Data Heterogeneity Limits the Scaling Effect of Pretraining Neural Data Transformers

Linxing Preston Jiang[1,*], **Shirui Chen**[2], **Emmanuel Tanumihardja**[1], **Xiaochuang Han**[1], **Weijia Shi**[1], **Eric Shea-Brown**[2], **Rajesh P. N. Rao**[1]

[1]Paul G. Allen School of Computer Science & Engineering
[2]Department of Applied Mathematics
University of Washington, Seattle, WA, USA
[*]prestonj@cs.washington.edu

Abstract

A key challenge in analyzing neuroscience datasets is the profound variability they exhibit across sessions, animals, and data modalities–i.e., heterogeneity. Several recent studies have demonstrated performance gains from pretraining neural foundation models on multi-session datasets, seemingly overcoming this challenge. However, these studies typically lack fine-grained data scaling analyses. It remains unclear how different sources of heterogeneity influence model performance as the amount of pretraining data increases, and whether all sessions contribute equally to downstream performance gains. In this work, we systematically investigate how data heterogeneity impacts the scaling behavior of neural data transformers (NDTs) in neural activity prediction. We found that explicit sources of heterogeneity, such as brain region mismatches among sessions, reduced scaling benefits of neuron- and region-level activity prediction performances. For tasks that do exhibit consistent scaling, we identified implicit data heterogeneity arising from cross-session variability. Through our proposed session-selection procedure, models pretrained on as few as five selected sessions outperformed those pretrained on the entire dataset of 84 sessions. Our findings challenge the direct applicability of traditional scaling laws to neural data and suggest that prior claims of multi-session scaling benefits may be premature. This work both highlights the importance of incremental data scaling analyses and suggests new avenues toward optimally selecting pretraining data when developing foundation models on large-scale neuroscience datasets.

1 Introduction

Recent advances in foundation models have revolutionized the modern machine learning paradigm. Across domains such as language and vision, it has been shown that "pretraining" a generic model on large-scale data before "finetuning" it to the actual tasks achieves much better performance than task-specific models [1, 2, 3]. This success has inspired similar efforts in systems neuroscience, where the goal is to develop foundation models trained on large, multi-session, multi-animal neural datasets of neural activity recordings. However, neural recordings pose unique challenges: data collected across brain regions, sessions, and individuals often exhibit substantial variability [4, 5, 6]. Even within the same recording session, stochasticity of neuronal firing and uncontrolled behavior can lead to significant trial-to-trial variability [7, 8, 9]. Furthermore, neural data can be non-stationary due to synaptic plasticity that induces gradual changes in population dynamics across days [10, 11]. These challenges raise a key question: Can neural foundation models overcome these sources of heterogeneity and learn more generalizable representations with more pretraining data?

Foundation Models for the Brain and Body NeurIPS 2025 Workshop

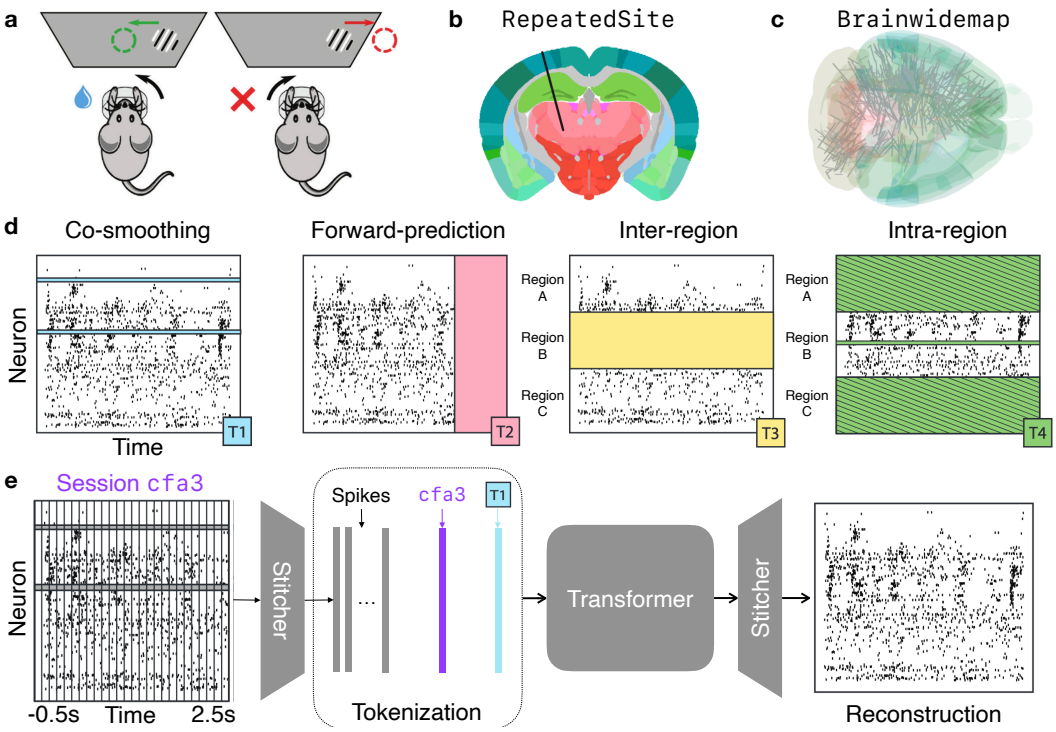

Figure 1: **Experimental Setup. (a)** Schematic of the visual decision-making task performed by mice. **(b)** Planned probe insertion location (black line) for all sessions in the `RepeatedSite` dataset. **(c)** Different probe insertion locations (gray lines) for different sessions in the `Brainwidemap` dataset. **(d)** Four different masking schemes of raw spike counts for model training. **(e)** The model architecture. Sub-figures adapted from [4, 5, 17, 14]. See text for details.

While several recent studies have demonstrated performance gains from multi-session pretraining on a wide range of encoding and decoding tasks, they typically lack fine-grained scaling analyses on the benefits of gradually increasing pretraining data [12, 13, 14, 15]. Most comparisons are limited to models trained on single sessions versus entire datasets with few increments in the middle, making it unclear how different sources of heterogeneity impact performance scaling. Moreover, it remains unknown whether all pretraining sessions contribute equally to downstream performance improvements. As pretraining scales to thousands of sessions and hours of data [16, 13], understanding the scaling behaviors of the model becomes increasingly critical.

In this work, we systematically investigate how data heterogeneity affects the scaling behavior of neural data transformers (NDTs) [18, 14, 16] using two large-scale datasets released from the International Brain Laboratory [5, 17]. These datasets differ in the consistency of recorded brain regions across sessions, offering us an opportunity to study how different levels of brain region heterogeneity in pretraining affect scaling. We further examine the effects of implicit heterogeneity such as session-to-session variability. Through a proposed session-selection procedure, we identify the impact of each pretraining session on downstream performance improvements. Our main findings include:

- We found that greater region-wise heterogeneity in pretraining data led to reduced improvements of neuron- and region-level activity prediction performances.
- To identify implicit heterogeneity, we propose a session-selection procedure based on single-session finetuning performances that effectively identified most beneficial single sessions for pretraining.
- Models trained with as few as five selected sessions outperformed those with randomly chosen sessions even when the full dataset was used, demonstrating the impact of session-to-session variability in performance scaling.

Together, these findings suggest that previous claims regarding the scaling benefits of pretraining without detailed incremental experiments may be premature, pointing to the need for rigorous scaling analyses in future work on neural foundation models to accurately assess the promise of large-scale pretraining.

2 Experimental Setup

Figure 1 summarizes the experimental setup used throughout our study, which mostly follows Zhang et al. [14] whose experiments were conducted on a subset of the same `RepeatedSite` dataset we used. We discuss the datasets, training pipeline, and evaluation metrics in detail below. More details are included in Appendix A.

2.1 Datasets

We used two multi-brain-region, multi-animal/session datasets from the International Brain Lab (IBL) collected from mice. Animals performed a visual decision-making task where they detected the presence of a visual grating (of varying contrast) to their left or right and rotated a wheel to bring the stimulus to the center (Fig. 1(a)). The main difference between the two IBL datasets lies in how often brain regions were repeatedly recorded across sessions. In the `RepeatedSite` dataset (henceforth `RS`), each session attempted to record from the *same* brain regions (Fig. 1(b), black line shows planned electrode insertion position). In contrast, the `Brainwidemap` dataset (henceforth `BWM`) aimed to cover as many *different* brain regions as possible (Fig. 1(c), gray lines show planned insertion positions), leading to little repetition of regions across sessions.

We used 89 out of 91 sessions in `RS`, excluding two sessions with fewer than one hundred trials. Five out of 89 sessions were held out for finetuning and evaluation. We randomly selected 200 sessions out of 460 sessions in `BWM` for pretraining and 10 sessions for finetuning and evaluation. Trials within each session were randomly split into training, validation, and test sets using an 8:1:1 ratio. Each trial included three seconds of neural activity, spanning from 0.5 seconds before to 2.5 seconds after stimulus onset with 20 ms bins for spike counts. The data from each session is thus a three-dimensional (trials $\times$ timesteps $\times$ neurons) tensor of integer spike counts.

2.2 Training pipeline

Multi-masking scheme During training, input spike count vectors were masked in one of four ways, as illustrated in Figure 1(d): (1) Co-smoothing: selected neurons' activities are masked; (2) Forward-prediction: all neurons' selected timesteps are masked; (3) Inter-region: all neurons' activities in a selected brain region are masked; (4) Intra-region: selected neurons in a selected brain region, along with all other neurons in other brain regions, are masked. Models were trained to reconstruct masked input from the unmasked [1] by maximizing the log likelihood of the Poisson distribution, with the model outputs as the predicted firing rates. We also applied causal attention in NDT if inputs were masked with forward-prediction to ensure no future neural activities were used to predict the present. Zhang et al. [14] showed that this multi-masking scheme significantly outperforms a scheme using forward-prediction masks alone in spike prediction tasks.

Model architecture We used the neural data transformer (NDT) architecture by Ye and Pandarinath [18] that has been widely applied to neural encoding and decoding tasks [19, 16, 15]. NDT also achieves state-of-the-art performance on the IBL dataset we used [14, 15]. Since different sessions have different numbers of neurons recorded, a session-specific linear layer (encoding "stitcher") maps raw spike counts to spike embeddings (Fig. 1(e) left) whose dimensions are shared across sessions [20]. A session embedding and a masking scheme embedding are also appended to input sequences. Lastly, another session-specific linear layer (decoding stitcher) maps the output of the transformer back to reconstructed spike rates (Fig. 1(e) right).

2.3 Evaluation

Baseline To show the effect of scaling up pretraining data, we directly trained single-session models on the training set of each heldout session as the baseline models.

Table 1: Evaluation metrics of baseline and best pretrained models on RS and BWM. Higher values indicate better performance.

	RS		BWM	
Task	Baseline	Pretrained	Baseline	Pretrained
Co-smoothing (BPS)	0.765	**0.774**	0.693	**0.711**
Forward-prediction (BPS)	0.256	**0.275**	0.309	**0.334**
Inter-region (BPS)	0.659	**0.689**	0.604	**0.640**
Intra-region (BPS)	0.644	**0.663**	0.688	**0.724**
Choice decoding (accuracy)	0.839	**0.841**	0.781	**0.823**

Metrics The bits-per-spike metric (BPS) is widely used to evaluate neural activity prediction performance [21, 22, 14, 15]:

$$\text{bits-per-spike}\left(\hat{\lambda}, \mathbf{X}\right) = \frac{1}{n_{sp}\log 2}\left(\mathcal{L}\left(\mathbf{X}; \hat{\lambda}\right) - \mathcal{L}\left(\mathbf{X}; \bar{\lambda}\right)\right), \tag{1}$$

where $\hat{\lambda}$ is the predicted spike rates by the model, $\mathbf{X}$ is the true spike counts, n_{sp} is the total spike count of $\mathbf{X}$, $\mathcal{L}$ is the log likelihood function of Poisson, and $\bar{\lambda}$ is the mean firing rate of $\mathbf{X}$. The BPS metric essentially evaluates the goodness-of-fit statistics of a model over the null model, normalized by the spike counts. Changes in BPS directly reflect changes in model log likelihood $\mathcal{L}\left(\mathbf{X}; \hat{\lambda}\right)$ when evaluated on the same dataset, as other terms remain constant. For all experiments, we report the metrics on the test sets of the heldout sessions after finetuning models to their training sets.

Tasks Each masking scheme used for training corresponds to a leave-one-out evaluation task of activity prediction, namely: (1) Co-smoothing: Activities of each neuron were predicted from all other neurons; (2) Forward-prediction: The model predicted neural activities in continuous, non-overlapping windows of 200 ms at a time (10 timesteps) given previous ground truth activities, starting from stimulus onset to 2.2 seconds after stimulus onset (110 timesteps in total); (3) Inter-region: Activities of all neurons in each region were predicted from all other regions; (4) Intra-region: Activities of each neuron in a region were predicted from all other neurons in that region. Repeat for each region.

3 Data heterogeneity limits NDT models' scaling behavior

As mentioned, the difference in brain region overlaps between RS and BWM provides a natural setting to study how pretraining data heterogeneity affects model scaling behavior. BWM sessions contain neural activity from largely non-overlapping brain regions, resulting in greater single-neuron and brain-region heterogeneity compared to RS. We hypothesize that increased heterogeneity in BWM will reduce the scaling benefits from pretraining on more sessions.

We conducted our scaling analysis as follows. Using the RS dataset, we pretrained NDT models on 20, 40, 60, and 84 sessions, then finetuned them to each of five heldout sessions. For BWM, we pretrained models on 20, 40, 60, 80, 100, 150 and 200 sessions (out of 460 total), then finetuned them on ten heldout sessions. During finetuning, a new session embedding and two session-specific stitchers were learned from scratch while the mask embedding and the core NDT parameters were initialized from pretraining.

3.1 Pretraining offers performance improvements across tasks over baseline models

Before analyzing whether pretrained models' performances scale with more pretraining data, we first confirm they indeed outperform baseline models. Table 1 presents the evaluation metrics achieved by the baseline models and the best pretrained-then-finetuned models on RS and BWM, averaged over all heldout sessions. Pretraining improves performance on both datasets, though gains are smaller for co-smoothing and choice decoding tasks than others. Notably, co-smoothing metrics are higher than intra-region in RS but not in BWM, suggesting that models trained on RS data benefited more from cross-region information. This is consistent with the distinct recording strategies of the two datasets.

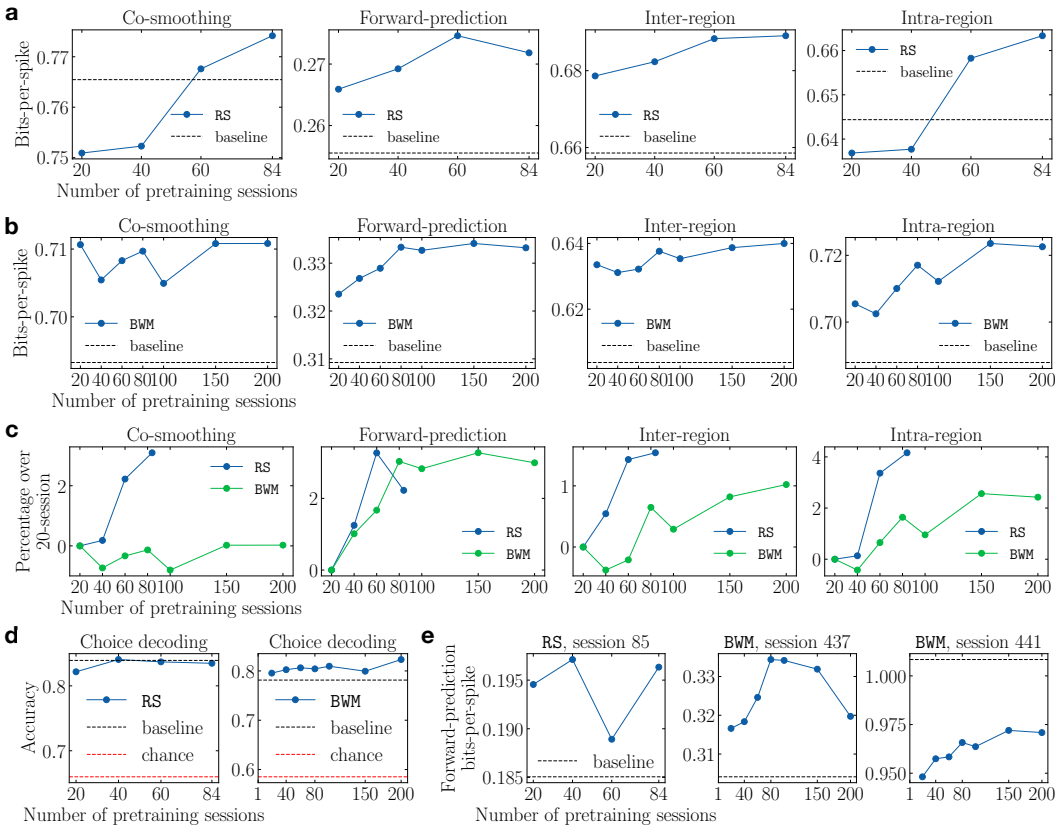

Figure 2: **Scaling Behavior of NDT Models.** Plots show pretrained NDT models' finetuning performances as the number of pretraining sessions increases. **(a)** Performances of each neural activity prediction task on RS data. Black dashed lines show the baseline models' performances. **(b)** Same as (a) but on BWM data. **(c)** Percentage improvements of models pretrained with more sessions over the 20-session model. **(d)** Choice decoding performances on RS (left) and BWM (right). Red dashed line shows the chance prediction accuracy. **(e)** Forward-prediction performance examples on individual heldout sessions from RS (first panel) and BWM (second & third panels).

3.2 Performance scaling is weaker in BWM than RS

Next, we investigate whether the task performances scale with increased pretraining data. Figure 2(a) and (b) show the evaluation results of models trained on RS and BWM data, respectively. Although performance generally improved with more pretraining data, the scaling effects were relatively modest. Figure 2(c) illustrates the percentage performance gains relative to the 20-session model, with the largest improvement of 4.2% observed on the intra-region task on RS using more than four times the amount of data. Notably, performance gains reduced across the co-smoothing, inter-region, and intra-region tasks on BWM compared to RS. For co-smoothing in particular, scaling benefits were negligible in BWM. In contrast, the forward-prediction performance scaled consistently on the two datasets. However, the forward-prediction performance also plateaued around 80 pretraining sessions, suggesting a potential upper limit on the performance improvements achievable through pretraining given the current scope of data. These results support our hypothesis that greater brain region heterogeneity in BWM limits the effectiveness of pretraining, particularly on single-neuron and region-level tasks.

To probe the quality of the NDT's internal representations, we trained a logistic regression classifier on the output of the third intermediate transformer block (out of five, see Appendix A) to predict the animals' decisions. Previous work typically performed such decoding analyses using reconstructed spikes rather than models' internal representations [22, 14]. Under the latter setup, improvements in choice decoding performance could reflect better spike reconstruction rather than better represen-

tations of choice-related latent states underlying the data. With our setup, decoding accuracies are higher in RS than BWM, as shown in Table 1. Figure 2(d) shows the changes in classification accuracy with increasing amounts of pretraining data. There was no clear scaling of decoding performance on either dataset. Taken together, these results highlight the importance of investigating the scaling behaviors of pretrained models with more fine-grained data increments.

We also observed large cross-session variabilities in the finetuning performances. Fig. 2(e) shows the forward-prediction performance of three heldout sessions (see Appendix B for all sessions). In addition to the substantial differences in absolute bits-per-spike values, their scaling trends deviate from the session averages (second panel in Fig. 2(a) & (b)). Such variability indicates a more implicit form of data heterogeneity that comes from individual differences among animals. Given this observation and our previous findings on the pretrained models' limited scaling behavior, a natural question arises: **Can we identify more beneficial sessions than others in the pretraining dataset for improving scaling performance?** We answer this question in the next section.

4 Identifying more beneficial single sessions for performance scaling

We hypothesize that each pretraining session exhibits varying degrees of distribution shift relative to a heldout session. We call this implicit data heterogeneity, which arises from subtle individual differences among animals and sessions that are harder to identify than explicit sources of heterogeneity such as task design and brain regions. We expect that models pretrained on sessions "closer" to the heldout sessions will achieve higher performances more data-efficiently than models pretrained with randomly selected sessions.

To test this, we conducted our experiments on the RS dataset, which allows us to control the brain region heterogeneity as discussed in the last section. We trained NDT models with only forward-prediction masking to focus on this evaluation task, which exhibits the most consistent scaling behavior on the two datasets (Fig. 2(c)). For more fine-grained scaling analysis, we pretrained models on 1, 2, 3, 4, 5, 10, 20, 30, 40, and all 84 sessions.

4.1 Ranking pretraining sessions by single-session finetuning performances

First, we propose using single-session finetuning performances as an estimate of the "closeness" between the data distributions of a pretraining session and a heldout session. Figure 3 illustrates this process: during the pretraining stage, we trained 84 single-session models, one for each pretraining session (Fig. 3(a)). During the finetuning stage, for a particular heldout session, we trained two new stitchers (for encoding and decoding) for each of the pretrained transformers while keeping the transformers' weights frozen (Fig. 3(b)). This ensures the finetuning performance maximally depends on the features learned from the pretraining session, as the only adjustable weights were the input/output linear layers that map the raw spike counts to the frozen feature space and back. Lastly, we report each model's forward-prediction performance on the heldout session's validation set, yielding 84 metric values – one per pretraining session. The pretraining stage (Fig. 3(a)) was performed once, while the finetuning and ranking stage (Fig. 3(b)) were repeated for each heldout session. See Appendix C for the ranked single-session finetuning performances.

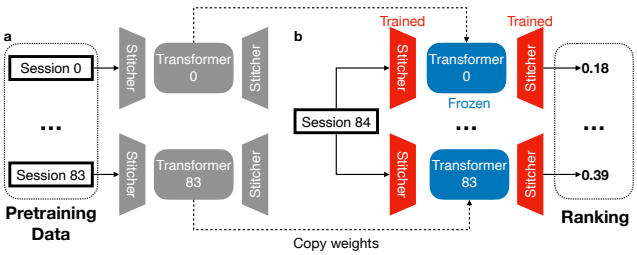

Figure 3: **Schematic of the ranking process. (a)** Pretraining stage: We trained 84 single-session models, each consisting of a transformer and two session-specific stitchers. **(b)** Finetuning stage: For each pretrained model, we trained two new stitchers on the heldout session's training set, keeping the transformer weights frozen. Models were ranked by their bits-per-spike metric on the heldout session's validation set.

We conducted our data scaling experiments by incrementally selecting more pretraining sessions in three session-selection orders: random, ranked (based on the procedure above), and reverse-ranked. To reduce variance, we used three random seeds for both ranking sessions and training models in the

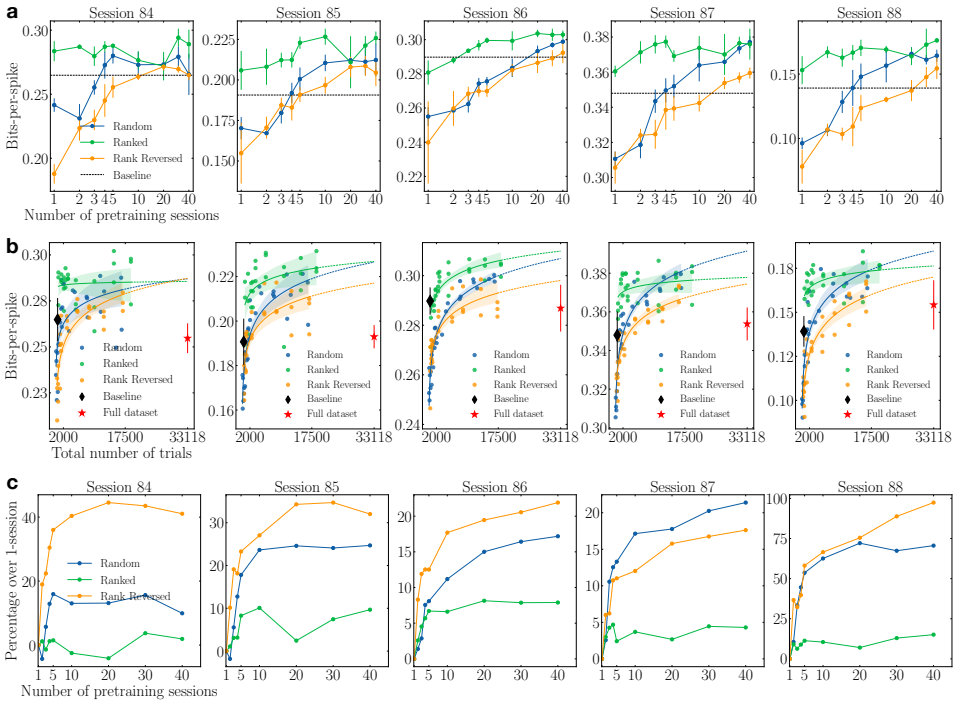

Figure 4: **Scaling performances under different session orders. (a)** Forward-prediction performances of each heldout session as we increased pretraining sessions according to random (blue), ranked (green), or reverse-ranked (orange) order. The error bars show the standard deviation over three seeds (ranked/reverse-ranked) or three shuffled orders (random). Black dashed lines show the baseline models' performance (averaged over three seeds). **(b)** Same as (a) but with the total number of trials as the x-axis. Linear regressions were fitted with logarithmic x values and dashed lines show extrapolated predictions. Shading shows the standard deviations. Red stars show the performances of the models pretrained with all pretraining sessions. **(c)** Percentage improvements of models pretrained with more sessions over the 1-session model.

scaling analysis, including the baseline models that were directly trained on the heldout sessions. For the random order, we used three different shuffles of the session list. The transformers' weights were frozen for all finetuning experiments across data orders to be consistent with the ranking procedure. We limited experiments to a maximum of 40 pretraining sessions (except for the full 84-session case) since more selected sessions overlap as we exhaust the pretraining data.

4.2 Pretraining on five top-ranked sessions outperforms all random sessions

Figure 4(a) shows the performances of our scaling analysis on each heldout session's test set with different session orders (see Appendix D for qualitative examples). The results clearly show that in all heldout sessions, models pretrained with ranked session order outperform those trained with randomly chosen sessions. Importantly, the models pretrained with reverse-ranked sessions achieved worse performances than random-order models, proving the validity of our ranking procedure based on single-session finetuning. Notably, the performance differences in ranked, random, and reverse-ranked settings are more pronounced in low-data regimes (fewer than ten sessions). Since the number of trials was different among sessions, we also plotted the model performances in Figure 4(a) against the total number of trials from the pretraining sessions for fairer comparisons. As shown in Figure 4(b), the same performance differences hold among the different session-selection orders given the same number of trials. We fit the models' performances using linear regression (with logarithmic input). In contrast to the success of "scaling laws" in machine learning [23, 24, 25], the actual pretraining performance using the entire 84 `RS` sessions (Fig. 4(b) red stars) is consistently lower than the extrapolated performance (Fig. 4(b) dashed lines), indicating limited scaling effects for the neural IBL data with the NDT model. This further supports our hypothesis that differences in pretraining and finetuning data distributions greatly affect the promises of neural data scaling.

Table 2: Percentage improvements over baseline with different session selection procedures.

	Session selection order (best of all experiments)			
Heldout session	Random	Ranked	Reverse-ranked	Ranked (top 5 sessions)
84	5.74%	**11.08%**	2.54%	8.69%
85	11.33%	**18.87%**	9.39%	16.92%
86	3.13%	**4.78%**	0.89%	3.39%
87	8.37%	**8.43%**	3.30%	8.43%
88	19.12%	**26.60%**	10.95%	22.44%
Average	9.54%	**13.95%**	5.42%	11.97%

Table 2 summarizes the best percentage improvements over the baseline models for each session selection order, along with the performance of models pretrained on five top-ranked sessions. On average, models using rank-ordered session data achieved a 4% greater improvement over the baseline than models using random-ordered session data. Remarkably, models trained on just five ranked sessions outperformed the best models trained on randomly selected sessions, indicating an over 8× gain in data efficiency (compared to 40 random session models, which outperformed the models trained on all sessions (Fig. 4(b)). However, this also implies a reduced scaling effect compared to randomly selected sessions. Figure 4(c) demonstrates that the percentage performance gains using more pretraining data relative to using one pretraining session under each session selection order. The scaling effects when using the ranked sessions were clearly weaker than when using random or reverse-ranked sessions. Indeed, models with five ranked sessions already achieved 86% of the best model performances with all 40 ranked sessions (Table 2), suggesting that most of the pretraining benefit is concentrated in the top few sessions.

We also examined how session-specific the rankings were. Figure 5 shows the number of sessions shared in all five top-k ranking of the heldout sessions. The top 20 sessions were highly session-specific: no session appeared in the top five for all held-out sessions, and only three were shared in the top 23. In contrast, rankings became increasingly similar beyond 23 sessions, with 43 sessions shared in all five top 53 rankings. These results suggest that only a small number of top-ranked sessions have a strong impact on performance, while the remaining pretraining sessions are more consistent across held-out sessions and affect model performance less significantly. Taken together, our analysis suggests that apparent scaling benefits in multi-session datasets can be highly sensitive to the specific sessions selected, due to substantial individual differences across sessions. Thus, it is extremely important for studies that claim scaling benefits to show detailed experimental results with fine-grained data increments.

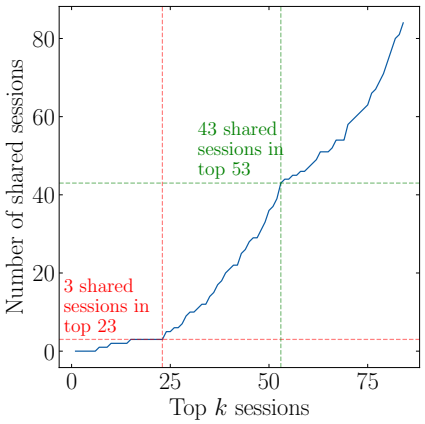

Figure 5: Number of shared sessions in the top-k ranking of all five heldout sessions.

5 Related Work

5.1 Foundation models in neuroscience

Foundation models represent a paradigm shift in artificial intelligence, allowing large-scale models pretrained on internet-scale data to be efficiently adapted to various downstream tasks through finetuning. These models demonstrate remarkable capability to learn versatile representations via self-supervised objectives and adapt effectively to downstream tasks [26, 27, 28, 29, 30, 31, 32, 33]. Motivated by these advances, the neuroscience community has begun adopting similar approaches, starting with non-invasive human neural data across modalities [34, 35, 36]. More recent work trained large attention-based models on invasive rodent and nonhuman primate data, such as POYO [12], POYO+ [13], and NDT model series [18, 37, 16]. Using calcium imaging data, Wang et al. [38]

explored combining recurrent architectures with attention modules for predicting neural activities from visual stimuli and locomotion.

More specifically, NDT models scaled attention-based models to multi-session spiking data by incorporating context embeddings and learning a shared latent space across sessions, enabling transfer to new recording conditions [37, 16]. Azabou et al. [12] extended the multi-subject pretraining paradigm to primate data and proposed single-spike tokenization through a PerceiverIO architecture [39]. Zhang et al. [14] employed multiple spike masking schemes on the IBL RS dataset, upon which we based our work. Similarly, Zhang et al. [15] proposed a novel multimodal training and masking method, demonstrating improved performance from multi-session training over single-session models.

5.2 Scaling behavior in foundation models for neuroscience

Foundation models in NLP have been empirically shown to follow scaling laws, where performance improves predictably with more data and parameters [23, 24]. In neuroscience, similar scaling effects have been explored through many pretraining studies. Azabou et al. [12] used primate data from motor and premotor cortices and demonstrated that pretraining on over 100 hours of data enables rapid adaptation to unseen sessions. Azabou et al. [13] extended their model to rodent visual cortex data and presented benefits of pretraining on over a thousand sessions. Using the IBL dataset, Zhang et al. [40] showed that reduced-rank regression models trained on hundreds of sessions across diverse brain regions outperform session-specific models, indicating benefits from multi-session data. Zhang et al. [15] similarly reported performance gains from their multi-task masking strategies when pretrained across multiple sessions. While these results suggest that pretraining on multi-session data is generally beneficial, they often lack fine-grained analyses of how performance trends evolve with incremental pretraining data. As shown by our results in Fig. 4, scaling behaviors of the model may drastically vary when pretrained on different subsets of the pretraining dataset. Therefore, it may be misleading to conclude that the model enjoys scaling benefits with just a few data increments. In fact, recent studies have begun to observe limited scaling effects in foundation models for motor decoding when finetuning data exceeds 100 minutes [16], similar to our results.

6 Discussion and Conclusion

Our results show that data heterogeneity in multi-session electrophysiology datasets fundamentally limits performance improvements expected from increasing pretraining data. Through detailed scaling analysis, we show that explicit sources of heterogeneity such as recording from different brain regions diminish performance scaling in the single-neuron activity prediction tasks and hinder performance scaling in other region-dependent tasks, such as inter- and intra-region tasks (Fig. 2). For the forward-prediction task where performances scaled consistently, we proposed a ranking procedure and used top-ranked pretraining sessions to reduce implicit heterogeneity due to session variabilities (Fig. 3). Our analysis reveals that as few as five sessions selected through ranking can outperform models pretrained with 40 randomly selected sessions (Fig. 4). This implies that simply changing session selection orders during pretraining can lead to drastically different scaling outcomes. Together, these findings suggest that previous claims about scaling benefits of pretraining without detailed incremental experiments may be premature.

Our study has several limitations. First, all experiments used the NDT architecture trained with self-supervised learning objectives. It remains to be seen whether our conclusions generalize to other architectures and objective functions (e.g., POYO with supervised loss [12, 13]). Second, different modalities of neural data recorded under different tasks and species may also affect model scaling behaviors. Concurrent work on motor decoding by Ye et al. [16] reports that the benefits from pretraining the model on 2000 hours of data are virtually nonexistent when finetuning datasets exceed 100 minutes, supporting our hypothesis that data heterogeneity issues extend beyond our dataset. Third, while our ranking procedure demonstrates the existence of a small subset of pretraining data that leads to optimal performance, it is computationally intensive. Future work can focus on identifying more general principles for pretraining data selection, particularly for experimental design where heldout sessions are not yet available.

In conclusion, our results show that pretraining neural encoding models with more sessions does not naturally lead to improved downstream performance. We strongly advocate for rigorous scaling analyses in future work on neuroscience foundation models to account for data heterogeneity effects.

Acknowledgments and Disclosure of Funding

We thank Shuchen Wu, Nick Steinmetz, Matt Golub, and Luke Zettlemoyer for discussions. This work was supported by National Science Foundation EFRI grant 2223495 (RPNR), a UW + Amazon Science Hub grant (RPNR), a Frameworks grant from the Templeton World Charity Foundation (RPNR) and the Air Force Office of Scientific Research under award number FA9550-24-1-0313 (RPNR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html`.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. ISSN 1533-7928. URL `http://jmlr.org/papers/v25/23-0870.html`.

[4] International Brain Laboratory, Valeria Aguillon-Rodriguez, Dora Angelaki, Hannah Bayer, Niccolo Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle Chapuis, Anne K Churchland, Yang Dan, Eric Dewitt, Mayo Faulkner, Hamish Forrest, Laura Haetzel, Michael Häusser, Sonja B Hofer, Fei Hu, Anup Khanal, Christopher Krasniak, Ines Laranjeira, Zachary F Mainen, Guido Meijer, Nathaniel J Miska, Thomas D Mrsic-Flogel, Masayoshi Murakami, Jean-Paul Noel, Alejandro Pan-Vazquez, Cyrille Rossant, Joshua Sanders, Karolina Socha, Rebecca Terry, Anne E Urai, Hernando Vergara, Miles Wells, Christian J Wilson, Ilana B Witten, Lauren E Wool, and Anthony M Zador. Standardized and reproducible measurement of decision-making in mice. *eLife*, 10:e63711, May 2021. ISSN 2050-084X. doi: 10.7554/eLife.63711. URL `https://doi.org/10.7554/eLife.63711`.

[5] International Brain Laboratory, Kush Banga, Julius Benson, Jai Bhagat, Dan Biderman, Daniel Birman, Niccolò Bonacchi, Sebastian A. Bruijns, Kelly Buchanan, Robert AA Campbell, Matteo Carandini, Gaëlle A. Chapuis, Anne K. Churchland, M. Felicia Davatolhagh, Hyun Dong Lee, Mayo Faulkner, Berk Gerçek, Fei Hu, Julia Huntenburg, Cole Hurwitz, Anup Khanal, Christopher Krasniak, Christopher Langfield, Petrina Lau, Nancy Mackenzie, Guido T. Meijer, Nathaniel J. Miska, Zeinab Mohammadi, Jean-Paul Noel, Liam Paninski, Alejandro Pan-Vazquez, Cyrille Rossant, Noam Roth, Michael Schartner, Karolina Socha, Nicholas A. Steinmetz, Karel Svoboda, Marsa Taheri, Anne E. Urai, Shuqi Wang, Miles Wells, Steven J. West, Matthew R. Whiteway, Olivier Winter, Ilana B. Witten, and Yizi Zhang. Reproducibility of in vivo electrophysiological measurements in mice. *eLife*, 13, March 2025. doi: 10.7554/eLife.100840.2. URL `https://elifesciences.org/reviewed-preprints/100840`.

[6] Leonhard Waschke, Niels A. Kloosterman, Jonas Obleser, and Douglas D. Garrett. Behavior needs neural variability. *Neuron*, 109(5):751–766, March 2021. ISSN 0896-6273. doi: 10.1016/j.neuron.2021.01.023. URL https://www.sciencedirect.com/science/article/pii/S0896627321000453.

[7] Kenneth D. Harris and Alexander Thiele. Cortical state and attention. *Nature Reviews Neuroscience*, 12(9):509–523, September 2011. ISSN 1471-0048. doi: 10.1038/nrn3084. URL https://www.nature.com/articles/nrn3084.

[8] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, April 2019. doi: 10.1126/science.aav7893. URL https://www.science.org/doi/full/10.1126/science.aav7893. Publisher: American Association for the Advancement of Science.

[9] Steven M. Peterson, Satpreet H. Singh, Nancy X. R. Wang, Rajesh P. N. Rao, and Bingni W. Brunton. Behavioral and Neural Variability of Naturalistic Arm Movements. *eNeuro*, 8 (3), May 2021. ISSN 2373-2822. doi: 10.1523/ENEURO.0007-21.2021. URL https://www.eneuro.org/content/8/3/ENEURO.0007-21.2021.

[10] Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147, October 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.08.005. URL https://www.sciencedirect.com/science/article/pii/S0959438819300303.

[11] Laura N. Driscoll, Lea Duncker, and Christopher D. Harvey. Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609, October 2022. ISSN 0959-4388. doi: 10.1016/j.conb.2022.102609. URL https://www.sciencedirect.com/science/article/pii/S0959438822001039.

[12] Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh B. Nachimuthu, Michael Jacob Mendelson, Blake Aaron Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A Unified, Scalable Framework for Neural Population Decoding. In *Advances in Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=sw2Y0sirtM.

[13] Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L. Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=IuU0wcOOmo.

[14] Yizi Zhang, Yanchen Wang, Donato M. Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, The International B. Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Towards a "Universal Translator" for Neural Dynamics at Single-Cell, Single-Spike Resolution. In *Advances in Neural Information Processing Systems*, volume 37, pages 80495–80521, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/934eb45b99eff8f16b5cb8e4d3cb5641-Abstract-Conference.html.

[15] Yizi Zhang, Yanchen Wang, Mehdi Azabou, Alexandre Andre, Zixuan Wang, Hanrui Lyu, The International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Neural Encoding and Decoding at Scale, April 2025. URL http://arxiv.org/abs/2504.08201.

[16] Joel Ye, Fabio Rizzoglio, Adam Smoulder, Hongwei Mao, Xuan Ma, Patrick Marino, Raeed Chowdhury, Dalton Moore, Gary Blumenthal, William Hockeimer, Nicolas G. Kunigk, J. Patrick Mayo, Aaron Batista, Steven Chase, Adam Rouse, Michael L. Boninger, Charles Greenspon, Andrew B. Schwartz, Nicholas G. Hatsopoulos, Lee E. Miller, Kristofer E. Bouchard, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. A Generalist Intracortical Motor Decoder, February 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.02.634313v1.

[17] International Brain Laboratory, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, Kcénia Bougrova, Sebastian A. Bruijns, Matteo Carandini, Joana A. Catarino,

Gaelle A. Chapuis, Anne K. Churchland, Yang Dan, Felicia Davatolhagh, Peter Dayan, Eric EJ DeWitt, Tatiana A. Engel, Michele Fabbri, Mayo Faulkner, Ila Rani Fiete, Charles Findling, Laura Freitas-Silva, Berk Gerçek, Kenneth D. Harris, Michael Häusser, Sonja B. Hofer, Fei Hu, Félix Hubert, Julia M. Huntenburg, Anup Khanal, Christopher Krasniak, Christopher Langdon, Petrina Y. P. Lau, Zachary F. Mainen, Guido T. Meijer, Nathaniel J. Miska, Thomas D. Mrsic-Flogel, Jean-Paul Noel, Kai Nylund, Alejandro Pan-Vazquez, Liam Paninski, Alexandre Pouget, Cyrille Rossant, Noam Roth, Rylan Schaeffer, Michael Schartner, Yanliang Shi, Karolina Z. Socha, Nicholas A. Steinmetz, Karel Svoboda, Anne E. Urai, Miles J. Wells, Steven Jon West, Matthew R. Whiteway, Olivier Winter, and Ilana B. Witten. A Brain-Wide Map of Neural Activity during Complex Behaviour, December 2024. URL https://www.biorxiv.org/content/10.1101/2023.07.04.547681v4.

[18] Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data analysis, and Theory*, 5(3):1–18, August 2021. doi: 10.51628/001c.27358. URL https://nbdt.scholasticahq.com/article/27358-representation-learning-for-neural-population-activity-with-neural-data-transformers.

[19] Trung Le and Eli Shlizerman. STNDT: Modeling Neural Population Activity with Spatiotemporal Transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 17926–17939, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/72163d1c3c1726f1c29157d06e9e93c1-Abstract-Conference.html.

[20] Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, October 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0109-9. URL https://www.nature.com/articles/s41592-018-0109-9. Number: 10 Publisher: Nature Publishing Group.

[21] Fred Rieke, Davd Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: exploring the neural code*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-18174-6.

[22] Felix C. Pei, Joel Ye, David M. Zoltowski, Anqi Wu, Raeed Hasan Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew Kaufman, Mark M. Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan W. Pillow, Il Memming Park, Eva L. Dyer, and Chethan 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)*, August 2021. URL https://openreview.net/forum?id=KVMS3fl4Rsv.

[23] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361.

[24] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

[25] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling Data-Constrained Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training, June 2018.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, February 2019.

[28] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations*, October 2021. URL `https://openreview.net/forum?id=gEZrGCozdqR`.

[29] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. URL `http://arxiv.org/abs/2303.08774`.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *The Ninth International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023. URL http://arxiv.org/abs/2304.02643.

[33] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent. *Transactions on Machine Learning Research*, August 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=1ikK0kHjvj.

[34] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-GPT: Towards A Foundation Model For EEG. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, May 2024. doi: 10.1109/ISBI56570.2024.10635453. URL https://ieeexplore.ieee.org/abstract/document/10635453. ISSN: 1945-8452.

[35] Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. In *Advances in neural information processing systems*, volume 35, pages 21255–21269, 2022.

[36] CTRL labs at Reality Labs, David Sussillo, and Thomas Reardon. A generic noninvasive neuromotor interface for human-computer interaction, July 2024. URL https://www.biorxiv.org/content/10.1101/2024.02.23.581779v2.

[37] Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural Data Transformer 2: Multi-context Pretraining for Neural Spiking Activity. In *Advances in Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=CBBtMnlTGq.

[38] Eric Y. Wang, Paul G. Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A. Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, Casey M. Schneider-Mizell, Nuno Maçarico da Costa, R. Clay Reid, Forrest Collman, Nuno Maçarico da Costa, Katrin Franke, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, Fabian H. Sinz, and Andreas S. Tolias. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058):470–477, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08829-y. URL https://www.nature.com/articles/s41586-025-08829-y.

[39] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *The Tenth International Conference on Learning Representations*, October 2021. URL https://openreview.net/forum?id=fILj7WpI-g.

[40] Yizi Zhang, Hanrui Lyu, Cole Hurwitz, Shuqi Wang, Charles Findling, Felix Hubert, Alexandre Pouget, International Brain Laboratory, Erdem Varol, and Liam Paninski. Exploiting correlations across trials and behavioral sessions to improve neural decoding, October 2024. URL https://www.biorxiv.org/content/10.1101/2024.09.14.613047v2.

[41] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *The Seventh International Conference on Learning Representations*, September 2018. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[42] Leslie N. Smith and Nicholay Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates, May 2018. URL http://arxiv.org/abs/1708.07120.

A Experimental setup details

In this section, we detail our experimental setup introduced in Section 2.

A.1 Model architecture

Our model follows the architecture of Zhang et al. [14]. At each time step t, a raw spike count vector $\mathbf{x}_t \in \mathbb{R}^{N_i}$ from session i (with N_i neurons recorded) is projected via a session-specific linear layer to a spike token with dimension d. Another linear layer with Softsign activation maps the spike tokens to embeddings. A session embedding and a masking scheme embedding were appended to the spike embedding sequence. Learned position embeddings are added to the input embeddings, making up the final input to the transformer block. Lastly, another session-specific linear layer maps the transformer output back to the spike count vectors of dimension N_i for session i.

A.2 Hyperparameter selection

We tuned learning rates, dropout rates, weight decay rates, and batch sizes using small-scale experiments on single- and ten-session models. Optimal values were chosen based on test losses from pretraining sessions in `RS`. Other hyperparameter values were inherited from the implementation of Zhang et al. [14]. Table 3 summarizes the hyperparameters we used for all experiments. We increased the spike token dimension to 1024 for `BWM` experiments to account for higher numbers of neurons. We used seed 42 for all experiments in Section 3, and seeds 10, 20, and 42 for experiments in Section 4 that required three seeds.

The two session-specific linear layers contain approximately 1.2 million parameters on average per session. The number of parameters in NDT (shared across sessions) is roughly 12 million with the values in Table 3, which achieved better performance than smaller 8 million models and larger 24/38 million models on `RS`. We did not run larger models on `BWM` due to computing resource constraints.

A.3 Compute

All models were trained on a single Nvidia A40 or L40 GPU. Single-session training and finetuning take about one hour to train on average. The full 84-session model on `RS` takes about 3.5 days, while the full 200-session model on `BWM` takes about a week to train. Finetuning jobs in Section 4 are significantly faster since we only train the two linear stitchers, with each taking about 20 minutes to finish on average.

B Cross-session variabilities in scaling behaviors

Here, we show detailed experimental results of the scaling analysis in Section 3. Figure 6 and Figure 7 show the finetuning performances on each heldout session in `RS` and `BWM`, respectively. As discussed in Section 3, there exist large cross-session variabilities in the finetuning performances among heldout sessions. These results suggest differential contributions of different sessions in pretraining, which we investigated in Section 4.

C Finetuning results from the session-selection procedure

Here, we show the results of the session-selection procedure described in Section 4.1. Figure 8 shows the single-session finetuning performances on the validation set of each heldout session, sorted from high to low. As the figure shows, there exist large differences between the best and worst single-session finetuning performances, which are more noticeable in Sessions 84, 85, and 88 than in others.

D Qualitative comparison on forward-prediction performances

Here, we qualitatively compare the forward-prediction results on some example neurons from Session 84. As shown in Figure 9, predictions made by the models trained on five top-ranked sessions (green)

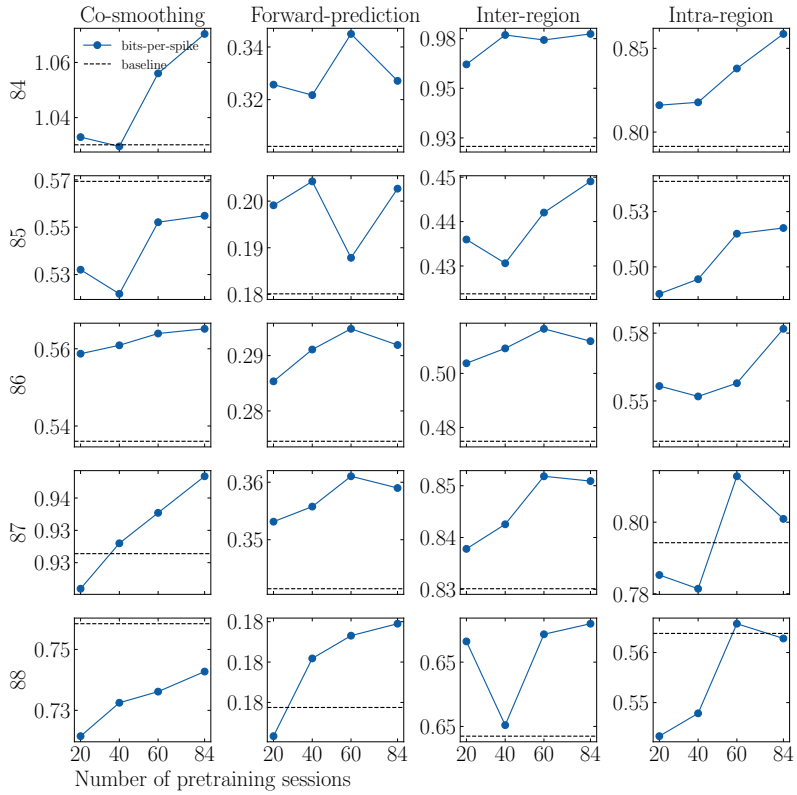

Figure 6: **Pretrained models' finetuning performances on each heldout session in** `RS`. Black dashed lines show the baseline models' performances.

match the ground truth activity dynamics (blue) much better than those trained by five reverse-ranked sessions (orange). Neurons were randomly selected from the 100 most active ones.

Table 3: Hyperparameter values across experiments

Hyperparameter	Value
Model	
Spike token dim	668
Embedding dim	512
Feedforward dim	1024
# attention heads	8
# transformer blocks	5
Dropout rate	0.2
Training	
Optimizer	AdamW [41]
Learning rate	1e-4
Learing rate scheduler	OneCycle [42]
Weight decay	0.01
Batch size	16
Gradient clipping	1
# Epochs	1000
Masking ratio	0.3

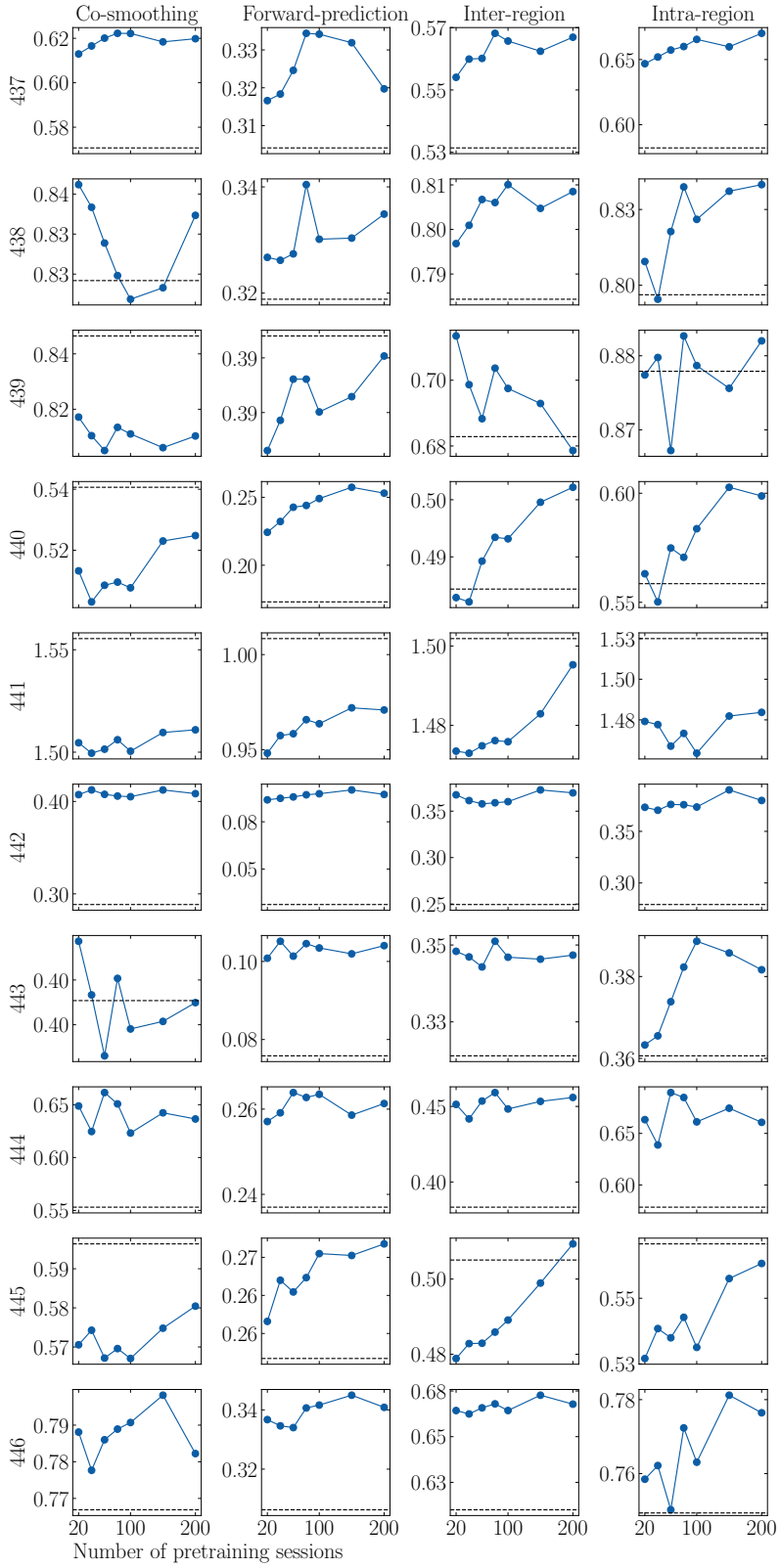

Figure 7: **Pretrained models' finetuning performances on each heldout session in** `BWM`. Black dashed lines show the baseline models' performances.

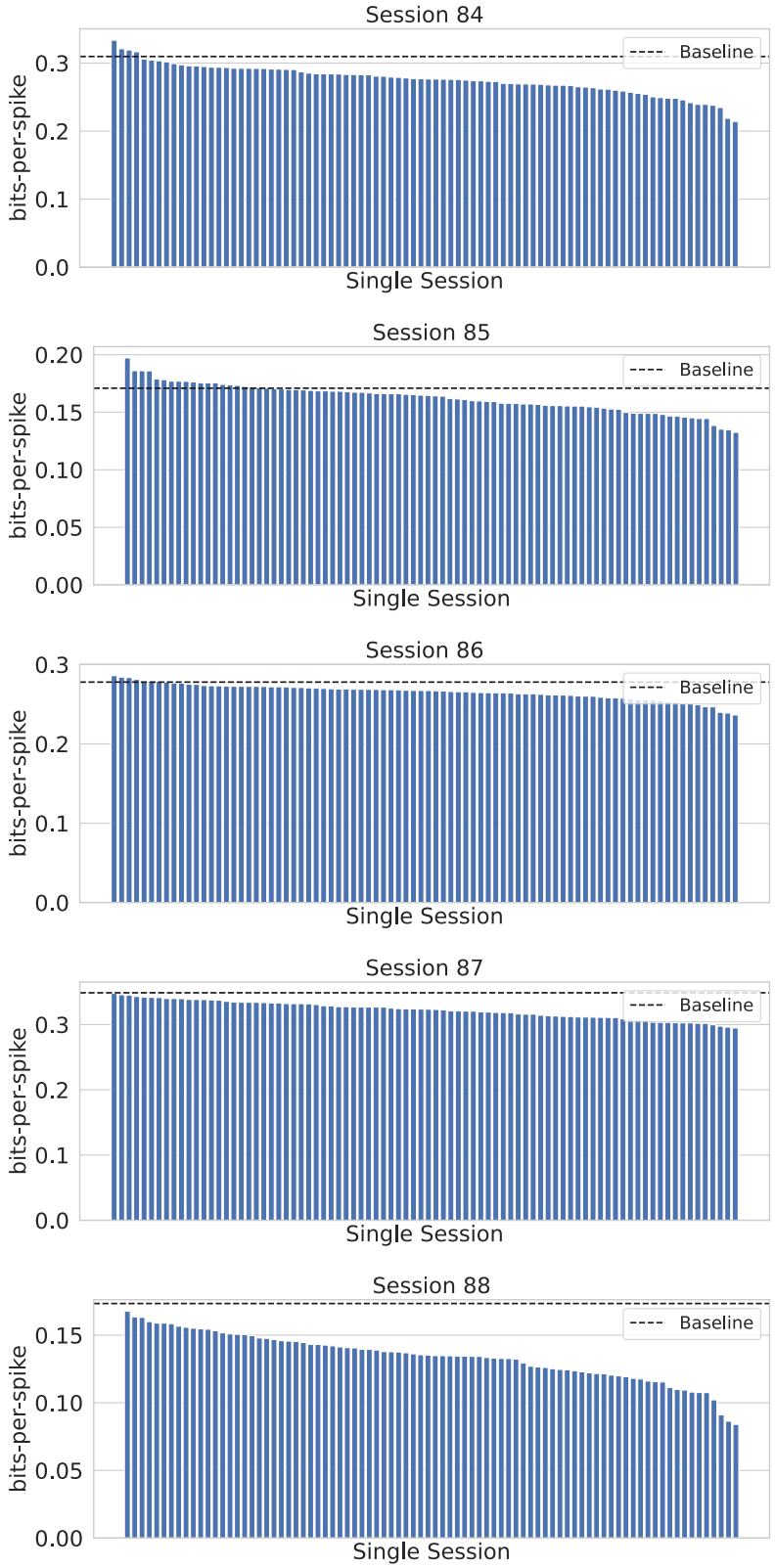

Figure 8: **Single-session finetuning performances on the validation set of the heldout sessions from the session-selection procedure.** Black dashed lines show the baseline models' performances.

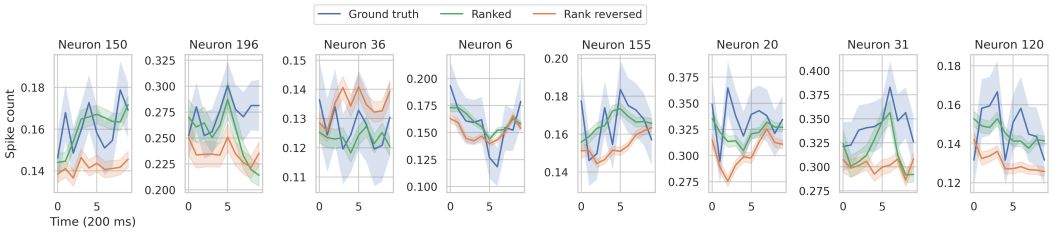

Figure 9: **Qualitative comparison of forward-prediction performances between ranked and reverse-ranked models.** Ground truth activities and predictions are averaged over trials. Shading shows standard error of the mean.