Restoring Fine Finger Control to Paralyzed Hands Using a Low-Power Brain-Controlled Functional Electrical Stimulation Neuroprosthesis

by

Samuel Ross Nason-Tomaszewski

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biomedical Engineering)
in the University of Michigan
2022

Doctoral Committee:

       Associate Professor Cynthia A. Chestek, Chair
       Professor David Blaauw
       Assistant Professor Kevin L. Kilgore, Case Western Reserve University
       Assistant Professor Scott F. Lempka
       Associate Professor Parag G. Patil

Samuel R. Nason-Tomaszewski

samnason@umich.edu

ORCID iD: 0000-0002-7127-0986

Dedication

I dedicate this dissertation to my family. To my Bubbe, Zayde, Grandma, and Grandpa, who I always found motivating even after their time with us. To my sisters, Livia and Abby, whose encouragement and support inside and outside of education has helped me stay true to my goals. To my parents, Carol and Michael, who, among countless things, committed to me throughout my life to support my love for science. And to my wife, Claire, who shared the entirety of this Ph.D. journey with me. I mean it when I say I could not have done this without you there, with me, every step of the way. We started on this path together, we are completing this step together, and now I look forward to the rest of our lives together.

Acknowledgements

As all science should be, everything presented in this dissertation is a result of the combined efforts of many people.

Firstly, I thank my family for their endless support during my Ph.D. My wife, Claire, who I likely would not have met if not for this program, has always supported my decisions related to my Ph.D. and career beyond. With the amount of hours I have talked your ear off looking for solutions to every research-related problem I faced, you deserve an honorary neural engineering Ph.D. in addition to your tissue engineering one. Thanks to my parents and sisters, Carol, Michael, Livia, and Abby. Your support and encouragement has kept me going the whole time, and your frequent departures from warmer Florida climates to visit us in the dead of winter will be memories I cherish forever. Thanks to the new extension to my family, the Tomaszewskis, who have always provided an entertaining distraction from education when I needed it most. I also want to thank my extended family for sending words of love at every opportunity. I cannot wait for the days when we can all safely be together again! And last but not least, my cats, Bean and Turkey. Thank you for the goodbyes when I would leave for lab during the day and the happy hellos when I got home. Gestures like these give something to look forward to at the end of a long day.

My advisor, Cindy Chestek, deserves particular recognition for her contributions to this dissertation. You believed in me from day 1, and I will forever feel indebted to you for how much you have helped me as a Ph.D. student in your lab, not to mention the endlessly helpful career advice and connections you have provided. I would consider myself lucky to become half as incredible of a Ph.D. advisor as you were to me, and you have been an incredible role model in

my efforts to become a better advisor to younger students. The creativity with which you approach science and the positivity with which you see every result is a state of mind I hope I picked up, at least a little bit, during my time studying with you. I will miss our frequent musings about the future of the brain-machine interface field, but I look forward to continuing them at conferences and other meet-ups in the future. I also owe a major thanks to Parag Patil, my unofficial co-advisor. Your confidence in yourself and in me helped me grow a little bit of self-confidence even during the most uncertain times. Thank you for the concrete surgical skills as well as the abstract surgeon's perspective to neural engineering. Thank you also for the career guidance and frequent recommendations to remember to take a quick breather whenever time was short and the work was piled high. I also want to extend thanks to the other members of my dissertation committee, David Blaauw, Kevin Kilgore, and Scott Lempka. Your opinions and scientific input had a substantial impact on the results of this dissertation, and I hope you see your footprints in the text that follows.

I would also like to thank my labmates, new, old, and unofficial, for their assistance with the science, for helping me develop professionally, and for their friendship. Paras Patel and Eric Kennedy, I consider myself lucky to be able to call you friends and colleagues. Thank you for the scientific and career advice, but also for always being around to watch a cooking show during lunch and stop for donuts at all hours of the night. Thank you Autumn Bullard, Charles Lu, Chrono Nu, Alex Vaskov, Philip Vu, Elissa Welle, and Matthew Willsey for teaching me all of the technical, experimental, and research skills I needed to conduct the work described in this dissertation. Thank you to my newer labmates Joey Costello, Joe Letner, Matt Mender, Julianna Richie, Hisham Temmar, and Dylan Wallace, for helping me learn to become a better mentor and for always providing your inputs to the science. I am so fortunate to call all of my labmates friends, and I cannot wait to see where life takes you in the coming years. And to all my personal friends

Table of Contents

List of Tables

List of Figures

List of Equations

List of Appendices

Abstract

Paralysis of the upper extremity is a devastating outcome of many neurological diseases and disorders. Brain-machine interfaces (BMIs) attempt to bypass the disability by recording information directly from the subject's brain and predicting the user's intentions to control a prosthetic device. Modern brain-machine interfaces have made limited translation to clinical use, where studies have not expanded far beyond controlled laboratory environments. Two of the primary hindrances to their widespread clinical translation is their dependence on stacks of power-hungry computers and performance compared to the able-bodied hand. The aim of this work is to establish low-power brain-machine interface technologies that restore fine control to paralyzed hands.

The first study presents the 300-1,000Hz spiking band power (SBP), which is a low power neural spiking feature that requires 90% less data than the standard threshold crossing rate (TCR) neural feature. In simulation, we found that SBP can extract accurate neural spiking patterns at lower signal-to-noise ratios and with greater unit specificity than TCR. Because of this, closed-loop decoders which used SBP performed as well or better than decoders using TCR in two rhesus macaques.

In the second study, we investigated whether BMIs could be implemented on embedded devices fit for implantation. We used three off-the-shelf low-power amplifiers controlled by a 32-bit microcontroller to perform SBP recording, feature extraction, and decoding. The device could achieve equivalent performance to our high-powered BMI when closed-loop predicting one-finger movements and comparable performance when predicting two-finger movements in a nonhuman

primate. To do so, the device required 58.4mW (equivalent to 11.3hr usage time with a standard 200mAh implantable battery), which we could compress to 12.5mW (52.8hr usage time) with an optimized processing pipeline implemented on an integrated circuit.

The third study showed, for the first time, that BMIs can control the simultaneous and independent movements of two finger groups in real-time with nonhuman primates. With the BMI, the primate could acquire targets at a rate of nearly 2 per second. Additionally, we found that cortical activity for independent finger movements and combined finger movements were similar. This allowed linear models to predict behaviors that were not used for training with a correlation coefficient at least 90% as high as a linear decoder trained on all behaviors.

In the fourth study, we investigated how well continuous finger movements could be restored with a brain-controlled functional electrical stimulation (BCFES) system. Following temporary paralysis delivered via nerve block, a nonhuman primate improved success rates to 89% with a 1.4s median target acquisition time in a one-finger task by using the BCFES system, up from 2.6% and a 9.5s median target acquisition time (near chance) when using his paralyzed native hand. Additionally, we allowed the monkey to use the BMI (no stimulation) to complete the two-finger version of the task following paralysis, and performance could be recovered by performing recalibrated feedback-intention training one time following paralysis, despite the absence of sensory feedback.

The results of this work demonstrate that low-power BMIs can restore substantial function in cases of upper extremity paralysis. All of the work presented here uses low-power technology that can simply be implemented on implantable devices. Next steps for this work will require validation of the hand control results in humans and development of completely implantable neuroprostheses to restore native hand functions to people with paralysis.

Chapter 1 Introduction

A version of this chapter "Restoring Upper Extremity Function with Brain-Machine Interfaces"

was published as chapter 8 in the book "Emerging Horizons in Neuromodulation: New Frontiers

in Brain and Spine Stimulation" in 2021 (Nason et al., 2021a).

The advent of digital signal processing and high-speed computation has opened the possibility of restoring lost upper-extremity function through brain-machine interface (BMI) neuroprosthetic devices. By capturing brain activity, predicting underlying intentions from that activity, and controlling a prosthetic device accordingly, BMI neuroprostheses emerge as a promising therapy for the 27 million existing patients worldwide with spinal cord injury and the nearly 1 million new patients injured annually (James et al., 2019). Presently, BMI devices can guide robotic arms and, in some cases, activate paralyzed extremities with volitional control in a limited way in a laboratory environment. Table 1.1 details some of the major accomplishments made by groups investigating brain-machine interfaces across the past 40 years, from studies in animals to present human clinical trials. Going forward, the promise of the field is encouraging.

BMI technologies have arisen from both publicly and privately sponsored research efforts. The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative of the National Institutes of Health (NIH) has funded hundreds of academic, neuroscientific, and neuroprosthetic research programs, yielding major advances in neural interface technologies and the understanding of brain signaling. In addition, research performed in the private sector has spawned numerous companies (Emotiv Inc., Neuralink Corp., Kernel, Paradromics Inc., InBrain

Table 1.1 Significant Milestones for the Development of Modern Brain-Machine Interfaces

Year	Subject	Milestone	Reference
1982	Monkey	Primary motor cortex neurons change activity levels depending on the directions of arm movements	(Georgopoulos et al., 1982)
1996	Cat	The Utah microelectrode array is presented for single unit recording with 100 electrodes	(Nordhausen et al., 1996)
1999	Rat	Real-time control of a robotic arm using intracortical brain activity	(Chapin et al., 1999)
2002-2003	Monkey	Robotic arms and computer cursors controlled in real-time using a brain-machine interface (BMI) with monkeys	(Carmena et al., 2003; Serruya et al., 2002; Taylor et al., 2002a)
2006	Monkey	First high-performance BMI	(Santhanam et al., 2006)
2006	Human	Human implanted with a Utah array uses a BMI	(Hochberg et al., 2006)
2008	Monkey	Monkey self-feeds using a brain-controlled robotic arm	(Velliste et al., 2008)
2008	Monkey	Monkey uses a single-neuron BMI to control functional electrical stimulation	(Moritz et al., 2008)
2011	Human	Continued Utah array functionality for 1,000 days after implantation in a human	(Simeral et al., 2011)
2012	Monkey	Monkeys intuitively control functional electrical stimulation of their paralyzed arms using a BMI	(Ethier et al., 2012)
2012	Human	Human uses a BMI to control a robotic arm for self-drinking	(Hochberg et al., 2012)
2015	Human	Human uses a BMI to control a 10-dimensional robotic arm	(Wodlinger et al., 2015)
2015	Human	Human uses a BMI recording from posterior parietal cortex to control a computer cursor	(Aflalo et al., 2015)
2015	Human	Human uses a BMI to virtually type on a computer	(Jarosiewicz et al., 2015)
2016	Human	Human receives intracortical electrical stimulation in somatosensory cortex to generate perception using Utah arrays	(Flesher et al., 2016)
2016	Monkey	Paralyzed monkeys use brain-controlled spinal cord stimulation to restore walking	(Capogrosso et al., 2016)
2016	Human	Human uses brain-controlled surface functional electrical stimulation to control hand function	(Bouton et al., 2016)
2017	Human	Human uses brain-controlled implanted functional electrical stimulation to restore whole-arm function	(Ajiboye et al., 2017)

Note: As of September 2018, 18 Utah arrays have been implanted chronically in humans (Bullard et al., 2020)

Neuroelectronics) and corporate ventures (Google Brain, Facebook Reality Labs) seeking to enable volitional interaction with prostheses, computers, vehicles, and video games through thought.

In this chapter, we will first explore various portable methods of extracting information from the central nervous system that are capable of predicting motor intention. Then, we will focus on the potential of high-bandwidth intracortical recording to provide the information needed to restore accurate and precise movement. Neuroprostheses have progressed drastically in the past twenty years, from slow predictions (Carmena et al., 2003; Serruya et al., 2002; Shenoy et al., 2003; Taylor et al., 2002a) to control systems using all limbs of the body (Willett et al., 2020). As such, we will conclude our chapter with some discussion of the newest neuroprosthetic technologies and some projections for the coming years.

1.1 Neural Signal Recordings

Diagnosing neurological disorders by recording the activity of the central nervous system has become common practice in medicine. Across the past several decades, several recording modalities have become critical components of many diagnostic procedures, some of which are illustrated in Figure 1.1. From non-invasive to invasive deep-brain methods, which are compared in Figure 1.2, each has prevalence in the realm of neuroprosthetic devices, which we discuss below.

1.1.1 Noninvasive Recording

Noninvasive recording modalities require minimal to no alteration of the person to record brain activity. Without crossing the skin barrier, noninvasive recordings avoid many of the infection, scarring, and surgical risks that are inherent to more invasive recording methods. This makes noninvasive recording methods the most readily accessible to any person, as the risks of complications are low. Technologies commonly used in clinical spaces, such as functional magnetic resonance imaging, functional near-infrared spectroscopy, and
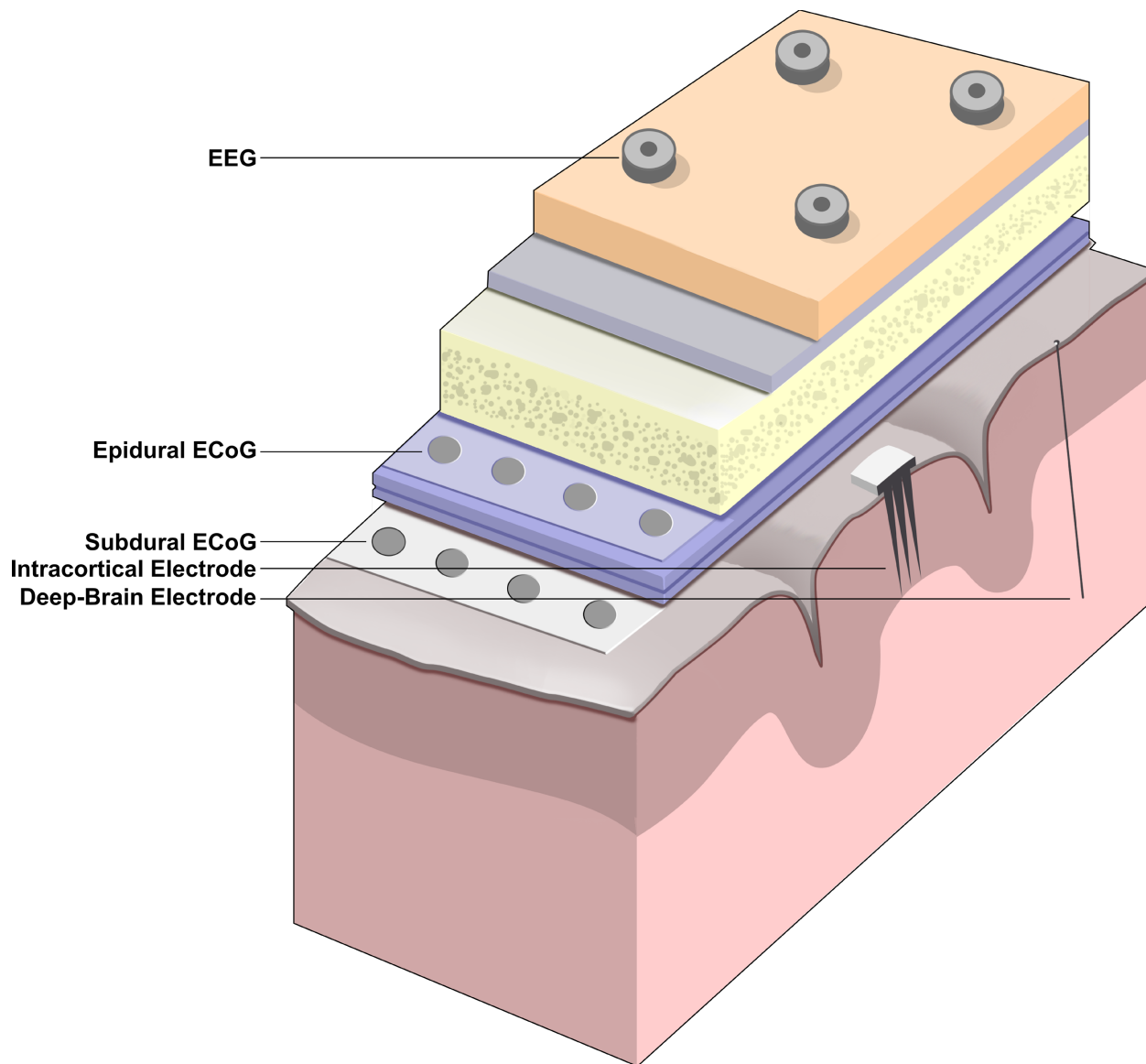
Figure 1.1 Relative invasiveness of various portable neural recording modalities. Scale of the drawing is not realistic. EEG: electroencephalography, ECoG: electrocorticography.

magnetoencephalography, have substantial literature demonstrating their capabilities to measure brain activity noninvasively. However, the seconds to minutes of latency between cognitive effort and representation in the measured signals, coupled with massive equipment needs, render these methods infeasible for real-time and portable neuroprostheses. Consequently, we will begin our discussion with electroencephalography (EEG).
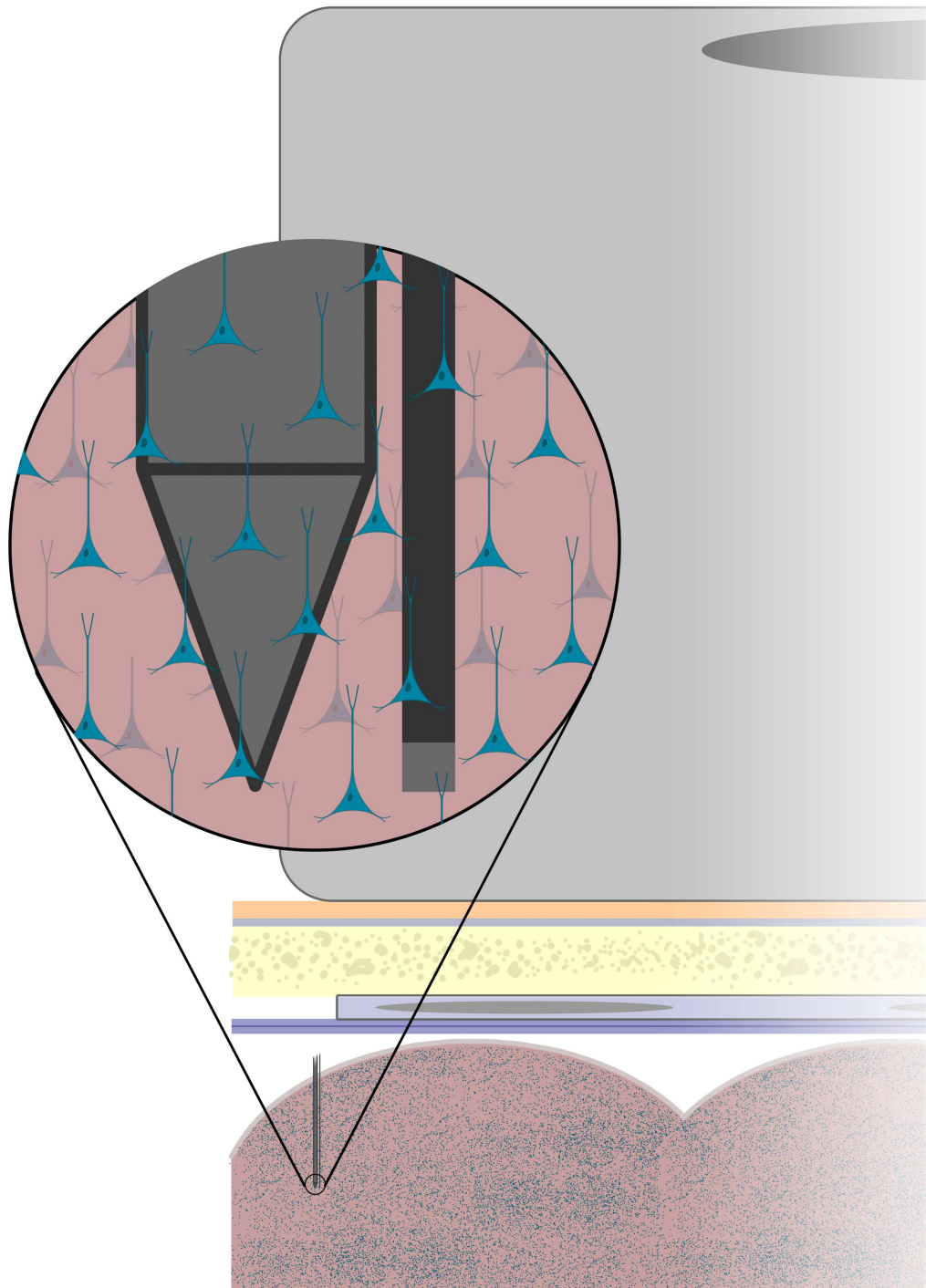
Figure 1.2 Relative specificity to single neurons of various portable electromagnetic neural recording modalities. Neurons as displayed in the inset are realistically dense at 40,000 neurons/mm^3. Neuron distributions and densities in the main image are not realistic.

Electroencephalography is a non-invasive recording technique that measures the electromagnetic byproducts of brain activity instantaneously. As neurons in the brain fire and

signals traverse axons and dendrites, small-amplitude electric currents are generated along with coupled magnetic fields, which are also weak in magnitude. When thousands to millions of neurons in close proximity become active in synchrony, which is typical during cognitive activation of a brain region, the magnetic fields from individual neurons superimpose to create larger-magnitude magnetic fields. These larger-magnitude magnetic fields generate small electrical currents on the scalp, which is measured by EEG. EEG signals are recorded using metallic electrodes several millimeters to centimeters in diameter that make electrically conductive connections with the scalp. Different frequency bands of the recorded activity, such as the delta (~<4Hz), theta (~4-8Hz), alpha (~8-15Hz), beta (~15-30Hz), and gamma (~30-40Hz) bands, are assumed to represent various types of cognitive effort, enabling a neuroprosthesis to predict intention of the user from strength, or power, of these bands.

Due to the simplicity of the EEG cap, EEG systems are lightweight, safe to equip, and portable enough for use with a neuroprosthesis. EEG-based BMIs were originally proposed in the 1980s, with the P300 speller as the first brain-controlled computer (Farwell and Donchin, 1988). With relative ease of implementation and use, they have been thoroughly explored and usually are limited to communication applications, which can accommodate high error rates. Classification accuracies of EEG-based BMIs alone can be fairly high in the range of 90%, but their information throughputs (a common measure of performance for all types of BMIs) are just 0.71 bits per second (Speier et al., 2014). This level of information is useful for rehabilitation after stroke (Hodkin et al., 2018) and some commercial uses, such as controlling basic computer functions and video games. Unfortunately, the relatively low throughput, reliance on conductive and gelled electrodes (which are difficult to place with paralysis), and requirement to be motionless (Kline et al., 2015) make EEG usage for functional restoration difficult.

1.1.2 Brain Surface Recording

Noninvasive recordings are substantially displaced from the source neurons, making it difficult to determine the specificity of a particular recording to some behavior. Invasive recordings attempt to address the challenges of spatial specificity by placing electrodes closer to neurons via surgical implantation. Up to 60% of people with paralysis, from spinal cord injury, ALS, and other neurological disorders, have reported they would be willing to undergo surgery to enable use of a neuroprosthesis (Blabe et al., 2015). All invasive recording methods developed to date capture electrophysiological signals from brain activity. The least invasive of such procedures involves placing an array of electrodes, similar in organization to an EEG array but with smaller electrode sizes, on the surface of the dura mater. Implanting the electrodes beneath dura and in direct contact with the surface of the brain is more invasive but brings the electrodes closer to the neurons being recorded.

These types of invasive recordings are called epidural and subdural electrocorticography (ECoG). ECoG electrode arrays are organized in similar fashion to EEG electrode arrays, but with smaller contact sizes, flatter materials to fit between the skull and the brain, and generally greater electrode quantities. Similar to EEG recordings, ECoG recordings represent the compound activity of neurons, but on the order of hundreds to thousands of cells as the electrode sizes are generally smaller and the electrodes are closer to the neural populations. The same general frequency bands of population activity are found in ECoG and EEG.

For BMI applications, ECoG is adept at detecting movement effort with nearly 100% accuracy due to the greater neuronal specificity. A popular application of ECoG-based BMIs is the prediction of hand and finger movements. While many continuous control experiments predict one degree of freedom (Chao et al., 2010; Flint et al., 2017; Hotson et al., 2016; Kubánek et al., 2009),

most of them rely on projecting a relatively binary movement/no movement signal onto a continuous movement dimension, instead of directly predicting position along that dimension. Others have attempted to classify attempted postures of the hand instead of the positions of the fingers, but required seconds of latency to make accurate predictions (Chestek et al., 2013; Pistohl et al., 2012). Alternatively, using ECoG recordings to predict speech has been quite successful, enabling prediction of speech in human clinical trials (Anumanchipalli et al., 2019; Makin et al., 2020). Since speech is spread across a wide area of cortex, the wide expanse of ECoG arrays makes speech prediction a particularly productive application for ECoG.

ECoG recording electrodes have also experienced some technological advancements in recent years. Not only are electrode sizes decreasing, enabling greater quantities of channels and greater specificity to neurons (Viventi et al., 2011), some sizes have become so small that single-neuron action potentials have been recorded from the surface (Khodagholy et al., 2015). It has been hypothesized using field modeling that this capability requires very small electrodes in very close contact to the brain, where the potentials are likely originating from superficial layer I neurons (Hill et al., 2018).

1.1.3 Penetrating Extracellular Recording

The most invasive neural recording options place electrodes among the cells in the brain. This enables simultaneous capture of population activity and the activity of individual neurons within the electrophysiological recordings. In contrast to the disc-shaped electrodes discussed thus far, electrodes that penetrate brain tissue are needle-shaped probes with one or more recording sites. As individual neurons are activated, they produce an action potential that creates a voltage gradient in the nearby extracellular space, which is detectable within approximately 100-150μm from the neuron (Henze et al., 2000; Moffitt and McIntyre, 2005). If a shank with a recording site

is within that range, its voltage recording may represent each action potential's voltage gradient with what are called "spikes" in the signal, or visible perturbations that are higher frequency compared to the population activity bands discussed for EEG and ECoG. By recording the spiking rates of individual neurons, these most invasive modalities of recording are also the most specific to behavior.

The most invasive penetrating electrodes are flexible cylinders tens of centimeters in length with multiple contacts. These electrodes provide deep-brain stimulation to structures such as the ventral intermediate nucleus of thalamus, subthalamic nucleus, and zona incerta to reduce symptoms of motor neurodegenerative disorders, such as Essential Tremor and Parkinson's Disease (Benabid et al., 2009). However, these electrodes can also provide electrophysiological recordings of these deep brain structures related to motor intentions (Swann et al., 2018). Due to the contact size of typical deep-brain stimulation macroelectrodes, the ability to distinguish the activity of individual neurons is difficult with so many active neurons nearby the large electrode area. Instead, often the local field potentials representing the aggregate activity of the nearby hundreds of neurons drives the voltage recordings. On the other hand, deep-brain microelectrodes have electrode areas that are small enough to record the spiking activity of individual neurons. Some studies have used these electrodes to obtain BMI control signals (Patil et al., 2004; Tan et al., 2016). Thalamus and neighboring structures are promising targets for deep-brain neuroprostheses. Descending cortical and cerebellar motor commands can be captured entirely within a few cubic millimeters, suggesting that just one or two deep-brain microelectrode arrays could capture all signals for a whole-body neuroprosthesis. However, limited site quantities and surgical risks associated with heavily-invasive deep-brain implants are challenges that must be overcome before widespread clinical translation (Bullard et al., 2020).

More superficial implants lower the surgical risk of extracellular recording. Michigan-style probes (Wise et al., 2004) allow many simultaneous recordings with hundreds to thousands of planar electrodes and have had tremendous success across the neuroscientific community. One shank with nearly one thousand electrodes, such as the new Neuropixels probe of similar design to the Michigan-style probe (Jun et al., 2017), can provide highly spatially specific signals along columns of neural processes. Recordings along columns of neurons are extremely helpful for neuroscientific investigations because neural signals traversing axons and dendrites from one neuron to another can be tracked precisely in space and time. However, for a high-dimensional neuroprosthesis, sampling many independent neurons across multiple columns that represent those behavioral dimensions is necessary.

Single-recording-site penetrating electrodes enable simultaneous recording from multiple places in the brain. Historically, microwires were implanted individually to obtain simultaneous recordings from separate cortical areas, and either left-in-place chronically or explanted following each experiment (Georgopoulos et al., 1984; Taylor et al., 2002a). Then, microfabricated devices like the Utah microelectrode array (or the Neuroport array approved for use in humans) were developed for simpler chronic implantation of electrodes by providing an organized bed of single-site shanks on a rectangular platform to cover a wide cortical area (Nordhausen et al., 1996). The shanks, which are generally all the same 1.1 or 1.5mm length with a 0.4mm pitch for cortical implants, target layer V neurons in motor cortex, which will tend to be pyramidal cells driving corticospinal neurons. Since the technology is already approved for acute use in humans to localize sources of epilepsy and it allows spatially specific recording of the spiking activity of many independent neurons, Utah arrays have become the standard implant for chronic intracortical BMI studies in nonhuman primates and humans, enabling self-feeding (Ajiboye et al., 2017; Hochberg

et al., 2012; Velliste et al., 2008), high-speed typing (Gilja et al., 2012, 2015; Kao et al., 2017; Pandarinath et al., 2017), and control of their own limbs despite paralysis (Ajiboye et al., 2017; Bouton et al., 2016; Capogrosso et al., 2016; Ethier et al., 2012).

1.2 Signal Processing and Prosthetic Control

Brain-machine interfaces have not yet translated to full-time use because of several inherent issues with experimental systems. Current clinical and experimental BMIs require a lab cart with all of the computers, require that the user remain tethered to the recording system with all of the high-bandwidth data being recorded, and consume hundreds of watts (several car batteries worth) of electrical power to run the necessary computations. This is an astounding amount of power in comparison to typical implantable medical devices like pacemakers and deep-brain stimulators, which require just tens of microwatts. Relaxation of all of these requirements is critical for therapeutic BMIs, and recent work has suggested promising solutions to many of these problems. Below, we discuss the minimum required components of a BMI in terms of their history, current investigations, and opportunities for improvement in the future.

Intracortical brain-machine interfaces are generally composed of several key components, which are portrayed in Figure 1.3. First, the tissue interface directly acquires electrophysiological signals from the brain, as discussed above. These voltage signals must be filtered and then amplified so they can be converted into digital signals. This allows general computing hardware to perform computations on those measurements. The first in the chain of computations is feature extraction, which extracts the components from neural signals, known as neural features, that are easiest to use to make predictions. The second is the prediction algorithm that converts the neural features into a behavioral prediction, usually referred to as a "decoder." Finally, those behavioral predictions are transmitted to a prosthesis to perform the intended command of the user.
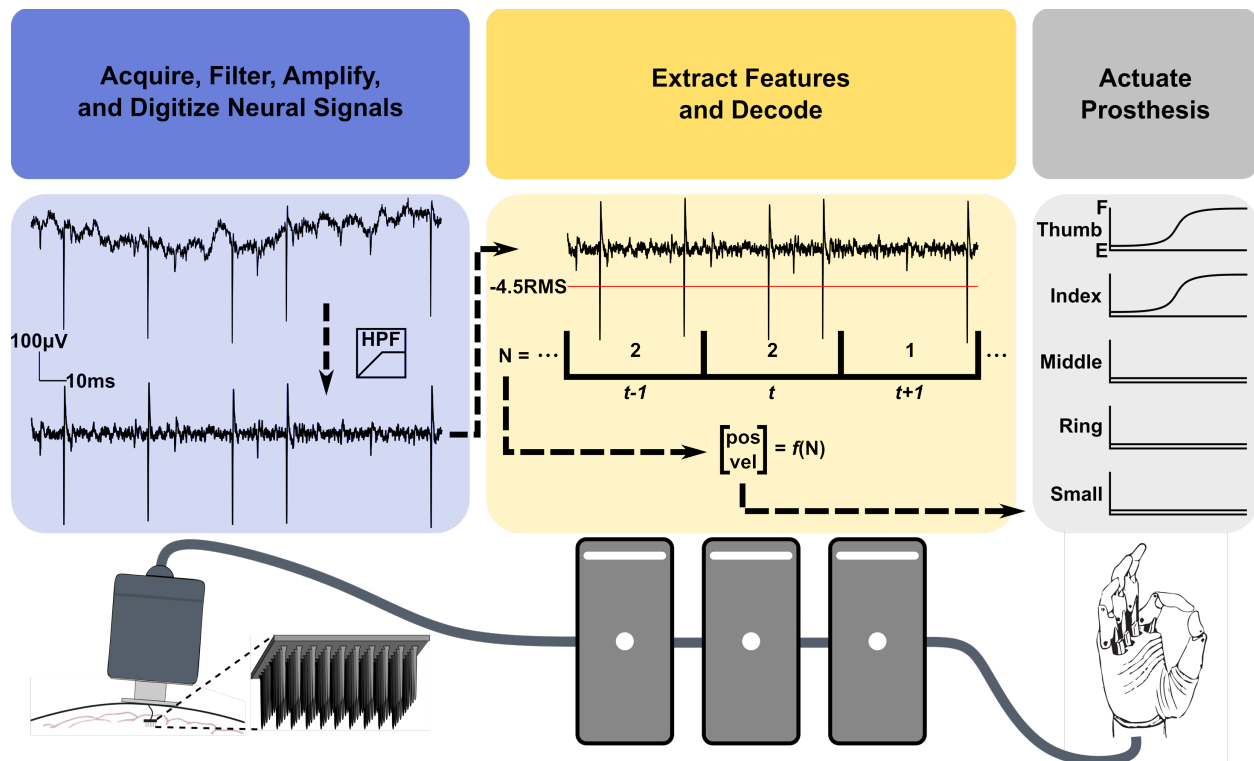
Figure 1.3 Flow of information through a brain machine interface. First in blue, the analog front-end acquires the raw intracortical signals from implanted electrodes, then amplifies, filters, and digitizes them. Next in yellow, crossings of a threshold are measured to estimate spiking rates of recorded neurons, which are decoded into the position and velocity intentions of the user. Finally in gray, a prosthesis is actuated according to the user's intentions.

1.2.1 Filtering, Amplifying, and Digitizing

Typically referred to as the analog front-end (AFE), the hardware required to filter, amplify, and digitize intracortical signals has been a large topic of investigation for decades. Although modern experimental BMIs employ several computers, recent simulations show that the power consumption of the minimum required BMI components (as discussed previously) is dominated by the AFE (Nason et al., 2020). Consequently, it is critical to reduce the power consumption of the AFE to translate BMIs into widespread clinical use.

At minimum, experimental AFEs first filter and amplify the incoming neural signals (Chandrakumar and Markovic, 2017a, 2017b; Chen et al., 2014; Harrison and Charles, 2003; Jang et al., 2018; Lopez et al., 2014; Mahajan et al., 2015; Muller et al., 2014; Ng and Xu, 2016; Shen

et al., 2018). The first filter in the chain is a band-pass filter that eliminates frequencies outside of those relevant to brain activity. Typical BMI AFEs inherently filter out frequencies below approximately 0.1Hz and above approximately 10kHz to permit passing of population activity that exists in the single to hundreds of Hz and single neuron activity that exists in the single kHz range. Any electromagnetic interference from radio, Wi-Fi, or cellular signals and the baseline voltage of the body's tissue are therefore attenuated by this first stage filter. Then, groups typically apply an additional set of filters to extract one or both of the population activity (several filters with all cutoffs below approximately 500Hz) and the single neuron activity (high-pass filter with a cutoff near 200-300Hz).

In addition to reducing noise, front-end filters also amplify their neural signal inputs. Amplification is necessary to bring the neural signals, which are typically on the order of tens to hundreds of microvolts in amplitude, to voltages at which the downstream electronic circuits generally operate, which are on the order of volts. Conveniently, this amplification factor can be incorporated into the design of one or both of the filter stages. We can further compress the circuitry to save electrical power consumption with another optimization. Instead of two stages of filters, which is helpful for on-the-fly configuration for experimentation, one appropriately configured filter would suffice in a clinical BMI. This eliminates some unnecessary hardware and can substantially cut the electrical power consumption. One final improvement to the efficiency of the amplifiers and filters is related to the inherent advancements in transistor technology. Improved fabrication technologies, reduced transistor sizes, and more electrically efficient materials have enabled drastic improvements in the power efficiencies of neural amplifiers across the past several decades.

The last stage of the AFE is the analog-to-digital converter (ADC). The ADC receives the filtered and amplified analog neural signal and periodically converts it to a digital one for use with computers. The rate at which conversions occur can be controlled, but it should be at least double the highest recorded frequency to enable valid and artifact-free reconstruction in the digital domain (Shannon, 1949). The resolution of the conversions can also be controlled with hardware customizations and is generally set to 16 bits. Such a high resolution allows the detection of sub-microvolt changes in the recorded voltages while maintaining a dynamic range large enough to compensate for millivolt scale recording artifacts, which can result from some movements or electrical stimulation (see subsequent sections for a discussion of this). Although ADCs are critical to the operation of a BMI, the fabrication and design changes that have resulted in improved power efficiency are largely independent of the field of neuroprostheses. As such, we discuss only the selection characteristics of ADCs that can provide power-savings.

In many BMI usage cases, there are rarely electrical artifacts in the signals. In usage cases that inherently generate artifacts, some of which will be discussed in later sections, there are methods to record neural signals while avoiding those artifacts. The 16-bits of ADC resolution typically used in BMI recording systems scales into the millivolt range to accommodate for these artifacts, which is an order of magnitude larger than the largest of signals recorded from intracortical electrodes. Consequently, 16-bits is a luxurious resolution that is generally overdesigned for most BMI applications. In one analysis of the ADC requirements for BMIs, it was found that 8 effective bits of resolution would suffice to accurately capture neural spikes (Even-Chen et al., 2020). A 50% reduction in the quantity of bits, from 16 to 8, would result in an estimated 99.997% reduction in the power consumption of each ADC (Murmann, 2020; 2004).

Additionally, whether by hardware or software, reduction in the quantity of bits reduces the computational complexity down-stream of the ADC, which will be discussed later.

The power consumption of both amplifiers and ADCs is directly tied to the bandwidth of the signals they are recording. Intracortical local field potentials require low recording and amplification bandwidths, similar to ECoG and EEG neural activity bands, which implies reduction in the power consumption for those amplifiers and ADCs. Unfortunately, since they represent the net activity of hundreds of cells, their relationship to behavior is more abstract than single unit spiking rates. This may explain why local field potential-based BMIs have not achieved the performance levels of spike-based BMIs (Flint et al., 2013; Milekovic et al., 2018; Perge et al., 2014; So et al., 2014; Stavisky et al., 2015; Wang et al., 2014). Alternatively, some groups have investigated reducing the bandwidth of spike recordings. One study found that the standard 0.25-7kHz bandwidth typically used to detect threshold crossings resulting from spikes is unnecessarily large, and a 1-2kHz bandwidth more than suffices to maintain high decoding performance (Even-Chen et al., 2020). We also previously found that a more reduced 0.3-1kHz bandwidth, denoted the spiking band power, predicts behavior comparably well to higher-bandwidth decoders (Irwin et al., 2016), with the additional benefits of greater signal-to-noise ratio and greater specificity to single neurons (Nason et al., 2020). By reducing amplifier and ADC bandwidth to the 0.3-1kHz spiking band, we found that the AFE could undergo a 90% reduction in power consumption with the custom circuits described here (Irwin et al., 2016; Nason et al., 2020).

Although custom analog circuits have great promise towards promoting next generation BMIs that can be used anywhere, challenges remain that must be addressed before they can be deemed safe for general use with humans. Customized amplifiers like those cited here are difficult to fabricate reliably without being thoroughly tested in a variety of conditions and licensed by a

commercial vendor. Only a few academic neural amplifiers have been used regularly in preclinical studies with animals. Cortera Neurotechnologies, Inc. (Berkeley, CA, USA), have developed one such device capable of recording and stimulation (Johnson et al., 2017). The RHD series of devices produced by Intan Technologies (Los Angeles, CA, USA) is another example of an integrated electrophysiological sensor that has translated to experimental use (Harrison and Charles, 2003; Harrison et al., 2007). The amplifiers have maintained their customizability while keeping consumption low, requiring 1.07mW/channel for typical 10kHz bandwidth or 0.14mW/channel for 0.3-1kHz spiking band power bandwidths. Although not yet used for many translational studies to date, they have been incorporated into embedded BMIs for animal studies (Bullard et al., 2019) and tested with humans (Wang et al., 2019). Existing and future analog circuits provide clear improvements to current clinical BMIs, but they require sufficient iteration and functional testing and validation prior to their commercialization and full-time use with BMIs.

1.2.2 Processing and Decoding

Once neural signals are digitized, BMIs employ computational algorithms that convert the recorded brain activity into a person's intentions. These algorithms are called decoders as they decode the encoded neural information into a more usable control signal. The encoded information for intracortical recordings is often some representation of each electrode's spiking or firing rate, broadly called features in the machine learning community, which are extracted via some processing discussed in the subsequent paragraph. The different effectors that can be controlled by these neural features will be discussed in detail below, but some brief examples include computer cursors and robotic or natural arms. BMIs often predict positions and velocities from the neural features to control these effectors, but in cases where limb movements are being predicted, BMIs can also predict the activations of the limb's muscles through a biomechanical model (Naufel et

al., 2019). These predictions are made at a physiologically relevant frequency (>10 predictions per second) so that movement of the effector appears continuous and the visual feedback is readily provided. As was discussed previously, neural data from most recording modalities have been used to predict these control signals, but intracortical recordings have been most successful in predicting accurate and precise real-time control signals due to the better time and spatial resolution.

The first processing stage in an intracortical BMI is the extraction of relevant neural features from the digitized neural data. The bandwidth of the digitized neural data often governs what types of neural features can be extracted. If the data is low bandwidth, or filtered below approximately 500Hz, then local field potentials, such as power in the delta, theta, alpha, beta, and gamma bands, are the primary feature. Filtering for these features can be present in the AFE or otherwise be implemented in processing. Within each band, some estimate of the signal power is calculated by integrating the squared signal over some time period for each channel. Then, a subset of these features can be used to predict movement in a BMI (Flint et al., 2012a). The period over which these calculations are computed generally corresponds to the frequency at which an updated prediction is required, often on the order of tens of milliseconds. Since the relationship between local field potentials and complex behavior is less directly related, as previously discussed, intracortical BMIs more often use spiking rates of the recorded neurons as the neural feature.

Spiking rates can be extracted from digitized signals that have been filtered between approximately 300Hz and 10kHz. The spiking rate of an individual unit, or one or more neurons firing in synchrony with morphologically similar spikes, can be extracted through a process called spike sorting, discussed in detail by others (Lewicki, 1998; Rey et al., 2015). After sorting, the spikes of each unit are counted in similar periods, yielding one feature per unit. Unfortunately, spike sorting requires substantial processing, either manual or algorithmic, that makes

implementation on translational BMIs difficult due to their limited space for computations. As such, many groups have transitioned to using just the threshold crossing rates of each electrode without sorting the spikes, as it is presumed that the neurons nearby one electrode encode similar information (Ventura, 2008). It has also been demonstrated that simply using the multiunit threshold crossing rates even for neuroscientific analyses is valid with negligible information losses (Trautmann et al., 2019). Unfortunately, extraction of threshold crossing rates requires the high power AFEs discussed in the previous section, making them difficult to fit in the budget for portable and low-power BMIs. Hence, lower bandwidth proxies for spiking rates, like the 0.3-1kHz spiking band power discussed in the previous section, may be more realistic features for a portable BMI.

After feature extraction, there are a number of different decoding algorithms to translate the observed neural data into a command signal. Before any particular algorithm can be used in a BMI, it first must undergo its own specific training procedure that optimizes the decoder's parameters for best prediction of those control signals. These parameters, sometimes called weights, represent how any particular measurement of neural activity can be translated to a control signal from some number of operations. Early on, it was observed that neurons modulate their firing rate with the direction of movement (Georgopoulos et al., 1982), which laid the groundwork that control signals could be predicted from weighted sums of the observed neural signals. The oldest decoding algorithm is perhaps the population vector algorithm based on this work (Georgopoulos et al., 1985). Early nonhuman primate BMI experiments used population vectors to control a cursor in real-time (Taylor et al 2002), but the algorithm was mostly replaced by other linear/weighted-sum decoders shortly after. Linear regression decoders, which no longer make the same population vector assumption that each recorded neuron contributes equally to a behavior

18

(Chase et al., 2009), allowed nonhuman primates to control computer cursors (Serruya et al., 2002) and a robotic arm (Carmena et al., 2003). Ridge regression decoders, which employ a regularization to the linear regression to train more representative weights and avoid some mathematical issues during training, have enabled human participants to control multiple degrees of freedom of a robotic arm (Collinger et al., 2013). Additionally, Kalman filters (Kalman, 1960), which optimally combine predictions based on prior behaviors and predictions based on current neural data, have demonstrated great success in predicting nonhuman primate (Gilja et al., 2012; Irwin et al., 2017; Mulliken et al., 2008; Wu et al., 2004a) and human behavior (Ajiboye et al., 2017; Gilja et al., 2015; Malik et al., 2011; Pandarinath et al., 2017; Simeral et al., 2011).

Linear decoding models have enabled impressive accomplishments with BMIs, but they maintain an inherent assumption that the relationship between brain activity and behavior is linear. They cannot accommodate the possibility where the behavior is not well described by a weighted sum of the neural activity measurements. It may ultimately be possible to achieve better performance with a nonlinear model that better represents what is likely to be a nonlinear relationship outside of a limited context. Some groups have already explored this by implementing computational neural networks for the prediction of behavior. Training and using neural network decoders are complex procedures, though the recent advancements in high-speed and parallel computing have made that process feasible for heavy computational machines. An early recurrent neural network demonstrated great cursor control in nonhuman primates (Sussillo et al., 2012), and more recent networks have improved BMI control in humans (Schwemmer et al., 2018). Though these studies have given neural networks promise for BMI control, their computational complexity must drop drastically before they can become relevant to portable or even implantable BMIs.

Whether an algorithm is linear or nonlinear, there are several additives that can potentially reduce the computational complexity and improve the decoding performance. Closing the feedback loop for real-time BMIs creates a very different control system than if the existence of feedback is ignored, providing an opportunity for further optimization of a BMI decoder's parameters. This has driven researchers to employ a technique called intention training that incorporates some assumptions about the user's behavioral intentions when performing some action. One example is recalibrated feedback intention training (ReFIT), originally proposed for usage with the Kalman filter but contains principles that can be applied to any BMI decoder (Gilja et al., 2012). After using the Kalman filter in a BMI task, the decoder's predictions, corrected based on the assumption that the user intended to acquire targets as quickly as possible, are then used to train an updated Kalman filter, theoretically with a better linear representation of how neural activity maps to intended behavior. Several variations of this technique have been tested (Orsborn et al., 2012; Willett et al., 2018), all showing improvements in performance compared to decoder models trained without consideration for visual feedback in monkeys and humans. These studies have motivated a piecewise linear model to better train BMI decoders on the first try, without needing an intention training step (Willett et al., 2019). As these techniques all occur during training, which can take advantage of computationally-intense and power-hungry machines, they remain very relevant for low-power portable BMIs.

Another computational addition to BMI decoders is dimensionality reduction, where the hundreds of neural features being recorded can be compressed to just the features that best represent the neural population's activity (Churchland et al., 2012; Cunningham and Yu, 2014; Shenoy et al., 2013). Principal components analysis (PCA) is an example of a dimensionality reduction technique, where the PCA dimensions would first be trained, then the incoming neural

features would first be decomposed into the PCA dimensions in real-time prior to being decoded into behavior. This can have the benefit of denoising the neural data and may reduce power consumption in the decoder by reducing the number of features to process. PCA and similar techniques have also been used to address waveform amplitude and firing rate changes of neurons in motor cortex (Chestek et al., 2007, 2011), which had traditionally made all decoder models lose their accuracy over time despite signals still being available across years (Simeral et al., 2011). By realigning neural features across days using a manifold/subset of neural population dimensions (Gallego et al., 2020), latent factor analysis via dynamical systems (LFADS) (Pandarinath et al., 2018), or other techniques (Degenhart et al., 2020), high-performance decoding using the originally trained weights remains possible. However, it remains unclear if the cuts in computational complexity resulting from compression of neural features are great enough to warrant the additional computations inherent to the dimensionality reduction algorithm.

To summarize, the advent of more advanced processing computers and decoding algorithms has sparked a remarkable improvement in decoding performance across the past two decades. Today, linear decoders have been established as the tried-and-true method for predicting behaviors from brain activity. The computational simplicity of linear decoders makes them promising candidates for implementation on a portable and low-power BMI, as the algorithms are the appropriate complexity to fit on embedded processors, similar to what is in a deep-brain stimulator. Newer decoders, such as neural networks, have shown promise for the future capabilities of BMI decoders, but they require substantial reductions in computational complexity before they can be implemented full-time in BMIs. Although including dimensionality reducing techniques like PCA may add some substantial computations to a BMI, perhaps reducing the

number of input features to a decoder may result in an overall lower complexity and provide room for more complex algorithms to be implemented on portable and low-power BMIs.

1.2.3 Controlling Prosthetic Actuators

The final stage of a brain-machine interface is the prosthesis. The prosthesis can take many forms: a surrogate arm and hand that is the colloquial usage of the term prosthesis, or a tool, such as a computer or a speaker, that provides some special functionality to the user. Ultimately, the goal of the prosthesis is to execute the intentions of the user. The following paragraphs discuss the different prostheses currently being investigated (along with images in Figure 1.4) as well as their projections for usage outside of laboratory environments.

Virtual environments provide a cheap, customizable, and informative medium through which to test and use BMIs. By simulating realistic environments, humans and nonhuman primates have been able to control virtual arms and hands (Ajiboye et al., 2017; Irwin et al., 2017) with nothing but their brain activity, creating an expectation of capability when testing with real limbs. Additionally, as personal computers and smartphones have become prevalent in modern life, it becomes increasingly important for a person using a neuroprosthesis to be able to efficiently interact with those devices. Instead of controlling a prosthesis to then manipulate electronic devices, several groups have investigated skipping the middleman and decoding intended body movements directly into control signals for computers. Some of the first mainstream neuroprosthetic investigations provided nonhuman primates with the ability to control a computer cursor in multiple dimensions (Carmena et al., 2003; Serruya et al., 2002; Taylor et al., 2002a). More recently, monkeys and humans have been able to type messages on keyboards at high character rates (Gilja et al., 2015; Nuyujukian et al., 2016; Pandarinath et al., 2017), generally by using imagined arm movements to control a cursor by way of the BMI. One study in particular

enabled a person to control a tablet computer through several applications with an intracortical BMI, demonstrating the usefulness of brain-computer interfaces for people with paralysis (Nuyujukian et al., 2018). Restoring computer usage returns substantial independence to people with paralysis in terms of interacting with digital media. However, the inability to directly manipulate the real world may leave something to be desired for BMI-controlled computer users.

Despite the unnaturalness of the colloquial prosthesis, it does enable interaction with the real world, and robotic limbs are one step more natural. Often affiliated with people with amputations, modern robotic limbs can perform most of the functions of the natural hand in terms of motor control and sensation. Foundational studies demonstrated the ability to control various robotic arms just with the decoder outputs (Carmena et al., 2003; Chapin et al., 1999; Hochberg et al., 2006, 2012). Many recent studies have employed the Modular Prosthetic Limb (Johns Hopkins University, Applied Physics Laboratory, Baltimore, MD, USA), which supports both motor output (Collinger et al., 2013; Wodlinger et al., 2015) and sensory feedback (Flesher et al., 2016). There are a handful of other robotic hands that could fit brain-machine interface applications (LUKE Arm, Mobius Bionics, LLC, Manchester, New Hampshire, USA; i-Limb, Össur hf., Reykjavík, Iceland; TASKA Hand, TASKA Prosthetics, Riccarton, Christchurch, New Zealand; BeBionic Hand, Ottobock Healthcare, Duderstadt, Germany; Ability Hand, Psyonic, Inc., Champaign, IL, USA), but require the attachment of arms for full mobility. A central motivator for the use of robotic upper extremities is that they do not require surgical intervention for repairs, making them lower risk prostheses to return some function to people with paralysis. However, the robotic limb, which is not necessarily lightweight, would require hefty mounting hardware for attachment to a wheelchair or even to the user if future rehabilitation technology enables independent mobility. The addition of a robotic limb while maintaining the natural limbs of the person is bulky, reducing

Virtual   **Robotic**   **Stimulation**

A — Cursor

B — Keyboard

C — Arm

D — Arm

E — Arm

F — Surface Hand

G — Impl. Upper Ext.

H — Implantable pulse generator / Epidural electrical stimulation — Impl. Spine
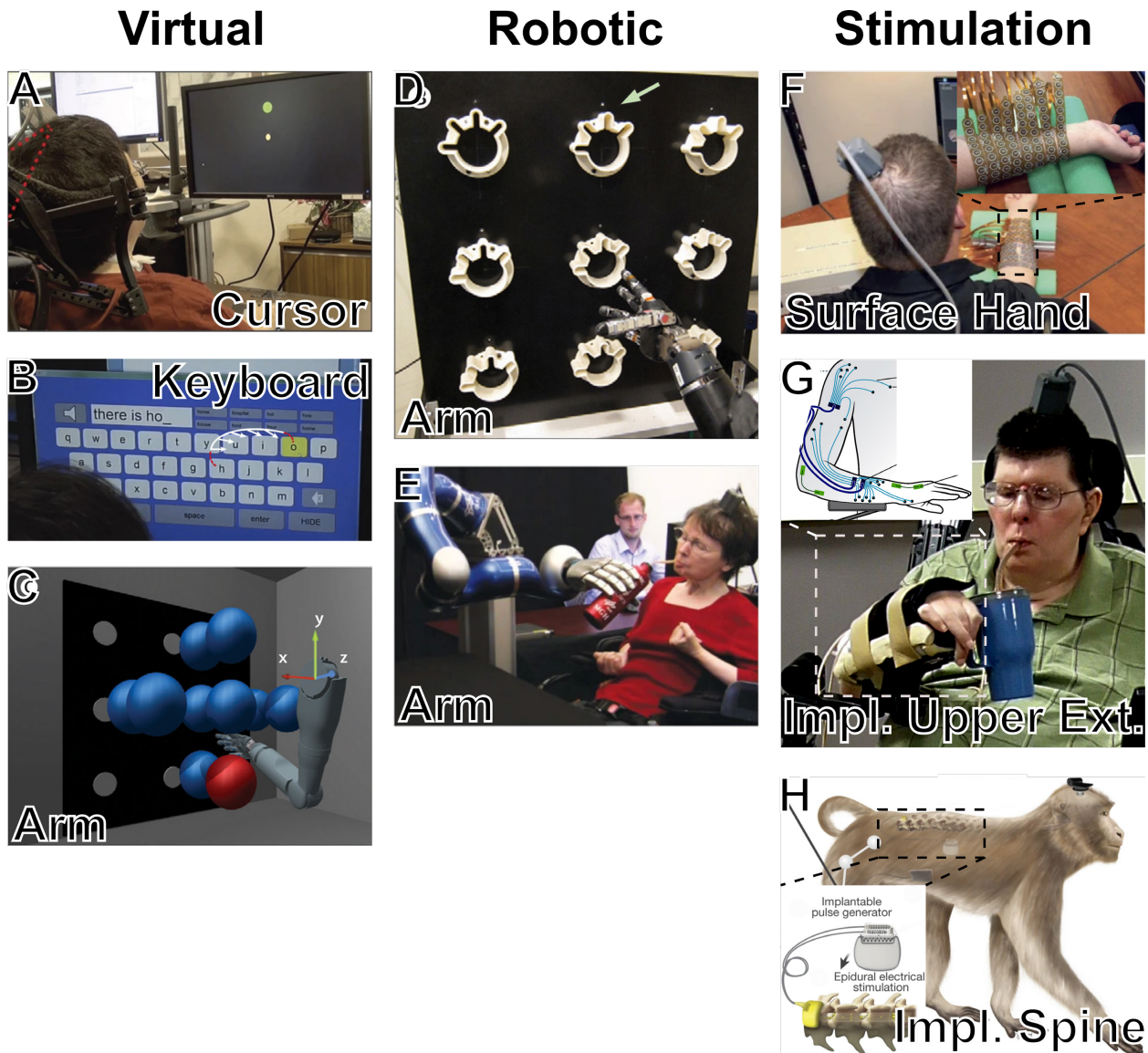
Figure 1.4 A variety of prosthetic actuators used in literature. Virtual prostheses include computer cursors (**A**, Gilja et al., 2015), keyboards (**B**, Jarosiewicz et al., 2015), and virtual limbs (**C**, Collinger et al., 2013). Robotic prostheses (**D**, Collinger et al., 2013, **E**, Hochberg et al., 2012) include various types of robotic arms. Stimulation prostheses activate muscular tissue via electrical stimulation on the skin surface (**F**, Bouton et al., 2016), via intramuscular electrodes (**G**, Ajiboye et al., 2017), or via spinal electrodes (**H**, Capogrosso et al., 2016).

maneuverability in smaller spaces and potentially making usage cumbersome. Further,

embodiment of a robotic limb is a difficult balance for engineers to maintain. Accurate and precise

motor control coupled with naturalistic sensory feedback that may result from continued research

may be sufficient for most to embody a robotic limb, but the current state of robotic arms and

hands do not match the capabilities of the natural upper extremity.

Consequently, several groups have investigated methods through which the natural limb can be reanimated for use, despite paralysis via the nervous system. Functional electrical stimulation (FES) is a technology that can reanimate paralyzed muscles by delivering electrical stimulation to a muscle or the nerve innervating a muscle. Coordinating contractions of several muscles via electrical stimulation can return function to any degree of freedom of the body. To date, several studies have demonstrated voluntary FES of the upper extremity using pattern recognition from residual functional muscle (Kilgore et al., 1989, 2008; Memberg et al., 2014; Peckham et al., 1980). More recently, groups have enabled nonhuman primates (Ethier et al., 2012; Moritz et al., 2008) and people with paralysis (Ajiboye et al., 2017; Bouton et al., 2016) to control their upper extremities using BMI-controlled FES. Recent work investigating simultaneous individuation of digits within the hand would work well with a BMI-controlled FES system, as the predictions of the finger movements could be used as a direct control signal for electrical stimulation (Aggarwal et al., 2008; Nason et al., 2021b). However, as mentioned previously, there are a number of challenges with recording brain activity during neuromuscular stimulation regarding the large amplitude artifacts each stimulation pulse generates. Fortunately, several means for mitigating the effects of artifacts have been presented. Intramuscular electrodes could be used in place of surface electrodes as they generate smaller amplitude stimulation artifacts in intracortical recordings (Young et al., 2018). Additionally, presence of the stimulation artifacts can often be predicted based on the status of the BMI, allowing artifact-free recording in between periods of stimulation. Otherwise, common-average referencing (Ludwig et al., 2009) or a linear regression reference (Young et al., 2018) have been shown to nearly completely eliminate the artifact in the intracortically-recorded signals for a relatively small increase in computational complexity.

There are a number of prosthetic devices that can be controlled by BMIs, ranging from computers to implantable hardware reanimating natural limbs. While present research aims to bring these technologies to full-time usage, perhaps the next stage of BMI-controlled prostheses will enable simultaneous usage of multiple functional devices.

1.3 Emerging Technologies

Existing technologies, if they were available in a wireless implantable form factor, can already begin to restore at least a few degrees of freedom of movement for at least a few years (Aflalo et al., 2015; Ajiboye et al., 2017; Bouton et al., 2016; Collinger et al., 2013; Hochberg et al., 2006, 2012; Pandarinath et al., 2017). This can be viewed as a promising start for the field when the first real-time control with visual feedback using multi electrode arrays was only two decades ago. There is again a wide variety of early pre-clinical technologies that may lead to large performance increases within the next two decades, some of which are showcased in Figure 1.5. There are opportunities for both integrated circuits and novel electrodes to enable order of magnitude improvements in brain machine interface performance. With the traditional neural signal circuit architecture described above, i.e. a low noise amplifier followed by a 16-bit ADC, followed by a method of spike extraction and often spike sorting, there are not a lot of remaining tools to improve the power consumption. Existing approaches are already fairly well optimized (Harrison et al., 2007). However, if you move outside of that architecture, for example by identifying low frequency correlates of spikes rather than spikes directly (Irwin et al., 2016; Lim et al., 2020; Nason et al., 2020) or by heavily optimizing sampling rates and communication bit rates (Even-Chen et al., 2020), order of magnitude reductions in power consumption are possible. Also, it is possible to use front ends with a very small number of transistors that may somewhat compromise signal quality in return for size, i.e. the "neural dust" approach (Seo et al., 2016), that
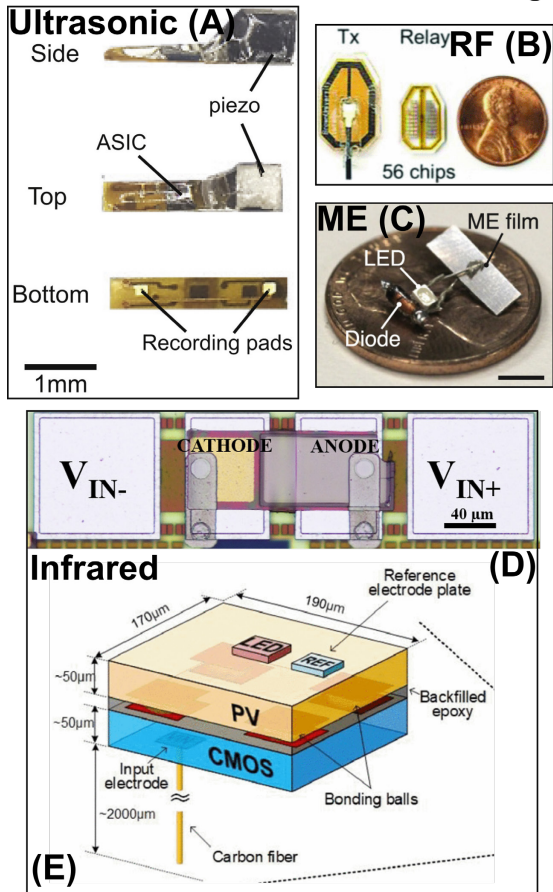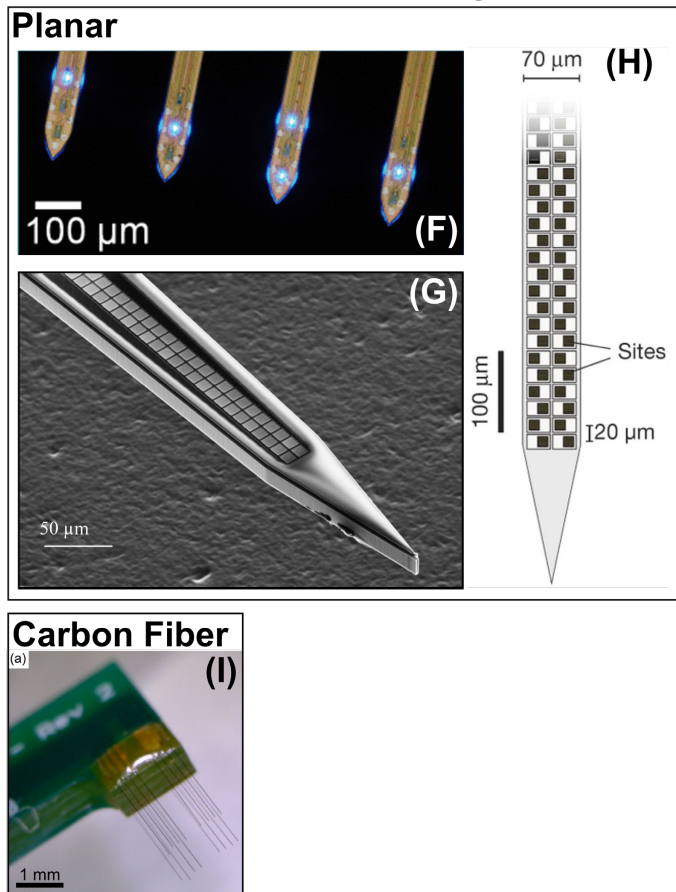
Figure 1.5 Emerging neural recording technologies. Novel small-scale wireless implantable electronic devices (*left*). (**A**) Ultrasonic neural dust (Seo et al., 2016). (**B**) Radio-frequency (RF) coupled Neurograins (J. Lee et al., 2019). (**C**) Magnetoelectric (ME) stimulators (Singer et al., 2020). (**D**) Infrared recording MOTEs (S. Lee et al., 2018). (**E**) Infrared recording probes (Lim et al., 2020). Novel electrode technologies (*right*). (**F**) High-density opto-electrophysiological silicon probes (Mendrela et al., 2018). (**G**) High-density oversampled silicon probes (Scholvin et al., 2016). (**H**) Neuropixels probe (Jun et al., 2017). (**I**) Carbon fiber microelectrode array (Patel et al., 2016).

could one day enable implanting many more channels. Specifically, prototype devices exist using

ultrasonic power (Seo et al., 2016), RF with optimized small antennas (Lee et al., 2019), magneto-

electric power (Singer et al., 2020) and optical power (Lee et al., 2018; Lim et al., 2020).

In terms of electrodes, the Utah array was developed in the late 90s (Nordhausen et al.,

1996). This is now in widespread use in humans, with 48 implants prior to 2018 (Bullard et al.,

2020). The robustness of this design has made it difficult to improve upon, since whatever

displaces it must have a lifetime measured in years (Chestek et al., 2011; Downey et al., 2018;

Simeral et al., 2011) and compete with many person years of existing safety data. However,

electrodes in pre-clinical testing have promising characteristics that could make their way into clinical use after a period of hardening over the next decade. Channel count is now far higher in rodent experiments than NHP or humans due to devices such as Neuropixels with 384 channels (Jun et al., 2017), and other high density silicon probes (Mendrela et al., 2018; Scholvin et al., 2016; Yang et al., 2020). Similarly, there are a large number of novel electrode designs that are smaller than neurons, after early work noted that this can enable a drastic reduction in bleeding and scarring (Nguyen et al., 2014; Seymour and Kipke, 2007). Specifically, "ultrasmall" electrodes have been created using carbon fibers (Guitchounts et al., 2013; Kozai et al., 2012; Patel et al., 2016), silicon carbide (Knaack et al., 2016), PDMS (Cohen-Karni et al., 2009), nanoelectronic thread (Luan et al., 2017), neuron-like electronics (Yang et al., 2019), parylene (Rodger et al., 2008), thermal drawing (Canales et al., 2015) and high density wire bundles (Obaid et al., 2020). This could one day enable multielectrode arrays that achieve very high channel count despite less overall damage than the 100 channel devices in use today. One can imagine a future where neural interfaces become a first line treatment for some illnesses rather than a last resort.

1.4 Summary of Thesis

In this thesis, I investigate the capabilities of low-power neural recording techniques to restore dexterous hand function in cases of hand paralysis.

In Chapter 2, I investigate the use of the 300-1,000Hz spiking band power (SBP) as a brain-machine interface feature for predicting fine hand movements. We have previously demonstrated that SBP can predict movement similarly to the standard 30kSps threshold crossing rate in basic behavioral tasks. However, the relationship between SBP and the underlying neural activity remains unclear, as does its capability of decoding complicated behavior. In simulation, I found that SBP is dominated by local single unit spikes with comparable or better spatial specificity and

maintains a higher correlation with the firing rates of lower signal-to-noise ratio units than threshold crossing rate. In online one-dimensional decoding of individual finger groups and offline two-dimensional cursor control in nonhuman primates, SBP always performed as well or better than threshold crossing rate, reaffirming the result of spatial specificity. These results suggest decoding performance can be enhanced while enabling drastic cuts in power consumption of neural interfaces.

In Chapter 3, I show that SBP enables end-to-end brain-machine interface compression into fully implantable device sizes. I first showed how optimizing firmware on an existing embedded brain-machine interface (Bullard et al., 2019) can enable it to record 96 channels of SBP in just 33.6mW, compared to one channel at 30kSps requiring 31.2mW. Then, I showed that the same embedded device with further firmware improvements can predict a nonhuman primate's closed-loop one-finger movements with higher performance than previously published using just 58.4mW. Additionally, the device can predict two-finger movements with good performance, not quite as high as our main brain-machine interface likely due to the reduction in recording signal-to-noise ratio. Finally, I demonstrated that a fully optimized integrated brain-machine interface can also predict closed-loop one- and two-finger movements in even less power, just 12.5mW. These results show that, with spiking band power, existing brain-machine interface technologies can be compressed into implantable packages with low enough power consumption for daily use.

In Chapter 4, I show that SBP can be used by a brain-machine interface to predict the simultaneous and independent movements of two finger groups in real-time. During online brain control, a ReFIT Kalman filter could predict a nonhuman primate's individuated finger group movements with mean acquisition times as low as 1.01s. Next, training ridge regression decoders with individual movements was sufficient to predict untrained combined movements, and vice

versa. Finally, I compared the postural and movement tuning of finger-related cortical activity to find that individual cortical units simultaneously encode multiple behavioral dimensions. These results suggest that linear decoders may be sufficient for BMIs to execute high-dimensional tasks with the performance levels required for naturalistic neural prostheses.

In Chapter 5, I show that low-power brain-machine interface techniques can be used to control functional electrical stimulation to restore continuous function to paralyzed hands. In a one-dimensional, continuous, one-finger target acquisition task, a nonhuman primate improved his success rate to 89% when using an SBP-enabled brain-controlled functional electrical stimulation system during temporary paralysis (1.4s median acquisition time) from 2.6% without prosthetic assistance (9.5s median acquisition time at the trial timeout). With the monkey fully anesthetized, I found that FES alone could control the monkey's fingers to rapidly reach targets in a median 0.8s, though there was some oscillation about the target using this simplified control scheme. Finally, when attempting to perform a two-finger continuous target acquisition task in brain-control mode following hand paralysis, acquisition time and success rate dropped substantially compared to before paralysis but could be recovered by performing recalibrated feedback-intention training one additional time. This suggests that BCFES can restore continuous finger function during paralysis using existing low-power technologies.

Finally, in Chapter 6, I will discuss the results of each study, present my projections for immediate next steps of this work, and discuss my visions for implantable brain-machine interfaces over the next few decades.

Chapter 2 Obtaining Single-Unit Spiking Activity with Power-Efficient Spiking Band Power to Exceed the Decoding Performance of Standard Neural Features

A version of this chapter "A Low-Power Band of Neuronal Spiking Activity Dominated by Local Single Units Improves the Performance of Brain-Machine Interfaces" was published in Nature Biomedical Engineering in 2020 (Nason et al., 2020)*.*

2.1 Introduction

Brain-machine interfaces have the potential to treat many neurological diseases and disorders. By translating brain activity into user intent, they can enable self-feeding, independent movement, and computer control to people with paralysis (Collinger et al., 2013; Gilja et al., 2015; Hochberg et al., 2012; Velliste et al., 2008). Some of these incredible advancements were demonstrated more than a decade ago, but they have not yet been translated to wide-spread clinical use. A fundamental barrier to clinical systems is the high power and computation requirements for high bandwidth neural recording. Consequently, modern high-performance solutions have required transcutaneous connections to a large cart of computers recording, processing, and decoding the neural activity. This renders existing clinical implantable recording systems, such as closed-loop deep-brain stimulation implants, unfeasible as replacements, since they record from too few channels with too small of a bandwidth.

To predict complicated behavior, brain-machine interfaces typically use estimates of the firing rates of relevant neurons. To extract these, they record from approximately 100 transcutaneous electrodes at 20kSps to isolate units via thresholding. In an attempt to eliminate the

infection risks associated with persistent transcutaneous connections and free the user from being tethered to the necessary computational hardware, several groups have investigated wireless, fully implantable neural interfaces. In particular, the devices presented by Rizk *et al.* and Borton *et al.* wirelessly transmitted broadband neural recordings from ~100 electrodes (Borton et al., 2013; Rizk et al., 2009). Unfortunately, power delivery generated unsafe tissue temperatures and the devices did not have decoding capabilities to translate the neural activity, which would require the user to remain within the wireless range of the processing computers. In the case of controlling functional electrical stimulation, additional wireless links would be required to interact with the implanted stimulating hardware, requiring more power-hungry, wireless, implantable devices.

To resolve the power consumption issue, many groups have developed application-specific integrated circuits (ASICs) to optimally perform the necessary computations (Abdelhalim et al., 2013; Aziz et al., 2009; Biederman et al., 2013; Borna and Najafi, 2014; Chae et al., 2008; Gao et al., 2012; Harrison et al., 2007; Karkare et al., 2013; Limnuson et al., 2014; Park et al., 2018; Shahrokhi et al., 2010; Wattanapanitch and Sarpeshkar, 2011). In particular, Harrison *et al.* demonstrated a promising 8mW wireless implantable chip accommodating acquisition, spike sorting, and transmission of 15.7kSps neural data from 100 electrodes (Harrison et al., 2009). However, the amplifiers used were nearly optimal between the noise efficiency factor (4.0) and current usage (16µA/channel), implying devices must optimize power consumption through alternative means (Harrison and Charles, 2003). Additionally, while radio frequency transmission of neural data and power has become a common solution for these groups, Seo *et al.* found that it cannot accommodate micron scale recording modules due to power loss through tissue and the inefficiencies of integrated components (Seo et al., 2013). It seems even the optimizations of customized integrated circuitry may still require new techniques to further cut power consumption.

These challenges have resulted in very few low power integrated solutions being functional enough for animal use before being considered for safety verification in humans.

One potential way to lower power is to identify neural features that maintain the decoding capabilities of intracortical spike recordings but using less bandwidth. Using threshold crossing events is a common solution (Ajiboye et al., 2017; Capogrosso et al., 2016; Ethier et al., 2012; Gilja et al., 2015; Irwin et al., 2017; Pandarinath et al., 2017; Shanechi et al., 2017; Wodlinger et al., 2015), but this only saves power after digitization, leaving the high bandwidth, low noise amplifier included. Local-field potentials have also been demonstrated to contain much movement information from recordings as invasive as threshold crossings (Bansal et al., 2011; Heldman et al., 2006; Mollazadeh et al., 2011). The frequency content of such signals is appropriately low-bandwidth for power reductions, and several groups have exhibited their usefulness towards open-loop decoding (Aggarwal et al., 2013; Baker et al., 2009; Flint et al., 2012a, 2012b; Perge et al., 2014; Stark and Abeles, 2007; Wang et al., 2014; Zhuang et al., 2010) and closed-loop decoding (Flint et al., 2013; Gilja et al., 2015; Milekovic et al., 2018; Pandarinath et al., 2017; So et al., 2014; Stavisky et al., 2015). However, due to their intracortical nature, it is inconclusive how much of the performance is impacted by low-pass filtered spikes instead of potentials alone. Therefore, although most studies show that local-field potential decoders can achieve peak performance levels comparable to but lower than spike-based features, we do not know if that is an artifact of spikes present in the signals.

Electrocorticography (2kSps sampling rate) has also been proposed as a less invasive neural feature for brain-machine interfaces. Several groups have shown open-loop (Chao et al., 2010; Chestek et al., 2013; Flint et al., 2017; Kubánek et al., 2009) and closed-loop (Hotson et al., 2016) decoding, and the average correlation coefficient of these studies is 0.67. While comparable

to spike-based neural features, these tasks were simple, where the multiple degrees of freedom (if present) moved mostly along one dimension to minimally varying displacements. This suggests the decoders were dominated by the presence of movement. It is unclear how well electrocorticography decoders will perform at predicting more complex tasks that have systematically varied target locations and displacements, such as the continuous center-out-and-back or the point-to-point tasks.

Another interesting neural feature is the integration of multiple unit activity, sometimes called "spiking band power" (SBP), which was originally suggested in the 1960s before the age of cheap electronics, leaving it uninvestigated during the rise of brain-machine interfaces (Schlag and Balvin, 1963). More recently, Stark and Abeles demonstrated that the power in the 300-6,000Hz band is not just equivalent, but superior at predicting movement direction compared to local field potentials, threshold crossing rate (TCR), and sorted unit firing rate (Stark and Abeles, 2007). Further, our group has demonstrated that the bandwidth over which the signal is integrated can be drastically lowered to 300-1,000Hz while maintaining 95% of the decoding performance (Irwin et al., 2016). However, these studies only investigated offline decoding of relatively simple tasks, where population activity may decode as well as single units. The spatial specificity of SBP is not well understood, which makes the ultimate decoding limitations difficult to predict from fundamentals.

In this manuscript, we use a combination of simulated and *in vivo* neural recordings from nonhuman primates and rats to compare the specificity of the 300-1,000Hz spiking band with signals containing spikes. These simulations indicate that SBP results primarily from spike waveforms themselves, weighted by their amplitude, which creates a highly spatially specific signal. Additionally, the simulations suggest that the spiking band can extract accurate spike

information at low signal-to-noise ratios (SNRs) below typical thresholds, implying that SBP may represent neural activity that can neither be extracted by TCR nor sorted units. We believe these analyses explain the findings of our past work and some of the work of Stark and Abeles (Irwin et al., 2016; Stark and Abeles, 2007). Further, the outcomes of these simulations suggest that this very low power approach will fare as well or better than unit-based methods at decoding more complex tasks. Thus, we evaluate this claim using a variety of decoded tasks in nonhuman primates: recalibrated feedback intention-trained (ReFIT) Kalman filters to control one-dimensional index and middle-ring-small fingers in closed-loop, support vector machines to classify two-dimensional individuated movements of unrestricted fingers in open-loop, and standard Kalman filters to predict two-dimensional center-out-and-back arm reaches in open-loop.

2.2 Methods

All procedures were approved by the University of Michigan Institutional Animal Care and Use Committee.

2.2.1 Array Implants

We implanted three male rhesus macaques with Utah microelectrode arrays (Blackrock Microsystems, Salt Lake City, UT) in the hand area of primary motor cortex, as described previously by Irwin *et al.* and Vaskov *et al.* (Irwin et al., 2017; Vaskov et al., 2018). Pictures of the implants are illustrated in Figure A.2. Only motor cortex arrays were used in this study, and only monkey N's most rostral array was used.

Additionally, we obtained previously recorded offline datasets (L120502 from Kao *et al.* 2017 and J121009, a center-out-and-back arm reaches behavioral task performed by monkey J using the same implants as described in Kao *et al.* 2017) for our two-dimensional cursor control

analysis. All surgical and animal care procedures for these animals were performed in accordance with National Institutes of Health guidelines and were approved by the Stanford University Institutional Animal Care and Use Committee. Monkeys J and L were adult male rhesus macaques implanted with 96 electrode Utah arrays using standard neurosurgical techniques. Electrode arrays were implanted in primary motor cortex (M1) as visually estimated from local anatomical landmarks. Monkey J had two arrays, one in M1 and one in dorsal premotor cortex (PMd; unused in this study), while Monkey L had one array implanted on the M1/PMd border.

We also obtained recordings using carbon fiber electrodes in the rat motor cortex (Patel et al., 2016). Each carbon fiber (~6.8µm) was insulated with parylene-C, cut to expose the carbon recording surface, and then coated to reduce impedance to near 30kΩ (Kozai et al., 2012; Patel et al., 2016). We obtained acute electrophysiology recordings from layer V of one male wild-type Long Evans rat's motor cortex following previously established methods (Kozai et al., 2012). Three minutes of electrophysiological data were acquired from two separate carbon fibers using a TDT RX5 Pentusa recording system (Tucker-Davis Technologies, Alachua, FL).

2.2.2 Feature Extraction

All processing was done in Matlab (Mathworks, Natick, MA), except where noted.

The band power neural features presented here are similar to those used by Schlag and Balvin and Stark and Abeles, but in less computationally expensive forms (Schlag and Balvin, 1963; Stark and Abeles, 2007). The raw data were collected from the monkeys using a Cerebus v1.0 using firmware version 6.03.01.00 (Blackrock Microsystems) digitizing at 30kSps and from the rat using the RX5 Pentusa digitizing at 24.414kSps. We first applied a 2nd order Butterworth filter to the raw data with a 300-1,000Hz pass-band for SBP. Then, we extracted the magnitude of the signal and downsampled it to 2kSps or 2.4414kSps for the Cerebus or RX5 Pentusa,

respectively. Lastly, we smoothed this signal using a 50ms Gaussian window for simulations or averaged the signal in non-overlapping 50ms bins for open-loop decoding.

For closed-loop decoding, we extracted SBP differently for simpler processing in real-time. We configured the Cerebus to band-pass filter the raw signals to 300-1,000Hz using the Digital Filter Editor feature included in the Central program (Blackrock Microsystems), then sampled at 2kSps for SBP. The continuous data was streamed to a computer running xPC Target (Mathworks), which took the absolute value of the incoming data and averaged it in 50ms bins for decoding.

To extract TCR in simulation, we first high-pass filtered the raw, simulated 30kSps recordings using a zero-phase 2nd order Butterworth filter with a 250Hz cutoff frequency. We set a threshold at -3.75RMS and smoothed the binary result using a 50ms Gaussian window. We optimized the -3.75RMS threshold for Figure 2.2 from the threshold that achieved similar low firing rate prediction performance to SBP (we found more negative thresholds further reduced TCR's low SNR prediction performance, which, at all thresholds, was worse than SBP).

To decode TCR, we configured the Cerebus to extract voltage snippets that crossed a -4.5RMS threshold. Then, these waveforms were streamed to the computer running xPC Target, which logged the time of each spike's arrival (for later open-loop decode processing) and counted the number of spikes in 50ms bins (for closed-loop decoding).

We also sought to compare SBP to a lower-bandwidth form of TCR to validate the improvements in decode performance were not artifacts of the restricted noise bands. To extract low-bandwidth TCR, we thresholded the absolute-value of the 300-1,000Hz band. We chose to threshold only the spiking band since our primary goal with using SBP is to reduce front-end bandwidth to save the most power, with a secondary goal of maintaining decode performance.

When comparing SBP to sorted units to mimic the multiple unit simulation, we imported the relevant recordings into Offline Sorter (Plexon, Dallas, TX). Then, we high-pass filtered the recordings with a 4-pole Butterworth filter with a cutoff frequency set to 250Hz. To sort units of all amplitudes for validation of the multiple unit simulation results, we manually chose a reasonable threshold and sorted the spikes crossing the threshold individually and in combination with principal component analysis and visual inspection. For the power spectrum analyses we performed, we decided it was necessary to extract the averaged units without any influence from the lower frequency, higher signal power 1-300 Hz band. The random phases of the spikes in that band could confound the averaged spike waveforms, despite the possibility that action potentials could contain power in that band.

To sort units for offline decoding, we imported into Offline Sorter spike waveforms and timings extracted by the Cerebus with a -4.5RMS threshold during a behavioral task. Then, we eliminated all waveforms which were clearly artifactual, determined by extraneously large amplitudes, extraneously fast timescales, or waveform shape. We manually sorted the remaining waveforms into single units by analyzing the three strongest principal components, the density of waveforms in the principal component space, and the visual patterns of the waveforms. For all channels with distinguishable units, we added one additional unit, termed the neural hash in this manuscript, containing the remaining unsorted waveforms that were not deemed artifact. After unit isolation, we synchronized the timings of the waveforms of each unit with the behavioral data and binned each unit's spiking rates in 50ms bins.

In all simulations, we compared each estimate of firing rate with the true firing rate of each simulated unit ("True FR" in Figure 2.1 and Figure 2.2). We defined true firing rate as the indicator function of the initiation of a spike smoothed by a 50ms Gaussian window, with the timings of all

simulated spikes known based on the methods detailed in the subsequent paragraph. Additionally, several simulations were also performed with 100ms Gaussian smoothing windows or exponential smoothing windows of various time lengths for all features, which yielded similar results and are subsequently not presented.

2.2.3 Simulation Design and Parameters

In Matlab, we simulated 30kSps recordings of neural activity in order to investigate the origins and characteristics of SBP. We simulated spontaneously firing neurons as point current sources with an action potential waveform averaged from a 3ms sorted unit from one of monkey W's intracortical recordings digitized at 30kSps. To emulate spontaneous neural firing patterns, we first selected a reasonable firing rate, FR, representative of typical neural activity and a simulation duration, T. Then, we subtracted the total number of samples occupied by spikes (i.e. 3ms of samples times the firing rate times the duration, or $3 \times 30 \times FR \times T$) from the $30,000 \times T$ samples in each simulation and randomly partitioned the remaining samples into $FR \times T + 1$ parts. Then, these randomly-sized partitions were placed between spikes to generate a random spiking pattern with known timings. As such, simulated single unit recordings never had overlapping spikes. When simulated spikes from multiple units overlapped in time on one simulated recording, we superimposed the recorded voltages as was done by Lempka *et al.* (Lempka et al., 2011).

We simulated our recording electrode as a point current sink. For a noise model, we generated additive 6.23μV thermal white noise as was determined typical of intracortical recording simulations by Lempka *et al.* (Lempka et al., 2011). We found this value similar to the noise level of our Cerebus recording hardware. We did not include any biological noise since such a noise source with frequency characteristics similar to the spikes being simulated could confound the results. Thus, the same simulated 30kSps signals, 6.23μV noise included, were used as inputs to

both the TCR and SBP feature extraction pipelines. For simulations, we defined SNR as the maximum of the absolute value of the noiseless signal divided by the RMS value of the isolated noise:

$$SNR = \frac{\max(|signal|)}{RMS(noise)} \tag{2.1}$$

This allowed us to evaluate simulated signal SNRs for unrectified broadband recordings and their corresponding rectified SBP using the same metric. Additionally, we have also performed all simulations using another SNR definition based on signal RMS to better represent the total spiking signal power of a recording (results in Appendix A section titled RMS-Based SNR Analysis):

$$SNR = \frac{RMS(signal)}{RMS(noise)}$$

To transition between a simulated unit's SNR and the estimated displacement from a recording electrode, we applied a variation of the monopole equation:

$$r = \frac{I_0}{4\pi\sigma \times SNR \times RMS(noise)} \tag{2.2}$$

where r is the estimated displacement between the point current source neuron and the point current sink recording electrode, I_0 is the maximum of the absolute value of the current source which we set to a realistic 5nA, and σ is the simulated conductivity of gray matter of 0.27S/m from Slutzky *et al.* (2010) (Slutzky et al., 2010). By this derivation of the monopole equation, we assume that displacement is inversely proportional to SNR, as was suggested by the simulation results of Holt and Koch (Holt and Koch, 1999). Thus, to convert to voltage when necessary, we multiplied the SNR value of each sample by the 6.23μV noise RMS level according to equation (5.3).

All simulations in this work were performed 100 times to avoid any effects of random noise and random spike timings, and the averages of those simulations are presented in these results. Correlations were computed from two window lengths into the simulated recording to two window

lengths from the end of the simulated recording to avoid any impacts of onset and offset transients on the filters. Each simulated neural recording was five seconds long, resulting in $n = 150{,}000$ samples at 30kSps and $n = 10{,}000$ samples at 2kSps per simulated recording.

2.2.4 Behavioral Task

We trained monkeys L, W, and N, to acquire virtual targets with virtual fingers by moving their physical fingers, as illustrated in Figure A.1. During all sessions, the monkeys sat in a shielded chamber with their right arms fixed at their sides flexed at 90 degrees at the elbow, resting on a table. Monkeys W and N had their hands placed in the manipulandum described by Vaskov *et al.*, while monkey L had his hand lightly restrained facing left in a simple stanchion (Vaskov et al., 2018). Finger movements were measured by flex sensors (FS-L-0073-103-ST, Spectra Symbol) attached to both doors of the manipulandum or directly to the index finger of monkey L's right hand. Position measurements were recorded by a computer running real-time xPC Target. The computer monitor directly in front of the monkey displayed a virtual monkey hand model (MusculoSkeletal Modeling Software) which was controlled by the xPC Target computer to mirror the monkey's hand movements (Davoodi et al., 2007).

For closed-loop decoding experiments presented in this work, monkeys W and N performed trials in which either the middle-ring-small door or the index door was locked at full extension and the other was free to move, or both doors were locked together to keep the monkeys' fingers moving together. Later, monkeys W and N were trained to acquire two targets simultaneously, one specifically with their index finger and the other specifically with their middle-ring-small fingers, to classify which fingers are being moved for each target acquisition. Both groups of fingers moved freely and independently in this movement style. Monkey L had no finger restrictions and voluntarily moved all four fingers together.

Each trial began with a spherical target(s) appearing along the path of the virtual finger(s) of interest, where each target occupied 15% of the full arc of motion of the virtual finger(s). For a successful trial, the monkey was required to move its fingers such that the corresponding virtual finger(s) of interest moved into the target(s) and remained there for 500-750ms. Upon successful trial completion, the monkeys received a juice reward. For closed-loop decoding, the targets were presented in a center-out pattern for one finger group, as detailed by Vaskov *et al.* (Vaskov et al., 2018). For finger movement classification, four target arrangements were presented in a pattern similar to center-out, where both targets defaulted to the rest position (50% between full extension and full flexion) and, randomly, either only the index, only the middle-ring-small, or both groups had targets presented at 80% flexion on the non-rest trials.

2.2.5 Open-Loop Decoding

For all open-loop decoding, we applied a standard position/velocity Kalman filter with a position/velocity neural tuning model and optimizations as described by Irwin *et al.* (2017) to predict the movements the monkeys made in the behavioral task (Irwin et al., 2017). We trained Kalman filters individually for each neural feature and conducted 10-fold cross-validation. In the analyses presented in Appendix A section titled Open-Loop Decoding Analysis, we performed open-loop decodes of all six neural features in Matlab: SBP, wideband SBP (i.e. 300-6,000Hz passband from Stark and Abeles 2007 (Stark and Abeles, 2007)), low-bandwidth TCR, TCR, single unit firing rate, and single unit firing rate with neural hash. All decodes used bin sizes of 50ms, no common-average referencing, the optimal quantity of additional historical bins per channel or unit determined for that set of data, the optimal bin lag between the neural activity and behavior determined for that set of data, and all available channels as previously described.

Additionally, to most fairly compare low-bandwidth TCR as the alternative power-efficient feature, we found the optimal RMS threshold for each set of data to compare to the other features. For each decode, we calculated Pearson's correlation coefficient and the root-mean-squared error between the predicted and actual finger positions using Matlab. We transformed all correlation coefficients between the predicted and actual movements to their Fisher's z-scores and performed two-tailed two-sample z-tests to compare the statistical differences between all features. To determine statistical significance between root-mean-squared errors of the different features, we calculated the root-mean-squared errors of bins of 10 distinct samples from the predictions/true movements, yielding vectors of errors representing subsets of the predictions. Then, between these vectors, we performed one-tailed two-sample Wilcoxon rank-sum tests to determine significance.

To estimate the value of each channel or unit for each decoder, we omitted each channel or unit from the decoder and calculated the corresponding correlation between the predicted and actual finger positions. Then, we subtracted the correlation of the decoder that had access to all neural data from the correlation of the decoder missing one channel or unit to gauge the impact on performance (Wahnoun et al., 2006).

2.2.6 Closed-Loop Decoding

To quantify how well a system decoding SBP would perform in a neuroprosthetic application, we gave monkeys W and N visual feedback of the decoders' outputs during the behavioral task. For each closed-loop experimental session, the monkeys began by completing at least 350 target presentations with the virtual hand controlled directly by the physical hand's movements. The monkeys were required to acquire and hold the target for 750ms continuously for a successful trial. The behavioral data (i.e. finger kinematics) were measured synchronously with the neural features by the xPC Target computer. Then, we trained a standard Kalman filter on this

data. Subsequently, we used it to predict the finger movements exclusively from the neural feature and actuate the virtual hand accordingly in real-time. These types of trials are what we call closed-loop experiments. For a successful closed-loop trial, the virtual finger had to remain within the target continuously for 500ms.

The monkeys completed at least 250 target presentations using the standard Kalman filter. Then, we trained a ReFIT Kalman filter using the data from the standard Kalman filter set of trials. ReFIT Kalman filter recalibration followed the same procedure and optimization as described by Vaskov *et al.* (2018) (Vaskov et al., 2018). Finally, we enabled the monkeys to complete at least 150 target presentations using the ReFIT Kalman filter in closed-loop form. An entire experimental session, from the initial hand-controlled training data through standard Kalman filter and ReFIT Kalman filter training and testing, occurred within a few consecutive hours, always on the same day.

We compared success rate, target acquisition time, and Fitts' Law bit rate of closed-loop decoders using exclusively SBP, TCR, or wideband SBP. All decoders used a bin size of 50ms, a bin lag of 0, 1, or 2 bins for monkey W and 2 or 3 bins for monkey N, and channel masking as needed, which were determined optimal for each decoder on each day during training (Irwin et al., 2017). We calculated Fitts' Law bit rate using the following formula:

$$\text{Bit Rate} = \frac{\log_2\left(1 + \frac{D_i - R_T}{2 \times R_T}\right)}{t_{acq}}$$

where D_i is the *n*-dimensional displacement between the finger's starting position to the center of the target, R_T is the radius of the target, and t_{acq} is the time the monkey took to successfully complete the trial minus the required holding time (Thompson et al., 2014). All failed trials were given a bit rate of 0. In order to best compare closed-loop decoding performance among the three

neural features, we aligned the training sets of data as best as possible. The SBP and TCR standard Kalman filters were trained on the same set of synchronized behavioral data and neural activity. As the Cerebus used could only apply one filter at a time to continuous data and the xPC Target computer could not additionally filter all incoming data quickly enough in real-time, the SBP and wideband SBP standard Kalman filters were compared using different synchronized behavioral and neural data acquired consecutively within the same experiment. We randomly switched the neural feature for which we would synchronize behavioral data first in a day's experiment to avoid any biases associated with the cumulative amount of training time for each set.

The closed-loop decode statistics were calculated with single-tailed two-sample *t*-tests between all combinations of virtual hand control methods. The statistics resulted from monkey W performing 4,659 trials using his hand, 1,084 trials using the TCR decoder, 2,042 trials using the SBP decoder, and 2,428 trials using the wideband SBP decoder and monkey N performing 1,556 trials using his hand, 1,256 trials using the TCR decoder, and 1,915 trials using the SBP decoder. Closed-loop experiments were performed in eight sessions across one month for monkey W and four sessions across nine days for monkey N.

2.2.7 Finger Classification

To characterize SBP in a more complex neural decoding scenario, we classified which of index or middle-ring-small groups were moving in a given trial. A support vector machine with a Gaussian kernel function and L_1 regularization was chosen as the classifier and tested using 10-fold cross validation in Matlab. Only trials in which the monkeys flexed one finger after both fingers were at the rest position were used for training and testing. The classifier translated 160ms bins of neural activity from active channels from 50ms before to 250ms after movement onset with 20ms steps between the start of each bin (Aggarwal et al., 2008). Movement onset was determined

by smoothing the moving finger's position using a 2nd order Savitzky-Golay filter with a frame duration of 200ms, deriving the velocity, and finding the first point of incline towards the maximum velocity, which is assumed to be involved in the initial movement towards a target. The results in Figure 2.4c were based on 88 trials from monkey W and 95 trials from monkey N.

2.3 Results

2.3.1 Visualizing Single Spikes in the Spiking Band

We began by investigating the frequency content of typical motor cortical biphasic spike waveforms to find a spectral basis for the 300-1,000Hz pass-band optimized by Irwin *et al.* (Irwin et al., 2016). We identified two units representing narrow and broad spikes, plotted in Figure 2.1a, from 30kSps recordings from a 96 channel Utah microelectrode array (Blackrock Microsystems, Salt Lake City, UT) implanted into primary motor cortex in a rhesus macaque (monkey W, see Figure A.2 for placement). Analyzing both narrow (putative inter-neurons with higher frequencies) and wide (putative pyramidal neurons with lower frequencies) spikes, which are expected waveforms and have been reported previously in premotor and motor areas in rhesus monkeys (Kaufman et al., 2010, 2013), should approximate the range of power belonging to the 300-1,000Hz spiking band for any motor cortical spike. We extracted the averaged spikes, processed their interpolated fast-Fourier transforms, and approximated the total amount of signal power belonging to each of the 1-300Hz, 0.3-1kHz, and 1-15kHz bands. Although most of the power in both spikes existed above 1kHz, we found that the 300-1,000Hz band contained a substantial amount, between 25-45% of the total, as shown in the green band of Figure 2.1a.

With substantial power in the spiking band, we expect at least high SNR single unit recordings to maintain strong spike representations after filtering to 300-1,000Hz. Figure 2.1b shows a simulated recording of the solid line unit from Figure 2.1a (which had the least amount of
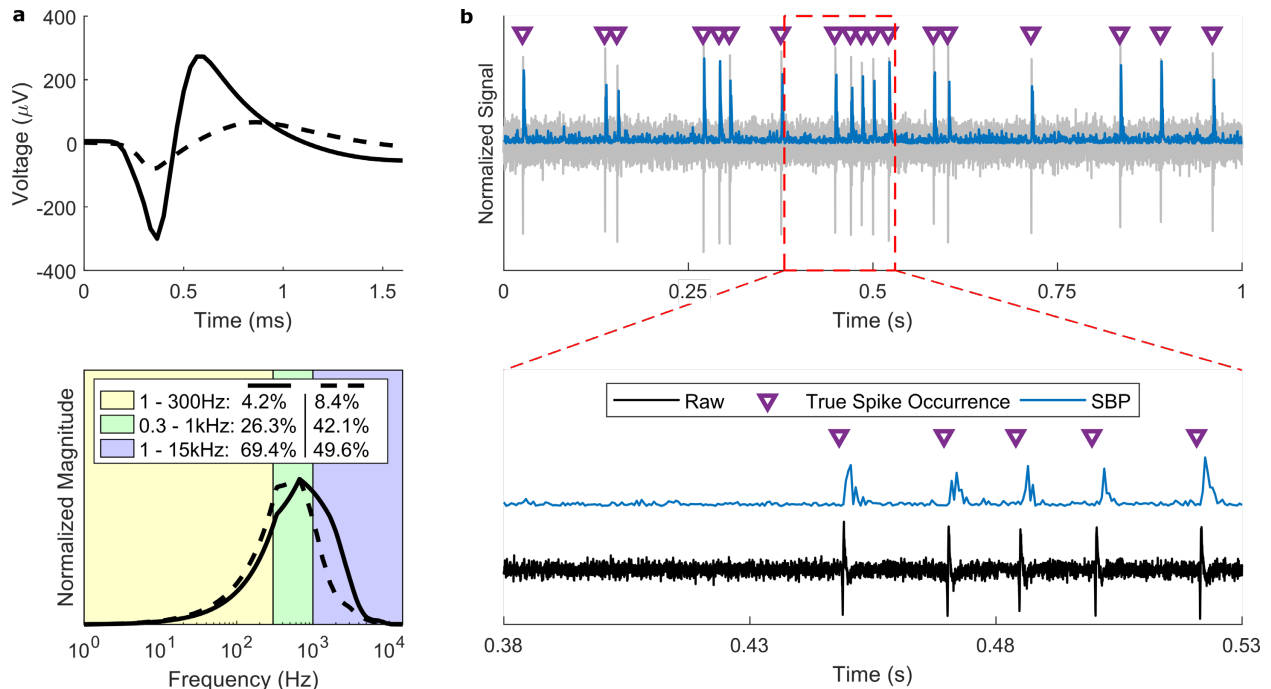
Figure 2.1 Representation of spikes in the 300-1,000Hz band. (**a**) Frequency spectrum of two averaged neural spikes extracted from intracortical nonhuman primate recordings. *Top:* Averaged neural spike waveforms. *Bottom:* Normalized interpolated fast Fourier transform of the above waveforms split by relevant frequency bands. The numbers in the legend indicate the amount of power belonging to each frequency band, with the solid line represented by the left most set of percentages. (**b**) Single unit simulated recording of the solid line unit in (**a**) at an SNR of 10. *Top:* Raw simulated noisy signal (gray) with spike occurrences (purple triangles) and SBP (blue) overlaid. *Bottom:* A snippet of the top plot with the SBP manually offset vertically to avoid occlusion.

power in the spiking band) firing spontaneously at an SNR of 10. The purple triangles represent

the true simulated spike occurrences of this unit, while the blue trace represents the absolute value

of the raw simulated signal filtered with a 300-1,000Hz 2nd-order Butterworth filter, re-sampled at

the Nyquist sampling rate of 2kSps. After smoothing both signals with a 50ms Gaussian window

filter, SBP exhibited a 0.95 correlation with the true firing rate of the unit. In the unsmoothed

signal energy on the bottom of Figure 2.1b, there is a clear burst at each instance of a spike with a

small lag resulting from the filter's non-zero phase characteristics. This resembles an envelope of

the broadband signal, showing how the neural activity is captured on a spike-by-spike basis. While

this first example represents a best-case-scenario, the SBP trace does not look unusual for a single

unit electrode and starts to suggest that the spikes are dominating the SBP signal. Interestingly,

though only obvious to us in hindsight, the SNR of the spiking band trace in Figure 2.1b is

approximately twice as high (SNR=19.52) as the SNR of the broadband trace. The small passband of SBP cuts the noise level by a factor of ~2 while taking advantage of the spike's energy above and below zero, suggesting SBP may be more robust to noise than single-ended broadband TCR.

We further explored this idea by simulating single neuron recordings under a variety of conditions to estimate the operational space of SBP for predicting true unit firing rates and compared it to that of TCR. In these simulations, the SBP and TCR feature extraction pipelines received the same simulated recording, including identical noise and spike timings. Additionally, our simulations did not estimate true firing rate by thresholding; rather, the true firing rate was known absolutely because the timings of the simulated spikes were known. Figure 2.2a shows these results in a grid for optimized TCR and SBP, where color represents the correlation with smoothed true firing rate. We found that if either the SNR or the firing rate is very low, the correlation is low for both features and vice versa (lower left and upper right corners). However, surprisingly, SBP maintains high correlation for lower SNRs and lower firing rates in comparison to optimized TCR.

However, we consider it unfair to directly compare SBP and TCR, since the reduced SBP passband eliminates much of the higher frequency noise components. Thus, we also compare SBP to a low-bandwidth form of TCR, which is extracted by thresholding the spiking band power. We found that SBP achieved similar low firing rate and low SNR prediction performance to low-bandwidth TCR. This suggests that the small bandwidth is indeed a major contributing factor towards the single unit firing rate estimation performance.

An example from this region of high SBP and low-bandwidth TCR correlation but low TCR correlation is shown in Figure 2.2b, where we recreated the same recording from Figure 2.1b but with an SNR of 2.25. SBP in blue and low-bandwidth TCR in red maintained relatively high
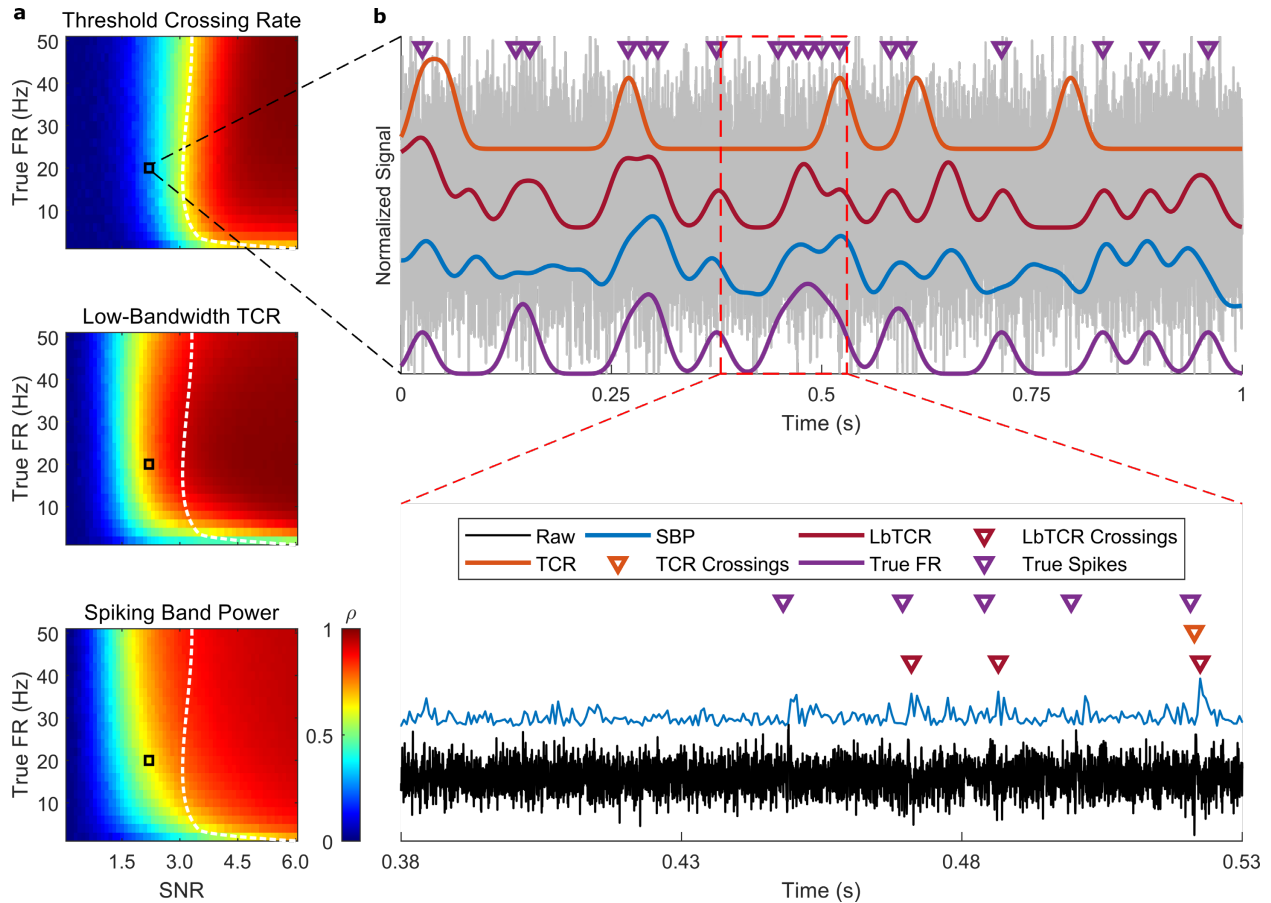
Figure 2.2 SBP, threshold crossing rate (TCR, optimized threshold set at $-3.75 \times$ root-mean-square [RMS]), and low-bandwidth TCR (LbTCR, optimized threshold set at 2.75RMS) prediction of true firing rate (True FR). (**a**) Predictive capability of the three features of various true firing rates from various recording SNRs. The color bar represents Pearson's correlation coefficient between the smoothed neural feature and the unit's smoothed true firing rate. Zero-phase smoothing used a 50ms Gaussian window. The black boxes indicate the simulated recording conditions in (**b**), and the white dotted lines track the yellow region of the TCR plot to better compare performance. (**b**) The same simulated unit plotted in Figure 2.1b but with a 2.25 SNR. Spike occurrences are overlaid as purple triangles along with smoothed TCR (orange), smoothed LbTCR (red), SBP (blue), and true firing rate (purple).

correlations with the unit's true firing rate in purple (0.62, 0.69 respectively) and are rarely falsely "activated" by noise. In comparison, the firing rate predicted by crossings of an optimized $-3.75 \times$ root-mean-square (RMS) threshold in purple had a 0.34 correlation and more frequently missed true spikes. Although this may seem unfair to TCR detecting just one of the five spikes in the bottom of Figure 2.2b, we found that lowering the threshold level to acquire more spikes resulted in more false detections. This reduced the overall predictive performance for TCR, even when the unit had a high SNR and a high firing rate. In addition, SBP required no trained parameters, while we found it necessary to optimize both TCR and low-bandwidth TCR to fairly compare the three

features (see Methods). To capture the low SNR spikes via thresholding, substantial time must be spent optimizing that threshold, whereas a fixed -4.5RMS threshold that is used frequently for closed-loop brain-machine interface experiments would miss virtually all of the low SNR spikes, a claim suggested by the worse low SNR prediction performance in the TCR plot of Figure 2.2a. Overall, this illustrates that the spiking band may be better able to extract firing rates of low SNR or low firing rate units despite its much lower bandwidth, ones that high-bandwidth detectors may ignore.

2.3.2 SBP Predicts Firing Rates of Local Units

The simulations presented thus far have only represented simulated recordings from one unit. In practice, electrodes record from many units simultaneously. Consequently, we investigated how well SBP tracks the true firing rates when there are multiple units surrounding an electrode. We first simulated recording of five neurons spiking spontaneously and independently at 20Hz in different arrangements surrounding the simulated recording electrode, similar to the work of Lempka *et al.* (Lempka et al., 2011). Each neuron's amplitude was attenuated based on its simulated displacement from the recording point according to the monopole equation (equation (2.2) in the methods) and the result of Holt and Koch's simulations (Holt and Koch, 1999). Figure 2.3 illustrates the results of these simulations, where a red color represents high correlation between that specific neuron's firing rate and the recording's SBP (blue is low correlation). First, we simulated a simple arrangement of the neurons cumulatively displaced by 50µm in Figure 2.3a. These waveform amplitudes fall off with the inverse of the distance according to the monopole equation. As a result, SBP strongly predicts the firing rate of just the highest amplitude unit, 50 um away. We then simulated a more realistic scenario with five units spaced out in a constant density around the electrode, which means that there are fewer close neurons and more distant

50
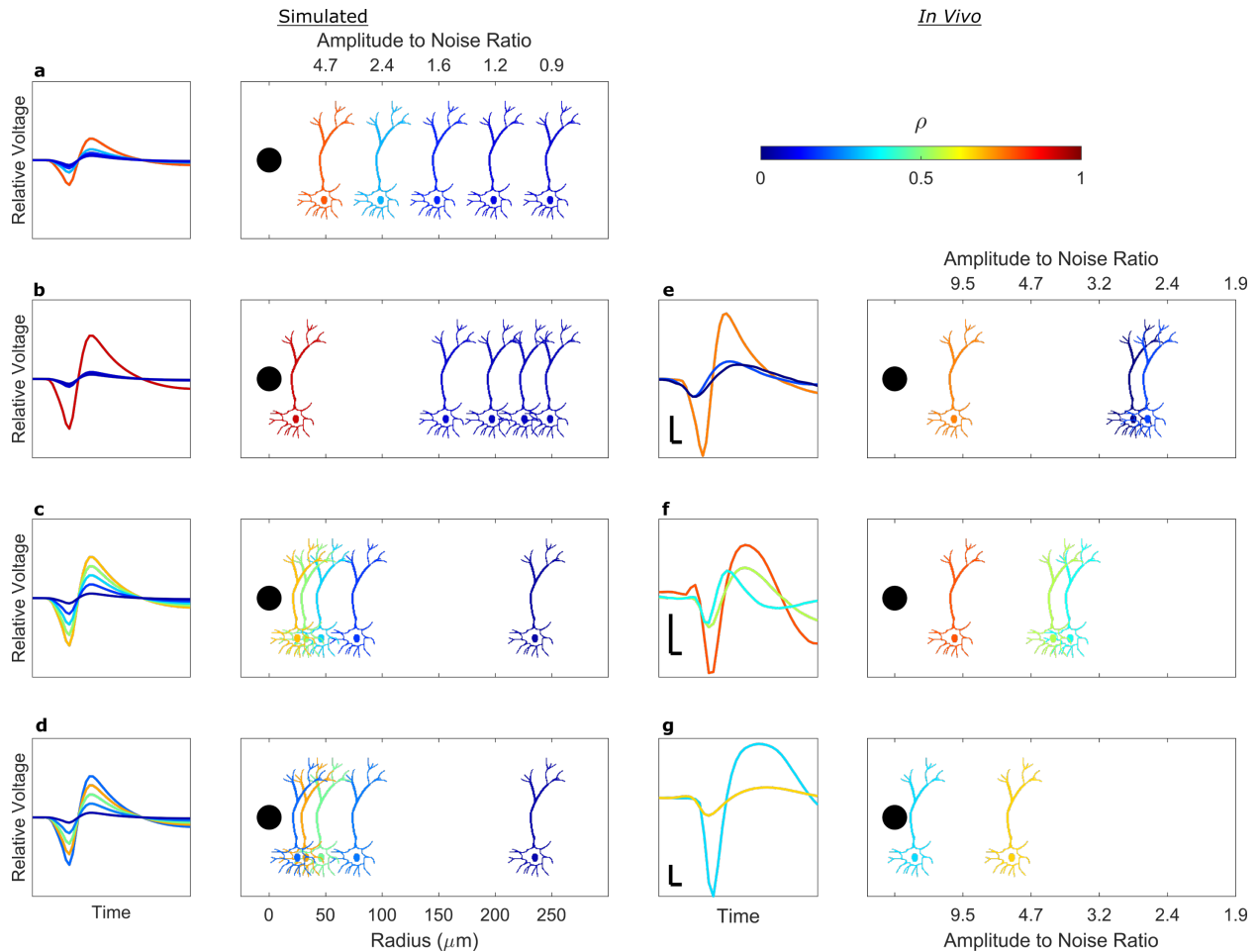
Figure 2.3 Correlation between the true firing rate of individual units and the field's SBP. Each rectangle represents an arrangement of neurons distributed as shown. The cartoon neurons and their corresponding spikes to the left are colored according to their firing rate's correlation with the recording's SBP. Simulated arrangements are plotted on the left (**a-d**); *in vivo* recording arrangements estimated by amplitude are on the right (**e-g**). Simulated neurons all fire at 20Hz, unless otherwise specified. The *in vivo* analyses on the right are similar in relative amplitude content to the simulations immediately to the left. Vertical and horizontal scale bars are 100µV and 0.1ms, respectively. Firing rates listed below are ordered from left to right as shown. (**a**) Cumulatively displaced by 50µm. (**b**) Constant volumetric density. (**c**) Linearly distributed by relative amplitudes from 9 to 1. (**d**) The same as (**c**) but the highest relative amplitude neuron fires at 1Hz. (**e**) *In vivo* recording from monkey W with units firing at 25, 34, and 7Hz. (**f**) *In vivo* recording from a carbon fiber electrode in rat motor cortex with units firing at 9, 12, and 9Hz. (**g**) *In vivo* recording from a carbon fiber electrode in rat motor cortex with the units firing at 0.5 and 15Hz.

neurons, and found an even stronger preference of SBP to the highest amplitude unit (Figure 2.3b).

Third, in Figure 2.3c, we simulated a highly active electrode, where multiple neurons correlated

with SBP. However, the correlation was weighted by amplitude towards the highest amplitude

units, leaving less representation of the lower amplitude units still visible above noise. Lastly, in

Figure 2.3d, to illustrate that firing rate is also factored into the representation, we simulated the

scenario in Figure 2.3c again, but with the highest amplitude unit spiking at just 1Hz. We found that SBP cannot detect the firing rate of the highest amplitude unit accurately (as predicted by lower firing rates in Figure 2.2a and the Appendix A section titled RMS-Based SNR Analysis) but shifts to a more specific prediction of the firing rate of the next largest amplitude unit that is active (or the highest RMS-based SNR unit according to the Appendix A section titled RMS-Based SNR Analysis). In all simulations, we found that SBP is a spatially local signal (see the monopole equation variant (2.2)] in Methods), dominated by the firing rates of the highest amplitude single units that are reasonably active, regardless of where they are with respect to any voltage threshold.

To compare with realistic voltage traces that include multiple sortable units, we used three recordings captured *in vivo* and performed the same analysis. These include one channel from monkey W's motor array (Figure 2.3e) and two from carbon fiber electrodes implanted acutely into rat motor cortex (Figure 2.3f, g). Overall, we found similar results to the simulations, validating that SBP is tuned strongly to the firing rates of higher amplitude units, while mostly ignoring low amplitude units. This suggests that while SBP may not be a single unit signal, it maintains the spatial specificity of broadband spike-related features like TCR or sorted units.

2.3.3 Improved Decoding with Spiking Band Power

In previous work, we and others demonstrated that SBP can be used to predict gross movements offline (Irwin et al., 2016; Stark and Abeles, 2007). Specifically for fingers, Irwin *et al.* showed that spiking band power from 300-1000Hz had very similar performance to TCR (Irwin et al., 2016). If the unit specificity suggested by the simulations above is correct, this would predict that the performance equivalence will be maintained in more complex tasks, specifically the control of multiple fingers, which we evaluate here. We trained three nonhuman primates, monkeys L, W, and N, to perform a one-dimensional task requiring them to move virtual fingers

into virtual targets by moving their physical fingers (see Figure A.1). In these datasets, monkeys L and W performed the task with all four fingers moving together and monkey N moving just his index finger. We first trained four Kalman filters for each monkey, one each for SBP, low-bandwidth TCR, TCR, and single unit firing rate, to decode the neural activity into finger positions offline. Each dataset was 10-fold cross-validated, and so, each decoder was always trained and tested on neural and behavioral data in which the monkey was moving the same finger(s). We found that SBP achieved correlations between its prediction and the actual finger movements statistically as high or higher than all other features ($p < 1 \times 10^{-4}$, Appendix A section titled Open-Loop Decoding Analysis). These results were comparable to those found by Vargas-Irwin *et al.*, Bansal *et al.*, and our group, but with a more challenging task to decode than our previous study due to more target locations and randomized target order (Bansal et al., 2011; Irwin et al., 2016; Vargas-Irwin et al., 2010). Offline, we also compared the importance of SBP, low-bandwidth TCR, and TCR channels for decoding in Figure A.5b and found that SBP generally had more channels that resulted in a drop in decode performance when individually excluded. This suggests SBP is extracting more valuable information from the available channels than TCR methods, which follows from the result of the low SNR simulations in Figure 2.2a.

With a functional offline SBP decoder, we tested SBP as a closed-loop neural feature in real-time to gauge its usability in neuroprosthetic applications. We trained monkeys W and N to perform the same one-dimensional finger task as the offline task, except using the index finger alone or the middle-ring-small fingers (MRS) as a group (as described by Vaskov *et al.*) (Vaskov et al., 2018). During each experiment, we trained ReFIT Kalman filters to decode TCR or SBP into finger positions and velocities (Vaskov et al., 2018). The closed-loop decoders the monkeys used for a specific finger group were always trained on neural data and behavioral data associated
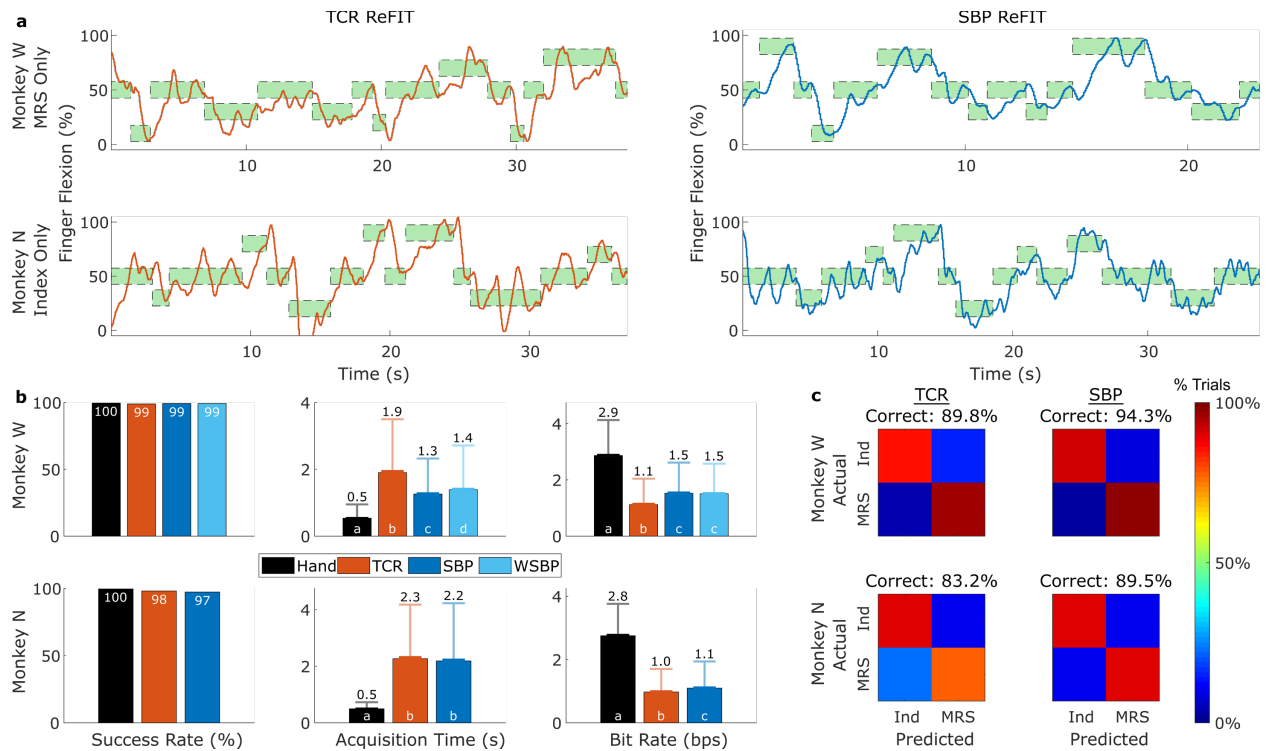
Figure 2.4 Comparison between the decoding performances of SBP and TCR. (**a**) Closed-loop decode traces from translating TCR in orange (*left*) or SBP in blue (*right*) for 15 targets of each decoder. Targets are green (for successful acquisition) and surrounded by the dashed line. (**b**) Success rate, acquisition time, and bit rate statistics for all methods of closed-loop virtual hand control. Columns represent mean with standard error shown by the darker error bars and standard deviation shown by the lighter error bars. Bars in a group labeled with the same letter are statistically similar. For example, in monkey W's bit rate statistics, hand control's bit rate was statistically different than all other control methods, so it was exclusively labelled 'a'. TCR control's bit rate was also statistically different than all others, so it was exclusively labelled 'b'. However, the bit rates for SBP and WSBP control were statistically similar to each other, so they share label 'c'. $p < 1 \times 10^{-3}$ between SBP and wideband SBP for monkey W acquisition time and between SBP and TCR for monkey N bit rate. $p < 1 \times 10^{-18}$ otherwise. The legend represents trials in which the virtual fingers were controlled by the monkey's physical hand movements or the TCR, SBP, or wideband SBP (300-6,000Hz passband) decoders, respectively. (**c**) Support vector machine confusion matrices from classifying finger movements as index (Ind) or middle-ring-small (MRS).

with movements of that finger group. Figure 2.4a shows predicted finger traces with visual feedback provided to monkey W using only his MRS and monkey N using only his index finger (recording conditions displayed in Figure A.3). Across all ReFIT decoders for any combination of finger groups, both monkeys achieved significantly higher bit rates (1.1bps to 1.5bps, $p < 1 \times 10^{-25}$ for monkey W, 1.0bps to 1.1bps, $p < 1 \times 10^{-4}$ for monkey N, single-tailed two-sample *t*-test) when using the SBP decoders in comparison to TCR decoders, though only monkey W acquired targets in significantly less time (1.9s to 1.3s, $p < 1 \times 10^{-30}$, single-tailed two-sample

t-test, Figure 2.4b). This suggests that ReFIT Kalman filters can decode SBP into individual finger motions at least as well as TCR. For completeness, monkey W also used a ReFIT Kalman filter trained to decode 300-6,000Hz SBP (i.e. from Stark and Abeles 2007 (Stark and Abeles, 2007)) to characterize any losses from the bandwidth reduction. With SBP allowing an improved acquisition time (1.4s to 1.3s, $p < 1 \times 10^{-3}$, single-tailed two-sample *t*-test) and no significant difference in bit rate (1.5bps, $p = 0.51$, two-tailed two-sample *t*-test) over 300-6,000Hz SBP, we suggest the 300-1,000Hz spiking band is sufficient for optimal closed-loop SBP decoding.

With the ability to use SBP to decode movements of individual finger groups, we sought to validate that SBP could maintain high performance in two-dimensional tasks. First, we attempted to classify which finger group was moving when both were given unrestricted mobility. We trained the monkeys to acquire two targets simultaneously in a two-dimensional finger task, one exclusively for each finger group, and presented targets such that only one group was moved from rest to a flexion position at a time while the other group remained at rest. We trained two support vector machines to classify which finger moved in a given trial, one using TCR and the other using SBP. The SBP machine achieved the higher prediction performance (average >5% improvement for both monkeys), as displayed in Figure 2.4c and similar to that found previously for TCR (Vaskov et al., 2018). Second, we predicted offline continuous hand position in a two-dimensional center-out-and-back cursor control task using a standard Kalman filter in nonhuman primates, similar to what was done by Kao *et al.* (Kao et al., 2017). We found that SBP achieved statistically higher correlation coefficients between the predicted and actual hand positions and statistically similar hand velocity correlation coefficients compared to TCR and low-bandwidth TCR (for SBP, TCR, and low-bandwidth TCR, respectively: 0.86, 0.84, 0.84 X position; 0.94, 0.93, 0,93 Y position; 0.74, 0.73, 0.74 X velocity; 0.84, 0.82, 0.82 Y velocity for monkey J; 0.84,

0.81, 0.81 X position; 0.77, 0.72, 0.75 Y position; 0.76, 0.71, 0.73 X velocity; 0.77, 0.73, 0.74 Y velocity for monkey L; $p < 1 \times 10^{-3}$, two-tailed two-sample z-test, Appendix A section titled Two-Dimensional Cursor Control).

These results, consistent with our one-dimensional finger task, suggest that SBP exceeds the performance of threshold crossing-based features at multidimensional control tasks despite the drastically reduced bandwidth, similar to what was found previously by Stark and Abeles (Stark and Abeles, 2007). We also investigated the impacts of increasing the recording noise on decoding performance in Appendix A section titled Two-Dimensional Cursor Control, and found that SBP could tolerate almost two times the noise level as TCR to maintain at least 80% of their no-noise decode performances.

2.4 Discussion

Here we have presented a new perspective for comparing single unit firing patterns and 300-1,000Hz spiking band power. Specifically, we have shown that the single unit specificity of spiking band power is comparable to sorted units despite the reduction in bandwidth, likely explaining the mechanism that yielded the outcomes of our past work (Irwin et al., 2016) and that of Stark and Abeles}. We also believe that the specificity of SBP improves its decoding performance to be comparable or better than standard high bandwidth methods, such as TCR, and comparable to sorted unit firing rates. Finally, we believe our results extend our previous findings as well as those of Stark and Abeles to suggest that the decode performance improvements of SBP are maintained in complicated behaviors, higher dimensions, and closed-loop.

The results of our simulations suggest that SBP, sampled at only 2kSps, is a highly spatially-specific neural feature. It is dominated by spikes, and to extract neural information, it requires spikes to be present. In comparison to TCR, SBP requires no specific tuning for it to

function optimally. It is "parameter-free," without requiring adjustable thresholds or any programming to implement spike sorting that must be done daily for optimal TCR performance. From a device perspective, TCR would require programmable and receiver circuitry to, at the least, optimize the thresholds that we found varied significantly between animals and across days within one animal (see Appendix A section titled Open-Loop Decoding Analysis). On the other hand, SBP would need no additional circuitry while maintaining performance levels. In a medical application, this implies a reduction in training time without a required daily parameter search of the best threshold and overall greater ease of use as a medical device. Additionally, lower device complexity and power consumption without any required receiver hardware would imply an increased battery life of SBP devices. As a result, SBP may become broadly applicable to other closed-loop low-power neural applications, such as bladder state monitoring, deep brain stimulation, and seizure prevention.

Importantly, these results suggest that SBP can accurately extract single unit firing rates at lower SNRs than TCR, suggesting that more activity can be extracted on channels that may not have visible spikes. This has major implications for the longevity of implanted microelectrode arrays in human patients, as it has been reported that the quantity of electrodes displaying spikes visible above noise wanes over time (Barrese et al., 2013). We performed a supplementary analysis (Appendix A section titled Two-Dimensional Cursor Control) in which we injected noise into the recorded neural activity to investigate the impacts of fewer visible neural spikes on decoding. Through this analysis, we validated that SBP maintains better decoding performance than TCR as visible spike presence on an array drops off. In fact, some groups conducting human clinical trials have already made use of this approach to extend the lifetime of their arrays (Ajiboye et al., 2017; Jarosiewicz et al., 2015; Willett et al., 2017). In particular, Ajiboye *et al.* added high frequency

band power (essentially 250-3,000Hz SBP) to the spike counts for their brain-controlled functional electrical stimulation prosthesis (Ajiboye et al., 2017). Jarosiewicz *et al.* also used 250-5,000Hz SBP in addition to spike counts for their virtual typing task in human patients (Jarosiewicz et al., 2015). In neighboring fields, groups that implant deep-brain electrodes utilize SBP to detect when electrodes have entered the gray matter of the subthalamic nucleus (Novak et al., 2011). We believe that our findings suggest a potential mechanism explaining why their methods work, and we hypothesize that SBP's capabilities of extracting low-SNR unit activity may extend the usable, high-performance lifetime of such a neuroprosthesis even once the spikes are indiscernible above broadband noise.

While most high-performance brain-machine interfaces rely on visible spikes to decode, many groups have experienced performance issues as the spike amplitudes wane over time or are invisible above noise (Chestek et al., 2011; Downey et al., 2018; Wang et al., 2014). Consequently, local-field potentials and electrocorticography have been investigated as replacement and supplemental feature sources, as they represent local neural activity even with lacking spikes. As many local-field potential studies have shown performances comparable to but lower than spike-based features (Aggarwal et al., 2013; Flint et al., 2012a, 2012b, 2013; Gilja et al., 2015; Milekovic et al., 2018; Pandarinath et al., 2017; Perge et al., 2014; So et al., 2014; Stark and Abeles, 2007; Stavisky et al., 2015; Zhuang et al., 2010), it is inconclusive how much of the performance is a result of low-pass filtered spikes. However, the results we have presented clearly demonstrate that SBP, while dependent on spikes, exceeds the performance of high-bandwidth spike-based features. Further, the average electrocorticography open-loop decode correlations of 0.67 (Chao et al., 2010; Chestek et al., 2013; Flint et al., 2017; Hotson et al., 2016; Kubánek et al., 2009) are similar to the offline analyses we have presented here for SBP and TCR. However, the simplicity of the tasks in

the electrocorticography studies suggests that the decoders may be predicting the presence of movement effort rather than the movement itself. In contrast, our one- and two-dimensional center-out-and-back decoding results using SBP demonstrate SBP's capability of predicting the movement directly with performance as good or better than TCR.

One of the major benefits of SBP as a reduced bandwidth, highly specific neural feature is the power savings it enables compared to higher bandwidth devices previously presented (Borton et al., 2013; Rizk et al., 2009; Yin et al., 2013). In our supplementary analysis presented in Appendix A section titled Integrated Circuit Simulations, we simulated the power consumptions of optimized, custom integrated circuits that extract 2kSps low-bandwidth features or 20kSps high-bandwidth features to estimate the power savings that can be attributed to the reduction in recording bandwidth. Based on those results, which showed a 90% reduction in power consumption from 11mW to 1.1mW, we believe the greatest improvement to the power consumption of neural interfaces can only be made through changes to the analog front-end. However, such a wide bandwidth is not optimized for power consumption and is much larger than what is necessary to accurately estimate firing rates with threshold crossings. While reducing the TCR bandwidth to the spiking band would save equivalent amounts of amplifier power, such a circuit would be sub-optimal, as the SBP decodes always out-performed low-bandwidth TCR. Additionally, the low-bandwidth TCR circuit would require a receiver and programmable circuits for the thresholds that SBP does not require. Such circuitry would increase the power consumption of a low-bandwidth TCR circuit (though not investigated here) and be a major hindrance in applicability to neural dust-like chiplets that have intermittent power supplies and would benefit greatly from having no programmable, power-consuming memory for threshold levels that would otherwise be forgotten each power cycle. The factor of 10 reduction in power consumption

predicted by the circuit simulations implies a substantial increase in battery life of SBP devices while enhancing decoding performance over the equivalent TCR circuit.

The lower power requirements of SBP would not only reduce the consumptions of current acquisition technologies, but also have implications for the future of cellular-scale neural recording. State-of-the-art front-end designs are being presented that require a small number of transistors (Seo et al., 2016; Steinmetz et al., 2018; Viventi et al., 2011). With simpler front-end amplifiers, it becomes much more plausible to have electronics mounted on the array itself or have modular, wireless "chiplets" with single unit level specificity (Lee et al., 2018; Seo et al., 2016). This also suggests that existing ASICs developed for local field potentials and electrocorticography may be sufficient for brain machine interface applications, so long as there are spikes present on the electrodes (Johnson et al., 2017; Mestais et al., 2015; Robinet et al., 2011). Overall, the minimal hardware and small supply requirements of SBP may help enable a path towards recording thousands to tens of thousands of channels with similar performance to single unit based systems.

2.5 Acknowledgements

Chapter 3 Embedded and Integrated Brain-Machine Interfaces for Predicting Finger Movements

This chapter combines my key contributions to three related bodies of work. First, this chapter explains my key contributions as second author on the manuscript titled, "Design and Testing of a 96-Channel Neural Interface Module for the Networked Neuroprosthesis System," published in Bioelectronic Medicine in 2019 (Bullard et al., 2019). Second, this chapter expands on that body of work and some of the results that were presented at the Society for Neuroscience 49th Annual Meeting in Chicago, IL, with additional details, analysis, and discussion. Third, this chapter also briefly explains my key contributions as co-first author to the manuscript titled, "A Power-Efficient Brain-Machine Interface System with a Sub-mW Feature Extraction and Decoding ASIC Demonstrated in Nonhuman Primates," which has been submitted to IEEE Transactions on Biomedical Circuits and Systems.

3.1 Introduction

Brain-machine interfaces (BMIs) have shown that they can restore function to people with paralysis. From restoring handwriting capabilities (Willett et al., 2021) to usage of one's own limb to eat and drink (Ajiboye et al., 2017), BMIs provide a more naturalistic solution to obtaining motor control signals than other methods, such as the usage of residual functional musculature to map intentions to assistive devices. However, most BMI studies to date have required carts of computers to handle the high data throughput required to accurately predict a user's intentions. While the capabilities are promising, substantial compression of BMI systems will be required

before they can be used as a general-purpose clinical therapy, one that can be used portably without being tethered to a large stack of machinery.

Accurate predictions originally required detection of the spiking rates of individual cortical units, which used high sampling rates at 30kSps, a ~250Hz to ~7,500Hz filter to remove low-frequency population activity, a threshold to detect spikes, and additional processing to sort the spikes into similar waveform templates (Carmena et al., 2003; Serruya et al., 2002; Shenoy et al., 2003; Taylor et al., 2002a). The first step towards compression was taken shortly afterwards, where bypassing the sorting step saved a substantial amount of post-processing without substantial losses in performance (Fraser et al., 2009; Ventura, 2008). It was later shown that neuroscientific population dynamics analyses could be done with just the threshold crossing rate as well (Trautmann et al., 2019). However, even with threshold crossing rate, the front-end amplifier that additionally filters the recorded neural data still strongly dominates BMI power consumption, as we determined previously in Appendix A (Nason et al., 2020). This remained so until it was found that the signal power in the 300-1,000Hz band of spiking activity, termed spiking band power (SBP), strongly represents the underlying spikes (Irwin et al., 2016; Nason et al., 2020). Not only would the amplifier save power, but the sampling rate could be reduced by 15x, substantially reducing the incoming data rate and the quantity of downstream computations. Despite all of the promising characteristics of SBP, no study has yet shown that scaling down to SBP results in substantial hardware compression of BMIs.

Here, we show that SBP enables low-power, compressed brain-machine interfaces as embedded and integrated devices. These devices take advantage of off-the-shelf front-end amplifiers and analog-to-digital converters (Intan RHD2132s) for optimal data acquisition, then use power-efficient processors to predict the intended finger movements of nonhuman primates in

real-time. Closed-loop decoding with the devices achieved similar performance to our previously-published one-finger decoding results (see Chapter 2) and comparable performance to our previously published two-finger decoding results (see Chapter 4), both using our experimental BMI rig.

3.2 Methods

3.2.1 Embedded Neural Signal Processing

First, we developed the device presented in (Bullard et al., 2019), referred to as the embedded BMI in this chapter. The device was comprised of three 32-channel Intan RHD2132 bioamplifiers (Intan Technologies, Los Angeles, CA, USA), which are off-the-shelf amplifiers and analog-to-digital converters optimized for power consumption. Three RHD2132s were used to yield 96 total channels, the quantity available on a standard Utah microelectrode array (Blackrock Microsystems, LLC., Salt Lake City, UT, USA). Sampling from these devices was coordinated by an Atmel AT32UC3C2256C microcontroller (Microchip Technology Inc., Chandler, AZ, USA), which supports 32-bit and floating-point operations. Processing results could be exported via a FESCAN interface for interaction with other implantable neuroprosthetic devices (Makowski et al., 2021), though a simple universal asynchronous receive/transmit interface (UART) was used for this study.

At the time of publication, the embedded BMI was able to extract SBP recorded directly from a Utah microelectrode array implanted in the hand area of primary motor cortex in a nonhuman primate. To do so, it used one of two operating modes. In high-bandwidth mode, it configured the RHD2132 amplifiers to filter one channel's neural signal from 0.1Hz to 7.5kHz and sampled it at 32kSps, which was the closest sampling rate above 30kSps (standard sampling rate of the Cerebus [Blackrock Microsystems]) that could be achieved while remaining agnostic

to timer-based interrupts (for firmware simplicity). Then, the microcontroller would transmit each sample over UART to a receiving computer. This mode was used to compare recording quality to standard recording systems.

In low-bandwidth mode, the microcontroller configured the RHD2132 amplifiers to filter all 96 channels' neural signals from 300-1,000Hz, with each channel sampled at 2.17kSps (for firmware simplicity). The RHD2132s include a helpful absolute-value function that assists with the computational efficiency of calculating signal power, which was used in low-bandwidth mode. Then, the microcontroller would average these absolute-valued samples in some customizable time period, which was set to 128 samples in duration (~58ms at 2.17kSps) during this study. This averaged activity represents the spiking band power for each channel and was transmitted over UART. We implemented two data buffers so that one buffer could be used for acquiring data while the other was being processed into SBP, leaving no gaps in recording time.

The Atmel AT32UC3C2256C includes a direct memory access controller, which enables peripherals to directly access random-access memory without requiring any oversight from the processing core. In this way, the direct memory access controller could manage all data sampling in both high- and low-bandwidth modes while the processing core remained asleep, saving substantial amounts of electrical power during usage. In high-bandwidth mode, the direct memory access controller grabbed all samples. In low-bandwidth mode, the direct memory access controller grabbed all samples from all channels per integration bin, then woke the processing core to calculate the average value of each channel in that bin.

3.2.2 Embedded Device Decoding

Since our publication illustrating low-power SBP recording, we made several improvements to the embedded BMI's firmware. The first improvement involved using the direct

memory access controller to grab one sample at a time from all channels prior to requiring reconfiguration from the processing core. While this may seem like a reduction in firmware efficiency, doing so enables more consistent sample timing.

This introduces the second improvement, in which an internal timer was configured to wake up the microcontroller every 500μs to coordinate all device operations. Upon wakeup, the processor addressed the timer's interrupt by restarting the direct memory access controller to collect the subsequent sample from all 96 channels. The firmware used a secondary counter to track the quantity of samples that had been acquired for the current bin, only beginning the averaging operations once a bin's worth of samples had been acquired for every channel.

The third improvement involves the capability of decoding the binned SBP in real-time. To do so, we implemented a steady-state Kalman filter (Malik et al., 2011), which offers lower computational complexity and fewer stored parameters in comparison to the standard Kalman filter without loss of accuracy. The Kalman gain converges to a steady-state value within a few seconds of use, which allows the steady-state Kalman filter to avoid a computationally expensive matrix inversion involved in its calculation. Instead, the product of this matrix inversion can be pre-computed during training, not during real-time operation. We also optimized parameter storage and computation by grouping similar matrix computations:

$$\hat{x}_t = (I - KC)A\hat{x}_{t-1} + Ky_t$$

where $\hat{x}_t$ is the prediction of the current state, i.e. position, velocity, etc.; $\hat{x}_{t-1}$ is the predicted state of the previous time-step; y_t is the current time-step's observation vector (i.e. 96 SBP measurements); K is the pre-computed Kalman gain; C is the observation matrix; and A is the state transition matrix.

3.2.3 Integrated Brain-Machine Interface

Our integrated BMI, named the Neural Recording and Decoding (NeuRAD) device, was designed to save more power beyond what was capable with the general-purpose hardware used in the embedded BMI. To do so, there were several major application-specific customizations incorporated into the design to increase computational and electrical efficiency. The device is described in brief detail below, with full details available in the completed publication (An et al., 2022, *in submission*).

The first major optimization was for calculating SBP. A hardware accelerator was designed that coordinates sampling from the 96 channels of the three Intan RHD2132s and computes the average activity of each channel in customizable bin sizes. The second major optimization was incorporating a matrix accelerator, which increased computational and power efficiency of executing matrix operations. Both the SBP and matrix accelerators offloaded substantial amounts of processing from the main processing core. The third major optimization was using a more power-efficient processing core, the M0 core, which required reducing the communication buses to 16-bit. Otherwise, the flow of information through the integrated BMI was exactly as described for the embedded BMI.

3.2.4 Closed-Loop Experiments

To test the embedded and integrated BMIs in application, we trained a nonhuman primate (Monkey N) to perform a finger target acquisition task, as discussed in-depth in Chapter 2 (Nason et al., 2020). For relevance to the two-dimensional brain-computer interfaces field involving the control of two-dimensional computer cursors, we also trained the monkey to perform a two-finger version of the finger task, discussed in-depth in Chapter 4 (Nason et al., 2021b). Briefly, the monkey controlled virtual fingers to acquire virtual targets by moving his physical fingers, with movements measured by a manipulandum. The monkey would complete this task in this fashion
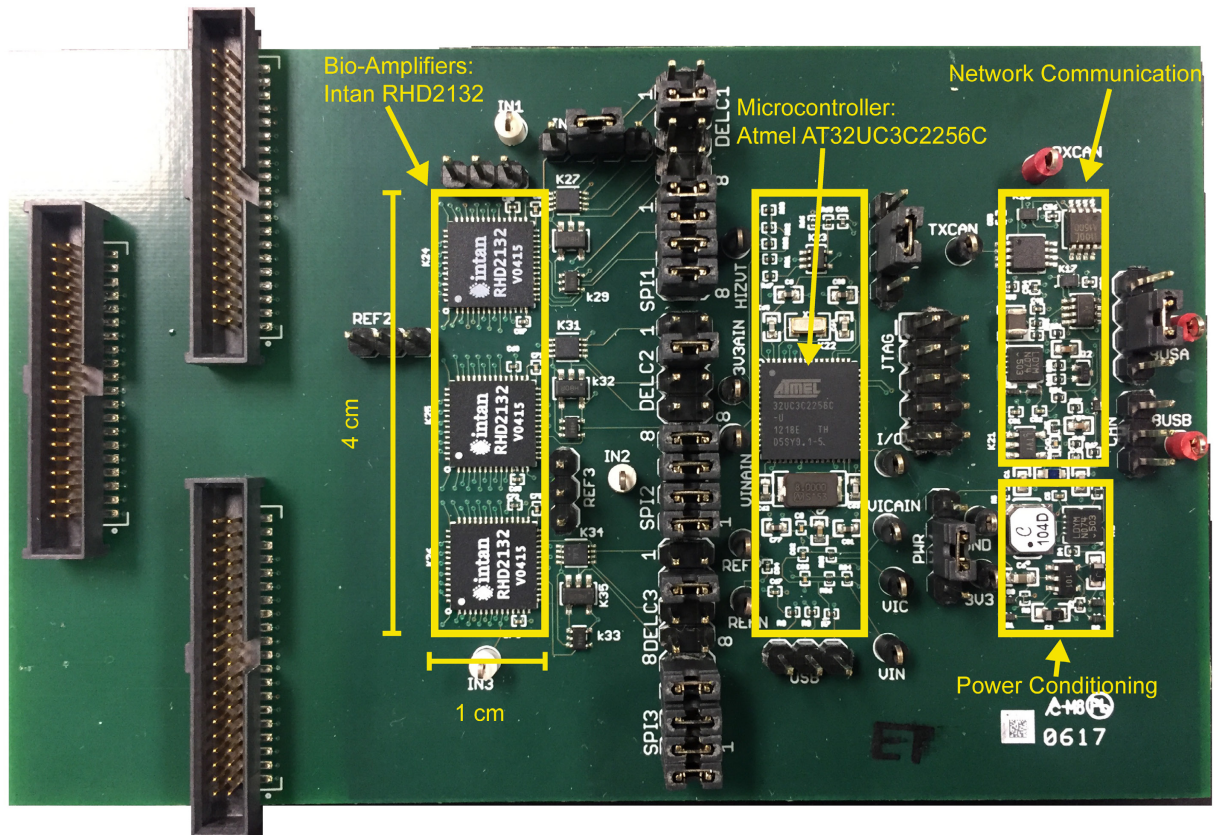
Figure 3.1 Embedded brain-machine interface as presented in (Bullard et al., 2019). In the left panel are the three Intan RHD2132s, in the middle panel is the Atmel AT32UC3C2256C, and in the right panel is all of the FESCAN interface circuitry. The circuitry in the panels contain the only required circuitry for complete device functionality, suggesting the device could be packaged in a 1x4cm container.

for approximately five to ten minutes while the embedded or integrated BMI was recording SBP from the Utah array implanted in the hand area of his primary motor cortex. The experiment was coordinated by a computer running xPC Target (Mathworks, Natick, MA, USA), which received the 16-bit, accumulated SBP measurements from the embedded or integrated BMI to be synchronized in real-time with the behavioral measurements. The devices used a 230,400 baud RS232 connection to transmit data to the xPC Target computer by way of a MAX3222E UART-to-RS232 converter (Maxim Integrated, San Jose, CA, USA).

Then, we trained a steady-state position/velocity Kalman filter from the manipulandum control trials and programmed the parameters to the devices. Finally, the devices used the steady-state Kalman filter to predict the fingers' positions and velocities in real-time using SBP (also

Table 3.1 Power consumed in different embedded BMI operating modes.

No. Channels	Operating Mode	System Power
1	30kSps	31.2mW
1	2kSps	22.1mW
96	2kSps DMA sampling & CPU awake	45.3mW
96	2kSps DMA sampling & Disabled Peripherals & CPU awake	40.0mW
96	2kSps DMA sampling & Disabled Peripherals & CPU asleep	33.6mW

Power consumption is presented by progressively adding a power saving technique. DMA means direct memory access controller. CPU means central processing unit, or the Atmel AT32UC3C2256C in this case.

calculated by the devices), then transmitted those predictions to the xPC Target computer. Those predictions were then used to control the virtual fingers instead of being controlled by measurements from the manipulandum. To mimic our closed-loop BMI results we have already published, we integrated velocity into position to control the virtual hand.

3.3 Results

First, we briefly present our embedded BMI's power consumption during recording, as established in (Bullard et al., 2019). Figure 3.1 presents the device as used in the study, and Table 3.1 presents the power consumption of the device under different operating modes. Notably, the embedded BMI consumed 31.2mW to acquire one channel at 30kSps, but consumed just 2.4mW more to acquire 95 additional channels of SBP, which provide greater unit specificity and better decoding performance, as established in Chapter 2.

3.3.1 Low-Power Decoding on an Embedded BMI

After the substantial optimizations made to the firmware presented in (Bullard et al., 2019), including decoding capabilities and improved sample timing, the new firmware's state machine is represented in Figure 3.2. Before we could investigate closed-loop prediction performance with the embedded BMI, we had to decide on a binning period. When a bin completes within the new firmware's state machine, a train of functions is called: swapping the buffer into which new data
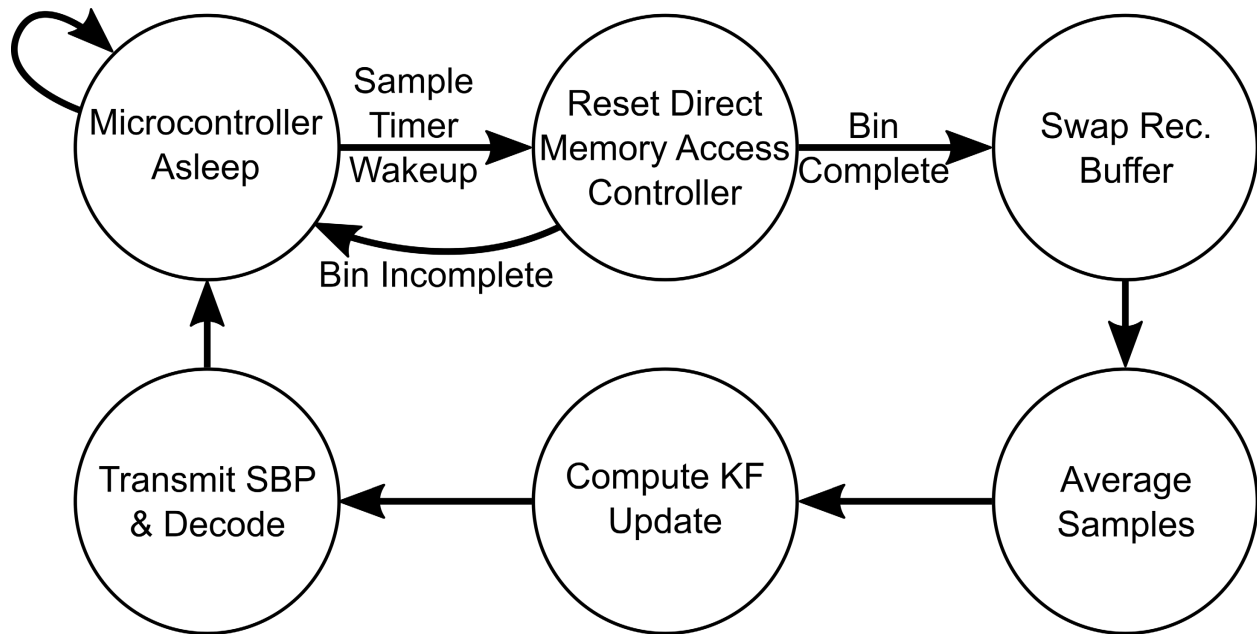
Figure 3.2 State machine of the embedded BMI during closed-loop decoding experiments. To optimize for power consumption, the microcontroller sleeps whenever continued functionality can support it. Therefore, the direct memory access controller coordinates all sampling.

is recorded, averaging the samples of the bin for each channel, computing a Kalman filter update, and transmitting the computed SBP and Kalman filter predictions to display on-screen. This means that, at the completion of a bin, the processing core will be awake for a substantial amount of time, which consumes more electrical power than when it is asleep. Smaller bin sizes imply faster update rates, meaning lower-latency visual feedback provided to the user, but more frequent periods of the processor computing updates. Larger bin sizes mean the opposite: slower feedback but lower power consumption.

We made this decision by first using our main, high-powered brain-machine interface rig to investigate the impact of bin size on closed-loop BMI performance. We trained a nonhuman primate to perform a two-finger target acquisition task (discussed in detail in Chapter 4), which was similar to the one-finger task presented in Chapter 2. The primary difference was that we trained Monkey N to split the movements of his index finger from his middle-ring-small fingers as a group to acquire one target presented for each group (two total targets). We trained Kalman
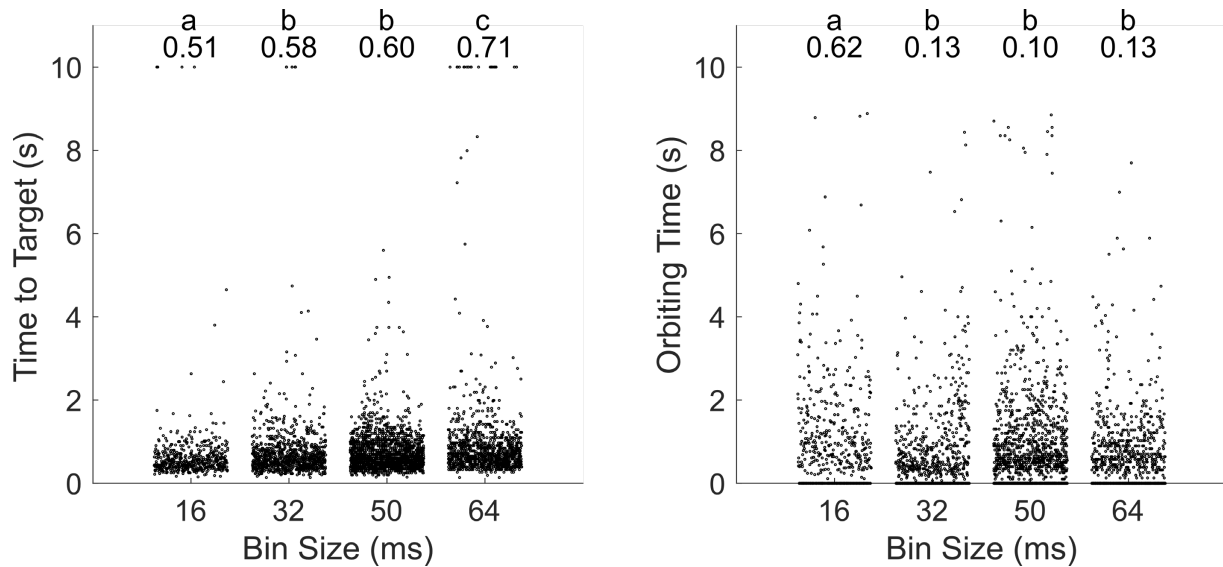
Figure 3.3 Comparison between different decoder update rates and SBP integration periods in a two-finger task (see Chapter 4 for more details). Time to target represents the time between the start of the trial and the time at which all fingers were in their corresponding targets. Orbiting time represents the time since the first instance all fingers touched their corresponding targets to the end of the trial, less the hold time. The numbers above each column represent the median of the data points below, and the associated letters represent statistically similar groups (i.e. any numbers with the same letter are not significantly different, $p > 1 \times 10^{-3}$, two-tailed Wilcoxon rank sum test).

filters to use SBP computed in a variety of bin sizes (16ms, 32ms, 50ms, 64ms, and 128ms) to predict Monkey N's movements of his index and middle-ring-small fingers in closed-loop form. The results of this experiment are showcased in Figure 3.3 (128ms was excluded because performance was substantially worse than the other bin sizes investigated). It seems that shorter bin sizes resulted in higher velocities corresponding to faster times to target (0.51s, 0.58s, 0.60s, and 0.71s, respectively), but also resulted in decoder instability, where the monkey spent a substantial amount of time trying to stop on the target after hitting it (0.62s, 0.13s, 0.10s, and 0.13s, respectively). Therefore, for computational simplicity on our embedded system, we settled for a 32ms update rate. 32ms provides a good balance between speed to get to the targets and short times spent stopping on the targets.

We then used a steady-state Kalman filter to predict one- and two-finger movements every 32ms using our embedded BMI with the predictions presented in real-time as visual feedback to Monkey N. Figure 3.4 illustrates closed-loop prediction traces and statistics for these experiments.
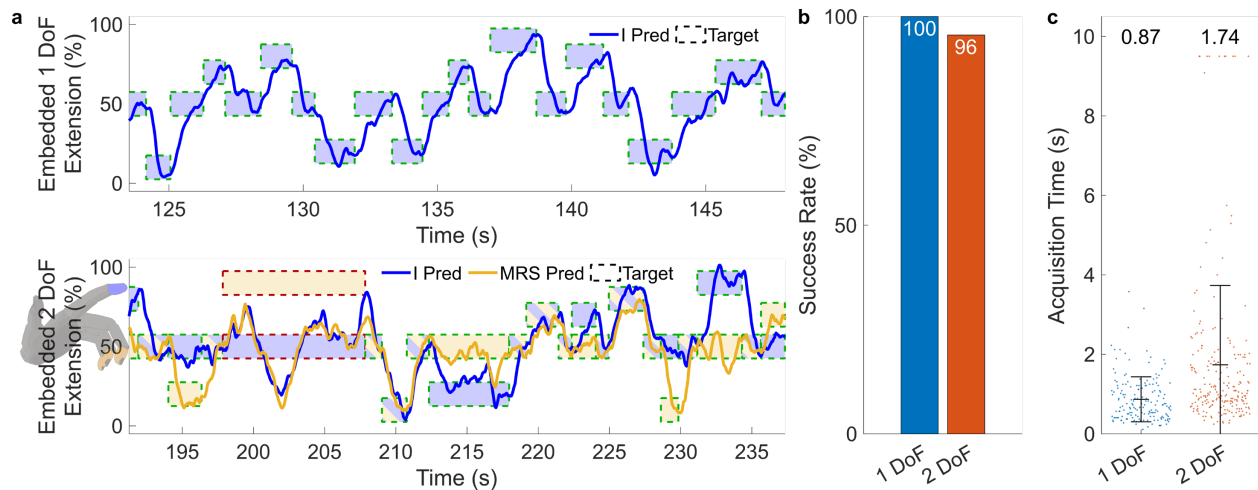
Figure 3.4 Closed-loop predictions and metrics using the embedded BMI. (**a**) Closed-loop prediction traces made by the embedded BMI with a 32ms update rate. Top represents one-finger predictions, bottom represents two-finger predictions. Targets are colored by the fingers for which they correspond, with a green border color if acquired successfully and a red border color if not. (**b**) Success rate statistics for each task. (**c**) Acquisition time statistics for each task. Error bars represent mean ± standard deviation, and the number centered over each column represents the mean.

Monkey N could achieve 0.87s and 1.74s mean target acquisition times for one- and two-finger tasks using the embedded BMI. To execute these predictions, the embedded BMI consumed just 58.4mW. For comparison, we also used a 64ms update rate to estimate the performance-to-power consumption trade-off with the one-finger task. We found that mean acquisition time increased to 1.14s (31% increase, $p < 1 \times 10^{-3}$, two-tailed Wilcoxon rank sum test) with a 0.6mW (~1%) power savings, which we think is not substantial enough to warrant such a loss of performance. It is also important to note that, compared to what we previously published in (Bullard et al., 2019), the power consumption is substantially greater due to the doubled clock speed that was required to additionally support real-time decoding.

3.3.2 Low-Power Decoding on an Integrated BMI

It is encouraging to find that brain-machine interfaces cannot only be compressed in size and power consumption when implemented using embedded hardware, but the performance levels remain high in one- and two-dimensional tasks. Unfortunately, 58.4mW would only give 11.3
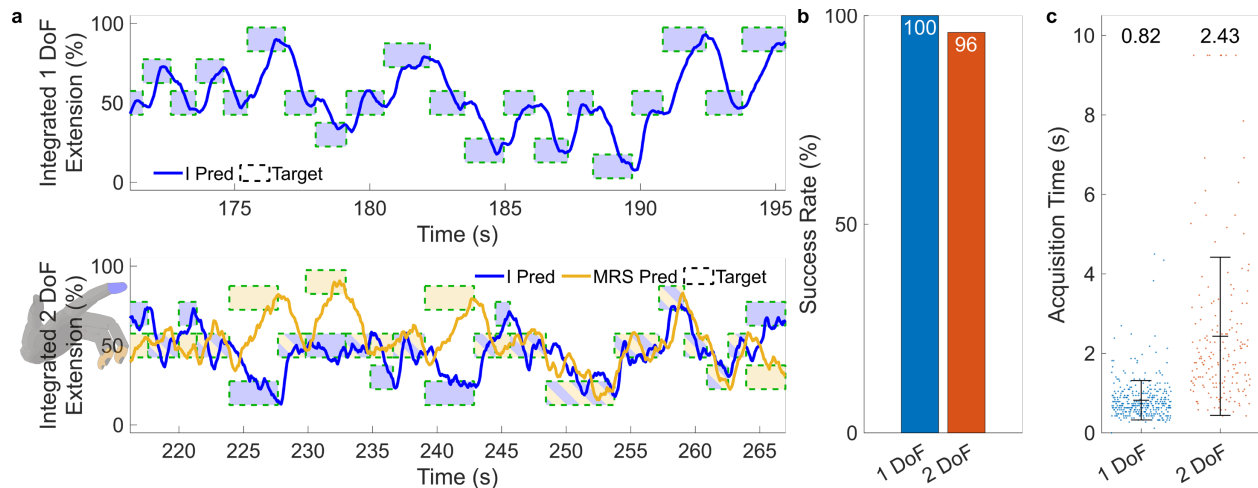
Figure 3.5 Closed-loop predictions and metrics using the NeuRAD. (a) Closed-loop prediction traces made by the NeuRAD with a 50ms update rate. Top represents one-finger predictions, bottom represents two-finger predictions. Targets are colored by the fingers for which they correspond **(b)** Success rate statistics for each task. **(c)** Acquisition time statistics for each task. Error bars represent mean ± standard deviation, and the number centered over each column represents the mean.

hours of usage with a standard 200mAh implantable battery. That battery life is too short, particularly if the total capacity additionally needs to power a prosthesis, such as the implantable Networked Neuroprosthesis functional electrical stimulation system (discussed more in Chapter 5). To this end, we sought to investigate how low we could push the power consumption of a brain-machine interface by designing an application-specific integrated circuit. Specifically, we designed the Neural Recording and Decoding (NeuRAD) device with SBP- and matrix math-focused hardware accelerators to determine the capabilities of custom circuitry.

Similar to the embedded BMI experiments, we wanted to validate that the NeuRAD could efficiently extract SBP and decode it into one- and two-finger movements in real-time. Approximately two years after the embedded BMI experiments with Monkey N, he used the NeuRAD to perform similar experiments with a 50ms bin size. Figure 3.5 illustrates closed-loop prediction traces and statistics for these experiments. Monkey N could achieve 0.82s and 2.43s mean target acquisition times for one- and two-finger tasks using the NeuRAD. To do so, the NeuRAD consumed approximately 12.5mW, a 78.6% reduction in power consumption to achieve

better performance for the one-finger task and lower but good performance for the two-finger task. Importantly, 12 of the total 12.5mW was consumed just by the Intan RHD2132s, meaning the processing alone dropped from 46.4mW to approximately 0.5mW, a 98.9% reduction in power consumption. Furthermore, at 12.5mW, a standard implantable 200mAh battery would last 52.8 hours with the NeuRAD, a 4.7x increase over the embedded BMI.

3.4 Discussion

Despite the promising functional restorations that have been provided by brain-machine interfaces to people with paralysis, translation of the technology to become a full-time clinical therapy has been severely hindered by device size and power consumption. Here, we present two devices that compress brain-machine interfaces into low-power, implantable packages. Our embedded and integrated BMIs were capable of predicting the simultaneous and independent movements of two finger groups in real-time with high performance, which makes the devices immediately applicable to the realm of BMI-controlled computers. To do so, the devices consumed 58.4mW and 12.5mW, respectively, which would provide substantial battery life with a standard 200mAh implantable battery. Importantly, to our knowledge, these two devices represent the first demonstration of an end-to-end brain-machine interface that can be completely implanted, implemented entirely using off-the-shelf hardware (for the embedded BMI) or an optimized application-specific integrated circuit (for the NeuRAD).

For the embedded BMI, one area of potential improvement is feedback latency. The embedded BMI's firmware was designed such that it would only compute the mean SBP activity on each channel once all samples belonging to a bin had been acquired. That period following a bin's completion where the averages are being computed takes up most of the time the processor is awake, substantially more than the time it takes to compute the Kalman filter update. As hinted

at in Appendix C, visual feedback latency can have a negative impact on the BMI's performance. By reducing the amount of time spent calculating channel means, the latency could be reduced, which means performance could be improved. This may become particularly crucial during usage of a prosthesis, where the prosthesis may inject a substantial amount of its own latency. It is important to note, however, that we do not think that the lower two-finger decoding performance exhibited by the NeuRAD is a reflection of its prediction latency (which was not substantially different than the embedded BMI). Rather, as signal-to-noise ratios of recorded units on a Utah array fall over time (Hughes et al., 2021) and the NeuRAD was investigated two years after the embedded BMI, it is more likely that the reduction in signal quality yielded worse real-time two-finger decoding performance.

The low-power BMIs presented here represent a substantial leap in compression of end-to-end BMIs without sacrificing performance. There have been no other devices presented that provide front-end amplification, feature extraction, and decoding at a continuous update rate in one low-power package, but many have attempted to tackle the issue of front-end power consumption. The most similar device is the Brown Wireless Device, which captures and transmits broadband activity from 96 channels at 20kSps with 12-bit resolution per channel in humans (Simeral et al., 2021). As presented in past work, this device consumes 51mW (Yin et al., 2013, 2014) or 90.6mW in its implantable form, likely driven by the need to wirelessly transport high-bandwidth neural data despite not including feature extraction or decoding capabilities. The Hermes series of devices have also demonstrated wireless broadband recording functions in nonhuman primates, consuming 63.2mW to record one channel at a time at 15.7kSps (or a newer version supporting 96 channels at 31.25kSps not tested with animals) (Chestek et al., 2009; Gao et al., 2012; Miranda et al., 2010; Santhanam et al., 2007). Recently, venture interest has enabled

substantial compression of wireless BMI technologies beyond what has been demonstrated in the past. In particular, the Neuralink integrated circuit recently showed high-bandwidth recording of 1,024 channels by wirelessly transmitting only the spikes detected on-chip (Yoon et al., 2021). Such interest is encouraging within the academic and industrial BMI field, and end-to-end BMIs like the embedded and integrated ones we have presented demonstrate how close the technology is to clinical viability.

3.5 Acknowledgements

Chapter 4 Decoding Simultaneous and Independent Finger Movements with Brain-Machine Interfaces

A version of this chapter "Real-Time Linear Prediction of Simultaneous and Independent Movements of Two Finger Groups Using an Intracortical Brain-Machine Interface" was published in Neuron in 2021 (Nason et al., 2021b)*.*

4.1 Introduction

Neural prostheses have the potential to return independence to many people with neurological disorders or injuries. In human clinical trials, laboratories have restored use of computers, self-feeding, and prosthetic hands using implants to translate electrophysiological signals into user intent (Ajiboye et al., 2017; Memberg et al., 2014; Nuyujukian et al., 2016; Pandarinath et al., 2017; Wodlinger et al., 2015). Of greatest interest to people with cervical-level spinal cord injury is the return of hand and arm function (Anderson, 2004). Although this has motivated many groups to study neural prostheses for hand control, only a few have been translated to use with people and none have been translated to full-time use outside of the laboratory. Functional electrical stimulation provides an avenue for outputting the intentions of the user to their natural limbs (Kilgore et al., 1989, 2008; Memberg et al., 2014; Smith et al., 2005), with commercial solutions having already existed such as the FreeHand System. Unfortunately, they typically rely on some external motion or myoelectric commands from residual functional muscles, which require learning and are generally unnatural to use.

This has driven many groups to use brain-machine interfaces to extract hand prosthesis control signals from a more natural source. In humans, various studies have attempted to characterize the relationship between finger movements and electrocorticography activity (Chestek et al., 2013; Hotson et al., 2016; Kubánek et al., 2009). However, that relationship was insufficiently strong to enable quick classifications or fully dexterous continuous movements. Some groups have used intracortical microelectrodes in monkeys to record activity patterns on the order of single neurons to investigate a more concrete connection between them and finger behaviors (Baker et al., 2009; Mollazadeh et al., 2011). These studies suggest such a relationship reliant on intracortical recordings is stronger than with electrocorticography, but classification of which finger is moving by itself does not provide enough understanding of that relationship to predict the quality of continuous control.

The ability to precisely control the positions (individuated placements of the hand degrees of freedom) of all individual fingers is a key characteristic of dexterous hand use in primates. As such, many groups have attempted to relate many, if not all 27, degrees of freedom (DoFs) within the hand to neural activity during reach-to-grasp tasks offline by continuously predicting hand posture (combined arrangements of different hand degrees of freedom; Aggarwal et al., 2013; Bansal et al., 2011; Okorokova et al., 2020; Vargas-Irwin et al., 2010). These studies found continuously predicting postures exceeded the performance of predicting the movement velocities, which contrasts with brain-machine interface studies for arm reaches. Additionally, although these studies showed very high offline correlation between firing rates and behaviors for many of those DoFs, most, if not all, of the DoFs presented showed highly correlated trajectories and may not have truly been independent. This makes it unclear how well this neural activity corresponds to those DoFs individually or if all the DoFs are moving so similarly that anything with a similar time

78

course will correlate well. Further, without evaluating online control of individual DoFs, the applicability to intuitive and naturalistic neural prostheses is uncertain.

Dexterous hand use in primates and humans has two classes of movement: prehensile (for grasping objects; Napier, 1956; Santello et al., 1998) and non-prehensile (for manipulating objects). While all of these studies have made great strides towards the ability to predict the intended movements of the hand, they have focused almost exclusively on prehensile movements (Napier, 1956; Santello et al., 1998), which has a clear importance towards the goal of returning hand function to those without it. There is a substantial gap in the literature surrounding the neural representation of continuous non-prehensile movements and whether that representation can be leveraged in a brain-machine interface. One study trained a primate to use a joystick to control non-prehensile movements of a virtual hand, eventually using an online decoder to move the virtual hand online (Rouse, 2016). However, it remains unclear how much of the brain control depended on the monkey's intentions to perform non-prehensile hand movements or if the monkey continued to think about its arm mapped to the virtual hand as the control system for the task.

Generally, the algorithms for online control of neural prostheses assume linear relationships between primary motor cortex neural activity and either the position and velocity, or expected muscle activations of prosthetic movements. Variants of Kalman filters (Ajiboye et al., 2017; Gilja et al., 2012; Malik et al., 2011; Wu et al., 2004b), ridge regressions (Collinger et al., 2013; Mulliken et al., 2008; Wodlinger et al., 2015), and Wiener filters (Ethier et al., 2012; Koyama et al., 2010; Sachs et al., 2015) have been used to control arms, hands, and fingers online. Linear online decoders are promising candidates for an out-of-laboratory clinical neural prosthesis due to their computational simplicity and high prediction performance. However, with limited quantities of recording electrodes, covariances between nearby neural signals, and increasing

numbers of DoFs required for finger control, linear decoders may be unable to accommodate multiple independent degrees of freedom.

It has been noted by several groups that the same neurons can covary with substantially different behaviors, which could make the prediction of finger movements particularly difficult. For example, primary motor cortex can simultaneously encode information about upper extremities, fingers, and speech, independent of body laterality (Cross et al., 2020; Diedrichsen et al., 2013; Heming et al., 2019; Jorge et al., 2020; Stavisky et al., 2019, 2020; Willett et al., 2020). As tasks increase in complexity, linear models may be unable to discriminate between neural states without sampling greater quantities of relevant neurons. Therefore, it is valuable to characterize the limits of linear models in discriminating neural states with truly simultaneous movement of independent DoFs.

Here, we show for the first time, fine, independent, and simultaneous online control of two systematically individuated groups of fingers within one hand to acquire two targets, one each for the index finger and the middle-ring-small (MRS) fingers, in a non-prehensile task using linear Kalman filters and an intracortical brain-machine interface in nonhuman primates. With intention-based retraining of the Kalman filters, we find that online brain control improves significantly. Then, we find that the magnitude of individual neural activations to particular non-prehensile movements, whether they correspond to movements of one group or combined movements of both groups, can be well predicted by the weighted sums of the most similar movements. This suggests that neural representations of continuous, non-prehensile finger movements are related by linear combinations of similar movements, enabling us to accurately predict untrained finger movements offline using ridge regression. Finally, we characterize the similarity between postural and

movement tuning of cortical spiking activity to fingers to find that the preferred movement

direction is rarely associated with the preferred posture, regardless of the beginning posture.

4.2 Methods

4.2.1 Implants

We implanted two male rhesus macaques (monkey N age 7 to 8, monkey W age 8 at the

time of data collection), with Utah microelectrode arrays (Blackrock Microsystems, Salt Lake

City, UT, USA) in the hand area of primary motor cortex, as described previously (Irwin et al.,

Table 4.1. Abbreviations

Abbr.	Complete Phrase	Description
SBP	Spiking Band Power	Mean-absolute value of the 300-1,000Hz filtered intracortical voltage recording
MRS	Middle/Ring/Small	One of the finger dimensions investigated, in which the middle, ring, and small digits moved together
R	Rest	A posture between flexion and extension where the monkey's digits were both in a resting state
IF	Index Flexion	A posture in which the index finger group was flexed and the MRS finger group was at rest
IF+MF	Index Flexion + MRS Flexion	A posture in which both the index and MRS finger groups were flexed
MF	MRS Flexion	A posture in which the index finger group was at rest and the MRS finger group was flexed
IE+MF	Index Extension + MRS Flexion	A posture in which the index finger group was extended and the MRS finger group was flexed
IE	Index Extension	A posture in which the index finger group was extended and the MRS finger group was at rest
IE+ME	Index Extension + MRS Extension	A posture in which both the index and MRS finger groups were extended
ME	MRS Extension	A posture in which the index finger group was at rest and the MRS finger group was extended
IF+ME	Index Flexion + MRS Extension	A posture in which the index finger group was flexed and the MRS finger group was extended

2017; Vaskov et al., 2018). Pictures of the implants are illustrated in Figure 4.1D. Only motor cortex arrays were used in this study. Monkey N's motor cortex arrays are two 64-channel arrays, explaining why they are smaller in area than his sensory array. Monkey N was between 117 days and 708 days post-cortical implant for all data analyzed. Monkey W was between 78 days and 100 days post-cortical implant for all data analyzed.

In a separate surgery, we implanted monkey N with chronic bipolar intramuscular electromyography recording electrodes (similar to PermaLoc^{TM} electrodes, Synapse Biomedical, Inc.). After induction of anesthesia, the monkey was positioned supine for easy access to his left arm. A single radial-volar incision was first used to access flexor muscles of the deep and superficial compartments of the forearm, following which a single dorsal-ulnar incision was used to access the extensor muscle of the forearm. For each muscle of interest, intra-operative neural stimulation of the muscle was performed to isolate finger-related and wrist-related actions of interest. Electrodes were secured intramuscularly using non-absorbable monofilament suture at a location in close proximity to the entry point of the innervating nerve. In instances where on neural stimulation, better isolation of intended muscular movement occurred at a site distal to the identified neural entry point, an additional electrode was secured at this distal site. Isolated muscles included the flexor digitorum profundus-index (1x near the nerve entry point, 1x distal near the wrist), flexor digitorum profundus-MRS (1x), flexor pollicis longus (1x, unused in this study), flexor carpi radialis (1x), flexor carpi ulnaris (1x), extensor digitorum communis (1x), extensor indicis proprius (1x), extensor carpi radialis brevis (1x), and extensor pollicis longus (1x, unused in this study). After all electrodes were secured to the muscles of interest, electrode wires were tunneled proximally to the upper arm using an incision on the posterior upper arm, posterior to the elbow, ensuring sufficient redundancy and laxity on the lengths of the wires to account for motion

82

at the elbow. An additional interscapular incision was used as an exit site for the tunneled wires which were then connected to the standard PermaLoc^TM connector. All incisions were closed in a layered fashion using absorbable sutures. Following the implantation of these electromyography recording electrodes, the monkey persistently wore a Primate jacket (Lomir Biomedical, Inc.). Monkey N was 36 days post-arm-implant for all electromyography data analyzed.

4.2.2 Feature Extraction

All processing was done in Matlab versions 2012b or 2018a (Mathworks, Natick, MA, USA), except where noted.

Threshold crossing rates were processed and synchronized in real-time during the experiments (see the subsequent section for a description of data flow). We configured the Cerebus neural signal processor (Blackrock Microsystems) to extract voltage snippets that crossed a -4.5 times the root-mean-square (RMS) threshold, customized to each channel. Then, these waveforms were streamed to a computer running xPC Target version 2012b (Mathworks), which logged the source channel of each spike and the time of each spike's arrival relative to all other real-time experimental information. Both monkeys had 96 channels of threshold crossing rate data analyzed, though for closed-loop decoding, channels were masked to those that were not clearly disconnected and had contained morphological spikes during the experiment or at some time in the past (see SBP section below for reasoning).

We also extracted sorted unit firing rates for offline analyses. We imported the relevant broadband (0.1Hz – 7.5kHz sampled at 30kSps) recordings into Offline Sorter version 3.3.5 (Plexon, Dallas, TX, USA). Then, we high-pass filtered the recordings with a 4-pole Butterworth filter with a cutoff frequency set to 250Hz. To sort clear units, we used the threshold level determined during the experiment by the Cerebus at -4.5RMS, then eliminated clearly artifactual

threshold crossings. Then, we sorted the remaining spikes crossing the threshold individually and in combination with principal component analysis, clusters as determined by k-means or Gaussian mixture model clustering (as implemented in Offline Sorter), and visual inspection. The spike timings of each sorted unit were then re-synchronized with the experimental data offline. After all sorting, monkey N had 47 units and monkey W had 68 sorted units.

Spiking band power was also acquired in real-time by the same experimental system as threshold crossing rates. We configured the Cerebus to band-pass filter the raw signals to 300-1,000Hz using the Digital Filter Editor feature included in the Central Software Suite version 6.5.4 (Blackrock Microsystems), then sampled at 2kSps for SBP. The continuous data was streamed to the computer running xPC Target, which took the magnitude of the incoming data, summed all magnitudes acquired in each 1ms iteration, and stored the 1ms sums as well as the quantity of samples received each 1ms synchronized with all other real-time experimental information. This allowed offline and online binning of the neural activity with 1ms precision. As with threshold crossing rate, we masked channels for closed-loop decoding to those that were not clearly disconnected and had contained morphological spikes during the experiment or at some time in the past, as SBP could possibly extract firing rates of low-SNR units remaining represented on such channels (Nason et al., 2020).

4.2.3 Experimental Setup

The experimental apparatus used for these experiments is the same as described previously (Irwin et al., 2017; Nason et al., 2020; Vaskov et al., 2018). Briefly, the monkeys' Utah arrays were connected to the patient cable (Blackrock Microsystems) and raw 0.1Hz-7.5kHz unfiltered broadband activity at 30kSps, 300-1,000Hz activity at 2kSps, and threshold crossings at a -4.5RMS threshold were extracted from the neural recordings by the Cerebus for storage. The 2kSps and

threshold crossing features were streamed to the xPC Target computer in real-time via a User Datagram Protocol packet structure. The xPC Target computer coordinated several components of the experiments. It binned threshold crossings and SBP in customizable bin sizes, coordinated target presentation, acquired measured finger group positions from one flex sensor per group (FS-L-0073-103-ST, Spectra Symbol, Salt Lake City, UT, USA), and transmitted finger positions along with target locations to an additional computer simulating movements of a virtual monkey hand (MusculoSkeletal Modeling Software) (Davoodi et al., 2007). Task parameters, states, and neural features were stored in real-time for later offline analysis.

4.2.4 Behavioral Task

We trained monkeys N and W to acquire virtual targets with virtual fingers by moving their physical fingers in a more complex version of the two-finger task we published previously (Nason et al., 2020). During all sessions, the monkeys sat in a shielded chamber with their arms fixed at their sides flexed at 90 degrees at the elbow, resting on a table. The monkeys had their left hands placed in the manipulandum described previously (Vaskov et al., 2018). Each monkey sat in front of a computer monitor displaying the virtual hand model and targets described previously. The monkeys were trained to move one finger group independent of another finger group. Monkey N preferred to perform the task with his index finger individuated from his MRS fingers, but monkey W preferred his index and middle fingers to be individuated from his ring and small fingers. The fingers of the virtual hand were split into an index finger group and a MRS finger group for both monkeys, where the fingers of each group moved together. Monkey N's index finger actuated the virtual index finger and his MRS fingers actuated the virtual MRS fingers. In monkey W's case, the movements of his index and middle finger actuated the virtual index finger while the movements of his ring and small fingers actuated the virtual MRS fingers. We later validated

anecdotally that there was no impact on performance due to the mismatch between monkey W's preferred finger split and the split of the virtual fingers by training monkey W to split his fingers in the same way as monkey N (index and MRS) and the virtual hand.

Each trial began with one spherical target appearing along the one-dimensional movement arc of each finger group, for a total of two simultaneous targets. Each target occupied 15% of the full arc of motion of the virtual fingers, except where indicated otherwise. Targets were presented in one of two patterns.

The first pattern represents a classical center-out-and-back pattern when viewed from the two-dimensional behavioral space illustrated in Figure 4.1B (Georgopoulos et al., 1982). Every other target was presented at a rest position, 50% between full flexion and full extension. The non-rest targets were pseudo-randomly selected from the postures in Figure 4.1B and a magnitude of movement was pseudo-randomly chosen (+20%, +30%, or +40% from rest). IE+MF and IF+ME postures did not have a +40% movement magnitude, as that was too far of a split in the finger groups to be performed reliably.

The second pattern represents random targets by pseudo-randomly placing targets for each finger group. First, a pseudo-random finger separation *fsep* was generated between -50% and +50% of the range of motion of each finger group. Then, a pseudo-random central position *cpos* was generated in the range of *abs(fsep)/2* to *1-abs(fsep)/2*. The index finger group was given a target at *cpos+fsep/2* and the MRS finger group was given a target at *cpos–fsep/2*.

The presentation order of the targets was random, though the same order was repeated across sets of trials. Approximately one year after data collection, several experiments with a true-random order of the targets validated no change in performance. For a successful trial, the monkey was required to move the virtual fingers into their respective targets and remain there for 752ms

continuously in manipulandum control or 502ms continuously in brain control, except where otherwise noted. The additional 2ms was an artifact of a minor but long-standing bug that required targets be held for 2ms longer than requested. Upon successful trial completion, the monkeys received a juice reward. Closed-loop decoding experiments used center-out target patterns without the IE+MF and IF+ME posture styles, as the data was collected prior to conception of those postures. At a later time, we trained and tested a ReFIT Kalman filter on the full set of targets in the center-out style (including the IE+MF and IF+ME targets). Tuning analyses used both center-out target patterns and random target patterns.

4.2.5 Neural Activity Normalization

All of our tuning analyses used normalized neural activity. For each sorted unit or SBP channel, we computed the mean, standard deviation, and root-mean-square activity level across the entire experiment in 300ms bins. Then, we eliminated all trials that were not center-to-out (unless the analysis used the random target pattern), that were unsuccessful, or had a duration longer than 3.0s for monkey N, indicating possible distraction, struggle, or other factors that may confound tuning analyses.

For each trial, we estimated the onset of native movements by the following procedure. First, for all finger groups that were not in the target at the beginning of the trial (which could happen in the random target pattern), we found each finger group's main movement time, calculated as the first point in time at which the finger group passed 20% of the distance from its starting position to its ending position. Then, we found each finger group's movement onset time, calculated as the point in time of the maximum jerk, or maximum change in acceleration, for each finger group prior to that group's main movement. The trial's movement onset time was selected as the earliest movement onset time across the active finger groups, where an active finger group

is one which starts outside of its corresponding target. For trials in which all finger groups had starting positions within their corresponding targets, all finger groups were considered active. We found this procedure matched estimated movement onsets visually well for nearly all observed trials.

Given the movement onset, then for each included trial, we extracted the activity for each sorted unit or SBP channel from the period of time that maximized tuning depth for each analysis for each monkey, as the analyses depended on representative tuning curves. In all cases, we analyzed 300ms of activity about movement onset with a starting time between -100 to +100ms relative to movement onset. The specific starting times used for each analysis are detailed in that analysis' methods section below. We then normalized these measurements by subtracting the unit's or SBP channel's mean then dividing by the standard deviation. This results in one normalized level of activity for each unit or SBP channel for each trial.

4.2.6 Computation of True and Predicted Tuning Curves

We calculated tuning curves for each monkey using one isolated experiment during which monkey N performed 2,130 and monkey W performed 1,836 manipulandum control trials of the center-out-and-back task, with monkey W performing the task with targets encompassing 16.5% of the range of motion of each finger group. We eliminated trials that were not center-to-out, that were unsuccessful, and, for monkey N only, that had a duration longer than 3.0s, then calculated the mean activity for each movement direction. For each monkey, we selected a start time relative to movement onset for the 300ms analysis window that maximized the tuning depth, measured as the difference between maximum and minimum mean activities across all movements. We found this start time to be 50ms after movement onset for monkey N and 100ms prior to movement onset for monkey W. Then, we further eliminated all trials that had movement onset times less than the

analysis window's start time, leaving 1,025 trials for monkey N and 819 trials for monkey W. All of these trials were used for the ridge regression analyses (see section below), while only trials in which the monkeys moved their fingers ±30% from rest were used for tuning analyses to guarantee consistency in behavior and ideally consistency in neural activity (all of monkey W's trials were ±30% from rest). We did not use the ±40% trials as we did not offer ±40% targets for the IE+MF nor the IF+ME postures due to difficulty of acquisition, leaving ±30% the greatest magnitude of movement with targets presented for all postures.

To compute sinusoidal tuning curves, we first fit the amplitude, phase, and offset by linear regression as was done by others (Chestek et al., 2007). However, we found that this resulted in few sinusoidal tuning curves that correlated significantly with the measured tuning curves (26 of 148 modulated SBP channels and 31 of 92 modulated sorted units were significantly correlated). We anecdotally found this was a result of the period of the optimal tuning curves being unequal to 360 degrees. Therefore, to better represent sinusoidal tuning curves, we fit the amplitude, period, phase, and offset by using Matlab 2018a's `fminsearch` function minimizing the total squared error over a maximum 1,000,000 iterations. These tuning curves yielded the results included above.

We tested three models of predicting the trial-averaged firing rates of all units corresponding to some movement given the trial-averaged activity of the units for other movements. The first naïvely assumes that the firing rates corresponding to one movement is the average of the firing rates of its component movements (i.e. the activity of IF+MF is the average of the activity of IF and the activity of MF). The second and third assume that the firing rate n of some movement m is a weighted sum of the firing rates corresponding to the other movements

along with some constant offset c_0. This can be modeled by the following linear equation for one channel with an arbitrary k number of movements:

$$n_m = c_0 + \sum_{i \in [1,k], i \neq m} c_i n_i$$

Let $n_{m,u}$ be the firing rate from neural unit u for movement m. Then N_m is the vector of firing rates from all neural units corresponding to movement m and N_D is the training data matrix containing firing rates from all neural units corresponding to all of the movements that are not movement m:

$$N_m = \begin{bmatrix} n_{m,1} \\ n_{m,2} \\ \vdots \end{bmatrix}$$

$$N_D = \begin{bmatrix} n_{1,1} & n_{2,1} & \cdots & n_{m-1,1} & n_{m+1,1} & \cdots & 1 \\ n_{1,2} & n_{2,2} & \cdots & n_{m-1,2} & n_{m+1,2} & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

As such, we can solve for the vector of coefficients for one movement m by solving the following linear regression equation using the data given by all valid units:

$$C_m = N_D^T (N_D N_D^T)^{-1} N_m$$

The three models we tested assumed that n_m is composed of some amounts of its component movements, which are n_{m-1} and n_{m+1} in this formulation. The first model (Avg) assumes n_m is the average of n_{m-1} and n_{m+1}, requiring no regression to learn weights, as illustrated in equation (5.3) below:

$$n_m = \frac{n_{m-1} + n_{m+1}}{2} \tag{4.1}$$

The second model (LR) assumed that n_m is composed of some learned weighted amounts of n_{m-1}, n_{m+1}, and a learned offset, as illustrated in equation (5.3) below

$$n_m = c_{m-1} \cdot n_{m-1} + c_{m+1} \cdot n_{m+1} + c_0 \tag{4.2}$$

where c_i are weights learned by regression. Finally, the third model (LRO) assumed that n_m is composed of some learned weighted amounts of n_{m-1}, n_{m+1}, a learned offset, and the firing rates corresponding to the opposite movement (included as a regression method due to the weak yellow diagonal four diagonals off of the main diagonal in the grids of Figure 4.5), as illustrated in equation (4.3) below:

$$n_m = c_{m-1} \cdot n_{m-1} + c_{m+1} \cdot n_{m+1} + c_{m+4} \cdot n_{m+4} + c_0 \tag{4.3}$$

where c_{m+4} is the weight learned for the opposite movement, in this case $+4$ due to there being eight movement directions in our center-out task. We performed these same procedures using SBP in place of the firing rates.

We predicted the neural activity for each movement based on the two regression models with leave-one-out cross-validation on the neural features. The cross-validated linear models for predicting sorted unit activity were trained only on the other sorted units within each monkey. The same was done for SBP channels. Sorted units and SBP channels that had no significantly tuned movements according to a two-sample two-tailed Kolmogorov-Smirnov test ($p > 0.001$) and sorted units with mean firing rates under 2Hz were excluded from analyses. This left a total of 148 SBP channels (92 from monkey N and 56 from monkey W) and 92 sorted units (35 from monkey N and 57 from monkey W). Significance of the predicted tuning curves was determined using the p-values associated with correlation against the null hypothesis that a predicted tuning curve was not significantly correlated with its measured tuning curve, corrected for false discovery rate.

4.2.7 Computation of Movement and Postural Tuning

We calculated movement and postural tuning curves for monkey N using one isolated manipulandum-control experiment and for monkey W using two isolated manipulandum-control experiments across two consecutive days. After eliminating unsuccessful trials, we found that

analysis window start times of 100ms after movement onset for monkey N and 0ms after movement onset for monkey W optimized tuning depth. Eliminating trials with movement onset times less than each monkey's analysis start time resulted in 1,353 center-out trials and 870 random trials for monkey N and 452 center-out trials and 352 random trials for monkey W.

To compare center-out and random tuning curves, we computed the preferred direction of each channel during each task via the following procedure. First, we normalized each channel's SBP activity across each set of trials in 100ms bins, then we computed the Kalman filter observational model (the C matrix in the section below) for each task. We used the coefficients trained for each finger group's velocity to estimate each channel's preferred movement direction within each task, then found the difference in angle between the preferred movement directions. To determine significance, we performed a bootstrap analysis similar to that performed previously (Chestek et al., 2007). Briefly, we took the normalized SBP activity from each task computed before, then resampled equal quantities of samples from the center-out and the random set of trials twice. We trained two Kalman filter observational models and calculated each channel's difference in preferred movement direction between the two trained models. We performed this procedure 1,000 times to estimate the null difference distribution for each SBP channel, then calculated the p-value of each channel's true difference in preferred movement direction between center-out and random tasks based on those distributions. We performed false-discovery rate corrections across the 80 channels with the largest velocity coefficients across monkeys, 40 from each.

The postural tuning surfaces were computed from the normalized SBP activity during the target acquisition periods. For monkey N, the activity was extracted from 252ms to 752ms after onset of the target acquisition period. For monkey W's center-out trials, the activity was extracted from 102ms to 302ms after onset of the target acquisition period (which was 302ms in duration),

and for his random trials, the activity was extracted from 252ms to 502ms after onset of the target acquisition period. These time periods were chosen based on the task parameters to maximize the amount of SBP activity analyzed but keep the true finger velocities nearest to 0. We sorted all trials into postural bins of width 10%, ranging from 0% to 100% flexion of each finger group, based on the mean posture of each trial during the same analysis periods detailed above. Each channel's SBP activity was averaged across all trials belonging to each postural bin, generating the postural tuning surface. The generated postural tuning surfaces were plotted using Matlab's *pcolor* function and linearly interpolated between neighboring postures. The preferred postural direction for each channel was calculated from the Kalman filter observational model's two trained coefficients for finger position.

To determine if the preferred postural direction was significantly different from the preferred movement direction, we performed a 1,000 iteration bootstrap analysis similar to what was detailed previously. Instead of computing preferred postural and movement directions from two different resamples of the random task data, the bootstrap preferred postural and movement directions were computed from the same set of resampled data to estimate the null difference distribution.

4.2.8 Decoding of Neural Activity

We executed two different algorithms to predict finger movements from neural activity.

4.2.8.1 Closed-Loop Kalman Filtering

First, to assess neural prosthetic performance in application, we gave the monkeys visual feedback of the decoders' outputs during the behavioral task. For each closed-loop experimental session, the monkeys began by completing at least 350 trials with the virtual hand controlled

93

directly by the movements of the manipulandum. The monkeys were required to acquire and hold the targets for 752ms continuously for a successful trial with a 10s trial timeout. The behavioral data (i.e. one-dimensional positions per finger group) were measured synchronously with the non-normalized neural features by the xPC Target computer. Then, we trained a standard position/velocity Kalman filter on this data binned at 32ms (which we found superior to other tested bin sizes), as described previously (Irwin et al., 2017), using Matlab version 2012b (Mathworks). For predictions of two finger dimensions, the Kalman filter assumed a kinematic state of one position and one velocity for each group:

$$x_t = \begin{bmatrix} P_I \\ P_{MRS} \\ V_I \\ V_{MRS} \\ I \end{bmatrix}$$

The Kalman filter predicts the state at each timestep based on an optimal combination of two different predictions. The first is a prediction made based on the state of the previous timestep, and the second is a prediction made based on a comparison between the measured neural activity and that predicted by the predicted kinematics of the current timestep. This can be summarized by the following equations:

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1}$$

$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(y_t - C\hat{x}_{t|t-1})$$

where y_t is a vector of neural features at the current time step, K_t is the Kalman gain balancing how much the neural activity should contribute to the final prediction, A is the state transition matrix, and C is a linear regression trained to convert kinematics to neural features. Training of these matrices was performed as described previously (Irwin et al., 2017) but extended to account for both index and MRS. The A matrix was fit to take the following form:

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & A_{V_I V_I} & A_{V_{MRS} V_I} & 0 \\ 0 & 0 & A_{V_I V_{MRS}} & A_{V_{MRS} V_{MRS}} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

After training the standard Kalman filter, we computed the Kalman filter's predictions in real-time to actuate the virtual hand independent of the monkeys' physical movements. For successful acquisition, the monkeys were required to acquire and hold the targets for 502ms continuously with a 10s trial timeout for monkey N or a 30s trial timeout for monkey W. The positions displayed with the virtual fingers were computed by adding the predicted velocity to the previous time step's predicted position, with an initial position taken as the true position of the fingers at the time the Kalman filter began running. Monkey N executed some brain control trials with his physical fingers restricted from movement and others without restriction. Monkey W did not have his finger movements restricted during brain control mode. Both monkeys used both the standard Kalman filter and the dual-stage training ReFIT Kalman filter. To train the ReFIT Kalman filters, the monkeys used the standard Kalman filter in closed-loop for at least 250 trials with their fingers unrestricted. Then, we used one of three intention-training paradigms to train the ReFIT Kalman filter.

1. The first is a two-dimensional form of the ReFIT Kalman filter we previously proposed (Vaskov et al., 2018), where we negated the incorrectly predicted one-dimensional velocity for each finger and set the velocities of fingers within their targets to zero before recomputing the regression matrices. This option assumes the monkey viewed the task as two separate one-dimensional problems and is highly related to the visual cues.

2. The second training method is very similar to the original ReFIT Kalman filter training method (Gilja et al., 2012). In the two-dimensional behavioral space illustrated in Figure

4.1B, we rotated the net two-dimensional velocity towards the target's center for every trial and calculated the component index and MRS velocities from the net velocity. This option assumes the monkey viewed the task as one two-dimensional problem and is highly related to our two-dimensional space in which the two one-dimensional targets are mapped to one two-dimensional target.

3. The third training method combines the first two. We concatenated the recalibrated predictions from each method to create a $2n \times d$ predictions matrix, where n is the number of timesteps in the training data and d is the dimensionality of the behavior (5 in this case). Then, we concatenated one copy of the $n \times f$ neural feature matrix (f being the number of features) to match the size of the first dimension of the predictions matrix to recompute the regressions. This option assumes the monkey viewed the task as some combination of two one-dimensional tasks and one-two dimensional task, which reflects some of the behavior seen in Figure B.1.

After training the ReFIT Kalman filter, we again computed the predictions in real-time using the new model and delivered visual feedback of the predictions to the monkey by actuating the virtual hand.

4.2.8.2 Stability of Trained Kalman Filter Parameters

To compare the least-squares regression parameters that best represent the neural activity in each type of virtual hand control (manipulandum, Kalman filter, and RFKF), we computed preferred movement directions for each SBP channel. First, we normalized the SBP activity during each control method, then computed the regression weights that would transform the kinematic measurements into normalized SBP activity. In the cases of the Kalman filter and RFKF, the predicted kinematics were intention-corrected via rotation (type 2) as described in the previous

section prior to regressing. The preferred movement directions for each SBP channel under each control method were extracted from the index and MRS velocity regression parameters.

4.2.8.3 Open-Loop Ridge Regression

When trying to gauge the generalization of linear models to untrained finger behaviors, we used ridge regression (with a regularization term $\lambda = 0.001$) to predict continuous finger positions as it does not depend on iteration for continued stability. The neural activity was binned every 100ms and 10 total bins were used per feature: one for the current time step and one for each of the previous 9 timesteps for a total 1s of neural data. Neural activity from before each trial's beginning was assumed to be 0 for sorted units and each channel's root-mean-square value for SBP, as had been done previously (Chestek et al., 2007).

After sorting all trials as previously described, we split them into two sets: individual and combined finger group movements. Then, we trained a regression model on one set and tested on the other. Additionally, we performed 10-fold cross-validation on the full dataset to gauge information loss without training on the full set of behaviors. To compare these, we performed a 100,000 iteration bootstrap analysis on the errors between the two decodes for all time points in each experiment. Then, if the error for any averaged sample was greater than the upper one-sided 95% confidence interval resulting from the bootstrap analysis, the split decode was deemed significantly different from the full decode. Correlation coefficients were computed as Pearson's r between the predicted and measured behavior, and variances accounted for were computed according to the following equation:

$$VAF = \left(1 - \frac{\mathrm{var}(y - \hat{y})}{\mathrm{var}(y)} \right)$$

where y is the measured ground truth behavior, $\hat{y}$ is the predicted behavior, and VAF is the variance of y accounted for by $\hat{y}$. The variance accounted for percentages, as written in the results, were taken as the ratio between the variance accounted for by the split-trained decoders divided by the variance accounted for by the full-trained decoder, multiplied by 100. The same computation was executed for the correlation coefficient ratios.

4.2.8.4 Performance Metrics

To evaluate performance, we used four metrics. First, success rate was computed as a percentage of the total number of trials for which the monkeys successfully acquired each trial's targets before the trial timed out (10s).

Second, path efficiency was computed as the straight-line distance between the starting position and the nearest point in the target in two-dimensional space (Figure 4.1B) divided by the total distance that was travelled in two-dimensional space until the first instance both finger groups were simultaneously in their respective targets. We implemented this metric to estimate the efficiency of all movements, though it could be used to assess the simultaneity of combined finger group movements. We have included Figure B.1 to indicate the average behavioral trajectories in two-dimensional space and demonstrate how simultaneous were the monkeys' multi-finger movements. Trials in which both finger groups were never in their respective targets simultaneously were given a path efficiency of 0%. Trials where the starting positions were inside the targets and successful trials immediately following failed trials were excluded from path efficiency analyses to avoid artificially increasing path efficiency.

Third, time to target was computed as the total time between a trial's beginning and the first instance both finger groups were simultaneously in their respective targets. Trials in which both finger groups never simultaneously reached their respective targets were given a time to target

equal to the trial timeout period (10s). Successful trials immediately following a failure and trials where all finger groups began inside their respective targets were excluded from time to target analyses to avoid artificially decreasing mean time to target.

Fourth, orbiting time was computed as the time between the first instance both finger groups were simultaneously in their respective targets and the end of the trial, minus the hold time. Orbiting time is a measure of the stopping ability of a controller, where the optimal orbiting time of 0s means the target was acquired and held beginning with the first instance both fingers were in their respective targets. To avoid artificially lowering the mean orbit time, unsuccessful trials that had a corresponding time to target later than 9 seconds into the trial were excluded from orbiting time analyses. Trials that did not have an instance where both finger groups were simultaneously in their respective targets, trials immediately following failed trials, or trials where the starting positions were inside the targets were also excluded from orbiting time analyses as these types of trials could artificially decrease the mean orbiting time.

We opted to exclude the hold time from the calculations so that the computed metrics better represented the time taken to reach the targets. Note that we required the monkey to hold targets for longer time periods when in manipulandum control mode so that there would be sufficient representation of neural data corresponding to stopping when training the online decoder.

4.2.9 Quantification and Statistical Analysis

Statistical analysis for closed-loop decoding experiments (Linear Two-Finger Decoding in Real-Time in the Results section) used a two-sided two-sample t-test between trial statistics in Matlab with significance level $p < 0.001$. Statistical analysis for determining the normality of neural activations for finger movements (Cortical Neurons Show Specificity to Individual Contractions in the Results section) used a two-sided one-sample Kolmogorov-Smirnov test in

Matlab with a significance level $p < 0.001$, corrected for false discovery rate. Statistical analysis

for determining movement tuning (Finger-Tuned Neural Activity Is Linear in the Results section)

used a two-sided two-sample Kolmogorov-Smirnov test compared to the mean activity across the

experiment in Matlab with significance level $p < 0.001$, corrected for false discovery rate.

Statistical analysis for fitting cosine and regression tuning curves to measured tuning curves

(Finger-Tuned Neural Activity Is Linear in the Results section) used the p-values associated with

Pearson's correlation computed in Matlab with a significance level $p < 0.05$, corrected for false

discovery rate. Statistical analysis for determining differences between the split-data decoders and

the full-data decoders (Linear Models Generalize to Predict Untrained Finger Movements in the

Results section) used a 100,000 iteration bootstrap analysis on the errors between the two decoders

in Matlab with a significance level of greater than a one-sided 95% confidence interval. Statistical

analysis for determining differences in tuning curves between the center-out and random tasks

(Misaligned Postural and Movement Tuning in Finger-Related Cortical Units in the Results

section) used a 1,000 iteration bootstrap analysis on the preferred directions for each channel

determined via regression with a significance level of $p < 0.01$, corrected for false discovery rate

among the 40 most-strongly-tuned channels across both monkeys (80 total). Statistical analysis for

determining differences between preferred posture and preferred movement direction (Misaligned

Postural and Movement Tuning in Finger-Related Cortical Units in the Results section) used a

1,000 iteration bootstrap analysis on the preferred postures and movement directions for each

channel determined via regression with a significance level of $p < 0.01$, corrected for false

discovery rate. Statistical analysis for determining differences between maximal and minimal

posture-related activations of channels (Misaligned Postural and Movement Tuning in Finger-

Related Cortical Units in the Results section) used a one-sided two-sample t-test with a significance level of $p < 0.01$.

4.3 Results

4.3.1 Linear Two-Finger Decoding in Real-Time

We first sought to validate that linear decoder models could individuate two systematically separated finger dimensions moving independently and often simultaneously throughout their entire ranges of motion. To do so, we trained two adult male able-bodied rhesus macaques, monkeys N and W, to perform a two-target two-finger task by using a manipulandum (Vaskov et al., 2018). The manipulandum consisted of two flat surfaces, one for each finger group, where each surface was free to rotate about a hinge at the metacarpophalangeal joints and was sized according to the finger group being used to push it (index vs. MRS). This task is illustrated in Figure 4.1A, with a drawing of the manipulandum included. Although the monkeys were presented with two one-dimensional targets, they can be visualized as one two-dimensional target in a two-dimensional space of % index flexion versus % MRS flexion, as shown in Figure 4.1B, without making any assumptions about cortical representations of these finger groups. For real-time decoding, targets were presented in a center-out style, in which every other target presented was located at rest. Non-rest targets were pseudo-randomly chosen from the postures in Figure 4.1B, then a magnitude of movement was pseudo-randomly chosen between +20%, +30%, or +40% flexion or extension from rest. These movement magnitudes can be imagined by further flexing each posture as shown in Figure 4.1B, as shown in Table B.1, and are realized on the vertical axes in Figure 4.2A and B and Figure 4.3A and B. This resulted in 19 total target combinations without IE+MF and IF+ME postures (3 magnitudes for 6 different postures with a central rest target; see 4.2 Methods). During the task, we synchronously recorded neural activity using 96 channels of
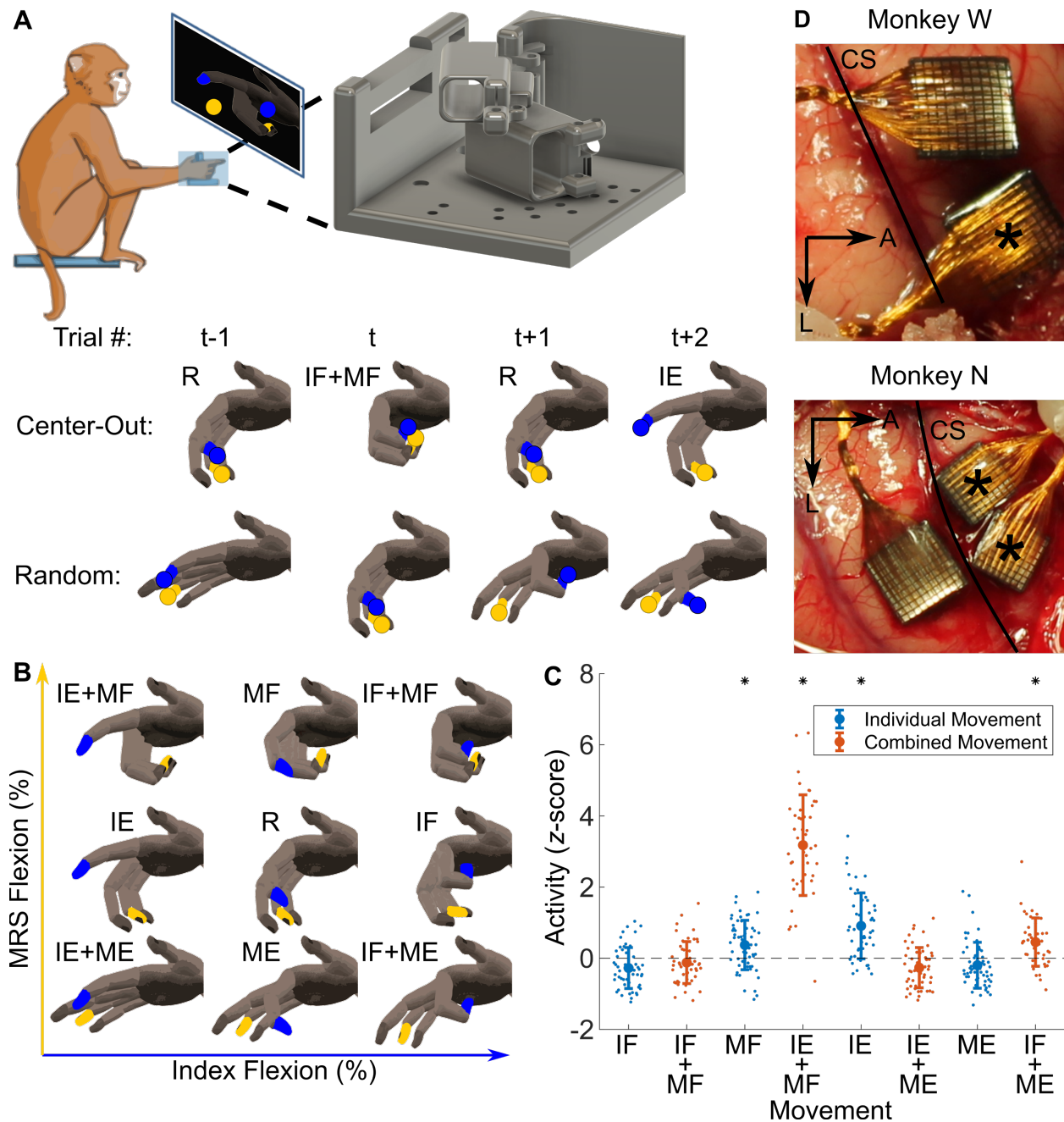
Figure 4.1 Experimental description. (**A**) The monkey was seated in front of a screen displaying a virtual hand with his left hand placed in a manipulandum. Positions of the index and middle/ring/small (MRS) finger groups were measured by the manipulandum (right side of **A**) synchronously with the neural activity. The position measurements or the decoded finger positions were used to actuate the virtual hand, depending on the stage of the experiment. Targets were pseudo-randomly presented in a center-out pattern based on the postures in **B** or in a pattern where target positions were pseudo-randomly placed along each finger's dimension, not separated by more than 50% of the range. (**B**) Two-dimensional space for visualizing the hand movements. Postures shown are at +30% compared to rest, which is at 50% between full flexion and full extension. I – index finger group, M – MRS finger group, F – flexion, E – extension, R – rest. (**C**) An example tuning curve from monkey N illustrating an SBP channel tuned to index extension and MRS flexion movements. Asterisks indicate significant difference from the average activity across the experiment (two-sided two-sample Kolmogorov-Smirnov test, $p < 0.001$, corrected for false discovery rate). (**D**) Photographs of monkey W's and monkey N's intracortical Utah microelectrode array implants. Both implants were in right hemisphere. * indicates arrays used in this study. A – anterior, L – lateral, CS – central sulcus.

implanted Utah silicon microelectrode arrays (Blackrock Microsystems, Salt Lake City, UT, USA)

from the hand area of primary motor cortex in each monkey (implant photographs in Figure 4.1D). Each experimental day, we collected a training dataset in which the monkey controlled the virtual hand using the manipulandum while synchronously recording 300-1,000Hz spiking band power (SBP). We have previously shown that SBP is well correlated with the firing rate of the largest amplitude single unit or units on an electrode and typically results in higher decoding performance than threshold crossing rate (Nason et al., 2020). Then, we trained a Kalman filter as detailed in the methods to predict fingertip velocities in real-time and tested it in closed-loop by actuating the virtual hand according to the predictions.

We found that the monkeys could successfully control the movements of both finger groups independently in real-time using a linear Kalman filter. Figure 4.2A and B show predicted finger traces from monkeys N and W using the SBP Kalman filter, respectively. The sections of predictions displayed were chosen specifically to show a wide variety of targets. These traces demonstrate that the monkeys could individuate the two fingers independently with smooth and controlled effort using the brain-machine interface. Figure 4.2C shows the closed-loop statistics for the two-finger SBP Kalman filter. Regarding path efficiency, monkey N's paths to targets were 61% efficient on average, less than the average 76% efficiency when controlling the virtual hand using the manipulandum. Monkey W acquired targets with an average efficiency of 53% using the decoder, also less than the average 63% efficiency when using the manipulandum. With respect to acquisition time, monkey N reached the target in an average 0.98s with an average orbiting time of 0.86s, greater than the average 0.51s time to target and 0.35s average orbiting time in manipulandum control. Monkey W achieved an average 1.3s time to target with a 1.0s average orbiting time using the Kalman filter, also higher than the average 0.79s time to target and 0.65s
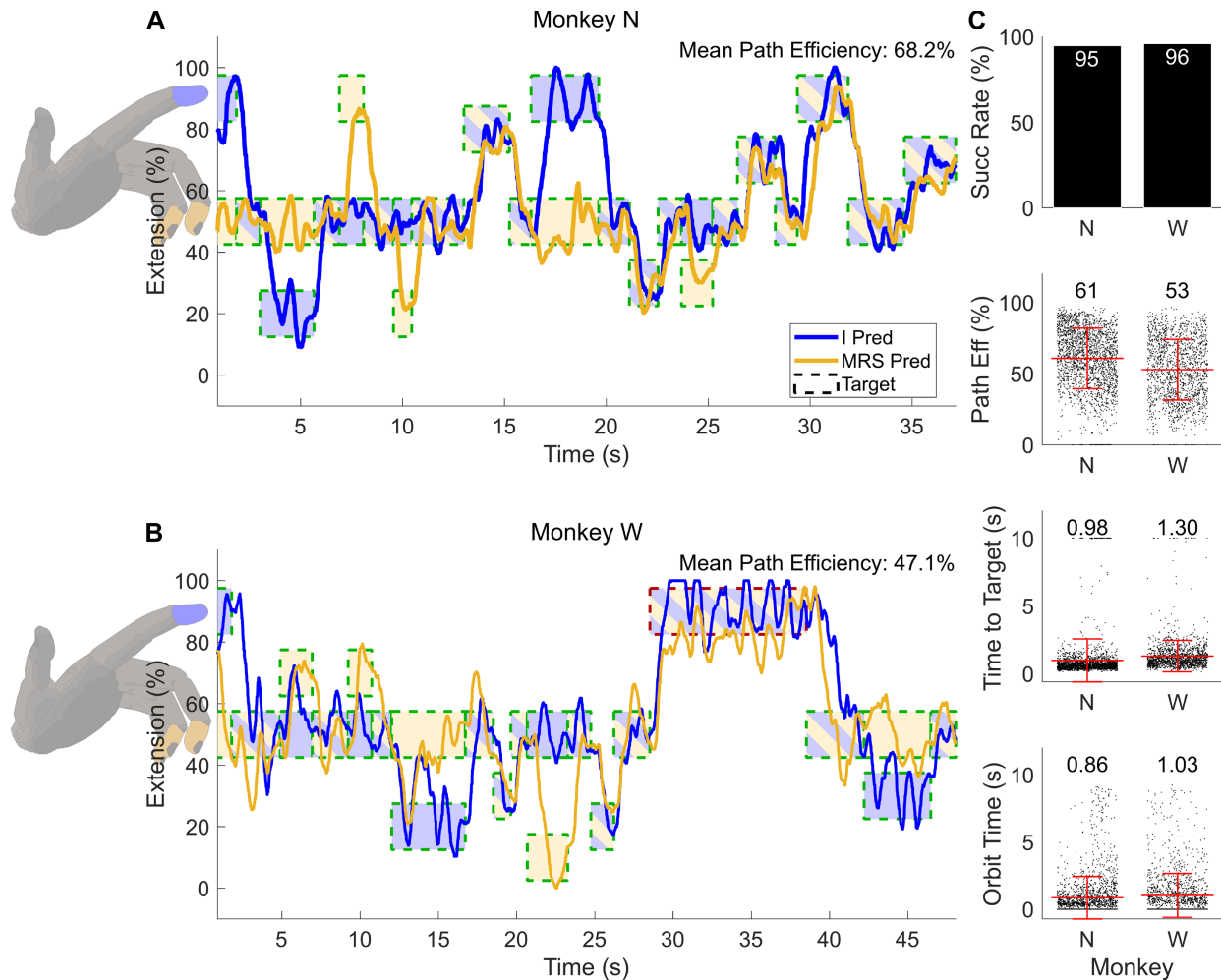
Figure 4.2 Two-finger closed-loop Kalman filter decodes using spiking band power (SBP). (**A**, **B**) Example closed-loop prediction traces from monkeys N and W using the standard Kalman filter, respectively. Targets are represented by the dashed boxes, internally colored to indicate the targeted finger with a border color representing whether the trial was acquired successfully (green, red if not). "I" means the index finger group and "MRS" means the middle/ring/small finger group. The mean path efficiency of the trials displayed in each window is presented at the top right. (**C**) Statistics for all closed-loop two-finger Kalman filter trials for monkeys N (left) and W (right). The red lines indicate the means, which are numerically displayed above each set of data, along with standard deviation. The statistic for each trial is represented by one dot in each plot, split into columns per monkey. "Succ Rate" means the percentage of total trials that were successfully acquired in time and "Path Eff" means path efficiency.

orbiting time when using the manipulandum. For completeness, Figure B.2 illustrates the same for

threshold crossing rate rather than SBP for monkey N.

Out of interest for applications to brain-machine interfaces, we attempted to maximize

closed-loop, two-finger decoding performance using the state-of-the-art recalibrated feedback

intention-trained (ReFIT) Kalman filter (RFKF; Gilja et al., 2012; Vaskov et al., 2018). The RFKF

training procedure occurs following the monkey's usage of the original Kalman filter. It assumes

Figure 4.3 Two-finger closed-loop ReFIT Kalman filter decodes. (**A**, **B**) Example closed-loop prediction traces from monkeys N and W, respectively, using the ReFIT Kalman filter. Targets are represented by the dashed boxes, internally colored to indicate the targeted finger. "I" means the index finger group and "MRS" means the middle/ring/small finger group. The mean path efficiency of the trials displayed is presented at the bottom right. (**C**) Statistics for all closed-loop two-finger ReFIT Kalman filter decodes for monkeys N (left) and W (right). The red lines indicate the means, which are numerically displayed above each set of data, along with standard deviation. The statistic for each trial is represented by one dot in each plot. "Succ Rate" means the percentage of total trials that were successfully acquired in time and "Path Eff" means path efficiency. (**D**) Statistics of each type of velocity reorientation for ReFIT training with monkey N (see Methods). "N" means velocities for each finger were negated if not pointing to that finger's target, "R" means the velocities were rotated in the two-dimensional finger space towards the target, and "B" means both reorientations were used by concatenating velocities modified by N and R and repeating the neural activity. Asterisks indicate significance ($p < 0.01$, two-tailed two-sample t-test).

that the neurons controlling the Kalman filter's predictions represented the intention of the monkey to optimally bring the fingers to the targets, regardless of the directions of the predictions. Then after reorienting predictions to match the presumed intentions of the monkey, the linear model is retrained. As discussed in the methods, the two interpretations of our finger task result in two frameworks for retraining: rotation of the net velocity in two-dimensional finger space to back-calculate each finger's intended velocity (similar to the original ReFIT method) or independent

negation of each finger's velocity if it is moving away from the target. The statistics comparing these two methods and a third combining both are shown in Figure 4.3D for monkey N. Ultimately, there were few statistical differences between the recalibration methods (p-values for the following order of comparisons: rotation and negation, negation and both, rotation and both; $p = 0.34, p = 0.72, p = 0.62$ for path efficiency comparisons; $p = 0.013, p = 0.80, p = 0.064$ for time to target comparisons; $p = 0.0044, p = 0.036, p = 0.81$ for orbiting time comparisons; two-tailed two-sample t-test).

For monkey N, the RFKF made a substantial improvement in decode performance over the standard Kalman filter. Figure 4.3A shows closed-loop prediction traces from monkey N using the SBP RFKF. In comparison to the Kalman filter, monkey N's ReFIT Kalman filters significantly improved prediction performance (68% versus 61% path efficiency, 0.73s versus 0.99s time to target, 0.28s vs 0.86s orbiting time, all significant with $p < 0.001$, two-tailed two-sample t-test). Additionally, the predictions are less oscillatory when attempting to stop on a target, and when it is oscillating, the amplitude is generally smaller than the standard SBP Kalman filter, aligning with what was previously reported (Gilja et al., 2012, 2015; Vaskov et al., 2018).

For monkey W, the RFKF did not improve performance over the standard Kalman filter (54% versus 53% path efficiency, $p = 0.12$, 1.23s versus 1.3s time to target, $p = 0.069$, 1.17s versus 1.0s orbiting time, $p = 0.013$, two-tailed two-sample t-test).

To explain this result, we analyzed the consistency of each monkey's SBP channels between the three modes of controlling the virtual hand: manipulandum control, Kalman filter control using SBP, and RFKF control using SBP. After z-scoring the SBP used to generate all of the data in Figure 4.2 and Figure 4.3, we recomputed three linear regression matrices: one between SBP and the finger kinematics during manipulandum control, the second between SBP and the
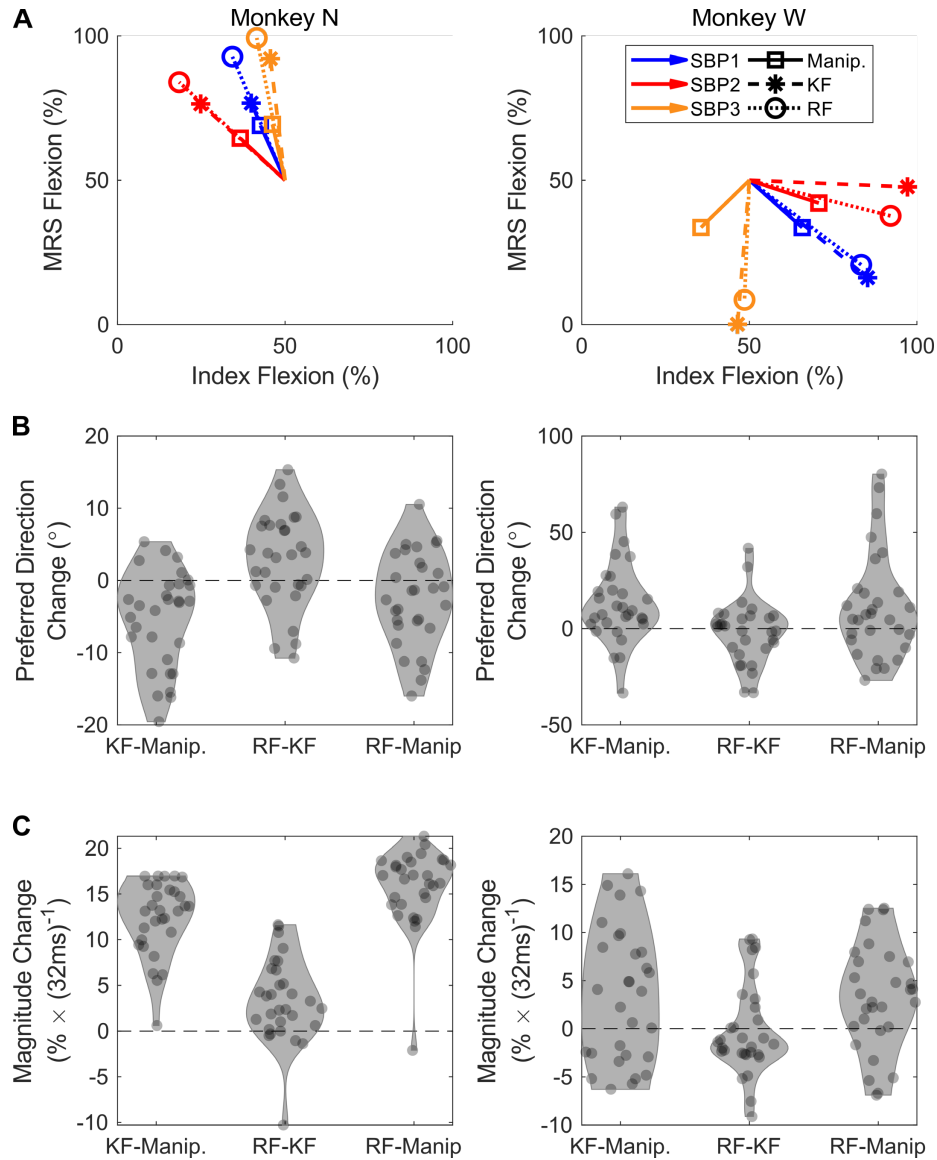
Figure 4.4 Analysis of changes in SBP channel tuning through different stages of online virtual hand control. (**A**) Preferred directions for the three normalized SBP channels with the highest manipulandum-control magnitude on one day for each monkey. Solid arrows with square markers are the preferred direction and magnitude for each channel during manipulandum control (Manip.), dashed arrows with asterisk markers are the same during Kalman filter control (KF), and dotted arrows with circle markers are the same during ReFIT Kalman filter control (RF). (**B**, **C**) Changes in preferred direction (**B**) and magnitude (**C**) of the 5 highest-magnitude manipulandum control channels for each of the 6 days for each monkey.

intention-corrected finger kinematics during Kalman filter control, and the third between SBP and

the intention-corrected finger kinematics during RFKF control. From the regression coefficients

trained for velocity, we estimated the preferred direction of each normalized SBP channel based

on the vector $[C_{n,I}, C_{n,MRS}]$ for coefficients C and channel n. Figure 4.4A illustrates the preferred

directions for the three manipulandum-control channels with the highest magnitudes in each monkey. These illustrative channels demonstrate that the channels that are most impactful for predicting kinematics do not substantially change their encoding preferences between manipulandum control and brain control modes (Figure 4.4B and C). For monkey W, however, there is small variation in the tuning preferences, which may have impacted his capability of using the RFKF to exceed the performance of the Kalman filter. We believe these variations are the direct result of substantially lower motivation to perform the task compared to that of monkey N.

4.3.2 Cortical Neurons Show Specificity to Individual Contractions

We found it surprising that such a complex hand task could be captured so well by a linear decoder. To look at whether these linear relationships hold within individual neurons, we constructed tuning curves of individual units across all eight finger postures at just the +30% magnitude to best guarantee consistent behavior and corresponding cortical activity. On one exclusive and representative day for each monkey, they performed the center-out task in manipulandum control for at least 30 continuous minutes while SBP and broadband activity were recorded synchronously. The broadband activity was spike sorted using Offline Sorter (Plexon, Dallas, TX, USA) to extract the firing rates belonging to sorted units. 388 trials for monkey N and 819 for monkey W were processed. We calculated the mean firing rate for each type of movement and plotted tuning curves. Illustrative tuning curves are displayed in the left plots of Figure B.3.

Initially, we characterized the tuning preferences of the 115 sorted units across both monkeys, which are summarized in Table B.2 and Table B.3. First, in Table B.2 regarding specificity of neural activation to flexion and extension, we found 42 of the 115 units showed specificity to one of the finger flexors or finger extensors, but not the other, suggesting cortical representation of distinct muscle contractions. Second, in Table B.3 regarding specificity to

individuated finger group movements, we found that 27 of the 115 units showed tuning to only one finger group. Among the units that showed specificity to one group, there were approximately equal quantities tuned to movements of the index and MRS finger groups. Most units showed Gaussian-distributed firing rates about the mean firing rate for each movement, with 46 of the 115 units having at least one movement for which the normalized activations were not normally distributed ($p < 0.001$, two-sided one-sample Kolmogorov-Smirnov test, corrected for false discovery rate).

In addition to performing all tuning analyses with the standard sorted units, we also included tuning analyses using the 300-1,000Hz spiking band power (SBP). We have previously shown that filtering spiking signals in the 300-1,000Hz band provides a signal that is highly correlated with the activity of the highest signal-to-noise ratio (SNR) units on an electrode (Nason et al., 2020). This is particularly effective for electrodes with low SNRs, as the 300-1,000Hz band was found to balance the tradeoff between signal and noise power (Irwin et al., 2016). Therefore, we constructed tuning curves from this band, which enabled us to perform the same analyses on sharper tuning curves from primarily multiunit electrode recordings. This resulted in more sources of unit activity than the limited quantities of sortable units on a Utah microelectrode array. As such, we obtained nearly 40% more tuned neural features, from 59 of 115 tuned sorted units to 82 of 192 tuned SBP channels, though there may be substantial overlap in information content between sorted units and SBP. In an anecdotal investigation, we also found that the cross-correlation of SBP features is greater than that of sorted units, likely due to the similarity in background noise (biological in the form of low-amplitude neural spikes but possibly also thermal in origin; Lempka et al., 2011). However, the effect of this appears minute, as the tuning curves for these channels that are included in the left plots of each pair in Figure B.3 as well as Table B.2

and Table B.3 showcase similar preferences to sorted units. Most channels showed specificity to one muscle group and one finger group, with more representation of MRS movements than index movements in monkey N and vice versa in monkey W. Lastly, similar to the sorted units, almost all SBP channels showed Gaussian-distributed power levels about their respective means for all movements across all trials (14 of 192 channels were not normally distributed, $p < 0.001$, two-sided one-sample Kolmogorov-Smirnov test, corrected for false discovery rate).

4.3.3 Finger-Tuned Neural Activity Is Linear

To investigate how the neural activity relates to similar finger movements, we began with a classical cosine tuning analysis. Of the 92 tuned sorted units and 148 tuned SBP channels, 60 and 94, respectively, had significantly correlated sinusoidal fits ($p < 0.05$, Pearson's correlation, corrected for false discovery rate). While sinusoidal tuning of a neuron to two-dimensional movements of a single limb may be reasonable (Georgopoulos et al., 1982), it may not be fundamental (Todorov, 2000), so the same rationale may not apply to movements of multiple limbs, or in our case, fingers. We have demonstrated that a linear Kalman filter enables online separation and individuation of two finger dimensions, so we sought to characterize the goodness of fit of several linear models.

First, we visually analyzed some tuning curves (like those in Figure 4.1C and Figure B.3) and noticed that the neural activity for any particular movement appeared to have a similar activation level for the most similar movements (the closest movements as plotted in Figure 4.1B and C). This motivated us to calculate how well the neural activity for each movement could linearly predict the neural activity for each other movement, independent of any cosine fits (see 4.2 Methods). Figure 4.5 illustrates this, where the activity on each grid's vertical axis is regressed from all of the other activities as indicated on the horizontal axis. In these grids, more yellow (or
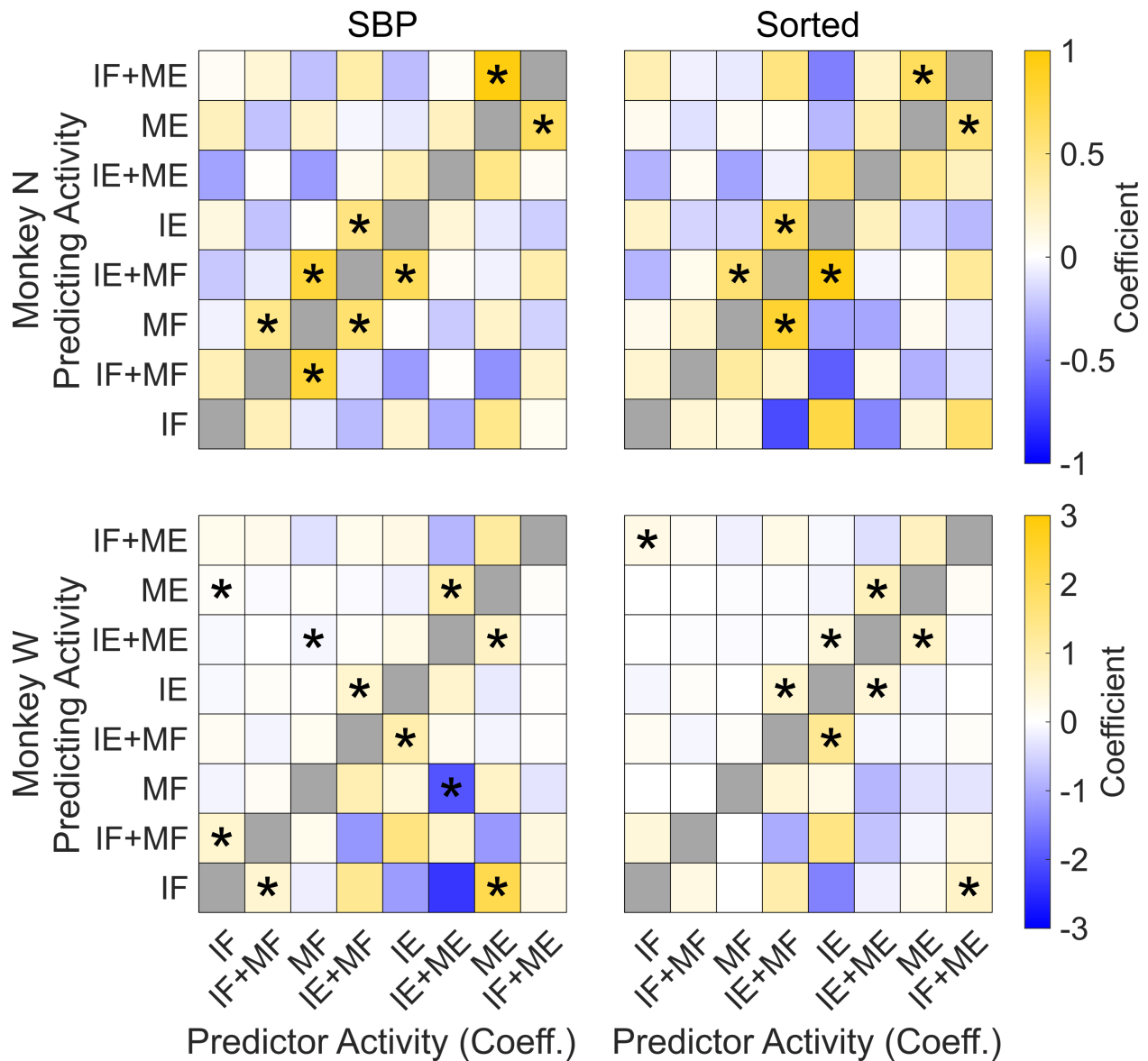
Figure 4.5 Predictability of the activity of each movement from the activities of other movements. The top and bottom rows of axes represent monkeys N and W, respectively, and the left and right columns of axes represent SBP and sorted unit results, respectively. A cell in each set of axes is colored based on the predictor's regression coefficient (horizontal axis) when predicting the activity of a movement (vertical axis). Asterisks indicate statistical significance (based on each coefficient's t-score, $p < 0.001$).

blue) cells indicate greater (or more negative) regression coefficients for the activities on the

horizontal axis to predict the activity on the vertical axis. Generally, we found that the activity

corresponding to most movements could be strongly predicted by the activity of its most related

movements (the yellow off-diagonal bands with substantial significance asterisks). For example,

the SBP activity corresponding to monkey N's MRS-only flexion (MF) can be predicted

significantly by the activities corresponding to index extension/MRS flexion (IE+MF) and index & MRS flexion (IF+MF). This suggests a number of potential linear models that explain the relationships between similar behaviors of different finger groups, three of which are investigated in Figure B.3 and Figure B.4. Overall, this suggests that cortical units represent the combined movements of multiple finger groups in a similar way to how those same units represent the individual movements of those finger groups. However, the optimal combinations of similar movements are to unequally weight their contributions (as illustrated in Figure 4.5 and Figure B.5), explaining why the average model in Figure B.4 failed to significantly predict any tuning curves.

4.3.4 Linear Models Generalize to Predict Untrained Finger Movements

From a decoding perspective, the similarity in neural activations between individual and combined movements may shed some light on the training data requirements for a high performance, multidimensional hand neural prosthesis. Specifically for our two-finger task, in addition to addressing the linearity of neural activity, might a decoder model need to be trained on both individual and combined finger group movements, or is the neural activity for one of the types of movements sufficient to generate a model that encompasses both?

To answer that question, we split the trials (1,018 from monkey N and 819 from monkey W) into two subsets: one representing individual finger group movements only and the other only combined movements of both finger groups. Then, we trained linear regressions with ridge regularization on each set of trials to predict finger positions using exclusively sorted units or SBP and tested them on the corresponding sorted units or SBP of the untrained set of trials. For example, we first trained regressions on *individual* movements of the index or MRS finger groups and used the trained algorithm to decode the *combined* movements of the index and MRS finger groups, and vice versa. Figure 4.6 illustrates the averaged predicted traces from split training for all trials
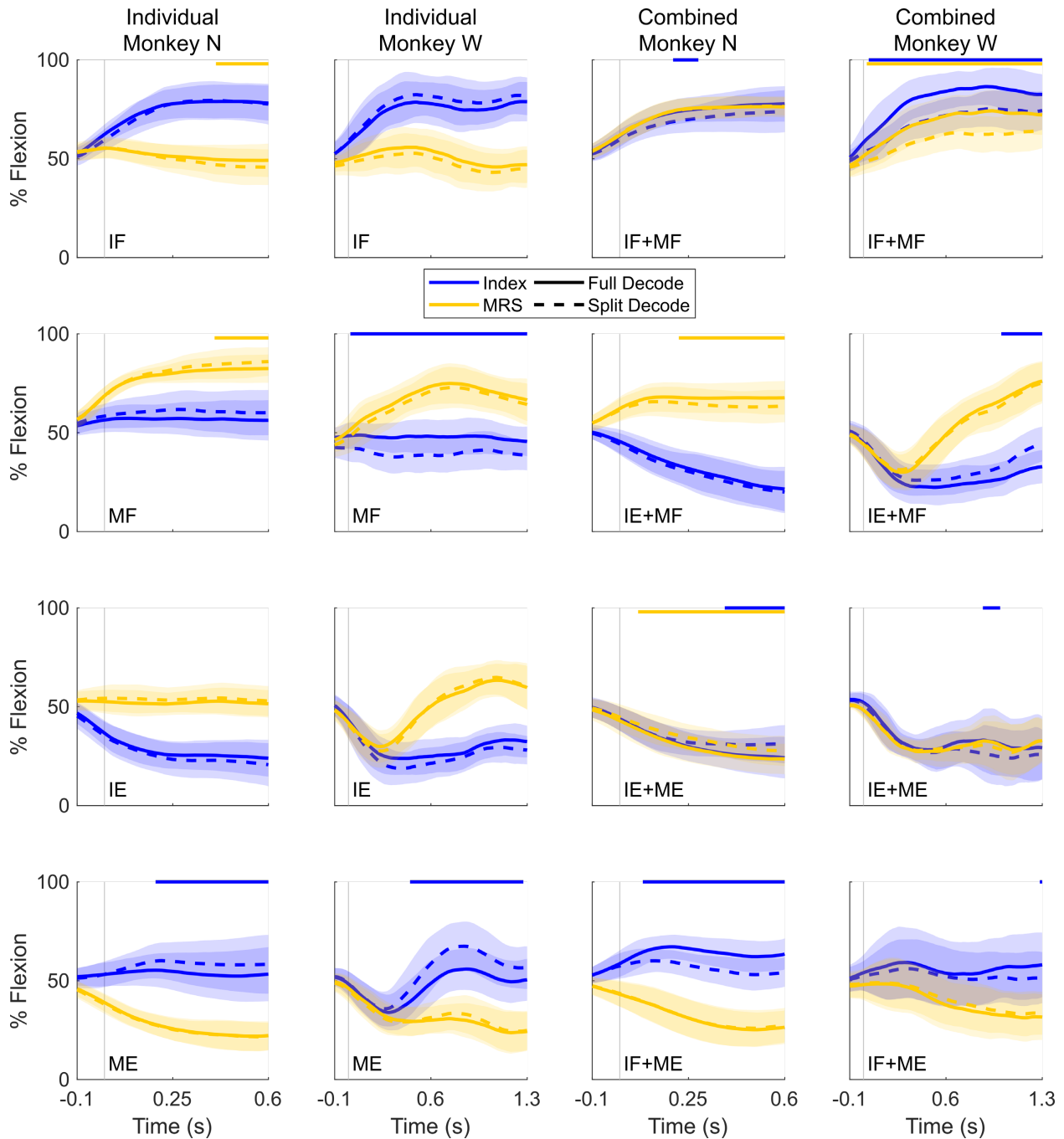
Figure 4.6 Offline ridge regression decoding of all SBP channels, trained on either individual or combined finger group movements. The central trace with shaded region are the average predicted behavior from all trials of the indicated movement ± standard deviation, aligned in time by movement onset (vertical gray line). Individual movements (two left columns) were decoded using a regression trained on combined movements (right two columns), and vice versa for combined movements ("Split Decode" dashed traces). The "Full Decode" solid traces represent the average decode given the full dataset to train the regression, with cross-validation. Blue traces correspond to the index group and yellow traces correspond to the MRS group. The yellow or blue lines near the top of each plot indicate significant differences between the mean predicted positions based on a bootstrap analysis on the differences (>95% one-sided confidence interval).

of a given movement overlaid on the averaged predicted traces for that same movement using a

decoder trained on all trials with cross-validation. Both monkeys' traces show that the split-trained decodes mostly overlay the full-trained decodes, which suggests that finger individuation information in the population's neural activity is preserved across movements, even for untrained behaviors.

However, there are differences during some of the predictions (greater than a one-sided 95% confidence interval of a bootstrap analysis on the errors). Differences in these predicted traces align with the results found during the tuning curve analyses, where the neural activity of particular movements requires unequal contributions of the component movements. As such, by the assumption of linearity, there should be some loss in performance associated with testing on untrained behaviors, which is shown in Figure 4.6. Numerically, for the two monkeys ordered as (N, W), the split-trained SBP decoders achieved (90.5%, 82.4%) of the variance accounted for (VAF) by the full-trained SBP decoder and (95.1%, 90.9%) of the correlation coefficient of the full-trained SBP decoder. For the sorted unit decoders (Figure B.6), the monkeys achieved (85.9%, 82.7%) of the VAF by the full-trained sorted unit decoder and (92.7%, 90.9%) of the correlation coefficient of the full-trained sorted unit decoder. We included the full set of averaged decodes for the eight movements using sorted units for both monkeys in Figure B.6. To avoid any confounds from averaged traces, we also included plots of the individual trials' decodes used to calculate the averages in Figure B.7 and Figure B.8.

4.3.5 Misaligned Postural and Movement Tuning in Finger-Related Cortical Units

It has been well described previously that finger postures can be predicted online from cortical units better than their movements (Aggarwal et al., 2013; Bansal et al., 2011; Okorokova et al., 2020; Vargas-Irwin et al., 2010), but our brain-machine interface used in Figure 4.2 and Figure 4.3 demonstrated high prediction performance for movements, not postures. This raises an

intriguing question related to how finger-related cortical processing circuits and units simultaneously represent movement and posture. Here, we provide a brief investigation into the similarity between cortical tuning to finger postures and cortical tuning to finger movements within our two-dimensional task.

To increase the range of postures the monkey explored at a variety of velocities, we modified the center-out task to be randomized, as illustrated in Figure 4.1A. In this random task, the starting and ending target postures for each trial were randomly determined along the ranges of motion of each finger group, not exceeding a separation of 50% of the range of motion. On one exclusive day for monkey N and across two consecutive days for monkey W, we compared the SBP movement tuning curves between the center-out and random tasks, with examples shown in Figure 4.7A. We found that many of the movement preferences did not substantially change between tasks (11/40 normalized SBP channels with the highest trained linear regression weights for monkey N and 18/40 for monkey W were significantly different in preferred direction, 1000-iteration bootstrap, $p < 0.01$, corrected for false discovery rate). Of those that were significantly different (four examples in Figure 4.7A), the preferred movement visually appeared similar, indicating that any significant differences in preferred movement are likely to be small.

To compare movement tuning with postural tuning, Figure 4.7B overlays the movement tuning curve on a two-dimensional postural tuning surface, where each two-dimensional point is colored based on the mean activity of that channel when the finger groups were stationary at the indicated posture. In those examples, many show alignments between the preferred movement direction (black line out from center) and the preferred postural direction (white line out from center), but surprisingly, most do not show such alignment. Figure 4.7C and D present some statistics for the 40 normalized SBP channels with the highest trained regression weights for each
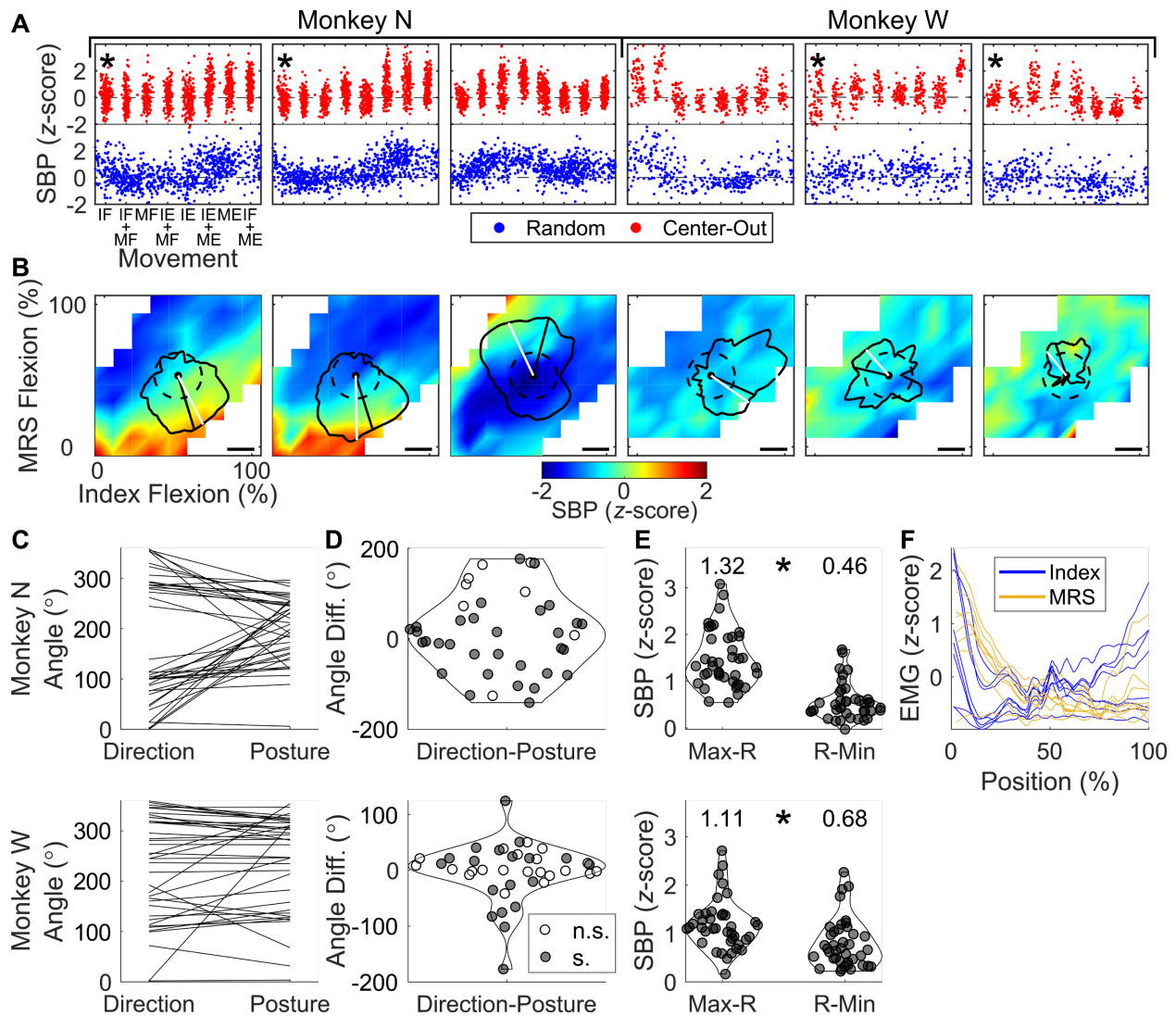
Figure 4.7 Comparison between postural and movement tunings of SBP during the random task with manipulandum control. Dashed traces in all plots represent zero *z*-score. (**A**) Tuning curves for three channels from each monkey. Asterisks indicate significant differences in preferred direction determined via 1,000-iteration bootstrap, $p < 0.01$, corrected for false discovery rate. (**B**) Movement tuning overlaid on postural tuning. The color at each location, or posture, is the SBP activity for the channel above extracted from the target holding period (i.e. at zero velocity). The white line represents the preferred postural direction. The solid black trace represents that channel's SBP activity for movements in each direction, smoothed across 10% of the trials. The black scale bar represents 0.5 *z*-score and the black line from center indicates the preferred movement direction. (**C-E**) Statistics of the plots in (**B**) for the 40 most impactful channels to a linear regression decoder. (**C**) Comparison between preferred movement and postural directions. Each line represents one channel. (**D**) Difference in angle between the preferred movement and postural directions. Each dot represents one channel. The width of the violin at each angle difference indicates smoothed relative density. n.s. means not significant, s. means significant with $p < 0.01$, corrected for false discovery rate. (**E**) Magnitude of difference between the maximum or minimum activity across all postures and rest. The numbers on top represent the median difference. Asterisks indicate statistical difference, $p < 0.01$, one-sided two-sample *t*-test. (**F**) Eight channels of *z*-scored electromyography across all postures during the hold periods.

monkey. 28 of monkey N's 40 channels and 14 of monkey W's 40 channels had a difference

between the preferred posture and preferred movement greater than 30°, despite the distributions

of differences appearing normal about 0°. Furthermore, 31 of monkey N's 40 channels and 19 of monkey W's 40 channels had significantly different preferred directions between posture and movement (1,000-iteration bootstrap, $p < 0.01$, corrected for false discovery rate), affirming misalignment in the preferred posture and the preferred movement direction of a given cortical unit.

Investigating these example postural tuning surfaces further suggests that the strongest cortical representation of the postures occurs at the limits of the range of motion. To characterize this, for each of the same 40 channels from each monkey, we calculated two differences: the first between the maximum activity across all postures and the activity at rest, and the second between the activity at rest and the minimum activity across all postures. Figure 4.7E illustrates these differences, indicating that resting activity is much closer to the minimum than the maximum activity across all postures ($p < 0.01$, one-sided two-sample t-test). This outcome suggests that cortical SBP channels are generally positively tuned towards their preferred posture and increase in activity as a posture changes from a resting state towards its maximum magnitude. This outcome is further emphasized in Figure 4.7F, which plots the smoothed electromyographic activity recorded from bipolar intramuscular electrodes within monkey N's forearm across all postures for each finger group. These traces are generally maximized at the edges of the range of motion with lower activity levels near 50% flexion, suggesting that our estimated resting posture is relatively close to the monkey's true resting posture.

4.4 Discussion

Modern hand neural prostheses have not yet been able to reproduce individuated finger movements across their entire ranges of motion. Here, we demonstrated that a linear Kalman filter with or without intention retraining is adept at decoding individuated, non-prehensile movements

117

of an index finger and a middle-ring-small finger group in real-time with high-performance, given sufficient number of tuned units. We then showed that both monkeys could maintain or substantially improve their control performance using the ReFIT Kalman filter, which trained consistent weights per input channel with the prior decoder models. To address how this complex task can be explained with a linear decoder, we found that the neural activity of any particular movement can be reliably predicted from the activity of its related movements via a weighted sum, providing an explanation for our high performance linear decoding model. We validated this claim by showing that combined finger group movements can be predicted with a decoder model trained on individual finger group movements and vice versa. Only a slight performance loss was realized when comparing predictions of decoders trained on the full set of behaviors and decoders trained only on trials representing individual or combined finger movements. Finally, using our non-prehensile task allowed us to begin to uncover how cortical neural activity simultaneously encodes information about postures and changing postures, suggesting that preferred movement directions and preferred postures are frequently misaligned.

The complexity of our 2D task proposes interesting questions regarding the interpretations and intentions of the monkeys performing it. By analyzing the tuning preferences of cortical spiking activity during manipulandum, Kalman filter, and RFKF control modes, it seems that cortical spiking activity was very consistent between all three. Any changes in decoder weights can likely be attributed to differences in the training data between control methods rather than neurons altering their tuning. Additionally, we proposed two frameworks for retraining linear Kalman filters based on the monkeys' intentions, one assuming each finger is its own task for two total objectives and the other assuming both fingers belong to one two-dimensional task. While both improved decoding performance substantially, neither strongly outperformed the other

despite making clearly different assumptions about the monkeys' interpretations of the task. As such, we think these data suggest two possible truths: the nuances of intention-based recalibration methods for the prediction of continuous two-finger movements will not significantly impact performance, or the optimal recalibration scheme for multiple finger movements remains unknown.

While others have found that decoding continuous postures of the hand achieves greater performance than decoding the movements (Aggarwal et al., 2013; Bansal et al., 2011; Okorokova et al., 2020; Vargas-Irwin et al., 2010), we were nonetheless able to achieve high performance by decoding the movements directly. The non-prehensile task explored here enabled us to begin to uncover the relationship between posture and movement representations in single cortical units. We surprisingly found that a unit's preferred posture is not necessarily related to that unit's preferred movement direction. From a decoding perspective, we hypothesize that this could explain the high-performance control the monkeys achieved using the Kalman filters. If the SBP activity can simultaneously represent movement along one behavioral dimension and posture along an orthogonal behavioral dimension, even simple decoders like linear regressions can parse each channel to provide two pieces of helpful information. Substantial additional work will be necessary to make such conclusions, particularly because the implants used in our study heavily under-sample cortical populations.

After determining that we could predict the level of cortical activation during combined movements as weighted sums of the activations during individual movements, independent of any cosine fits, we validated this finding by predicting untrained, combined finger movements using linear regressions trained only on individual finger movements. We did not expect the positions predicted by the split decoder and the positions predicted by the full decoder to be so similar, which

they were in many cases. We propose a few potential explanations. First, this could be attributed to using behaviors of different amplitudes (±20%, ±30%, and ±40% from rest, see Behavioral Task in Methods) for each direction, such that when plotting just the middle amplitude, the differences in predictions were negligible. Second, since the tuning curves showed graded activations to several related movements, combinations of the contributions from different units and SBP channels may be sufficient to predict movements of greater magnitude. Third, finger-related neural activity may have a minor non-linearity, either between units and SBP channels or with its relationship to the behavior, which has been reported previously (Naufel et al., 2019).

Our real-time decoding results suggest that linear models can accurately fit the movements of two independent fingers across their full ranges of motion from cortical spiking activity. Our previous study originally suggested that this may require nonlinear models (Vaskov et al., 2018). However, at that time, the monkeys had not been trained to voluntarily individuate their finger movements in manipulandum control mode and were not penalized for making undesired movements. In this manuscript, the monkeys received substantial additional training without using the brain-machine interface and with the necessary penalties for incorrect movements to voluntarily individuate their fingers, indicating linear models are sufficient for fitting such movements. Linear decoding of individuated finger movements has major implications for clinical neural prostheses. It has been previously suggested by unconstrained finger movements that primary motor cortex linearly predicts movements of the hand and fingers in monkeys (Aggarwal et al., 2013; Ethier et al., 2012; Kirsch et al., 2014; Okorokova et al., 2020) and humans (Ajiboye et al., 2017; Wodlinger et al., 2015) when many of the DoFs may move along very similar trajectories. Our results extend these to show that linear models can accurately predict the movements of two well-separated and independent finger DoFs across their continuous ranges of

motion. On the front of individuation, some other groups have demonstrated that intracortical arrays can classify movements of individuated digits in humans (Bouton et al., 2016; Jorge et al., 2020). Our methods and results, taking advantage of able-bodied primates and the precise tracking of finger movements with the manipulandum, bring the state-of-the-art in finger individuation from discrete decoding to the realm of continuous decoding. Finger-related neural activity in primary motor cortex recorded with human-grade Utah microelectrode arrays can be sufficient to individuate at least two systematically separate DoFs using linear models. Further, linear individuation of two DoFs within the hand, as shown here, suggests the possibility of linear models sufficiently individuating more of the 27 DoFs within the hand, as has been hinted for low-magnitude movements (Kirsch et al., 2014).

To close the gap between the capabilities of hand neural prostheses and the natural hand, further work is needed to completely characterize the cortical representation of non-prehensile movements as well as how cortex manages movements that are simultaneously prehensile and non-prehensile (i.e. playing a guitar or rapidly solving a puzzle cube). Co-contraction of finger-related muscles makes characterizing the relationship between motor cortex activity and muscle activity difficult (Häger-Ross and Schieber, 2000; Lang and Schieber, 2004). While we have briefly investigated how cortical spiking activity and the electromyography of finger muscles corresponds to our primarily kinematic non-prehensile task, further investigation is required to uncover the impact variable forces have on the relationships characterized here. Usage of the thumb is particularly central to this question, which will be difficult to address with macaques because of the restricted functionality of their thumbs compared to those of humans.

In terms of clinical viability of a neural prosthesis, where usage outside of a laboratory, hospital, or rehabilitation environment is the ultimate goal, the computational simplicity and

generality of linear models make them promising solutions. Decoders such as the Kalman filter, Wiener filter, or ridge regression (Collinger et al., 2013; Ethier et al., 2012; Malik et al., 2011) require surprisingly few computations per iteration, opening the possibility of implementation on portable or implantable devices. Additionally, our results suggesting that the activities of individual movements can be linearly combined into the activities of more complex movements hint that decoders may not need to be trained on the full suite of behaviors that they will be used to predict. Instead, training decoders on orthogonal behaviors that span the full behavioral space (such as a center-out task), with representative neural activity, may be all that is required. This may cut the 5-10 minutes of decoder training time drastically, potentially streamlining the daily calibration of an outside-of-laboratory neural prosthesis. Generally, the results presented here suggest that naturalistic hand and finger neural prostheses with many DoFs may be close to clinical translation using simple linear decoder models and simple training procedures.

Though the results presented here show that linear decoding models can predict the movements of individuated fingers with high performance, the decoders were unable to achieve the level of precision and control of the able-bodied hand. This is despite the monkeys partially moving their fingers during brain-control mode, hypothetically to assist with controlling the virtual fingers. To bridge that gap, decoders may need to account for a nonlinear relationship between cortical activity and behavior. Several nonlinear neural networks have been tested for brain-machine interfaces (Hosman et al., 2019; Pandarinath et al., 2018), though few have transitioned to testing online. An early online recurrent neural network (Sussillo et al., 2012) showed that neural architectures have promise in online decoding. However, due to the heavy computational requirements of neural networks per online prediction update, they must be optimized and significantly compressed before being considered for out-of-laboratory, portable, and implantable

brain-machine interfaces. For example, the online recurrent neural network presented previously (Sussillo et al., 2012) required an estimated 144,000 to 225,000 multiplication operations per prediction update for their two-dimensional center-out arm reaches task, which is substantially more than the 505 multiplication operations required by a two-dimensional steady-state Kalman filter similar to what was used in this work (Malik et al., 2011). Therefore, it is valuable to characterize the limits of linear models in discriminating neural states with truly simultaneous movement of independent DoFs, as was presented with fingers in this work.

Importantly, the finger-related neural tuning models give insight into how primary motor cortex may represent the activity of a variety of related movements. Most neural units presented in this study did not show nonlinear specificity to one movement, but graded tuning to several related movements. In fact, most of the tuning curves fit the classical cosine tuning model demonstrated for arm reaches (Georgopoulos et al., 1982) and finger movements (Georgopoulos et al., 1999). Despite directional tuning with arm reaches fitting logically (the angles and magnitudes of movement can all be referenced to one limb in a radial task), we do not believe cosine tuning can represent the movements of multiple fingers. The physiological assumptions are broken by our task with two independent finger dimensions that have their own relatively independent muscles. Our task employs what are essentially two limbs (or fingers) traversing their own spaces with any given neuron capable of being tuned to both limbs, making musculoskeletal models seem like the more relevant explanation for the underlying neural activity (Todorov, 2000). While it is intuitive to conclude that the cosine tuning model expands to our two-finger task, we think our regression models (LR and LRO above) better explain how neurons can simultaneously encode movements of multiple independent limbs, though further investigation is required to investigate the role of sensory feedback in these consistent and linear relationships.

This raises a major question: how far can this weighted-average model be extended as we consider more DoFs? Considering the addition of the thumb, which is critical to hand use in primates and humans, the regressions may simply require fitting the new similar movements (i.e. thumb flexion in the case of predicting IF+MF+ thumb flexion). Further investigation will be required to determine if the regression model will hold in this case, as perhaps the addition of DoFs will require an exponentially increasing number of required component movements. For example, for IF+MF+thumb flexion, perhaps all of IF, MF, thumb flexion, IF+MF, IF+thumb flexion, and MF+thumb flexion may be required for accurate activity prediction. On the surface, these conclusions may appear to contrast with previous findings that cortical units strongly represent muscular synergies within the hand during awake use and during cortical microstimulation (Overduin et al., 2014, 2015; Saleh et al., 2010). We hypothesize that our findings complement this literature, where the shortcomings of our linear models (i.e. reduced performance when predicting combined finger movements with a model trained on individual movements) could possibly be better addressed with synergistic modeling. An alternative hypothesis is that the linear relationships we found between finger dimensions may arise from an alternative operating mode of cortex during non-prehensile movements, which may require more focused control to execute. In both instances, further work will be necessary to compare independent vs. synergistic cortical models related to finger movements (Kirsch et al., 2014; Mollazadeh et al., 2014).

In our work and those just mentioned, it is clear that finger-related cortical units often encode information about multiple digits. Several groups have investigated how neurons tuned to multiple behaviors respond to tasks requiring those behaviors for different extremities and across various species (Cross et al., 2020; Diedrichsen et al., 2013; Heming et al., 2019; Jorge et al., 2020; Stavisky et al., 2019, 2020; Willett et al., 2020). One way to explain this is that the neural activity

underlying separate and combined behaviors can be explained by multiple orthogonal subspaces. Here, we have evidence to suggest that linear models can effectively combine at least two finger subspaces that are substantially related. Provided sufficient quantities of neurons representing those subspaces, we think compound movements may continue to be well represented by linear combinations of their component subspaces, up to the full dimensionality of the hand and beyond to the entire motor system.

4.5 Acknowledgements

Chapter 5 Restoring Continuous Hand Function with Brain-Controlled Functional Electrical Stimulation

A subset of these results were presented as a virtual poster at the Society for Neuroscience 50th Annual Meeting.

5.1 Introduction

Neural prostheses have the potential to restore function and independence to people with neurological disorders and injuries. By interfacing with the nervous system directly, they can extract a user's intention information from the native controlling circuits or reanimate paralyzed limbs. Particularly in tetraplegia, it has been found that restoration of hand and arm function is of greatest importance (Anderson, 2004). Furthermore, for the purposes of upper extremity restoration, people with paralysis would prefer usage of their natural arm and hand over external prostheses (Blabe et al., 2015).

Functional electrical stimulation (FES) has shown promise in restoring function to paralyzed arms and hands for reaching and grasping (Kilgore et al., 2006, 2008; Knutson et al., 2013; Memberg et al., 2014). By delivering electrical current to partially or completely non-functional musculature, commercial devices like the Freehand System (Taylor et al., 2002b) have already been used with people with upper extremity paralysis to restore movement of native arms and hands. More recently, systems like the Networked Neuroprosthesis (NNP) have shown promise in primarily research-focused studies, having been implanted in five people with spinal cord injury to date (Makowski et al., 2021; Smith et al., 2005). In cases of upper extremity

paralysis, FES systems are often controlled by residual functional musculature to cycle between and activate grips. While successfully restoring functional grasps to paralyzed hands, there may be alternative controllers that feel more natural to use.

Meanwhile, brain-machine interfaces (BMIs) have allowed people with paralysis to control a variety of end effectors, including computers and robotic hands (Ajiboye et al., 2017; Hochberg et al., 2012; Pandarinath et al., 2017; Wodlinger et al., 2015). Since BMIs extract control signals from the natural planning and execution circuits in the brain, they have the potential to provide a more intuitive control source for FES. Multi-dimensional control of robotic arms via BMIs has enabled people with paralysis to independently manipulate objects, self-feed, and drink (Hochberg et al., 2012; Wodlinger et al., 2015). Recently, BMIs have predicted complicated hand and arm movements involved in handwriting (Willett et al., 2021), suggesting the upper bound on control capabilities continues to expand. Specifically with regards to the hand, BMIs can classify independent finger movements in people with paralysis (Bouton et al., 2016; Jorge et al., 2020). In monkeys, we have recently shown that BMIs can predict the simultaneous and independent movements of two finger groups continuously (Nason et al., 2021b), though it remains to be seen if the techniques translate to a paralyzed subject. Despite these impressive accomplishments in people, very few have translated to use with functional electrical stimulation in humans or monkeys.

Those that have translated BMIs to use with FES have successfully shown restoration of function in paralysis with monkeys (Capogrosso et al., 2016; Ethier et al., 2012; Moritz et al., 2008) and people (Ajiboye et al., 2017; Bouton et al., 2016). However, in most FES-related upper extremity studies, including those mentioned that employ brain-controlled functional electrical stimulation (BCFES), functional restoration comes in the form of discrete grasps. While important

for completing common activities of daily living, there is insufficient understanding of how precisely FES can continuously control movements of the hand.

In this work, we demonstrate a brain-controlled functional electrical stimulation system that controls the continuous one-dimensional movements of prehensile fingers in nonhuman primates (NHP). In a state of temporary wrist and hand paralysis, the monkey was able to use the BCFES system to acquire virtual finger targets at a rate substantially higher than he could with his paralyzed hand. In an anesthetized state, we used target information to control FES directly and found that hysteresis, failure of muscular recruitment, and feedback latency substantially reduce target acquisition speed in a continuous finger task. Finally, we demonstrate that the simultaneous and independent movements of multiple finger groups can be controlled virtually by a brain-machine interface even when the monkey's hand is paralyzed. Despite the performance reduction that was likely a result of the absence of sensory feedback, recalibrating the brain-machine interface using recalibrated feedback intention-training (ReFIT) restored performance to levels achieved by the brain-machine interface prior to paralysis.

5.2 Methods

All procedures were approved by the University of Michigan Institutional Animal Care and Use Committee.

5.2.1 Implants

We implanted one adult male rhesus macaque (Monkey N age 9 at the time of data collection) with Utah microelectrode arrays (Blackrock Microsystems, Salt Lake City, UT, USA) in the hand area of primary motor cortex, as described previously (Irwin et al., 2017; Nason et al., 2021b; Vaskov et al., 2018). Pictures of the implant can be found in Figure C.1. All motor cortex

arrays and the sensory cortex array were used in this study. Monkey N was between 867 and 1,071 days post cortical implant for all data collected in this study.

In a separate surgery, we implanted Monkey N with 86cm chronic bipolar intramuscular electromyography recording electrodes (similar to PermaLoc^{TM} electrodes, Synapse Biomedical, Inc., Oberlin, OH, USA). These electrodes were limited to 5μC of charge per pulse. Electrodes were implanted as described previously (Nason et al., 2021b). Briefly, a single radial-volar incision was used to access flexor muscles, and electrodes were placed in flexor digitorum profundus-index (1x near the nerve entry point, 1x distal near the wrist), flexor digitorum profundus-MRS (1x), flexor pollicis longus (1x, unused in this study), flexor carpi radialis (1x, unused in this study), and flexor carpi ulnaris (1x, unused in this study). Electrodes were secured intramuscularly using non-absorbable monofilament suture. After closing the radial-volar incision, a single dorsal-ulnar incision was used to access the extensors. Electrodes were placed in extensor digitorum communis (1x), extensor indicis proprius (1x, unused in this study), extensor carpi radialis brevis (1x, unused in this study), and extensor pollicis longus (1x, unused in this study). Electrodes were tunneled proximally to an interscapular exit site and connected to the standard PermaLoc^{TM} connector. All incisions were closed in a layered fashion using absorbable sutures. Following implantation, the monkey persistently wore a Primate jacket (Lomir Biomedical, Inc., Malone, NY, USA). Monkey N was between 195 and 399 days post FES electrode implant for all data collected in this study.

5.2.2 Feature Extraction

All processing was done in MATLAB versions 2012b or 2019b (Mathworks, Natick, MA, USA), except where noted.

Motor cortex spiking band power was acquired in real-time during the experiments (see the subsequent section for a description of data flow). We configured the Cerebus neural signal

processor (Blackrock Microsystems) to band-pass filter the raw signals to 300-1,000Hz using the Digital Filter Editor feature included in the Central Software Suite version 6.5.4 (Blackrock Microsystems), then sampled at 2kSps for SBP. The continuous data was streamed to a computer running xPC Target version 2012b (Mathworks), which took the magnitude of the incoming data, summed all magnitudes acquired in each 1ms iteration, and stored the 1ms sums as well as the quantity of samples received each 1ms synchronized with all other real-time experimental information. This allowed offline and online binning of the neural activity to create larger bin sizes, such as the 32ms used in this work, with 1ms precision. We masked channels for closed-loop decoding to those that were not clearly disconnected and had contained morphological spikes during the experiment or at some time in the past, as SBP could possibly extract firing rates of low signal-to-noise ratio units remaining represented on such channels (Nason et al., 2020).

Sensory cortex SBP was recorded to disk for later offline synchronization. Monkey N's sensory cortex pedestal was connected to a CerePlex Direct (Blackrock Microsystems) via a CerePlex E (Blackrock Microsystems), which band-pass filtered the incoming signals to 300-1,000Hz using the Digital Filter Editor feature, then sampled at 2kSps for SBP. To synchronize the sensory SBP activity after the experiment, we used the Sync Pulse functionality included in Central. The unique Sync Pulses were recorded by both the Cerebus and the CerePlex direct, enabling synchronization of the recordings from both systems offline. Then, sensory SBP was absolute valued and accumulated in equivalent windows as the motor SBP used for comparison using MATLAB R2019b.

5.2.3 Experimental Setup

The experimental apparatus used for these experiments is similar to what was described previously (Irwin et al., 2017; Nason et al., 2021b; Vaskov et al., 2018). Briefly, the monkeys'

Utah arrays were connected to the patient cable (Blackrock Microsystems) and 300-1,000Hz filtered 2kSps neural activity was extracted by the Cerebus and CerePlex Direct for storage. The 2kSps features from the Cerebus were streamed to the xPC Target computer in real-time via a User Datagram Protocol packet structure. The xPC target computer coordinated several components of the experiments. It binned SBP in customizable bin sizes, coordinated target presentation, acquired measured finger group positions from one flex sensor per group (FS-L-0073-103-ST, Spectra Symbol, Salt Lake City, UT, USA), and transmitted finger positions along with target locations to an additional computer simulating movements of a virtual monkey hand (MusculoSkeletal Modeling Software; Davoodi et al., 2007). Task parameters, states, and neural features were stored in real-time for later offline analysis.

One additional functionality was implemented in the xPC Target computer to facilitate real-time FES control. We used an RS-232 interface to the Networked Neuroprosthesis Access Point that comprised of an MSP-EXP430F5529LP evaluation board (Texas Instruments Inc., Dallas, TX, USA), an Evaluation Module Adapter Board (Texas Instruments), a CC1101EMK433 evaluation kit (Texas Instruments), and a MAX3222E RS-232 level shifter (Maxim Integrated, San Jose, CA, USA). Stimulation commands for one pattern were transmitted at 115,200 baud, which were sent wirelessly to the NNP power module to be configured into pulses delivered (see Functional Electrical Stimulation and Patterns sections below).

5.2.4 Behavioral Task

We trained Monkey N to acquire virtual targets with virtual fingers by moving his physical fingers in a one- or two-finger task, similar to what we have previously published (Irwin et al., 2017; Nason et al., 2021b). During all sessions, Monkey N sat in a shielded chamber with his arms fixed at his side flexed at 90 degrees at the elbow, resting on a table. The monkey had his left hand

placed in the manipulandum described previously (Nason et al., 2021b). The monkey sat in front of a computer monitor displaying the virtual hand model and targets described previously. The monkey was trained to either move his four fingers together with a target presented for just the index finger, or his index finger separately from his MRS fingers, where one target was presented for each finger group simultaneously for two total targets.

Each trial began with one spherical target per finger degree-of-freedom appearing along the one-dimensional movement arc. Each target occupied 15% of the full arc of motion of the virtual fingers and were presented in a center-out pattern. In the center-out pattern, every other target was presented at rest (halfway between full flexion and full extension or 50% as illustrated in figures), and the non-rest targets were randomly chosen between 20%, 30%, or 40% flexion or extension from rest. For a successful trial, the monkey was required to move the virtual fingers into their respective targets and remain there for 750ms continuously in able-bodied manipulandum control or 500ms continuously in brain- and BCFES control modes. If the monkey could not acquire and hold the target within 10s, the trial was deemed unsuccessful and the target was placed at center repeatedly until successfully acquired. Upon successful target acquisition, the monkey received a juice reward, which was modulated to maintain Monkey N's motivation levels.

5.2.5 Nerve Block Procedure

Temporary paralysis of the hand was achieved by delivering a solution of lidocaine (2%) and epinephrine (1:100,000) intramuscularly at three peripheral nerves (radial, median, ulnar), in the upper arm just proximal to the elbow. The solution was delivered so that it completely surrounded each nerve. Figure C.2 presents some example snapshots of the median, radial, and ulnar nerves from Monkey N under ultrasound. The lidocaine/epinephrine solution was either purchased premixed (NDC 0409-3182-11) or compounded using stock solutions of lidocaine (2%)

and epinephrine (1mg/mL). Delivery was done under ultrasound (Lumify L12-4 broadband linear array transducer, Philips Healthcare, Best, NL, paired with Samsung Galaxy Tab S6, Samsung Electronics, Suwon, KOR) using a 20 gauge echogenic needle (B.Braun #33642, B.Braun, Melsungen, DE). The NHP was placed in a restraint chair and the arm was manually restrained. In some experiments, prior to delivering the block, pure lidocaine (2%) was injected subcutaneously to the target injection sites to ease with comfort during the injection, and approximately 6-7mg/kg of lidocaine in the lidocaine/epinephrine solution (less than that was ineffective) was used cumulatively for each blocking procedure. Block onset typically occurred after 20-30 minutes and lasted up to 2 hours. The monkey underwent at least two days between blocking procedures.

To guarantee the nerve block was still active at any time during an experiment, we would occasionally disable the BCFES system and the decoder, allowing the monkey to complete trials with his native anatomy as capable.

5.2.6 Propofol Anesthesia

To investigate the capabilities of functional electrical stimulation alone without voluntary activation from the monkey, we lightly anesthetized Monkey N with propofol.

After placing the monkey in the primate chair, comfortably restraining his head, and comfortably restraining his arms, we placed an IV catheter in the cephalic vein just distal to the monkey's right elbow, contralateral to the arm being used for FES experiments. We then flushed the catheter with sterile saline to ensure patency. We delivered a 2.5-3.0mg/kg bolus with additive 0.2mg/kg boluses of propofol until the primate was visibly unconscious to induce a light plane of anesthesia. Following induction, light anesthesia was maintained with a constant rate infusion of propofol at 7.5mg/kg/hr and supplemental boluses of 0.2mg/kg were given as needed to maintain the desired plane of anesthesia.

Oxygen supplementation was provided with a nasal cannula. Heart rate, respiratory rate, body temperature, and blood oxygen levels were monitored throughout the experiment with a pulse oximeter and pediatric ear thermometer. Blood oxygen levels never dropped below 90% and were always greater than or equal to 98% after providing supplemental oxygen.

Anesthetic depth was measured with jaw tension, respiratory rate, and heart rate. After a stable and light plane of anesthesia was reached, we proceeded with functional electrical stimulation. Anesthetic events/procedures occurred no more than once per week and for no longer than two hours with most lasting one hour.

5.2.7 Functional Electrical Stimulation

We delivered stimulation using the Networked Neuroprosthesis Evaluation System (Makowski et al., 2021). The evaluation system contains the exact circuitry used as implants in people (NCT02329652) but in a form factor more conducive of experimentation and debugging. Our evaluation system consisted of one power module, including three 1,000mAh batteries to power the system and deliver stimulation (PRT-13813, SparkFun Electronics, Niwot, CO, USA), and three 4-channel pulse generator modules. The system could deliver 12 total monopolar stimulation channels. All stimulation was current-controlled and delivered at 10mA with a 32ms inter-pulse interval. Pulses were charge-balanced biphasic with a square cathodic pulse delivered first and the subsequent anodic pulse exponentially balancing charge. Pulse delivery was staggered within each pulse generator module to avoid too large of a current draw than could be provided by the power module. The pulse for channel 1 of a given pulse generator module was delivered 20ms following the reception of the stimulation command (the time required to prepare the pulse based on the command), and each subsequent channel delivered its pulse 1ms after the prior channel. All pulse generator modules received the stimulation command simultaneously.

To adapt our bipolar leads to function with monopolar stimulation, we only used one electrode of each pair to deliver current. The electrode that was implanted further inside the muscle was used for monopolar stimulation. The second electrode of each bipolar pair that was further outside the muscle was tied with all other second electrodes connected to one pulse generator module. Each group of four electrodes that were further outside the muscle were connected to that pulse generator module's current return.

Prior to usage of the BCFES or target-controlled FES systems, the flex sensors were recalibrated to the range of motion that could be achieved independently by FES. During some BCFES and target-controlled FES experiments, small wooden stints were taped to the monkey's fingers following the nerve block. FES would occasionally flex the distal and proximal interphalangeal joints within the manipulandum in a way that could not be measured by the manipulandum and prevent extension of the manipulandum. Without electrodes implanted in all hand muscles, FES could not re-extend the fingers independently. The stints helped with keeping the fingers straight and flexing primarily at the metacarpophalangeal joint. Additionally, during BCFES experiments, we would occasionally stop sets of BCFES trials early in the event of an incompatible finger state (as previously described) or if the decoder consistently predicted a maximum flexion or extension state for at least one trial continuously. This was done in an attempt to avoid the muscles fatiguing too soon into the experiment. In such instances, the decoder was reinitialized and BCFES was re-enabled.

5.2.8 Patterns

To keep communication bandwidth low, stimulation was delivered by coordinated patterns (Ajiboye et al., 2017; Kilgore et al., 1989). In a patterned stimulation paradigm, electrodes belonging to a pattern have their pulse parameters governed by a one-dimensional variable, given

the range 0-100% in this manuscript. In this way, one command can govern stimulation parameters on any number of electrodes.

In our one-dimensional implementation, we used one pattern to represent stimulation parameters for the electrodes in the FDPi, FDPmrs, and EDC muscles. As implemented on the NNP, patterned stimulation allows one to control pulse width and current amplitude at each command value. Command values were computed by the xPC Target computer from the finger positions predicted by the RKF or the target controller during BCFES and target-controlled FES experiments, respectively. The patterns we implemented did not modulate stimulation amplitude with command value (constant 10mA) but did modulate pulse width between 0 and 255µs. Our pattern delivered maximum stimulation to EDC at 0% command and maximum stimulation to FDPi and FDPmrs at 100% command. In the region around 50% command value, all of EDC, FDPi, and FDPmrs received stimulation at approximately 30% of the maximum pulse width according to the pattern. Generating co-contracting stimulation when not at the limits of range of motion can more consistently create the desired position at that particular command value and combat hysteresis. Figure C.3 illustrates an example pattern used during a BCFES experiment.

5.2.9 Determining Stimulation Command Value

5.2.9.1 Closed-Loop ReFIT Kalman Filter

For brain-controlled functional electrical stimulation, we used a ReFIT Kalman filter to predict Monkey N's intended one-dimensional finger movements (Vaskov et al., 2018). To train the ReFIT Kalman filter, Monkey N first performed at least 350 trials in manipulandum control mode with a 750ms target hold time. Then we trained a position/velocity Kalman filter (Irwin et al., 2017), which the monkey used in a closed-loop fashion for at least 250 trials with a 500ms hold time. We trained a ReFIT Kalman filter using those closed-loop standard Kalman filter trials

and tested it for at least 100 trials prior to the experiment's nerve block. Following the nerve block, the same ReFIT Kalman filter was used in RKF trials and BCFES trials, where the RKF output was mapped directly to stimulation command value. During brain-control trials, the positions of the fingers as displayed on the screen were integrated from the RKF's predicted velocities. Manual velocity biases were added as needed to the RKF's predicted velocities following the nerve block to assist with any changes in noise level resulting from disconnecting and reconnecting the recording hardware.

We implemented a two-finger position/velocity ReFIT Kalman filter in a similar fashion to what we have previously published (Nason et al., 2021b). The two-finger RKF was never used to control stimulation, as it was limited to 1 degree-of-freedom control. Training the two-finger RKF followed the same procedure as the one-finger RKF but using two-finger manipulandum control and standard Kalman filter control trials. Following usage of the standard Kalman filter, the ReFIT Kalman filter was trained by rotating the predicted velocity vector towards the single two-dimensional target generated by combining the one-dimensional target for each finger group, similar to the original ReFIT method (Gilja et al., 2012).

5.2.9.2 Target-Controlled FES

We used target information to control stimulation to better understand the capabilities and challenges when performing continuous functional electrical stimulation. The state machine diagram in Figure 5.2a illustrates this controller. At the beginning of each trial, the stimulation command value is set to the target's center position. Then, a stimulation update was delivered prior to every pulse (32ms between updates) depending on the target's relative location to the current finger position. If the target required more finger flexion, stimulation command value was increased by approximately 1%. If the target required more finger extension, stimulation command

137

value was decreased by approximately 1%. Updates were sent continuously every 32ms in this fashion until the finger moved to within the middle 75% of the target. Once in the target, stimulation command value was held constant until the finger inadvertently moved out of the middle 75% of the target, the target was successfully acquired, or the trial timeout was reached.

5.2.10 Performance Metrics

5.2.10.1 Closed-Loop Performance Metrics

We estimated closed-loop performance with success rate and acquisition time, occasionally split into measures of time to target and orbiting time. Success rate was calculated as the total number of targets acquired successfully divided by the total number of targets presented. Acquisition time was computed as the total amount of time from the beginning of the trial to the time the target was successfully acquired and held, less the hold time. Failed trials were given acquisition times equal to the trial timeout less the hold time. Time to target was computed as the time from the beginning of the trial to the first instance the target was touched by the finger. Trials in which the finger never touched its target were given acquisition times and times to target equal to the trial timeout less the hold time. Orbiting time was computed as the time from first touching the target to the time the target was successfully acquired and held, less the hold time. The sum of a trial's time to target and orbiting time equals its acquisition time.

In all metrics, the following conditions were excluded from analysis to avoid biasing the results. Successful trials immediately following a failure were excluded to avoid situations in which the monkey would wait at the central position to acquire the central target once it returned following a failed outer target. Trials immediately following a change in control method were also excluded, as the state of the controller could not be determined prior to usage. For example, when switching from RKF control to manipulandum control, the displayed finger position could jump

from outside of the target straight to the subsequent target, resulting in an unrealistically small acquisition time.

For the orbiting time metric, trials in which the finger never touched its target were excluded, as orbiting time could not be computed.

5.2.10.2 Open-Loop Performance Metrics

We estimated open-loop prediction performance using the intra-class correlation coefficient. As opposed to Pearson's correlation, intra-class correlation is a better measure of agreement between two conditions. Intra-class correlation is computed via the following equation (5.3):

$$\rho = \frac{\sigma_b^2 - \sigma_w^2}{\sigma_b^2 + \sigma_w^2} \tag{5.1}$$

where σ_b^2 is the variance between conditions and σ_w^2 being the variance within conditions. Confidence intervals were computed via the following formulas and equation (5.2) (Donner, 1986):

$$V(\rho) = \frac{(1-\rho)^2}{2}\left(\frac{(1+\rho)^2}{n} + \frac{(1-\rho)*(1+3\rho)+4\rho^2}{n-1}\right)$$

$$\theta = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

$$SE = \frac{\sqrt{V(\rho)}}{(1+\rho)(1-\rho)}$$

$$(\theta_L, \theta_U) = \theta \pm Z_{1-\frac{\alpha}{2}}SE$$

$$(\rho_L, \rho_U) = \frac{e^{2(\theta_L,\theta_U)} - 1}{e^{2(\theta_L,\theta_U)} + 1} \tag{5.2}$$

where $Z_{1-\frac{\alpha}{2}}$ is the z-score at the confidence level $1 - \frac{\alpha}{2}$. Significant differences between intra-class correlations were determined via the Fisher's transformation of θ for each intra-class correlation via equation (5.3):

$$z(\theta_1, \theta_2) = \frac{\theta_1 - \theta_2}{\sqrt{SE_1^2 + SE_2^2}} \qquad (5.3)$$

The p-value for z was calculated from a normal distribution.

5.3 Results

5.3.1 Restoration of Continuous Hand Function with Brain-Controlled Functional Electrical Stimulation in Paralysis

We first sought to estimate a nonhuman primate's capability of continuously controlling his own paralyzed hand using a brain-controlled functional electrical stimulation neuroprosthesis. To do so, we trained a nonhuman primate to perform a finger task in which they controlled a virtual hand model using proportional movements of their physical fingers. A virtual target encompassing 15% of the range of motion was presented for movements of all four prehensile fingers together, and the monkey was required to move his physical fingers such that the virtual fingers entered their targets and held them continuously for 750ms. Then, we trained and tested a ReFIT Kalman filter (RKF) to predict the primate's intended finger movements from spiking band power recorded from Utah microelectrode arrays implanted in the hand area of primary motor cortex in real-time (Vaskov et al., 2018). We used spiking band power as we have previously shown it can extract firing patterns at lower signal-to-noise ratios, is more specific to the spiking rates of single units,
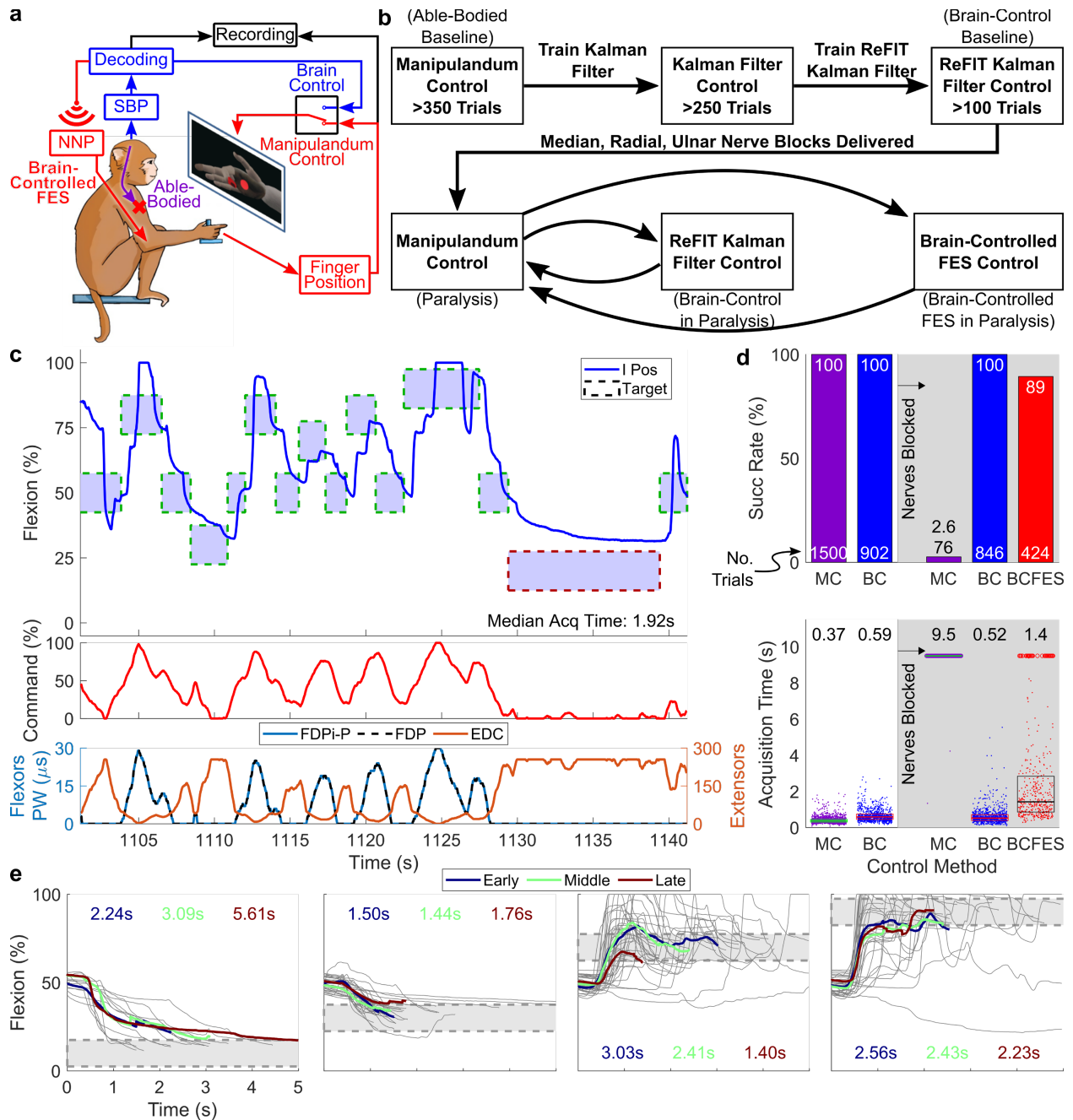
Figure 5.1 Brain-controlled functional electrical stimulation restores continuous hand function following paralysis. (**a**) Diagram portraying the experimental control options. Purple represents able-bodied control using the manipulandum. Blue represents brain control in which decoded intentions control the virtual hand. Red represents brain-controlled functional electrical stimulation (BCFES) following nerve block, in which the decoder's predictions control FES to the monkey's arm electrodes. (**b**) Flow diagram of experimental procedures. (**c**) Example control traces under BCFES with stimulation command and stimulation pulse widths displayed in parallel. Green-border targets were successfully acquired, and red-border targets failed to trial timeout. I means index position, for which the target was presented despite the whole hand being controlled by stimulation. (**d**) Target acquisition statistics for each method of control. Boxes represent the 25th and 75th percentiles, with the midway line representing the median. Dots represent successful trials, and circles represent trials failed to timeout. (**e**) All BCFES trial traces split by closest and furthest outer targets. The blue, green, and red colored traces represent mean traces during early, middle, and late trial epochs, respectively. Median acquisition times for each target are displayed and colored according to the epoch they represent.

and achieves as good or better prediction performance than threshold crossing rates (Nason et al.,

2020).

To demonstrate that we could control finger movements continuously with FES, we implanted chronic bipolar stimulating electrodes into finger-related muscles of the forearm (see Methods for a list of electrode locations and quantities). Then, we delivered a lidocaine with epinephrine solution to surround the median, radial, and ulnar nerves just proximal to the elbow to temporarily block voluntary muscle contractions of the hand. We delivered electrical stimulation using the Networked Neuroprosthesis Evaluation System (NNP; Makowski et al., 2021), which is a benchtop version of a human implantable device, controlling the aperture of the hand's fingers with stimulation patterns (Kilgore et al., 1989). Then, we tested the monkey's capability of controlling the virtual hand using his physical hand movements after the nerve block (purple route in Figure 5.1a), using the RKF's predictions to control the virtual hand directly (blue route in Figure 5.1a), and using the RKF to control the stimulation delivered by the NNP with the virtual hand controlled directly by his physical hand movements (red route in Figure 5.1a).

Figure 5.1d shows single-trial statistics from all four experiment days under all control methods: manipulandum and RKF control prior to nerve block and manipulandum, RKF, and BCFES control following the nerve block. Using the BCFES system, monkey N could acquire 89% of the presented targets (424 trials) in a median acquisition time of 1.4s. BCFES substantially improved control capabilities over monkey N's native hand movements following the nerve block, which had a 2.6% success rate (76 trials) and a median acquisition time of 9.5s, which is the trial timeout. Following the nerve block, monkey N's usage of the ReFIT Kalman filter to control the animated fingers did not change substantially, maintaining a 100% success rate with a median acquisition time of 0.52s (846 trials).

Figure 5.1c shows a representative BCFES control example with decoded commands and delivered stimulus pulse widths included. Most failures of the BCFES were towards further extension targets, as monkey N's extensors fatigued quickly during BCFES usage. This is clearly showcased in Figure 5.1e, which displays individual successful trial traces to the closest and furthest outer targets with early, middle, and late trials averaged into three highlighted traces of median length. The median acquisition times to extension targets during late epochs were generally longer than the durations during early and middle epochs (1.4s and 0.98s higher median acquisition times than early and middle epochs, respectively). Recruitment of Monkey N's EDC muscle was challenging even at maximum stimulation, which resulted in rapid fatiguing and his reliance on gradual recruitment to hit extension targets.

5.3.2 High-Speed Time to Target with Target-Controlled FES

We next sought to better understand how well our FES solution worked as a standalone prosthesis without BMI control. To do so, we anesthetized Monkeys N and W with propofol and used target information to control the stimulation command. Figure 5.2a illustrates the simple proportional controller, where at the beginning of a trial, the stimulation command is set to the target position. Then, to provide any necessary minor corrections ideally without overshooting the target, the stimulation command was increased or decreased by approximately 1% every iteration depending on the direction of the target from the current position of the finger.

Figure 5.2b illustrates acquisition time statistics for all trials with target-controlled stimulation, with colors indicating the direction of movement in the center-out task. Overall, times to target were quick (median 0.8s). There were no substantial differences between acquisition times to flexion or central targets, though extension targets took longer to acquire likely due to rapid muscle fatigue.
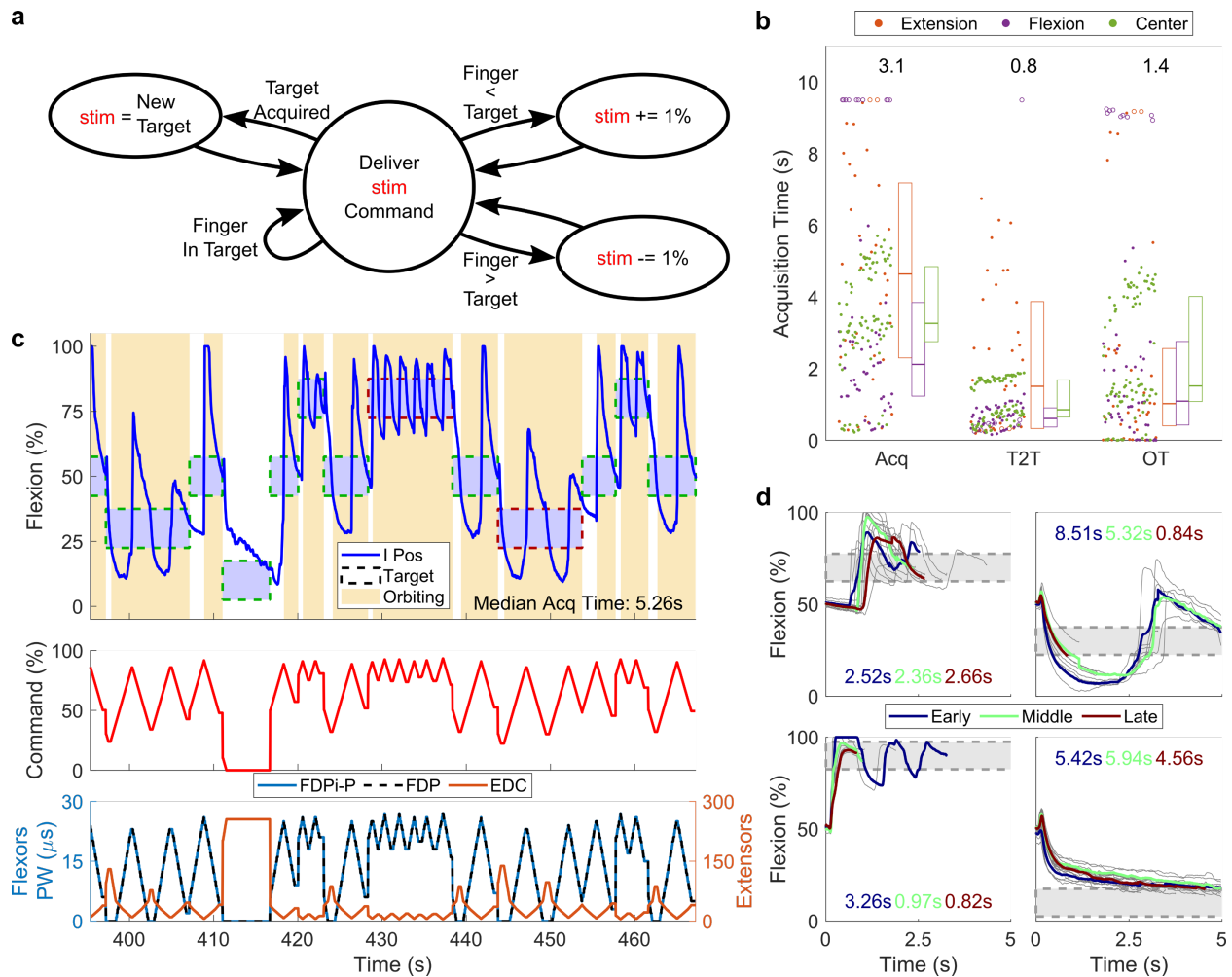
Figure 5.2 Target-controlled functional electrical stimulation performance. (**a**) State machine diagram for target-controlled stimulation. (**b**) Target acquisition statistics for target-controlled stimulation. T2T means the time to target, or the amount of time to first touch the target. OT means the orbiting time, or the amount of time between first touching the target and successfully acquiring it, less the required hold time. Acq means the total acquisition time, or the sum of T2T and OT for each trial. Boxes to the right represent the 25th and 75th percentiles, with the midway line representing the median for each type of target. Dots represent successful trials, and circles represent trials failed to timeout. (**c**) Example control traces under target controlled stimulation with stimulation command and delivered stimulation pulse widths displayed in parallel. Green-border targets were successfully acquired, and red-border targets failed to trial timeout. Yellow regions represent periods of orbiting the target. I means index position, for which the target was presented despite the whole hand being controlled by stimulation. (**d**) All target-controlled FES trials split by closest and furthest outer targets. The blue, green, and red colored traces represent mean traces during early, middle, and late trial epochs, respectively. Median acquisition times for each target are displayed and colored according to the epoch they represent.

Interestingly, orbiting times were drastically longer than the times to target (median 1.4s).

For movements back towards a central target, there are two regions of acquisition times: one for

extensions back to center and one for flexions back to center. For flexions out of center, the times

to target were generally quick for Monkey N (median 0.61s), as his FDP and FDPi muscles were

easily recruited by stimulation. However, there was a wide spread of orbiting times due to hysteresis in muscular activation from stimulation (0.43s to 2.76s for the 25th to 75th percentiles, respectively). This is clearly showcased in Figure 5.2c, which displays the movement traces for representative trials under target-controlled stimulation. The yellow regions indicate times where the finger was orbiting the target, calculated as the time period between leaving a target after first hitting it to the end of the trial. While extensions, flexions, and central targets all show substantial amounts of orbiting, the frequency of target crossings when attempting to acquire flexion targets was generally higher than for other targets. Orbiting resulted in 9 out of the 10 failures when attempting to acquire flexion targets. In all cases, a well-tuned proportional-integral-derivative controller could substantially shorten many of the orbiting periods.

For extensions out of center, times to target were much more variable due to the poor recruitment of the EDC muscle (1.51s median, 0.33s to 3.87s for the 25th to 75th percentiles, respectively). Most targets more extended than the central target could only be acquired with the gradual recruitment of the muscle under sustained stimulation. Extension targets also generally had long orbiting times (1.03s median, 0.40s to 2.57s for the 25th to 75th percentiles, respectively), where orbiting resulted in both failures to acquire extension targets. Due to the poor recruitment of the EDC muscle even at maximum stimulation, the frequency of orbiting was generally slower than the orbiting frequency for acquiring flexion targets. Figure 5.2d illustrates this slower orbiting frequency, and only trials to the nearest extension showed any orbiting.

5.3.3 High-Performance Two-Finger Brain-Control in Paralysis

To close the gap towards a multi-finger brain-controlled functional electrical stimulation neuroprosthesis, we next wanted to investigate how well our brain-machine interface systems could predict multi-finger movements in a paretic scenario. Similar to our one-dimensional finger

task discussed thus far, we trained a two-dimensional RKF to predict the simultaneous and independent movements of two finger groups (the index finger separate from the middle, ring, and small fingers as a group) prior to blocking the nerves (Nason et al., 2021b). Then, we used the RKF to predict the two-dimensional finger movements in real-time following the nerve block on two separate days.

Figure 5.3a illustrates the results of these experiments, where prior to the nerve block, the monkey's capabilities of completing the task with his physical hand and the RKF were high (100% success rate for both and 0.4s and 0.75s median acquisition times, respectively). Following the nerve block, the monkey's acquisition rate when using his physical hand dropped drastically (9.3% success rate due to chance acquisition, with 9.5s median acquisition time), as expected due to the inability to physically move his fingers.

Then, we transitioned back to investigating usage of the BMI following the nerve block. When testing the RKF, we surprisingly found a substantial drop in performance (89% success rate with 1.5s median acquisition time) despite no direct cortical interventions. In an attempt to restore performance, we performed the ReFIT training procedure an additional time using the closed-loop RKF trials Monkey N performed following the nerve block. We termed this second-stage RKF the Re-RKF, indicated in green in Figure 5.3. After recalibrating, closed-loop control with the Re-RKF returned to near the level of performance achieved by the RKF prior to the nerve block (Figure 5.3a, 100% vs. 99% success rate and 0.75s vs. 0.81s between RKF prior to the nerve block and Re-RKF, respectively).

The surprising drop in performance following the nerve block motivated us to investigate what may have been its cause. When we blocked median, radial, and ulnar nerves, not only were the descending motor commands prevented from reaching their muscular targets, but perhaps
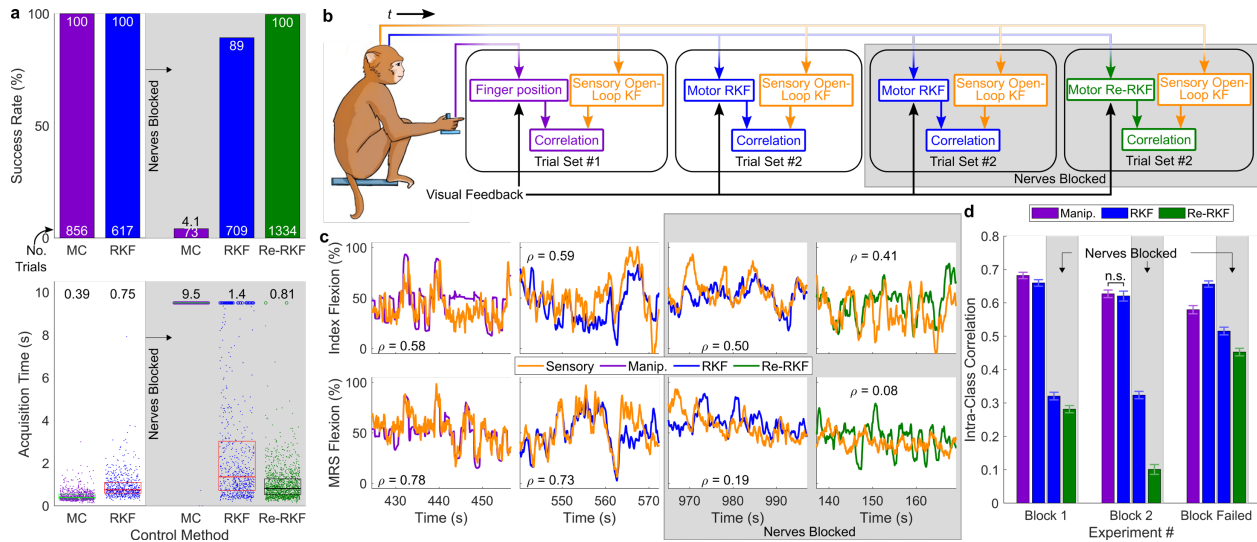
146

Figure 5.3 RKF can control multi-finger movements following nerve block despite the absence of sensory feedback. (**a**) Success rates and acquisition times for manipulandum, RKF, and recalibrated RKF (Re-RKF) two-finger control before and after nerve blocks. (**b**) Description of analysis. SBP was recorded from Monkey N's sensory implant synchronously with manipulandum measurements and the SBP from Monkey N's motor implant. For each comparison, we trained an open-loop Kalman filter to predict the virtual control outputs (manipulandum, RKF, or Re-RKF) from sensory SBP recorded at the same time. The predictions from the sensory Kalman filter were intra-class correlated with the control outputs during each set of trials. (**c**) Example cross-validated open-loop position predictions from the sensory Kalman filter (orange) overlaid on the control outputs. The horizontal time axis represents the seconds into each usage instance of each closed-loop control method. Index and MRS were controlled simultaneously by the monkey with traces split here for clarity. The included correlations represent each snippet's intra-class correlation. (**d**) Intra-class correlation coefficients between the sensory Kalman filter predictions and each control method, with positions combined across fingers and all collected data from each closed-loop control method. Error bars represent the upper and lower bounds with an alpha level of 0.001. Data represent experiments with nerve blocks on two separate days with one additional day in which the nerve block failed (87% success rate with manipulandum control). All comparisons within each day were significantly different ($p < 0.001$, one-sided two-sample z-test on the Fisher-transformed intra-class correlation coefficients), except those labelled with n.s.

proprioceptive and sensory feedback were also partially or completely blocked from transduction back to cortex. We hypothesized that the absence of this feedback information to motor cortex resulted in minute changes in cortical tuning in the task's environment, enough to increase difficulty of using the RKF. To investigate this hypothesis, we analyzed how well cortical spiking activity in sensory cortex (recorded with a Utah array implanted in Monkey N's sensory cortex but never used for BMI control) represented the two-dimensional finger task the monkey was performing. For each task condition, we trained an open-loop Kalman filter to use SBP recorded from sensory cortex to predict either the measured finger movements in manipulandum control

mode or the closed-loop movements predicted by the RKF or Re-RKF from motor cortex SBP during online control. This procedure is illustrated in Figure 5.3b.

Figure 5.3c and d show that SBP in sensory cortex can accurately predict multi-dimensional finger positions during manipulandum control mode (0.68 and 0.63 intra-class correlations across both testing days). During RKF control prior to the nerve block, sensory SBP also highly predicted the RKF's closed-loop predictions (0.66 and 0.62 intra-class correlation between sensory and motor predictions). In this case, the maintenance of high correlation coefficient suggests sensorimotor synchronicity during the task, as the monkey was free to move his physical hand during RKF use.

Following the nerve block, however, the capability of sensory SBP to predict the motor RKF predictions dropped substantially (0.32 and 0.37 intra-class correlation). The drop in intra-class correlation was similar to the closed-loop performance drop, which may suggest a loss of synchronicity between sensory and motor cortices. SBP prediction performance continued to drop further during Re-RKF use (green bar in Figure 5.3d, 0.28 and 0.06 intra-class correlation between sensory and Re-RKF predictions), despite the restoration of performance when using Re-RKF. Interestingly, in a third experiment day where a nerve block was attempted but failed to substantially block signal transduction (87% success rate with manipulandum control post-block), the sensory Kalman filter maintained the capability of predicting the online RKF and Re-RKF predictions from motor cortex (0.58, 0.66, 0.51, and 0.45 intra-class correlation to manipulandum control, RKF control pre-block, RKF control post-block, and Re-RKF control, respectively).

5.4 Discussion

Modern FES has primarily focused on restoring discrete grasps and grips but the extent to which FES can continuously control end effectors is not well understood. In this work, we have

demonstrated that an NHP can use a brain-controlled FES system to control his finger movements continuously despite temporary paralysis of his wrist and hand. The BCFES system allowed the monkey to exceed the performance of his native hand while paralyzed, though it could not achieve the levels of performance of the able-bodied hand prior to paralysis nor the BMI. By using information about the presented target to control stimulation without voluntary control from the monkey, we found that hysteresis, muscle recruitment failure, and controller latency provide systematic challenges in using a BCFES prosthesis. Fortunately, these challenges can be partially mitigated by the user via brain-control or with more advanced controllers such as a proportional-integral-derivative controller. Finally, we have shown that the simultaneous and independent movements of multiple finger groups can be continuously predicted with high performance when the wrist and hand are paralyzed despite the absence of proprioceptive and sensory feedback. Recalibrating a ReFIT Kalman filter restored performance of the two-finger decoder to similar levels as the original ReFIT Kalman filter used prior to paralysis.

Our prior work which demonstrated brain control of continuous, multi-finger movements had used unrestricted, able-bodied monkeys, so it remained unclear if the result would translate to paralysis. Here, following induction of temporary paralysis, performance of the RKF trained prior to paralysis dropped substantially but recovered by recalibrating with the ReFIT procedure an additional time. Importantly, it is the first demonstration that continuous finger individuation can be predicted with a BMI in paralysis despite the absence of proprioceptive and sensory feedback. There was no need to recalibrate the decoder for the one-dimensional task, which may suggest that sensory information plays a more substantial role in more complex finger behaviors.

Interestingly, our analysis of the sensorimotor synchronicity following the nerve block suggests that the loss of sensory feedback may have been the cause for the drop in performance.

When using sensory cortex activity to predict the closed-loop RKF predictions following induction of paralysis, we noticed a substantial drop in intra-class correlation compared to before paralysis. That intra-class correlation dropped further during usage of the Re-RKF, which paints a picture of how sensory and motor cortices might interact with a multi-finger brain-machine interface. We hypothesize that some of the motor SBP channels that the RKF found valuable contained some information transferred from sensory cortex. Without that sensory information following the nerve block, RKF lost some of the information it used to accurately predict multi-finger movements, resulting in the closed-loop RKF performance loss as well as the loss of intra-class correlation between sensory SBP and closed-loop RKF predictions. Since motor cortex did not have any sensory inputs during usage of the RKF following the nerve block, training the Re-RKF decoder used purely motor cortex activity, which we think is what allowed it to function with high performance in the absence of sensory feedback. Although this hypothetical explanation is intriguing, further experimental validation will be necessary to support it.

Implanting electrodes among the muscles directly gives more targeted muscle activation than surface FES. While the work presented here only targeted one-dimensional hand aperture, precision in the placement of electrodes could selectively activate flexors and extensors for individual digits, enabling native individual finger control for all digits as published by others (Jorge et al., 2020). Placing electrodes within the hand is challenging with NHPs, with a high risk of infection, but intrinsic muscles could restore more function and gripping capabilities to the hand. Of particular importance are the flexor pollicis brevis and adductor pollicis, which would restore thumb opposition. Even with the greatest electrode placement precision, stimulation bleeding into neighboring muscles adds a challenge to independent and continuous FES. Nonlinear stimulation patterns could assist here as well, where nonlinear combinations of stimulation across many

electrodes or patterns could generate unique contractions that could not be accomplished with a linear stimulation pattern, as was investigated here.

When investigating the capability of FES to continuously control finger movements without voluntary intervention from the monkey, we found substantial orbiting to reach flexion and extension targets that was likely an outcome of hysteresis and failure of muscle recruitment. Another systematic issue with FES is the muscle fatigue that restricts FES functionality over time. The statistics in Figure 5.2b show a noticeable but insignificant incline in the total acquisition times, times to target, and orbiting times as time progressed through the experiment. This suggests that muscle fatigue resulted in a general increase in acquisition times as cumulative stimulation time increased. Furthermore, latency in between a command and the behavioral response resulting from stimulation could cause severe oscillations about a target, such as those illustrated in Figure 5.2c. During BCFES usage, Figure C.4 shows that the optimal latency between activity in motor cortex and the behavioral outputs was substantially higher than in manipulandum and RKF control conditions. In this implementation, stimulus commands were updated every 32ms, which means that the wireless latency, the time to prepare each stimulation pulse, and the latency for the tissue to generate a behavioral response from the stimulation may delay feedback to the controller by multiple iterations. Particularly in the case of Monkey N's FDP and FDPi electrodes, the nonlinearity between muscle recruitment and stimulus magnitude amplifies this problem. Fortunately, learning to use the BCFES system as shown in Figure 5.1 enabled Monkey N to reduce these oscillations, which may be accomplished by a target-based controller with a well-tuned proportional-integral-derivative controller.

Using a target-based controller for BCFES may help improve performance overall. In this implementation, the RKF output limited to the [0, 1] range was bijectively mapped to the

stimulation command range [0, 100]%, but such a linear mapping may not represent the stimulation necessary to bring fingers to the commanded positions. FES hysteresis makes this particularly challenging. However, if the present positions of the fingers are known, a feedback controller could be added to the BCFES processing chain to nonlinearly alter stimulation based on the error between the desired (i.e. commanded by the decoder) and measured finger positions. To do so, the user would only need to wear a glove that measures finger positions and relays them to a feedback controller, such as a proportional-integral-derivative controller (Crago et al., 1980). Such an addition might improve kinematic and kinetic control capabilities without additional conscious effort from the user.

The BCFES results presented here expand upon the BCFES accomplishments of others (Ajiboye et al., 2017; Bouton et al., 2016; Ethier et al., 2012; Moritz et al., 2008) by beginning to investigate the realm of continuous control with FES, using similar technologies as what was previously used in clinical trials. By adding continuous control to grasp and force, a user may be able to perform more natural hand movements and use the device for a wider variety of tasks. Furthermore, combining these techniques with sensory feedback via intracortical microstimulation (Flesher et al., 2016, 2021) may enhance control capabilities. Although we have shown sensory feedback is not required for cortical control of dexterous finger movements, it may help with the accuracy of BCFES movements as it has been shown to improve motor control (Flesher et al., 2021). Including a hand-state measurement glove, as discussed in the previous paragraph, could acquire the necessary signals to deliver valuable proprioceptive and force-related feedback in addition to improving BCFES performance.

For the purposes of translating BCFES for functional restoration beyond laboratory use, the technologies presented in this manuscript all represent low-power solutions end-to-end. The

stimulation system used in this work is an evaluation system for the Networked Neuroprosthesis, a low-power, wireless, completely implantable stimulator presently undergoing a clinical trial (NCT02329652). The brain-machine interface used the 300-1,000Hz spiking band power, a low-power neural feature that is specific to single-unit activity despite a ten-fold reduction in processing requirements. We have already demonstrated SBP's efficacy on embedded and integrated systems (Bullard et al., 2019; Lim et al., 2020, 2021), and others have used it in similar brain-machine interface and BCFES applications (Ajiboye et al., 2017; Jarosiewicz et al., 2015; Willett et al., 2017). The pieces of a power-efficient, fully implantable brain-controlled functional electrical stimulation neuroprosthesis are in the literature, what remains is assembly and safety validation for full-time use in people with paralysis.

5.5 Acknowledgements

Chapter 6 Discussion

6.1 Low-Power Brain-Machine Interfaces

Brain-machine interfaces have restored substantial function to people with paralysis in the form of computer control (Jarosiewicz et al., 2015; Pandarinath et al., 2017; Willett et al., 2021), robotic arm control (Flesher et al., 2021; Wodlinger et al., 2015), and restoration of native movements (Ajiboye et al., 2017; Bouton et al., 2016). While impressive, these technologies have not undergone widespread clinical translation due to a number of remaining challenges, one being the large power consumption of the computers needed to execute the BMI computations. Traditional intracortical BMIs have remained tied to the usage of high-power, high-bandwidth neural features, like threshold crossing rate, to achieve the highest BMI performance standards presented. Deviations to lower-bandwidth features such as local field potentials (Flint et al., 2012a, 2013; Milekovic et al., 2018; Stavisky et al., 2015) or less spatially specific recording modalities like electrocorticography (Chao et al., 2010; Chestek et al., 2013; Flint et al., 2017; Hotson et al., 2016; Kubánek et al., 2009; Schalk et al., 2007) have resulted in substantial reductions in BMI performance.

In Chapter 2, we presented the 300-1,000Hz spiking band power as a replacement neural feature for BMIs that can extract spiking-level features using at least 90% less data than traditional methods. First, in simulation, we demonstrated that SBP can extract the spiking rates of individual units at lower signal-to-noise ratios than the standard high-bandwidth threshold crossing rate. This is particularly important for the usage of Utah microelectrode arrays (the only human-approved intracortical electrode array for chronic implantation) as it has been shown that the signal-to-noise

ratios of recorded units wane over time (Barrese et al., 2013; Hughes et al., 2021). Additionally, we also found that SBP can extract spiking information with greater single-unit specificity than the standard threshold crossing rate, as its computation gives greater weight to larger amplitude units recorded on a given electrode. This has significant implications for real-time BMI usage, where signals that are least corrupted by noise and potentially conflicting information from other nearby neural units will likely be more valuable to decoders. We then demonstrated that SBP exceeds the decoding performance of threshold crossing rate alone in a closed-loop one-finger BMI task with nonhuman primates (later validated again for two-finger BMI tasks in Chapter 4 and Appendix B).

From an engineering perspective, SBP provides a viable solution to one of the biggest hindrances to clinical translation of BMIs: power consumption. Requiring 90% less data per channel while simultaneously providing the aforementioned benefits leaves space for other required BMI processes to be incorporated into a portable device. Using SBP, we developed embedded and integrated BMIs, presented in Chapter 3, to show that end-to-end brain-machine interfaces can be made with current technology. Our embedded and integrated BMIs predicted the intended one-dimensional finger movements of a nonhuman primate implanted with a Utah array in real-time with performance that exceeds what we previously published (discussed in Chapter 2) and achieved reasonable performance in a two-finger version of the task compared to what we previously presented (discussed in Chapter 4). Such a result is immediately relevant to medical device manufacturers, where off-the-shelf technology can be used to produce a two-dimensional BMI, or efforts can be put towards an integrated solution that has the potential to deliver substantially longer battery life.

While SBP is the primary advancement in this work that led to BMI compression to embedded and integrated systems, there is a large optimization space for portable or implantable BMIs. One example involves decoding. The steady-state Kalman filter used with our embedded and integrated BMIs is a computationally efficient algorithm that provided great prediction performance, but there was still residual computational budget that could allow for greater complexity. We have recently found evidence that relatively simple neural network decoders can outperform the ReFIT Kalman filter (Willsey et al., 2021), and recurrent networks have been shown to model the inherent cortical dynamics underlying the spiking activity being recorded (Pandarinath et al., 2018; Sussillo et al., 2012, 2016). It remains to be seen if the performance improvements from these decoders warrants the increase in computational complexity and its associated increase in power consumption. A second potential optimization to the devices presented in Chapter 3 could revolve around the power-efficient off-the-shelf amplifiers (Intan RHD2132). While amplification and digitization of all 96 channels required just 12mW, it remains unclear if alternative amplifiers with greater input impedances (to yield higher signal-to-noise ratios) could improve performance without substantially increasing power consumption. Despite these potential optimizations, the devices as presented in Chapter 3 require small enough footprints that they could be considered fit as implantable devices, so next steps involve developing a hermetic package to guarantee continued device functionality following implantation.

6.2 Functional Restoration of Hand Movements with Brain-Machine Interfaces

Low-power and portable BMIs have the potential to strongly improve the lives of people with paralysis, but that impact is strongly dependent on the prosthesis the BMI is controlling. Restoration of upper extremity function is of the highest priority to people with tetraplegia (Anderson, 2004), and many people with tetraplegia would be willing to undergo invasive brain

surgery to implant a device such as a Utah array if it restored substantial upper extremity function (Blabe et al., 2015). However, BMI-controlled hand function has remained limited to whole hand grasping patterns (Aggarwal et al., 2013; Bansal et al., 2011; Okorokova et al., 2020; Vargas-Irwin et al., 2010), which assists with restoring discrete grasps but may not be sufficient to return dexterous, individual finger function.

In Chapter 4, we showed that nonhuman primates can use a brain-machine interface to control the simultaneous and independent movements of two finger groups, an important step towards the restoration of dexterous hand function in cases of high tetraplegia. Such a result is particularly relevant to groups working to restore hand function with people, where continuous and individuated finger function could complement the established grasp results already presented (Ajiboye et al., 2017; Collinger et al., 2013; Flesher et al., 2021; Wodlinger et al., 2015). Furthermore, the simplicity of the linear decoder (Kalman filter) used for closed-loop control implies immediate relevance to groups interested in portable or implantable BMIs, as we discussed in Chapter 3. Interestingly, we found that the cortical activity representing combined movements of both finger groups could be well represented by a linear combination of the cortical activity representing the independent movements of those finger groups. This allowed us to accurately decode all combined and independent behaviors offline without training on all of those behaviors, suggesting that simple decoders for complex but related behaviors may not require training on all permutations of those behaviors.

In Chapter 5, we restored continuous voluntary hand function to a temporarily paralyzed nonhuman primate using brain-controlled functional electrical stimulation. Brain-controlled functional electrical stimulation of the upper extremity is not novel, with a small number of studies having presented nonhuman primate and human results across the past 1.5 decades (Ajiboye et al.,

2017; Bouton et al., 2016; Ethier et al., 2012; Moritz et al., 2008). As impactful as these results were, functional restoration was primarily limited to grasps and did not investigate dexterous hand function. The continuous brain-controlled functional electrical stimulation results we presented here could complement grasp-focused neuroprostheses by providing fine motor control, bringing upper extremity neuroprostheses closer to native capabilities. Importantly, we demonstrated for the first time that independent and continuous two-finger movements can be predicted from cortical activity even in a state of paralysis. By performing recalibrated feedback-intention training one time following paralysis, a Kalman filter was capable of achieving the same closed-loop BMI performance as it could prior to paralysis for predicting individuated finger movements. Evidence suggests that the loss of sensory feedback was the primary cause of performance reduction following paralysis, but sensory feedback was not necessary to individuate finger movements from cortical activity.

While the restoration of two finger degrees of freedom from motor cortex activity can restore substantial function to people with paralysis (thumb and/or finger flexion for object manipulation), completely natural hand function will eventually require restoration of more hand degrees of freedom. For example, without wrist function, the formation of a variety of grips would restore limited object interaction capabilities. Furthermore, while it is encouraging that the control of multiple finger degrees-of-freedom can be restored with a BMI, physical interactions with objects will require a non-virtual prosthesis. As important as it is to increase the quantity of hand degrees-of-freedom that can be controlled by a BMI, it is equally valuable to demonstrate that those degrees of freedom can be restored during prosthesis use, either via functional electrical stimulation or a robotic arm. Finally, primate models of brain-controlled functional electrical stimulation are critical for rapid development of these technologies for use with people who would

159

benefit most from them. Beyond abstract actions like continuous finger movements, restoring grasp and continuous finger function together to a temporarily paralyzed monkey might provide a neuroprosthesis with which he could grab treats or complete dexterity enrichment puzzles. Demonstrating tangible benefits of brain-controlled functional electrical stimulation in primate models can inform good practices for designing similar systems for use with people.

6.3 Future Directions

The results outlined in this thesis suggest that fully implantable, end-to-end neuroprostheses can be made with existing technologies. The Networked Neuroprosthesis (Makowski et al., 2021) used for functional electrical stimulation in this work has already been implanted in five participants in a clinical trial (NCT02329652), and the embedded BMI already incorporates the necessary network hardware to send stimulation commands to the existing NNP modules. Immediate next steps involve using the embedded BMI to perform end-to-end neuroprosthetic control of a paralyzed primate hand with completely implantable hardware. Looking further into the future, benchtop systems already exist that are viable for implantation, but implantable BMIs will require development of high-channel count hermetic feedthroughs to support transmission of hundreds or thousands of signals. I expect the first-in-human clinical trials will begin for fully implanted brain-controlled functional electrical stimulation systems within ten years.

Despite functional electrical stimulation having been used to restore function to paralyzed musculature for nearly four decades, some of the original problems persist, like muscle fatigue and hysteresis. Somewhat recently, cortical optogenetic stimulation, in which optical pulses are delivered to genetically-modified neurons to inhibit or excite them, has allowed fascinating neuroscientific studies in which individual cortical neurons can be perturbed in their natural

processing circuits (Boyden et al., 2005). More recently, functional optogenetic stimulation was shown to be capable of restoring limb flexion in mice with more natural recruitment and less muscle fatiguing than functional electrical stimulation (Srinivasan et al., 2018), reducing or potentially resolving some of the long standing issues of electrical stimulation. In its current technological form, optogenetics requires viral interventions to alter the ion channels on target neurons. Without immunosuppression, the native immune system might reject the necessary genetic change, which is a major health concern for pre-clinical primate studies before consideration in human use. However, if safe methodologies are developed, functional optogenetic stimulation may provide a more viable avenue for widespread motor restoration in paralysis.

From the brain-machine interface perspective, the Utah array (or NeuroPort array as approved for human use) has been and is key in providing countless animal and human neuroscientific and technological advances since its introduction (Maynard et al., 1997), but the technology has not substantially advanced since. Fortunately, animals and people who have been implanted with Utah arrays rarely show lasting cognitive or functional deficits, but there have been studies illustrating a substantial loss of neurons surrounding each intracortical electrode of the array (McCreery et al., 2016; Potter et al., 2012; Sharma et al., 2015). Novel electrode materials and designs are being proposed that may reduce the immune response to implantation, increase recording yield, extend the functional lifetime of recording, and increase recording density (Canales et al., 2015; Cohen-Karni et al., 2009; Guitchounts et al., 2013; Knaack et al., 2016; Kozai et al., 2012; Luan et al., 2017; Obaid et al., 2020; Patel et al., 2016; Rodger et al., 2008; Yang et al., 2019). While many of these technologies are presently targeting neuroscience studies, within ten years, some may be challenging the Utah array for chronic implantation in humans for use with neuroprosthetic devices.

A recently popular neurotechnological research area is the development of micron-scale, wireless recording devices for chronic, untethered implantation. Infection risks associated with chronic percutaneous pedestals, such as what is required for the Utah array, could be entirely avoided if the scalp skin can be resealed following an intracortical implant. Several groups are proposing these "chiplet"-style systems that have one or few recording sites per 100μm-1mm-scale device. Each device is untethered, floating on the surface of the brain, with power and recorded samples transmitted wirelessly to and from a receiver or repeater replacing the skull or on the scalp. Many wireless communication media are being investigated, between ultrasound (Seo et al., 2016), radio frequency (Lee et al., 2019), magneto-electric (Singer et al., 2020), and optics(Lee et al., 2018; Lim et al., 2021), and it remains unknown which will best support expansion to thousands of recording channels. Only one of these devices has accomplished *in vivo* recording (Lee et al., 2021), but a decade from now might see these devices entering pre-clinical nonhuman primate implantations with the intentions of informing a future human-grade, wireless, intracortical implant.

Some of our work has taken advantage of spiking band power to develop ReMotes, our optically-powered wireless neural recording chiplets, where the electrode providing the tissue interface is made of a carbon fiber. Of the chiplet devices proposed to date, none but the ReMote are simultaneously taking advantage of recent electrode technologies to further improve yield and health prospects. We are hopeful that carbon fiber ReMotes can provide higher-density and higher quantity recordings, but the marriage of these two next-generation technologies paints a picture of what chiplet-style recording technologies may be proposed a decade from now. Current chiplets are generally shaped like a lollipop: a thin electrode that traverses >1mm into cortical tissue with a large integrated circuit on top containing the amplifier, analog-to-digital converter, and wireless

communication and power delivery circuitry. With Moore's law continuing to somewhat accurately predict the compression of digital circuits combined with the probable relaxation of neural recording requirements (similar to spiking band power), chiplets may look less like lollipops and more like straws. With enough compression, one could envision the required amplifier, digitization, and telemetry circuitry fitting within the length of the probe itself, keeping the active recording contact in the small ~10μm space at the distal tip. This could substantially increase recording density while keeping immune response and cell death to a minimum, particularly without the large box of circuitry at the top of each device.

On a bit shorter of a timescale, recent venture interest in the neurotechnology field has enabled some devices to expand beyond one-and-done functionality to become viable products that may translate to human use in the near future. Neuralink provides one example (Musk, 2019), where one coin-sized device can record from 1,024 channels and transmit threshold crossing timings wirelessly over a Bluetooth link (Yoon et al., 2021). Translation of Neuralink's device to safe usage in people would almost certainly be very impactful to people with paralysis or other neurological deficits, but will also advance the technology used in academic neuroscience and neural engineering. Part of the substantial success of the Utah array is due to its immediate relevance to humans: if a result is found with a Utah array in animals, then it can be immediately investigated with a NeuroPort array in humans. A higher channel count, wireless, completely implantable neural interface like Neuralink's device could advance the world of human-relevant animal studies, from advancing motor neuroprosthetic studies to making cage-side neuroscience easier.

From a neuroprosthetic perspective, a Bluetooth-capable neural recording device could generate a major leap forward in terms of the decoding algorithms relevant to portable devices.

The embedded and integrated BMIs presented in this work used simple linear Kalman filters with little room for more computational complexity. Although neural network decoders have begun showing promise in improving prediction performance (Sussillo et al., 2016; Willett et al., 2021; Willsey et al., 2021), they have been loosely relevant to portable or implantable devices due to their computational complexity. In contrast, a Bluetooth-capable device could access the processing capabilities of a person's cell phone or general-purpose computer, providing an avenue through which complicated neural network decoders can become portable. Within ten years, whether it be Neuralink's device or another, there may be people with paralysis that have implanted, Bluetooth-capable brain-machine interfaces that utilize their cell phones for computing high-performance decoder updates. The technological requirements to do so already exist: high-definition Bluetooth audio already supports 48kHz sampling rates at 16-bit resolution, yielding a communication bandwidth of 768kbps. For a 1,024 channel system at 12-bit resolution (the effective resolution we used with our integrated BMI, the NeuRAD), 62.5 features per second per channel can be transmitted with the same bandwidth, which is faster than most decoder update rates in use with neural networks or linear decoders to date.

Technological advancements often come in one of two varieties: enhancement or compression, with the occasional case that accomplishes both. Either the performance is increased and requires greater complexity, or an optimization is introduced that reduces complexity without sacrificing performance. Neuroprosthetic advancements follow the same trend, and a Bluetooth-capable neural interface that uses external devices for high-performance, complex decoding could be one of the rare advancements that accomplishes both enhancement and compression. The future of neural engineering and neurotechnology is bright, and within ten years, I expect to see the first

clinical trial of a fully-implanted brain-controlled neuroprosthesis for the restoration of upper extremity function.

Appendix

Supplementary Material for Chapter 2, Chapter 4, and Chapter 5

Appendix A Supplement to Chapter 2
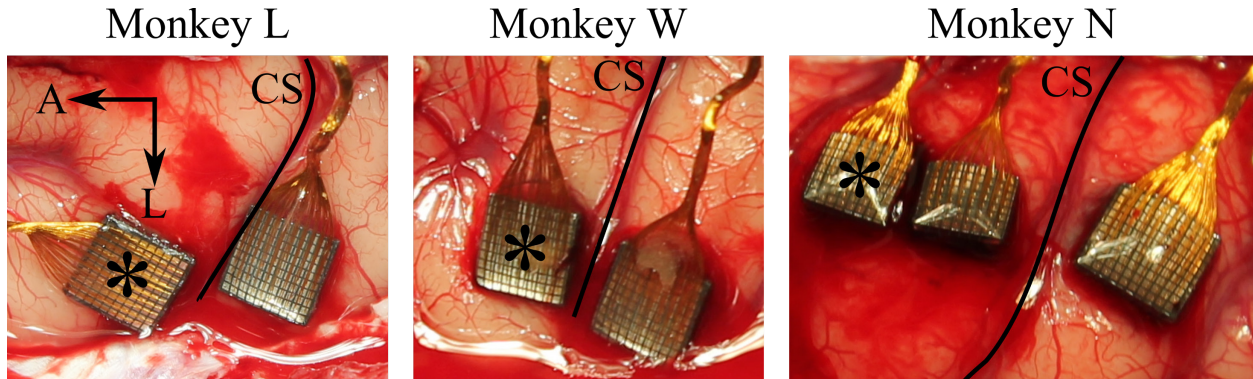
A.1 Supplementary Methods Figures

Figure A.1 Surgical photographs of microelectrode array implants. The arrays labeled with an asterisk are the only arrays used in this study. CS indicates central sulcus, A indicates anterior direction, L indicates lateral direction.
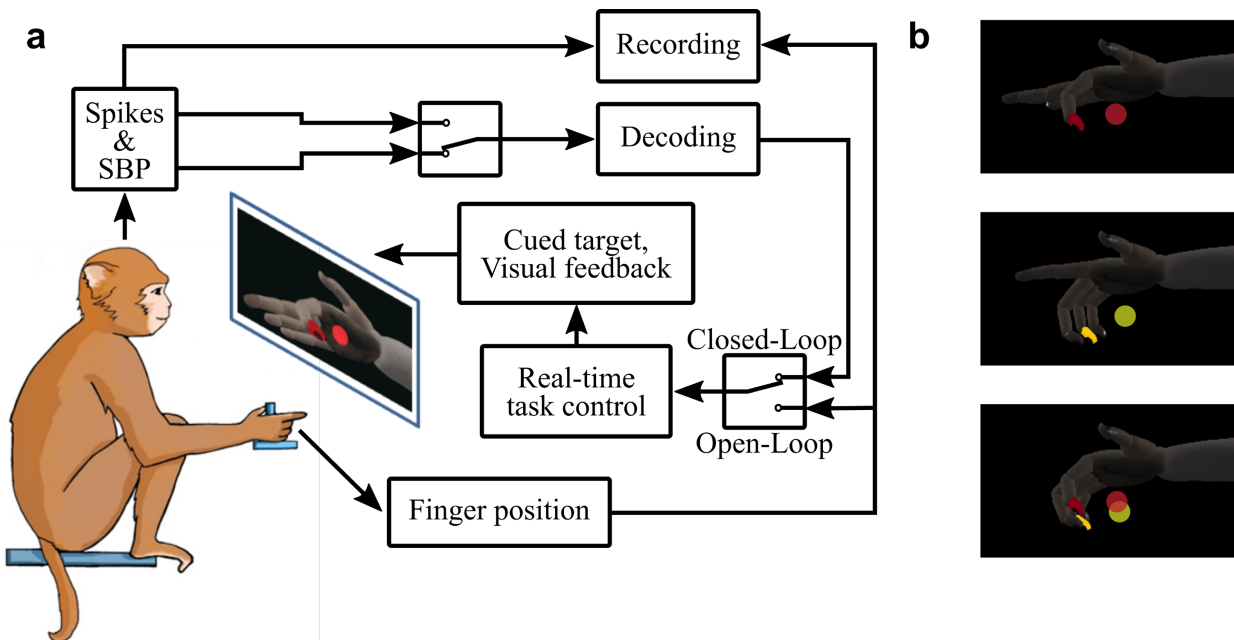
Figure A.2 Diagram of behavioral task. The monkey moved separate groups of fingers (either index alone, middle-ring-small (MRS) alone, or all four fingers together) to hit virtual targets presented on a computer screen. The virtual fingers were controlled by the monkey's physical finger movements or the decoder's translation of the brain activity. Later, we trained monkeys W and N to acquire two targets simultaneously, one in red for the index finger and one in yellow for the MRS finger group.
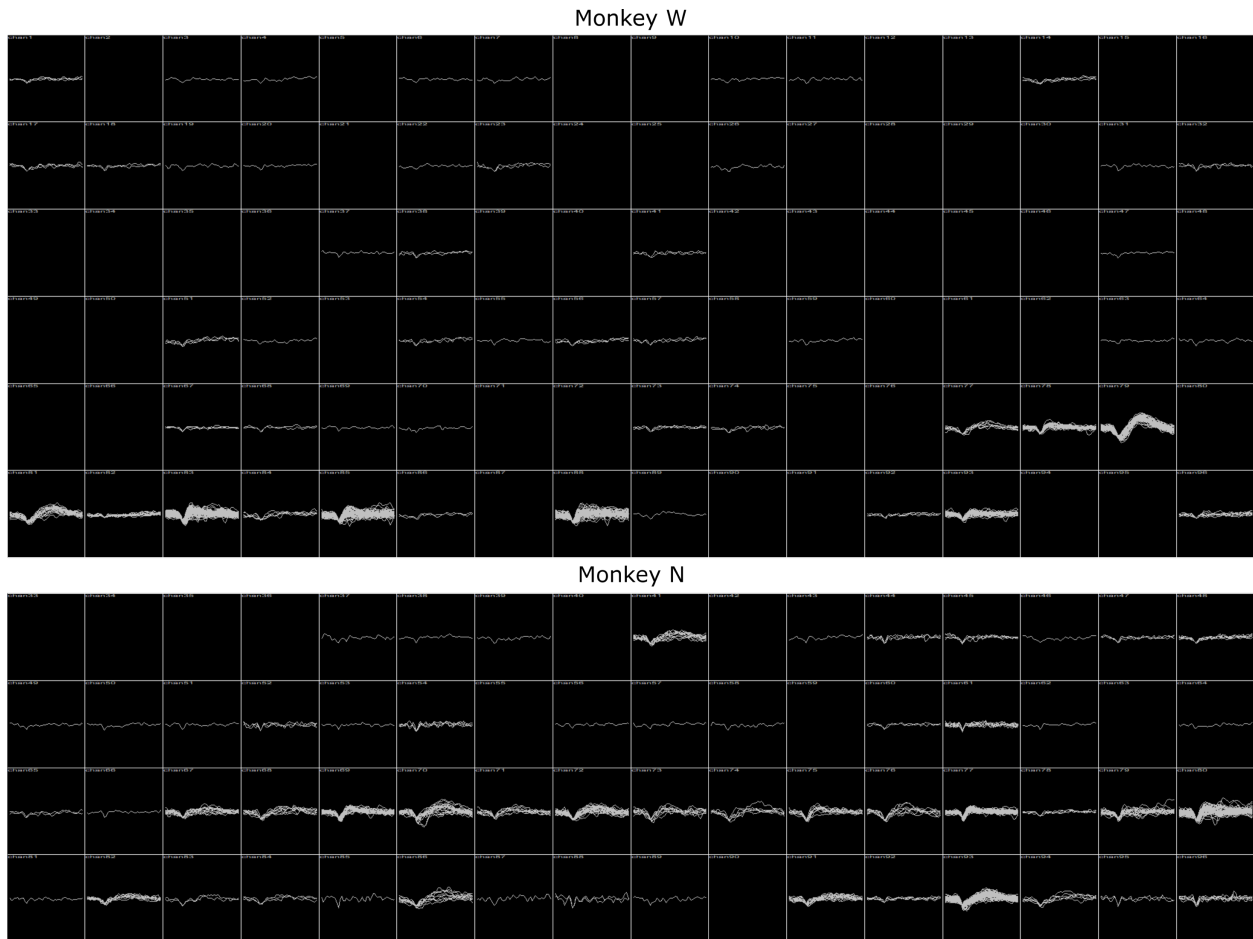
Figure A.3 Snapshots of monkey W's (top) and monkey N's (bottom) arrays on the days of decoding presented in Figure 2.4a with thresholds set at −4.5 × root-mean-square (RMS). Monkey W's array recorded from no more than 10 potential units to achieve the presented decode performance, and monkey N's array exhibited the units as shown from premotor areas.

A.2 Supplementary Data

A.2.1 RMS-Based SNR Analysis

The definition of SNR used in the main-text is commonly used to represent the SNRs of single units (Byun et al., 2017; Wark et al., 2013; Xu et al., 2018). However, taking a signal processing approach to the discussion of SBP's capability of extracting low-SNR single units provides a different perspective of the spiking power in a signal. To relate the root-mean-square (RMS)-based SNR definition to neural activity, we must re-evaluate the RMS formula for neurons,
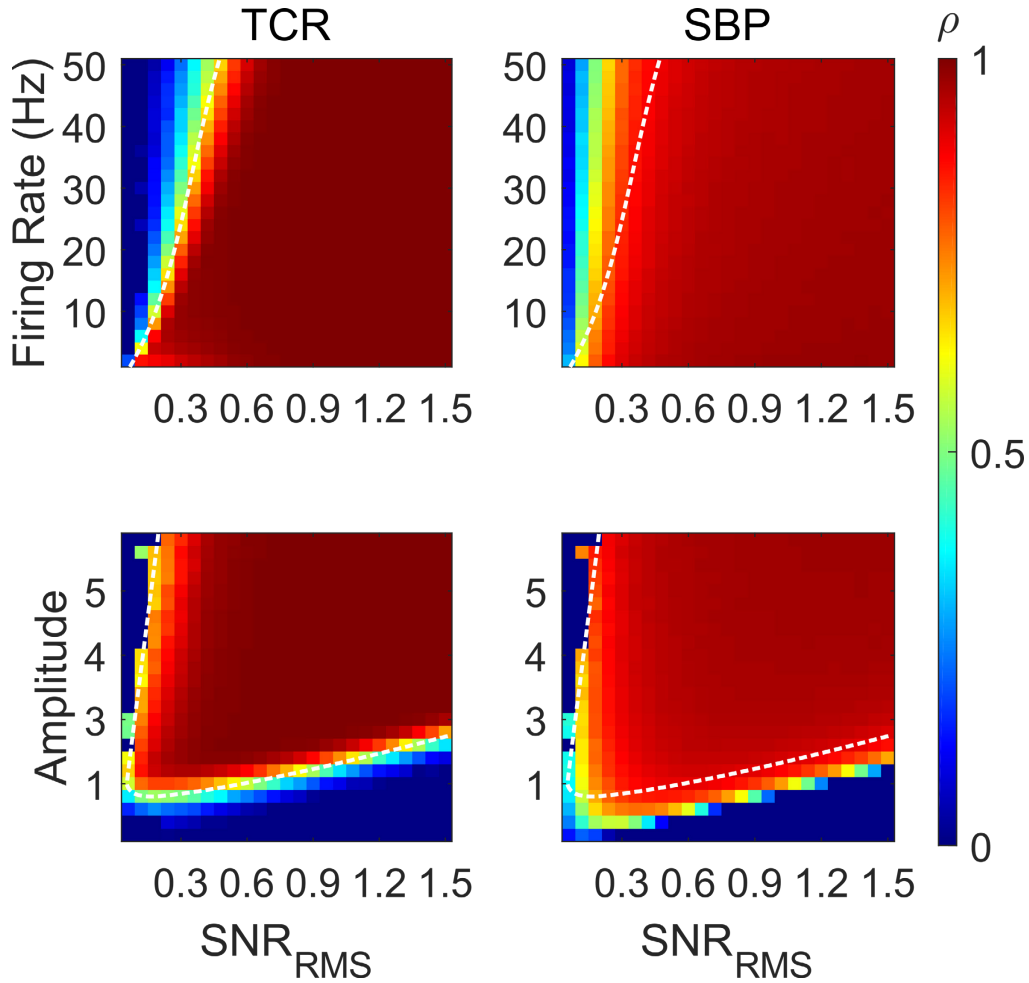
Figure A.4 SBP and TCR prediction of true firing rate using the signal RMS-based SNR definition. Color is correlation between the true firing rate and that predicted by each feature. The white dotted lines track the yellow region of the left-most TCR plot to better compare low SNR_{RMS} performance. *Top:* Firing rate prediction performance with changing firing rate, where the spike amplitude is varied to generate the different SNR_{RMS} values for a given firing rate. *Bottom:* Firing rate prediction performance with changing spike amplitude to noise ratio, where the firing rate is varied to generate the different SNR_{RMS} values for a given ratio. The dark blue regions to the left of all three plots resulted from a firing rate of 0.

whose activity is not defined as a continuous time signal but rather as a relatively sparse firing rate. Thus, we propose that the RMS value of a noiseless spiking signal is the square root of the mean value of one spike's energy, multiplied by the quantity of spikes in the signal, divided by the total number of samples in the signal. Thus, we can represent our SNR_{RMS} definition as

$$\text{SNR}_{\text{RMS}} = \frac{\text{RMS(signal)}}{\text{RMS(noise)}} = \frac{\sqrt{n_{\text{spikes}} \sum_{i=1}^{\ell_{\text{spike}}} \left(A \cdot \text{spike}(i)\right)^2}}{\text{RMS(noise)}}$$
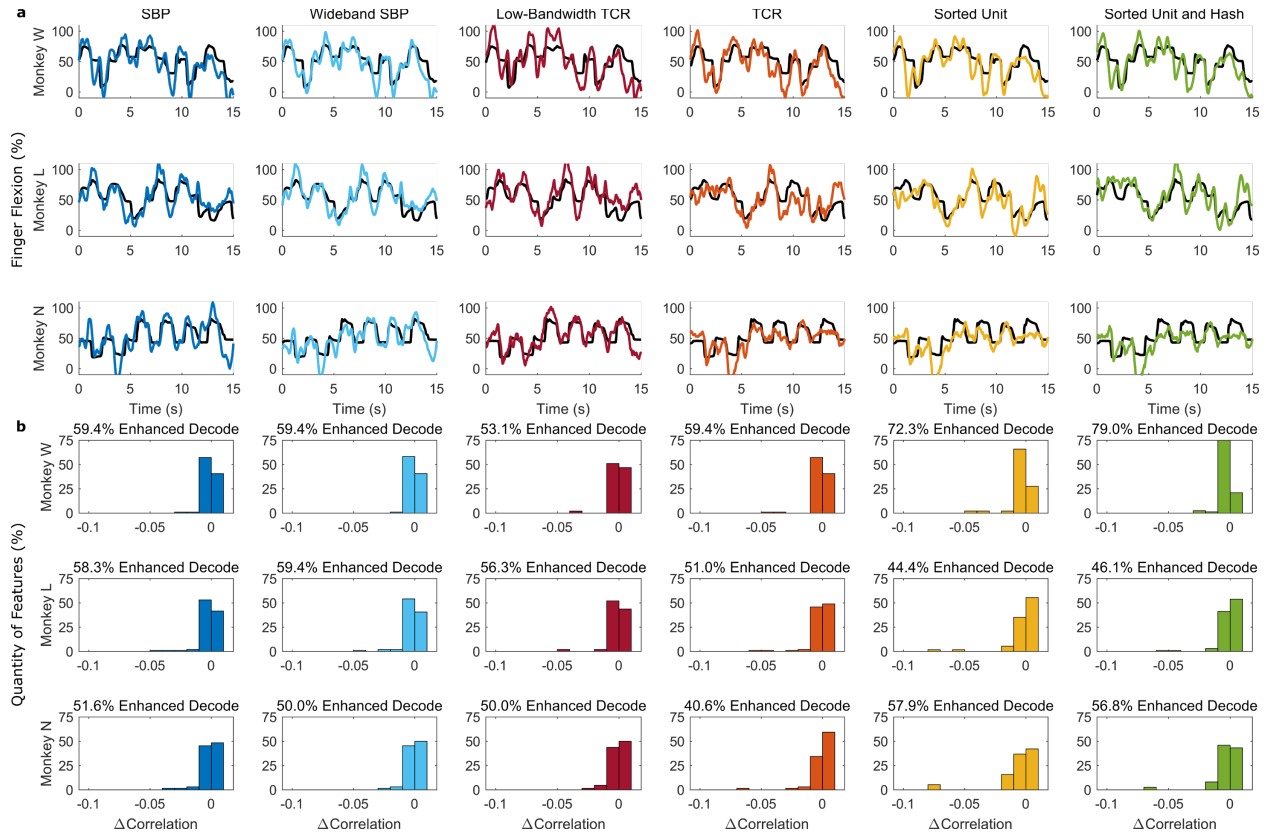
Figure A.5 Open-loop decode results using a standard Kalman filter to decode SBP, 300-6,000Hz wideband SBP, low-bandwidth TCR, TCR, single unit firing rate, and single unit firing rate with neural hash. (a) Decode traces overlaid on the actual finger positions in black. **(b)** Performance lost by omitting individual channels or units during decodes. The histograms represent the percentage of total channels or units that, when left out of the decode individually, resulted in a correlation loss within each bin. Lower numbers indicate the omitted channel was more valuable towards the overall decode performance. Above each histogram is the percentage of available features with negative correlation differences.

where we included the term A to maintain the capability of scaling the neuron's relative amplitude. Consequently, the SNR_{RMS} values corresponding to realistic firing rates and amplitudes seen in motor cortex, such as those presented in Figure 2.2, will be small. As we now have a four-dimensional problem (firing rate, relative amplitude, SNR_{RMS}, and correlation coefficient), we present results in Figure A.4 in a similar format to Figure 2.2 but in parameterized plots.

The results in the top of Figure A.4 suggest that SBP is capable of accurately predicting neural activity at lower SNR_{RMS} values than TCR, except for small, rare motor cortical spiking frequencies in the bottom left of the plots (e.g. average firing rates below 2Hz are not informative

Table A.1 Optimized open-loop decode results.

Subject ρ RMSE	SBP	WSBP	LbTCR	TCR	SU	SUH
Monkey W	0.730^a	0.770^b	0.694^c	0.727^a	0.722^a	0.750^d
$n = 45,124$	0.176^a	0.161^b	0.190^c	0.168^b	0.178^a	0.167^b
Monkey L	0.760^a	0.782^b	0.715^c	0.694^d	0.665^e	$0.695^{c,d}$
$n = 37,203$	0.149^a	0.143^a	0.163^b	0.164^b	0.188^c	$0.177^{b,c}$
Monkey N	$0.708^{a,b}$	0.732^a	0.684^b	0.689^b	0.676^b	0.686^b
$n = 7,547$	0.194^a	0.185^a	0.173^a	0.171^a	0.174^a	0.172^a

Correlation ρ is Pearson's correlation coefficient between the actual finger trace and that predicted by each feature. Root-mean-squared error (RMSE) was also calculated between the actual and predicted finger traces. n is the number of samples 50ms in size used for each calculation. Given that the optimal amount of historical bins and offset between the neural data and behavior differed slightly between features, the minimum number of bins per dataset is shown. The letters noted by each statistic indicate statistically similar numbers within one subject (i.e. numbers with the same letter are not statistically different, $p < 1 \times 10^{-4}$, two-tailed two-sample z-tests on Fisher's z-scores for the correlation coefficients comparisons, one-tailed two-sample Wilcoxon rank-sum tests for the RMSE comparisons). WSBP: wideband SBP, LbTCR: low-bandwidth TCR, SU: sorted unit firing rate, SUH: sorted unit firing rate with neural hash rate.

for decoders). The bottom of Figure A.4 suggests that SBP can extract the firing patterns of smaller amplitude units (equivalent to the results from the amplitude-based SNR definition in Figure 2.2a), though not at the extremely low SNR_{RMS} values (bottom-left corner of the plots) that TCR is better able to extract.

A.2.2 Open-Loop Decoding Analysis

We performed open-loop Kalman filter decodes to compare SBP to 300-6,000Hz wideband SBP (i.e. from Stark and Abeles 2007), low-bandwidth TCR, TCR, sorted unit firing rates, and combined sorted unit firing rate with hash rates. We optimized two decoder aspects for each feature and each set of data: the number of historical neural data bins, and the offset between neural data and behavior, anywhere between 0 and 5 bins. For low-bandwidth TCR, we also optimized the threshold for each set of data. Figure A.5a illustrates predictions made by each optimized decoder along with the true hand position overlaid. SBP achieved statistically equivalent or better correlation coefficients than low-bandwidth TCR, TCR, and single unit firing rate in all animals

Table A.2 Parameters optimized per dataset for open-loop decoding.

Optimized Parameter	Monkey	SBP	WSBP	LbTCR	TCR	SU	SUH
No. Additional Bins	W	5	5	4	1	5	5
	L	1	1	1	1	4	4
	N_1^*	4	5	5	0	0	0
	N_2	5	5	0	0	0	0
Bin Offset	W	0	0	0	0	0	0
	L	0	0	0	0	0	0
	N_1^*	0	0	0	3	2	3
	N_2	0	0	3	2	2	2
Threshold	W	-	-	2.40	-	-	-
	L	-	-	2.25	-	-	-
	N_1^*	-	-	2.20	-	-	-
	N_2	-	-	2.15	-	-	-

Monkey N's dataset 1 (marked by *) was not used for any other analyses, but included here to demonstrate variability in optimal parameters within one animal. "No. Additional Bins" is the quantity of additional historical neural bins, between 0 and 5, that achieved the highest correlation coefficient. "Bin Offset" is the delay, in bins from 0 to 5, between the neural activity and the behavior that achieved the highest correlation coefficient. "Threshold" is the RMS threshold that achieved the highest correlation coefficient.

and additionally single unit firing rate with hash rate in monkeys L and N (results in Table A.1 with optimized parameters in Table A.2). We also found that wideband SBP outperformed all other neural features, which complies with the result found by Stark and Abeles, and that there is a small drop in performance resulting from the bandwidth reduction from ~6kHz to ~1kHz as was demonstrated by Irwin *et al.* (Irwin et al., 2016; Stark and Abeles, 2007).

Since SBP can extract unit-specific neural activity from SNRs below typical threshold levels, we also investigated the quantity of recorded channels valuable for decoding. We re-processed the optimized decodes, except trained and tested on neural data lacking one feature at a time. Then, we calculated the new correlation between the finger position predicted by the decoder lacking a feature and the actual finger position and found the difference between it and the correlation resulting from all features included. Thus, a more negative difference means the excluded feature was more valuable to the decoder. The histograms in Figure A.5b represent these
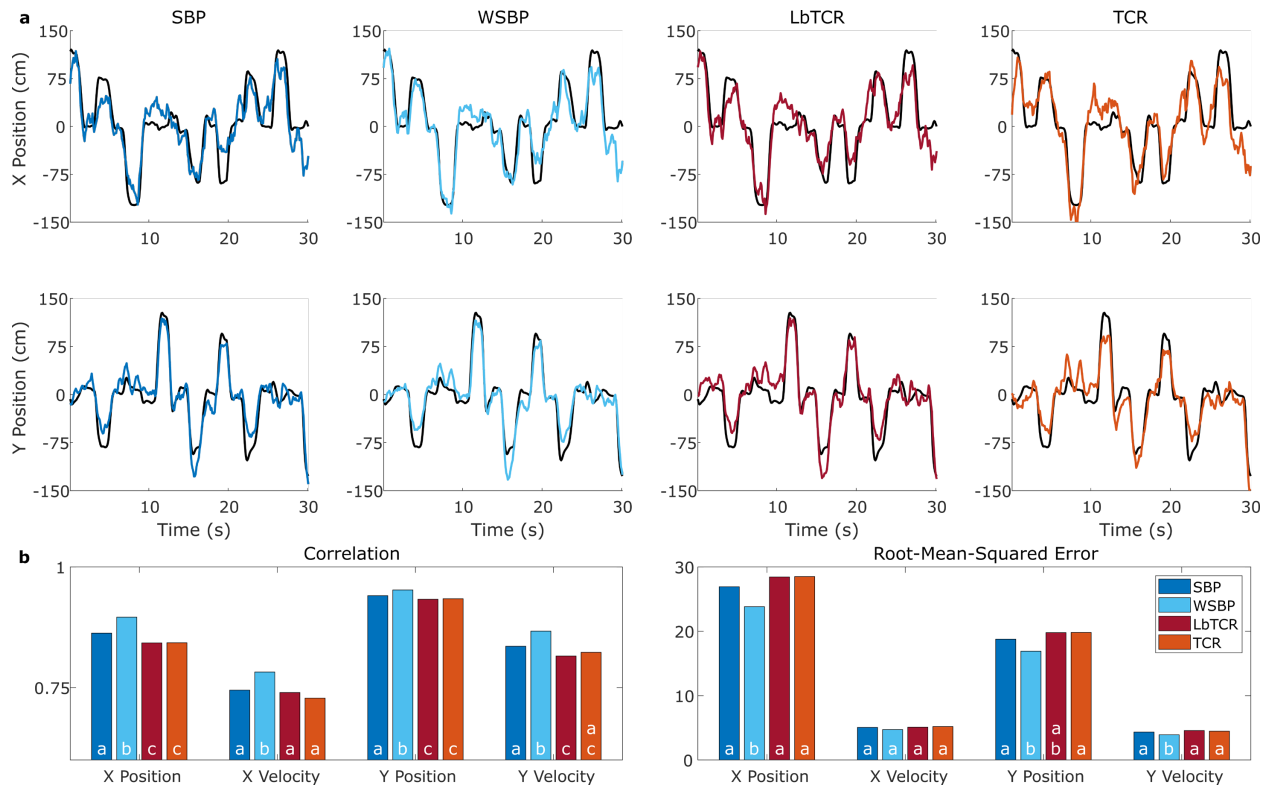
Figure A.6 Open-loop two-dimensional cursor control task for monkey J over 506 trials, or $n = 9,295$ samples 50ms in size. (**a**) Example decodes for SBP, wideband SBP, low-bandwidth TCR, and TCR. (**b**) Decode statistics. Bars in a group labelled with different letters are statistically different. $p < 1 \times 10^{-3}$, two-sided two-tailed z-test for Pearson's correlation coefficients. $p < 1 \times 10^{-3}$, two-tailed two-sample Wilcoxon rank-sum test for root-mean-squared errors.

differences in correlation. As follows from Figure 2.2, there were generally more SBP features that resulted in negative correlation differences than threshold crossings, suggesting SBP may extract decodable neural information from channels that threshold crossings may find invaluable. Sorted unit features expectedly had some higher values, as manual feature extraction excludes lots of invaluable information.

A.2.3 Two-Dimensional Cursor Control

To validate that SBP maintains performance in a clinically-relevant state-of-the-art task, we compared SBP, wideband SBP, TCR (threshold at -4.5RMS), and low-bandwidth TCR (thresholded at 2.25RMS, the averaged optimal value from the finger offline decodes) at predicting two-dimensional cursor control. Specifically, we analyzed some offline datasets from monkey L
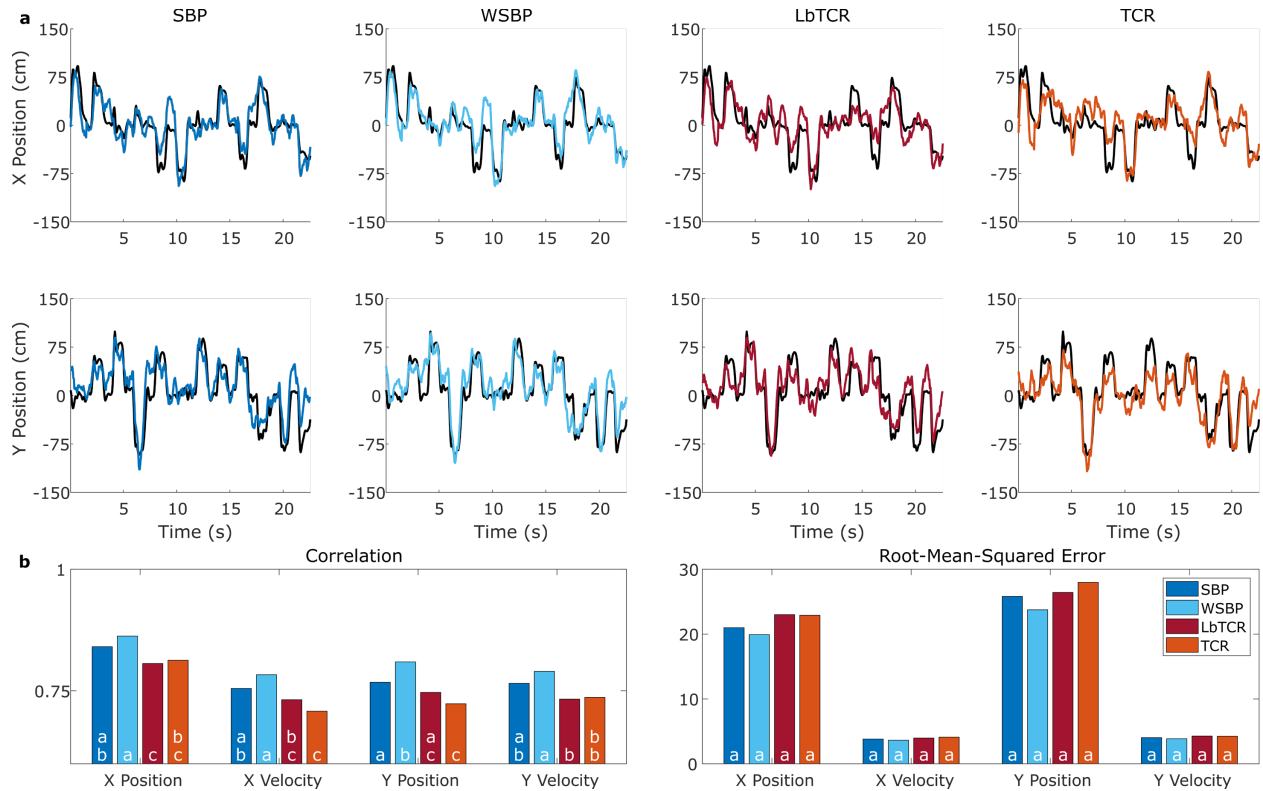
Figure A.7 Open-loop two-dimensional cursor control task for monkey L over 117 trials, or *n* = 2,283 samples 50ms in size. (**a**) Example decodes for SBP, wideband SBP, low-bandwidth TCR, and TCR. (**b**) Decode statistics. Bars in a group labelled with different letters are statistically different. $p < 1 \times 10^{-3}$, two-sided two-tailed *z*-test for Pearson's correlation coefficients. $p < 1 \times 10^{-3}$, two-tailed two-sample Wilcoxon rank-sum test for root-mean-squared errors.

from a prior study (L120502) and monkey J (J121009), which is the same animal with the same implants as described previously (Kao et al., 2017). Briefly, the monkey was trained to perform center-out-and-back arm reaches along 8 directions while 30kSps neural activity was recorded.

We used a standard Kalman filter to decode 506 trials within one day for monkey J using all 96 channels and 117 trials within one day for monkey L using 95 channels. We excluded channels without units from all decodes of all features, and all features were binned in 50ms bins. To avoid over-fitting, we 10-fold cross-validated each dataset. Statistical tests followed the procedures defined in the open-loop decode methods, section 2.2.5.

We found that SBP achieved statistically as good or better decoding performance than TCR and low-bandwidth TCR, which is consistent with the results of our offline one-dimensional finger
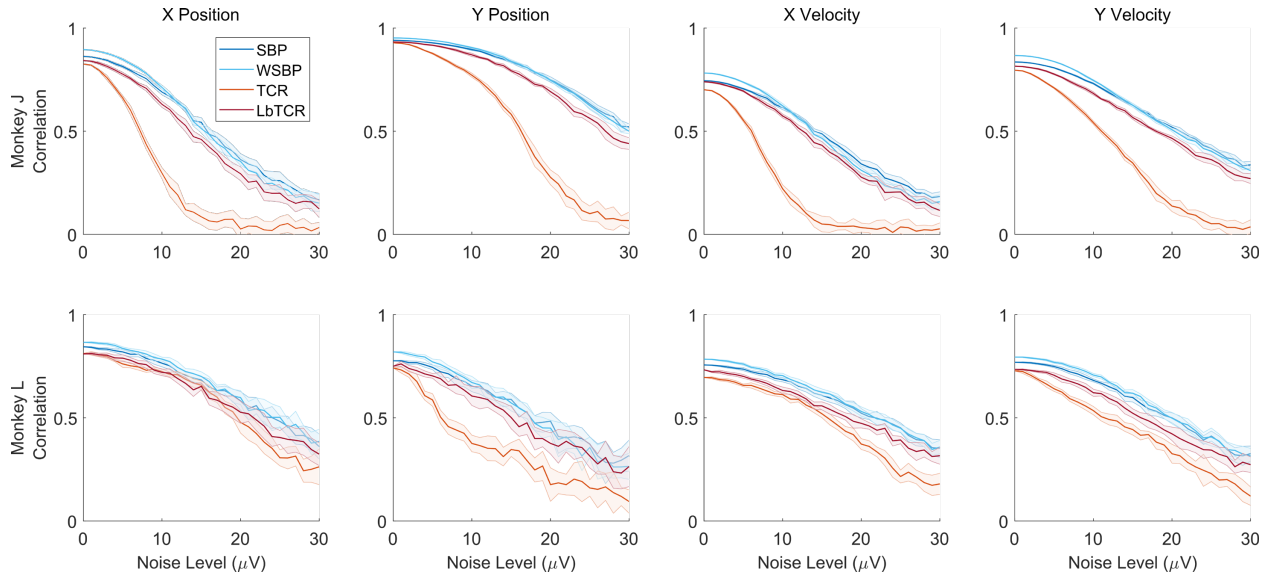
Figure A.8 Open-loop two-dimensional cursor control correlation coefficients for monkeys J (top) and L (bottom) with manually injected white noise at various levels from 0 to 30µV. Solid lines represent the mean correlation coefficient of each feature's decode at each injected noise level. Shaded area surrounding each solid line represents standard deviation across the 10 trials of randomly generated noise.

tasks. Wideband SBP achieved the highest performance of any feature, as was expected from others' results and our one-dimensional results (Stark and Abeles, 2007). Figure A.6 and Figure A.7 compare the decodes for a few movements along with statistics. Monkey J's SBP had significantly higher positional correlation coefficients and as good or better velocity correlation coefficients than TCR features, and monkey L's SBP correlation coefficients were as good or better than TCR features ($p < 1 \times 10^{-3}$, two-sided two-sample z-test). Root-mean-squared errors were all statistically similar between SBP and TCR features, though numerically, SBP root-mean-squared errors were lower than the TCR features for both monkeys ($p < 1 \times 10^{-3}$, two-tailed two-sample Wilcoxon rank-sum test). Extrapolating these results to the outcomes of the one-dimensional open-loop and closed-loop decodes suggests that SBP may maintain performance at least as good as TCR in multi-dimensional closed-loop tasks.

Using these two-dimensional datasets, we can also investigate how robust each neural feature is to various amounts of noise. The results of section 2.3.1 suggest that low-bandwidth

Table A.3 The amount of noise added to each channel before extracting features that resulted in a 20% loss in performance in μV, nV/√Hz.

Neural	Monkey J				Monkey L			
Feature	X Pos.	Y Pos.	X Vel.	Y Vel.	X Pos.	Y Pos.	X Vel.	Y Vel.
SBP	9, 340	19, 718	10, 378	12, 454	15, 567	12, 454	15, 567	13, 491
WSBP	9, 119	19, 252	9, 119	12, 159	15, 199	11, 146	14, 185	13, 172
LbTCR	8, 302	17, 643	9, 340	11, 416	15, 567	10, 378	13, 491	12, 454
TCR	5, 66.2	11, 146	4, 53.0	7, 92.7	15, 199	4, 53.0	13, 172	7, 92.7

The header above the table reads: Noise Added to Generate 20% Performance Loss $\mu V, \frac{nV}{\sqrt{Hz}}$

Higher numbers imply more noise could be tolerated to maintain 80% of the maximum performance. Results averaged from 10 random generations of noise for each noise level.

features are more robust to noise than higher bandwidth threshold crossings when predicting unit firing rates, so we superimposed white noise at RMS levels from 0 to 30μV in increments of 1μV before extracting features. We selected a maximum of 30μV of noise because, at that noise level, decode correlations were around 0.5, a number low enough to suggest poor performance. We generated random noise ten times at each noise level to obtain the average effects of noise on decoding performance.

The averaged results are plotted in Figure A.8. We found that broadband TCR showed the most rapid dropoff of decode performance with increasing noise level, where the other features remained in the same neighborhood of performance. Additionally, SBP always maintained a higher decoding performance than low-bandwidth TCR at any level of injected noise, suggesting the performance enhancements of SBP are maintained even in lower-quality recordings. Lastly, Table A.3 demonstrates that the low-bandwidth features can accommodate almost three times the amount of noise as the higher bandwidth features to maintain 80% of the maximum performance, with SBP able to withstand at least as much noise as low-bandwidth TCR. Overall, this suggests that eliminating noise by restricting the passband to the spiking band increases low-bandwidth TCR and SBP's tolerance of noisy recordings, with the optimal low-bandwidth feature being SBP as it persistently achieved higher performance.

Table A.4 Summary of analog parameters for integrated circuit simulations.

Circuit Type	NEF	Noise V_{RMS}	U_T	k	T	Bandwidth
Low-Bandwidth	4.0	2μV	26.7mV	1.38×10^{-23}	310K	700Hz
High-Bandwidth						5,750Hz

NEF is the noise-efficiency factor, U_T is the thermal voltage, k is the Boltzmann Constant, T is temperature in Kelvin.

A.2.4 Integrated Circuit Simulations

Many groups have attempted to reduce circuit power by using lower power process nodes, optimized digital hardware, and data compression (Borton et al., 2013; Harrison et al., 2007, 2009; Rizk et al., 2009; Yin et al., 2013). To better realize the applicable benefits of low-bandwidth neural recording, we simulated fully customized low- and high-bandwidth-recording integrated neural interfaces using established equations to compare their estimated power consumptions. We defined the following requirements for our simulated circuits: acquire data, extract neural features, decode the features, and transmit the decoded values off-chip. The necessary components to accomplish those goals included analog amplifiers and filters, analog to digital converters (ADCs), feature extraction circuitry at a 50ms bin size, a Kalman filter hardware accelerator to decode, and a controller area network (CAN) interface operating at 100kbaud to communicate the predictions.

To estimate the power consumption of the analog front-ends, we first calculated the power of the amplifier front-end chain using the noise efficiency factor (NEF) formula proposed by Steyaert *et al.* (Steyaert et al., 1987):

$$I = \left(\frac{NEF}{V_{RMS}}\right)^2 \frac{\pi \cdot U_T \cdot 4\kappa T \cdot BW}{2}$$

with a thermal voltage U_T of 26.7mV, Boltzmann constant κ of 1.38×10^{-23}, temperature T of 310K, and signal bandwidth BW. These parameters are summarized in Table A.4. Even though some amplifiers (Shen et al., 2017) have approached the ideal NEF of 1, the NEF of a single bipolar junction transistor, neural recording amplifiers typically show NEF values larger than 4 because
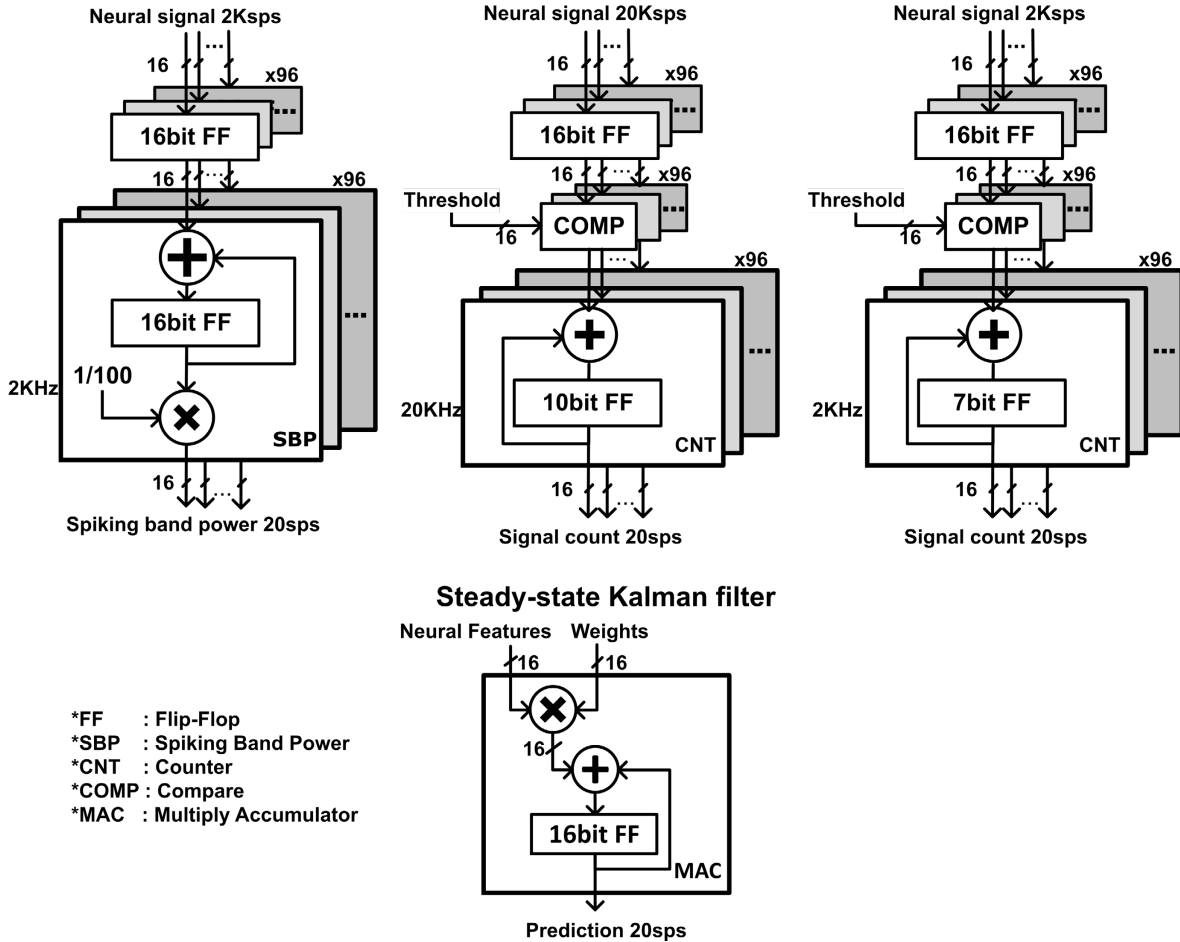
Figure A.9 Digital circuit designs. (*Left*) Digital circuit designs for SBP feature extraction with 2kSps neural recording. 16-bit data of 96 channels are stored in flip-flops to be accumulated and averaged for SBP. This logic executes at 2kHz to match the sampling rate of SBP. (*Middle*) Digital circuit design for threshold crossing rate with 20kSps neural recording. 16-bit data of 96 channels are stored in flip-flops to count the number of events where the sampled data has crossed the threshold value. This logic must execute at 20kHz to match the sampling rate of the broadband recordings. (*Right*) Digital circuit design for low-bandwidth TCR with 2kSps neural recording. 16-bit data of 96 channels are stored in flip-flops to count the number of events where the 2kSps data has crossed the threshold value. This logic executes at 2kHz to match the sampling rate of the low-bandwidth recordings. (*Bottom*) Computational unit for the steady-state Kalman filter using a 16-bit multiply-accumulator unit (MAC).

of other stringent specifications for their robust operation, such as high input impedance, high common mode ratio, and high power supply rejection ratio (Chandrakumar and Markovic, 2017a, 2017b; Mahajan et al., 2015; Muller et al., 2014). Therefore, we assumed an NEF of 4.0. For the robust detection of neural spikes whose amplitude ranges around 100μV, 2μV input referred noise is assumed. The signal bandwidth was set to 300-1,000Hz for the low-bandwidth amplifier and

250-6,000Hz for the high-bandwidth amplifier for accurate spike detection (which is in the range of similar devices presented by Harrison *et al.*, Borton *et al.*, and Yin *et al.*) (Borton et al., 2013; Harrison et al., 2007, 2009; Yin et al., 2013). We chose a primary supply voltage of 3.3V for a high-resolution ADC (described later) and a matched last-stage amplifier to support high amplifier linearity, with an optional 1.2V supply for the amplifier chain given multiple supply domains. We computed the power consumption by multiplying the estimated current consumption by the voltage supply level and by 96 for the total quantity of channels. We estimated the power consumption of each 16-bit analog to digital converter using the Schreier Figure-of-Merit (FoM_S) formula, whose state-of-the-art envelope at a sampling rate less than 5MHz is 178dB (Murmann, 2020):

$$FoM_S = SNDR(dB) + 10 \log_{10} \left(\frac{BW}{P} \right)$$

We assumed a signal-to-noise-and-distortion-ratio (SNDR) of 96dB in anticipation of potential movement or electrical stimulation artifacts (which can be as large as 20mV (Young et al., 2018)), resulting in an effective number of bits of 15.6 to accommodate artifacts and neural recordings with the more typical low SNR units or the occasional high SNR units. Additionally, we assumed sampling frequencies of 2kHz for the low-bandwidth circuit or 20kHz for the high-bandwidth circuit and one converter for each of the 96 channels. The analog to digital converters' outputs were assumed to be transmitted in parallel to the digital circuitry.

We designed the digital circuits in Verilog HDL (IEEE Computer Society, 2001), synthesized using Synopsys Design Compiler Version L-2016.03-SP2 in TSMC 0.18 UM MIXED SIGNAL GENERAL PURPOSE II 1P6M/1P5M SALICIDE 1.8V/3.3V technology, and simulated using Cadence NCverilog simulator Version 15.20-s005, with power estimated using Synopsys PrimeTime PX Version M-2017.06. To represent the digital circuits' power

179

Table A.5 Breakdown of power estimation for each integrated circuit component.

Circuit Type	Analog (mW)		Digital (µW)			Total (mW)
	Amplifier	ADC	Feature Extractor	3-DoF SSKF	CAN	
	(3.3V, 1.2V)		(180nm, 40nm)			(3.3V, 1.2V)
Low-Bandwidth (SBP)	(0.64, 0.23)	0.42	(0.67, 0.22)	(0.64, 0.22)	0.032	(1.1, 0.65)
Low-Bandwidth (TCR)	(0.64, 0.23)	0.42	(0.14, 0.05)	(0.64, 0.22)	0.032	(1.1, 0.65)
High-Bandwidth (TCR)	(5.2, 1.9)	6.1	(0.89, 0.30)	(0.64, 0.22)	0.032	(11, 8.0)

(3.3V, 1.2V) indicates the power consumptions of the amplifiers calculated at a 3.3V supply or a 1.2V supply. (180nm, 40nm) indicates the digital circuit power consumption when simulated at a 180nm or a 40nm process node. TCR is threshold crossing rate, ADC is analog-to-digital converter, DoF is degrees of freedom, SSKF is steady-state Kalman filter, CAN is controller area network.

consumptions in a more advanced process node, we have additionally performed all digital simulations in 40nm technology. It should be noted that the analog components' noise constraints result in larger than minimum size analog transistors, even using the older 180nm technology. This means that the analog domain will not gain any power advantage by using a smaller process size. Figure A.9 contains diagrams illustrating the components of our circuit designs. The input vectors of the gate-level simulations were randomized 2kSps neural signals for the low-bandwidth circuit and randomized 20kSps neural signals for the high-bandwidth circuit. For relevance to this manuscript and the devices presented in literature, we simulated two low-bandwidth circuits, one that extracted SBP as the neural feature and the other that extracted low-bandwidth TCR, while the high-bandwidth circuit extracted TCR as the neural feature. The low-bandwidth SBP circuit was composed of 96 copies of 16-bit registers, a 16-bit adder, and 96 copies of average generators. The TCR circuits were composed of 96 copies of 16-bit registers, a threshold comparator, and 96 copies of 10-bit counters for the high-bandwidth circuit or 7-bit counters for the low-bandwidth circuit. We estimated the power consumptions for the Kalman filter and CAN blocks using the same simulation software. For all digital logic simulated, we used a near-threshold supply voltage

180

of 0.6V, which we found to be low power while adequate for the performance requirement and logic depth.

Our simulations and calculations demonstrated that the overall power consumption of the high-bandwidth circuit is dominated by the front-end, needing 11mW for a 3.3V amplifier supply or 8.0mW for a 1.2V amplifier supply and 1.6μW for the digital back-end components (in 180nm, 0.55μW in 40nm). When we reduced the front-end bandwidth to the spiking band, the simulation predicted a drastic cut in the power consumption to 1.1mW for a 3.3V amplifier supply or 0.65mW for a 1.2V amplifier supply, while the consumption of the digital components remained low at 1.3μW (in 180nm, 0.47μW in 40nm). The power consumptions of the individual circuit components are listed in Table A.5. The results of these simulations suggest that, regardless of the process node, the power consumption of integrated neural devices is dominated by the analog front-end, which can be cut by at least 90% when restricted to the 300-1,000Hz band, despite maintaining a power efficiency factor (PEF) near the state-of-the-art at 19.2 with a 1.2V supply (Ng and Xu, 2016). However, TCR can accurately predict firing patterns at much narrower bandwidths, as demonstrated in Figure 2.2. Simulating the low-bandwidth TCR circuit revealed that it can further reduce the power consumption of the digital components to a total of 0.81μW (in 180nm, 0.30μW in 40nm). While extracting low-bandwidth TCR using the spiking band would result in an equivalent noise spectral density as SBP and save equivalent amounts of amplifier power and more digital power than the low-bandwidth SBP circuit, the analog front-end remains the dominating power-hungry component. Thus, since the decoding results in Figure A.5 - Figure A.8 suggest that SBP persistently achieves higher decoding performances than low-bandwidth TCR and low-bandwidth TCR enables a small 0.05% power savings over SBP, we conclude that SBP would remain the highest performing neural feature for low-power decoding in all cases.

Appendix B Supplement to Chapter 4
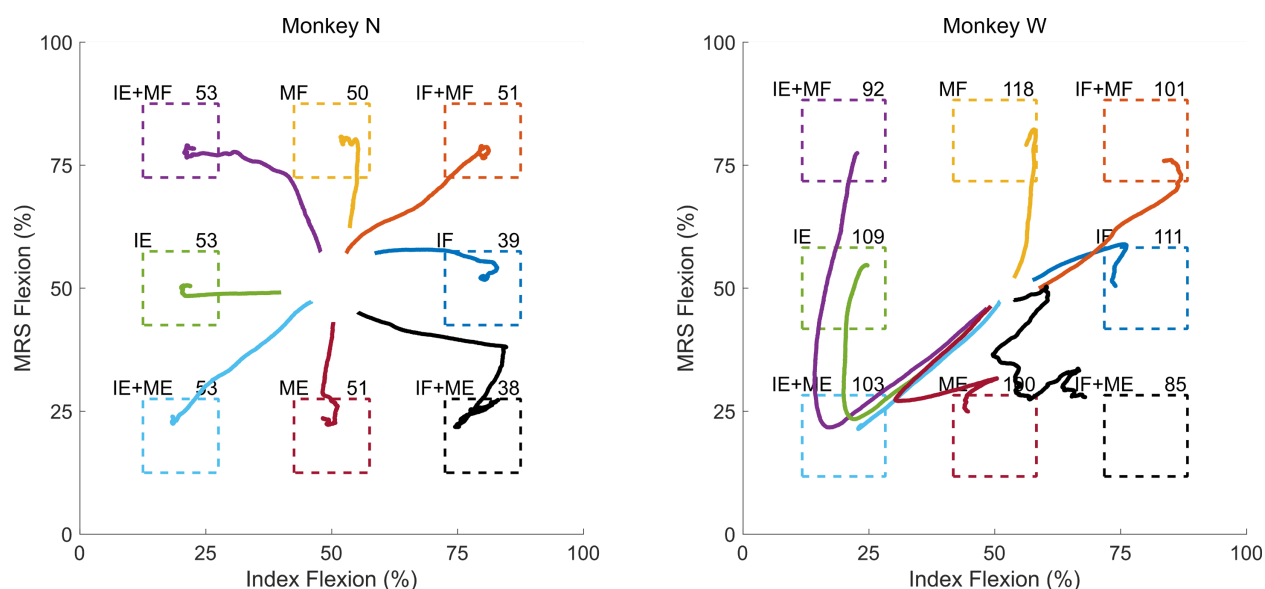
B.1 Supplementary Figures and Tables
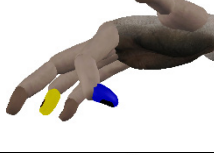
Figure B.1 Averaged example behaviors for each monkey (N left, W right), related to 4.2 Methods. Each two-dimensional trace is plotted as MRS flexion percentage versus index flexion percentage, in the same space as Figure 4.1B. Targets are the dashed boxes, color coordinated to the traces corresponding to attempts to acquire that target. The letters to the top left of each target detail the movement type in the two-dimensional space, and the numbers to the top right indicate the number of trials used to obtain each averaged trace. The behaviors to generate this figure were from the same data sets upon which the tuning analyses were performed. Note that most trajectories are directed towards the targets as if this were one two-dimensional task, but the IF+ME and IE+MF trajectories for both monkeys and the extension trajectories for monkey W suggest the monkeys may have viewed the task as two one-dimensional tasks.

Table B.1 Illustrations of all center-out postures, ordered according to Figure 4.1.

Posture	+20% from Rest	+30% from Rest	+40% from Rest	Relevant Analyses
Index Flexion (IF)				Online Tuning

Index Flexion + MRS Flexion (IF+MF)				Online Tuning
MRS Flexion (MF)				Online Tuning
Index Extension + MRS Flexion (IE+MF)			N/A (split too far for reliable performance)	Tuning
Index Extension (IE)				Online Tuning
Index Extension + MRS Extension (IE+ME)				Online Tuning
MRS Extension (ME)				Online Tuning
Index Flexion + MRS Extension (IF+ME)			N/A (split too far for reliable performance)	Tuning
Rest (R)		N/A	N/A	Online Tuning

"Online" refers to use in online brain-machine interface experiments (Figure 4.2, Figure 4.3, Figure 4.4). "Tuning refers to use in neural tuning analyses (Figure 4.5, Figure 4.6, Figure 4.7).

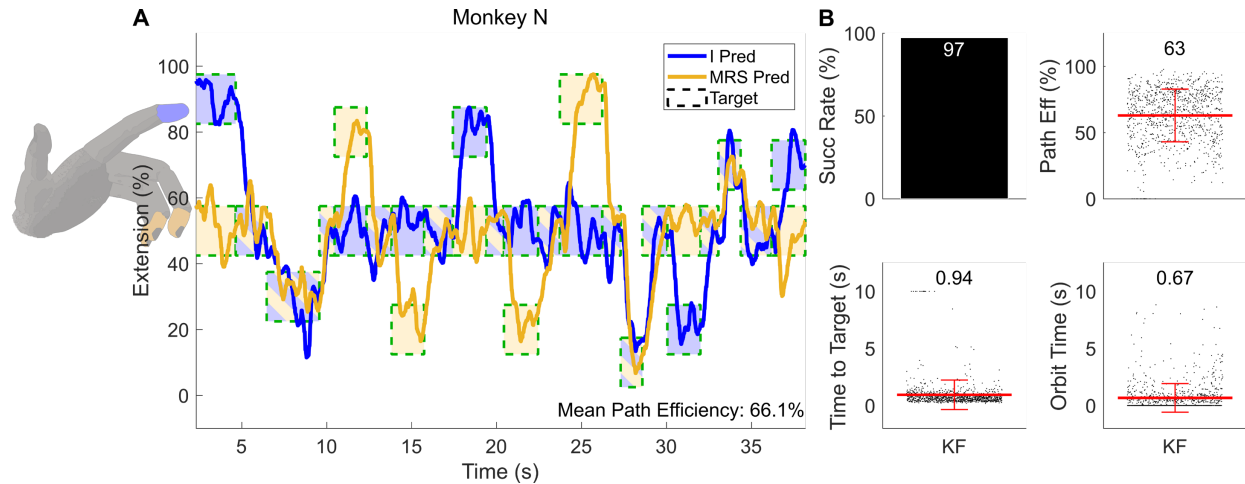**Figure B.2 Two-finger closed-loop Kalman filter decode using threshold crossing rates, related to Figure 4.2.** (**A**) Example closed-loop prediction traces from monkey N using the standard Kalman filter. Targets are represented by the dashed boxes, internally colored to indicate the targeted finger with a border color representing whether the trial was acquired successfully. "I" means the index finger group and "MRS" means the middle/ring/small finger group. The mean path efficiency of the trials displayed is presented at the bottom right. (**B**) Statistics for all closed-loop two-finger threshold crossing rate Kalman filter decodes. The red lines indicate the means, which are numerically displayed above each set of data. The statistic for each trial is represented by one dot in each plot. "Succ Rate" means the percentage of total trials that were successfully acquired in time and "Path Eff" means the two-dimensional path efficiency.

Table B.2 Summary of neural feature tunings to flexion versus extension, related to 4.2 Methods.

Monkey	Feature	Qty	Muscular Group				Untuned
			Flexion	Extension	Opposed	Multiple	
N	SU	47	4	2	12	22	7
	SBP	96	2	12	33	33	16
W	SU	68	36	0	5	18	9
	SBP	96	32	1	2	26	35

The quantity (Qty) column indicates the total number of sources for each feature. SU means sorted units, SBP means spiking band power. Classifications were determined by $p < 0.001$ with a two-tailed two-sample Kolmogorov-Smirnov test between movement types. Opposed tunings showed activity levels different from baseline during movements where the two finger groups moved in opposite directions (IF+ME or IE+MF). Multiple tunings showed activity levels different from baseline for more than half but not all of the movements, indicating activation but perhaps not specificity. Decisions were made from 388 monkey N trials and 819 monkey W trials.

Table B.3 Summary of neural feature tunings to index versus MRS group movements, related to 4.2 Methods.

Monkey	Feature	Qty	Finger Group				Untuned
			Index	MRS	Both	Multiple	
N	SU	47	2	2	14	22	7
	SBP	96	0	6	41	33	16
W	SU	68	10	13	18	18	9
	SBP	96	22	6	7	26	35

The quantity (Qty) column indicates the total number of sources for each feature. SU means sorted units, SBP means spiking band power. Classifications were determined by $p < 0.001$ with a two-tailed two-sample Kolmogorov-Smirnov test between movement types. Here, multiple tunings showed activity levels different from baseline for more than half but not all of the movements, indicating activation but perhaps not specificity. Decisions were made from 388 monkey N trials and 819 monkey W trials.
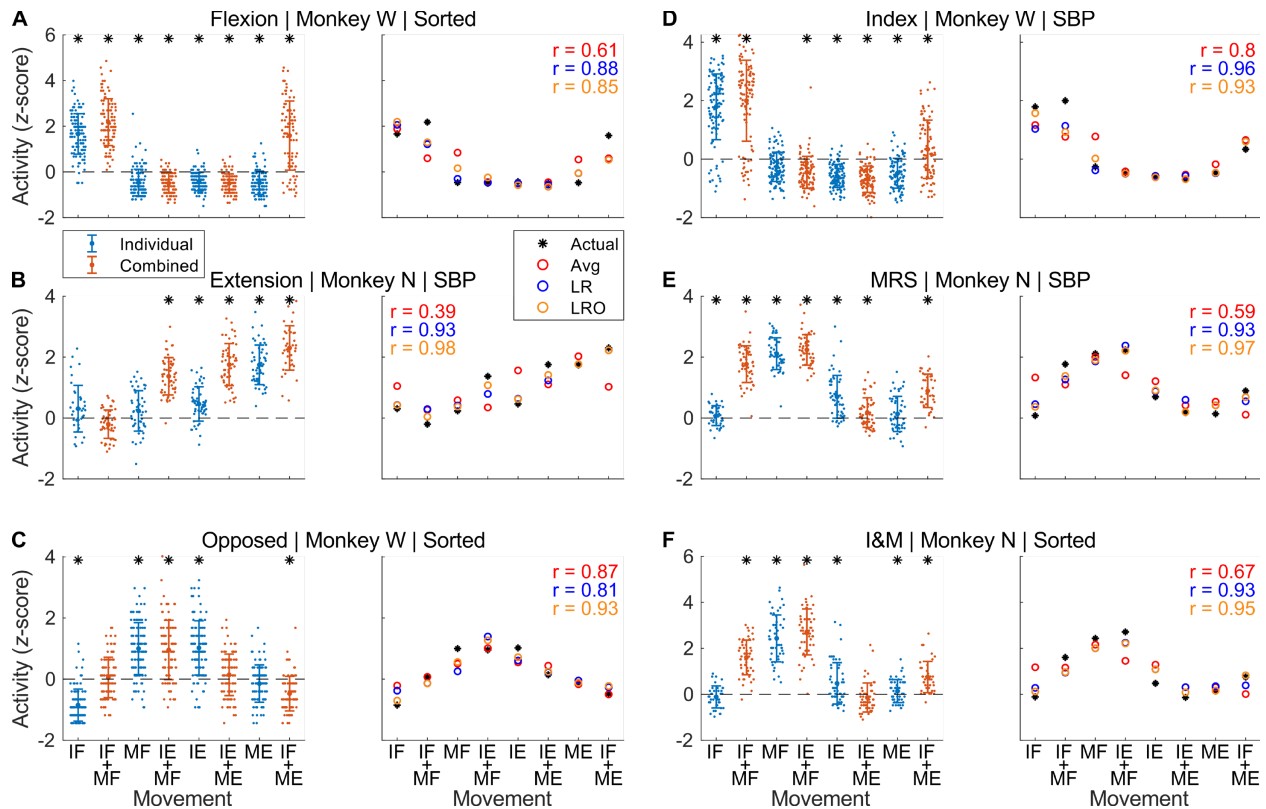
Figure B.3 Exemplary tuning curves and linear predictability of the activity, related to Figure 4.5. The text above each pair of plots first indicates the tuning preference of a sorted unit or SBP channel, followed by the subject from which the activity was recorded, then the feature type. I&M suggests preference to both the index and MRS groups. The left plot of each pair displays the true tuning curve of the given sorted unit or SBP channel. The large dot and error bars represent the mean and standard deviation for the peri-movement activities across all trials of each type of movement. The smaller dots represent the peri-movement activity for each trial of each type of movement. Asterisks indicate significant difference from the unit's or channel's mean activity, or zero z-score, across the experiment (two-sided two-sample Kolmogorov-Smirnov test, $p < 0.001$, corrected for false discovery rate). Individual group movements are plotted in blue, combined group movements are plotted in orange. The right plot of each pair displays the mean activity for each movement (copied from the left plot in asterisks) as well as the predictions of the activity for each movement. For example, for IF in **C**, the black asterisk is the mean dot copied from the left hand plot, the red circle (Avg) is the average of the asterisks for IF+MF and IF+ME, the blue circle (LR) is the weighted sum of the asterisks for IF+MF and IF+ME, and the yellow circle (LRO) is the weighted sums of the asterisks for IF+MF, IF+ME, and IE. Pearson's correlation coefficients between the predictions and the true activity are displayed in each plot.
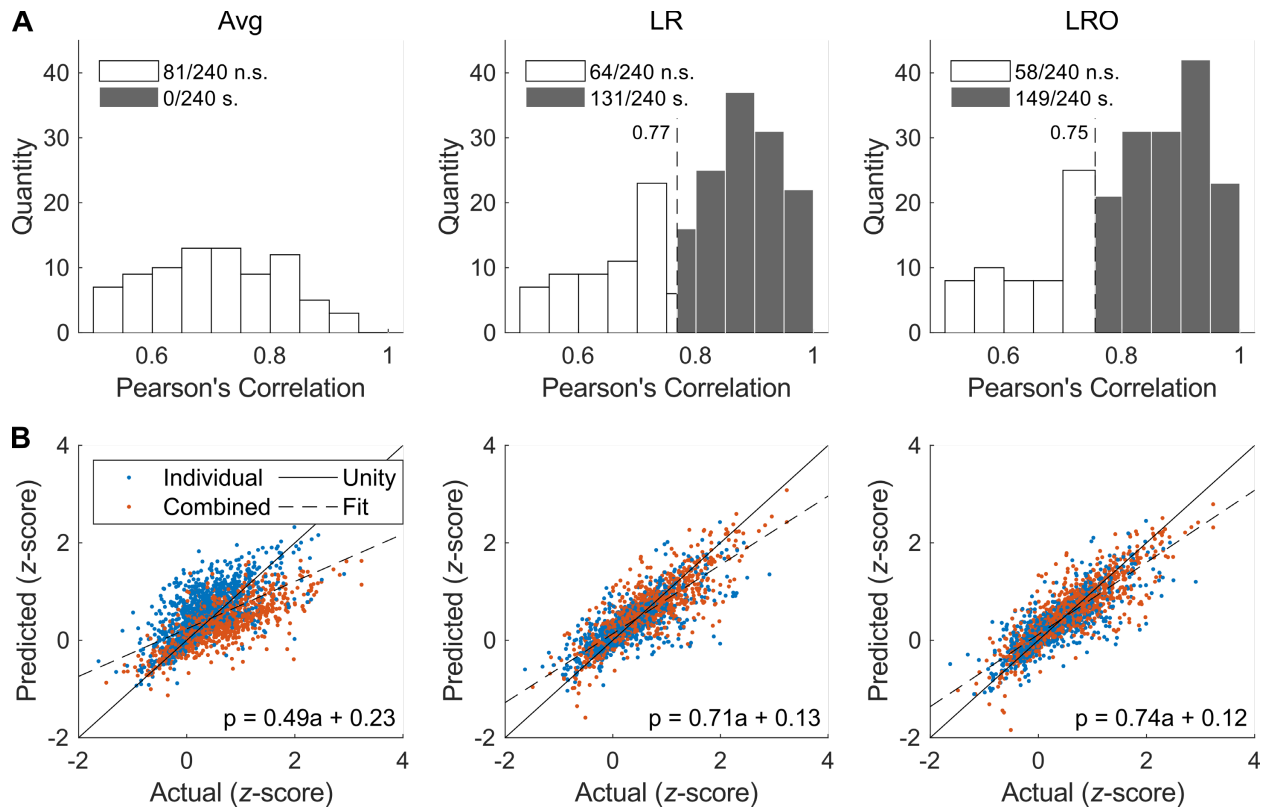
Figure B.4 Statistics of the linearity in tuning curves, related to Figure 4.5. Sorted units and SBP channels from monkeys N and W that had at least one movement significantly different from the average activity and an average firing rate greater than 2Hz (for sorted units) are included in analysis. The first column represents results from averaging the neighbors, and the second and third columns represent results from linearly regressing the neighbors without and with the opposite movement, respectively. (**A**) Histogram of the correlation coefficients (Pearson's *r*) between the predicted tuning curves and the actual tuning curves. The histogram is limited to the 0.5-1 range as we found this range of tuning curves to be visually linear. The dotted line with the attached number indicates the cutoff for significance after correction for false discovery rate. n.s. means not significant, s. means significant with $p < 0.05$. Significance was determined based on the null hypothesis that a particular predicted tuning curve did not significantly correlate with the true tuning curve. (**B**) Scatter plots of the predicted activity vs. the true activity for each movement. Blue dots represent individual group movements and orange dots represent combined group movements. The solid line is unity, where the predicted activity would equal the true activity. The dashed line is fit to the data.
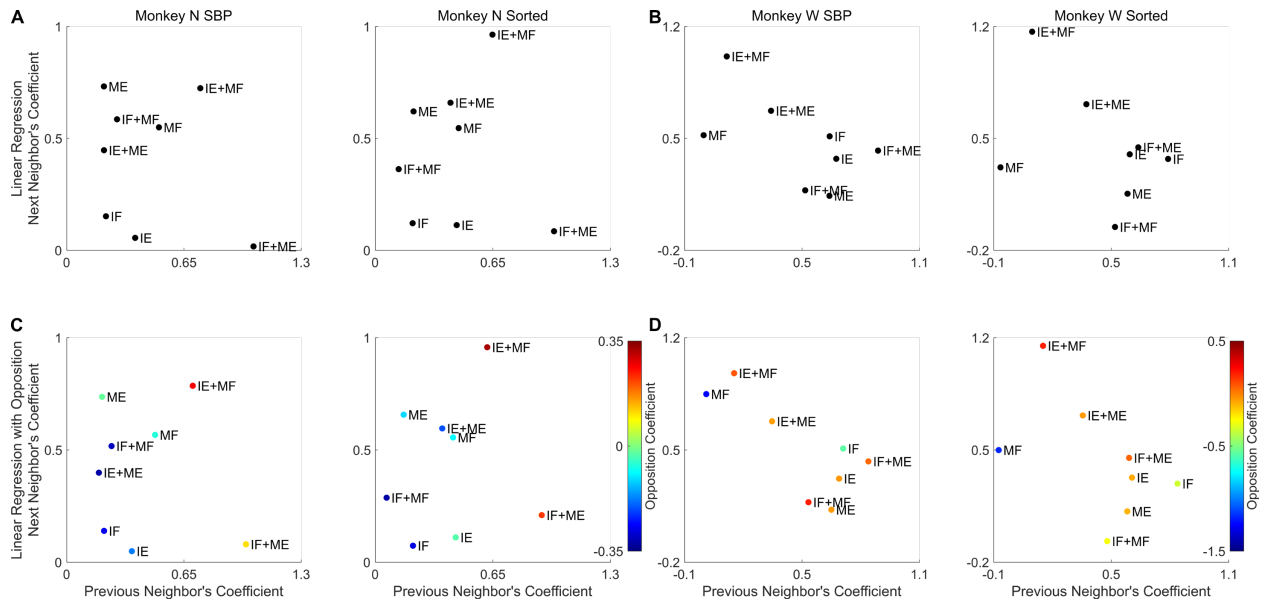
Figure B.5 Regression coefficients to predict the neural activity associated with certain movements from the activity of others, related to Figure 4.5. Each point is labelled with the movement whose activity would be predicted by the activity of other movements with (x, y) position indicating the trained coefficients of neighboring movements and color indicating the coefficient of the opposition movement. I – index finger group, M – MRS finger group, F – flexion, E – extension. (**A**, **B**) Coefficients for regressions under the assumption that the activity of a movement is a weighted sum of the activities of its neighbors for monkeys N and W, respectively. (**C**, **D**) Coefficients for regressions under the assumption that the activity of a movement is a weighted sum of the activities of its neighbors and the opposite movement for monkeys N and W, respectively. Points are colored based on the best-fit coefficient by which to multiply the opposite movement.

Figure B.6 Offline ridge regression decoding for sorted units, trained on either individual or combined finger group movements, related to Figure 4.6. The central trace is the average predicted behavior from all trials of the indicated movement and the shaded region is the standard deviation, all aligned in time by the movement onset (vertical gray line). Individual finger group movements (the two left columns of plots) were decoded using a regression model trained on combined finger group movements (the right two columns of plots), and vice versa for the combined finger group movements. These are represented by the "Split Decode" dashed traces. The "Full Decode" solid line traces represent the average decode given the full dataset to train the regression model, with cross-validation. The blue traces correspond to the index group and the yellow traces correspond to the MRS group. The yellow or blue lines near the top of each plot indicate significant differences between the two predicted positions based on a bootstrap analysis on the differences (greater than a one-sided 95% confidence interval).

189
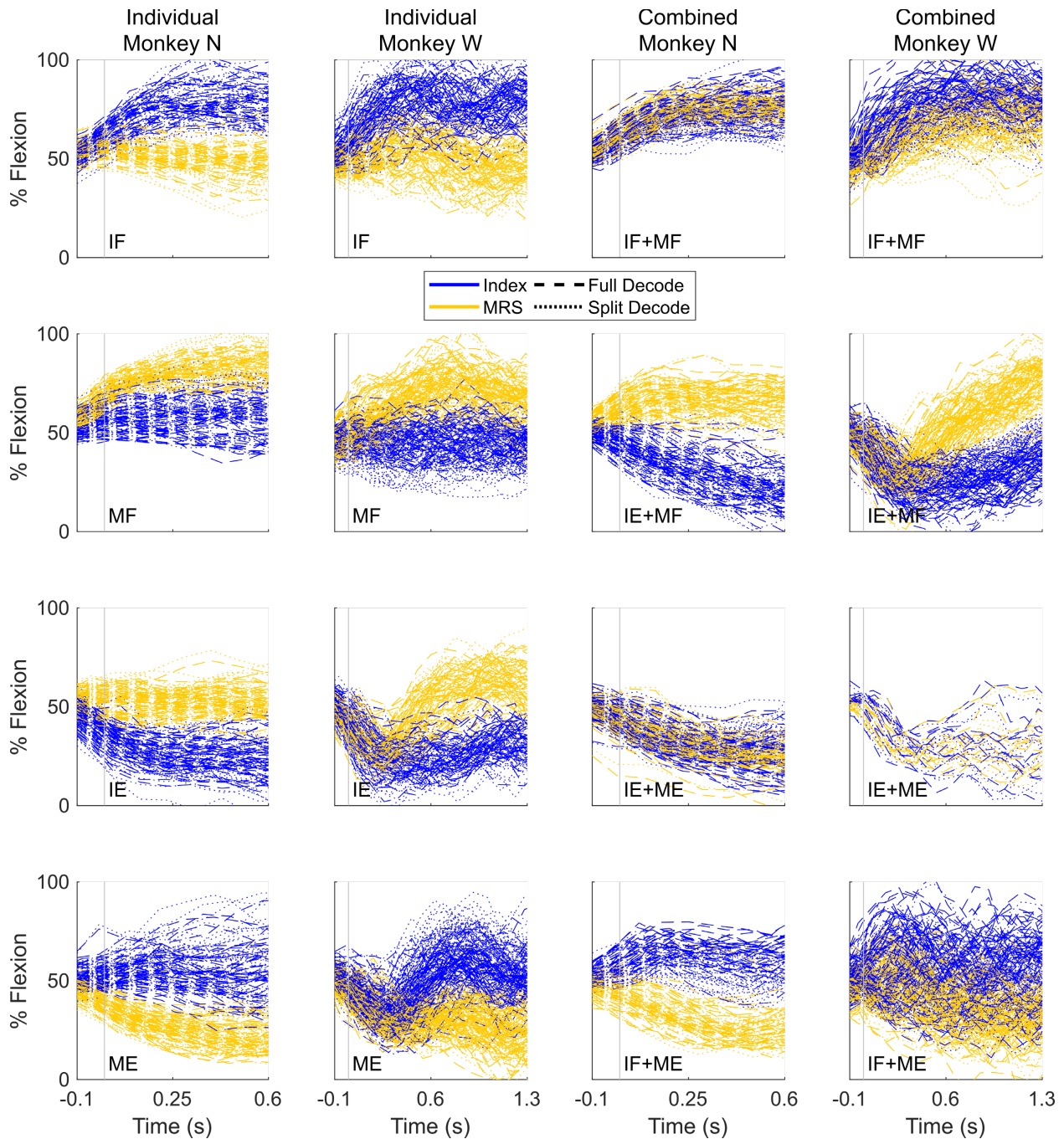
Figure B.7 Individual trial offline ridge regression decoding of all SBP channels, trained on either individual or combined finger group movements, related to Figure 4.6. The vertical gray line indicates movement onset. Individual finger group movements (the two left columns of plots) were decoded using a regression model trained on combined finger group movements (the two right columns of plots), and vice versa for the combined finger group movements. These are represented by the "Split Decode" dotted traces. The "Full Decode" dashed traces represent the average decode given the full dataset to train the regression model, with cross-validation. The blue traces correspond to the index group and the yellow traces correspond to the MRS group.
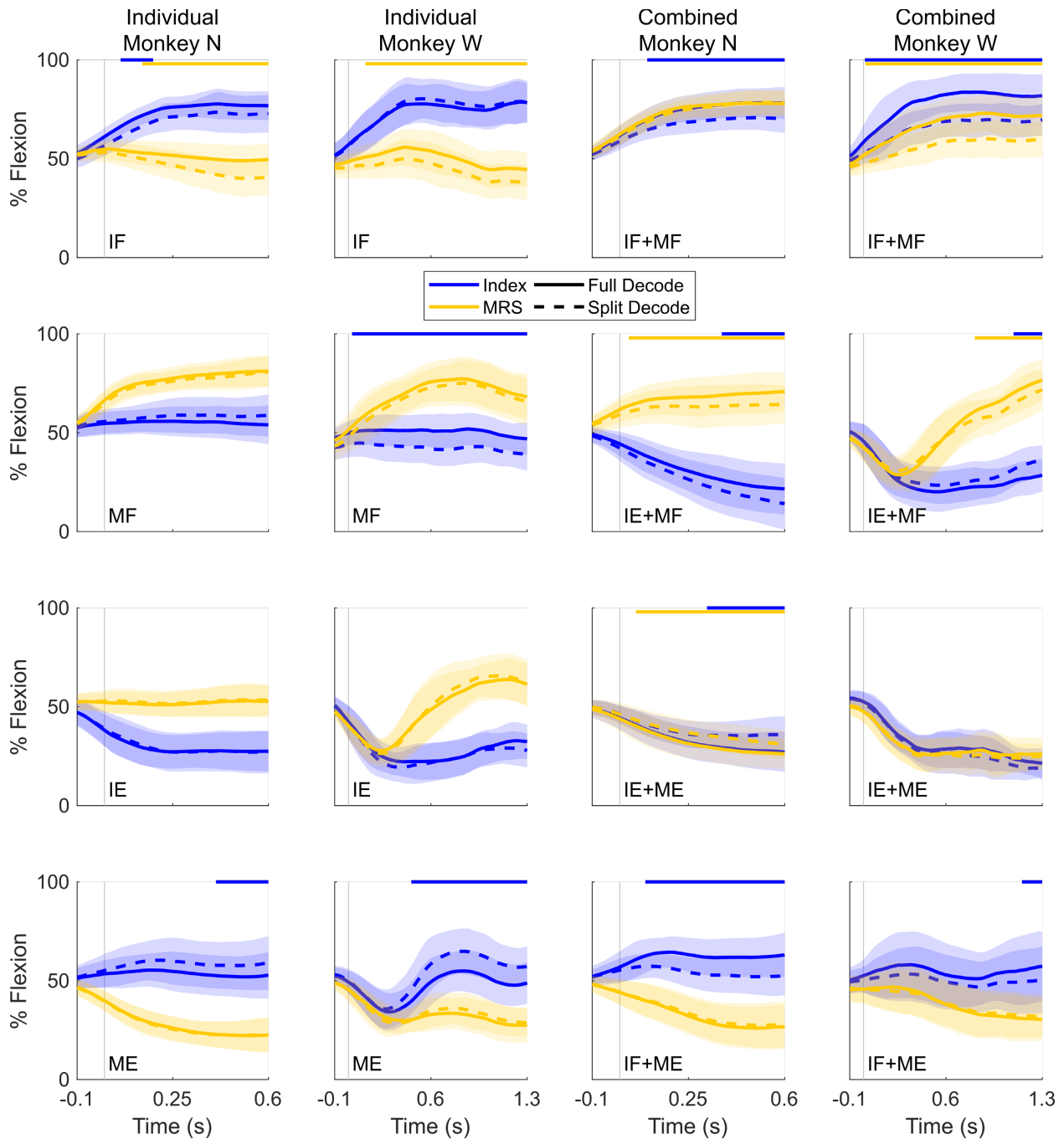
Figure B.8 Individual trial offline ridge regression decoding of all sorted units, trained on either individual or combined finger group movements, related to Figure 4.6. The vertical gray line indicates movement onset. Individual finger group movements (the two left columns of plots) were decoded using a regression model trained on combined finger group movements (the two right columns of plots), and vice versa for the combined finger group movements. These are represented by the "Split Decode" dotted traces. The "Full Decode" dashed traces represent the average decode given the full dataset to train the regression model, with cross-validation. The blue traces correspond to the index group and the yellow traces correspond to the MRS group.

Appendix C Supplement to Chapter 5

C.1 Supplementary Methods

C.1.1 Cortical Implants

To perform brain-controlled functional electrical stimulation for the restoration of continuous hand function, we implanted Monkey N with three Utah microelectrode arrays in the hand areas of primary motor and sensory cortices. Figure C.1 presents a photograph of the implants.

C.1.2 Ultrasound Examples

To guarantee that the monkey's capable movements were purely a result of the BCFES system and not the monkey's native efforts, we delivered nerve blocks to the median, radial, and
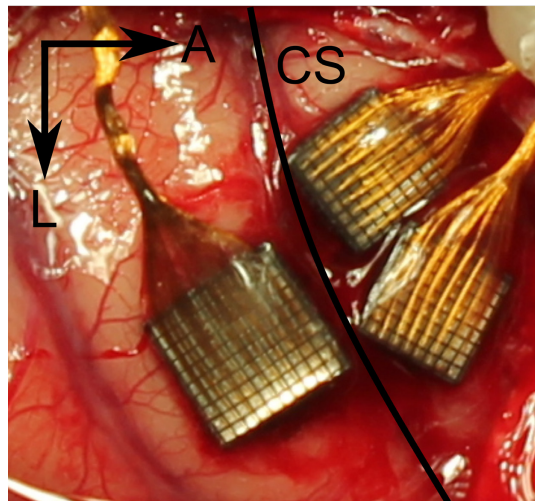
Figure C.1 Photograph of Monkey N's right-hemisphere Utah microelectrode array implant locations. The posterior array is a 10x10 array with 96 active electrodes implanted in the hand area of primary sensory cortex. The two anterior arrays are each 8x8 arrays with 64 active electrodes each implanted in the hand area of primary motor cortex.

Median	Radial	Ulnar

Figure C.2 Example ultrasound images taken during a nerve block procedure. All images were captured along a transverse plane just proximal to the elbow.

ulnar nerves of his left arm to temporarily block signal transduction to the arm. We used ultrasound to deliver a 2% lidocaine with 1:100,000 epinephrine to the tissue surrounding the nerves to block activity. Figure C.2 displays some example images captured from one experiment.

C.1.3 Functional Electrical Stimulation Pattern

To deliver electrical stimulation in a low-bandwidth communication space, we used stimulation patterns to govern parameters for all electrodes innervating the hand. The controller commanded a value on the horizontal axis in Figure C.3, and the Networked Neuroprosthesis commanded the corresponding pulse widths for each of the active electrodes.

C.2 Supplementary Analysis

C.2.1 Brain-Controlled FES Performance May Be Improved with Lower Latency

Muscle fatigue and hysteresis may increase the difficulty of acquiring targets, particularly with simple controllers like the one investigated in Figure 5.2a. In the case of any feedback controller, latency in the feedback loop can also increase difficulty by introducing ringing (or orbiting as we refer to it in the context of target acquisition) without proper dampening. We next sought to investigate the impact of controller latency on Monkey N's capability of controlling the virtual hand using able-bodied manipulandum control and the BCFES system.
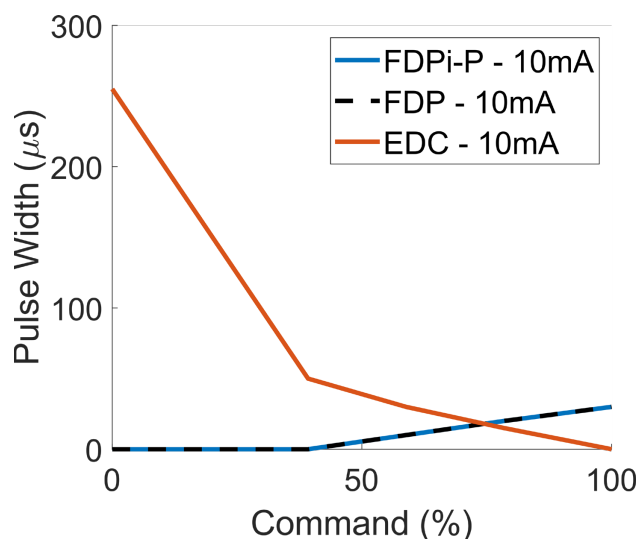
Figure C.3 An example stimulation pattern used during a brain-controlled functional electrical stimulation experiment with Monkey N.

First, we estimated the relative latencies of our various control methods for the 1D task. To do so, we offset the measured or predicted behavior and SBP from -200ms (behavior precedes cortical activity by 200ms) to +400ms (cortical activity precedes behavior by 400ms) in 1ms steps and used ridge regression ($\lambda = 0.0001$) to predict the movement velocities at each offset. The optimal latency for each control mode was calculated as the maximum intra-class correlation across all offsets. We compared latencies of three implants in three cortical hemispheres in two monkeys, data from Monkeys N and W in our past publication (Nason et al., 2020) and Monkey N data from this publication.

In all cases, we found reasonable latencies between cortical SBP and behavioral outputs on the order of 16 to 37ms (25th and 75th percentiles, respectively, $n = 26$) in manipulandum control mode. Note these estimates do not take into consideration the processing and communication latencies of the BMI hardware, so are not necessarily representative of true spinal cord transduction speeds but rather can be compared relatively. Figure C.4a presents histograms illustrating the optimal latencies of all datasets analyzed for each implant. Data from Monkeys N and W from our
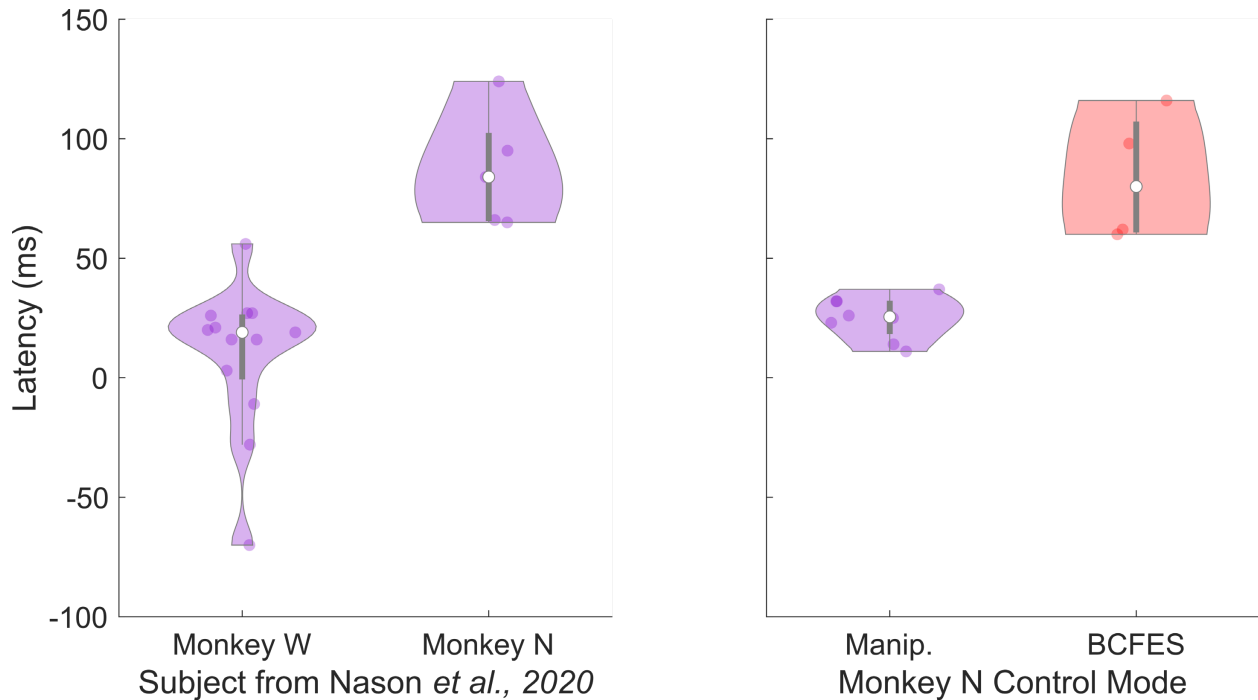
Figure C.4 Latency of BCFES compared to manipulandum control. Violin plots illustrate optimal latencies for various implants in various control settings. Each colored dot represents one day of recording. The white dot represents the median and the gray bar represents the 25th and 75th percentiles. (*Left*) Optimal manipulandum control latencies for Monkeys W and N from Nason *et al.*, 2020. (*Right*) Optimal latencies for Monkey N using the manipulandum and the BCFES system.

past work show similar latencies between 16 and 65ms (25th and 75th percentiles, respectively, $n = 18$), though Monkey N's RKF latencies were higher likely due to the more anterior placement of the recording array. Monkey N's latencies from this work were very consistently between 18.5-32ms (25th and 75th percentiles, respectively, $n = 8$). At their optimal latencies, we found that the peak velocity intra-class correlations were reasonably high for ridge regressions at 0.18 to 0.39 (25th and 75th percentiles, respectively, $n = 26$).

When Monkey N used the BCFES system, we found that the optimal latency between cortical activity and the resulting FES-controlled behavior was substantially higher at 61 to 107ms (25th and 75th percentiles, respectively, $n = 4$). The substantial increase in latency compared to manipulandum control is likely due to the wireless transmission of stimulation commands, real-

time computation of stimulation updates in the embedded hardware, and generation of a contraction response in the target musculature.

Bibliography

Abdelhalim, K., Kokarovtseva, L., Perez Velazquez, J.L., and Genov, R. (2013). 915-MHz FSK/OOK Wireless Neural Recording SoC With 64 Mixed-Signal FIR Filters. IEEE J. Solid-State Circuits *48*, 2478–2493.

Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., Shanfield, K., Hayes-jackson, S., Aisen, M., Heck, C., et al. (2015). Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. Science (80-.). *348*, 906–910.

Aggarwal, V., Acharya, S., Tenore, F., Shin, H.C., Etienne-Cummings, R., Schieber, M.H., and Thakor, N. V. (2008). Asynchronous decoding of dexterous finger movements using M1 neurons. IEEE Trans. Neural Syst. Rehabil. Eng. *16*, 3–14.

Aggarwal, V., Mollazadeh, M., Davidson, A.G., Schieber, M.H., and Thakor, N. V. (2013). State-based decoding of hand and finger kinematics using neuronal ensemble and LFP activity during dexterous reach-to-grasp movements. J. Neurophysiol. *109*, 3067–3081.

Ajiboye, A.B., Willett, F.R., Young, D.R., Memberg, W.D., Murphy, B.A., Miller, J.P., Walter, B.L., Sweet, J.A., Hoyen, H.A., Keith, M.W., et al. (2017). Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. Lancet *389*, 1821–1830.

Anderson, K.D. (2004). Targeting recovery: Priorities of the spinal cord-injured population. J. Neurotrauma *21*, 1371–1383.

Anumanchipalli, G.K., Chartier, J., and Chang, E.F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature *568*, 493–498.

Aziz, J.N.Y., Abdelhalim, K., Shulyzki, R., Genov, R., Bardakjian, B.L., Derchansky, M., Serletis, D., and Carlen, P.L. (2009). 256-channel neural recording and delta compression microsystem with 3D electrodes. IEEE J. Solid-State Circuits *44*, 995–1005.

Baker, J., Bishop, W., Kellis, S., Levy, T., House, P., and Greger, B. (2009). Multi-scale recordings for neuroprosthetic control of finger movements. Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. Eng. Futur. Biomed. EMBC 2009 4573–4577.

Bansal, A.K., Vargas-Irwin, C.E., Truccolo, W., and Donoghue, J.P. (2011). Relationships among low-frequency local field potentials, spiking activity, and three-dimensional reach and grasp kinematics in primary motor and ventral premotor cortices. J. Neurophysiol. *105*, 1603–1619.

Barrese, J.C., Rao, N., Paroo, K., Triebwasser, C., Vargas-Irwin, C., Franquemont, L., and Donoghue, J.P. (2013). Failure mode analysis of silicon-based intracortical microelectrode arrays in non-human primates. J. Neural Eng. *10*.

Benabid, A.L., Chabardes, S., Mitrofanis, J., and Pollak, P. (2009). Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease. Lancet Neurol. *8*, 67–81.

Biederman, W., Yeager, D.J., Narevsky, N., Koralek, A.C., Carmena, J.M., Alon, E., and Rabaey, J.M. (2013). A Fully-Integrated, Miniaturized (0.125 mm^2) 10.5 µW Wireless Neural Sensor. IEEE J. Solid-State Circuits *48*, 960–970.

Blabe, C.H., Gilja, V., Chestek, C.A., Shenoy, K. V, Anderson, K.D., and Henderson, J.M. (2015). Assessment of brain–machine interfaces from the perspective of people with paralysis. J. Neural Eng. *12*, 9.

Borna, A., and Najafi, K. (2014). A low power light weight wireless multichannel microsystem for reliable neural recording. IEEE J. Solid-State Circuits *49*, 439–451.

Borton, D.A., Yin, M., Aceros, J., and Nurmikko, A. (2013). An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. J. Neural Eng. *10*.

Bouton, C.E., Shaikhouni, A., Annetta, N. V., Bockbrader, M.A., Friedenberg, D.A., Nielson, D.M., Sharma, G., Sederberg, P.B., Glenn, B.C., Mysiw, W.J., et al. (2016). Restoring cortical control of functional movement in a human with quadriplegia. Nature *533*, 247–250.

Boyden, E.S., Zhang, F., Bamberg, E., Nagel, G., and Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. Nat. Neurosci. *8*, 1263–1268.

Bullard, A.J., Nason, S.R., Irwin, Z.T., Nu, C.S., Smith, B., Campean, A., Peckham, P.H., Kilgore, K.L., Willsey, M.S., Patil, P.G., et al. (2019). Design and testing of a 96-channel neural interface module for the Networked Neuroprosthesis system. Bioelectron. Med. *5*, 1–14.

Bullard, A.J., Hutchison, B.C., Lee, J., Chestek, C.A., and Patil, P.G. (2020). Estimating Risk for Future Intracranial, Fully Implanted, Modular Neuroprosthetic Systems : A Systematic Review of Hardware Complications in Clinical Deep Brain Stimulation and Experimental Human Intracortical Arrays. Neuromodulation Technol. Neural Interface *23*, 411–426.

Byun, D., Cho, S.J., Lee, B.H., Min, J., Lee, J.H., and Kim, S. (2017). Recording nerve signals in canine sciatic nerves with a flexible penetrating microelectrode array. J. Neural Eng. *14*.

Canales, A., Jia, X., Froriep, U.P., Koppes, R.A., Tringides, C.M., Selvidge, J., Lu, C., Hou, C., Wei, L., Fink, Y., et al. (2015). Multifunctional fibers for simultaneous optical, electrical and chemical interrogation of neural circuits in vivo. Nat. Biotechnol. *33*, 277–284.

Capogrosso, M., Milekovic, T., Borton, D., Wagner, F., Moraud, E.M., Mignardot, J.B., Buse, N., Gandar, J., Barraud, Q., Xing, D., et al. (2016). A brain-spine interface alleviating gait deficits

after spinal cord injury in primates. Nature *539*, 284–288.

Carmena, J.M., Lebedev, M.A., Crist, R.E., O'Doherty, J.E., Santucci, D.M., Dimitrov, D.F., Patil, P.G., Henriquez, C.S., and Nicolelis, M.A.L. (2003). Learning to control a brain-machine interface for reaching and grasping by primates. PLoS Biol. *1*, 193–208.

Chae, M., Liu, W., Yang, Z., Chen, T., Kim, J., Sivaprakasam, M., and Yuce, M. (2008). A 128-channel 6mW wireless neural recording IC with on-the-fly spike sorting and UWB Tansmitter. IEEE Int. Solid-State Circuits Conf. *51*, 146–148.

Chandrakumar, H., and Markovic, D. (2017a). A High Dynamic-Range Neural Recording Chopper Amplifier for Simultaneous Neural Recording and Stimulation. IEEE J. Solid-State Circuits *52*, 645–656.

Chandrakumar, H., and Markovic, D. (2017b). An 80-mVpp linear-input range, 1.6-G Ω input impedance, low-power chopper amplifier for closed-loop neural recording that is tolerant to 650-mVpp common-mode interference. IEEE J. Solid-State Circuits *52*, 2811–2828.

Chao, Z.C., Nagasaka, Y., and Fujii, N. (2010). Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Front. Neuroeng. *3*.

Chapin, J.K., Moxon, K.A., Markowitz, R.S., and Nicolelis, M.A.L. (1999). Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. Nat. Neurosci. *2*, 664–670.

Chase, S.M., Schwartz, A.B., and Kass, R.E. (2009). Bias, optimal linear estimation, and the differences between open-loop simulation and closed-loop performance of spiking-based brain-computer interface algorithms. Neural Networks *22*, 1203–1213.

Chen, Y.-P., Blaauw, D., and Sylvester, D. (2014). A 266nW Multi-Chopper Amplifier with 1.38 Noise Efficiency Factor for Neural Signal Recording. In Symposium on VLSI Circuits Digest of Technical Papers, (IEEE), p. 2.

Chestek, C.A., Batista, A.P., Santhanam, G., Yu, B.M., Afshar, A., Cunningham, J.P., Gilja, V., Ryu, S.I., Churchland, M.M., and Shenoy, K. V. (2007). Single-neuron stability during repeated reaching in macaque premotor cortex. J. Neurosci. *27*, 10742–10750.

Chestek, C.A., Gilja, V., Nuyujukian, P., Kier, R.J., Solzbacher, F., Ryu, S.I., Harrison, R.R., and Shenoy, K. V. (2009). HermesC: Low-power wireless neural recording system for freely moving primates. IEEE Trans. Neural Syst. Rehabil. Eng. *17*, 330–338.

Chestek, C.A., Gilja, V., Nuyujukian, P., Foster, J.D., Fan, J.M., Kaufman, M.T., Churchland, M.M., Rivera-Alvidrez, Z., Cunningham, J.P., Ryu, S.I., et al. (2011). Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. J. Neural Eng. *8*.

Chestek, C.A., Gilja, V., Blabe, C.H., Foster, B.L., Shenoy, K. V., Parvizi, J., and Henderson, J.M. (2013). Hand posture classification using electrocorticography signals in the gamma band over human sensorimotor brain areas. J. Neural Eng. *10*.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K. V (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

Cohen-Karni, T., Timko, B.P., Weiss, L.E., and Lieber, C.M. (2009). Flexible electrical recording from cells using nanowire transistor arrays. Proc. Natl. Acad. Sci. U. S. A. *106*, 7309–7313.

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J.C., Velliste, M., Boninger, M.L., and Schwartz, A.B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. Lancet *381*, 557–564.

Crago, P.E., Mortimer, J.T., and Peckham, P.H. (1980). Closed-Loop Control of Force During Electrical Stimulation of Muscle. IEEE Trans. Biomed. Eng. *BME-27*, 306–312.

Cross, K.P., Heming, E.A., Cook, D.J., and Scott, S.H. (2020). Maintained Representations of the Ipsilateral and Contralateral Limbs during Bimanual Control in Primary Motor Cortex. J. Neurosci. *40*, 6732–6747.

Cunningham, J.P., and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. *17*, 1500–1509.

Davoodi, R., Urata, C., Hauschild, M., Khachani, M., and Loeb, G.E. (2007). Model-based development of neural prostheses for movement. IEEE Trans. Biomed. Eng. *54*, 1909–1918.

Degenhart, A.D., Bishop, W.E., Oby, E.R., Tyler-Kabara, E.C., Chase, S.M., Batista, A.P., and Yu, B.M. (2020). Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. Nat. Biomed. Eng. *4*, 672–685.

Diedrichsen, J., Wiestler, T., and Krakauer, J.W. (2013). Two distinct ipsilateral cortical representations for individuated finger movements. Cereb. Cortex *23*, 1362–1377.

Donner, A. (1986). A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. Int. Stat. Rev. *54*, 67–82.

Downey, J.E., Schwed, N., Chase, S.M., Schwartz, A.B., and Collinger, J.L. (2018). Intracortical recording stability in human brain-computer interface users. J. Neural Eng. *15*.

Ethier, C., Oby, E.R., Bauman, M.J., and Miller, L.E. (2012). Restoration of grasp following paralysis through brain-controlled stimulation of muscles. Nature *485*, 368–371.

Even-Chen, N., Muratore, D.G., Stavisky, S.D., Hochberg, L.R., Henderson, J.M., Murmann, B., and Shenoy, K. V (2020). Power-saving design opportunities for wireless intracortical brain–computer interfaces. Nat. Biomed. Eng. *4*, 984–996.

Farwell, L.A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr. Clin. Neurophysiol. *70*, 510–523.

Flesher, S.N., Collinger, J.L., Foldes, S.T., Weiss, J.M., Downey, J.E., Tyler-Kabara, E.C., Bensmaia, S.J., Schwartz, A.B., Boninger, M.L., and Gaunt, R.A. (2016). Intracortical microstimulation of human somatosensory cortex. Sci. Transl. Med. *8*, 1–10.

Flesher, S.N., Downey, J.E., Weiss, J.M., Hughes, C.L., Herrera, A.J., Tyler-Kabara, E.C., Boninger, M.L., Collinger, J.L., and Gaunt, R.A. (2021). A brain-computer interface that evokes tactile sensations improves robotic arm control. Science (80-.). *372*, 831–836.

Flint, R.D., Lindberg, E.W., Jordan, L.R., Miller, L.E., and Slutzky, M.W. (2012a). Accurate decoding of reaching movements from field potentials in the absence of spikes. J. Neural Eng. *9*.

Flint, R.D., Ethier, C., Oby, E.R., Miller, L.E., and Slutzky, M.W. (2012b). Local field potentials allow accurate decoding of muscle activity. J. Neurophysiol. *108*, 18–24.

Flint, R.D., Wright, Z.A., Scheid, M.R., and Slutzky, M.W. (2013). Long term, stable brain machine interface performance using local field potentials and multiunit spikes. J. Neural Eng. *10*.

Flint, R.D., Rosenow, J.M., Tate, M.C., and Slutzky, M.W. (2017). Continuous decoding of human grasp kinematics using epidural and subdural signals. J. Neural Eng. *14*.

Fraser, G.W., Chase, S.M., Whitford, A., and Schwartz, A.B. (2009). Control of a brain-computer interface without spike sorting. J. Neural Eng. *6*.

Gallego, J.A., Perich, M.G., Chowdhury, R.H., Solla, S.A., and Miller, L.E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. Nat. Neurosci. *23*, 260–270.

Gao, H., Walker, R.M., Nuyujukian, P., Makinwa, K.A.A., Shenoy, K. V., Murmann, B., and Meng, T.H. (2012). HermesE: A 96-channel full data rate direct neural interface in 0.13 μm CMOS. IEEE J. Solid-State Circuits *47*, 1043–1055.

Georgopoulos, A.P., Kalaska, J.F., Caminiti, R., and Massey, J.T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. J. Neurosci. *2*, 1527–1537.

Georgopoulos, A.P., Kalaska, J.F., Crutcher, M.D., Caminiti, R., and Massey, J.T. (1984). The representation of movement direction in the motor cortex: single cell and population studies. Dyn. Asp. Neocortical Funct. 501–524.

Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1985). Neuronal Population Coding of Movement Direction. Science (80-.). *233*, 1416–1419.

Georgopoulos, A.P., Pellizzer, G., Poliakov, A. V., and Schieber, M.H. (1999). Neural coding of

finger and wrist movements. J. Comput. Neurosci. *6*, 279–288.

Gilja, V., Nuyujukian, P., Chestek, C.A., Cunningham, J.P., Yu, B.M., Fan, J.M., Churchland, M.M., Kaufman, M.T., Kao, J.C., Ryu, S.I., et al. (2012). A high-performance neural prosthesis enabled by control algorithm design. Nat. Neurosci. *15*, 1752–1757.

Gilja, V., Pandarinath, C., Blabe, C.H., Nuyujukian, P., Simeral, J.D., Sarma, A.A., Sorice, B.L., Perge, J.A., Jarosiewicz, B., Hochberg, L.R., et al. (2015). Clinical translation of a high-performance neural prosthesis. Nat. Med. *21*, 1142–1145.

Guitchounts, G., Markowitz, J.E., Liberti, W.A., and Gardner, T.J. (2013). A carbon-fiber electrode array for long-term neural recording. J. Neural Eng. *10*, 13.

Häger-Ross, C., and Schieber, M.H. (2000). Quantifying the independence of human finger movements: Comparisons of digits, hands, and movement frequencies. J. Neurosci. *20*, 8542–8550.

Harrison, R.R., and Charles, C. (2003). A low-power low-noise CMOS amplifier for neural recording applications. IEEE J. Solid-State Circuits *38*, 958–965.

Harrison, R.R., Watkins, P.T., Kier, R.J., Lovejoy, R.O., Black, D.J., Greger, B., and Solzbacher, F. (2007). A low-power integrated circuit for a wireless 100-electrode neural recording system. IEEE J. Solid-State Circuits *42*, 123–133.

Harrison, R.R., Kier, R.J., Chestek, C.A., Gilja, V., Nuyujukian, P., Ryu, S., Greger, B., Solzbacher, F., and Shenoy, K. V (2009). Wireless Neural Recording With Single Low-Power Integrated Circuit. IEEE Trans. Neural Syst. Rehabil. Eng. *17*, 322–329.

Heldman, D.A., Wang, W., Chan, S.S., and Moran, D.W. (2006). Local field potential spectral tuning in motor cortex during reaching. IEEE Trans. Neural Syst. Rehabil. Eng. *14*, 180–183.

Heming, E.A., Cross, K.P., Takei, T., Cook, D.J., and Scott, S.H. (2019). Independent representations of ipsilateral and contralateral limbs in primary motor cortex. Elife *8*, 1–26.

Henze, D.A., Borhegyi, Z., Csicsvari, J., Mamiya, A., Harris, K.D., and Buzsáki, G. (2000). Intracellular Features Predicted by Extracellular Recordings in the Hippocampus In Vivo. J. Neurophysiol. *84*, 390–400.

Hill, M., Rios, E., Sudhakar, S.K., Lempka, S.F., and Chestek, C.A. (2018). Quantitative simulation of extracellular single unit recording from the surface of cortex. J. Neural Eng. *15*, 12.

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., and Donoghue, J.P. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature *442*, 164–171.

Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S.,

Liu, J., Cash, S.S., Van Der Smagt, P., et al. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature *485*, 372–375.

Hodkin, E.F., Lei, Y., Humby, J., Glover, I.S., Choudhury, S., Kumar, H., Perez, M.A., Rodgers, H., and Jackson, A. (2018). Automated FES for Upper Limb Rehabilitation Following Stroke and Spinal Cord Injury. IEEE Trans. Neural Syst. Rehabil. Eng. *26*, 1067–1074.

Holt, G.R., and Koch, C. (1999). Electrical Interactions via the Extracellular Potential Near Cell Bodies. J. Comput. Neurosci. *6*, 169–184.

Hosman, T., Vilela, M., Milstein, D., Kelemen, J.N., Brandman, D.M., Hochberg, L.R., and Simeral, J.D. (2019). BCI decoder performance comparison of an LSTM recurrent neural network and a Kalman filter in retrospective simulation. Int. IEEE/EMBS Conf. Neural Eng. NER *2019-March*, 1066–1071.

Hotson, G., Mcmullen, D.P., Fifer, M.S., Johannes, M.S., Katyal, K.D., Para, M.P., Armiger, R., Anderson, W.S., Thakor, N. V, Wester, B.A., et al. (2016). Individual finger control of a modular prosthetic limb using high-density electrocorticography in a human subject. J. Neural Eng. *13*.

Hughes, C.L., Flesher, S.N., Weiss, J.M., Downey, J.E., Boninger, M., Collinger, J.L., and Gaunt, R.A. (2021). Neural stimulation and recording performance in human sensorimotor cortex over 1500 days. J. Neural Eng. *18*.

IEEE Computer Society (2001). IEEE Standard Verilog Hardware Description Language.

Irwin, Z.T., Thompson, D.E., Schroeder, K.E., Tat, D.M., Hassani, A., Bullard, A.J., Woo, S.L., Urbanchek, M.G., Sachs, A.J., Cederna, P.S., et al. (2016). Enabling Low-Power, Multi-Modal Neural Interfaces Through a Common, Low-Bandwidth Feature Space. IEEE Trans. Neural Syst. Rehabil. Eng. *24*, 521–531.

Irwin, Z.T., Schroeder, K.E., Vu, P.P., Bullard, A.J., Tat, D.M., Nu, C.S., Vaskov, A., Nason, S.R., Thompson, D.E., Bentley, J.N., et al. (2017). Neural control of finger movement via intracortical brain-machine interface. J. Neural Eng. *14*.

James, S.L., Bannick, M.S., Montjoy-Venning, W.C., Lucchesi, L.R., Dandona, L., Dandona, R., Hawley, C., Hay, S.I., Jakovljevic, M., Khalil, I., et al. (2019). Global, regional, and national burden of traumatic brain injury and spinal cord injury, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. *18*, 56–87.

Jang, T., Lim, J., Choo, K., Nason, S., Lee, J., Oh, S., Chestek, C., Sylvester, D., and Blaauw, D. (2018). A Noise-Efficient Neural Recording Amplifier Using Discrete-Time Parametric Amplification. IEEE Solid-State Circuits Lett. *1*, 203–206.

Jarosiewicz, B., Sarma, A.A., Bacher, D., Masse, N.Y., Simeral, J.D., Sorice, B., Oakley, E.M., Blabe, C., Pandarinath, C., Gilja, V., et al. (2015). Virtual typing by people with tetraplegia using

a self-calibrating intracortical brain-computer interface. Sci. Transl. Med. *7*, 1–11.

Johnson, B.C., Gambini, S., Izyumin, I., Moin, A., Zhou, A., Alexandrov, G., Santacruz, S.R., Rabaey, J.M., Carmena, J.M., and Muller, R. (2017). An implantable 700μW 64-channel neuromodulation IC for simultaneous recording and stimulation with rapid artifact recovery. IEEE Symp. VLSI Circuits, Dig. Tech. Pap. C48–C49.

Jorge, A., Royston, D.A., Tyler-Kabara, E.C., Boninger, M.L., and Collinger, J.L. (2020). Classification of Individual Finger Movements Using Intracortical Recordings in Human Motor Cortex. Neurosurgery *0*, 1–9.

Jun, J.J., Steinmetz, N.A., Siegle, J.H., Denman, D.J., Bauza, M., Barbarits, B., Lee, A.K., Anastassiou, C.A., Andrei, A., Aydın, Ç., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. Nature *551*, 232–236.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. J. Fluids Eng. Trans. ASME *82*, 35–45.

Kao, J.C., Nuyujukian, P., Ryu, S.I., and Shenoy, K. V. (2017). A High-Performance Neural Prosthesis Incorporating Discrete State Selection with Hidden Markov Models. IEEE Trans. Biomed. Eng. *64*, 935–945.

Karkare, V., Gibson, S., and Markovic, D. (2013). A 75-μW, 16-channel neural spike-sorting processor with unsupervised clustering. IEEE J. Solid-State Circuits *48*, 2230–2238.

Kaufman, M.T., Churchland, M.M., Santhanam, G., Yu, B.M., Afshar, A., Ryu, S.I., and Shenoy, K. V. (2010). Roles of Monkey Premotor Neuron Classes in Movement Preparation and Execution. J. Neurophysiol. *104*, 799–810.

Kaufman, M.T., Churchland, M.M., and Shenoy, K. V. (2013). The roles of monkey M1 neuron classes in movement preparation and execution. J. Neurophysiol. *110*, 817–825.

Kester, W. (Ed.). (2004). The Data Conversion Handbook (Elsevier Science & Technology).

Khodagholy, D., Gelinas, J.N., Thesen, T., Doyle, W., Devinsky, O., Malliaras, G.G., and Buzsáki, G. (2015). NeuroGrid: Recording action potentials from the surface of the brain. Nat. Neurosci. *18*, 310–315.

Kilgore, K.L., Peckham, P.H., Thrope, G.B., Keith, M.W., and Gallaher-Stone, K.A. (1989). Synthesis of Hand Grasp Using Functional Neuromuscular Stimulation. IEEE Trans. Biomed. Eng. *36*, 761–770.

Kilgore, K.L., Hart, R.L., Montague, F.W., Bryden, A.M., Keith, M.W., Hoyen, H.A., Sams, C.J., and Peckham, P.H. (2006). An Implanted Myoelectrically-Controlled Neuroprosthesis for Upper Extremity Function in Spinal Cord Injury. In IEEE EMBS Annual International Conference, (IEEE), pp. 1630–1633.

Kilgore, K.L., Hoyen, H.A., Bryden, A.M., Hart, R.L., Keith, M.W., and Peckham, P.H. (2008). An Implanted Upper-Extremity Neuroprosthesis Using Myoelectric Control. J. Hand Surg. Am. *33*, 539–550.

Kirsch, E., Rivlis, G., and Schieber, M.H. (2014). Primary motor cortex neurons during individuated finger and wrist movements: Correlation of spike firing rates with the motion of individual digits versus their principal components. Front. Neurol. *5*, 1–11.

Kline, J.E., Huang, H.J., Snyder, K.L., and Ferris, D.P. (2015). Isolating gait-related movement artifacts in electroencephalography during human walking. J. Neural Eng. *12*, 16.

Knaack, G.L., McHail, D.G., Borda, G., Koo, B., Peixoto, N., Cogan, S.F., Dumas, T.C., and Pancrazio, J.J. (2016). In vivo Characterization of Amorphous Silicon Carbide As a Biomaterial for Chronic Neural Interfaces. Front. Neurosci. *10*, 1–12.

Knutson, J.S., Chae, J., Hart, R.L., Keith, M.W., Hoyen, H.A., Harley, M.Y., Hisel, T.Z., Bryden, A.M., Kilgore, K.L., and Peckham, P.H. (2013). Implanted Neuroprosthesis for Assisting Arm and Hand Function after Stroke: A Case Study. J. Reh *49*, 1505–1516.

Koyama, S., Chase, S.M., Whitford, A.S., Velliste, M., Schwartz, A.B., and Kass, R.E. (2010). Comparison of brain-computer interface decoding algorithms in open-loop and closed-loop control. J. Comput. Neurosci. *29*, 73–87.

Kozai, T.D.Y., Langhals, N.B., Patel, P.R., Deng, X., Zhang, H., Smith, K.L., Lahann, J., Kotov, N.A., and Kipke, D.R. (2012). Ultrasmall implantable composite microelectrodes with bioactive surfaces for chronic neural interfaces. Nat. Mater. *11*, 1065–1073.

Kubánek, J., Miller, K.J., Ojemann, J.G., Wolpaw, J.R., and Schalk, G. (2009). Decoding flexion of individual fingers using electrocorticographic signals in humans. J. Neural Eng. *6*.

Lang, C.E., and Schieber, M.H. (2004). Human finger independence: Limitations due to passive mechanical coupling versus active neuromuscular control. J. Neurophysiol. *92*, 2802–2810.

Lee, J., Mok, E., Huang, J., Cui, L., Lee, A., Leung, V., Shellhammer, S., Larson, L., Asbeck, P., Rao, R., et al. (2019). An Implantable Wireless Network of Distributed Microscale Sensors for Neural Applications. In IEEE EMBS Conference on Neural Engineering, (IEEE), pp. 871–874.

Lee, J., Leung, V., Lee, A.H., Huang, J., Asbeck, P., Mercier, P.P., Shellhammer, S., Larson, L., Laiwalla, F., and Nurmikko, A. (2021). Neural recording and stimulation using wireless networks of microimplants. Nat. Electron. *4*, 604–614.

Lee, S., Cortese, A.J., Gandhi, A.P., Agger, E.R., McEuen, P.L., and Molnar, A.C. (2018). A 250 μm × 57 μm Microscale Opto-electronically Transduced Electrodes (MOTEs) for Neural Recording. IEEE Trans. Biomed. Circuits Syst. *12*, 1256–1266.

Lempka, S.F., Johnson, M.D., Moffitt, M.A., Otto, K.J., Kipke, D.R., and McIntyre, C.C. (2011).

Theoretical analysis of intracortical microelectrode recordings. J. Neural Eng. *8*.

Lewicki, M.S. (1998). A review of methods for spike sorting: the detection and classification of neural action potentials. Netw. Comput. Neural Syst. *9*, R53–R78.

Lim, J., Moon, E., Barrow, M., Nason, S.R., Patel, P.R., Patil, P.G., Oh, S., Lee, I., Kim, H., Sylvester, D., et al. (2020). A 0.19×0.17mm^2 Wireless Neural Recording IC for Motor Prediction with Near-Infrared-Based Power and Data Telemetry. In IEEE International Solid-State Circuits Conference, (IEEE), pp. 416–418.

Lim, J., Lee, J., Moon, E., Barrow, M., Atzeni, G., Letner, J., Costello, J., Nason, S.R., Patel, P.R., Patil, P.G., et al. (2021). A Light Tolerant Neural Recording IC for Near-Infrared-Powered Free Floating Motes. IEEE Symp. VLSI Circuits, Dig. Tech. Pap. *2021-June*, 2021–2022.

Limnuson, K., Lu, H., Chiel, H.J., and Mohseni, P. (2014). A bidirectional neural interface SoC with an integrated spike recorder, microstimulator, and low-power processor for real-time stimulus artifact rejection. Proc. IEEE 2014 Cust. Integr. Circuits Conf. CICC 2014 1–4.

Lopez, C.M., Andrei, A., Mitra, S., Welkenhuysen, M., Eberle, W., Bartic, C., Puers, R., Yazicioglu, R.F., and Gielen, G.G.E. (2014). An implantable 455-active-electrode 52-channel CMOS neural probe. IEEE J. Solid-State Circuits *49*, 248–261.

Luan, L., Wei, X., Zhao, Z., Siegel, J.J., Potnis, O., Tuppen, C.A., Lin, S., Kazmi, S., Fowler, R.A., Holloway, S., et al. (2017). Ultraflexible nanoelectronic probes form reliable , glial scar–free neural integration. Sci. Adv. *3*, 9.

Ludwig, K.A., Miriani, R.M., Langhals, N.B., Joseph, M.D., Anderson, D.J., and Kipke, D.R. (2009). Using a common average reference to improve cortical neuron recordings from microelectrode arrays. J. Neurophysiol. *101*, 1679–1689.

Mahajan, A., Bidhendi, A.K., Wang, P.T., McCrimmon, C.M., Liu, C.Y., Nenadic, Z., Do, A.H., and Heydari, P. (2015). A 64-channel ultra-low power bioelectric signal acquisition system for brain-computer interface. IEEE Biomed. Circuits Syst. Conf. 3–6.

Makin, J.G., Moses, D.A., and Chang, E.F. (2020). Machine translation of cortical activity to text with an encoder-decoder framework. Nat. Neurosci. *23*, 575–582.

Makowski, N., Campean, A., Lambrecht, J., Buckett, J., Coburn, J., Hart, R., Miller, M., Montague, F., Crish, T., Fu, M., et al. (2021). Design and Testing of Stimulation and Myoelectric Recording Modules in an Implanted Distributed Neuroprosthetic System. IEEE Trans. Biomed. Circuits Syst. *15*, 281–293.

Malik, W.Q., Truccolo, W., Brown, E.N., and Hochberg, L.R. (2011). Efficient decoding with steady-state kalman filter in neural interface systems. IEEE Trans. Neural Syst. Rehabil. Eng. *19*, 25–34.

Maynard, E.M., Nordhausen, C.T., and Normann, R.A. (1997). The Utah Intracortical Electrode Array: A recording structure for potential brain-computer interfaces. Electroencephalogr. Clin. Neurophysiol. *102*, 228–239.

McCreery, D., Cogan, S., Kane, S., and Pikov, V. (2016). Correlations between histology and neuronal activity recorded by microelectrodes implanted chronically in the cerebral cortex. J. Neural Eng. *13*, 1–17.

Memberg, W.D., Polasek, K.H., Hart, R.L., Bryden, A.M., Kilgore, K.L., Nemunaitis, G.A., Hoyen, H.A., Keith, M.W., and Kirsch, R.F. (2014). Implanted neuroprosthesis for restoring arm and hand function in people with high level tetraplegia. Arch. Phys. Med. Rehabil. *95*, 1201-1211.e1.

Mendrela, A.E., Kim, K., English, D., McKenzie, S., Seymour, J.P., Buzsáki, G., and Yoon, E. (2018). A High-Resolution Opto-Electrophysiology System With a Miniature Integrated Headstage. IEEE Trans. Biomed. Circuits Syst. *12*, 1065–1075.

Mestais, C.S., Charvet, G., Sauter-Starace, F., Foerster, M., Ratel, D., and Benabid, A.L. (2015). WIMAGINE: Wireless 64-channel ECoG recording implant for long term clinical applications. IEEE Trans. Neural Syst. Rehabil. Eng. *23*, 10–21.

Milekovic, T., Sarma, A.A., Bacher, D., Simeral, J.D., Saab, J., Pandarinath, C., Sorice, B.L., Blabe, C., Oakley, E.M., Tringale, K.R., et al. (2018). Stable long-term BCI-enabled communication in ALS and locked-in syndrome using LFP signals. J. Neurophysiol. *120*, 343–360.

Miranda, H., Gilja, V., Chestek, C.A., Shenoy, K. V., and Meng, T.H. (2010). HermesD: A high-rate long-range wireless transmission system for simultaneous multichannel neural recording applications. IEEE Trans. Biomed. Circuits Syst. *4*, 181–191.

Moffitt, M.A., and McIntyre, C.C. (2005). Model-based analysis of cortical recording with silicon microelectrodes. Clin. Neurophysiol. *116*, 2240–2250.

Mollazadeh, M., Aggarwal, V., Davidson, A.G., Law, A.J., Thakor, N. V., and Schieber, M.H. (2011). Spatiotemporal variation of multiple neurophysiological signals in the primary motor cortex during dexterous reach-to-Grasp movements. J. Neurosci. *31*, 15531–15543.

Mollazadeh, M., Aggarwal, V., Thakor, N. V., and Schieber, M.H. (2014). Principal components of hand kinematics and neurophysiological signals in motor cortex during reach to grasp movements. J. Neurophysiol. *112*, 1857–1870.

Moritz, C.T., Perlmutter, S.I., and Fetz, E.E. (2008). Direct control of paralysed muscles by cortical neurons. Nature *456*, 639–642.

Muller, R., Le, H.P., Li, W., Ledochowitsch, P., Gambini, S., Bjorninen, T., Koralek, A., Carmena,

J.M., Maharbiz, M.M., Alon, E., et al. (2014). A miniaturized 64-channel 225µW wireless electrocorticographic neural sensor. Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf. *57*, 412–413.

Mulliken, G.H., Musallam, S., and Andersen, R.A. (2008). Decoding trajectories from posterior parietal cortex ensembles. J. Neurosci. *28*, 12913–12926.

Murmann, B. (2020). ADC Performance Survey 1997-2020.

Musk, E. (2019). An integrated brain-machine interface platform with thousands of channels. J. Med. Internet Res. *21*, 1–14.

Napier, J.R. (1956). The prehensile movements of the human hand. J. Bone Jt. Surg. *38-B*, 902–913.

Nason, S.R., Vaskov, A.K., Willsey, M.S., Welle, E.J., An, H., Vu, P.P., Bullard, A.J., Nu, C.S., Kao, J.C., Shenoy, K. V., et al. (2020). A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces. Nat. Biomed. Eng. *4*, 973–983.

Nason, S.R., Mender, M.J., Letner, J.G., Chestek, C.A., and Patil, P.G. (2021a). Restoring upper extremity function with brain-machine interfaces. Emerging Horizons in Neuromodulation: New Frontiers in Brain and Spine Stimulation *159*, 153-186.

Nason, S.R., Mender, M.J., Vaskov, A.K., Willsey, M.S., Ganesh Kumar, N., Kung, T.A., Patil, P.G., and Chestek, C.A. (2021b). Real-time linear prediction of simultaneous and independent movements of two finger groups using an intracortical brain-machine interface. Neuron *109*, 3164–3177.

Naufel, S., Glaser, J.I., Kording, K.P., Perreault, E.J., and Miller, L.E. (2019). A muscle-activity-dependent gain between motor cortex and EMG. J. Neurophysiol. *121*, 61–73.

Ng, K.A., and Xu, Y.P. (2016). A Low-Power, High CMRR Neural Amplifier System Employing CMOS Inverter-Based OTAs With CMFB Through Supply Rails. IEEE J. Solid-State Circuits *51*, 724–737.

Nguyen, J.K., Park, D.J., Skousen, J.L., Hess-Dunning, A.E., Tyler, D.J., Rowan, S.J., Weder, C., and Capadona, J.R. (2014). Mechanically-compliant intracortical implants reduce the neuroinflammatory response. J. Neural Eng. *11*, 15.

Nordhausen, C.T., Maynard, E.M., and Normann, R.A. (1996). Single unit recording capabilities of a 100 microelectrode array. Brain Res. *726*, 129–140.

Novak, P., Przybyszewski, A.W., Barborica, A., Ravin, P., Margolin, L., and Pilitsis, J.G. (2011). Localization of the subthalamic nucleus in Parkinson disease using multiunit activity. J. Neurol. Sci. *310*, 44–49.

Nuyujukian, P., Kao, J.C., Ryu, S.I., and Shenoy, K. V. (2016). A Nonhuman Primate Brain-Computer Typing Interface. Proc. IEEE *105*, 1–7.

Nuyujukian, P., Albites Sanabria, J., Saab, J., Pandarinath, C., Jarosiewicz, B., Blabe, C.H., Franco, B., Mernoff, S.T., Eskandar, E.N., Simeral, J.D., et al. (2018). Cortical control of a tablet computer by people with paralysis. PLoS One *13*, 1–16.

Obaid, A., Hanna, M., Wu, Y., Kollo, M., Racz, R., Angle, M.R., Müller, J., Brackbill, N., Wray, W., Franke, F., et al. (2020). Massively parallel microwire arrays integrated with CMOS chips for neural recording. Sci. Adv. *6*, 10.

Okorokova, E. V., Goodman, J.M., Hatsopoulos, N.G., and Bensmaia, S.J. (2020). Decoding hand kinematics from population responses in sensorimotor cortex during grasping. J. Neural Eng. *17*, 12.

Orsborn, A.L., Dangi, S., Moorman, H.G., and Carmena, J.M. (2012). Closed-Loop Decoder Adaptation on Intermediate Time-Scales Facilitates Rapid BMI Performance Improvements Independent of Decoder Initialization Conditions. IEEE Trans. Neural Syst. Rehabil. Eng. *20*, 468–477.

Overduin, S.A., d'Avella, A., Carmena, J.M., and Bizzi, E. (2014). Muscle synergies evoked by microstimulation are preferentially encoded during behavior. Front. Comput. Neurosci. *8*, 1–10.

Overduin, S.A., D'Avella, A., Roh, J., Carmena, J.M., and Bizzi, E. (2015). Representation of muscle synergies in the primate brain. J. Neurosci. *35*, 12615–12624.

Pandarinath, C., Nuyujukian, P., Blabe, C.H., Sorice, B.L., Saab, J., Willett, F.R., Hochberg, L.R., Shenoy, K. V., and Henderson, J.M. (2017). High performance communication by people with paralysis using an intracortical brain-computer interface. Elife *6*, 1–27.

Pandarinath, C., O'Shea, D.J., Collins, J., Jozefowicz, R., Stavisky, S.D., Kao, J.C., Trautmann, E.M., Kaufman, M.T., Ryu, S.I., Hochberg, L.R., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods *15*, 805–815.

Park, S.Y., Cho, J., Na, K., and Yoon, E. (2018). Modular 128-Channel Δ - ΔΣ Analog Front-End Architecture Using Spectrum Equalization Scheme for 1024-Channel 3-D Neural Recording Microsystems. IEEE J. Solid-State Circuits *53*, 501–514.

Patel, P.R., Zhang, H., Robbins, M.T., Nofar, J.B., Marshall, S.P., Kobylarek, M.J., Kozai, T.D.Y., Kotov, N.A., and Chestek, C.A. (2016). Chronic in vivo stability assessment of carbon fiber microelectrode arrays. J. Neural Eng. *13*.

Patil, P.G., Carmena, J.M., Nicolelis, M.A.L., and Turner, D.A. (2004). Ensemble Recordings of Human Subcortical Neurons as a Source of Motor Control Signals for a Brain-Machine Interface. Neurosurgery *55*, 27–38.

Peckham, P.H., Mortimer, J.T., and Marsolais, E.B. (1980). Controlled prehension and release in the C5 quadriplegic elicited by functional electrical stimulation of the paralyzed forearm musculature. Ann. Biomed. Eng. *8*, 369–388.

Perge, J.A., Zhang, S., Malik, W.Q., Homer, M.L., Cash, S., Friehs, G., Eskandar, E.N., Donoghue, J.P., and Hochberg, L.R. (2014). Reliability of directional information in unsorted spikes and local field potentials recorded in human motor cortex. J. Neural Eng. *11*.

Pistohl, T., Schulze-Bonhage, A., Aertsen, A., Mehring, C., and Ball, T. (2012). Decoding natural grasp types from human ECoG. Neuroimage *59*, 248–260.

Potter, K.A., Buck, A.C., Self, W.K., and Capadona, J.R. (2012). Stab injury and device implantation within the brain results in inversely multiphasic neuroinflammatory and neurodegenerative responses. J. Neural Eng. *9*.

Rey, H.G., Pedreira, C., and Quiroga, R.Q. (2015). Past, present and future of spike sorting techniques. Brain Res. Bull. *119*, 106–117.

Rizk, M., Bossetti, C.A., Jochum, T.A., Callender, S.H., Nicolelis, M.A.L., Turner, D.A., and Wolf, P.D. (2009). A fully implantable 96-channel neural data acquisition system. J. Neural Eng. *6*, 026002.

Robinet, S., Audebert, P., Regis, G., Zongo, B., Beche, J.F., Condemine, C., Filipe, S., and Charvet, G. (2011). A Low-Power 0.7 µVrms 32-Channel Mixed-Signal Circuit for ECoG Recordings. IEEE J. Emerg. Sel. Top. Circuits Syst. *1*, 451–460.

Rodger, D.C., Fong, A.J., Li, W., Ameri, H., Ahuja, A.K., Gutierrez, C., Lavrov, I., Zhong, H., Menon, P.R., Meng, E., et al. (2008). Flexible parylene-based multielectrode array technology for high-density neural stimulation and recording. Sensors Actuators B Chem. *132*, 449–460.

Rouse, A.G. (2016). A four-dimensional virtual hand brain-machine interface using active dimension selection. J. Neural Eng. *13*.

Sachs, N.A., Ruiz-Torres, R., Perreault, E.J., and Miller, L.E. (2015). Brain-state classification and a dual-state decoder dramatically improve the control of cursor movement through a brain-machine interface. J. Neural Eng. *13*, 16009.

Saleh, M., Takahashi, K., Amit, Y., and Hatsopoulos, N.G. (2010). Encoding of coordinated grasp trajectories in primary motor cortex. J. Neurosci. *30*, 17079–17090.

Santello, M., Flanders, M., and Soechting, J.F. (1998). Postural hand synergies for tool use. J. Neurosci. *18*, 10105–10115.

Santhanam, G., Ryu, S.I., Yu, B.M., Afshar, A., and Shenoy, K. V. (2006). A high-performance brain-computer interface. Nature *442*, 195–198.

Santhanam, G., Linderman, M.D., Gilja, V., Afshar, A., Ryu, S.I., Meng, T.H., and Shenoy, K. V. (2007). HermesB: A continuous neural recording system for freely behaving primates. IEEE Trans. Biomed. Eng. *54*, 2037–2050.

Schalk, G., Kubánek, J., Miller, K.J., Anderson, N.R., Leuthardt, E.C., Ojemann, J.G., Limbrick, D., Moran, D., Gerhardt, L.A., and Wolpaw, J.R. (2007). Decoding two-dimensional movement trajectories using electrocorticographic signals in humans. J. Neural Eng. *4*, 264–275.

Schlag, J., and Balvin, R. (1963). Background activity in the cerebral cortex and reticular formation in relation with the electroencephalogram. Exp. Neurol. *8*, 203–219.

Scholvin, J., Kinney, J.P., Bernstein, J.G., Moore-kochlacs, C., Kopell, N., Fonstad, C.G., and Boyden, E.S. (2016). Close-Packed Silicon Microelectrodes for Scalable Spatially Oversampled Neural Recording. IEEE Trans. Biomed. Eng. *63*, 120–130.

Schwemmer, M.A., Skomrock, N.D., Sederberg, P.B., Ting, J.E., Sharma, G., Bockbrader, M.A., and Friedenberg, D.A. (2018). Meeting brain–computer interface user performance expectations using a deep neural network decoding framework. Nat. Med. *24*, 1669–1676.

Seo, D., Carmena, J.M., Rabaey, J.M., Alon, E., and Maharbiz, M.M. (2013). Neural Dust: An Ultrasonic, Low Power Solution for Chronic Brain-Machine Interfaces.

Seo, D., Neely, R.M., Shen, K., Singhal, U., Alon, E., Rabaey, J.M., Carmena, J.M., and Maharbiz, M.M. (2016). Wireless Recording in the Peripheral Nervous System with Ultrasonic Neural Dust. Neuron *91*, 529–539.

Serruya, M.D., Hatsopoulos, N.G., Paninski, L., Fellows, M.R., and Donoghue, J.P. (2002). Instant neural control of a movement signal. Nature *416*, 141–142.

Seymour, J.P., and Kipke, D.R. (2007). Neural probe design for reduced tissue encapsulation in CNS. Biomaterials *28*, 3594–3607.

Shahrokhi, F., Abdelhalim, K., Serletis, D., Carlen, P.L., and Genov, R. (2010). The 128-channel fully differential digital integrated neural recording and stimulation interface. IEEE Trans. Biomed. Circuits Syst. *4*, 149–161.

Shanechi, M.M., Orsborn, A.L., Moorman, H.G., Gowda, S., Dangi, S., and Carmena, J.M. (2017). Rapid control and feedback rates enhance neuroprosthetic control. Nat. Commun. *8*, 1–10.

Shannon, C.E. (1949). Communication in the Presence Noise. Proc. IRE *37*, 10–21.

Sharma, G., Annetta, N., Friedenberg, D., Blanco, T., Vasconcelos, D., Shaikhouni, A., Rezai, A.R., and Bouton, C. (2015). Time Stability and Coherence Analysis of Multiunit, Single-Unit and Local Field Potential Neuronal Signals in Chronically Implanted Brain Electrodes. Bioelectron. Med. *2*, 63–71.

Shen, L., Lu, N., and Sun, N. (2017). A 1V 0.25uW Inverter-Stacking Amplifier with 1.07 Noise Efficiency Factor. In 2017 Symposium on VLSI Circuits, pp. C140–C141.

Shen, L., Lu, N., and Sun, N. (2018). A 1-V 0.25-μW Inverter Stacking Amplifier With 1.07 Noise Efficiency Factor. IEEE J. Solid-State Circuits *53*, 896–905.

Shenoy, K. V., Meeker, D., Cao, S., Kureshi, S.A., Pesaran, B., Buneo, C.A., Batista, A.P., Mitra, P.P., Burdick, J.W., and Andersen, R.A. (2003). Neural prosthetic control signals from plan activity. Neuroreport *14*, 591–596.

Shenoy, K. V, Sahani, M., and Churchland, M.M. (2013). Cortical Control of Arm Movements: A Dynamical Systems Perspective. Annu. Rev. Neurosci. *36*, 337–359.

Simeral, J.D., Kim, S.-P., Black, M.J., Donoghue, J.P., and Hochberg, L.R. (2011). Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. J. Neural Eng. *8*, 24.

Simeral, J.D., Hosman, T., Saab, J., Flesher, S.N., Vilela, M., Franco, B., Kelemen, J., Brandman, D.M., Ciancibello, J.G., Rezaii, P.G., et al. (2021). Home Use of a Percutaneous Wireless Intracortical Brain-Computer Interface by Individuals With Tetraplegia. IEEE Trans. Biomed. Eng. *68*, 2313–2325.

Singer, A., Dutta, S., Lewis, E., Chen, Z., Chen, J.C., Verma, N., Avants, B., Feldman, A.K., O'Malley, J., Beierlein, M., et al. (2020). Magnetoelectric Materials for Miniature , Wireless Neural Stimulation at Therapeutic Frequencies. Neuron *107*, 631–643.

Slutzky, M.W., Jordan, L.R., Krieg, T., Chen, M., Mogul, D.J., and Miller, L.E. (2010). Optimal spacing of surface electrode arrays for brain-machine interface applications. J. Neural Eng. *7*.

Smith, B., Crish, T.J., Buckett, J.R., Kilgore, K.L., and Peckham, P.H. (2005). Development of an implantable networked neuroprosthesis. 2nd Int. IEEE EMBS Conf. Neural Eng. *2005*, 454–457.

So, K., Dangi, S., Orsborn, A.L., Gastpar, M.C., and Carmena, J.M. (2014). Subject-specific modulation of local field potential spectral power during brain-machine interface control in primates. J. Neural Eng. *11*.

Speier, W., Arnold, C., Lu, J., Deshpande, A., and Pouratian, N. (2014). Integrating Language Information With a Hidden Markov Model to Improve Communication Rate in the P300 Speller. IEEE Trans. Neural Syst. Rehabil. Eng. *22*, 678–684.

Srinivasan, S.S., Maimon, B.E., Diaz, M., Song, H., and Herr, H.M. (2018). Closed-loop functional optogenetic stimulation. Nat. Commun. *9*, 1–10.

Stark, E., and Abeles, M. (2007). Predicting movement from multiunit activity. J. Neurosci. *27*, 8387–8394.

Stavisky, S.D., Kao, J.C., Nuyujukian, P., Ryu, S.I., and Shenoy, K. V. (2015). A high performing brain-machine interface driven by low-frequency local field potentials alone and together with spikes. J. Neural Eng. *12*.

Stavisky, S.D., Willett, F.R., Wilson, G.H., Murphy, B.A., Rezaii, P., Avansino, D.T., Memberg, W.D., Miller, J.P., Kirsch, R.F., Hochberg, L.R., et al. (2019). Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis. Elife *8*, 1–31.

Stavisky, S.D., Willett, F.R., Avansino, D.T., Hochberg, L.R., Shenoy, K. V., and Henderson, J.M. (2020). Speech-related dorsal motor cortex activity does not interfere with iBCI cursor control. J. Neural Eng. *17*.

Steinmetz, N.A., Koch, C., Harris, K.D., and Carandini, M. (2018). Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. Curr. Opin. Neurobiol. *50*, 92–100.

Steyaert, M.S.J., Sansen, W.M.C., and Zhongyuan, C. (1987). A Micropower Low-Noise Monolithic Instrumentation Amplifier For Medical Purposes. IEEE J. Solid-State Circuits *22*, 1163–1168.

Sussillo, D., Nuyujukian, P., Fan, J.M., Kao, J.C., Stavisky, S.D., Ryu, S., and Shenoy, K. (2012). A recurrent neural network for closed-loop intracortical brain-machine interface decoders. J. Neural Eng. *9*.

Sussillo, D., Stavisky, S.D., Kao, J.C., Ryu, S.I., and Shenoy, K. V. (2016). Making brain-machine interfaces robust to future neural variability. Nat. Commun. *7*, 1–12.

Swann, N.C., De Hemptinne, C., Miocinovic, S., Qasim, S., Ostrem, J.L., Galifianakis, N.B., Luciano, M.S., Wang, S.S., Ziman, N., Taylor, R., et al. (2018). Chronic multisite brain recordings from a totally implantable bidirectional neural interface: Experience in 5 patients with Parkinson's disease. J. Neurosurg. *128*, 605–616.

Tan, H., Pogosyan, A., Ashkan, K., Green, A.L., Aziz, T., Foltynie, T., Limousin, P., Zrinzo, L., Hariz, M., and Brown, P. (2016). Decoding gripping force based on local field potentials recorded from subthalamic nucleus in humans. Elife *5*, 1–24.

Taylor, D.M., Tillery, S.I.H., and Schwartz, A.B. (2002a). Direct cortical control of 3D neuroprosthetic devices. Science (80-.). *296*, 1829–1832.

Taylor, P., Esnouf, J., and Hobby, J. (2002b). The functional impact of the Freehand system on tetraplegic hand function. Clinical results. Nat. Spinal Cord *40*, 560–566.

Thompson, D.E., Quitadamo, L.R., Mainardi, L., Laghari, K.U.R., Gao, S., Kindermans, P.J., Simeral, J.D., Fazel-Rezai, R., Matteucci, M., Falk, T.H., et al. (2014). Performance measurement for brain-computer or brain-machine interfaces: A tutorial. J. Neural Eng. *11*.

Todorov, E. (2000). Direct cortical control of muscle activation in voluntary arm movements: A

model. Nat. Neurosci. *3*, 391–398.

Trautmann, E.M., Stavisky, S.D., Lahiri, S., Ames, K.C., Kaufman, M.T., O'Shea, D.J., Vyas, S., Sun, X., Ryu, S.I., Ganguli, S., et al. (2019). Accurate Estimation of Neural Population Dynamics without Spike Sorting. Neuron *103*, 292-308.e4.

Vargas-Irwin, C.E., Shakhnarovich, G., Yadollahpour, P., Mislow, J.M.K., Black, M.J., and Donoghue, J.P. (2010). Decoding complete reach and grasp actions from local primary motor cortex populations. J. Neurosci. *30*, 9659–9669.

Vaskov, A.K., Irwin, Z.T., Nason, S.R., Vu, P.P., Nu, C.S., Bullard, A.J., Hill, M., North, N., Patil, P.G., and Chestek, C.A. (2018). Cortical decoding of individual finger group motions using ReFIT Kalman filter. Front. Neurosci. *12*.

Velliste, M., Perel, S., Spalding, M.C., Whitford, A.S., and Schwartz, A.B. (2008). Cortical control of a prosthetic arm for self-feeding. Nature *453*, 1098–1101.

Ventura, V. (2008). Spike Train Decoding Without Spike Sorting. Neural Comput. *20*, 923–963.

Viventi, J., Kim, D.H., Vigeland, L., Frechette, E.S., Blanco, J.A., Kim, Y.S., Avrin, A.E., Tiruvadi, V.R., Hwang, S.W., Vanleer, A.C., et al. (2011). Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. Nat. Neurosci. *14*, 1599–1605.

Wahnoun, R., He, J., and Helms Tillery, S.I. (2006). Selection and parameterization of cortical neurons for neuroprosthetic control. J. Neural Eng. *3*, 162–171.

Wang, D., Zhang, Q., Li, Y., Wang, Y., Zhu, J., Zhang, S., and Zheng, X. (2014). Long-term decoding stability of local field potentials from silicon arrays in primate motor cortex during a 2D center out task. J. Neural Eng. *11*, 049501.

Wang, P.T., Camacho, E., Wang, M., Li, Y., Shaw, S.J., Armacost, M., Gong, H., Kramer, D., Lee, B., Andersen, R.A., et al. (2019). A benchtop system to assess the feasibility of a fully independent and implantable brain-machine interface. J. Neural Eng. *16*.

Wark, H.A.C., Sharma, R., Mathews, K.S., Fernandez, E., Yoo, J., Christensen, B., Tresco, P., Rieth, L., Solzbacher, F., Normann, R.A., et al. (2013). A new high-density (25 electrodes/mm2) penetrating microelectrode array for recording and stimulating sub-millimeter neuroanatomical structures. J. Neural Eng. *10*.

Wattanapanitch, W., and Sarpeshkar, R. (2011). A low-power 32-channel digitally programmable neural recording integrated circuit. IEEE Trans. Biomed. Circuits Syst. *5*, 592–602.

Willett, F.R., Pandarinath, C., Jarosiewicz, B., Murphy, B.A., Memberg, W.D., Blabe, C.H., Saab, J., Walter, B.L., Sweet, J.A., Miller, J.P., et al. (2017). Feedback control policies employed by people using intracortical brain-computer interfaces. J. Neural Eng. *14*.

Willett, F.R., Murphy, B.A., Young, D., Memberg, W.D., Blabe, C.H., Pandarinath, C., Franco, B., Saab, J., Walter, B.L., Sweet, J.A., et al. (2018). A comparison of intention estimation methods for decoder calibration in intracortical brain-computer interfaces. IEEE Trans. Biomed. Eng. *65*, 2066–2078.

Willett, F.R., Young, D.R., Murphy, B.A., Memberg, W.D., Blabe, C.H., Pandarinath, C., Stavisky, S.D., Rezaii, P., Saab, J., Walter, B.L., et al. (2019). Principled BCI Decoder Design and Parameter Selection Using a Feedback Control Model. Sci. Rep. *9*, 1–17.

Willett, F.R., Deo, D.R., Avansino, D.T., Rezaii, P., Hochberg, L.R., Henderson, J.M., and Shenoy, K. V. (2020). Hand Knob Area of Premotor Cortex Represents the Whole Body in a Compositional Way. Cell *181*, 396-409.e26.

Willett, F.R., Avansino, D.T., Hochberg, L.R., Henderson, J.M., and Shenoy, K. V (2021). High-performance brain-to-text communication via handwriting. Nature *593*, 249–254.

Willsey, M.S., Nason, S.R., Ensel, S.R., Temmar, H., Mender, M.J., Costello, J.T., Patil, P.G., and Chestek, C.A. (2021). Real-Time Brain-Machine Interface Achieves High-Velocity Prosthetic Finger Movements Using a Biologically-Inspired Neural Network Decoder. BioRxiv.

Wise, K.D., Anderson, D.J., Hetke, J.F., Kipke, D.R., and Najafi, K. (2004). Wireless implantable microsystems: High-density electronic interfaces to the nervous system. Proc. IEEE *92*, 76–97.

Wodlinger, B., Downey, J.E., Tyler-Kabara, E.C., Schwartz, A.B., Boninger, M.L., and Collinger, J.L. (2015). Ten-dimensional anthropomorphic arm control in a human brain−machine interface: difficulties, solutions, and limitations. J. Neural Eng. *12*.

Wu, W., Shaikhouni, A., Donoghue, J.P., and Black, M.J. (2004a). Closed-loop neural control of cursor motion using a kalman filter. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. *26*, 4126–4129.

Wu, W., Black, M.J., Mumford, D., Gao, Y., Bienenstock, E., and Donoghue, J.P. (2004b). Modeling and decoding motor cortical activity using a switching Kalman filter. IEEE Trans. Biomed. Eng. *51*, 933–942.

Xu, H., Hirschberg, A.W., Scholten, K., Berger, T.W., Song, D., and Meng, E. (2018). Acute in vivo testing of a conformal polymer microelectrode array for multi-region hippocampal recordings. J. Neural Eng. *15*.

Yang, L., Lee, K., Villagracia, J., and Masmanidis, S.C. (2020). Open source silicon microprobes for high throughput neural recording. J. Neural Eng. *17*.

Yang, X., Zhou, T., Zwang, T.J., Hong, G., Zhao, Y., Viveros, R.D., Fu, T.M., Gao, T., and Lieber, C.M. (2019). Bioinspired neuron-like electronics. Nat. Mater. *18*, 510–517.

Yin, M., Li, H., Bull, C., Borton, D.A., Aceros, J., Larson, L., and Nurmikko, A. V. (2013). An externally head-mounted wireless neural recording device for laboratory animal research and

possible human clinical use. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS 3109–3114.

Yin, M., Borton, D.A., Komar, J., Agha, N., Lu, Y., Li, H., Laurens, J., Lang, Y., Li, Q., Bull, C., et al. (2014). Wireless neurosensor for full-spectrum electrophysiology recordings during free behavior. Neuron *84*, 1170–1182.

Yoon, D.Y., Pinto, S., Chung, S.W., Merolla, P., Koh, T.W., and Seo, D. (2021). A 1024-Channel Simultaneous Recording Neural SoC with Stimulation and Real-Time Spike Detection. IEEE Symp. VLSI Circuits, Dig. Tech. Pap. 2020–2021.

Young, D., Willett, F., Memberg, W.D., Murphy, B., Walter, B., Sweet, J., Miller, J., Hochberg, L.R., Kirsch, R.F., and Ajiboye, A.B. (2018). Signal processing methods for reducing artifacts in microelectrode brain recordings caused by functional electrical stimulation. J. Neural Eng. *15*.

Zhuang, J., Truccolo, W., Vargas-Irwin, C., and Donoghue, J.P. (2010). Decoding 3-D reach and grasp kinematics from high-frequency local field potentials in primate primary motor cortex. IEEE Trans. Biomed. Eng. *57*, 1774–1784.