A GENERALIST INTRACORTICAL MOTOR DECODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Mapping the relationship between neural activity and motor behavior is a central aim of sensorimotor neuroscience and neurotechnology. Most progress to this end has relied on restricting complexity: studying specific simple behaviors, in limited subjects, with interpretable computational models. However, current trends in deep learning suggest that modeling a breadth of neural and behavioral data all at once is not only possible, but that such a model would also benefit downstream analysis of related data. We accordingly developed Neural Data Transformer 3 (NDT3) as a foundation model for motor decoding of neural data from intracortical microelectrodes. We pretrained NDT3 with 2000 hours of neural population spiking activity paired with diverse motor covariates from over 30 monkeys and humans from 10 labs. Pretrained NDT3 is broadly useful, benefiting decoding on 8 downstream decoding tasks and generalizing to a variety of neural distribution shifts. However, we find signs that scaling over diverse neural datasets may be challenging, as scaling from 200 to 2000 hours already requires increasing model size to 350M parameters to avoid model saturation, and several downstream datasets scarcely benefit from scale. We provide two demonstrations that this scaling is at least partially limited by variability in input and output spaces across neural datasets, which pretraining alone may not resolve.

1 INTRODUCTION

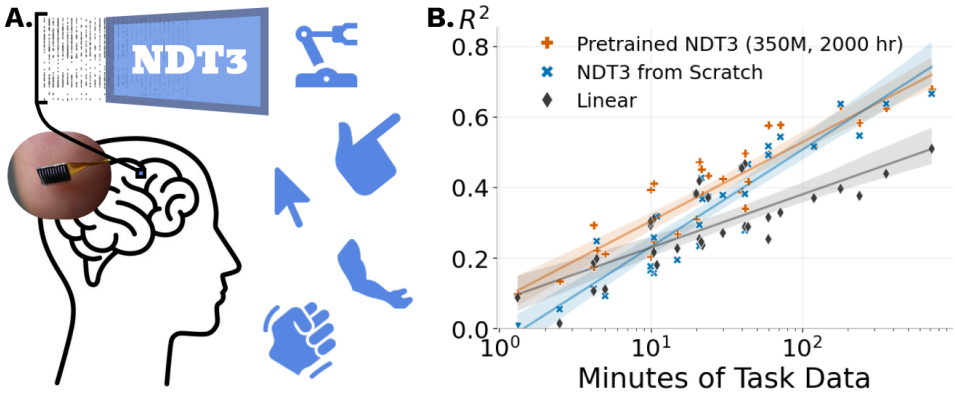

Figure 1. A. NDT3 is a deep network for decoding intracortical spiking activity into low-dimensional time series for various motor effectors[1]. **B.** We aggregate decoding performance on downstream tasks with variable amounts of data (from Fig. 11). While from-scratch models only reliably outperforms a linear baseline after 15 minutes of data, tuning a 350M param. NDT3 pretrained with 2000 hours of data is consistently better at all scales.

Intracortical neural data collection is growing rapidly. This growth comprises not only larger individual datasets with more neurons and higher behavioral complexity (Urai et al., 2022; Stevenson, 2023), but also an increase in the collective number of datasets. This wealth of data presents an opportunity to develop insights and applications that span multiple datasets at once, provided we can reconcile their inherent diversity. Large deep networks appear very suitable for this task, so much

[1]Photo courtesy of REDACT and The Chicago Tribune.

so that the creation of deep networks operating on broad domain data has been termed foundation modeling (Bommasani et al., 2022). Efforts to create foundation models are now proliferating beyond their origins in natural language processing (NLP) and computer vision (CV) into many domains of engineering and science (Wang et al., 2023c). Here, we propose a foundation model for motor decoding from intracortical spiking activity, which we call Neural Data Transformer 3 (NDT3).

Motor decoding is a valuable initial domain for characterizing neural data foundation models. Academic, clinical, and industrial efforts to create iBCIs for neuroprosthetics provide a path for scaling data collection from hundreds to thousands of subject-hours, and also fuel a need for pretrained models that generalize quickly and perform robustly for new users and settings. Behavior prediction metrics for BCI performance are also more intuitive for benchmarking progress than neural data prediction or the abstract goal of providing scientific insight (e.g. with latent variable models or in silico models) (Pei et al., 2021; Wang et al., 2023b). Finally, recent work has shown that deep networks are able to transfer learn across motor cortical datasets collected at different timepoints, subjects, or tasks (Azabou et al., 2024; Ye et al., 2023; Schneider et al., 2023). These ingredients provide the motivation and means for scaling neural data modeling.

However, scaling may be constrained by the design and heterogeneity of modern neural datasets. By design, we refer to the fact that many neural datasets restrict behavioral complexity to probe specific hypotheses. These restrictions impoverish not only the behavioral signals but also the observed neural data (Gao and Ganguli, 2015), providing us a narrow window through which to understand the general relationship between neural activity and behavior. Beyond the limitations of individual datasets, each neural dataset inherently contains unique variability distinguishing them from others. This is most salient when comparing across the datasets aggregated in pretraining, where different neurons are recorded in each subject and distinct output dimensions are required for each effector. To illustrate why these factors together challenge scaling, consider a 2-neuron toy setting, where both neurons fire noisily. One neuron fires on leftward motion and the other fires on rightward motion. No amount of scaling on other datasets could reduce the data needed from this setting to determine which each neuron's preferred direction, but neither is the problem trivial due to stochastic firing.

To assess the value of scaled pretraining on heterogeneous spiking activity, we developed Neural Data Transformer 3 (NDT3). NDT3 uses simple tokenization strategies to enable pretraining over diverse datasets and fine-tuning to new tasks without introducing any new parameters (Fig. 1A). We pretrained NDT3 using up to 2000 hours of neural and behavioral data from motor neuroscience experiments with monkeys and clinical iBCI trials with humans. We then evaluated NDT3's decoding performance on eight diverse motor tasks (Section 3.1) and find that tuning NDT3 yields models that either improve or match task-specific models trained from scratch, with prominent gains when task data is under 1.5 hours (Fig. 1B). Further, these gains persist under a number of distribution shifts (Section 3.3). These benefits may enable both more complex experimental design and potentially decrease the burden of decoder training for people using iBCIs. However, our results also suggest that neural data heterogeneity may be limiting scaling. Scaling pretraining data to 2K hours required raising model capacity to 350M parameters to mitigate performance saturation, and some tasks accrue no benefits from scale at all. We identify NDT3's sensitivity to the specific inputs and outputs seen during fine-tuning as two limits to be overcome for more productive neural data foundation models.

2 Approach

2.1 Data

NDT3 models datasets of paired neural spiking activity and behavior (Fig. 2). Given our focus on motor decoding, most of the data comes from devices implanted in motor cortex of various monkeys and humans (Fig. 2A). These devices are intracortical multielectrode arrays or probes that record 30 kHz extracellular potentials. Spikes are extracted from these potentials, typically by bandpass-filtering the data between 300 and 3000 Hz, and marking a spike when the voltage signal crosses a preset threshold value. The neural data in our pretraining are diverse (Fig. 2B top). Data can have markedly different profiles across electrodes due to being from different electrode arrays in the same subject (left), have many silent channels (middle), or be densely active due to noise (right).

The typical behaviors in the pretraining data are different types of upper-limb reaching and grasping, nearly all from experimental paradigms that consist of short, repeated trials. While neural data were

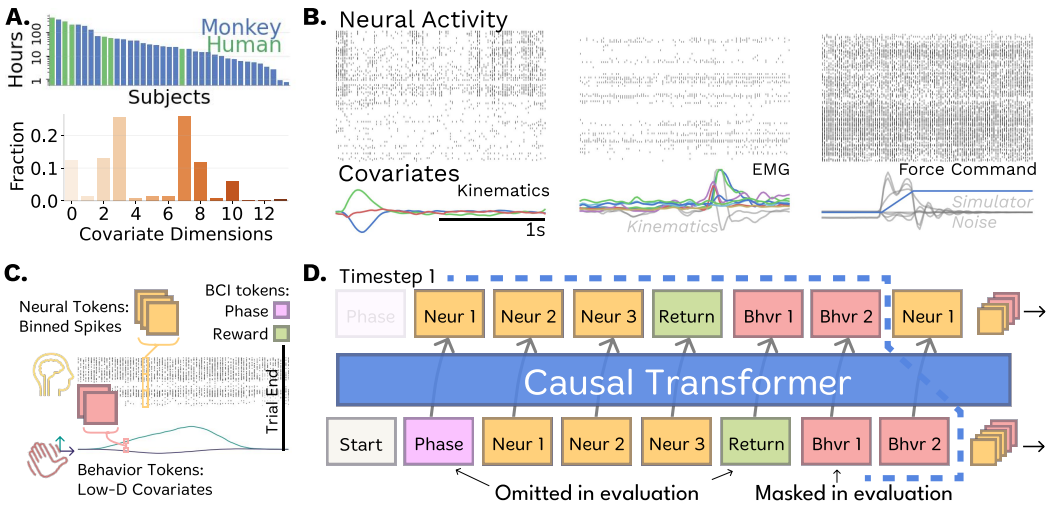

Figure 2. NDT3 Data and Model Design: A. NDT3 models paired neural spiking activity and behavioral covariate timeseries. We plot the distribution of 2000 hours of pretraining data volume by subjects (top) and covariate dimensionality (bottom). **B.** Examples of the neural and behavioral data for each of the three types of behavioral covariates in pretraining: kinematics, EMG (electromyography), or forces. Not all modeled dimensions in data are meaningfully task-related (right, grey behavior). **C.** Neural spiking activity is tokenized in time by binning the number of spikes every 20 ms, and in "space" using patches of channels (usually 32), as in NDT2 (Ye et al., 2023). Behavior is low-dimensional in our data, so we use 1 token per behavior dimension, also per 20 ms timestep. NDT3 also pretrains on data from BCI control, which we annotate with two additional tokens. The phase token indicates whether the user is controlling or observing the behavior and the reward token indicates if the BCI task was completed. **D.** NDT3 models tokens in a single flat stream with linear readins and readouts. Every real-world timestep (shown by the blue cutout) yields several tokens, which we order to allow causal decoding in evaluation. In evaluation, we omit return and phase tokens and zero-mask behavior tokens.

always recorded from microelectrodes, motor covariate signals came from various sensors. In monkey datasets, these sensors measure actual limb activity (e.g., Fig. 2B, left: limb kinematics from optical tracking, middle: electromyography (EMG)). In human datasets, physical movements are typically not possible, so the data's behavior signals are programmatically generated. These signals are "paired" with the neural data in that they are cued or otherwise instructed to the person, who will attempt or imagine the corresponding behavior, such as grasping at a specified force level (Fig. 2B right). This force panel also shows that in pretraining, we cannot always automatically discern the primary task covariates (e.g., blue line, force, in the panel) from other recorded behavioral variables (grey). Thus, some behavior variables may unpredictable. Finally, we include closed loop iBCI data, where some behavior is generated by an iBCI decoder (not NDT3, see modeling strategy in Section 2.2).

The pretraining datasets are composed of archives from several experimental labs and some public datasets, and contain data from non-human primate neuroscience experiments and human clinical trials for neuroprosthetics. The grassroots nature of this aggregate dataset presents a heterogeneity in neural data processing, motor effectors, and experimental setup, most comparable to aggregate robotics datasets (OpenX et al., 2024). We detail the pretraining data composition in Section C.4.

We minimize preprocessing of these data to maximize the applicability of our generalist model. Kinematic signals are typically converted to velocities, and all behavior (kinematics, EMG, force) is normalized per dataset such that the maximum absolute value of each variable is one. Data are cut or concatenated into fixed length sequences, without additional annotation of data discontinuity. This strategy, common in language modeling (Geiping and Goldstein, 2022), homogenizes the data for improved GPU utilization while maintaining throughput of real data. We used a length of two seconds as it is roughly the timescale of the behavior in our data (Fig. 2A). Sequences with no spikes or covariate variability are discarded. In total, this yielded about 3 million sequences, 1 billion neural tokens, or 1750 hours, which we sample uniformly for pretraining. We round this to 2 khrs in subsequent text for simplicity.

2.2 MODEL

NDT3 is a causal Transformer with linear readin and readout layers for its various modalities, similar to GATO or TDMs (Reed et al., 2022; Schubert et al., 2023; Chameleon, 2024). For use with a Transformer, the data must be tokenized (Fig. 2C). We tokenize neural data by patching spike counts (Ye et al., 2023); each token is a flattened vector of the binned spikes in a chosen temporal resolution (20 ms) and spatial dimension (32 channels). For example, neural activity sampled from an electrode array with 100 channels would patch into $4 = \lceil 100/32 \rceil$ 32D neural tokens per 20 ms timestep. As the behavioral variables are already low-dimensional, we simply assign 1 token per dimension at the same temporal resolution as neural data. Finally, we add tokens marking whether the behavior are generated by a BCI system or by physical limb movement. While measured kinematics, EMG, or force will reflect a natural relationship with neural activity, behavioral data from BCI tasks are controlled by a program or learned decoder. We frame BCI-driven behavior as a suboptimal demonstration (Merel et al., 2016), and adopt a scheme inspired by Decision Transformers (Chen et al., 2021; Lee et al., 2022). In this scheme, we use a Phase token to track the timesteps where behavior is at least driven by neural activity and under decoder control, or only under programmatic, open loop control. We also use a Return token reflecting controller quality based on task completion. Note that these signals are only considered for pretraining, and are ablated entirely from the model at evaluation. Similarly, input behavior tokens are masked out in inference, so that the model input only indicates how many tokens must be predicted. NDT3 is trained with mean-squared error for prediction of behavioral variables, and categorical cross-entropy losses for prediction of neural spike count and reward.

All modalities are flattened into a single token stream, with the order of tokens in each real-world timestep respecting a canonical order required for control (Fig. 2D). As in GATO, individual tokens are annotated with learned position embeddings identifying token modality and sub-modality "position." We additionally use rotary embeddings (Su et al., 2023) to track real-world timesteps.

Pretraining and Fine-Tuning We pretrain NDT3 models over variable pretraining data and in sizes of 45M and 350M parameters to assess the impact of data and model scaling. Pretraining is early stopped according to validation loss or terminated at a maximum of 400 epochs. The 200 hour, 45M model trains for 480 A100-hours while the 2000 hour (2kh) 350M model takes 20K A100-hours. Fine-tuning maintains the pretraining objectives and updates all parameters.

2.3 EVALUATION STRATEGY

Evaluation datasets and tuning Our main evaluation (Section 3.1) uses four human and four monkey datasets sampling varied upper limb movements, which we detail in Section C.4. Each dataset contains multiple sessions of data, typically from a single monkey or human. We will refer to each such setting as a "task," distinguished from the behavioral procedure performed in each dataset. Each session has unique variability, so fine-tuning procedure may greatly impact decoding results. Prior work (Azabou et al., 2024; Ye et al., 2023; Zhang et al., 2024) ran focused evaluations by tuning and evaluating separately for each evaluation session. To manage compute and storage demands and to reflect that real world datasets are rarely collected or analyzed in isolation, we fine-tune NDT3 jointly over data combined from multiple evaluation sessions. Fig. 12 shows this joint tuning outperforms focused tuning for multi-session data.

Downstream Hyperparameters All tuned and from-scratch NDT3 models use a search over 3 learning rates. This sweep is limited to make computational demands tractable, but also demonstrates the versatility of the base model. Importantly, the same search space is used for all tasks; we list the space and show its sufficiency relative to wider sweeps in Section C.2. The best learning rate is chosen based on average validation performance across three random seeds, and we report their mean on the evaluation data.

Baselines We compare against Wiener filters (WF) and multi-session NDT2. WFs are a conventional linear method for both motor decoding and control in iBCI devices (Pandarinath and Bensmaia, 2022), and we implement them as ridge regression with multi-timestep history. NDT2 is a Transformer that uses MAE-style (He et al., 2021) self-supervision and has been demonstrated to improve with heterogeneous neural datasets, including the multi-session data we use it for here. Other Transformer variants have been proposed for pretraining on spiking data (Azabou et al., 2024; Zhang et al., 2024), but the field yet lacks consensus benchmarks to distinguish the most promising proposal to scale.

3 RESULTS

NDT3's pretraining effort advances prior intracortical models an order of magnitude in data and model scale, from 200 to 2000 hours and 10M+ to 100M+ parameters. In Section 3.1, we show the increased data scale saturates aggregate downstream performance unless simultaneously increasing model scale. We propose that the performance drop from scaling data alone is due to high variability across intracortical motor decoding datasets. In Section 3.2, we show how this variability reflects in NDT3's sensitivity to shifts in data input or output. Section 3.3 concludes by showing that despite this challenge for further data scaling, NDT3's pretraining already provides gains that generalize to various novel settings, establishing NDT3 as a useful foundation for motor decoding.

3.1 MULTI-SCALE EVALUATION ACROSS MOTOR DECODING TASKS

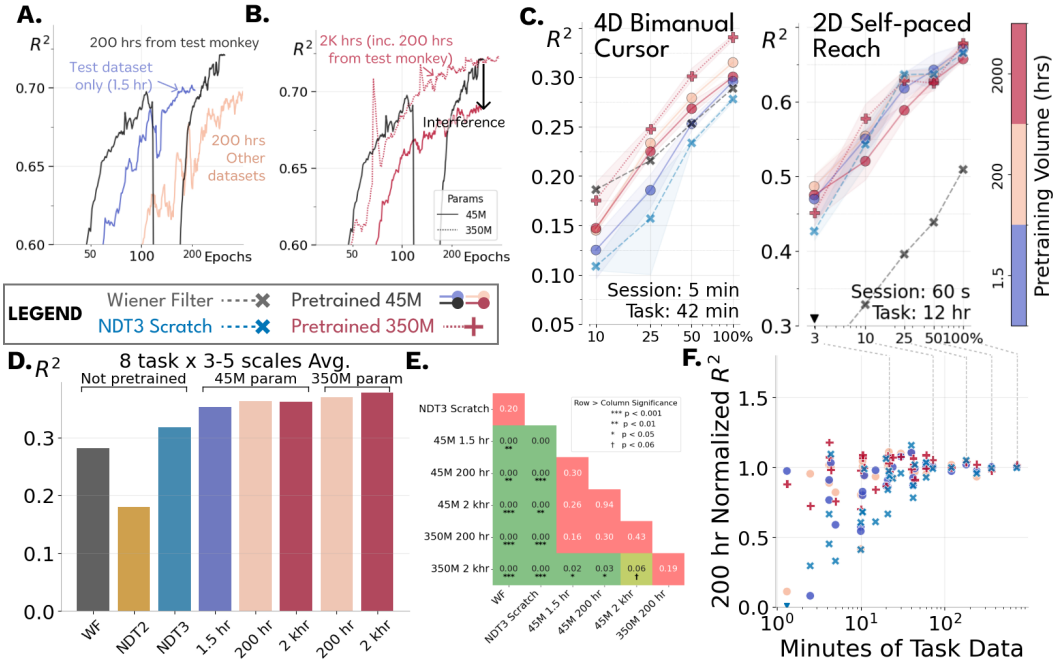

Figure 3. Evaluation on diverse motor tasks: A single legend and color scheme is used throughout. **A.** Test-split pretraining R^2 compared for 3 models. All model pretraining data includes 1.5 hours of calibration data for the test dataset. We compare a model with just this data (Test dataset only) vs using 200 hours of additional data either from the test monkey or from over 10 other monkeys. Only the additional test monkey data improves over the calibration model. Models terminate at different points due to early stopping. **B.** Pretraining R^2 for models with up to 2000 hours (2 khrs) of pretraining data. The 2 khr model degrades in performance vs the 200 hr model at 45M parameters and merely maintains performance at 350M parameters. **C.** Examples of good and bad data-scaling in downstream multiscale evaluation on two datasets. The bottom right text shows time in each evaluation session and total time in each dataset. The x-axis scales this full dataset down by random subsampling. Shading shows standard deviation on 3 tuning seeds. Increasing pretraining data yields performance gains at all downstream scales in the 4D task, but effects are unclear in the self-paced reach task. ▼ indicate outliers clipped for clarity. **D.** Downstream performance averaged for 31 settings comprised of different downstream datasets and scales, for different NDT3s and baselines. 45M NDT3s improve with data from 1.5 hrs to 200 hrs but saturate at 2 khrs. Increasing model size to 350M parameters enables further gain. **E.** p-values computed from FDR-corrected pairwise t-tests for each pair of models. The 350M 2 khr NDT3 significantly outperforms other pretrained NDT3s, except the 350M 200 hr NDT3, and is the only model to do so. NDT2 is omitted for brevity, full figure in Fig. 13. **F.** Per-task performance, normalized by the 350M 200 hr NDT3 performance, is shown against task time for different NDT3 models. Each vertical band shows models trained on the same evaluation setting, e.g. dashed lines show the evaluations from the self-paced reaching dataset. Model variability vanishes by 1.5 hours.

To set expectations for how data scale and model size will impact model performance, we first examine pretraining curves computed on a test split. This test split contains multiple sessions of 2D reaching behavior mainly from one monkey. From the dataset this test split was drawn from, we sampled 1.5 hours of data and included it in the pretraining of all models to allow learning this specific

test task. Fig. 3A shows the test performance of a model using just this 1.5 hours of "calibration" data, and two models using 200 hours of data. One of these 200 hour models used data from over 10 other monkeys performing a variety of reaching tasks, but did not improve over the minimal 1.5 hour model on the test data. In contrast, using 200 hours more from the test monkey (from a separate set of experiments with similar behavior) achieved a small improvement in performance. Thus, only closely related data appears to benefit a model that already uses sufficient task-specific data, in this case 1.5 hours. To further emphasize dataset dissimilarity, Fig. 3B shows that a 2 khr model that sees the same extra test data degrades in performance, indicating additional data can interfere in the learning of the test task. This interference can be mitigated by increasing model size to 350M parameters, which matches prior work showing the importance of increasing model size and dataset size in tandem (Dosovitskiy et al., 2021; Kolesnikov et al., 2020; Aghajanyan et al., 2023). However, we still do not see gains on the test task beyond what is achieved by providing 200 hours from the test monkey.

This upstream saturation motivated a downstream evaluation conducted at multiple data scales. We illustrate this evaluation for two tasks in Fig. 3C. These two datasets are from a human performing open loop iBCI calibration for bimanual cursor use (Deo et al., 2024), and a monkey performing self-paced reach to random targets (O'Doherty et al., 2017). In both cases, the individuals are held-out from pretraining entirely, so the task-specific data is only seen in tuning. In the bimanual task, NDT3 performance improves with increased pretraining data at all downstream data scales. The saturation in the 2000 hour model and subsequent rescue by increased model size is also replicated. The self-paced reach shows a much less clear result. For example, the from-scratch NDT3 achieves nearly the best performance at most data scales.

These two tasks reflect just two of several different downstream trends on the eight different evaluation datasets we study. We defer discussion of individual tasks and their variability to Section B.5, and next consider summary performance in Fig. 3D. This summary, produced by tuning over 2000 models in 31 evaluation settings, identifies an overall benefit of pretraining scale. Consistent with findings in computer vision (Entezari et al., 2023), we find that minimal pretraining (1.5 hrs) already achieves a large performance gain, but scaling data to 200 hours yields further improvements. Further scaling to 2000 hours is also helpful, but only when paired with increased model size to 350M parameters, as in upstream evaluation. The gain of the 350M 2 khr model over other models is statistically significant for all except the 350M 200 hr model (Fig. 3E). Other pairs of pretrained NDT3s are not significantly different, though we suspect this could be achieved with even more evaluations.

Scaling's positive overall trend is not supported by uniform improvements in each downstream setting (Fig. 11). While we have already seen that comparing benefits by experimental task is challenging, it is known that pretraining is particularly beneficial at low downstream data scale. We reproduce this finding in our setting by plotting task performance against downstream scale in Fig. 3F. The distinction between pretrained models and from-scratch models vanishes by 1.5 hours. This result is also supported in the raw, un-normalized performance (Fig. 1B) and consistent with upstream results (Fig. 3B). Estimating that a behavior in an experiment ranges from 1 to 10 seconds, 1.5 hours would comprise about 0.5-5K data points. Pretraining only being productive up to 5K data points is relatively impotent by CV or NLP standards (Kolesnikov et al., 2020; Wang et al., 2019), and in practical terms can be exceeded after a few sessions of data collection. Importantly, we cannot attribute this saturation to limits of the neural data signal itself (Abnar et al., 2021), as scaling downstream data productively increases task performance. This leaves uncertainty on whether NDT3's pretraining is inherently limited.

3.2 Pretraining is limited by input and output variability in tuning

In Section 3.1 we observe that the benefit of scaling pretraining appears to saturate at relatively low downstream data scales. It is possible that this limit reflects intrinsic variability across neural datasets, with a long tail of specialized features that are needed to the best performance in each dataset. Alternatively, modeling decisions around architecture, hyperparameters, and post-training, may all significantly influence scaling. As a first step towards understanding whether the scaling we observe is limited by data or methodology, we next analyze NDT3's sensitivity to the specific neural inputs and covariate outputs seen in tuning.

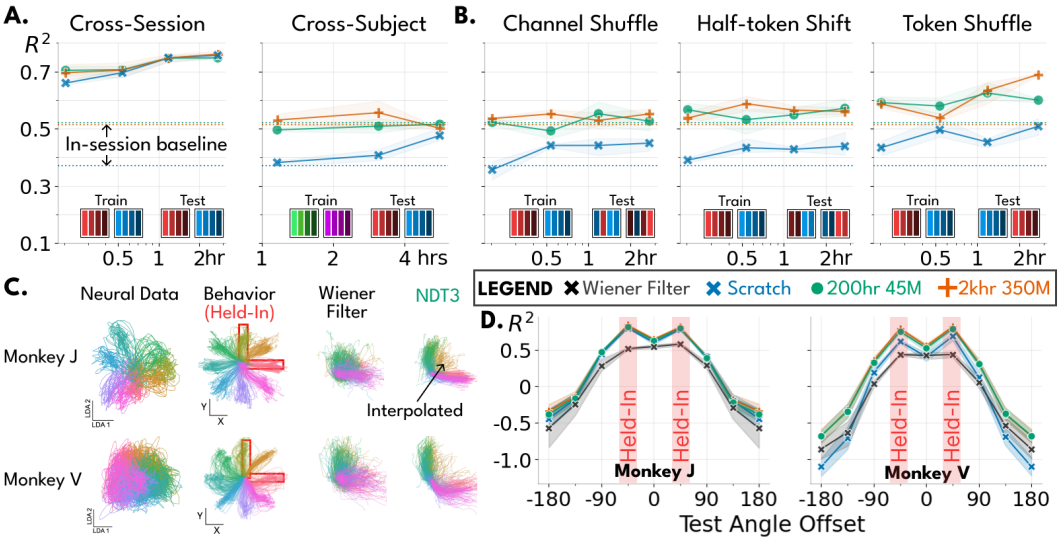

Figure 4. NDT3 fails in certain novel input and output configurations. A: Cross-session transfer persists after pretraining, but cross-subject does not. We test NDT3 in a downstream task with one evaluation session from a monkey self-paced reaching dataset. Training uses 1 minute from the evaluation session and additional data from other sessions (Cross-Session) with the same monkey or from sessions from a different monkey for the same behavior (Cross-Subject). **B: Shuffling inputs ablates cross-session data to resemble cross-subject transfer.** uses the same cross-session neural data but permutes input dimensions (recording channels). Shuffle channel randomly permutes inputs, half-token shift rolls channels so that each channel i uses data from $i + 16$, and shuffle token permutes data patchwise, keeping channels from the same patch together. Channel shuffling and half-token shifts both are sufficient to reduce cross-session transfer to the same level as cross-subject transfer. All panels show the baseline performance achieved by the model with just 1 minute of test-session data, x-axis shows additional cross-context data provided. **C. Pretraining does not improve angular extrapolation.** Two different monkey 2D center-out datasets with 8 angular conditions. LDA on all neural data show monkey J's data is distinctly more separable than monkey V's. We then test generalization of behavioral decoding to held-out angles after training on 2 of 8 angles (boxed in red). Both WF and NDT3 predictions are constrained between the held-out angles. NDT3's predictions are visible for the interpolated angle only in monkey J. **D.** We quantify performance for decoding on each angle with respect to distance from the angle between the held-in angles. We average performance on all 8 central angles.

Input order sensitivity may limit cross-subject transfer. In neural data, the effectiveness of transfer learning is greatly reduced when using cross-subject data compared to cross-session data (Ye et al., 2023). This suggests limited scaling may be caused in part by the vanishing utility of cross-subject data, even when the data is collected in identical experimental setups and thus controlled for other variables. We can illustrate this by comparing cross-session and cross-subject transfer after large-scale pretraining. On a monkey 2D reaching dataset (O'Doherty et al., 2017) in Fig. 4A, we tune NDT3 with calibration data from one test session and additional cross-session or cross-subject data. As in Section 3.1, cross-session data is still highly beneficial even after pretraining, but cross-subject data is only helpful for from-scratch models. Pretrained models in the cross-subject setting instead begin and plateau at a performance that is just slightly better than the best cross-subject performance in from-scratch models. These results suggest that NDT3's pretraining has already learned the features that cross-subject transfer provides in this task, supporting the idea that scaling is limited by a low (task-dependent) ceiling on cross-subject transfer.

Despite this ceiling, NDT3 achieves moderate cross-subject transfer, suggesting that the limitations may stem from data variability rather than methodology. On the other hand, transfer may also be inhibited by NDT3's design and inductive biases. To probe this distinction, we observe that since cross-subject data contain different neurons, cross-subject transfer must at least accommodate changes in the specific semantics of each data dimension, a problem which we term "sensor variability." We can isolate how effectively NDT3 resolves sensor variability by transferring with cross-session data while permuting the test-session's neural dimensions. Fig. 4B Channel shuffle shows that input permutation cripples the ability of NDT3, whether pretrained or not, to learn from cross-session data.

To sharpen what distinguishes cross-session from cross-subject transfer, we apply structured ablations of input order, as in Neyshabur et al. (2020). We find that even the small alteration of a half-token

shift in channel order is sufficient to reproduce the effect of full shuffling (Fig. 4B center). This shows NDT3's cross-session transfer depends greatly on the specific token dimension semantics, and may drive the observed limits in scaling. Finally, we consider a shuffle that only alters the test session's neural token order with respect to cross-session data, hypothesizing Transformers can more easily correct token ordering alterations. Indeed, pretrained and from-scratch models recover some gains when cross-session data is only token-shuffled (Fig. 4B right). This manipulation shows the influence of model design on transfer. For reference, POYO (Azabou et al., 2024) adopts a graph-based Perceiver design motivated by this sensor variability challenge, making it an appealing candidate for scaling up. We note that to combat potential confounds from jointly tuning on heterogeneous data, we evaluated both sequential and joint tuning strategies in these experiments and reported the better approach for each panel. A full comparison is given in Section B.6.

Outputs seen in tuning restrict NDT3 predictions. The increased sample efficiency of pretrained models suggests that NDT3 could map a new individual's neural activity to behavior without sampling the full range of neural-behavior data. For example, having seen many instances of radial reaches, NDT3 may generalize to unseen reach directions in a new subject better than non-pretrained decoders, which fail completely (Rizzoglio et al., 2022). Structured radial reaches are a particularly simple litmus test for such generalization as their underlying neural activity is easily visualized in terms of a planar subspace (Churchland et al., 2012). We assess angular generalization in Fig. 4C, by evaluating reach decoding in an isometric, force-based (Monkey J) setting and a manipulandum-based (Monkey J) setting. Beyond the change in effectors, the neural data in these datasets also vary in their separability as visualized by linear discriminant analysis (LDA). We next train decoders on every pair of angles separated by 90 degrees (one shown) and plot predictions on held-out trials from all angles. The specific failure of both WFs and NDT3 is that their predictions do not extrapolate to held-out angles, consistent with Rizzoglio et al. (2022). While NDT3 produces more organized trajectories than the WF, held-out angles are constrained to the held-in conditions. Intriguingly, even the intermediate, interpolated angle appears to disappear for the less separable monkey V dataset. We quantify prediction performance in Fig. 4D. Notably, the two datasets achieve similar R^2 patterns despite significantly different LDA projections, with the only subtle difference being that from-scratch models are slightly worse for monkey V.

These results show that NDT3 fails to generalize to held-out directions, but also hint at the reason why. One appealing interpretation is that NDT3 is incentivized to sacrifice any angular generalization to maximize held-in performance, consistent with reports that RNNs exploit attractor structures to improve performance on noisy neural data (Costello et al., 2023). The fact that in the more challenging monkey V dataset, the interpolated angle between held-in angles appears degraded, and pretrained models slightly outperform from-scratch models, suggest that pretraining may even incentivize this tradeoff. Importantly, we show in Section B.1 that a constrained linear decoder class (more so than WFs) can yield held-out generalization. Thus while NDT3 fails to generalize to held-out directions, it remains unclear whether this implies that pretraining failed to learn a viable prior for radial reach (such as a low-dimensional linear constraint), or whether we have merely failed to elicit this structure (i.e. through post-training). Even in the easiest setting of planar reaches, it can be challenging to directly probe whether the model has learned what we expect it to.

Our manipulations on model input and output highlight the difficulty of pinpointing whether scaling is limited by data or methodology. However, they do provide basic tests of model capability, namely robustness to channel shifts and generalization to unseen behaviors, that we expect future approaches will need to overcome to achieve better neural data foundation models.

3.3 Where does pretraining help?

Despite challenges for scaling, NDT3's pretraining learns a prior from hundreds of hours of neural data. We conclude by showing how this prior generalizes beyond the main evaluation in Section 3.1.

Neural distribution shifts Neural data is nonstationary, with shifts rising from a mix of controlled experimental variables to more speculative factors. For example, the firing rate of different channels will evolve over the course of an hour, implying a distribution shift associated with change in time (Fig. 5A top left). Shifts also occur between activity in two arm postures or whether finger motion occurs under spring load or not (Fig. 5A top middle and right). Since these shifts are common in neural data, pretraining gains should ideally be robust to their effect. We thus tune models on data

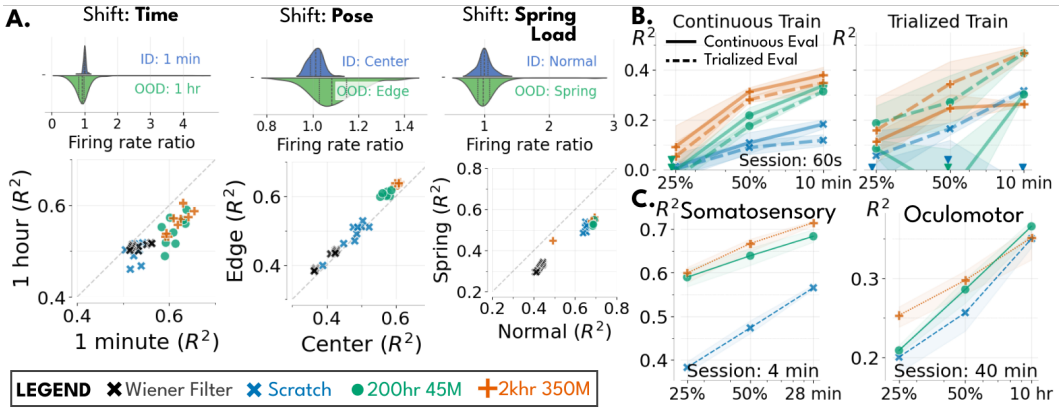

Figure 5. Generalizability of pretraining gains. A. Models fine-tuned in one distribution of data are evaluated in-distribution (ID) and out-of-distribution (OOD). Top plots show the distribution across channels of neural firing rates from OOD and ID trials, normalized by average ID firing rates. Lower plots scatter OOD vs ID performance, with each point being a single model with different hyperparameters. . The **time** shift uses two human cursor datasets collected one hour apart. Models were tuned in each block and were evaluated in the second block. **Pose** shift uses a monkey center-out reach task which was performed with the hand starting in different locations in the workspace. **Spring Load** uses a dataset of monkey 1D finger motion with or without spring force feedback. **B.** 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. **C.** Multiscale fine-tuning performance of NDT3 on datasets recorded outside motor cortex, namely S1 (Somatosensory) and FEF/MT (Oculomotor).

from one setting (in-distribution, ID) measure the performance of models in that same setting and the shifted setting (out-of-distribution, OOD). Positive correlation of performance in all cases imply the ID gains conferred by pretraining persist OOD. This ID-OOD correlation is consistent with Miller et al. (2021), implying a potentially fruitful relationship between the distribution shifts characterized in neural data and those studied in computer vision. More practically, these examples suggest that pretraining benefits are not dependent on narrow features specific to the choice of tuning dataset.

Trial structure DNNs have been observed to overfit the temporal structure of experimental trials in datasets, which poses a serious challenge for applying DNNs to continuous control in iBCI use (Deo et al., 2024; Costello et al., 2023). For example, a deep network might learn there is always no motion before the start of a trial, independent of the neural activity. However, these claims have been studied exclusively on un-pretrained deep networks. In Fig. 5B, we assess how pretraining impacts such overfit to trial structure in open loop human cursor control data. We do this by comparing a continuous and trialized setting. Continuous training and evaluation follows our standard data preparation, where we cut random one second intervals of data for training and continuously stream up to one second of data in evaluation. Trialized training and evaluation provides data formatted to respect trial boundaries, such that the data always begins with the initiation of behavior. The superior performance in all models when combining trialized training and evaluation over continuous training and evaluation shows that all models can learn to exploit trial structure. However, while trialized from-scratch models become subtrivial under continuous evaluation (solid blue line is off-panel), pretrained models degrade more gracefully. For example, the 350M 2 khr model evaluated continuously only performs slightly worse with trialized tuning than with continuous tuning. Thus, we conclude there is a reduced dependence on trial structure in pretrained NDT3s, which should benefit both data analysis and iBCI control. Note however the contrast in these results with Fig. 4C, which show that DNNs clearly do overfit to tuning data in some cases. These overfitting subtleties imply that future work needs to take particular care in evaluating model generalization.

New brain areas In Fig. 5C, we return to multiscale fine-tuning to test how NDT3, pretrained on motor cortex, performs in somatosensory cortex (S1) and oculomotor areas (FEF and MT). Pretraining provides a large boost over from-scratch models is high in S1, but also nontrivial in the Oculomotor dataset. While the former can be attributed to the close interaction of sensorimotor areas, the latter

implies NDT3 has learned a broader prior. While encouraging, our results thus far suggest this prior could reflect neurophysiology (e.g. declining subject focus over time (Steinmetz et al., 2019)), but might also reflect common experimental artifacts like trial structure. For example, this Oculomotor dataset contains 4 behavioral conditions, which may benefit from the tendency to learn classifiers shown in Fig. 4C rather than a prior on neural dynamics.

4 Discussion

Many fields are now pursuing large scale deep learning as "a tide that lifts all boats" (Abnar et al., 2021), with the hope that improvements in effective pretraining will yield field-wide, downstream improvements. Such a unifying abstraction may be timely for neuroscience, given the increasing volume, diversity, and complexity of modern neural data. Joining other pretraining efforts on varied modalities of neural data (Section A), we trained NDT3 on 2000 hours of paired neural population activity from motor cortex and behavior, and then conducted a broad downstream decoding evaluation. Consistent with the broad foundation modeling narrative, we found the best aggregate performance from increasing data scale and model size jointly. However, these benefits from pretraining vary with the downstream dataset, with several datasets having minimal improvements from scale (Section B.5). This result may stem in various ways from our approach, for example in insufficient breadth of hyperparameter sweeps, or too narrow of a focus on decoding metrics. Alternatively, we have highlighted how improving downstream gains may require new architectural innovations robust to input ordering shifts, and possible new training strategies to promote generalization. Overall, NDT3 establishes a strong baseline foundation model for intracortical decoding from spiking activity, but highlights important directions for future scaling.

More broadly, we advocate for further consideration of how neural data can contribute to and gain from the ongoing cross-disciplinary conversation on foundation modeling. For example, our input and output sensitivity analyses were inspired by ML (Neyshabur et al., 2020; Pham et al., 2021) and neuroscientific literature (Gallego et al., 2020; Sadtler et al., 2014), respectively. Challenges to scale in neural data could deeply resemble interference in multimodal models (Aghajanyan et al., 2023; Liu et al., 2024). Inversely, neural distribution shifts have the advantage of being carefully characterized, and so the appearance of correlated ID-OOD performance, as also appears in CV, NLP, and other AI domains (Taori et al., 2020), may refine our understanding of when such correlation will occur, and thus when foundation models will be effective. Our hypothesized challenge of sensor variability should be particularly interesting to compare across the biosignals community, which must overcome analogous variability to achieve our shared goal of achieving user-general models.

4.1 Ethics Statement

The animal datasets used in this work were collected for other studies that were approved by Institutional Animal Care and Use Committees. Human datasets were also collected for other studies, with Institutional Review Board approval and as part of clinical trials conducted under FDA Investigational Device Exemptions. Informed consent was obtained prior to any experimental procedures. We discuss the potential for NDT3 to reduce user burden for iBCI-based neuroprosthetics, though the dissemination of pretrained models on these data raise the risk that the original human data may be recoverable from model weights. Since this seems technically challenging at this point, and since the source data are restricted to binned spiking activity to begin with, we deem the risk low enough to justify the potential scientific benefit of sharing our pretrained models.

4.2 Reproducibility Statement

Advancing neural data foundation modeling will require a flourishing open-source ecosystem, including data, models, and evaluations. While we will release our models and codebase, our work currently has limited reproducibility given our inability to release pretraining data. Similarly, we have tried to use open evaluations where possible, but several evaluation datasets remain private. We expect that field-wide trends toward open data releases, and larger scale academic (Koch et al., 2022) or academic-industrial collaborations, can alleviate this limitation in the near future.

REFERENCES

S. Abnar, M. Dehghani, B. Neyshabur, and H. Sedghi. Exploring the limits of large scale pre-training, 2021. URL https://arxiv.org/abs/2110.02095.

E. H. Adelson, J. R. Bergen, et al. *The plenoptic function and the elements of early vision*, volume 2. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of . . . , 1991.

A. Aghajanyan, L. Yu, A. Conneau, W.-N. Hsu, K. Hambardzumyan, S. Zhang, S. Roller, N. Goyal, O. Levy, and L. Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023.

R. Antonello, A. Vaidya, and A. G. Huth. Scaling laws for language encoding models in fmri, 2024. URL https://arxiv.org/abs/2305.11863.

M. Azabou, V. Arora, V. Ganesh, X. Mao, S. Nachimuthu, M. Mendelson, B. Richards, M. Perich, G. Lajoie, and E. Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36, 2024.

G. Bachmann and V. Nagarajan. The pitfalls of next-token prediction, 2024. URL https://arxiv.org/abs/2403.06963.

T. Benster, G. Wilson, R. Elisha, F. R. Willett, and S. Druckmann. A cross-modal approach to silent speech with llm-enhanced recognition. *arXiv preprint arXiv:2403.05583*, 2024.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

M. P. Branco, L. M. De Boer, N. F. Ramsey, and M. J. Vansteensel. Encoding of kinetic and kinematic movement parameters in the sensorimotor cortex: A brain-computer interface perspective. *European Journal of Neuroscience*, 50(5):2755–2772, Sept. 2019. ISSN 0953-816X, 1460-9568. doi: 10.1111/ejn.14342. URL https://onlinelibrary.wiley.com/doi/10.1111/ejn.14342.

J. O. Caro, A. H. de Oliveira Fonseca, S. A. Rizvi, M. Rosati, C. Averill, J. L. Cross, P. Mittal, E. Zappala, R. M. Dhodapkar, C. Abdallah, and D. van Dijk. BrainLM: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RwI7ZEfR27.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

T. Chameleon. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

G. Chau, C. Wang, S. Talukder, V. Subramaniam, S. Soedarmadji, Y. Yue, B. Katz, and A. Barbu. Population transformer: Learning population-level representations of intracranial activity, 2024. URL https://arxiv.org/abs/2406.03044.

L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL https://arxiv.org/abs/2106.01345.

M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

J. T. Costello, H. Temmar, L. H. Cubillos, M. J. Mender, D. M. Wallace, M. S. Willsey, P. G. Patil, and C. A. Chestek. Balancing memorization and generalization in rnns for high performance brain-machine interfaces. *bioRxiv*, pages 2023–05, 2023.

T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. van Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. P. Collier, A. Gritsenko, V. Birodkar, C. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetić, D. Tran, T. Kipf, M. Lučić, X. Zhai, D. Keysers, J. Harmsen, and N. Houlsby. Scaling vision transformers to 22 billion parameters, 2023. URL https://arxiv.org/abs/2302.05442.

D. R. Deo, F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy. Brain control of bimanual movement enabled by recurrent neural networks. *Scientific Reports*, 14(1):1598, 2024.

A. Doerig, R. Sommers, K. Seeliger, B. Richards, J. Ismael, G. Lindsay, K. Kording, T. Konkle, M. A. J. V. Gerven, N. Kriegeskorte, and T. C. Kietzmann. The neuroconnectionist research programme, 2022. URL https://arxiv.org/abs/2209.03718.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 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.

R. Entezari, M. Wortsman, O. Saukh, M. M. Shariatnia, H. Sedghi, and L. Schmidt. The role of pre-training data in transfer learning. *arXiv preprint arXiv:2302.13602*, 2023.

C. Fan, N. Hahn, F. Kamdar, D. Avansino, G. H. Wilson, L. Hochberg, K. V. Shenoy, J. M. Henderson, and F. R. Willett. 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.

J. Farebrother, J. Orbay, Q. Vuong, A. A. Taïga, Y. Chebotar, T. Xiao, A. Irpan, S. Levine, P. S. Castro, A. Faust, A. Kumar, and R. Agarwal. Stop regressing: Training value functions via classification for scalable deep rl, 2024. URL https://arxiv.org/abs/2403.03950.

R. D. Flint, E. W. Lindberg, L. R. Jordan, L. E. Miller, and M. W. Slutzky. Accurate decoding of reaching movements from field potentials in the absence of spikes. *Journal of neural engineering*, 9(4):046006, 2012.

J. A. Gallego, M. G. Perich, R. H. Chowdhury, S. A. Solla, and L. E. Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270, Feb 2020. ISSN 1546-1726. doi: 10.1038/s41593-019-0555-4. URL https://www.nature.com/articles/s41593-019-0555-4.

P. Gao and S. Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

J. Geiping and T. Goldstein. Cramming: Training a language model on a single gpu in one day, 2022. URL https://arxiv.org/abs/2212.14034.

M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories, 2021. URL https://arxiv.org/abs/2012.14913.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

B. Jarosiewicz, A. A. Sarma, D. Bacher, N. Y. Masse, J. D. Simeral, B. Sorice, E. M. Oakley, C. Blabe, C. Pandarinath, V. Gilja, et al. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science translational medicine*, 7(313):313ra179–313ra179, 2015.

W. Jiang, L. Zhao, and B. liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

B. M. Karpowicz, J. Ye, C. Fan, P. Tostado-Marcos, F. Rizzoglio, C. Washington, T. Scodeler, D. de Lucena, S. R. Nason-Tomaszewski, M. J. Mender, X. Ma, E. M. Arneodo, L. R. Hochberg, C. A. Chestek, J. M. Henderson, T. Q. Gentner, V. Gilja, L. E. Miller, A. G. Rouse, R. A. Gaunt, J. L. Collinger, and C. Pandarinath. Few-shot algorithms for consistent neural decoding (falcon) benchmark. *bioRxiv*, 2024. doi: 10.1101/2024.09.15.613126. URL https://www.biorxiv.org/content/early/2024/09/16/2024.09.15.613126.

C. Koch, K. Svoboda, A. Bernard, M. A. Basso, A. K. Churchland, A. L. Fairhall, P. A. Groblewski, J. A. Lecoq, Z. F. Mainen, M. W. Mathis, et al. Next-generation brain observatories. *Neuron*, 110(22):3661–3666, 2022.

A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning, 2020. URL https://arxiv.org/abs/1912.11370.

D. Kostas, S. Aroca-Ouellette, and F. Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15, 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/articles/10.3389/fnhum.2021.653659.

C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation, 2016. URL https://arxiv.org/abs/1608.08242.

K.-H. Lee, O. Nachum, M. Yang, L. Lee, D. Freeman, W. Xu, S. Guadarrama, I. Fischer, E. Jang, H. Michalewski, and I. Mordatch. Multi-game decision transformers, 2022. URL https://arxiv.org/abs/2205.15241.

J. Liu, T. Wang, P. Cui, and H. Namkoong. On the need of a modeling language for distribution shifts: Illustrations on tabular datasets, 2024. URL https://arxiv.org/abs/2307.05284.

X. Ma, F. Rizzoglio, E. J. Perreault, L. E. Miller, and A. Kennedy. Using adversarial networks to extend brain computer interface decoding accuracy over time. Aug 2022. doi: 10.1101/2022.08.26.504777. URL https://www.biorxiv.org/content/10.1101/2022.08.26.504777v1.

P. J. Marino, L. Bahureksa, C. Fernández Fisac, E. R. Oby, A. L. Smoulder, A. Motiwala, A. D. Degenhart, E. M. Grigsby, W. M. Joiner, S. M. Chase, et al. A posture subspace in primary motor cortex. *bioRxiv*, pages 2024–08, 2024.

M. J. Mender, S. R. Nason-Tomaszewski, H. Temmar, J. T. Costello, D. M. Wallace, M. S. Willsey, N. Ganesh Kumar, T. A. Kung, P. Patil, and C. A. Chestek. The impact of task context on predicting finger movements in a brain-machine interface. *eLife*, 12:e82598, jun 2023. ISSN 2050-084X. doi: 10.7554/eLife.82598. URL https://doi.org/10.7554/eLife.82598.

J. Merel, D. Carlson, L. Paninski, and J. P. Cunningham. Neuroprosthetic decoder training as imitation learning. *PLoS computational biology*, 12(5):e1004948, 2016.

J. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization, 2021. URL https://arxiv.org/abs/2107.04649.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM, Jan. 2019. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

K. K. Noneman and J. Patrick Mayo. Decoding continuous tracking eye movements from cortical spiking activity. *International Journal of Neural Systems*, page S0129065724500709, Oct. 2024. ISSN 0129-0657, 1793-6462. doi: 10.1142/S0129065724500709. URL https://www.worldscientific.com/doi/10.1142/S0129065724500709.

J. E. O'Doherty, M. M. B. Cardoso, J. G. Makin, and P. N. Sabes. Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology, May 2017. URL https://doi.org/10.5281/zenodo.788569.

C. OpenX, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Mart'in-Mart'in, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL https://arxiv.org/abs/2310.08864.

C. Pandarinath and S. J. Bensmaia. The science and engineering behind sensitized brain-controlled bionic hands. *Physiological Reviews*, 102(2):551–604, 2022.

C. Pandarinath, P. Nuyujukian, C. H. Blabe, B. L. Sorice, J. Saab, F. R. Willett, L. R. Hochberg, K. V. Shenoy, and J. M. Henderson. High performance communication by people with paralysis using an intracortical brain-computer interface. *elife*, 6:e18554, 2017.

F. C. Pei, J. Ye, D. M. Zoltowski, A. Wu, R. H. Chowdhury, H. Sohn, J. E. O'Doherty, K. V. Shenoy, M. Kaufman, M. M. Churchland, M. Jazayeri, L. E. Miller, J. W. Pillow, I. M. Park, E. L. Dyer, and C. Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=KVMS3fl4Rsv.

S. M. Peterson, S. H. Singh, B. Dichter, K. Tan, C. DiBartolomeo, D. Theogarajan, P. Fisher, and J. Parvizi. Ajile12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific Data*, 9(1):184, 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01280-y. URL https://doi.org/10.1038/s41597-022-01280-y.

T. M. Pham, T. Bui, L. Mai, and A. Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks?, 2021. URL https://arxiv.org/abs/2012.15180.

K. M. Quick, J. L. Mischel, P. J. Loughlin, and A. P. Batista. The critical stability task: quantifying sensory-motor control during ongoing movement in nonhuman primates. *Journal of Neurophysiology*, 120(5):2164–2181, 2018.

S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A generalist agent, 2022. URL https://arxiv.org/abs/2205.06175.

F. Rizzoglio, E. Altan, X. Ma, K. L. Bodkin, B. M. Dekleva, S. A. Solla, A. Kennedy, and L. E. Miller. Monkey-to-human transfer of brain-computer interface decoders. *bioRxiv*, 2022. doi: 10.1101/2022.11.12.515040. URL https://www.biorxiv.org/content/early/2022/11/13/2022.11.12.515040.

A. C. Rodriguez, M. G. Perich, L. E. Miller, and M. D. Humphries. Motor cortex latent dynamics encode spatial and temporal arm movement parameters independently. *Journal of Neuroscience*, 44(35), 2024.

P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, B. M. Yu, and A. P. Batista. Neural constraints on learning. *Nature*, 512(7515):423–426, 2014.

M. Sato, K. Tomeoka, I. Horiguchi, K. Arulkumaran, R. Kanai, and S. Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data, 2024. URL https://arxiv.org/abs/2407.07595.

S. Schneider, J. H. Lee, and M. W. Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://doi.org/10.1038/s41586-023-06031-6.

I. Schubert, J. Zhang, J. Bruce, S. Bechtle, E. Parisotto, M. Riedmiller, J. T. Springenberg, A. Byravan, L. Hasenclever, and N. Heess. A generalist dynamics model for control, 2023. URL https://arxiv.org/abs/2305.10912.

N. M. M. Shafiullah, Z. J. Cui, A. Altanzaya, and L. Pinto. Behavior transformers: Cloning k modes with one stone, 2022. URL https://arxiv.org/abs/2206.11251.

Q. Simeon, L. Venâncio, M. A. Skuhersky, A. Nayebi, E. S. Boyden, and G. R. Yang. Scaling properties for artificial neural network models of a small nervous system. In *SoutheastCon 2024*, pages 516–524. IEEE, 2024.

N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

I. H. Stevenson. Tracking advances in neural recording. Statistical Neuroscience Lab, University of Connecticut, 2023. URL https://stevenson.lab.uconn.edu/scaling/. Accessed September 6, 2024.

J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

S. J. Talukder, J. J. Sun, M. K. Leonard, B. W. Brunton, and Y. Yue. Deep neural imputation: A framework for recovering incomplete brain recordings. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022. URL https://openreview.net/forum?id=c9qFg8UrIcn.

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification, 2020. URL https://arxiv.org/abs/2007.00644.

R. Thapa, B. He, M. R. Kjaer, H. Moore, G. Ganjoo, E. Mignot, and J. Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals, 2024. URL https://arxiv.org/abs/2405.17766.

A. W. Thomas, C. Ré, and R. A. Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data, 2023.

A. E. Urai, B. Doiron, A. M. Leifer, and A. K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25:11–19, 2022. doi: 10.1038/s41593-021-00980-9.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=xmcYx_reUn6.

E. Y. Wang, P. G. Fahey, K. Ponder, Z. Ding, A. Chang, T. Muhammad, S. Patel, Z. Ding, D. Tran, J. Fu, et al. Towards a foundation model of the mouse visual cortex. *bioRxiv*, 2023b.

H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023c.

T. Wang, A. Roberts, D. Hesslow, T. L. Scao, H. W. Chung, I. Beltagy, J. Launay, and C. Raffel. What language model architecture and pretraining objective work best for zero-shot generalization?, 2022. URL https://arxiv.org/abs/2204.05832.

F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03506-2. URL https://www.nature.com/articles/s41586-021-03506-2.

J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

M. Wortsman, P. J. Liu, L. Xiao, K. E. Everett, A. A. Alemi, B. Adlam, J. D. Co-Reyes, I. Gur, A. Kumar, R. Novak, J. Pennington, J. Sohl-Dickstein, K. Xu, J. Lee, J. Gilmer, and S. Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ.

W. Xia, R. de Charette, C. Öztireli, and J.-H. Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision (ECCV)*, 2024.

C. Yang, M. B. Westover, and J. Sun. Biot: Cross-data biosignal learning in the wild, 2023. URL https://arxiv.org/abs/2305.10351.

J. Ye, J. L. Collinger, L. Wehbe, and R. Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=CBBtMnlTGq.

Z. Yuan, F. Shen, M. Li, Y. Yu, C. Tan, and Y. Yang. Brainwave: A brain signal foundation model for clinical applications, 2024. URL https://arxiv.org/abs/2402.10251.

Y. Zhang, Y. Wang, D. Jimenez-Beneto, Z. Wang, M. Azabou, B. Richards, O. Winter, T. I. B. Laboratory, E. Dyer, L. Paninski, et al. Towards a" universal translator" for neural dynamics at single-cell, single-spike resolution. *arXiv preprint arXiv:2407.14668*, 2024.

A Related works and a proposed taxonomy

Neural data is sufficiently diverse so as to support many distinct efforts to train large neural data models. The scale of pretraining is somewhat larger in the non-implanted modalities, where data is more abundant. The largest EEG models have reached a scale of 2.5K (Jiang et al., 2024) to 40K hours of data (Yuan et al., 2024), or higher volumes if also considering non-brain biosignals (EKG) (Yang et al., 2023; Thapa et al., 2024). Current fMRI models operate in the 1K (Thomas et al., 2023) to 7K (Caro et al., 2024) hour range. The largest models in these studies are in the 0.1B-1B parameter range. Intracranial modalities, including sEEG (Wang et al., 2023a; Chau et al., 2024), ECoG (Talukder et al., 2022; Peterson et al., 2022), and spiking activity (Wang et al., 2023b), have thus far been studied at an order of magnitude smaller scales of data and model size (20-1000 hours, <0.1B parameters).

Direct scaling on neural data modeling should be distinguished from NeuroAI efforts (Doerig et al., 2022) to measure how models of the human sensorimotor experience (e.g. language, vision, audio models) predict neural data (Antonello et al., 2024). However, as multimodal efforts begin to blur this distinction (Benster et al., 2024; Xia et al., 2024), care will be required to distinguish advances in modeling neural data, embodied data, or their interaction.

Comparing neural data models Current efforts to understand scaling in neural data Simeon et al. (2024); Sato et al. (2024) will have their reach limited by the specificity of every neural dataset. A meta-challenge for the field is understanding how different parameters (species, brain area, modality, task) impact scaling properties. This would be greatly aided by development of reporting practices for different neural data models. To facilitate comparison, we create a model card (Mitchell et al., 2019) for NDT3 in Section D. In addition to the standard model card, we propose reporting an additional taxonomy to aid comparisons across neural data models, using two concepts.

First: neural data models can be conceptualized as modeling slices of the *plenneural function*, inspired by the plenoptic function in vision (Adelson et al., 1991). The plenoptic function is a model of an idealized eye which parameterizes all possible images with 7 dimensions: 4D to describe the global spacetime of the view, 2D to describe viewing angle (spherical) or coordinate (Cartesian) of the image, and 1D for wavelength. Since neural data models are primarily interested in circumscribed systems rather than the physical world, a similar global coordinate system (e.g. 4D for all possible electric potentials) would be uninformative. We thus propose reporting more qualitative coordinates:

1. Identity: The network or individual being recorded.
2. Task: The behavior, stimuli, or other activity the network is reflecting.
3. Spacetime: Coordinates specified in a network-local coordinate frame (e.g. brain area).

Second: The modeled extent of this plenneural function is conveniently discretized in three resolutions in a Transformer-like sequence modeling framework: the token, the sequence, and the full training data. The token is the most granular unit of data being modeled; NDT3 models neural populations 32 neurons at a time, in 20ms bins. At the sequence input level, NDT3 models inputs from single humans or monkeys, across 128-256 neurons in 2 second snippets, while performing effectively one "movement." Finally, NDT3's pretraining spans dozens of individuals, records motor and premotor areas over 2.5K hours, over a variety of arm and hand movements.

B Supplementary Results

B.1 Extrapolation in center-out decoding is possible with restricted linear decoders.

In Section 3.2, we proposed that pretrained models like NDT3 should be able to generalize to held-out, and in particular, extrapolated reach angles. We hypothesized this due to the frequent appearance of 2D linear projections of high dimensional neural activity showing clear separation by reach angle (e.g. (Rodriguez et al., 2024)), which would imply NDT3 meta-learning of an explicit planar prior would enable our desired generalization. Here we make this intuition explicit, by constructing a linear decoder that generalizes to held-out reach directions, and thus illustrate that NDT3's failure to extrapolate is not due to an inherently unconstrained generalization task but is rather due to a

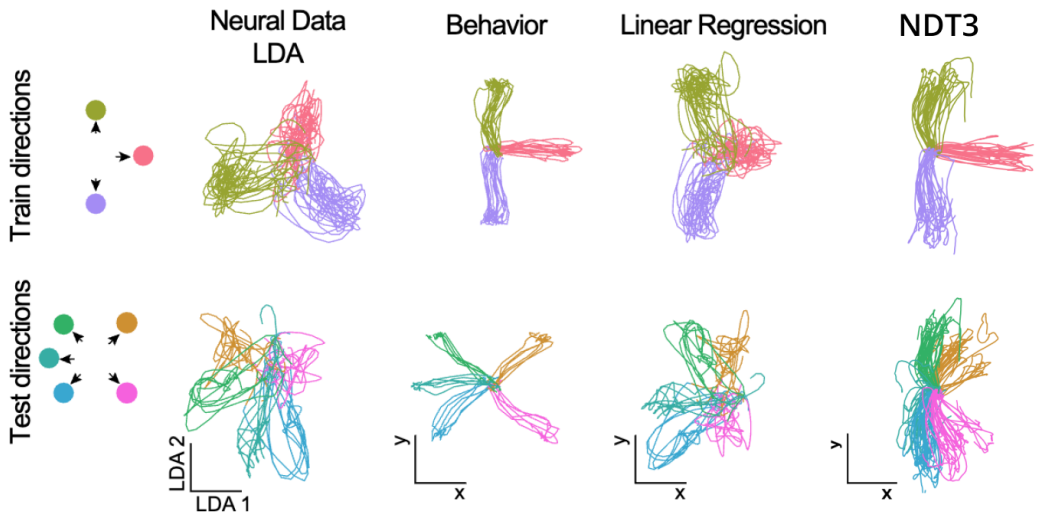

Figure 6. We illustrate how explicitly restricting to a linear decoder plane allow reconstruction of held-out behavior (Linear Regression), in contrast with NDT3 predictions which are restricted to held-in conditions even when shown three cardinal directions.

lack of proper objective (i.e. requires post-training to make generalization desirable) or a failure of pretraining.

To do this, we follow a classic neural data analysis procedure on a single session of the monkey J dataset (as in Fig. 4). As a reminder, the behavior here is an isometric task where the monkey generates isometric forces against a small box equipped with a six-degree-of-freedom load cell (JR3 Inc., CA). The forces were linearly mapped to control cursor movement: wrist flexion/extension moved the cursor left/right, and radial/ulnar deviation moved it up/down. The monkey had to move the cursor from a central position toward one of eight peripheral targets, in a classic center-out task. We extracted successful trials for each target from 0.5 seconds before to 1 second after movement onset.

With this preparation, we can now project the high-dimensional (96) neural activity at each timestep onto a candidate plane that reflects reach related variance, after which a simple rotation would allow generalized decoding. If we were fitting neural activity from all conditions, the top PCs from Principal Components Analysis (PCA) would typically be sufficient to identify this plane. Since we would like to find our general decoding plane without fitting all conditions, PCA alone is not reliable enough to extract the plane we desire. We thus first fit PCA to the three held-in conditions, and then used Linear Discriminant Analysis (LDA) on the top 10 PCs to find the 2D plane that best separated the three directions. This yields a plane where neural activity is well separated by their reach direction in a consistent manner for train and test directions (Fig. 6 Neural Data and Behavior). This allows a ridge regression to generalize to from held-in trajectories (variance accounted for (VAF): 0.70 ± 0.0) to both the interpolated (VAF: 0.48 ± 0.02) and extrapolated held-out (VAF: 0.44 ± 0.05) directions (Fig. 6 Linear Regression). In contrast, NDT3 trajectories are, as before, clearly constrained to held-in directions.

B.2 Ablations

We ablate the major design decisions made to enable NDT3's large scale pretraining. These ablations give us confidence that NDT3 overcomes the basic challenges we encountered in development, but compute restrictions prevent more exhaustive comparisons or exploration of model design space. We encourage further work exploring the influence of different hyperparameters. In these plots, we distinguish validation split performance and evaluation split performance, which is computed by batch-mode prediction (not the costly streaming evaluation used throughout main experiments).

Covariate dropout We find the default next-step prediction objective fails for learning decoding of highly autocorrelated covariate timeseries, perhaps because simply relying on teacher-forced

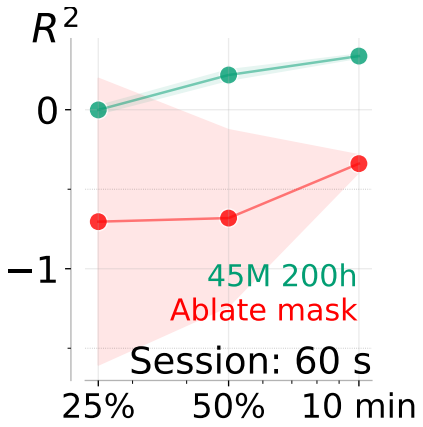

Figure 7. Ablation of covariate masking on an open 2D Cursor + Click dataset. Covariate inputs are completely masked in inference for the default NDT3, and autoregressively generated in the ablation.

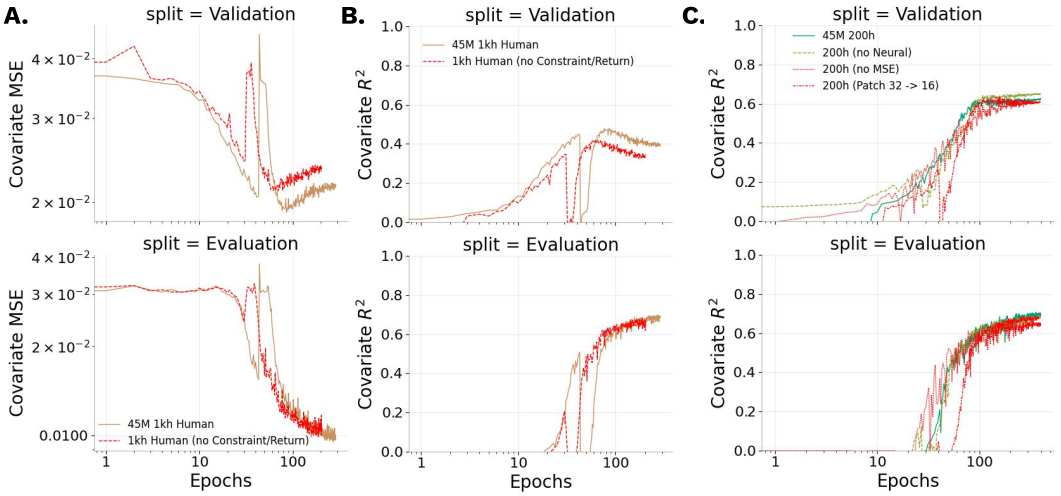

Figure 8. Ablations evaluated based on upstream evaluation split. **A., B.** Ablation of BCI control tokens **C.** Ablation of neural objective and covariate MSE objective in favor of classification over quantized covariates.

behavioral inputs provides a severe shortcut that prevents learning of a proper neural to behavior decoding map (Bachmann and Nagarajan, 2024). Different time-series models have addressed this by adopting convolutional input-output layers (Lea et al., 2016), tokenizing along temporal dimensions (Das et al., 2024), or learning with contrastive objectives (Chau et al., 2024; Kostas et al., 2021). We avoid introducing architectural modifications and instead adopt a simple dropout procedure that masks a portion of covariate inputs some fraction of the time. Specifically, on every training batch, two random numbers are drawn. The first, $M \sim U[0, 1]$, determines what fraction of covariate inputs should be masked. On 90% of batches, we also sample $T \sim U[0, 2]$ seconds, such that the mask is only applied after timestep T. That is, on 90% of batches, the model is provided a prefix-prompt. We do not block losses on this prefix as in prefix-LMs (Wang et al., 2022). Pretraining metrics for validation and evaluation are always computed with a prefix and full masking of non-prefix timesteps. In Fig. 7, we ablate covariate masking (which also removes the prefix logic), and tune on a 2D Cursor + Click task. The ablated model performs subtrivially with student-forced predictions provided as input at test time. Note that the ablated model performs trivially with masked inputs (not shown).

BCI-phase and return conditioning NDT3's pretraining includes several hundred hours of BCI control data, where the covariates were set by another decoder. We introduced phase and return conditioning tokens to differentiate the several types of BCI control data from recorded behavior. Specifically, in BCI data, NDT3 receives input tokens specifying what fraction of the behavior reflects

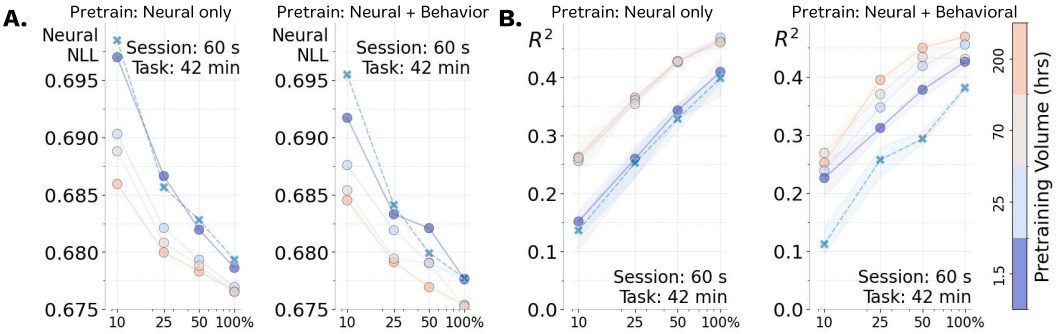

Figure 9. Scaling of unsupervised pretraining After pretraining models up to 200 hours with only the neural reconstruction objective, we fine-tune models in a similar multiscale evaluation as in Section 3.1, with a newly initialized behavioral readout. We use the CST task here. **A.** left shows neural loss scales with neural-only pretraining. The right panel plots the neural loss, also present in the standard evaluation, also scales with joint pretraining. **B.** left shows that with neural pretraining, downstream decoding performance saturates at a flat performance after 25 hours. This is compared against the nonsaturated scaling from joint pretraining. Colorbar is common for all plots, and Xs are from-scratch NDT3 models.

neural input (BCI control is on) vs programmatic input (BCI control is off, as in open loop BCI calibration). Further, we provide inputs encoding reward (trial success) when trials change, and return (future reward over a 10 second horizon, which crosses data boundaries). This design is intended to evaluate the potential for a Decision-Transformer like offline learning strategy for improved online control, but we do not discuss this in this work. In Fig. 8A, B, we focus on whether these inputs improves pretraining loss and R^2 in validation splits, which has contains BCI data, and the held-out evaluation split containing only monkey behavior. The figures show that the ablation significantly decreases validation split performance, and causes a slightly earlier stopping point leading to worse evaluation performance. Note both models early before the full training budget of 400 epochs.

Neural reconstruction objective All main NDT3 models used a neural reconstruction objective inherited from the self-supervised learning pretraining from NDT2. We ablate this choice post-hoc and see it may actually minorly harm pretraining (validation split), though the neural objective doesn't harm evaluation split decoding (Fig. 8C). Note the scalar weighting of neural vs covariate objectives were set to be roughly balanced in pretraining. Section B.3 provides a downstream analysis on the standalone value of the neural reconstruction objective.

MSE over classification In robotics and certain generalist models (Schubert et al., 2023), continuous action spaces are sometimes better decoded and controlled when quantized (Shafiullah et al., 2022). This is because MSE is an insufficient objective when the output distribution is multimodal (e.g. one of two possible paths in robotics). While it seems unlikely that the close relationship between movement behavior and motor cortex is multimodal, multimodal behavior may be appropriate when pretrained on heterogeneous data, i.e. when similar neural activity corresponds to different behavior in two datasets. We attempted such a quantization, including HL-Gauss smoothing (Farebrother et al., 2024) which we found to help; but this does not recover the performance of the default MSE objective (Fig. 8C) on the evaluation split. We found this performance gap persisted under fine-tuning (not shown). This suggests that NDT3 is differentiates neural data inputs from different datasets.

Patch size NDT2 and NDT3 both tokenize neural data by patching them into fixed size clusters. It is unclear whether transfer learning might occur for sub-token features, which motivates the use of smaller tokens in larger datasets that might afford it (Caron et al., 2021). We change patch size to 16 and show this performs slightly worse in the 45M 200h model (Fig. 8)C. Smaller patches (and subsequent increased neural tokens) may be more beneficial in the larger scale models, but their benefit must be weighed against their increased compute burden.

B.3 Isolated scaling in neural data

Due to NDT3's joint modeling of behavior and neural data, it is difficult to dissociate whether scaling gains in behavior come from improved behavioral or neural priors. To assess whether NDT3 can scale solely from neural data modeling, we pretrain a new set of 45M parameter models up to 200

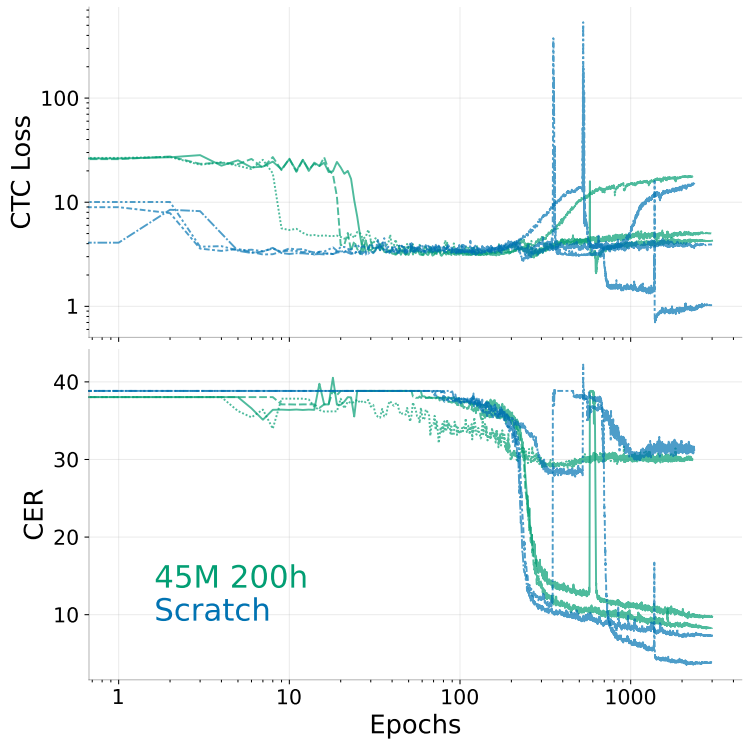

Figure 10. Three regimes of NDT3 training for handwriting decoding. We show validation loss and character error rates for example runs of from-scratch and fine-tuned NDT3s.

hours with only causal neural data modeling objective. As before, we then tune to a downstream decoding task, in this case, a Critical Stability Task dataset (CST). From a representation probing perspective, improved downstream performance implies higher quality neural representations. We use the standard single-stream autoregressive modeling objective as in the rest of this work in the downstream setting, we find direct linear probing of neural representations perform worse. Fig. 9 compares the scaling on downstream neural and behavioral metrics after the standard fine-tuning procedure.

Fig. 9A shows that downstream neural reconstruction improves with increased pretraining data either when using only the neural objective or both neural and behavioral objectives (as in the standard setting). The joint pretraining achieves advances neural metrics in all settings, illustrating that decoding behavior is a complementary objective to neural data reconstruction even for representation learning.

Fig. 9B contrasts decoding curves in the two pretraining settings, in that neural pretraining has saturated decoding after just 25 hours of pretraining. This is consistent with the interpretation that the behavioral readout reflects only one aspect of the neural data. Together with the neural metric plots, this analysis shows scaling over solely neural data is possible, but also that decoding behavior is a complementary pretraining objective for improving neural representation learning and decoding.

B.4 Pretraining does not benefit FALCON H2 (Handwriting)

We also evaluated NDT3 for decoding of letters in a human-open loop handwriting task (FALCON H2). Although this is also a motor cortical decoding task, we excluded H2 from NDT3's aggregate evaluation since it is a sequence-to-sequence as opposed to continuous task. To apply NDT3 to this task, we pool neural tokens at each timestep and add a linear projection and optimize with a CTC loss (Graves et al., 2006). We maintain the default neural reconstruction loss and causal attention mask, and do not apply data augmentation.

Note that RNNs are the current standard architecture for communication tasks like H2 (Karpowicz et al., 2024; Willett et al., 2021). Training and tuning was less stable than for our continuous decoding

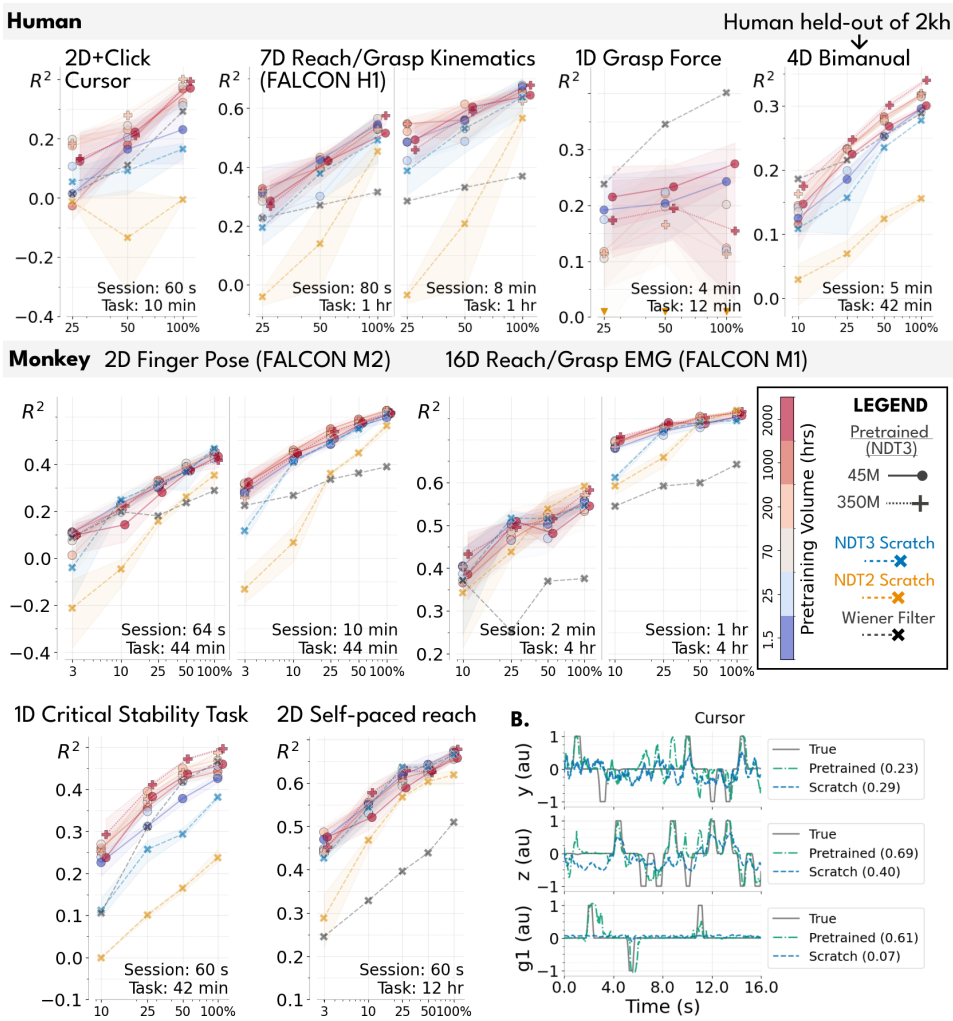

Figure 11. A. Fine-tuning evaluations for individual datasets. Performance on both held-out (left) and held-in (right) splits are shown side by side by FALCON datasets. We shade the standard deviation of 3 model seeds in fine-tuning. Different tasks show substantial variability in benefit from pretraining. B. We show example predictions of a pretrained (45M 200h) and from-scratch NDT3 for the 2D + Click Cursor task to give a sense of what different prediction performances mean in terms of open loop data prediction. Numbers in legend are the R^2 for that model's predictions in the shown snippet.

tasks and required more extensive hyperparameter tuning, perhaps because the overall dataset size remains small (<1k samples), specific parameters are listed in the codebase. We observe three regimes in both training and fine-tuning. First, the model can fail to achieve an initial learning period. Second, the model can achieve reasonable nontrivial solutions, comparable to expected performance for unaugmented RNNs (though we do not quantify this). Third, some models will exhibit learning instabilities that resolve in significantly improved performance. We illustrate these regimes in example validation curves below. Overall, the third regime is rarely achieved. More relevant to the main narrative of this work, fine-tuning appears to degrade both final solution quality and reduces the range of nontrivial hyperparameters (not shown). Investigating a sequence to sequence objective over CTC loss would be valuable future work.

B.5 Multiscale decoding on individual motor tasks

Fig. 11A plots model performance for each of the 31 evaluation settings we study in the eight primary evaluation datasets we use. Studying any individual dataset will yield variable conclusions on whether pretraining structure is helpful, underscoring the need for proposed foundation models

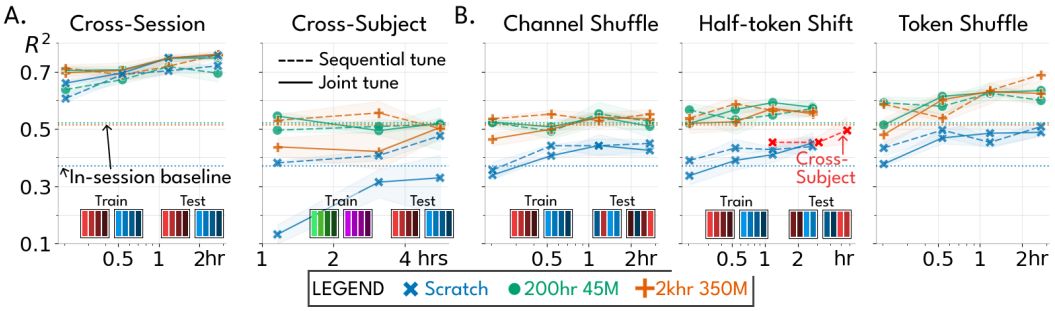

Figure 12. A replication of Fig. 4A/B, but showing results for both sequential and joint fine-tuning in each setting. As before, each model here tunes with some data from other settings (e.g. cross-subject data for the cross-subject panel), and a fixed amount of data on the test session. Sequential tuning first tunes on other-setting data, and then tunes on test session data. Joint tuning uses all data at once. In Fig. 4A/B, we only showed the better choice for each panel, i.e. we showed joint tuning for cross-session data, and sequential tuning for the other settings. In addition, we overlay how the cross-subject tuned models perform when applied to half-token shifted test data, confirming that subject and session data transfer similarly to shifted test data.

to be evaluated across many different datasets. Here specifically we see the most clear scaling with pretraining data (color gradient with red on top) in the Critical Stability Task and Bimanual Task. FALCON tasks and Self-paced Reach appear minimally affected by scaling pretraining data, in that either pretrained models are generally slightly above a from scratch model at all data scales with no particular best pretrained model. The 2D + Click and Grasp datasets uniquely show high variability in model performance and strong degradation of the 350M 2 khr model at low data scales. Grasp instability was so high that we trained 9 seeds instead of the standard 3 to better estimate model performance. We propose this degradation is due to the instability of full fine-tuning of large models at the extremely low data scales these datasets present (e.g. 2.5 minutes at the 25% scaling). Finally, we remind that the 2D + Click, FALCON H1, and 1D Grasp Force tasks are datasets from human participants that are included in the 2 khr pretraining. Surprisingly, we see no particular benefit to the 2 khr model.

These scaling plots also provide more precise context for baseline performance. NDT2 performs particularly poorly in the low data regime, while Wiener Filters perform poorly in the high data regimes.

In Fig. 11B, we illustrate qualitative predictions on private datasets. These visualizations show a diversity in covariate timescales and structure. They also illustrate that the summary R^2 obscure several features of model predictions. For example, pretrained models in Cursor Y tend have false positive deflections in movement. R^2 also is not easily comparable in tasks with continuous dynamics (CST) vs. transient dynamics (Cursor G1).

B.6 Sequential tuning is similar to joint tuning

In Section 3.2, we showed that channel shuffling and half-token shifts were sufficient to reduce cross-session transfer to the extent of cross-subject transfer. Here we add a methodological subtlety on how the tuning is done. Given cross-context data and a test session, we can either jointly tune on all data (as we do in our primary evaluation), or sequentially tune on the cross-context data and test session. We find sequential tuning is particularly necessary for successful subject transfer of from-scratch NDT3 models, but that it slightly underperformed joint tuning on cross-session models (Fig. 12A). Seeing that sequential tuning is mainly advantageous for from-scratch models, we speculate that sequential tuning is particularly helpful for filtering learning signals in highly heterogeneous data. In Fig. 4A/B, for clarity, we reported jointly tuned results for cross-session data and sequentially tuned results otherwise.

B.7 Aggregate performance on all NDT3 models with significance tests

We additionally report the average performance of 25 hr and 70 hr pretrained NDT3 models, placed in context with the models from Fig. 3D, in Fig. 13A. Note each successively larger data scale uses a strict superset of data from smaller scales. We also provide the p-values computed for the significance

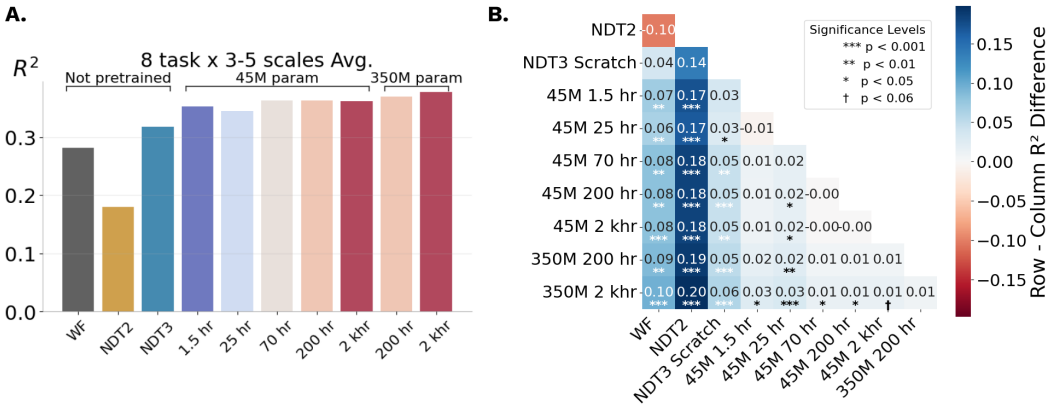

Figure 13. A. A replication of Fig. 3D including an additional 25 hr and 70 hr model. The two additional models show the precise performance we measure may be noisily related to to pretraining data scale. B. Heatmap of differences between performances of pairs of models with significance tested with FDR-corrected pairwise t-tests. Note that coloring is used here to indicate differences, not significance. Positive numbers with significance indicate the row model outperforms the column model.

of the difference between each pair of models in Fig. 13B. P-values are computed as FDR-corrected pairwise t-tests. The 350M 2 khr model has $p < 0.06$ improvements over all but the 350M 200 hr model. Interestingly, all other pretrained models, except the 25 hr model, appear equivalent, at least statistically. We presume this is due to the fact that our evaluation of 31 task settings may be insufficiently large.

C METHODS

C.1 METRICS AND EVALUATION

Throughout this work we evaluate offline decoding of continuous covariates timeseries. The metric we specifically use is the coefficient of determination, R^2, as computed by scikit-learn's *r2_score* function. R^2 is a useful metric over MSE as 1 represents perfect prediction and 0 is the score achieved by best-guess baseline, the mean of the data. In pretraining, R^2 is computed over the flat average of all covariate dimensions, since each datapoint has differing covariate dimensionalities. In evaluation, R^2 is computed as a variance-weighted average of R^2s in each covariate dimension. Another difference between training and evaluation metrics is that training predictions are made over batched data, while evaluation predictions are mostly computed in a *streaming* fashion. Streaming requires continuous neural data across different behavioral epochs, and so cannot be performed for the Oculomotor and CST datasets. We also omit it for the motor cortex self-paced reach dataset, which has a very large evaluation split. Streaming allows timesteps at the beginning of each sequence to leverage neural context from the preceding sequence, which raises performance slightly, as shown in the continuous vs trialized analysis (Section 3.3). We limit history in streaming evaluations to the max history seen in tuning (1 second).

C.2 TRAINING

Pretraining hyperparameters were manually tuned in preliminary experiments at the 45M parameter models on small datasets. 350M models diverged at the chosen $4e-4$ peak LR, so we lowered peak LR to $1e-4$. For tuning, the explored LRs are $1e-4, 3e-4, 5e-4$ for training from scratch and $3e-5, 1e-4, 4e-4$ for fine-tuning. While this is far from an exhaustive search, we show in Fig. 14 that other regularization hyperparameters are set to reasonable defaults such that this sweep finds near optimal results for both a from scratch model and fine-tuning the 45M model. Fine-tuning, like pretraining, is early stopped with a patience of 100 epochs. Batch size is uniformly set to 16K in pretraining, and scaled to be roughly 10-20% of dataset size in fine-tuning. NDT3 from-scratch models were trained at the 11M parameter range. Exact model configurations for different experiments are documented in the codebase.

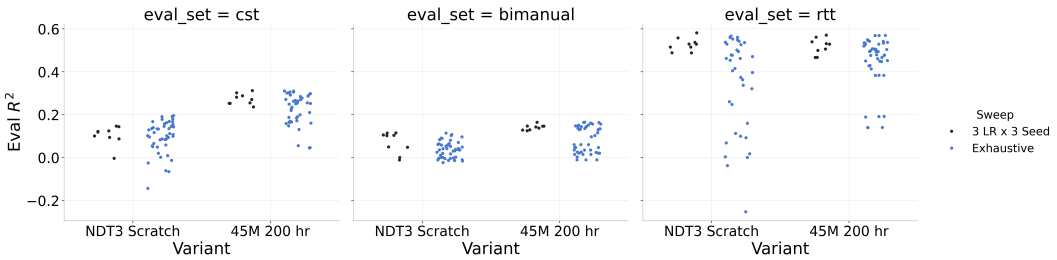

Figure 14. For 3 monkeys datasets at 10% scale, we extend a HP sweep to 5 LRs and dropout in $[0.0, 0.1, 0.3]$ (vs default 0.1). For fine-tuning, we also sweep weight decay in $[0.0001, 0.01, 0.1]$ (vs default 0.1), while for from-scratch models we also sweep Transformer width ($[256, 512, 1024]$) vs default 512. This yields a 45-model sweep on 1 seed. We compare the range of scores achieved by this larger sweep against the standard 3 LR x 3 seed sweep.

NDT3's simple architectural design allows us to train on batches from different tasks and dimensionalities. To avoid excess padding in training, we concatenate pretraining data that is otherwise discontinuous (trialized) into 2 second data. We do not add any separator tokens, as this does not appear to have a performance impact for language models (Geiping and Goldstein, 2022). With mixed-precision training, the 350M parameter NDT3 can fit the 4-8K tokens in each input context in the memory of 40G NVIDIA A100 GPUs. Thus we can restrict NDT3's pretraining parallelism to data-parallelism.

Using Kaplan et al. (2020)'s equation for FLOP computation, $C_{\text{forward}} = 2N + 2n_{\text{layer}}n_{\text{ctx}}d_{\text{attn}}$, we compute the footprint of the 350M 2kh model. We use about 0.9B FLOPs per token in the forward pass, and about 0.9T neural tokens processed over training, which yields a pretraining footprint of about 2.4e21 FLOPs.

C.3 Baselines

Wiener Filter The Wiener Filter baseline was cross-validated over regularization strength. We also swept history of neural input up to the max length provided to NDT, and reported the R^2 of the best WF according to test data in primary evaluation (slightly advantaging the WF). Generalization plots in Section 3.3 report the performance of WF models at these different histories. For evaluating angular generalization, WFs were only swept up to 1s history due to memory limits; performance was not varying substantially with history so we do not expect this to have impacted conclusions. The WF was for simplicity directly fit on the concatenated trial data, which may have slightly negatively impacted its performance in trialized datasets (Oculomotor, CST, Generalization analyses).

In the primary evaluations in Section 3.1, we considered WFs fit either independently per session in a dataset or jointly on all sessions, which is helpful for sessions in very low data regimes. We report the better of the 2. In generalization analyses, for simplicity, we only report joint fits, which may cause a slight downward bias in performance.

Dataset	**Patience**	**Held-In R^2**	**Held-Out R^2**
H1	100	$0.567_{\pm 0.034}$	$0.453_{\pm 0.030}$
H1 (reproduction)	250	$0.628_{\pm 0.011}$	$0.517_{\pm 0.016}$
H1 ((Karpowicz et al., 2024))	250	0.62	0.52
M2	100	$0.563_{\pm 0.015}$	$0.352_{\pm 0.028}$
M2 (reproduction)	250	$0.582_{\pm 0.002}$	$0.391_{\pm 0.009}$
M2 ((Karpowicz et al., 2024))	250	0.63	0.43

Table 1. NDT2 H1 and M2 results when trained with 100 epochs of patience (this work) in fine-tuning vs 250 as in Karpowicz et al. (2024). We report mean and standard deviation of 3 model seeds on the FALCON evaluation (which is in turn a cross-session mean).

NDT2 NDT2 baselines were prepared with its public codebase. Max context length and patience were held constant across the models. This restriction to a patience of 100 accounts for some difference with the reported FALCON benchmark results in Karpowicz et al. (2024), as we note in Table 1. Other choices were left to NDT2 defaults. For example, NDT2 uses z-score normalization, which we

kept. A major change to the NDT2 approach, for simplicity, is that we jointly trained NDT2 with its neural reconstruction loss (masking of 25%) and supervised decoding loss. This is true for all eight evaluation tasks except CST, where we used only the supervised decoding loss as the token dropout used in reconstruction can dropout all neural input. NDT2 hyperparameters were not explored widely, which likely is a source of its mediocre performance in this work. We did however sweep NDT2 over 2 model sizes (20M and 72M parameters) in addition to the standard 3 learning rates, which provides it twice the budget as NDT3.

C.4 Pretraining and Evaluation datasets

Pretraining datasets were comprised of historical data from several labs, the rough composition of which is shown in Fig. 2B. The evaluation behavior used during pretraining was reaching in 2 monkeys. The first monkey dataset came from a public release (Flint et al., 2012), and the second from a private dataset (REDACT lab). The latter had center-out reach in standard conditions and under visual feedback perturbations. The monkey in the second dataset is also present in the 1khr monkey and 2kh and up model dataset sizes, though performing in a different set of experiments.

Inherent to the process of large-scale scraping is a loss of detail on what precise tasks were used, so we only have a qualitative description of tasks we believe are well represented. NDT3 trains on a wide variety of reaching behaviors from relatively constrained (2D center-out reaching to fixed number of targets) to relatively unconstrained (self-paced, more targets, potentially 3D) and under experimental manipulations (delayed onset, multiple targets, different error thresholds requiring more precision). These reaching behaviors are described in both endpoint kinematics and as EMG. A smaller fraction of pretraining data are isometric and force related (force exerted against manipulandums) for wrist and arm motion. Human datasets contain a variety of iBCI tasks, with closed loop datasets reflecting both high and low quality control. These tasks include reach and grasp behavior from 1-10 degrees of freedom, as well as some individuated finger tasks for clicking.

We detail evaluation datasets in Table 2. Three datasets come from the FALCON benchmark (Karpowicz et al., 2024), two are based on public datasets ((O'Doherty et al., 2017; Deo et al., 2024)), and three are private. Note we avoid the Neural Latents Benchmark (Pei et al., 2021) as it does not directly measure decoding performance. For each evaluation dataset, we specify a tuning split and an evaluation split. Only tuning split data is changed when varying data scale. Tuning and evaluation splits are block-contiguous, i.e. trials are not interleaved, for better downstream applicability.

C.5 Generalization analyses and Further evaluations

Intra-session generalization Posture, spring, and angular generalization evaluate OOD performance in the standard setup of comparing in-distribution and out-of-distribution performance directly (with changes in the underlying evaluation dataset) The intra-session temporal shift analysis is evaluated in an inverted, slightly more rigorous setting. Specifically, we trained two sets of models on the two different temporal blocks, and evaluated on an evaluation split in the later block, rather than only training on the early block and evaluating on both blocks. This way, the OOD shift is measured with respect to the same evaluation dataset.

C.6 Architectural Details

NDT3 adopts several architectural innovations used in recent Transformer models. These were compared against baselines in preliminary experiments, but formal ablations in the final experimental setting were not conducted. We defer full description of the Transformer dimensions to the public codebase.

- FlashAttention 2 (Dao, 2023) is used to increase training and inference speeds. On the NERSC Perlmutter cluster, with FA2, 45M NDT3 trained at about 270M neural tokens per 40G A100 hour, 350M NDT3 trained at about 70M neural tokens per A100 hour. Note, FA2 also enables use of the 350M model for real-time (<20ms) inference latency.
- Positional Embeddings (Su et al., 2023): Rotary embeddings are applied to indicate the real-world timestep of every input token. Additionally, 48 categorical learned embeddings are reserved to distinguish token modality and position within a timestep (10 for neural,

16 for covariates, 16 for covariate constraints, 1 for reward/return, 1 for dummy tokens, remainder unused).

- QK Normalization (Dehghani et al., 2023; Wortsman et al., 2024): An additional layer norm is applied to the query and key embeddings, before the rotary embeddings, which helped stabilize training of the 350M parameter models.
- No context embeddings (Ye et al., 2023): Differing from NDT2, no learned embeddings for disambiguating input datasets were prepended to each input. This was removed for simplicity. Per GATO (Reed et al., 2022) and language modeling practices, we instead leave task / dataset disambiguation to the modeling process: In pretraining, the covariate maskout strategy allows for many tasks to be specified in-context (as later behavior can be inferred on the basis of earlier neural-behavioral token relationships). In fine-tuning, the tuning dataset already uniquely specifies the function to be learned.
- Cross entropy loss for spiking data prediction: We used the standard cross entropy loss to classify spike count over the Poisson loss common in many neural data architectures. Since the overall ablation of neural objective shows no large impact in this work, it is likely that this decision should be evaluated with neural data related tasks rather than decoding.

We document the Transformer model shapes considered in our work in Table 3. This shape is not systematically explored in our work, and is by historical artifact, slightly different than the shapes used in NLP/CV. Embedding parameters are negligible. One possible area of interest is that the feedforward expansion factor is 1 in our model, i.e. the MLP dimension is low. If MLPs do serve as memory stores in Transformers (Geva et al., 2021), increasing this shape may yield more performant model size scaling, given the heterogeneity of our datasets.

D NDT3 Model Card

The card is currently only provided in the codebase.

Table 2. Evaluation datasets used for multiscale decoding and generalization analyses. The references provide extended description of the behavioral task. Dashed line separates datasets for Section 3.1 and for analysis. Datasets use unsorted multi-unit activity and are processed in 1s chops unless otherwise mentioned.

Dataset	Description
FALCON H1, M1, M2 (Karpowicz et al., 2024)	3 separate single-subject multi-session datasets for different iBCI tasks. Data comes in a high data split (held-in), and a low-data split (held-out), with the intention on identifying methods that can achieve parity in the two settings. H1 is an open loop human dataset for calibrating 7D reach-and-grasp in a robot arm. M1 is a monkey reach-and-grasp task to different objects with EMG recordings. M2 is a monkey 2D finger movement task with manipulandum-measured kinematics. Scaling scores are reported on the test set.
Self-paced reach (RTT) (O'Doherty et al., 2017)	Monkeys reach for random targets one at a time in a small planar workspace. We decode 2D arm velocity in monkey Indy. Has neural data from M1 and S1, we use M1 in Section 3.1 and Section 3.2 and S1 in Section 3.3.
Bimanual Cursor Control (Deo et al., 2024)	A human open loop dataset where the participant attempts movement of one or both hands to control two cursors.
2D Cursor + Click (private)	Cursor control is a classic iBCI endpoint (Pandarinath et al., 2017; Wolpaw et al., 2002; Jarosiewicz et al., 2015). Two human participants attempt movement according to visually cued cursor movement and audiovisual click cues. We also use this dataset for trial structure analysis in Section 3.3.
Grasp force (private)	A open-loop dataset with two human participants attempting isometric power grasps. Specifically, participants were asked to match force output according to visual cues in a Mujoco environment. Grasps cued were both static (instant onset, hold, and offset) or dynamic (gradually increasing force). This dataset is valuable for human iBCI study because force modulation is required in many motor behaviors, and grasp force has primarily only been characterized in monkeys until now (Branco et al., 2019). Uses 2 second intervals due to long behavior timescale. We expect this dataset can be released by end of 2024.
Critical Stability Task (Quick et al., 2018) (private, trialized, sorted)	A monkey dataset collected to study continuous control relative to ballistic movement. The monkey balances a virtual cursor on a 1D workspace for up to 6 seconds.
Posture-varied Center-Out (Marino et al., 2024) (private, trialized, sorted)	A monkey center-out task, but the monkey's hand is adjusted to one of 6 different starting positions. We use the central position as center and the rest as edge.
Spring-load (Mender et al., 2023)	A monkey moves fingers, clamped together in a manipulandum for effective 1DoF, is neutral or under spring load.
Center-out, Monkey J (Ma et al., 2022) (trialized)	Used in Section 3.2. A monkey performs an isometric center out task. Forces are measured by the manipulandum and converted to cursor velocity signals.
Center-out, Monkey V (private, trialized)	Used in Section 3.2. A monkey reaches to one of 8 radially arranged targets by moving a manipulandum (Kinarm).
Oculomotor pursuit (Noneman and Patrick Mayo, 2024) (private, trialized, sorted)	A monkey visually tracks (via smooth pursuit) a target that moves from center of workspace to one of four directions. A few dozen neurons are recorded on probes in each of frontal eye field (FEF) and area MT. We decode pupil velocity. The small number of neurons in this dataset required resetting NDT3 neural readin/readout layers.
FALCON H2	Human open loop dataset where a participant attempts movement to write letters cued on a screen (Willett et al., 2021; Fan et al., 2023). The large number of timesteps in this dataset required resetting NDT3 neural readin/readout layers (to use fewer neural tokens).

Model	Layers	Width	MLP Size	Heads	Parameters (M)
NDT3 Base	6	1024	1024	8	45
NDT3 Big	12	2048	2048	16	350

Table 3. Transformer Model Shapes used in this work.