Opportunities in AI/ML for CCC

Auralee Edelen

(with input/examples from many colleagues, especially:
R. Roussel, C. Emma, J. Duris, A. Hanuka, C. Mayes, D. Ratner, A. Scheinker,
N. Neveu, L. Gupta, B. O'Shea, E. Cropp, P. Musumeci, A. Mishra)

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Places for AI/ML to contribute

Design optimization

- More efficient search of computationally-expensive simulations *(e.g. multi-objective, multi-fidelity Bayesian optimization)*

- Fast upstream models to aid start-to-end optimization

- Can leverage standards + uniform tools for data and I/O of accelerator simulations being used in AI/ML *(e.g. LUME, xopt)*

Online modeling and control

- Fast feed-forward corrections *(e.g. RF, trajectory; can also help reduce RF costs)*

- Sample-efficient online characterization and optimization

- Finding sources of systematic error between simulations and real machine, tracking time-varying deviations *(e.g. can aid meeting of desired tolerances and improve physics models)*

- Online models to provide additional diagnostic information

Fault detection and prediction

- Exclude faulty read-backs from feedback *(e.g. BPMs)*

- Identify (and possibly compensate for) impending RF trips

Simulation and Modeling Infrastructure

Standards for easy interfacing of simulations and optimizers

Xopt

CNSGA, Bayesian algorithms, sampler

https://christophermayes.github.io/Xopt/index.html

Optimizer

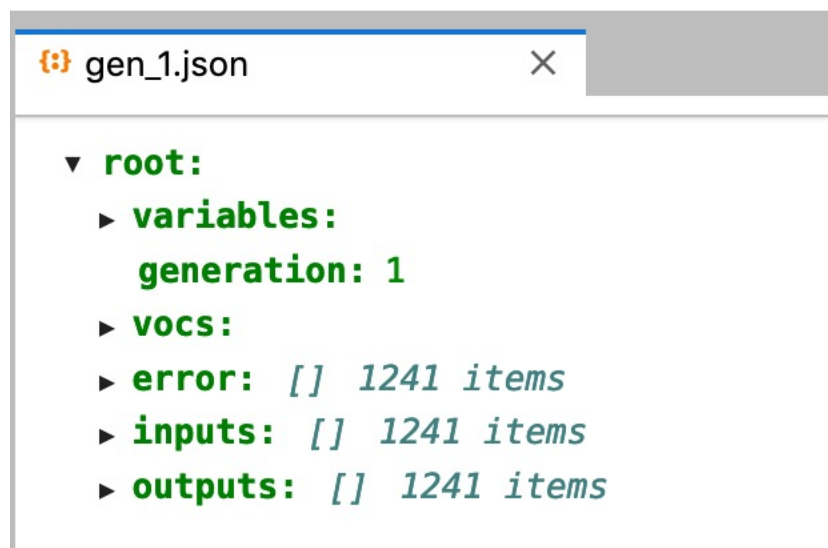
Simulation

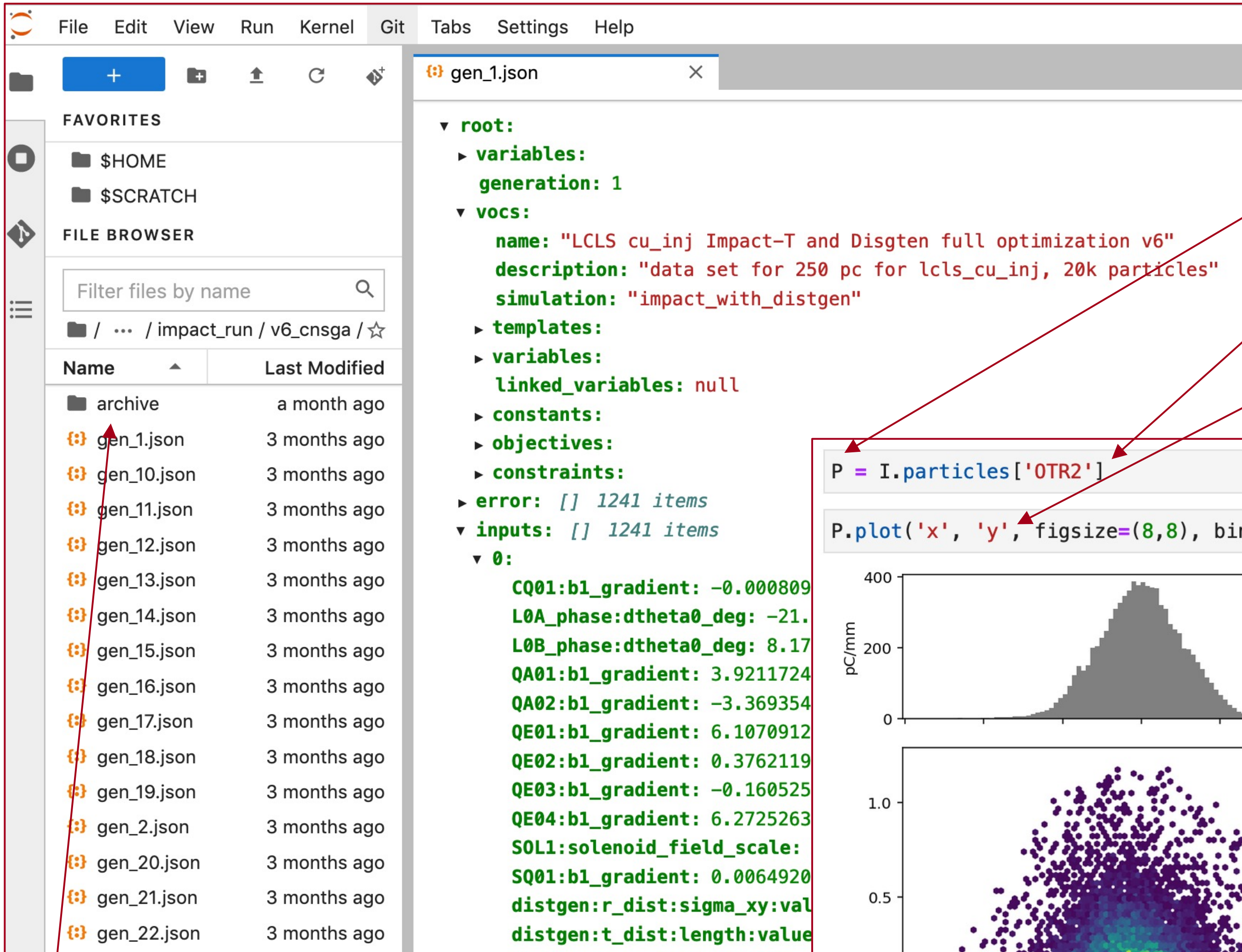
Data Set

LUME – light source unified modeling environment

https://www.lume.science/

Impact
ASTRA
GPT
Bmad
Genesis
SRW
work in progress:
elegant

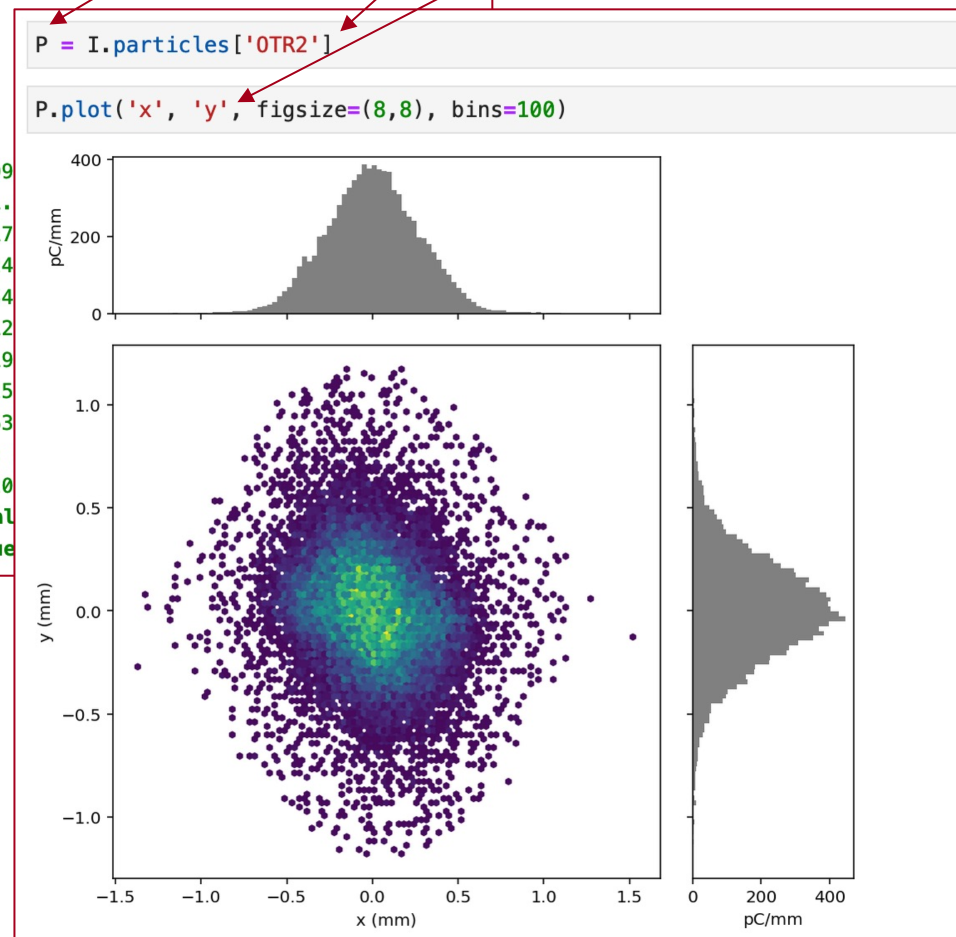




particle group

location

select projection to plot

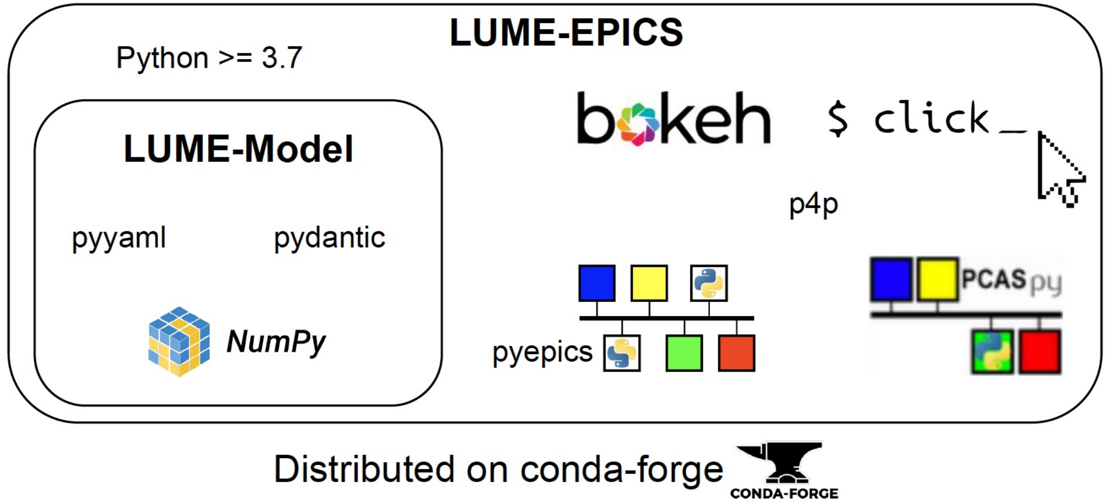


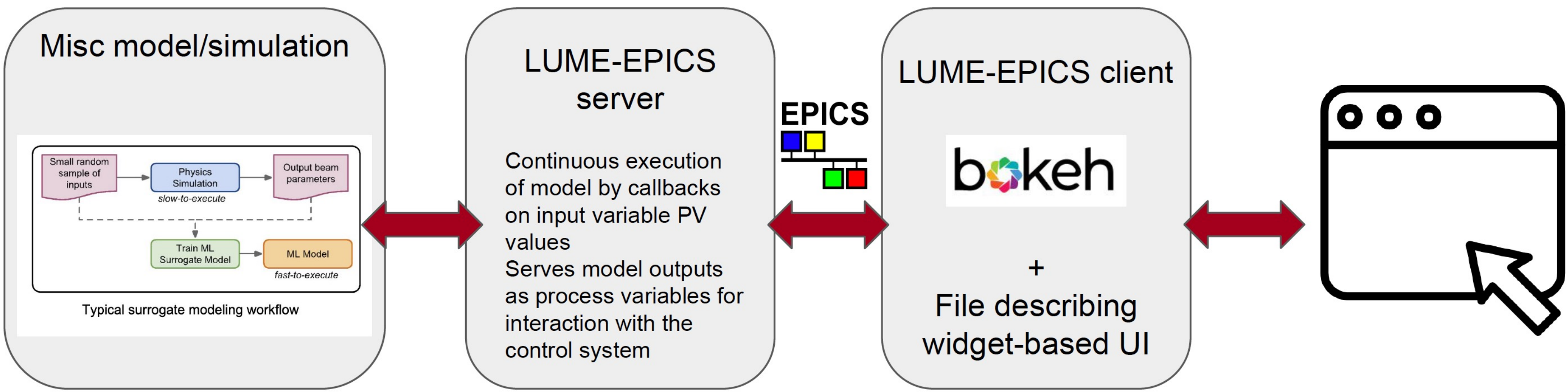
h5 files with beam distributions

→ easy to use with open-pmd-beamphysics

https://github.com/ChristopherMayes/openPMD-beamphysics

LUME-EPICS: Towards online execution

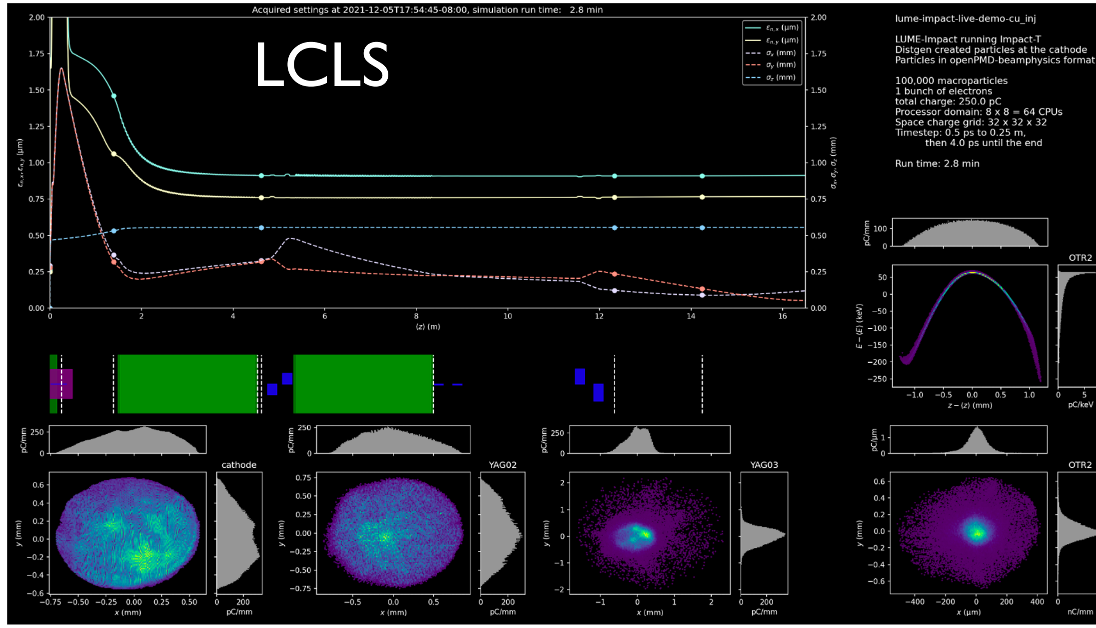


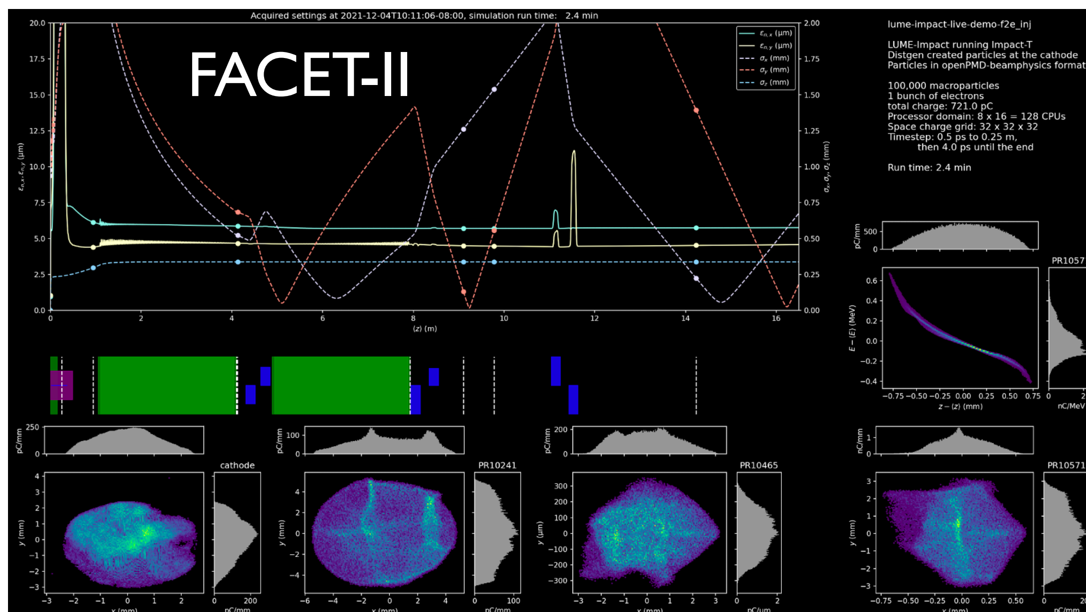


IMPACT-T models running online *(LCLS and FACET-II injectors)*

Read inputs online (including laser distribution)

Standard interfaces make this easily extendable to new systems





Optimization Methods

Optimization approaches can leverage different amounts of data

assumed knowledge of machine

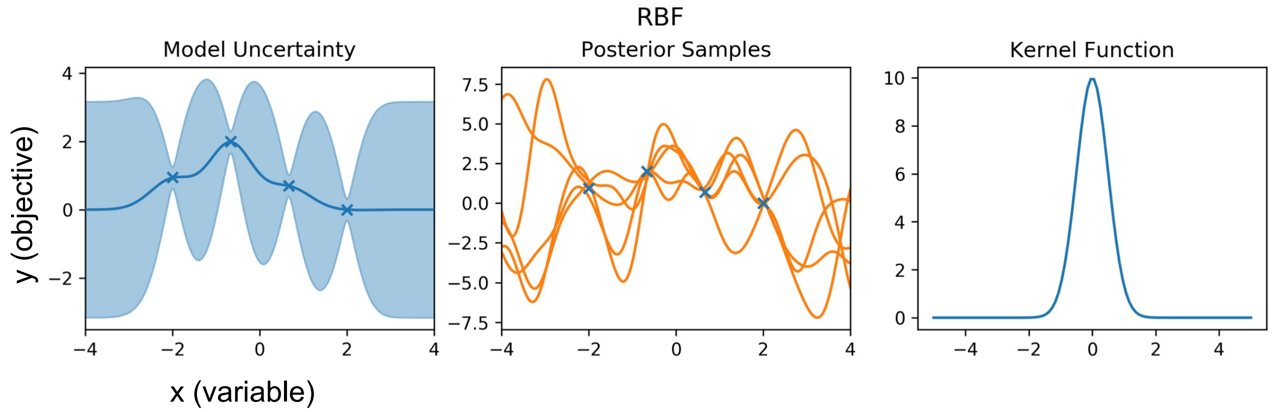
less ←→ more

Model-Free Optimization	Model-guided Optimization	Global Modeling + Feedforward Corrections
Observe performance change after a setting adjustment	*Update a model during each search step*	*Make fast / accurate system model*
→ *estimate direction toward improvement*	→ *use model to help select the next point*	→ *provide guess for good settings* → *make predictions about machine*
gradient descent simplex	Bayesian optimization Reinforcement learning	ML system models + inverse models

Bayesian Optimization

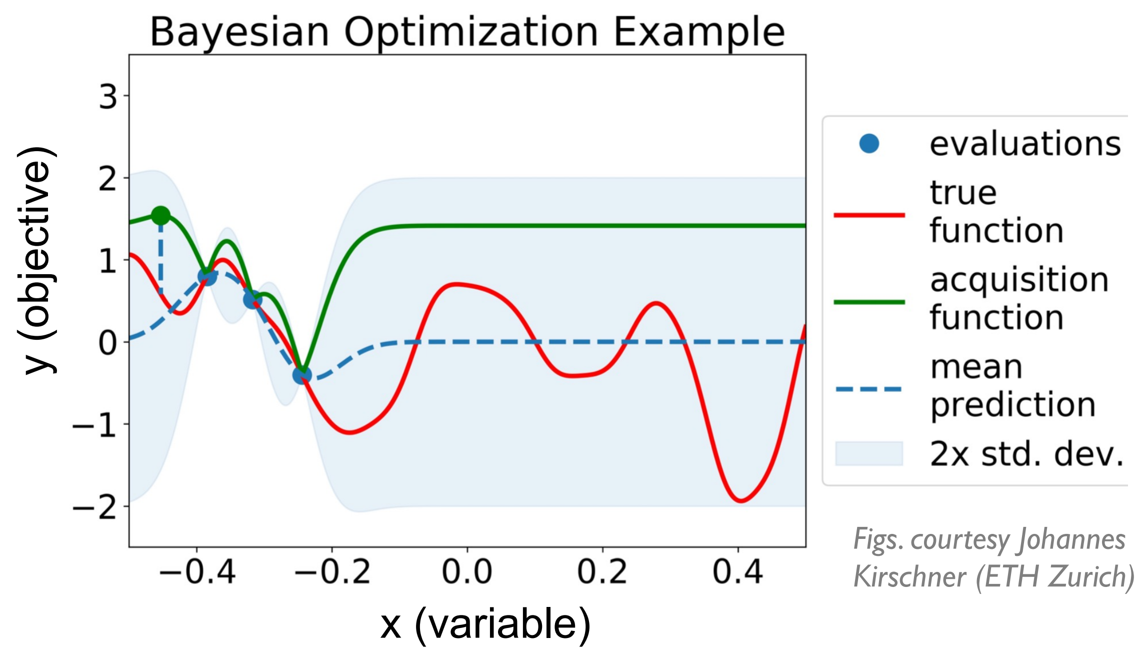
Set up probabilistic model

→ e.g. Gaussian Process



Iteratively refit model while sampling new points

Use model predictions and uncertainty to guide search for optimum while sampling



Figs. courtesy Johannes Kirschner (ETH Zurich)

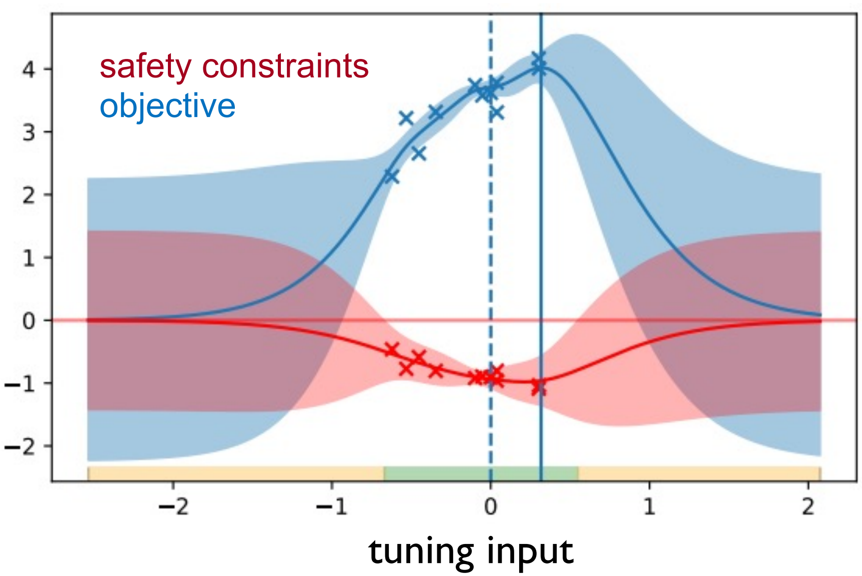
J. Kirschner et al., ICML 2019

Safe Optimization: Example on SwissFEL

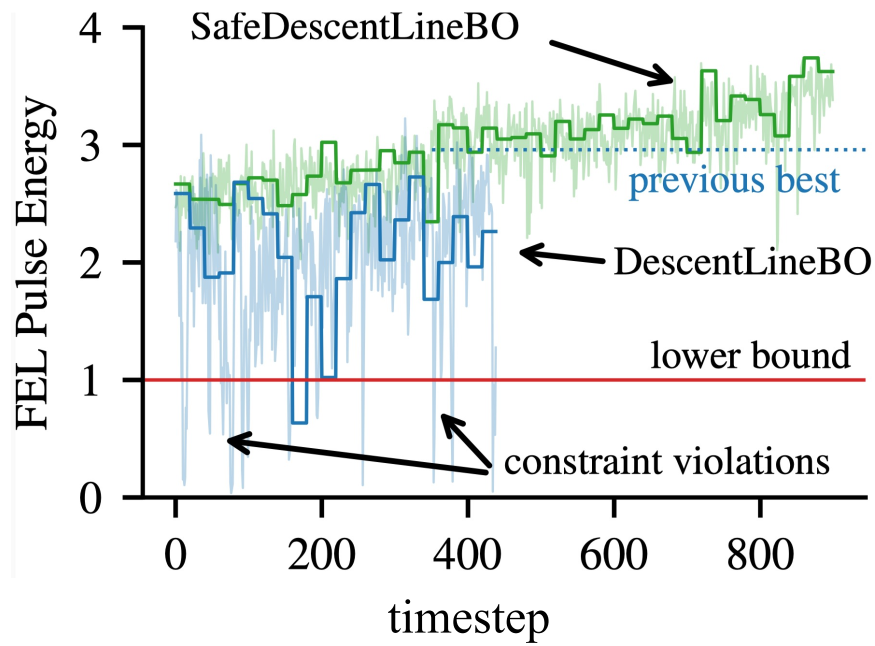
Don't just want to maximize FEL energy → we have other requirements

- pulse energy drops below certain level → *angry users!*
- beam losses go above a certain threshold → *damage machine!*

Add these requirements as safety constraints in Bayesian optimization



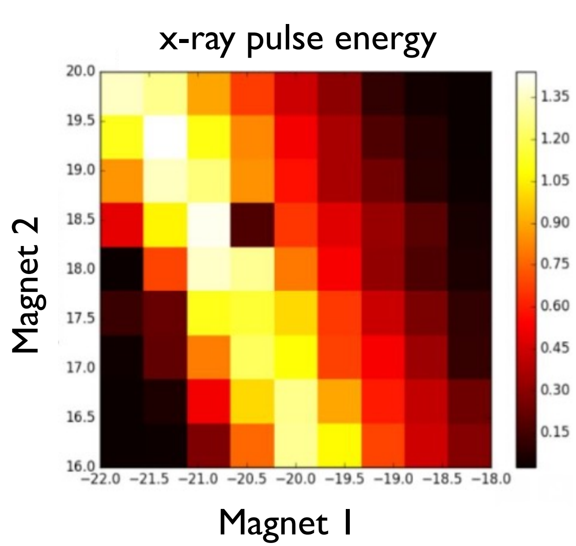
GP output for one timestep

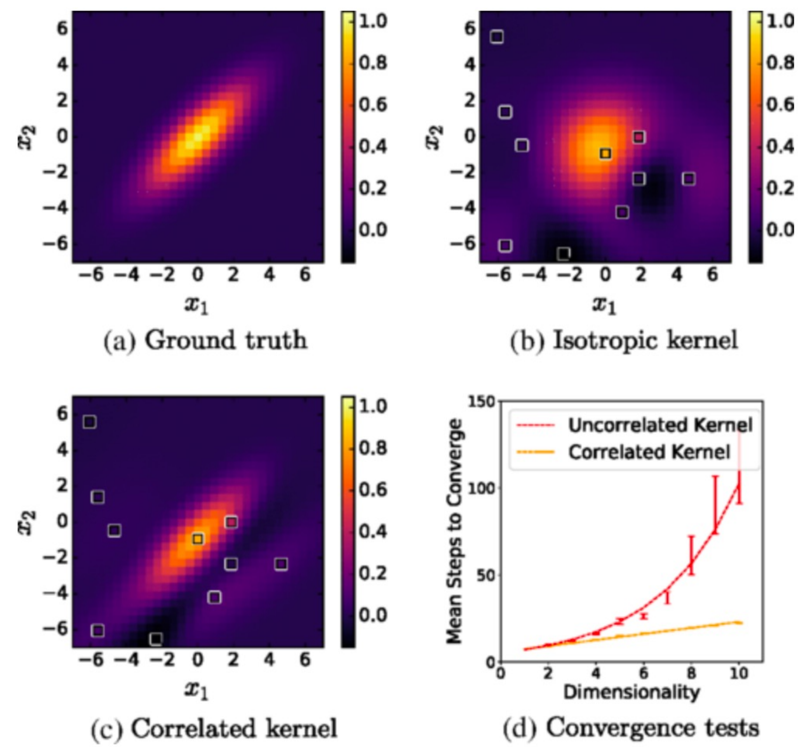


Model-informed Bayesian optimization

Can design GP kernel based on expected physics

- GP optimization at LCLS→ tune focusing magnets to maximize FEL pulse energy
- Make GP kernel informed by how quads correlate with FEL





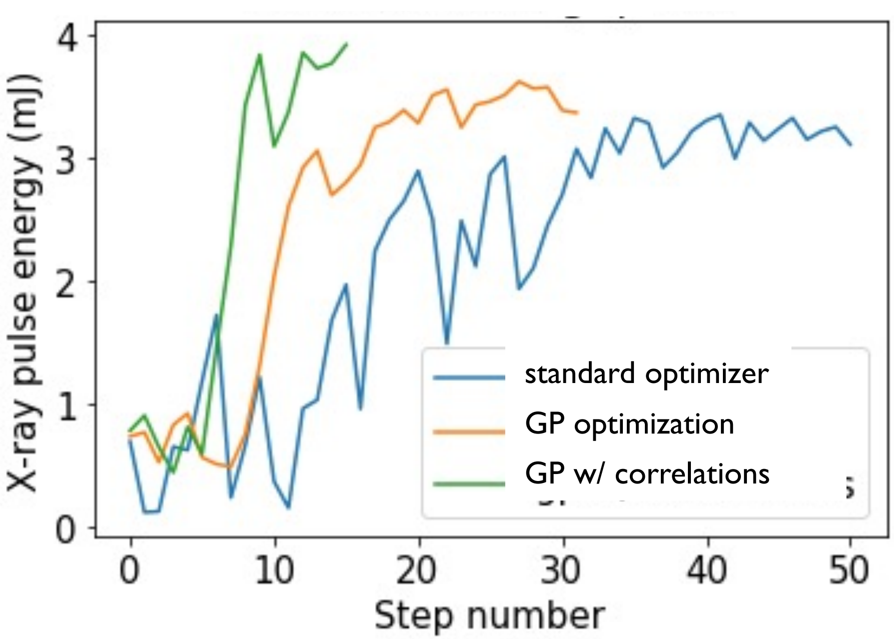
Including expected correlation improves ability to model the data with fewer samples

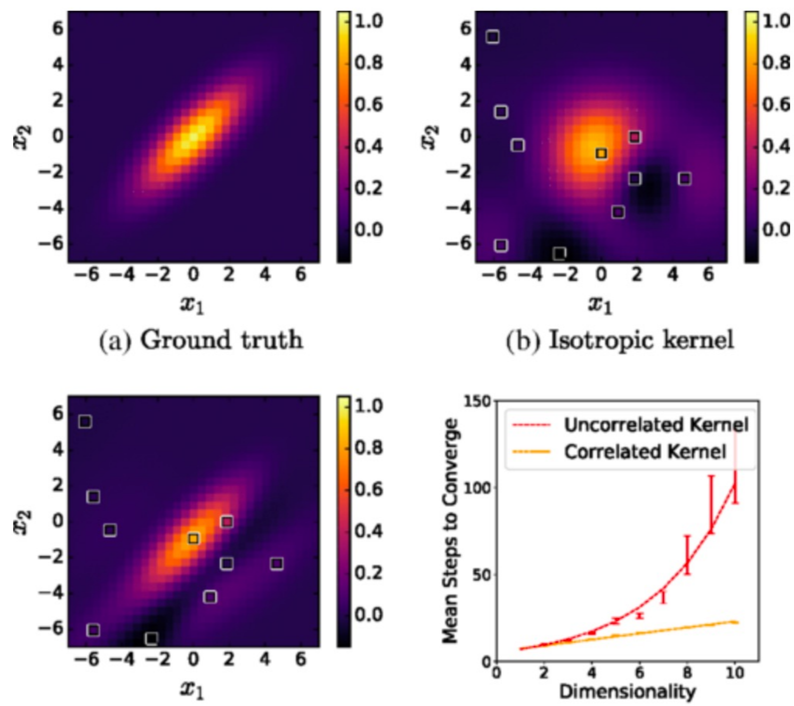
J. Duris et al., PRL, 2020
https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.124.124801

Model-informed Bayesian optimization

Can design GP kernel based on expected physics

- GP optimization at LCLS→ tune focusing magnets to maximize FEL pulse energy
- Make GP kernel informed by how quads correlate with FEL





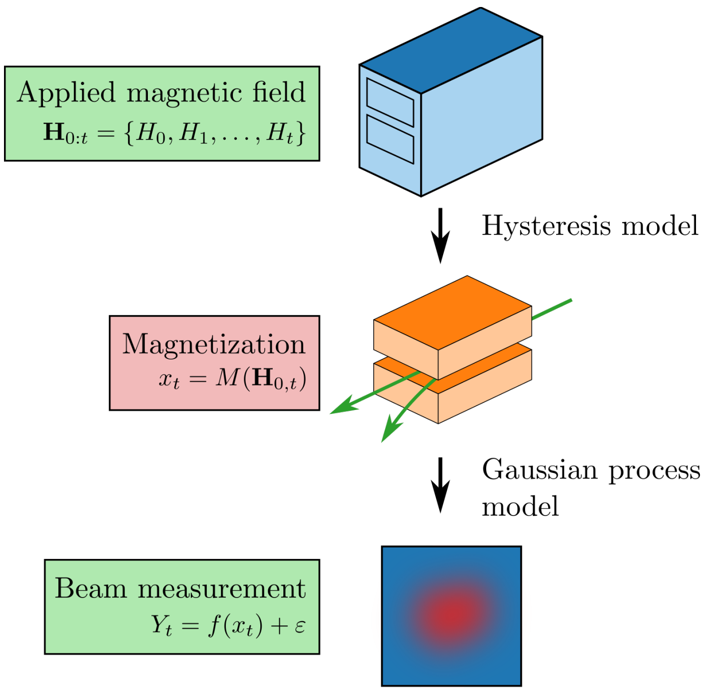
(a) Ground truth (b) Isotropic kernel

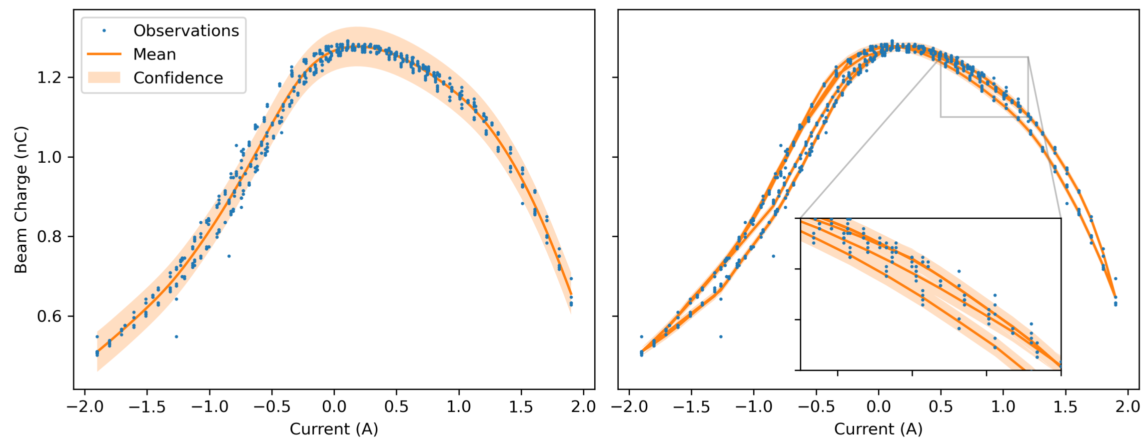
(c) Correlated kernel (d) Convergence tests

Including expected correlation improves ability to model the data with fewer samples

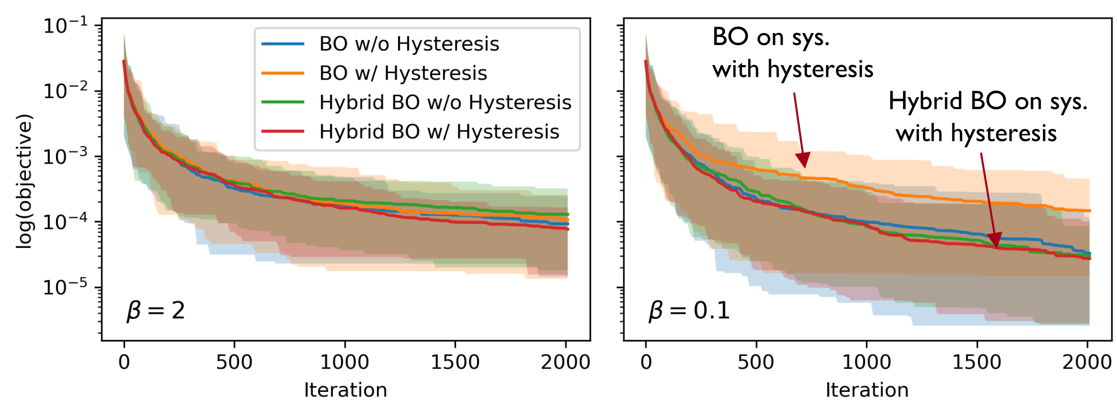
J. Duris et al., PRL, 2020
https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.124.124801

Differentiable Hysteresis Modeling for Accelerators





Joint modeling of hysteresis and beam propagation



Optimization improvements when including hysteresis

R. Roussel

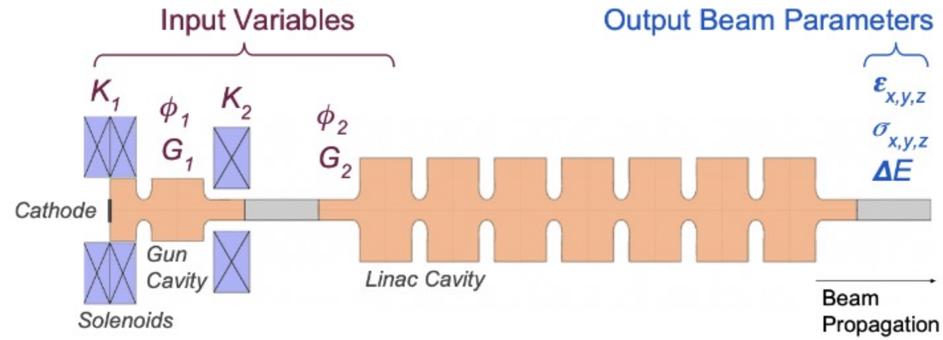
Multi-objective Bayesian optimization

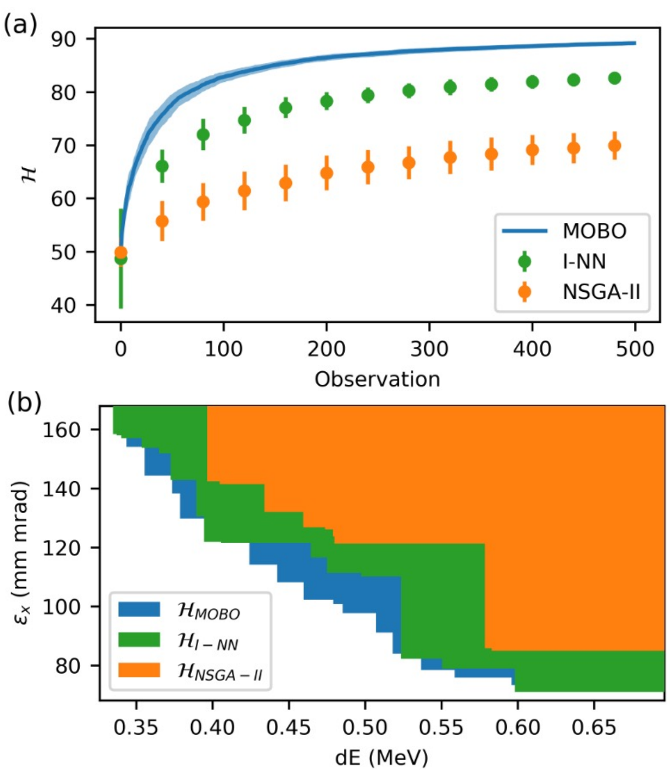
Use Bayesian optimization for **serial online multi-objective optimization**

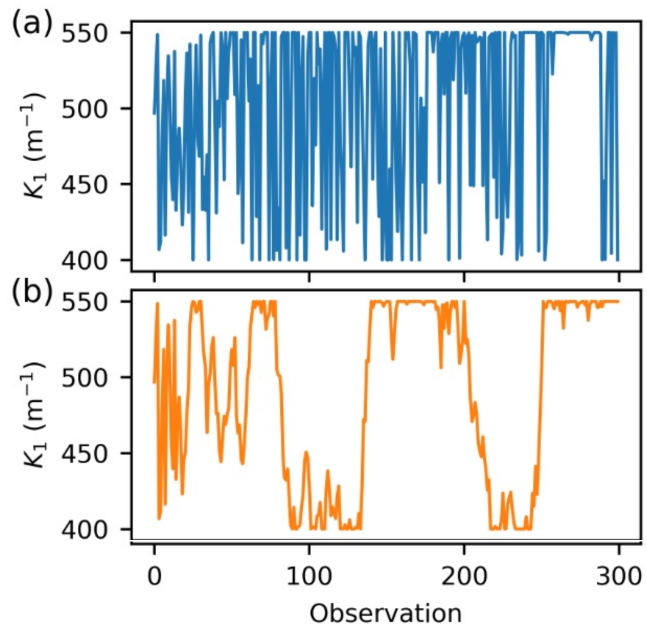
More sample-efficient and fills out front efficiently than other methods

→ Extremely useful for characterization

→ Experimental demos have been done at AWA and LCLS photoinjectors







Can enforce smooth exploration

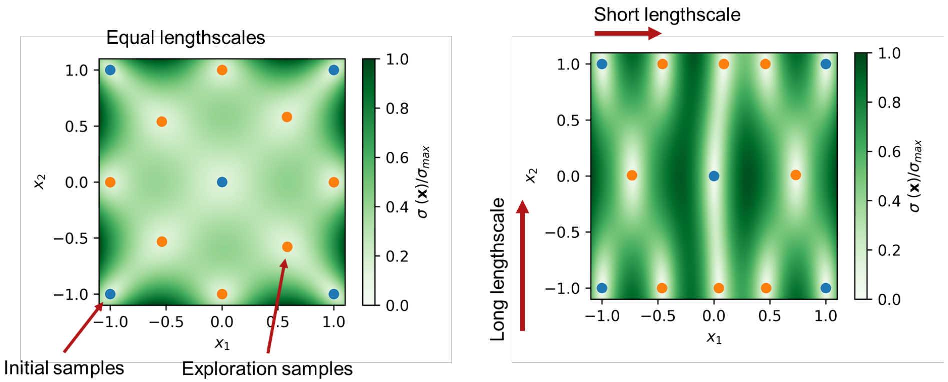
(no wild changes in input settings)

R. Roussel, et al., PRAB (2021)
https://journals.aps.org/prab/pdf/10.1103/PhysRevAccelBeams.24.062801

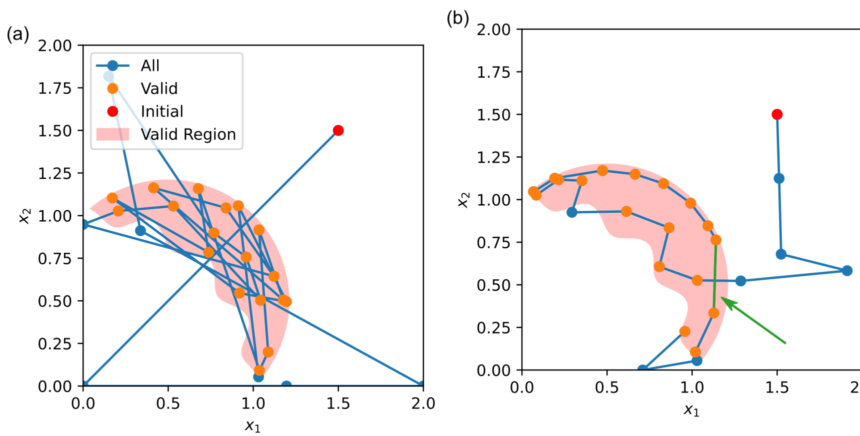
Bayesian Exploration

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

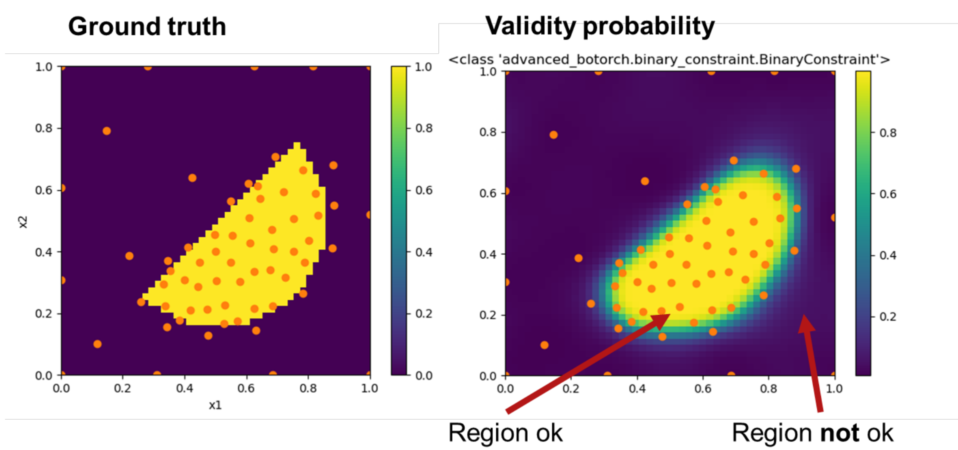
Adaptive sampling



Proximal biasing



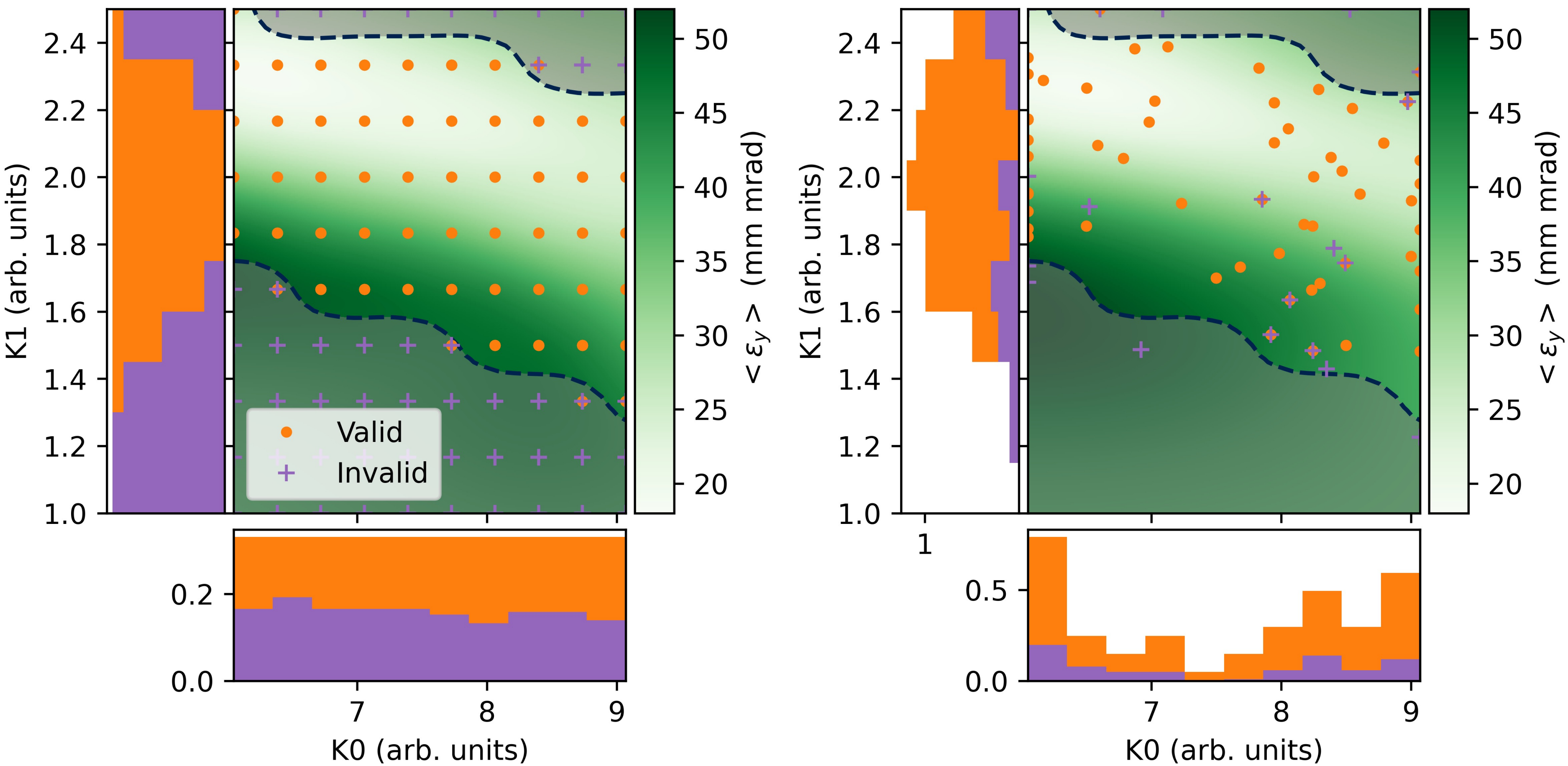
Unknown constraints



R. Roussel

Roussel et. Al. *Nat. Comm.* **2021**

Characterizing Photoinjector Emittance at AWA



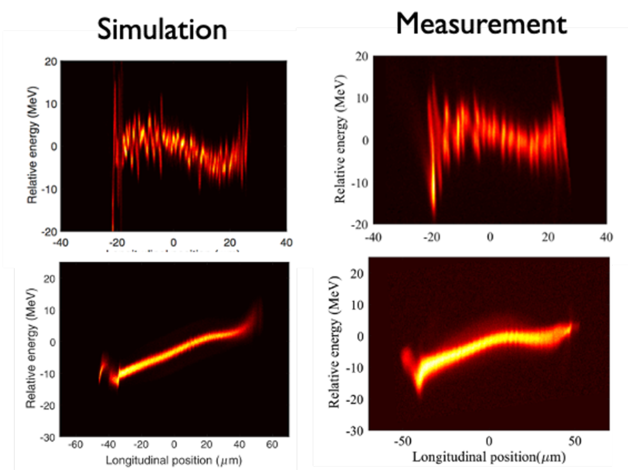
Was also recently used at FACET-II to characterize a 10-dimensional input space wrt emittance and beam matching parameters

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

Fast / Accurate Modeling

Fast Modeling

Accelerator simulations including nonlinear and collective effects are powerful tools…

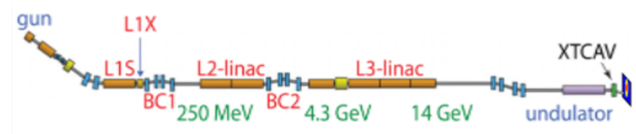


J. Qiang, et al., PRSTAB30, 054402, 2017

"10 hours on thousands of cores at the NERSC"

…but are computationally expensive

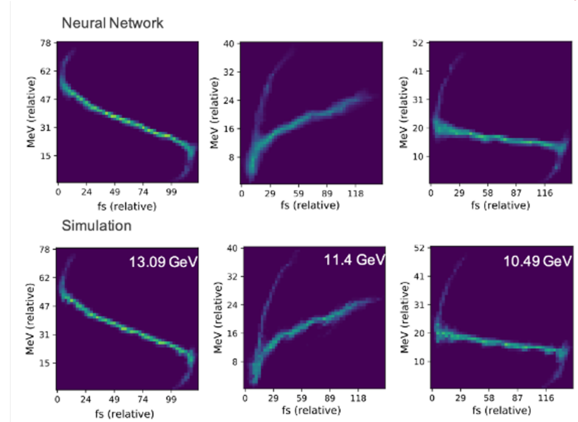
ML models can provide fast approximations for end-to-end simulations



Linac sim in Bmad with collective beam effects

Scan of 6 settings in simulation

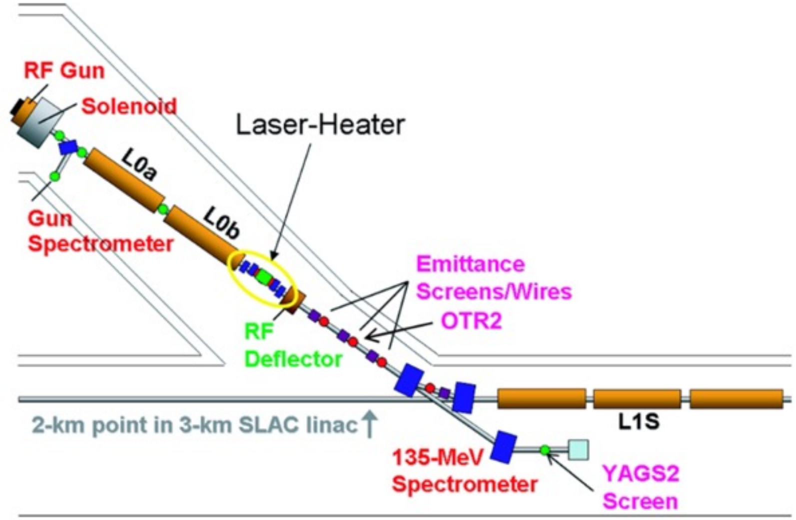
Variable	Min	Max	Nominal	Unit
L1 Phase	-40	-20	-25.1	deg
L2 Phase	-50	0	-41.4	deg
L3 Phase	-10	10	0	deg
L1 Voltage	50	110	100	percent
L2 Voltage	50	110	100	percent
L3 Voltage	50	110	100	percent



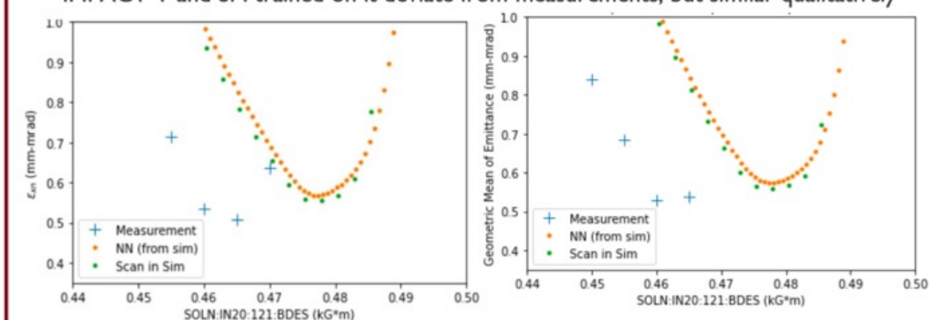
< ms execution speed

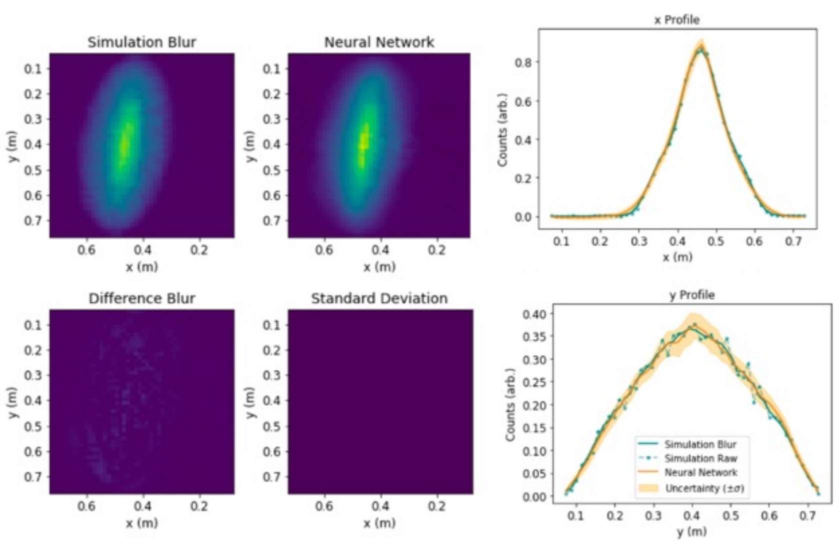
LCLS Injector Surrogate Model

- **Many versions (predict phase space, evolution along z etc); including one with scalar outputs of interest at OTR2**
    - **Inputs:** *laser length + spot size, L0A/B phases, Solenoid, SQ quad, CQ quad, 6matching quads*
    - **Outputs:** *emittances, bunch length, spot sizes, covariances (for Twiss calc), energy*
- Neural network trained on IMPACT-T sims
- Set up to take machine inputs in PV units
- Focused on interpolation to sim vs. exact match to measurements
- Using in tuning algorithm + code testing

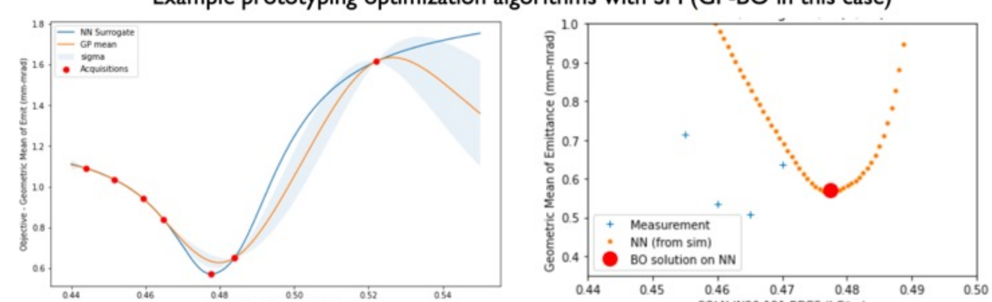


IMPACT-T and SM trained on it deviate from measurements, but similar qualitatively





Example prototyping optimization algorithms with SM (GP-BO in this case)

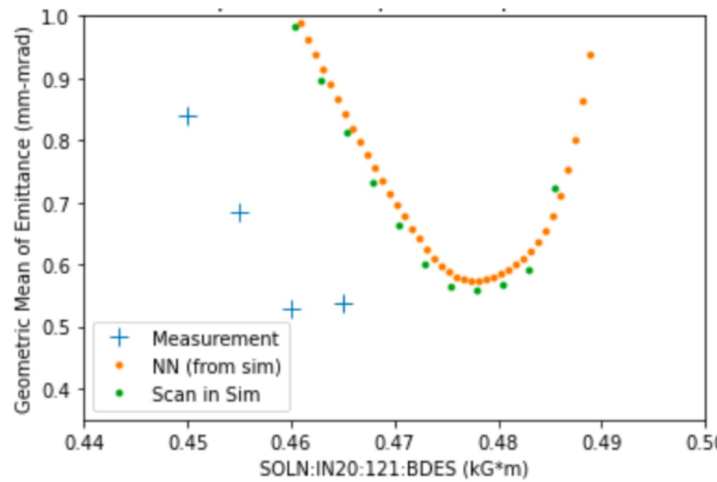


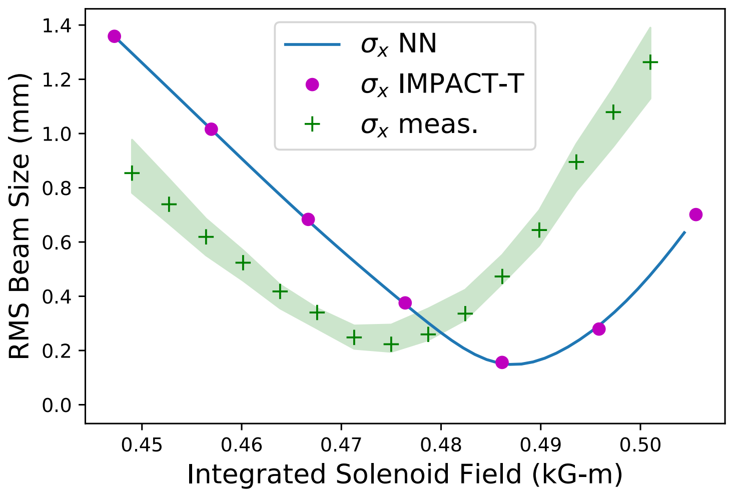
Finding Sources of Systematic Error Between Simulations and Measurement

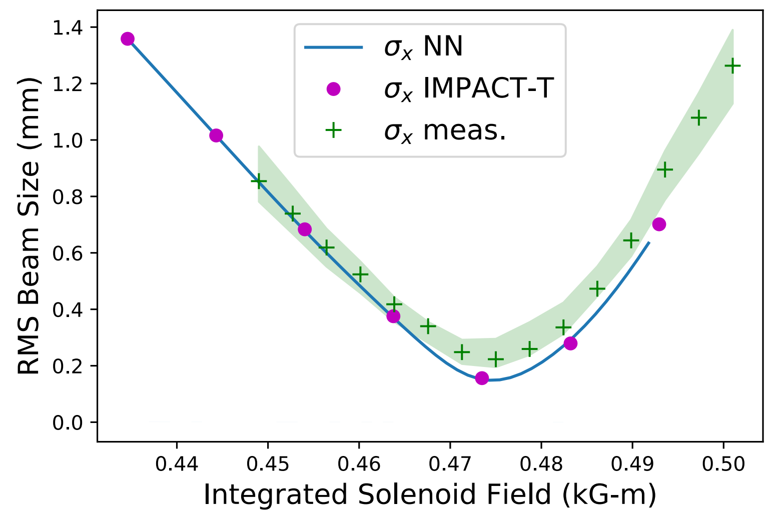
Many non-idealities and miscalibrations are not included in physics simulations → ***identifying these can help correct them and improve meeting of tolerances***

→ *ML model allows fast / automatic exploration of possible error sources*

→ *Can be applied to time-varying changes as well*





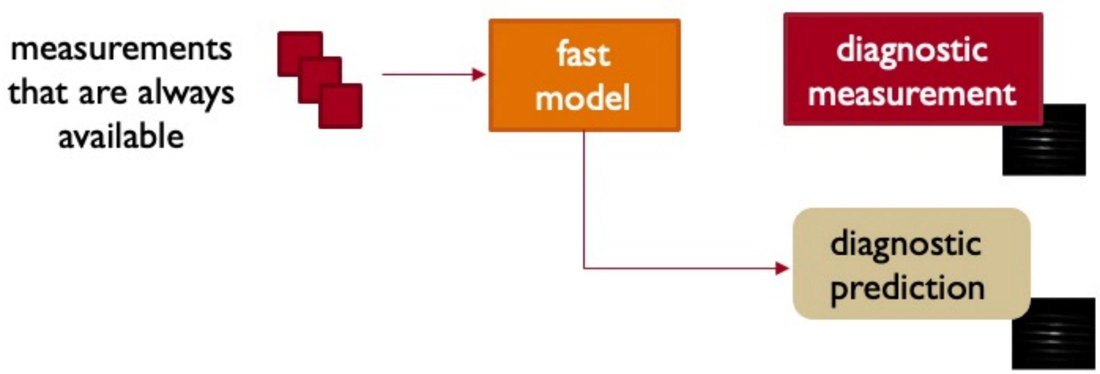


Here: calibration offset in solenoid strength found automatically with neural network model (trained first in simulation, then calibrated to machine)

Virtual Diagnostics

Real diagnostic not always available:

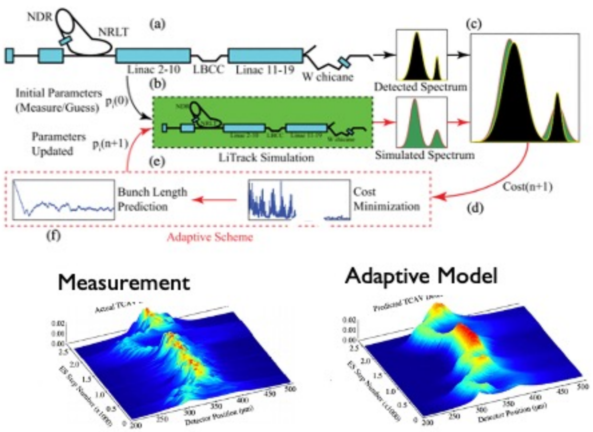
- *destructive, cannot use during user operations*
- *not sensitive in entire operating range*
- *slower update rate than desired*
- *moved to another location*



Can use a physics simulation if fast / accurate enough
→ without this, can use a learned model

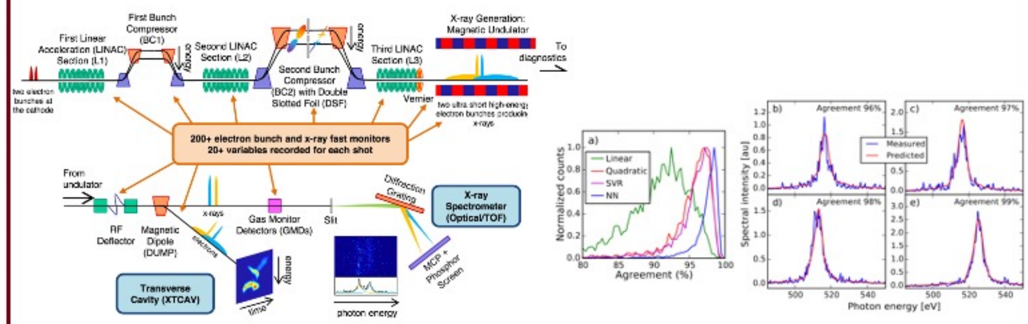
Examples of virtual diagnostics for longitudinal phase space: mix of adaptively calibrated physics models and ML-based prediction…

Adaptively tune a simple physics model



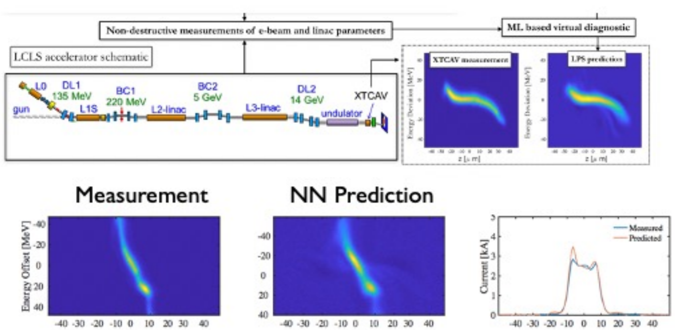
A. Scheinker, S.Gessner, PRAB 18, 102801 (2015)

Fill in shots: use archive data to learn correlation between fast and slow diagnostics



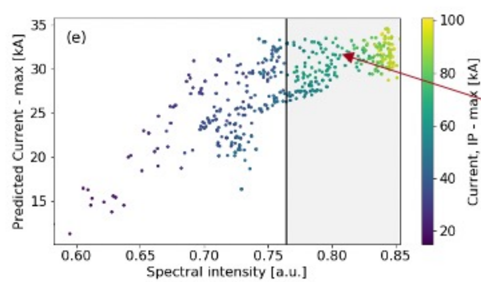
A. Sanchez-Gonzalez, et al., Nature Comms (2017)

Predict with a trained neural network



C. Emma, A. Edelen, et al., PRAB21, 112802 (2018)

Can use spectral information as input to predict beyond typical diagnostic resolution



Shots are beyond the TCAV resolution

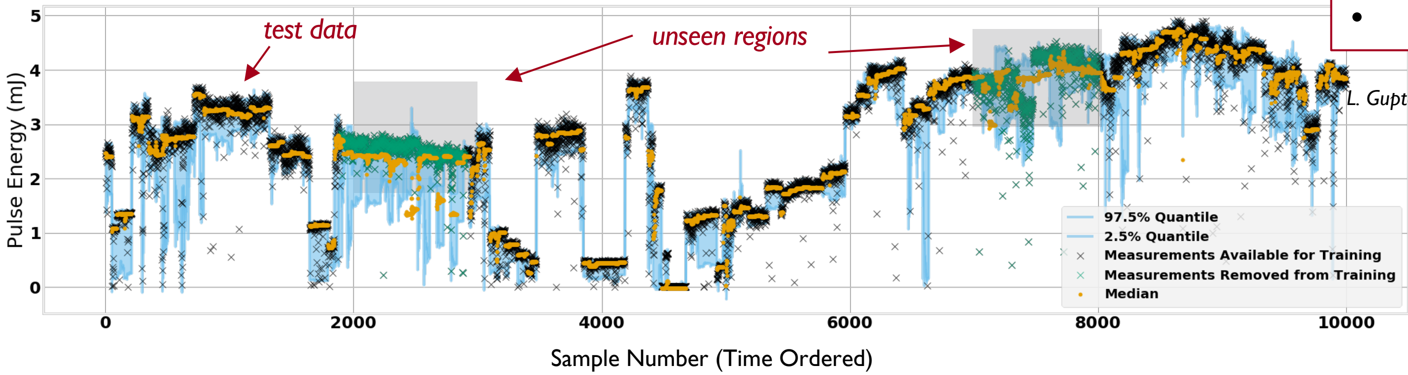
A. Hanuka, et al. 2009.12835 [accepted to Nature Scientific Reports]

ML-based Uncertainty Quantification

Prediction uncertainties can be leveraged in online modeling and control

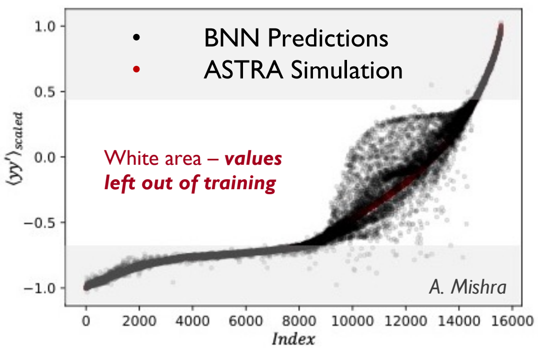
Can also help identify and correct for drifting inputs

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



Neural network with quantile regression predicting FEL pulse energy at LCLS

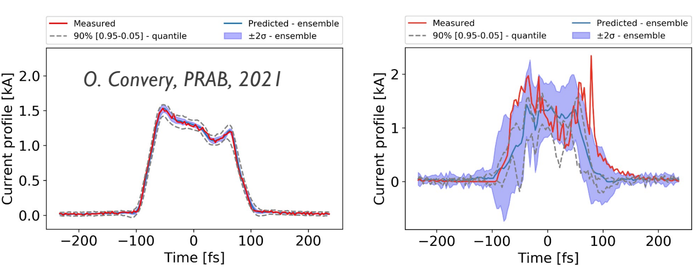
https://github.com/lipiqupta/FEL-UQ/blob/main/notebooks/QR--Interp-2.ipynb



Bayesian neural network predicting scalar parameters for the LCLS-II injector

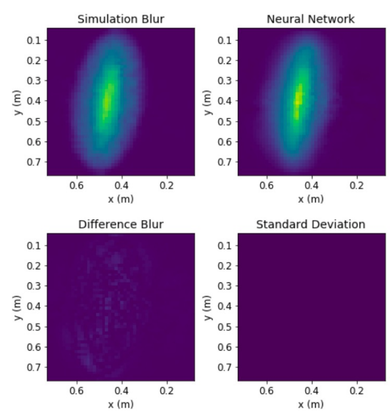
Test shot within trained distribution

Out-of-distribution

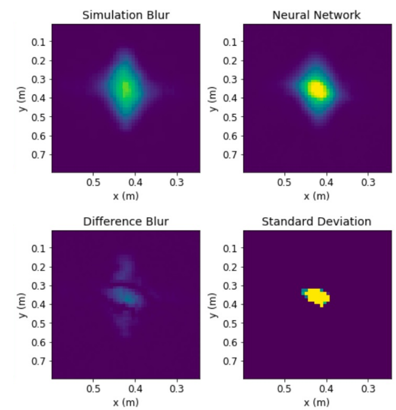


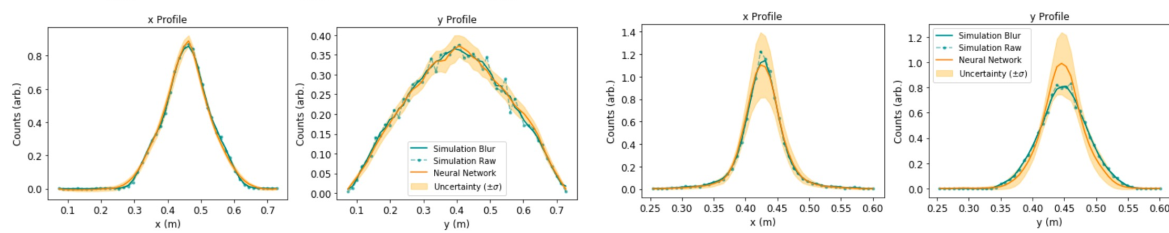
Longitudinal phase space beam profiles

in-distribution



out-of-distribution





LCLS injector transverse distributions on out-of-training distribution shots, neural network ensemble

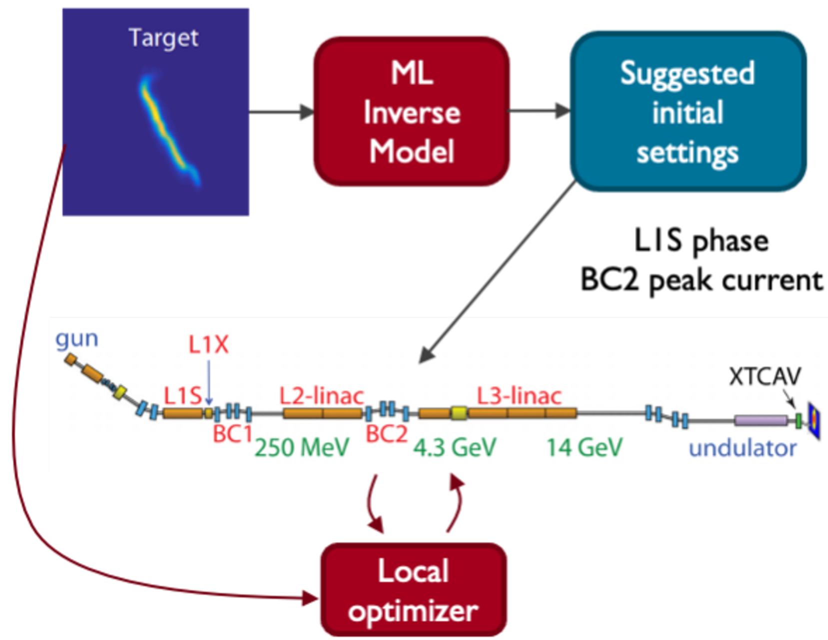
Faster optimization with warm starts from global models

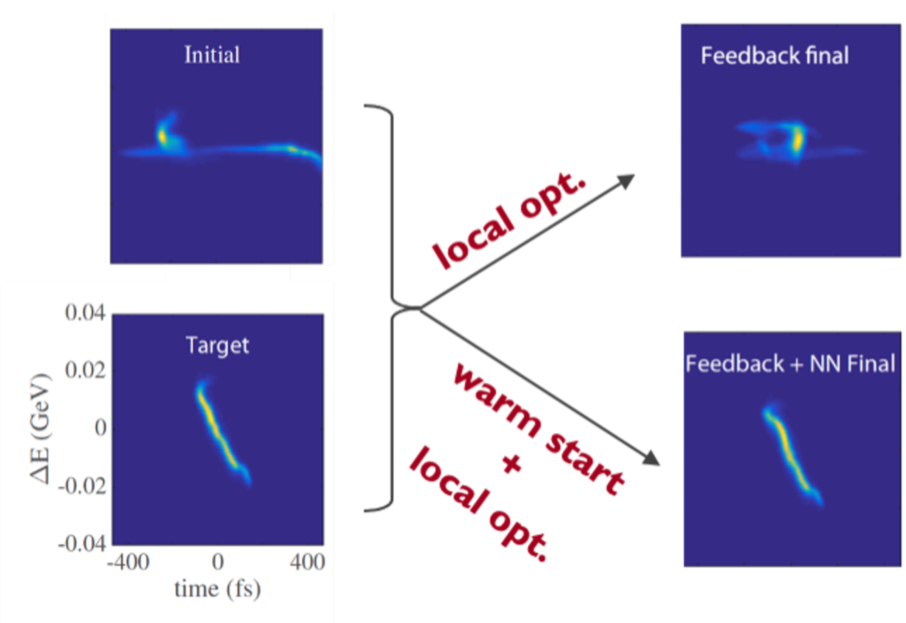
What if we are far away from some target beam parameters and want to switch between configurations quickly?

→ *Use global model to give an initial guess at settings, then refine with local optimization ("warm start")*

Example at LCLS:

- Two settings scanned *(L1S phase, BC2 peak current)*; trained neural network model to map longitudinal phase space to settings
- Compared optimization algorithm with/without warm start

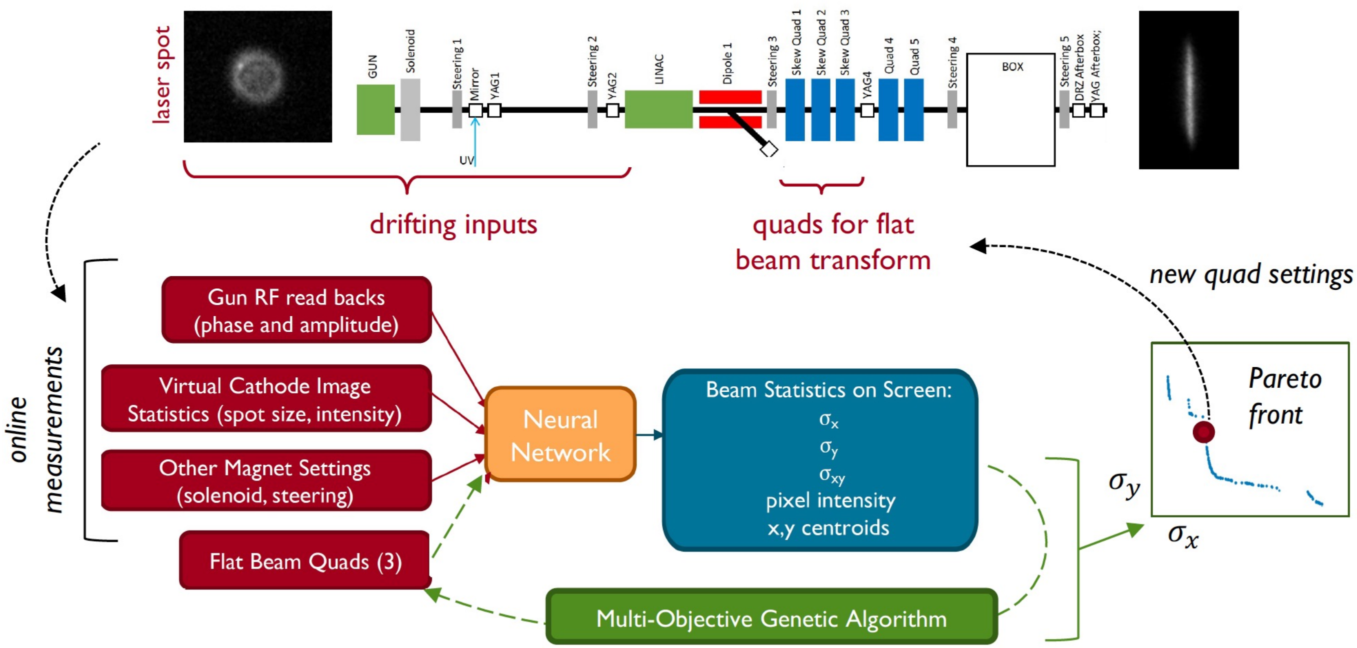




Local optimizer alone was unable to converge → **able to converge after initial settings from neural network**

A. Scheinker, A. Edelen, et al., PRL 121, 044801 (2018)
sim study w/ a THz FEL: A. Edelen, et al., FEL'17

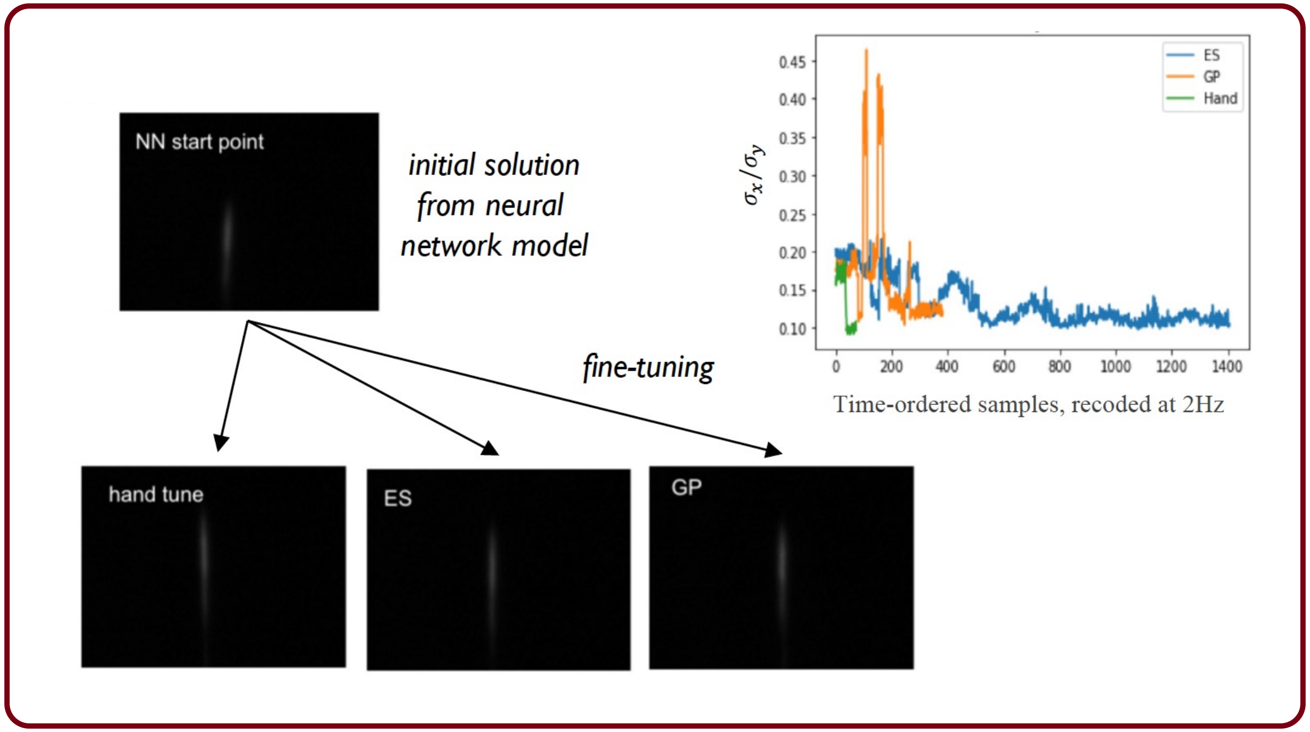
Another way: run optimizer on learned online model



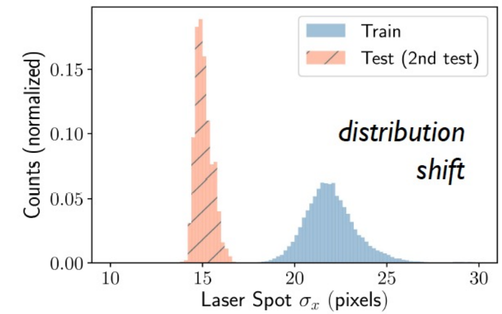
- Round to flat beam transforms are challenging to optimize
- Took measured scan data at Pegasus (UCLA)
- Trained neural network model to predict fits to beam image
- Tested online multi-objective optimization over model (3 quad settings) given present readings of other inputs

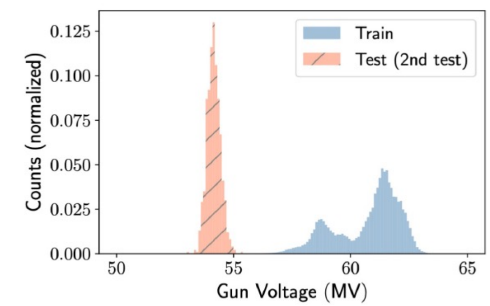
E. Cropp et al., in preparation

Can use neural network to provide first guess at solution, then fine tune with other methods…



can work even under distribution shift in some cases





Hand-tuning in seconds vs. tens of minutes

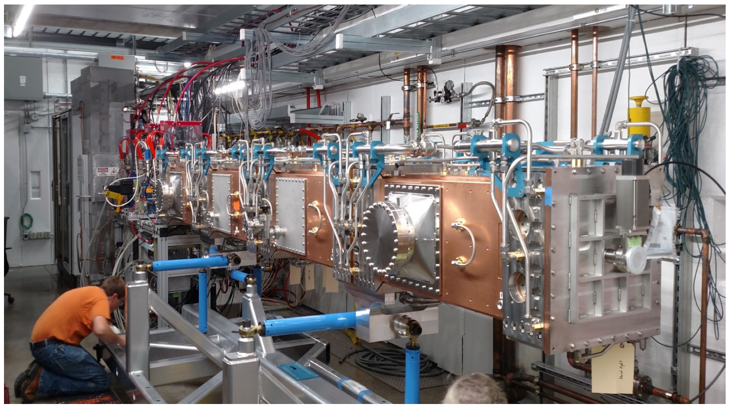
Boost in convergence speed for other algorithms

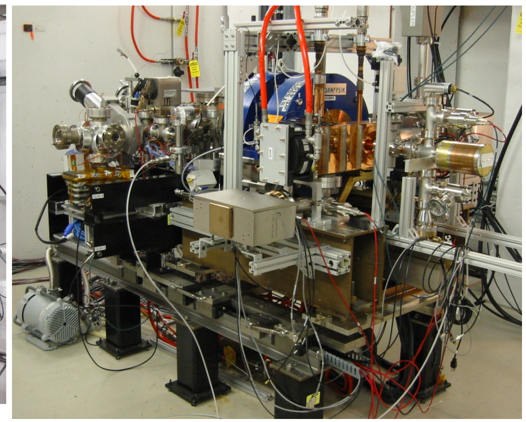
E. Cropp et al., in preparation

RF system control

For RF control, water or cryogenic based cooling systems need to be controlled too

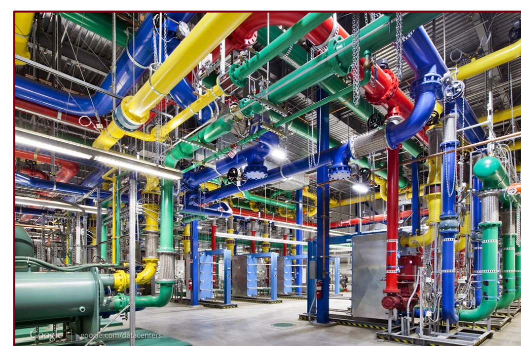
→ Fluctuations can impact RF resonant frequency (compensated with increased forward power)
→ RF is a major driver of machine costs (both in designing RF overhead and in operational costs)





Transport delays, variable heat load, complex dynamics

Transport delays, variable heat load
Efficient servers were not enough
→ needed better control of cooling system



https://googleblog.blogspot.com

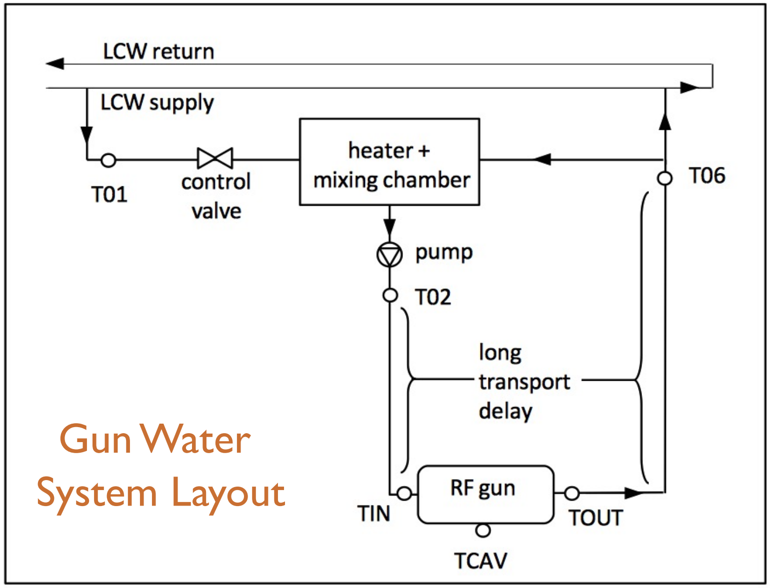
Example from FAST RF gun

Resonant frequency controlled via temperature

- *Long transport delays and thermal responses*
- *Two controllable variables: heater power + flow valve aperture*

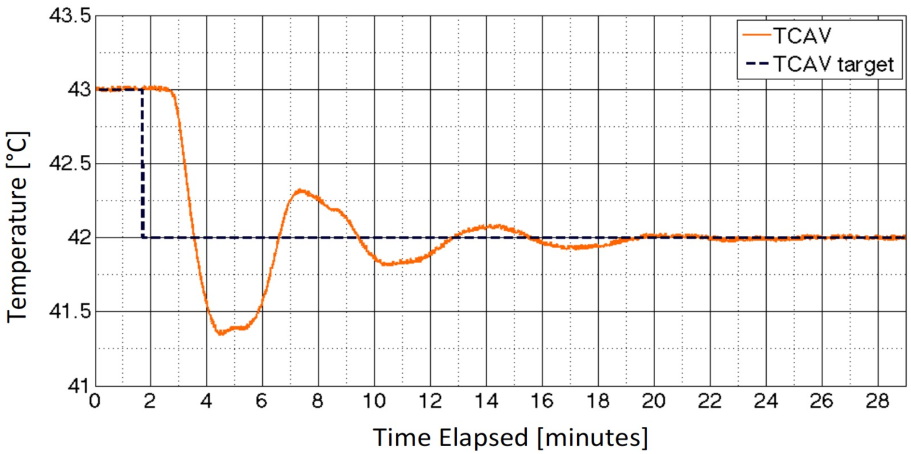
Applied model predictive control with a neural network model trained on measured data

~ 5x faster settling time + no large overshoot **(reduce RF costs)**

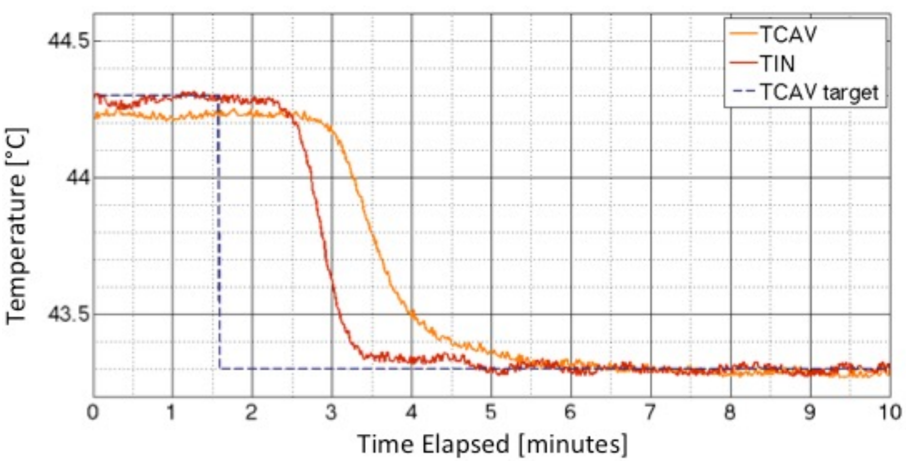


Gun Water System Layout

Existing Feedforward/PID Controller



Model Predictive Controller



Note that the oscillations are largely due to the transport delays and water recirculation, rather than PID gains

Similar techniques can be applied to cryogenic systems

Edelen, IPAC'15 ; Edelen, TNS, 2016

Classifying SRF Trips

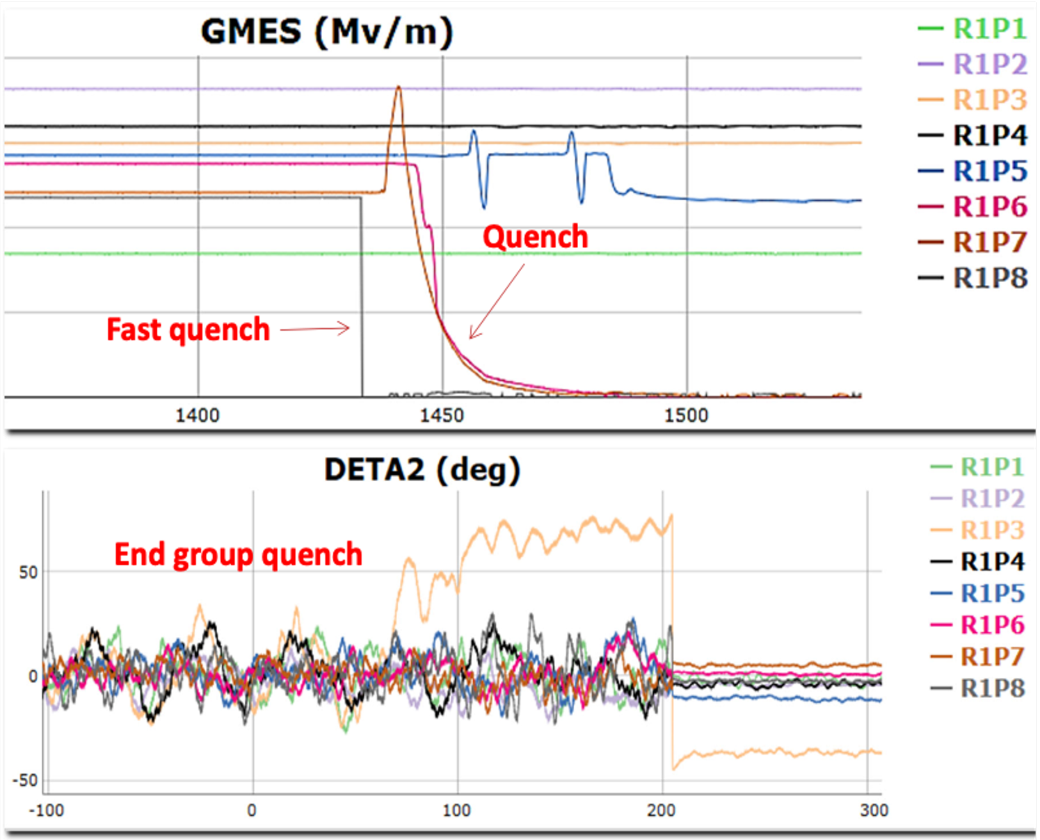
Cavities can trip in a variety of ways

(fast quench, thermal quench, end group quench, microphonics)

Experts identify type of trip from RF waveform data

Instead, use automatic classification:

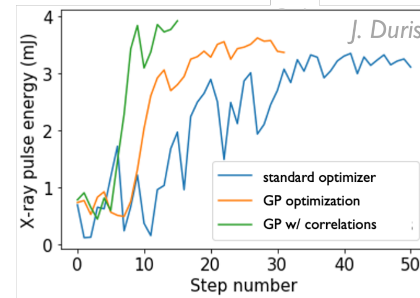
- *Enables more systematic study of trips and effectiveness of recovery strategies*
- *Quickly informs a proper response in the control room*



A. Solopova, et al., IPAC'19

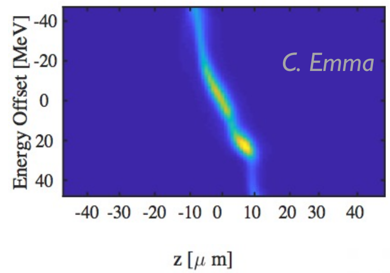
Several major areas for ML to play a role

automated control
+ optimization

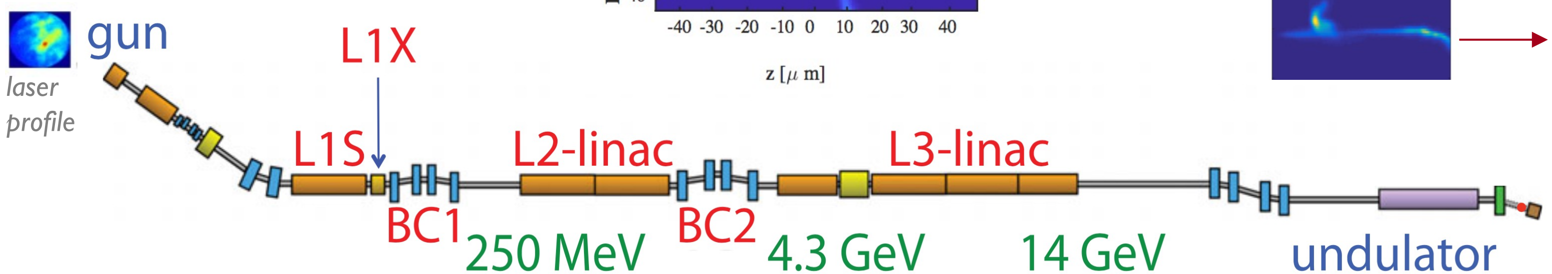


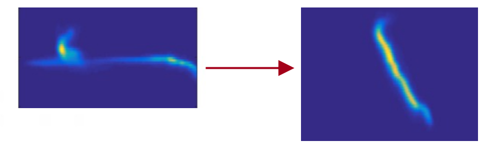
diagnostics

(reconstruct / analyze beam)

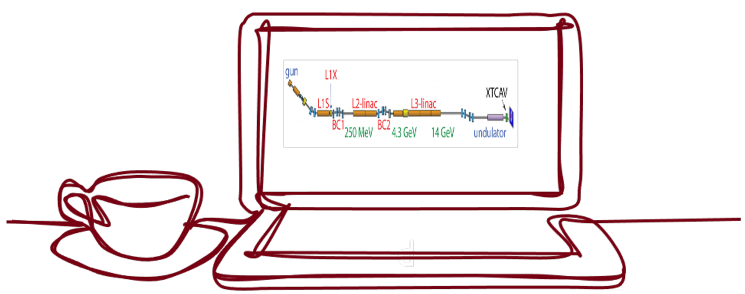


anomaly detection
failure prediction





incorporate
physics
information



extract unexpected
relationships

(feed into control / design)

digital twins + online modeling

(planning, model-based control, finding differences between sim/machine)

+ need uncertainty
quantification for all

Integration of AI/ML and Online Accelerator Modeling / Control

- Many proof-of-principle results for AI/ML modeling and control of accelerators → *usually in limited ranges of operating conditions or addressing isolated problems (e.g. only optimization, only modeling)*

- **Now need to address integration into dedicated operation:**
  - Need a comprehensive ***facility-agnostic*** software/hardware ecosystem that can couple HPC, online simulation, and AI/ML
  - Need to assess/address robustness challenges of dedicated operation and coupling different types of AI/ML tasks together
  - Coupling of AI/ML, traditional algorithms, and human-in-the-loop operations *(provide useful/actionable information rather than add to information overload)*

→ Prototyping a comprehensive AI/ML ecosystem for online modeling/control at smaller-scale test facilities would (1) provide substantial benefit in bringing this technology to maturity and (2) provide a roadmap for scaling it up to larger facilities