Integration of neural optical recording and stimulation on minimally invasive, deep-brain implantable CMOS

Adriaan Johannes Taal

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022
Abstract

Integration of neural optical recording and stimulation on minimally invasive, deep-brain implantable CMOS

Adriaan Johannes Taal

This thesis describes the development of a minimally invasive integrated platform for all-optical neural stimulation and measurement (OptoSAM). The OptoSAM platform is a single mixed-signal complementary metal-oxide semiconductor (CMOS) chip. After design, the chip is postprocessed to contain the necessary optical filters and emitters to enable both fluorescent detection and neural stimulation. Finally, the chip is packaged in a probe form factor for minimally-invasive implantation into neural tissue.

The thesis describes how the OptoSAM is engineered for two applications: optical fluorescent imaging on one hand and optogenetic stimulation on the other. For either application, constraints and tradeoffs are described that guide design specifications.

For fluorescent neural detection, this thesis focuses on improvements made in lens-less image reconstruction and optical filtering. It describes circuit design for the lens-less, filter-less fluorescent imaging subsystem and characterizes the resulting imaging performance. The lack of on-chip filters precludes reliable imaging of fluorescent targets both in-vivo and ex-vivo.

To address these limitations, the metal-insulator-metal angle sensitive pixel (MIMASP) is introduced, a novel nanophotonic structure that integrates lens-less imaging and optical filtering in an ultrathin (<5µm) frontend. The MIMASP offers three advantages over previously published angle sensitive pixels. First, it orthogonally modulates the detection light field for two arbitrary wavelengths, enabling the separation and detection of colors in the image. Secondly, each layer is
constructed from optical long-pass filters, rejecting the blue excitation light. Third, an analytical framework is created that allows to optimize the ensemble image reconstruction resolution as a function of the available per-pixel geometries. The angle sensitive pixels are a promising lens-less imaging method for situations where both the number of pixels and the permitted device dimensions are extremely constrained. Equipped with the MIMASP frontend, the imager is demonstrated in scattering media to successfully separate fluorescent targets based on color, fluorescent lifetime and even environmental pH. The experiments are extended to fluorescent detection in \textit{ex-vivo} acute brain slices.

For optogenetic stimulation, we equip the OptoSAM platform with organic light emitting diodes (OLEDs) as thin-film emitters. \textit{In-vivo} results show how the OLED probe can evoke neural activity in a fully scalable fashion. Using synchronized groups of OLEDs, large neural populations can be synchronously activated. Simultaneously, single neurons can be manipulated by emission from single OLEDs at a 25µm pitch. We demonstrate single-unit manipulation and separation of both pyramidal and interneurons. A custom flexible, transparent multi-electrode array (MEA) provides the electrophysiological recording for cross-validation in the deep-brain. Measurements show how local field potentials (LFPs) are evoked at both 300µm and 1.2mm deep, and how the LFP magnitude roll-off proves locality of the induced activity. Compared to previously published state-of-the-art, the OLED-on-CMOS approach provides a two orders of magnitude larger field of view (FoV) while improving resolution by $3\times$. Pixel pitch and count can be fully scaled to provide arbitrary fields of view and resolution.

The OptoSAM platform proves a pathway towards behavioral studies in awake mice. These studies could address multiple brain regions independently with a single device insertion. This
provides neuroscientists with the tools to study relationship between distant regions with single-neuron resolution.

While the detection and stimulation are separately optimized and validated, the chip is a promising platform for future integration of both modalities. The thesis discusses recommendations to achieve full integration of both modalities. To this end, it proposes three future chip designs, each with their respective strengths. The proposals also provide potential solutions to the challenges associated with the design and fabrication. The thesis concludes with recommendations for future experiments, both for the OptoSAM platform and for future designs.
# Table of Contents

List of Figures ......................................................................................................................... iv

List of Tables ............................................................................................................................. xv

1. Introduction ............................................................................................................................. 1

   1.1. Brain computer interfaces ............................................................................................... 1

   1.2. The ideal BCI and sensor modalities ............................................................................... 3

   1.3. Access to deep-brain ...................................................................................................... 5

   1.4. The optical paradigm for deep-brain interfaces ............................................................ 6

   1.5. Goals of this thesis ......................................................................................................... 7

   1.6. Structure of this thesis .................................................................................................... 7

2. Design space for the OptoSAM all-optical deep-brain neural interface ................................ 10

   2.1. Potential applications for all-optical deep-brain BCIs ..................................................... 10

   2.2. State-of-the-art for integrated optical neural interfaces ............................................... 13

   2.3. Project assumptions and trade-offs .............................................................................. 17

   2.4. Design specifications for the OptoSAM system .............................................................. 23

3. A fluorescent lifetime imager shank .................................................................................. 25

   3.1. Introduction ..................................................................................................................... 25

   3.2. System overview ............................................................................................................ 28

   3.3. Time-gated fluorescence imaging .................................................................................. 31
3.4. Lens-less imaging using near-field angle-sensitive detection ........................................... 42

3.5. Imaging performance ........................................................................................................ 46

3.6. Discussion ...................................................................................................................... 50

4. Integration of optical color filters and compressive sensing in an ultrathin frontend ........ 51

4.1. Introduction .................................................................................................................... 51

4.2. Results .......................................................................................................................... 54

4.3. Discussion ...................................................................................................................... 72

5. DISTINCT: An Algorithm for fast higher order separable Basis Pursuit Denoising ........ 76

5.1. Introduction .................................................................................................................... 76

5.2. The DISTINCT Algorithm for compressed sensing with input and output modulation 80

5.3. DISTINCT accuracy and computation time versus competing algorithms .................. 83

5.4. Convergence Guarantee and Requirements ................................................................... 87

5.5. Discussion ...................................................................................................................... 91


6.1. Introduction .................................................................................................................... 92

6.2. Results .......................................................................................................................... 94

6.3. Discussion ...................................................................................................................... 101

7. Integration of OLEDs on high-density CMOS for in-vivo optogenetic stimulation ........ 103

7.1. Introduction .................................................................................................................... 103
7.2. A 1024-emitter optogenetic stimulation probe in CMOS ................................................. 106
7.3. OLED characterization ........................................................................................................ 109
7.4. Large-FOV, localized optogenetic stimulation in-vivo ...................................................... 112
7.5. Single-neuron, single-OLED stimulation with high temporal resolution ......................... 114
7.6. Modulation of local field potentials in-vivo ....................................................................... 116
7.7. Discussion .......................................................................................................................... 118
8. Conclusions .......................................................................................................................... 121
  8.1. Summary of results and contributions ............................................................................... 121
  8.2. Review of specifications .................................................................................................... 123
  8.3. Discussion of results ......................................................................................................... 125
  8.4. Future engineering proposals ........................................................................................... 136
  8.5. Future biological applications .......................................................................................... 143
References ................................................................................................................................... 145
List of Figures

Figure 3.1 Size and imaging depth comparison of (a) multi-photon microscope, (b) implantable microscope with inserted GRIN lens, and (c) imaging probe resolving 4.1 mm-deep in brain tissue with an external excitation laser. Fluorescent emission is collected by 512 SPAD pixels located on two shanks. ................................................................. 25

Figure 3.2 (a) Chip micrograph with shank IC highlighted. (b) System architecture consisting of optoelectronic IC, read-out FPGA and computer for image reconstruction. ................................. 26

Figure 3.3 (a) Post-processing of CMOS die into implantable imaging probe. (b) Scanning electron microscopy image of post-processed imaging probe with 120 µm width and 40 µm thickness, packaged for insertion (inset). ................................................................. 27

Figure 3.4 (a) Circuit diagram of active-reset quenching circuit. (b) Timing diagram of laser-synced time-gate operation. (c) Layout of single pixel................................................................. 28

Figure 3.5 (a) Dark count rate of 512 pixels. (b) Median PDP versus incident wavelength for 350 nW isotropic illumination. (c) FoM at three different fluences plotted versus $V_{OV}$ (d) SPAD count linearity versus incident light intensity. Error bars indicate the standard deviation over 512 pixels. ................................................................. 29

Figure 3.6 (a) Circuit diagram of DTC. (b) Timing diagram with $F_{CLKIN}=32\times F_{TRIGGER}$ and tuneable counter width=1 and delay=0. (c) Ideal and synthesized ON signal delays and (d) conversion nonlinearity. .............................................................................. 30

Figure 3.7 (a) Diagram of fluorescence measurement setup showing fluorescent and scattered excitation photons reaching the SPAD array located along the x axis. (b) Photo of 10 µM fluorescein in 0.6% agarose, excited by a collimated 488 nm pulsed laser. (c) Time-gated photon counts synchronized to the laser pulse at $t=0$ s for pure agarose (blue) and fluorescein in agarose
(FL, green) plotted against the left axis; respective SBR (Fluorescein:Agarose) is plotted on the right axis (red).................................................................................................................. 32

Figure 3.8: Fluorescence emission of EGFP-expressing somata in model tissue compared with background and dark count as a function of tissue depth. (b) Combined photon shot noise and dark count shot noise, plotted with SNBR as a function of tissue depth. ................................................. 32

Figure 3.9 (a) Illustration of an angular sensitive SPAD implemented in CMOS. (b) Measured angular modulation of sixteen different grating structures in direction (x, y), frequency (High, Low), and quadrature phase (0°, 90°, 180°, 270°), for 520 nm illumination (c) Simulated angular responses of grating combinations y/L/180° (top left) and y/H/90° (bottom left), with respective φ=0° cross sections measured and parametrically fitted (right). ................................................. 34

Figure 3.10: PSF of a single point source located at coordinates [2000 µm, 0 µm, 200 µm] portrayed as (a) xy-, (b) xz-, (c) yz-plane cross sections of a 3D pseudoinverse backprojection; voxels are 5 µm across all dimensions; (d) PSF FWHM in x, y, z with respect to with increasing distance of point source from imager; (e) xz-plane cross section of imager FoV. Dotted yellow lines in (a) and (e) mark the location of the imager.................................................................................................................. 35

Figure 3.11: A monochromatic light source hovering 200 µm above the imager was placed at x locations of 500 µm (p1), 1500 µm (p2), and 2500 µm (p3). (a) Raw 512-pixel counts along two shanks and (b) pseudoinverse backprojection in plane z=400 µm with source locations (*) and probe (--) overlaid.................................................................................................................. 36

Figure 3.12 (a) 10 µm-diameter microsphere clusters deposited on cover glass placed 400 µm above imager (dotted line overlay); epifluorescence microscopy image(top) and pseudoinverse backprojection(bottom). (b) Same with larger clusters. (c) Two sphere types, fluorescent (left) and
scattering (right), showing selective imaging of fluorescent target; top-fluorescent, middle-brightfield, bottom-backprojection. All scale bars are 200 μm. ............................................ 37

Figure 3.13 (a) Distribution of localization error in spatial coordinates. (b) Standard deviation when accounted for 2% false positive / false negative outliers. (c) Bias of localization error. .... 45

Figure 3.14 (a) 3D reconstruction of microsphere clusters in Fig. 12b; two most significant solutions are marked in red circles and the following 40 in black, with diameters depicting amplitude estimate; the overlaid ground truth image indicates the location of microspheres 400 μm above the chip. (b) L2-norm of residual error decreasing with added sources, red and gray corresponding to first 2 and next 40 voxels in order of highest amplitude estimate. ............... 45

Figure 4.1: MIMAS filter design. (a) The Ag-HfO2-Ag MIM spectral filter selectively transmits a single resonant wavelength by tuning the oxide thickness. Interleaving both color filters with wavelength scale pitch constructs the grating structure. (b) Transmission spectrum of planar bandpass filters and interleaved grating filter over the visible spectrum for measured (solid) and FDTD simulated (dash) devices. The planar bandpass filters selectively transmit a single color, whereas the interleaved structure transmits a superposition of the two. (c) A 500 nm cut-on longpass filter dye is combined with the MIM filter, attenuating wavelengths under 480 nm by four orders of magnitude. (d) Verification of non-overlapping microscale Talbot self-image using near-field scanning optical microscopy (NSOM) in transmission mode, at wavelengths of 532 and 660 nm. ................................................................. 56

Figure 4.2: MIMAS filter integrated on CMOS. (a) Fabrication of MIMAS on CMOS (b) FIB-SEM sideview showing front end fabrication, with annotations for SPAD, electrical contacts and CMOS metal stack. (c) The dual grating structure selectively transmits light according to incident angle, illustrating the point spread function from a single detector for either wavelength. (d) Two-
dimensional angular response for a single device for both resonant wavelengths. A plot at $\varphi=0^\circ$ compares measurement (solid) to simulation (dash). Resulting parameters of interest are the angular period $360^\circ \beta^{-1}$, peak transmissivity offset from zero $\alpha$, and modulation index $m$.

Figure 4.3: MIMAS-CMOS implantable shanks. (a) The implantable imager consists of four shanks, each employing 2x128 SPAD pixels. (b) The chip is packaged and mounted on a micromanipulator for implantation. (c) Photomicrograph showing the 16 front end geometries distributed over the imager. Rotations ($\chi$) are swept along the $x$-direction, grating pitch ($p$) and offsets ($\delta$) placed adjacent. (d) The linear separability and relative photon count (compared with a SPAD with no mask) for all voxels in a $1000\times1000\times150$ µm subset of the field of view. The MIMAS (current work) is compared with a conventional metal-based angle-sensitive design where each pixel carries separate red and green filters for each pixel alternatively. (e) PSF in the XY plane at a height of 150 µm above the pixel ensemble with detail along $y=400$ µm. A conventional angle sensitive array without dual bandpass filter shows distorted PSF, as its effective pixel density is halved.

Figure 4.4: Multi-wavelength lens-less imaging using angle sensitive pixels. (a) Photomicrograph of the fluorescent pinhole array overlaid on the chip. 30 µMol l-1 ATTO 465 and ATTO 490LS fluorescent dye is dispersed in 1.5% agar simulating scattering tissue. The agar is overlaid on a pinhole array and illuminated from the back. (b) Recorded data from MIMAS imager chip. (c) Calibration of recorded lifetimes for 30 µMol l-1 green, red and undyed agar. Time series for the mean recorded data, separated in green and red color channels through blind source separation. (d) Data separated into color channels after blind source separation. (e) Inverse image for both wavelength channels with detail plots representing the summation in the X and Y dimensions.
Figure 4.5: Localization in scattering medium. (a) Photomicrograph of 45 µm polymer beads containing Fluoresbrite YG, dispersed in 1.5% agar simulating scattering tissue. The probe is inserted at \( z = 0 \) orthogonal to the direction of excitation light. (b) Recorded data from MIMAS imager chip at 0ns delay. (c) Time series for mean recorded data, estimated fluorescent signal and excitation background through blind source separation. (d) Estimated scattered illumination background and fluorescent data after blind source separation. (e) Volumetric localization through an \( L=6 \) sparse representation, compared with fluorescent bead ground truth. (f) Lifetime calibration for FITC as a function of environment pH. The lifetime fingerprint of pixel 370 reveals the Fluoresbrite is encapsulated in an environment of pH=9. 

Figure 4.6: Ex-vivo recording. (a) Brightfield image of the whole brain slice laminated over the chip. The fluorescence image reveals the beads distributed through the tissue. Brightfield image of the location over the chip. (b) Recorded data, plotted alongside the background recorded for illumination of brain slice in the absence of beads. (c) Inverse image of the data recorded, compared to the true locations denoted by red crosses. The contour is drawn around all values corresponding to 10% of the maximum value of the image.

Fig. 5.1: Image vision system applying compression at both the emitter and detector side. The computational mask compresses the three-dimensional field of view onto the two-dimensional photodetector array. An ensemble of individually switchable micro-LEDs sparsely illuminates the field of view.

Fig. 5.2. Numerical benchmark performance of DISTINCT and select competitors as a function of the number of sources \( L \). The system carries \( M=1024 \) detectors and emitters, and the dimensionality of the problem is \( N=6468 \). (a) Error rate, defined as 1-accuracy (b) Computational
time required by each of the algorithms to solve the image reconstruction problem. (c) Figure of Merit, defined as the inverse of the error rate times computation time. .............................................. 81

Fig. 5.3: Numerical benchmark performance of DISTINCT and select competitors as a function of the problem complexity $N$. The system carries $M=1024$ detectors and emitters. (a). Error rate, defined as 1-accuracy (b) Computational time required by each of the algorithms to solve the image reconstruction problem. (c) Figure of Merit, defined as the inverse of the error rate multiplied by the computation time........................................................................................................... 84

Figure 6.1: OLED-electrophysiology frontend. (a) Device constructed by the vertical integration of recording electrode stack and an OLED stack. (b) Photomicrograph of the electrodes hydrostatically laminated on the OLED. The transparent PEDOT:PSS opening is $19 \times 19 \mu m$, through which we observe OLED photoemission. (c)................................................................. 94

Figure 6.2: Flexible transparent electrodes. (a) Impedance characteristic over the bandwidth of interest for different thickness of PSS:PEDOT on gold electrodes. (b) Noise power spectral densities of the stack comparing OLED on ($V_f=5V$) and off ($V_f=0V$). (c) Optical transmission in the visible spectrum for select PEDOT:PSS thicknesses. (d) Transmission versus PEDOT:PSS thickness for select wavelengths. (e) Artifacts induced by increasing levels of illumination at 490nm. (f) Ten electrical stimulation pulses are injected at $t=0.5s$, and recorded from the electrodes neighboring the stimulation electrode. The injected current saturates the amplifiers and obscures the recording for upto a second after the stimulation current has been turned off. ................. 96

Figure 6.3: Spatial characterization of OLEDs. (a) Confocal images for (i) OLED in solution and (ii) neural tissue with thickness $t_{slice} = 10 \mu m$ and (iii) thickness $t_{slice} = 100 \mu m$. (b) Normalized photon count detail at $Y=50 \mu m$ indicating scattering profile increasing with slice thickness. (c) Absolute count rate across $Y=50 \mu m$. (d) Integrated photon count over a circular area with $15 \mu m$
diameter centered above the OLED, indicating the photon flux on the neuron surface as a function of depth. (e) OLED optical emission as a function of applied forward voltage. (f) The emission spectrum is tuned to the channel rhodopsin absorption spectrum.

Figure 6.4: Optogenetic stimulation and recording of neural spike trains. (a) Slice laminated on top of the OLED and electrode sandwich. OLEDs turned on over the region covered by the electrodes. (b) Raw data of electrophysiological recordings obtained by PEDOT:PSS (red) and tungsten (black) electrodes. (c) Recording time series. The green shaded regions mark where we enabled the OLED for 10 seconds at a stimulation frequency of 40 Hz. Spectrogram based on the Gabor wavelet transform reveals enhanced activity in the 0.1-100 Hz bands. Detailed plots for each stimulation window shows the activation with delayed onsets. (d) Recorded multi-unit activity during stimulation. Top left panel is the waveform of the activity averaged from 50 traces and top right panel is raw data of sample spikes. Bottom panel shows the number of spikes before and during stimulation. (e) Wild-type brain slice control experiment. Apart from spontaneous spikes, no global activity is observed.

Figure 7.1: OLED-on-CMOS stimulation probe. (a) microphotograph of CMOS die of four-shank probe containing 256 OLED anode drivers and three cathodes per shank. The CMOS chip is bonded to a PCB which doubles as an insertion guide for neural implantation. (b) The CMOS circuitry is contained underneath each anode pad, and contains local addressing, level shifting and decoupling capacitance. (c) Fabrication flow from CMOS die to implantable probe form factor, with annotated silicon (light purple), active CMOS (dark purple), copper interconnect (orange) and aluminum contacts (gray). (i) Trench definition by laser cutting. (ii) Backside thinning removes excess silicon down to 50 µm. (iii) Direct evaporation of OLEDs on shanks. (iv) Encapsulation with nanolaminate and parylene. (d) Layer architecture of the orange and blue
emitting OLEDs integrated on the shanks. (e) SEM/FIB cross-section of the completed OLED-on-
CMOS stack with Al contact pad covering the Cu interconnect, OLED stack, and thin film
encapsulation consisting of two pairs of parylene-C and oxide nanolaminate. ......................... 107

Figure 7.2: Characterization of OLEDs of shank probe (a) Photograph of orange probe with 512
of the 1024 OLEDs activated over the four shanks in a checkerboard pattern. (b) Photograph of
blue probe with all OLEDs on. (c) False color micrograph showing emission intensity over a single
OLED pixel on the probe. (d) The emission spectra for the orange and blue probes are optimized
for the channelrhodopsin variants ChRmine and ChR2, respectively. (e) Emitted power and current
as a function of forward voltage for representative orange and blue pixels on CMOS devices. (f)
OLED emission power as a function of pulsing frequency. (g) Monte Carlo simulation of the
optical power projected into scattering neural tissue from a single blue OLED (left) and a cluster
of four OLEDs (right) for operation at 7V. Contours correspond to different light intensities in the
tissue are annotated. (h) Detailed plot of the projected optical power as a function of distance from
the probe for the conditions in (g). The required optical power densities for ChR2 and ChRmine
are annotated. (i) Temporal evolution of normalized optical power for a blue OLED driven at 6.25
MHz pulsing frequency. (j) Stability of OLED emission power over a 12-hour stress test. (k)
Histogram of optical power density emitted by orange pixels operated at 7 V forward voltage.

.......................................................... 108

Figure 7.3: Large scale, hyper-localized optogenetic stimulation in vivo. (a) Photomicrograph
of neural insertion of blue OLED shank. Co-implantation of the flexible multi electrode array and
rigid blue OLED shank is illustrated. The inset shows the multi electrode array with 32 transparent
electrodes at a 49 µm pitch. (b) The noise spectra of the electrodes before and during blue OLED
activation as a function of frequency. (c) Induced single-unit activity during OLED operation. Left
side illustrates the gliding illumination window achieved by sequential activation of blue OLED groups from surface to deeper layers. Spike waveforms acquired from electrode sites 3 and 4 show synchronized superficial layer 1 single-unit interneuron activity (green and cyan circles). In proximity to electrode 10 in layer 2/3, an excitatory pyramidal neuron (red triangle) is activated synchronously to the blue OLED pulsing. Detailed waveforms are averaged over 50 spiking events. Corresponding autocorrelograms of each spike is plotted on the bottom part. (d) Matrix of the firing histogram of each neuron during stimulation pulses applied by each blue OLED group. Dashed blue squares represent the applied stimulation pulses with 100 ms duration. Every pulse is applied for a total count of 100 times. For the corresponding OLED group activation, the responsivity of neuron 1, neuron 2 and neuron 3 were 63%, 61% and 88% respectively. 

Figure 7.4: Single orange OLED stimulating a single neuron with ChRmine. (a) Spike waveform of the stimulated pyramidal neuron as a function of electrodes placed with 49 µm pitch. (b) Spiking histogram of the corresponding neuron averaged over 200 pulses with 10 ms width. (c) Illustration of the firing rate view of the corresponding neuron during a 900 seconds window. Three separate trains of 200 orange OLED stimulation pulses at 2.5Hz (illustrated on top part) are applied for a duration of 80 seconds each.

Figure 7.5: Local field potential (LFP) modulation in vivo. (a). LFP modulations in the delta, theta and alpha bands are reproducibly induced by simultaneous activation of 128 orange OLEDs in ChRmine injected mice. Square pulses represent the time of applied stimulation. Corresponding time-frequency colormap is placed on bottom left. LFP power spectra as a function of frequency before (black) and during (red) stimulation within 0-50 Hz range is plotted on bottom right. (b) Induced LFP modulations as a function of applied power emission for 650 nW, 1600 nW and 3000 nW is plotted respectively. Square pulses represent the time of applied stimulation. Bottom row
illustrates corresponding LFP power spectra as a function of frequency before (black) and during (red) stimulation within 0-50 Hz range. (c) Figure illustrates LFP modulation fallout as a function of distance from the active OLED groups over a distance of 1500 µm. Corresponding power spectrums are located at bottom showing the LFP power at top electrode (black) and bottom electrode (red). ................................................................. 118

Figure 8.1: A photonic crystal (PC) simulated on top of the OLED shapes the emitted light field. (a) 3D view of the photonic crystal etched into the passivation layer. (b) Sideview of the anode-OLED-cathode-passivation stack with the crystal etched into the passivation layer. (c) Simulated E-field inside the photonic crystal and resulting far-field emission profile for different pitch values of pitch = 150nm, (d) pitch = 200nm, (e) pitch = 250nm. ............................................................... 126

Figure 8.2: Time-gating measurement using the LinoSPAD from (a) the NKT EXU6-pp mode locked laser (b) the Cree GaN LED .................................................................................................................. 129

Figure 8.3: SPAD optimization in a design of experiments (DOE) characterization for the four best performing layer structures. T1-4 are chamfered square, T5-T8 are circular geometry. The layer structure, Photon Detection Probability (PDP) and I-V curve is drawn for (a) Virtual P-epi Guard Ring (b) Standard P-Well Guard Ring (c) Deeply implanted High-Voltage-P-Well guard ring (d) High-Voltage-P-Well guard ring (e) Instrument Response function for all types. (f) By connecting the square N-doped buried layer (NBL) across all SPADs, and sharing the P+ substrate connection, the fill factor is increased at the cost of cross-talk. ................................................. 132

Figure 8.4: Combined OLED and electrophysiology system on CMOS................................. 137

Figure 8.5: µLED & photodetector shank. ........................................................................... 139

Figure 8.6: autonomous system for freely behaving mice. The OLED shank and recording electrodes are co-implanted and cemented on the head. A flex cable connects these two front ends
to a battery-powered controller PCB. A computer sends stimulation patterns and receives compressed data events.
List of Tables

Table 1: Electronic sensor modalities for sensing and inducing neural activity....................... 4
Table 2: Specification targets for design of the OptoSAM system ........................................ 23
Table 3: Comparison table with prior spad imagers. PDP and DCR are reported at respective excess bias voltages. PDP is reported as the maximum value over all wavelengths. .......................... 31
Table 4: The DISTINCT algorithm ......................................................................................... 78
Table 5: Achieved results versus target specifications for the OptoSAM system...................... 123
1. Introduction

Neuroscience is the scientific backbone for neuromedicine, which holds great promise for improving people’s wellbeing in the 21st century. As society has developed to provide more physical comfort, many individuals now experience their minds ageing faster than their bodies. Additionally, advancing neuroscience provides the basis for treating genetically or trauma-induced mental illnesses, estimated by the NIH to afflict approximately 1 in 4 humans throughout their lifetime.

While very little is known about brain function and the exact origin of mental illnesses, one popular view states that specific brain regions are tied to specific functions. Understanding the neuronal pathways inside and between particular regions could therefore lead to understanding of the corresponding function and disfunction. In order to understand these pathways of interest, neuroscientists require the correct investigational tools.

1.1. Brain computer interfaces

The advent of computers allows any type of sensor platform to be recorded, processed and stored in a massive fashion. As the brain contains an estimated hundred billion neurons with one quadrillion interconnects, computers are essential to be able to make sense of even the tiniest fraction of the brain. The emerging field of neuroelectronics tries to build brain computer interfaces (BCIs), capable of performing elementary operations directly on neural circuits. This field is very young and as of 2022, the most widely deployed neuroelectronics are those interfacing only with the peripheral nervous system. As many as 1 million people yearly receive either a pacemaker, cochlear implant, or an assisted breathing implant. These implants are relatively
easily inserted and can be explanted if necessary, letting the surrounding inflammation heal again without too much permanent damage.

For the brain however, any implantation procedure carries enormous acute risks of hemorrhage. Long-term effects of implants on mental function are unknown. Therefore, implanted neural interfaces that target deep-brain modulation for alleviating Parkinson’s disease or depression, are proof-of-concept devices deployed very sparsely.

The progress of useful brain-computer interfaces is an archetypical chicken-egg problem: due to a lack of useful devices, we refrain from human experimentation. Vice versa, a lack of use cases obfuscates the efficient design of new devices. The field of neuroelectronics hopes to break through this challenge. By continuously improving the underlying technology, someday the performance will be sufficient to tip the risk-reward ratio in favor of human experimentation.

None of the devices proposed in this thesis are designed for human experimentation. Instead, experimentation is performed on mice, which carries four advantages. First, mice can be genetically altered to enable the optical experiments presented in this thesis. Second, the small form factor greatly simplifies surgical and behavioral experimental setups. Third, mice exhibit simplified versions of the essential cognitive and emotional functionality found in humans. Fourth, fast reproduction lowers costs and provides large experimental numbers to ensure statistical relevance.

On the other hand, mice are wildly different from humans. Neuron sizes and structure are different for each species. Neural firing patterns, revealed by electrophysiological oscillations, have completely different origins and effects between mice and humans. Size matters too, as the probe presented in this thesis can access the entire mouse deep-brain but barely scratches the human brain surface. And most importantly, while optogenetics and fluorescent tagging in
humans has been explored, they are very far from being commonly available. BCIs for humans are a completely different class from those presented in this thesis. Still, I believe this thesis presents a minor though crucial step towards the ultimate goal of realizing the ideal BCI.

1.2. The ideal BCI and sensor modalities

The ideal BCI would have the following specifications:

- Crosstalk-free neural readout and activation
- Soma-level resolution (5–20µm)
- Temporal resolution matching the maximum neuron firing rate (~1ms)
- Trauma-free implantation
- Chronic operation over a lifespan without neural damage
- Field-of-view throughout the entire brain
- Low power consumption for long battery life

To achieve the interface between the electromagnetic properties of the neurons and the computer itself, there are many potential modalities. Table 1 gives an abridged overview of current electronic neural interrogation and activation modalities with their respective state-of-the-art performance metrics.
Table 1: Electronic sensor modalities for sensing and inducing neural activity

<table>
<thead>
<tr>
<th>Sensor modality</th>
<th>Neural interface</th>
<th>Invasive-ness</th>
<th>Resolution</th>
<th>Field of view</th>
<th>Common usages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic</td>
<td>Hemoglobin levels in blood vessels</td>
<td>None</td>
<td>~1mm(^{13,14})</td>
<td>Whole brain</td>
<td>Structural MRI for tumor detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Functional MRI for readout of regional activity</td>
</tr>
<tr>
<td>Electrical EEG</td>
<td>Low-frequency synchronized local field potentials</td>
<td>None</td>
<td>~1cm(^{15})</td>
<td>Whole brain</td>
<td>Seizure detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Electrical stimulation of large neuronal clusters(^{16})</td>
</tr>
<tr>
<td>Electro-physiology</td>
<td>Extracellular action potentials, local field potentials</td>
<td>High; needs direct access to brain</td>
<td>Single neuron</td>
<td>~50μm from sensor for single neuron activity</td>
<td>Measuring direct single neuronal response</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Activation of small neuronal clusters non-selectively</td>
</tr>
<tr>
<td>Optical fluorescent imaging</td>
<td>Fluorescent proteins: light sensitive cell membrane proteins</td>
<td>Very high; needs direct access to brain and requires fluorescent marker expression</td>
<td>Single neuron</td>
<td>~200μm from sensor for 1-photon imaging, &gt;1mm for 2-photon imaging</td>
<td>Imaging of fundamental structures</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High resolution functional imaging</td>
</tr>
<tr>
<td>Optical optogenetic stimulation</td>
<td>Channelrhodopsins: photocurrent generating cell membrane proteins</td>
<td>Very high, needs direct access to brain and rhodopsin expression</td>
<td>Single neuron</td>
<td>~200μm from emitter for single neuron.</td>
<td>Activation of specific neuron types and regions</td>
</tr>
<tr>
<td>Optical fNIRS</td>
<td>Absorption of infrared light as a function of Hemoglobin</td>
<td>None</td>
<td>~10mm(^{17})</td>
<td>Only superficial layers due to absorption</td>
<td>Readout of activity from superficial layers</td>
</tr>
<tr>
<td>Ultrasound</td>
<td>Acoustic impedance differences between tissue layers</td>
<td>Low, removal of hair often required</td>
<td>~5 mm(^{18})</td>
<td>Whole brain</td>
<td>Stimulation of activity(^{19}) Opening of blood brain barrier for chemical transfer(^{20})</td>
</tr>
</tbody>
</table>
1.3. Access to deep-brain

From Table 1, we find a trend across sensor modalities. Resolution and invasiveness are strongly inversely related to the field of view (FoV). To infer the behavior of single neurons, we need a direct interface to harvest their very weak emitted signals. To activate a single isolated neuron, direct access is equally necessary to avoid loss of focus due to reflection and absorption on the interfaces between the multiple layers of skin, cranium, spinal fluids and tissue\textsuperscript{21}.

After single neuron resolution is achieved on the brain surface, accessing the deep brain involves miniaturizing the same interface into an implantable form factor. For example, the recording electrodes on an electrophysiology system can be miniaturized and made biocompatible\textsuperscript{22}. These electrodes can then read from large neural clusters simultaneously\textsuperscript{23}. While lacking spatial discernibility, the fast readout speeds and low noise aid in spike sorting\textsuperscript{24}. Scaling of CMOS will progressively allow more pixels and higher performance electronics on smaller packages.

Miniaturization of optical microscopes seems far less likely, as they require large lasers, focusing optics, and filters which displace significant tissue volume. Systems that employ these three components in their traditional form have reached a scaling limit\textsuperscript{25,26}. Implantable graded index (GRIN) lenses and multimode fibers\textsuperscript{27} can extend the objective several millimeters into the tissue. However, these cylindrical devices have a poor field of view (FoV) to displaced tissue ratio’s as they only image from the lens/fiber distal end.

These challenges have lagged the development of optical interfaces behind those using the electrophysiological paradigm\textsuperscript{28}. From these limitations, it might appear that the optical paradigm is infeasible for realization of direct neural interfaces inside the deep brain.
1.4. The optical paradigm for deep-brain interfaces

However, a deeper comparison between the electrophysiological and optical paradigms for deep brain access shows how the optical domain carries three benefits over electrophysiology. The first benefit pertains to the optical interface itself. As neurons are not optical sources, the optical interface starts with a protein expressed in the neuron membrane. In the case of functional imaging, these fluorescent proteins are quenched in the absence of sodium ions, i.e., when the cell is not actively firing. To activate neurons optically, channelrhodopsins are expressed as transmembrane proteins which open to conduct cations out of the cell\textsuperscript{29,30}. Irradiance with a sufficient light intensity will transfer enough ions to trigger a neuronal depolarization. As neurons require genetic modification to express these proteins, the functional and regional specificity can be genetically encoded as well. This encoding allows to precisely isolate the neurons of interest, tremendously aiding the investigative efforts.

Secondly, as opposed to electrical signals, the optical signal can be focused and selectively received. This focusing further aids in spatial selectivity, and forms the basis for image reconstruction. This effect can effectively extend the field of view in the axial direction, as it allows to isolate signals originating further away from the sensor.

Thirdly, spectral sensitivities can be exploited to build nonoverlapping channels within the same device. For example, functional imaging can be implemented in the blue and green spectra while channelrhodopsins sensitive to yellow and red are responsible for targeted neural activation.
1.5. Goals of this thesis

Recognizing the major advantages that optical neural interfaces promise, the goal of this thesis is twofold. The first goal is to understand the challenges that obstruct the use of optics in the alive deep brain. The second goal is to engineer and validate solutions to these challenges.

In this thesis, we design, fabricate and validate the OptoSAM platform, a mixed-signal complementary metal-oxide semiconductor (CMOS) chip. The thesis describes how the OptoSAM contains two subsystems, each engineered for its separate application: optical fluorescent imaging on one hand and optogenetic stimulation on the other. To achieve necessary performance for each subsystem, the chip is enhanced through novel nanofabrication and data processing schemes.

1.6. Structure of this thesis

Chapter 2 establishes the design space for the OptoSAM chip. This design space is constrained by first looking at the desired neuroscientific applications. Then, the state of the art is reviewed to understand past successes. More importantly, shortcomings in the state of the art are enumerated. It analyzes which shortcomings are feasible to overcome, in order to maximize the novelty of the OptoSAM platform. A specification table is drafted to summarize all requirements of both the OptoSAM chip and its accompanying postprocessing and data processing schemes.

Chapter 3 describes the design and characterization of the 1024-pixel SPAD imager subsystem. The chapter shows how the CMOS die is transformed into a probe and packaged for neural implantation. It describes the digital and analog circuitry necessary to perform time-gating for excitation light rejection. It analyzes the image reconstruction performance of integrated metal angle sensitive pixel grating frontends. The imaging performance limitations of the chip are discussed.
Chapter 4 improves upon the limitations of the imager subsystem from chapter 3. It describes the development of color filters integrated with the ASP into a novel MIMASP frontend structure. It expands the analysis for improving the MIMASP layout to optimize image reconstruction performance. It demonstrates how the time-gating technique can demix the different lifetimes of different fluorescent proteins. This improves the separability, as well as measuring other quantities such as pH through lifetime modulation. It compares imaging quality of fluorescent beads between free space, agar and mouse brain tissue ex-vivo.

Chapter 5 describes a novel reconstruction algorithm for increasing the imaging resolution of the device. The DISTINCT algorithm incorporates a priori known information from both the illumination and detection light fields. It extends the detector-side light field optimization analysis from chapter 4 to optimization of the illumination patterns.

Chapter 6 describes the design and characterization of a novel structure necessary for validating optogenetic stimulation in-vivo. The vertically integrated pixel is able to simultaneous activate neural activity and record artifact-free electrophysiological signals. The transparent organic electrode conductor (PEDOT:PSS) allows vertical integration and enables the underlying OLED to emit through the electrodes. This increases density and simplifies measurement analysis, as emitter and detector are in the exact same location. Both OLED and recording stacks are highly biocompatible due to their fully organic and flexible nature, making them ideal for long term implantation.

Chapter 7 integrates 1024 OLED emitters onto the OptoSAM subsystem for optogenetic stimulation. It describes the fabrication process necessary to integrate either red or blue OLEDs, and fully characterizes the resulting optical performance. It demonstrates how activation of a single
OLED induces deep-brain single-neuron spiking activity. This activation is scalable over a $4 \times 1$ mm field of view. We validate the results in mice hippocampus *in-vivo* labeled with both ChR2 and ChRmine, an ultra-sensitive red-shifted channelrhodopsin. We demonstrate independent control of multiple single unit neurons within a single insertion. By increasing the emission power, a deeper stimulation field is achieved up to $150 \mu m$ which results in the activation of synchronized multi-unit activity. By scaling the number of active OLEDs to 128, we activate globally synchronized activity, as revealed by low-frequency LFPs.

Chapter 8 ties together results and conclusions from chapters 3-7. The engineering results from chapters 3-7 are first compiled, and compared against the specification table as presented in chapter 2. The discussion reviews in-depth all specifications that failed to meet. It also discusses any challenges faced commonly throughout chapters 3-7. From these conclusions, the proposals for future research are twofold. First, in section 8.4, three sensing platforms are proposed. These platforms incorporate the strengths of the work in this thesis and improve upon its limitations. Second, section 8.5 proposes future experimental questions to explore. These experimental questions could be answered by either the platform presented in this thesis, or by the three proposed sensing platforms.
2. Design space for the OptoSAM all-optical deep-brain neural interface

In this chapter, we start creating the design space for the OptoSAM system looking at the desired neuroscientific applications. Afterwards, the state of the art is reviewed to understand past successes. Shortcomings in the state of the art provide a list of challenges for the design. Challenges which are deemed feasible to be addressed by this thesis are selected. With the desired applications and improvements in mind, quantified trade-offs and assumptions are made that guide implementation. The result from this chapter is a specification table, summarizing all requirements of both the OptoSAM chip and its accompanying postprocessing and data processing schemes.

2.1. Potential applications for all-optical deep-brain BCIs

Before diving into the design space, we need to consider the desired applications. The applications are split between imaging of fluorescently tagged neurons on one hand, and optogenetic activation of neurons on the other. As neurons are neither optical sources nor optically receptive, the neuron itself has to be modified to achieve both types of optical sensitivity.

2.1.1. Fluorescent imaging

Optical neural imaging was first enabled through fluorescent tags such as EGFP, a fluorescent protein distilled from jellyfish. The genetic code for EGFP can be connected to a genetic promotort that drives expression in the somas of interest. By performing structural imaging chronically, neuroscientists could study neural structural plasticity during long-term learning experiments. Or, real-time monitoring of structural neurodegeneration could give insights into
progression of mental degeneration. Alternatively, if tumor cells could be uniquely fluorescently tagged, an optical imaging probe could guide tumor removal on the cellular level\textsuperscript{11}.

Additionally, neural activity can be imaged by Green Calcium Modulating Protein (GCaMP), a calcium-sensitive version of EGFP. As extracellular calcium is a proxy for activity, this GCaMP is engineered to fold its protein structure and thus quench its fluorescence in the absence of calcium. However, this folding dynamic is slow compared to the actual intracellular calcium pumping. This gives GCaMP a decay rate of about 150 ms\textsuperscript{31}. Individual spikes cannot be revealed, which limits feasibility of blind source separation techniques (BSS)\textsuperscript{28}. On the other hand, the high brightness and longer allowed integration times makes GCaMP the functional indicator of choice for many neuroscientific experiments. Within the GCaMP\textsubscript{6}\textsuperscript{32} and GCaMP\textsubscript{7}\textsuperscript{33} family, there are versions for higher bandwidth (GCaMPxf) and enhanced brightness (GCaMPxs).

Imaging GCaMP in the deep brain can aid neuroscience. For example, exact associations between behavior and neural location and type can be inferred. Fluorescence imaging in the amygdala\textsuperscript{34} has shown what cell types are involved in trauma extinction.

On the other hand, genetically encoded voltage indicators (GEVIs) possess kilohertz bandwidths at the cost of reduced brightness\textsuperscript{35}. These GEVIs reveal individual spikes, enabling blind source separation. BSS is a strategy often used in electrophysiology\textsuperscript{24,36}, requiring temporal oversampling of 10\times the spike bandwidth. As neurons cannot refire before completing a repolarization time of approximately 1-2ms, two slightly delayed spikes registered within this time frame therefore have to originate from two distinct neurons. They can be separated even if these two neurons are spaced closer than the imager’s spatial resolution. This separability is further enhanced by identifying neuron physiology from their characteristic spike waveform\textsuperscript{37}. If GEVI
neurons can be accurately imaged, spatial and temporal resolution and be dramatically enhanced, allowing for imaging on a much denser scale.

2.1.2. Optogenetic stimulation

The understanding of the link between cell-level neural activity and behavior is rapidly progressing. Many discoveries over the past decade have been enabled by technologies capable of determining how neural circuit activity controls behavior. Researchers can now control activity in pre-defined neuronal populations while examining the consequences on behavior and physiology\textsuperscript{38,39}. Repeated optogenetic stimulation modulates the desire for social interactions\textsuperscript{40}, albeit in a highly constrained environment. Neurons in the V1 visual cortex are directly mapped to retinal photoreceptor cells, and neural stimulation restores the perception of vision\textsuperscript{41}. In smaller animals, laser beams can completely control locomotion\textsuperscript{42,43}, demonstrating potential for manipulation of locomotion in larger animals.

In recent years, advances in new generation opsins with fast kinetics, high photoinduced currents, cell specific and sub-cellularly restricted expression made it possible to target specific neural function\textsuperscript{44}. Any electrical stimulation experiment can therefore be ported to optogenetics to exploit the specificity and resolution of optical platform. The development of optical probes lags behind probes for electrical stimulation, explaining the lag in popularity for optogenetics in freely behaving animals.
2.2. State-of-the-art for integrated optical neural interfaces

This section describes recent seminal publications, and reviews their strengths to import into our system. More importantly, the shortcomings of the state of the art are analyzed. This analysis is executed separately for the fluorescent imaging and optogenetic stimulation subsystems. The chapters, corresponding to where improvements are made upon those shortcomings, are referenced.

The outcome of this section is a list, guiding how our design should focus efforts to maximize the systems novelty. While this section is concise and high-level, a more detailed state-of-the-art literature review is made within each chapter, pertaining to the specific design challenges in that chapter.

2.2.1. Fluorescent imaging

Conventional fluorescent microscopes use four microscope elements; an objective forms images by focusing light, a light source excites the fluorescent target, a detector harvests fluorescent signal, and an optical filter removes excitation light from the detector. Miniaturized versions of this conventional microscope assist in freely behaving animals but are severely limited in depth by light scattering and absorption in tissue. Imaging can be extended at depth by implanting graded index (GRIN) lenses or optical fibers. This method involves implanting an optical fiber in targeted regions of the brain, and then employing it for one-photon fluorescence excitation of reporters at the distal end of the fiber. However, the resulting displaced tissue volume is high for these devices. Micro-endoscopes typically have diameters around 500 µm, causing significant destruction of tissue along the implantation path. Although micro-endoscopy resolves the issue of light delivery from within remote brain tissue, illumination typically occurs only within
one optical plane near the tip of the endoscope. The approach is feasible for acute measurements in the location of the tip of the fiber, using direct CCD-imaging\(^{48}\) or probe-based confocal laser micro-endoscopy\(^{49}\) but expanding to large volumes will cause more implantation damage.

Fully integrated imagers\(^{2}\) aim to miniaturize all four microscope elements, as well as the need for external connectivity. Coded aperture masks are shown to replace the objective with a 200 µm thick integrated frontend\(^{50,51}\), moving the burden of image reconstruction into the digital domain. Angle sensitive pixels eliminate the focal distance altogether by integrating compressive sensing in-pixel\(^{52-55}\).

SPAD imagers\(^{56}\) are promising photodetector devices as they enable time-correlated lifetime measurement of fluorescent markers\(^{57}\). While their SNR may be equivalent to or higher than conventional photodiodes, SPADs can resolve weak signals in photon starved environments such as biological imaging. Integrated µLEDs are necessary to provide excitation light adjacent to the photodetector array\(^{58}\), and to uniformly illuminate the tissue of interest. Finally, thin-film filters\(^{59-61}\) are necessary to separate the fluorescent signal from the excitation light source in a minimal form factor.

### 2.2.2. Optogenetic stimulation

While imaging requires light delivery and detection, optogenetics necessitates only the former. Probes for deep-brain optogenetics have therefore enjoyed considerably more progress than imagers\(^{62}\). Compared to microscopes, a probe’s ratio of field of view to displaced volume can be many orders of magnitude better, as emitting sites populate the entire implanted surface area. This dramatically increases the total field of view, while the total probe cross-section is strongly decreased by thinning down the probes in order to remove unused substrate volume.
Integrated waveguides can be used to direct an externally coupled laser through shank probes\(^{63,64}\). While these methods enjoy advantages provided by recent advances in integrated photonics, they lack scalability as the waveguide density is fundamentally limited by the size of the optical mode. For optogenetic stimulation studies, power requirements do not justify the cost and complexity of external laser setups. Dual color (405/635nm) integrated laser diodes can couple into local waveguides\(^6\) for an all-electronic, fiber-less package. To eliminate the fiber connector, the design implements a separate laser diode for each channel, an intrinsically unscalable method.

Electronic routing does not suffer fundamentally from density limitations, and integrated methods exploit this advantage by converting electrons to photons in-pixel. GaN \(\mu\)LEDs\(^6\) have been used at the neural interface for efficient electron to blue photon conversion. Single LEDs with neuronal dimensions were proven to achieve sufficient optical pulse power and repetition rate for optogenetic activation \textit{in-vivo}\(^6\). However, integration of LEDs was demonstrated in passive configurations\(^6\)–\(^7\), i.e. without an active backplane. Due to space limitations in a narrow shank configuration, a large-scale parallel connection of multiple emission sites to the base is not feasible. Passive addressing limits the number of LEDs therefore, precludes large pixel count due to lack of on-probe multiplexing.

\textbf{2.2.3. Shortcomings in current state of the art}

None of the fully integrated imagers have successfully demonstrated deep-brain neural imaging \textit{in-vivo}. An integrated imager that can image fluorescent targets in scattering medium is presented in chapters 3 and 4.

All state-of-the-art integrated imagers lack means of deep-brain light delivery. Without integrated light delivery, the deep brain remains obscured as light scattering and absorption
constitute a combined exponential decay length of around 200µm. The optogenetic fiber-based devices are not scalable. They rely on expensive, bulky, external optical equipment which prohibits deployment en masse. Experimentally, fiber connectors are generally unpreferred since they are fragile, difficult to align and often bulkier than electronic connections. In general, state-of-the-art optogenetic probes sacrifice precise spatial selectivity versus a large field of view. The GaN-on-Si probes lack large fields of view which prohibits investigating multiple brain regions independently and simultaneously. High-power emission from a few small sites to synchronize large neural populations can locally heat tissue. Distributed emission over a field of view that matches the size of these populations has not been demonstrated. Integration of ultra-dense, scalable, large field of view light delivery in the deep brain is presented in chapter 7.

As deep-brain imagers lack the illumination capability, there are no sparse reconstruction algorithms that specifically exploit co-optimization of structured illumination and detection light fields. An algorithm that improves fluorescent imaging performance under these circumstances is presented in chapter 5.

Optogenetics needs to be combined with a recording modality to reliably cross-validate the results. However, no proposed devices achieve this in a fully scalable way. Freely behaving optogenetic probes depend on direct behavioral verification, limiting the usefulness in situations where certain negative behavior needs to be suppressed. Furthermore, the combination of integrated optogenetics and electrophysiological recording modality is only shown in acute experiments. Using recording and optogenetics in a full feedback loop in freely behaving mice remains undemonstrated with integrated photonics, only through fiber-coupled methods. Integration of electrophysiological verification with optogenetic stimulation in-situ is presented in chapter 6.
2.3. Project assumptions and trade-offs

In this chapter so far, we have discussed both the target applications for optical neural interfaces, and the achievements and shortcomings in the state-of-the-art. With this knowledge, we can start writing specifications for our deep-brain neural imager and stimulator. These specifications require making trade-offs and assumptions along different themes. The goal of this section is to make quantified assumptions and trade-offs in order to establish a list of specifications for the neural interface. This section also selects methods that will be used in the implementation and verification chapters (3-7) of this thesis.

2.3.1. Minimally invasive form factor

To preserve the cell quality, the device needs to prevent scar tissue formation, blood clotting and cell death. The device cross-section requires to be minimum to prevent vessel puncturing. A linear array probe with active surface along the side face provides the highest device surface area for any given cross section. While the width contributes to useful device area, the thickness does not and should be minimized until warping arises from internal material stress. When fabricated with ultrasmooth sidewalls and a sharp insertion tip, these probes are proven to provide healthy cell function for up to 8 weeks after implantation\textsuperscript{74}. The minimum shank-to-shank spacing required to minimize dimpling is approximately 150 \( \mu \text{m} \)\textsuperscript{75}, at the cost of pixel density. To access the mouse cortex and hippocampus at depths up to 4mm, the shank should measure approximately 6mm in length to account for geometric margins.

Chronic studies outside the lab demand system integration into fully portable form factors. Therefore, we need the device to be as autonomous as possible. Any accompanying electronics should have a clear future path towards miniaturization.
### 2.3.2. Limited real estate for detectors and emitters

In many probe designs, passive routing inhibits aggressively scaling the number of pixels\(^{63,64,70}\). The CMOS platform allows ultra-dense integration of filtering, amplification and storage inside each pixel. Digital switches can multiplex the signals onto a bus for successive readout without loss of quality. The TSMC 130BCD technology node provides a good tradeoff between transistor size, cost and availability of photodetector devices\(^ {76,77}\).

The number of recording channels needs to maximized to decorrelate as many neurons as possible from the recorded detector ensemble data. The theoretically maximum number of distinguishable point sources\(^ {78}\) is equal to the detector channel count \(M\). In practical reconstruction algorithms however\(^ {79}\), this number of reliably resolvable neurons is approximately \(\sqrt{M}\).

To form an image from raw optical recordings, each pixel needs a unique spatial sampling profile\(^ {80}\). The image is then reconstructed by inverting the sampling profile. Conventionally, a lens performs this function by uniquely mapping each point on the focal plane to a pixel, removing the need for inversion. However, even the thinnest plasmonic lenses\(^ {81}\) require a focal distance between lens and detector. Furthermore, a lens functions for a pixel ensemble. In the linear array with low pixel count, the lens effect is strongly diminished.

Instead, we can give pixels individual spatial diversity by implementing angle sensitive pixels\(^ {53,82}\) (ASPs). These ultrathin (<5 µm) diffraction gratings modulate the detector light fields and are freely tunable in angular frequency, rotation and phase offset. However, these ASPs are not spectrally sensitive\(^ {83}\), requiring extra color filters to discriminate between different fluorescent spectral channels. Furthermore, they have not yet shown to achieve spatial resolution better than 50um\(^ {55}\).
Neurons measure 5-20 µm in diameter\(^8\), with larger neurons providing a larger optical signal. When pixels are spaced with a pitch smaller than the neurons of interest, signals from neighboring pixels will be highly correlated. This diminished return on scaling establishes a lower bound for pixel pitch. Additionally, larger pixels provide a higher fill factor on the probe, increasing total signal yield. Larger, fewer pixels require less unnecessary system complexity. Finally, to measure angles up to 45° with low cross-talk, we require a pixel width of twice the vertical passivation height, measuring 10µm in TSMC 130BCD\(^84\).

Fluorescent signals of transmembrane proteins are extremely weak. A 15 µm diameter pyramidal EGFP-stained soma is estimated to emit with a brightness\(^85\) equivalent to 10 μMol·l\(^{-1}\) of fluorescein\(^86\). To detect the fluorescence from distances of at least 100µm away, extremely high photocurrent amplification is required with very low input referred noise. Single Photon Avalanche Diodes\(^56\) are metastable, reverse biased photodiodes that generate very large reverse avalanche currents when a free carrier is energized to the conduction band. With the correct architecture, these devices reliably generate avalanche events from single incident photons\(^87\), while minimizing dark count avalanches from thermal events\(^88\). While regular photodiodes might have similar or better signal to noise ratio’s, the detection threshold for SPADs is orders of magnitude lower\(^89\).

While static structural imaging does not constrain the framerate, the Green Calcium Modulating Protein (GCAMP6f) exhibits rise time (\(\tau_{\text{rise}}=10\)ms) and fall time (\(\tau_{\text{fall}}=100\)ms) dynamics\(^90\), bringing the Nyquist sampling rate to \(f > 2 \times \frac{0.35}{\tau_{\text{fall}}} = 7\)fps. Much faster voltage indicators have bandwidths of 1KHz\(^91\), requiring multiple Kfps framerates. At these very short integration times however, the optical signal risks to be dominated by shot noise\(^92\). Exploiting the fast readout speeds of the CMOS platform, we can flexibly design for very high framerates and average individual frames if necessary.
2.3.3. Spectral rejection of excitation light

A fluorescent microscope uses dichroic mirrors and emission filters to reject the excitation light from the fluorescent signal with a rejection factor of $10^6$-$10^8$. For GCaMP6, the fluorescent emission peak wavelength is stokes shifted by approximately 30nm from the peak excitation wavelength. An integrated 500nm longpass filter will reject blue excitation light, while transmitting GFP/GCaMP6 fluorescent signal. Such a longpass filter is compatible with other, longer stokes shift fluorescent proteins.

In addition to spectral filtering, we can exploit the fact that GFP fluorescent lifetime decay ($\tau_{\text{GFP}}=4\text{ns}$) is much longer than the laser rolloff ($\tau_{\text{laser}}<100\text{ps}$). By disabling the SPADs during laser pulsing and reenable during the fluorescent decay, we hope to reject significant amounts of excitation light. The SPADs need excellent instrument response function (IRF) and the analog quenching electronics need to be matched to the capacitance and series resistance of the SPAD. Additionally, by rolling this window with a resolution of $\tau_{\text{GFP}}/20=200\text{ps}$, we should be able to estimate the lifetime decay with high reliability from a histogram. Construction of a histogram requires many repeated photon counts. Synchronizing the detection window to a mode-locked laser with repetition rate of 20MHz will give a maximum of $2.85\times10^6$ histogram elements at 7fps.

This combination of spectral filtering and time-domain rejection could be sufficient to distinguish fluorescent signal ($N_{\text{fluo}}$) from the background excitation counts ($N_{\text{BG}}$) and dark count noise ($N_{\text{DCR}}$). While the average of background and dark count noise can be subtracted, their shot noise variability cannot. The fluorescent signal is also shot noise corrupted. The total system signal-to-noise-ratio (SNR) is described as:
\[ SNR = \frac{N_{\text{flu}}}{\sqrt{N_{\text{flu}} + N_{\text{BG}} + N_{\text{DCR}}}} \]

2.3.4. Integrated light delivery in the deep brain

After integration of focusing elements, photodetectors and filters, the miniaturization of the excitation light source is the fourth and final step in miniaturizing all components of an optical microscope. For shallow imaging experiments through no more than 200 µm of tissue, we use external mode-locked lasers projected onto the FoV. At depths beyond 200 µm, scattering distorts the fluorescent signal such that cells of 15 µm can no longer be resolved\(^{96}\). Densely packed µLEDs can therefore illuminate the deep brain up to a distance of 200 µm from the shank.

The illumination light fields can be structured to aid image reconstruction performance\(^{97}\). A sparse reconstruction algorithm needs to be designed that optimally incorporates the illumination patterns in conjunction with the detection light fields. As the reconstruction requires a static image across all exposures within a frame, the effective frame rate will be reduced by \( K \), the number of illumination patterns.

For optogenetics, the µLEDs should match the spectrum, optical power and bandwidth of the used channelrhodopsins\(^{98-100}\). Channelrhodopsins generally have wide spectral sensitivities of over 50 nm. At the same time, we can employ red-shifted channelrhodopsins to decouple the optogenetic subsystem from the blue/green GCaMP6 imaging subsystem. The 0.1 mW·mm\(^{-2}\) optical threshold of Chrim\(^{101}\) allows us to aggressively downsize the power requirements on the LED compared to stimulation requirements for other channelrhodopsins, saving power dissipation and thus tissue heating. These channelrhodopsins have maximum bandwidths of about 40Hz\(^{102}\). However, switching with large current densities will introduce noise coupling into an electrophysiological recording system. To filter out these artifacts, the LED driver could pulse
with a frequency of 20KHz, an order of magnitude higher than the bandwidth of interest for electrophysiology (1Hz-3000Hz)\textsuperscript{23}.

Heating of brain tissue strongly modulates activity which is indistinguishable from and far less local than optogenetic activation results\textsuperscript{103}. Since blue photons are able to travel an average of 150 μm through tissue before absorption\textsuperscript{96}, large photon fluxes are permitted before direct photoheating becomes a problem\textsuperscript{68}. Instead, resistive loads inside the CMOS probe will heat significantly at high current consumption, prompting a point source model\textsuperscript{104} where a 1mW dissipation from a 10x10μm active area keeps the surrounding tissue within 38 °C.

2.3.5. Integration of concurrent recording and stimulation for in-situ verification

The described system depends on novel implementations on both stimulation and detection side. To correctly cross-validate either experimental subsystem, we need to implement them concurrently with off-the-shelf ground truth systems. Only then can issues be isolated and the performance be correctly compared to the state of the art.

To validate correct functionality of the optogenetic subsystem in-vivo, we will rely on a co-inserted electrophysiology array. The electrical recording site pitch can be on the same dimensions as the μLED pixel pitch. The electrophysiological readout needs to be in very close proximity to the shank, to emulate conditions as if they were integrated on the CMOS shank itself. To validate correct functionality of the fluorescent imaging subsystem, we use an Olympus BX51 epifluorescent microscope in 1-photon with the same illumination conditions. Verification of imaging in deep tissue is an unsolved issue and is out of scope for this thesis.
2.4. Design specifications for the OptoSAM system

After having carefully reviewed the trade-offs and assumptions in section 2.3, we enumerate a list of specifications needed to perform the applications described in section 2.1. This specification list, shown in Table 2, is an overarching set of design goals.

With these specifications, the goal of chapters 3-7 is to implement these specifications step-by-step and characterize the resulting performance. Each chapter also verifies the performance in biological environments by demonstrating the relevant biological experiments. Afterwards, we will review the specification list in section 8.1 according to the same table.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shank length</td>
<td>6mm</td>
</tr>
<tr>
<td>Shank width</td>
<td>100 µm</td>
</tr>
<tr>
<td>Shank thickness</td>
<td>50 µm</td>
</tr>
<tr>
<td>Technology node</td>
<td>TSMC 130nm BCDMOS</td>
</tr>
<tr>
<td>Pixel pitch</td>
<td>20 µm</td>
</tr>
<tr>
<td>Target fluorescent protein (excitation, emission wavelength)</td>
<td>GFP / GCAMP / FITC (ex. 480nm, em. 515nm)</td>
</tr>
<tr>
<td>Framerate</td>
<td>5 Kfps</td>
</tr>
<tr>
<td>SPAD Repetition rate</td>
<td>20MHz</td>
</tr>
<tr>
<td>Temporal resolution</td>
<td>200ps</td>
</tr>
<tr>
<td>Fluorescent excitation source</td>
<td>Mode-locked laser</td>
</tr>
<tr>
<td>Excitation source rejection</td>
<td>1:10⁶ contrast (OD 6) between excitation and fluorescent signal</td>
</tr>
<tr>
<td>Spectral sensitivity</td>
<td>Postprocessed color filters</td>
</tr>
<tr>
<td>Fluorescent imaging validation method</td>
<td>1-photon microscope</td>
</tr>
<tr>
<td>Reconstruction algorithm</td>
<td>Incorporate illumination and/or lifetime a priori knowledge</td>
</tr>
<tr>
<td>Photodetector type</td>
<td>Single Photon Avalanche Diode</td>
</tr>
<tr>
<td>Spatial resolution</td>
<td>20µm</td>
</tr>
<tr>
<td>Optogenetic emitter type</td>
<td>Organic Light Emitting Diode</td>
</tr>
<tr>
<td>Target optogenetic protein (activation wavelength)</td>
<td>ChRmine (590nm)</td>
</tr>
<tr>
<td>OLED repetition rate</td>
<td>20KHz</td>
</tr>
<tr>
<td>Reconstruction algorithm</td>
<td>Incorporate <em>a priori</em> knowledge about illumination structure and/or lifetime decay</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Validation method for optogenetic activation</td>
<td>Deep brain electrophysiology</td>
</tr>
<tr>
<td>Maximum tissue heating</td>
<td>1°C, or 1mW in-pixel consumption</td>
</tr>
</tbody>
</table>
3. A fluorescent lifetime imager shank

3.1. Introduction

Optical functional neural imaging has revolutionized neuroscience with optical reporters that enable single-cell-resolved monitoring of neuronal activity \textit{in-vivo}. State-of-the-art microscopy methods, however, are fundamentally limited in imaging depth by optical scattering in tissue even with the use of the most advanced multiphoton microscopy techniques (Fig. 3.1a) \cite{105,106}.

Instead, if a camera itself can be inserted into the brain, then imaging at arbitrary depth is possible. One method to enhance depth imaging from a multi-photon microscope is to implant a relay graded-index (GRIN) lens \cite{107}. While these lenses allow one to couple laser power into deep nuclei at the expense of the cerebral tissue above the interrogation volume, and enable functional Calcium imaging, their displaced tissue volume is very large (> 0.1 mm$^3$) compared to the imaging volume (> 0.01 mm$^3$). These methods achieve deep brain imaging at the cost of significant neural network trauma and still require free-space microscopes for read-out.

Figure 3.1 Size and imaging depth comparison of (a) multi-photon microscope, (b) implantable microscope with inserted GRIN lens, and (c) imaging probe resolving 4.1 mm-deep in brain tissue with an external excitation laser. Fluorescent emission is collected by 512 SPAD pixels located on two shanks.
In addition to solving the imaging depth problem, there is interest in scaling down the size of the epifluorescence microscopes themselves (Fig. 3.1b)\textsuperscript{108,109}. These efforts have primarily focused on small-form-factor implementations of full microscopes, including lenses and optical filters. This continued reliance on lenses and filters challenges miniaturization.

More recently, there has been significant efforts directed toward lens-less imaging. These approaches have been based on far-field masking (either phase or amplitude) to produce a spatially diverse illumination pattern on the imager which can be used to computationally reconstruct an image\textsuperscript{50}. The requirement for far-field positioning of the mask, however, precludes use of these approaches in shanks, where shank thicknesses want to be maintained to a minimum to prevent tissue displacement. The requirement for optical spectral filters in the scheme of fluorescence microscopy adds to the probe thickness.

In this work, we take an entirely different approach in which the “camera” itself can be inserted into the brain, allowing imaging at arbitrary depth by collecting fluorescence signal very close to the neuron of interest, without its signal having to scatter and attenuate along its insertion path.
depth. In particular, we present a monolithically integrated single-photon imager in the form of a CMOS, shank-based optical image sensor array that can be inserted into the brain (Fig. 3.1c) \(^{110}\). Single-photon avalanche diodes (SPADs) fabricated in CMOS are used as proximal light detectors. Lenses are replaced with near-field diffraction gratings placed on each SPAD, giving each of them a unique angularly modulated field of view (FoV) \(^{111}\). The mapping of SPAD counts into a volumetric scene allows the imager to determine the location of light sources. This additional information provides a multi-source localization capability similar to a far-field lens-less imaging approach. An external 480-nm pulsed pico-second laser delivers fluorescence excitation at the insertion point of the shanks by means of a fiber-coupled collimator. Spectral filters are replaced with time-gated operation of the SPADs in which detectors are turned on immediately after the pulsed excitation light has been shut off.

This chapter is structured as follows. In Section 3.2, we present the system design, and explain the post-processing into the shank form factor. The SPAD device and quenching circuit
that allows for time-gated, filter-less fluorescence photometry is described in Section 3.3. Section 3.4 describes the implementation of lens-less imaging through angle-sensitive detection, and Section 3.5 demonstrates the computational multi-source localization of phantom targets.

### 3.2. System overview

Fig. 3.2a shows the die photo of the implantable imaging probe with two shanks, each 4.1 mm-long and 120 μm-wide, connected to a 2-mm-by-1.2-mm head. Geometrically, this differs from our earlier design, extending the shanks by approximately 650 μm in length and reducing the shank separation from 250 μm to 175 μm to increase overlap of individual shank FoVs, aiding reconstruction performance. As a result, it is used to accommodate buffers and decoupling capacitors for delivering signal and power along the shank length. Each shank employs two rows of 128 SPADs operating in Geiger mode at 25.3 μm pitch and 7.7 μm active diameter, delivering a 6.3% fill factor. Each pixel contains a SPAD detector with a quenching circuit, 6-bit memory and addressing logic. Because the shanks have partially overlapping FoVs, a total imaging volume of 3.4 mm × 600 μm × 400 μm is achieved. The probe head contains a digital-to-time converter.
(DTC) for on-chip control of pixel gating and metal-oxide-metal (MOM) capacitors for decoupling the SPAD bias voltage. In subsequent discussions, we use the x direction to denote distance along the length of the shank, the z direction to denote vertical distance away from the shanks, and the y direction to denote distance along the width of the shank. Fabricated in a 130-nm high-voltage process, the system is subdivided into four voltage domains including a 1.5-V digital core domain, a 3.3-V digital input and output (IO) domain, a 5-V domain for the SPAD quenching circuits, and a >16-V domain for biasing the SPAD cathodes.

3.2.1. System Architecture

A system level block diagram is shown in Fig. 3.2b. An FPGA is used to synchronize time-gated photon counting with an external pulsed laser. A phase-locked loop (PLL) implemented within the FPGA uses a photodetector-produced laser pulse waveform to deliver a synchronization signal to the DTC, which then produces a global time-gate signal with programmable delay and pulse width. Photon counts are accumulated with this global shutter for a desired number of laser
pulses and read out by addressing each pixel’s 6-bit memory serially with a 31.25 MHz read-out clock.

This read-out clock frequency is ultimately limited by the rate at which data can be read from the pixel-level memory. A per-pixel sequential read-out rate of 31.25 MHz, combined with a minimum image acquisition time of 3.2 μs when used in synchronization with a 20 MHz laser repetition rate, leads to a maximum possible frame rate of 51 kilo-frames per second (kfps). The 6-bit dynamic range at this frame rate can be maximally utilized at high signal conditions, which corresponds to a 520-nm wavelength irradiance of $2.2 \times 10^6$ photons/s/um$^2$ at the pixel surface. The data is streamed out with a USB 3.0 link, and image reconstruction is performed on an external computer.
3.2.2. Post-Processing

In order to further miniaturize the imaging probe’s volume displacement, excess silicon surrounding the shank is removed to yield a 40 µm-thin structure ready for implantation. Fig. 3.3a shows the fabrication sequence starting with (i) photoresist (AZ P4620, MicroChemicals) patterning, followed by an O2-CF4 reactive ion etch through the (ii) passivation and (iii) CMOS stack. (iv) A Bosch process etches through 100 µm of silicon substrate, after which (v) photoresist is stripped, and (vi) the die is mechanically milled (X-PREP, Allied High Tech Products) to remove 250 µm of substrate silicon, fully releasing the shanks (Fig. 3.3b). The imaging probe is finally wire-bonded and packaged to a printed circuit board with a flat flexible cable connection and passivated with an epoxy (Fig. 3.3b inset).

3.3. Time-gated fluorescence imaging

Optical multi-dielectric spectral filters are an important part of any fluorescence microscope. They are required to reject the excitation intensity and allow only the fluorescence response to reach the detector. It is typically accomplished in the spectral domain by three
components, an emission filter, an excitation filter, and a dichroic mirror, combined usually in a “cube” to provide an optical density (OD) > 6 rejection of the excitation wavelength. Achieving the same levels of rejection with an integrated filter alone on our shank-based imagers remains difficult and is compounded by the dependency of filter properties on incident angle in the case of thin-film interference filters. Pigment-based absorption filters, as commonly used in color displays and CMOS imagers, do not have adequate rejection ratios\textsuperscript{112}. Interference filters are difficult to integrate on CMOS and have problems in the rejection of non-orthogonal incident light\textsuperscript{113,114}.

As an alternative to spectral filters, time-domain rejection is a powerful alternative by time-stamping the excitation source (pulse) and time-gating the detectors such that the fluorescence response is detected after the excitation has been turned off. This relies on the fact that fluorescent reporters have a time decay of fluorescence intensity characterized by their lifetimes. For commonly used fluorescent dyes, these lifetimes are in the 1-10 ns range (4.1 ns for EGFP, 4.0 ns

Figure 3.7 (a) Diagram of fluorescence measurement setup showing fluorescent and scattered excitation photons reaching the SPAD array located along the x axis. (b) Photo of 10 μM fluorescein in 0.6% agarose, excited by a collimated 488 nm pulsed laser. (c) Time-gated photon counts synchronized to the laser pulse at t=0 s for pure agarose (blue) and fluorescein in agarose (FL, green) plotted against the left axis; respective SBR (Fluorescein:Agarose) is plotted on the right axis (red).
for fluorescein, and 1.68 ns for Rhodamine B). The signal-to-background ratio (SBR) of a time-gated fluorescence image is determined by the ratio of the integrated photon count coming from the fluorescence emission to that remnant from the excitation source. The exact position of the time-gate relative to the laser turn-off time is chosen to maximize this SBR for the given laser turn-off and lifetime characteristics of the fluorescence.

3.3.1. Pixel Quenching and Reset Circuit

The in-pixel quench-and-reset circuit and pixel layout are shown in Figs. 3.4a,c. Each of the 512 SPADs is reverse biased between a global cathode and individual anode (AN). The cathode is held at a constant voltage above breakdown (VBD), while the anode (AN) moves between zero and the reset voltage (VRST) to put the SPAD into and out of Geiger mode. The rising edge of the ON signal provided by the DTC triggers edge detection. The large reset NMOS transistor (M1) begins discharging AN until it drops below the input threshold (\(~750\) mV) of the inverter-based comparator. When the inverter output goes high, the internal feedback clears the flip-flop-based edge detector, turning M1 off and setting AN to a high-impedance state. The SPAD is then biased in Geiger mode and ready for avalanche breakdown at an incoming photon. The self-timed reset,
a design change from the earlier version, optimizes the pulse width of the RST signal and internally minimizes the reset time for an arbitrary value of $V_{RST}$ between 1.5 V and 5 V without the need of an externally programmed pulse.

The long-channel half-latch transistor (M2) provides high-impedance at the AN terminal of the SPAD. During avalanche breakdown, a large current flow causes a build-up of voltage at the input of the comparator. The output of the comparator is flipped and triggers the event detection flip-flop, incrementing the counter by one. Only events within the ON signal are detected and at most one photon can be detected per cycle. When the counter reaches 63 counts, the Full signal is asserted, and the counter stops incrementing. The timing diagram in Fig. 3.4b demonstrates a cycle...
with zero counts and a cycle with one count. During the zero-count cycle, the internal delay of the self-timing reset mechanism and the falling edge of ON show the “effective” time-gate. This can be compensated with minor adjustments to the duty-cycle of the ON signal. During the one-count cycle, the avalanche is quenched and the counter is incremented by one.

3.3.2. Pixel Performance

The P-type (N-implant) shallow-junction SPAD was manufactured in a 130-nm process with a $V_{BD}$ of 15.5 V and its key performance metrics were measured in an integrating sphere\textsuperscript{110}. In Fig. 3.5a, the cumulative percentage of pixels under a given dark count rate (DCR) is plotted for increasing overvoltage ($V_{OV} = V_{CATH} - V_{BD}$), showing a “hot-pixel” percentage of 2%; hot pixels are defined as those with more than five times the median DCR. The median of the array-wise photon detection probability (PDP) is plotted as a function of wavelength in Fig. 3.5b. The SPADs show a peak sensitivity at 575 nm.

![Figure 3.10: PSF of a single point source located at coordinates $[2000 \, \mu m, 0 \, \mu m, 200 \, \mu m]$ portrayed as (a) xy-, (b) xz-, (c) yz-plane cross sections of a 3D pseudoinverse backprojection; voxels are 5 $\mu m$ across all dimensions; (d) PSF FWHM in $x$, $y$, $z$ with respect to with increasing distance of point source from imager; (e) xz-plane cross section of imager FoV. Dotted yellow lines in (a) and (e) mark the location of the imager.](image-url)
Figs. 3.5a,b show an increase in both DCR and PDP, respectively, with increasing $V_{OV}$, as measured with a monochromatic light source centered at 520 nm which matches the emission peak of EGFP. We use a figure-of-merit (FoM: Eq. 2 of 115) based on photon counting statistics for fluorescence imaging to evaluate SPAD performance. This FoM selects the highest probability of detecting photon events while avoiding false positives and noise events. In Fig. 3.5c, this FoM is plotted for fluences over a range of $10^3$-$10^5$ photons per second, equivalent to a photon flux of 0.4-40 fW/μm² at 520-nm wavelength. Due to the relatively low DCR compared to fluence, the optimal operating value of $V_{OV}$ is found to be 1 V. Fig. 3.5d demonstrates the linearity of photon counts in response to increasing fluence at 520 nm. The linearity is limited by the dark count at low fluence and by photon pile-up at high fluence.

$V_{OV}$ at 1 V also has the benefit of reduced afterpulsing events 116. Afterpulsing is estimated by measuring DCR while sweeping the active-reset SPAD dead time from 300 ns to 10 ns. The
median 40-Hz dark count holds constant across this range, indicating that afterpulsing is not a significant contributor to DCR.

Table 3 reports comparison with prior time-gated SPAD imagers. While much of the novelty of this work lies in the form factor and application of this imager, this design compares favorably with other time-gated SPAD imagers in conventional planar array formats \(^{114,117,118}\), while displacing a volume of only 0.03 mm\(^3\) with the implanted shanks.

### 3.3.3. Time-Gating Circuitry

The system employs a DTC to synthesize the global shutter ON signal with a tunable delay with respect to the laser pulse. Fig. 3.6 shows its architecture and timing diagram. TRIGGER determines the ON repetition rate. The Johnson counter creates an 8-bit sequence signal, CNT, where each bit corresponds to an 8×-down-sampled version of CLKIN, each 45 degrees delayed and fed into both phase interpolators (PIs). A user-defined 8-bit programmed phase is decoded to select any combination of three phases of CNT. These are equally summed into Σ, low-pass filtered and buffered \(^{119}\). Whereas CLK is generated in the first PI, REF is generated from a second PI to
flexibly compensate for internal delays, and its rising edge asserts \textit{READY} to allow for the next synthesized \textit{CLK} to propagate to the output \textit{ON}. \textit{READY} is pulled low when \textit{ON} is activated and asserted again on the rising edge of \textit{REF} when \textit{TRIGGER} is high. The dynamic range of pulse width and delay is extended through the final tunable counter stage. A serial peripheral interface (SPI) slave programs phase, width and delay from the control FPGA.

A variable output duty cycle of 20-50\% for \textit{ON} allows the SPAD to be moved out of Geiger mode after sufficient decay of the fluorescence signal, avoiding dark count and unnecessary power dissipation. As \textit{CLK} and \textit{REF} are 8×-down-sampled, 50\% duty cycle, phase-delayed versions of \textit{CLKin}, the \textit{CLKin} frequency determines the duty cycle (D) resolution of \textit{ON}. The \textit{CLKin} frequency must always be greater than or equal to \(8 \times \frac{F_{\text{Trigger}}}{D}\), where \(F_{\text{Trigger}}\) is equal to the laser repetition rate, in order to ensure correct phase and duty cycle for \textit{ON}, with higher frequencies allowing higher resolution phase control. All tests were performed at a 10 MHz \textit{TRIGGER} repetition rate. If one keeps the \textit{REF} edge constant, the DTC allows \textit{CLK} timing to be tuned relative to \textit{REF} with an 8-bit resolution in a range of the \textit{REF} and \textit{CLK} period. In the case of fluorescence decay lifetimes on the order of 2-4 nanoseconds, \textit{REF}-to-\textit{CLK} delays up to 12.5 ns are required, equivalent to 50\% of the available range (seven bits of the DTC). \textit{CLKin} is set to a 320 MHz frequency to synthesize a 40 MHz \textit{CLK}, allowing for a 25\% duty cycle for \textit{ON}. Fig. 3.6c shows how 128 LSB (seven bits) for \textit{CLK} results in 97-ps resolution. Fig. 3.6d shows the measured linearity performance of the converter, with DNL staying within [-1, +2.5] LSB and INL staying within [-4, +2] LSB, with standard deviations over the nonlinearity curve of 0.6 and 1.3 LSB, respectively. From the measurement data in Fig. 3.6c, we calculate the signal-to-noise-and-distortion (SINAD) of the DTC as \(^{120}\):

\[
SINAD = 20 \log_{10}\left(\frac{T_{fs}}{\sigma_e}\right) = 40.95 \text{ dB}
\]
where \( T_{fs} \) is the full-scale conversion time (12.5 ns), and \( \sigma_e = 0.112 \) ns represents the standard deviation of the output error. This leads to an effective number of bits (ENOB) of 6.5 bit, or an effective resolution of 137 ps.

### 3.3.4. Time-Gating Performance

To demonstrate the effectiveness of time-gated fluorescence imaging, the imaging probe was placed under media with properties similar to a fluorescence-expressing mouse brain (Fig. 3.7a). We used a 0.6% agarose gel (A7777, Teknova) to imitate optical scattering of neural tissue \(^{121-123}\), and a range of fluorescein concentrations were added to it to simulate bulk fluorescence expression (Fig. 3.7b). A fluorescein concentration of 10 μM is equivalent in brightness to cytoplasmic expression of GCaMP in a neuron \(^8^6\). A picosecond-pulsed laser (Fianium SC450-pp, NKT Photonics) delivers excitation light through a collimated beam 500 μM above the imager with 670 μW average power \((1.6 \times 10^6 \text{ photon/s})\), and the time-gate \( ON \) signal with a 10-ns on-time is moved in steps of 280 ps over the range of 12 ns after the laser pulse.

In the scattering medium, the excitation laser light reaches the pixels after traveling along a Mie scattering path, which we find representative of neural tissue without the presence of fluorescence. The resulting photon count (Fig. 3.7c, \textit{Agar}) is the convolution of the SPADs’ 10 ns time-gate with the 50 ps-wide laser pulse, which is denoted as the instrument response function (IRF) of the time-gated quenching circuit. Its sub-ns exponential decay characteristic is the result of minority carriers, created by the high excitation photon flux, diffusing into the multiplication region of the SPAD and creating an avalanche breakdown in the absence of an incoming photon \(^{124}\). The time-gated rejection ratio 2 ns after the pulse is 98%, corresponding to an excitation OD of 1.7.
In the case of a fluorescent medium representative of an EGFP-expressing mouse cortex (Fig. 3.7c, FL), direct laser intensity is not observed on the imager because most of the laser light is converted to fluorescence emission before it reaches the imager. Instead, we see a slow rise and 4.7 ns lifetime fall produced by a convolution of the excitation light time-gate and the 4.1 ns fluorescence lifetime of fluorescein.

Time-gated fluorescence imaging of fluorescent markers in a scattering environment requires a time-window in which the fluorescence response exceeds that of the scattered excitation light. In the case of 10 μM fluorescein in agarose, the SBR reaches 31.6 dB two nanoseconds after the laser pulse. SBR and excitation optical density (OD) continue to increase past 2 ns, reaching 105 dB and 3.2 OD, respectively, at 12 ns, however, this is achieved at the cost of reduced signal yield (5% at 12 ns). While time-gated filtering alone does not match the OD of epifluorescent microscopes, relatively low-OD spectral filters can be added on top of the imager to supplement the rejection from time gating.

When the imager is illuminated with the output-saturating photon flux of $2.2 \times 10^6$ photons/s/μm² and reads out at the maximum frame rate of 51 kfps, operating at maximum power consumption, it consumes a total of 6.24 mW. The charging of the SPAD diode capacitance at quenching events consumes 2.94 mW on the $V_{RST}$ and SPAD Cathode nodes together. The 1.5 V digital core consumes 3.3 mW. The 3.3 V IO power was supplied through the FPGA. With the majority of the digital logic located on the base of the shank, we estimate a power consumption of ~3 mW for the implantable regions of the imager, comparable to levels reported by other implantable CMOS neural recording shanks $^{23}$. 
3.3.5. Photometry In Scattering Neural Tissue

To test the imaging performance on a fluorescent target with known size and properties, a Monte Carlo photon trajectory simulation $^{125,126}$ is performed in a tissue model characteristic of EGFP expression in gray matter. The tissue is modeled with anisotropy factor $g=0.88$, refractive index $n=1.37$, scattering coefficient $\mu_s=21 \text{ mm}^{-1}$, and absorption coefficient $\mu_a=0.06 \text{ mm}^{-1}$ $^{127}$, while EGFP is modeled with a quantum yield of 0.6, extinction coefficient of $55000 \text{ M}^{-1}\text{cm}^{-1}$ $^{128}$, and lifetime of 4.1 ns. Using a simulation setup similar to Fig. 3.7a, we model an imager implanted along the x axis with pixels located between $x=0$ and 3.2 mm, in parallel with a collimated pulsed laser light source illuminating the tissue boundary of $x=0 \mu\text{m}$, at a distance $z=100 \mu\text{m}$ away from the imager with 700 $\mu\text{W}$ average power. A group of fluorescently labeled somata contained in a 50 $\mu\text{m}$-radius located at height $z=100 \mu\text{m}$, expressing EGFP with an equivalent concentration of 10 $\mu\text{M}$ $^{86}$, is swept along the x axis. The model assumes a duty-cycle of 50% and a 12-ns time-gate delay after the excitation pulse, which is the shortest delay at which SBR exceeds one at all depths. At this delay the filter achieves an OD of 3.2 and the fluorescence emission falls to 5% of its peak intensity.

We perform these simulations at a frame rate of 1 fps. For a laser repetition rate of 20 MHz, this means that photon counts for $2 \times 10^7$ laser pulses are integrated. Fig. 3.8a shows resulting counts for the time-gated excitation source, the fluorescent emission, and the dark count. The fluorescent emission assumes that the group of somata is positioned directly above the sensing SPAD at each depth. Fluorescence emission is found to overcome scattered excitation at all simulated depths, although by a small margin. Fig. 3.8b plots the signal-to-noise-and-background-ratio (SNBR), defined as the ratio of fluorescent signal counts to the background, photon-shot-noise and dark-count-shot-noise. For time gate delays longer than 12 ns, the drop in signal yield
for a given frame rate leads to an increase in shot noise relative to signal. A frame rate of 1 fps is necessary to reduce the shot noise to yield SNBR larger than 3 dB. SNBR is largest at tissue depths below 500 µm because scattering has not yet become significant. When the imager is deeper in tissue, excitation collimation is lost and more blue excitation light scatters into the detector. However, due to the limited acceptance angle of 60° (see Section 3.4.1), the majority is still rejected, and the excitation background and fluorescent signal reduce proportionally between 500 µm and 2.5 mm.

Beyond 2.5 mm, exponential extinction of the excitation light places the detector into dark-count-limited operation. In the absence of better excitation rejection, the SNBR of our imager is sufficient up to a depth limit, and deep brain imaging becomes a problem of light delivery. Potential solutions have been demonstrated including implantable waveguides 129–131, densely integrated GaN µLEDs 132,133, or multimode fiber endoscopes 134.

Another solution is to incorporate spectral filters that work together with time gating. In Fig. 3.8b, we repeat the same analysis with the addition of a filter which is able to block the excitation wavelength at an OD of only 0.6 135, but which acts to further reduce the background and photon shot noise associated with this background. The SNBR in this case is sufficient to allow imaging at full shank depths. Imaging at the hardware limit of 51 kfps would be possible given high enough illumination. The frame rate is in direct trade-off with shot-noise-limited SNBR.

3.4. Lens-less imaging using near-field angle-sensitive detection

The lack of refractive focusing optics poses two challenges in image formation. The first is low photon yield, which is addressed in part through the use of sensitive SPAD detectors. The second is the lack of spatial resolution. Simply relying on the limited numerical aperture of the
detectors in the array itself (which is an acceptance cone of approximately 60° single-sided) would allow an x- or y- spatial resolution of only ~450 μm to be achieved 200 μm from the detector. Instead, we rely on near-field diffraction gratings on each pixel to provide an angle sensitivity diversity that can be used for computational image reconstruction. Use of these gratings comes at a cost in transmission efficiency, which averages only 3% of incident photons. This can be improved to 29% using non-blocking dielectric phase gratings to achieve the same diffraction gratings instead of metal wires. The inverse imaging problem in this case is formulated as a linear system from the scene $x$ to the array photon count $y$:

$$y = Ax + \epsilon \quad (2)$$

The sensing matrix $A$ compresses the scene into a much lower dimensionality determined by the pixel count. Dark count and background $\epsilon$ are assumed to be spatially and temporally uncorrelated to the sources, and afterpulsing was estimated to be negligible. Image quality is improved by constructing a sensing matrix $A$ that maximizes information extraction from the scene $x$. Such an optimized $A$ allows for the best reconstruction of the inverted image $\hat{x}$. In particular, we seek an $A$ that has maximally incoherent columns, indicating that each location in space is compressed onto the imager’s response in a maximally distinct way.

### 3.4.1. Angle-Sensitive Pixel Design

$A$ is determined by a diversity of angle-sensitivity introduced into each pixel with near-field diffraction gratings as shown in Fig. 3.9. These gratings, formed in two layers of back-end metal separated by 1.14 μm (Fig. 3.9a), constitute sixteen angular-sensitivity variations (which take up a total distance along the length of the shank of 200 μm) consisting of orthogonal combinations of two angular modulation frequencies, two rotations, and four quadrature phases.
The gratings are employed across both shanks in a repeating pattern of two by eight pixels to ensure that a source located 100 µm above the center axis of the shank and positioned (in x) at the midpoint of the 16-pixel group, is situated in the 60° FoV of each pixel variation.

The angular response for each grating combination was measured by illuminating the imager with a collimated monochromatic light source at 520 nm and varying the angles of incidence along θ and φ, two azimuthal angles pivoting around the x and y axes respectively, using a rotational and tilting stage under the imager (Fig. 3.9b). Two examples of grating structures, each with a low and high angular frequency, are shown in Fig. 3.9c, along with the parametric fit.

The higher angular modulation amplitudes at extreme angles in Fig. 3.9b reflect the fact that the Talbot self-image is not generated at the exact height of the secondary grating. At even higher incident angles, the CMOS metal interconnect starts occluding photons for incident angles larger than 45°, and rejects completely at 60°. Additionally, the PDP at green wavelengths is lower for obliquely incident photons due to a deeper effective multiplication layer, shifting peak PDP away to a higher wavelength.
3.4.2. Point Spread Function

The point spread function (PSF) describes the system response to a point source, and its width determines how close two point sources can be placed while still being individually reconstructed. The PSF of this imaging system can be computed as a pseudoinverse backprojection of a single voxel:

\[ x_{PSF} = (A^+A)x_{source} \]  

(3)

where \( A^+ \) is the left pseudoinverse of \( A \) and \( x_{source} \) is a \( n_{voxel} \)-by-1 array which is zero except at a single voxel. \( n_{voxel} \) is \( \sim \)2 million in the case of using a voxel grid of 5 \( \mu \)m. The PSF of a single voxel located at Cartesian coordinates of \([2000 \, \mu \text{m}, 0 \, \mu \text{m}, 200 \, \mu \text{m}]\) is plotted in Figs. 3.9a-c, showing the correlated nature of voxels in this highly underdetermined system. Fig. 3.10d displays the resolution in the form of full-width at half-maximum (FWHM) of the one-dimensional PSF.
profile. For continuity of the plot, data points in which the local minima closest to the peak did not fall below the half-maximum were omitted.

The broadening of the FWHM with increasing source distance results from an approximately constant angular resolution translating to a larger spatial spread at further distances. The minimum resolution is measured to be 64 µm, 26 µm, and 65 µm in the x, y, z directions respectively. In comparison to a GRIN lens based miniature microscope \(^{107}\), while the imager resolution is 1-2 orders of magnitude worse, it images a brain volume larger by the same factor. The imaging volume can be sculpted further by restricting the illumination profile to a volume of interest, drastically reducing the image background.

3.5. Imaging performance

3.5.1. Single Point Source Imaging

To mimic a cluster of neuron somata, a diffuser tip of 100 µm radius (Thorlabs CFDSB20), coupled to a monochromatic green LED (300 µW, \(7.9\times10^{14}\) photons/s), was placed 200 µm above the imager and translated along its length at three different positions (Fig. 3.11, \(p1, p2, p3\)). The image was acquired with a 0.8 fps frame rate and photon accumulation time of 200 ms, yielding a maximum of 41k counts at the pixel directly under the light source. The repeating pattern of sixteen gratings is responsible for the periodicity of raw counts in Fig. 3.11a and contribute to localization of the point source. Pseudoinverse backprojections (\(\chi_{proj} = A^+ y\)) sectioned at \(z=200\) µm show the imager’s ability to find centers of brightness within a 30 µm standard deviation (Fig. 3.11b).

3.5.2. Multisource Phantom Imaging

A multisource phantom fluorescent target was constructed by depositing fluorescent microspheres (F8836 10 µm, Thermofisher) on a cover glass, which was then placed 400 µm away
from the imager. A picosecond-pulsed excitation (Fianium SC450-pp, NKT Photonics) centered at 480 nm and carrying 700 µW of average power (1.7×10^{15} photons·s^{-1}) with a 20-MHz repetition rate was delivered in a direction parallel to the shank with a 500 µm diameter collimator. This parallel incident angle for the excitation power minimizes the image background produced by incomplete time-gate rejection. Figs. 3.11a,b show two distinct scenes with a difference in the size and arrangement of the imaging target. The fluorescence microscope image is displayed alongside the reconstructed pseudoinverse backprojections of SPAD counts. Photon counts were accumulated for 200 ms at 1 fps with a time-gate delay of 550 ps after the laser pulse, resulting in a maximum SBR of 15 dB. Voxel sizes are 20 µm in x, y, and z dimensions to optimize for computation speed.

To show the efficacy of time-gated fluorescence imaging, a third scene (Fig. 3.12c) is constructed with two microsphere clusters, one composed of fluorescent spheres and another with scattering latex spheres. Despite the visibility of both clusters under brightfield imaging, the imager selectively detects only the fluorescent microspheres.

### 3.5.3. Source Localization by L1-norm Minimization

Due to the highly compressed nature of the imaging system which maps ~10^5 voxels to 512 pixels when constructing a three-dimensional (3D) image with 20 µm voxels, a direct backprojection image will inevitably have limited spatial resolution and appear blurred. Furthermore, high frequency noise is amplified into the image through high spatial sampling frequencies, associated with low magnitude singular values of the sensing matrix A. By incorporating a sparsity constraint, the brightest few source locations can be solved for in a least-mean-squares manner. Performing this basis pursuit denoising optimization allows an estimation
of best-fit source locations while penalizing the number of total sources with a weighting parameter $\lambda$. This optimization problem employs a cost function that includes the L1-norm of sources ($f_{L1}$) and the L2-norm of the residual error ($f_{L2}$):

$$\hat{x} = \arg\min_x (\lambda\|x\|_1 + \frac{1}{2}\|Ax - y\|_2^2) = \arg\min_x (f_{L1} + f_{L2})$$

(4)

To solve this optimization problem, we use the in-crowd algorithm to solve for sparse locations within the imager’s FoV in real time. Fig. 3.13a shows the 3D reconstructed image of the microsphere arrangement in Fig. 3.12b solved with 20 $\mu$m voxel sizes. Raw imager data is drawn as a heat map at the plane of the imager ($z=0 \mu$m). The largest magnitude contributors to the solution ($\hat{x}$) are displayed in circles, while circle diameters indicate the magnitude estimate. The two voxels with largest magnitude, colored red, show the imager’s ability to locate the two microsphere clusters at their correct volumetric locations. The next highest 40, colored black, are also found to gravitate around the two clusters with decreasing amplitude estimates. As more voxels are added to the solution $\hat{x}$, the L2-norm of the residual error ($f_{L2}$) declines as shown in Fig. 3.13b, but each with a diminishing contribution to residual error minimization. By combining this information from residual error with the appearance of clusters of estimated sources, we can heuristically infer sparsity and location without relying on an external verification of ground truth.

3.5.4. Limit of Reconstruction Accuracy

We perform a Monte Carlo simulation to determine the maximum number of sources the imaging probe can resolve simultaneously. Sparse reconstruction theory states that the maximum number of detectable sources is the number of pixels $M$. However, for L1-minimization solvers that assume no constraints on the sensing matrix $A$, phase transition occurs around $\sqrt{M}$, meaning that a sharp decline in reconstruction reliability occurs above that density. In the case of the
imaging probe, the columns of $\mathbf{A}$ have a repetitive structure determined by the angular sensitive gratings. As a result, we expect the maximum allowable sparsity to be below $\sqrt{M}$.

We place $K$ sources randomly in the imaging volume of $3.2 \text{ mm} \times 200 \mu\text{m} \times 150 \mu\text{m}$ above a single shank at $250 \mu\text{m}$ depth. Photon shot noise is applied to the simulated fluorescent data, and additional excitation background is added with the expected SBR of $3.5 \text{ dB}$. Source placement and subsequent localization is repeated $N=1000$ for each value of $K$. To correctly compare the localization result against the ground truth, an exhaustive search method links found sources with their closest true locations, where computational tractability limits the analysis to $K=10$. We analyze the localization error in terms of mean square error (MSE), allowing a decomposition into the average error squared and variance of the error:

$$
MSE(x, \hat{x}) = \frac{1}{N} \| x - \hat{x} \|_2^2 = \left( \frac{1}{N} \sum (x - \hat{x}) \right)^2 + \sigma_{x-\hat{x}}^2
$$

Fig. 3.14 shows how localization performance is affected by the source cardinality, $K$. Fig. 3.14a shows the distribution of the error over all iterations and source cardinalities in each dimension. Large outliers are essentially ground truth locations that are not found, and the larger imaging volume in the $x$ direction permits larger outliers along that axis. The error is distributed with 98% of sources found within $130 \mu\text{m}$, or around two times the largest FWHM at $100 \mu\text{m}$ distance, of the true location. The compressive sensing method has a 2% likelihood for introducing false positives and negatives, which need to be accounted for in subsequent imaging target analyses.

The total ensemble standard deviation of all errors in all directions reduces to $17 \mu\text{m}$ (Fig. 3.14b) when corrected for these outliers. The error increases with more sources due to increased noise sensitivity at higher source densities. Both the standard deviation and bias of the
error (Fig. 3.14c) are smaller than the pixel pitch in all directions x, y, and z. The sign of the data point indicates an estimation bias in the respective direction.

Even at the reported large variance, a low error bias indicates the imager’s strength at localizing volumetric centers of brightness with micrometer accuracy, where precision improves with repeated measurements and averaging over multiple reconstructions. We, therefore, expect that a single shank with 256 pixels can localize up to ten point sources simultaneously with 98% certainty at SBR levels comparable with those estimated for an in-vivo environment.

3.6. Discussion

In this work, we present an implantable neural shank-based imager with minimal tissue displacement of only 0.03 mm³, while monitoring a volume of 0.4 mm³. The minimally invasive imager is able to localize fluorescent objects at depths beyond those of conventional imaging systems. It is capable of distinguishing multiple fluorescent objects with a resolution of 64 µm, 26 µm, and 65 µm in the x, y, and z directions, respectively, without the use of spectral filters and refractive focusing optics. Time-gating adds two orders of magnitude of excitation light rejection, adding to what can be provided by thin-film absorptive optical filters. The imager’s modest pixel count of 512 trades off spatial resolution with fast image acquisition; its high maximum frame rate (51 kfps), synchronized with an external pulsed excitation source, allows the investigation of sparsely expressed biomarkers for in vivo neural imaging.
4. Integration of optical color filters and compressive sensing in an ultrathin frontend

4.1. Introduction

Optical imaging has revolutionized neuroscience by allowing recording of neural function in vivo. However, light scattering and absorption in tissue fundamentally limits the depth at which fluorescence microscopes can detect labelled neurons. To overcome this depth limitation, techniques to date have focused on microendoscopy, in which an optical fiber is implanted in targeted brain regions, sometimes with a miniature lens or prism to allow imaging at the fiber distal end. These approaches, however, result in a significantly restricted field of view for a significant amount of tissue damage. At the same time, several realizations of head-mounted microscopes for 1p and 2p calcium imaging in mice have proven the feasibility of fluorescence microscopy with compact form factors. To achieve imaging at depth, these instruments also require implantation of their requisite GRIN lenses (typically, with 0.3 - 2 mm diameter); this results in even greater displacement of brain tissue while still delivering only restricted fields of view. The limitations with existing state-of-the-art imaging approaches at depth has led to the search for alternatives in which the imager itself is given a shank form factor and directly inserted into the tissue.

In this case, minimizing the thickness of the device is of prime importance to reduce displaced tissue volume. The lenses and filters associated with traditional microscopes are incompatible with this thin form factor. Instead, interest has turned to plenoptic imaging, in which light conditioning is introduced before the photodetector to modulate its response based on incident angle, wavelength, arrival time, or polarization. Lens-less imaging is one form of light-field microscopy (LFM), which has recently benefited substantially from advances in
plenoptic imaging capabilities\textsuperscript{148}. When a diversity of pixels with maximally orthogonally modulating responses are combined computationally and the number of sources is constrained to be below the number of detectors\textsuperscript{79,149}, compressive sensing algorithms allow for unambiguous reconstruction of the sources. Targeted labeling of functional neural reporters can provide the spatial-temporal source sparsity required to guarantee reconstruction. In general, the challenge for implantable imagers is to maximize signal conditioning to allow for application-specific image reconstruction, while aggressively minimizing the system footprint. While implantable lens-less methods lack the light gathering capabilities of lenses, they can image sources close to the device to compensate for diminished photon yield, not limited by the working distances associated with refractive optics.

One approach to modulate light to enable lens-less, volumetric, light-field imaging is with coded apertures\textsuperscript{50,58}, either in phase or amplitude. These approaches allow for high-resolution imaging but still rely on placing these masks many wavelengths away from the photodetector. Pixel-to-pixel orthogonality is determined by the mask feature size and separation distance between the mask and the detector. The mask minimum feature size is limited by diffraction and thus imaging resolution will decrease when a thinner device is required.

A near-field diffraction approach based on the Talbot effect places diffraction gratings in front of the detectors\textsuperscript{84,150}. Angle sensitivity is achieved with a pair of gratings: a diffractive grating and an analyzer grating. The top grating diffracts incident light into a Talbot diffraction pattern\textsuperscript{111}, while the bottom analyzer grating selectively transmits or rejects the formed peak intensities, giving rise to angular modulation. In this case, distances between mask and detector are on the order of the wavelength, allowing lens-less imaging to be achieved with much thinner devices, opening up applications for flexible\textsuperscript{151}, stackable\textsuperscript{152} or implantable imagers\textsuperscript{110}. Amplitude-
modulating gratings, which can be fabricated directly in the back-end metal of a conventional semiconductor fabrication process, produce this angular sensitivity but with significant optical loss. Phase gratings do not produce this signal loss but generally require an extra post-processing step\textsuperscript{84}. These near-field diffraction approaches have an angular sensitivity that operates at a given wavelength with other wavelengths operating at degraded resolution\textsuperscript{83}. Approaches to preserve resolution with shifted wavelengths such as odd-symmetry phase gratings\textsuperscript{153} show reduced angle-sensitive modulation generally.

In this work, we make these near-field conditioning masks angle-sensitive and multispectral, allowing multiple colors to be detected with a single photodetector within the context of angle-sensitive lens-less imaging. In this case, simultaneous spectral and spatial filter can be achieved. This is done by replacing the analyzer grating in the traditional angle-sensitive design with a dual-bandpass filter based on a metal-insulator-metal (MIM) Fabry-Perot structure\textsuperscript{61,135,154,155}. We apply these masks to a complementary metal-oxide-semiconductor (CMOS) imager array, based on time-gated single-photon avalanche photodiode (SPAD) detectors\textsuperscript{110}. The 1024-pixel imager consists of four 130 µm wide shanks (with 256 SPADs per shank), thinned to 50 µm thickness and packaged for neural implantation. We demonstrate how the combination of angular and spectral sensing allows for simultaneous source separation and localization of multispectral fluorescent sources. We demonstrate the localization of 45 µm-diameter green fluorescent beads embedded in scattering medium. We also demonstrate lifetime imaging capabilities which further enhance source separability and provide the ability to estimate environmental pH using fluorescein lifetime. Imagers based on these techniques will eventually enable many interesting modalities for \textit{in-vivo} neural imaging.
4.2. Results

4.2.1. Diffraction grating constructed from metal-insulator-metal resonance filters

Key to this design is the MIM analyzer grating. As opposed to plasmonic hole and nanodisk array filters\textsuperscript{60,113,114,156}, these MIM structures do not depend on spatial periodicity and thus can be laterally patterned on the wavelength scale. We begin with two MIM filters, green and red bandpass filters designed with HfO\textsubscript{2} dielectric thicknesses of $d_1 = 70$ nm and $d_2 = 100$ nm, respectively, sandwiched between two 30 nm thick silver layers (Fig. 4.1a). We define the depth difference, $\Delta d = d_2 - d_1$. The optical cavity selectively transmits light with a wavelength twice the optical path length, which is determined by the product of the insulator refractive index and thickness. In this case, we create a dual bandpass filter that also functions as a diffraction grating by turning the top mirror of the constituent filter design into a grating (chosen to have a pitch, $p$, of 1 \textmu m pitch) and interleaving these gratings for the red and green filters.

Fig. 4.1b shows the measured spectral transmission efficiency for each structure in Fig. 4.1a under normal incidence (with the incident angle $\theta = 0^\circ$). As expected, the interleaved grating transmits a superposition of the individual planar MIM bandpass filters. Finite-difference time-domain (FDTD) simulations (ANSYS Ltd, Waterloo, Canada) show similar performance (Fig. 4.1b) for these structures with peak transmissivity in excess of 25% at thicknesses of 150 nm. More importantly, the red filter rejects green light by a factor of 10, and vice versa, minimizing cross-talk to allow for robust separation of these color channels. Fabrication nonidealities lead to lower rejection in the blue regime experimentally for the interleaved filter. This can be compensated, in part, by the introduction of a 3.5 \textmu m thick layer of SU-8 epoxy containing an absorptive dye which provides a 500 nm cut-on longpass filter characteristic. As shown in Fig. 4.1c, wavelengths under 475 nm are attenuated up to four orders of magnitude. The MIM filter additionally blocks
auto-fluorescence generated by the absorptive dye. This SU-8 layer also acts as the spacer between the diffractive and analyzer gratings in the full angle-sensitive implementation.

The full width at half maximum (FWHM) determines how closely spaced two resonant colors can be designed while maintaining two well-defined bandpass characteristics. The filter Q-factor (defined as the ratio of peak-transmission wavelength, \( \lambda_0 \), to the FWHM) is determined by metal reflectivity (\( R \)) as:

\[
Q = \frac{\lambda_0}{FWHM} = \frac{\pi \sqrt{R}}{1 - R} \tag{4.1}
\]

For the green and red bandpass filters of Fig. 4.1b, Q-factors of 8.7 and 10.4, respectively, are achieved, consistent with the high reflectivity (\( R \approx 0.7 \)) of the silver films. These Q-factors match the emission spectrum of popular fluorescent proteins, such as EGFP, Alexa 488, and FITC. HfO\(_2\) (\( n_{MIM} = 2.1 \)) was chosen over SiO\(_2\) as the insulator material to decrease the sensitivity of the transmission spectrum to \( \theta \) and \( \rho \) variations.

The near-field diffraction produced by these interleaved filters can be directly measured through the use of near-field scanning optical microscopy (NSOM)\(^{157} \). In these measurements, local electric fields are transmitted through a transparent-tip cantilever and coupled into a fiber through a 100\( \times \) objective. The light intensity is recorded using a photomultiplier tube. NSOM measurements at 532 nm and 660 nm produce non-overlapping Talbot diffraction effects in a 4 by 4 \( \mu \text{m} \) area above the gratings (Fig. 4.1d).

### 4.2.2. Angle-spectral-sensitive pixel front end on CMOS

To create these MIM angle-sensitive (MIMAS) conditioning masks, the MIM filter grating structure is used as the analyzer grating, while a phase, rather than amplitude, diffraction grating is used, doubling transmission efficiency and minimizing internal reflections to produce sharper
angular modulation. We fabricate such grating structures on top of a CMOS SPAD photodetector\textsuperscript{110} as shown in Fig. 4.2a.
To fabricate this, we sputter through a shadow mask a 30 nm thick Ag layer and a 100 nm thick HfO$_2$ insulator. Subsequently, we etch the HfO$_2$ after defining the pixel-specific mask in PMMA through electron-beam lithography. Then, we sputter deposit 25 nm thick Ag capped by 30 nm of HfO$_2$ to prevent oxidation of the silver. The LP500 filter dye (Adam Gates Company, Hillsborough Township, NJ, USA) is dissolved in 3.5 µm thick SU-8, separating the MIM grating from the phase Talbot grating directly patterned in 220 nm thick mr-EBL negative-tone electron beam resist$^{158}$. This resist provides the high refractive index ($n_{PG} \geq 1.6$) required for a suitable refractive index contrast with the source medium ($n_{PG} - n_s$), either air or aqueous media.

Fig. 4.2b shows the scanning-electron microscope (SEM) image of a postprocessed SPAD taken after focused-ion-beam-milling cross-sectioning (FIB-SEM). The SPAD, MIM grating, SU-8 separation layer, and phase grating are annotated. Despite the thickness of the silver, its high reflectivity makes the MIM grating structure visible. The refractive indices of the surrounding medium, phase grating, separation layer, and MIM dielectric are given by $n_s$, $n_{PG}$, $n_Z$, and $n_{MIM}$, respectively. The epoxies SU-8 and mr-EBL have been studied for cytotoxicity, with results varying from fully biocompatible to slightly cytotoxic$^{159,160}$. While the SU-8 layer can be encapsulated, additional encapsulation is difficult because of the need to have abrupt index features in the mr-EBL layer.
In conventional lens-based imaging, each pixel is mapped to a unique point on the object focal plane. In contrast, a lens-less imager that uses angle sensitive pixels compressively samples the scene using co-sinusoidal basis functions (Fig. 4.2c), similar to a discrete cosine transform\textsuperscript{161}. The function describing this pixel-level modulated sampling is defined as $T(\theta, \lambda)$ where $\theta$ denotes the elevation angle pivoting around the $x$ axis. We can parameterize $T(\theta, \lambda)$ with angular frequency $\beta$, peak offset from zero $\alpha$, modulation index $m$, and grating rotation $\chi$ as:

\[ T(\theta, \lambda) = \sin[\beta \theta + \alpha] \cdot \cos[m \lambda + \chi] \]

Figure 4.2: MIMAS filter integrated on CMOS. (a) Fabrication of MIMAS on CMOS (b) FIB-SEM sideview showing front end fabrication, with annotations for SPAD, electrical contacts and CMOS metal stack. (c) The dual grating structure selectively transmits light according to incident angle, illustrating the point spread function from a single detector for either wavelength. (d) Two-dimensional angular response for a single device for both resonant wavelengths. A plot at $\phi=0^\circ$ compares measurement (solid) to simulation (dash). Resulting parameters of interest are the angular period $360^\circ \beta^{-1}$, peak transmissivity offset from zero $\alpha$, and modulation index $m$. 

In conventional lens-based imaging, each pixel is mapped to a unique point on the object focal plane. In contrast, a lens-less imager that uses angle sensitive pixels compressively samples the scene using co-sinusoidal basis functions (Fig. 4.2c), similar to a discrete cosine transform\textsuperscript{161}. The function describing this pixel-level modulated sampling is defined as $T(\theta, \lambda)$ where $\theta$ denotes the elevation angle pivoting around the $x$ axis. We can parameterize $T(\theta, \lambda)$ with angular frequency $\beta$, peak offset from zero $\alpha$, modulation index $m$, and grating rotation $\chi$ as:
\[ T(\theta, \lambda) = \frac{1}{2} + \frac{m}{2} \cos[\beta(\lambda) \cdot \theta(\chi) + \alpha(\lambda)] \] (4.2)

after we convert the volume of interest, relative to the pixel location, from Cartesian to spherical coordinates.

The pixel response is fully characterized by a far-field angular measurement. The device is mounted on two rotational stages to measure 2D angular modulation. Fig. 4.2d shows measured angular (\(\theta\) and \(\phi\)) dependency and detailed simulated and measured \(\theta\)-angle-dependent transmission efficiency (\(\chi=0\)) at \(\phi=0^\circ\). The measurements are normalized against a SPAD imager without fabricated gratings, showing a 32% peak transmission efficiency, compared to the 35% predicted in simulation. A 25% average transmission over all angles is observed. In these measurements, we observe the generation of orthogonal angular-sensitivity functions at each color. The green (530 nm) and red (660 nm) responses are phase shifted with respect to each other by \(\alpha=180^\circ\), a direct result of the interleaved color filter gratings. The experimentally found values for \(\beta\) (11 and 12.5 for green and red, respectively) compare well to simulation. Incident light is correctly modulated up to \(\pm 45^\circ\). At higher incident angles, transmissivity is decreased due to occlusion from the CMOS metal stack. In aqueous environments, the angular frequency \(\beta\) increases\(^{84}\) by a factor \(n_{s}n_{air}^{-1}=1.33\). The modulation index, \(m\), determines the sensitivity of the response to changes in angle of incidence. We achieve \(m\approx0.65\), close to those achieved in simulation (\(m\approx0.8\)). These results show we can achieve a well-defined, pixel-level angular-spectral modulation of incident light with these masks. The pixel-to-pixel standard deviation in \(m\) for a given angular-sensitive geometry is typically less than 0.06. As the angle of maximum intensity recorded from each pixel does not see significant pixel-to-pixel variability and this has a more dominant influence on reconstruction quality, we do not see significant effects on image
reconstruction associated with using the average transmission for each incident angle for parameterization.

### 4.2.3. Array-wise response of imager array

We fabricate these MIMAS frontends on an implantable CMOS imager, designed in a TSMC 130 nm high-voltage technology, which consists of four shanks (Fig. 4.3a). Each shank contains $2 \times 128$ SPAD pixels, with 24.5 µm and 92 µm pixel pitch in the x and y directions, respectively. The pixel pitch is chosen to support SPAD quenching and detection circuit, five-bit in-pixel memory, and global shuttering and readout. The SPAD geometries are chamfered squares with 14 µm vertices for a fill factor of 8%. The 700 µm long tip ensures a sharp insertion profile...
but does not carry detectors. The shank imager is wirebonded onto a circuit board, ready for neural implantation (Fig. 4.3b).

Each front end’s angular frequency, phase, and rotation correspond to a unique point in Fourier space. For the imager to sample the scene in the complete Fourier space, a diversity of pixels is required which orthogonally sample angular frequency ($\beta$), phase ($\alpha$) and rotation ($\chi$). While high angular frequencies $\beta$, implemented by small grating pitch, provide high spatial resolution, low angular frequencies are needed to complete the sampling spectrum$^{162}$ and result

Figure 4.4: Multi-wavelength lens-less imaging using angle sensitive pixels. (a) Photomicrograph of the fluorescent pinhole array overlaid on the chip. 30 µMol l-1 ATTO 465 and ATTO 490LS fluorescent dye is dispersed in 1.5% agar simulating scattering tissue. The agar is overlaid on a pinhole array and illuminated from the back. (b) Recorded data from MIMAS imager chip. (c) Calibration of recorded lifetimes for 30 µMol l-1 green, red and undyed agar. Time series for the mean recorded data, separated in green and red color channels through blind source separation. (d) Data separated into color channels after blind source separation. (e) Inverse image for both wavelength channels with detail plots representing the summation in the X and Y dimensions.
from larger grating pitches. Simulated and fabricated values abide to a previously derived analytical model for $\beta^{84}$. The implementation of grating rotations $\chi$ allows one to resolve rotated features ($\theta(\chi)$). Changing the grating offset $\delta$ allows the sampling of angular frequency with different phase ($\alpha$). We design a repeating ensemble of maximally orthogonal front-end geometries, repeating every sixteen pixels along the array. This repetition ensures that a source located 150 $\mu$m above the shank is in the field-of-view (FoV) of each geometry.

Fig. 4.3c shows the tip of the shank with the pixel array starting 600 $\mu$m from the tip, to ensure smooth insertion. The 16 different grating structures are created from two pitch sizes, two grating offsets, and eight grating rotations as structural design parameters. The 2D angular response for each pixel front end is shown next to the pixel for both red and green wavelength. A large diversity of rotations is found to be more beneficial in achieving orthogonality than pitch and offset variation.

In this form factor, each of four shanks displaces a tissue volume of 120×4000×50 $\mu$m while yielding an imaging FoV of 200×4000×150 $\mu$m. In this shank geometry, the FoV and volume scale equally with an increasing number of pixels. Gradient refractive index (GRIN) lenses$^{107,163}$ and implantable multimode fibers$^{164}$ allow imaging at depth with external optics, imagers, and light sources. However, they image only from the end face, leading to a displaced volume that scales only linearly with both FoV and desired depth. Furthermore, because using these lenses still requires bulky external objectives and image sensor systems, fully implantable versions of these systems remain elusive. For the current shanks, the FoV-to-volume-ratio is approximately five, two orders of magnitude better than what is achievable with implantable GRIN lenses. This will only improve with further advances in pixel sensitivity and with shank width and thickness optimizations.
4.2.4. Multispectral sparse reconstruction formulation

Before imaging, we establish the computational framework for image reconstruction, making use of compressive sensing techniques. In this multi-spectral, lens-less imaging, the imaging volume is compressively sampled through a spatial sensing matrix $A$. Each element $A_{ij}$ represents the response from voxel $j$ to pixel $i$. As such, the rows $A_i$ correspond to the co-sinusoidal modulation of each pixel and are constructed from the angular modulation function $T_i(\theta, \phi, \lambda)$. The response is converted from spherical to Cartesian coordinates and translated to the pixel’s position on the imager to allow inverse imaging in an arbitrary volume voxel grid. The sensing matrix $A$ compresses the image $x$ onto the recorded data $y$ for each color ($A_g, A_r$). By solving for the image, the solution explicitly contains a source color classification ($x_g, x_r$):

$$Y(t) = Ax(t) + \epsilon = [A_g \ A_r][x_g(t), x_r(t)] + \epsilon \quad (4.3)$$

where $\epsilon$ represents photon and dark-count shot noise, ignoring spectral cross-talk. While the data from each pixel $y_i$ are determined from multiple sources $x_j$, sparse reconstruction theory states that a given number of recording sites allows localization of a smaller or equal number of sources, even though the number of voxels, and thus the dimensionality of the problem, may be much higher. For the MIMAS imager, the out-of-phase angular transmission spectrum forces orthogonality of the $A_g$ and $A_r$ submatrices, supplying the information required to correctly identify point sources of multiple colors.
The quality of localization of each voxel within the field of view can be quantified using two metrics, the photon count and the linear separability. The photon count from a voxel (relative to a SPAD imager in the absence of the MIMAS mask) determines the achievable signal-to-noise-ratio (SNR). The linear separability for each voxel $j$ is quantified by the cosine of the angle between the $j$th column of the sensing matrix $A$, $A_j$, and the hyperplane spanned by all other vectors in the

Figure 4.5: Localization in scattering medium. (a) Photomicrograph of 45 µm polymer beads containing Fluoresbrite YG, dispersed in 1.5% agar simulating scattering tissue. The probe is inserted at $z = 0$ orthogonal to the direction of excitation light. (b) Recorded data from MIMAS imager chip at 0ns delay (c) Time series for mean recorded data, estimated fluorescent signal and excitation background through blind source separation. (d) Estimated scattered illumination background and fluorescent data after blind source separation. (e) Volumetric localization through an $L=6$ sparse representation, compared with fluorescent bead ground truth. (f) Lifetime calibration for FITC as a function of environment pH. The lifetime fingerprint of pixel 370 reveals the Fluoresbrite is encapsulated in an environment of pH=9.

The quality of localization of each voxel within the field of view can be quantified using two metrics, the photon count and the linear separability. The photon count from a voxel (relative to a SPAD imager in the absence of the MIMAS mask) determines the achievable signal-to-noise-ratio (SNR). The linear separability for each voxel $j$ is quantified by the cosine of the angle between the $j$th column of the sensing matrix $A$, $A_j$, and the hyperplane spanned by all other vectors in the
matrix. Separability is expected to decrease with voxel density. We calculate both metrics for each voxel within a subset of the field of view, generated from a 1000×1000×150 µm volume at 15 µm rectangular grid.

Fig. 4.3d compares linear separability and photon count between the current MIMAS design and a metal-grating-based angle-sensitive design. The dual-bandpass solution provides double the effective pixel density for each color channel as compared to a metal-based angle-sensitive array with separate red and green filters for each pixel. For MIMAS, the spectral sensitivity and added orthogonality between both color channels provides a 95% higher separability value. A 25% transmission efficiency over all angles, compared with 3% previously reported for metal gratings\textsuperscript{144}, improves signal photon count by over 8×.

For the metal grating-based angle-sensitive implementation, there is already limited ability to perform spectral decomposition. However, the green ($A_g$) and the red ($A_r$) sensing submatrices are highly correlated. This provides an insufficient amount of color sensitivity to cope with the doubled dimensionality of the multispectral inverse imaging problem.

The minimum distance between two neighboring point sources, however, as recovered by a sparsity-regularized least-squares solver, is limited by the dimension of the point spread function (PSF). We calculate the PSF, as seen by the pixel ensemble, around a point of interest $x$ by applying the projection matrix $P_A$:

$$\text{PSF}_g = P_{A_g} \cdot x = A_g^T (A_g A_g^T)^{-1} A_g \cdot x \quad (4.4)$$

Fig. 3e shows how a simulated point source, centered above the pixel ensemble at a height of 150 µm, resolves to a point spread function for the green color. The structure of the PSF is a combination of co-sinusoids as a result of the angular sensitivity, shown in detail along the $x$ dimension. Consequently, the PSF confinement in the $xy$-plane depends strongly on the $z$ location.
Objects closer to the imager will, therefore, enjoy higher resolution\textsuperscript{144}. While the PSF has a FWHM of 52 µm, its central peak at the source location is unbiased, enabling sparse source localization. Fig. 4.3e also shows the severely aliased spatial confinement of the PSF of an array where each pixel is alternatively red and green filtered.

4.2.5. Fluorophore demixing and imaging

In order to validate the MIMAS imager in a context relevant to neural imaging, we demonstrate simultaneous detection of multispectral fluorescent sources. We do this first by placing a pinhole array carrying agar stained with a fluorescent dye 400 µm above the imager (Fig. 4.4a). The pinholes are alternatively stained with ATTO 465 (maximum emission 515 nm, lifetime 5.0 ns) and ATTO 490LS (maximum emission 660 nm, lifetime 2.6 ns) as proxies for fluorophores such as GFP and RFP.

An NKT EXU-6PP mode-locked supercontinuum laser carries 1.5mW of average power at 19.5 MHz repetition rate, and illuminates the field of view through the pinhole array. The pinhole array location and rotation are oriented such that the individual pinholes are within the field of view of a maximum number of pixels. The output from the SuperK Varia acousto-optic transmission filter was tuned to 460 nm with 40 nm bandwidth to approximately equally excite ATTO 465 and 490LS, accounting for respective excitation spectra and quantum yields. Even though the two dyes exhibit a 5% emission overlap in the yellow regime, the MIM filter attenuates cross-talk to less than 1% of total signal power. The majority of blue excitation light is not converted to fluorescent signal and is thus projected on the imager as background. In order to
maximize signal yield, we increase the laser power to the point at which we begin to observe pile-up in the SPADs, which typically occurs when a photon arrival event occurs in more than 3% of all SPAD frames\textsuperscript{144}.

The imager records the data $Y$ (Fig. 4.4b) with a framerate of 10 Hz. We select this framerate as it is the minimum required for functional neural imaging. Data from “hot pixels” (more than five times median dark counts) are discarded. Regions in the data with both high intensity and strong modulation are indicative of a source. The combination of the MIM and absorption filter rejects blue excitation light with OD 3, while the angular sensitivity rejects the spatially uniform background excitation light with an additional OD 0.8. In contrast, fluorescent light, entering at an angle of maximum transmission, is transmitted through the MIMAS with over 30% efficiency. A baseline recording in absence of fluorescent sources is plotted for comparison.

Figure 4.6: Ex-vivo recording. (a) Brightfield image of the whole brain slice laminated over the chip. The fluorescence image reveals the beads distributed through the tissue. Brightfield image of the location over the chip. (b) Recorded data, plotted alongside the background recorded for illumination of brain slice in the absence of beads. (c) Inverse image of the data recorded, compared to the true locations denoted by red crosses. The contour is drawn around all values corresponding to 10% of the maximum value of the image.
By time-gating the SPAD detectors (enabling them after the laser has been shut off) with a shifting 20% duty-cycle enable signal, we can measure fluorescence lifetime\textsuperscript{124,144}. Fig. 4.4c shows calibration lifetime measurements for uniformly stained 10 µMol l\textsuperscript{-1} ATTO 465 and 490LS fluorescent dye in 1.5% w/v agar in phosphate buffered saline (PBS, pH 7.4), representative of neural tissue densely labeled with EGFP\textsuperscript{85}. The estimated lifetimes for the green (5.3 ns) and red (3.4 ns) dyes are a convolution of the dye lifetimes (5.0 ns, 2.6 ns) and the lifetime of the SPAD instrument response function (IRF) (1.2 ns).

The lifetime characteristics of the fluorophore aid in robust localization of the fluorescent sources even in the presence of insufficient excitation light rejection. Separation is further aided by the presence of more than one spectrally distinct fluorophore of different lifetimes. Using orthogonal non-negative matrix factorization (O-NNMF)\textsuperscript{165}, we incorporate the calibrated lifetimes as \textit{a priori} information to correctly separate the recorded data \(Y(t)\) into green and red channels in \(S(t)\) (Fig. 4.4c) and their associated data in \(H_g\) and \(H_r\) (Fig. 4.4d):

\[
Y(t) \xrightarrow{O-NNMF} H \cdot S(t) + \epsilon = [H_g \quad H_r] \begin{bmatrix} S_g(t) \\ S_r(t) \end{bmatrix} + \epsilon \quad (4.5)
\]

Even though neither \(S(t)\) nor \(H\) supply any spatial information, the source separation acts as a matched filter in the temporal domain, removing background excitation photon counts in \(\epsilon\) and enhancing SNR in \(H\). This blind source separation (BSS) is only possible as \(H_g\) and \(H_r\) span orthogonal subspaces because of the properties of the MIMAS conditioning masks. The separated red and green channel have a total correlation of 0.46 over the entire data sequence. This correlation is directly related to the correlation between the two lifetimes. The estimated background is small and has a fast lifetime decay (1.2 ns), equal to the SPAD IRF.
We make a best guess for $L$, the total number of fluorophore locations, by looking at the singular value decomposition of $Y$. We localize the fluorophore locations for each color channel separately using the sparsity-constrained in-crowd algorithm\textsuperscript{138}. Afterwards, an inverse image is formed by applying the projection matrix to the $L$ most likely locations. Fig. 4.4e shows the inverse image for $L=8$. All pinhole locations are found and show local maxima at the true locations. The circular aliasing artifacts in the image are a direct result of the combination of rotated angle sensitive pixel responses, as seen in the structure of the PSF. The solution is extremely robust against variation in input parameter $L$.

4.2.6. Fluorescent beads in scattering medium

We next use a phantom model to verify that the MIMAS CMOS imager is capable of detecting localized fluorescent sources in scattering media. We insert the shank into 1.5% w/v agarose (Fig. 4.5a). The imager is positioned orthogonal to the direction of illumination (at $z=0$ in Fig. 4.5a) to minimize excitation laser light directly entering the SPADs. Fluoresbrite YG beads with 45 µm diameter are randomly dispersed inside the agar. These beads provide the highest brightness while constraining dimensions within the resolution limit of the imager.

The laser illuminates the field of view through a 20× water immersion objective. The brightfield image at four different depths reveals six beads inside the field of view of the shanks. The data, recorded at a framerate of 10 Hz, is plotted in Fig. 4.5b. As the 20× objective does not cover the entire field of view and agar scatters photons anisotropically, the illumination is not uniform. This will be characteristic of systems employing any kind of spatially localized illumination \textsuperscript{4}. The emitter matrix, $E$, describes the illumination photon flux for each voxel and is incorporated in the linear system of equations:

$$Y(t) = A_g x(t) E + \epsilon \quad (4.6)$$
A simple model is computed using ValoMC\textsuperscript{166} based on the Henyey-Greenstein phase function for anisotropic scattering. We use a scattering length of 2.86 mm\textsuperscript{-1}, refractive index of 1.34, and anisotropy parameter of 0.9\textsuperscript{167,168}, closely resembling neural tissue\textsuperscript{127,169}. Absorption inside the agar is considered negligible. We model a conical source with a 450 nm wavelength and 1000 µm FWHM, entering the agar from water with index 1.33.

The source-separation technique for background estimation and removal is repeated. The long fluorescence decay is used to separate the data into fluorescence (9 ns) and background (3 ns) lifetimes (Fig. 4.5c). The average lifetime of the estimated background is longer than calibrated, as the inherently imperfect separation algorithm attributes some fluorescent photon counts to the background. The estimated fluorescent and background contributions are plotted in Fig. 4.5d. We localize the fluorescent sources using the sparsity-constrained least squares solver. Fig. 4.5e shows the localized fluorescent targets compared to ground-truth locations. The mean absolute Euclidian error averaged across the five identified sources is 99 µm, less than twice the FWHM. A sixth location is too far away to be found.

The effectiveness of temporal fluorescence separation is limited by the signal-to-noise ratio (SNR), given by:

\[
SNR = \frac{N_{flu}}{\sqrt{N_{flu} + N_{BG} + N_{DCR}}} \quad (4.7)
\]

where \(N_{flu}\), \(N_{BG}\) and \(N_{DCR}\) are the fluorescence, excitation background and dark noise photon counts under Poisson photon arrival statistics, respectively\textsuperscript{92,170}. At best, we can subtract a per-pixel estimate for the background but not its induced shot noise. The pixel with peak signal (pixel 370) registers a maximum SNR of 32 dB. Over the whole imager the average SNR is 17 dB. The spatial reconstruction accuracy of a single fluorescent source is limited by the SNR, the MIM filter FWHM, and fluorophore emission spectrum.
Finally, we demonstrate how the chip can exploit lifetime imaging to optically measure pH levels\textsuperscript{171}. Functional groups of fluorescein are modulated by pH, modulating the fluorescent lifetime\textsuperscript{172,173}. Inside neural tissue, local variations in pH could indicate glial cell misfunction\textsuperscript{174}, activation of pain signaling\textsuperscript{175}, or can even be used to directly image synaptic activity\textsuperscript{176}. State-of-the-art implantable pH sensors based on electrodes offer low electrode count in a narrow field of view\textsuperscript{177}, while sensors based on diffusion dynamics suffer from unpractically long acquisition times\textsuperscript{178,179}. The optical paradigm could offer fast framerates and a large field of view.

As a calibration, we dissolve 10 \textmu Mol l\textsuperscript{-1} in 1.5% w/v agar using PBS with pH levels of 5.8 and 9.0. Fig. 4.5e shows how the FITC lifetime increases strongly with environment pH. The lifetime as seen by the imager ranges from 4.5 ns (pH=5.8) to 7.8 ns (pH=9). We compare these calibrations with the lifetime registered by pixel 370 in the Fluoresbrite experiment, which correlates strongest with the calibrated lifetime for pH=9. By these means, the imager can simultaneously image and infer environmental pH of a fluorescent target.

\subsection*{4.2.7. Ex-vivo fluorescent imaging}

We image fluorescent beads \textit{ex-vivo} through a whole brain slice to verify operation inside strongly scattering tissue. The 150 \textmu m thick coronal mouse brain slices are contained within a 3D printed mold and perfused with artificial cortical spinal fluid (ACSF). We position 45 \textmu m FITC beads on the side of the tissue opposite the shank (Fig. 4.6a) to emulate clusters of densely intracellularly dyed neurons\textsuperscript{180}. The brightfield image reveals 15 beads within the red annotated field of view. The tissue structures are also clearly visible. For excitation, a supercontinuum laser is tuned to a spectrum of 480-490 nm and is projected onto the slice through a Nikon 4x 0.15NA objective. As can be seen in the fluorescent image, the illumination profile is constant over an 800 \textmu m diameter area. Uniform illumination is key when imaging over very large fields of view.
We compare photon counts from slices both with and without beads present (Fig. 4.6b), and find that the background (primarily resulting from unfiltered excitation photon scattered into the detectors) is about 10× dimmer than the fluorescent signal from the beads. We estimate the green fluorescent signal (“green channel”) by subtracting the background from the raw data for a maximum estimated SNR of 22 dB.

The resulting image of 15 source locations is shown in Fig. 4.6c. We find 10 sources with an average error of 200 μm, which corresponds to less than 2× the measured PSF FWHM for a single 45 μm bead in scattering tissue. The other five beads are found as false positives in larger clusters next to correct locations. The imager also struggles to correctly identify beads illuminated less brightly by the laser. It is important to note that these 45 μm beads used here have a fluorescence that is approximately 100× brighter than what would be expected from the soma of a GCAMP6f transgenic neuron, for example. Further improvements in background rejection (approximately an additional OD2) will be necessary to observe fluorescence of the intensity required for GCAMP6f imaging.

4.3. Discussion

In this work, we presented an implantable imager, based on an optical sensor front end that combines optical bandpass filters with angular selectivity for compressive sensing of two colors. The MIM grating, fabricated by interleaving two MIM bandpass filters, creates a superposition of two independent Talbot diffraction patterns. The MIMAS front end, fabricated by superimposing a phase grating above the MIM grating, provides well-defined, orthogonal angular modulation for two wavelengths.
The MIMAS conditioning mask increases multicolor sparse localization over conventional angle-sensitive designs employing metals for both diffractive and analyzer grating in three important ways. First, the interleaved color filter reduces color channel cross-talk within the same photodetector by a factor of 10, improving location and color classification of fluorescence point sources. Secondly, because of the use of phase diffractive grating, the color filter phase grating design is 15% efficient over all angles, compared to 3% using metal gratings. Third, the orthogonality between the two-color channel sensing submatrices enables blind source separation of fluorescent lifetimes, greatly enhancing removal of background fluorescence which is characterized by a different lifetime.

By augmenting a CMOS SPAD array with a diversity of MIMAS geometries, the correct localization and labeling of green and red fluorescent dyes was demonstrated. The MIMAS SPAD array was implanted into an optically scattering medium. The blind source separation technique is able to robustly identify fluorescent beads even in the presence of incomplete filtering of the scattered excitation light.

Additionally, measurement of fluorescence lifetime enables the imager to be used as an implantable sensor with the same resolution and field-of-view achieved without lifetime discrimination. In particular, we use the dependence of fluorescent lifetime on pH of fluorescein to demonstrate pH sensing capabilities. As fluorescein suffers from reduced quantum yield with decreasing pH, the development of improved pH sensitive fluorescent labels will aid in measurements of large cell populations.

The effect of tissue scattering was investigated ex-vivo through 150 μm thick whole brain slices. Here, we observe an approximate 2× reduction in resolution for the same fluorescent
sources when moving from agar to neural tissue. The false count rate also increases by approximately 33%.

Ultimately, incomplete rejection of the laser excitation background photons restricts resolution, frame rate and source density. Improved spectral filters are required to image smaller, less bright fluorescent sources. We estimate that an excitation rejection of OD6 is ultimately necessary for detecting somas labeled with GFP in-vivo, requiring an additional OD2 of filtering on top of the current implementation. The resulting increased SNR should allow framerates above 10 Hz. Better spectral filters will also allow for better temporal separation of more similar lifetimes, by shifting the burden of background removal from the temporal to the amplitude domain, improving sensing resolution based on fluorescent lifetime changes. Background rejection through temporal means can be further improved by reducing autofluorescence and autofluorescence lifetime in the filter dye material, improving SPAD IRF, and exploiting larger fluorophore lifetimes.

In neural tissue, scattering and absorption limit the depth of field\textsuperscript{127}; optical power typically drops 90\% at a distance of 200 \( \mu \)m from the source. This makes light delivery at depth challenging. These rigid shanks could be co-inserted with a cannula, waveguide arrays\textsuperscript{181} or tapered fiber\textsuperscript{182,183} coupled to a mode-locked laser. Alternatively, light emitters, such as light-emitting diodes, could be cointegrated on the shank. In all cases, increasing the optical power is only of benefit if the excitation-wavelength background due to scattered photons can be sufficiently rejected. The imaging depth of field is also limited by the angular resolution of the pixels, degrading resolution at large distances from the shank.

MIMAS enables multispectral detection without requiring separate pixels for each color, as is typically done in imagers. MIMAS combines light-field imaging with pixel-to-pixel color filtering, greatly enhancing resolution over conventional angle-sensitive designs, while blocking
background excitation light in the context of fluorescent imaging. MIMAS sets no requirements on the detector and adds negligible thickness, enabling many spectral sensing applications for light-field imaging, especially in applications where total imager volume is highly constrained.
5. DISTINCT: An Algorithm for fast higher order separable Basis Pursuit Denoising

5.1. Introduction

Compressed sensing is a powerful signal processing tool for reconstruction of signals of interest from undersampled data. It relies on the parameter space of the signal of interest to be far sparser than the dimensionality of the sampling space. Exact and computationally efficient reconstruction of this sparse representation is a central problem in compressed sensing. In general, compressive sensing relies heavily on \textit{a priori} knowledge of the model generating the data. In turn, that knowledge can be exploited to improve recovery algorithms for that exact application. Reconstruction algorithms have been improved by exploiting spatial-temporal data structures, by incorporating noise estimates, or using prior estimates for the signal sparsity. Most receiving systems are modeled with a simple linear forward transform:

\[ y = Dx \quad (5.1) \]

In such cases, the receiving end has no direct control over preconditioning of the signal \( x \). The receiver can only optimize the detection matrix \( D \) to correctly reconstruct the signal from the data \( y \):

\[ \hat{x} = \arg \min_x \|x\|_0 \text{ s.t. } y = Dx \quad (5.2) \]

Conversely, in systems where direct control over signal preconditioning would be possible, the preconditioning optimization opens up a new dimension for extracting information about the signal of interest. Systems exploiting compression on the input side in addition to the detection side are found in multiple-input-multiple-output (MIMO) telecommunications\(^{184}\), range-doppler RADAR
holographic imaging. However, the formulation is not limited to applications where data is sent back and forth to the same user. Image compression and encryption can benefit from two-dimensional paradigm. In video capture, correlations in the temporal dimension can be exploited to recover neural spike trains. The formulation with modulation on both the system input and output is called two-dimensional compressive sensing (2D-CS).

Consider the case of the fluorescence image vision system shown in Fig. 5.1. For biological imaging applications, the small thicknesses of the chip to minimize tissue trauma precludes the use of lenses. Furthermore, the dimensionality of the three-dimensional field of view, at a 50µm resolution, is much higher than the number of photodetectors and photoemitters. Such a system therefore has to rely on compressive sensing to image point sources. A well-designed compressive sensor can detect a number of features equal to \( M \), the number of photodetectors.

However, compressive sensing is not limited to the detector side. The illumination light can be modulated to induce extra sparsity to each recording frame. In matrix terms, the field of

![Fig. 5.1: Image vision system applying compression at both the emitter and detector side. The computational mask compresses the three-dimensional field of view onto the two-dimensional photodetector array. An ensemble of individually switchable micro-LEDs sparsely illuminates the field of view.](image-url)
view $x$ is two-dimensionally compressed through a left detection matrix $D$ and right illumination matrix $E$:

$$Y = DXE^T, \ X = \text{diag}(x) \quad (5.3)$$

The data $Y_{m,k}$ recorded from each pixel $m$ for each illumination pattern $k$ can be described as:

$$Y(m, k) = \sum_{i=1}^{N} D(m, i) \cdot x_i \cdot E(i, k)$$

While $D$ compresses the signal space, the $E$ matrix serves as an oversampling method. As each element in $x$ is a unique location in space, the diagonal constraint appears. A location in space cannot be illuminated as one, and detected as another. The problem is completely separable in $D$ and $E$.

### Table 4: The DISTINCT algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initialize</td>
<td>$I = 0$, $k = 0$, $g = \text{diag}(DYE) - \lambda$, $u = g$</td>
</tr>
<tr>
<td>2</td>
<td>Add indices whose likelihood surpasses 0 to active set</td>
<td>$I \equiv \begin{bmatrix} I \mid (u &gt; 0) \end{bmatrix}$</td>
</tr>
<tr>
<td>3</td>
<td>Solve least squares on the active set $I_c$ using the SVD or solve least squares in terms of the Hessian</td>
<td>$\hat{x}<em>j = \arg \min</em>{x_j} | U_P S_j V_j^T - D_I X_j E_j |_F^2$</td>
</tr>
<tr>
<td>4</td>
<td>Remove zero entries from the active set</td>
<td>$\hat{x}_j = (</td>
</tr>
<tr>
<td>5</td>
<td>Update usefulness</td>
<td>$u = g - H_l \hat{x}_l$</td>
</tr>
<tr>
<td>6</td>
<td>Update cost function</td>
<td>$J(\hat{x}) = | \hat{Y} |_F^2 + (\hat{x} - g - \lambda)^T (\hat{x} - g - \lambda)$</td>
</tr>
<tr>
<td>7</td>
<td>Check for convergence return to step 2 otherwise</td>
<td>$\frac{J(\hat{x})}{J^{old}} &lt; \varepsilon_1$, $\frac{| \hat{x} |_F}{| x |_F} &lt; \varepsilon_2$</td>
</tr>
</tbody>
</table>
We record k unique frames of a static scene x, each frame illuminated by different combinations of \(\mu\)LEDs as defined in columns of ET. To recover a sparse image, we modify eq. 5.2 to incorporate the additional matrix product:

\[
\hat{x} = \arg \min_x \|x\|_0 \text{ s.t. } X = \text{diag}(x), \ Y = DXE^T \ (5.4)
\]

which we call two-dimensional separable compressed sensing (2D-SCS).

There are numerous algorithms for sparse recovery of the 2D-CS problem. NIHT \(^{191}\) is adapted to 2D \(^{192,193}\). The 2D Smooth L0 (2D-SL0) \(^{194,195}\) is a highly accurate algorithm. The Iterative adaptive approach requires a signal model \(^{184,196}\), as does 2D MUSIC \(^{197}\). Often, the imaging problem is treated in one-dimensional form \(^{28}\) to apply standard sparse recovery techniques.

However, these methods suffer from slow execution times, limiting their applicability for real-time image reconstruction. First, none of these algorithms explicitly exploit the diagonal matrix structure that arises in applications with separable input and output modulation. Secondly, these algorithms optimize over the entire solution space. For applications where the solution is sparse, active set methods have provided significant speedup.

Here, we present DISTINCT, a two-dimensional basis pursuit denoising algorithm for solving the 2D-SCS. Compared to existing algorithms, DISTINCT provides the solution with the highest accuracy compared to the true image. Additionally, we find an order of magnitude computational speedup compared to all existing algorithms. We demonstrate the added value of the two-dimensional sparse recovery problem by comparing to the case in the absence of illumination modulation. We apply DISTINCT to imaging data of fluorescent point sources, and find DISTINCT to provide both the best image quality and an order of magnitude speedup compared to existing algorithms.
This paper is structured as follows. In Section 5.2, we discuss basis pursuit denoising and how the two-dimensional adaptation gives rise to the DISTINCT algorithm. In Section 5.3 we compare reconstruction accuracy and computational efficiency versus existing algorithms for sparse imaging problems. Section 5.4 derives matrix orthogonality requirements for guaranteeing exact recovery of $x$, and proofs for converging to the exact solution.

5.2. The DISTINCT Algorithm for compressed sensing with input and output modulation

A popular method for approximate sparse recovery is basis pursuit denoising (BPDN). To avoid non-convexity, the $l_0$-norm in the cost function (eq. 5.4), is relaxed to the $l_1$-norm. Solving for the $l_1$-norm produces the sparsest possible result while guaranteeing convexity. Data fidelity is incorporated as a minimization of the mean square error (MSE). The tradeoff between sparsity and data fidelity is controlled by a parameter $\lambda$:

$$\hat{x} = \arg \min_x \lambda \|x\|_1 + \|Y - DXE^T\|_F^2 \quad (5.5)$$

with a larger $\lambda$ resulting in a sparser solution. To adapt BPDN to the 2D matrix structure, we choose to convert an existing algorithm. Active set methods are preferred fast solvers, as they restrict large matrix operations to a subset, and adjust the subset cardinality iteratively. The in-crowd algorithm\textsuperscript{138} adds further speed-up by exploiting separability between the $l_1$ and $l_2$-norm. Only indices that exceed a likelihood prior are fed into the exact least mean squares solver. Following similar steps, we outline the DISTINCT (Diagonalized Shrinkage and selection Through In-Crowd Translation) algorithm in table 4.
First, the active set $I_C$ is initialized to zero. Second, indices exceeding $\lambda$ on the likelihood function are proposed candidate indices for the active set. The user can choose to add either all or $p$ most likely candidates. The signal on the active set $x_C$ is then recovered in the least squares sense. This problem is small, dense and overdetermined. We present two least squares formulations, one via the singular value decomposition and one via the Hessian. The singular value solver solves the matrix equation:

Fig. 5.2. Numerical benchmark performance of DISTINCT and select competitors as a function of the number of sources $L$. The system carries $M=1024$ detectors and emitters, and the dimensionality of the problem is $N=6468$. (a) Error rate, defined as 1-accuracy (b) Computational time required by each of the algorithms to solve the image reconstruction problem. (c) Figure of Merit, defined as the inverse of the error rate times computation time.
\[ \bar{x}_C = \arg\min \| U_Y \Sigma_Y V_Y^T - D_I X_I E_I^T + \lambda \| ^2 \] (5.6)

Here, both \( \Sigma_Y \) and \( X_C \) are diagonal, \( \text{col}(U_Y) = \text{col}(D_I) \), \( \text{col}(V_Y) = \text{col}(E_I) \). We solve eq. 5.6 by finding the correlation between left and right singular vectors and \( D \) and \( E \), and multiplying with the corresponding singular value:

\[ \bar{x}_I = [D_I^T (U_Y) \circ (V_Y^T (E_Y^T)^+)^T] (\sigma_Y - \sigma_{\lambda}) \] (5.7)

Where \( \sigma_Y = \text{diag}(\Sigma_Y) \) and \( \sigma_{\lambda} = [(D^T U_Y)^+ \circ (V_Y^T E_Y^+)^+] \lambda \) is a transformed version of \( \lambda \) to scale the singular values. This method requires the SVD to be calculated for each dataset \( Y \). In low noise cases however, the SVD reveals an estimate for the solution sparsity \( L \). Matrices \( U_Y, \Sigma_Y, V_Y \) can be truncated to \( L \) columns, strongly decreasing the computational effort. Having an estimate for \( L \) helps selecting \( \lambda \) and \( p \), reducing the number of iterations required for adjusting the size of the in-crowd.

Our second formulation solves the normal equations, stemming from setting the gradient to zero:

\[ \bar{x}_I = \arg\min \| H_C x_I - \text{diag} \{ D_I^T Y E_I \} + \lambda \| _F ^2 \] (5.8)

With \( H_C \) the Hessian on the active set. Fast gradient descent methods can be used to solve the normal equations, denoted as DISTINCT-GD. The practical computational complexity of eq. 5.8 depends on the condition number of \( H \). In the ideal case, when the Hessian closely resembles the identity matrix, inversion is simply element-wise division. Furthermore, the solver can be warm started with the solution from previous iterations, as opposed to the SVD method. Secondly, while the SVD method uses full-sized columns to represent the data, the dimensions of the normal equations are only the size of the in-crowd. The normal equations use a pre-compressed version of
the data in \( \text{diag}\{D_C^T Y C\} \). In noisy situations, the threshold for removal of small elements in \( \hat{X}_l \) can be increased, shrinking the active set and decreasing computation time.

5.3. DISTINCT accuracy and computation time versus competing algorithms

To benchmark the performance of DISTINCT versus existing algorithms, we use a Monte Carlo simulation. We generate a two-dimensional compressive sensing optimization problem and apply the algorithm to solve for the optimal image.

The optimization problem is defined as follows. We define a three-dimensional field of view with dimensions \( 3800 \times 1350 \times 150 \) µm, at a 50 µm grid spacing, for a total of \( N=6468 \) potential locations. We define an imager with \( M=1024 \) detectors and \( K=1024 \) emitters (Fig 5.1). The detectors and emitters are placed at a 25 µm pitch. In previous work, we have demonstrated computational imaging using an imager with a similar form factor \(^{198}\).

The detectors are equipped with angle sensitive frontends enabling them to sample the image with cosinusoidal basis functions \(^{84}\). Each angle sensitive detector has a unique front-end geometry and therefore has a unique field of view. The information provided by the overlap and non-overlap of the detector fields allows to reconstruct the image. The detector matrix \( D \) exists of \( M \) rows, each row maps the field of view to photon counts on each pixel. While the voxel dimensionality (\( N=6468 \)) is much larger than the detector count, we expect the sparse recovery scheme to succeed when the cardinality of \( x \) is smaller than \( M \).

The emitters are placed neighboring the photodetectors. Each emitter also illuminates the image with a unique illumination pattern. The columns of \( E^T \) map the spatial illumination from each emitter.
While the detector matrix works as a discrete cosine transform (DCT), the emitter matrix is similar to a Haar transform. While the DCT is global and smooth, the Haar elements are localized and sharp. We expect therefore the combination of DCT and Haar to maximize computational reconstruction.

We benchmark the algorithms via two performance metrics: the accuracy and computational efficiency of the image reconstruction. Accuracy is defined by the linear correlation...
between the reconstructed image $\hat{x}$ and the ground truth $x$. We measure computational efficiency through CPU time. The Monte Carlo simulation is repeated 100 times to establish statistical significance. For each repetition, we generate a new ground truth image $x$ by randomly distributing $L$ point sources in the field of view. We only consider the noiseless case to compare algorithm performance under ideal conditions. To solve the image reconstruction problem, we apply the DISTINCT algorithm, as well as adapted versions of 2D-NIHT, 2D-SL0 to the same image. We also compare to DISTINCT-GD, where we solve eq. 5.8 with an iterative gradient descent method instead of using the analytical SVD solution of eq. 5.7. Finally, we compare results with the performance of sparse reconstruction in the absence of the $E$-matrix. We include the In-Crowd algorithm to solve eq. 5.2. In that formulation, the field of view is illuminated uniformly.

### 5.3.1. Error rate

The accuracy ($\rho$) is defined as correlation between the true and reconstructed images. The error rate ($\epsilon$) is derived from the accuracy:

$$\epsilon(L) = 1 - \rho = 1 - corr(\hat{x}, x) \quad (5.9)$$

We find that DISTINCT shows the lowest error rate for sparse problems ($L \leq 32 = \sqrt{M}$), increasing linearly with source density $L$. The Gradient Descent solution to the DISTINCT problem extends this operation of high accuracy to $L=128$. In that regime, the 2D-SL0 delivers an error rate between three to ten times higher than DISTINCT. For $L$ beyond 128, it performs better. We find 2D-NIHT fails more often than not and is unsuitable for very sparse problems. The maximum accuracy occurs for $L=12$ and drops afterwards. When omitting sparse illumination in the 1D formulation, and solving using the In-Crowd algorithm, the error rate quickly increases to
unity. The reconstruction quality for the In-Crowd algorithm is similar to previously reported values\textsuperscript{138}.

The separable matrix structure improves locality and thus conditioning in $H$. The DISTINCT algorithm is able to exploit conditioning gains in both accuracy and computational speed. The fast increase in error rate for $L > \sqrt{M}$ due to phase transition is a limitation shared by all L1-norm minimization algorithms\textsuperscript{79}.

### 5.3.2. Computational efficiency

Fig. 5.2b shows the averaged required computing time versus density for each algorithm. We fit a polynomial to each curve and display the relationship. For DISTINCT, computing time increases logarithmically with $L$. The gradient descent version of DISTINCT increases with the square root of $L$. At low $L$ however, the overhead of calculations outside the internal solver equalizes the total computational time for both methods. The generally high computing time for NIHT is agnostic of $L$, as most time is consumed calculating computationally expensive matrix norms for checking convergence. We find the DISTINCT algorithm to be a three to four orders of magnitude faster than SL0, whose computational complexity increases linearly with $L$. The computational performance for the In-crowd algorithm is almost independent of $L$. Solving on a subset provides major speedup as compared to other two-dimensional sparse reconstruction algorithms. We find additional speedup compared to the in-crowd algorithm by omitting calculation of the residual which is implicit in the usefulness parameter.

### 5.3.3. Figure of Merit

We define a Figure of Merit ($FoM$) as:
\begin{equation}
FoM(L) = \frac{1}{\text{error rate} \times \text{computing time} \times L}
\end{equation}

where higher is better, giving equal weight to error rate and computing time. The FOM (Fig. 5.2c) reveals the DISTINCT algorithm yields a three orders of magnitude improvement in terms of combined accuracy and computational complexity for the sparsest applications. The fast computation of the 1-dimensional In-Crowd algorithm yields a similar figure of merit only for dense problems, at the cost of poor accuracy. Only when very high accuracy is required for very dense problems does 2D-SL0 justify the strongly increased computational requirement.

5.3.4. Complexity

We perform the reconstruction problem with varying problem complexities. We keep the source density constant \((L = 32 = \sqrt{M})\) while increasing the matrix size. The error rate increases slowest for DISTINCT (Fig. 5.3a). The speedup for DISTINCT also increases with dimensionality relative to its most accurate competitor, 2D-SL0. For each dimensionality, the DISTINCT-GD solver is faster and slightly more accurate. As such, its FOM is always higher than the SVD solver and two orders of magnitude higher than any other algorithm.

5.4. Convergence Guarantee and Requirements

Here we describe the convergence criteria for DISTINCT. First, we prove convexity of the cost function in eq. 5.5 and establish criteria on matrices \(D\) and \(E\). Then, we prove the local optimality of the solution after convergence. Finally, we show how every iteration of DISTINCT decreases the cost function. Adding these three proofs together, it follows logically that DISTINCT converges monotonically to a global optimum.
5.4.1. Convexity

In many applications such as computer vision, the compression matrices are constrained by physical design parameters. Here we establish criteria on the compression matrices $D$ and $E$ to in order for the cost function to be convex.

$$J(x) = J_F(x) + J_1(x) = \|Y - DXE^T\|_F^2 + \lambda |x|_1$$

For a quadratic cost function, the convexity can be verified through the second derivative. In this analysis, we omit the $l1$-norm part of the cost function, as its second derivative is zero. We start by obtaining the first derivative of the Frobenius norm part ($J_F(x)$):

$$\nabla_x(J_F) = \frac{\partial J_F(x)}{\partial x} = diag\{2D^TDXE^T - 2D^TYE\}$$

Since $X$ is diagonal, we exploit the cyclical nature of the Hadamard product to simplify the expression for the gradient:

$$\nabla_x(J_F) = 2(D^TD \circ E^TE)x - 2diag\{D^TYE\}$$

The Hessian is then given by:

$$\frac{\partial^2 J(x)}{\partial x^2} = H = D^TD \circ E^TE$$

The conclusion that the cost function is convex can be drawn when the Hessian is positive definite:

$$H > 0$$

A sufficient requirement for the Hessian to be positive definite is both $D^TD$ and $E^TE$ are positive definite, as indicated by the Schur Product Theorem. However, in this case, the Hessian is composed of the Hadamard product of matrices $D^TD$ and $E^TE$. Therefore, the Hessian can be positive definite even if $D$ and $E$ do not have linearly independent columns.

A poorly conditioned detector matrix $D$ can be counterbalanced by the addition of a well-conditioned $E$. The eigenvalues of the Hessian can be positive even if those of both $D^TD$ and $E^TE$
are not. Any subsets of columns of $D$ that are non-orthogonal, can be repaired by orthogonal columns in those indices in $E$, and vice versa.

5.4.2. Convergence criteria

For a given $\lambda$, the subdifferential of the cost function $J(x)$ is given by

$$
\partial J(x) = 2\text{diag}\{-D^T(Y - DXE^T)E\} + \lambda \partial |x|_1
$$

$$
\partial J(x) = 2Hx - 2\text{diag}\{D^TYE\} + \lambda \partial |x|_1
$$

For the case of $x > 0$, we can simplify the subdifferential of the $l_1$ norm:

$$
\lambda \partial |x|_1 = \lambda \text{sgn}(x_k) = \lambda
$$

Let constant $g = \text{diag}\{D^TYE\}$. As the true $x$ is sparse, only the subset $k \in N$ lies on the support of $x$. The indices $j \in N \sim \{k\}$ lie off the support. The optimality condition for the solution $\hat{x}$ requires $0 \in J(\hat{x})$, giving rise to our likelihood function $u(x)$:

$$
u_n(\hat{x}) = g_n - H_n \hat{x}_n - \frac{\lambda}{2} = 0
$$

For all indices $n \in N$. Off the support of $x$, $x_j = 0$, and therefore:

$$
u_j = g_j - \frac{\lambda}{2} = 0
$$

On the support of $x$, $x_k \neq 0$,

$$
u_k(x) = g_k - H_k x_k - \frac{\lambda}{2} = 0
$$

As we initialize $x^{t=0} = 0$ and $I^{t=0} = \emptyset$, we identify new candidate indices for the active set $I$ where the likelihood function supersedes $\lambda$:

$$I^{t+1} = I^t \cup \{n \in N \mid u_n(\hat{x}^t) > 0\}
$$

After identification of the active set, we solve for the exact values on that active set. The optimal $x$ is given when $\nabla_J$ is equal to zero, and can be solved in terms of the Hessian:
\[ \hat{x}_i = H_i^{-1} \cdot (g_i - \frac{\lambda}{2}) \]

As more indices are added to the in-crowd, more indices the likelihood function are decreasing towards \( \lambda \). With the addition of new candidate indices, the Frobenius norm part of the cost function \( J_F(x) \) decreases faster than the l1-norm part \( J_1(x) \) increases. Instead of using the Frobenius norm, we can rewrite the cost function in terms of the indefinite integral of its derivative:

\[
J(\hat{x}^t) = \hat{x}^T \nabla_x (f) = \|Y\|_F^2 + \hat{x}^T H \hat{x}^t - 2 \hat{x}^T \left( g - \frac{\lambda}{2} \right)
\]

For each iteration, we require updating \( u^t = g - Hx^t - \frac{\lambda}{2} \). Using those variables, we can efficiently calculate the cost function by:

\[
J(x) = \|Y\|_F^2 - x^T (u + g)
\]

We found convergence, relative to our initial error \( J(x^{t=0}) = \|Y\|_F^2 \), within some tolerance \( \epsilon \) when we achieve:

\[
\frac{J(x^t)}{\|Y\|_F^2} = 1 - \frac{x^T (u^t + g)}{\|Y\|_F^2} < \epsilon
\]

### 5.4.3. Point spread function

The resolution of the imager determines how close two point sources can be spaced and still be fully resolved. It is defined by the full width at half max (FWHM) of the point spread function (PSF). We calculate the PSF around an arbitrary location by projecting onto the column spaces of both \( D \) and \( E \):

\[
PSF_n = P_D \cdot X_n \cdot P_E = (D^T D)^{-1} D^T D \cdot X_n \cdot EE^T (EE^T)^{-1} = P_D(:,n) \cdot P_E(n,:) \quad (8)
\]
Where $P$ denotes the projection matrix, and $X_n$ denotes only the n’th diagonal index in $X$ contains a source. The inner product, as a result of the separable formulation, implies that the PSF can be separately optimized for the $D$ and $E$ matrices.

### 5.4.4. Extension to Higher Order Systems

The separable formulation can be extended to higher order systems. When data is constructed from three matrices, we consider the image $X$ to be a three-dimensional matrix with nonzero elements only on its diagonal.

$$Y = (DXE^T) \otimes Q$$

For this case, the usefulness can be found by multiplying the data with each matrix along their respective dimension. Only the elements on the tensor diagonal are considered:

$$u = diag\{(D^TYE) \otimes Q^T\}$$

The Hessian is a 2-dimensional matrix constructed from the Hadamard multiplication of all compression matrices:

$$H = D^TD \circ E^TE \circ \ldots \circ Q^TQ$$

And such the solution on the in-crowd is found by

$$\hat{x}_C = H_C^{-1}diag\{(D^TYE) \otimes Q^T\}$$

### 5.5. Discussion

We presented the DISTINCT algorithm for fast optimization of 2D separable BPDN. The modulation of the input signal, where available, adds tremendous signal recovery capability to the sparse reconstruction optimization problem. With DISTINCT, the fastest, most accurate solver is provided for the 2D-BPDN.

6.1. Introduction

Simultaneous stimulation and recordings of neuronal circuits is essential for feedback regulated manipulation of particular brain regions and neural pathways. Realization of large-scale systems with cellular resolution offers the potential to explore how neural ensembles contribute to complex behavioral processes. Current bidirectional neural interfaces are primarily electrophysiology-based, but the current levels required for stimulation saturate the recording amplifiers and prevent one from performing simultaneous recording. Even the most sophisticated techniques are unable to resolve neural activity from this stimulation background. Furthermore, with electrical approaches, stimulation is indiscriminate and lacks cell-specific targeting.

To address these limitations, bidirectional neural interfaces can instead be achieved by stimulating optically and recording electrically. Optical stimulation, of course, relies on transmembrane proteins, called rhodopsins, which become permeable to ions when exposed to specific optical wavelengths. A targeted selection of neurons can be genetically transduced by use of recombinant replication-defective viral vectors or genomic manipulation. In this way, targeted in-vivo neuronal populations can be polarized or depolarized on the cellular-level with a light-controllable transmembrane ion flux. State-of-the-art neural bidirectional optogenetic systems primarily utilize microscopy techniques to project external light sources onto the tissue. Utilizing spatial light modulators it is possible to split a single laser beam into multiple points and shape the wavefront. This way, simultaneous targets at user defined positions in the sample can be accessed while imaging the corresponding response. However, despite the excellent spatiotemporal
resolution of stimulation with these techniques, stimulation is confined to the superficial layers of brain. Current all-optical approaches to achieve optical stimulation at depth include implantable graded index (GRIN) lenses or multimode fibers. However, these approaches do not easily scale to large numbers of light sources and result in significant tissue displacement on insertion. Electrooptic approaches, such as the use of monolithically integrated GaN LEDs, are more promising for this scaling. Fabrication approaches using solid-state LEDs are complex and generally incompatible with the flexible packaging approaches employed here. In addition, producing many wavelengths for emission to allow multiplexed stimulation of different reporters is challenging.

Here, we develop mechanically flexible implantable shanks that contain high-capacitance recording electrodes and organic LEDs (OLEDs) fashioned on either side of a flexible polyimide layer. OLEDs have recently been recognized as having favorable properties as scalable, high-density light emitters for optogenetics. In OLEDs, the active layers of organic photon emitters are sandwiched between two thin-film encapsulation layers consisting of inorganic nanolaminates and parylene-C. In our case, OLEDs are positioned opposite on the polyimide substrates to transparent conducting polymer, poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate) (PEDOT:PSS) electrodes, which allows the 50-nW LEDs, illuminating through the PEDOT:PSS to act as a simultaneous recording and stimulating site. We demonstrate instantaneous, artifact-free recording during optical stimulation in mouse brain slices. Through adjustments in the OLED power, we can precisely control the field of stimulation. The materials properties of the PEDOT:PSS are critical to making this possible. In an aqueous environment, the polymer absorbs ions inside its bulk which contribute to formation of Helmholtz double layer, enhancing the capacitance of the electrodes. In addition, the optical extinction coefficient of PEDOT:PSS is less
than 0.005 in the wavelengths of interest for optogenetic applications, allowing effective light transmission through these electrodes.

6.2. Results

Fig. 6.1a shows the layered stackup of the optoelectronic devices developed here, which consists of an OLED sub-stack for light emission and an electrophysiology sub-stack for recording. The OLED stack consists of six layers based on a p-i-n architecture for a total thickness of 200 nm. These structures are optimized for high current densities at low voltage operation. The OLED functional area is defined by patterning a silver anode contact. While the OLED stack is deposited conformally over the entire die, the silver anode Ohmic contacts define the OLED positions through localized charge injection. The short lateral diffusion distance inside the organic material (<<3 µm) enables the spacing between OLEDs to be reduced to less than 3µm, while still spatially separating their light emission. The cathode is a 25-nm silver layer that is deposited with conventional thermal evaporation using a wetting-layer approach. This ultra-thin silver provides sufficient conductivity while transmitting 70% of the photons. The OLED microcavity
optical thickness is tuned to act as an optimal Fabry-Perot etalon, minimizing losses from standing wave effects\textsuperscript{213,214}. These OLEDs are encapsulated against degradation through corrosion and oxygenation by alternating layers of parylene-C and zirconia/alumina nanolaminate\textsuperscript{206}. The recording stack consists of the transparent multielectrode array (MEA) encapsulated between two layers of thin-film parylene substrates (2 µm), one acting as the mechanical substrate and the other as a passivation layer to define the electrodes. Due to the low electrical conductivity of PEDOT:PSS (370 S/cm), gold is used for interconnects. The overlap between PEDOT:PSS and gold interconnects is larger than 5 µm to reduce Ohmic losses in this contact. Fig. 6.1b shows a 50x magnified microphotograph of an active OLED with dimensions of 21 µm ×21 µm, emitting light through the recording electrode.

For applicability to a wide range of channelrhodopsins, the electrode stack requires high optical transmissivity through the entire visible spectrum while minimizing electrode impedances. The capacitance of the PEDOT:PSS scales linearly with the layer volume; a thicker layer of PEDOT:PSS, therefore, results in lower impedance. However, increasing this coating over the active pixel comes at the cost of lower optical transmissivity efficiency. Fig. 6.2 explores this tradeoff. The electrode thermal noise, given by $v_{\text{th}} = \sqrt{4k_bTRB}$, where $k_b$ is Boltzmann’s constant, $T$ is the temperature (310 K), $R$ is the resistance, and $B$ is the measurement bandwidth, is required to be lower than 5 µV rms over the bandwidth of interest between the slow dynamic local field potentials (LFPs) (>0.5 Hz) and the fast neural spikes (<3 kHz). This constrains the resistance to a maximum of 125 kΩ, assuming the electrode is followed by a first-order low pass filter with cutoff at 7KHz.

We note that the electrodes are almost purely capacitive, with the complex impedance carrying a phase of -75° (Fig. 6.2a). A resistance of 125 kW can be achieved for a PEDOT:PSS
thickness equal to or larger than 75 nm. For a 75-nm PEDOT:PSS layer, we extract a series capacitance of approximately 500 pF with a system total input referred root-mean-square noise over the 3-kHz bandwidth of 6.5 μV (Fig. 6.2b). At this thickness, we achieve a total transmission efficiency in excess of 90% over the entire visible spectrum (Fig. 6.2c). Beyond a thickness of 100 nm, transmission efficiency degrades quickly, most notably for wavelengths longer than 650 nm (Fig. 6.2d).
We further investigated electrode noise during OLED emission. Traditional metal electrodes suffer from photoelectric effects which produce recording artifacts. Highly transparent PEDOT:PSS electrodes strongly mitigate these effects. We project an external LED on the electrodes with pulse widths of 10ms at a repetition rate of 40 Hz. Fig. 6.2e shows recording artefacts as a function of LED optical power. Artefacts are not observed until optical powers exceed 5 mW/mm². However, at 5 mW/mm², stimulation thresholds for ChR2 can be reached at distances up to 300 μm.

Figure 6.3: Spatial characterization of OLEDs. (a) Confocal images for (i) OLED in solution and (ii) neural tissue with thickness $t_{slice} = 10 \, \mu m$ and (iii) thickness $t_{slice} = 100 \, \mu m$. (b) Normalized photon count detail at $Y=50 \, \mu m$ indicating scattering profile increasing with slice thickness. (c) Absolute count rate across $Y=50 \, \mu m$. (d) Integrated photon count over a circular area with 15μm diameter centered above the OLED, indicating the photon flux on the neuron surface as a function of depth. (e) OLED optical emission as a function of applied forward voltage. (f) The emission spectrum is tuned to the channel rhodopsin absorption spectrum.
To compare the artefacts results from optical stimulation with those resulting from electrical stimulation, we use the electrode to provide a biphasic stimulation waveform with a magnitude of 5 μA, which is on the lower end of what is typically used for extracellular stimulation of neuronal activity. Even for the electrodes located as far as 300 μm from the stimulating electrodes, current pulses lead to high stimulation artefacts during the time of stimulation followed by long lasting capacitive charging transients (Fig. 6.2f).

Inside tissue, optical scattering limits spatial selectivity and energy transfer efficiency. To understand these effects, we extracted brain slices of various thickness \( t_{\text{slice}} \) from wild-type mice and placed them on devices. Using a confocal microscope (Fig. 6.3a), we measured the distribution of the light intensity from a single OLED. Fig. 6.3b shows in detail the expanding light contour with distance travelled through tissue. The spatial intensity profile shows strong scattering at \( t_{\text{slice}}=100 \) μm with a 10% intensity contour width of 42.5 μm. The confocal image at a slice thickness of \( t_{\text{slice}}=30 \) μm is better confined with a 10% intensity contour dimension of 33 μm. In Fig. 6.3d, we show the average optical power density collected in a 15-mm-diameter area centered over the OLED at varying distances in tissue. For a distance of 30 μm in tissue, approximately 22% of total photons transmit to a 15-μm diameter area, comparable to the spatial extent of a soma, agreeing with simulations using the ValoMC Monte Carlo simulation package in a tissue scattering model. At an operating voltage of 6.9 V (Fig. 6.3e), the OLEDs have an output power of 200 nW, allowing them to overcome the stimulation threshold for ChR2-H134R of approximately 0.5mW/mm² at a distance of up to 20 μm.

The OLED emission spectrum can be freely tuned through choice of materials, doping implants and dimensions of the OLED microcavity. Here, we adjust the OLED output spectrum (Fig. 6.3f) spectrum to maximally overlap with Chr2-H134R, generating maximum
output emission at wavelengths between 450 nm to 490 nm (Fig. 6.3f). For optogenetic stimulation, we use the OLEDs in pulsed mode, which facilitates operation at higher current density without device degradation. Enabling and disabling the OLEDs for 10ms and 15ms respectively, also closely matched ChR2 opening and closing dynamics, resulting in the maximum stimulation bandwidth of 40Hz for ChR2.

After confirming that light delivery to the tissue is over the stimulation threshold with artifact-free recording through coplanar PEDOT:PSS electrodes, we perform simultaneous optogenetic stimulation and electrophysical recording in brain slices. We extracted hippocampus preparations from mouse (P21-P28) that possess channelrhodopsin-2 (ChR2-H134R). We mounted the 300-μm-thick brain slices on the target area of the OLED array (Fig. 6.4a) and recorded the subsequent broadband electrophysiological activity. The OLEDs were enabled for 10 seconds with a pulse repetition frequency of 40 Hz and pulse width of 10ms. The OLEDs were then disabled for 30 seconds to allow the neural activity to decay. The recordings, which were simultaneously obtained across multiple channels, had signal quality comparable to conventional tungsten recording electrode placed adjacent to the planar electrodes (Fig. 6.4b). Upon illumination of the neural populations through 32 channels, we could observe an immediate increase in the frequency response of field potentials. Multiple cycles of the same stimulation protocol showed high reproducibility in the invoked neural activity. The four unfiltered activation cycles are plot in detail in Fig. 6.4c, including two seconds before and after OLED emission. A delay between 0 and 500ms from OLED emission to LFP onset is observed. Moreover, due to high electrode recording density, it was possible to access different forms of activity across the tissue. We observed an increase in multiunit activity (Fig. 6.4d) on electrodes that are in direct contact with active somatic bodies. This multiunit activity represents simultaneous firing of multiple
Figure 6.4: Optogenetic stimulation and recording of neural spike trains. (a) Slice laminated on top of the OLED and electrode sandwich. OLEDs turned on over the region covered by the electrodes. (b) Raw data of electrophysiological recordings obtained by PEDOT:PSS (red) and tungsten (black) electrodes. (c) Recording time series. The green shaded regions mark where we enabled the OLED for 10 seconds at a stimulation frequency of 40 Hz. Spectrogram based on the Gabor wavelet transform reveals enhanced activity in the 0.1-100 Hz bands. Detailed plots for each stimulation window shows the activation with delayed onsets (d) Recorded multi-unit activity during stimulation. Top left panel is the waveform of the activity averaged from 50 traces and top right panel is raw data of sample spikes. Bottom panel shows the number of spikes before and during stimulation. (e) Wild-type brain slice control experiment. Apart from spontaneous spikes, no global activity is observed.
these neuronal populations during active stimulation validates the bidirectional communication capability down to single neuron spikes. As a control, we repeated the same experiments using wild type animals (Fig. 6.4e). The observed spontaneous activity baseline demonstrates intact slices and excellent electrode coupling. The corresponding neural activity did not show any change before or during the stimulation, proving induced hyperexcitability occurs on in the presence of the opsins.

6.3. Discussion

In this work, we presented thin-film polymer neural probes that combine optogenetic and electrophysiological modalities in a single device. By achieving multimode operation in a mechanical flexible device, we bring advantage over existing approaches that rely on either flexible electrophysiological electrodes or rigid solid-state probes that cointegrate LEDs and recording electrodes. Furthermore, the lamination approach used here allows the light emission to come directly from the electrodes themselves, going beyond simple coplanar integration. We have demonstrated the ability to tune these devices to single neuron stimulation resolution. Fabrication approaches for both the OLEDs and electrodes and their interconnection involve spin-coating and chemical vapor deposition (CVD) all of which can be performed at room temperature, making this compatible with polymer substrates. Such thin-film photoelectronic devices can also exploit surface topology to selectively emit in different spatial directions, further increasing spatial selectivity. Additionally, OLED spectral tunability allows the straightforward fabrication of multiple colors for multispectral stimulation of specific neural populations.
7. Integration of OLEDs on high-density CMOS for in-vivo optogenetic stimulation

7.1. Introduction

Large-scale recording and stimulation of neuronal populations is key both to detection of neural circuitries underlying behavior\textsuperscript{38,39} and to manipulating them selectively. Current approaches to achieve this are primarily either electrical or optical. In the electrophysiological modality, integrated circuits based on complementary metal-oxide-semiconductor (CMOS) technology have provided a successful path for microfabrication of implanted electrode arrays that allow to scale the density of active sites to match that of the neural networks in their proximity.\textsuperscript{23,220,221} Despite the excellent spatiotemporal resolution of these tools for neural recordings, electrical stimulation lack cell-type specificity and leads to high sparsity. In contrast, for optical techniques, cell-type-targeted promoters can be used to introduce opsins.\textsuperscript{44,72} However, targeted stimulation of opsin expressing neuronal ensembles necessitates light delivery systems that provide large fields-of-view while maintaining single-cell spatial and sub-millisecond temporal resolution.

Methods for light delivery have relied primarily on refinement of microscopy techniques, including the use of spatial light modulators.\textsuperscript{222} This superficial light delivery can be extended to the deep brain with graded index (GRIN) lenses\textsuperscript{46} or optical fibers\textsuperscript{27,47,223}. Illumination typically occurs only within the optical plane near the tip of the endoscope with significant tissue damage because of diameters in the 0.3 – 1 mm range. Devices based on integrated photonics to multiplex or shape light delivery also lead to bulky devices, while maintaining the need for external lasers.\textsuperscript{28,224} Instead, fully implantable optoelectronic devices, leveraging CMOS technology, would allow controllable light delivery with large fields-of-view and minimal displaced tissue
volume. Such devices could closely resemble their electrophysiological counterparts with electronic signaling, converting to light only at the point of transduction to the neural tissue. To that end, face-to-face bonding of GaN μLEDs provides more than sufficient optical pulse power (10s of mW/mm² or more) and repetition rate for optogenetics. Despite these successes, the mismatch between GaN and Si prevents growing defect-free GaN films directly onto CMOS substrates, thus precluding monolithic integration of μLEDs on CMOS. Alternatives, like face-to-face bonding of thinned GaN/InGaN μLEDs to the CMOS substrate, have low yield and require larger device sizes. Transfer printing via elastomer stamps has been suggested as an alternative but involves multiple complicated steps for fabrication as well as transfer. In addition, to achieve fine-tuned emission spectra matched to the excitation spectrum of the opsins, solid state μLEDs require either transfer printing from multiple wafers, or complex post-processed filter structures. As a result, integration of conventional LEDs for implantable devices has only been demonstrated in passive structures, in which each LED is hard-wired to its own external contact pad, leading to a limited number of sources (generally no more than 16 per shank) due to wiring limitations, with LED pitches typically constrained to no smaller than 50 μm. Given the limited emissive area of μLEDs and low fill factor, this in turn also limits the integral optical power emitted by such devices.

Integrated pixels based on organic LEDs (OLEDs) are a promising technology for scalable on-chip emission of light. Due to the amorphous structure of the organic semiconductor materials used in OLEDs, they can be monolithically integrated on silicon chips and other substrates without any lattice matching restrictions. The emission spectrum of OLEDs can be readily tuned across the entire visible spectrum by modifying the chemical structure of the emitter material. There have already been substantial research efforts to use OLED-on-silicon technology particularly for
microdisplay applications where devices with >10^6 individually addressable pixels were demonstrated.\textsuperscript{232–234} To adapt this technology for use in implantable optoelectronic arrays for optogenetic stimulation, two key innovations were required. First, significantly higher optical output powers of 0.1 - 1 mW/mm^2 or \sim 5,000 - 50,000 cd/m^2 are required, compared to the < 0.01 mW/mm^2 or \sim 300 cd/m^2 typical for most display applications. Unlike for GaN based LEDs, achieving sufficient output powers is non-trivial with OLEDs, but recent work has demonstrated optical power density in excess of 1 mW/mm^2 and robust optogenetic stimulation with OLEDs on glass and on passive silicon substrates.\textsuperscript{42,207,235–237} Secondly, the devices must be encapsulated in such a way that they can be immersed in aqueous salt solutions without degradation, remarkably challenging because of the sensitivity of the materials used in OLEDs to water and oxygen. Very recently, we have developed a chemical vapor deposition (CVD) based passivation process for OLEDs that combines an Al_2O_3/ZrO_2 nanolaminate with a parylene-based organic coating to provide a resilient barrier to water and oxygen and thus facilitates prolonged exposure of OLEDs to physiological conditions (e.g., no loss in function after weeks in saline buffer and cell culture medium at 37°C).\textsuperscript{206,238,239}

Here, we present a monolithically integrated, high-density, large field-of-view, optical probe for neural stimulation that incorporates a total of 1024 OLEDs with two different colors (orange and blue) and 256 devices on each of four shanks. The OLEDs measure 21 \mu m \times 19 \mu m at a 24.5 \mu m pitch and satisfy the brightness and repetition rate requirements for optogenetic stimulation of both red and blue channelrhodopsins. We show the probe is capable of inducing localized activity at the level of individual pyramidal neurons from a single OLED in mice \textit{in-vivo}.  

7.2. A 1024-emitter optogenetic stimulation probe in CMOS

The device is designed on an integrated circuit chip in a 130-nm high-voltage CMOS process. A 2 mm × 1 mm base contains decoupling capacitance and addressing logic. Four shanks, extending 6 mm from the base and at a pitch between shanks of 250 μm, carry two rows of 128 anode drivers each (Fig. 1a). 1024 aluminum pads, each measuring 21 μm × 19 μm at a 24.5 μm pitch, serve as anode contacts for the OLEDs. As the 1-mm-long tip of each shank contains only decoupling capacitance, the furthest anode can be implanted as deep as 5 mm. Each anode contains in-driver logic for local addressing (Fig. 1b). The chip is processed into an implantable shank form factor with a shank width of 100 μm and a thickness of 55 μm by trenching and backside milling (Fig. 1c, see Online Methods).

Top-emitting microcavity blue or orange OLEDs are monolithically integrated on these fully packaged probes by high-vacuum, temperature-controlled thermal evaporation of the required stack of organic and metallic films (see Online Methods). The OLED stack on top of the aluminum anode contact pad consists of a 3-nm-thick silver layer, a 1-nm-thick MoO₃ layer, a p-doped hole transport layer (HTL) of Spiro-TTB:F₆TNAP (190 nm for orange-emitting OLEDs, 150 nm for blue-emitting OLEDs), an electron blocking layer (EBL) of NPB (10 nm, for orange) or Spiro-TTB (10 nm, for blue), an emission layer (EML) of NPB doped with the phosphorescent orange emitter Ir(MDQ)₂acac (40 nm) or MADN doped with the fluorescent blue emitter TBPe (20 nm), a hole blocking layer (HBL) of BAlq (10 nm), an electron transport layer (ETL) of Cs-doped BPhen (60 nm), and a semi-transparent silver top cathode (20 nm) (Fig. 1d). The total thickness of the OLED microcavity is tuned to the second optical maximum for efficiency, forward directed emission, and robustness to substrate roughness (Supplementary Section S1). After OLED integration, the entire device is encapsulated using our recently developed CVD based passivation.
process that combines Al₂O₃/ZrO₂ nanolaminates and parylene layers. The final device structure is visualized by scanning electrode microscopy (SEM) after opening up cross-sections through the sample with focused ion beam (FIB) milling (Fig. 7.1e).

**Figure 7.1:** OLED-on-CMOS stimulation probe. (a) microphotograph of CMOS die of four-shank probe containing 256 OLED anode drivers and three cathodes per shank. The CMOS chip is bonded to a PCB which doubles as an insertion guide for neural implantation.

(b) The CMOS circuitry is contained underneath each anode pad, and contains local addressing, level shifting and decoupling capacitance. (c) Fabrication flow from CMOS die to implantable probe form factor, with annotated silicon (light purple), active CMOS (dark purple), copper interconnect (orange) and aluminum contacts (gray). (i) Trench definition by laser cutting, (ii) Backside thinning removes excess silicon down to 50 µm. (iii) Direct evaporation of OLEDs on shanks. (iv) Encapsulation with nanolaminate and parylene. (d) Layer architecture of the orange and blue emitting OLEDs integrated on the shanks. (e) SEM/FIB cross-section of the completed OLED-on-CMOS stack with Al contact pad covering the Cu interconnect, OLED stack, and thin film encapsulation consisting of two pairs of parylene-C and oxide nanolaminate.
Figure 7.2: Characterization of OLEDs of shank probe (a) Photograph of orange probe with 512 of the 1024 OLEDs activated over the four shanks in a checkerboard pattern. (b) Photograph of blue probe with all OLEDs on. (c) False color micrograph showing emission intensity over a single OLED pixel on the probe. (d) The emission spectra for the orange and blue probes are optimized for the channelrhodopsin variants ChRmine and ChR2, respectively. (e) Emitted power and current as a function of forward voltage for representative orange and blue pixels on CMOS devices. (f) OLED emission power as a function of pulsing frequency. (g) Monte Carlo simulation of the optical power projected into scattering neural tissue from a single blue OLED (left) and a cluster of four OLEDs (right) for operation at 7V. Contours correspond to different light intensities in the tissue are annotated. (h) Detailed plot of the projected optical power as a function of distance from the probe for the conditions in (g). The required optical power densities for ChR2 and ChRmine are annotated. (i) Temporal evolution of normalized optical power for a blue OLED driven at 6.25 MHz pulsing frequency. (j) Stability of OLED emission power over a 12-hour stress test. (k) Histogram of optical power density emitted by orange pixels operated at 7 V forward voltage.
7.3. OLED characterization

The fabricated probes with 1024 monolithically integrated OLEDs can be individually addressed through the corresponding anode contact pad. Fig. 2a shows a photograph of an orange shank device driven with a checkerboard pattern with 512 OLEDs switched ON. Fig. 2b shows a blue OLED shank with all OLEDs ON (see also Supplementary Video 1). Supporting electronics control the OLED illumination sequence (Supplementary Section S2).

Fig. 2c displays a micrograph of the emission from a single OLED on the shank. The variations observed in light emission over the surface of the electrode are largely the result of surface roughness on the aluminum contacts on the foundry CMOS process (see Supplementary Section S3). Because the average surface condition of the anode determines device performance, micrometer-scale variations in morphology and brightness within individual pixels did not affect the ability of the device to perform targeted optogenetic stimulation. The emission spectra of the blue and orange OLED pixels match the activation spectra of ChR2 and ChRmine, respectively (Fig. 2d).

Fig. 2e shows the emitted optical power and forward current as a function of forward voltage for both the orange and blue CMOS-integrated OLEDs. Like conventional GaN based LEDs, OLEDs suffer from a reduction in efficiency at high brightness (known as ‘droop’ or ‘roll-off’). Roll-off and Ohmic resistance dominate at supply voltages above 5 V, and the measured supply current becomes a linear function of applied voltage. Pulsed operation of the OLEDs at 10 kHz, (see Online Methods), is used to maximize emission power (Fig. 2f). Under these conditions, the blue OLED pixels emit an average power of 100 nW (0.25 mW/mm²) whereas the orange OLEDs reach an average emitted power of 40 nW (0.1 mW/mm²). These optical powers were measured with the shank in air. Due to the low refractive index of the OLED materials (nOLED ≈ 1.75) and
the thin film encapsulation \( n_{\text{parylene}} \approx 1.64 \), a substantial increase in the efficiency of light-extraction from the device is expected when it is in direct contact with aqueous tissue \( n_{\text{tissue}} \approx 1.37 \); to a first approximation, this increase is proportional to \( (n_{\text{tissue}})^2 \approx 1.9 \)-fold.

We perform Monte Carlo simulation in scattering tissue\(^{216} \) to predict how much light a neuron receives as a function of distance from the OLED (see Online Methods). The microcavity formed by the electrodes surrounding the organic layers of both the blue and orange OLED pixels causes the light to be emitted with an emission profile narrower than Lambertian (Supplementary Section S4). We consider the case of the blue OLEDs, which are assumed to be operating at 7 V, i.e. at an average optical power density of 0.25 mW/mm\(^2\). As shown in Fig. 2g, due to its small size, a single OLED resembles a point source, for which the projected light falls off quadratically with distance from the OLED. This rapid fall-off in optical power can be offset by turning on more OLEDs to contribute to the illumination field at the cost of reduced spatial resolution, also shown in Fig. 2g for four OLEDs. Enabling clusters of contiguous OLEDs increases both the volume of stimulation and the stimulation intensity.

To induce spikes with a 10% probability, ChR2 (H134R) requires an irradiance of approximately 0.1 mW/mm\(^2\); for 100% probability the power density requirement is 0.5 mW/mm\(^2\).\(^{70,100,240} \) For the more sensitive red-shifted ChRmine opsin \(^{21,101} \) these irradiance requirements are relaxed to 5-50 \( \mu \)W/mm\(^2\) for 10-100% probability. For a single OLED, the ChR2 threshold power densities can only be realized for neurons in immediate proximity to the shank (Fig. 2h); enabling four contiguous OLEDs allows activating neurons stained with ChR2 at distance up to 100 \( \mu \)m in the near-threshold regime. However, for ChRmine, this distance can be enhanced to 225 \( \mu \)m by illumination from only a single OLED.
Direct integration of OLEDs with CMOS drivers minimizes the parasitic capacitances associated with longer interconnects, which allows us to achieve OLED rise and fall times of 40 ns and 20 ns, respectively. (Fig. 2i) These rise and fall times are substantially faster than the response time typically achieved for passively switched LEDs. In addition to setting the color to match the absorption spectra, this creates the opportunity for OLEDs to be carefully matched to the temporal dynamics of the opsins.

With maximum supply currents of 30 µA at 7 V and at 10 kHz pulsed operation, the maximum power consumption per pixel is 210 µW. Comparing these power consumptions to FEM-based simulations of heat dissipation indicates that this results in negligible heating in the tissue surrounding the pixel (Supplementary Section S5). As local heating also leads to potential degradation of the organic emitter materials, we emulated a long-term neural stimulation experiment by enabling the OLED for repeated cycles in which the OLEDs are on for five seconds (using 10 kHz pulsed operation) and subsequently off for five seconds. Over a 12-hour-long stress test, we find no significant degradation (Fig. 2j). The observed variance is instead associated with fluctuations (<±5%) in the power supply and ambient temperature variations.

To characterize pixel-to-pixel variance, we plot the distribution of average optical power density generated by each OLED under 7 V pulsed operation (Fig. 2k). For the orange OLEDs, the ensemble is approximately normally distributed with mean and standard deviation of 0.1 and 0.025 mW/mm², respectively. Similar variability is observed for the blue OLEDs (Supplementary Section S6). We attribute the observed intensity variations primarily to the inhomogeneity of the aluminum anode contacts (Supplementary Section S3), which is negatively impacted by some of the shank post-processing that precedes OLED deposition.
7.4. Large-FOV, localized optogenetic stimulation in-vivo

To enable electrical recording during optogenetic stimulation, we laminate the OLED-on-CMOS probes on insertion with an optically transparent, mechanically flexible multielectrode array (MEA), as shown in Fig. 3a (see Methods and Supplementary Section S7). The noise power spectral density in vivo in the presence of OLEDs pulsing at 10 kHz shows only a minor increase in the integrated noise from 11.9 to 12.8 µV rms (over the 0.15 Hz to 6 kHz band) (Fig. 3b).

In all of our experiments, a 3 mm × 3 mm cranial window was created over the somatosensory cortex (SC) of a head-fixed anesthetized mouse using a titanium head-plate; the dura mater was removed to minimize insertion resistance. The recording probe was fabricated to have a through-hole located at the tip, whose outer dimension matches the tip of the OLED shank. During insertion, the OLED shank attaches to the flexible recording probe, and they are inserted together as a laminated structure aided by strong hydrophobic interactions. Post-implantation images demonstrate successful alignment with ±8 µm precision (Supplementary Section S8). Because the passive electrodes only cover half of the OLED shank, the shanks were implanted to a depth of 1.3 mm providing an interface with 128 OLEDs and 32 electrodes. In all cases, we operate with 10-kHz pulses, 50 % duty cycle, and with a 7 V bias.

We first experiment with transgenic mice expressing ChR2 over the entire cortex. In this case, to reliably stimulate single neurons with blue OLEDs, we turn on eight OLEDs synchronously, starting from the brain surface and gradually moving towards greater depth. We keep the center of stimulation over a recording electrode, turning on four OLEDs and turning off four to obtain a moving stimulation window. Stimulation is applied for 100 ms at a 3.3 Hz repetition rate for a total of 100 “on” envelope pulses before activating the next group in the moving stimulation window (Supplementary Section S9). In the experiments, we observe a spike
success rate of over 60%. Therefore, we choose a minimum of 100 pulses to gather a statistically significant number of spikes.

We extract spike clusters at each electrode and detect the ones that show an increase at the time of OLED stimulation. Fig. 3c shows three separate observed spike clusters that exhibit a major increase in spiking frequency during stimulation for at least 50 separate cycles. Two of them are detected around 150 µm below the surface in Layer 2 and show characteristic waveforms and autocorrelograms of interneurons. The third spike cluster occurs at depths associated with Layer 5 neurons. The fast firing rates and waveform structure of this third detected neuron exhibits the characteristics of a pyramidal neuron. In every case, spike waveforms recorded by adjacent electrodes show strong full attenuation at distances beyond 100 µm. Control experiments performed with wild type (WT) mice without opsin expression showed a homogeneous spike distribution before and after stimulation onset (Supplementary Section S10).

We observe a significant increase in synchronized spikes exclusively from the neuron located close to their respective OLED block. (Fig. 3d) The fast OLED rise and fall times create the opportunity to study the spike latency, the time from the turn-on of the pulse envelope to the peak voltage of the spike. We plotted the histogram of responsive spikes synchronized to the applied stimulation pulse envelopes. For ChR2, we observe a wide range of delays for the observed spikes after the stimulation onset. These variations are consistent with the dynamics of ChR2 at the relatively low illumination power achieved here, on the order of 0.1 mW/mm². At these power densities, photon shot noise and ChR2 latency variation contribute significantly to spiking temporal uncertainty.
Figure 7.3: Large scale, hyper-localized optogenetic stimulation in vivo. (a) Photomicrograph of neural insertion of blue OLED shank. Co-implantation of the flexible multi electrode array and rigid blue OLED shank is illustrated. The inset shows the multi electrode array with 32 transparent electrodes at a 49 µm pitch. (b) The noise spectra of the electrodes before and during blue OLED activation as a function of frequency. (c) Induced single-unit activity during OLED operation. Left side illustrates the gliding illumination window achieved by sequential activation of blue OLED groups from surface to deeper layers. Spike waveforms acquired from electrode sites 3 and 4 show synchronized superficial layer 1 single-unit interneuron activity (green and cyan circles). In proximity to electrode 10 in layer 2/3, an excitatory pyramidal neuron (red triangle) is activated synchronously to the blue OLED pulsing. Detailed waveforms are averaged over 50 spiking events. Corresponding autocorrelograms of each spike is plotted on the bottom part. (d) Matrix of the firing histogram of each neuron during stimulation pulses applied by each blue OLED group. Dashed blue squares represent the applied stimulation pulses with 100 ms duration. Every pulse is applied for a total count of 100 times. For the corresponding OLED group activation, the responsivity of neuron 1, neuron 2 and neuron 3 were 63%, 61% and 88% respectively.

7.5. Single-neuron, single-OLED stimulation with high temporal resolution

To further validate the ultra-fast switching speed of our OLEDs, we change the opsin model from ChR2 to red-shifted ChRmine, which possesses rapid on-kinetics suitable for millisecond-scale control of neural activity.\(^\text{101}\) In addition, optical modelling indicates that the enhanced sensitivity of ChRmine should allow us to activate neurons with illumination from a single orange
OLED (Fig. 2g). To test this, we co-insert the orange OLED shank and recording shanks at a depth of 1.5 mm and activate single OLEDs with 10-ms-long envelope pulses at a repetition rate of 2.5 Hz. Two hundred pulses are applied before moving the excitation window. Fig. 4a shows a representative result of one of these experiments in which a pyramidal neuron, adjacent to the emitting OLED, is observed to fire synchronous to the stimulation pulses. Additional examples with similar kinetics are provided in Supplementary Section S11. Over 70% of the corresponding spike distribution (Fig. 4b) is centered within the first 2 ms after the onset of stimulation, band limited by the rise and fall times of the opsin and reflecting the faster kinetics of ChRmine compared to ChR2. We apply no stimulation pulses for five minutes to let activity return to baseline, after which we repeat another two hundred stimulation pulses. Fig. 4c shows the spike train over 15 minutes, demonstrating repeatable local spike induction for each successive pulse train.

Figure 7.4: Single orange OLED stimulating a single neuron with ChRmine. (a) Spike waveform of the stimulated pyramidal neuron as a function of electrodes placed with 49 µm pitch. (b) Spiking histogram of the corresponding neuron averaged over 200 pulses with 10 ms width. (c) Illustration of the firing rate view of the corresponding neuron during a
900 seconds window. Three separate trains of 200 orange OLED stimulation pulses at 2.5Hz (illustrated on top part) are applied for a duration of 80 seconds each.

7.6. Modulation of local field potentials in-vivo

The ability to controllably expand the emission field provides the opportunity for stimulating larger neural populations and modulating local field potentials. Here, we seek to create a tunable activation window to access groups of neurons selectively by employing simultaneous stimulation from multiple OLEDs and tuning their emission power.

To validate the ability to create reliable and repetitive global stimulation, we first program 128 contiguous OLEDs to apply stimulation cycles consisting of five seconds of illumination followed by 30 seconds of rest (Fig. 5a). Such illumination timing proved an optimal scheme for reproducible LFP induction, in which the stimulation window is long enough for further analysis of frequency components of the induced activity and delay time is optimal for the excited activity to stabilize to its spontaneous phase post stimulation (Supplementary Section S12). We record from the flexible, transparent electrodes and establish the periodograms using a Gabor wavelet transform (see Online Methods). We observe a highly reproducible, three-fold increase in the corresponding activity, primarily in the 1-25 Hz band, with stimulation that indicates large scale modulation of neural activity. To confirm that these effects are due to optical stimulation and not heating from the device, control experiments are performed in mice without opsin expression. The corresponding LFP activity does not show any change during OLED activation (Supplementary Section S13).

While all the results presented thus far have been for maximum applied voltage of 7 V, the addressable stimulation depth can also be adjusted by tuning the applied OLED anode potential from 5 to 7V (Fig. 5b). When illuminating synchronously from 16 contiguous OLEDs, these
voltages correspond to total emitted optical power of 0.15 µW, 0.35 µW, and 0.65 µW, respectively, which is in direct proportion with the effective field of stimulation (Supplementary Section S14). To be able to precisely validate the controlled LFP modulation, we ensure the brain is in the same state of sleep prior to application of stimulation pulses and that the spectral domains acquired by computing power spectra of the signals overlap. All pre-stimulation spectra exhibit a peak in the 1-2 Hz range, a marker of slow-wave sleep-like activity obtained under deep anesthesia²⁴², and a 7-10 Hz peak characteristic of the somatosensory cortex in this state of anesthesia. Modulation with 0.65 µW enhances the power spectra between 10 to 40 Hz, largely in the beta band. Increasing the emission power levels to 1.6 and 3 µW, respectively, enhances LFP modulation over a larger frequency range of 0-40 Hz with power spectra that varies in proportion to the incident optical power density.

For more focused induction of LFPs, we instead apply stimulation pulses only to the top and bottom 16 OLEDs of a 128 OLED group that is integrated with 16 electrodes spaced at 100 µm pitch. Fig. 5c shows the recording from two separate experiments with mice that were injected at 300 µm and 1.2 mm depth. Induced activity, observable as far as 1.5 mm away from the point of stimulation, is primarily concentrated around the active OLEDs. Corresponding LFP power analysis shows a major decay as a function of distance towards distant electrodes (1.44 mm) which is consistent with the optical stimulation being the input source. These results show the light emission field can be extended both laterally by expanding the number of OLEDs and axially by increasing the emitted optical power.
Figure 7.5: Local field potential (LFP) modulation in vivo. (a) LFP modulations in the delta, theta and alpha bands are reproducibly induced by simultaneous activation of 128 orange OLEDs in ChRmine injected mice. Square pulses represent the time of applied stimulation. Corresponding time-frequency colormap is placed on bottom left. LFP power spectra as a function of frequency before (black) and during (red) stimulation within 0-50 Hz range is plotted on bottom right. (b) Induced LFP modulations as a function of applied power emission for 650 nW, 1600 nW and 3000 nW is plotted respectively. Square pulses represent the time of applied stimulation. Bottom row illustrates corresponding LFP power spectra as a function of frequency before (black) and during (red) stimulation within 0-50 Hz range. (c) Figure illustrates LFP modulation fallout as a function of distance from the active OLED groups over a distance of 1500 µm. Corresponding power spectrums are located at bottom showing the LFP power at top electrode (black) and bottom electrode (red).

7.7. Discussion

The OLED-on-CMOS design presented here is part of a new class of implantable optoelectronic devices that use CMOS technology to create volumetrically efficient implantables for optical brain interfaces. An important need in these devices, both for optogenetic stimulation and for imaging,
is to shape the illumination field spatially and temporally. OLEDs bring substantial fabrication advantages and unlike other LED choices allow emitter density (and pixel pitch) to be pushed to neuron-scale densities. In fact, due to the low horizontal conductivity of the OLED material, the size of each OLED and the OLED fill factor on the device are only determined by the dimensions and arrangements of the anode contacts on the CMOS platform. Due to the active CMOS control, our completed devices can be set to create arbitrary illumination patterns, within the limits of the lateral resolution of the anode contacts (i.e., 24.5 μm in this study). Using this flexibility, we demonstrated the ability to stimulate neuron populations at various temporal and spatial scales using different patterns of structured illumination, from individual neurons with single OLEDs to large populations of neurons with synchronized excitation of larger OLED groups. While in earlier generation devices, LEDs were often separated from their driver by significant interconnect networks, in our case the CMOS drivers are immediately beneath the OLEDs and tuned to drive the OLED capacitance. The resulting fast rise and fall times are well-suited for high-frequency modulation of neuronal activity with opsins offering sub-millisecond kinetics, such as Chronos or ChroME2f.

One of the key advantages of optogenetic stimulation is the possibility of bidirectional stimulation, i.e., activation and inactivation of neuronal activity. Here, we have demonstrated two different OLED colors, blue and orange, that overlap with the characteristics of ChR2 and ChRmine. In future applications, multiple colors can be integrated on the same shank by optimizing the masking process. Very high resolution (<5 μm) shadow-masking methods can be leveraged to pattern different colors on the pixel level. This approach would allow bidirectional stimulation of opsins, such as SOUL, with non-overlapping activation and inhibition at orange and blue wavelengths, respectively.
In some of our experiments, we pulsed small groups of OLEDs, running the active pixels successively along the entire length of the shank, which resulted in three hour long experiments. In addition, we performed multiple implantations of the same shank in separate animals. In aggregate, shanks were operated and remained functional in vivo for at least 24 hours.

In the current study, we used a second laminated probe to validate the stimulation capabilities of the OLED shank, but in the future these electrodes and the associated recording electronics can be integrated onto the CMOS shank itself. In order to further increase the addressable volume, the OLED brightness can be further improved through improvements to the planarity and surface resistance of the foundry provided aluminum pad. Relative to the OLEDs-on-CMOS reported here, OLEDs with optimized electrode contacts deposited on glass or passive silicon substrates have been shown to reach up to 10-fold higher brightness levels at the same drive voltage. Finally, while we performed die-level OLED-on-CMOS fabrication here, the process can be scaled to be wafer-level if the thermal budgets during dicing and postprocessing can be managed to maintain OLED performance. Wafer-level processing of OLEDs on CMOS has been demonstrated for microdisplays, albeit with lower-brightness OLEDs.

In summary, we presented a shank-form-factor CMOS platform containing 1024 emitter channels for localized optogenetic stimulation in the deep brain. The spatiotemporal stimulation resolution of our device is comparable to advanced recording modalities, offering single-neuron-resolution targeted stimulation. Along with its tunable field of stimulation and large-scale integration, OLED-on-CMOS technology provides a new tool for the development of optical brain-machine interfaces.
8. Conclusions

8.1. Summary of results and contributions

The thesis presents a body of work contributing to the state-of-the-art across multiple disciplines. The core of the work is the OptoSAM, a CMOS chip that contains both 1024 SPAD pixels for fluorescent imaging, in addition to 1024 OLED drivers for optogenetic stimulation. Beyond the chip design as presented in chapter 3, this thesis implemented four prerequisites to enable successful fluorescent imaging and optogenetic stimulation. First, chapter 4 presented the MIMAS frontend that supplies the necessary filters and image reconstruction for decoding of multicolor fluorescent targets. Second, chapter 5 presented DISTINCT, a novel ultrafast algorithm that incorporates illumination structure to strongly improve imaging resolution. Third, to cross-validate the effects of optogenetic activation using artifact-free electrophysiological recording, chapter 6 presented a fully organic vertically integrated pixel consisting of an OLED and a transparent recording electrode. Fourth, chapter 7 presented the integration of OLEDs on the OptoSAM chip for hyperlocal and scalable optogenetic stimulation in mice in-vivo. Together, these contributions have produced the following papers:


*Contributed equally

Changhyuk Lee, Adriaan J. Taal, Jaebin Choi, Kukjoo Kim, Kevin Tien, Laurent Moreaux, Michael L. Roukes, Kenneth L. Shepard, “A 512-Pixel 3kHz-Frame-Rate Dual-Shank Lensless
8.2. Review of specifications

In order to assess whether the results match the desired performance, we need to review them with the initial specifications. This allows us to open discussion regarding avenues for future improvements. Table 5 compares the quantitative results achieved throughout chapters 3-7, versus the desired specifications postulated in chapter 2. The table references the corresponding chapter where the result was first achieved.

Table 5: Achieved results versus target specifications for the OptoSAM system.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Target</th>
<th>Result (chapter)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shank length</td>
<td>6mm</td>
<td>6mm (ch. 3)</td>
<td>800 μm inactive tip, 3.5 mm of active devices, 1.7mm of inactive base</td>
</tr>
<tr>
<td>Shank width</td>
<td>100 μm</td>
<td>120 μm (ch. 3)</td>
<td>100 μm active, 10 μm silicon margin for fabrication</td>
</tr>
<tr>
<td>Shank thickness</td>
<td>50 μm</td>
<td>75 μm (ch. 3)</td>
<td>55 μm silicon, 10μm oxide, 10 μm OLED passivation</td>
</tr>
<tr>
<td>Technology node</td>
<td>TSMC 130nm BCD</td>
<td>TSMC 130nm BCD</td>
<td>Requires special Fab 12 cybershuttle for SPAD fabrication</td>
</tr>
<tr>
<td>Pixel pitch</td>
<td>20 μm</td>
<td>25 μm (ch. 3)</td>
<td>This pitch size allows in-pixel counter, addressing and sufficient decoupling capacitor for the OLED drivers</td>
</tr>
<tr>
<td>Target fluorescent protein (excitation, emission wavelength)</td>
<td>GFP / GCAMP / FITC (ex. 480nm, em. 515nm)</td>
<td>FITC (ch. 4)</td>
<td>Used FITC for interesting pH dependent lifetime properties, and availability of ultrabright, 45μm diameter beads</td>
</tr>
<tr>
<td>Framerate</td>
<td>5 Kfps</td>
<td>51 Kfps theoretically (ch. 3)</td>
<td>51kfps for maximum illumination conditions and dominated by the readout overhead. Practically, 20 fps works well for moderate illumination conditions.</td>
</tr>
<tr>
<td>SPAD Repetition rate</td>
<td>20MHz</td>
<td>80 MHz (ch. 3)</td>
<td>Tested with 80MHz Ti:Sapphire laser</td>
</tr>
<tr>
<td>Temporal resolution</td>
<td>200ps</td>
<td>138ps (ch. 3)</td>
<td>FPGA PLL achieves a 1-degree phase resolution on a 50ns clock.</td>
</tr>
<tr>
<td><strong>Fluorescent excitation source</strong></td>
<td>Mode-locked laser</td>
<td>NKT EXU6-PP for variable spectrum and repetition rate (ch. 3)</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------</td>
<td>-------------------------------------------------</td>
<td>---</td>
</tr>
<tr>
<td><strong>Excitation source rejection</strong></td>
<td>OD 8</td>
<td>OD 4 (ch. 4)</td>
<td>~OD 3 from the MIMAS and LP500 integrated filters. ~OD 1 from time-gating.</td>
</tr>
<tr>
<td><strong>Spatial resolution</strong></td>
<td>20µm</td>
<td>50µm (ch. 4)</td>
<td>~50µm resolution over the FoV upto distances of ~150µm away</td>
</tr>
<tr>
<td><strong>Spectral sensitivity</strong></td>
<td>Contrast of OD 1</td>
<td>~OD 1 through MIMAS structure (ch. 4)</td>
<td>Novel optical frontend that provides orthogonal light fields for two wavelengths in the same detector</td>
</tr>
<tr>
<td><strong>Fluorescent imaging validation method</strong></td>
<td>1-photon microscope</td>
<td>Brain slice ground truth images (ch. 4)</td>
<td>Used for MIMAS verification</td>
</tr>
<tr>
<td><strong>Reconstruction algorithm</strong></td>
<td>Incorporate illumination and/or lifetime a priori knowledge</td>
<td>DISTINCT algorithm (ch. 5)</td>
<td>Improves the imaging quality when illuminating with $K$ illumination patterns. Faster and more accurate than existing algorithms that only use the sparsity on the detector side.</td>
</tr>
<tr>
<td><strong>Photodetector type</strong></td>
<td>Single Photon Avalanche Diode</td>
<td>Improved upon TSMC SPAD designs (ch. 4)</td>
<td>Improved fill factor by $4\times$, and selected SPADs with lowest IRF compared to the first tapeout$^{10}$</td>
</tr>
<tr>
<td><strong>Spatial modulation</strong></td>
<td>Angle-sensitive pixels</td>
<td>MIMAS Structure (ch. 4)</td>
<td>Improved the ASP structure. Expanded the resulting image quality versus geometry analysis for future design tradeoffs.</td>
</tr>
<tr>
<td><strong>Optogenetic emitter type</strong></td>
<td>Organic Light Emitting Diode</td>
<td>OLEDs in red (615nm) and blue (490nm) (ch. 7)</td>
<td>Designed and optimized by Malte Gather group.</td>
</tr>
<tr>
<td><strong>OLED emission power</strong></td>
<td>1mW/mm$^2$</td>
<td>0.1mW/mm$^2$ (ch. 7)</td>
<td>0.1mW/mm$^2$ from the OLED surface translates to 0.05mW/mm$^2$ on a neuron 50µm away. That power density induces spikes in ChRmine-stained neurons with a probability of 100%.</td>
</tr>
<tr>
<td><strong>Optogenetic protein (activation wavelength)</strong></td>
<td>ChRmine (590nm)</td>
<td>Successful expression via pAAV injections (ch. 7)</td>
<td></td>
</tr>
<tr>
<td><strong>ChR2 (470nm)</strong></td>
<td></td>
<td>Successful expression via transgenic mouse line (ch. 7)</td>
<td></td>
</tr>
<tr>
<td><strong>SOUL (470nm enable, 580nm disable)</strong></td>
<td>Unsuccessful</td>
<td>We tried two batches of injections, 2 mice each. In neither batch did we observe any optical response.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLED repetition rate</td>
<td>&gt;6MHz (ch. 7)</td>
<td>Peak optical emission at 10 KHz</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------</td>
<td>---------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Optogenetic validation</td>
<td>20KHz</td>
<td>Artifact free spike detection (ch. 6)</td>
<td>No recording artifacts from the digital CMOS registered within the 0.5 Hz &lt; B &lt; 5KHz bandwidth</td>
</tr>
<tr>
<td>method</td>
<td>&gt;6MHz (ch. 7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum tissue heating</td>
<td>1C, or 100mW dissipation from 10x15μm area</td>
<td>Assumed negligible (ch. 7)</td>
<td>We lacked the dedicated thermal microscopy tools to measure. Our OLEDs dissipate a maximum of 210μW from a 20x20 μm area.</td>
</tr>
</tbody>
</table>

### 8.3. Discussion of results

Here, we discuss the results versus the desired specifications. For each subsection, it describes which exact shortcomings in the specifications relate to limitations in performance. The goal of this section is to highlight what exact improvements need to be made to improve future experimental results. Some background work is introduced as ideas for improvement.

#### 8.3.1. Signal yield and addressable volume

The optical paradigm promised a deeper field of view (200 μm) versus the electrical recording and stimulation paradigm (50 μm). We achieved an imaging distance of 150 μm on the detector side by manipulating the light field using the MIMAS frontend.

In order to address these volumes with the OLED emitters, we switched to the ultra-sensitive red-shifted channelrhodopsin ChRmine. We estimated from Monte Carlo simulations that a 200 nW OLED can overcome the 0.05mW·mm⁻² activation threshold at a distance of 150 μm. Scattering and absorption measurements of OLED light in brain slices confirmed this hypothesis.

The OLED devices remained too dim to activate single ChR2 neurons beyond distances of 30 μm. A fivefold increase in brightness is needed to achieve the ChR2 threshold of 0.5mW·mm⁻² at a distance of 100 μm from the shank. The OLEDs also require spatial emission patterns if
single neuron precision is to be preserved at larger distances. These patterns could be achieved in two ways. First, the aluminum anode contact could be patterned into a grating that translates into the OLED stack an passivation layers, generating a waveguide structure that only emits at angles, by diffracting out the transverse mode and suppressing the vertical mode\textsuperscript{219}. Secondly, a photonic crystal can be placed on top of the OLED to diffract the emitted light in specified directions\textsuperscript{250}. The photonic crystal acts as a Bragg reflector for light travelling at certain angles, thus allows light propagation only in specific directions. Similar to a Bragg reflector, the more crystal periods the light travels through, the better the directionality. Photonic crystals need therefore be maximally extended beyond the OLED emitting region. The light diffracts into the medium in a structured fashion. The resulting far field\textsuperscript{251,252} is computed in FDTD-solutions. Fig. 8.1a shows a possible quadratic photonic crystal layout, etched into the OLED passivation layer. Fig. 8.1b shows the
sideview. The air holes (n=1) are a periodic refractive index contrast with the passivation layer (n=1.9).

The photonic crystal light molding characteristic strongly depends on the hole pitch, hole radius, refractive indices, and hole depth. Holes that extend all the way towards the OLED stack generate stronger directionality, at the cost of a weaker passivation that may fail sooner.

Fig. 8.1c-e shows examples how far field emission profile strongly depends on pitch alone. For p=150nm and p=200nm, the periodic grating only allows light propagation in horizontal and vertical directions. With pitch increased to p=250nm, light can only travel along the diagonals. The photonic crystal 90° rotational symmetry results in the same symmetry in the far field projection.

8.3.2. Frame rates, integration times and optical filtering

The SPAD imager achieved a 51kfps when synchronized to a mode-locked laser with 20MHz repetition rate. Such frame rates assume a registered photon for each laser repetition, filling up the 6-bit in-pixel counter within 3.2 µs under global shutter operation. Readout overhead of 16 µs brings the maximum framerate to 51kfps. Implementation of a rolling shutter can improve the ratio of integration versus readout time.

In our experiments however, the maximum useable framerate was limited to 20fps. The SPADs registered approximately 1 photon per 1000 laser repetitions, as a result of the picowatt-level fluorescent signal, the lack of focusing optics, and photon loss due to optical filtering and limited PDP. Typical photon counts were on the order of 10,000-100,000 per second. Three types of noise were present: signal shot noise, background shot noise, and dark count. First, dark count rate was negligible at around 100 counts∙s⁻¹, rendering these SPADs therefore very suitable to these very
low photon flux environments. Background shot noise was very high compared to signal levels (Fig. 3.8 in chapter 3). While the fluorescent signal was emitted from dim microspheres, the excitation background was uniformly injected into the tissue at mW-level powers. The longpass filter presented in chapter 4 was sufficient to suppress the excitation background below the signal level of 45 µm Fluoresbrite beads. However, the signal from smaller, dimmer fluorescent sources such as neurons stained with GCaMP6s will be completely obscured by excitation background counts.

Long integration times are required to accumulate enough photons in order to reduce the relative variance of the shot noise. Chapters 3 and 4 described how the resulting SNR increases with the square root of the integration time. When the background shot noise variance is low compared to the signal and background magnitude, we can reliably subtract the background from the signal. This method assumes the imager operates linearly within its dynamic range. However, exact knowledge of the background counts in the absence of fluorescent targets is impossible. The shank insertion changes the tissue morphology so dramatically that the background scatters in wildly different ways from experiment to experiment. Depending on the application, such as in Fig. 6 of chapter 4, using a simple recording in the absence of any biological sample may give some useful information on the structure of the background counts. The only method to reliably remove background counts is therefore through optical filtering. The demonstrated frame rate of 20fps is sufficient to image GCaMP6s dynamics. To maintain that framerate for imaging dimmer fluorescent neurons, the excitation laser power needs to increase, in turn requiring a better excitation filter. We estimate that GCaMP6s imaging in-vivo requires an excitation filter of at least OD 8 without further sacrificing signal photon yield. The excitation filter needs to be at least OD 10 to image voltage indicators at a 2Kfps framerate.
8.3.3. Time-gating

The circuits for time-gating of the SPADs were introduced in chapter 3. Initially, we designed the time-gating to remove excitation light with OD 6. Unfortunately, only OD 2 was achieved. Electron-hole pairs generated in the doped silicon by laser illumination while the SPAD was out of Geiger mode (14V bias) would generate an avalanche when switching the SPAD into Geiger mode (17V bias). The lifetime of these generated carriers determines the instrument response function (IRF).

This method however only works when the fluorescent lifetime is significantly longer than both the SPAD IRF, the illumination source IRF, and any excitation filter autofluorescence. The illumination source IRF is on the order of picoseconds for a mode-locked laser as used in chapters 3 and 4. LED sources on the other hand cannot be used for time-gated measurements. The junction resistance and diffusion capacitance of free carriers in the diode creates a slow turn-off dynamic with a decay time on the order of several nanoseconds, completely obscuring the fluorescent lifetime. Figure 8.2 shows the difference in IRF between an ultrafast laser and an LED, as measured by the LinoSPAD.87 Measurement of the mode-locked laser in Fig. 8.2a is dominated

---

**Figure 8.2**: Time-gating measurement using the LinoSPAD from (a) the NKT EXU6-pp mode locked laser (b) the CREE GaN LED
by the LinoSPAD IRF. After 1 ns, only 1% of photon counts are registered. The same measurement performed for the CREE LED (Fig. 8.2b) shows that a delay of 3 ns is required to reduce the photon counts to 1%.

Instead of relying on time-gating to reject excitation light, in chapter 4 it was shown that time-gating can be used for lifetime estimation and demixing. The lifetime is extracted from the photon count versus time-gate delay histogram. To assemble this temporal histogram, the image needs to be repeatedly recorded for the number of time-gate delays, diminishing the effective frame rate. This method is a temporal oversampling and requires a static image throughout the histogram acquisition. It is recommended to prioritize measurement of the shorter delays where the SNR is highest. Recorded photon counts at delays longer than ~8 ns have very low SNR and can provide false information to the exponential fitting algorithm.

For dynamic imaging where a high frame rate is crucial, assembling the histogram for exponential fitting is not feasible. Instead, a constant time-gating delay should be selected that yields the best fluorescent signal yield to excitation background, around 1-2 ns after the laser pulse for fluorescent markers with a lifetime of ~4ns such as GFP and GCaMP.

8.3.4. SPAD optimization

The collaboration with TSMC provided resources to optimize the SPAD performance in a design of experiments (DOE). We constructed 9 different layer types with variations on implant types, implant depths, and guard ring doping and widths. The SPAD performance was evaluated in terms of PDP, IRF and dark count, using the same methods as in Fig. 3.5 and 3.7 of chapter 3. Side views of the layers are drawn in Fig. 8.3a-d for the four best performing SPADs, denoted by T1-
4. SPAD architectures T5-8 use the same layers as T1-4 respectively. T5-8 are circular and T1-4 are chamfered squares as seen from above. T6 is the same SPAD as characterized in chapter 3. In order to maximize time-gating performance, we evaluated the SPAD IRF in Fig. 8.3e. Surprisingly, the original SPAD already provided the best IRF, with second-best performance being SPAD T4. Other SPADs have multiple, delayed peaks in the IRF. Such peaks can arise from carrier generation in multiple layers simultaneously, each with different lifetimes and potential to create avalanches at separate delays. This leads to undesired effects in the fluorescent lifetime measurement and estimation.

Ultimately, the chamfered square is preferred for two reasons. First, due to reduced dead area it has an active surface area of 171 µm², versus 38 µm² for a circular SPAD, at the same 24.5 µm pitch. Secondly, fill factor can be increased further by connecting the buried N-doped layer and substrate connections as in Fig. 8.3f. The T4 SPAD is selected for implementation on the second-generation chip as presented in chapter 4 and chapter 7. Future SPAD optimization can further improve fill factor, photon sensitivity, and IRF.
8.3.5. Imaging resolution

The resulting imaging resolution is a function of three parameters: detector pixel pitch, detector mask feature size, and array-wise mask layout. Chapter 4, Fig. 4.3e showed how the detector mask is fine-tuned to provide a spatial resolution of approximately 50 µm. The angle sensitive pixels were tuned to provide this resolution over the 200×200×3200 µm field of view. The number of fluorescent targets needs to be constrained well under the number of pixels \((M=1024)\) to guarantee this resolution. Within the scattering medium of neural tissue, the resolution degraded to 100 µm. Higher pixel count and lower pixel pitch provide better resolution only when they are equipped with finer grating pitches and thus provide higher spatial frequency.
information. If denser pixels are not equipped with maximally diverse frontends, the signal recorded from two neighboring pixels will be highly correlated.

When we add sparsity on the illumination side, emitter pitch and emitter mask geometry become new parameters, independent of the detection side. This synergy is explored in chapter 5 and provided a fast, accurate algorithm for image reconstruction that optimally uses detector-side compression and emitter-side oversampling. An optimal case was achieved when the detector matrix $D$ is constituted from global, smooth functions such as a discrete cosine transform. An example implementation is the improved angle sensitive pixel (MIMASP) from chapter 4. The emitter matrix $E$ then provides sharp, local features as a Haar wavelet matrix. Example implementations are the photonic crystal of Fig. 8.1, or the MURA mask$^{50}$.

This combination results in a Hessian matrix $(H = D^T D \circ E^T E)$ closely resembling an identity matrix, albeit with some correlated rows. The rank of the Hessian is $\text{rank}(H) < M \cdot K$, with $K$ the number of illumination fields. Theoretically, the number of simultaneously resolvable fluorescent targets would scale linearly with $K$. With such a well-conditioned Hessian, the solver also converges faster, decreasing image reconstruction time.

An improved spatial resolution helps discerning dense groups of neurons. But higher framerates also contribute. In this thesis, only static imaging was demonstrated. Signal yields were too low to attempt functional imaging of GCaMP6s or GEVIs. Temporal oversampling above the Nyquist rate will help separating two neurons that are situated closer than the spatial resolution, but firing at slightly delayed times. In chapter 7 we demonstrated electrophysiological spike sorting, based on a 10× temporal oversampling. This technique allowed to separate spikes that occur too close for them to be originating from a single neuron.
The integrated optical imager is a platform that provides both scalable imaging resolution and scalable framerate. This combination will ultimately enable the separability of a larger number of neurons than pixels.

### 8.3.6. Infrastructure for neuroscience experiments

The OpoSAM chip demonstrated highly localized stimulation of single neurons in chapter 7. At a 25µm pitch, the emitters have a pitch size similar to the neurons themselves. Theoretically, the probe therefore provides the user with independent control of 1024 neurons. In reality, this control is only valid for neurons which are not only expressing the channelrhodopsin but are also free of post-insertion damage. The electrophysiology array provides the knowledge of successful optogenetic activation. Once the spiking neurons have been identified and spatially mapped, the user can focus efforts to modulate those viable neurons.

However, the current state of the art lacks real-time spike detection for feedback to the user. During experimentation, we successively enabled all possible clusters of OLEDs. Yet, only spike sorting revealed which OLEDs were successfully modulating neurons. Significant time is thus wasted activating OLEDs not in the vicinity of healthy, tagged neurons. Determining what OLED channels to activate is called the selection problem. The selection problem increases exponentially with channel count, and is a nonconvex optimization problem requiring real-time adaptive solvers. To fully exploit the potential of these very high-density stimulation probes, we identify the need for a real-time spike sorting and feedback system to guide the user to focus efforts.
8.3.7. OLED improvement

In chapter 7, we demonstrated how the OLED-on-CMOS is an effective tool for single-neuron resolution optogenetic activation. However, in our experiments we did not achieve independent activation of 1024 neurons. This is partly caused by the sparse expression and neural death, and partly caused by brightness and yield issues of the OLEDs.

Further brightness and yield optimization is necessary to increase success of localized stimulation. By optimizing the fabrication processes we recover the power density lost when migrating the ideal OLED fabrication onto the CMOS. Ideal OLEDs are an order of magnitude brighter than those characterized on the CMOS, bringing the emitted power density for a single OLED well above the ChR2 stimulation threshold. Lower resistance will have a secondary advantage of reducing thermal heating, and thus brightness degradation, due to ohmic losses.

Brightness and yield are reduced by three factors: oxidation of the aluminum anode, contamination of anodes with residues, and anode surface roughness.

First, reducing temperatures for wirebonding and epoxy curing to 100 °C showed to decrease oxide formation. Second, we found that processing with deionized water (DI-H2O) instead of organic solvents reduced the amount of residue. Extra aluminum oxidation formed in water is removed under an etch step before OLED evaporation. Third, the anode surface could be planarized by chemical mechanical polishing. The copper and aluminum top layers (respectively 3 µm and 1.5 µm thick) are not optimized for surface roughness by TSMC.

8.3.8. Full integration of light delivery, detection and optogenetics

We found the CMOS shank to be an effective platform for multiple types of optical experiments. The integration of multiple modalities and measurement capabilities has merit for
two reasons. First, it provides flexibility to perform novel combinations of measurements to drive the field of biosensing forward. An example is simultaneous 3D localization and non-contact pH sensing using the lifetime dependence. Another example is static localization with demixing of wavelengths based on the lifetime. Secondly, it also allows for cross-validation of each experimental measurement technique. A multifunctional SPAD and electrophysiology probe can compare the two datasets and draw stronger conclusions from the commonalities and differences in the datasets.

8.4. Future engineering proposals

Based on the merits offered by the integration of multiple modalities on a single platform, here we propose three classes of optical shank systems. The device-, circuit- and fabrication-level recommendations are included where relevant.

8.4.1. Proposed System 1: Integration of OLEDs and electrophysiology on CMOS

This system, drawn in Fig. 8.4, will be a combination of the OptoSAM and the Neuropixels\textsuperscript{23}. It will dramatically increase the number of electrophysiology recording sites compared to our passive co-inserted flexible MEA (chapter 7). While the local density will be lower than the vertically stacked OLED-electrodes (chapter 6), the photon transfer efficiency from OLED to neuron will increase due to fewer absorptive and reflective layers. The current pixel pitch of 25 µm does not need to change. For these unfocused light emission and electronic detection, pixels smaller than the neurons themselves will impose more system complexity and overhead, while providing an insignificant increase in information. Required emission brightness and electrode SNR will
determine the necessary fill factor. Rather, maintaining a large field of view is more useful for experiments spanning across multiple functional brain areas.

There are three challenges. First, we found electromagnetic OLED digital switching noise coupling into the external electrodes (chapter 7). Since the flexible MEA was separated by 10 µm of passivation, we expect noise to couple stronger when tightly integrated into the same CMOS shank. Proper shielding of the analog power and signal lines is eminent to preserve signal quality. The system will require three grounds: a digital OLED ground, an analog biasing and amplifier ground for the electrophysiology, and a shielding ground separating the two. In-pixel preamplification as opposed to a unity buffered multiplexing can help to drastically reduce noise contributions along the shank.
Second, the OLEDs rely on passivation layers to prevent oxidative degradation, whereas the electrodes Conversely need a low impedance interface to the tissue. This conflict brings a fabrication challenge of locally opening the passivation above the electrodes while not compromising OLED passivation. This challenge is exacerbated by the 80 °C temperature constraint to prevent OLED degradation.

Thirdly, the OLEDs are evaporated onto the shank through a shadow mask with a practical resolution of 50 µm. These organic layers need to be removed from the electrode contact for proper low impedance contact with the neural medium.

8.4.2. Proposed System 2: Integration of excitation light source on CMOS

Integrating the excitation light sources onto the shank is the only method for achieving a constant light field across a large field of view inside the deep brain. The CMOS platform defines the microscale contacts while the µLEDs themselves need be post-fabricated on top. µLEDs might be directly fabricated in future CMOS processes256, their emission wavelengths defined by doping conditions.

As the emitters are placed on the shank itself (Fig. 8.5), very low emission powers are needed to excite neurons close to the shank. This generates some degree of locality as the weak fluorescent signals can only be registered by nearby photodetectors. LED photon transfer efficiency through scattering and absorptive tissue is approximately 20% at a distance of 50µm, as shown in chapter 6. However, fluorescent dye return efficiency is much
lower as the fluorescence radiates isotropically, as opposed to the narrow top-emitting LED light.

A 15μm diameter SPAD harvests 3% of light from a 50μm diameter emission sphere. A roundtrip (including 50% fluorescent dye quantum efficiency) could harvest 0.2 x 0.03 x 0.5 = 3‰ of emitted photons as fluorescent signal. A minimum lower bound of OD 4 excitation rejection will then provide a theoretical idealized contrast of 30x. The brightness of neurons however will scale down quadratically with their radius. Practical situations will rather require excitation rejection ratios of OD 6 to OD 8, similar to filters found in fluorescent microscopes.

The very tight integration of emitters and detectors poses very hard challenges for excitation filtering. The currently used optical filters are between 4 μm (LP500 in SU-8) and 10μm (interference filters, ECI) thick. If the emitter to detector pitch is smaller, incomplete filtering occurs. Furthermore, the interference filter peak rejection wavelength is reduced with incident angle according to:
\[ \lambda(\theta) = \lambda(\theta = 0^\circ) \sqrt{1 - \frac{n_s^2 \sin^2(\theta)}{n_{\text{MIM}}^2}} \]

With \( n_s \) and \( n_{\text{MIM}} \) the refractive indices of the fluorescent source medium and the interference filter respectively. A 500nm longpass interference filter becomes transparent for 480nm excitation light that enters at an angle of 30°. Furthermore, when LED excitation light incidents at a 90° angle with respect to the filters, light could freely couple into the waveguide structure and diffract into the photodetectors. LEDs force the experiment to forgo lifetime estimation measurements and to operate the device in a continuous wave operation. In continuous wave operation, the integrated filter carriers the full burden of excitation light rejection.

Laser diodes on the other hand don’t suffer from diffusion capacitance, as the stimulated emission effect always depletes free excited carriers into emitted photons. As opposed to LEDs, lasers have no free carriers that need to be extracted in order to bring the forward voltage below the threshold voltage. Integrated microlaser diodes, especially Vertical Cavity Surface Emitting Lasers\(^{257} \) (VCSELs) are a promising new technology to deliver solutions to the problems posed above.

### 8.4.3. Proposed System 3: autonomous driver for freely-behaving mice

To open up avenues for chronic studies, we need a system that allows the mice to behave freely. Chronic implantation also raises the reliability requirements on the shank system. The high-level design is shown in Fig. 8.6. The rigid OLED shank from chapter 7 is flip-chip bonded to a polyimide flexible cable on the 2mm wide base. This 5cm long flex cable carries the 25 essential connections for addressing and pulsing the OLEDs, and connects to a 2x1cm PCB carried by the mouse on the back. This PCB contains a TI CC2640 low power microcontroller running the OLED
control state machine. The entire system is powered by 7V batteries and require charging every few days. A PC transfers the OLED patterns to the CC2640 via an RF motherboard. The electrophysiology recording electrodes will be hydrostatically laminated to the shank during co-insertion. This system can be integrated with the Intan record controller chip.

Experiments for chronically implanted OLEDs include long-term fear inhibition by activation and suppression of the amygdala\textsuperscript{258}, risk-reward modulation\textsuperscript{259}, study of starvation and hunger\textsuperscript{260}. The combined OLED and electrophysiology provides a true real-time feedback loop, with much higher channel count than previously published devices\textsuperscript{260}. Power consumption is kept low by minimizing the wireless data telemetry. The following engineering issues need to be addressed in order to succeed this project:

- We need to understand the emitted optical power requirements to modulate behavior.
- The OLED fabrication needs to be improved to meet these emission requirements.
• Know the current densities required to drive the OLEDs to those emission powers. Then find a battery that can supply these for the length of the experiments.

• Building a closed-loop seizure detection scheme in the C-code. A simple detection scheme can be based on the ARM FFT library, as seizures reveal themselves as LFPs with much higher frequencies than baseline neural activity.

• The current laser trenching fabrication method creates shanks with sidewalls too rough for chronic implantation. A more time consuming ICP etch trenching method creates sidewalls smooth enough to avoid long-term tissue damage. The effect of this fabrication on OLED anode quality is unknown.
8.5. Future biological applications

Using the current and future systems, I would raise the following experimental questions. Progressing with the OLED subsystem as presented in chapter 7, we can ask experimental questions concerning more elaborate single-cell stimulation experiments:

- How do stimulation parameters differ for different neuron types and functional regions? Are different neuron types characterized by different delays between OLED pulse and spike onset, which we can reveal using the ultra-fast pulsing?

- Can concurrent electrical stimulation induce neural hyperexcitability, lowering the stimulation threshold? Can neurons further from the shank be stimulated this way in a synchronized fashion, increasing the total addressable volume?

- Can we reliably transfect two different channelrhodopsins in the brain, and activate them independently with two different OLED colors? How is the spectral crosstalk for two neighboring cells?

- How long can these shanks be inserted before cell death disables optogenetic activation? How can this duration be improved?

Next, we can ask experimental questions regarding the synchronized stimulation of large areas:

- When using the step-function opsin called SOUL\textsuperscript{245}, how many pulses and power will it require to induce behavior? Will continuous activity from the SOUL-labeled neurons spill over to non-illuminated neurons, or even to non-labeled neurons?
• How specifically can we spatially confine the stimulation with the highly sensitive ChRmine? Can we cross-validate the stimulation with three-dimensional 2-photon imaging? What happens with connected neurons that aren’t illuminated?

Once we better understand the requirements of behavioral stimulation, very complicated experiments can be devised around the following research questions:

• When inserting the OLED shank to intersect multiple regions, how many areas can we reliably target simultaneously? When taking the paw/finger area, can we modulate the finger movement in precise direction, frequency? What is the locomotive “crosstalk” between neighboring fingers when stimulating a single finger? Can we find distinct neuron populations that move certain fingers and only those fingers? How large is this cross-talk in both space and time?

• When a seizure happens, what latencies are permitted to prevent the seizure from spreading throughout the brain? And is spatial control along a linear shank array sufficient, or does seizure inhibition require three-dimensional addressability?
References


87. Burri, S., Homulle, H., Bruschini, C. & Charbon, E. LinoSPAD: a time-resolved 256×1 CMOS SPAD line sensor system featuring 64 FPGA-based TDC channels running at up to 8.5 giga-events per second. in *Optical Sensing and Detection IV* vol. 9899 98990D (2016).


102. Bansal, H., Gupta, N. & Roy, S. Theoretical analysis of optogenetic spiking with ChRmine,


112. Koo, H. S., Chen, M. & Pan, P. C. LCD-based color filter films fabricated by a pigment-


151. Sun, P., Ishihara, R. & Charbon, E. Flexible ultrathin-body single-photon avalanche diode


186. Pisano, F. et al. Depth-resolved fiber photometry with a single tapered optical fiber implant.


241. Stark, E., Koos, T. & Buzsáki, G. Diode probes for spatiotemporal optical control of


