ML Commons

Science MLCommons Working Group

Science Working Group contribution to MLCommons Community Meeting December 9 2021

Gregg Barrett,
Wahid Bhimji,
Bala Desinghu,
Murali Emani,
Geoffrey Fox,
Grigori Fursin,
Tony Hey,
David Kanter,
Christine Kirkpatrick,
Hai Ah Nam,
Juri Papay,
Amit Ruhela,
Mallikarjun Shankar,
Jeyan Thiyagalingam
Aristeidis Tsaris,
Gregor von Laszewski,
Feiyi Wang,
Junqi Yin,

attended September--November meetings;
Blue co-chairs

MLCommons (MLPerf) Consortium Deep Learning Benchmarks

- Major effort of 52 companies to produce benchmarks with ongoing challenges
- Training at V1.0 (fourth release)
- Fox set up Science Working Group with co-chair Tony Hey who has a significant benchmarking group SciML
  - Identified ~12 science benchmarks including light source, satellite, surrogate and time series
- MLCommons aims to accelerate machine learning innovation to benefit everyone. Benchmarking, Datasets, Best Practices **Total Effort ~50 FTE**

Some Relevant Working Groups

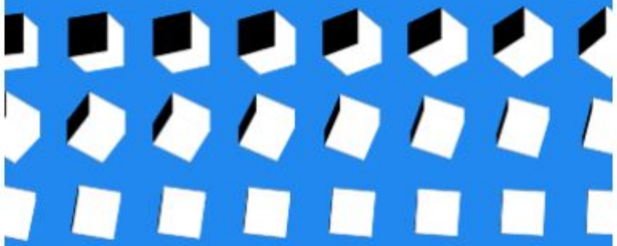
- Training
- Inference (Batch and Streaming)
- TinyML (embedded)
- Power
- Datasets
- HPC (Supercomputer Implementations)
- Research (Academic-Industry Links)
- Science (AI for Science)
- Best Practice (Software)
- Logging/Infrastructure (metadata)

MLCommons (MLPerf) Consortium Activity Areas

Benchmarking

Benchmarks provide consistent measurements of accuracy, speed, and efficiency. Consistent measurements enable engineers to design reliable products and services, and enable researchers to compare innovations and choose the best ideas to drive the solutions of tomorrow.

Datasets



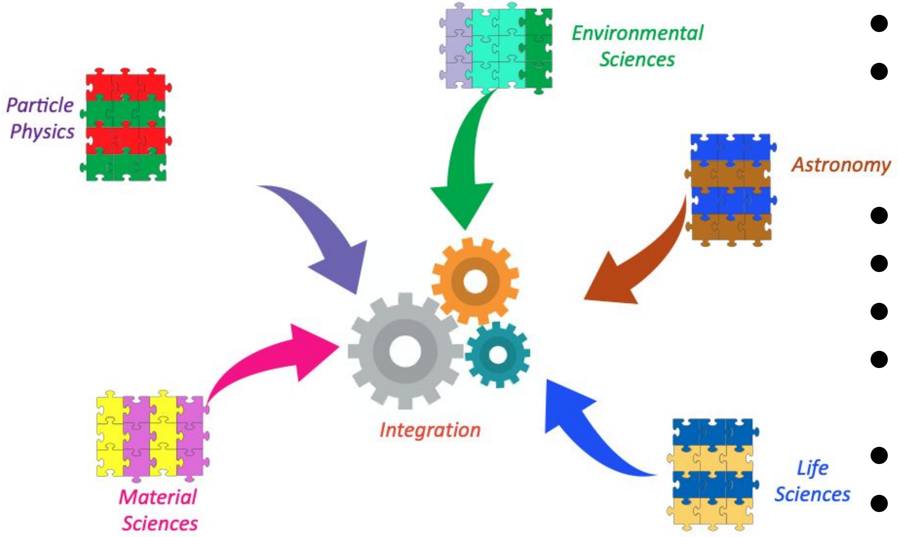
Datasets are the raw materials for all of machine learning. Models are only as good as the data they are trained on. Academics and entrepreneurs in particular depend on public datasets to create new technologies and new companies.

Best Practices

Best Practices empower researchers and engineers to more easily exchange models, reproduce experiments, and build applications that leverages machine learning. Improving best practices accelerates progress in, and grows the market for, machine learning.

Science Research MLCommons working group

- Science like industry involves edge and data-center issues, end-to-end systems, inference, and training, There are some similarities in the datasets and analytics as both industry and science involve image data but also differences; science data associated with simulations and particle physics experiments are quite different from most industry exemplars
- When fully contributed, the benchmark suite will cover (at least) the following domains: **material sciences, environmental sciences, life sciences, fusion, particle physics, astronomy, earthquake and earth sciences**, with more than one representative problem from each of these domains



- https://mlcommons.org/en/groups/research-science/
- One aim is to provide a mechanism for assessing the capability of different ML models in addressing different scientific problem
- i.e. **one benchmark measure is Scientific Discovery**
- Cover rich range of problem classes
- "End-to-end" is one class
- Provide common environment to store and run benchmarks (Software)
- 4 Initial Benchmarks (2 from DOE labs, 1 UK, 1 UVA)
- Surrogates Included (1 from LLNL next round)
- Lead use of FAIR metadata for MLCommons

Science-based Metrics

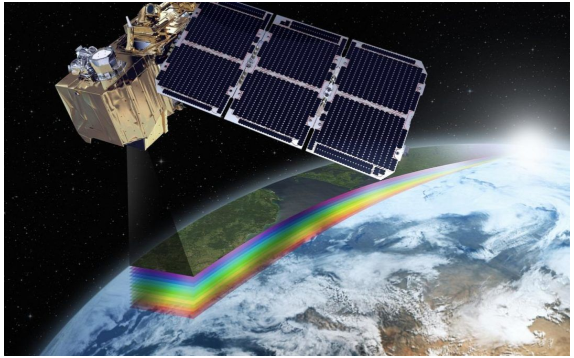
- Metrics will include those measuring **performance on science discovery,** e.g., could be one or more of:
  - Accuracy achieved
  - Time to solution (to meet a specific accuracy target)
  - Top-1 or Top-5 score
  - Chance your home will suffer a big earthquake …..
- Goal of our benchmarks is to **stimulate development of new methods relevant for scientific outcomes**. We aim to:
  - Offer well-defined “science data” sets
  - Provide a reference implementation - to help others overcome any format/interpretation/usage hurdles
  - Specify target benchmark metrics (to outperform)
  - Require a description of the improved method or code used by respondents
- *The science data should have enough richness to allow experimentation with innovative approaches.*
- Also include **traditional system performance benchmarks**

Benchmark	Science	Task	Owner Institute	Specific Benchmark Issues
CloudMask	Climate	Segmentation	RAL	Classify cloud pixels in images
STEMDL	Material	Classification	ORNL	Classifying the space groups of materials from their electron diffraction patterns
CANDLE-UNO	Medicine	Classification	ANL	Cancer prediction at cellular, molecular and population levels.
TEvolOp Forecasting	Earthquake	Regression	Virginia	Predict Earthquake Activity from recorded event data
ICF or Inertial Confinement Fusion	Plasma Physics	Simulation surrogate	LLNL	There are other possible LLNL benchmarks from collection of 10

Benchmark contains Datasets, Science Goals, Reference Implementations; hosted at SDSC or RAL
Specification of 4 Benchmarks https://drive.google.com/file/d/1BeefJTj4ZZL4Wa5c3zNz1l5nzQN-ktGR/view?usp=sharing

RAL Cloud Masking – Benchmark Overview

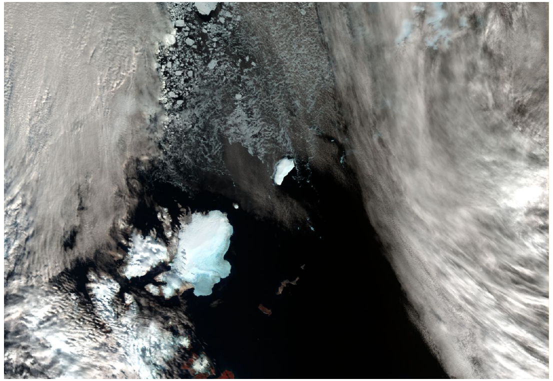
- Problem of identifying individual pixels of cloud from satellite imagery necessary for estimating the sea or land surface temperature
- This benchmark focusses on this particular ‘Cloud Masking’ task
- Relies on Sentinel-3 satellite data, particularly the Data from the Sea Land Surface Temperature Radiometer (SLSTR) instrument



Sam Jackson, Caroline Cox, Jeyan Thiyagalingam and Tony Hey

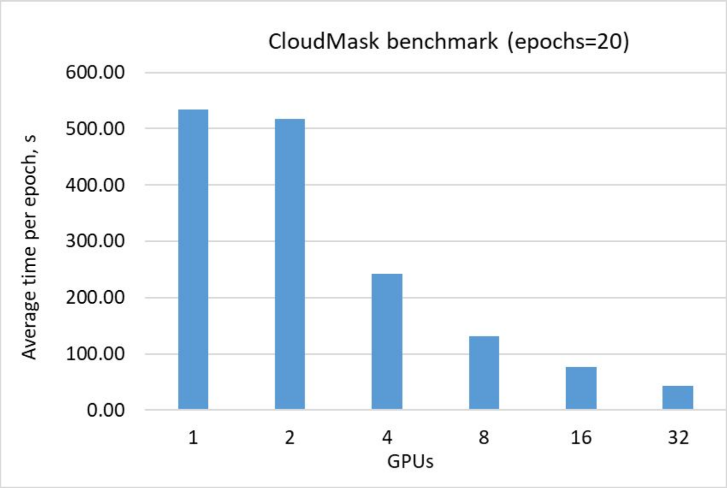
Cloud Masking – Benchmark Challenges

- Problem is challenging because cloud identification can be confused by several conditions
  - Snow, sea-ice, sun-glint, smoke, dust, …
- Traditional solution is thresholding or Bayesian filtering
- This benchmark uses a U-Net-based deep neural network
- Dataset
  - Around 200GB
  - Reflectance (6 channels, 2400 x 3000 pixels)
  - Brightness temperature (3 channels, 1200 x 1500 pixels)



Cloud Masking – Benchmark Status

- Benchmark Status: Ready to package
- Some initial results are being collected
- First version will have one dataset (**200GB**), but we intend to include another (**>1TB**)
- Implementation: Python, TensorFlow 2 (with Horovod)
- Metrics: Classification accuracy (among others)

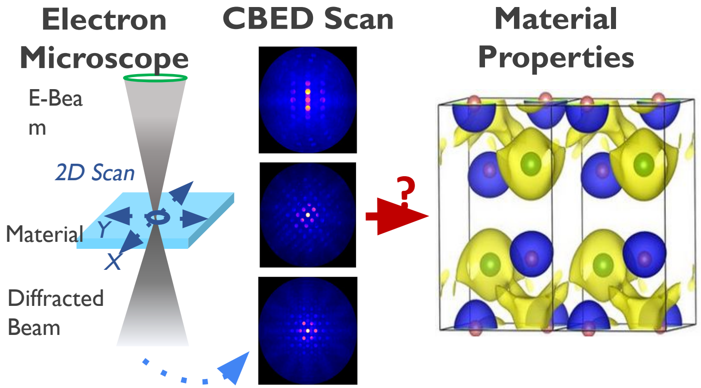


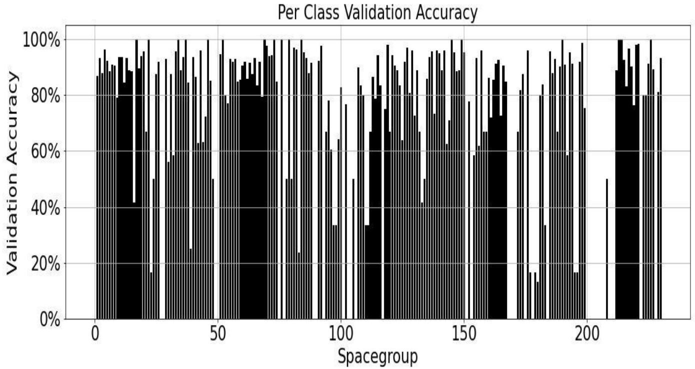
Average Training Time on V100s (per epoch)

Science and Technology Facilities Council

ORNL STEMDL – Benchmark Overview

- Classifying the space groups of materials from their electron diffraction patterns
- Reference based on ResNet50-based model
- Implementation: Python, PyTorch (with Horovod)
- Metrics: F1-Score and per-class accuracy (among others)



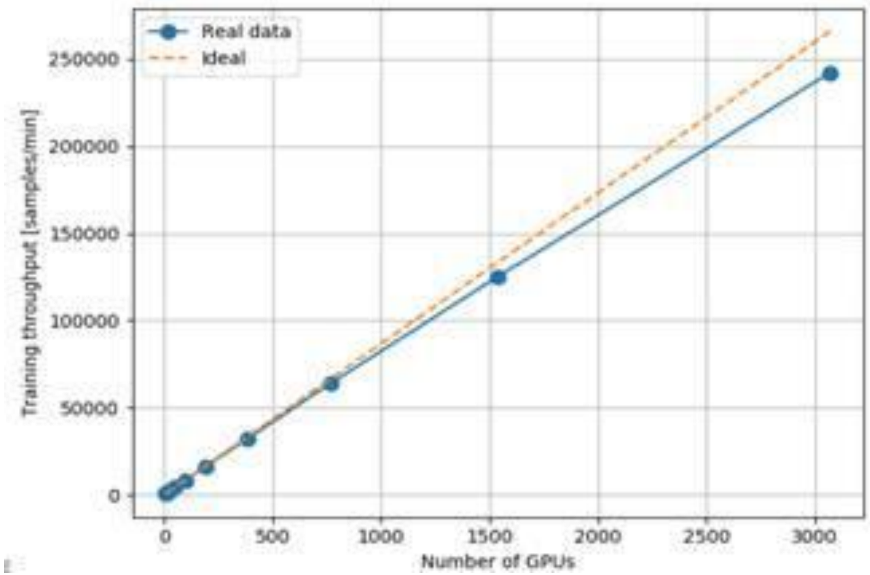


OAK RIDGE National Laboratory

Junqi Yin, Sajal Dash, Aristeidis Tsaris, Feiyi Wang, Mallikarjun Shankar

STEMDL – Benchmark Status

- Benchmark Status: Ready to package
- Some initial results are being collected
- Data: electron diffraction patterns for over 60,000 materials in material project database 10.13139/OLCF/1510313 (~**550GB**)



STEMDL Performance

OAK RIDGE National Laboratory

ANL CANDLE – Benchmark Overview

- Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer
- Implement deep learning architectures that are relevant to problems in cancer. These architectures address problems at three biological scales: cellular (Pilot1 P1), molecular (Pilot P2) and population (Pilot3)
- This benchmark focuses on **Uno from Pilot1 (P1):** The high-level goal of the problem is to **predict drug response** based on molecular features of tumor cells across multiple data sources.

- The goal of Uno is to build neural network-based models to predict tumor response to single and paired drugs, based on molecular features of tumor cells.
- It implements a deep learning architecture with 21M parameters in Python, Tensorflow 2, Keras
- **Metrics:** Time-to-solution (training time till a validation criteria, loss threshold, is reached) and throughput (samples/sec)
  - Need to add science metric

Murali Emani and Venkatram Vishwanath

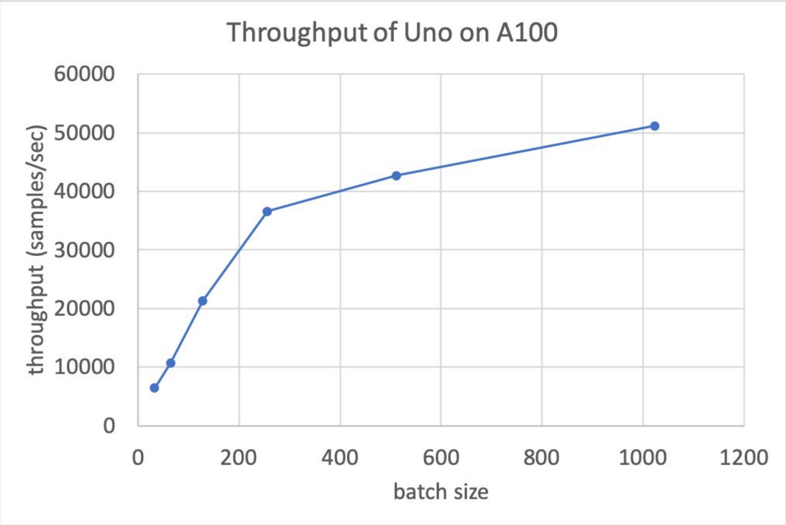
Argonne NATIONAL LABORATORY

CANDLE – Status

- ❖ Benchmark Status: Ready to Package
- ❖ Code is already available on GitHub

https://github.com/ECP-CANDLE/Benchmarks/Pilot1/Uno

- ❖ Relevant datasets are automatically downloaded
- ❖ Collecting some initial results (Theta @ ANL)
- ❖ Has collection of data engineering steps so could be end-to-end benchmark
- ❖ 3070 unique samples and 53520 unique drugs

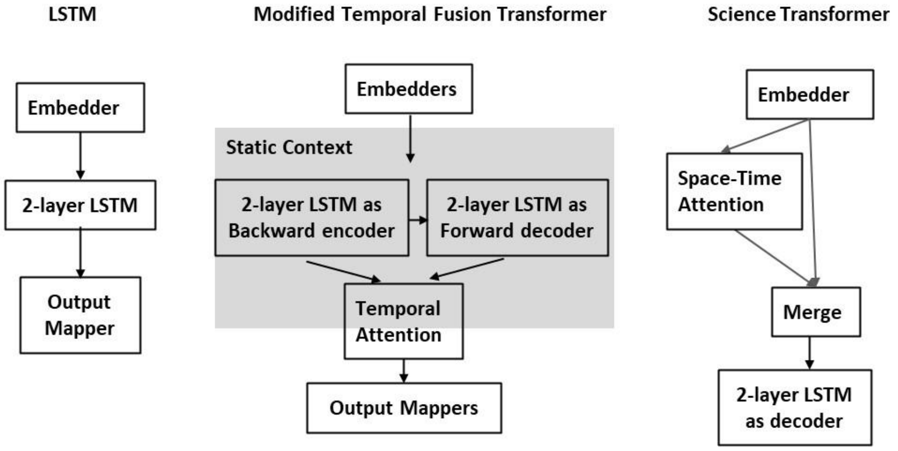


Throughput vs. batchsize on a single A100 GPU in a ThetaGPU node

Argonne NATIONAL LABORATORY

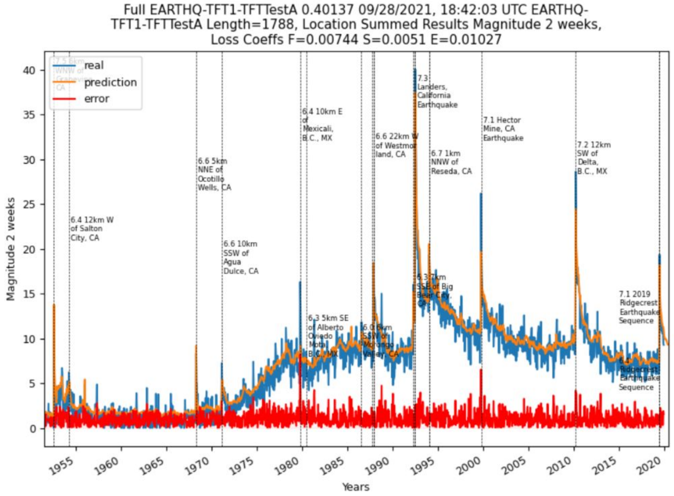
UVA TEvolOp Benchmark - Overview

- Time Series Evolution Operator
- Focuses on extracting the evolution in earthquake time series data
- **Earthquake data from 1950-now from USGS**
  - California; faults tagged -- typical results shown in figure
- Contains three reference models
  - **LSTM**
  - **Temporal Fusion Transformer** (Google/NVIDIA modified by UVA)
  - **Science Transformer** (University of Virginia)
- Metrics: multi-year forecasts of Earthquake activity as a function of time
  - Nash-Sutcliffe Efficiency
- Related to extreme events in stock market

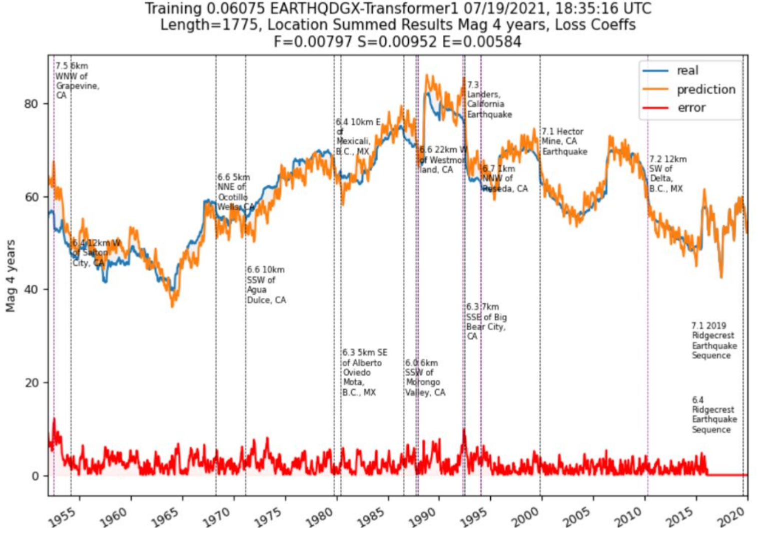


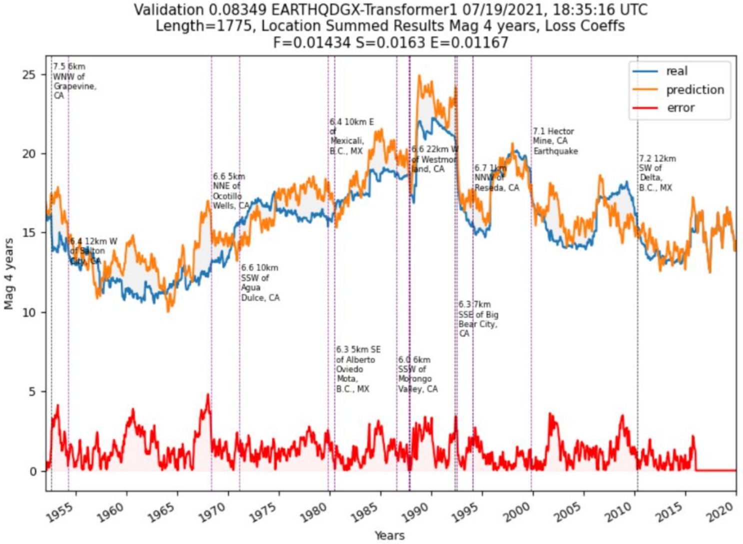
TEvolOp Benchmark - Status

- Benchmark Status: Ready to Package
- Implementation: Python - Jupyter Notebook, TensorFlow, Container
- 1790 time bins (2 weeks but input data daily), 2400 locations, ~12 measurements of magnitude, energy, depth, multiplicity
- 5 GB raw data
- Choice of Validation set -- time or space



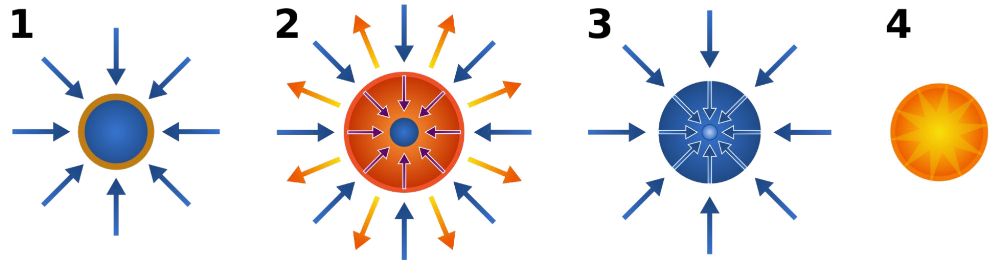
Figures are
On right: 4 year Training and Validation
On lef: 2 weeks



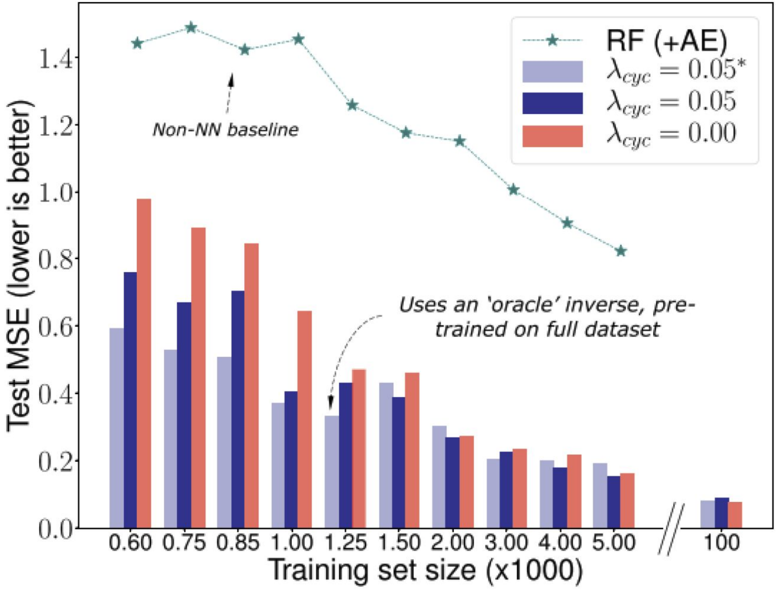


LLNL Inertial Confinement Fusion Simulation Surrogates I

- Allows generation of ensembles of simulations of final stages of an implosion (compression of (fusion target)
- https://www.pnas.org/content/pnas/117/18/9741.full.pdf
- 10K (getting much larger) training set with 5 input parameters
- Output is 22 scalars and 4 images from different energies



Schematic of the stages of inertial confinement fusion using lasers. The blue arrows represent radiation; orange is blowoff; purple is inwardly transported thermal energy.



- In similar state to other benchmarks with good write-up and GitHub https://github.com/rushilanirudh/macc
- Needs integration with Science WG and LLNL review

LLNL Inertial Confinement Fusion Simulation Surrogates II

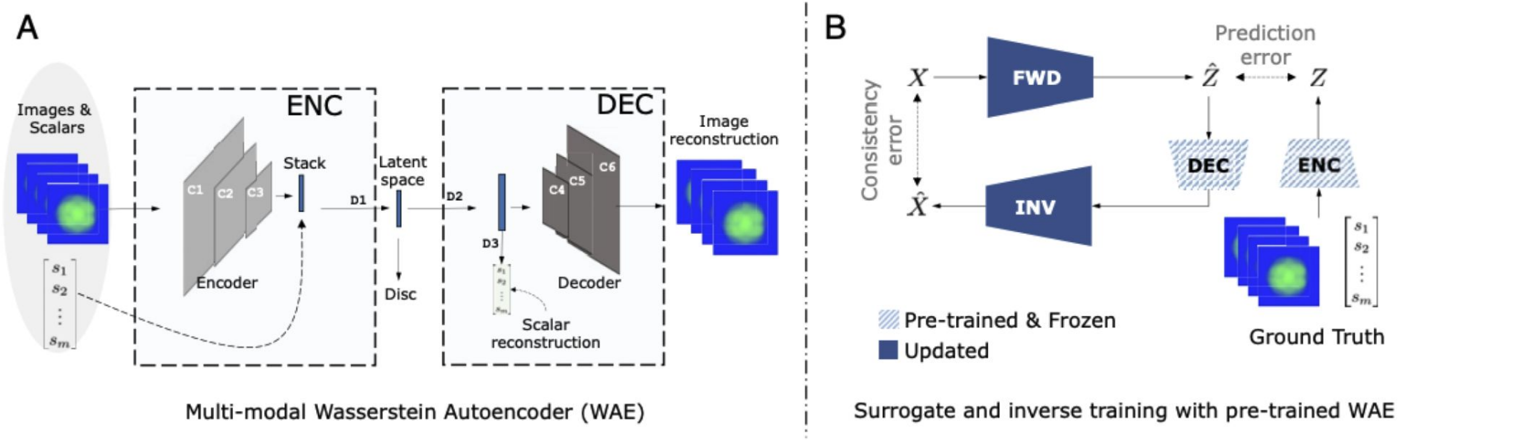
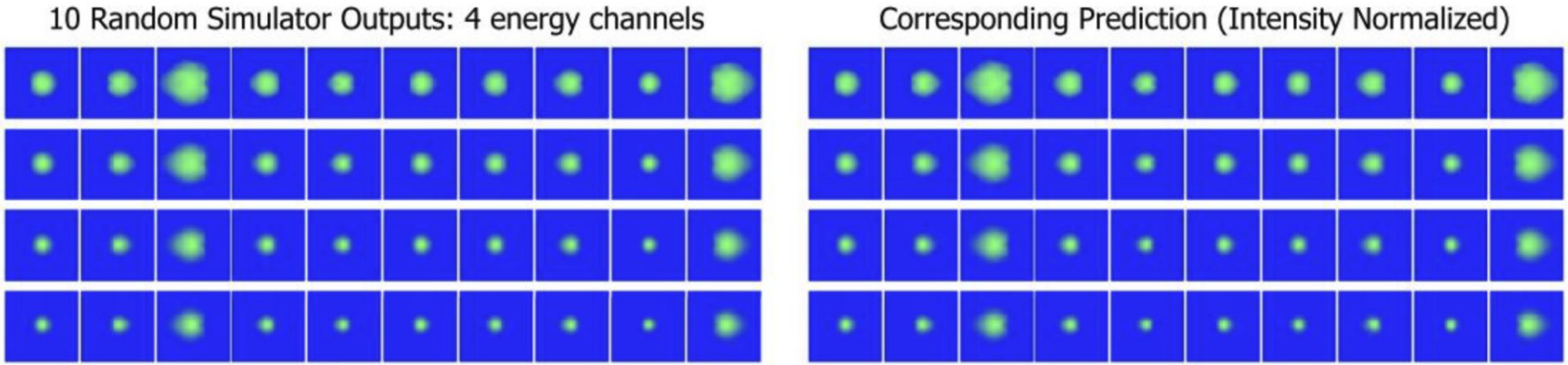


Fig. 1. MaCC surrogates. The proposed architecture uses a pretrained autoencoder (*A*) for ensuring manifold consistency and an inverse model (*B*) for cyclical consistency and robustness. ENC, encoder; DEC, decoder; FWD, forward; INV, inverse.



A. Random predictions from the proposed surrogate model

Packaging Software and FAIR Metadata

As well as benchmarks themselves, group is interested in technology for benchmarking

- We will develop FAIR metadata ontologies
- We will release the source code for all (through the working group's page) using:
  - MLCube ™
    - Quick Plug and Play
  - SciMLBench (Release 1.0+)
    - Open-Source Benchmarking Framework for AI for Science
    - Supports multiple nodes, containers and full customizability
- https://github.com/stfc-sciml/sciml-bench

Current Science WG Benchmark Status

- 4+1 Benchmarks available with datasets, reference implementations and preliminary goals
- The benchmarks are ready except for uniform MLCommons structures and specific submission formats.
- The formal submission process is not yet precisely defined but there will be
  - Open Division: Metric is Scientific Discovery
  - Closed Division: Metric is System Performance
- Access at
- MLCommonsScienceBenchmarks.pdf
- https://github.com/rushilanirudh/macc
- Join Working group https://mlcommons.org/en/groups/research-science/ at https://mlcommons.org/en/get-involved/
- See minutes at https://docs.google.com/document/d/167m7FK6-Ud4M5gXta5cIc1hKqaRHkk2B1GyKasdeQLc/edit?usp=sharing