

Real-time reinforcement learning on coupling correction

Yihao Gong, Shunqiang Tian^{✉,*} and Linglong Mao

*Shanghai Advanced Research Institute Chinese Academy of Sciences, Shanghai, China
and Shanghai Institute of Applied Physics Chinese Academy of Sciences, Shanghai, China*

Xinzhong Liu, Xuan Shouzhi[✉], and Liyuan Tan[✉]

Shanghai Advanced Research Institute Chinese Academy of Sciences, Shanghai, China



(Received 11 January 2026; accepted 4 September 2026; published 20 October 2026)

The stability of vertical emittance is a critical performance metric for modern high-brightness synchrotron light sources. This paper presents a real-time, reinforcement learning (RL)-based correction framework based on deep reinforcement learning by employing dueling deep Q-network with a discrete action space to dynamically adjust skew quadrupole magnets. To ensure operational safety and accelerate convergence, we introduce a curriculum learning strategy and leverage a high-fidelity simulation of the Shanghai Synchrotron Radiation Facility for systematic hyperparameter optimization. Experimental results demonstrate that the agent, trained *in situ* on the real machine, successfully achieves stabilization of the beam-size readback at the submicrometer fluctuation level and exhibits rapid adaptability to target changes. This work validates the feasibility of using RL for robust, online accelerator optimization during user beam delivery.

DOI: [10.1103/49ns-fryv](https://doi.org/10.1103/49ns-fryv)

I. INTRODUCTION

High-brightness synchrotron radiation sources have become indispensable tools for a wide range of scientific disciplines, encompassing materials science, biology, and industry. The scientific performance of these storage ring light sources hinges critically on the stability and precise control of beam parameters, particularly the beam emittance and transverse coupling. With the advent of fourth-generation diffraction-limited storage rings, such as ESRF-EBS, MAX IV, ALS-U, and HEPS [1–4], achieving and maintaining ultralow emittance [5,6] and designed coupling ratio has become a paramount design and operational goal. However, operating under such stringent conditions introduces substantial challenges in maintaining beam size stability against dynamic disturbances and varying user configurations.

A persistent challenge in these systems is the fluctuation of brightness caused by insertion devices (IDs). The operation of IDs introduces skew quadrupole fields, which inevitably perturb the beam coupling and degrade the vertical emittance [7–9]. Traditional compensation strategies predominantly rely on feedforward correction

schemes [8] derived from lookup tables. While effective to a degree, these methods are time-consuming to commission and struggle to adapt to machine drift or complex, nonlinear interactions between multiple IDs, often resulting in residual coupling errors.

To address these limitations, data-driven approaches, particularly machine learning, have recently been explored to augment conventional correction methods [10]. A prominent example is the work by Leemann *et al.* at the Advanced Light Source (ALS), where a neural network-based feedforward system was successfully deployed to stabilize source properties based on historical machine data [11,12]. While such supervised learning models have proven highly effective in compensating for reproducible perturbations captured within their training datasets, they are inherently reliant on the quality and coverage of that data. Consequently, these static models may require retraining or recalibration when facing significant machine drifts or novel operational conditions that deviate from the historical distribution. Bayesian optimization (BO) has also been successfully applied to accelerator tuning and online beam optimization [13–16]. In the present work, we focus on RL not because it is assumed to be universally superior to BO, but because the target problem is formulated as a sequential online feedback task under evolving machine conditions. In this setting, RL is attractive because it can learn a state-dependent control policy for repeated local corrections, whereas BO is more naturally suited to iterative operating-point optimization. Moreover, BO typically requires exploratory sampling of the

*Contact author: tiansq@sari.ac.cn

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

objective landscape, which can be difficult to perform during user operation when one seeks nonintrusive stabilization rather than repeated perturbative search.

In contrast, reinforcement learning (RL) [17,18] offers a paradigm where an agent learns an optimal control policy through continuous interaction with the accelerator environment, without the need for labeled training data. This makes RL particularly suitable for the dynamic and complex problem of coupling correction. Recent accelerator applications of RL and related adaptive control methods include tune and coupling feedback in the LHC [19], online policy-gradient control at FERMI [20], and FPGA-deployed real-time fast orbit feedback at KARA [21]. Related online accelerator-control approaches also include model-independent extremum-seeking methods for beam optimization and recent model-based RL formulations for higher-dimensional feedback problems [22,23].

This work presents an application of a dueling deep Q-network (DQN) [24,25] for real-time, online correction of vertical coupling in synchrotron storage rings. The proposed method utilizes beam size measurements and vertical dispersion estimates as state observations. To shorten the training time and improve sample efficiency, we employ a multidimensional discrete-action space. Rather than relying on offline training, we leverage a high-fidelity simulation environment based on the Shanghai Synchrotron Radiation Facility (SSRF) [26] lattice to perform extensive hyperparameter optimization. The identified robust hyperparameters are then applied to the live agent, which learns the control policy from scratch on the real machine. This approach is validated via experiments at the SSRF machine, demonstrating that the RL agent can adaptively stabilize the beam-size readback at the submicrometer fluctuation level, as a preparatory study toward the tighter beam-quality control demands expected in future fourth-generation light sources.

II. COUPLING CORRECTION STRATEGY

The stability of photon beam brightness is a crucial performance metric for modern high-brightness synchrotron light sources, and even more so for the emerging fourth-generation diffraction-limited storage rings. The photon brightness is inversely proportional to the vertical beam size σ_y [27,28]. This parameter is sensitive to both the vertical emittance ε_y and the local vertical dispersion η_y , as described by

$$\sigma_y(s) = \sqrt{\beta_y(s)\varepsilon_y + [\eta_y(s)\sigma_\delta]^2}, \quad (1)$$

where β_y is the vertical beta function and σ_δ is the energy spread. To maintain the designed ultralow emittance and ensure consistent user experiments, precise real-time correction of vertical emittance growth and spurious vertical dispersion is essential.

TABLE I. Comparison between conventional optimization and the proposed RL-based optimization.

Conventional correction	RL-based correction
Relies on invasive measurements (e.g., ORM) requiring beam interruption	Performs nonintrusive online correction during user beam delivery
Based on static linear models (vulnerable to model mismatch)	Data-driven and RL policy (learns from real nonlinear machine response)
Iterative calculation (slower convergence)	Direct policy inference (real-time response)

However, real-time coupling correction presents significant challenges for traditional control strategies. Conventional global correction methods, such as those based on the orbit response matrix (ORM) or LOCO [29], are typically iterative and invasive, requiring beam interruptions that are incompatible with user beam delivery. On the other hand, feedforward compensation using lookup tables often struggles with the complex nonlinearities introduced by the varying gaps and phases of multiple IDs. Table I summarizes the key differences between these conventional approaches and the proposed RL framework.

To address these limitations, we employ a data-driven approach. A critical practical constraint in storage rings is the scarcity of diagnostics; typically, highly sensitive beam-size monitors (e.g., interferometers) are installed at only one or two locations. As illustrated in Fig. 1, a correction strategy relying solely on minimizing σ_y at a single observation point can effectively reduce the local beam size (orange line). However, this "local" optimization often

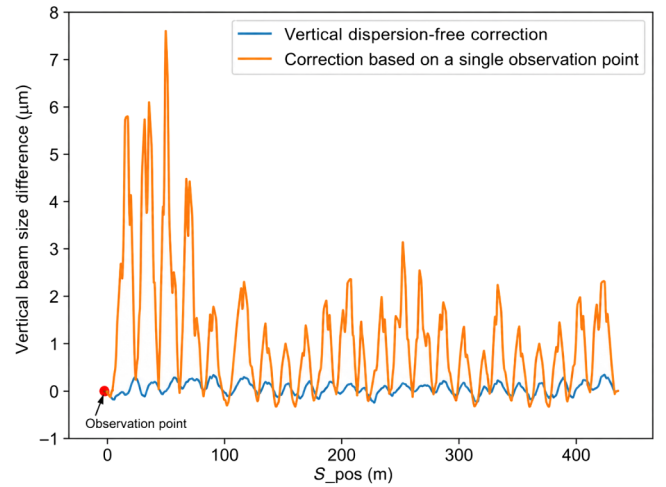


FIG. 1. Comparison of vertical beam size distribution along the storage ring. The local correction method (orange) effectively minimizes σ_y at the observation point ($s = 0$) but induces significant oscillations globally due to uncompensated vertical dispersion. In contrast, the global correction scheme (blue) achieves a uniform suppression of beam size across the entire ring.

fails to suppress—and may even amplify—the vertical dispersion elsewhere in the ring, leading to global emittance degradation [30].

Therefore, the correction scheme must be formulated as a multiobjective optimization problem that infers global stability from limited observations. The objective is to dynamically adjust the skew quadrupole strengths to simultaneously minimize the global vertical dispersion and maintain the vertical beam size at a specific target value σ_{target} . It is important to note that σ_{target} represents a calibrated operational setpoint chosen to ensure consistent flux density for users, rather than the theoretical minimum limit of the lattice. We address this complex control task using a reinforcement learning framework, where an agent learns to balance these coupled objectives through continuous interaction with the machine.

III. RL FRAMEWORK AND AGENT DESIGN

To address the complexity of real-time coupling correction, we formulate the problem as a Markov decision process and solve it using an RL approach [31]. Specifically, we implement a DQN adapted for a multi-dimensional discrete action space. This section details the environment definition and the customized neural network architecture designed for this control task.

A. RL environment definition

The interaction environment is built using the Accelerator Toolbox [32] to accurately emulate the beam dynamics of the SSRF storage ring [33]. We wrap this physics engine in an OpenAI Gym-compatible interface [34] to facilitate interaction with standard RL algorithms. Figure 2 shows the SSRF DBA cell optics and the locations of skew quadrupoles on sextupoles.

Observation space: The state vector s_t is designed to provide the agent with a comprehensive snapshot of the system status, combining beam quality metrics with actuator states. It is defined as

$$s_t = [\Delta\sigma_y, \text{rms}_{\text{disp}}, k_1, k_2, \dots, k_6], \quad (2)$$

where:

- (i) $\Delta\sigma_y = \sigma_y(t) - \sigma_{\text{target}}$ represents the deviation of the vertical beam size from the target value.
- (ii) rms_{disp} denotes the global vertical dispersion proxy. Since direct dispersion measurement is invasive during user operation, we employ the premeasured response matrix $R_{y,k}$ and the active skew quadrupole settings k_i :

$$\text{rms}_{\text{disp}} = \frac{1}{F_{\text{norm}}} \sqrt{\frac{1}{N} \sum_i (R_{y,k}^i k_i)^2}, \quad (3)$$

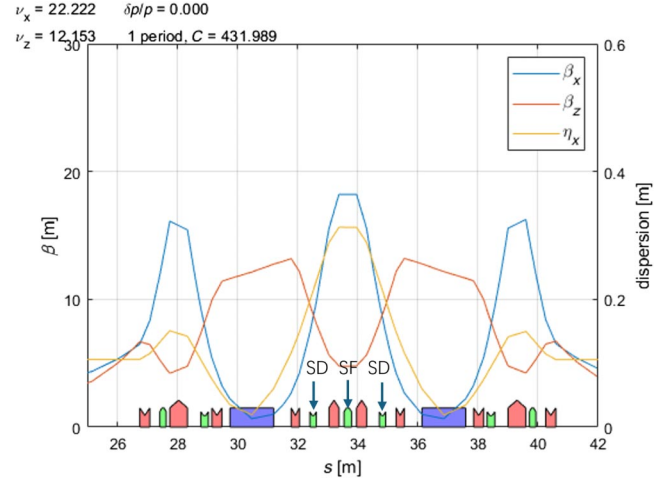


FIG. 2. Optical functions (β_x, β_y, η_x) of the SSRF lattice cell used in the simulation environment. The skew quadrupole correctors are integrated into the sextupole SD/SF shown in the lattice structure. The remaining sextupoles are used for nonlinear correction.

where F_{norm} denotes the value of the global vertical-dispersion proxy corresponding to a $1 \mu\text{m}$ global beam-size maximum change (peak to peak), obtained by introducing skew-quadrupole errors while keeping the lattice functions and vertical equilibrium emittance unchanged for the given optics configuration.

- (iii) k_i denotes the strength of the i th skew-quadrupole corrector. Specifically, in the present study, the RL action space consists of six independent skew-quadrupole control knobs, corresponding to the three SD/SF sextupole locations in SSRF cell 1 and cell 2.

Action space: To reduce the complexity of the exploration space and shorten the training duration, we define a discrete action space. Discrete action spaces significantly reduce the exploration complexity, enabling the agent to converge to a safe policy within a practical time frame. At each time step, the agent can only modify the strength of each skew quadrupole by a small, fixed increment Δk :

$$\mathcal{A} = \{a_1, a_2, \dots, a_6\}, \quad a_i \in \{-1, 0, +1\}. \quad (4)$$

Here, actions $-1, 0, +1$ correspond to decreasing, maintaining, or increasing the magnet strength by Δk (set to $2 \times 10^{-4} \text{ m}^{-2}$), respectively. Δk was chosen as a conservative minimum actuation step such that a single skew adjustment still produces an observable beam-size response at the resolution level of the online diagnostic. This results in a total action space size of $3^6 = 729$.

Reward: To guide the agent toward the multiobjective goal while accelerating convergence, we design a composite reward function r_t . The reward at time step t is

formulated as

$$r_t = -(|\Delta\sigma_y| + \text{rms}_{\text{disp}}) + B(|\Delta\sigma_y|, \text{rms}_{\text{disp}}), \quad (5)$$

$$B(\cdot) = \begin{cases} 3.0, & |\Delta\sigma_y| < 0.3 \text{ } \mu\text{m} \text{ \& } \text{rms}_{\text{disp}} < 0.2, \\ 1.0, & |\Delta\sigma_y| < 1.0 \text{ } \mu\text{m} \text{ \& } \text{rms}_{\text{disp}} < 0.3, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where the first term penalizes state error, the second term rewards incremental convergence (progress). The term $B(\cdot)$ is a discrete bonus function that returns positive values only when the beam parameters satisfy strict tolerance thresholds. This hybrid design is essential for effective lattice control: the continuous penalties provide dense gradients to guide the agent rapidly out of poor states, while the sparse bonus acts as a strong attractor to prevent policy oscillation and lock the beam size precisely at the target.

Episode termination: Each episode was capped at 600 environment steps in the main experiments. An episode terminated earlier if the absolute beam-size deviation or if the global dispersion proxy exceeded the safety bound. In the implementation, these conditions correspond to $|\Delta\sigma_y| > 3.0 \text{ } \mu\text{m}$ or $\text{rms}_{\text{disp}} > 0.5$.

B. Dueling network architecture

Standard DQN architectures often struggle to differentiate between state values and the specific advantages of individual actions, particularly in high-dimensional spaces. To overcome this, we employ a dueling network architecture, which explicitly separates these estimates.

As depicted in Fig. 3, our network features a customized split-input design:

- (1) Input encoders: The state vector is split into two streams. A beam dynamics encoder processes the

error signals ($[\Delta\sigma_y, \text{rms}_{\text{disp}}]$), while a skew encoder processes the magnet settings ($[k_1, \dots, k_6]$). This separation allows the network to extract distinct features for the beam status and the actuator configuration.

- (2) Feature fusion: The outputs of these encoders are concatenated and passed through a shared fully connected layer to form a global feature representation.
- (3) Dueling heads: The network then branches into:
 - (a) A single value head, $V(s)$, estimating the overall quality of the current state.
 - (b) Six independent advantage heads, one for each skew quadrupole. Each head i outputs a vector $A_i(s, a)$ representing the relative utility of the three possible actions $\{-1, 0, +1\}$ for that specific magnet.

The final Q value for action a on magnet i is aggregated using the dueling formula:

$$Q_i(s, a) = V(s) + \left(A_i(s, a) - \frac{1}{|\mathcal{A}_i|} \sum_{a' \in \mathcal{A}_i} A_i(s, a') \right). \quad (7)$$

This architecture allows the agent to learn which specific magnets are most effective for correcting a given error, significantly enhancing learning efficiency and control precision.

The network was trained using an experience replay buffer with a capacity of 100,000 transitions. In the final optimized setting, target-network synchronization was implemented by soft updates with $\tau = 0.001$ and an update interval of 8 environment steps, while action selection followed an ϵ -greedy strategy with $\epsilon_0 = 1.0$, $\epsilon_{\min} = 0.01$, and $\epsilon_{\text{decay}} = 0.995$, the decay applied once at the end of each training episode. ReLU activations with LayerNorm

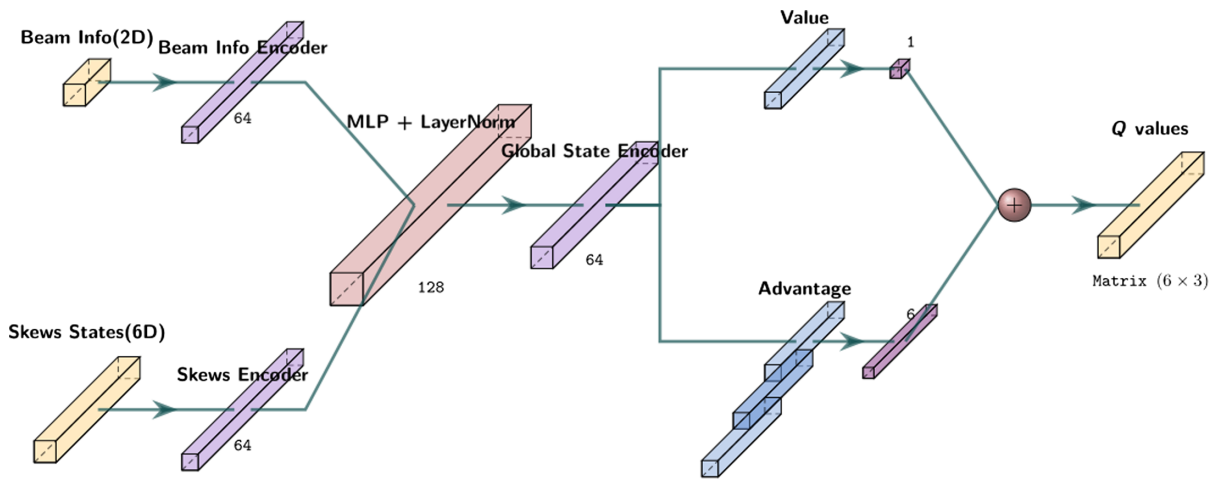


FIG. 3. Schematic of the dueling DQN architecture. The network splits the state input into beam dynamics metrics and actuator settings, processes them via separate encoders, and aggregates the values using a shared multilayer perceptron (MLP), i.e., the shared fully connected layer, before the dueling head outputs discrete action commands.

were used throughout the network with a dropout layer of probability 0.1 applied in the physics-constrained feature block, and optimization was performed using Adam with weight decay 10^{-5} and Huber loss.

IV. SIMULATION STUDY AND HYPERPARAMETER OPTIMIZATION

Before deploying the agent on the real machine, we conducted extensive studies in the simulation environment. This phase served two critical purposes: optimizing the agent's hyperparameters to ensure robust convergence and validating the efficacy of the curriculum learning strategy [35].

A. Curriculum learning strategy

To ensure convergence efficiency while strictly adhering to machine protection constraints, we employ a curriculum learning framework that progressively refines the optimization landscape. Related accelerator RL studies have also used progressively tightened success or termination thresholds during training, as well as simulation-assisted sim-to-real workflows [36,37]. This approach incorporates physics-informed constraints by implementing an adaptive deadband mechanism within the control loop. As illustrated in Fig. 4, the agent executes a null actuation (i.e., maintaining the current magnet setpoints, $a_i = 0$) whenever the observed beam state falls within the dynamic acceptance region defined by

$$\begin{aligned} |\Delta\sigma_y| &< \Delta\sigma_0\tau(t) \\ \text{rms}_{\text{disp}} &< \text{rms}_{\text{disp}0}\tau(t). \end{aligned} \quad (8)$$

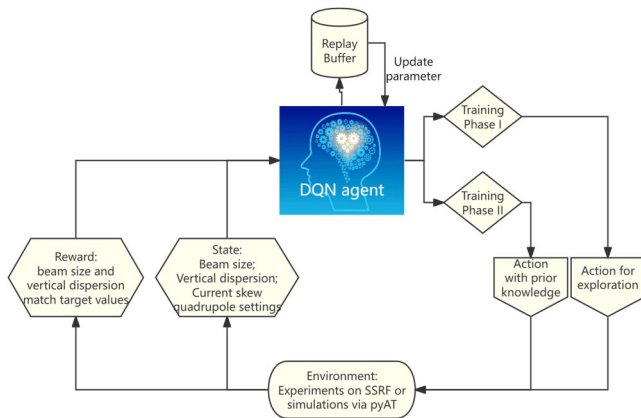


FIG. 4. Overview of the curriculum learning workflow. The agent interacts with the environment to update its policy via the replay buffer. The training process is structured into three phases: pretraining, enhanced, and fine-tuning. In the latter two phases, the "Action with prior knowledge" branch imposes dynamic thresholds to strictly limit beam size fluctuations.

For episode $t \geq t_0$, this threshold multiplier $\tau(t)$ is updated as

$$\tau(t) = \max(\tau_{\text{init}} - \beta(t - t_0), \tau_{\text{min}}), \quad (9)$$

where t_0 marks the start of the decay phase, τ_{init} is the starting threshold, β is the constant decay rate, and τ_{min} is the lower bound plateau for this phase.

Here, $\tau(t)$ denotes time-dependent thresholds that decay monotonically during the training process. This strategy effectively governs the transition from a coarse-grained correction phase—focused on global stability—to a high-precision feedback regime, thereby minimizing unnecessary magnet jitter once submicron beam quality is achieved.

B. Hyperparameter optimization

The performance of deep RL is notoriously sensitive to hyperparameters. Tuning these on a physical accelerator is impractical due to limited beam time. Therefore, we leveraged the high-fidelity simulation to perform a systematic hyperparameter optimization (HPO) [38].

We evaluated key parameters, including the learning rate, discount factor (γ), and batch size. Figure 5 visualizes the results of this search using a parallel coordinates plot.

The analysis identified a robust set of hyperparameters (learning rate = 3×10^{-4} , $\gamma = 0.995$, batch size = 128). As indicated by the high-scoring clusters in Fig. 5, this configuration maximized the training stability, allowing the agent to consistently achieve cumulative rewards exceeding 640 within the stabilization phase. These validated values were subsequently frozen and used for the live experiments at SSRF.

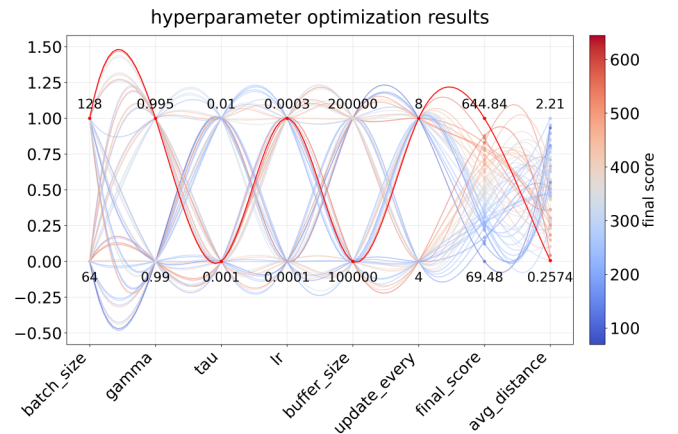


FIG. 5. HPO results in simulation. The parallel coordinates plot illustrates the correlation between various hyperparameters (e.g., learning rate, batch size) and the final training score (color scale). We selected the configuration that yielded the highest stable reward for the on-machine experiments.

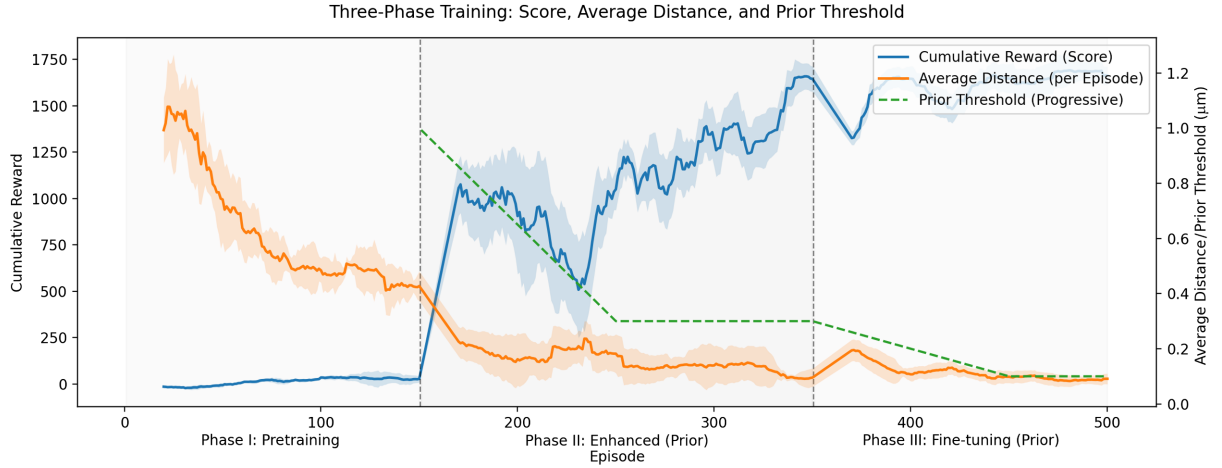


FIG. 6. Training progression of the dueling DQN agent in simulation. The plot shows the smoothed cumulative reward (blue) and the average beam size deviation (orange) over 500 episodes. The progressive tightening of the curriculum thresholds (dashed line) effectively guides the agent toward convergence. The results are obtained from 10 independent full training runs with different random seeds. A 20-episode moving average is first applied to each training trajectory; the solid curves represent the across-run mean, and the shaded bands indicate ± 1 standard deviation across the 10 smoothed trajectories.

C. Simulation results

The training progression of the agent in the simulation environment is shown in Fig. 6. The learning curve demonstrates three distinct phases corresponding to the curriculum schedule, and the specific phase-wise settings of τ_{init} , β , and τ_{min} used in the present implementation are summarized in Table II: (i) Pretraining phase (episodes 0–150): The agent explores the state space, resulting in fluctuating rewards. In the present implementation, the transition to the next phase is scheduled after 150 episodes as an empirical choice based on the observed flattening of the average distance reduction. (ii) Enhanced phase (episodes 150–350): Prior thresholding is introduced in a stage-wise manner. In this phase, gradual threshold reduction is combined with short optimization intervals under locally fixed threshold levels, allowing the agent to adapt and stabilize under each acceptance band before further tightening, which leads to a rapid increase in reward. (iii) Fine-tuning phase (episodes 350–500): The threshold is further tightened while still preserving short fixed-threshold optimization windows, enabling control accuracy without destabilizing the learning process.

To further illustrate the control performance, Fig. 7 presents the detailed beam evolution during a single validation episode after training.

TABLE II. Phase-wise threshold schedule used in curriculum learning.

Phase	Episodes	τ_{init}	β (/100 episodes)	τ_{min}
I	0–150	1.0	0	1.0
II	150–350	1.0	0.7	0.3
III	350–500	0.3	0.2	0.1

The comparison clearly highlights the agent’s capability to actively compensate for dynamic errors, achieving a stability significantly superior to the uncontrolled baseline.

Crucially, the simulation results confirm that the proposed strategy successfully addresses the global optimization challenge. Comparative analysis shows that the convergence episode was 413.2 ± 15.3 with curriculum learning and 1251.5 ± 73.4 without curriculum learning (mean $\pm$ standard deviation over 10 independent runs), representing a $67.0\% \pm 2.3\%$ reduction in the mean convergence episode. To quantify the benefit of curriculum learning, convergence is defined as the first training episode in which the beam remains within the prescribed target-distance band for at least 570 steps (95% of the maximum episode length of 600 steps), and this condition is maintained for 3 consecutive episodes. This criterion is adopted because a temporary reduction in the average distance alone does not necessarily indicate stable control performance, whereas sustained operation within the target band provides a more meaningful measure of successful convergence.

The discrete-action dueling DQN not only stabilized the local vertical beam size to $\sigma_y \approx 5 \mu\text{m}$ (simulation target used here as an idealized low-coupling validation case in the lattice model) but also suppressed the global vertical dispersion throughout the ring, which verifies the global coupling control discussed in Sec. II, thus providing a reliable baseline for on-machine implementation. Here, the simulation environment captures the dominant beam-response behavior of the coupling-correction task, although sensor noise, control latency, and magnet hysteresis are not explicitly modeled.

Finally, to provide a quantitative comparison with conventional correction, an SVD-based controller was additionally evaluated in the same simulation environment.

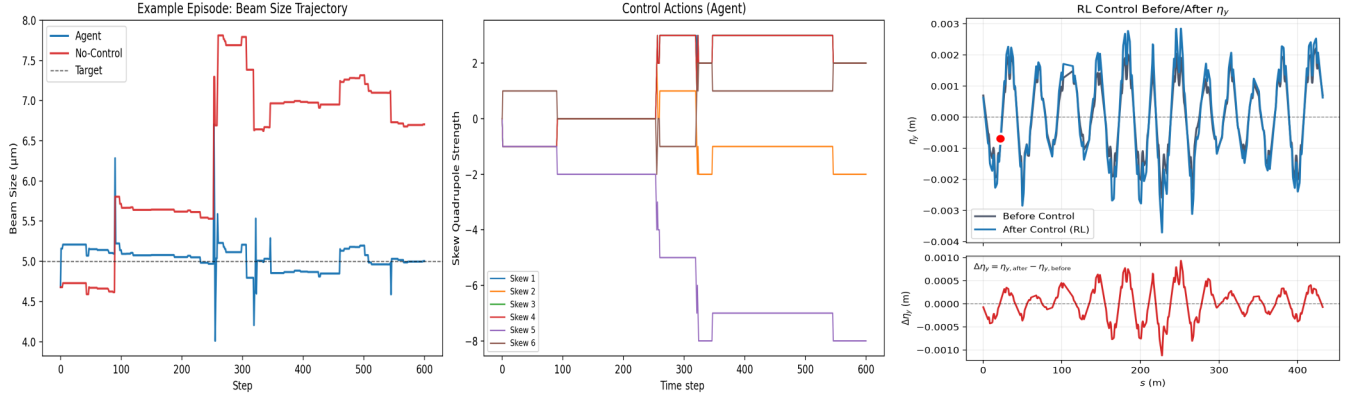


FIG. 7. Representative simulation of a single correction episode. Left: evolution of vertical beam size. The RL agent (blue) rapidly suppresses the initial error and stabilizes the beam size at the target value ($5 \mu\text{m}$), whereas the uncorrected beam (red) exhibits large fluctuations due to dynamic disturbances. Middle: the corresponding discrete control actions taken by the agent for the six skew quadrupoles, showing a distinct policy of aggressive initial correction followed by minor adjustments. Right: vertical-dispersion distribution $\eta_y(s)$ around the full ring before and after RL correction and their difference $\Delta\eta_y = \eta_{y,\text{after}} - \eta_{y,\text{before}}$ (bottom). The red marker denotes the beam-size observation point at the second bending magnet of cell 01.

A joint response matrix was constructed for the six skew quadrupoles, combining their responses to the local vertical beam size and the vertical dispersion around the full ring. The resulting SVD correction was subject to the same discrete actuator constraints as DQN. Under the same initial perturbation and 600-step disturbance sequence, the mean absolute beam-size error was $0.118 \mu\text{m}$ for DQN and $0.275 \mu\text{m}$ for SVD.

V. ON-MACHINE EXPERIMENTAL RESULTS

Following the validation in simulation, the RL correction framework was deployed on the SSRF storage ring during user operation. The primary objective was to verify the agent's ability to learn effective control policies in the real-world environment, characterized by measurement noise.

A. Experimental setup and training

The DQN agent was integrated with the SSRF control system via a PYTHON interface. It read real-time data from the diagnostic system, including the vertical beam size from the interferometer and the ORM-based dispersion proxy defined in Eq. (3), and sent discrete increment commands to the power supplies of the six skew quadrupoles. This ORM is premeasured and kept fixed during the experiment.

Crucially, the agent was not preloaded with simulation weights. Instead, it learned from scratch on the live machine, utilizing the robust hyperparameters identified in Sec. IV. This approach tests the algorithm's rapid adaptability.

Figure 8 depicts the temporal evolution of the on-machine training process over 200 episodes, clearly

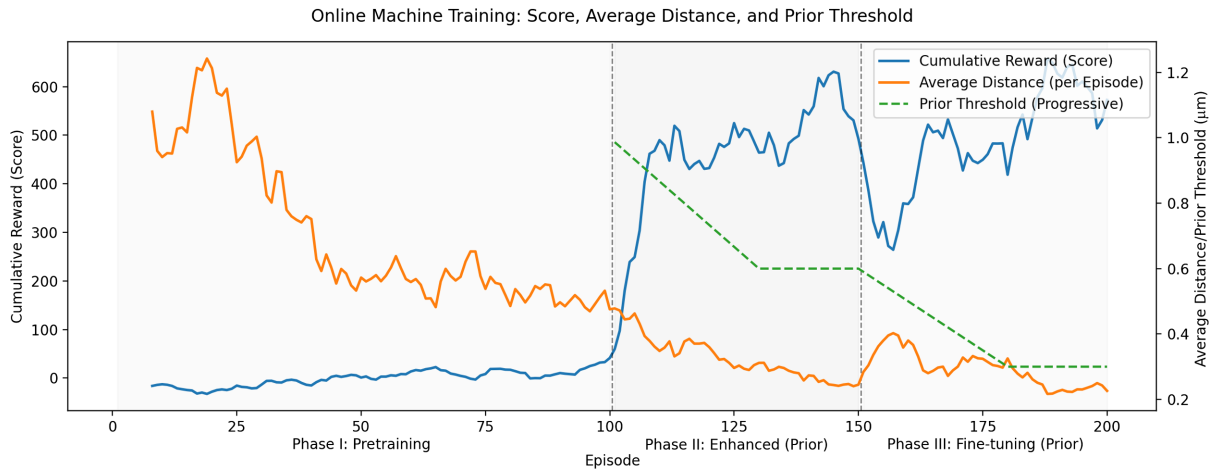


FIG. 8. On-machine training progression at SSRF. The plot shows the cumulative reward (blue) and average beam size deviation (orange) recorded during live operation. Despite initial fluctuations due to exploration, the agent converged to a stable high-reward policy within approximately 100 episodes, guided by the curriculum thresholds (dashed line).

TABLE III. Phase-wise threshold schedule used in on-machine curriculum learning.

Phase	Episodes	τ_{init}	β (/30 episodes)	τ_{min}
I	0–100	1.0	0	1.0
II	100–150	1.0	0.4	0.6
III	150–200	0.6	0.3	0.3

delineating three distinct operational regimes. The corresponding phase-wise threshold settings are summarized in Table III.

During on-machine training, each episode was limited to a maximum of 300 control steps and terminated earlier if the beam-size deviation or dispersion constraint exceeded the prescribed safety bound. The mean episode lengths were 37.0 ± 4.1 , 242.6 ± 15.0 , and 291.9 ± 3.4 steps for the pretraining, enhanced, and fine-tuning phases, respectively. At the measured closed-loop period of 4.8 s per step, these correspond to approximately 3.0, 19.4, and 23.3 min per episode. The full 200-episode on-machine training campaign therefore corresponds to approximately 30,000 control steps, or 40 h of beam time.

During phase I (pretraining, episodes 0–100), the agent engages in initial exploration. While the average distance to the target (orange solid line) decreases significantly, the cumulative reward (blue solid line) remains suppressed, indicating that the agent has not yet mastered the precision required for high rewards.

The transition to phase II (enhanced, episodes 100–150) is marked by the activation of the curriculum strategy. Here, the physics-informed acceptance threshold (green dashed line) is introduced and progressively tightened. This constraint effectively guides the agent, resulting in an increase in cumulative reward.

Finally, in phase III (fine-tuning, episodes 150–200), the threshold settles at its minimum value. The agent achieves a steady-state performance characterized by consistently high rewards and minimal rms error, demonstrating that the policy has successfully converged to a robust solution within the submicron tolerance window.

B. Steady-state stability

Once trained, the agent's performance was evaluated under steady-state conditions. Figure 9 presents the time evolution of the vertical beam size and the corresponding control actions over 4500 time steps (6 h), validating the control policy's effectiveness in terms of convergence, actuator protection, and robustness. The cycle time corresponds to an effective closed-loop period of approximately 4.8 s per step. This period is determined by power-supply response, the beam-size monitor refresh time, and the waiting time required to ensure that the observation is taken after the machine response has settled. Initially, the agent successfully converged the vertical beam size to the immediate vicinity of the target value (18 μm). Here, the baseline refers to normal SSRF operation with the insertion-device coupling feedforward remaining active, but without RL correction. Over separate 6-h evaluation windows under comparable machine conditions, the corresponding rms fluctuation was 0.15 μm under RL feedback, compared with 0.62 μm under the baseline SSRF operating condition. The 0.15 μm residual is comparable to the typical short-term unperturbed readback variation of 0.1–0.2 μm , indicating that the measured closed-loop stability approaches the diagnostic noise floor. A distinguishing feature of the control strategy, as illustrated in the right panel, is the sparse and discrete nature of the skew quadrupole actuations. This steplike behavior confirms the

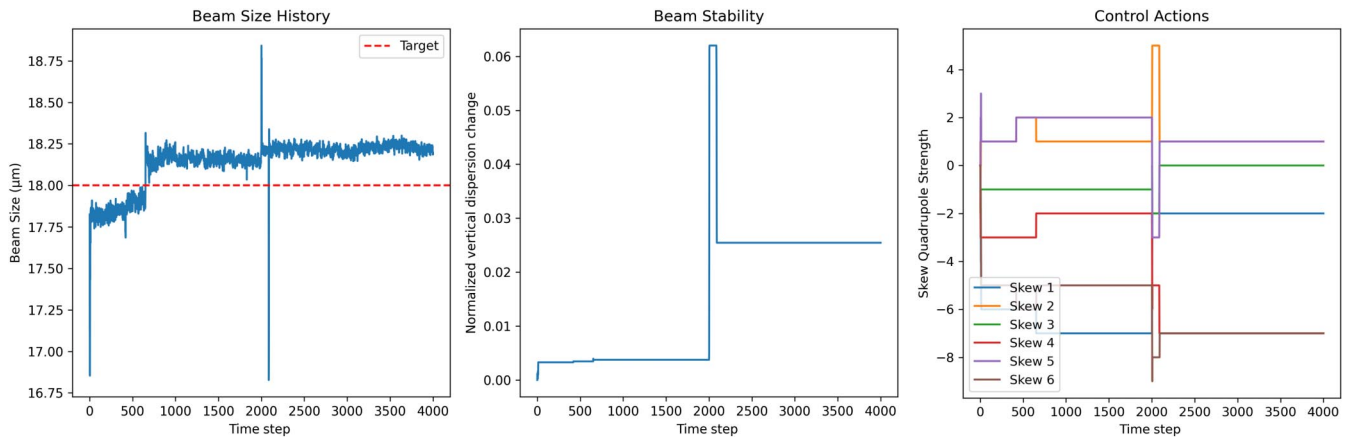


FIG. 9. Experimental verification of beam size stabilization. Left: the vertical beam size (blue) is tightly maintained around the target (red dashed line) by the RL agent. Middle: the normalized vertical dispersion change remains minimal. Right: history of skew quadrupole actions. The agent effectively identifies and applies necessary corrections while maintaining a stable configuration for extended periods. The step at $t = 2000$ corresponds to a scheduled ID movement for testing.

successful deployment of the adaptive deadband mechanism (described in Sec. IV A), which prevents high-frequency chattering and thereby mitigates fatigue in the magnet power supplies. Furthermore, to rigorously evaluate the agent's adaptability to dynamic lattice perturbations, the configuration of an ID was intentionally altered at step $t = 2000$, resulting in a beam size change. In response to this external disturbance, the agent immediately identified the state change and re-optimized the skew quadrupole strengths to a new equilibrium. This rapid compensation demonstrates the policy's capability to autonomously counteract ID-induced linear optics distortions and maintain beam stability without human intervention.

The decision to employ a discrete action space rather than a continuous policy (e.g., proximal policy optimization or soft actor-critic) was driven by the strict constraints of online accelerator operation. Training an agent on a live machine is extremely expensive in terms of beam time. Continuous action spaces typically require significantly more samples to explore the infinite dimensionality of the control output effectively. By quantizing the actions into $\{-1, 0, +1\}$ increments, we drastically reduced the exploration volume, enabling the agent to converge to a safe policy. For the six-skew-quadrupole task considered here, the resulting $3^6 = 729$ joint actions remain tractable in both simulation and online operation. However, the joint action space scales as 3^n with the number of independently controlled correctors and would become prohibitive for substantially larger systems. In such cases, more structured control formulations, such as factorized, hierarchical, or model-assisted approaches, would be required to improve scalability.

Our steady-state results confirm that with the deadband mechanism, this discretization does not compromise the submicron stability required for user experiments.

C. Adaptability to target changes

A key advantage of the proposed RL framework is its operational flexibility. In routine user operations, the storage ring may need to switch between different coupling modes (e.g., high-brightness mode vs timing mode) to satisfy varying experimental requirements. This necessitates a control policy capable of tracking dynamic targets rather than being fixed to a single setpoint.

To validate this capability, we dynamically changed the target beam size from 18 to 20 μm during active beam delivery. As shown in Fig. 10, the agent responded immediately to the step change in the target signal. Although brief transient fluctuations were observed immediately before the setpoint change—attributed to the agent's exploration of the gradient in the new state space—these oscillations were rapidly suppressed. Without any manual intervention or retraining, the policy autonomously adjusted the skew quadrupoles to converge to the new setpoint.

This step response confirms that the learned policy captures the underlying physics of the coupling response. However, it is worth noting that this adaptability is most effective for perturbative adjustments within the local linear region of the lattice optics (e.g., adjustment ranges of a few micrometers). For large-scale deviations (e.g., increasing the beam size from 18 to 100 μm), the required magnet strengths might exceed the incremental action

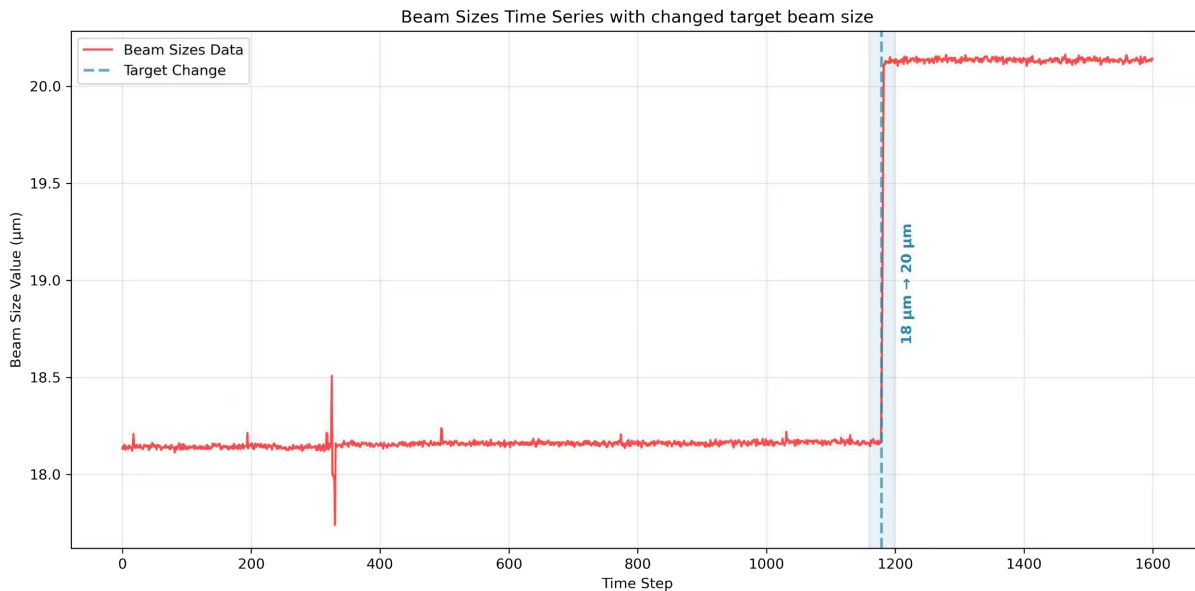


FIG. 10. Step response of the RL agent to a change in target beam size. When the target setpoint was increased from 18 to 20 μm (blue dashed line), the agent autonomously adjusted the magnets to track the new target (red line). Despite minor initial transients, the beam size stabilized at the new value with a fast settling time. The spike around step 330 was caused by a manual ID movement during the experiment and represents the beam response to an external perturbation rather than the nominal steady-state fluctuation.

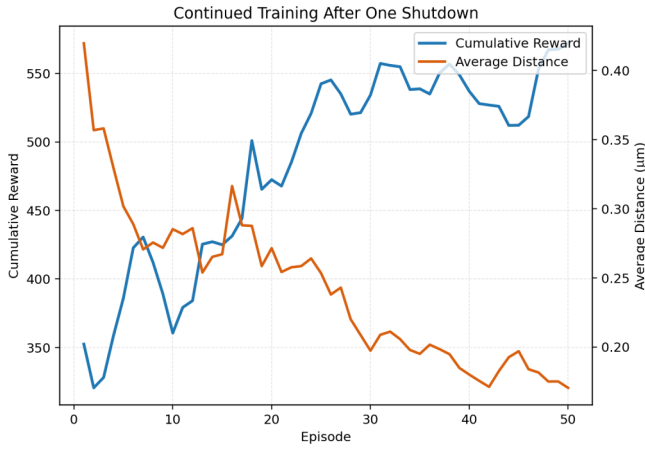


FIG. 11. Continued on-machine training after one winter shutdown and restart. The agent recovers useful performance within about 30–50 episodes, as indicated by the increasing cumulative reward and decreasing average distance.

limits or push the lattice into a significantly different nonlinear regime. In such scenarios, the recommended procedure is to first reconfigure the lattice to approach the new target, and then restart the RL agent for feedback correction. For the moderate optics changes encountered in our operation, such as after a shutdown and restart, the measured response matrices before and after the optics change remained nearly identical. Under these conditions, continued on-machine training allowed the agent to recover useful performance within about 30–50 episodes, requiring approximately 18 h of additional beam time, without requiring a new simulation campaign or re-optimization of hyperparameters, as illustrated in Fig. 11. For substantially different optics modes, however, this transferability cannot be assumed. In such cases, the response characterization, target setting, and training conditions may need to be re-established and revalidated. At the same time, we do not claim that the present DQN-based framework solves the general problem of continual learning under strong or arbitrarily long-term drift. If the machine response changes too far from the regime represented by recent experience, replay-based Q-learning may become less reliable.

VI. CONCLUSION

In this work, we have developed and experimentally validated a reinforcement learning framework for real-time coupling correction in storage rings. By reformulating the global minimization of vertical beam size and dispersion as a sequential decision-making problem, we demonstrated that a dueling DQN agent can autonomously learn effective control policies in a model-free manner at the policy-learning level, while using a premeasured ORM-based dispersion proxy as part of the observation.

A key contribution of this study is the integration of a high-fidelity simulation for hyperparameter optimization and curriculum design. This strategy effectively bridged the gap between simulation and reality, allowing the agent to learn from scratch on the physical machine with high stability and convergence speed. Experimental results at the SSRF confirmed that the proposed agent could stabilize the vertical beam size at the micrometer level and rapidly adapt to varying target constraints during user operation.

Compared to traditional response-matrix-based methods, the RL approach offers a nonintrusive alternative that naturally handles system nonlinearities. Future work will extend this framework to continuous action spaces for smoother control and explore multiagent reinforcement learning for the simultaneous optimization of multiple beam parameters, paving the way for fully autonomous accelerator operations.

DATA AVAILABILITY

The data supporting this study are not available for sharing.

- [1] P. Raimondi, N. Carmignani, L. R. Carver *et al.*, Commissioning of the hybrid multibend achromat lattice at the European synchrotron radiation facility, *Phys. Rev. Accel. Beams* **24**, 110701 (2021).
- [2] P. F. Tavares, E. Al-Dmour, Åke Andersson, F. Cullinan, B. N. Jensen, D. Olsson, D. K. Olsson, M. Sjöström, H. Tarawneh, S. Thorin, and A. Vorozhtsov, Commissioning and first-year operational results of the MAX IV 3 GeV ring, *J. Synchrotron Radiat.* **25**, 1291 (2018).
- [3] C. Steier, A. Allézy, A. Anders *et al.*, Status of the conceptual design of ALS-U, in *Proc. 9th Int. Particle Accelerator Conf. (IPAC'18), Vancouver, BC, Canada* (JACoW Publishing, Geneva, Switzerland, 2018), pp. 4136–4139.
- [4] Y. Jiao, G. Xu, X. H. Cui *et al.*, The HEPS project, *J. Synchrotron Radiat.* **25**, 1611 (2018).
- [5] M. Aiba, M. Böge, and A. Streun, Ultra-low vertical emittance at SLS, in *Proc. 3rd Int. Particle Accelerator Conf. (IPAC'12), New Orleans, LA, USA* (JACoW Publishing, Geneva, Switzerland, 2012), pp. 843–845.
- [6] X. Liu, S. Tian, L. Tan, Y. Gong, S. Xuan, and Z. Zhao, Applications of vertical damping wigglers in an x-ray diffraction limited storage ring, *Nucl. Instrum. Methods Phys. Res., Sect. A* **1056**, 168653 (2023).
- [7] A. Franchi, L. Farvacque, J. Chavanne, F. Ewald, B. Nash, K. Scheidt, and R. Tomás, Vertical emittance reduction and preservation in electron storage rings via resonance driving terms correction, *Phys. Rev. ST Accel. Beams* **14**, 034002 (2011).
- [8] X.-Z. Liu, S.-Q. Tian, X. Wu, M. Wang, Z.-T. Zhao, and B.-Y. Feng, Feedforward compensation of the insertion

- devices effects in the SSRF storage ring, *Nucl. Sci. Tech.* **33**, 70 (2022).
- [9] Y. Hidaka, Y. Li, T. Shaftan, T. Tanabe, Y. Tian, and G. Wang, Correction of ID-induced transverse linear coupling at NSLS-II, in *9th International Particle Accelerator Conference* (JACoW Publishing, Geneva, Switzerland, 2018).
- [10] A. L. Edelen, S. G. Biedron, B. E. Chase, D. Edstrom, S. V. Milton, and P. Stabile, Neural networks for modeling and control of particle accelerators, *IEEE Trans. Nucl. Sci.* **63**, 878 (2016).
- [11] S. C. Leemann, S. Liu, A. Hexemer, M. A. Marcus, C. N. Melton, H. Nishimura, and C. Sun, Demonstration of machine learning-based model-independent stabilization of source properties in synchrotron light sources, *Phys. Rev. Lett.* **123**, 194801 (2019).
- [12] T. Hellert, T. Ford, S. C. Leemann, H. Nishimura, M. Venturini, and A. Pollastro, Application of deep learning methods for beam size control during user operation at the advanced light source, *Phys. Rev. Accel. Beams* **27**, 074602 (2024).
- [13] J. Duris, D. Kennedy, A. Hanuka, J. Shtalenkova, A. Edelen, P. Baxevanis, A. Egger, T. Cope, M. McIntire, S. Ermon, and D. Ratner, Bayesian optimization of a free-electron laser, *Phys. Rev. Lett.* **124**, 124801 (2020).
- [14] R. Roussel, A. Hanuka, and A. Edelen, Multiobjective Bayesian optimization for online accelerator tuning, *Phys. Rev. Accel. Beams* **24**, 062801 (2021).
- [15] R. Roussel *et al.*, Bayesian optimization algorithms for accelerator physics, *Phys. Rev. Accel. Beams* **27**, 084801 (2024).
- [16] J. Kaiser, C. Xu, A. Eichler, A. S. Garcia, O. Stein, E. Bründermann, W. Kuroepka, H. Dinter, F. Mayet, T. Vinatier, F. Burkart, and H. Schlarb, Reinforcement learning-trained optimisers and Bayesian optimisation for online particle accelerator tuning, *Sci. Rep.* **14**, 15733 (2024).
- [17] V. Kain, S. Hirlander, B. Goddard, F. M. Velotti, G. Z. Della Porta, N. Bruchon, and G. Valentino, Sample-efficient reinforcement learning for CERN accelerator control, *Phys. Rev. Accel. Beams* **23**, 124801 (2020).
- [18] X. Chen, Y. Jia, X. Qi, Z. Wang, and Y. He, Orbit correction based on improved reinforcement learning algorithm, *Phys. Rev. Accel. Beams* **26**, 044601 (2023).
- [19] L. Grech, G. Valentino, D. Alves, and S. Hirlander, Application of reinforcement learning in the LHC tune feedback, *Front. Phys.* **10**, 929064 (2022).
- [20] F. H. O'Shea, N. Bruchon, and G. Gaio, Policy gradient methods for free-electron laser and terahertz source optimization and stabilization at the fermi free-electron laser at Elettra, *Phys. Rev. Accel. Beams* **23**, 122802 (2020).
- [21] W. Wang, M. Caselle, T. Boltz, E. Blomley, M. Brosi, T. Dritschler, A. Ebersoldt, A. Kopmann, A. Santamaria Garcia, P. Schreiber, E. Bründermann, M. Weber, A.-S. Müller, and Y. Fang, Accelerated deep reinforcement learning for fast feedback of beam dynamics at KARA, *IEEE Trans. Nucl. Sci.* **68**, 1794 (2021).
- [22] A. Scheinker, A. Edelen, D. Bohler, C. Emma, and A. Lutman, Demonstration of model-independent control of the longitudinal phase space of electron beams in the linac-coherent light source with femtosecond resolution, *Phys. Rev. Lett.* **121**, 044801 (2018).
- [23] A. Scheinker, S. Hirlander, F. M. Velotti, S. Gessner, G. Z. Della Porta, V. Kain, B. Goddard, and R. Ramjiawan, Online multi-objective particle accelerator optimization of the awake electron beam line for simultaneous emittance and orbit control, *AIP Adv.* **10**, 055320 (2020).
- [24] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, Human-level control through deep reinforcement learning, *Nature (London)* **518**, 529 (2015).
- [25] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, Dueling network architectures for deep reinforcement learning, in *Proceedings of the 33rd International Conference on Machine Learning (ICML)* (PMLR, New York, USA, 2016), Vol. 48, pp. 1995–2003.
- [26] H. Xu and Z. Zhao, Current status and progresses of SSRF project, *Nucl. Sci. Tech.* **19**, 1 (2008).
- [27] A. Wolski, Low-emittance storage rings, CERN, Tech. Rep. CERN-2014-009, 2014.
- [28] C. L. Li, S. Q. Tian, and W. Z. Zhang, Study on the vertical emittance for the SSRF phase II project, *Nucl. Sci. Tech.* **39**, 050102 (2016).
- [29] J. Safranek, Experimental determination of storage ring optics using orbit response measurements, *Nucl. Instrum. Methods Phys. Res., Sect. A* **388**, 27 (1997).
- [30] D. A. Edwards and L. C. Teng, Parametrization of linear coupled motion in periodic systems, *IEEE Trans. Nucl. Sci.* **20**, 885 (1973).
- [31] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, Cambridge, MA, USA, 2018).
- [32] A. Terebilo, Accelerator toolbox for MATLAB, SLAC National Accelerator Laboratory, Tech. Rep. SLAC-PUB-8732, 2001.
- [33] X. Wu, S. Tian, X. Liu, W. Zhang, and Z. Zhao, Design and commissioning of the new SSRF storage ring lattice with asymmetric optics, *Nucl. Instrum. Methods Phys. Res., Sect. A* **1025**, 166098 (2022).
- [34] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, Openai gym, [arXiv:1606.01540](https://arxiv.org/abs/1606.01540).
- [35] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, Curriculum learning, in *Proceedings of the 25th International Conference on Machine Learning (ICML)* (Association for Computing Machinery, New York, NY, USA, 2009), pp. 41–48.
- [36] F. M. Velotti, B. Goddard, V. Kain, R. Ramjiawan, G. Zevi Della Porta, and S. Hirlander, Towards automatic setup of 18 MeV electron beamline using machine learning, *Mach. Learn.* **4**, 025016 (2023).
- [37] A. Awal, J. Hetzel, R. Gebel, and J. Pretz, Injection optimization at particle accelerators via reinforcement learning: From simulation to real-world application, *Phys. Rev. Accel. Beams* **28**, 034601 (2025).
- [38] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 2623–2631.