OAK RIDGE National Laboratory

ORNL's AI Initiative: Advancing Secure, Trustworthy, and Energy-Efficient AI for Science

Massimiliano (Max) Lupo-Pasini

Prasanna Balaprakash

Oak Ridge National Laboratory

September 10, 2024

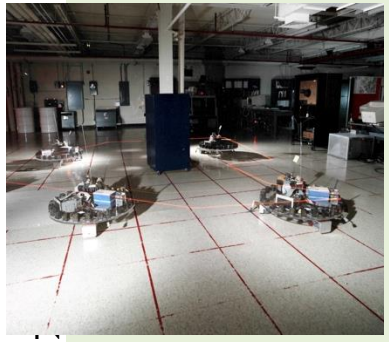
15th International Conference on Parallel Processing & Applied Mathematics

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

ORNL has a rich history leveraging AI for science

1979
Oak Ridge Applied Artificial Intelligence Project



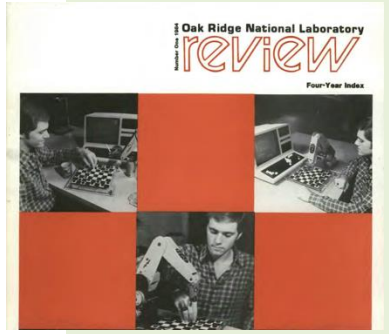
1991
Automated machines

Current
Frontier

- #1 HPL-MxP @10 exaflops for AI
- #2 on GREEN500
- 2021 ACM A.M. Turing Award
- Scaled to 1T+ parameter AI model training

1940–1970 | 1980 | 1990 | 2000

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



1983
Robotics

2017
Summit: World's "smartest" supercomputer optimized for AI

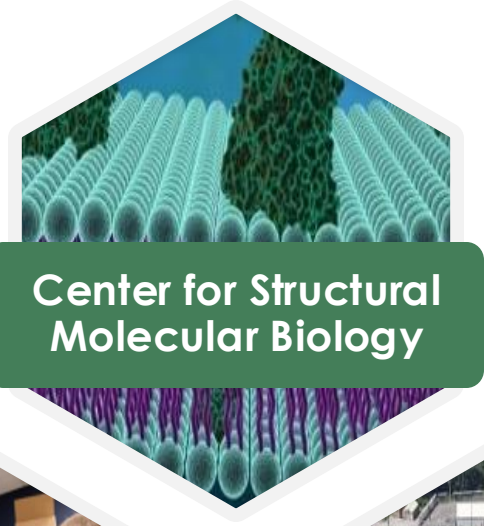
OAK RIDGE National Laboratory

AI transforming science and national security

ORNL facilities, expertise enable AI revolution

Spallation Neutron Source

Manufacturing Demonstration Facility



Center for Structural Molecular Biology

Oak Ridge Leadership Computing Facility

Cyber Science Research Facility

High Flux Isotope Reactor

Grand challenges in AI

The promise of AI is challenged by serious concerns about:

- Safety
- Security
- Trustworthiness
- Energy consumption

Comment | Published: 07 March 2022

Dual use of artificial-intelligence-powered drug discovery

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins

Nature Machin

120k Access

An internati
technologie
weapons. A

Forbes

FORBES > INNOVATION

10 Ways Cybercriminals Can Abuse Large Language Models

Mich
Forb
COU

AI fakes raise election risks as lawmakers and tech companies scramble to catch up

FEBRUARY 8, 2024 · 5:00 AM ET

HEARD ON MORNING EDITION

Shannon Bond

4-Minute Listen + PLAYLIST

SEARCH FORTUNE SIGN IN Subscribe Now

Home News Tech Finance Leadership Well Recommends Fortune 500

TECH · BRAINSTORM AI

AI could gobble up a quarter of all electricity in the U.S. by 2030 if it doesn't break its energy addiction, says Arm Holdings exec

OAK RIDGE National Laboratory

Presidential AI executive order

OCTOBER 30, 2023

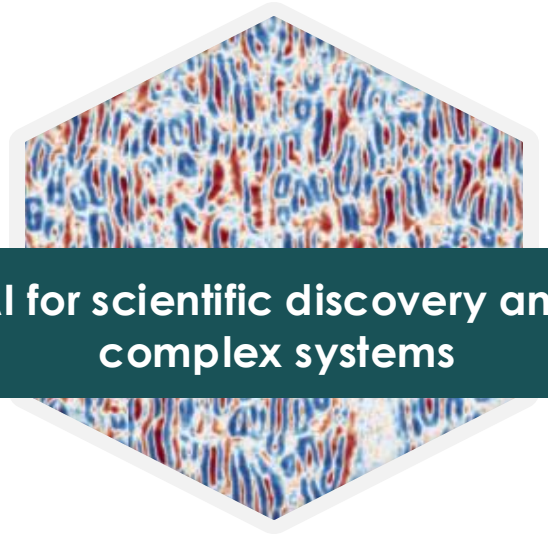
Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM ▸ PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

ORNL's AI initiative

Secure, trustworthy, and energy-efficient AI



AI for scientific discovery and complex systems

AI for experimental facilities

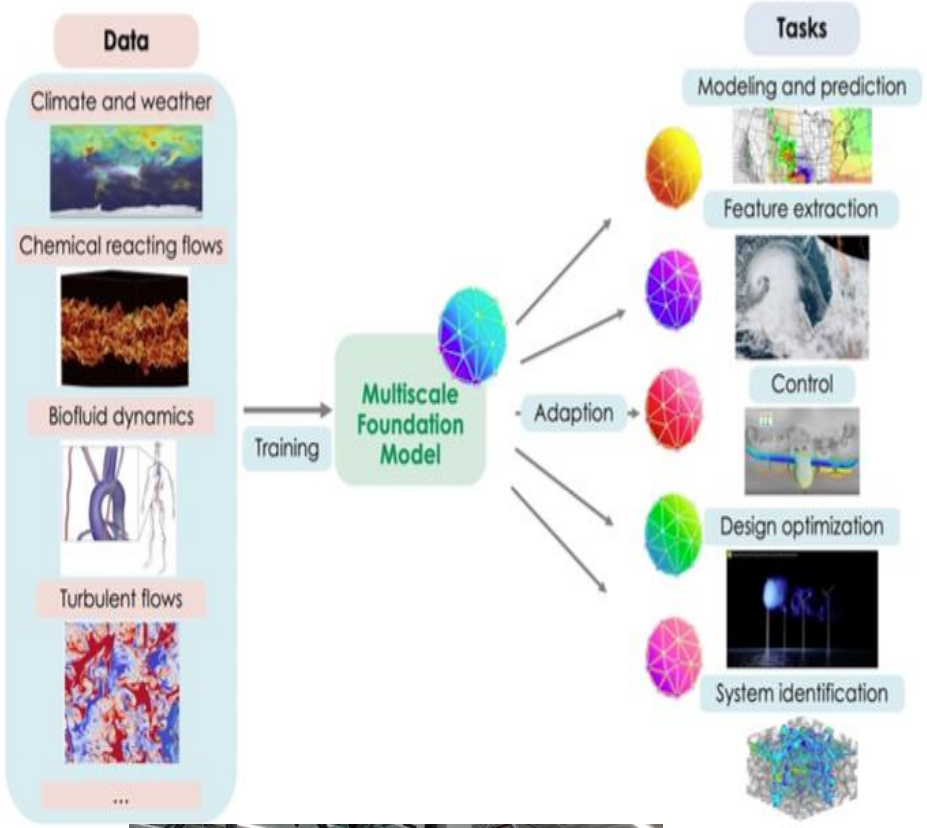
AI for national security

Secure	Trustworthy	Energy efficient
Alignment	Validation and verification	Scalability
Cybersecurity	Uncertainty qualifications	Edge
Robustness	Casual reasoning	Co-design

Foundation AI model(s) for science

Scientific & Engineering Datasets

Materials
Chemistry
Computer Science
Climate
Fusion Energy
Accelerators
Reactors
Energy Systems
Manufacturing



Hypotheses Generation

Digital Twins

Inverse Design

Optimization

Accelerated Simulations

Autonomous Experiments

First open-source instantiation of a trillion-parameter model on Frontier

Since the model is too large to fit in one GPU's memory, we distributed it across multiple GPUs using multi-dimensional parallelization

- First time setting and on non-NVIDIA hardware in open science
- Achieved more than 80% efficiency (best use of hardware and thus energy efficient)
- Democratized recipe for the benefit of the scientific community

Frontier trained a ChatGPT-sized large language model with only 3,000 of its 37,888 Radeon GPUs — the world's fastest supercomputer blasts through one trillion parameter model with only 8 percent of its MI250X GPUs

News By Matthew Connatser published January 07, 2024

Now you're playing with AI power!

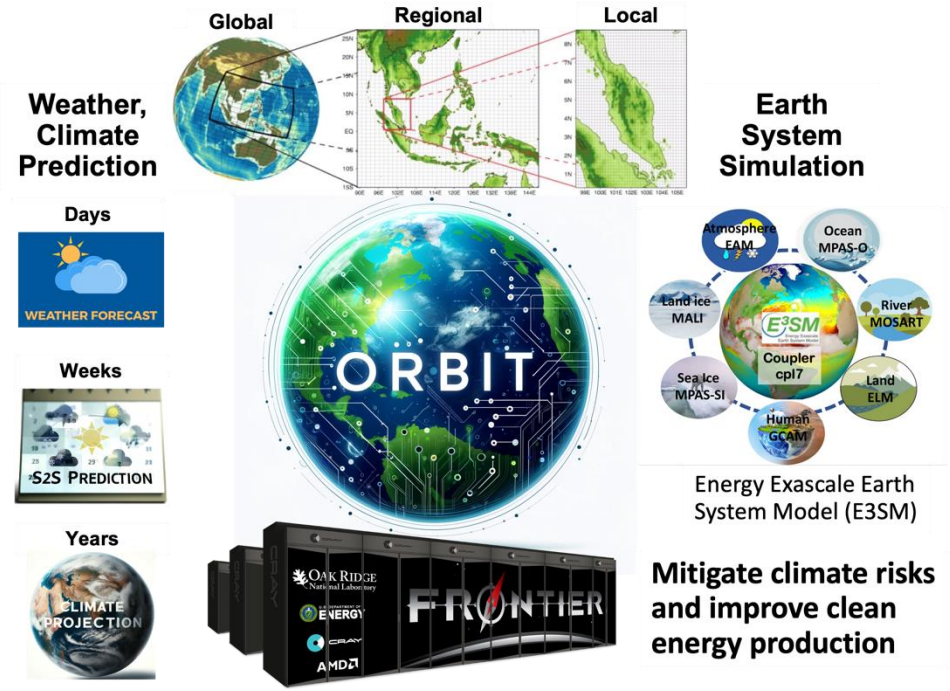
Comments (19)

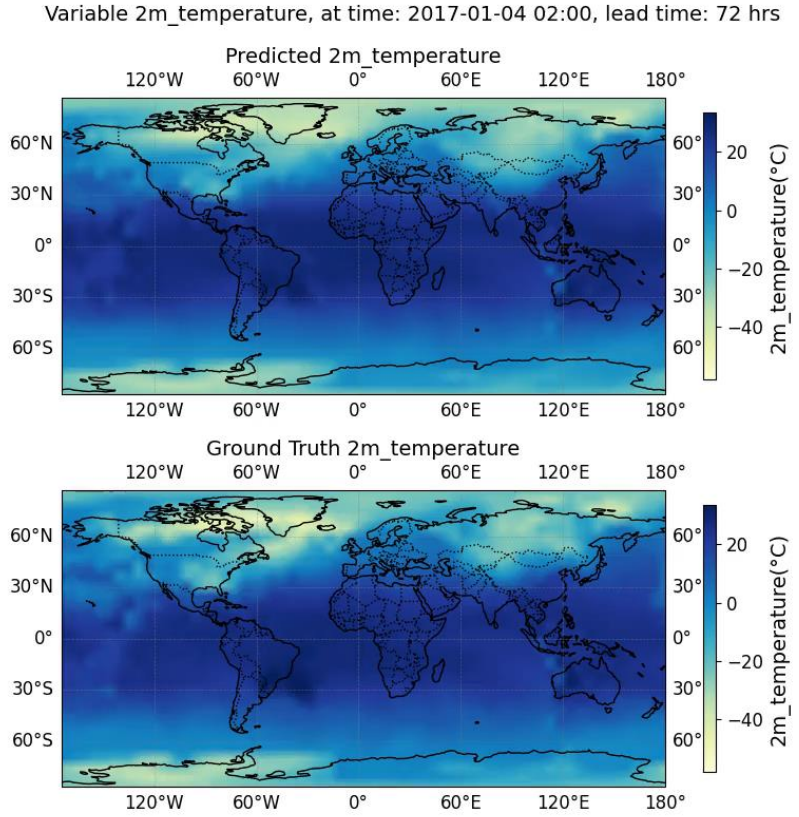
(Image credit: ORNL)

Researchers at Oak Ridge National Laboratory trained a large language model (LLM) the size of ChatGPT on the Frontier supercomputer and only needed 3,072 of its 37,888 GPUs to do it. The team published a research paper that details

OAK RIDGE National Laboratory

ORNL AI model for accurate weather and climate predictions



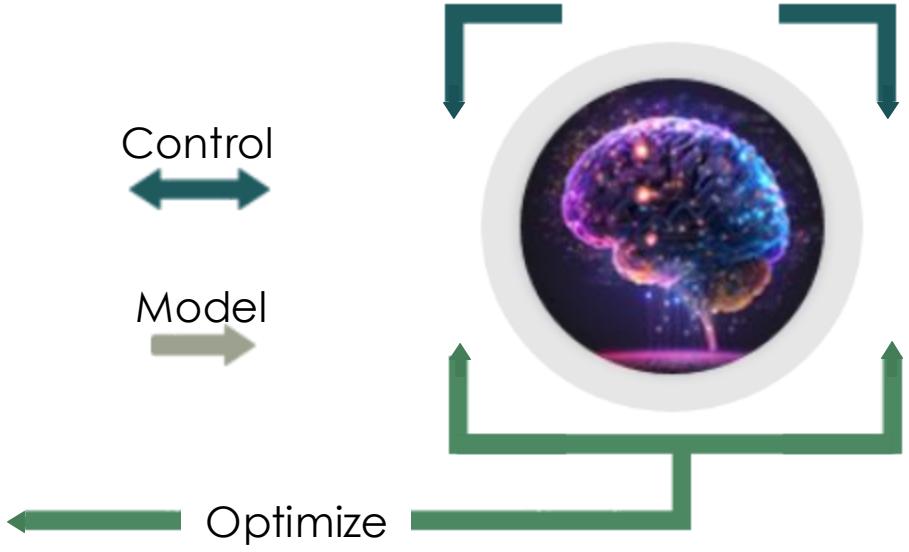


- ORBIT provides 72 hours weather forecasts
- ORBIT with 100 Billion parameters scales on 24K GPUs on Frontier supercomputer

AI will revolutionize climate science and safeguard the nation from climate risks

OAK RIDGE National Laboratory

ORNL AI model for robust and scalable decision making in complex dynamic systems



AI will transform the grid to be secure, resilient, and future-ready, leading to U.S. leadership in energy and climate risk solutions

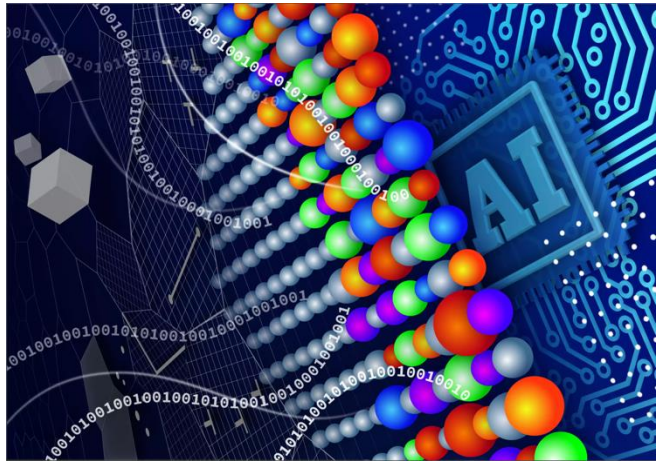
Accelerated materials science via AI models on Frontier

- **Characterization, discovery, and design of new materials** with desired properties and new chemical processes to control chemical reactions **is crucial to U.S. competitiveness and national security**, especially for energy generation, transportation, electronics, and information technology.

- **Labor-intensive experiments and/or computationally expensive first-principles calculations** are unaffordable for effective explorations of high-dimensional materials spaces.

- Once properly trained on experimental an/or simulation data, **AI** models provide faster, less expensive, and sufficiently accurate estimates of materials properties, thereby **enabling an accelerated exploration** of high-dimensional **materials spaces**.

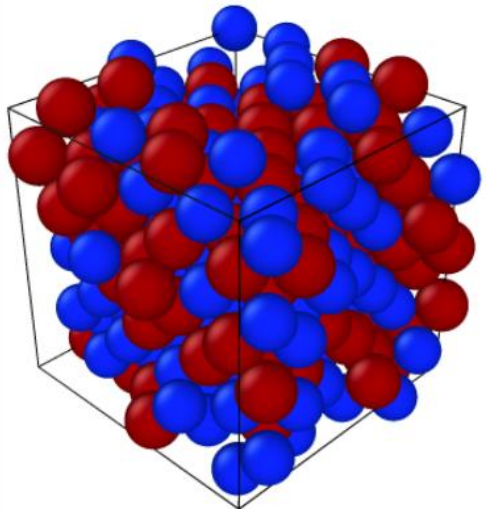
Picture from https://www.printedelectronicsnow.com/contents/view_breaking-news/2018-12-24/ornl-new-composite-advances-lignin-as-renewable-3d-printing-material/



Picture from https://www.mpg.de/20096180/artificial-intelligence-in-material-design

Graph representation of materials at different scales

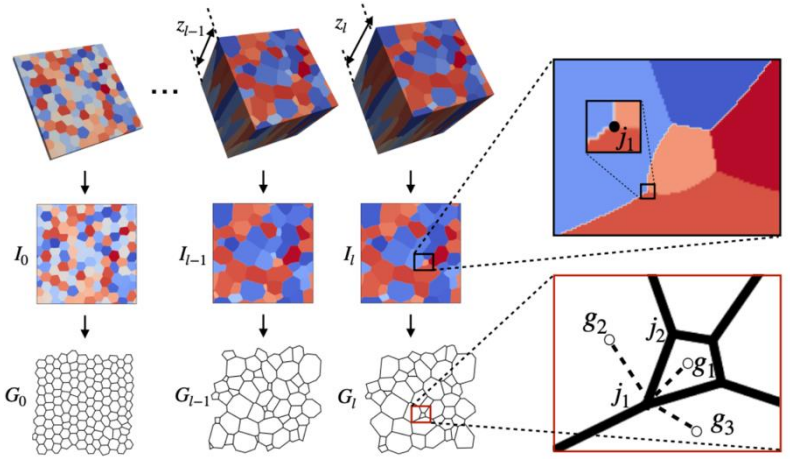
Atomistic scale



Nodes = atoms

Edges = interatomic bonds

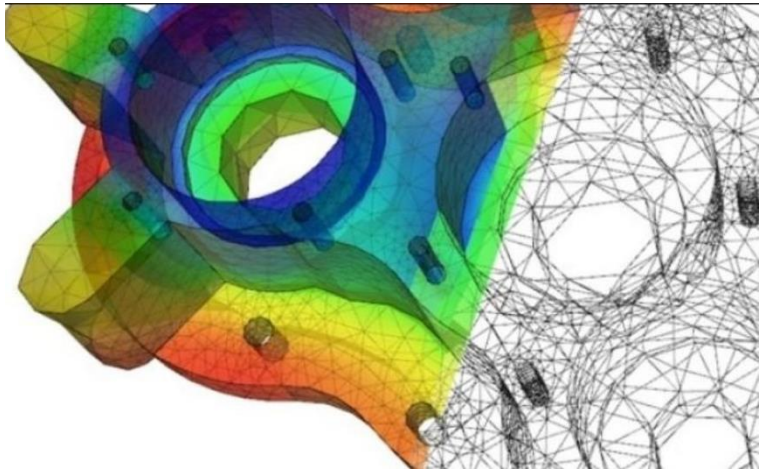
Mesoscale



Nodes = Voronoi centers

Edges = connection between Voronoi centers

Continuum scale



Nodes = vertices of the finite element mesh

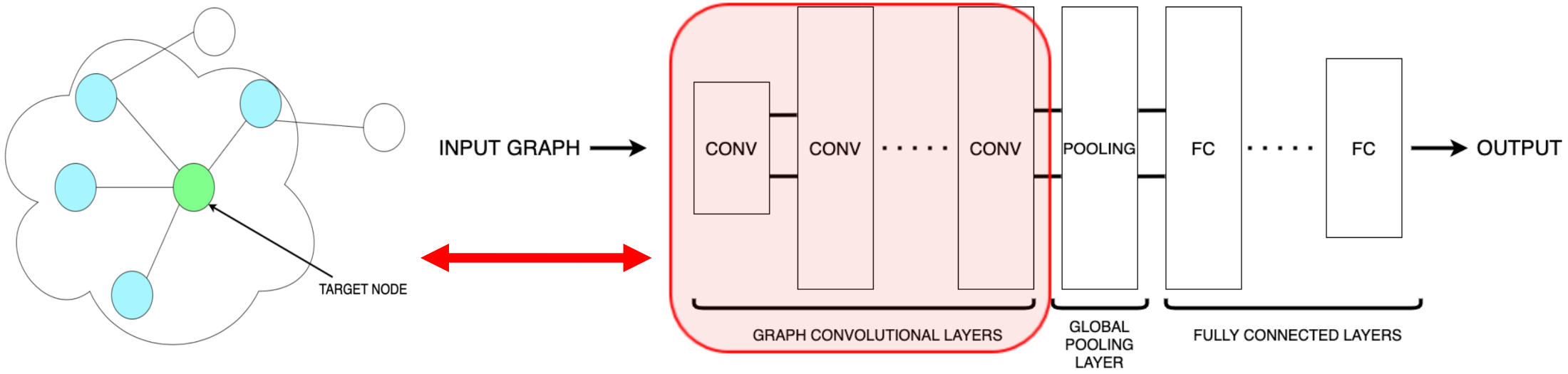
Edges = edges of the finite element mesh

Graph structured data maps naturally onto graph neural networks (GNNs)

Graph neural networks (GNNs)

The architecture of a GNN is made of:

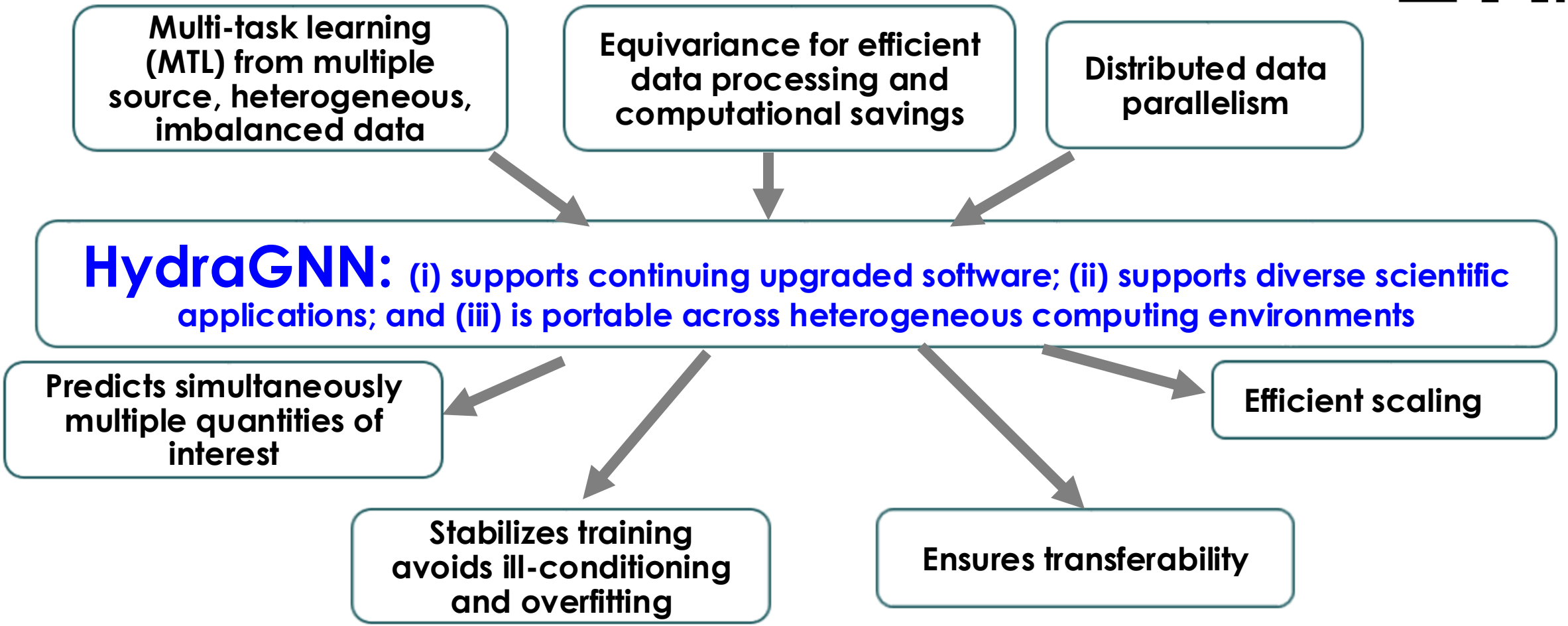
1. a graph embedding layer
2. hidden graph layers to capture short range interactions between nodes in the graph
3. pooling layers interleaved with graph layers to synthetize information related to adjacent nodes via aggregation
4. fully connected (FC) dense layers at the end of the architecture to capture global features of the properties of interest



Convolutional operations aggregate information from neighboring nodes, thereby enabling transferability of local information to larger scales

HydraGNN: a scalable GNN architecture for materials science applications

https://www.osti.gov/doecode/biblio/65891
https://github.com/ORNL/HydraGNN

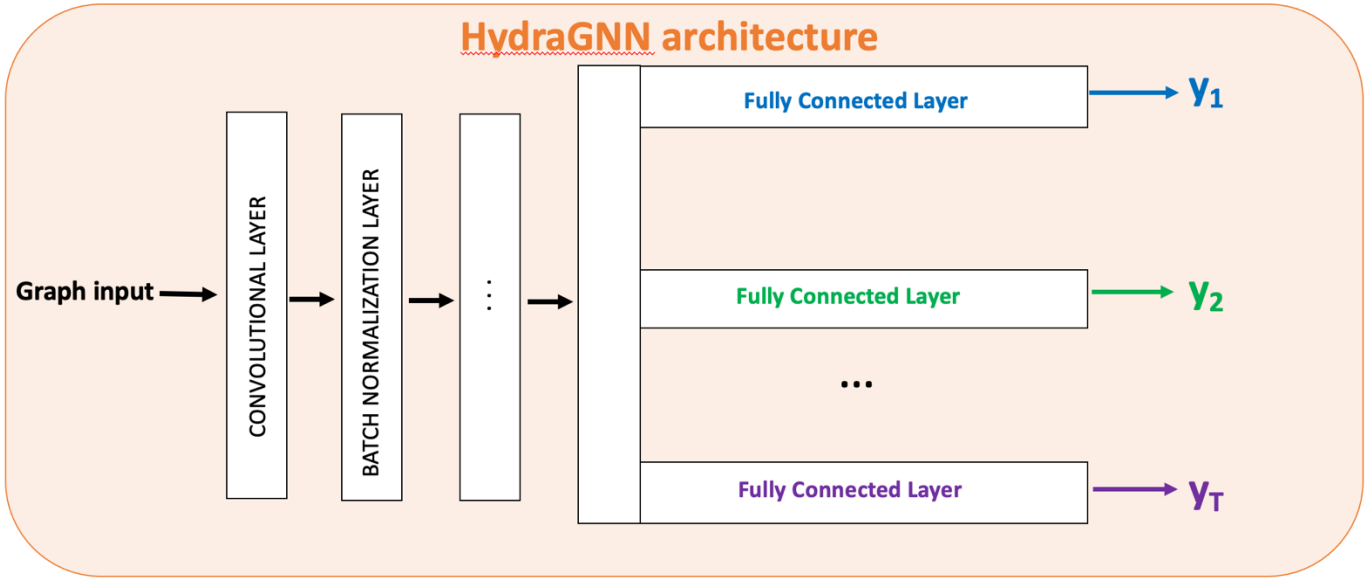


HydraGNN: multi-task learning

Multi-Task Learning stabilizes predictions of multiple properties
Each property operates as a mutual regularizer on the other properties

Quantities simultaneously predicted:

- **Property y_1**
- **Property y_2**
- ...
- **Property y_T**



$\mathbf{w}$ = parameters of the neural network to optimize during training

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y}_{\text{predict},1}(\mathbf{w}) - \mathbf{y}_1\|_2^2 + \|\mathbf{y}_{\text{predict},2}(\mathbf{w}) - \mathbf{y}_2\|_2^2 + \ldots + \|\mathbf{y}_{\text{predict},T}(\mathbf{w}) - \mathbf{y}_T\|_2^2$$

Global Multi-Task Training Loss Function

HydraGNN: equivariance

Equivariance is the property that, under Euclidean transformations, maintains consistency between the geometric structure and the physical properties associated with it.
This property is stronger than regular invariance that maintains only geometric properties.

Equivariance collapses the whole class of structurally and functionally equivalent compounds into just one representative.

Implementing equivariance in the message passing layers acts as an **inductive bias.**
It eliminates data redundancy and reduces the computational cost to reach the desired accuracy. ***This is expected also to reduce energy consumption.***

Examples of invariant material properties:
HOMO-LUMO gap, free energy, vibrational spectrum, electronic excitation spectrum

Examples of equivariant material properties:
Electron charge density, atomic forces, magnetic moment

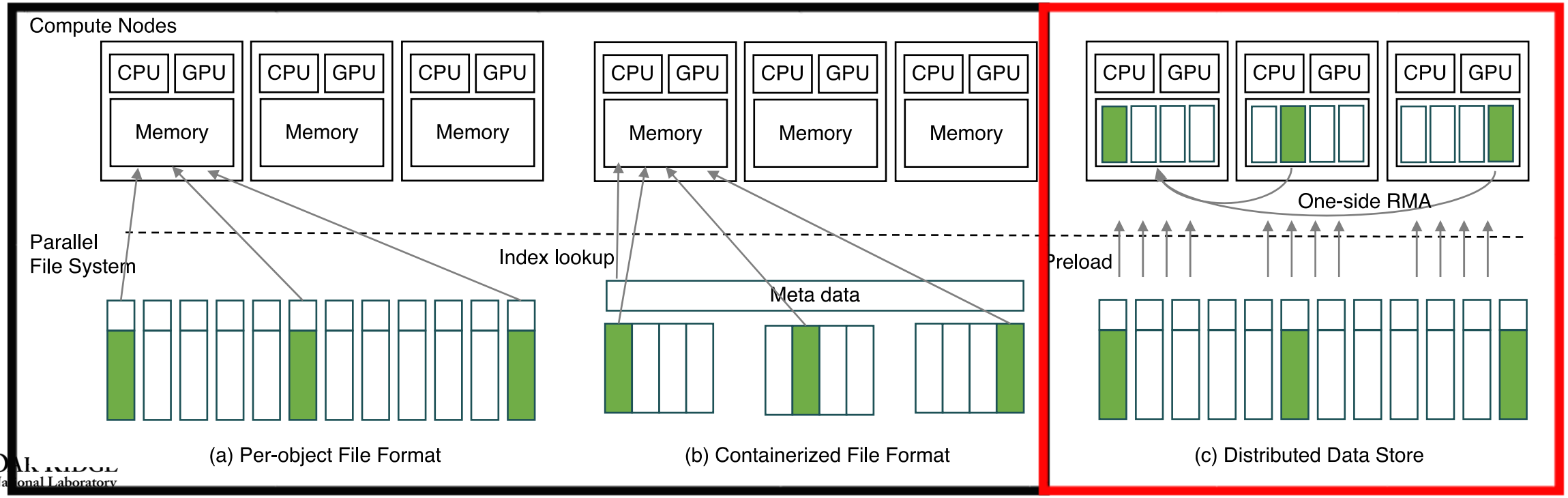
HydraGNN: scalable training with distributed data parallelism (DDP)

Traditional form of DDP perform one of the following operations:

- Move all data onto the memory of one compute unit, which is possible only when the data is relatively small to fit in it
- Periodically pull data from the parallel file system (PFS), which may lead to traffic congestion on the communication network

Our Distributed Data Store (DDStore) library partitions data in chunks and moves each chunk from the PFS to the compute memory of each node. This:

- **Helps scaling DDP for data that CANNOT be stored within the memory of one compute unit**
- **Avoids frequent communications with the PFS**



Applications of HydraGNN to materials science

Prediction of energetic properties of solid solution alloys
in collaboration with:
- CSED (CCSD)
- OLCF (CCSD)
- MSTD (PSD)

Prediction of opto-electronic properties of organic molecules
in collaboration with:
- CSED
- CSMD (CCSD)

Prediction of vibrational spectra of organic molecules
in collaboration with SNS

Geometry optimization of structures with defects
in collaboration with CNMS

Predictive graph foundation models (GFMs)
in collaboration with:
- OLCF
- AMD
- Lawrence Berkeley National Laboratory

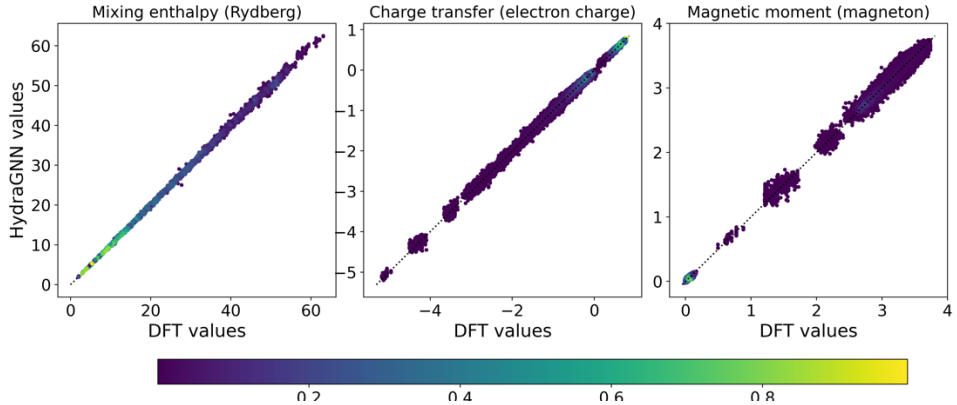
OAK RIDGE National Laboratory

1. Prediction of atomic partial charge and magnetic moment of solid solution alloys

M. L. P. and M. Eisenbach
FePt binary alloy with 32 atoms - LSMS-3 data. United States: N. p., 2021.
https://www.osti.gov/biblio/1762742

Markus Eisenbach

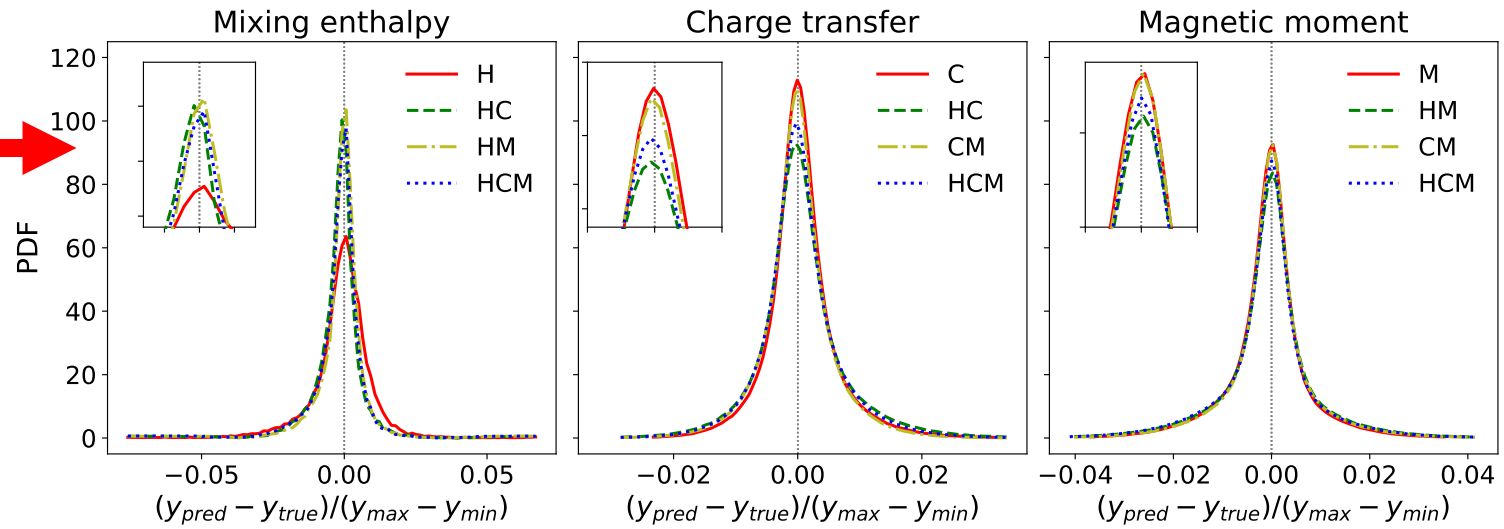
M. L. P., et al., *Multi-task graph neural networks for simultaneous prediction of global and atomic properties in ferromagnetic systems*, https://iopscience.iop.org/article/10.1088/2632-2153/ac6a51/meta

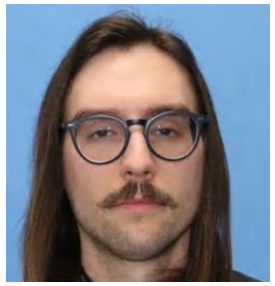


Pei Zhang

Jong Youl Choi

Multi-task learning enhances stability and accuracy, and extracts underlying physics correlation





Sam Reeve

OAK RIDGE National Laboratory

2. Transferable predictions of atomization energy of solid solution alloys across lattices of increasing size

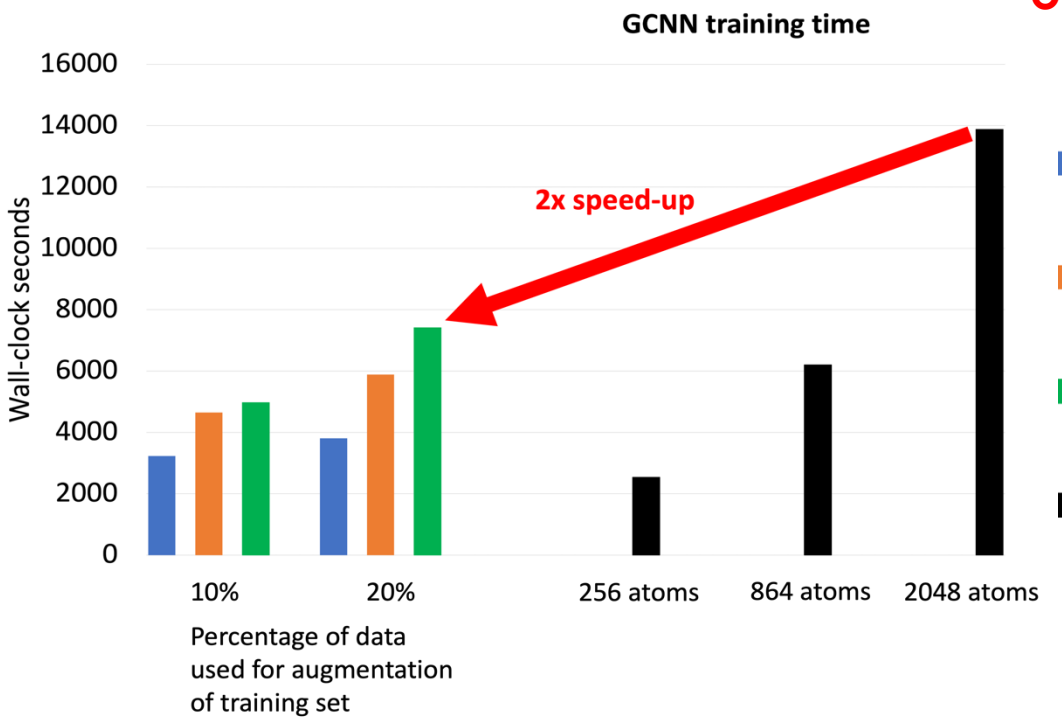
ORNL_AISD_NiPt iopen-source dataset
https://www.osti.gov/biblio/1958172

M. L. P. et al,
https://iopscience.iop.org/article/10.1088/2632-2153/ad3d2c/meta

Each atomic sample has a disordered phase obtained running geometry optimization that starts from an initial regular crystal structure of type face-centered cubic (FCC) crystal structure.
65,046 atomic structures with **256 atoms**
63,936 atomic structures with **864 atoms**
61,997 atomic structures with **2,048 atoms**

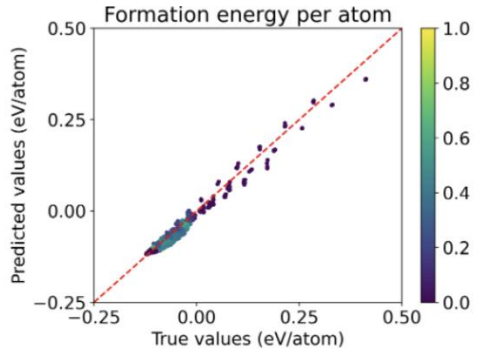
Markus Eisenbach

Mariia Karabin

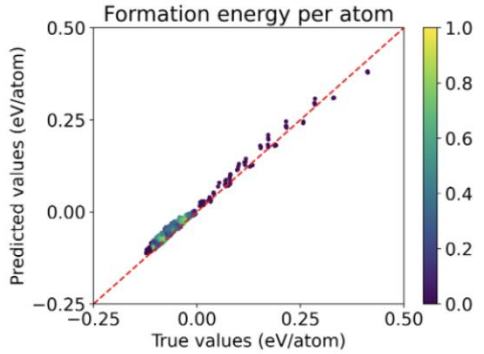


HydraGNN uses the short-range interactions learnt on small lattices to transfer the learning on larger lattices

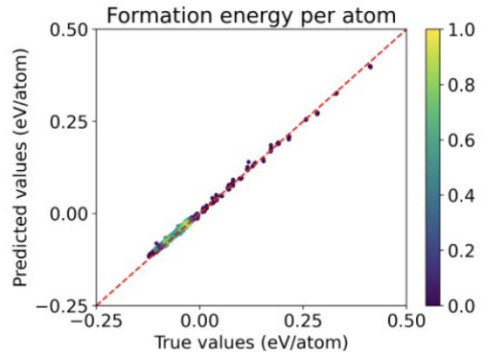
- Augment training data with crystals of 864 atoms
- Augment training data with crystals of 2048 atoms
- Augment training data with crystals of 864 atoms and 2048 atoms
- Entire training set of crystals of the same size used for training



Train on 256 atoms - Test on 2,048 atoms



Train on 256 atoms + 10% 864 atoms
Test on 2,048 atoms



Train on 256 atoms + 20% 864 atoms
+ 20% 2,048 atoms
Test on 2,048 atoms

3. Transferable predictions of atomization energy of solid solution alloys across chemical compositions

High-accuracy DFT data for solid solution alloys Nb-Ta-V

M.L.P., et al. https://doi.org/10.1038/s41597-024-03720-3

Open-source datasets uploaded to OLCF Data Constellation:

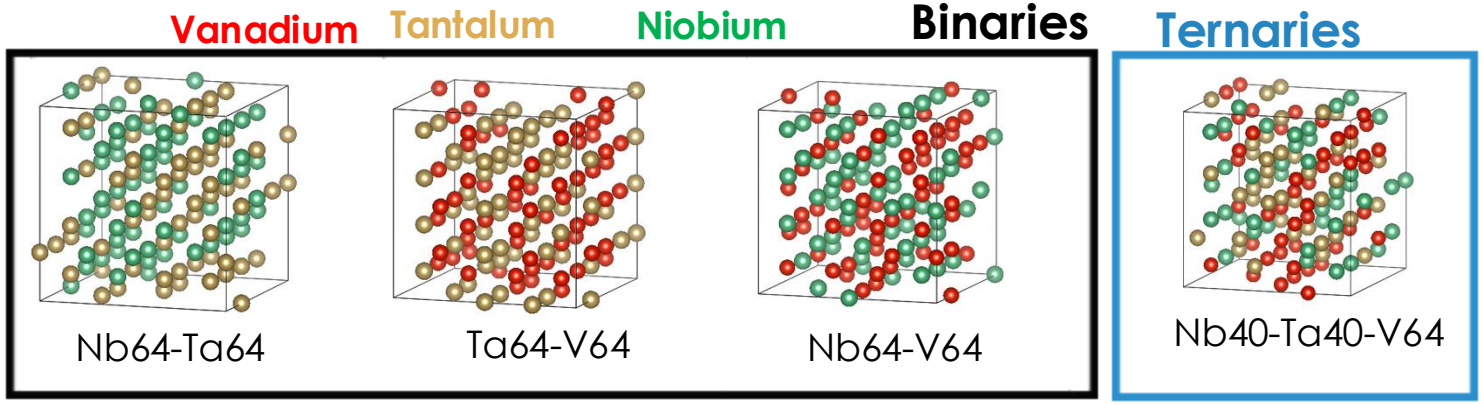
M. L. P. et al., *TaV_BCC_SolidSolution_128atoms_VASP6*
https://www.osti.gov/biblio/2222910
M. L. P. et al., *NbV_BCC_SolidSolution_128atoms_VASP6*
https://www.osti.gov/biblio/2228839
M. L. P. et al., *NbTa_BCC_SolidSolution_128atoms_VASP6*
https://www.osti.gov/biblio/2222906
M. L. P. et al., *NbTaV_BCC_SolidSolution_128atoms_VASP6*
https://www.osti.gov/biblio/2217644

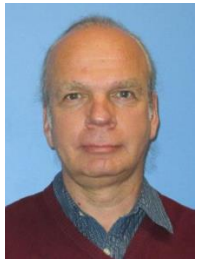
Body centered-cubic lattices with 128 atoms

Total amount of data

- 3,100 atomic configurations for each binary
- 10,500 atomic configurations for ternaries

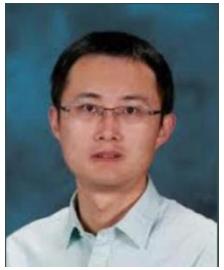
Geometry optimization on each atomistic structure




German Samolyuk


Jong Youl Choi


Markus Eisenbach


Junqi Yin

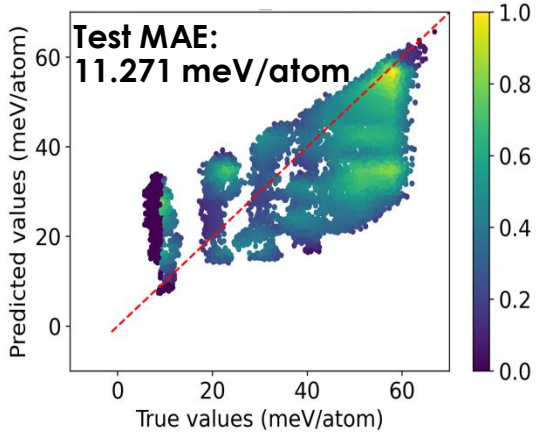

Pei Zhang

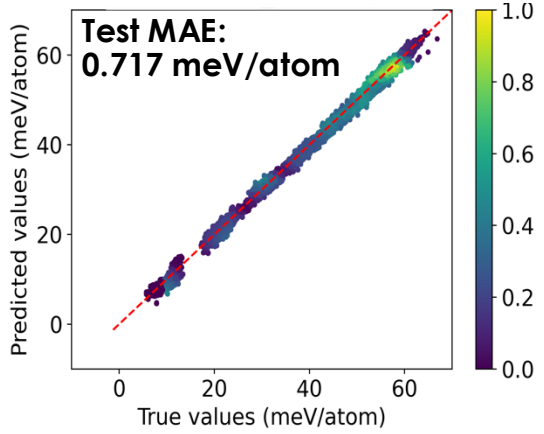

David Rogers


Ying Yang

Zero-shot extrapolations from binaries to ternaries

HydraGNN learns features from binaries and transfers them to ternaries with minimal amount of additional ternary data



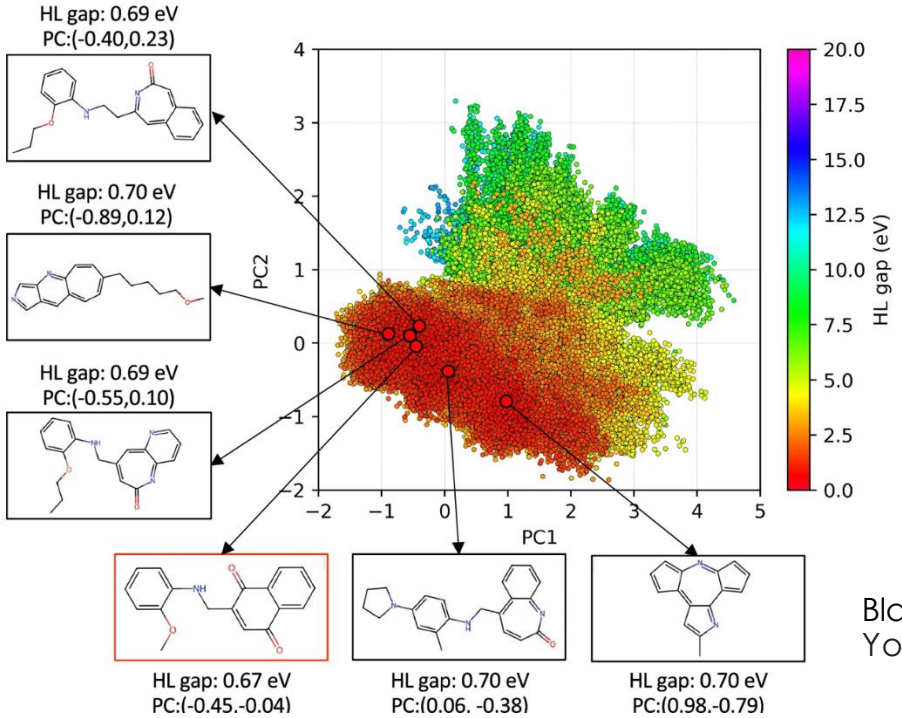


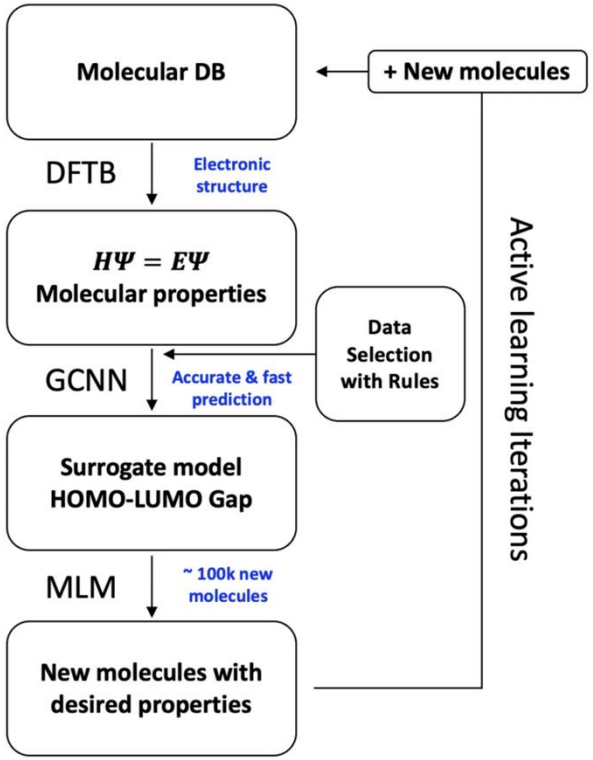
Augmentation of training data with 1 ternary configuration for each chemical composition

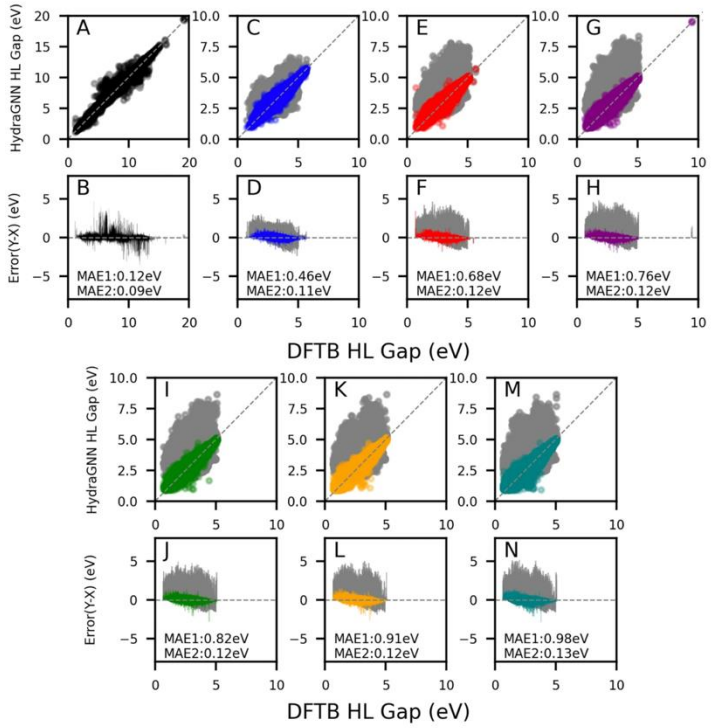
OAK RIDGE National Laboratory

4. Design of organic molecules with desired HOMO-LUMO gap

HydraGNN combined with generative algorithms enable an accelerated screening of high-dimensional parameter spaces for discovery of molecules with desired optoelectronic properties







Release of generated molecules: P. Yoo, et al.. United States: N. p., 2023. Web. https://www.osti.gov/biblio/1996925

Blanchard, A.E. et al. (2022). https://doi.org/10.1007/978-3-031-23606-8_1
Yoo, P., Bhowmik, D., Mehta, K. et al. (2023). https://doi.org/10.1038/s41598-023-45385-9

A. Blanchard, et al. *AISD HOMO-LUMO*. United States: N. p., 2022. Web. https://www.osti.gov/dataexplorer/biblio/dataset/1869409-aisd-homo-lumo


Pilsun Yoo

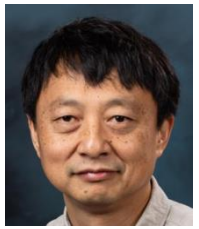

John Gounley

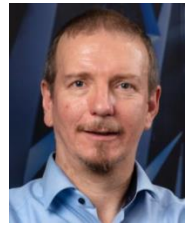

Andrew Blanchard


Debsindhu Bhowmik


Kshitij Mehta


Pei Zhang

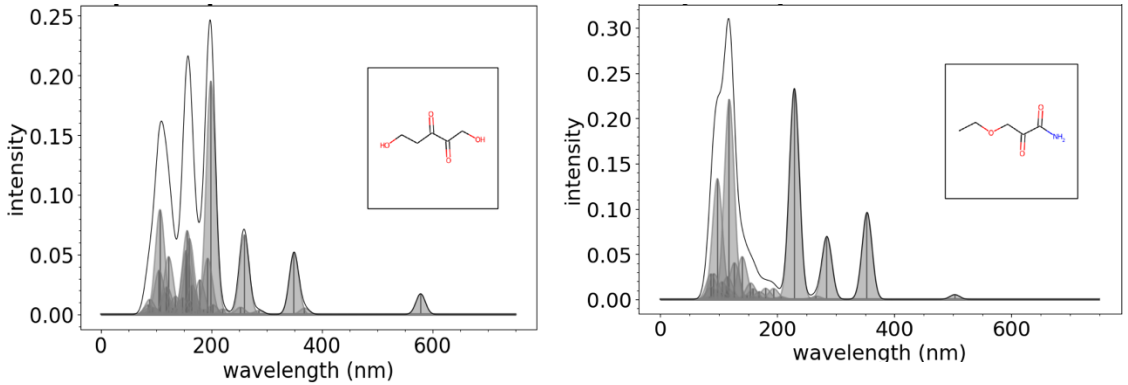

Frank Liu

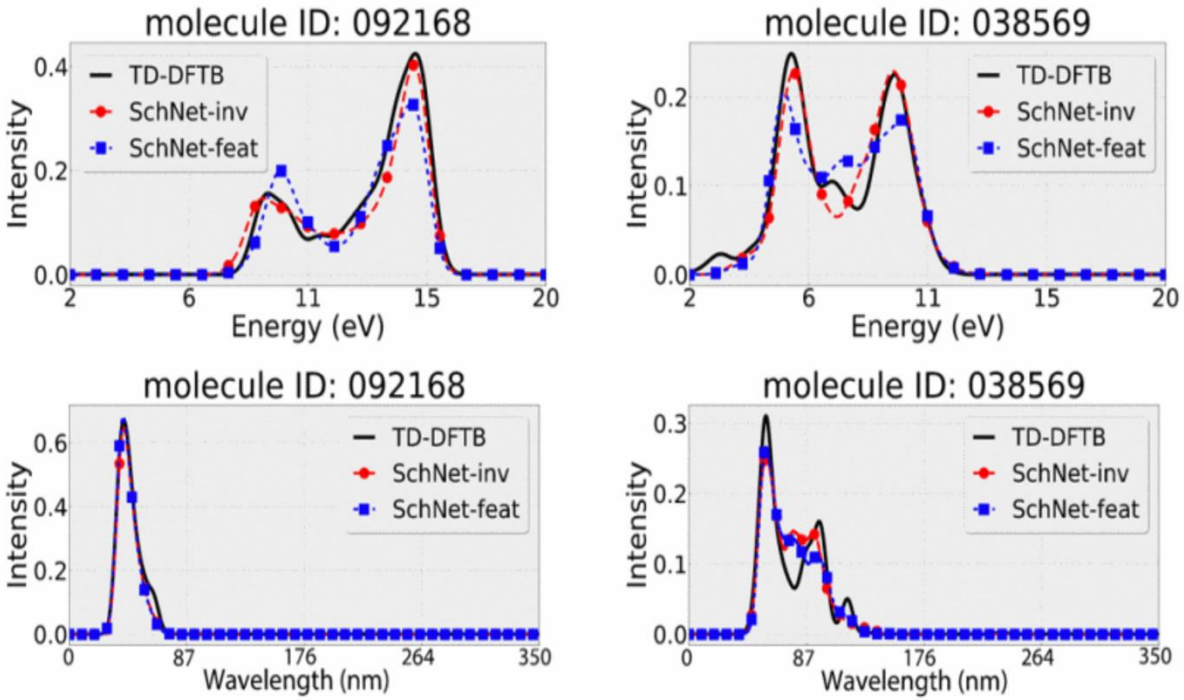

Stephan Irle

OAK RIDGE National Laboratory

5. Prediction of UV-vis spectra of organic molecules

M. L. P., et al .(2023). https://doi.org/10.1038/s41597-023-02408-4
M. L. P., et al. 2022. https://www.osti.gov/biblio/1890227
M. L. P., et al.. 2023. https://www.osti.gov/biblio/1907919
J. Baker, M. L.P., and C. Hauck, 2024, https://ieeexplore.ieee.org/abstract/document/10500060





Pilsun Yoo **Kshitij Mehta** **Stephan Irle**

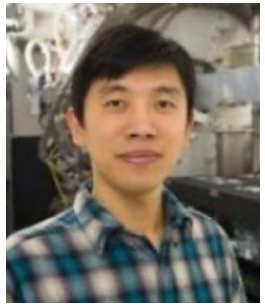
Cory Hauck

Justin Baker
University of Utah, Salt Lake City

Invariant-preserving HydraGNN models are more data-efficient and therefore achieve higher accuracy when trained on a fixed volume of data

6. Prediction of vibrational spectra of organic molecules

Bowen Han



Yongqiang Cheng

Kshitij Mehta

Fully optimized molecular geometry

↓

Prediction of total energy E

↓

Calculation of atomic forces $\mathbf{f}_i = -\nabla_{x_i} E$

↓

Calculation of the Hessian $H_{i,j} = \frac{\partial^2 E}{\partial x_i \partial x_j}$

↓

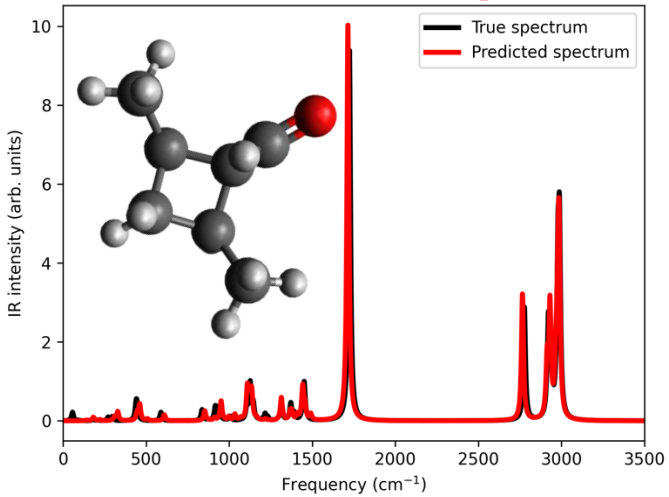
Spectral decomposition of the Hessian

Eigenvalues = Vibrational frequencies
Eigenvectors = Vibrational modes (patterns of atomic motion)

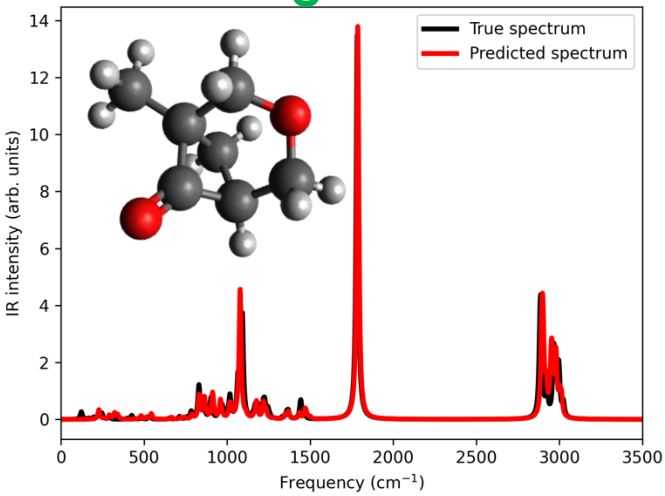
NequIP model **trained and validated on QM8 molecules**
tested on QM9 molecules

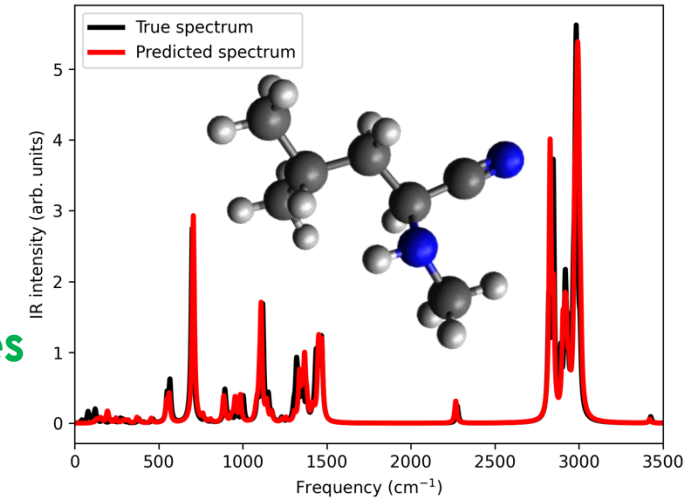
Infrared spectra

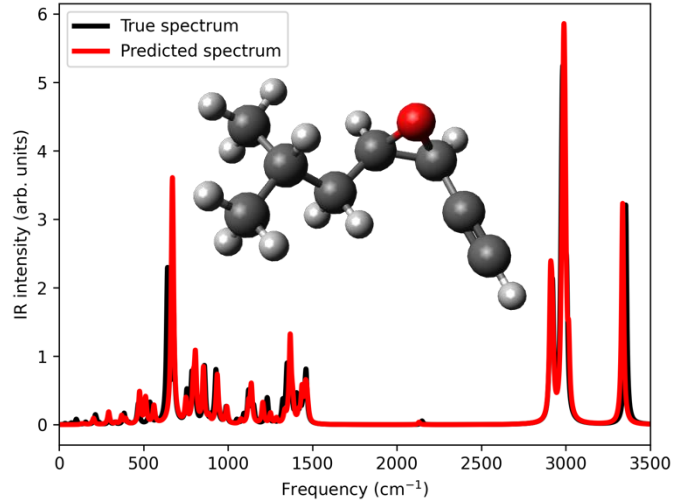
Validation on QM8



Testing on QM9







7. GNN-accelerated geometry optimization of atomistic structures with defects

Goal: Identify 3D atomistic structures associated with experimentally observed with 2D scanning transmission electron microscopy (STEM)

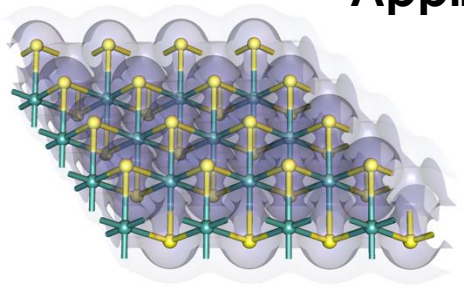
Approach:

- Train GNN surrogate model for accurate and fast predictions of energies and forces
- Use ensemble of models to stabilize predictions via averaging
- Integrate GNN into geometry optimizer to enable rapid screening of several 3D structures with defects
- Use GNN-accelerated structure optimizer to solve inverse problem by comparing simulated 2D STEM image with experimentally observed one

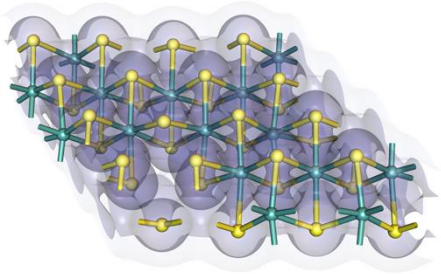
Paul Kent

Addis Fuhr

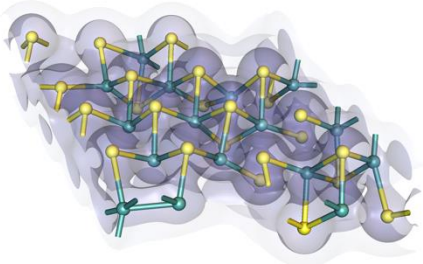
Application to 2D dichalcogenides



Defect-free structure



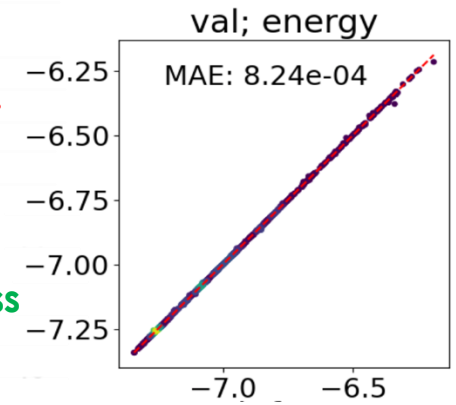
Single vacancy

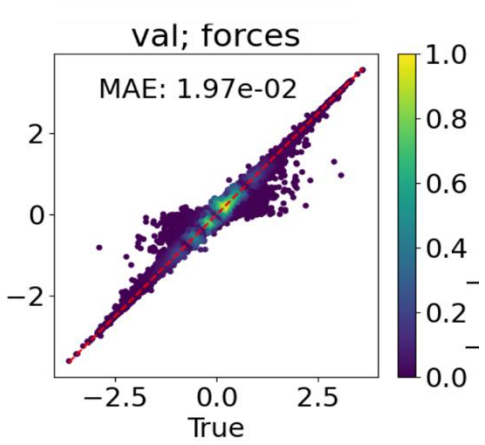


Many defects

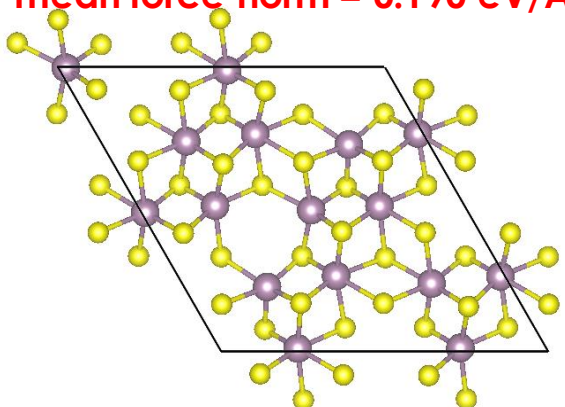
Individual models do not reach desired accuracy.

Ensemble models dampen stochasticity of training and improve robustness of predictions.

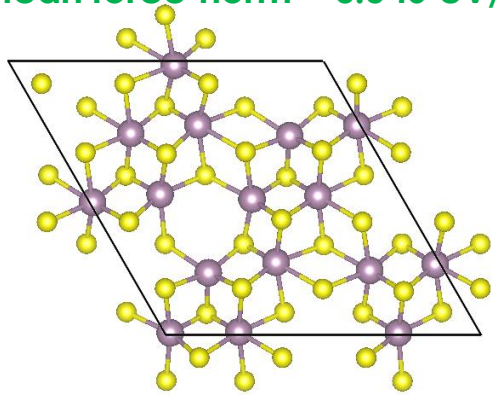




Initial geometry:
mean force norm = 0.190 eV/Å



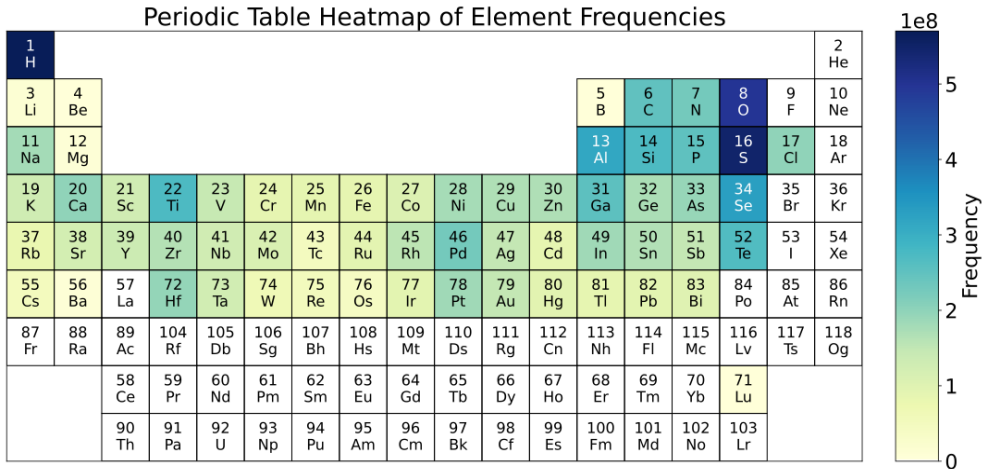
Final geometry:
mean force norm = 0.046 eV/Å



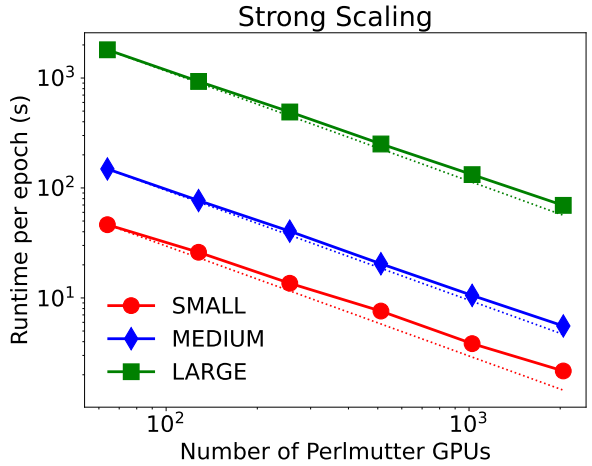
OAK RIDGE National Laboratory

8. Scalable predictive graph foundation models (GFMs)

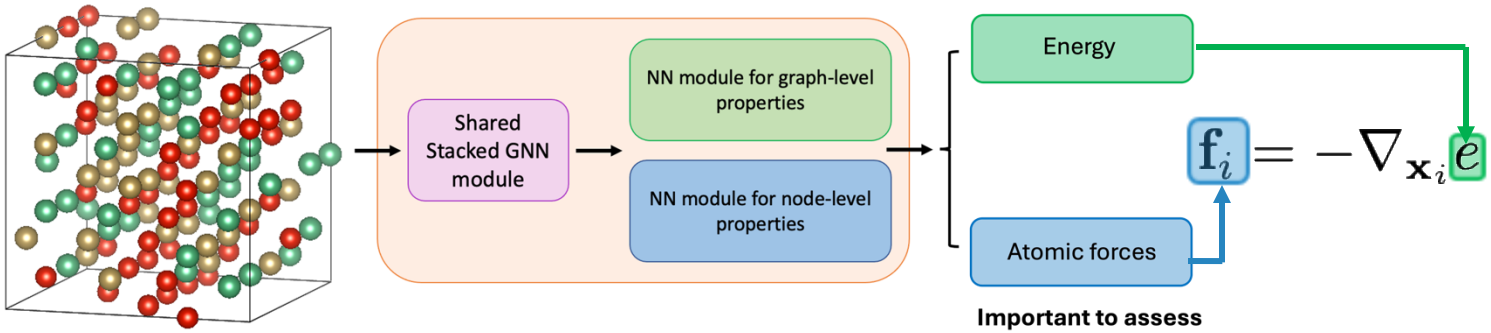
Dataset	Type of compounds	# data samples
Open Catalyst 2020	Alloy slabs with interacting catalysts on the surface	134,929,018
Open Catalyst 2022	Alloy slabs with interacting catalysts on the surface	8,847,031
Materials Project Trajectory	Bulk metals and alloys	1,580,395
ANI1x	Organic molecules	4,956,005
QM7x	Organic molecules	4,195,237
Total		**154,507,686**



SMALL model ~ 54 K parameters
MEDIUM model ~ 16 M parameters
LARGE model ~ 164 M parameters

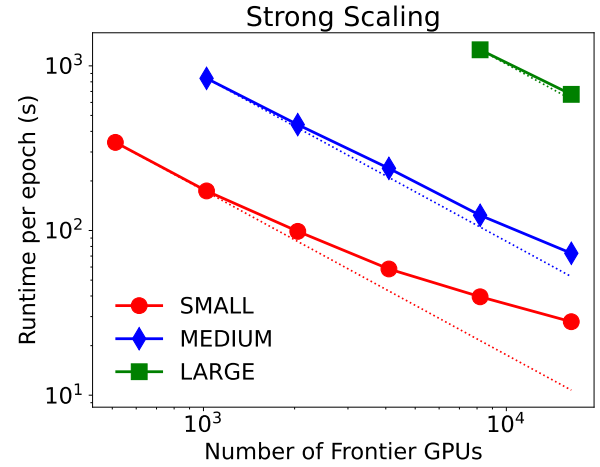


Important to assess chemical stability of an atomistic structure



Important to assess dynamical stability of an atomistic structure

Prasanna Balaprakash



Jong Youl Choi
OAK RIDGE National Laboratory

Pei Zhang

Kshitij Mehta

David Rogers

Khaled Ibrahim
LBNL

Karl W. Schulz
AMD Research

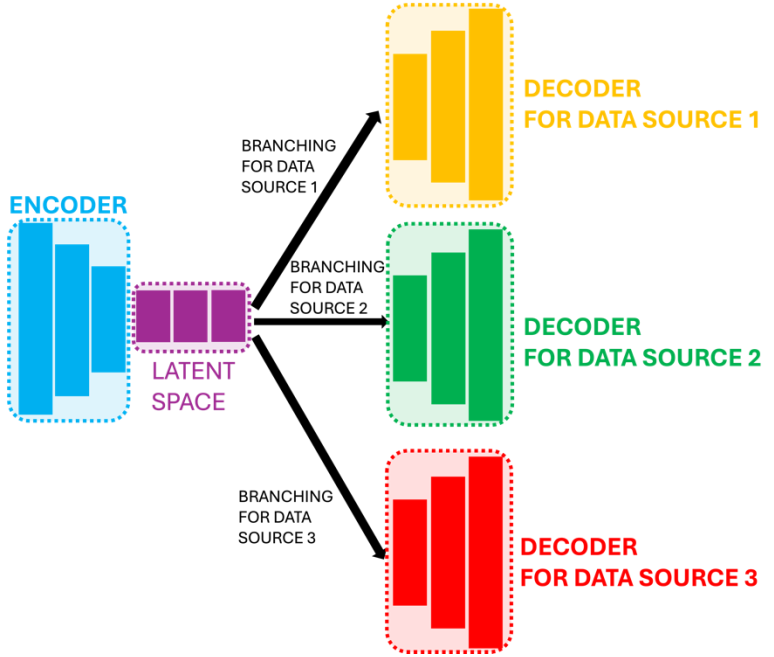
Ashwin Aji
AMD Research

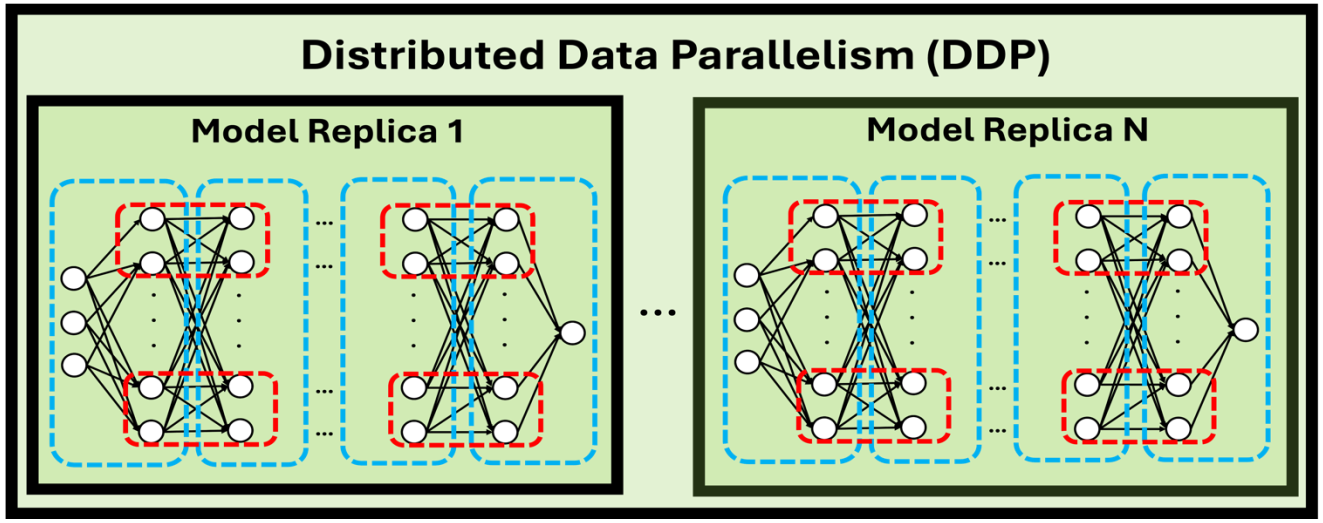
Jorda Polo
AMD Research

M.L.P. *Scalable Training of Graph Foundation Models for Atomistic Materials Modeling: A Case Study with HydraGNN*, https://arxiv.org/abs/2406.12909

Future work

- Develop equivariant generative diffusion models with masking techniques
- Integrate DDP, model pipeline parallelism (MPP), and model tensor parallelism (MTP)
- Generalize MTL for stable and energy-efficient training
- Apply HydraGNN to modeling evolution of grain structure at mesoscopic scale



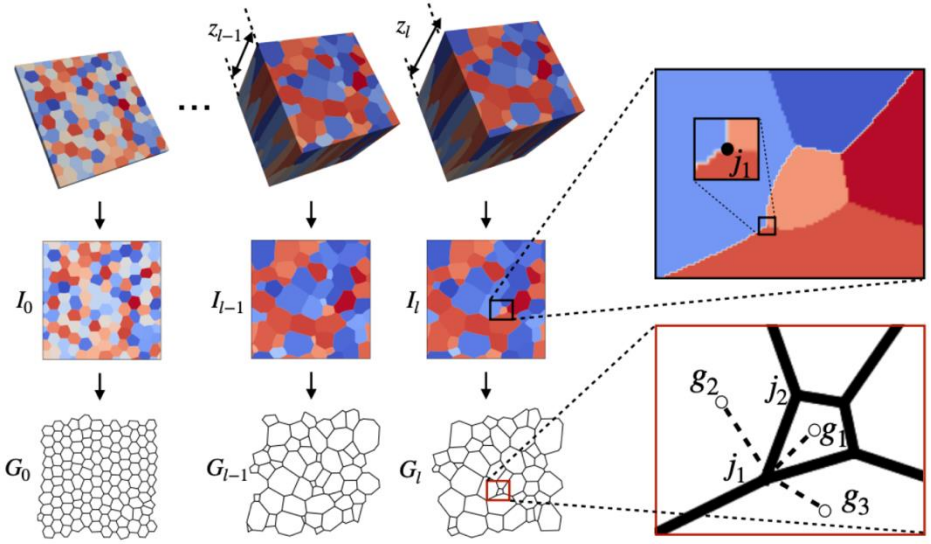


Model Pipeline Parallelism (MPP)

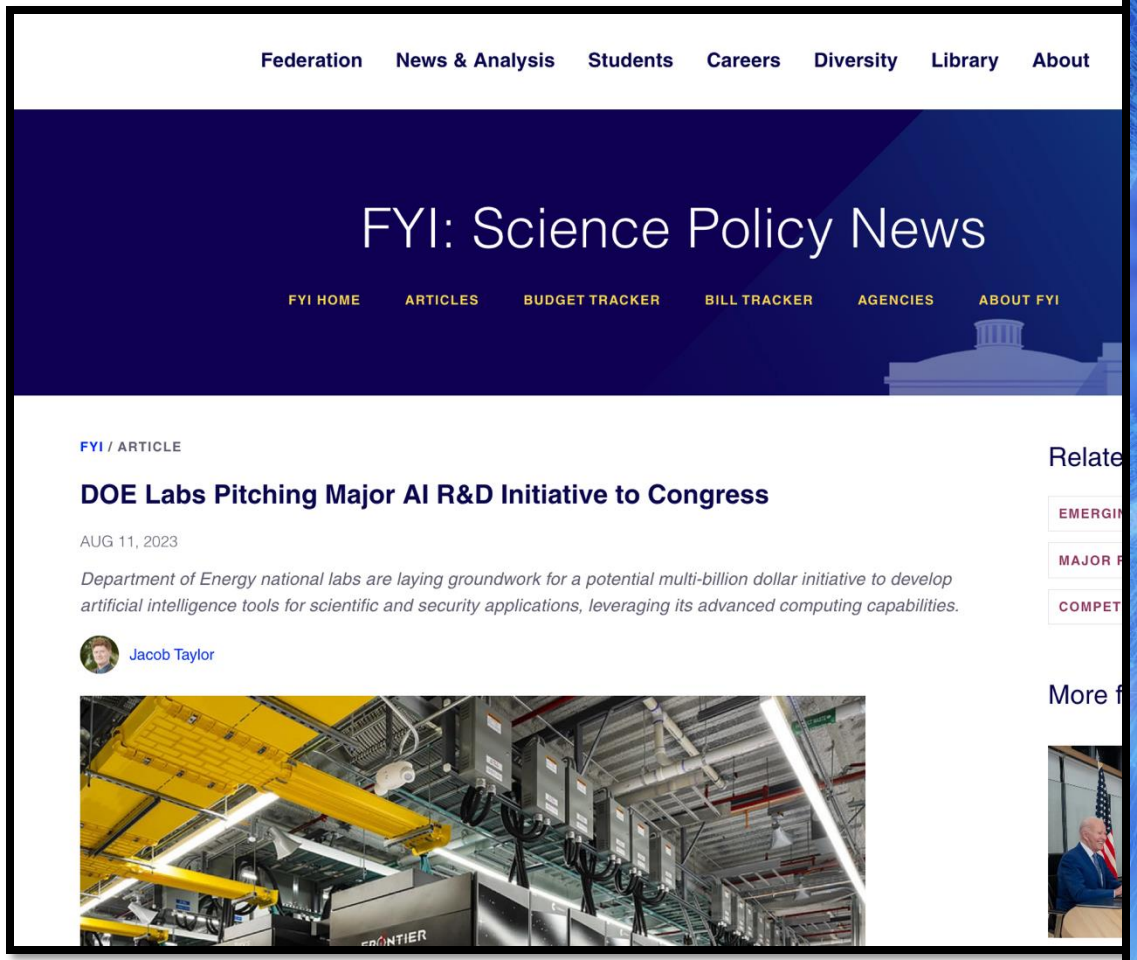
Model Tensor Parallelism (MTP)

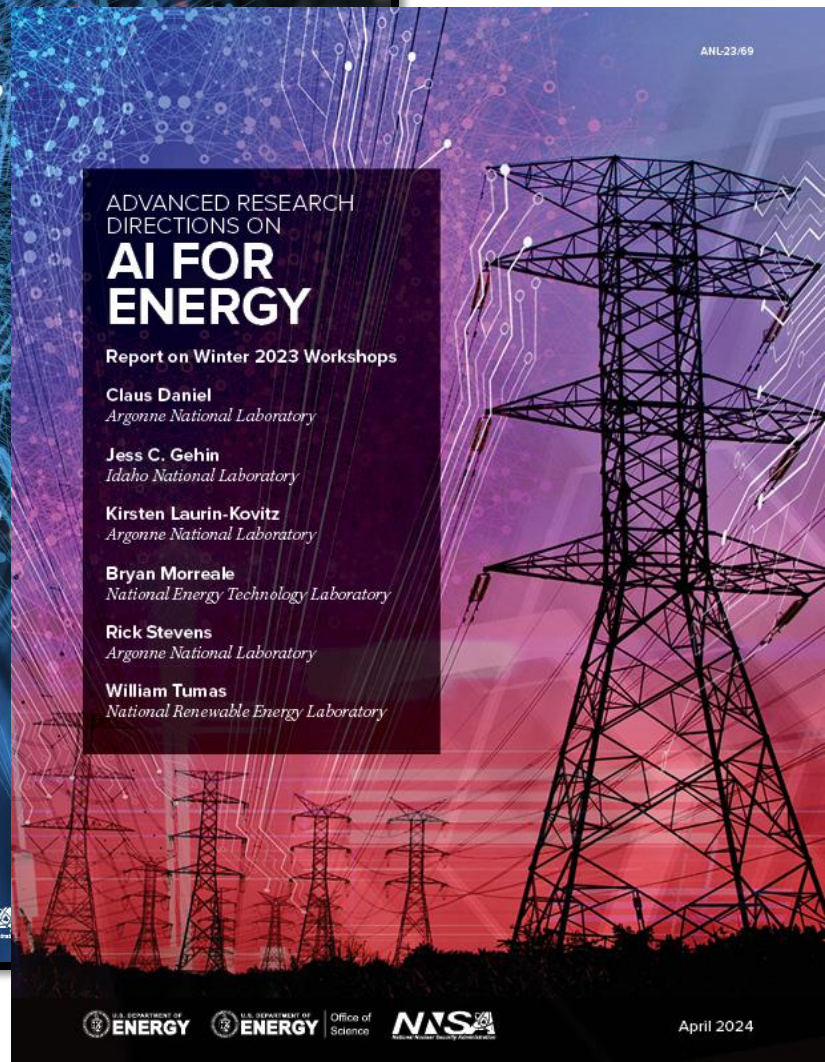
Mesoscale

Nodes = Voronoi centers
Edges = connection between Voronoi centers



FASST: Frontiers in AI for science, security, and technology





Acknowledgments

This research is partially supported by:

- The Artificial Intelligence Initiative as part of the Laboratory Directed Research and Development (LDRD) Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725.

- The SciDAC Institute for Computer Science, Data, and Artificial Intelligence (RAPIDS), Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

- The US-DOE Advanced Scientific Computing Research (ASCR) under contract DE-SC0023490.

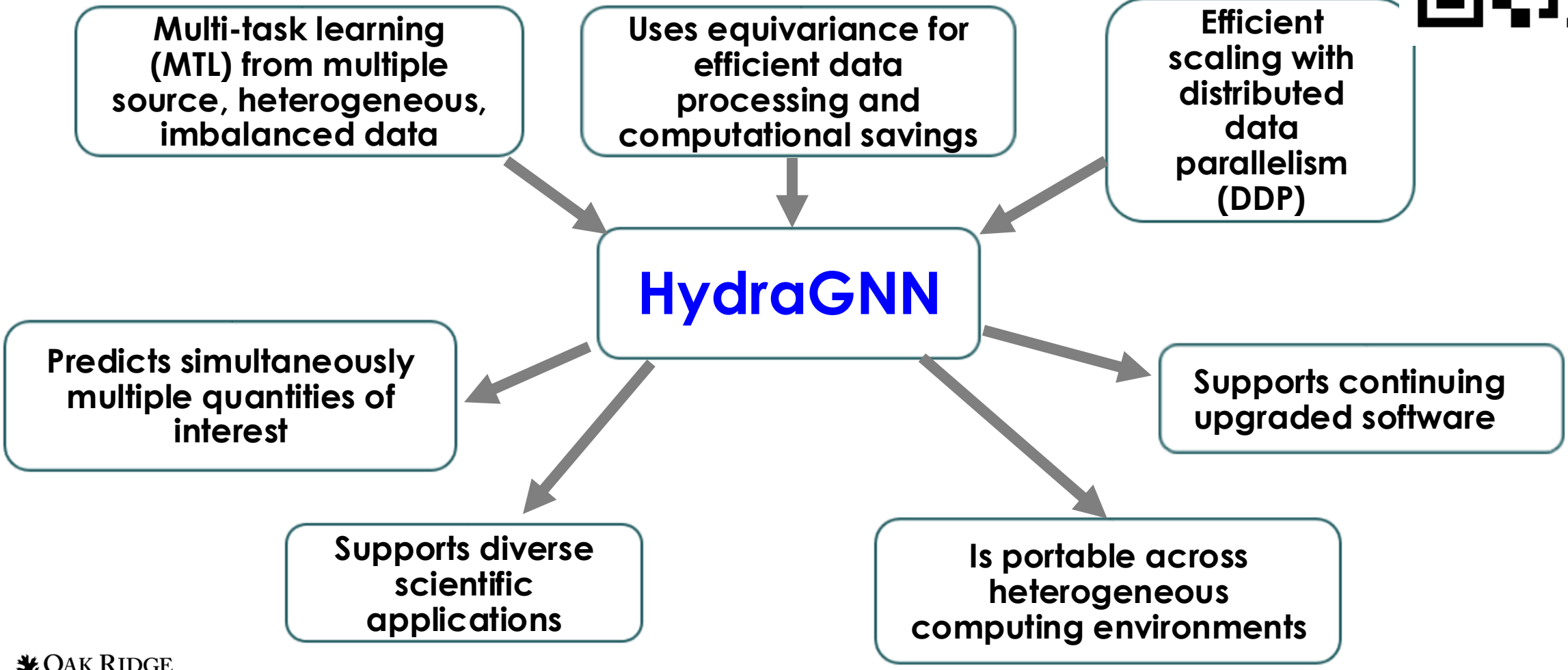
This research used resources of the Oak Ridge Leadership Computing Facility and of the Edge Computing program at the Oak Ridge National Laboratory. Computer time was provided by the INCITE program using the OLCF award CPH161 (AY2024), the SummitPLUS program using MAT272, and OLCF Director's Discretion Project program using the OLCF awards MAT250 (AY2022) and LRN026 (AY2023).

This research also used resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 and No. DE-AC02-05CH11231, using NERSC awards ASCR-ERCAP0022058 (AY2022), ASCR-ERCAP0025216 (AY2023) and ASCR-ERCAP0027259 (AY2024).


Thank you!
Questions?

HydraGNN: a scalable GNN architecture for materials science applications

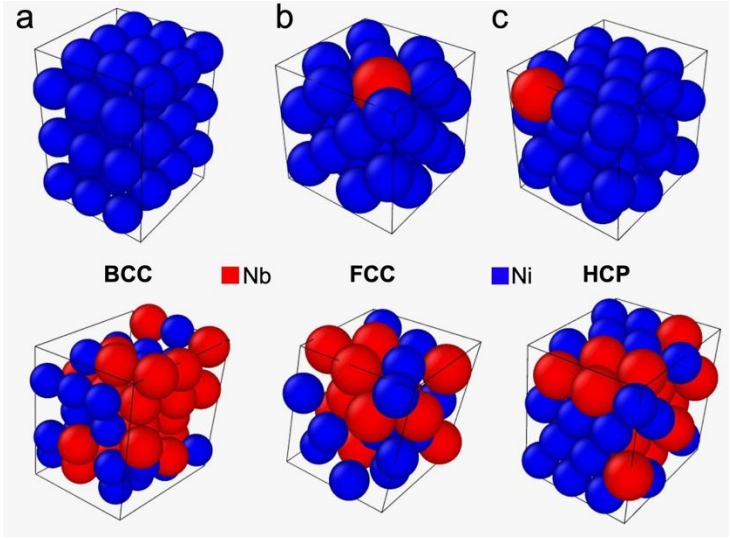
https://www.osti.gov/doecode/biblio/65891
https://github.com/ORNL/HydraGNN



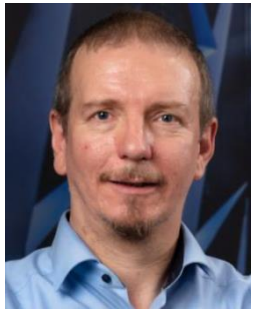
Prediction of mechanical properties of solid solution alloys

Open-source dataset:
ORNL_AISD_NiNb
https://www.osti.gov/dataexplorer/biblio/dataset/1890159

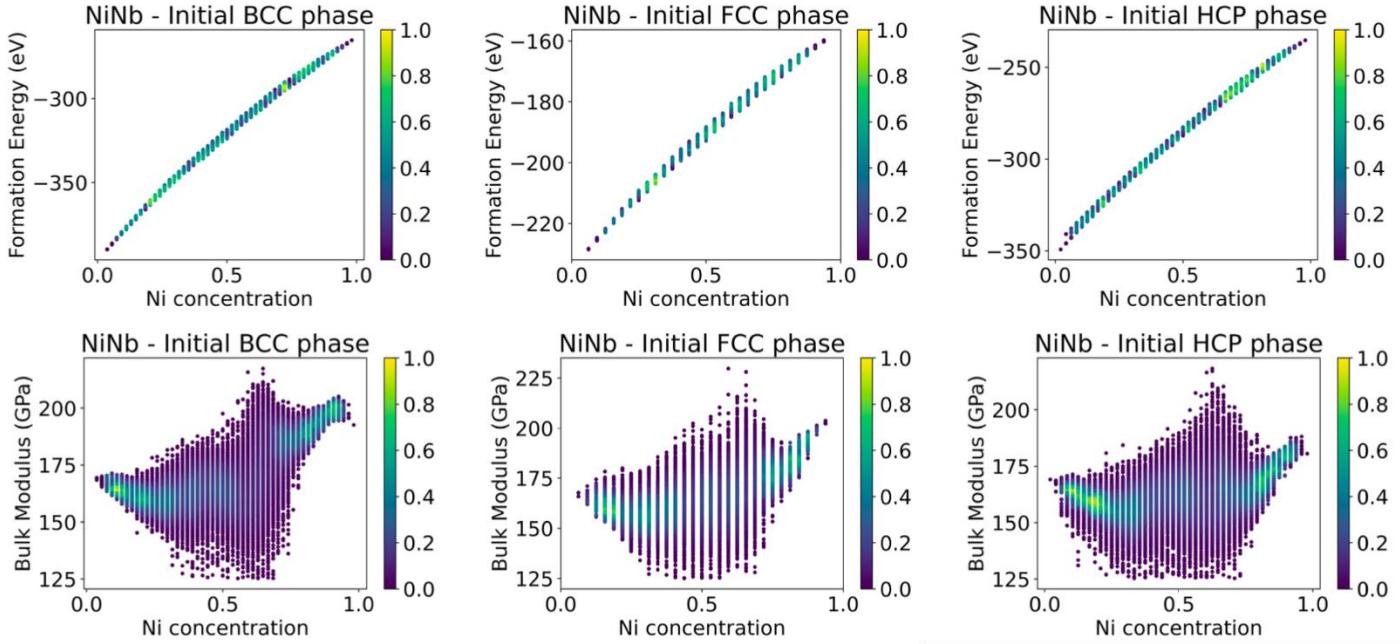
Each atomic sample has a disordered phase obtained running geometry optimization that starts from an initial regular crystal structure of body-centered cubic (BCC), face-centered cubic (FCC), or hexagonal compact packed (HCP).
46,086 BCC structures with 54 atoms
24,543 FCC structures with 32 atoms
39,303 HCP structures with 48 atoms

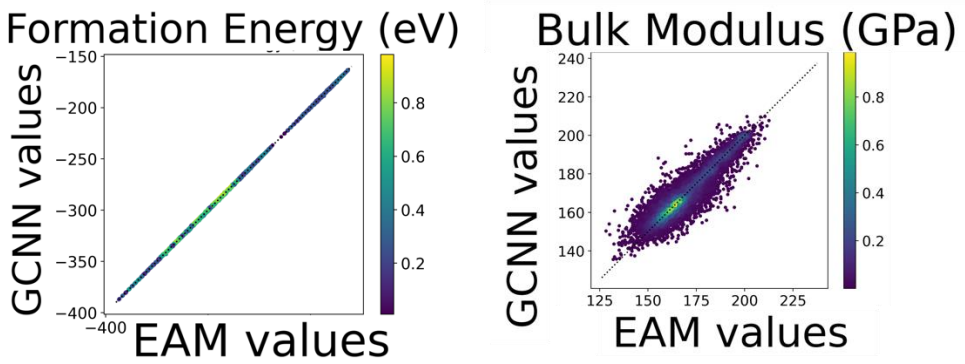


GS Jung



Stephan Irle





Formation Energy Data:
- varies significantly across chemical composition
- varies only mildly across atomic configurations for the same chemical composition

Bulk Modulus Data:
- varies significantly across chemical composition **AND** atomic configurations for the same chemical composition

→ The broader variability of the data makes the bulk modulus much more difficult to predict

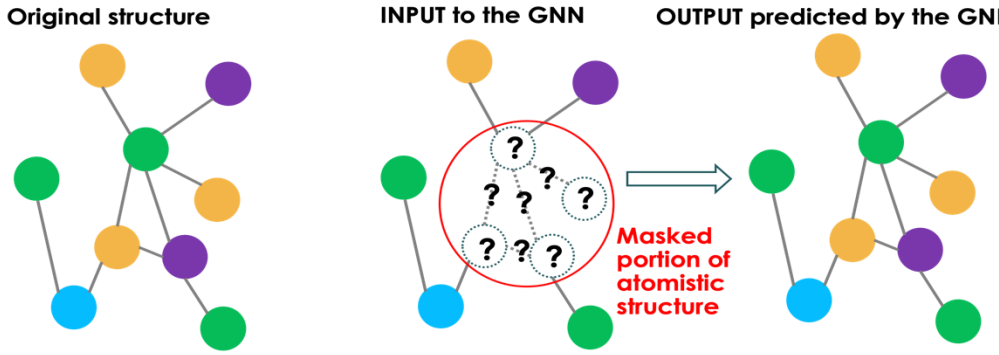
M. L. P., G. S. Jung, S. Irle, *Graph neural networks predict energetic and mechanical properties for models of solid solution metal alloy phases*, Computational Materials Science, Volume 224, 2023,112141, ISSN 0927-0256, https://doi.org/10.1016/j.commatsci.2023.112141

Future Work: Equivariant Generative Diffusion Models with Masking Techniques

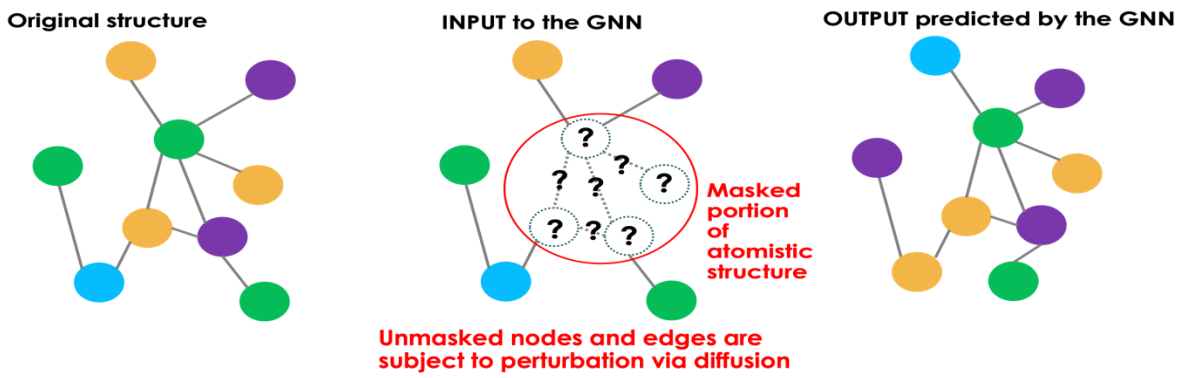
Approach: Perform autoregressive graph masking at each iterative step of the reversed diffusion process

- Impose **equivariance** constraints on the diffusion process to eliminate redundancies
- Use **global attention mechanisms** with GraphGPS to account for long-range interactions
- Use the **mask size as a tunable parameter**, to find the best compromise between exploration and exploitation

Standard Masking



Hybrid Diffusion + Masking



Expected outcome:

(a) **efficient exploration** of the material space

(b) the combined effect of equivariance + the robustness and computational efficiency of autoregressive graph masking will result in **computational and energy savings**

Future Work: Integrate Distributed Data Parallelism (DDP), Model Pipeline Parallelism (MPP), and Model Tensor Parallelism (TMP)

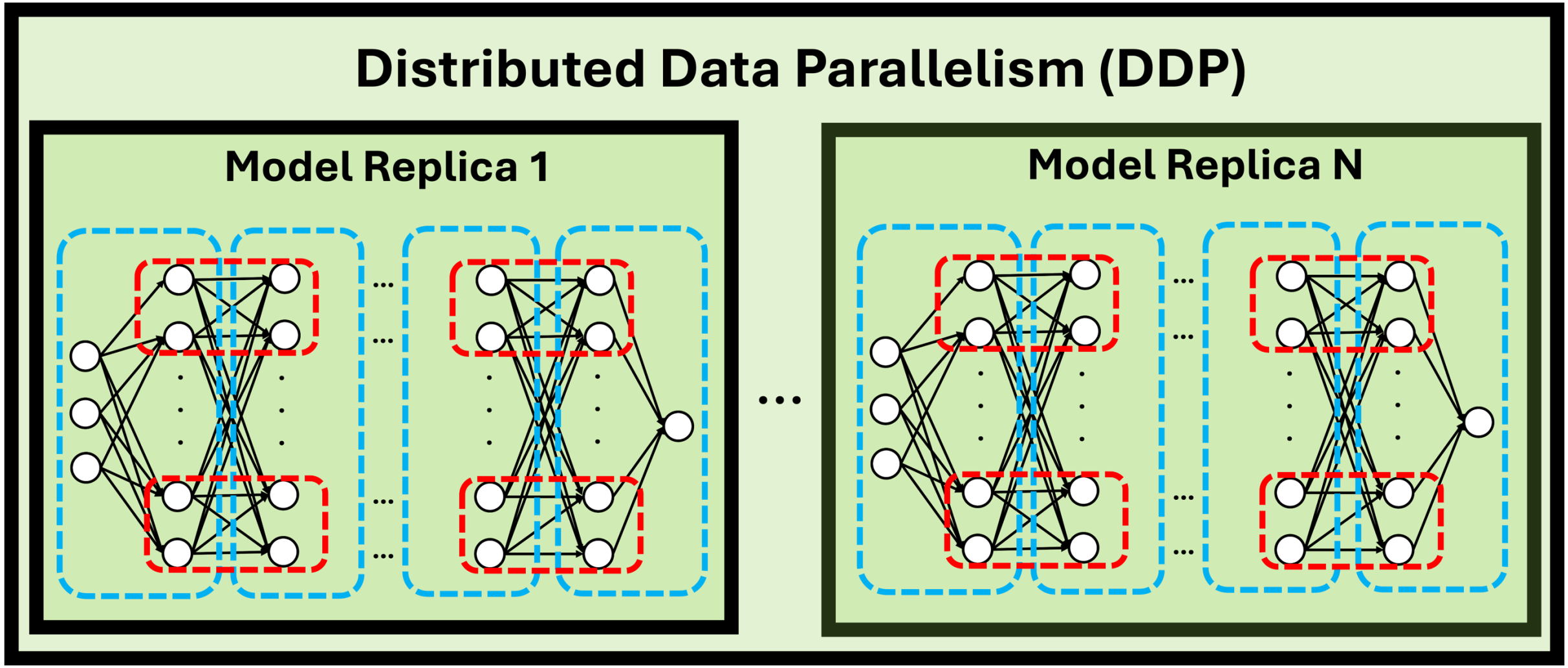
Approach: Hierarchically integrate DDP, MPP, and MTP

- If the data is too large, **DDP** will be used to partition the data across multiple GPUs.
- If the data has also broad variability, larger models may be needed to properly capture it. If the number of layers in the model is too large to fit in a single GPU, **MPP** will be used to split different layers across separate GPUs.
- If the number of neurons in each layer makes the model too large to fit in a single GPU, **MTP** will be used to split different neurons across separate GPUs.

Expected outcome:

- GNN architecture that simultaneously combines DDP, MPP, and MTP allowing training of **large, complex GNN models**
- Optimized latency and memory overhead and efficient training of GNN models at very large scale resulting in **energy savings**

Future Work: Integrate Distributed Data Parallelism, Model Pipeline Parallelism, and Model Tensor Parallelism (cont.'d)



Future Work: Generalize MTL for Stable and Energy Efficient Data Processing

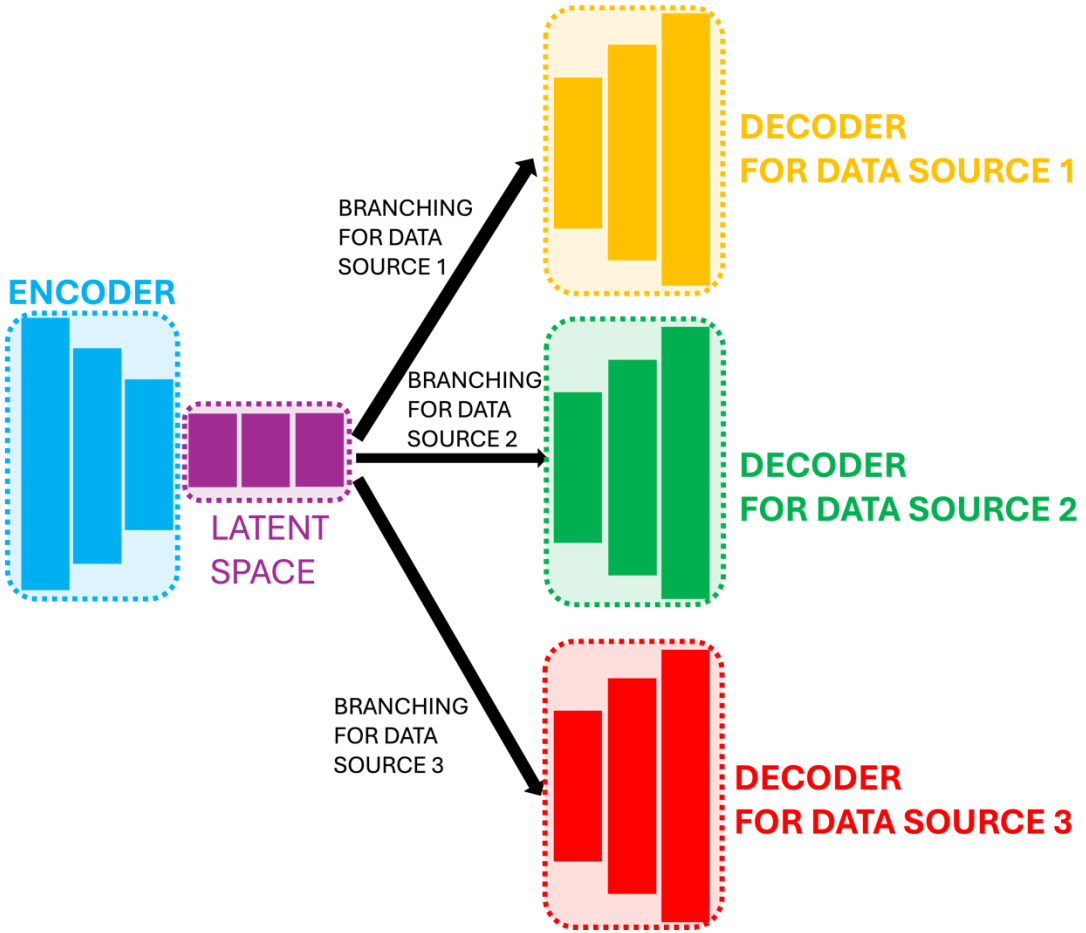
Approach: Dedicate different heads of the GFM to process data from different sources.

- Ensure compatibility between implementation of hard parameter sharing and 3D parallelization
- New hard parameter sharing implementation:
  - uncovers the correlation between the data in the latent space,
  - sends to the heads only the data that each of them will process.

Expected outcome: Reduced number of parameters and unnecessary calculations, leading to:

(i) reduced computational resources and energy consumption; and

(ii) increased numerical stability and resilience against perturbations to the model's parameters

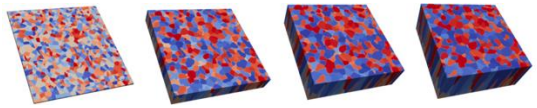


Future Work: GNN Application to MDF

Steve De Witt

Microstructural Calculations

Evolution of grains



(Y. Qin et al, https://arxiv.org/pdf/2401.03661)

Oak Ridge Leadership Computing Facility (OLCF)

Computational User Facility

GNNs

Experimental User Facilities

Manufacturing Demonstration Facility (MDF)

Spallation Neutron Source (SNS)

Center for Nanophase Materials Sciences (CNMS)