9th Asia-Pacific Conference on Plasma Physics, 21-26 Sep, 2025 at Fukuoka

Transferable and interpretable disruption prediction based on physics-guided machine learning approaches

C. Shen[1], W. Zheng[1], B. Guo[2], Y. Ding[1], D. Chen[2], X. Ai[1], F. Xue[1], Y. Zhong[1], R, Luo[1], B. Xiao[2], Z. Chen[1]

[1] State Key Laboratory of Advanced Electromagnetic Engineering and Technology, International Joint Research Laboratory of Magnetic Confinement Fusion and Plasma Physics, School of Electrical and Electronic Engineering, Huazhong University of Science and Technology

[2]Institute of Plasma Physics, Hefei Institutes of Physical Science, Chinese Academy of Sciences

e-mail (speaker): shenchengshuo@hust.edu.cn

Transferability and interpretability are becoming increasingly important in disruption prediction research. Since it is expensive to collect a large amount of disruptive data on future tokamaks (such as ITER), it is necessary to transfer existing disruption prediction models to these devices. In addition, future tokamaks have a low tolerance for disruptions, so it is crucial to ensure the reliability of prediction models. To achieve this, models must be interpretable. Interpretability not only helps build trust in model predictions but also allows researchers to explore the causes of disruptions and develop targeted avoidance strategies.

In our research on J-TEXT and EAST [1–3], we applied physics-guided machine learning methods to achieve both transferability and interpretability. The model in the source domain called interpretable disruption prediction based on physics-guided feature extraction (IPD-PGFE).

Specifically, we used three approaches: PGFE, improving the quality of disruption labelling, and estimating standardized parameters in the target domain. Using these methods, we trained a model on J-TEXT data and achieved an AUC of 0.893 on EAST without using any EAST data during training. Based on this, we applied a domain adaptation method called CORAL (CORrelation ALignment) to further improve performance. The models of two target data situations, few shots and no shot from the target tokamak, are shown in figure1. With only 10 disruptive discharges from EAST for training, the model achieved an AUC of 0.947.

The interpretability work is based on PGFE and SHAP algorithms. On J-TEXT, IDP-PGFE showed that the model had learned correct disruption-related knowledge, which supports its reliability. For cross-machine prediction, interpretability results also showed that the model learned disruption features common to both J-TEXT and EAST. We also applied the model to support physics experiments. As shown in figure2, in J-TEXT high-density disruption experiments, the model helped identify that 3/1 and 4/1 RMPs (Resonant Magnetic Perturbations) could raise the density limit by affecting the radiation distribution.

Moreover, the model was able to capture the Greenwald density limit behavior. However, our goal is to help the model learn more about the physics of high-density disruptions, such as MARFE (Multifaceted Asymmetric Radiation from the Edge) and turbulence. To do this, we developed a new disruption prediction model designed to study the contributions of MARFE and turbulence, without including the Greenwald fraction as an input.

References
[1] C. Shen, et al., Nucl. Fusion 63 (2023) 046024.
[2] C. Shen, et al., Nucl. Fusion 64 (2024) 066036.
[3] C. Shen, et al. 49th EPS Conference on Contr. Fusion and Plasma Phys., P4.016 (2023)

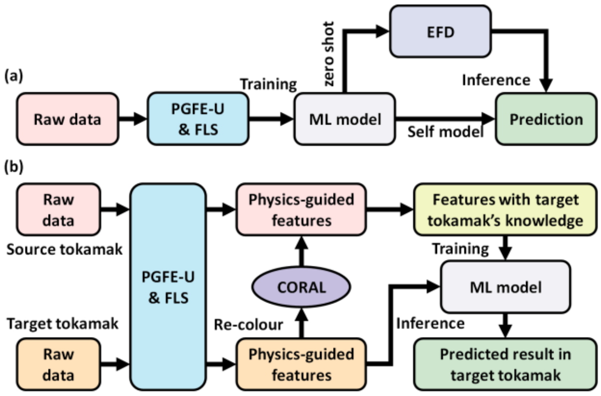
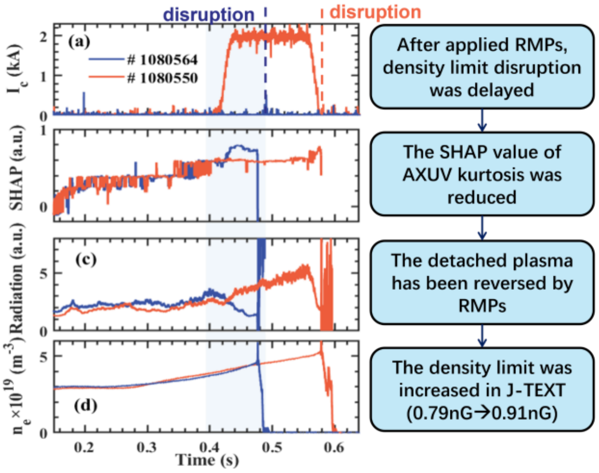


Figure 1 Model structures of the cross-tokamak disruption models. (a) is the model structure with no shot from the target tokamak. (b) is the model structure with few shots from the target tokamak.



After applied RMPs, density limit disruption was delayed

The SHAP value of AXUV kurtosis was reduced

The detached plasma has been reversed by RMPs

The density limit was increased in J-TEXT (0.79nG→0.91nG)

Figure 2 IDP-PGFE empowers AI discovery