Deep Learning for Scalable Sensorimotor Brain-Computer Interfaces

Joel Ye

PhD Thesis Proposal

July 30, 2025

Neural Computation, Neuroscience Institute
Machine Learning Department, School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Robert GAUNT (co-chair)

Leila WEHBE (co-chair)

Jennifer COLLINGER

Aran NAYEBI

Chethan PANDARINATH (Emory University)

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Abstract

The increasing ambition of neuroscience and neurotechnology both supplies and demands vast new quantities of neural data, which creates a critical need for methods that can operate effectively on neural data at scale. This need could potentially be addressed by deep learning (DL). Here, we assess this program for sensorimotor intracortical brain-computer interfaces (iBCI). iBCI systems have traditionally been built by collecting short datasets to relate a user's neural activity with their bodily state. Although this approach can enable both motor control and sensory feedback in research settings, iBCI's real-world adoption will further depend on achieving high performance with exquisite reliability and convenience. In other words, iBCI systems must, too, evolve to operate effectively on BCI data at real-world scales. To this end, I present one completed study and two proposed projects that show how deep learning models can provide a platform for scaling BCI systems. Through this platform, I promote the view that BCI models can and should be built to reflect the BCI system's lifelong data collection.

Aim 1: Large scale pretraining improves modeling of intracortical motor datasets [in submission].

The driver of deep learning's efficacy across domains is its ability to leverage conserved statistical structure across datasets, primarily enabled by a stage of model preparation on large scale data called pretraining. Previous work has identified the requisite conserved structures across motor cortical datasets, and so we systematically measure the efficacy of deep neural network (DNN) pretraining on such datasets. We establish that DNN efficacy improves with neural data scale on both multi-subject and multi-behavior datasets. Yet, this scaling has diminishing returns as data in the downstream, target setting grows, rendering the scaling less impactful for long term BCI applications. We conclude that pretrained networks may accelerate initial BCI calibration speeds but will not fundamentally remove the need for continuous BCI data collection.

Aim 2: Pretrained deep networks enable rapid and scaleable upper limb neuroprosthetic control [proposed].

High degree of freedom control of a robotic arm and hand is possible with current iBCIs, but requires extensive daily recalibration and experimenter intervention. Based on successful deep network use in speech BCIs, we propose that an NDT-based controller which accumulates calibration data across days can address both of these challenges. We will evaluate NDT for 7-degree of freedom neuroprosthetic upper limb control in up to two human participants, aiming to demonstrate rapidly calibrated robotic upper limb control. We further aim to demonstrate the model's scalability by extending this control to additional degrees of freedom in the hand without changes to the model design.

Aim 3: Surveying the neural response to intracortical microstimulation [proposed].

Sensory feedback is integral to native motor control, and can be provided in iBCIs through in-

tracortical microstimulation (ICMS) of the somatosensory cortex. Yet, current characterization of different ICMS patterns relies nearly entirely on expensive and noisy behavioral reports, instead of through its impact on ongoing neural activity. A predictive model of the neural response to ICMS is needed to accelerate the development of sophisticated ICMS protocols. Towards this goal, we first introduce a method to recover spiking activity from artifacted recordings, and then taxonomize the sensory neural response to ICMS through transfer learning experiments. In this taxonomy, we show that the neural response to temporally varied stimulation shares transferable structure, but responses across different stimulation channels are much more idiosyncratic. These results inform the design of protocols to accurately map the neural response to ICMS.

Contents

1 Introduction and Background

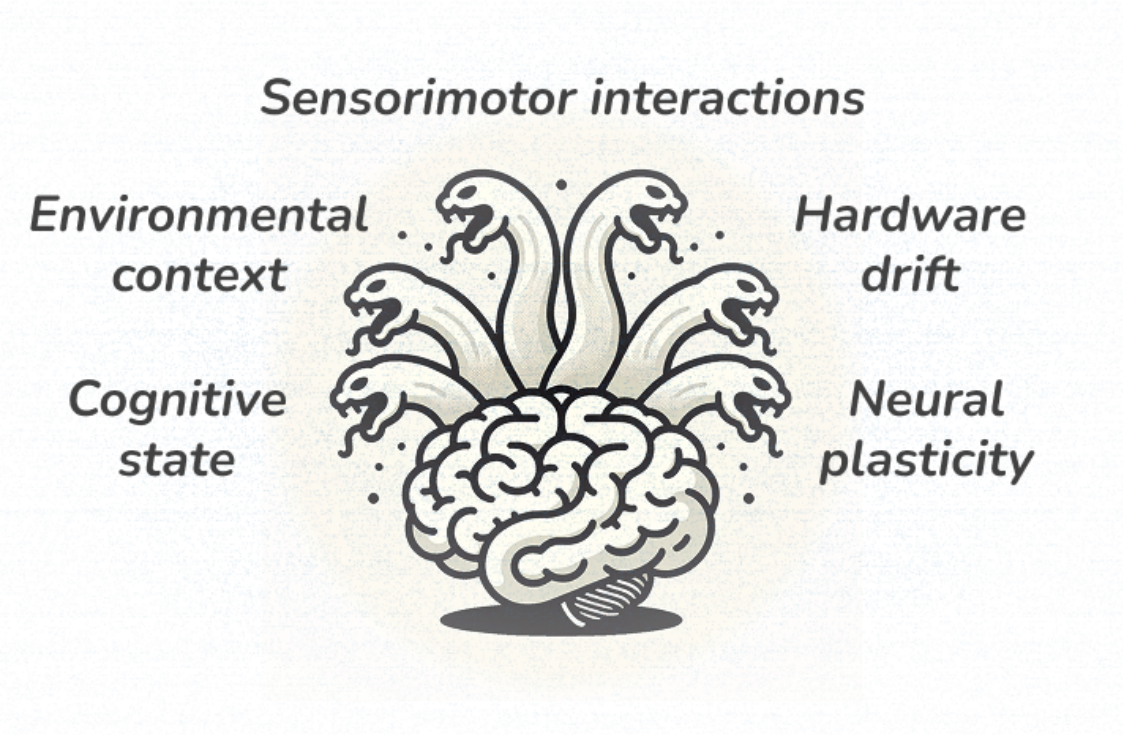

Figure 1.1: Models of neural data must robustly relate neural activity to certain variables of interest while accounting for the influence of numerous other factors that also modulate the same neural activity.

Today, neuroscientists seek to study the brain in natural, dynamic environments, and neurotechnology companies are similarly racing to produce a device that will work robustly in the real world. In this increasing scope, a great variety of factors can affect the neural data that BCI devices collect, and consequently, the performance of the models built on that data. In the best cases, we have some mechanistic account of these factors, such as in our characterization of the physical degradation at the neural-electrode interface [1, 2]. In other cases, well-controlled experiments provide a strong descriptive characterization, as in the case of cognitive states [3, 4] and body posture [5]. Most often, we lump the remaining factors in the catch-all term of "context," as often arises in studying similar motor behaviors under variable task requirements [6–8]. Designing BCI systems that perform at scale is thus daunting because it should require accounting for all of these known factors and presumably many more unknown ones.

Fortunately, progress in BCIs has rarely hinged on theoretical clarity. The first population vector decoders were enabled by the observation that motor neurons fired reliably according to cosine tuning curves [9], not from a mechanistic understanding of why they did so. A caricatured workflow for mitigating the myriad impacts of the above factors is to collect data and use this data to update

our models. A deeper understanding can support expectations of model robustness, but ultimately, we must design BCIs that recover gracefully when our understanding eventually fails. Likely, the rate of BCI progress will hinge on how quickly our methods can adapt to empirical failures as they arise.

To minimize the Sisyphean burden of this task, many domains faced with similar challenges have turned to deep neural networks, deprioritizing interpretable theories in favor of radical empiricism [10]. In this proposal, I aim to establish such a data-driven framework for thinking about sensorimotor BCIs, describing three ways in which we can use deep networks to relate and aggregate BCI datasets. To contextualize this work, I will introduce brain computer interfaces and deep learning's application to BCIs in turn.

1.1 Brain-computer interface models

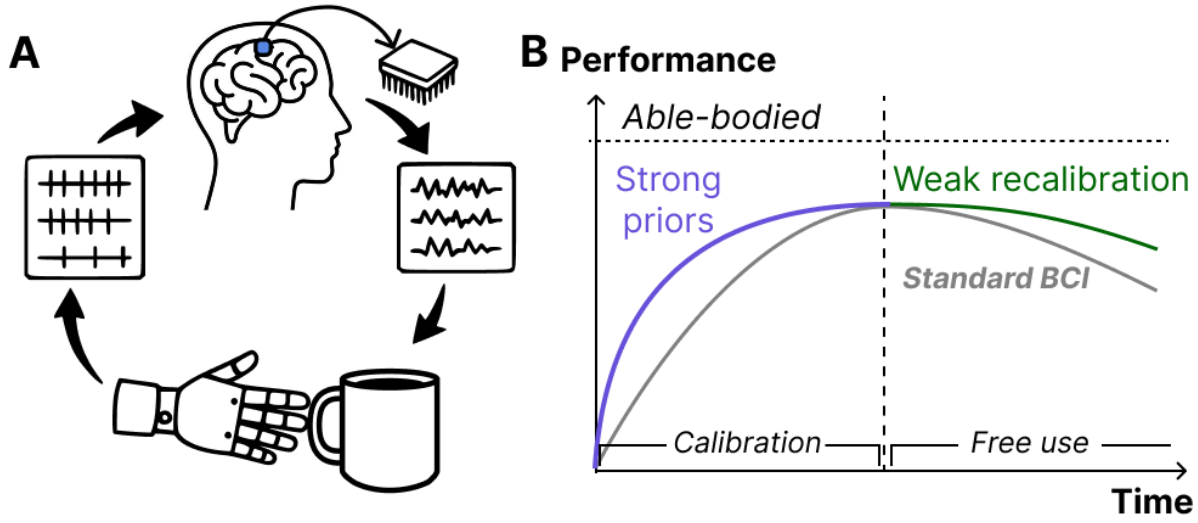

Figure 1.2: A) A sensorimotor BCI-enabled control loop. By recording neural activity from an implanted microelectrode array, this BCI can allow the user to control a robotic hand. When touch sensors on the hand contact a physical object, we can deliver stimulation pulses back to the brain to provide the user sensory feedback. B) Motor BCI models must be calibrated to relate neural activity at the implant interface with the user's intentions. Performance increases over an explicit calibration phase, which can be accelerated with learned priors, but will decline over time during free use of the system. This decline can be mitigated with varied recalibration methods.

A brain-computer interface (BCI) is a system that allows observation and control of neural activity. In the rehabilitative setting, for example, BCIs can record from motor cortex to predict movement intentions and can control sensory cortex to evoke sensations. Together, the premise of a sensorimotor BCI is that we might directly relate a user's neural activity with their sensorimotor experience, and thereby restore a degree of sensorimotor function to a paralyzed user (Fig. 1.2A).

Over the last two decades, a number of studies have demonstrated the efficacy of human BCIs

based on implanted microelectrode arrays [1, 11]. With implants in motor cortex recording neuronal spiking activity, users have achieved control of digital cursors, robotic arms, and language and speech generation [12–15]. With stimulation of implants in somatosensory cortex, users perceive tactile sensations on their hand, which is valuable as an end in itself but can also be utilized functionally for more skillful motor control [16–19]. BCI algorithms research now unfolds on several axes, naturally including expanding BCI capabilities to new functional milestones [14, 20], but also making the existing repertoire reliably performant across users and time [21, 22], and minimizing user burden for system adoption [1, 23].

Motor BCI challenges: Motor innovations can be organized by evaluating impact to BCI performance over time (Fig. 1.2B). BCI models are initialized with a calibration protocol where experimenters collect by paired neural activity and motor intention labels by prompting the user to behave according to experimental cues. The gray arc shows how BCI performance will increase with this calibration period [24], but then decline over time due to signal nonstationarities that render the calibration data less relevant [25]. Performance gain from calibration can be accelerated with strong priors, like those given by data from recent days, degraded only by light nonstationarity. Given that practical constraints often end calibration before performance saturation, such priors may improve the maximum attainable performance. Next, preserving model performance over time is a major priority for real world convenience, as the worst case strategy of recalibrating models from scratch imposes substantial user burden. One strategy to achieve persistent performance is collect a large corpus of calibration data across multiple days. This strategy has been used to provide multiple months of continuous performance in speech BCIs and 2D cursor control BCIs [13, 21, 26, 27]. Performance can also be explicitly maintained by using minimal supervised recalibration or behavioral priors (Weak recalibration) [22, 28]. Multi-day data modeling has also enabled stable reach and grasp performance in an electrocorticographic (ECoG) device [29]. However, these methods have not yet been proven for continuous multi-DoF upper limb control, which requires a greater variety of behavior at higher temporal precision and ostensibly poses a greater challenge to model stability.

Sensory BCI challenges: In contrast, sensory BCIs currently struggle less in model stability and more in building precise models of evoked sensations to begin with. That is, while any given stimuli's evoked percept location and quality appear relatively stable over time [30], how these features of the evoked sensations relate to electrode choice and stimulation parameters are difficult to predict [31]. The consequence is that sensory BCI capabilities are currently developed by collecting a library of responses to varied stimulation patterns and selecting a subset of relevant patterns for use in functional tasks. One promising framework to guide future stimulation design is the principle of biomimicry. Biomimetic ICMS asserts that stimulation should be designed to evoke neural responses resembling those of natural touch, rather than directly matching stimulation frequencies or amplitudes to physical parameters [32]. However, this aspiration is hamstringed on both ends, as it is difficult both to observe the neural response to ICMS and to observe the neural response

to natural touch in BCI participants who typically have impaired somatosensory pathways. Current demonstrations of biomimetic ICMS have compromised by matching coarse, subject-general features of natural touch and assuming the neural response to ICMS is proportional to either frequency or amplitude inputs [30, 33, 34]. Realizing biomimetic ICMS in higher fidelity through neural activity, and thereby presumably improving somatosensory BCI efficacy overall, will require addressing both these assumptions.

1.2 Deep network models of BCI data

Modern deep learning organizes applications research by the data being modeled, the model architecture, and the optimization objective. We will take this approach to introduce deep networks in BCI.

Data: The neural data recorded from our BCIs are multichannel voltage timeseries. Most traditionally, and in the majority of this proposal, this broadband voltage will first be filtered and thresholded to extract rapid, transient deflections, which mark putative spikes from a neuron nearby the recording electrode. A typical BCI dataset will include 100 to 200 electrode channels of data with dozens of high amplitude spiking waveforms, along with many more channels recording multiunit activity that is is not easily seperable into distinctive waveforms. Our motor datasets will also contain upper limb covariates, for example in the form of kinematics or electromyography that varies on the timescale of seconds. In non-human primate (NHP) datasets, these behaviors generally derive from physical sensors, but as mentioned in Section 1.1, a different strategy must be used to create human BCI datasets since human BCI users have impaired control of their native limb. Human BCI datasets instead assert a certain behavior and experimentally cue the user to attempt the behavior. This artificial association means that the behavior label is likely to be temporally warped and otherwise imprecise [35] relative to the unobserved motor intention of the user. The electrical stimuli that we use in sensory datasets comprise trains of biphasic current-controlled pulses, where each individual pulse's timing and amplitude can be varied. Perceptual reports to ICMS, to the extent discussed in this proposal, will be described in terms of binary detection or scalar ratings of intensity.

Architecture: The mainstays of broader deep learning architectures that have flourished in the last decade, including multi-layer perceptrons (MLPs) [36], convolutional neural networks (CNNs) [37], recurrent neural networks (RNNs) [38], and Transformers [39], have all been applied to BCI data. Unlike in other domains, BCI has yet to reach consensus on the most performant architectures. This proposal presumes the view of the Bitter Lesson [40], which states that performance differences across architectures may be minor relative to gains from increasing data scale. Adopting this view, we contrast here only the RNN, which many neuroscientists may have familiarity with, and the Transformer, which has empirically dominated the architectural landscape in machine learning (ML) domains, including even other timeseries domains like audio processing [41]. RNNs process

timeseries data one timestep at a time, maintaining an internal state that evolves with the input data. An RNN provides an appealing interpretative lens for neural data, as its iterations describe the evolution of the neural dynamical system. The Transformer, in contrast, does not maintain a centralized state but instead learns the relationship between fragments of the data. Each of these fragments is called a token, and they could be for example the population activity vector at the start of an experimental trial, abstract metadata about the experiment, or the vertical velocity of a user's arm.

Objective: This proposal focuses on predictive objectives, namely regression and classification. This will be true despite variety in the qualitative function of our different models. When the work's objective is to model neural data in isolation, we say we are interested in the model's representation learning. At other times, we say we are building decoders of neural data to predict behavior, or encoders of stimuli to predict neural activity. With the lens of deep networks in particular, these different terms all implementationally overlap, in that the underlying models are operating between one or two data modalities.

Challenges: BCI datasets feature a number of challenges that uniquely interact with deep learning. First, by experimental design, there is often a large component of the neural data that appears strongly connected to the covariate of interest, implying there is not an immediately clear role for nonlinear DNNs. For example, in the motor cortex, we can identify single neurons with firing activity that can be well characterized by a cosine tuning curve with two parameters, a preferred arm direction and depth of firing rate modulation to movement in that direction. To balance this simplicity, BCI datasets will often have nonstationarities [42–44], as discussed in the introduction. New cognitive states, neurophysiological changes at the electrode interface, and BCI system hardware state can all introduce shifts in the observed activity. These externalities are difficult to directly observe, so we must build systems that are implicitly robust to these shifts so as to enable stable BCI performance [27].

Foundation models: A modern trend in deep learning applications research is to create a model that provides broadly competent performance across tasks in a field, at which point the model can be considered a "foundation model". Foundation models were most deservingly coined in the context of natural language processing (NLP) [45], where a large scale initial training phase (pretraining) across internet text created models that quickly dominated all mainstream benchmarks [46]. Critically, model performance correlates reliably with the scale of pretraining, providing the impetus for many fields to also measure the empirical "scaling" of model performance with their own domain's data [47]. The merit of this paradigm for neuroscience writ large is still under debate, but this has not deterred a number of different efforts to create these large models in BCI domains [48]. Aim 1 discusses precisely this effort for intracortical motor BCI data.

2 Aim 1: Large-scale pretraining for intracortical neural datasets

2.1 Summary and Significance

Deep learning's greatest successes have depended on exploiting large and varied datasets. None of these efforts have occurred in a vacuum, but rather, have depended on the gradual scaling of model training and continual, critical evaluation. Growing the deep learning paradigm in neural data will also require quantifying the benefits from scaling the size and diversity of data used in model training. Fortunately, we approach this problem at a time when the neuroscience field has already identified structural relationships in data collected across time periods, subjects, and behavioral tasks, which delineate a natural scope for model training data.

We perform two scaling studies on Transformer pretraining across motor cortical datasets. In the first study, we measure the returns on increased pretraining data along the three aformentioned axes of data diversity. We find that, under the lens of the model we used, cross-session data scales nearly as well as data collected on the same day, and that cross-subject and cross-task data lag behind but still have productive returns to scaling. The second study executes on this premise, fitting a single large Transformer to 2000 hours of neural activity pooled from over 40 implanted monkeys and humans. Here, we see continued benefits from increasing pretraining data, but add the important caveat that benefits decline inversely with increasing data availability in deployment settings.

Significance: This work provides a baseline quantification of the benefit of scaling deep learning for motor cortical decoding, grounding narratives on currently feasible data scales and projecting returns on larger ones. To this end, we show that the standard expectations of pretraining are met, in that scaling up provides increased data efficiency in a broad class of new datasets. Moreover, the work dissects the distinct benefits of the different axes of neural data diversity, which all scaled pretraining on neural data must contend with. Under our Transformer models, each neural dataset has considerable individual variability, and transfer learning drops off greatly when using data from different subjects. Thus, while pretraining with up to 2000 hours of data is useful, it does not qualitatively reduce the data needed to achieve high performance in most downstream settings, and therefore does not fundamentally alter data collection strategies for high performance BCI.

Papers.

- Ye *et al.*, Neural Data Transformer 2. NeurIPS 2023.
- Ye *et al.*, A Generalist Intracortical Motor Decoder. (In submission)

2.2 Approach

The full pipeline needed to prepare foundation models like language models is complex beyond the scope of this proposal. Here, we study the earliest stage of model preparation, known as pretraining, and the pretrained model's immediate adaptation to a new setting, known as fine-tuning. Pretraining requires the availability of some large and loosely related volume of datasets in the domain of interest, while fine-tuning assumes the availability of some small amount of data that is much more closely related to the evaluation data. For example, if we were to assess a language model's ability to produce scientific-sounding text, the pretraining dataset might be all of Wikipedia, or all of the Internet's language content, and the fine-tuning data may be given as an article from a recent journal. For neural data, we must shortly define what the loosely related pretraining data might mean relative to our target domain of intracortical human BCI. In fine-tuning, we use a random temporal subset of our actual evaluation datasets.

We aim to characterize the impact of both data and model design parameters on final model performance. In both pretraining and fine-tuning, the data varies qualitatively in how training data and evaluation data relate, and quantitatively in the training data volume. The deep learning literature identifies two model factors as critical in scaling studies: the number of parameters in the deep network model and the compute used in training the deep network.

Data: These studies use previously collected datasets that were either released publicly or internal to Rehab Neural Engineering Labs (RNEL) and our collaborators. Data varies in the lab and experimental hardware that was used for collection, in the NHP or human subject that produced the data, time period of collection, and behavior in the task. The behaviors largely comprise reaching and grasping from experimental paradigms containing short, repeated trials, but still contain large diversity in the number of experimental conditions and behavior sensors.

Models: The great variety of datasets we aim to model at once in pretraining requires flexible model designs, as the identity and count of neurons in each dataset's population activity will differ across datasets. Here we discuss models based on the Transformer to meet this challenge. In Fig. 2.1, we review the Transformer's data processing in the specific contexts of the NDT2 and NDT3 architectures. The Transformer model requires input data to be fragmented into a number of tokens. Once these tokens are defined, they are simultaneously processed through a stack of alternating attention and MLP layers. Attention layers updates each token as a function of all tokens, while MLP layers update tokens individually. NDT's major design obligation is to decide how to perform this tokenization. In both models, we tokenize spiking timeseries in time with a 20 ms binning operation, and in space with a patching operation. This patching operation divides the N-dimensional input, i.e. all electrodes in a subject, into groups of K channels, padding the input with zeros to resolve remainder channels. This design follows Transformers in computer vision, which similarly tokenizes 2D images into smaller patches of pixels [49, 50]. By analogy, smaller patches should increase model expressivity but incur increased computation and data requirements.

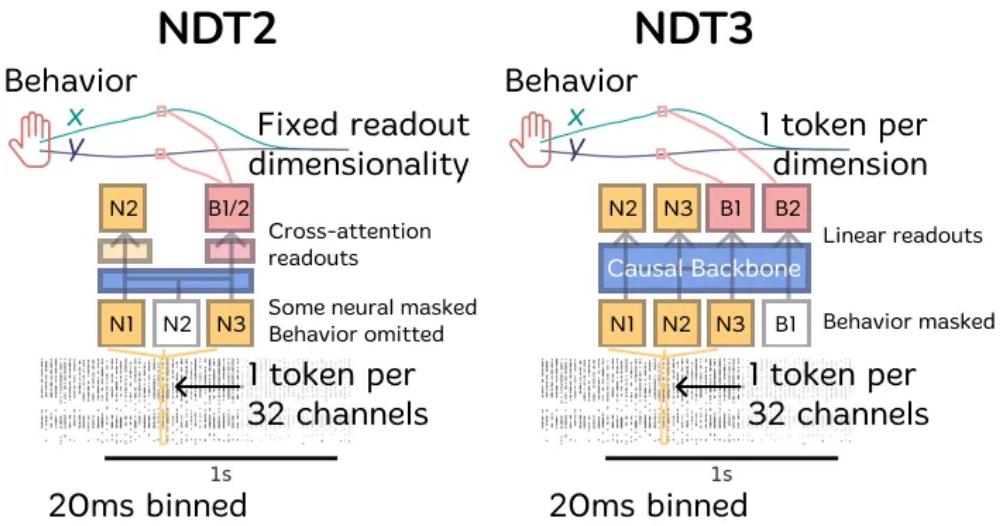

Figure 2.1: NDT2 and NDT3 both take input neural data and output neural data and behavior predictions. Both models tokenize neural data in time with 20ms timebins and in space by dividing the population activity into subsets of fixed size (32 in this figure). NDT2 emits one behavior token for prediction at each timestep, while NDT3 emits one behavior token per behavior dimension and timestep, which enables streamlined prediction of behavioral data of varied dimensionality. NDT2, inspired by a Masked Autoencoder design [49], masks out a fraction of neural token inputs and only predicts this fraction. NDT3 adopts the autoregressive modeling framework and allows prediction of all neural tokens conditioned on neural tokens from previous timesteps.

Note that the two model designs reflect their distinct motivations. NDT2 was developed to evaluate whether DNNs could learn transferable representations of neural spiking activity. It thus exclusively accepts neural data as input and extracts behavioral representations at each timestep with a linear layer of fixed output dimensionality. NDT3 was designed to accept highly variable data in order to scale pretraining as widely as possible. To do so, NDT3 accepts unidimensional behavioral tokens as inputs in order to specify how many behavior dimensions the model needs to predict.

2.3 Results

NDT2 shows that DNNs can transfer learn across datasets of spiking activity from different sessions, subjects, and behaviors. Prior work to NDT2 demonstrated this transfer can be achieved with explicit mechanisms to align data from one setting to another, such as canonical correlation analysis [51] or joint training with alignment layers [38, 52]. NDT2 demonstrates the same is possible merely through its population patching scheme. In subsequent experiments, we show the relative value of scaling pretraining trials from the different data sources, showing that cross-session data transfers much better than equivalently sized cross-subject or cross-task datasets (Fig. 2.2A). Nonetheless, the positively scaling benefits from increased pretraining on all these varied datasets suggests that we should expect benefits from aggregating a large and diverse pretraining dataset.

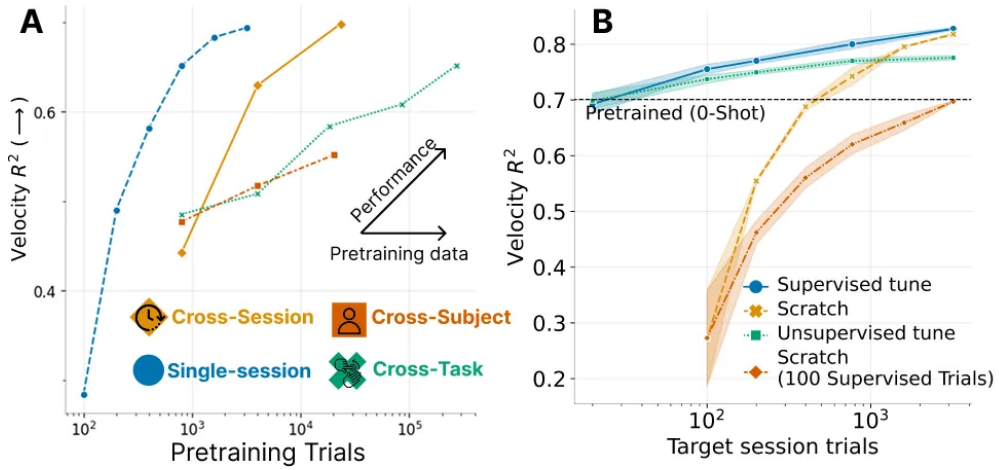

Figure 2.2: A. NDT2 decoding of monkey reach velocity scales with increasing pretraining data from different sessions, subjects, and behavioral tasks. The pretraining abscissa for the single-session curve indicates the total training data available to the model. For the order curves, the abscissa indicates the pretraining data scale. These pretrained models fine-tune to the target session with 100 trials of data. B. Pretrained multisession models can be deployed on new sessions for immediate zero-shot performance. They can also be tuned either through supervised or unsupervised objectives to improve performance, which outperforms models trained from scratch across tuning dataset size.

Fig. 2.2B shows that a pretrained NDT2 model retains high performance for immediate use on a new day, without any data renormalization (0-shot). Further, NDT2 can be fine-tuned for a new day either through unsupervised or supervised calibration. This strategy outperforms models that are trained from scratch at all fine-tuning data scales, so the implication for BCI deployment is that pretraining should always be used when possible.

NDT3 builds on NDT2's scaling results and pretrains Transformers on up to 2000 hours of motor cortical spiking activity and behavior. We conduct an evaluation commensurate with this pretraining by fine-tuning pretrained NDTs on eight downstream datasets varying in behavior, subject, and length. Fig. 2.3 quantifies the two expectations we have of pretraining. First, we expect that pretraining should be beneficial at low downstream data scales, and eventually converge with from-scratch models once downstream data scales are high. Fig. 2.3A shows that in our evaluation, this convergence point is around 90 minutes. Second, Fig. 2.3B shows that scaling pretraining inputs should yield improvements in downstream tasks, while in the <90 minute regime.

Data efficiency evaluations have shown that NDT3 has learned generically useful priors for predicting motor behavior. We next evaluate generalization to test whether these priors also yield alignment with our normative desires for neural data models. As an example, we show in Fig. 2.4 how models trained on trialized data generalize to continuous data. In this analysis, the training and evaluation data only differ in formatting. Trialized data is segmented and presented to the model so that the model's first and last timestep always coincide with the endpoints of a reaching movement. Continuous data presents random 1 second chops of this same data. Fig. 2.4 shows

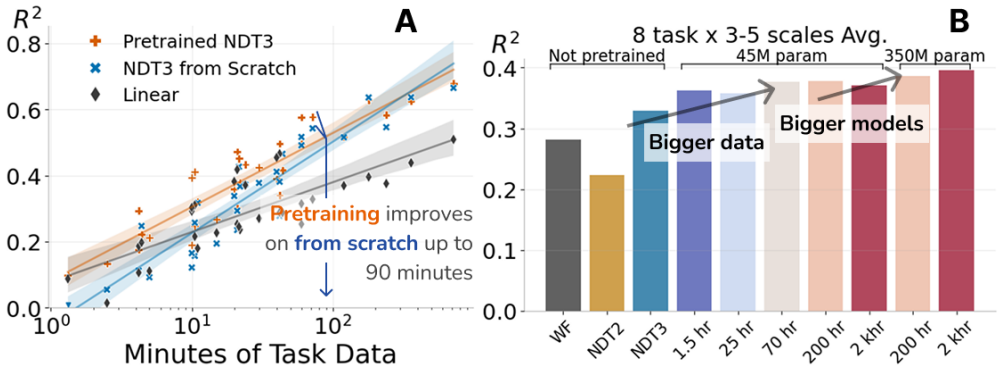

Figure 2.3: NDT3 evaluations. **A**. When evaluations are organized by downstream data scale (Minutes of Task Data), a pretrained model is better than or equal to non-pretrained models and a linear baseline at all data scales. However, the non-pretrained NDT3 model matches the pretrained model when downstream task data is sufficiently large, here at 90 minutes. **B**: Collapsing all evaluations into a single average, we see that scaling pretraining dataset size and model size improves summary performance.

that when trained on trialized data and evaluated on continuous data, from-scratch models degrade substantially while pretrained models only degrade minorly. As desired for BCI control, pretraining may guide models to decode in a more robust and generalizable manner, regardless of neural data formatting.

NDT3's results are limited in that we only evaluate the specific tokenization mechanism validated by NDT2, and the ratio of data between pretraining and fine-tuning is far smaller than foundation models in other domains, which likely drives the rapid convergence of pretrained and non-pretrained model scores. Nonetheless, it provides a baseline for future improvements, derived from innovations on model design or the next order of magnitude of pretraining data. For deployment to online BCI control, NDT3 clearly provides a better pretrained model than NDT2, though our extended generalization analyses show that our offline evaluations only scratch the surface of aspects of model behavior relevant for online control.

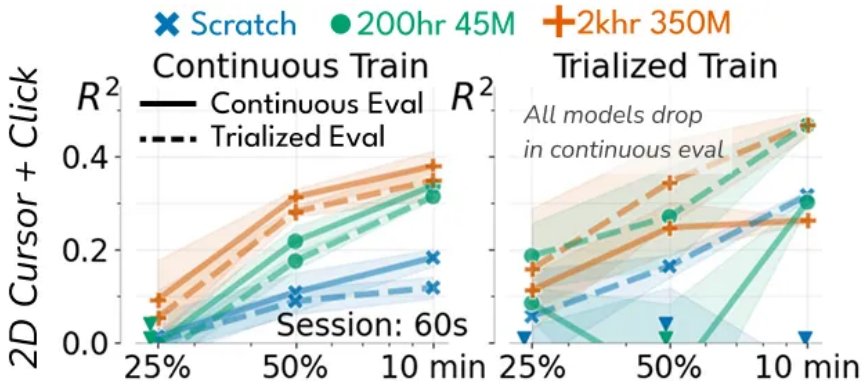

Figure 2.4: Models are evaluated on a human open-loop cursor dataset prepared in two ways. Trialized training receives inputs according to trial boundaries, varying from 2–4 seconds in length. Continuous training receives random 1 second snippets (that can cross trial boundaries). Trialized evaluation matches trialized training, and continuous evaluation is done by streaming up to 1 second of history. ▼ indicates points below 0.0. Continuously trained models perform well in both evaluation settings, while models trained on trialized data fail in continuous evaluation.

3 Aim 2: Pretrained deep networks enable continuous and scalable upper-limb neuroprosthetic control

3.1 Summary and Significance

Motor cortical BCIs record neural activity that can be used to control a variety of upper limb behaviors, like arm translation, wrist rotation, and hand grasp. However, this high degree-of-freedom (DoF) control is currently difficult to maintain, requiring daily recalibration and frequent experimenter intervention. The increased complexity of high DoF BCI control highlights a basic need for BCI models that can be flexibly developed, maintained, and extended. We propose tuning pretrained NDT3 for this purpose, given its ability to aggregate data of potentially varied formats, from multiple days, into a holistic controller. We will measure NDT3's ability to provide day-over-day performance of 7 DoF, and then 9 DoF control of the upper limb in two human BCI participants. To support this goal, we have conducted pilot assessments of online control. We show that for 2D cursor control, a fine-tuned NDT3 performs comparably with linear baselines, and for 4D virtual limb control combining translation, rotation, and grasp, NDT3 outperforms linear baselines, and improves with data across days.

Significance: This aim assesses whether data aggregation, through its sublimation into the NDT3 model, can meet two critical needs for upper limb neuroprosthetics. First, we assess whether data aggregation can enable continuous and convenient high DoF upper limb control. A reliable and minimally burdensome system is necessary to translate upper limb neuroprosthetics from the lab towards independent device use. Second, we assess whether aggregation is useful to enable the extension of an existing BCI to new behavioral repertoires. If successful, these results would further the case for BCI models that accumulate calibration data across the device's lifetime.

Hypotheses:

- Scaling calibration data will enable performant high-DoF DNN-based BCI control.
- Data aggregation across multiple days will reduce the need for extensive per-day calibration.

3.2 Approach

We aim to use an NDT3-based deep network controller to enable convenient high DoF control. To this end, our first goal is to enable quick high DoF control across days, by aggregating calibration data across days. Our second goal is to assess the viability of amending and extending control through targeted calibration, primarily by adding more DoF to control.

3.2.1 Rapid calibration on new days

The requisite time to sample sufficient behavioral diversity to calibrate a high DoF controller grows with behavior dimensionality, reaching 15 minutes for 7 DoF [12] and 20 minutes for 10 DoF [53]. However, this calibration burden need not be repeated daily, as the relationship between neural activity and behavior only partially changes across days. In speech BCIs and simpler movement BCIs, decoders trained with multiple days of data can be adapted for high performance on new days with minimal, if any, calibration [21–23, 27]. We aim to replicate this strategy with high DoF motor control. We will begin these experiments with 30 minutes of calibration in each of the first two sessions, which has historically enabled at least 80% success rate under brain control in the virtual task environment. On new days, we will then alternately collect calibration data and evaluate control in increments of 5 minutes, until the same 80% success rate criteria is reached. Our expectation is that new days will require at most 2 of these calibration blocks.

Evaluation: To compare with prior studies, we will evaluate functional performance with timed object transfer trials and the Action Research Arm Test (ARAT), for comparison with prior results in high DoF upper limb control [12, 53]. Both protocols assess grasping and carrying competence. To evaluate the continuous stability of system performance, we will target at least 8 experimental sessions over the course of a month. In a subset of these sessions, we will compare against the previous state of the art, 7D control with linear controllers. Due to restrictions on experiment time, comparisons against other baselines and system design choices will use offline prediction-based analyses. In addition to these functional evaluations, we will include a number of virtual brain-controlled reach-grasp-carry sequences in the same setting as the virtual calibration data. These virtual evaluations will allow us to assess more fine-grained control metrics like path efficiency and movement phase-wise performance.

3.2.2 Scaling to higher DoF control

We initially target 7 DoF for robot arm and hand control. For these 7 dimensions, 3 dimensions will be for hand endpoint translation, 3 dimensions will be for wrist rotation, and 1 dimension will control whole hand power grasp. This control scheme has been historically used in our lab and is readily supported by current robot arm and hand capabilities [12]. However, dexterous upper limb control will require more degrees of freedom in the hand.

To expand hand control, we will replicate our strategy for obtaining 7DoF control. We will first collect open-loop calibration data where the arm and wrist are in neutral positions and different finger degrees of freedom that have been decoded in other human BCI trials are varied [20, 54]: the flexion and extension of individual fingers and groups of adjacent fingers. For the thumb, we will additionally assess abduction/adduction. Given the strong separability of the thumb from the other fingers in humans with implants in the hand knob area, we expect to achieve at least 3 DoF control in the hand with this strategy (2D thumb, 1D for the other 4 fingers as one group). For the

selected hand DoFs, we will then perform a unified calibration task that requires the participants to sequentially control each DoF, analogous to the 7D grasp and carry task.

Finally, in preparation for combined control of the hand, wrist, and arm, we will collect a reduced series of hand calibration datasets in different arm and wrist postures, and arm/wrist datasets in new hand postures. If these datasets are modeled well by decoders trained on neutral-posture calibration datasets, we will proceed with combined control evaluations. If not, we will collect full length calibration datasets in these new combined conditions For simplicity, as for 7D control, we do not plan on collecting calibration datasets where multiple DoF are varying simultaneously.

Evaluation The newly enabled hand control will be evaluated in virtual environments. Preserved control of arm and wrist will similarly be evaluated by comparing control in virtual 7D grasp and carry tasks with and without tuning on the new hand calibration data.

If individual virtual evaluations are successful, we will assess combined arm and hand control with two robot-based functional evaluations. We will begin with ARAT. Previous work extending hand control from whole hand grasp into 4 hand-shape basis dimensions [53] showed that the BCI user could still accomplish the ARAT task without the new dimensions, and rather preferred not to use the extra dimensions. Overall, the extra dimensions did not provide higher clinical gain of function relative to the 1D whole hand grasp. Given these results, we do not expect higher ARAT scores with increased hand DoFs. Thus, we will also evaluate object transfer trials of Southampton Hand Assessment Procedure (SHAP) objects [55], which require more precise hand postures beyond power grasp.

3.3 Results & Remaining Work

We have collected pilot experiments that demonstrate the efficacy of NDT3 and its use in multiday aggregation for lower dimensional control, namely for 2D cursor control in three participants and for 4D grasp and carry tasks in Mujoco with one participant.

2D Cursor control: We have compared NDT3 against linear controllers for 2D cursor control with three participants, each in two separate sessions. These controllers are directly created from training on open loop calibration data alone, without further closed loop tuning. Qualitatively, the NDT3 model often produces slower, steadier movement compared to the linear controller (Fig. 3.1A), whereas quantiative measures of path efficiency and acquisition time show rough similarity between the linear baseline and NDT3 models (Fig. 3.1B). Given the qualitative difference in control, we note that this measured equivalence likely depends on the particular distance and precision requirements in our evaluation task [56].

Cursor controllers are often adapted after blocks of closed loop control to improve performance [56–58], and so evaluating models after blocks of closed loop tuning likely better reflects real world performance limits. We use the ReFIT intention estimation strategy [58] to enable closed loop

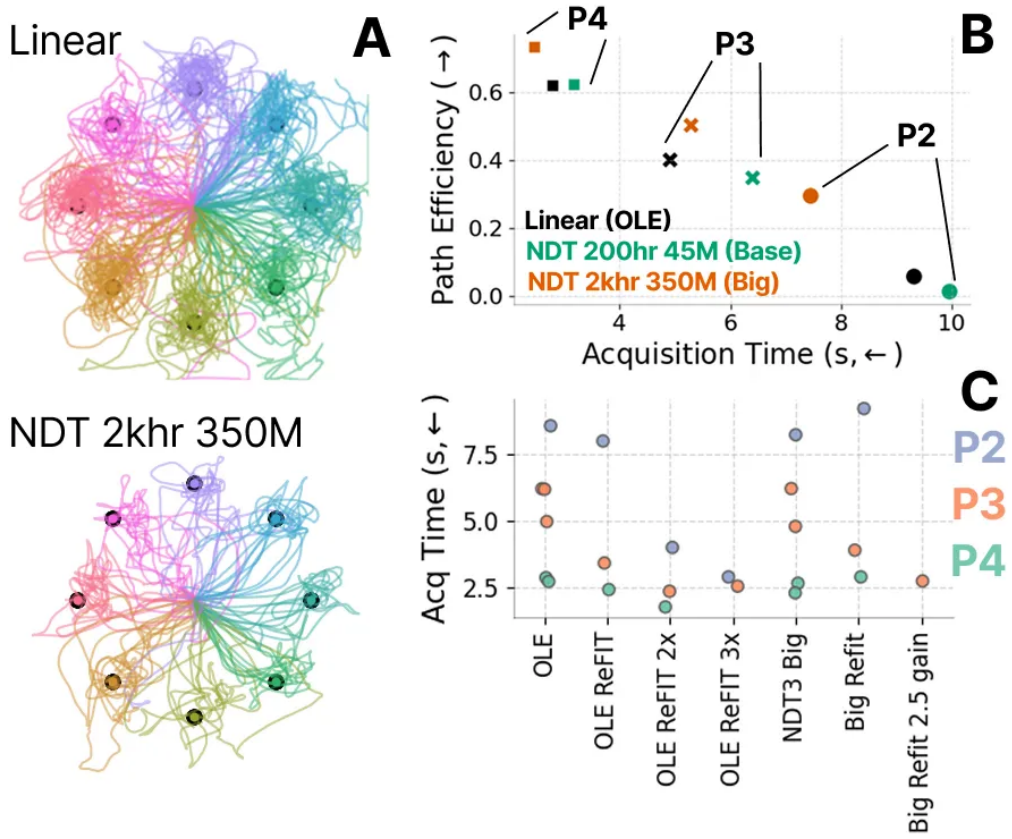

Figure 3.1: A. OLE and NDT brain-control trajectories in one human participant. B. The largest NDT model matches linear model performance for 2D cursor control when all models are trained on an open loop calibration block. B. With closed loop tuning, linear control steadily improves, but NDT does not. This gap can be closed by increasing NDT's gain manually.

adaptation of both linear and NDT controllers. We replicate that control reliably improves with ReFIT across participants, but that NDT control does not improve as significantly (Fig. 3.1C). NDT does not benefit as greatly from reducing angular decoding errors and experiences a degradation in control gain. Observing this, we observed greatly increased performance from manually increasing NDT gain. We speculate that NDT suffers more significantly from the label noise introduced across the multiple calibration datasets. This result motivates care in the use of closed loop tuning for NDT. We do not anticipate this result to prohibit our high DoF study, as historical high DoF experiments were reported without intention estimation, and closed loop tuning does not majorly affect performance in our most recent participant with the highest signal quality recordings (P4), which is the setting we will study for high DoF control.

4D Mujoco control: We have also collected several sessions of 4D control, with 2 arm translation dimensions, wrist roll, and whole hand grasp, calibrated and evaluated in the Mujoco simulator. All trials are sequenced movement tasks that have distinct phases that require isolated translation, rotation, and grasp. These experiments derisk two aspects of DNN control we target in our high-

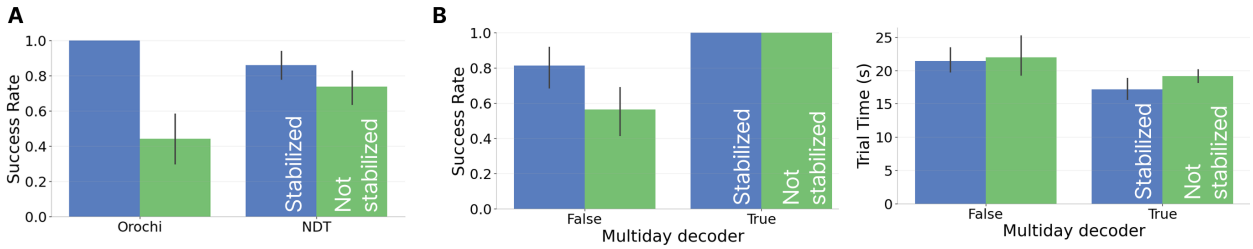

Figure 3.2: A. Orochi linear decoder compared against NDT DNN decoder for Mujoco 4D FBC Sequenced movement trials. Error bars show standard deviation across trials. Evaluations are conducted in default (not stabilized) and stabilized settings. Stabilized evaluations only allow one of arm translation, wrist rotation, or hand grasp to be active at a time. B. Multiday NDT decoders achieve high success and faster completion times relative to decoders trained from scratch.

DoF experiments.

First, NDT provides distinctive stability benefits for high-D control. By stability, we specifically mean it is hard for the virtual effector to remain on target for a sufficient holding period to pass the trial. This failure mode degrades the linear decoder in regular evaluations where all DoF are allowed to vary in all phases, despite good performance in a stabilized evaluation where irrelevant DoFs are frozen in each phase. This lack of stability is a known phenomena of linear models and has motivated prior works to explore nonlinear output scaling for 2D cursor control to enable easier stopping [23, 59]. The increased severity of instability at higher dimensions has not been saliently discussed in prior works, perhaps because no alternative was available at the time. In contrast with linear decoders, NDT does not suffer a large performance drop from removing movement constraints, due to its inherent minimal movement in off-target dimensions. This stabilizing aspect of DNNs is corroborated in DNN-enabled 4D bimanual control [60], and suggests that DNNs like NDT will provide qualitatively superior control over linear controllers in high dimensions due to a tendency of DNNs to isolate DoF, despite their quantitative similarity for cursor control. For clarity as to why this happens, we can turn to two monkey studies [37, 61] that suggest that RNN decoders will tend to more precisely replicate training label distributions, which in our case have isolated translation, rotation, and grasp. This result does raise the potential of requiring calibration that varies multiple DoF simultaneously, though that has not yet limited our 4D brain control experiments.

Second, NDT supports multiday adaptation for arm and hand control. Fig. 3.2B shows these brief multiday results for 4D, showing 2 evaluation days where the NDT decoder had either 2 or 3 days of full calibration data. Thiese multiday decoders enjoy a moderate performance gain by both success rate and acquisition time. We take this as evidence that a multiday NDT decoder will likely sustain a fixed level of performance with reduced calibration data.

The main high-D experiments have not yet begun. The main set of proposed experiments will likely only be viable to test with a new participant we expect to implant in late August due to array degradation or participant health in our other participants. We must first show that we can

achieve day-over-day 7D control for Mujoco environments and robot arm control, and that the stated stability properties are beneficial for functional tasks. We target demonstration of sustained control for up to a month. Conditions permitting, we will try to enable sustained 7D control for a second participant with arrays that are 2 years old. I will then target finger control calibration and evaluation, anticipating around 6 sessions of data collection.

4 Aim 3: Modeling the neural response to intracortical microstimulation

4.1 Summary and Significance

Today's neural interfaces provide the ability to intervene on ongoing neural activity through stimulation. This ability can be valuably applied in intracortical BCIs to provide sensory information about the user's environment. The evoked sensory response to different stimulation parameters can be characterized and composed in simple manners to convey basic tactile features, and integrated to improve grasping efficiency [18, 33]. The challenge is to now advance beyond this initial characterization and to more thoroughly relate the exponentially large and artificial parameter space of electrical stimulation patterns to the high dimensional and qualitative space of sensory perception.

To enable a data-driven approach for this challenge, we must augment stimulation characterization beyond the current norm of noisy and slow perceptual reports. This can be done by decomposing the current stimulation modeling problem into two: relating stimulation and the evoked local neural response, and relating the local neural response and subsequent percept. This decomposition strategy relies on two critical assumptions. First, given the spatiotemporal locality of the immediate neural response to ICMS [62–64], we assume that the controllable variability in the subsequent perceptual response to stimulation is mediated by the locally observed neural response. Second, to scale data collection for the neural response model, we assume that the neural and perceptual response to passive task-free stimulation is informative of the response in functional task contexts. With these assumptions, we also defer the challenge of relating the neural response to perceptual response, and focus in this proposal on two challenges in modeling the local neural response to passive stimulation.

First, we introduce a new method for electrical artifact removal that allows the recovery of spiking activity on the stimulating electrode array, where neural responses relevant to the perceptual report are likely to be. Second, we characterize how the neural response to different spatiotemporal stimulation patterns relate to each other, under the lens of transfer learning. We identify that responses remain largely predictable under temporally varied stimulation, but responses across different channels are idiosyncratic. These results inform a strategy for accumulating a model of the neural response to stimulation through passive stimulation.

Significance: Modeling the neural response to stimulation may provide a critical foothold to developing sophisticated intracortical stimulation strategies. Our artifact removal method enables this, by enabling the recovery of short-timescale neural responses. The latter study, which aims to taxonomize the neural response to passive stimulation, provides basic recommendations for how

to build a tractable model of the neural response to stimulation. The utility of such a model will rest on the strength of the relationship between the neural and perceptual response. For example, our neural taxonomy identifies channel specific idiosyncrasies in the neural response to ICMS, which is also reported at the level of behavioral response [30, 31]. If the two phenomena can be connected, it would imply our neural response model should greatly inform our ultimate functional goal of modeling the perceptual response to stimulation.

Hypotheses:

- DELETE provides a generic method for artifact removal that enables recovery of the neural response to ICMS.
- The short term neural response to ICMS is predictably structured with respect to varied stimulation pulse time and amplitude, but idiosyncratic across channels.

4.2 Approach

Recovering the neural response with DELETE

The first obstacle to modeling the neural response to stimulation is simply observing the neural response to stimulation. The hardware required to enable simultaneous electrophysiological recording through stimulation has only recently arrived in NHP and human labs, and on these hardware, large electrical artifacts will often distort the putative spikes extracted by conventional signal processing. The artifacts induced in our experimental setup are particularly severe and warranted the development of a new artifact rejection approach which we call DELETE (Denoising Electrical Events with a Transformer Encoder). DELETE is a nonlinear generalization of population reconstruction-based methods including ERAASR [65] and linear regression rereferencing (LRR) [66]. We will show that DELETE outperforms these methods and other baselines both in recovering peri-stimulation neural responses and in generically denoising broadband activity.

DELETE Design: DELETE is a deep network optimized to reconstruct broadband multichannel neural activity. Reconstructed activity is assumed to be artifact due to its rough similarity across channels, and subtracted to recover putative neural activity. While similar reasoning has motivated previous PCA-based methods [65], DNNs have renown as universal function approximators, which makes it unclear whether DELETE might also reconstruct and remove neural activity in addition to electrical artifact. We rely here on info-theoretic intuitions to constrain DELETE's behavior, restricting its input window (e.g. 6 ms) and forcing its training across broad data under a point-estimate loss (mean-squared error). Combined, these restrictions imply a network could only erase neural activity by knowing the precise firing characteristics of a specific channel when deployed on an arbitrary new 6 ms window. Ultimately, like other artifact rejection methods, DELETE is best justified empirically, from our broad validation.

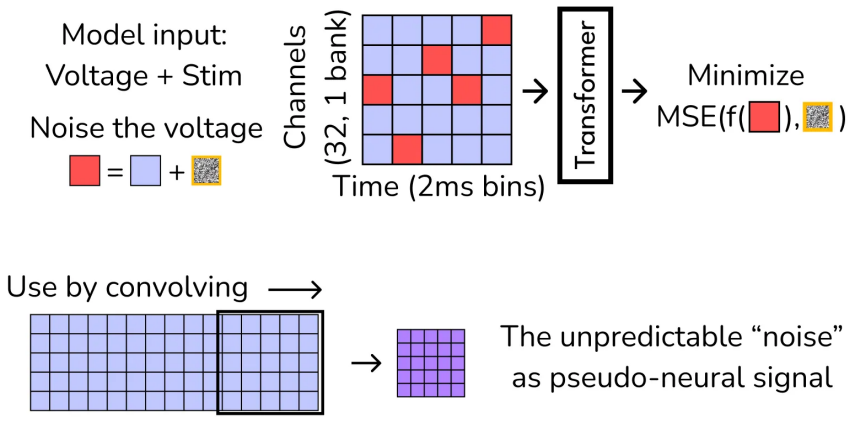

Figure 4.1: DELETE schematic. DELETE is a denoising autoencoder that is optimized to predict the noise component of data that has been injected into fractions of its input. At test time, DELETE is applied by convolution on raw broadband activity, and the model's predicted noise components form putative de-artifacted, neural signal.

DELETE joins a number of related methods using deep learning to process raw scientific data [10]. DELETE relates to the use of autoencoding for anomaly detection to remove uninteresting background events, as in e.g. particle collision experiments, so as to highlight the residual signal. DELETE critically should be distinguished from similar-sounding denoising autoencoders [67, 68], which *preserve* the common component of multidimensional data and discard the residual as noise. Beyond this distinction in usage, DELETE is distinguished in design by its use of Gaussian noise to ease the learning of the task, as the artifact is otherwise too variable for satisfactory estimation. The model does not use fully masked inputs to directly reconstruct broadband activity, but rather receives input activity distorted by additive Gaussian noise, and must predict the injected noise.

Analysis: Our analysis of the DELETE method has three components. We first evaluate its performance as a stimulation artifact rejector, focusing on activity recovered in short periods after stimulation pulses. This analysis will be performed on a variety of ICMS datasets, discussed further in the next section. We next evaluate its performance as a generic denoiser, using synthetically noised datasets and datasets with movement artifacts. Finally, we conduct isolated studies of model behavior over input and architectural ablations, to build an understanding of model function.

Taxonomizing passive ICMS responses

To productively scale data collection to model the neural response to ICMS, we must first build a coarse understanding of the impact of different stimulation parameters. The full parameter space is exponential, but the presence of some structure in the behavioral response to ICMS suggests that there are natural axes of data diversity in the neural response as well. Thus, we taxonomize which sets of ICMS parameters evoke structurally similar responses, by collecting datasets of passive

ICMS stimulation and observing how models (RNNs or Transformers) trained on one subset of the data generalize to other subsets [69]. We first collect multiple single electrode ICMS datasets that vary pulse timing and amplitude, either using the fixed parameter trains commonly used in ICMS protocols to date, or with random amplitudes and Poisson timing. We also collect corresponding single and multi-channel datasets to assess whether we can back out performant models of single channel stimulation from multi-channel datasets. Finally, we scale spatiotemporally randomized multichannel stimulation trains over months to illustrate that multiday models can productively accumulate these diverse data, similar to our strategy for motor models.

4.3 Results & Remaining Work

Thus far we have trained and evaluated the DELETE estimator on ICMS datasets using both simple, single-channel fixed frequency and amplitude trains and multi-channel, spatiotemporally diverse trains. The taxonomy project has proceeded under the assumption that DELETE has properly de-artifacted our ICMS datasets. The requisite passive ICMS datasets were collected in 2022 to 2023, in 2 participants.

DELETE for ICMS artifact rejection: As we began passive ICMS data collection, we quickly observed stimulation artifacts in our data that were more severe than those typically seen in the literature. Specifically, our stimulation induces not only a large, high amplitude transient on each stimulation pulse, but also includes a post-transient distortion that decays on the timescale of a second (Fig. 4.2A). Our task is to remove this artifact so as to extract the neural spiking activity. Ideally, the extracted spiking activity will be verifiable based on the spike's waveform shape, though this may be challenging given that the spike is typically much smaller in amplitude than even the artifact's slow decay. An example of a channel with one distinctly preserved waveform is shown to the right. The challenge is to do this reliably, across stimulation parameters and channels. In Fig. 4.2B, we show example waveform recovery summaries from a single stimulation dataset for two different artifact estimation methods. Polynomial estimation has many high amplitude waveforms recovered in peri-stimulation period, even for otherwise very quiet channels, due to artifact leakage. We seek methods that have high correlation across channels and datasets, and DELETE greatly outperforms the other baselines we have developed. In Fig. 4.2C, we show an synthetic sensitivity analysis computed on a dataset processed by the DELETE method. To create these recall curves, we injected synthetic spikes into broadband activity at different offsets from a stimulation pulse onset. We prohibit spike recovery shortly after each pulse and in a timeframe around 3ms after each pulse due to high artifact severity, but outside of these blanked periods, nearly all injected spikes are recovered. We can compute area-under-curve metrics for different methods in this analysis, and DELETE typically outperforms baselines here as well.

DELETE as a generic denoiser: DELETE is a method that convolves over all timepoints and therefore applies to non-stimulation datasets as well. Its objective implies it may be a useful

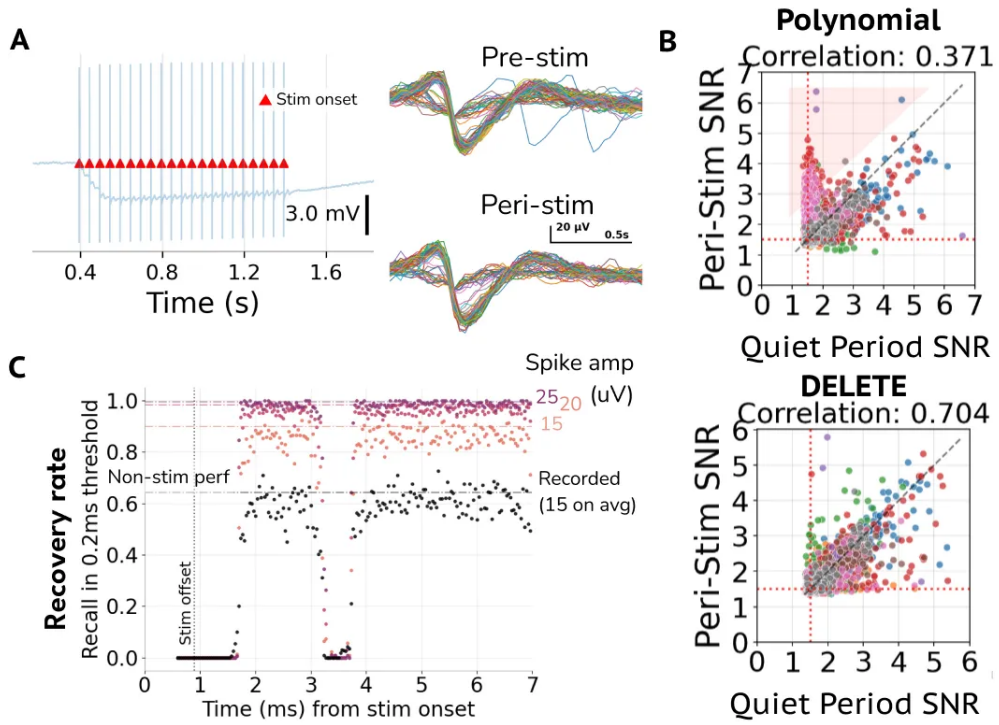

Figure 4.2: A. Left: A sample broadband recording from a channel during a stimulation trial. The displayed channel is on the electrode array that is stimulated. Right: Spike waveforms recovered from the broadband activity both prior to stimulation onset and in the artifacted peri-stimulation period. B. To summarize waveform recovery, we compute channel SNR as a ratio of peak waveform amplitude against background noise amplitude. The color of each dot indicates the physical array the channel is on. We expect physiological waveform SNRs to be conserved through stimulation, so high correlation of quiet period and peri-stim SNR provides a heuristic for good de-artifacting. Dashed red lines indicate the 4.5x background noise threshold we use to determine spike presence. C. We evaluate model sensitivity by injecting synthetic spikes into the broadband activity and evaluating whether they are recovered. The colored curves differ in the precise waveform injected, which were either recorded and extracted from the data or simulated. Horizontal lines indicate the noise ceiling computed as the recovery rate of spikes injected outside of stimulation periods.

generic denoiser to remove non-neurophysiological signals that degrade our recordings, so we try to characterize this intuition in Fig. 4.3. In panel A we show two raster plots computed from the same minute of resting state activity recorded from one of our BCI participants. The Butterworth filtered data show several streaks of transient high firing across multiple channels that typically indicate electrical noise caused by participant movement. The DELETE pre-filter produces an overall quieter view of this same activity, and notably also does not contain these periods of correlated firing.

The broad change in statistics may be concerning, but similar changes can occur even as the result of changing classical filter settings. We still cannot evaluate against any underlying ground truth, but we can rely on the principle that the activity of high amplitude and therefore high quality spikes should be well conserved. Fig. 4.3B and C show that DELETE does satisfy this requirement.

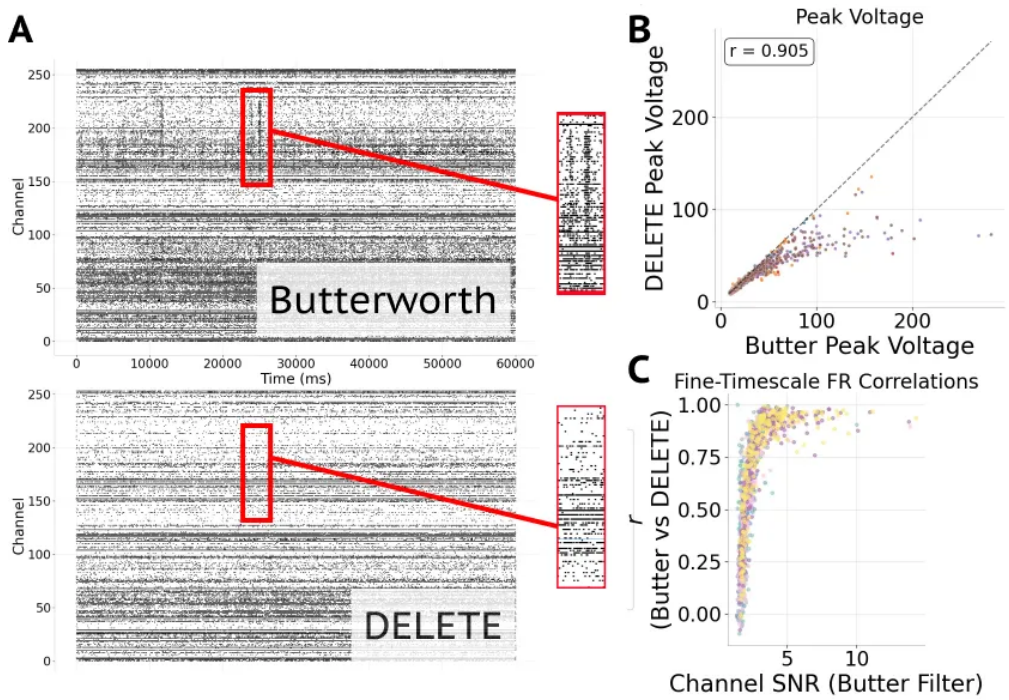

Figure 4.3: A. Top: Butteworth filter extracted spiking activity from a minute of resting state recordings. Inset shows putative movement-induced artifact. Bottom: The same data pre-filtered by DELETE, inset no longer shows correlated firing across channels. B. Peak voltages of waveforms extracted by DELETE and a standard Butterworth filter are compared. There is high correlation, but DELETE peaks saturate by 100μV. C. Firing rates are determined by convolving a 50ms Gaussian kernel against Butterworth or DELETE filtered spiking activity. These per-channel firing rates are correlated across a number of different channels and datasets. Color of dots indicate datasets. High SNR channels correlate well.

B shows that DELETE roughly conserves the peak voltage of spike waveforms, so large spike waveforms will remain large regardless of DELETE pre-filtering. There is a notable saturation in DELETE-processed waveforms at around 100μV, which likely corresponds to the amplitude of the noise used in DELETE's training. We can also show that single channel firing is conserved at a more detailed level, by correlating the firing rates identified with or without the DELETE pre-filter. In doing so, we see that as hoped for, high SNR channels again have highly conserved firing rate timecourses.

ICMS Taxonomy: To characterize how the neural response to different ICMS trains relate to each other, we collected a variety of ICMS datasets and modeled them with RNNs (chosen for efficacy at very small data scales). This flexible model-based characterization of data relationships are necessary as we are interested in the response to spatiotemporally diverse ICMS, as opposed to only fixed parameter trains. One such temporally varying ICMS train is shown in Fig. 4.4A. Modulated responses on both stimulated and non-stimulated arrays can be modeled with an RNN model. Panel B shows an example generalization assay that tests whether models trained on data from one family of ICMS data will generalize to another. Responses to random-amplitude poisson (RAP)

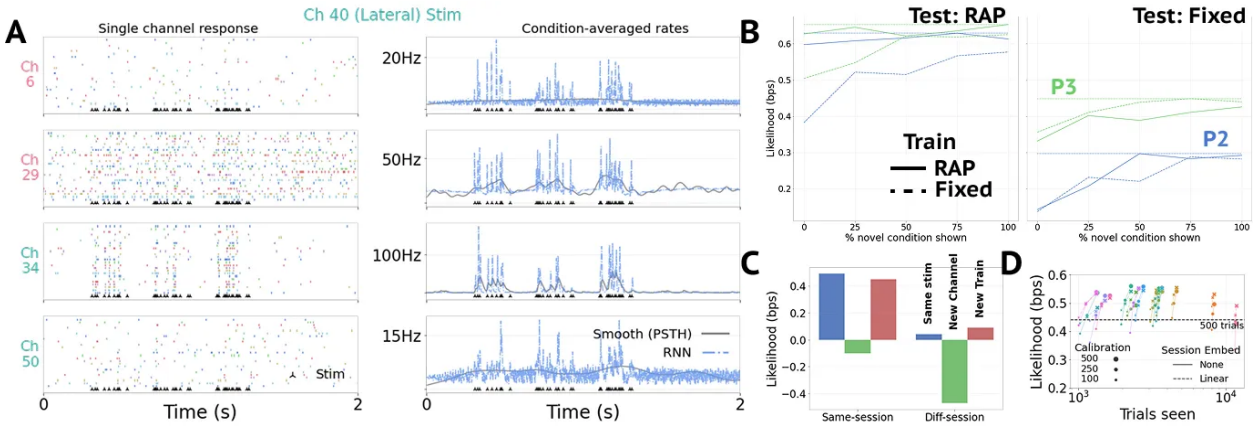

Figure 4.4: A. Example rasters to stimulation on a single channel (40). Modulated responses are visible both on the stimulated array (bottom two rows) and another sensory array (top two rows). Firing rates as inferred by computing PSTH and RNN models are given on the right. B. Example analysis testing generalization of RNN models to novel single channel stimulation trains. In these plots, models are first trained on either random amplitude Possion (RAP) or fixed frequency and amplitude trains (Fixed). They are then gradually exposed to increasing levels (% novel condition shown) of a new set of ICMS again from either RAP or fixed stimuli (Test condition). The flat dotted line marks the max performance achieved by any in-distribution model. The closer the other curves are to this flat line, the better the model's generalization to new stimuli. C. We train one model on one type of RAP ICMS and assess its generalization to 3 stimulation conditions from either the same or different experimental session. D. We conduct a brief scaled pretraining analysis. We vary the pretraining scale (Trials seen) and subsequent fine-tuning scale (Calibration), and evaluate on a new session. Session identity is either not modeled (Session Embed None) or modeled with a linear readin layer (Linear). The reference line shows session performance using only 500 calibration trials from the evaluation dataset. Most pretrained models exceed this level, suggesting positive transfer.

trains are well-predicted if models have already trained with other RAP trains, but generalization to unseen fixed parameter trains is poor regardless of whether the model has first trained on RAP or fixed parameter ICMS. We have run a number of similar analyses testing generalization across stimulation channels and timing, and tentatively believe that RAP trains provide a good basis to learn the broad response to stimulation on a given channel, but there appears to be little correspondence between the evoked response on different channels. The next two panels address model generalization across days. Fig. 4.4C first restates that generalization to repeats (Same stim) and to new RAP trains (new train) is high, but generalization to new channels is poor. The same relative ordering holds when the data is collected from a different experimental session, but all scores are greatly reduced. Fig. 4.4D shows that despite this session-wise drift, we can train an RNN model on increasing volumes of ICMS data to achieve high model performance on new days. The decline in performance at the high end of pretraining scale currently challenges a straightforward ICMS foundation model built with RNNs, and prompted the long detour to develop multisession modeling capability with the NDT models. We have yet to assess whether NDT models scale resolve these scaling difficulties.

The remaining work for DELETE is to polish the existing per-dataset profiles into a summary

of DELETE's efficacy. This entails characterizing a polynomial estimator baseline and linear regression rereferencing, and developing a synthetic battery. We also aim to ground DELETE's performance in plausible neurophysiology. We can verify this for the generic denoising case by applying DELETE to motor datasets and verifying it conserves population structure and decodable information. For the ICMS case, we will check that DELETE's recovered distribution of evoked responses resembles those from external datasets collected in settings without our severe artifacts.

The taxonomy work is in a similar state, where the technical proof of concept is nearly complete, but the analysis needs extension to many datasets. As we already have one more participant and may soon have another, we may consider further data collection to improve the strength of our conclusions. For example, to strengthen the relevance of this study to rehabilitative BCI, we should demonstrate whether the fully passive data collection protocol also generalizes to data collected in functional (but non-motoric) tasks where participants must report sensation. Beyond this, I see three possible directions to extend the work: generalizing the findings by extending the analysis to new brain areas (e.g. V1 or PFC, by collaboration), translating the model for ICMS control by developing an inverse model, or grounding the response in sensory experience by using it to decode participant reports. I have left time in my schedule in Q2 2026 to allow these potential extensions.

5 Schedule

I began my doctoral research with a stimulation focus (Aim 3), but put it on hiatus to study how we might aggregate neural datasets (Aim 1). My remaining schedule begins with data collection for NDT-control experiments (Aim 2). As those experiments wind down I expect to be able to focus on writing those results up while running analysis and writeup for Aim 3. Note I expect Aim 3 to produce two publications, one on the artifact removal work and one on the ICMS response taxonomy work.

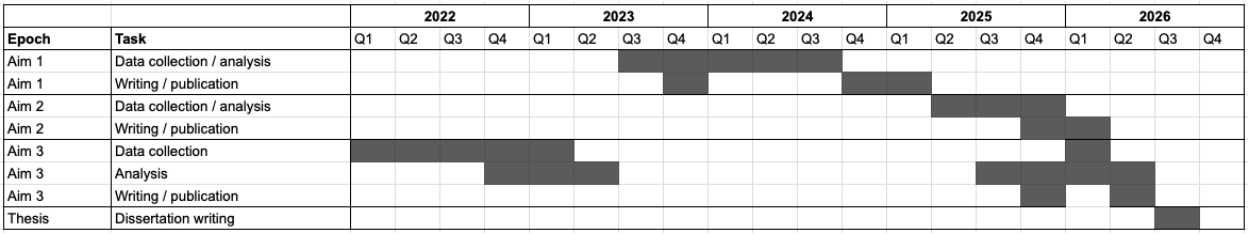

Figure 5.1: Gantt chart overview of project priorities throughout the PhD.

Bibliography

[1] Chethan Pandarinath and Sliman J Bensmaia. "The science and engineering behind sensitized brain-controlled bionic hands". In: *Physiological Reviews* 102.2 (2022), pp. 551–604.

[2] and others. "Quantifying physical degradation alongside recording and stimulation performance of 980 intracortical microelectrodes chronically implanted in three humans for 956–2246 days". In: *medRxiv* (2024).

[3] Benjamin R Cowley et al. "Slow drift of neural activity as a signature of impulsivity in macaque visual and prefrontal cortex". In: *Neuron* 108.3 (2020), pp. 551–567.

[4] Adam L Smoulder et al. "A neural basis of choking under pressure". In: *Neuron* 112.20 (2024), pp. 3424–3433.

[5] Patrick J Marino et al. "A posture subspace in primary motor cortex". In: *bioRxiv* (2024), pp. 2024–08.

[6] Matthew J Mender et al. "The impact of task context on predicting finger movements in a brain-machine interface". In: *eLife* 12 (June 2023). Ed. by J Andrew Pruszynski, Tamar R Makin, and Christian Éthier, e82598. ISSN: 2050-084X. DOI: 10.7554/eLife.82598. URL: https://doi.org/10.7554/eLife.82598.

[7] Xuan Ma, Kevin L Bodkin, and Lee E Miller. "Population activity in motor cortex is influenced by the contexts of the motor behavior". In: *10th International IEEE/EMBS Conference on Neural Engineering*. 2021.

[8] John E Downey et al. "Motor cortical activity changes during neuroprosthetic-controlled object interaction". In: *Scientific reports* 7.1 (2017), p. 16947.

[9] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. "Neuronal population coding of movement direction". In: *Science* 233.4771 (1986), pp. 1416–1419. DOI: 10.1126/science.3749885.

[10] Hanchen Wang et al. "Scientific discovery in the age of artificial intelligence". In: *Nature* 620.7972 (2023), pp. 47–60.

[11] David M. Brandman, Sydney S. Cash, and Leigh R. Hochberg. "Review: Human Intracortical Recording and Neural Decoding for Brain-Computer Interfaces". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2017). Epub ahead of print March 2, 2017, pp. 1687–1696. DOI: 10.1109/TNSRE.2017.2677443.

[12] Jennifer L Collinger et al. "High-performance neuroprosthetic control by an individual with tetraplegia". In: *The Lancet* 381.9866 (2013), pp. 557–564.

[13] Francis R Willett et al. “A high-performance speech neuroprosthesis”. In: *Nature* 620.7976 (2023), pp. 1031–1036.

[14] Maitreyee Wairagkar et al. “An instantaneous voice-synthesis neuroprosthesis”. In: *Nature* 644 (2025), pp. 145–152. DOI: 10.1038/s41586-025-09127-3.

[15] Chethan Pandarinath et al. “High performance communication by people with paralysis using an intracortical brain–computer interface”. In: *eLife* 6 (2017), e18554. DOI: 10.7554/eLife.18554.

[16] Sharlene N. Flesher, Jennifer L. Collinger, Stephen T. Foldes, et al. “Intracortical Microstimulation of Human Somatosensory Cortex”. In: *Science Translational Medicine* 8.361 (Oct. 2016), 361ra141. DOI: 10.1126/scitranslmed.aaf8083.

[17] Emily Graczyk et al. “Clinical Applications and Future Translation of Somatosensory Neuroprostheses”. In: *Journal of Neuroscience* 44.40 (2024).

[18] Sharlene N Flesher et al. “A brain-computer interface that evokes tactile sensations improves robotic arm control”. In: *Science* 372.6544 (2021), pp. 831–836.

[19] Michelle Armenta Salas et al. “Proprioceptive and cutaneous sensations in humans elicited by intracortical microstimulation”. In: *eLife* 7 (2018). Published April 10, 2018, e32904. DOI: 10.7554/eLife.32904.

[20] Matthew S Willsey et al. “A real-time, high-performance brain-computer interface for finger decoding and quadcopter control”. In: *bioRxiv* (2024), pp. 2024–02.

[21] Thomas Hosman et al. “Months-long high-performance fixed LSTM decoder for cursor control in human intracortical brain-computer interfaces”. In: *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2023, pp. 1–5.

[22] Chaofei Fan et al. “Plug-and-Play Stability for Intracortical Brain-Computer Interfaces: A One-Year Demonstration of Seamless Brain-to-Text Communication”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=STqaMqhtDi.

[23] Nicholas S Card et al. “Long-term independent use of an intracortical brain-computer interface for speech and cursor control”. In: *bioRxiv* (2025), pp. 2025–06.

[24] Joseph G Makin et al. “Superior arm-movement decoding from cortex with a new, unsupervised-learning algorithm”. In: *Journal of Neural Engineering* 15.2 (Apr. 2018), p. 026010. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/aa9e95. URL: https://iopscience.iop.org/article/10.1088/1741-2552/aa9e95.

[25] Brianna M. Karpowicz, Rafal Jozefowicz, Chethan Pandarinath, et al. “Stabilizing Brain-Computer Interfaces through Alignment of Latent Dynamics”. In: *bioRxiv* (Apr. 2022). DOI: 10.1101/2022.04.06.487388.

[26] Nicholas S Card et al. “An accurate and rapidly calibrating speech neuroprosthesis”. In: *New England Journal of Medicine* 391.7 (2024), pp. 609–618.

[27] David Sussillo et al. “Making brain–machine interfaces robust to future neural variability”. en. In: 7 (Dec. 2016), p. 13749. ISSN: 2041-1723. DOI: 10.1038/ncomms13749. URL: https://www.nature.com/articles/ncomms13749.

[28] Guy H Wilson et al. “Long-term unsupervised recalibration of cursor BCIs”. In: *bioRxiv* (2023), pp. 2023–02.

[29] Nikhilesh Natraj et al. “Sampling representational plasticity of simple imagined movements across days enables long-term neuroprosthetic control”. In: *Cell* 188.5 (2025), 1208–1225.e32. DOI: 10.1016/j.cell.2025.02.001.

[30] Charles M. Greenspon et al. “Evoking stable and precise tactile sensations via multi-electrode intracortical microstimulation of the somatosensory cortex”. In: *Nature Biomedical Engineering* 9 (2025), pp. 935–951. DOI: 10.1038/s41551-024-01299-z.

[31] Christopher L. Hughes, Sharlene N. Flesher, Jennifer L. Collinger, et al. “Perception of Microstimulation Frequency in Human Somatosensory Cortex”. In: *eLife* 10 (July 2021), e65128. DOI: 10.7554/eLife.65128.

[32] Sliman J Bensmaia, Dustin J Tyler, and Silvestro Micera. “Restoration of sensory information via bionic hands”. In: *Nature Biomedical Engineering* 7.4 (2023), pp. 443–455.

[33] Giacomo Valle et al. “Tactile edges and motion via patterned microstimulation of the human somatosensory cortex”. In: *Science* 387.6731 (2025), pp. 315–322.

[34] Taylor G Hobbs et al. “Biomimetic stimulation patterns drive natural artificial touch percepts using intracortical microstimulation in humans”. In: *Journal of neural engineering* 22.3 (2025), p. 036014.

[35] Josh Merel et al. “Neuroprosthetic decoder training as imitation learning”. In: *PLoS computational biology* 12.5 (2016), e1004948.

[36] M. S. Willsey et al. “Real-time brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder”. In: *Nature Communications* 13 (2022), p. 6899. DOI: 10.1038/s41467-022-34452-w.

[37] Hisham Temmar et al. “Artificial neural network for brain-machine interface consistently produces more naturalistic finger movements than linear methods”. In: *bioRxiv* (2024), pp. 2024–03.

[38] Chethan Pandarinath et al. “Inferring single-trial neural population dynamics using sequential auto-encoders”. en. In: *Nature Methods* 15.10 (Oct. 2018), pp. 805–815. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0109-9. URL: https://www.nature.com/articles/s41592-018-0109-9.

[39] Joel Ye and Chethan Pandarinath. *Representation learning for neural population activity with Neural Data Transformers*. 2021. arXiv: 2108.01210 [q-bio.NC]. URL: https://doi.org/10.51628/001c.27358.

[40] Richard S. Sutton. "The Bitter Lesson". In: *In* Incomplete Ideas. Essay. 2019.

[41] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356.

[42] John E. Downey et al. "Intracortical recording stability in human brain-computer interface users". eng. In: *Journal of Neural Engineering* 15.4 (Aug. 2018), p. 046016. ISSN: 1741-2552. DOI: 10.1088/1741-2552/aab7a0.

[43] Lahiru N Wimalasena, Lee E Miller, and Chethan Pandarinath. "From unstable input to robust output". In: *Nature Biomedical Engineering* 4.7 (2020), pp. 665–667.

[44] János A Perge et al. "Intra-day signal instabilities affect decoding performance in an intracortical neural interface system". In: *Journal of neural engineering* 10.3 (2013), p. 036004.

[45] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

[46] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[47] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG].

[48] Eva L. Dyer and Blake A. Richards. *Accepting the Bitter Lesson and Embracing the Brain's Complexity*. Opinion piece. The Transmitter: Neuroscience News and Perspectives. Mar. 26, 2025. URL: https://www.thetransmitter.org/neuroai/accepting-the-bitter-lesson-and-embracing-the-brains-complexity/ (visited on 08/10/2025).

[49] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV]. URL: https://arxiv.org/abs/2111.06377.

[50] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[51] Juan A. Gallego et al. "Cortical population activity within a preserved neural manifold underlies multiple motor behaviors". en. In: *Nature Communications* 9.1 (Oct. 2018), p. 4233. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06560-z. URL: https://www.nature.com/articles/s41467-018-06560-z.

[52] Srini Turaga et al. “Inferring neural population dynamics from multiple partial recordings of the same neural circuit”. In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/01386bd6d8e091c2ab4c7c7de644d37b-Paper.pdf.

[53] B Wodlinger et al. “Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations”. In: *Journal of Neural Engineering* 12.1 (Feb. 2015), p. 016011. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2560/12/1/016011. URL: https://iopscience.iop.org/article/10.1088/1741-2560/12/1/016011.

[54] Nishal P Shah et al. “Pseudo-linear summation explains neural geometry of multi-finger movements in human premotor cortex”. In: *Nature Communications* 16.1 (2025), p. 5008.

[55] Peter J. Kyberd et al. “Case studies to demonstrate the range of applications of the Southampton Hand Assessment Procedure”. In: *British Journal of Occupational Therapy* 72.5 (May 2009), pp. 212–218.

[56] Francis R Willett et al. “A comparison of intention estimation methods for decoder calibration in intracortical brain–computer interfaces”. In: *IEEE Transactions on Biomedical Engineering* 65.9 (2017), pp. 2066–2078.

[57] Beata Jarosiewicz et al. “Advantages of closed-loop calibration in intracortical brain–computer interfaces for people with tetraplegia”. In: *Journal of Neural Engineering* 10.4 (2013), p. 046012. DOI: 10.1088/1741-2560/10/4/046012.

[58] Vikash Gilja et al. “A high-performance neural prosthesis enabled by control algorithm design”. In: *Nature Neuroscience* 15.12 (2012), pp. 1752–1757. DOI: 10.1038/nn.3265.

[59] Nicholas A Sachs et al. “Brain-state classification and a dual-state decoder dramatically improve the control of cursor movement through a brain-machine interface”. In: *Journal of neural engineering* 13.1 (2015), p. 016009.

[60] Darrel R Deo et al. “Brain control of bimanual movement enabled by recurrent neural networks”. In: *Scientific Reports* 14.1 (2024), p. 1598.

[61] Joseph T Costello et al. “Balancing Memorization and Generalization in RNNs for High Performance Brain-Machine Interfaces”. In: *bioRxiv* (2023), pp. 2023–05.

[62] James R. Eles, Takuya D. Y. Kozai, Antonio L. Vazquez, et al. “The Temporal Pattern of Intracortical Microstimulation Pulses Elicits Distinct Temporal and Spatial Recruitment of Cortical Neuropil and Neurons”. In: *Journal of Neural Engineering* 18.1 (Feb. 2021), p. 015001. DOI: 10.1088/1741-2552/abc29c.

[63] Karthik Kumaravelu et al. “Stoney vs. Histed: Quantifying the spatial effects of intracortical microstimulation”. In: *Brain Stimulation* 15.1 (Jan. 2022), pp. 141–151. ISSN: 1935-861X. DOI: 10.1016/j.brs.2021.11.015.

[64] Karthik Kumaravelu and Warren M. Grill. “Neural mechanisms of the temporal response of cortical neurons to intracortical microstimulation”. In: *Brain Stimulation* 17.2 (Mar. 2024), pp. 365–381. DOI: 10.1016/j.brs.2024.03.012.

[65] Daniel J. O’Shea and Krishna V. Shenoy. “ERAASR: An Algorithm for Removing Electrical Stimulation Artifacts from Multielectrode Array Recordings”. In: *Journal of Neural Engineering* 15.2 (Apr. 2018), p. 026020. DOI: 10.1088/1741-2552/aaa365.

[66] D. Young et al. “Signal processing methods for reducing artifacts in microelectrode brain recordings caused by functional electrical stimulation”. In: *Journal of Neural Engineering* 15.2 (2018), p. 026014. DOI: 10.1088/1741-2552/aa9ee8.

[67] Jérôme Lecoq et al. “Removing independent noise in systems neuroscience data using DeepInterpolation”. In: *Nature Methods* 18.11 (2021), pp. 1401–1408. DOI: 10.1038/s41592-021-01285-2.

[68] Tie Liu et al. “Astronomical image denoising by self-supervised deep learning and restoration processes”. In: *Nature Astronomy* 9 (2025), pp. 608–615. DOI: 10.1038/s41550-025-02484-z.

[69] Amir Zamir et al. *Taskonomy: Disentangling Task Transfer Learning*. 2018. arXiv: 1804.08328 [cs.CV].