Interpretable data-driven disruption predictors to trigger avoidance and mitigation actuators on different tokamaks

Cristina Rea[1]

with K.J. Montes[1], R.A. Tinguely[1], J.X. Zhu[1], R.S. Granetz[1], W. Hu[2], B. Shen[2], Q.P. Yuan[2], B.J. Xiao[2], J. Barr[3], E. Olofsson[3], T. Yokoyama[4], J. Lee[5], J. Kim[5], K. Erickson[6], G. Dong[6], W. Tang[6]

[1]MIT PSFC, Cambridge, MA, USA
[2]ASIPP, Hefei, China
[3]General Atomics, San Diego, CA, USA
[4]GSFS, Univ. of Tokyo and Research Fellowships for Young Scientist, JSPS
[5]NFRI, Daejeon, Korea
[6]PPPL, Princeton, NJ, USA

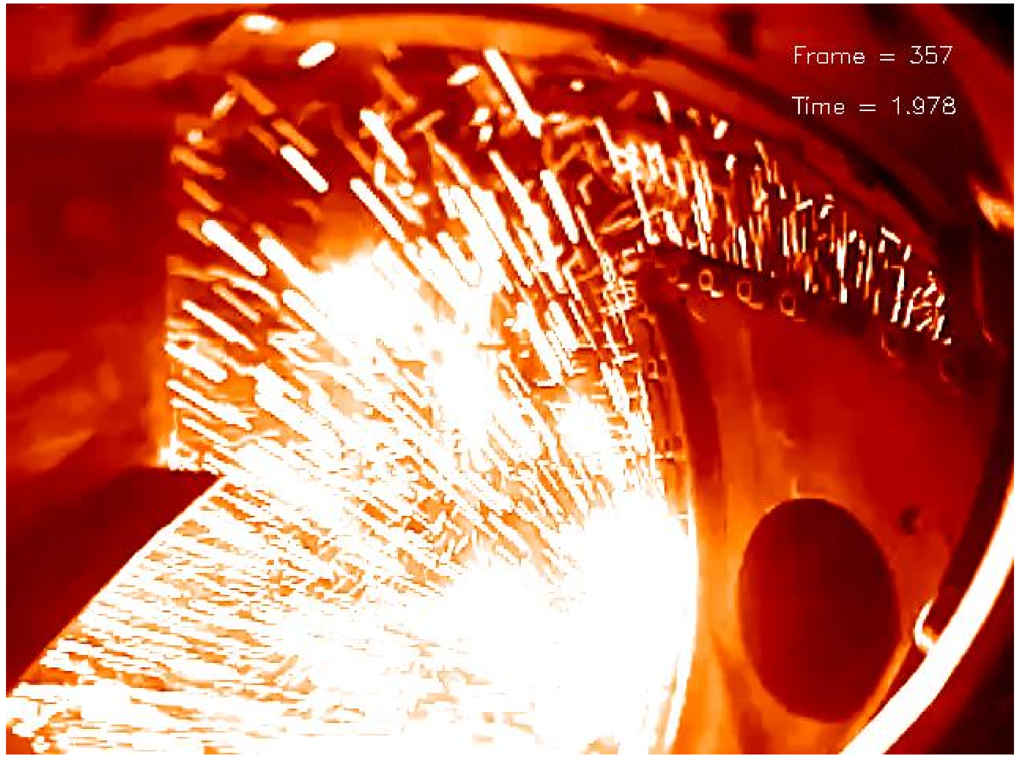
1st IAEA Technical Meeting on Plasma Disruptions and their Mitigation
July 20-23, 2020

Outline

- **Disruption Prediction**
  - Intro and motivations
- **Overview Of Interpretable Algorithms Across Devices**
  - DIII-D
  - EAST
  - KSTAR
  - JT-60U
- **Summary And Conclusions**

Plasma pushed close to operational limits often leads to instabilities onset or control faults: unintentional disruptions

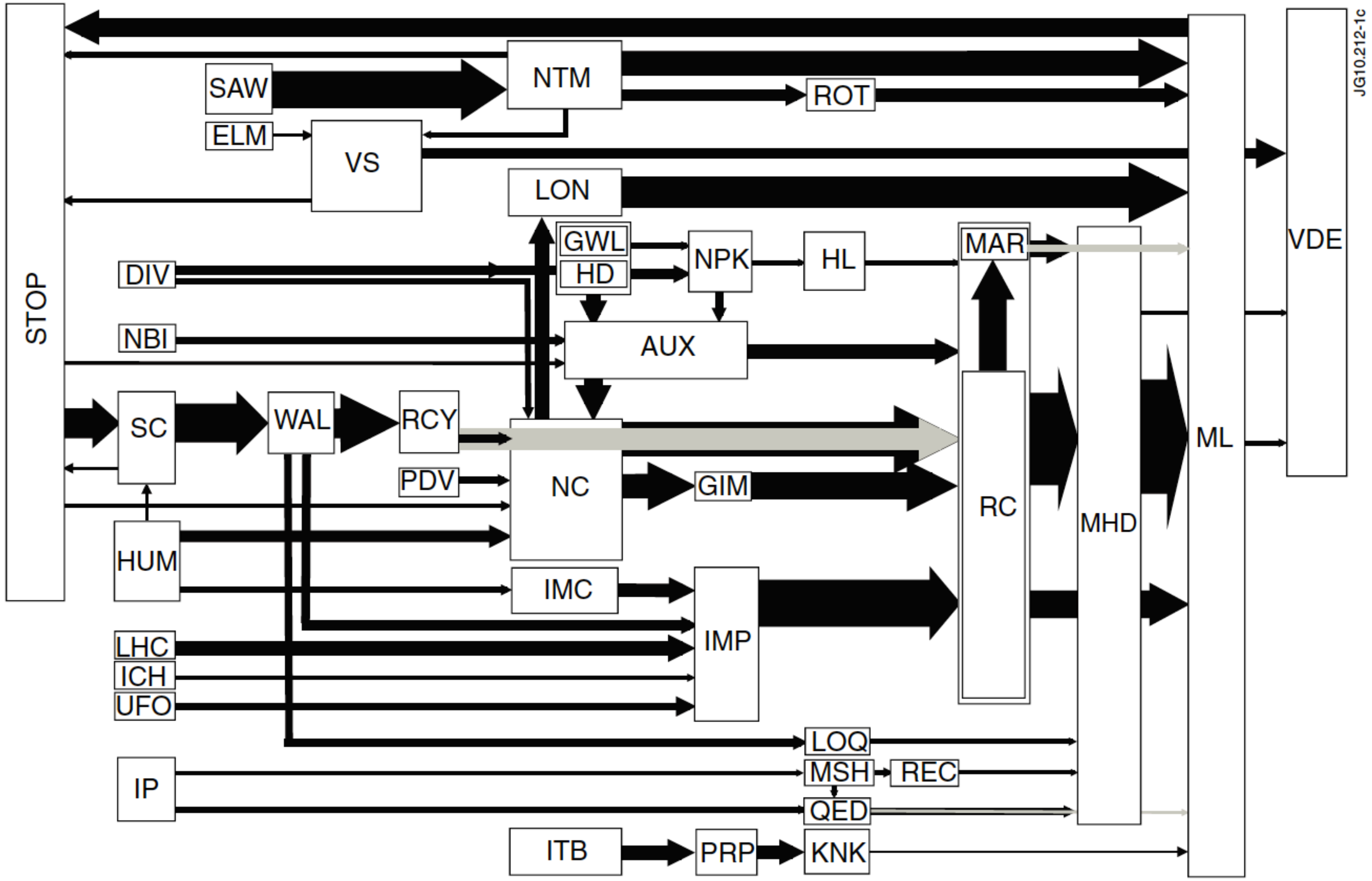
- Disruptions related to **peak plasma performances:** higher stored energy, longer confinement times…
- **Real-time prediction** and **avoidance**, with **mitigation**, mandatory when scaling to reactor sizes and forces.



View from visible camera of disruption on Alcator C-Mod.

Statistical studies show complex chains of events: disruption precursors

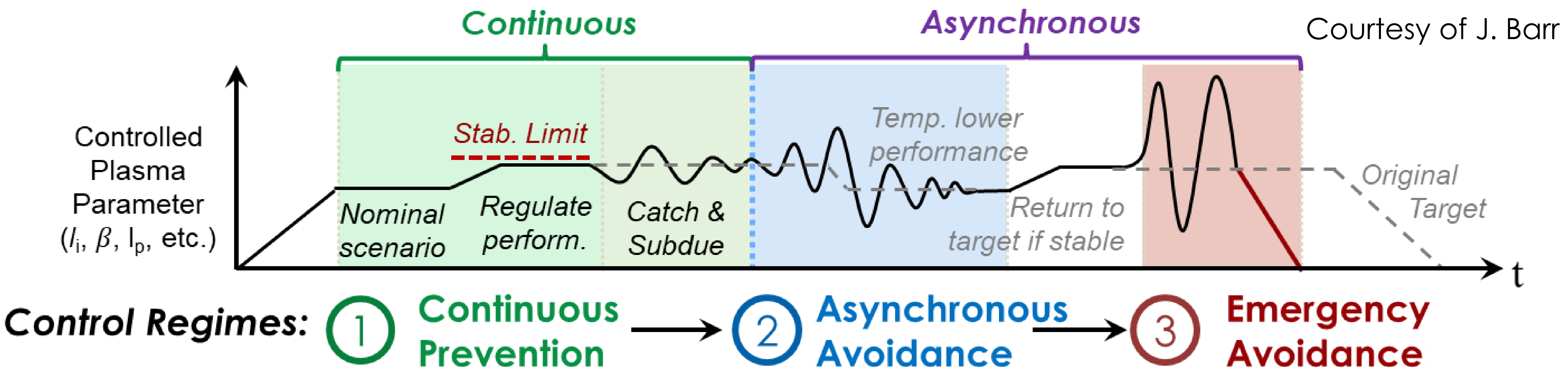
possible disruptive chains of events



Disruptions as final loss of control: successful precursors identification can inform plasma controllers on proper actuators.

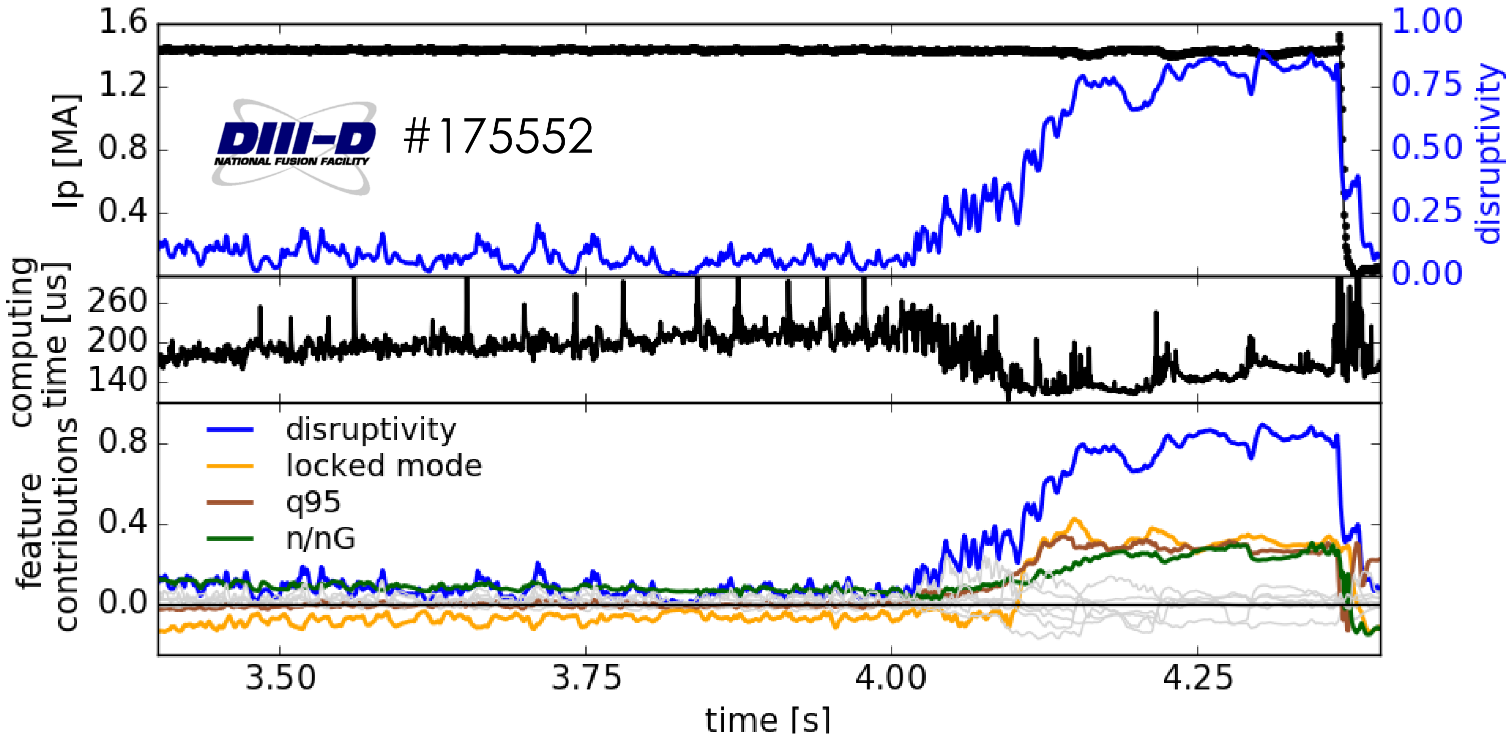
 De Vries et al. NF 51 (2011) 053018 "Survey of disruption causes at JET"

Active monitoring and prediction of soft/hard limits necessary to inform transition across ops boundaries



Proximity to stability boundaries need to be actively controlled: different control regimes

Interpretable ML models for disruption prediction useful resources to identify in real-time stability boundaries



- On **DIII-D** and **EAST**, the Disruption Prediction via Random Forest algorithm (DPRF) computes disruptivity and interprets its drivers in real-time.
- On **KSTAR**, similar exploration through Random Forest.
- **JT-60U**: Sparse Modeling by Exhaustive Search and Support Vector Machine.

Outline

- **Disruption Prediction**
  - Intro and motivations
- **Overview Of Interpretable Algorithms Across Devices**
  - DIII-D
    Deep Learning
    DPRF
    Survival Analysis
  - EAST
  - KSTAR
  - JT-60U
- **Summary And Conclusions**

Deep Learning extracts general representations of disruptive behavior across devices

J.X. Zhu et al, "A new Deep Learning architecture for general disruption prediction across tokamaks", *this meeting*

- Numerical experiments with aggregated **DIII-D**, **C-Mod**, and **EAST** data show DL learns disruptive characteristics: **device-independent knowledge**.
- Non disruptive data results **device-specific**, not improving performances.
- **Limited disruptive** data from target device still needed for prediction, as well as **all available non-disruptive** data.

Fusion Recurrent Neural Network (FRNN) with 0-D scalar inputs installed in DIII-D control system (PCS)

- FRNN Long Short-term Memory block implemented in DIII-D PCS.

  J. Kates-Harbeck, et al., Nature (2019)

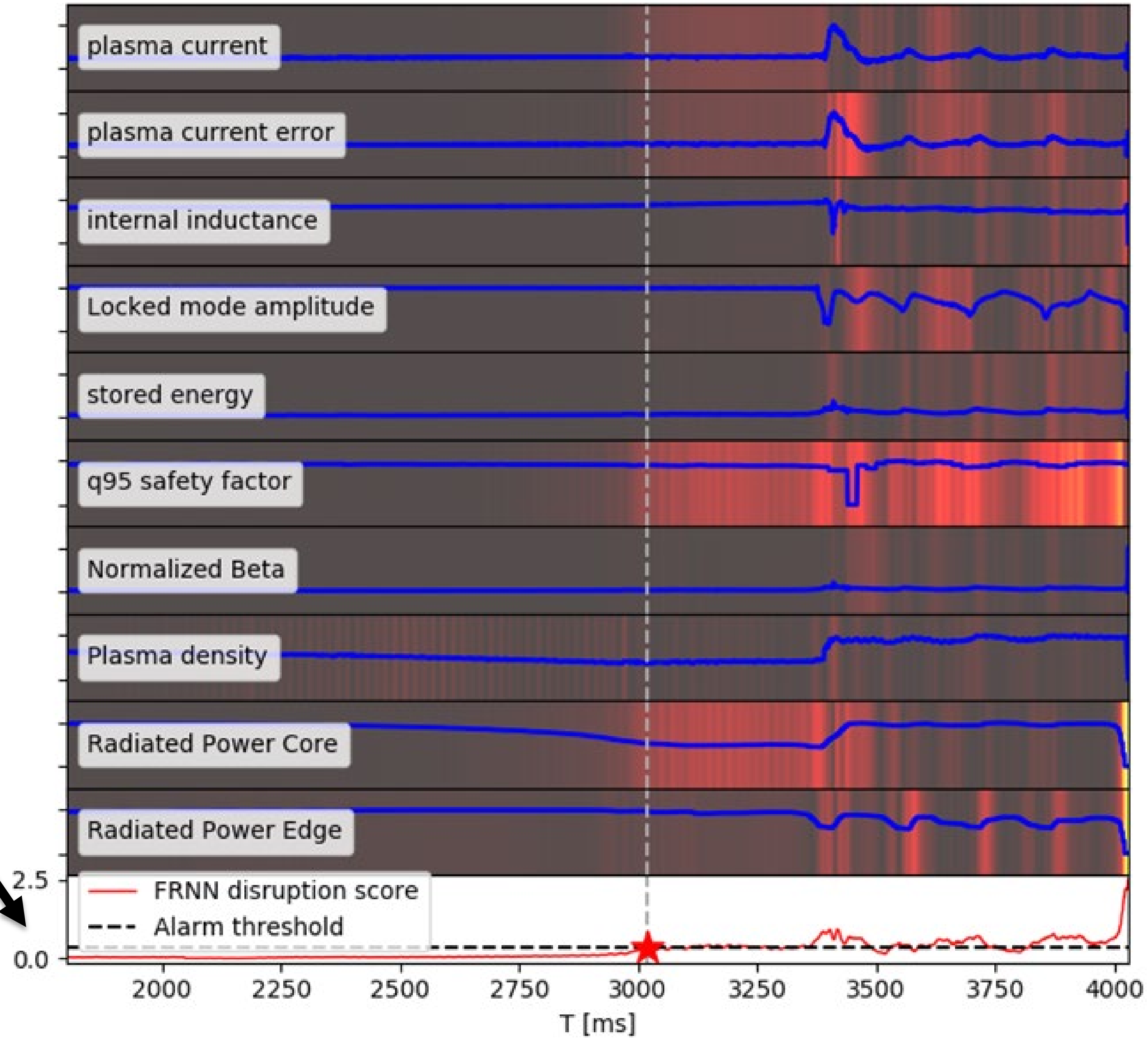
- **Computing time < 2ms** for real-time eval.

- Associated actuator response studies in progress.

- FRNN **heat map** shows *disruption score at alarm time* most sensitive to **radiated core power** and **q95**.

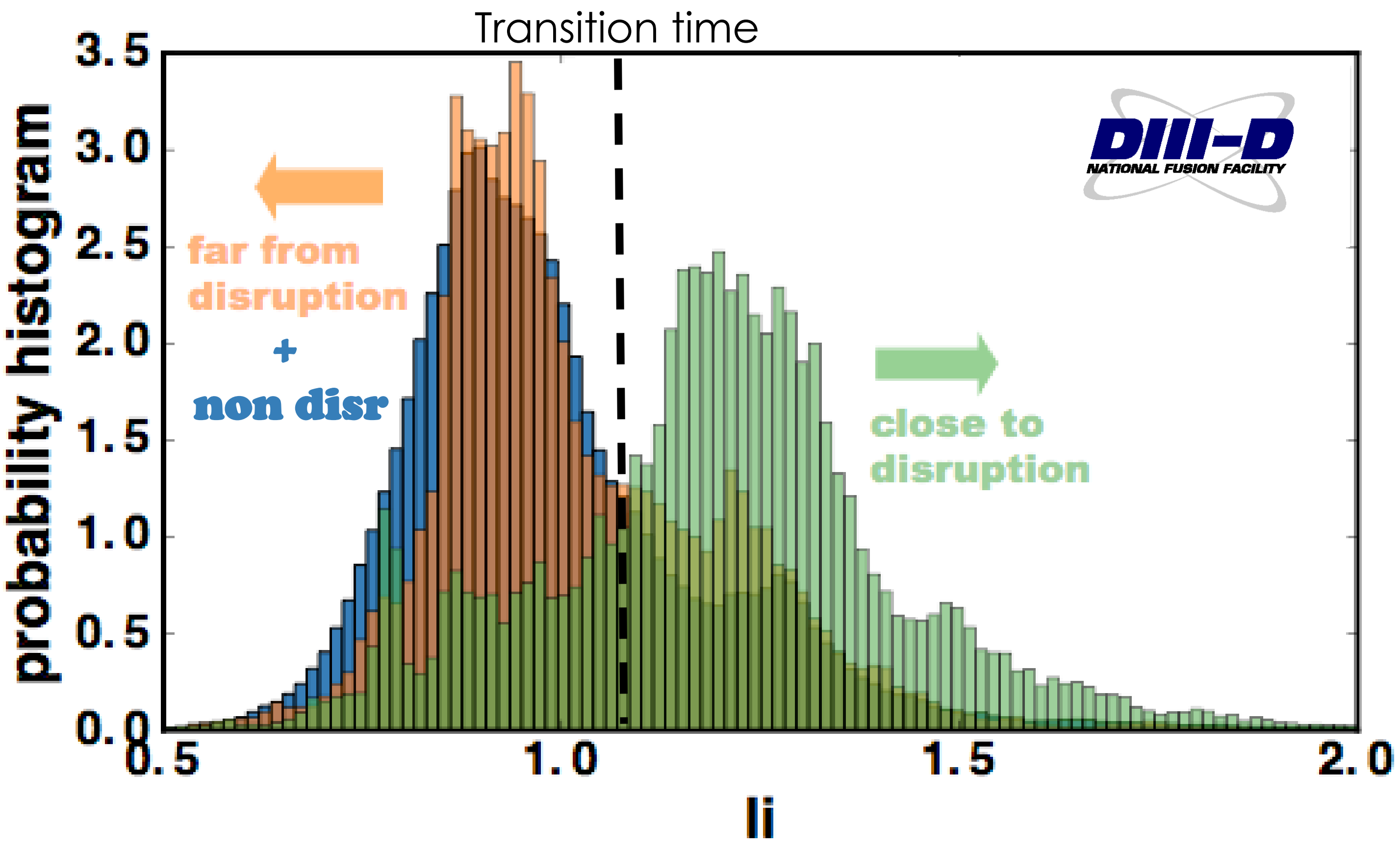
W. Tang et al., accepted 2020 IAEA FEC paper TH/7-1Ra

FRNN Sensitivity Study – shot 164582



DIII-D NATIONAL FUSION FACILITY

DPRF supervised binary classification algorithm: identify transition non disruptive – disruptive phases



DPRF is based on the **Random Forest** ensemble algorithm → collection of decision trees:

Provides **metrics of interpretability**.

- **Fixed time** for **transition** from safe to disruptive operational space.
- Training set thousands of discharges, **agnostic to disruption type**.
- **Offline cross-machine** investigation **0-D features** (**flattop** data).

C. Rea and R.S. Granetz, Fus. Science Tech. 74 (2018)
C. Rea et al., Plasma Phys. Control. Fusion 60 (2018)
C. Rea et al., Nucl. Fusion 59 (2019)
K. Montes, C. Rea et al., *Nucl. Fusion* 59 (2019)

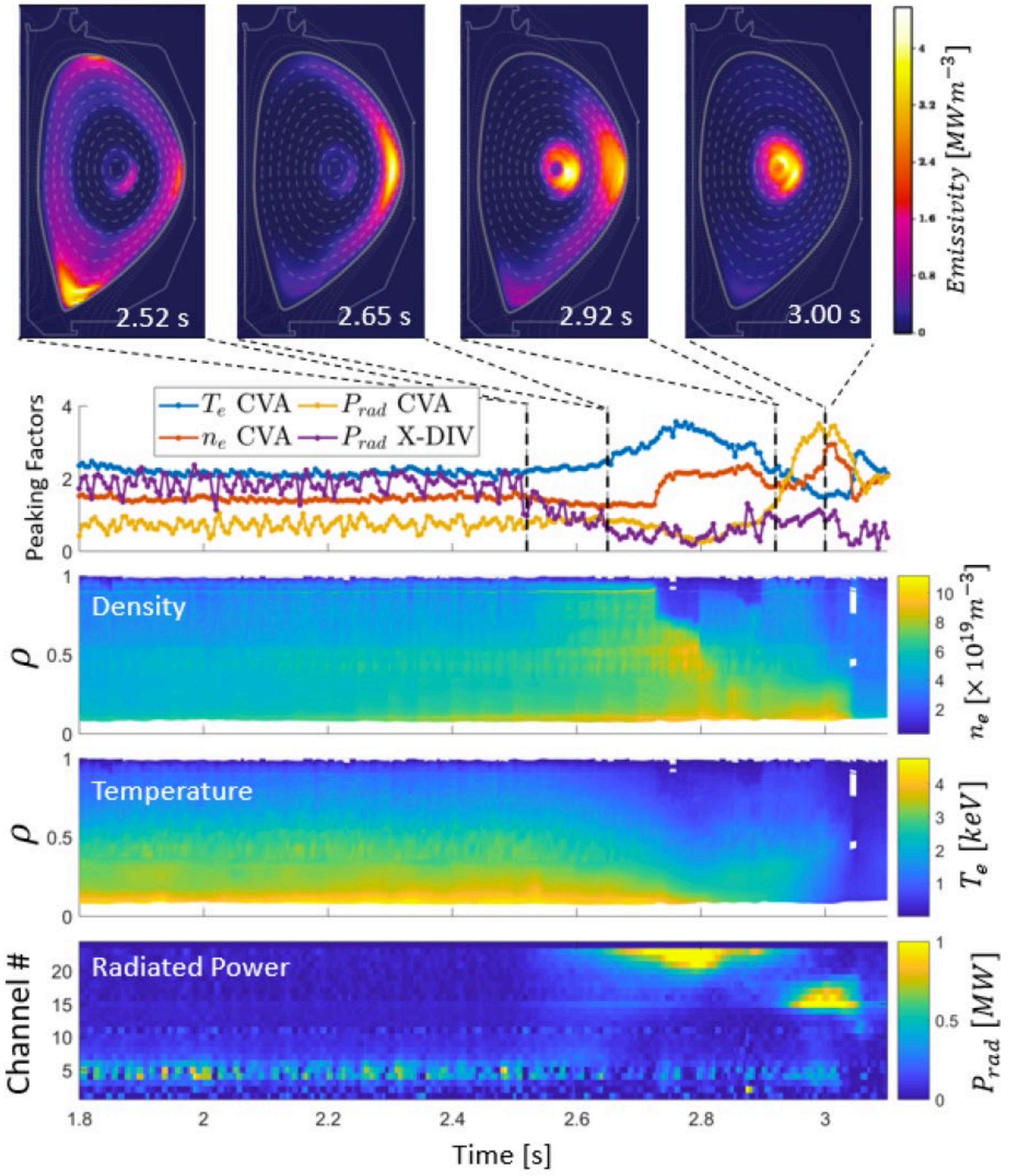
→ DIII-D DPRF 2.0

DPRF 2.0: to detect earlier disruptive precursors, feature engineering and dimensionality reduction

- **1D/2D profile information** compressed into **peaking factors**.

- Profile diagnostics mapped onto flux surfaces or core / divertor regions.

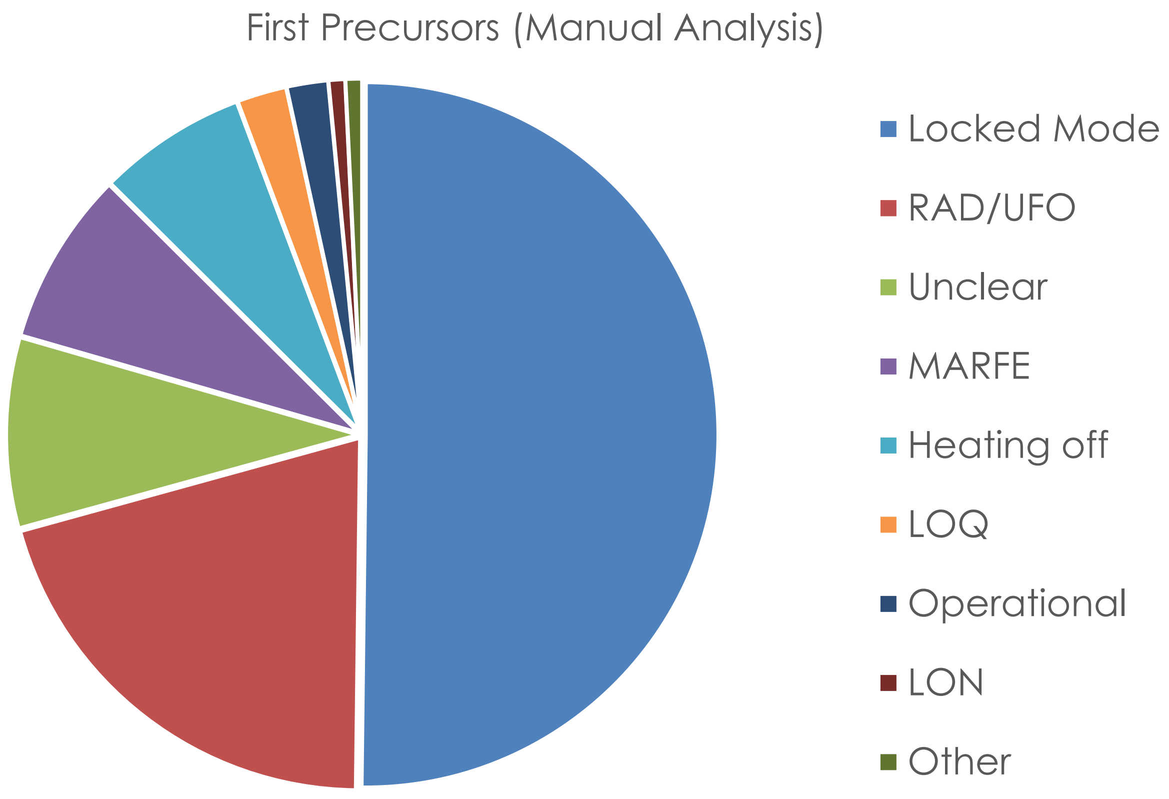
Peaking factors are interpretable, easy to calculate in real-time



A. Pau et al., *IEEE TPS*, 46 (2018)
A. Pau et al., *Nucl. Fusion* 59 (2019)
C. Rea, K.J. Montes, A. Pau, R.S. Granetz, O. Sauter, "Progress Towards Interpretable Machine Learning-based Disruption Predictors Across Tokamaks", *Fus. Science Tech.* (2020)

DPRF 2.0: improved label classification by detecting transition between specific operational boundaries

- **First disruptive precursors manually identified** for hundreds of discharges → Transition into unstable operational space: scenario detection.



- ML algorithms: training composition can skew the sensitivity of the model towards certain scenarios.
- **Need for (automated) identification of disruption causes.**

K. Montes et al, "Accelerating Disruption Database Studies with Semi-Supervised Learning", *this meeting*

S. Sabbagh et al, "Progress on Tokamak Disruption Event Characterization and Forecasting Research and Expansion to Real-Time Application", *this meeting*

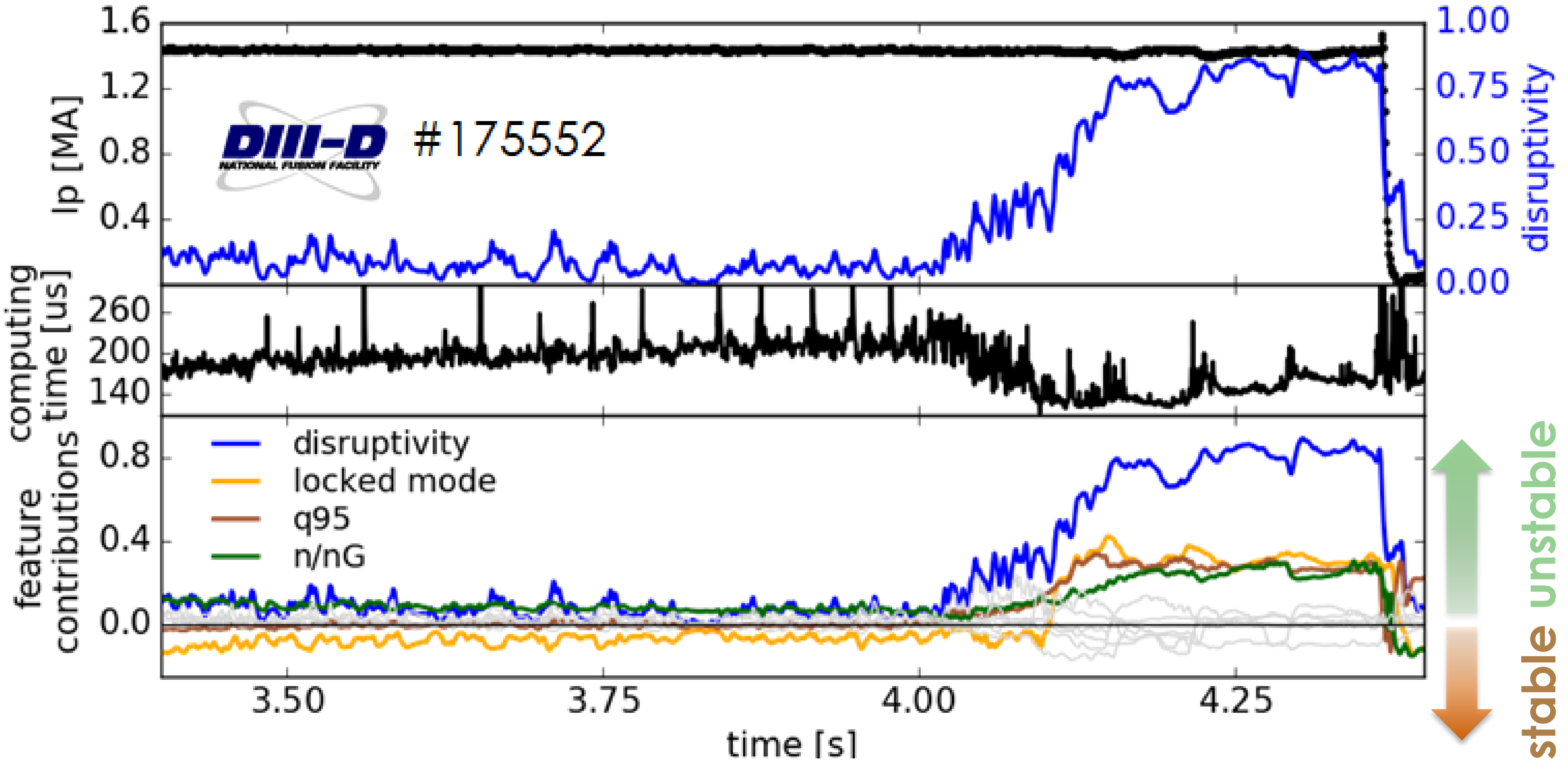
DIII-D DPRF 2.0 – peaking factors added to 0-D inputs
Feature contributions to explain disruptivity drivers

DPRF 2.0
n/n_G
W_{mhd}
β_n
$(I_p - I_{prog})/I_{prog}$
ℓ_i
$B_r^{n=1}/B_\phi$ Locked mode
κ Elongation
δ Triangularity
ζ Squareness
V_{loop}
T_e peaking
n_e peaking
P_{rad} Core Peaking
P_{rad} Divertor Peaking

Decision paths in DPRF trees provide average **measures of explainability** by assigning (±) **contributions** to input features during inference.

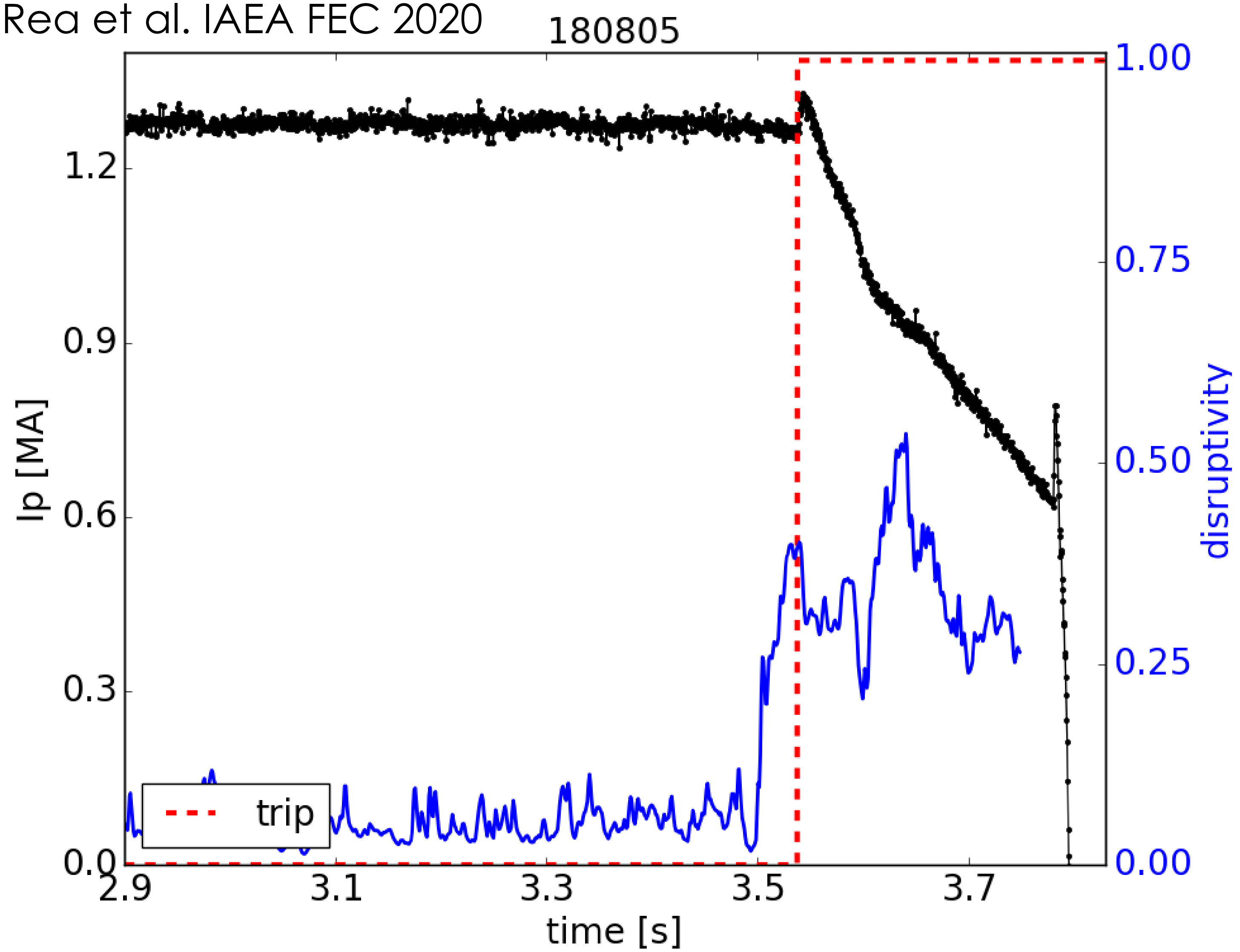
(see example in additional slides)

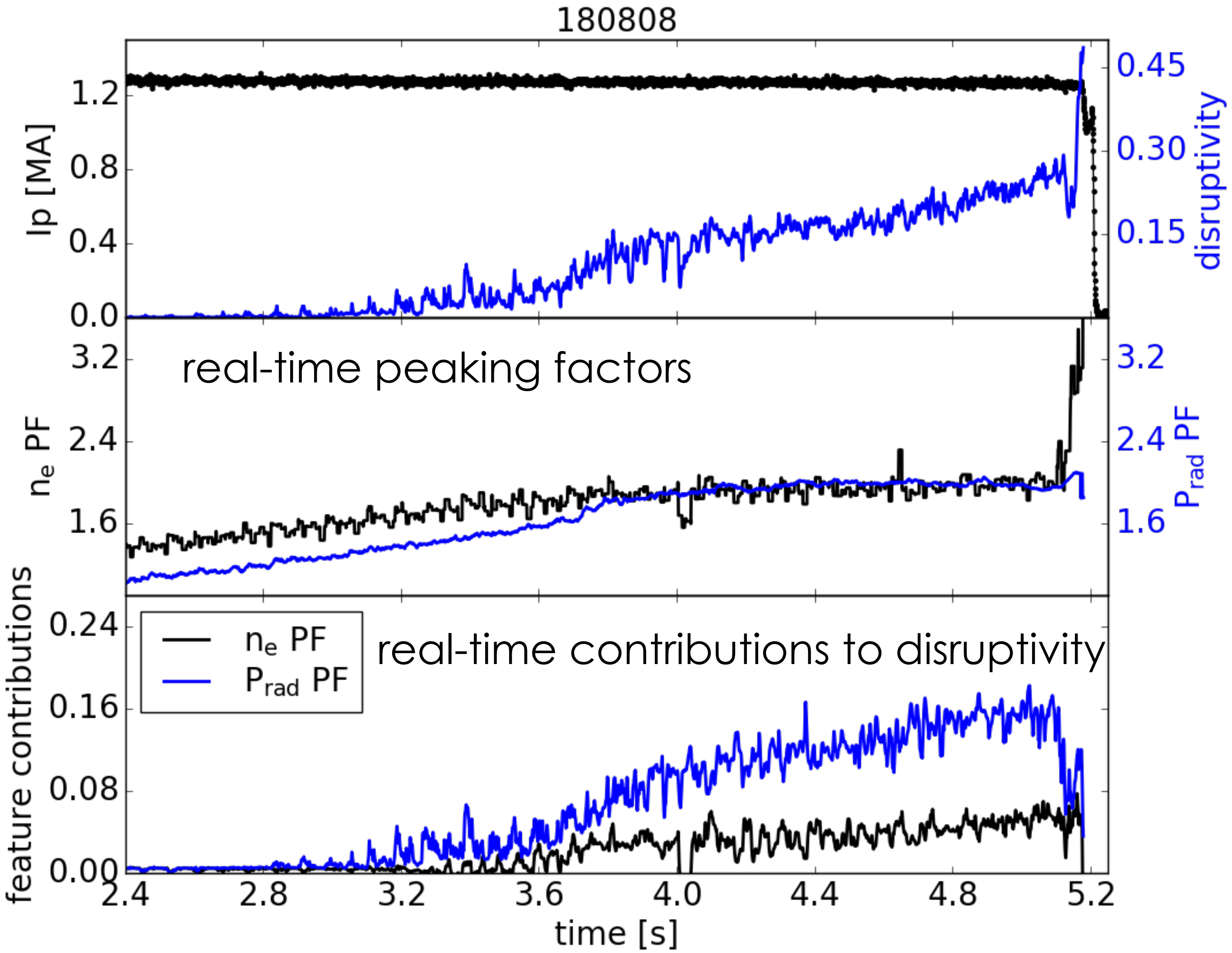
Access to disruptivity drivers in real-time: monitoring of unstable plasma features



DPRF 2.0 shows real-time feature contribution computation (~ 200 μs) and successful ONFR* integration

C. Rea et al. IAEA FEC 2020



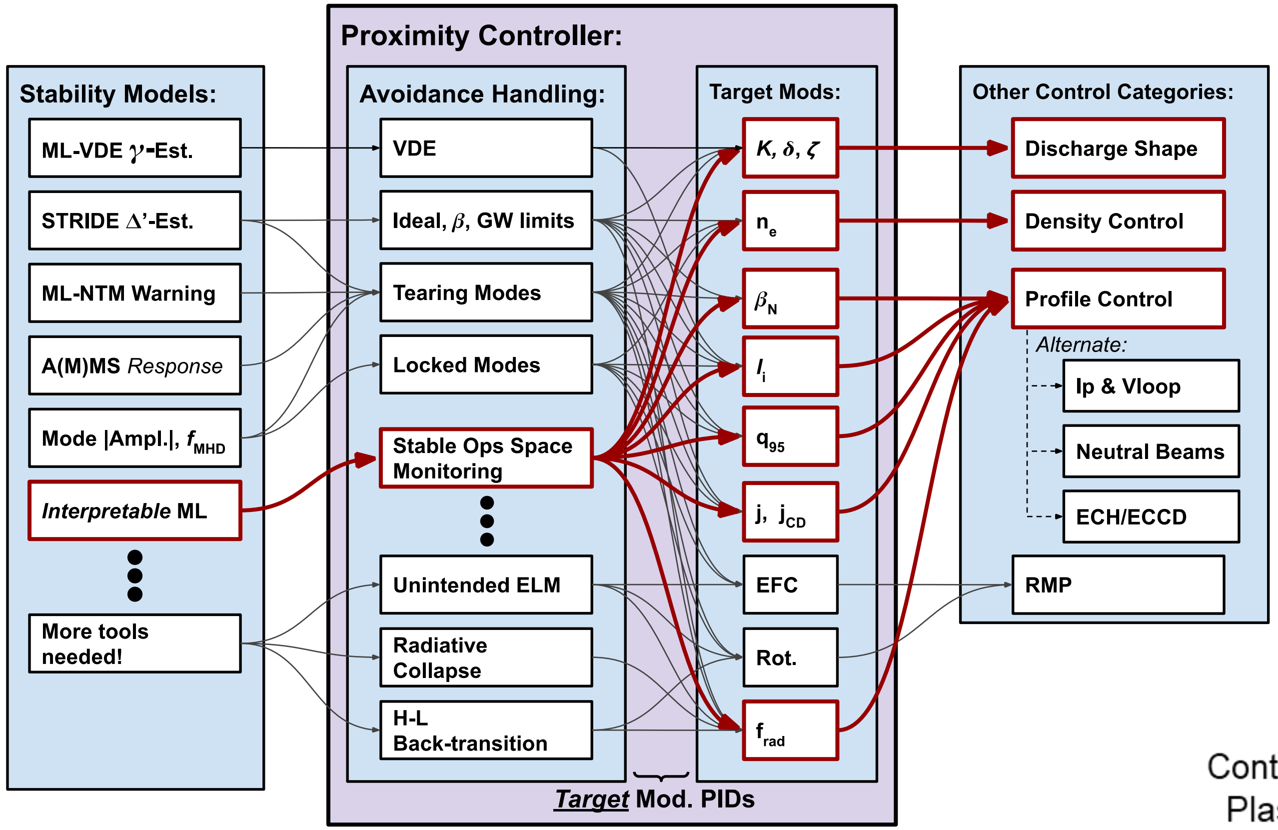


Closed the loop in the PCS by triggering early rapid shutdown, MGI, and ECH

Assessed peaking factors as relevant metrics in scenario ~ ITER baseline

*Off-Normal Fault Response → Asynchronous and Emergency response.
N. Eidietis et al., 2018 *Nucl. Fusion* 58 056023

In progress: include DPRF 2.0 in DIII-D proximity control architecture to regulate stability and avoidance



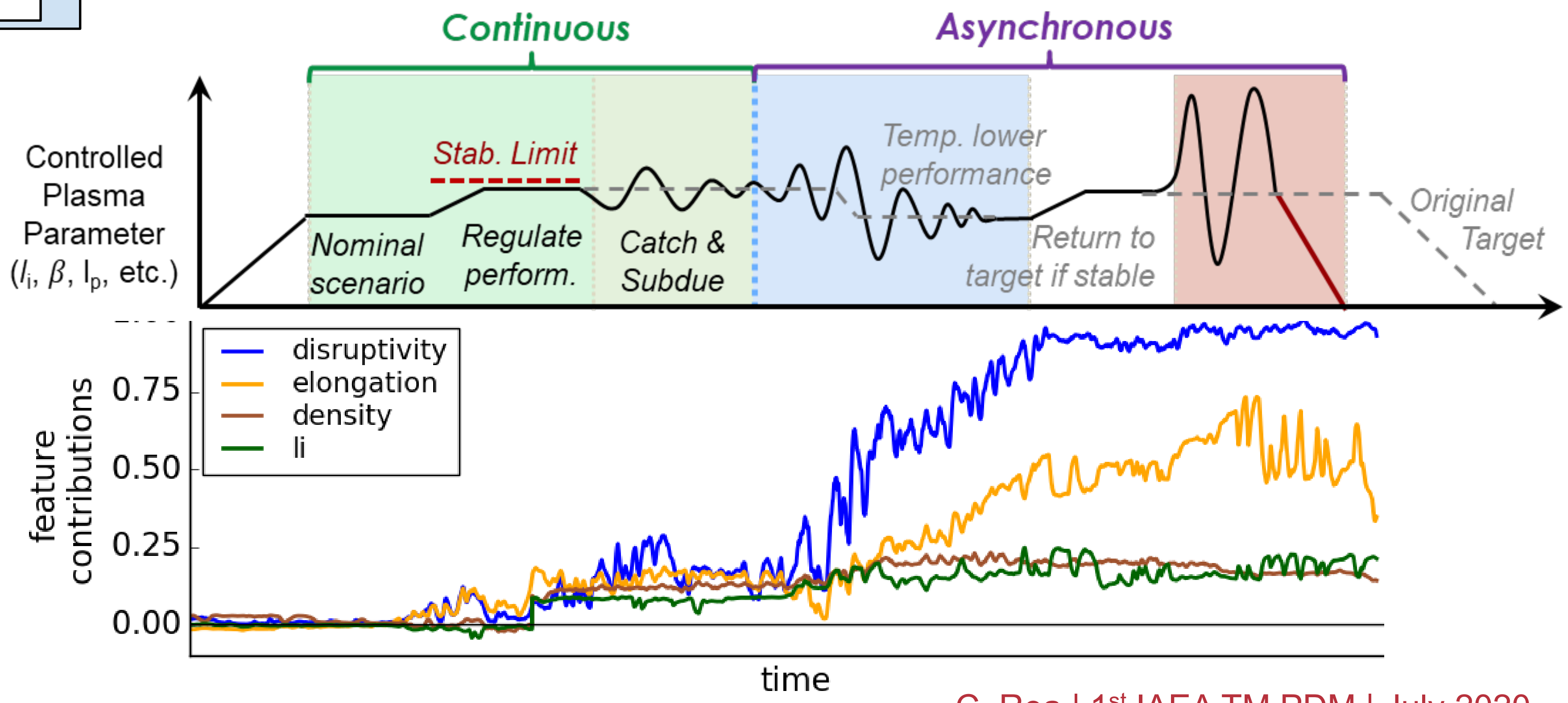
$$\Delta\kappa = f_{danger} * f_{\kappa,contrib} * sign\left(\frac{d\kappa}{dt}\right)\frac{\Delta\kappa_{target}}{\Delta f_{\kappa,contrib}}$$

J. Barr, "Control Solutions Supporting Disruption Free Operation on DIII-D and EAST", *this meeting*

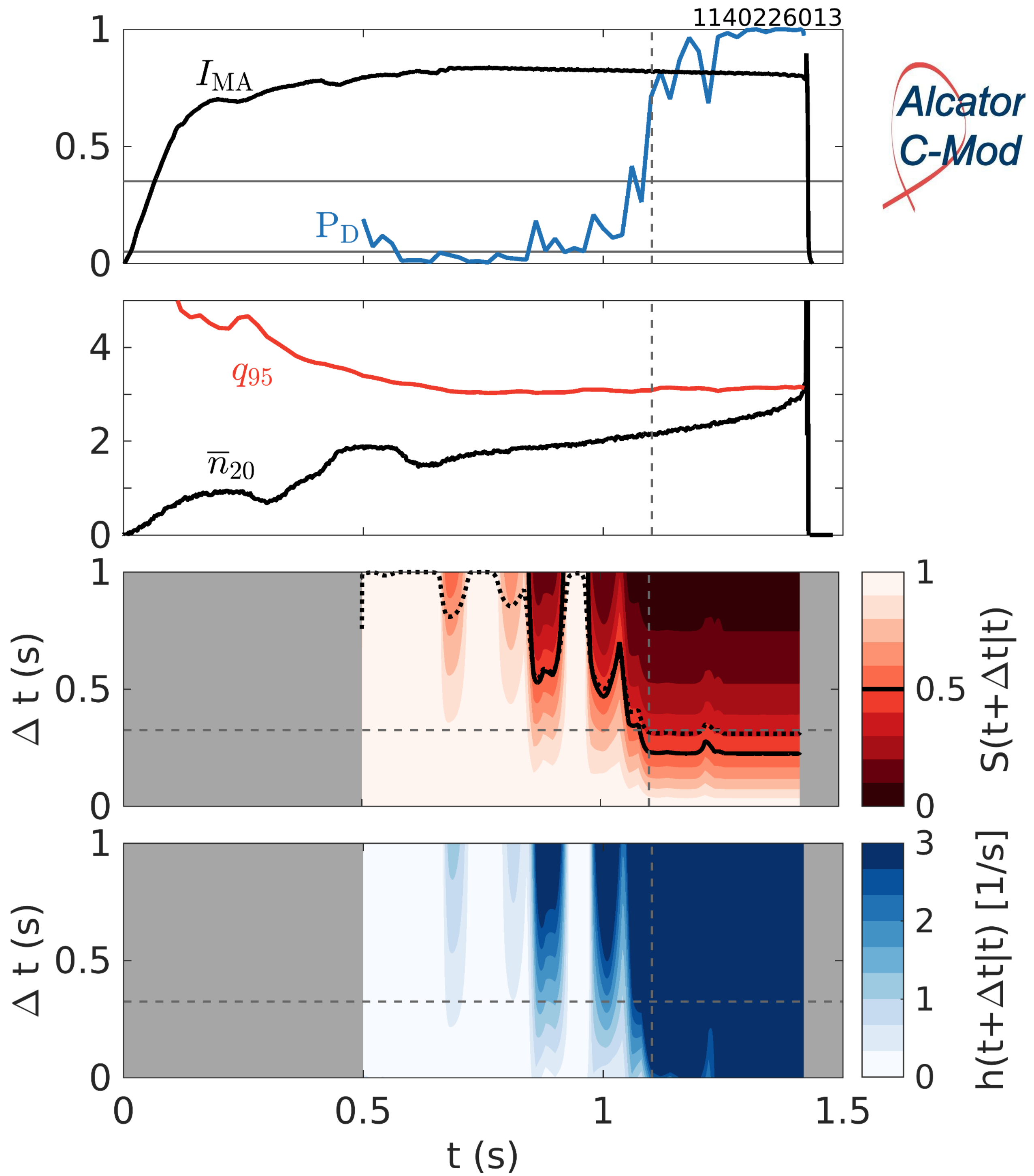
Disruptivity as general proximity of current plasma state to unstable ops space

DIII-D NATIONAL FUSION FACILITY

Feature contributions can be mapped onto controllable plasma parameters to regulate stability



DPRF disruptivity analogous to current probability of membership to disruptive class



Alcator C-Mod data used as proof of concept to combine DPRF with survival analysis.

The disruptivity P_D can be used to:

- Predict the **future probability** of **plasma survival** $S(t + \Delta t \mid t)$ [1] or
- Model the **instantaneous hazard** [2,3] $h = \mathrm{d} \ln S / \mathrm{d}t$ to be used as **probability generator**.

[1] RA Tinguely et al 2019 PPCF 61
[2] KEJ Olofsson et al 2018 PPCF 60
[3] KEJ Olofsson et al 2018 FED 146

Hazard function modeling connects dynamical systems and risk-aware control design by probability generation

Survival function for future event

Dynamical system: a(x) drift, b(x) diffusion

$$\Pr\left[T > t | X_0 = x\right] = \mathbb{E}\left\{\exp\left(-\int_0^t d\tau h(X_\tau)\right)\right\} \text{ s.t. } dX = a(X)dt + b(X)dW$$

ML-enabled direct hazard function h(x)

- Dynamical system *(a, b)* either by ML or first principles or a combination; plasma state x.
- Dependence on future actuation makes **future event probability conditional**: control design.
- Hazard function directly corresponds to (probabilistically calibrated) operational boundaries.
- **Underutilized approach**: only tearing mode events analyzed (in DIII-D) to date.

KEJ Olofsson et al 2018 PPCF 60
KEJ Olofsson et al 2018 FED 146

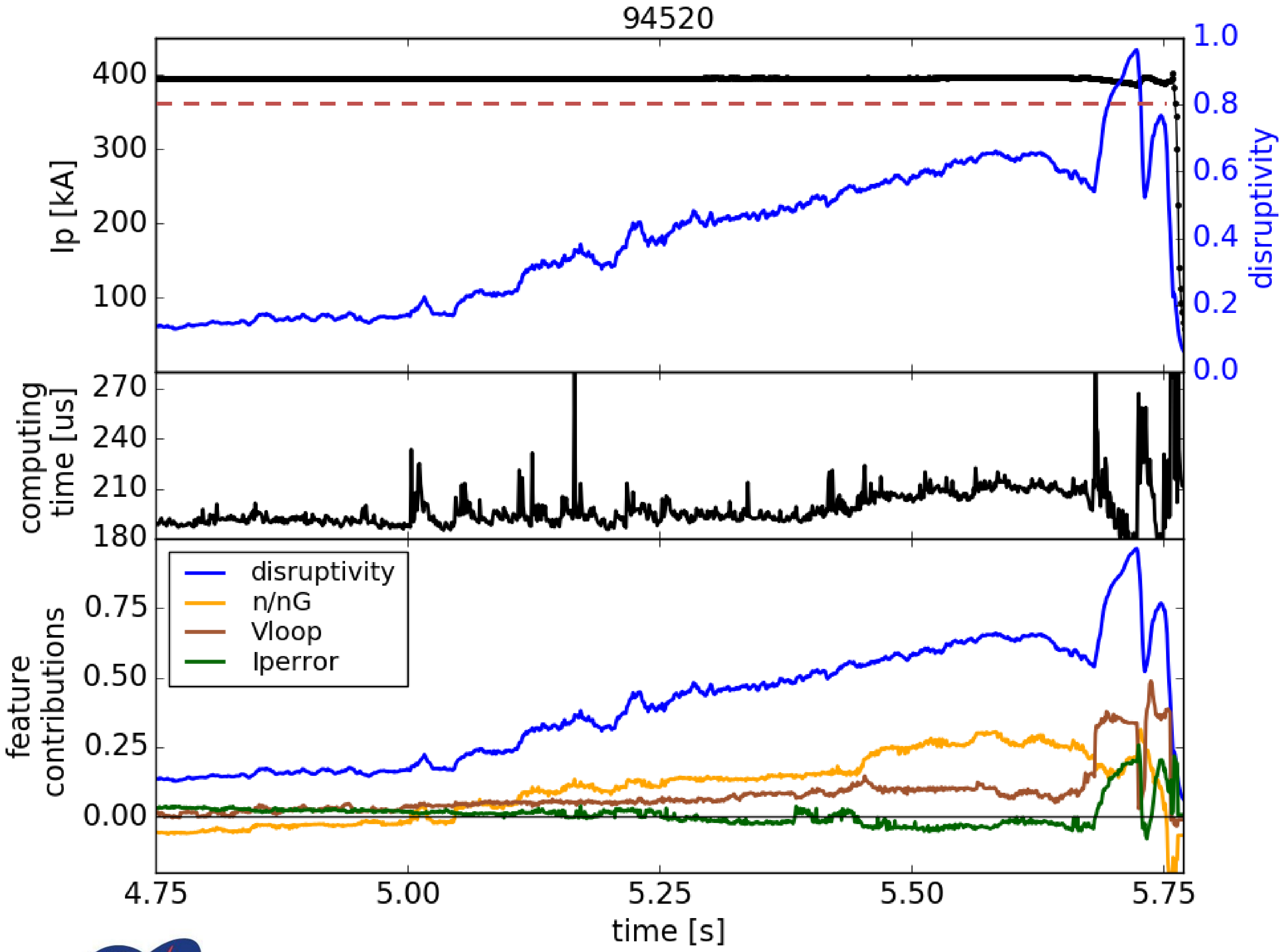
Outline

- Disruption Prediction
  - Intro and motivations
- Overview Of Interpretable Algorithms Across Devices
  - DIII-D
  - EAST
  - KSTAR
  - JT-60U

DPRF
RF
SVM

- Summary And Conclusions

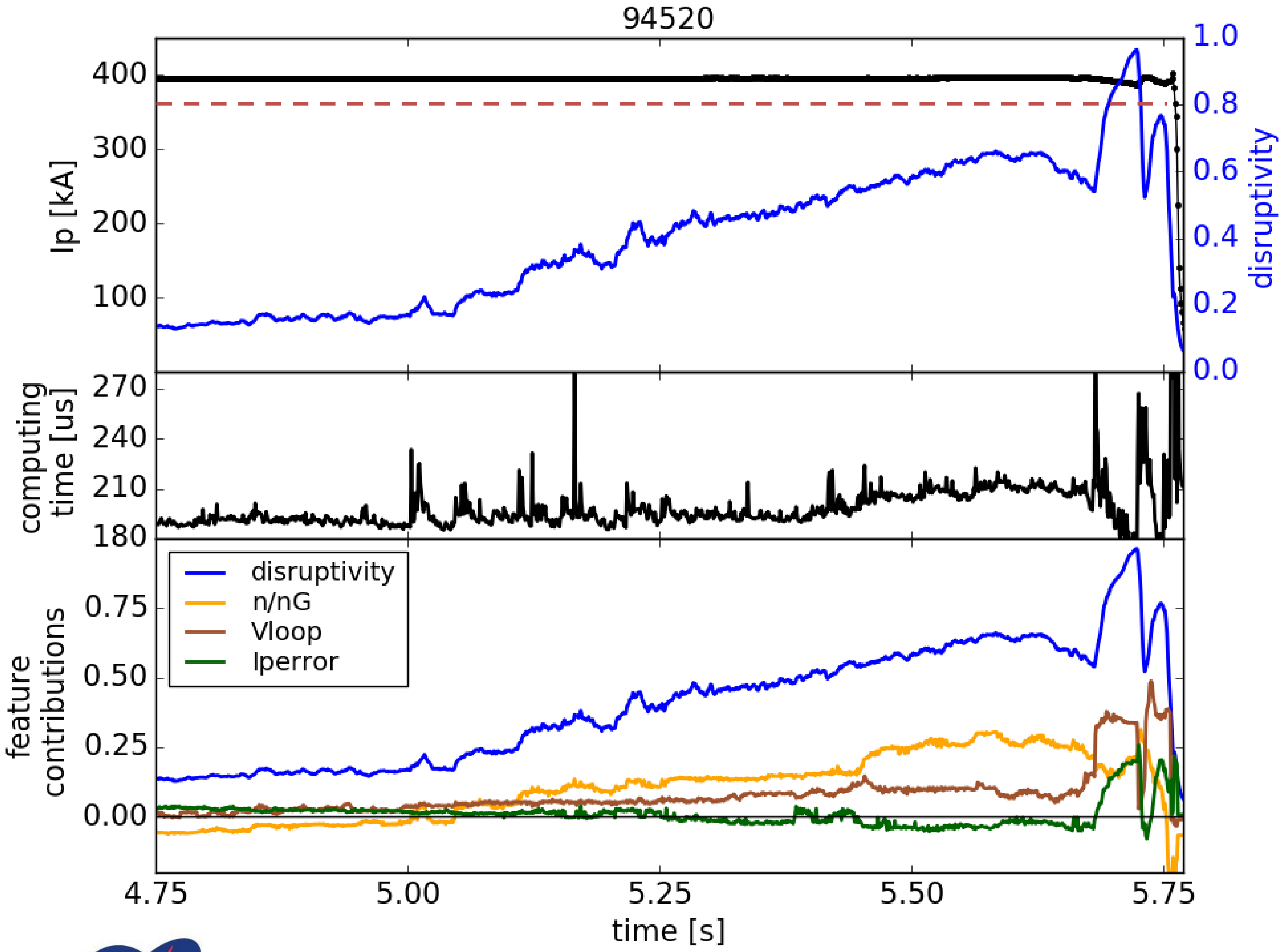
DPRF installed in EAST PCS: feature contributions and disruptivity calculated in real-time in < 200 μs



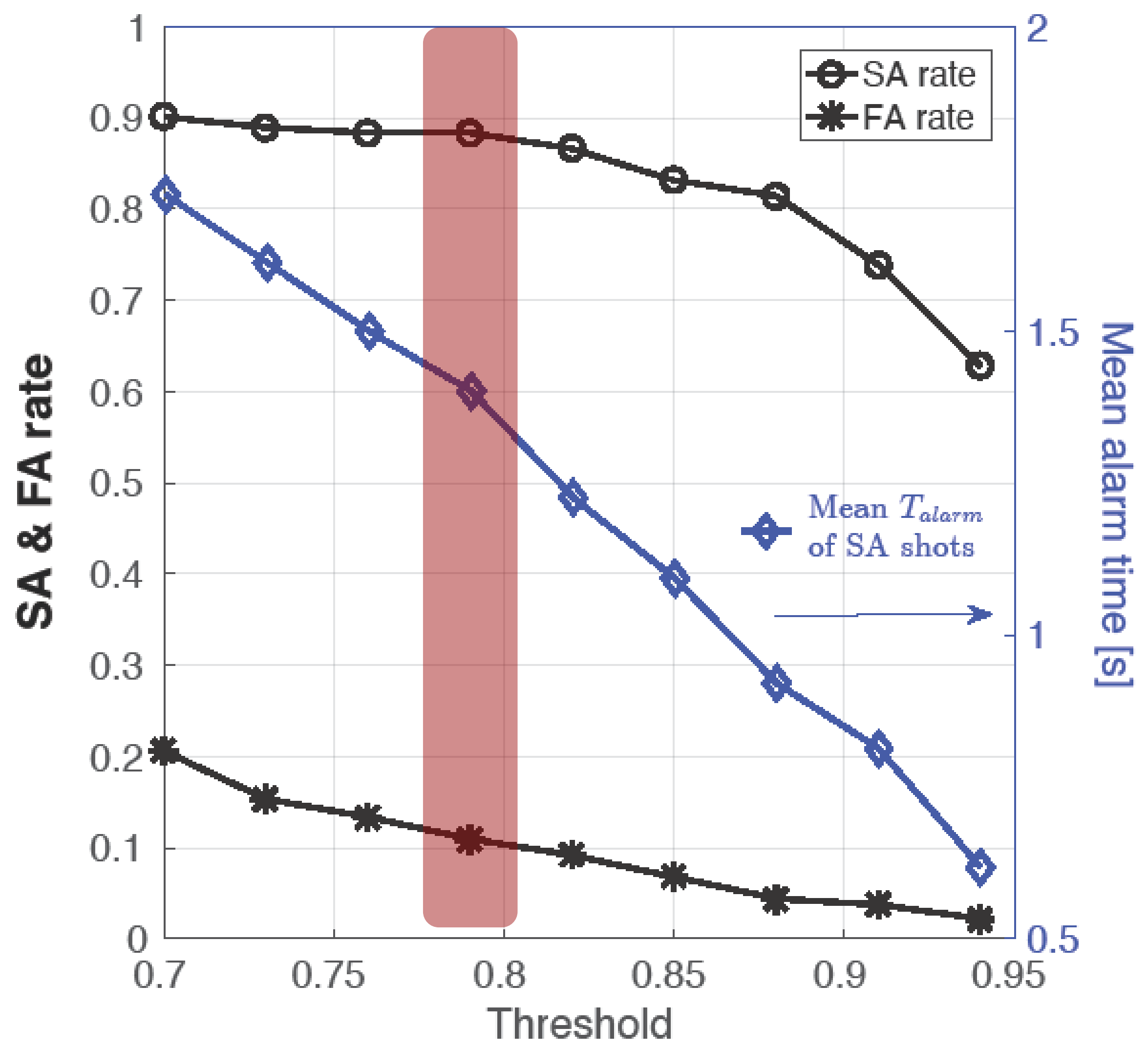
- DPRF trained using 400 high-density (ne/nG > 0.8) disruptions and 400 non-disruptive data.
- Tested in real-time on 172 disruptive and 456 non-disruptive discharges.
- Tested in **closed-loop to fire mitigation system.**

C. Rea et al. IAEA FEC 2020
W. Hu, C. Rea et al., in preparation 2020

EAST DPRF: disruptivity threshold of 0.8 guarantees SA ~89% and FA ~9% and alarm > 1 s



- SA: successful alarm, disruption detected in advance;
- FA: false alarm, alarm triggered for non-disruptive discharge.



C. Rea et al. IAEA FEC 2020
W. Hu, C. Rea et al., in preparation 2020

Development of data-driven disruption prediction system using random forest method in KSTAR

- Object
  Development of disruption prediction system based on data-driven machine-learning methodology using KSTAR database

- Database
  - Total 1054 disruption shots from 2015 to 2018 KSTAR campaign
  - Label (disruptive / non-disruptive) based on 40 ms prior to thermal quench (40 ms: required time to activate disruption mitigation system, such as MGI or SPI)
  - Dataset: $I_{p,error}$, f_{GW}, δB_{LM}, Z_0, q_{95}, V_{loop}, and l_i

- Training result
  - Random forest, binary classification
  - Confusion matrix:

True label	Predicted label: Non-disruptive	Predicted label: Disruptive
Non-disruptive	14766 (0.463)	1533 (0.048)
Disruptive	2033 (0.064)	13531 (0.425)

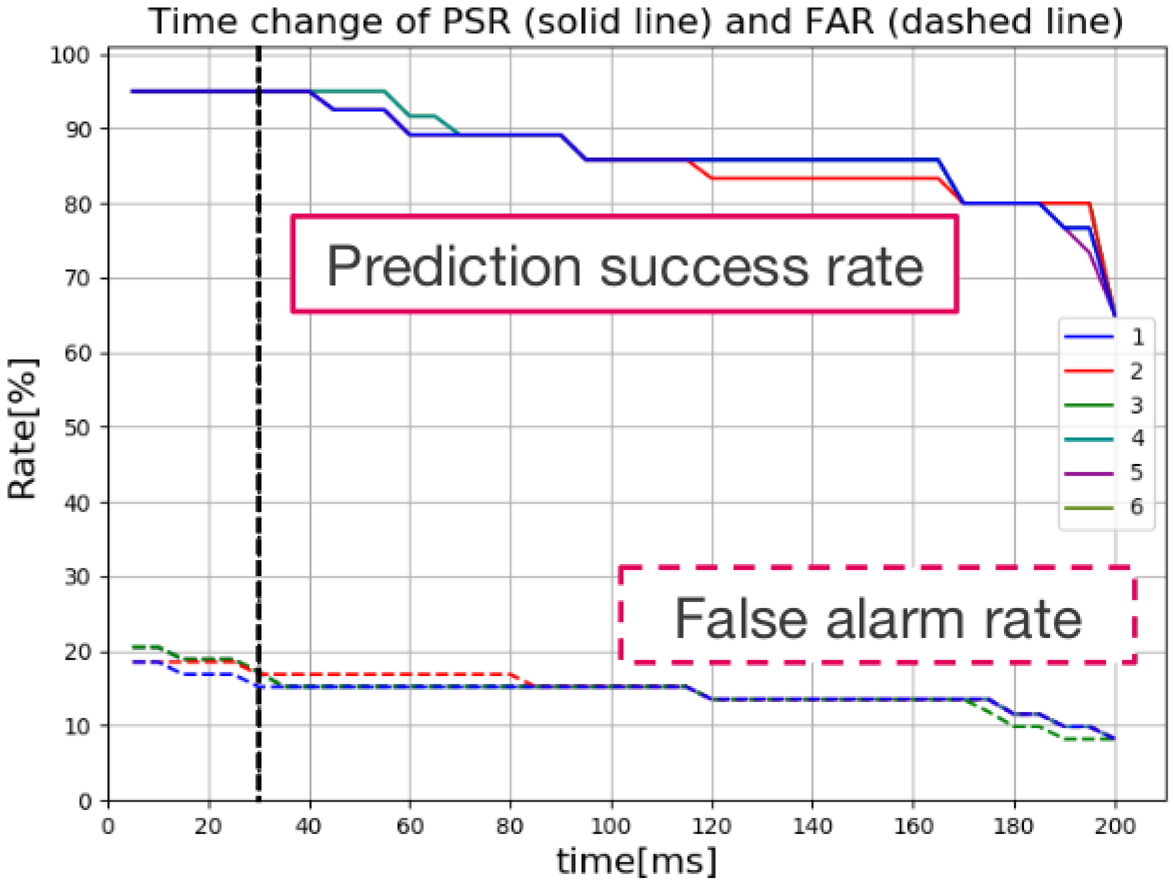
Accuracy on non-disruptive class: 90.6%
Accuracy on disruptive class: 86.7%

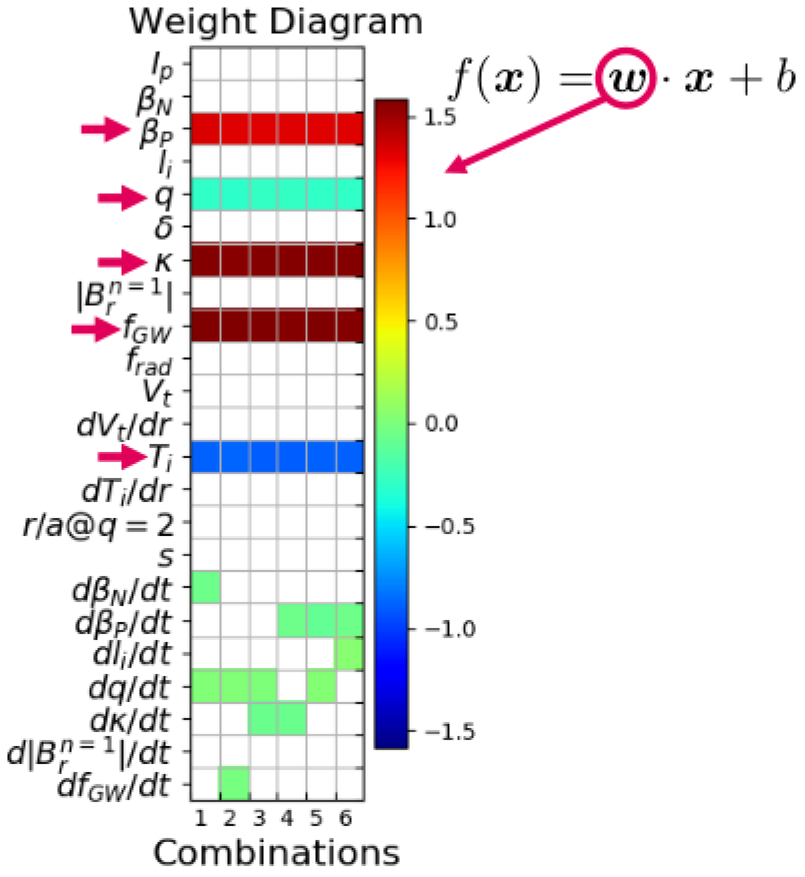
J. Lee et al., *private communications*

NFRI

High-beta disruption prediction in JT-60U through exhaustive search and SVM

- Feature extraction via Sparse Modeling → **K-sparse Exhaustive Search**



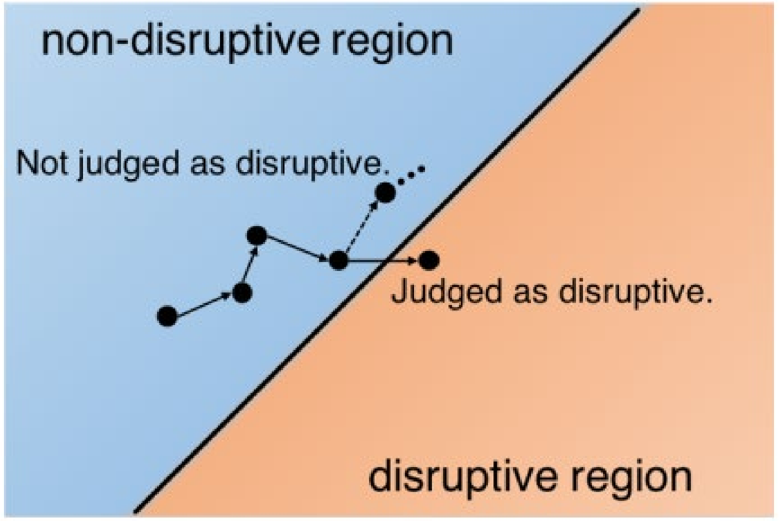


Top combination:

$$\boxed{\beta_P, q_{95}, \kappa, f_{GW}, T_i}$$

Results in **PSR ~ 95%**, **FAR ~ 15%** at **30 ms** before the disruption.

- **Binary classification** through linear Support Vector Machine (**SVM**) to extract decision function for the **boundary**: $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$



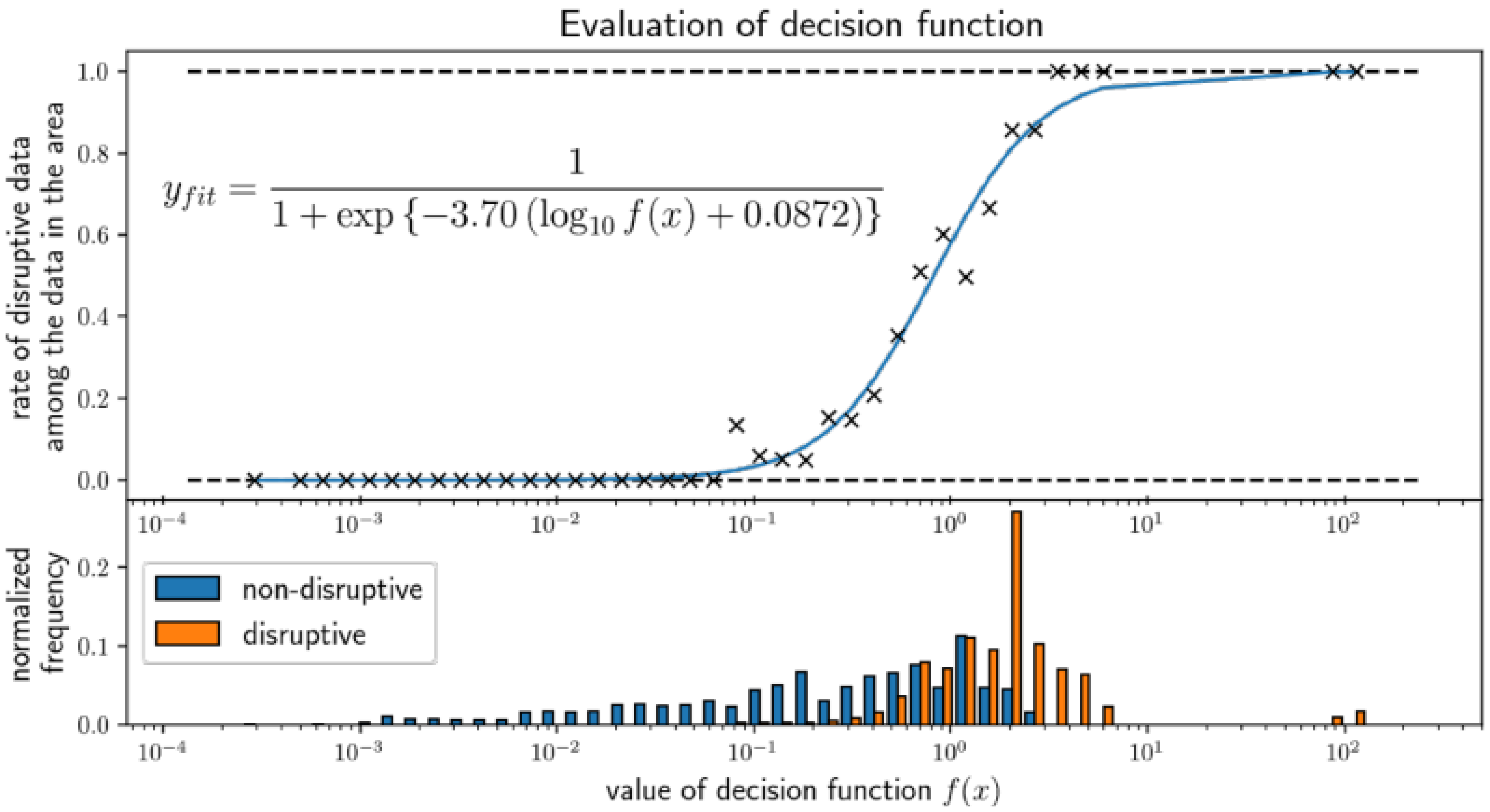
T. Yokoyama et al., *Fus. Eng. Design* 140 (2019) 67–80

High-beta disruption prediction in JT-60U through exhaustive search and SVM

- Decision function obtained by retraining the SVM, after taking the log of the training data:

$$f_{\exp}(\boldsymbol{x}) = e^{7.45}\beta_{\mathrm{P}}^{5.39}q_{95}^{-8.29}\kappa^{7.40}f_{\mathrm{GW}}^{4.50}T_{\mathrm{i}}^{-0.120}$$

Decision function parametrized from top combination of features enables disruption likelihood estimate



T. Yokoyama et al., Data-driven study of high-beta disruption prediction in JT-60U using exhaustive search, AAPPS 2019

T. Yokoyama et al., *Fus. Eng. Design* 140 (2019) 67–80

Outline

- **Disruption Prediction**
  - Intro and motivations
- **Overview Of Interpretable Algorithms Across Devices**
  - DIII-D
  - EAST
  - KSTAR
  - JT-60U
- **Summary And Conclusions**

More than 20 years of research in disruption prediction have produced voluminous literature

Device	References (incomplete list)
ADITYA	Sengupta and Ranjan 2000 NF 40 Sengupta and Ranjan 2001 NF 41
Alcator C-Mod	Rea et al 2018 PPCF 60 Montes et al 2019 NF 59 Tinguely et al 2019 PPCF 61
ASDEX-U	Pautasso et al 2002 NF 42 Windsor et al 2005 NF 45 Aledda et al 2015 FED 96-97
DIII-D	Wroblewski et al 1997 NF 37 Rea and Granetz 2018 FST 74 Rea et al 2018 PPCF 60 Montes et al 2019 NF 59 Rea et al 2019 NF 59 Kates-Harbeck et al 2019 Nature 568
EAST	Montes et al 2019 NF 59
JET	Windsor et al 2005 NF 45 Cannas et al 2004 NF 44 Cannas et al 2007 FED 82 Murari et al 2008 NF 48 Murari et al 2009 NF 49 Ratta' et al 2010 NF 50 De Vries et al 2011 NF 51 Vega et al 2013 FED 88 Cannas et al 2014 PPCF 56 Ratta' et al 2014 PPCF 56 Murari et al 2018 NF 58 Pau et al 2018 IEEE TPS 46 Kates-Harbeck et al 2019 Nature 568 Pau et al 2019 NF 59
JT-60U	Yoshino 2003 NF 43 Yoshino 2005 NF 45 Yokoyama et al. 2019 FED 140
J-TEXT	Wang et al 2016 PPCF 58 Zheng et al 2018 NF 58
NSTX	Gerhardt et al 2013 PPCF 60

Data-driven predictors to be adopted as last line of defense for disruption mitigation but…

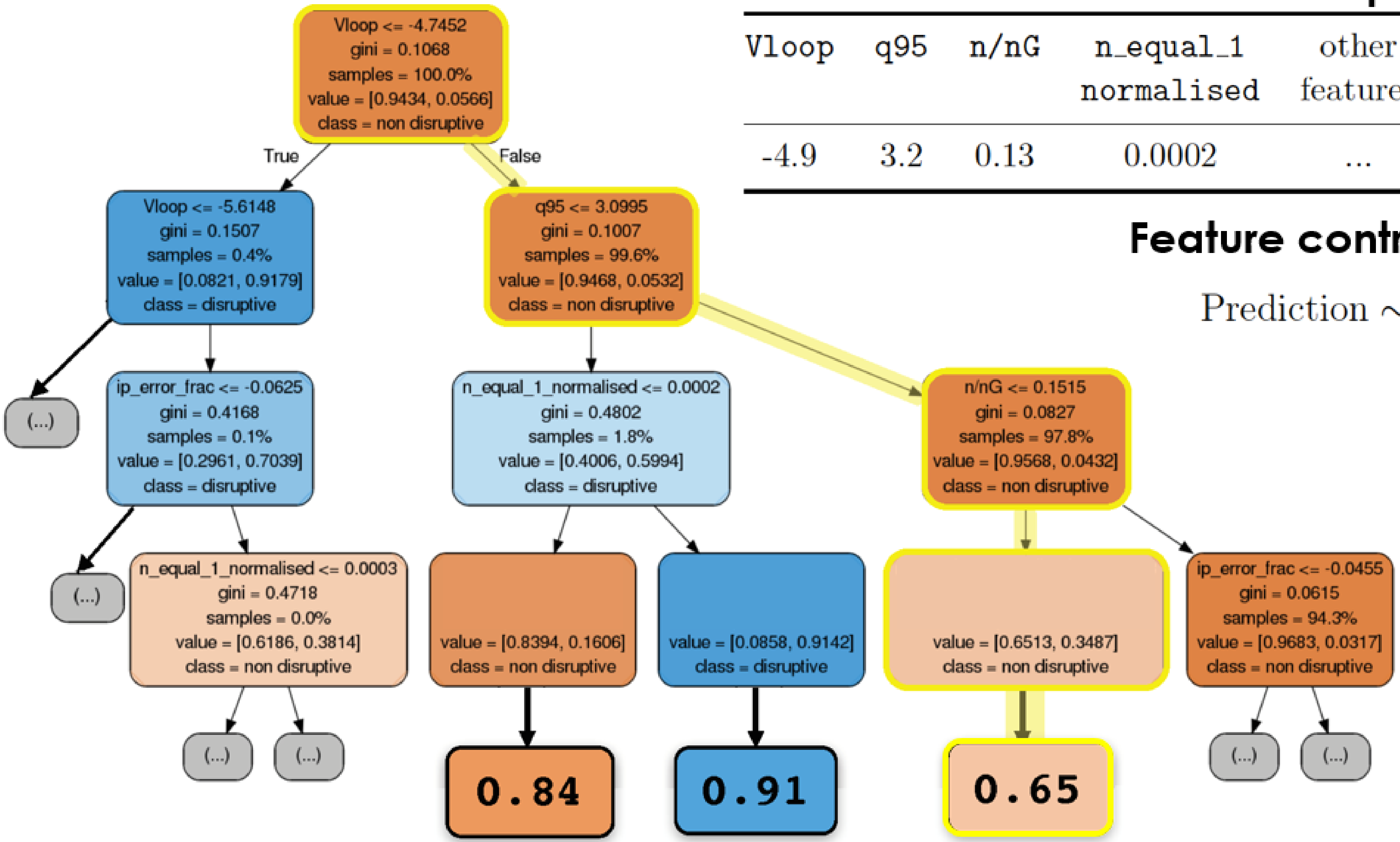
- **Interpretable output** combined with **control** algorithms can inform the PCS on **disruption precursors** and be employed in **avoidance** schemes.
  - Frameworks exist to extract plasma **future survival** → *Tinguely et al.*
    or **instantaneous hazard** (as probability generator) for instabilities → *Olofsson et al.*
- **DPRF** provides **explainable predictions** – tested on **C-Mod, EAST, DIII-D**:
  - Works as **real-time scenario detector** (DIII-D, EAST).
  - To be integrated with **proximity controller** for continuous avoidance (DIII-D).
- Analogous efforts ongoing at international facilities:
  - *J. Lee and J. Kim* @ KSTAR
  - *T. Yokoyama* @ JT-60U;
  - *A. Pau and others* @ JET, TCV, AUG;
  - *G. Dong et al.* @ DIII-D.
- **Ongoing work to design predictor for ITER:**
  - Few ITER disruptions might still be needed to design effective data-driven solutions.

→ *J.X. Zhu et al.*
→ *J. Kates-Harbeck et al.*

Funding acknowledgement, disclaimer, and additional slides

Acknowledgement:
This work was supported, in part, by the U.S. Department of Energy under DE-FC02-04ER54698, DE-FC02-99ER54512, DE-SC0014264, DE-SC0010720, DE-SC0010492.

Decision paths in DPRF trees provide local measures of explainability through information gain and loss



https://github.com/andosa/treeinterpreter, A. Saabas,
A. Palczewska et al., Integration of Reusable Systems (2014).

Feature vector example

Vloop	q95	n/nG	n_equal_1 normalised	other features	prediction
-4.9	3.2	0.13	0.0002	...	~ 0.65

Feature contribution breakdown

$$\begin{aligned} \text{Prediction} \sim 0.651 \sim\ & 0.943 \quad \text{(bias term)} \\ & + 0.003 \quad (\text{gain, } \texttt{Vloop}) \\ & + 0.010 \quad (\text{gain, } \texttt{q95}) \\ & - 0.305 \quad (\text{loss, } \texttt{n/nG}) \end{aligned}$$

contributions, or information gain (+) and loss (-)

Predictions for forest of M trees can be **decomposed** in the K **contributions** from each evaluated input feature:

$$F(x) = \frac{1}{M}\sum_{m=1}^{M} \text{bias}_m + \sum_{k=1}^{K}\left(\frac{1}{M}\sum_{m=1}^{M} \text{contrib}_m(x,k)\right)$$

DPRF 0-D scalar input features – DIII-D and EAST

DIII-D
$B_r^{n=1}/B_\phi$ Locked Mode proxy
q_{95}
n/n_G
$(I_p - I_{prog})/I_{prog}$
ℓ_i
β_p
V_{loop}
W_{mhd}
$r_{HWHM}(T_e)/a$
P_{rad}/P_{inp}

EAST
n/n_G
V_{loop}
$(I_p - I_{prog})/I_{prog}$
ℓ_i
q_{95}
W_{mhd}
$(z_{cur} - z_{prog})/a$
β_n
κ