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

Towards Open Machine Learning Datasets for Fusion with AI Driven Annotation & Visualization

Samuel Jackson[1], Nitesh Bhatia[1], Niraj Bhujel[2], Josh Blake[1] Rui Costa[1], Nathan Cummings[1], Matthew Field[1], Saiful Khan[2], Abdullah Saleem[1], Stanislas Pamela[1]
[1] United Kingdom Atomic Energy Authority, [2] Science & Technology Facilities Council
e-mail (speaker): samuel.jackson@ukaea.uk

Machine learning (ML) has become a powerful tool for aiding and enhancing data analysis in plasma science[1], presenting exciting opportunities to automate workflows and accelerate scientific discovery. However, machine learning is famously data hungry, requiring costly annotation and curation, which is exacerbated by the specialist knowledge & expertise required within the fusion domain. Furthermore, the general lack of open and widely accessible fusion diagnostic data for benchmarking the ML models chronically hinders development and comparison across experimental facilities.

The FAIR-MAST[2, 3] project at the UKAEA aimed to curate and release an extensive, open catalog of historical diagnostic data from the MAST machine following FAIR[4] principles. However, high quality associated metadata for describing 1) key plasma events occurring within each discharge (disruptions, confinement mode, ELMs, MHD modes, sawteeth, UFOs etc.), and 2) the validity of the diagnostic data is lacking. Such historical metadata is not only crucial for efficient user search and retrieval but is also fundamental towards automating analysis and scientific workflows at experimental facilities. Furthermore, acquiring such a collection of metadata on the MAST machine would provide a valuable open data for model comparison within the fusion domain.

To address these challenges, we present progress on the development of an AI driven, human-in-the-loop web UI platform for the annotation and curation of diagnostic data, enabling us to bring together domain expertise with AI tools to accelerate and automate its collection. Crucially, our platform provides a framework to plug in custom models for automatically annotating regions of interest within diagnostic data traces. Annotation models could be anything between classical signal processing to ML models. Our platform is built on a modern web stack and fully open source. The platform is designed to be machine agnostic and extensible, supporting a wide variety of data access layers, including both domain specific and general formats. Metadata captured by the platform is stored in a machine-readable database accessible through a REST API.

We will discuss future plans for the development of a public database of annotated data and downstream tasks towards providing the community with a dataset for model intercomparison within the FAIR MAST project as well as how we may “close the loop” on the annotator training.

References

[1] R. Anirudh, *et al*, IEEE Trans. Plasma. Sci., doi: 10.1109/TPS.2023.3268170 (2022)
[2] S. Jackson *et al,* IEEE Trans. Plasma. Sci., doi: 10.1109/TPS.2025.3583419 (2025)
[3] S. Jackson *et al,* Software X, doi: 10.1016/j.softx.2024.101869 (2024)
[4] Wilkinson, *et al.* Nature Sci. Data, doi: 10.1038/sdata.2016.18

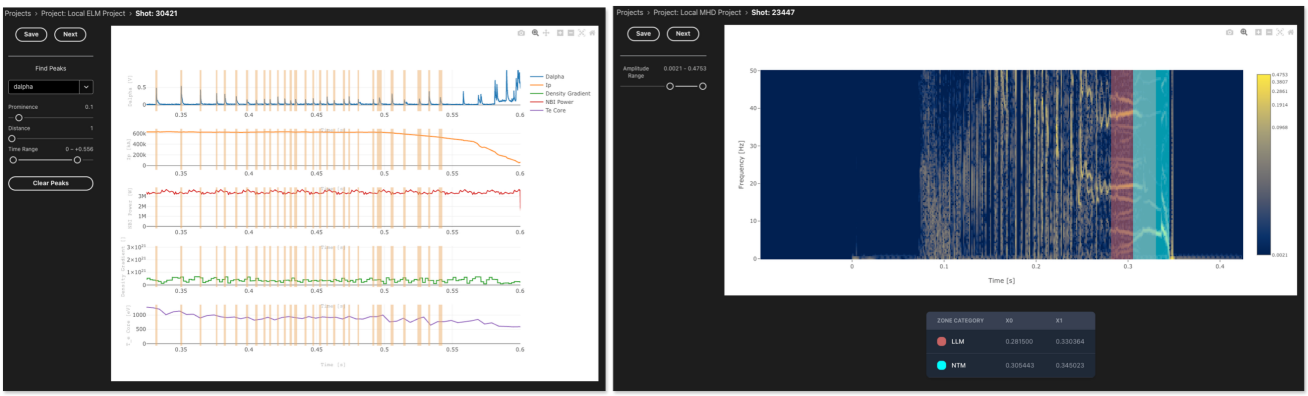

Figure 1: Examples of our UI for fusion diagnostic data. The UI provides the diagnostician with the ability to interactively annotate complex multi-model diagnostic data. *Left*: semi-automatic annotation of ELM peaks in relevant diagnostic traces (Dα, Ip, T_e, etc.). *Right:* interactive annotation of NTM and LLM modes in Mirnov coil spectrograms.