EUROfusion

Swiss Plasma Center

E-TASC 2nd Meeting - Garching

A Scalable Production Framework for Multimachine Databases

Alessandro Pau and team

Swiss Plasma Center - EPFL

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Multimachine Database: a challenge

Swiss Plasma Center

Fusion experiments generate vast, heterogeneous, multi-modal data across devices with different diagnostics, naming conventions, and legacy systems.

This fragmentation is a **major bottleneck** for:

- **Downstream tasks** — AI/ML training, control-related tasks, simulations, etc.
- **Scientific exploitation** — Cross-machine statistics, scaling laws, extrapolation to ITER
- **Reproducibility** — Provenance tracking, versioned workflows, FAIR compliance
- **Collaborative validation** — Expert consensus at scale, standardized labeling and metadata templates

What the Community Needs

→ Standardized data curation tools — not ad hoc per-device processing

→ Interoperable data mapping tools

→ Scalable infrastructure for heterogeneous data at production scale

→ End-to-end provenance and metadata for reproducibility

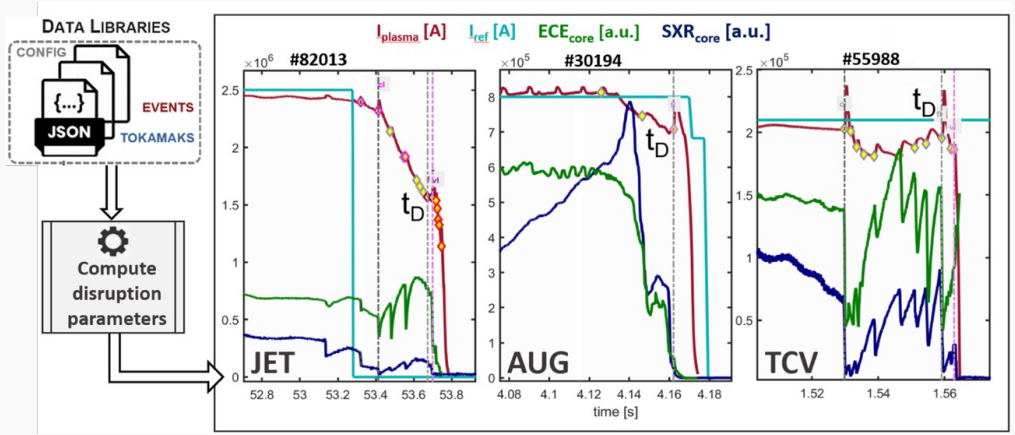
→ Tools for collaborative validation orchestrated across a mixture of physics experts

→ AI-ready labeled datasets with quantified uncertainty

Multi-machine databases, main use cases

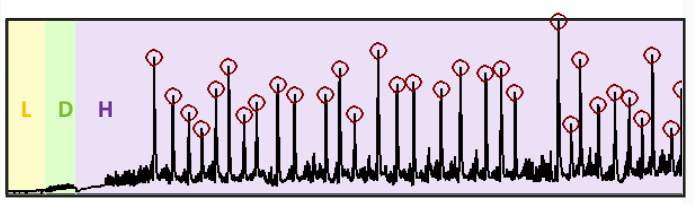
Swiss Plasma Center

Disruption Database



Validated multi-machine disruption database covering TCV, JET-ILW, AUG, MAST ~ several thousands of disruptions validated.

L-H transition database

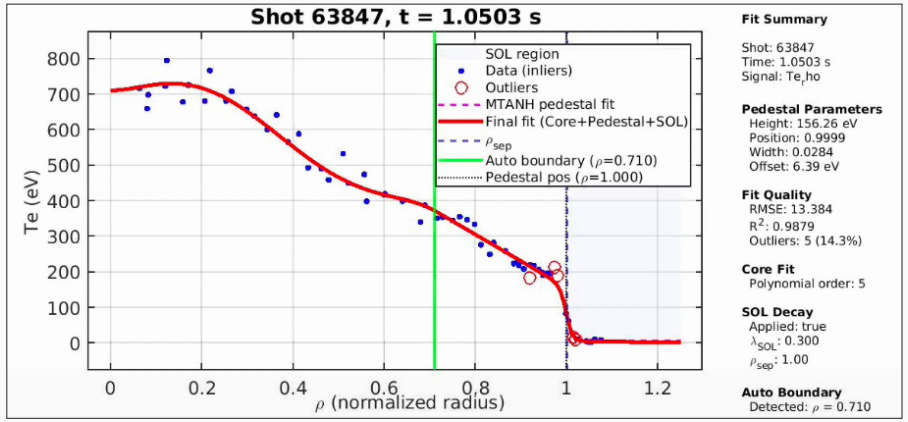


Conf. state database of ~1k discharges including JET & TCV experiments. Data-driven models with calibrated uncertainty (JET-PPF)

https://zenodo.org/records/14996664
https://zenodo.org/records/16631053

DEFUSE

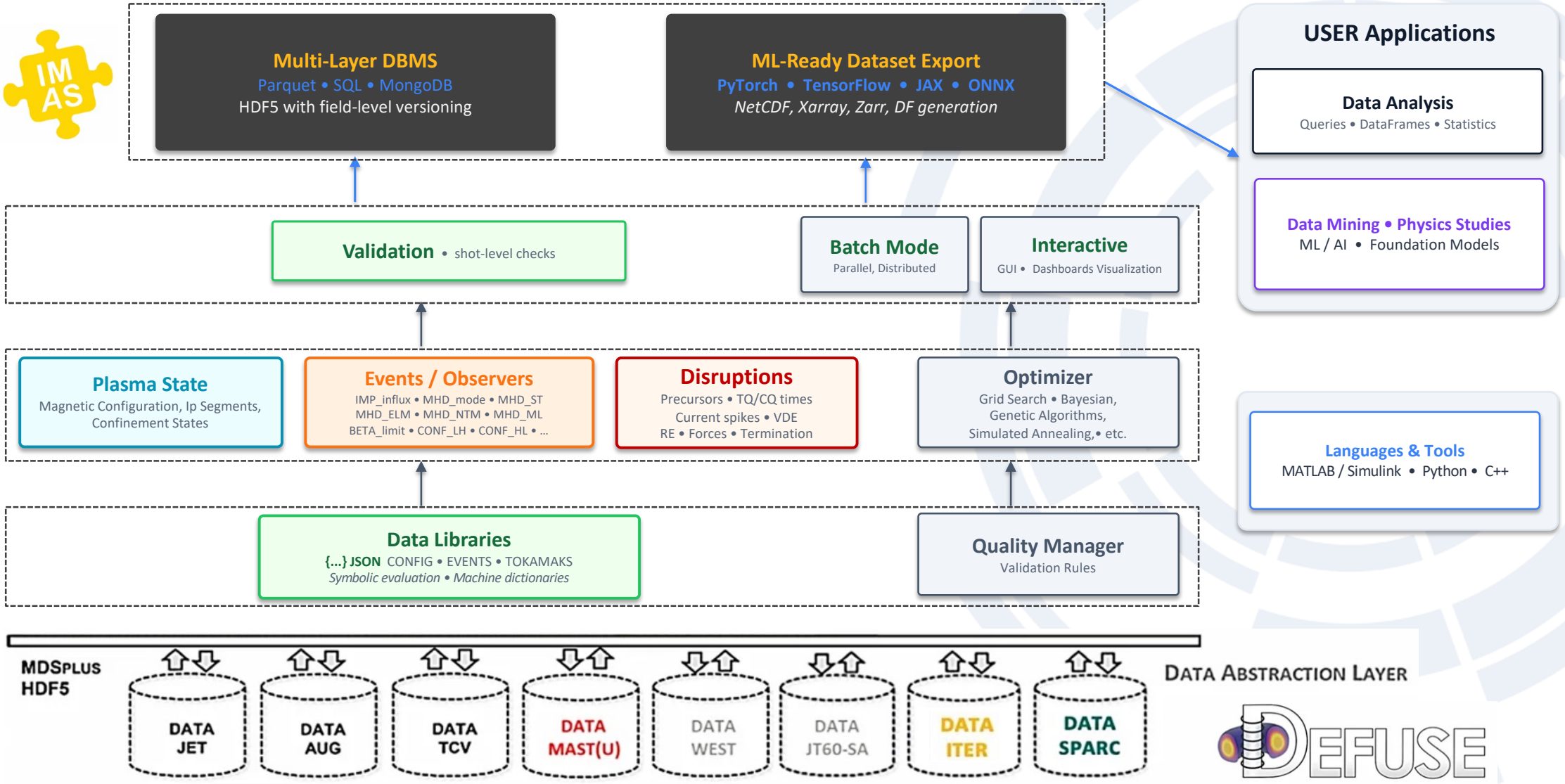
Pedestal database



Validation for EUROfusion WPTE campaigns & IMAS-ready for Gateway storage.
Automated initial fit and outlier rejection based on Gaussian Process Regression (GPR) and principled Uncertainty Quantification (UQ)

DEFUSE a scalable framework for fusion experiment

Swiss Plasma Center



Data libraries in DEFUSE

Swiss Plasma Center

TOKAMAKS


```
AUG  {4}
JET  {5}
MAST  {5}
TCV  {5}
    DATA_dictionary {333}
    DATA_processing {5}
    DIAG_dictionary {22}
    Data_access : local
    Data_format : hdf5
Dictionaries_info {6}
Table_DATA {35}
Release {5}
version : 20
```


Data Dictionary


```
BETAN {8}
    Name : BETAN
    Format : 1D
    Description : normalized toroidal beta
    Units : a.u.
    ErrorBar : null
    Origin {3}
        variables [4]
            0  : BZERO
            1  : a_minor
            2  : IPLA
            3  : BETAT
        coefficients {1}
            k    : 100000000
        expressions [1]
            0  : k * abs((BETAT ./ IPLA) .* BZERO .* a_minor)
    IMAS : /equilibrium/time_slice(itime)/global_quantities/beta_normal
    MachineID : 3
```


- **Template DATA dictionary**
  main source reference for SQL & MongoDB
- **Symbolic evaluation**
  dynamic and abstract representation and transformation of data.
- **Data mapping**
  Custom plugins (IMAS, etc.)

Data libraries in DEFUSE

Swiss Plasma Center

EVENTS

```
Catalog {27}
Detectors {4}
Discharge {8}
Disruptions {5}
Release {5}
version : 2
```

Detectors & Observer

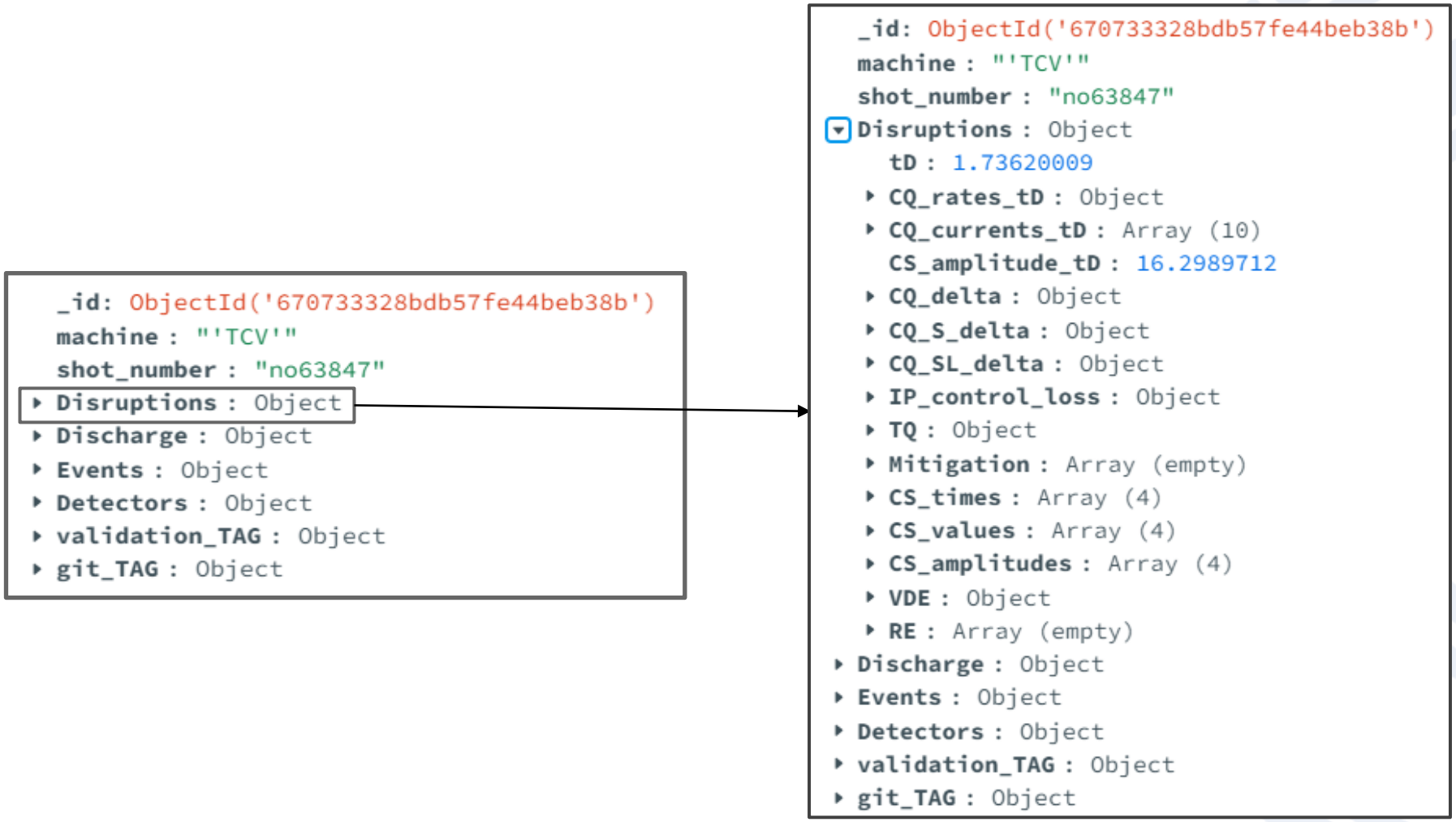
```
Detectors {4}
    info_tag : dictionary for detector main parameter settings
    Dictionary {12}
    Catalog_with_dictionary {40}
        Fields [37]
        DUMMY {12}
            label : DUMMY
            machine : JET
            Type : OffNormal
            ComparisonMethod : Equal
            Threshold : 1
            timeRef : null
            AssertionTime : 1
            minDelta : 1
            maxDelta : 1
            t_offset : 1
            diag : value
            model : physics-based
```

- **Dictionaries and Catalogs**

  encoding the description of disruption characteristic parameters, off-normal events & causes, and plasma state (main discharge segments)

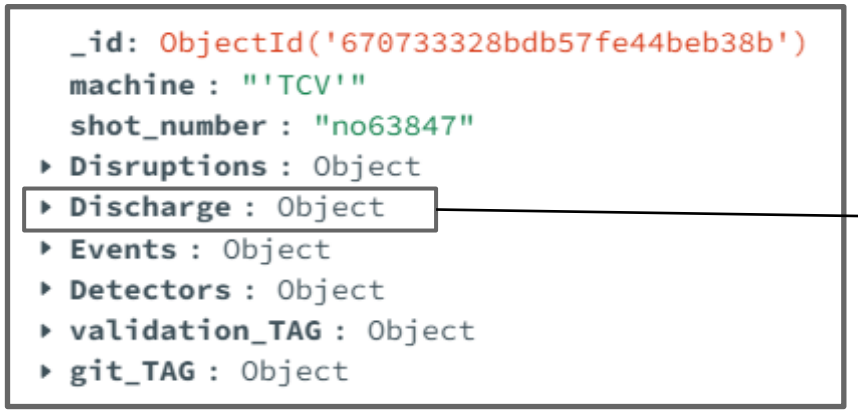
Database entries

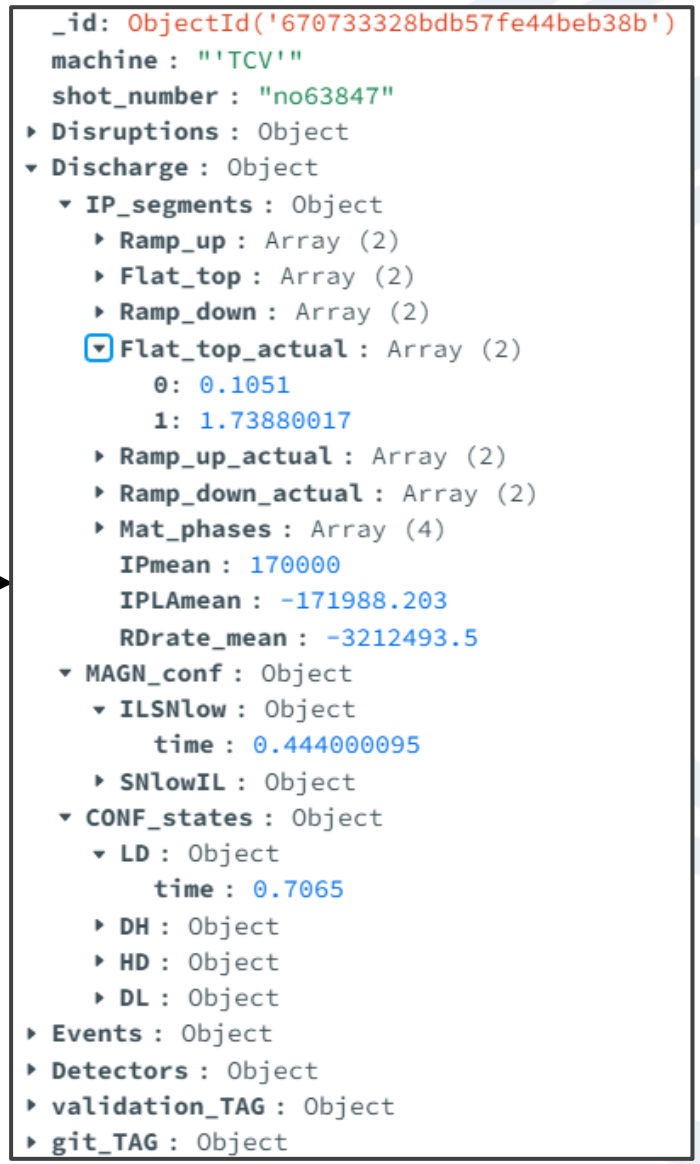
Swiss Plasma Center



Database entries

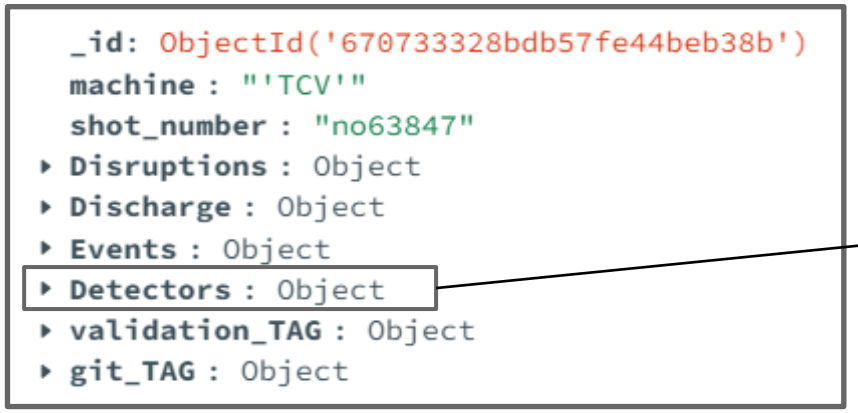
Swiss Plasma Center

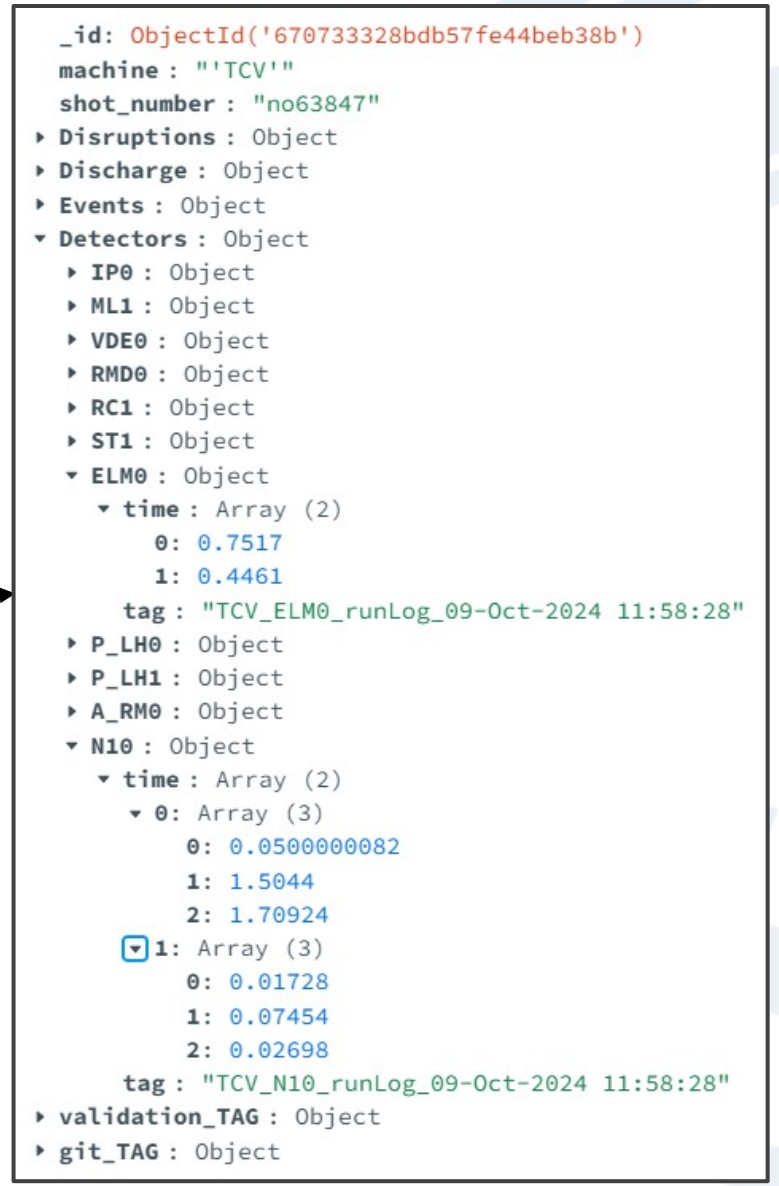




Database entries

Swiss Plasma Center

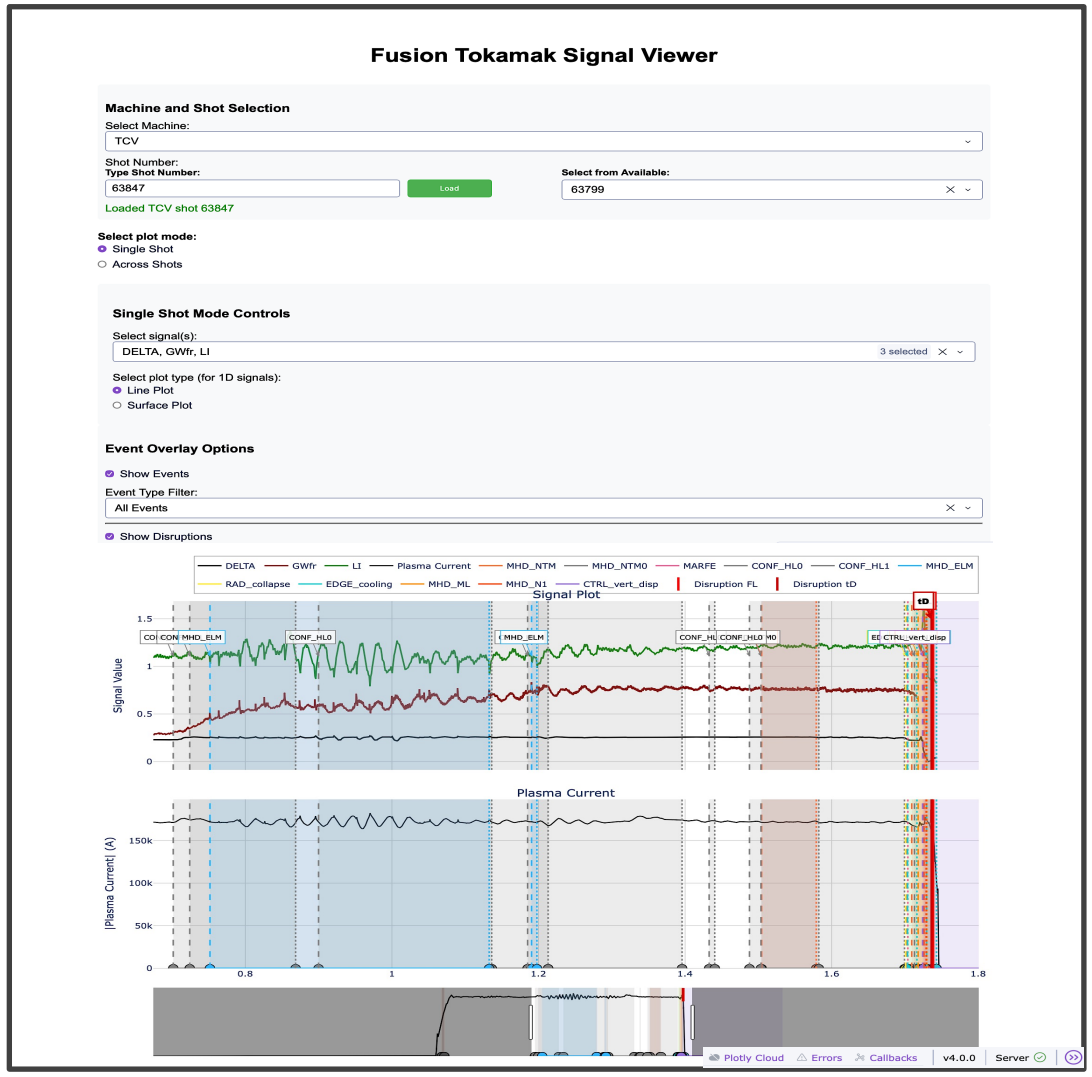




Collaborative Environment for Validation

Swiss Plasma Center

Interactive Tools for the Community



Why Collaborative?

Mixture of physics experts

Validation requires device-specific domain expertise. No single person can label all events accurately — orchestrated collaborative review is essential.

Interactive validation

Automated detectors provide initial labels; experts validate and correct through interactive GUIs and SQL-baked dashboards for advanced visualization.

User annotation & review

Multiple experts contribute annotations on specific discharges. Iterative refinement with versioning ensures long-term database quality and traceability.

Integration with DMP-IMAS catalogues

Web-based dashboards and advanced visualization utilities can be integrated with DMP-IMAS catalogues

Outlook: Towards Trustworthy Fusion Data

Swiss Plasma Center

Standards & Procedures

Data Curation & labeling at scale
Scalable infrastructure for heterogeneous data at production scale; Standardized data curation tools — not ad hoc per-device processing

Standardized Metadata
Enforce consistent metadata schemas across devices. Machine descriptions, diagnostic specs, and processing parameters must be interoperable and self-documenting.

Validation Orchestration
Enable orchestrated validation from a mixture of physics experts — device-specific knowledge with cross-machine consistency checks. Not a single-group effort.

Database Governance
Standardize versioning, backup, and maintenance procedures. Long-term sustainability requires clear governance, access policies, and community-driven data stewardship.

Uncertainty & Reproducibility & AI

VVUQ & End-to-End Uncertainty
Data-driven models must propagate uncertainty from experimental measurements, through processing pipelines, to final labels and predictions. Calibrated uncertainty is non-negotiable.

Reproducibility in Science
Track full data provenance: from raw signals through processing workflows to curated datasets. Versioned pipelines, atomic workflow snapshots, persistent identifiers (DOIs).

Open & FAIR Science
Open-source tooling, public repositories, standardized outputs. Broad community reuse, citation, and contribution. Lower barriers for new devices and research groups.

AI-assisted workflows
AI-assisted curation, automated pipeline development, and LLM-based tools can dramatically accelerate database development, scientific exploitation, and code generation.

Backup slides

Machine-Agnostic Data Curation & Labeling at Scale

Swiss Plasma Center

INGEST Automated loaders from MDSplus / HDF5 / IMAS / UDA with caching

▼

VALIDATE Quality assessment: gaps, saturation, outliers, SNR, …

▼

PROCESS Resampling, rescaling, feature extraction with full provenance

▼

LABEL Automated + expert-validated labels for system states, events, etc.

▼

EXPORT ML-ready dataset and efficient data-loaders

Labeling & Curation Strategy

Hybrid model-assisted labeling and validation

Plasma state and events detection framework (MHD-modes, sawteeth, ELMs, thermal collapses, off-normal events) provides automated labels. Physics-based & data-driven models with configs encoded in data libraries for reproducibility.

Expert validation through GUIs

GUIs for interactive review and correction. Confinement states, L-H transitions, disruption event chains validated by domain experts (automated labeling will never be perfect, but we can flag what is “suspicious”).

Scalable batch execution

Parallel computation with checkpoint-based batch processing. support. Performance: few tens of seconds for full shot characterization.

Standardized output formats

Standardized dataframes with labels. HDF5 with rich metadata. Full traceability from raw signal to final label, including the entire data processing pipeline.

Collaborative Environment for Validation

Swiss Plasma Center

CI/CD Pipelines

GitLab CI/CD with Jenkins integration. Automated build, test (~80% coverage target), deployment. Pre-commit validation blocks regressions.

Multi-Layer DBMS

SQL (SQLite) relational model for IMAS mapping with role-based auth. NoSQL (MongoDB) for events & plasma state. JSON1 for nested structures.

Parallel Processing

Configurable workers for batch execution. Checkpoint-based with merge & recovery. Target: <30s/shot for full disruption characterization.

Multi-Format I/O

HDF5 (IMAS backend), Parquet/Xarray (analytics), NetCDF. Automated table generation: DATA, COORD, LABELS, NORM variants per shot.

Versioning & Provenance

Shot-level validation with rollback & logging. Role-based auth. Field-level provenance tracking. Unit testing framework with regression detection.

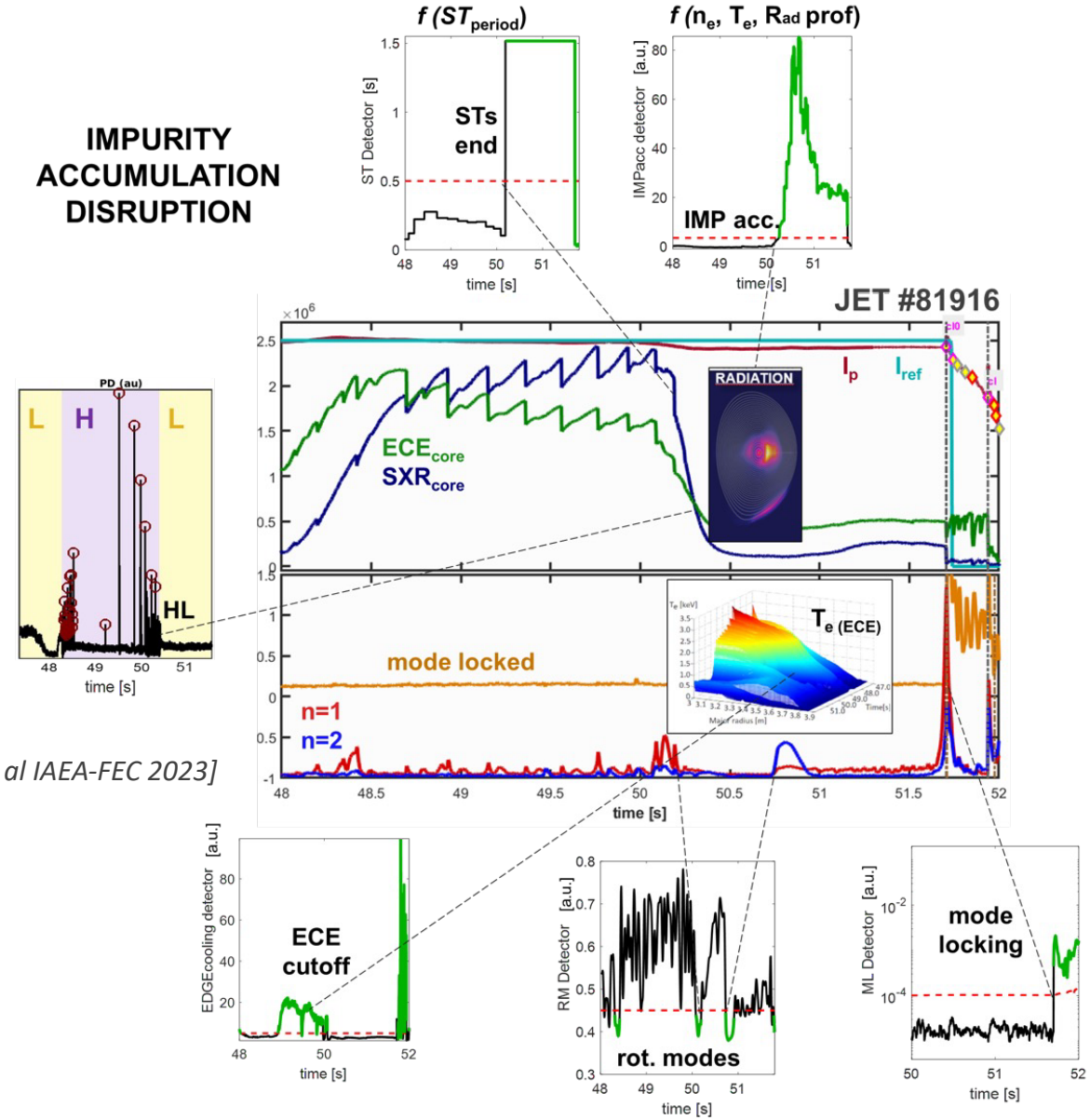
Gateway Migration

Stack migrating to new EUROfusion Gateway. 20–50 TB storage. IMAS-compliant backend. Web service interfaces for collaborative access.

EUROfusion DDB: Events and Modelling

Swiss Plasma Center

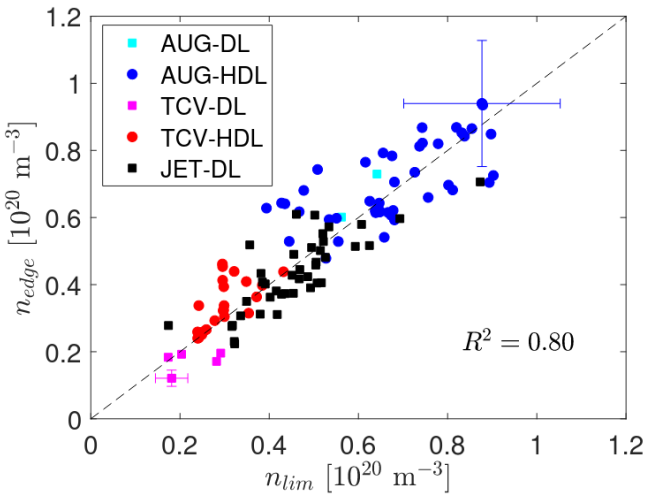
IMPURITY ACCUMULATION DISRUPTION



REF: [A. Pau et al IAEA-FEC 2023]

Density Limit First-Principles Scaling

REF: [M. Giacomin et al PRL 2022]



1. P_{SOL} dependence of n_{eLIM}
2. $n_{e_{LIM}}{}^{ITER} \simeq 2 \cdot n_{e_{GW}}{}^{ITER}$