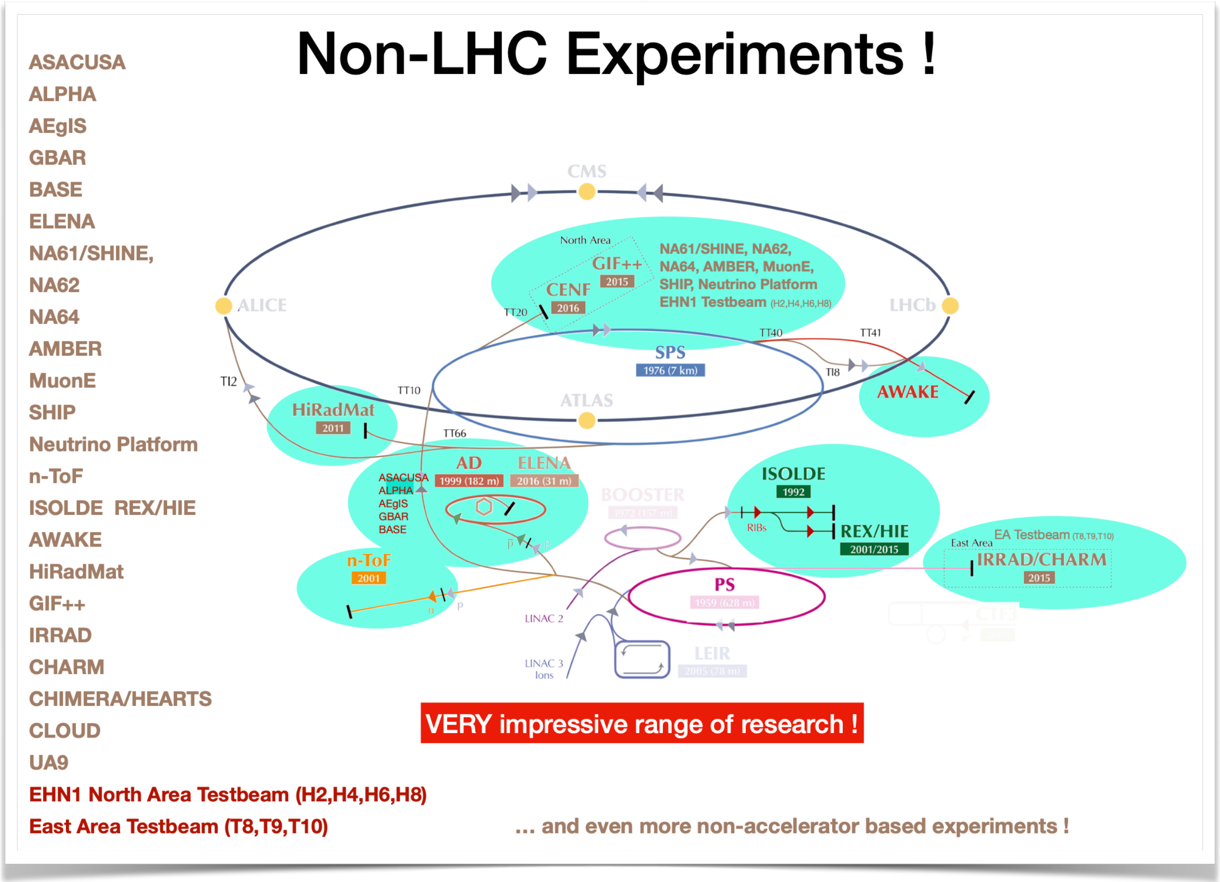
ML approaches for the operation of the CERN accelerator complex

V. Kain, N. Burchon, A. Calia, L. Felsberger, J.C. Garnier, R. Gorbonosov, M. Hostettler, A. Huschauer, F. Irannejad, D. Jacquet, N. Madysa, B. Rodriguez Mateos, K. Papastergiou, C. Petrone, M. Remta, M. Schenk, M. Sobieszek, G. Trad, F. Velotti, J. Wulff

Low Emittance Rings workshop, V. Kain, 15-Feb-2024

LHC and non-LHC physics

Not only LHC physics! → Many different beam types, production schemes,…



Screenshot from recent Chamonix CERN Accelerator Performance workshop

Why AI/ML for CERN accelerators? Examples...

Summary talk *CERN Injector and Experimental Facility Workshop (IEF) '21*

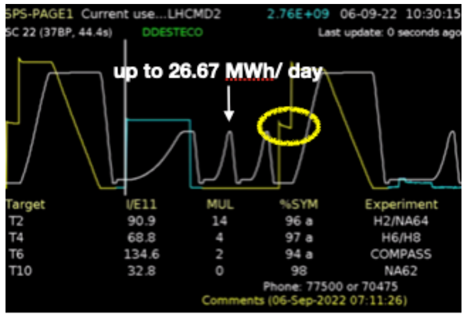
2. Address reproducibility and availability

- Availability OK, under control of Groups. **Reproducibility** is critical concern with increasing flexibility and multi-destination operation
- Transmission problems and instability in beam delivery in many locations. "Need more time in 2022"→have to ensure this is there (add in schedule?) #A
- Addressing reproducibility relies on many factors including equipment, accelerator modelling and high-level controls approach

Input from CERN Joint Accelerator Performance Workshop'22

→ **Hysteresis** is severe limitation for efficiency and flexibility in most machines, current mitigation methods wasting energy

- ~ 15 % of yearly cost of SPS fixed target cycle for "waste" cycles and quasi-degauss Cycle MD1



Efficiency Think Tank (ETT): Oct '22 - Mar '23

ETT = Body for brainstorming for strategy definition for more efficient CERN accelerator exploitation

Response to *Injector and Experimental Facility Workshop 2021* concerns with efficiency and reproducibility

- large extended team for community discussions, small core team to synthesise directions

Wide range of efficiency topics touched

- shorter turn-around, more flexibility, energy efficiency,…

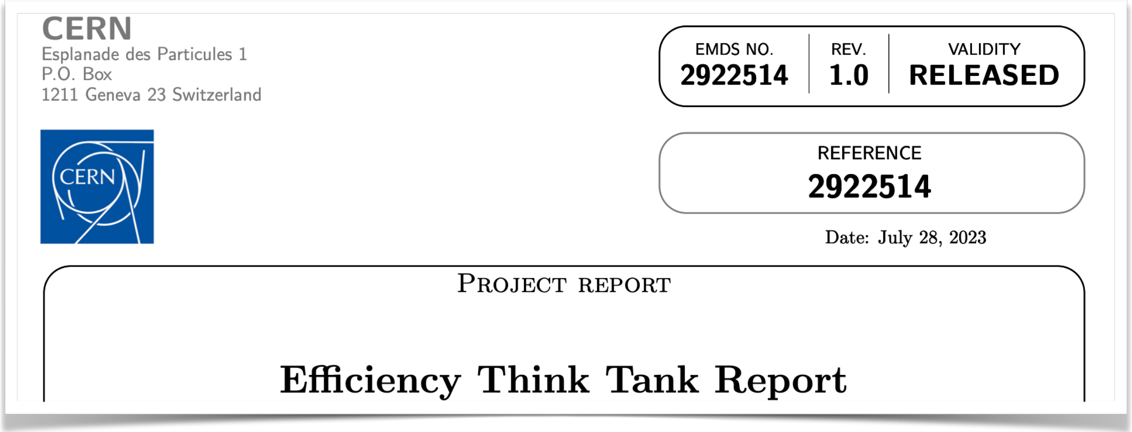
Key target areas identified

- 7 high priority recommendations

→ CERN accelerator sectorwide project to implement recommendations

- **Efficient Particle Accelerators (EPA) project:** 5 year project → improvements ready for HL-LHC

7 recommendations → Automating exploitation


CERN
Esplanade des Particules 1
P.O. Box
1211 Geneva 23 Switzerland

EMDS NO.	REV.	VALIDITY
2922514	1.0	RELEASED

REFERENCE
2922514

Date: July 28, 2023

PROJECT REPORT

Efficiency Think Tank Report

1. **Hysteresis compensation**
2. **Automatic and dynamic beam scheduling**
3. **Automatic LHC filling**
4. **Auto-pilots**
5. **Automatic fault analysis, recovery and prevention**
6. **Automatic testing and sequencing**
7. **Automatic parameter optimisation**

7 recommendations → Automating exploitation

CERN
Esplanade des Particules 1
P.O. Box
1211 Geneva 23 Switzerland

EMDS NO.	REV.	VALIDITY
2922514	1.0	RELEASED

REFERENCE
2922514

Date: July 28, 2023

PROJECT REPORT

Efficiency Think Tank Report

1. **Hysteresis compensation**
2. **Automatic and dynamic beam scheduling**
3. **Automatic LHC filling**
4. **Auto-pilots**
5. **Automatic fault analysis, recovery and prevention**

→ **Automating standard physics operation**

6. **Automatic testing and sequencing**
7. **Automatic parameter optimisation**

→ **Automating commissioning**

7 recommendations → Automating exploitation

CERN
Esplanade des Particules 1
P.O. Box
1211 Geneva 23 Switzerland

EMDS NO.	REV.	VALIDITY
2922514	1.0	RELEASED

REFERENCE
2922514

Date: July 28, 2023

PROJECT REPORT

Efficiency Think Tank Report

→ **Automating standard physics operation**

- AI Hysteresis compensation
- 2. Automatic and dynamic beam scheduling
- AI Automatic LHC filling
- AI Auto-pilots
- AI Automatic fault analysis, recovery and prevention

→ **Automating commissioning**

- 6. Automatic testing and sequencing
- AI Automatic parameter optimisation

The waves of automation @ CERN

The **current** automation efforts are based on three threads

- **Automation wave 1** (2006 -)
  - reduce complexity through models (LSA)
  - high level parameter control, sequencers, software interlock system, classic control algorithms in feedforward and feedback (SVD, COSE,…)

The waves of automation @ CERN

The **current** automation efforts are based on three threads

- **Automation wave 1** (2006 -)
  - ✳ reduce complexity through models (LSA)
  - ✳ high level parameter control, sequencers, software interlock system, classic control algorithms in feedforward and feedback (SVD, COSE,…)
- **Automation wave 2** (2018 -)
  - ✳ → provide clever solutions if models are not available. E.g. Learn them…
  - ✳ Python into the control room
  - ✳ **Optimisers, ML**…on demand

The waves of automation @ CERN

The **current** automation efforts are based on three threads

- **Automation wave 1** (2006 -)
  - reduce complexity through models (LSA)
  - high level parameter control, sequencers, software interlock system, classic control algorithms in feedforward and feedback (SVD, COSE,...)
- **Automation wave 2** (2018 -)
  - → provide clever solutions if models are not available. E.g. Learn them...
  - Python into the control room
  - **Optimisers, ML**...on demand
- **Automation wave 3** (2021 -)
  - → close the loop
  - **frameworks** (Generic Optimisation Framework (GeOFF), Machine Learning Platform)
  - **auto**-launch correction, **auto**-resets, **auto**-analysis
  - → auto-pilots

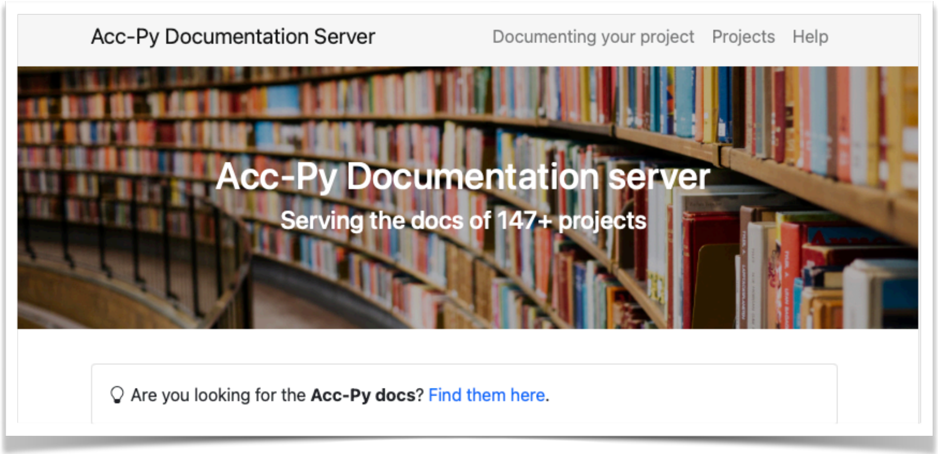
Acc-Py - Python for CERN accelerators

Python distribution fully integrated with control system:

- Equipment access: pyjapc → pyda, next: event building
- Access to settings and archived data
- GUI framework

Acc-Py package repository and centrally managed deployment

acc-py-repo.cern.ch



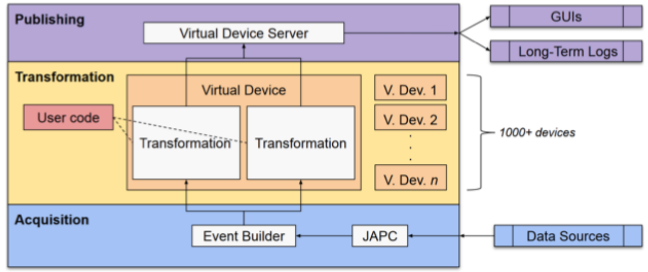
acc-py.web.cern.ch

UCAP

Unified Controls Acquisition and Processing (“Virtual Device Service”) → servers on-the-fly in JAVA or Python

→ Plug&Play framework for analysis (UCAP transformations)

→ continuously running in the background, triggered by “events "



Provides:

- tools for development, test and deployment of “transformations" (+actors)
- infrastructure to run transformations - no servers to buy; new: UCAP node with GPUs

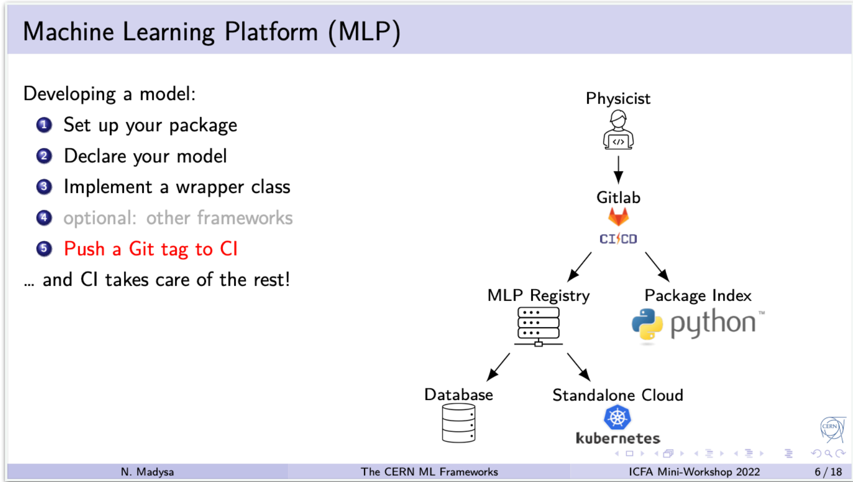
Key features:

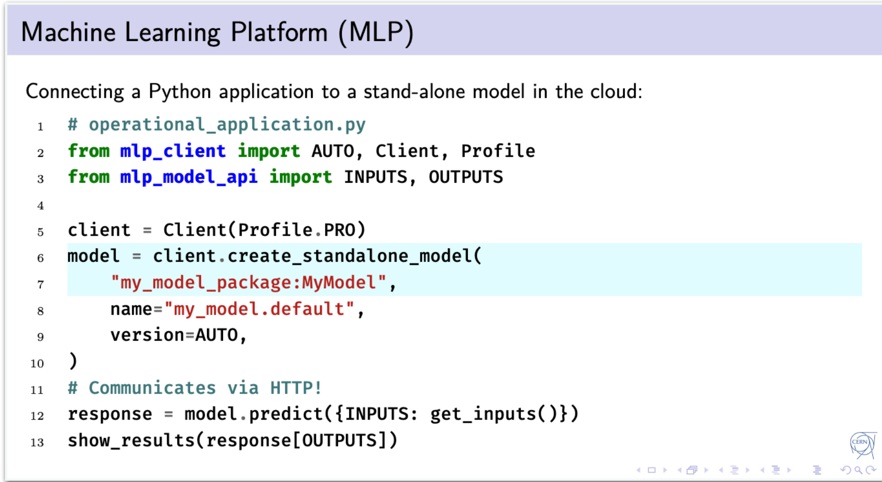
→ event building (e.g. buffered data,…)

→ others can subscribe to “transformation" result (i.e. applications, controllers,…)

Machine Learning Platform (MLP)

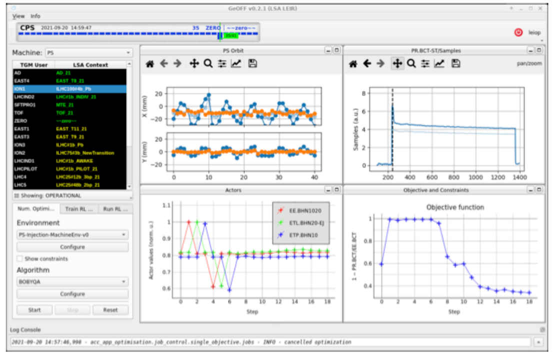
Storing and sharing trained ML models (e.g. ANNs) between users and applications of different languages. Stand-alone models hosted in the CERN cloud, ready for inference.





Successfully used for deploying AccGPT - our AI assistant: challenge of very large Llama2 model

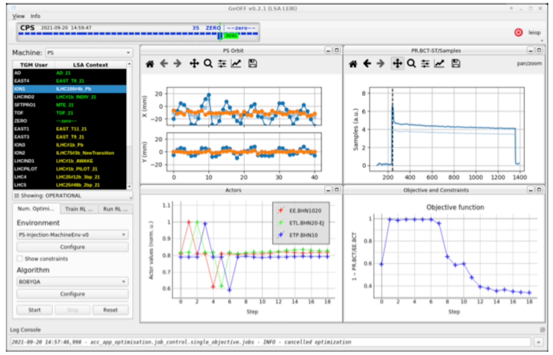
Optimisation Infrastructure



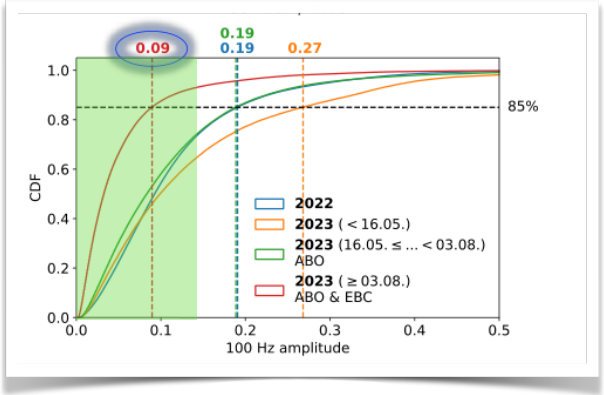
- **Generic Optimisation Framework GeOFF**
  - Manual scans and grid scans are inefficient for multi-parameter problems → optimisation algorithms
  - GeOFF = easy and flexible parameter optimisation in the control room
  - To date: > 20 parameter optimisation problems automated across complex

Optimisation Infrastructure

- **Generic Optimisation Framework GeOFF**
  - Manual scans and grid scans are inefficient for multi-parameter problems → optimisation algorithms
  - GeOFF = easy and flexible parameter optimisation in the control room
  - To date: > 20 parameter optimisation problems automated across complex

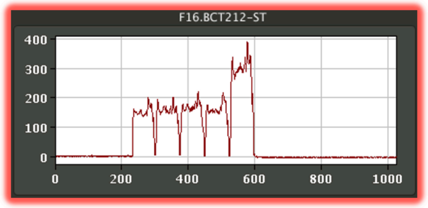


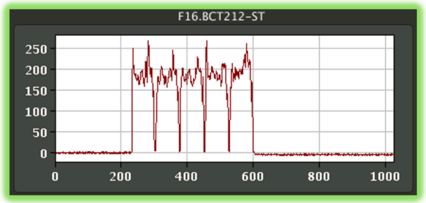
100 Hz content of NA spill with ABO and EBC



- **Optimisation framework for auto-pilots**
  - GeOFF on UCAP → acc-geoff4ucap released in summer 2023.
  - Operational: $n \times 50$ Hz control with Adaptive Bayesian Optimisation (ABO) for North Area spill with **GPUs on UCAP**
  - To come in **2024:**
    - automated PS2SPS trajectory steering
    - Multi-turn extraction (MTE) efficiency drift stabilisation

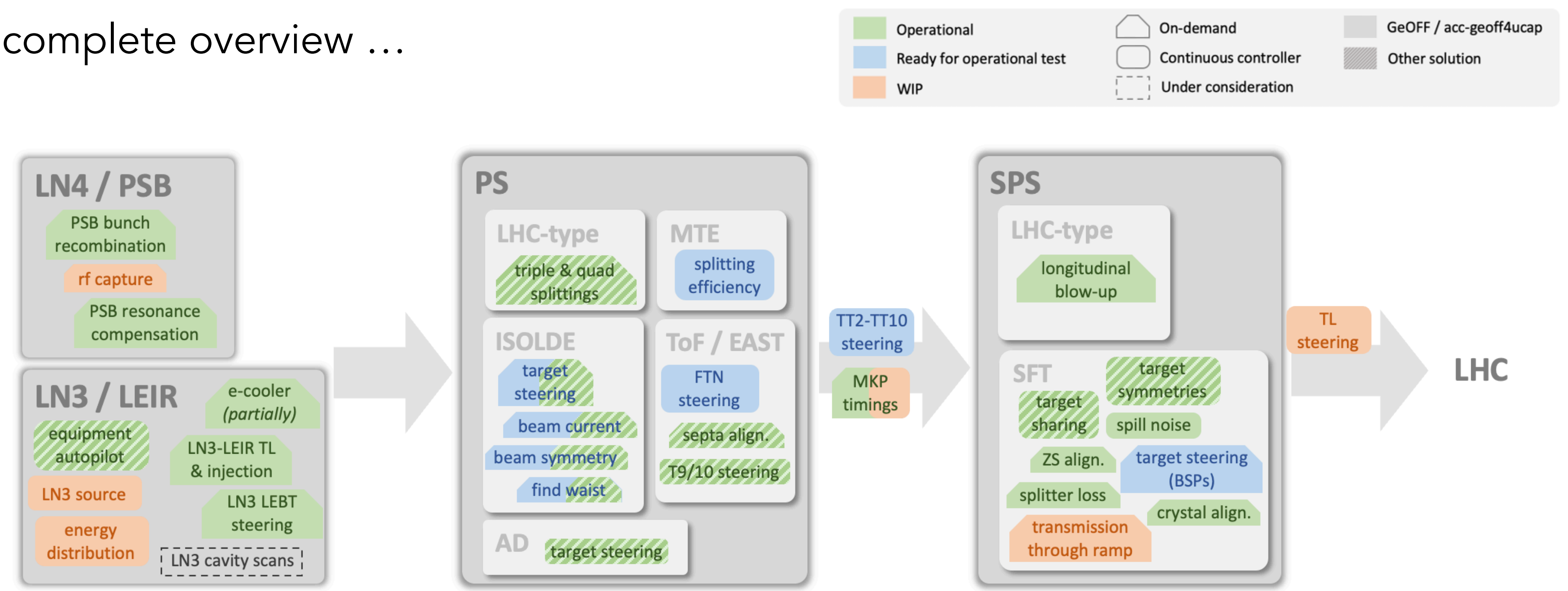
MTE island intensities to be equal with core





Auto-pilots, optimisers,...

An incomplete overview …



Courtesy M. Schenk

Status 2023: multiple auto-pilots / optimisers used operationally

Trends 2024: on-demand → continuous (UCAP) I some new auto-pilots

Until end of run 3: automation of all typical optimisation and continuous control problems

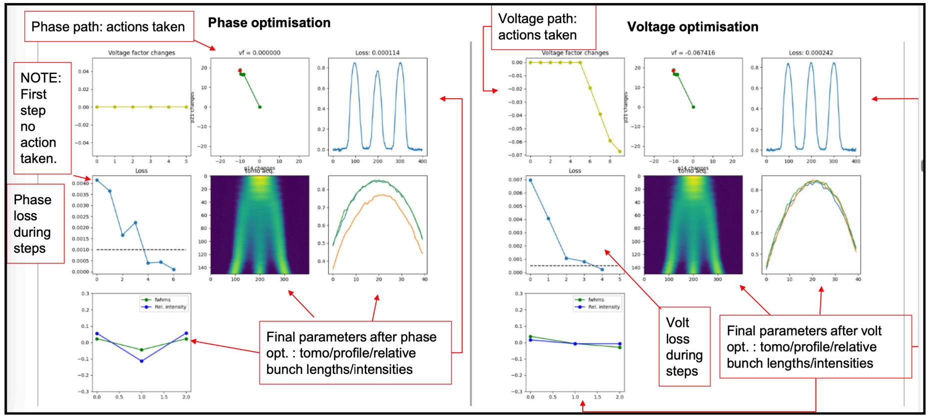
RL in the control room

RL agent to correct **RF phase and voltage** to produce uniform RF splitting in PS for LHC beams

★**Trained on simulation and successfully transferred to control room → fully operational**

★RL algorithm: Soft Actor-Critic (SAC); multi-agent algorithm using CNN to define initial set point

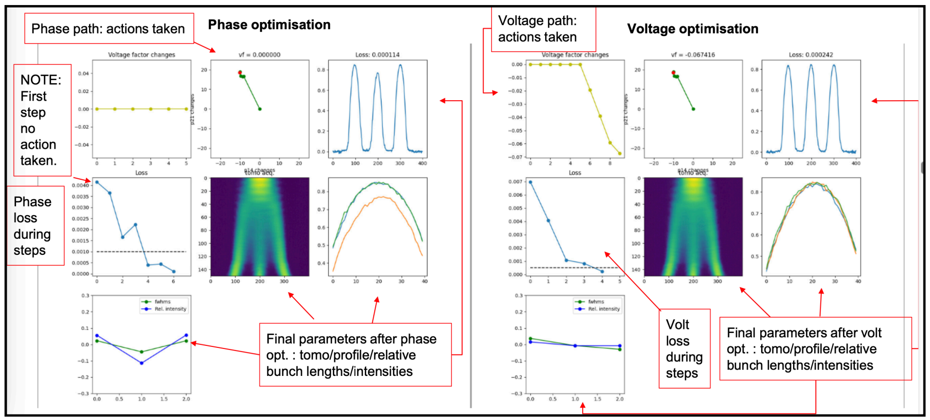
★Next step: from on-demand to continuous: → UCAP

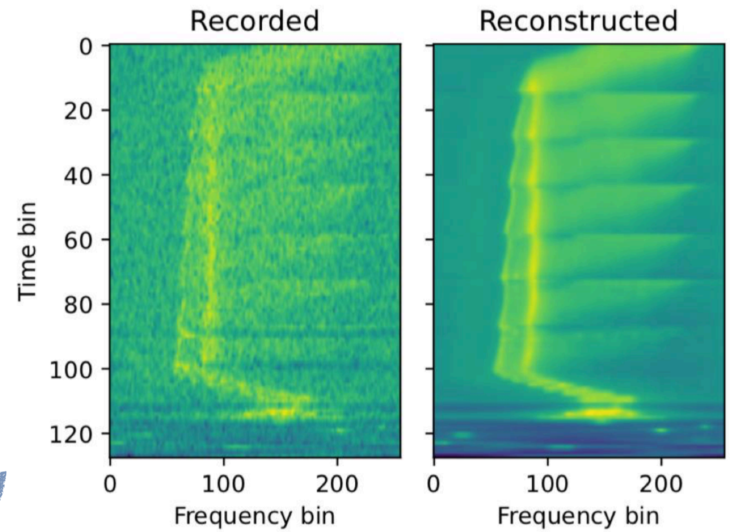


RL in the control room

RL agent to correct **RF phase and voltage** to produce uniform RF splitting in PS for LHC beams

- ★**Trained on simulation and successfully transferred to control room → fully operational**
- ★RL algorithm: Soft Actor-Critic (SAC); multi-agent algorithm using CNN to define initial set point
- ★Next step: from on-demand to continuous: → UCAP



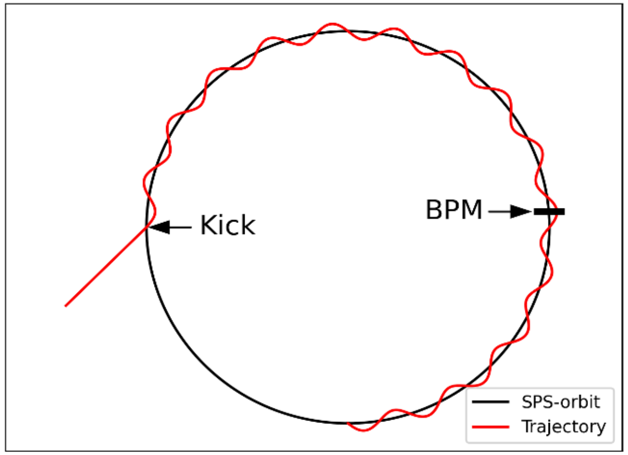


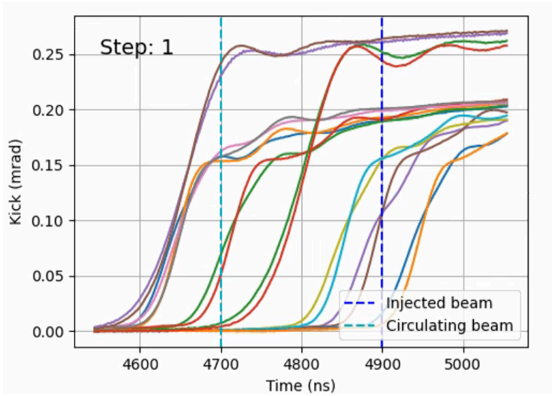
PhD ongoing for: control r**amping and debunching cavity in LINAC3** for optimal injection efficiency into LEIR, based on **Schottky spectrum. Trained on data-driven dynamics**

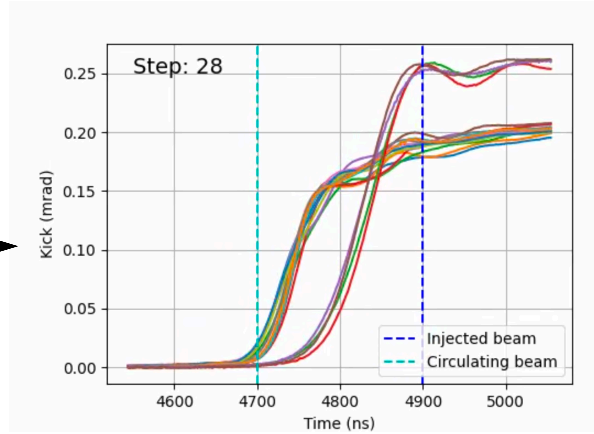
RL in the control room

Ready for transfer test: Adjusting the fine delays of SPS injection kicker with RL for LHC beams

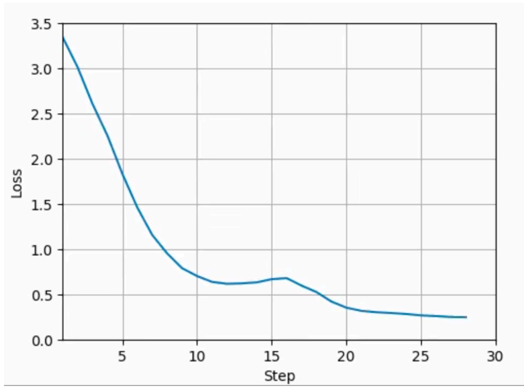
Trained on data-driven dynamics model: RL algorithm PPO







$$r = -\text{ loss } = -(x_{inj}^2 + x_{circ}^2 + (x_{inj} - x_{circ})^2)$$



Courtesy M. Remta, F. Velotti

Predicting magnetic hysteresis and eddy current effects

Potentially game-changing!

Time-series forecasting problem: need magnets to be measured on test bench
$[B_t, B_{t+1}, \ldots, B_{t+n-1}], [I_t, I_{t+1}, \ldots, I_{t+n+N}] \rightarrow [B_{t+n}, B_{t+n+1}, \ldots, B_{t+n+N}]$

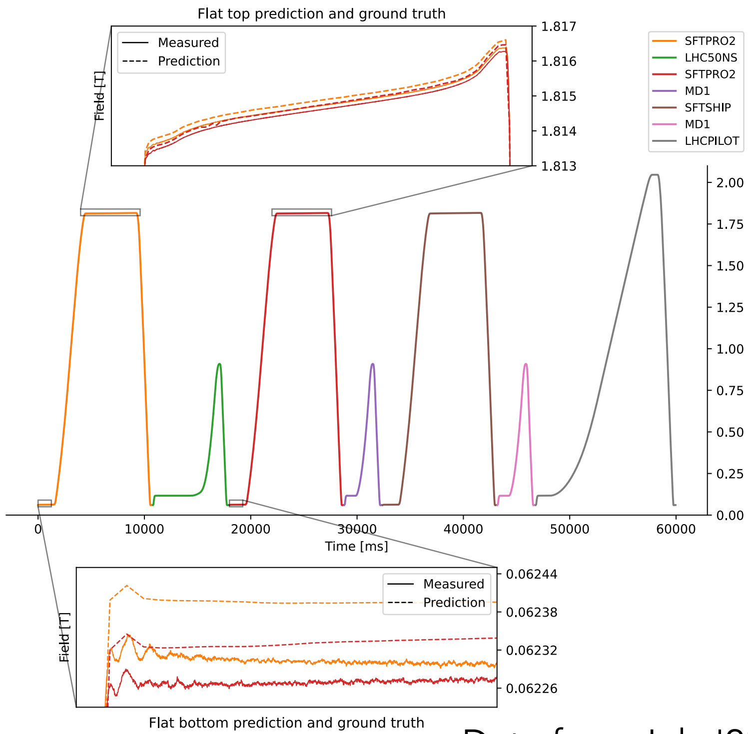
First operational experience:

- **feedforward correction** triggered before every cycle
- accuracy not sufficient yet

First results PhyLSTM for SPS main dipoles assuming
$\ddot{B} + g(B, \dot{B}) = \Gamma I(t)$, next: Transformers

Courtesy A. Lu

SPS main dipole field prediction vs measured, for fixed target cycles



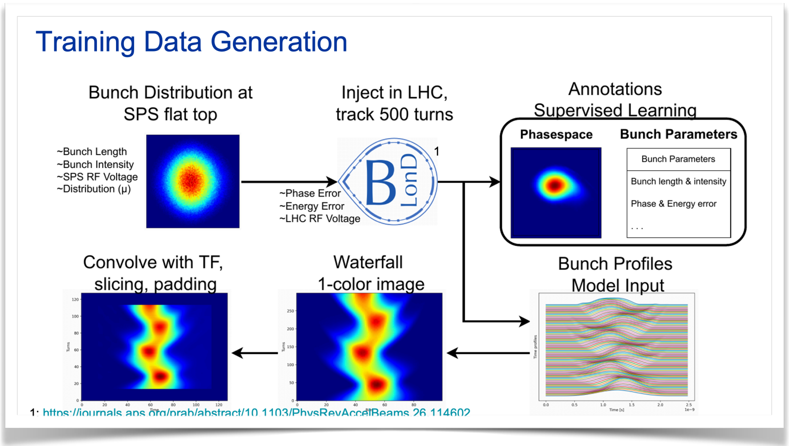
Data from July '23

Enhancing diagnostics - ML tomoscope in the LHC

Speeding up tomographic bunch-by-bunch reconstruction in the LHC: using Auto-encoder ensemble: without AI bunch-by-bunch **not** possible.

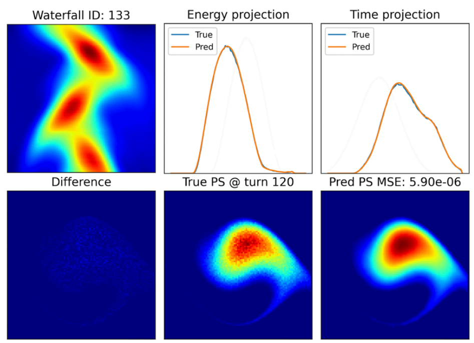
→ fully operational (on MLP). Used in the LHC control room to measure **injection errors**

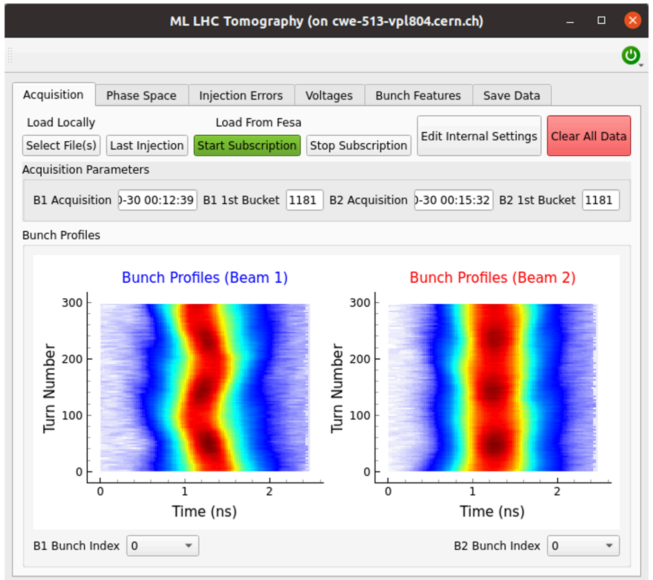
Trained with simulator



Tomoscope Evaluation

- MAE: 0.001 (1‰)
- Visually indistinguishable





Courtesy K. Iliakis, T. Argyropoulos

Conclusions

CERN is now heavily investing into AI/ML solutions in the control room with the **Efficient Particle Accelerators** project.

- → AI-ready infrastructure
- 5 year project to be ready for the HL-LHC era

Conclusions

CERN is now heavily investing into AI/ML solutions in the control room with the **Efficient Particle Accelerators** project.

- → AI-ready infrastructure
- 5 year project to be ready for the HL-LHC era

Main drivers for embracing AI solutions:

- Automation
- Improved - i.e. data-driven - modelling (e.g. hysteresis compensation)
- Speed up of computation: surrogate models, tomography,...

Conclusions

CERN is now heavily investing into AI/ML solutions in the control room with the **Efficient Particle Accelerators** project.

- → AI-ready infrastructure
- 5 year project to be ready for the HL-LHC era

Main drivers for embracing AI solutions:

- Automation
- Improved - i.e. data-driven - modelling (e.g. hysteresis compensation)
- Speed up of computation: surrogate models, tomography,…

To come still at scale:

- ML for fault analysis and anomaly detection for preventive maintenance