2024–2025 ALCC Project Closeout Report

Date Submitted: 07-07-2025

A. General Information

Project Name:	Foundation Neuroscience AI Model-NeuroX
Principal Investigator (PI):	Shinjae Yoo
PI Institution:	Brookhaven National Laboratory

Utilization ***(OLCF will complete this table)***

Allocation Year	Total Allocation	Frontier Final Usage	Andes Final Usage
2024–2025	152,000	296002	—

B. Project Results

Executive Summary

This project on Frontier successfully achieved its goal of establishing a robust methodology for training multi-billion parameter foundation models for neuroscience. We progressed from overcoming initial GPU memory challenges to validating advanced scaling techniques and demonstrating readiness for models up to 14.5B parameters. While scaling was a technical success, critical scientific insights are emerging from ongoing downstream task evaluations. Early results indicate that pre-trained 51M parameter models consistently outperform models trained from scratch on challenging tasks. The campaign culminated in the successful pre-training of 3B-parameter models. Our ongoing evaluations of these larger models are yielding a crucial scientific insight: performance does not scale trivially with model size in this complex domain. Characterizing this non-linear relationship and understanding the interplay between model capacity, data scale, and fine-tuning strategy is now a primary focus of our analysis and directly informs our future scaling strategy. This work not only proved the viability of our approach but also served as a crucial de-risking campaign for a recent INCITE proposal, providing the empirical foundation and critical questions needed to target next-generation models effectively.

Summarize the significant scientific and/or engineering accomplishments achieved through this ALCC award.

Our accomplishments span both the development of a highly efficient, scalable training framework and the significant advancement of the SwiFT v2 model's scientific capabilities.

- **Established an Efficient, Scalable Training Recipe:** We successfully completed our primary objective on Frontier by creating an optimal methodology for large-scale neuroscience models. We systematically tested various distributed training setting like a memory-constrained approach (DeepSpeed ZeRO-3 with CPU Offload) and a highly performant recipe using DeepSpeed Stage 1 with BF16. This investigation dramatically improved throughput and established an optimal recipe for multi-billion parameter models, further enhanced by validating advanced techniques like μTransfer for efficient hyperparameter tuning.

- **Engineered a Temporally-Aware Model Architecture:** To enhance the model's capacity to learn the intricate temporal dynamics of fMRI data, we implemented two key architectural innovations. We replaced standard positional embeddings with Rotary Positional Embedding (RoPE) to better represent relative temporal relationships, and we decoupled the Swin Transformer block to process spatial and temporal features sequentially (**Figure 1**). These enhancements were highly effective, enabling the model to faithfully capture the temporal context of the fMRI signal, a success validated by autocorrelation analysis of its embeddings.

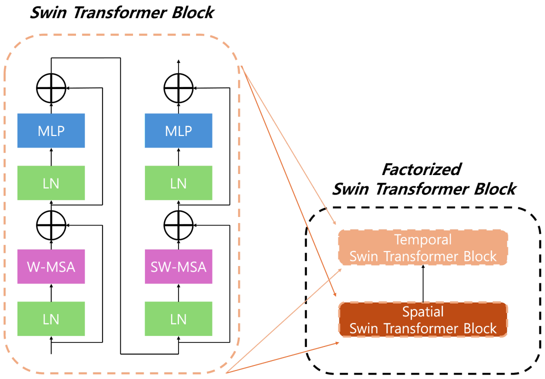


Figure 1. Schematic of the factorized Swin Transformer block. This block sequentially applies shifted window attention to spatial and then temporal dimensions, thereby decoupling the processing of spatio-temporal information.

- **Demonstrated Extreme-Scale Readiness on Frontier:** We validated our core scaling strategy through a series of landmark achievements. A key success was the efficient training of our 14.5B-parameter models on 64 nodes, proving the viability of our approach at a significant scale (Figure 2).

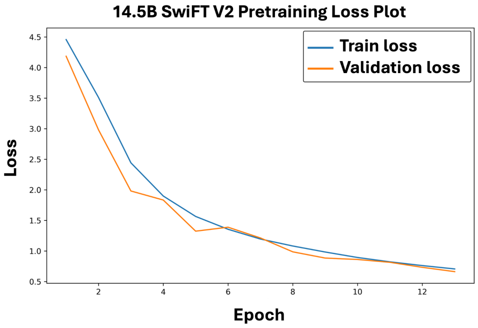


Figure 2. Pretraining loss plot of 14.5B SwiFT V2 model.

- Furthermore, we established a critical scaling performance baseline on Frontier by conducting strong scaling tests up to 2,048 nodes (Figure 3). This benchmark data was instrumental, serving as a direct comparison against other leadership-class systems like Aurora. These cross-system benchmarks not only demonstrated Frontier's robust performance but also allowed us to quantify the unique architectural strengths of different LCF systems for our specific workload.

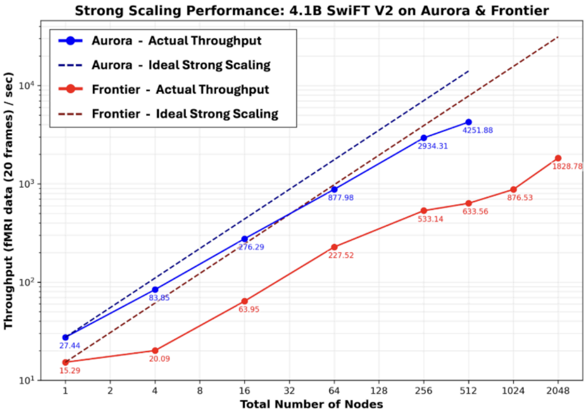


Figure 3. Strong scaling performance of the 4.1B SwifT V2 model on Aurora and Frontier. We set the throughput of an experiment using one node as an ideal strong scaling baseline.

- **Peer-Reviewed Validation of Methodology:** Key methods and initial findings from this project, including the successful scaling of the SwiFT V2 model and the validation of neural scaling laws for fMRI, were compiled and accepted as an abstract for the 2025 Cognitive Computational Neuroscience (CCN) conference (Figure 4). This acceptance signifies formal peer-reviewed validation of our scientific approach and the significance of our initial results.

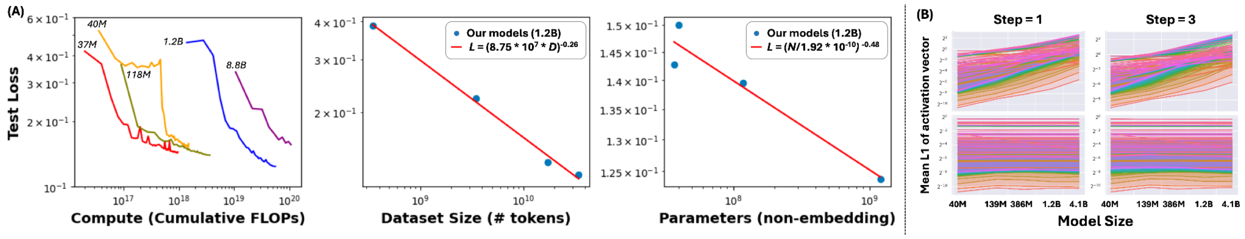


Figure 4. (A) Neural scaling laws of SwiFT V2 and (B) training stability with µTransfer. (A) SwiFT V2, fMRI swin transformer model's pretraining loss improves with more compute, data, and parameters. (B) Training stability, as verified by convergence of the norm of the activation vector, is accomplished by implementing µTransfer (top row - without µTransfer, bottom row - with µTransfer).

- **Uncovered Complex Interplay of Model Configuration, Model Scale, and Downstream Performance:** Our downstream evaluations revealed a complex interplay between model configuration, scale, and the ultimate benefits of pre-training. We systematically investigated this relationship across numerous configurations and two distinct model scales.
- At the 51M-parameter scale, optimal configuration was key to unlocking performance gains. Through extensive testing of parameters like mask ratios, patch sizes, and sequence lengths, we identified configurations where pre-training delivered significant and consistent improvements. These optimized models robustly outperformed their from-scratch counterparts, not only on in-distribution tasks (UKB) but also on challenging external validation datasets (HBN, EMBARC, ADNI), proving the generalizability of the pre-training benefits when model scale and configuration are properly aligned (**Table 1**).
- However, this relationship did not hold at the 3B-parameter scale. In stark contrast, the larger model consistently underperformed its from-scratch counterpart, indicating that the benefits of pre-training did not successfully transfer. This critical finding demonstrates that simply increasing model scale is insufficient; the optimal configuration is scale-dependent and must be re-established to effectively leverage the capacity of larger models. This insight fundamentally shapes our future research, prioritizing strategies for co-designing model architecture and training methods across different scales.

Table 1. Downstream performance comparison of small (51M encoder) SwiFT V2 models. All experiments are repeated three times, and the average test metric is presented. We only show the best model from the fine-tuned models of all configurations. (UKB: UK Biobank; HBN: Healthy Brain Network; EMBARC: Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care; ADNI: Alzheimer's Disease Neuroimaging Initiative)

Method	UKB		HBN	EMBARC	ADNI
	Depression (AUROC ↑)	Intelligence (MAE ↓)	Emotion valence prediction (MAE ↓)	MDD treatment response (AUROC ↑)	MCI-to-AD conversion (AUROC ↑)
From scratch	0.645 ± 0.11	1.589 ± 0.01	0.671	0.603 ± 0.06	0.690 ± 0.14
Fine-tuning	**0.708 ± 0.01**	**1.577 ± 0.01**	**0.544**	**0.662 ± 0.03**	**0.741 ± 0.08**

Describe how access to leadership computing resources, data capabilities, your liaison, and/or staff at the Oak Ridge Leadership Computing Facility (OLCF) enabled these results.

Access to Frontier was indispensable for the success of this project. Its large number of powerful AMD GPUs and substantial node memory were essential for testing models at the multi-billion parameter scale (1.2B, 8.8B, 14.5B), and for validating advanced techniques like mu-parameterization and µTransfer, which would have been infeasible on smaller systems.

Furthermore, the stability and performance of the Frontier system were outstanding, allowing us to utilize 296,002 node-hours, significantly exceeding our initial allocation of 152,000 hours. This expanded access was made possible by crucial OLCF allocation policies, specifically the batch priority increase and the waiving of over-allocation penalties for ALCC projects. These policies enabled the extensive, high-throughput experimentation required to systematically test numerous configurations and complete our milestone objectives ahead of schedule. Finally, OLCF documentation and seminars on HPC best practices (e.g., sbcast, RCCL plugins) provided crucial context for optimizing our multi-node workflow.

C. Application Performance

Primary code(s) used by this project

SwiFT v2, a private, in-house developed 4D fMRI Transformer model based on PyTorch and DeepSpeed.

Application Performance

Summarize the performance (e.g., percent of peak, scalability) of your project's simulation codes used in the allocations for this project.

The performance of SwiFT v2 on Frontier scaled effectively, providing a crucial baseline for our INCITE proposal benchmarks. As shown in Figure 2, our 4.1B parameter model demonstrated robust strong scaling on Frontier up to 2048 nodes, serving as a direct comparison to validate Aurora's performance for our workload and suggesting the necessity for I/O optimization. In separate large-model tests, our 8.8B parameter model achieved 93.3% GPU utilization on 16 nodes and we were abled to train 14.5B parameter model on 64 nodes without nan loss, demonstrating high computational efficiency when the model size is sufficient to saturate the GPUs and overcome communication overhead.

What progress was made in improving the application's performance on LCF architecture during this project?

Performance optimization was a central achievement of this project. Our primary progress was evolving the application from an initial memory-constrained state to a highly efficient training framework. We began by using DeepSpeed ZeRO-3 with CPU Offload to overcome memory limitations, but transitioned to a more performant recipe using DeepSpeed Stage 1 BF16 without offloading. This shift significantly reduced computational overhead and improved throughput, especially for models in the 1B-10B parameter range.

This optimization process revealed crucial lessons about scaling dynamics on LCF architectures. While benchmarks demonstrated excellent strong scaling (Figure 3), our experiments also uncovered that at larger node counts (e.g., 64 nodes), the performance of smaller models (1.2B) was bottlenecked by relative I/O and communication overhead, resulting in lower GPU utilization. In contrast, the higher computational intensity of larger models (e.g., the 8.8B test) successfully saturated the GPUs, achieving high efficiency. This insight—that I/O, not just computation, becomes the primary bottleneck at scale—is a critical finding that has fundamentally shaped our strategy for data management and model training in future large-scale campaigns.

What challenges (if any) remain?

A key remaining challenge is the interoperability of data formats and filesystems across different Leadership Computing Facilities. While we successfully engineered a highly efficient HDF5-based data pipeline on Aurora to overcome I/O bottlenecks, the logistics of transferring and utilizing this optimized dataset on Frontier's filesystem for comparative scaling tests proved prohibitive within the project's scope. This highlights the ongoing challenge of seamless data portability for large-scale scientific campaigns across multiple sites.

List the technical risks and challenges that were confronted by your project (overcome or not) this year. Were they anticipated?

- Managing the substantial GPU memory required by 4D fMRI data, which was successfully overcome using DeepSpeed ZeRO optimizations. It was anticipated
- Intermittent nan loss issues, which were traced back to data path and transfer problems early in the allocation. It was not anticipated, but addressed through systematic debugging

- Unexpected node failures of specific experiments, which were discussed with OLCF staff (Logan W. Gillum) recently. We could not overcome this problem. It was not anticipated.
- Scaling inefficiencies due to heavy I/O workloads. We could not overcome this problem. It was anticipated.

Leadership Usage (OLCF will complete this table)

Allocation Year	**Percent of Usage from Batch Jobs Requesting Less than 20% of Systems' Available Resources**	**Percent of Usage from Batch Jobs Requesting Between 20% and 60% of Systems' Available Resources**	**Percent of Usage from Batch Jobs Requesting Greater than 60% of Systems' Available Resources**
2024–2025	%	%	%

D. Impact of Research

Impact Statement

The work on Frontier was instrumental in de-risking and defining the strategy for extreme-scale neuro-AI. By developing and validating a successful methodology for training models up to 14.5B parameters, we provided a robust blueprint for the community. The quantitative scaling benchmarks and readiness demonstrations on Frontier were a cornerstone of our successful 2025 INCITE proposal, transforming this foundational research into a concrete plan for building a 130B-parameter unified brain model. Crucially, our findings that smaller, 51M parameter encoder models show superior and more generalizable downstream performance, while initial results from larger 3B encoder models are still under investigation, provide an invaluable lesson for the community, guiding future research to explore more than a pure "bigger is better" approach and focus on nuanced and efficient strategies for building truly effective foundation models for neuroscience. When the final paper is submitted, we plan to share our code and model checkpoints with the research community.

Technical Accomplishments

- Development of an efficient, scalable training framework for 4D fMRI Transformers on AMD GPUs.
- Successful validation of mu-parameterization and mu-transfer for stable and efficient large-model training in neuroscience.
- Creation of optimized multi-node job scripts for Frontier utilizing sbcast and integrated GPU monitoring.

- Architectural enhancement of the SwiFT v2 model through the implementation of Rotary Positional Embedding (RoPE) and decoupled spatio-temporal attention blocks to improve temporal learning.

Publications and Presentations

Published

Kwon, J., Seo, J., Wang, H., Moon, T., Yoo, S., & Cha, J. (2025). Predicting task-related brain activity from resting-state brain dynamics with fMRI Transformer. Imaging Neuroscience, 3, imag_a_00440.

Accepted

Choi, J., Wang, H., Kwon, J., Yoo, S., & Cha, J. (2025). SwiFT V2: Towards Large-scale Foundation Model for Functional MRI. Abstract accepted for presentation at the Cognitive Computational Neuroscience (CCN) 2025 conference.

Han, D. D., Lee, A. L., Lee, T., & Cha, J. (2025). DIVER-0: A fully channel equivariant EEG foundation model. Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Li, S., Yoo, S., & Yang, Y. (2025). Maximal Update Parametrization and Zero-Shot Hyperparameter Transfer for Fourier Neural Operators. ICML.

Submitted

Park, J., Kim, P., Cha, J., Yoo, S, & Moon, T. (2025). SEED: Towards More Accurate Semantic Evaluation for Visual Brain Decoding https://arxiv.org/abs/2503.06437 (Planning to change acknowledgment soon)

Lee, D., Park, J., & Moon, T. (2025). DEAL: Decoupled Classifier with Adaptive Linear Modulation for Group Robust Early Diagnosis of MCI to AD Conversion

Invited Talks

None

Presentations

None

Software Releases

None (will release once the paper is accepted)

Dissertation

None

Post-doctoral Fellows and Students

- Jubin Choi, PhD student, Seoul National University
- Heehwan Wang, PhD student, Seoul National University

E. Highlights

xx produced an OLCF highlight related to this work.

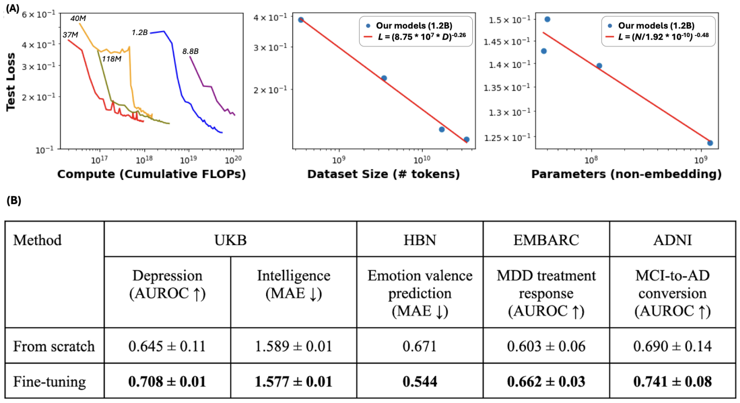
[Insert Photo]

F. One-Page PowerPoint (PPT) Elements (Required at CLOSEOUT)

a. **The Science** – The immense complexity of the human brain has made it difficult to identify reliable biomarkers for mental and neurological disorders. Our project aims to create a 'foundation model' for neuroscience by learning the fundamental principles of brain function directly from large-scale 4D fMRI data. By leveraging AI and exascale computing, we have developed a model capable of understanding the brain's intricate spatio-temporal dynamics. This foundational approach allows us to decode brain activity and predict cognitive and clinical outcomes, moving beyond the limitations of previous methods.

b. **The Impact** – This project delivered a validated, scalable recipe for training large-scale AI models on fMRI data by implementing key architectural improvements and advanced scaling techniques on Frontier. By successfully training models up to 14.5 B parameters, we provided a blueprint that the neuroscience community can adopt. Critically, our work revealed that smaller, more efficient 51M models consistently outperform from-scratch models on challenging clinical prediction tasks. Our initial evaluations of larger 3B models are underway, and these pivotal experiments are challenging common AI scaling assumptions, providing a new, more efficient path forward for developing foundation models capable of discovering novel biomarkers.

c. **Image**



Method	UKB		HBN	EMBARC	ADNI
	Depression (AUROC ↑)	Intelligence (MAE ↓)	Emotion valence prediction (MAE ↓)	MDD treatment response (AUROC ↑)	MCI-to-AD conversion (AUROC ↑)
From scratch	0.645 ± 0.11	1.589 ± 0.01	0.671	0.603 ± 0.06	0.690 ± 0.14
Fine-tuning	**0.708 ± 0.01**	**1.577 ± 0.01**	**0.544**	**0.662 ± 0.03**	**0.741 ± 0.08**

d. **Caption for Image** – Figure: (A) Neural scaling laws of SwiFT V2 and (B) Downstream task performance of 51M SwiFT V2. (A) SwiFT V2, fMRI swin transformer model's pretraining loss improves with more compute, data, and parameters. (B) Pretrained 51M SwiFT V2 models outperformed models trained from scratch not only on the in-distribution dataset (UKB), but also on external validation datasets (HBN, EMBARC, ADNI).

e. **Related Publication Citation.** - Choi, J., Wang, H., Kwon, J., Yoo, S., & Cha, J. (2025). SwiFT V2: Towards Large-scale Foundation Model for Functional MRI. Abstract accepted for presentation at the Cognitive Computational Neuroscience (CCN) 2025 conference.