Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Surrogate model driven RL optimization of the TWOCRYST crystal angular alignment

Leander Grech

leander.grech@um.edu.mt

University of Malta

Gianluca Valentino

University of Malta

Oliver Aberle

European Organization for Nuclear Research

Federico Alessio

European Organization for Nuclear Research

Chiara Antuono

European Organization for Nuclear Research

Gianluigi Arduini

European Organization for Nuclear Research

Laura Bandiera

Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara

Massimo Benettoni

Istituto Nazionale di Fisica Nucleare, Sezione di Padova

Marco Calviani

European Organization for Nuclear Research

Sara Cesare

Istituto Nazionale di Fisica Nucleare, Sezione di Milano

Victor Coco

European Organization for Nuclear Research

Simone Coelli

Istituto Nazionale di Fisica Nucleare, Sezione di Milano

Georges Daher

European Organization for Nuclear Research

Quentin Demassieux

European Organization for Nuclear Research

Davide De Salvador

Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali di Legnaro

Mario Di Castro

European Organization for Nuclear Research

Luigi Esposito
European Organization for Nuclear Research

Massimiliano Ferro-Luzzi
European Organization for Nuclear Research

Alex Fomin
IJCLab

Jianling Fu
University of Chinese Academy of Sciences

Paolo Gandini
Istituto Nazionale di Fisica Nucleare, Sezione di Milano

Vincenzo Guidi
Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara

Hana Havlikova
European Organization for Nuclear Research

Pascal Hermes
European Organization for Nuclear Research

Sergio Jaimes Elles
Instituto de Física Corpuscular

Sune Jakobsen
European Organization for Nuclear Research

Krzysztof Korcyl
Institute of Nuclear Physics, Polish Academy of Sciences

Gianluca Lamanna
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa

Simone Libralon
Instituto de Física Corpuscular

Chiara Maccani
European Organization for Nuclear Research

Lorenzo Malagutti
Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara

Daniele Marangotto
Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali di Legnaro

Fernando Martinez Vidal
Instituto de Física Corpuscular

Eloise Matheson
European Organization for Nuclear Research

Jose Mazorra de Cos
Instituto de Física Corpuscular

Andrea Mazzolari
Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara
Andrea Merli
Istituto Nazionale di Fisica Nucleare, Sezione di Milano
Haixing Miao
University of Chinese Academy of Sciences
Daniele Mirarchi
European Organization for Nuclear Research
Riccardo Negrello
Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara
Nicola Neri
Istituto Nazionale di Fisica Nucleare, Sezione di Milano
Marcin Patecki
University of Warsaw
Antonio Perillo Marcone
European Organization for Nuclear Research
Jacopo Pinzino
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa
Stefano Redaelli
European Organization for Nuclear Research
Patrick Robbe
IJCLab
Marco Romagnoni
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa
Benoit Salvant
European Organization for Nuclear Research
Izaac Sanderswood
Instituto de Física Corpuscular
Philippe Schoofs
European Organization for Nuclear Research
Regis Seidenbinder
European Organization for Nuclear Research
Gabriele Simi
Istituto Nazionale di Fisica Nucleare, Sezione di Padova
Santiago Solis Paiva
European Organization for Nuclear Research
Enrica Soria
European Organization for Nuclear Research
Marco Sozzi

Istituto Nazionale di Fisica Nucleare, Sezione di Pisa
Elisabetta Spadaro Norella
Istituto Nazionale di Fisica Nucleare, Sezione di Milano
Achille Stocchi
IJCLab
Katie Taylor
European Organization for Nuclear Research
Giorgia Tonani
Istituto Nazionale di Fisica Nucleare, Sezione di Milano
Andrea Triossi
Istituto Nazionale di Fisica Nucleare, Sezione di Padova
Nicola Turini
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa
Santiago Vico Gil
Instituto de Física Corpuscular
Chiara Vinotto
Istituto Nazionale di Fisica Nucleare, Sezione di Pisa
Volodymyr Svintozelskyi
Instituto de Física Corpuscular
Tianyu Xing
University of Chinese Academy of Sciences
Marco Zanetti
Istituto Nazionale di Fisica Nucleare, Sezione di Padova
Federico Zangari
Istituto Nazionale di Fisica Nucleare, Sezione di Milano
Carlo Zannini
European Organization for Nuclear Research

Research Article

Keywords: Reinforcement learning, Crystal channeling, Particle accelerator control, TWOCRYST, LHC, Autonomous alignment

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10868983/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Surrogate model driven RL optimization of the TWOCRYST crystal angular alignment

O. Aberle[1], F. Alessio[1], C. Antuono[1], G. Arduini[1], L. Bandiera[4,5], M. Benettoni[9,10], M. Calviani[1], S. Cesare[7,8], V. Coco[1], S. Coelli[7,8], G. Daeher[1], Q. Demassieux[1], D. De Salvador[6,10], M. Di Castro[1], L. Esposito[1], M. Ferro-Luzzi[1], A. Fomin[3], J. Fu[14], P. Gandini[7,8], L. Grech[15*], V. Guidi[4, 5], H. Havlikova[1], P. Hermes[1], S.J. Jaimes Elles[2], S. Jakobsen[1], K. Korcyl[13], G. Lamanna[11,12], S. Libralon[2], C. Maccani[1,10], L. Malagutti[4, 5], D. Marangotto[7,8], F. Martinez Vidal[2], E. Matheson[1], J. Mazorra de Cos[2], A. Mazzolari[4, 5], A. Merli[7,8,18], H. Miao[14], D. Mirarchi[1], R. Negrello[4,5], N. Neri[7, 8], M. Patecki[17], A. Perillo Marcone[1], J. Pinzino[11, 12], S. Redaelli[1], P. Robbe[3], M. Romagnoni[4, 5], B. Salvant[1], I. Sanderswood[2], P. Schoofs[1], R. Seidenbinder[1], G. Simi[9,10], S. Solis Paiva[1], E. Soria[1], M. Sozzi[11,12], E. Spadaro Norella[7,8,19], A. Stocchi[3], K. Taylor[1], G. Tonani[7,8], A. Triossi[9,10], N. Turini[11,16], G. Valentino[15], S. Vico Gil[2], C. Vinotto[11,12], V. Svintozelskyi[2], T. Xing[14], M. Zanetti[9,10], F. Zangari[7,8], C. Zannini[1]

[1]CERN, Geneva, Switzerland.
[2]IFIC, University of Valencia-CSIC, Valencia, Spain.
[3]IJCLab, Orsay, France.
[4]INFN Ferrara, Italy.
[5]Università degli Studi di Ferrara, Ferrara, Italy.
[6]INFN Sezione di Legnaro, Legnaro, Italy.
[7]INFN Sezione di Milano, Milano, Italy.
[8]Università degli Studi di Milano, Milano, Italy.
[9]INFN Sezione di Padova, Padova, Italy.
[10]Università di Padova, Padova, Italy.
[11]INFN Sezione di Pisa, Pisa, Italy.
[12]Università di Pisa, Pisa, Italy.

[13]Institute of Nuclear Physics – Polish Academy of Sciences, Krakow, Poland.
[14]University of Chinese Academy of Sciences (UCAS), People's Republic of China.
[15]University of Malta, Msida, Malta.
[16]Università di Siena, Siena, Italy.
[17]Warsaw University of Technology, Warsaw, Poland.
[18]now at EPFL, Lausanne, Switzerland.
[19]now at Università degli Studi di Genova, Genova, Italy.

*Corresponding author(s). E-mail(s): leander.grech@um.edu.mt;

Abstract

The TWOCRYST double-crystal experimental setup exploits planar channeling to extract halo protons from the LHC circulating beam and steer them onto a target followed by a crystal. TWOCRYST paves the way for possible future experiments at the LHC based on similar machine layout, such as the ALADDIN experiment that aims to measure the electric and magnetic dipole moments of short-lived charm baryons, in particular $\mathbf{\Lambda}_c^+$. Optimal double-channeling performance requires microradian-level alignment of the Target Collimator Crystal for Precession (TCCP), a task currently performed manually over extensive parameter sweeps. We develop MDHaloEnv, a Gymnasium-based reinforcement learning environment that synthesizes multi-modal sensor data from LHC beam tests at beam energies from 450 GeV to 3 TeV, with the aim of improving future commissioning efficiency of double-channeling experiments. We propose theoretically grounded improvements to the Markov Decision Process formulation, including a reward function that mitigates credit assignment ambiguity, noise-derived convergence thresholds based on impulse noise characteristics, and a smoothness penalty scaled to observation noise variance. Proximal Policy Optimization agents were trained on the data-driven environment and evaluated on the June 2025 450 GeV dataset. This work presents the application of deep reinforcement learning to TCCP crystal alignment in the TWOCRYST experiment, and establishes the data-driven surrogate paradigm as a foundation for future multi-energy and multi-actuator extensions.

Keywords: Reinforcement learning, Crystal channeling, Particle accelerator control, TWOCRYST, LHC, Autonomous alignment

1 Introduction

The TWOCRYST experiment [1], installed in the CERN Large Hadron Collider (LHC) in 2025, serves as a proof-of-principle demonstrator for a future experiment

under consideration at the LHC to measure the electromagnetic dipole moments of the charmed baryon Λ_c^+ [2]. The experiment exploits planar channeling in bent silicon crystals, where charged particles remain trapped within the potential wells between crystallographic planes [3] and follow the macroscopic curvature of the crystal [4].

TWOCRYST employs a double-crystal scheme installed in the LHC at Insertion Region 3. Figure 1 shows a schematic of the TWOCRYST experiment. In double-channeling configuration, a first crystal (TCCS—Target Collimator Crystal for Splitting; 4 mm length, ∼50 µrad bending angle) extracts protons from the beam halo and directs them onto a second crystal (TCCP—Target Collimator Crystal for Precession; 70 mm length, ∼7 mrad bending angle), which further channels the protons arriving onto a downstream VeloPix hybrid pixel detector (PIXEL) [5, 6] and a TWOCRYST Fibre Tracker [7] detector (TFT), both housed in separate Roman pots, which image the deflected particles, with the double-channeled particles appearing in a distinct upper region of the detector. Three beam loss monitors (BLMs) at the TCCP, TCCS, and downstream absorber (TCLA) positions provide complementary information on channeling efficiency. The possibility to add a target in front of the TCCP is also available.

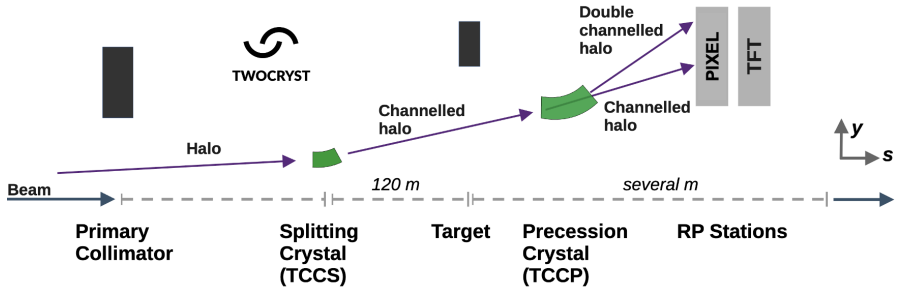


Fig. 1: TWOCRYST experiment schematic.

Achieving efficient double-channeling requires angular alignment of the TCCP with microradian precision with respect to the channeled particles from the TCCS. This is challenging because of the small beams and low signals from BLMs which are typically used to establish the channeling. The optimization objective is to maximize the fraction of total detected intensity concentrated in the upper detector region, which is directly proportional to the double-channeling efficiency of the TCCP crystal. The TCCP angular acceptance is $\theta_L \approx 13\,\mu\text{rad}$, and the goniometer provides 0.1 µrad resolution over an operational range of several mrad. Figure 2 shows the resulting effect of a TCCP rotational sweep on the PIXEL detector frames, illustrating the appearance and varying intensity of the double-channeled spot in the upper detector region; obtained from the data collected during the TWOCRYST beam test in June 2025. Current alignment procedures rely on expert operators performing systematic rotational scans over a >1 mrad range, requiring 2–4 hours per complete sweep, which is a significant amount of beam time for an eventual experiment. The channeling

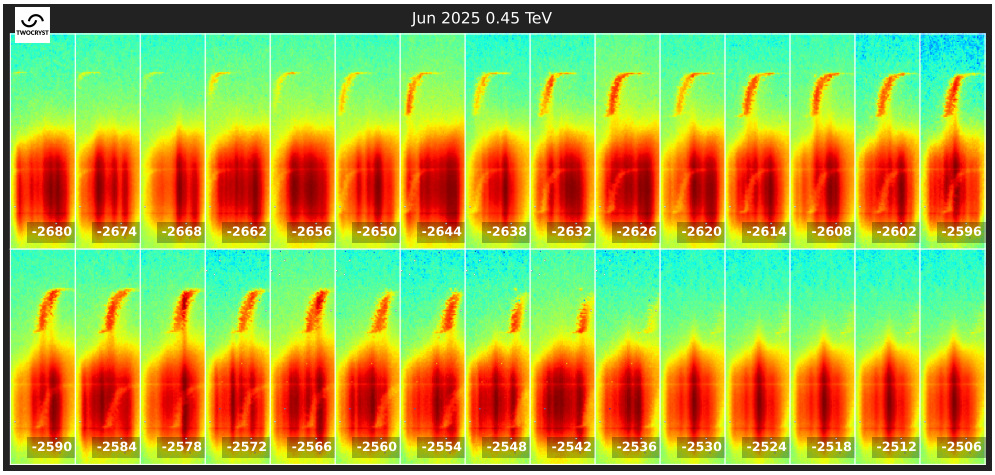


Fig. 2: TCCP goniometer rotational sweep showing varying intensity of the double-channeled spot. Each image is an X-Y distribution of event rate on the PIXEL detector. The boxed number on each frame indicates the goniometer angle in µrad.

needs to be re-established in the event of changing machine configurations (e.g. beam optics), as well as beam conditions.

Reinforcement learning (RL) offers a compelling alternative for this control problem. Modern RL algorithms can learn effective policies after 10^5–10^6 environment interactions and provide continuous adaptation as beam conditions evolve. Previous applications of RL to accelerator control have demonstrated human-expert-level performance in beam emittance optimization [8], FEL pulse property optimization [9], and orbit stability control [10]. However, these applications primarily address problems with well-defined forward models or rapid online learning opportunities. Crystal alignment presents distinct challenges: limited training data availability (few beam test sessions per year), no accurate forward model for possible crystal torsion effects, dechanneling, and surface quality effects, and machine protection constraints.

In this paper, we develop a data-driven RL environment, called MDHaloEnv, that synthesizes observations from historical beam test data, propose theoretically grounded improvements to the MDP formulation, and train and evaluate agents on the primary 450 GeV injection energy dataset, and identify multi-energy generalisation and multi-actuator control as natural extensions of the framework. Four ablation configurations of the reward formulation are evaluated to isolate the contribution of geometric reward decay and action regularisation to within-episode reward variance. The approach learns directly from real accelerator data without requiring analytical physics models or beam dynamics simulators.

The paper is organized as follows. Section 2 describes the beam test data acquisition campaigns and the characteristics of the collected datasets. Section 3 details the data processing pipeline, environment design, and MDP formulation including the proposed theoretical improvements. Section 4 presents training methodology and

experimental results. Section 5 concludes with a discussion of findings and future directions.

2 Beam Tests

2.1 Layout and Instrumentation

The devices and instrumentation which comprise the TWOCRYST setup in IR3 of the LHC include:

- **PIXEL detector**: VeloPix hybrid pixel detector based on the LHCb upgraded Vertex Locator [11] (768×256 pixels, 55 µm pitch) positioned downstream of TCCP. The detector images both double-channeled particles (upper region) and single-channeled or volume-reflected particles (lower region). Readout rate: 0.3–0.7 Hz.
- **TFT detector**: Fibre tracker detector (1500×1500 pixels, 30 µm resolution), positioned downstream of the PIXEL detector. Given that more rotational sweeps were performed with PIXEL, resulting in more data for the surrogate model, this detector was not used to train the RL agent.
- **Beam loss monitors**: Three ionization chambers—BLMTI.04L3 (at TCCP), BLMTI.05L3 (at absorber TCLA), and BLMTI.05R3 (at TCCS)—sampled at $\sim$1 Hz.
- **TCCS and TCCP goniometers**: Piezo-actuated systems with five degrees of freedom (0.1 µrad rotational resolution). The rotation angle serves as the primary RL control variable.

Additionally, the transverse damper [12] (ADTH/ADTV) located in IR4 was used to excite the beam in the horizontal & vertical planes to generate enough resolution in the BLMs and the detectors. All devices are synchronized via the CERN timing network with sub-millisecond accuracy and recorded in the logging system (NXCALS) at 10 min intervals. Table 1 summarizes the key characteristics of the dataset.

2.2 Injection energy (450 GeV) beam test - June 2025

The primary training dataset used in this paper was collected on 22 June 2025 during a beam test at 450 GeV injection energy. Data acquisition spanned approximately 3.5 hours with PIXEL images collected during two windows (00:00–01:00 and 15:00–17:30 UTC). The TCCP rotational sweep covered the range $[-2686, -2280]$ µrad, yielding 366 unique angle bins (after rounding to the nearest integer µrad) with a statistical significance threshold of ≥ 2 PIXEL samples per angle bin. Five double-channeling periods were manually identified, totalling approximately 13 minutes of captured events. The double-channeling TCCP range for the most populated static crystal configuration spans $[-2674, -2530]$ µrad ($\sim$140 µrad range). The sample distribution across angle bins is non-uniform, with a median of $\sim$5 samples per bin and one particular TCCP angle (-2570 µrad) containing approximately 650 samples due to extended dwell time to collect more data for double-channeling efficiency calculations.

Table 1: Summary of the beam test dataset collected in June 2025 at 450 GeV, including the LHC Beam 2 (B2) intensity I_b and vertical emittance ϵ_V. The double-channeling (DC) range indicates the TCCP angular region where double-channeled spots were observed on the PIXEL detector.

TCCP range [µrad]	**Angle bins**	**PIXEL frames**	**Samples/bin** (median)	**DC range** [µrad]	I_b charges	ϵ_V µm
$[-2686, -2280]$	366	1266	$\sim$5	$[-2674, -2530]$	2.78×10^{11}	2.88

Note that a strongly non-uniform angular sampling, in particular the overrepresentation of -2570 µrad, was observed in the dataset, which creates a potential bias that the RL environment must handle through careful sampling strategies.

3 Surrogate model and MDP

To clarify the reinforcement learning (RL) nomenclature and the control framework, Figure 3 illustrates the core Markov Decision Process (MDP) feedback loop. In this paradigm, the alignment optimization is framed as a continuous interaction between a learning *agent* (`PPO`) and the surrogate *environment* (`MDHaloEnv`).

At each discrete time step, the agent observes the current *state* of the system—comprising a rolling buffer of downsampled PIXEL detector images and recent action history—and outputs an *action* corresponding to a rotational adjustment of the TCCP goniometer. The environment applies this continuous action within its transition dynamics to compute the new crystal angle, querying the data synthesizer (`MDDataSynth`) for the corresponding historical sensor data.

To ensure policy robustness and simulate hardware uncertainty, the raw PIXEL detector images undergo a shared salt-and-pepper noise injection before splitting into two distinct processing paths:

- *Observation Path*: Applies random spatial jiggling and downsampling to build the next state representation, preventing the agent from overfitting to exact pixel coordinates.
- *Reward Path*: Extracts unjiggled, exact-crop intensities to calculate the immediate Reward. This reward evaluates the double-channeling efficiency via an upper-to-total intensity ratio, applies a geometric decay weighting to smooth readout fluctuations, and subtracts an l_1 penalty to discourage excessive actuation.

This continuous cycle of action, environmental transition, and reward-state feedback provides the signal necessary for the Proximal Policy Optimization (PPO) agent to learn a stable and autonomous alignment policy.

3.1 Temporal alignment strategy

The fundamental challenge in constructing a coherent training dataset from TWOCRYST instrumentation is the asynchronous nature of the data streams. The

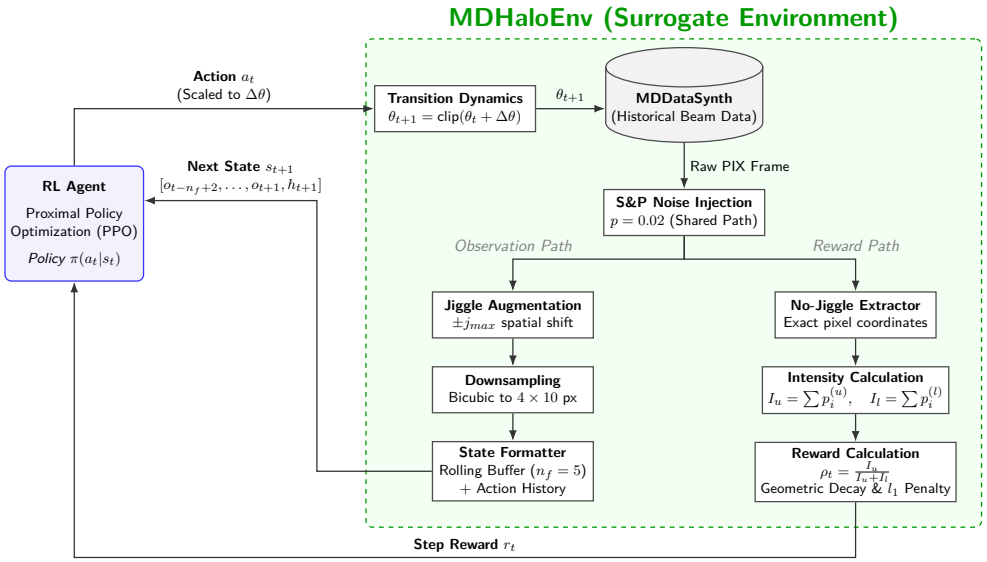


Fig. 3: Architecture of the Markov Decision Process (MDP) feedback loop. The RL Agent issues continuous angular adjustments ($\Delta\theta$) to the `MDHaloEnv` surrogate environment. The environment synthesizes historical detector observations, applies shared stochastic noise, and splits the signal into an augmented observation path for robust state representation and a rigid reward path for precise double-channeling efficiency calculation.

PIXEL detector transmits images to the CERN Control Centre for visualization purposes at $\sim$0.5 Hz while scalar sensors (BLMs, goniometers) operate at $\sim$1 Hz, creating a 2$\times$ frequency mismatch.

For RL environment construction, strict causality must be maintained: the agent must never observe information that would be unavailable during actual online deployment. This excludes naive alignment approaches such as nearest-neighbor interpolation (which may select future measurements), linear interpolation (which requires future data points), and synchronous resampling (which introduces non-physical synthetic data).

We employ an *as-of merge* (point-in-time join) strategy, where each PIXEL timestamp serves as a temporal anchor and all scalar streams are aligned by selecting the most recent measurement with timestamp $t_s \leq t_{\text{PIXEL}}$. Formally, for TCCP rotation data serving as the base time series and each additional data stream k:

$$v_k(t_{\text{base}}) = v_k(\max\{t \in T_k : t \leq t_{\text{base}}\}) \tag{1}$$

where T_k is the set of timestamps for stream k. This operation is implemented using `pandas.merge_asof` with `direction='forward'` from the TCCP rotation timestamps to the PIXEL and BLM streams, preserving causality while maximizing data utilization.

After temporal alignment, TCCP rotation values are binned at 1 µrad intervals to ensure sufficient statistics on which to build a robust surrogate model, and bins with fewer than 2 PIXEL samples are discarded. The resulting filtered dataset is stored as a pandas DataFrame indexed by angle bin, enabling efficient on-demand sampling during RL training.

3.2 Top-k frame selection

To mitigate the strong non-uniformity in per-angle bin sample counts and improve training stability, a top-k filtering step is applied to each angle bin before environment construction. For each bin, the k PIXEL frames with the highest upper-spot intensity (computed over the nominal, unjiggled upper slice) are retained and the remainder discarded.

This biases the surrogate distribution toward frames that contain a visible double-channeled spot, reducing the proportion of low-signal frames that would otherwise dominate the stochastic reward signal at angles near the double channeling peak. The selection is deliberately optimistic: it does not reflect the true marginal distribution of PIXEL frames at each angle, but instead focuses the agent's training experience on the high-reward regime where the learning signal is richest.

The practical effect is improved reward signal-to-noise during early training and faster initial policy improvement, at the cost of some fidelity in representing the full variability of detector observations. This work presents results obtained with $k = 10$, which is a soft cap that only filters the heavily overrepresented bins (principally at TCCP goniometer rotational angle of$-2570\,\mu$rad).

3.3 MDDataSynth: data synthesizer

The MDDataSynth module provides the data backend for the RL environment. Given a requested TCCP angle θ, it returns a randomly sampled data record from the corresponding angle bin, including the PIXEL image, BLM readings, and ADTH/V positions. All sensor data undergo a normalization pipeline:

1. **Log-scale transformation**: BLM and PIXEL intensities are transformed as $v \mapsto \log_{10}(v)$ for $x > 0$, with zero values mapped to the minimum observed log-intensity.
2. **Min-max normalization**: All scalar sensor and pixel values are scaled to the range $[-1, 1]$ using pre-computed dataset statistics:

$$\hat{v} = 2 \cdot \frac{v - v_{\min}}{v_{\max} - v_{\min}} - 1 \tag{2}$$

where v denotes the raw (or log-transformed) feature value, and v_{min} and v_{max} are the minimum and maximum values of that feature computed across the entire training dataset.

Raw PIXEL images are processed through a multi-stage pipeline to construct observation vectors suitable for RL:

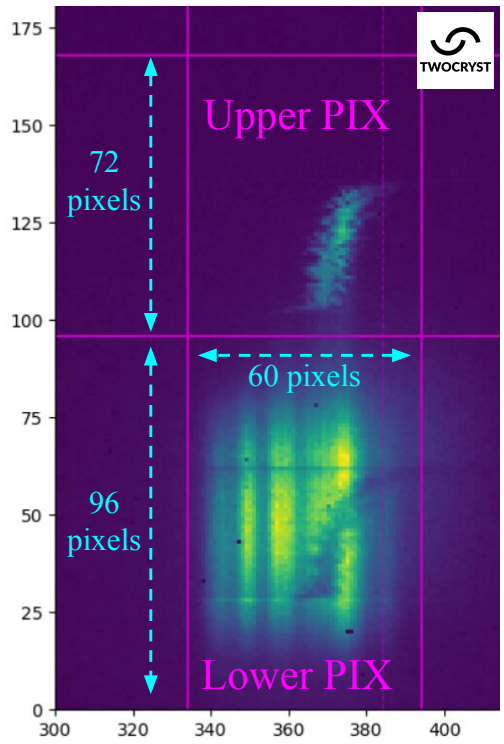


Fig. 4: Upper and lower PIXEL slice definition.

Region-of-interest extraction

Two regions are sliced from the full PIXEL acquisition as shown in Figure 4, with an example of a 60×72 pixels upper and a 60×96 pixels lower regions: an upper slice (pixel coordinates $y \in [y_{upper,1}, y_{upper,2}]$, $x \in [x_{upper,1}, x_{upper,2}]$) capturing the double-channeled particle spot, and a lower slice ($y \in [y_{lower,1}, y_{lower,2}]$, $x \in [x_{lower,1}, x_{lower,2}]$) capturing single-channeled and volume-reflected components. These regions were identified from expert analysis of PIXEL images during confirmed double-channeling periods.

Jiggle augmentation

At each episode reset, a random translation (j_x, j_y) is sampled uniformly from $[-j_{\max}, j_{\max}]$ (typically $j_{\max} = 5$ pixels) and applied to the slicing window coordinates. The same jiggle is applied to both upper and lower slices within an episode to maintain spatial consistency. This augmentation promotes translation invariance and prevents overfitting to exact pixel positions.

Noise injection

Salt-and-pepper noise with corruption probability $p = 0.02$ is applied to each slice before downsampling. This simulates detector readout noise and improves policy robustness to measurement uncertainty.

Downsampling

Both upper and lower noisy slices are resized to 4×10 pixels using bicubic interpolation with anti-aliasing, yielding a 40-dimensional feature vector per PIXEL slice.

3.4 MDHaloEnv: Gymnasium environment

The MDHaloEnv implements the standard Gymnasium [13] interface and encapsulates the complete MDP formulation for TCCP alignment optimization.

3.4.1 State space

The observation vector consists of two parts: a rolling buffer of n_f single-frame observations (default $n_f = 5$), followed by an action history vector appended outside the per-frame structure:

$$\mathbf{s}_t = \left[\mathbf{o}_{t-n_f+1}, \ldots, \mathbf{o}_t,\ \mathbf{h}_t\right] \in \mathbb{R}^{n_f \cdot d + n_f} \tag{3}$$

where $\mathbf{h}_t = [\Delta\theta_{t-n_f+1}, \ldots, \Delta\theta_t] \in \mathbb{R}^{n_f}$ records the n_f most recent normalised angle deltas. The action history h_t records the n_f most recent normalised angle deltas and is appended as a single flat vector outside the per-frame structure, therefore it is not replicated per frame. Its inclusion is motivated by the partial observability inherent in the raw pixel observations: a single PIXEL frame does not uniquely determine the agent's recent trajectory, and without knowledge of prior actions the environment is effectively non-Markovian. Augmenting the observation with the action history restores a fuller state representation, allowing the policy to condition on both the current sensor signal and the sequence of goniometer adjustments that produced it.

Each single-frame observation o_t is composed of a flattened upper PIXEL slice $p_u \in \mathbb{R}^{40}$ and a flattened lower PIXEL slice $p_l \in \mathbb{R}^{40}$, both produced by bicubic downsampling to 4×10 pixels, together with a normalised episode progress signal $\hat{t} = 2t/(T_{max} - 1)$. Several additional signals are available in the environment but are deliberately excluded from the active training configuration: the three BLM readings $b_{TCCP}, b_{TCCS}, b_{TCLA}$; the normalised absolute TCCP rotation $\hat{\theta}$; and the normalised spatial jiggle offsets (j_x/j_{max}, j_y/j_{max}).

The exclusion of the absolute TCCP rotation is a design choice to prevent the agent from bypassing the primary learning objective. Were $\hat{\theta}$ included, the agent could exploit angle memorisation, i.e. learning to map a specific goniometer reading directly to an action without ever developing sensitivity to the pixel intensity content of the detector images. Excluding it forces the policy to ground its decisions entirely in the PIXEL signal, which is the operationally meaningful quantity. Similarly, the jiggle offsets are withheld so that the observation is decoupled from the precise spatial

location of the extraction window; the agent must respond to intensity patterns rather than to an implicit encoding of slice position. The BLM readings are excluded for two reasons: the first is to prevent overfitting to beam-condition-specific loss values that may not generalise across beam tests or beam energies; and second, because the BLM signals recorded during the available beam tests were observed to be highly noisy, raising the risk that the agent learns to exploit spurious patterns in the one-dimensional scalar stream rather than the richer, spatially structured PIXEL content that constitutes the true alignment signal.

In the active training configuration ($n_f = 5$; BLMs, TCCP rotation, and jiggle disabled; episode time and action history enabled), the observation dimension is $(40+40+1)\times 5+5 = 410$. All observations lie in $[-1, 1]$ by construction.

3.4.2 Action space

The environment supports both continuous and discrete action spaces. In the continuous formulation, $a \in [-1, 1]$ is scaled to a TCCP rotation change:

$$\Delta\theta = \text{round}(\text{deadzone}(a) \cdot s_a) \quad [\mu\text{rad}] \tag{4}$$

where s_a is the action scale (typically 3 µrad) and the deadzone function zeroes actions below a configurable threshold to prevent unnecessary jitter. In the discrete formulation, three actions map to $\{-1, 0, +1\} \times s_a$ µrad.

The continuous action formulation is used throughout this work. The discrete formulation, while simpler to optimise, restricts the agent to fixed angular increments of $\{-s_a, 0, +s_a\}$ μrad and cannot produce sub-step corrections. This resolution limitation is undesirable given the ~ 13 µrad angular acceptance window of the TCCP, where fine-grained dithering near the channeling peak is operationally important. The continuous formulation retains this capability through the deadzone-clipped scaling of Equation (4), while using only one action dimension.

3.4.3 Transition dynamics

At the start of each episode, the initial TCCP angle is sampled uniformly from the interior of the data range, with a 5% edge buffer on each side: $\theta_0 \sim \mathcal{U}[\theta_{\min}+\delta,\ \theta_{\max}-\delta]$ where $\delta = 0.05(\theta_{\max}-\theta_{\min})$. This prevents the agent from starting at boundary states where only one direction of movement is available.

At each step, the environment: (1) computes the new angle $\theta_{t+1} = \text{clip}(\theta_t + \Delta\theta,\ \theta_{\min},\ \theta_{\max})$; (2) queries MDDataSynth for a PIXEL image and BLM values at θ_{t+1}; (3) applies salt-and-pepper noise to the full PIXEL acquisition (shared between the observation and reward paths); (4) constructs the single-frame observation $\mathbf{o}_{t+1}$ using the *jiggled, resized* upper and lower crops (Section 3.3) from the noisy image; (5) computes the step reward using *unjiggled, unresized* crops from the same noisy image via a dedicated no-jiggle extractor (see Section 3.4.4); (6) appends $\mathbf{o}_{t+1}$ to the rolling buffer and returns $[\mathbf{o}_{t-n_f+2}, \ldots, \mathbf{o}_{t+1}, \mathbf{h}_{t+1}]$ as the new state.

This separation between the observation path (jiggled, resized) and the reward path (unjiggled, raw pixel counts) is deliberate: the spatial augmentation improves

policy robustness, while the unjiggled crop gives a stable intensity reference for the reward signal.

Episodes terminate when the maximum step count $T_{\max}$ is reached (default $T_{\max} = 100$) or when the angle exceeds the allowed bounds $[\theta_{\min}, \theta_{\max}]$, which triggers truncation. In the current configuration, out-of-bounds truncation applies the same reward value as the boundary step (`rew_truncation: same`) rather than a fixed penalty.

3.4.4 Reward function

The reward is based on the PIXEL frame upper-to-total intensity ratio:

$$\rho_t = \frac{I_u^{(t)}}{I_u^{(t)} + I_l^{(t)}} \tag{5}$$

where $I_u^{(t)}$ and $I_l^{(t)}$ are the instantaneous pixel-sum intensities of the upper and lower crops at step t. The optimization objective is to maximize the fraction of total detected intensity concentrated in the upper detector region, which corresponds directly to the double-channeling efficiency of the TCCP crystal. The ratio formulation is preferable to maximizing $I_u^{(t)}$ alone because the total beam intensity reaching the detector varies with machine conditions; the ratio is insensitive to such global fluctuations and rewards angular configurations that genuinely enhance double-channeling rather than simply high-intensity beam periods.

The raw per-step ratio ρ_t is not applied directly as the step reward. Instead, a geometrically weighted average over the rolling history of raw rewards is computed:

$$r_t = c_s \cdot \frac{\sum_{k=0}^{n_f-1} \gamma^{n_f-1-k} \rho_{t-k}}{\sum_{k=0}^{n_f-1} \gamma^{n_f-1-k}} \tag{6}$$

where $\gamma \in (0, 1)$ is a decay factor and the weights are re-normalised over the available history during episode warm-up. This assigns the highest weight to the most recent frame while smoothing over short-timescale PIXEL readout fluctuations, reducing the variance of the reward signal without reintroducing the credit assignment ambiguity of a uniform rolling mean. In the limit $\gamma \to 0$ the formulation recovers the purely current-frame reward while in $\gamma \to 1$ it recovers a uniform mean; in practice $\gamma = 0.9$ provides a modest smoothing effect across the $n_f = 5$ frame buffer. c_s is a reward scale factor set to $c_s = 10$.

A practical limitation of this formulation is that the reward inherits the frame-to-frame variance present in the beam test datasets: at a fixed TCCP angle, consecutive PIXEL frames can differ substantially due to beam intensity fluctuations, halo population changes, and detector readout noise. This stochasticity manifests as reward noise that cannot be fully mitigated by the ratio normalization, since both $I_u^{(t)}$ and $I_l^{(t)}$ co-vary with global beam conditions. The noise-derived convergence threshold (Section 3.4.7) provides a principled characterization of this irreducible variance and is used to assess when the agent has reached a stable angular position.

Reward path

A critical implementation detail is that the intensities are computed from *unjiggled, unresized* crops, not from the observation slices. At each step, after S&P noise is applied to the full PIXEL acquisition, a dedicated no-jiggle extractor (zero spatial offset) cuts the calibrated upper and lower regions at their nominal pixel coordinates. The intensity is then the raw pixel sum over that crop:

$$I_u^{(t)} = \sum_i p_i^{(u)}, \qquad I_l^{(t)} = \sum_i p_i^{(l)} \tag{7}$$

where $p_i^{(u)}$ and $p_i^{(l)}$ are the pixel values of the unjiggled, unresized upper and lower crops taken from the noisy image (which has been shifted to $[0, 1]$ prior to noise injection). Using unjiggled crops for the reward ensures that the intensity reference is spatially consistent across episodes regardless of the jiggle offset assigned at reset, avoiding a confound between spatial augmentation and reward magnitude.

3.4.5 Smoothness penalty

To discourage unnecessarily large or oscillatory goniometer adjustments, an action cost term is subtracted from the step reward when enabled:

$$r_t \leftarrow r_t - \frac{1}{\lambda_a} \sum_{k=0}^{n_f-1} |\Delta\theta_{t-k}|$$

where $\Delta\theta_{t-k}$ are the angle deltas in the rolling action history and $\lambda_a = 100$ is a penalty scale hyperparameter. The penalty accumulates over the same n_f-step window as the observation buffer, ensuring the agent is penalised for sustained high-magnitude actuation rather than individual steps in isolation. This formulation is equivalent to an ℓ_1 regulariser on recent actions, promoting sparse, deliberate corrections over continuous dithering at a cost proportional to the magnitude of angular displacement.

3.4.6 Observation buffer initialization

A subtle but important implementation detail concerns the initialization of the rolling observation buffer at episode start. The naive approach fills the buffer as $[\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \mathbf{o}_0]$, creating a non-stationary artifact: consecutive zero frames are encountered *only* at episode initialization, allowing the policy to trivially identify episode boundaries—an information leak absent during deployment.

We instead initialize the buffer by repeating the first observation: $[\mathbf{o}_0, \mathbf{o}_0, \ldots, \mathbf{o}_0]$. This eliminates the artificial distribution shift and ensures that the initial state distribution is consistent with the steady-state distribution encountered during rollouts.

3.4.7 Noise-derived convergence threshold

A theoretically grounded convergence threshold from the injected noise model is derived, which provides a lower bound on the irreducible reward variance at any fixed TCCP angle. Salt-and-pepper noise at corruption probability p induces an intensity variance of:

$$\sigma_I^2 = 4Np(1-p) \tag{8}$$

for a crop of N pixels. Propagating through the ratio reward (Eq. 5) via standard error propagation gives the noise-floor reward standard deviation:

$$\sigma_r = \frac{2\sqrt{p(1-p)} \cdot \sqrt{I_l^2 H_u W_u + I_u^2 H_l W_l}}{(I_u + I_l)^2} \tag{9}$$

where (H_u, W_u) and (H_l, W_l) are the upper and lower crop dimensions. For a sliding window of n_w rewards, the corresponding threshold is:

$$\tau_{\text{conv}} = z_\alpha \cdot \frac{\sigma_r}{\sqrt{n_w}} \tag{10}$$

The threshold is parametrised by the S&P corruption probability p, the empirical mean upper and lower slice intensities I_u and I_l, the window length n_w, and the chosen confidence level z_α. For the ablation study presented in Section 4.3, these are set to $p = 0.2$, $I_u = 60$, $I_l = 700$, $n_w = 50$, and $c_s = 10$, yielding noise-floor thresholds of $\tau_{\text{conv}}^{(80)} = 0.089$, $\tau_{\text{conv}}^{(95)} = 0.136$, and $\tau_{\text{conv}}^{(99)} = 0.179$. The elevated I_l relative to I_u reflects the dominance of the single-channeled and volume-reflected population in the lower detection region across most of the operational TCCP range, as seen in Figure 5.

In practice, applying this threshold as an early-termination criterion requires that the reward distribution at each angle bin be well-characterized, which in turn demands a sufficient number of PIXEL samples per angle bin to separate genuine angular-dependent variation from frame-to-frame stochasticity. The beam test datasets available for this work do not consistently provide the sample density required for reliable per-bin convergence assessment: as detailed in Section 2, the median sample count is as low as 2 frames per bin in the September campaigns. Consequently, the convergence criterion is retained in the environment implementation but is not used to trigger early episode termination. Instead, τ_{conv} is employed as a diagnostic metric during training analysis, allowing the reward variance at different angular positions and across ablation configurations to be compared against the noise-floor baseline.

4 Training and results

4.1 Reinforcement learning algorithm

We train agents using Proximal Policy Optimization (PPO) [14], implemented via Stable-Baselines3 [15], selected for its robustness to hyperparameter choices and stability with high-dimensional continuous observations.

Network architecture

Both the policy (actor) and value (critic) networks are independent two-layer MLPs with hidden dimensions $410 \to 128 \to 128$, using Tanh activations throughout. The policy head produces a mean action $\mu(s) \in [-1, 1]$ via a final Tanh activation, with a learnable log-standard-deviation parameter defining the Gaussian exploration noise. The value head outputs an unbounded scalar $V(s) \in \mathbb{R}$.

Hyperparameters

Table 2 summarises the PPO hyperparameters used throughout. Training runs for 10^6 total environment steps on a single CPU process. The rollout buffer accumulates $n_{\text{steps}} = 1024$ transitions per update; each update performs 10 gradient epochs over mini-batches of size 64, giving $\lfloor 1024/64 \rfloor = 16$ mini-batch updates per epoch.

Table 2: PPO hyperparameters

Hyperparameter	**Value**
Learning rate α	3×10^{-4} (constant)
Rollout steps n_{steps}	1024
Mini-batch size	64
PPO epochs per update	10
Discount factor γ	0.99
GAE parameter λ	0.95
Clip range ε	0.2
Entropy coefficient	0.01
Value function coefficient	0.5
Max. gradient norm	0.5
Parallel environments	1
Random seed	123

Callbacks and checkpointing

A custom `EvalCallback` fires every $\max(n_{\text{steps}}, f_{\text{eval}}) = 1024$ timesteps. At each evaluation point, one deterministic episode (mean action, no exploration noise) is rolled out on a held-out evaluation environment instance. The callback logs per-step reward mean and standard deviation within the episode, episode length, and the settled and convergence flags described in Section 3.4.7. The model weights achieving the highest mean best-step reward across the ten most recent evaluation checkpoints are saved to disk as the best model.

4.2 Reward landscape

Prior to training, the reward landscape was characterised by sweeping the initial TCCP angle through the full range of the June 2025 dataset and saturating the rolling observation buffer at each angle before computing the reward via Equation (5).

Figure 5 presents the mean reward r_t together with the contributing intensity components I_u and I_l as a function of the TCCP rotation angle.

The mean reward profile features a broad, fluctuating peak within the roughly $[-2580, -2570]$ µrad interval, aligning perfectly with the double-channeling region identified in the June 2025 beam test data (Section 2.2). Outside this optimal window, the reward decays toward a stable baseline dominated by the volume-reflected and single-channeled particle populations. This confirms that Equation (5) successfully

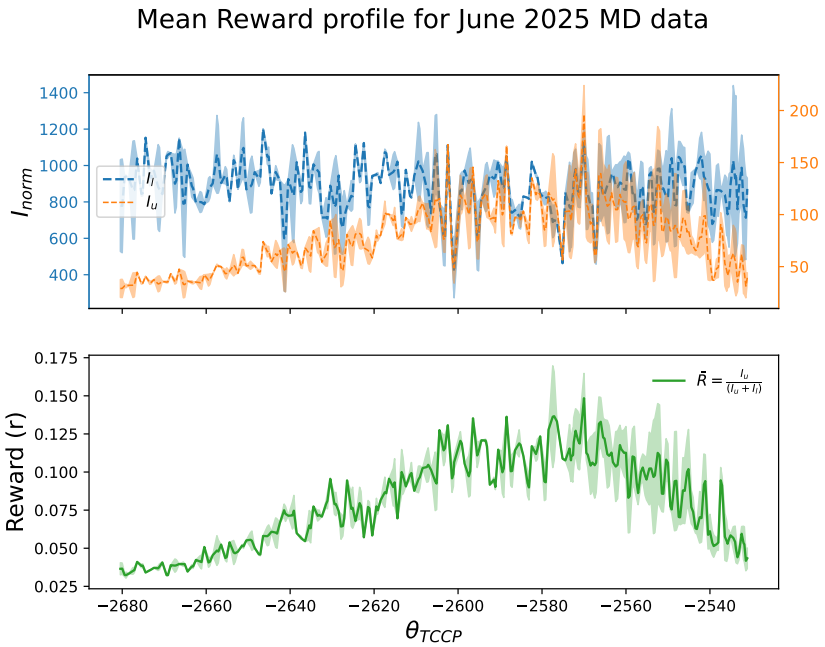


Fig. 5: Reward landscape sweep over the June 2025 450 GeV beam test dataset. *Top*: Mean lower-region intensity I_l (blue, left axis) and mean upper-spot intensity I_u (orange, right axis) as a function of TCCP rotation angle. The two are drawn on separate axes because I_l exceeds I_u by roughly an order of magnitude and would otherwise mask its structure. *Bottom*: Mean per-frame intensity ratio $\rho_t = I_u/(I_u + I_l)$ of Equation (5), with the rolling observation buffer saturated prior to evaluation. Shaded bands denote $\pm 1\sigma$ across the frames sampled in each angle bin. A noisy peak at $\approx [-2580, -2570]$ µrad coincides with the double-channeling region confirmed by expert analysis of the beam test data. The upper intensity I_u closely tracks the reward profile, rising by a factor ≈ 3 across the sweep, whereas I_l fluctuates about a constant level with no comparable angular structure, indicating that the reward signal is predominantly driven by changes in the double-channeled particle population in the upper detector region. The reward spread widens correspondingly inside the channeling window ($\bar{\sigma}_r \approx 0.14$), reflecting the larger frame-to-frame variability of the double-channeled population and setting the irreducible variance floor characterised in Section 3.4.7.

shapes a smooth, unimodal optimization landscape essential for stable gradient-driven policy search.

A closer examination of the individual intensity traces reveals a key asymmetry. The upper-spot intensity I_u closely tracks the reward signal-rising sharply within the channeling window and settling at a low baseline elsewhere. Conversely, the lower-region intensity I_l exhibits minimal structured variation, behaving instead like broadband noise across the entire sweep. Consequently, the ratio reward is effectively numerator-dominated: $r_t \approx I_u^{(t)} / \left(I_u^{(t)} + I_l^{(t)}\right)$ is primarily driven by changes in I_u. The I_l term acts mostly as a constant denominator offset that scales the reward magnitude without artificially sharpening or broadening the topological peak.

To learn a robust navigation policy, the agent relies on meaningful variation in the distribution of PIXEL upper-spot images across TCCP angles, allowing it to infer peak proximity purely from spatial pixel features. However, as detailed in Section 2.2, the available beam test datasets for surrogate environment construction suffer from highly constrained size and angular coverage. With a median of just $\sim$ 5 PIXEL samples per angle bin in the June dataset—and as few as 2 in the September multi-energy tests—the synthetic environment provides limited upper-spot diversity at any specific TCCP setting, even when applying salt-and-pepper noise augmentations.

Consequently, the surrogate MDP likely undersamples the true observation distribution at the double-channeling peak. In this regime, the upper spot is bright, yet its spatial morphology fluctuates alongside halo population and real-time beam conditions on timescales much faster than the PIXEL readout period. Together with the noisiness of I_l, this data sparsity constrains the effective signal-to-noise ratio of the reward signal and motivates the noise-derived convergence threshold introduced in Section 3.4.7 as a means of characterising the irreducible variance floor imposed by the surrogate environment.

4.3 Ablation configurations

Four reward formulation variants are evaluated to isolate the contribution of the decay weighting and action-cost penalty introduced in Sections 3.4.4 and 3.4.5:

1. **Current-frame reward**: the base formulation $r_t = c_s \cdot \rho_t$ with no decay and no action penalty, corresponding to $\gamma \to 0$ and $\lambda_a^{-1} = 0$.
2. **Decay reward**: geometric weighting over the rolling reward history with $\gamma = 0.9$, no action penalty.
3. **Act-cost penalty**: current-frame reward with the ℓ_1 action regulariser enabled ($\lambda_a = 100$), no decay.
4. **Act-cost + Decay**: both modifications active simultaneously, $\gamma = 0.9$ and $\lambda_a = 100$.

All other hyperparameters (Table 2) are held fixed across configurations. Results of the variance diagnostic across these configurations are presented in Section 4.4. Each variant is trained with $n = 1$ run; the results in Section 4.4, therefore, reflect single-run behaviour rather than a seeded average and are used to compare reward smoothness properties rather than to establish statistical significance.

4.4 Reward variance evolution against convergence thresholds

The within-episode reward standard deviation $\hat{\sigma}_r$ logged at each evaluation checkpoint is compared against the noise-floor convergence thresholds $\tau_{\text{conv}}(z_\alpha)$ derived in Section 3.4.7 at three confidence levels:

$$\tau_{\text{conv}}^{(\alpha)} = z_\alpha \cdot \frac{\sigma_r}{\sqrt{n_w}}, \qquad z_\alpha \in \{1.282,\ 1.960,\ 2.576\} \tag{11}$$

corresponding to 80%, 95%, and 99% confidence intervals respectively.

The three thresholds form a nested set of noise-consistency bounds and should not be read as significance levels for rejecting convergence. Rather, each $\tau^{(\alpha)}$ answers the question: *what within-episode reward standard deviation would pure injected noise*

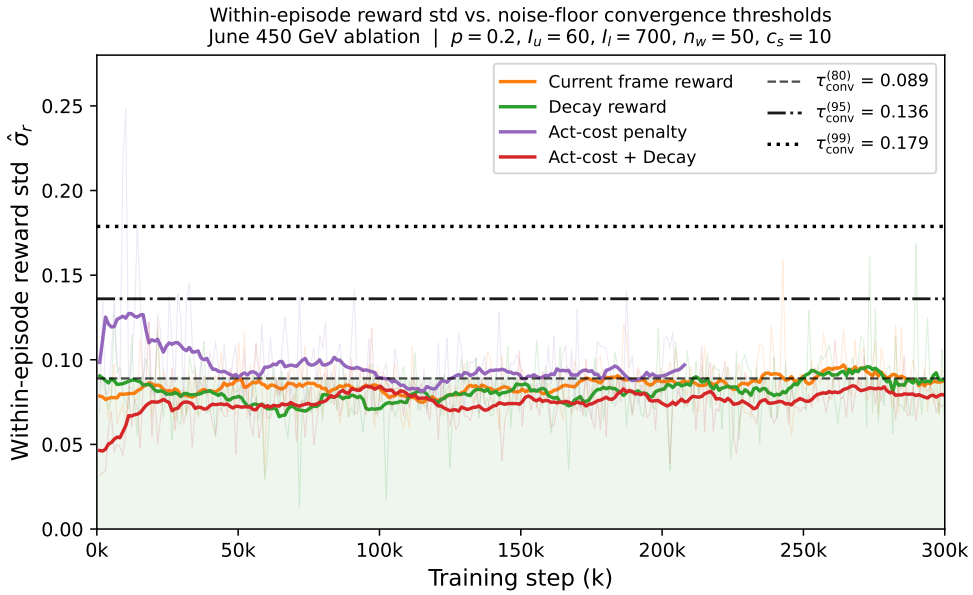


Fig. 6: Within-episode reward standard deviation $\hat{\sigma}_r$ as a function of training step for four ablation configurations, compared against noise-floor convergence thresholds $\tau_{\text{conv}}^{(\alpha)}$ at 80%, 95%, and 99% confidence levels ($p = 0.2$, $I_u = 60$, $I_l = 700$, $n_w = 50$, $c_s = 10$). The three thresholds form a nested noise-consistency envelope: $\tau^{(80)}$ is the strictest bound, admitting only the lowest-variance noise realisations, while $\tau^{(99)}$ defines the ceiling of noise-consistent behaviour. All configurations converge near or below $\tau^{(80)}$ after approximately 10^5 steps, confirming that residual within-episode variance is attributable to injected salt-and-pepper noise rather than policy instability. The act-cost penalty variant exhibits elevated variance during early training before descending to the noise floor, while the combined act-cost and decay configuration achieves the lowest $\hat{\sigma}_r$ throughout, reflecting the smoothing effect of geometric reward weighting and action regularisation.

alone produce, with probability α, at a fixed TCCP angle? A consequence of this construction is that the directionality is inverted relative to conventional hypothesis testing: $\tau^{(80)}$ is the *strictest* criterion, admitting only the quietest noise realisations, while $\tau^{(99)}$ is the most permissive, accepting higher variance as still noise-consistent. Convergence is declared when $\hat{\sigma}_r \leq \tau^{(\alpha)}$; falling below $\tau^{(80)}$ therefore requires that the within-episode reward variance has dropped to the lower tail of what the noise floor alone would produce.

Figure 6 shows the evolution of $\hat{\sigma}_r$ across four ablation configurations over the course of training. All configurations converge toward or below $\tau^{(80)}$ after approximately 10^5 steps, indicating that residual within-episode variance is dominated by frame-to-frame measurement noise rather than continued policy-driven angular excursions. The act-cost penalty variant begins the training run above $\tau^{(95)}$, reflecting the additional reward variance introduced by the penalty term itself before the policy stabilises; it subsequently descends toward the noise floor in line with the other configurations. The combined act-cost and decay variant exhibits the lowest $\hat{\sigma}_r$ throughout training, consistent with the dual smoothing effect of geometric reward weighting and action regularisation. That all trained policies ultimately cluster near $\tau^{(80)}$ confirms that the convergence threshold has diagnosed the correct noise regime: the irreducible variance floor imposed by the surrogate environment has been reached, and further training cannot be expected to reduce within-episode reward fluctuations below this bound.

As noted in Section 3.4.7, the threshold is retained in the current implementation as a diagnostic metric rather than an operational early-termination criterion, owing to insufficient per-angle sample density in the available beam test datasets. The convergence of all ablation variants to the noise floor nevertheless validates the theoretical derivation of $\tau^{(\alpha)}_{\text{conv}}$ and motivates the collection of denser per-angle PIXEL samples in future beam test campaigns, which would enable the criterion to be activated as a live termination signal during online deployment.

4.5 Training performance (Act-cost + Decay)

The agent was trained on the June 2025 beam test dataset, which provides the richest angular coverage and highest sample counts among the available beam test datasets. At every evaluation checkpoint (every 1024 timesteps), the mean and standard deviation of per-step rewards within the deterministic evaluation episode are recorded. The mean episodic return is plotted as a function of training timesteps to characterise the learning dynamics and assess whether the policy converges to a stable solution within the 10^6-step budget.

Figure 7 shows the learning curves averaged over $n = 4$ independent runs. The mean episodic return rises from approximately 1.68 at initialisation to a plateau around 1.72 within the first 5×10^4 steps and remains stable thereafter, confirming consistent convergence across seeds. The mean best-step reward (dashed) reaches 1.90–2.05, indicating that the agent reliably identifies angular configurations in the double-channeling region within each episode. The mean worst-step reward (dotted) stabilises near 1.50–1.55, reflecting the reward floor at sub-optimal angles adjacent to the target region.

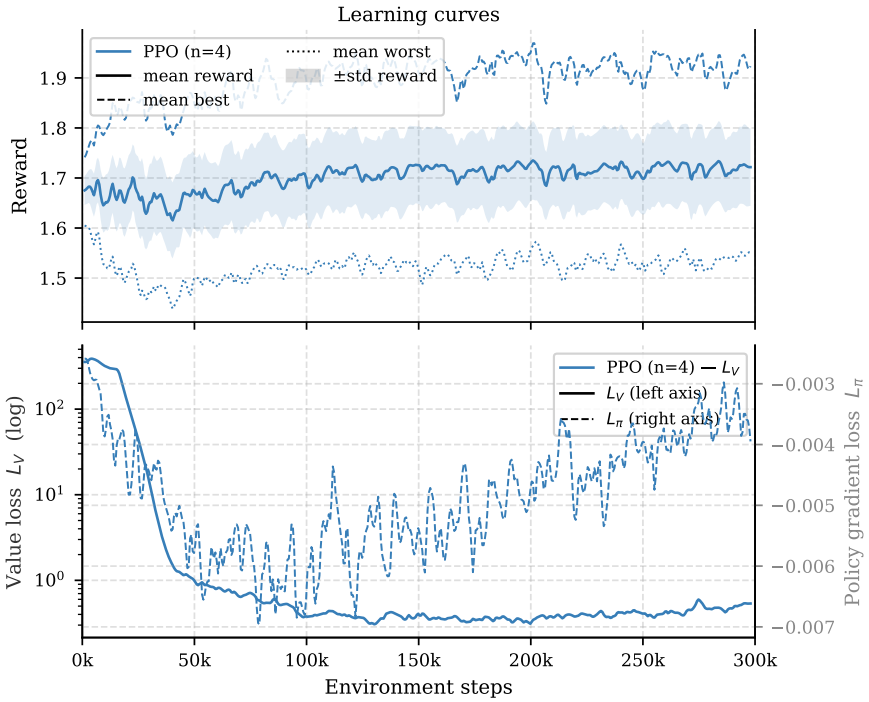


Fig. 7: Learning curves for the PPO agent trained on the June 2025 450 GeV dataset, averaged over $n = 4$ independent runs. *Top:* Mean episodic return (solid), mean best-step reward (dashed), and mean worst-step reward (dotted) as a function of training steps; the shaded band indicates ± 1 standard deviation of the per-step reward within the evaluation episode. *Bottom:* Value loss L_V (left axis, log scale) and policy gradient loss L_π (right axis) as logged by PPO during training.

The $\pm 1\sigma$ band remains stable over training, consistent with the policy converging to a strategy where the residual reward variance is attributable to inter-PIXEL-sample stochasticity at the occupied TCCP angle rather than policy instability. Value loss L_V decreases by approximately three orders of magnitude (from $\sim$300 to $\sim$0.5) and plateaus by 1×10^5 steps, indicating that the critic has learned a stable value estimate. Policy gradient loss L_π remains small and negative throughout, with an exploratory stage during the first $\sim 1 \times 10^5$ steps in which it decreases steadily in tandem with the critic, reflecting continued policy improvement driven by the improving value baseline. Once the critic plateaus, L_π also begins to settle toward smaller magnitudes, consistent with the policy entering a refinement regime where only incremental improvements to the action distribution remain. No clear convergence plateau is visible within the 3×10^5-step window shown; however, evaluation of individual agents from checkpoints beyond 2×10^5 steps does not yield measurably better performance than those obtained from the 1×10^5–2×10^5 step region, suggesting that the useful learning signal is effectively exhausted well before the end of the training budget.

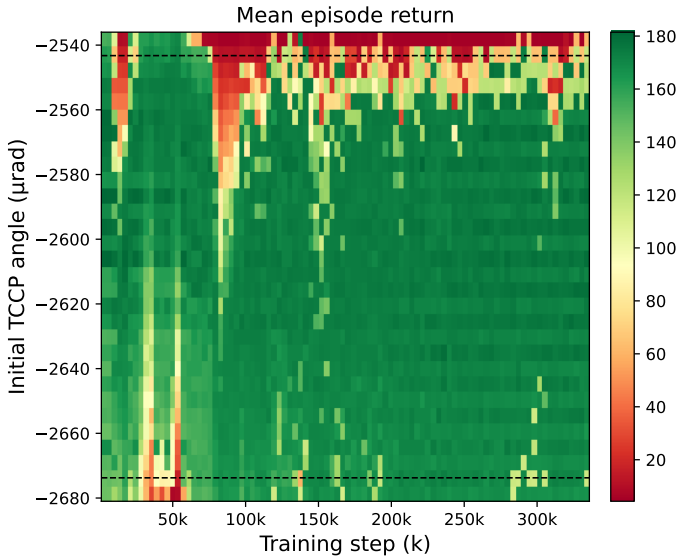


Fig. 8: Mean episodic return as a function of training step (x-axis) and initial TCCP angle (y-axis), averaged over three evaluation episodes per checkpoint. Early checkpoints ($< 5 \times 10^4$ steps) yield returns proportional to proximity to the double-channeling region, consistent with a passive policy. A high-return band (green, $\sim$ 150–180) emerges after 2×10^5 steps for initial angles in $[-2680, -2560]$ µrad, while angles above -2560 µrad exhibit persistently low returns attributable to the 5% edge buffer applied during training. Dashed lines indicate the edge buffer boundary.

4.6 Checkpoint sweep analysis

To assess how policy performance evolves jointly with training progress and initial conditions, a systematic sweep evaluation was conducted at every saved checkpoint across the full training run. At each checkpoint, three episodes were executed for a range of initial TCCP angles reported in Section 2.2, and the mean return, per-step standard deviation, and best-step return were recorded per episode. Figures 8 to 10, respectively, present these three quantities as heatmaps over training step (x-axis) and initial TCCP angle (y-axis). As an example, the colour of a cell in Figure 8 corresponds to the averaged return (sum of all episode rewards) of all three episodes rolled out for that specific agent checkpoint and initial TCCP angle at which the evaluation episodes were initialised.

The mean return heatmap shows that early checkpoints (before 5×10^4 steps) yield returns proportional to the angle difference at which double-channeling is to be found (~ -2570 µrad). This is consistent with a policy that has not yet learned directed navigation, and instead settles at the initial TCCP angle and collects passive rewards. Beyond 5×10^4 steps shows an exploratory period, where episodes with intial TCCP angle at < -2560 µrad mark a significant improvement in episodic returns, at the cost of > -2560 µrad initial TCCP angles to suffer consistently low returns

throughout the rest of training. A high-return band (green, $\sim\ 150 - 180$) emerges after 2×10^5 steps and stabilises for initial angles in the range $[-2560, -2680]$ µrad. The low performance above TCCP initial angle of ~ -2560 µrad can be attributed to the initial TCCP angle buffer of 5% the allowable TCCP range, from the edge of the allowable TCCP angle values, in order to avoid training at the edges of the dataset.

The standard deviation heatmap exhibits elevated variance during the transitional phase of training (roughly 50k–200k steps). The converging pattern observed in the mean returns heatmap is repeated with the return standard deviation heatmap between TCCP angle of $[-2680, -2560]$ µrad after training step 2×10^5 steps, where the agent settles on a policy which obtains high returns with low variance.

The best-step return heatmap provides an optimistic view of policy capability: it asks not whether the agent converges reliably, but whether it visits the double channeling region at all during an episode. Here the green band spans the full angular range from early in training onwards, indicating that the policy consistently finds high-reward configurations within an episode even from distant starting angles where the mean return is moderate. Notably, the residual red region concentrated near -2540 µrad in early checkpoints — where the policy overshoots the peak — largely disappears by 3×10^5 steps, reflecting a refinement of the dithering strategy.

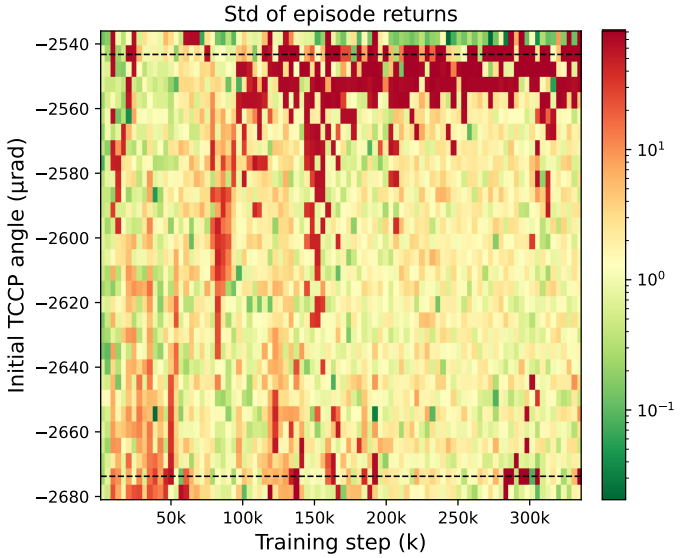


Fig. 9: Standard deviation of episodic returns as a function of training step and initial TCCP angle (log scale). Elevated variance during the transitional phase (50k–200k steps) reflects the policy's exploratory behaviour. For initial angles in $[-2680, -2560]$ µrad, variance decreases markedly after $2e5$ steps, converging toward the noise floor and confirming that the policy stabilises onto a consistent high-return strategy in this region. Dashed lines indicate the 5% edge buffer boundary.

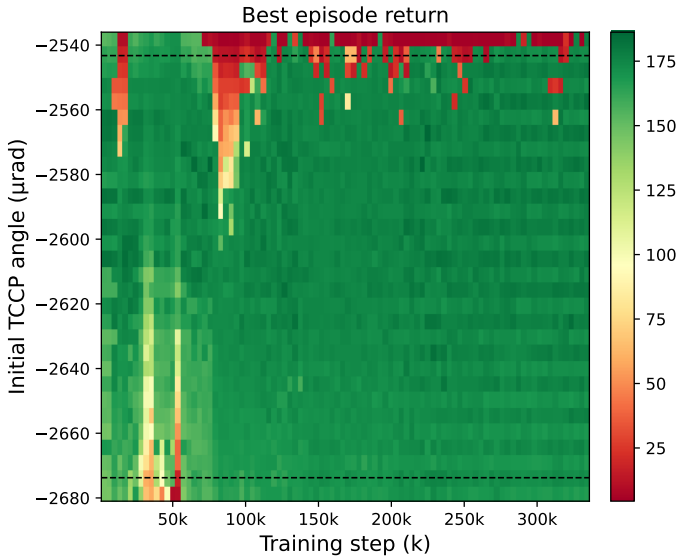


Fig. 10: Best per-episode return (maximum single-step reward within each episode) as a function of training step and initial TCCP angle. The green band spanning the full angular range from $> 1.5e5$ training steps indicates that the policy reliably visits the double-channeling region within an episode even from distant starting angles. The residual low-return region above $-2540\,\mu\text{rad}$, where the policy overshoots the reward peak, largely decreases by 3×10^5 steps. Dashed lines indicate the 5% edge buffer boundary.

4.7 TCCP angular sweep evaluation

To characterise the spatial generalisation of the trained policy, a systematic sweep evaluation was conducted using the best checkpoint obtained when training with the action cost (Section 3.4.5) and reward decay (Equation (6) in Section 3.4.4), saved at $\sim 3.36 \times 10^5$ training steps. A single deterministic episode was executed from each integer initial TCCP angle spanning the full data range $[\theta_{\min}, \theta_{\max}]$, yielding one trajectory per starting position. No exploration noise was applied during evaluation (`deterministic=True`), so the results reflect the learned policy's mean action at each state.

The upper panel of Figure 11 shows the mean (blue), maximum (green), and minimum (red) per-step reward achieved within each episode, plotted against the initial TCCP angle. The mean reward is broadly stable across the full sweep range, lying in the interval $[1.6, 1.8]$ with a standard deviation band of approximately ± 0.1. The maximum per-step reward reaches values around 2.2–2.3 across most starting positions, indicating that the policy successfully visits the double-channeling region during the episode regardless of where it begins. The minimum reward floor is consistently near 1.4, reflecting the baseline intensity ratio observed outside the channeling peak. Notably, no strong systematic dependence of mean reward on initial angle is apparent

across the $[-2670, -2560]$ µrad range, suggesting that the policy has learned a robust navigation strategy rather than one that degrades for starting positions far from the optimum.

The lower-left panel displays the full TCCP rotation trajectory for each episode, colour-coded from blue (most negative initial angles) to red (most positive). Two distinct behavioural regimes are visible. Episodes initialised below the channeling peak (blue trajectories) exhibit a monotonic positive drift upward toward it, whereas those initialised above the peak (red trajectories) drift downward. It is also noted how it is more likely for episodes to go outside the allowable TCCP range when initialised in the red range, which can be attributed to the peak of the reward profile being skewed towards this range (see Figure 5), resulting in larger changes in the PIXEL profile for the same TCCP angle change in this operating region of MDHaloEnv. Once within the double-chaneling band, the policy exhibits characteristic chattering behaviour—alternating single-step positive and negative actions—indicative of a learned dithering strategy to maintain proximity to the reward peak under the stochastic environment dynamics.

The histogram of episode-final TCCP angles (lower-right panel, computed as the median of the last ten steps of each trajectory) shows that the large majority of episodes terminate in the range $[-2596, -2560]$ µrad, with the modes at

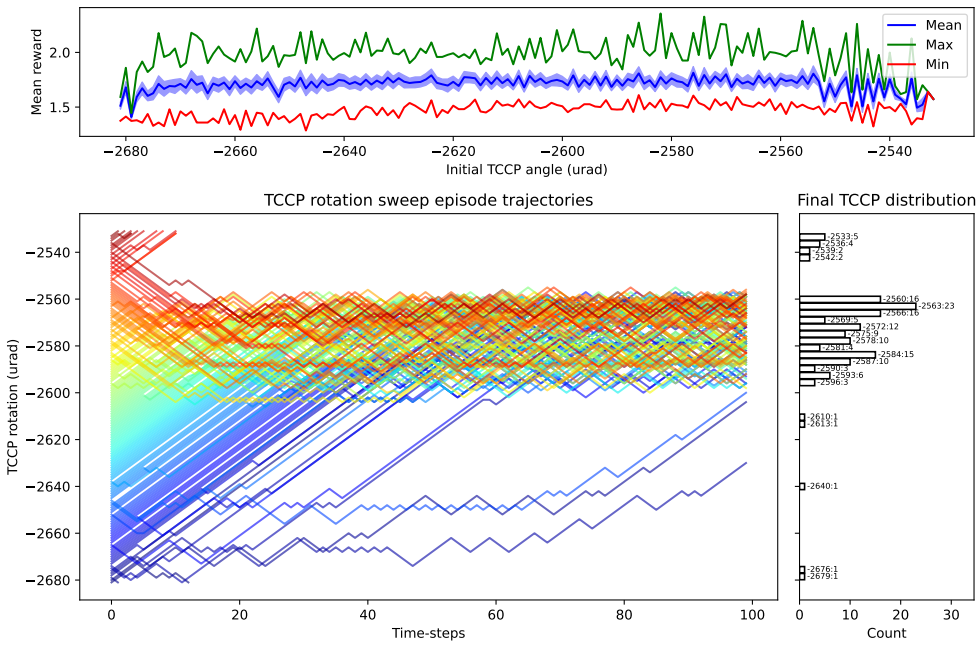


Fig. 11: Evaluation sweep over all integer initial TCCP angles in the operational range. **Top:** Mean, maximum, and minimum per-step rewards per episode. **Lower left:** Full TCCP rotation trajectories. **Lower right:** Histogram of episode-final TCCP angles.

$-2563\,\mu\text{rad}$ ($\sim$ 25 episodes) and $-2584\,\mu\text{rad}$ ($\sim$ 15 episodes). This $\sim 36\,\mu\text{rad}$ band falls within the double-channeling range identified from the June 2025 beam test data ($[-2674, -2530]$ µrad), and is centred near the reward peak observed in the landscape sweep. The sparse tail of episodes settling above $-2540\,\mu\text{rad}$ (near -2542 to $-2533\,\mu\text{rad}$) corresponds to trajectories that overshot the peak, while the isolated counts below $-2600\,\mu\text{rad}$ represent episodes with initial angles far from the channeling region which did not converge within the 100-step episode budget.

Taken together, these results demonstrate that the trained PPO agent has learned a generalised navigation policy that successfully identifies and converges to the double-channeling angular region from the large majority of initial conditions spanning the full operational range. The primary failure mode of non-convergence from the most negative starting angles is attributable to the finite episode horizon rather than a policy deficiency, and could be addressed either by increasing $T_{\max}$ or by implementing a curriculum over initial angle distributions during training. Another mode of failure is overshooting the allowable TCCP angle range, such as the episodes starting in the red-most range, which corresponds to episodes initialised with a goniometer angle to the right of the reward profile peak.

5 Conclusion and outlook

This paper presented training and evaluation results from a data-driven reinforcement learning system for autonomous crystal collimator alignment in the TWOCRYST experiment at the CERN LHC. The system introduces MDHaloEnv, the first RL environment for bent crystal alignment constructed entirely from real accelerator data. Key contributions include:

1. A temporal alignment strategy based on as-of merges that preserves causality across asynchronous sensor streams spanning a $2\times$ frequency ratio between the PIXEL detector and scalar sensors.
2. Theoretically grounded MDP improvements including a current-frame intensity-ratio reward that eliminates credit assignment ambiguity, proper rolling-buffer initialisation that removes episode-boundary information leakage, a geometrically decayed reward formulation that smooths frame-to-frame PIXEL variance without reintroducing temporal credit ambiguity, an ℓ_1 action regulariser that penalises sustained high-magnitude actuation, and noise-derived convergence thresholds that characterise the irreducible variance floor of the surrogate environment.
3. An ablation study across four reward formulation variants, demonstrating that the combined act-cost and decay configuration achieves the lowest within-episode reward variance and that all trained policies converge to the theoretically predicted noise floor after approximately 10^5 training steps.
4. Characterisation of the trained policy's spatial generalisation across the full operational TCCP range, demonstrating robust convergence to the double-channeling angular region from the large majority of initial conditions.

The approach demonstrates the feasibility of learning crystal alignment policies directly from limited beam test data without analytical physics models. This data-driven paradigm is particularly valuable for crystal channeling applications where simulator fidelity is limited by poorly characterized crystal imperfections (torsion, non-uniform bending, surface miscut).

Several objectives identified during this work are deferred to future investigation. Multi-energy generalisation, i.e. training agents that condition on beam energy and transfer across the 450 GeV–3 TeV LHC operational range, requires denser per-energy beam test datasets than are currently available. Beam tests conducted in September 2025 provide a starting point but are too sparse for reliable surrogate construction at 1 TeV and 2 TeV. Multi-task agents with explicit energy conditioning will be investigated once richer datasets are available from future beam test campaigns. Similarly, an ablation study with multiple random seeds per configuration is deferred to when dataset size permits a more controlled experimental design.

Several directions for future work are envisioned. In the short term, the system should be extended to multi-actuator control (simultaneous TCCS and TCCP optimization) and validated through hardware-in-the-loop testing with realistic actuator dynamics. The inference pipeline, already implemented via PyJAPC integration, will be deployed in shadow mode during the next available beam test to compare RL decisions against expert operator actions in real time. Medium-term goals include safe RL formulations with hard BLM threshold constraints, model-based RL using learned surrogate models $f(\theta_{\mathrm{TCCP}}, E) \rightarrow (I_u, I_l)$ for improved sample efficiency, and meta-learning approaches for rapid adaptation to new beam conditions. Ultimately, this work will pave the way for eventual autonomous crystal alignment for the ALADDIN experiment, which will utilize such precession crystals for precision measurements of charmed baryon electric and magnetic dipole moments.

Acknowledgements

TWOCRYST is carried out under the Physics Beyond Colliders (PBC) Fixed-Target Working Group and is a collaboration of the institutes CERN, INFN (Italy), IFIC (University of Valencia-CSIC, Spain), IJCLab (France), the University of Malta, the Institute of Nuclear Physics of the Polish Academy of Sciences, Warsaw University of Technology (Poland), and UCAS (China). The project acknowledges support from the PBC project; the HL-LHC project; the ATLAS-ALFA collaboration; the ERC SELDOM Grant No. 771642; the INFN CSN5 project OREO; the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU under Project Title PRIN 202277EWLW; contributions from the AICRYSCON project funded by Xjenza Malta through the FUSION: R&I Research Excellence Programme; support from MICIU, AEI, and GVA (Spain) under projects PID2022-139842NB-C22, CIPROM/2022/36, ASFAE/2022/030, and NextGenerationEU; and funding from the National Science Centre, Poland (project No. 2021/43/D/ST2/02761). The TWOCRYST collaboration also acknowledges support from the LHCb VELO and ONLINE groups.

References

[1] Bandiera, L., Cai, R., Carsi, S., Cesare, S., Dewhurst, K.A., D'Andrea, M., Salvador, D.D., Gandini, P., Guidi, V., Hermes, P., Lezzani, G., Malagutti, L., Marangotto, D., Maccani, C., Mazzolari, A., Merli, A., Mirarchi, D., Monti-Guarnieri, P., Montanari, C.E., Negrello, R., Neri, N., Prest, M., Redaelli, S., Romagnoni, M., Selmi, A., Tonani, G., Vallazza, E., Veres, D., Zangari, F., Zielińska, M.: Performance of short and long bent crystals for the TWOCRYST experiment at the large hadron collider. Eur. Phys. J. C Part. Fields **85**(12), 1373 (2025)

[2] Akiba, K., Alessio, F., Benettoni, M., Bizzeti, M., Borgato, F., Bucci, F., Cardinale, R., Cesare, S., Citterio, M., Coco, V., Collins, P., Dall'Occo, E., Ferro-Luzzi, M., Fomin, A., Forty, R., Fu, J., Gandini, P., Giorgi, M., Grabowski, J., Jaimes Elles, S.J., Jakobsen, S., Kou, E., Lamanna, G., Li, H., Libralon, S., Marangotto, D., Martinez Vidal, F., Cos, J., Merli, A., Miao, H., Neri, N., Neubert, S., Petrolini, A., Pilloni, A., Pinzino, J., Prest, M., Robbe, P., Rossi, L., Ruiz-Vidal, J., Sanderswood, I., Sergi, A., Simi, G., Sorbi, M., Sozzi, M., Spadaro Norella, E., Stocchi, A., Tonani, G., Tork, T., Triossi, A., Turini, N., Vallazza, E., Vico Gil, S., Wang, Z., Wang, M., Xing, T., Zanetti, M., Zangari, F.: ALADDIN: An Lhc Apparatus for Direct Dipole moments INvestigation. Technical report, CERN, Geneva (2024). https://doi.org/10.17181/CERN.2G4V.0YAO . https://cds.cern.ch/record/2905467

[3] Lindhard, J.: Influence of crystal lattice on motion of energetic charged particles. Kongelige Danske Videnskabernes Selskab **34**(14), 64 (1965)

[4] Tsyganov, E.N.: Estimates of cooling and bending processes for charged particle penetration through a monocrystal. Technical Report FERMILAB-TM-0684, Fermi National Accelerator Laboratory, Batavia, Illinois (September 1976). https://lss.fnal.gov/archive/test-tm/0000/fermilab-tm-0684.pdf

[5] Cesare, S., Neri, N.: PROOF-OF-PRINCIPLE TEST FOR AN EXPERIMENT FOR SHORT-LIVED PARTICLE DIPOLE MOMENTS AT LHC. CERN (2025). https://doi.org/10.17181/y2499-6bf38 . https://doi.org/10.17181/y2499-6bf38

[6] Zangari, F.: The pixel detector of the TWOCRYST experiment at the LHC. University of Milan (2025)

[7] Maccani, C., Valentino, G., Korcyl, K., Hermes, P., Jakobsen, S.: The TWOCRYST fibre tracker: A detector to characterize precession crystals at the LHC. JACoW IPAC **2025**, 020 (2025) https://doi.org/10.18429/JACoW-IPAC2025-MOPS020

[8] Edelen, A., Neveu, N., Frey, M., Huber, Y., Mayes, C., Adelmann, A.: Machine learning for orders of magnitude speedup in multiobjective optimization of particle accelerator systems. Phys. Rev. Accel. Beams **23**(4), 044601 (2020)

[9] Bruchon, N., Fenu, G., Gaio, G., Lonza, M., O'Shea, F.H., Pellegrino, F.A., Salvato, E.: Basic reinforcement learning techniques to control the intensity of a seeded free-electron laser. Electronics **9**(5), 781 (2020)

[10] Eichler, A., Bründermann, E., Burkart, F., Kaiser, J., Kuropka, W., Santamaria Garcia, A., Stein, O., Xu, C.: First Steps Toward an Autonomous Accelerator, a Common Project Between DESY and KIT. In: 12th International Particle Accelerator Conference (2021). https://doi.org/10.18429/JACoW-IPAC2021-TUPAB298

[11] Aaij, R., *et al.*: The LHCb Upgrade I. Journal of Instrumentation **19**(05), 05065 (2024) https://doi.org/10.1088/1748-0221/19/05/P05065

[12] Holfe, W., Valuch, D., Assmann, R., Redaelli, S., Schmidt, R., Wollmann, D., Zerlauth, M.: Emittance blow-up and loss maps in the LHC using the transverse damper as exciter (2012)

[13] Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., Cola, G.D., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J.J., Tan, H., Younis, O.G.: Gymnasium: A Standard Interface for Reinforcement Learning Environments (2024). https://arxiv.org/abs/2407.17032

[14] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv [cs.LG] (2017) [cs.LG]

[15] Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., Wu, Y.: Stable Baselines. GitHub (2018). https://github.com/hill-a/stable-baselines