Oak Ridge National Laboratory

Aug 5, 2025

Advancing Scalable and Trustworthy Foundational Models for Science at Oak Ridge National Laboratory

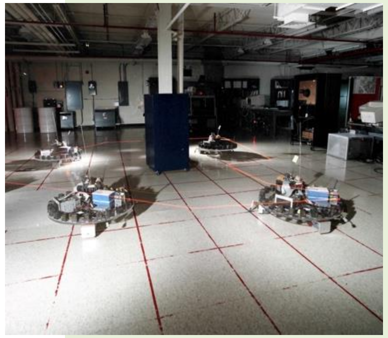
Prasanna Balaprakash

U.S. DEPARTMENT OF ENERGY

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY

ORNL has a rich history leveraging AI for science

1979
Oak Ridge Applied Artificial Intelligence Project



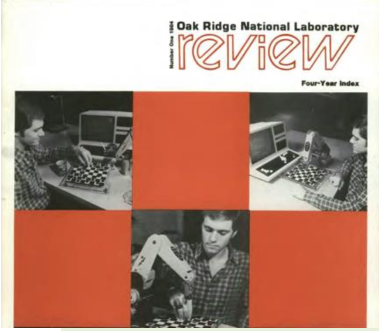
1991
Automated machines

Current
Frontier

- #2 HPL-MxP @10 exaflops for AI
- Scaled to 1T+ parameter AI model training

1980 1990 2000

1981
AI infrastructure supports spectroscopy, environmental management, nuclear fuel reprocessing, and programming assistance



1983
Robotics

2012
Titan:
First GPU-powered supercomputer

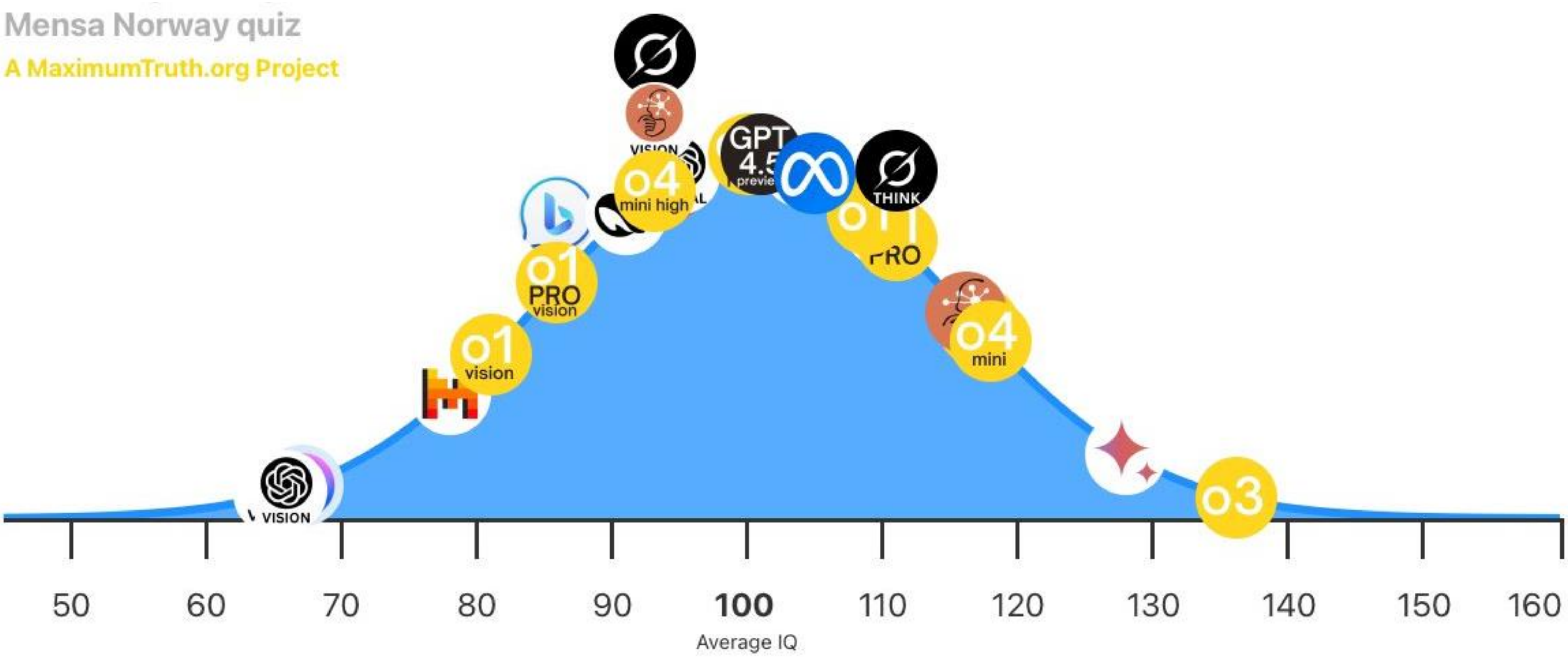
Oak Ridge National Laboratory

AI transforming science and national security

ORNL facilities, expertise enable AI revolution

OAK RIDGE National Laboratory

What is industry doing?



The Gentle Singularity — Sam Altman, 2025

- **The Takeoff Has Begun**
  - Humanity has crossed the AI event horizon.
  - AI systems (e.g. GPT-4, o3) already outperform humans in key areas.
  - Scientific breakthroughs behind them were the hardest part—momentum is now self-reinforcing.
- **AI's Transformative Impact**
  - 2025: Agents doing real cognitive work (coding, scientific discovery).
  - By 2027: Robots may enter the physical world.
  - Productivity, creativity, and research velocity are surging—scientists report 2–3X output boosts.
- **Self-Reinforcing Progress Loops**
  - Recursive AI research: AI used to build better AI.
  - Economic flywheel: AI → Value → Infrastructure → More AI.
  - Automation of datacenters, supply chains, and robot manufacturing looms ahead.
- **Toward Abundant Intelligence & Energy**
  - Intelligence cost converging to energy cost.
  - Exponential gains may redefine what we consider "real work" or "progress."
  - Wonders become routine—and then table stakes.

Zuckerberg said the company plans on investing "hundreds of billions of dollars" to power AI

GE VERNOVA

Businesses News Investors Sustainability Company

PRESS RELEASE • NUCLEAR POWER

Tennessee Valley Authority submits application for construction of first BWRX-300 small modular reactor in the U.S.

• 5 min read

KNOXVILLE, Tenn. (May 20, 2025) – Tennessee Valley Authority (TVA) has submitted an application to the U.S. Nuclear Regulatory Commission to construct a GE Vernova Hitachi Nuclear Energy (GVH) BWRX-300 small modular reactor (SMR) at the Clinch River site in Oak Ridge, Tennessee. It is the first construction permit application for a BWRX-300 in the U.S.

U.S. DEPARTMENT of ENERGY

About Our Mission New Horizons Topics Consumer Savings Services & Opportunities

Home › Topics › Unleash American Energy Innovation › DOE Announces Site Selection for AI Data Center and Energy Infrastructure Development on Federal Lands

DOE Announces Site Selection for AI Data Center and Energy Infrastructure Development on Federal Lands

The forthcoming solicitations will drive innovation in reliable energy technologies, contribute to lower energy costs, and strengthen American leadership in artificial intelligence.

Energy.gov

July 24, 2025

Scientific AI demand richer world models

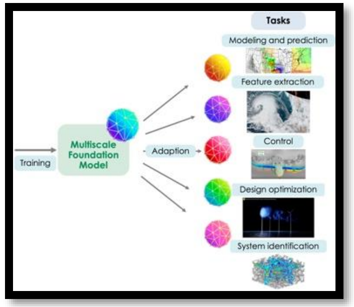
LLMs Today
- Trained on ~2.0E13 text tokens (~6.0E13 bytes) on static data
- Reading equivalent would take a human 300,000 year
- Text input along often lacks physical grounding

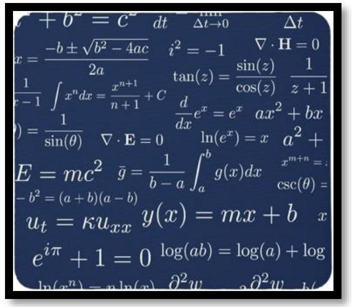
4 years old child
- 16K waking hours in 4 years
- 1.1E14 bytes of multimodal, real-time data
- Vision, touch, language, causality

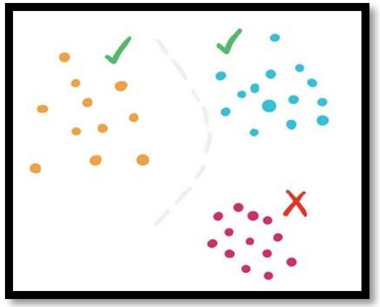
Even a child's learning is vastly more multimodal than today's LLMs.
Science demands even more!!

Yann LeCun, Josiah Willard Gibbs Lecture, the 2025 Joint Mathematics Meetings on "Mathematical Obstacles on the Way to Human-Level AI", 2025

How can we capitalize?







Exascale and AI Integration

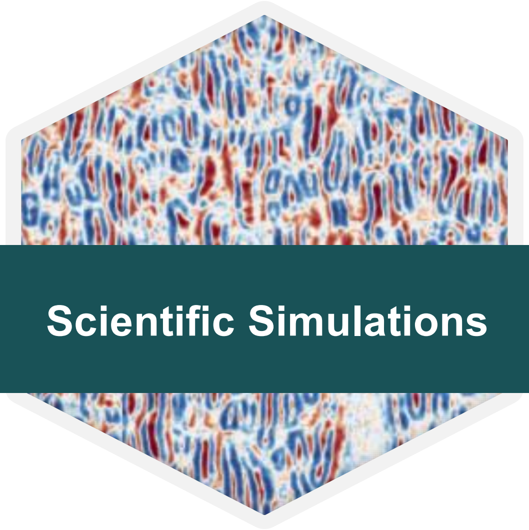
Multimodal AI Models

Domain-Knowledge

Robust Validation Frameworks

Secure, Usable AI Infrastructure

ORNL's AI initiative
Secure, assured, and efficient



Scientific Simulations

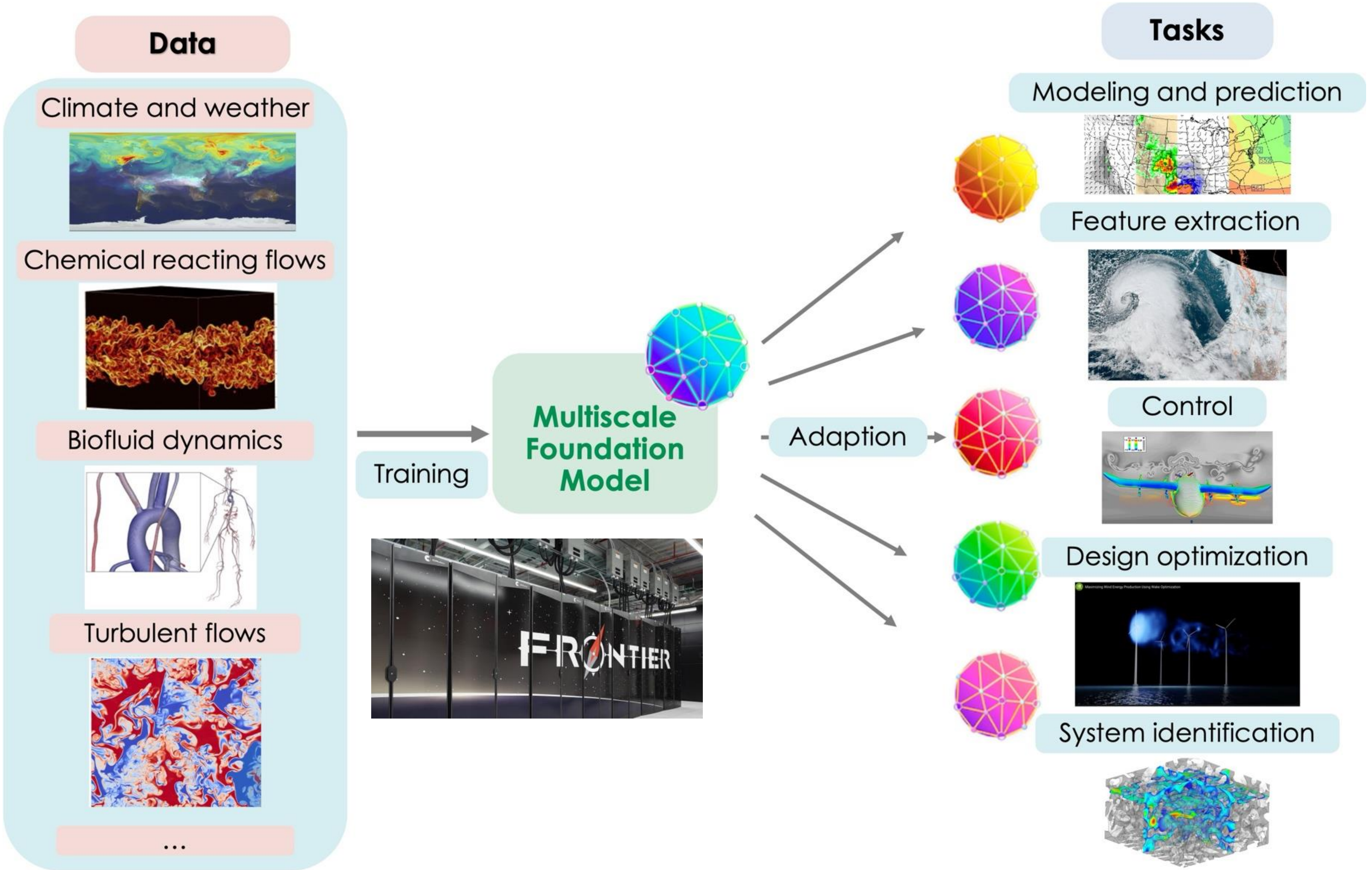
Experimental Facilities

National Security

Secure	Assurance	Efficient
Alignment	Validation and verification	Scalability
Cybersecurity	Uncertainty qualifications	Edge
Robustness	Causal reasoning	Novel AI hardware

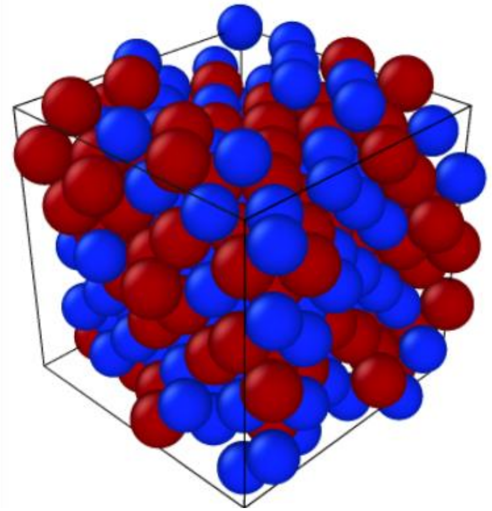
The initiative's portfolio comprises 15 advanced AI projects and involves over 50 researchers from 5 different directorates across the lab.

Foundation AI model(s) for science

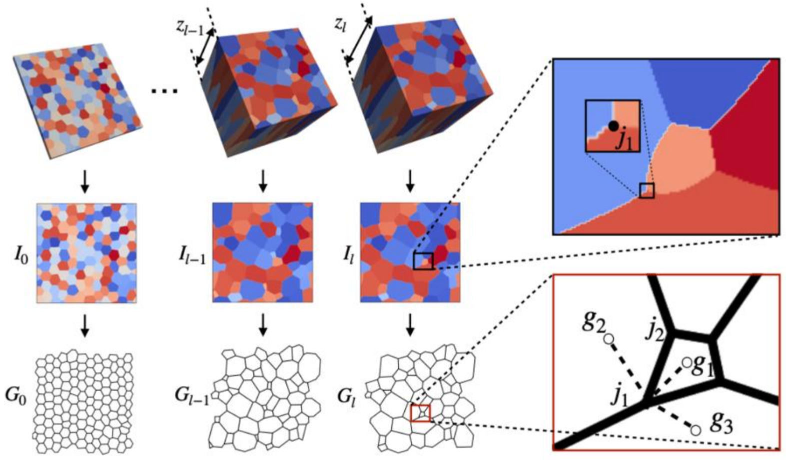


Graph representation of materials at different scales

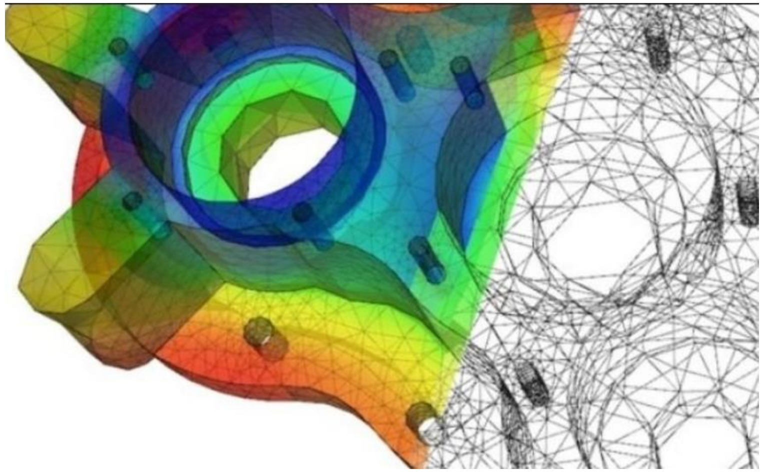
Atomistic scale



Mesoscale



Continuum scale



Nodes = atoms

Edges = interatomic bonds

Nodes = Voronoi centers

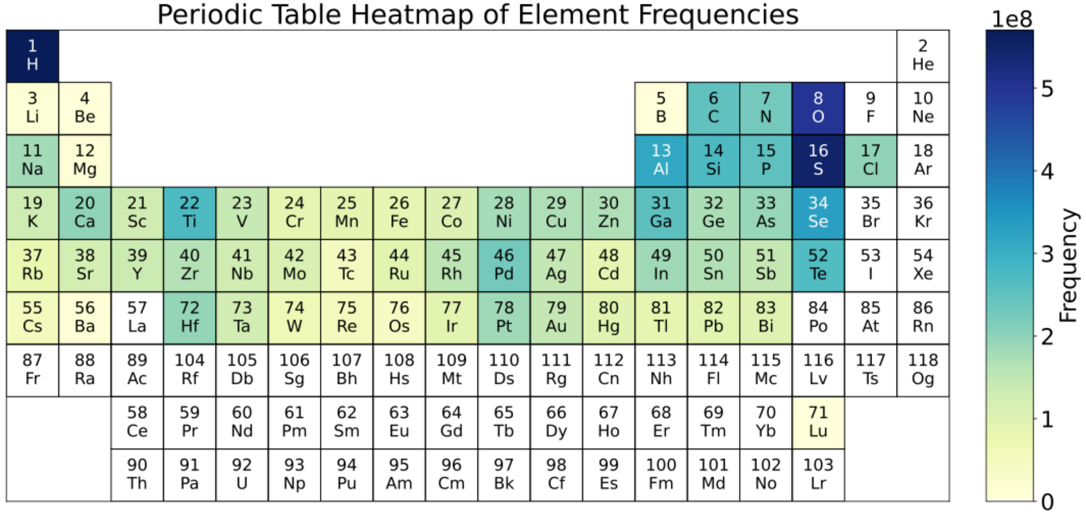
Edges = connection between Voronoi centers

Nodes = vertices of the finite element mesh

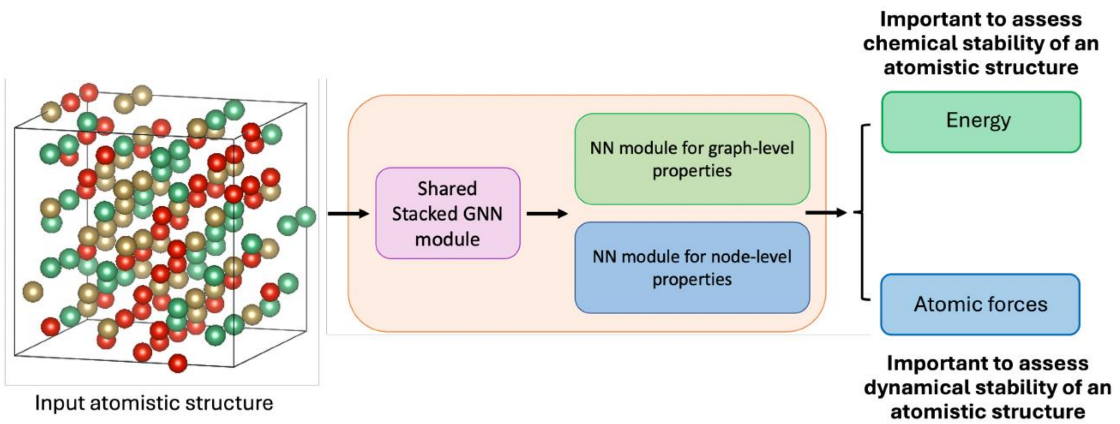
Edges = edges of the finite element mesh

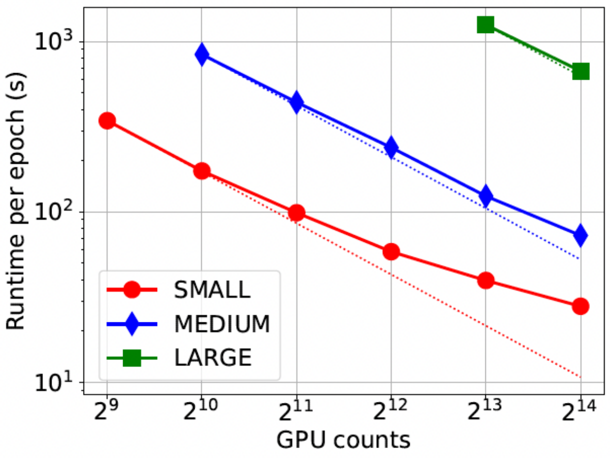
Graph structured data maps naturally onto graph neural networks

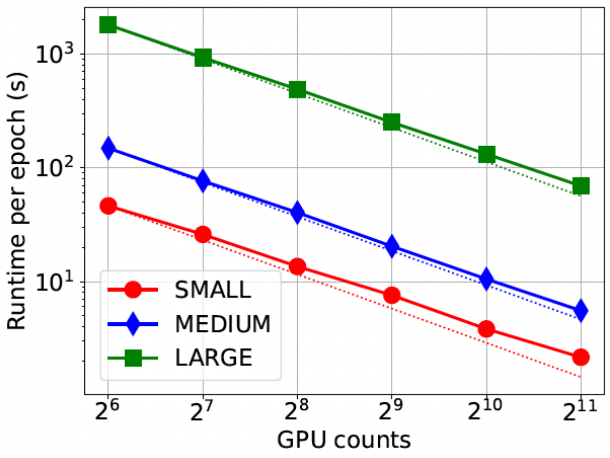
Scalable training of graph foundation model for materials science



Dataset	Number of data samples	Size
ANI1x [63]	4,956,005	5.3 GB
QM7-X [64]	4,195,237	23 GB
OC2020 [39]	134,929,018	4.3 TB
OC2022 [40]	8,847,031	648 GB
MPTrj [37]	1,580,395	17 GB
Total	154,507,686	5.2 TB

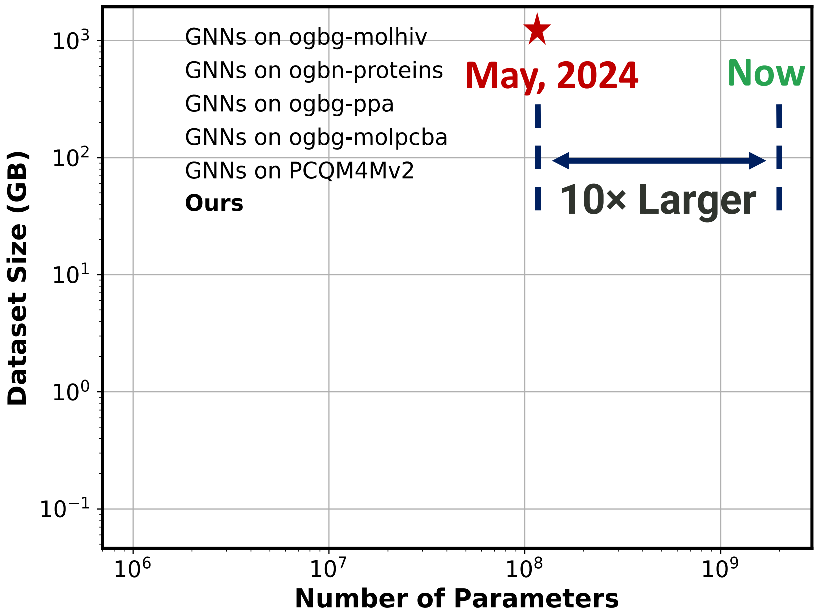






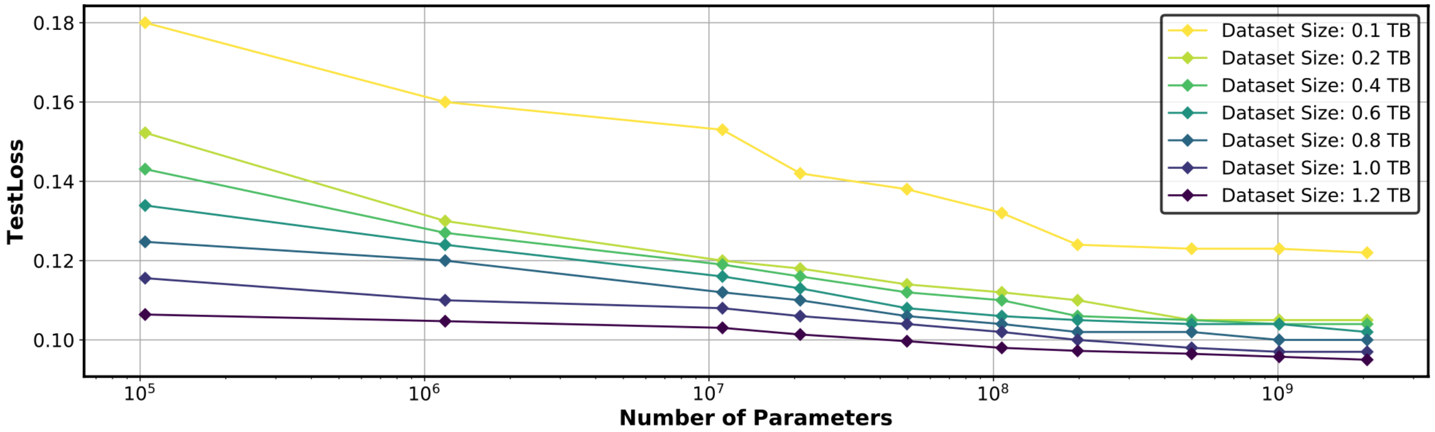
Strong scaling of HydraGNN multitasking pre-training on a problem of 120 million graphs on Frontier 16K GPUs and 2 million graphs on Perlmutter 2K GPUs with three GNN model sizes.

Scaling laws for GNNs

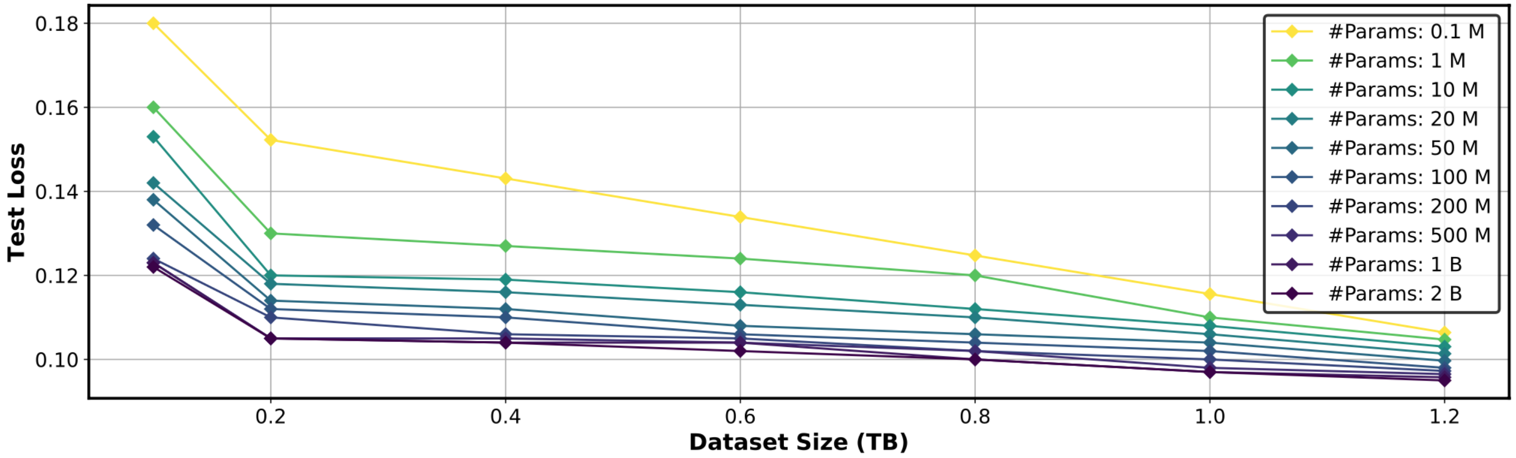


HydraGNN with 2 billion (10× larger than previous state-of-the-art) parameters on 1.2 TB of data.

Scaling law of GNN accuracy as a function of model size



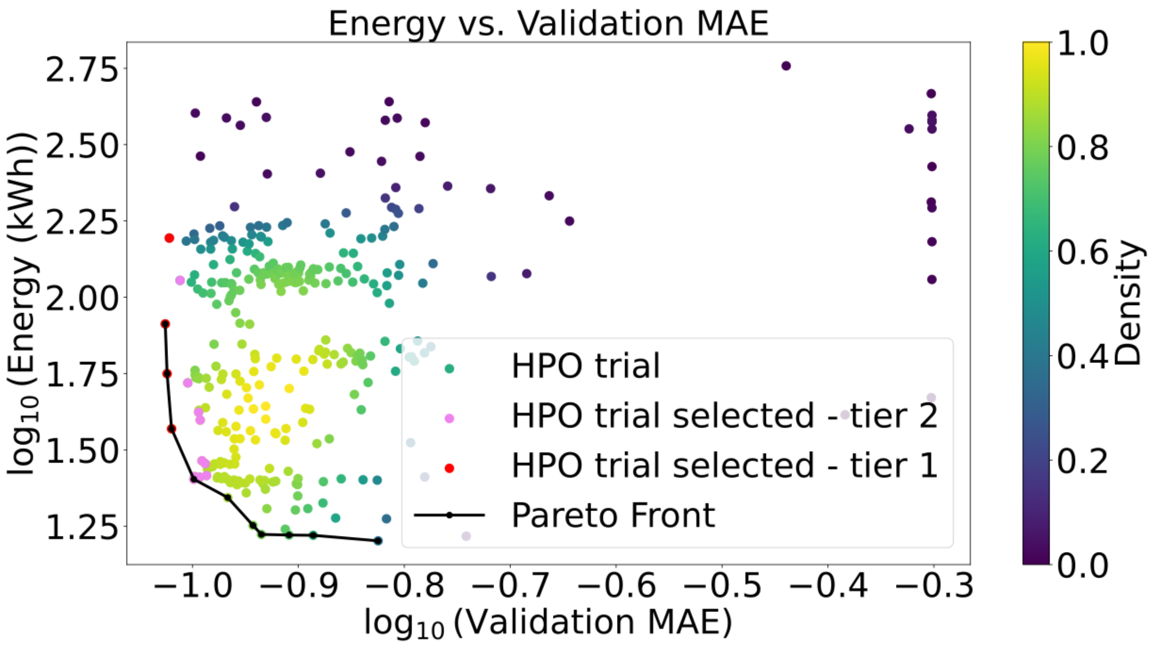
Scaling law of GNN accuracy as a function of data size

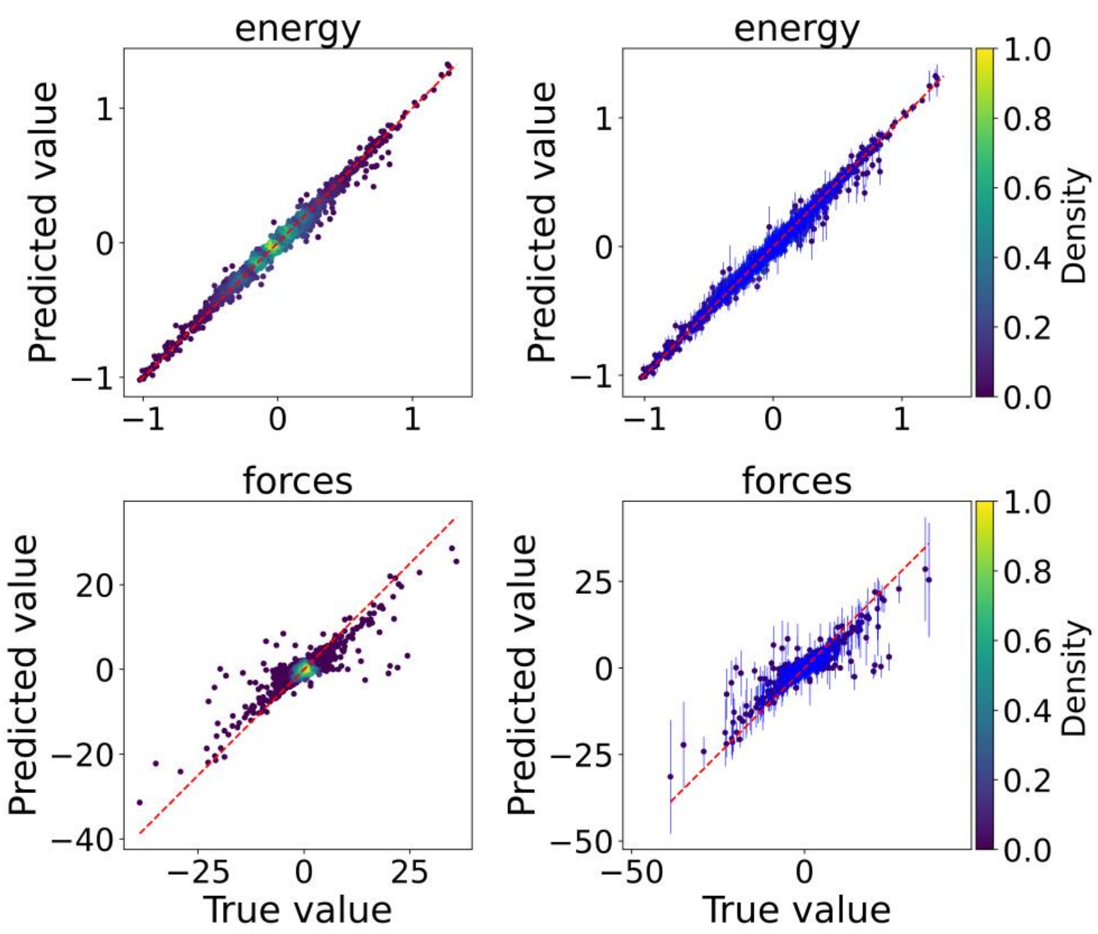


OAK RIDGE National Laboratory

Bi-objective optimization and ensemble uncertainty quantification

Hyperparameter	Type	Admissible values
# Type of MPNN layer	Categorical	{PNA, EGNN, SchNet}
# MPNN layers	Integer	{1,. . . ,6}
# neurons in MPNN layers	Integer	{100, . . . , 2,000}
# FC layers	Integer	{2,3}
# neurons in FC layers	Integer	{300, . . . , 1,000}
# batch size	Integer	{16, . . . , 128}





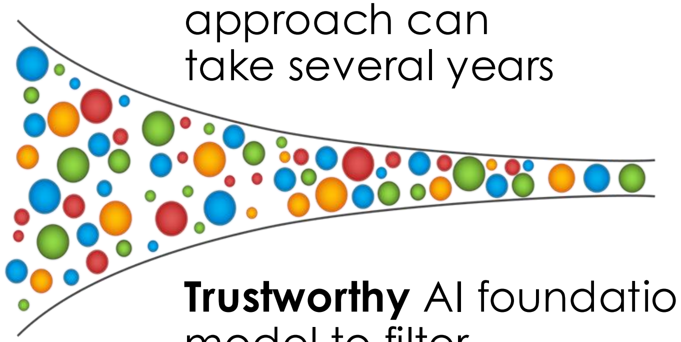
Oak Ridge National Laboratory

[Model weights: https://huggingface.co/mlupopa/HydraGNN_Predictive_GFM_2024](https://huggingface.co/mlupopa/HydraGNN_Predictive_GFM_2024)

Accelerated materials discovery via trustworthy AI models on Frontier

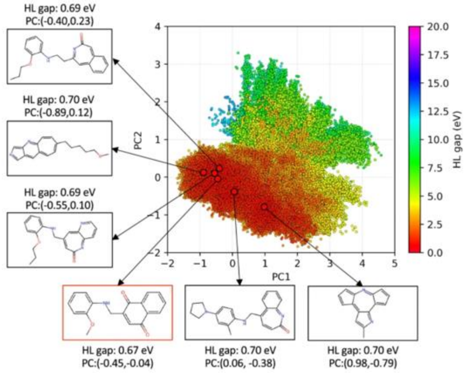
Billions of potential candidates

Traditional approach can take several years

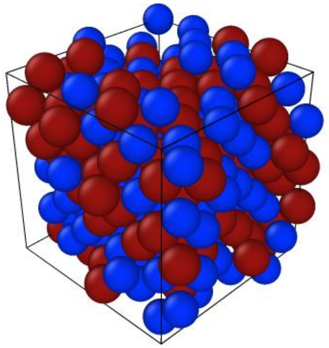


Trustworthy AI foundation model to filter in mins or hours

Tens or hundreds of promising candidates



Chromophores
Biomedical, MRIs, Quantum circuits

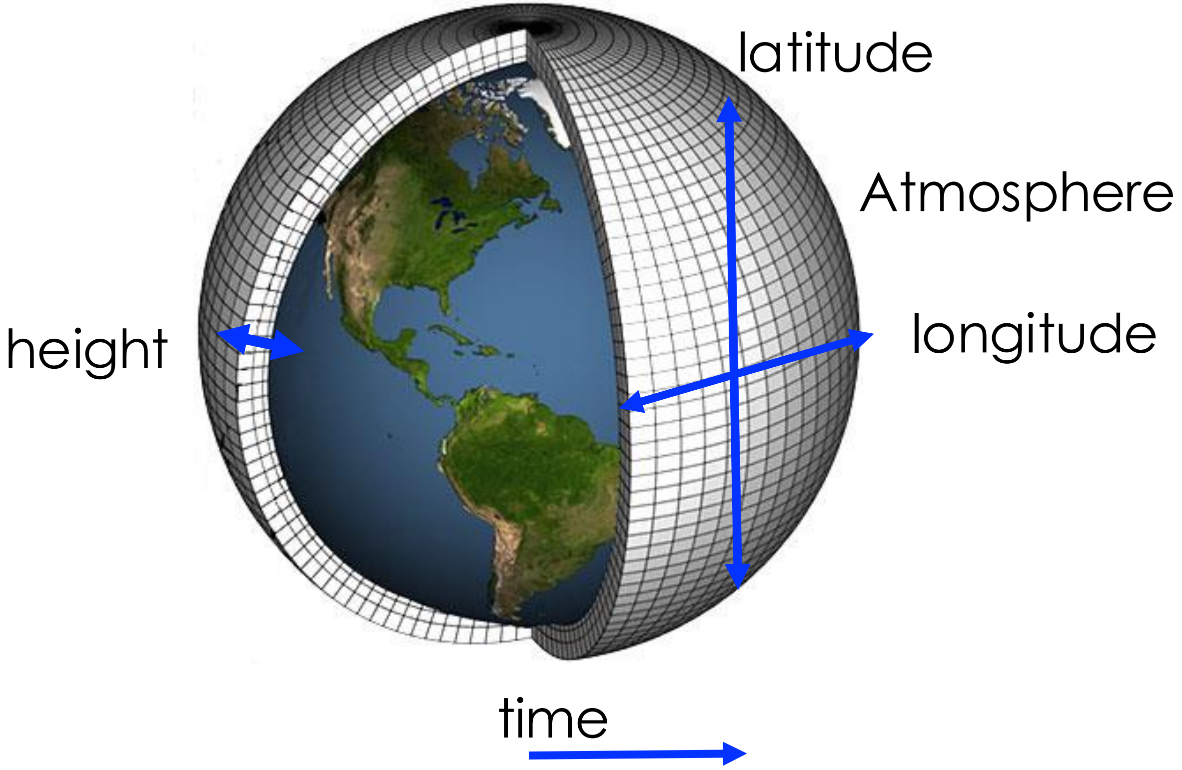


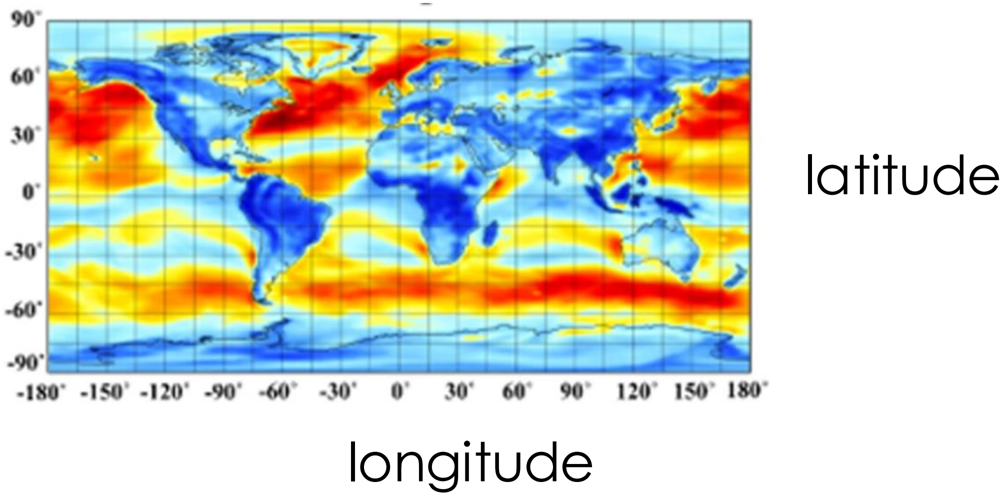
High-entropy alloy
Aerospace, mechanical manufacturing, biomedicine

Massimiliano Lupo Pasini and team

Spatiotemporal data

4D+X
longitude, latitude, time, (#height, #weather/climate variables) (91)





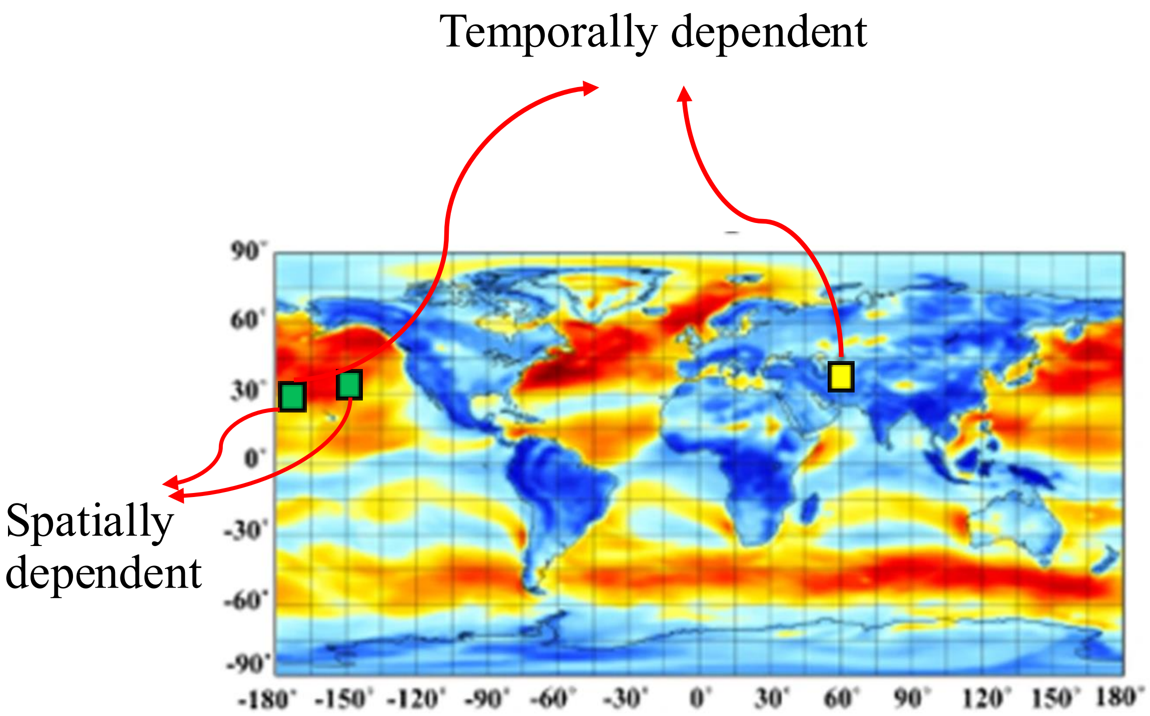
2D visualization for a weather variable at a fixed height and time point

Challenges for scaling up foundation models for spatiotemporal data

Data Challenge:
- 95D (4D+X) spatial-temporal data
- Non-linear memory and computing increase with resolution
- Complex spatial and temporal dependency

Model Challenge:
- Larger activations, parameters, gradients, optimization states
- Non-linear memory and computing increase

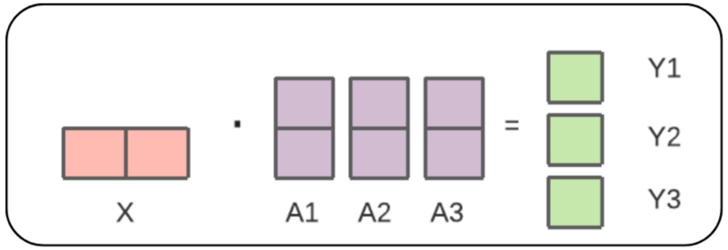
Industry solutions: Pipeline, Tensor, FSDP
- Not optimized for this modality
- Limited scalability



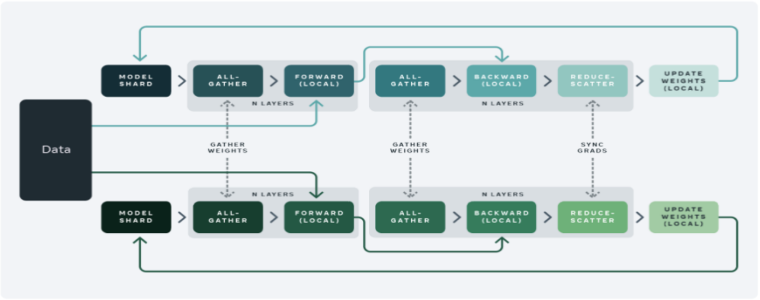
Implications:
(1) Much more expensive than LLM
(2) Small Vision Transformer Model Size
- largest dense ViT has _22 billion parameters_
- largest climate ViT AI model has _115 to 500 million_ parameters

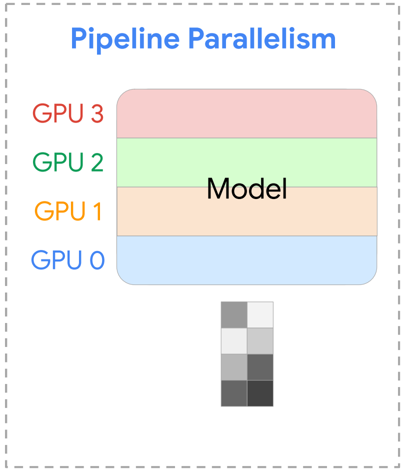
Hybrid sharded tensor-orthogonal parallelism

- Tensor parallelism to ensure that compute-heavy layers don't bottleneck.
  - Dividing a single matrix multiplication across GPUs.
- Fully Sharded Data Parallelism to eliminate replicas
  - Every GPU holds just a slice of the model's weights
- Pipeline Parallelism to minimize GPUs idleness
  - the model is sliced into stages. While GPU 1 is working on Layer 1 of Batch A, GPU 2 is crunching Layer 2 of Batch B

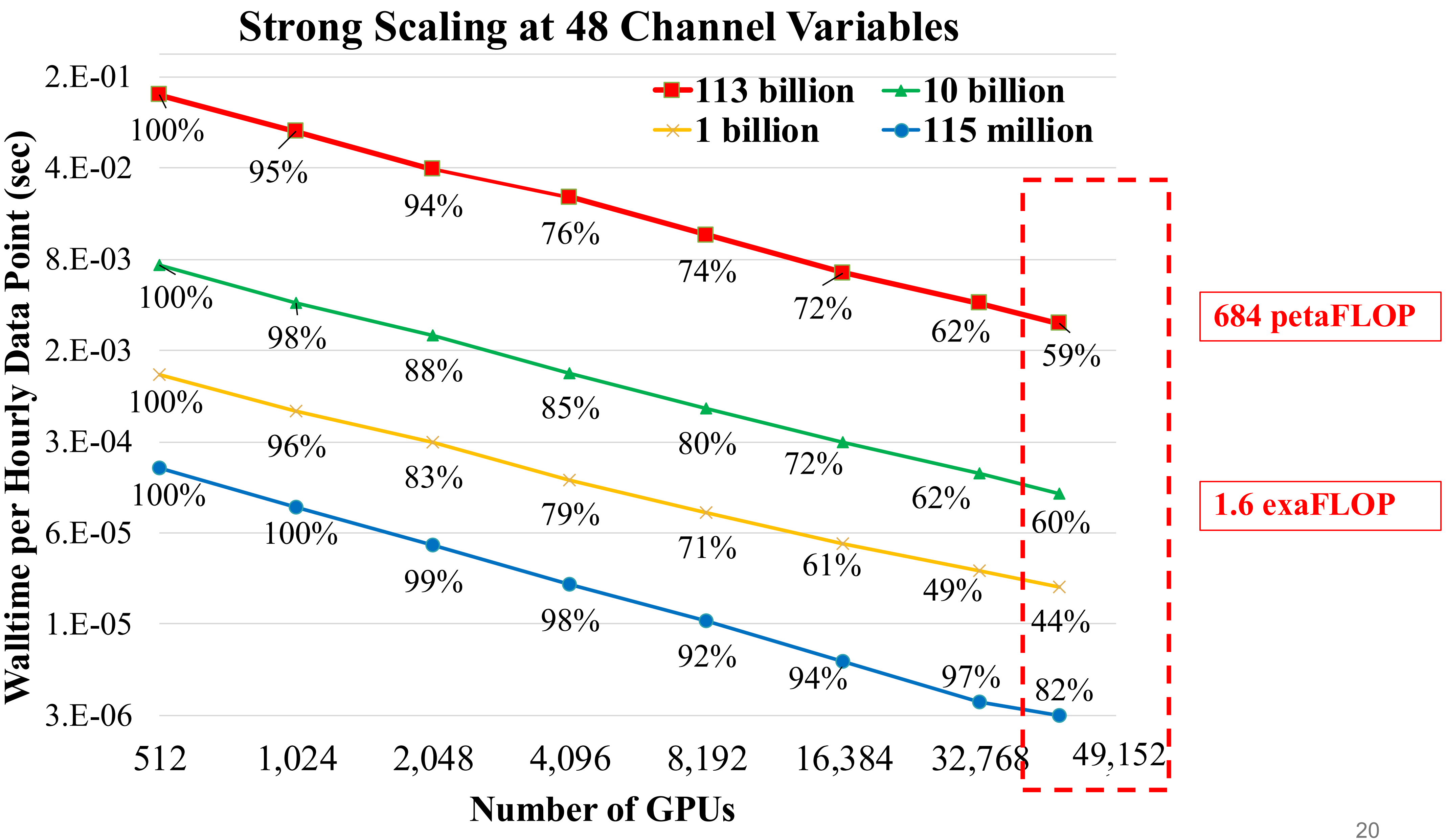


Fully sharded data parallel training





Pretraining achieves ExaFLOP throughput with CMIP



OAK RIDGE National Laboratory

ORBIT inferencing enables near real-time prediction

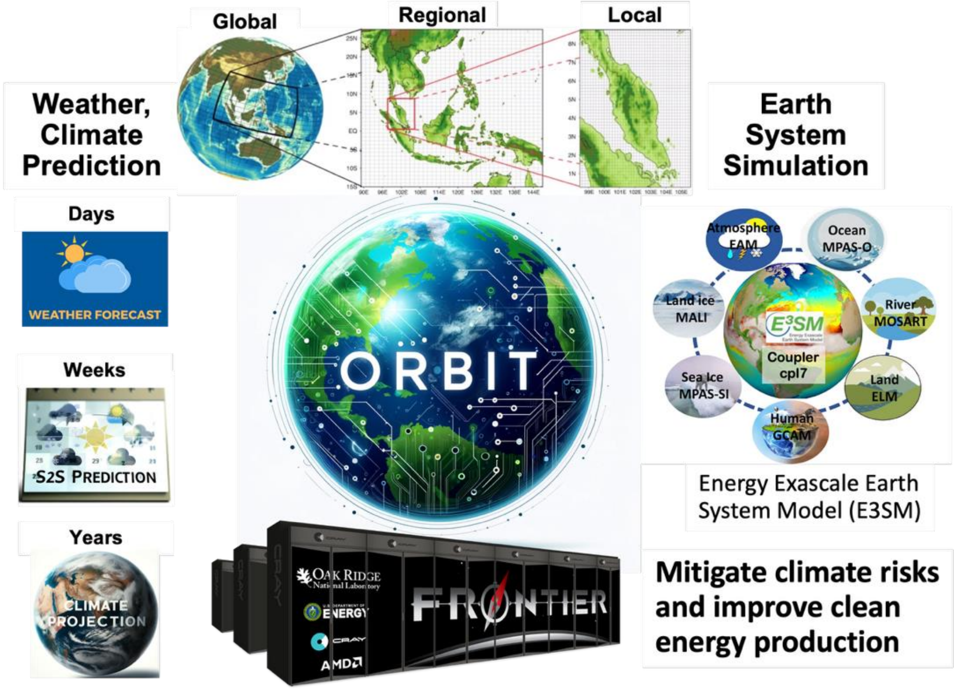
Model Size	115 million	1 billion	10 billion	113 billion
GPUs	1 GPU	1 GPU	8 GPUs	80 GPUs
Inference Speed (sec)	0.04 sec	0.24 sec	0.16 sec	0.5 sec

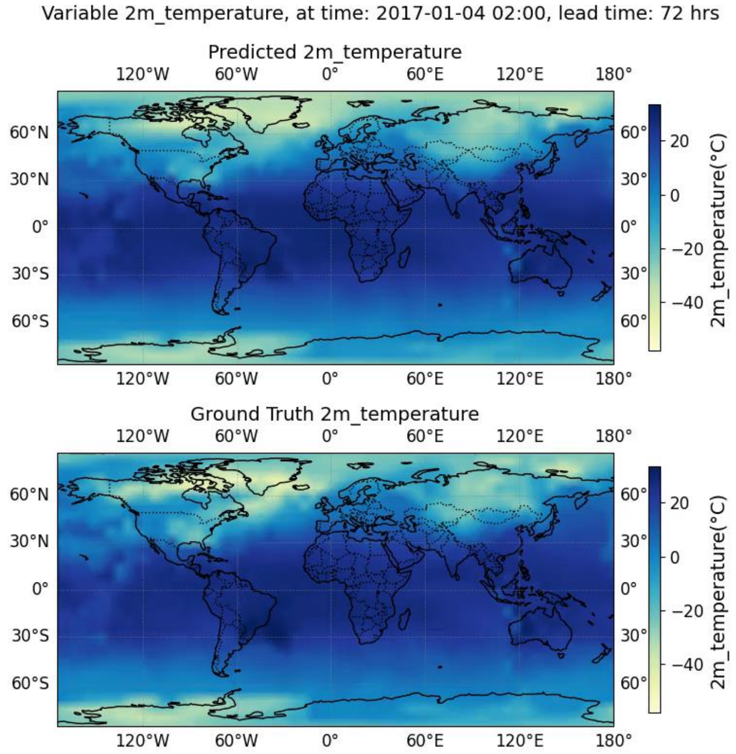
E3SM Atmosphere Model
1.26 simulated year per day on Frontier supercomputer

Potential for 500x speedup on limited resource

Oak Ridge National Laboratory

ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability

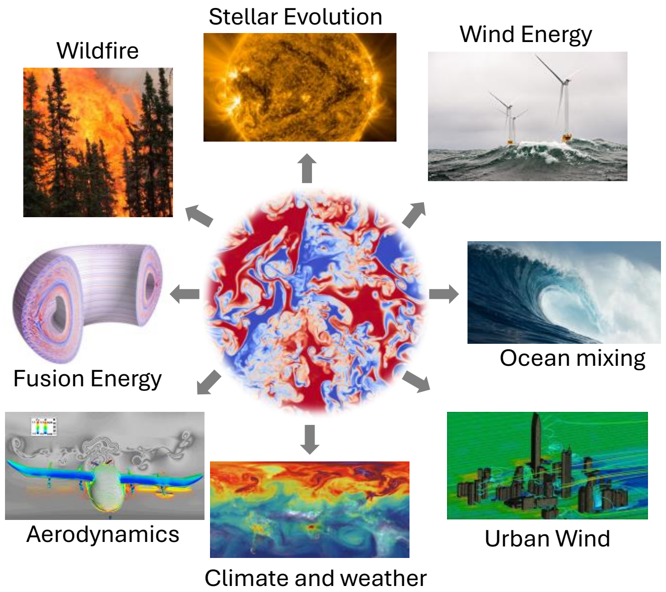




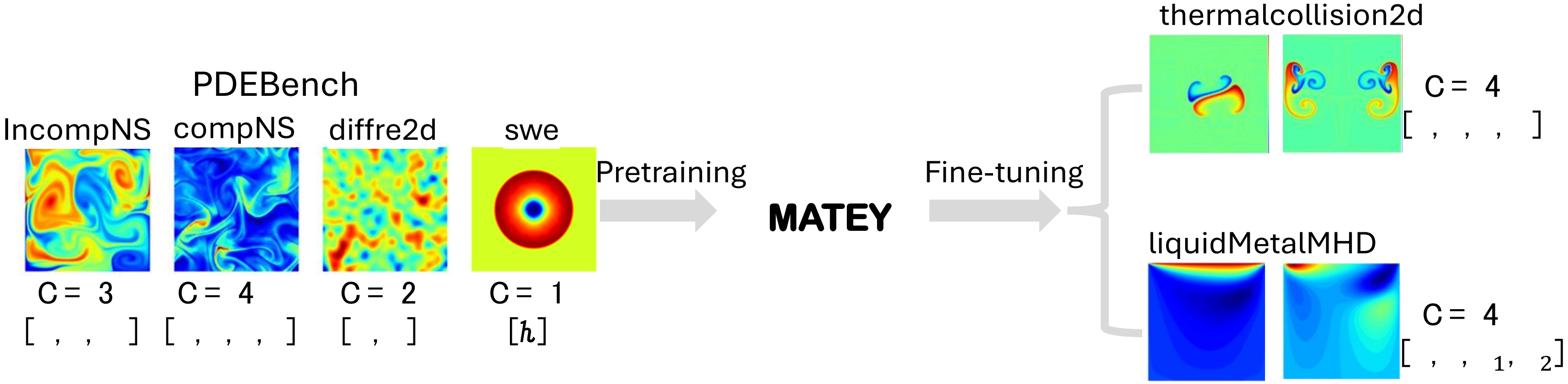
- Developed AI foundation model (FM), pretrained on CMIP6 model simulation data and adaptable to various Earth system modeling tasks.
- Using 49,152 GPUs on 6,144 Frontier nodes, ORBIT achieves 70% scaling efficiency with a computing throughput of 1.6 exaflops (finalist for the 2024 and 2025 Gordon Bell Prize for Climate Modelling; 2025 SC best paper nomination).
- ORBIT-2 achieves competitive or better accuracy for super resolution task


Xiao Wang, Dan Lu, and team

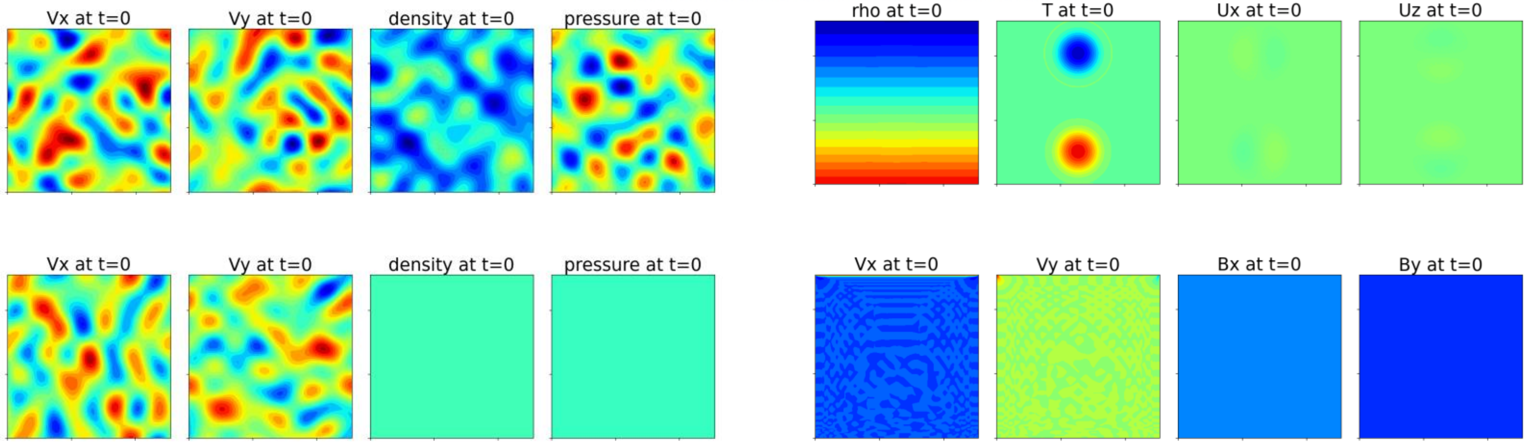
MATEY: multiscale adaptive foundation models for spatiotemporal physical systems



Diverse applications characterized by the same core physics: turbulence



$$\mathbf{u}_{t+t_{\text{lead}}} \approx \mathbf{f}_{\mathbf{w}}\left(\underbrace{\mathbf{u}_{t-T+1}, \ldots, \mathbf{u}_t}_{\mathbf{U}_k \in \mathbb{R}^{T \times H \times W \times C_k}}; t_{\text{lead}}\right)$$



Oak Ridge National Laboratory

Pei Zhang and team

The DOE American Science Cloud

OBBB also includes forward-looking provisions that position government data as a strategic asset for American competitiveness:

- ***The American Science Cloud:*** The Department of Energy's $150 million investment aims to mobilize National Laboratories to structure and preprocess scientific data for AI and machine learning applications. Importantly, the legislation requires these models to be made available to the broader scientific community through a cloud-based platform—embodying the principles of open data and collaborative innovation.

AI-driven autonomy is reshaping the scientific workflow

Labs of the future: Interconnected network of multi-domain research hubs to drive new investigative approaches that **combine human creativity with evolving artificial intelligence (AI)**

Scientist Controlled Flexible Automation and Autonomy

Self-optimizing protocols, digital twins, cross knowledge integration

Multi-Domain research hubs cut across user facilities and labs

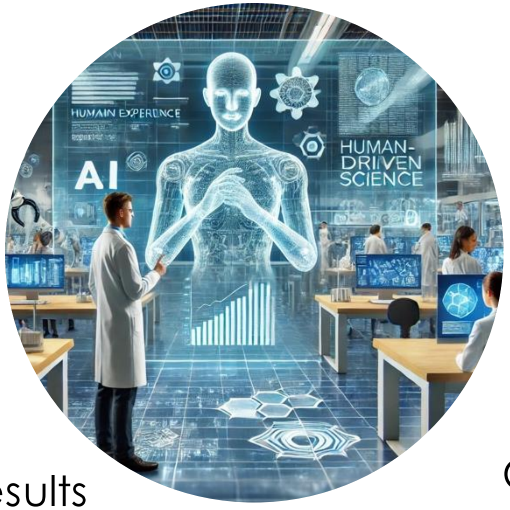
Generate **hypotheses**

Design Experiments

Conduct **experiments across labs**

Analyze results in real time with **AI, theory, and simulation**

Iteratively repeat



Interconnected AI tools, mod-sim, theory, and analysis

Autonomous Collection and Curation of Data

Interoperable AI Driven Workflows Across Directorates

AI Agent drive autonomous decisions
Robotic Platforms automate experiments
Humans in/out of the workflow loop