E331: FY23 Progress and Plans for FY24

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

---

Auralee Edelen on behalf of E331 / SLAC National Accelerator Laboratory

FACET-II PAC Meeting, 18 October 2023, SLAC

FACET-II

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University | U.S. DEPARTMENT OF ENERGY

E331 Science Motivation

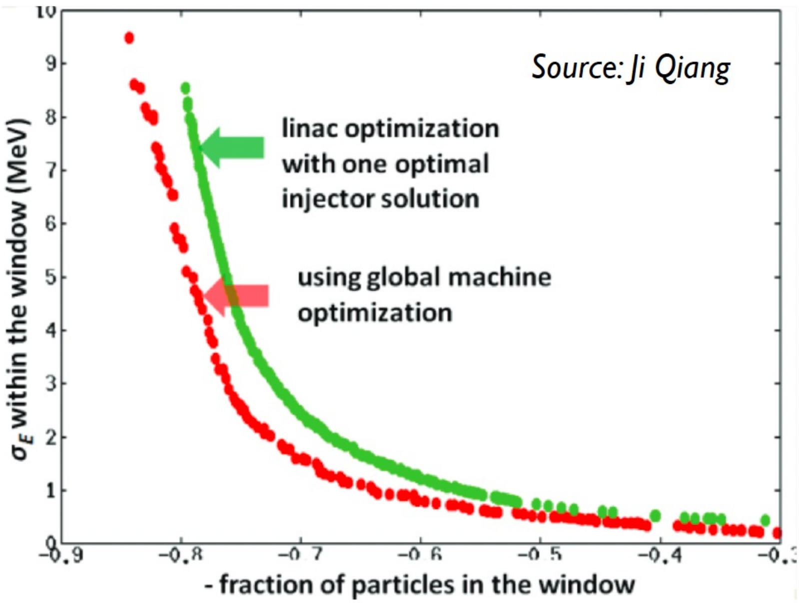
Major limitations in the way accelerator tuning is done:

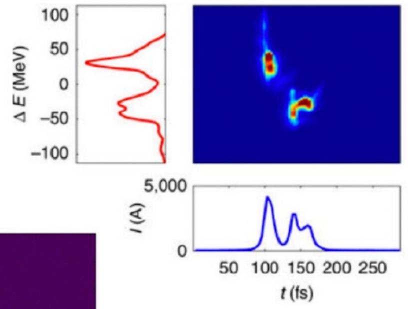
- Piecemeal tuning of subsystems *(known to be sub-optimal)*
- Indirect use of high-dimensional diagnostics *(e.g. images)*
- Often a lack of accurate online models

  → *Potentially limiting factors in control of extreme beams*

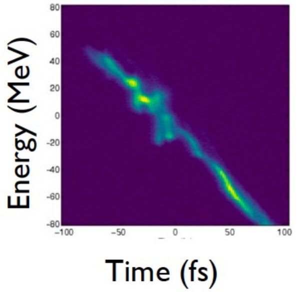
More global view can enable better control:

- 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)



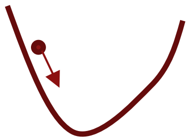
A. Marinelli, IPAC'18

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

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

less ← assumed knowledge of machine → more

Model-Free Optimization

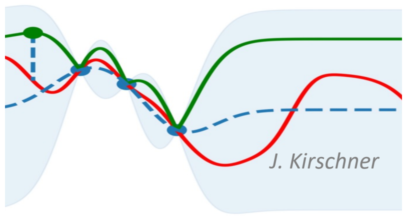


Observe performance change after setting adjustments

→ estimate direction or apply heuristics toward improvement

gradient descent
simplex
ES

Model-guided Optimization

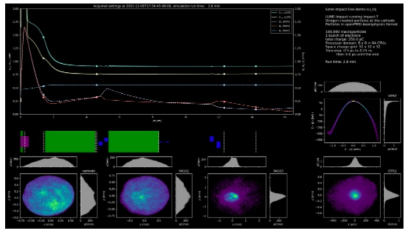


Update a model at each step

→ 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

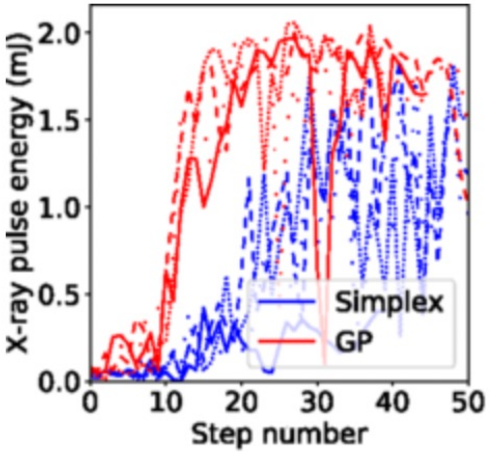
Tuning research aimed at combining the strengths of different approaches.

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

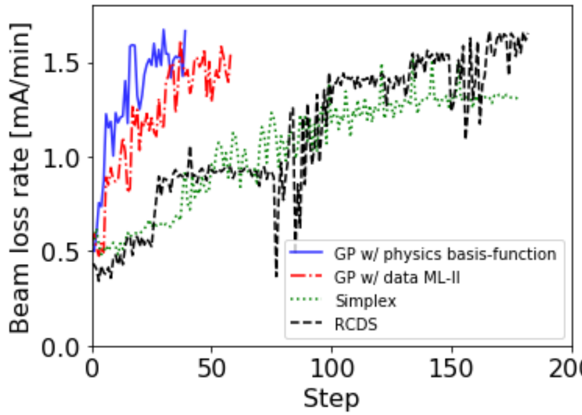
(+ improvements)

FEL pulse energy tuning at LCLS



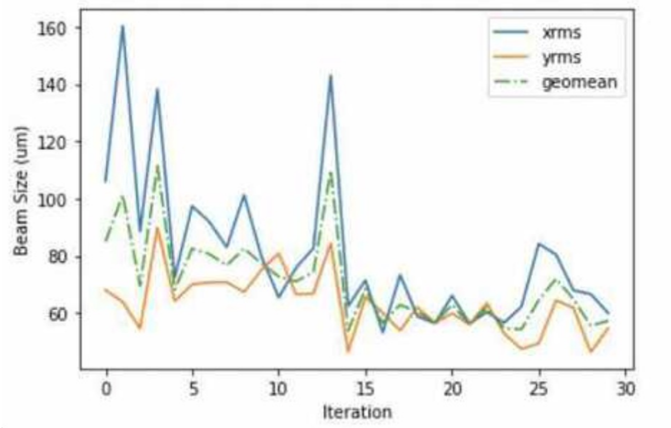
*Duris et. al. PRL , **2020***

Loss rate tuning at SPEAR3

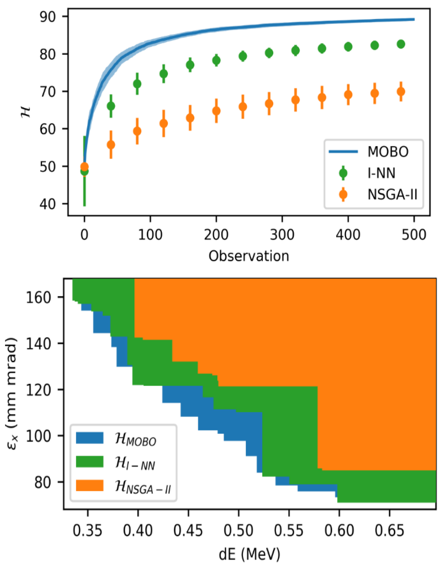


*Hanuka et. al. PRAB , **2021***

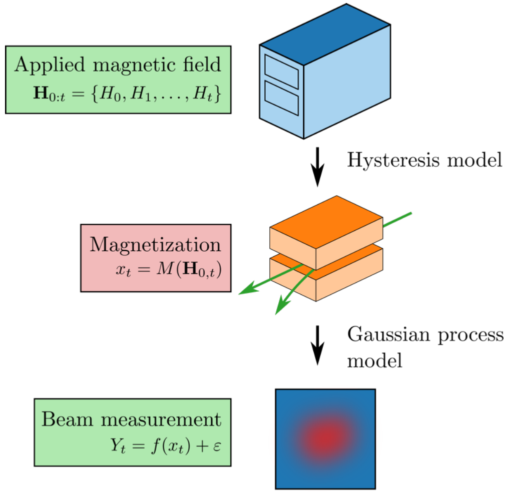
Sextupole tuning for IP at FACET-II



Multi-objective Bayesian Optimization

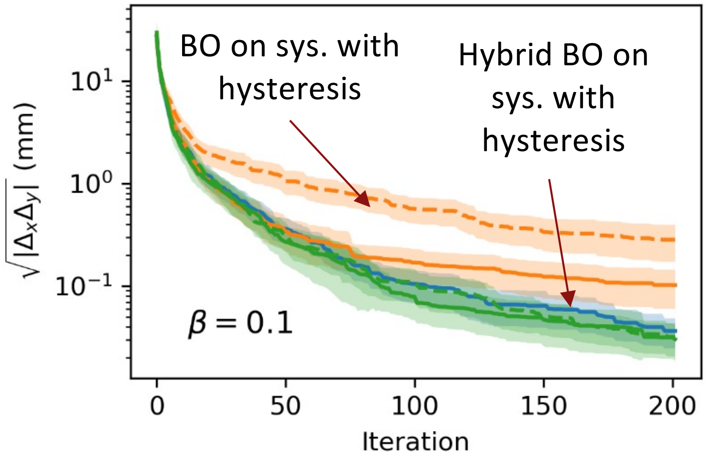


*Roussel et. al. PRAB , **2021***

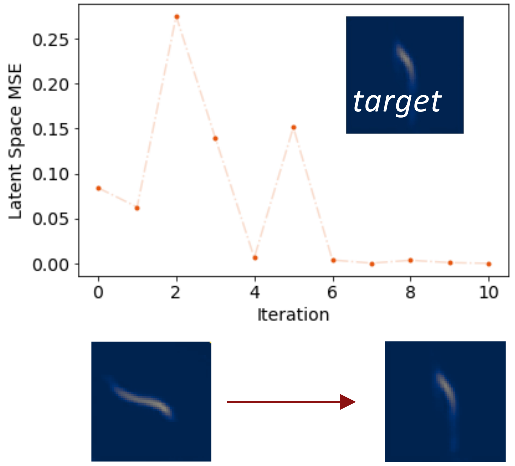


*Roussel et. al. PRL , **2022***

Higher-precision optimization possible when including hysteresis effects in model



Longitudinal phase space tuning on LCLS

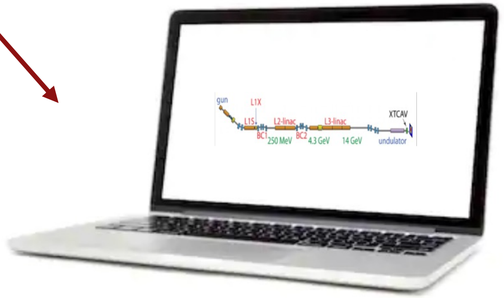


Algorithms being implemented/distributed in Xopt: https://github.com/ChristopherMayes/Xopt

Fast-Executing, Accurate System Models

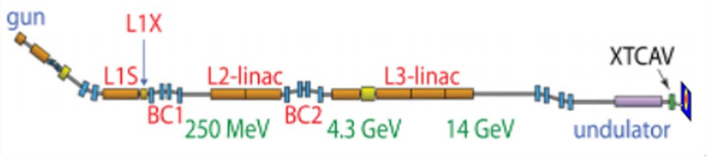
Bringing simulation tools from HPC systems to online/local compute

Online prediction
Model-based control



Control prototyping
Experiment planning

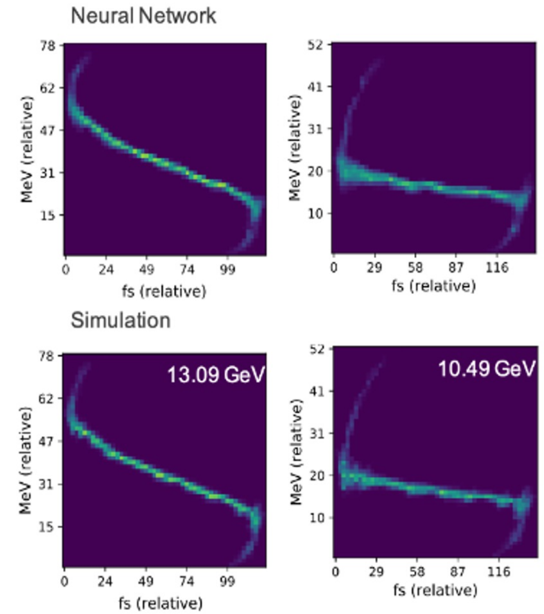
ML models are able to provide fast approximations to simulations ("surrogate models")



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

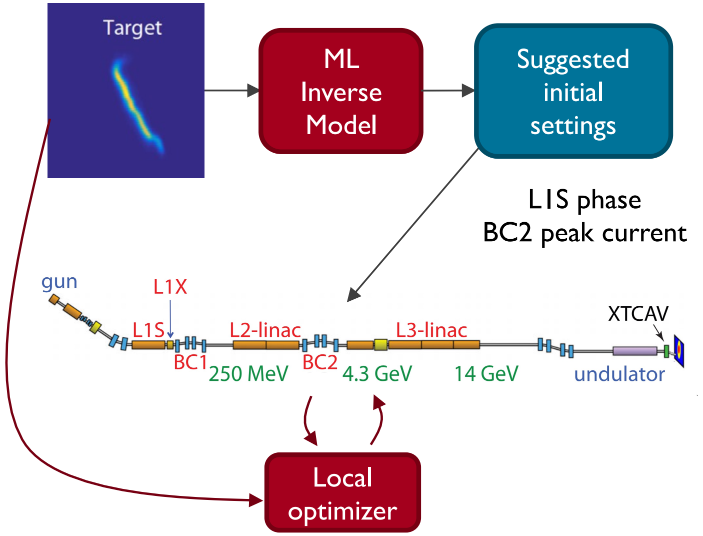
10^6 times speedup

Edelen et al., NeurIPS 2019

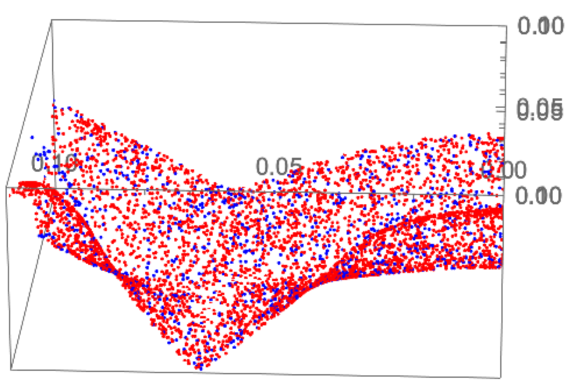
ML modeling enables accurate predictions of system responses with unprecedented speeds, opening up new avenues for high-fidelity online prediction, tracking of machine behavior, and model-based control

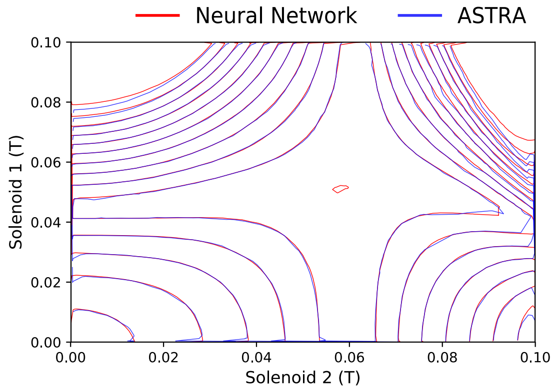
Warm starts for optimization

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

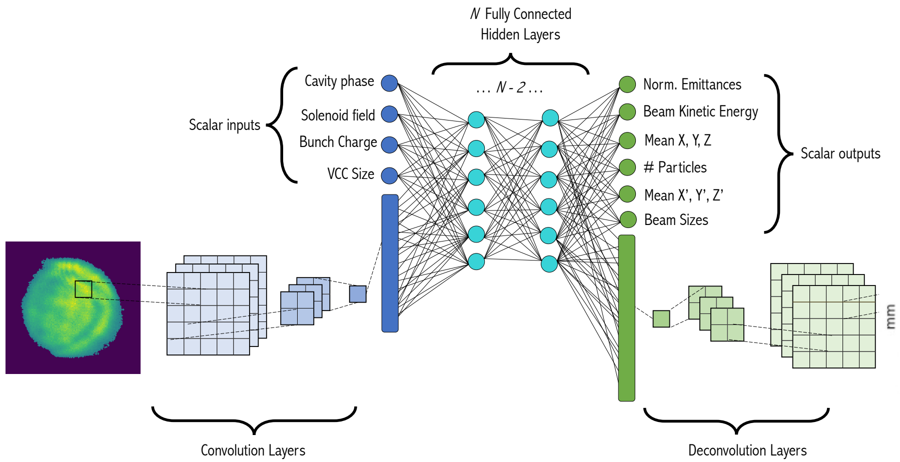


Smooth interpolation
Example σ_x surface from 2D scan, LCLS-II Injector

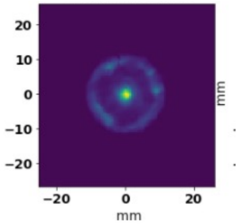




A. Edelen et al., NeurIPS 2019

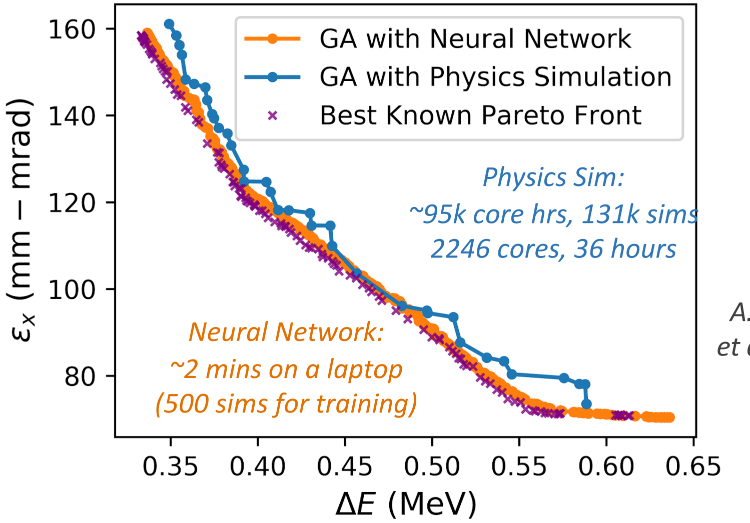


L. Gupta, et al, MLST, 2021

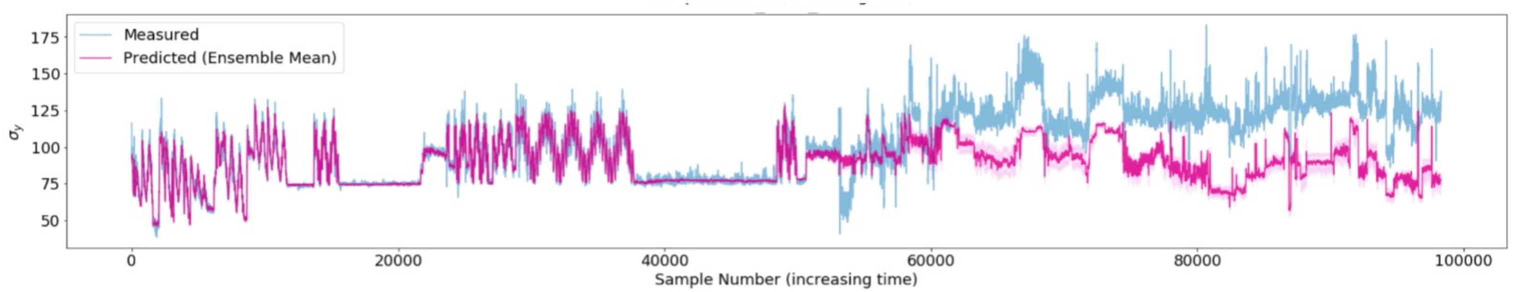


Include high-dimensional input information → better output predictions

Surrogate-boosted design optimization

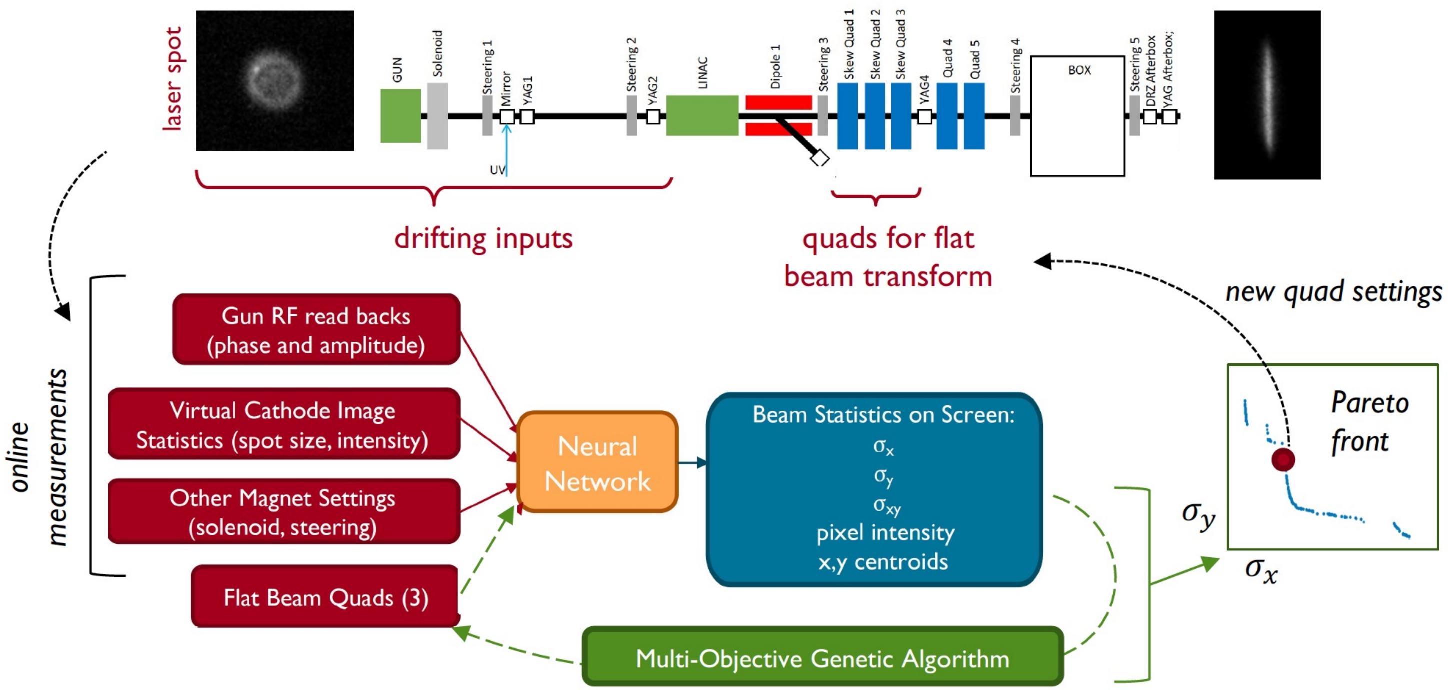


A. Edelen et al., PRAB, 2020

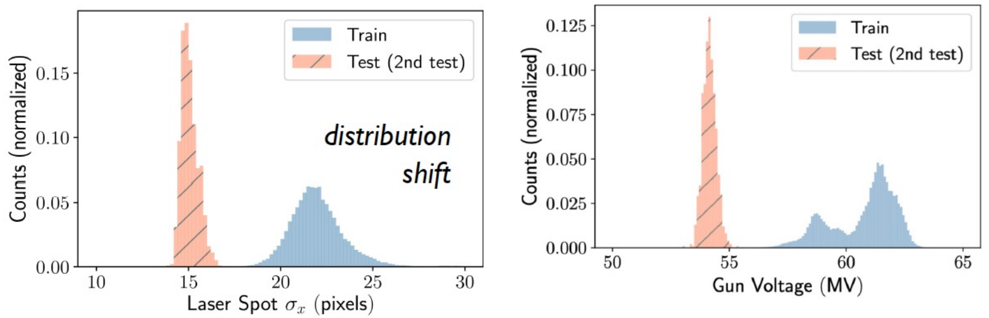


Relative uncertainty estimates indicate when to retrain

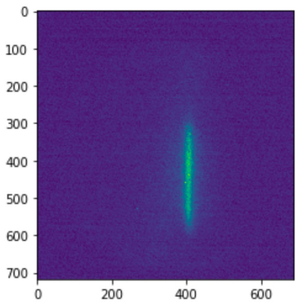
Example: Warm Starts from Online Models

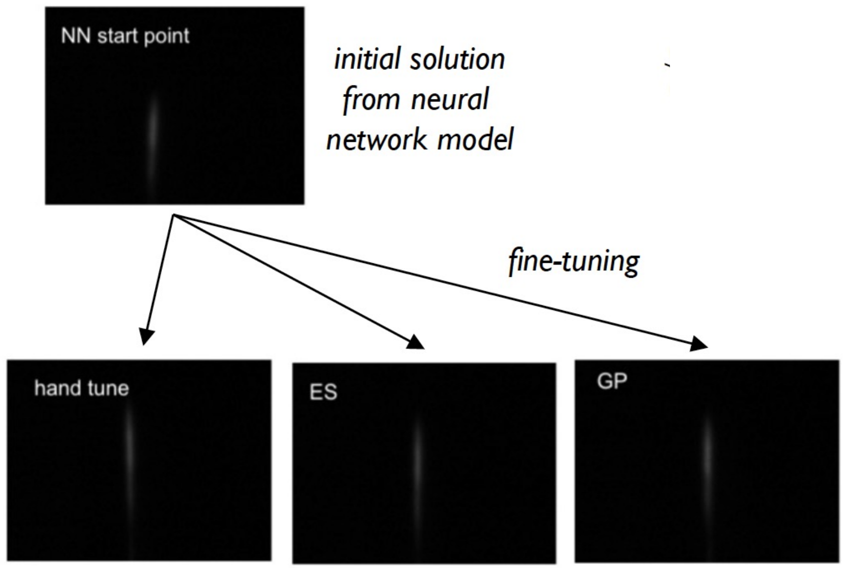


Can work even under distribution shift



- 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 DDPG Reinforcement Learning agent and tested on machine under different conditions than training



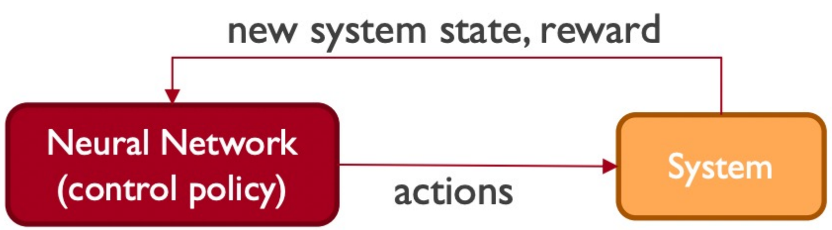


Hand-tuning in seconds vs. tens of minutes

Boost in convergence speed for other algorithms

SLAC

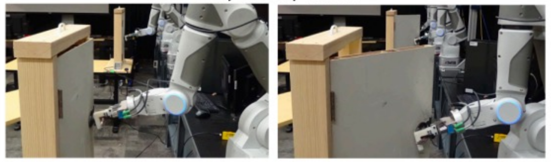
Deep Reinforcement Learning



- Control policy maps states to actions
- Policy is learned over time based on performance *(quantified by the "reward")*
- Neural network enables use of diverse signal types *(e.g. scalars, images, time series)*
- Often learns a system model simultaneously *(map states + actions to expected reward)*

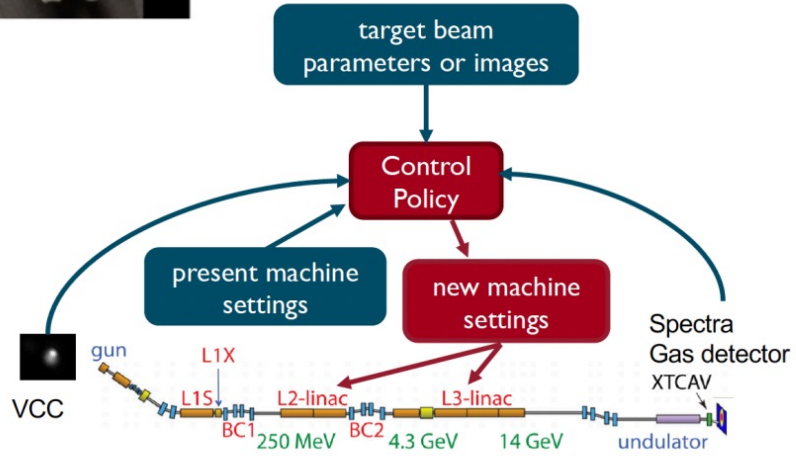
Nagabandi, et al., 2019

Gu, et al., 2016



Appeal for accelerator control:

- Suitable for large, nonlinear systems
- Exploit machine-wide sensitivities + directly use complicated diagnostic information
- Leverage information from past observations
- Transfer between similar designs
- Well-established in other fields (e.g. robotic control)
  → but accelerators have unique challenges



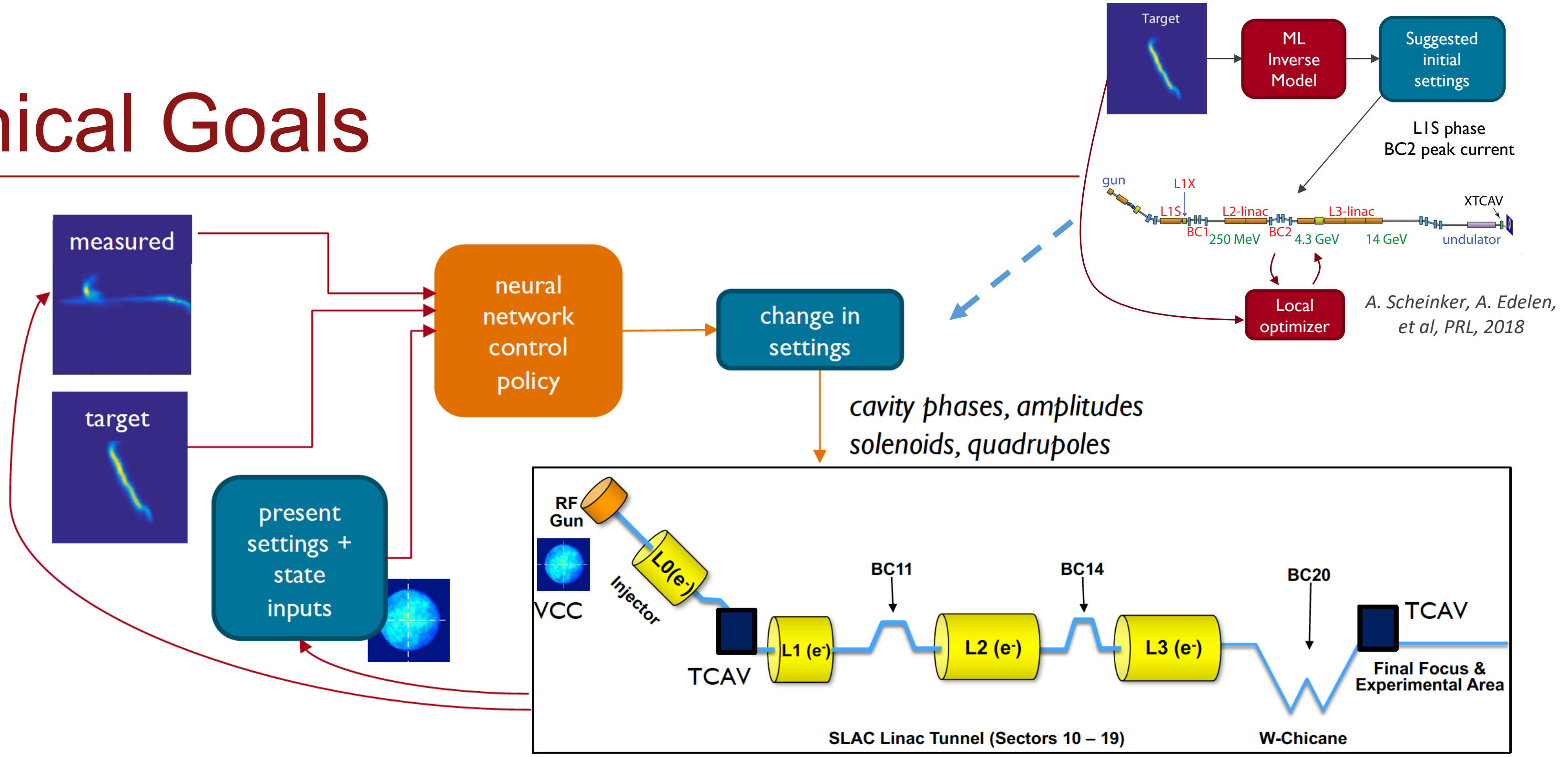
Deep RL is well-suited to accelerator control, but dedicated R&D is needed to bring it to full fruition

SLAC

E331 Science/Technical Goals

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**

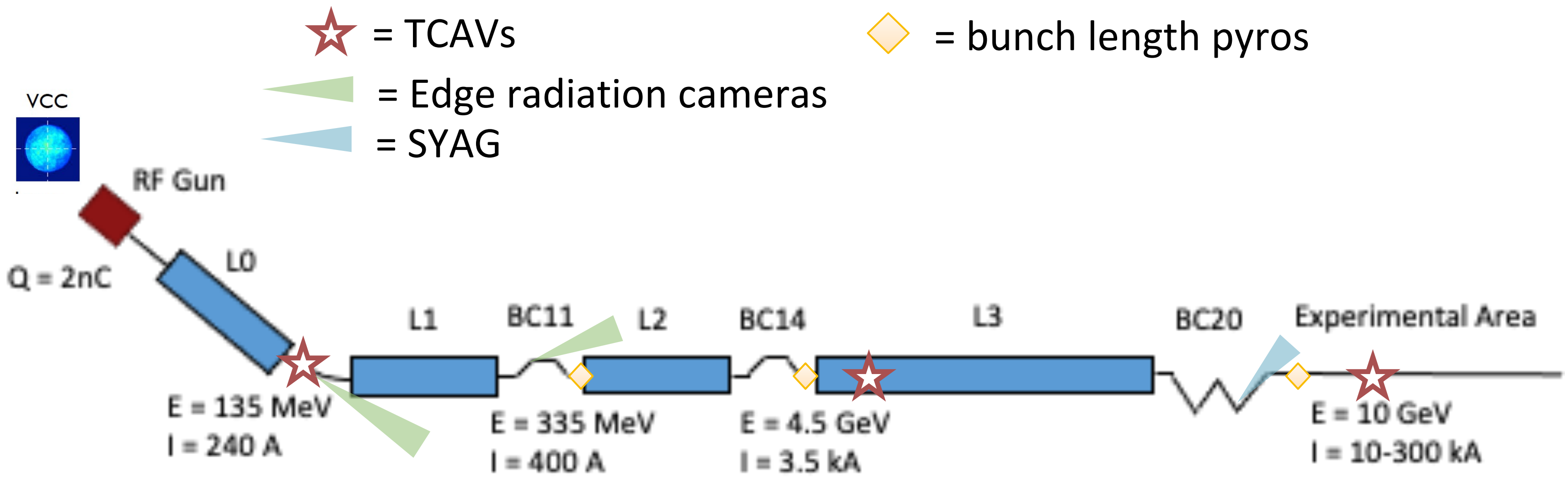
(build up incrementally to machine-wide neural network-based reinforcement learning)



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	1-3 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 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 transition into operations

E331 Diagnostic and Observables



Similar diagnostic needs to E327

- 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: ~150 scalar diagnostics (e.g. BPMs, toroids, RF readbacks, BLEN pyros) and multiple image diagnostics (SYAG, EOS, TCAV)

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

Numerous diagnostics to inform tuning or be used as tuning targets

FY22-FY23 Progress - shift timeline

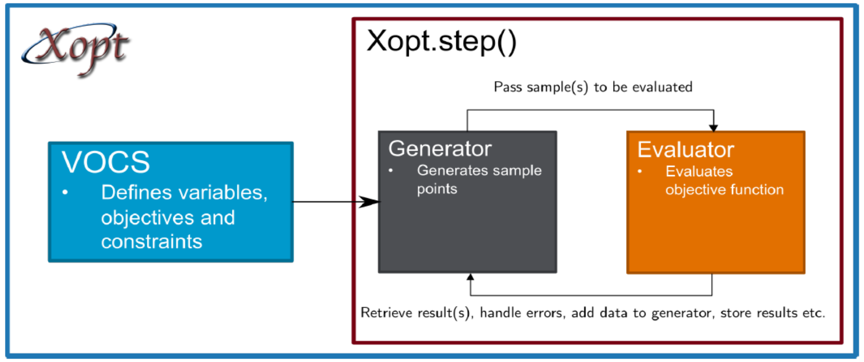
Shift Summary Date	Experiment Num	Shift Start Time	Shift End Time	Useful Beam Time	Accelerator Downtime	User Downtime	Brief Summary	
11/17/21	E327	11/17/21 11:44	11/17/21 17:44	6	0	0	Tested software. Ran ND scan characterizing injector emittance vs sol,buck, cq,sq	
11/20/21	E327	11/19/21 20:12	11/20/21	12	0	0	Gathered training data for ML optimiization of injector emittance. Tested software.	
11/29/21	E327	11/28/21 18:17	11/29/21	8	0	0	Test Bayes Exp for injector emittance	
12/4/21	E331	12/3/21 12:00	12/4/21 0:00	12	0	0	Ran Bayes Exp on emittance + bmag	
12/11/21	E331	12/10/21 20:25	12/11/21 12:25	16	0	0	Ran Bayes Exp on emittance + bmag	
12/17/21	E327	12/16/21 20:13	12/17/21 8:13	12	0	0	TCAV measurements scanning L2 phase data gathered. Inj opt data gathered with match	TCAV
2/27/22	E331	2/27/22 11:59	2/27/22 23:59	12	0	0	Compared opt methods for injector emittance + match at new laser wavelength of 253nm with 266 nm prior data	
5/14/22	E331	5/13/22 18:02	5/14/22 5:02	11	0	0	Characterize emittance at 1.8 nC with Bayes Ex. Optimize with BO + other methods and gather comparative data	
8/22/22	E331	8/21/22 14:30	8/22/22 2:30	6	6	0	Ran Bayesian Optimization on Sextupole movers. Gathered TCAV, EOS and wire scanner data at different sextupole mover positions	TCAV
			Sum Total Hrs	95	6	0		

- Shared beam time with E327
- Deployed initial software tools for measurements and optimization
- Characterized injector under different charge settings and laser parameters
- Tested new ML algorithms for efficient characterization and tuning (applied to injector emittance and IP spot size tuning)
- Next steps: continue scaling up + use data gathered to move toward more comprehensive model-based approaches; incorporate TCAVs in tuning

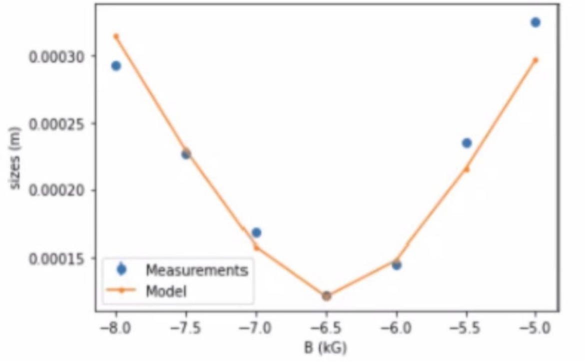
First shifts demonstrated utility of ML optimization tools → data gathered will be used in next phases of project

E331 Progress: Practicalities and Infrastructure

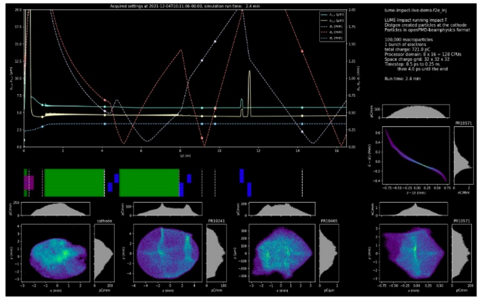
- Vetted **adaptive emittance measurement method** for use in automated emittance optimization (PyEmittance) https://github.com/slaclab/PyEmittance
  - *Need to re-evaluate in new machine config, extend to downstream emittance measurements*
- Integrated **Xopt into FACET-II control system** → *aids algorithm transfer between systems and will make it easy to test new algorithms on FACET-II*
- **Deployed online LUME-IMPACT model of injector** (live reading from machine and making predictions)
  - *Particle-in-cell code includes space charge, uses VCC image*
  - *Same infrastructure for deploying online ML models we plan to use in model-based tuning*
- Next steps: **Badger user interface for optimization** *(also saves tuning runs → useful data for developing model-based algorithms)*



Xopt running on FACET-II for easy ML algorithm deployment on different tuning problems



Adaptive quad scan emittance measurement deployed for robust measurements



FACET-II Injector model running online using LUME-IMPACT

https://www.lume.science/

Badger GUI: useful for online optimization AND archiving of useful data

Variety of tools for online modeling and optimization. Optimization software useful for algorithm testing, deployment into ops, and collection of useful data for more comprehensive model training.

E331 Progress: ML for Efficient Characterization

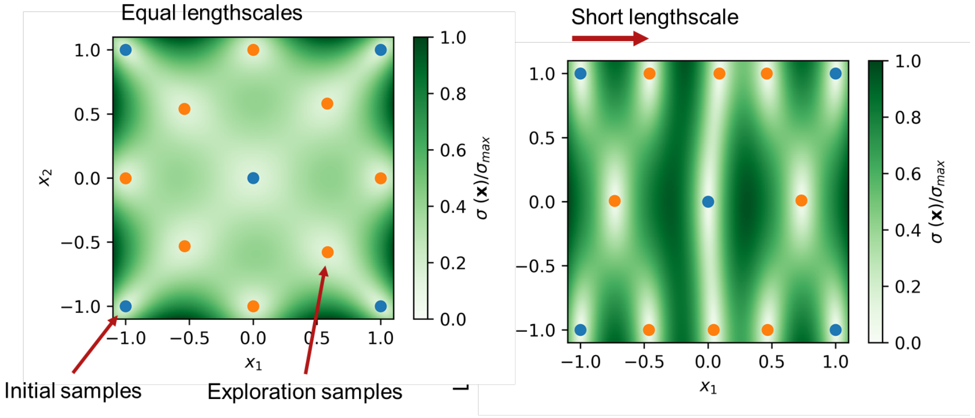
Better Data Sampling: Bayesian Exploration

$$\alpha(\boldsymbol{x}) = \sigma(\boldsymbol{x}) \prod_{i=1}^{N} p_i(g_i(\boldsymbol{x}) \geq h_i)\, \Psi(\boldsymbol{x}, \boldsymbol{x_0})$$

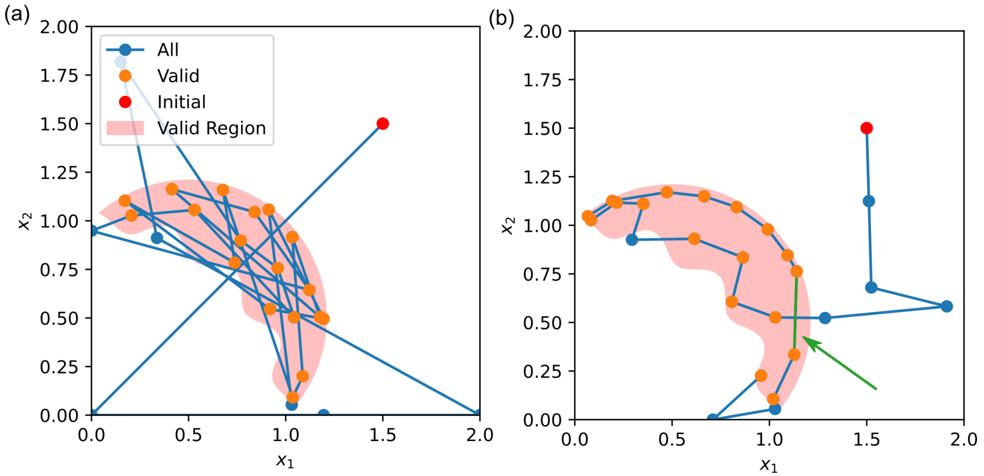
R. Roussel et. al.
Nat. Comm. **2021**

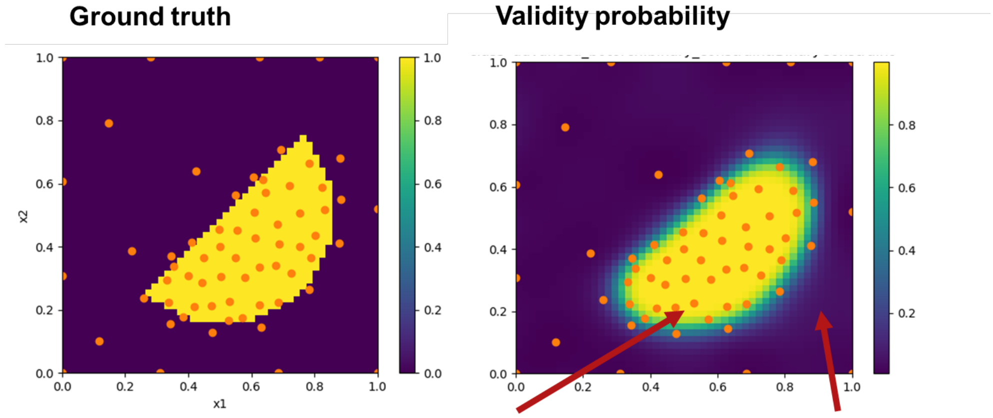
proximal biasing

adaptive sampling



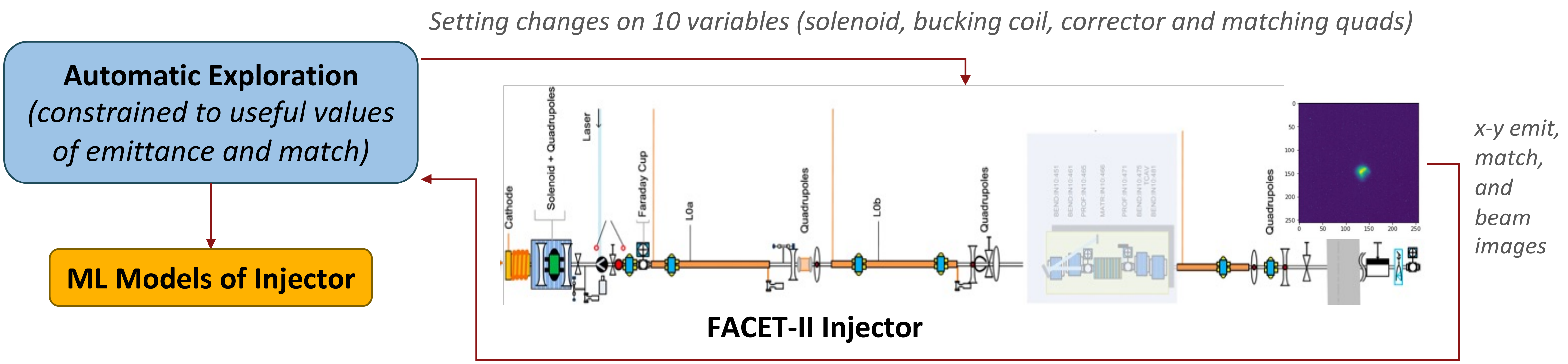
learning constraints





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

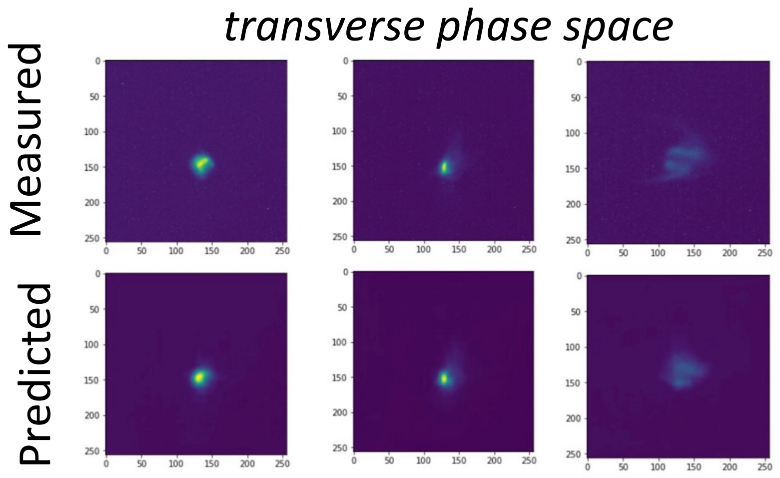
E331 Progress: ML for Efficient Characterization



- 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.

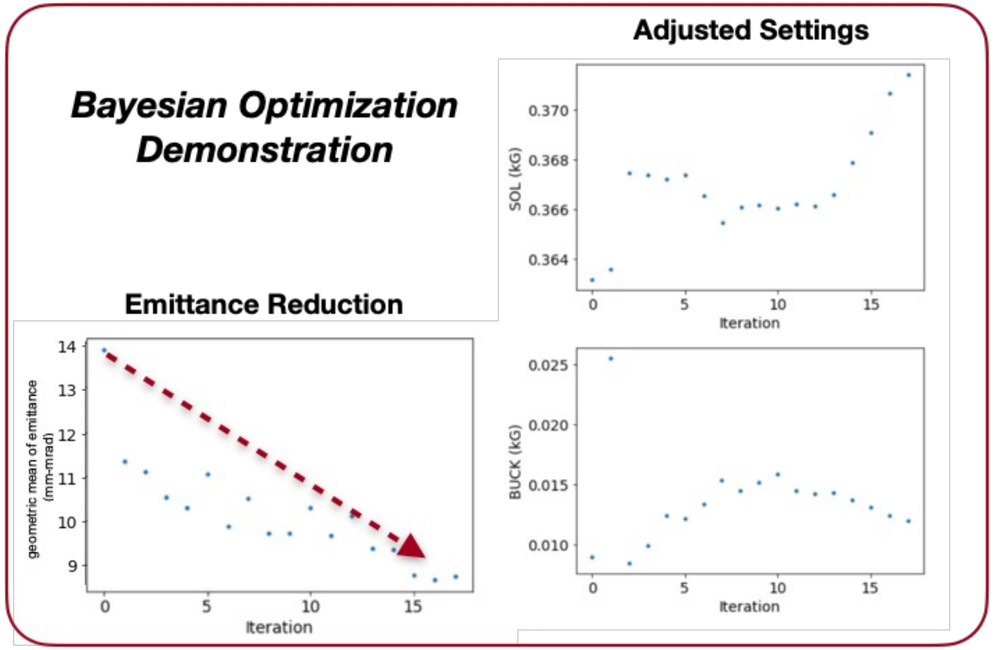
- **Example of integrated cycle between characterization, modeling, and optimization → now want to extend to larger system sections and new setups**

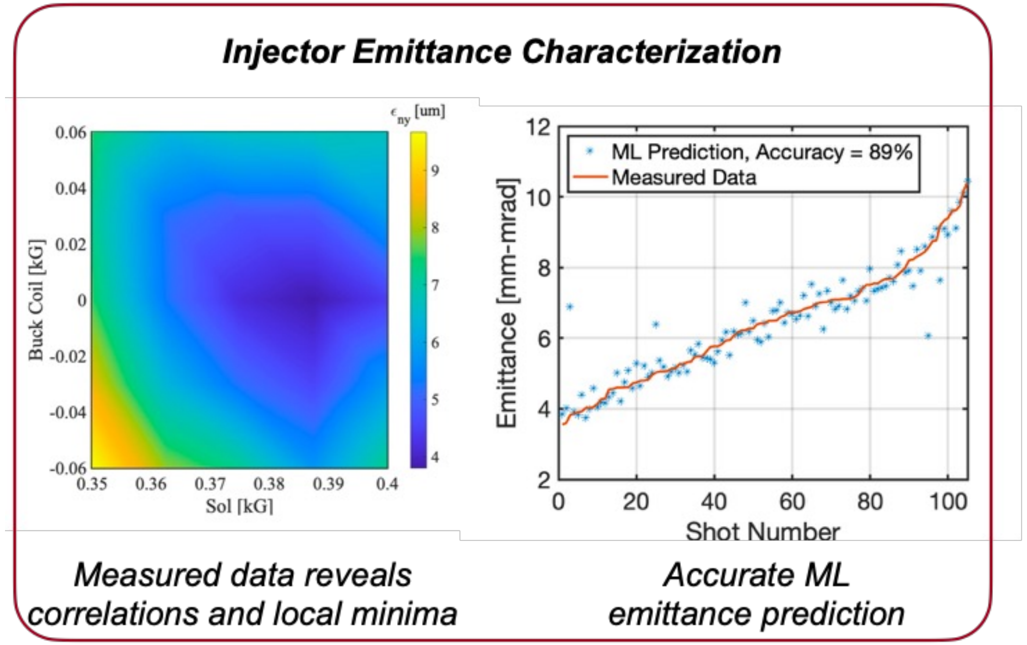


Use of Bayesian exploration to generate training data was sample-efficient, reduced burden of data cleaning, and resulted in a well-balanced distribution for the training data set over the input space. ML models were immediately useful for optimization.

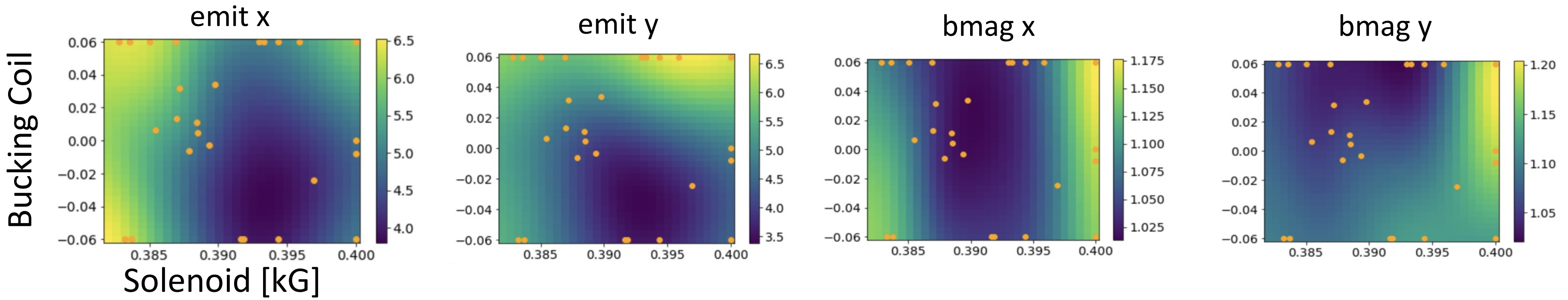
E331 Progress: Bayesian Optimization and Characterization of Injector

- Demonstrations of Bayesian optimization on the injector with up to 10 variables
- Extensive data obtained from characterization studies at 2nC and 700pC
- ML models from data give insight into machine behavior → still exploring this extensively





1.8 nC



ML model predictions and new sample locations

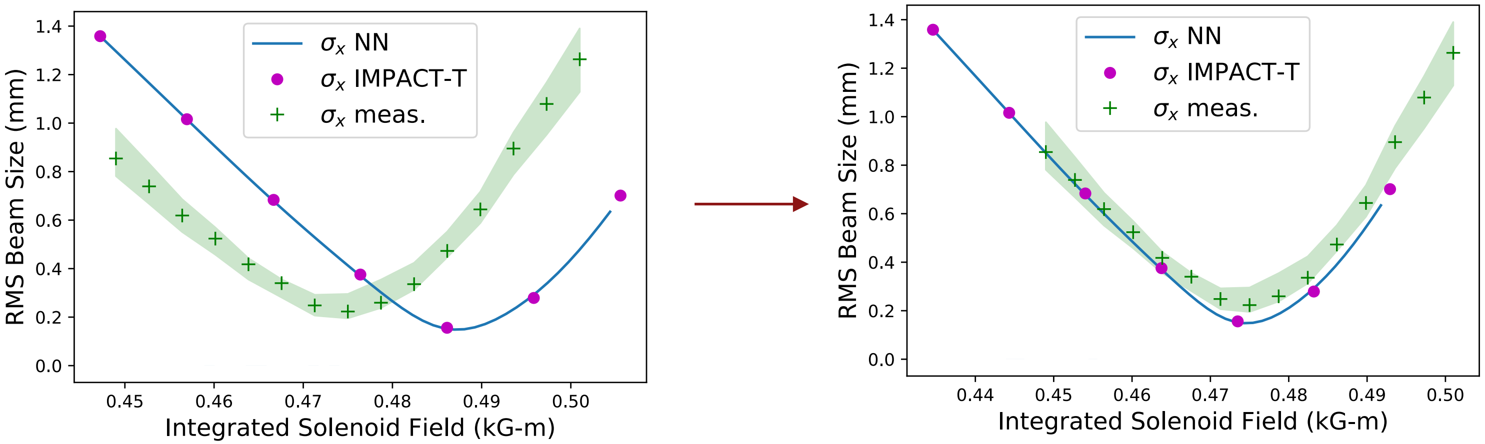
(learning to balance tradeoffs between outputs)

E331 Progress: Bayesian Optimization and Characterization of Injector

- Demonstrations
- Extensive dat
- ML models

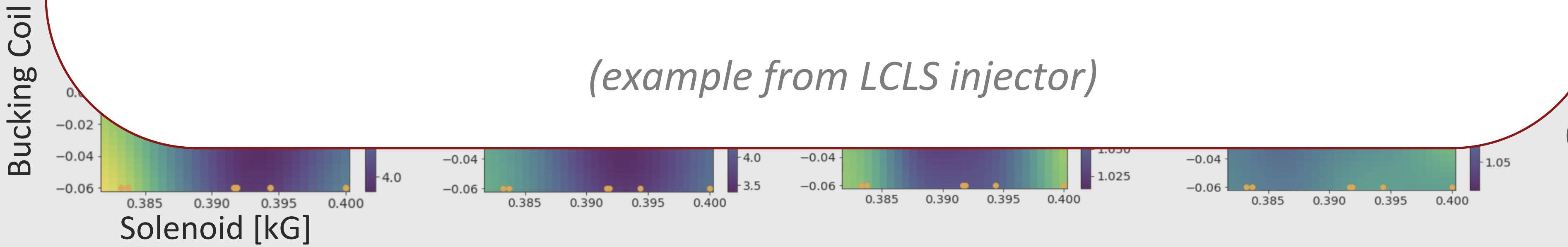
Bay

Working on using measured data gathered from these experiments to make comprehensive injector model and do model calibration to find sources of error and better match machine



(example from LCLS injector)

1.8 nC

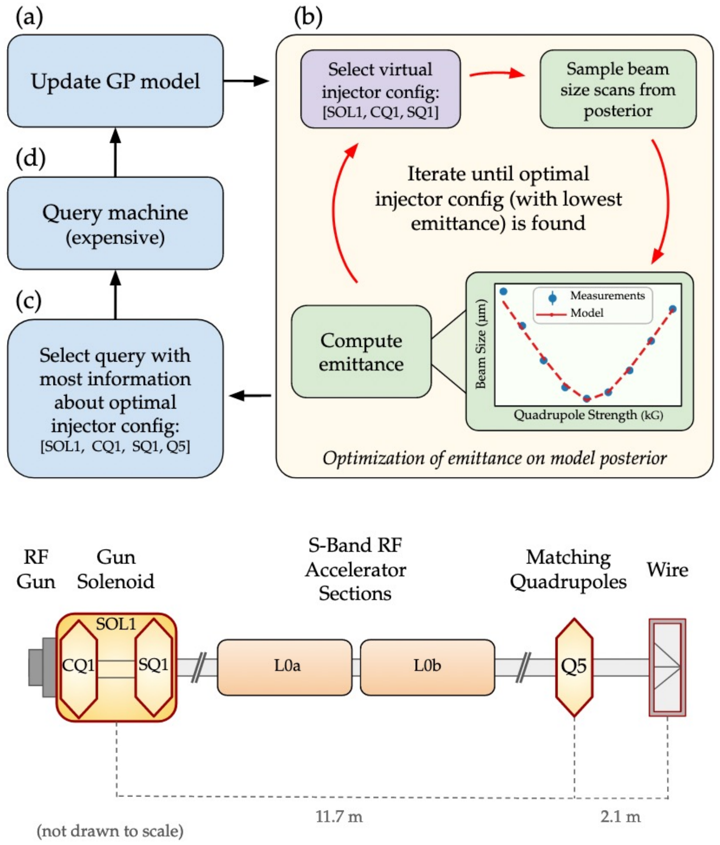


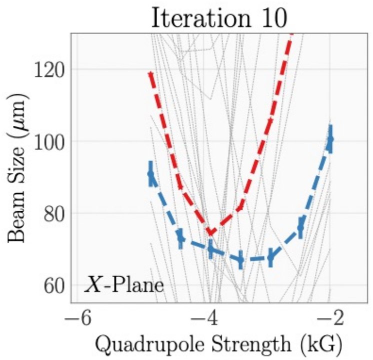
model predictions and new sample locations

(learning to balance tradeoffs between outputs)

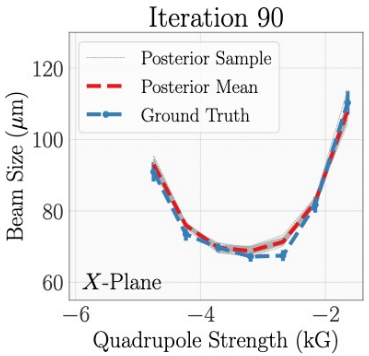
E331 Progress:
Efficient Emittance Optimization with Partial Measurements

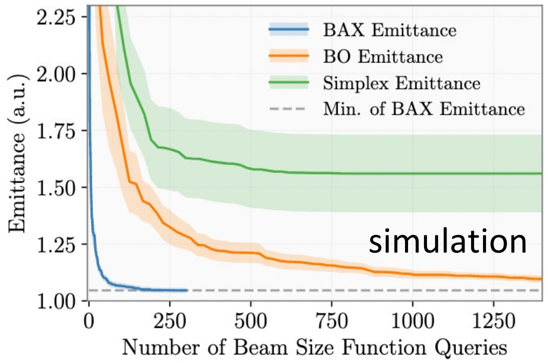
- **Instead of tuning on costly emittance measurements directly, learn a fast-executing model online for beam size while optimizing**
- **Demonstrated new algorithmic paradigm leveraging "Bayesian Algorithm Execution" (BAX) for 20x speedup in tuning** → *learn on direct observables (e.g. beam size); do inferred "measurements" (e.g. emittance) much more quickly on the model than would be possible on the machine*

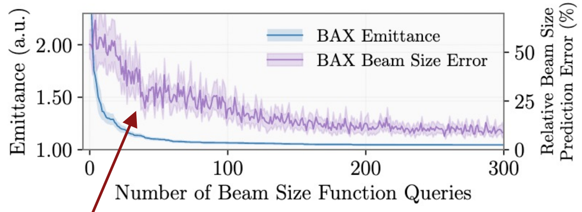




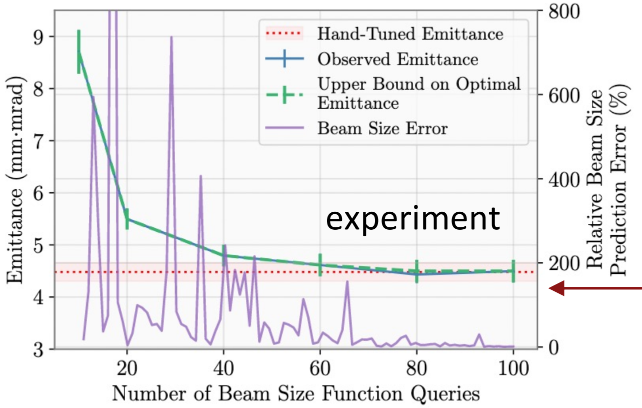
model is learned on-the-fly







Convergence of beam size prediction error gives practical indicator of optimization convergence *(no need to do direct emittance measurement until the end)*



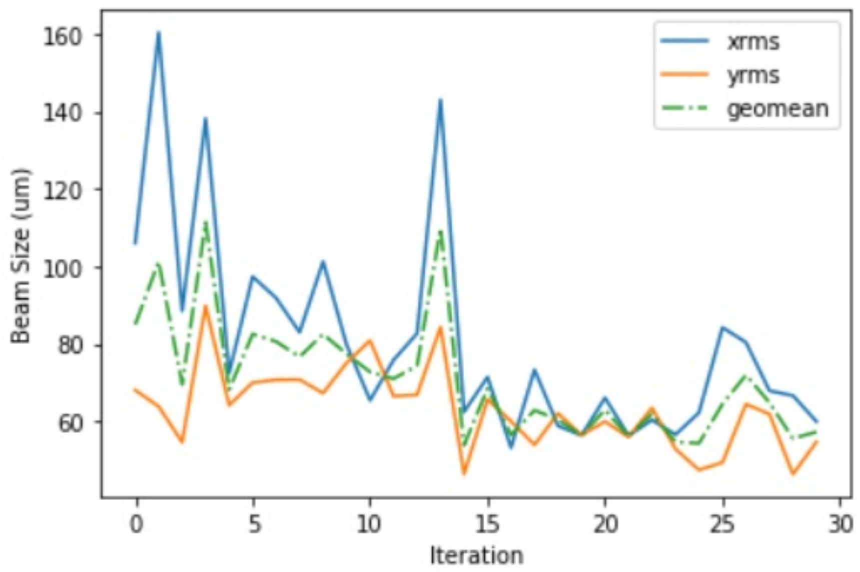
Found equivalent quality to hand-tuning in about 70 iterations *(just a few minutes with computationally optimized routine)*

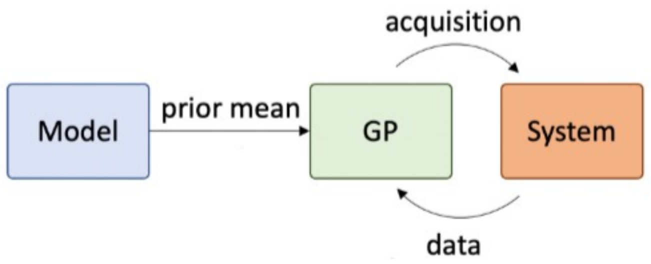
https://arxiv.org/abs/2209.04587

New method demonstrated at FACET-II has 20x speed improvement over standard emittance optimization method. Paradigm shift in how tuning on indirectly computed beam measurements (such as emittance) is done.

E331 Progress: 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*





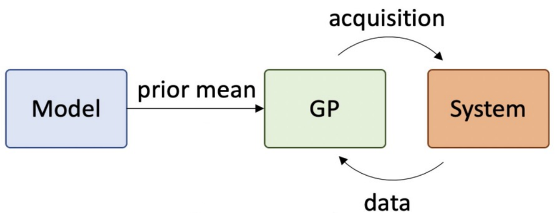
Next:
- Want to use on both IPs (with multi-objective optimization) and use greater number of variables
- Use data to inform faster subsequent optimization

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.

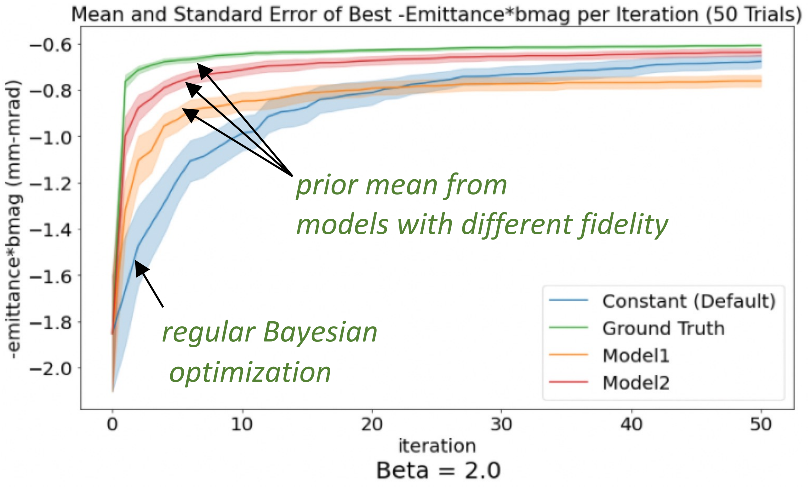
Next Steps: NN Prior

Combining neural networks with BO → important for scaling BO up to higher-dimensional tuning problems

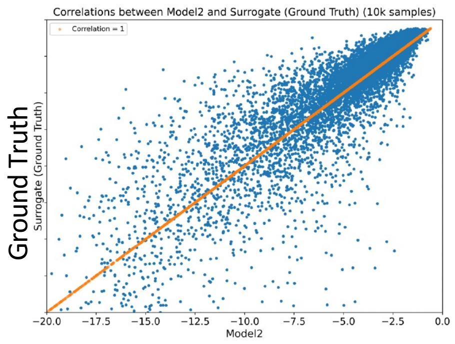
Good first step from previous work: use neural network system model to provide a prior mean for a GP



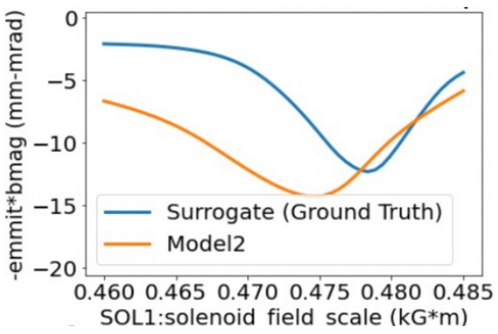
Used LCLS injector surrogate model for prototyping
variables: *solenoid, 2 corrector quads, 6 matching quads*
objective: *minimize emittance and matching parameter*



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



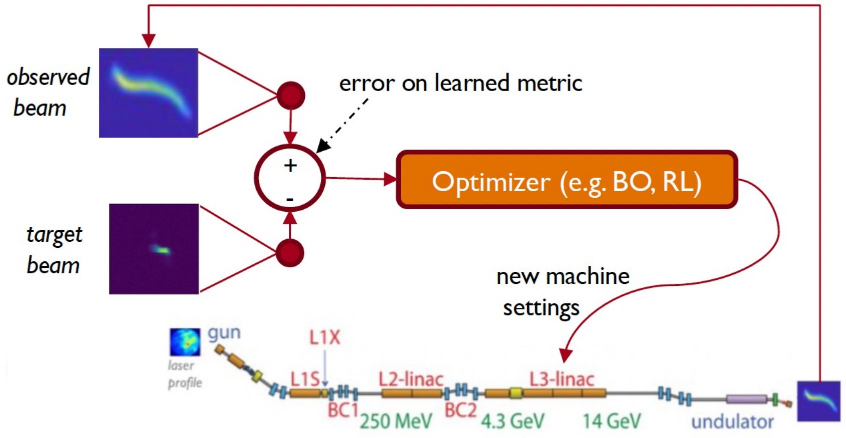
Model 2



NeurIPS proceeding: https://arxiv.org/abs/2211.09028

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

Next Steps: LPS Tuning



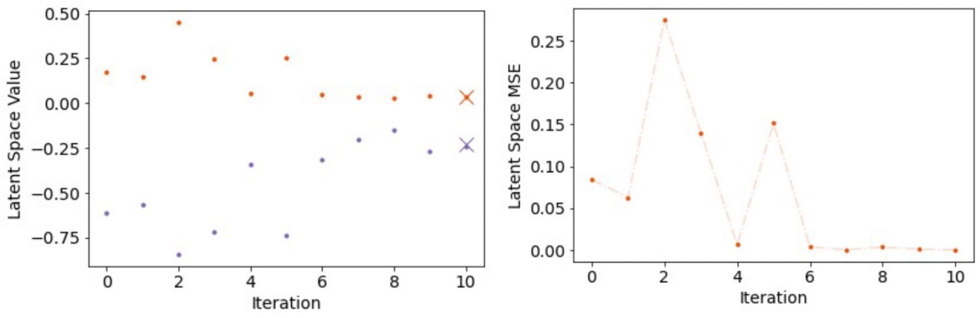
Demonstrated Bayesian optimization for LPS tuning on LCLS for several variants of problem setup:

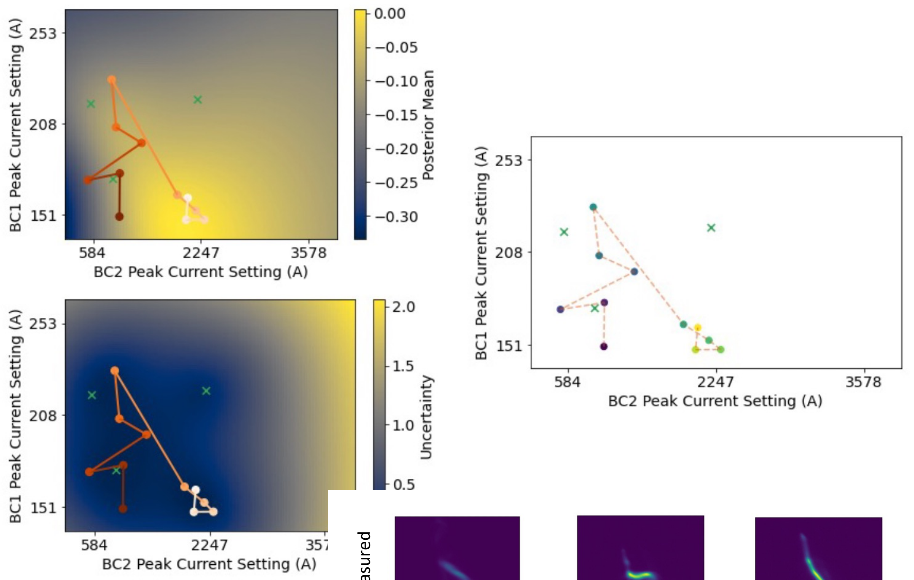
- *2 peak current settings, 6 phases and amplitudes*
- *Target phase space, minimize energy spread and bunch length*

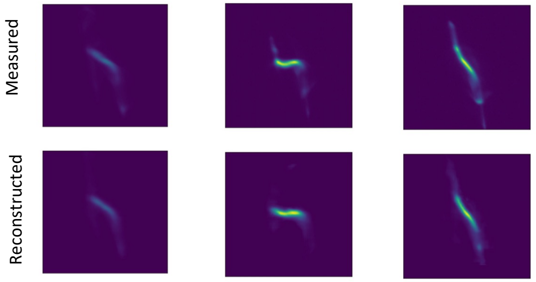
→ Want to test on FACET-II as first step toward more comprehensive neural network based control for LPS

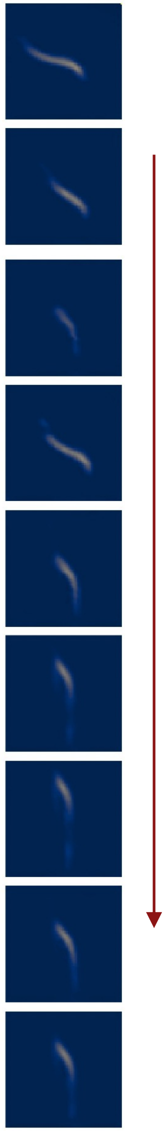
→ Data gathered during BO-based tuning will be useful for next steps *(model calibration, neural network control policy + reinforcement learning)*

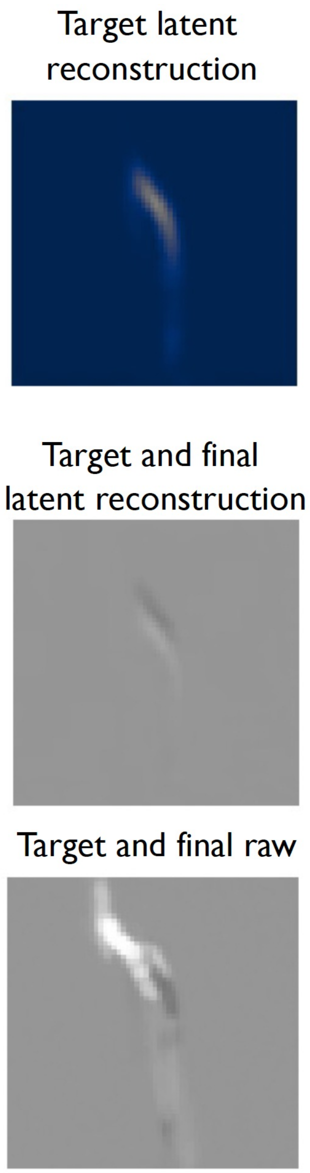
Example from LCLS











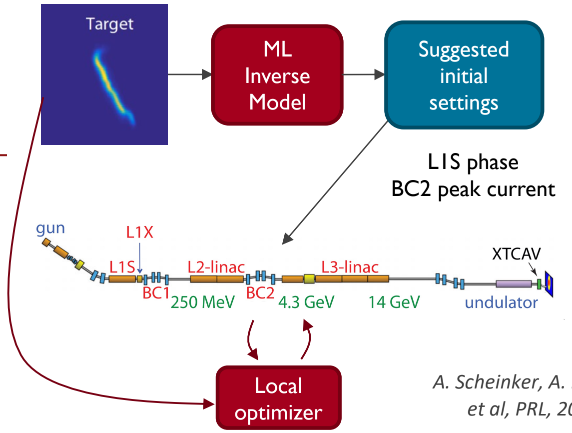
Future Work

- Next steps:
  - Simultaneous optimization of the beam spot at both IPs (adjusting sextupole movers and other variables in S20), optimization to reduce emittance growth

    Can use trust region BO and then NN prior + BO

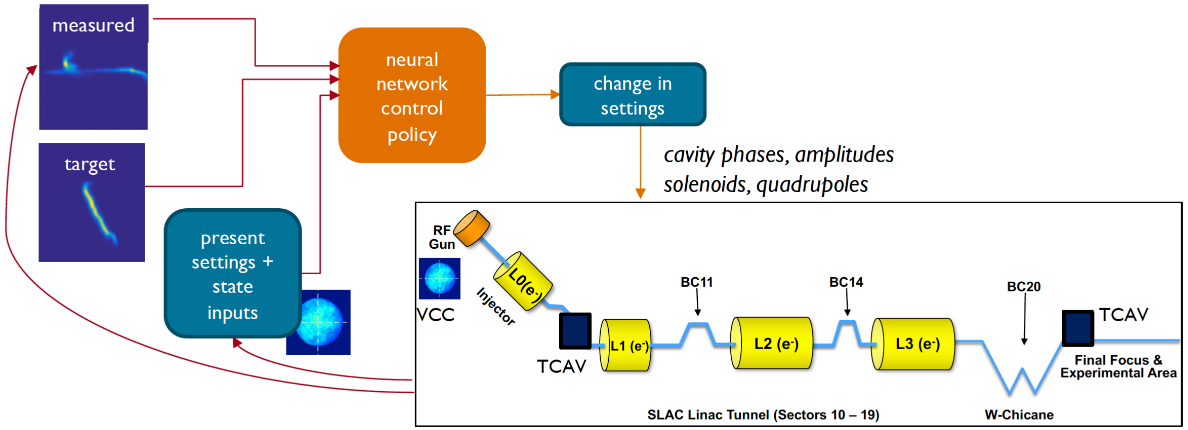
  - Incorporate TCAVs in tuning for longitudinal phase space optimization
  - Use data gathered for comprehensive model-based approaches *(calibrate global models, use neural network prior mean to speed up Bayesian optimization, extend to reinforcement learning)*

    Aim to use for fast switching between configurations and fine-tuning



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

RL is a complementary approach to model-based warm starts



- Farther in the future:
  - Drive and witness bunch optimization
  - PWFA optimization
    - Reduction of beam jitter (synergy with E325 + E327)
    - Can leverage virtual diagnostic from E327 as additional tuning output
  - ML aided LPS shaping with the laser heater (synergy with E325 + E327)

Desired facility upgrades

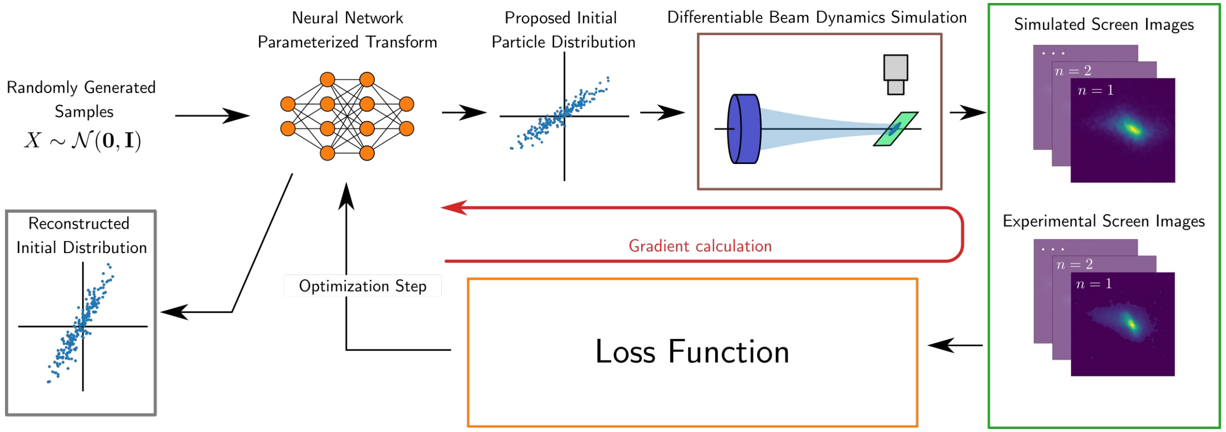
- Computing
  - GPU integration into online compute resources with read and write permissions to machine (S3DF, controls network, or local compute)
  - Working on getting links to S3DF with limited write access (with TID/EED)
  - November ‘23 Jingchen and others will start looking into suitable GPUs for controls network
  - Have a standalone GPU box → would like to get write access as a temporary measure in the interim (but has met with resistance)

R. Roussel et al, PRL 2023

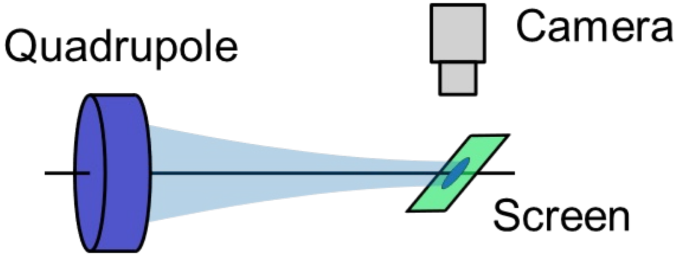
Phase Space Reconstruction with Differentiable Tracking Simulations

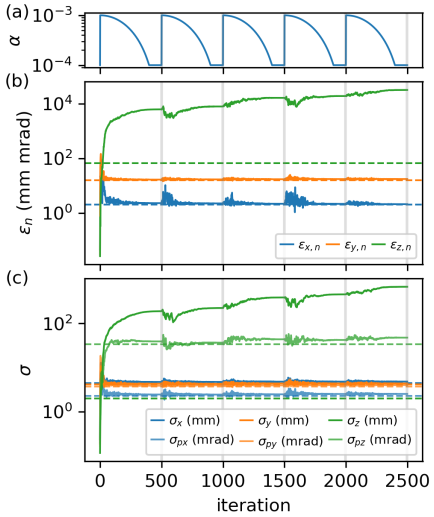
Bmad-X

Differentiable pipeline for reconstructing 6D phase space distribution using neural network parameterization

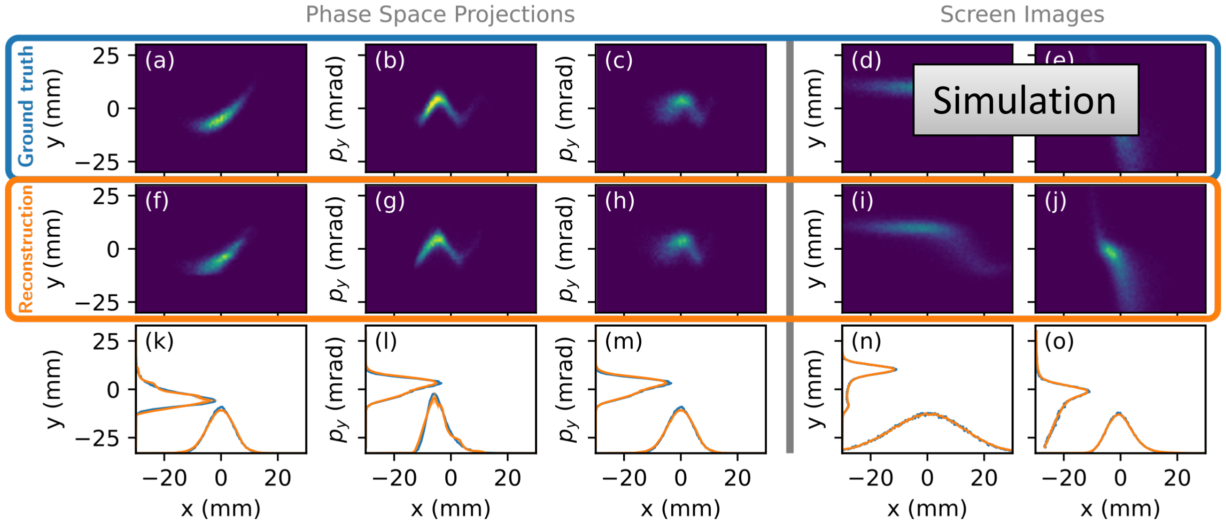


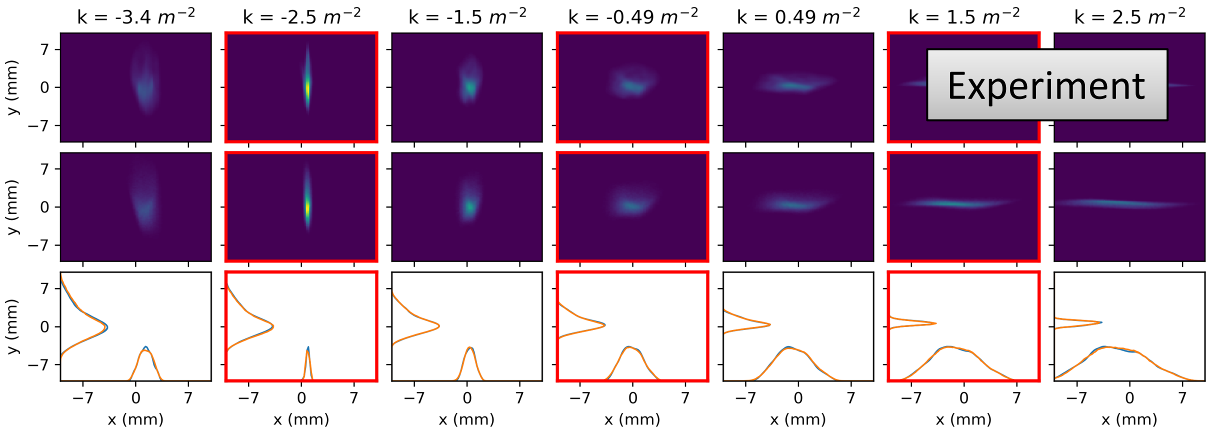
Reconstruct 4D phase space distribution + approx. energy spread from simple beamline diagnostic and 10 measurements





Confidence estimates

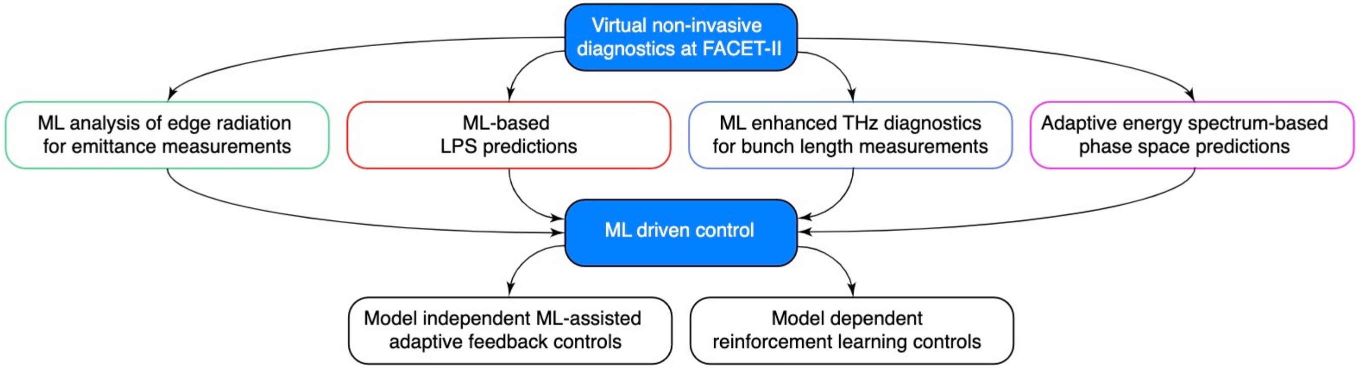




ML combined with differentiable simulations opens up a new paradigm for constructing detailed phase space diagnostics in a way that is computationally-efficient and sample-efficient

Thanks to the team and collaborators!

A. Edelen, C. Emma, R. Roussel, S. Miskovich, W. Neiswanger, G. White, S. Gessner, A. Scheinker, C. Mayes, D. Ratner, B. O’Shea, Z. Zhang, T. Boltz, J. P. Gonzalez-Aguilera, D. Kennedy and many others



Backups

Machine Control and Understanding

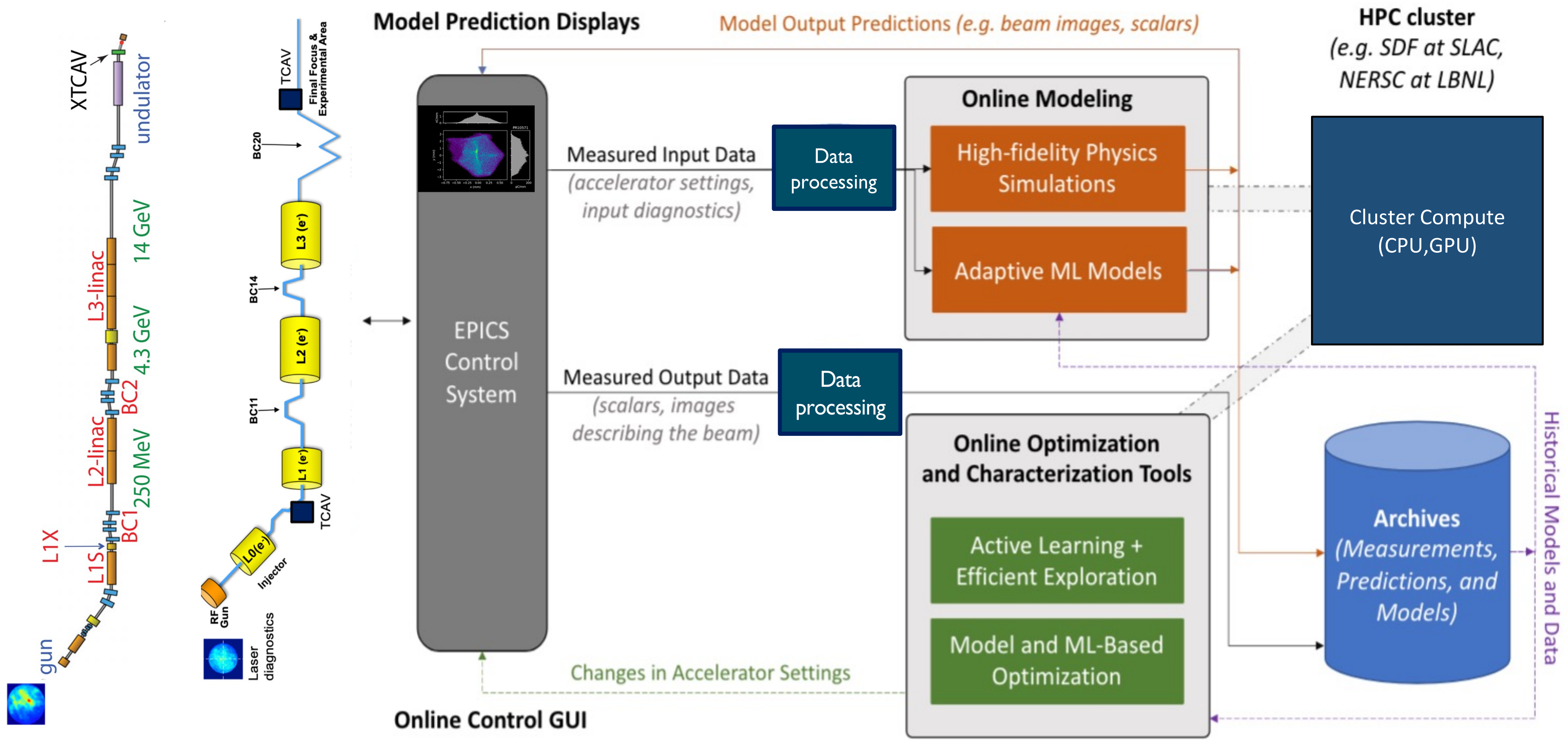
Diagnostics (Need information to make decisions)		Control (How to make decisions)	
Edge Radiation Emittance Diagnostics (E326)	Virtual TCAV Predictive Diagnostics (E327)	Adaptive Feedback (E325)	Learned Control (E331)
Non-destructive, single shot continuous monitoring of emittance of high-current beams	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

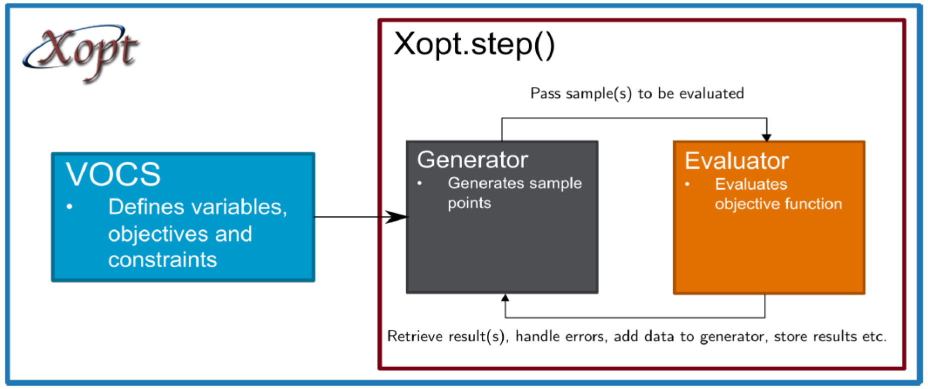
Will enable system-wide application to aid operations, and help drive AI/ML development *(e.g. higher dimensionality, robustness, combining algorithms efficiently)*



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

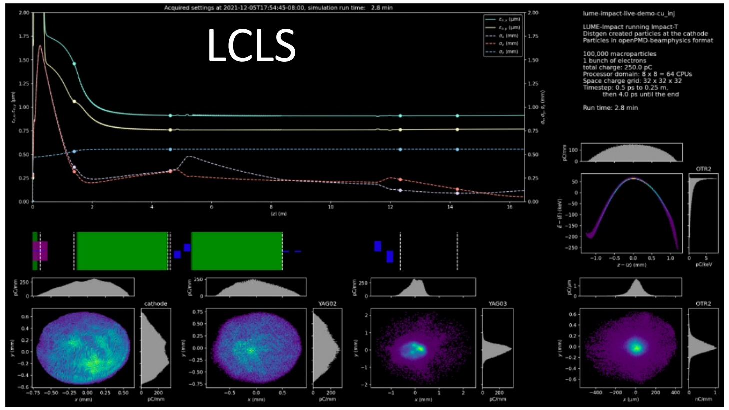
Modular, Open-Source Software Development

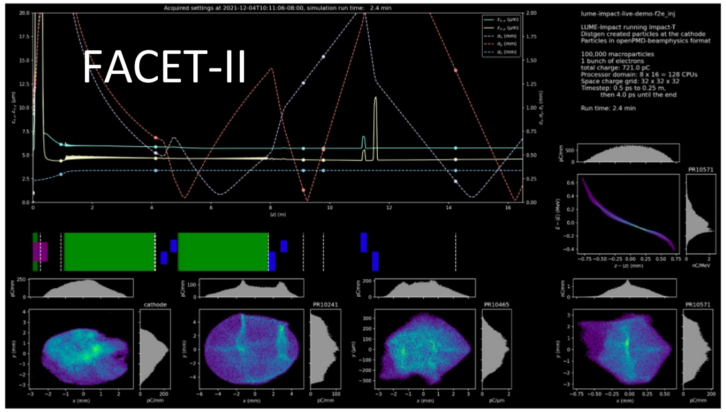
- Community development of **re-usable, reliable, flexible software tools** for AI/ML workflows has been essential to maximize return on investment and ensure transferability between systems
- **Modularity has been key**: separating different parts of the workflow + using shared standards




```
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

Different software for different tasks:

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

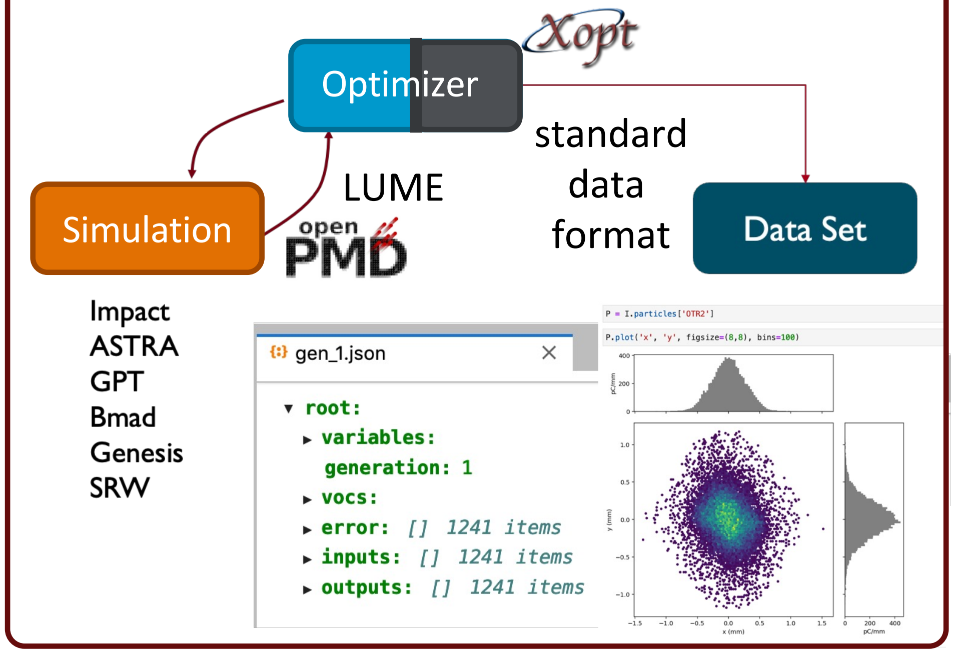
Visual control room interface *(e.g. Badger)*

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/*](https://www.lume.science/)

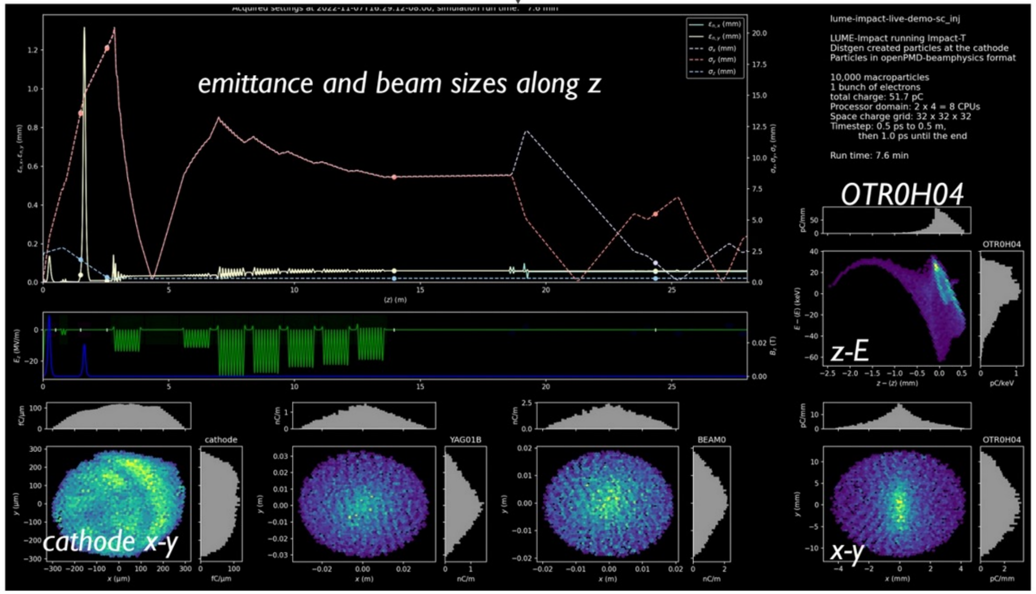


Modular open-source software has been essential for our work. We welcome new users and contributors.

Example: Online Models and Bayesian Optimization in Operations

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

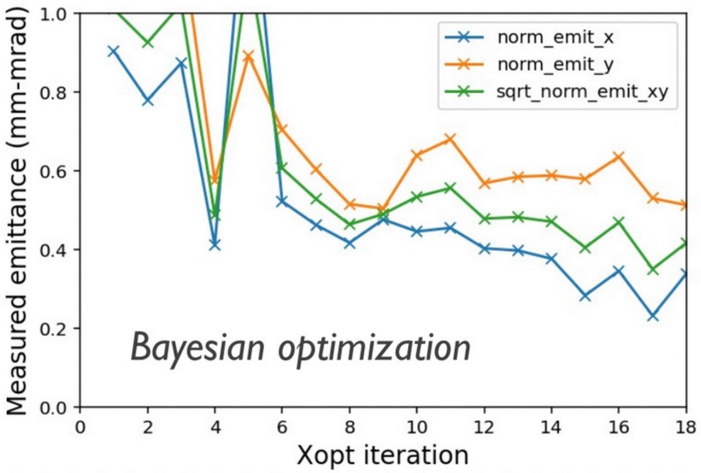
Readings from machine via EPICS
injector settings, laser profile from VCC image

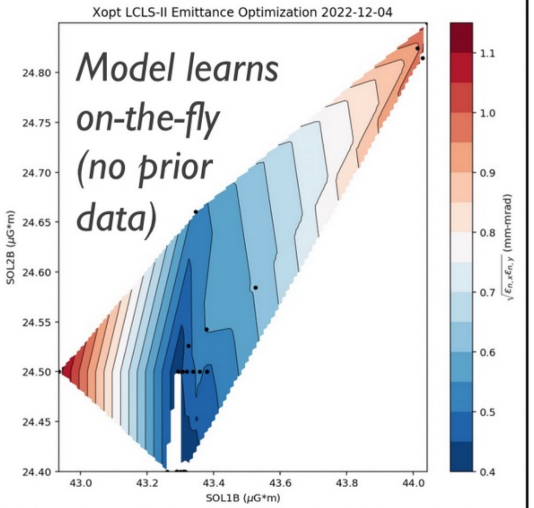


LCLS-II live sim: run on HPC and display in control room
Updates every 3-8 mins, space charge included, uses LUME-IMPACT

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

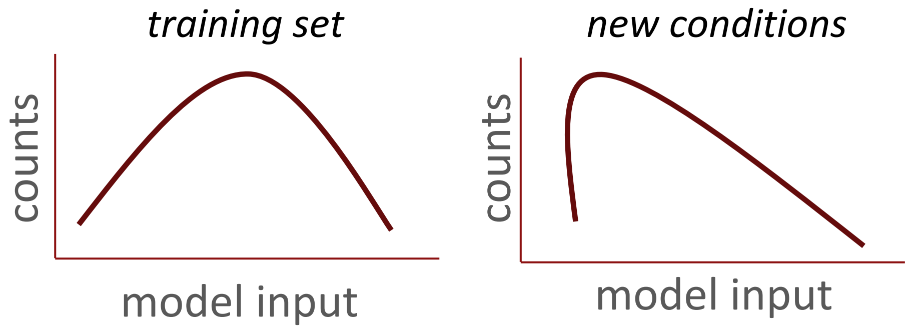
despite extensive previous hand-tuning

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

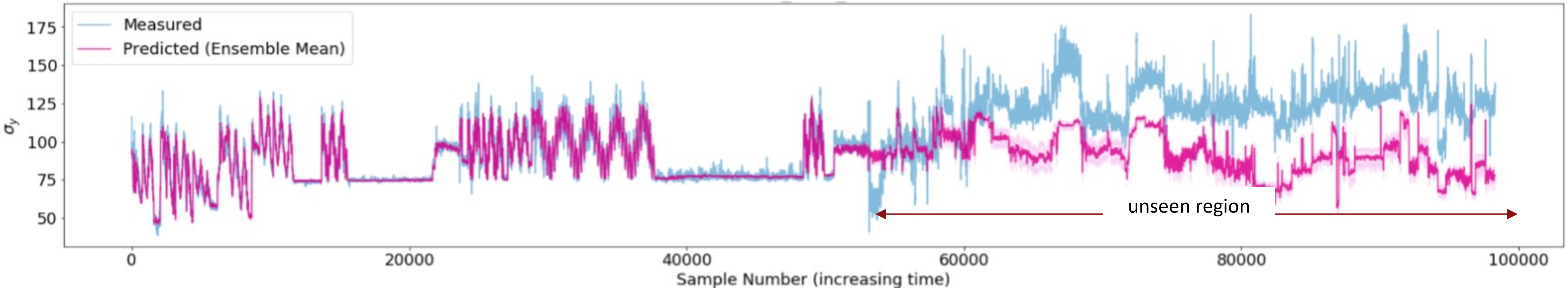
HPC enables fundamentally new capabilities in what can be realistically simulated online

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

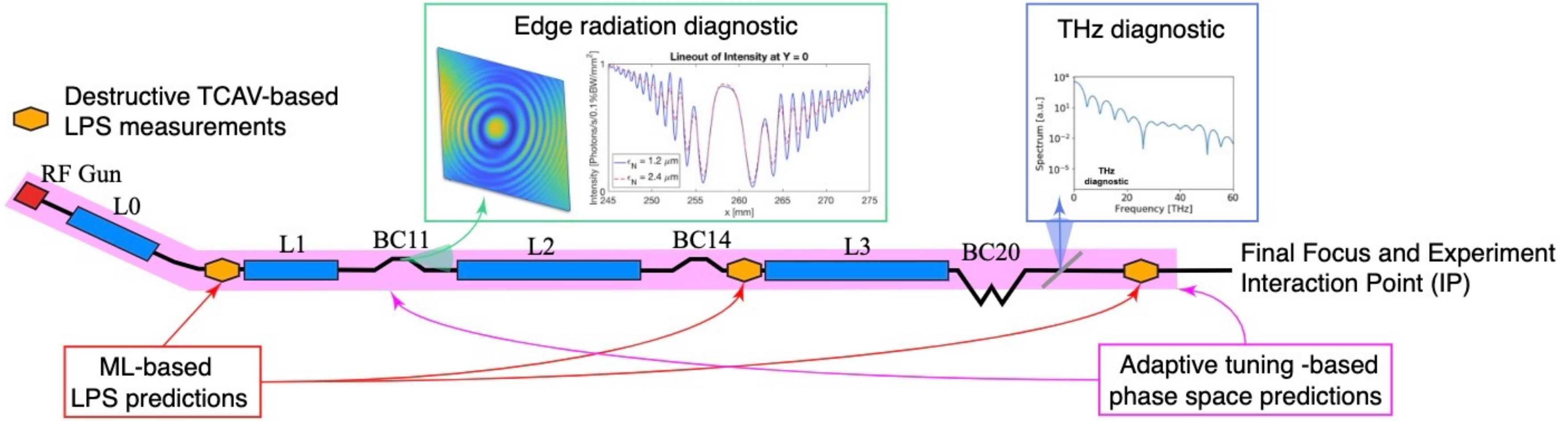


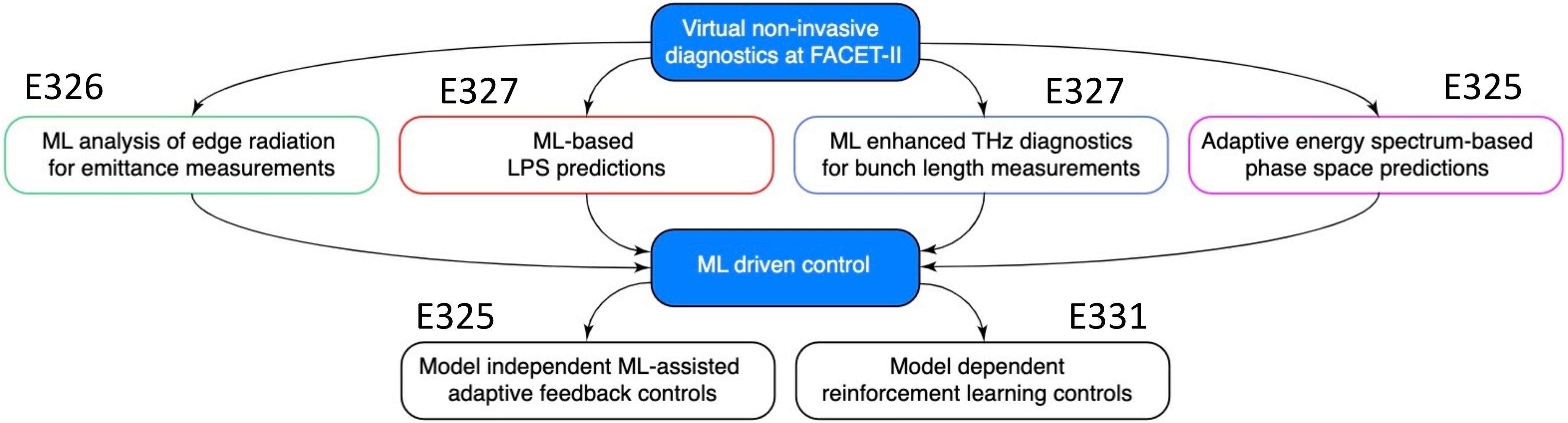
Example: beam size prediction and uncertainty estimates under drift from a neural network
Uncertainty estimate from neural network ensemble does not cover prediction error, but does give a qualitative metric for uncertainty



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

Landscape of AI/ML Activities at FACET-II





Synergistic experiments, individual success enhances all research + facility operation