogenous contrast agents that alter fluorescence in some way after a spike (e.g. by changing brightness). The purpose of using optical systems is to directly image individual neurons, avoiding the spike sorting problem altogether, although scattering in tissue presents other challenges. Optical recording methods are undergoing a great period of innovation, circumventing conventional imaging limitations through ingenious optical systems[22, 23], and constantly producing new contrast agents to overcome limitations of sensitivity, temporal resolution, brightness, or targeting specificity[24–26]. Optical recording systems are an extremely useful and complementary tool to electrophysiology. Yet for some applications,
the constraints of complex optical systems are too burdensome, or contrast agents haven’t achieved a necessary specification, or recordings need to be performed in human subjects. Therefore, it is critical to continue to develop electrophysiological recording as a platform for scalable neural recording.

1.3.1 Spike Sorting - Clustering and ICA

Action potentials are events triggered by excitatory input, inducing rapid movement of ions across cell membranes, creating an electrical signal which propagates down neural processes and causes neurotransmitter release at synaptic junctions. This movement of charges creates short, stereotyped voltage changes on electrodes with a large amplitude compared to background activity, called spikes. It is easy to detect when a neuron has spiked near an electrode just by thresholding the recording; however, attributing each spike recorded on an electrode to a particular neuron is a major challenge. This is because an electrode can detect spikes from neurons up to 100-150 microns away, a volume which may contain as many as 100 neurons. The standard method for sorting spikes to neurons is to decompose each spike into a set of features based on amplitude and shape, and then look at all of the spikes in this feature space, and distinguish neurons as clusters of spikes in this space. Spikes from different neurons may look very similar when recorded on a single electrode, so it helps to have multiple electrodes close enough to see some of the same spikes[27, 28].

Even so, it can be hard to decide where are the boundaries of clusters or even how many clusters there are, and even well-trained humans in the same lab disagree about how many neurons are in a dataset[29]. Automating spike sorting can help[19], but all clustering algorithms make mistakes, by some combination of splitting neurons into multiple units, merging multiple neurons into a single unit, and adding or subtracting spikes at the boundaries of clusters. Perfect cluster-based spike sorting is an impossible problem for many neurons recorded, because spike waveforms from individual neurons exhibit large variability[19], due to intrinsic variability and co-occurring activity, which often overwhelms the difference in spike waveforms between neurons that are close together and relatively far from a recording site. For these neurons, co-occurring activity is often the biggest problem for spike sorting.

Independent Component Analysis[30] (ICA) is a blind source separation technique we
applied to remove the influence of co-occurring activity on spike waveforms. Returning to
the metaphor of microphones in a crowded room, ICA works when there are many micro-
phones dispersed in the room. Each microphone picks up each speaker in the room with
a different amplitude, determined by distance from each speaker to each microphone. ICA
uses statistical methods to calculate the original signals coming from the speakers, as well
as the relative amplitudes of all the speakers and microphones (and thus their relative posi-
tions). Applied to extracellular electrophysiology data, ICA should separate the spikes from
individual neurons into unique components based on the spatial origin of each signal.

The idea to use ICA for spike sorting has been around for many years[31], but it could
not be implemented without probes with very closely spaced electrodes, so that each neuron
can be recorded by many electrodes. While neural data violates the assumption that the
same signal from one neuron is picked up on each detector with only a difference in linear
scaling[19], extensive tests of ICA on simulations of noisy, nonlinear spike waveforms[32]
suggested that it would deal gracefully with the nonideal conditions of neural recording.

1.4 Outline of Chapters

Chapter 2 contains a description of the complete neural recording system we designed. Chap-
ter 3 describes the results of experiments to validate spike sorting using ICA with data
recorded from our system. Chapter 4 is adapted from previous work[33] on devices for
multisite optical neuromodulation.


Chapter 2

Hardware for Scalable Electrophysiology

2.1 Introduction

Designing an electrophysiology recording system for scalability presents a particular set of challenges. A recording system is comprised of a set of devices that form a signal chain, from a set of conductive surfaces in the brain that sense voltages induced by nearby neural activity, through insulated wires to circuits that amplify and digitize the signals into series of bits, which are then transmitted to a data acquisition system which can store, analyze, and visualize the data. Hardware at each stop in the signal chain must be optimized for its task and designed to interface with the rest of the chain.

For electrode design, the main goal of scalability is to increase the number of recording sites in an electrode array without increasing the spatial dimensions of the array. The trick is to create a sufficiently large exposed surface for each electrode while minimizing the volume of the wire and any substrate which may be supporting it. The best available way to construct electrode arrays with maximal ratio of surface area to volume is to pattern metal electrode sites and wires on a micromachined silicon substrate, which is a process that began development in the 1970's using lithography tools developed by the electronics industry, and has since been continuously developed and commercialized[1–4].

To record a signal from an electrode, the voltage needs to be amplified and digitized. The
main challenge to scalability is getting enough amplifiers within a short enough distance to minimize noise pickup, and the main strategy for increasing channel counts is amplifier miniaturization, enabled by integrated circuit neural amplifiers[5]. For a thousand channel recording system, commercially-available packaged amplifier chips can be put on separate circuit boards (colloquially called ‘headstages’ due to their location during recording) and connected to probes with off-the-shelf connectors and cables. The bandwidth of digitized signals from large neural recordings also presents challenges to data acquisition and analysis. Modern recording systems generate data near the limit of what a single computer can receive and store (16 bits per channel at 30kS/s generates 480 Mb/s of data), presenting a good opportunity to rethink approaches to system architecture to make it simple to add multiple recording nodes in parallel.

The scaling of electrophysiology recording capacity has been driven by advances in electrode fabrication techniques, amplifier miniaturization, and computer systems for data storage and analysis. Notably, all of these components can be fabricated on silicon with tools developed to build computer chips, and the capacity of components built by the semiconductor industry has increased much faster than the capacity of neural recording systems, suggesting that neural recordings systems today are underpowered compared to the state of the art in electronics. There is a clear roadmap for leveraging the steady advances in fabrication technology driven by the computer industry to develop neural recording systems with capacity that is several orders of magnitude greater than the current state of the art.

2.2 Results

We designed a complete 1000-channel extracellular neural recording system, from probes to headstages to data acquisition, in order to demonstrate the principles of scalable recording system design, and test out new algorithms for spike-sorting with close-packed electrodes, as discussed in Chapter 3. As shown in Figure 2-1, we kept the three parts of this system separate and modular, to enable continuous development of each component.
2.2.1 System Design

Probes

We developed a close-packed silicon microelectrode technology that enables a tight continuum of recording sites along the length of the shank. This arrangement thus enables tetrode-like spatial oversampling continuously running down the shank, so that sorting of spikes recorded by such densely packed electrodes can be facilitated for all the sites of the probe simultaneously. We use advanced lithography tools to create these close-packed arrangements in a scalable fashion, demonstrating probes with 1000 electrode pads situated on 5 shanks with 200 recording sites per shank (Figure 2-2). A key challenge with traditional silicon electrode designs has been that wiring to the recording sites occupied a large fraction of the available shank area. This is problematic because scaling up the number of recording
sites forces the shank geometry to widen, placing an upper bound on the number of sites practical for a single shank of a given width. We here reduced the wire geometry, and thus the overhead that the wiring places on the shank width, by creating submicron wires with high-speed electron beam lithography (EBL), which enables us to create feature sizes of 200 nm. The accuracy of this technology also enables a 1.5 μm space between individual recording sites, resulting in a close-packed design in a simple single-metal layer process.

![Figure 2-2: Close-packed electrodes manufactured on a thin silicon substrate. A. Photo of a close-packed 1,000-channel probe, shown with penny, for scale. The sensor area is outlined by the dashed orange line, and the silicon extends to the rectangular region below the penny where wirebonds connect the silicon to the printed circuit board (PCB). The rest of the PCB's area is devoted to wiring and connectors for routing the signals to neural amplifiers on a separate PCB. B. SEM image of sensor area of a 2-shank, 240-channel probe, with electrode sites false-colored gold. Electrode sites are 9 μm × 9 μm, with 11 μm spacing center-to-center. Electrode shanks taper from 120 μm wide at their widest point to 57 μm at the bottom row of electrodes. The silicon substrate beneath the sensors is 15 μm thick. C. SEM image of a cross-section of a silicon probe, with different material layers highlighted in false-color: gold for gold, cyan for silicon dioxide, and blue for silicon. The gold layer is 150 nm thick. The opening in the insulating silicon dioxide for the gold electrode site, right, is 10 μm. The gold wiring encapsulated in silicon dioxide, left, is 200 nm wide on a 400 nm center-to-center spacing.

We adapted a hybrid lithography method, using electron beam lithography (EBL) to
write the fine structures on the probes (for example, the finest wiring along the implanted component) while using classical MEMS fabrication techniques for larger structures (for example, the wirebond pads, the non-fine wiring, and the cutout of the probes). Breaking the metallization into these two steps allows us to utilize the submicron capability of electron beam lithography for dense wiring on the shanks, while maintaining the high throughput of optical lithography to define the less dense wiring outside of the shanks. The highest wiring density is required when routing alongside the recording sites, because here the shank needs to accommodate both the wiring as well as the recording sites, while remaining as narrow as possible to reduce brain displacement or damage. Once the wiring leaves the shank and enters the main body of the probe, significantly more space is available and the wiring pitch can be relaxed, and transitioned to optical lithography. Overall, a 150 mm wafer required about 4 to 5 hours of EBL tool time to complete, with approximately 6000 recording sites spread across a variety of different probe styles of 64, 128, 256 and 1000 channels.

Headstages

We designed custom headstages to amplify and digitize signals between our custom probes and data acquisition system because no commercially-available system offered the density or flexibility to work with our hardware. Headstages employ specially designed neural amplifiers because the signal recorded from extracellular electrodes is small and weak (as sensed at an electrode site, spikes are often less than 100μV peak amplitude, with an impedance greater than 100 kΩ at 1 kHz). These weak signals are very sensitive to noise pickup from many sources (e.g. conductive coupling from muscle activity, capacitive coupling from nearby electrical equipment, and inductive coupling from power lines). This places constraints on the distance amplifiers can be placed from the electrodes and the design of recording circuits. Fortunately, the principles of amplifier design for recording neural signals and PCB layout for mixed signal devices are well known[5], and headstages can be designed with commercial, off-the-shelf (COTS) components.

While designing our first generation headstages, we focused on understanding mixed-signal circuit design, defining the special needs of a 1000-channel system, and producing something with which to test our complete signal chain. In this generation, we produced
Figure 2-3: First generation 512-Channel signal-conditioning headstage. Utilizing 16 32-channel analog neural amplifiers (bottom red box) positioned on front and back of PCB. A 16 bit, 1 MS/s analog-to-digital converter (ADC, middle red box) reads in the multiplexed analog data from the neural amplifier, and outputs single-ended pulses to a differential converter (top red box). Clock pulses are transmitted differentially from the data acquisition module and converted to single-ended signals to control to the amplifiers and ADCs. Standard mixed-signal PCB layout techniques are used to prevent crosstalk from the digital signals on the analog recordings. The headstage circuit board is 6 inches wide.

512-channel headstages that were just 6 inches wide, using 8 32-channel analog multiplexed neural amplifiers (Intan RHA2132) on both the front and back of the board, connected to the probe interface board over flat flexibles cables (FFCs) with a 0.3 mm wire pitch (Figure 2-3). We used mixed-signal circuit design techniques to achieve an input-referred noise of 3 μV per channel, measured through the complete data acquisition system, very close to the amplifier's intrinsic noise level of 2.5 μV.

The second generation of headstage design focused on usability improvements. We adopted a more modular design, using 128-channel headstages with a newer amplifier design (Intan RHD2132) with built-in ADCs and greater functionality. Our headstages included
auxiliary connections for analog input, which were used to synchronize amplified patch-clamp signals and experimental metadata during colocalized recordings (Chapter 3). We also used the headstages in a custom electroplating system to reduce the impedance of our electrodes. The headstages switch between providing electroplating current to the electrodes and measuring the resulting impedance, enabling feedback-controlled specification of impedance. Other usability improvements included mounting points on the circuit boards, indicator LEDs, and power switches, which proved very useful when the headstages were used for colocalized recordings, relied on daily in a complex experimental procedure.

Data Acquisition

We built a minimalist DAQ module (Figure 2-4) centered around an FPGA (Spartan-6 LX150T, Xilinx, Inc.) that acquires neural data directly from analog-to-digital converters downstream from amplifiers headstages. Existing DAQ systems commonly rely on conventional desktop or server machines for data storage. This works well for small numbers of neural DAQ channels, but at higher channel counts a direct communication design based on FGPAs, such as has been implemented for massive DAQ projects in high energy physics[6] and astronomy[7], offers several advantages. Writing data directly to hard drive yields a cost savings, as well as a reduction in complexity. Unlike a CPU, on the FPGA each of the data pathways (connecting sensors, hard drive, and Ethernet) are processed in parallel on dedicated circuitry and have no effect on each other’s performance. With no operating system adding a source of complexity and non-determinism, basic performance guarantees of an FPGA are straightforward and do not require sophisticated techniques such as real time kernels. Furthermore, an FPGA, once programmed, employs dedicated logic for each of the system tasks such as acquiring and storing data, yielding exceptionally high data throughput due to the intrinsically parallel structure of how data streams can flow. Of course, this architecture could also be implemented in an application specific integrated circuit (ASIC), but FPGAs are off-the-shelf parts, reducing entry costs to create such devices, while also allowing for easy modification of the firmware and implemented circuits.

Each of these modules can acquire data from 1024 channels at 16-bit depth at 30 kHz. The module stores data directly to an attached SATA storage device and simultaneously
distributes a copy of the neural data over a high speed (1 Gb/s) Ethernet wired network, e.g., to a visualization computer for real-time inspection of neural data. The module is controlled remotely by sending basic user commands, e.g., start or stop recording, over the network from the visualization computer. In addition to neural data, the module can store experiment protocol information or experimental data, e.g., triggers from beam breaks or for laser pulses, accessible by the experimenter and useful for coordinating overall system operation. To assist in reconstructing a complete neural recording data set from multiple modules being used in the same acquisition session, each module tags neural data with meta-data such as hardware identifiers and time stamps. The total data rate to acquire 1024 channels of data, including auxiliary data, meta-data and zero padding, is 122.88 MB/s. At this data rate, a 512 GB hard drive, for example, can store up to 70 min of data from 1024 channels.

Our implementation uses 62% of the available resources of the FPGA chosen. With the unutilized resources it may be possible to add further functionality such as real-time lossless compression, which can enable both a higher data transmission density and an increased storage capacity (and thus improved utilization of the hard drives). In principle the duration of recording by the data acquisition system could be made longer by increasing the size and/or number of hard drives. Multiple independent SATA controllers could be embedded in the FPGA firmware to support multiple attached hard drives (for instance, our implementation allows two attached hard drives), allowing for increased recording times or for preventing data loss after a drive failure via redundancy. Furthermore, upgrading the embedded controller to operate at SATA II for faster data write speeds (i.e., 3 Gb/s) to, for example, accommodate more than 1024 channels of neural data, would be straightforward. In addition, extremely low latency, closed loop control experiments are feasible if control logic is implemented within the FPGA[8]. Furthermore, it is conceivable that analysis could be performed on the FPGA, with the level of difficulty depending on the complexity of the algorithm. Notably, spike sorting on the FPGA, and only storing the resultant spike information, could in principle yield a dramatic reduction in the rate of output data by several orders of magnitude[6, 9].
Figure 2-4: Implementation of direct communication design of data acquisition (DAQ) module. Photograph of the acquisition module with key parts labeled. The field-programmable gate array (FPGA) board measures 15cm x 16 cm x 2 cm.
2.3 Discussion

We have developed a 1000-channel neural recording system optimized for scalability. Our probes utilize state-of-the-art fabrication tools to maximally cover the surface of the probe with recording sites. Headstages leverage small neural amplifiers chips to pack a large amount of signal conditioning capacity in a small area, which is important for signal integrity and usability. Our data-acquisition systems are contained in a purpose-built custom module to reduce cost and minimize the complexity of using multiple modules simultaneously. All of these components can leverage the continuous development of microelectronics fabrication tools to increase channel counts, decrease system size, and expand the computational capabilities of the system.

One particular lesson we have learned from extensive user testing is that it's very important to make good choices for the physical definition of interconnects between modules. Because sticking with these choices is critical to maintaining modularity, they can persist much longer than any other design choice in the system. We chose to use FFCs with zero-insertion-force (ZIF) connectors between the probes and headstages because they have very dense wiring and mechanically decouple the two systems, be we found it annoying to connect so many FFCs when swapping probes in and out. If we were to respecify this interconnect, we would use connectors that are high-channel-count as well as high density, and mount them on custom flexible PCBs to maintain the mechanical affordances of the system. Likewise, between the headstage and data acquisition system we would employ fewer cables, by serializing outputs. We use the twisted pairs of our HDMI cables to carry differential signal outputs, but they are much lower bandwidth than the cables are rated for. With serialization, all of the data for a 1000-channel recording could be carried over a single HDMI cable. Both of these improvements could be achieved with COTS components, minimizing the cost and complexity of the implementation.

Yet for order-of-magnitude improvements in system capacity, we will need to make some changes to the architecture. Probes and amplifiers will require tighter integration, no longer considered to be separate modules. Probes will become three-dimensional, in order to improve spike sorting (see Chapter 3) and more fully map neural circuits. Amplifiers will be
directly attached to neural probes, leveraging finer wire widths and better packaging options, as shown in Figure 2-5. Looking ever further ahead, we can even use more advanced amplifier architectures that achieve greater density[cite], and even fabricate amplifiers on the same silicon as the probes[2], a strategy that should scale to tens or hundreds of thousands of channels.

The modular, parallel architecture of the data acquisition system will enable scaling by unit increments of 1024 channels, to an arbitrarily large number (Figure 2-6). This parallelism splits up the original data, and the data can if desired later be recombined on the analysis side, during readout from the individual units. This simply requires synchronization across individual DAQ modules, which can be achieved with a shared clock pulse. Visualization could take place on multiple computers if desired, each receiving a number of data streams from several modules. Data analysis will be decentralized as well, on acquisition systems, so that results can be pooled with significant data compression.
Figure 2-5: Methods for scaling and integrating probes and headstages. A) 3D probe array. B) Silicon probe with bare silicon amplifier chips (top left dashed box) wire-bonded directly to silicon interposer. Using bare silicon amplifiers dramatically reduces amplifier area (compare to size of packaged neural amplifier, top right dashed box) and wiring size (compare bottom two dashed boxes). Digital I/O and power lines are connected to red headstage PCB via wirebonds.
Figure 2-6: **Schematic for using multiple data acquisition systems in parallel.** Bottom to top - signals from electrode array are split into separate pipelines that feed through headstages and into acquisition modules which can record and preprocess the data. A network switch connecting all of the acquisition modules allows a single workstation to communicate with all parts of the system, which may include extra nodes for computation.
2.4 Methods

2.4.1 Probes

The fabrication of our devices was carried out on 150 mm silicon-on-insulator (SOI) wafers (Ultrasil Corp., Hayward CA). The top ("device") layer thickness of the SOI wafer defines the probe shank thickness, and therefore a wide range of precise thicknesses is possible. We chose a thickness of 15 μm, comparable to typical silicon-based electrodes [16]. The process consists of two key parts: first, the formation of the insulated metal wiring and exposed metal pads, and second, the deep reactive ion etching (DRIE) steps that define the shape of the probe and the thin shanks. Fabrication starts with deposition of an insulating dielectric. We then use electron beam lithography to define the recording sites and high density metal wiring. Exposure and development of electron beam lithography resist is followed by metal evaporation of a film stack of 10 nm Ti, 150 nm Au, and 5 nm Ti, where the Au is the conductor and the Ti serves as an adhesion layer to the SiO2. After a liftoff procedure in acetone, a second metallization step follows, using standard optical (UV) contact lithography with a 2 μm feature size. Metal for liftoff is deposited by evaporating a 10 nm Ti, 250 nm Au, and 5 nm Ti film stack, and metal liftoff is again performed in acetone. After the metal processing on the front side, an SiO2 layer is deposited and patterned to expose only the recording sites while insulating the wiring. To pattern the dielectric layer and to open up the electrode and wirebond sites, we again use a combination of EBL and optical lithography. We first pattern and etch open the recording sites with precisely aligned EBL to achieve submicron alignment accuracy to the recording sites, with 1 to 2 μm spacing between them. Next, the large wirebond pads are patterned and etched with optical lithography.

Finally, we etch out the probe shapes using two deep reactive ion etch (DRIE) steps, once on the front side of the probe, and then again on the backside. The dual-sided DRIE etching allows us to create shanks of precisely defined thickness, as determined by the SOI wafer, and also precisely defined shank outlines, as determined by lithography. The parts of the probe that are not going to be implanted can remain at the original wafer thickness, which facilitates handling and packaging. A portion of the shanks themselves can also remain at full wafer thickness if those sections will not enter the brain. This can help to reduce space
constraints during in vivo use and enabling a better visibility during the probe insertion into the brain, while avoiding unnecessarily long thinned shanks, which can bend due to thin-film stress. We have fabricated thin shanks with lengths up to 7.5 mm (at a thickness of 15 μm), but, we expect more careful stress balancing is needed for shanks with a more aggressive aspect ratio (either longer than 7.5 mm or thinner than 15 μm).

To acquire neural signals recorded with the probes, we need to package and connect the probes to electronic amplifiers and digitizers, which are commonly referred to as “headstages”. The purpose of the packaging is to provide the intermediary routing between the silicon probe and the headstages, and to protect the fragile silicon shanks by providing a mechanical body for handling. Both of these goals are achieved through the use of printed circuit boards (PCBs). We directly attach the probes onto a PCB and connect to the probe’s wirebond pads using a gold ball bonder. To allow wirebonding, the PCB has an electroless nickel immersion gold finish (ENIG). We used a multilayer PCB to achieve a dense routing of the signals (10 layers, with 8 inner routing layers and 3 mil feature sizes, Advanced Circuits, Aurora CO). These specifications allow us to connect the very densely laid out wirebond pads on the 1000 channel probe. However, for the PCB design, the bottleneck in the layout is the diameter of the via that connects the surface pads to the individual routing layers (in our design, a 14 mil diameter ring with a 4 mil via hole). These dimensions constrained our wirebond pad pitch to 10 mil. Using a more aggressive (albeit more expensive) PCB technology can help reduce this pitch as needed, enabling either higher channel counts or reducing the chip size. However, with our choice of dimensions we find an adequate tradeoff between PCB cost and the required probe size for 1000 channels. Because the metal we wirebond to on the probe is thin, with only 250 nm of gold, and because it sits on top of a relatively thin layer of 1 μm SiO2, choosing the right wirebonder settings is important to avoid cratering through the oxide and short circuiting the pads. Typical chip pads in the semiconductor industry often employ a much thicker metal and dielectric stack that is more robust. However, by appropriately choosing the wirebonding conditions (in our case, power and time settings), we found that we can reliably bond all 1000 channels without any pad damage. The routing across 8 layers in the PCB then takes place below the silicon chip. While lower channel count probes do not require this many PCB layers, we fabricated them
on the same PCB run and therefore benefit from the reduced amount of space needed for PCB routing. The 64 channel probe seen in Fig. 10 illustrates the small form factor that can be achieved. Once the PCB wiring exits the immediate area around the probe, the wiring density can be relaxed as we route towards Flat Flexible Cable (FFC) connectors (Molex 5025983393) that attach to the amplifier headstages. We chose FFC connectors for this design because of the low cost and the ability to use flexible cables that help to mechanically decouple the probe from the headstages. But, many other high-density connector types exist, and they can readily be substituted for the FFC connectors. The additional space overhead is acceptable for the acute, head-fixed in vivo experiments that we are targeting with this technology, for example for mice in a head-fixed virtual reality environment [23]. Overall, the parasitic capacitances are small (10 pF) when compared to the electrode impedance (500 pF).

2.4.2 Headstages

PCB layout utilized mixed signal layout techniques in a 4-layer board to prevent crosstalk from digital I/O lines to amplifier pickups. This included using split ground planes, ground vias for current return paths when signals switched layers, and differential signaling wherever possible. Noise was measured by connecting a probe to the amplifier and immersing it in saline to achieve realistic input impedance. Digitized signals from the data acquisition system were analyzed and RMS noise values were extracted.

2.4.3 Data Acquisition

The FPGA core architecture is shown in Figure 2-7. To transfer data onto and off the FPGA, we took advantage of built-in, high-speed, serial transceivers, and implemented the industry standard SATA 1 interface (187.5 MB/s) in the FPGA itself as a SATA core. This allows the module to store data directly to an attached SATA storage device at data rate sufficient to capture 1024 neural recording channels. In addition, a copy of the neural data can be distributed over a network for relay to a visualization computer. To achieve 1 Gb/s data rates over Ethernet, neural data is encapsulated as UDP packets by the Ethernet/UDP
core. This core interfaces with an onboard Ethernet controller (Marvell MV88E1111, which implements an Ethernet PHY layer).

Figure 2-7: **Block diagram of the FPGA circuitry.** The programmable FPGA circuitry implements a DAQ core to interface with headstages and add meta-data, a SATA core for direct-to-drive data storage, an Ethernet/UDP core for high-speed data transmission over Ethernet, and a Control core containing a bank of registers for controlling the system. One gigabyte of DDR3 memory buffers acquired data before it is stored to hard drive. The TCP protocol is implemented outside of the FPGA using a separate microcontroller.

The 1024 channels of data are here brought on board using 32 front-end neural amplifier digital-output chips (Intan Technologies Inc., RHD2132). Each chip is capable of 32 channels of amplification and digitization, and the individual SPI signals from each chip are aggregated on a simple interface board. The DAQ core includes drivers to communicate over a low-level SPI bus protocol to individual chips (according to the RHD2132 specification sheet), higher-level drivers to initialize the chips and acquire data, and an aggregation layer to combine data from all of the chips and make this aggregate data available to other cores through a RAM interface. In addition to sampling the 32 input channels of each chip, the core allows access to a digital-to-analog waveform generation circuit (useful for electrode impedance measurements), auxiliary analog inputs, and the full set of SPI controllable features of the Intan chip.

The Ethernet/UDP, DAQ and SATA cores are governed by a Control core which mediates
data flow timing and provides a unified configuration interface. This Control core exposes configuration and control registers over TCP/IP to the network with the help of a simple off-the-shelf, on-board, microcontroller board (Maple Mini, LeafLabs, Cambridge, MA, USA). Offloading the TCP/IP implementation simplifies the FPGA firmware and frees up valuable FPGA resources to be used for other data processing tasks or could allow for the use of a smaller/cheaper FPGA. In contrast to the TCP protocol, under which the sender keeps a copy of each packet to ensure reliable communication, in the UDP protocol no copy is kept, so delivery of a packet to a user-controlled software application is not guaranteed by the UDP protocol operating alone. Thus, commands sent as TCP packets to the Control core are always delivered, but neural data sent as UDP packets to real-time visualization devices on the network may not. Regardless, a complete record of all acquired data is stored on attached hard drive(s) for full offline analysis.

The FPGA is able to acquire data from up to 82 SPI interfaces simultaneously, each running with data rates of 400 Mbits per second. In our implementation, each of 32 headstage circuits (RDH2132) supply 16.8 Mbits of data per second, consisting of both the 32 input channels as well as three auxiliary channels which can be chosen arbitrarily from any of the various registers in the RHD2132 chip, or from the current state of 16 GPIO channels on the DAQ module. By default the FPGA is programmed to automatically record the state of all GPIO pins at 2 kHz sample rate. In aggregate, this amounts to 1120 channels at 30 kS/s and 16-bit resolution. To these 1120 16-bit neural and auxiliary samples, extra meta-data is prepended to assist in reconstructing a complete data set, i.e., all data collected in the same acquisition session. The meta-data identifies the acquisition session, the recording module identity, the sequence in which the samples were acquired, and which neural channel contributed to the recording. In our implementation, this meta-data amounts to 0.9% of the channel data rate and includes an 8-byte experiment identifier, a 4-byte module identifier, a 4-byte sample index, and a 4-byte chip live status.

The purpose of the hard drive is solely to store the data as a single stream. Therefore, a random-access file system is not required, and instead the data is written in a single sequential manner. This helps to improve read/write performance and storage efficiency. When sending data to the hard drive, currently the data is padded with zeros at the end.
This padding amounts to 44.8% of the channel data rate and was initially put into place to maintain constant write speeds. However, with the addition to the module of extra memory (e.g., 1 GB DDR3 RAM), this zero padding is not necessary, and could instead be used to store more data in future implementations (up to an additional 831 neural channels and associated auxiliary channels and meta-data, in this implementation).
2.5 Contribution

The high-level neural recording hardware system architecture and specifications were conceived in group discussions with all members of the wired team involved during the project's initial phases, including Ed Boyden, Jorg Scholvin, Justin Kinney, Caroline Moore-Kochlacs, Scott Arfin, Christian Wentz, our collaborators at Leafabls, and I. Lower-level specifications and design choices evolved during the course of the project. Jorg Scholvin performed all design, fabrication, and benchtop testing of the silicon probes. Data acquisition system design was handled by Leafabls, with primary oversight by Justin Kinney, who also engineered a prototype data acquisition system using an off-the-shelf FPGA development board programmed by Jack Scheuer. I performed design, PCB layout, and benchtop testing of all headstage electronics and front-end interfaces to our data acquisition systems. I designed the protocol, hardware, and software for electroplating silicon probes with PEDOT. I specified the file formats and data analysis pipeline for our validation experiments, and wrote the software leading up to invocation of ICA.

Sections of the results and methods pertaining to the silicon probes and data acquisition system were copied or modified from our previous publications[10–12].
Bibliography


Chapter 3

Spike Sorting with Close-Packed Electrode Arrays

3.1 Introduction

Trusted, automated spike sorting is necessary to enable scalable extracellular electrophysiology recordings, but conventional methods are faced with an impossible clustering problem and a difficult choice between the quality and quantity of extracted single units. In order to extract the most information from the data produced from available probes and data acquisition systems, the most sophisticated modern algorithms require human supervision\cite{1} to make these decisions, but this approach strategy has limited scalability. Not only is manual spike sorting a major bottleneck to the analysis of large datasets, but analyses of network activity that are a primary goal of scalable recording become more sensitive to errors as more neurons are included\cite{2-4}. Thus the highest priority for scalable neural recording is to quantify and minimize the error rates in extracted single units in an automated fashion. Working backwards from this goal presents opportunities for new developments to hardware and software based on understanding the problems facing spike sorting today.

The challenge of spike sorting is to assign individual action potentials recorded from an extracellular electrode to an unknown number of neurons. The conventional approach is to extract features from the spike waveforms, and divide spikes into clusters in this feature space, with each cluster representing a different neuron. In theory, spikes from different
neurons should appear in distinct clusters in this feature space when three conditions are met during a recording: action potentials from each neuron have a consistent waveform, the spatiotemporal profiles (a.k.a. spike fields) from different neurons are significantly different, and spikes are sparse enough in time that extracellular spike waveforms recorded from a single neuron are uncorrupted by co-occurring activity. Spike sorting is improved when multiple electrodes record spikes from the same neurons\cite{5, 6}, which provides more chances to sample a point in space where spike waveforms from different neurons are discriminable. Yet in practice, multi-electrode recordings still produce many clusters of spikes from neurons with overlapping boundaries, such that even theoretically optimal divisions result in significant error rates\cite{1}, and it is often unclear just how many neurons there are to cluster. Thus, knowing when and how these conditions are violated is a prerequisite to developing methods to improve spike sorting.

Fortunately, action potentials and spike fields can be described by mathematical models\cite{7-9} that help explain why some clusters are well-separated and some are not. Because most of a neuron’s spike field is generated by currents very close to its soma, it can be treated approximately as a point source, and multiple electrodes can essentially triangulate the origin of the signal when neurons are nearby\cite{10}. But triangulation is less effective at greater distances, because the electrodes all see a similar part of the spike field, and the amplitude decreases quickly (with the square of the distance), rendering amplitude measurements more sensitive to noise. Thus, as neurons get farther away, spike sorters must rely more heavily on the consistency of each neuron’s action potential waveforms and differences in spike shapes from different neurons. At some point spike sorting is rendered impossible, and so the critical question is whether this is sufficient to explain the overlapping clusters found in recorded data.

Yet clusters are not just observed to overlap, they also have very large outliers\cite{11}, which violates the commonly-held assumption that clusters should be Gaussian. While it’s possible this is due to some complexity to in-vivo spiking that has yet to be modeled or observed, another explanation is that recording noise is non-Gaussian, which could be caused by relatively large and sparse noise events, i.e. co-occurring spikes. If co-occurring spikes are large enough to significantly alter spike cluster distributions, then they could be a major cause of
overlapping clusters, which means the assumption of sparsity does not hold. This presents the opportunity to improve spike sorting by finding a way to separate out co-occurring activity.

Fortunately, there exist methods for doing this called blind source separation algorithms. We hypothesized that Independent Component Analysis (ICA) could be used for spike sorting. ICA assumes that recorded signals are a linear mixture of many sources, and searches for a linear unmixing matrix that produces components believed to match the original sources[12, 13]. While real neural signals do not strictly match the assumptions of ICA, simulations suggest it can still be an effective tool for removing noise from co-occurring activity and separating out single units into independent components[14]. We sought to show whether this would work in real recordings, using close-packed electrodes.

3.2 Results

3.2.1 ICA on Ground-Truth Recordings

To test the application of ICA to spike-sorting recordings with close-packed electrodes, we needed a dataset with ground truth knowledge of individual neuron spiking activity. Since patch-clamp recording is the gold standard for electrophysiology at the single neuron level, we performed in vivo colocalized patch-clamp and MEA recordings, using the electrodes and neural recording system described in Chapter 2, and the autopatching robot described in [15], as shown in Figure 3-1. An MEA recorded activity through all layers of visual cortex in a headfixed mouse, and patch pipettes were stereotaxically targeted to search for neurons in layers 2-5 near the MEA. As extracellular spike amplitudes quickly decrease with distance from the soma, it was critical to find neurons just tens of microns from the surface of the MEA. We used voltage pulses emitted from the patch pipette to get feedback on position during the neuron hunt (see Section 3.4.4), and we looked at a patched neuron’s spike-triggered median waveforms on the MEA to assess colocalization post-hoc. With 51 cells patched in 19 mice, the amplitude of the maximum spike amplitude across all recording sites ranged from 5 to 100 uV, with one outlier neuron > 1 mV.
Figure 3-1: Colocalized patch-clamp and multielectrode array recording. A) Diagram of recording setup, using a 4-column MEA with 11 micron pad pitch (electrode sites colored gold). The triangles represent neurons, placed in a 50 micron tall, 50 micron radius cylinder of average cortical neural density. The glass patch pipette, clear-gray, records from a neuron 30-50 microns from the MEA surface. B) 5 second snippet of raw patch recording, top, and bandpass-filtered electrode recording, bottom, from electrode pad with largest amplitude signal. C) Schematic of 256-channel MEA used in this recording, left, with arrow from B pointing towards electrode where signal was recorded. Patch-triggered median spike waveforms, right, on all recording sites, with the same row and column positions as recording electrodes. Red dot signifies proximal electrode, and nearby pads with no signal were from electrodes excluded from analysis due to manufacturing defects in prototype devices used for recordings.
Next, we applied the infomax implementation of ICA to these recordings. Under ideal conditions (i.e. when signal mixing is linear across all detectors, and there are more detectors than signal sources) each component returned by ICA should contain the signal emitted from a single source, if the algorithm converges to the global optimum. Simulations on synthetic datasets of extracellular recordings show that even when these conditions are violated, ICA can return independent components (ICs) that are spike-for-spike accurate with a single neuron[14], which we call single units. In this case, spike identification is performed akin to whole-cell patch clamping, setting a threshold and counting every spike which exceeds it; for a perfect single unit, every spike above the threshold matches a spike in the patched neuron within a narrow time window, and no spikes in the patched neuron fall below threshold in the IC. One important finding from these simulations was that in ICs containing single units, the spikes are well separated from the noise floor, with a large separation in amplitude from spikes from any other neurons. One caveat of these simulations is that they did not model bursting, which dramatically changes extracellular spike waveforms (see Figure 3-7), and so we excluded all spikes with an inter-spike interval (ISI) less than 30 ms (see Figure 3-8) from our spike-sorting tests. In three of the recordings, ICA returned an IC matching the patched neuron with near-perfect spike-for-spike accuracy using this thresholding method. We also found a number of ICs across our recordings with good spike separation in our datasets, indicating separation of a single unit which was not the neuron being patched. These well-separated units can be identified with a simple algorithm, enabling automated spike sorting of some units with good confidence (see Discussion for explanation of blind separation algorithm, simulation results, and methods for excluding and recovering bursts within this framework).

In the tradeoff between spike-for-spike accuracy of extracted units and number of neurons extracted, this algorithm is very strict, and recordings where the patched neuron was not found in a single unit IC might cause concern that this approach does not produce satisfactory yield. Simulations suggest directions for probe development that will increase the number of single units recovered with ICA, but if an alternative spike-sorting method could extract more good neurons from the same recordings, then perhaps probes should be designed to optimize the performance of those methods instead. Multi-unit ICs, which make up most of
ground-truth dataset, are an ideal test case for this tradeoff. They reveal the cases in which spike sorting is most challenging, present an opportunity to test and refine spike-sorting algorithms, and provide a sensitive measure of spike sorting performance.

### 3.2.2 Analysis of Multi-Unit Independent Components

Figure 3-2 shows results from a recording in which the best matching component to the patched neuron was multi-unit. We can view spike-sorting neurons based on IC spike amplitude as solving a one-dimensional clustering problem, defining a boundary to decide which spikes come from the patched neuron; sometimes the distribution of spike amplitudes of the patched neuron overlaps with the distribution of spike amplitudes of another neuron, in which case we get errors. Using an optimal threshold (derived from ground-truth spike knowledge), the minimum combined error rate for spike sorting this component is 9%, measured by the number of false positives (FPs, spikes from the other neuron included in the cluster) and false negatives (FNs, spikes from the patched neuron excluded from the cluster) divided by the total number of spikes; this optimal threshold sets the upper limit on spike sorting performance using thresholded ICA.

Yet it is not necessary to restrict spike sorting on ICs to spike amplitude; waveform shapes may also be informative, just as in conventional spike sorting. We extracted the first three principal components of the IC spike waveforms as features to cluster, as shown in Figure 3-3. To find an upper limit on performance, we used Linear Discriminant Analysis (LDA) to find an optimal boundary, resulting in a 0.8% error rate. In contrast, using LDA on commonly-used spike waveform features [1] achieved a 1.8% error rate. Moreover, clusters of IC features had dramatically fewer outliers and were better matched to a Gaussian distribution, as shown in Figure 3-13, indicating that ICA effectively removed noise from co-occurring spikes. Finally, we note that spike-sorting IC features can be achieved in a much lower-dimensional space, as shown in Figure 3-12, which is known to make clustering more reliable[16].

50
Figure 3-2: Multi-unit component containing spikes from patched neuron. Top, comparison of component trace, red, to patched neuron, blue. Bottom left, receiver operating characteristic (ROC) curve comparing number of true positives (TPs) to false positives (FPs) as a function of component spike threshold. ROC curve axes are numbers instead of rates because there is no clear specification for negative events. Bottom right, plot of FPs, false negatives (FNs), and combined errors (FPs+FNs) as a function of component spike threshold. No threshold results in error-free spike sorting, but an optimum threshold exists for minimizing combined errors.
Figure 3-3: **Comparison of clustering on electrode and IC waveforms.** A-C) Analyses run on electrode waveforms. D-F) Analyses run on IC waveforms. A,D) 200 ms window of spike waveform from proximal electrode (A) and matching IC (D). The last three spikes were from the patched neuron. B) Overlay of all true positive (TP) spikes (black) and false positive spikes (red), measured at two recording sites, the proximal electrode and the most discriminant electrode. TPs and FPs were set by optimal threshold of the IC. E) Average of all TP, FP, and FN spike waveforms, left, and overlay of all TP and FP waveforms, right. C,F) Optimal separation of first three principal components of electrode spike features (C) and IC waveforms (F).
3.3 Discussion

We have shown that Independent Component Analysis (ICA) can reduce the variability of action potential waveforms recorded extracellularly on close-packed multi-electrode arrays sufficiently to enable near-error-free spike sorting that cannot be replicated using conventional methods. Our results provide further evidence that co-occurring activity of nearby neurons is a major impediment to spike sorting, motivating methods that do not require sparse spiking to achieve good spike sorting. Furthermore, ICA produces clusters with more normal distributions that are also discriminable in lower dimensional spaces, which should help enable trusted automation, a critical need for scalable neural recording. Thus, blind source separation methods such as ICA are an important tool for spike sorting.

Yet there is also a wide range in effectiveness of ICA on spike sorting different neurons. Some neurons get extracted perfectly, with a very clear signature in the spike amplitude histogram of the IC. In other cases, several neurons appear with similar amplitudes in a single IC, even when some recording sites measure significantly different average spike waveforms between the neurons. In these cases it would appear that losing the spike field information from all recording sites and clustering in a lower-dimensional space would hinder spike sorting, but if ICA does a sufficient job removing spike waveform noise from co-occurring activity, then the regularity of the spike waveform can more than compensate for the loss of spatial information. This is encouraging, because there will likely always be a distance at which triangulation is insufficient to discriminate between neighboring neurons, so methods that improve clustering on waveform shapes between nearby neurons are essential.

However, some ICs contain multiple neurons with large outliers, where ICA seems to be failing to separate co-occurring activity. Understanding why this happens provides an opportunity for further improvement of spike sorting methods. It is helpful to see ICA as a mathematical framework that renders the underconstrained problem of blind source separation tractable and computationally feasible by choosing a certain set of assumptions and simplifications, which have a variety of implications for the analysis of neural data. Two assumptions that are baked into the formulation of ICA - that transformation from sources to detectors is linear, and that there are no more sources than detectors - are clearly violated.
in extracellular recordings, which can result in solutions that merge spikes from multiple neurons into one IC or share spikes from a single neuron across multiple ICs[14]. While we expect these violations of assumptions will set limits on the performance of spike sorting with ICA, the fact that neurons recorded with similar amplitudes and spike fields achieve drastically different spike sorting performance suggests another mechanism is responsible for limiting performance. It seems likely that in many cases, ICA is not converging to the global optimum (i.e. it is unable to find the weights matrix that produces the best independent components). Several metrics exist to measure the performance of ICA against its own heuristics[17, 18], but it is much harder to know how that compares against what is possible. The search space for the optimal weights matrix is very large, and different implementations of ICA use different choices for heuristic, seeding, and gradient descent, which must balance the exhaustiveness of the search with the cost of computation. Thus, there are many opportunities to try different implementations of ICA to improve spike sorting.

One phenomenenon that poses a particular challenge to spike sorting is bursting. When a neuron produces a fast volley of action potentials (greater than 50 Hz), the waveforms demonstrate a complex history dependence[9, 19, 20]. In particular, with slower rise times because sodium channels need to recover from inactivation, extracellular spikes, which are roughly the derivative of the membrane voltage, significantly decrease in amplitude, often disappearing into the noise floor. In practice, burst spikes are so far removed in feature space from non-burst spikes from the same neuron that they don’t contribute to the problem of overlapping clusters. Therefore, detecting and sorting burst spikes poses a separate challenge, which future algorithms should address with contextual information about the bursting behavior of different cell types.

The results from the current work can be translated into a blind spike sorting method. It begins by finding all of the spikes in each IC using a low threshold. Analyzing the histogram of spike amplitudes, a higher threshold is found to exclude all spikes which can be confidently excluded from the first cluster. All putative burst spikes are then set aside from analysis, and the histogram is reexamined. If a single cluster with good separation is found, then the cluster is judged to be a single unit. If there is no good separation in spike amplitudes, a Gaussian mixture model (GMM) is used to cluster the principal components of IC spike
waveforms, and a metric such as the Mahalanobis distance can be used to keep only clusters that achieve good separation. Finally, burst spikes that were removed directly after a spike from the identified single unit are added back to the cluster.

Our dataset, which contains ground-truth information about spiking activity in primary visual cortex across a range of anatomical positions, stimulus types, and states of consciousness, provides for excellent tests of spike sorting algorithms. In a general sense, we would like to use it to map the space of what neurons are spike sortable, by understanding what covariates affect spike sorting. These covariates include local neuron density, firing rates, synchrony, bursting, spike field nonlinearity, spike variability, and cell type. If we can combine a better understanding of the impact of these covariates with a metric for estimating the quality of extracted units, then we can optimize automated spike sorting algorithms for a variety of goals. We can also use this information to optimize probe design, along with tests of the impact of multishank and 3-D recordings, which should improve neuron discrimination in the non-axial dimensions. We believe that advances to spike sorting algorithms and recording probe designs will enable truly scalable neural recording.

3.4 Methods

3.4.1 Surgical Procedures

All surgical procedures were performed following guidelines laid out by the Committee on Animal Care at MIT, under isoflurane anesthesia (1.5-2.5%) with administration of multiple analgesics. First, stainless steel headplates were attached to male C57Bl/6 mice of 8-12 weeks of age. Two stainless steel screws with attached stainless steel wires were implanted in the skull over the cerebellum and frontal cortex, to act as electrical reference and to provide extra stability for the head plate. The head plate was secured with Metabond dental cement, and the skull was sealed in with Kwikcast silicon elastomer.

After 1-5 weeks of recovery, on the day of recording or one day before, craniotomies were drilled under isoflurane anesthesia. 200-300 um diameter circular craniotomies were drilled at stereotaxically-defined coordinates above the visual cortex, either with a hand drill or with
the autodriller robotic system[21]. The craniotomy for insertion of the MEA was drilled at -2.8 mm A/P, 3.0 mm M/L to bregma, and separate craniotomies for patch pipettes targeting layers 2/3 and layers 4/5 were drilled 0.5 mm and 1.0 mm lateral to the MEA craniotomy. Craniotomies were periodically doused with saline or lactated Ringer’s solution to prevent dehydration, and were sealed up with Kwikcast.

3.4.2 Colocalized Electrophysiology Recordings

For the subset of awake recordings, mice were habituated to head-restraint over 3 days for 15, 30, and then 45 minutes with periodic condensed milk reward. At the start of a recording session, a mouse was initially anesthetized with isoflurane in an induction chamber and affixed by his head plate to a metal holder, with his body fit snugly inside a 3D-printed tube. A fluidic heating pad was placed in contact with the tube to maintain core body temperature. A cone was placed over the nose for continuous delivery of isoflurane anesthesia. A thin layer of eye lube was placed over the eyes to maintain moisture. For anesthetized recording sessions, isoflurane was tuned to the lowest value between 0.5-1.2% that did not result in any motion of the animal. For an awake session, a mouse anesthetized as above was woken up by the cessation of isoflurane delivery.

Electrical ground was shared between the patch and MEA recording apparatuses through a silver chloride pellet placed above the skull in saline or lactated Ringer’s solution. Electrical reference for the MEA recording was connected to either the cerebellum or cortex skull screw wire. The MEA was then inserted perpendicular to the brain surface by a software-controlled, motor-driven linear stage to its target depth, and allowed to settle for at least 10 minutes.

Patch recordings were performed with the aid of the autopatcher robotic system[15] with pipettes pulled to 4-8 MOhm filled with internal solution containing 0.4% biocytin. Upon insertion of the patch pipette into the patch amplifier, the pipette was lowered while monitored under a microscope to the center of the inserted MEA, just above the brain surface using a Siskiyou manipulator, and then touched down to the patch craniotomies to check alignment. If the pipette tip did not line up with roughly the center of the craniotomy, the angle in the plane of the ground was changed and the process was repeated until success.

Patch pipettes were inserted at a 63 degree angle relative to the MEA, aligned to converge
with the surface of the MEA at either layers 2/3 or layers 4/5. The neuron hunt portion of
the patching session typically began from 200um from the target depth and proceeded until
the autopatcher detected a neuron close enough to attempt to form a seal. If the pipette
was estimated to be within 50um of the target portion of the MEA, a seal attempt was
made, otherwise the autopatcher program was directed not to attempt a seal, and either
re-commence the neuron hunt or withdraw the pipette for another attempt. For many
initial recordings, the distance estimation was made stereotaxically, aided by monitoring the
amplitude of a square wave voltage pulse signal emitted from the pipette and measured on
the MEA. These voltage pulses were produced by the autopatcher to monitor the pipette’s
series resistance to the neuropil while searching for neurons, and their amplitude was increase
(to +/- 100mV, corresponding to a current of 10-15nA) to allow for better detection on the
MEA. Later recordings utilized a 1/r physical model of voltage spread from the pulse volleys
to obtain more accurate estimates of pipette position.

The autopatcher contains a standard protocol for attempting to establish a seal. If a
seal was formed but a membrane break-in was not achieved through suction pulses, the
neuron was recorded from in cell-attached mode. If a break-in was achieved, membrane
resistance and capacitance were estimated in voltage clamp mode, and the patching system
was switched to current clamp mode, where 1s long hyperpolarizing and depolarizing current
steps were injected to measure cell electrical properties and spiking threshold. A visual
stimulus was then played to elicit activity. If a neuron did not spike within a few minutes
of the onset of the visual stimulus presentation, 50-150nA of current was injected in steady
state to increase excitation. The visual stimulus would then be played again. In some cases,
a non-spiking neuron that became active with current stimulation was then re-recorded from
without current stimulation, when its basal activity had increased.

A small computer screen was placed at a roughly 45 degree angle in the mouse’s right
visual field. 8 minutes of a visual stimulus were played of either sinusoidal drifting gratings
or a natural scene of reeds blowing in the wind (Chicago motion database) to elicit neural
activity. A photodiode was placed in the lower left of the screen and was digitized with the
patch signal, for synchronization. A typical recording session consisted of 1 or more 8 minute
presentations in succession.
Mice were perfused post-experiment with 4% paraformaldehyde in 1x PBS. Brains were harvested from 2 mice post-experiment for subsequent reconstruction of the location of the patched neuron and dye track from the MEA. Tissue was sliced on a vibratome for coronal sections of 100um thickness and washed in 1X PBS and 1X PBS with 0.5% Triton-X100 and 100mM Glycine (PBT). Slices were then put on a shaker overnight in a solution of PBT and 1:200 Steptavidin-Alexa488 and washed the following day in PBT and then PBS. Slices were mounted with Vectashield medium and imaged on a confocal microscope. Maximum intensity projections were made in ImageJ.

3.4.3 Data Acquisition and Analysis Pipeline

In early recordings, the patch and MEA signals were acquired with separate acquisition systems, a Multiclamp 700B and Intan RHD2000 eval board, respectively, sampling at 25kHz. To account for clock drift, a 25Hz sync pulse was recorded by both systems, for post-hoc temporal alignment. Later recordings acquired both the MEA and patch signals with the Willow system (Leaflabs) sampling at 30kHz, with the patch signal preamplified by an Axoclamp 2B amplifier and routed to an auxiliary analog input of a Willow system headstage for synchronized digitization. The photodiode signal which monitored the visual stimulus presentation was digitized by either the Multiclamp 700B or Willow system and synchronized with the MEA recordings the same way as the patch signal.

On the willow system, MEA data was recorded continuously in a raw (filesystem-free) format to 0.5 TB or 1 TB solid state hard drives (SSDs), which could record for over an hour before needing to be replaced. When experiments were complete, the data on the hard drives was downloaded and compiled into HDF5 files, several hundred GBs in size, which allowed for random access of all data and metadata. Typical 8 minute stimulus presentations generated roughly 10 GB of data from the MEA recording, which could be found inside the larger recording files by comparing timestamps on the willow system and computers controlling the experiments, and locating synchronization pulses in the recording metadata.

I developed a data analysis pipeline in MATLAB and a data model to manage the data generated by our recordings and downstream analyses, schematized in Figure 3-4. The pipeline starts with compiling all of the data and metadata from the experiment, including
Figure 3-4: **Data Analysis Flow Chart.** A canonical dataset is compiled into a standardized hdf5 file format with raw data from a colocalized recording, as well as a record of visual stimulus presentation and metadata about the experimental conditions (peach boxes). Preprocessing steps analyze the recordings and save the results back to the canonical dataset. Spike sorting algorithms are applied to the processed data, and the results are also saved back to the canonical dataset. Methods are provided for search for datasets by metadata features, extracting results from the datasets and summarizing results across recordings, and analyzing many datasets with the same algorithm, either in a serial batch process on a single computer, or simultaneously across a large computing cluster.

The data model serves several important functions. It helps to tame the complexity of tracking variables from experiment to experiment - once data has been compiled to conform with the data model, it does not matter which data acquisition system was used to generate it, even though several different sets of code were required to compile all of our experiments. It supports a robust suite of tools, for searching for datasets that match particular metadata criteria, and deploying batches of analyses on a local machine or remote computing cluster. It also aids metanalysis, pooling results across recordings to identify clusters and trends.
3.4.4 Pipette Position Estimation

Voltage pulses emitted by the patch pipette during the neuron hunt phase of the colocalized recordings were used to estimate the position of the tip of the pipette relative to the MEA. During the neuron hunt, the pipette moves in 2 micron increments, pausing after each step to measure the pipette’s access resistance to the neuropil, which increases when it is pressed against a neuron. The voltage pulses are detected by the MEA; as the pipette moves towards the MEA, the signal increases, and the signal distribution across the pads changes. The pipette tip can be modeled as a unipolar point current source, given that the silver chloride pellet that acts as ground for the patch recordings is much farther from the pipette than the MEA. A unipolar current source in a resistive medium produces a $\frac{1}{r}$ voltage falloff (this is dual to the electrical potential of point charge in free space[22]), and so the voltage distribution measured across all pads of the MEA can be fit to this distribution to estimate the probe position.

To produce this fit, the voltage pulses are first located in the patch signal. They are volleys of 20 Hz square waves lasting 0.75 s per step, with 1.25 s between each step (triggers for the beginning and end of the voltage pulses are software controlled, so the durations are variable). A 0.5 s window is automatically selected from the center of each volley, a Fourier transform is performed over this time window on each electrode pad, and the amplitude of the 20 Hz time bin is extracted (because the signal is stationary and precisely 20 Hz, the Fourier transform provides a more accurate estimate of signal strength than other methods).

The amplitude of the voltage pulse versus electrode position is then fit to a five-parameter model of the $\frac{1}{r}$ voltage falloff, showing in Equation 3.1: three parameters for position $(x_0, y_0, z_0)$, one parameter setting proportionality constant $(K)$, and one parameter for added noise $(c$, which significantly improved fits, because signal strength did not decay towards zero).

$$A(x, y) = \frac{K}{\sqrt{(x - x_0)^2 + (y - y_0)^2 + z_0^2}} + c$$

(3.1)

An example fit is shown in Figure 3-5. Even in the presence of an extraneous signal near the top of the probe (possibly due to muscle activity, which can form electrical continuity through saline to the surface of the brain), the fit matches the amplitude of measured voltage pulse...
Figure 3-5: **Estimate of pipette position by curve fit to $\frac{1}{r}$ falloff.** The 20 Hz amplitude measured on each of detectors after a single step of the patch pipette is plotted as a function of detector position along MEA, with blue circles representing the true measurements, and colored manifold displaying the curve fit. Note that the fit is robust to an artifact recorded near the top of the probe, likely due to electrical conduction through saline from muscle to the brain surface.

very closely. The fit is found using MATLAB’s curve fitting toolbox, which readily converges to excellent fits, often with R-squared values greater than 0.99, when initial conditions are set with zero noise and the pipette position $(x_0, y_0, z_0)$ 50 microns above the electrode pad measuring the largest voltage pulse amplitude. In initial tests, I allowed the exponent in the falloff curve to be a free parameter, i.e. fitting $n$ for $A \propto \frac{1}{r^n}$, and found that fits were best for $n$ very close to 1, confirming the model of a unipolar current source.

We gain increased confidence in our fits by tracking the pipette through its complete neuron hunt, as summarized in Figure 3-6. The pipette position after each step is estimated independently, and then a best-fit line is found across all of the points. The step size and angle of pipette insertion derived from the best fit line are found to be in very close agreement to the mechanical specifications of the probe insertion system, and the residuals, calculated as the distance between each estimated step position and the point along the best
Figure 3-6: **Probe track across complete neuron hunt.** Three orthogonal views of probe track estimates, top. The independent position estimate after each step is denoted with a red star, and red lines connect each step to the next. Green line shows the best-fit line through the points. Middle, distance from each step position estimate to estimated progress along green line (note this is not simply the shortest distance to the green line). Bottom, R-squared fit value for each step.
fit line corresponding to that step, often average about 5 microns. At this time, it is unclear what are the relative contributions of fit uncertainty, fine-scale tissue inhomogeneity, and pipette deformation to these residuals. Regardless, 5 micron uncertainty in pipette position is comparable to the uncertainty in the location on a neuron where a patch is formed, and is much less than the average spacing between neurons, so this measure provides as much accuracy as is necessary to theoretically identify a patched neuron in imaged tissue.

3.4.5 Spike Identification in Patch Signal and Burst Exclusion

The first step in assessing spike sorting is to identify spikes and extract spike times in the patch trace, to produce average spiking maps of the patched neuron on the MEA, and provide a ground truth reference for extracted component. We developed algorithms for automatically detecting spikes, but visually inspected all results. In rare cases, burst spikes of cell attached neurons were unclear, but these did not affect downstream analyses. Spikes in whole cells recordings were characterized by consistent spike amplitudes, though recordings showed occasional drift. Spikes were detected by threshold crossing, usually set halfway between the spike peak and resting potential. Spike thresholds were increased as necessary, if the membrane did not fall below the original threshold during bursts. Cell-attached recordings displayed a greater variety of behaviors. We found that the peak of the derivative was a more consistent, though noisier indicator for spikes. We applied a match filter to enhance the spike waveforms of the derivative of the patch voltage, and suppress spurious spikes resulting from signal jumps in a single time bin.

We found that the peak of the derivative of the patch voltage was closer in time to peak of extracellular waveform (Figure 3-7), and average spiking maps triggered on the peak of the derivative produced shorter waveforms with higher peaks, indicating that its timing was more consistently matched to the extracellular waveform than the peak of the patch voltage. This finding agrees with the theory that extracellular waveform is produced by currents through the membrane, whereas the membrane acts as a capacitor.

Bursts were a prominent feature in our datasets. We observed burst to have many features matching what is reported in the literature [11, 23, 24], including increased resting membrane potential in between bursts seen in patched neuron, and slower rise times, a decrease in peak
Figure 3-7: **Comparison of intracellular and extracellular waveform during a burst.** Membrane voltage, in blue, reaches approximately same peak during every spike, though waveform shape changes and rise times increase. Matching ICA component, in red, shows decreasing spike amplitude during the burst, and returns to initial height after a 100 ms interval. The ICA component is displayed rather than directly-measured extracellular voltage in order to minimize noise from co-occurring activity.

derivative, and a decrease in extracellular spike amplitude. Extra logic implemented to detect bursts, but want to exclude for the purposes of evaluating ICA for spike sorting.

To find an ISI threshold for excluding burst spikes, we examined the spike amplitude as a function of ISI for the second spike in the burst. We saw a dramatic decrease for ISIs less than 10 ms, with measurable decreases up to 30 ms, so we set a conservative minimum of 30 ms ISI for evaluating spike sorting. The amplitude of the third spike in a burst showed a dependence on ISI of both previous spikes, indicating a complex history dependence. We also found that spikes steadily decreased throughout a burst.
Figure 3-8: **Impact of ISI on burst spike amplitude.** Spikes were coded by order in a burst and binned by ISI. Spikes that came second in a burst were roughly half the size of the first spike in the burst. As ISI increased, so did amplitude, and spikes with greater than 30 ms ISI showed negligible changes in spike height.

### 3.4.6 Filtering and Artifact Subtraction in MEA Signal

It is common practice to high-pass or band-pass filter extracellular signal before spike sorting, in order to remove high-amplitude, low-frequency signals. We investigated optimal filter settings to preserve action potential waveform shape. While it is common to use corner frequencies as high as 500 Hz[1], we found that a 100 Hz pole was necessary to prevent artifactual positive deflection before and after spike, as shown in Figure 3-9. We chose to
use a noncausal second-order IIR butterworth bandpass 100-6k Hz filter, run forwards and backwards (using MATLAB's 'filtfilt' function) to avoid shifting the signal in time. We investigated more complex filters such as FIR filters using optimal Parks-McClellan filter design, but found them to have less faithful waveform reproduction and to run slower.

Figure 3-9: Effect of filter parameters on extracellular spike waveform. Top, 150 ms of unfiltered recording from a single electrode, showing a number of spikes of various amplitudes, and large low-frequency features. Middle, the same signal, bandpass filtered from 300 Hz - 6 kHz with a second order butterworth IIR. Positive deflections bracket each spike which are not present in the unfiltered signal. Bottom, the same signal, bandpass filtered from 100 Hz - 6 kHz with a second order butterworth IIR. Spike waveforms look qualitatively identical to the unfiltered recording, and low-frequency features are almost completely removed.

In some recordings, an artifact from motion from mouse is strongly coupled into the signal, likely due to continuity from muscle to brain surface through saline. In this case, the artifact shows up on all channels, and can be larger than the spikes. To remove this
motion artifact, we would calculate the average waveform across all good channels, calculate its projection onto each channel (which could vary by approximately 10% due to electrode position and impedance), and subtract off the artifact, as shown in Figure 3-10.

Figure 3-10: **Motion artifact subtraction.** Top, signal from one electrode sensing a motion artifact, which is much larger than recorded spikes. Bottom, motion artifact, extracted from mean signal recorded by all channels. Second row, signal from above, with motion artifact subtracted, using mean of all channels to act as a virtual reference. Third row, a component trace containing the same spikes, for comparison. ICA filters out the motion artifact without relying on a virtual reference.
3.4.7 ICA Implementation

We used the infomax implementation of ICA, which uses a maximum entropy heuristic that minimizes mutual information under certain assumptions. We used the runica implementation, and allowed it to run until the weights change was less than 1e6, which took 3 hours for 8 minutes of 256-channel data using 10 cores and 100 GB ram on a computing cluster.

3.4.8 Matching Independent Components to Patched Neuron

To match ICs to a patched neuron’s spike times, we extracted all spikes six times higher than the median of the absolute value of each IC, which consistently allowed small spikes above threshold while excluding noise. We looked for ICs with spike times matching the peak of the derivative of the patch neuron’s voltage within a 1 ms window. If multiple ICs contained nearly all of the spikes of the patched neuron, we chose the IC with the best ROC curve.

3.4.9 Feature Extraction and Discriminant Analysis

We used linear and quadratic discriminant analysis (LDA and QDA) to assess the separability of spikes using features extracted from the waveforms measured on the electrodes and from the best independent component. Spiketimes were extracted from the independent component best matching the patched neuron, using the maximum threshold that resulted in zero false negatives. The top three principal components were extracted as features from the thirteen rows of the MEA centered on the detector that recorded the largest amplitude spike in the median waveform, excluding detectors that were part of short- or open-circuits. This produced 90 features for the recording analyzed in Figures 3-2,3-3,3-11,3-12, and 3-13. We applied PCA to these features and, separately, to the component spike waveforms, to extract the top ten principal components (PCs) from detector space and component space.

We used 5-fold cross-validation LDA on the three sets of features (detector space features, detector space PCs, and component space PCs) to compare error rates achievable by clustering the data with Gaussian mixture models (GMMs). We used the MATLAB function 'cvshrink' to perform a search over two parameters (gamma, which regularizes the measured covariance matrix, and delta, which sets a minimum threshold on discriminant dimensions)
Figure 3-11: Comparison of median TP and FP waveforms on MEA for neuron shown in Figures 3-2 and 3-3. TP waveforms are shown in black, and FP waveforms are shown in red. The figure shows the 11 rows of the MEA with largest signals for this neuron. Time windows are 4 ms around spike peak on largest channel.
Figure 3-12: **Comparison of error rates vs. number of predictors for linear discriminants across three sets of features.** Feature sets were extracted using PCA from waveforms measured on the MEA or the best independent component. PCA was also performed on the MEA feature set to achieve dimensionality reduction. 5-fold cross-validation LDA was used to estimate error rates, and two parameters were swept to exclude the least informative features from the discriminant model. Figure shows a scatter plot of the error rate vs number of predictors for each parameter pair tested for each feature set.

We tested how well the clusters matched normal distributions, by generating QQ plots, shown in Figure 3-13.

### 3.5 Contribution

Brian Allen performed all of the colocalized neural recordings. Brian discovered that he could use the voltage pulses as a beacon for locating the patch pipette relative to the MEA, and I performed the curve fitting that produced precise estimates of the pipette location. Leaflabs provided functions for copying data from recordings to an analysis computer. I designed and implemented the data analysis pipeline and the analyses detailed in this chapter.
Figure 3-13: QQ-plots of neuron clusters in IC space and MEA feature space. Optimal model parameters for TP (black dot) and FP (red dot) clusters were found using Quadratic Discriminant Analysis (QDA). Left, using the top two principal components (PCs), and right, using the top four PCs. Top, clusters in IC space match a normal distribution more closely than clusters in MEA space, bottom.
Bibliography


Chapter 4

Devices for 3D Neuromodulation

4.1 Introduction

Optogenetics has been widely applied to the control of single or small numbers of deep structures in the brain (e.g., using a fiber-coupled laser[1–4]), as well as to patterned control of superficial brain areas (e.g., using scanning lasers, LED arrays, and other 2-D patterning strategies (e.g., [5–7])). Recently we have engaged in developing devices that exhibit both the scalability to high target counts exhibited by 2-D arrays of light sources, and the deep structure targetability of optical fibers, by delivering the light from a 2-D array of custom-placed sub-millimeter-sized LEDs, into a set of custom-length optical fibers that are individually docked to LEDs [8, 9], even in wireless fashion [10]. A key advantage of this methodology is that these devices can be built and tested by individual groups using simple machining and assembly techniques. We here present the process for design and construction of LED arrays and LED-coupled optical fiber arrays, demonstrating the key engineering principles of design and fabrication Such devices are compact and lightweight, and are easily carried by freely moving mice. Our design is centered around a procedure in which a 2-D LED array is assembled, and a set of custom-length fibers are docked to it, in a single step, thus enabling easy end-user customization and fabrication of a set of arrays in a matter of days, using inexpensive computer-based automated machining tools. We enable device operation for behaviorally relevant timescales, and can support electrophysiological recording concurrent with optical illumination. We describe new tools to systematize the surgery, facilitating...
4.2 Results

4.2.1 Fiber array design, fabrication, and operation

We have previously shown that arrays of LEDs, patterned on a substrate to match the shape of target brain regions and implanted above the surface of the cortex, can be used to control behavior with optogenetic stimulation[10]. These arrays, wirelessly controlled and powered, allow a freedom of movement impossible to achieve with optical stimulators tethered to distal lasers. Additionally, the method used to pattern the LEDs on the substrate achieves a packing density (<1 mm center-to-center) and precision (features to mount LEDs machined with 25 μm resolution) that would be extremely difficult to achieve with traditional stereotaxic methods of implanting individual optical fibers. Thus, we sought to develop a method for fabricating arrays of optical fibers to be coupled to our arrays of LEDs, in order to target deep brain structures with the same precision we have employed to target surface structures.

One core innovation used here is an efficient, precise way of docking a large set of optical fibers to a large set of LEDs arrayed on a planar surface, so that the resultant device can safely and effectively deliver light independently to a set of sites distributed in a three dimensional fashion in the mammalian brain (Figure 4-1). We chose our fiber-LED coupling strategy to maximize light output from an optical fiber of fixed width, using optical fibers with high numerical aperture LEDs with high surface irradiance. According to a theorem derived in ref. [11], maximal light coupling from an LED into an optical fiber is achieved when a fiber is in direct contact with an LED, and the surface area of the LED is greater than the fiber. This coupling method simplifies device design by removing the need for additional optics, and the Lambertian emission pattern of LEDs tolerates alignment errors of several degrees thanks to the angle-independent irradiance of LEDs, in contrast to lasers which must be precisely aligned.

Using this coupling method, placement of a 200 micron optical fiber (0.48 NA) onto a
Figure 4-1: **Design and fabrication of optical fiber arrays.** A, Schematic, in exploded view, of a fiber array with fibers pointing upwards, with inset zoomed in on LEDs and their connections, adapted from refs. [8-10]. Vertical dashed lines denote points at which components dock together when the device is assembled. Numbers refer to key components: 1, optical fiber; 2, LED; 3, LED pedestal (carved out of LED base plate, 11); 4, wire bond; 5, fiber alignment plate; 6, fiber fitting; 7, reflector plate; 8, device assembly guidepost (to be removed after final assembly, but before implantation); 9, reflector plate heat conduit; 10, circuit board; 11, LED base plate—; 12, circuit board connector. B, Key structural components, numbered the same as in A. C, Photograph of a relatively dense hippocampal CA1-targeted fiber array device (schematized in A), appropriate for silencing the entire hippocampus for example, with fibers pointing downwards, with a penny for scale. D, An 8-fiber hippocampal array, appropriate for stimulating multiple points in the hippocampus, shown with optional cooling module before encapsulation with biocompatible epoxy.
600 micron x 600 micron LED (465 nm wavelength Cree EZBright600), with the small gap bridged by an index-matched adhesive (see Section 4.4.3 ), yielded a maximum irradiance at the free end of the fiber of 220 + 10 mW/mm² (mean + standard deviation; LED run at the less-than-maximal current of 500 mA; n = 7 LED-coupled fibers, using the testing procedure described in Section 4.4.5), matching fiber tip irradiances commonly used in vivo for safe illumination of opsin-bearing cells [1-3]. At this tip irradiance, a fiber can illuminate the edges of a >1 mm³ volume to an irradiance of >1 mW/mm² (as reflected by computational models and experiments in refs. [1, 2, 12]), a light level at which many commonly used microbial opsins are significantly activated [12–15]. Thus, depending on the scientific goal at hand, LED-coupled fibers can be packed together to guarantee ‘tiling’ of brain structures (750 microns apart, as in Figure 4-1A-C, a “dense” bilateral CA1 hippocampus targeting array), or spaced sparsely (1.5 mm apart, as in Figure 4-1D, a “sparse” bilateral CA1 hippocampus targeting array).

To fabricate a 3-D fiber array, we first fabricate a 2-D array of LEDs, using computer-aided circuit design and manufacturing tools, and then align and dock the set of optical fibers to the set of LEDs, in a single positioning step. We here describe the principles of the design; detailed step-by-step instructions for semi-automatically manufacturing the components and performing the assembly are given in Sections sections 4.4.1 to 4.4.3. Notably, only two relatively inexpensive machines not commonly found in an ordinary neurophysiology lab are required, a tabletop mill and a wirebonder. Assembly proceeds in two phases: in the first phase of construction, raw die LEDs (component 2 in Figure 4-1A) self-align onto solder-coated planar pedestals (component 3) to form a 2-D array of LEDs on a copper LED base plate (component 11), and then LED terminals are wirebonded (component 4) to an attached circuit board (component 10); the circuit board bears copper traces for provision of power to individual LEDs, and a connector to connect the LED array to the outside world (component 12). In the second phase of construction, an alignment plate (component 5) and reflector plate (component 7) that firmly hold, via fitted pieces of tubing (component 6), a set of optical fibers (component 1) cleaved to lengths that correspond to the dorsoventral coordinates of brain targets, is then lowered onto the 2-D LED array along a set of assembly guideposts (component 8); optical adhesive is used to couple fibers to their
corresponding LEDs. The reflector plate is then thermally connected to the LED base plate via copper heat conduits (component 9). Finally, exposed surfaces of the device are coated with biocompatible epoxy. For surgical practice, fiber arrays without LEDs can be very rapidly fabricated, as shown in Figure 4-2Bi, and described in Section 4.4.9.

Several key mechanical, optical, and electrical engineering innovations equip the device with the properties of easy end-user design and fabrication, good optical performance, compact size and small weight, good thermal management, and low electrical noise production. For example, after brain targets are chosen by the user, the four key structural plates (components 5, 7, 10, and 11 in Figure 4-1A) are all semi-automatically machined out of their respective materials by a tabletop mill, which is controlled by a PC that runs custom scripts (e.g., in MATLAB) that convert LED coordinates into mill instructions. Precision machining of all core structural plates from a single set of parameters insures high precision of fiber location along medio-lateral and anterior-posterior axes (<10 micron variability). To ensure a similar precision for the positioning of small raw LED dies onto the LED base plate, we utilized a self-alignment strategy in which the surface tension of solder on the LED pedestals (component 3 in Figure 4-1A) automatically positions the LEDs in the proper geometry. Once assembled, such a device occupies <0.5 cm³ in volume, and weighs <1 g, and thus is easily carried by a freely moving mouse on its head.

To achieve the compactness of our device, we use the LED base plate (component 11 in Figure 4-1A) as not only the mechanical LED mount and heat sink, but also as the common anode for all the LEDs. To power LEDs wired in this way, we designed a custom LED driver circuit board to receive TTL pulses (from any commodity source, e.g., a NI-DAQ board) and then to trigger the board to transmit the appropriate power to control the LEDs (see Section 4.4.4 for a description of the circuit). We designed a novel geometry for a multichannel coaxial cable (visible in Figure 4-2C-D, and schematized in Figure 4-3B) to provide power to the LEDs on the fiber array; this cable uniquely minimizes both capacitive and inductive coupling of electrical signals to nearby structures (e.g., recording electrodes, Figure 4-3A) while allowing for dozens of independent current-carrying LED control lines in a flexible cable format; this cable is described in more detail in Section 4.4.4. We use pulleys and a counterweight to offset the weight of the electrical cable (and, optionally, of the cooling
Figure 4-2: Accessory devices, and modifications, for implantation and utilization of fiber arrays. A, Fluidic cooling system, added to the back of the fiber array, appropriate for increasing the amount of time the array can be run continuously. i, Schematic, in exploded view, of the cooling system, added to the back of the fiber array. Numbers refer to key components: 11, LED base plate; 13, fluidic cooling channel plate; 14, fluidic cooling channel backing plate; 15, barbed fluidic connector. ii, Photograph of a cooled hippocampal CA1-targeted fiber array (based on the design in Figure 4-3A). B, Accessory devices to aid fiber array implantation. i, Practice array, with the same targets as the array schematized in Figure 4-2A, shown during fabrication before encapsulation with biocompatible epoxy. ii, Parallelized craniotomy marker, consisting of freely-moving hypodermic steel tubing with inked tips, appropriate for conforming to the contours of the skull and marking sites of craniotomies for fibers.
Figure 4-3: A custom coaxial cable design enables decoupled LED activation and recording. A, Raw traces of three electrode recordings on a tetrode (the fourth electrode here acts as amplifier reference), of neurons in the cortex of an awake headfixed mouse (non ChR2-expressing) during 50 Hz LED operation (each pulse was for 10ms at 500mA, indicated by blue bars). B, Average, across trials and electrodes, of 33 traces obtained in each of the saline (left) and in vivo (middle) conditions, displayed as mean (solid lines) + standard deviation between electrodes (shaded area). Right, average of the 33 traces obtained in vivo, after each trace was preprocessed by subtracting off the saline-characterized artifact recorded on the same electrode (and averaged across all trials for that electrode, e.g. 11 trials). C, Schematic of the “spatially averaged” multichannel coaxial cable utilized to deliver power to LEDs while minimizing capacitive and inductive coupling to recording electrodes.

tubing, described below), that enables practically free mouse movement (and could be used in conjunction with electrical and fluidic commutators, if necessary).

Thermal management is an important consideration for these devices because our optical coupling strategy, chosen to maximize the light output from optical fibers, has a tradeoff in power efficiency. There are two causes for this. As the drive current to an LED increases, its efficiency (measured in light output power over electrical input power) decreases. We chose to use a 500 mA drive current, which is near the knee of the power efficiency curve. The second efficiency loss is due to the optics governing the coupling of LED light to an optical fiber – by using an LED larger than the optical fiber, we make sure to fill the surface of the fiber with light, but also create a lot of light beyond the edges of the fiber which cannot be coupled in. Ultimately, we use almost 2 W of electrical power to generate roughly 300 mW
of light power, of which 7 mW is coupled into an optical fiber.

We designed the LED base plate to function as a high heat-capacity heat sink to pull heat away from the mouse skull. Since the LEDs are soldered directly to this heat sink, heat that is generated during light production has a very low resistance path into the heat sink. To account for optical energy that is not coupled into the fibers, we chose to direct light power away from the head and back towards the heatsink with a highly reflective aluminum reflector plate (component 7) attached to the fiber alignment plate with thermally insulating epoxy. Aluminum absorbs much less blue light than copper, and calculations indicate that this strategy reduces the heat delivered towards the skull by stray LED light by approximately two orders of magnitude, vs. a design without a reflector.

To determine the safe operating regime for a fiber array, we devised an experiment to measure the temperature that would occur near the skull of mouse with an implanted fiber array during fiber array operation. We determined that for a conservative restriction of 1 °C temperature increase, the heatsink could absorb 7 seconds worth of LED operation, which may distributed over longer time periods (e.g. running one LED for 14 seconds at 50% duty cycle), or divided among multiple LEDs (e.g. running ten LEDs for 0.7 seconds). To expand the range of fiber array protocols of operation possible, we devised a supplementary cooling module (shown in Figure 4-2A) that increases the maximum amount of energy that can be consumed on the device by an order of magnitude or more (dotted lines and filled diamonds in Figure 4-4). In this modified design, a microfluidic cooling channel plate (component 13, with microfluidic cooling channel backing plate, component 14, in Figure 4-2Ai) attaches to the back of the LED base plate to support water cooling of the device (via connectors, component 15 in Figure 4-2Ai, that connect to tubing that leads to a peristaltic pump). A modest cooling system (using 750 micron diameter channels, and room temperature water flowing at a constant rate of 0.5 mL/sec) enables a 9x increase in the duration of the protocol of operation for one LED, or a 5x increase in the duration of the protocol of operation for 10 LEDs. After the fiber array is run on a protocol of operation, the array cools to baseline temperature with a time constant of $3.8 \pm 0.1$ minutes (mean + standard deviation; n = 2 arrays), meaning that the available 7000 LED-ms of device operation can be re-utilized every few minutes, appropriate for many behavioral paradigms. Note that if a fiber array
protocol of operation involves slow heat generation compared to this time constant—that is, the LEDs are run at a low power level, or the LEDs are pulsed at a very low frequency, then such a protocol of operation can continue indefinitely (Figure 4-4, steep upper left part of each curve). Additionally, the time constant of array cooling speeds up to just 1.5 ± .5 minutes (measured for n = 2 arrays).

Figure 4-4: Thermal characteristics of fiber array use. Maximum temperature increase at a point in the material (dental acrylic) holding an array to the skull of an anesthetized mouse, for uncooled (solid lines, circles, n = 2 arrays) or cooled (dotted lines, diamonds, n = 2 arrays) fiber arrays, plotted as a function of the total fiber array power (x-axis, expressed in units of the number of active LEDs running at 500 mA) and the duration of fiber array operation at that given total fiber array power (y-axis). Shown for clarity is the behaviorally relevant portion of the temperature increase dataset acquired (circles, diamonds), as well as mathematically fit isothermal contours of temperature increase by 0.5, 1.0, and 1.5 °C.
4.2.2 Implantation, utilization, and validation of fiber arrays in behaving mice

Surgical implantation of fiber arrays into the mouse brain takes place under anesthesia, after the insertion of three anchor screws into the skull[16]. We have begun the process of automating the surgery; for example, we mark all of the craniotomy sites needed for fiber array insertion in a single step by lowering onto the skull an array of free-sliding micro-pens so that each marks the location of a craniotomy with an ink dot (see Figure 4-3Bii and Section 4.4.9)

4.2.3 Integration of electrophysiological recording and fiber array illumination

The fast pulses of current that power the LEDs of a fiber array can potentially result in artifacts on nearby conductors via capacitive and inductive coupling. To solve this problem, we developed a lightweight multichannel coaxial cable to convey power to the LEDs (schematized in Figure 4-3C). The key design innovation is to have the internal wires of the cable spiral about the central axis of the cable, so that as seen from outside the cable, each individual internal wire spatially averages so as to simulate a true coaxial cable, in conjunction with an outer copper mesh. Currents (500 mA, 10 ms long) delivered to an LED on an array resulted in little noise on electrodes running along the corresponding fiber on the array. The small, low frequency artifacts that did couple in (average peak of 15 + 4 µV, avg + std. dev. across electrodes, n = 3 electrodes) were identical between recordings performed in saline and in vivo, enabling high-fidelity subtraction, leaving undetectable noise (Figure 4-3B).

4.3 Discussion

We here present the design and fabrication methodology for a device that comprises an array of custom-length optical fibers docked to a 2-D array of LEDs[8–10], which enables independent light delivery to sites distributed in a three-dimensional pattern throughout
the mammalian brain, important for targeting realistically-shaped brain circuits for neural perturbation. Our device possesses a form factor that is compact and lightweight enough (e.g., 1 gram, 0.5 cm3) to be borne by freely moving mice. Each individual fiber can easily address a volume on the order of a cubic millimeter with irradiances on the order of 1 mW/mm2; greater volumes are addressable at lower light powers. The device can be operated for periods of time appropriate for many behavioral experiments; when used in conjunction with a supplementary fluidic cooling apparatus, it can support experiments in which large numbers of LEDs are operated for long durations. For power delivery, our device utilizes a novel multichannel coaxial cable, which minimizes capacitive and inductive artifacts on adjacent neural recording electrodes. Finally, we devised new tools to systematize surgery, enabling large numbers of fibers to be inserted through targeted craniotomies. We demonstrate how to build these devices using just two inexpensive manufacturing devices in addition to what might be found in a modern neurophysiology laboratory. With some practice, creating a batch of fiber arrays, from design to final assembly, takes just a few days. Building devices in the lab is advantageous for rapid prototyping when designs may change often, but fabrication and assembly could also be outsourced. The maturation of 3D printers may also soon offer sufficient resolution to simplify many parts of the fabrication process.

To simultaneously achieve in a single device the many specifications required—three-dimensional targeting of light with good optical performance, easy end-user design and fabrication, compact size and small weight, low heating and electrical noise—required a large number of optical, mechanical, electrical, and thermal variables to be simultaneously optimized. For example, the coupling strategy between LEDs and fibers not only maximizes light throughput, but also enables single-step alignment of a large number of optical fibers to a 2-D LED array. The core structural plates of the device enable not only this single-step alignment process, but also provide for heat management, and support passive electrical noise cancellation.

Some performance tradeoff decisions are required when specifying the device design. While our light-coupling strategy maximizes light output from our optical fibers, it sacrifices overall power efficiency—3% of the light output from an 600 micron LED is coupled into a 200 micron optical fiber— which also places demands on thermal management for the devices.
Efficiency could be increased by decreasing LED size (decreasing overall fiber output power and increasing power output variability between fibers due to manufacturing tolerances) or using laser diodes instead of LEDs (increasing device complexity). Thermal management could be extended to support many more LEDs with the use of low-temperature fluids for array cooling, in conjunction with active thermistor feedback. Thus, with different choices for what parameters to optimize, the design principles for these arrays can be used to support applications with many different kinds of constraints. There are also many possibilities for how these fiber arrays may be customized and augmented. Our device uses blue LEDs, appropriate for activation of not only opsins whose action spectra peak in the blue such as ChR2, Mac, and Chronos[17], but also many other opsins such as Arch that have action spectra peaks in the green or yellow but are also well activated by blue light. In the case of Arch, for example, its sensitivity to light at 470 nm is approximately 50% of its peak sensitivity, still enabling very significant silencing in a region[12]. Other colors of LEDs may also be used, such as high-power 625 nm LEDs to drive the red-light sensitive silencer Jaws[18].

We expect that as consumer applications drive improvements in the components that make up the fiber array – e.g., LEDs become more efficient, digital fabrication becomes cheaper – the fiber array device itself will ride these technology development curves, becoming more powerful. Commonly used neural recording devices can also be integrated with our array: for example, tetrode bundles or silicon probes could be passed down through holes strategically placed in the array. In the years to come, we anticipate that new technologies, such as bundles of waveguides[19] coupled to miniaturized lasers, and novel surgical technologies such as machines that automatically drill craniotomies, will pave the way towards systematic deployment of such devices at even denser levels throughout distributed neural circuits. Such devices may also serve, in the future, as prototype optical neural control prosthetics, supporting the development of novel treatment strategies for intractable neurological and psychiatric disorders.
4.4 Methods

4.4.1 LED and Fiber-coupled LED Array Fabrication: Design and Preparation of Key Structural Components

Fiber arrays are made of an array of optical fibers (components 1 in Figure 4-1A), which are docked to a planar set of LEDs (components 2 in Figure 4-1A). The alignment of the fibers to LEDs is achieved via a stack of structural components (the fiber alignment plate, reflector plate, and LED base plate, components 5, 7, and 11 in Figure 4-1A, respectively) that hold the optical elements (LEDs and optical fibers) in precise positions (within 10 microns) relative to guide holes on the structural components. The guide holes are aligned with device assembly guideposts (component 8 in Figure 4-1A) to hold the stack of structural components in the proper position. Detailed assembly instructions are given in the following sections, and the purpose of each component is explained in the Results, Section 4.2.1.

To facilitate the design and creation of fiber arrays, we developed a pipeline of computer aided design and fabrication tools. We use EAGLE, a free CAD program, to graphically lay out all of the components in the fiber array, and we use a tabletop, computer-controlled mill (MDX-15, from Roland DGA) to cut components 5, 7, 10, and 11 (see Figure 4-1A) out of stock materials. MATLAB scripts act as a bridge between the design and fabrication processes, translating specifications for the arrays extracted from EAGLE into machine code readable by the mill. Crucially, EAGLE provides methods for automated data input, through script files, as well as automated data output, through its CAM processor. Thus, array specifications stored in a MATLAB script can be visualized and then adjusted in EAGLE, and changes made in EAGLE can be recorded in the MATLAB script.

4.4.2 Fiber Array Fabrication: Preparation of Other Components

A number of other components must be prepared prior to assembly into a fiber array. Optical fibers (component 1 in Figure 4-1A) are prepared from 0.48 NA, 200 μm-diameter core optical fiber (Thorlabs) by sectioning into 1-inch-long segments with a razor blade and removing all jacketing with a wire stripper (Stripmaster). In order to facilitate reliable and repeated
cleaving of fibers to desired length, a diamond fiber cleaver is used, in conjunction with fiber length trimming shims, which are small blocks that fit in a slot of the diamond cleaver and are milled to the length of each fiber. The fiber segment is inserted into a diamond fiber cleaver (Delaware Diamond Knives) against the back of the cleaver, and cleaved once. The fiber is then pushed forward within the cleaver the length of the shim of desired fiber length (i.e., by placing the shim between the fiber and the back of the cleaver), and cleaved again, producing an optical fiber the length of the shim with optically smooth cleaves on both ends. This process is repeated for each desired fiber. Guide posts (component 8 in Figure 4-1A) are made of .020” stainless steel wire (Small Parts), cut into 1 cm lengths (e.g., using a Dremel abrasive wheel). Heat conduits (component 9 in Figure 4-1A) are made of 1/32” copper wire (McMaster-Carr), cut into 3 mm lengths. When fluidic cooling is used, cooling channels and backing plates (components 13 and 14 in Figure 4-2Ai) are cut from 1/16” thick copper plates with waterjet cutters, and holes in the backing plates are threaded with an 0-80 tap. Fluidic barb connectors (component 15 in Figure 4-2Ai) are 3D printed, and the ends are threaded with an 0-80 die.

4.4.3 Fiber Array Fabrication: Assembly

The first set of steps of device assembly is to put the LEDs and the circuit board (components 2 and 10 in Figure 4-1A) onto the LED base plate (component 11 in Figure 4-1A). The LED pedestals (component 3 in Figure 4-1A) are pre-tinned by depositing a thin layer of solder paste on top and then heating the LED base plates on a hot plate set to 235oC for 20 seconds. Then additional solder paste, as well as raw die LEDs (e.g., Cree EZBright Gen II EZ500, EZ600, EZ700, or EZ1000 chips), are put on the pedestals. At this time, we also place the circuit board (component 10 in Figure 4-1A) atop the LED base plate, with a thin sandwich of solder paste in between. The device assembly guideposts (component 8 in Figure 4-1A) are then inserted through the guidepost holes so as to align the circuit board to the LED base plate. Optionally, the fluidic cooling channel plate (component 13 in Figure 4-2A) and fluidic cooling channel backing plate (component 14 in Figure 4-2A) are attached to the LED base plate with solder paste, forming a water-tight seal. This assembly is heated on the hot plate for another 20 seconds, during which LEDs self-align to the edges of the pedestals.
due to the surface tension of the liquid solder. Guideposts are removed after the device has cooled. The electrical connector (component 12 in Figure 4-1A) (i.e. Samtec FTE10) is then soldered to the circuit board. A copper wire is inserted through the via (the top hole in the circuit board, Figure 4-1B) and soldered to both sides of the circuit board; this connects the positive voltage supply, coming from the electrical connector on the top side of the circuit board, to the plane of copper on the bottom side of the circuit board and the LED base plate, which provides the positive voltage source common to all the LEDs soldered onto the LED base plate. Finally, the assembled apparatus is cleaned of flux residue in a small plastic tube filled with isopropyl alcohol inside an ultrasonic bath.

The assembled apparatus is next hot glued to a glass slide with the coolant backing plate (for cooled arrays) or LED base plate (for uncooled arrays) flat against the slide. LED bond pads are wedge bonded to copper circuit board traces via 0.001” aluminum wire (using a wire bonder, West Bond 747677E-79C).

The remainder of the key structural parts, the alignment plate and the reflector plate (components 5 and 7 in Figure 4-1A), are linked by the four device assembly guideposts, and epoxied around the edges to insure a small (e.g., 1 mm) gap for thermal insulation. Reflector plate heat conduits are inserted into the reflector plate and attached with thermal epoxy. Fiber fittings (component 6 in Figure 4-1, made from PEEK tubing 0.010” inner diameter, 0.018” outer diameter for the plates as designed above) are inserted through the alignment plate and reflector plate and cut flush with a razor blade on both sides, then set in place by filling the space between the alignment plate and reflector plate with epoxy. Then the optical fibers are loaded into the PEEK fiber fittings. This second assembly is then lowered onto the LED base plate so that the guideposts insert into the guidepost holes of the LED base plate, and the optical fibers are just above (100 microns) the LEDs. Optics glue (Thorlabs) is then applied to the LED-fiber interface and cured with a UV lamp. The reflector plate heat conduits are bonded to the LED base plate with thermally conductive epoxy. The spaces in between the reflector plate, the circuit board, and the LED base plate are filled with 5-minute epoxy, and after curing, the guide posts are rotated about their own axes gently to loosen them from the epoxy, then pulled out of the device. To seal the outside of the device to insure biocompatibility, to prevent electrical shorts, and to block stray light.
from emitting, the exposed leads of the electrical connector are sealed with hot glue, and the
surface of the device is coated with black epoxy. Finished arrays are released from the glass
slides by placing the slides on a hot plate at 100°C until the hot glue attaching the array to
the slide (described in the previous paragraph) becomes soft (this procedure does not affect
the hot glue used to seal the exposed leads of the electrical connector). Finally, if cooling is
used, barbed fluidic connectors (component 15 in Figure 4-2A, 3D-printed out of acrylic and
threaded on the outside with a 0-80 die) are screwed into the holes in the coolant backing
plate and sealed to the plate with epoxy.

For the bilateral CA1 14-fiber design (Figure 4-1A-C), we chose fiber termini (listed in
the following list, in units of millimeters anterior, lateral, and ventral to bregma) of: (-1.70,
±0.60, 1.25), (-1.70, ±1.30, 1.00), (-2.40, ±1.50, 0.90), (-2.40, ±2.20, 1.10), (-3.10, ±4.10,
4.25), (-3.10, ±2.50, 1.20), (-3.80, ±3.85, 2.75), so that 14 LEDs were used with centered
fibers; arrays were constructed both with and without cooling modules.

4.4.4 LED Driver Circuit Design and Operation

The LEDs can be controlled using a standard LED driver circuit with a transistor and
current-limiting resistor connected to the cathode of each LED. Importantly, the anodes of
the LEDs are all connected together on the LED base plate, and it is desirable to tie the
LED base plate to earth ground. Therefore, the LED cathodes must be connected to a
negative voltage supply. We built a custom circuit board to facilitate controlling the LEDs
with ground-referenced digital logic signals.

4.4.5 Fiber Array Testing: Light Power

We measured light power output with n = 7 Cree EZ600 LEDs, each coupled to a 200 μm core-
diameter, 0.48 NA optical fiber. Power measurements were performed with an integrating
sphere photometer (Thorlabs). Arrays were fit in place over the integrating sphere (with
original connector removed) with a lasercut adapter with a dock for the alignment plate,
which blocked stray light from entering from the integrating sphere. Individual LEDs were
run at 500 mA with water cooling (room temperature, 30 mL/min, with a peristaltic pump)
for several seconds to obtain a stable measurement. Measured power was divided by optical fiber tip surface area to obtain irradiance.

4.4.6 Fiber Array Testing: Thermal Testing In Vivo

We measured temperature within the dental cement between a fiber array and the skull of a mouse (an acute experiment under isoflurane anesthesia, body temperature maintained with a heating pad). We tested two cooled hippocampal arrays (room temperature water coolant circulated through the cooling module with a peristaltic pump, 30 mL/min) and two uncooled arrays with two LEDs each, in each case with a disposable thermistor (Digikey part 495-2159-ND) embedded in the dental acrylic between the array and the skull. The thermistor was placed in a voltage divider circuit, and the thermistor voltage acquired by a NI-DAQ board, digitized at 10 Hz. We sampled the space of fiber array powers and fiber array durations of operation over a wide range of protocols of interest to neuroscientists. To facilitate further extrapolation to protocols not explicitly tested here, we additionally performed a curve fit for each of the two datasets (the cooled array dataset and the uncooled array set), modeling the system as undergoing two processes: fast-timescale dumping of heat into the system, and slower-timescale heat dissipation to the environment. A convenient approximation to these two guiding principles is simply to fit the empirical data with a two parameter equation that both captures the kinetics of heat generation and of heat dissipation; to first order, this can be written as:

$$T = \frac{a \cdot P \cdot D}{1 + b \cdot D}$$

expressing temperature increase, $T$, as a function of the total fiber array power (as utilized in Figure 4-3A), $P$, the duration of fiber array operation, $D$, and two free parameters, $a$ and $b$.

4.4.7 Fiber Array Fabrication: Optimization for Electrophysiology

Two sets of cables are needed for joint optical fiber array perturbation with concurrent recording: one to carry power to the LEDs and one to carry neural signals back (for this, a
conventional cable is fine). For the power cables, we developed a novel style of coaxial cable (Figure 4-3C), which is designed to eliminate inductive and capacitive coupling to nearby surfaces and wires. Briefly, the outside comprises a tubular copper mesh (Daburn), held at ground (which eliminates capacitive coupling), that provides a current path from a power supply to the common anode of the LEDs; running inside the mesh is a bundle of twisted copper strands (33 gauge magnet wire, MWS Industries, twisted with a power drill) that serve as individual current return paths for each LED’s cathode. For the internal copper wires, determining how the pins of the two male connectors correspond can be accomplished with a multimeter. Importantly, the helical copper strands and the tubular copper mesh are coaxial and carry equal and opposite currents, minimizing the inductive coupling between the cable and nearby recording devices.

4.4.8 Electrophysiological Recording and Data Analysis

We performed acute cortical recordings in an awake, headfixed mouse, using a LED-coupled fiber with a tetrode epoxied to the side of the fiber. To prevent photoelectrochemical artifacts [3, 11] from corrupting our measurement, during the construction of the device a thin opaque FR1 plate was placed between the LED and the end of the fiber. Neural signals were amplified 20x by a headstage amplifier and 50x by a second stage amplifier (Plexon), digitized at 30 kHz with a Digidata (Molecular Devices), and analyzed with pClamp and MATLAB. After neural recordings were established, LEDs were then driven at 500mA for 10 ms pulse durations at 50 Hz for 300 ms train durations (i.e., 15 pulse repetitions), with a pulse train every five seconds, for 1 minute sessions. The same measurements were repeated with the tetrode and fiber tips in saline. For analysis of LED-electrode coupling, we analyzed the data obtained from 11 LED pulses that were delivered at 10 ms duration at maximum power, after at least 50 ms of darkness (3 electrodes, for a total of 33 total traces). We calculated average and standard deviation traces in MATLAB, aligning traces on the LED pulses. We computed the difference between the neural and saline artifact by subtracting the averaged saline artifact from each raw neural trace, and then averaging the resultant subtracted neural traces. To look for action potentials, traces were bandpass filtered between 270 and 8000 Hz.
4.4.9 Fiber Array Fabrication: Accessories for Surgery Practice and Surgery Facilitation

Molds for assembling practice arrays are created from 1/16” FR1 epoxy laminate plates, each plate of which is milled separately (see Supplementary Methods S3.4 for details on milling), and which are then stacked on top of each other, aligned with guideposts that go through holes milled in each of the four corners (and that stick out, for use in practice array assembly). Plates are secured to each other on the sides with hot glue. The key design feature of the practice array assembly mold, facilitated by the computer aided design and fabrication system, is that each component plate is milled to have holes that line up, in the final stacked assembly, to form columns through which fibers can be inserted, so that each fiber is immersed in the mold to a depth which equals its ultimate desired depth in the brain. To enable fibers to be inserted into these columns so that the centers of the fibers are precisely aligned, pieces of PEEK tubing (i.e., identical to the PEEK fiber fitting utilized above, component 6 in Figure 4-1A) are inserted into the plates as described in the above section; to facilitate replacement of PEEK tubing (e.g., if damaged) without having to remake the whole practice array mold, the PEEK tubing pieces can be mechanically constrained (e.g., by being inserted into a plate that is flanked above and below by plates with smaller holes than the outer diameter of the PEEK tubing).

Practice arrays (Figure 4-2Bi) are constructed by placing a practice array plate (essentially a fiber alignment plate, with holes drilled where the fibers pass through, with an extra tab sticking out, to facilitate later attachment of an electrical connector to enable mounting on the stereotaxic arm) over the practice array assembly mold, so that the guideposts pass through the guidepost holes in the practice array plate. Then, PEEK fiber fittings (component 6 in Figure 4-1A) are inserted into the holes in the practice array plate, and trimmed to the thickness of the plate. Optical fibers are stripped of their jacketing and inserted through the PEEK fiber fittings in the practice array plate, down into the practice array assembly mold until the fibers touch the bottoms of the columns that have been machined to insure that the appropriate lengths of fiber extends below the end of the practice array plate. Optical fibers are cut so that a few millimeters of fiber extend above the practice plate, and
then hot glue is placed on top to hold the fibers and PEEK fittings to each other and to the plate. An electrical connector (identical to component 12 in Figure 4-1A) is epoxied onto the top of the practice plate, to help with later holding by the stereotax during surgery. The practice array plate is encapsulated in biocompatible epoxy.

Parallelized craniotomy markers (Figure 4-2Bii), to aid in the stereotactic determination of coordinates for drilling craniotomies to insert fiber arrays into the brain, are constructed by supergluing PEEK fiber fittings into the holes of a practice array plate. Hypodermic tubing (32 gauge, 1 cm long) is cut and then inserted into the PEEK fiber fittings. An electrical connector (identical to component 12 in Figure 4-1A) is epoxied onto the top of the automated craniotomy marker, to help with later holding by the stereotax during surgery.

4.4.10 Fiber array implantation

All animal procedures were in accordance with the National Institutes of Health Guide for the care and use of Laboratory Animals and approved by the Massachusetts Institute of Technology Animal Care and Use Committee. Mice were anesthetized with isoflurane in oxygen, and then administered buprenorphine and Meloxicam for analgesia. After revealing the skull, and leveling bregma and lambda to the same vertical position, the sites for craniotomy opening were indicated by using a parallelized craniotomy marker (essentially an alignment plate with freely-moving hypodermic tubing inserted within, and dipped into sterile, biocompatible ink, see 2.9 for details) held by the stereotax and lowered, to mark the locations of all the drill sites in one step. Then three small screws (size 000, 3/32" long, J.I. Morriss) were implanted to make a broad-based tripod for attachment of the fiber array atop the skull [12]. Small craniotomies in the skull were made, over each ink-stained drill site, with a hand drill (e.g., Barrett pin vise, with HSD-75 drill bit); alternatively, a large craniotomy was made around the outline of the ink-stained drill sites with a dental drill. Then, the optical fiber array to be implanted was attached to a custom holder (essentially a Samtec connector on a post that plugged into the electrical connector of the fiber array, component 12 in Figure 4-1A), and then lowered using a stereotaxic apparatus (e.g., Kopf) so that the tips of the optical fibers were 250 microns above the target coordinates, as explained above.

Then, the fiber array was secured to the three skull screws with dental acrylic; it is important
to “loop” the wet acrylic over the array to mechanically secure everything together.
Bibliography


2. Bernstein, J. G. et al. Prosthetic systems for therapeutic optical activation and silencing of genetically targeted neurons in *Biomedical Optics (BiOS)* 2008 (2008), 68540H–68540H.


