PhySiViT : A Physics Simulation Vision Transformer

Jessica Ezemba[1], James Afful[2], Mei-Yu Wang[3] (Advisor)

[1] Carnegie Mellon University, Pittsburgh, PA, USA; [2] Iowa State University, Ames, IA, USA; [3] Pittsburgh Supercomputing Center, Pittsburgh, PA, USA

PhySiViT HuggingFace

The Well

Introduction

Scientific simulations generate petabytes of complex, multi-channel, time-evolving data. Current machine learning models are narrow, domain-specific, and unable to generalize across physics types**.**

Vision transformers (ViTs) have revolutionized natural image understanding via models like CLIP and DINO, but **no equivalent foundation models exist for scientific spatiotemporal data**

Objective

To train a foundational image encoder model (***PhySiViT***) for scientific spatiotemporal data, enabling downstream tasks such as **classification**, **temporal forecasting**, and **embedding visualization.**

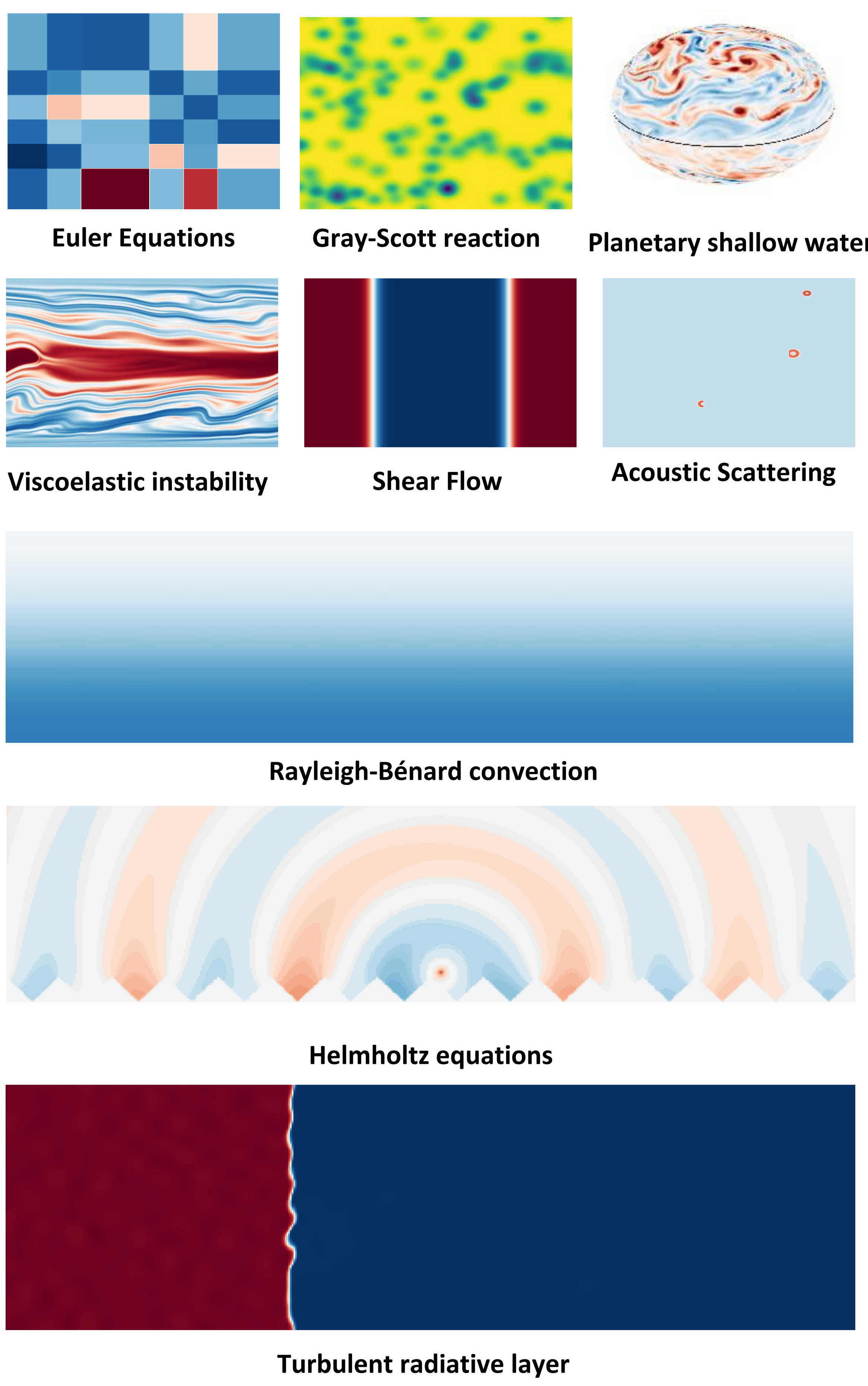

*Figure 1: The following simulation classes were selected from **The Well**, which resulted in 7M 2D images to train the Vision Transformer.*

For downstream task evaluation:

- ***Classification***: logistic regression for domain labels (accuracy).
- ***Temporal forecasting***: ridge regression for next timestep prediction (R^2/MSE).
- ***Embedding visualization***: silhouette score for domain separation.

Methods

PhySiViT uses a standard Huge Vision Transformer architecture[2], summarized below. Custom augmentations, such as temporal splits and color schemes, were used on input data. physics

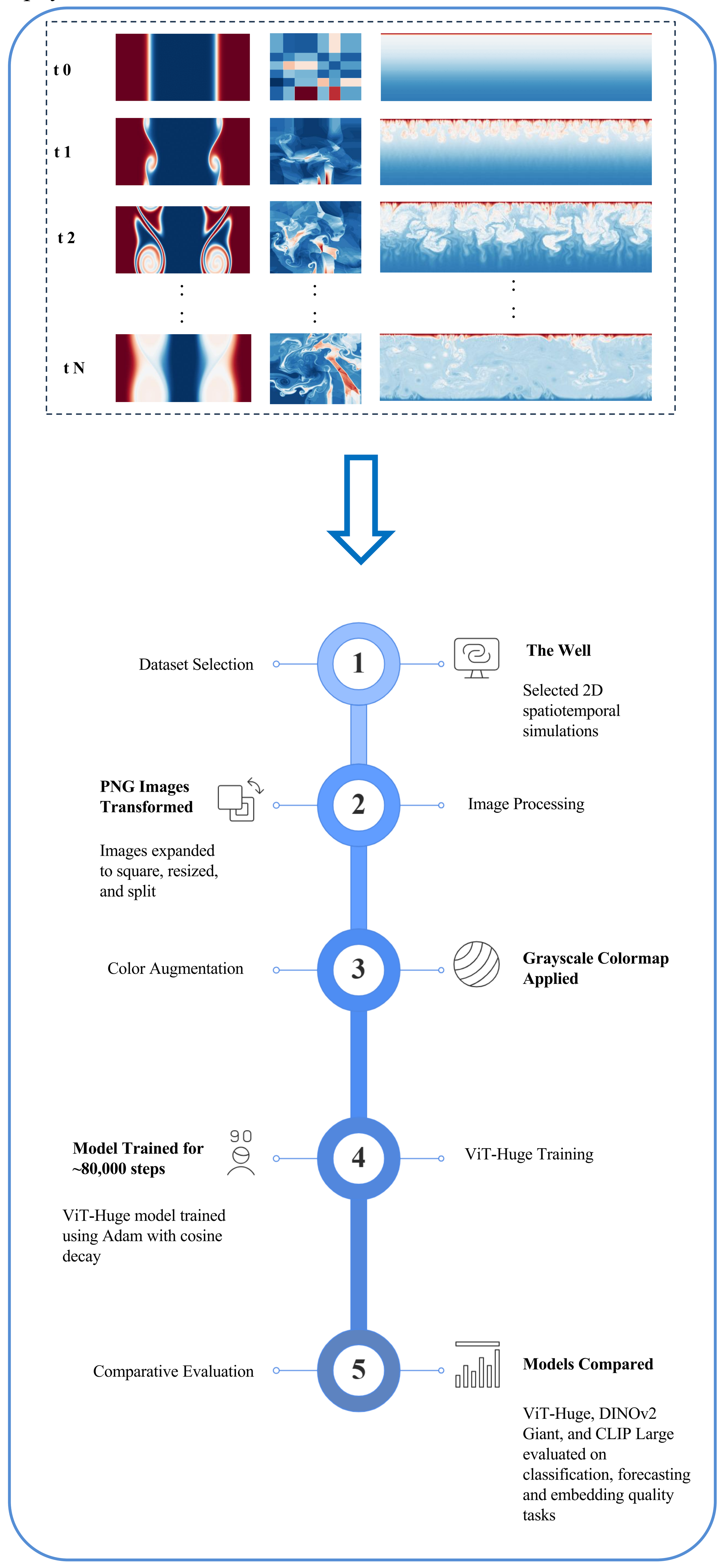

Figure 2: Model training and evaluation pipeline

Results

PhySiViT outperforms large natural-image foundation models on physics-relevant tasks, with notable gains in temporal forecasting and embedding separability that highlight the value of domain-specific pretraining.

Model	Accuracy (↑)	R^2 (↑)	MSE (↓)	Silhouette (↑)
PhySiViT	0.98	**0.33**	**0.57**	**0.23**
DINOv2 Giant	**0.99**	0.23	0.62	0.20
CLIP-ViT Large	0.99	0.22	0.63	0.19

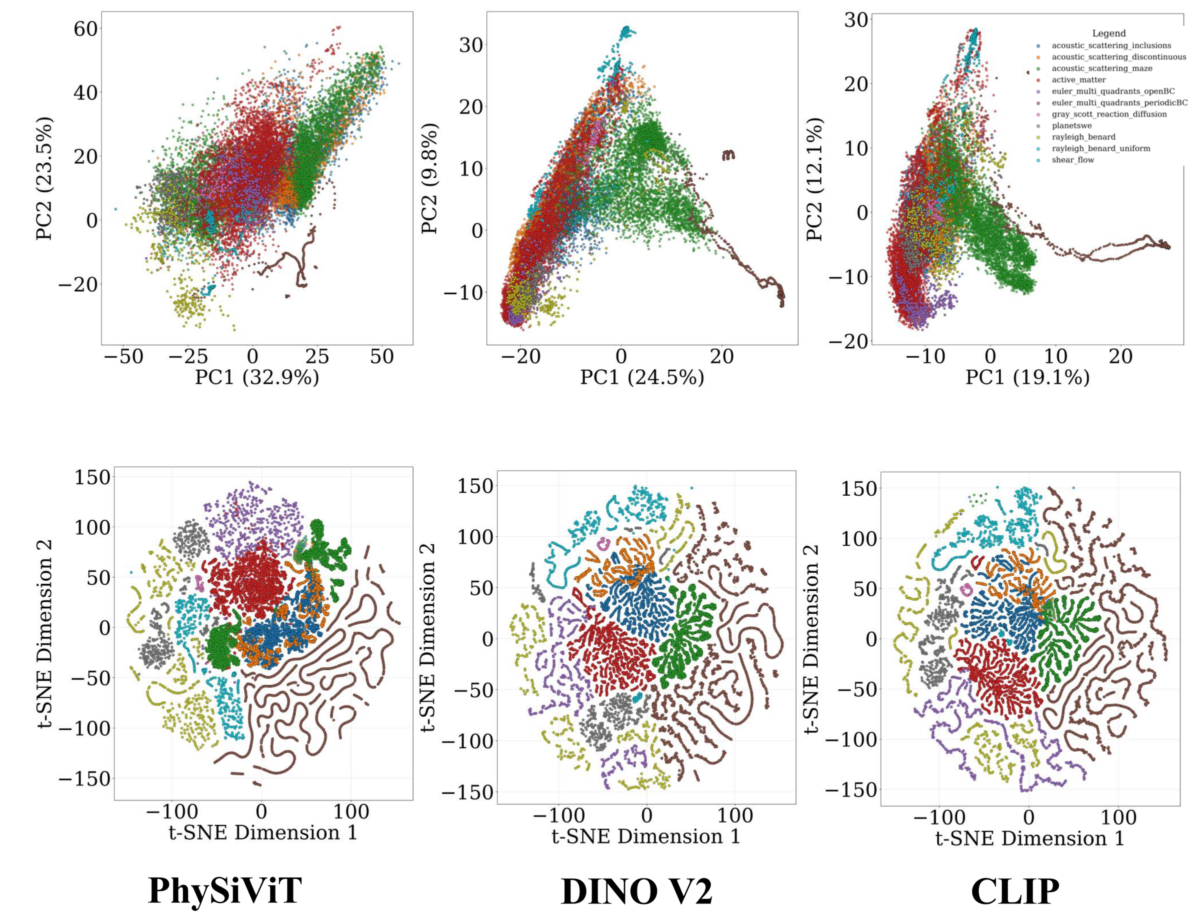

Figure 3: Principal-component visualization of learned embeddings

Training Considerations

Training ***PhySiViT*** on the **Cerebras CS-3** system achieved the fastest throughput; processing ~70 k images in only **0.20 hours**, compared with 2.50 hours on an NVIDIA H100 GPU and 14.03 hours on an AMD EPYC CPU.

PhySiViT was trained on approximately 7 million images from The Well using a Cerebras CS-3, reaching convergence in 22.83 hours with a batch size of 1,470.

Training Speed Comparison (Single Device with 70 k Images)

Compute Device	Total Time (hours)
AMD EPYC 7702P (CPU)	14.03
NVIDIA H100 80GB (GPU)	2.50
Cerebras CS-3	**0.20**

Conclusion & Future Work

PhySiViT is better at **physics-related tasks**, such as temporal prediction, and is comparable in general classification tasks, particularly silhouette scores represent ***clearer cluster of physical simulation domains in embedding space***.

References

1. Brashear, W., Chakravorty, D., He, Z., O'Connor, D., Siegmann, E., Buitrago, P. A., & Sanielevici, S. (2025). ByteBoost: An advanced cybertraining program designed to enhance research on testbed systems. In *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration* (pp. 1-5).
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR)
3. Ohana, R., McCabe, M., Meyer, L., Morel, R., Agocs, F., Beneitez, M., ... & Ho, S. (2024). The well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, *37*, 44989-45037.
4. Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research.
5. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.
6. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., ... & El-Nouby, A. (2025). Multimodal autoregressive pre-training of large vision encoders. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 9641-9654).

This work was made possible thanks to the ByteBoost cybertraining program which is funded by the National Science Foundation Cybertraining awards: 2320990, 2320991, and 2320992, and the Neocortex project, the ACES platform, and the Ookami cluster.

The Neocortex project is supported by National Science Foundation award number 2005597.

The ACES (Accelerating Computing for Emerging Sciences) platform was funded by National Science Foundation award number 2112356.

The Ookami cluster is supported by National Science Foundation award number 1927880.