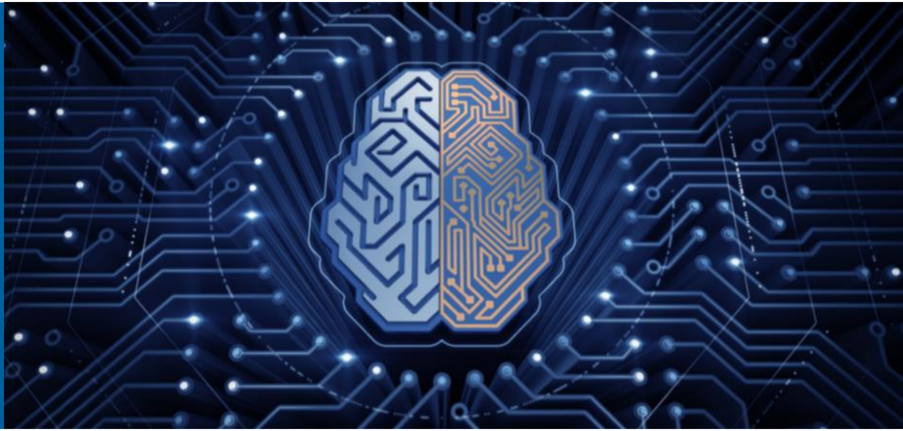
2021 DOE/NP ACCELERATOR R&D AND AI-ML PI EXCHANGE MEETING

Argonne NATIONAL LABORATORY 75 1946–2021

USE OF AI TO OPTIMIZE ACCELERATOR OPERATIONS & IMPROVE MACHINE PERFORMANCE



PRESENTER

Brahim Mustapha
Physics Division
Argonne National Laboratory

CONTRIBUTORS

Jose Martinez, Postdoc (ANL/PHY)
Ian Sugrue, Student (NIU, ANL/HEP)
Anthony Tran, Student (MSU/FRIB)

John Power (ANL/HEP)
Philippe Piot (NIU, ANL/APS)
Yue Hao (MSU/FRIB)

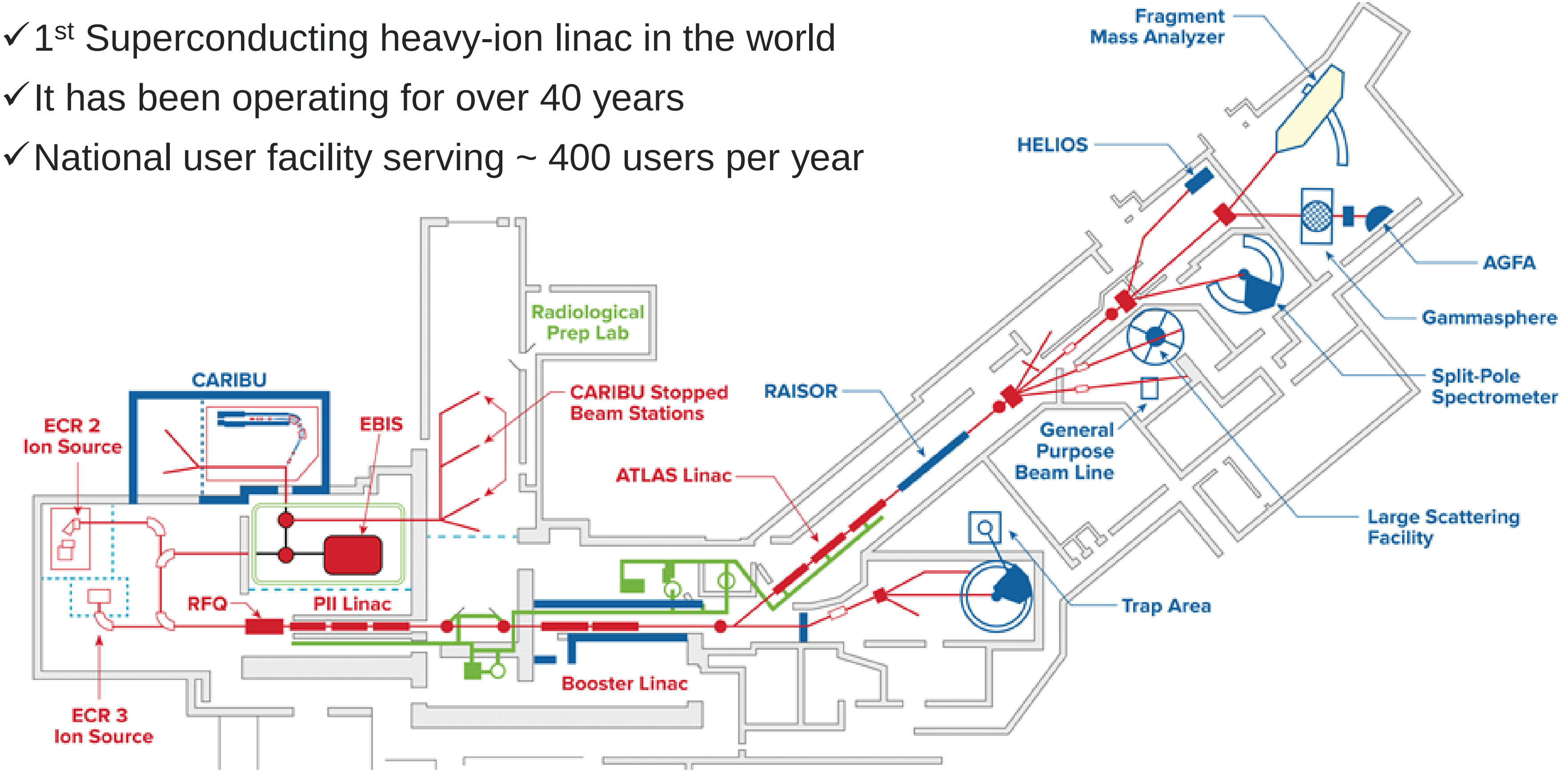
November 30th, 2021
DOE/NP (Via Zoom)

OUTLINE

- ❑Brief Description of the ATLAS AI Project
- ❑The Team / Collaboration
- ❑Project Status: Budget and Summary of Progress
- ❑Progress Highlights at ATLAS, AWA and FRIB

ATLAS: ARGONNE TANDEM LINEAR ACCELERATOR SYSTEM

- ✓1st Superconducting heavy-ion linac in the world
- ✓It has been operating for over 40 years
- ✓National user facility serving ~ 400 users per year



THE ATLAS AI / ML PROJECT:

Use of artificial intelligence to optimize accelerator operations and improve machine performance

- ❑ At ATLAS, we switch ion beam species every 3-4 days … → Using AI could streamline beam tuning & help improve machine performance
- ❑ The main project goals are:
  - Data collection, organization and classification, towards a fully automatic and electronic data collection for both machine and beam data
  - Online tuning model to optimize operations and shorten beam tuning time in order to make more beam time available for the experimental program
  - Virtual model to enhance our understanding of the machine behavior in order to improve performance and optimize particular and new operating modes

THE TEAM / COLLABORATION

- ANL / PHY: B. Blomberg, D. Stanton, J. Martinez and C. Dickerson
  - J. Martinez is a postdoc focused on the ATLAS project
- MSU / FRIB: Y. Hao and A. Tran (PhD student started in May)
  - ATLAS and FRIB have a lot in common, any development for ATLAS will be useful for FRIB and vice versa
- ANL / AWA: J. Power, P. Piot and I. Sugrue (PhD student started in Jan.)
  - AWA can serve as test bed for AI tools development and testing. Being a test facility, more beam time is available for testing tools useful for ATLAS
- ANL / DSL & ALCF: A. Ramanathan and V. Vishwanath
  - Consult & advise on ML/AI modeling, HP computing and data storage at ALCF

BUDGET SUMMARY & EXPENDITURE

	FY-20	FY-21	Total
Funds allocated	$280k	$280k	$560k
Actual costs to date	$120k	$0k	$120k

- ✓ Project officially started in January 2021
- ✓ Budget table above is as of end of September 2021

PROGRESS ON ATLAS WORK

- Data collection effort: … towards fully electronic and automatic collection
  - Beam current readings digitized, saved at request and with machine settings
  - Beam profiles digitized, saved at request and with machine settings
  - Starting a new database that combines both beam and machine data
  - Python wrapper was developed to interact with ATLAS control system to read (data collection) and set (tuning and optimization) machine settings
- Online tuning model:
  - A simulation-based model was developed, it uses Gaussian processes and Bayesian optimization to tune for maximum beam transmission
  - Model expanded to MHB-RFQ section of ATLAS using old tunes as starting data
  - Work in progress to add misalignment and steering for online test early next year
- Virtual machine model:
  - A surrogate model for the RFQ was already developed, it's significantly faster than TRACK simulations, allowing real-time comparison with the machine
  - Work is in progress to develop a model for the PII section of the linac

PROGRESS ON AWA WORK

AWA as a testbed for ML-based machine tuning and virtual diagnostics development

- ❑ Surrogate NN model mapping beam images to input lattice parameters
  - Simulation data generated using the OPAL code for the main AWA line
  - 9 Lattice parameters varied to generate ~ 10 k of beam images (YAG images)
  - Beam images analyzed and reduced using Partial Component Analysis (PCA)
  - NN model built, different loss functions tested, PCA norm converges well.
- ❑ Solving the inverse problem using the surrogate Model:
  - Goal: reproduce a “nice” beam image with unknown or uncertain lattice settings
  - The surrogate model is very fast and was used to fit the desired beam image and get the corresponding lattice settings
  - Work in progress to test this procedure experimentally.
- ❑ Experimental side: … data collection and YAG image processing …
  - Developed and tested scripts to acquire beam YAG images and accelerator settings
  - Currently exploring image-size reduction from 1440x1080 to more manageable pixels
  - PCA technique provides a digital filter and removes some image noise and artifacts

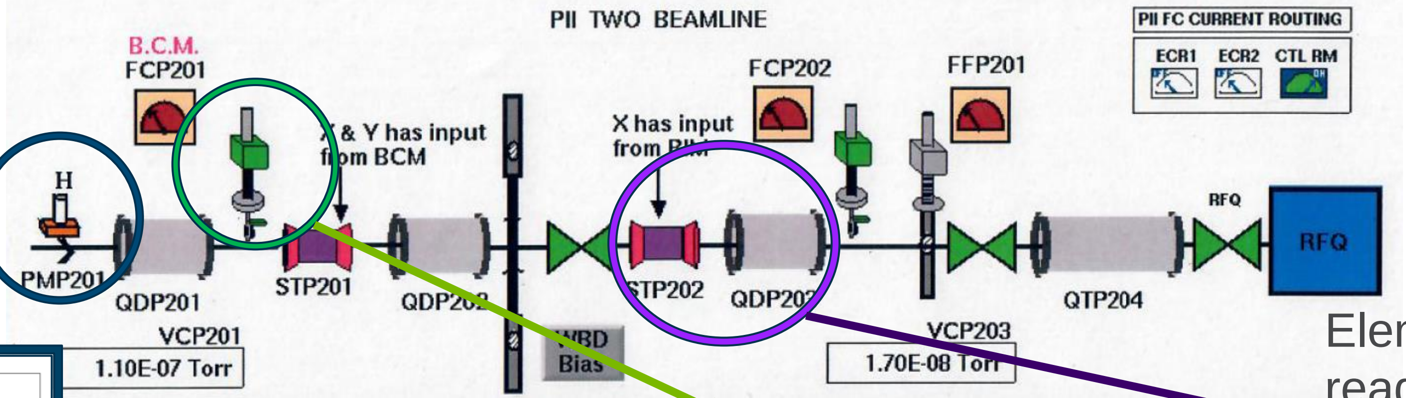
PROGRESS ON FRIB WORK

Transfer learning between ATLAS and FRIB based on similarities

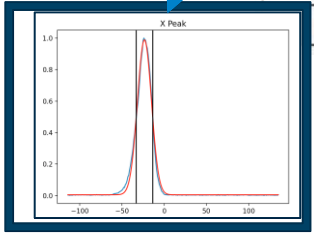
- Development of surrogate models for beam emittance and particle loss
  - Simulation data generated using TRACK code for a short ATLAS section
  - NN model was trained to reproduce transverse emittance growth and beam loss
  - A Gaussian process model was developed and compared to NN model
  - Gaussian process model is more useful for Bayesian optimization
- Bayesian optimization for single and multiple objective
  - Single objective optimization for both emittance and beam loss → beam loss dominate
  - Multiple objective optimization for emittance and beam loss separately
  - Work in progress …
- Modeling with initial beam distribution using convolutional neural network (CNN)
  - Data generated using TRACK with different initial distributions and lattice settings
  - Model trained using images of projections of initial distributions in phase space
  - Goal is to see if the model can predict beam transmission for an irregular distribution

PROGRESS HIGHLIGHTS - ATLAS

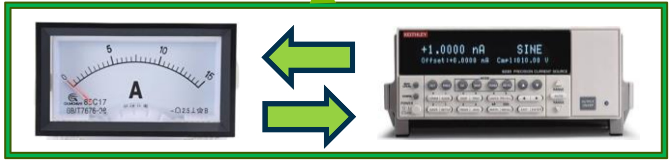
DATA COLLECTION & ACCESS TO CONTROL SYSTEM



Elements: read/set



BPM: digitized, read

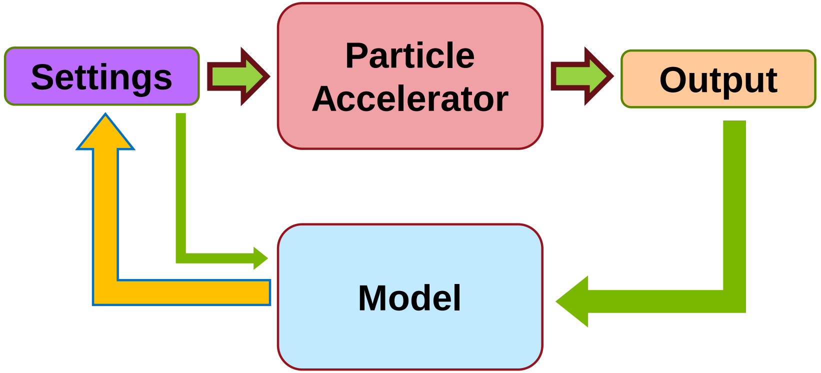


FC: digitized, insert, read

BAYESIAN OPTIMIZATION FOR ONLINE BEAM TUNING

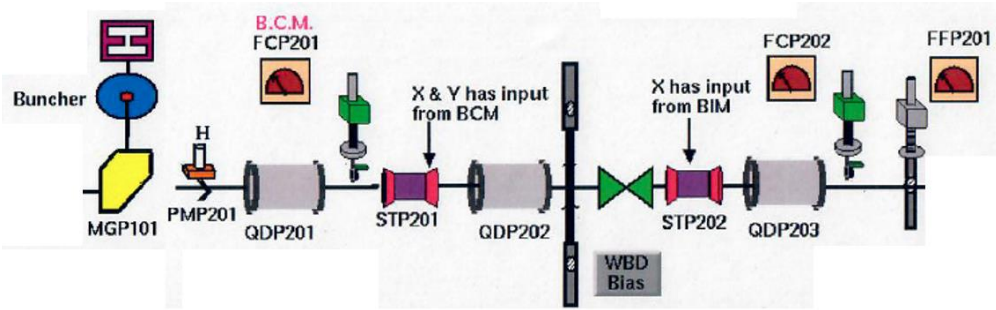
Goal: Find the global optimum in minimum number of iteration steps

Principle of Bayesian Optimization



- ❑ Explicitly unknown objective function: f(x)?
- ❑ First: Build a probabilistic model (Surrogate Model) based on initial data sample
- ❑ Second: Choose next point to improve objective and decrease uncertainty (Acquisition Function)
- ❑ Third: Sample new point, update the model and repeat until convergence (Optimization Loop)

Applied to a subsection of ATLAS Linac (MHB to RFQ)

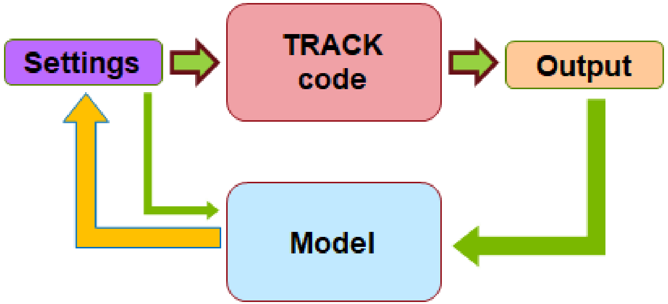


- ✓ Maximize transmission by varying voltages of 6 quadrupoles
- ✓ Not using the machine yet → TRACK code acts as machine
- ✓ Quads limited from -9 kV to 9 kV, data normalized for training
- ✓ Surrogate Model: Gaussian Process with Matern Kernel and Gaussian likelihood (Works with a limited data sample)
- ✓ Acquisition Function: Expected Improvement
- ✓ GPyTorch used for Gaussian Process
- ✓ BoTorch Library used for Bayesian Optimization

ONLINE TUNING MODEL – FIRST RESULTS

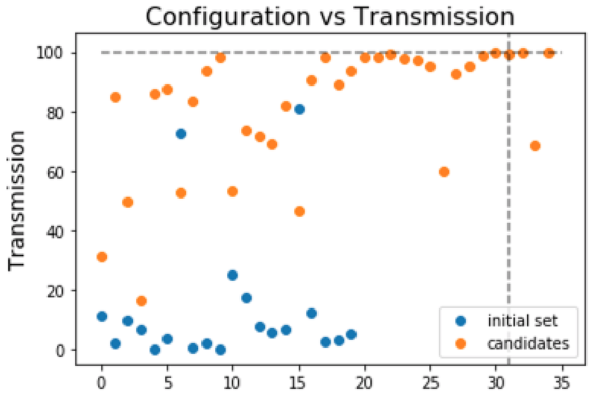
Goal: Find the best tune in a minimum number of setting changes

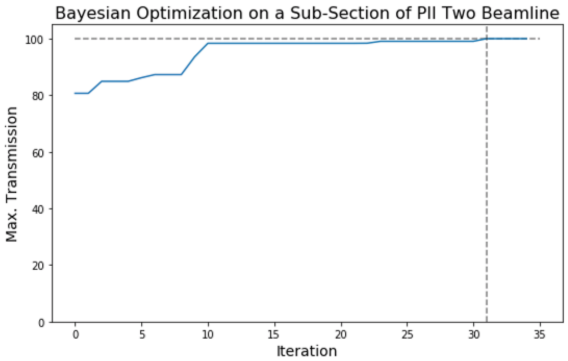
The Actual Model, Simulation based



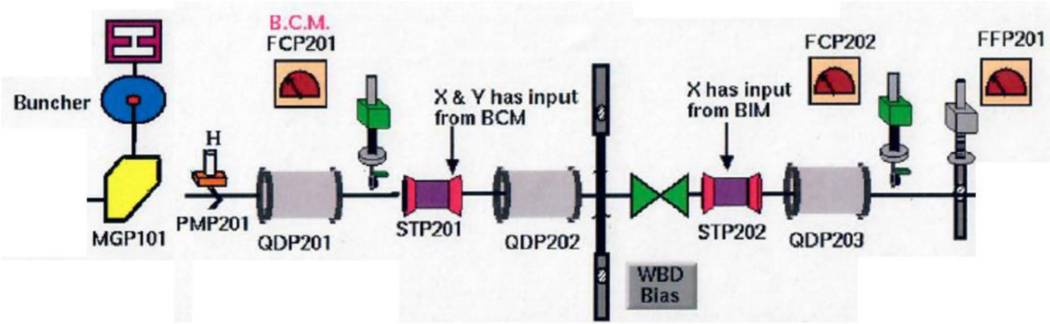
- Initial Data: 20 random settings of 6-quads and corresponding beam transmissions (can be old tunes from the tunes/beam database)
- Target function: loss rate (1-transmission) from TRACK using input quads setting
- Converged in 31 iterations, but depends on size and quality of initial sample
- Convergence from ~ 80% to 100% in ~ 3 min

The Results

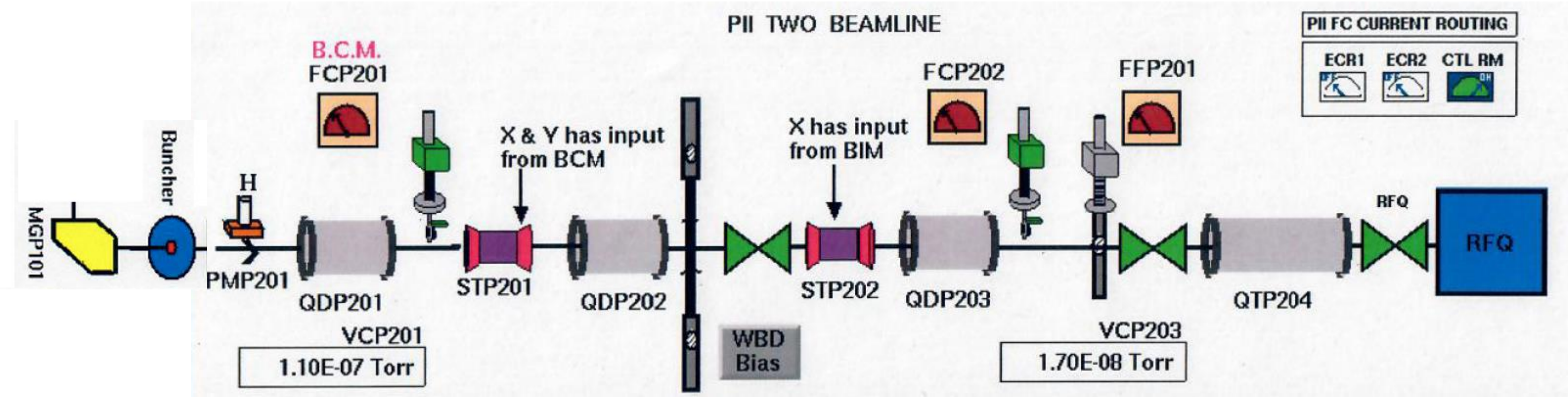




ONLINE TUNING MODEL – EXPANDED TO INCLUDE RFQ



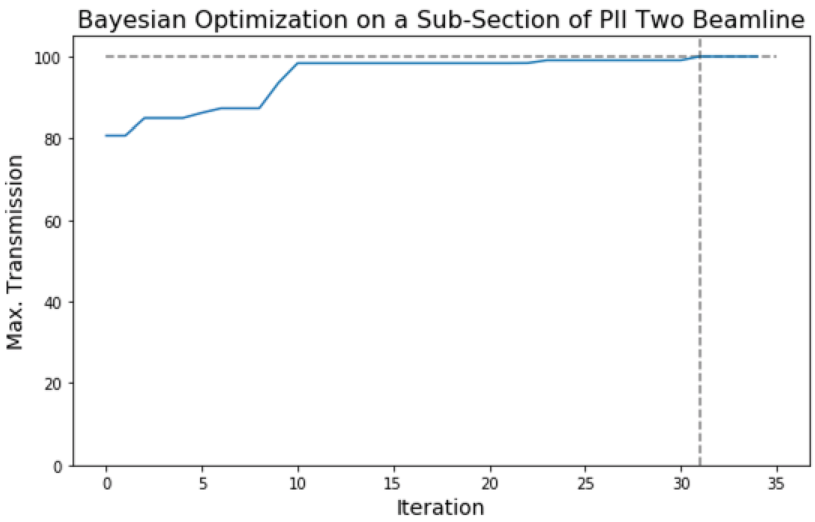
3 e-quad. doublets



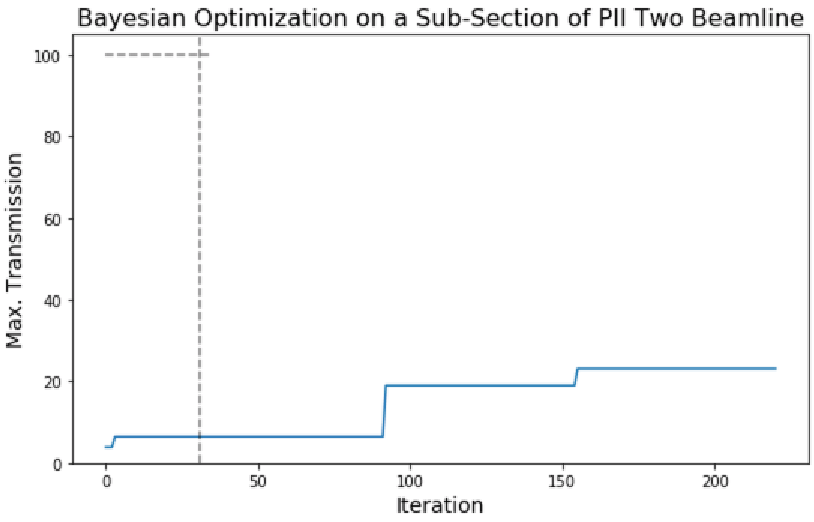
3 e-doublets
+ e-triplet
+ RFQ

ONLINE TUNING MODEL – WITH RFQ

No RFQ



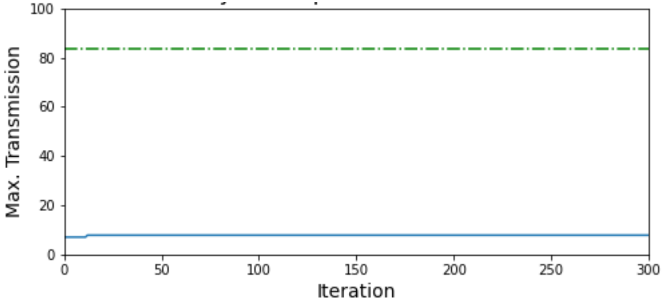
With RFQ



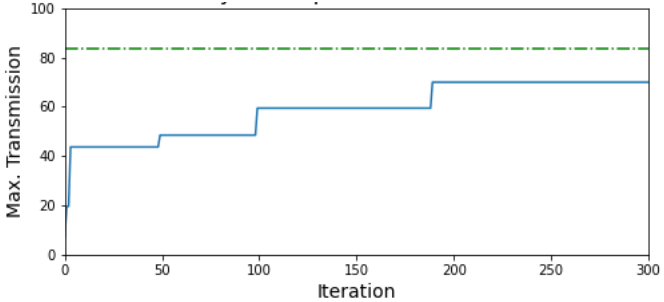
- ➢ The original model is not working with the RFQ, starting with low transmission and converging very slowly!
- ➢ The RFQ requires accurate transverse beam matching, highly constraining the quad settings
- ➢ Starting with a random or unconstrained settings will take forever to converge → use known settings
- ➢ Need to use existing tunes data and known constraints …

ONLINE TUNING MODEL – USE DATA & CONSTRAINTS

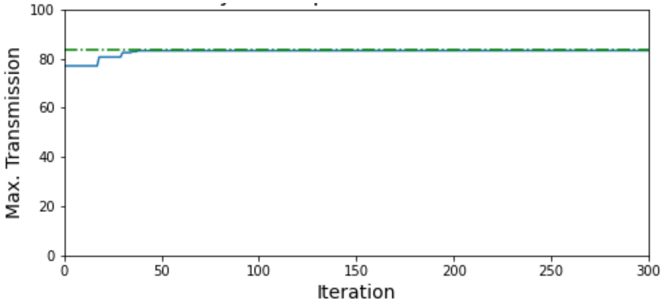
Initial data: 50 random settings
New quad settings: unconstrained



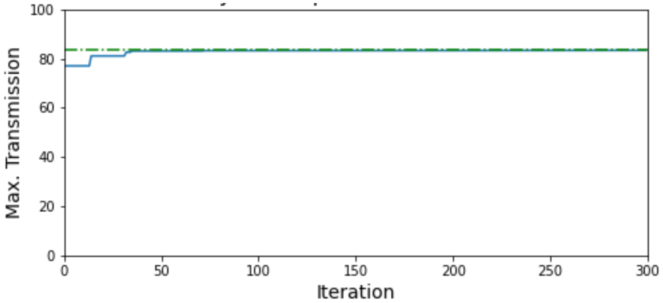
Initial data: 50 random settings
New quad settings: constrained



Initial data: 29 old tunes (scaled)
New quad settings: unconstrained

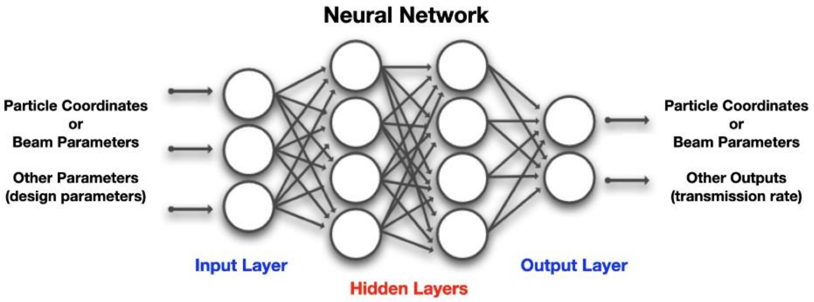


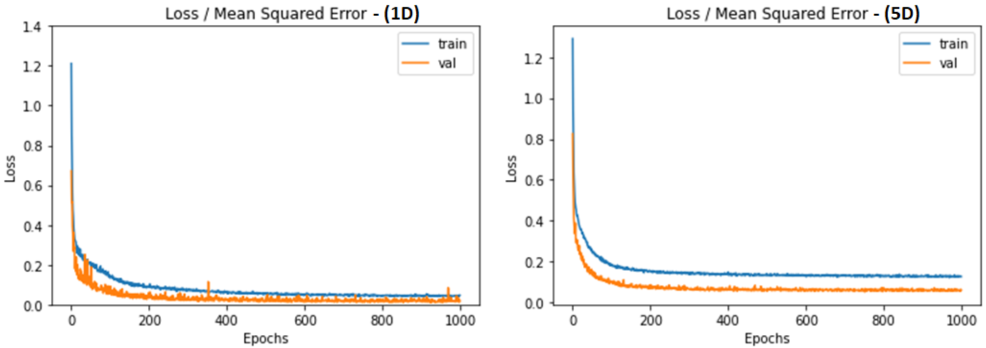
Initial data: 29 old tunes (scaled)
New quad settings: constrained

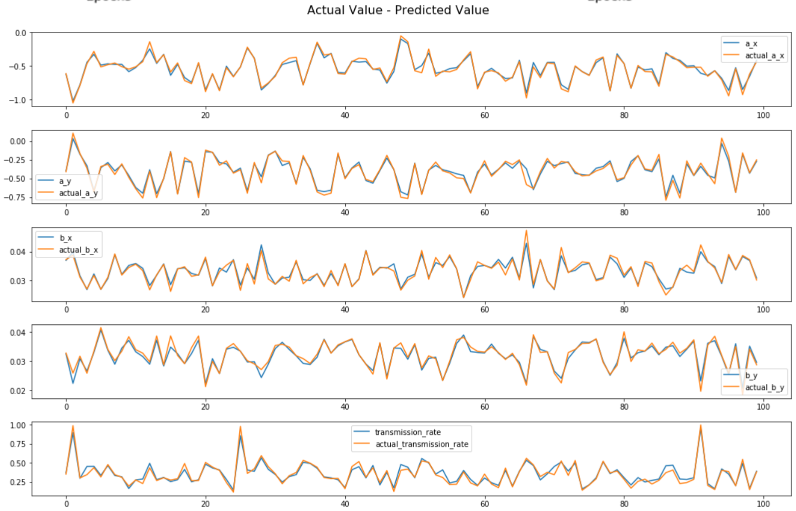


- The model improvement is clear when using existing tunes and known setting constraints!

SURROGATE ML MODEL FOR THE ATLAS RFQ







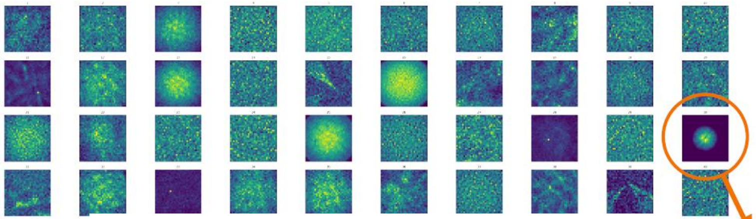
- ❑ We used a neural network for this model, which is fully based on simulations data
- ❑ Excellent convergence for 1D results, will need more data for the 5D case!
- ❑ Excellent agreement with TRACK 3D beam simulations, similar to # codes comparison!
- ❑ Much much faster than TRACK, speed-up factor ~ 30,000 → can use online

PROGRESS HIGHLIGHTS - AWA

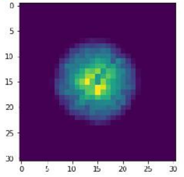
BEAM IMAGE ANALYSIS: PRINCIPAL COMPONENT ANALYSIS

Goal: Reduce a large set of images to be represented by a vector base

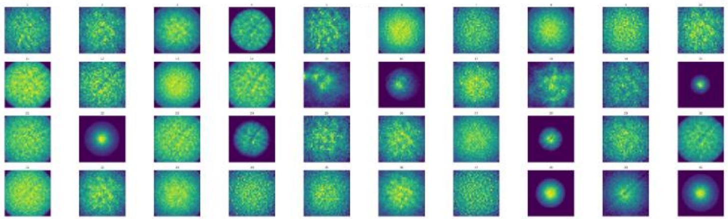
- Step-0: Generate/load data in the form of beam images, here generated by simulations
- Step-1: Select a "nice" representative image of a beam, let's call it X_0
- Step-2: Use k-nearest neighbor method to select k "nicest" images
  - Let X_i represent the i'th image in our dataset.
  - For all i, calculate the Euclidean norm $| X_i - X_0|$
  - Sort the X_i's from smallest to largest norm
  - Take the first k images and their parameters
- Step-3: Perform PCA by matrix SVD
  - Center the data: subtract mean values, add after
  - Perform singular value decomposition on X matrix: $X = U\Sigma V^T$; U and V orthogonal, Σ diagonal matrix of singular values
  - The orthogonal basis of images are columns of V



Original Images

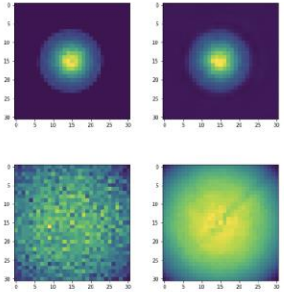


A "nice" image

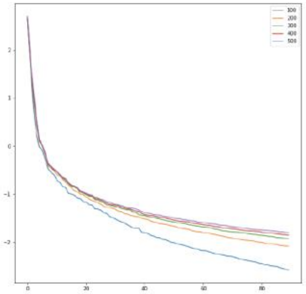


Sample of k=400 nearest neighbors

Original PCA-15



SVD as function of k

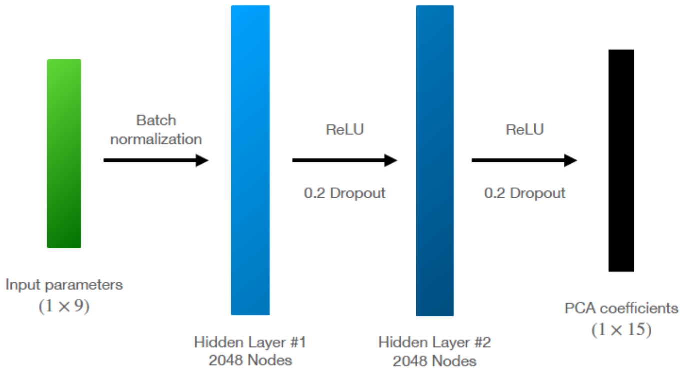


SVD values saturation at k ~ 400

SURROGATE MODEL: NN MAPPING IMAGE TO INPUT PARAM.

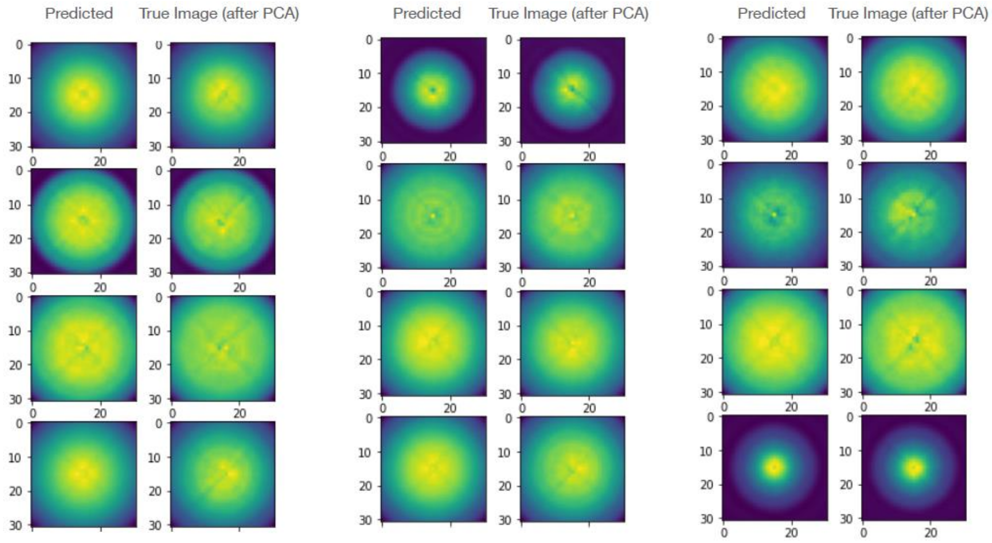
Goal: Associate a given image to given input lattice parameters

Neural Network architecture



- ❑ 9 Input lattice parameters
- ❑ Images reduced to 15 PCA components
- ❑ Two hidden layers of 2048 nodes each
- ❑ ~ 500 epochs, default batch size (32), MSE loss function

Preliminary results: Predicted vs. PCA images

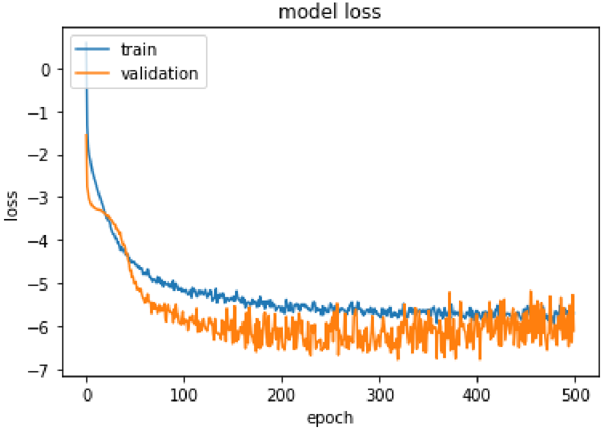


- ✓ Very good results given the complexity of the problem
- ✓ Surrogate model is ready to solve the inverse problem: Reproduce a nice beam image for which the lattice settings were not saved, or uncertain due to drift in time.

SURROGATE MODEL: NN MAPPING IMAGE TO INPUT PARAM.

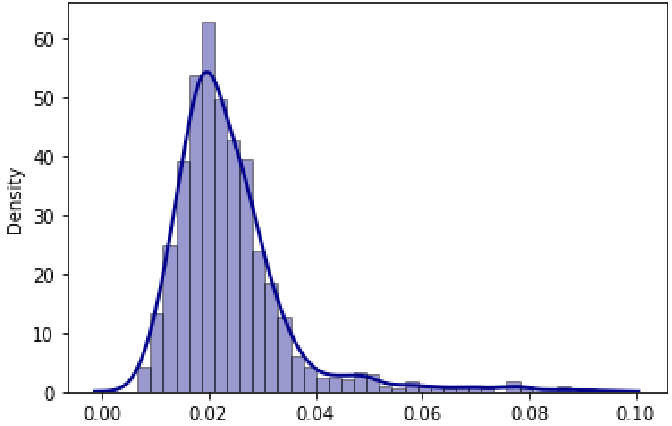
Goal: What are the lattice parameters for best beam quality / image?

Neural Network convergence



Training (blue trace) and validation (orange trace) loss in logarithm unit as a function of number of epochs used in the surrogate model.

Least mean square fit of desired beam image



Histogram/density plot of the relative error of the non-linear least squares problem after 1000 simulations and least square fit for retrieval of control parameters. Most results indicate a relative error of ~ 2% is attained

PROGRESS HIGHLIGHTS - FRIB

GAUSSIAN PROCESS – SINGLE & MULTIPLE OBJECTIVES

Prepared data: 10000 TRACK simulations

Lattice
1. drift
2. eq3d
3. eq3d
4. drift
5. eq3d
6. eq3d
7. drift
8. eq3d
9. eq3d
10. drift

Input
1. Vf1
2. Vf2
3. Vf3
4. Vf4
5. Vf5
6. Vf6

Single Output
1. 4*εx_rms

Multiple Output
1. 4*εx_rms
2. 4*εx_rms
3. Beam loss

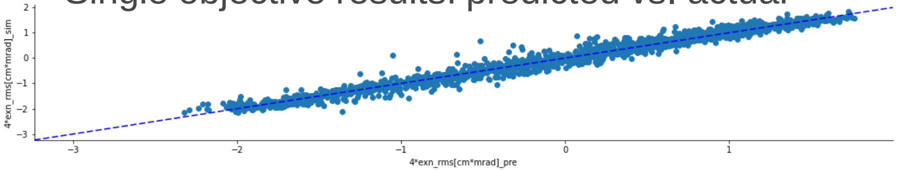
Training set:

80% of data

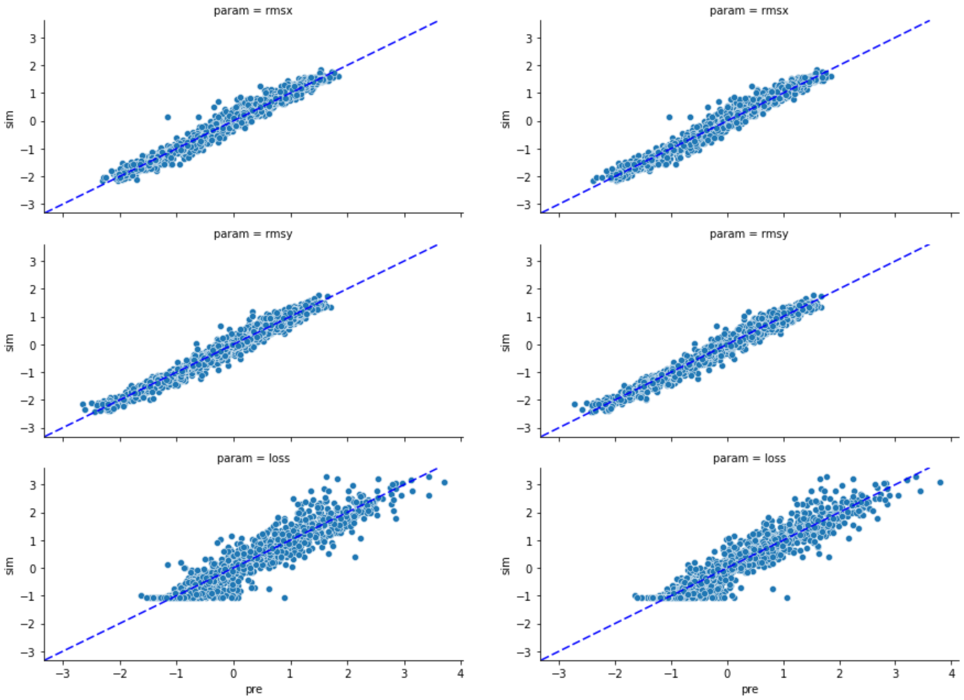
Test set:

20% of data

Single objective results: predicted vs. actual



Multitask: Correlated output | **Multibatch: Independent output**

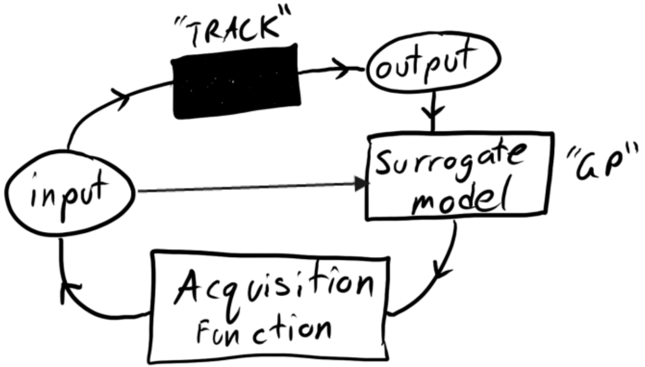


Multiple objective results: predicted vs. actual

BAYESIAN OPTIMIZATION: SINGLE VS. MULTI-OBJECTIVE

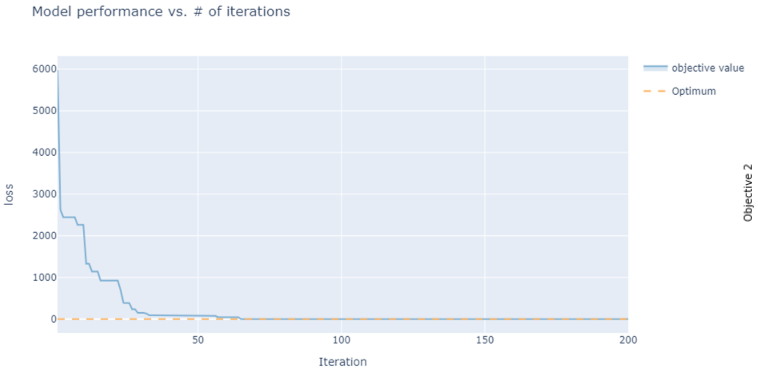
Basic Idea/Analogy: **Multi-armed bandit**

- A gambler at a row of slot machines. A slot machine is one-armed bandit.
- Each machine gives rewards according to a probability distribution specific to that machine, but at the start this is unknown.
- You have to trade between exploration vs exploitation.



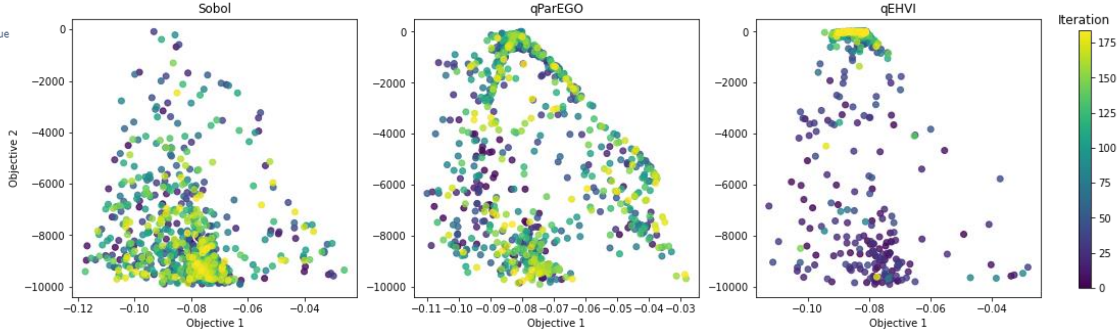
Single Objective:

- Loss Function = 4*rmsx + 4*rmsy + particle loss
- Particle loss dominates in this case



Muti-Objective:

- Find the **pareto front**: the optimal set of non-dominated points where no objective can be improved without sacrificing at least one other objective.
- 4*rmsx, 4*rmsy, and particle loss are the three objectives.



MODELING TRACK WITH CNN

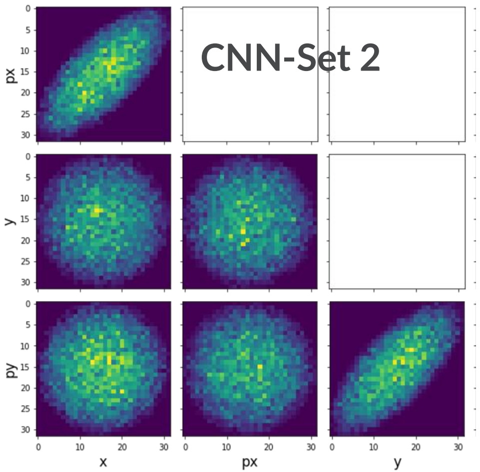
The initial distribution affects directly the beam transmission and emittance.

Use 9 different distribution created in TRACK. ex/ Waterbag, Uniform, Gaussian, KV

Created around 100,000 samples.

Deposit particles on NxN grid for each phase space pair.

Images of 6-d phase space. 4D Waterbag distribution →



Training and Test datasets

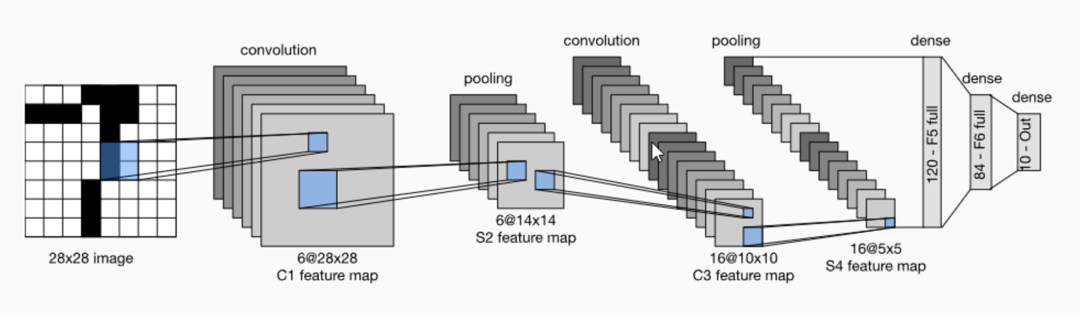
Set 1: 200,000 samples
Input: 6 quad. voltages
Output: Particles left out of 10k
Set 2: 100,000 samples
Input: 6 voltages & 9 distributions
Output: Number of particles left
Set 3: 25,000 samples
Input: 6 voltages &
1 non-ideal distribution
Output: Number of particles left

CNN: Convolutional Neural Network Method:

Use images as input.

Extract features and use these features in a NN. →

Add voltage input along with the features from the images in a NN to obtain number of particles left.



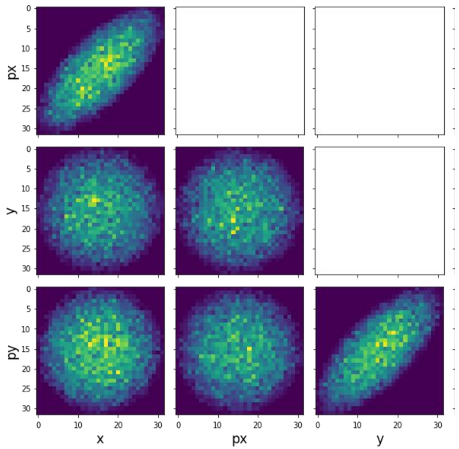
6.6. Convolutional Neural Networks (LeNet) — Dive into Deep Learning 0.17.0 documentation (d2l.ai)

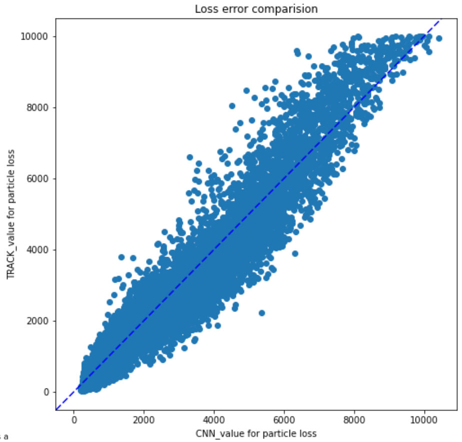
RESULTS AND TEST ON NON-IDEAL DISTRIBUTION

Model trained on 80,000 samples of regular beam distributions →

It was tested on 13,000 samples of similar distributions

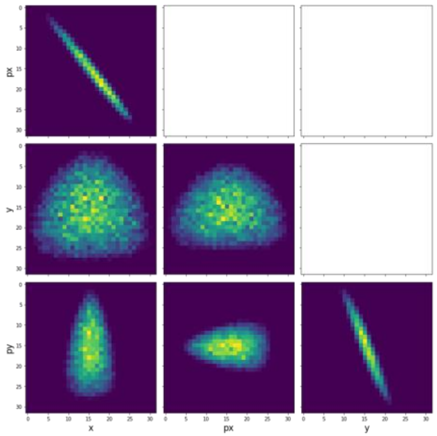
↓

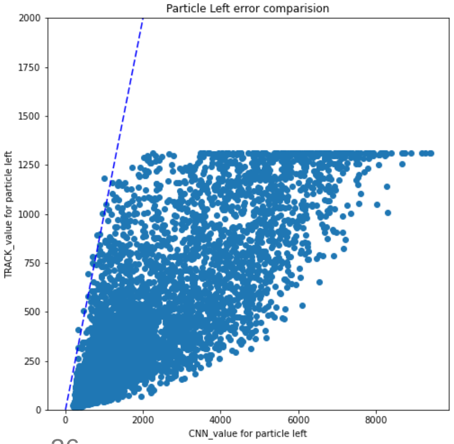




CNN-Set 2
Regular distribution
Mean error: 400
Std error: 450

→ Reasonable





Question: How well does the model generalize to data it never seen before?

← Distribution generated using a combination of quadrupole, sextupole, and drift → non-ideal distribution, images of odd phase space projections

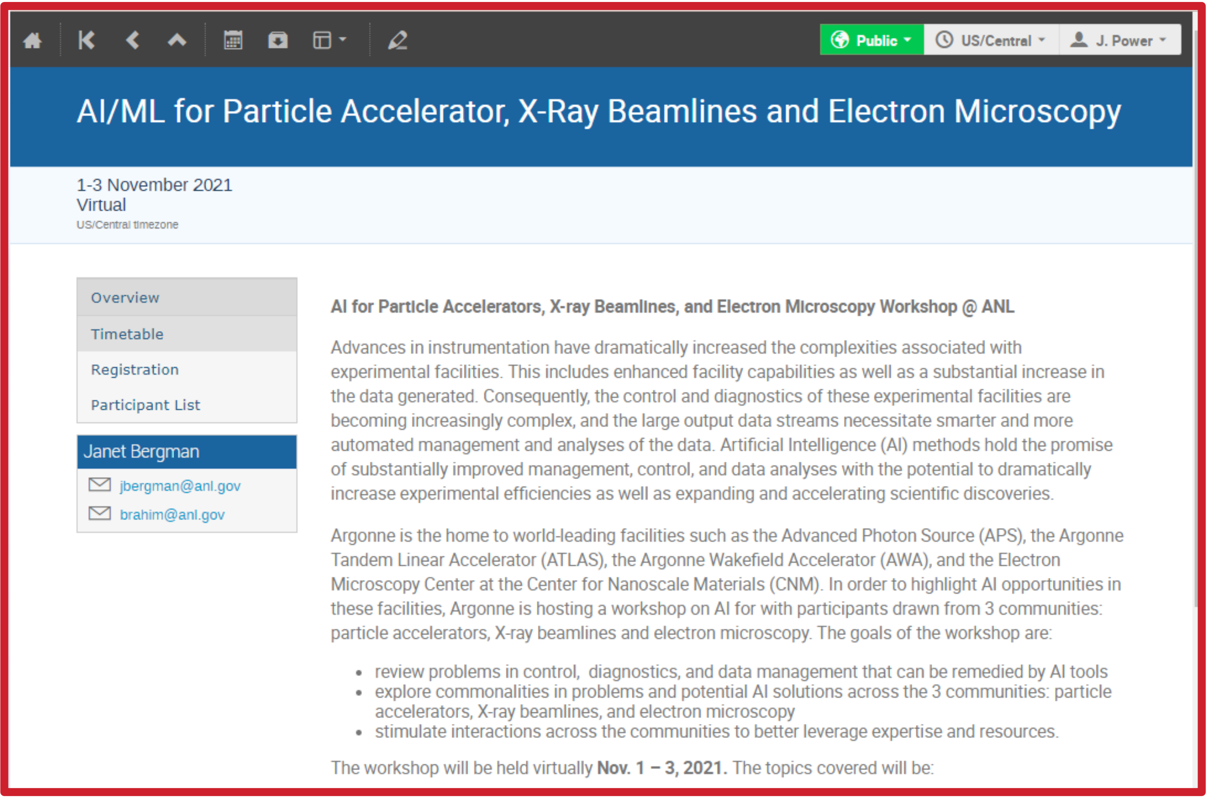
Created around 25,000 samples using this initial distribution, but different voltages → CNN Set 3

CNN-Set 3
Irregular beam distribution
Mean error: 1604
Std error: 1170

→ Not good, more work is needed!

RECENT AI-ML WORKSHOP AT ARGONNE

AI/ML WORKSHOP @ ARGONNE



https://indico.fnal.gov/event/50731/overview

Workshop logistics

- ANL hosted a 3-day workshop
- November 1st - 3rd, 2021
- Public indico page
- All/most talks posted
- All sessions recorded
- 144 participants

Labs- (ANL, BNL, FNAL, LBNL, LANL, TJNAF, ORNL, PNNL, SLAC, Canadian light source)
Industry- (Euclid, RadiaSoft)
Universities- (MIT, MCS, MSU, Cornell University, NIU, Northwestern University, UChicago, University of California, Santa Barbara, University of Illinois at Urbana-Champaign, University of Illinois at Urbana-Champaign, University of Illinois at Urbana-Champaign, University of Michigan University of Pennsylvania, University of Wisconsin, Bucknell University)

AI/ML WORKSHOP @ ARGONNE

Workshop objectives & goals

- **First local meeting** of three different communities to compare notes about AI/ML efforts, invited speakers from other National Labs and Universities
- **Review** the different AI/ML methods and techniques developed and applied in the 3 communities
  - Particle Accelerators
  - X-ray Beamlines
  - Electron Microscopy
- **Learn** about (new or different) AI/ML techniques being applied in other communities.
  - Although the specific problems are different, applicable AI/ML methods may be similar.
  - Exchange ideas and explore ways to work together.

AI/ML WORKSHOP @ ARGONNE

Cross cutting ANL organizing committee

WORKSHOP ORGANIZERS FROM EACH OF THE THREE COMMUNITIES

Particle Accelerators	X-ray Beamlines	Electron Microscopy
Michael Borland (APS)	Olle Heinonen (PSE)	Jianguo Wen (NST)
Brahim Mustapha (PHY)	Nicholas Schwarz (APS)	Charudatta Phatak (MSD)
John Power (HEP)	Martin Holt (NST)	

AI/ML WORKSHOP @ ARGONNE

Six sessions

	Monday	*Tuesday*	*Wednesday*
AM	**Automated tuning and control**	**Imaging and Data Processing**	**Autonomous Discovery**
PM	**Data Analytics**	**Failure Detection, Virtual Diagnostics, and Digital Twins**	**Data, Computing, and Modeling**

Each session ended with ~ 30 min panel discussion.


Argonne
NATIONAL LABORATORY
1946
75
2021

THANK YOU ALL