End of Project Summary

PI NAME : Shinjae Yoo

YEAR : 2025

PROGRAM : ALCC

Project Usage

The utilization of our ALCC allocation was strategically phased to maximize both scientific discovery and technical enablement on ALCF's leadership-class systems. The project commenced with a period of relatively low utilization on Polaris, which was dedicated to essential preparatory work. This included transferring and organizing massive fMRI datasets for over 50,000 subjects and undertaking significant code development to ensure our framework was compatible with modern libraries like PyTorch and DeepSpeed, and PBS job scheduler, building on work from a prior 2024 ALCF GPU Hackathon.

The project's focus then pivoted to a high-utilization campaign on Aurora. A critical performance barrier was encountered early in this phase: our data pipeline, which relied on millions of small *.pt* files, created a severe I/O bottleneck on Aurora's parallel filesystem, extending epoch times to over two hours. The primary barrier to scaling was resolved by re-architecting the pipeline to a per-subject HDF5 file format with Gzip compression. This optimization was transformative, **reducing epoch times to approximately 10 minutes** and enabling the high-throughput experimentation necessary for our large-scale pre-training and systematic downstream task evaluations.

Notably, we adapted our experimental plan to leverage Aurora's capabilities more fully. While our original proposal designated Polaris for downstream fine-tuning tasks, we found that conducting these evaluations directly on Aurora was significantly more efficient. The flexibility of using as few as a single node on Aurora (compared to a 10-node minimum on Polaris at the time) allowed for rapid, iterative testing without the overhead of transferring large model checkpoints between systems. Consequently, our resource usage on Aurora peaked during these intensive, combined pre-training and fine-tuning campaigns.

Report on Project Milestones

Original Milestones from Proposal:

- Milestone 1: Finding Optimal NeuroX Configurations.
- Milestone 2: Pretraining NeuroX.
- Milestone 3: Fine-Tuning NeuroX.

Status of Milestones:

Significant progress has been made on all milestones. All Milestones have been achieved, with their direction informed by the key scientific and computational discoveries made during the project.

- **Milestone 1 & 2**
  We successfully executed a comprehensive, three-phase experimental plan on Aurora to identify optimal configurations and pre-train our model, SwiFT v2 (under the project name NeuroX). This involved systematically evaluating numerous model configurations and learning rates (**Table 1**). A key success was the empirical validation that our fMRI foundation model adheres to **neural scaling laws**, with pre-training loss consistently improving with more compute, data, and parameters. This foundational result has been accepted for presentation at CCN 2025. This work culminated in the successful pre-training of models of various sizes up to a **4.1B parameter model (3B encoder)**. Exploratory runs with larger models (up to 13.4B parameters) were instrumental in identifying I/O bottlenecks at larger node counts (e.g., >128 nodes) even with HDF5 files, which has informed our ongoing work.

Table 1. Pre-training and evaluation on various downstream tasks to find the optimal model configurations using a 51M-parameter model (Milestone 1 & 2). We investigated the performance of eight types of model configurations (input scaling method, masking patch size, masking ratio, masking type, patch size, overlap ratio between input segments, sequence lengths, and stride between input segments) on various downstream tasks and selected the best model configuration, masking ratio 0.6.

Models	ZnormSum_noTOPS	RANK_ZnormSum_noTOPS	ZnormSum_noTOPS_addADNI_HBN	RANK_noTOPS_addADNI_HBN	ZnormSum_noTOPS_addADNI	RANK_noTOPS_addADNI
From_scratch	0.344628711	11	1.529021992	5	1.108059671	8
P1_51M_epoch80 (default)	0.817847012	7	-0.759274822	12	1.05931269	9
P1_51M_ISM_minmax	-2.587367088	25		#N/A	-2.459372124	24
P1_51M_ISM_znorm_zeroback	-2.361719016	23		#N/A	-3.334389971	25
P1_51M_MPS_12_12_12_2	1.237023585	6	2.026961875	3	1.523877548	6
P1_51M_MPS_12_12_12_4	-0.014410399	14	0.648448598	6	0.227055279	13
P1_51M_MPS_24_24_24_2	-1.607393963	22	-3.175365836	14	-1.365928285	20
P1_51M_MPS_6_6_6_4	0.302071053	12	0.077759865	10	-0.443658475	15
P1_51M_MR_0p2	0.675888517	9	-0.421128993	11	-0.716624078	18
P1_51M_MR_0p4	1.604965669	5	0.121867446	9	1.94855499	5
P1_51M_MR_0p6	2.976098146	**1**	3.400942754	2	2.945234112	3
P1_51M_MR_0p9	2.356871839	**2**	**3.920237699**	**1**	**3.483409081**	**1**
P1_51M_MT_temporal	-0.219261643	16	0.437718273	8	-0.034531322	14
P1_51M_MT_tube	0.713645384	8	1.689937159	4	1.216093703	7
P1_51M_PS_12_12_12_1	-0.851147481	20	0.54497267	7	-0.507558161	16
P1_51M_PS_12_12_12_2	-2.405096215	24	-1.943985605	13	-2.345183678	23
P1_51M_PS_12_12_12_4	-0.805300341	19		#N/A	0.536831257	11
P1_51M_PS_6_6_6_1	2.211497869	**3**		#N/A	3.133787827	2
P1_51M_PS_6_6_6_4	-2.798333743	26		#N/A	-3.87312834	26
P1_51M_SBS_0p5	2.071985868	**4**		#N/A	0.872373486	10
P1_51M_SBS_0p75	-0.950421441	21		#N/A	-1.151491545	19
P1_51M_SL_10	-0.208262253	15		#N/A	-2.020616487	22
P1_51M_SL_40	0.480413129	10		#N/A	-1.683700316	21
P1_51M_SWS_2	-0.477720207	17		#N/A	0.43322268	12
P1_51M_SWS_3	0.029893619	13		#N/A	1.984767069	4
P1_51M_SWS_6	-0.53639661	18		#N/A	-0.53639661	17

- **Milestone 3**
  This milestone is the current focus of the project. We are conducting extensive downstream task fine-tuning and evaluation on Aurora, leveraging the checkpoints from our pre-training runs. This phase has already yielded a crucial insight: our smaller, pre-trained 51M parameter models robustly outperform larger models on a majority of downstream tasks. Based on this, our fine-tuning efforts have become more sophisticated. We are now systematically evaluating advanced, parameter-efficient fine-tuning (PEFT) techniques—such as **LoRA, prompt tuning, and using complex**

task-specific heads with a frozen encoder—to determine the most effective methods for transferring knowledge from our pre-trained models. This work will be applied across our suite of models and an expanding number of diverse downstream datasets.

Key Scientific Findings and Current Status of the Project:

This project has yielded significant insights into the effectiveness of pre-training for downstream neuroscience tasks, moving from foundational validation to the discovery of complex, nuanced behaviors in large-scale models.

Finding 1: Pre-training is a Highly Effective Foundation

A primary and robust finding is the powerful performance of our pre-trained 51M parameter model. Our systematic evaluations consistently revealed that **fine-tuning only 30% of the total parameters of this 51M model led to superior performance compared to fully training an identical model from scratch.** As detailed in **Table 2** and **Figure 1**, this result was validated across various in-distribution and external datasets. This confirms that our pre-training methodology is highly effective at this scale for creating models that generalize well.

Table 2. Downstream performance comparison of the 51M SwiFT V2 model. All experiments are repeated three times, and the average test metric is presented.

Method	UKB		HCP	HBN	EMBARC	ADNI	Woo_ToPS
	Depression (AUROC ↑)	Intelligence (MAE ↓)	Intelligence (MAE ↓)	Emotion valence prediction (MAE ↓)	MDD treatment response (AUROC ↑)	MCI - AD conversion (AUROC ↑)	Task condition prediction (AUROC ↑)
From scratch	0.640 ± 0.07	1.638 ± 0.06	17.593 ± 0.16	0.783	0.603 ± 0.06	0.690 ± 0.14	**0.954 ± 0.05**
Fine-tuned last block	**0.708 ± 0.01 (10.62%↑)**	**1.572 ± 0.02 (4.03%↓)**	**16.596 ± 0.03 (5.67%↓)**	**0.544 (30.52%↓)**	**0.662 ± 0.03 (9.78%↑)**	**0.741 ± 0.08 (7.39%↑)**	0.898 ± 0.04 (5.87%↓)

Abbreviations for the dataset and prediction targets are as follows. UKB: UK Biobank; HBN: Healthy Brain Network; EMBARC: Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care; ADNI: Alzheimer's Disease Neuroimaging Initiative; Woo_ToPS: Tonic Pain Signature fMRI data from Woo and Lee et al. 2021[ref]. MDD: Major Depressive Disorder; MCI: Mild Cognitive Impairment; AD: Alzheimer's Disease.

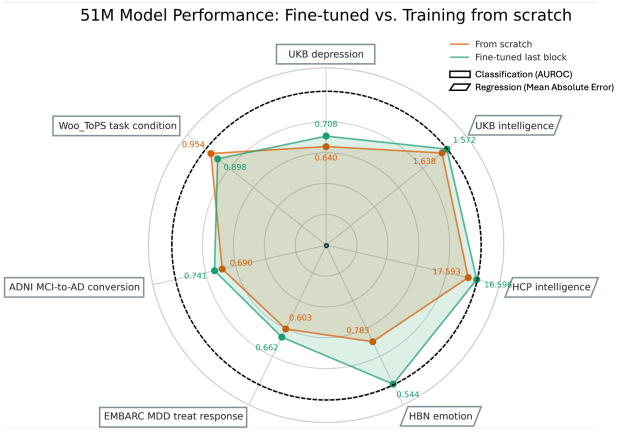


Figure 1. Downstream task performance of the 51M parameter SwiFT V2 model. The radar plot compares the performance of a model trained from scratch against a pre-trained model that was fine-tuned on each task. The outer black dashed circle represents the ideal performance for each task (an AUROC of 1.0 for classification and the minimum Mean Absolute Error for regression), with values closer to the circle indicating better performance.

Finding 2: Data Diversity is Crucial for Predictable Scaling

Our work in scaling these models revealed a critical requirement for achieving predictable performance improvements. We discovered that pre-training on a single, large dataset (UK Biobank) was insufficient to produce a clear scaling law with respect to pre-training loss. However, by diversifying the pre-training corpus to include the ABCD and HCP datasets, a robust and predictable scaling law emerged, where pre-training loss consistently decreased as more data, compute, and parameters were added (**Figure 2**). This finding underscores the importance of data diversity for the stable training of large foundation models in this domain.

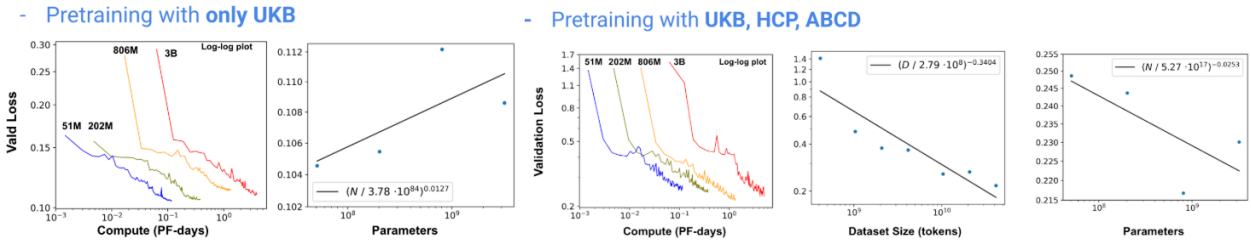


Figure 2. Neural scaling laws of SwiFT V2. The model's pre-training loss improves with more compute, data, and parameters when using a diverse dataset (right), but not when trained on only the UKB data (left).

Finding 3: "Larger is Not Always Better" — Uncovering Complex Generalization

Table 3. Downstream performance comparison of SwiFT V2 models of varying parameter sizes. All experiments were repeated three times, with the average and standard deviation of the test metric presented. For each model size, performance is compared between training from scratch and fine-tuning the last block of a pre-trained model. The percentage change from fine-tuning is shown in parentheses. Within each model-size pairing, the bolded value indicates the superior result. The bolded and underlined value indicates the best result among all methods.

Method		UKB		HCP	EMBARC	Woo_ToPS
		Depression (AUROC ↑)	Intelligence (MAE ↓)	Intelligence (MAE ↓)	MDD treatment response (AUROC ↑)	Task condition prediction (AUROC ↑)
51M	From scratch	**0.640 ± 0.07**	1.638 ± 0.06	17.593 ± 0.16	0.603 ± 0.06	**0.954 ± 0.05**
	Fine-tuned last block	0.525 ± 0.03 (17.97% ↓)	**<u>1.572 ± 0.02</u> (4.03% ↓)**	**<u>16.596 ± 0.03</u> (5.67% ↓)**	**0.611 ± 0.06 (1.32% ↑)**	0.898 ± 0.04 (5.87%↓)
202M	From scratch	**<u>0.679 ± 0.02</u>**	1.628 ± 0.01	17.940 ± 0.44	-	-
	Fine-tuned last block	0.483 ± 0.03 (28.87% ↓)	**1.579 ± 0.01 (3.01% ↓)**	**17.398 ± 0.97 (3.02% ↓)**	0.590 ± 0.05	0.467 ± 0.04
806M	From scratch	**0.649 ± 0.01**	1.621± 0.00	**17.273 ± 0.10**	-	-
	Fine-tuned last block	0.582 ± 0.03 (10.32% ↓)	**1.586 ± 0.00 (2.16% ↓)**	18.694 ± 0.53 (8.23% ↑)	0.629 ± 0.06	**<u>0.962 ± 0.03</u>**
3B	From scratch	**0.636 ± 0.05**	**1.589 ± 0.01**	17.689 ± 0.37	-	-
	Fine-tuned last block	0.461 ± 0.03 (27.52% ↓)	1.591 ± 0.00 (0.13% ↑)	**16.748 ± 0.20 (5.32% ↓)**	**<u>0.637 ± 0.03</u>**	0.913 ± 0.04

While larger models performed better during the pre-training phase (as predicted by the scaling law), this advantage did not consistently translate to superior performance on downstream tasks (**Table 3**). It has been a key scientific discovery of our project, revealing that the optimal model size is highly task-dependent:

- Pre-trained **51M** parameter model achieved the best overall performance on both **UKB and HCP intelligence prediction** when fine-tuned.
- For **UKB depression classification**, the best result came from a medium-sized 202M model that was trained from scratch, outperforming all fine-tuned models for that task.
- However, for the external datasets, larger models showed better results. For the **EMBARC MDD treatment response task**, the largest fine-tuned **3B** parameter model performed best, while the **806M** model was the top performer for the **Woo_ToPS task condition prediction**.

Finding 4: A New Paradigm for fMRI Analysis: Direct Sequence Modeling with NeuroMamba

Motivated by the rigid, grid-based limitations of hierarchical models, we pioneered a more flexible and powerful approach by developing **NeuroMamba**, a foundation model for the direct sequence modeling of 4D whole-brain fMRI. Using a Mamba state-space backbone trained with an autoregressive objective, this model treats fMRI data as a sequence of spatiotemporal patches, enabling a key innovation: the adaptive removal of computationally expensive background signals **(Figure 3)**.

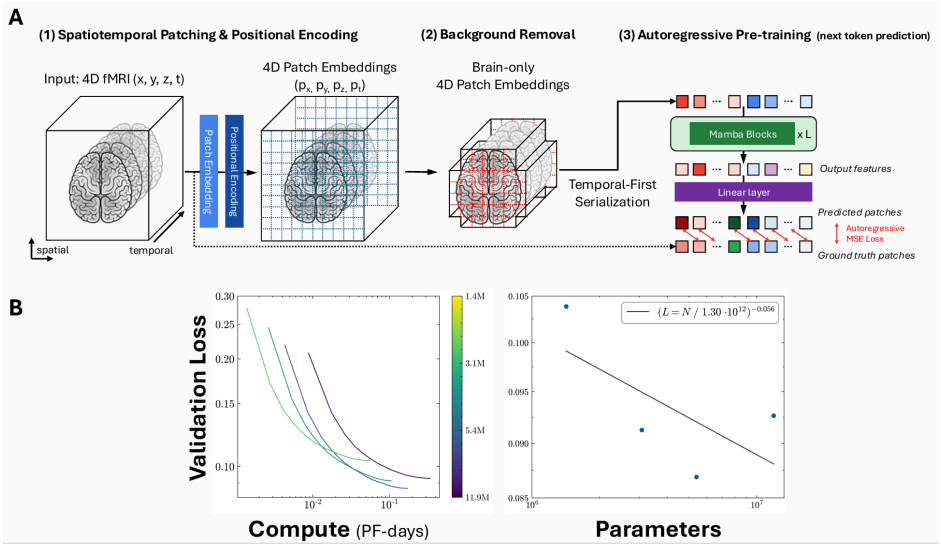


Figure 3. Overview of NeuroMamba workflow and neural scaling laws. (A) The pre-training pipeline converts 4D fMRI data into a sequence of spatiotemporal patches, adaptively removes non-brain background tokens, and trains a Mamba backbone using an autoregressive next-token prediction objective. (B) The model demonstrates neural scaling laws, as validation loss consistently decreases with more compute (left) and shows a power-law relationship with an increasing number of parameters (right).

This new paradigm yielded substantial improvements in both efficiency and effectiveness. By training only on brain-relevant tokens, we reduced the computational cost (total FLOPs per epoch) by **46.5%** while simultaneously improving training stability and model performance. This culminated in NeuroMamba achieving a new **state-of-the-art accuracy of 94.9%** on the challenging HCP sex classification benchmark, surpassing prior methods. The success of NeuroMamba validates direct sequence modeling as a viable and potent direction for the future of fMRI analysis. We submitted this work to Foundation Models for the Brain and Body NeurIPS 2025 Workshop.

Current Status

Scale up SwiFT v2. The project is currently in the midst of Milestone 3. We have successfully pre-trained models up to the 4.1B parameter scale, and initial downstream evaluations for these models are now underway. Guided by the findings above, our work is now focused on applying diverse and advanced fine-tuning strategies to clarify the relationship between model scale and effective knowledge transfer in this domain. We are also investigating the I/O workflow to enable training the 10B model on larger nodes.

Improving the model architecture of SwiFT v2. Although our results have successfully demonstrated the potential of scaling the SwiFT v2 model, several architectural limitations hinder its ability to model the long-term temporal context of fMRI data. First, the parameter size of the 4-dimensional spatio-temporal window attention increases exponentially as the temporal window size expands. Second, the use of absolute positional embeddings restricts the model's capacity to capture relative temporal dynamics, necessitating training on extremely long sequences to learn complex dependencies. To address these limitations, which create a significant GPU memory bottleneck, we have developed two key improvements for the SwiFT v2 architecture: factorizing the unified spatio-temporal attention into separate spatial and temporal mechanisms, and adopting Rotary Positional Embeddings (RoPE) to effectively capture both absolute and relative positional information (see **Figure 4**).

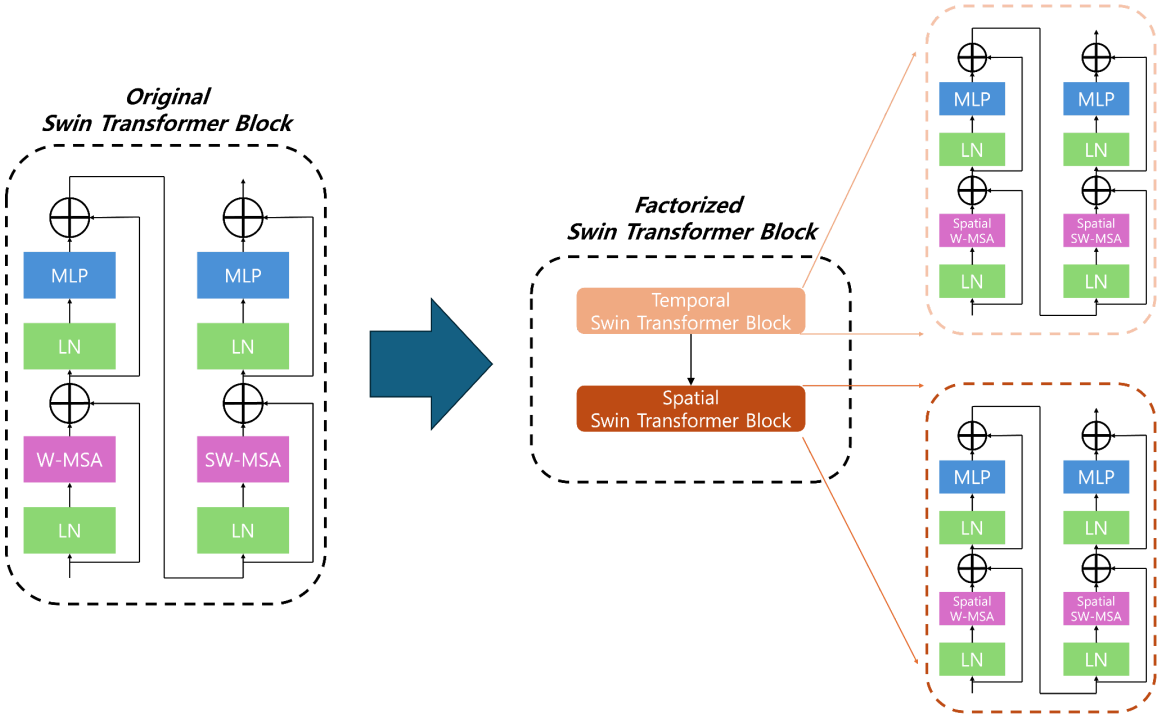
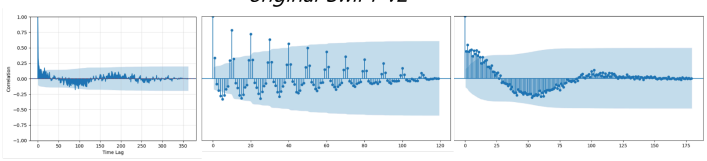


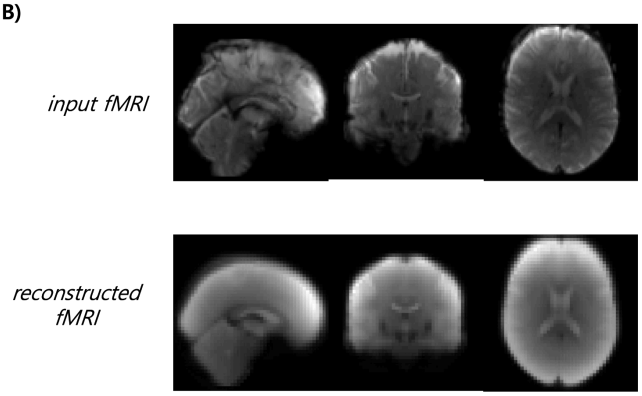
Figure 4. Factorization of spatial/temporal shifted-window attention and Neural scaling laws of SwiFT V2. Revised SwiFT v2 factorizes 4-dimensional spatio-temporal shifted-window attention into 1+3 dimensional spatial/temporal shifted-window attention. Also, we replaced the absolute positional embeddings of the original SwiFT v2 with rotary positional embeddings.

To validate these enhancements, we performed a preliminary analysis of the revised SwiFT v2. An autocorrelation analysis of its embeddings revealed a smoother correlation pattern, more akin to the original input data and devoid of the artificial peaks seen in the original model, which suggests an improved ability to capture long-term temporal dependencies (**Figure 5-A**). In addition, the revised model demonstrated high-fidelity fMRI reconstruction that successfully preserved spatial patterns (**Figure 5-B**). Critically, these improvements were achieved through attention factorization, a technique that also enables the model to be trained on longer fMRI sequences within the same GPU memory footprint (**Figure 5-C**). In a preliminary performance evaluation on the challenging task of classifying Major Depressive Disorder (MDD) in adolescents, the revised SwiFT v2 achieved an encouraging 20.5% increase in AUC compared to the original model (see **Table 4**). Building on these positive results, we plan to conduct a full-scale pre-training of the revised model using a masked image modeling approach. Subsequently, we will systematically evaluate its performance across the comprehensive suite of downstream tasks established in our prior work to quantify the improvements.

A)



B)



C)

	Parameter Size	Time Length	Memory Usage (GB)
original SwiFT v2	**807M**	**20**	**18.38**
revised SwiFT v2	**839M**	**20**	**17.45 (5% ↓)**

Figure 5. Validation of revised SwiFT V2 Architectural Enhancements. A) shows the autocorrelation function of data used for pre-training and its embeddings from revised SwiFT v2. B) shows a frame of reconstructed fMRI during revised SwiFT v2 training with masked image modeling, indicating that revised SwiFT v2 learned the spatial pattern of fMRI images. C) shows the comparison between original SwiFT v2's memory usage and revised SwiFT v2's memory usage.

Table 4. Model performance of original and revised SwiFT v2 on classifying children's prior diagnoses of Major Depressive Disorder.
All experiments were repeated three times, with the average and standard deviation of the test metric presented. Within each model-size pairing, the bolded value indicates the superior result. The bolded and underlined value indicates the best result among all methods.

Method		Dataset	Depression (AUROC ↑)
original SwiFT v2	806M	ABCD	0.5022 ± 0.032
revised SwiFT v2	837M		**<u>0.6052 ± 0.019</u> (20.5% ↑)**

Additional major accomplishments:

- **Full Enablement and Scaling on Aurora:**
  We successfully ported and validated the entire SwiFT v2 framework on the Aurora XPU architecture. This was a significant undertaking that required developing custom PyTorch Lightning components and debugging complex multi-node issues related to the XPU environment. This effort culminated in a successful 13.4B parameter scalability test run on 256 nodes, confirming our code's readiness for large-scale science campaigns on Aurora.

- **Critical I/O Pipeline Optimization**:
  We successfully diagnosed and resolved a major I/O bottleneck that initially hindered our scaling on Aurora. By re-architecting our data storage from millions of individual files to a per-subject HDF5 file with Gzip compression, we stabilized performance and dramatically improved efficiency. This optimization was the single most crucial step that enabled all subsequent large-scale experiments on Aurora.

- **Foundational Results and Peer-Reviewed Dissemination:**
  Our initial experiments, which focused on establishing the viability of our approach, including advanced techniques like *μTransfer*, yielded promising results. These preliminary findings, which demonstrated the potential of our model, formed the basis of a conference abstract that was accepted for presentation at the 2025 CCN **(Figure 6)**.

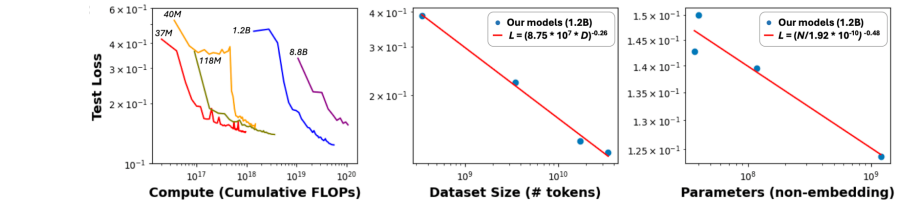


Table 1. Downstream Task Performance
(ACC: Accuracy; MAE: Mean Absolute Error)

Method	ABCD		HCP			UKB		
	Sex (ACC)	Intelligence (MAE)	Sex (ACC)	Age (MAE)	Intelligence (MAE)	Sex (ACC)	Age (MAE)	Intelligence (MAE)
From scratch	50.0	0.86	50.0	3.17	16.2	50.6	7.09	1.74
Fine-tuned	**50.2**	**0.85**	**57.6**	**3.13**	**16.1**	**56.9**	**6.44**	**1.72**

Figure 6. Key results from the accepted CCN 2025 abstract. It demonstrates initial scaling laws (top), successful implementation of *μTransfer* (bottom left), and downstream task performance of using only 100 training subjects (bottom right)

- **High-Performance Computing Scalability:**
  We conducted rigorous strong scaling benchmarks of our optimized 4.1B parameter SwiFT v2 model. As shown in **Figure 7** and **Table 5**, the results demonstrate exceptional performance on Aurora, achieving 6.7 times the throughput of Frontier and maintaining 60.5% scaling efficiency at 512 nodes (when re-baselining at 64 nodes). This validates our code's high level

of optimization for the ALCF architecture and its capability to efficiently leverage leadership-class resources.

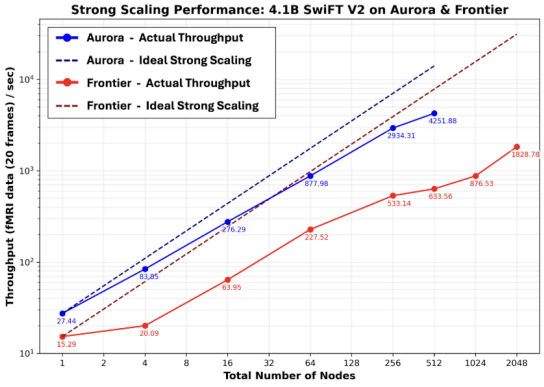


Figure 7. Strong scaling performance of the 4.1B SwifT V2 model on Aurora and Frontier.

Table 4. Details of the scaling performance of the 4.1B SwiFT model training on Aurora

Nodes	# GPU-ranks	Throughput (sample/s)	Ideal scaling & Efficiency (Baseline = N1)	Ideal scaling & Efficiency (Baseline = N64)
1	12	27.44	-	-
4	48	83.85	109.77 (76.4%)	-
16	192	276.29	439.07 (62.9%)	-
64	768	877.98	1756.28 (50.0%)	-
256	3072	2934.31	7025.14 (41.8%)	3511.91 (83.6%)
512	6144	4251.88	14050.27 (30.3%)	7023.83 (60.5%)

Project Productivity

- Accepted Conference Abstract:
  - Choi, J., Wang, H., Kwon, J., Yoo, S., & Cha, J. (2025). SwiFT V2: Towards Large-scale Foundation Model for Functional MRI. Abstract accepted for the Cognitive Computational Neuroscience (CCN) 2025 conference. This work was a two-week-long science campaign on a maximum of 256 nodes of Aurora.
  - 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.

    These results were obtained using resources from an ALCC project.
- Journal Publication (Supported by project methods):
  - 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.

This work utilized methods and insights developed during the ALCC project.

- Submitted papers
    - **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
      These results were obtained using resources from an ALCC project. We are planning to change the 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**
      These results were obtained using resources from an ALCC project.
    - **Choi, J., Park, D., Kwon, J., Yoo, S., & Cha, J. (2025). NeuroMamba: A State-Space Foundation Model for Functional MRI.**
      Submitted to the Foundation Models for the Brain and Body NeurIPS 2025 Workshop. These results were obtained using resources from an ALCC project.
    - **Park, J., Kim, P., Park, J., Choi, J., Seo, J., Cha, J., & Moon, T. (2025). Processing fMRI Brain Signals Using Latents from Natural Image Autoencoders.**
      Submitted to the Foundation Models for the Brain and Body NeurIPS 2025 Workshop. These results were obtained using resources from an ALCC project.

Code Description

The SwiFT v2 code underwent significant evolution throughout this project. It matured from a research-level GPU code into a robust, portable HPC framework. The most substantial improvements included:

1. **Full Adaptation for Aurora's XPU Architecture**: This required creating custom backend integrations for PyTorch Lightning and DeepSpeed (XPUAccelerator, XPUDeepSpeedStrategy) to ensure seamless execution.
2. **Integration of Advanced Scaling Techniques**: The codebase was enhanced with mu-parameterization to ensure stable training of multi-billion parameter models across different hardware platforms.
3. **Data Pipeline Re-architecture**: The data loading modules were fundamentally changed to support a highly efficient HDF5-based data format, resolving critical I/O bottlenecks encountered on large-scale parallel filesystems. This change was the single most impactful performance improvement for this project on ALCF resources.

Next Steps

The unexpected yet crucial findings of this project have clearly defined our next steps.

- **Upcoming Analysis**: Our analysis will focus on understanding the mechanisms of effective knowledge transfer in neuroimaging AI. We will investigate why the 51M parameter models generalize so effectively to downstream tasks, while also exploring the optimization and convergence dynamics of the 3B encoder models to better understand their initial downstream performance. This will involve investigating representational similarity, catastrophic forgetting during fine-tuning, and the complexity of the features learned by models of different scales to determine the optimal strategies for leveraging pre-trained models.
- **Future Papers/Presentations**: The results of our three-phase experimental plan, particularly the superior performance of smaller models and the disconnect between pre-training and downstream objectives, will form the basis of a major upcoming manuscript. The successful scaling on Aurora and the I/O optimization strategy will also be detailed. After the paper is submitted, we will share our code and model checkpoints with the research community.
- **Future Work & Allocations**: The scientific direction of this project has pivoted based on our key findings. We recently submitted a 2026 INCITE proposal to continue this research. If awarded, we plan to leverage the validated and highly efficient SwiFT v2 framework on ALCF resources to:
    a. Perform targeted, long-duration pre-training of the large-size models (e.g., 1.5B, 10B) to ensure full convergence and maximize their downstream potential.
    b. Systematically explore alternative fine-tuning strategies (e.g., LoRA) to improve knowledge transfer, particularly for larger-scale models.
    c. Improve overall scaling efficiency by optimizing the I/O workflow using DAOS and by systematically profiling performance bottlenecks in multi-node training experiments to prepare for extending our model from an unimodal model to a multimodal MRI model.

Other Comments

The assistance from ALCF support teams, particularly guidance from **Kyle Felker** and **Kaushik Velusamy**, was beneficial. Early discussions about interpreting system warnings in our code and providing informative sources during the Aurora enablement phase were crucial for maintaining momentum and making informed decisions, which ultimately led to resolving key technical hurdles. Furthermore, recent discussions about our I/O workflow led directly to the HDF5 data re-architecture, which produced significant improvements in throughput.

Also, we are extremely grateful for the ALCF's flexible and supportive allocation management. The extension of the ALCC award and the policies that increased batch priority and waived overburn penalties were instrumental to our success. These policies provided us with the necessary computational resources and runtime flexibility to complete our comprehensive three-phase experimental plan, including the many systematic evaluations that led to our project's most critical scientific discoveries. This support was essential for maximizing the scientific return on our allocation.