



BERKELEY LAB

Bringing Science Solutions to the World



Machine Learning Optimization Upstream and Downstream of the Accelerator: The Cases of VENUS and GRETA

Funded under FY2023 Lab FOA

Heather Crawford, Chris Campbell, Mario Cromaz,
Marco Salathe, and Damon Todd

NP AI/ML PI Exchange Meeting
December 5, 2024

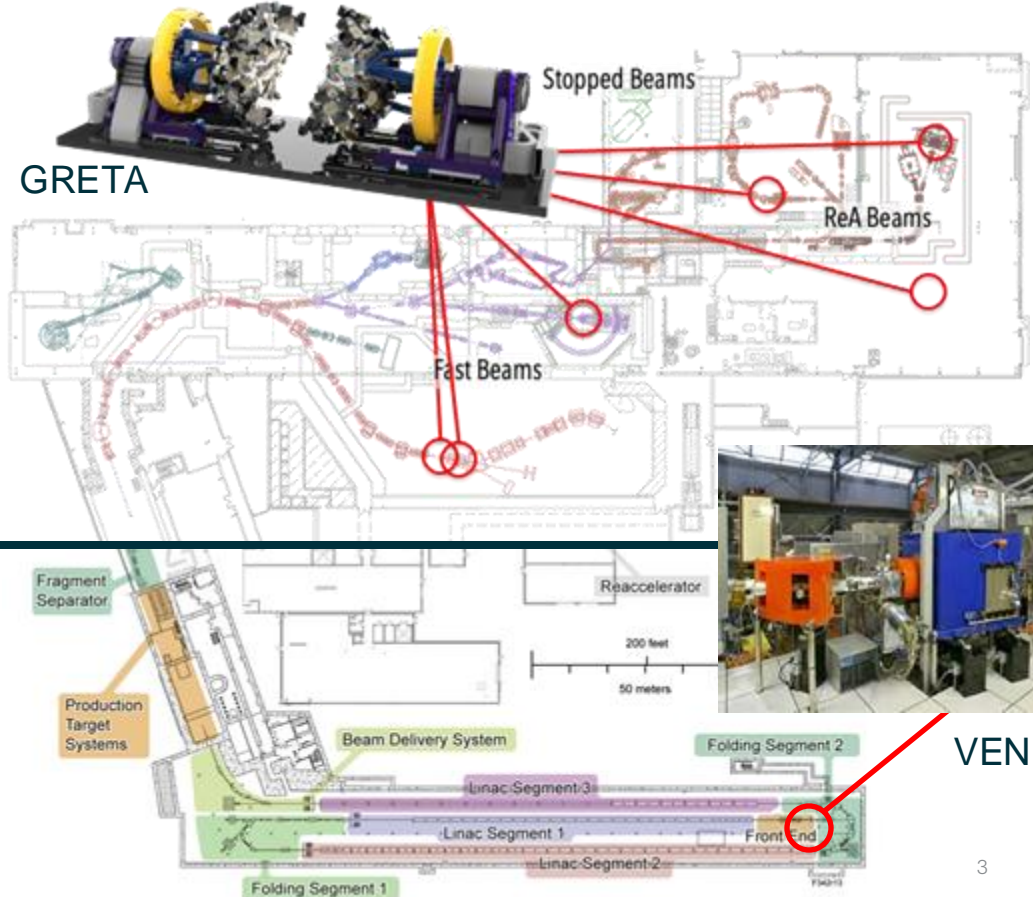
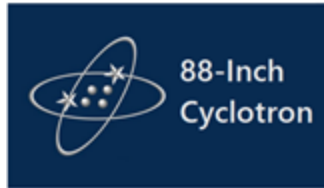
Optimizing the Front-End and Experimental End-Station



The effective operation of any accelerator facility is not limited to the accelerator itself – fully optimized operation is realized by optimizing all parts of an experiment, and reducing down-time along the entire facility chain.

We focus on the front end and the end-point of a facility - the VENUS ion source and GRETA experiment.

Applying Machine Learning to LBNL Systems to Impact 88" and FRIB Operations



Original FY21 ML/AI VENUS+GRETA Project

- The original effort was funded with a FY21 award, \$1M split evenly across two years
- First effort focused on:
 - Readyng VENUS for application of ML techniques – no data was recorded regularly, combination of EPICS and LabView interfaces needed to be made/re-written
 - Accumulating data from VENUS from human-driven tuning and source baking to provide a starting data set for ML applications
 - Automation of the frequent “baking” operation to reduce human time and improve efficiency
 - Initial demonstration of Bayesian optimized tuning within limited parameter space
 - For GRETA focused on automating the optimization of the electronics signal chains for resolution, and providing complete calibration of a crystal; required interfacing to GRETA EPICS systems and hardware and explored optimization approaches (traditional Nelder-Mead, GPR, ...)
- Ended grant period with \$360k of carryover (postdoc joined 10 months into the award period)

Budget (FY23 Award)

	FY23 (\$k)	FY24 (\$k)	Total (\$k)
Funds Allocated	228	870	1,098
Actual Costs to Date	588**	107	695

** \$360k in carryover from previous award

Research Team - Staff and Postdocs



Chris Campbell
Scientific Engineer
LE Program / GRETA



Mario Cromaz
*Applied Physicist Staff
Scientist*
LE Program / GRETA



Marco Salathe
*Applied Physicist
Research Scientist*
ANP Program



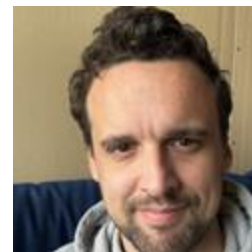
Damon Todd
*Principal Scientific
Engineering Associate*
88 Ops Program



Nico Abgrall
*Senior Scientific
Engineering Associate*
ANP Program



Yue Shi Lai
*Applied Physicist
Research Scientist*
ANP Program



Victor Watson
*Postdoctoral
Researcher*
ML Project

Research Team - Undergraduate Researchers



Ezra Apple
UCBerkeley
Electrical
Engineering/Computer
Science
Class of 2025



Julia Dreiling
University of Ohio
Data Analytics
Class of 2024
- now pursuing MSc at
University of St.
Andrews



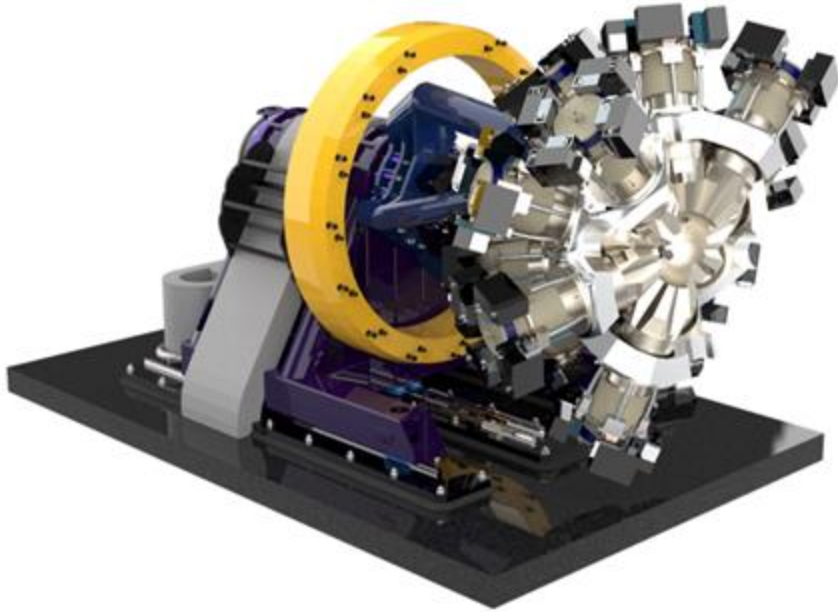
Alex Kireeff
Carnegie Mellon
University,
Electrical Computer
Engineering
Class of 2024



Arin Manohar
UCBerkeley
Physics, Computer
Science, Mathematics
Class of 2026

GRETA

Gamma-Ray Energy Tracking Array, GRETA



- U.S. implementation of a gamma-ray tracking array
- Complete 4π solid angle coverage of active high-purity germanium (HPGe), consisting of 120 individual detector crystals, each with 37 electrical signals
- Gamma-ray tracking and Compton suppression is enabled by signal decomposition algorithm which localized gamma-ray scatter events to within $\sim\text{mm}^3$ volumes

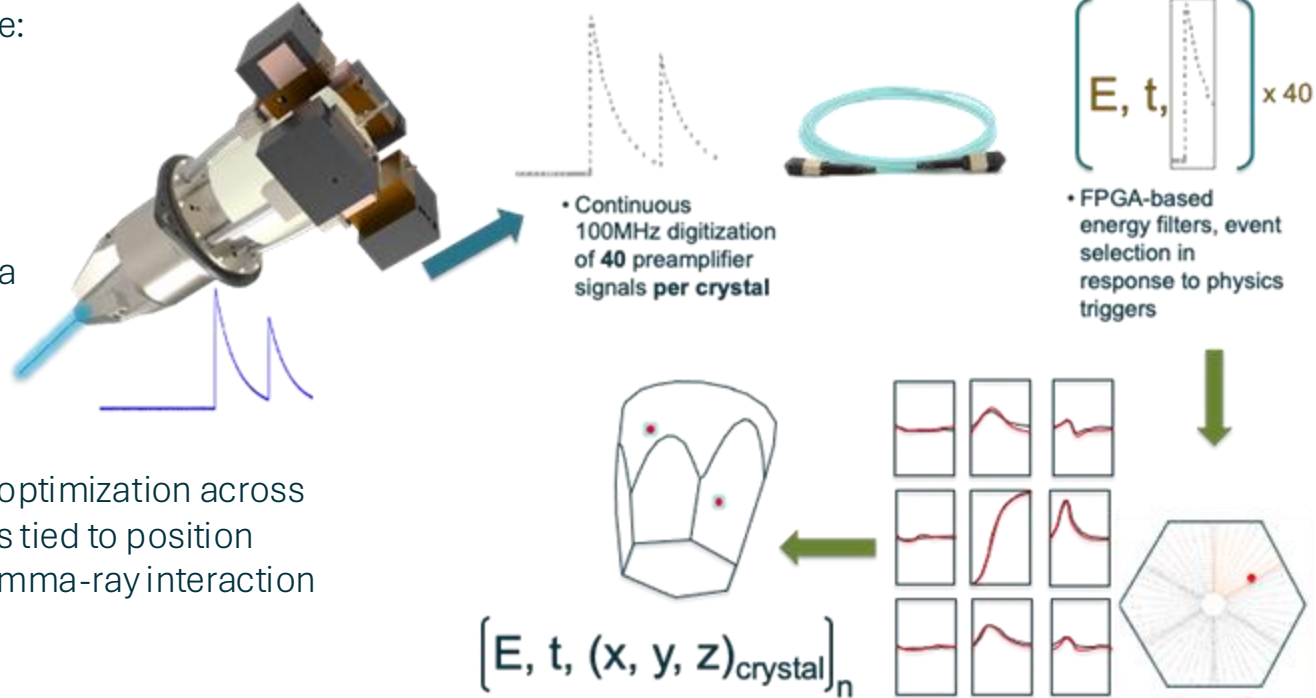
GRETA will be the world-leading gamma-ray spectrometer once delivered to FRIB in 2025, where it will be an experimental physics workhorse

GRETA Optimizations

Simple control parameters include:

- 4-6+ energy filter parameters per channel
- 2+ calibration parameters per channel

~ 30k knobs just for energy spectra



In addition to energy resolution optimization across the array, GRETA performance is tied to position resolution for reconstructing gamma-ray interaction points.

Position resolution depends on the fidelity of the calculated response of the HPGe crystals.

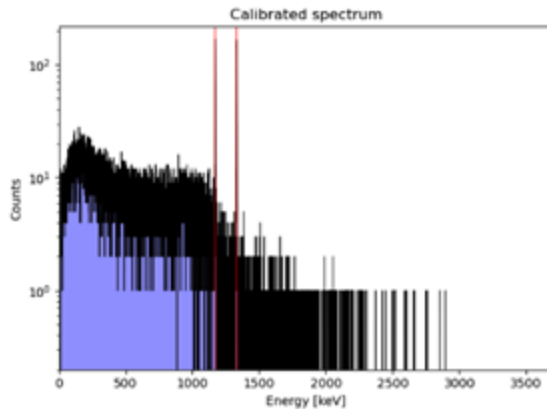
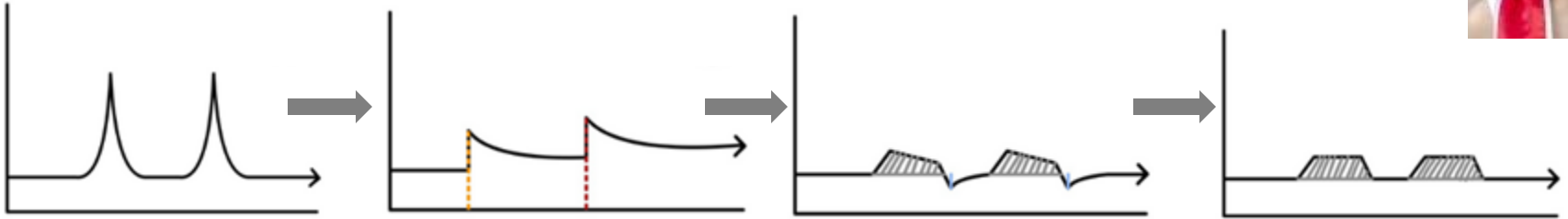
GRETA Goals and Status

- Automated optimization and calibration of all 4800 channels of electronics
- Improvement of the signal basis used for the process of signal decomposition, exploring improvement in the position resolution of interactions in GRETA by improving the calculated signals used in the fit through an ML-driven global optimization.
 - The automated optimization online is complete for 1 crystal and being extended to all 120 crystals (the full GRETA array)
 - Signal basis refinement is currently at the stage of refactoring and updating the software chain to enable an iterative optimization – e.g. superpulse fitting, for electronics response characterization, is now working in Python and electronics response function is being evaluated

GRETA Signal Chain Optimization Energy Resolution and Calibration

Automated Optimization of GRETA Signal Chain Resolution and Calibration

Work performed by Julia Dreiling (undergraduate student researcher)



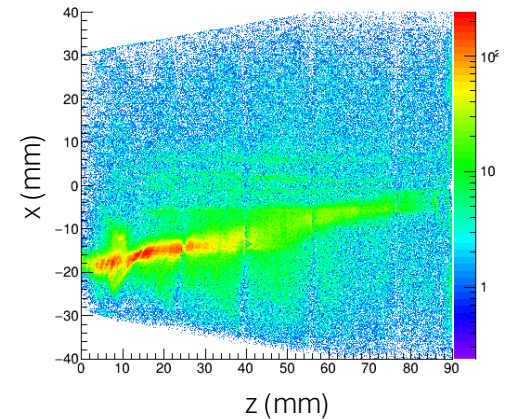
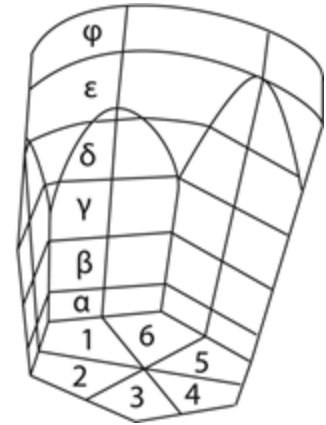
- Optimization of online data completed with interface to the GRETA signal filter board hardware for 1 crystal using Gaussian process regressor
- Optimization is now being expanded for simultaneous application to all 120 crystals
- LBNL-developed Becquerel package used to calibrate all signals (core + segments) at each step

GRETA Basis Creation Optimization

GRETA Basis Creation Overview

The GRETA basis production has two distinct steps:

1. Pristine basis calculation and signal generation
 - a. A calculation of the HPGe semiconductor is used to calculate the electric fields and weighting potentials within each crystal, and from this the shapes of signals on all crystal electrodes based on quantities e.g. **material impurity (profile), temperature, bias voltage, dead layers** (dozens of parameters)
2. Electronics response correction
 - a. The real data folds the innate crystal response with the response of the signal processing electronics – includes **shaping times, cross-talk (integral + differential), rise times** (includes several hundred parameters)



GRETAPulseGen: an integrative development framework

Non-concurrent, CPU-bound C/C++ libraries currently used to cover each step of the signal basis generation:

- (1) Calculate electric fields and weighting potentials;
- (2) Generate grid and raw basis signals;
- (3) Produce simulated superpulse (SP) using the raw signal basis;
- (4) Fit simulated SP against experimental SP to determine cross-talk parameters;
- (5) Generate cross-talk corrected basis.

Pros: well tested pipeline

Cons: cumbersome, time consuming, not suitable for iterative studies
and ML applications

GRETAPulseGen: an integrative development framework

Migration toward integrative framework:

(1) GRETAPulseGen Engine

Provides backend and HW acceleration. Features

include:

- C++ / CMake build generator
- Memory management
- Parallelization from CUDA support
- Thread pool for CPU bound parallelizable applications (WIP)

(1) GRETAPulseGen Application

Development API for user applications

GRETAPulseGen: an integrative development framework

Current and future activities:

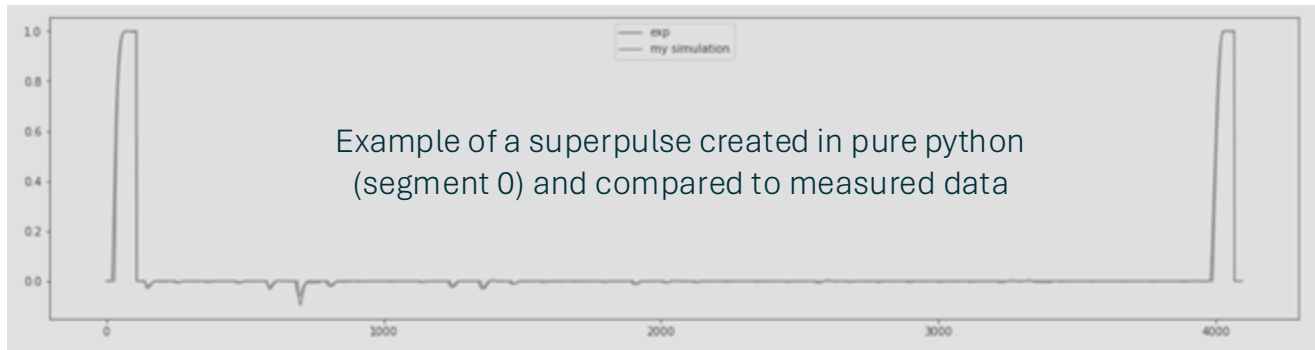
- (1) Migrate current pipeline to GRETAPulseGen framework
 - Leverage CUDA kernels for Poisson solver
 - Parallelization of CPU bound tasks
- (1) Application development
 - “Classic” optimization of crystal parameters and electronics response
 - Development of ML-based optimization techniques

Superpulse Fitting and Detector Response

Work performed by Arin
Manohar
(undergraduate student
researcher)









- Goal is to streamline the procedure from Geant4 and basis points data to fitting the superpulse, to automate procedure and improve the current fit model
 - Convert Geant4 and Basis files to python readable input (done)
 - Reconstruct superpulse generation pipeline in python (mostly done, cross checking)



- Convert fitting routine to python (currently ongoing)
- Develop updated and simplified physics-driven electronics response function (ongoing)
 - Parameter importance investigation

GRETA Project Goals and Status

WBS	Milestone	Description	
2	GRETA Staffing Requirements Met	Advertise and hire an undergraduate student and postdoc to work on GRETA scope.	
2.2.1	Develop Python utilities for signal basis representation	Develop a library of Python tools for signal basis representation and visualization, including pulses at individual interaction points.	
2.2.1	Define electronics response function	Define a parameterization for the electronics response function for basis generation.	
2.2.1	Explore sensitivity of superpulse types to parameters	Characterize the sensitivity of different measurement types (superpulse types) to parameters in the electronics response function.	
2.2.2	Evaluate hyperparameter search tools for use in GRETA case	Explore the available hyperparameter search tools that we can consider for use in optimizing the electronics response and crystal parameters.	
2.1	Demonstrate (up to) 120 crystal simultaneous optimization	Extend the optimization and calibration code to tackle 120 crystals at once.	
2.2.1	Implement updated signal basis generation tool chain	Implement and configure complete signal basis generation tool chain with updated utilities for automated basis generation.	
2.2.2	Develop parameterization for crystal description	Define a parameterization of the crystal properties such as impurity profile etc.	
2.3	Evaluate opportunities for direct ML inference of basis signals	Look into techniques that can generate a signal basis without the crystal properties calculation based on data only.	
2.2.2	Complete final code base for open-source distribution.	Assuming success for previous steps, clean up code and package for open-source distribution following LBNL policies.	

VENUS

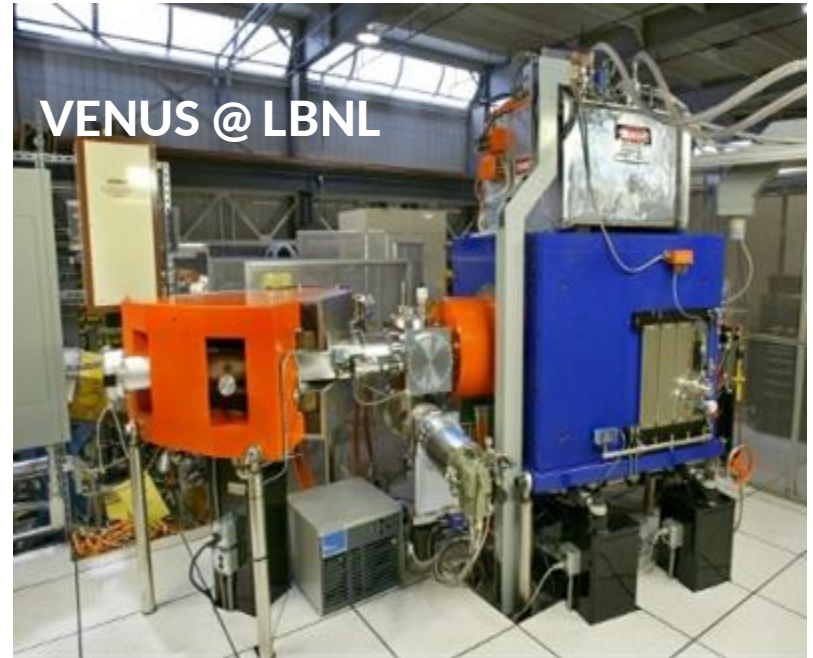
The Electron Cyclotron Resonance (ECR) Ion Source VENUS

VENUS:

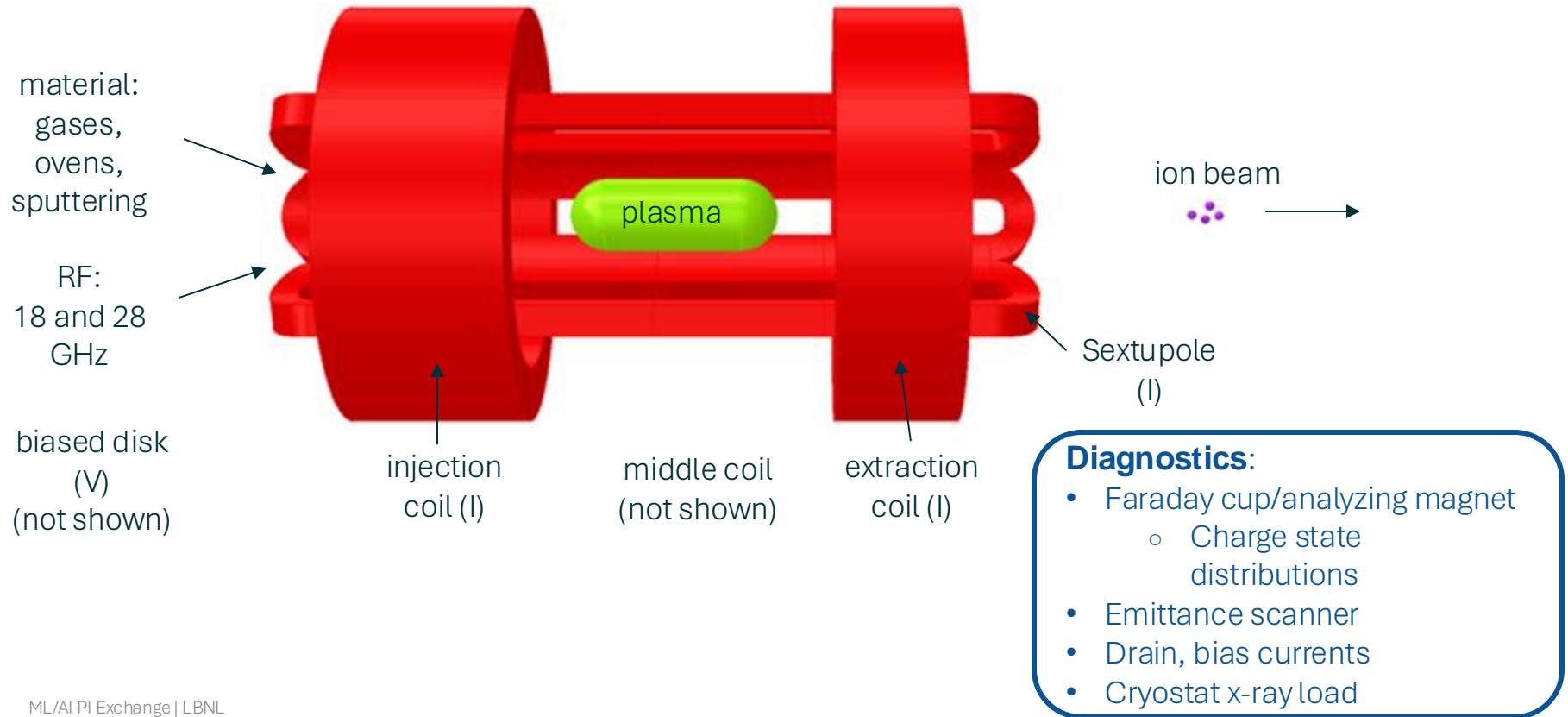
- World's first fully-superconducting ECR ion source designed for 28 GHz operation
- One of the world's two highest-performing ECR ion sources
- Injector for LBNL's 88" Cyclotron
- Prototype ECR ion source for FRIB, where a near-identical copy has been installed

Example beams:

- $> 4.7 \text{ mA O}^{6+}$, $> 20 \text{ mA He}^+$ from source
- $> 2 \text{ p}\mu\text{A}$, $5 \text{ MeV/u } ^{48}\text{Ca}^{11+}$ and $> 1.4 \text{ p}\mu\text{A } ^{48}\text{Ti}^{11+}$ from cyclotron for superheavy element research
- Beams up to U, Xe and Au (v. high charge states, demanding tunes)



VENUS Primary Control and Diagnostic Parameters

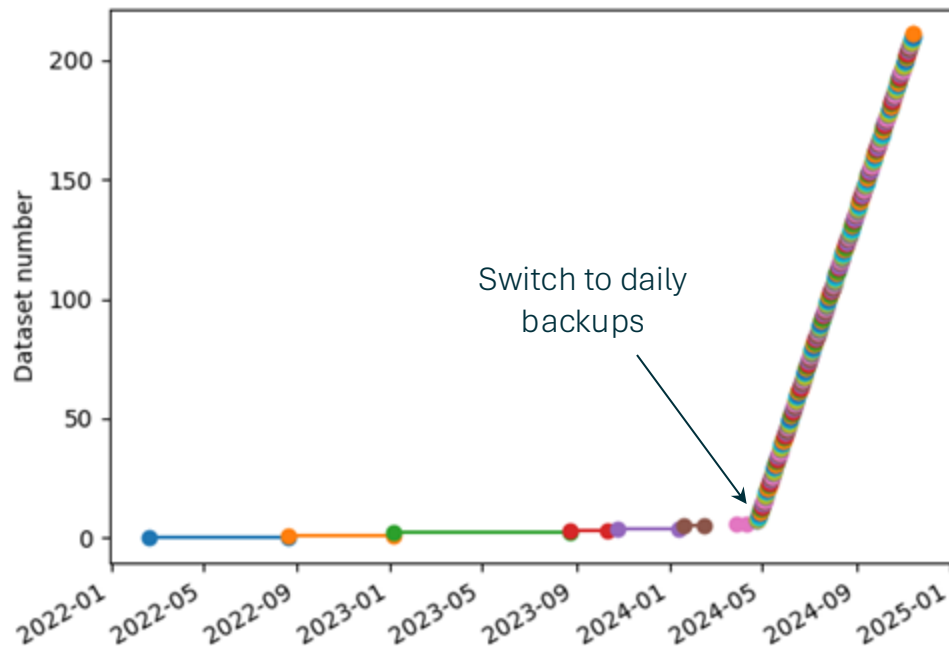


VENUS Project Goals and Status

- Enhance VENUS capabilities by adding two hardware systems enabling fast measurement of beam emittance and charge state distributions, and a non-interruptive measure of beam current using a flying-wire system – enable full use of all VENUS operations data for algorithm training, and to provide key information to develop a more fully optimized cost function for VENUS
- Extend optimization of the VENUS system to the full parameter space available, utilizing tools of reinforcement learning
- Explore methods for running VENUS in a continuously optimized and stable configuration
 - Fast charge state distributions are now enabled, as well as faster beam current measurements; emittance and flying wire are in design stages
 - Offline reinforcement learning and random forest show promise for improved optimization
 - Algorithm for stable and continuously optimized operation is ready to test

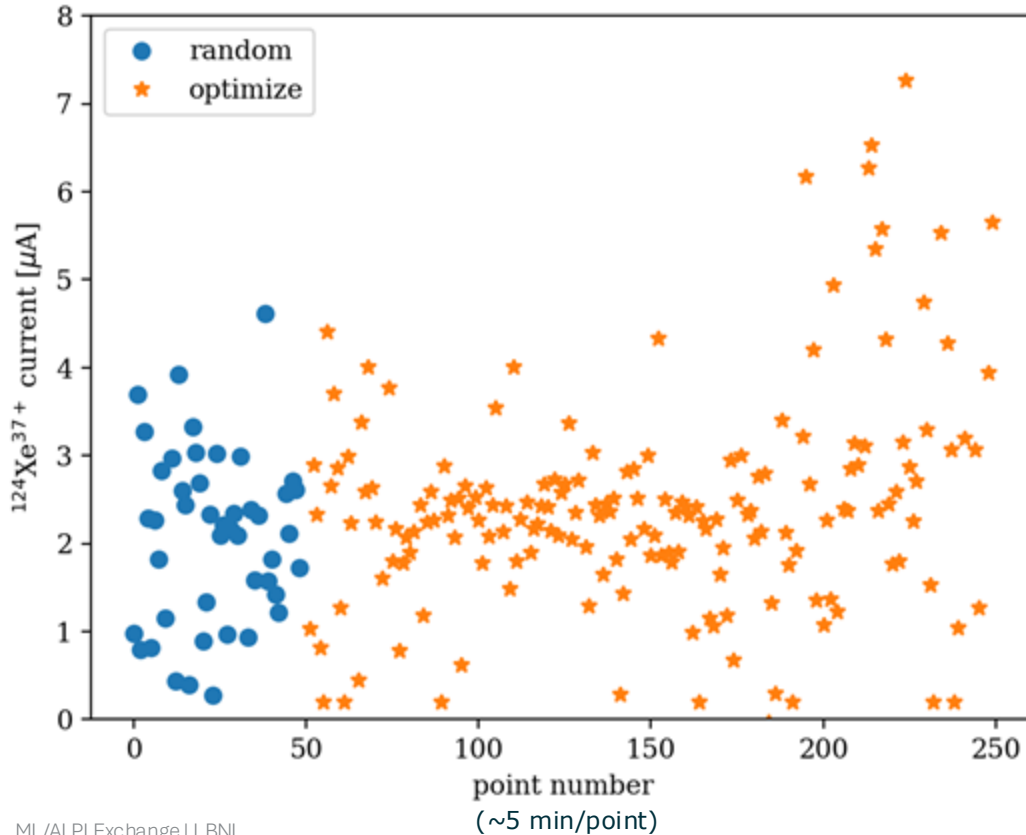
VENUS database

- Almost continuous recording of VENUS data to SQLite database
 - More than 880 days of data
 - More than a dozen scans of parameter space (2 – 4 parameters)
- Outlook
 - Add full charge state distribution data to database, whenever one is run
 - Possible switch to time series database (Prometheus)



VENUS Bayesian Optimization and Improved Charge State Distributions

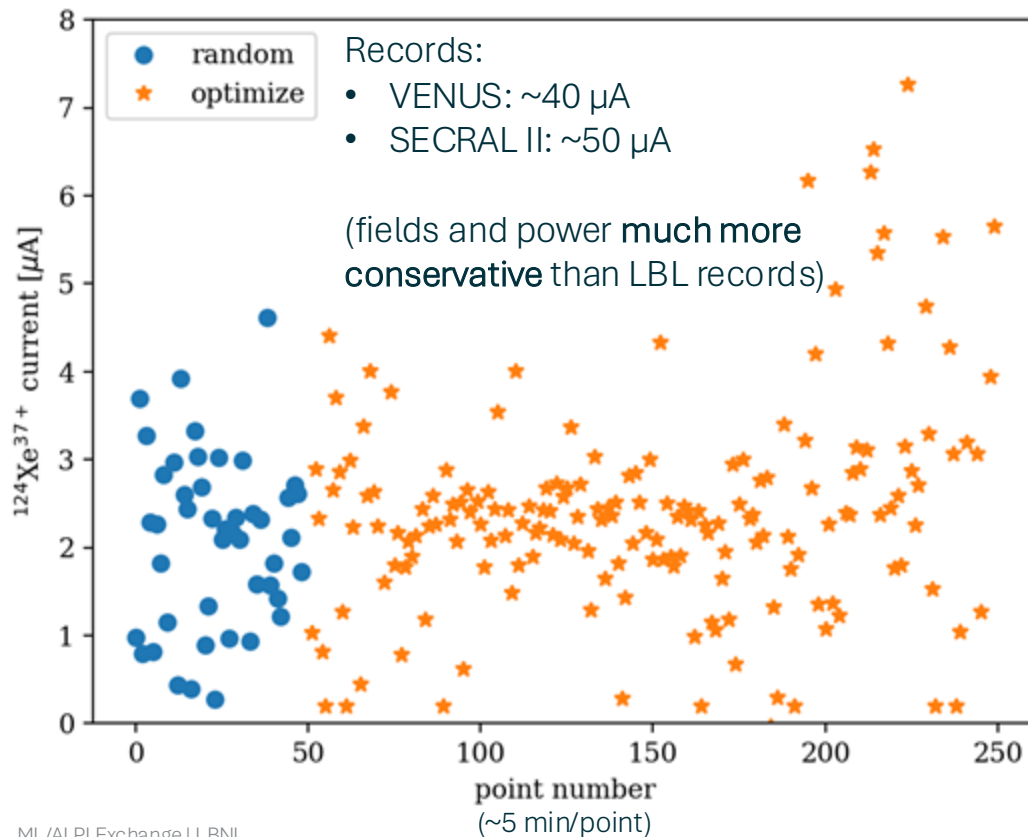
Machine Learning: Full Bayesian Optimization of $^{124}\text{Xe}^{37+}$



Parameter	Min	Max
Bias voltage [V]	40	105
Oxygen valve	11.6	12.5
Xenon valve	8.0	13.0
Inj coil [A]	185.6	186.0
Ext coil [A]	136.6	136.8
Mid coil [A]	152.0	152.3
Sext coil [A]	430.3	430.5
18 GHz [kW]	1.4	1.8
28 GHz [kW]	5.2	6.0

- VENUS completely under computer control
- Computer “knows” nothing about VENUS

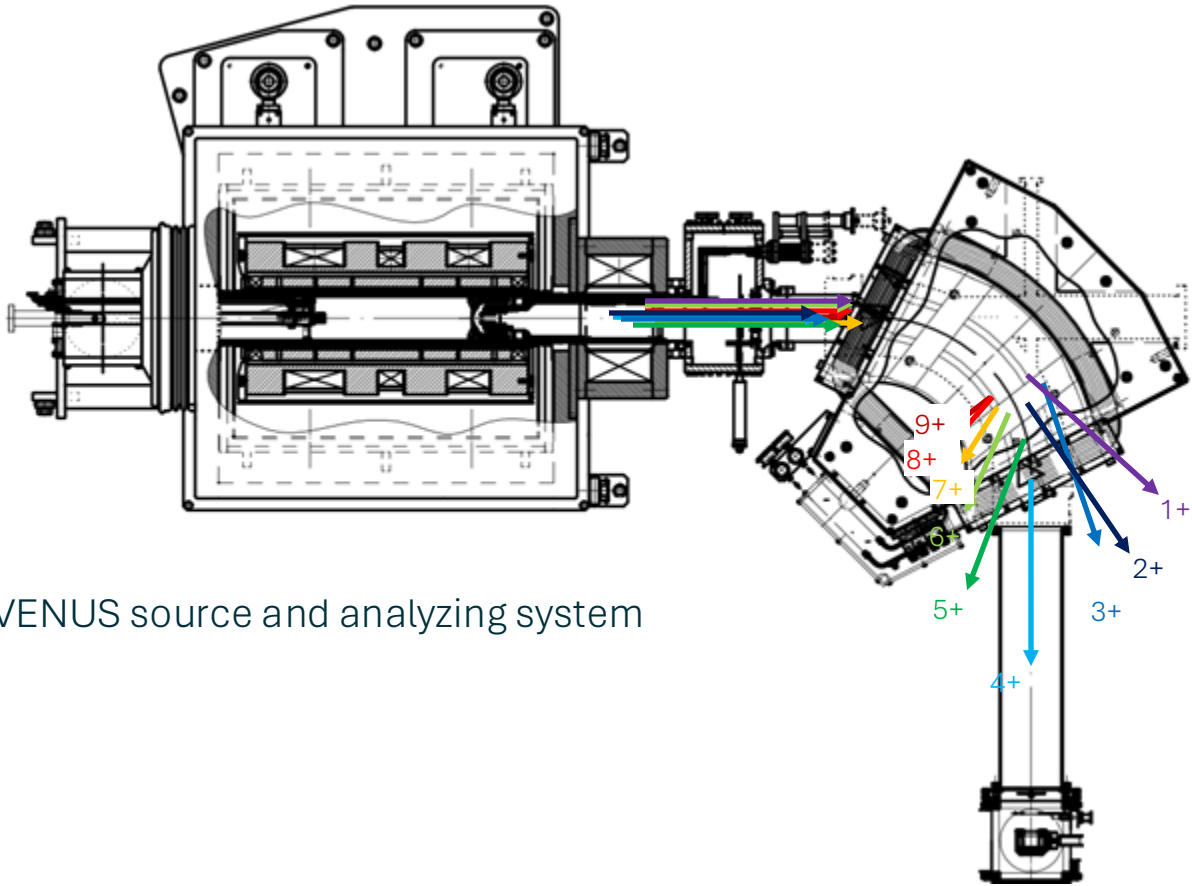
Machine Learning: Full Bayesian Optimization of $^{124}\text{Xe}^{37+}$



Parameter	Min	Max
Bias voltage [V]	40	105
Oxygen valve	11.6	12.5
Xenon valve	8.0	13.0
Inj coil [A]	185.6	186.0
Ext coil [A]	136.6	136.8
Mid coil [A]	152.0	152.3
Sext coil [A]	430.3	430.5
18 GHz [kW]	1.4	1.8
28 GHz [kW]	5.2	6.0

- VENUS completely under computer control
- Computer “knows” nothing about VENUS

What the experiment looks like



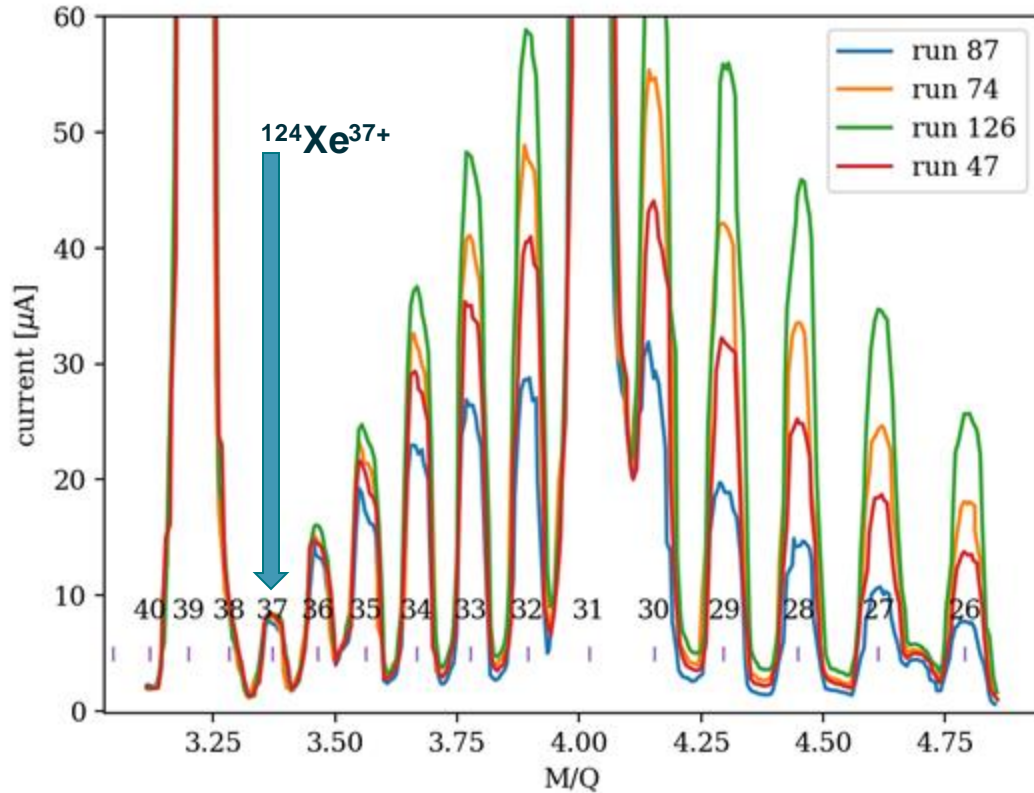
VENUS source and analyzing system

Many ways to same result

Later search maximized $^{124}\text{Xe}^{37+}$ to ~ 10 uA

- However, CSD shows many ways to get there

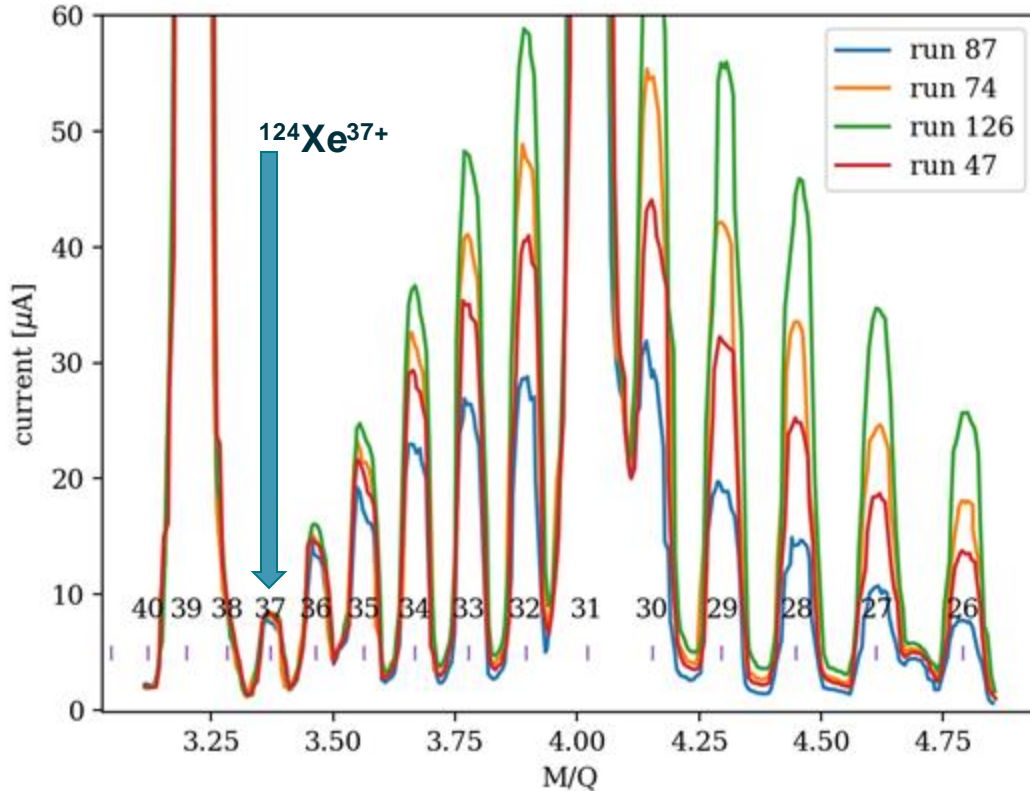
Many ways to same result



Later search maximized $^{124}\text{Xe}^{37+}$ to $\sim 10 \mu\text{A}$

- However, CSD shows many ways to get there

Many ways to same result



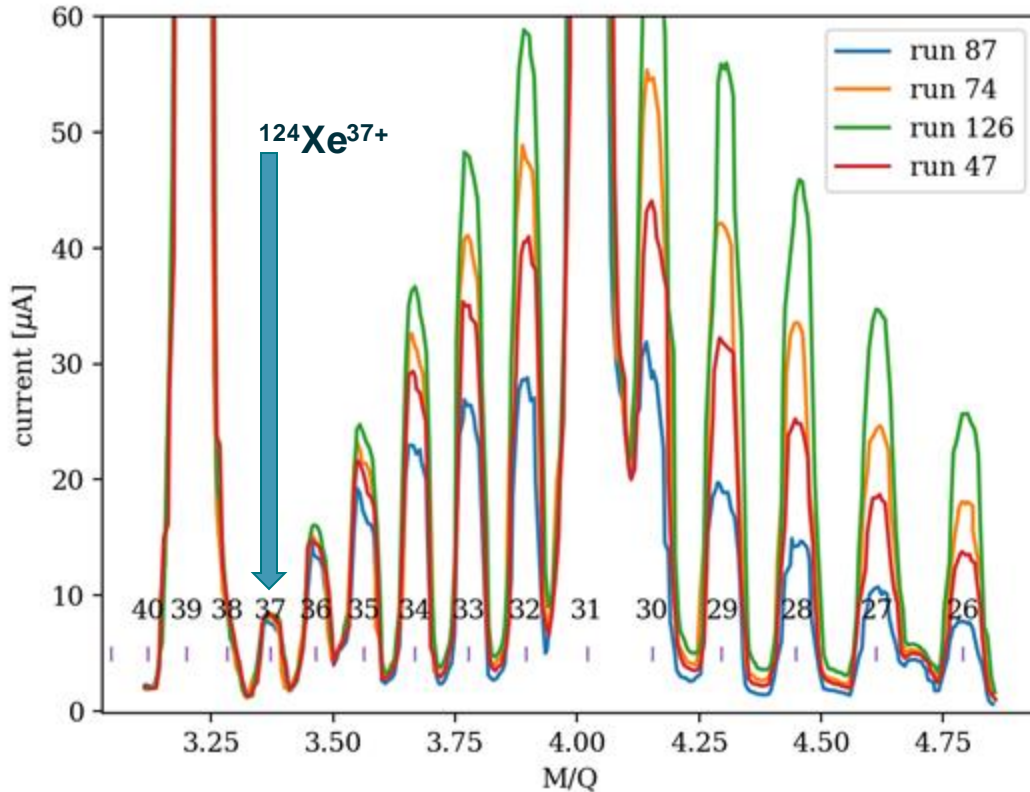
Later search maximized $^{124}\text{Xe}^{37+}$ to $\sim 10 \mu\text{A}$

- However, CSD shows many ways to get there

Takeaways:

- optimizing charge state's current without CSD knowledge is restricting
- CSD is slow: $\sim 2\text{-}3$ minutes each
- Even beam statistics are slow: ~ 3 Hz

Many ways to same result



Later search maximized $^{124}\text{Xe}^{37+}$ to $\sim 10 \mu\text{A}$

- However, CSD shows many ways to get there

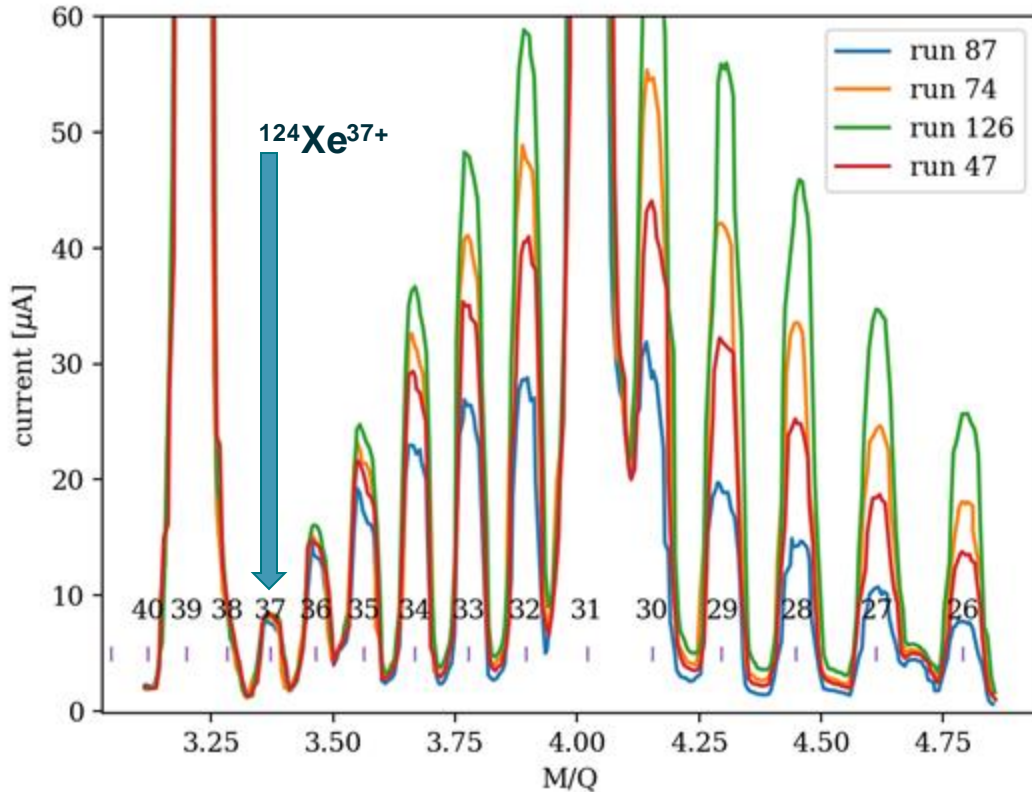
Takeaways:

- optimizing charge state's current without CSD knowledge is restricting
- CSD is slow: $\sim 2\text{-}3$ minutes each
- Even beam statistics are slow: ~ 3 Hz

Goals:

- Let computer try for record beam
- Speed up data gathering

Many ways to same result



Later search maximized $^{124}\text{Xe}^{37+}$ to ~10 uA

- However, CSD shows many ways to get there

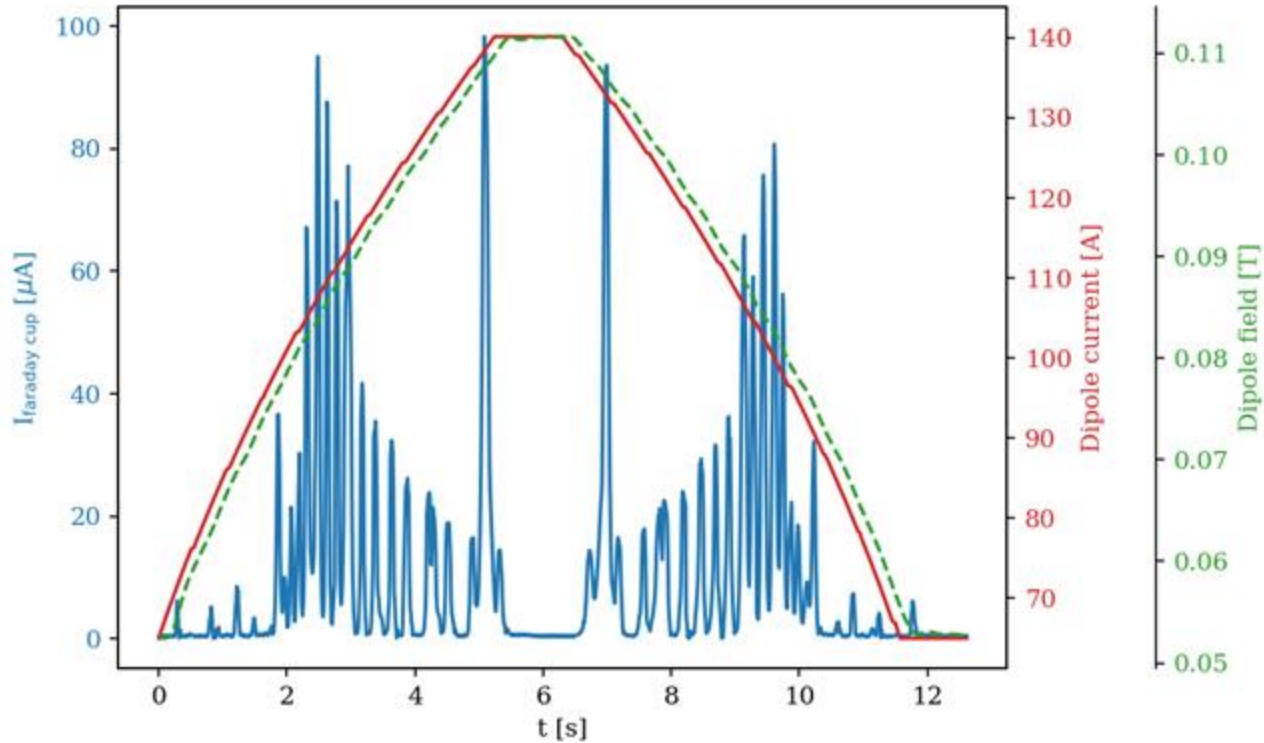
Takeaways:

- optimizing charge state's current without CSD knowledge is restricting
- CSD is slow: ~2-3 minutes each
- Even beam statistics are slow: ~3 Hz

Goals:

- Let computer try for record beam ✗
- Speed up data gathering ✓

Faster beam measurements



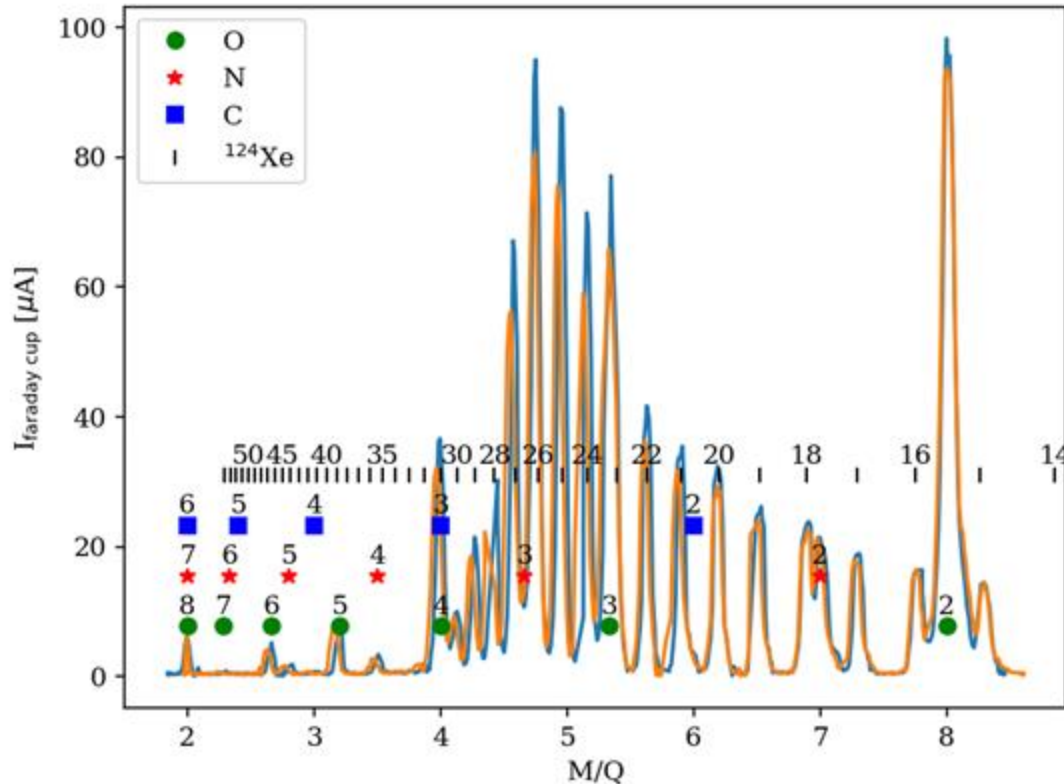
At 100 Hz:

- Set dipole current
- Read dipole's hall probe
- Read beam current

Faster CSDs

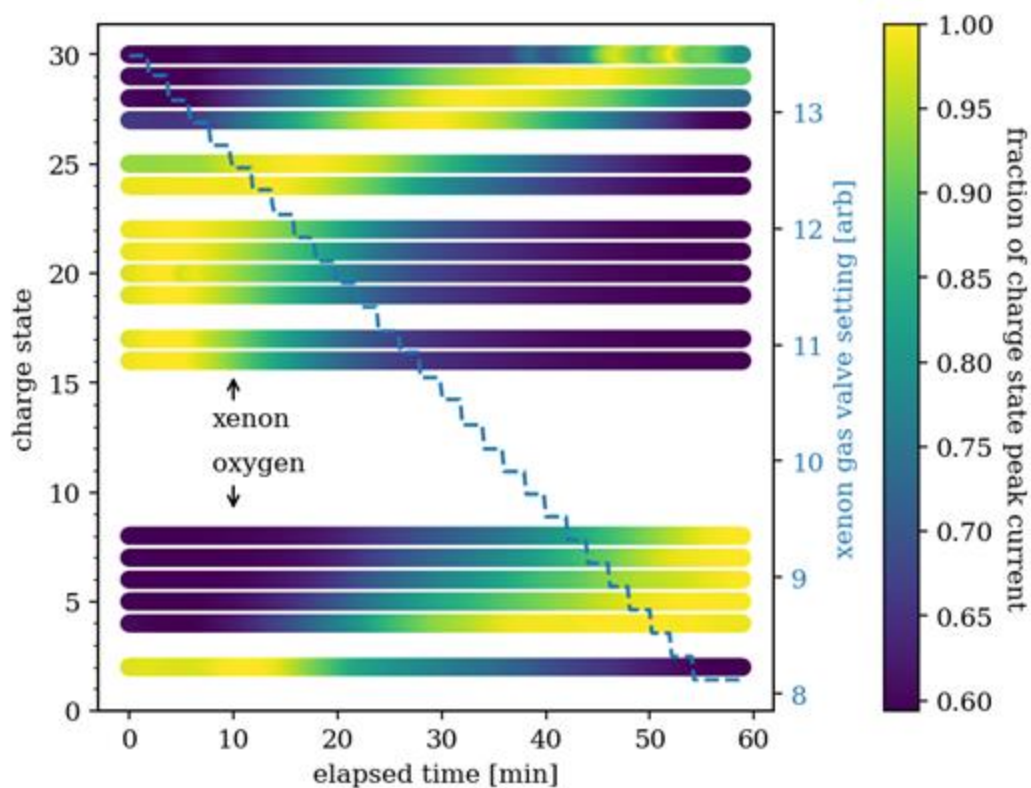
- agree well with slower ones

Unwrapping fast CSDs



- Reduction in measured current on return, especially for high charge states
- Reduction recovers by the next CSD sweep
- Use only “increasing current” CSDs for comparison

Visualizing Dynamic CSD Information



Note: this visualization problem is ours only. Machine learning can deal with multiple dimensional arrays, etc.

Frequent CSD are now a reality for VENUS and are being incorporated now into our existing Bayesian optimizers and other optimization approaches.

The next target is to similarly optimize emittance scanning to incorporate this information.

Modeling VENUS Across Data Sets with Random Forests

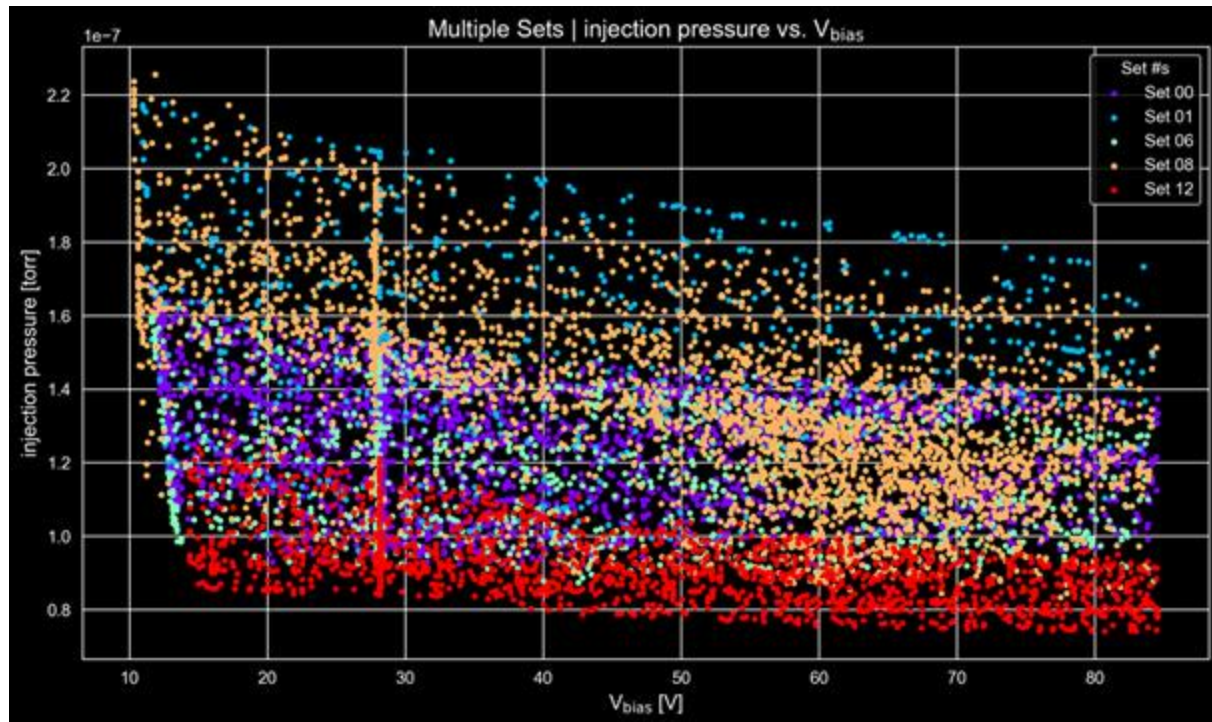
Variation Across VENUS Data Sets

Work performed by Ezra
Apple (undergraduate
student researcher)



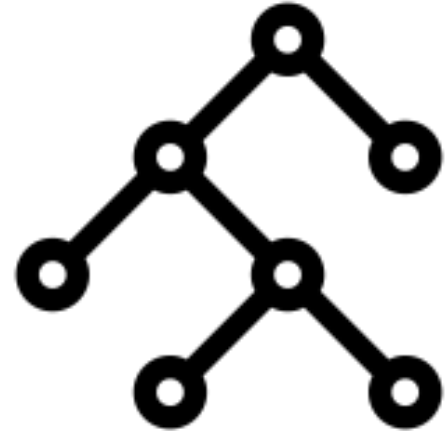
After pre-selection of data for stability and “settled-ness”, the variation in data sets all optimized for O^{7+} becomes apparent – there are similarities in features, but high variability.

This variation presents challenges for training models robustly.

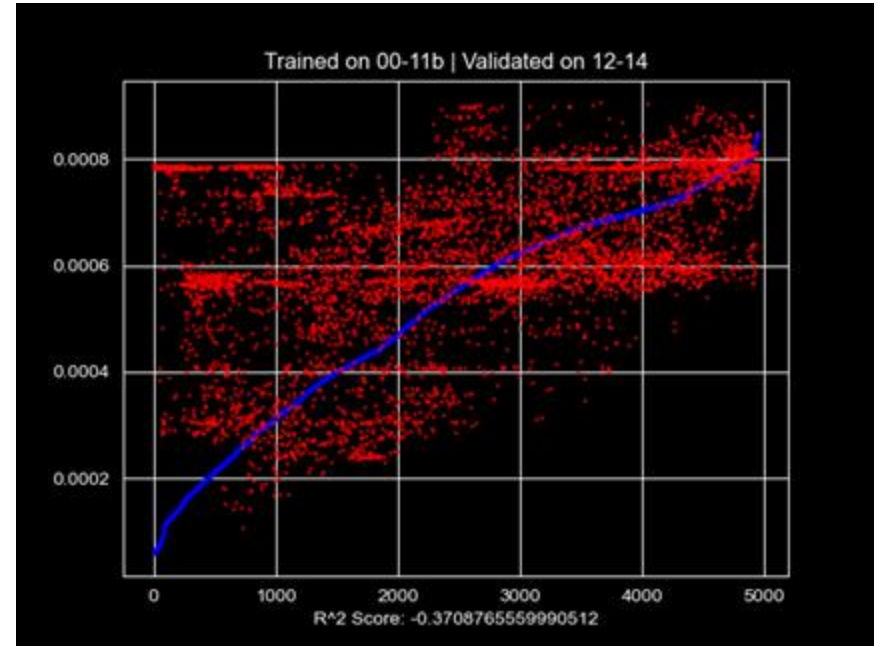
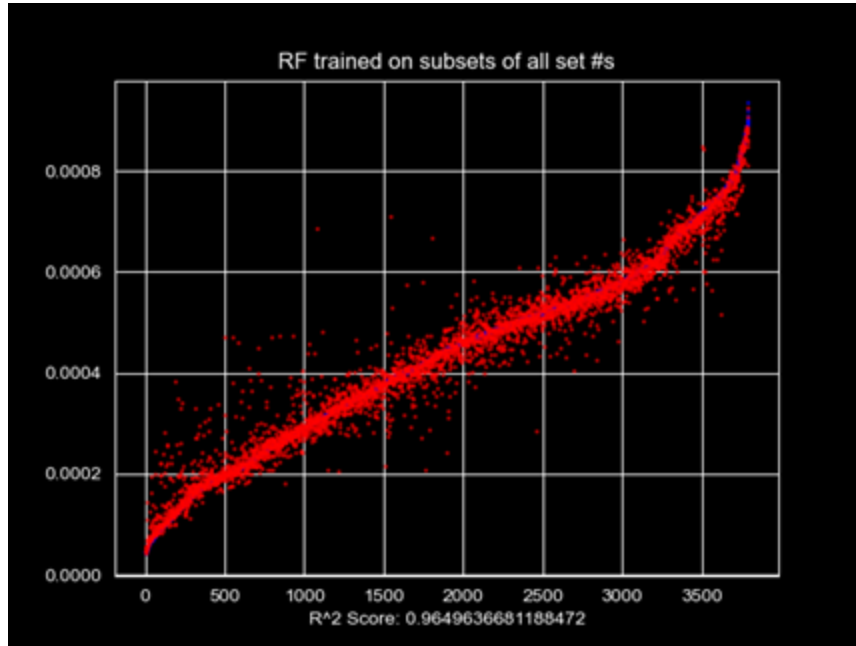


Modeling with Random Forests

- Random forests were chosen to be explored for their robustness and expressiveness
- They are also quick to train, and the inherent randomness helps prevent over-fitting
- Applied random forest for regression – uses an ensemble method using multiple decision trees, and makes predictions using the average output of all trees
- Very effective for modeling complex non-linear relationships with the possibility for feature importance analysis

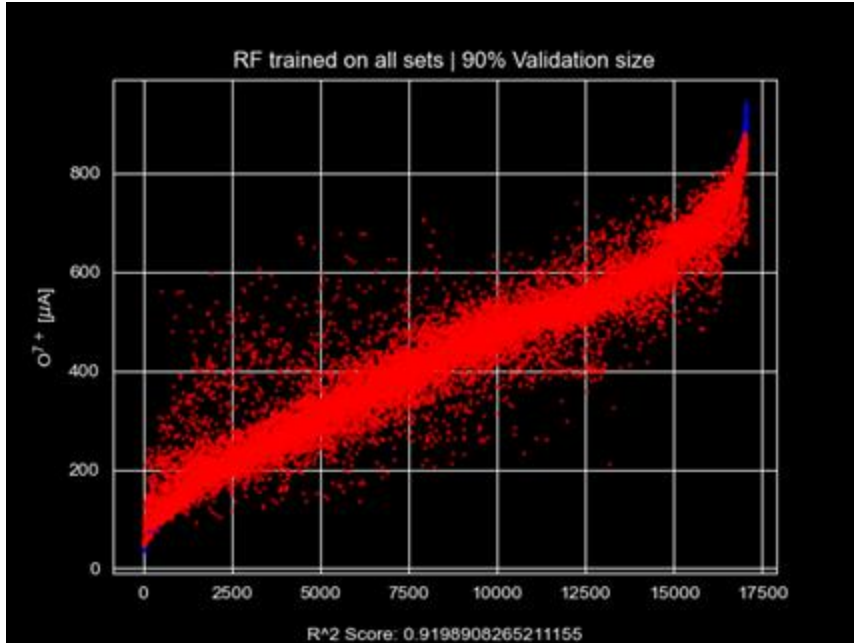


Predictive Performance with Random Forest Modeling



The predictive performance depends strongly on training data – dramatically worse performance when predicting on unseen data sets.

Predictive Performance with Minimal Training Data



Even when setting aside 90% of the data to validate with, the model performs well if it has trained on at least part of the set.

We are exploring how much training data is required to achieve predictive power – given the speed of Random Forest training, it is possible to update the predictions in real time and couple to our current methods to improve optimization of the source.

Offline Reinforcement Learning for VENUS

Why Offline Reinforcement Learning?

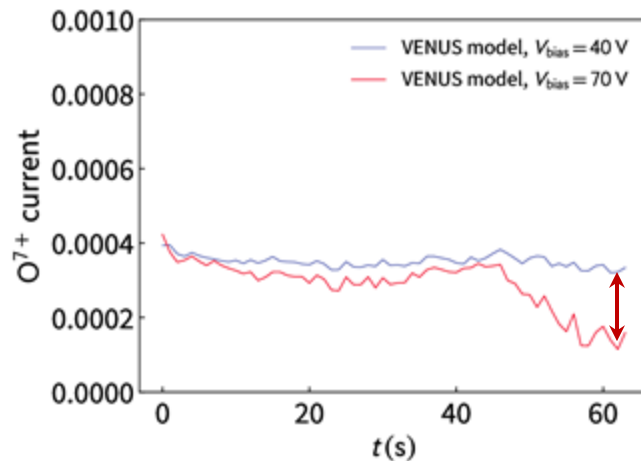
Compared to traditional methods:

- Fully models expected VENUS feedback, including long time horizon effects (stability)
- Ignores long stretches of suboptimal control/inactivity (unlike behavior cloning)
- Can be potentially adapted to online reinforcement learning

Compared to online reinforcement learning:

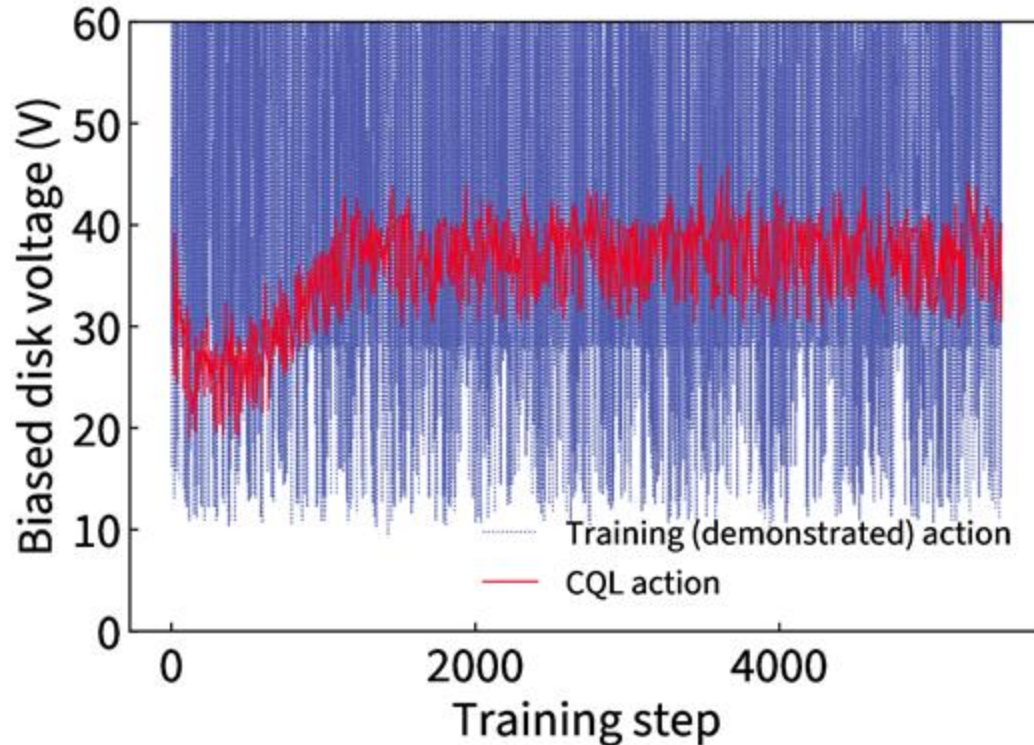
- Can be applied directly to vast (years) of collected human control

Tools used: Google JAX + DeepMind ACME's implementation of conservative Q-learning (CQL), customized for regularization



Delayed loss of current in simulated surrogate model

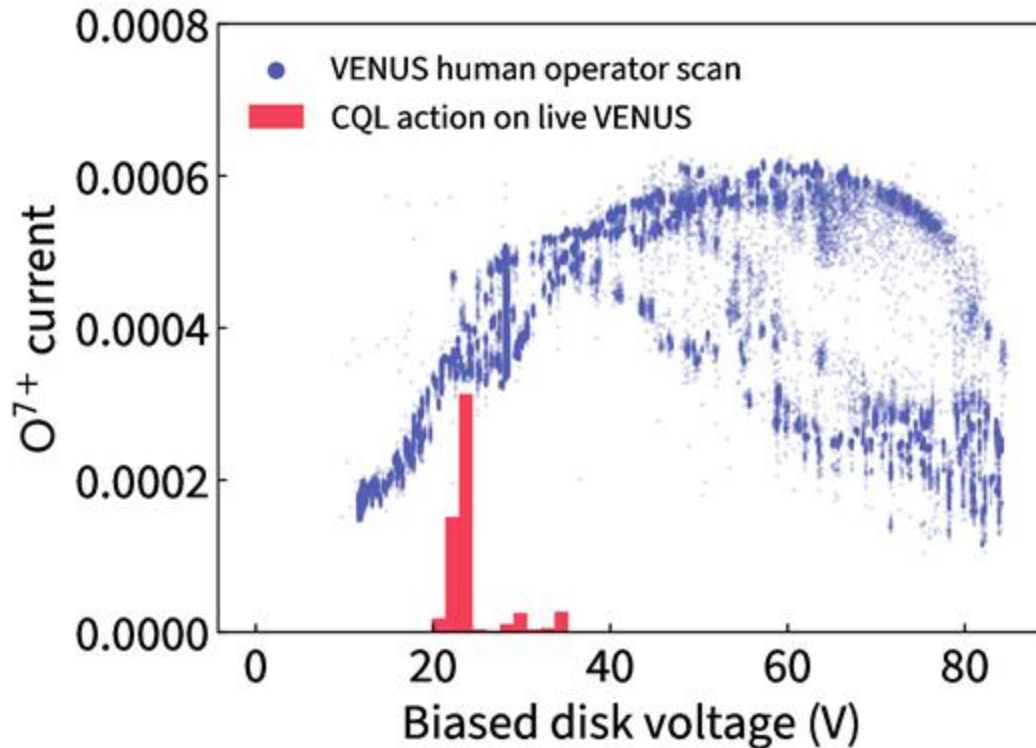
Offline reinforcement learning



Offline reinforcement learning training

- The **first known offline reinforcement learning (RL)** for ion source
- Utilizes recorded human control, but actively learns strategies with high beam current reward (e.g. ignores time ranges with low activity)
- Bypasses the high cost of performing sample-inefficient online RL on live VENUS
- Converges to a reasonable biased disk voltage with random samples from a human operator scan

Offline reinforcement learning



Older live VENUS running

- RL trained with long time horizon, including possible instability at high biased disk voltage
- RL control generally chooses to stay at a safe reward region

Next Steps

- A surrogate for VENUS has now been developed using recurrent NN – hopefully will allow larger scale off-line RL for VENUS
- Future tests with on-line running are planned over the next several months

Lai *et al.*, in preparation.

VENUS Stable and Optimized Operation

Dynamics of the VENUS System

- One of the challenges of the VENUS source is that the optimal conditions evolve with time, so the maximum beam current moves in parameter space as do source instabilities
- The state of the system is described by a couple of unknown functions $f(\cdot)$ and $g(\cdot)$ depending on a set of parameters θ and varying along time t
- While the system runs, the search for optimum is constant and must obey two constraints:
 - Objective function $f(\theta, t) > \alpha_0$ at every running point
 - System stays under control and that the control function $g(\theta, t) < \alpha_c$ at every running point
- We developed a Bayesian optimizer for non invariant dynamic systems (NIDSO) creates an adaptive statistical model of $f(\cdot)$ and $g(\cdot)$ while the system is running and optimizes it while respecting the constraints

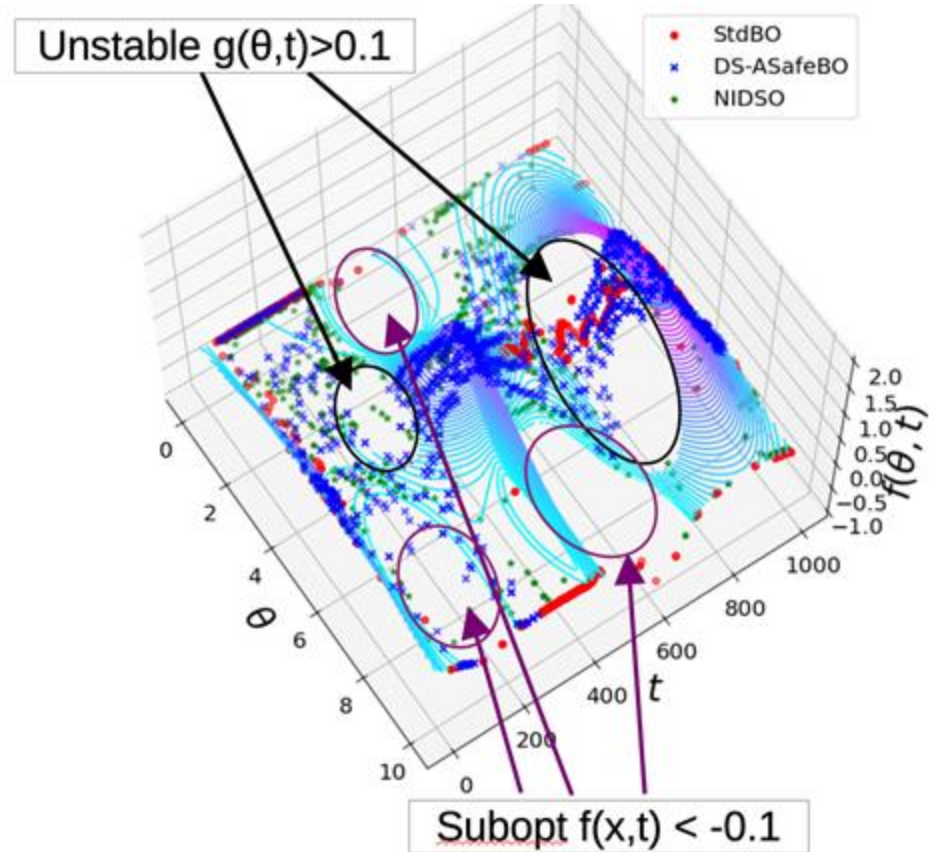
Toy Model for Stable VENUS Operations

- We can mock-up the VENUS dynamics by defining a surface with maximum(s) in $f(\theta, t)$ and negative-going regions of instability, $g(\theta, t)$

- In order to capture dynamics, an isotropic kernel is used which depends on the distance between evaluated points

$$k_M(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{l} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}r}{l} \right)$$

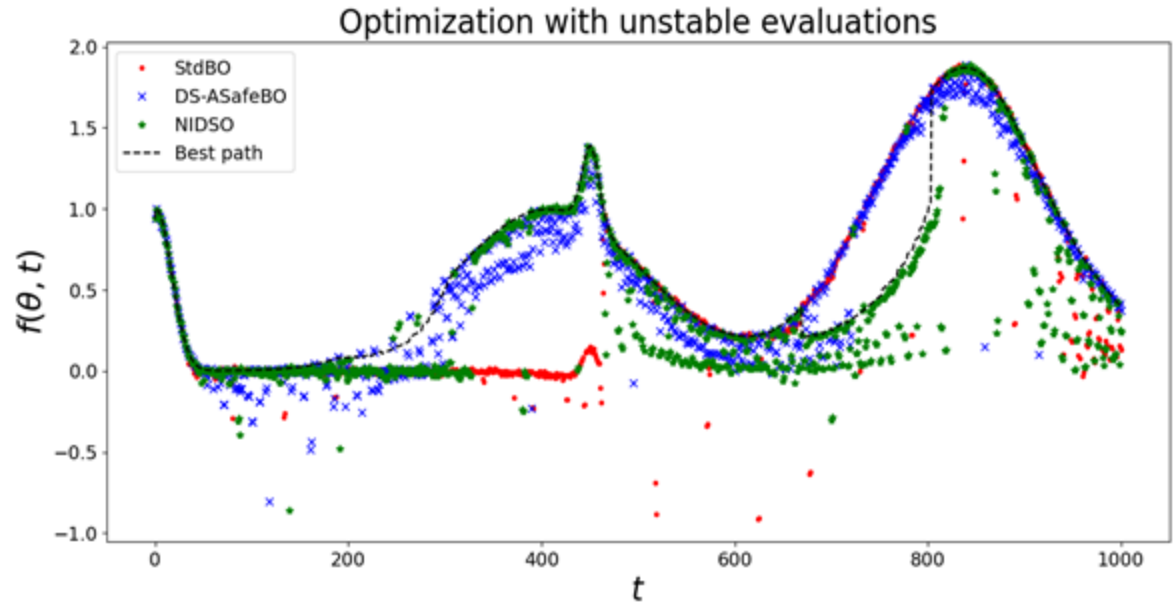
- As the Gaussian process update with each new measurement, the expected improvement acquisition function provides the next point



Results

Comparison of three methods:

- ‘standard’ Bayesian optimization with no instability awareness (**StdBO**);
- the newly-developed method of **NIDSO**;
- adaptation of **ASafeBO** (Han et al.) which provides a mechanism to stay out of high-risk areas but had to be adapted to force exploration of the parameter space when the maximum drifts

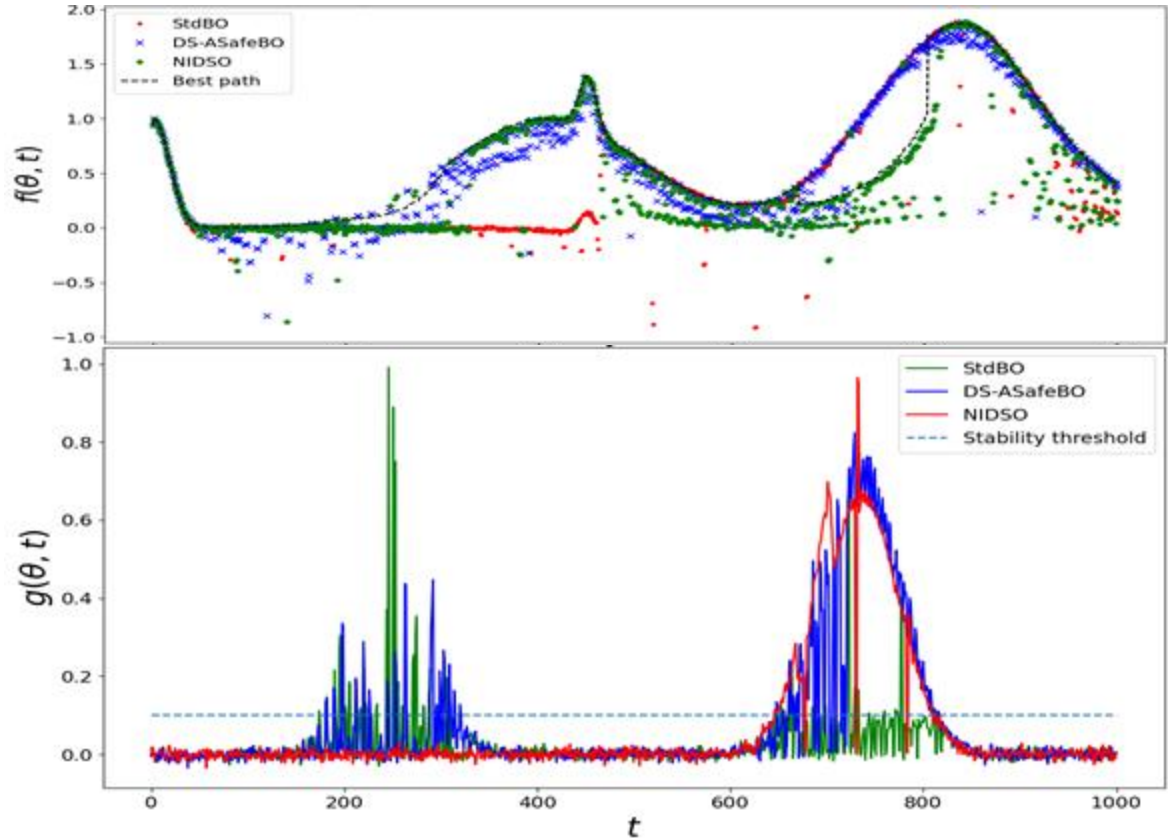


Han, G., Jeong, J., and Kim, J.-H. (2023). Adaptive bayesian optimization for fast exploration under safety constraints. IEEE Access, 11:42949–42969.

Results

Comparison of three methods:

- ‘standard’ Bayesian optimization with no instability awareness (StdBO);
- the newly-developed method of NIDSO;
- adaptation of ASafeBO (Han et al.) which provides a mechanism to stay out of high-risk areas but had to be adapted to force exploration of the parameter space when the maximum drifts

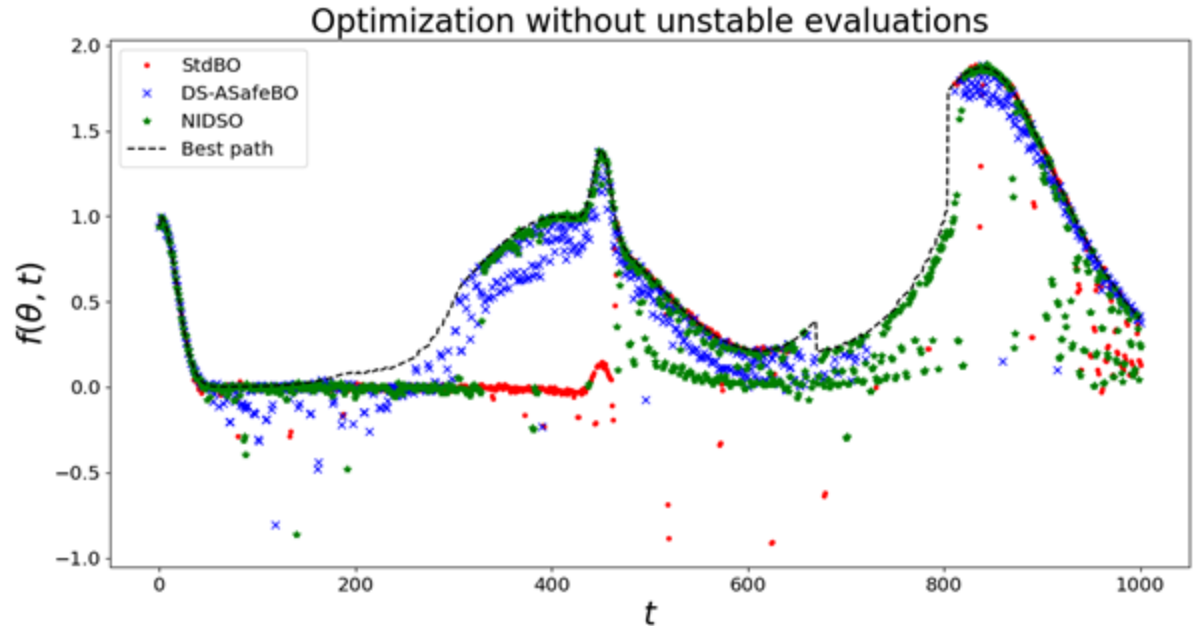


Han, G., Jeong, J., and Kim, J.-H. (2023). Adaptive bayesian optimization for fast exploration under safety constraints. IEEE Access, 11:42949–42969.

Results

Comparison of three methods:

- ‘standard’ Bayesian optimization with no instability awareness (**StdBO**);
- the newly-developed method of **NIDSO**;
- adaptation of **ASafeBO** (Han et al.) which provides a mechanism to stay out of high-risk areas but had to be adapted to force exploration of the parameter space when the maximum drifts



Han, G., Jeong, J., and Kim, J.-H. (2023). Adaptive bayesian optimization for fast exploration under safety constraints. IEEE Access, 11:42949–42969.

Stable and Optimized Performance

- The newly-developed NIDSO is able to optimize the system close to in a difficult case where $g(\cdot)$ is close to the threshold
- Both the adapted ASafeBO and our NIDSO give rather good result in terms of following the maximum

Method	Standard Bayesian Optimization	Adapted ASafeBO	NIDSO
Performance failure $f(\theta, t)$ below threshold	2.2%	3.5%	0.9%
Instability failure $g(\theta, t)$ above threshold	16%	17.5%	7.1%

- The new method will be implemented on VENUS – the challenge is to introduce test instabilities in a predictable way to evaluate performance

VENUS Project Goals and Status

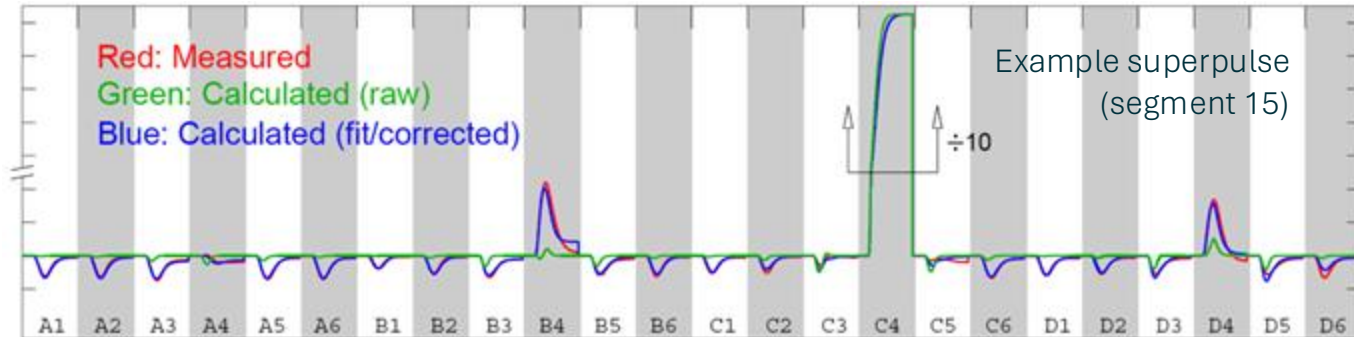
WBS	Milestone	Description
1.1	Implement a monitoring code to predict/warn of instabilities	Based on training with recorded data, implement an online stability monitoring program for VENUS.
1.3	Implement ML-driven baking for VENUS	Implement an ML-based program for baking VENUS and benchmark performance against human and automated script.
1.2	Incorporate emittance scanning into VENUS optimization	Following upgrade of emittance scanner hardware incorporated into optimization as a separate parameter to optimize.



Thank you

Automatic superpulse generation

Superpulses are a compressed/filtered representation (shape: 37x36x50) that allows to compare simulation to measurements => Input to a complex fitting routine of >500 parameters



- Measured superpulse:
 - Data are collected with a radioactive source, signals are filtered and summed
- Simulated superpulse:
 - Basis generation: simulate expected signals on a grid (5000 points) in the germanium crystal
 - Geant4 simulation: interaction pattern of gamma rays from a radioactive source in the crystal
 - Modeling of expected pulse shape for each of the simulated events
 - Filtering and summing signals according to procedure done with measured superpulse