OAK RIDGE National Laboratory

Artificial Intelligence for Prediction of Multi-Scale Material Properties

Massimiliano (Max) Lupo Pasini

7th Annual Commonwealth Computational Summit

October 16 & 17, 2023 The University of Kentucky's Center for Computational Sciences (CCS) and ITS/Research Computing Infrastructure (ITS/RCI)

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

U.S. DEPARTMENT OF ENERGY

Goal and Challenges

Goal: efficiently explore high-dimensional chemical spaces for fast and accurate discovery and design of materials with desired functional properties

The high-dimensionality of the chemical space is the result of the diverse characterization of materials across **multiple scales**:

- Macroscale → chemical composition (concentration of atoms of different elements)
- Micro-scale → size and orientation of micro-structures
- Nano-scale → Arrangement of atoms of different elements and orientation of interatomic bonds

Challenges: state-of-the-art experimental (in-vivo) and computational (in-vitro) approaches are impractical to explore high-dimensional chemical spaces

- Experiments are labor-intensive and time-consuming
- Computational methods are expensive

Proposed Solution: Surrogate Models

- **Surrogate models significantly lower the computational cost** of expensive explorations of high-dimensional chemical spaces **while maintaining sufficient accuracy**

- For many applications, the structure of the **physical system can be mapped onto a graph**

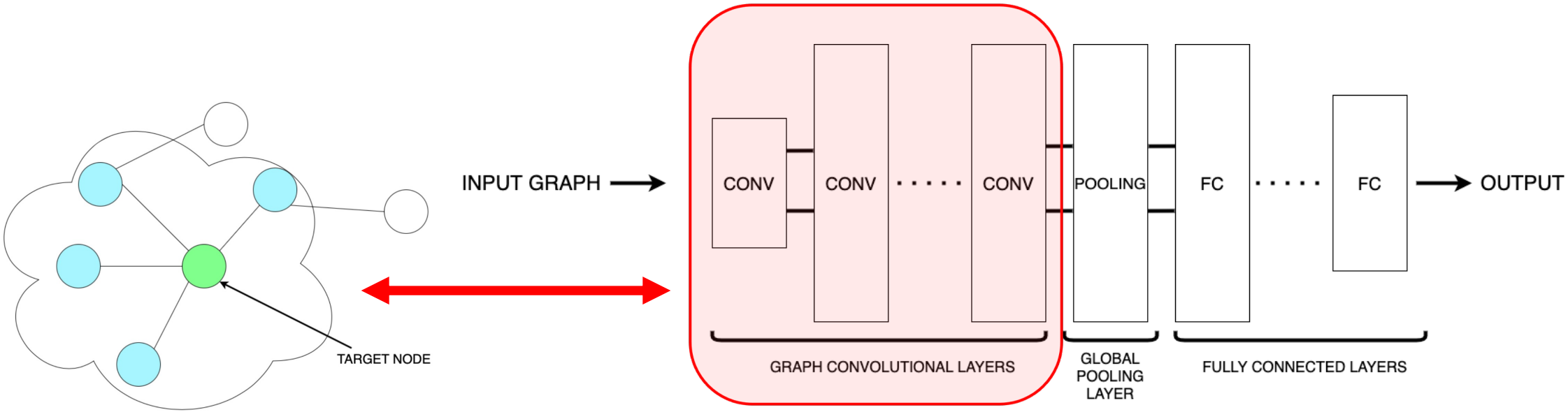
  Examples:
  - **Atomic modeling (addressed in this talk):** nodes of the graph = atoms
    edges of the graph = interatomic bonds
  - Finite element simulations**:** nodes of the graph = vertices of the mesh
    edges of the graph = connectivity of the mesh
  - Transportation: nodes of the graph = neighborhoods or cities
    edges of the graph = roads or highways

- Whenever the data can be expressed in the format of a graph, **graph neural networks (GNNs)** have been identified as promising tools to **extract relevant nodal and graph-level features** that describe the dynamics of the physical system

Graph Neural Networks (GNNs)

The architecture of a GNN is made of:

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



Convolutional operations aggregate information from neighboring nodes

HydraGNN: Distributed PyTorch Implementation of Multi-Headed GNNs

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

HydraGNN simultaneously enables the following computational capabilities:

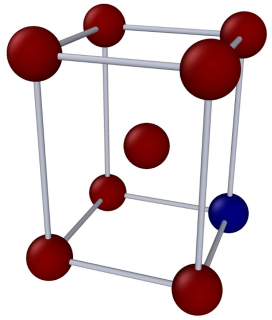
- **Multi-task learning (MTL)** for stabilization by extracting physics correlations between multiple target properties of interest
- **Equivariant message passing layers** to take advantage of symmetries in the data
- **Transferable Learning** for extrapolation of accurate predictions from smaller to larger atomic systems
- **Scalable training with Distribute Data Parallelism (DDP)** for large scale training on massive volumes of data

HydraGNN: Multi-Task Learning (MTL)

HydraGNN: Multi-Task Learning

Condensed Matter Physics: Solid Solution Alloys

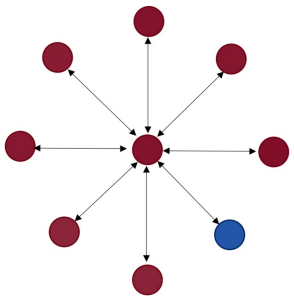
L.P., M. et al., *Multi-task graph neural networks for simultaneous prediction of global and atomic properties in ferromagnetic systems,* https://iopscience.iop.org/article/10.1088/2632-2153/ac6a51/meta

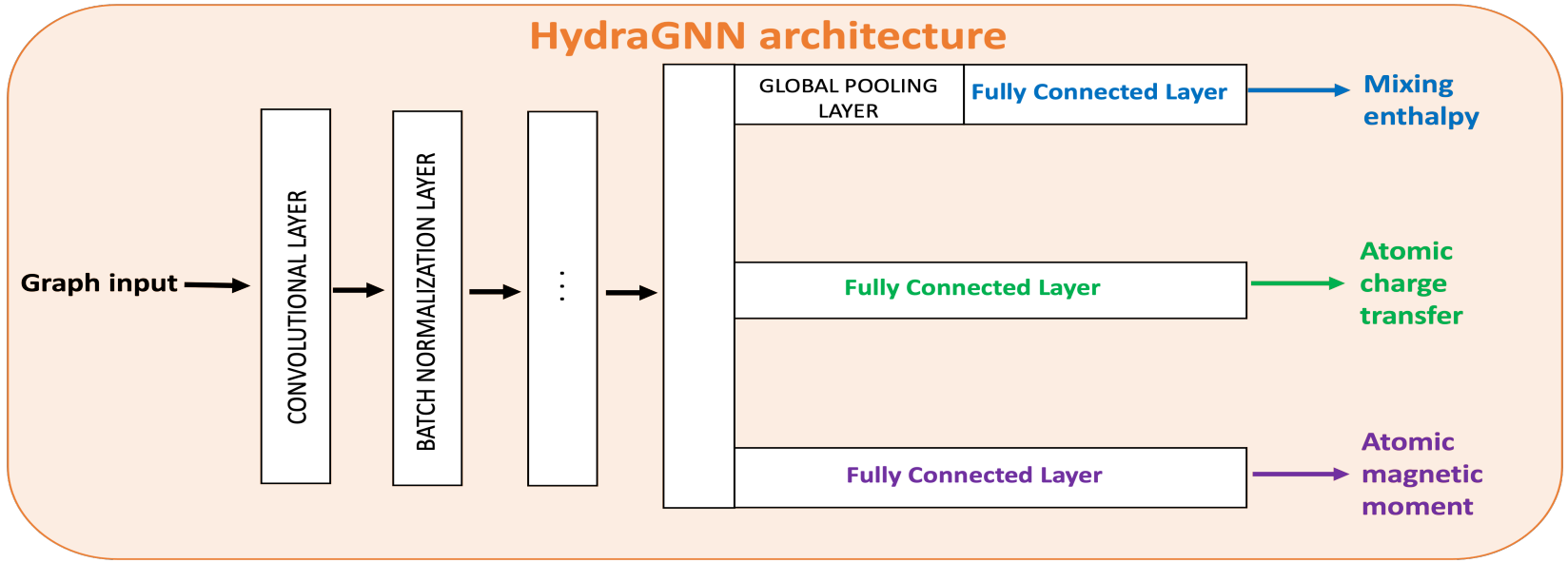


MTL uses one single HydraGNN to *simultaneously* predict all material properties of interest

MTL uses each property as a mutual regularizer of the others, thereby counteracting the curse of dimensionality when data is provided in small volume and defined in a high dimensional space

Conversion of lattice structure into a graph





HydraGNN: Multi-Task Learning - Numerical Results - FePt

The GCNN models are implemented in PyTorch

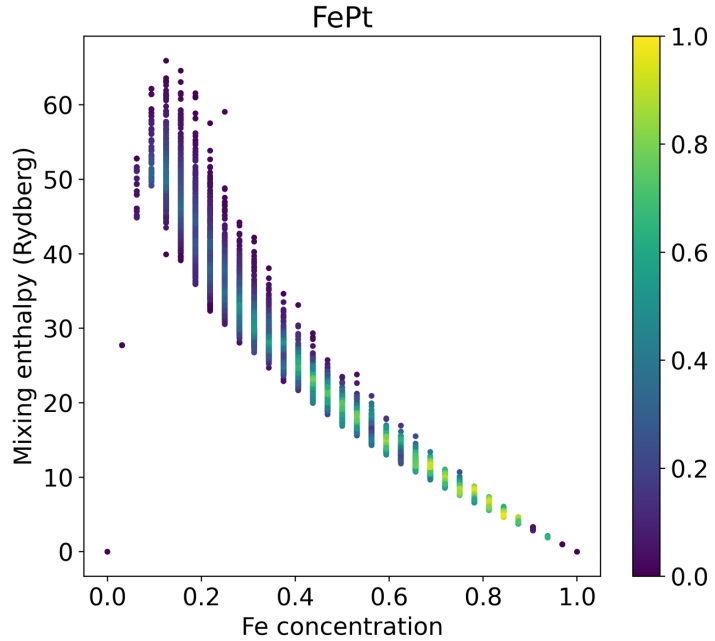
70 % - 15 % - 15 % splitting of dataset between training, validation, and testing

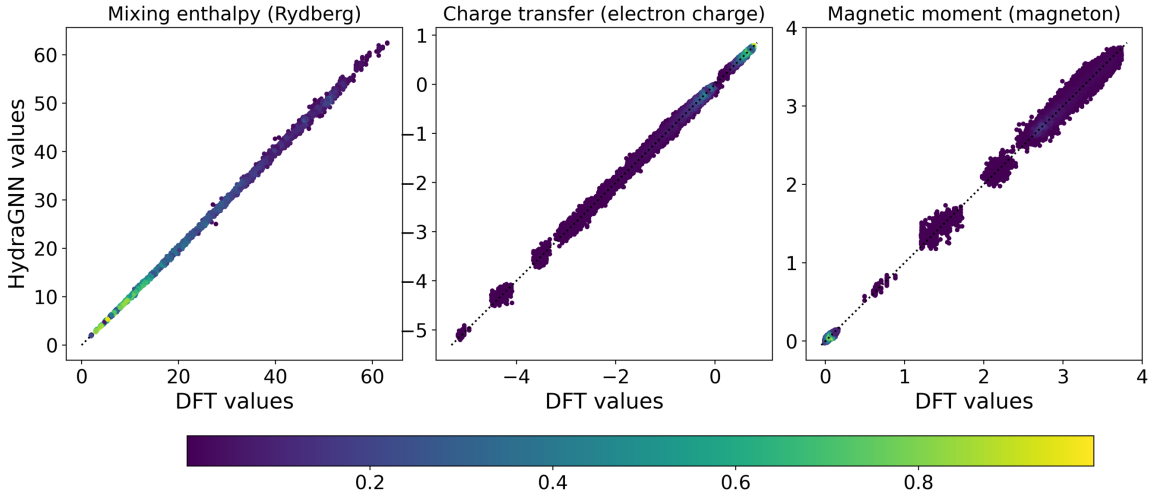
The splitting is stratified across compositions

Architecture: 6 Conv layers, 20 channels for each conv layer
every head: 2 MLP layers, 50 neurons per layer

Adam used as optimizer for training with learning rate = 1e-4

No hyperparameter tuning performed





PNA aggregation outperformed GIN and GAT in accuracy

	Test RMSE		
Training method	**Mixing enthalpy**	**Charge transfer**	**Magnetic moment**
MTL, HCM	$7.54e{-3} \pm 8.70e{-4}$	$6.77e{-3} \pm 3.59e{-4}$	$1.04e{-2} \pm 4.94e{-4}$
MTL, HC	$7.33e{-3} \pm 4.77e{-4}$	$7.36e{-3} \pm 3.23e{-4}$	-
MTL, HM	$6.64e{-3} \pm 5.08e{-4}$	-	$1.02e{-2} \pm 5.23e{-4}$
MTL, CM	-	$5.94e{-3} \pm 3.02e{-4}$	$9.30e{-3} \pm 4.12e{-4}$
STL, H	$1.02e{-2} \pm 1.16e{-3}$	-	-
STL, C	-	$5.94e{-3} \pm 4.39e{-4}$	-
STL, M	-	-	$8.77e{-3} \pm 3.18e{-4}$

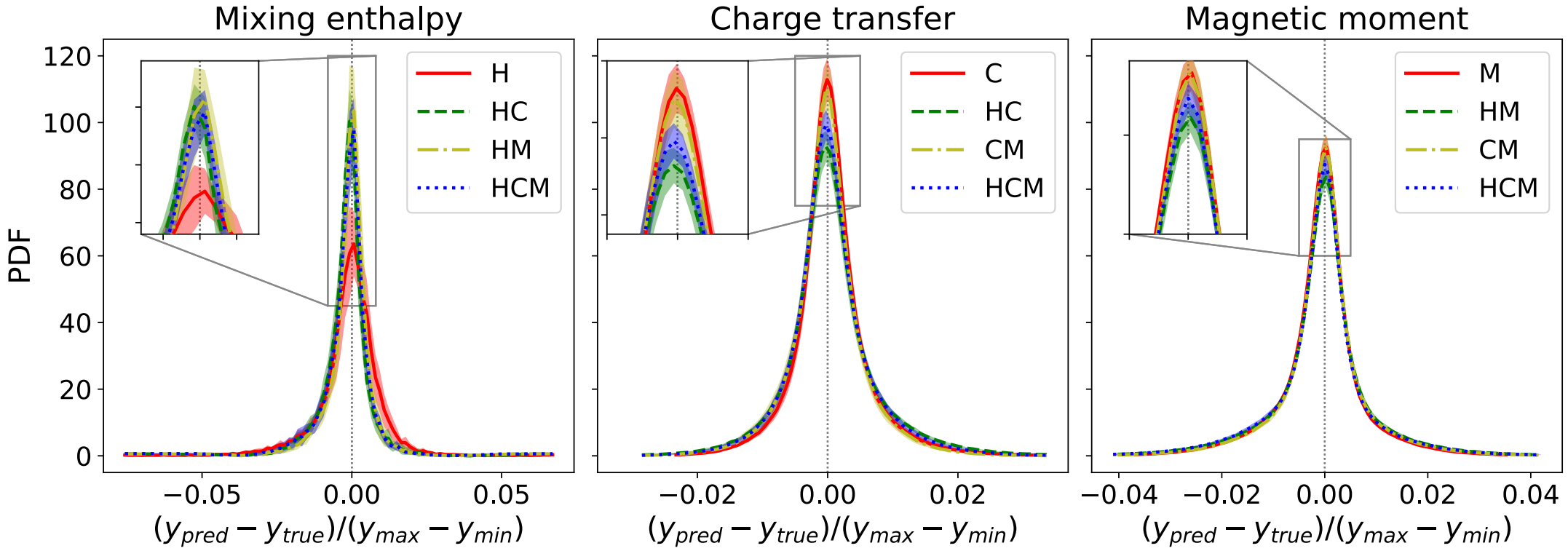
HydraGNN: Multi-Task Learning

Numerical Results - FePt

Probability distributions functions of prediction errors for mixing enthalpy (left), atomic charge transfer (center), and atomic magnetic moment (right)

H = single prediction on mixing enthalpy
C = single prediction on charge transfer
M = single prediction on magnetic moment

HM = prediction on mixing enthalpy + magnetic moment
HC = prediction on mixing enthalpy + charge transfer
HCM = prediction on all three properties



HydraGNN: Equivariant Architecture

HydraGNN: Equivariant Architecture

Numerical results – Ultraviolet-visible (UV-vis) spectrum of organic molecules

DL architectures are equivariant if their behavior is consistent under rotations and symmetries applied to the input graph.

Equivariant features can reduce the amount of data required by DL models to reach a desired accuracy. Even if UV-vis spectrum is invariant, enforcing equivariance in the message passing allows to transfer the learnt DL embedding to other predictive tasks where there target property is equivariant.

UV-vis spectroscopy measures the amount of light at ultraviolet or visible wavelengths absorbed by or transmitted through a sample (in our case, a molecule)
Experimental measurement of UV-vis spectrum are obtained by hitting a molecule with beams of light in the UV-vis range (vis. range: [380 nm, 750nm] – UV range: [100nm, 400 nm])

The UV-vis spectrum is used to:

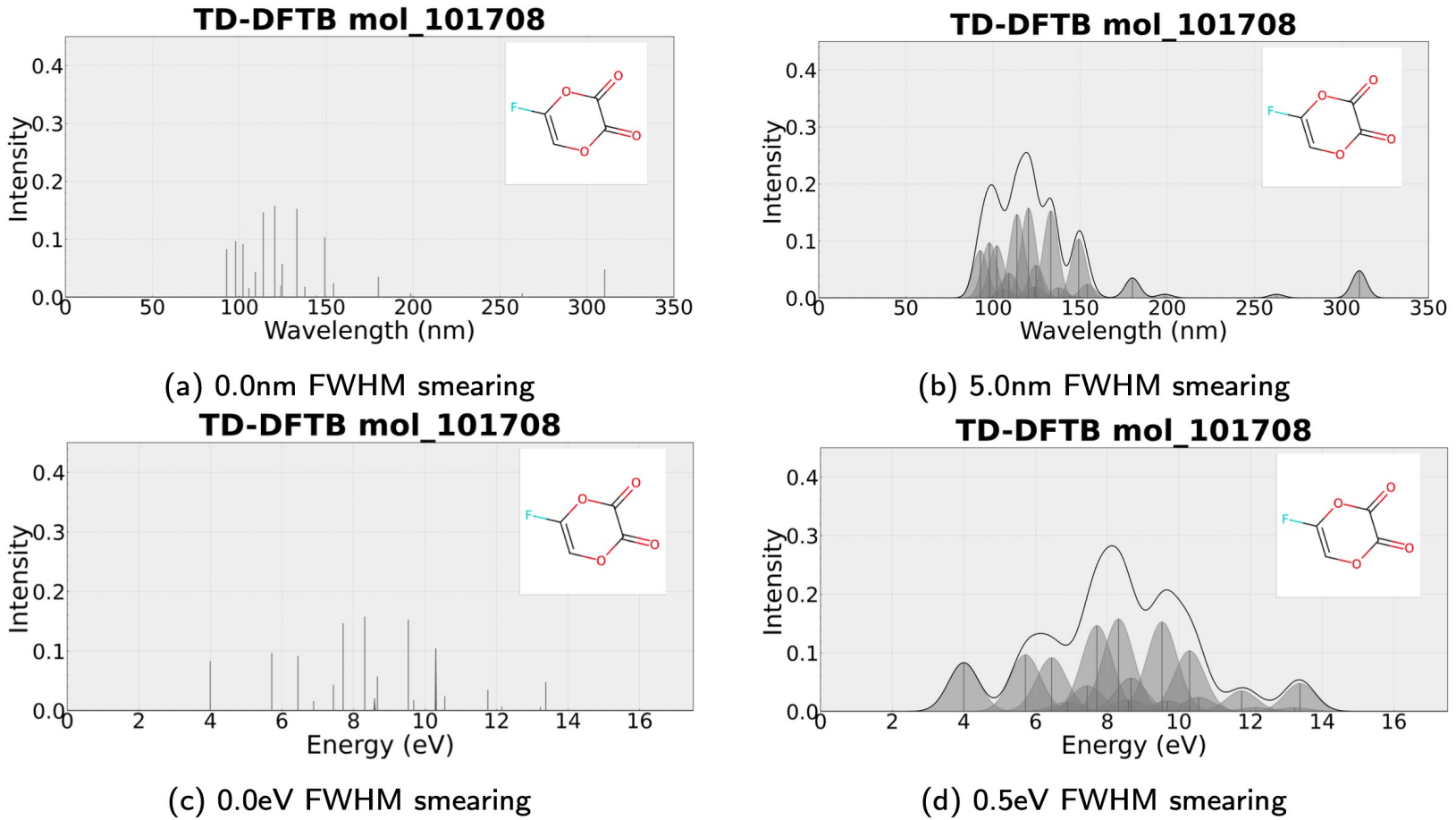
- Identify chemical species present in a material or produced during a chemical reactions
- Grow bacteria for study
- Perform drug design by identifying molecules that are easy to engage in chemical reactions
- Quantify the number of nucleic acids (DNA or RNA) to determine their average concentrations in a mixture, as well as their purity

Numerical methods of UV-vis spectrum require running time-dependent density functional theory calculations

HydraGNN: Equivariant Architecture

Numerical results – Ultraviolet-visible (UV-vis) spectrum of organic molecules

GDB-9-Ex : Electronic excitation spectrum for GDB-9 molecules https://www.osti.gov/biblio/1890227
M. Lupo Pasini et al. *Two excited-state datasets for quantum chemical UV-vis spectra of organic molecules,* Nature Scientific Data, Volume 10, Issue 546



(a) 0.0nm FWHM smearing

(b) 5.0nm FWHM smearing

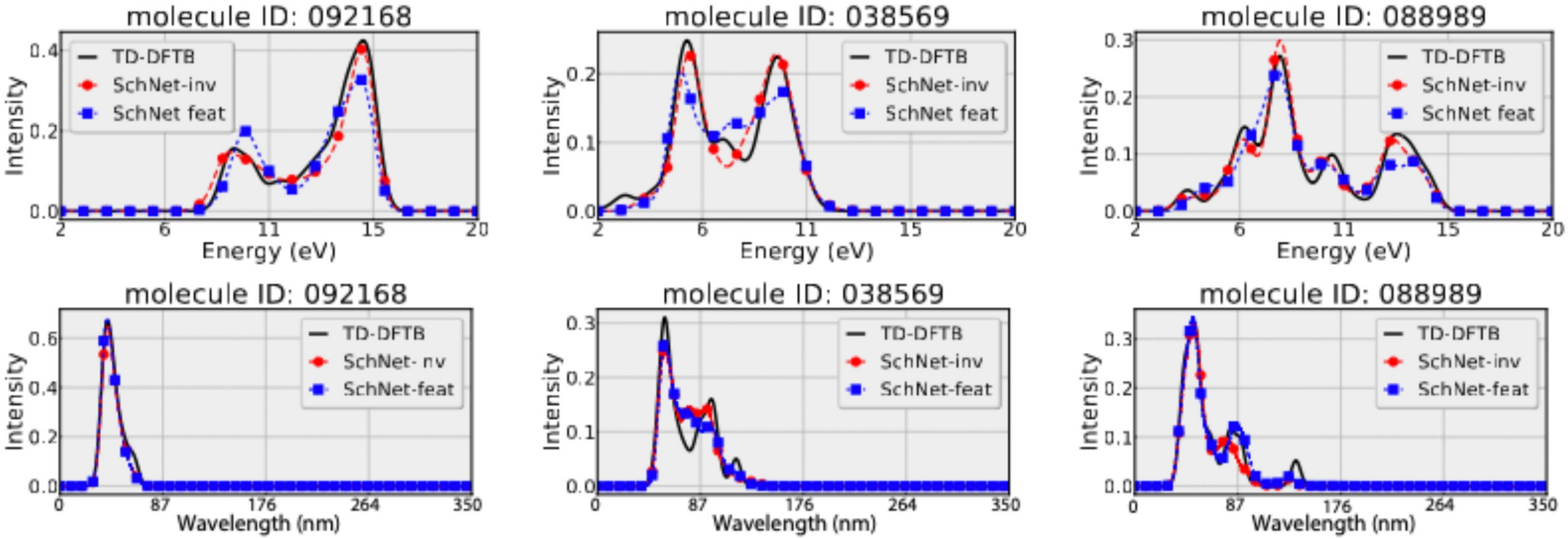
(c) 0.0eV FWHM smearing

(d) 0.5eV FWHM smearing

Figure: Electronic excitation spectrum of C_4HFO_4 plotted in wavelength(top) and energy(bottom) with exact peaks (left) and Full Width Half Maximum (FWHM) smearing (right).

HydraGNN: Equivariant Architecture

Numerical results – Ultraviolet-visible (UV-vis) spectrum of organic molecules



HydraGNN: Transferable Learning

HydraGNN: Transferable Learning

Open-source dataset

Solid Solution Nickel-Platinum (NiPt)

ORNL_AISD_NiPt https://www.osti.gov/biblio/1958172

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**

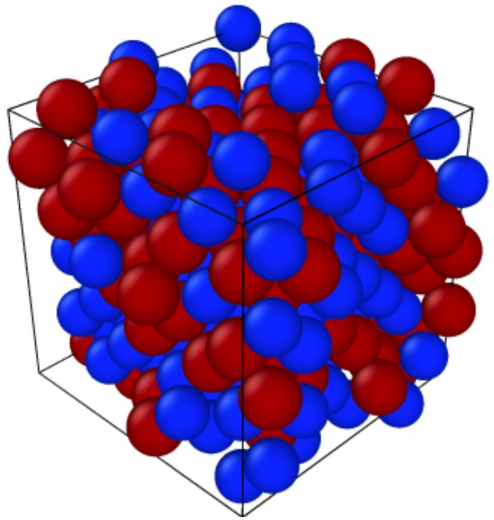
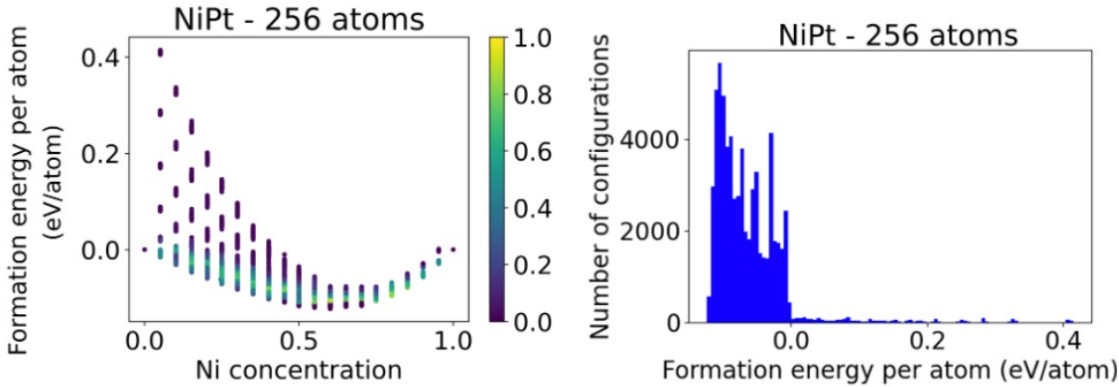
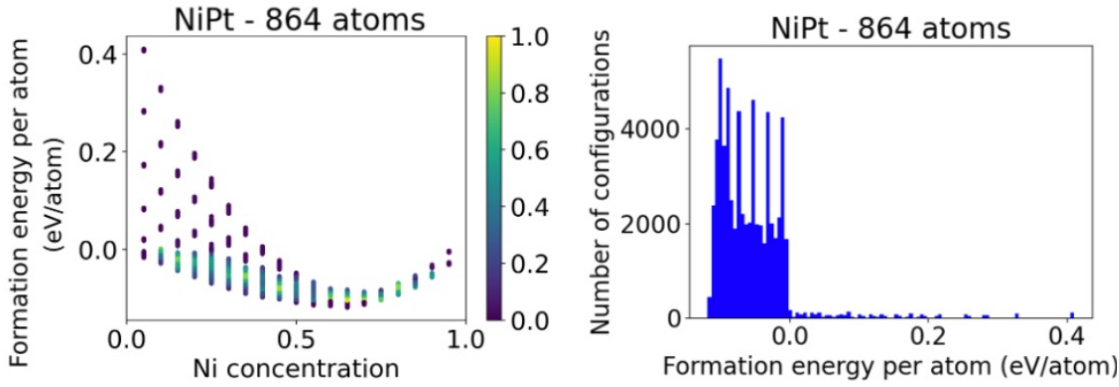
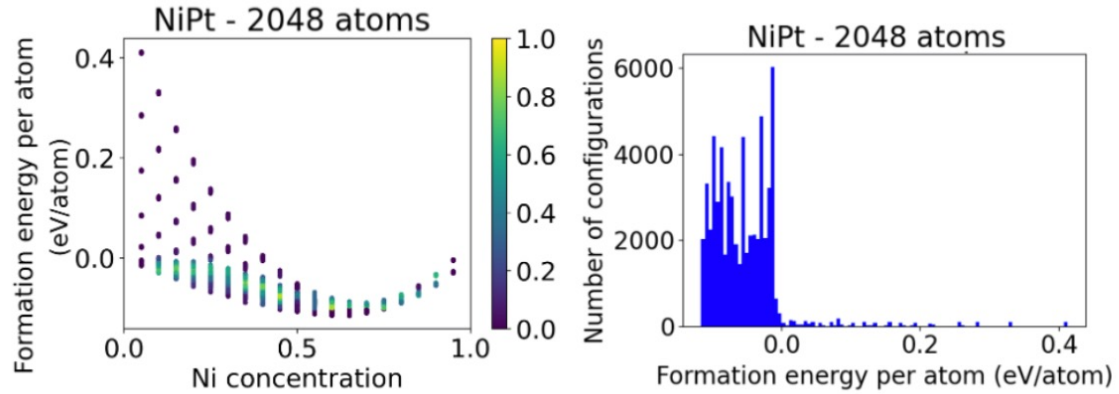


Figure: nickel-platinum (NiPt) solid solution binary alloy with 50 % chemical concentration of Ni and 50 % chemical concentration of Pt.



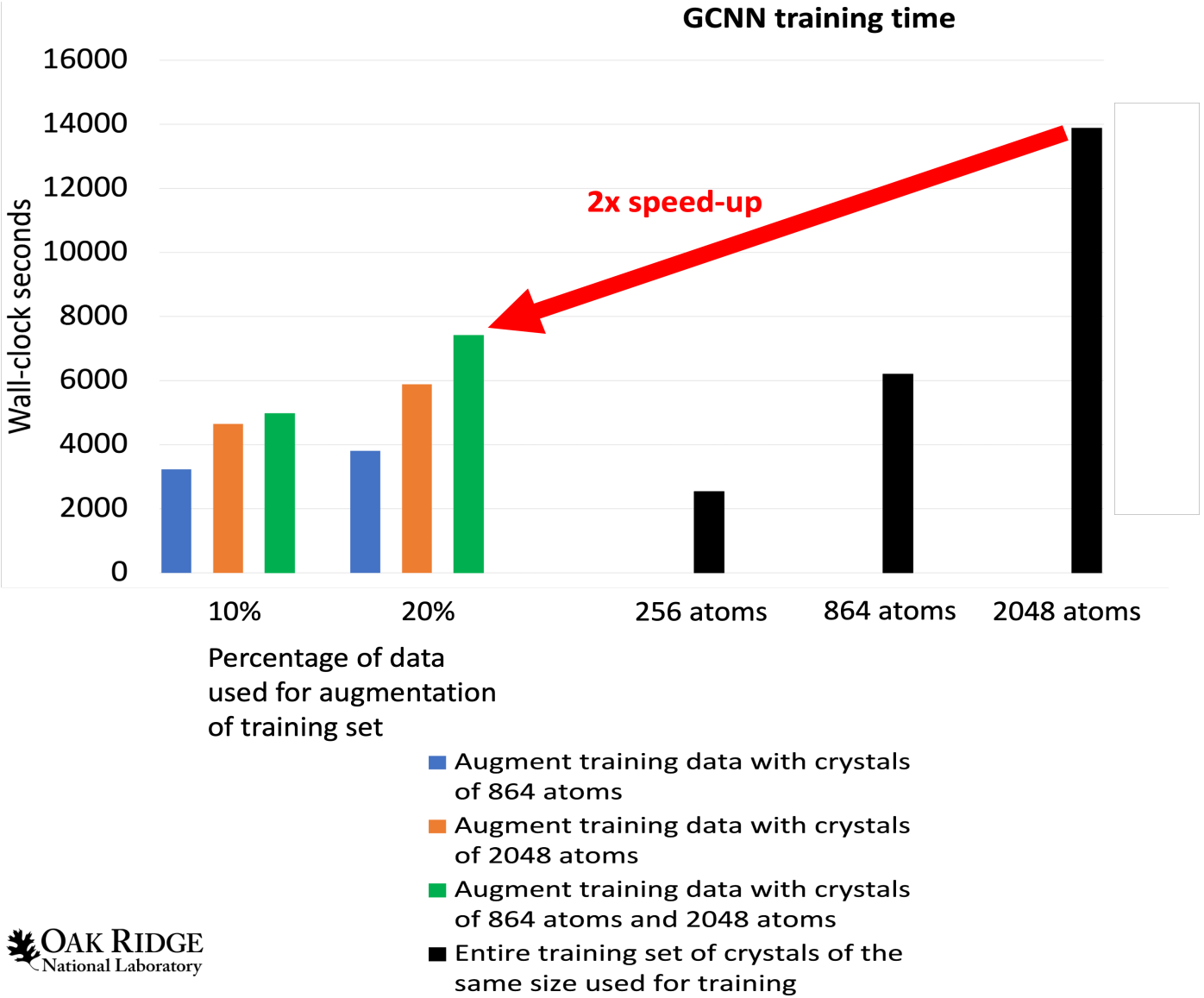


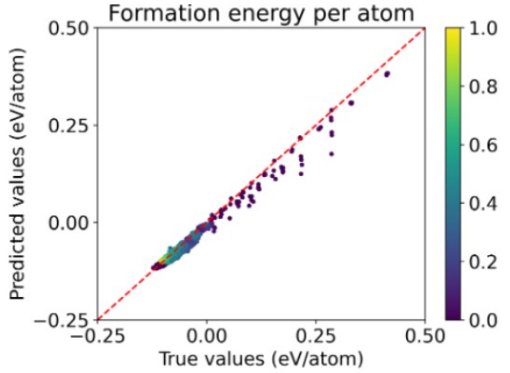


OAK RIDGE National Laboratory

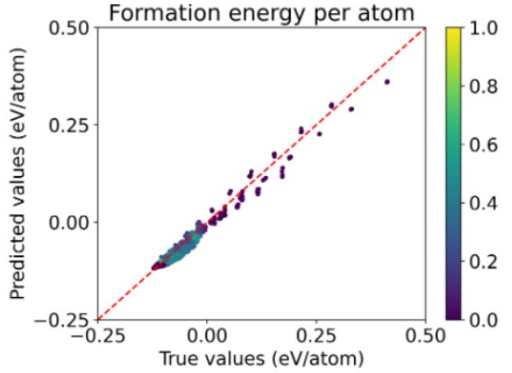
Numerical Results

Transfer predictions of GCNN trained from smaller lattices to larger lattices extrapolate

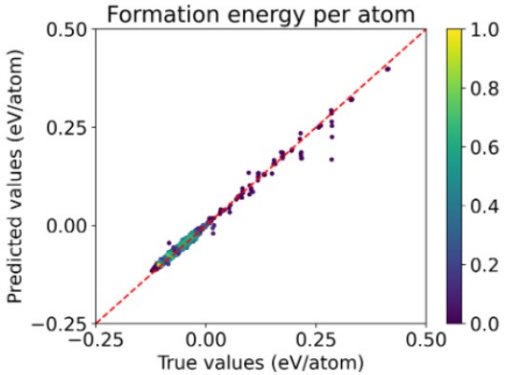




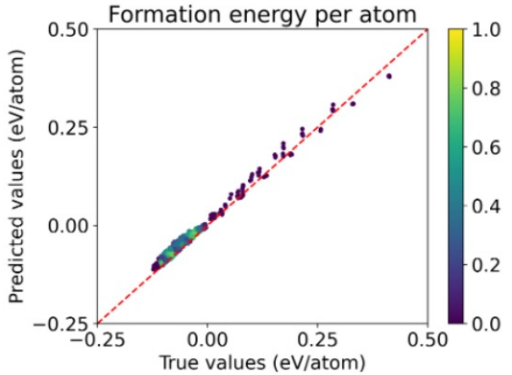
Train on 256 atoms - Test on 864 atoms



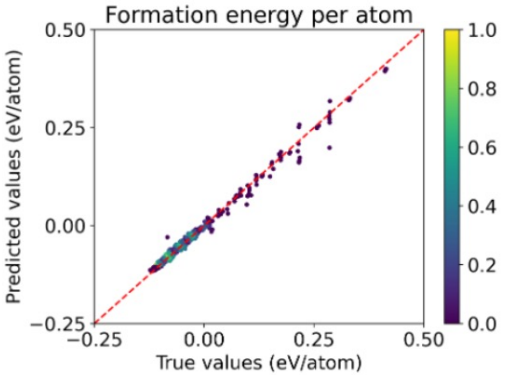
Train on 256 atoms - Test on 2,048 atoms



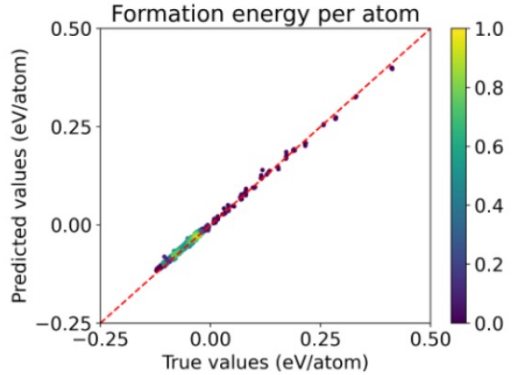
Train on 256 atoms + 10% 864 atoms
Test on 864 atoms



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



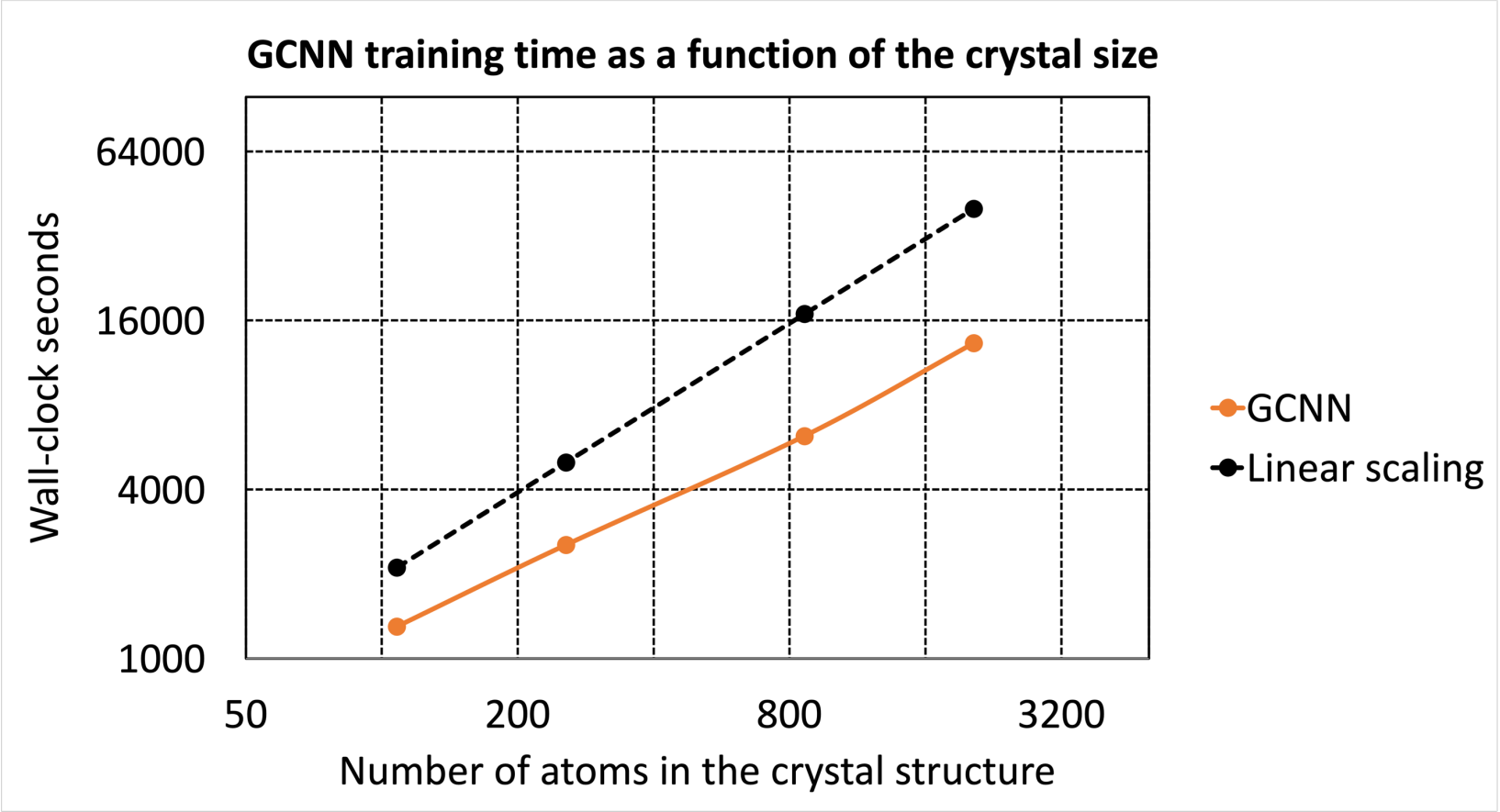
Train on 256 atoms + 20% 864 atoms
+ 20% 2,048 atoms
Test on 864 atoms



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

HydraGNN: Transferable Learning
Numerical Results: Solid Solution NiPt

Linear scaling of GCNN training time with respect to lattices of increasing size

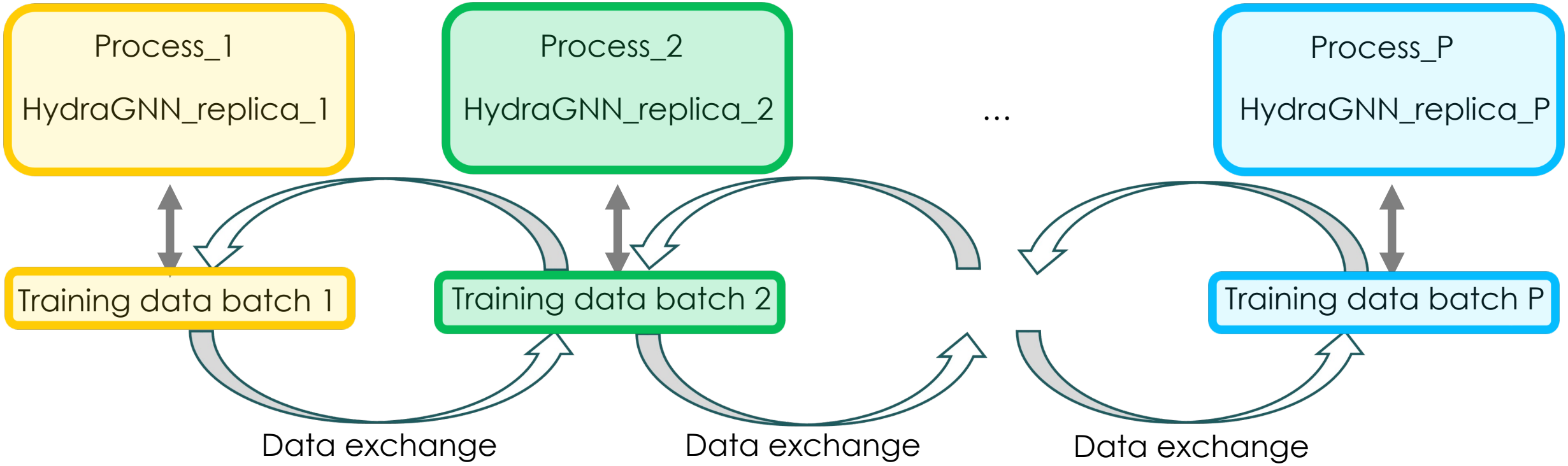


HydraGNN: Scalable Training

HydraGNN: Distributed Training with Distributed Data Parallelism

Intrinsic Multi-Tasking allows for Algorithmic Scalability

Distributed Data Parallelism allows for High-Performance Computing (HPC) scalability



HydraGNN: Distributed Training with Distributed Data Parallelism
Portability and scalability across diverse HPC environments

Datasets:

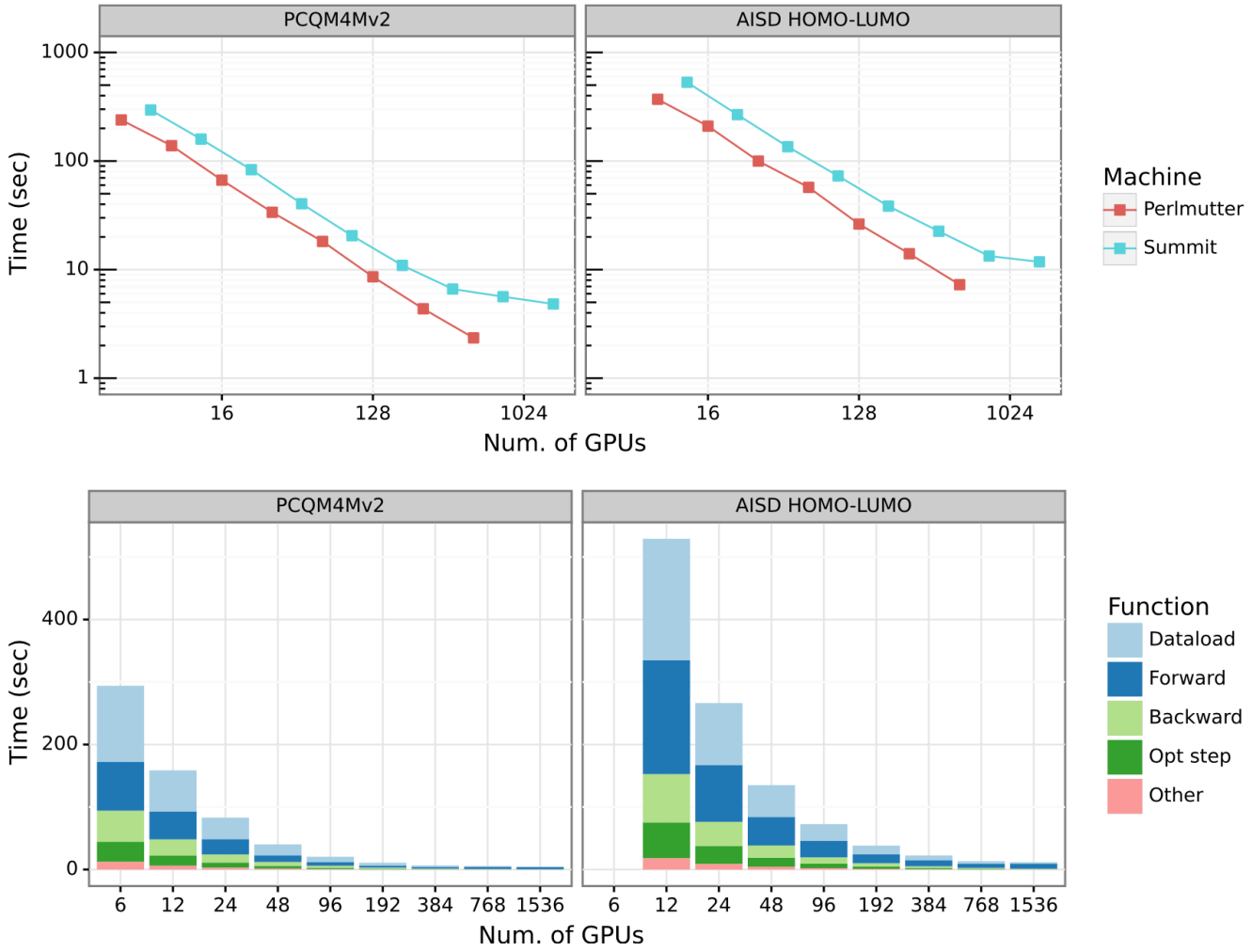
- **PCQM4Mv2 (~ 3 million molecules)**
https://ogb.stanford.edu/docs/lsc/pcqm4mv2/

- **AISD HOMO-LUMO (~ 10 million molecules)**
https://www.osti.gov/dataexplorer/biblio/dataset/1869409

ADIOS2 library is used for scalable data reading

Distributed Data Parallelism is used for scalable training of the HydraGNN model

Results: linear scaling of data reading + training using up to 1,024 NVIDIA V100 GPUs on OLCF Summit and 1,024 NVIDIA GPUs on NERSC Perlmutter



Summary

HydraGNN is an ORNL branded AI architecture developed within the AISD thrust of the AI Initiative that simultaneously:

- Performs multi-task learning https://iopscience.iop.org/article/10.1088/2632-2153/ac6a51/meta
- Exploits both explicit and implicit correlations to stabilize the training
- Allows for flexible choice of information exchange policies implemented in message passing layers
- Includes equivariance
- Scales linearly on leadership-class supercomputing facilities
- Ports training seamlessly on various computing platforms

HydraGNN has been used for:

- Successful applications to material science, molecular design, neutron spectroscopy, and structural engineering

These enhancements and applications show HydraGNN's relevance to the lab's mission

- Strong integration with the Design product, the ASCR program, and ORNL facilities

Future Work

HydraGNN training for imbalanced multi-source multi-fidelity data

- Train HydraGNN model to predict UV-vis spectra of organic molecules using
  - Low fidelity data → time-dependent density functional tight-binding (TD-DFTB)
  - Intermediate fidelity data → time-dependent density functional theory (TD-DFT)
  - High fidelity data → Equation of motion coupled cluster singles and doubles (EOM-CCSD)

Generative Models

- Develop generative diffusion models in HydraGNN to perform scalable and robust exploration of new chemical compounds

Collaborators

Oak Ridge National Laboratory
- Frank Liu
- Stephan Irle
- Ayana Gosh
- Kadir Amasyali
- Pilsun Yoo
- Kshitij Mehta
- Paul Laiu
- Pei Zhang
- Jong Youl Choi
- Brett Eiffert
- Zach Fox
- Debsindhu Bhowmik
- John Gounley
- Gang Seob Jung
- Dongwon Shin
- Ying Yang
- German Samolyuk
- Samuel T. Reeve
- Andrew E. Blanchard
- Junqi Yin
- Cory Hauck

Lawrence Berkeley National Laboratory
- Khaleed Ibrahim
- Jonghyun Bae

University of Utah (Salt Lake City)
- Justin Baker

Politecnico di Milano
- Simona Perotto
- Nicola Ferro

Acknowledgments

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

Research partially sponsored by the US-DOE Advanced Scientific Computing Research (ASCR) under contract DE-SC0023490.

This work used resources of the Oak Ridge Leadership Computing Facility (OLCF) and of the Edge Computing program at ORNL. Computer time was provided by the OLCF Director's Discretion Project program under the OLCF awards MAT250 and LRN026.

Thank you!

Questions?

Examples of interest to US-DOE

1. **Organic molecular compounds**
   - Small molecules (with at most hundreds of atoms)
   - Proteins (typically contain at least thousands of atoms)
   - Nucleic acids

2. **Inorganic compounds**
   - Alloys
   - Zeolites

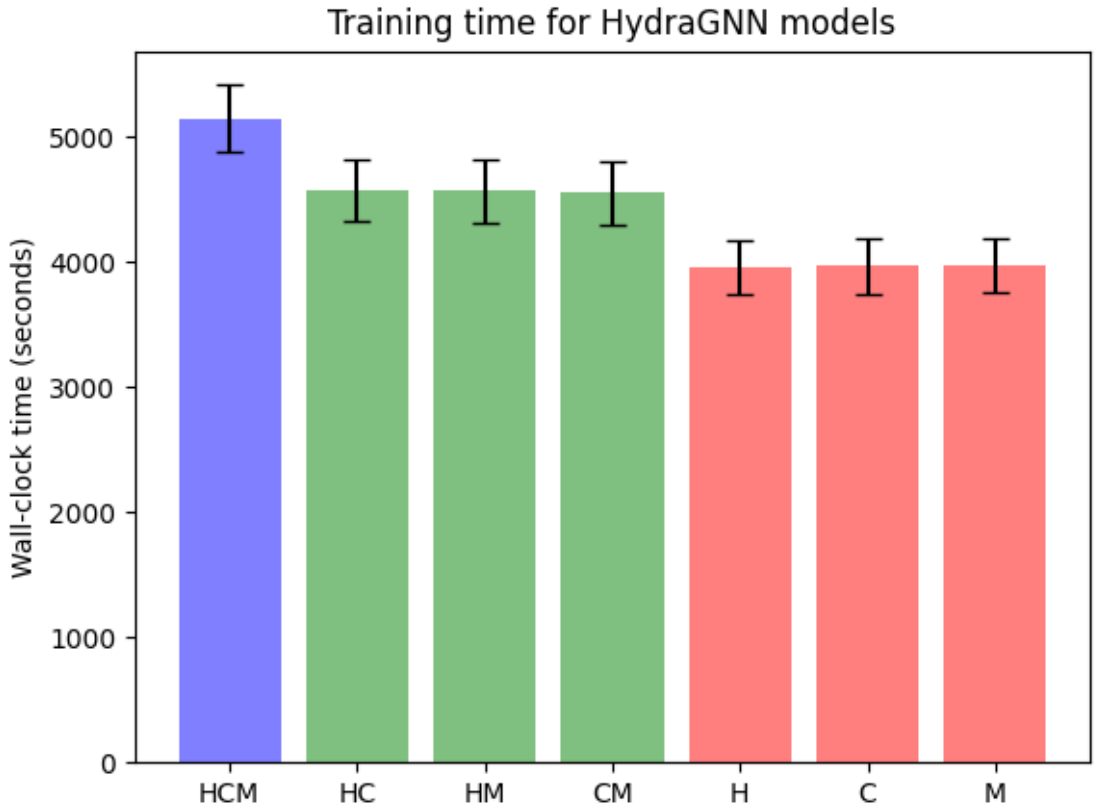
Examples of targeted applications include, but are not limited to:

i. renewable energy (e.g., solar cells, organic photovoltaics, and organic light-emitting diodes)
ii. energy storage (e.g., batteries and supercapacitors)
iii. carbon capture and sequestration
iv. materials innovation (e.g., drugs, or materials with desired conductivity, thermal stability, and catalytic activity)
v. Genomics (protein synthesis)

HydraGNN: Multi-Task Learning
Numerical Results - FePt

Tot. training time = **Training time of convolutional layers** + (**Number of heads** X **training time for each head**)

Training time of convolutional layers does NOT change with the number of properties simultaneously predicted → **computational savings**



H = single prediction on mixing enthalpy
C = single prediction on charge transfer
M = single prediction on magnetic moment

HM = prediction on mixing enthalpy + magnetic moment
HC = prediction on mixing enthalpy + charge transfer
HCM = prediction on all three properties

Blue: wall-clock time to train HydraGNN on all three material properties

Green: wall-clock time to train HydraGNN on two material properties

Red: single task training

HydraGNN: Equivariant Architecture

Definition: E(3): Euclidean Group on $\mathbb{R}^3$

The Euclidean group on $\mathbb{R}^3$ is the collection of all **reflections, rotations and translations**.

Elements of E(3) can be represented by

- orthogonal rotation matrices $\boldsymbol{Q} \in \mathbb{R}^{3\times 3}$
- translational vectors $\boldsymbol{b} \in \mathbb{R}^3$.

Application: Molecular Graphs

For a molecular structure, we are given the following:

- a connectivity graph $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ representing atoms and atomic interactions as nodes $\mathcal{V}$ and edges $\mathcal{E}$ respectively, where $n = |\mathcal{V}|$.
- several (h) chemical properties of each atom $\boldsymbol{H} \in \mathbb{R}^{h\times n}$.
- the position of each atom $\boldsymbol{X} \in \mathbb{R}^{3\times n}$.

HydraGNN: Equivariant Architecture

Definition: Invariance and Equivariance

Recall that the E(3) group action is defined by rotation, reflections and translations. Given the GNN update equation

$$\boldsymbol{H}^{l+1}, \boldsymbol{X}^{l+1} = \text{GNN}(\boldsymbol{H}^{l}, \boldsymbol{X}^{l}, \boldsymbol{A})$$

A GNN is **invariant** under the E(3) group action if it satisfies the following.

$$\boldsymbol{H}^{l+1}, \boldsymbol{X}^{l+1} = \text{GNN}(\boldsymbol{H}^{l}, \boldsymbol{Q}\boldsymbol{X}^{l} + \boldsymbol{b}, \boldsymbol{A})$$

A GNN is **equivariant** under the E(3) group action if it satisfies the following.

$$\boldsymbol{H}^{l+1}, \boldsymbol{Q}\boldsymbol{X}^{l+1} + \boldsymbol{b} = \text{GNN}(\boldsymbol{H}^{l}, \boldsymbol{Q}\boldsymbol{X}^{l} + \boldsymbol{b}, \boldsymbol{A})$$

Application: Molecular Properties

- **Invariant Properties:** HOMO-LUMO gap, free energy, excitation spectrum, ...
- **Equivariant Properties:** electron charge density, inter molecular forces, ...

HydraGNN: Transferable Learning
Numerical Results: Solid Solution NiPt

Baseline training with GCNN model trained and validated on lattices of the same size

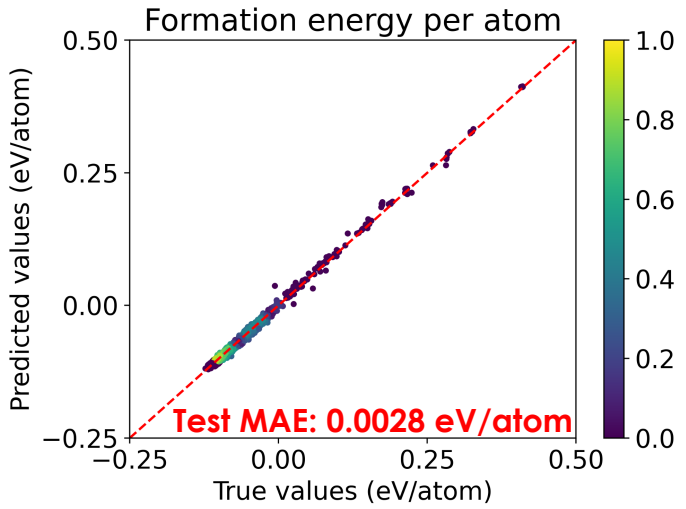
Dataset splitting: 80% training – 10% validation – 10% testing

GCNN architecture:

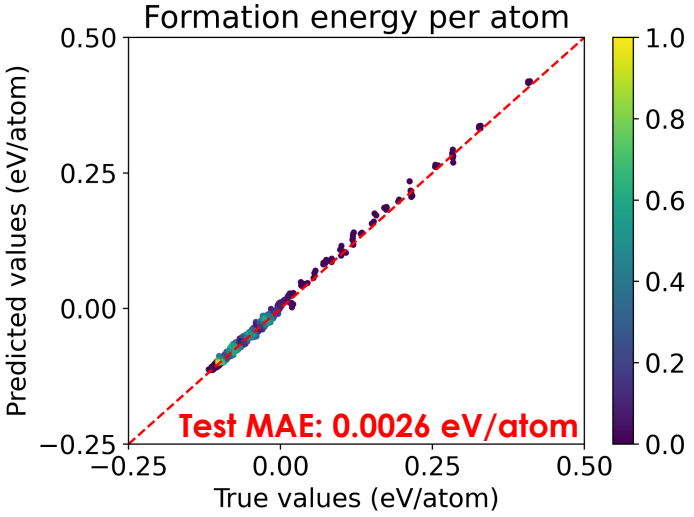
- Radius cutoff of 5.0 Angstrom for graph connectivity
- 6 principal neighborhood aggregation (PNA) convolutional layers
- 3 fully connected layers
- 50 neurons for each hidden layer

Training:

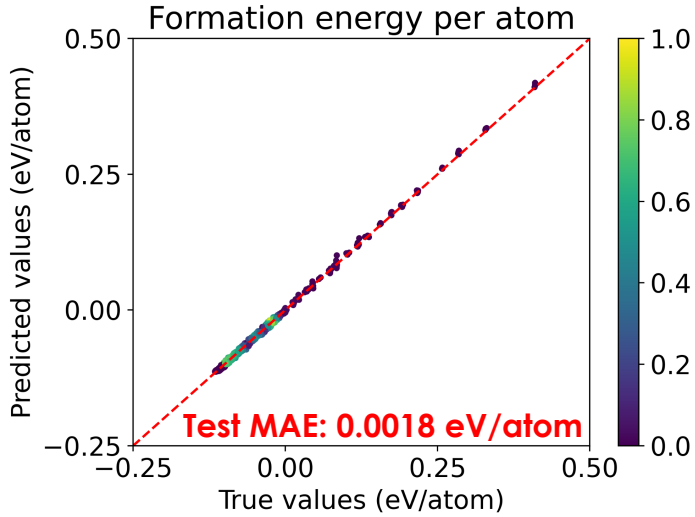
- Stochastic optimizer: AdamW
- Number of epochs: 50
- Batch size: 32 samples
- Early stopping patience: 10 epochs



256 atoms



864 atoms



2,048 atoms

6. HydraGNN: reliable uncertainty quantification over vast regions of the chemical space

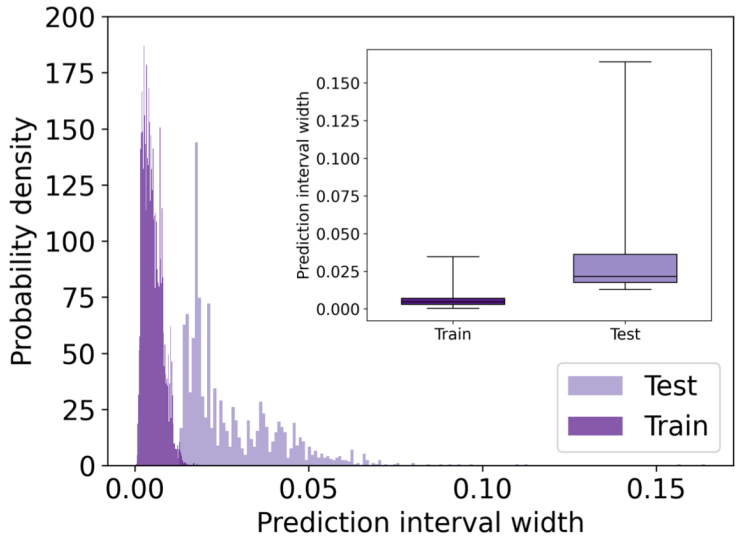
Uncertainty quantification for prediction of formation energy of solid solution alloys and organic molecules

AISD Surrogates members: Max Lupo Pasini, Pei Zhang
Assurance Thrust: Samuel Temple Reeve, Siyan Liu, Dan Lu

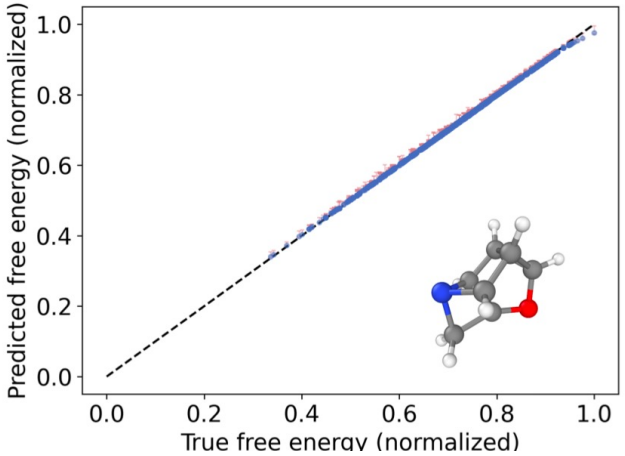
Datasets: Organic molecules - QM9

Molecules with fluorine were not used for training

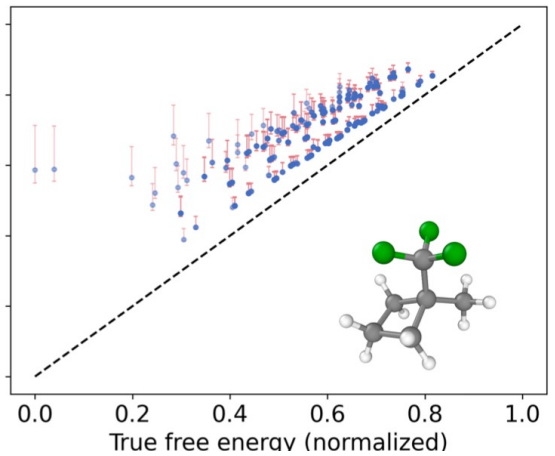
Results:
During validation, PI3NN helps HydraGNN capture out of distribution samples when molecules with fluorine are passed to the model for inference



In-distribution UQ for molecules **without fluorine**



Out-of-distribution UQ for molecules **with fluorine**



7. Released Datasets from May 2022 through December 2022

- GS Jung, M. Lupo Pasini, S. Irle, ORNL_AISD_NiNb, https://www.osti.gov/dataexplorer/biblio/dataset/1890159 United States: N. p., 2022. Web. doi:10.13139/OLCF/1890159.

  Contains 109,932 atomic configurations
- M. Lupo Pasini, P. Yoo, K. Mehta S. Irle, GDB-9-Ex: Quantum chemical prediction of UV/Vis absorption spectra for GDB-9 molecules. https://www.osti.gov/dataexplorer/biblio/dataset/1890227

  Contains 96,766 organic molecules
- M. Lupo Pasini, K. Mehta, P. Yoo, S. Irle, ORNL_AISD-Ex: Quantum chemical prediction of UV/Vis absorption spectra for over 10 million molecules. (Submitted for review to OLCF)

  Contains 10,502,000 million molecules
- M. Karabin, M. Lupo Pasini, Markus Eisenbach, ORNL_AISD_NiPt: data for solid solution binary alloy NiPt (In prerapation)

  Contains 240,000 atomic configurations

Publications from October 2021 through May 2022

Collaboration with OLCF on scalable training of generative models:

M. Lupo Pasini, J. Yin, *Stable Parallel Training of Wasserstein Conditional Generative Adversarial Neural Networks,* **Published**, Computational Science & Computational Intelligence (CSCI'21) https://ieeexplore.ieee.org/document/9799213

Surrogates model for microscale - HydraGNN:

1. M. Lupo Pasini, M. Burcul, S. T. Reeve, M. Eisenbach, S. Perotto, *Fast and accurate predictions of total energy for solid solution alloys with graph convolutional neural networks*, **Published**, Smoky Mountain Conference 2021 https://link.springer.com/chapter/10.1007/978-3-030-96498-6_5
2. M. Lupo Pasini, V. Reshniak, M. Stoyanov, *Anderson Acceleration for Distributed Training of Deep Learning Models*, **Published**, IEEE South Eeast Conference 2022 https://ieeexplore.ieee.org/document/9763953
3. M. Lupo Pasini, P. Zhang, S. T. Reeve, J. Y. Choi, *Multi-task graph neural networks for simultaneous prediction of global and atomic properties in ferromagnetic systems,* **Published**, Machine Learning: Science and Technology https://iopscience.iop.org/article/10.1088/2632-2153/ac6a51/meta
4. P. Laiu, Y. Yang, J. Y. Choi, M. Lupo Pasini, D. Shin, *A Neural Network Approach to Predict Gibbs Free Energy of Multi-component Solid Solutions*, **Published,** Journal of Phase Equilibria and Diffusion Kinetics
5. A, E. Blanchard, P. Zhang, K. Mehta, D. Bhowmik, J. Gounley, S. T. Reeve, S. Irle, and M, Lupo Pasini, *Computational Workflow for Accelerated Molecular Design Using Quantum Chemical Simulations and Deep Learning Models*, Smoky Mountain Conference 2022, **Accepted,** Smoky Mountain Conference 2022
6. M. Eisenbach, M. Karabin, M. Lupo Pasini, J. Yin, *Statistical Mechanics of Materials using First Principles Calculations and Machine Learning*, Smoky Mountain Conference 2022, **Accepted,** Smoky Mountain Conference 2022

Publications from May 2022 through December 2022

Surrogate models for mesoscale:

Paul Laiu, Ying Yang, Massimiliano Lupo Pasini, Jong Youl Choi, Dongwon Shin,
A Neural Network Approach to Predict Gibbs Free Energy of Ternary Solid Solutions, Journal of Phase Equilibria and Diffusion, **Accepted**

Surrogates model for microscale - HydraGNN:

1. Jong Youl Choi, P. Zhang, K. Mehta, A. Blanchard, M. Lupo Pasini, *Scalable training of graph convolutional neural networks for fast and accurate predictions of HOMO-LUMO gap in molecules,* **Published**, Journal of Cheminformatics 14(70), 2022. https://doi.org/10.1186/s13321-022-00652-1
2. M. Eisenbach, M. Karabin, M. Lupo Pasini, J. Yin, *Statistical Mechanics of Materials using First Principles Calculations and Machine Learning*, Smoky Mountain Conference 2022, **Accepted**
3. S. T. Reeve, P. Zhang, M. Lupo Pasini, D. Lu, *Uncertainty quantification for atomic predictions from graph convolutional neural networks*, **Submitted** to Modelling and Simulation in Materials Science and Engineering
4. M. Lupo Pasini, GS Jung, S. Irle, *Graph neural networks predict energetic and mechanical properties for models of solid solution metal alloy phases*, **Submitted** to Computational Materials Science