Online Autonomous Tuning of the FRIB Accelerator Using Machine Learning

Peter N. Ostroumov, ostroumov@frib.msu.edu

Co-PIs: Kilean Hwang, Dean Lee, Alexander Scheinker

Contributors: Kei Fukushima, Tomofumi Maruta, Alexander Plastun, Jinyu Wan, Tong Zhang

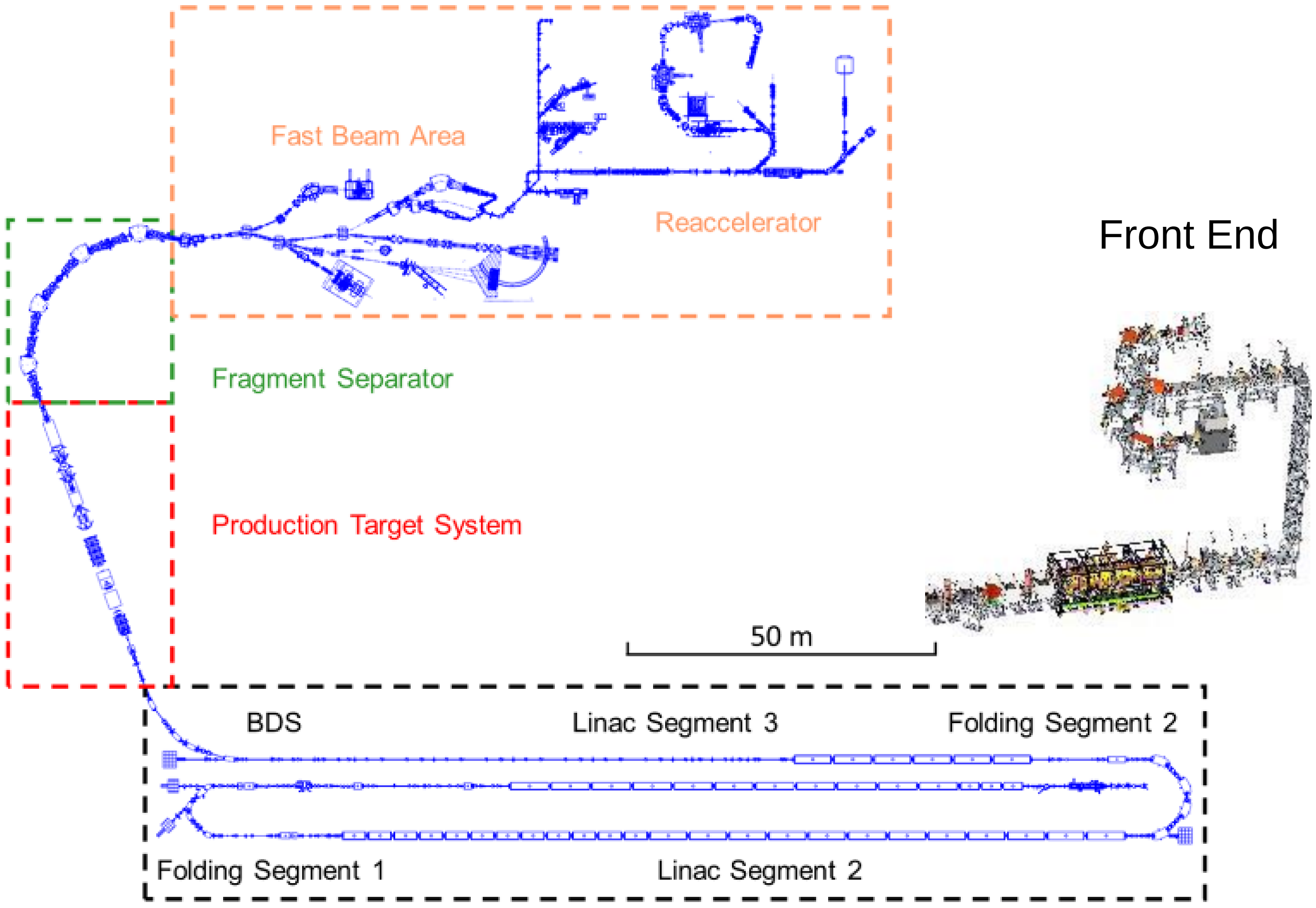
2024 NP AI-ML PI Exchange Meeting

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics, under Award Number DE-SC0024707 and used resources of the FRIB Operations, which is a DOE Office of Science User Facility under Award Number DE-SC0023633.

Content

- FRIB Accelerators
- Accelerator tuning approach
- Examples of ML tasks
- Bayesian optimization (BO)
- Customized BO
- Virtual 4D Phase Space Diagnostics
- Generic BO application for accelerator tuning
- Fast NN model trained on physics equations for online application
- Virtual diagnostics for bunch length measurements
- Virtual diagnostics for beam quadrupolar moments and calculation of Courant Snyder parameters

FRIB Layout

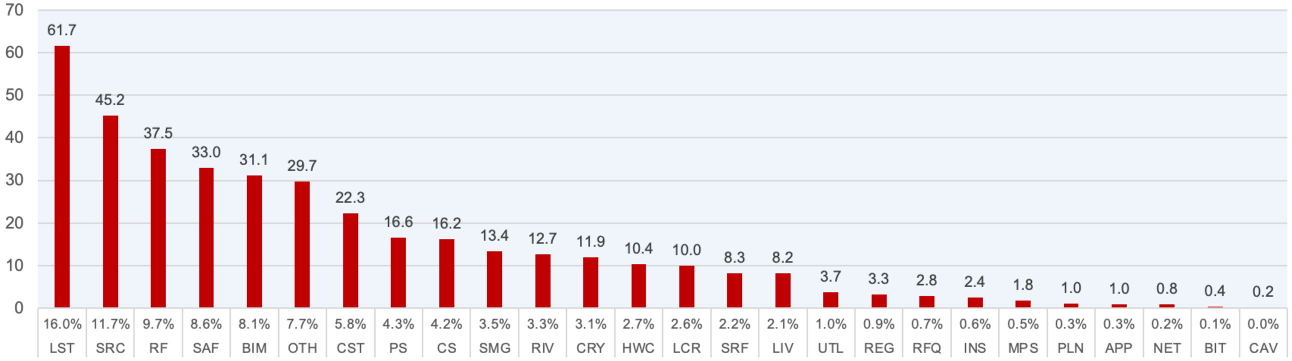


- CW Linear accelerator
  - 200 MeV/u - U, 320 MeV/u - O
  - RFQ, q/A=1/7
  - 324 SC cavities in 46 cryomodules
- 80-meter-long space for upgrade to 400 MeV/u of uranium
- Fragmentation target/Beam dump
- 2-stage fragment separator
- Fast beamlines
- 6 MeV/u Re-accelerator
- In operation since May 2022
- 44 nuclear science experiments conducted since the commencement of user operation

FRIB Operation

- Commencement of User Operation – May 2022
- The first two years of operation
  - Delivered 8500 beam hours at full energy for science and 4000 hours at lower energies, up to 40 MeV/u for Single Event Experiments (SEE)
    - » Beam availability in the first year was 92% and in the second year - 94%
  - 44 science experiments were carried out; the results were reported in multiple PRL papers.
  - More than 270 unstable isotopes produced with primary beams of ^{18}O, ^{20}Ne, ^{28}Si, ^{36}Ar, ^{40}Ar, ^{48}Ca, ^{64}Zn, ^{70}Zn, ^{82}Se, ^{86}Kr, ^{124}Xe,^{198}Pt and ^{238}U accelerated up to 300 MeV/u
- Both primary and secondary beams are extremely stable during the experiment

Breakdown hours by systems

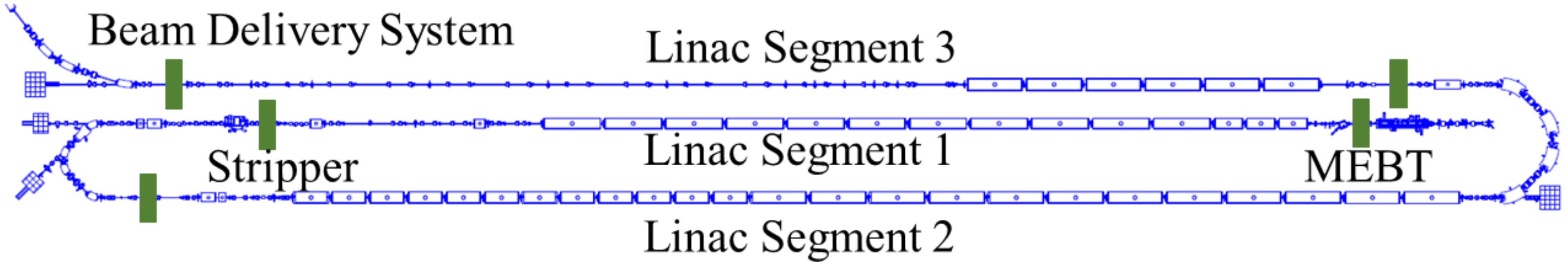


Main systems contributing to downtime

- Lithium stripper
- Ion Source
- RF power
- Safety systems
- Beam instrumentation

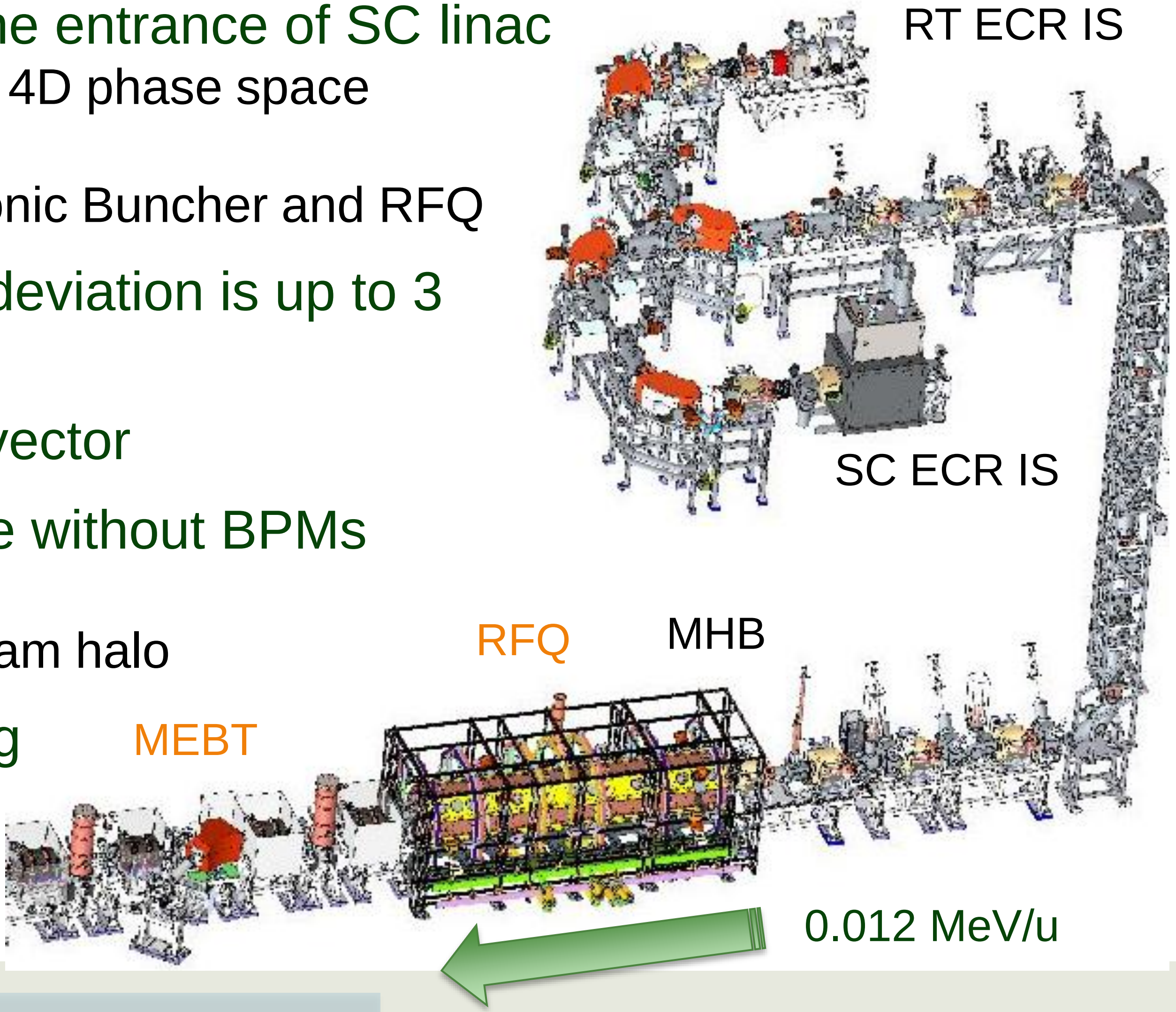
Accelerator Tune Development

- New tunes are required for new beams, more charge states, and transition to higher power
- Virtual accelerator is based on second order envelope code and simulations of the longitudinal dynamics in realistic fields
  - Equivalent to the digital twin of the linac
  - Provides pre-setting of the entire accelerator
- Low-loss accelerator tune is achieved by transverse matching in several sections
  - The setting is verified by envelope mapping in both longitudinal and transverse planes
- Frequent ion species switching; typically occurs every week
  - Rapid beam tuning essential to provide more beam time for science
- Our Solution: Advanced beam tuning applications based on
  - Physics simulation
  - Machine Learning (ML)
  - Classical Optimization (e.g. Nelder-Mead)



Examples of ML Task in Front End

- Problem in restoring previously tuned beam at the entrance of SC linac
  - ECR produces slightly different beam distribution in 4D phase space
  - The beam central trajectory deviates
  - Transverse-longitudinal coupling in the Multi-Harmonic Buncher and RFQ
- Result: at the MEBT, transverse beam centroid deviation is up to 3 mm; phase deviation is up to 4 degrees
- Task: match beam centroid to previously tuned vector
- Steer beam through two apertures in straight line without BPMs
  - Limited fast beam centroid diagnostics.
  - Maximize beam current through apertures ⇒ cut beam halo
- RFQ transmission and MEBT 6D beam centering
- LS1 longitudinal beam center restoration



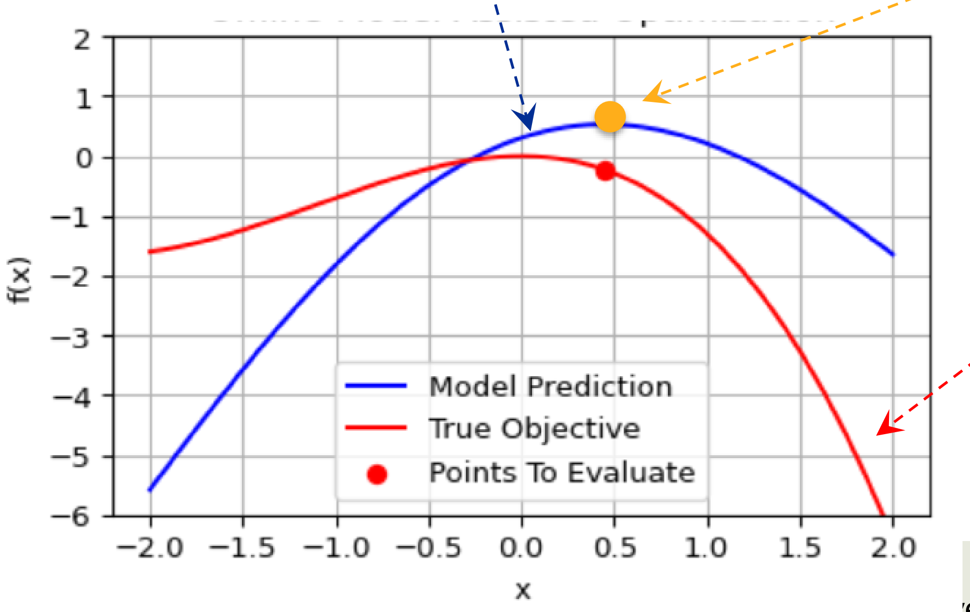
Surrogate Model Assisted Optimization

- We consider problem of finding optimum set of control x^* that maximize an objective $f(x)$

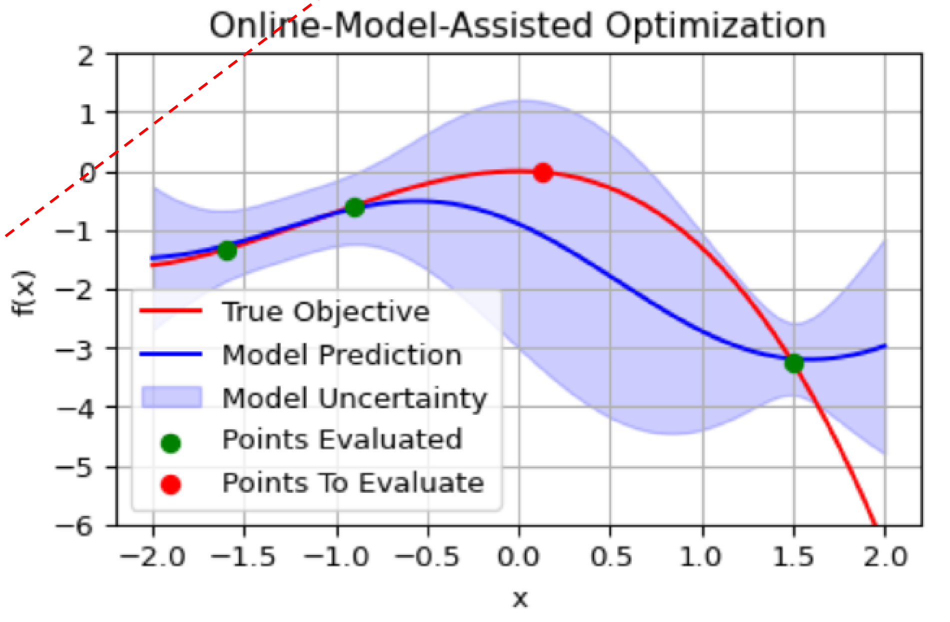
$$x^* = \text{argmax}\, f(x)$$

- If we have a model $m(x) \simeq f(x)$, can find solution candidate $x_0 = \text{argmax}\, m(x)$

Optimization based on the physics or data-based deterministic surrogate model



Online Bayesian optimization



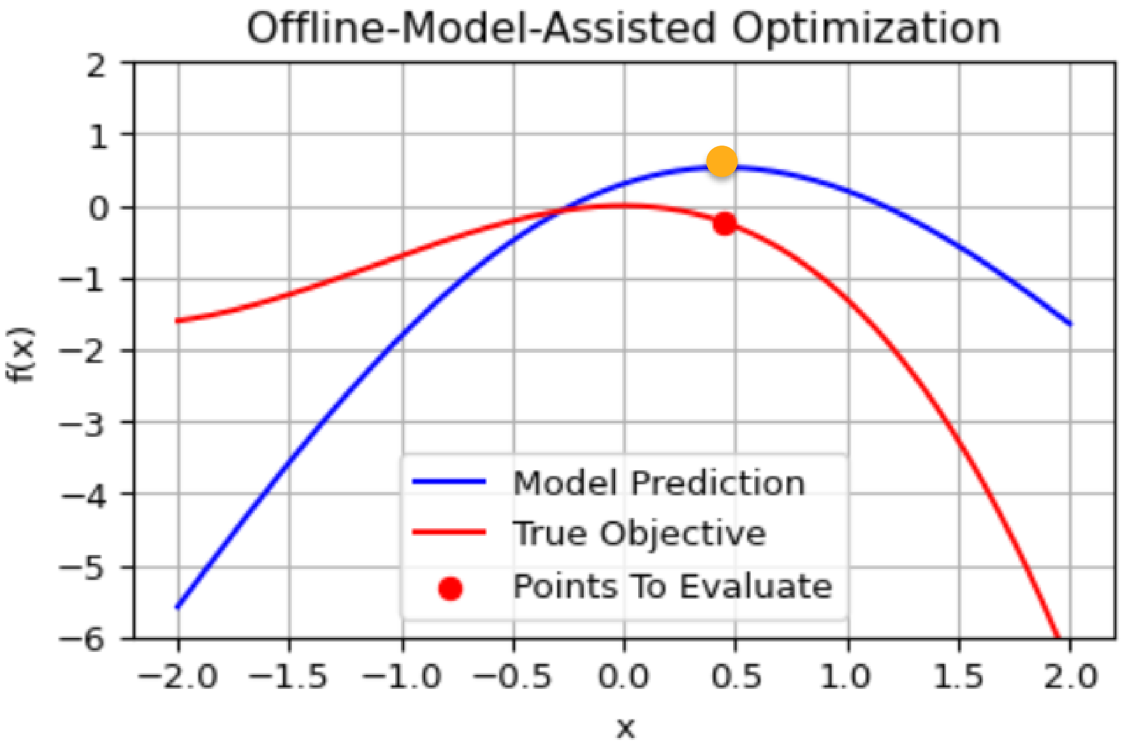
FRIB
U.S. Department of Energy Office of Science | Michigan State University
640 South Shaw Lane • East Lansing, MI 48824, USA
frib.msu.edu

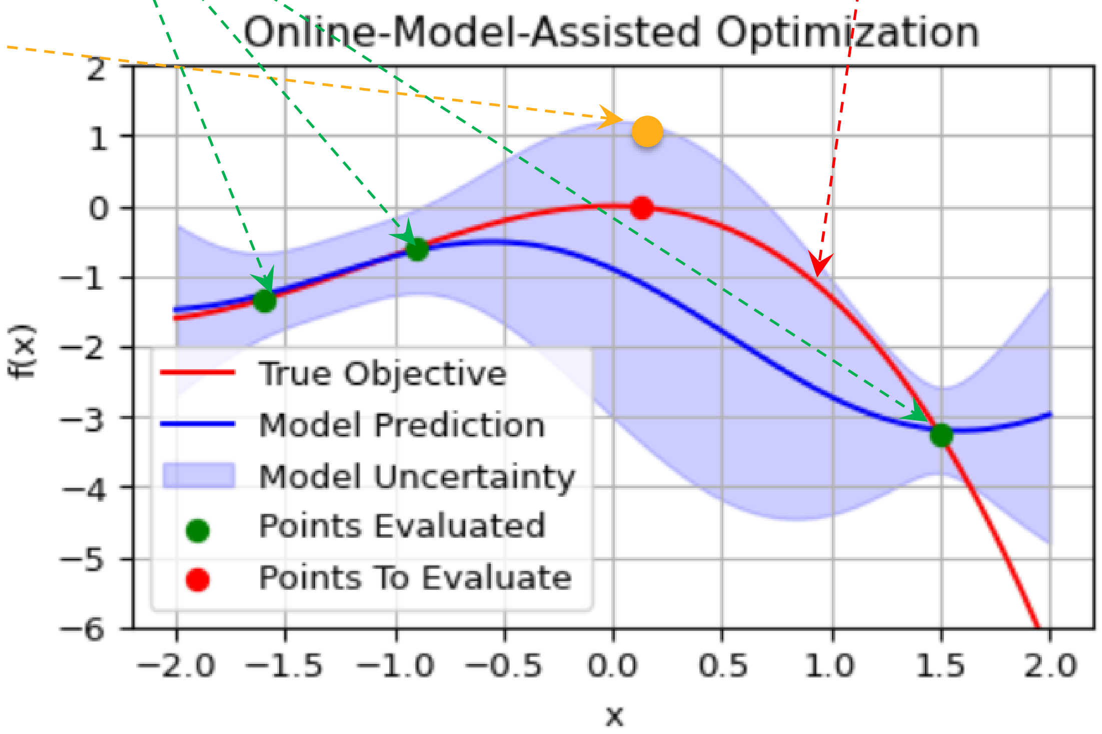
Surrogate Model Assisted Optimization

- We consider problem of finding optimum set of control x^* that maximize an objective $f(x)$

$$x^* = \text{argmax}\, f(x)$$

- If we have a model $m(x) \simeq f(x)$, can find solution candidate $x_0 = \text{argmax}\, m(x)$
- If we have a data-based probabilistic model $\mathcal{P}(y|x, \mathcal{D})$, we can query next candidate x_n that can account for model uncertainty.

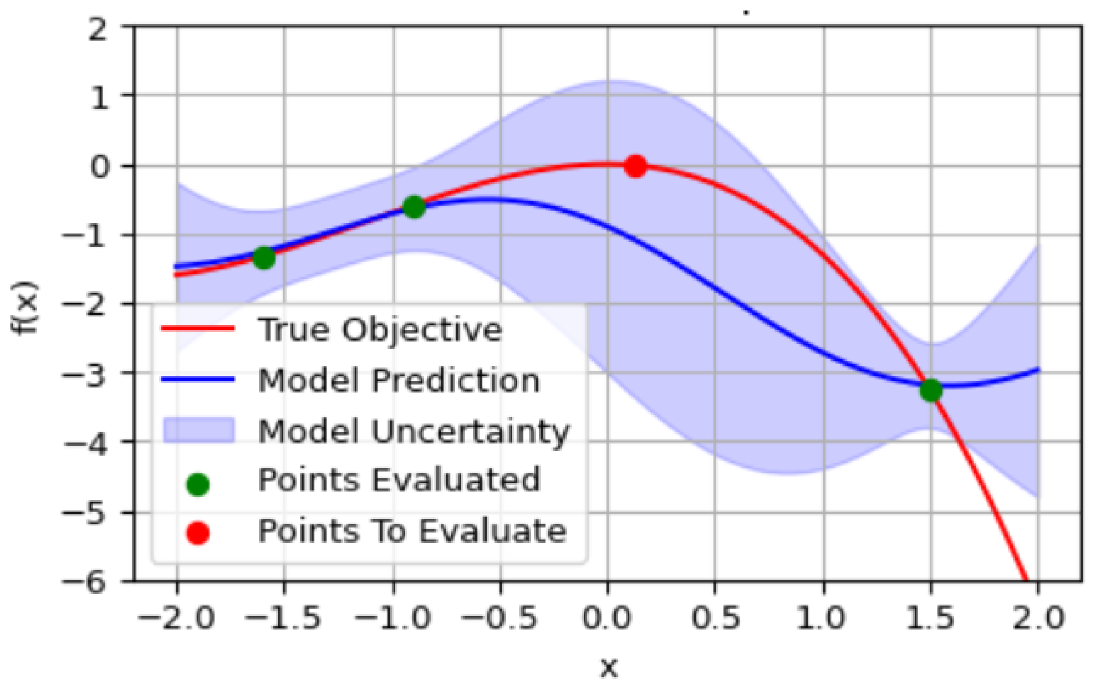




Bayesian Optimization

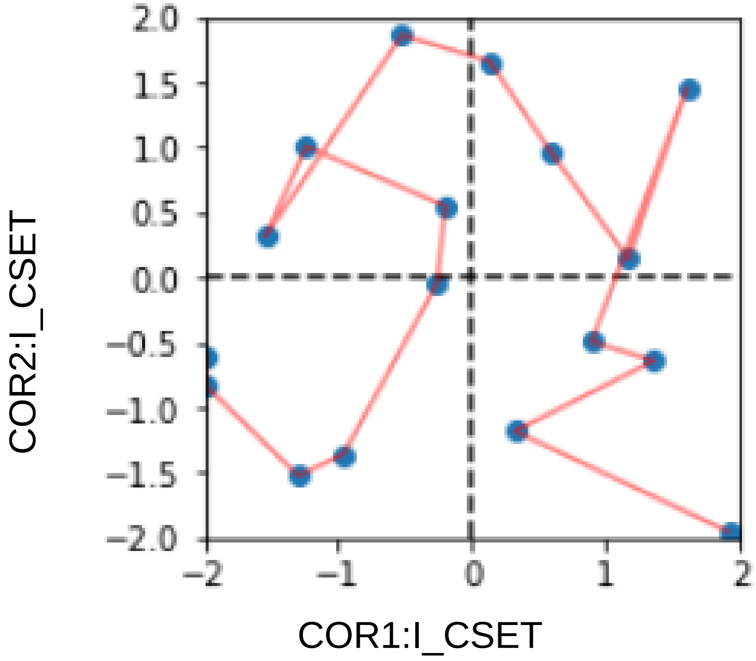
- We consider the objectives that takes few to several seconds to evaluate
  - » Power supply ramping from old control x_i to new control x_{i+1} : generally a few sec but can take up to 30 sec
  - » Average out noisy measurements for a few sec.
    - Due the ramping time, a few sec of averaging cost will benefit for training surrogate model than noisier data.
  - Depending on the problem size, the overall time budget ranges from 2 to 20 minutes.

- Given the few to several seconds of evaluation time, BO is an appropriate choice
  - Due to its sample efficiency.
  - A few seconds of numerical cost (for less than 10 dim, and less than 200 data points)
  - The scalability issue of computational complexity of BO won't be a limiting factor as the maximum number of function evaluation will not exceed a few hundreds due to tight time budget.



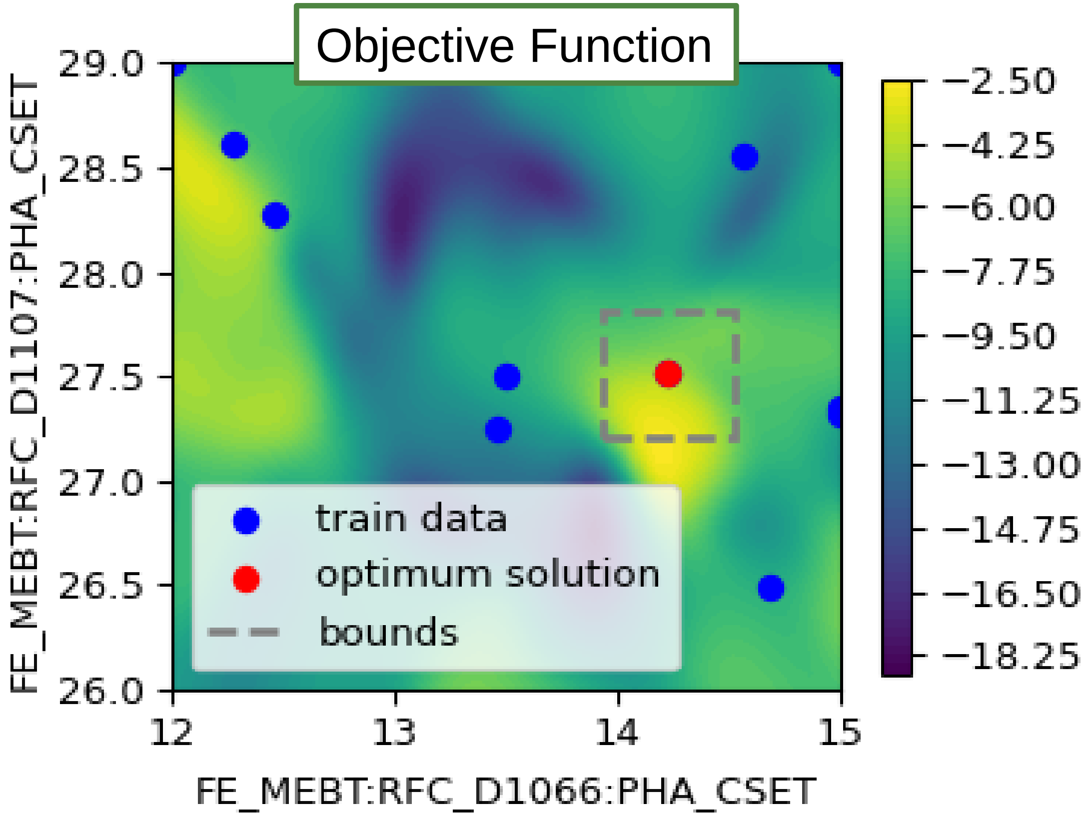
Customized BO

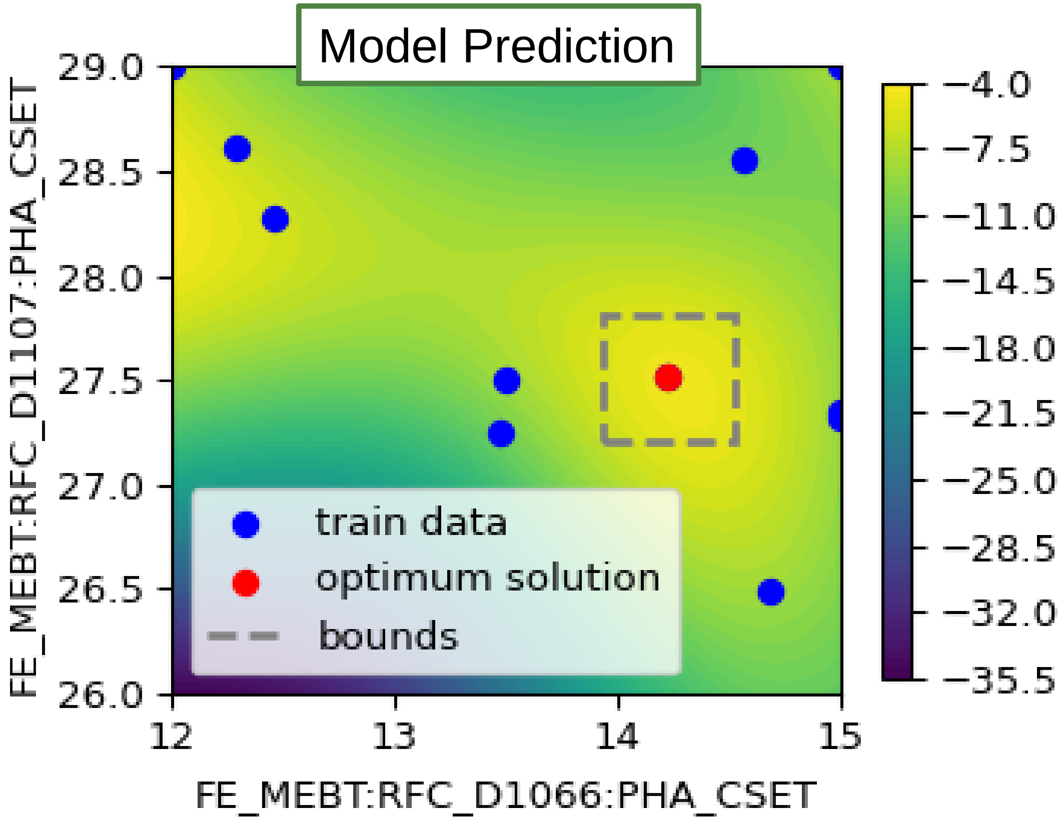
- Tightly avoid accelerator idle time, threaded computation
  - Simultaneously queries candidate solutions while the beam measurements are ongoing
  - Candidate search terminates upon objective evaluation completion by machine
- Resetting time awareness for max beam time utilization
  - Particularly, power supply of corrector polarity change.
- Order evaluation candidates (initialization points or multi-batch query) to minimize resetting cost



Customized BO: Global

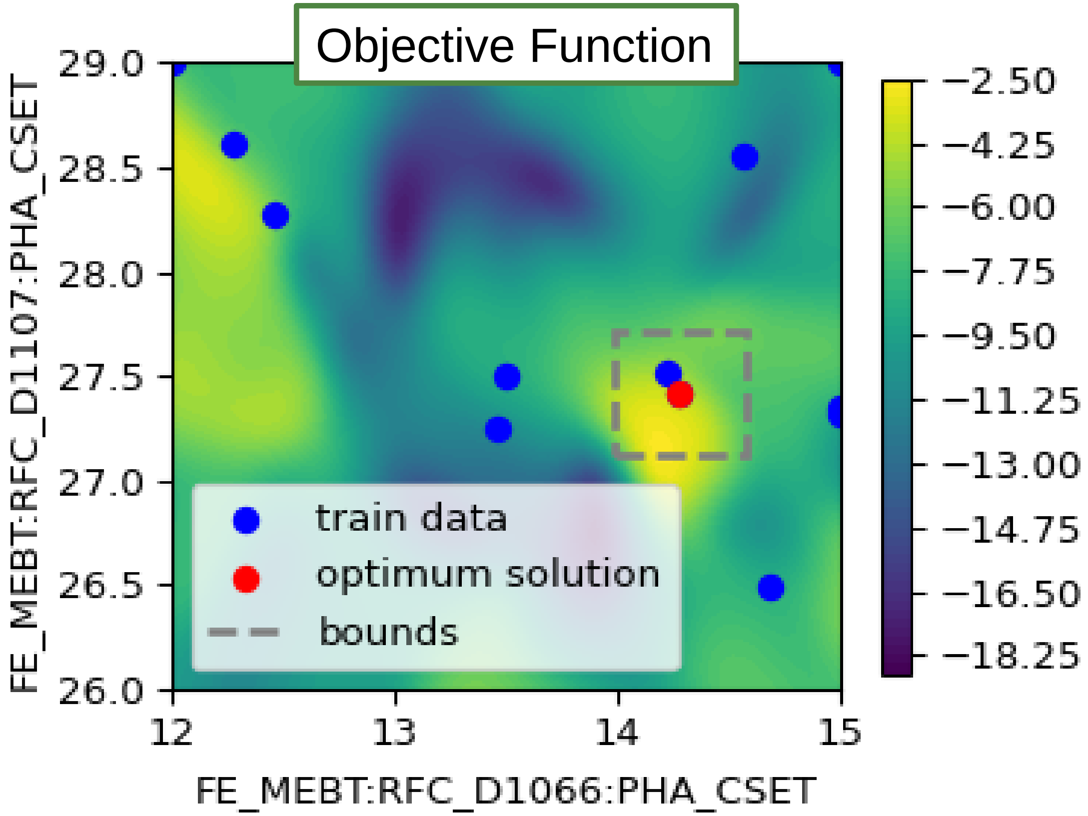
- Transit from Global (over whole control domain) to Local (over narrow domain) optimization
  - GlobalBO in a limited beam-time remains challenging ⇒ Completion through LocalBO
- Incorporate archived data or simulation model
  - Through prior mean assisted BO (pmBO)

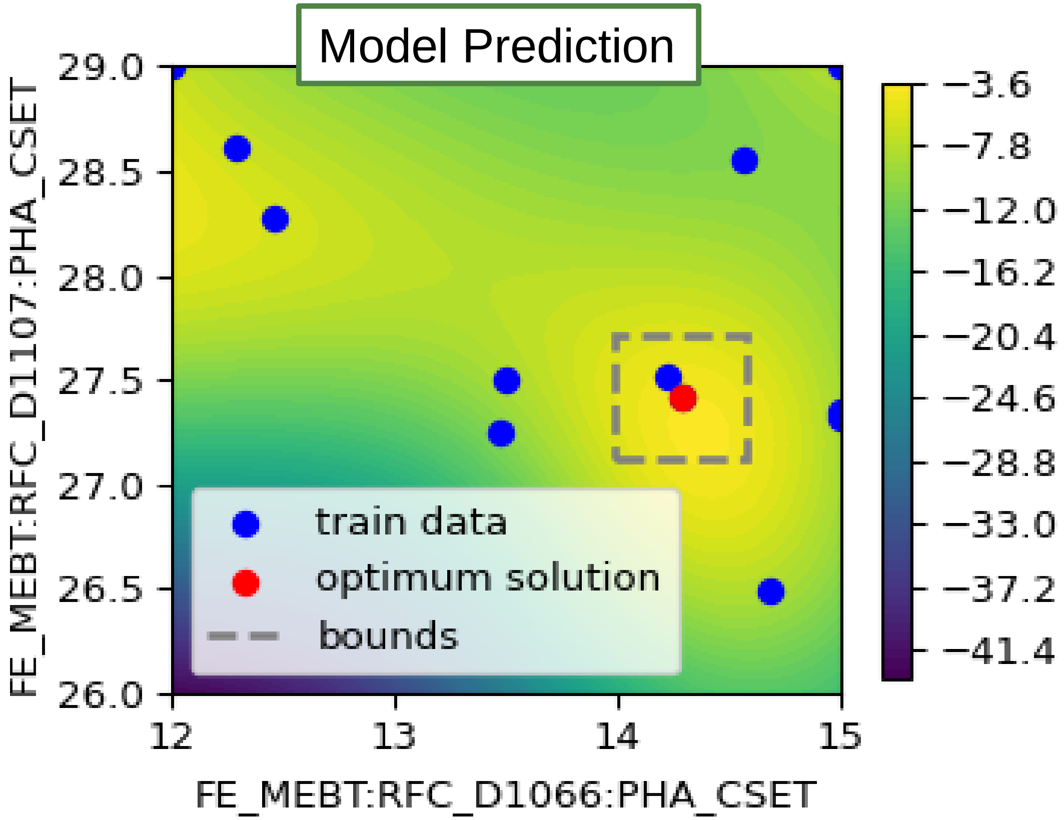




Customized BO: Global to Local

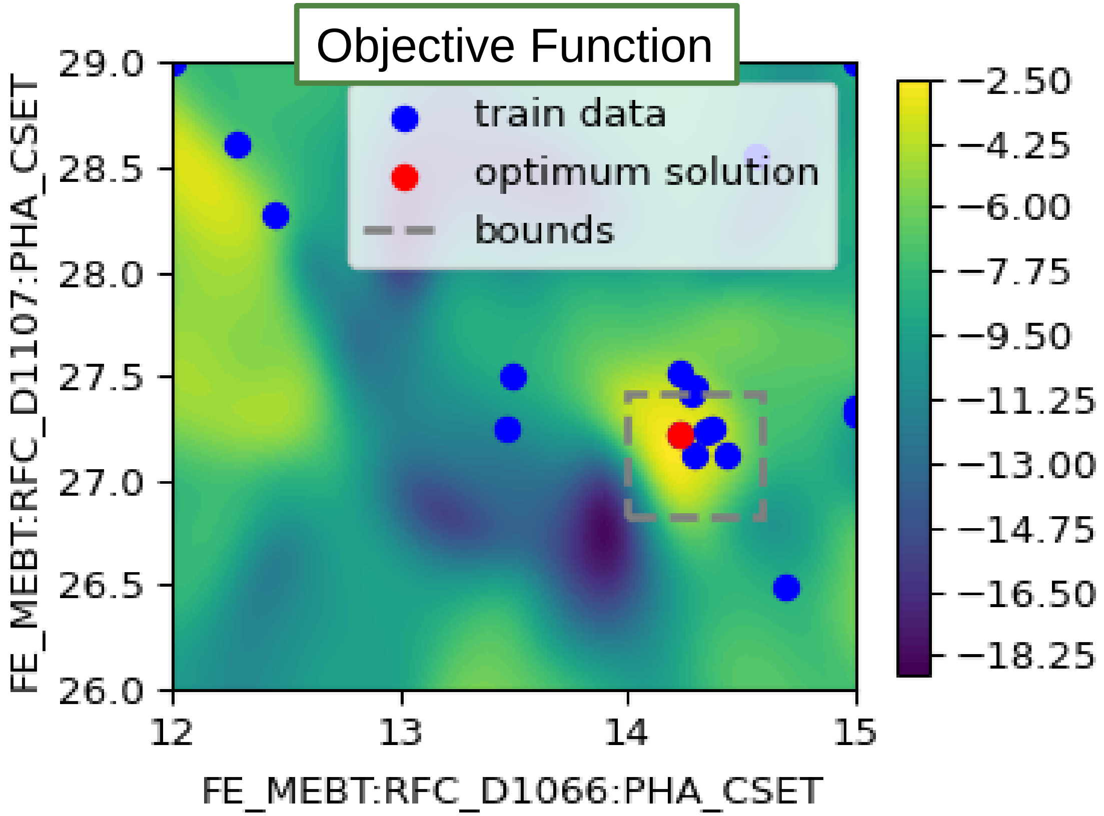
- Transit from Global (over whole control domain) to Local (over narrow domain) optimization
  - GlobalBO in a limited beam-time remains challenging ⇒ Completion through LocalBO
- Incorporate archived data or simulation model
  - Through prior mean assisted BO (pmBO)

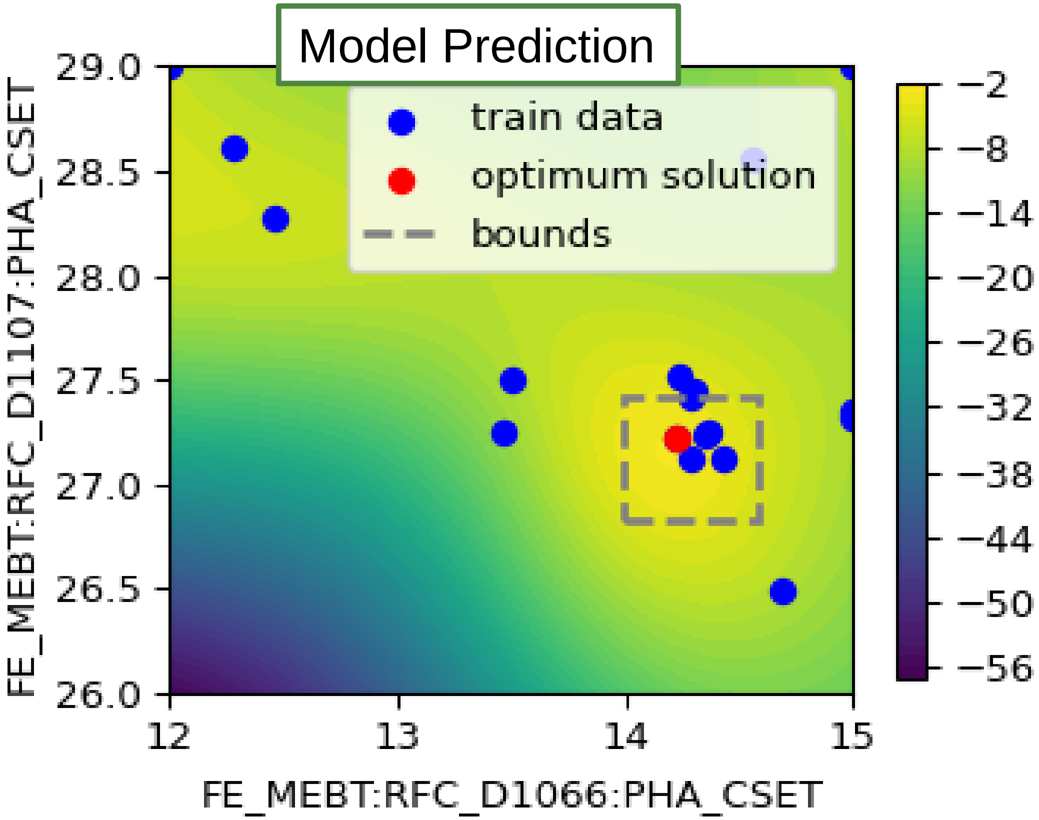




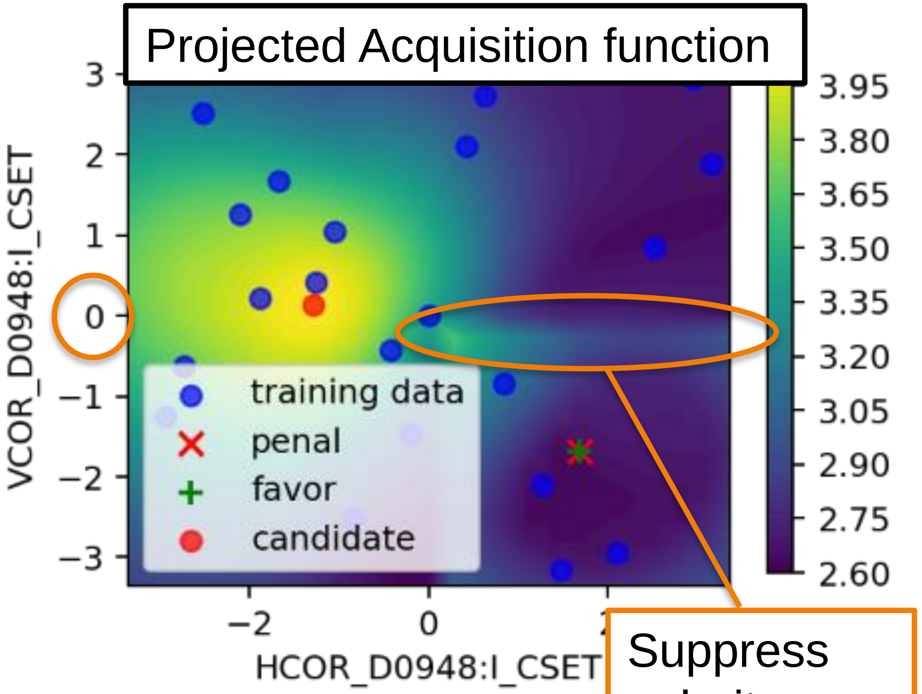
Customized BO: Global to Local

- Transit from Global (over whole control domain) to Local (over narrow domain) optimization
  - GlobalBO in a limited beam-time remains challenging ⇒ Completion through LocalBO
- Incorporate archived data or simulation model
  - Through prior mean assisted BO (pmBO)





Customized BO: Maximize Transmission through Two Apertures



Suppress polarity change by favoring current quadrant

- Tightly avoid machine idle time
  - Simultaneously queries candidate solutions while the machine evaluates objectives.
  - Candidate search terminates upon objective evaluation completion by machine
- Ramping time awareness for max beam time utilization
  - Particularly, power supply of corrector polarity change.
- Order evaluation candidates (initialization points or multi-batch query) to minimize ramping cost

$$f_{penal} = -C_{penal} e^{-(x-x_{penal})^2/L_{penal}^2}$$

$$f_{favor} = +C_{favor} e^{-(x-x_{favor})^2/L_{favor}^2}$$

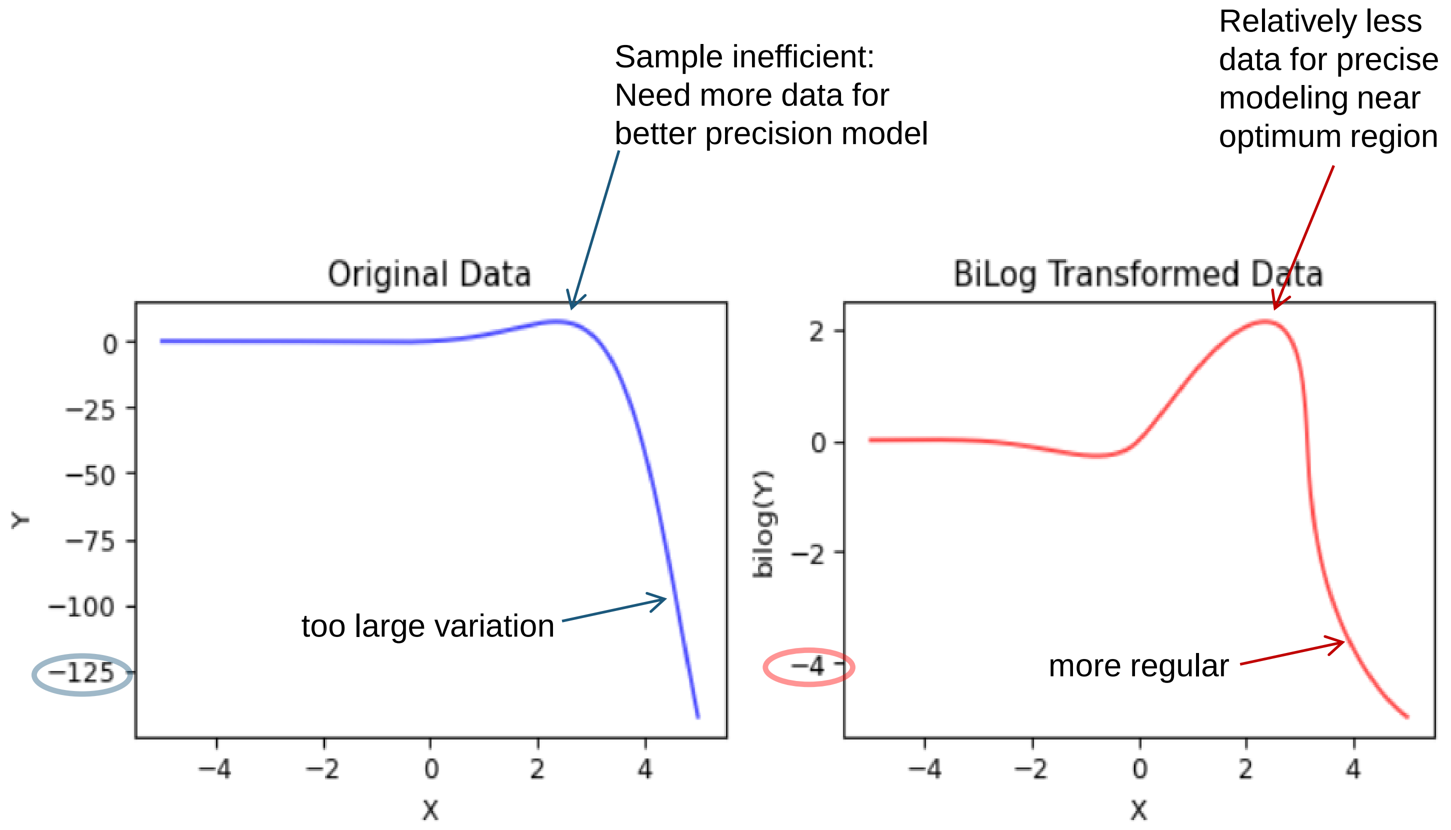
$$f_{polarity} = \begin{cases} +C_{polaity} & \text{if } sign(x) = sign(x_{current}) \\ 0 & \text{else} \end{cases}$$

Objective Function Construction

- Regularity of the objective function strongly effect optimization performance
- Objective function construction template saves beam time!
- We implemented intuitive scalarization of multiple objectives.

$$Obj = \sum_i w_i \varphi\left(\frac{v_{i,meas} - v_{i,target}}{tolerance_i}\right)$$

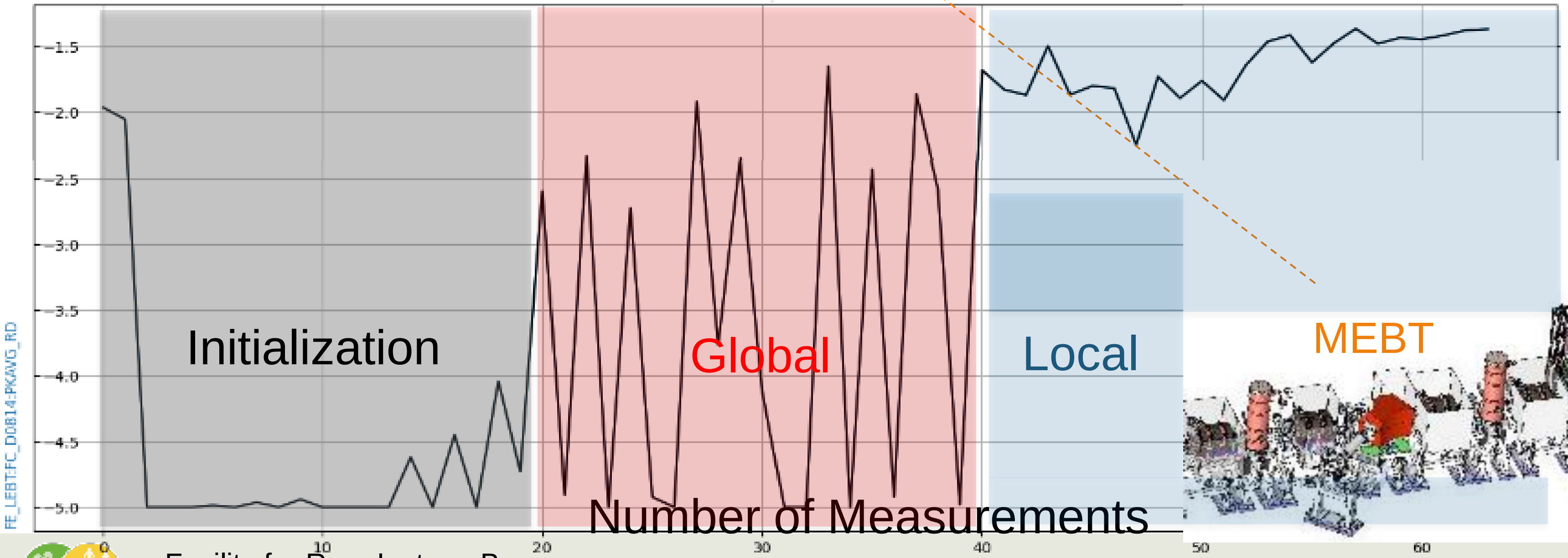
$$Obj_{bilog} = \mathrm{sign}(Obj)\log(1 + |Obj|)$$

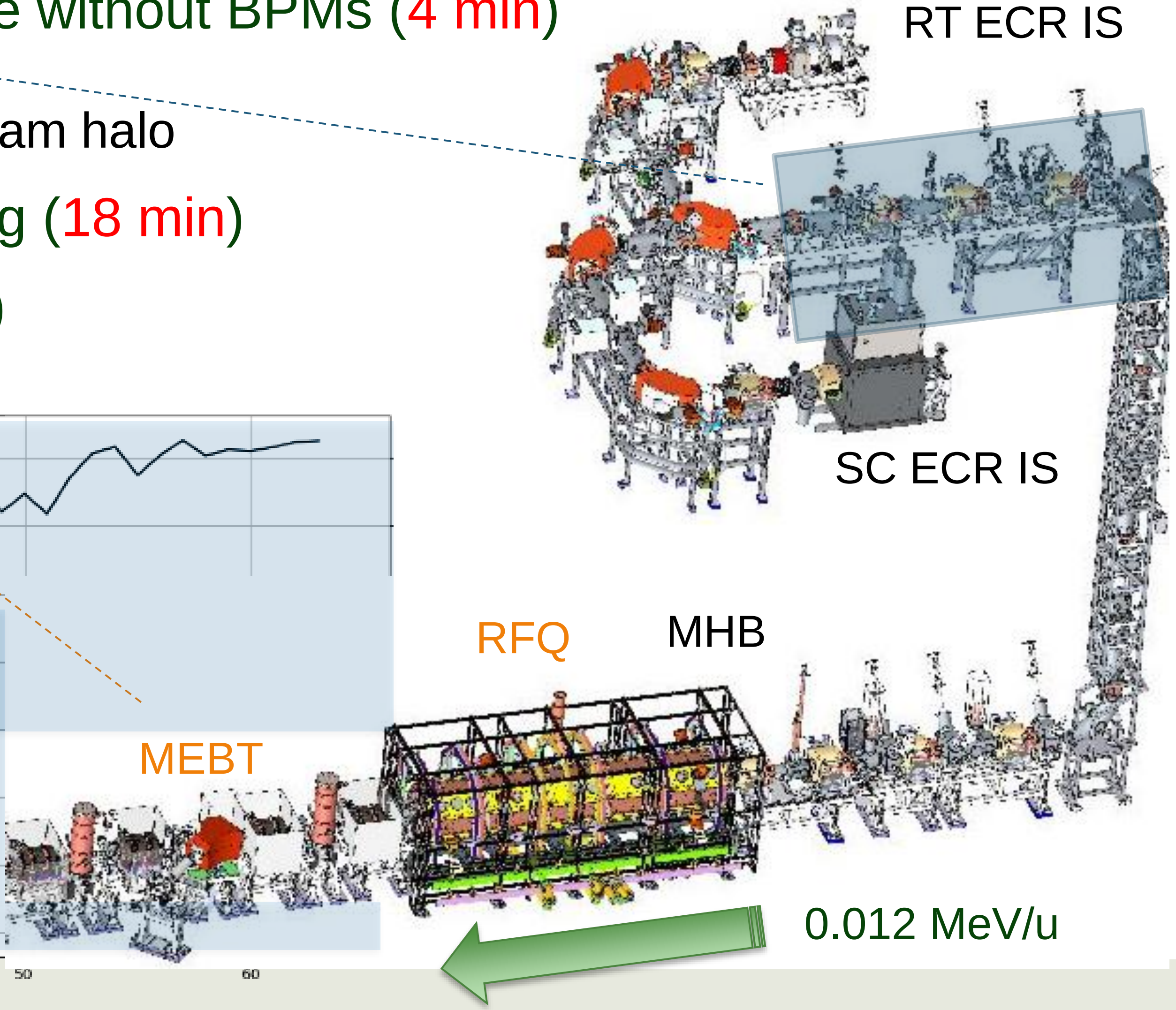


Front End Tuning

- Steer beam through two apertures in straight line without BPMs (4 min)
  - Limited fast beam centroid diagnostics.
  - Maximize beam current through apertures ⇒ cut beam halo
- RFQ transmission and MEBT 6D beam centering (18 min)
- LS1 longitudinal beam center restoration (2 min)

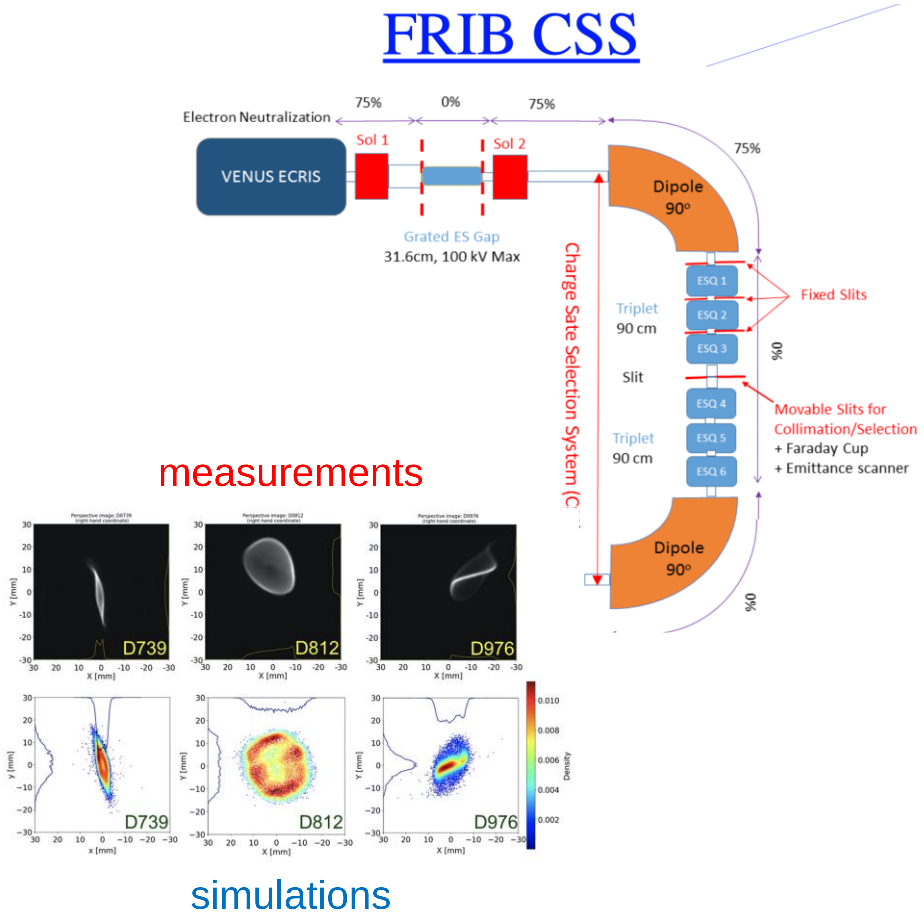
Objective of FC reading after two apertures





Motivation: Accelerators are complex time-varying systems (especially FRIB!)

- Collective effects: space charge forces.



Dataset for AI training summary

13 beam species

Charge states:
26, 22, 23, 24, 25, 27, 28, 1, 2, 3, 4, 5, 6

Ion mass numbers:
124, 124, 124, 124, 124, 124, 124, 16, 16, 16, 16, 16, 16

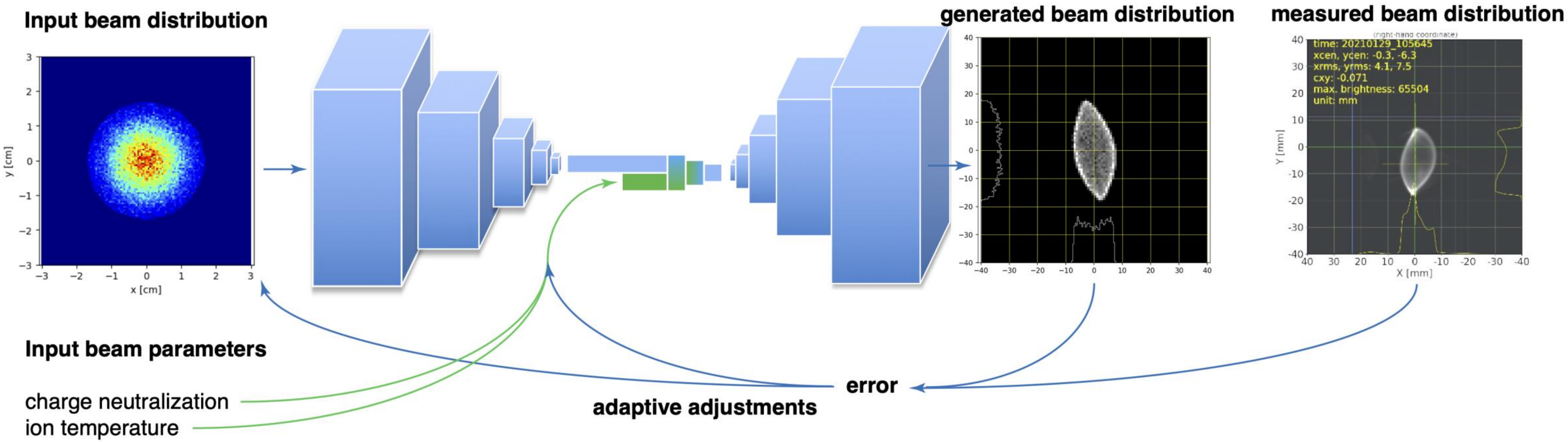
Macroparticles per charge state: 300 K

Total macroparticles per simulation: 3.9 M

Time for 1 simulation, using 96 cores: ~6 hours

Total simulations run: ~420

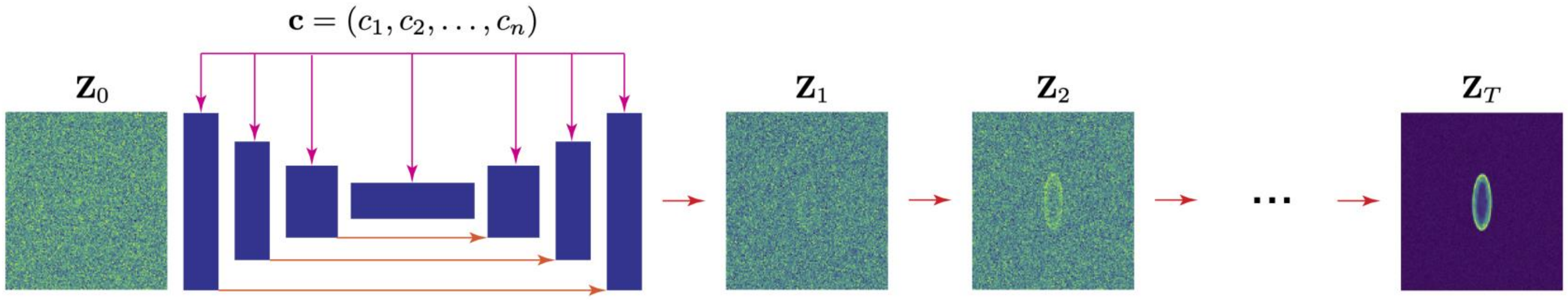
Goal: Create Adaptive Generative Deep Learning Tools for Virtual 4D Phase Space Diagnostics



Conditional Guided Diffusion for Creating 4D Phase Space

- Generative diffusion is the state-of-the-art AI-based method for creating high resolution representations of complex objects, such as all 6 unique projections of the FRIB beam's 4D phase space (assuming a beam uniform in z with little/no energy spread).
- Conditional vector *c* contains lattice parameters (such as magnet settings), which of the beam's 2D phase space projections to generate, and where along the lattice to generate it.

Sample from mean 0, unit variance gaussian noise distribution, $\mathbf{Z}_0 \sim N(\mathbf{0},\mathbf{I})$



Conditionally guided iterative generative Unet utilizing attention mechanisms at lower-dimensional representations.

Iterative generative process.

Conditional Guided Diffusion for Creating 4D Phase Space

Iterative Steps

$$\mathbf{z}_1 = \sqrt{1-\beta_1}\mathbf{Y} + \sqrt{\beta_1}\varepsilon_1, \quad \mathbf{z}_2 = \sqrt{1-\beta_2}\mathbf{z}_1 + \sqrt{\beta_2}\varepsilon_2$$

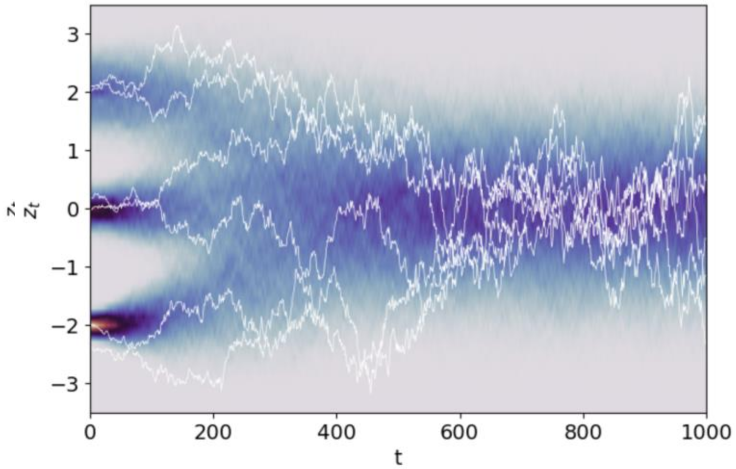
$$\vdots$$

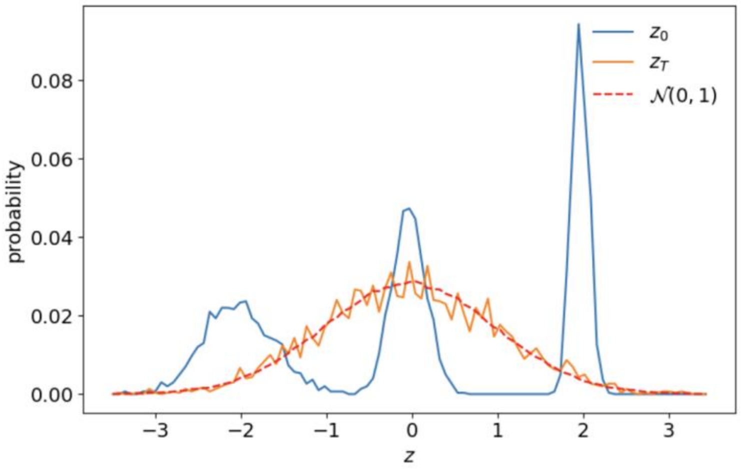
$$\mathbf{z}_t = \sqrt{1-\beta_t}\mathbf{z}_{t-1} + \sqrt{\beta_t}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \quad \beta_t \in [0,1], \quad t \in \{1,2,\ldots,T\}$$

Probability Distributions of a Markov Chain

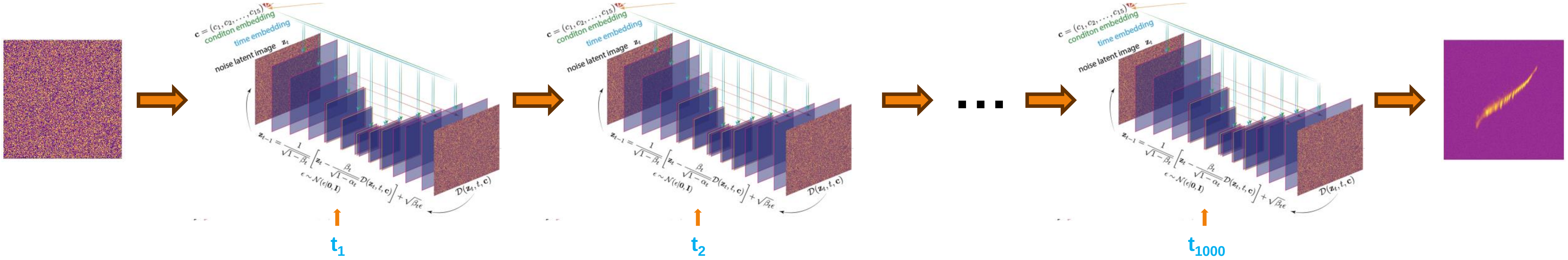
$$q(\mathbf{z}_1|\mathbf{Y}) = \mathcal{N}_{\mathbf{z}_1}\left[\sqrt{1-\beta_1}\mathbf{Y}, \beta_1\mathbf{I}\right], \quad q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}_{\mathbf{z}_t}\left[\sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right] \qquad q(\mathbf{z}_{1\ldots T}|\mathbf{Y}) = q(\mathbf{z}_1|\mathbf{Y})\prod_{t=2}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1})$$

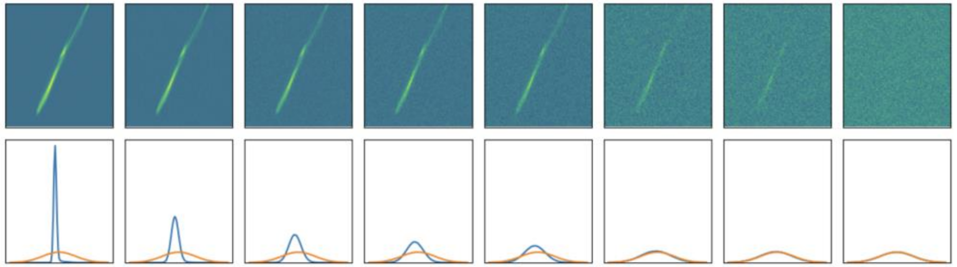
$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{Y} + \sqrt{1-\alpha_t}\varepsilon, \quad \alpha_t = \prod_{s=1}^{t}(1-\beta_s) \quad \Longrightarrow \quad q(\mathbf{z}_t|\mathbf{Y}) = \mathcal{N}_{\mathbf{z}_t}\left[\sqrt{\alpha_t}\mathbf{Y}, (1-\alpha_t)\mathbf{I}\right] \quad \rightarrow \quad \mathcal{N}(0,1)$$





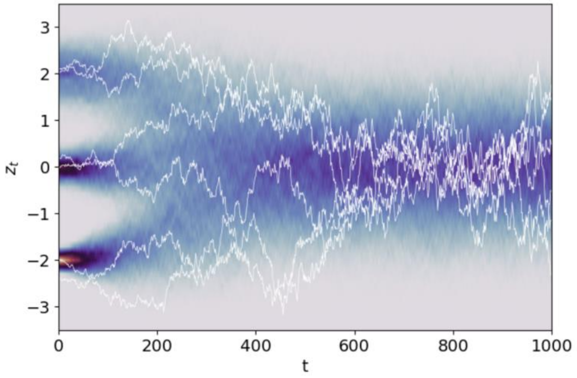
Conditional Guided Diffusion for Creating 4D Phase Space





$$d\mathbf{z} = \mathbf{f}(\mathbf{z}, t)dt + g(t)d\mathbf{w}, \quad \text{SDE}$$

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

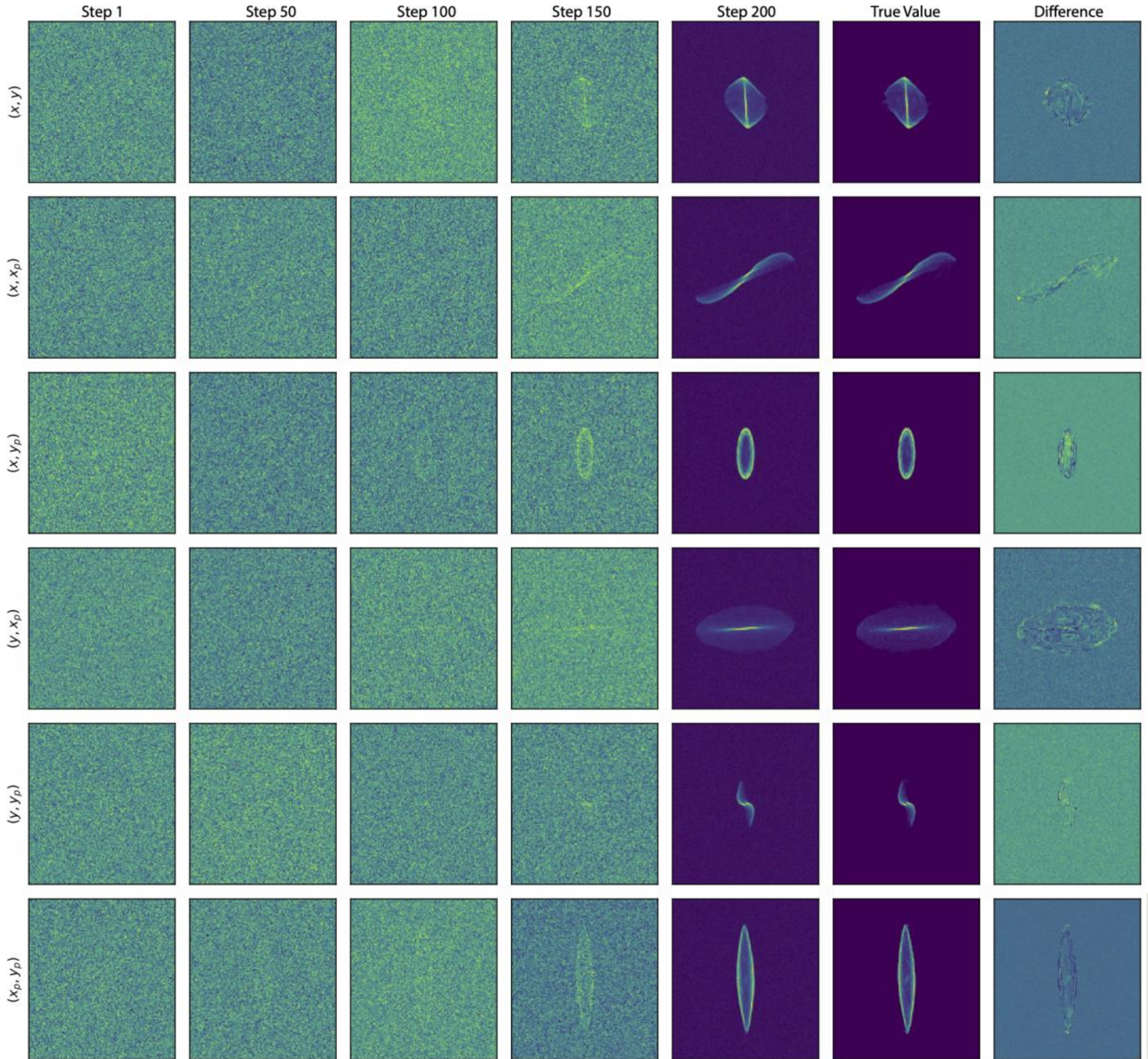


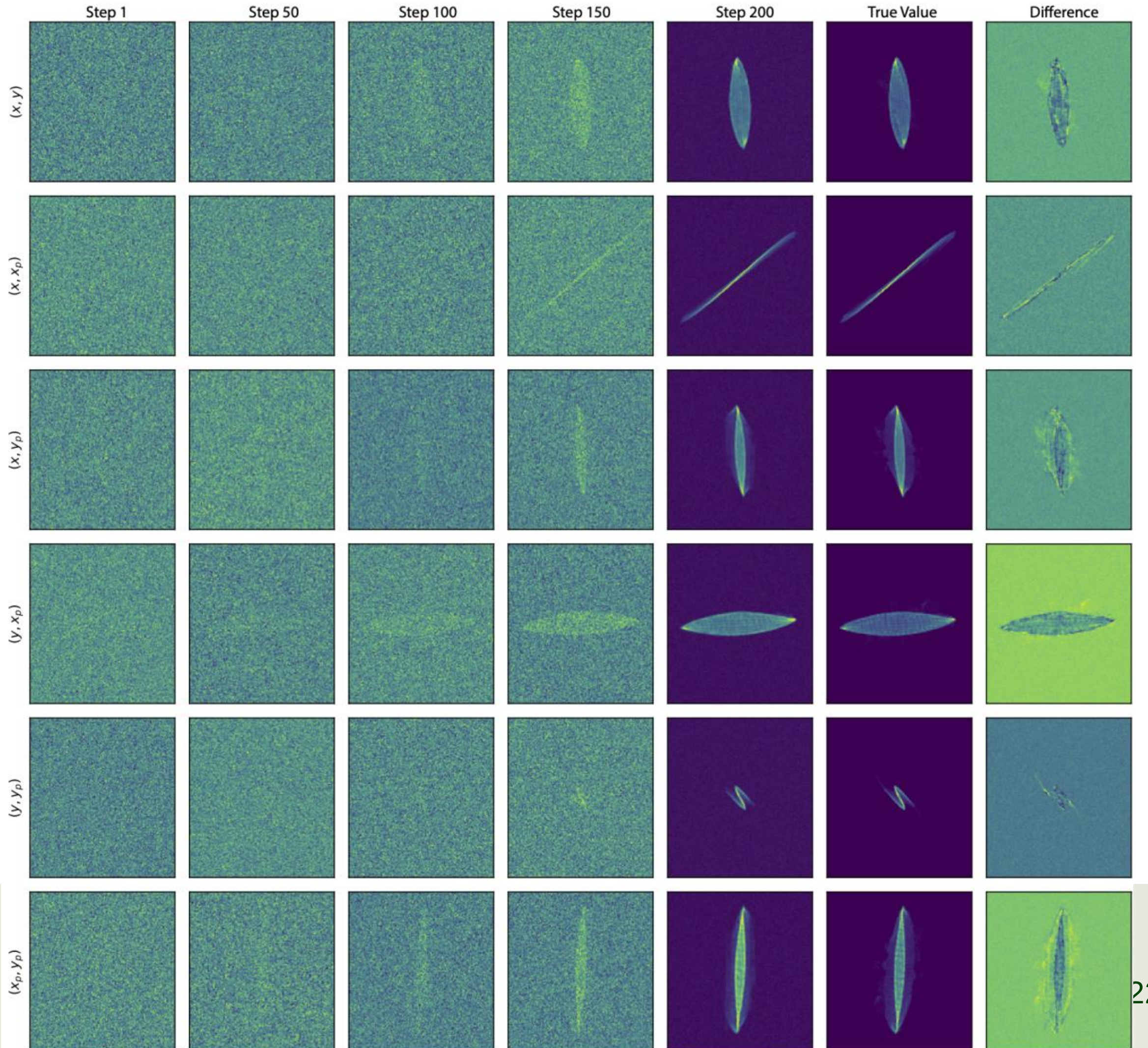
$$i\hbar \frac{\partial}{\partial t}\Psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m}\nabla^2\Psi(\mathbf{r}, t) + V(\mathbf{r})\Psi(\mathbf{r}, t)$$

infinite?

Preliminary Results: Conditional Guided Diffusion for Creating 4D Phase Space

- Generative diffusion is the state-of-the-art AI-based method for creating high resolution representations of complex objects, such as all 6 unique projections of the FRIB beam’s 4D phase space (assuming a beam uniform in z with little/no energy spread).

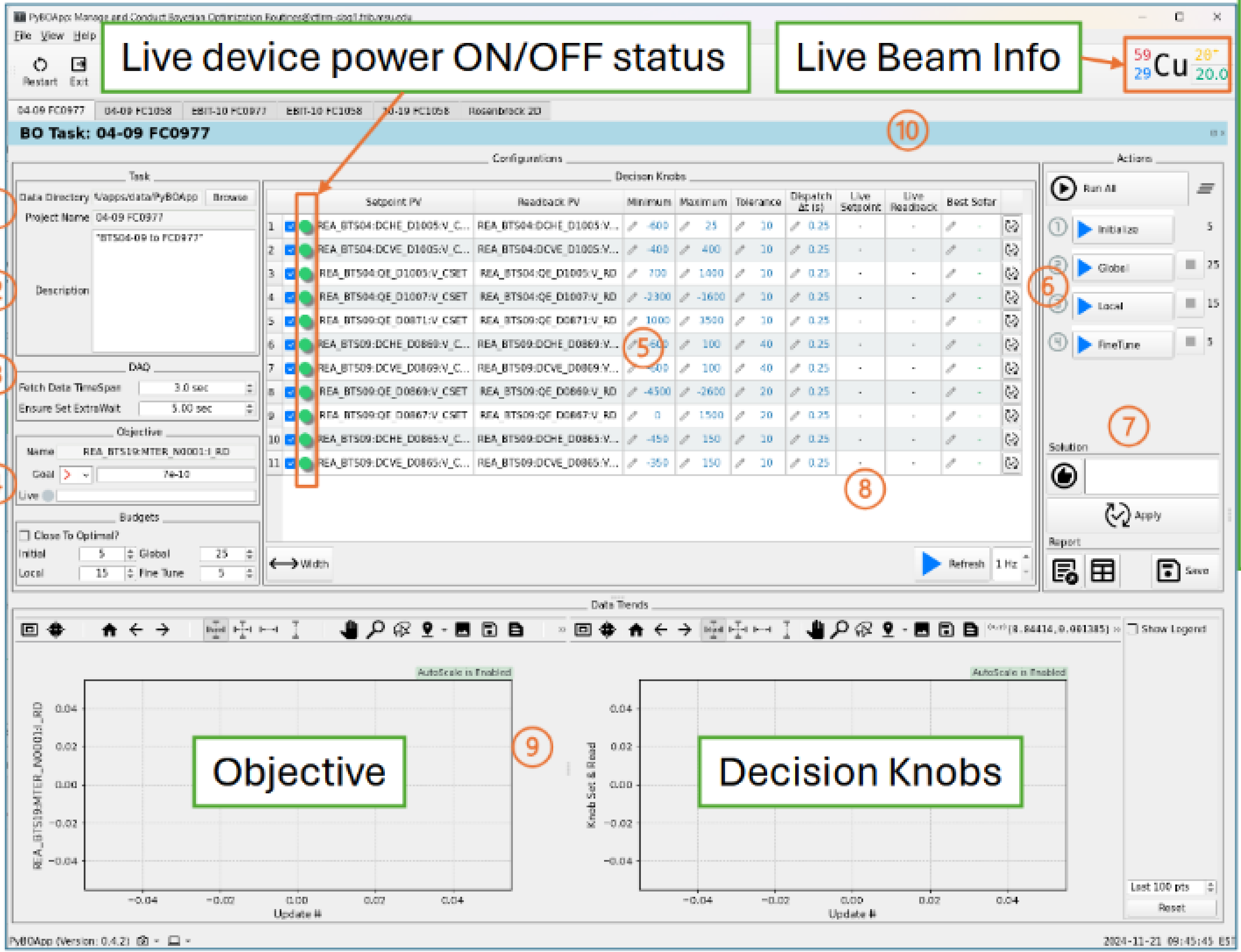




General GUI App for Bayesian Optimization

- PyBOApp is a general GUI app built upon the PHANTASY (Virtual Accelerator) framework, specifically by using its UI components and libraries.
- The App is integrated into App Launcher, an app manages the physics applications for the accelerator commissioning and operations.
- Manages Bayesian Optimization (BO) tasks with a configuration file.
- Command line interface tools are provided for testing and development.
- Debian packages were created and deployed in the Linux system

PyBOApp: User Interface and Interactions

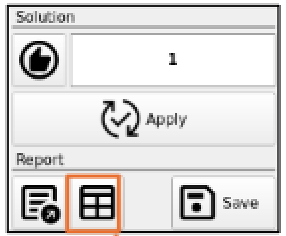


Each task is presented as a clickable tab

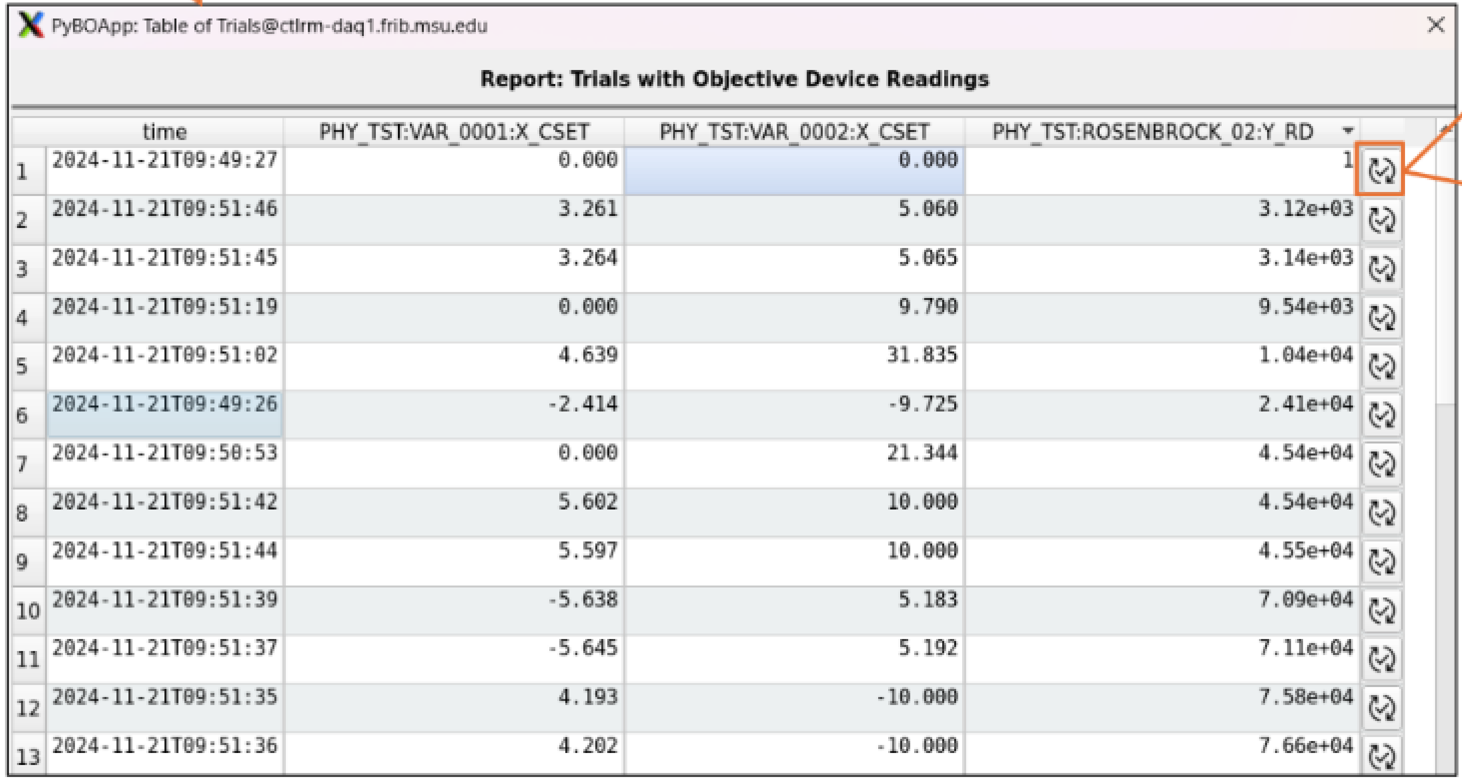
1. Defines where the data files to be saved
2. Defines how the data is set and get
3. Defines the objective
4. Defines the budgets
5. Defines the decision knobs
6. Run the BO routines all or stage by stage
7. Get the solution, task report, save the data
8. Activate the data/figure updating
9. Data visualization
10. Undock by double-clicking

- Executing the task in a single button click [Run All]
- Or, iterating each stage individually [Initialize] [Global] [Local] [FineTune]
- Support Abort the running

PyBOApp: Solution Tracking and Applying

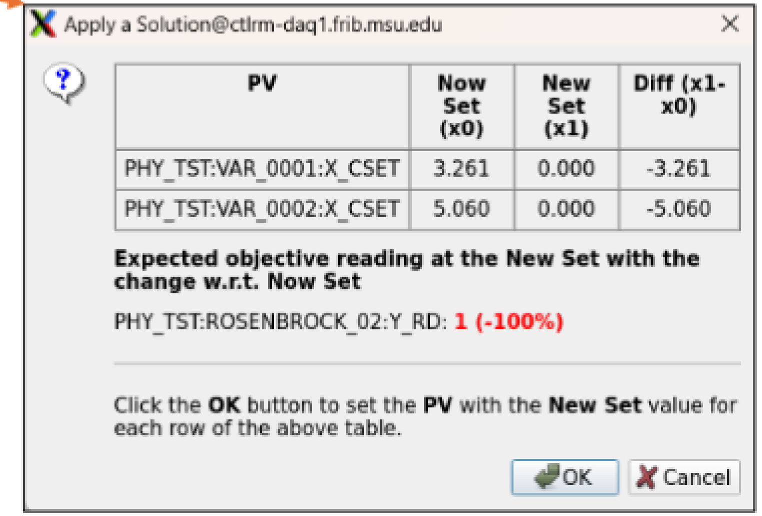


- The app keeps all the solutions during the BO run in tabular format, one can click to read the table
- The table supports sorting, e.g. on the objective device readings
- Each row of solution could be applied to the system by clicking the button
    - The pop-up dialog indicates how much objective change is expected
    - Press OK to apply the changes



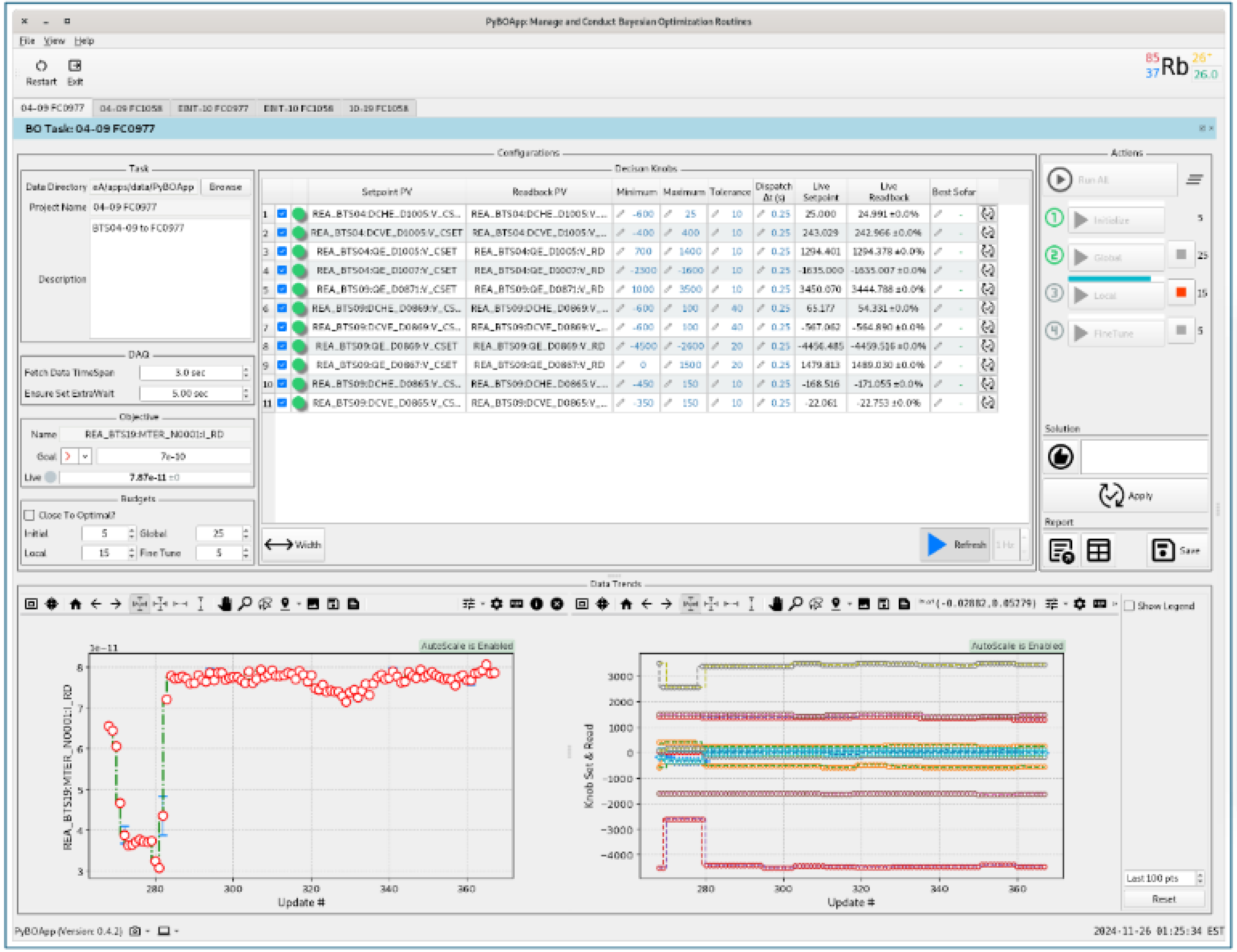
	time	PHY_TST:VAR_0001:X_CSET	PHY_TST:VAR_0002:X_CSET	PHY_TST:ROSENBROCK_02:Y_RD
1	2024-11-21T09:49:27	0.000	0.000	1
2	2024-11-21T09:51:46	3.261	5.060	3.12e+03
3	2024-11-21T09:51:45	3.264	5.065	3.14e+03
4	2024-11-21T09:51:19	0.000	9.790	9.54e+03
5	2024-11-21T09:51:02	4.639	31.835	1.04e+04
6	2024-11-21T09:49:26	-2.414	-9.725	2.41e+04
7	2024-11-21T09:50:53	0.000	21.344	4.54e+04
8	2024-11-21T09:51:42	5.602	10.000	4.54e+04
9	2024-11-21T09:51:44	5.597	10.000	4.55e+04
10	2024-11-21T09:51:39	-5.638	5.183	7.09e+04
11	2024-11-21T09:51:37	-5.645	5.192	7.11e+04
12	2024-11-21T09:51:35	4.193	-10.000	7.58e+04
13	2024-11-21T09:51:36	4.202	-10.000	7.66e+04

Since it (1) is the minimal (best), it is equivalent as click Apply, which applies the best-so-far solution.

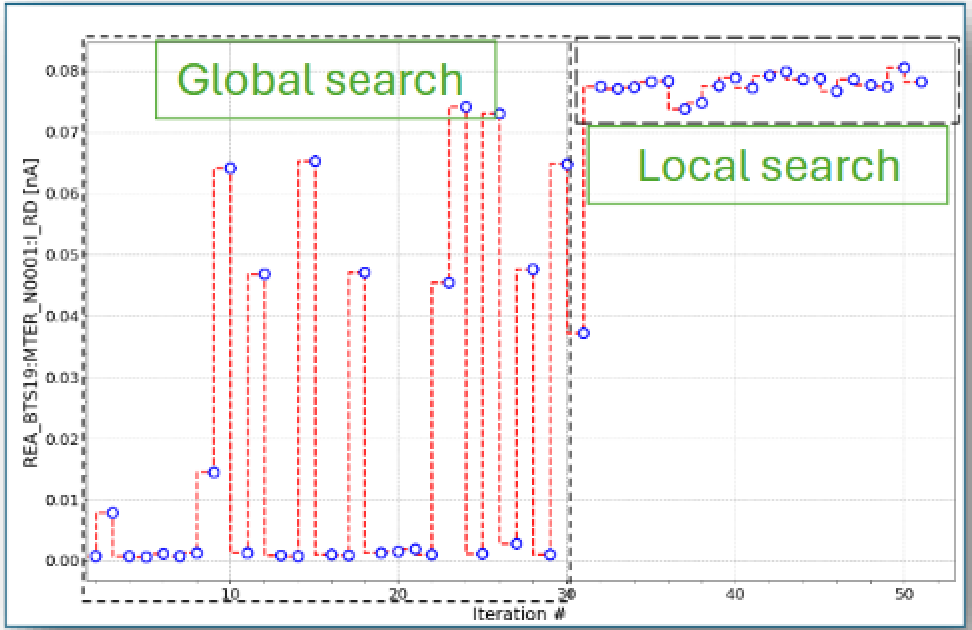


PV	Now Set (x0)	New Set (x1)	Diff (x1-x0)
PHY_TST:VAR_0001:X_CSET	3.261	0.000	-3.261
PHY_TST:VAR_0002:X_CSET	5.060	0.000	-5.060

Application on Re-Accelerator: Source Tuning



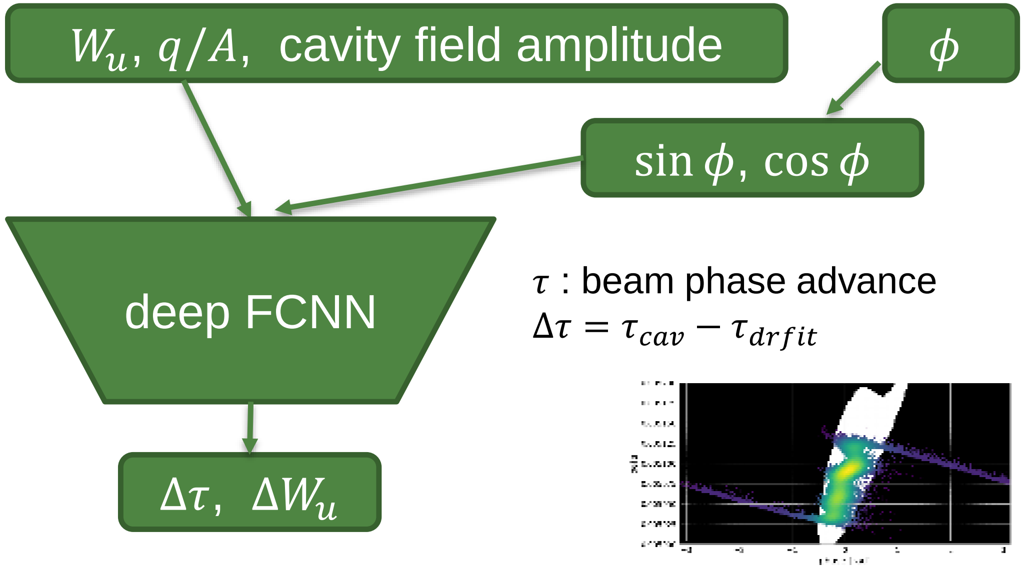
Goal: Maximize a Faraday Cup reading (transmission) by tuning 3 pairs of correctors (horizontal & vertical) and 5 quadrupoles, that is 11 devices in total in the upstream beamline.



Iterating the Global, Local, FineTune run stages to find the optimal solutions

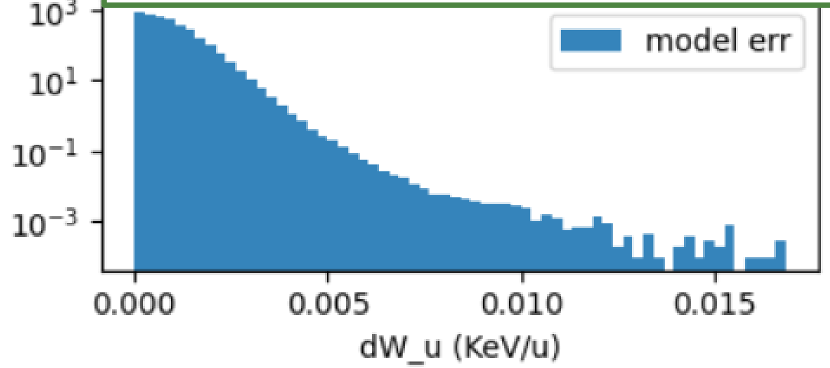
Fast Physics Model is essential for Online Optimization

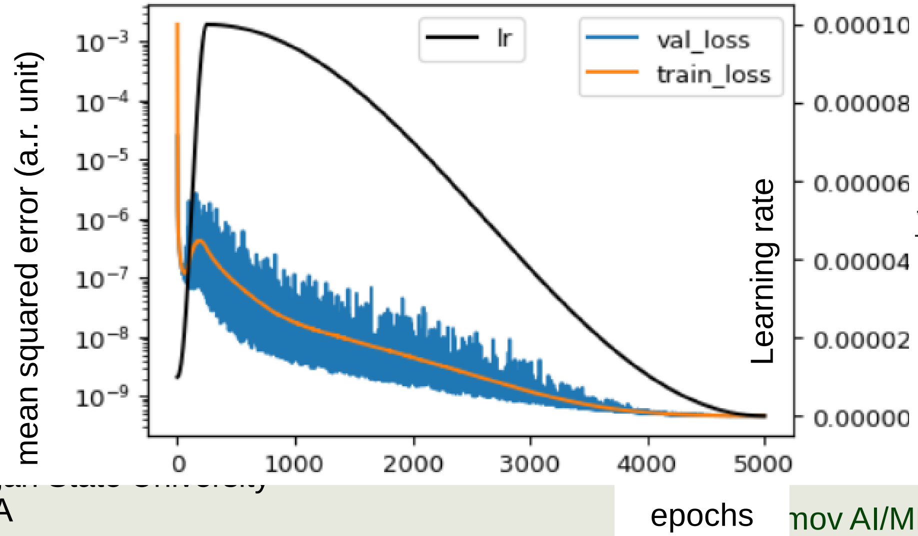
- BO can exploit physics model for online beam tuning provided the model is fast enough
  - Otherwise, constructing online model through online data is easier and faster
- Development of rapid surrogate models of physics simulators is underway.
  - surrogate modeling of 1D longitudinal RF cavity simulator achieved good accuracy and speed (30x faster ~ 100 us)



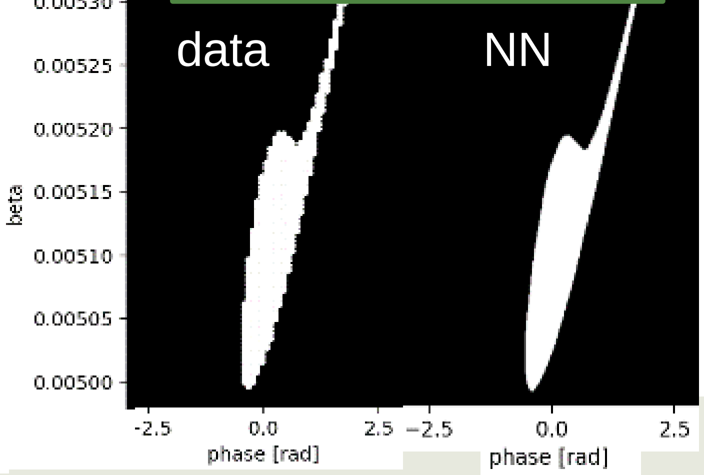
τ : beam phase advance
$\Delta\tau = \tau_{cav} - \tau_{drfit}$

Histogram of surrogate model (NN) prediction error of beam energy output over test data





RFQ acceptance surrogate

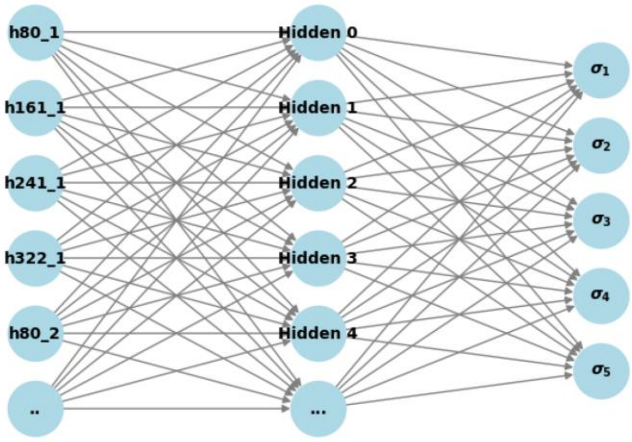


Bunch Length from Beam Position Monitor (BPM)

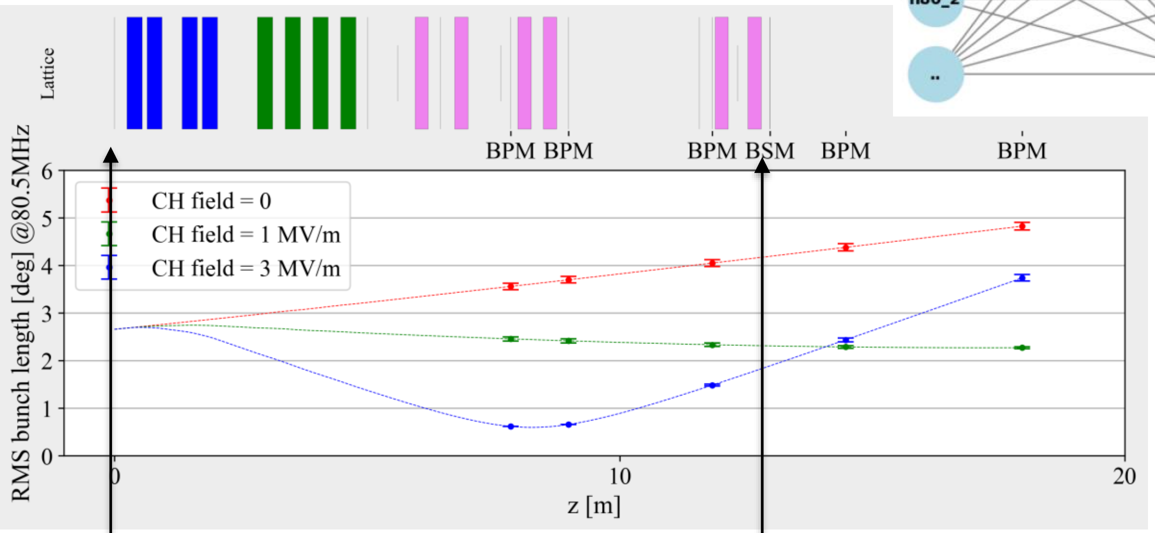
TUPB006
Jinyu Wan

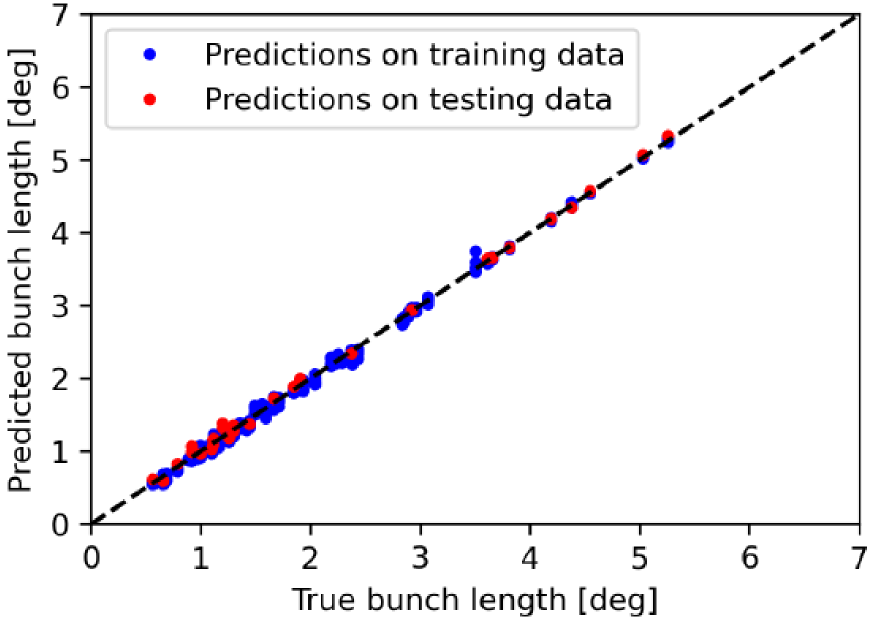
A neural network is trained to learn the correlation of the BPM harmonics and the bunch lengths

Input:
BPM
harmonics



Output:
bunch lengths at the
BPM locations





Fitted beam distribution

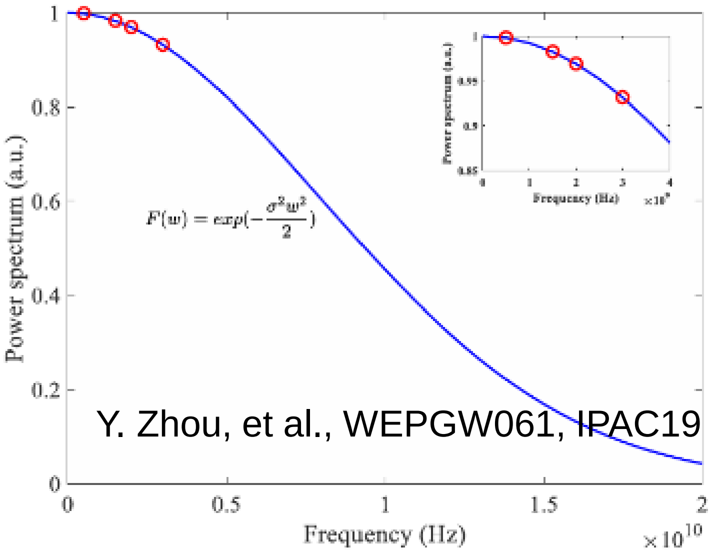
Measurement location

BPM Harmonics

For a Gaussian beam, the BPM harmonics are determined by

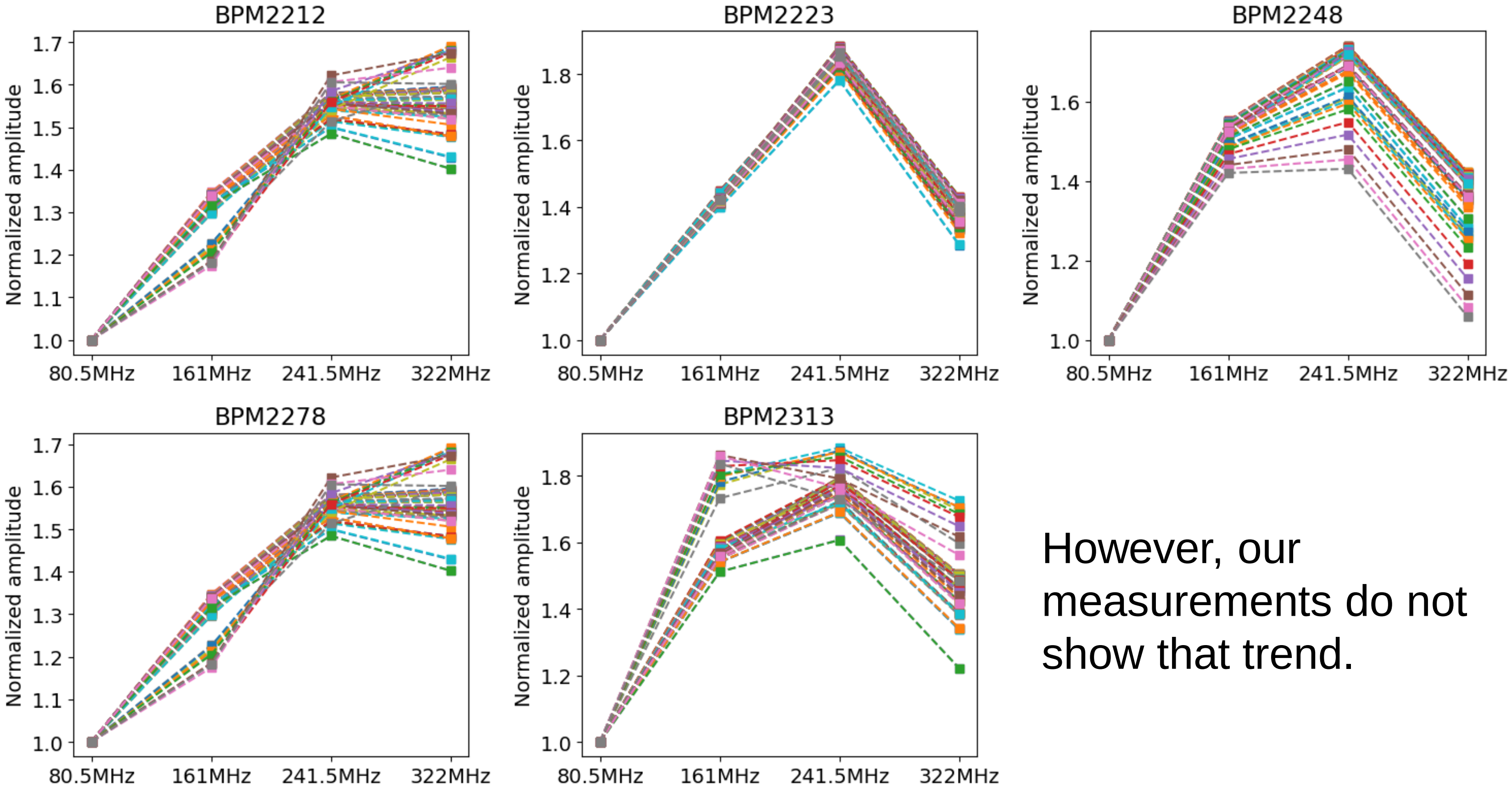
$$A_n = A_0 \exp\left(-\frac{(\omega_n \sigma_t)^2}{2}\right)$$

$$\frac{-\ln\left(\frac{A_n}{A_m}\right)}{(n^2 - m^2)2\pi^2 f_0^2} = \sigma_t^2$$



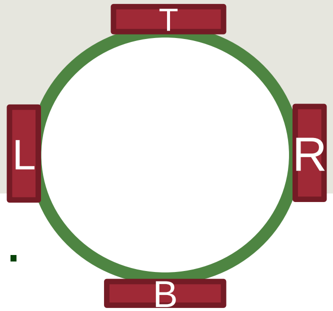
Y. Zhou, et al., WEPGW061, IPAC19

Figure 1: Power spectrum for Gaussian bunches.



However, our measurements do not show that trend.

Beam Quadrupole Moment from BPM



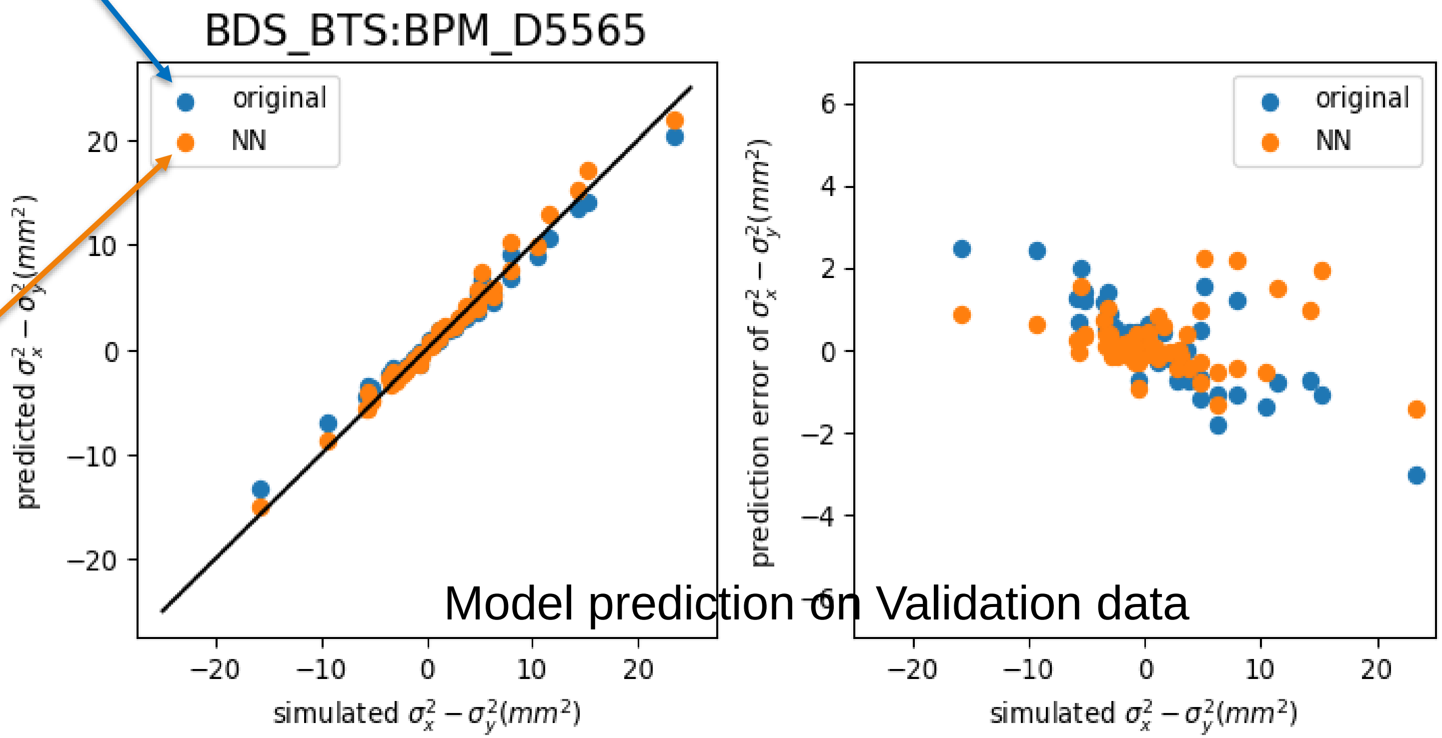
- BPM consists of 4 pick-ups to measure beam-induced signals U_R, U_L, U_T, U_B .

geometric factor

$$BPMQ = G\frac{U_R + U_L - (U_T + U_B)}{U_R + U_L + U_T + U_B} - (x^2 - y^2) \simeq \sigma_x^2 - \sigma_y^2$$

- BPM is designed to measure beam position, not the quadrupole moments
- Quadrupole moment signal strength is the 2nd order → Weak signal
  - Small signal error (calibration error, noises) leads to big BPMQ error
- Use NN to calibrate. Accuracy of data is the key for accurate NN model

$$\sigma_x^2 - \sigma_y^2 = \mathrm{NN}(U_R, U_L, U_T, U_B)$$

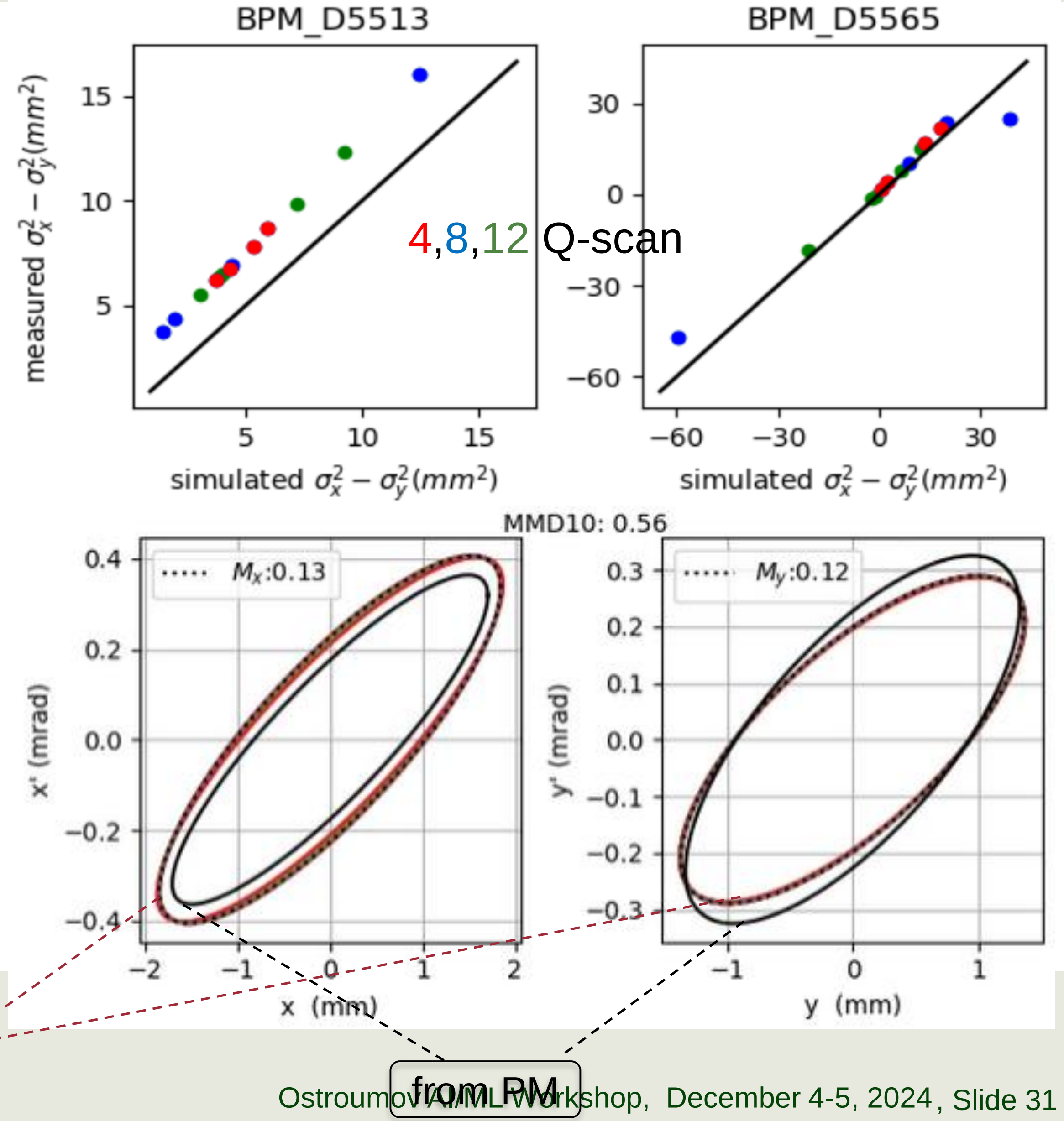


Model prediction on Validation data

feature engineering, outlier rejection, etc

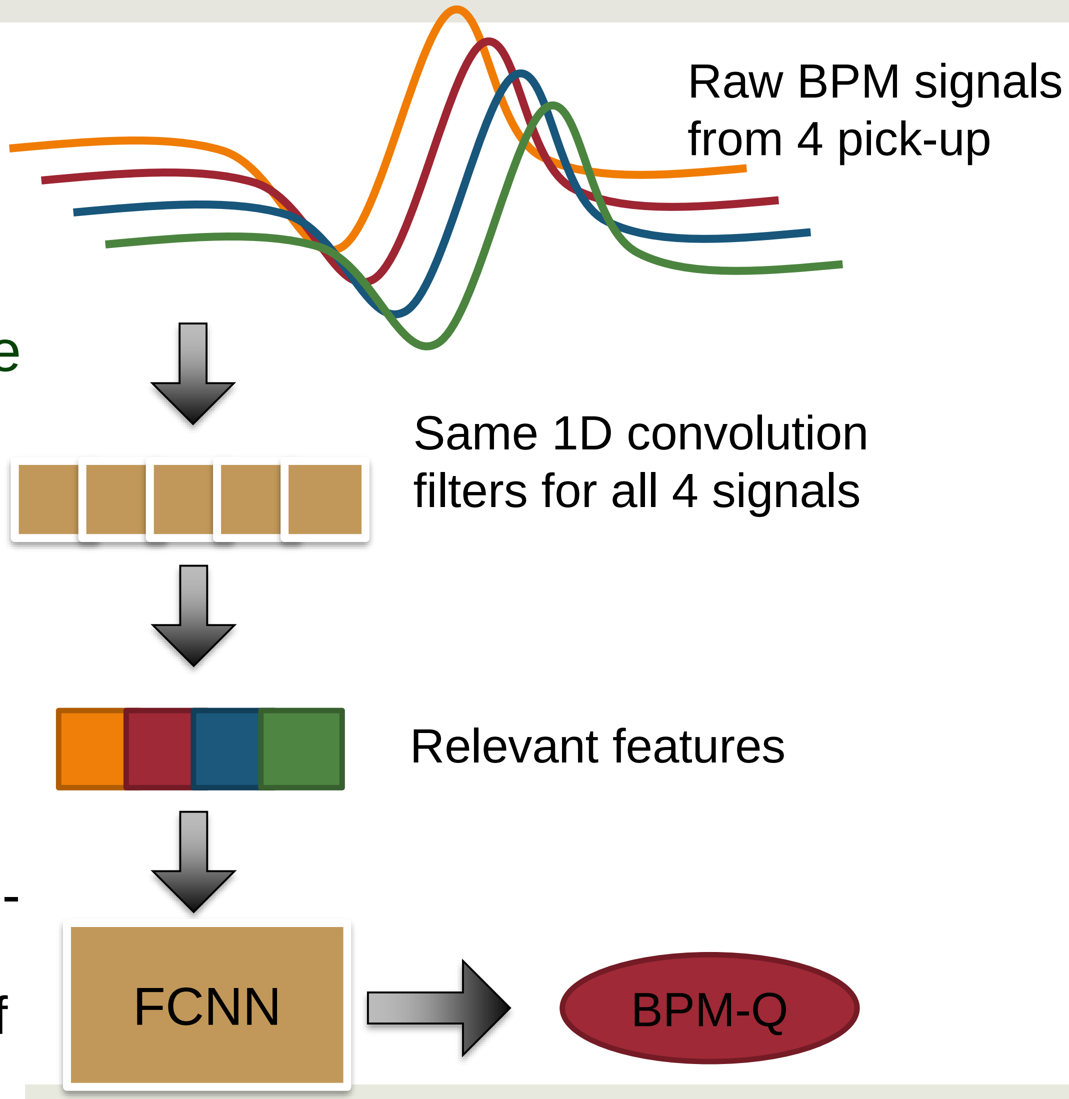
Courant-Snyder (CS) Parameters Reconstruction from BPMQ Using Bayesian Active Learning (BAL)

- CS parameter reconstruction using beam quadrupole moments at a few BPMs requires efficient quadrupole magnet scan.
- Ensemble of a backward-differentiable envelope simulation (with pyTorch) for surrogate model of BAL
  - Candidate quadrupole settings are queried to maximize surrogate model uncertainty of BPMQs.
- Beam test: $^{16}O^{6+}$ ion beam (with 12 epoch of q-Scan).
  - Despite the model prediction shows some error
    - » NN model for BPM_D5565 shows better accuracy
      - training data is more reliable due to PM5567
  - Good enough agreement between reconstructed CS using BPMQ with BAL vs PM.
    - » with BPM5565 and emittance prior



Recent Development: Improved BPM-Q Modeling

- Amplitude of the 2nd harmonic (161MHz) of beam frequency of induced signal at BPM 4-pickups are utilized for beam position and quadrupole moment to avoid cross-talk from many other RF devices.
- Instead of the manually engineered input feature – the 161MHz component, we use NN to extract the features relevant to the BPM-Q from the full raw signal in time domain:
  - To mitigate overfitting caused by the high-dimensional input data and a relatively small training dataset, we use an encoder architecture composed of 1D convolutional layers. These layers serve to extract meaningful features by reducing input dimensionality, followed by a Fully Connected Neural Network (FCNN) that outputs the BPM-Q.
  - This resulted in about 25% improved accuracy in terms of BPM-Q error on validation data

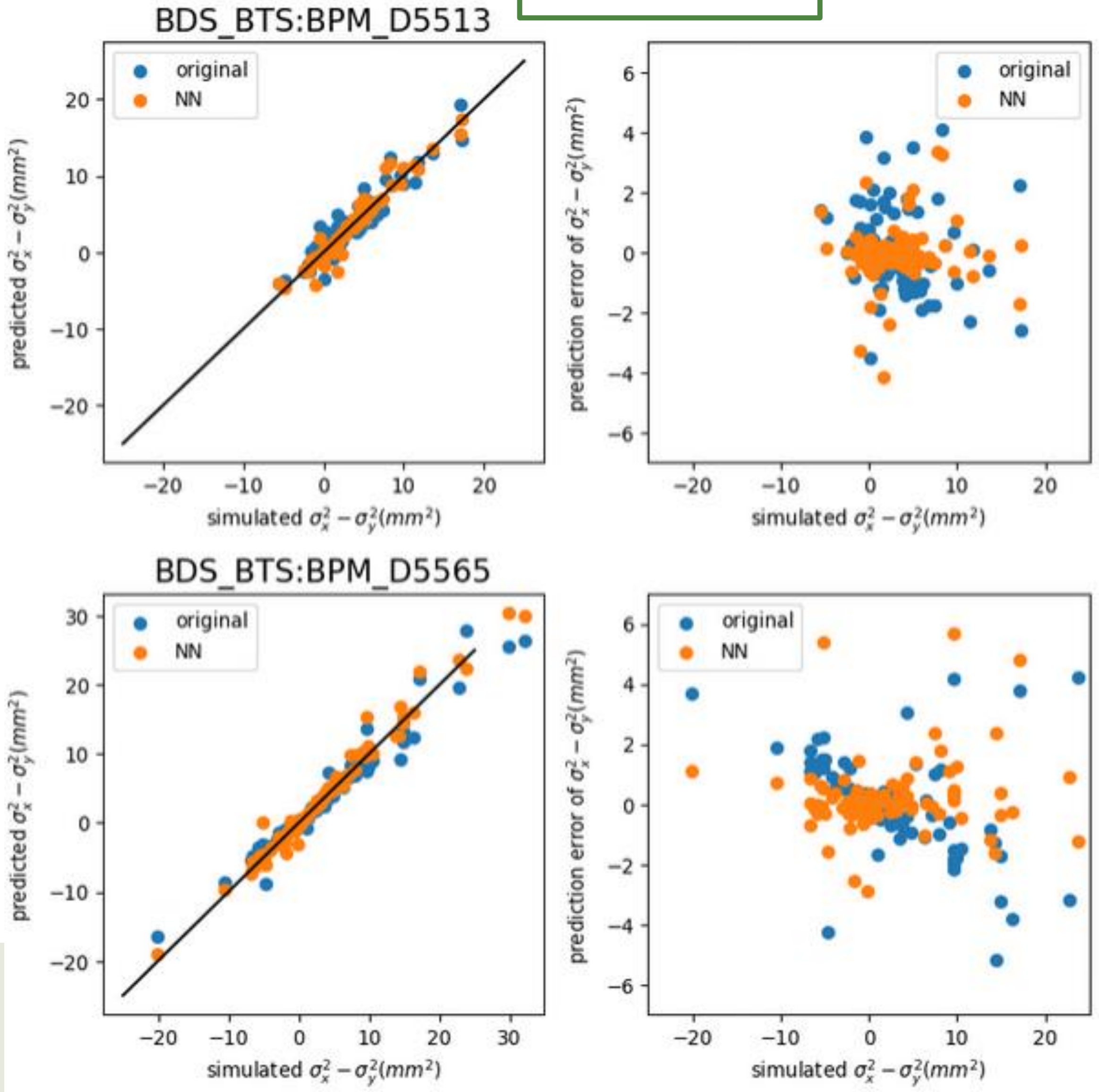


Conclusion

- Fast beam tuning is critical for the FRIB mission
- Customized Bayesian Optimization (BO) for FRIB
  - Maximum utilization of beam time
  - Routine tasks are being established
  - Enhanced Automation, Visualization, and UI based on user feedback
- Surrogate modeling of physics simulators
  - Surrogate modeling of 1D longitudinal RF cavity simulator achieved good accuracy and speed
- Virtual Diagnostics
  - Bunch length measurement using BPM
  - Beam quadrupole moment measurement using BPM
  - Model accuracy depends on the data accuracy and quantity.
  - Strategies of CS reconstruction accuracy improvement with measurement error from virtual diagnostics is under development
    - » Even if BPMQ error is large, CS could be reconstructed within tolerable accuracy by
      - Emittance prior
      - A large number of Q-scan to avoid over-fit CS to BPM-Q error
      - Convolution filters applied for 1D bunch signals from BPM

BPM-Q Model with Full Profile vs 161MHz Amplitude Input

Full Profile



161 MHz amplutude