Not All Sessions Are Equal: Data Selection for Multi-Session Pretraining in 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
*prestonj@cs.washington.edu

Abstract

A key challenge in analyzing neuroscience datasets is the profound variability they exhibit across sessions, animals, and data modalities. 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 whether all sessions contribute equally to downstream performance gains. In this work, we systematically investigate how cross-session variability impacts the scaling behavior of neural data transformers (NDTs) in neural activity prediction. We propose a session selection procedure based on single-session finetuning performances. Through this 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 multi-session scaling benefits may need to be re-examined in the light of session-to-session variability. 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 variability and learn more generalizable representations with more pretraining data?

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

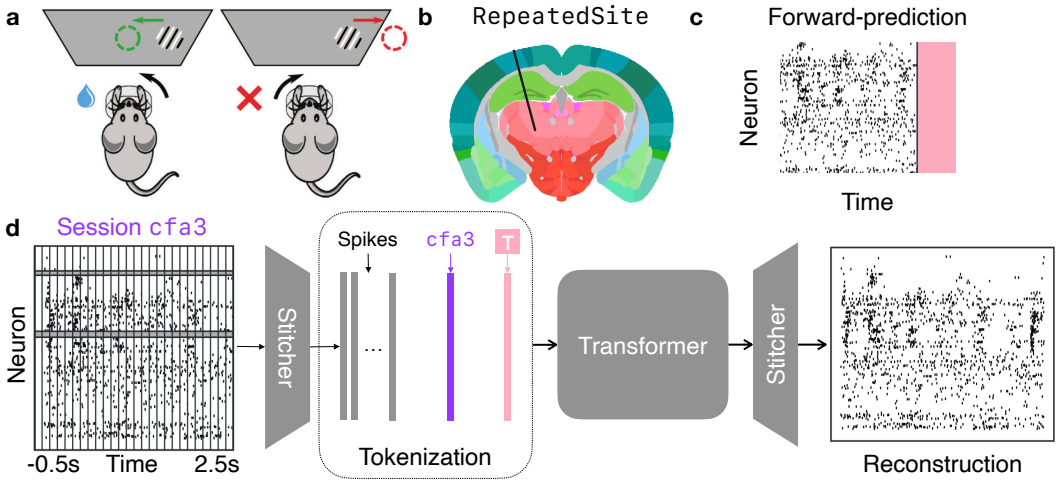

Figure 1: **Experimental Setup. (a)** Schematic of the visual decision-making task performed by mice. **(b)** Planned probe insertion location (black line) for all sessions in the `RepeatedSite` dataset. **(c)** Forward-prediction training task. **(d)** The model architecture. Sub-figures adapted from [4, 5, 18, 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 data scaling impacts downstream performances. 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 cross-session variability affects the scaling behavior of neural data transformers (NDTs) [17, 14, 16] using the `RepeatedSite` dataset released from the International Brain Laboratory [5, 18]. Through a proposed session-selection procedure based on single-session finetuning performances, we identified the impact of each pretraining session on downstream performance improvements. We found that 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. These findings point to the need for rigorous scaling analyses in future work on neural foundation models to accurately assess the effect of data scaling and 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.

Dataset We used the multi-brain-region, multi-animal/session `RepeatedSite` (henceforth `RS`) dataset 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)). Each session attempted to record from the *same* brain regions (Fig. 1(b), black line shows planned electrode insertion position). 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. 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.

Model We used the neural data transformer (NDT) architecture by Ye and Pandarinath [17] that has been widely applied to neural encoding and decoding tasks [19, 16, 15]. During training, the model is trained to predict future activities of all neurons from previous activities with causal attention masks (Fig. 1(c)). 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(d) left) whose dimensions are shared across sessions [20]. A session embedding and a masking scheme embedding [14] are also appended to input sequences [1]. Lastly, another session-specific linear layer (decoding stitcher) maps the output of the transformer back to reconstructed spike rates (Fig. 1(d) right).

Evaluation 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. The models' neural activity prediction performances are evaluated with the widely used bits-per-spike (BPS) metric:[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.

3 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, which arises from subtle, implicit individual differences among animals and sessions. 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.

3.1 Ranking pretraining sessions by single-session finetuning performances

To test this hypothesis, we first 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 2 illustrates this process: during the pretraining stage, we trained 84 single-session models, one for each pretraining session (Fig. 2(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. 2(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

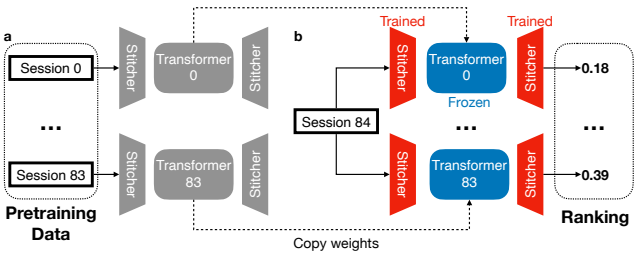

Figure 2: **Schematic of the ranking process. (a)** Pretraining stage: We trained 84 single-session models, each consisting of a transformer and two session-specific stitchers. **(b)** Finetuning stage: For each pretrained model, we trained two new stitchers on the heldout session's training set, keeping the transformer weights frozen. Models were ranked by their bits-per-spike metric on the heldout session's validation set.

[1] This setup was inherited from a training pipeline that used other masking schemes as in [14]. Since our work here only uses the forward-prediction task, this masking scheme embedding is not strictly necessary.

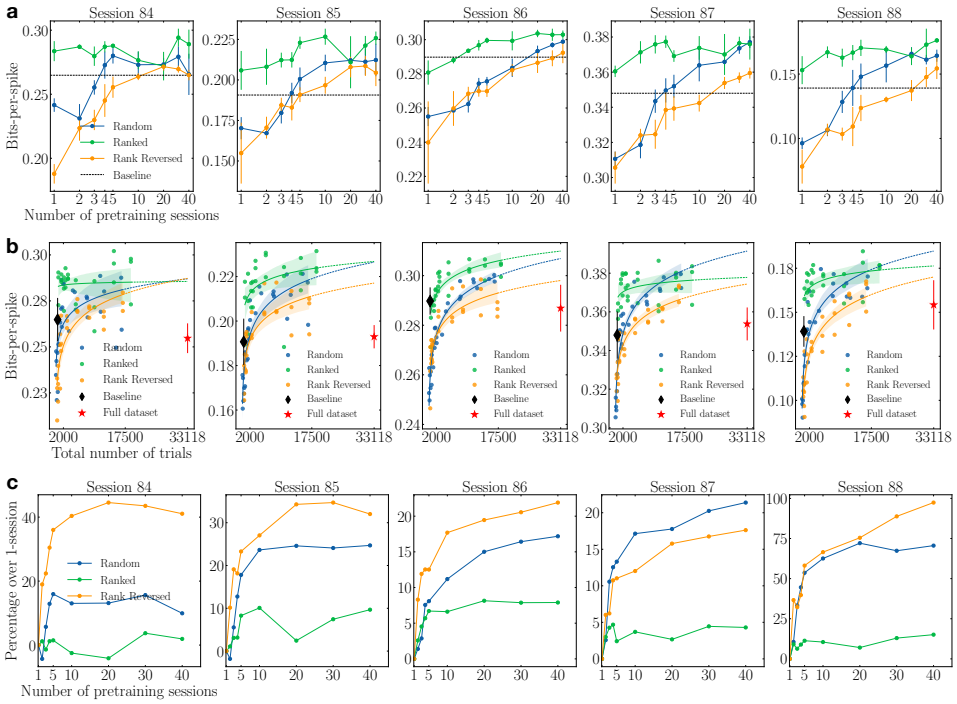

Figure 3: **Scaling performances under different session orders. (a)** Forward-prediction performances of each heldout session as we increased pretraining sessions according to random (blue), ranked (green), or reverse-ranked (orange) order. The error bars show the standard deviation over three seeds (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.

performance on the heldout session's validation set, yielding 84 metric values – one per pretraining session. The pretraining stage (Fig. 2(a)) was performed once, while the finetuning and ranking stage (Fig. 2(b)) were repeated for each heldout session. See Appendix B and Fig. A1 for the ranked single-session finetuning performances.

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 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.

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

Figure 3(a) shows the performances of our scaling analysis on each heldout session's test set with different session orders. 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, see statistical tests in Table A2). Since the number of trials was different among sessions, we also plotted the model performances in Figure 3(a) against the

Table 1: 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%

total number of trials from the pretraining sessions for fairer comparisons. As shown in Figure 3(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. 3(b) red stars) is consistently lower than the extrapolated performance (Fig. 3(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 1 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. 3(b)). However, this also implies a reduced scaling effect compared to randomly selected sessions. Figure 3(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 1), suggesting that most of the pretraining benefit is concentrated in the top few sessions.

3.3 Rankings are session-specific and cannot be predicted from session metadata

Since the ranking procedure was performance-based, it is inherently model-dependent. Are there model-independent measures that could be used to recover these rankings for more efficient session selection? To answer this question, we attempted to use simple heuristics (i.e., metadata about the session) to predict the rankings we computed through single-session finetuning. Out of the different types of session metadata [5], we identified three types that correlate with the performance-based ranking: number of trials, number of neurons, and number of "good quality" neurons ($p < 0.05$). However, through regression analyses, these types of metadata can only explain 24% of the variance in ranking or finetuning bits-per-spike performances. This further shows that the intrinsic variability between sessions contributes to the model's scaling behavior, and such variability cannot be solely determined by simple heuristics.

We also examined how session-specific the rankings were. Figure A2 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.

4 Conclusion

In conclusion, 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 and variability across sessions. Thus, it is extremely important for studies that claim scaling benefits to show detailed experimental results with fine-grained data increments.

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. We gratefully acknowledge InVirtualis for their support and for providing the computational resources to this research.

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=IuU0wcO0mo.

[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] 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.

[18] 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.

[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, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*, October 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

[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] 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.

[27] 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 A1 summarizes the hyperparameters we used for all experiments. We used seeds 10, 20, and 42 for experiments in Section 3 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 A1, which achieved better performance than smaller 8 million models and larger 24/38 million models on RS.

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. Finetuning jobs in Section 3 are significantly faster since we only train the two linear stitchers, with each taking about 20 minutes to finish on average.

B Finetuning results from the session-selection procedure

Here, we show the results of the session-selection procedure described in Section 3.1. Figure A1 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.

Table A2: p values from t-test on Fig. 3

Session 84									
Number of Sessions	1	2	3	4	5	10	20	30	40
Random vs. Ranked	0.003	0.002	0.022	0.148	0.146	0.707	0.887	0.118	0.166
Random vs. Rank Reversed	0.001	0.509	0.021	0.078	0.028	0.281	0.709	0.181	0.967
Ranked vs. Rank Reversed	0.000	0.001	0.003	0.012	0.005	0.072	0.965	0.010	0.048
Session 85									
Number of Sessions	1	2	3	4	5	10	20	30	40
Random vs. Ranked	0.022	0.007	0.004	0.022	0.017	0.032	0.922	0.235	0.151
Random vs. Rank Reversed	0.337	0.537	0.446	0.279	0.182	0.055	0.412	0.731	0.440
Ranked vs. Rank Reversed	0.032	0.015	0.004	0.011	0.002	0.004	0.819	0.105	0.030
Session 86									
Number of Sessions	1	2	3	4	5	10	20	30	40
Random vs. Ranked	0.016	0.000	0.001	0.002	0.000	0.045	0.019	0.094	0.096
Random vs. Rank Reversed	0.433	0.884	0.251	0.247	0.124	0.789	0.228	0.087	0.214
Ranked vs. Rank Reversed	0.082	0.018	0.001	0.001	0.000	0.021	0.019	0.029	0.086
Session 87									
Number of Sessions	1	2	3	4	5	10	20	30	40
Random vs. Ranked	0.000	0.002	0.006	0.007	0.029	0.284	0.640	0.435	0.872
Random vs. Rank Reversed	0.517	0.421	0.066	0.298	0.135	0.042	0.012	0.005	0.006
Ranked vs. Rank Reversed	0.001	0.001	0.002	0.009	0.005	0.006	0.132	0.004	0.067
Session 88									
Number of Sessions	1	2	3	4	5	10	20	30	40
Random vs. Ranked	0.002	0.000	0.012	0.063	0.058	0.253	0.769	0.271	0.024
Random vs. Rank Reversed	0.141	0.937	0.024	0.109	0.076	0.048	0.014	0.092	0.214
Ranked vs. Rank Reversed	0.003	0.000	0.001	0.008	0.005	0.000	0.024	0.072	0.020

Table A1: 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 [26]
Learning rate	1e-4
Learing rate scheduler	OneCycle [27]
Weight decay	0.01
Batch size	16
Gradient clipping	1
# Epochs	1000
Masking ratio	0.3

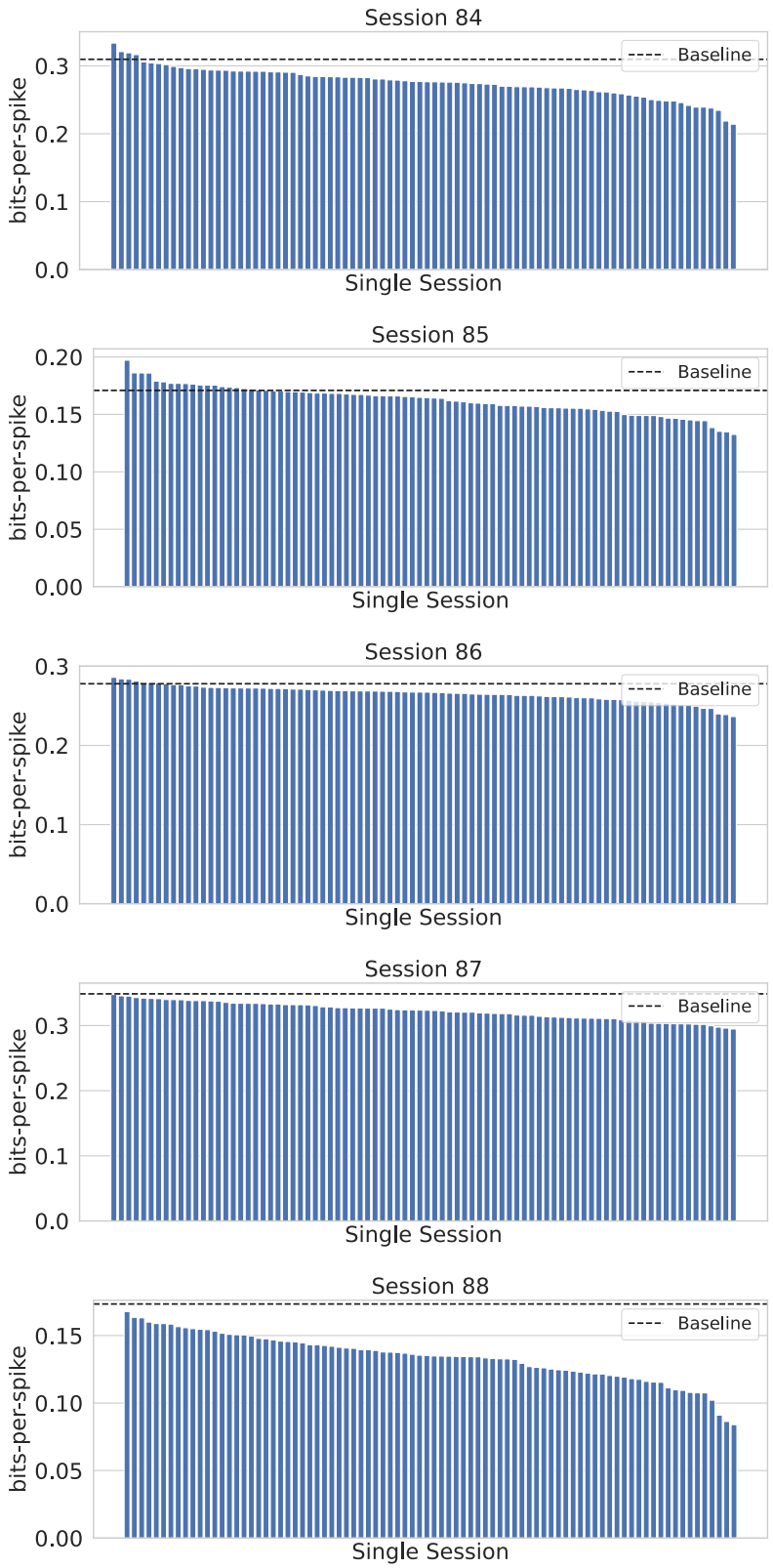

Figure A1: **Single-session finetuning performances on the validation set of the heldout sessions from the session-selection procedure.** Black dashed lines show the baseline models' performances.

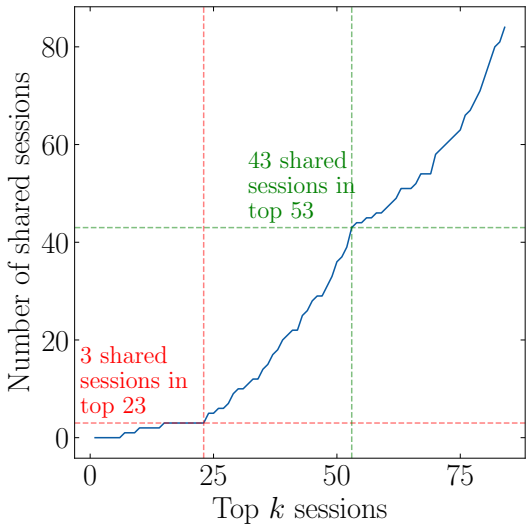

Figure A2: Number of shared sessions in the top-k ranking of all five heldout sessions.