

<https://doi.org/10.1038/s42005-025-02302-y>

High-fidelity data-driven dynamics model for reinforcement learning-based control in HL-3 tokamak

Check for updates

Niannian Wu^{1,2,4}, Zongyu Yang^{2,4}, Rongpeng Li^{1,4}✉, Ning Wei^{3,4}, Yihang Chen², Qianyun Dong^{1,2}, Jiyuan Li², Guohui Zheng², Xinwen Gong², Feng Gao³, Bo Li², Min Xu², Zhifeng Zhao^{1,3}✉ & Wulyu Zhong²✉

Reinforcement learning (RL)-based control in tokamaks offers improved flexibility for nuclear fusion, but typically depends on simulators that can accurately evolve the high-dimensional plasma state. First-principle simulators are often too computationally intensive for efficient RL training. Here, we develop a fully data-driven simulator that mitigates compounding errors caused by its autoregressive nature. This high-fidelity model enables rapid training of an RL agent that generates engineering-reasonable actuator commands to reach long-term plasma configuration targets. Combined with a neural network surrogate for equilibrium fitting (EFITNN), the agent maintains a 400-ms, 1 kHz control trajectory on the HL-3 tokamak, accurately tracking plasma current and boundary shape. It also adapts to changes in triangularity without retraining, demonstrating robustness. These results show that data-driven dynamics models can support fast and reliable RL-based control, meeting anticipated engineering requirements for routine operation in future fusion devices such as ITER.

The objective of the tokamak, a torus-shaped nuclear fusion device, is to provide sustainable energy by effectively confining high-temperature plasmas using magnetic fields. The stable confinement of plasma in a tokamak depends significantly on well-calibrated control approaches, which can be developed through a thorough understanding of the underlying dynamics^{1–3}. Rather than providing an exhaustive interpretation of fundamental physics, control-oriented dynamics models deliver significant advantages in simplicity with reasonable accuracy⁴, assisting the precise manipulation of various actuators toward desired configurations. For instance, the reliable prediction of rapidly growing plasma instabilities contributes to proactively suppressing disruptions^{5–8}, while an accurate long-term evolution of the plasma dynamics supports Reinforcement Learning (RL)-based controller to meticulously regulate high-dimensional, mutually coupled magnetic coils toward target plasma current and shape^{9,10}.

Classically, a high-fidelity, autoregression-capable dynamics model^{11,12} can be built on first-principles-based computations. A notable example is the Forward Grad-Shafranov Evolution (FGE) model, which solves the free-boundary equilibrium while accounting for the plasma current evolution coupled to controller dynamics¹³. Through accurately synthesizing high-dimensional magnetic flux and local magnetic field measurements as outputs of the tokamak sensors, this model enables the

offline training of an RL-based magnetic controller⁹. Although this physics-driven approach can give highly detailed simulations, an over-reliance on such simulators without incorporating sufficient real-world data may reduce the generalizability of RL policies¹⁴. Additionally, each training step has to await the FGE simulator for the cumbersome computations of lumped-parameter equations⁹, making the training of RL policies possibly take hours¹⁰. In contrast, ITER, the International Thermonuclear Experimental Reactor, is expected to be capable of executing the next pulse within a minimum of 30 min¹⁵. Despite progress to accelerate training, the physics-driven simulator is still not competent for training RL-based control policies at a sufficiently fast pace¹⁰, an anticipated engineering requirement in daily discharge practices for ITER.

Compared to first-principles-based approaches, historical discharge data offers the potential to develop models with significantly faster speed. For example, experimental data facilitates the adoption of linearized system identification techniques to approximately reproduce the response of coupled plasma profiles under variations of given actuators^{14,16}. Nevertheless, the validity of the linearized dynamics model is limited to the vicinity of the plasma equilibrium that needs to be tracked¹⁷. An alternative approach resorts to Artificial Intelligence (AI) models, as Deep Neural Network (DNN)-based models demonstrate satisfactory

¹Zhejiang University, Hangzhou, China. ²Southwestern Institute of Physics, Chengdu, China. ³Zhejiang Lab, Hangzhou, China. ⁴These authors contributed equally: Niannian Wu, Zongyu Yang, Rongpeng Li, Ning Wei. ✉e-mail: lirongpeng@zju.edu.cn; zhaozf@zhejianglab.org; zhongwl@swip.ac.cn

accuracy^{18–22} with preliminary dynamics modeling results^{5,11,12}. Prominently, a DNN model that competently forecasts the future tearing likelihood from various plasma profiles and tokamak actuations has successfully contributed to the prediction-driven planning of a high-level policy to yield desired beam power and plasma triangularity target for tearing avoidance⁵. Though such a process disentangles the predictive power of the dynamics model from the policy learning procedure, building a high-fidelity, data-driven simulator for training a high-dimensional, high-frequency RL controller remains challenging. Unless optimizations are adopted to make long-range predictions, the autoregressive nature of such simulators can lead to compounding errors, which gradually accumulate over long-term multiple-step evolution, resulting in significant discrepancies and potentially catastrophic outcomes²³.

To combat the long-term divergence issue, one intuitive means is to unroll the single-step prediction to incorporate multi-prediction losses during training^{24–28}; while another line of work lies in implicitly^{29–31} or explicitly^{32,33} learning a specific dynamics model for every n -step prediction. Albeit the practicability in low-dimensional scenarios^{25–32}, both strategies imply significant training difficulties with limited scalability for a magnetic control that usually lasts for hundreds or even thousands of steps. Therefore, on top of single-step prediction, several techniques like ensembling^{34,35}, principal component analysis³⁶, and noise layer^{11,12} have been leveraged to ameliorate this issue. However, to our best knowledge, the data-driven dynamics model is still far from perfect for long-term, high-dimensional trajectory control. Inevitably, Proportional-Integral-Derivative (PID) controllers^{5,11,12}, historically logged multi-variate commands^{36,37}, or online Model Predictive Control (MPC)³⁵ indispensable complement for achieving localized short-term, low-level magnetic control. Nevertheless, PID controllers imply tremendous engineering and design efforts before deployment³, while online MPC suffers from the cumbersome online autoregression of the dynamics model for trajectory sampling^{6,35}. Consequently, due to the rare readiness of either low-level controllers or physical simulators, the limited accuracy with only short-term guarantees in existing data-driven dynamic models hinders the learning of high-dimensional, long-term trajectory control policies.

In this paper, we aim to provide a data-driven dynamics model that exclusively learns from the discharge logs collected through the interaction between PID controllers and the tokamak device. The dynamics model is designed to efficiently support the fast training and practical deployment of a long-term, low-level RL-based agent, capable of directly generating actuator commands. The system architecture is shown in Fig. 1. As shown in Fig. 1a, EFITNN³⁸, a surrogate model of the magnetic equilibrium reconstruction code EFIT^{39,40}, is leveraged to meaningfully transform the high-dimensional magnetic measurements to lower-dimensional plasma shape information while maintaining the physical interpretability and supporting real-time deployment. Subsequently, a Long Short-Term Memory Network (LSTM)-based⁴¹ dynamics model is developed to track the evolution of plasma shape and current, based on accessible zero-dimensional plasma parameters and executed actuations. Benefiting from effectively leveraging several useful techniques, including scheduling sampling^{42,43}, adaptive weight adjustment^{44,45}, self-attention⁴⁶, and model ensembling⁴⁷, the trained dynamics model demonstrates exceptional generalization capability on unseen plasma current regions during training and sensitively captures shape variations in typical scenarios, such as transitioning from limiter to (advanced) divertor configuration⁴⁸, and soft ramp-down⁴⁹ during discharges. Unlike localized short-term planning^{5,11,12} shown in Supplementary Fig. 1a, this data-driven dynamics model enables fully RL-based 400-ms, 1-kHz low-level trajectory control of up to 17 magnetic coils, achieving the desired Last Closed Flux Surface (LCFS)⁵⁰ via six shape and position parameters; while, as in Supplementary Fig. 1b, the training of 250,000 iterations completes within 22 min on the customer-grade Nvidia GeForce RTX 4090 GPU, fully satisfying the anticipation for RL-based magnetic control in ITER. It even achieves zero-shot triangularity adjustment, validating the effectiveness and reliability of the data-driven dynamics model.

Results

Construction of a data-driven dynamics model

To train a 1-kHz RL-based control policy, it is essential to build a high-fidelity data-driven dynamics model, which can autoregressively produce accurate predictions for up to 25 variates. These predictive variates encompass the plasma current, 6 plasma shape and position parameters, and 18 channels of feedforward coil currents, which together correspond to a 44-dimensional indispensable input (Supplementary Table 2) and support the training of an engineering-reasonable policy for controlling 17 magnetic coil voltages. In comparison to yielding 100 ms-averaged evolution¹², forecasting single-variate tearability⁵, or only tuning the total power and torque injected from the neutral beams³⁶, the significantly higher resolution and a larger number of coupling actuators implies to pose considerable challenges and require a more comprehensive design.

During training, to mitigate the compounding errors, scheduled sampling, which blends the historical logged data and model output^{42,43}, is leveraged. Fig. 2 shows that this easy-to-implement technique ensures the training scalability and brings appealing results; in the “Methods” section, a more thorough comparison with alternative solutions, such as noise layer¹², shows the flexibility of this training approach. On top, training the dynamics model incorporates several useful techniques. In particular, adaptive weight adjustment^{44,45}, which simplifies the tuning of weights for combining multi-task loss functions⁴⁵ and guarantees resilient and steady model learning⁴⁴, is adopted to capture the implicit relationship among plasma shape and position parameters. Meanwhile, auxiliary techniques, such as self-attention⁴⁶ and model ensembling⁴⁷, further enhance accuracy through a concerted effort. Fig. 3a presents the ablation study in terms of Mean Absolute Error (MAE) and demonstrates the incremental contribution of each technique to train the dynamics model. Besides, the model maintains high consistency across different tasks, achieving at least a 58.8% reduction in MAE than the standard LSTM-based training. In addition, Fig. 3b provides the statistical measure in terms of the coefficient of determination R^2 and shows that the autoregressive output of the dynamics model can fit the actual data well.

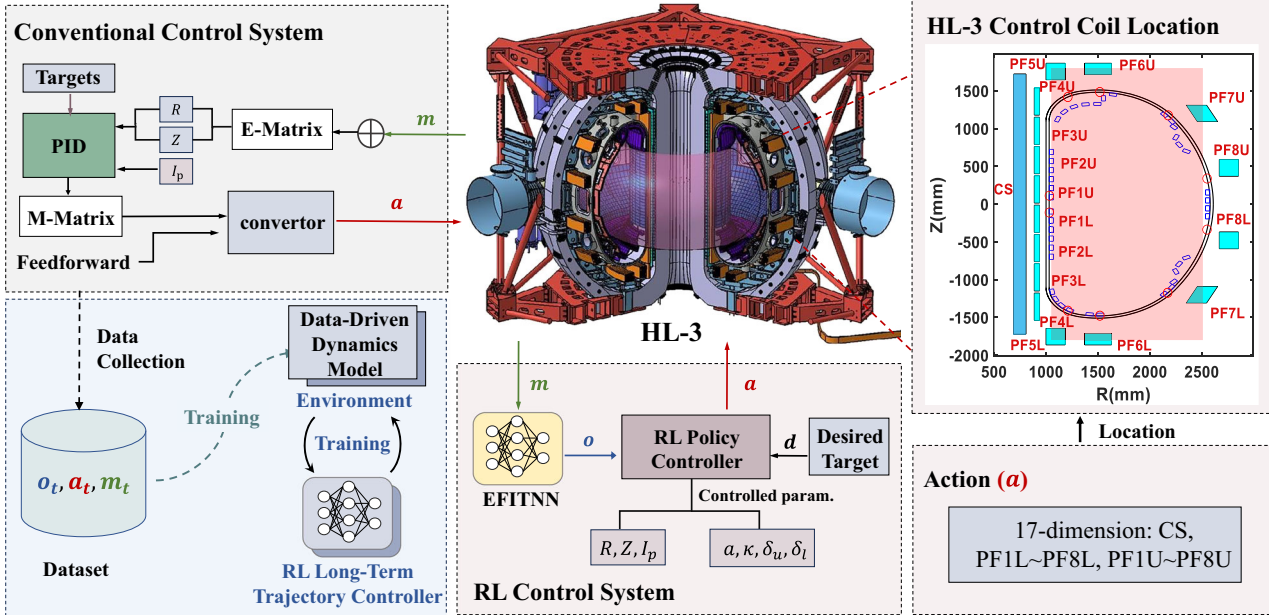
Prediction accuracy under abrupt shape variations

In the previous section, we demonstrated the capability of the data-driven dynamics model in learning the high-dimensional evolution. In this section, we switch to investigate the prediction accuracy under abrupt plasma shape variations, which are commonly encountered due to the transition from limiter to divertor shape, the advanced divertor configuration transformation, or the initialization of soft ramp-down. Taking the example of shot #6221, an H-mode discharge with edge localized modes (ELMs), Fig. 4d manifests the significant impact of these activities on the plasma LCFS. Specifically, following the shift from the limiter to the divertor configuration, the plasma’s elongation κ , upper triangularity δ_u , and lower triangularity δ_l surge abruptly. Fortunately, as shown in Fig. 4a, the dynamics model sensitively captures this variation. Similar robustness in predicting significant fluctuations in elongation and horizontal displacement can be observed in the initial segment of the ramp-down phase, where magnetic control still dominates despite the plasma’s strong interaction with the wall, producing significant non-magnetic effects. On the other hand, Fig. 4b and c validate the accuracy of the dynamics model to predict abrupt shape variations in both advanced divertor configuration and Internal Transport Barrier (ITB) discharges^{51,52}.

Extrapolation capability on unseen plasma current

Consistent with the staged research plan in ITER to gradually increase plasma current (I_p)¹⁵, it is crucial to assess the extrapolation capability of a dynamics model trained on lower- I_p datasets when applied to a higher I_p plasma discharge scenario. Therefore, we investigate the marginal contributions of incorporating various numbers of additional 700 kA discharges to fine-tune a dynamics model initially trained on data with plasma currents ranging from 300 to 600 kA.

a Deployment



b Data-Driven Dynamics Model Training

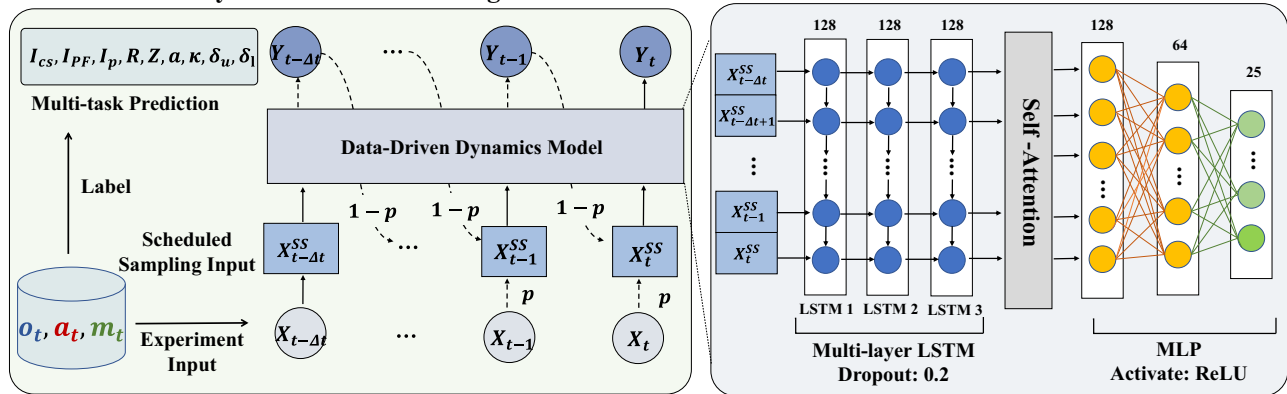


Fig. 1 | The overall architecture of the plasma current and shape control in the HL-3 tokamak. a Deployment schematic for plasma control on the HL-3 tokamak, illustrating two control paradigms. The Conventional Control System utilizes a Proportional-Integral-Derivative (PID) controller. It operates on the error between desired targets and a measured plasma state derived from the direct diagnostic measurements m via the M-Matrix, which calculates the R and Z positions for control, to ultimately generate commands that contribute to the final actuator voltages a . The Reinforcement Learning (RL) Control System employs a trained policy controller. This controller receives its state o from the diagnostic measurements m processed by the EFITNN model. This state o is composed of the plasma current I_p as well as six shape and position parameters: the minor radius a , elongation κ , upper

triangularity δ_u , lower triangularity δ_l , and the radial and vertical positions, R and Z , of the plasma geometric center. The controller then takes this state o and the desired target state d as inputs to directly generate the action a , which represents the voltage commands for the 17 magnetic coils. Historical data from the conventional system, including direct diagnostic measurements m_t , derived state observations o_t , and the executed actuator commands a_t , are collected to train the data-driven dynamics model. This model then serves as a high-fidelity simulated environment for the offline training of the RL controller. **b** The proposed training diagram of the dynamics model, including scheduling sampling, multi-task adaptive weight adjustment, self-attention, and model ensembling.

Figure 5 presents the corresponding results averaging over a testing dataset consisting of 30 shots of 700 kA discharges, where the zero-shot training implies no involvement of data from 700 kA discharges, while full-shot learning indicates using data from all available 67 shots. As depicted in Fig. 6, under zero-shot training, the model gives a 152.6 kA discrepancy and fails to reach the maximum plasma current value of 700 kA, highlighting the generalization difficulty to higher currents without any prior information. Nevertheless, with the introduction of just 2 or 4 additional shots of 700 kA data for training, the prediction accuracy improves significantly, with a 90% reduction in MAE. Meanwhile, both Figs. 5 and 6 indicate that with the inclusion of 6 additional shots, the performance closely approaches that of full-shot training, demonstrating the model's strong extrapolation capability even with a limited dataset. On the other hand, for the geometric

parameters in the 700 kA discharge, a zero-shot model maintains an acceptably low prediction error, suggesting more robust extrapolative abilities for shape and position parameters. These experiments carry encouraging implications for both the predictive and control aspects of staged discharge within the upcoming ITER device.

RL-based on-device control results

In line with the long-term trajectory, magnetic control in ref. 9, we aim to learn an RL-based policy to generate direct coil voltage commands for the target plasma shape. This policy is trained through interactions with an autoregressively evolving dynamics model. During training, an actor-critic-based Proximal Policy Optimization (PPO)⁵³ is leveraged for learning to achieve the desired plasma current and shape, which is identical to the

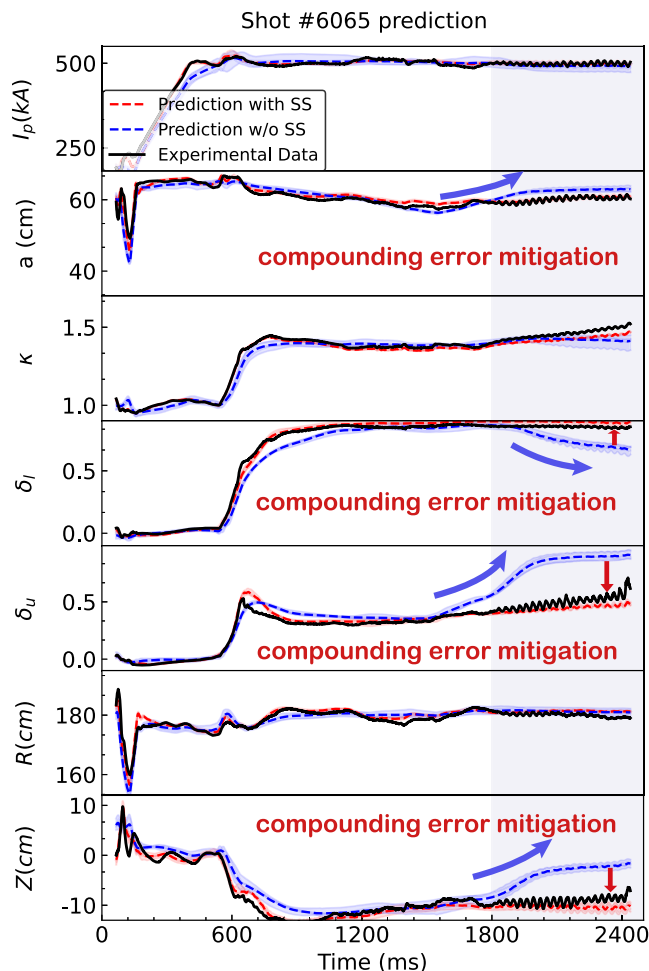


Fig. 2 | The demonstration case to predict plasma shape and position parameters for shot #6065 with and without the involvement of Scheduled Sampling. Compared with actual discharge data (black solid line), a model trained with scheduled sampling (red dashed line) more effectively mitigates the compounding error than one without scheduled sampling (blue dashed line). The shaded area represents the results of ensembling 5 independent models.

discharge result between 1500 and 2000 ms of shot #6226 of the HL-3 tokamak, an H-mode discharge with ELMs and a plasma current of 300 kA. In PPO⁵³, the “actor” utilizes a policy to generate physics-reasonable magnetic coil voltage commands (see “Methods” section) and modulate the plasma to the configuration of interest, while the “critic” evaluates the effectiveness of the current policy by estimating state values or advantages for potential policy refinement. Fig. 7a demonstrates the target tracking accuracy of the well-trained PPO actor within 1500–2000 ms⁵ simulations. The control error of the plasma current is 2.55 kA, or equivalently, 0.85% of the target for a 300 kA discharge. The control error for minor radius a is 0.34 cm, which corresponds to 0.52% of the 65 cm minor radius characteristic of the HL-3 tokamak device. Additionally, the control errors for the geometric center coordinates, both R and Z , are constrained within a maximum of 1 cm. The results indicate that the RL agent, supported by a data-driven dynamics model, can effectively explore and assimilate strategies that are advantageous for achieving discharge control objectives.

Inspired by these promising results, we further evaluate the performance of the RL-based policy by integrating the well-trained PPO actor with the Plasma Control System (PCS) of the HL-3 tokamak. Consistent with ref. 9, experiments are executed without further tuning the weights of the PPO actor after training. In other words, there is a “zero-shot” transfer from the simulation to the real device. Meanwhile, to provide a real-time plasma shape diagnostic and complete the shape-control feedback loop, we utilize

our previously developed EFITNN³⁸, a surrogate model of the EFIT code^{39,40} consisting of several fully connected neural network layers with residual connections⁵⁴. In comparison with offline EFIT, EFITNN yields a relative error of less than 1% and consumes just 0.08 ms on NVIDIA A100 Tensor Core GPU with TensorRT. The adoption of data-driven EFITNN contributes to effectively reducing the learning expenses of the dynamics model and facilitating real-time control. A summarized comparison between EFITNN and offline EFIT is presented in Fig. 8.

RL-based 400-ms magnetic control for target tracking. To thoroughly evaluate the flexibility of RL-based target tracking, we adopt a progressive approach. Initially, we assess the agent’s target tracking ability and ensure it aligns with the simulations. For this purpose, the experimental target remains identical to the reference shot #6226 for PPO training, and the PPO actor responds to the variations in plasma shape and position parameters diagnosed by EFITNN. Despite the consistency in targets, noticeable differences arise during the PID control due to the inherent uncertainties in tokamak discharge and simultaneously coupled operations for HL-3 maintenance and testing. A series of RL control experiments has been performed on HL-3.

Typically, in shot #12781, alongside the traditional PID control, we conduct an RL-based magnetic control from 1600 ms to 2000 ms with 30 ms transition periods at the start and end. Fig. 7b provides the corresponding control results, and Table 1 details the evolution of elongation (κ) and upper and lower triangularities (δ_u and δ_l) during the RL control phase. During its operational phase, the RL controller demonstrates excellent tracking performance for the four plasma shape parameters (a , κ , δ_u , and δ_l). It rapidly adjusts these parameters to their target values and subsequently maintains their stability with high precision. On par with PID, the controller capably maintains the radial (R) and vertical (Z) positions within a stable range, with precision slightly lower than that of shape parameters. Additionally, the control handover induces a brief fluctuation in the plasma current (I_p), which the RL controller promptly corrects, restoring it to the target level. The experimental results confirm that the RL controller possesses the capability for precise plasma shape regulation. In contrast to the conventional PID system, which is limited to controlling plasma positions (R and Z) and current (I_p), the RL controller successfully extends the closed-loop scope to the simultaneous regulation of four additional key shape parameters (a , κ , δ_u , and δ_l).

RL-based control for zero-shot adaptation to a changed triangularity target. Triangularity is a critical shape parameter in tokamak plasmas, significantly affecting both vertical stability and the structure of the plasma boundary⁵⁵. Consequently, the ability to control triangularity is crucial for optimizing tokamak plasma equilibria and performance. To validate the flexibility of RL, we intentionally decrease the upper and lower triangular shape parameters of the plasma by 0.2 from their target values during the training phase. The performance of the PPO actor on triangularity targets not encountered during training allows us to evaluate the agent’s adaptability to varying control objectives.

In shot #6698, similar to the procedure described in the preceding subsection on target tracking control, RL assumes the discharge control task from PID from 1600 ms to 1700 ms with 10 ms transition periods at the start and end. Fig. 9 and Table 2 illustrate the corresponding control results of the upper and lower triangularity. Both the lower triangularity δ_l and the upper triangularity δ_u successfully reach the new target values. Specifically, the lower triangularity δ_l swiftly progresses toward the target following a transition period of approximately 10 ms, finally reaching it around 1650 ms. Meanwhile, the upper triangularity δ_u follows a gradual trajectory toward the target during the initial 10-ms transition phase, accelerates its adjustment after around 20 ms, and aligns with the target at approximately 1650 ms. It then maintains alignment with the target trajectory for the remaining 50 ms. These zero-shot triangularity control results verify the versatility of RL toward effective control in previously unencountered scenarios. Furthermore, the post-experiment replay of the dynamics model

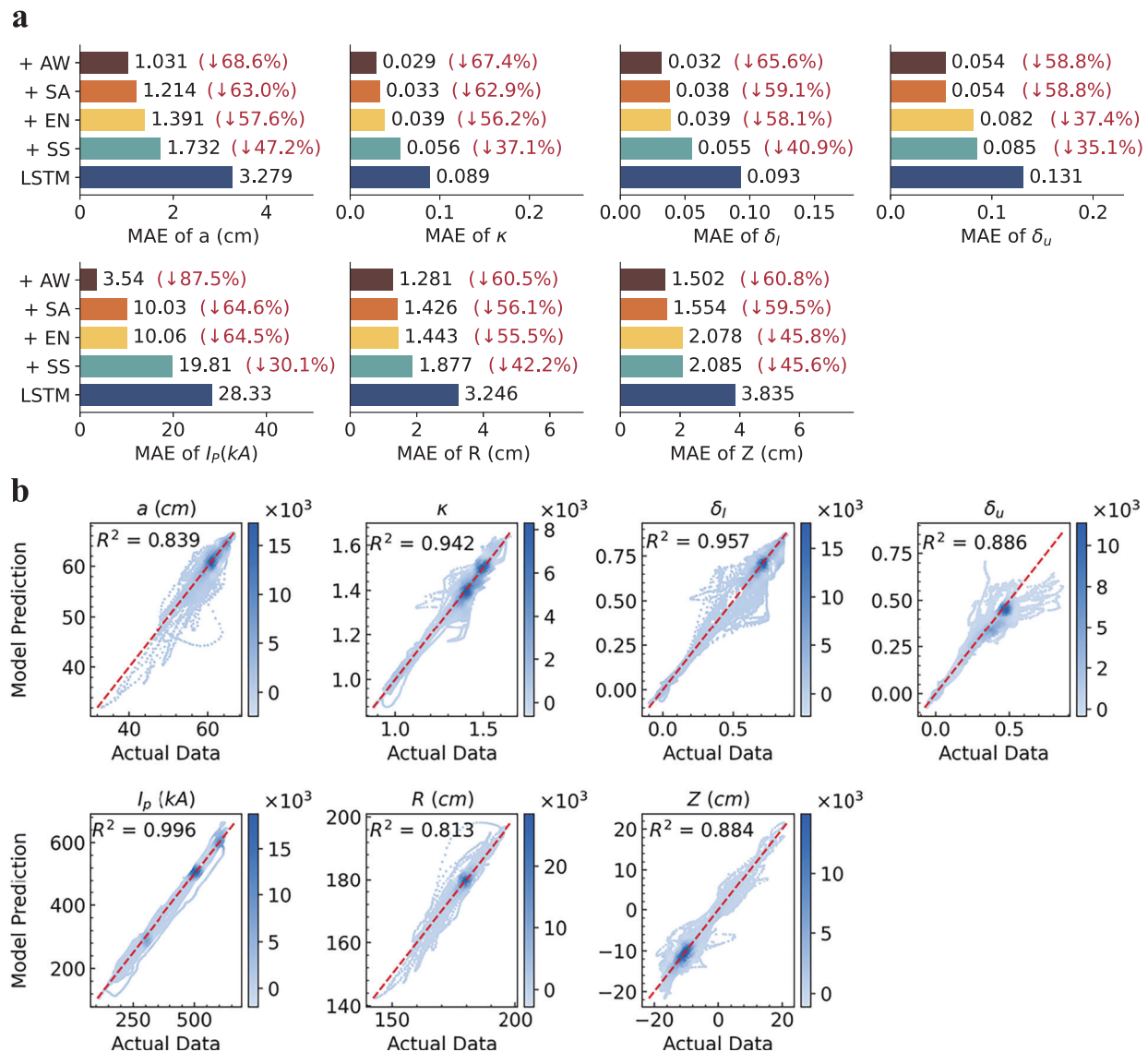


Fig. 3 | Accuracy of the dynamics model. a Ablations for key techniques in terms of MAE. SS: scheduled sampling, EN: ensemble, SA: self-attention, AW: adaptive weight. Each group depicted here is derived by incorporating the corresponding module on the basis of the preceding group. For instance, the notation “+EN”

signifies the integration of the Ensemble into the LSTM model that already incorporates scheduled sampling (“+SS”). **b** Goodness-of-fit of the dynamics model in terms of the coefficient of determination R^2 . The color map indicates the density of data points, and the red dashed line represents the line of perfect agreement.

from 1600 ms to 1700 ms reveals a consistent trend, underscoring the model’s generalizability and robustness for supporting RL in exploratory applications.

Conclusions

We present a technique to learn a fully data-driven dynamics model to characterize the evolution of plasma in a tokamak device. The dynamics model, which leverages multiple components to mitigate the compounding error issue and improve accuracy during autoregression, exhibits high fidelity and appealing extrapolation capability. It effectively supports the fast training of an RL-based high-dimensional trajectory control policy, which, together with the real-time plasma diagnostic empowered by EFITNN, can directly manipulate the magnetic coil voltages to transform the plasma state toward desired configurations.

Our work demonstrates the feasibility of data-driven, end-to-end learning and control of plasma current and shape. By avoiding reliance on first-principle-based simulators⁹ and auxiliary controlling commands from historical data or controllers^{5,11,12,35–37}, our approach significantly

differs from existing works. High-dimensional control over a learned model falls into the general scope of model-based reinforcement learning (MBRL)²⁴. In this regard, as in Supplementary Fig. 1a, instead of evolving the plasma state, ref. 5 mainly uses the output of the learned dynamics model to generate a competent reward function, which limits its applicability to localized reactive control only. Meanwhile, as in Fig. 10, ref. 35 conducts a receding (limited-term) horizon control through computation-intensive MPC, which involves cumbersome online autoregression of the dynamics model for trajectory sampling, making real-time deployment challenging⁵⁶. On the contrary, as in Supplementary Fig. 1b, our work provides a proof-of-concept study on long-term high-dimensional dynamics model learning, which further enables the training and deployment of the independent PPO actor for complex magnetic control.

Our research presents a promising direction for routine discharge operations in future tokamak devices. In the future, we will focus on further improving the overall performance and generalization capability of the control model. Building upon this foundation, we aim to address more challenging control tasks, including the intelligent avoidance of

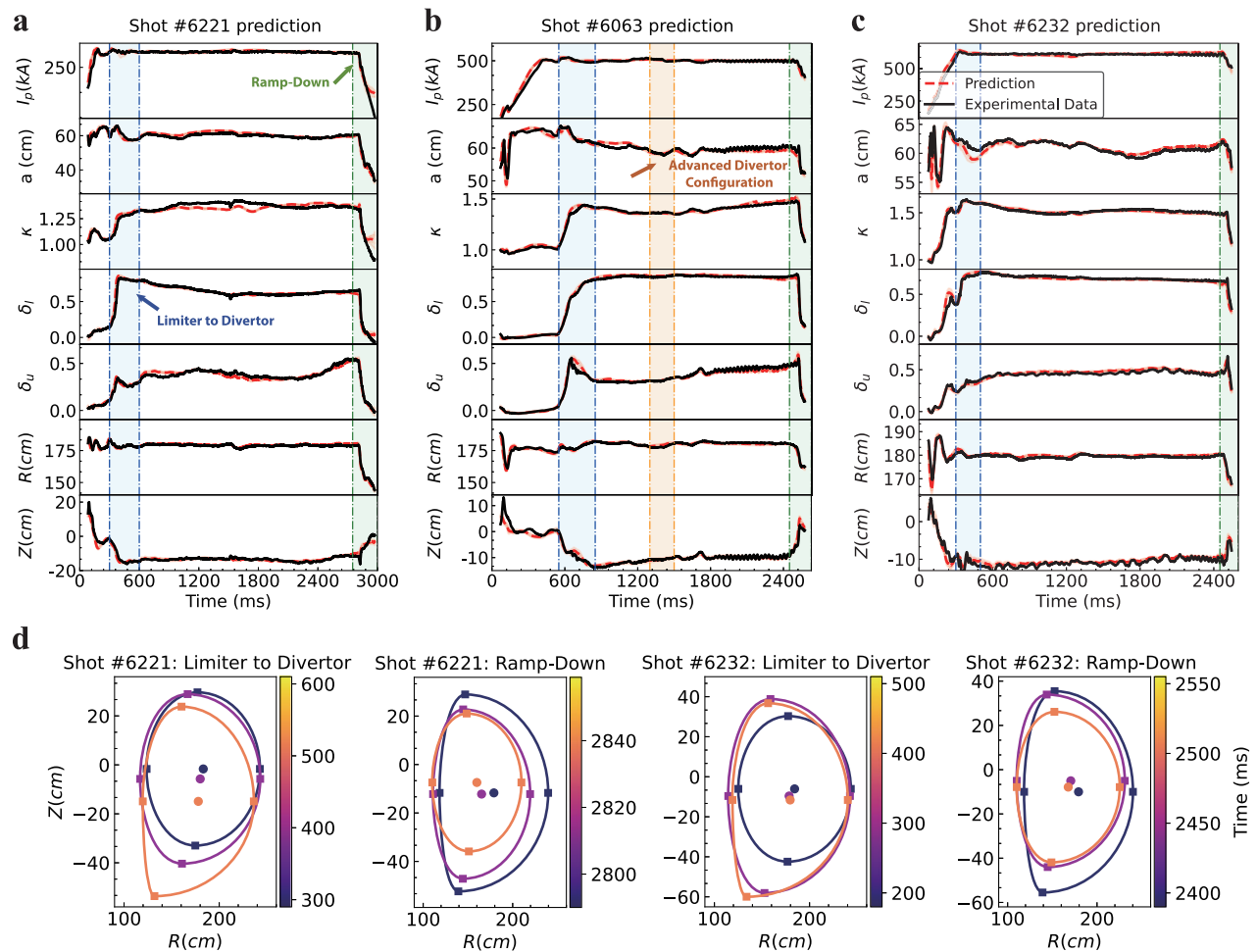


Fig. 4 | The accuracy of the dynamics model to predict abrupt plasma shape variations. In (a–c), the model’s prediction (red dashed line) is compared with experimental data (black solid line). The shaded regions highlight key operational phases: the light blue region marks the transition from the limiter to divertor configuration, the light orange region indicates the advanced divertor configuration transformation phase, and the light green region denotes the plasma current ramp-down phase. **a** Shot #6221, an H-mode discharge with ELMs and a plasma current of

300 kA. **b** Shot #6063, an advanced divertor configuration discharge with a plasma current of 500 kA, with the advanced divertor configuration undergoing a complete transformation between 1300 ms and 1500 ms. **c** Shot #6232, an ITB discharge with a plasma current of 600 kA. **d** The Last Closed Flux Surface (LCFS) variations correspond to the transition from limiter to divertor shape and soft ramp-down phases for shot #6221 (300 kA) and shot #6232 (600 kA), respectively.

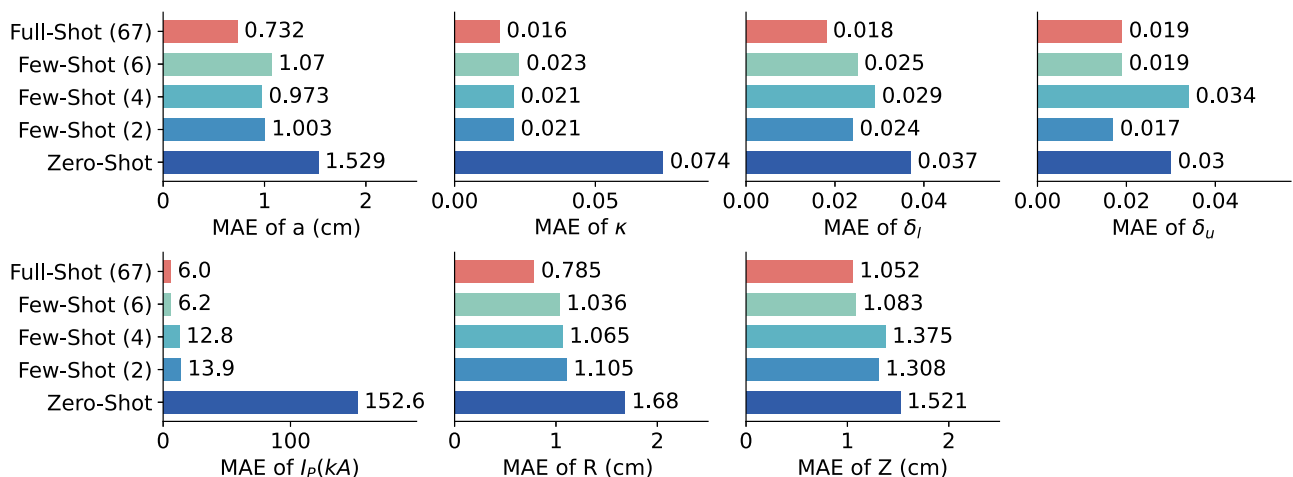


Fig. 5 | Model error on 700 kA discharges. Mean Absolute Error (MAE) averaging over 30 shots of 700 kA discharges. The numbers in the brackets correspond to the number of 700 kA discharges included in the training dataset.

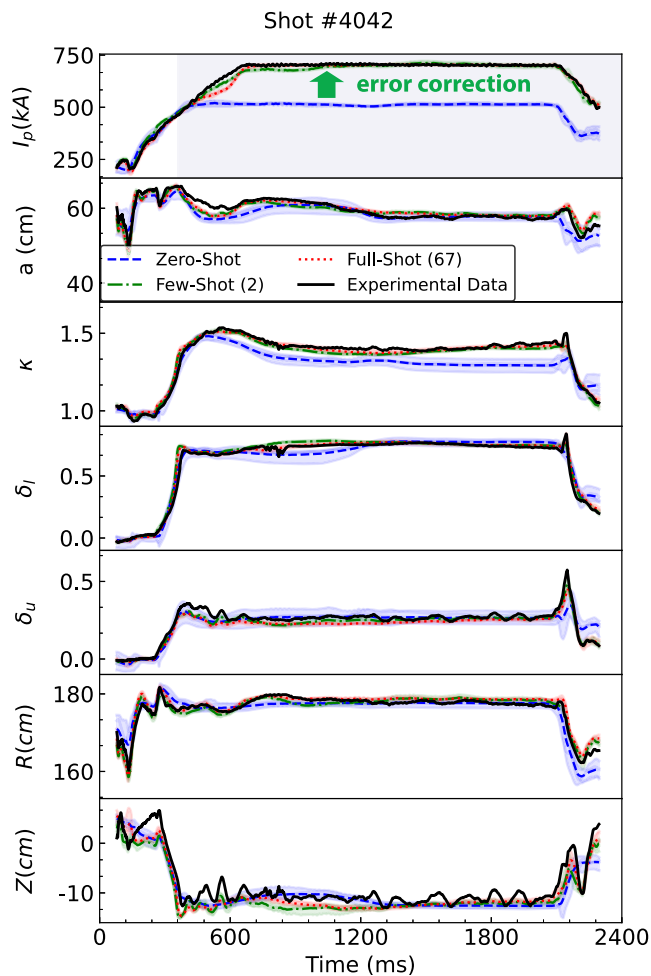


Fig. 6 | Few-shot learning performance on an unseen discharge. Test results of shot #4042 under the dynamics model from zero-shot (blue dashed line), 2-shot (green dash-dot line), and full-shot training (red dotted line). The shaded area represents the results of the model ensemble, while actual discharge data is given in a black solid line.

plasma instabilities and disruptions^{57–59}. Moreover, we plan to conduct transfer learning across tokamak devices⁵⁷, further verifying the potential advantages of a fully data-driven method over first-principle-based simulators.

Methods

HL-3

The HL-3 (previously named HL-2M) Tokamak is an experimental fusion device constructed at the Southwestern Institute of Physics in China. It is designed to operate at plasma current $I_p = 3$ MA, toroidal field $B_t = 3$ T, major radius $R = 1.78$ m, minor radius $a = 0.65$ m, elongation $\kappa \leq 1.8$, triangularity $\delta > 0.5$ ⁶⁰. HL-3 has recently focused on high-current, high β_N discharges to achieve high-performance plasma, establishing critical foundations for burning plasma research. As shown in Fig. 1a, the poloidal field configuration of HL-3 is intricately managed by a Central Solenoid (CS) coil, complemented by an ensemble of eight pairs of Poloidal Field (PF) coils strategically distributed as inner coils (PF1–PF4), top-bottom coils (PF5–PF6), and outer coils (PF7–PF8), with the latter trio playing a pivotal role in establishing the divertor configuration, hence their nomenclature as divertor coils^{61–63}. In particular, HL-3 directly manipulates the voltage of 17 coils, comprising 1 CS coil and 16 PF coils. This manipulation indirectly influences the coil currents and regulates the magnetic field within the vacuum chamber, enabling the confinement and discharge of plasma.

Data pre-processing

Typically, the plasma discharge process can be divided into three distinct phases (i.e., ramp-up, flat-top, and ramp-down phases). However, for the ramp-up phase's plasma, the relatively low current and the difficulty in calculating the eddy currents in passive conductor structures (such as the vacuum chamber and the first wall) often result in inaccurate shape calculations, weakening the correlation between the coil voltage commands and plasma shape. Therefore, considering that the primary objective of the dynamics model is to capture the complex evolution of the plasma state and ultimately facilitate efficient RL training, model training is exclusively based on data from the ramp-up phase, where the plasma current surpasses 100 kA, and the flat-top phase with a duration of at least 500 ms. Based on the historical discharge logs of HL-3 across campaigns from 2022 to 2024, after such a meticulous selection procedure, a dataset comprising 832 successful discharges has been curated, with a temporal resolution of 1 ms, totaling 1,833,516 time slices. The categorization of these 832 shots by flat-top plasma current is detailed in Supplementary Table 1.

Acknowledging the importance of the historical plasma state and coil commands, the dynamics model integrates a time series as its input to better capture the plasma evolution toward equilibrium. Specifically, given the 1-kHz control frequency, the model uses a 30-ms historical dataset for single-step prediction of plasma current, shape, and position parameters, as well as the coil currents. A summarized list of the input and output variables of the dynamics model can be found in Supplementary Table 2. Briefly, this model learns up to 25 variables related to plasma current, shape, and position, as well as coil current from a 44-dimensional input, which includes autoregressive predictions of these 25 variables, 17 executed coil voltage configurations, the toroidal magnetic field, and loop voltage. In addition, $n = 30$ represents the length of the time series, and $\Delta t = 1$ ms is the temporal resolution. As the CS coil is composed of two parallel connected sets of coils, the dimension of coil current (I_c) equals that of coil voltage (U) plus 1.

Throughout the discharge experiments in HL-3, engineering conditions change continuously. As a result, we utilize the most recent discharges for testing while earlier datasets are used for training. This methodology contributes to learning from updated data distributions and adapts to the latest discharge environment. To prevent model overfitting, we designate 50 discharges from shot #5893 to shot #6055 as the validation set, and 70 discharges, including shots #6056–#6265 and #5095–#5105, as the test set. Note that the range from shot #5095 to #5105 corresponds to 10 discharges of the advanced divertor configuration, which exhibit more pronounced morphological variations than standard discharges. During training, a 15-ms smoothing process is implemented on the training data to effectively mitigate non-electromagnetic interference. Since the training dataset suffers from a paucity of similar data instances, the corresponding results in Supplementary Fig. 4 could partially validate the accuracy and extrapolation capability of the dynamics model.

DNN design

DNN structure for the dynamics model. Due to the practical demands, plasma discharges are often carried out with identical or similar parameter sets to replicate specific physical outcomes, resulting in a lack of diversity in the dataset. Therefore, although the Transformer excels in parallel computation and can apprehend holistic information owing to its self-attention mechanism⁴⁶, it faces overfitting issues under limited and narrow data distribution. The testing MAE for both the Transformer and LSTM networks, as shown in Supplementary Fig. 2, suggests that under the same training methods, dynamics models utilizing LSTM networks typically outperform those based on the Transformer. As a result, we opt for LSTM as the backbone of our dynamics model.

We utilize a stack of DNNs, including a three-layer LSTM network, a self-attention module, and a three-layer fully connected MultiLayer Perceptron (MLP). Specifically, each LSTM layer comprises 128 hidden neurons with a dropout rate of 0.2. For the 30 consecutive steps' 44-dimensional input, as in Supplementary Table 2, the third LSTM layer produces an intermediate hidden layer output with dimensions of 30×128 .

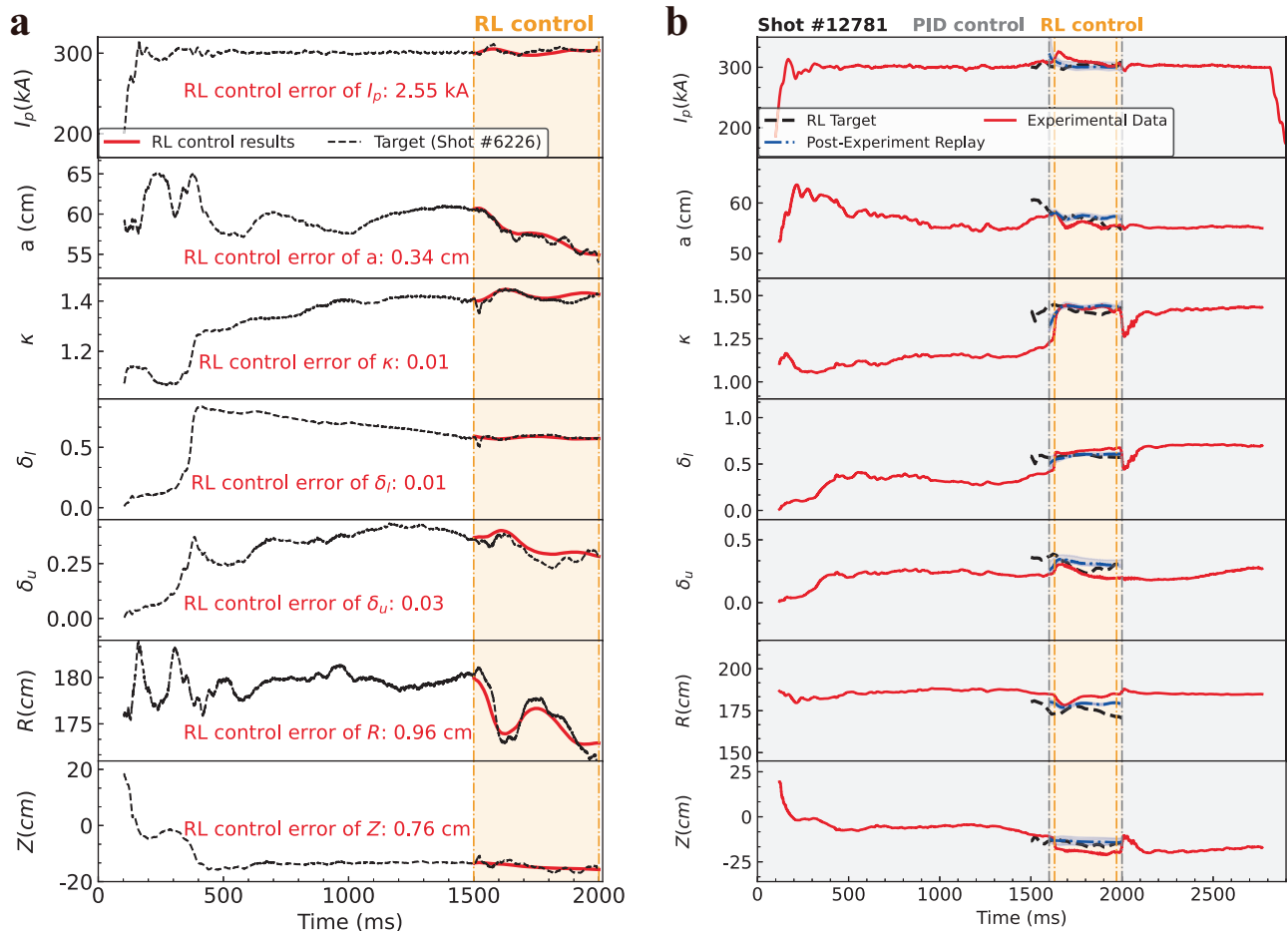


Fig. 7 | Target tracking control results of RL on both simulation and actual devices. **a** Simulated control results of plasma current I_p , shape and position parameters a , κ , δ_u , δ_l , R , and Z . The RL control results (red solid line) track the target waveform from shot #6226 (black dashed line), with the handover time set at 1500 ms. **b** In shot #12781, the experimental data (red solid line) and the post-

experiment replay of the dynamics model (blue dashdot line), as well as the target (black dashed line). The gray shaded area indicates the phase under PID control, while the orange shaded area represents the phase under RL control. The vertical dashed lines mark the start and end of the 30 ms handover period between the two controllers.

Subsequently, the inner product between each 128-length tensor and the last one is computed, followed by a softmax-based normalization operation to scale the values between 0 and 1. These self-attention values serve as attention weights to aggregate the original LSTM outputs as a 128-length vector. The integration of a self-attention mechanism into the LSTM allows the assignment of distinctive weights to the intermediate output, rather than placing equal emphasis and neglecting the inter-differences. Fig. 3a validates the effectiveness of this modification. Afterward, the self-attention-processed vector undergoes three MLPs with 128, 64, and 25 neurons and a ReLU activation function, culminating in the single-step prediction of 25 variables as specified in Supplementary Table 2.

DNN structure for EFITNN. EFITNN³⁸ is a surrogate model trained with input and output provided by an offline Equilibrium FIT (EFIT) code^{39,40}. It accepts input signals from a poloidal array of pick-up coils, flux loops, and coil currents, and concatenates them into a 68-element vector. EFITNN supports the prediction of 8 scalar values, including six plasma shape parameters, as well as the safety factor in 95% minor radius q_{95} and plasma inductance l_i . Besides, it can produce predictions for poloidal magnetic flux distribution Ψ_{rz} and toroidal current density distribution J_p with a resolution up to 129×129 . For these heterogeneous outputs, EFITNN adopts a shared 4-layer MLPs. Subsequently, for scalar prediction, another 4-layer MLPs with residual connections⁵⁴ are leveraged, while Ψ_{rz} and J_p rely on deconvolutional networks⁶⁴. Note that each MLP layer consists of 512 neurons with a GeLU activation function. EFITNN

adopts simultaneous training for this multi-task learning, and we kindly refer to ref. 38 for the training details. Fig. 8 manifests the prediction accuracy after training. In addition, if only plasma shape and position parameters are predicted, EFITNN, without the use of deconvolutional networks, consumes 0.08 ms on NVIDIA A100 Tensor Core GPU with TensorRT, fully supporting the real-time deployment.

DNN structure for PPO. We train the RL agent using PPO⁵³, which, as an actor-critic framework, updates both the policy and value networks through interactions with the well-trained dynamics model. In our experiments, both networks are structured as three-layer MLPs. The policy network uses a tanh activation function and consists of 64 neurons per layer, while the value network employs a ReLU activation function and comprises 128 neurons per layer.

Training of high-fidelity dynamics models

Scheduled sampling. Without any preventive training measures, the autoregressive nature of dynamics models leads to a gradually increasing deviation between the actual data and model predictions due to compounding error, as illustrated in Supplementary Fig. 3. Directly applying such a model in practice can result in exposure bias⁶⁵, incurring reduced efficacy and severe misalignment between the training and testing modalities. Therefore, during our training, we incorporate scheduled sampling^{66,67} to stochastically determine the input, by selecting between the model's past predictive outputs and the current-step actual data

Fig. 8 | Error between EFITNN and offline EFIT. Histograms of the prediction error between the EFITNN model and the offline EFIT code for key plasma parameters over all time slices in the test set. The Mean Absolute Error (MAE) for each parameter is shown in the title of its respective panel.

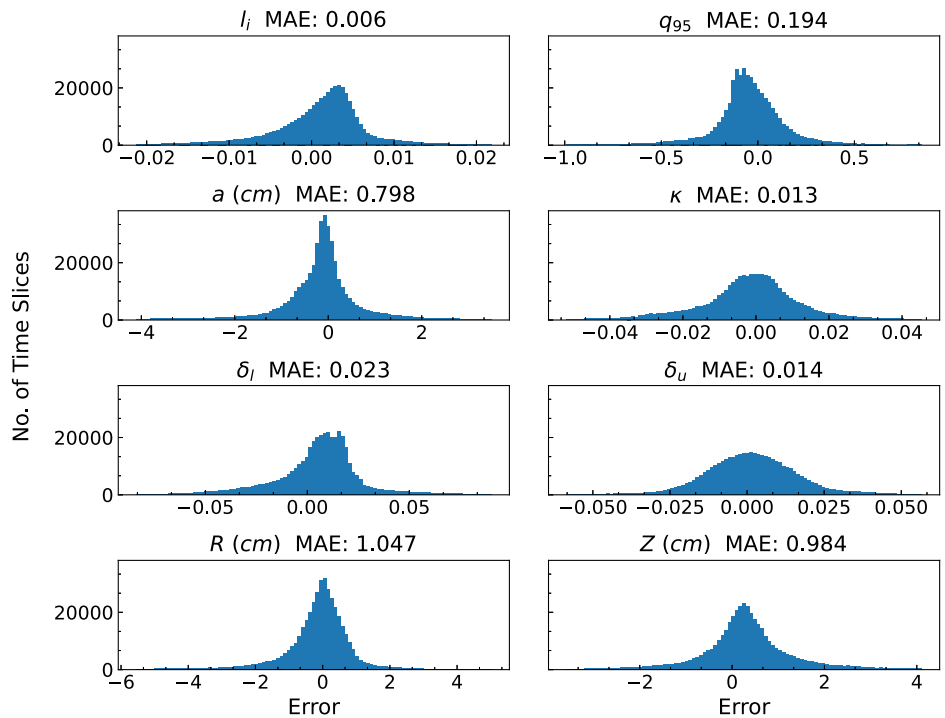


Table 1 | Plasma shape parameters during RL control for Shot #12781

Time (ms)	Elongation (κ)	Upper triangularity (δ_u)	Lower triangularity (δ_l)
1600	1.30	0.316	0.518
1630	1.40	0.347	0.612
1700	1.46	0.339	0.611
1800	1.42	0.269	0.639
1900	1.42	0.229	0.661
2000	1.38	0.182	0.620

according to a teacher-forcing ratio $p \in [0, 1]$. In other words, as the ratio p approaches 1, a greater proportion of the real input data will be used. Conversely, as p turns to 0, the model transitions to a strictly autoregressive mode, solely relying on the preceding steps' predictions. During the training, we initialize the ratio p to 1 and linearly decay it to 0 over 300 epochs. This approach ensures rapid convergence during training and facilitates a seamless shift toward an autoregressive scheme for training RL.

In contrast,¹¹ adopts a different approach to intentionally add a zero-mean Gaussian noise layer to the training data. Nevertheless, apart from the conflict with the data smoothing operation during training, the comparison in Supplementary Fig. 4 also suggests its inferiority to scheduled sampling. Specifically, the diverse impact on high-dimensional multi-task learning makes selecting the appropriate variance of added Gaussian noise challenging. For example, with an amplitude of $0.1\sigma_x$, where σ_x indicates the standard deviation of the input data, the model exhibits the lowest error in predicting κ but the highest error in predicting I_p . Conversely, the incorporation of scheduled sampling maintains stability in significantly reducing prediction errors for 25 variables. Therefore, we primarily take account of scheduled sampling in this paper.

Adaptive weight adjustment. To unanimously provide multi-task predictions for 25 variables, in Supplementary Table 2, the calibration of relative weights assigned to the predictive losses of individual variables could heavily affect the model performance. Mathematically, for a space

\mathcal{V} containing 25 variables, the aggregated loss function given the 44-dimensional input \mathbf{x} shall be a scalar $L_{\mathcal{V}}(\mathbf{y}_{\mathcal{V}}, \hat{\mathbf{y}}_{\mathcal{V}}; \mathbf{x})$ computed from 25 individual loss functions $L_v(\mathbf{y}_v, \hat{\mathbf{y}}_v; \mathbf{x})$ weighted by c_v , where \mathbf{y}_v and $\hat{\mathbf{y}}_v$ correspond to actual data and model prediction for the variable $v \in \mathcal{V}$, respectively, while $\mathbf{y}_{\mathcal{V}}$ and $\hat{\mathbf{y}}_{\mathcal{V}}$ denote the concatenation of \mathbf{y}_v and $\hat{\mathbf{y}}_v$, $\forall v \in \mathcal{V}$. Due to the discrepancies in distribution and noise among the variables, a straightforward averaging of individual loss functions can adversely impact the model's accuracy. Meanwhile, manual weight tuning is an arduous and time-intensive effort. Therefore, we utilize an adaptive weight adjustment technique by regarding the loss weight c_v as a trainable single variate⁴⁵. To prevent the loss weights from taking negative values during the optimization process, which could lead to negative loss values, we also employ a non-negative regularization constraint⁴⁴ to enhance the learning robustness. To sum up, the final loss function for training can be calculated as

$$L_{\mathcal{V}}(\mathbf{y}_{\mathcal{V}}, \hat{\mathbf{y}}_{\mathcal{V}}; \mathbf{x}) = \sum_{v \in \mathcal{V}} \left[\frac{1}{2 \cdot c_v^2} \cdot L_v(\mathbf{y}_v, \hat{\mathbf{y}}_v; \mathbf{x}) + \ln(1 + c_v^2) \right]. \quad (1)$$

Training details. In training the dynamics model, we initially segment all the discharge data of training shots into consecutive groups with a fixed length of 30 ms, followed by a shuffling process on all the segmented data. We use DataLoader, a PyTorch library, to load the processed training data, with a batch size set as 4096. In each training round, we randomly utilize a batch of data without replacement for training, and an epoch completes when all groups have been utilized. We employ the Adam optimizer for optimizing network parameters, with an initial learning rate of 0.001 for both the dynamics model and the adaptive weight adjustment. To prevent training instability due to an excessively high learning rate in the later stages of training, we adjust the learning rate using an exponential decay method. Specifically, we reduce the learning rate by a decay factor of 0.95 every 5 epochs, over a total of 300 training epochs.

Finally, in addition to Fig. 3, Supplementary Fig. 5 shows the accuracy for predicting all 18 coil currents.

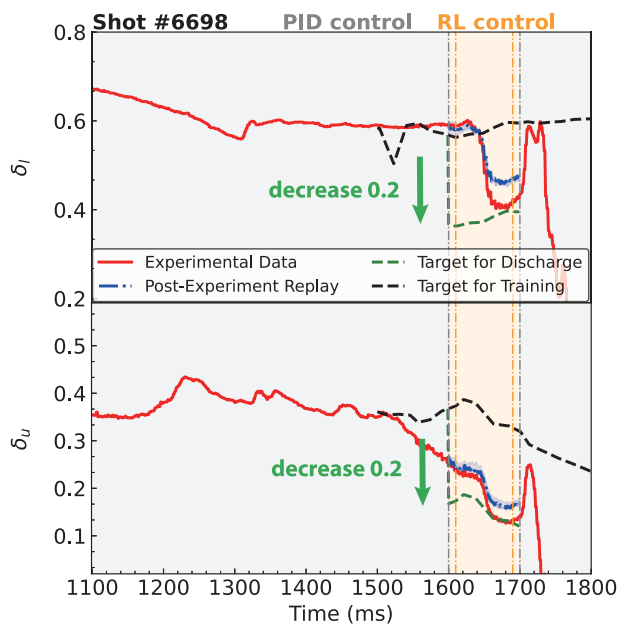


Fig. 9 | Zero-shot triangularity control in shot #6698. The upper and lower triangular control results of shot #6698. Compared to the target used for RL training (black dashed line), the target for this discharge (green dashed line) has a triangularity decreased by 0.2. The red solid line illustrates the experimental result, and the blue dash-dot line represents the post-experiment replay of our dynamics model. The gray shaded area indicates the phase under PID control, while the orange shaded area represents the phase under RL control. The vertical dashed lines mark the start and end of the 10 ms handover period between the two controllers.

Table 2 | Triangularity parameters during RL control for shot #6698

Time (ms)	Upper triangularity (δ_u)	Lower triangularity (δ_l)
1610	0.235	0.589
1630	0.224	0.597
1640	0.220	0.577
1650	0.189	0.482
1660	0.144	0.412
1690	0.131	0.423
Target	0.128	0.396

Training and deployment of RL-based magnetic controller

State, action, and reward of RL. In this study, we utilize RL to control the plasma current (I_p) and six shape and position parameters (a , κ , δ_l , δ_u , R , Z), and denote the concatenation of current observations as \mathbf{o} . To ensure the integrity and stability of HL-3 plasma discharges, we select the actual waveform of shot #6226, an H-mode discharge with ELMs and a plasma current of 300 kA, as the desired target \mathbf{d} for control. As shown in Supplementary Table 3, we consider a 14-dimensional state \mathbf{s} by amalgamating the observations \mathbf{o} and desired target \mathbf{d} , while the 17-dimensional action \mathbf{a} encompasses voltage commands of 1 CS coil and 16 PF coils. To ensure control safety and hardware integrity, we impose operational constraints by stipulating that the dynamic rate of change between two consecutive voltage commands must not exceed 20 V ms^{-1} . In other words, for any coil, any voltage command that changes beyond this limit will be clipped to adhere to the safety constraints.

Typically, RL relies on explicit targets for reward function design. In fusion experiments, however, diverse targets from varied discharge templates complicate reward engineering and can undermine model efficacy.

Therefore, we devise the reward function $r(\mathbf{s}, \mathbf{a})$ from two orthogonal perspectives. First, consistent with refs. 9,10, the reward function primarily takes account of the element-wise divergence between the current observation $\mathbf{o} \triangleq [o^{(1)}, \dots, o^{(7)}]$ and desired target $\mathbf{d} \triangleq [d^{(1)}, \dots, d^{(7)}]$. To merge the divergence components into a scalar reward and emphasize those control variables with sub-optimal control results, a smooth-max function¹⁰ is leveraged. Second, owing to the constraints of the power module, the coil current is permitted to cross zero only once. Therefore, contingent on the coil current predicted by the dynamics model, we impose a penalty term to inhibit the agent from exploring control commands that would result in multiple zero crossings of the coil current. To sum up, the reward for a state-action pair (\mathbf{s}, \mathbf{a}) is given as

$$r(\mathbf{s}, \mathbf{a}; \boldsymbol{\omega}, \alpha) = -\frac{\sum_{i=1}^7 \omega_i |o^{(i)} - d^{(i)}| \exp(-\alpha |o^{(i)} - d^{(i)}|)}{\sum_{i=1}^7 \omega_i \exp(-\alpha |o^{(i)} - d^{(i)}|)} + \sum_{i=1}^{17} p_i, \quad (2)$$

where weights $\boldsymbol{\omega} \triangleq [\omega_1, \dots, \omega_7]$ in the softmax function provide the (relative) importance of each component and are initialized as [3, 2, 2, 3, 1, 2, 2]. The modulation factor α , which affects the trade-offs between “easy” and “hard” (to satisfy) components¹⁰, is set as -1 in our experiment. Besides, for 17 coils, the penalty term p_i for coil i is designed as -1 when the coil current exceeds one zero-crossing, while nulls otherwise.

Dynamics model-based training. We utilize the trained dynamics model as the interaction environment, through which the RL agent explores suitable actions to achieve the desired plasma current and shape. Specifically, as the HL-3 tokamak is a preliminarily established device and initiates the global open campaigns in 2024 and 2025, a significant number of discharge shots have been scheduled for international joint physical experiments. Consequently, the proposal for RL-based control has to share the discharges with other experiments. The choice of shot #6226 for target tracking is also motivated by the potential to maximize RL’s takeover opportunities, considering the frequent use of similar shot templates in planned experiments on the HL-3 tokamak. As highlighted in Fig. 9, the promising outcomes underscore the inherent generalization capability, which suggests robust applicability to a variety of other discharge templates. In addition, given the potential uncertainty in freshly deploying RL-based control on HL-3, the early phase of a discharge, prior to 1500 ms, is primarily allocated to other physics experiments. Therefore, during training, RL assumes the 1-kHz control since 1500 ms of the discharge, and the initial observations \mathbf{o} are synchronized with those of shot #6226, ensuring a smooth transition between PID control and RL control. Furthermore, we take episodic training from 1500 ms to 2000 ms of the discharge. Without loss of generality, at each time step t , the RL agent determines an action \mathbf{a}_t according to the state \mathbf{s}_t , while the dynamics model yields next-step plasma current as well as the six shape and position parameters \mathbf{o}_{t+1} based on \mathbf{s}_t and \mathbf{a}_t . Note that given the slow-changing nature of the toroidal magnetic field and loop voltage, these two values are considered invariant and maintain the same values as the handover time of shot #6226. Hence, a reward $r(\mathbf{s}_t, \mathbf{a}_t)$ can be computed as in Equation (2). Finally, the episodic reward is derived from the discounted cumulative reward with a discount factor γ set as 0.98. To strike a balance between the variance and bias in the value function estimation, we employ the Generalized Advantage Estimation (GAE)⁶⁸ to compute the advantage function. In addition, other parameters for the PPO are set in a standard way. Specifically, we set the entropy coefficient, the clipping parameter ϵ , and the λ parameter for GAE as 0.05, 0.2, and 0.95, respectively. Following this, the PPO algorithm⁵³ can iteratively update the agent’s actor and critic networks. Fig. 7a illustrates the tracking accuracy of the trained agent toward the target waveform within the simulation setting and suggests that the RL agent, supported by a data-driven dynamics model, has effectively explored and assimilated strategies that are beneficial for achieving discharge control objectives.

The aforementioned single-template training necessitates retraining the model with a new target for changed operational scenarios (e.g.,

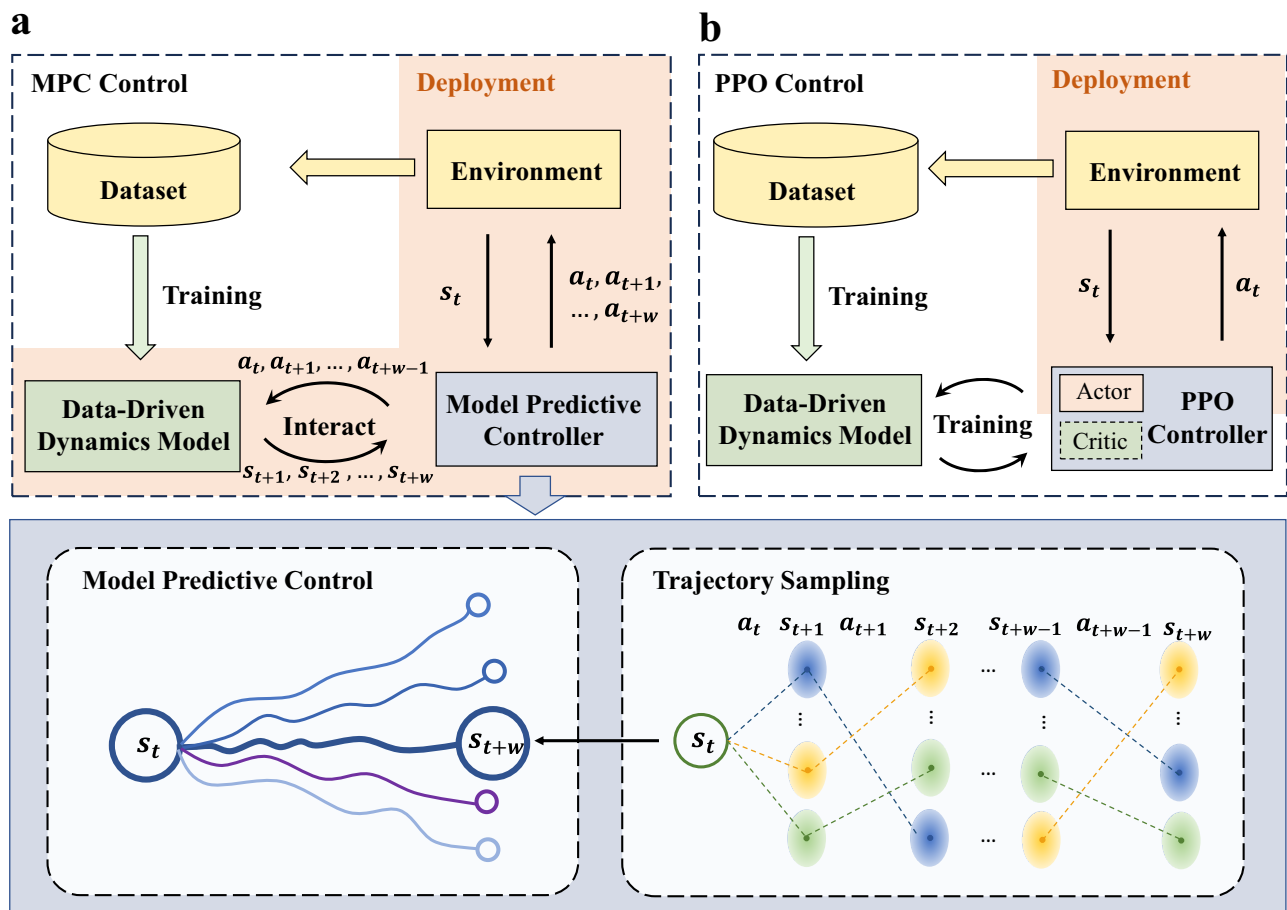


Fig. 10 | Comparison of the training and deployment between MPC (Model Predictive Control) and PPO (Proximal Policy Optimization). **a** MPC with online, limited-term dynamics model autoregression, and **b** PPO interacting with the dynamics model for training only. The bottom panel serves as an inset for **(a)**, illustrating the core logic of the Model Predictive Controller. At the current state, the controller performs trajectory sampling by simulating various possible future action

sequences using the dynamics model. These simulations generate multiple potential future state trajectories (visualized as different colored paths). The controller evaluates these trajectories and selects the optimal one (the bold blue path) that best achieves a predefined objective. Finally, only the first action of this optimal sequence is executed, and this entire planning process is repeated at the next state.

targeting a 700 kA plasma current). Despite this requirement for retraining, the process is efficient. Systematic evaluations across different GPU platforms (Supplementary Fig. 9 and Supplementary Table 4) substantiate this efficiency. The PPO-based controller converges within 500 training episodes (25,000 interaction steps). Utilizing a standard configuration (PyTorch 2.1.0, CUDA 12.2, batch size 512), training on an NVIDIA GeForce RTX 4090 completes in just 22 min. Such rapid training underscores the framework's practical adaptability, even with the need for template-specific retraining.

On-device deployment. To ascertain the practicability of the trained RL agent as well as the accuracy of the dynamics model, we integrate the PPO actor into the PCS of HL-3 to undertake RL-based control in real discharges. Since the PCS operates under the Red Hat 6.9 system and performs real-time computations in C language, we convert the Pytorch-based PPO actor into a compatible ONNX (Open Neural Network Exchange)⁶⁹ format, which offers cross-platform deployment capability and high execution efficiency. During real discharges, the RL agent receives plasma shape and position parameters from EFITNN³⁸ while the plasma current is measured by the CODIS-RTC system via reflective memory in real time. The agent then generates coil voltage commands and delivers them to actuators through the PCS, completing a closed-loop control period of 1 ms, synchronized with that for the PID control.

As mentioned earlier, the open campaigns on HL-3 have attracted a considerable number of international joint experiments, limiting the

opportunities for testing RL-based control. Eventually, in a shared discharge, the handover to RL-based control has to be delayed until other physics experiments involving auxiliary heating are completed. To avoid any potential plasma disturbances caused by shutting down the auxiliary heating at 1500 ms, which could impact the analysis of experimental results, we have further postponed the start of RL-based control to 1600 ms. In other words, we activate the RL control at 1600 ms and revert to the PID control after a predefined window. To ensure stability incurred during these mode switches, we establish a transition period, during which a weighted sum of commands from both modes is undertaken, and the weight of the superseding control mode linearly increases.

Our initial RL control experiments are conducted during the #6691-#6698 shot series on HL-3. Due to the coincidence with the development of the fast vertical position control system, the experiments undergo severely unstable discharge conditions, with only three effective discharges (i.e., shots #6691 (Supplementary Fig. 6a), #6693 (Supplementary Fig. 7), and #6698 (Supplementary Fig. 6b)) sustaining the flat-top phase beyond 1600 ms. Furthermore, to ensure operational safety, the RL control window is limited to 100 ms, with 10 ms transition periods at its start and end. Unfortunately, the conventional PID and fast vertical position control still exhibit limited capability to suppress emerging voltage variations after reverting from RL to PID control, frequently resulting in plasma disruptions. More recently, following the maturation of the fast vertical position control system for routine operations, we replicated these experiments during the #12777-#12781 shot series by adopting the same actor model. This second

campaign yields three effective RL-controlled discharges: shots #12777 (100 ms), #12779 (260 ms), and #12781 (400 ms). For these shots, 30 ms transition periods are implemented at the start and end. All three effective discharges are completed with no disruptive events. Fig. 7b illustrates the complete discharge for shot #12781. The complete discharges for the remaining five successful RL-controlled shots (from both experimental campaigns) are presented in Supplementary Figs. 6, 7, and 8. To quantitatively assess the performance differences between RL-based control and fixed voltage control, we conduct a comparative experiment using shot #12777. For shot #12777, it adopts an RL control solely until 1700 ms without the last 30-ms transition period, while from 1700 ms to 1730 ms, the control voltages are deliberately fixed at their 1700 ms values. As depicted in Supplementary Fig. 8, the resulting oscillations in plasma current and shape confirm that, without closed-loop control, plasma stability is significantly compromised, demonstrating the critical need for an effective closed-loop control. Supplementary Fig. 10 illustrates the plasma current for the seven shots from these campaigns in which RL control is not executed.

Post-experiment analyses. As illustrated in Figs. 7b and 9a, the post-experiment replay for shot #12781 and shot #6698 is calculated using the dynamics model. Specifically, the dynamics model starts to evolve according to the input variables in Supplementary Table 2, set as practical values at the handover time, as well as the practically executed coil voltage commands, which are generated by RL during on-device deployment. In other words, such a post-experiment replay represents the autoregressive output of the dynamics model, and its closeness to the EFITNN-based diagnostic measurement contributes to verifying the accuracy of the dynamic model.

On the other hand, as shown in Fig. 4d, to visually display the prediction and control effect on the plasma shape and position, we calculate the four extreme points of the cross-section based on the six shape and position parameters obtained from the EFITNN computation. The plasma boundary is then approximated as the connection of four-quarter ellipses⁷⁰.

To better understand the control logic of the RL controller, we perform a more in-depth analysis of the experimental results from shot #12781. As illustrated in Fig. 7b, while the controller demonstrates high-fidelity tracking of the primary shape parameters (a , κ , δ_w , and δ_i), the positional parameters (R and Z) exhibit persistent steady-state deviations from their targets. This differential control efficacy is primarily attributed to three factors: (1) the inherent strong coupling among the shape-defining parameters (a , κ , δ_i , and δ_w), which are co-regulated by the PF coils; (2) the non-negligible influence of non-electromagnetic experimental disturbances, such as gas puffing and auxiliary heating, which could significantly affect the evolution of R and Z ; and (3) an implicit control priority mechanism within the RL policy – the policy prioritizes the correction of κ , δ_i , δ_w , and R due to the larger deviations therein. In summary, compared to the standard PID controller, the RL controller obtains similar precision in regulating R and Z , but significantly extends this capability by simultaneously providing closed-loop control over minor radius a , upper and lower triangularity δ_w , δ_i , and elongation κ .

Data availability

The source data for all graphs and charts presented in the main figures are available in Supplementary Data 1. Any other datasets generated during the current study are available from the corresponding author on reasonable request.

Code availability

Codes used during the current study are available from the corresponding author on reasonable request.

Received: 9 December 2024; Accepted: 29 August 2025;

Published online: 03 October 2025

References

1. He, Z.-H. et al. Coherent control of plasma dynamics. *Nat. Commun.* **6**, 7156 (2015).

2. Fortov, V., Ivlev, A., Khrapak, S., Khrapak, A. & Morfill, G. Complex (dusty) plasmas: current status, open issues, perspectives. *Phys. Rep.* **421**, 1–103 (2005).
3. Yuan, Q. et al. Plasma current, position and shape feedback control on EAST. *Nucl. Fusion* **53**, 043009 (2013).
4. Walker, M. et al. The ITER plasma control system simulation platform. *Fusion Eng. Des.* **96**, 716–719 (2015).
5. Seo, J. et al. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature* **626**, 746–751 (2024).
6. Maljaars, E. et al. Control of the tokamak safety factor profile with time-varying constraints using MPC. *Nucl. Fusion* **55**, 023001 (2015).
7. Felici, F. et al. Real-time-capable prediction of temperature and density profiles in a tokamak using RAPTOR and a first-principle-based transport model. *Nucl. Fusion* **58**, 096006 (2018).
8. Van Mulders, S. et al. Scenario optimization for the tokamak ramp-down phase in RAPTOR: part a. Analysis and model validation on ASDEX Upgrade. *Plasma Phys. Control. Fusion* **66**, 025006 (2023).
9. Degraeve, J. et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* **602**, 414–419 (2022).
10. Tracey, B. D. et al. Towards practical reinforcement learning for tokamak magnetic control. *Fusion Eng. Des.* **200**, 114161 (2024).
11. Seo, J. et al. Feedforward beta control in the KSTAR tokamak by deep reinforcement learning. *Nucl. Fusion* **61**, 106010 (2021).
12. Seo, J. et al. Development of an operation trajectory design algorithm for control of multiple 0D parameters using deep reinforcement learning in KSTAR. *Nucl. Fusion* **62**, 086049 (2022).
13. Carpanese, F. *Development of Free-Boundary Equilibrium and Transport Solvers for Simulation and Real-Time Interpretation of Tokamak Experiments*. Ph.D. thesis, EPFL, Lausanne (2021).
14. Zhang, Y., Wang, S., Yuan, Q., Xiao, B. & Huang, Y. Real-time feedback control of β_p based on deep reinforcement learning on EAST. *Plasma Phys. Control. Fusion* **66**, 055014 (2024).
15. Campbell, D. et al. ITER research plan within the staged approach (level iii - final version). <https://www.iter.org/sites/default/files/media/2024-04/itr-24-005-ok1.pdf> (2024).
16. Moreau, D. et al. Integrated magnetic and kinetic control of advanced tokamak plasmas on DIII-D based on data-driven models. *Nucl. Fusion* **53**, 063020 (2013).
17. Moreau, D. et al. Plasma models for real-time control of advanced tokamak scenarios. *Nucl. Fusion* **51**, 063009 (2011).
18. Kates-Harbeck, J., Svyatkovskiy, A. & Tang, W. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature* **568**, 526–531 (2019).
19. Dong, G. et al. Deep learning based surrogate models for first-principles global simulations of fusion plasmas. *Nucl. Fusion* **61**, 126061 (2021).
20. Li, H., Li, J. Q., Fu, Y. L., Wang, Z. X. & Jiang, M. Simulation prediction of micro-instability transition and associated particle transport in tokamak plasmas. *Nucl. Fusion* **62**, 036014 (2022).
21. Wan, C. et al. Predict the last closed-flux surface evolution without physical simulation. *Nucl. Fusion* **64**, 026014 (2024).
22. Zhu, B. et al. Data-driven model for divertor plasma detachment prediction. *J. Plasma Phys.* **88**, 895880504 (2022).
23. Asadi, K., Misra, D. & Littman, M. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning* 264–273 (PMLR, 2018).
24. Moerland, T. M., Broekens, J., Plaat, A. & Jonker, C. M. Model-based reinforcement learning: a survey. *Found. Trends Mach. Learn.* **16**, 1–118 (2023).
25. Ke, N. R. et al. Modeling the long term future in model-based reinforcement learning. In *Proc. International Conference on Learning Representations* (ICLR, 2019).
26. Hafner, D. et al. Learning latent dynamics for planning from pixels. In *Proc. International Conference on Machine Learning* 2555–2565 (PMLR, 2019).

27. Chiappa, S., Racaniere, S., Wierstra, D. & Mohamed, S. Recurrent environment simulators. In *Proc. International Conference on Learning Representations (ICLR)*, 2017).
28. Abbeel, P. & Ng, A. Learning first-order Markov models for control. In *Proc. Advances in Neural Information Processing Systems (ACM)*, 2004).
29. Lambert, N., Wilcox, A., Zhang, H., Pister, K. S. & Calandra, R. Learning accurate long-term dynamics for model-based reinforcement learning. In *2021 60th IEEE Conference on decision and control (CDC)* 2880–2887 (IEEE, 2021).
30. Lambert, N., Pister, K. & Calandra, R. Investigating compounding prediction errors in learned dynamics models. Preprint at <https://arxiv.org/abs/2203.09637> (2022).
31. Mishra, N., Abbeel, P. & Mordatch, I. Prediction and control with temporal segment models. In *Proc. International Conference on Machine Learning* 2459–2468 (PMLR, 2017).
32. Asadi, K., Cater, E., Misra, D. & Littman, M. L. Towards a simple approach to multi-step model-based reinforcement learning. Preprint at <https://arxiv.org/abs/1811.00128> (2018).
33. Char, I., Chung, Y., Abbate, J., Kolemen, E. & Schneider, J. Full shot predictions for the DIII-D tokamak via deep recurrent networks. Preprint at <https://arxiv.org/abs/2404.12416> (2024).
34. Abbate, J., Conlin, R. & Kolemen, E. Data-driven profile prediction for DIII-D. *Nucl. Fusion* **61**, 046027 (2021).
35. Abbate, J., Conlin, R., Shousha, R., Erickson, K. & Kolemen, E. A general infrastructure for data-driven control design and implementation in tokamaks. *J. Plasma Phys.* **89**, 895890102 (2023).
36. Char, I. et al. Offline model-based reinforcement learning for tokamak control. In *Learning for Dynamics and Control Conference* 1357–1372 (PMLR, 2023).
37. Char, I. & Schneider, J. PID-inspired inductive biases for deep reinforcement learning in partially observable control tasks. *Adv. Neural Inf. Process Syst.* **36**, 59425–59463 (2023).
38. Zheng, G. et al. Real-time equilibrium reconstruction by multi-task learning neural network based on HL-3 tokamak. *Nucl. Fusion* **64**, 126041 (2024).
39. Lao, L. et al. Equilibrium analysis of current profiles in tokamaks. *Nucl. Fusion* **30**, 1035 (1990).
40. Lao, L., John, H. S., Stambaugh, R., Kellman, A. & Pfeiffer, W. Reconstruction of current profile parameters and plasma shapes in tokamaks. *Nucl. Fusion* **25**, 1611 (1985).
41. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. LSTM: a search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 2222–2232 (2017).
42. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proc. Advances in Neural Information Processing Systems*, vol. 28 (ACM, 2015).
43. Venkatraman, A., Hebert, M. & Bagnell, J. Improving multi-step prediction of learned time series models. In *Proc. AAAI Conference on Artificial Intelligence*, vol. 29 (ACM, 2015).
44. Liebel, L. & Körner, M. Auxiliary tasks in multi-task learning. Preprint at <https://arxiv.org/abs/1805.06334> (2018).
45. Kendall, A., Gal, Y. & Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 7482–7491 (IEEE, 2018).
46. Vaswani, A. et al. Attention is all you need. *Adv. Neural inf. Process Syst.* **30**, 5998–6008 (2017).
47. Rozante, J., Moreira, D., Godoy, R. & Fernandes, A. Multi-model ensemble: Technique and validation. *Geosci. Model Dev.* **7**, 2333–2343 (2014).
48. Ryutov, D., Cohen, R., Rognlien, T. & Umansky, M. A snowflake divertor: a possible solution to the power exhaust problem for tokamaks. *Plasma Phys. Control. Fusion* **54**, 124050 (2012).
49. de Vries, P. C. et al. Multi-machine analysis of termination scenarios with comparison to simulations of controlled shutdown of ITER discharges. *Nucl. Fusion* **58**, 026019 (2017).
50. Han, W. et al. Tracking blobs in the turbulent edge plasma of a tokamak fusion device. *Sci. Rep.* **12**, 18142 (2022).
51. Wang, S., Wang, Z. & Wu, T. Self-organized evolution of the internal transport barrier in ion-temperature-gradient driven gyrokinetic turbulence. *Phys. Rev. Lett.* **132**, 065106 (2024).
52. Staebler, G. Theory of internal and edge transport barriers. *Plasma Phys. Control. Fusion* **40**, 569 (1998).
53. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at <https://arxiv.org/abs/1707.06347> (2017).
54. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
55. Schuster, E. & Ariola, M. The role of controls in nuclear fusion. In *Proceedings of the 45th IEEE Conference on Decision and Control* 2263–2270 (IEEE, 2006).
56. Lore, J. et al. Time-dependent SOLPS-ITER simulations of the tokamak plasma boundary for model predictive control using SINDy*. *Nucl. Fusion* **63**, 046015 (2023).
57. Murari, A., Rossi, R., Craciunescu, T., Vega, J. & Gelfusa, M. A control oriented strategy of disruption prediction to avoid the configuration collapse of tokamak reactors. *Nat. Commun.* **15**, 2424 (2024).
58. Sheikh, U. et al. Disruption avoidance through the prevention of NTM destabilization in TCV. *Nucl. Fusion* **58**, 106026 (2018).
59. Barr, J. L. et al. Development and experimental qualification of novel disruption prevention techniques on DIII-D. *Nucl. Fusion* **61**, 126019 (2021).
60. Zhong, W. China's HL-3 tokamak achieves h-mode operation with 1 MA plasma current. *Innovation* **5**, 100555 (2024).
61. Song, X. et al. Plasma initiation and preliminary magnetic control in the HL-2M tokamak. *Nucl. Fusion* **61**, 086010 (2021).
62. Song, X., Li, J., Leuer, J., Zhang, J. & Song, X. First plasma scenario development for HL-2M. *Fusion Eng. Des.* **147**, 111254 (2019).
63. Xue, L. et al. Integrated plasma scenario analysis for the HL-2M tokamak. *Nucl. Fusion* **60**, 016016 (2019).
64. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision*, 818–833 (Springer International Publishing, Zurich, Switzerland, 2014).
65. Schmidt, F. Generalization in generation: a closer look at exposure bias. In *Proc. 3rd Workshop on Neural Generation and Translation* (Association for Computational Linguistics, Hong Kong, 2019).
66. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *Adv. Neural inf. Process Syst.* **28**, 1171–1179 (2015).
67. Vlachas, P. R. & Koumoutsakos, P. Learning on predictions: fusing training and autoregressive inference for long-term spatiotemporal forecasts. *Phys. D.* **470**, 134371 (2024).
68. Fujimoto, S., Hoof, H. & Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning* 1587–1596 (PMLR, 2018).
69. Jin, T. et al. Compiling ONNX neural network models using MLIR. Preprint at <https://arxiv.org/abs/2008.08272> (2020).
70. Dong, Q. et al. Adapted Swin transformer-based real-time plasma shape detection and control in HL-3. *Nucl. Fusion* **65**, 026031 (2025).

Acknowledgements

The authors would like to thank the entire HL-3 team for providing experimental data and engineering support for the verification experiments in the HL-3 tokamak. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFE03020001, the Zhejiang Provincial Natural Science Foundation of

China under Grant LR23F010005, and the National Natural Science Foundation of China under Grant No. U21A20440 and No. 12405253.

Author contributions

N.N.W., Z.Y., R.L., and N.W. contributed equally to this work. R.L., M.X., Z.Z., and W.Z. conceived this collaboration. N.N.W., R.L., and N.W. led the design & implementation of the dynamics models and reinforcement learning approach; Z.Y. and G.Z. designed and implemented EFITNN; Y.C., X.G., F.G., and B.L. contributed to the real-time on-device implementation. N.N.W., Z.Y., R.L., and N.W. discussed the results; Q.D. and J.L. contributed to the analyses and illustration of some results. N.N.W. and R.L. wrote this manuscript; Z.Y. and N.W. revised the manuscript; all other authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42005-025-02302-y>.

Correspondence and requests for materials should be addressed to Rongpeng Li, Zhifeng Zhao or Wulyu Zhong.

Peer review information *Communications Physics* thanks Jaemin Seo and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025