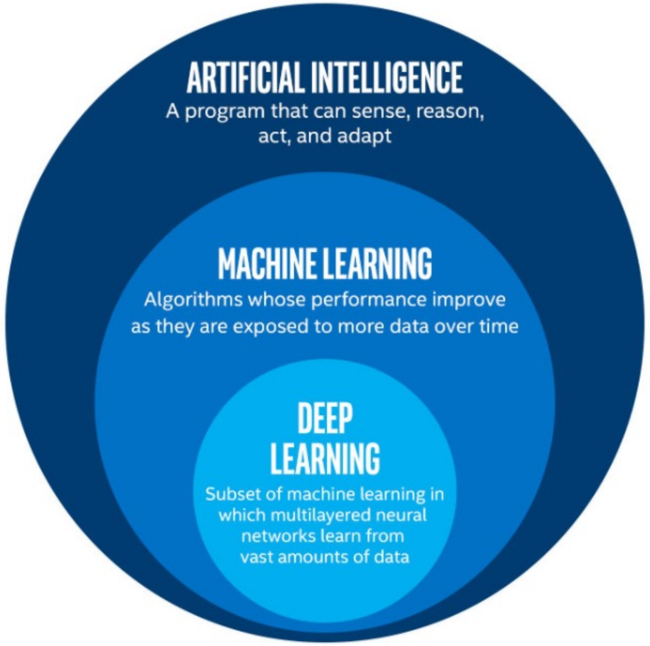
Data Science Department



Malachi Schram, Ph.D.
Head of the Data Science Department
Thomas Jefferson National Accelerator Facility

Overview

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. Provide world-class solutions to scientific challenges
    2. Provide real-time optimization solutions for complex operational instruments
    3. Champion education and research opportunities with regional Universities and industry
    4. Reduce the carbon footprint by optimizing the data science workflow and algorithms

FY22 LAB S&T Agenda, Milestones – Data Science & AI/ML

Overall Goals:

- Establish JLab as the data science hub for nuclear physics and regional scientific applications
- Develop a core capability targeting key elements of the Basic Research Needs (BRNs) defined in the SciML report, etc.
- Provide education and research opportunities to regional universities and industries

Yearly Activities and Milestones:

- Participate in HEP, NP, ASCR, and SciDAC proposal calls (expand as opportunities become available)
- Participate in DOE data science related workshops and BRN reports
- Continuously reevaluate based on State of the Art (SOTA)
- Expand scientific applications and collaborations

FY22:

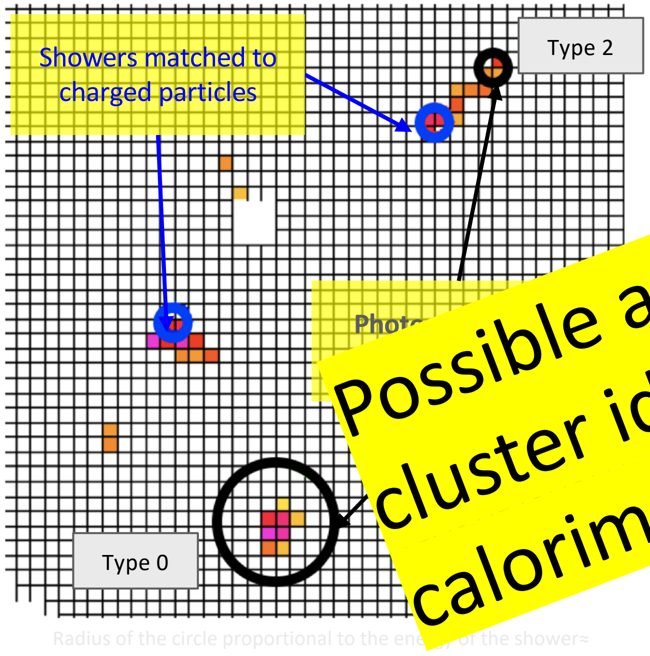
- **Establish a data science priority research needs and data availability for JLab:**
  - **Experimental Halls, Theory, Accelerator, and Facilities**
- Establish collaborations with regional universities and national laboratories
- Develop core capabilities:
  - *Infrastructure*:
    - Evaluate latest SOTA workflow tools for data science
    - Evaluate datasets and model repositories
    - Identify/evaluate needs for digital twin
  - *Methods & Algorithms (addressing the needs of the SciML BRN):*
    - Expand capability in ML-based ***uncertainty quantification*** techniques
    - Develop ***interpretability*** techniques
    - Expand optimization research for ***design and control***
    - Domain-aware ML

- **Potential Applications:**
  - **Clustering and Particle Identification**
  - **Anomaly/similarity detection/prediction**
  - **Calibration**
  - **Design and control**

GlueX FCAL Cluster Identification

R. Barsotti and M.R. Shepherd 2020 *JINST* 15 P05021

- Separation of **electromagnetic** and **hadronic** interactions (i.e., low energy vs split-offs) in the GlueX forward calorimeter (2800 lead glass modules)
- Algorithm trained on using $\omega \rightarrow \pi^+\pi^-\pi^0(\gamma\gamma)$ meson decays which contain both true photons and charged particles interacting with the calorimeter



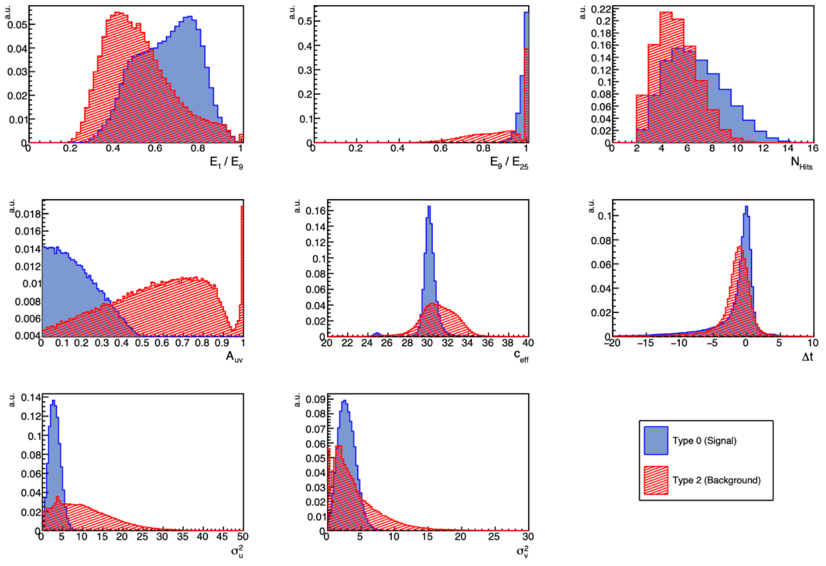
Showers classifi…

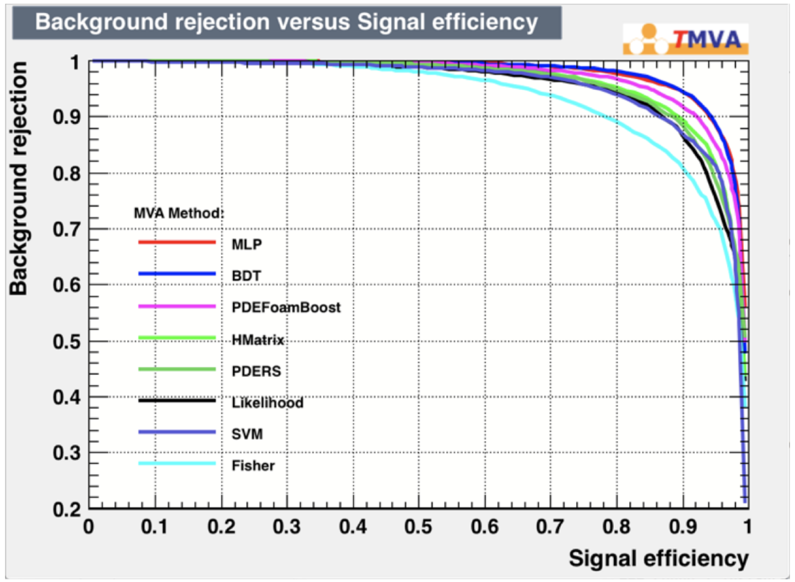
…g., π^0)

…Type 1 or

…energy deposit, hits, time and …etry of clusters

Possible application in Hall C for cluster identification in new calorimeters, e.g., the NPS?





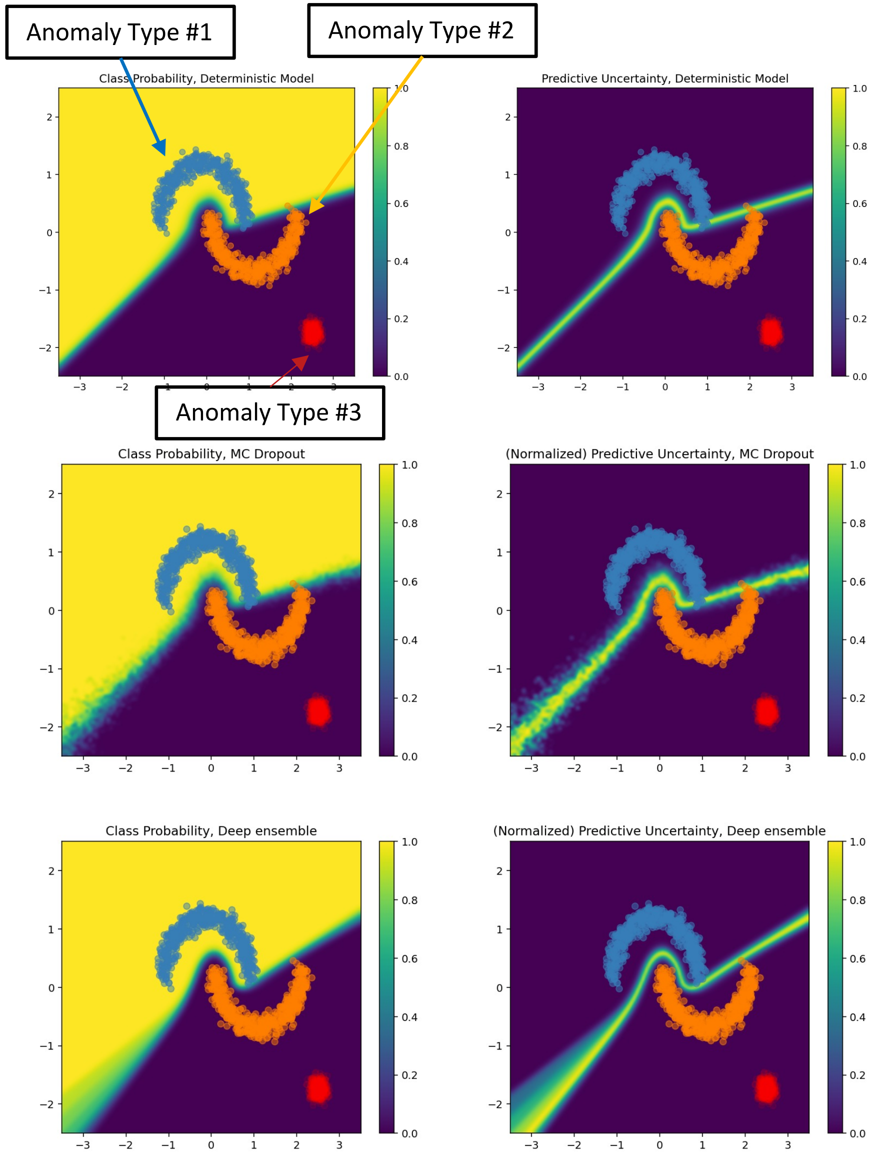
- MLP selected due to ease of implementation within the GlueX software framework
- Thorough data/MC comparison studies
- Background reduction of 60% and signal retention of 85% on inclusive π^0 data. MC and data efficiency when studied with ω events agree within the statistical precision

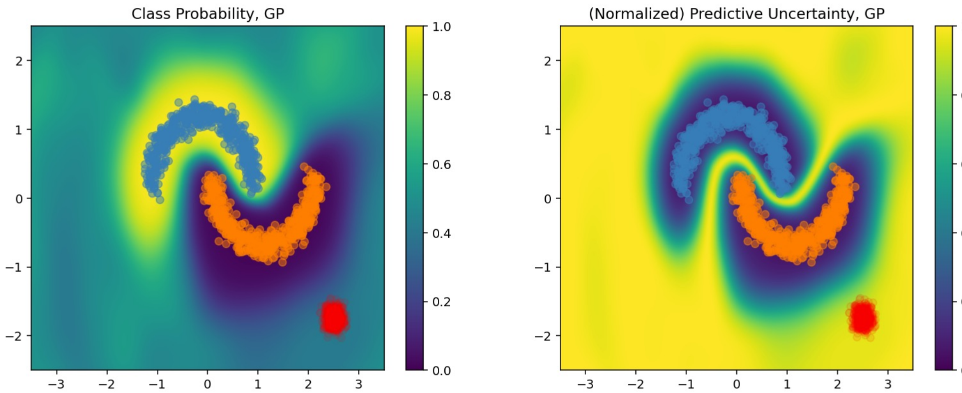
Understand what your model knows and doesn't know

Different method yield vastly different **classification predictions**, some examples:

- Deterministic
- MC Dropout
- Deep Ensemble
- Gaussian Processes
- Bayesian Neural Networks

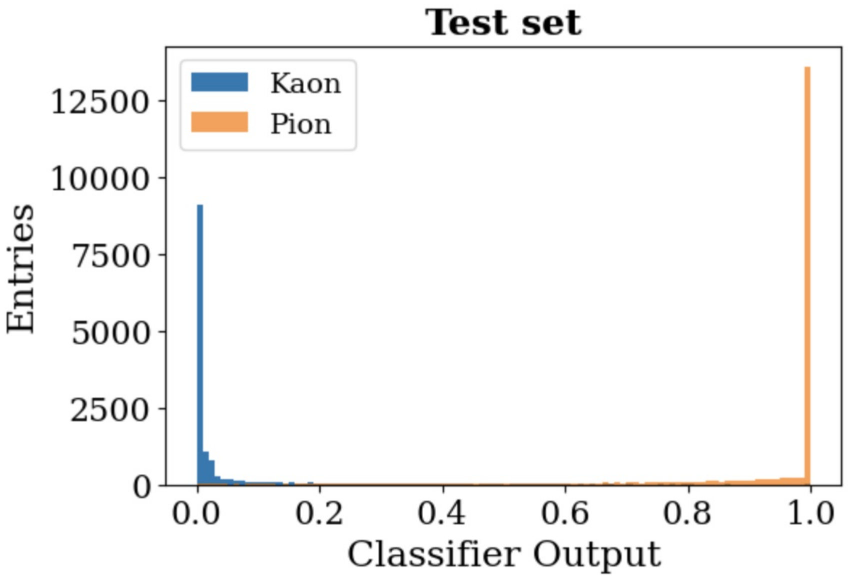
- Different models architectures can yield better results if you do not know all classifications

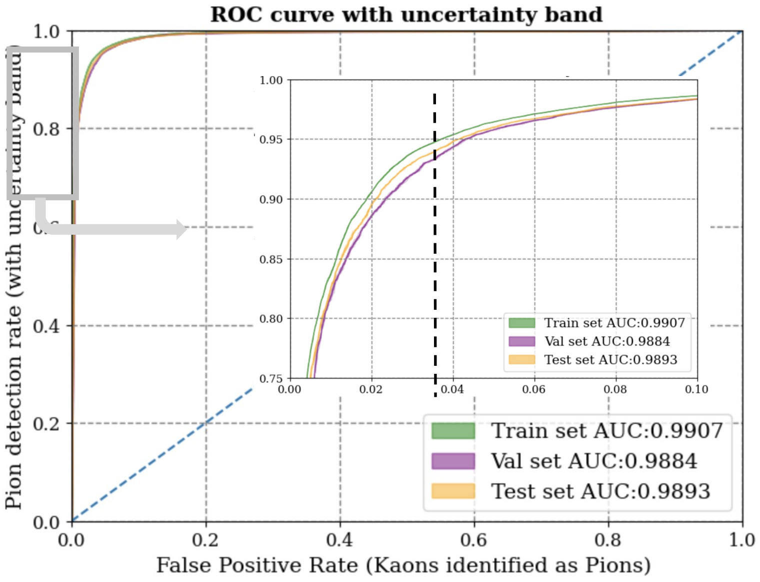


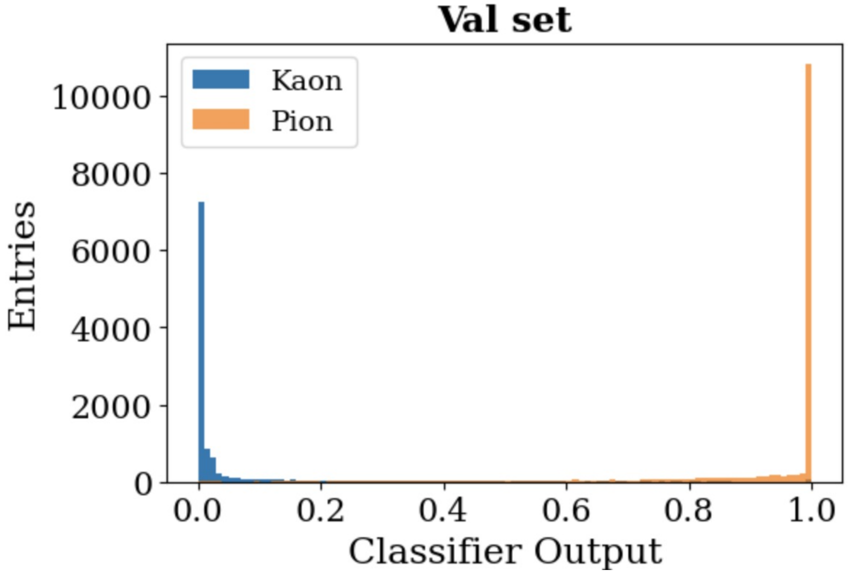


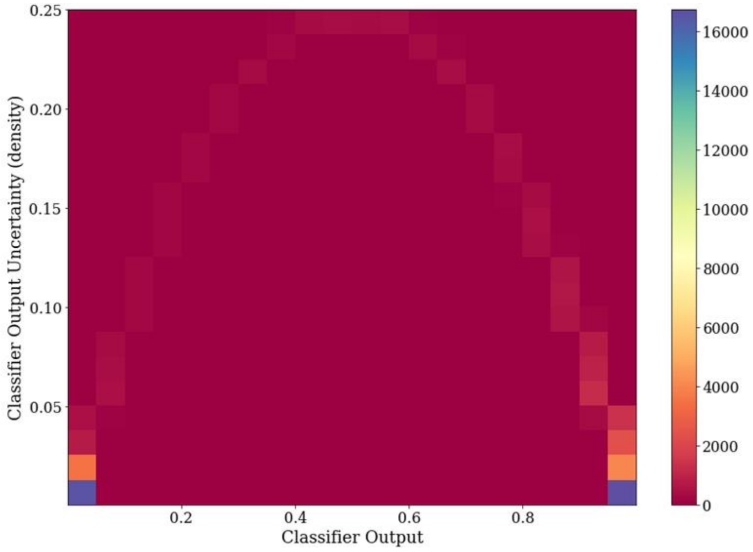
Particle identification for SoLID with uncertainty quantification

- The initial goal was to achieve >95% pion efficiency while keeping false positive below 5%
- Distance aware model provides uncertainty values associated with each output
- Performing hyper parameter optimization can potentially improve the accuracy of the model
- We are able to achieve the initial goal on the most difficult kinematic and smallest resolution readout
- With increase in the resolution using other readouts, we believe it is possible to further improve the efficiency





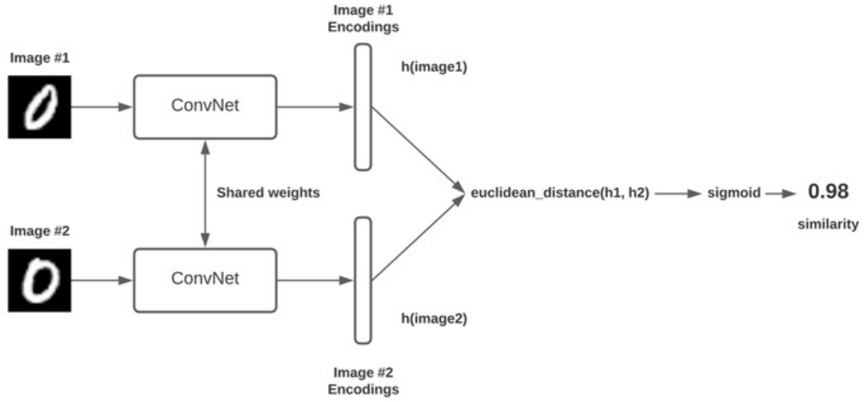


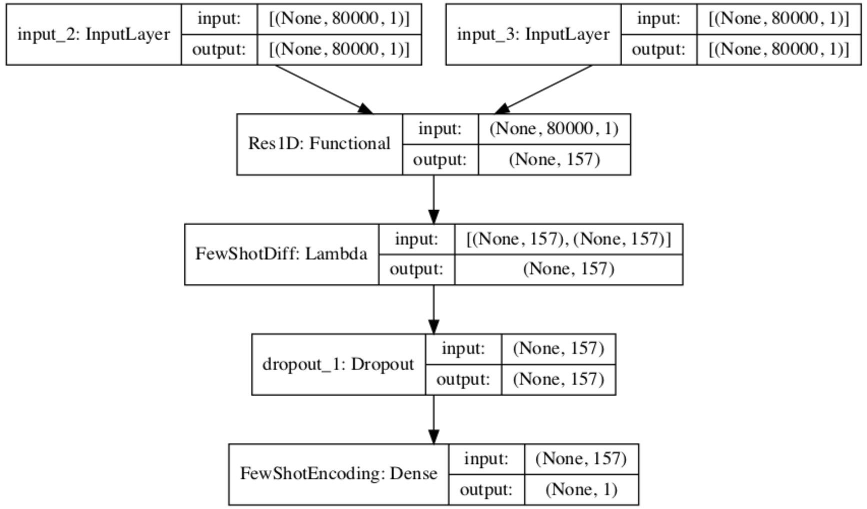


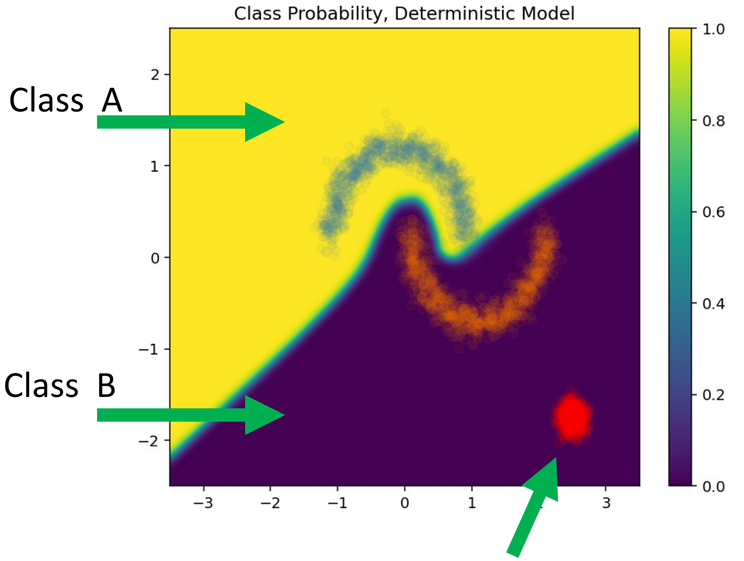
Leveraging similarities to identify anomalies

Problem definition:
- Predict anomalous pulses before they happen

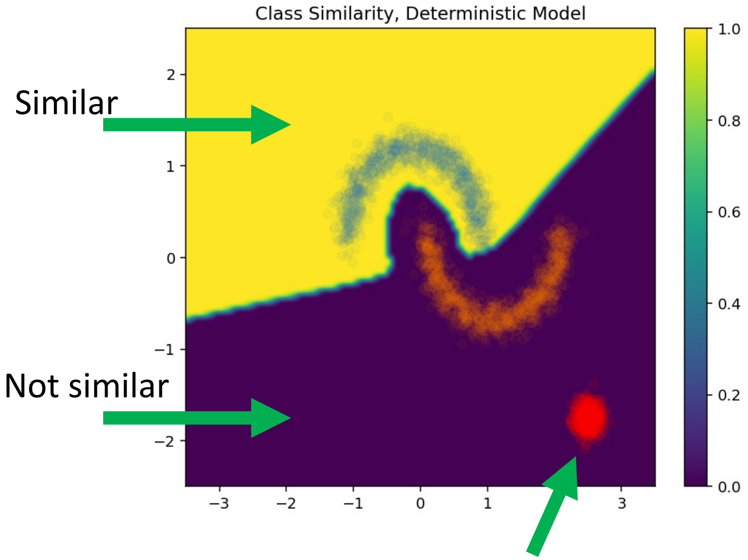
Proposed solution
- We are using a Siamese model since we want to focus on the **similarity** between a reference pulse and the current pulse
- Siamese model does not explicitly model the classification but focuses on the similarities
- We enhanced our Siamese model by adding a GP layer that provides an uncertainty estimate





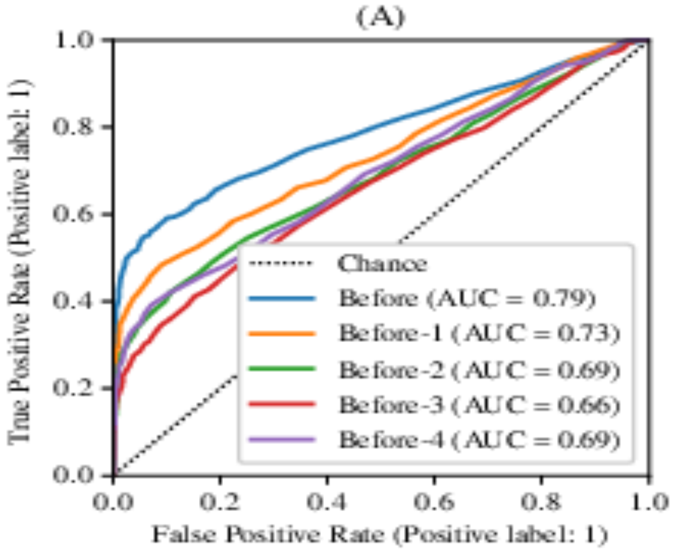


Completely miss identified the out of domain sample.



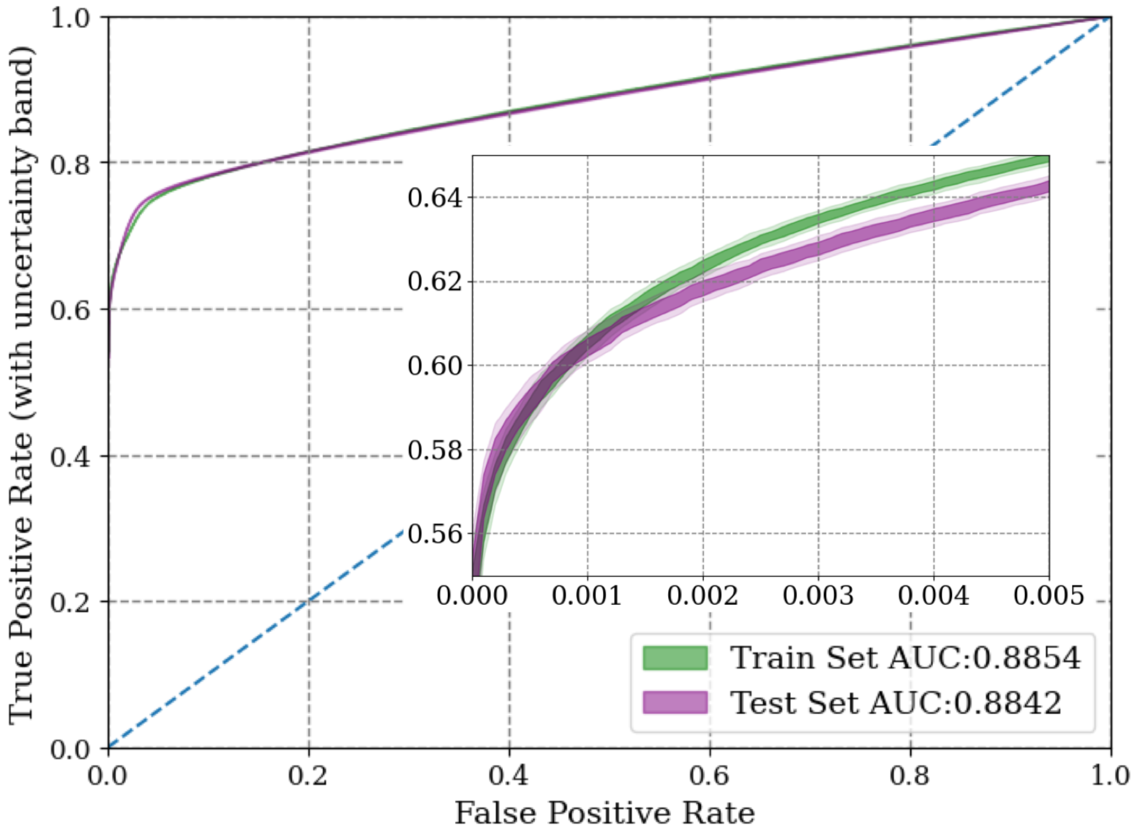
Correctly identified as different from the reference data

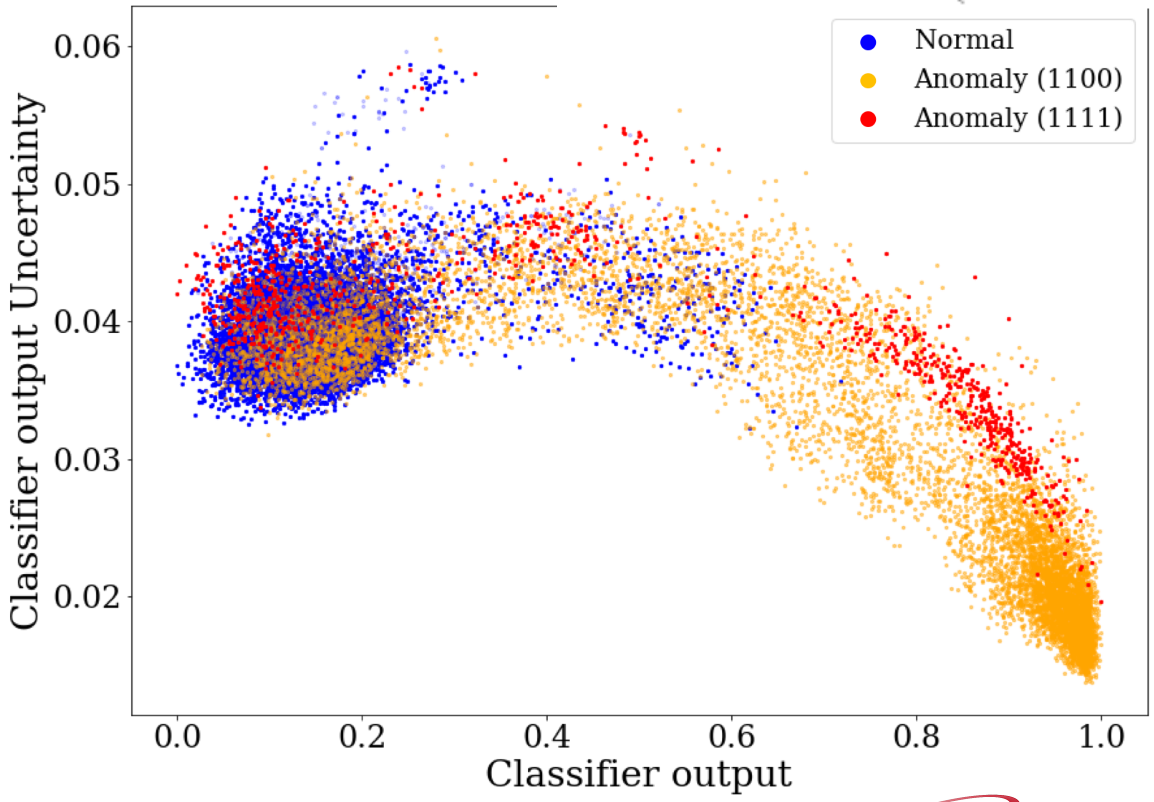
Uncertainty quantification for accelerator pulses



Results

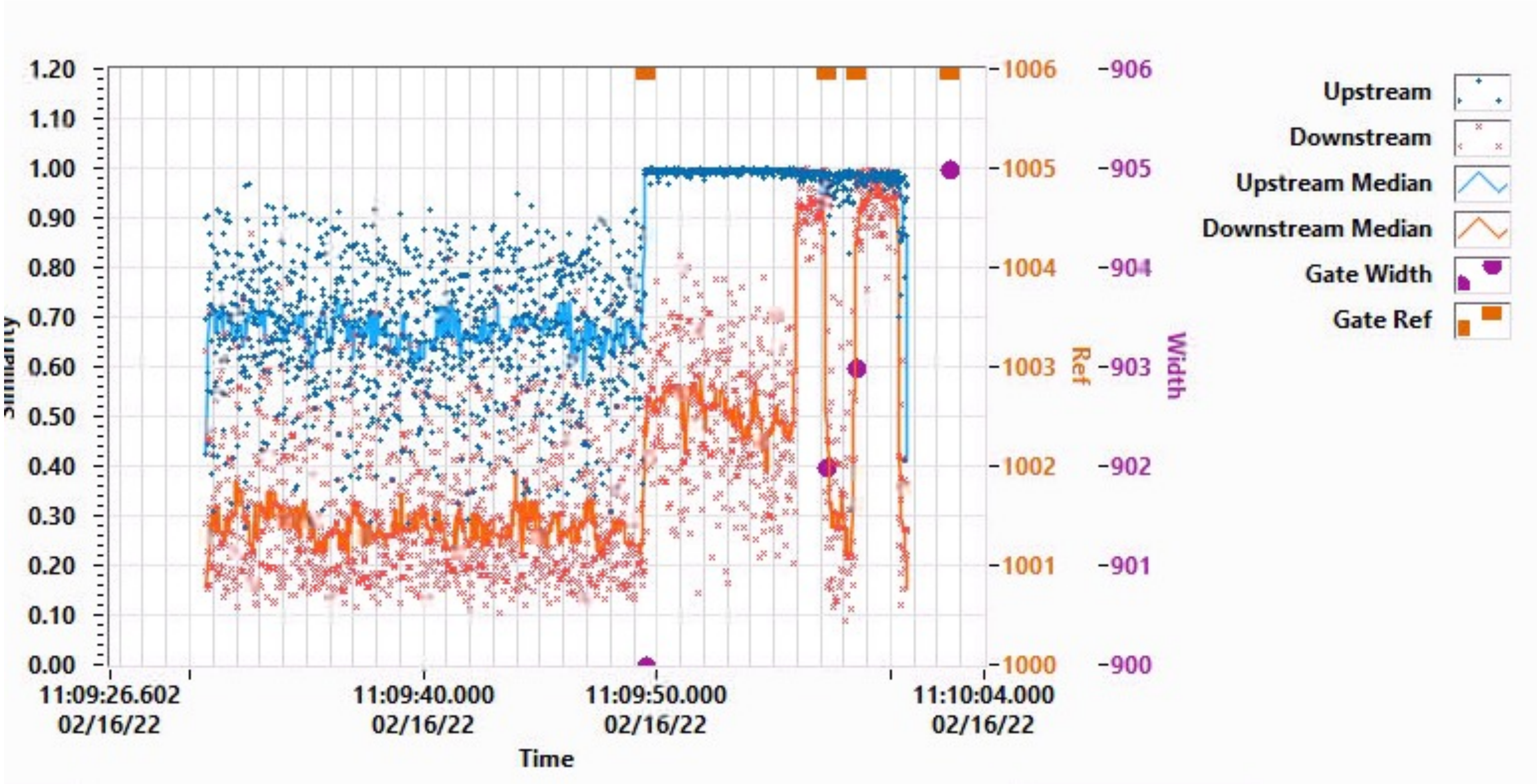
- The similarity model has ~4x better performance than published classification model
- The ROC curve shows nearly the same level of performance (not optimized)
- We introduced an **out-of-domain anomaly**, labelled 1111 (red), the UQ-based model correctly identified the anomaly and indicated high uncertainty.





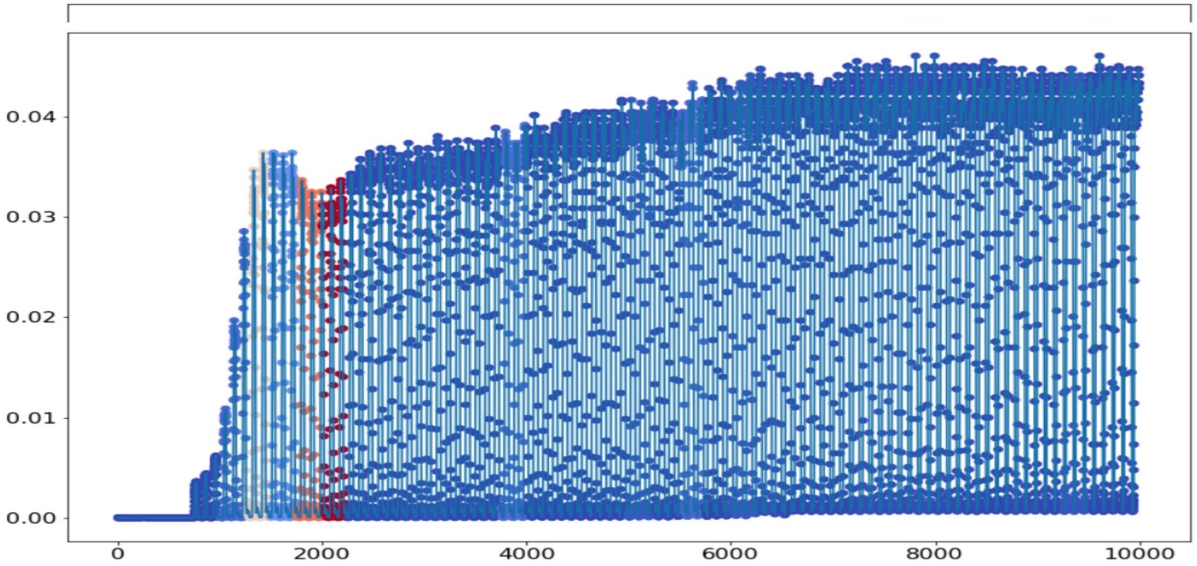
Online monitoring for system changes and anomalies/similarities

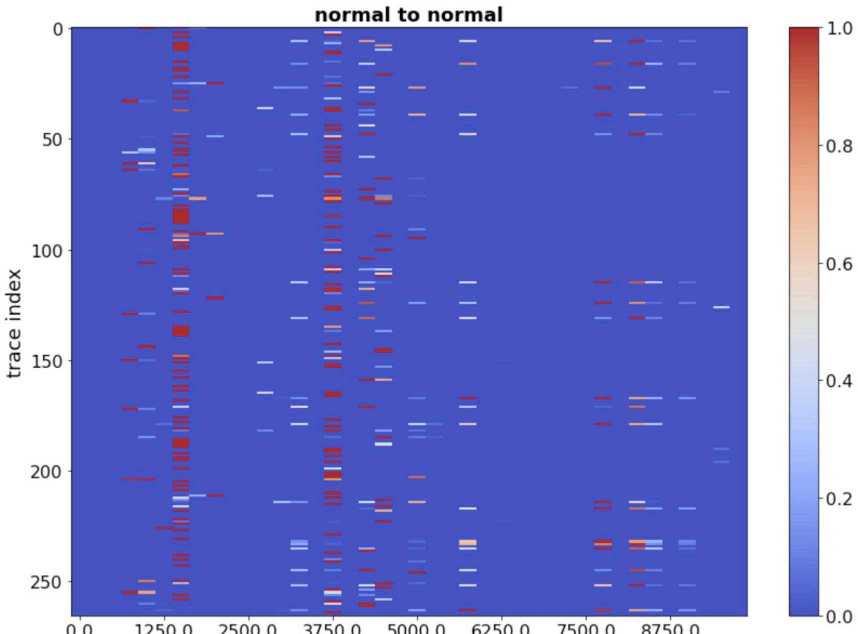
- Real-time ML-based techniques for monitoring and system change identification
- Real-time calibration using multi-model data sets

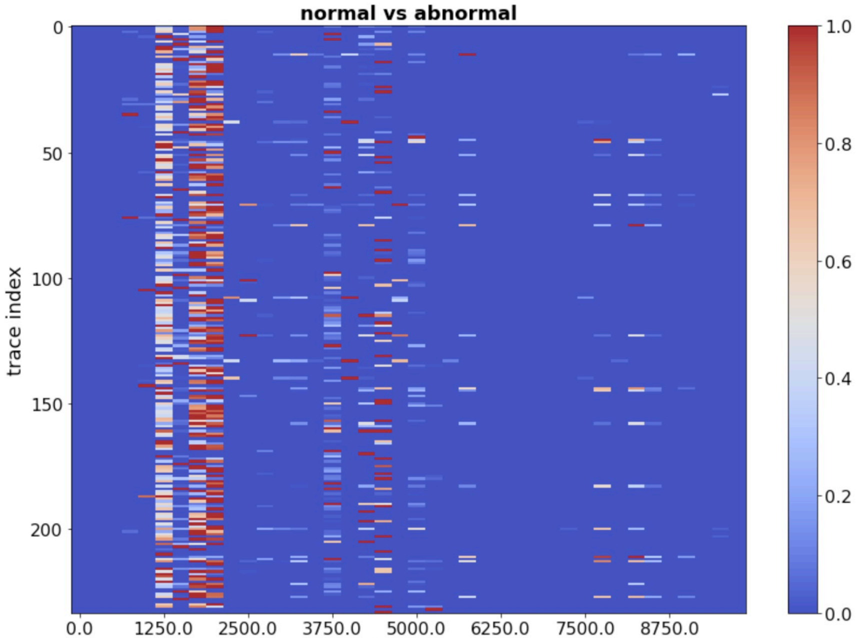


Trying to interpret the model

- We are exploring techniques such as Gradient Class Activation Map (GradCAM).
- This enables us to ‘see’ important activation regions in the model.
- These results tell us the regions in the data that are most relevant when distinguishing normal and abnormal operations conditions.
- It also allows us to reduce the input data size which reduces the inference time. This is critical for online decision making.
- Clustering on the GradCAM results can be used to classify the anomalies







Data science tools for experiment control and calibration

Scientific Achievement

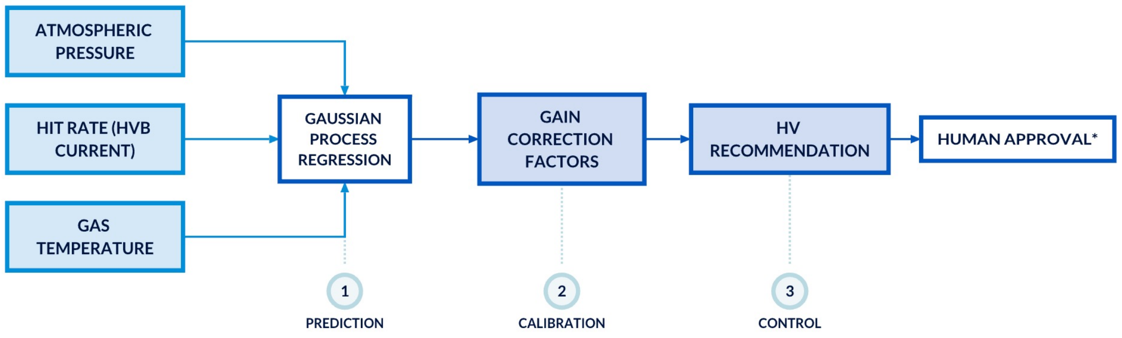
An data science system to calibrate and control the GlueX Central Drift Chamber (CDC).

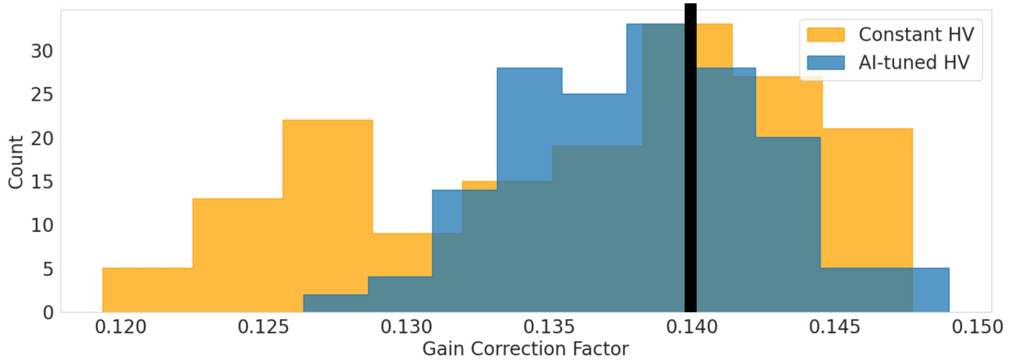
Significance and Impact

The Gaussian Process Regression using atmospheric pressure and CDC specific features. The data science system recommends anode voltages and calibration values for the detector in response to changing pressure. Controlling CDC anode voltage stabilizes the response of the detector throughout the duration of the experiment and reduces the calibration efforts required afterwards.

Research Details

- Developed by team of Data Scientists and Physicists
- The system uses detector specific and environmental measurements as input features to Gaussian Process Regression model to predict calibration values and determine voltage settings throughout the experiment.
- Shift takers can accept or reject the voltage recommendation.
- Successfully implemented during PrimEx run period.
- Understanding systematic and model uncertainty is in progress.
- Application to other detector systems at the Lab in progress.





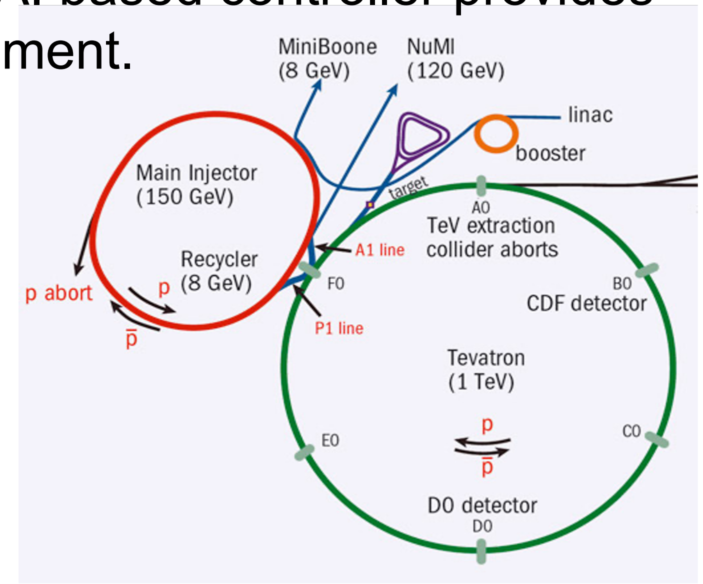
Top plot: Schematic of the data science system. Bottom plot: CDC Gain Correction Factor with GP-tuned (orange) and constant (blue) high voltage. Black line is target gain. Using the tuned HV settings produced less variation in the Gain Correction Factor compared to constant HV running conditions.

Autonomous Control:
Reinforcement Learning for Accelerator Control at FNAL

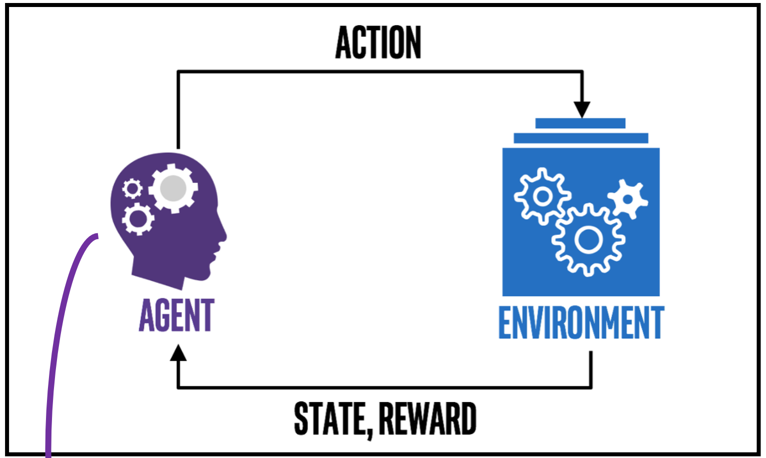
Problem definition:
Reduce beam losses in the FNAL Booster by developing a Machine Learning (ML) model that provides an optimal set of actions for accelerator controls

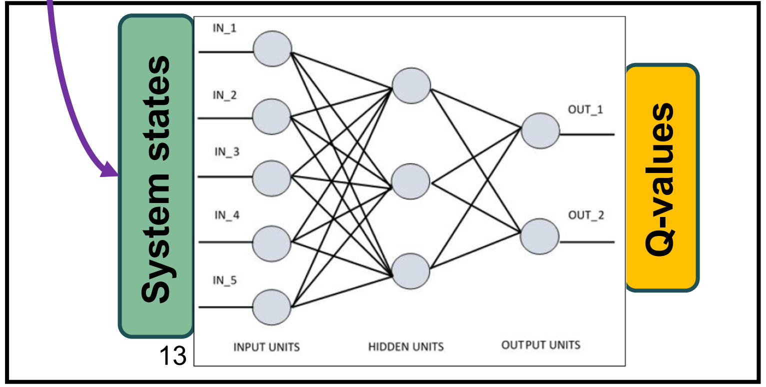
Approach:
Developed a surrogate model and a reinforcement learning policy model for online control system

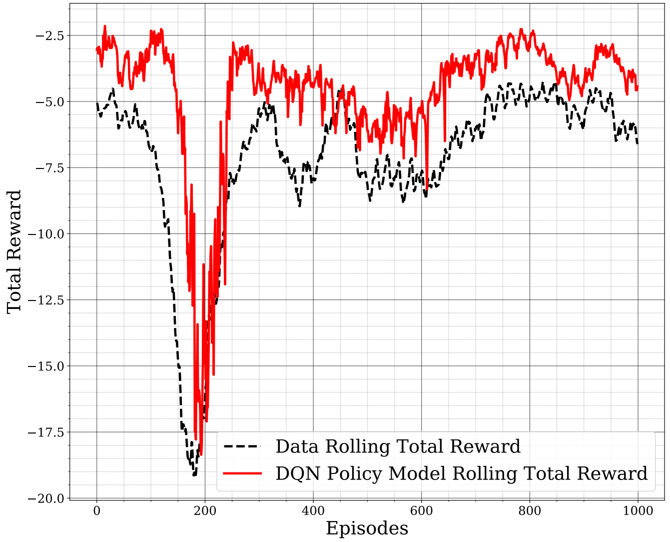
Results:
The current AI based controller provides ~2x improvement.

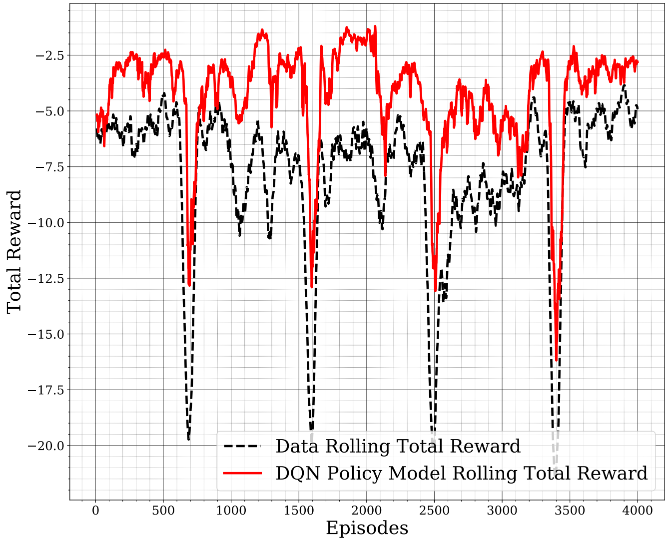


Courtesy: Christian Herwig

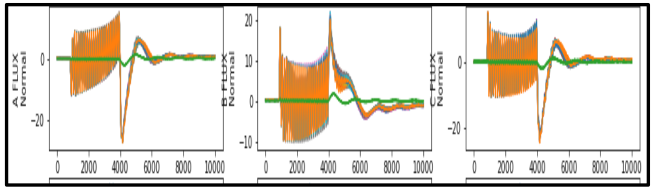


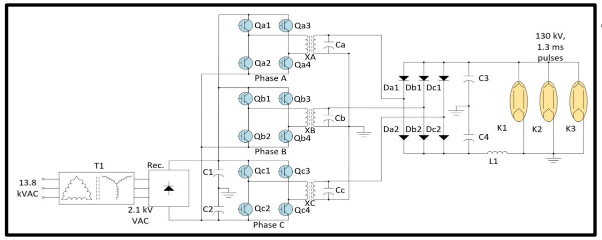


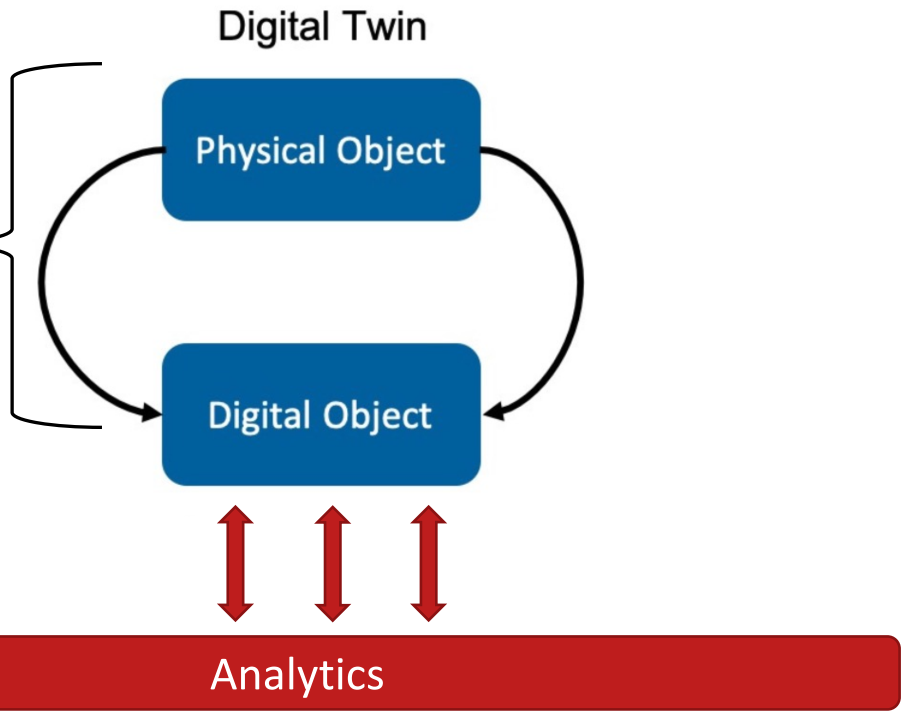




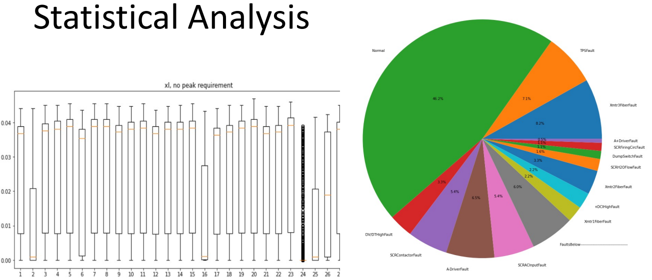
Example of integrated data science analytics workflow



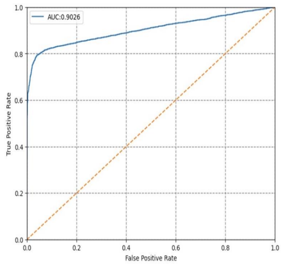




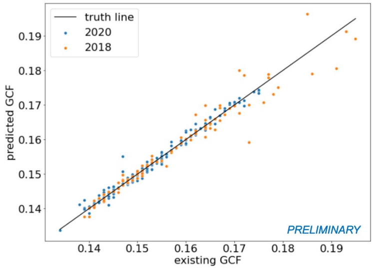
Statistical Analysis



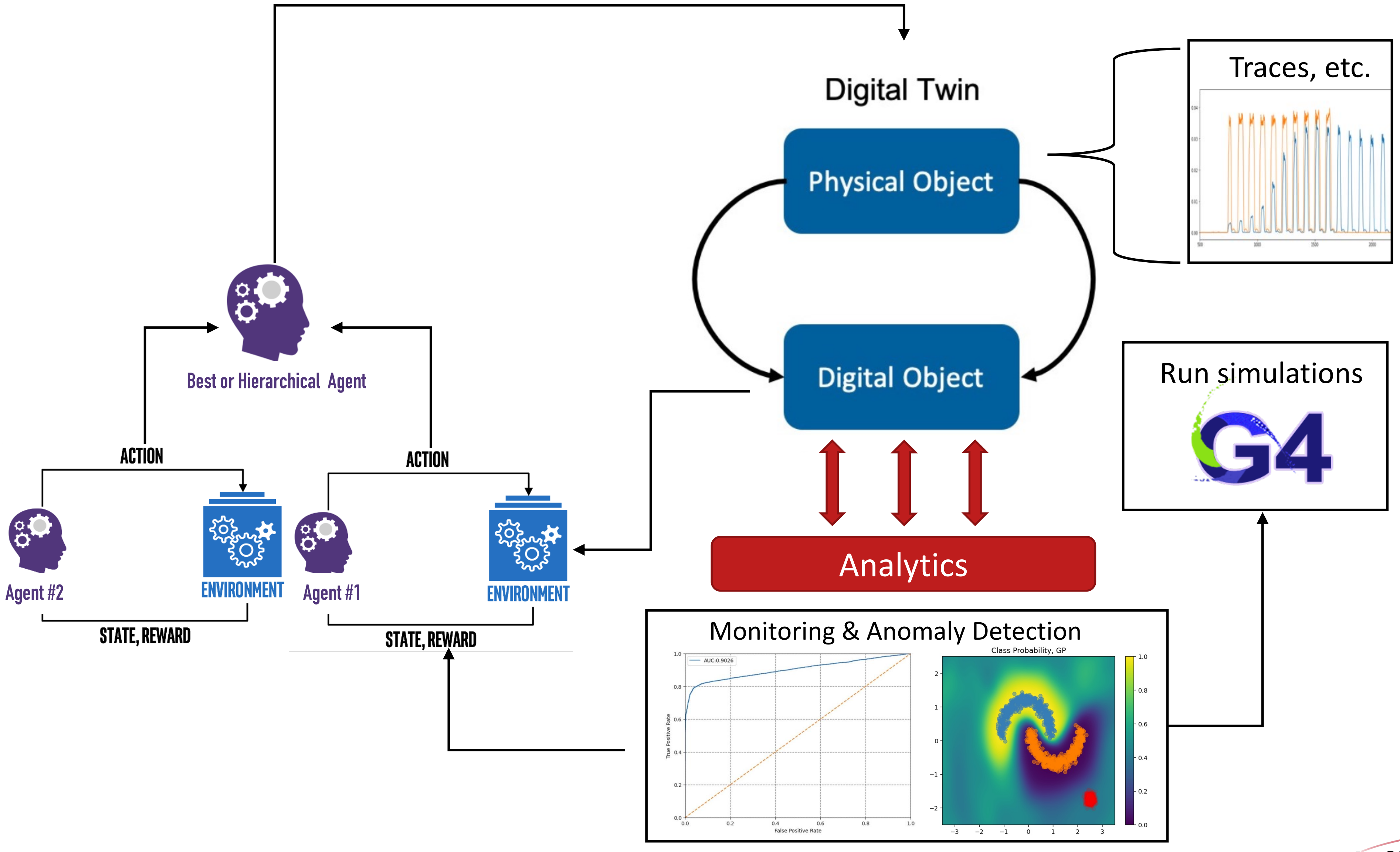
Monitoring & Anomaly Detection



Calibration



Autonomous experimental monitoring, anomaly detection/prediction, and control



Jefferson Lab

- New Data Science department (~6 months)
- Multiple internal meetings to identify areas where data science can improve the scientific mission
- Multiple external meetings to identify areas for new collaborations with regional universities and laboratories
- Building a core infrastructure to improve efficiency, reproducibility, and collaborations across divisions
- Building core technical foundations to address identified needs within the DOE scientific portfolio

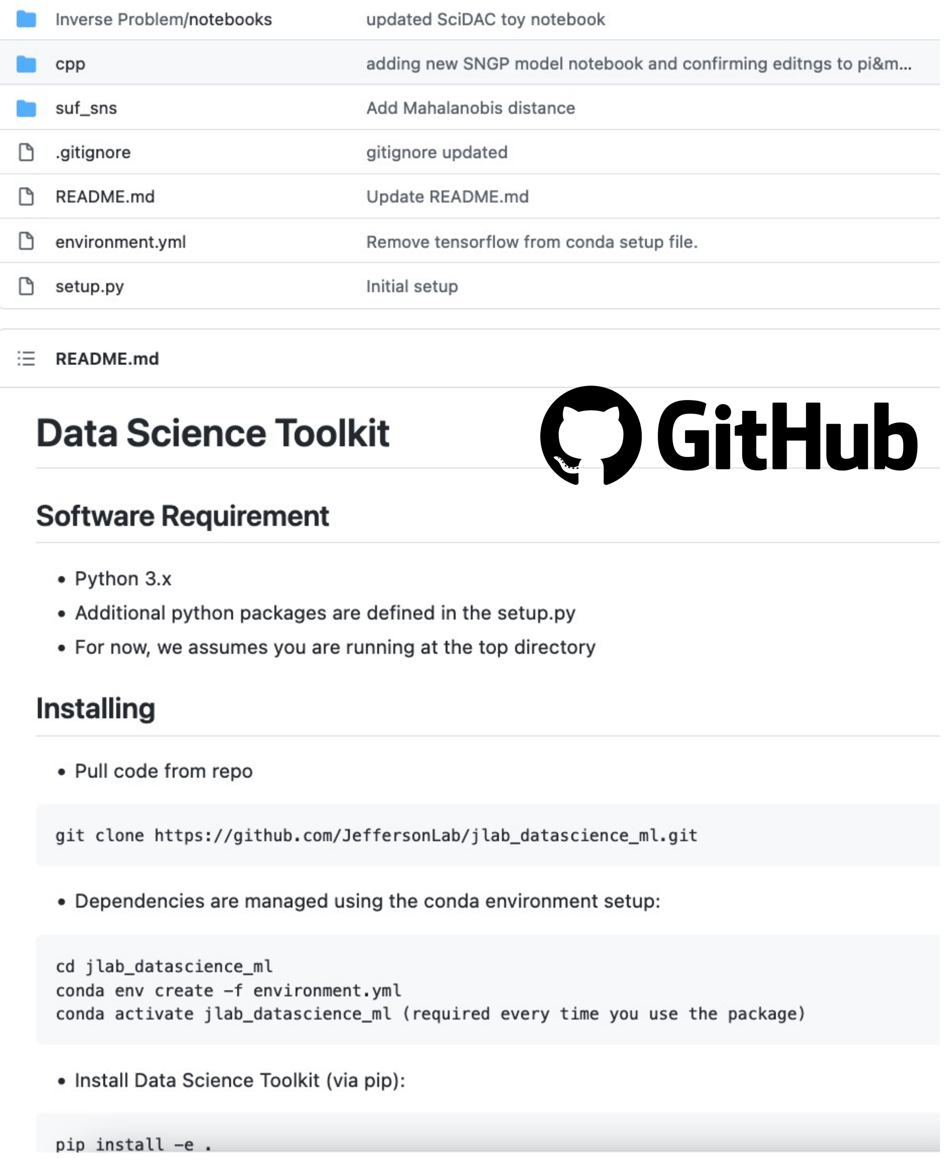
We look forward to having more detailed discussions with Hall C

THANK YOU

- **Develop core capabilities:**
  - **Infrastructure:**
    - **Setup data science software repository and best practice**
    - **Evaluate/implement datasets and model repositories**

Developing a production data science infrastructure for NP and beyond

- Software repository:
  - Create a collaborative environment and shared knowledge
  - Reusability of common algorithms and tools
  - Better overall efficiency
- Unit test:
  - Ensures that all code blocks work as designed, captures code complexity, enables continuous integration
- Continuous integrations:
  - Ensure that new changes in the package are tested and identifies issues early in the development cycle



Inverse Problem/notebooks	updated SciDAC toy notebook
cpp	adding new SNGP model notebook and confirming editngs to pi&m...
suf_sns	Add Mahalanobis distance
.gitignore	gitignore updated
README.md	Update README.md
environment.yml	Remove tensorflow from conda setup file.
setup.py	Initial setup

README.md

GitHub

Data Science Toolkit

Software Requirement

- Python 3.x
- Additional python packages are defined in the setup.py
- For now, we assumes you are running at the top directory

Installing

- Pull code from repo

```
git clone https://github.com/JeffersonLab/jlab_datascience_ml.git
```

- Dependencies are managed using the conda environment setup:

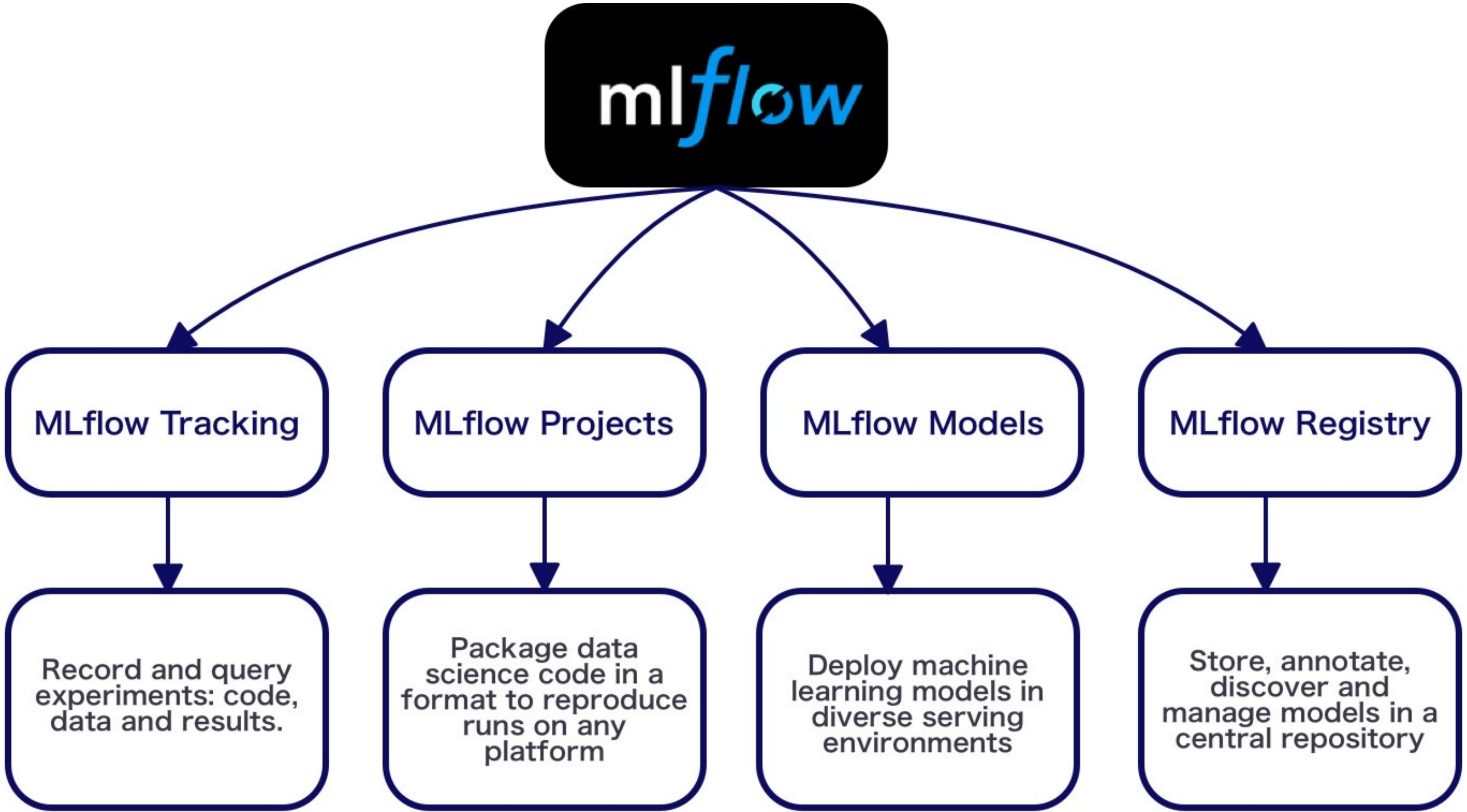
```
cd jlab_datascience_ml
conda env create -f environment.yml
conda activate jlab_datascience_ml (required every time you use the package)
```

- Install Data Science Toolkit (via pip):

```
pip install -e .
```

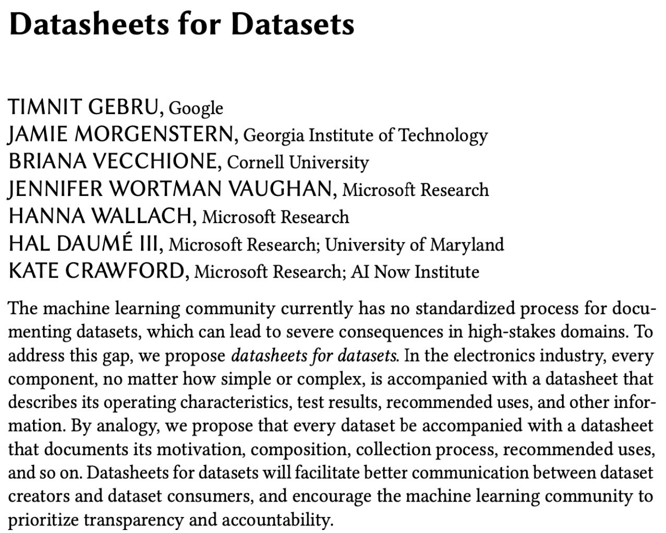
Across JLab: ML repositories

- JLab scientists are developing a lot of ML models from scratch
- There is currently no central system to capture models and artifacts or to find pre-existing models
- There is an opportunity to improve productivity and reduce the labs carbon footprint
- JLab needs a common repository to capture the ML provenance for all models used for operations, etc.
- We are currently evaluating www.mlflow.org as a service to satisfy this requirement



Across JLab: Dataset

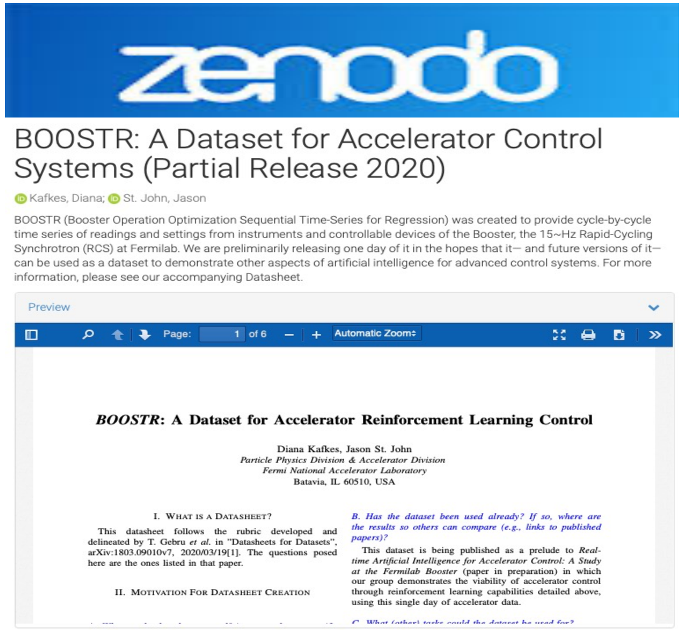
- JLab has a lot of datasets that can be shared for algorithm development
- Sharing datasets can provide insights in their commonalities
- We are working to develop a private collection of JLab specific datasets that will allow us to easily collaborate and quickly evaluate algorithms
- We are currently exploring https://inveniosoftware.org/



Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets.* In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.



Jefferson Lab

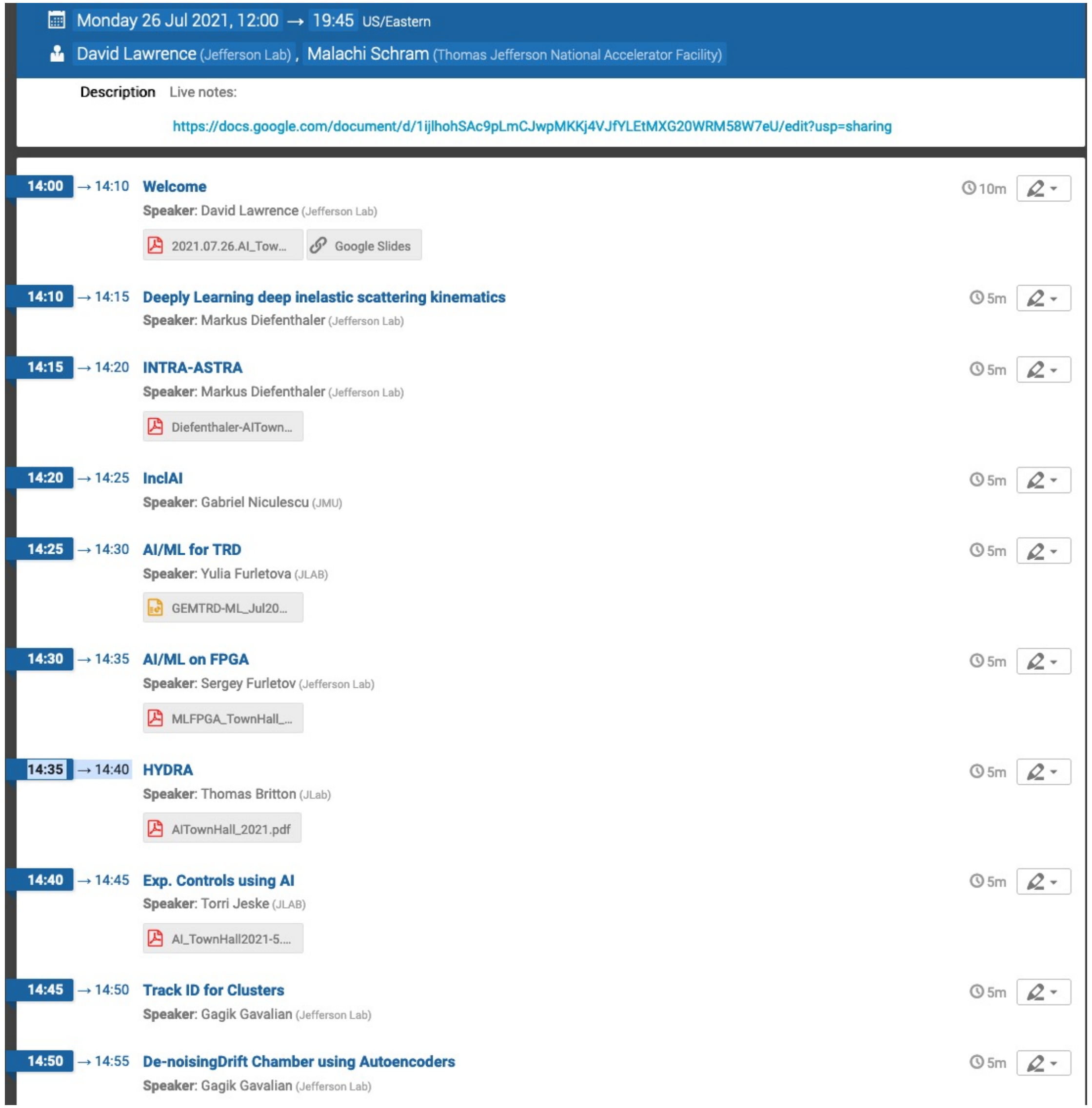
Developing a Data Science priorities to support JLab science:

- Experimental Halls and Theory
- Accelerator and Facilities

AI/ML in Physics and Theory Division

- AI Townhall (July 26th 2021)
  - EPSCI (lead) and Data Science Dept. (contributed) to the second AI/ML Townhall to review the current activities in the experimental halls and theory.
- Numerous follow-up meetings with each experimental hall. Some initial ideas for collaborations:
  - Hall A:
    - SoLID tracking and particle identification
  - Hall D:
    - CPP particle identification
    - Calibration
    - Anomaly detection
  - Theory

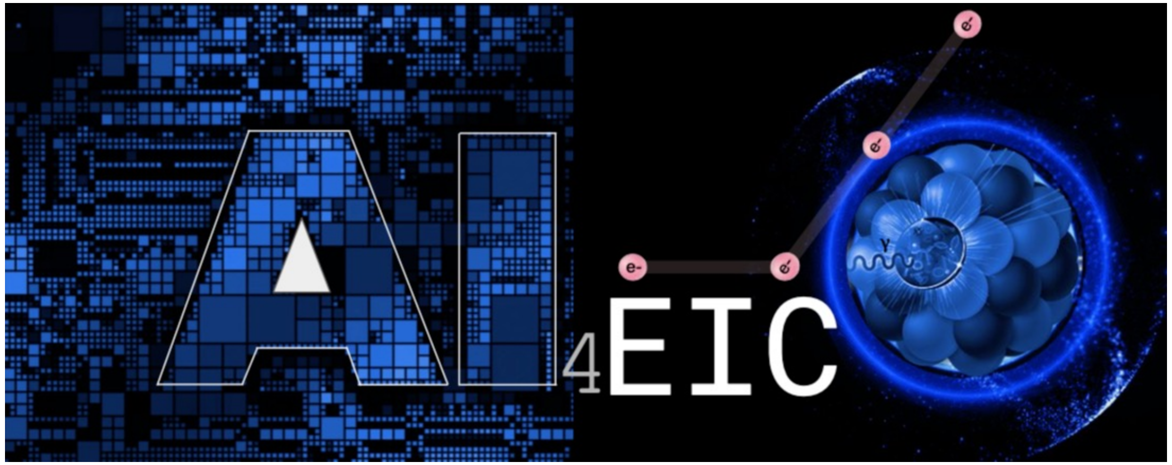
Some common needs have been identified

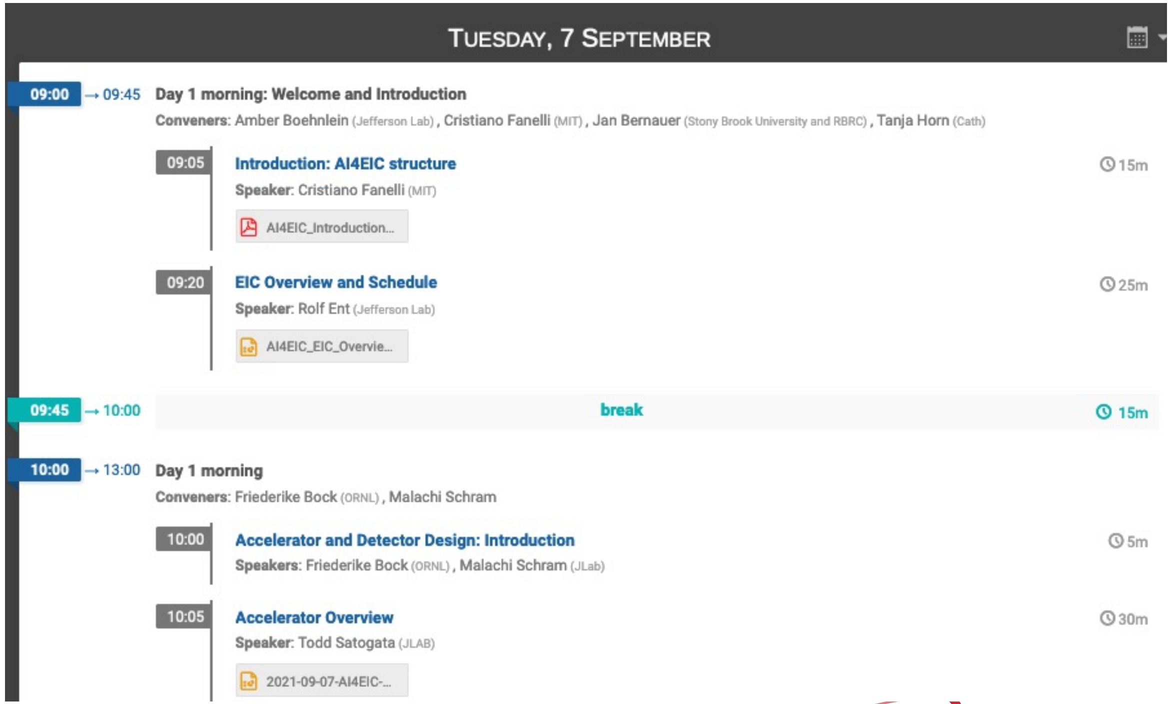


Hall E (EIC)

- AI4EIC Workshop (Sept. 7-10th)
  - Data science dept. participated in the first AI focused workshop for EIC covering:
    - Accelerator and Detector Design
    - Simulations
    - Reconstruction and Analysis
    - Accelerator and Detector Control
    - Detector Readout
    - Real-time AI tracking and tagging
    - Computing Frontiers

- No definitive research plans have been developed but will resume now that the detector proposals are done





AI/ML in Accelerator Division

- AI Townhall (Nov. 12th 2021):
  - Accelerator division (lead) and Data Science Dept. (contributed) to review the current activities
  - Contributed several talks
- Co-lead a LDRD proposal on MORL for CEBAF with accelerator division
  - Project started Q1 FY22
- Contributed to STTR proposal
- Initiated data science bridge position with ODU with a focus on accelerator physics
  - Position now open

13:00	**Welcome and Introduction** *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Christopher Tennant* 13:00 - 13:15
	AI for Sparse-to-Dense Mapping of Site Radiation Dose *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Adam Stavola* 13:15 - 13:23
	LLRF DAQ for AI/ML *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Rama Bachimanchi* 13:23 - 13:31
	SRF Cavity Instability Detection *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Dennis Turner* 13:31 - 13:39
	C100 Fault Prediction *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Lasitha Vidyaratne* 13:39 - 13:47
	Uncertainty Aware Anomaly Detection for SNS *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Malachi Schram* 13:47 - 13:55
14:00	**Reinforcement Learning for Accelerator Control for FNAL** *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Malachi Schram* 13:55 - 14:03
	Multi Objective Optimization of Cryogenic Heat Load and Trip Rates *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Kishan Rajput* 14:03 - 14:11
	Global Orbit Locks *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Adam Carpenter* 14:11 - 14:19
	Field Emission Management *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Adam Carpenter* 14:19 - 14:27
	Thoughts to Improve the Performance of Polarized Electron Sources by AI/ML *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Shukui Zhang* 14:27 - 14:35
	Smart Alarm for the CEBAF Injector *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Christopher Tennant* 14:35 - 14:43
	Graph Analytics for CEBAF Operations *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Christopher Tennant* 14:43 - 14:51
	Semi-Autonomous Mobile Diagnostic *(Click on Go to Map for Bluejeans connection), BlueJeans*	*Christopher Tennant* 14:51 - 14:59
15:00		

AIEC: Using ML to Predict Calibrations for the Central Drift Chamber in GlueX

Objective:

Predict and/or adjust the controls of a sensitive detector to reduce or eliminate the need for calibration, i.e., Maintain consistent detector response to changing environmental/experimental conditions by adjusting CDC HV

Current SOTA:

Sensitive detectors need to be calibrated to obtain optimal resolution

Calibrations cause a delay between data collection and analysis (months)

- Requires multiple iterations are needed to converge to final set of constants
- MONTHS of time for traditional calibration using reconstructed data.

Proposed approach:

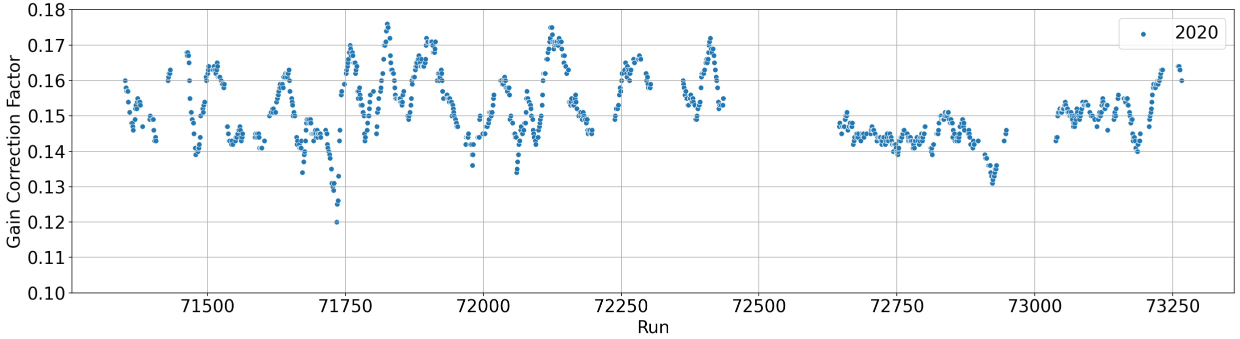
Used Gaussian Process Regression to predicted the desired calibrations and estimate the prediction uncertainty using environmental data to accelerate time to complete.

Targets: existing Gain Correction Factor (GCF) from GlueX 2020 run period

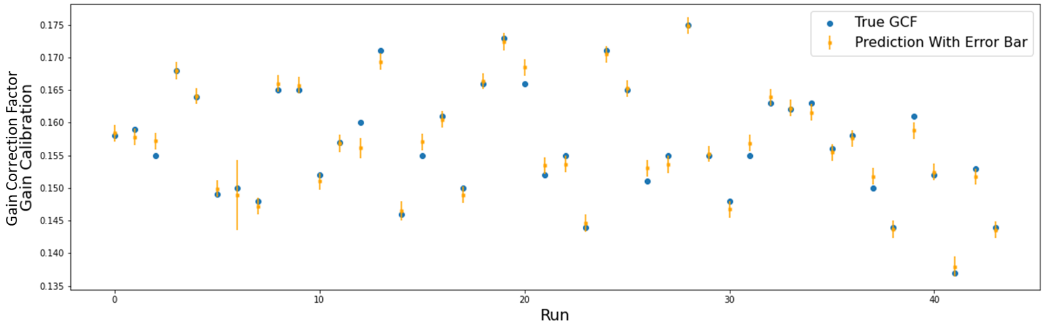
Initially Trained using GlueX 2020 data

Training variable used: Atmospheric pressure, HV board current, CDC gas temperature from previous minute to predict HV needed to produce relative gain.

Raw GlueX 2020 data:



Calibrated results for test data set:



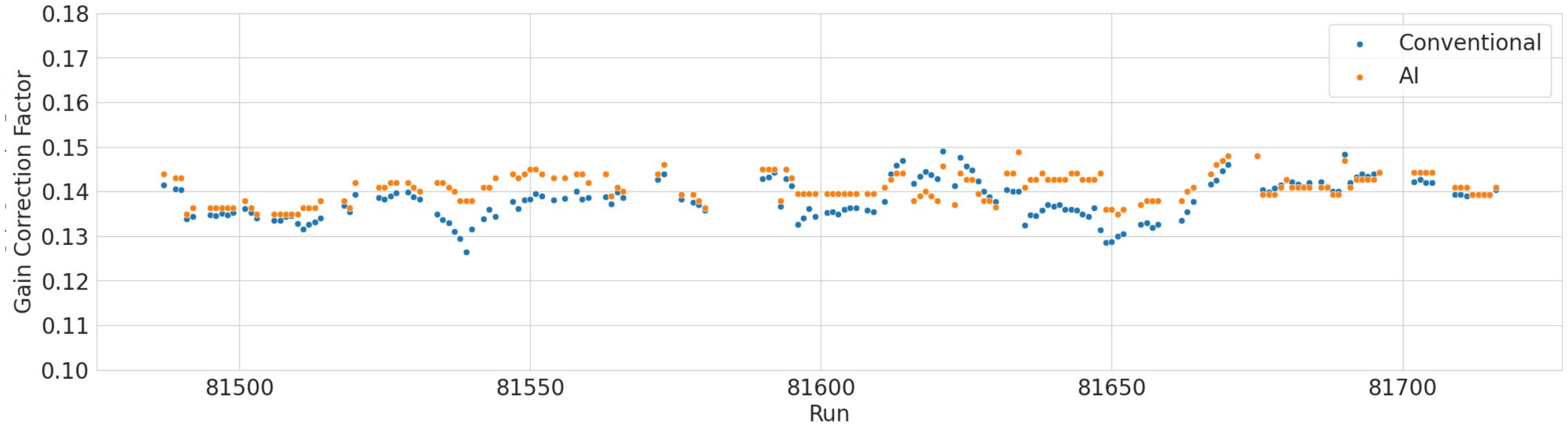
GPR based results are consistent to the baseline approach.

AIEC: Calibrations with ML: Gain with PrimeX-η

Once the model was training using GlueX dataset, we provide corrections to the new streaming PrimeX-η data for the shifter to apply.

Figure illustrates good agreement between the baseline and the GPR approach.

And it appears changing HV stabilizes relative gain.

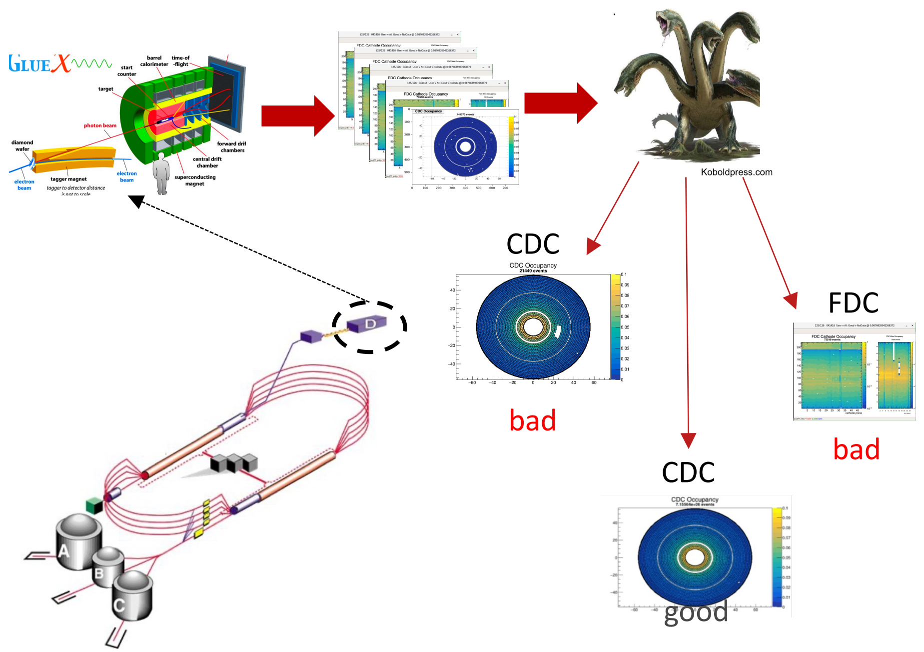


Hydra

Problem definition:

Detectors are complex, and delicate, the enormous amount of data generated by them are needed to be monitored to make ensure the quality of the physics research. Manual monitoring is tedious task and does not provide the granularity and efficiency as machine.

Solution (Hydra)

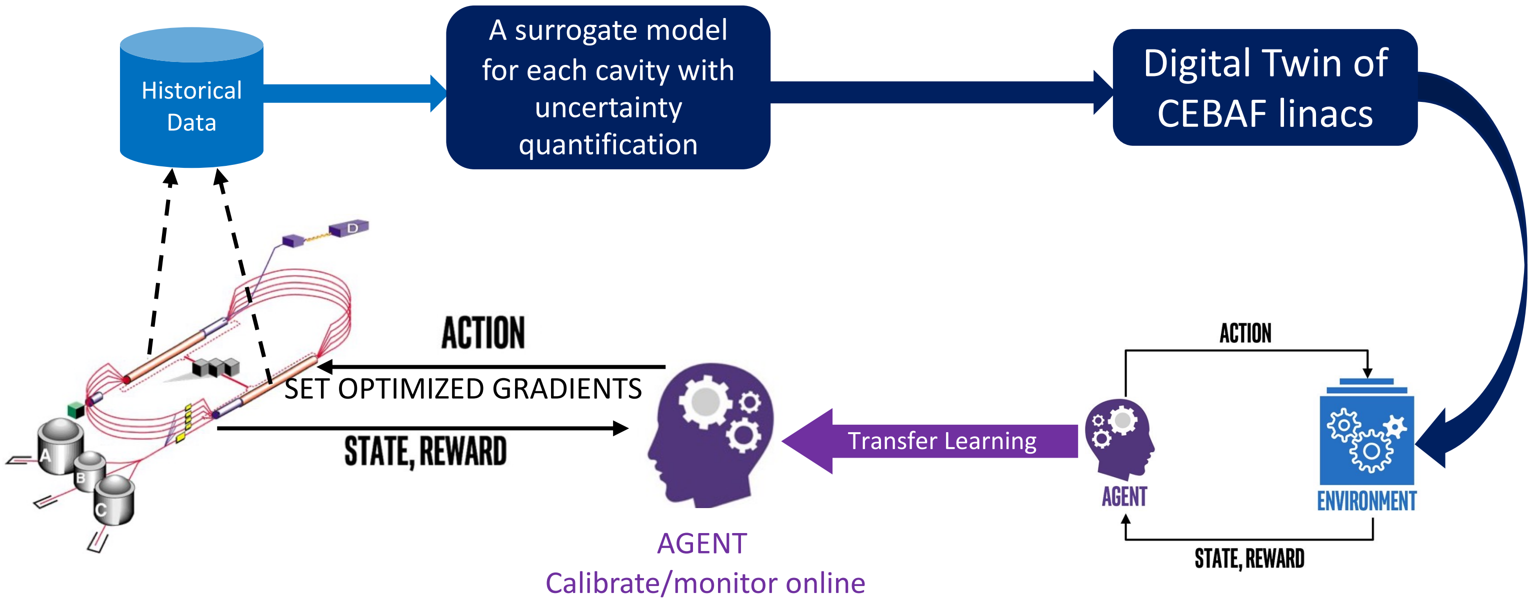


- ❑ Hydra was originally developed by Thomas Britton, a post-doc in Hall-D, with the help of off-the-shelf ML techniques (Inception-V3 model)
  - ❑ Enhanced/put in production by the EPSCI group with Data Science Dept.
- ❑ It has detected some of the issues that were missed by the detector experts.
- ❑ Future plan:
  - ❑ Enable diagnosis. (Instead of just saying "good" vs "bad", it would be able to point out possible issues in the detector)
  - ❑ Enhance the prediction capability and convert it into a toolkit for wider usage

Autonomous Control:
Multi-Objective Optimization of Cryogenic Heat Load and Trip Rates

Problem definition:
Optimize heat load and trip rate in CEBAF by setting optimal gradients in the SRF cavities
Proposed solution:
Explore the use of modern techniques such as reinforcement learning to provide optimization solution

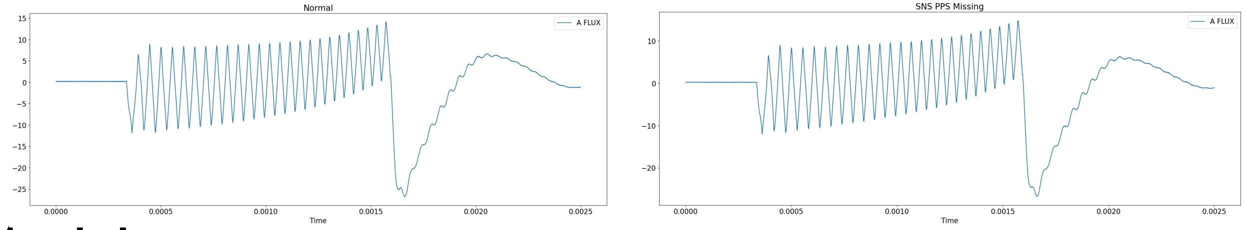
- Generic toolkit to use at other SRF based DOE facilities with similar challenges
- Making toolkit applicable to other multi-objective optimization problems in other domains
- Develop a digital twin of CEBAF LINACS that can further be utilized for future projects



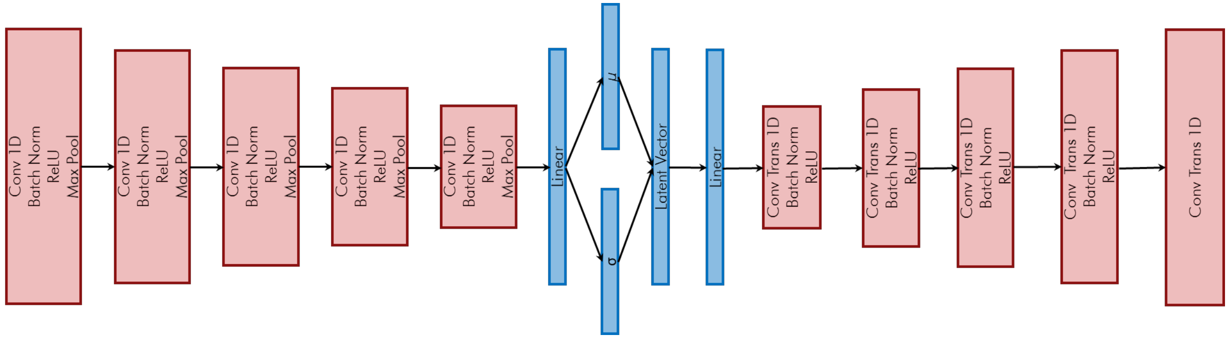
HVCM Anomaly Detection at ORNL

- **Problem Definition**
  - Investigate the possibility of predicting anomalies in the high voltage converter modules (HVCM) using ML techniques
- **Approach**
  - Autoencoder based anomaly detection scheme. Develop a 1D CNN variational autoencoder (VAE) capable of re-constructing multiple signals from HVCM modules
- **Preliminary Results**
  - With limited number of examples from an HVCM module, the VAE can identify approximately 60% of anomalies at 10% (acceptable) false positive rate

Data

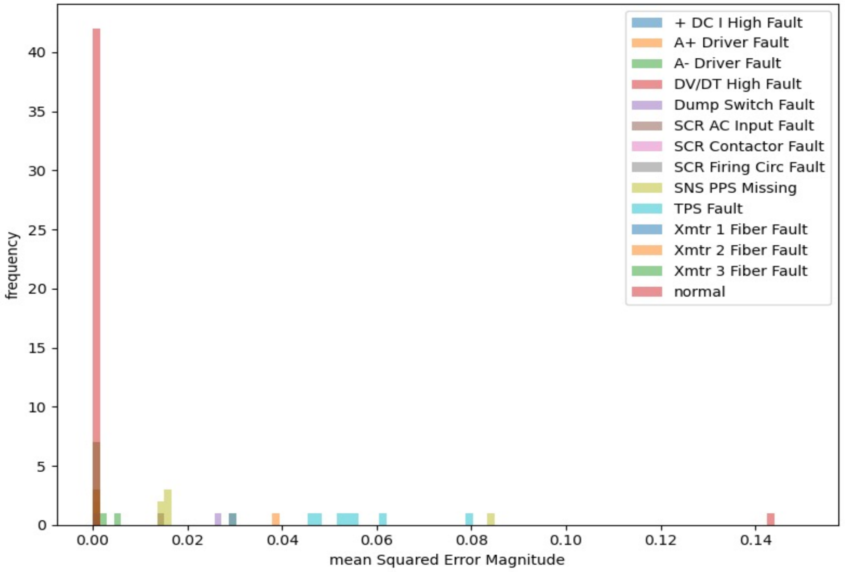


Model

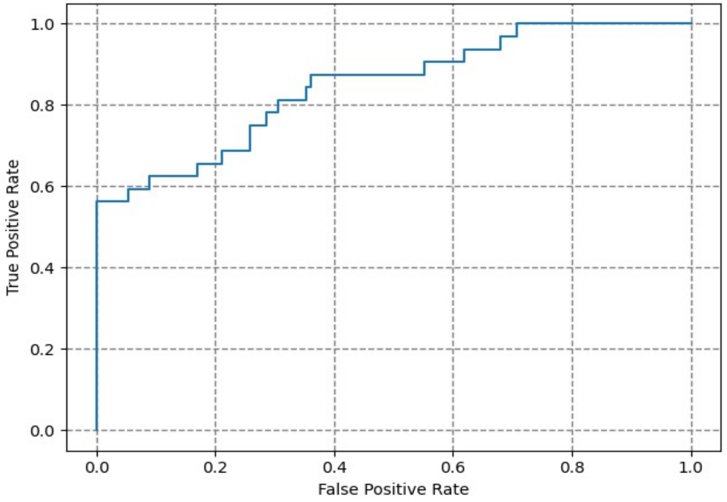


Results

Reconstruction error for normal and anomalous pulses



ROC curve: True positive rate versus false positive rate for anomaly detection



Jefferson Lab

Establish collaborations with others

- Regional Universities
- National Laboratories

Establish collaborations with regional universities

- Old Dominion University
  - They submitted a proposal for the CST bridge position proposal call
  - Proposal was accepted and we finalized the MOU
    - Faculty position on data science for accelerator physics is open
  - Started a new data science project with the ODU Virginia Modeling, Analysis, and Simulation Center
  - Currently establishing a research associate professor position within ODU VMASC
- William & Mary
  - They submitted a proposal for the CST bridge position proposal call
  - Proposal was accepted and we finalized the MOU
    - Faculty position is open and they will start reviewing applicants early calendar year
  - Currently establishing two affiliate faculty positions
- University of Virginia:
  - Submitted a capstone proposal which was accepted
  - Students are working on anomaly detection for accelerators

Establish collaborations with regional universities

- New project titled “Hampton Roads Digital Twins”
- Collaboration between Jefferson Lab and Old Dominion University (ODU) designed to leverage separate and complementary scientific expertise to find an innovative approach to address the health inequities of the Hampton Roads region.
- Benefits/Impacts:
  – General: Data science could provide new techniques to address public health issues which would be a paradigm shift in how data is used for public health research. Additionally, this effort will establish a vision and direction for a partnership between JLab and ODU/Virginia Modeling, Analysis, and Simulation Center (VMASC)/Hampton Roads Biomedical Research Consortium (HRBRC).
  – To ODU: This project will establish a partnership with JLab to provide a data science program in public health. JLab will provide expertise in data science methods used to address targeted public health research efforts.
  – To Jefferson Lab: This project will establish a partnership with ODU/VMASC/HRBRC focused on public health and the new NIH strategic plan. This project will provide new opportunities to expand the data science department portfolio.

Establish new collaborations with national laboratories, etc.

- High Energy Physics (HEP) program:
  - “Real-time Artificial Intelligence for Accelerator Control”
    - Partner: JLab and FNAL
    - Led the offline surrogate model and reinforcement learning effort
    - Advising on updates to the surrogate model and reinforcement learning in preparation for online deployment
    - Two published papers and one conference talk
- Advanced Scientific Computing Research (ASCR) program:
  - “Data-Driven Decision Control for Complex Systems”
    - Partners: JLab, ORNL, PNNL, Arizona State University, University of California
    - Leading the Reinforcement Learning Thrust
- Basic Energy Science (BES) program:
  - “Machine Learning for Improving Accelerator and Target Performance”
    - Partners: JLab and ORNL
    - Leading data science and machine learning efforts for specific application
    - One published paper, one submitted, and two conference talks
- Small Business Technology Transfer (STTR) program:
  - “Precise Control of Particle Orbit in Accelerator using Machine Learning Technology”
    - Partners: JLab, Raytum Photonics, and University of Washington
    - In review

Example: Machine Learning Grant from BES

Collaboration between SNS, ORNL, and Jefferson Lab.

1) Develop ML capabilities for time series analysis and prediction:
   a) Use unsupervised and semi-supervised learning with existing data to develop ML models with multiple timescales to predict failure scenarios.
   b) Develop ML-based anomaly detection capabilities for accelerator operations.
   c) Develop Uncertainty Quantification (UQ) capabilities for robust and reliable time-series analysis.
   d) Develop causal analysis capabilities relating failure predictions and anomalies to sensor measurements and system operating parameters.

2) Demonstrate Objective (1) on accelerator and target systems to monitor condition, detect anomalies, and predict failure, using information from system instrumentation and beam-based signatures.

3) Demonstrate ML-based surrogate modeling to optimize parameters and inform design choices.