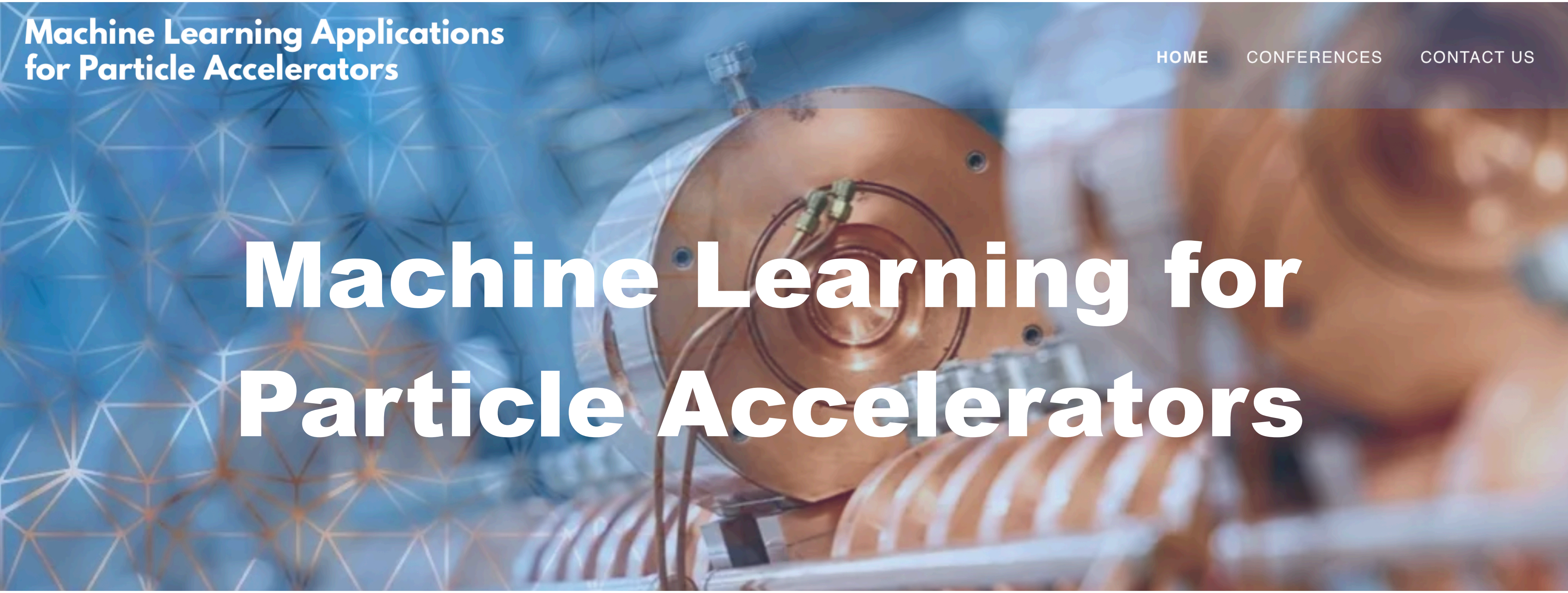

Machine Learning Applications
for Particle Accelerators
HOME
CONFERENCES
CONTACT US
Machine Learning for
Particle Accelerators

Nothing new under the sun...

CERN-ISR-CO/80-29

TOWARDS FULL AUTOMATION OF ACCELERATORS THROUGH COMPUTER CONTROL

by

J. Gamble, J.-Y. Hemery, D. Kemp, R. Keyser, J.-P. Koutchouk,

P. Martucci, L. Tausch, L. Vos

Nothing new under the sun...

CERN-ISR-CO/80-29

TOWARDS FULL AUTOMATION OF AC

J. Gamble, J.-Y. Hemery, D. Kemp, R. Keyser, J.-P. Koutchouk,
P. Martucci, L. Tausch, L. Vos

The second line of development encounters problems of a global nature. Once all the steps in the setting up of the accelerator have modules associated with them and the operators' principal task is to analyze graphical outputs and to handle error conditions, then full automation seems to be near. Unfortunately this is not the case. One of the most difficult aspects in progressing from partial automation to full automation is in quantifying the analysis of the various displays. For example, at the ISR the closed orbit can be presented to the operator who, after inspection decides whether or not to call the closed orbit correction module. This decision is based on the criteria from many different inputs other than the closed orbit display. The closed orbit may be acceptable (i.e. within recognized limits) but due to a background level slightly higher than normal in part of the ring, the closed orbit is viewed with a more critical eye and may be corrected slightly. These considerations represent one of the major problems of full automation.

Thus another important aspect of full automation is the ability to correlate the outputs of many different modules to arrive at a particular decision.

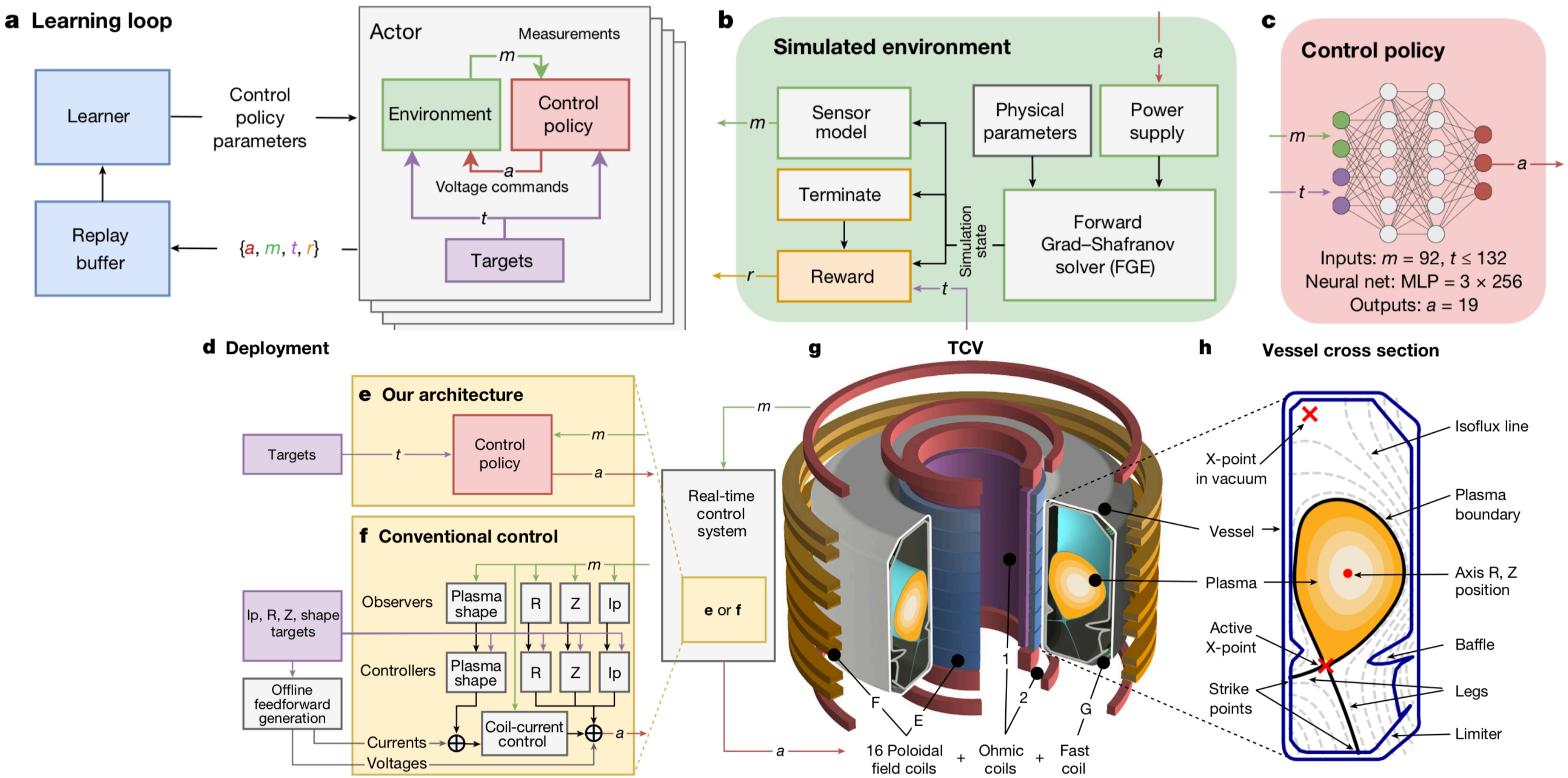
What can AI do for accelerators?

Something like this maybe…

Magnetic control of tokamak plasmas through deep reinforcement learning

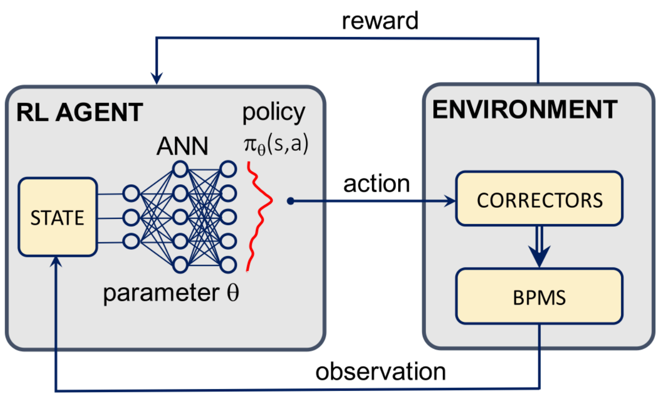
https://doi.org/10.1038/s41586-021-04301-9

Time-varying, non-linear, multi-variate control problem solved with deep Reinforcement Learning



Reinforcement Learning (RL)

Learn dynamics (once and for all) through trial-and-error, no exploration after training!



RL setup for trajectory steering

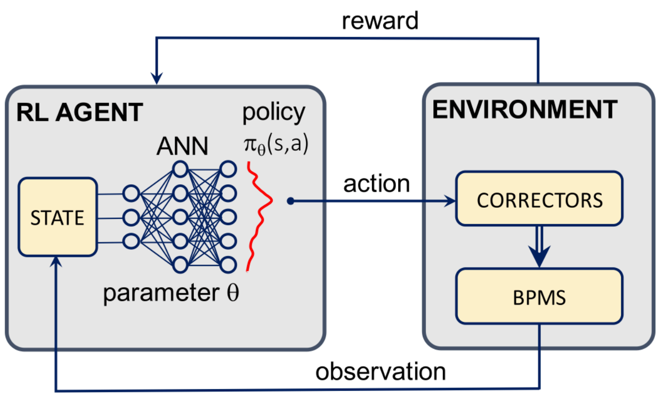
RL elegant (if not ideal) solution, but **online training** often not possible!

- Not sample-efficient enough
- Safety constraints

→ **RL (like MPC) needs to be built into accelerator design**.

Reinforcement Learning (RL)

Learn dynamics (once and for all) through trial-and-error, no exploration after training!



RL setup for trajectory steering

Systems to be built <u>for</u> RL:

→ fast executing (accurate) simulation / digital twin for training

→ instrumentation designed with control algorithm

RL elegant (if not ideal) solution, but **online training** often not possible!

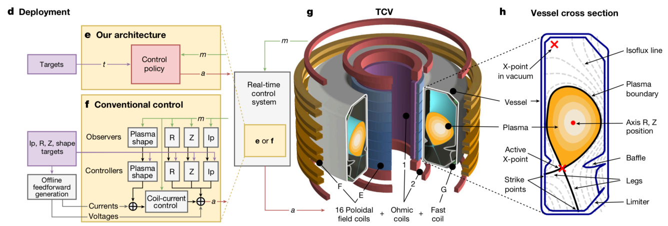
- Not sample-efficient enough
- Safety constraints

→ **RL (like MPC) needs to be built into accelerator design**.

Key ingredients...

Magnetic control of tokamak plasmas through deep reinforcement learning

https://doi.org/10.1038/s41586-021-04301-9



Accurate simulators

- Full or meta training in simulation, sim2real transfer or very sample-efficient adaptation

Adequate, non-destructive instrumentation as state information

- In the case of Tokamak control: 92 input state features (plus targets)

Optimised Reinforcement Learning algorithms → available

- Need to be easy to tune and some guarantees of convergence
- Sample-efficiency "less" strict requirement

From Tokamaks to Accelerators…

Fully autonomous control with RL for small low intensity linear accelerator with fast simulations: ARES @ DESY

Courtesy O. Stein, J. Kaiser, A. Eichler

ARES Our Testbed

Small research accelerator at DESY’s SINBAD facility

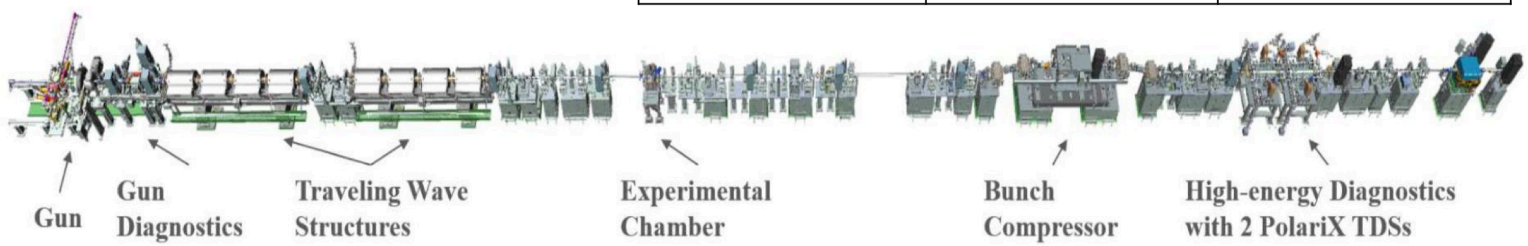
ARES a small (~60m) linear electron accelerator.

Goal: Characterize ultrashort e- bunches for applications (advanced & compact long. diag. and accelerating structures development, FLASH radiotherapy, etc.).

ARES is an easy accessible accelerator posing as a testbed for testing new developments.

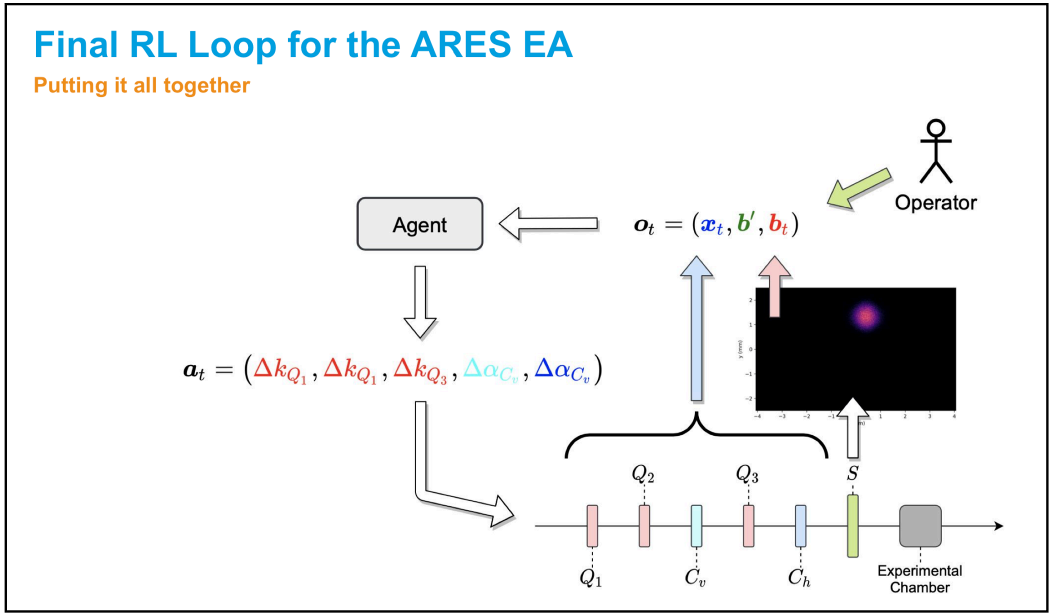
Frequent (almost weekly) beam times at ARES with close collaboration with the ARES team.

Properties	Target Values	Status
Charge	0.1 - 100 pC	0.1 – 50 pC
Momentum	50 -100 Mev/c	50 – 156 MeV/c
Momentum Spread	1,00E-04	1,00E-04
Transverse emittance	< 0.8 π.mm.mrad	≈ 0.4 π.mm.mrad
Duration	Sub-fs to ≈ 10 fs	≈ 40 fs



From Tokamaks to Accelerators...

Fully autonomous control with RL for small low intensity linear accelerator with fast simulations: ARES @ DESY



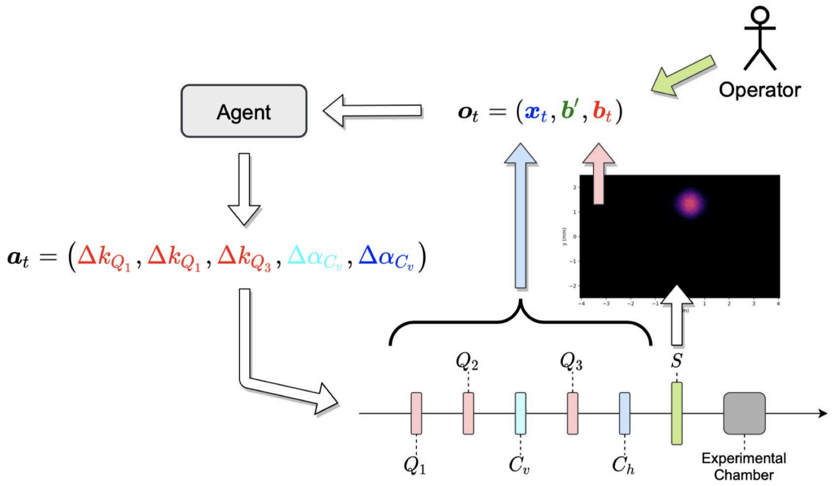
Courtesy O. Stein, J. Kaiser, A. Eichler

From Tokamaks to Accelerators...

Fully autonomous control with RL for small low intensity linear accelerator with fast simulations: ARES @ DESY

Final RL Loop for the ARES EA

Putting it all together

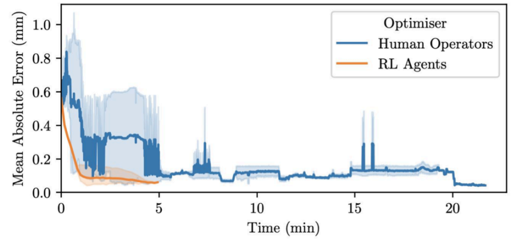


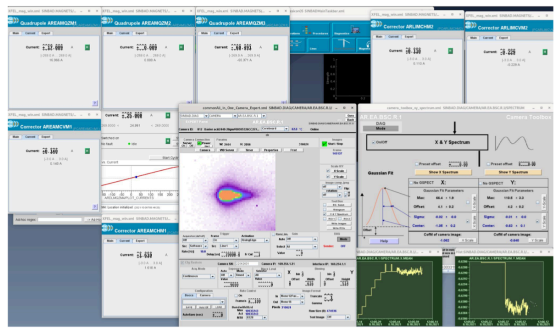
Courtesy O. Stein, J. Kaiser, A. Eichler

Human vs. Machine

Can John Connor beat Polished Donkey?

- Two experienced human operators were tasked to solve the same problem as three sibling RL agents.
- Initially RL agents converge slightly faster than operators. Operators take a long time to fine tune to a competitive result but eventually achieve better result than RL agents.
- RL agent speed severely limited by accelerator hardware speed in ARES EA. Likely much faster on other hardware.
- RL agent performance much more consistent than operator performance.
- Both follow sensible strategies but humans simplify (use only two quadrupoles and focus on one parameter at a time) as opposed to RL agents who are agnostic to simplicity of solutions.





Operators took time ahead of timing them to set up well, though.

One step further… "Talk to the accelerators"

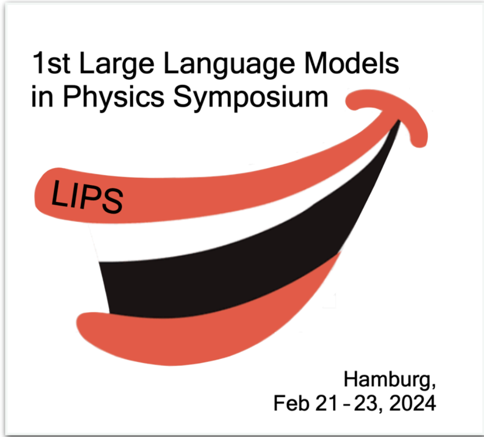
"Conversational Tuning" for transverse beam parameter tuning at ARES

Objective: $\text{objective} = |\mu_x - \mu'_x| + |\mu_y - \mu'_y| + |\sigma_x - \sigma'_x| + |\sigma_y - \sigma'_y|$

Courtesy J. Kaiser et al

```
Human: Now you will help me minimise a function with five input variables Q1, Q2, CV, Q3 and
CH. I have some (Q1, Q2, CV, Q3, CH) pairs and the corresponding function values at those
points. The samples are arranged in descending order based on their function values, where
lower values are better.

Inputs:
```json
{
  "Q1": -13.50,
  "Q2": -9.00,
  "CV": -3.00,
```



Prompt engineering very important.

```
  "Q3": -9.00,
  "CH": -6.00
}
```
Objective value = 2.37

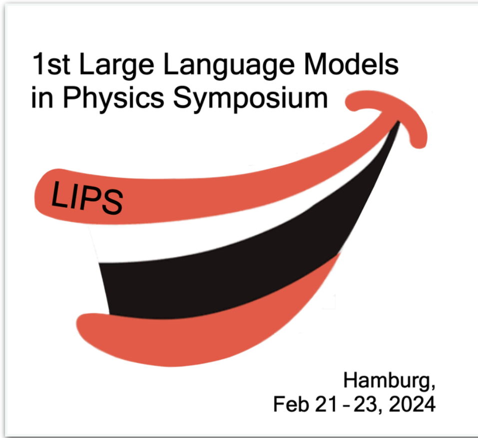
Inputs:
```json
{
  "Q1": -13.25,
  "Q2": -8.85,
  "CV": -2.80,
  "Q3": -8.90,
  "CH": -5.70
}
```
Objective value = 2.28

Give me a new sample (Q1, Q2, CV, Q3, CH) that is different from all pairs above, and has a
function value lower than any of the above.

The output should be a markdown code snippet formatted in the following schema, including
the leading and trailing "```json" and "```":

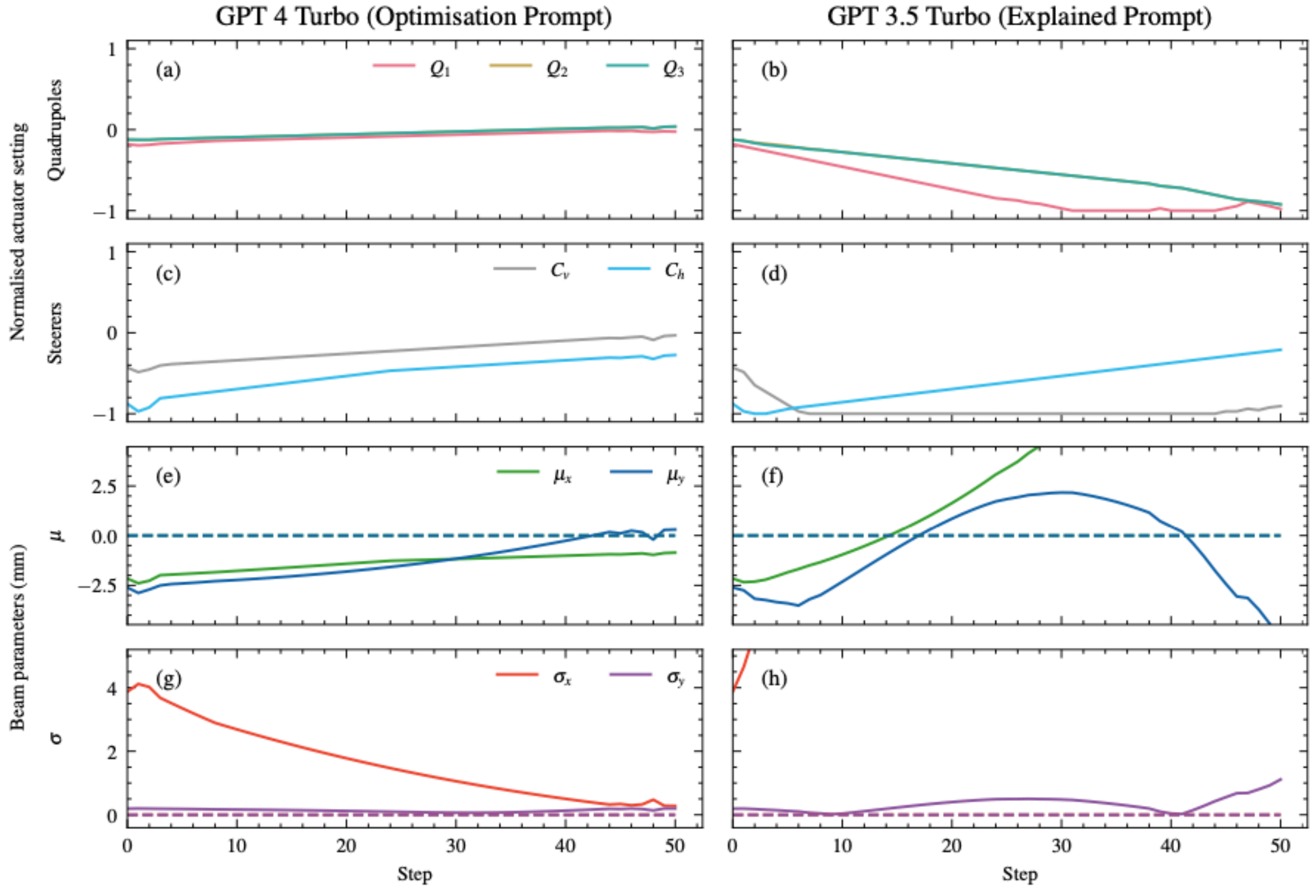
```json
{
  "Q1": float // First input
  "Q2": float // Second input
  "CV": float // Third input
  "Q3": float // Fourth input
  "CH": float // Fifth input
}
```
```

One step further… “Talk to the accelerators”

"Conversational Tuning” for transverse beam parameter tuning at ARES

Prompting scheme: optimisation prompt

Ideally expect answer like this:

````
```json
{
  "Q1":  -14.30,
  "Q2":  -9.70,
  "CV":  -2.55,
  "Q3":  -8.10,
  "CH":  -5.21
}
```
I suggest decreasing Q1 slightly to bring down the horizontal beam position, while keeping
the other quadrupole magnets at their previous values to maintain the vertical beam position
and focusing.  I also kept the steering magnet settings close to their last values for
smoothness.
````

Courtesy J. Kaiser

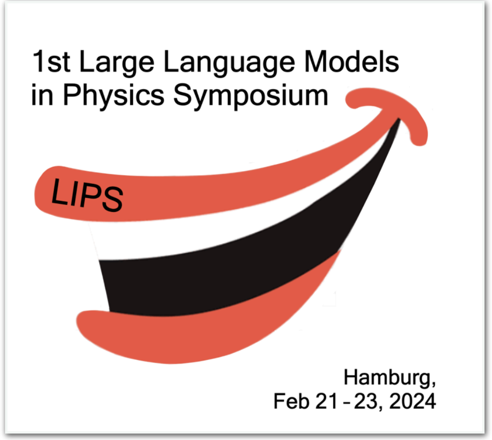


One step further… “Talk to the accelerators”

"Conversational Tuning" for transverse beam parameter tuning at ARES



Courtesy J. Kaiser



General AI assistants in the control room?

LLMs to use tools…

Using ReACT prompting scheme = Reasoning + Actions: chain-of-thought prompting + information injection via actions

- Needs large enough models (+ hardware resources)

General AI Assistant for Intelligent Accelerator Operations (GAIA)

Link to talk

F. Mayet, DESY, 2024

General AI assistants in the control room?

Using ReACT prompting scheme = Reasoning + Actions: chain-of-thought prompting + information injection via actions

- Needs large enough models (+ hardware resources)

```
Terminal
fmayet@desymac42 ~ % ./gaia —verbose True

Welcome to GAIA (served on max-wng055)

What can I do for you?
Can you summarize the last operations meeting?

> Entering new AgentExecutor chain…
Thought: To answer this question, I need to use the "Operation Meeting Summarizer" tool and
provide the date of the meeting as input. I will ask the user for the date if it has not
been provided.

Action:
```
{
  "action": "User Input Tool",
  "action_input": {
    "question": "What is the date of the operation meeting you want me to summarize?"
  }
}
```
What is the date of the operation meeting you want me to summarize?
29.01.2024
```

```
Terminal
Observation: 29.01.2024
Thought: Based on the user input, I can now use the "Operation Meeting Summarizer" tool to
generate a summary of the operations meeting that took place on the given date.

Action:
```json
{
  "action": "Operation Meeting Summarizer",
  "action_input": {
    "date": "2024-01-29"
  }
}
```

Observation: The ARES Operation Meeting summary from January 29, 2024 covers […]

> Finished chain.
```

F. Mayet, DESY, 2024

What can AI do for accelerators? The vision..

Autonomous accelerators

Key words: optimal control and optimisation, anomaly detection and prescriptive maintenance, surrogate modelling, differentiable simulations, virtual diagnostics,…

Optimised accelerator design

Key words: fast-executing simulations for optimisation algorithms, differentiable simulations,…

Generic AI for efficient research and development

Key words: AI assistants for code development, knowledge retrievable,…

Future accelerators = AI-ready accelerators

Input from the FCC operational model discussion:

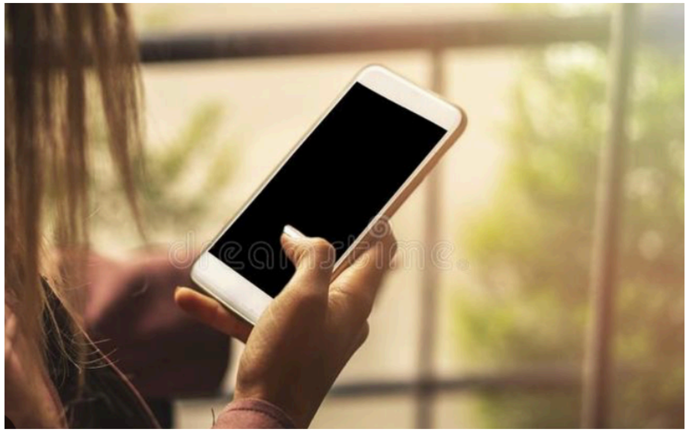
The ***business-as-usual*** solution: FCC just larger LHC

- Brute force scale-up → using helicopters to reduce intervention times, more people, more sites,...
- (Financially excluded, luckily)

The ***elegant*** solution: FCC to be run like a space telescope.

- Reinvent exploitation paradigm: hierarchical autonomous systems
- AI is key technology
- Management's preferred option

The (obvious) new equipment paradigm



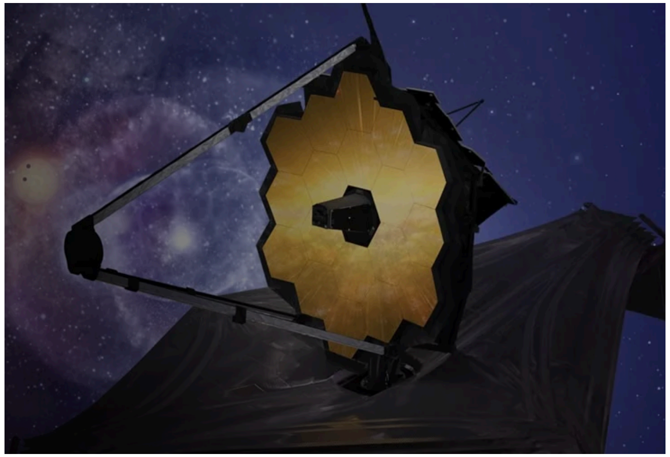
Think banking apps, heating systems,….
All digital, all remote controllable/ analysable

Simulations will be key.
Fast-executing, differentiable.
Digital twins…

Exploit **automation** at every level.

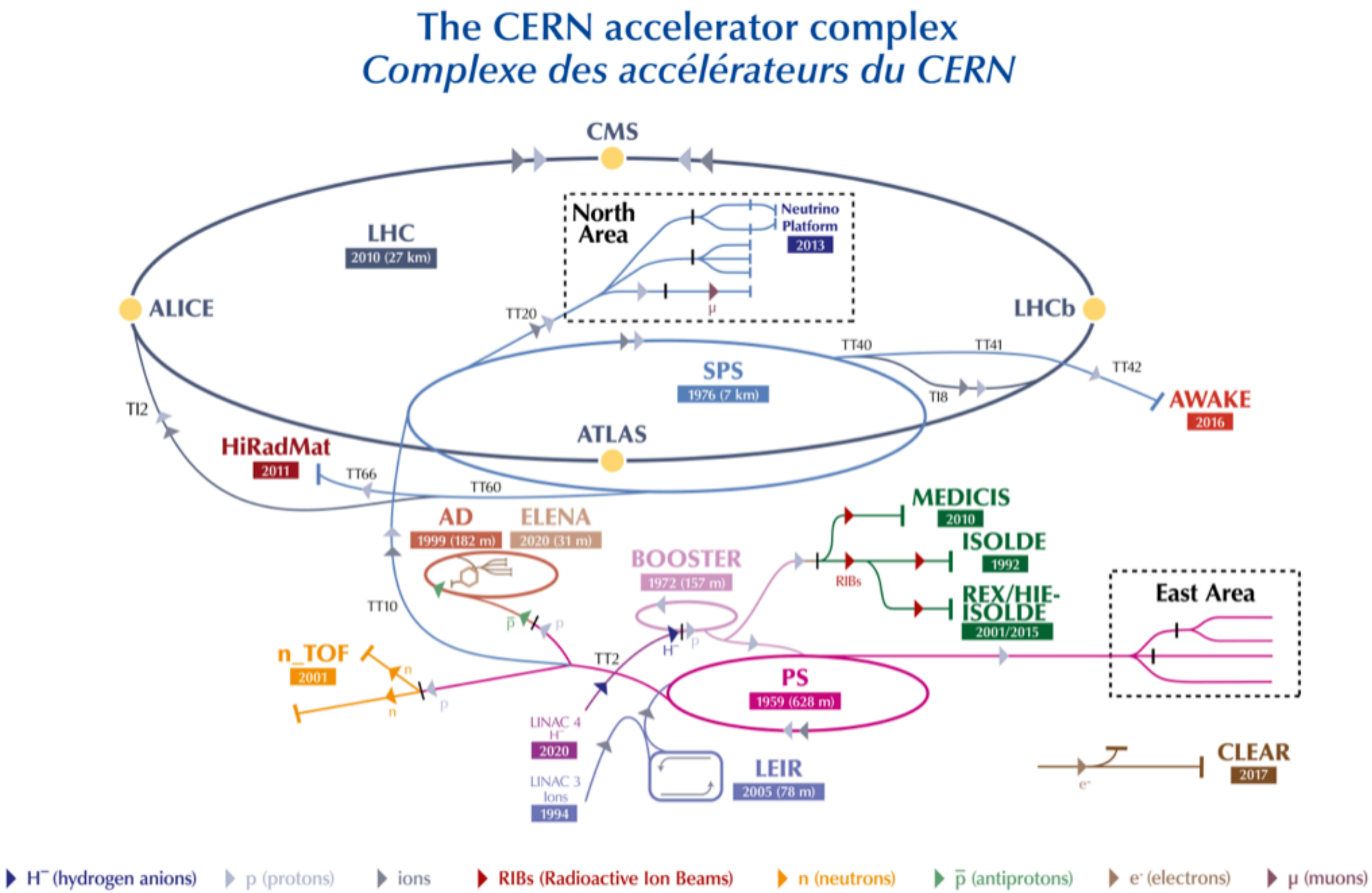
Automation across systems.
Automation within given system.
→ different players to implement automation

All equipment designed with automation in mind:
auto-configure, auto-stabilize, auto-analyse, auto-recover,…

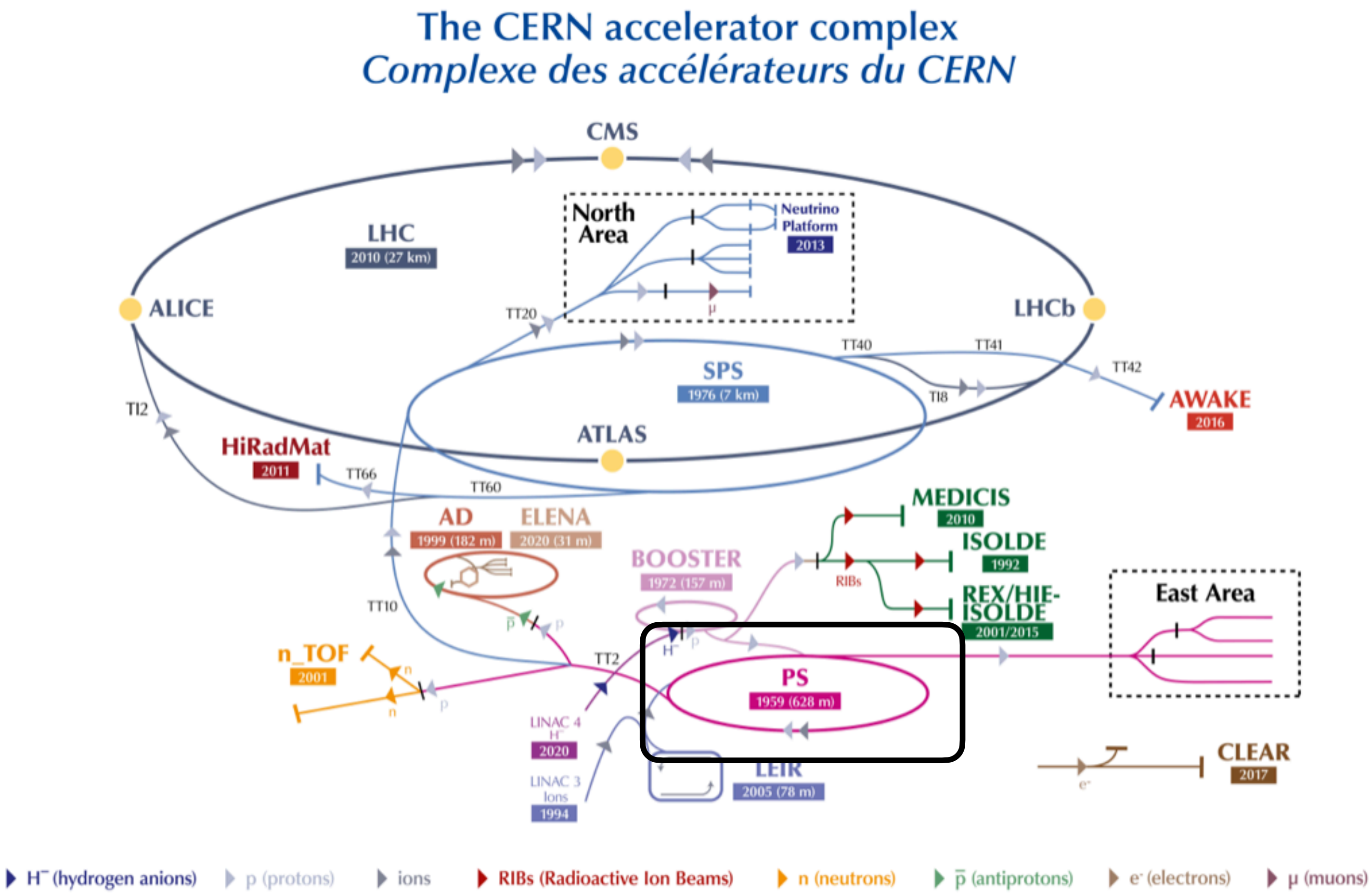


You cannot go there to fix it…
Redundancy, robotics,…

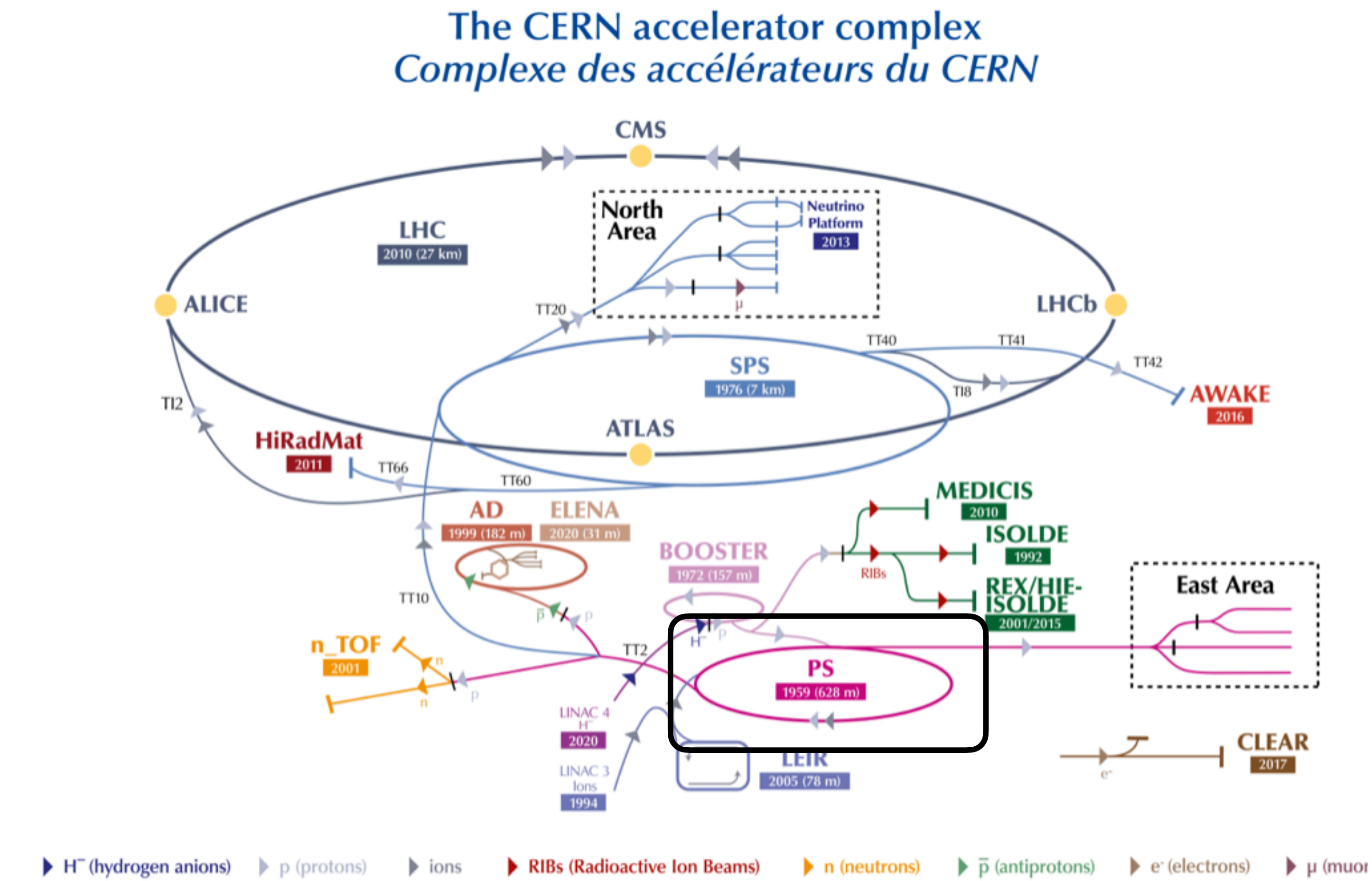
Retrofit the old stuff?



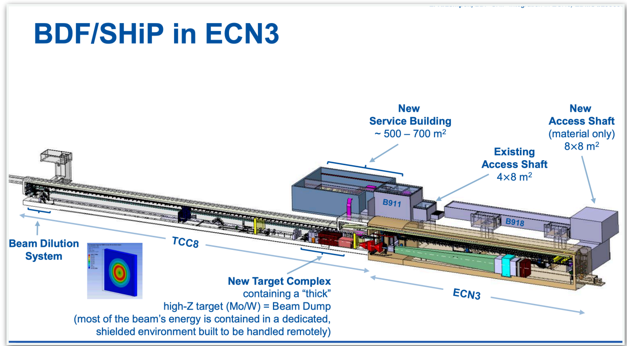
Retrofit the old stuff?



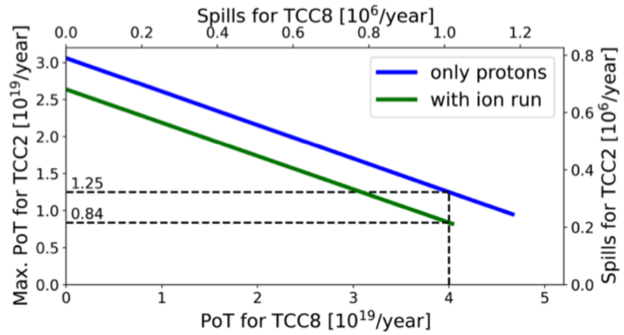
Retrofit the old stuff?



SHiP - Search for Hidden Particles
Start operation ~2030



Many challenges: losses, proton sharing,…



Efficient Particle Accelerators (EPA) project @ CERN

→ automating accelerator exploitation - AI and classical means

Approved in autumn 2023 after pre-study in Efficiency Think Tank (ETT)

9 work packages: ETT recommendations and controls infrastructure evolution.

WP1 Dynamic Beam Scheduling	WP2 Automated LHC Filling	WP3 Automated Parameter Control & Optimisation
WP4 Hysteresis Compensation	WP5 Next Generation Sequencer	WP6 Efficient Settings Management
WP7 Automated Equipment Testing	WP8 Automate Equipment	WP9 Data Processing Framework

EPA goals

Focus is on **automation** → to increase efficiency, reproducibility, flexibility and performance

WP1 **Dynamic Beam Scheduling**

- Automatically and dynamically schedule beams

WP2 **Automated LHC Filling**

- Automate and standardise LHC beam preparation and filling; reduce impact on fixed target users and LHC turn-around time

WP3 **Automated Parameter Control and Optimisation:**

- Automate parameter optimisation, automatically contain drifts

WP4 **Hysteresis Compensation:**

- Deterministic field control, decouple cycles

WP7 **Automated Equipment Testing:**

- AccTesting for “all" equipment for injectors and LHC, fully automated Hardware Commissioning

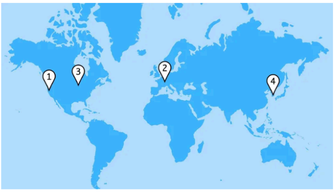
WP8 **Automate Equipment:**

- Automatic equipment setup; automate fault analysis, recovery; towards preventive maintenance

What are the trends in the community?

Machine Learning Applications for Particle Accelerators

HOME CONFERENCES CONTACT US



4TH MACHINE LEARNING APPLICATIONS FOR PARTICLE ACCELERATORS (2024), GYEONGJU, SOUTH KOREA. HOSTED BY PAL

3RD ICFA BEAM DYNAMICS MINI-WORKSHOP ON MACHINE LEARNING APPLICATIONS FOR PARTICLE ACCELERATORS (2022), CHICAGO, USA. HOSTED BY BNL.

2ND ICFA MINI-WORKSHOP ON MACHINE LEARNING FOR CHARGED PARTICLE ACCELERATORS (2019), VILLIGEN, SWITZERLAND. HOSTED BY PSI.

1ST MACHINE LEARNING APPLICATIONS FOR PARTICLE ACCELERATORS (2018), MENLO PARK, USA. HOSTED BY SLAC.

Workshop 2025 will be at CERN!

Topics of ML workshop '24

We are pleased to announce the **4th ICFA Beam Dynamics Mini-Workshop on Machine Learning for Particle Accelerators** will be held in ***Gyeongju, South Korea***. The goal of this workshop is to help build a world-wide community of researchers interested in applying machine learning techniques to particle accelerators.

The workshop will consist of six topics:

1. Analysis & Diagnostics
2. Anomaly Detection / Failure Prediction
3. Infrastructure / Deployment Workflows
4. Optimization & Control
5. Modeling Approaches
6. Lessons Learned

Tutorials:

1. Reinforcement Learning
2. Model Adaptation / Up-keep
3. Transformers for Timeseries Prediction

Talks will include both accelerator physicists and computer scientists. This workshop has the following goals:

- Collect and unify the community's understanding of the relevant state-of-the-art ML techniques.
- Provide a simple tutorial of machine learning for accelerator physicists and engineers.
- Seed collaborations between laboratories, academia, and industry.

Please contact the organizers if you are interested in attending.

Accelerator ML community - Trends...

Focus shifting slowly from R&D to AI at scale with full life cycle management.

→ Infrastructure/Deployment Workflows (MLOps) one of the longest sessions at last workshop

→ Discussion about standards: e.g. optimisation problem definition standards

→ Non-trivial life cycle management questions becoming important: "continual learning"

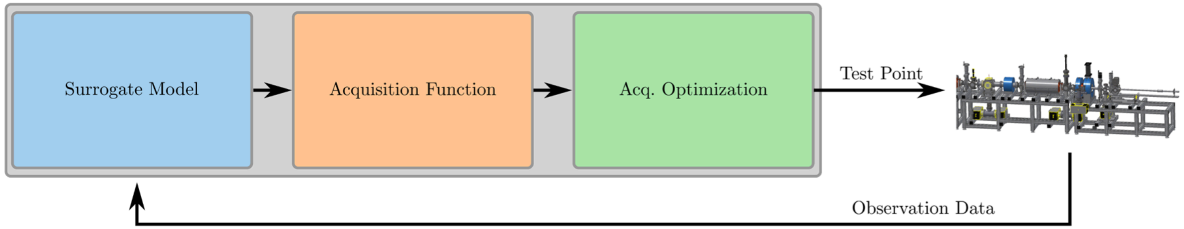
- Full tutorial about it.

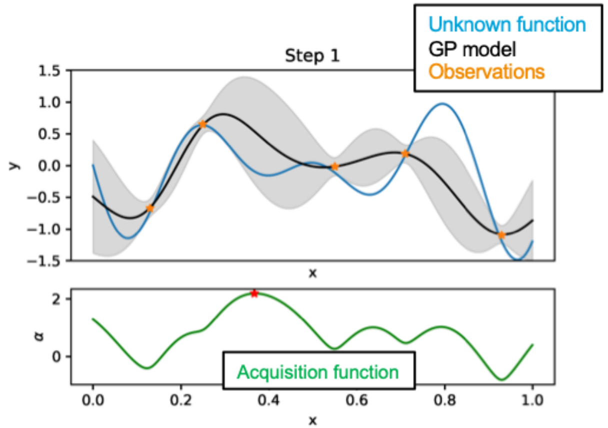
The big new theme: LLMs...PACuna, Logbook search, AI assistants in the control room

Progress on all fronts...

Bayesian Optimisation (BO)

Majority of talks in "optimisation and control" session about BO





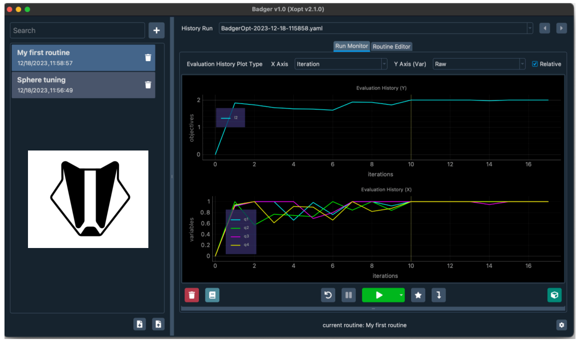
Recent publication of review paper:
Bayesian optimisation algorithms for accelerator physics
Phys. Rev. Accel. Beams 27, 084801

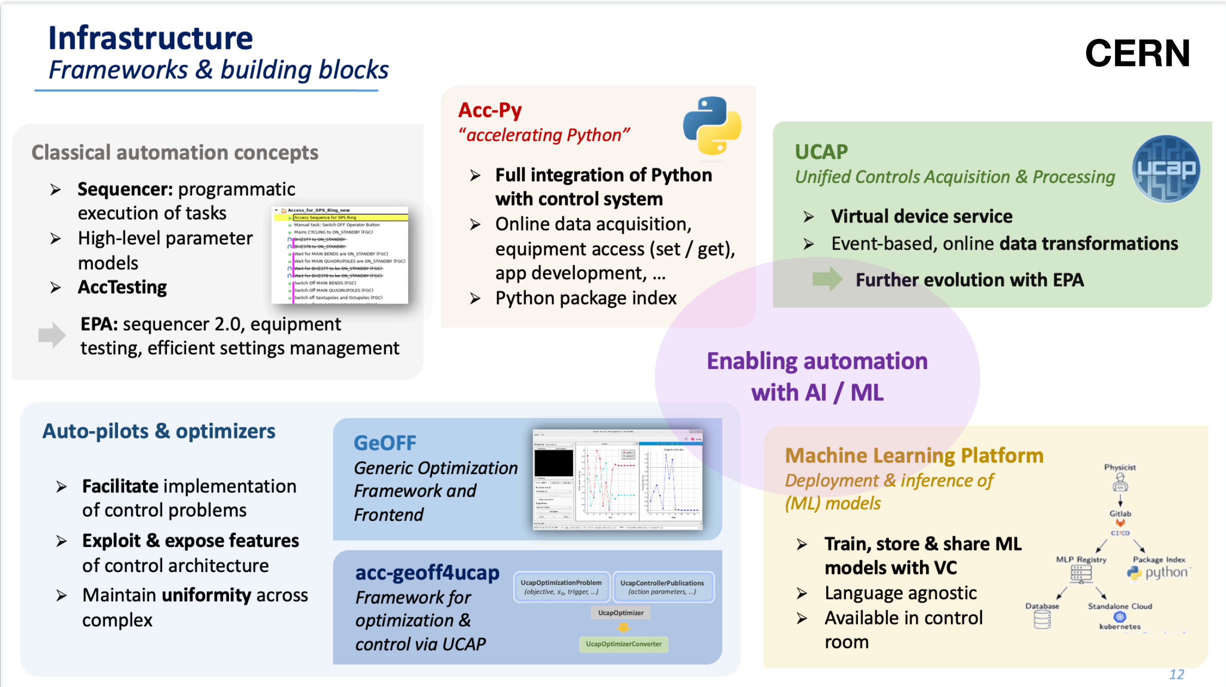
State-of-the-art BO:

- Model-based priors for various applications
- Multi-fidelity BO for laser plasma accelerators
- SafeOpt to include safety constraints - faster convergence with ModSafeOpt
- Information-based Bayesian Optimisation with virtual objectives
- …

Common tools and frameworks were key

SLAC





State-of-the-art BO with **BoTorch**

- GPU accelerated
- versatile
- fully integrated with PyTorch, GPyTorch

BoTorch
BAYESIAN OPTIMIZATION IN PYTORCH

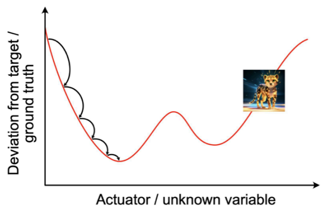
Differentiable simulation codes

Optimisation algorithms work best and are most sample-efficient with gradient information of the objective function.

Gradient-based Tuning

Transverse beam tuning at ARES

- Tune magnet settings or lattice parameters using the **gradient of the beam dynamics model** computed through **automatic differentiation**.
- Seamless **integration with PyTorch** tools tuning neural networks.
- Becomes very useful for **high-dimensional tuning tasks** (see neural network training).



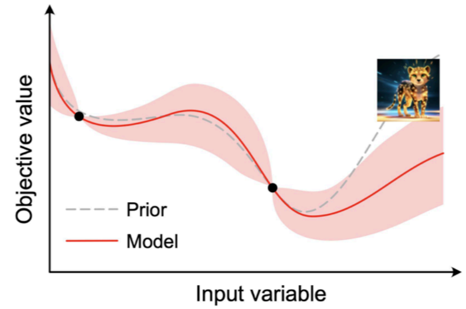

```
ares_ea.AREAMQZM1.k1 = nn.Parameter(0.0)
ares_ea.AREAMQZM2.k1 = nn.Parameter(0.0)
ares_ea.AREAMCVM1.angle = nn.Parameter(0.0)
ares_ea.AREAMQZM3.k1 = nn.Parameter(0.0)
ares_ea.AREAMCHM1.angle = nn.Parameter(0.0)

optimizer = Adam(ares_ea.parameters())

for _ in range(42):
    ougoing = ares_ea.track(incoming)
    loss = loss_fn(outgoing)

    loss.backward()
    optimizer.step()
    optimizer.zero_grad()
```

- A physics-informed prior can help **improve the performance of BO** by preventing over-exploitation.
- Cheetah's differentiability allows **efficient acquisition function optimisation** using gradient descent methods in modern BO packages like BoTorch.
- Has well-defined behaviour and **does not need data** to train like neural network priors.
- Can be used in **combination with gradient-based system identification** to overcome model inaccuracies.

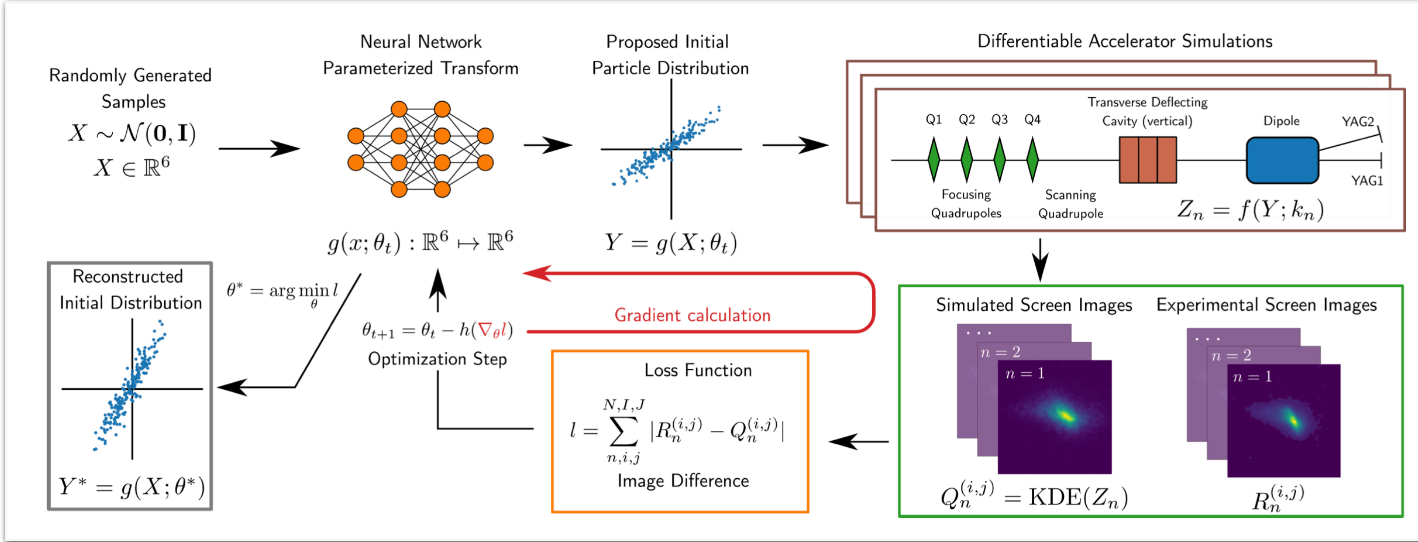


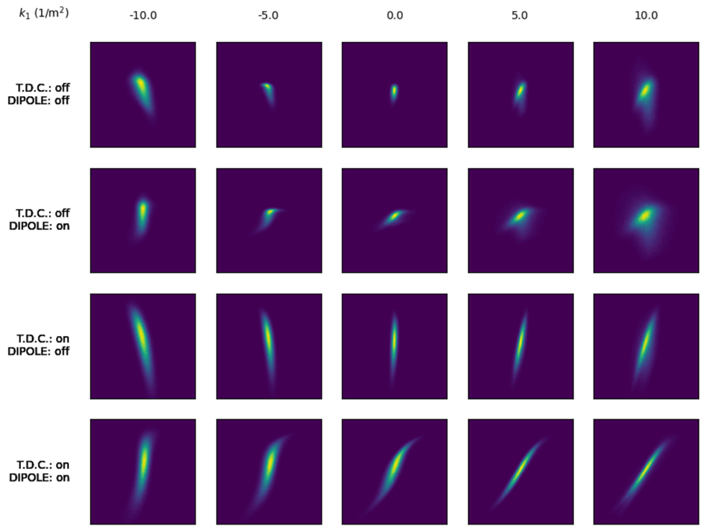
If you have differentiable codes....

Example: generative phase-space reconstruction in 6D

These measurements are normally rarely done, too time consuming…

E.g. Spallation Neutron Source (SNS): 5×10^6 measurements over 36 h





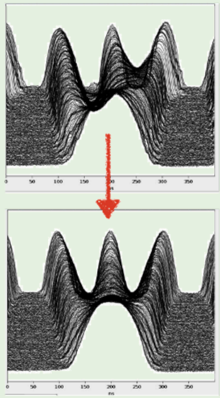
Courtesy R. Roussel et al arXiv:2404.10853

Tested at Argonne Wakefield Accelerator (AWA): **only 20 measurements** for full 6 D reconstruction.

RL4AA - workshop

Pushing the frontiers of RL for accelerators → autonomous accelerators.

RL @ CERN - all trained on simulation or surrogate



PS

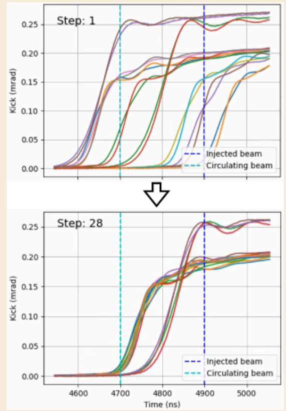
- Correct RF **phase & voltage** for **uniform bunch splitting** (LHC beams)
- Successful **sim2real** & fully **operational**
- **Multi-agent** (SAC) & **CNN** for initial guess
- **Next: continuous** controller (UCAP)

A. Lasheen, J. Wulff

PS to SPS

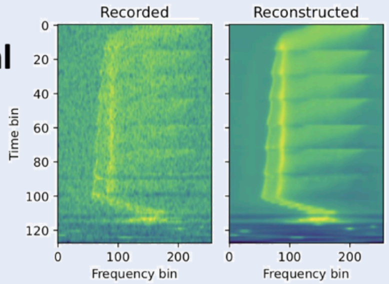
- Adjust **fine delays** of SPS **injection kicker**
- RL agent (PPO) trained on **data-driven dynamics model**
- Ready for **sim2real test**

M. Remta, F. Velotti

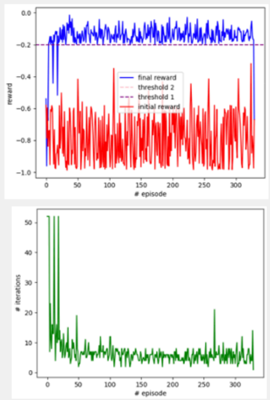


LINAC3 / LEIR

- **PhD project** *(B. Rodriguez)*: control LINAC3 cavities for **optimal injection efficiency** into LEIR
- RL state based on **VAE-encoded Schottky** spectra
- Agent trained on **data-driven dynamics model**



V. Kain, N. Madysa



SPS

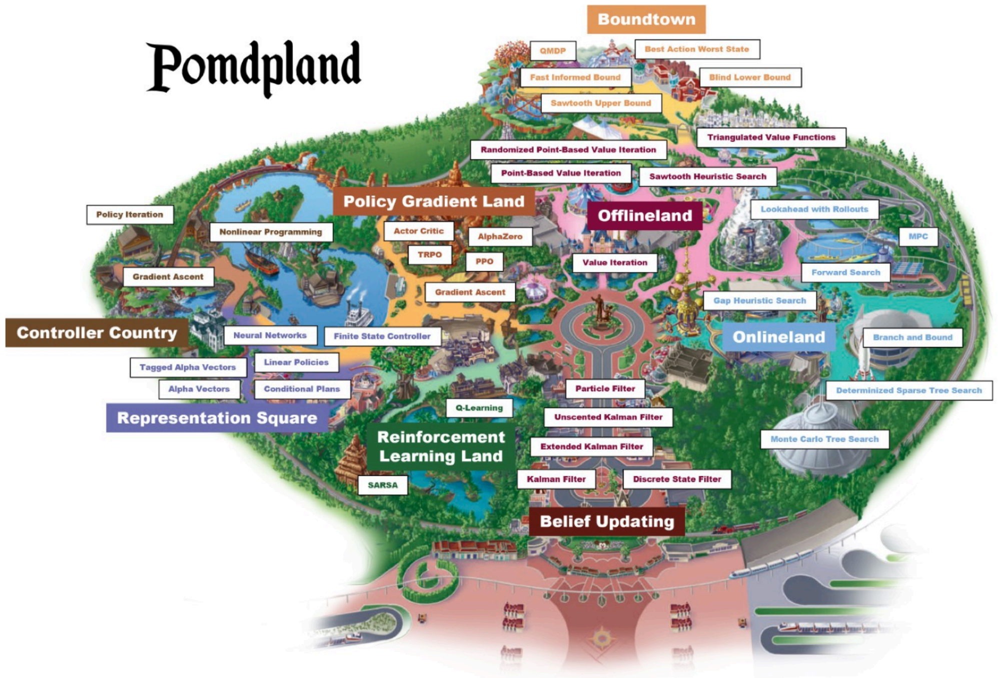
- **Steer DC beams** in TT20 TL using **split-foil** secondary emission **monitors**
- Works well in simulations, **with noise and varying emittances**
- Ready for **sim2real test**

N. Bruchon, V. Kain

Courtesy M. Schenk

Advanced RL concepts

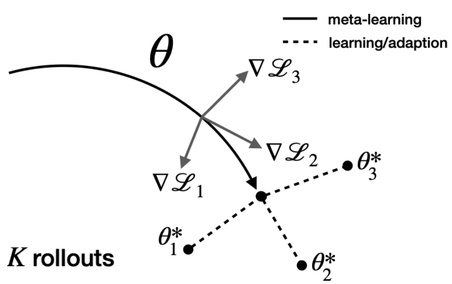
How to deal with time-varying systems, partially observable systems (POMDP).



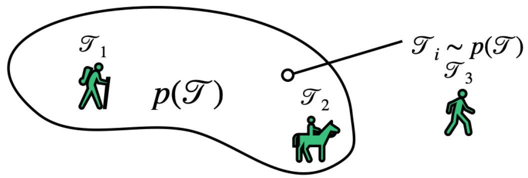
From Mykel Kochenderfer

META-RL i.e. MAML

Require $p(\mathcal{T})$: distribution over tasks
Require α, β: step size hyper-parameters
1. randomly initialise θ
2. **while** not done **do**
3. sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4. **for each** $\mathcal{T}_i$ **do**
5. Sample with policy θ: $\mathcal{D}^{tr}_{\mathcal{T}_i} \sim \mathcal{D}_{\mathcal{T}_i}$
6. Evaluate $\nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{tr}_{\mathcal{T}_i})$ with respect to K rollouts
7. Compute adapted parameters with gradient descent: $\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{tr}_{\mathcal{T}_i})$
8. Sample with new policy ϕ_i: $\mathcal{D}^{test}_{\mathcal{T}_i} \sim \mathcal{D}_{\mathcal{T}_i}$
9. Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}(\phi_i, \mathcal{D}^{test}_{\mathcal{T}_i})$



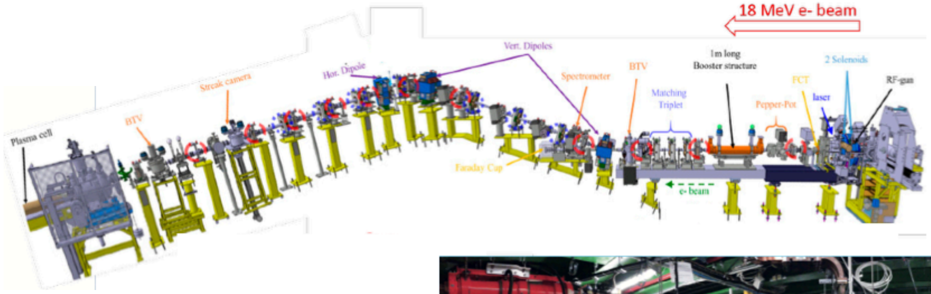
$$\mathcal{L}_{\mathcal{T}_i}(\theta) = -\mathbb{E}_{s_t, a_t \sim \pi_\theta, \mathcal{T}_i}\left[\sum_{t=1}^{H} R_i(s_t, a_t)\right]$$

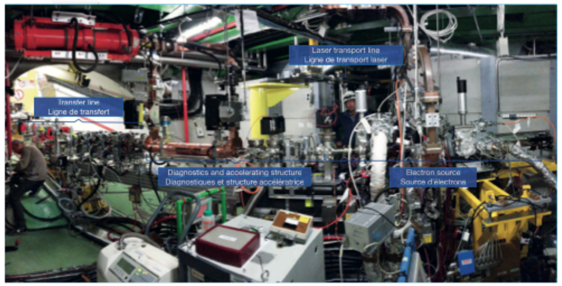


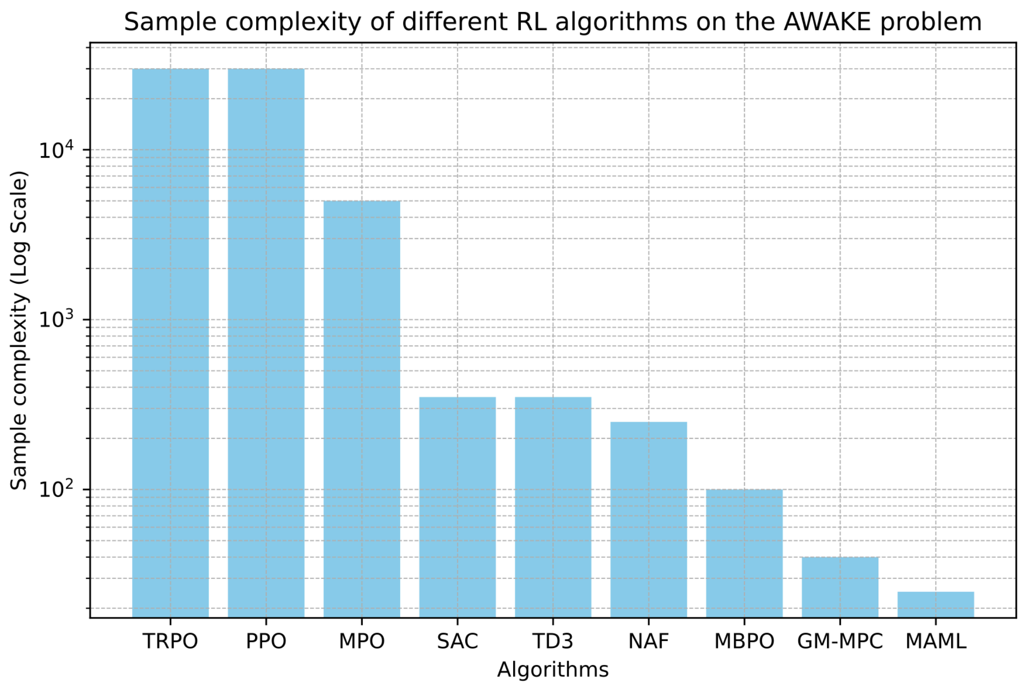
META-RL allows for few shot adaptation

META-RL for accelerators

Tested algorithm on for AWAKE electron line steering at CERN.







Include the physics you know… PINNs

PDEs play a crucial role in many applications in accelerators…

E.g. Finite elements modelling (FEM) methods are standard solvers for various design question: magnets and other accelerator components

FEM limitations when used for design optimisation: computational cost depending on mesh size.

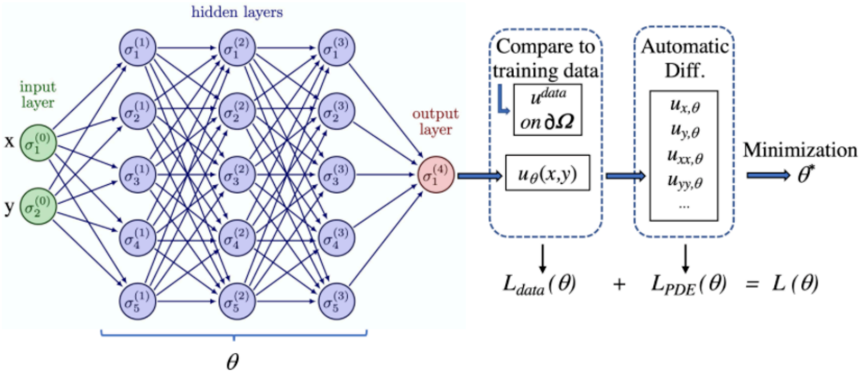
Can solve PDEs with NNs

→ Physics-inspired NNs PINNs: use automatic differentiation and add terms in loss function

Example: $\frac{df}{dt} = Rf(t)(1 - f(t))$

→ to be minimised: $L_{ODE} = \frac{df_{NN}}{dt} - Rf_{NN}(t)(1 - f_{NN}(t))$

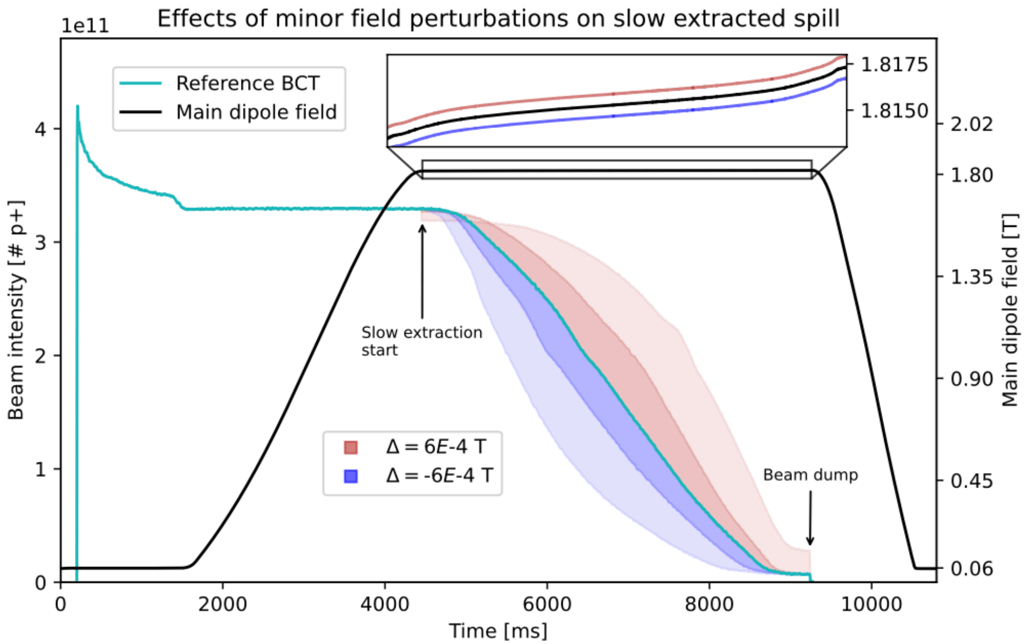
Total loss function: $L = L_{colocation} + L_{ODE}$

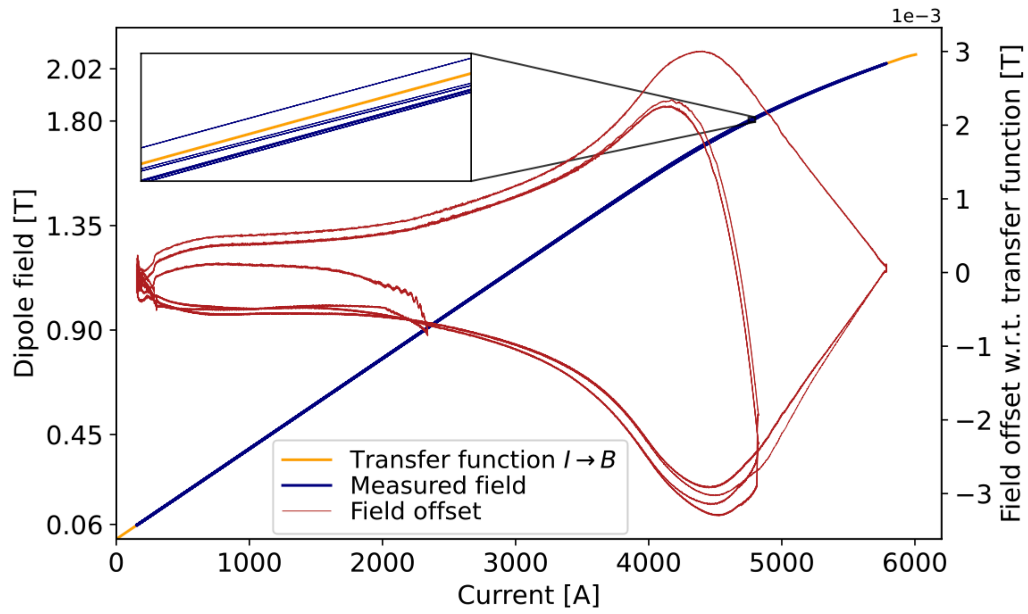


From arXiv:2403.00599

Example: PhyLSTM hysteresis compensation

Hysteresis of magnets in the SPS has impact on slow extracted spill quality





Effects on slope of intensity decrease with varying main dipole field

Example: PhyLSTM hysteresis compensation

Hysteresis modelling rather challenging, no closed form solution.

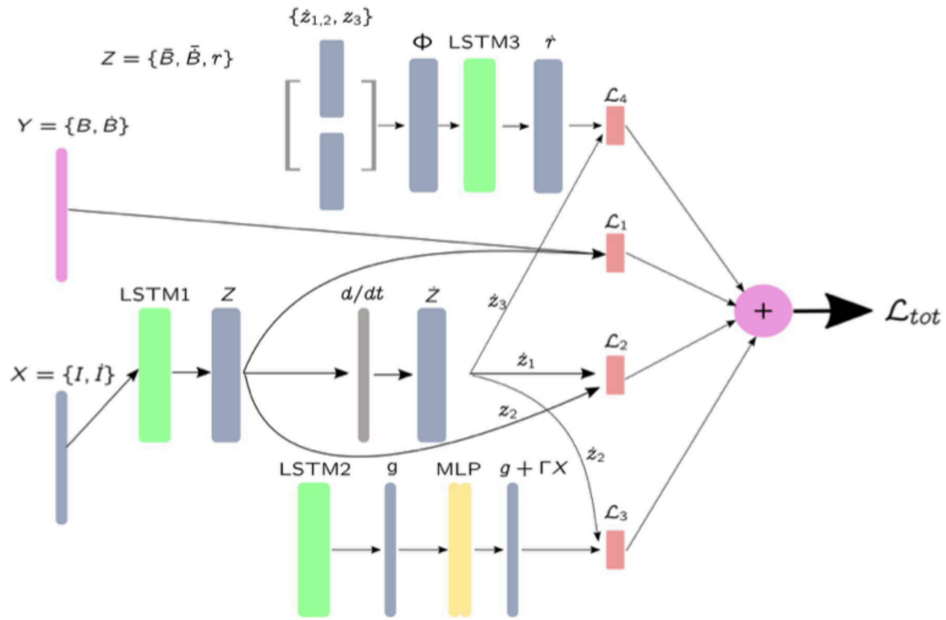
First idea: use Bouc-Wen model $a\ddot{y}(t) + b(y, \dot{y}) + r(y, \dot{y}, y(\tau)) = \Gamma x(t)$, $\ddot{y} + g = \Gamma x$

Input $x = \{I, dI/dt\}$, output $y = \{B, dB/dt\}$

$$\mathcal{L}_1 = MSE(z_1(\theta_1) - y_1) + MSE(z_2(\theta_1) - y_2)$$
$$\mathcal{L}_2 = MSE(\dot{z}_1(\theta_1) - z_2(\theta_1))$$
$$\mathcal{L}_3 = MSE(\dot{z}_2(\theta_1) + MLP(g(\theta_1, \theta_2), x_1))$$
$$\mathcal{L}_4 = MSE(\dot{r}(\theta_1, \theta_3) - \dot{z}_3(\theta_1)); \dot{r} = f(\Phi); \Phi = \{\Delta z_2, r\}$$

$$\Downarrow$$

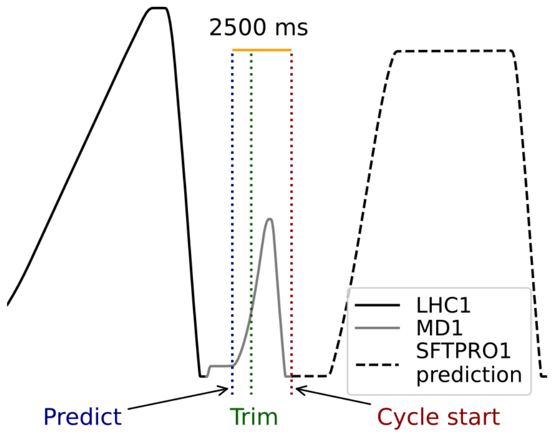
$$\mathcal{L}_{tot} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3 + \eta\mathcal{L}_4$$



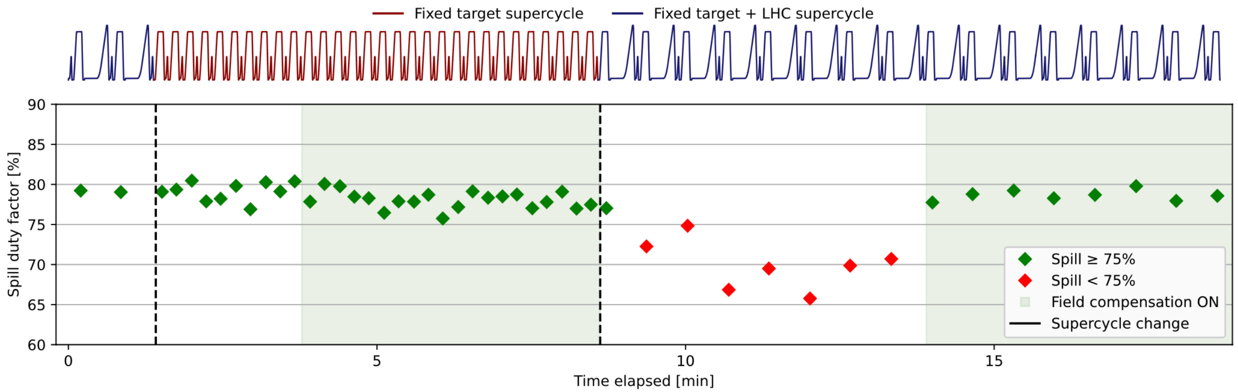
State-of-the-art hysteresis modelling

Best results with transformer-based timeseries forecasting models (purely data-driven):
Temporal Fusion Transformer

$$B(t_0, \ldots, t_N), I(t_0, \ldots, t_{N+M}) \rightarrow B(t_{N+1}, \ldots, t_{N+M})$$

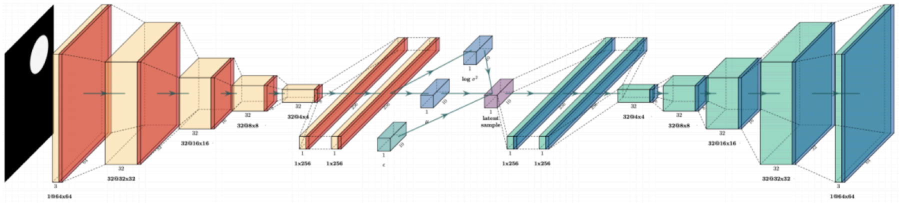


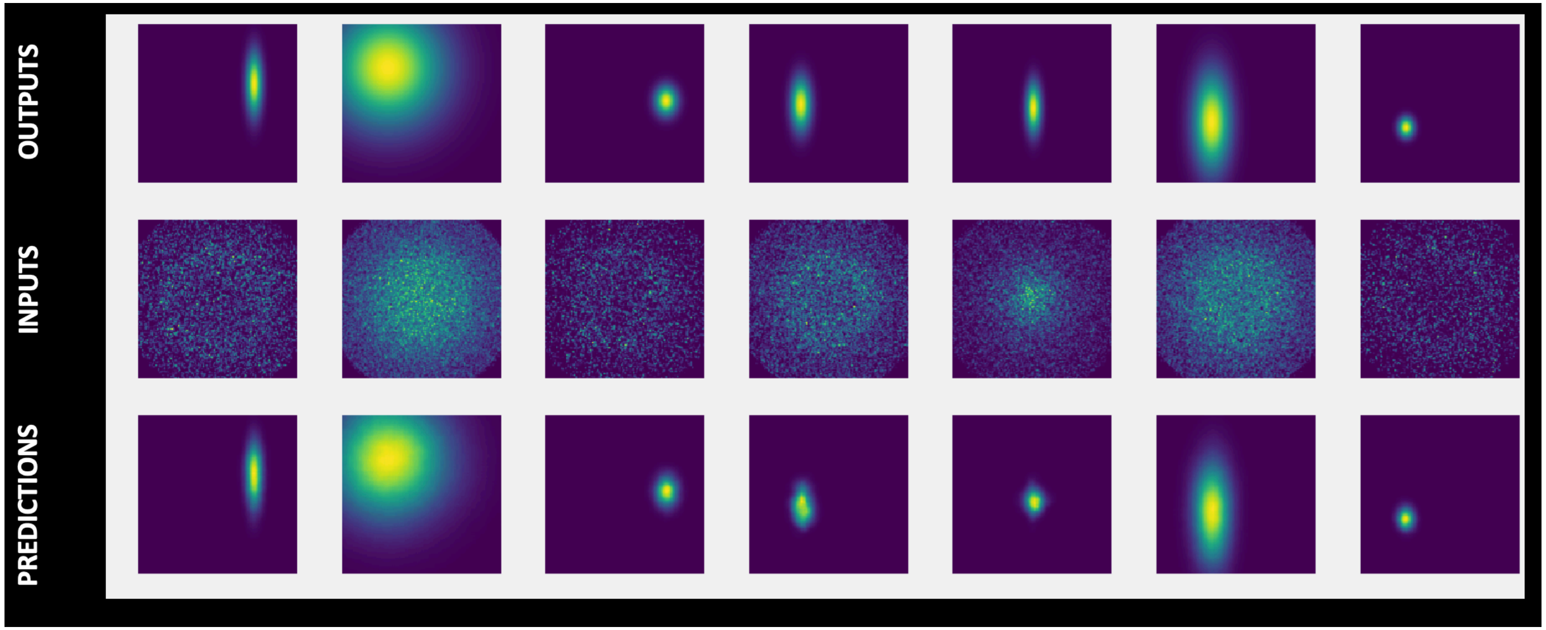
Feed forward correction scheme in control room



Diagnostics and Analysis - Computer vision

Variational auto-encoders for radiation hard **Optical Fibre Imaging** → next generation beam profile monitors?

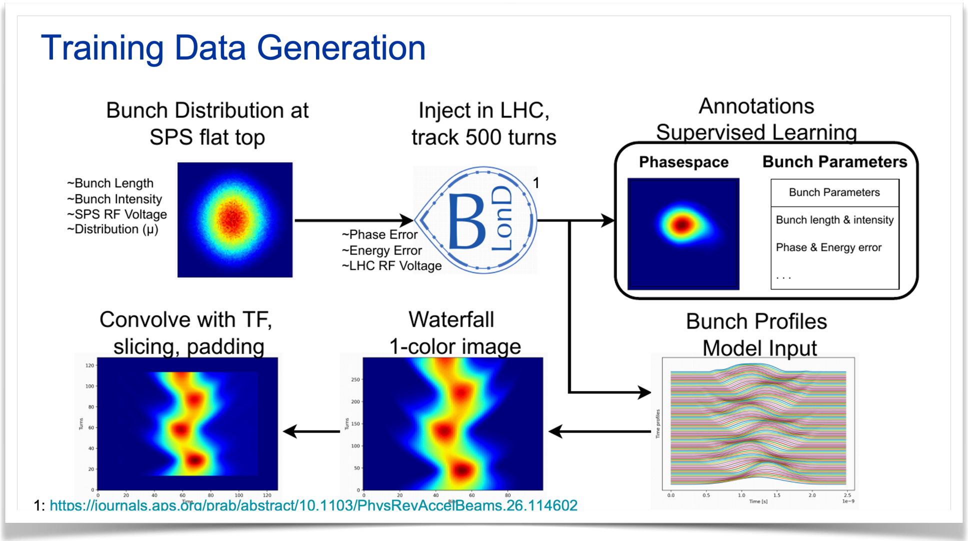


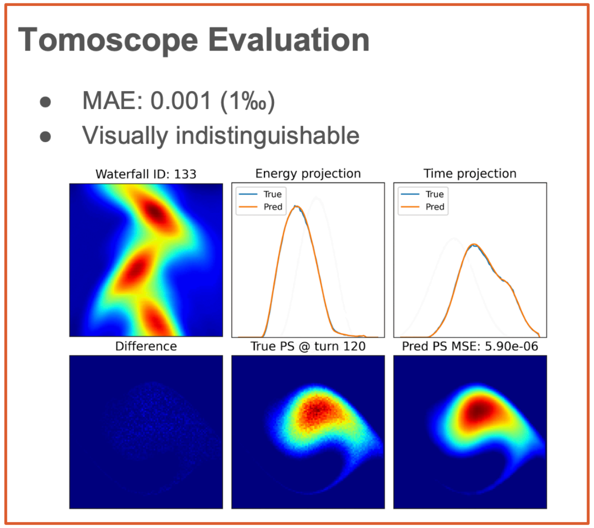


G.Trad

Diagnostics and Analysis - Computer vision

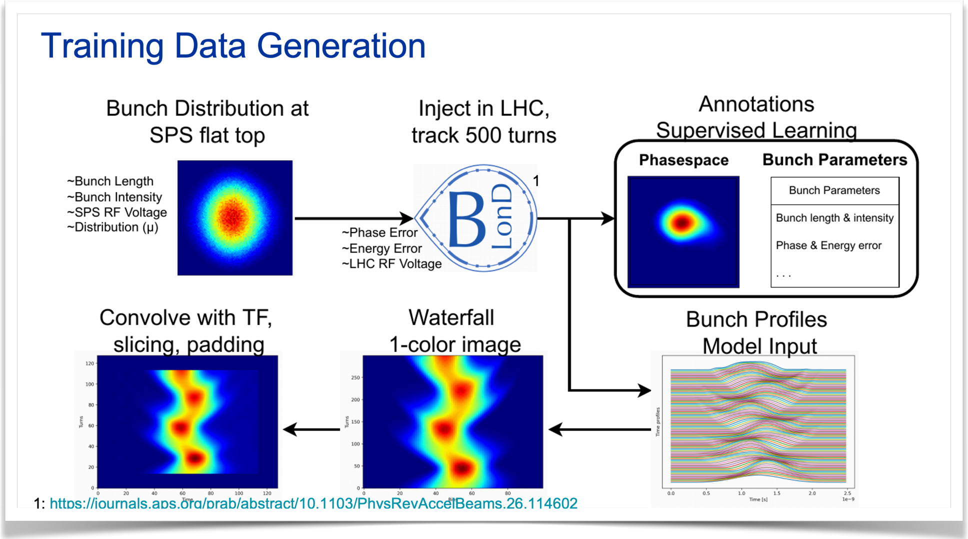
Example: bunch-by-bunch tomographic reconstruction in the LHC; ensemble of auto-encoders trained in simulation on turn-by-turn longitudinal bunch profiles

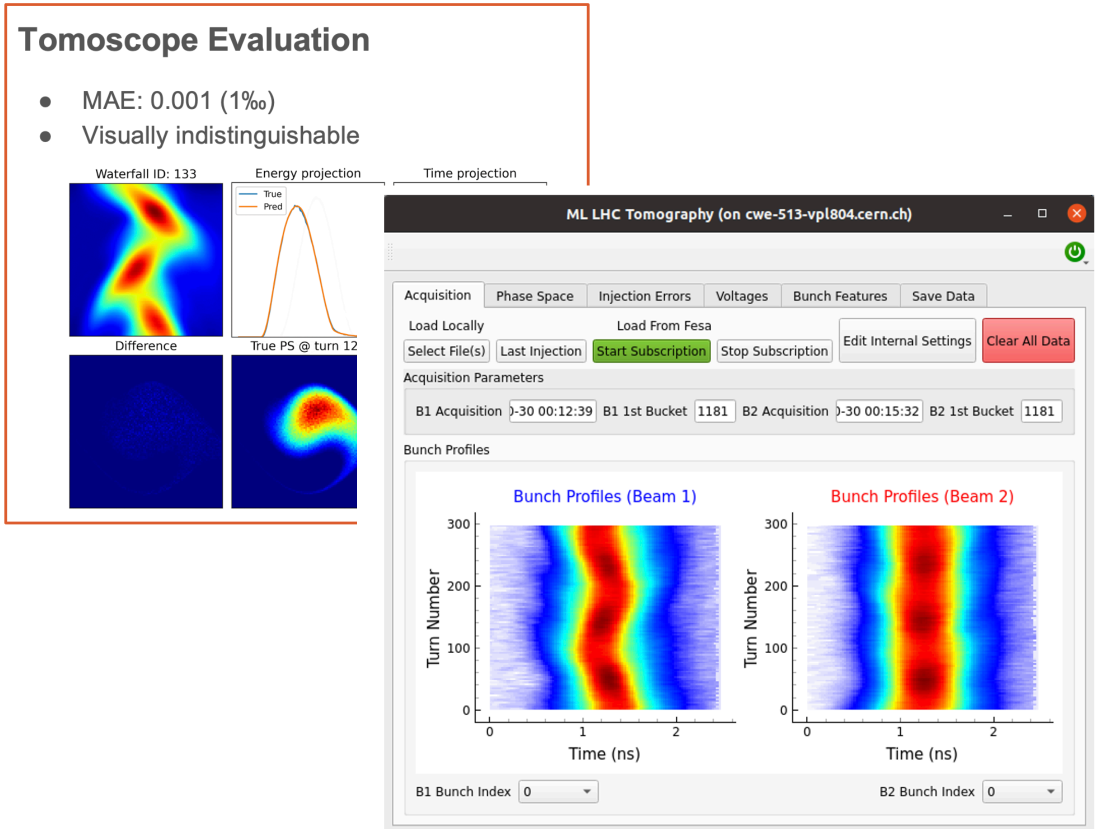




Diagnostics and Analysis - Computer vision

Example: bunch-by-bunch tomographic reconstruction in the LHC; ensemble of auto-encoders trained in simulation on turn-by-turn longitudinal bunch profiles





Final words...

CONCLUSIONS

This paper has described one methodology suitable for the introduction of full automation. Barriers are recognized to exist but are not unsurmountable. Often all that is needed to overcome these barriers is the belief that the process can and should be fully automated.

CERN-ISR-CO/80-29

What's next...

AI is changing how we exploit particle accelerators and will drive how we build new ones

Many different use cases at particle accelerators, for many different types of AI/ML algorithms.

In this lecture series will focus on optimisation and control aspects.

You will get an introduction into

- Bayesian Optimisation
- Reinforcement Learning