Physics-informed ML for polarization increase in injectors (FOA)

Georg Hoffstaetter de Torquat
Collider-Accelerator Department, BNL and Cornell University
Georg.Hoffstaetter@cornell.edu

A collaboration of BNL, Cornell, TJNAF, SLAC, RPI

@BrookhavenLab

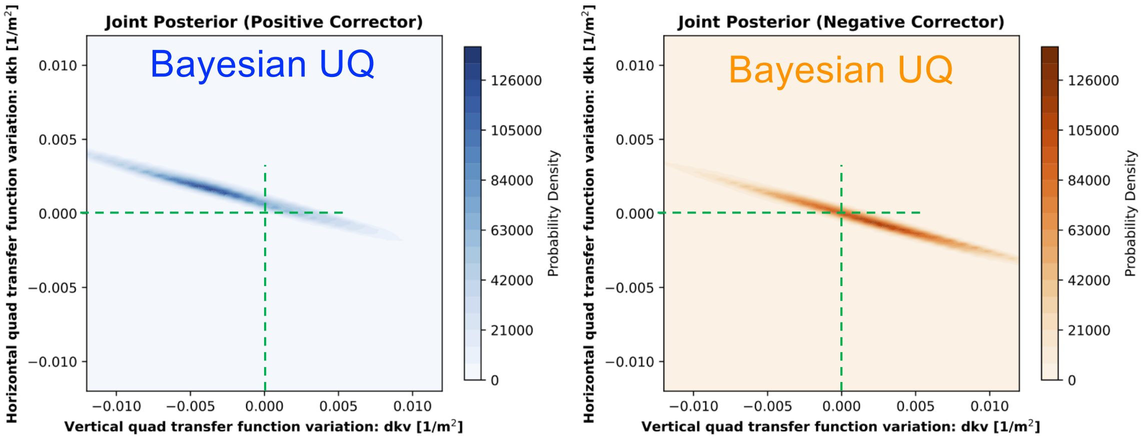
2024 C-AD MAC

December 17, 2024

Responses to the last MAC

Recommendation 7 (ML for beam polarization increase (FOA)):
Include at least the first-order effect of field overlap between nearby magnets (change in effective magnet field length) in the model.

➔ We have analyzed orbit responses to corrector changes in the Booster and subjected them to Uncertainty Quantification by Bayesian analysis. This showed that an effective magnet length error common to all quadrupoles cannot be identified as responsible for discrepancies between measurement and simulation. We are continuing the UQ to identify the source of the Booster's orbit responses.

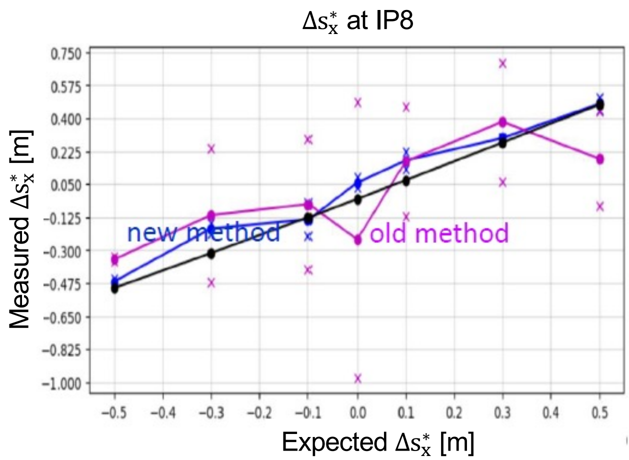


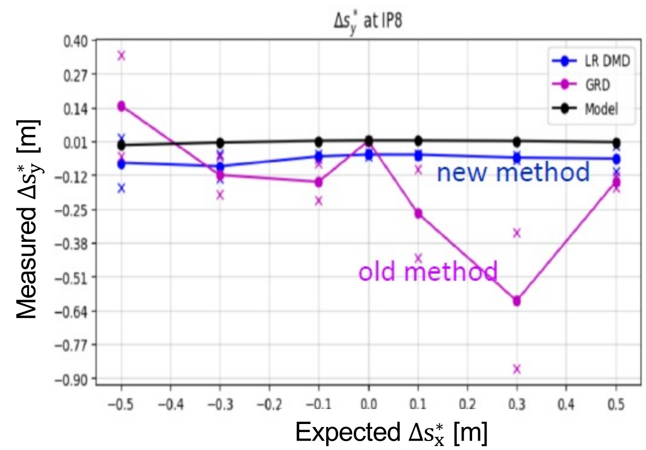
For this MAC we focus on
ii) the plans to for maintaining and upgrading the hadron injector complex for EIC, and (iii) the presented R&D efforts.

Responses to the last MAC

Recommendation 8 (ML for luminosity maximization, FOA with PI Xiofeng Gu):
Pay more attention to the results of the optimization runs in order to gain insight of the process as well as better characterize the system being optimized (sPHENIX luminosity, or EBIS as the case may be).

➔ The GPTune optimization framework was integrated into the control software and experimental measurement loop, which was successfully tested during the first Accelerator Physics Experiments (APEX) session. Exactly following the recommendation, during this optimization process, it was observed that the s* control algorithm required adjustments. A new s* control method was developed, and a detailed description is being prepared for publication in a journal paper.





Meanwhile, the global accelerator parameters (beam losses and orbit for example) were well maintained during the APEX session. An observed tune variations of 0.005 can be controlled with tune feedback. The system is available for s^* optimization as needed.

DE-FOA-0002875 : ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR AUTONOMOUS OPTIMIZATION AND CONTROL OF ACCELERATORS AND DETECTORS

Title: Beam polarization increase in the BNL hadron injectors through physics-informed Bayesian Learning

Collaborators: BNL, Cornell, SLAC, JLAB, RPI

Budget: $1.5M, 09/01/2023 to 8/31/2025

Funding through DOE-NP DE SC-0024287, contr.# 2023-BNL-AD060-FUND

Funding officer Manouchehr Farkhondeh

FOA requested topic:

- Address the challenges of autonomous control and experimentation
- Efficiency of operation of accelerators and scientific instruments

New NOFO (DE-FOA-0003458, Artificial Intelligence and Machine Learning Applied to Nuclear Science and Technology) for a continuation proposal.

Desired result: higher proton polarization

- What high-impact operational challenge can be addressed by MI/AI? ➔ Polarized protons.
- From the source to high energy RHIC experiments, more than 20% polarization is lost.
- The EIC asks for 70% proton polarization, which is 5% higher than even a good RHIC run.
- Polarized luminosity for longitudinal collisions scales with P^4, i.e., a factor of 2 reduction!
- The proton polarization chain depends on many delicate accelerator settings form Linac to the Booster, the AGS, and the RHIC ramp.
- Even 5% more polarization would be a significant achievement.
- Approximately 2/3 of the polarization loss is in the injector chain.
- Accelerator time in RHIC is much less available than in the injector chain.

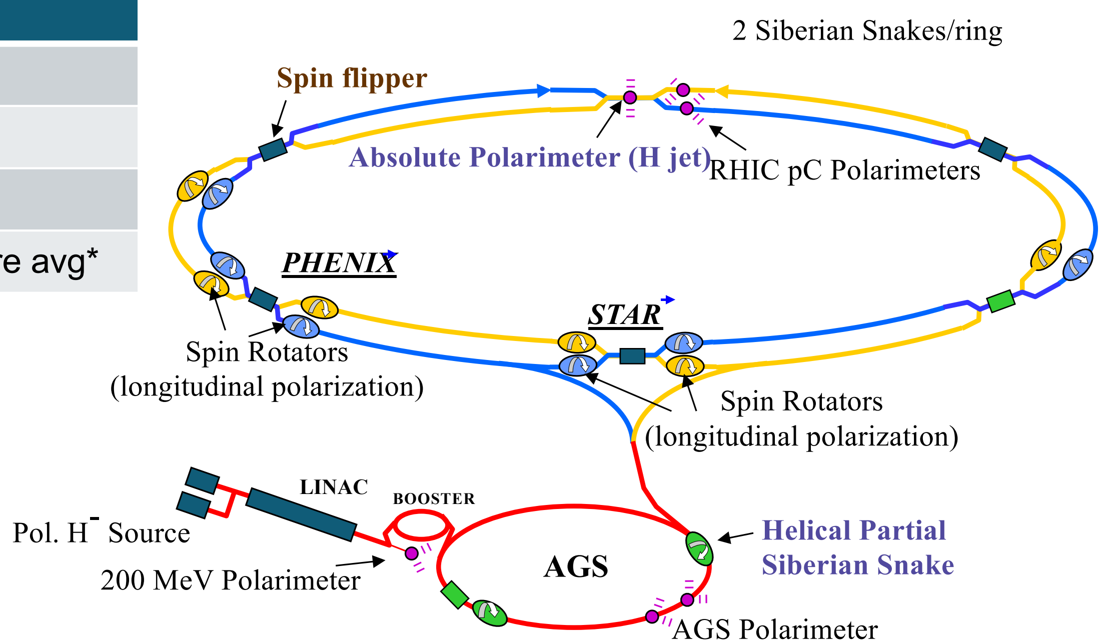
➔ Focus: polarization increase from the injector chain.

The polarized proton accelerator chain

RHIC Polarized Beam Complex

	Max tot. Energy [GeV]	Pol. At Max Energy [%]	Polarimeter
Source+Linac	1.1	82-84	
Booster	2.5	~80-84	
AGS	23.8	67-70	p-Carbon
RHIC	255	55-60	Jet, full store avg*

* Includes both ramp loss and store decay



	Relative Ramp Polarization Loss (Run 17, full run avg)
AGS	17 %
RHIC	8 %

Topics that can improve polarization

- (1) Emittance reduction
- (2) More accurate timing of tune jumps
- (3) Reduction of resonance driving terms

Optimizers for different applications

less ← assumed knowledge of machine → more

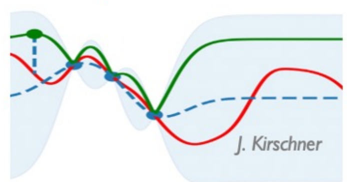
Model-Free Optimization

Observe performance change after a setting adjustment

→ estimate direction or apply heuristics toward improvement

gradient descent
simplex
ES

Model-guided Optimization

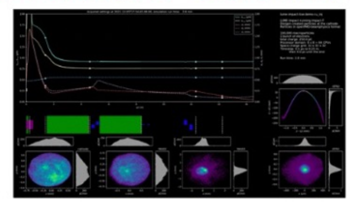


Update a model at each step

→ use model to help select the next point

Bayesian optimization
reinforcement learning

Global Modeling + Feed-forward Corrections



Make fast system model

→ provide initial guess (i.e. warm start) for settings or fast compensation

ML system models +
inverse models

Courtesy Auralee Edelen

Characteristics of polarization optimizations

1. Optimal parameter settings are hard to find, and the optimum is difficult to maintain.

2. The data to optimize on has significant uncertainties.

3. Good, approximate models of the accelerator exist.

4. A history of much data is available.

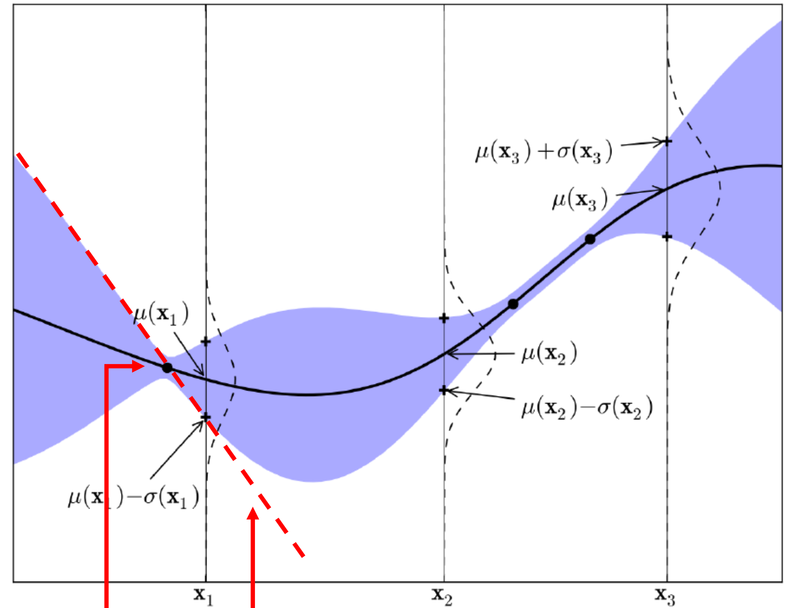
Is this type of problem suitable for Machine Learning?

Why would ML be better suited than other optimizers and feedbacks?

Optimization with Gaussian Processes

- GP model built with scikit-learn library
- A probability distribution over possible functions that fit a set of points
- Mean function + Covariance function

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$



2. The data to optimize on has significant uncertainties.
3. Models of the accelerator exist

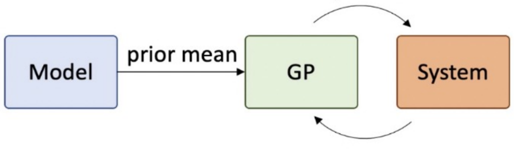
- Kernel: covariance function $k(x_i, x_j)$ of the input variables
- Covariance matrix $\mathrm{K} = k(X, X) = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_t) \\ \vdots & \ddots & \vdots \\ k(x_t, x_1) & \cdots & k(x_t, x_t) \end{bmatrix}$
- At a sample point x_i, Gaussian process returns mean $\mu(x_i|X) = m(x_i) + k(x_i, X)K^{-1}\big(f(X) - m(X)\big)$ and variance $\sigma^2(x_i|X) = k(x_i, x_i) - k(x_i, X)K^{-1}k(X, x_i)$

Merit of physics-informed optimization

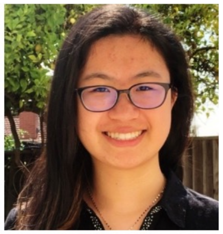
Neural Network System Models + Bayesian Optimization

Combining more expressive models with BO → **important for scaling up to higher-dimensional tuning problems (more variables)**

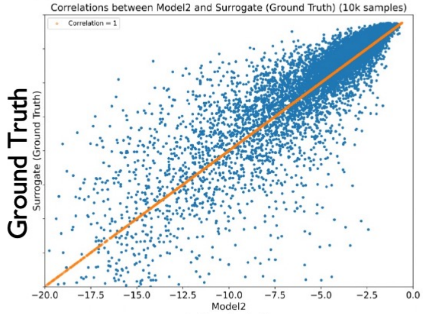
Good first step from previous work: use neural network system model to provide a prior mean for a GP

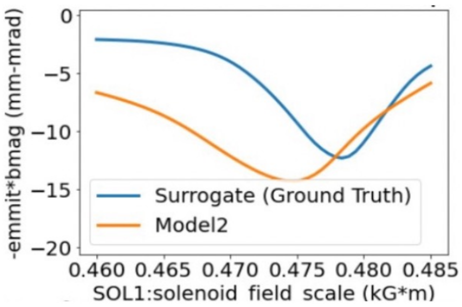


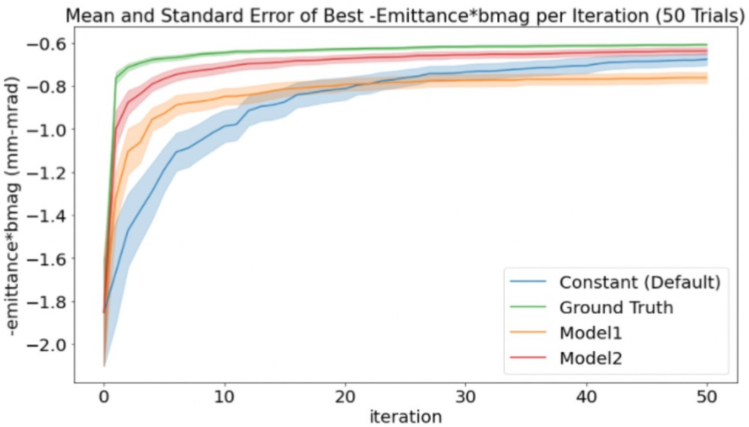
Used the LCLS injector surrogate model for prototyping
variables: *solenoid, 2 corrector quads, 6 matching quads*
objective: *minimize emittance and matching parameter*



Summer '22 undergrad intern
Connie Xu







Even prior mean models with substantial inaccuracies provide a boost in initial convergence
→ now testing on machine and refining approach

Forthcoming paper at NeurIPS ML for Physical Science

Courtesy
Auralee Edelen

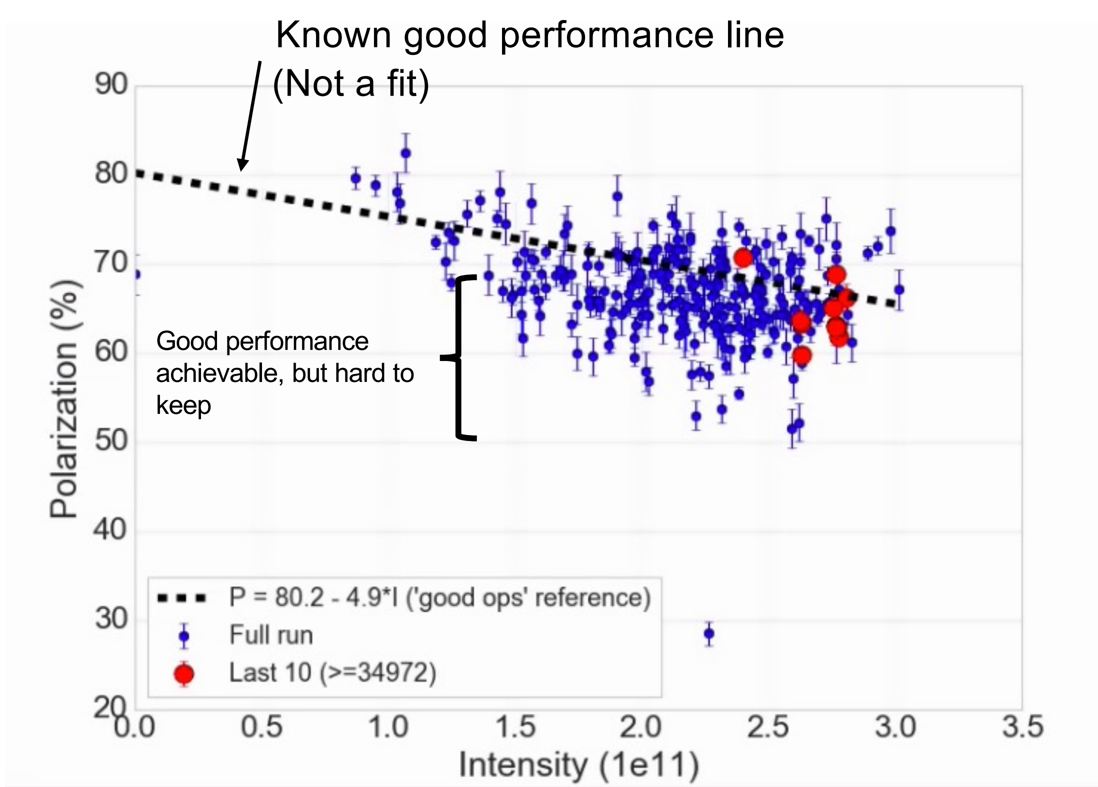
AGS Performance

Highest AGS performance is difficult to achieve ***and*** *maintain*

There is much value in just holding a known optimum

Our optimization approach combines
(a) minimizing and maintaining emittances and
(b) direct polarization interventions

AGS Polarization vs intensity for RHIC fills (Run 24)



Polarized collider performance vs. beam intensity

Collider luminosity, $\mathcal{L}$

$$\mathcal{L} \propto \frac{N^2}{\varepsilon}$$

N = intensity/ bunch
ε = tran. emittance

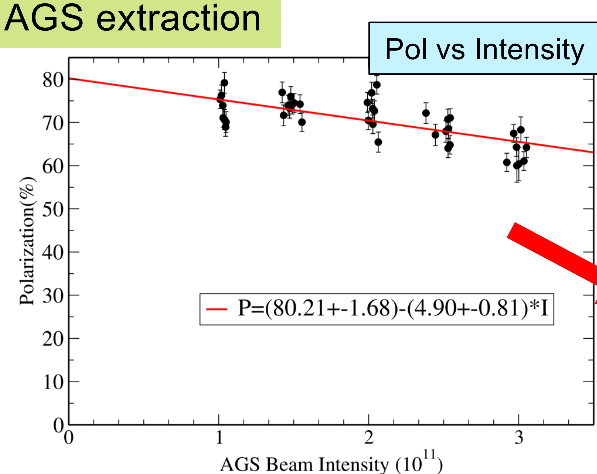
Polarized collider figure of merit (for polarization P):

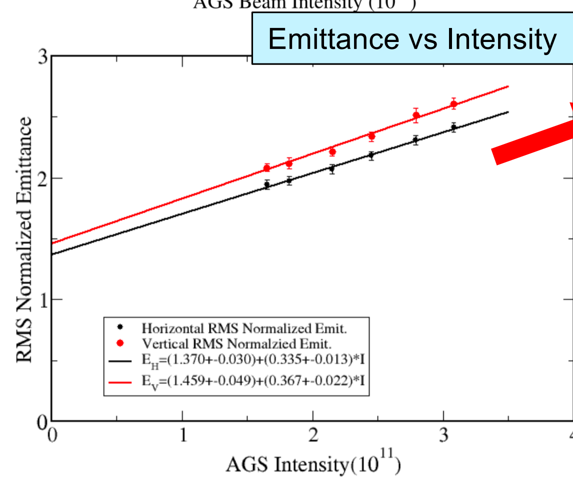
$$\text{FoM} = \begin{cases} \mathcal{L}\, P^2 & \text{transverse spin} \\ \mathcal{L}\, P^4 & \text{longitudinal spin} \end{cases}$$

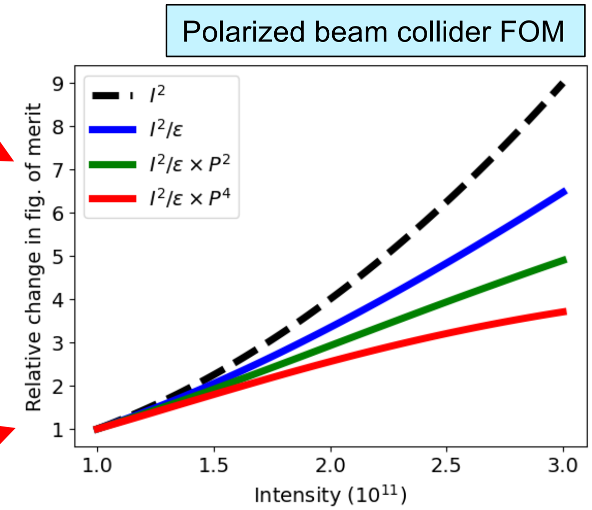
Since both emittance and polarization degrade with intensity figure of merit decreases rapidly

FoM dependence on intensity closer to linear in N than quadratic.

AGS extraction







Impact of intensity increase on FoM given emittance and polarization dependence at AGS extraction

14

Emittance reduction ➔ less depolarization

To reduce and maintain emittances we

- optimize Linac to Booster transfer
- optimize Booster to AGS transfer
- correct optics and orbit in Booster and AGS
- use orbit responses to calibrate models of Booster and AGS.
- split bunches in the Booster for space charge reduction and re-merge them at AGS top energy.

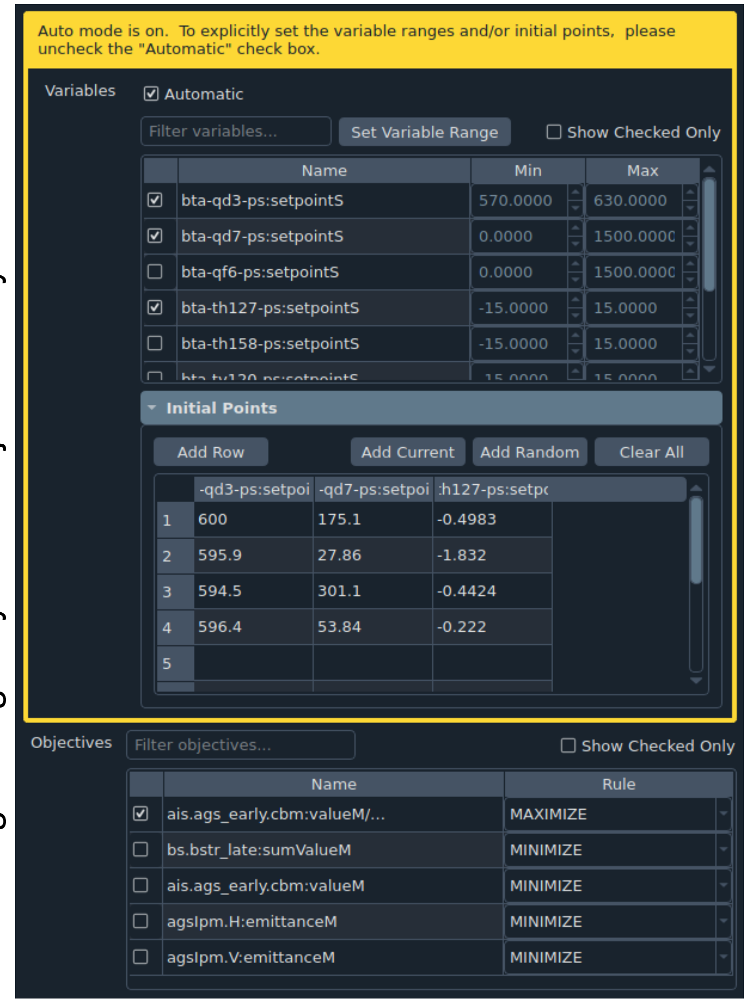
Developed so far, for the control room …

- Bayesian Optimization of Booster injection, available in the control room
- Bayesian Optimization of AGS injection, available in the control room
- Reinforcement Learning routine of AGS bunch merges, partially tested

This allows maintaining optimal emittances automatically.

These were not yet ready for the last RHIC run.

Through Badger by Leve Hajdu and Ryan Russel



Brookhaven National Laboratory

Booster injection

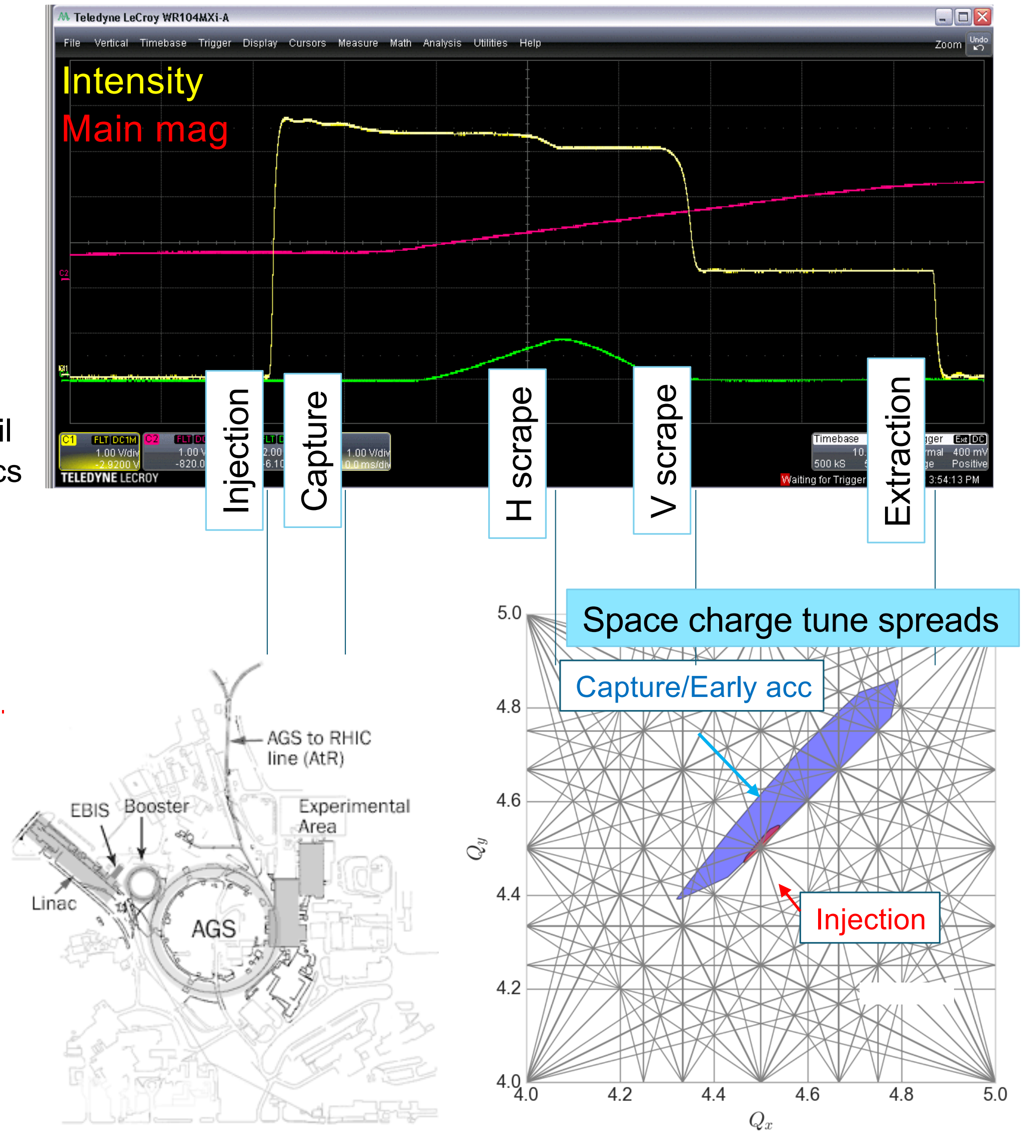
Booster injection/early acceleration process sets maximum beam brightness for rest of acceleration though RHIC

Goal: Get most current past fixed V and H scrapers.

Model building (complex): Injection is complex, incl. ionization foil with 300 turn scattering, H and V scraping, space-charge dynamics an acceleration.

Optimization: Many ”knobs” are available, incl. transfer line magnets, puls length vs. height, RF capture parameters, Booster orbit and optics.
So far, our Bayesian Optimization only uses transfer line magnets.

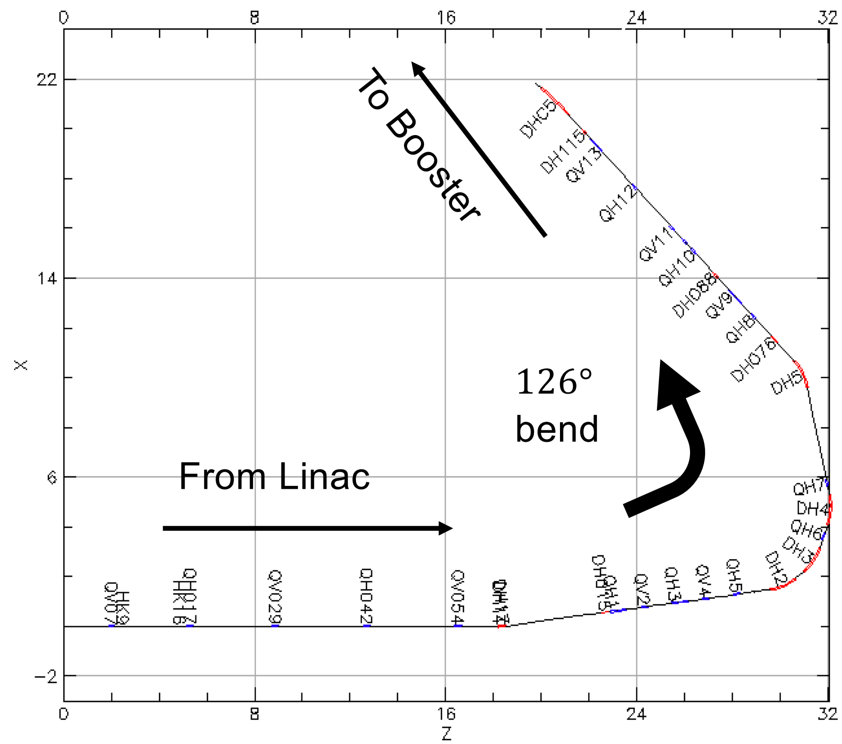
Instrumentation (complex):
- WCM, BPMs don’t work until after capture
- No transverse profile monitor in Booster
- Scraping efficiency as proxy for brightness optimization
- Emittance only measurable in the extraction line via multiwire



Model building for Booster injection

Booster injection process sets maximum beam brightness for rest of acceleration through RHIC

- Transfer line, including complex injection magnet modeled.
- Ionization foil and 300 turn scattering modeled
- Acceleration to H and V scrapers modeled
- Acceleration modeled, not yet with space charge

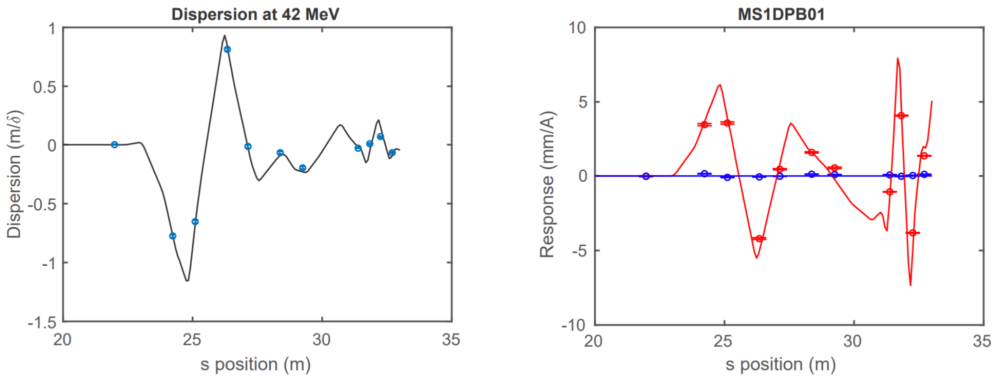


Goals:
(a) set up a digital twin to streamline operations

(b) make Bayesian Optimization physics informed by the model.

Digital Twin for hadron injector sections

A Digital Twin is a bi-directional connection between an accelerator's physics model and its control system.

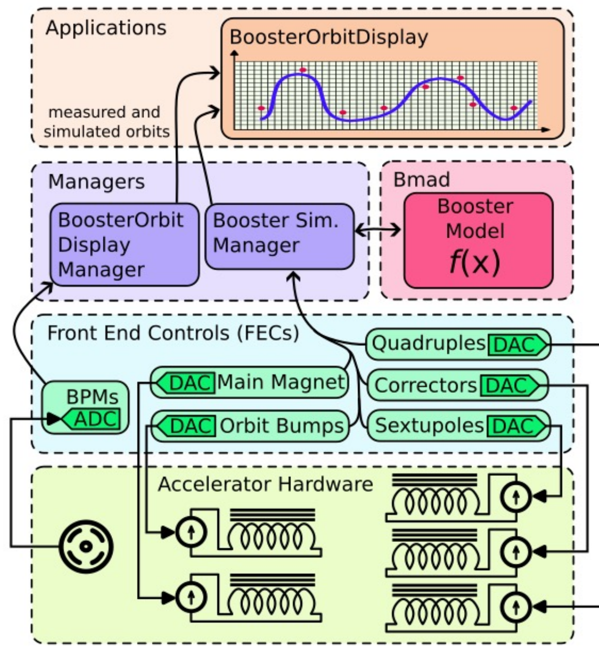


Example digital-twins for CBETA: combine Bmad with EPICS bidirectionally.

- Bmad ➔ control system: DT results are displayed by the control system, just like measured accelerator data.
- Control system ➔ DT: Power supply settings automatically load into the physics model.

Great for continuous comparison of operations and model.
Great for offline development of operations procedures.
Great for virtual diagnostics.

- Additional benefit: Neural network can be trained to predict slow to simulate beam behavior in operations time, e.g. space charge dynamics.
- ML control routines always have the up-to-date physics model available.



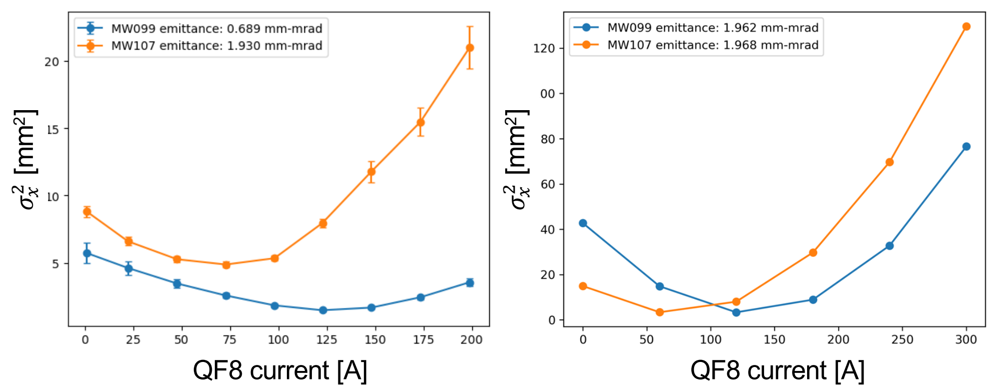
DT currently being prepared for the Booster.

Speed up of BO with physics information

BO of emittance minimization already works, but it could be faster with model information.

- To model injection into the booster, the beam's phase space distribution in the LtB line needs to be known.
- While a NN can be trained to determine the beam's phase space distribution from tomography, the current diagnostics does not permit to resolve x-y coupling.
- Polarized proton beam has such coupling because it is created in a solenoid field.

➔ We will use skew quads in the booster and tilted multi wire detectors to resolve x-y coupling.
➔ Then our BO can be extended by a physics informed model.

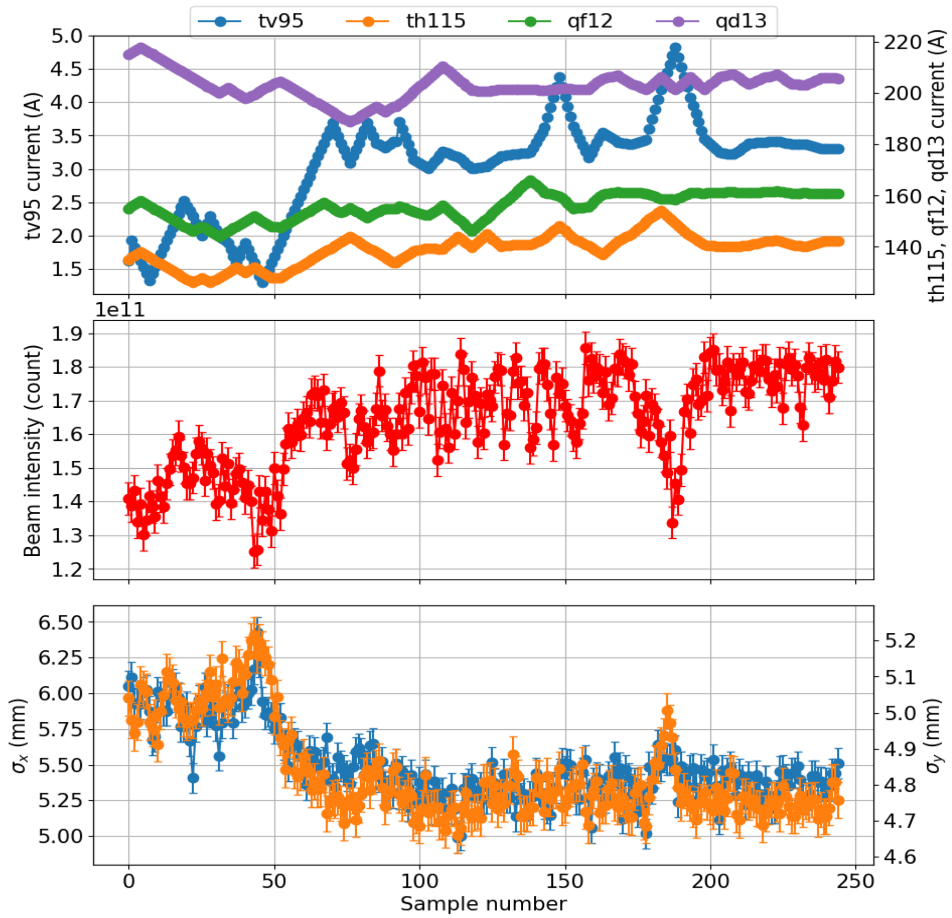


Simulated (left) and measured (right) quadrupole scan results for horizontal quad QF8 observed at two multi-wires (MW099, MW107) in the LtB line.

➔ The x/y projected emittances change along the transfer line, i.e., coupling needs to be considered.

Result: Automatic BO for Booster injection

- Controls: Power supply currents of two correctors and two quadrupoles at the end of the LtB line
- Beam size decrease in both planes in the BtA line in correspondence with intensity increase



Bayesian optimization of the Booster injection process.

Top: power supply currents of two correctors (tv95, th115) and two quadrupoles (qf12, qd13) in the LtB line.

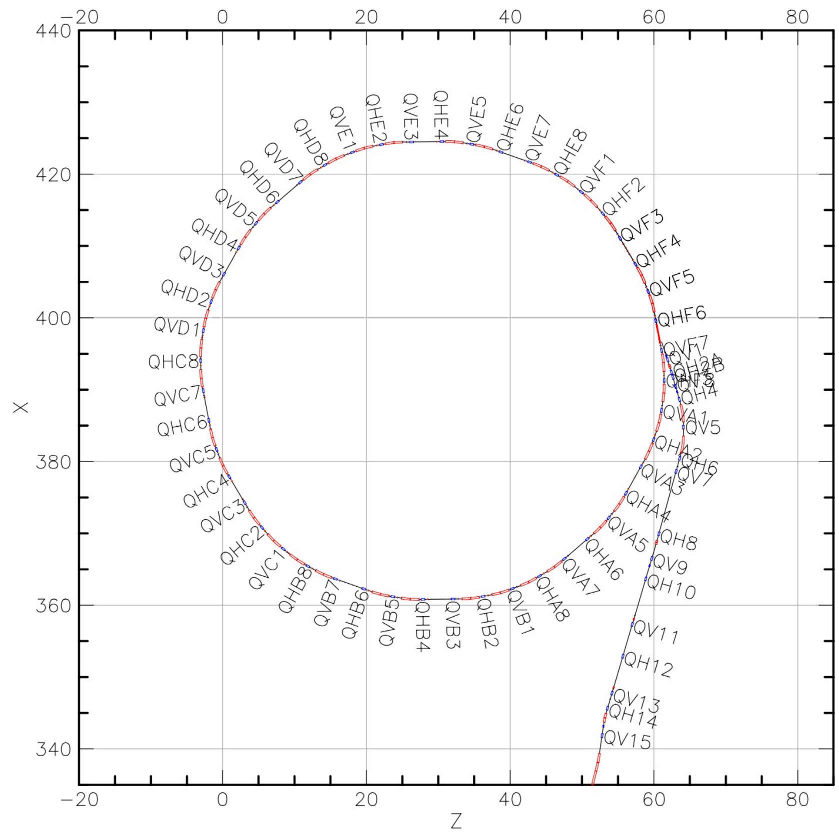
Middle: beam intensity after Booster injection, scaping, and acceleration.

Bottom: Beam size measurements in the BtA line during Bayesian optimization.

Control system: This Bayesian Optimization is now available as a control system application to operators.

BtA Transfer Line Structure in Bmad

- Lattice can be divided into branches connected with forks to simulate connection to a transfer line
- Require documented coordinates for elements to construct correct geometry
- Beam parameters from the end of one branch is automatically inherited by the start of downstream branch → continuous tracking
- BtA universe with three branches
  - 1st branch: Booster ring with extraction bumps
  - 2nd branch: Extraction line from F2 to F6 septum with F3 kicker on
  - 3rd branch: BtA transfer line



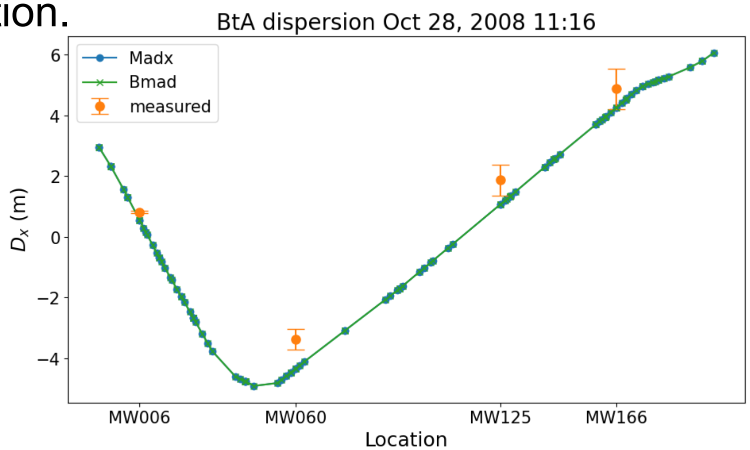
BtA modeling and data comparison

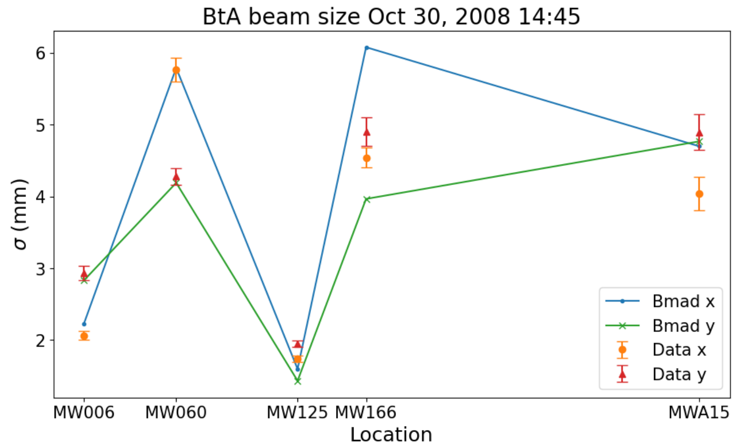
Goals:

(a) set up a digital twin to streamline operations

(b) make Bayesian Optimization physics informed by the model.

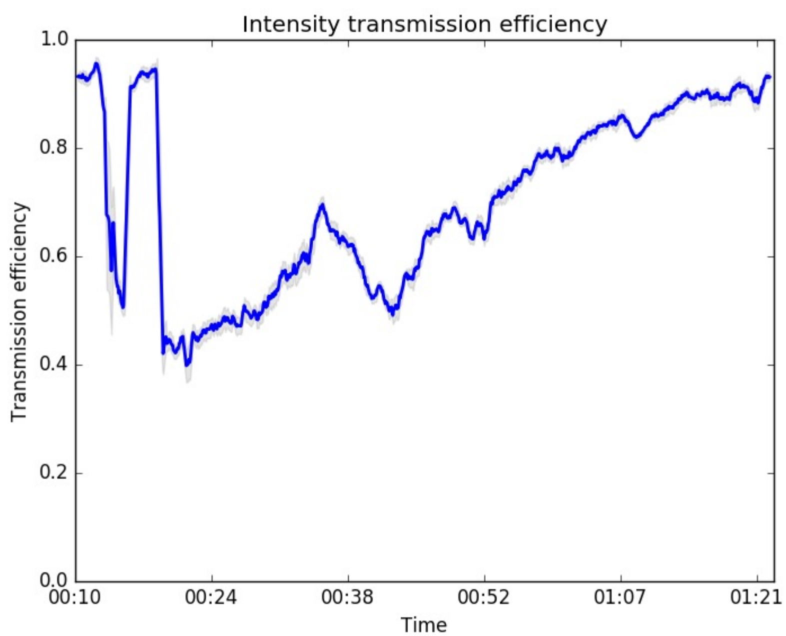
- Bmad tracking leads to horizontal dispersion matching measurements
- Beam size values from bunch tracking show agreements for upstream multi-wire measurements, disagreement downstream needs further investigation
- BO of emittance minimization already works, but it could be faster with model information.

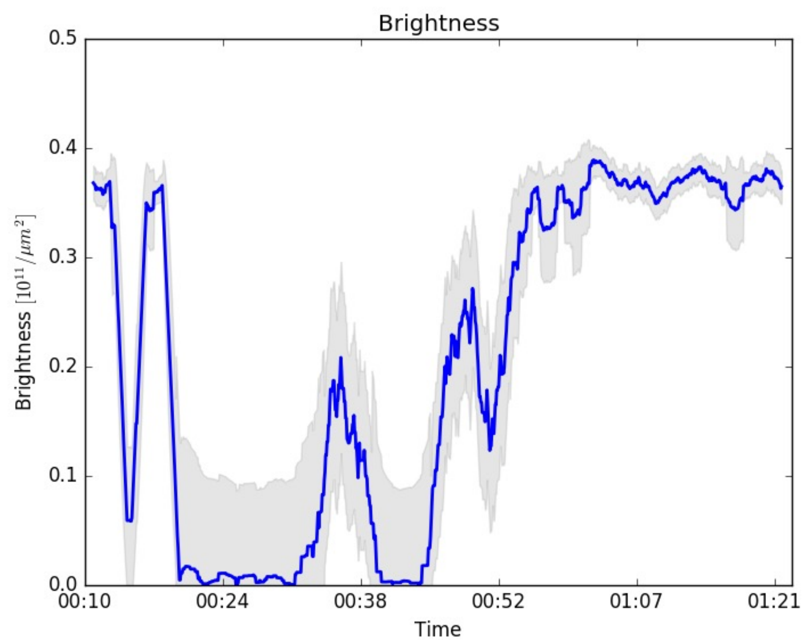




Result: Automatic BO for AGS injection

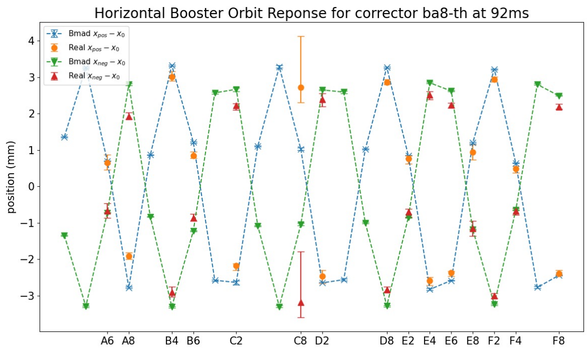
Algorithm efficiently found settings that were different, but at least as good as the previously optimized ones, automatically maintain the AGS injection at optimal performance without human intervention.

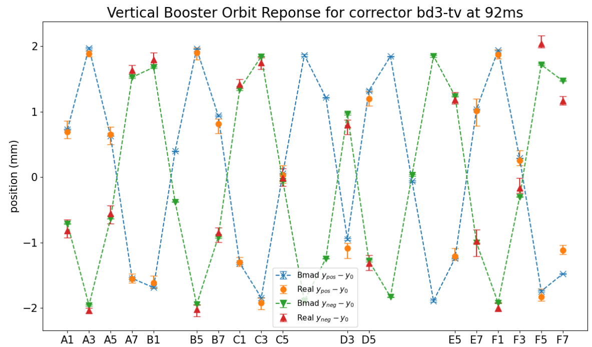




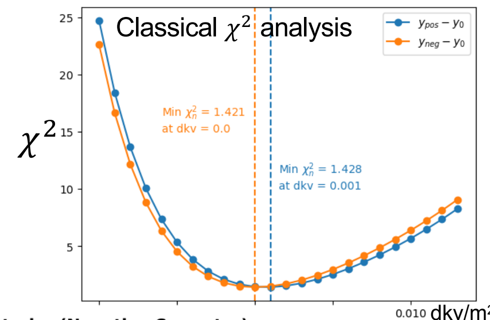
➔ Optimization of current while observing the brightness.

Uncertainty Quantification from orbit responses in the Booster





Orbit response data can be used to find and quantify unknown parameters (e.g., power supply scaling factors, magnet misalignment etc.) in real accelerators,
by Lucy Lin (from C-AD) and Nathan Urban (from CSI)

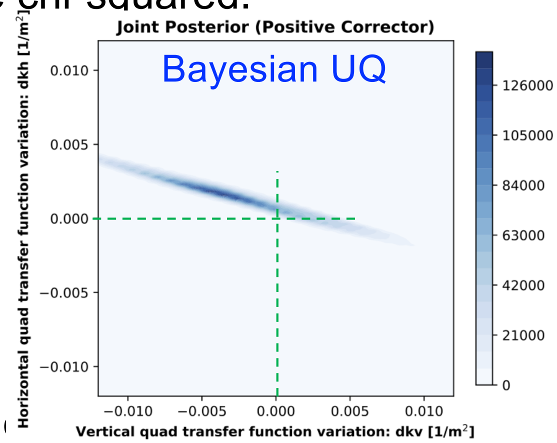


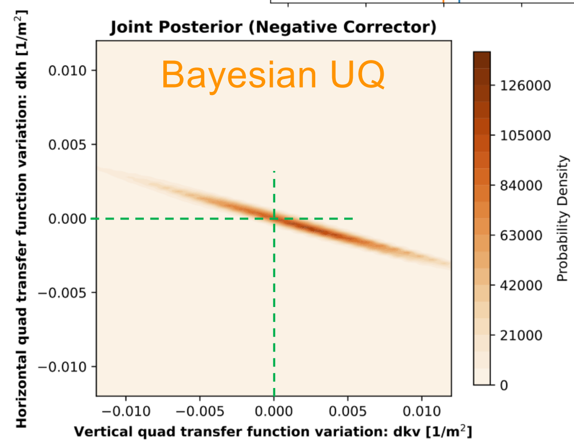
➔ Good agreements between Booster data and Bmad model are reached, with small discrepancies between model and measurement (within 1 mm)

➔ chi-squared/DF = 1.4 for model-experiment. Reasons are analyzed by
(a) Least square fitting to reduce chi-squared.
(b) Uncertainty Quantification.

➔ The main power supply transfer functions (a) do not reduce χ^2, (b) their UQ is consistent with 0

➔ Other error sources are being analyzed.





Georg.Hoffstaetter@cornell.edu

Reinforcement Learning (RL)

Extension of our initial goal of using physics-informed Bayesian Optimization: Can RL have advantages over BO for accelerator controls?

+ RL uses more data about the accelerator (as state variables) even if relationship to the optimum is not known.

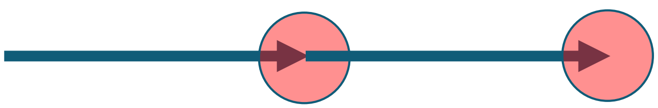
+ RL follows an optimum setting, even when the system changes ➔ accelerator control not only optimization.

-- RL requires millions of data points and may seem inapplicable to accelerators, but with an improved model may deliver many of these points, making RL feasible.

Many system parameters can be measured and may be implicitly related to the the optimum.

Example: Drive along a road with one viewpoint that has uncertainty

A second viewpoint increases the directional accuracy, even if it has a similar error.



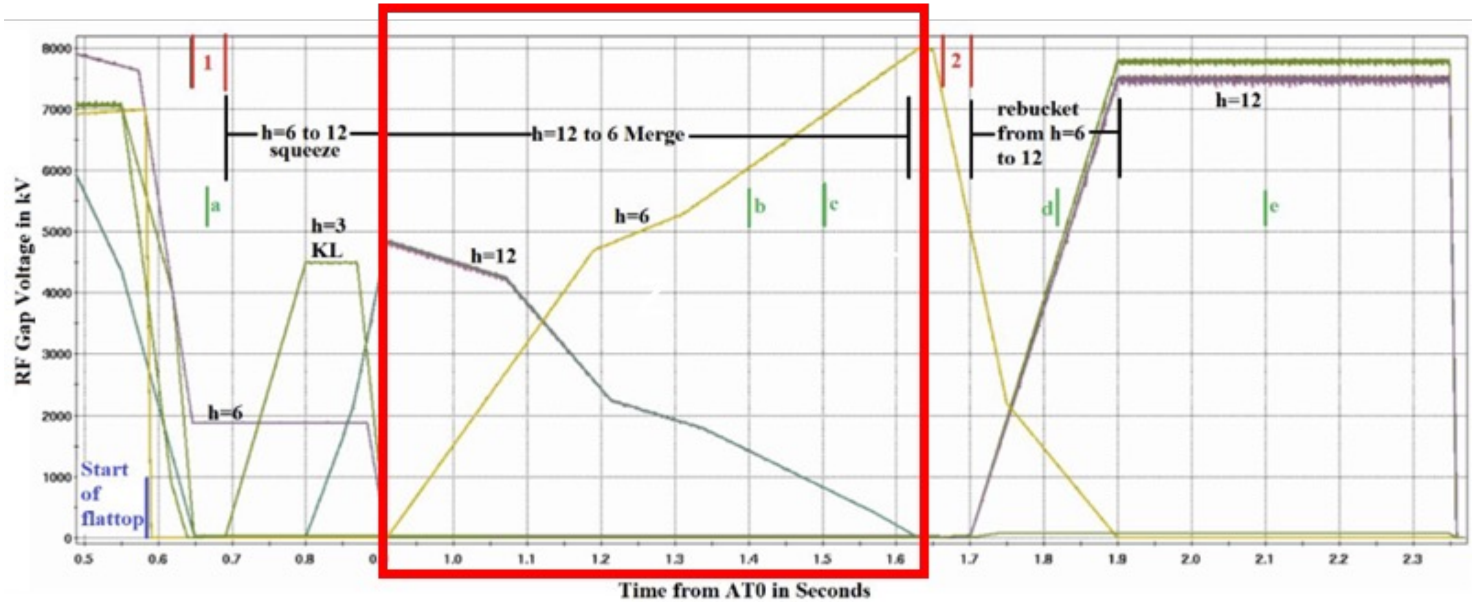
This relies on knowing how the second measurement is related to the direction.

Reinforcement Learning empirically learns this relationship!

This relationship remains useful, even under small changes ➔ follows optimum.

Bunch splitting in Booster / merging in AGS

Splitting in the Booster and merging after AGS accelerator reduces space charge and emittance growth ➔ more polarization



Three RF amplitudes (h=3, 6, 12) in the AGS during bucket manipulation and merging.

➔ We have set up Reinforcement Learning for the merging section.

Reinforcement Learning Tuning
test - varying 6 voltage points for each RF system

RF voltages | Bunch traces | Last trace

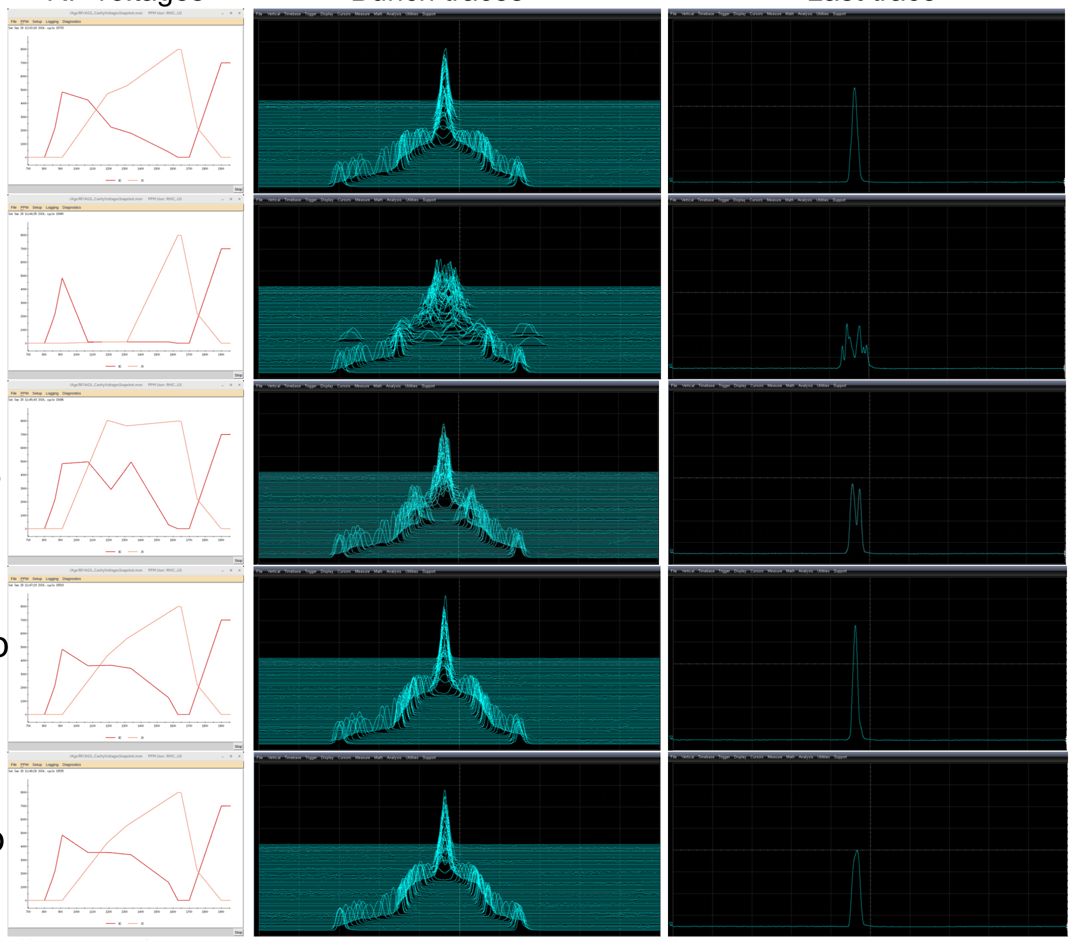
Original

Start

1st step

2nd step

3rd step



Goal: minimize the longitudinal emittance after bunch merging

RF amplitudes as function of time have been optimized in experiments.

Automatic readout of longitudinal emittance not yet available, test used simulated bunch lengths as reward.

Plan: check whether Reinforcement Learning has advantages over BO.

Plan: Include also RF phases as actors and coherent oscillations as state variables.

Determine useful state variables

- measurable
- related to the reward

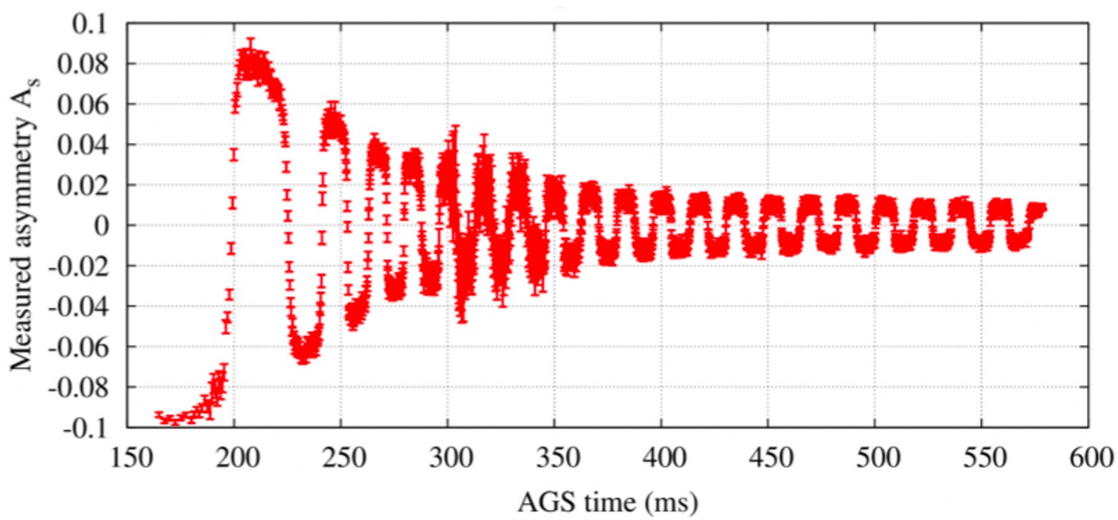
Timing of tune jumps

The G-gamma meter and accurate energy vs. time

(1) Measure the energy by orbit + revolution frequency measurement

(2) Measure of energy by field + revolution frequency measurement

(3) Measure energy by spin flip at every integer spin tune



Combined optimization

➔ measurement with reduced uncertainty at every energy

➔ better timing

➔ higher polarization

Still being worked on. It is less critical with new skew-quad resonance minimization.

Model building for the AGS

- Proton energy range 2.5 GeV -> 23 GeV
- Polarization preserved using
  - helical dipole snakes
  - + horizontal tune jump
  - Resonance correction in development (would replace tune jump)

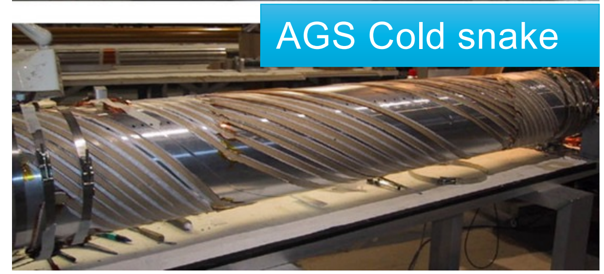
- Requires “near integer” tune
  - Orbit, optics unusually sensitive to errors

- Helical dipoles are complicated magnets
  - Large optical effects at low energy
  - Many related magnetic elements for compensation orbit/optics

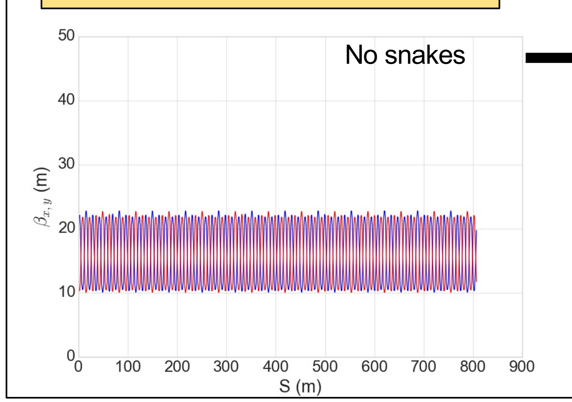
- The complex fields and lattice + high tune requirements are a challenge to modeling (Eiad Hamwi’s work)

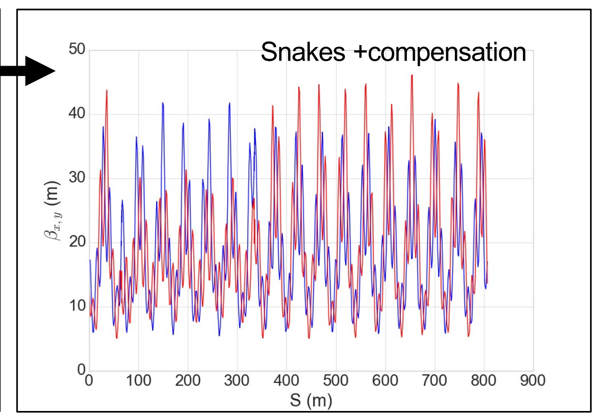
AGS Warm snake

AGS Cold snake



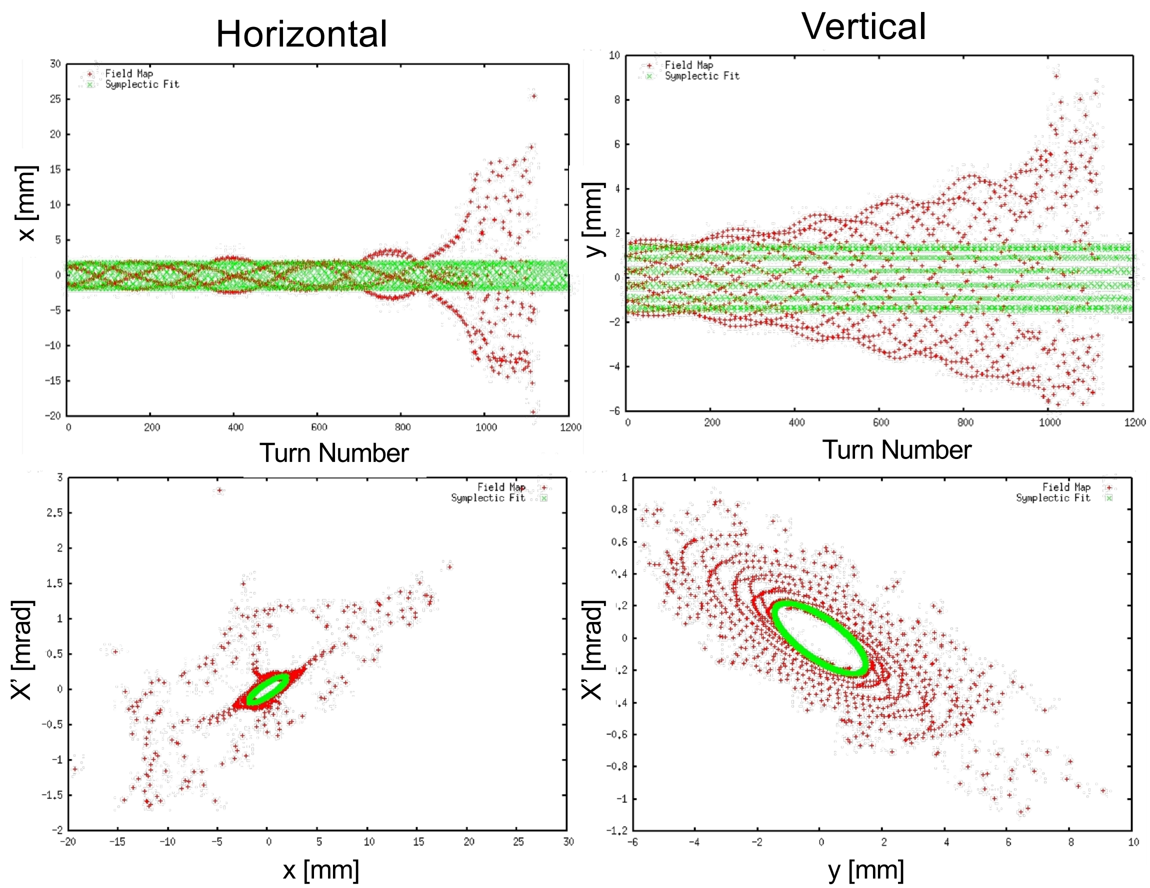
AGS Optics (low energy)





Symplectic AGS Siberian Snake modeling

- AGS Siberian snake field maps violates symplecticity, especially at AGS injection energy
- Symplectic tracking (green) is stable for over 10,000 turns



Reduction of AGS resonance driving terms

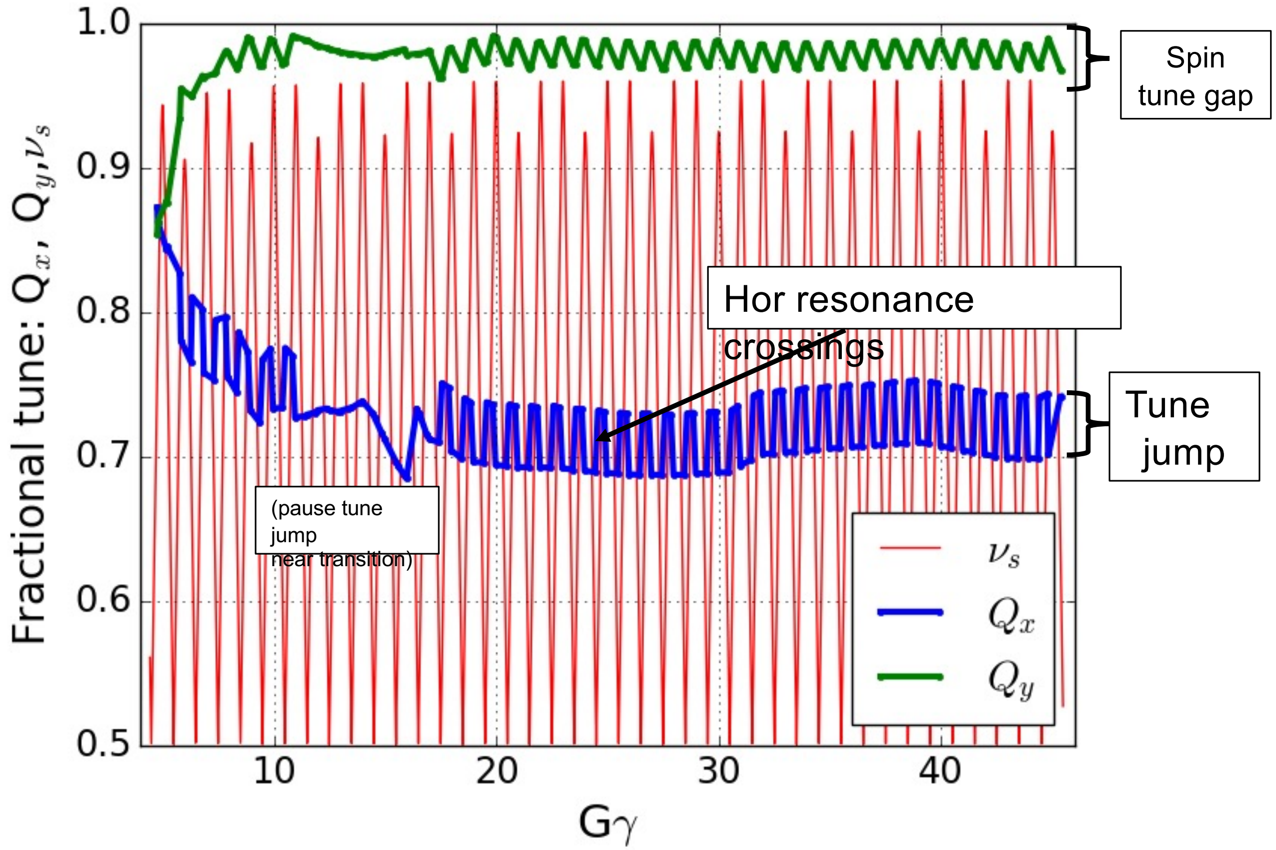
Polarization is preserved in the AGS with two partial helical dipole snakes (10% and 6% rotation)

Provides spin tune ‘gap’ where imperfection and vertical intrinsic resonance condition are never met

- $\nu_s \neq N$ (full spin flips)
- $\nu_s \neq N \pm Q_y$

Horizontal resonance condition still met

- $\nu_s = N \pm Q_x$
- Horizontal resonance are weak, but many (82 crossings)
- Currently handled with fast tune jump
  $\Delta Q_x = 0.04$, 100 μs



Partial snakes drive horizontal depolarizing resonances

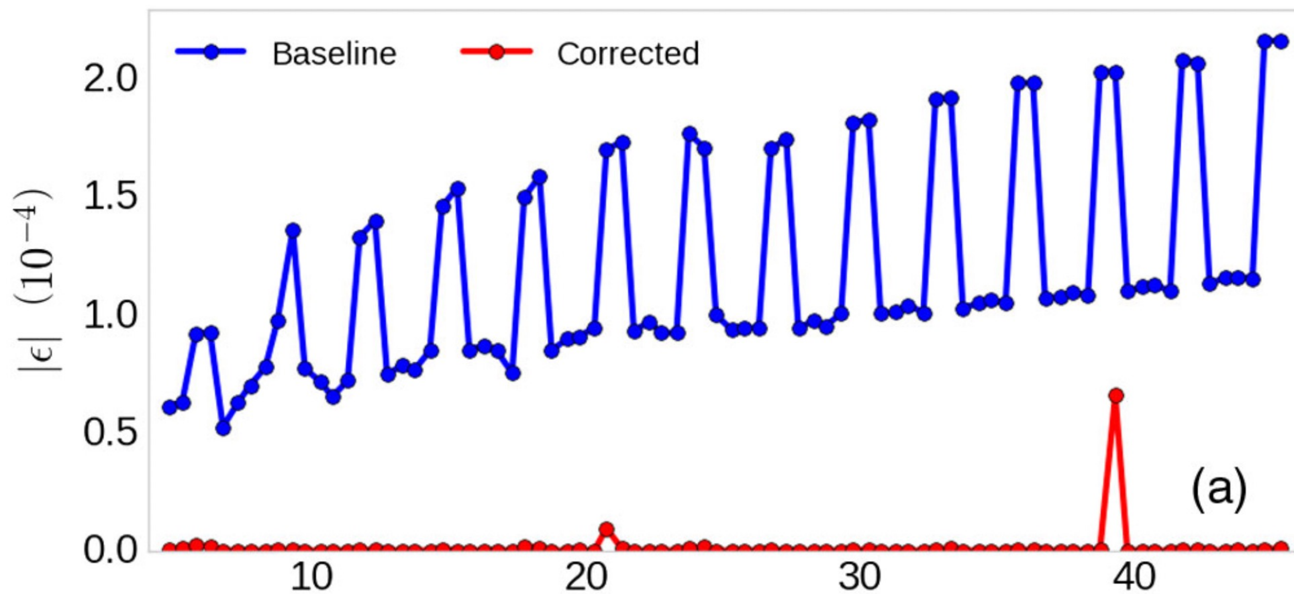
➔ Compensate by other coupling elements, e.g., skew quads

Reduction of AGS resonance driving terms

- Two snakes, separated by 1/3 circumference
  - Modulated resonance amplitude highest near Gɣ = 3N (when snakes add constructively)
- Horizontal resonances occur every 4-5 ms at the standard AGS acceleration rate

ML/AI:
Physics informed Learning of the optimal skew quad strength + optimal timing.

Horizontal Resonance Amplitudes in AGS



AGS Spin Resonance Correction Skew Quadrupoles

- A set of 15 pulsed skew quadrupoles, each with an individual power supply
- Designed to excite coupling resonance to compensate the 82 depolarizing resonances associated with horizontal betatron motion in the AGS partial snakes

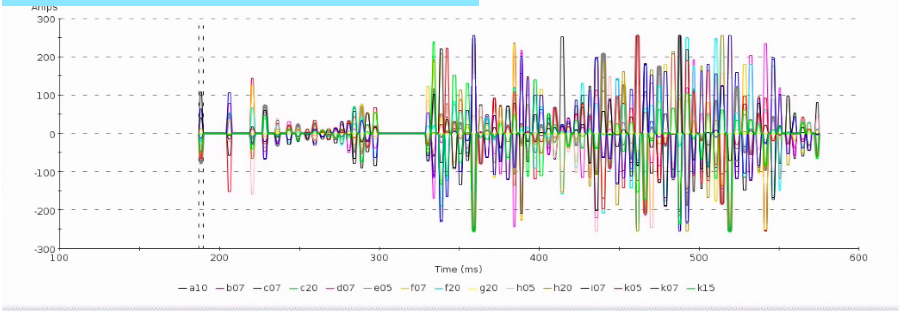
- 15 knobs, 82 different resonances
  - Expected effect is 10-15% gain in polarization
  - A +/-2% measurement takes 5-10 minutes

See presentation by Vincent Schoefer

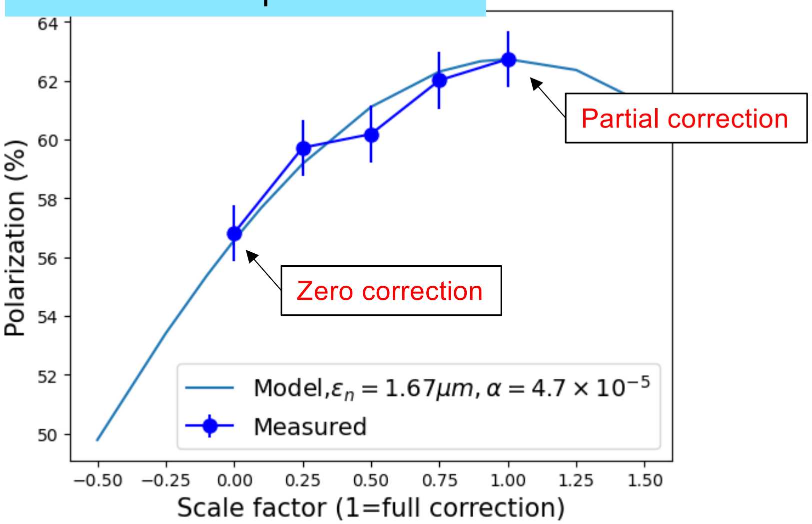
- Run 24: Observation of polarization gain factor (+10%) during acceleration (similar to existing tune jump), with the about the top half the pulses enabled)

- Plans for further improvements (enabling more pulses for 5-10% more gain):
  - Addressing model inaccuracies at low energy
  - Iteration on orbit centering
  - Possible optimizations based on ML methods

Skew quad current pulses



Correction amplitude scan



SciBmad a ML-oriented Toolkits (Libraries)

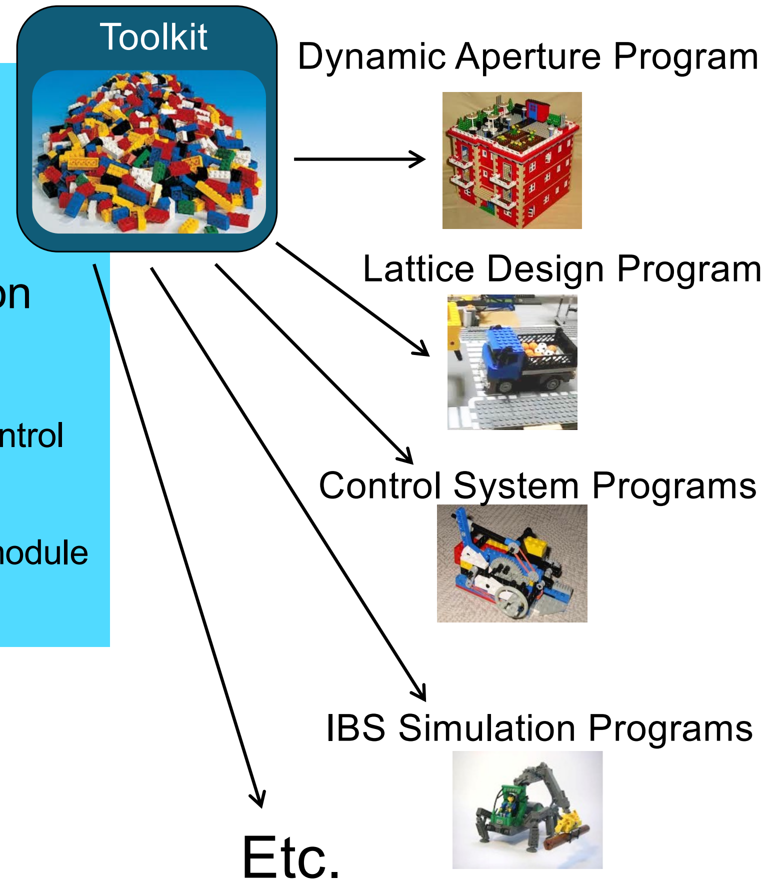
Advantages the toolkit:

Fully differentiable (reverse and forward)

➔ excellent for Neural Network optimizations

➔ Excellent for Bayesian optimization with slope information

- Cuts down on the *time* needed to develop programs.
- Cuts down on programming *errors* (via module reuse).
- Provides a simple mechanism for lattice function calculations from within control system programs.
- *Standardizes* sharing of lattice information between programs.
- Increased *safety*: Modular code provides a firewall. For example, a buggy module introduced into the toolkit will not affect programs that do not use it.



This project is

- funded by DOE-HEP
- has a growing list of collaborators
- has a weekly wise people meetings
- ➔ **is looking for collaborators**

Brookhaven National Laboratory

Summary

- DOE-NP funded project for the enhancement of proton polarization using ML/AI. Goal: 5%.
- Bayesian Optimization for **automatically optimizing and maintaining emittances** in Booster and AGS are available in the control room.
- **Reinforcement learning** routings for optimal bunch merging are being developed and have been tested in the control room.
- Models of LtB line, Booster, BtA line, and AGS have been much improved, will be the basis for **differentiable digital twins**.
- **Excellent team** from BNL, Cornell, JLAB, SLAC, and RPI has formed, two PhD students have graduated (Bohong Huang and Lucy Lin. She is now postdoc in C-AD).
- **Reduction of resonance driving terms** already works above transition energy, may use ML belove γ_t.
- A **continuation proposal** is being written extending to (a) polarized sources and linac, (b) Reinforcement Learning, (c) differentiable Digital Twins, (d) emittances of unpolarized beams.

Dominant Participants

BNL: Kevin Brown, Weinin Dai, Bhawin Dhital, Yuan Gao, Levente Hajdu, Kiel Hock, Bohong Huang, Natalie Isenberg, Nguyen Linh, Chuyu Liu, Vincent Schoefer, Nathan Urban

Cornell: Georg Hoffstaetter de Torquat (also BNL), Lucy Lin, Eiad Hamwi, David Sagan, Matt Signorelli

SLAC: Auralee Edelen

JLAB: Malachi Schram, Armen Kasparian

RPI: Yinan Wang

Radiasoft: Nathan Cook, Jon Edelen, Chris Hall

Thank you and Questions?