Surrogate Modeling for Charged Particle Accelerator Beam Dynamics

Auralee Edelen, Nicole Neveu, Andreas Adelmann, Yannick Huber

ICAP 2018, Key West, Florida

SLAC NATIONAL ACCELERATOR LABORATORY
Argonne NATIONAL LABORATORY
PAUL SCHERRER INSTITUT PSI
U.S. DEPARTMENT OF ENERGY Office of Science

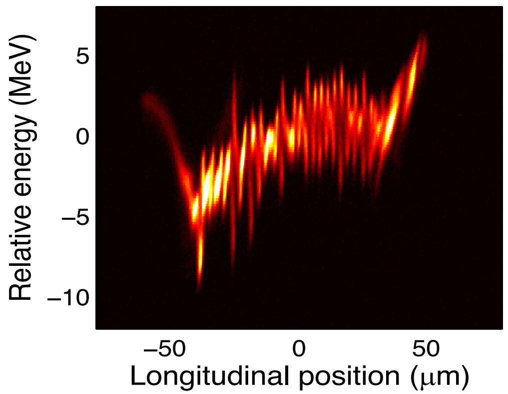
Accelerator simulations that include nonlinear / collective effects are powerful tools, but they can be very slow to execute

Impedes start-to-end optimization

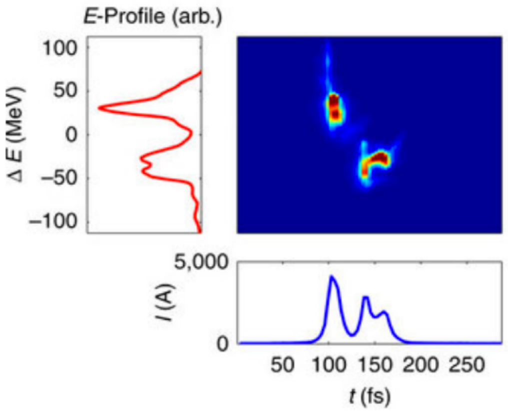
Impedes use as an online model / virtual diagnostic

Impedes use in control / control development

Often takes much effort to replicate real machine behavior



D. Ratner, et al., PRSTAB18, 030704 (2015)



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

→ *especially for complicated setups and acceleration schemes (e.g. plasma-based)*

Accelerator simulations that include nonlinear / collective effects are powerful tools, but they can be very slow to execute

Impedes start-to-end optimization

Impedes use as an online model / virtual diagnostic

Impedes use in control / control development

Often takes much effort to replicate real machine behavior

One approach: faster modeling codes

Simpler models (tradeoff with accuracy)

analytic calculations — *e. g. J. Galambos, et al., HPPA5, 2007*

Parallelization and GPU-acceleration of existing codes

HPSim/PARMILA — *X. Pang, PAC13, MOPMA13*

elegant — *I.V. Pogorelov, et al., IPAC15, MOPMA035*

Improvements to modeling algorithms

Lorentz-boosted frame — *J.-L. Vay, Phys. Rev. Lett. 98 (2007) 130405*

Accelerator simulations that include nonlinear / collective effects are powerful tools, but they can be very slow to execute

Impedes start-to-end optimization

Impedes use as an online model / virtual diagnostic

Impedes use in control / control development

Often takes much effort to replicate real machine behavior

One approach: faster modeling codes

Simpler models (tradeoff with accuracy)

analytic calculations — *e. g. J. Galambos, et al., HPPA5, 2007*

Parallelization and GPU-acceleration of existing codes

HPSim/PARMILA — *X. Pang, PAC13, MOPMA13*

elegant — *I.V. Pogorelov, et al., IPAC15, MOPMA035*

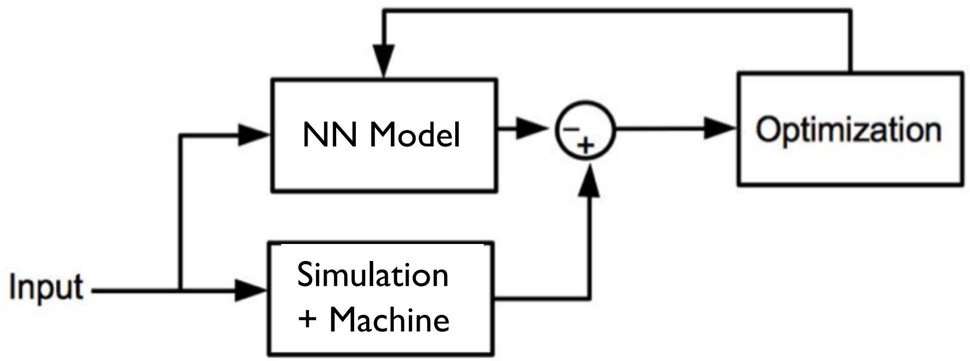
Improvements to modeling algorithms

Lorentz-boosted frame — *J.-L. Vay, Phys. Rev. Lett. 98 (2007) 130405*

Another approach: machine learning model

Once trained, neural networks can execute quickly

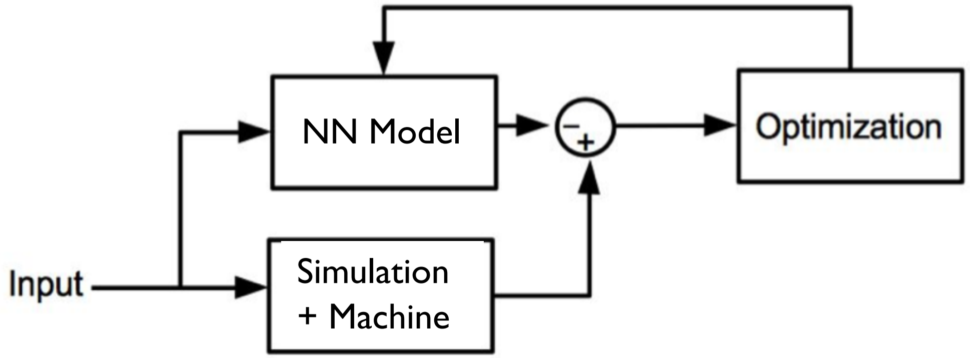
Train on data from slow, high-fidelity simulations
+
Train on measured data



Another approach: machine learning model

Once trained, neural networks can execute quickly

Train on data from slow, high-fidelity simulations
+
Train on measured data

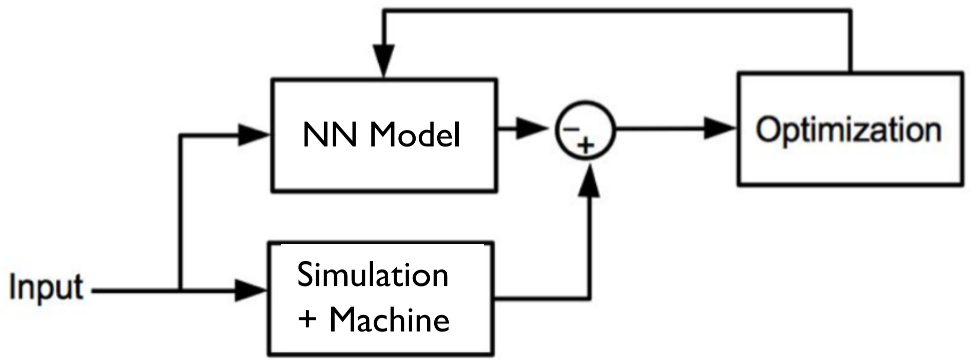


- Train from high-fidelity simulation results → *orders of magnitude speedup*
- Update with measured data → *bridge gap between sims and real machine*
- Use as a virtual diagnostic → *predict what a diagnostic would show when it is unavailable*
- Use to facilitate control → *model-based control, use with online optimization, use as a platform for controls development*
- Can use for design studies → *new setups on existing machines + designing downstream components*

Another approach: machine learning model

Once trained, neural networks can execute quickly

Train on data from slow, high-fidelity simulations
+
Train on measured data

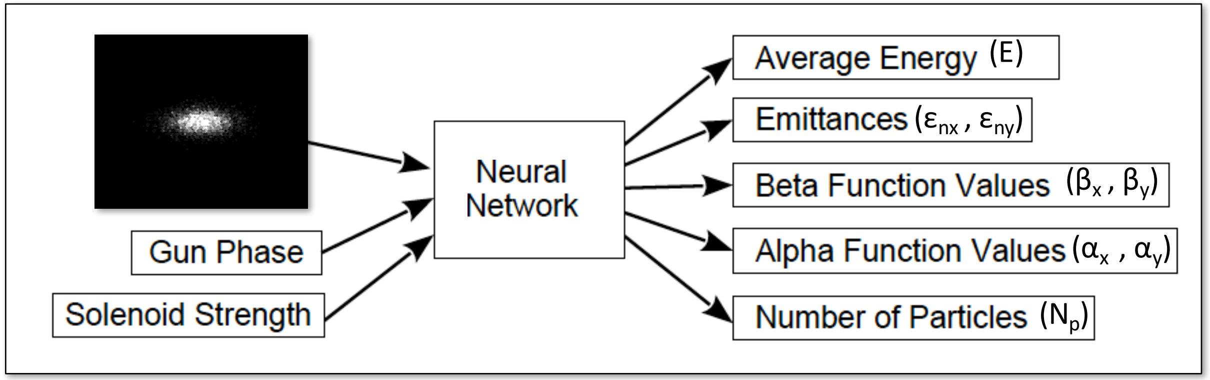


- Train from high-fidelity simulation results → *orders of magnitude speedup*
- Update with measured data → *bridge gap between sims and real machine*
- Use as a virtual diagnostic → *predict what a diagnostic would show when it is unavailable*
- Use to facilitate control → *model-based control, use with online optimization, use as a platform for controls development*
- Can use for design studies → *new setups on existing machines + designing downstream components*

An initial study at Fermilab:

A. L. Edelen, J.P. Edelen. D. Edstrom, et al. NAPAC16, TUPOA51

PARMELA with 2-D space charge routine: ~ 20 mins
Neural network model: ~ a millisecond



All mean absolute errors between 0.9% and 3.1% of the parameter ranges

But can we really trust these models in optimization, and what are the limitations?

But can we really trust these models in optimization, and what are the limitations?

Decided to investigate this with the Argonne Wakefield Accelerator

-- extensive simulation work for the AWA already done by N. Neveu

-- computing resources to do GA study in simulation

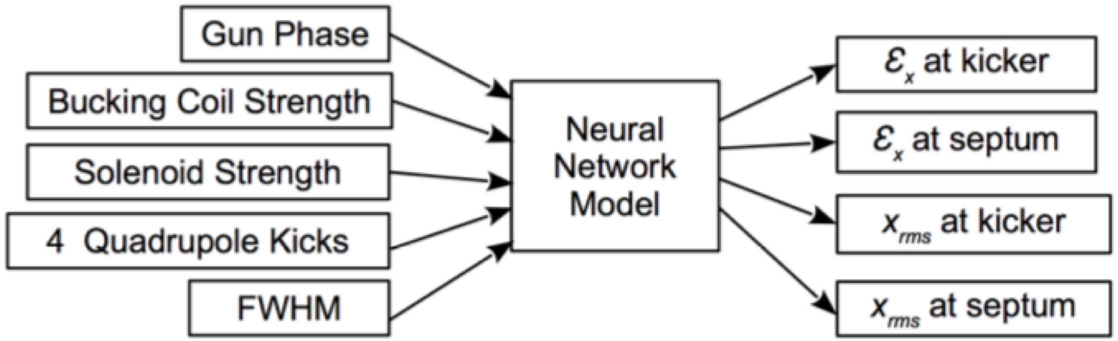
-- OPAL head developer A. Adelmann already collaborating with AWA + past work on polynomial chaos expansion (PCe) surrogates for a cyclotron

(https://arxiv.org/pdf/1509.08130.pdf)

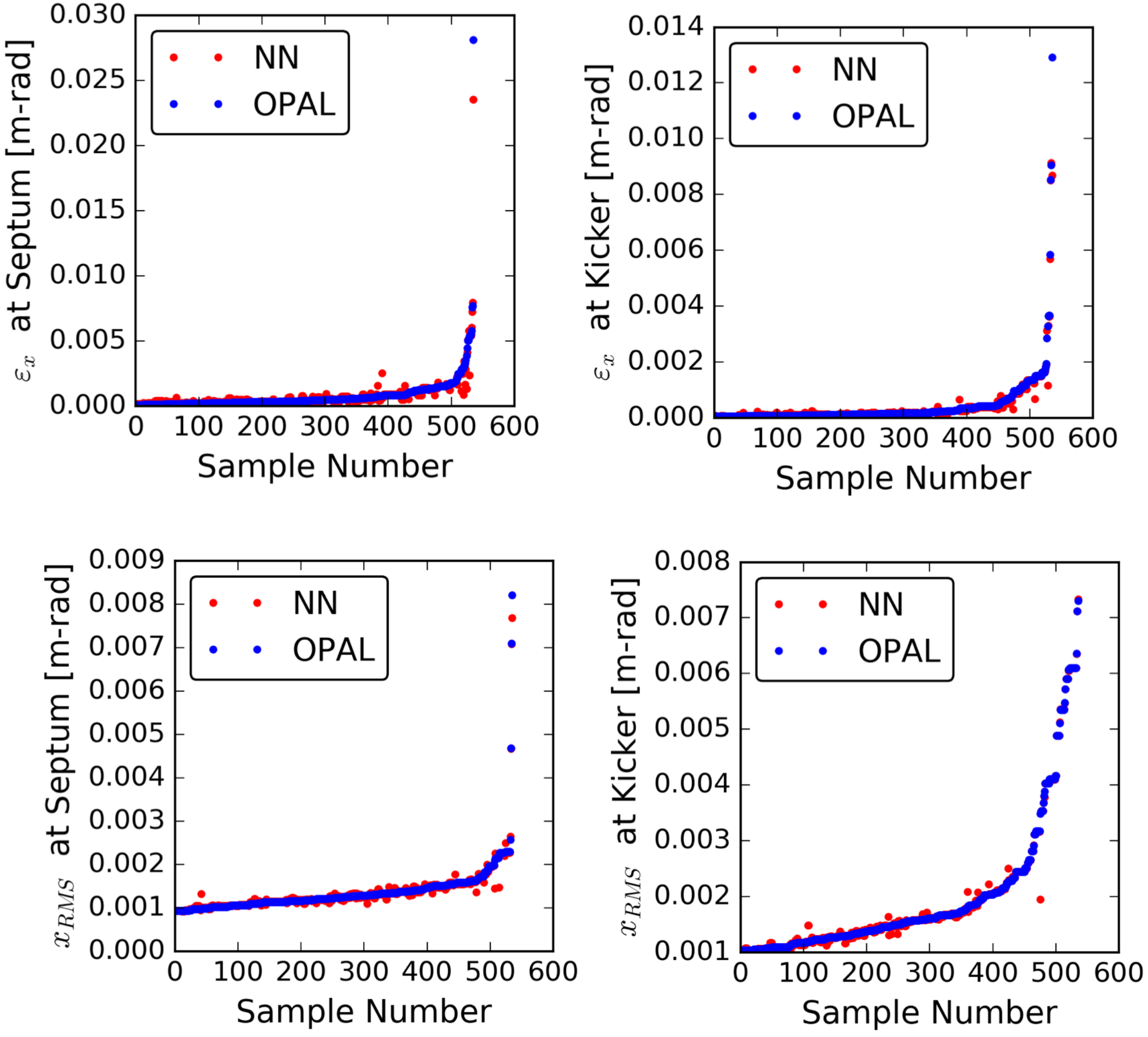
Surrogate Modeling for the AWA: Small Initial Study

SLAC

Trained on ~30k iterations of output from optimization of injector / beamline in OPAL



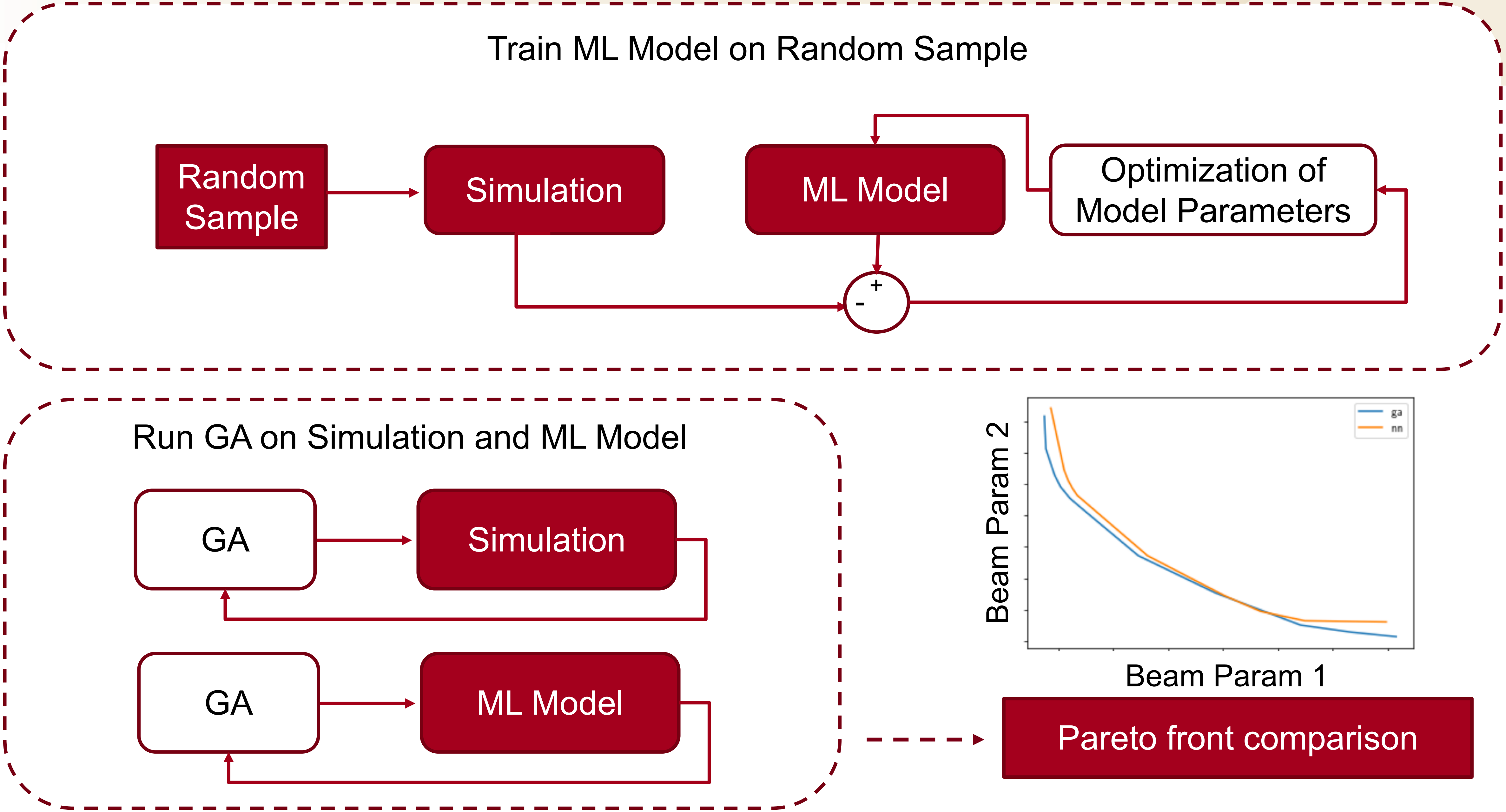
Variable	***Unit***	***Range***
Bunch FWHM	[ps]	0.05 – 25.1
ϕ	[°]	-39.1 – 6.7
I_{bs}	[A]	72 – 638
I_s	[A]	173 – 266
Q1	[m^{-1}]	-10.0 – 12.0
Q2	[m^{-1}]	-12.5 – 13.7
Q3	[m^{-1}]	-10.4 – 13.1
Q4	[m^{-1}]	-12.2 – 7.9



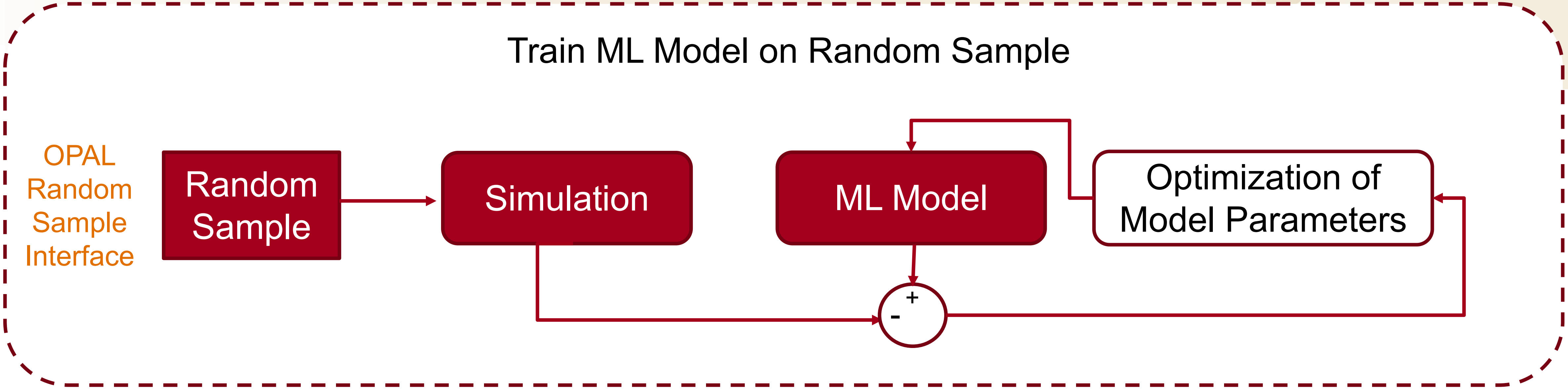
Follow-up study:

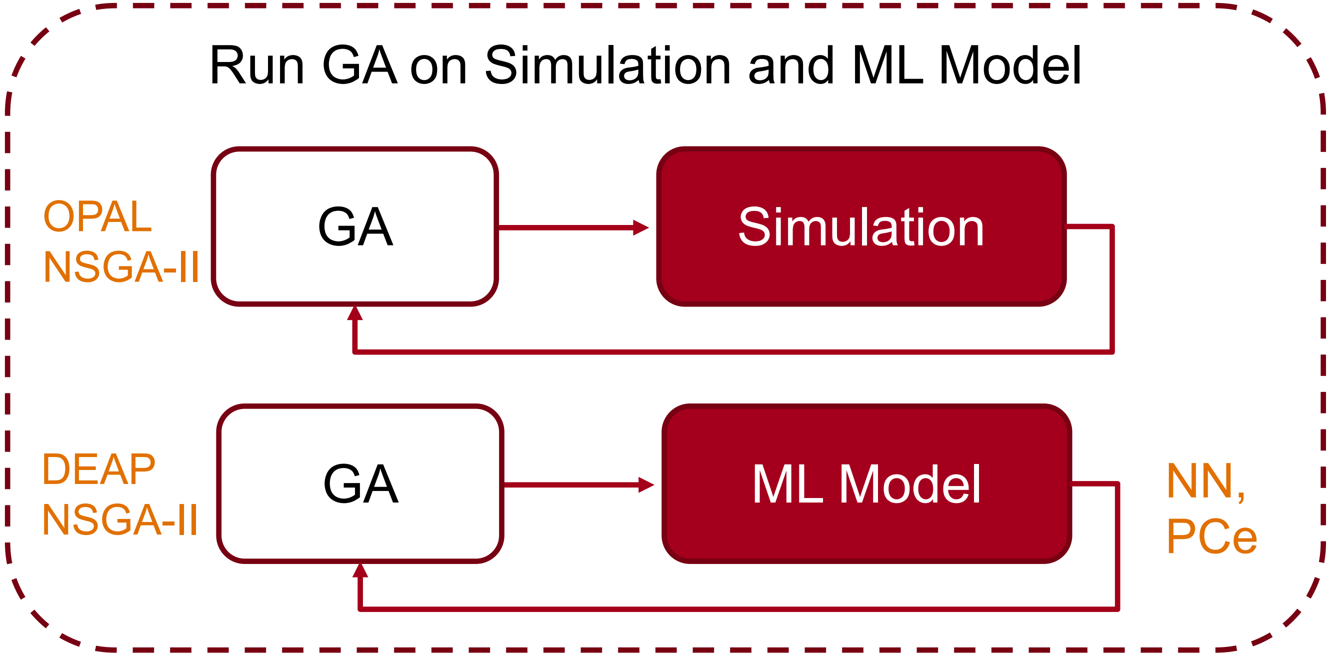
focus on pareto fronts

Workflow for Assessing Comparison with GA

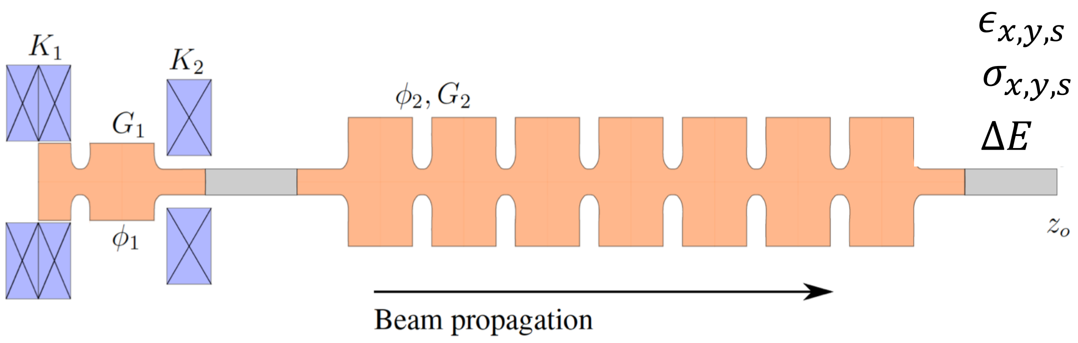


Workflow for Assessing Comparison with GA

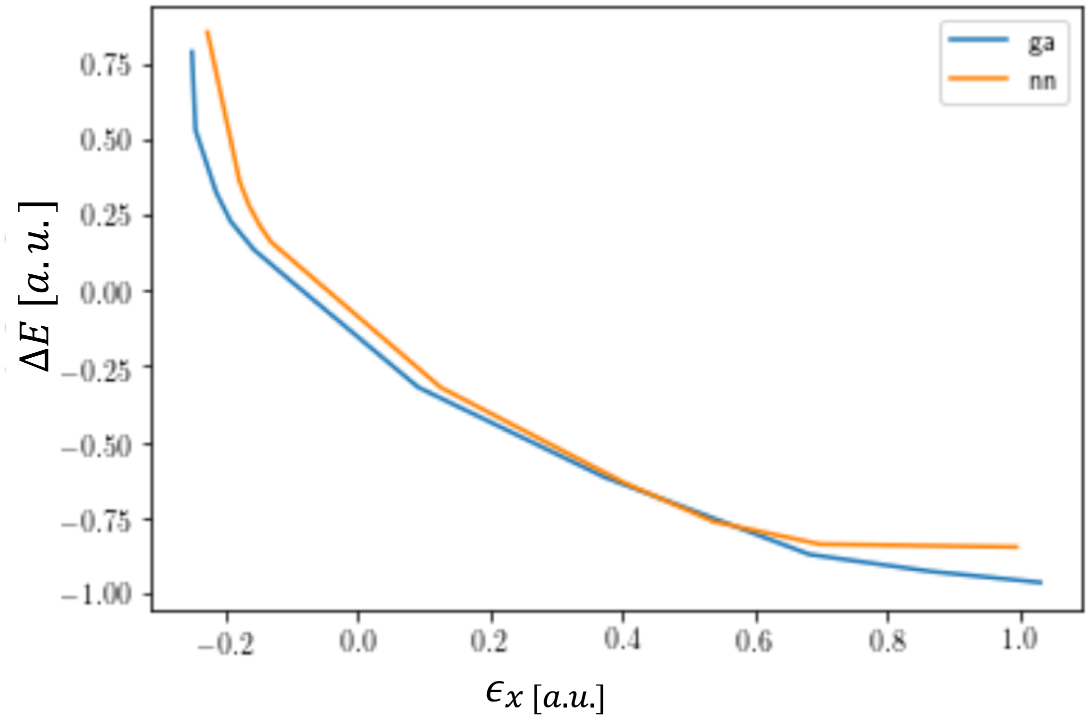


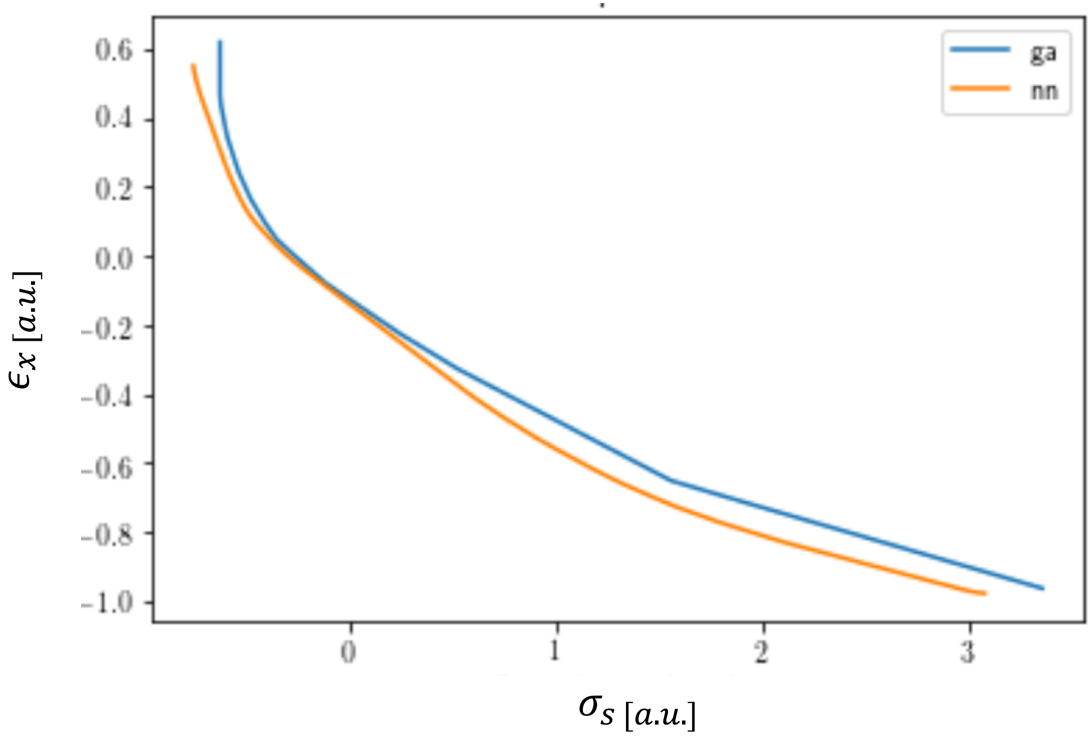


Adjust six design variables over known good range
Evaluate seven beam parameters



Comparison of Pareto Fronts





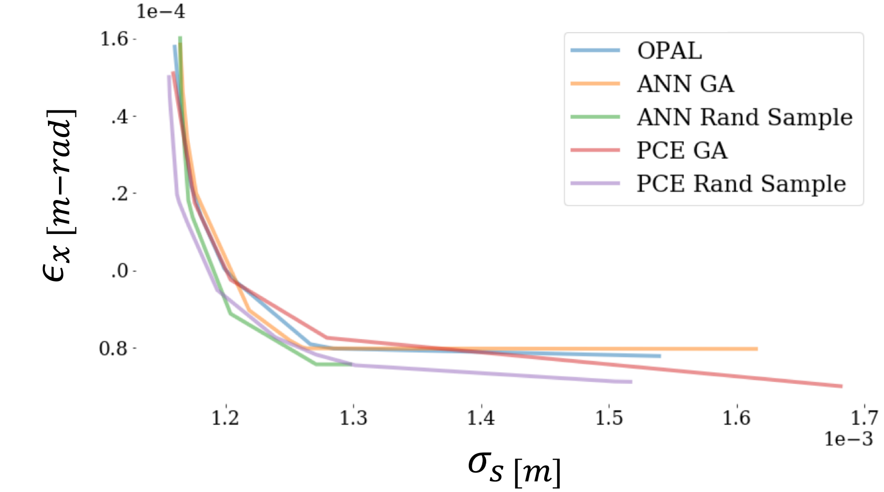
OPAL GA: ~42,510 core hours at ALCF
~16.2 hours
130,865 simulation evaluations

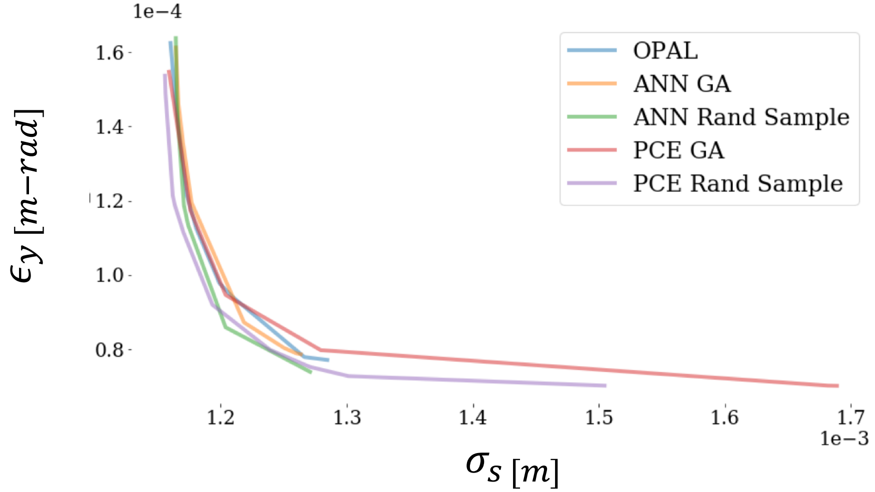
(for each new optimization)

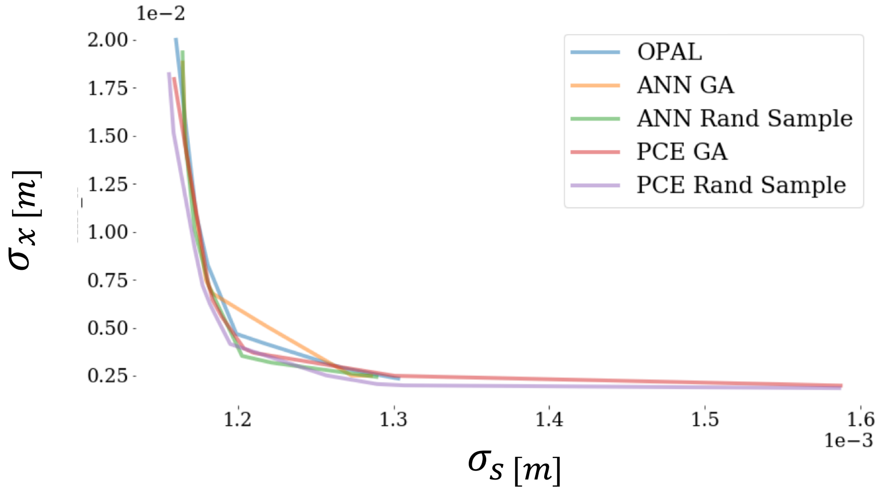
NN Surrogate: ~2 minutes on laptop

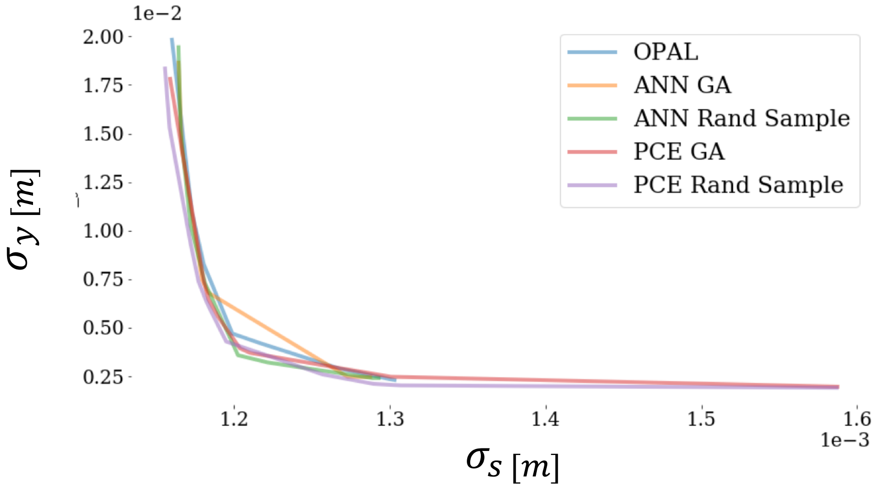
(hidden cost: ~70k initial simulations for training, but in principle only need to do once, and might be able to use smaller data set)

Comparison of Pareto Fronts





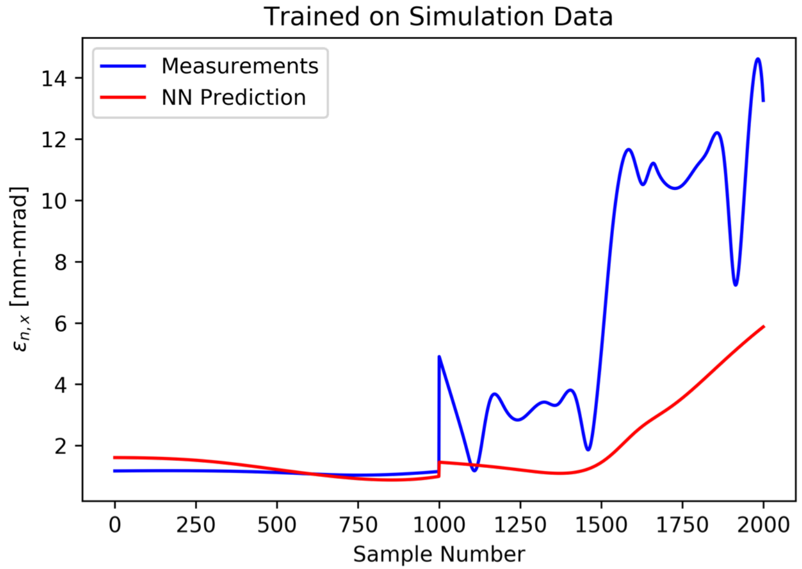


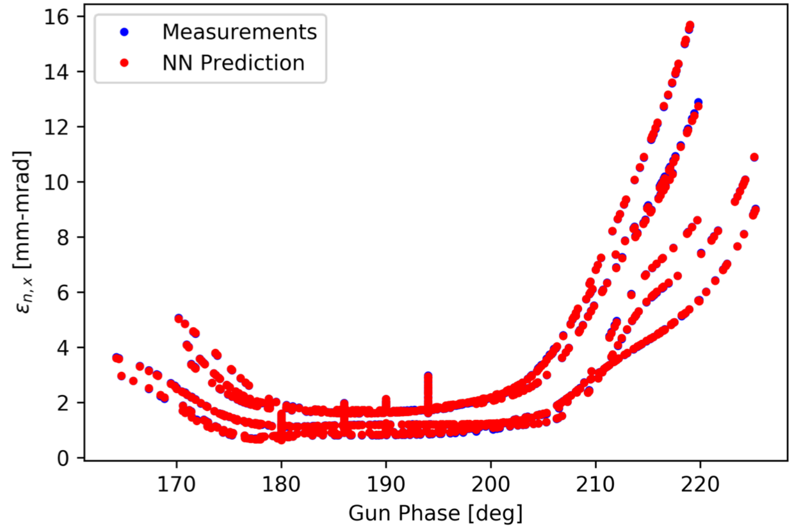


Training on imperfect simulations: ML model only as good as the simulation relative to the real machine

Poor agreement between simulation and measured data for some input/output relationships, but good for others

→ *can we update the NN model with measured data without disrupting the good predictions?*

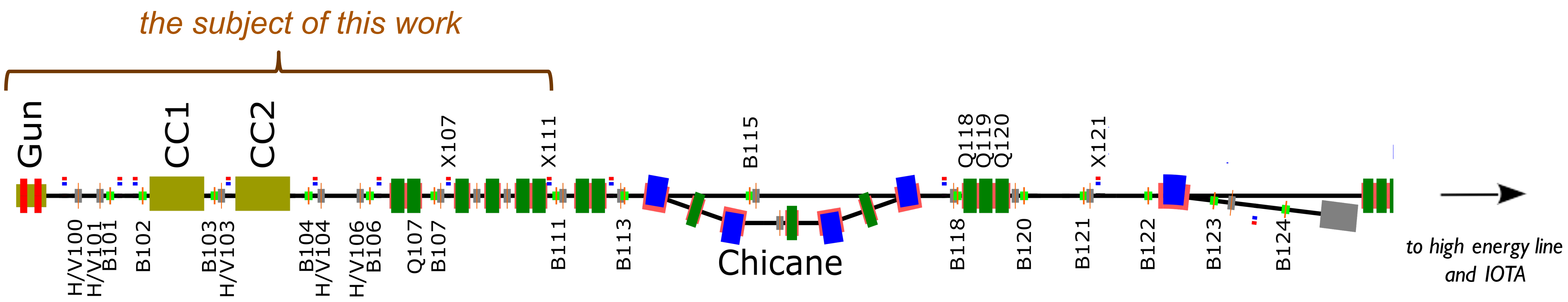


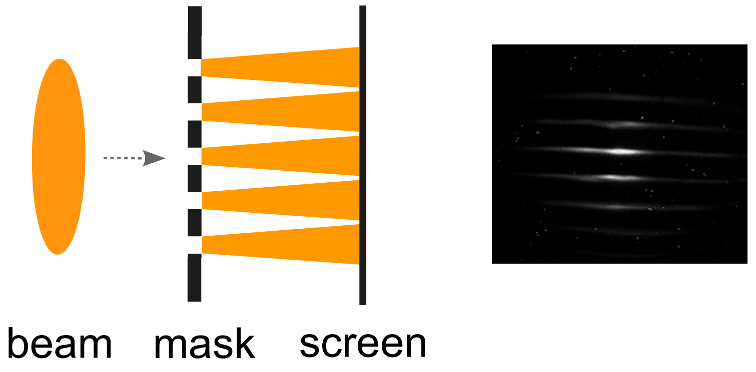


Work with J.P Edelen, D. Edstrom, J. Ruan

A. L. Edelen, et al. IPAC18, WEPAF040

Example from Fermilab's FAST Facility





→ fit to obtain subset of phase space parameters

$\langle y^2 \rangle$

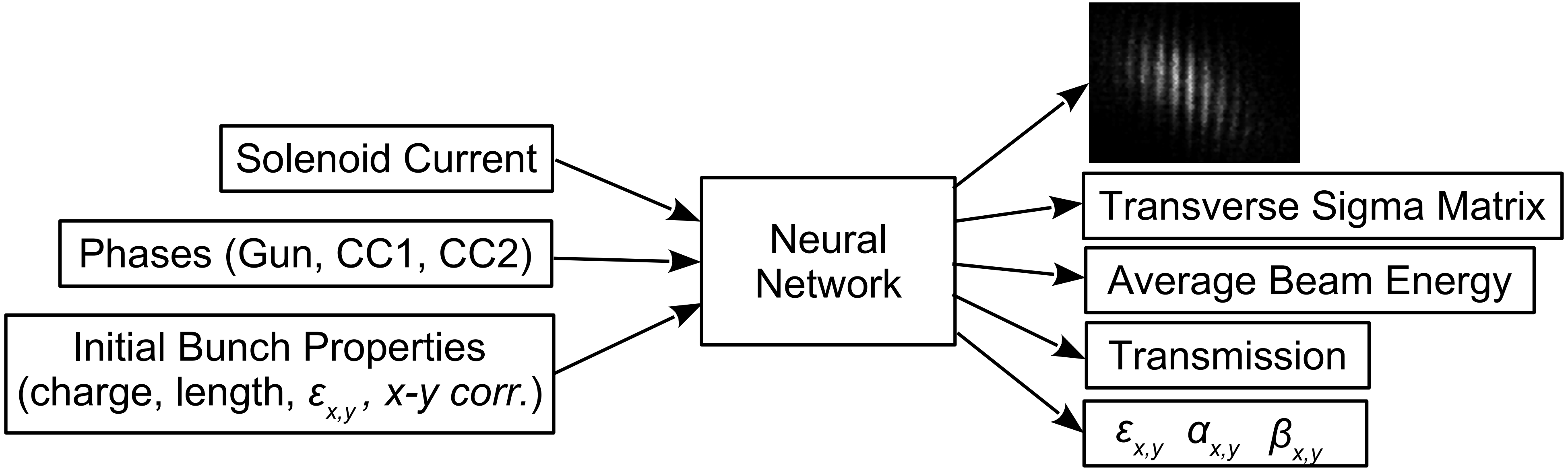
$\langle yy' \rangle$

$\langle y'^2 \rangle$

Multi-slit emittance measurement after the second capture cavity (X107 to X111) takes 10-15 seconds

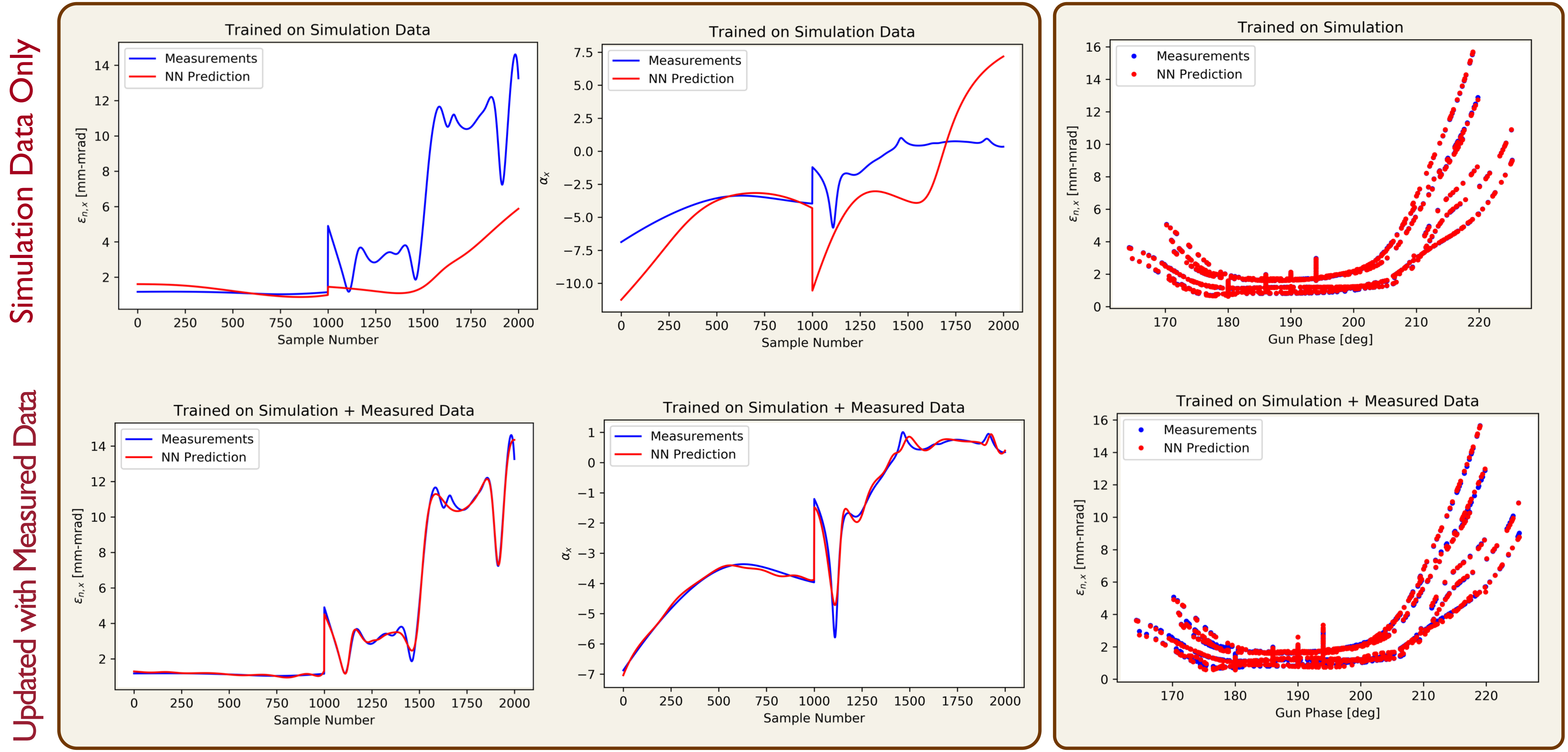
→ can we get an online prediction of what this intercepting diagnostic would show?

Example from FAST



Solenoid Scan

Phase Scan

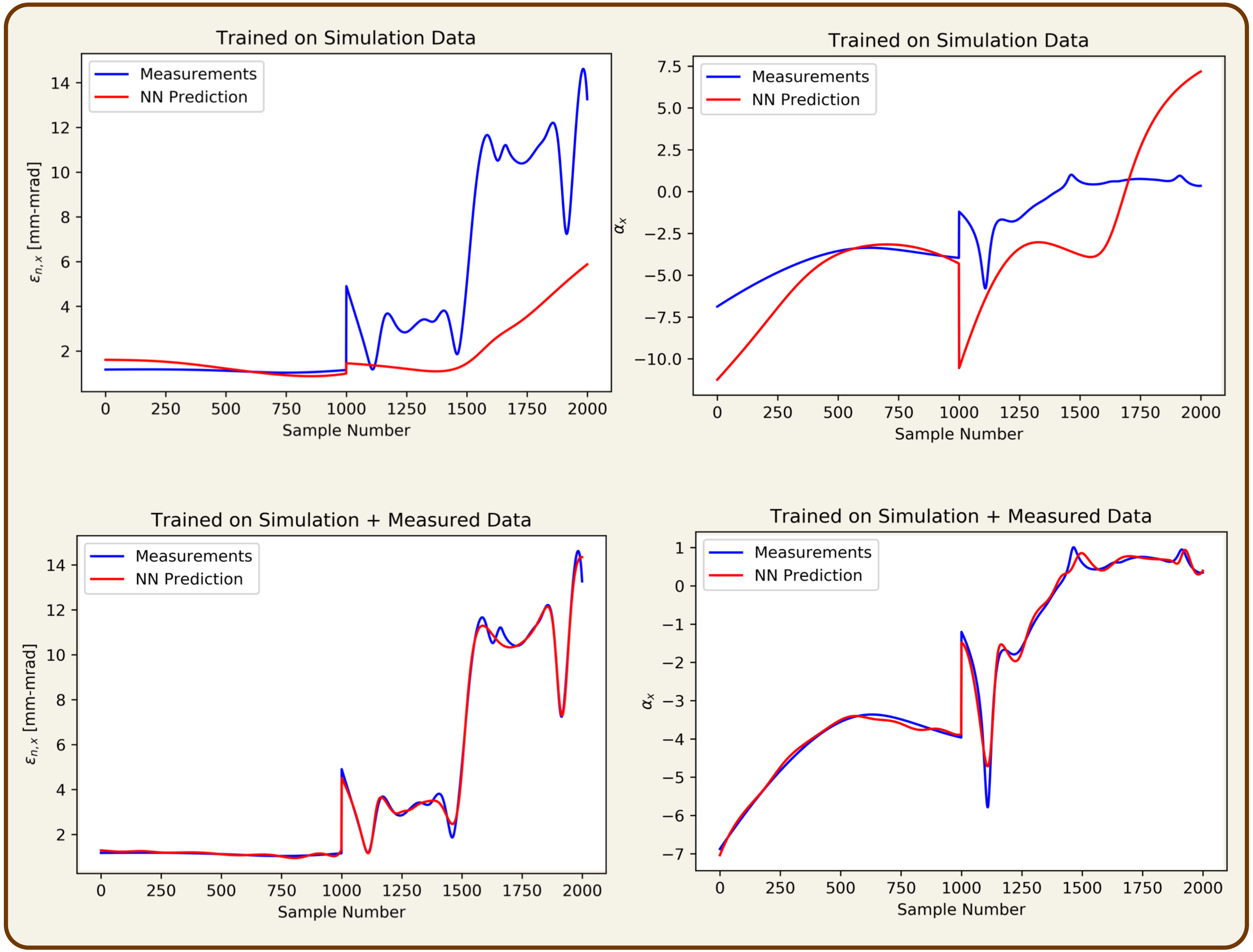


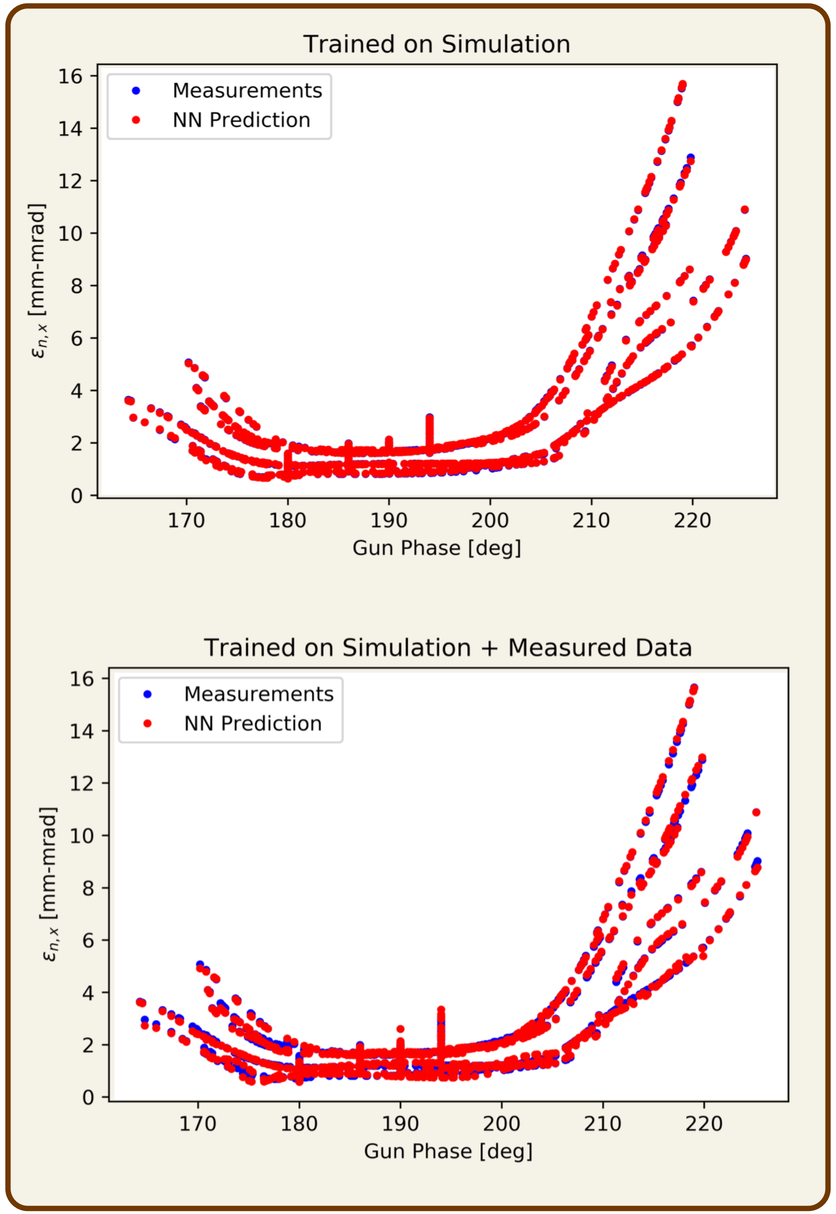
Solenoid Scan

Phase Scan

Simulation Data Only

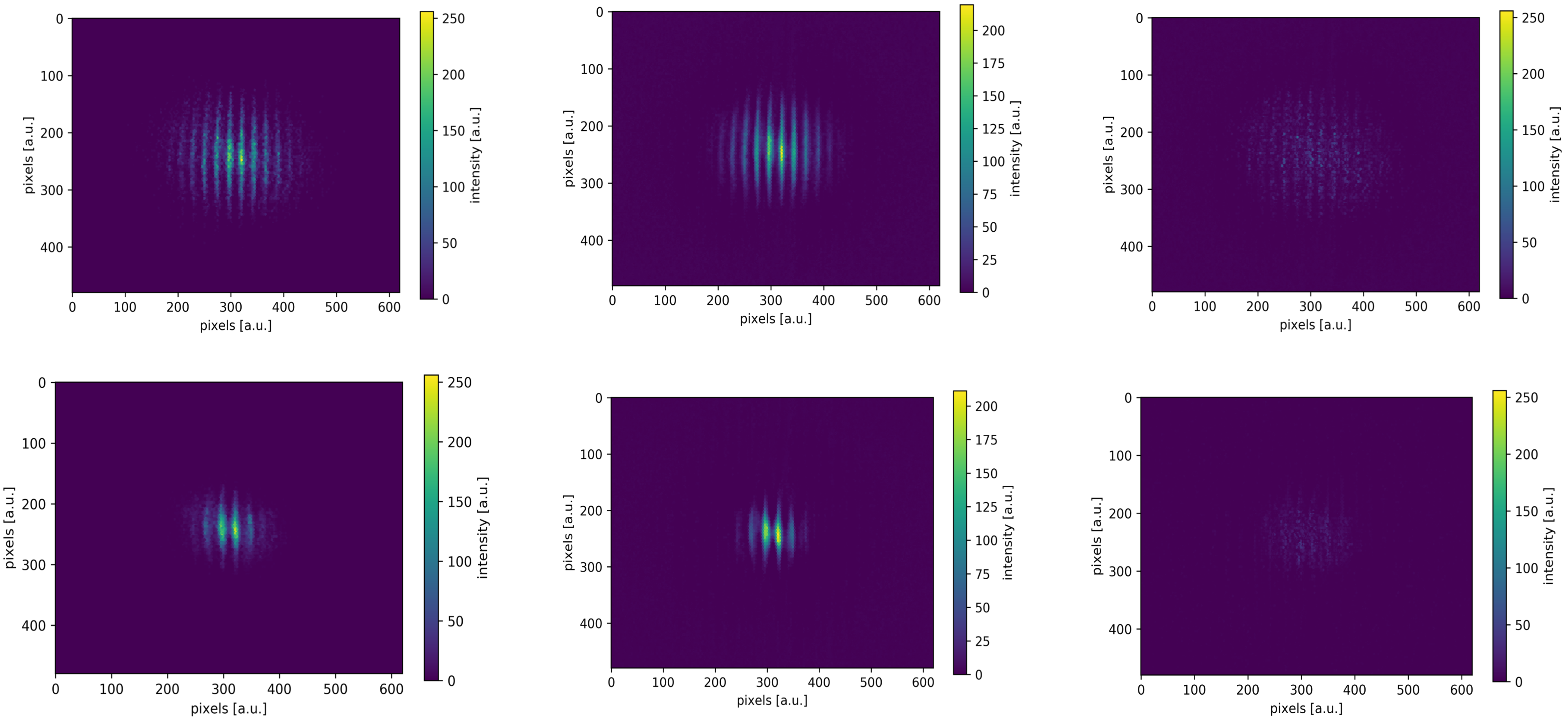
Updated with Measured Data





Why bother with simulation at all? → Rough initial solution facilitates training with small amount of measured data

Predicting Image Output Directly



Simulated

NN Predictions

Difference

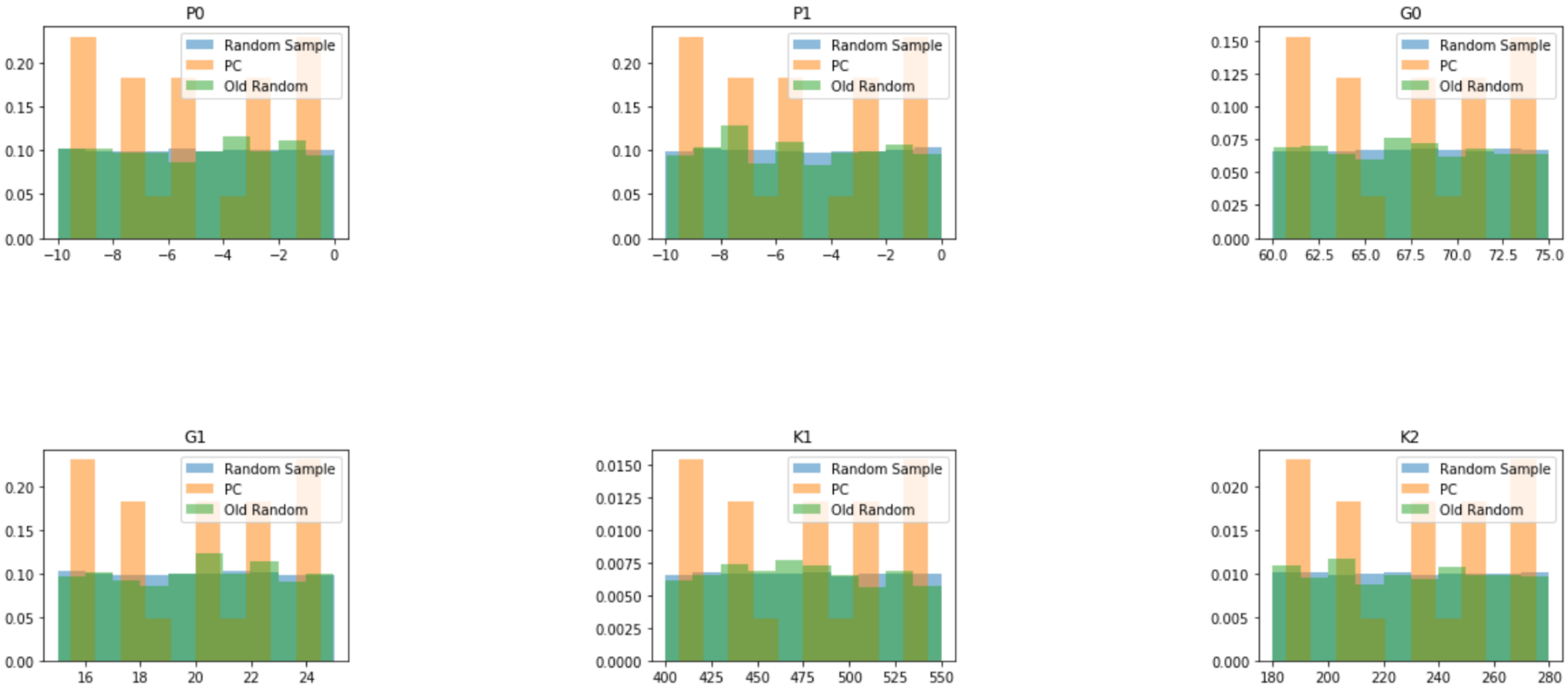
Summary

- Results from AWA look promising with regard to using surrogate model in optimization

- Results from FAST show promise in updating surrogate trained in simulation with measured data + predicting image output directly as a virtual diagnostic

- Still needs more thorough study
    - How to ensure sampling is sufficient to capture behavior
    - Robustness with wider parameter ranges *(for AWA case didn't include cases with particle losses)*
    - Comparison with other models *(looked mainly at NN and PCe)*
    - Prediction uncertainty + sensitivity analysis *(get prediction uncertainty for 'free' with PCe model)*

→ *New initiative at SLAC (with D. Ratner, C. Mayes, N. Neveu) in surrogate modeling for LCLS, LCLS-II + ongoing collaboration between PSI and SLAC*


P0
Random Sample
PC
Old Random
P1
Random Sample
PC
Old Random
G0
Random Sample
PC
Old Random
G1
Random Sample
PC
Old Random
K1
Random Sample
PC
Old Random
K2
Random Sample
PC
Old Random

Initial population: 656

Min population: 328

Cores used: 2624

Nodes (64 cores each): 41

Number of gens: 200

Total time: 16.2 hours

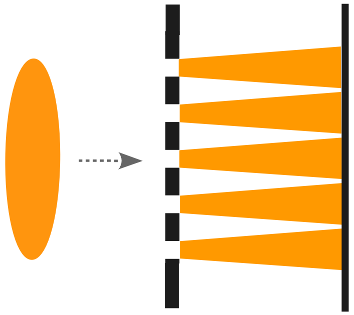
Core hours: ~42,510

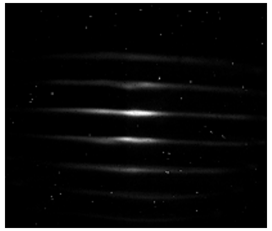
Could in principle use measured data alone, but want to be efficient with machine time

→ *use simulation data to fill out the training set*

gun phase scans
solenoid current scans
(with two different laser intensities)

→





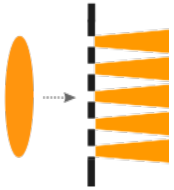
→

fit to obtain subset of phase space parameters

$\langle y^2 \rangle$
$\langle yy' \rangle$
$\langle y'^2 \rangle$

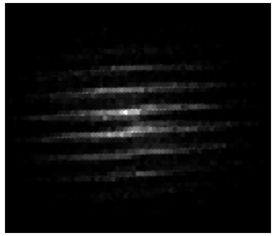
other setting combinations →

OPAL



elegant

→



\+ full sigma matrix

cathode → CC2
with 3-D space charge routine