E331: FY24 Progress and Plans for FY25

Neural network based tuning to exploit machine-wide sensitivities in pursuit of high beam quality

---

E331 collaboration

FACET-II PAC Meeting, 20 November 2024, SLAC

Auralee Edelen, Ryan Roussel, Eric Cropp, Claudio Emma, Brendan O’Shea, Sara Miskovich, Nathan Majernik, Robert Ariniello, Chaojie Zhang, Willie Neiswanger, and many others

FACET-II

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University | U.S. DEPARTMENT OF ENERGY

Outline

- Background
- E331 goals and timeline
- Progress on infrastructure
- 2023-2024 progress
- Next steps and goals for 2025

E331 Science Motivation

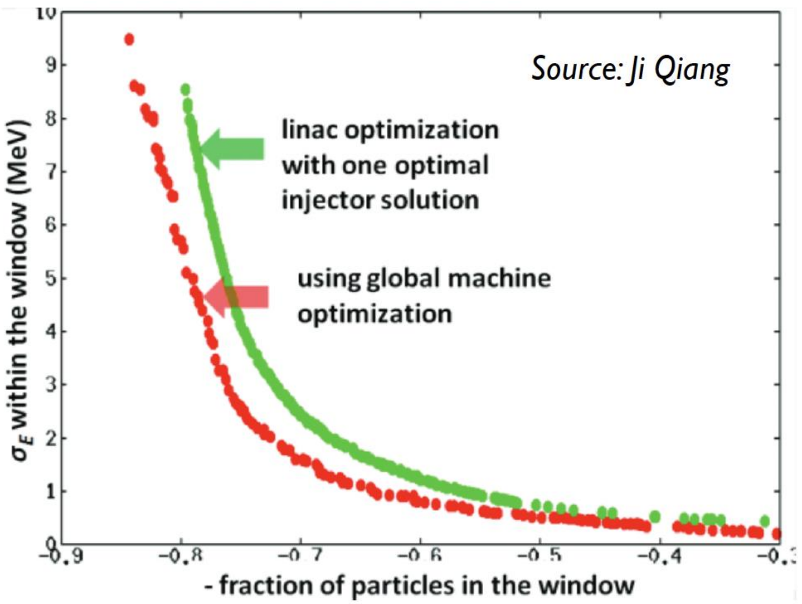
Major limitations in the way accelerator tuning is done:

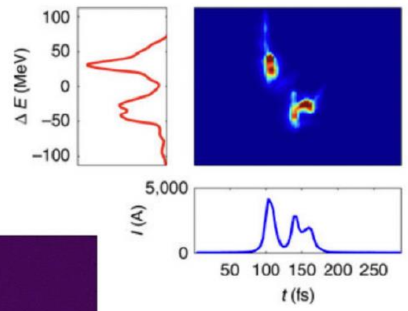
- Piecemeal tuning of subsystems *(known to be sub-optimal)*
- Indirect use of high-dimensional diagnostics *(e.g. images)*
- Often a lack of accurate online models

→ *Potentially limiting factors in control of extreme beams*

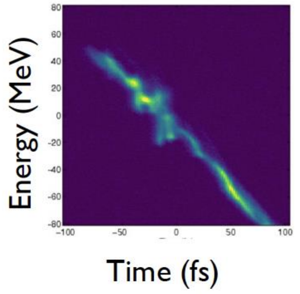
More global view can enable better control:

- Fully exploit unknown system-wide sensitivities + nonlinearities
- Faster switching between setups *(if using global representation of machine)*
- Better handling of parameter tradeoffs *(e.g. jitter, matching, longitudinal phase space)*





A. Marinelli, et al., Nat. Commun. **6**, 6369 (2015)



A. Marinelli, IPAC'18

Comprehensive, system-wide control is likely to be a key factor in improving custom control of extreme beams, but this is a difficult task

Neural Network Reinforcement Learning

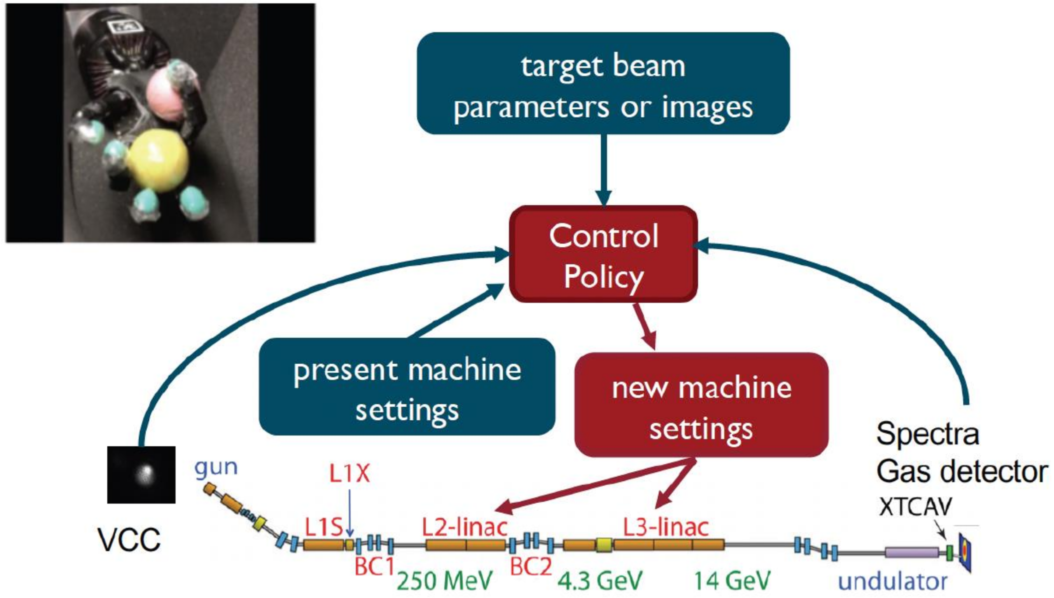
RL can help address a different set of needs than BO:

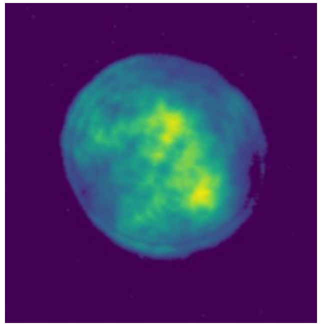
- Use global machine information, more historical data
- Treat as a dynamical system *(many time-dependent processes/feedbacks + drift)*
- Address demands for fast dynamic control from users

Suitability of accelerator tuning problems for RL:

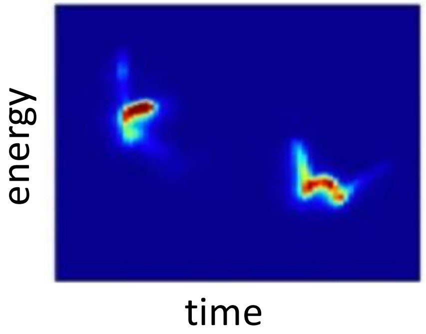
- Many variables, multi-modal signals (images, scalars, time series)
- Continuous state/action spaces (similar to robotics)
- Have physics models/simulators for many problems

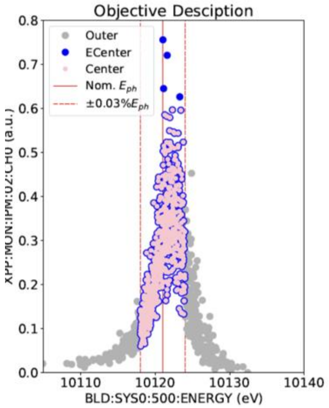
Nagabandi, et al., 2019



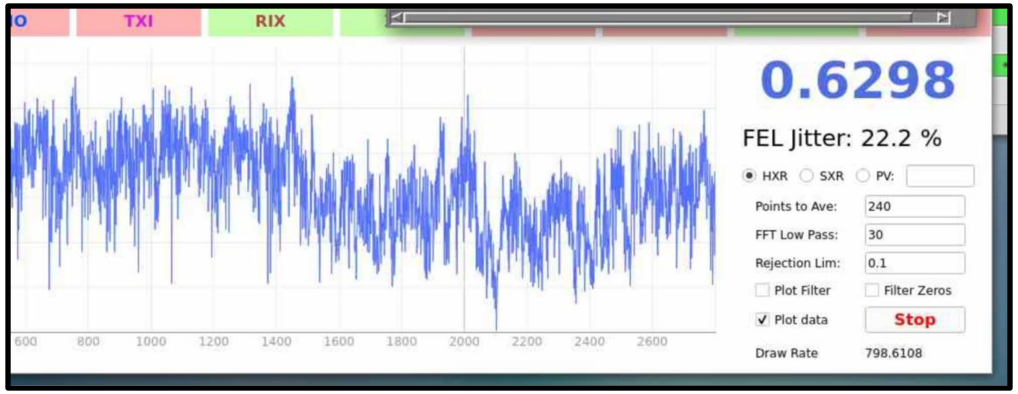


x-y laser





Variety of high dimensional signals for states, objectives



120 Hz FEL pulse intensity

Nonlinear instability → sensitive to dynamic processes

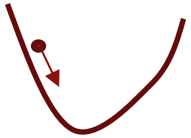
(e.g. trajectory feedback, cooling, LLRF control)

SLAC

Tuning approaches leverage different amounts of data / previous knowledge
→ suitable under different circumstances

less ← assumed knowledge of machine → more

Model-Free Optimization

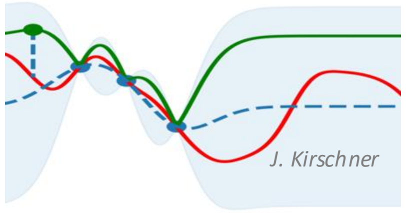


Observe performance change after setting adjustments

→ estimate direction or apply heuristics toward improvement

gradient descent
simplex
ES

Model-guided Optimization

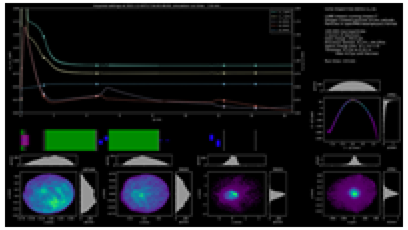


Update a model at each step

→ 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

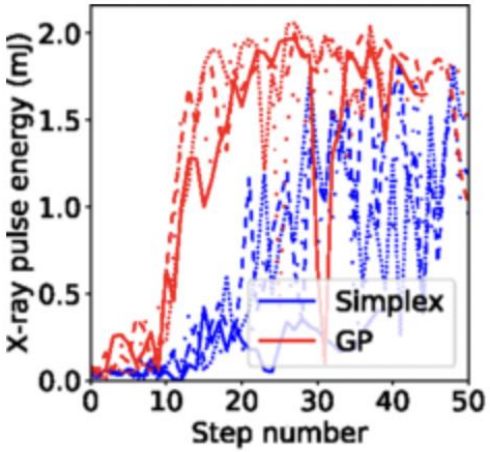
Model-based warm start

General strategy: start with sample-efficient methods that do well on new systems, then build up to more data-intensive and heavily model-informed approaches.

Many successes with Bayesian Optimization

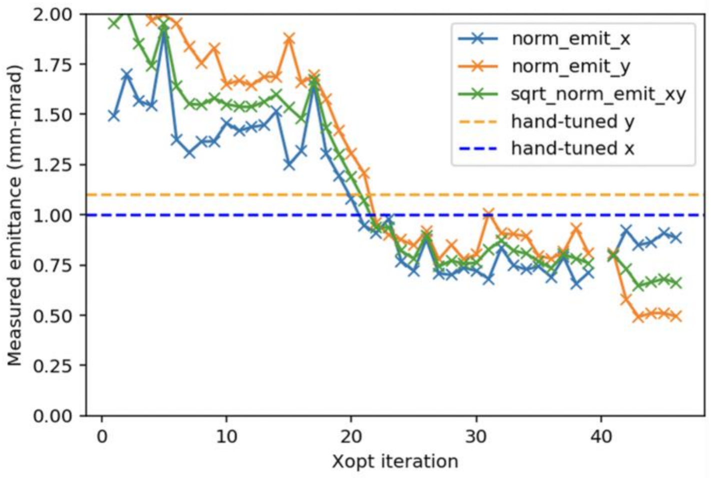
(+ improvements)

FEL pulse energy tuning at LCLS

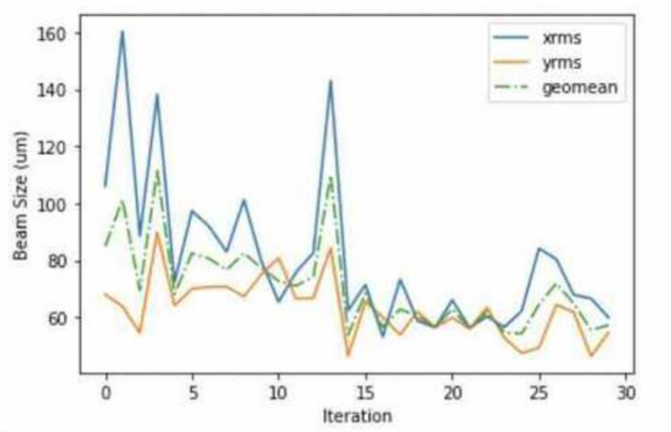


*Duris et. al. PRL , **2020***

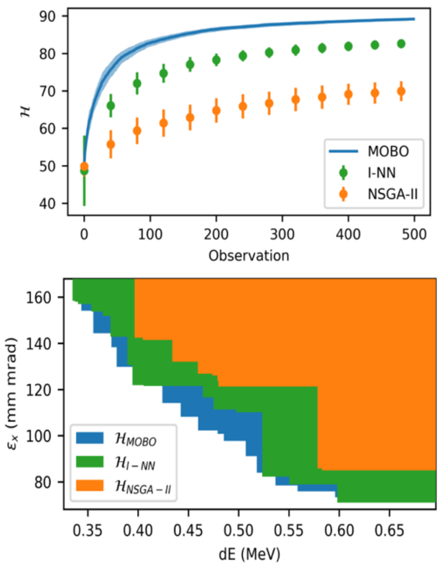
Beam emittance tuning for LCLS-II injector



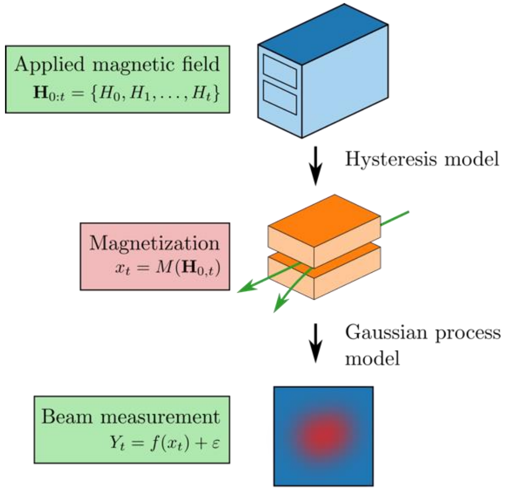
Sextupole tuning for IP at FACET-II



Multi-objective Bayesian Optimization

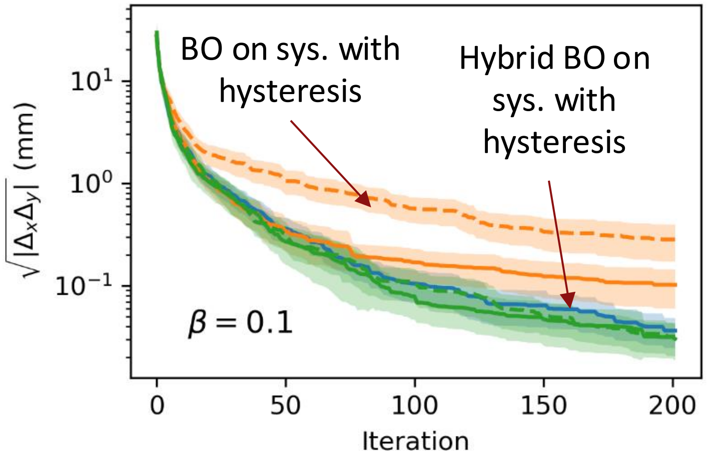


*Roussel et. al. PRAB , **2021***

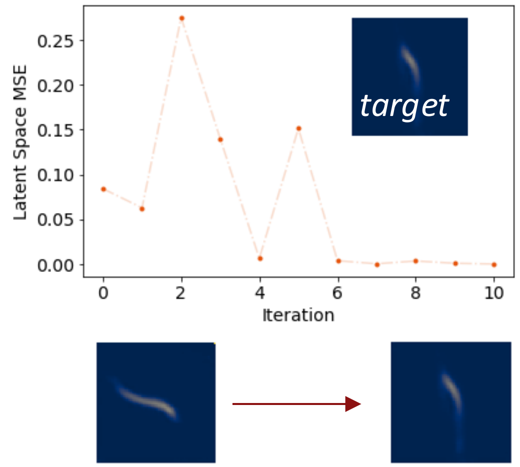


*Roussel et. al. PRL , **2022***

Higher-precision optimization possible when including hysteresis effects in model



Longitudinal phase space tuning on LCLS



Algorithms being implemented/distributed in Xopt: https://github.com/xopt-org/

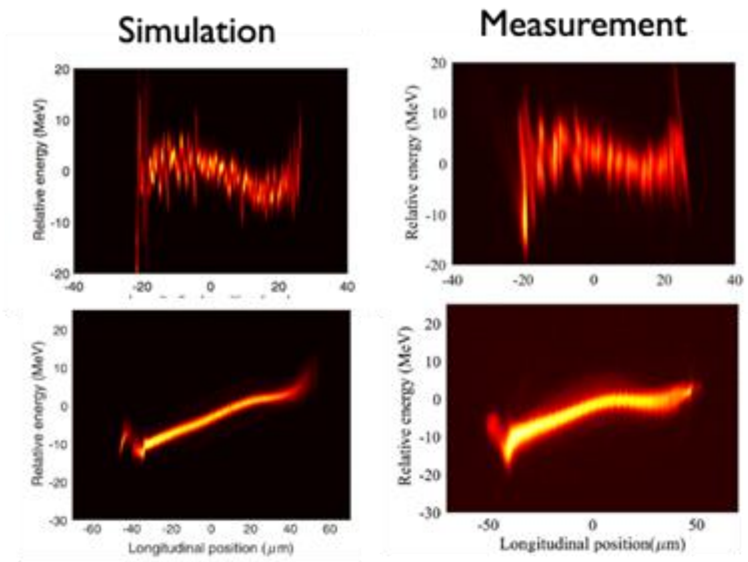
Comprehensive review of advanced BO for particle accelerators: https://arxiv.org/html/2312.05667v2

SLAC

Xopt

Fast-Executing, Accurate System Models

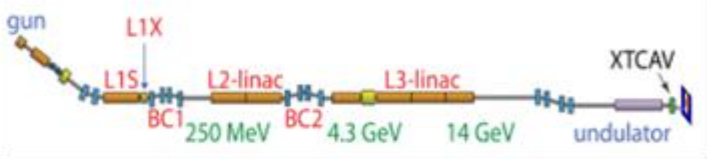
Accelerator simulations that include nonlinear and collective effects are powerful tools, but they can be computationally expensive



J. Qiang, et al., PRSTAB30, 054402, 2017

10 hours on thousands of cores at NERSC!

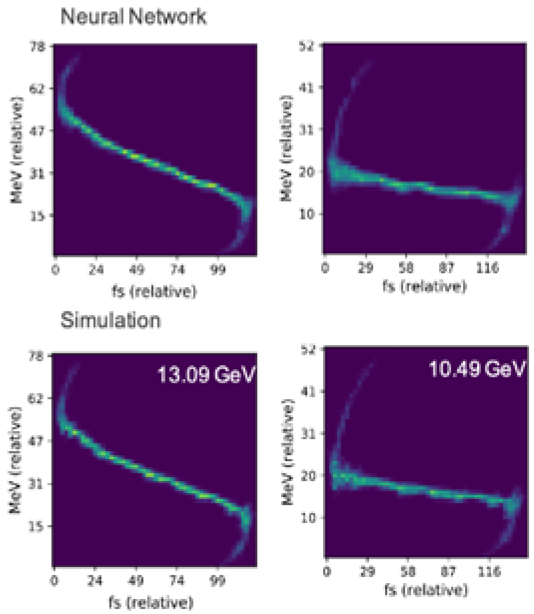
ML models are able to provide fast approximations to simulations ("surrogate models")



Linac sim in Bmad with collective beam effects

Scan of 6 settings in simulation

Variable	Min	Max	Nominal	Unit
L1 Phase	-40	-20	-25.1	deg
L2 Phase	-50	0	-41.4	deg
L3 Phase	-10	10	0	deg
L1 Voltage	50	110	100	percent
L2 Voltage	50	110	100	percent
L3 Voltage	50	110	100	percent



< ms execution speed

10^6 times speedup

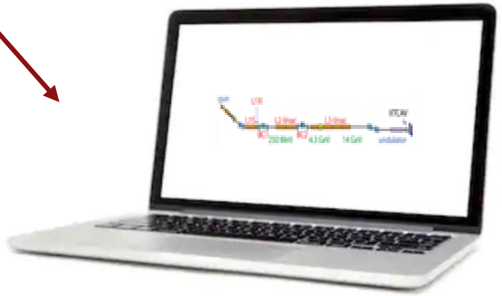
Edelen et al., NeurIPS 2019

ML modeling enables accurate predictions of system responses with unprecedented speeds, opening up new avenues for high-fidelity online prediction, tracking of machine behavior, and model-based control

Fast-Executing, Accurate System Models

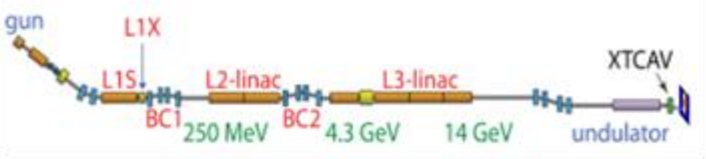
Bringing simulation tools from HPC systems to online/local compute

Online prediction
Model-based control



Control prototyping
Experiment planning

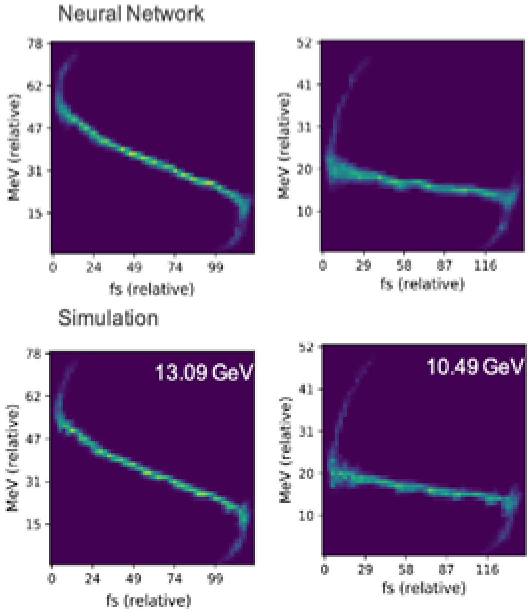
ML models are able to provide fast approximations to simulations ("surrogate models")



Linac sim in Bmad with collective beam effects

Scan of 6 settings in simulation

Variable	Min	Max	Nominal	Unit
L1 Phase	-40	-20	-25.1	deg
L2 Phase	-50	0	-41.4	deg
L3 Phase	-10	10	0	deg
L1 Voltage	50	110	100	percent
L2 Voltage	50	110	100	percent
L3 Voltage	50	110	100	percent



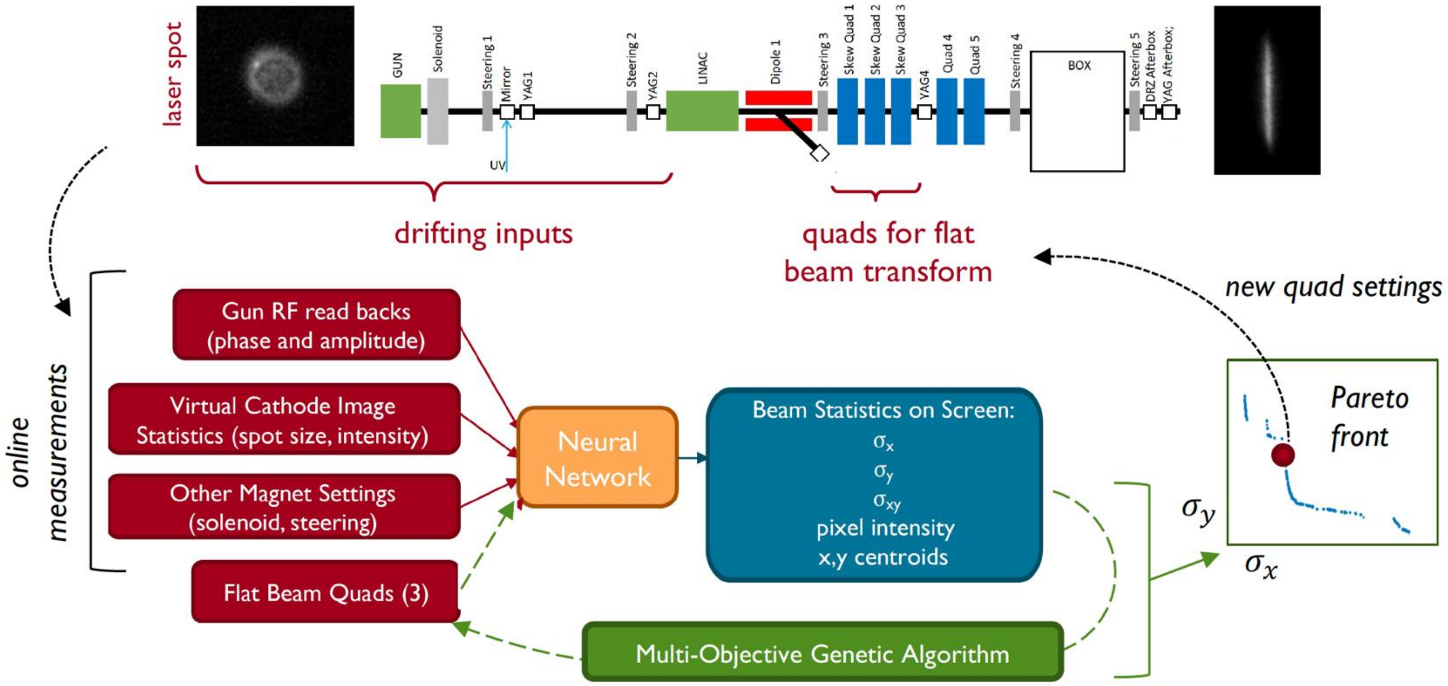
< ms execution speed

10^6 times speedup

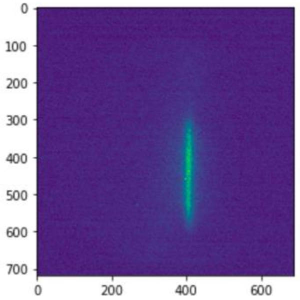
[*Edelen et al., NeurIPS 2019*]

ML modeling enables accurate predictions of system responses with unprecedented speeds, opening up new avenues for high-fidelity online prediction, tracking of machine behavior, and model-based control

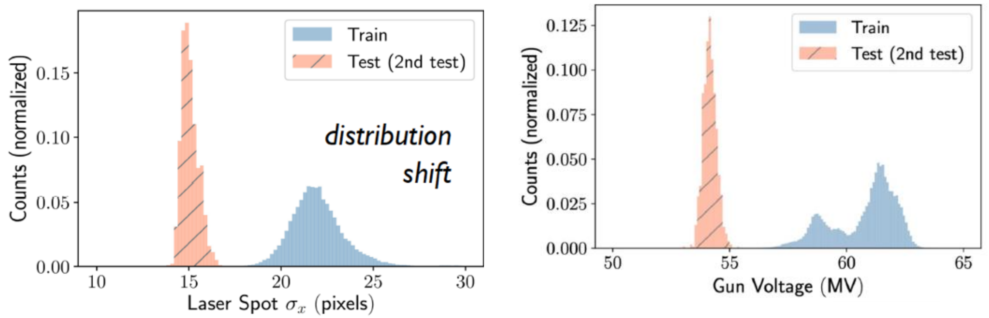
Example: Warm Starts from Online Models

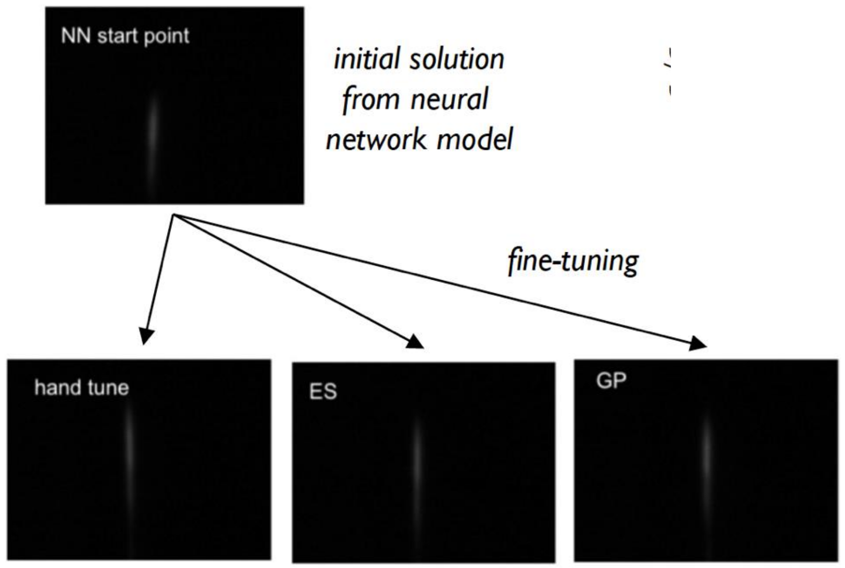


- Round-to-flat beam transforms are challenging to optimize → 2019 study explored ability of a learned model to help
- Trained neural network model to predict fits to beam image, based on archived data
- Tested online multi-objective optimization over model (3 quad settings) given present readings of other inputs
- Used as warm start for other optimizers
- Trained Reinforcement Learning agent and tested



Can work even under drift





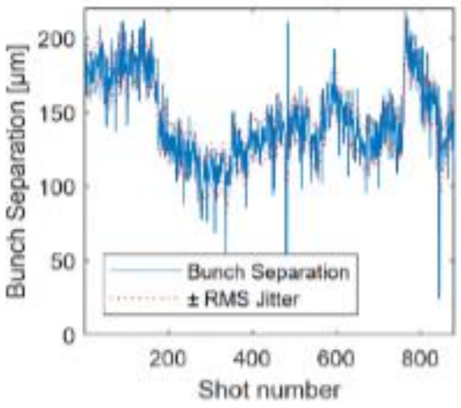
Hand-tuning in seconds vs. tens of minutes

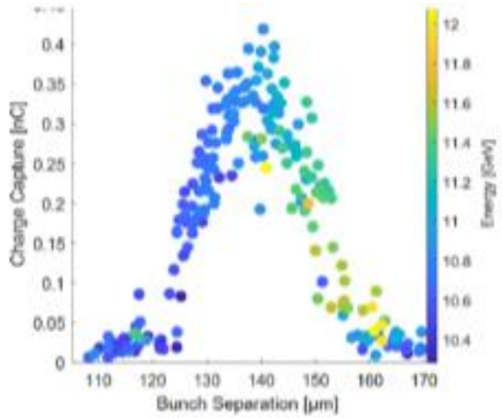
Boost in convergence speed for other algorithms

Drift / Distribution Shift is a Major Challenge

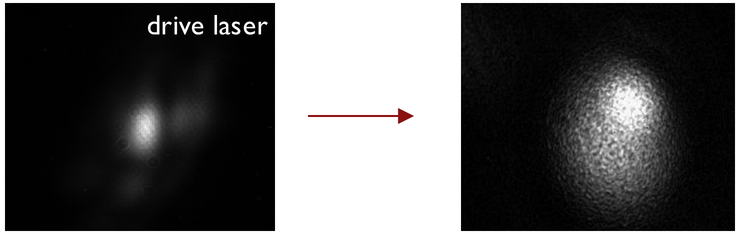
Many sources of change over time:

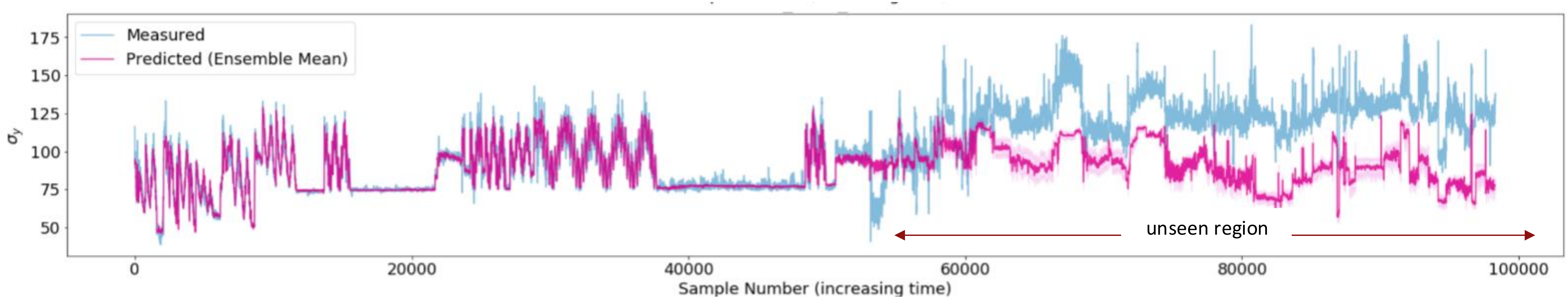
- Deliberate changes in beam configuration (e.g. beam charge)
- Unintended drift in initial conditions (including in unobservable variables), diurnal temperature/humidity changes, etc
- Time-dependent action of feedback systems





B. O'Shea





Example: beam size prediction and uncertainty estimates under drift from a neural network

Uncertainty estimate from neural network ensemble does not cover prediction error, but does give a qualitative metric for uncertainty

Reliable uncertainty estimates and model adaptation methods are key for putting online models to use operationally

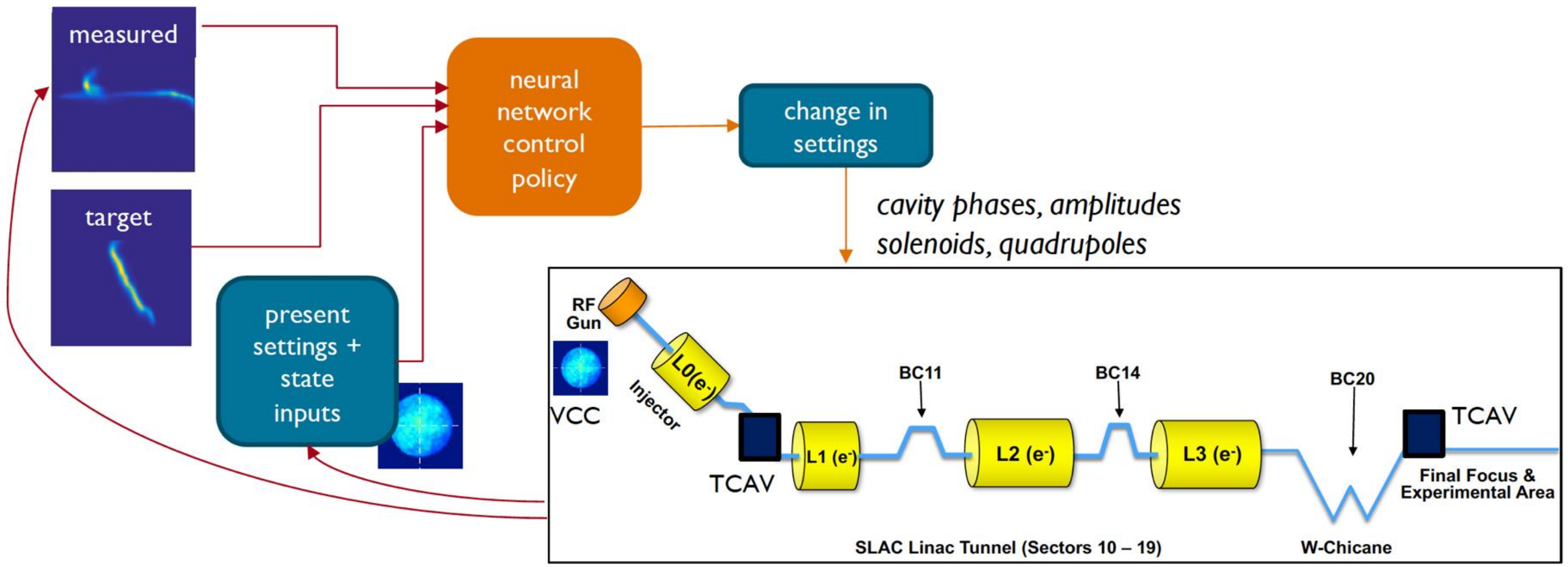
E331 Science/Technical Goals

AI/ML R&D for complex system control

high-impact science cases

Main goal: develop and demonstrate methods to leverage global learned system responses to aid **fast, high-quality tuning** of beams under challenging conditions and aid **switching between setups**

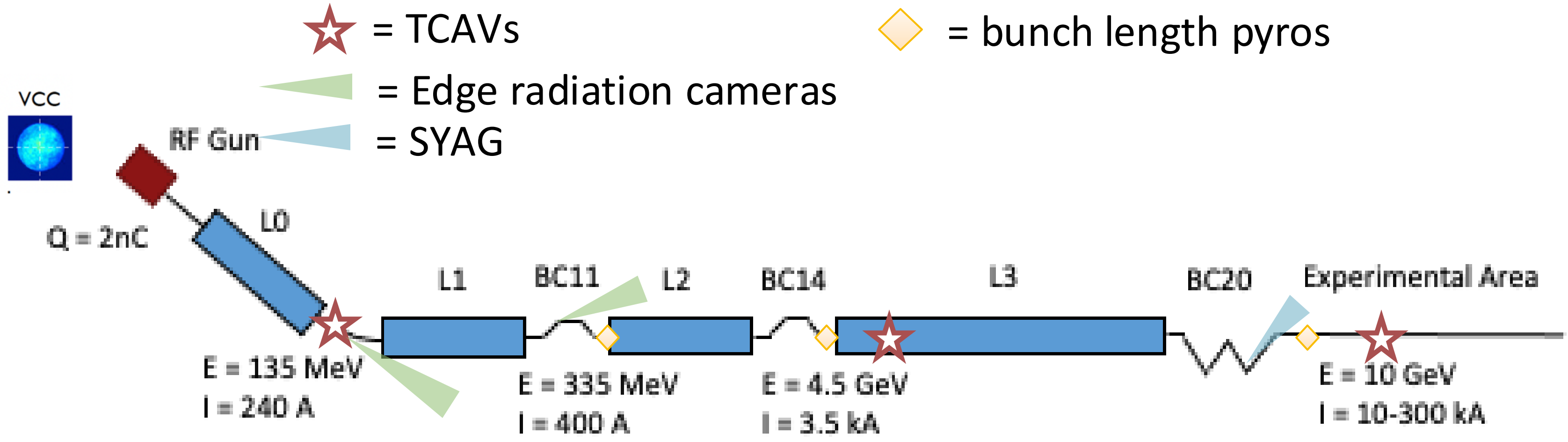
(build up incrementally to machine-wide neural network-based reinforcement learning)



Science/Technical Goal	Target Time	Definition of Success
Evaluate methods for high-dimensional, high-quality control over beams using learned responses, starting with small-scale problems + single-bunch mode	3+ years	Automated tuning of transverse emittance and longitudinal phase space: faster, higher-quality tuning than standard methods, new capabilities in control
High-quality control over extreme beams and plasma experiments, two-bunch mode	3-5 years	Same as above but for more challenging setups/target beams
Deliver algorithms and interfaces for regular operation	continual	Tools incorporated into regular use + transitioned to operations

Staged approach gradually increases complexity, goes from sample-efficient methods that learn on-the-fly to comprehensive model-based methods that use variety of machine data → success determined by improvements in tuning quality and speed, and TTO

E331 Diagnostic and Observables



- LPS diagnostics (e.g. injector + downstream TCAVs)
- Emittance measurements, x-y beam sizes from wires, transverse phase space from screens
- Upstream inputs: virtual cathode camera, QE map once available, laser diagnostics
- Readbacks from settings (gun solenoid, gun and linac phases/amplitudes etc)
- DAQ: scalar diagnostics (e.g. BPMs, toroids, RF readbacks, BLEN pyros) and multiple image diagnostics (SYAG, EOS, TCAV)
- Plasma diagnostics

→ *Flexibility in E331 enables adaptation to installation / commissioning schedule for different diagnostics*

Numerous diagnostics to inform tuning or be used as tuning targets

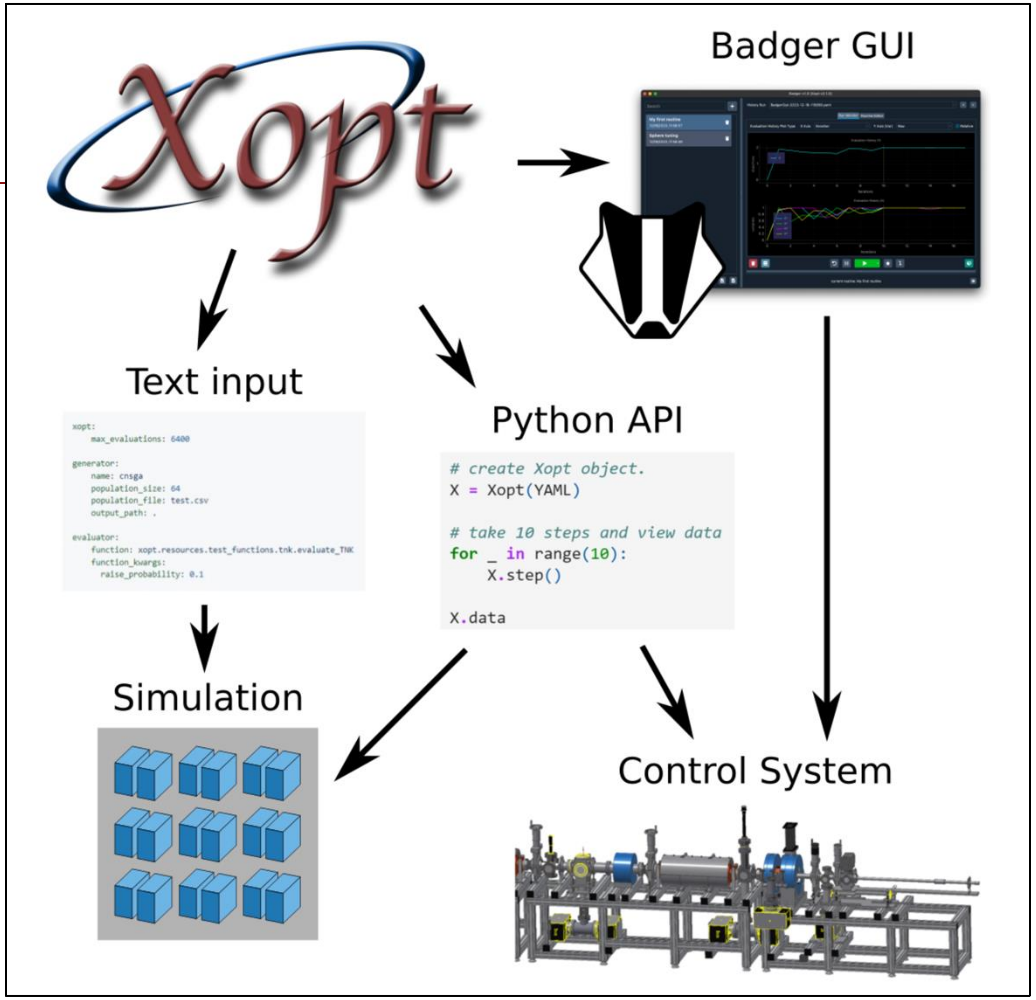
E331 Practicalities and Infrastructure

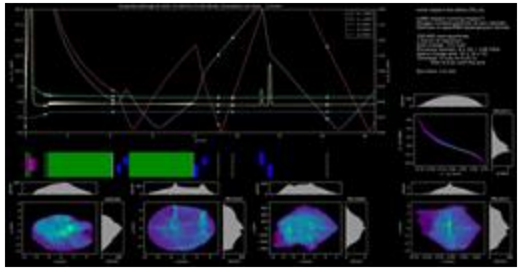
- Integrated our **Xopt software into FACET-II control system**
  - *Aids algorithm transfer between experiments; easier to test new algorithms*
  - *Xopt being used very broadly in accelerator community*

- **Have improved Badger user interface.** *Next need to add FACET-II tuning cases (e.g. sextupole mover tuning, injector emittance tuning)*

- Found **computing is a major bottleneck**
  - *20-40x slower per compute step on controls network than laptops*
  - Working on several solutions: (1) GPU on controls network, (2) HPC integration with S3DF (broader laboratory infrastructure)
  - Addressing this has been an extremely slow process due to administrative, funding, and engineering personnel constraints (above the FACET-II level); have now (FY25) gotten better laboratory support for this

- **Modeling infrastructure** *(need for model-based control / RL)*
  - Been making progress on start-to-end simulation tools and infrastructure for online ML and physics model deployment
  - Want improved model accuracy wrt real machine behavior + to have reliable software infrastructure for online model deployment





FACET-II Injector model running online using LUME-IMPACT

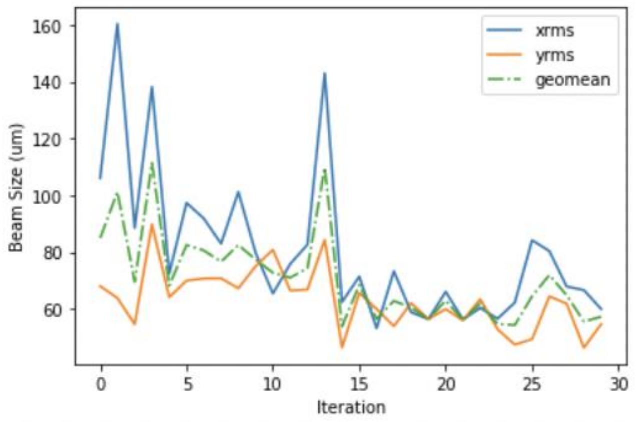
Variety of tools for online modeling and optimization. Integration of these into FACET-II is progressing.

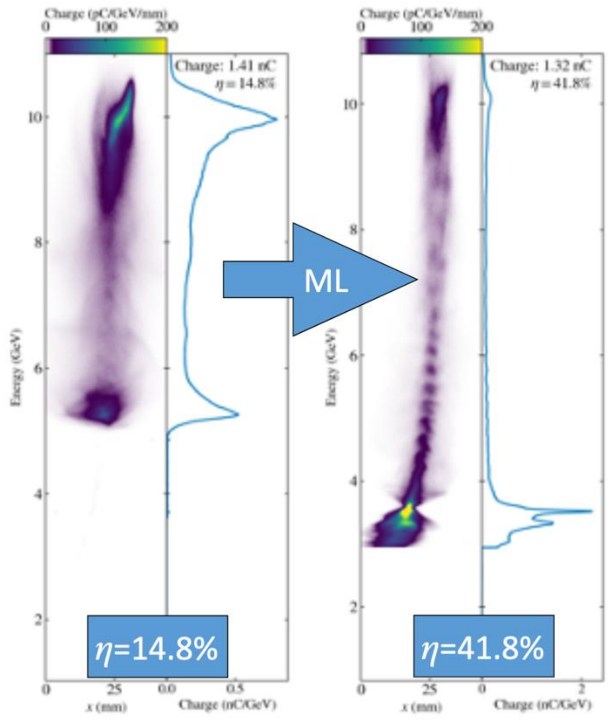
E331 Progress 2023-2024

- **Several data-efficient methods for tuning at FACET-II ready for TTO**
  - Injector emittance tuning → *mostly ready for transition to operations; needs some additional software/interface work and robustness testing*
  - Sextupole mover tuning (beam size, plasma parameters) → *have transitioned to other user experiments; need to add to Badger UI to aid ease-of-use*
  - Automatic "smart sampling" for characterization → *ready for general use; need to add to Badger UI and consider DAQ interface*

- **First transfer to other experiments with E300 for sextupole tuning (see Chan's talk yesterday)!**
  - Initial development with E331 *(8 sextupole movers to minimize S20 beam size, then expanded to eloss and energy gain)*
  - Helped FACET-II/E300 team get set up with the algorithm and provided initial guidance during shifts; joint iteration on how best to set up metrics, data processing

- Progress on ML-based model calibration methods *(getting simulation to match machine in a data-efficient way)*





Algorithm developed/tested in E331 transferred to E300!

Bayesian Optimization and Characterization of Injector

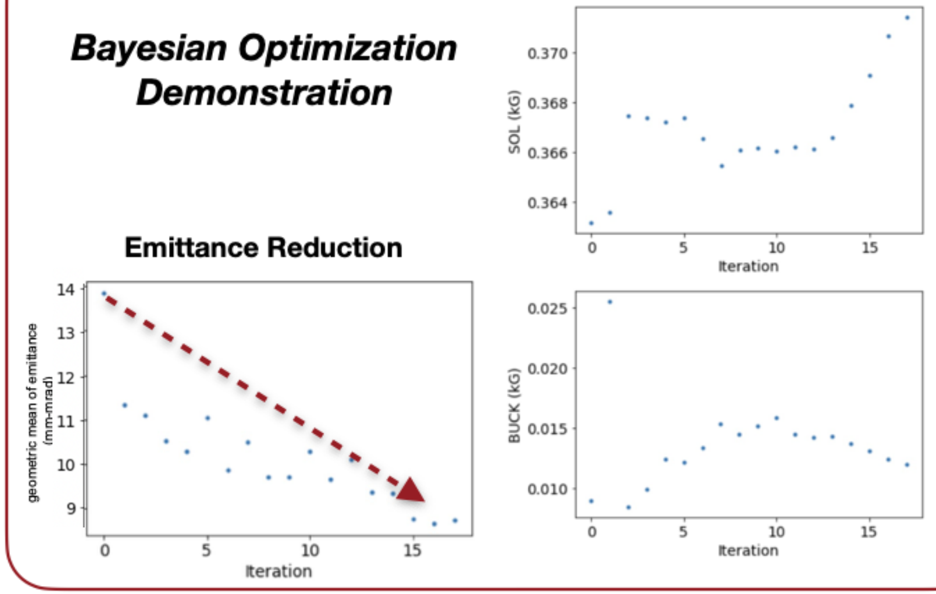
- Bayesian optimization on the injector with up to 10 variables
- Extensive data obtained from characterization studies at 3nC, 2nC, and 700pC

Use of BO results in a local model that can be interrogated → visualization being incorporated into Badger

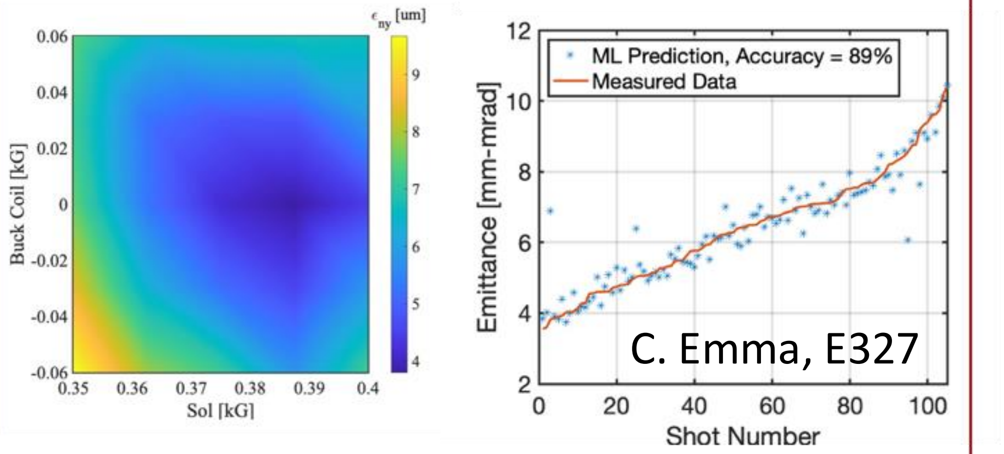
Bayesian Optimization Demonstration

Adjusted Settings

Emittance Reduction



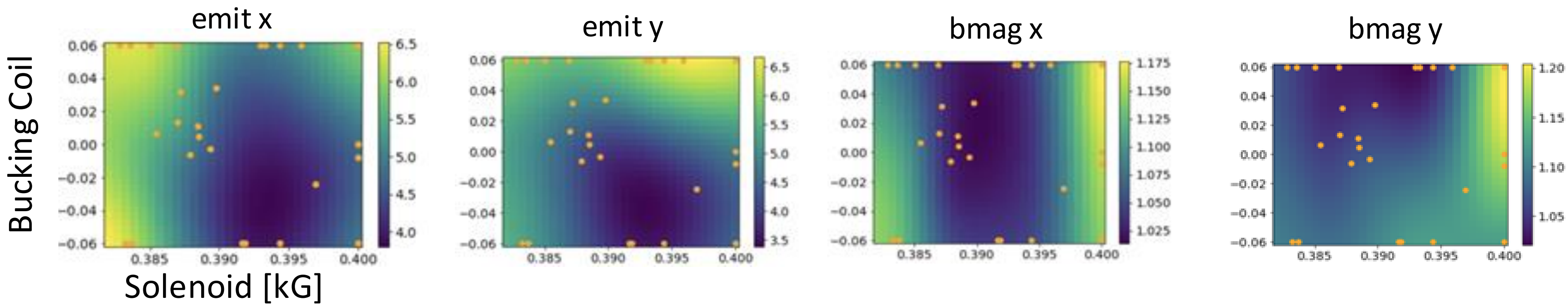
Injector Emittance Characterization



Measured data reveals correlations and local minima

Accurate ML emittance prediction

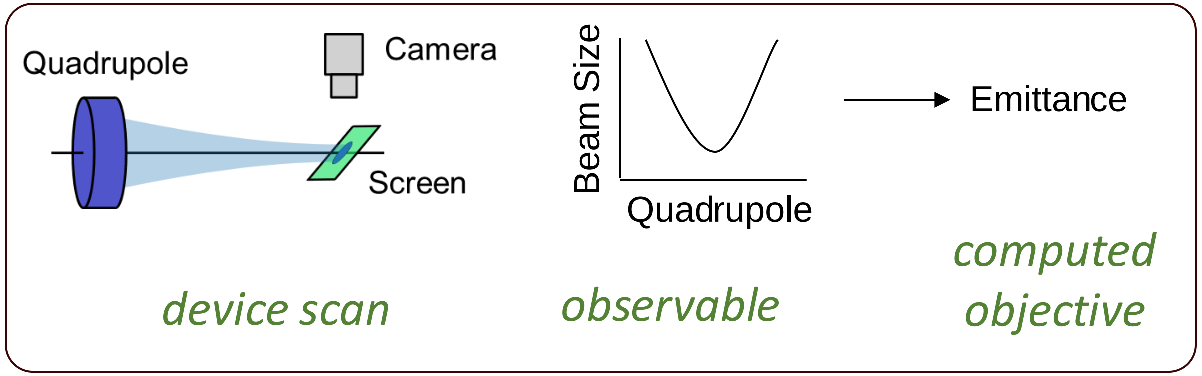
1.8 nC

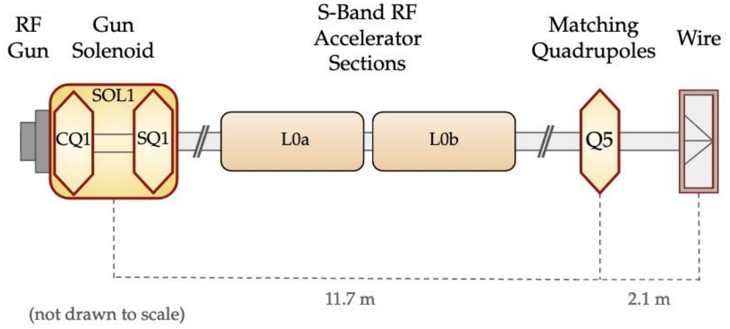


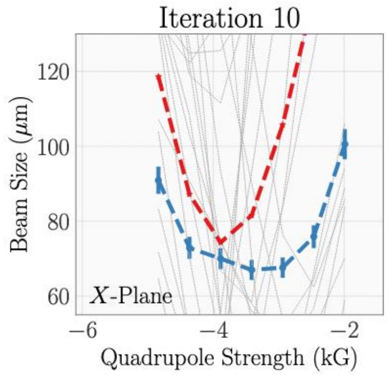
ML model predictions and sample locations

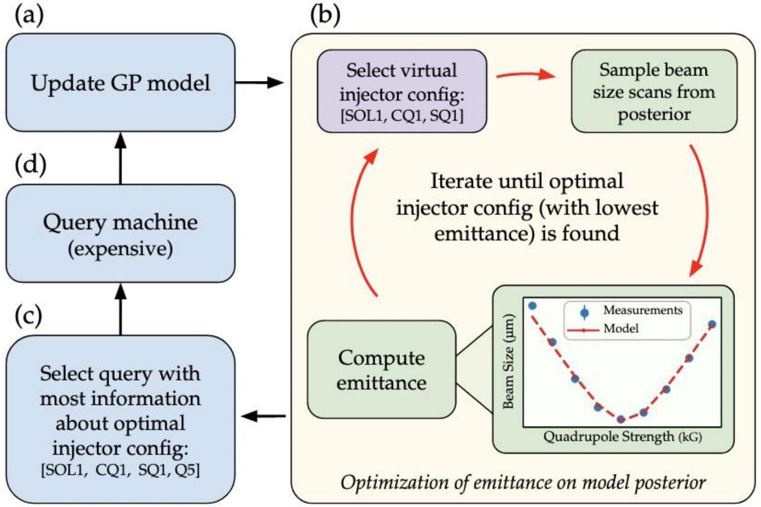
Optimization with Virtual Objectives

- Many objectives require layered scans or optimization problems
- Instead learn model from scratch online and do scan on model
- Bayesian Algorithm Execution (BAX) → 20x speedup in tuning

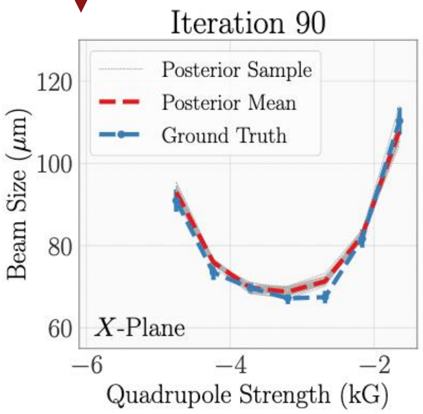


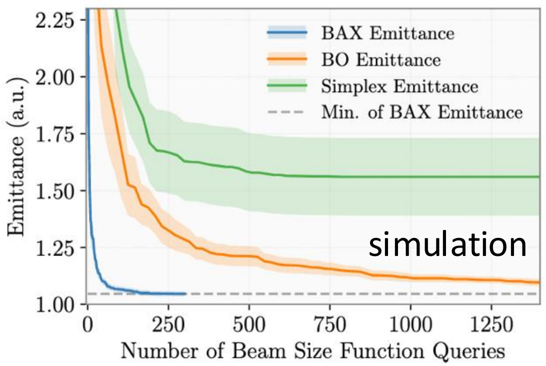


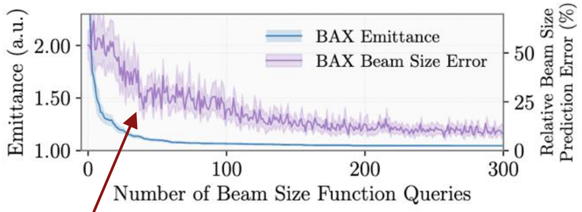




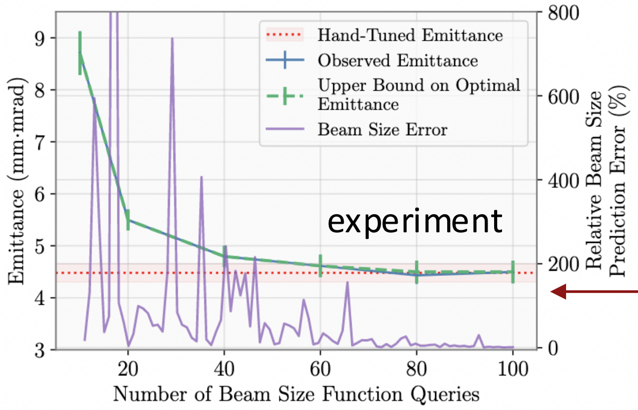
model is learned on-the-fly







Convergence of beam size prediction error gives practical indicator of convergence



20x faster tuning than standard BO, equivalent or better solution than hand-tuning

S. Miskovich, MLST, 2024

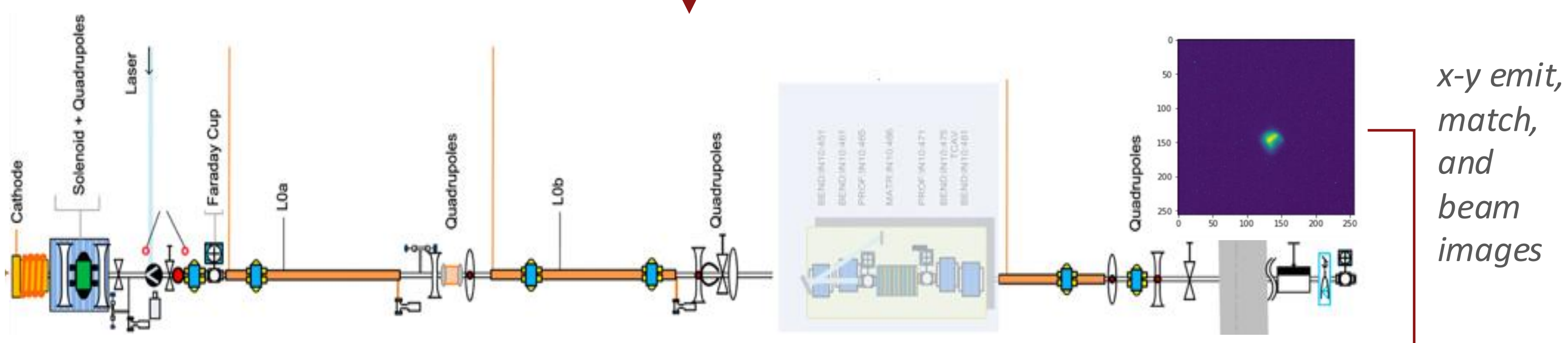
BAX enables a paradigm shift in how optimization problems with complicated scans or other indirect measurements are handled
Demoed at FACET-II injector → now working to get set up for routine use in operations

ML for Efficient Characterization

Automatic Exploration
(constrained to useful values of emittance and match)

ML Models of Injector

Setting changes on 10 variables (solenoid, bucking coil, corrector and matching quads)



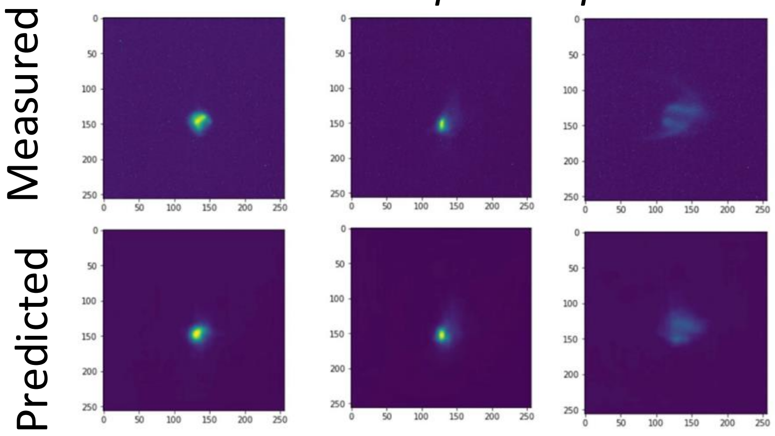
FACET-II Injector

- Used Bayesian Exploration for efficient high-dimensional characterization (10 variables) of emittance and match at 700pC: **2 hrs for 10 variables compared to 5 hrs for 4 variables with N-D parameter scan**

- Data was used to train neural network model of injector response predicting x-y beam images. GP ML model from exploration predicts emittance and match.

- **Example of integrated cycle between characterization, modeling, and optimization → now want to extend to larger system sections and new setups**

transverse phase space



Have used extensively now at FACET-II for injector to aid model calibration studies
→ now want to help expand to other experiments and use downstream in linac

Finding Sources of Error Between Simulations and Measurements

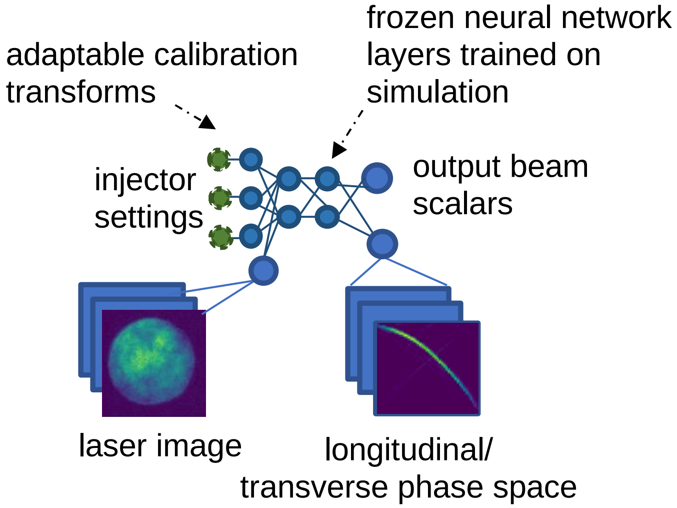
Many non-idealities not included in physics simulations:

- **static error sources** (e.g. magnetic field nonlinearities, physical offsets)
- **time-varying changes** (e.g. temperature-induced phase calibrations)

Want to identify these to get better understanding of machine performance

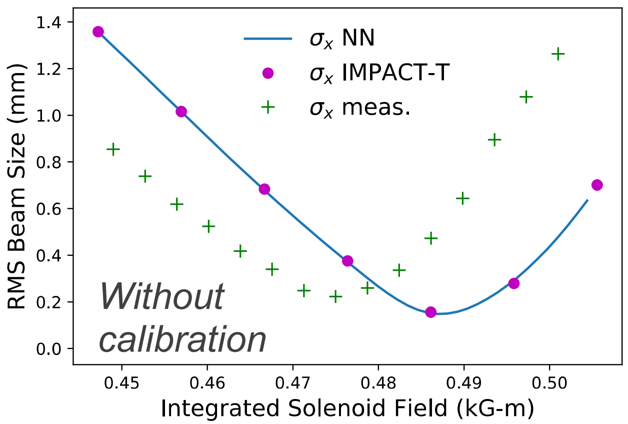
à ML model allows fast / automatic exploration of error sources in high dimension

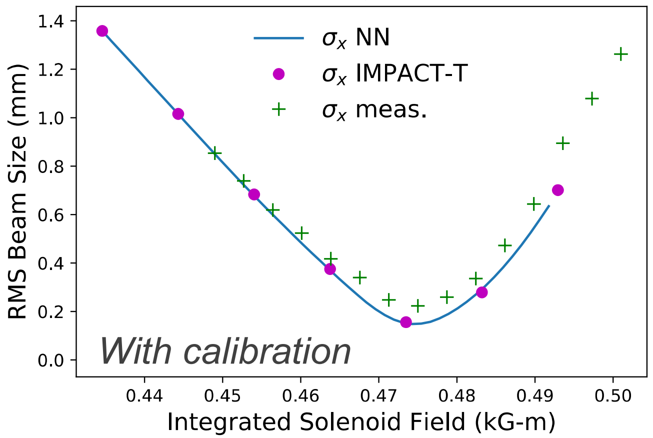
Example: calibration offset in injector solenoid strength found automatically with neural network model (trained first in simulation, then calibrated to machine)



Inputs
Laser radius
Laser spot sizes
Pulse length
Charge
Solenoid
L0A phase
L0B phase
SQ quad
CQ quad
6 matching quads

Outputs
Beam size (x,y)
Emittance (x,y)
Bunch length





Speed of ML models enables rapid identification of error sources between idealized physics simulations and real machine

→ path toward an accurate system model suitable for model-based control and training reinforcement learning control

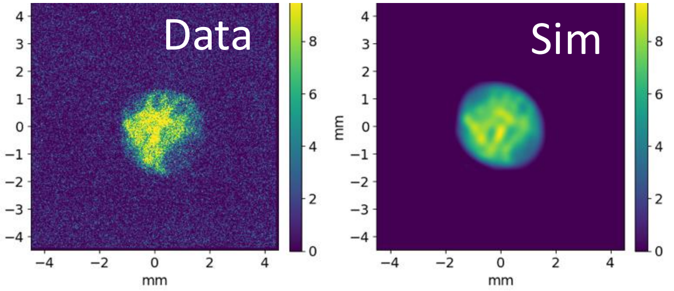
Model Calibration

- Want methods that scale efficiently to high dimension (e.g. injector + linac), are minimally data hungry, and give reliable uncertainty estimates

Neural network approach alone is not sufficient!

- Now have better simulation pipeline and working on several methods with the FACET-II injector and linac
  - Neural network models *(e.g. previous slide)*
  - Bayesian approaches *(e.g. multi-fidelity optimization)*
  - Differentiable simulations *(see Ryan's talk today)*

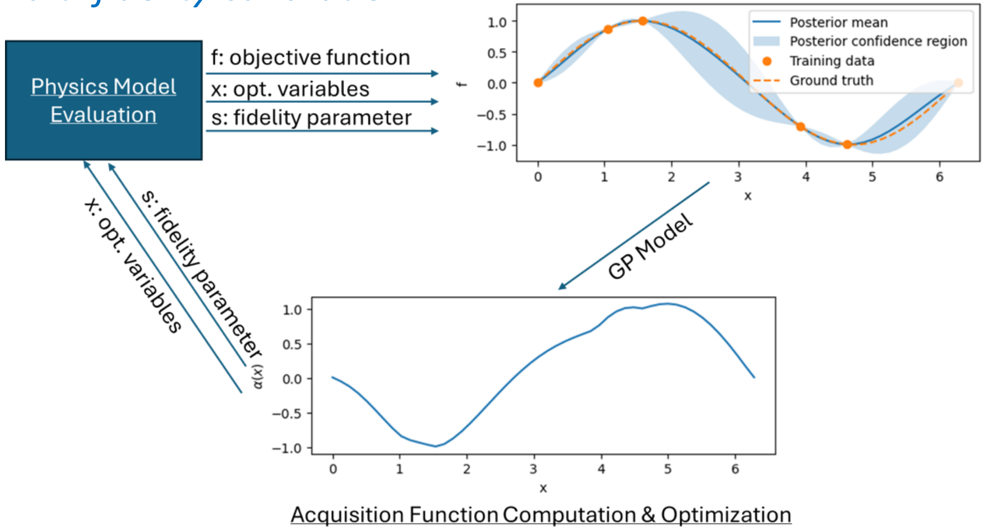
Parameter	Nominal	Optimal
Gun Phase	30 Degrees	31.9 Degrees
Gun Amplitude	120 MV/m	118.99 MV/m
Pulse Length	1 ps	0.847 ps

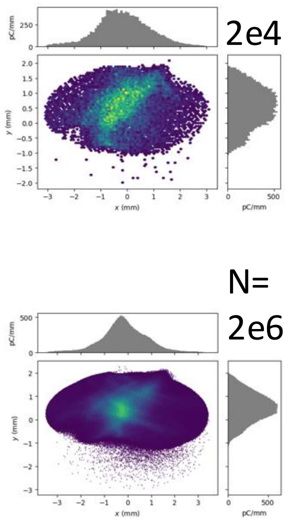


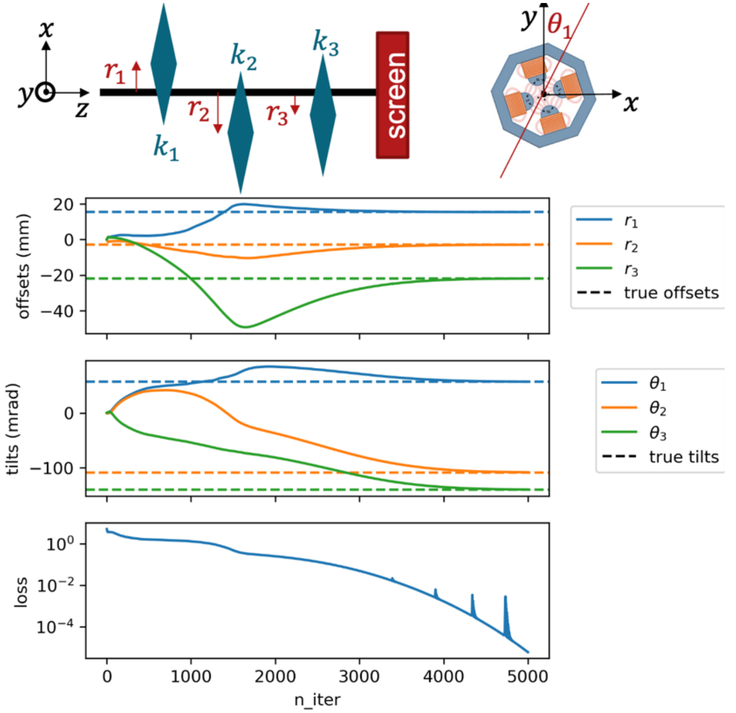
E. Cropp

Preliminary, low-dimensional example of multi-fidelity calibration for injector

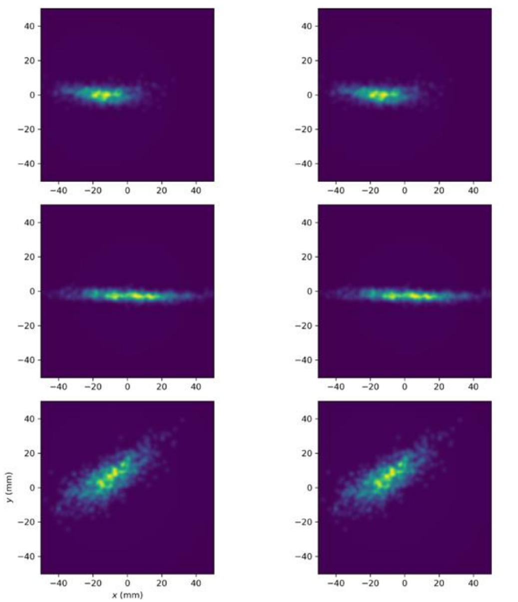
Multi-fidelity calibration







Differentiable physics simulation example



J.P. Gonzalez-Aguilera

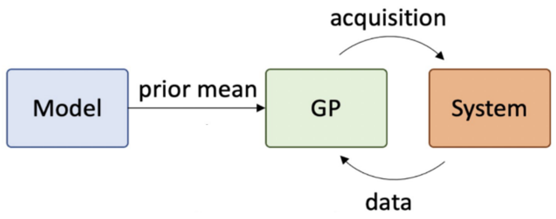
https://accelconf.web.cern.ch/ipac2023/pdf/WEPA065.pdf

SLAC

E331 Next Steps: Neural Network Prior

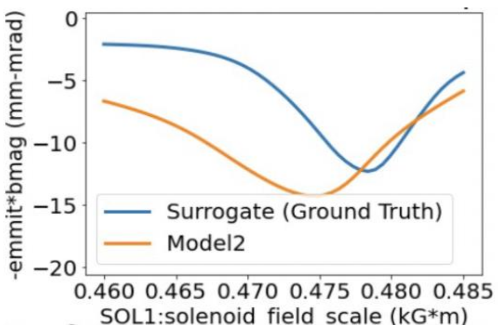
Combining system models with BO → important for scaling BO up to higher-dimensional tuning problems

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



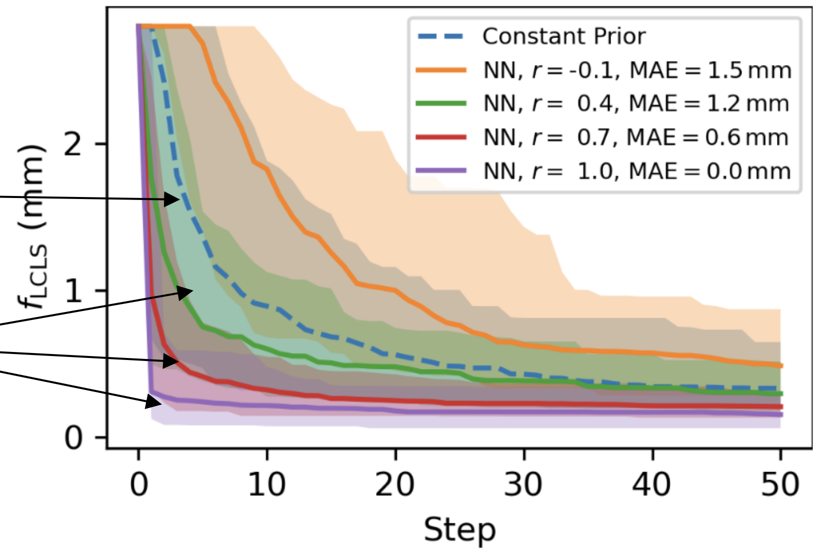
Used LCLS injector surrogate model for prototyping
variables: *solenoid, 2 corrector quads, 6 matching quads*
objective: *minimize emittance and matching parameter*

Even prior mean models with substantial inaccuracies provide a boost in initial convergence



regular Bayesian optimization

prior mean from models with different fidelity

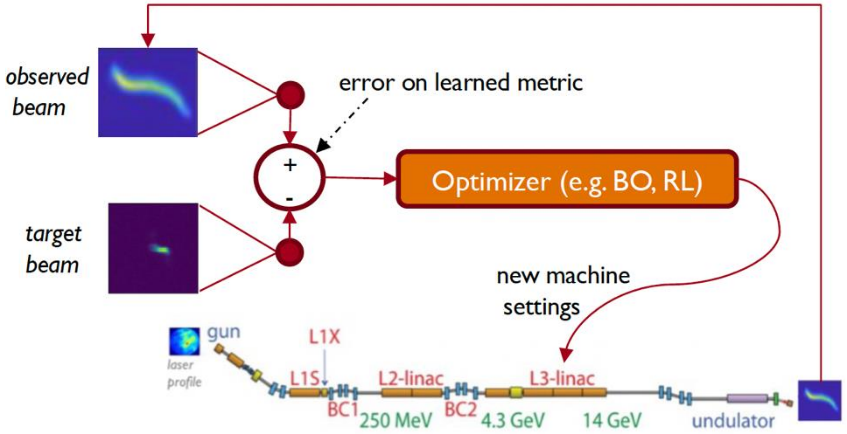


- Want to apply this to with sextupole tuning, injector and linac tuning, etc at FACET-II → would help significantly with high-dimensional tuning
- Should work well in cases where machine drifts but qualitative response is similar

https://arxiv.org/abs/2403.03225
https://arxiv.org/abs/2211.09028

SLAC

E331 Next Steps:
Longitudinal Phase Space Tuning



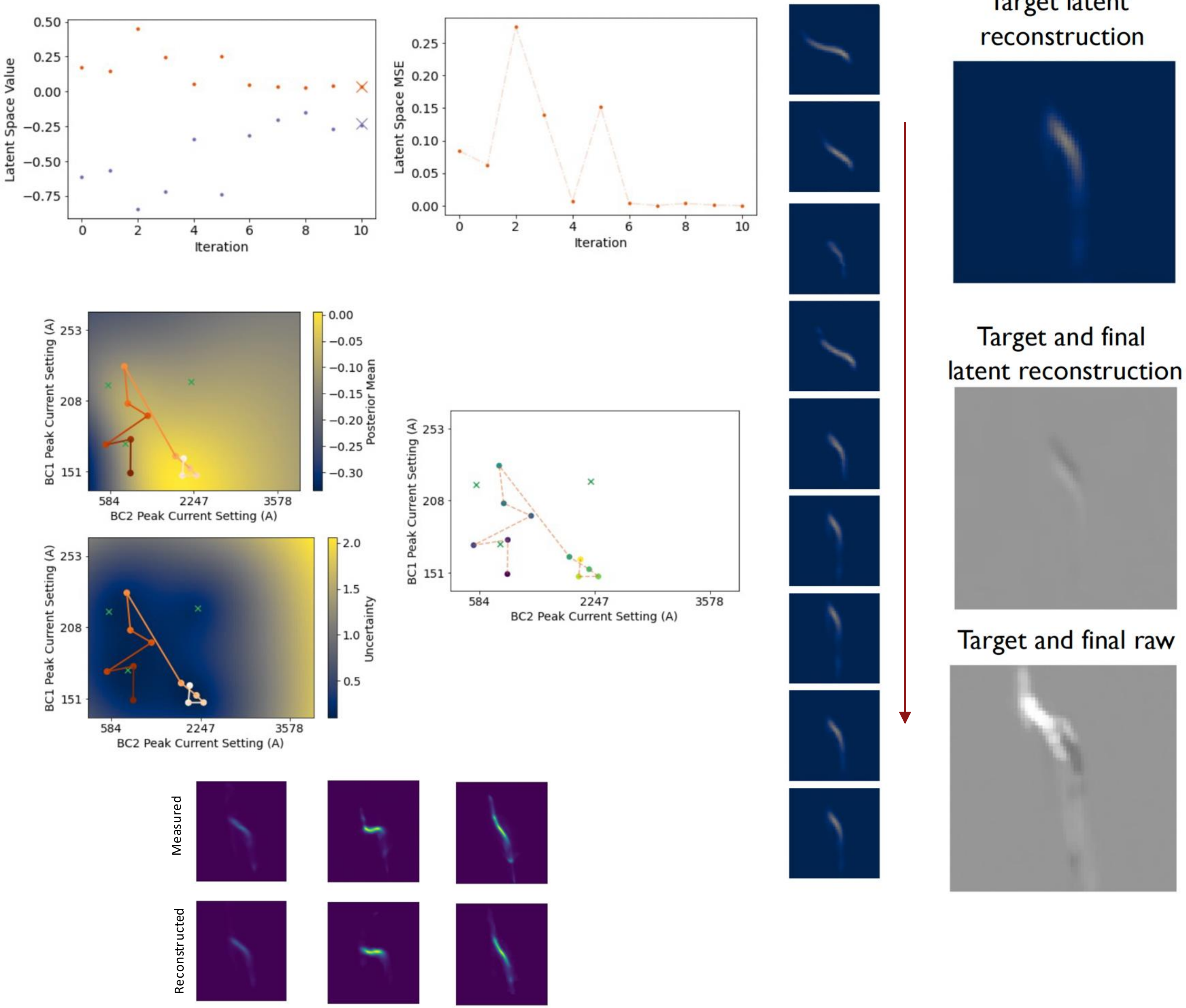
Demonstrated Bayesian optimization for LPS tuning on LCLS for several variants of problem setup:

- *2 peak current settings, 6 phases and amplitudes*
- *Target phase space, minimize energy spread and bunch length*

→ Want to expand on this work at FACET-II

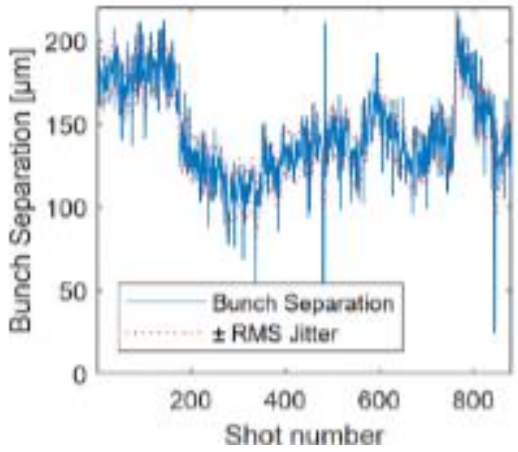
→ Data gathered during BO-based tuning will be useful for next steps *(model calibration, neural network control policy + reinforcement learning)*

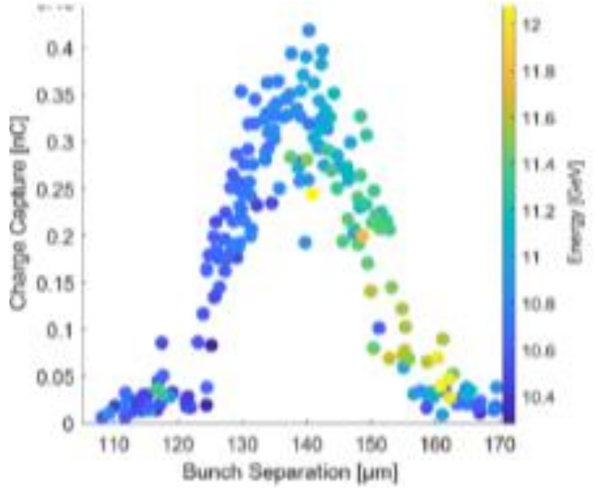
Example from LCLS



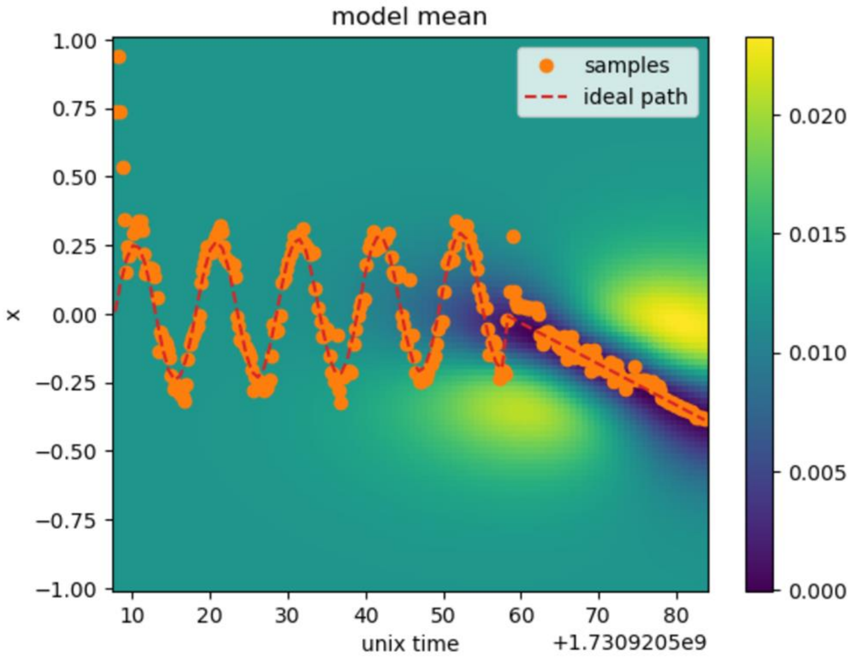
SLAC

Next steps: drift compensation

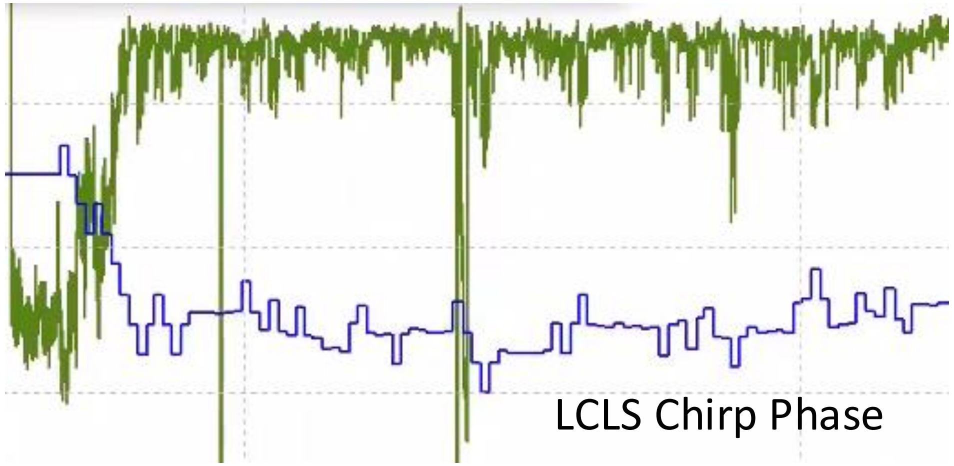




Synthetic problem



LCLS example: FEL pulse intensity



Time Dependent Bayesian Optimization

Time dependent BO, contextual BO, and model-based RL are all good candidates

Would like to help address the major pain points for other experiments wrt drift

SLAC

Next plans for E331

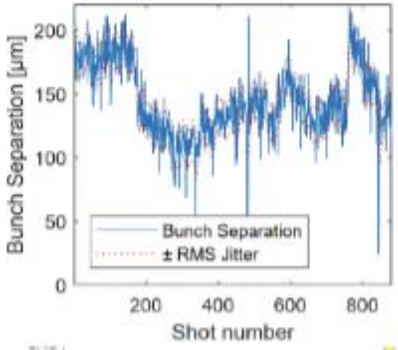
- **Expand tuning scope being addressed with Bayesian approaches**:
  - Incorporate information about upstream variations (e.g. charge, bunch length) to aid **compensation for drift**
  - Expand to **emittance growth minimization for linac**
  - Expand sextupole mover tuning w/ eloss and energy gain to additional controllable variables across linac
  - Expand to **longitudinal phase space tuning** (e.g. dynamic control over two bunch separation)
  - Expand to **jitter reduction**?
  - Continue expanding to **multiple plasma output metrics**

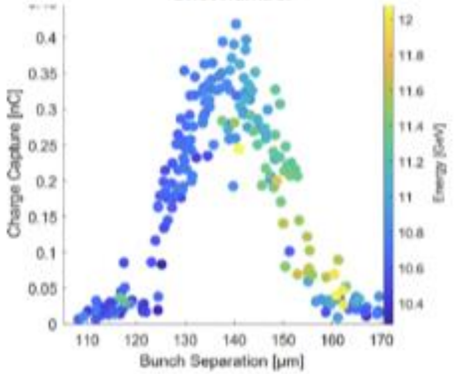
Will use trust region BO, neural network prior + BO, contextual BO, time-dependent BO

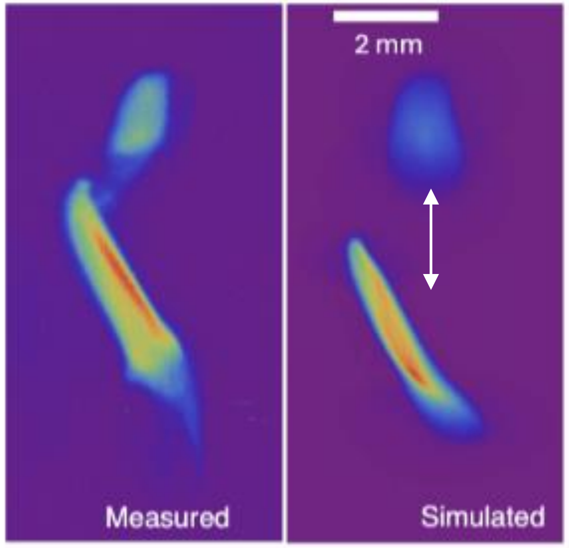
- **Incorporate Badger UI into FACET-II** and **expand work with other experiments** to leverage and improve these optimization tools, explore high-impact use-cases

- Finish **system model calibration** studies with injector and linac, expand to adaptive model calibration and use online for tuning

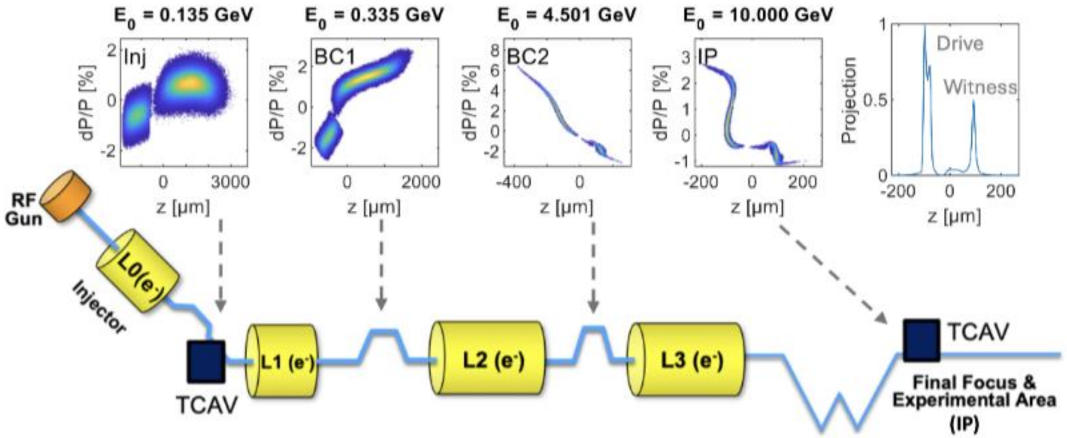
- Continue development on **reinforcement learning approaches for more comprehensive continuous control**, fast switching between setups







screen images



Desired facility upgrades

Computing

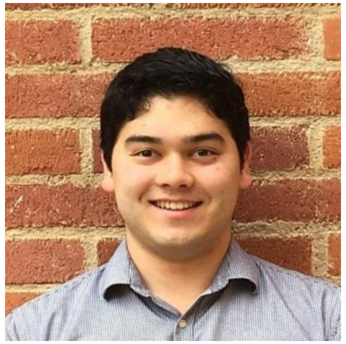
- GPU integration into controls network
  - *GPU purchased / on-site → now need to add to network (EED)*

- Working on getting read/write links between S3DF (on-site HPC), a different GPU system, and the controls network
  - *TID promises an initial solution this week*

Expect algorithms to be ***20-30x faster per iteration*** *once better compute is established*

→ critical step (right now computing is a major bottleneck for experiment time and TTO)








Thanks to the team involved in E331!

Questions?

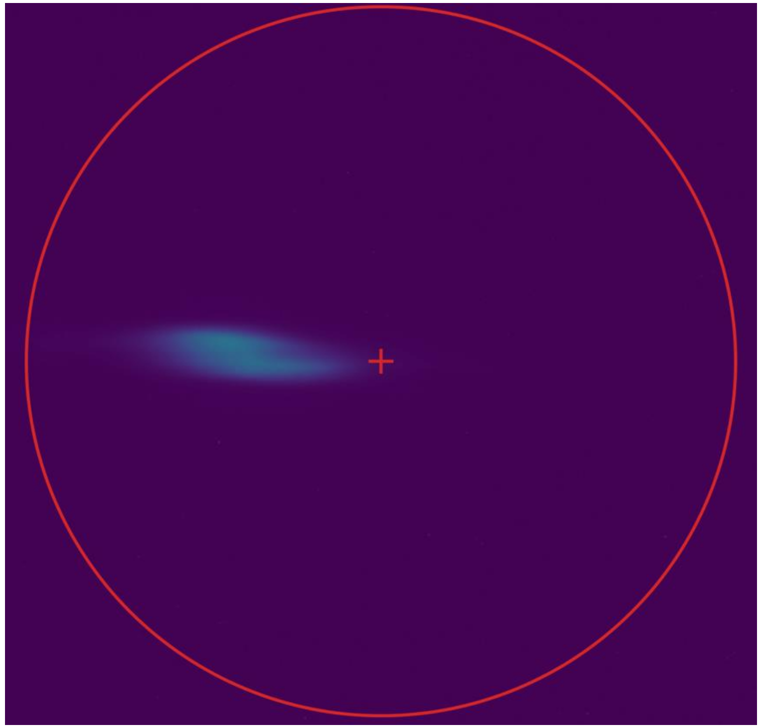
backups

Incorporating Constraints

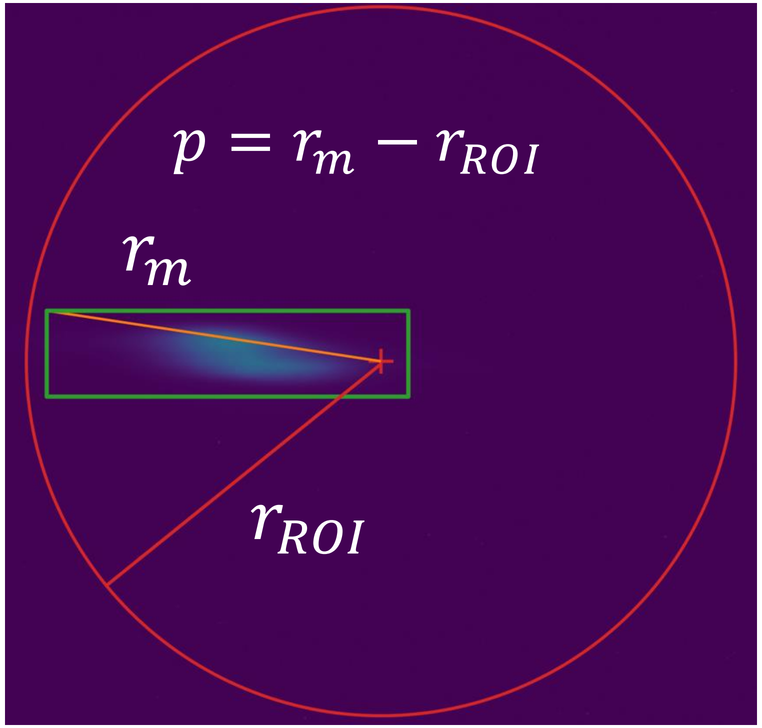
We want to ensure during measurements that the beam stays on screen

→ Define a **smoothly varying** penalty function to act as a constraint

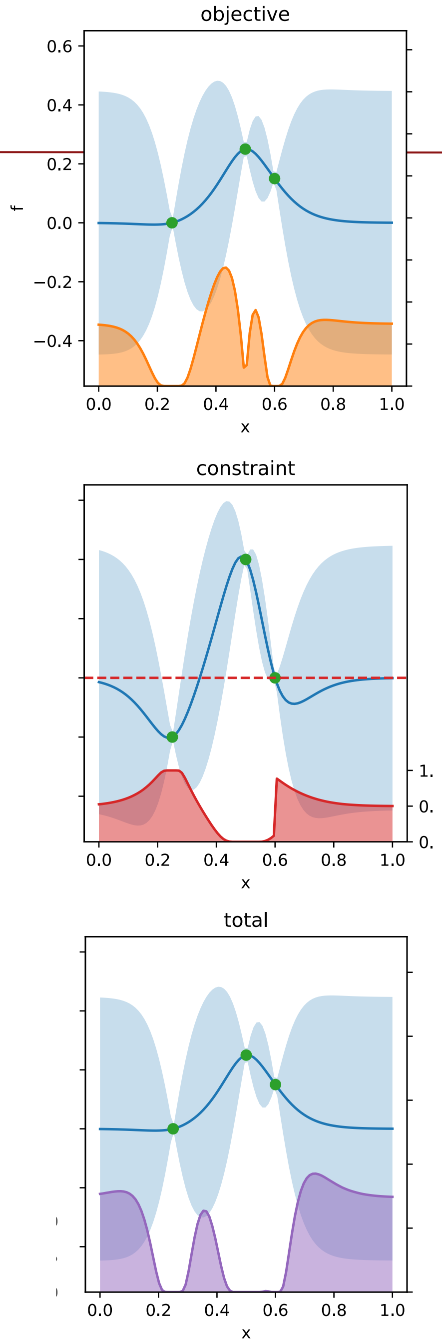
Define a circular ROI



Measure maximum distance from the ROI center to bounding box



Constraint: $p \leq 0$

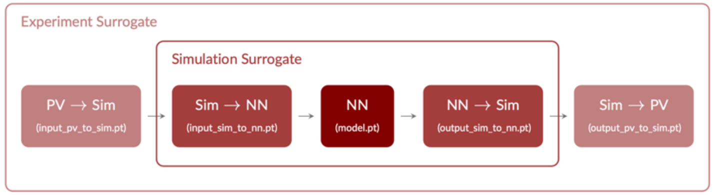


Other examples: Beam losses, dark current production, emittance, etc.

See R. Roussel et al., PRAB (2024) https://journals.aps.org/prab/abstract/10.1103/PhysRevAccelBeams.27.084801

Gardner et. al. ICML 2014

Digital Twin Infrastructure



Ecosystem of modular tools (can use independently)

LUME – simulation interfaces/wrappers in Python

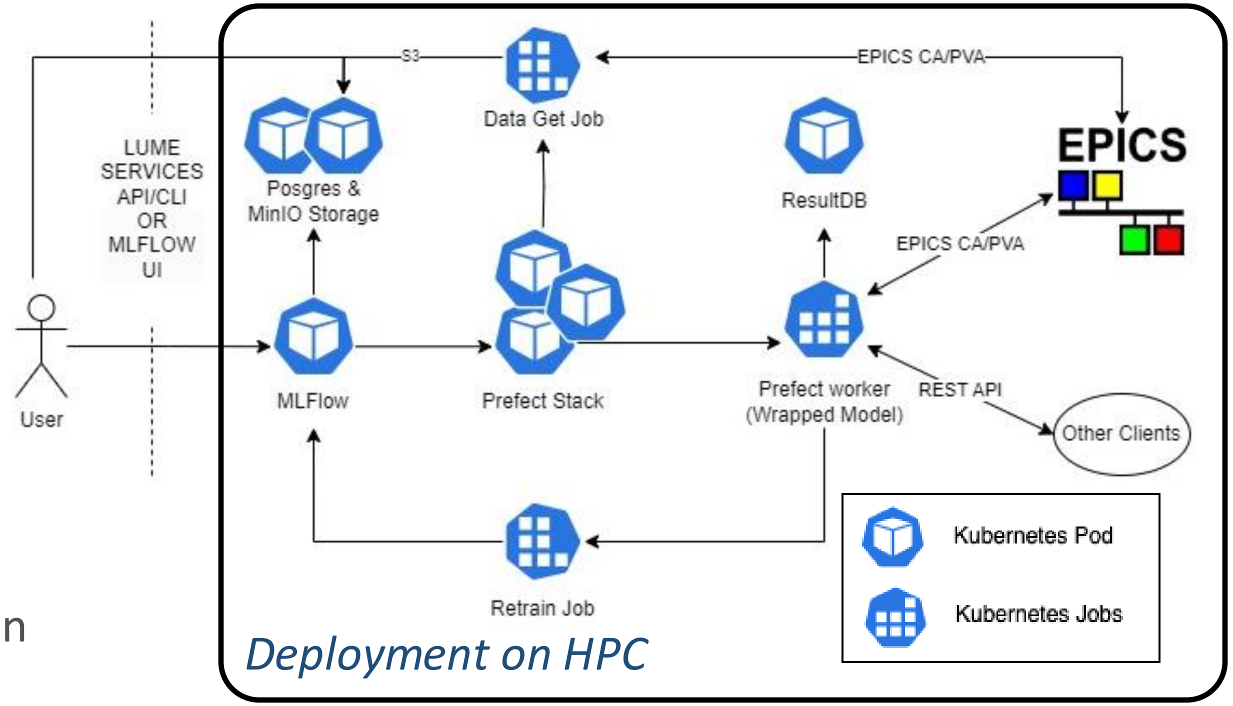
lume-model – wraps ML models, facilitates calibration

lume-services – online model deployment and orchestration

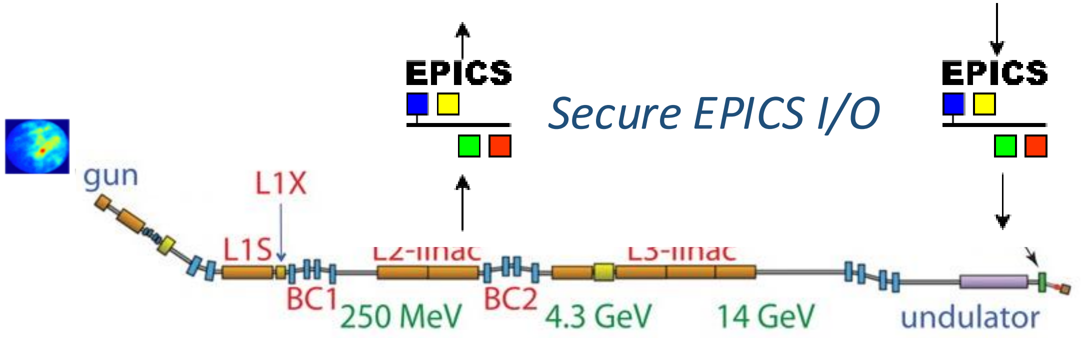
distgen – flexible creation of beam distributions

Integration with MLFlow for MLOps

https://www.lume.science/



- Live physics simulations and ML models now linked between SLAC's HPC system (S3DF) and control system
  → *run with Kubernetes and Prefect*
- Working with NERSC to swap between S3DF/NERSC resources
- Beginning work on MLOps aspects that will be used in continual learning research



Substantial progress on deploying ML and Physics-based models and integrating with HPC in a portable way

E331 Progress: ML for Efficient Characterization

R. Roussel et. al.
Nat. Comm. **2021**

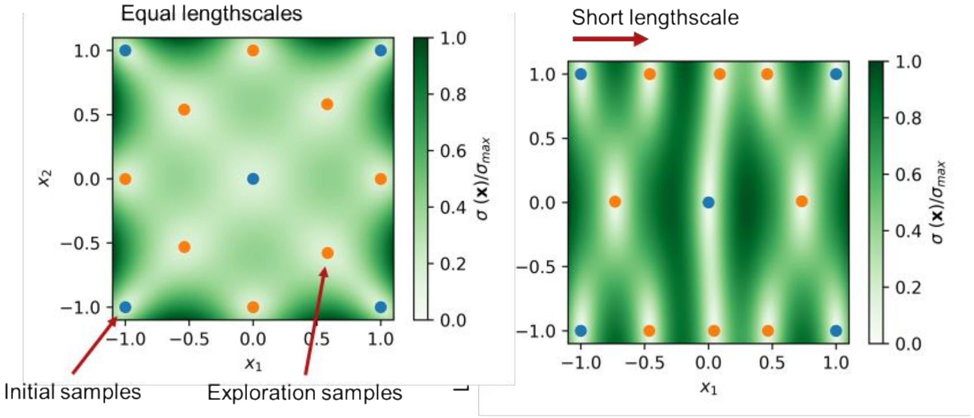
Better Data Sampling: Bayesian Exploration

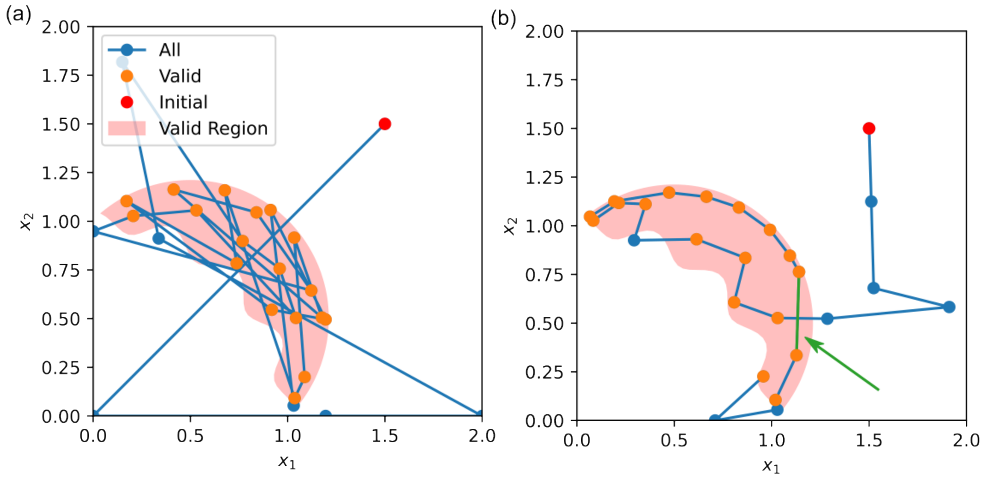
"
! (") = $(") % &(' !(") ≥ h!) Ψ(" , " %)
!#$

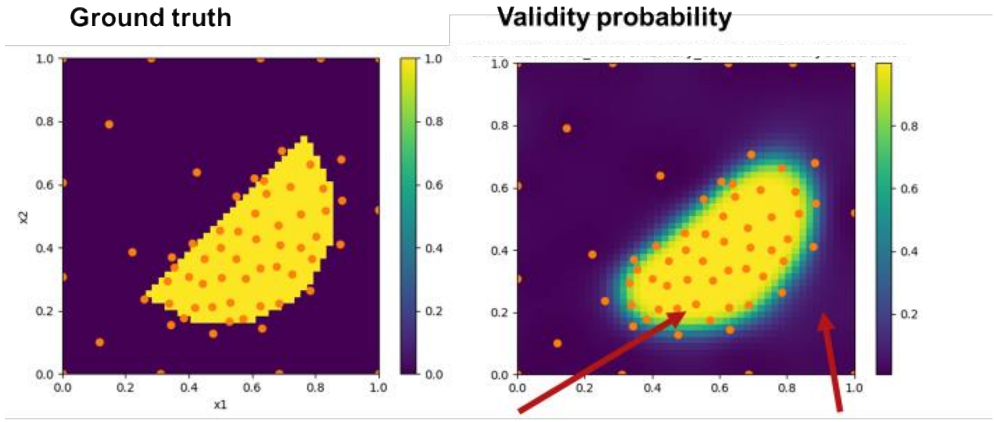
adaptive sampling

proximal biasing

learning constraints



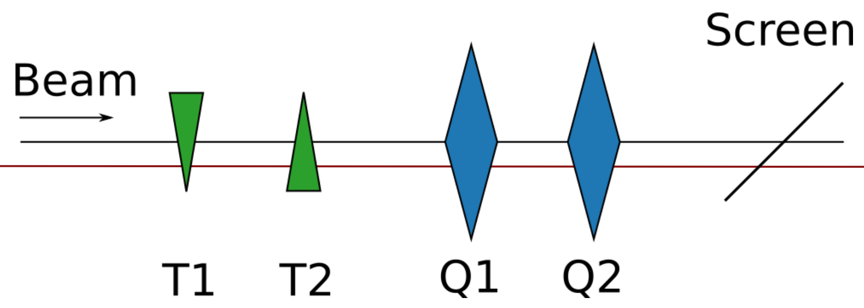




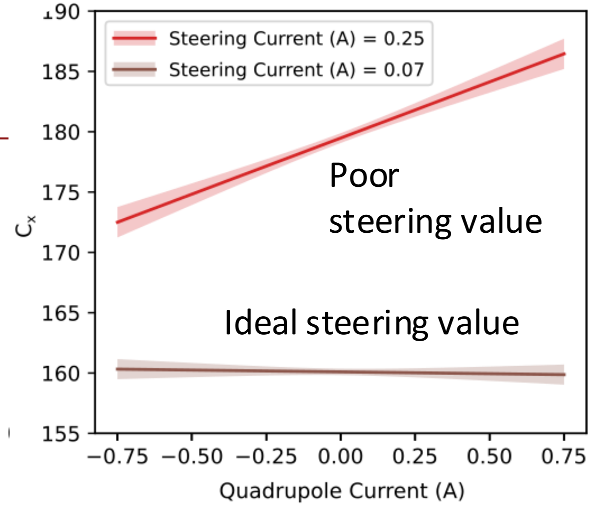
Enables sample-efficient characterization of high-dimensional spaces, while respecting both input and output constraints

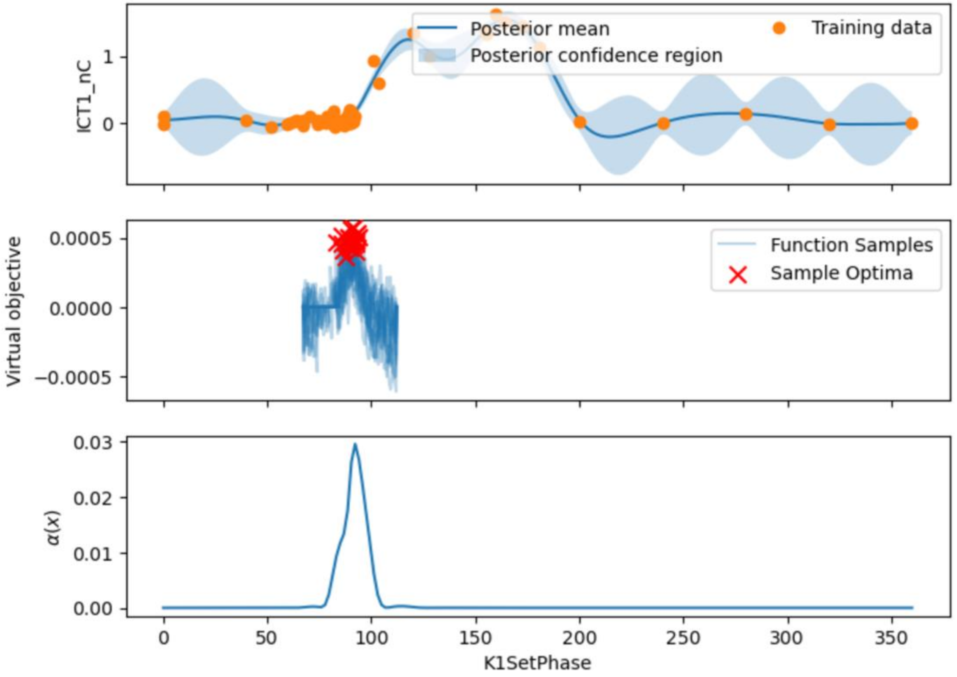
Further Automation

- Chaining together automation of sub-tasks and measurements
- RF /laser timing scans, beamline alignment, smart sampling for measurements

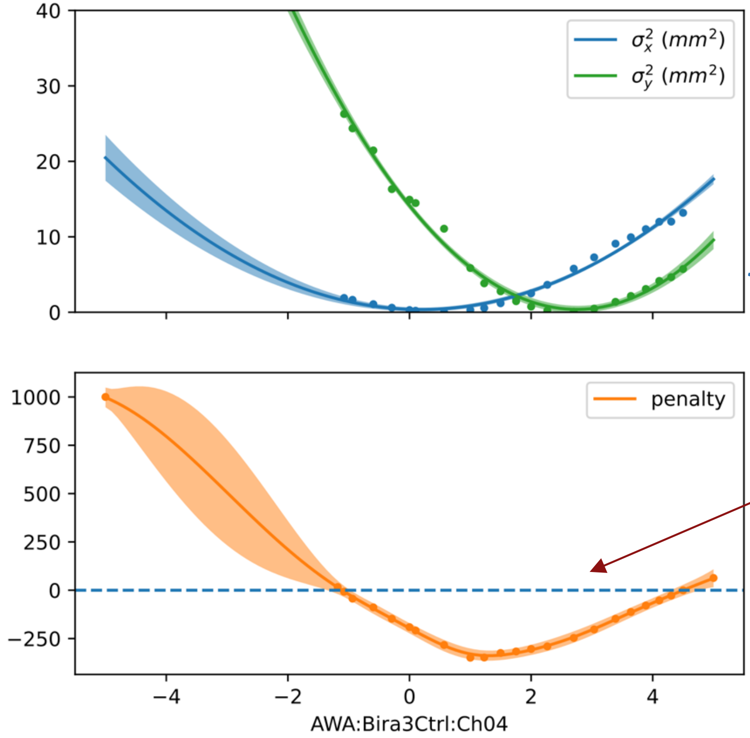


Automated beam alignment
→ 20-30 minutes by hand
→ 5 minutes with BAX





Automated determination of gun phase with BAX



Smart sampling for emittance measurements with Bayesian Exploration

Beam bounding box penalty

SLAC

R. Roussel, D. Kennedy

Synergies between ML experiments

Machine Control and Understanding

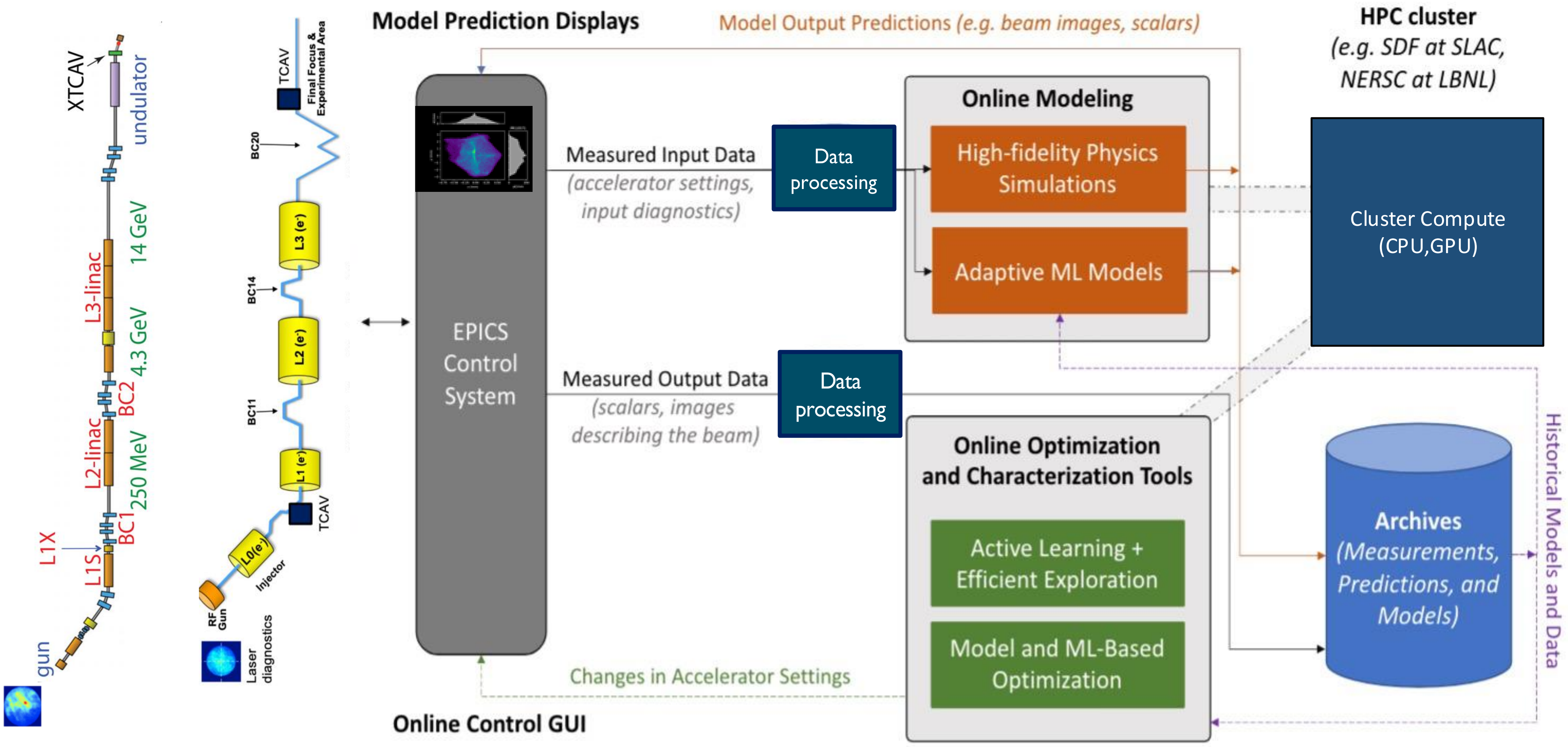
Diagnostics (Need information to make decisions)		Control (How to make decisions)	
Edge Radiation Emittance Diagnostics (E326)	Virtual TCAV Predictive Diagnostics (E327)	Adaptive Feedback (E325)	Learned Control (E331)
Non-destructive, single shot continuous monitoring of emittance of high-current beams	Longitudinal phase space diagnostics, always on, and for extremely short bunches	Stable, high-quality beams through control of unmodeled accelerator behavior	Fast, high-quality control of extreme beams by exploiting learned FACET-II responses

Synergistic experiments, individual success enhances all research

Goal: Full Integration of AI/ML Optimization, Data-Driven Modeling, and Physics Simulations

Working on a ***facility-agnostic*** ecosystem for online simulation, ML modeling, and AI/ML driven characterization/optimization

Will enable system-wide application to aid operations, and help drive AI/ML development *(e.g. higher dimensionality, robustness, combining algorithms efficiently)*



Making good progress toward this vision with open-source, modular software tools

Modular, Open-Source Software Development

- Community development of **re-usable, reliable, flexible software tools** for AI/ML workflows has been essential to maximize return on investment and ensure transferability between systems
- **Modularity has been key**: separating different parts of the workflow + using shared standards

Different software for different tasks:

Optimization algorithm driver *(e.g. Xopt)*

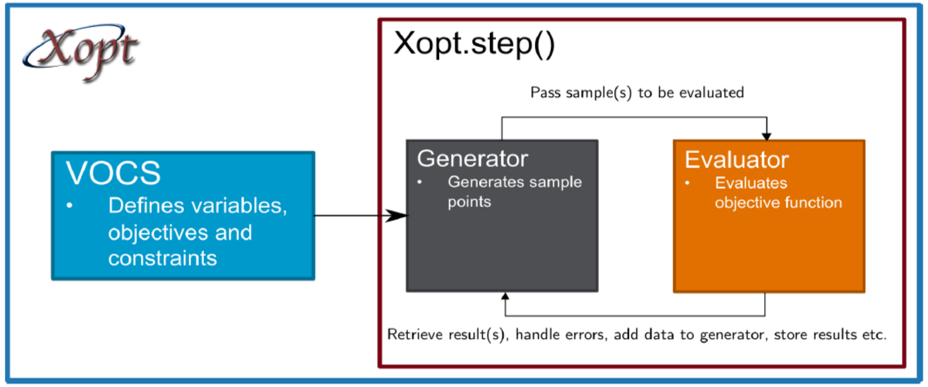
Visual control room interface *(e.g. Badger)*

Simulation drivers *(e.g. LUME)*

Standards model descriptions, data formats, and software interfaces *(e.g. openPMD)*

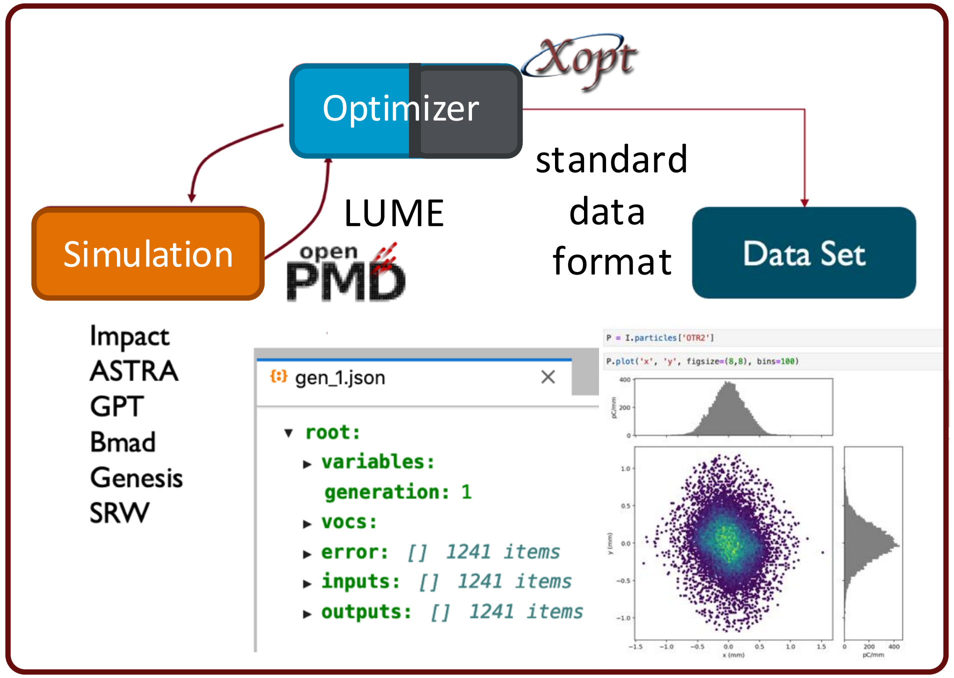
Online model deployment *(LUME-services)*

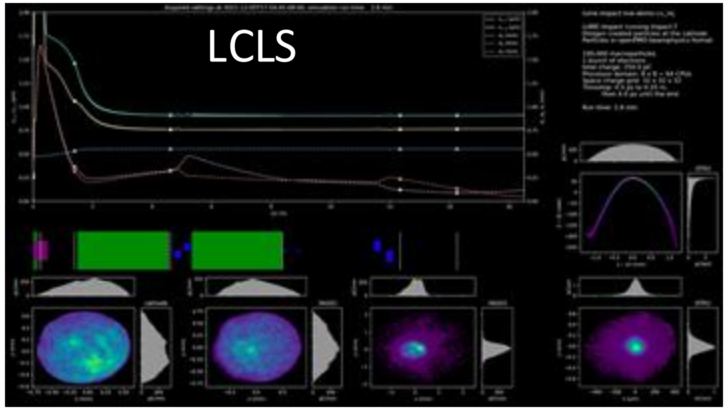
More details at https://www.lume.science/

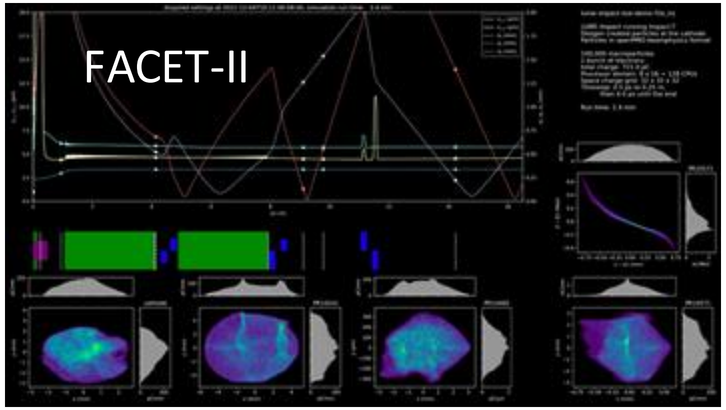



```
vocs:
  name: TNK_test
  variables:
    x1: [0, 3.14159]
    x2: [0, 3.14159]
  objectives: {y1: MINIMIZE}
  constraints:
    c1: [GREATER_THAN, 0]
    c2: ['LESS_THAN', 0.5]
```

```
algorithm:
  name: bayesian_exploration
  options:
      n_initial_samples: 5
      n_steps: 25
      generator_options:
          batch_size: 1
          #sigma: [[0.01, 0.0],
          use_gpu: False
```







Online Impact-T simulation and live display; trivial to get running on FACET-II using same software tools as the LCLS injector

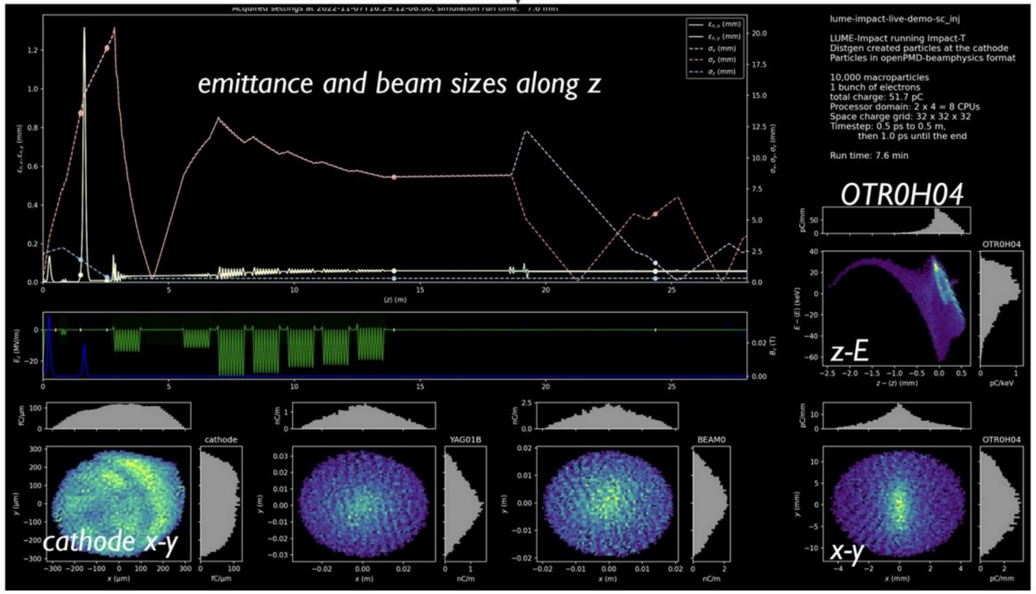
Modular open-source software has been essential for our work. We welcome new users and contributors.

Example: Online Models and Bayesian Optimization in Operations

Used combination of online physics simulation and Bayesian optimization algorithms to aid LCLS-II injector commissioning

Readings from machine via EPICS

injector settings, laser profile from VCC image

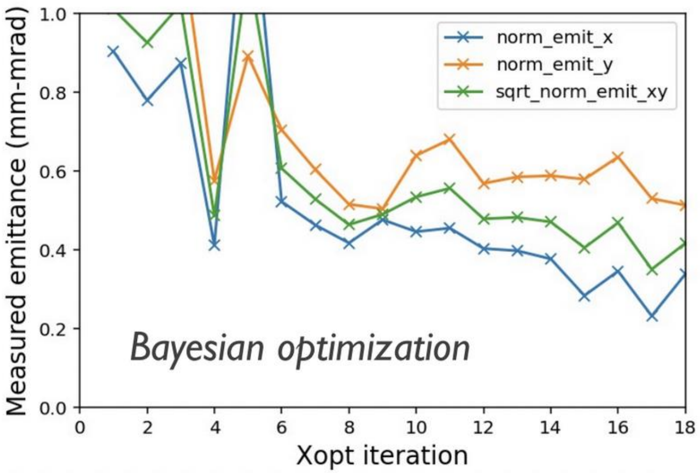


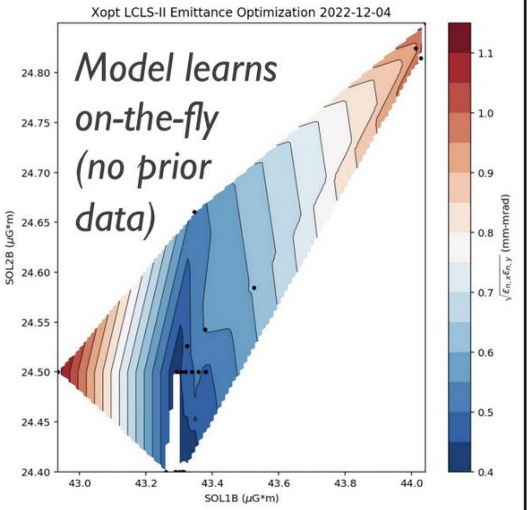
LCLS-II live sim: run on HPC and display in control room

Updates every 3-8 mins, space charge included, uses LUME-IMPACT

Adjust settings / ranges with insight from predictions

Hand over to ML-based optimization for fine tuning





Best emittance yet obtained during LCLS-II injector commissioning

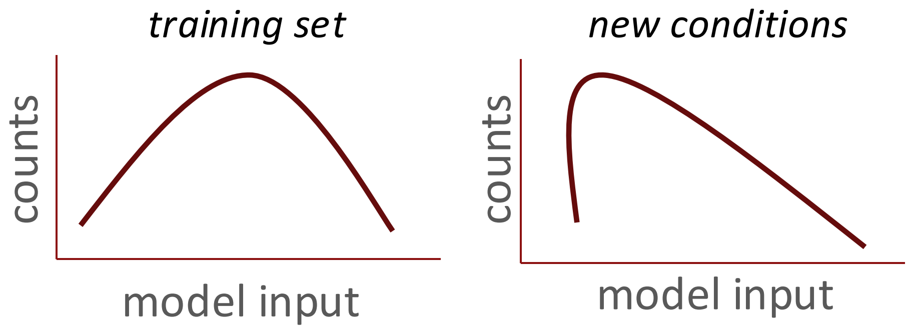
despite extensive previous hand-tuning

Physicists' intuition aided by detailed online physics model → simple example of how a "virtual accelerator" can aid tuning

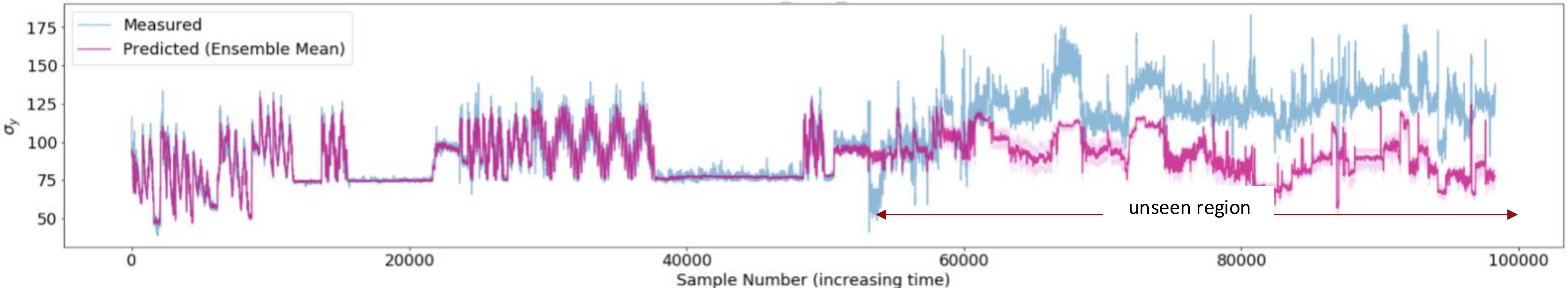
HPC enables fundamentally new capabilities in what can be realistically simulated online

Uncertainty Quantification / Robust Modeling / Model Adaptation

- Major area of AI/ML research: statistical distribution shift between training and test data degrades prediction
- Distribution shift is extremely common in accelerators, due to both deliberate changes in beam configuration and uncontrolled or hidden variables



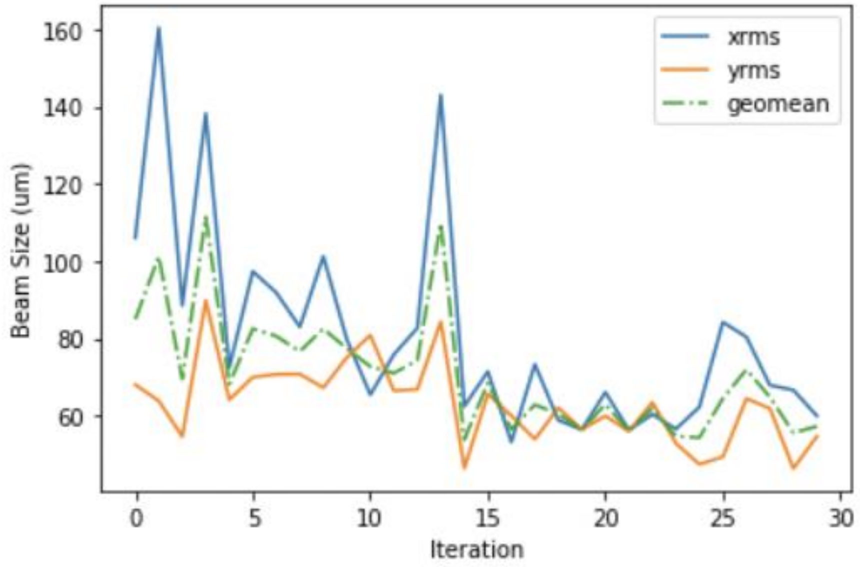
Example: beam size prediction and uncertainty estimates under drift from a neural network
Uncertainty estimate from neural network ensemble does not cover prediction error, but does give a qualitative metric for uncertainty

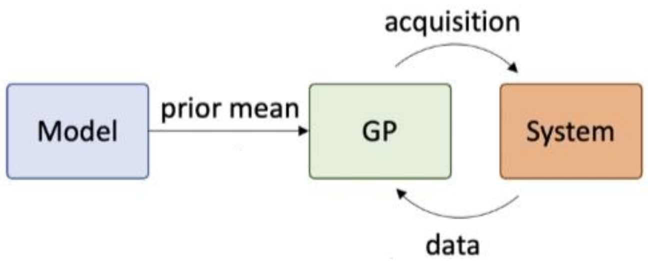


Reliable uncertainty estimates and model adaptation methods are key for putting online models to use operationally

Optimization of Sextupoles for Spot Size at IP

- Ran constrained Bayesian optimization on the sextupole movers (8 variables total) to minimize spot size as measured on the wires in S20
- Recorded auxiliary data (TCAV and EOS, BSA)
- First step toward more comprehensive tuning in S20
- Used software, Xopt, established for previous runs with little need for adjustment to this specific problem → *nice demonstration of extensibility*



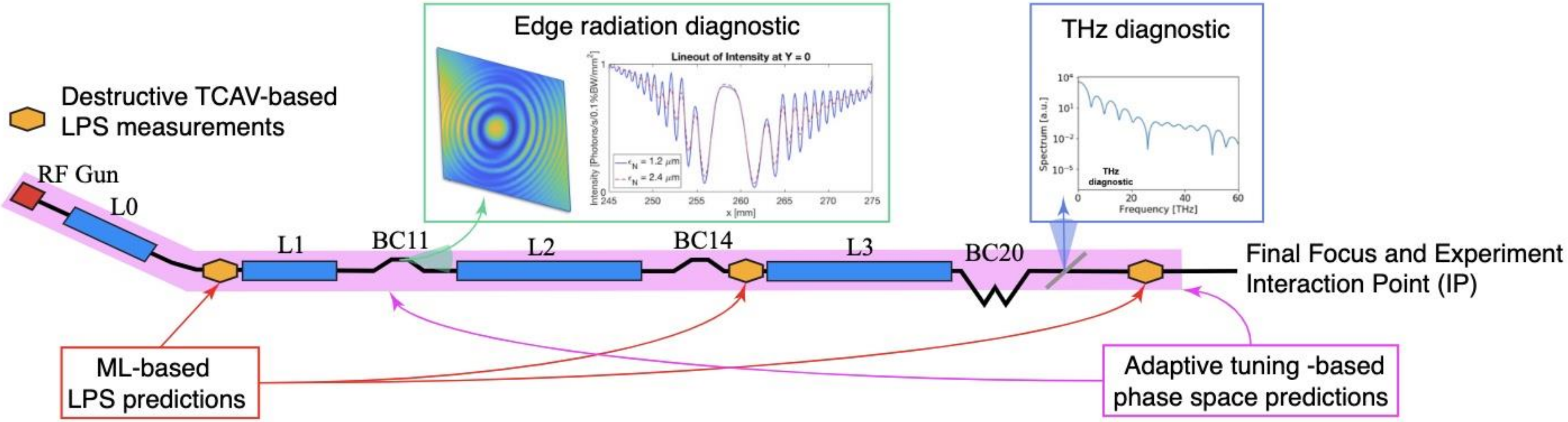


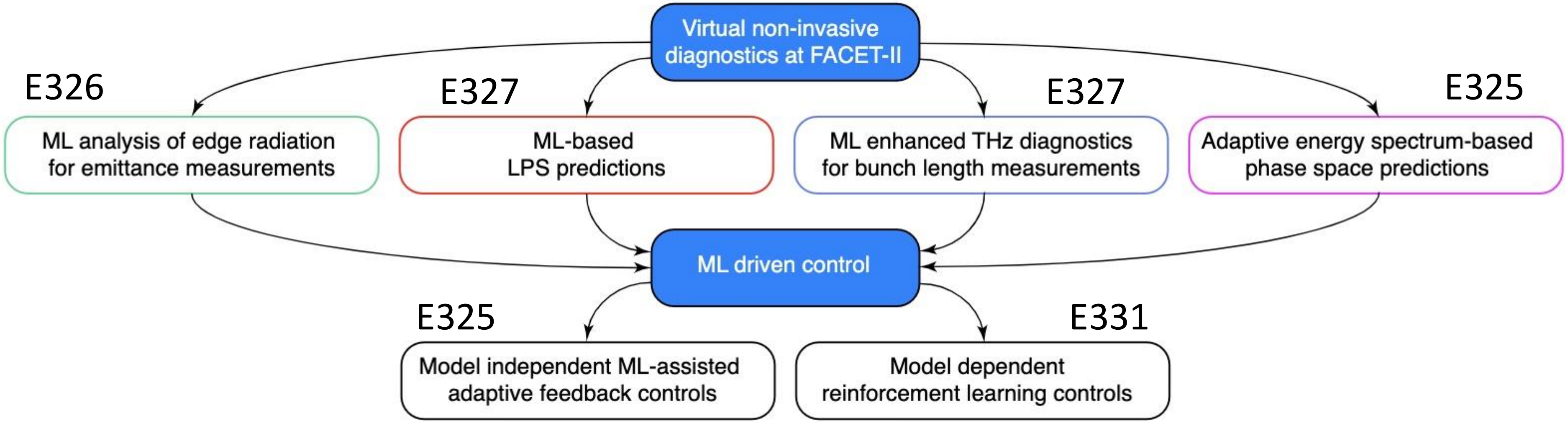
Next:

- Want to use on both IPs (with multi-objective optimization) and use greater number of variables
- Use data to inform faster subsequent optimization

Automatically tuned for a small, round beam at the IP using sextupole movers. Ready for next steps in tuning both IPs and with broader set of variables.

Landscape of AI/ML Activities at FACET-II

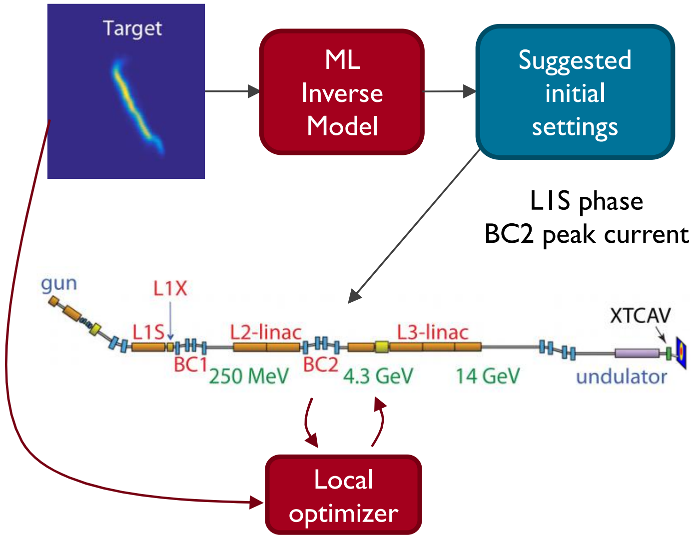




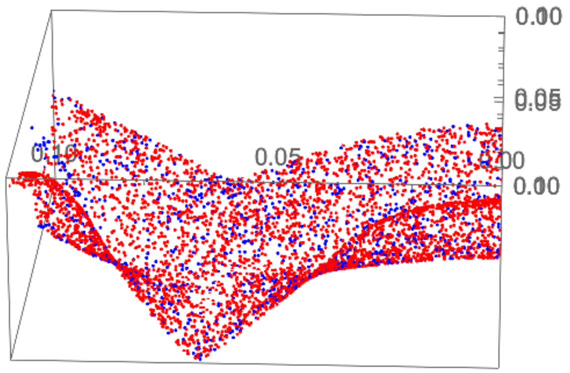
Synergistic experiments, individual success enhances all research + facility operation

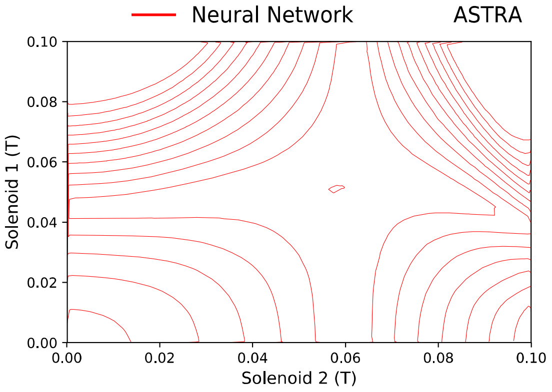
Warm starts for optimization

A. Scheinker, A. Edelen, et al, PRL, 2018

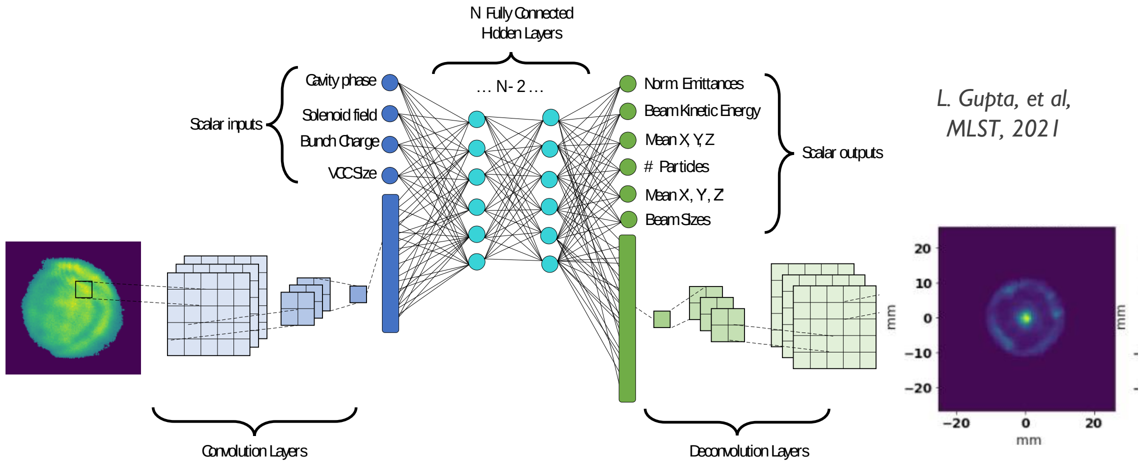


Smooth interpolation
Example σ_x surface from 2D scan, LCLS-II Injector



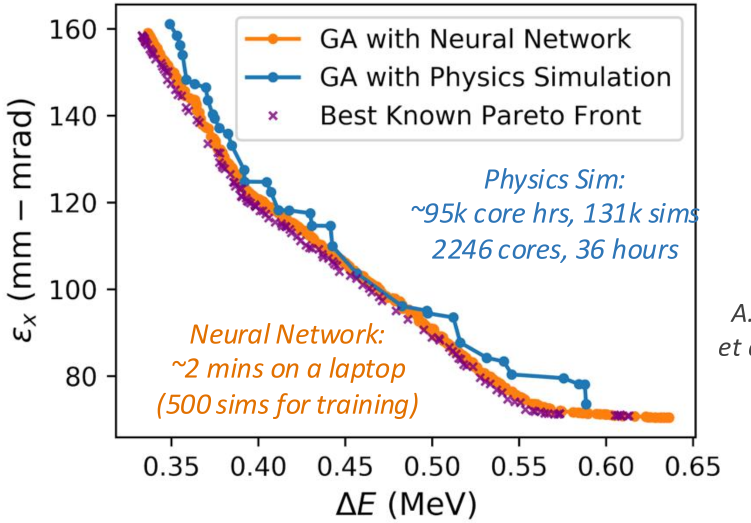


A. Edelen et al., NeurIPS 2019



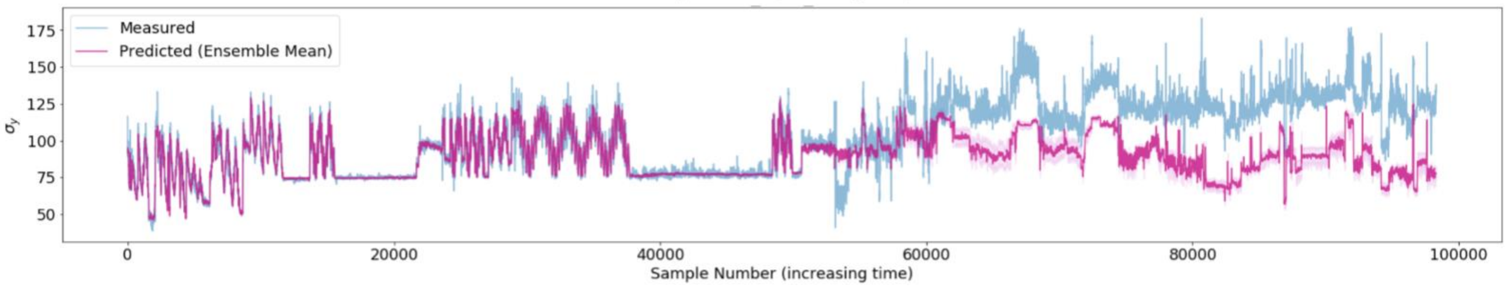
L. Gupta, et al, MLST, 2021

Surrogate-boosted design optimization



A. Edelen et al., PRAB, 2020

Include high-dimensional input information à better output predictions



Relative uncertainty estimates indicate when to retrain