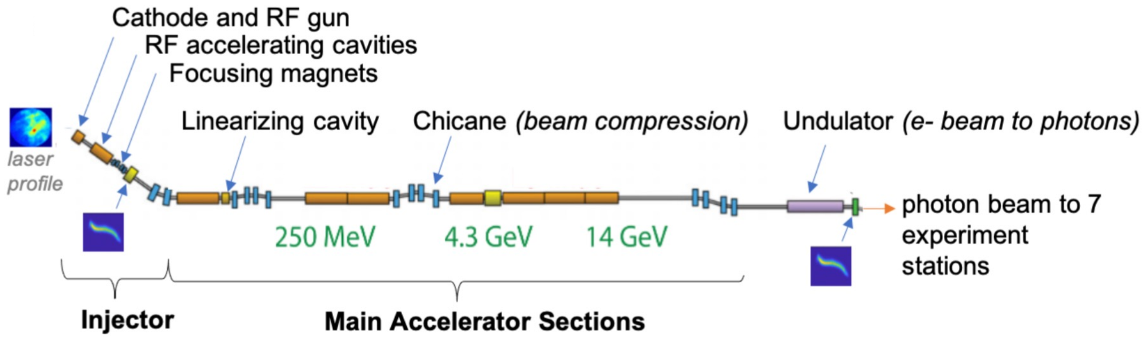
FCC AIML Mini Workshop
June 13, 2024

Progress on Combining Digital Twins and Machine Learning Based Control for Accelerators at SLAC

Auralee Edelen
edelen@slac.stanford.edu

Many tuning problems require detailed beam phase space customization for different experiments

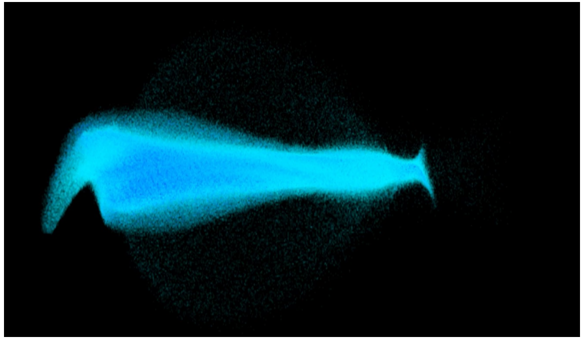


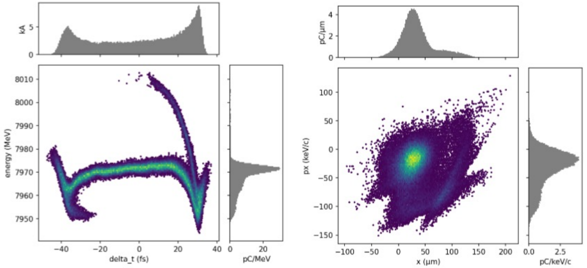
Beam exists in 6-D position-momentum phase space

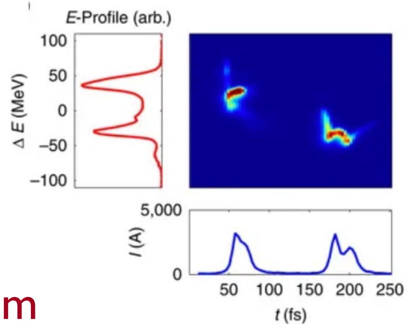
Have incomplete information: *measure 2-D projections or reconstruct based on perturbations of upstream controls (e.g. tomography, quad scans)*

Have dozens-to-hundreds of controllable variables and hundreds-of-thousands to monitor

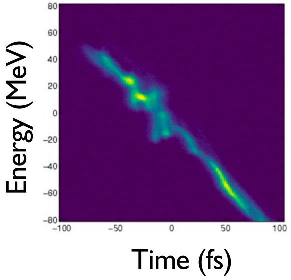
Nonlinear, high-dimensional optimization/control problem





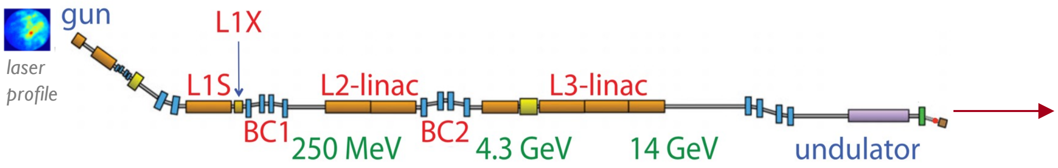


A. Marinelli, et al., Nat. Commun. **6**, 6369 (2015)



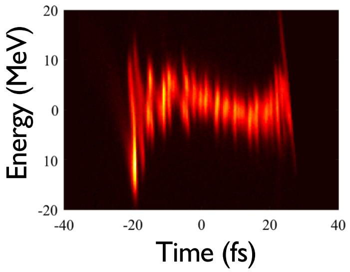
A. Marinelli, IPAC'18

wide spectrum of tuning needs

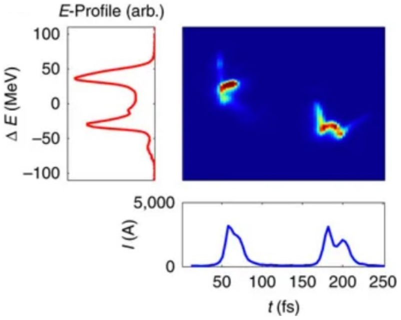


photon beam to 7 experiment stations

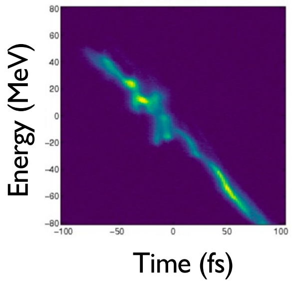
J. Qiang et al, PRAB (2017)



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



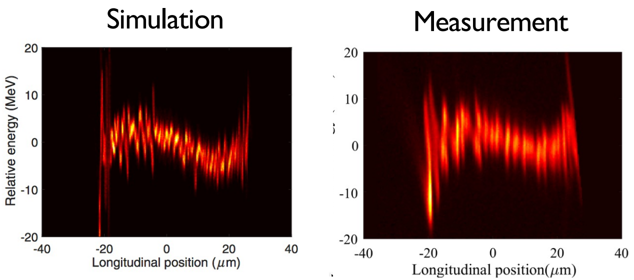
A. Marinelli, IPAC'18



Rapid beam customization

Achieve new configurations + unprecedented beam parameters

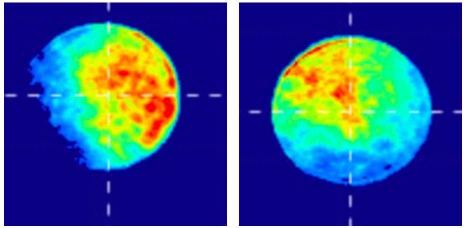
Fine control to maintain stability within tolerances *computationally expensive simulations*



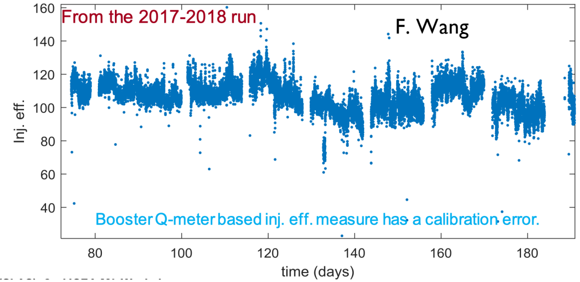
"10 hours on thousands of cores at the NERSC"

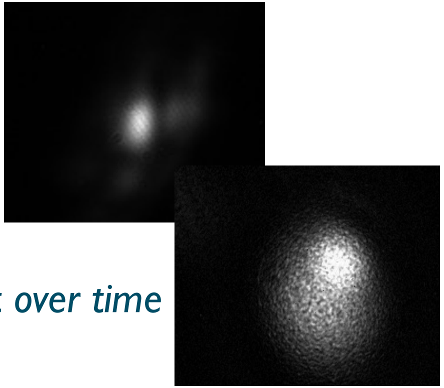
J. Qiang, et al., PRSTAB30, 054402, 2017

fluctuations/noise (e.g. initial beam conditions)

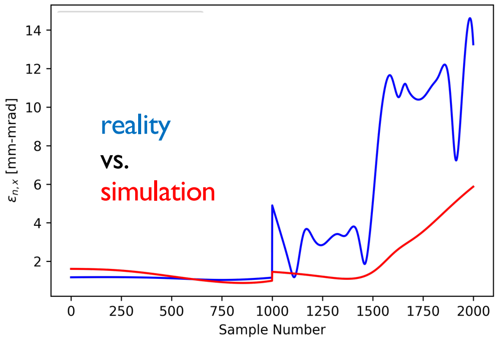


hidden variables / sensitivities

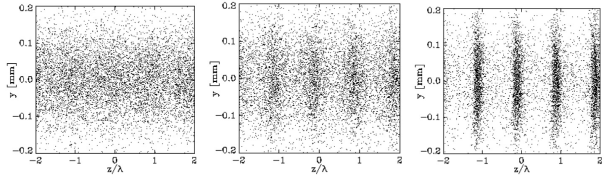




drift over time



many small, compounding sources of uncertainty

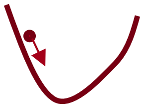


nonlinear effects / instabilities

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

less ← assumed knowledge of machine → more

Model-Free Optimization

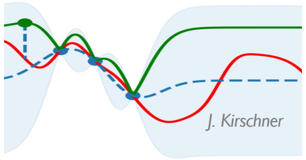


Observe performance change after a setting adjustment

→ estimate direction or apply heuristics toward improvement

gradient descent
simplex
ES

Model-guided Optimization

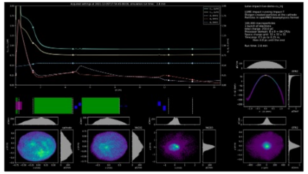


Update a model at each step

→ use model to help select the next point

Bayesian optimization
reinforcement learning

Global Modeling + Feed-forward Corrections



→ provide initial guess (i.e. warm start)
→ provide insight to operators
→ model-based control

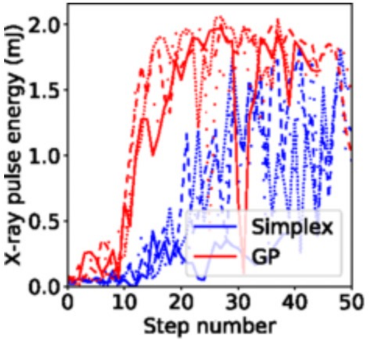
ML system models +
inverse models

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

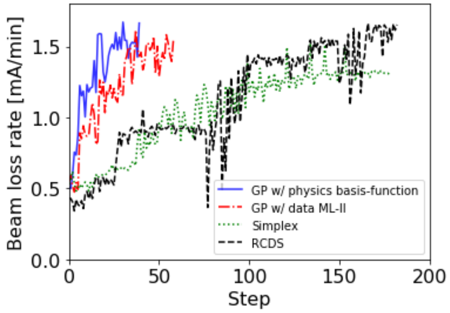
(+ improvements)

FEL pulse energy tuning at LCLS



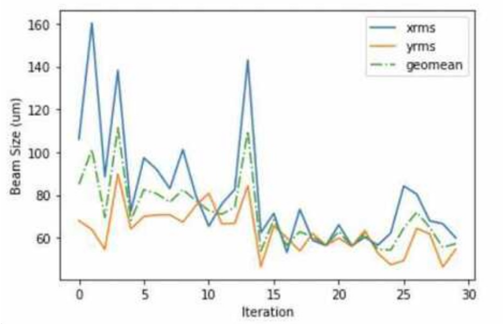
Duris et. al. PRL , **2020**

Loss rate tuning at SPEAR3

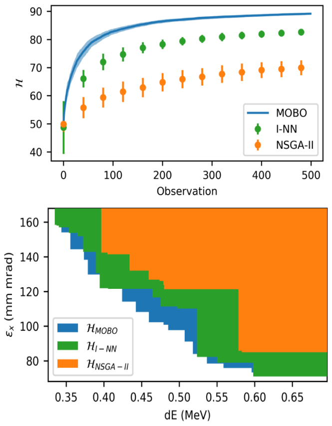


Hanuka et. al. PRAB , **2021**

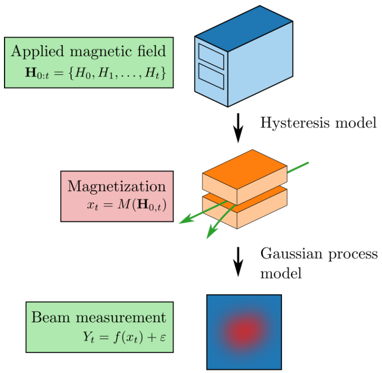
Sextupole tuning for IP at FACET-II



Multi-objective Bayesian Optimization

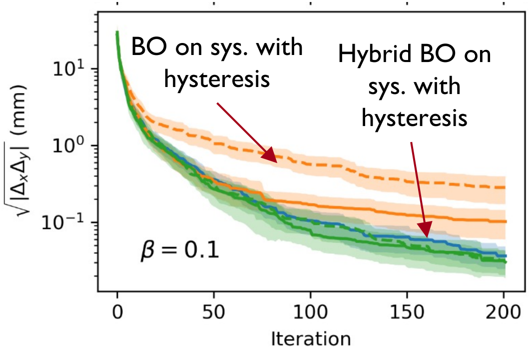


Roussel et. al. PRAB , **2021**

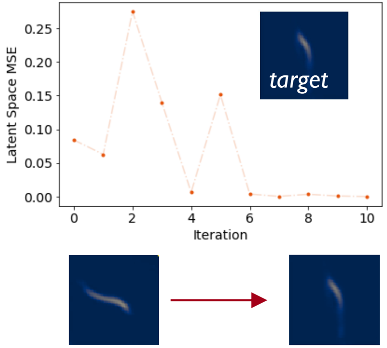


Roussel et. al. PRL , **2022**

Higher-precision optimization possible when including hysteresis effects in model



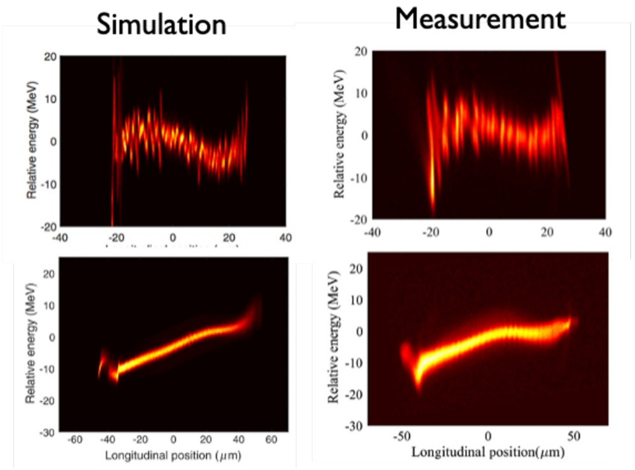
Longitudinal phase space tuning on LCLS



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

Fast-Executing, Accurate System Models

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



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

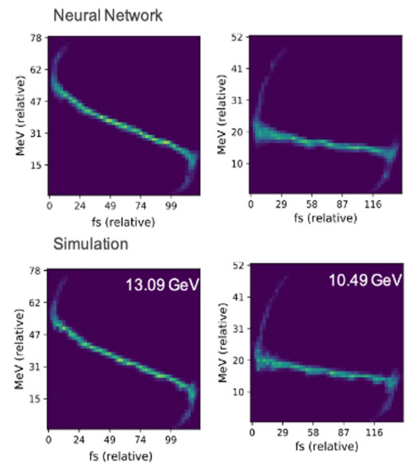
10 hours on thousands of cores at NERSC!

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

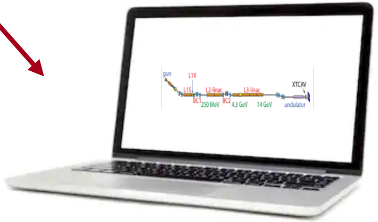
Edelen et al., NeurIPS 2019

Long history now of using ML modeling to enable accurate predictions of accelerator system responses with unprecedented speeds

Fast-Executing, Accurate System Models

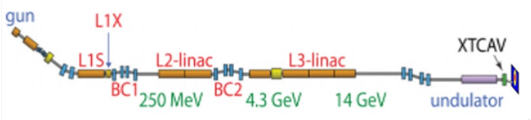
Bringing simulation tools from HPC systems to online/local compute

Online prediction
Model-based control



Control prototyping
Experiment planning

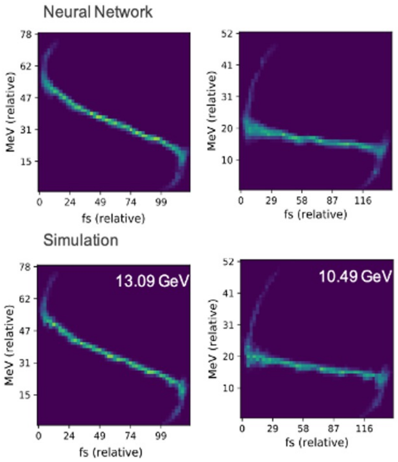
ML models are able to provide fast approximations to simulations ("surrogate models")



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



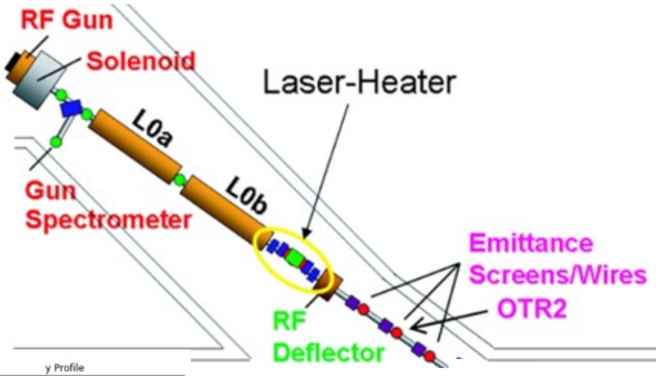
< ms execution speed

10^6 times speedup

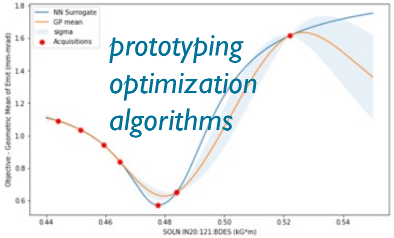
Edelen et al., NeurIPS 2019

Long history now of using ML modeling to enable accurate predictions of accelerator system responses with unprecedented speeds

In Regular Use: Injector Surrogate Model at LCLS

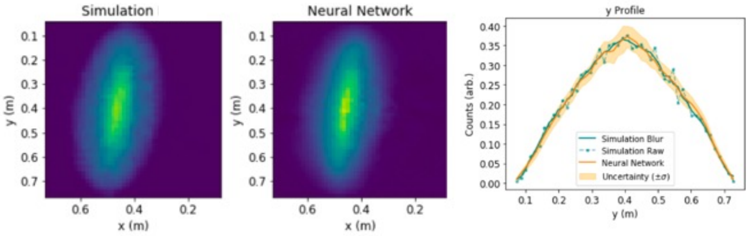


- ML models trained on detailed physics simulations with nonlinear collective effects
- Accurate over a wide range of settings → calibrate to match machine measurements
- Used to develop/prototype new algorithms before testing online
  (e.g. BAX w/ 20x speedup in emittance tuning https://arxiv.org/abs/2209.04587)
- Will provide initial parameters for downstream model

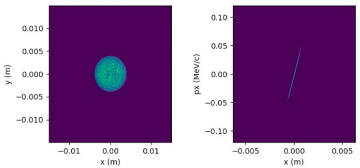


prototyping optimization algorithms

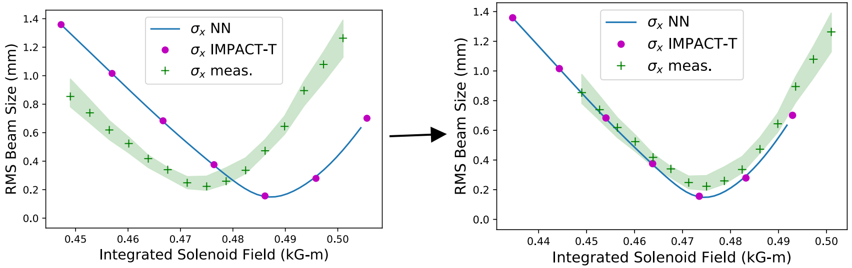
ML model matches simulation under interpolation



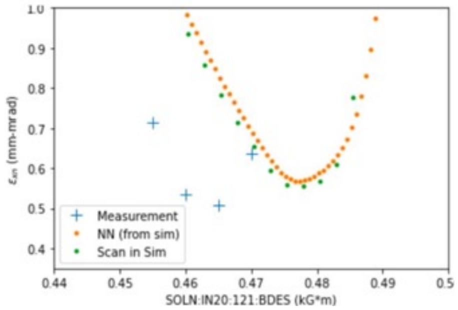
Simulation and ML model trained on it are qualitatively similar to measurements under interpolation (setting combinations reasonable distance from training set)



interactive model widget and visualization tools



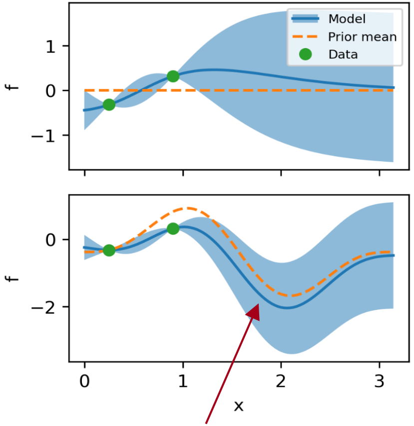
Automatic adaptation of models and identification of sources of deviation between simulations and as-built machine



ML models trained on simulations and measurements have enabled fast prototyping of new optimization algorithms, facilitated rapid model adaptation under new conditions, and can directly aid online tuning and operator decision making

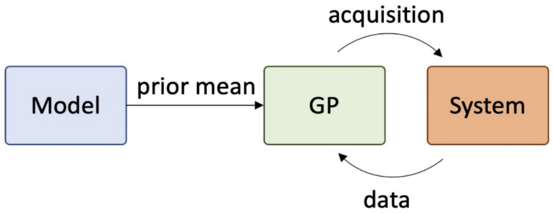
Leveraging Online Models for Faster Optimization

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



model prediction returns to prior

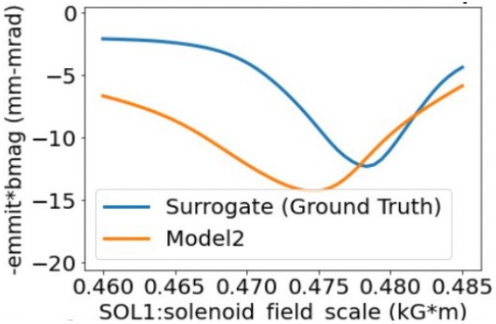
Good first step from previous work: use neural network system model to provide a prior mean for a GP



Used the LCLS injector surrogate model for prototyping

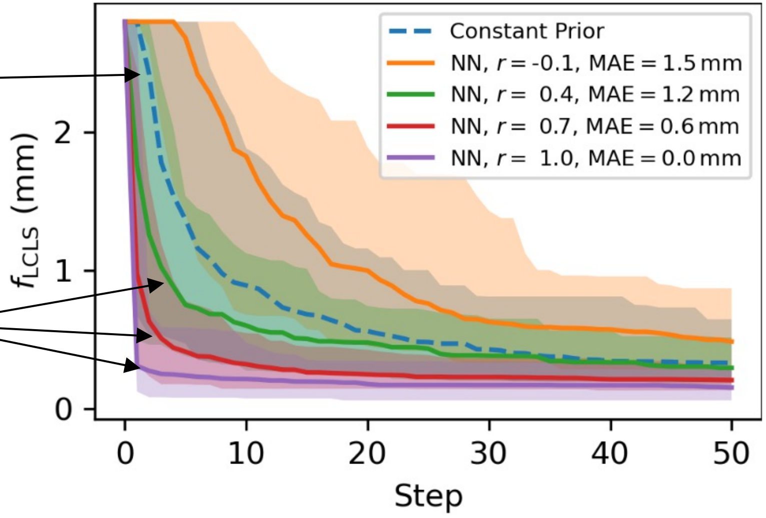
variables: *solenoid, 2 corrector quads, 6 matching quads*

objective: *minimize emittance and matching parameter*



regular Bayesian optimization

prior mean from models with different fidelity

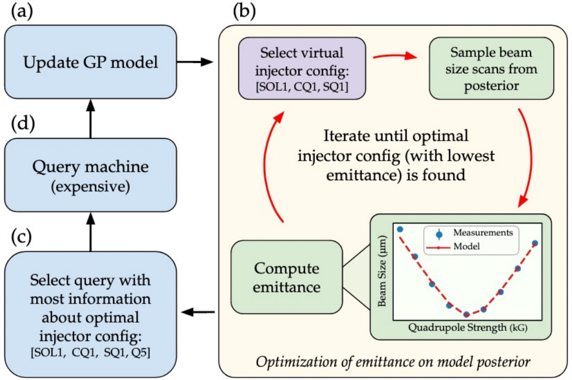


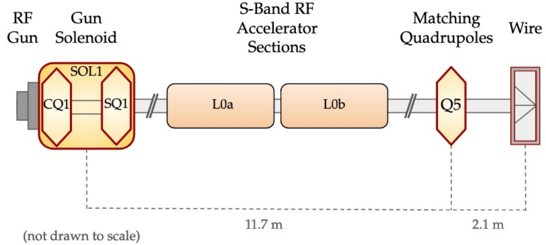
Even prior mean models with substantial inaccuracies provide a boost in optimization speed

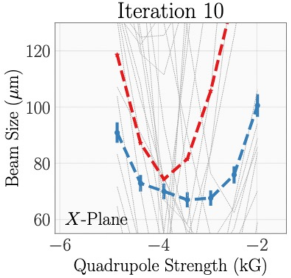
https://arxiv.org/abs/2403.03225 https://arxiv.org/abs/2211.09028

Efficient Emittance Optimization with Virtual Objectives

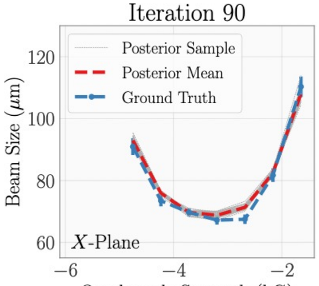
- **Instead of tuning on costly emittance measurements directly: learn a fast-executing model online for beam size while optimizing** → *learn on direct observables (e.g. beam size); do inferred "measurements" (e.g. emittance)*
- **New algorithmic paradigm leveraging "Bayesian Algorithm Execution" (BAX) for 20x speedup in tuning**

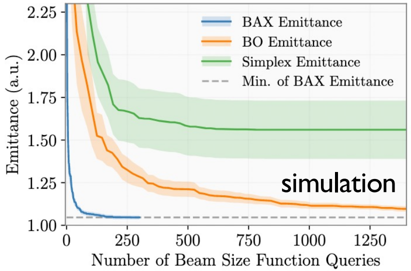


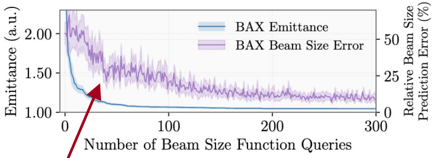




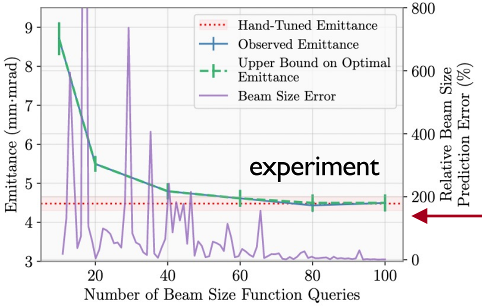
model is learned on-the-fly







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



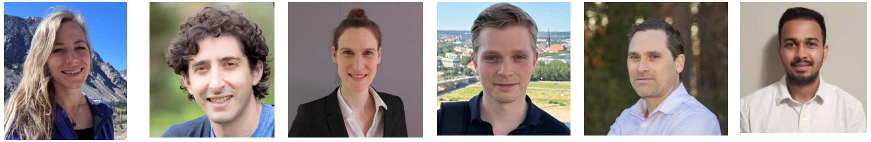
Found equivalent quality to hand-tuning in about 70 iterations *(estimate this would take a few minutes with computationally optimized routine)*

https://arxiv.org/abs/2209.04587

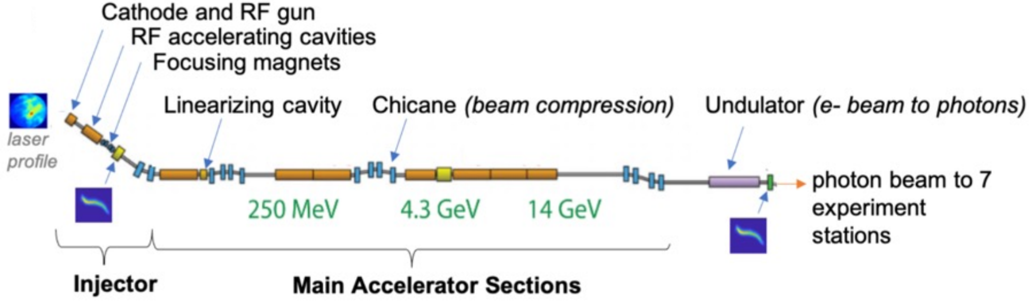
Paradigm shift in how tuning on indirectly computed beam measurements (such as emittance) is done, with 20x improvement over standard method for emittance tuning. → ***Now working to integrate into operations.***

→ ***Also now working to incorporate more informative global models /priors rather than learning the model from scratch each time.***

RL for LCLS Accelerator

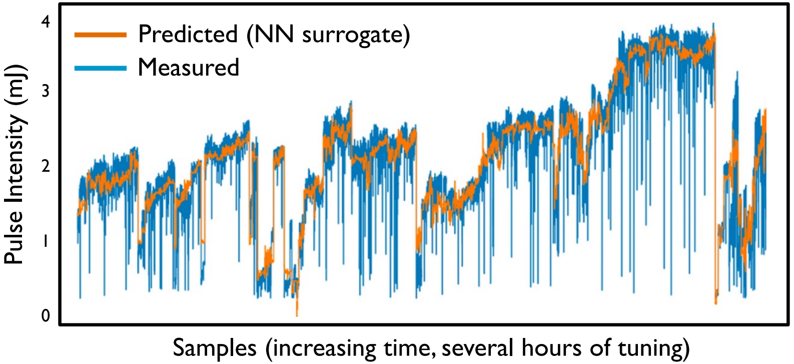


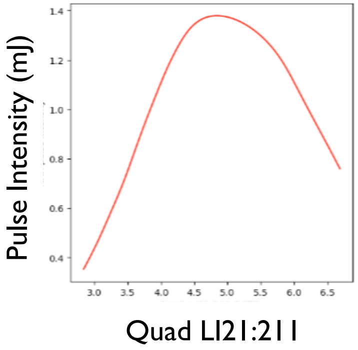
- Focusing on FEL pulse intensity tuning and quadrupole magnets first
- FEL is sensitive to focusing, trajectory; perturbing beam/feedbacks too much results in beam losses
- Using data-driven surrogates and differentiable sims (Cheetah and Bmad) to train agents (TD3, PPO)
- Iteratively add more data and variables:
  - Longitudinal phase space, spectra
  - RF phases and amp., undulator taper
  - Combine with photon beamline, trajectory control



~28 focusing magnets for FEL pulse intensity

(many more variables to include: steering, rf, taper, drive laser)





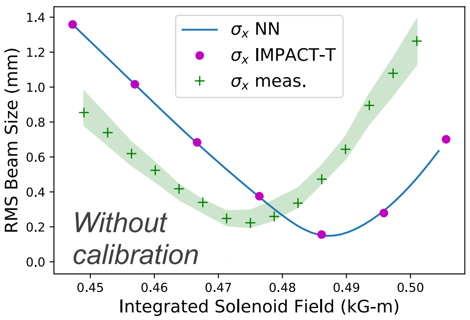
Finding Sources of Error Between Simulations and Measurements

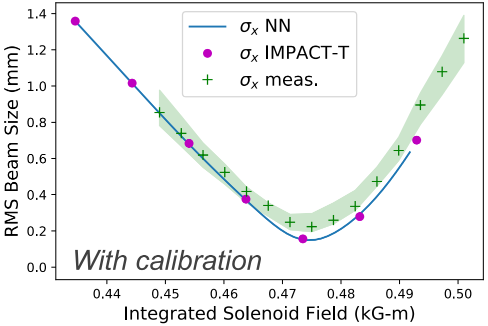
Many non-idealities not included in physics simulations:

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

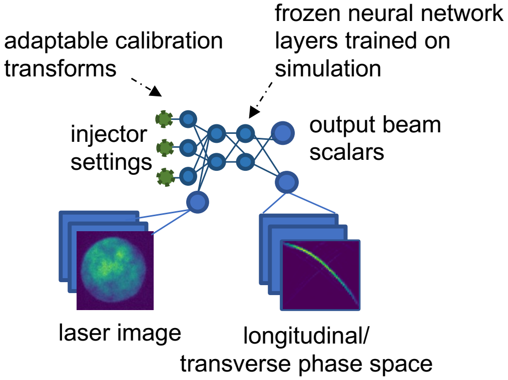
Want to identify these to get better understanding of machine performance

→ ***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)



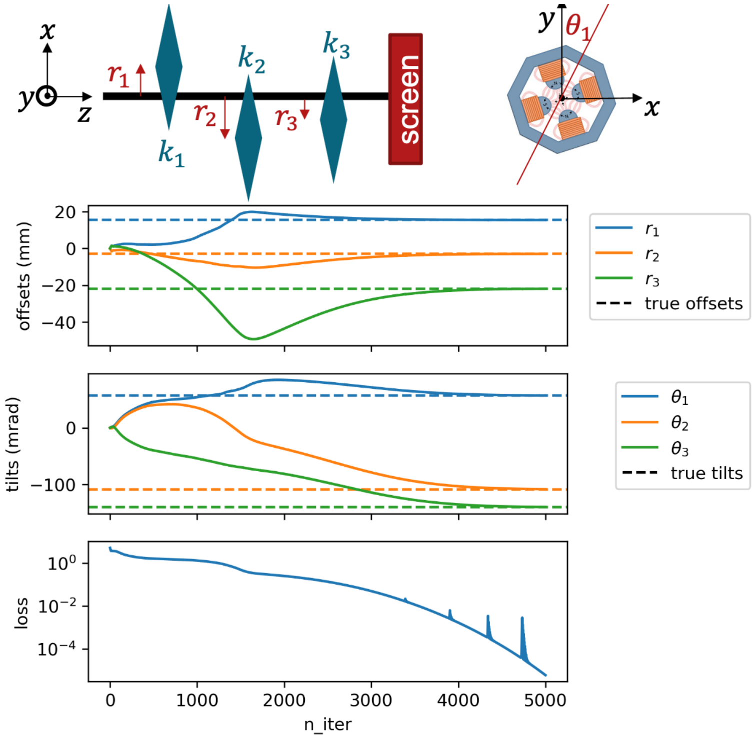
Inputs
Laser radius
Laser spot sizes
Pulse length
Charge
Solenoid
L0A phase
L0B phase
SQ quad
CQ quad
6 matching quads

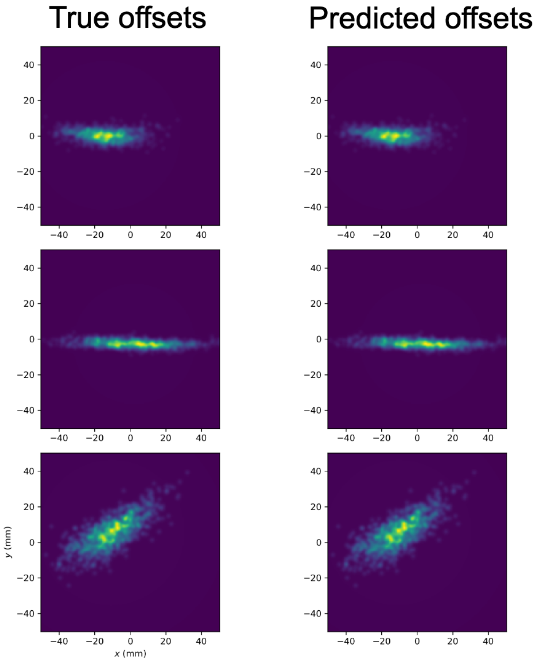
Outputs
Beam size (x,y)
Emittance (x,y)
Bunch length

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

Finding Sources of Error Between Simulations and Measurements

Same approach can be used with differentiable physics simulations





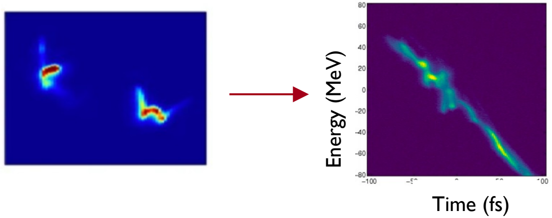
J.P. Gonzalez-Aguilera
https://accelconf.web.cern.ch/ipac2023/pdf/WEPA065.pdf

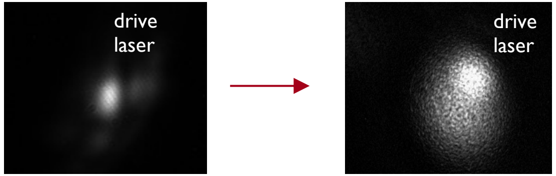
Differentiable simulations allow direct learning of calibrations while being constrained by the expected physics

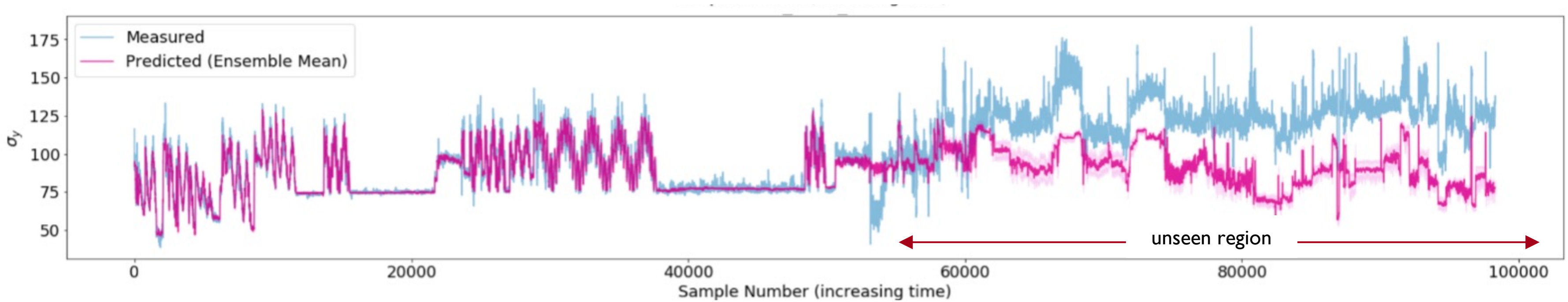
Distribution Shift is a Major Challenge in Particle Accelerators

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







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

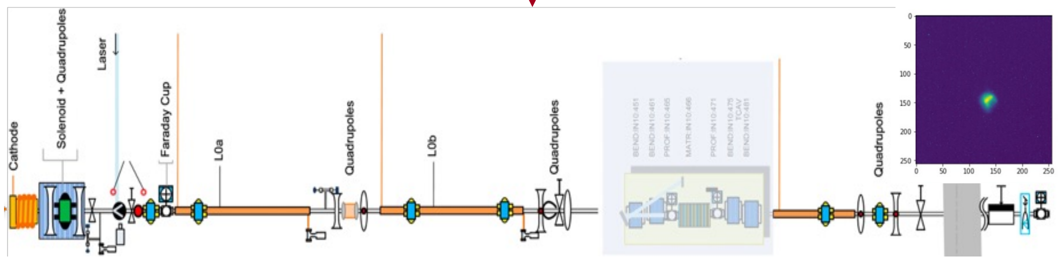
Reliable uncertainty estimates and model adaptation methods are key for putting online models to use operationally
Need fast ways of obtaining characterization data from accelerator

"Bayesian Exploration" for Efficient Characterization

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

Comprehensive ML Models of Injector

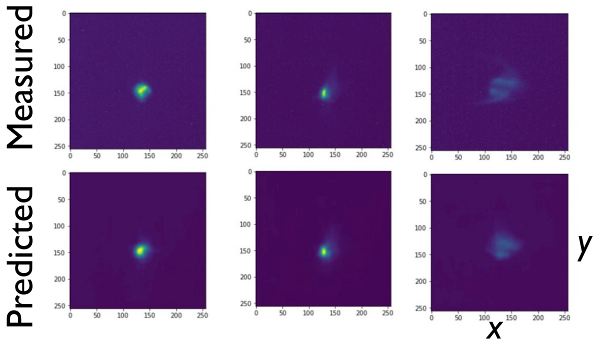
Setting changes on 10 variables (solenoid, bucking coil, corrector quads and matching quads)



FACET-II Injector

x-y emit, match, and beam images

transverse phase space



https://www.nature.com/articles/s41467-021-25757-3

- 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**

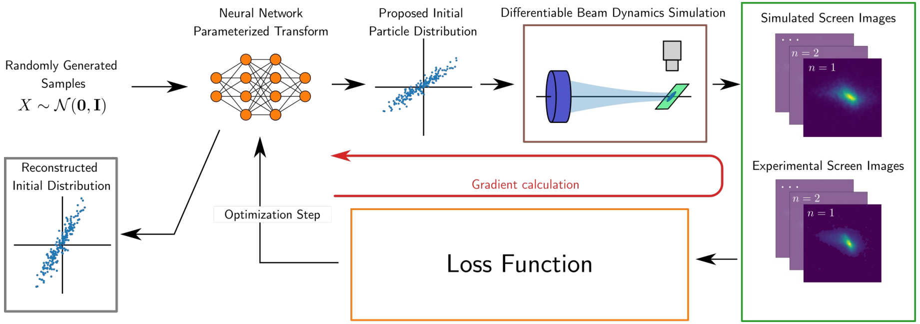
Use of Bayesian exploration to generate training data was sample-efficient, reduced burden of data cleaning, and resulted in a well-balanced distribution for the training data set over the input space. ML models were immediately useful for optimization.

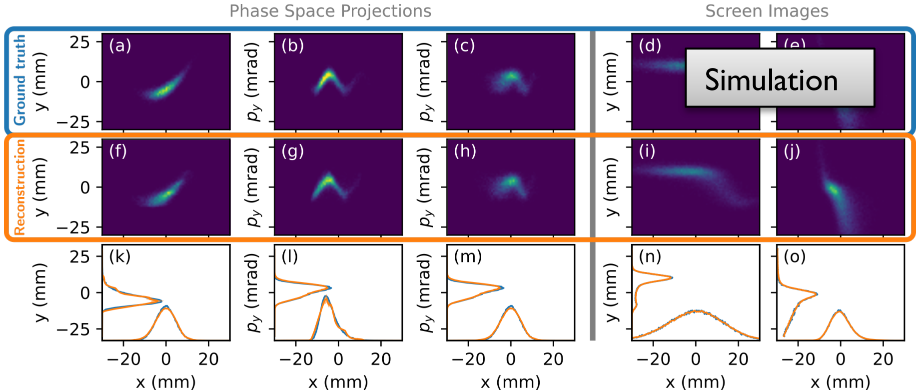
Phase Space Reconstruction with Differentiable Tracking Simulations

https://arxiv.org/abs/2404.10853

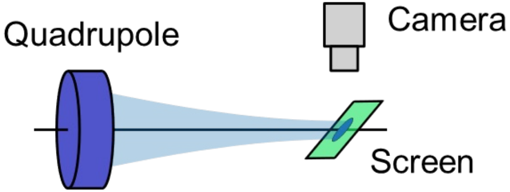
https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.130.145001

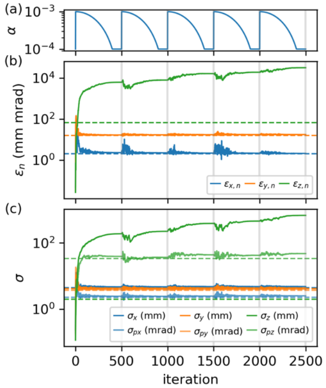
Differentiable pipeline for reconstructing 6D phase space distribution using neural network parameterization



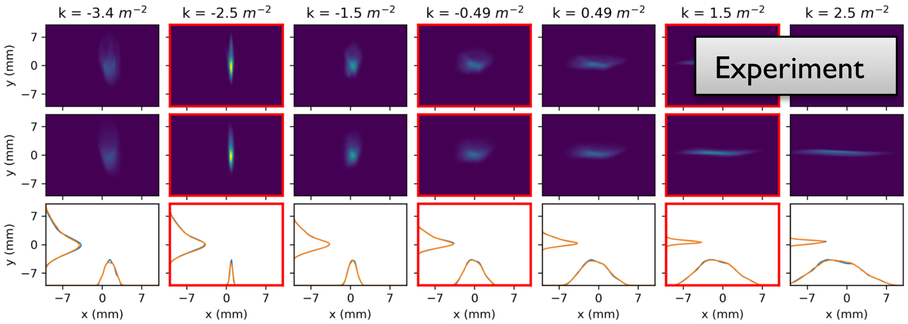


Reconstruct 4D phase space distribution + approx. energy spread from simple beamline diagnostic and 10 measurements





Confidence estimates

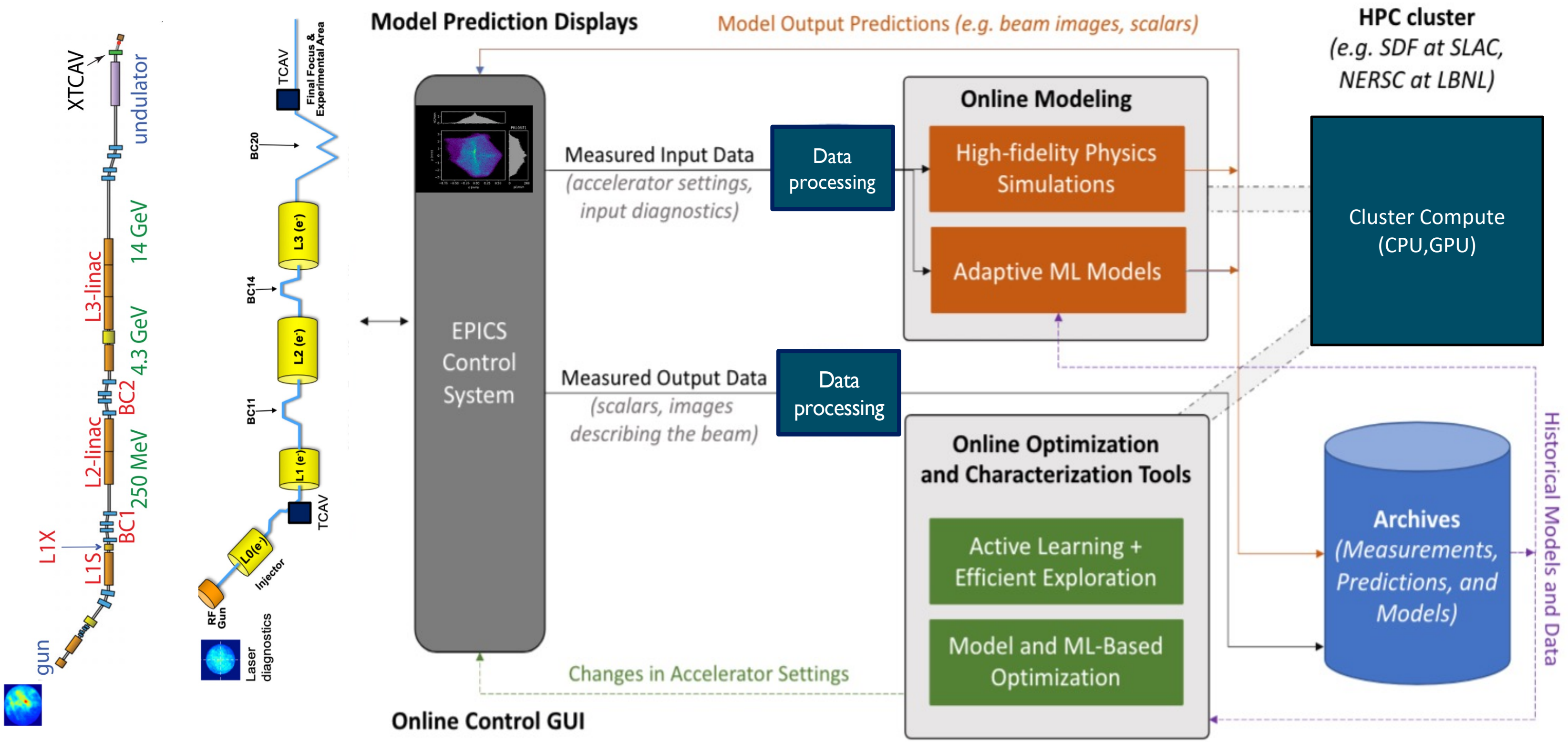


ML combined with differentiable simulations opens up a new paradigm for constructing detailed phase space diagnostics in a way that is computationally-efficient and sample-efficient

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

Digital Twin Infrastructure

Ecosystem of modular tools (can use independently)

LUME – simulation interfaces/wrappers in Python

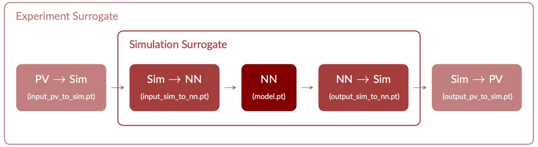
lume-model – wraps ML models, facilitates calibration

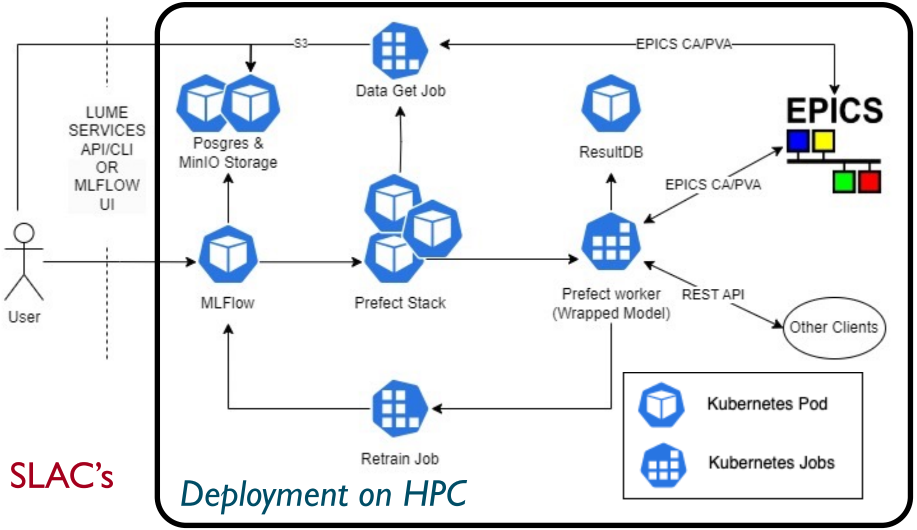
lume-services – online model deployment and orchestration

distgen – flexible creation of beam distributions

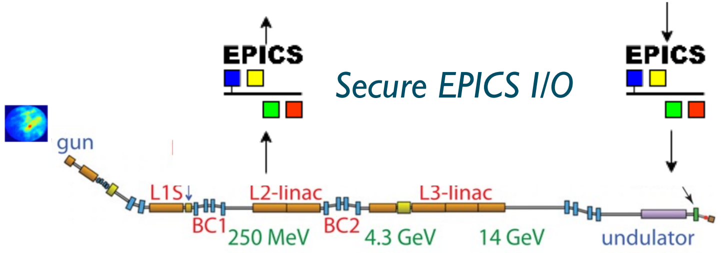
Integration with MLFlow for MLOps

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
- Collaboration with LBNL through SciDAC on "virtual accelerators"



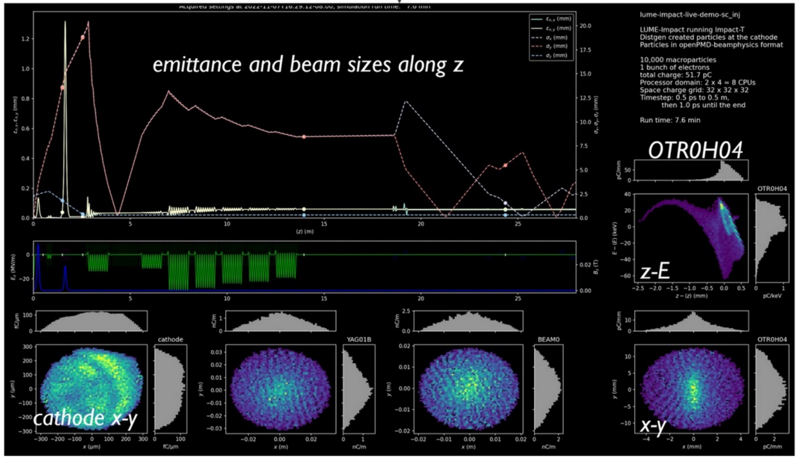
Substantial progress on deploying ML and Physics-based models and integrating with HPC in a portable way

Combining BO with Warm Starts from Online Physics Models

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

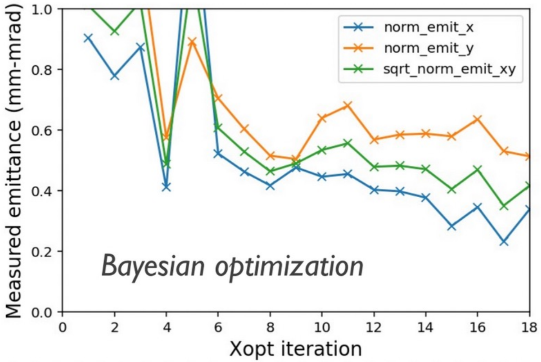


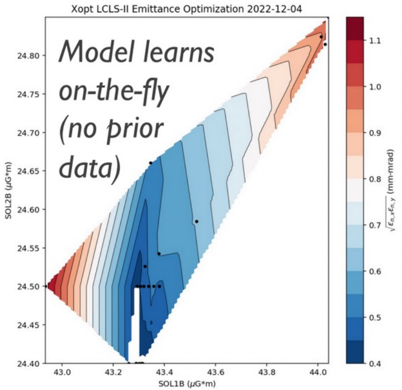
LCLS-II live sim: run on HPC and display in control room

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

Summary/Conclusions

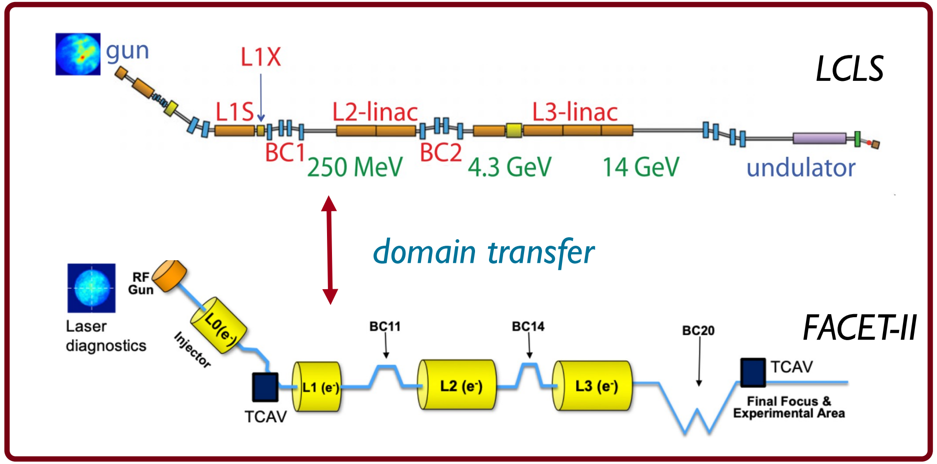
- Particle accelerators stand to benefit substantially from the development and deployment of modern digital twins
    - Faster optimization, new capabilities in beam customization, human-AI interaction
    - High impact for science that is supported by particle accelerators (and translations to industry/medicine)

- SLAC and collaborating labs (LBNL, JLab, FNAL, ANL) are building out infrastructure to deploy detailed physics simulations and ML models "online" with the control system → *community open source software is essential!*

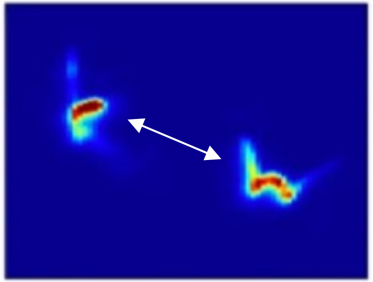
- Now scaling up small-scale demos of combining ML surrogate models, adaptive model calibration, automatic characterization, and integration into online control

→ **Many interesting problems to tackle**

→ **Accelerators are also interesting platforms for AIML research!**



fast dynamic beam customization

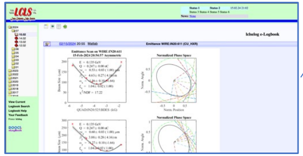


AIML + human feedback

Aim to tie together AIML to aid many different tasks toward autonomous accelerator control

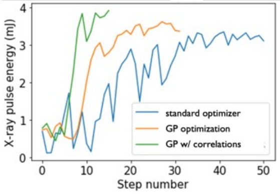
Human-computer interaction

Language modeling / multi-modal data
(e.g. electronic logbook)

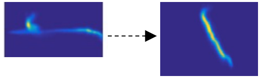


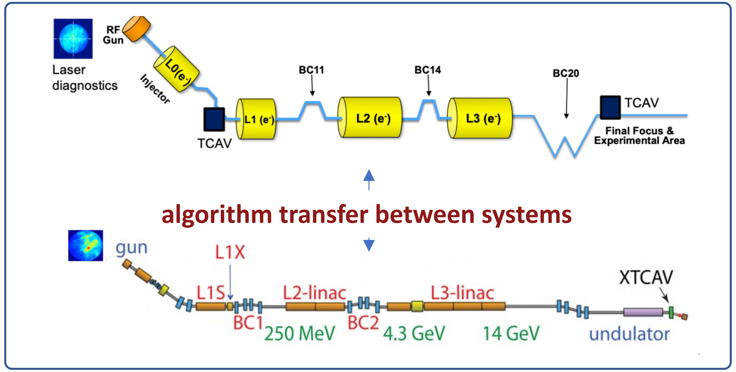
Data reduction/rejection *(kHz/MHz data streams)*
Event triggering

Automated control + optimization

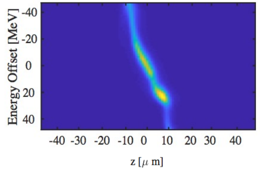


J. Duris et al., PRL, 2020



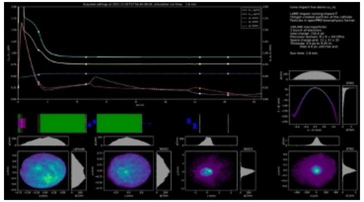


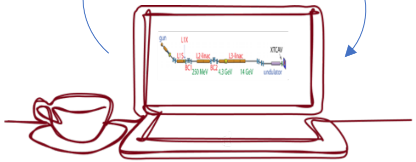
ML-enhanced diagnostics
(provide insight at faster rate, at higher resolution, non-invasively)



C. Emma et al., PRAB, 2018

Anomaly detection failure prediction
(plan maintenance; alert to changes in machine; alert to interesting science)

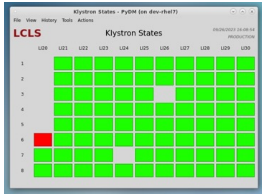




Digital twins + online modeling
(fast sims, differentiable sims, model calibration, model adaptation)

Extract unknown relationships + correlations
(feed into future control / design)

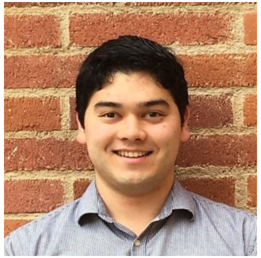
+ need uncertainty quantification for all
+ can incorporate physics information in all



Backups










Thanks to the core team at SLAC working on various digital twin and AIML technologies and infrastructure, and many other collaborators at other labs!

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*)

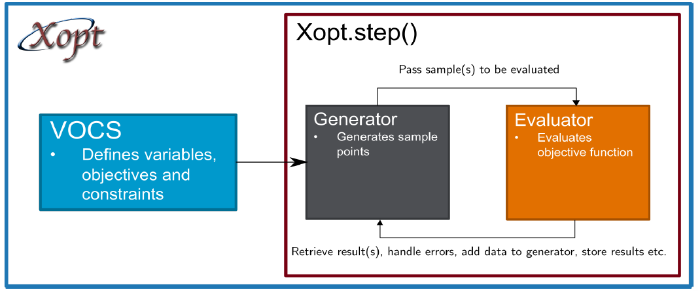
Visual control room interface (*e.g. Badger*)

Simulation drivers (*e.g. LUME*)

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

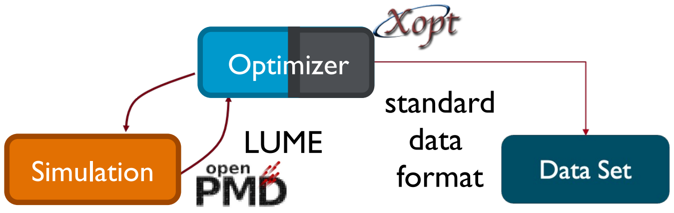
Online model deployment (*LUME-services*)

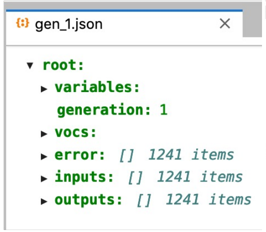
More details at https://www.lume.science/

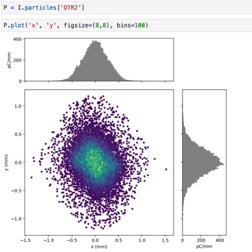


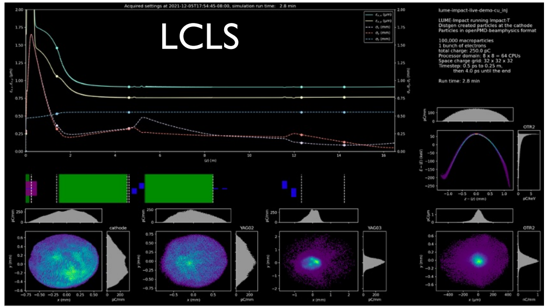

```
vocs:
  name: TNK_test
  variables:
    x1: [0, 3.14159]
    x2: [0, 3.14159]
  objectives: {y1: MINIMIZE}
  constraints:
    c1: [GREATER_THAN, 0]
    c2: ['LESS_THAN', 0.5]
```

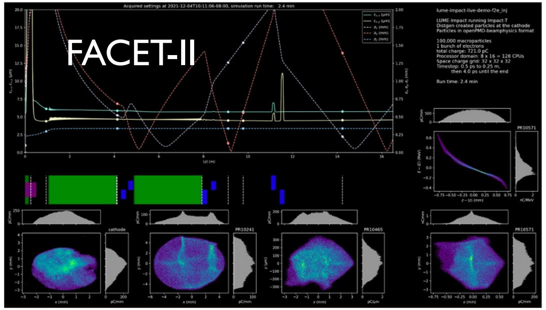
```
algorithm:
  name: bayesian_exploration
  options:
      n_initial_samples: 5
      n_steps: 25
      generator_options:
          batch_size: 1
          #sigma: [[0.01, 0.0],
          use_gpu: False
```











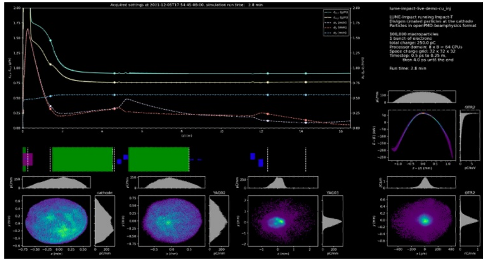
Online Impact-T simulation and live display; trivial to get running on FACET-II using same software tools as the LCLS injector

Modular open-source software has been essential for our work.

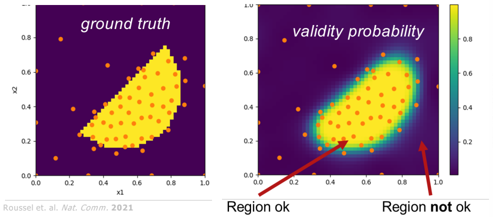
Broad Research Program at SLAC in AI/ML for Accelerators

(1) Developing new approaches for accelerator optimization/characterization and faster higher-fidelity system modeling, (2) developing portable software tools to support end-to-end AI/ML workflows, (3) helping integrating these into regular use

Online prediction with physics sims and **fast/accurate ML system models**



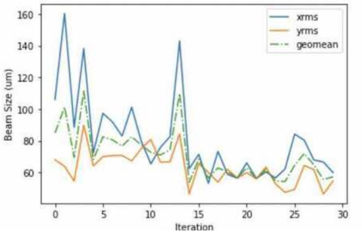
Efficient, safe optimization algorithms



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

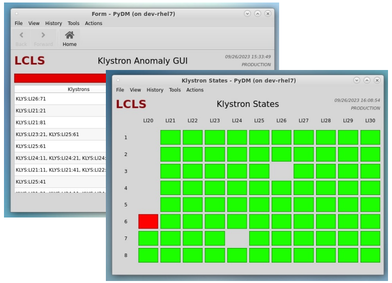
Output constraints learned on-the-fly

Adhere to constraints and balance multiple targets

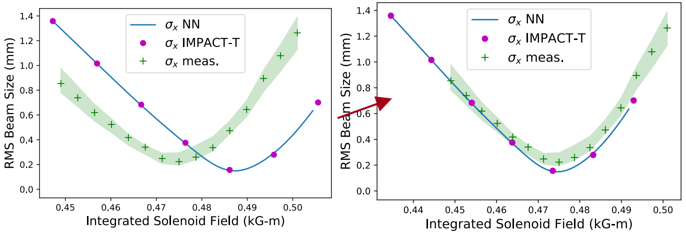


Challenging problems: e.g. sextupole tuning

Anomaly detection

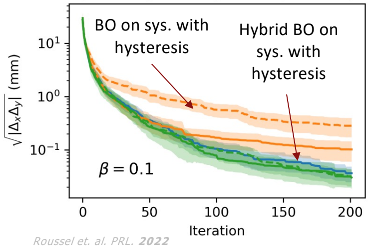


Adaptation of models and **identification of sources of deviation** between simulations and as-built machine



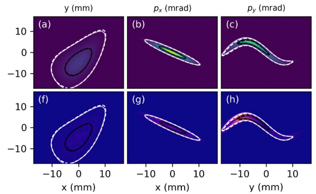
Combining physics and ML for better performance

Hysteresis-aware optimization



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

Differentiable simulations + ML for 6D phase space reconstruction

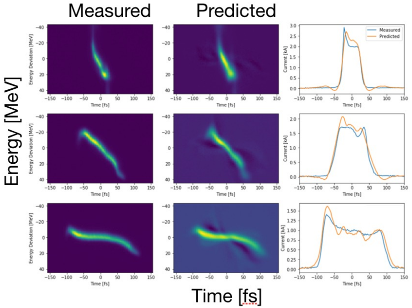


Roussel et. al. PRL. ***2023***

ML-enhanced diagnostics

Rapid analysis/virtual diagnostics

Shot-to-shot predictions at beam rate



C. Emma, et al. – PRAB ***21****, 112802 (2018)*

Many solutions put into reusable open-source software (e.g. Xopt/Badger) demoed at many facilities

AI/ML enables fundamentally new capabilities across a broad range of applications → highly promising from initial demos.

Opportunities for AIML Accelerator Research
(mix of needs from science side + compelling areas in AIML)

- Pushing to higher-dimensional algorithms (more comprehensive, precise tuning); incorporation of multiple signals on photon side to characterize beam quality

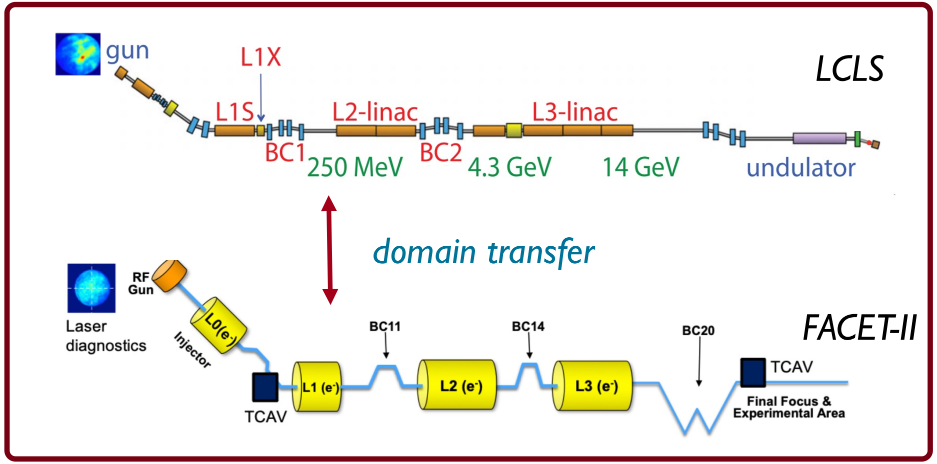
- Sample-efficient adaptation across setups needed
  (different charges, beam phase space, multi-bunch)

- Enabling fundamentally new capabilities in beam physics / photon science
  - *FACET-II "extreme beams"; highly sensitive*
  - *Photon science requiring precise dynamic control*

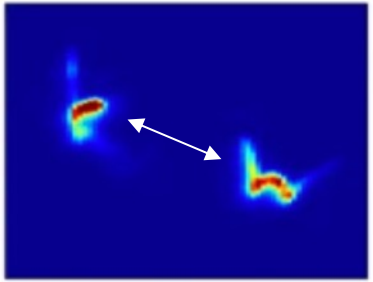
- Comprehensive online system modeling + ML-based optimization
  - *Physics sims + ML surrogates being deployed on local HPC connected to control system*
  - *"digital twins" + "outer-loop" applications of interest to ASCR*

- AI with human feedback → *human-AI interaction in the control room is a current area of study*

- Transfer learning between LCLS/LCLS-II/FACET-II
  → *Similar layouts, component design, beam diagnostics, user needs (e.g. scan two bunches)*



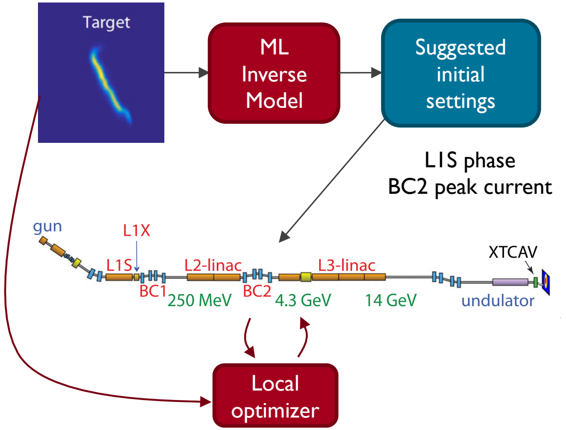
fast dynamic beam customization



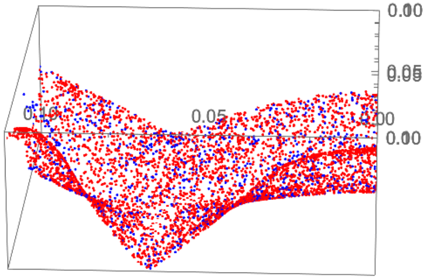
AIML + human feedback

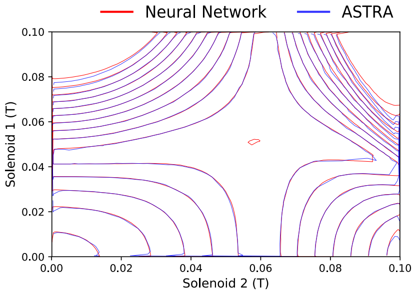
Warm starts for optimization

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

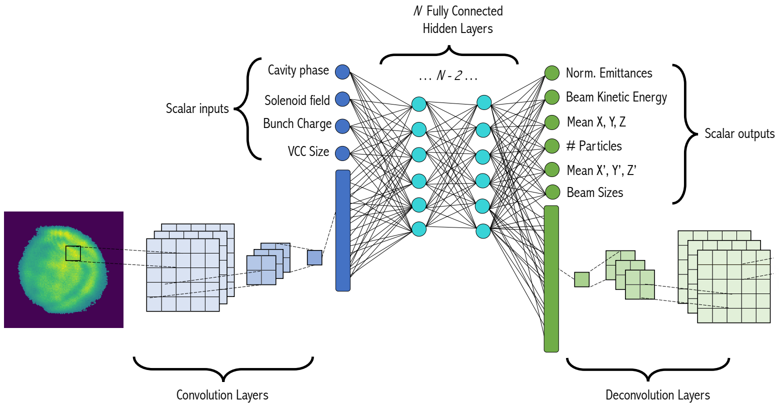


Smooth interpolation
Example σ_x surface from 2D scan, LCLS-II Injector

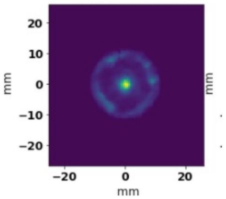


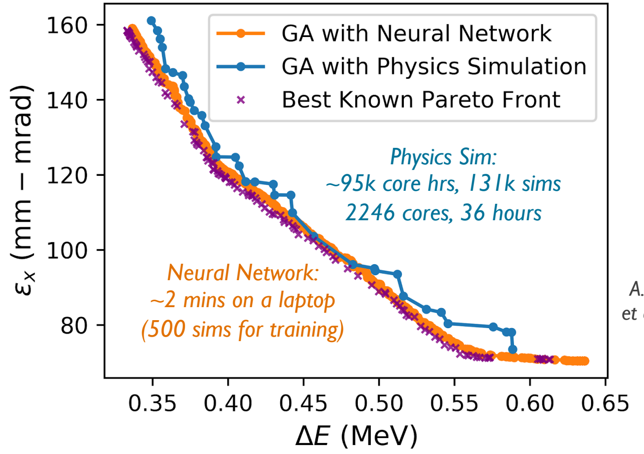


A. Edelen et al., NeurIPS 2019



L. Gupta, et al, MLST, 2021





A. Edelen et al., PRAB, 2020

Include high-dimensional input information → better output predictions

Surrogate-boosted design optimization

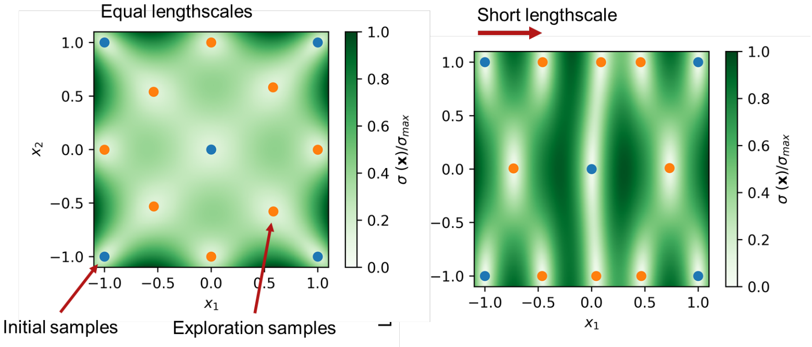
Efficient Characterization with Bayesian Exploration

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

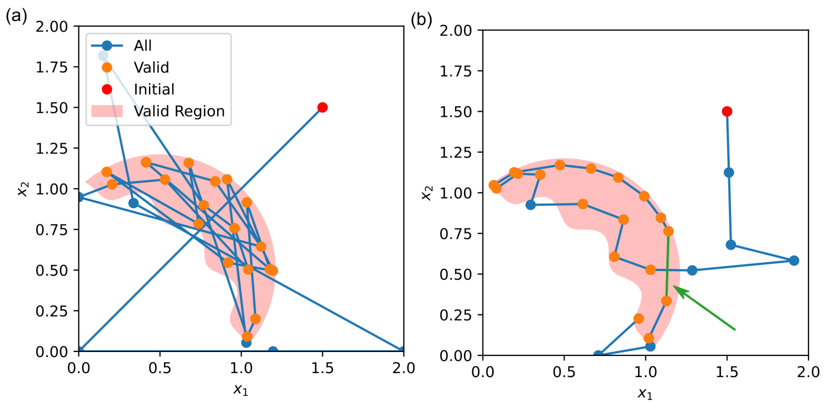
R. Roussel et. al.
Nat. Comm. **2021**

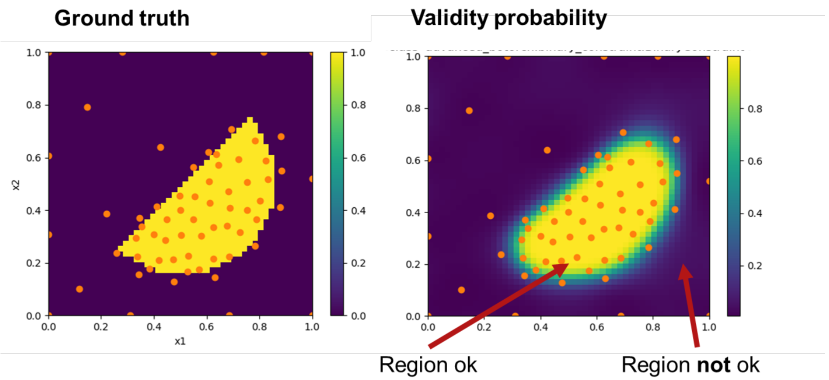
proximal biasing

adaptive sampling



learning constraints

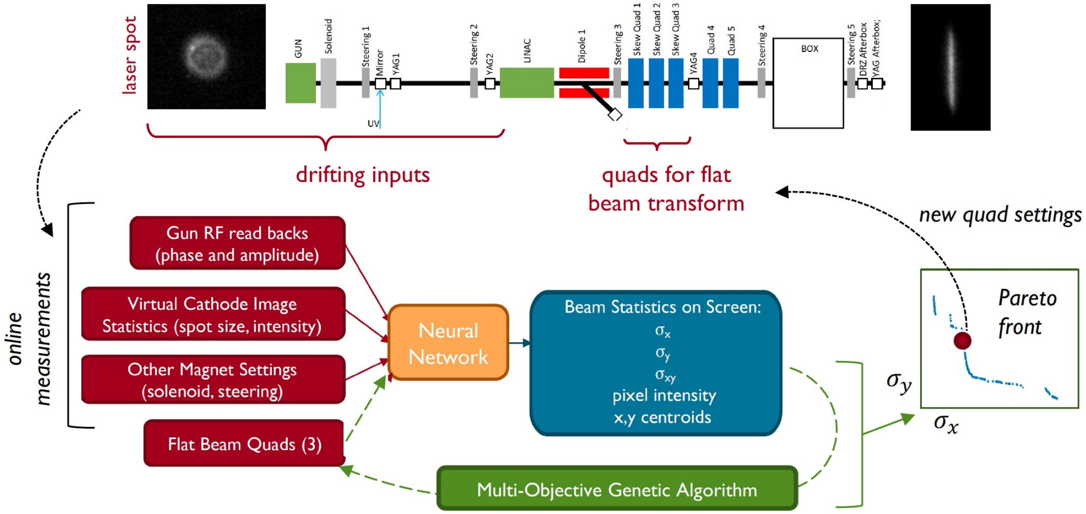




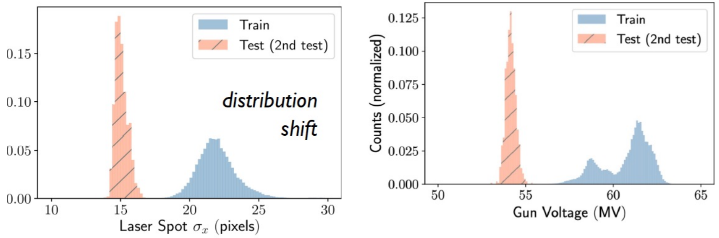
Enables sample-efficient characterization of high-dimensional spaces, while respecting both input and output constraints

Example: Warm Starts from Online Models

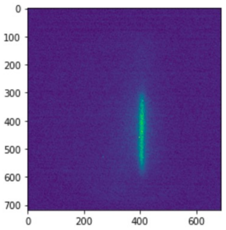
E. Cropp et al., in preparation

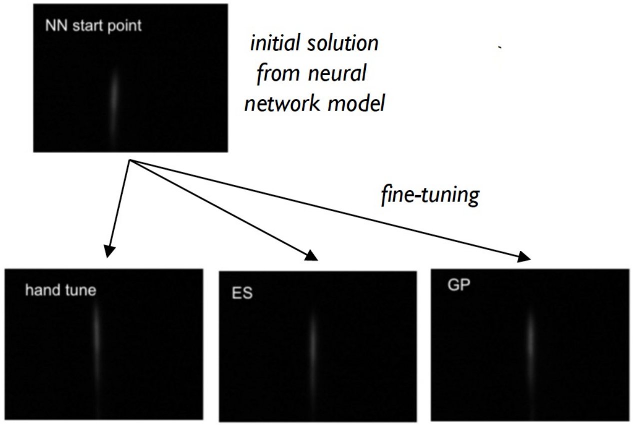


Can work even under distribution shift



- 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



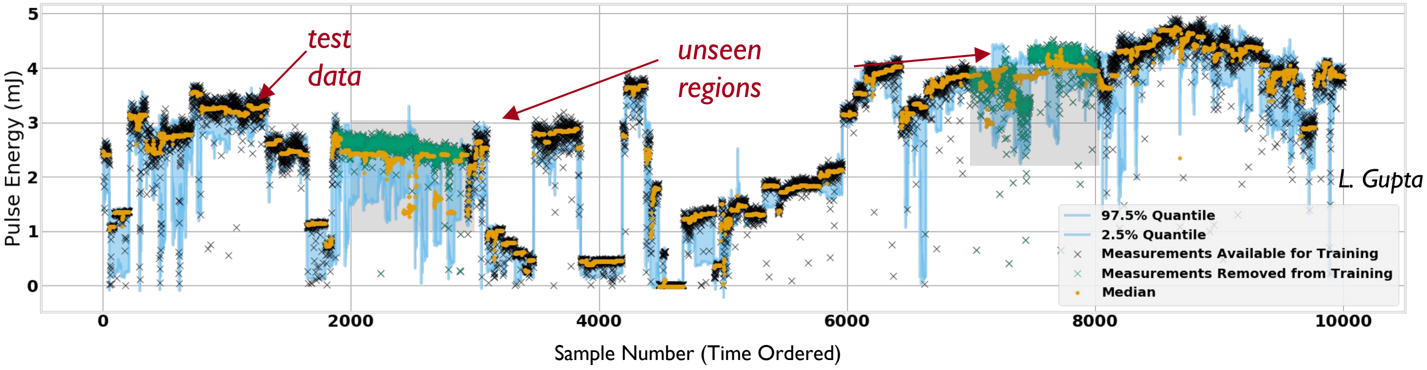


Hand-tuning in seconds vs. tens of minutes

Boost in convergence speed for other algorithms

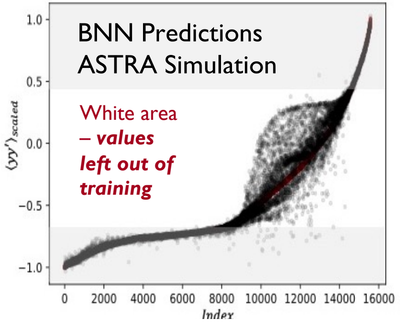
Uncertainty Quantification / Robust Modeling

Essential for decision making under uncertainty (e.g. safe opt., intelligent sampling, virtual diagnostics)



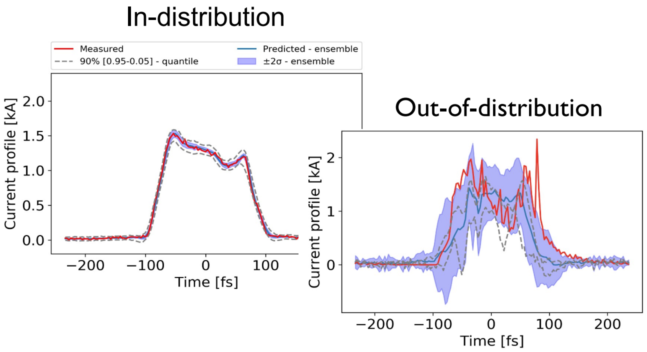
Current approaches
- Ensembles
- Gaussian Processes
- Bayesian NNs
- Quantile Regression

Neural network with quantile regression predicting FEL pulse energy at LCLS



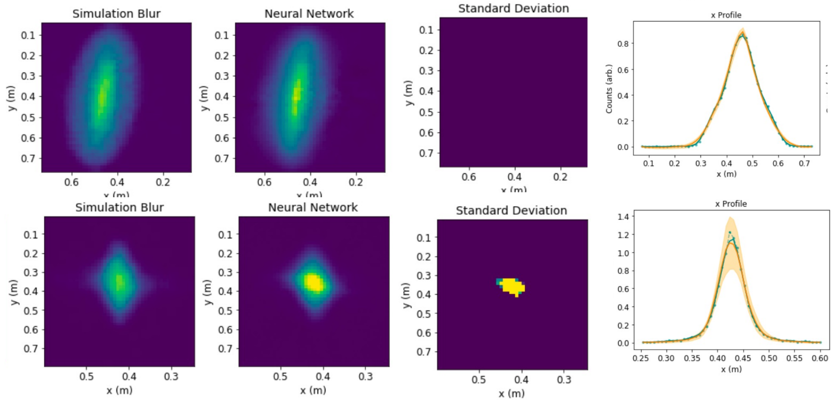
Scalar parameters for the LCLS-II injector (Bayesian neural network)

A. Mishra et. al., PRAB, 2021



longitudinal phase space (quantile regression + ensemble)

O. Convery, et al., PRAB, 2021



LCLS injector transverse phase space (ensemble)