Data Science Department Overview

Malachi Schram
Head of the Data Science Department
Newport News, Virginia
January 6th, 2023

Data Science at Jefferson Lab

Mission:

- Provide world-class data science solutions to advance research in nuclear physics by working with the subject matter experts at Jefferson Lab, partnering universities and Labs, and the Department of Energy.
- Provide world-class data science solutions to scientific applications relevant to the regional scientific community

Vision:

- Expand the capability and capacity of data science at JLab
- Create a collaborative data science research hub to:
    1. Work with regional partners on challenging scientific problems
    2. Champion education and research opportunities with regional universities and industry
    3. Reduce the carbon footprint by optimizing the data science workflow and algorithms

Current Portfolio

DOE Nuclear Physics:

- Working with the experimental Halls (Particle Identification, etc.)
- Data Science contributing effort for AIEC (lead by EPSCI)
- Quantom SciDAC (with ANL)

DOE Basic Energy Science:

- Machine Learning for Improving Accelerator and Target Performance (with ORNL)
- Collaborating with SLAC on application of ML-based controls for accelerators

DOE Advanced Scientific Computing Research:

- Data-Driven Decision Control for Complex Systems (with PNNL, ORNL, UC)

Non-DOE:

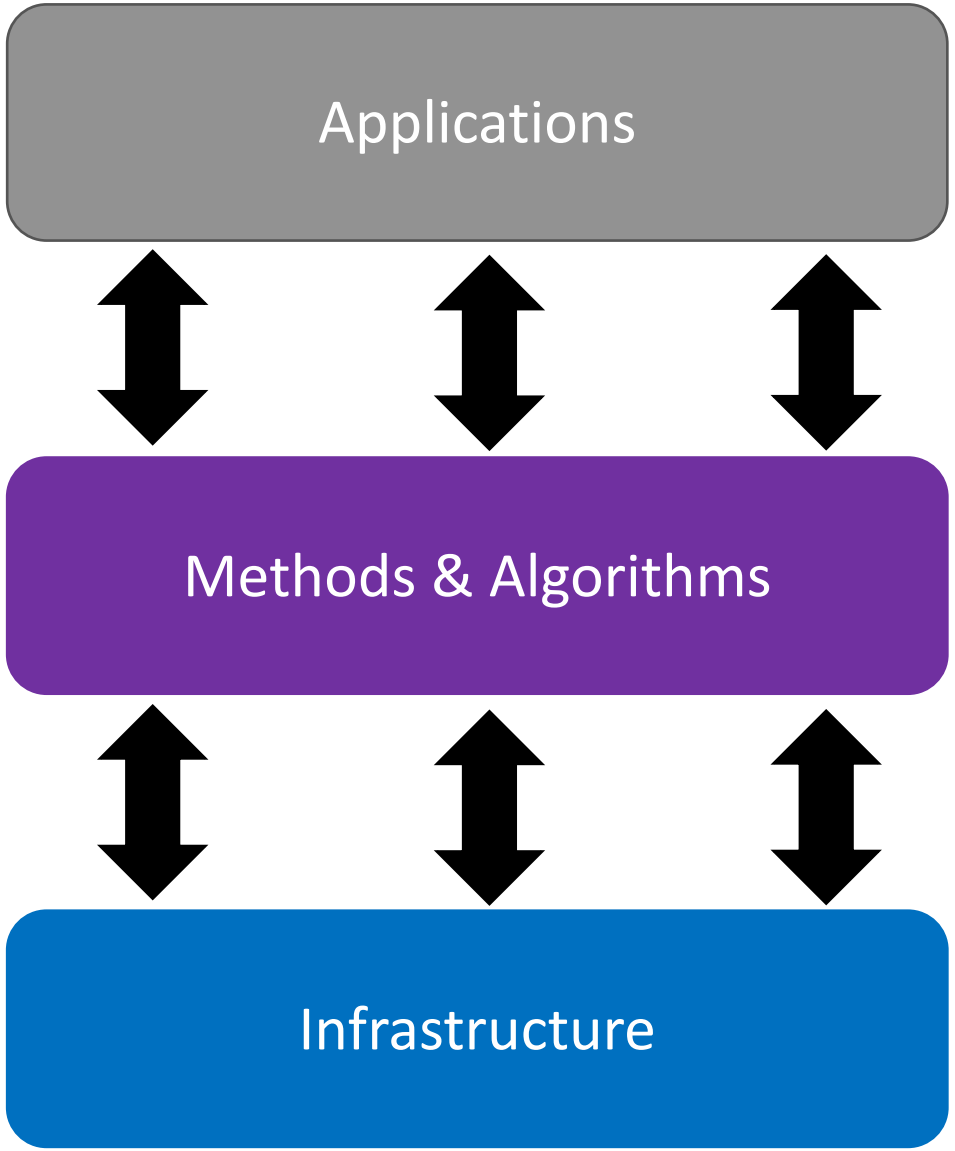
- Hampton Roads Digital Twin (with ODU)

Laboratory Directed Research & Development (LDRD):

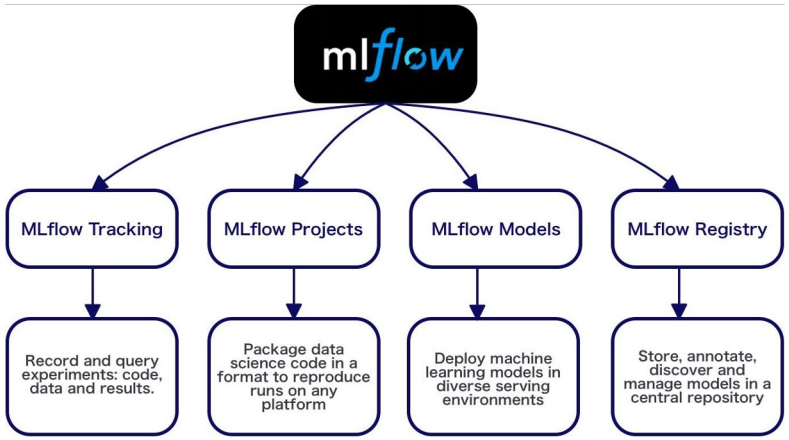
- Multi-objective Optimization of Heat Load and Trip Rates in CEBAF (FY22)
- Adaptive Strategies for Optimal Computing Availability (FY23)

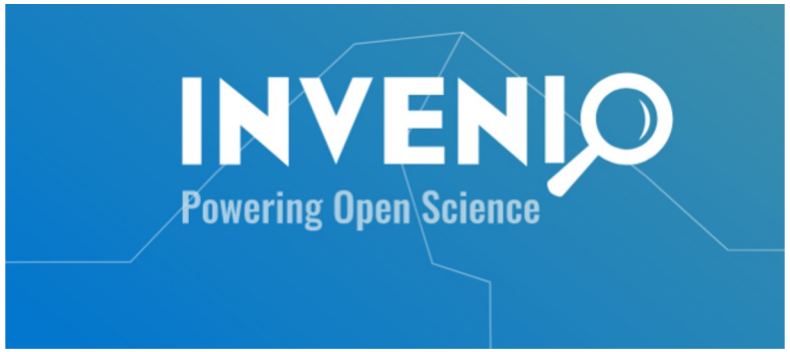
JLab Data Science Pillars

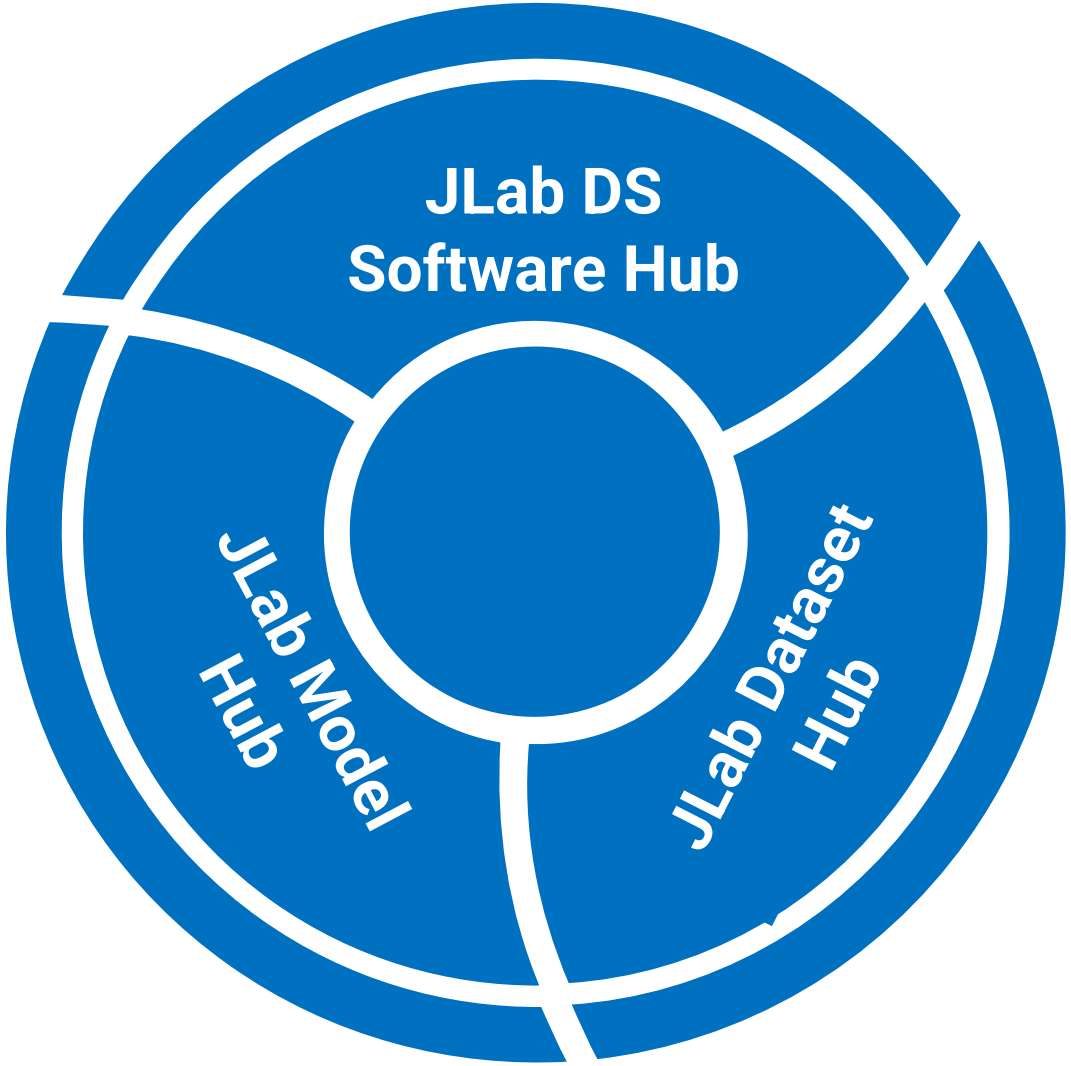
- **Applications**:
  - Nuclear Physics
  - Advanced Scientific Computing
  - Health & Climate
- **Focused Methods & Algorithms**:
  - Uncertainty Quantification
  - Interpretability and Explainability
  - Design & Control
- **Infrastructure:**
  - JLab ML & Data Hub
  - JLab Data Science software



Data Science Infrastructure

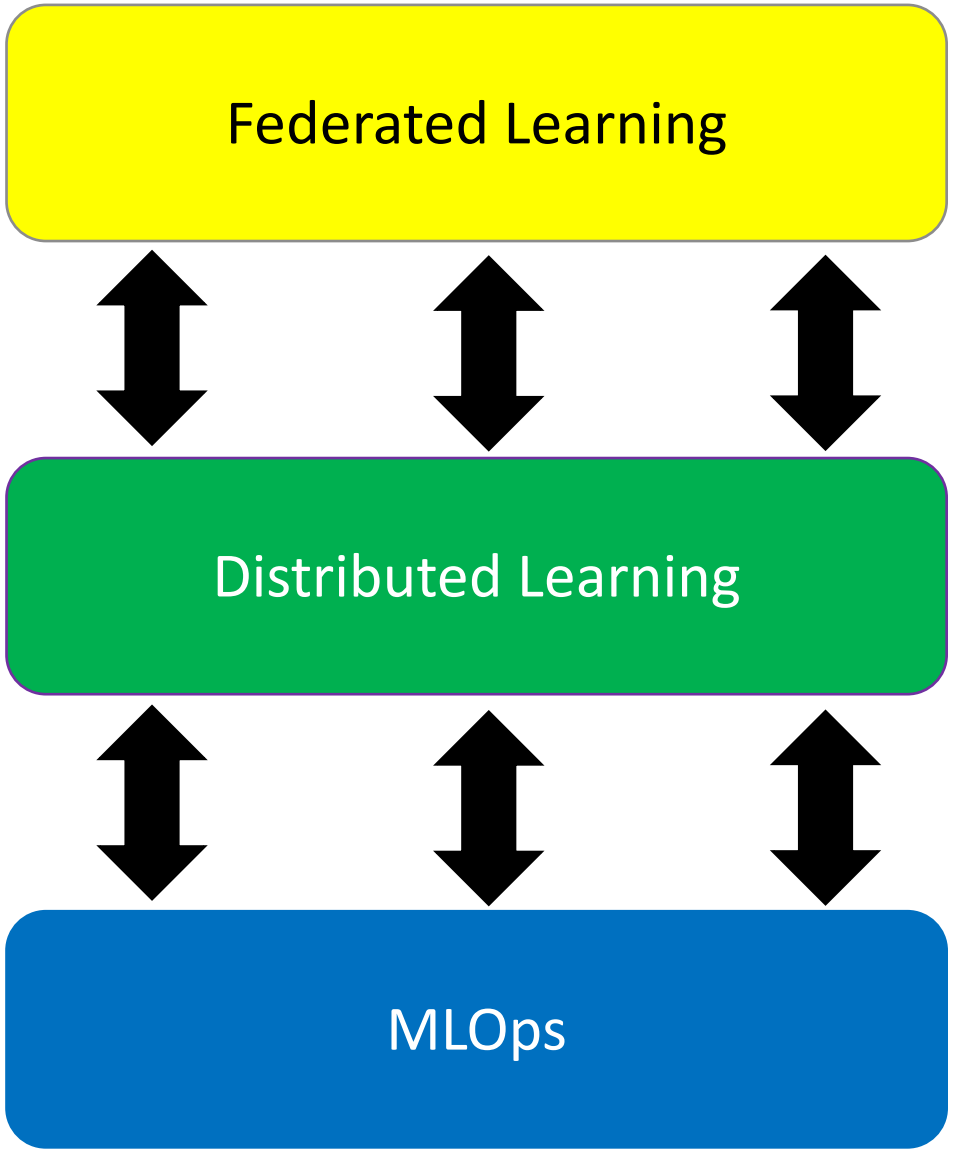






Jefferson Lab

Data Science Infrastructure



Phase 3: Extension to provide privacy preserving capabilities

Phase 2: Allows for local and geographically distributed learning

Phase 1: Provides the core infrastructure to capture the full lifecycle

Data Science Methods & Algorithms

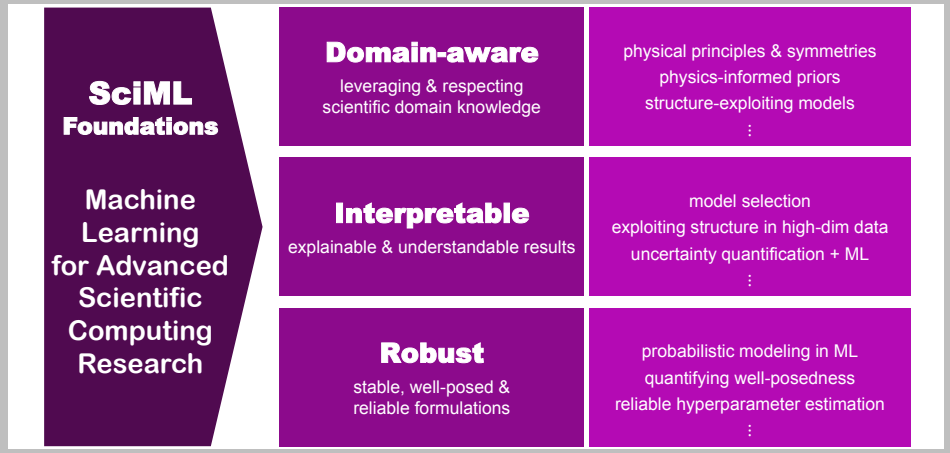


Figure 1: Foundational research themes of SciML must tackle the challenges of creating domain-aware, interpretable, and robust ML formulations, methods, and algorithms.

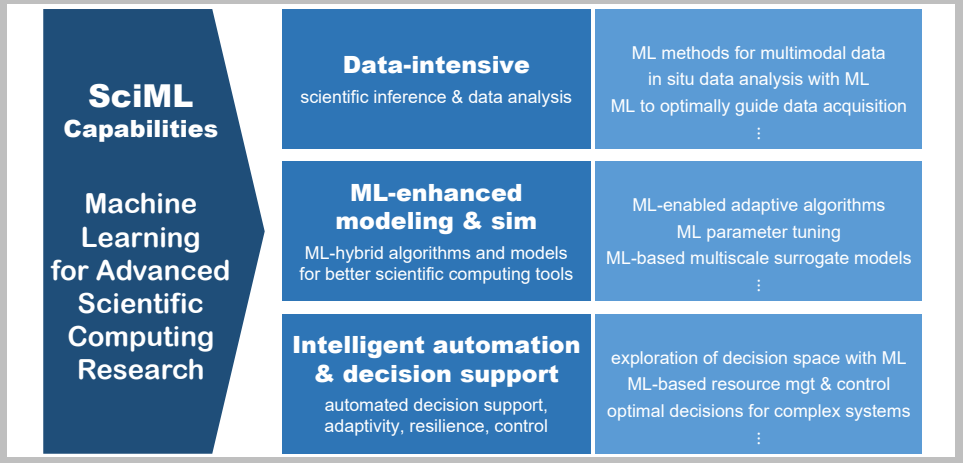
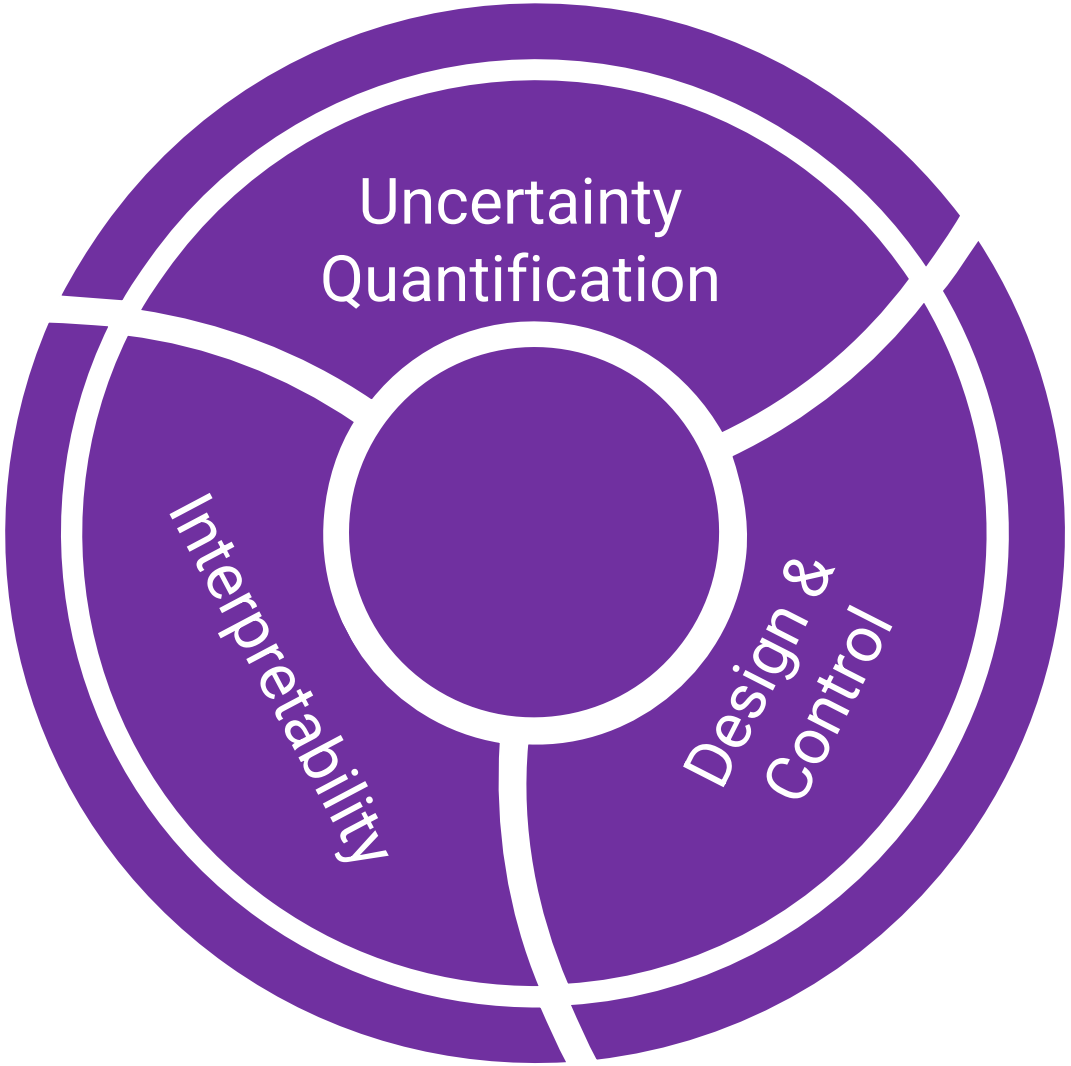


Figure 2: Opportunities for SciML impact arise in scientific inference and data analysis; in ML-enhanced modeling and simulation; in intelligent automation and decision support; and in related applications.



Jefferson Lab

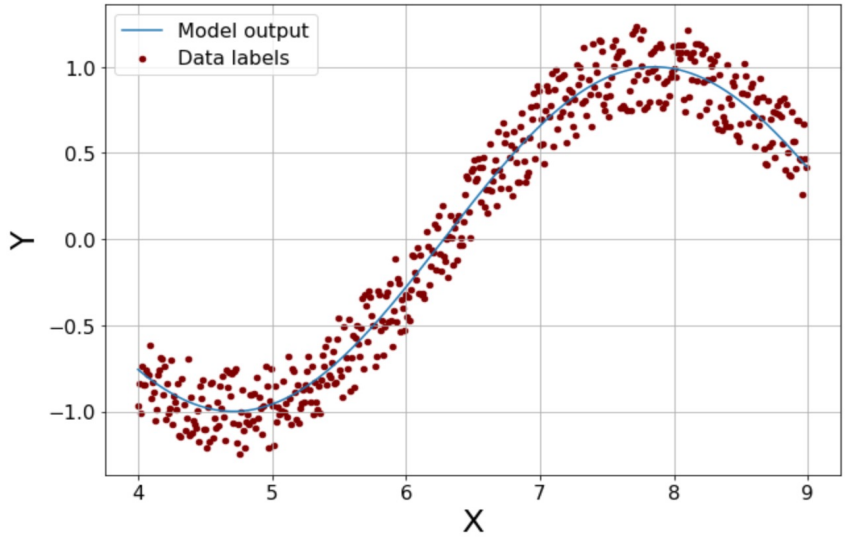
UQ PROBLEM DEFINITION

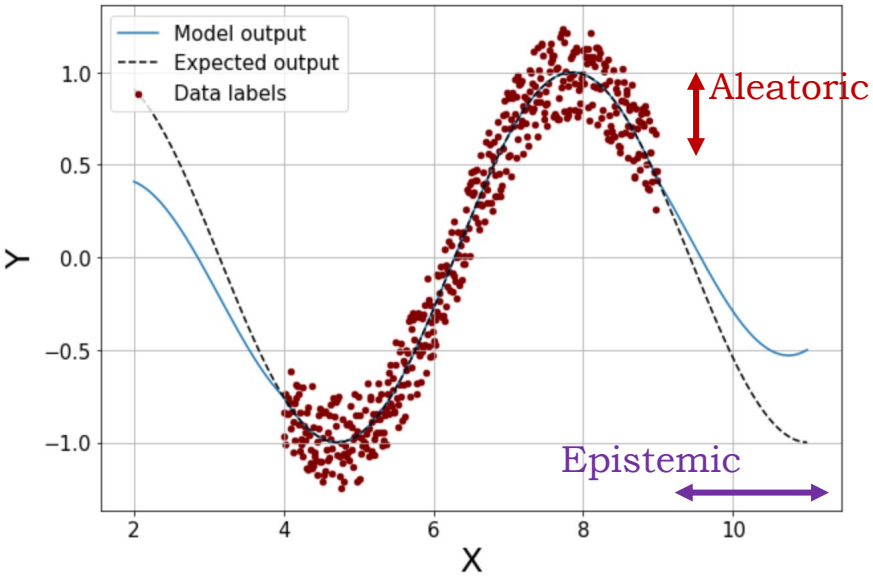
- We are focused on:
  1. Applications with **high-dimensional** continuous input features
  2. Focused on **large data sets** for DOE applications
  3. **Safety constraints** that should never or at least rarely be violated.
  4. **Inference** that must happen in **real-time** at the control frequency of the system.

- To tackle some of these points would need:
  - Integration of uncertainty quantification (UQ) to provide safety
    - Including **out-of-distribution** uncertainty
  - **Single inferences** model estimation with UQ

UNCERTAINTY QUANTIFICATION

- Deep Learning (DL) models are deterministic transformation functions from an input to the output
- DL models are very powerful and expressive
- It is important to know the confidence associated with each prediction from a DL models for decision making

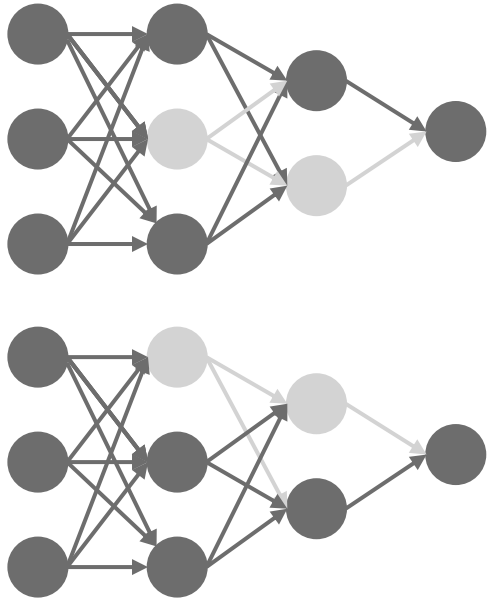




Uncertainty Types: Aleatoric vs Epistemic uncertainties

- Aleatoric → Data uncertainties
- Epistemic → Out of training distribution uncertainty (OOD)

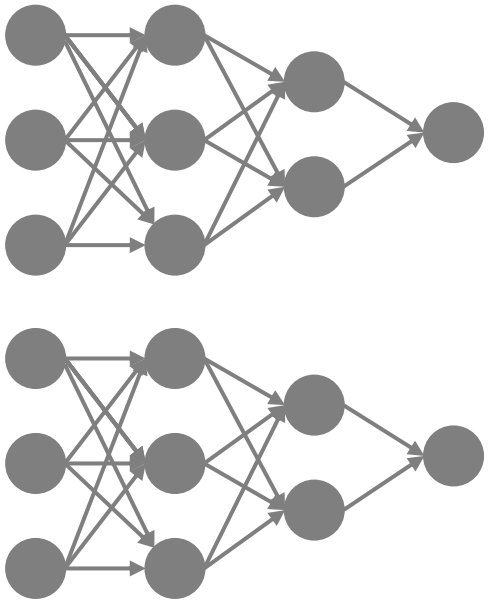
COMMON UNCERTAINTY ESTIMATION METHODS IN DEEP LEARNING



(a) MC Dropout

Use MC dropout during inference with dropout layers on can provide uncertainty prediction.

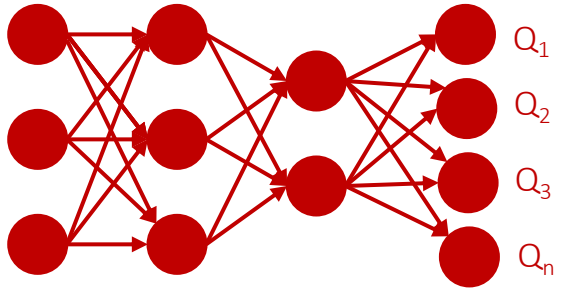
However, it slow and requires offline calibration.



(b) Ensemble

Create multiple copies of the same model architecture trained with different parameters initialization.

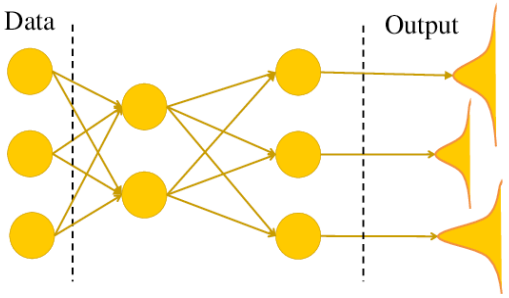
However, it's requires a lot more memory, it's slower (aggregate results) and requires calibration after training.



(c) Quantile Regression

Model is trained to predict quantiles for the regression problem.

However, we'll see it doesn't account for out-of-distribution uncertainty.

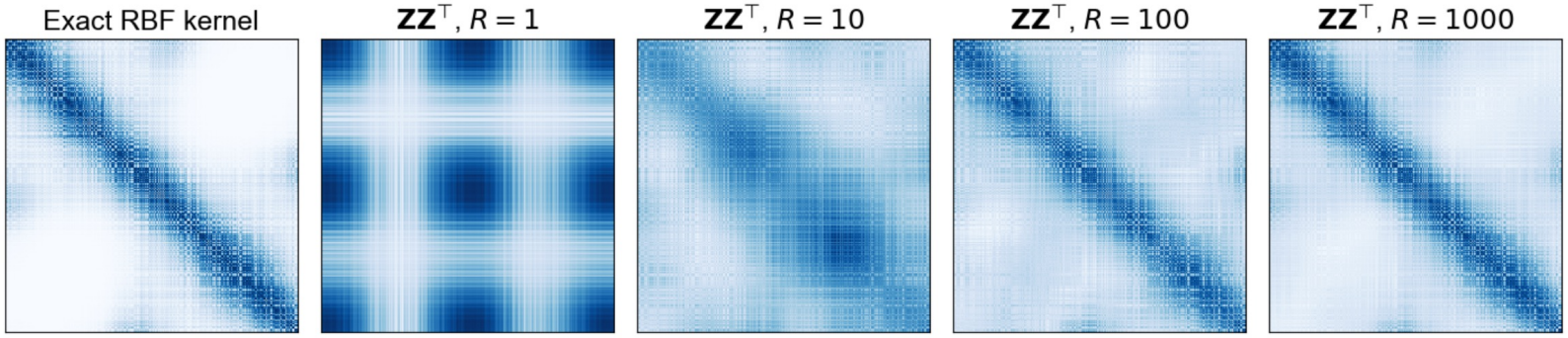


Jefferson Lab

GAUSSIAN PROCESSES AND RANDOM FEATURES

- Gaussian processes scales very poorly with high dimensions and large datasets
- Random Fourier Features have been used to approximate the kernel (for specific conditions) to significantly resource the computational cost for large dimension and big data problem

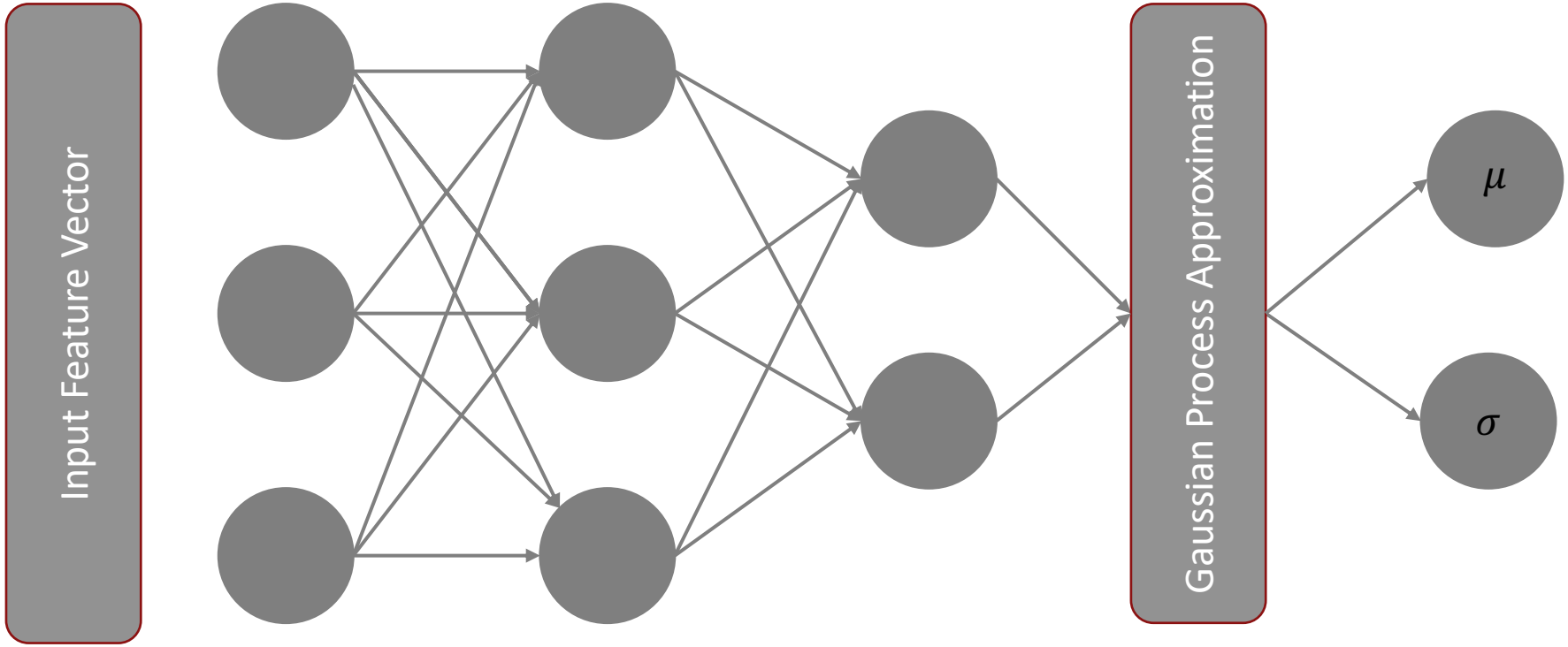
$$k(x, y) \approx z^T(x)z(y)$$



- Select research on reducing the high dimension using deep model:
  - Random Features for Large-Scale Kernel Machines (https://proceedings.neurips.cc/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf)
  - Deep Kernel Learning (https://arxiv.org/abs/1511.02222)
  - Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness (https://arxiv.org/abs/2006.10108)
  - On Feature Collapse and Deep Kernel Learning for Single Forward Pass Uncertainty (https://arxiv.org/abs/2102.11409)

DEEP GAUSSIAN PROCESS APPROXIMATION

1. Reduce the high dimensional input feature vector using a neural network



2. Take the reduced latent space as input to the Gaussian Process approximation

$$k(h, h') \approx z^T(h)z(h')$$

BI-LIPSCHITZ CONSTRAINT AND FEATURE COLLAPSE

- A problem when introducing a deep model to reduce latent space is it doesn't guarantee that the distance between the input features is preserved in the latent space

- This is handled using the bi-Lipschitz constraint:

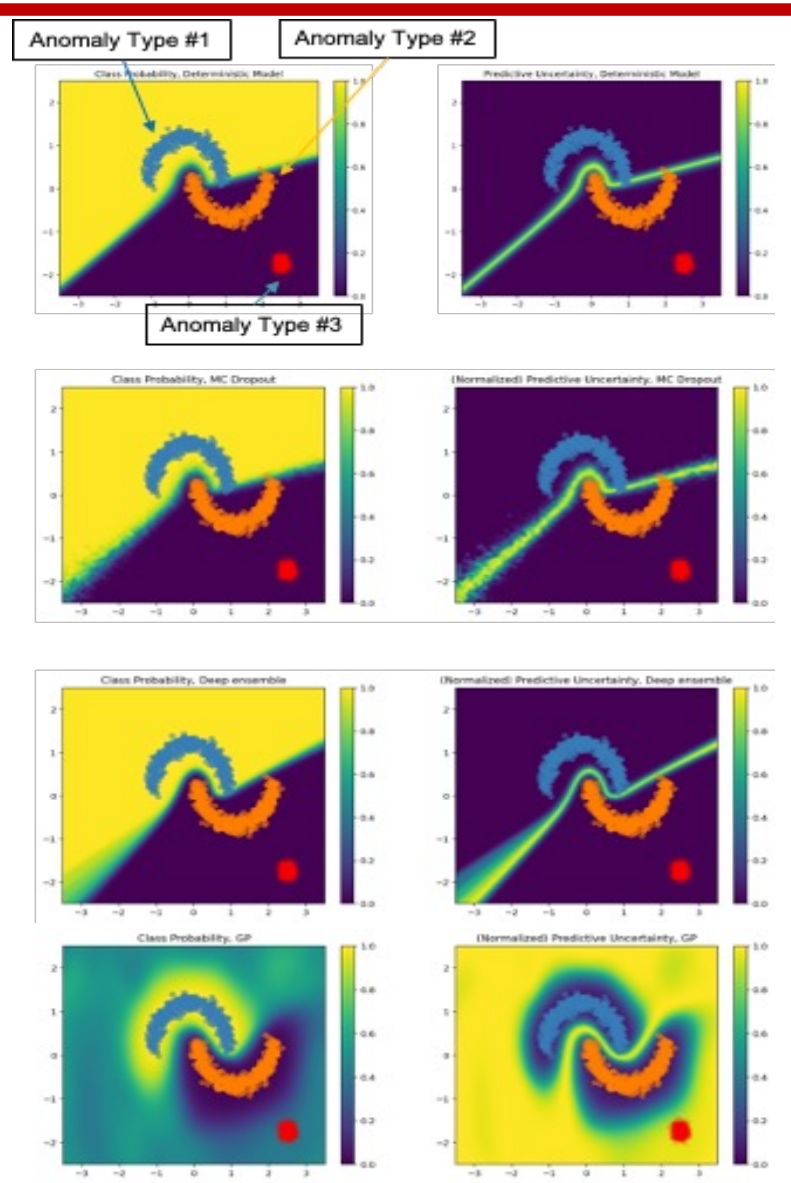
$$L_1 * ||x_1 - x_2||_X \leq ||h_1 - h_2||_H \leq L_2 * ||x_1 - x_2||_X$$

- The lower bound avoid feature collapse
- The upper bound ensure feature similarity
- We enforce this constraint using a loss penalty but will revisit other techniques

Uncertainty Quantification for ML

Develop methods that include uncertainty estimates in machine learning models

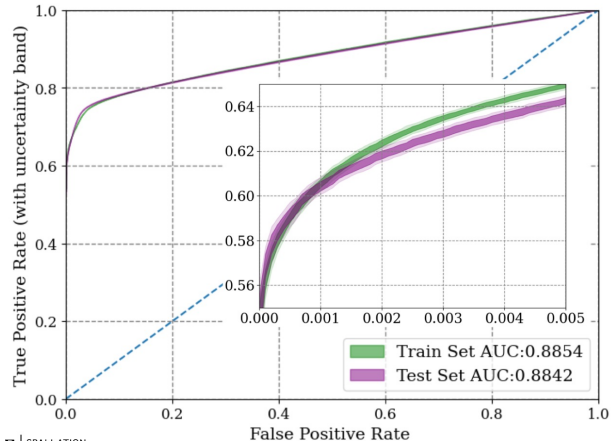
- Applications:
  - Data driven ML-based surrogate models
  - Real time controller
  - Anomaly detections
- Requirements:
  - Out-of-distribution uncertainties
  - Auto-calibration
  - Single inference
- Hardware considerations:
  - Memory
  - Inference time
  - Performance trade-off due to approximations

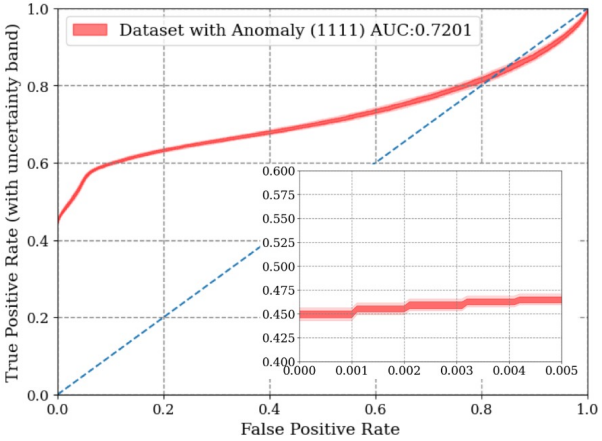


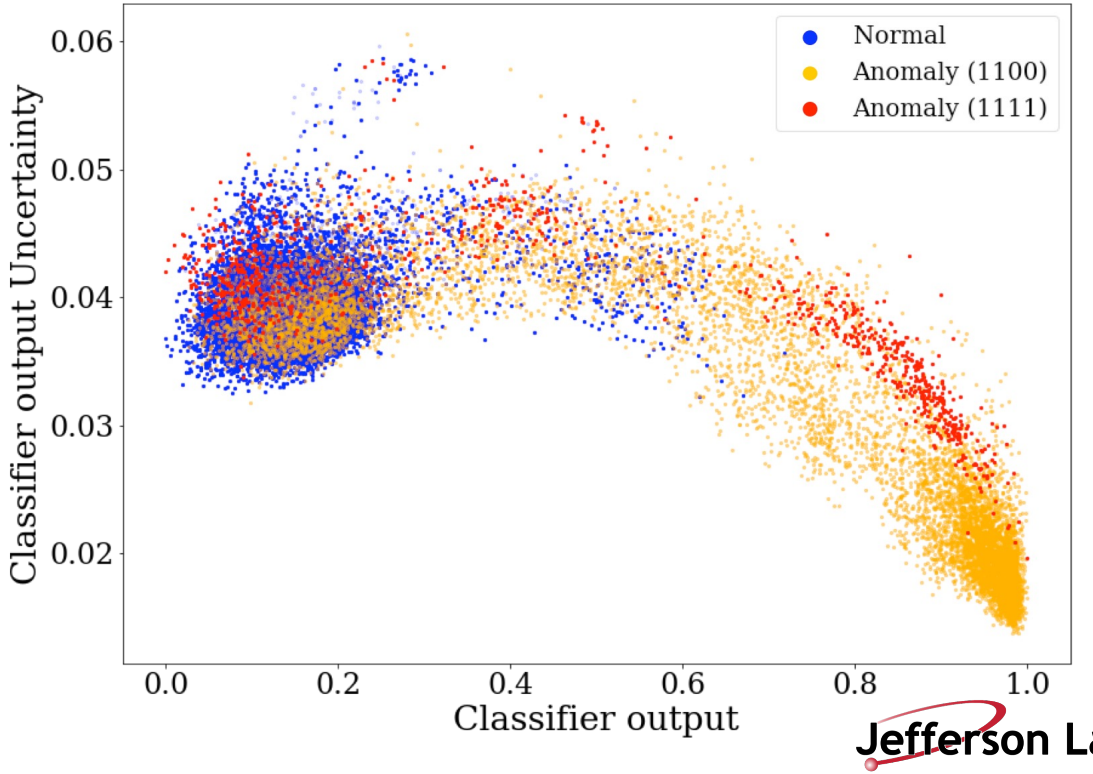
Jefferson Lab

Uncertainty Aware Siamese Model ("Classification")

- We enhanced our models by adding GP approximation layer which provides the uncertainty estimate
- Results from similarity model showed a ~4x improvement in performance over previously published results, it is also much better than a vanilla Auto-encoder
- The ROC curves show true fault detection rate above 60% while keeping the false alarms below 0.5% (not optimized)
- We introduced an out-of-domain anomaly, labelled 1111 (red), the UQ-based model performed similar in classifying the anomalies and indicated high uncertainty (as expected)



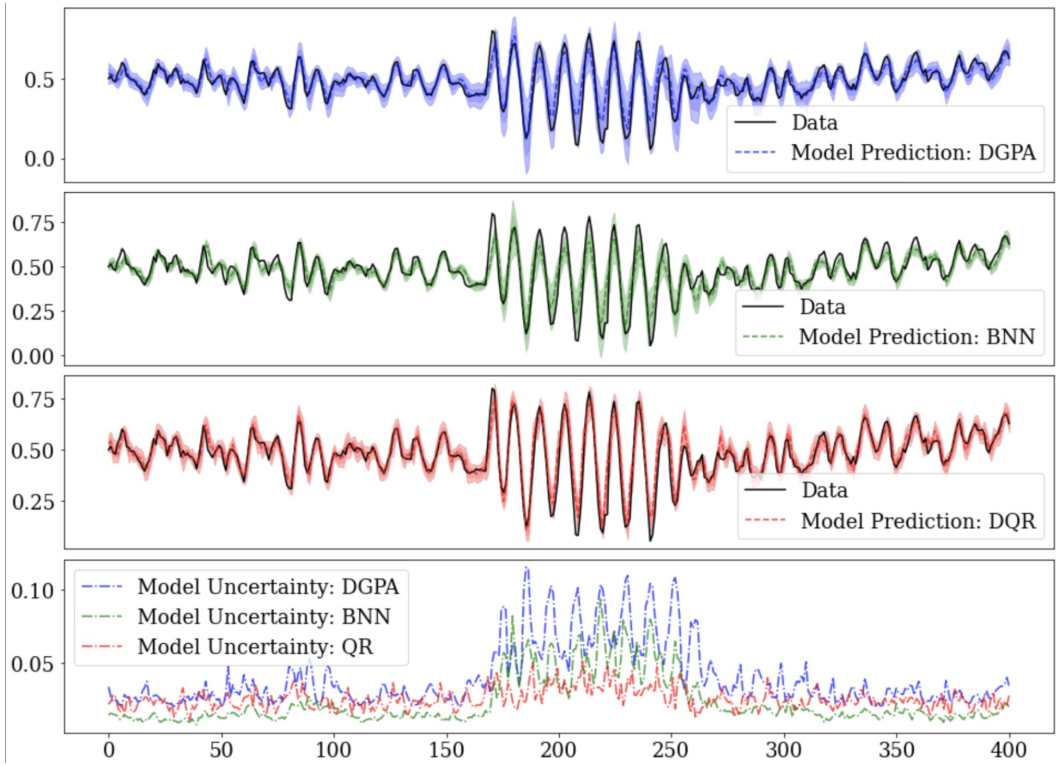


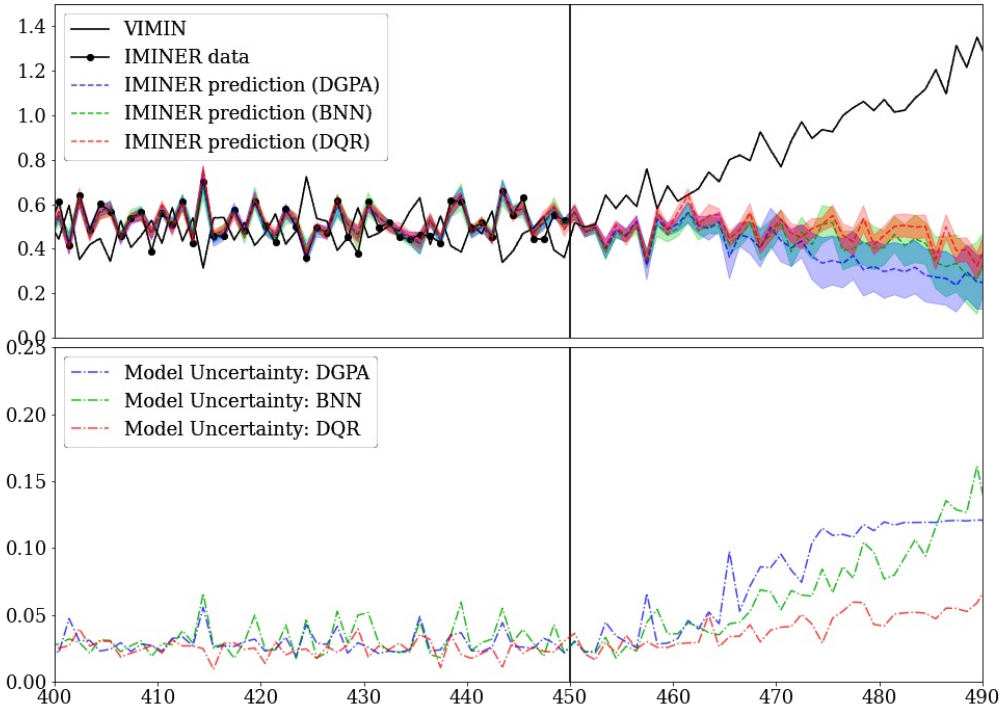


Jefferson Lab

Data Driven UQ ML-based Surrogate Models (Regression)

- Compare different techniques: DQR, BNN, DGPA
- DQR models have great performance for training distribution but not for OOD
- BNN models do a better job to estimate OOD
- DGPA models are distance aware by design resulting in better OOD estimation



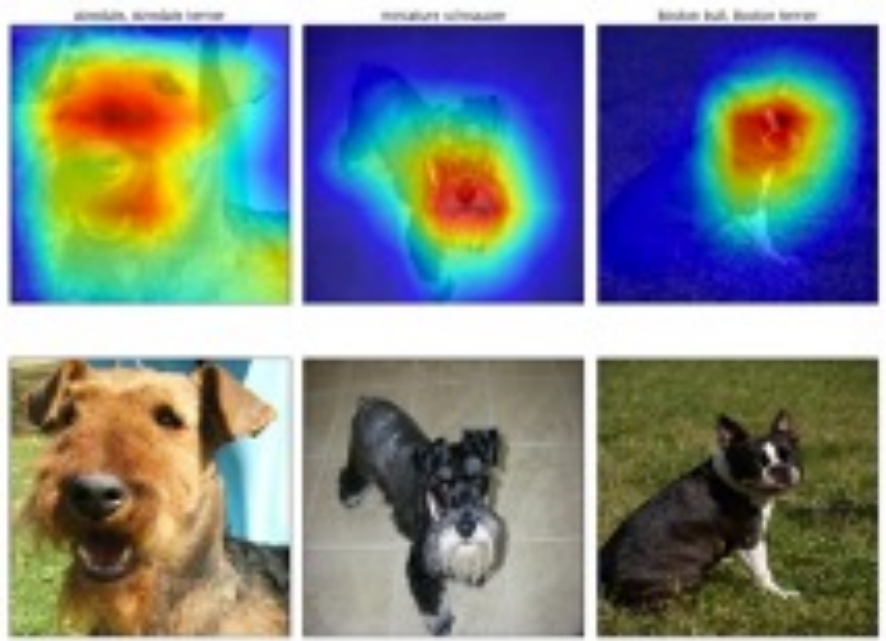


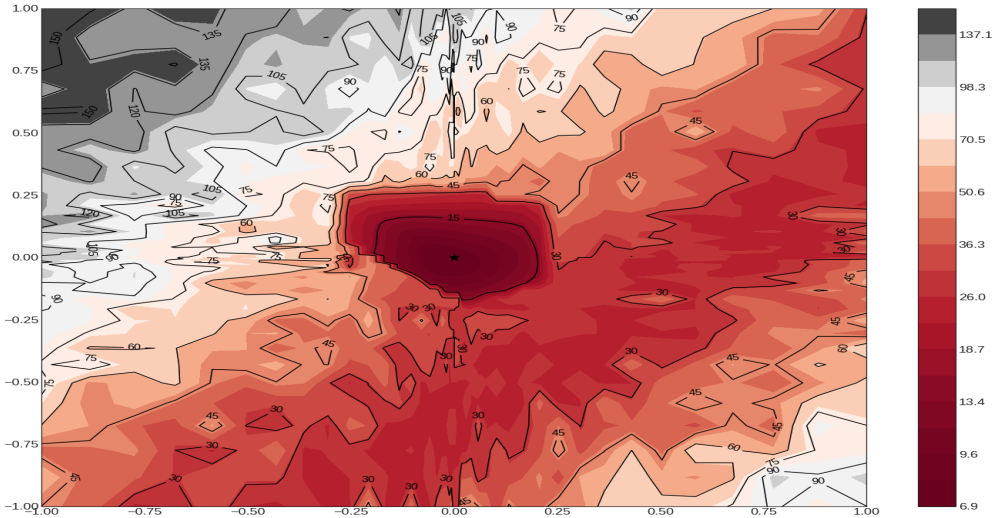
Jefferson Lab

Interpretability, Explainability, and Robustness

Applying and developing techniques to better understand model predictions and stability:

- Gradient activation studies to understand what the model is focusing on
- Gradient model layer studies to understand where the model is learning
- Loss landscape analysis to better understand the model stability





Loss Landscape for FNAL system dynamic model

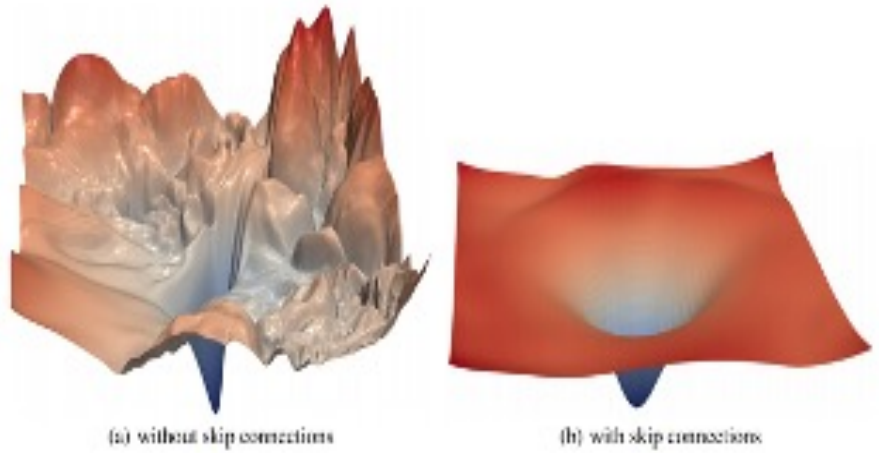


Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

Jefferson Lab

Gradient Class Activation Mapping (GradCAM)

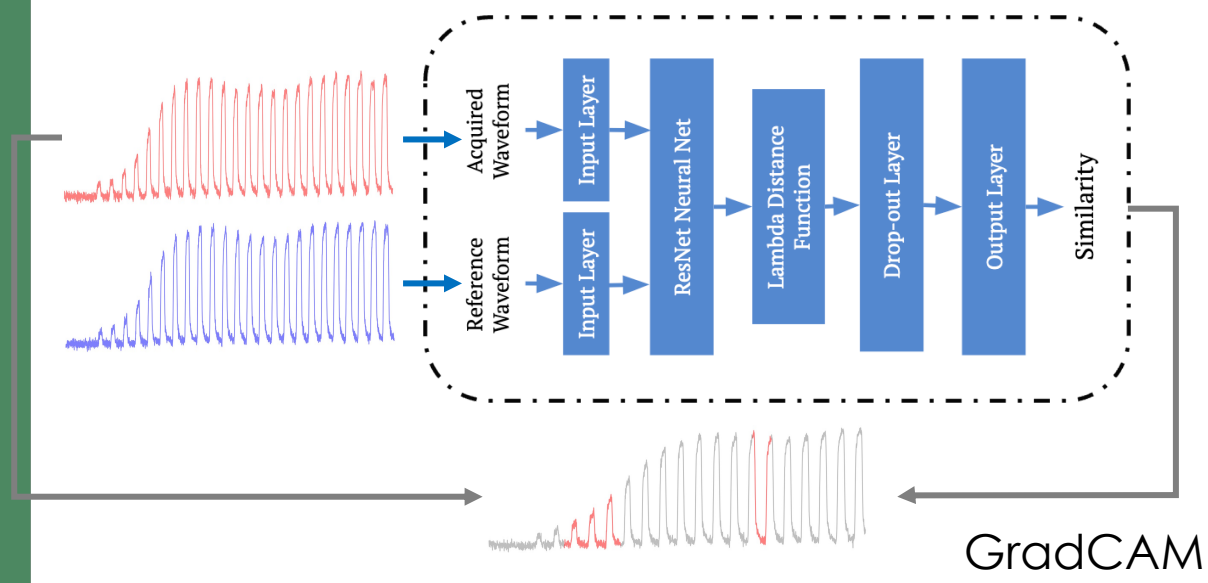
- GradCAM provides mapping between the the model output to the features in the input that the model thinks are the most relevant
- Extracts the most active features in the last convolutional layer and maps them back to the input

Original Image | Grad-CAM 'Cat' | Grad-CAM 'Dog'

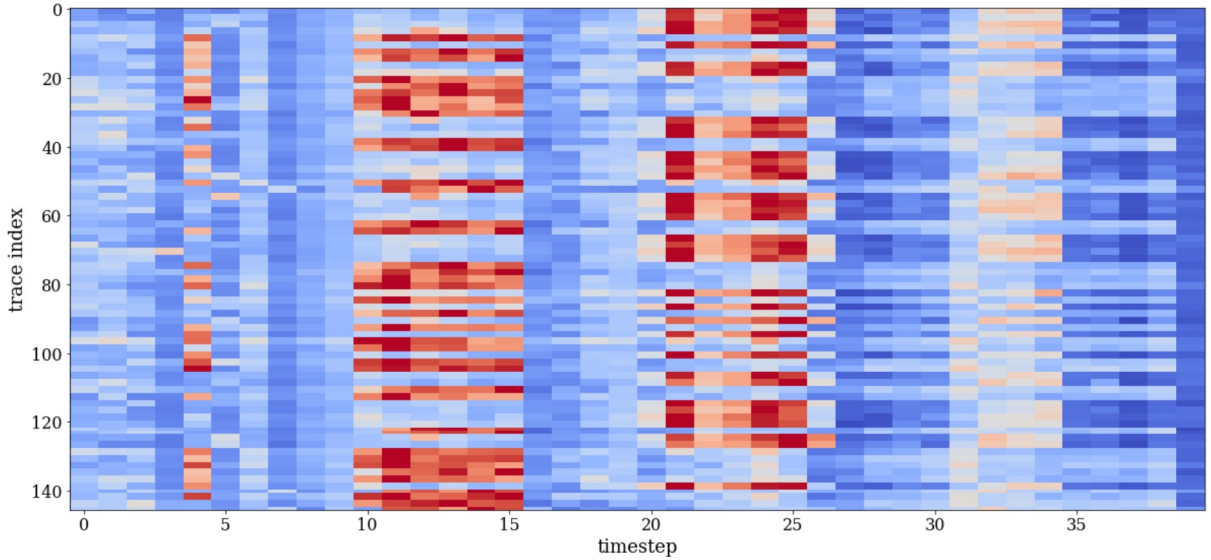
https://arxiv.org/abs/1610.02391

Equipment Fault Classification using GradCAM

- Applied GradCAM on SNN model trained to predict Errant Beam Pulses
- It identified sections of the waveform most relevant for a particular decision from the model

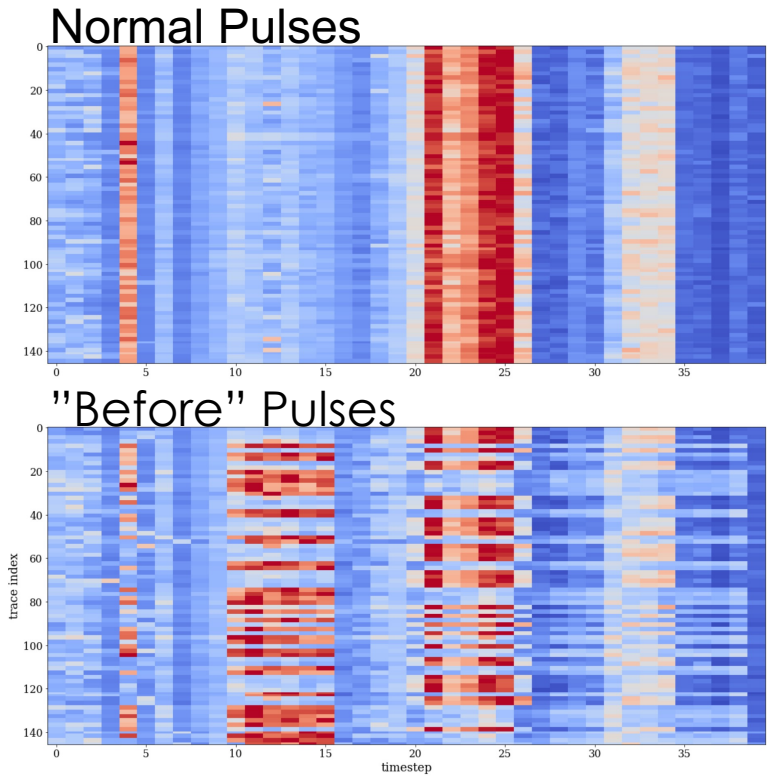


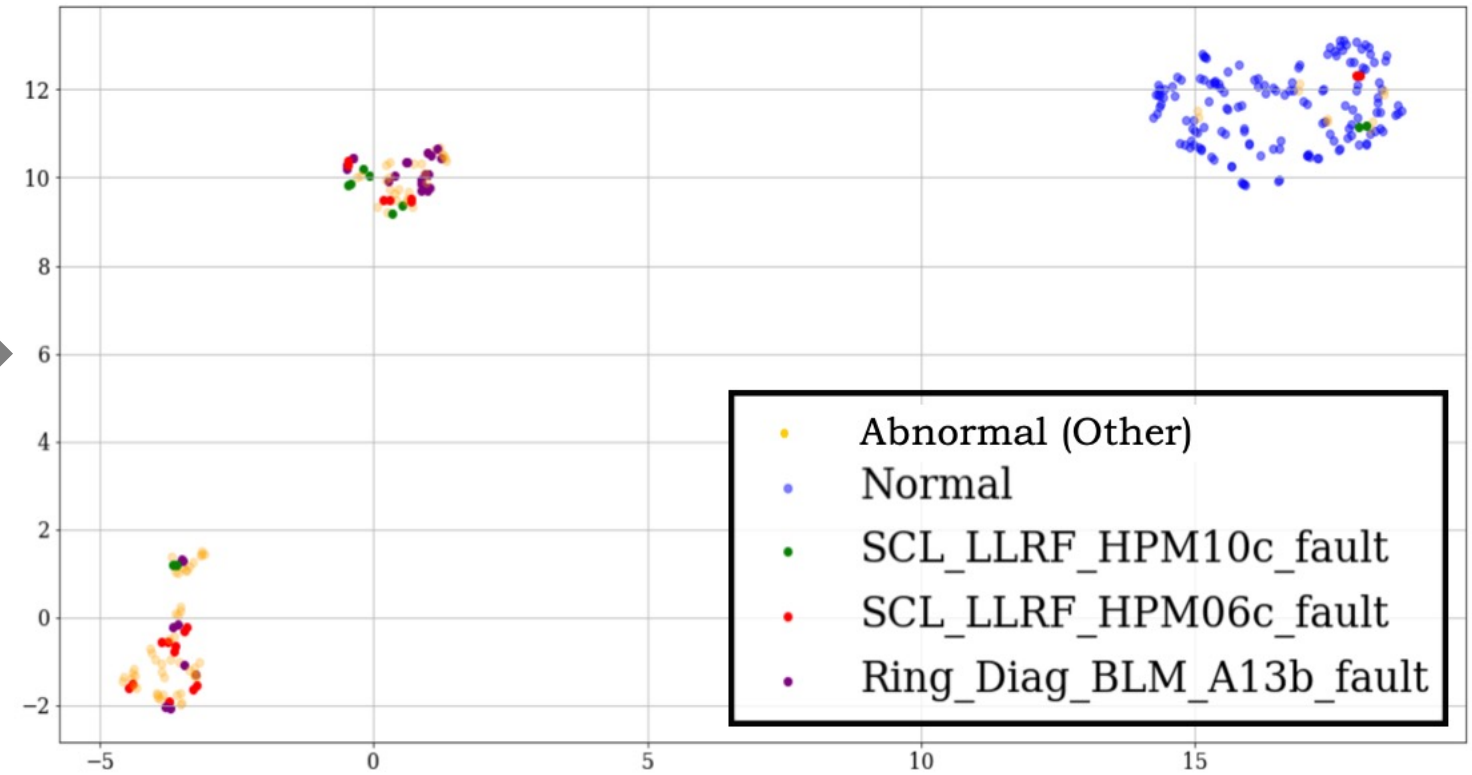
Stacked Relevance Vectors ("Before" Pulses)



Equipment Fault Classification using GradCAM

- The salient feature vectors are reduced to 2-dimensional space using UMAP*
- Studying how the cluster location from anomalies relate to specific equipment failures





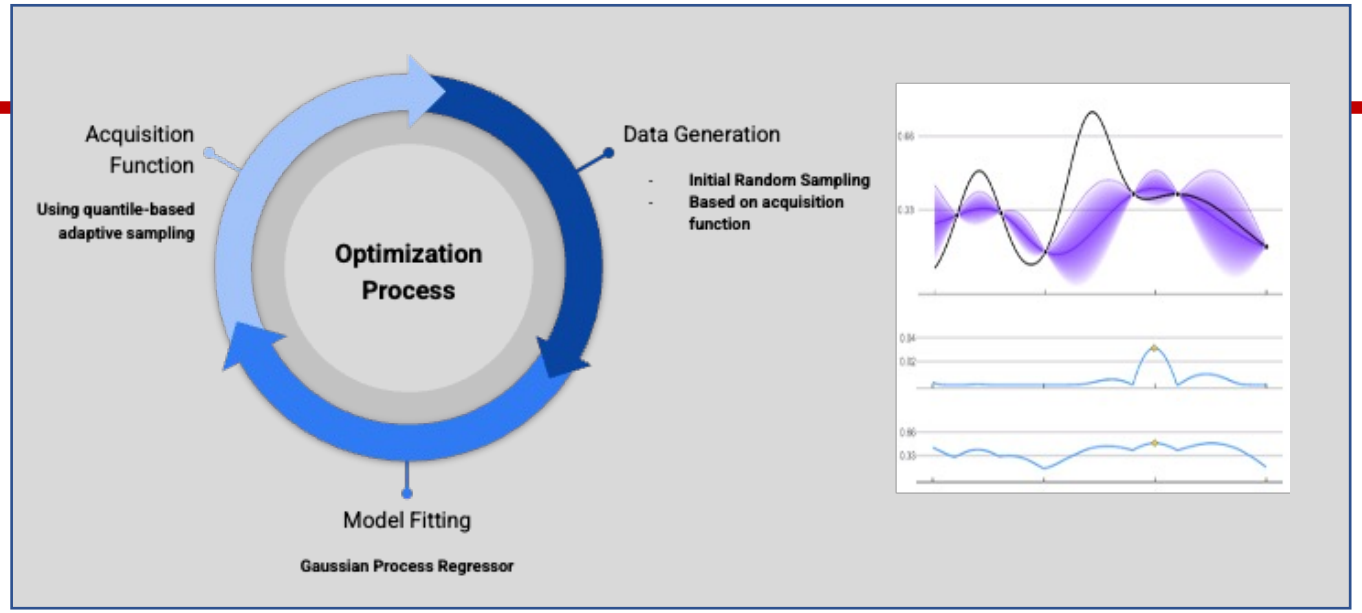
* https://arxiv.org/abs/1802.03426

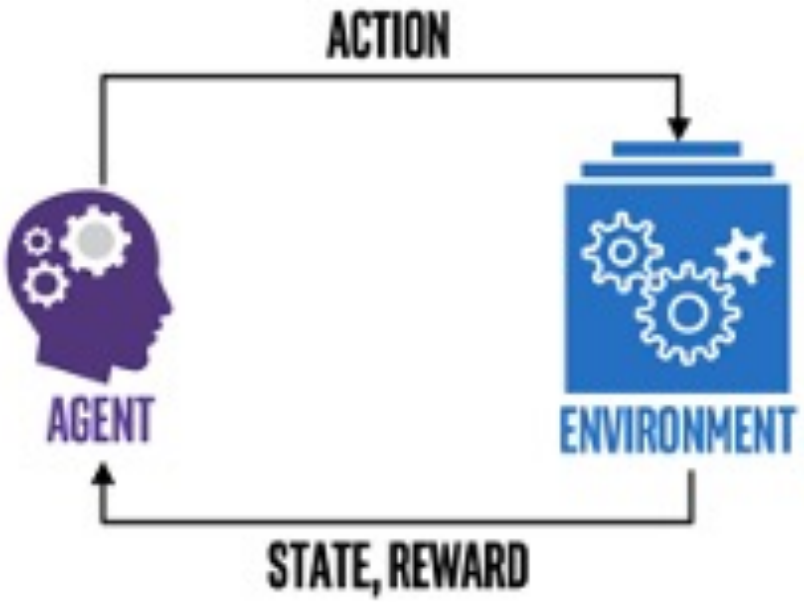
OAK RIDGE National Laboratory | SPALLATION NEUTRON SOURCE

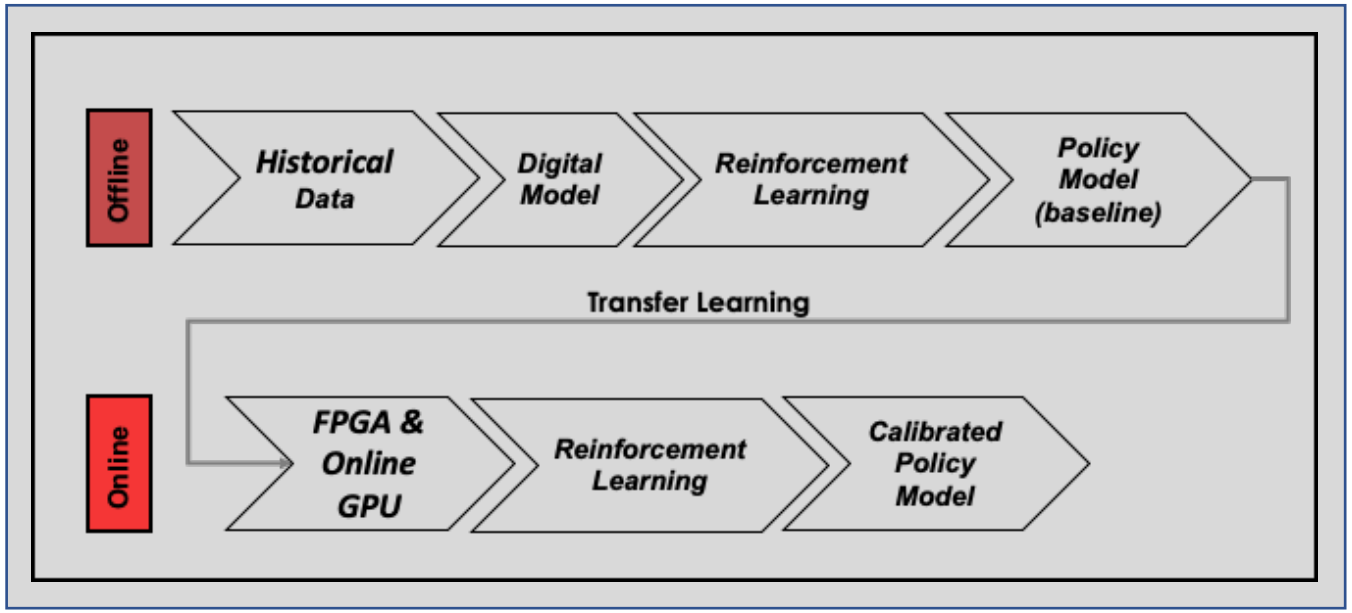
Jefferson Lab

Design & Control

- Advanced applications for design & control
  - Bayesian Optimization
  - Genetic Algorithms
  - Model Predictive Control
  - Reinforcement Learning





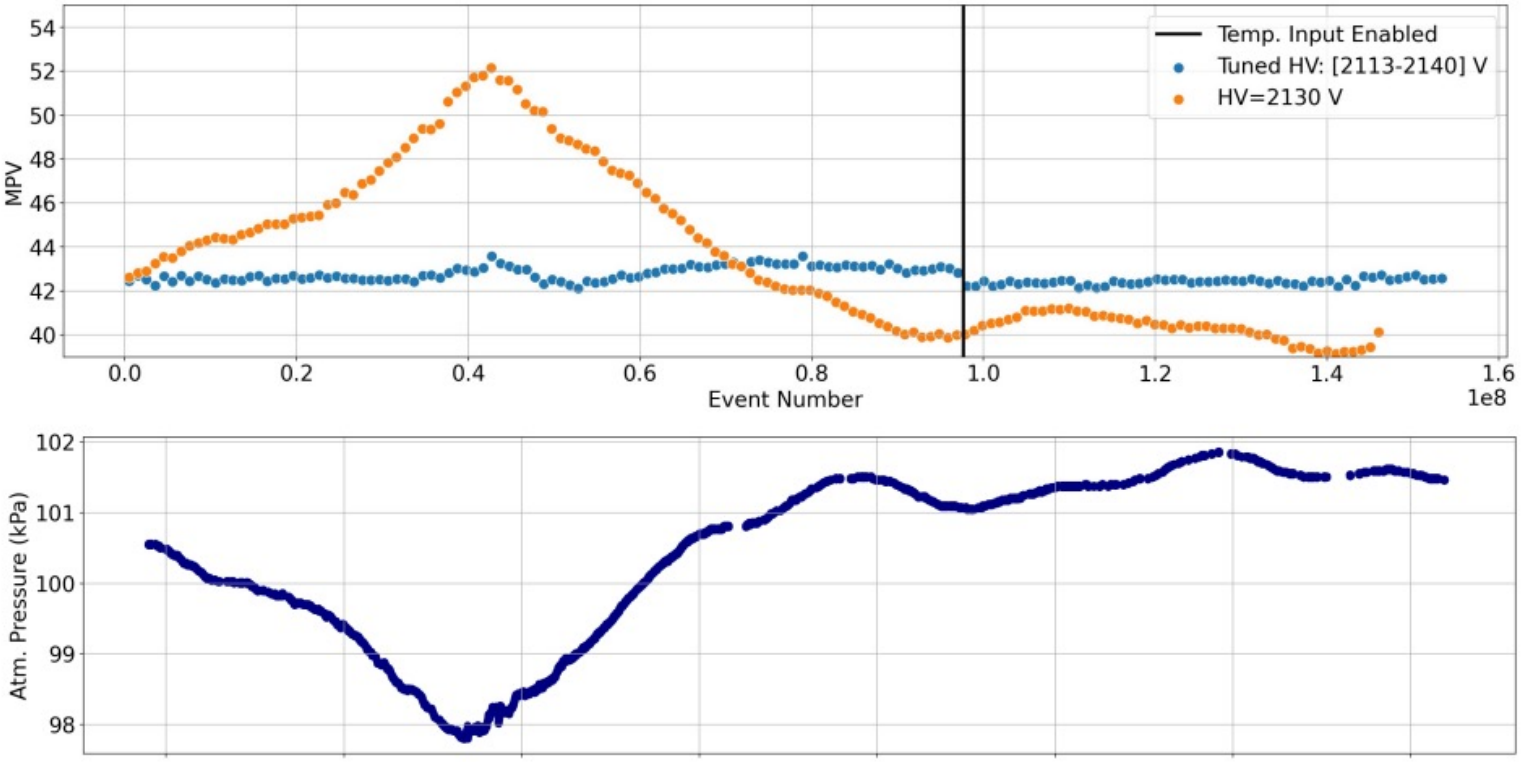


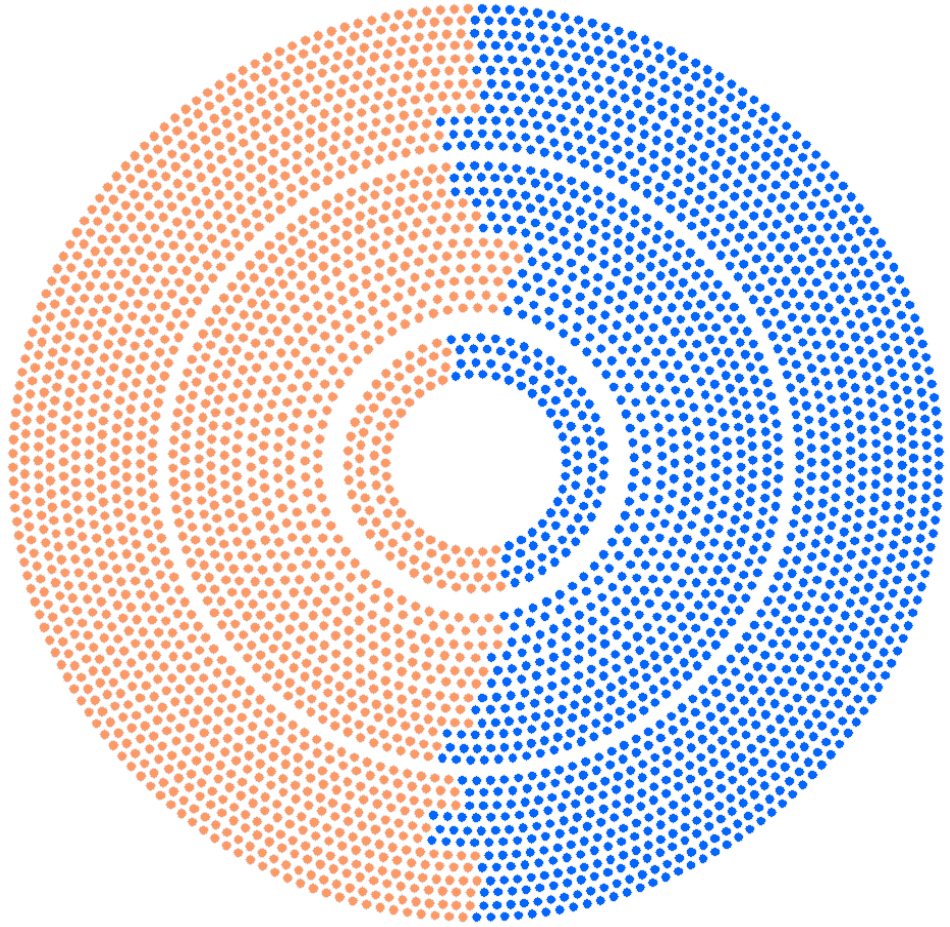
Jefferson Lab

Controls for Detectors

Accelerate the calibration from month(s) to minute(s):

- **Peak heights from Gaussian Process side** of the CDC show dramatic reduction in pressure dependence compared to **constant HV**



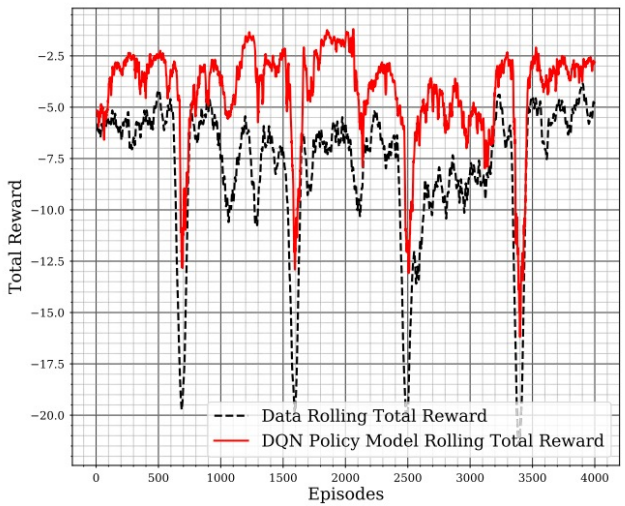


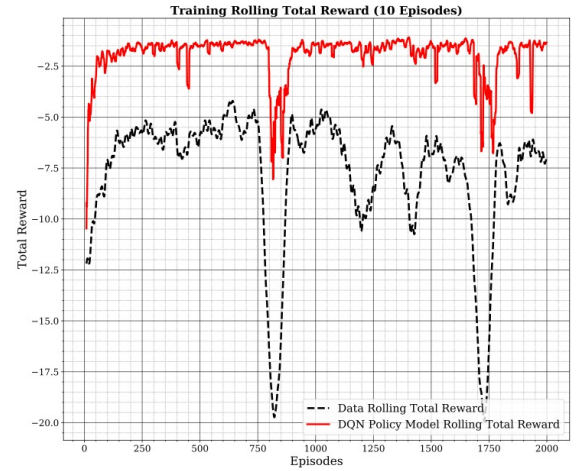
Schematic of downstream view of CDC, with straws HV control status indicated.

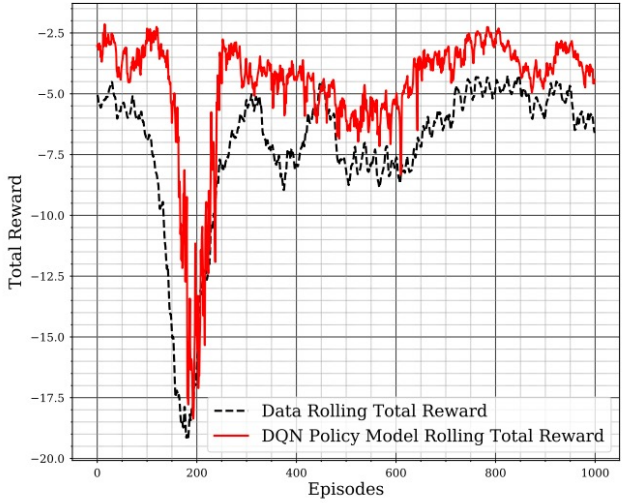
Jefferson Lab

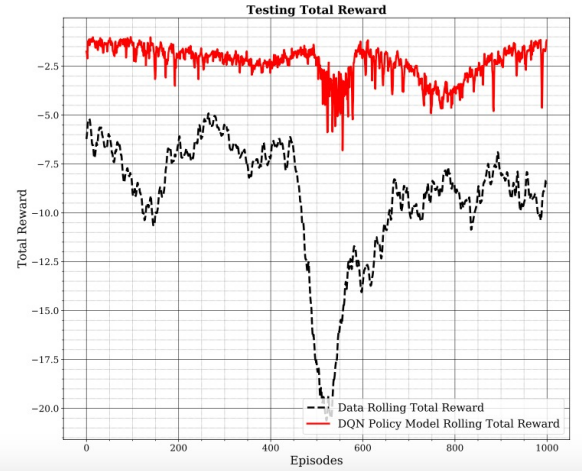
Controls for Accelerators

- We used a DDQN RL agent
- Original results used a stacked LSTM model yielding ~2x improvement over the original control system
  - Real-time artificial intelligence for accelerator control: A study at the Fermilab Booster (https://journals.aps.org/prab/abstract/10.1103/PhysRevAccelBeams.24.104601)
- Second study used a BNN to incorporate uncertainty quantification (calibrated) and showed improved results and stability:
  - Developing Robust Digital Twins and reinforcement learning for accelerator control systems at the Fermilab Booster (https://arxiv.org/abs/2105.12847)





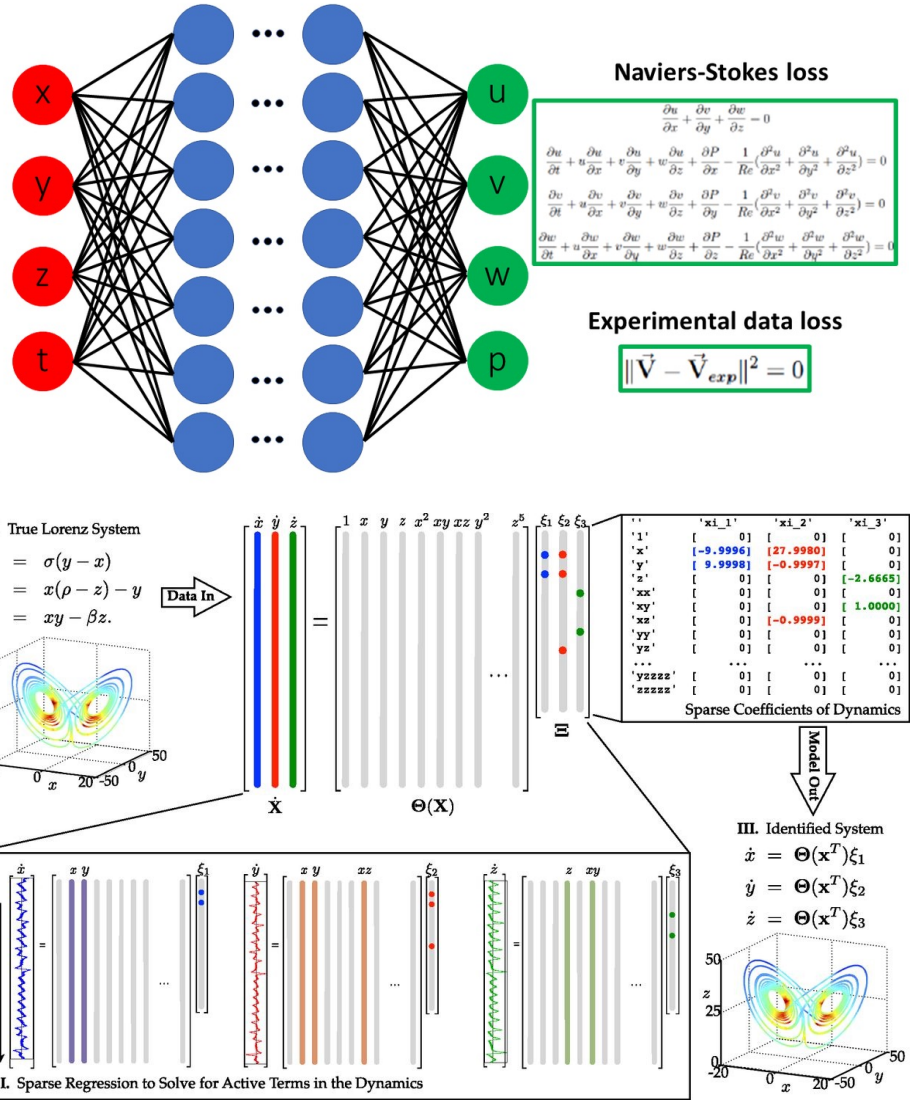




Jefferson Lab

A lot of interesting ideas to look forward to

- **Uncertainty quantification for deep learning models**
  - Detailed study of uncertainty estimation techniques for AI/ML in NP applications as it relates to higher dimensionality and unique modalities.
  - Applications of UQ ML models on edge hardware (under constraint)
- **Scalable Distributed Learning**
  - In order to efficiently train over large datasets, the need for a distributed learning computing infrastructure will likely be required
- **Techniques to advance scientific discovery**
  - Sparse Identification of Nonlinear Dynamical Systems (SINDy) is an algorithm to discover governing dynamical equations
  - Similar techniques could be applied to research at JLab
- **Techniques for explicit physics knowledge integration**
  - Applications of automatic differentiation through known physics equations into the ML models
  - Low energy nuclear physics examples have shown some improved results



Jefferson Lab


Thank you