Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Anomaly Detection with Machine Learning on Time Series Data from the Fermi Anti-Coincidence Detector

Andrea Adelfio, Sara Cutini, Stefano Germani (INFN Perugia) , Simone Maldera (INFN Torino), Francesco Longo (INFN Trieste) and Riccardo Crupi (Univeristy of Udine)

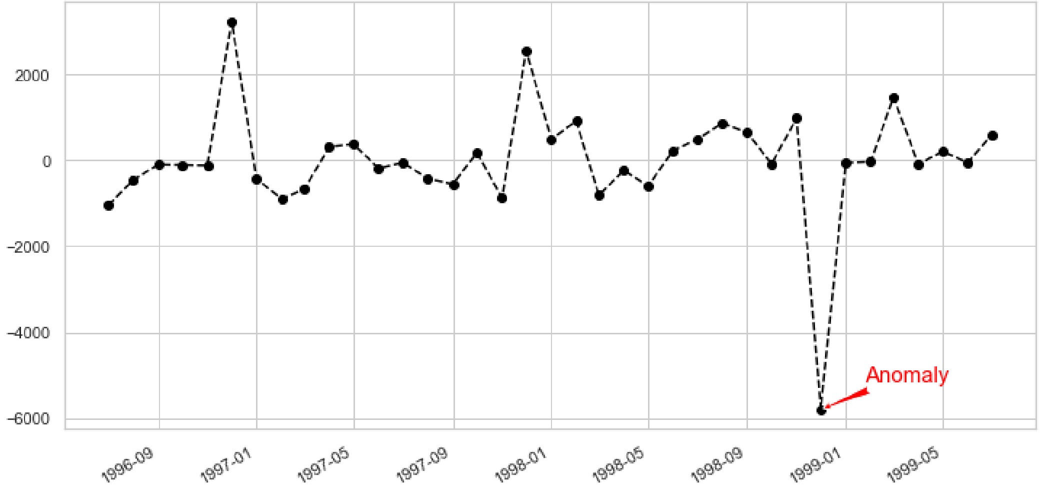
Spoke 3 General Meeting, Elba 5-9 / 05, 2024

Anomaly Detection with Machine Learning on Time [illegible]erie[illegible] Data from the Fermi Anti-Coincidence Detector

Andrea Adelfio, Sara Cutini, Stefano Germani (INFN Perugia) , Simone Maldera (INF[illegible]no),
Francesco Longo (INFN Trieste) and Riccardo Crupi (Univeristy of Udine[illegible]

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024

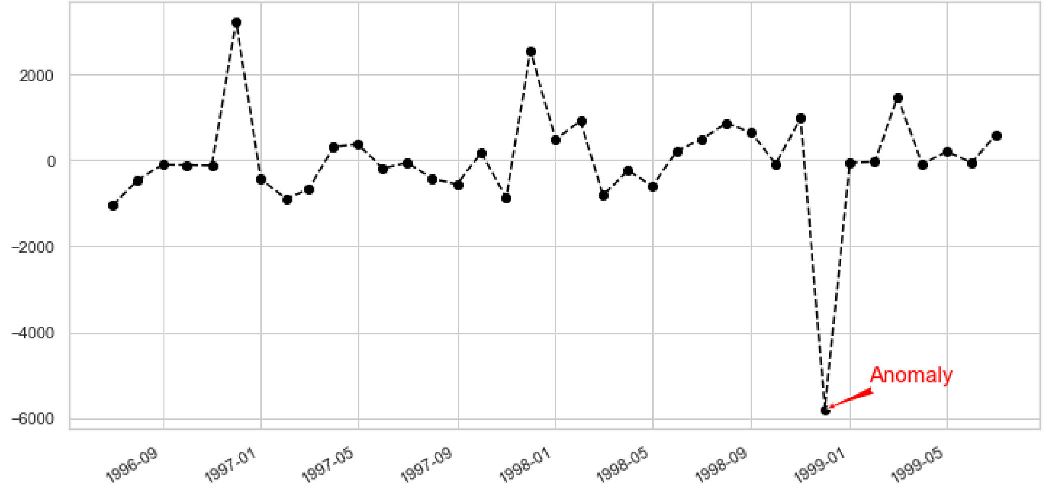
Scientific Rationale



-To develop an Anomaly Detection algorithm for Time Series data using Machine Learning techniques;

-Apply it to create a pipeline for the astrophysical data from the Fermi Anti-Coincidence Detector (ACD);

Scientific Rationale



- To develop an Anomaly Detection algorithm for Time Series data using Machine Learning techniques;

- Apply it to create a pipeline for the astrophysical data from the Fermi Anti-Coincidence Detector (ACD);

Technical Objectives, Methodologies and Solutions

The functionality of this framework can be summarized in two points:

1. **Get a baseline prediction $\hat{Y}$ of a signal Y, given a set of context variables X and a corresponding value of $\hat{Y} = f(X)$.**

2. **To use the prediction for the background of Y to find significant deviations in the signal with an efficient algorithm.**

Technical Objectives, Methodologies and Solutions

The functionality of this framework can be summarized in two points:

1. **Get a baseline prediction $\hat{Y}$ of a signal Y, given a set of context variables X and a corresponding value of $\hat{Y} = f(X)$, with a function that can be modeled with some Machine Learning technique.**

2. **To use the prediction for the background of Y to find significant deviations in the signal with an efficient algorithm.**

Technical Objectives, Methodologies and Solutions

The functionality of this framework can be summarized in two points:

1. **Get a baseline prediction $\widehat{Y}$ of a signal Y, given a set of context variables X and a corresponding value of $\widehat{Y} = f(X)$, with a function that can be modeled with some Machine Learning technique.**

2. **To use the prediction for the background of Y to find significant deviations in the signal with an efficient algorithm, the Functional Online CuSUM (FOCuS) algorithm.**

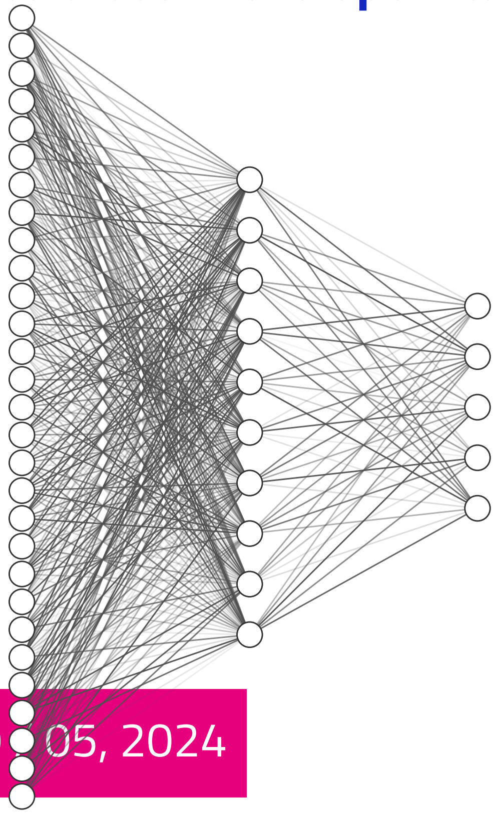
Feed Forward Neural Network

We used a Feed Forward Neural Network to find the best model that fits the background signal.

We have initiated preliminary analysis to discern the optimal structure to train the NN model.

The base structure consists of M dense hidden layers with N nodes.

The use of a Batch Normalization Layer and a Dropout Layer has been considered.



The used Loss function is the mean absolute error:

$$\mathrm{MAE}(z, y) = \frac{1}{n}\sum_{i=1}^{n} | y_i - z_i |$$

A Neural Network family

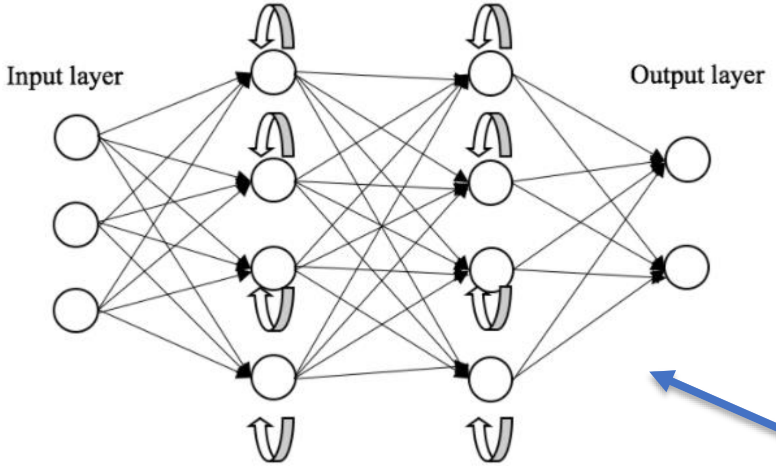
ural Network to find the best model that f

y analysis to discern the optimal structur

of
N nodes.

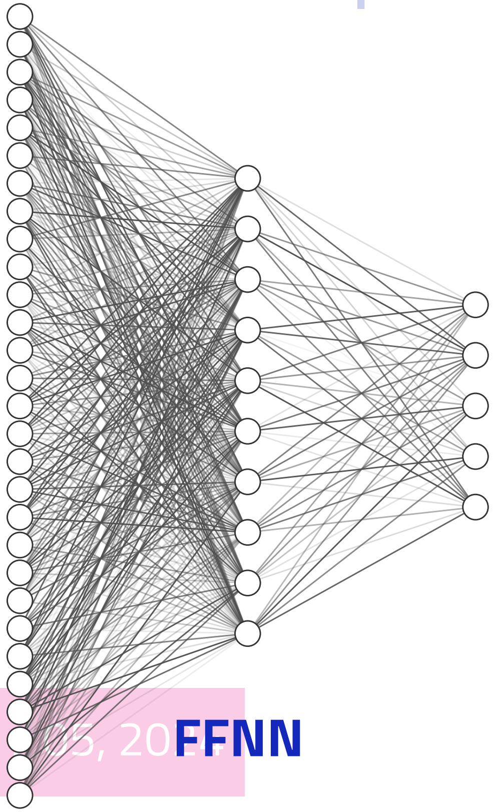
The
mean absolute error:

$$MAE(z, y) = \frac{1}{n}\sum^{n} |y_i - z_i|$$

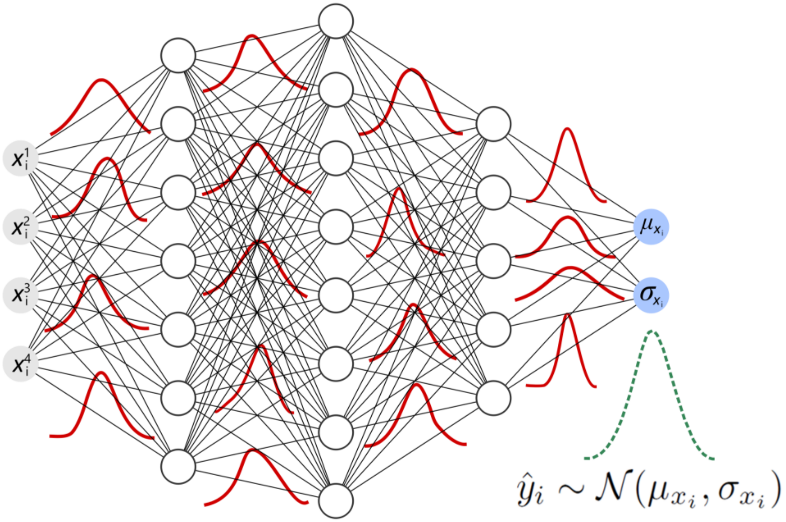


RNN

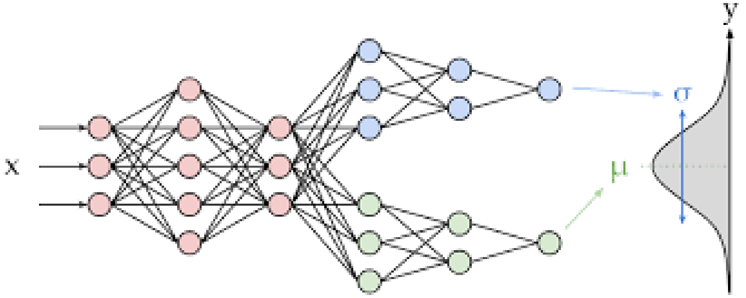
The use of a Batch Normalization Layer and a Dropout Layer has been considered.



FFNN



PBNN

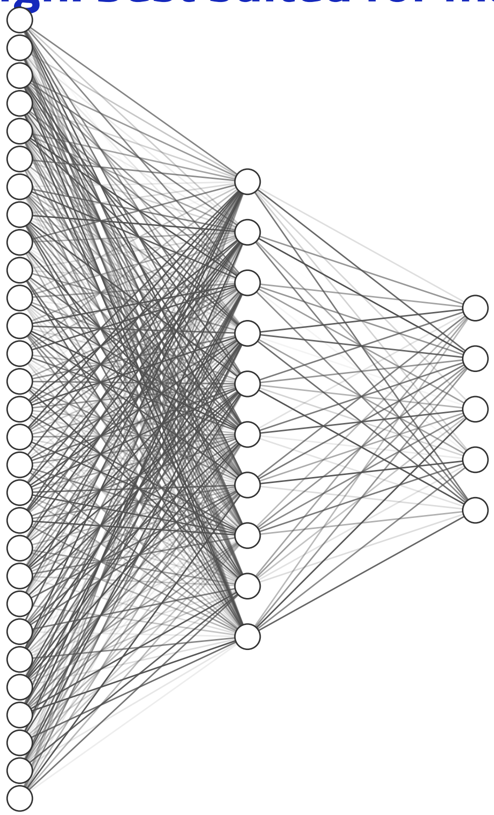


BNN

A Neural Network family: Feed Forward NN

Feed Forward Neural Network to find the best model that fits the background signal.

Key design: best suited for mapping input features to output values, given a complete set of inputs.



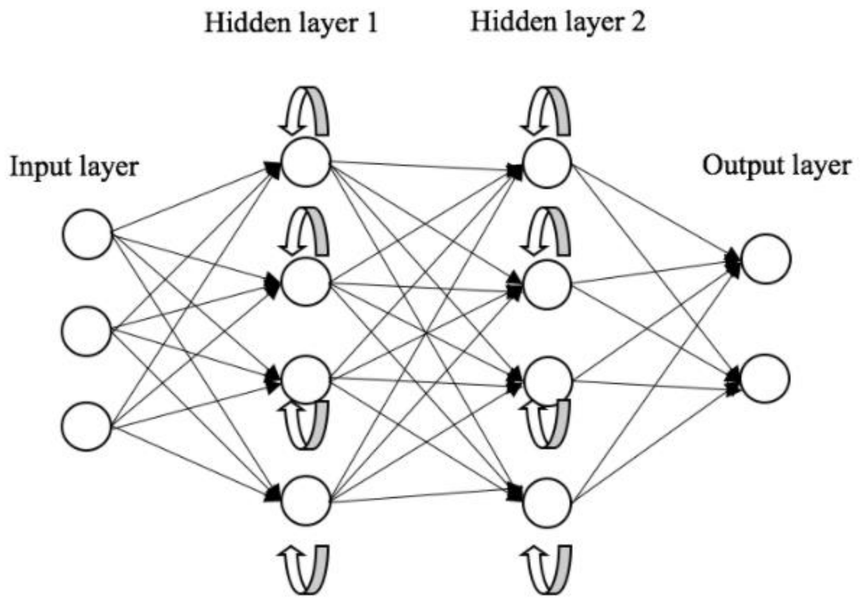
Loss function is the mean absolute error:

$$\mathrm{MAE}(z, y) = \frac{1}{n}\sum_{i=1}^{n} | y_i - z_i |$$

A Neural Network family: Recurrent NN

Recurrent Neural Network: typically used on time series data for its ability to identify temporal patterns in the data (trends, seasonality, cyclicity...).

Key design: uses previous timesteps $(i-n, \dots, i-1)$ together with current instant i to estimate $f(X_{i-n}, \dots, X_i) = Y_i$



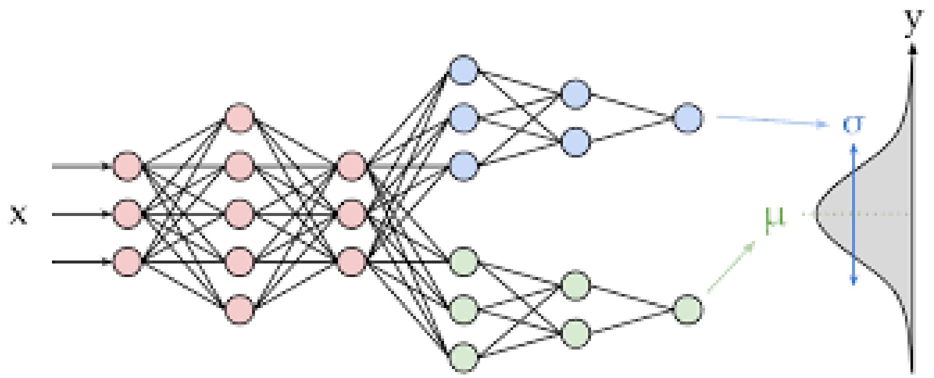
Loss function is the mean absolute error:

$$\mathrm{MAE}(z, y) = \frac{1}{n}\sum_{i=1}^{n} |\ y_i - z_i\ |$$

A Neural Network family: Bayesian NN

Bayesian Neural Network to find the best model that describes the distribution of the data.

Key design: separate outputs, a first half dedicated to estimate the values of $\hat{Y}$ and the second half dedicated to estimate the σ_Y, maximizing the likelihood of the model.



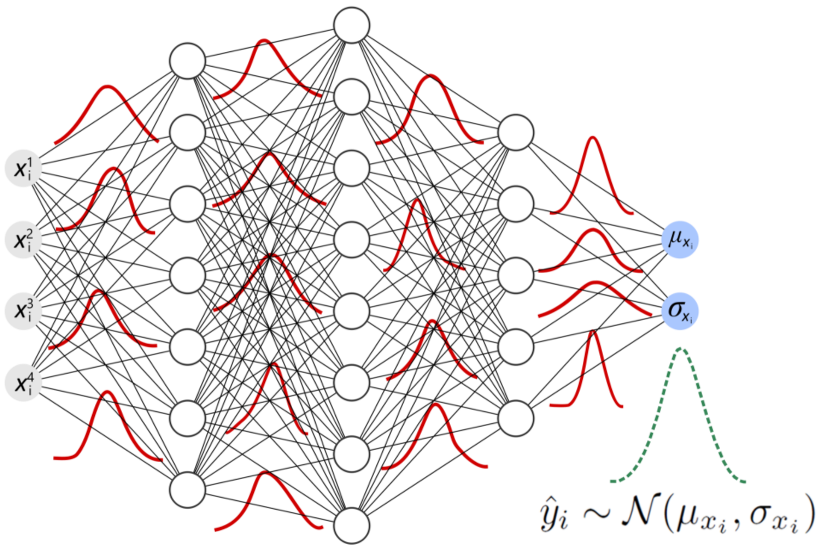
Loss function is the negative log likelihood:

$$l(\theta) = -\sum_{i=1}^{n}\left(y_i \log \hat{y}_{\theta,i} + (1-y_i)\log(1-\hat{y}_{\theta,i})\right)$$

A Neural Network family: Probabilistic Bayesian NN

Probabilistic Bayesian Neural Network to find the best model that describes the distribution of the data.

Key design: each weight and bias in the nodes have an associated probability distribution; separate outputs, a first half dedicated to estimate the values of $\widehat{Y}$ and the second half dedicated to estimate the σ_Y, maximizing the likelihood of the model.



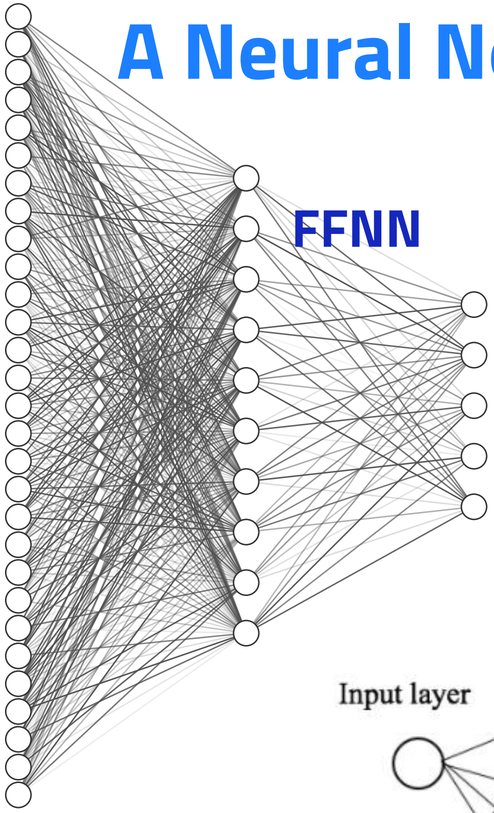
Loss function is the negative log likelihood:

$$l(\theta) = -\sum_{i=1}^{n} \left(y_i \log \hat{y}_{\theta,i} + (1 - y_i) \log (1 - \hat{y}_{\theta,i}) \right)$$

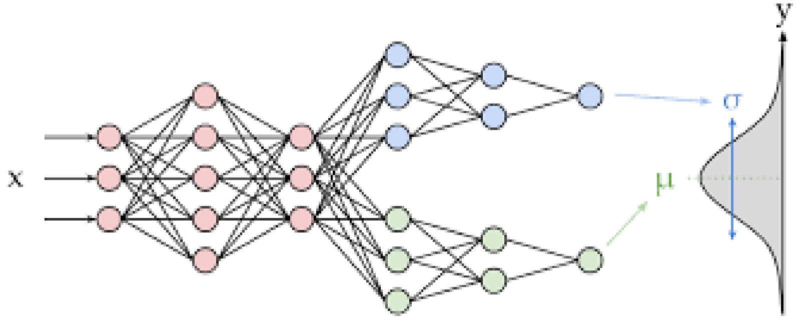
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

A Neural Network family: Choice of the Model

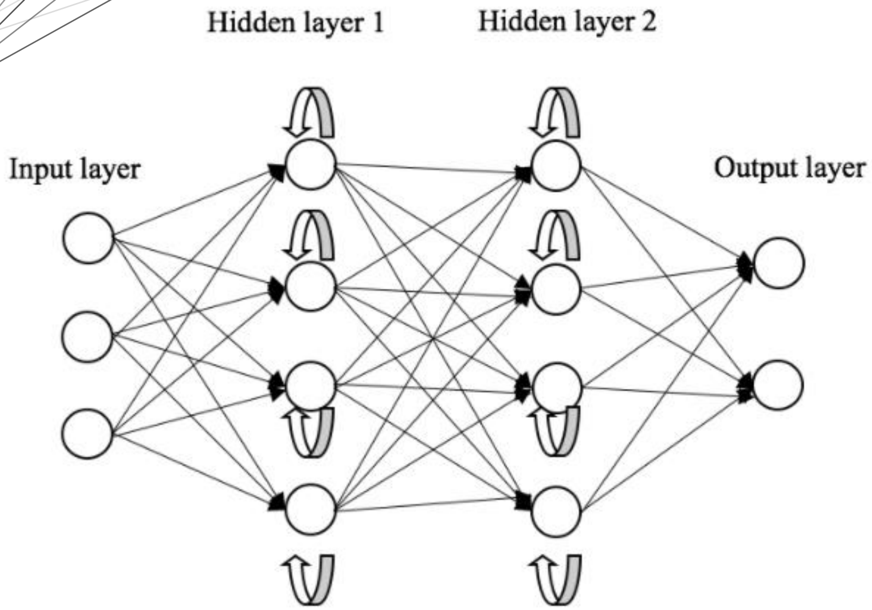
FFNN



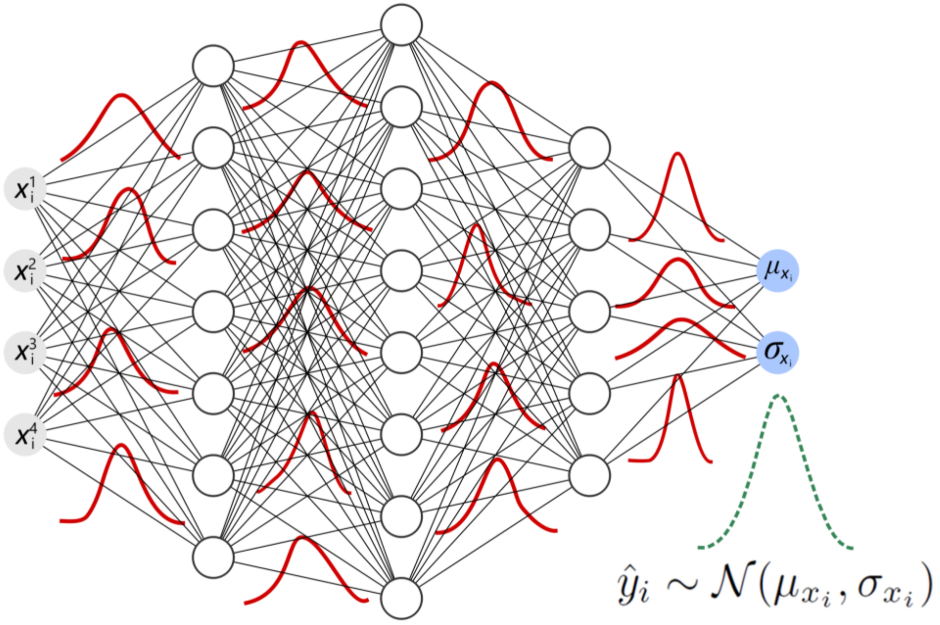
BNN



RNN



PBNN



Triggering Algorithm

A problem with typical online triggering algorithms is the need to choose a window size around a data point to estimate the background in that instance and the threshold above which a data point is labelled as an anomaly.

Triggering Algorithm: FOCuS

Identify the anomalies in the data as deviations from the background, quantifying the significance of each anomaly.

The Functional Online CuSUM (FOCuS) is a fast and efficient algorithm based on the computation of the cumulative sum of the score statistics of the data.

***Efficient*: computes the sum of score statistics and compares it to a threshold. Efficient at identifying change points in the data set.**

***Fast*: only records score statistics of data points that deviate from the distribution.**

$$S(s,n) = \sum_{i=s+1}^{N} H(x_i, \mu_0)$$

Triggering Algorithm: FOCuS

Identify the anomalies in the data as deviations from the background, quantifying the significance of each anomaly.

The Functional Online CuSUM (FOCuS) is a fast and efficient algorithm based on the computation of the cumulative sum of the score statistics of the data.

Can be used in *flavours*:
-Poisson-FOCuS: assumes a Poisson-like distribution of data; can be used for count rates data.

-Gaussian-FOCuS: assumes a Gaussian distribution; can be used for varying signals (temperature…)

-Non-parametric-FOCuS: no assumptions on the type of data.

Anomaly Detection Software

Can be used in the form of an online/offline pipeline to analyse time series data.

The software is available in a public github repository (will be added in the ICSC-Spoke3 repo).

The documentation is available, together with a set of examples, both in modular form and in the form of a pipeline.

You can also find a poster on its application on the Fermi ACD data.

SCAN ME

Timescale and KPIs

- Hired in September 2023;
- October to December, study of scientific literature;
- January to April, development of code to prepare the dataset and preliminary algorithm to fit the signal.

Code made available on github.

Timescale and KPIs

- **Study of new additional NN models typically used in Time Series data;**
- **Implementation of those models (*RNN, BNN, PBNN*).**
- **Implementation of the *Poisson-FOCuS* and *Gaussian-FOCuS* triggering algorithm.**
- **Explainability (WORK IN PROGRESS).**
- ***DataGenerator/ DASK* implementation for training of larger-than-memory Datasets (WORK IN PROGRESS).**

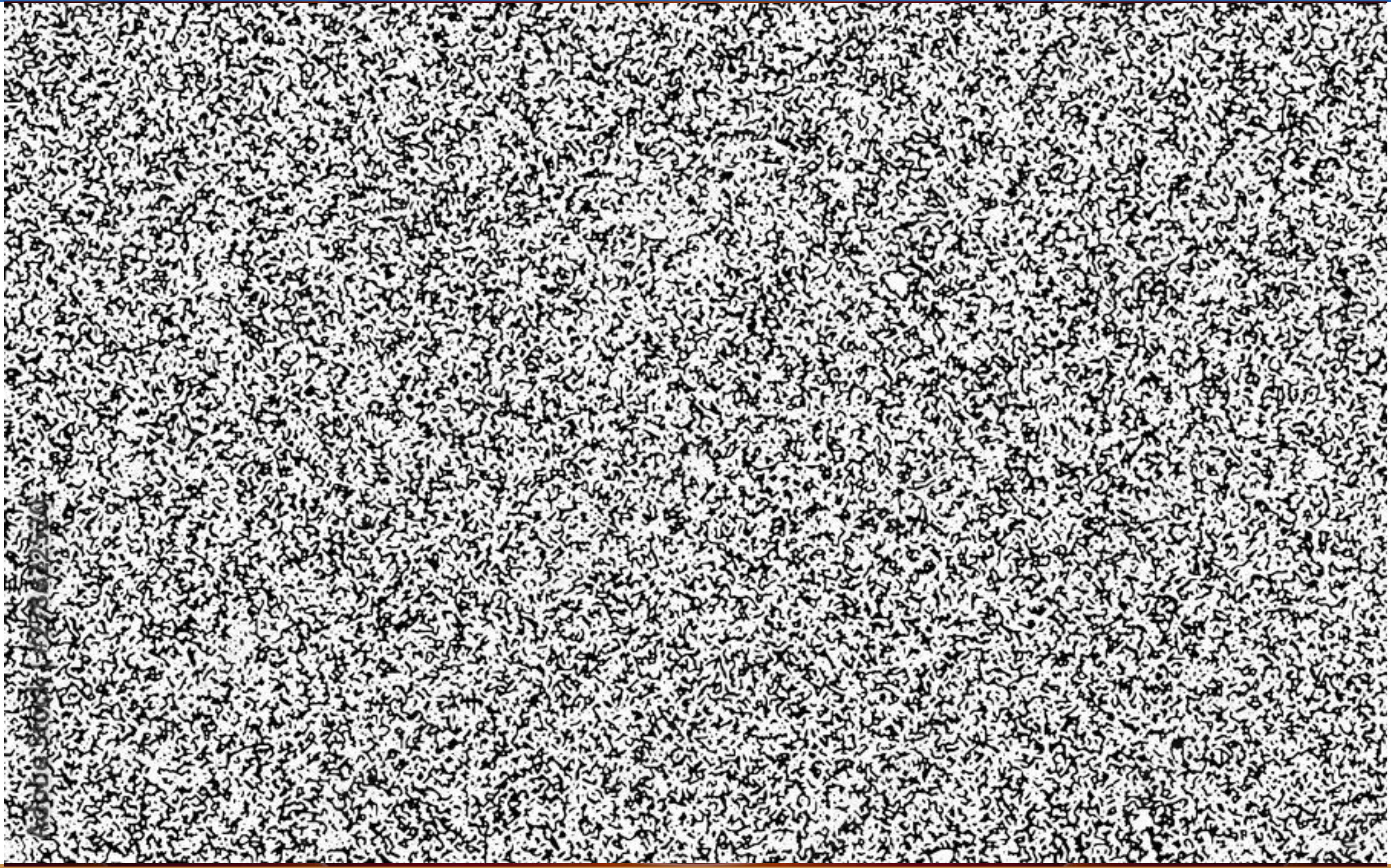
Percentage

70% - 80% ?

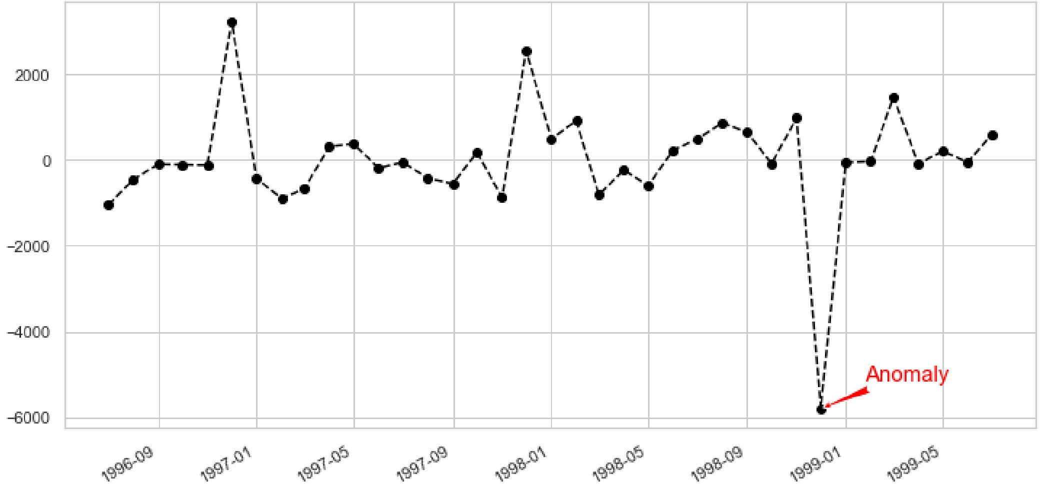
Next Steps

- **Study and implementation of the Non-parametric-FOCuS.**
- **Explainability.**
- ***DataGenerator/ DASK* implementation for training of larger-than-memory Datasets.**
- **Container distribution and optimization for cloud use (portability, scalability, enhanced management).**


That's all Folks!
SUPERB wallpapers



Scientific Rationale



-To develop an Anomaly Detection algorithm for Time Series data using Machine Learning techniques;

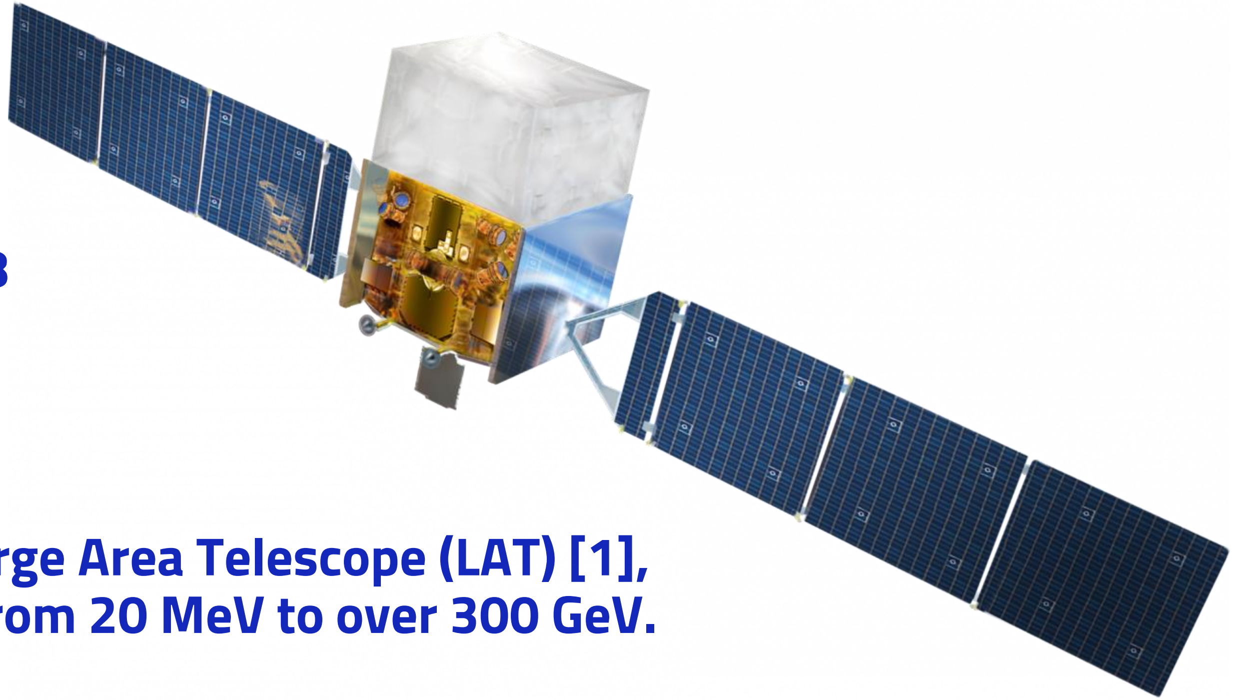
-Apply it to create a pipeline for the astrophysical data from the Fermi Anti-Coincidence Detector (ACD);

Fermi ACD application

-To develop an Anomaly Detection algorithm for Time Series data using Machine Learning techniques;

-Apply it to create a pipeline for the astrophysical data from the Fermi Anti-Coincidence Detector (ACD);

Fermi satellite and ACD



The Fermi Gamma-ray Space Telescope
is a space observatory launched by NASA in 2008
to study high-energy gamma rays.

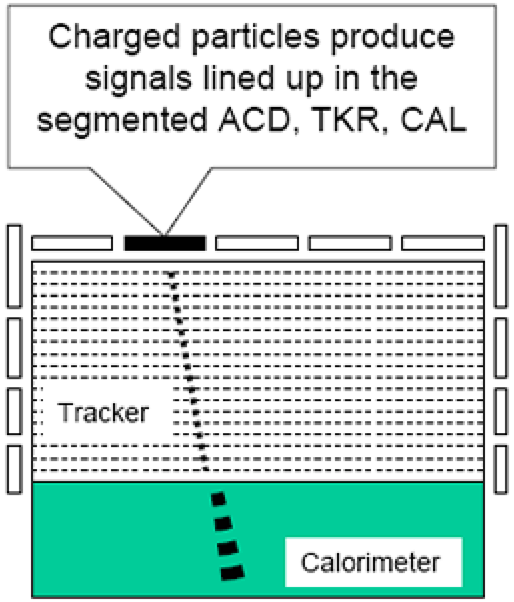
The primary instrument on board Fermi is the Large Area Telescope (LAT) [1],
which detects gamma rays in the energy range from 20 MeV to over 300 GeV.

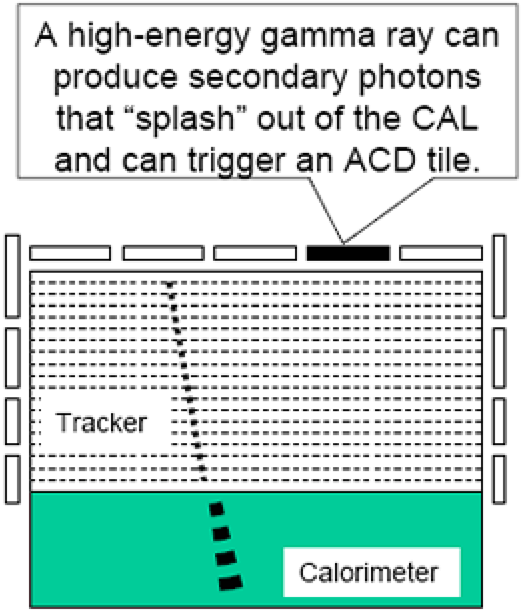
The Gamma-ray Burst Monitor (GBM) [2], designed to observe gamma-ray bursts
in the energy range from 8 keV to 40 MeV.

(1) Atwood 2009 - THE LARGE AREA TELESCOPE ON THE FERMI GAMMA-RAY SPACE TELESCOPE MISSION
(2) Meegan 2009 - THE FERMI GAMMA-RAY BURST MONITOR

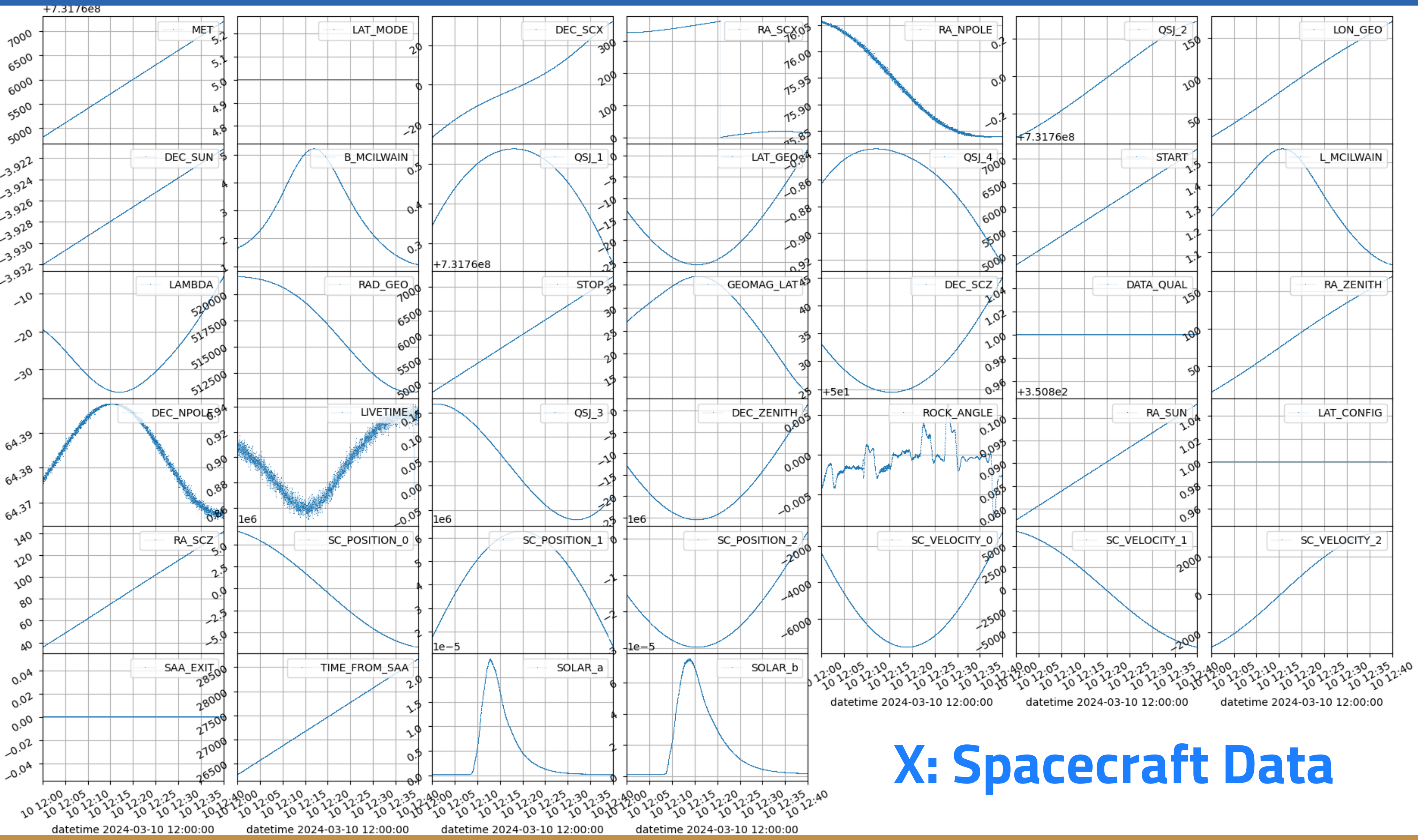
Fermi satellite and ACD

The LAT instrument is surrounded by its Anti-Coincidence Detector (ACD), used to filter out unwanted signals, such as cosmic rays, that can mimic gamma-ray signatures.





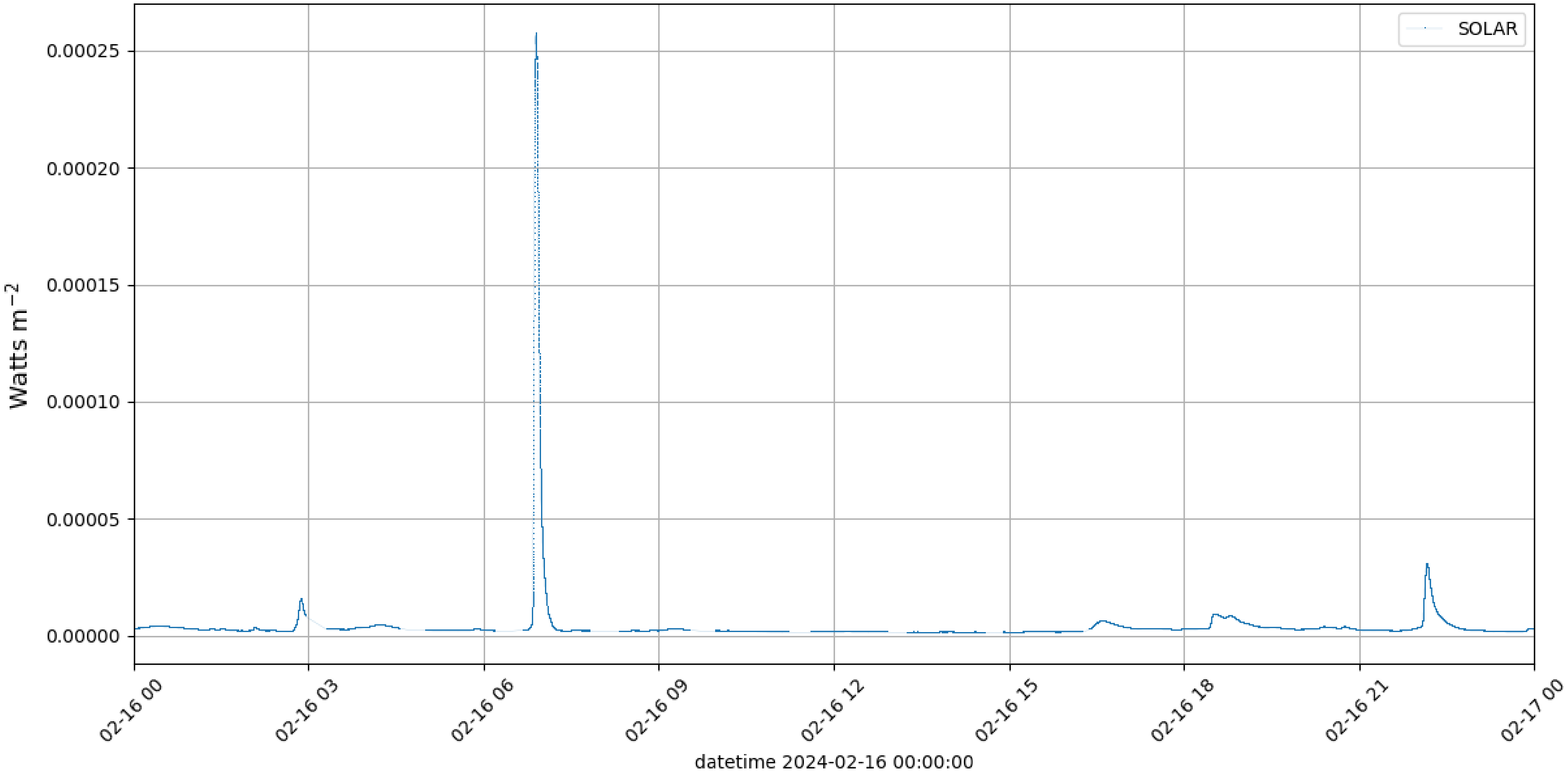
The ACD consists of an array of plastic scintillator tiles, which emit light when traversed by charged particles. By detecting these particles, the ACD helps identify and reject events caused by charged particles, allowing the LAT to focus on gamma-ray signals.



X: Spacecraft Data

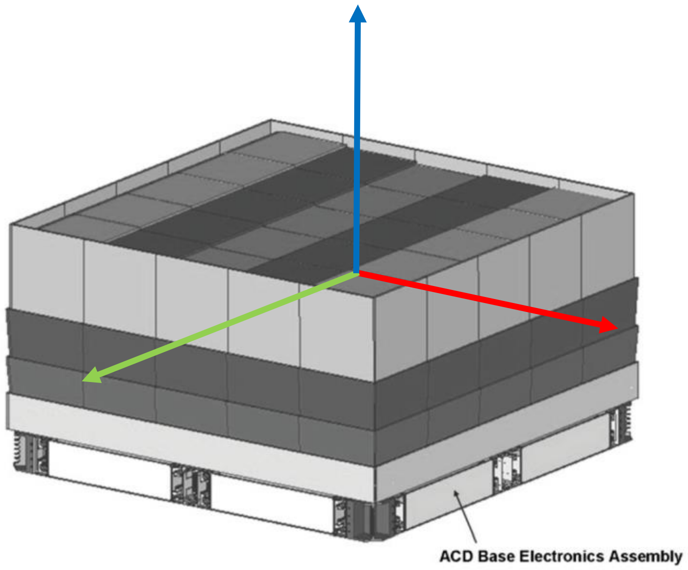
X: Solar Activity Data from GOES X-Ray Sensor (XRS)

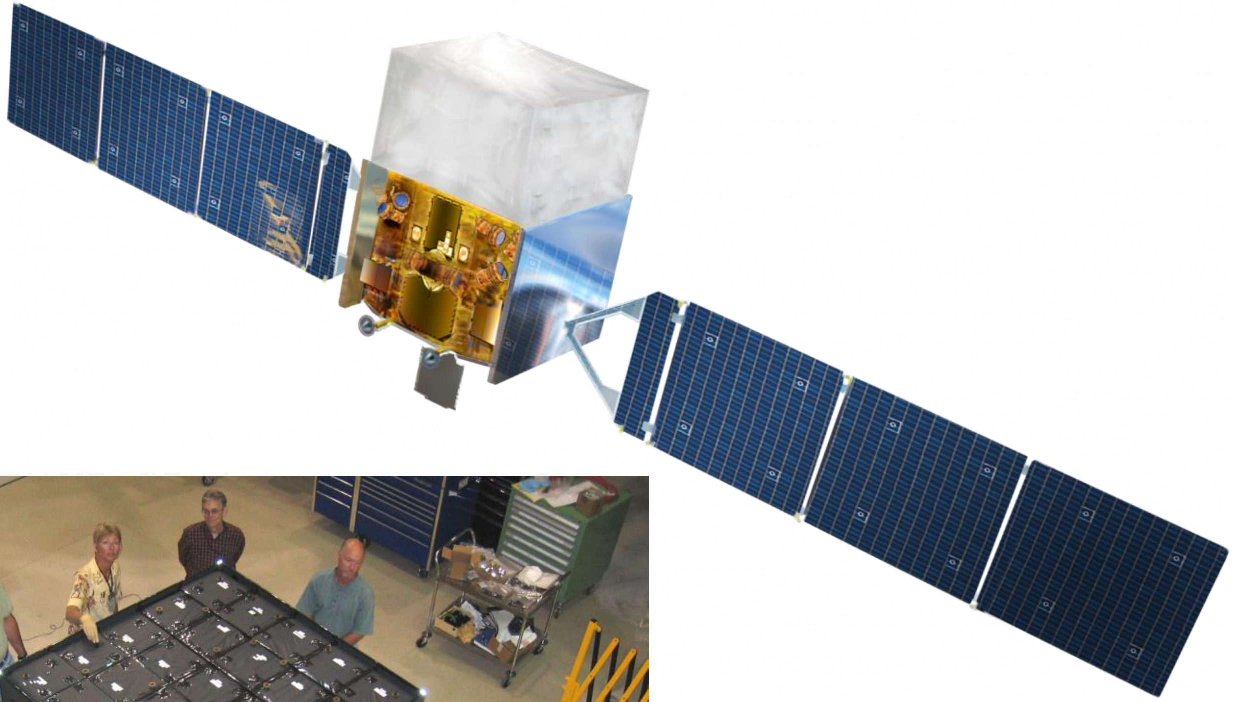
It describes the flux of X-rays coming from the Sun.



Y: ACD Data

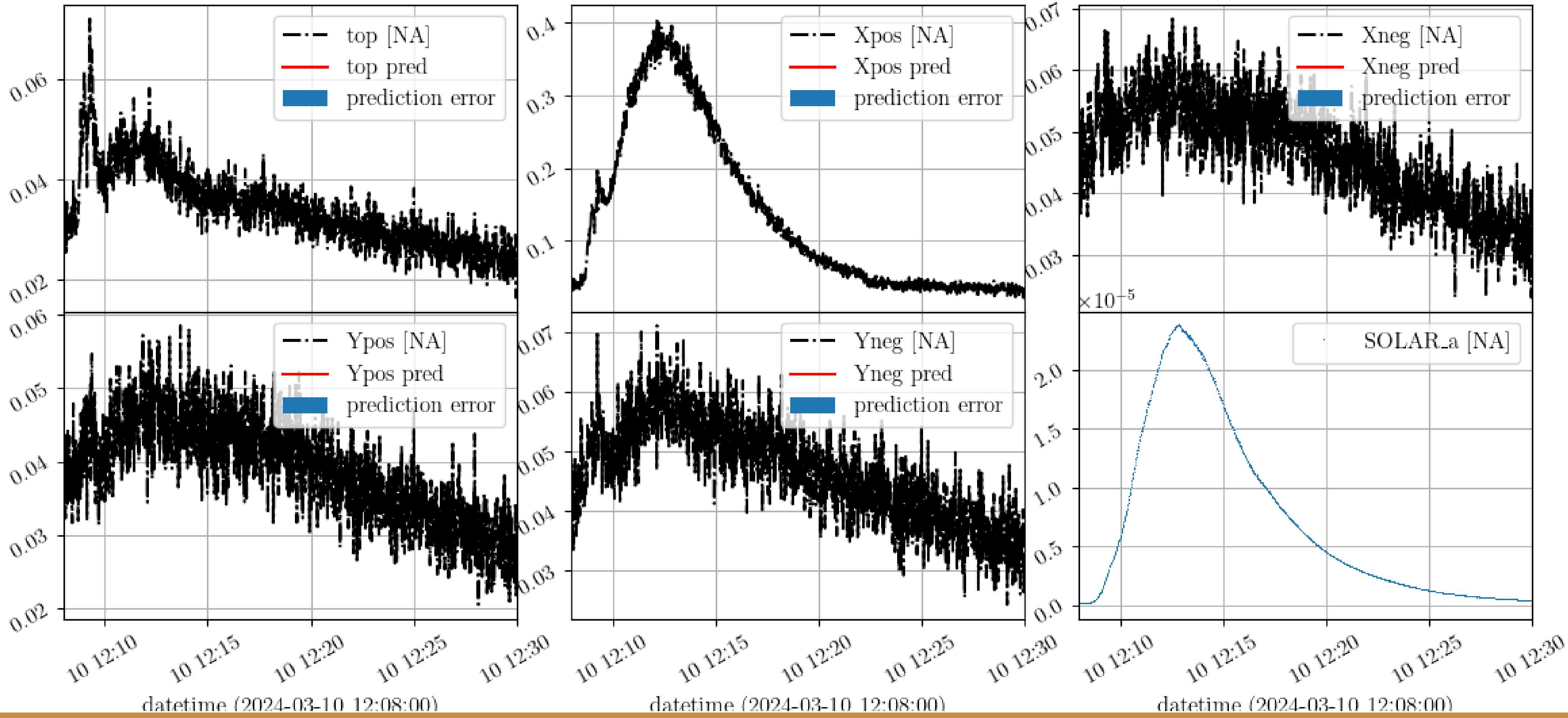
- -top (Z)
- -Xpos (X+)
- -Xneg (X-)
- -Ypos (Y+)
- -Yneg (Y-)





Y: ACD Data

- -top (Z)
- -Xpos (X+)
- -Xneg (X-)
- -Ypos (Y+)
- -Yneg (Y-)

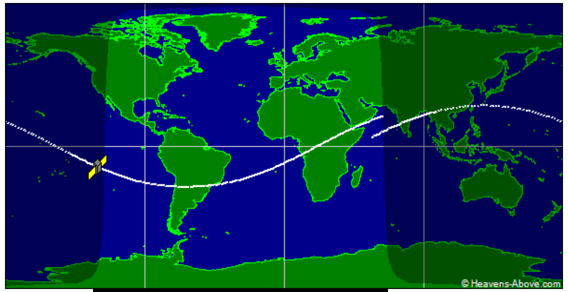


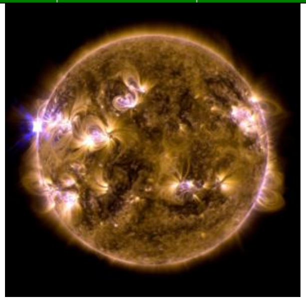
Dataset

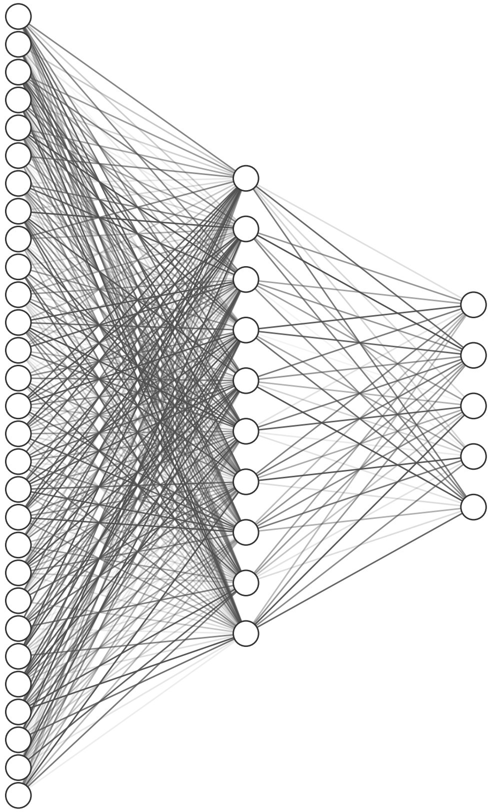
It is divided in around 30 input parameters from Spacecraft files + 1 from GOES data for the solar activity, and the signals from the 5 faces of the ACD.

Input parameters (FT2):

START
STOP
SC_POSITION
SC_VELOCITY
LAT_GEO
LON_GEO
...
SOLAR ACTIVITY

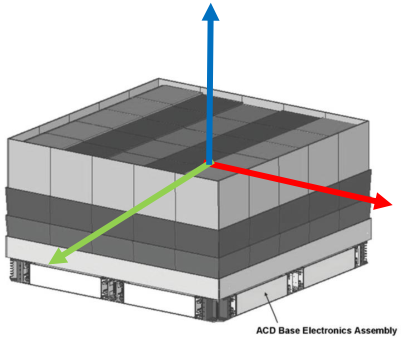






Output parameters:

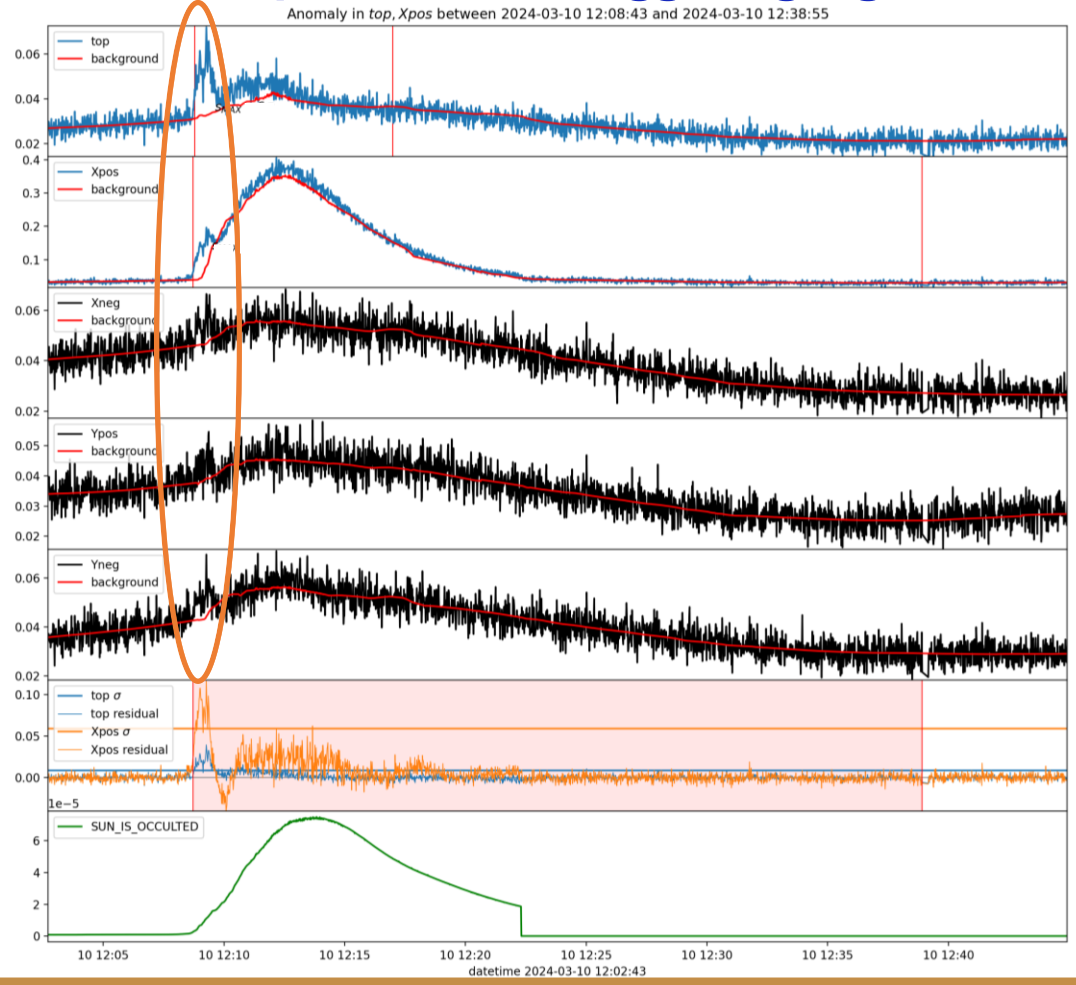
top signal
Xpos signal
Xneg signal
Ypos signal
Yneg signal

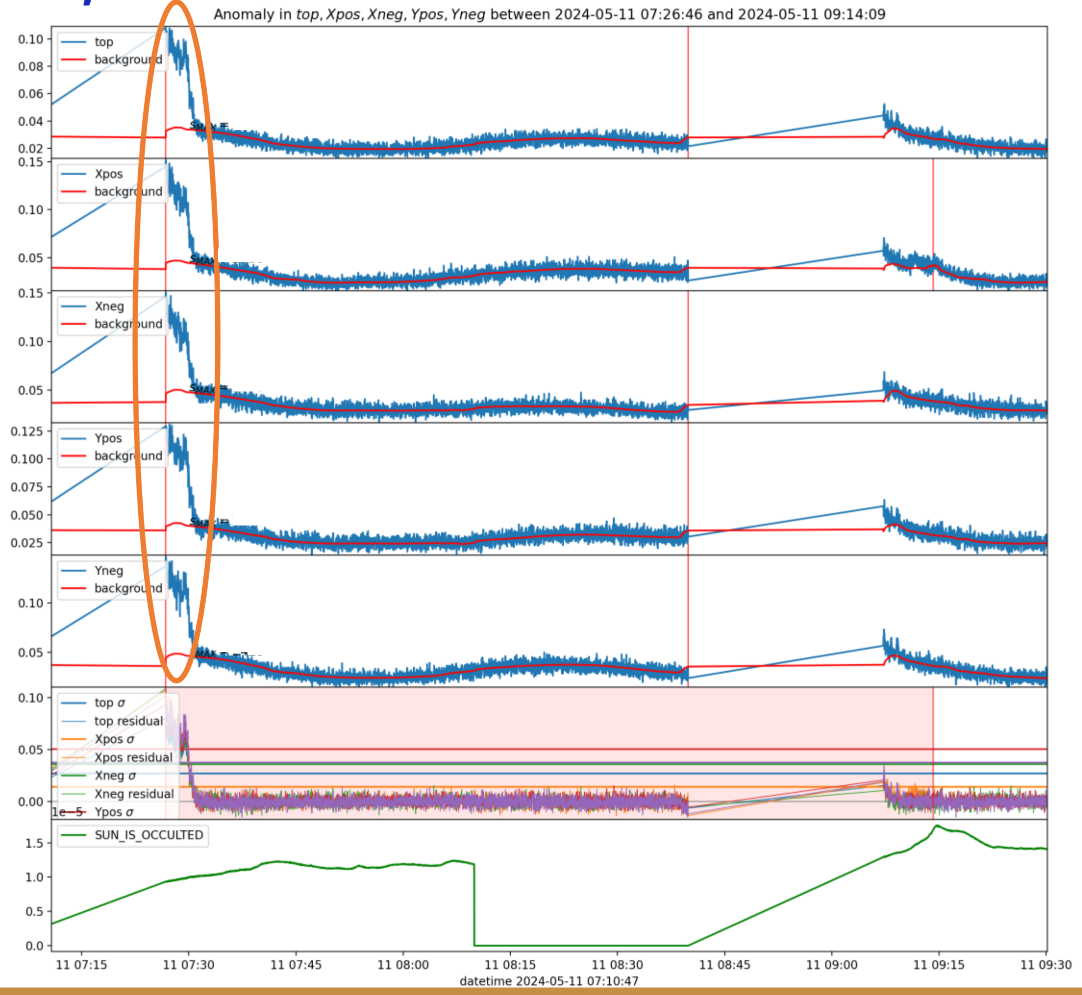


Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

September 2024

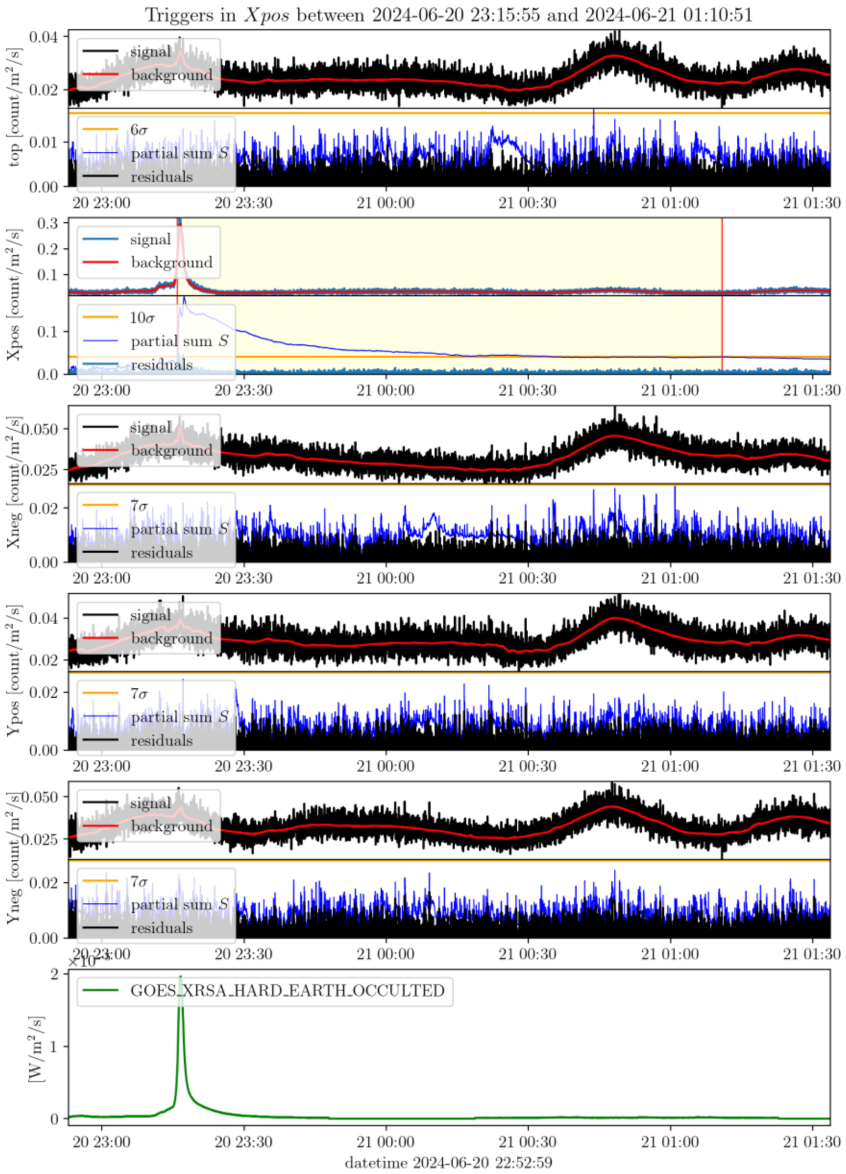
We implemented a triggering algorithm, the Gaussian-FOCuS, here some tests:

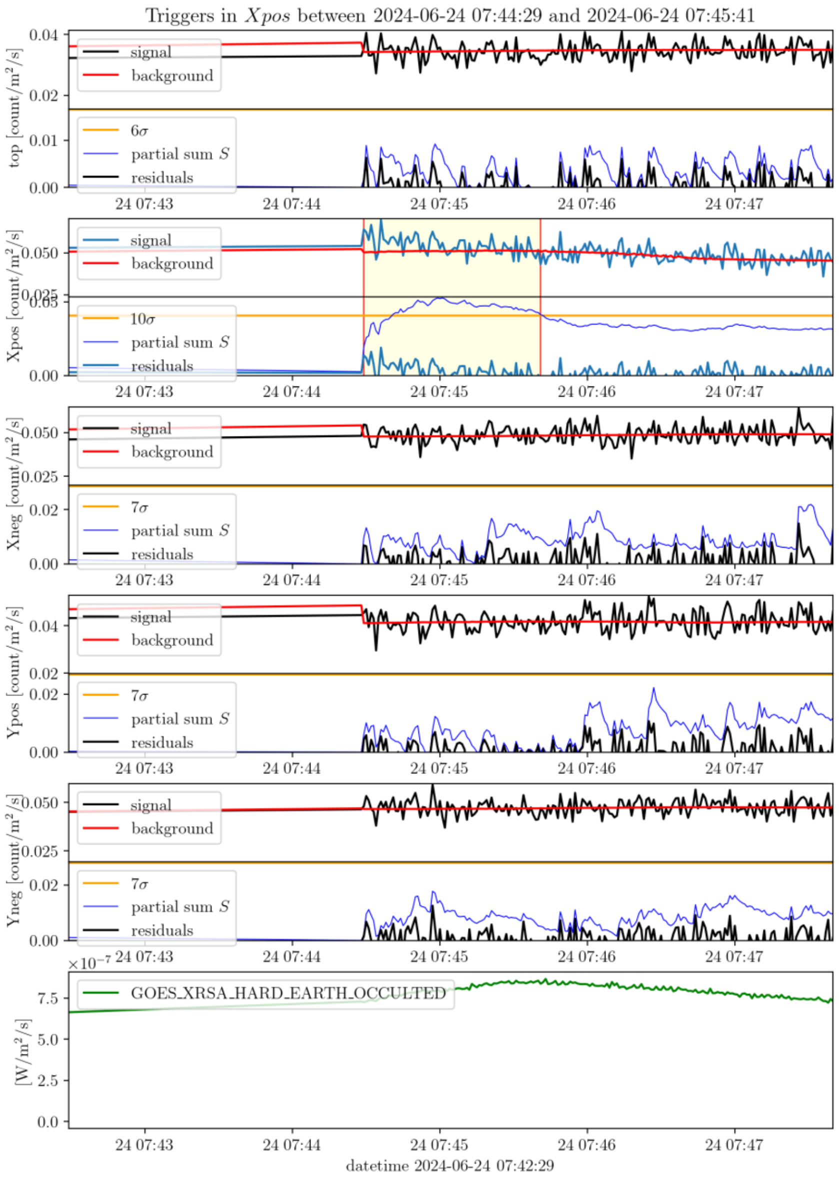




Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

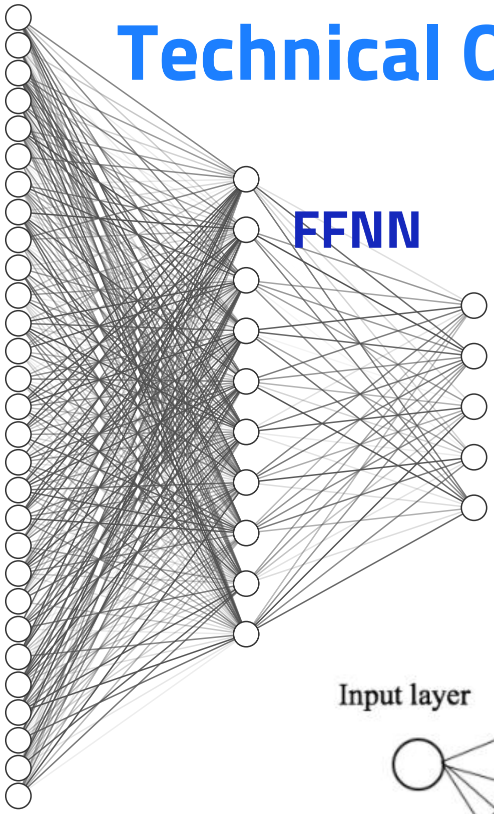
Gaussian-FOCuS Results



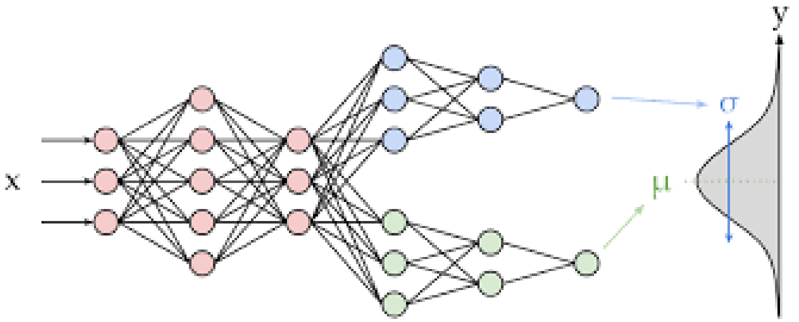


Technical Objectives, Methodologies and Solutions

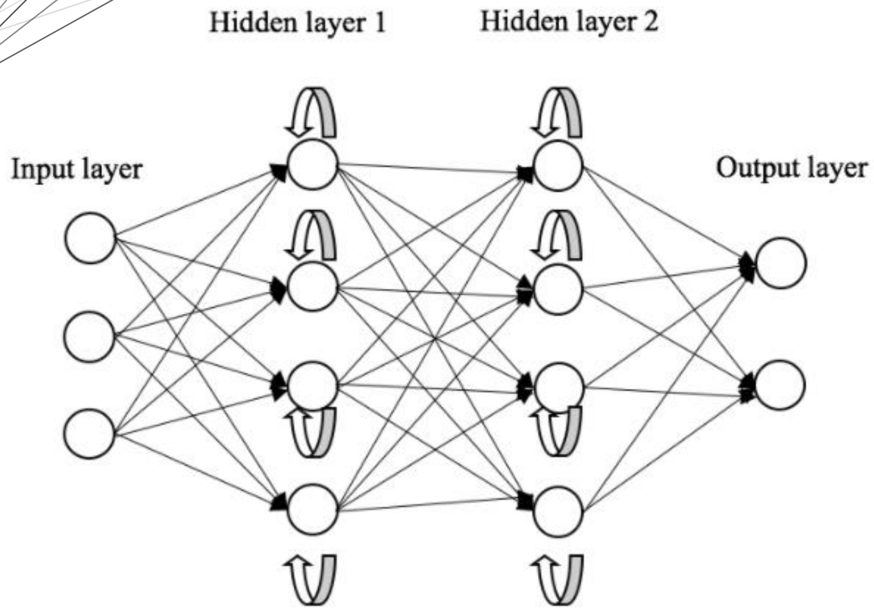
FFNN



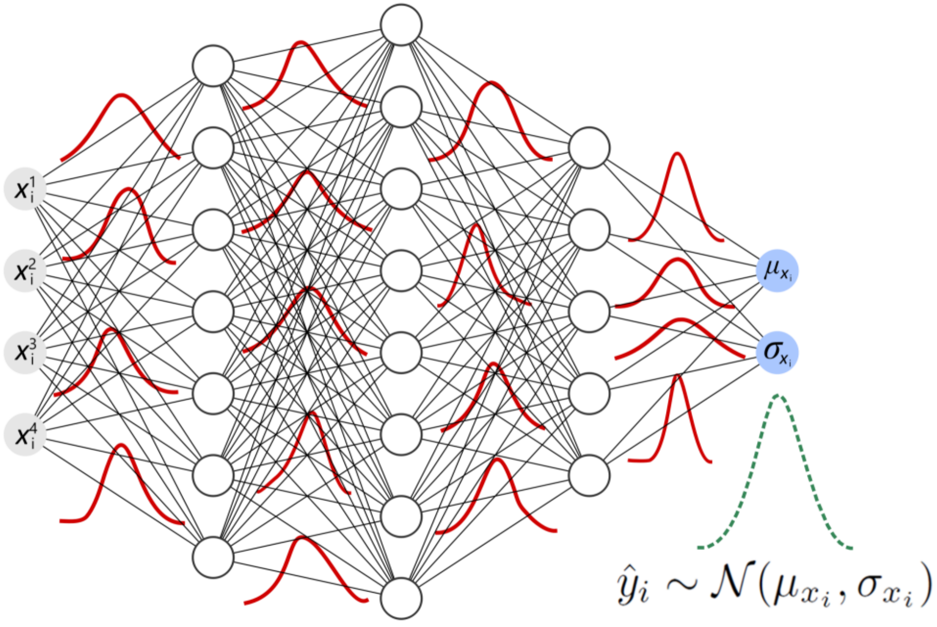
BNN



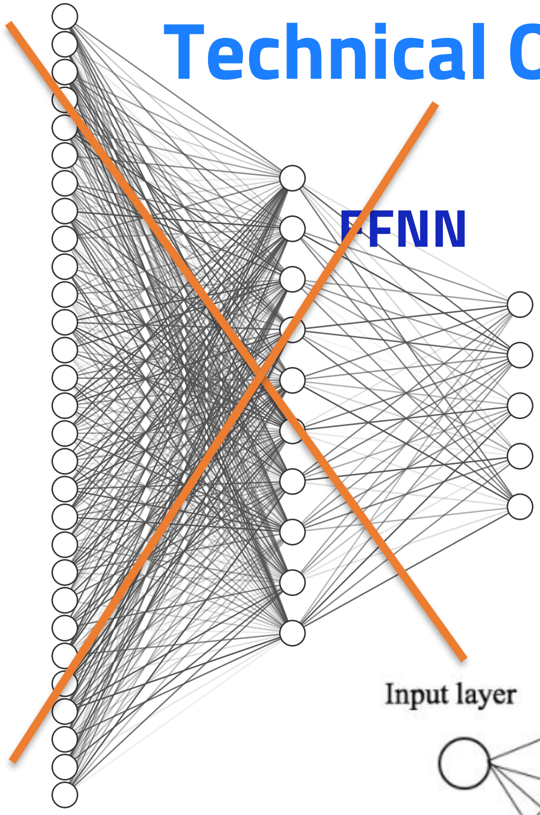
RNN



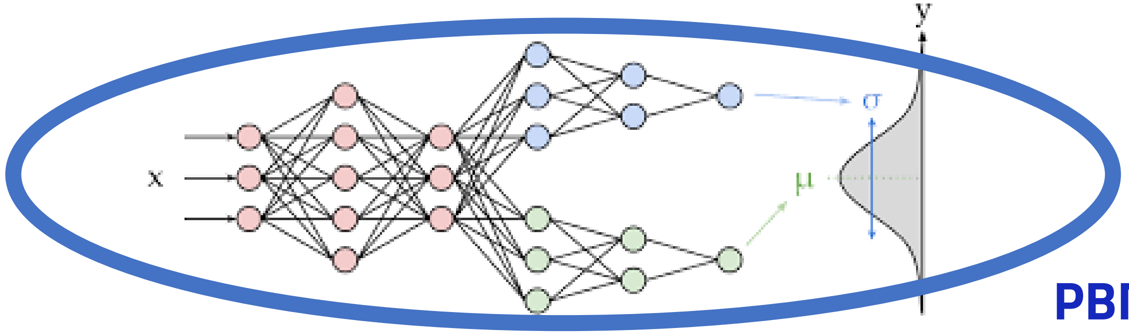
PBNN



Technical Objectives, Methodologies and Solutions

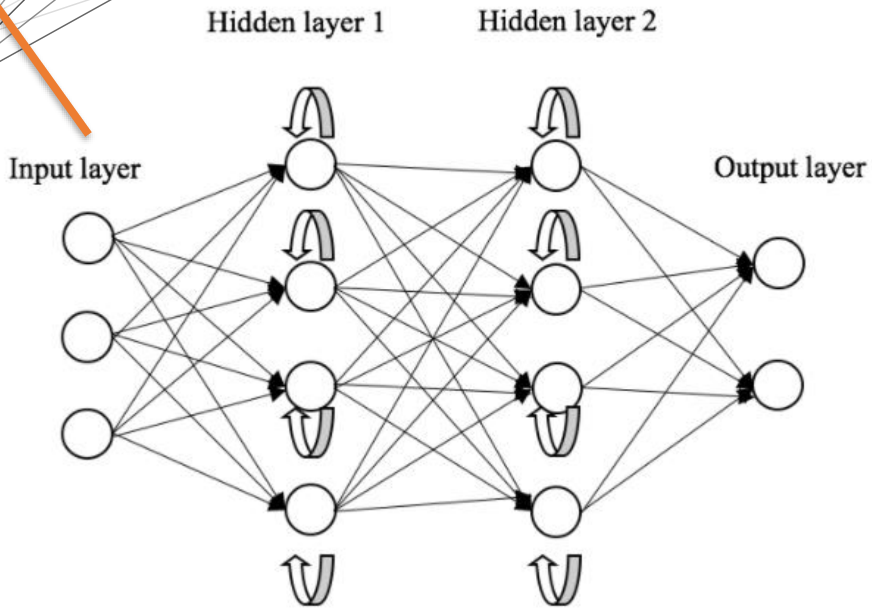


FFNN

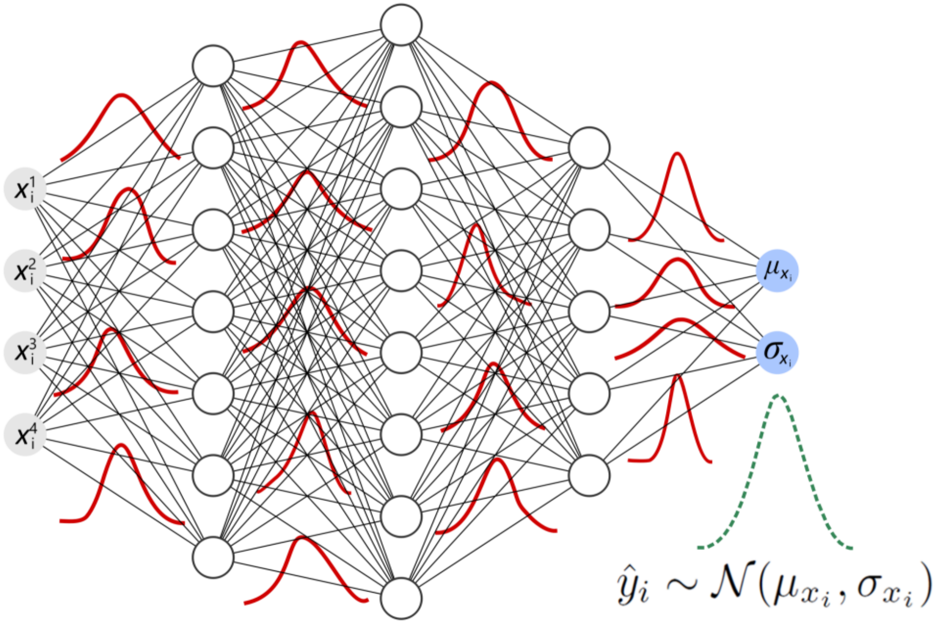


BNN

RNN



PBNN

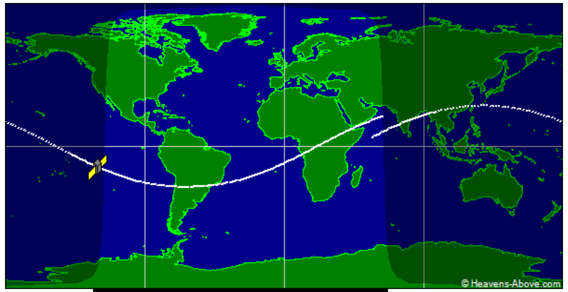


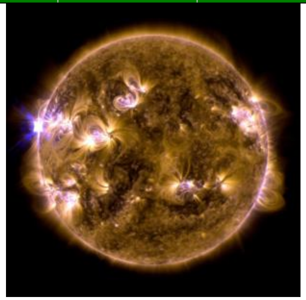
Dataset

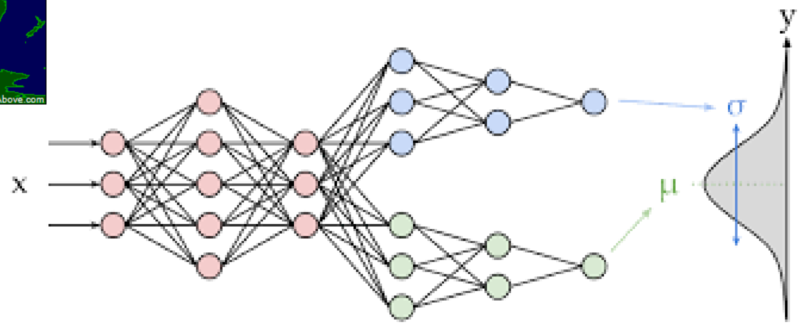
It is divided in around 30 input parameters from Spacecraft files + 1 from GOES data for the solar activity, and the signals from the 5 faces of the ACD.

Input parameters (FT2):

START
STOP
SC_POSITION
SC_VELOCITY
LAT_GEO
LON_GEO
...
SOLAR ACTIVITY



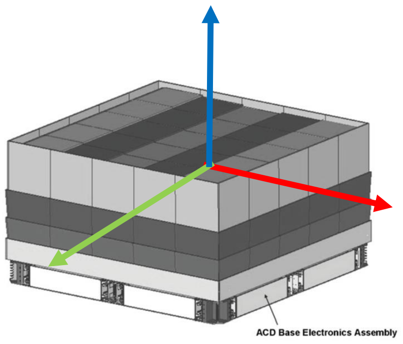




Output parameters:

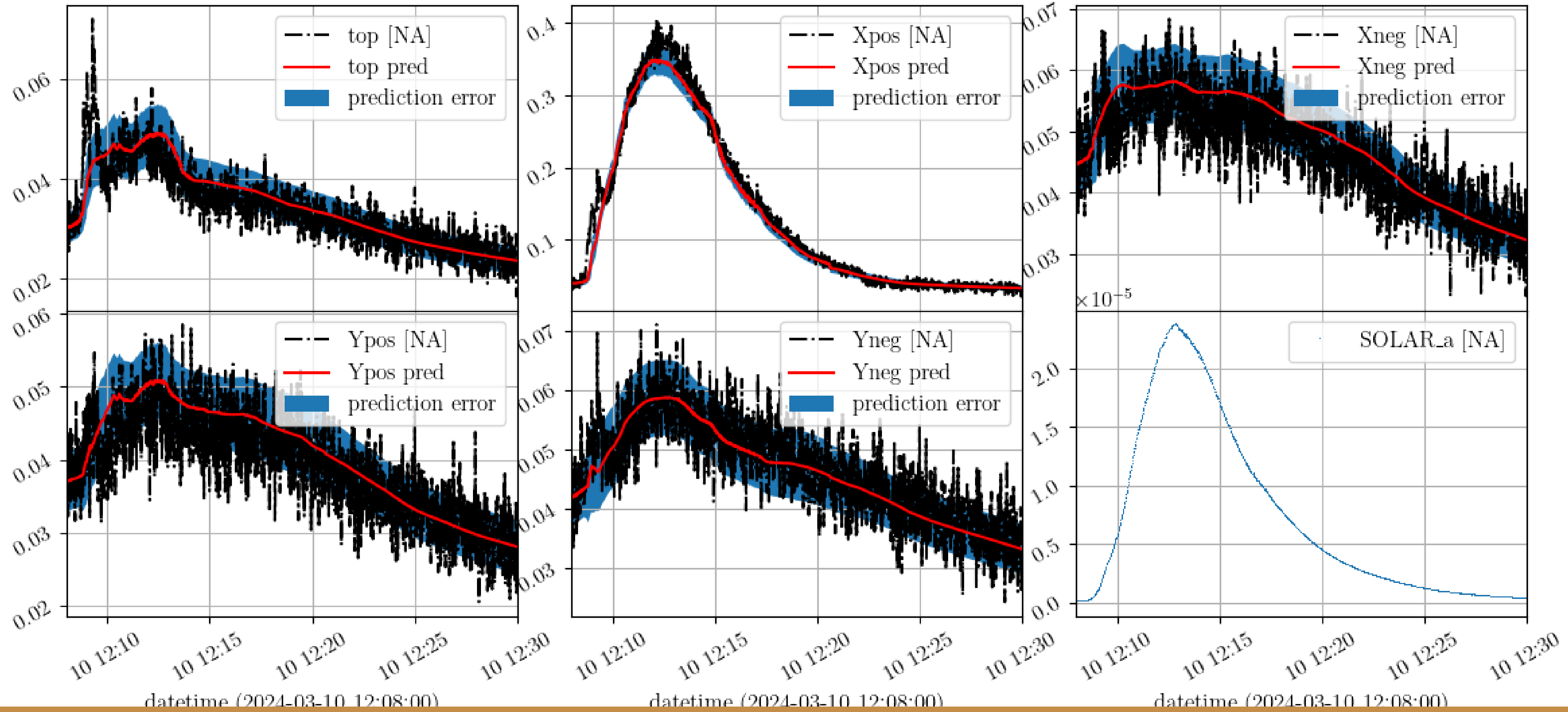
top std
Xpos std
Xneg std
Ypos std
Yneg std

top mean
Xpos mean
Xneg mean
Ypos mean
Yneg mean



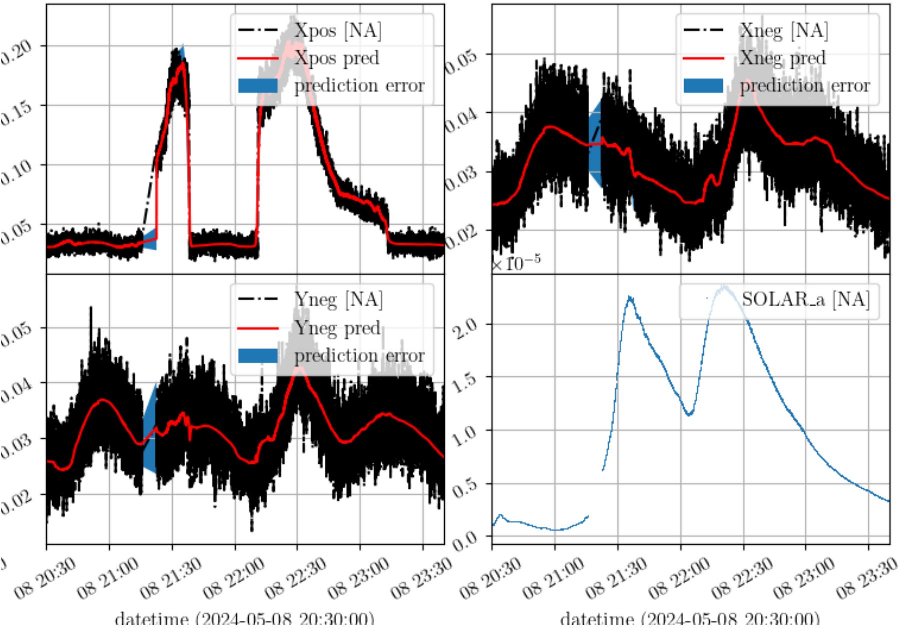
NN Results

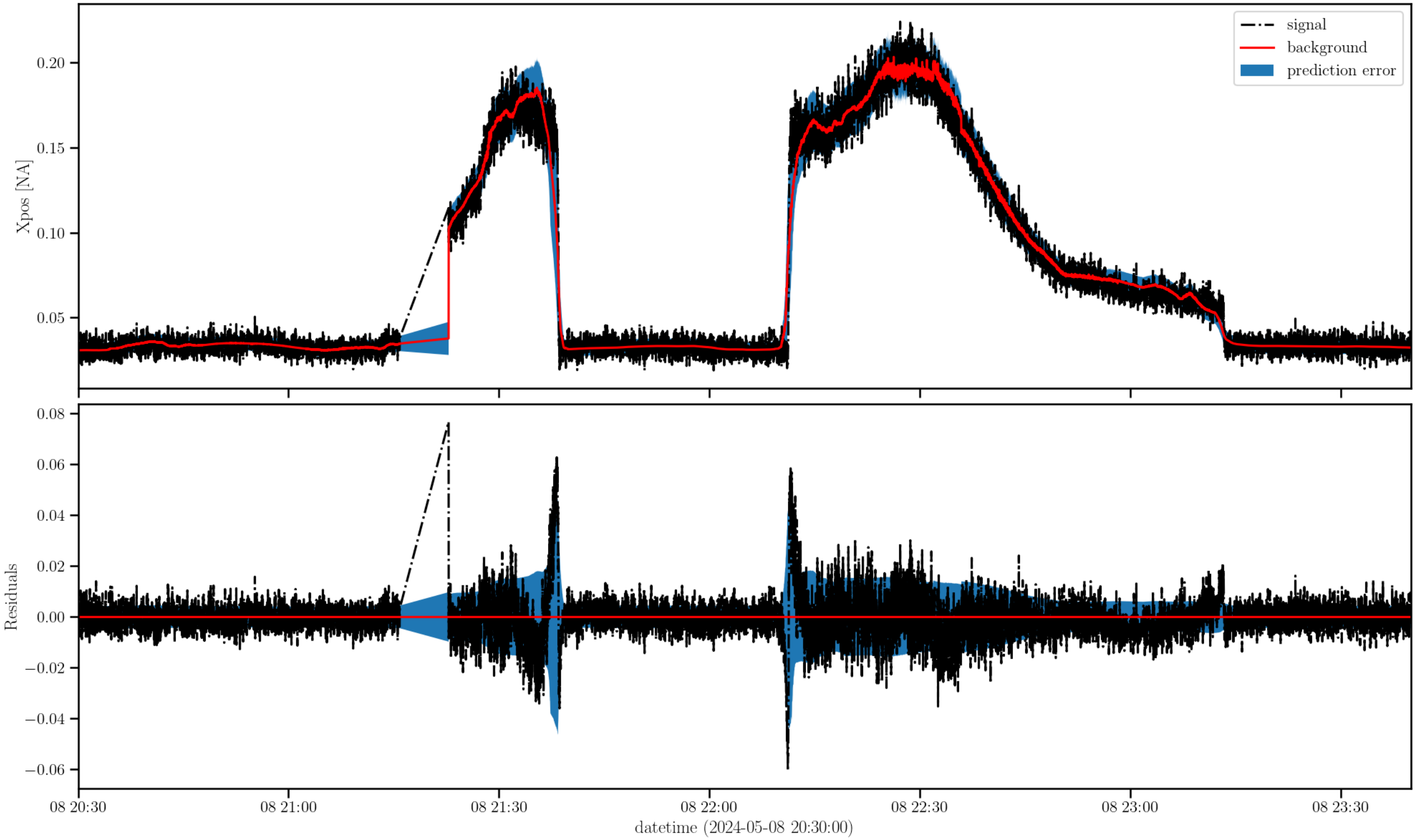
This is the prediction of the model for the Xpos signal.



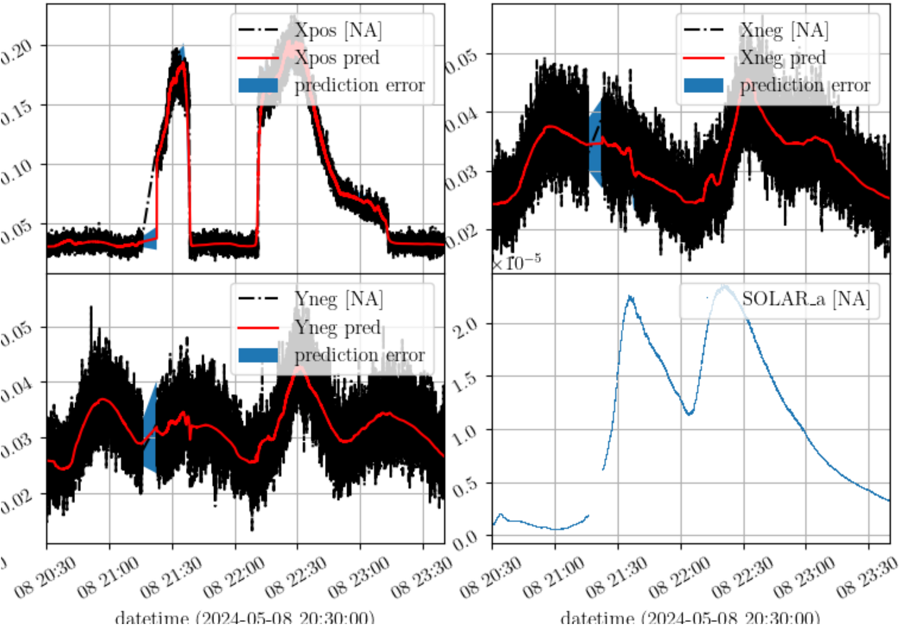
NN Results

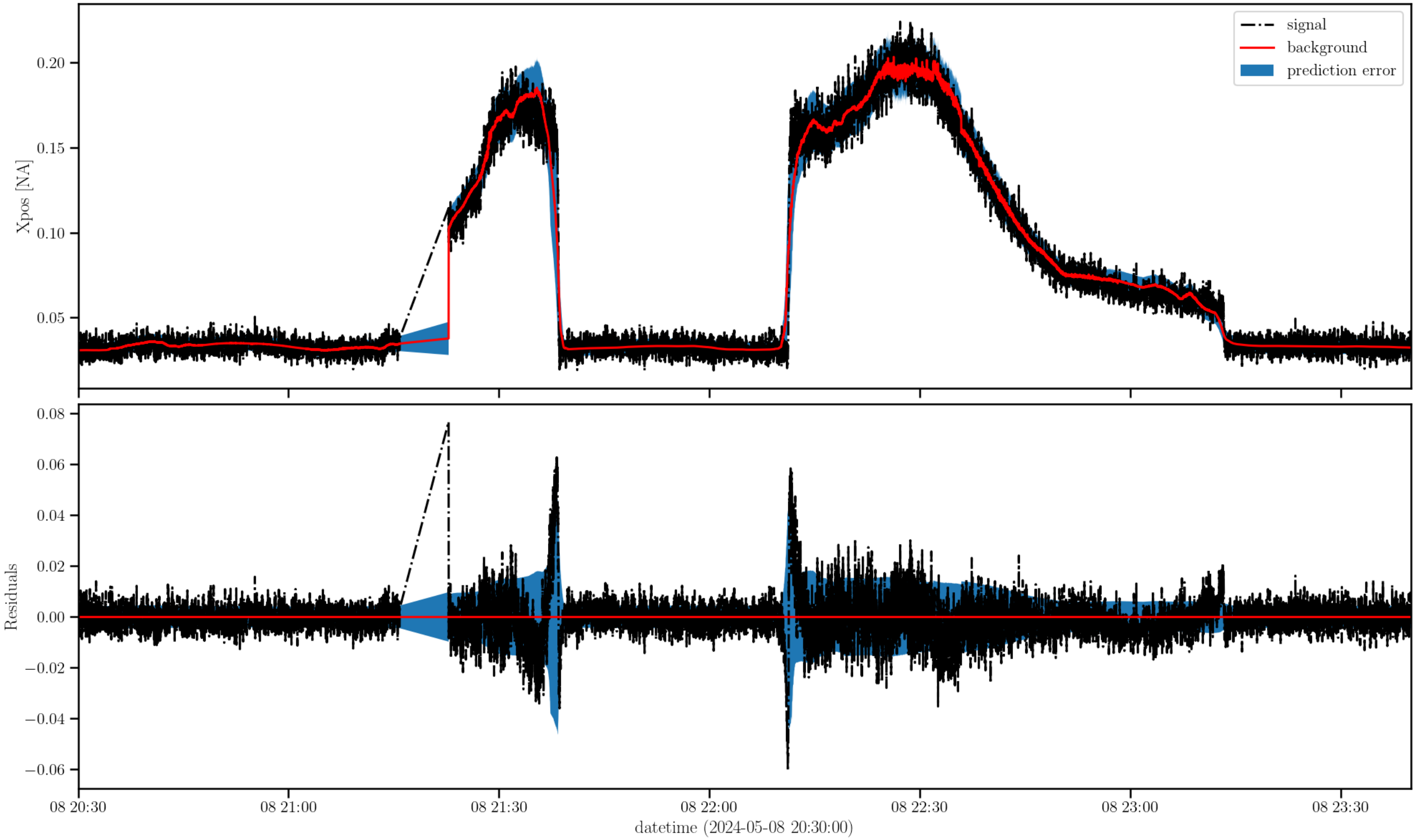
This is the prediction of the model for the Xpos signal.





Standardization of the data distribution?

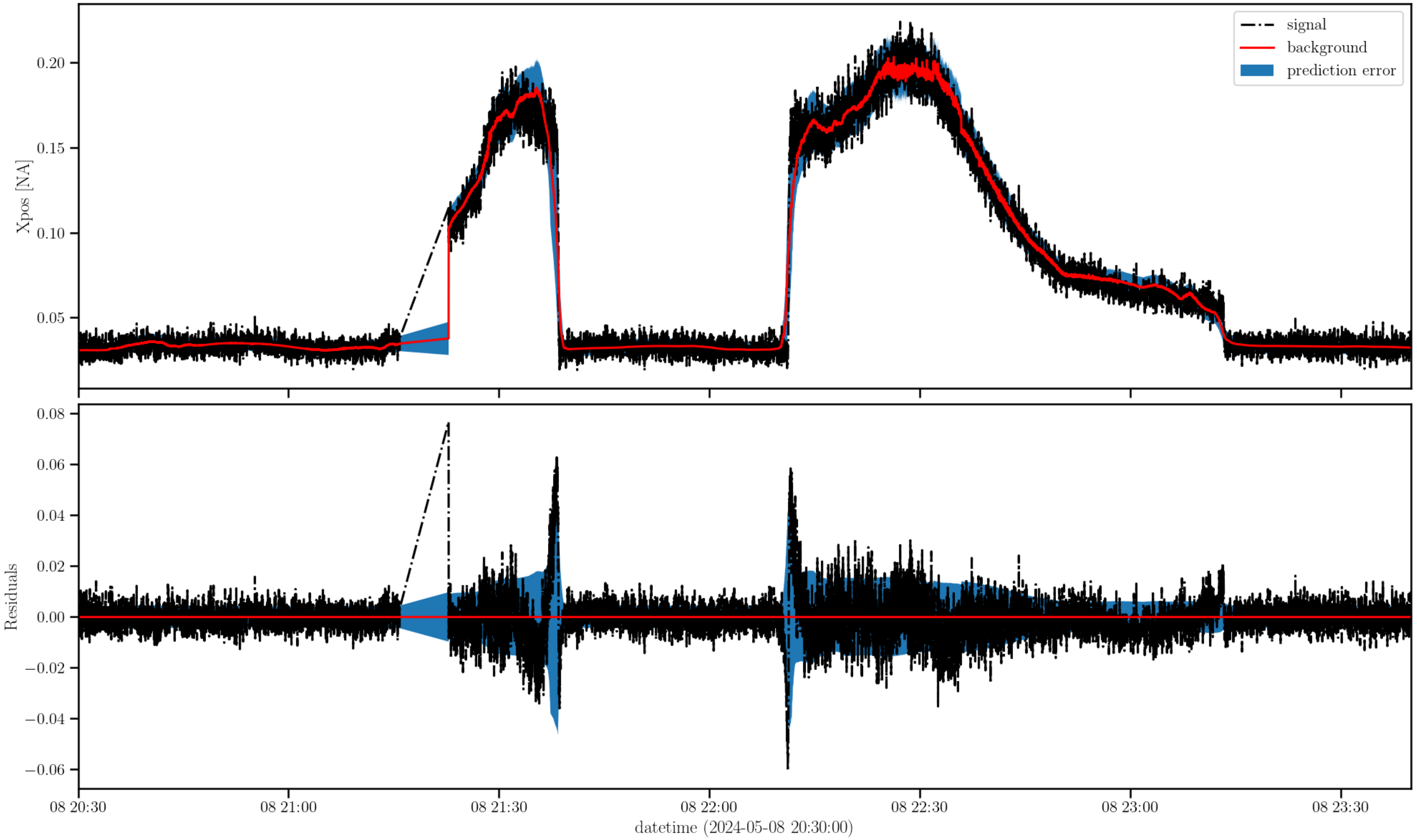


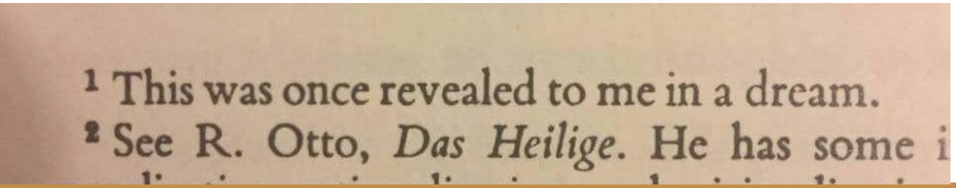


Standardization of the data distribution?

Such an approach would require the variance estimates to be very accurate.

This would be the way to go if we decided to stick with Gaussian-FOCuS.




[1] This was once revealed to me in a dream.
[2] See R. Otto, *Das Heilige*. He has some i

Triggering Algorithm: FOCuS

Identify the anomalies in the data as deviations from the background, quantifying the significance of each anomaly.

The Functional Online CuSUM (FOCuS) is a fast and efficient algorithm based on the computation of the cumulative sum of the score statistics of the data.

Can be used in *flavours*:

-Poisson-FOCuS: assumes a Poisson-like distribution of data; can be used for count rates data.

-Gaussian-FOCuS: assumes a Gaussian distribution; can be used for varying signals (temperature...)

-Non-parametric-FOCuS: no assumptions on the parameters that describe the data distribution.

Next Steps

-Study and implementation of the Non-parametric-FOCuS.

-Explainability.

-*DataGenerator/ DASK* implementation for training of larger-than-memory Datasets (WORK IN PROGRESS).

-Distribution for on-cloud use.

Percentage

60%

Next Steps

- Study and implementation of the Non-parametric-FOCuS.

- Standardize the dataset and feed it to the triggering algorithm to save all the results.

- Complete the second version of the dataset (third version with cut in energy?)

- Karaoke night at *Empire Pub*, 18 Wed. at 22.30

- Clean the slides from trashy images!


SUPERB wallpapers
That's all Folks!
FOR REAL