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 in Time Series on Fermi ACD Data with ML techniques

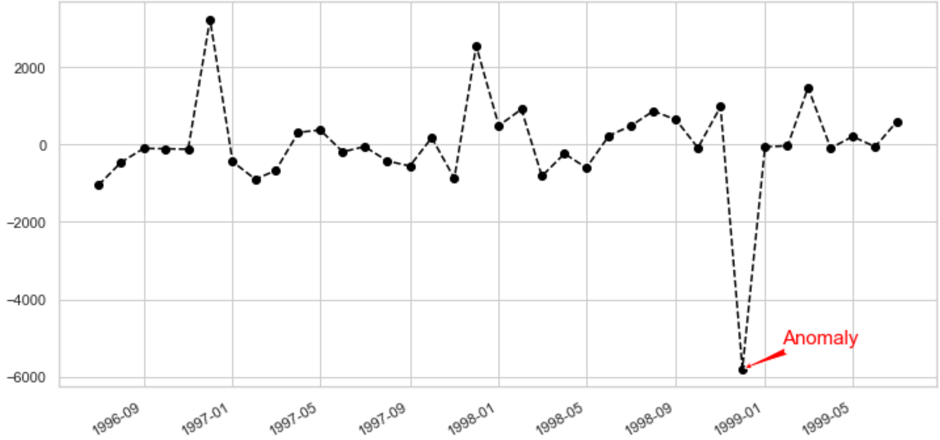
Andrea Adelfio and Sara Cutini for WP3 (INFN Perugia)

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

Scientific Rationale

- To develop an Anomaly Detection algorithm for Time Series data;
- By means of Astrophysical Time Series data from the Anti-Coincidence Detector (ACD) on board of the Fermi satellite;
- Use of Machine Learning techniques to get a baseline prediction of the background signal;
- Implementation of a triggering algorithm for detection of the anomalies in the time series;
- In collaboration with F. Longo (UniTS) for Scientific purposes;
- In collaboration with R. Crupi from Intesa Sanpaolo to work on Banking Time Series data.

Scientific Rationale



- To develop an Anomaly Detection algorithm for Time Series data;

- By means of Astrophysical Time Series data from the Anti-Coincidence Detector (ACD) on board of the Fermi satellite;

- Use of Machine Learning techniques to get a baseline prediction of the background signal;

- Implementation of a triggering algorithm for detection of the anomalies in the time series;

- In collaboration with F. Longo (UniTS) for Scientific purposes;

- In collaboration with R. Crupi from Intesa Sanpaolo to work on Banking Time Series data.

Scientific Rationale

- To develop an Anomaly Detection algorithm for Time Series data;

- By means of Astrophysical Time Series data from the Anti-Coincidence Detector (ACD) on board of the Fermi satellite;

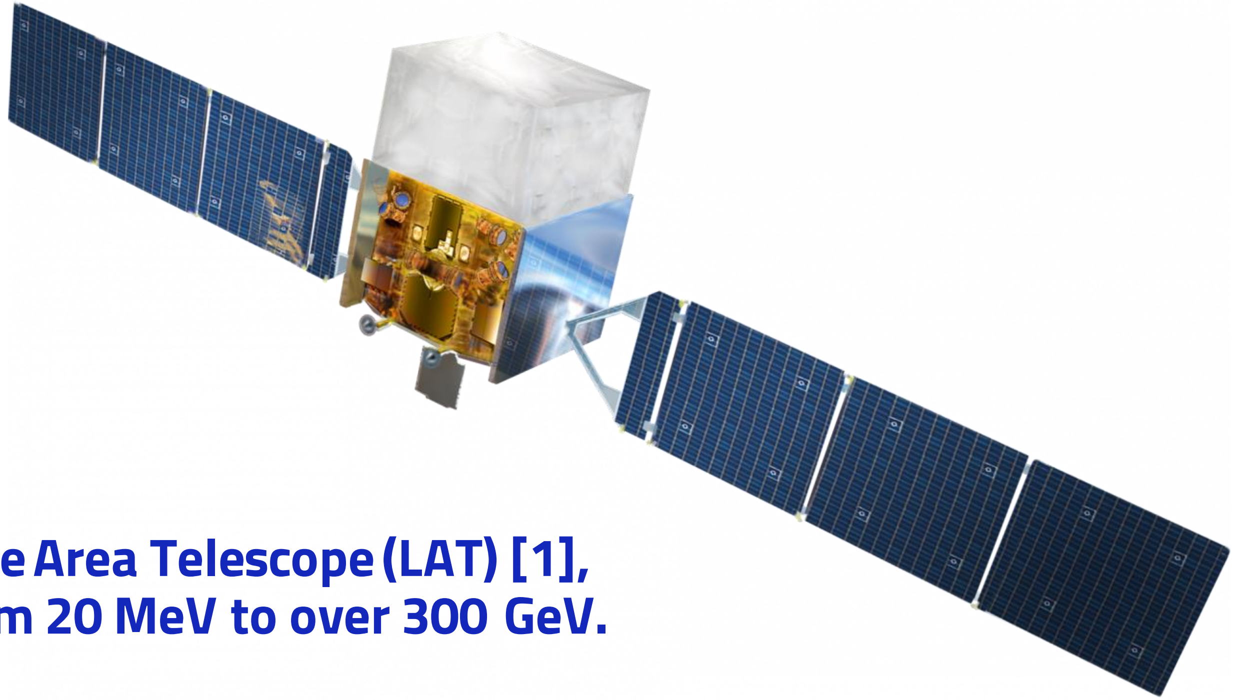
- Use of Machine Learning techniques to get a baseline prediction of the background signal;

- Implementation of a triggering algorithm for detection of the anomalies in the time series;

- In collaboration with F. Longo (UniTS) for Scientific purposes;

- In collaboration with R. Crupi from Intesa Sanpaolo to work on Banking Time Series data.

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.

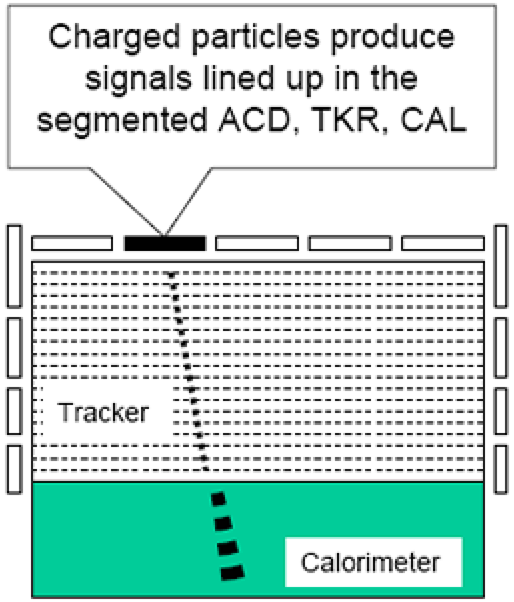
The primary instrument on board Fermi is the Large Area Telescope (LAT) [1],
which detects gamma rays in the energy range from 20 MeV to over 300 GeV.

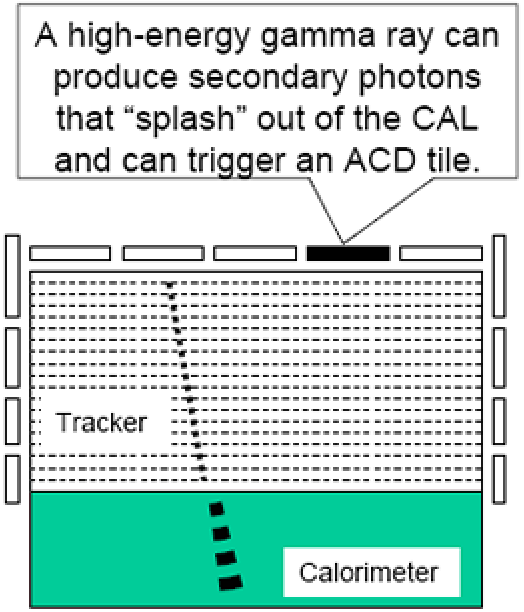
The Gamma-ray Burst Monitor (GBM) [2], designed to observe gamma-ray bursts
in the energy range from 8 keV to 40 MeV.

(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.

Data

We start with a small 52 days long dataset, from the 2023-12-05 to 2024-01-26.

It consists of data with a time resolution of 1 second:

-Particles count rates for each tile of the ACD system (~800 MB/week of data before reduction);

-Weekly Spacecraft data (from Fermi FTP), which contains all the parameters describing the spatial configuration of the satellite, with other parameters such as the geomagnetic flux description along the orbit (~80 MB/week);

-Solar Activity from the Geostationary Operational Environmental Satellite (GOES) X-Ray Sensor (XRS) (~35 MB/week).

Particle Count Rates Data

The reduced dataset contains the count rate of particles in each face of the ACD. Based on the three axes of the cube (X,Y,Z) we call the faces:

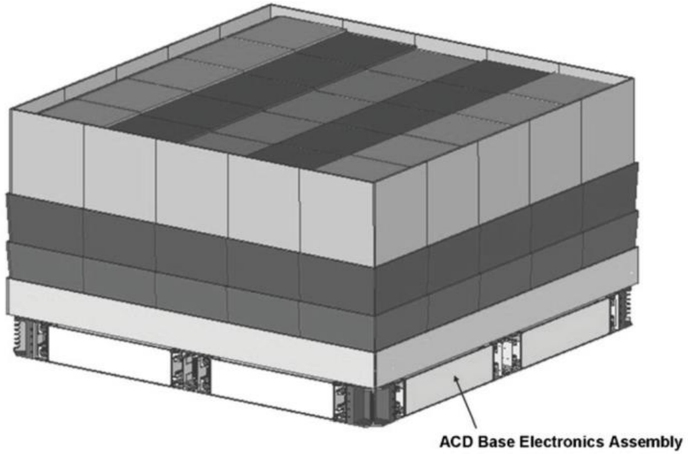
-top (Z)

-Xpos (X+)

-Xneg (X-)

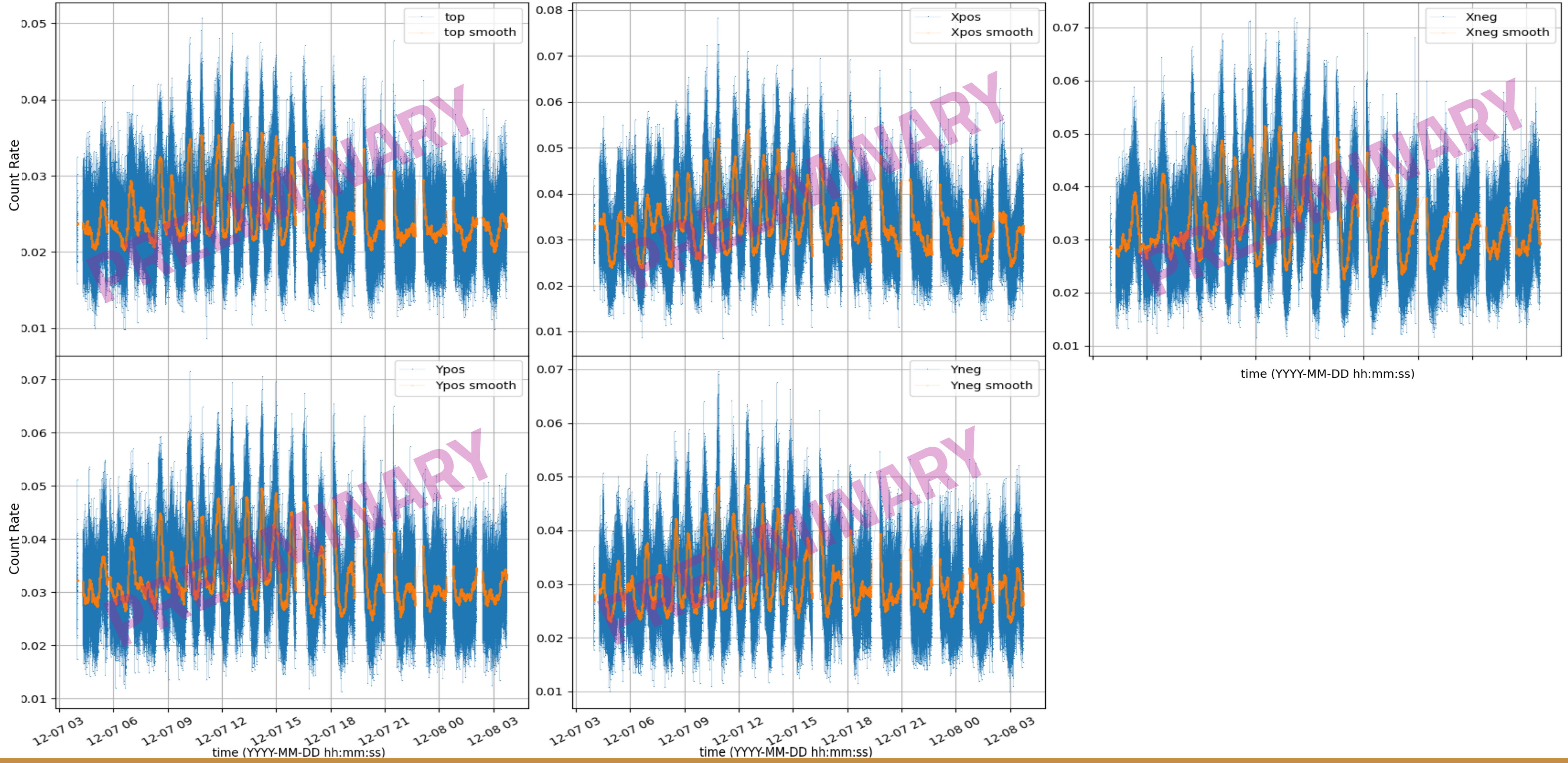
-Ypos (Y+)

-Yneg (Y-)



Particle Count Rates Data

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



Spacecraft Data

START (seconds) -> Mission Elapsed Time of start of interval

STOP (seconds) -> Mission Elapsed Time of end of interval

SC_POSITION (meters) -> Three element array giving the position (x, y, z) of spacecraft in inertial (ECI) coordinates at START

SC_VELOCITY (m/s) -> Three element array giving the spacecraft velocity in the same coordinate frame as SC_POSITION at START

LAT_GEO (deg) -> ground point latitude

LON_GEO (deg) -> ground point longitude

RAD_GEO (m) -> spacecraft altitude

RA_ZENITH (deg) -> RA of zenith direction at START

DEC_ZENITH (deg) -> Dec of zenith direction at START

...

Spacecraft Data

B_MCILWAIN (Gauss) -> McIlwain B parameter, magnitude of the magnetic field at START

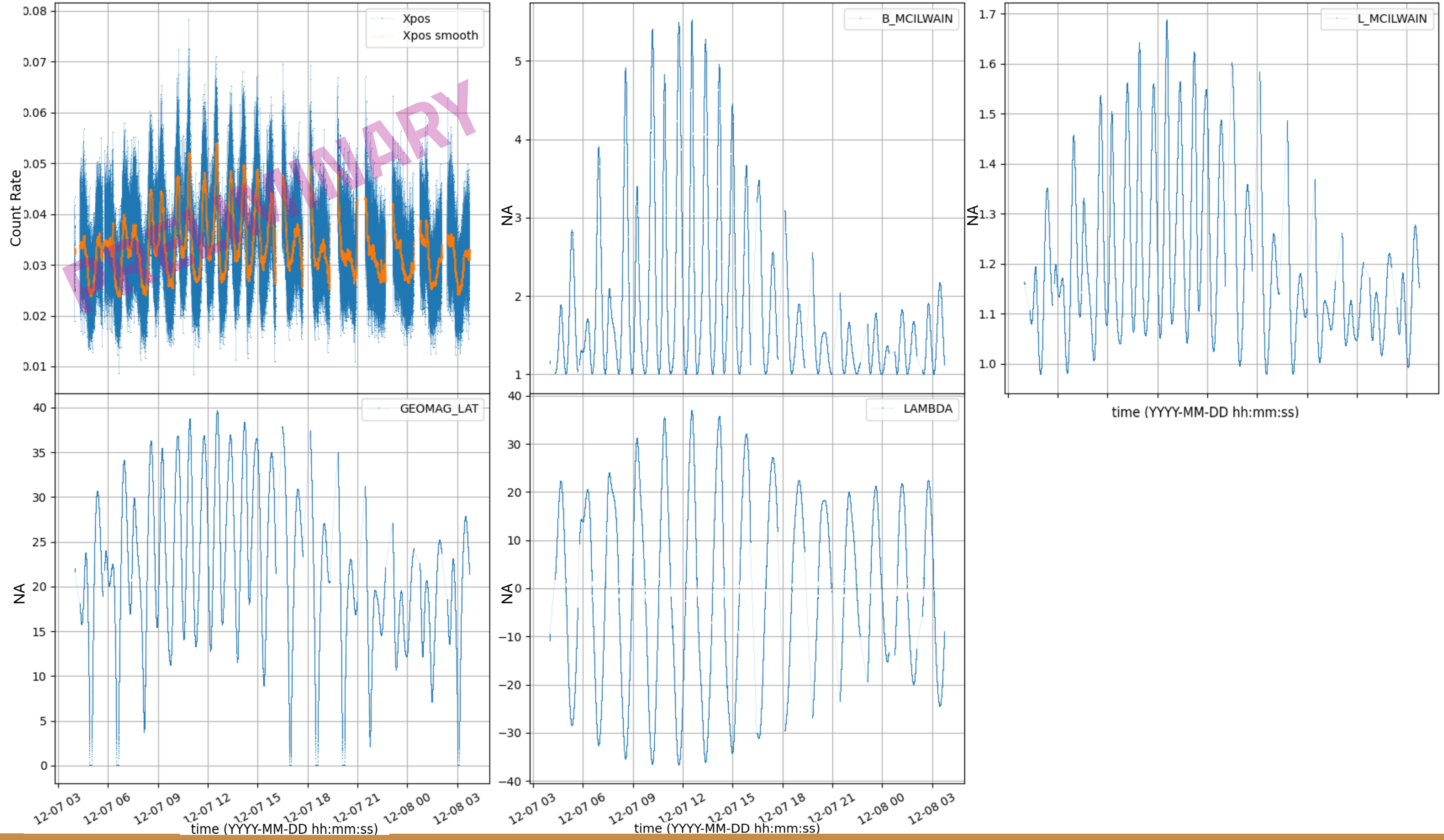
L_MCILWAIN (Earth_Radii) -> McIlwain L parameter, distance/shell value at START
'https://www.spenvis.oma.be/help/background/magfield/bl.html'

GEOMAG_LAT (deg) -> invariant geomagnetic latitude

LAMBDA (deg) -> effective geomagnetic latitude

IN_SAA -> whether spacecraft was in SAA

Spacecraft Data



Spacecraft Data

LAT_MODE -> Spacecraft GNC mode, where the 3 nominal modes are 3 (inertial point), 4 (Maneuver) and 5 (zenithpoint/survey). Other modes include 1 and 2 (capture and sunpoint - rarely used) and 6 and 7 (reentry modes).

LAT_CONFIG -> Flag for the configuration of the LAT (1 = nominal science configuration, 0 = not recommended for analysis)

DATA_QUAL -> Signed integer value indicating the quality of the LAT data

LIVETIME -> Accumulated livetime of the LAT during the interval from START to STOP

Spacecraft Data

QSJ_1 -> First component of SC attitude quaternion

QSJ_2 -> Second component of SC attitude quaternion

QSJ_3 -> Third component of SC attitude quaternion

QSJ_4 -> Fourth component of SC attitude quaternion

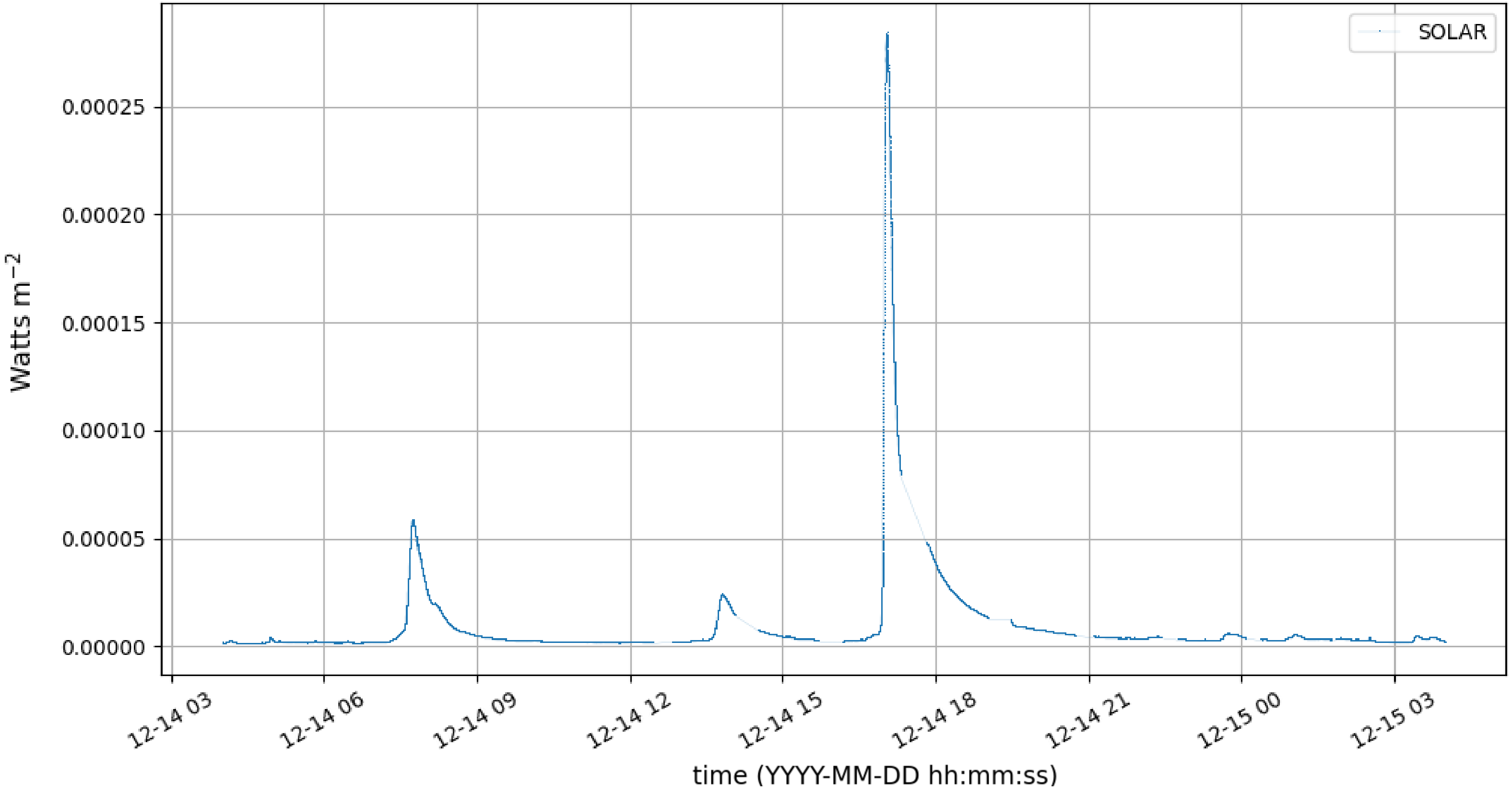
Spacecraft Data

RA_SUN (deg) -> RA of Sun

DEC_SUN (deg) -> DEC of Sun

Solar Activity Data

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



Solar Activity in the ACD count rate

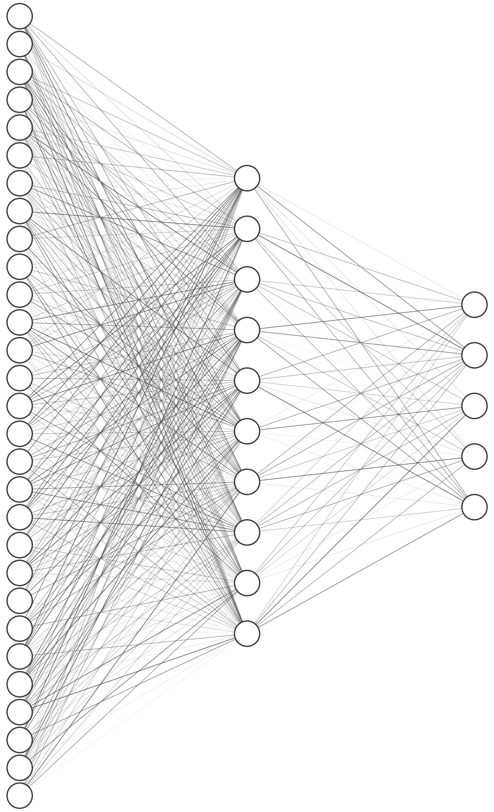
- -top (Z)
- -Xpos (X+)
- -Xneg (X-)
- -Ypos (Y+)
- -Yneg (Y-)

Dataset for the Neural Network

It is divided in 30 input parameters from Spacecraft data and 1 from Solar Activity data, and the signals from the 5 faces of the ACD.

Input parameters:

START
STOP
SC_POSITION
SC_VELOCITY
LAT_GEO
LON_GEO
…
SOLAR ACTIVITY



Output parameters:

top count rates
Xpos count rates
Xneg count rates
Ypos count rates
Yneg count rates

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

Technical Objectives, Methodologies and Solutions

- Our aim is to get a model of the count rate of the ACD signal given the geospatial configuration of the satellite and where the solar activity is also part of the background (i.e. it's not an anomaly).

We do this with the help of Machine Learning techniques:

- K-Nearest Neighbors

- Feed Forward Neural Network

Timescale and KPIs

- Hired in September 2023;

- October to December, study of scientific literature;

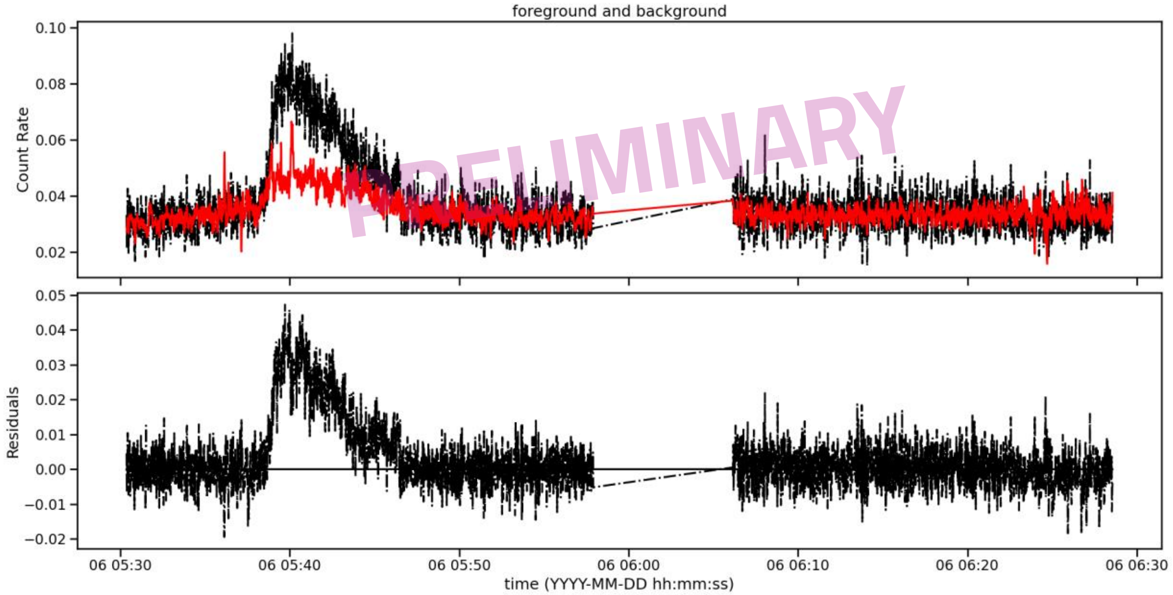
- January to April, development of code to prepare the dataset and preliminary algorithm to fit the signal.

Code available on github.

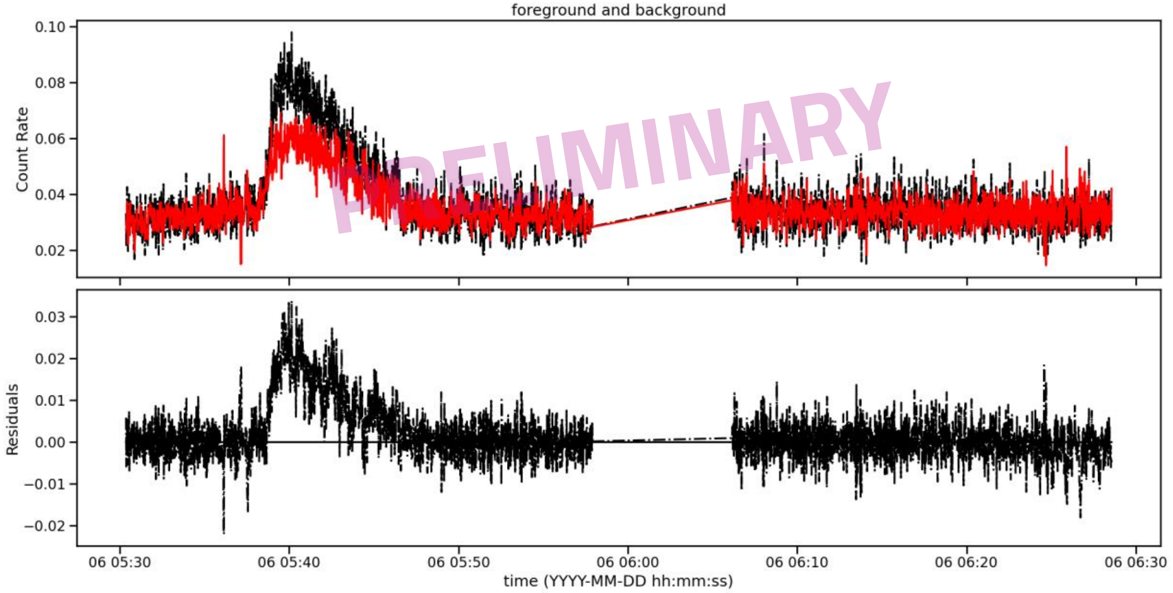
K-Nearest Neighbors

Via KNeighborsRegressor from scikit
We tried with k = 2, 5, 20

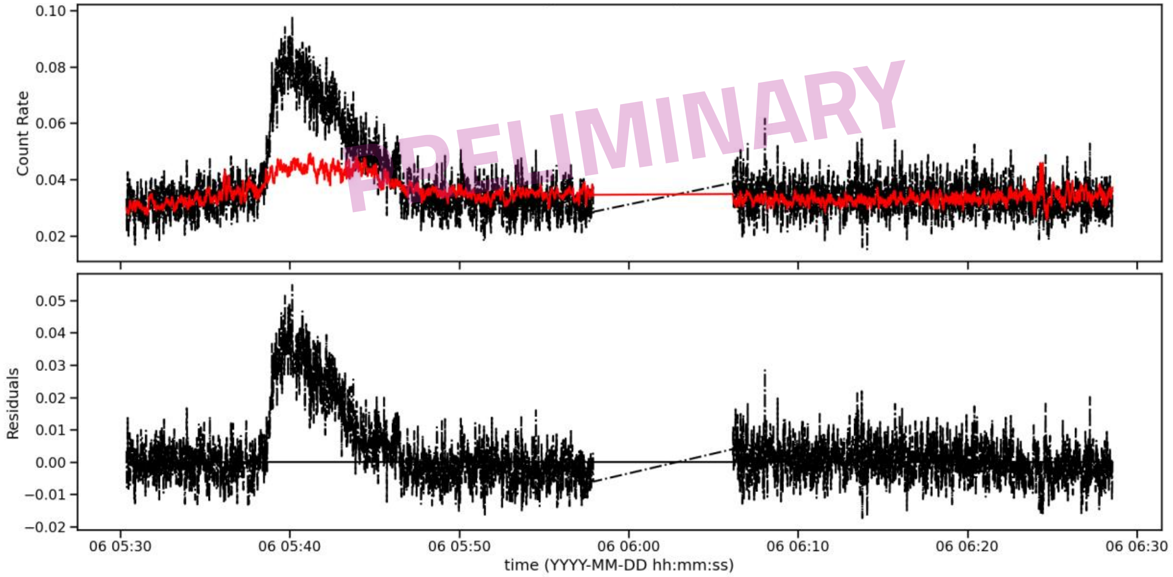
5



2

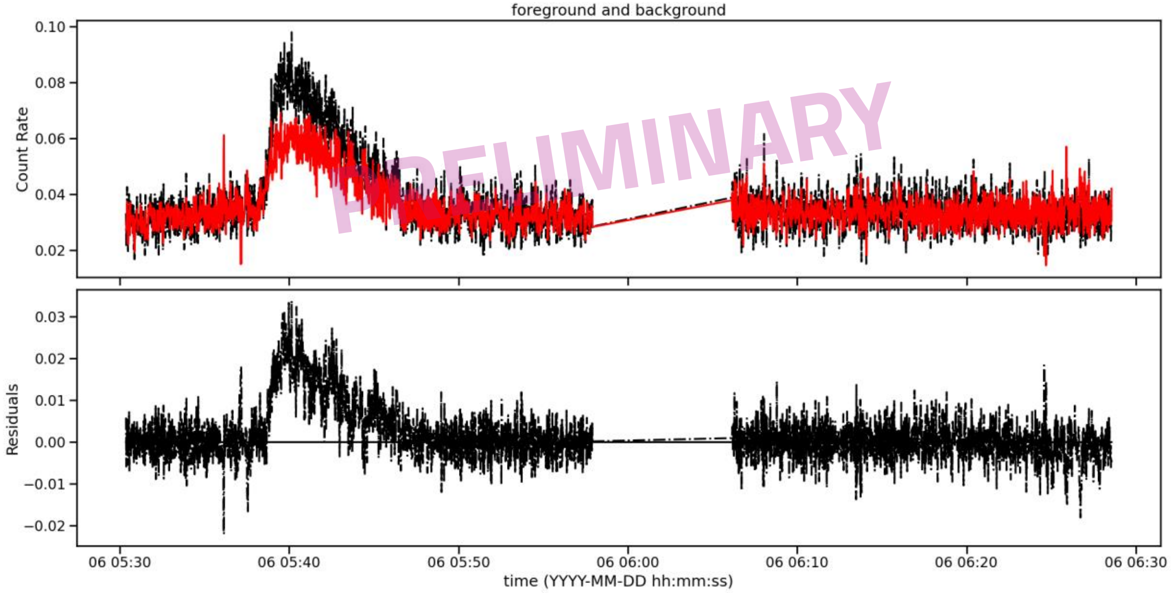


20



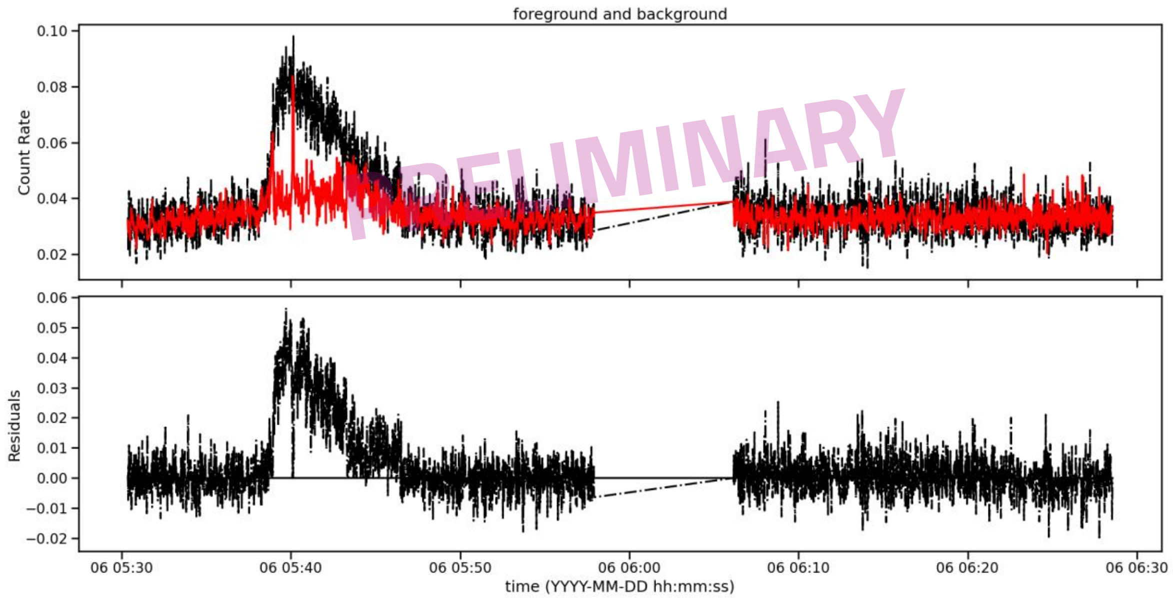
K-Nearest Neighbors

Via KNeighborsRegressor from scikit
We tried with k = 2, 5, 20

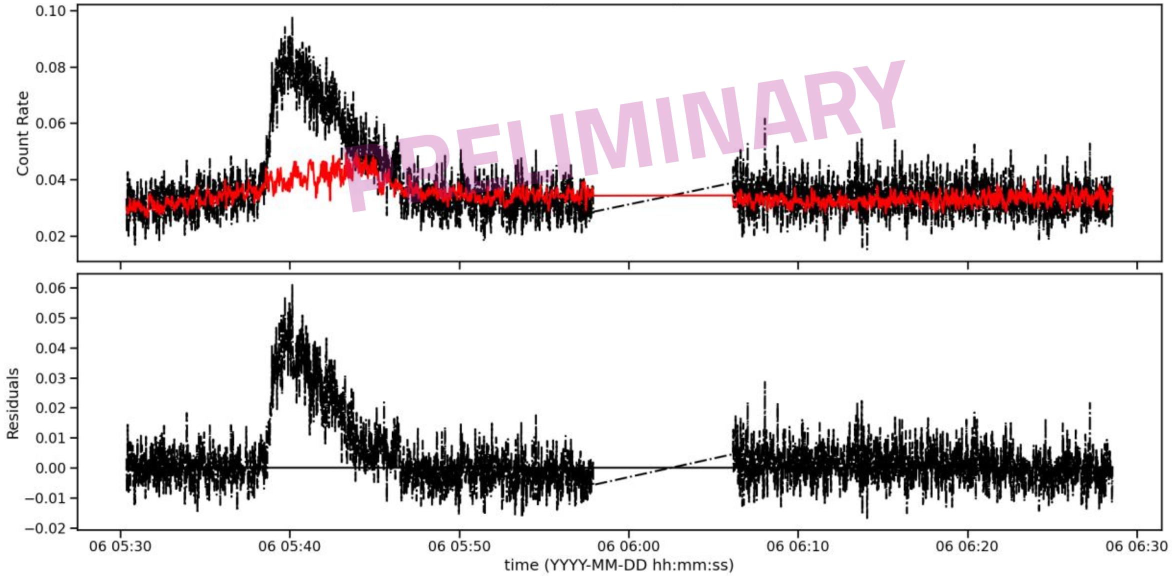


2

5



20



... changing from mean to median when calculating the best y between the k neighbors.

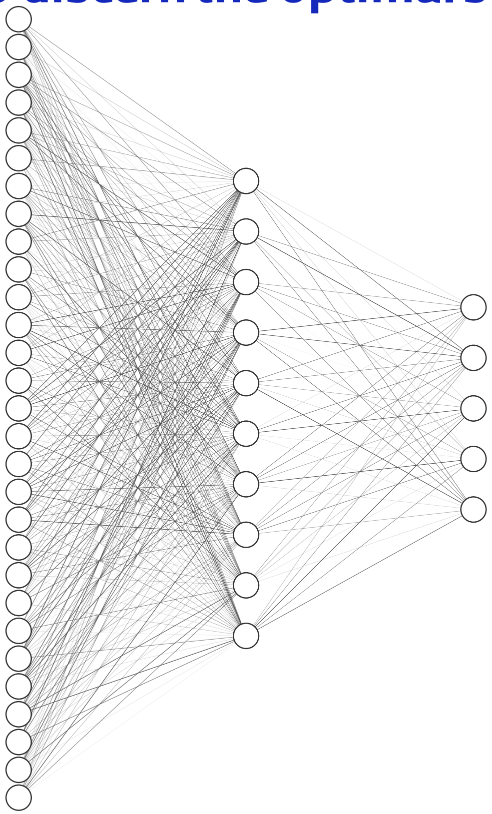
Feed Forward Neural Network

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:

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

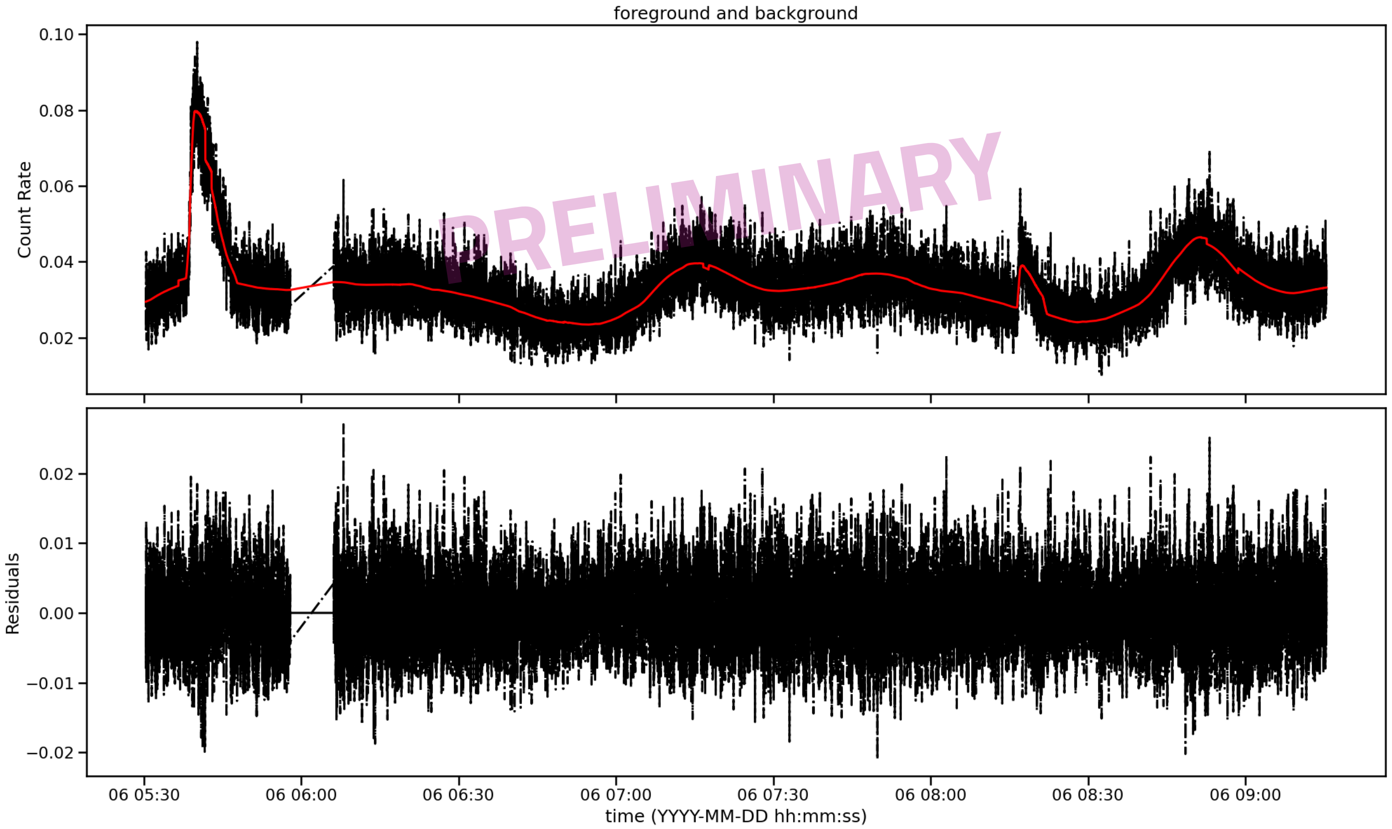
NN Training and results

The best models appear to be the three layers models without Normalization and without Dropout.

	model_id	units_1	units_2	units_3	norm	drop	...	loss_type	top	Xpos	Xneg	Ypos	Yneg
92	92	50	90	70	0	0	...	mae	0.002998	0.004502	0.004454	0.004312	0.004225
96	96	50	90	90	0	0	...	mae	0.003004	0.004509	0.004462	0.004321	0.004232
108	108	90	50	50	0	0	...	mae	0.002993	0.004507	0.004442	0.004300	0.004217
124	124	90	90	30	0	0	...	mae	0.003024	0.004504	0.004466	0.004317	0.004229
132	132	90	90	70	0	0	...	mae	0.002998	0.004499	0.004454	0.004312	0.004224

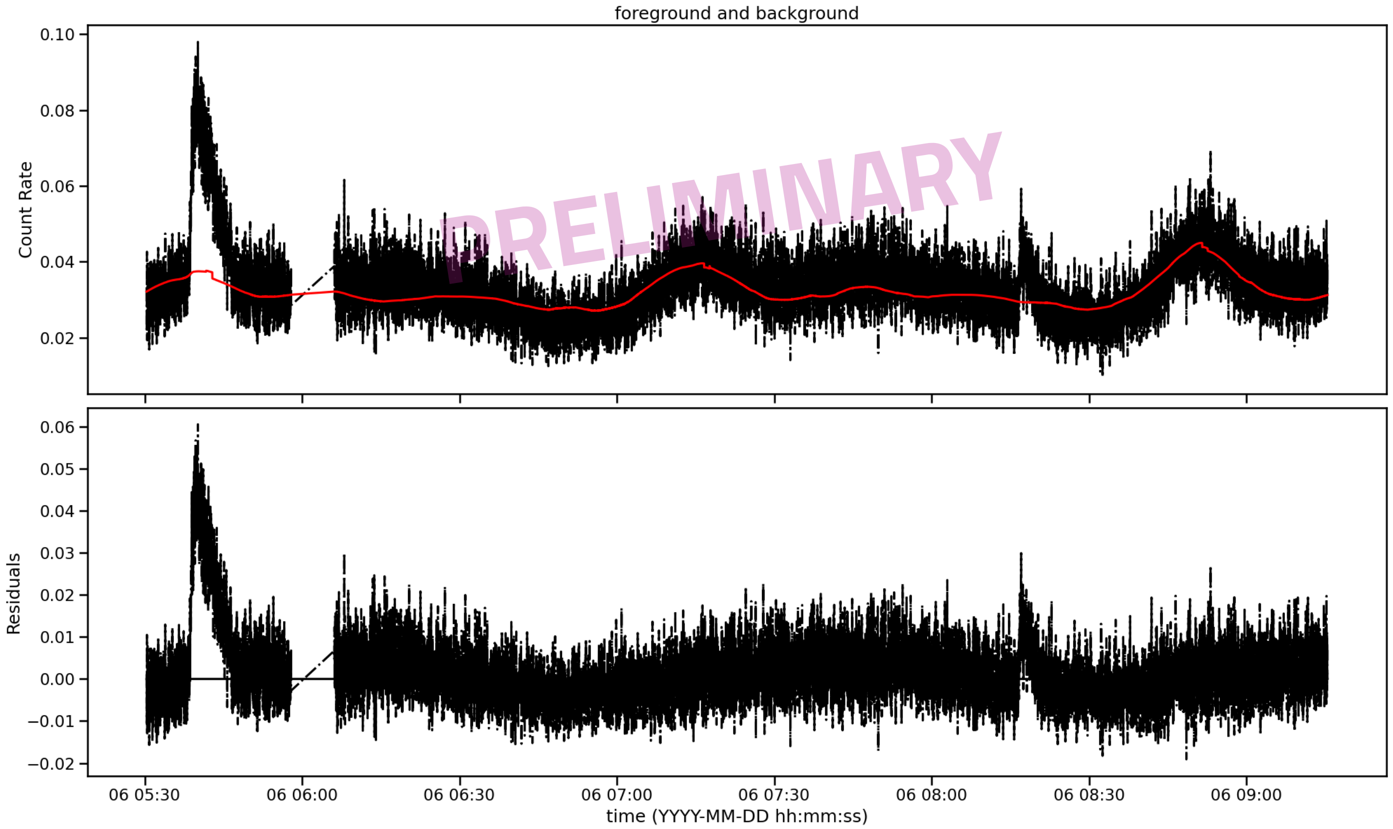
Accomplished work, Results

The best models appear to be the three layers models without Normalization and without Dropout. In particular, the best model gives this fit:



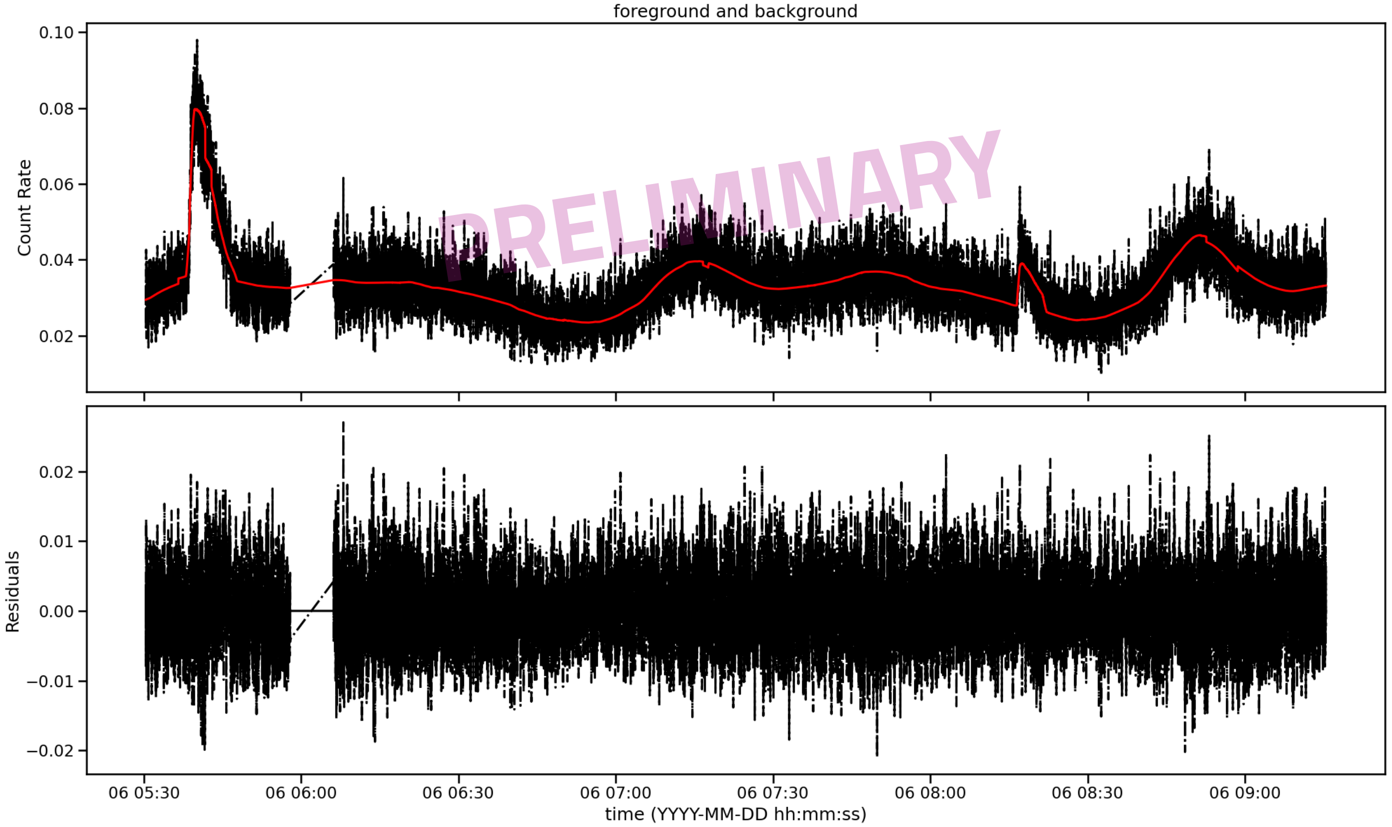
Accomplished work, Results

The FFNN was trained without solar activity in the dataset:



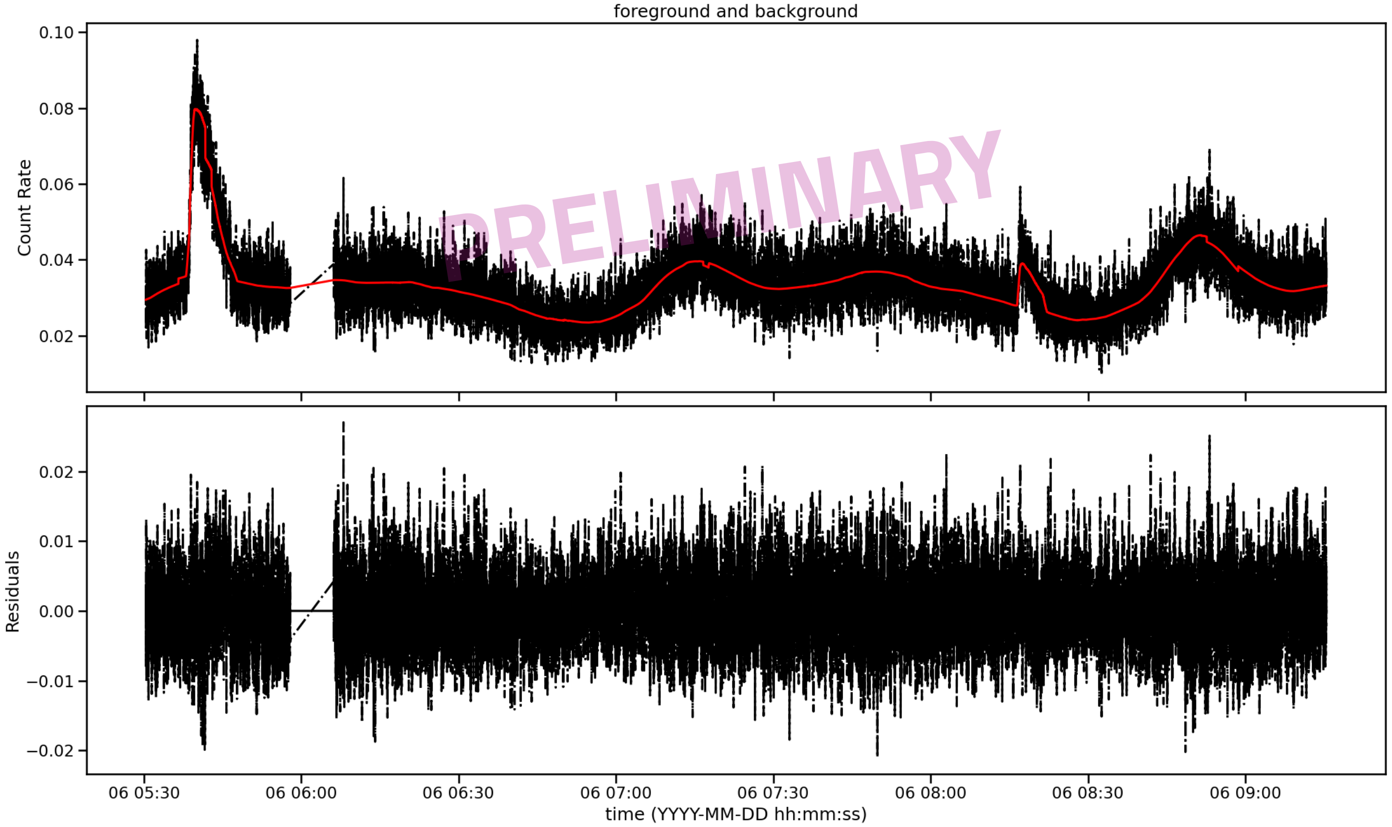
Next Steps

We will continue this analysis with more data (more than two months) and we will consider more complicated models, such as Recurrent Neural Networks.



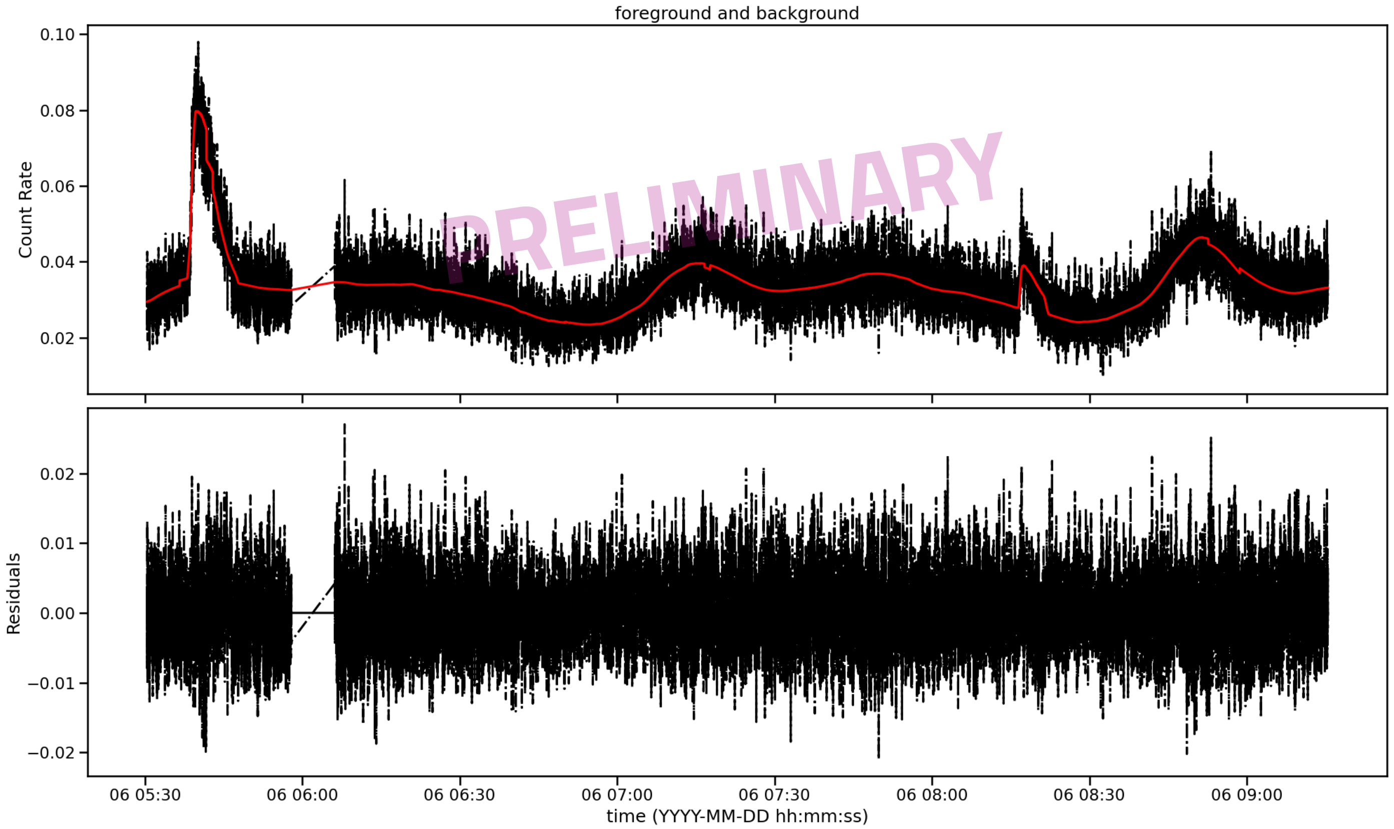
Next Steps

Also, the addition of new input parameters, such as a description of the flux of cosmic rays, could help to better assess the background signal in case of excess of cosmic rays.



Next Steps and Expected Results

Finally, we will implement a trigger algorithm to automate the identification of significant anomalies



Thank you

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

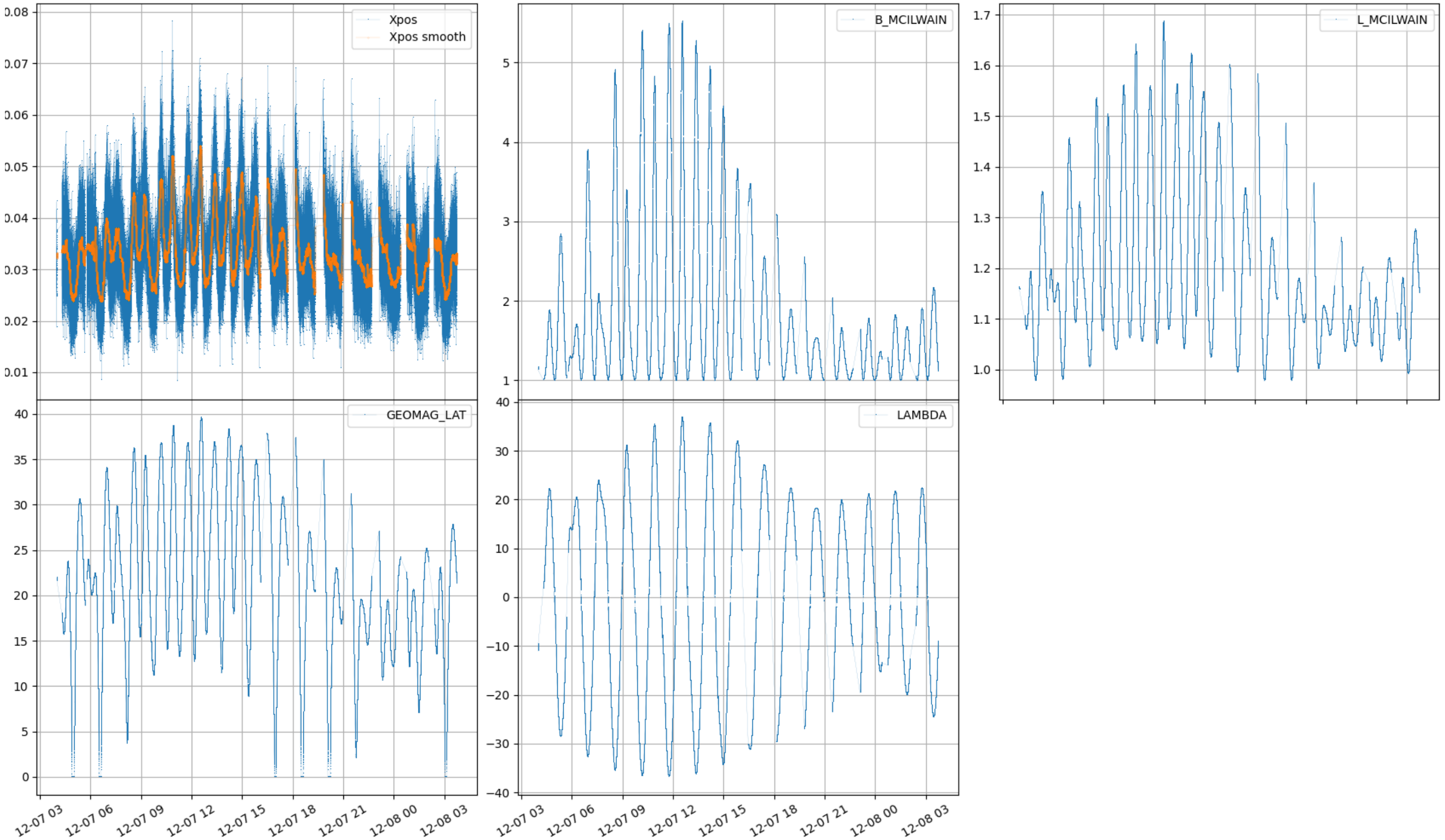
(2) Meegan 2009 - THE FERMI GAMMA-RAY BURST MONITOR

(3) Crupi, R., Dilillo, G., Bissaldi, E. et al. Searching for long faint astronomical high energy transients: a data driven approach

BACKUP

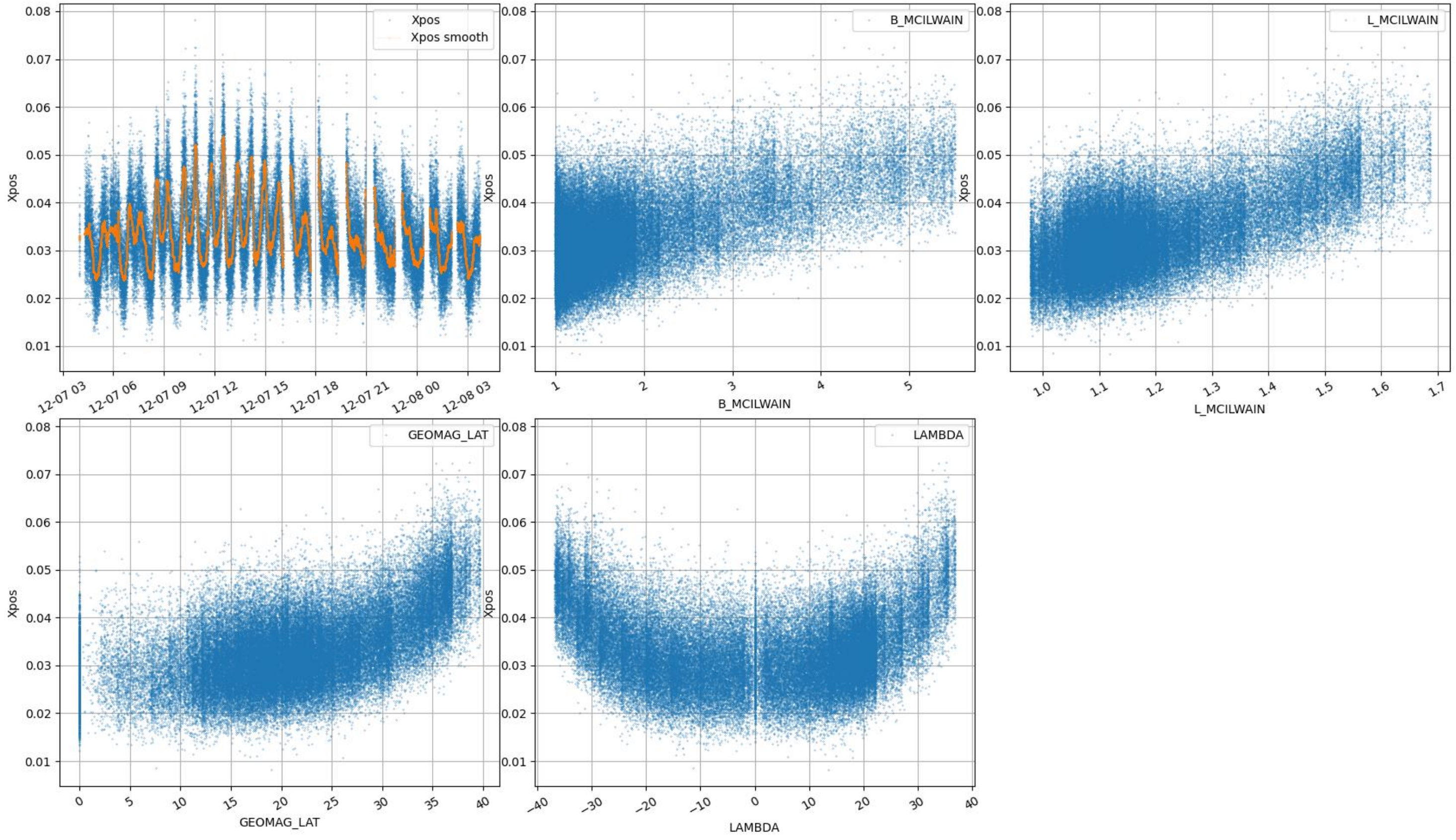
Dataset for the Neural Network

From the data, we can observe that some parameters have a significant influence on the number of photons observed:



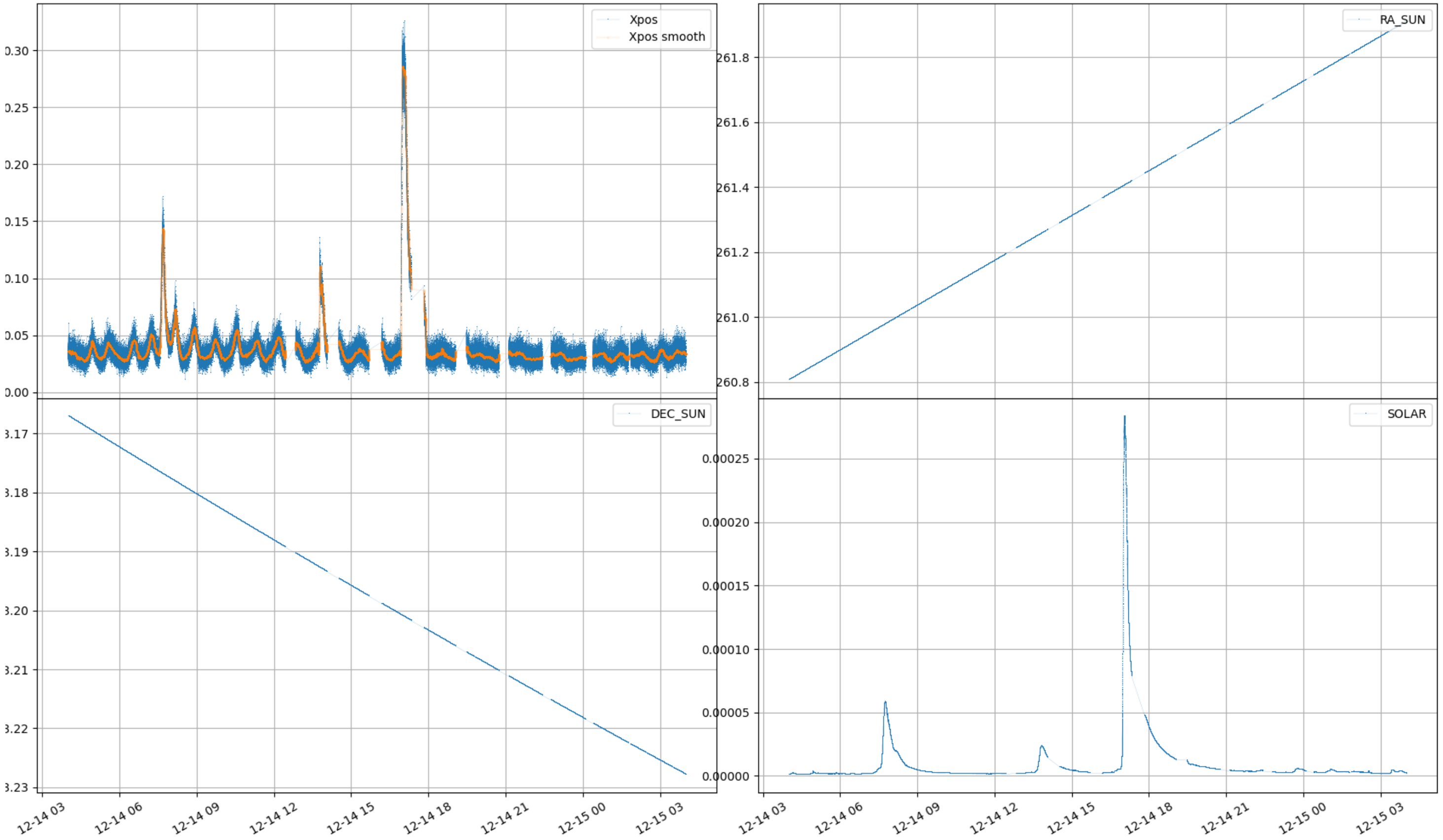
Dataset for the Neural Network

From the data, we can observe that some parameters have a significant influence on the number of photons observed.



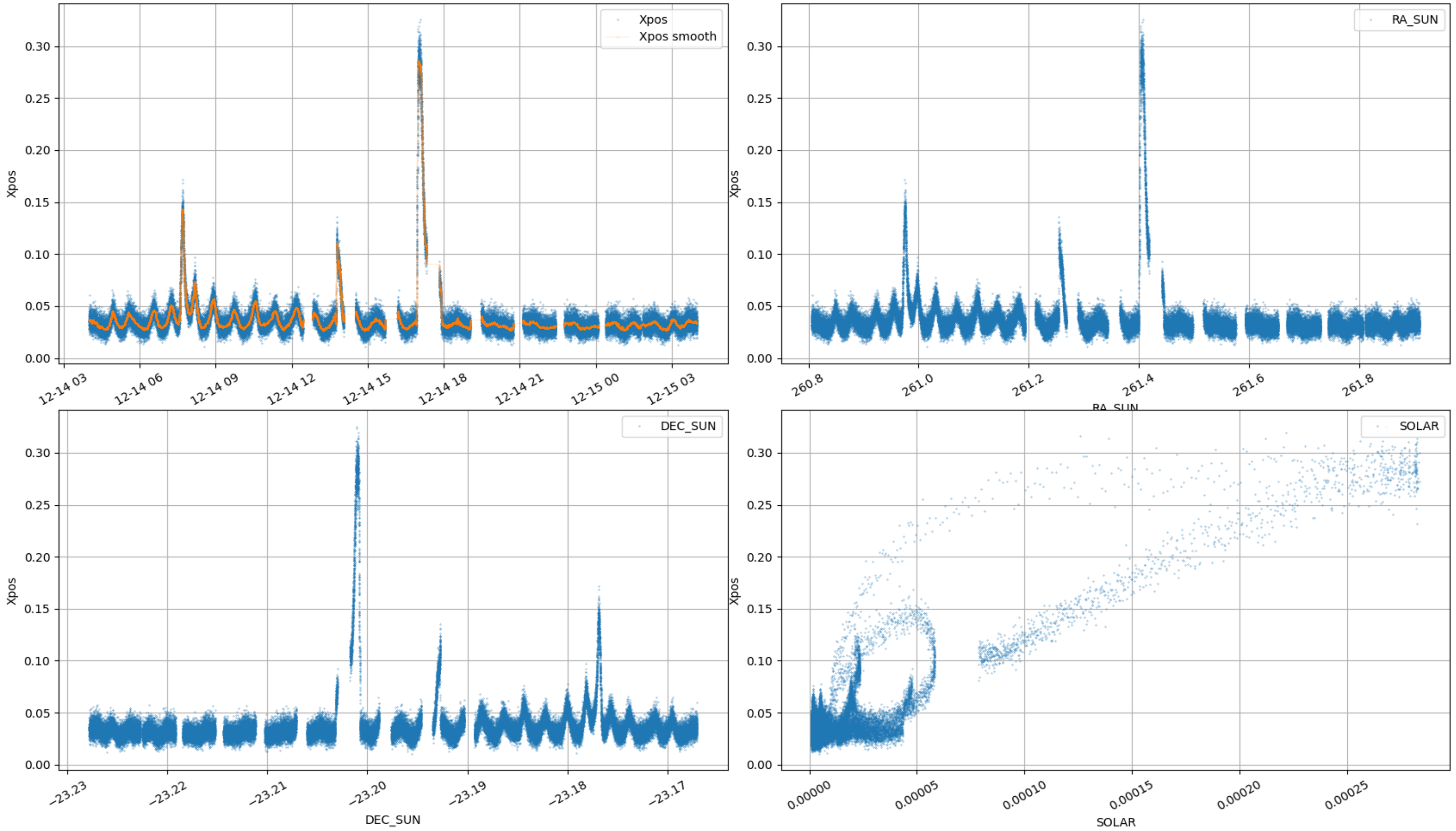
Dataset for the Neural Network

From the data, we can observe that some parameters have a significant influence on the number of photons observed:



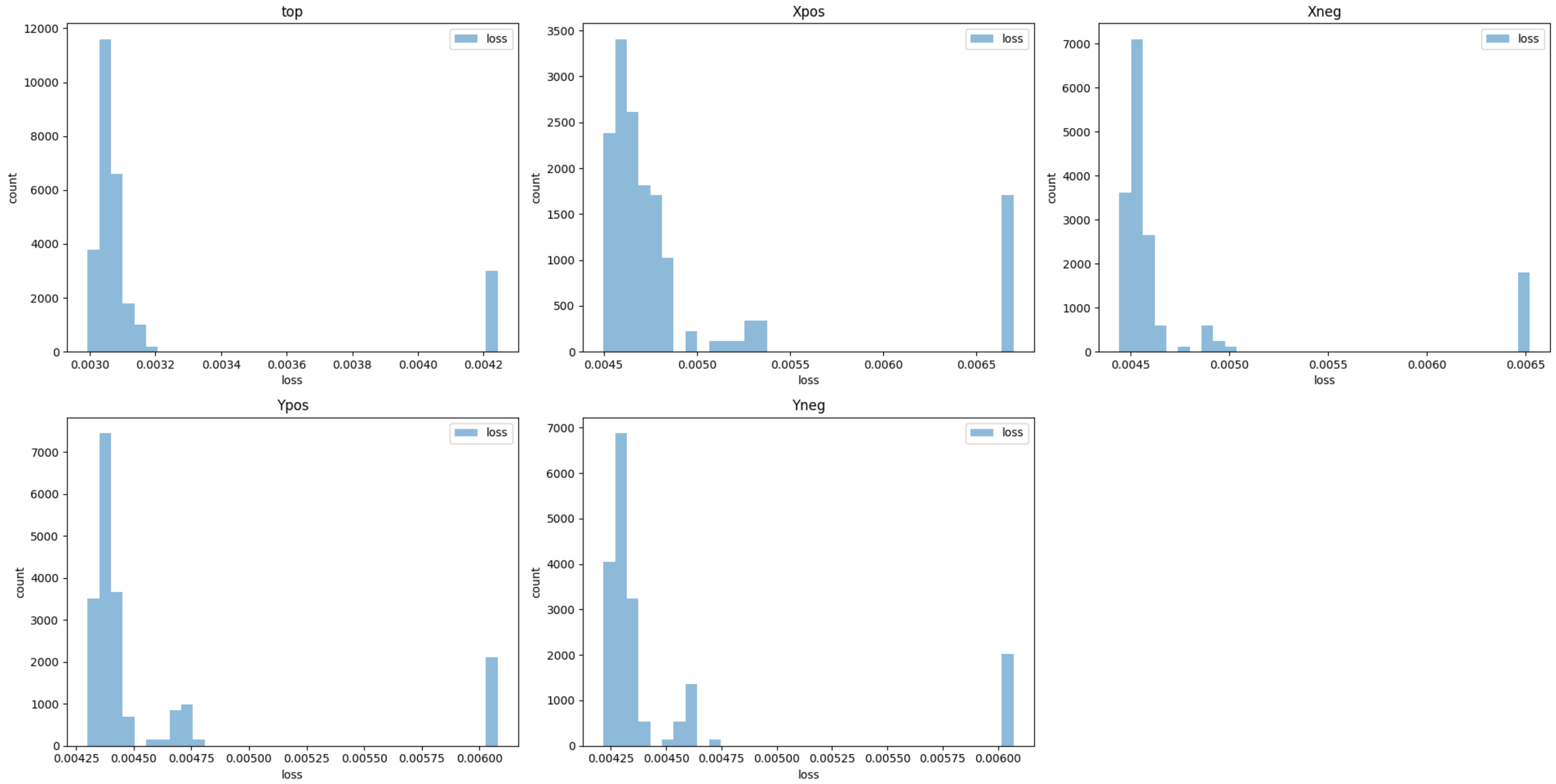
Dataset for the Neural Network

From the data, we can observe that some parameters have a significant influence on the number of photons observed:



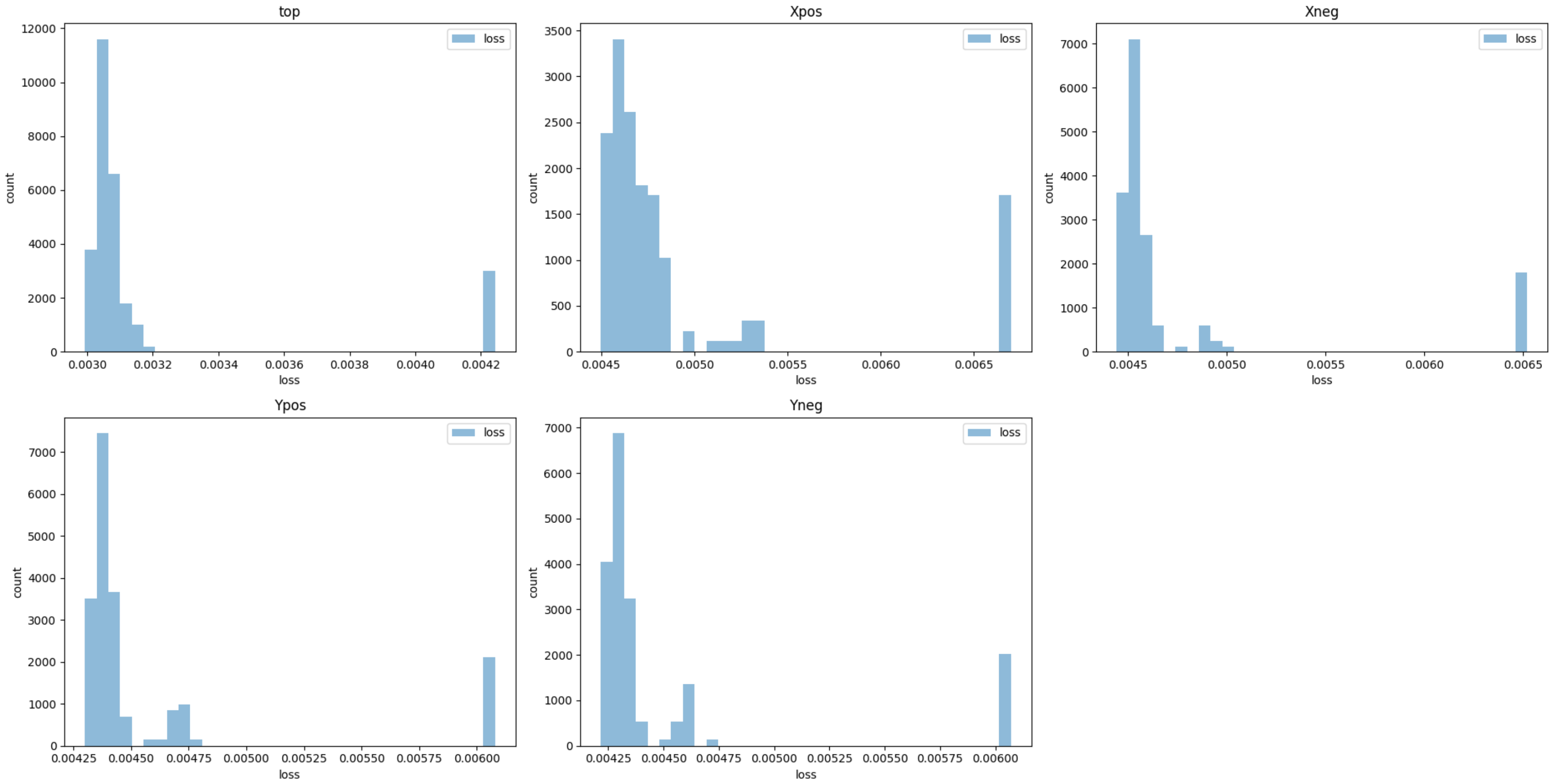
NN Training and results

Each combination of these layers has been trained for 60 epochs.



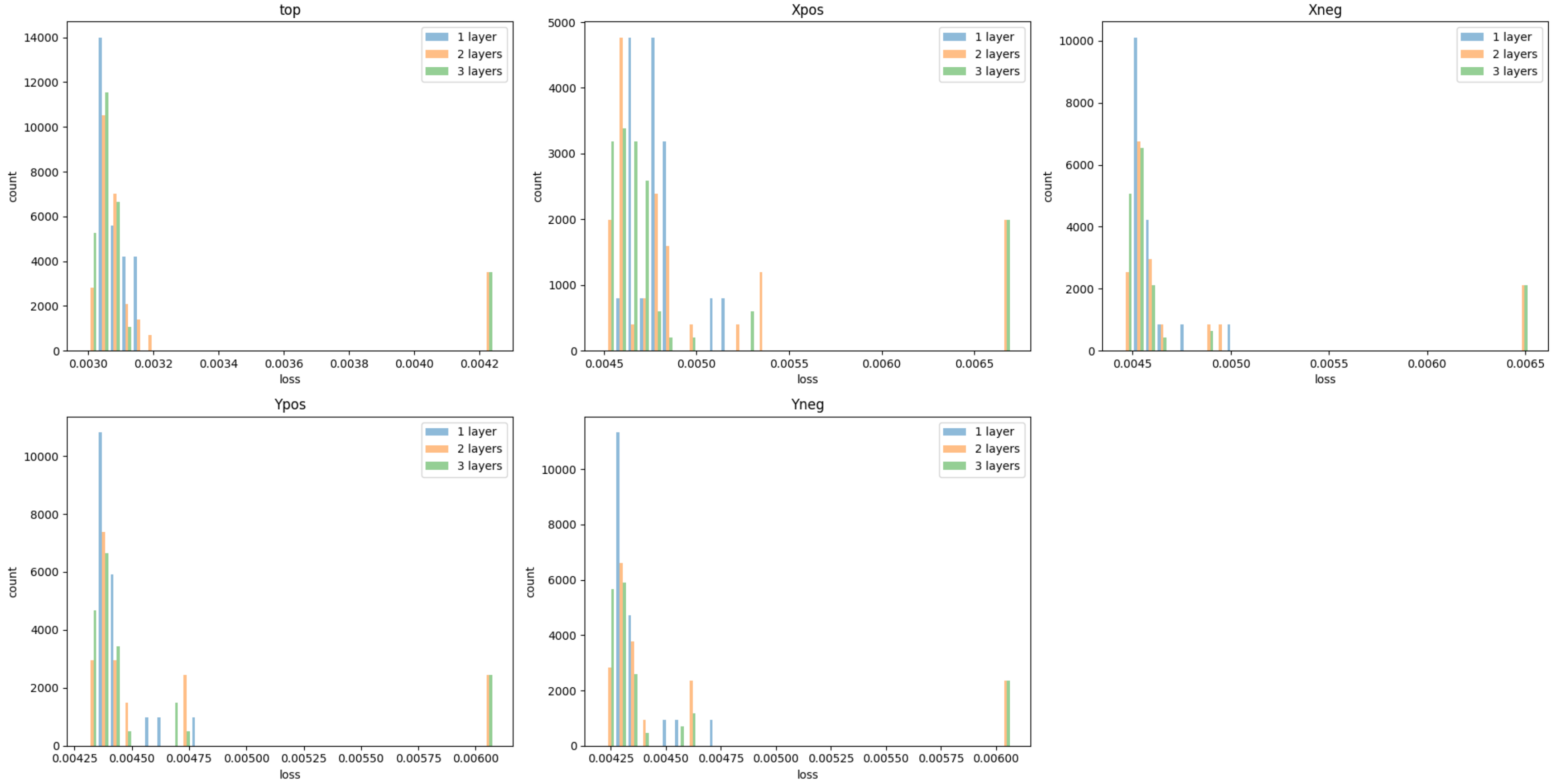
NN Training and results

These histograms have been subdivided based on the number of layers.



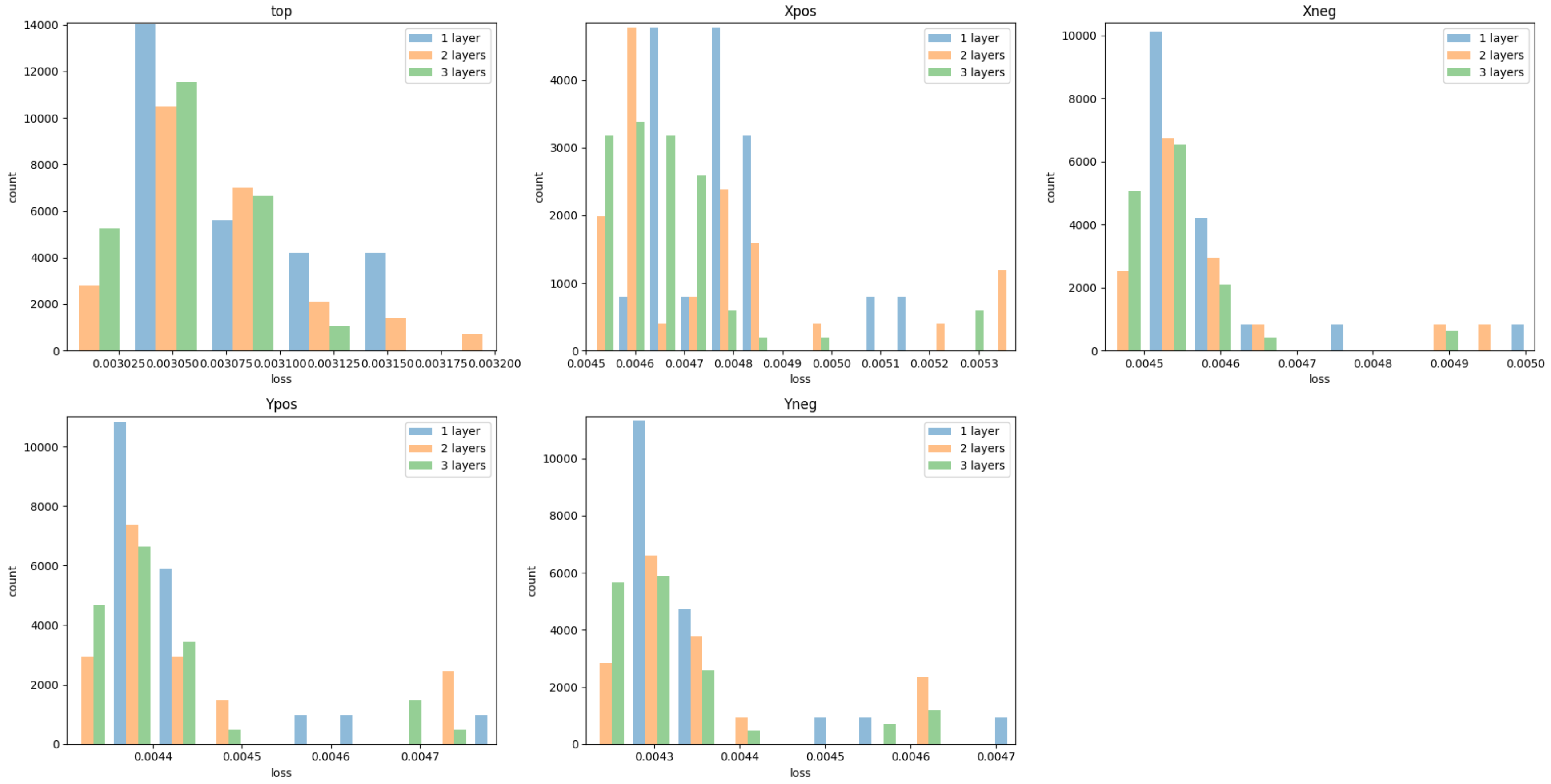
NN Training and results

These histograms have been subdivided based on the number of layers.



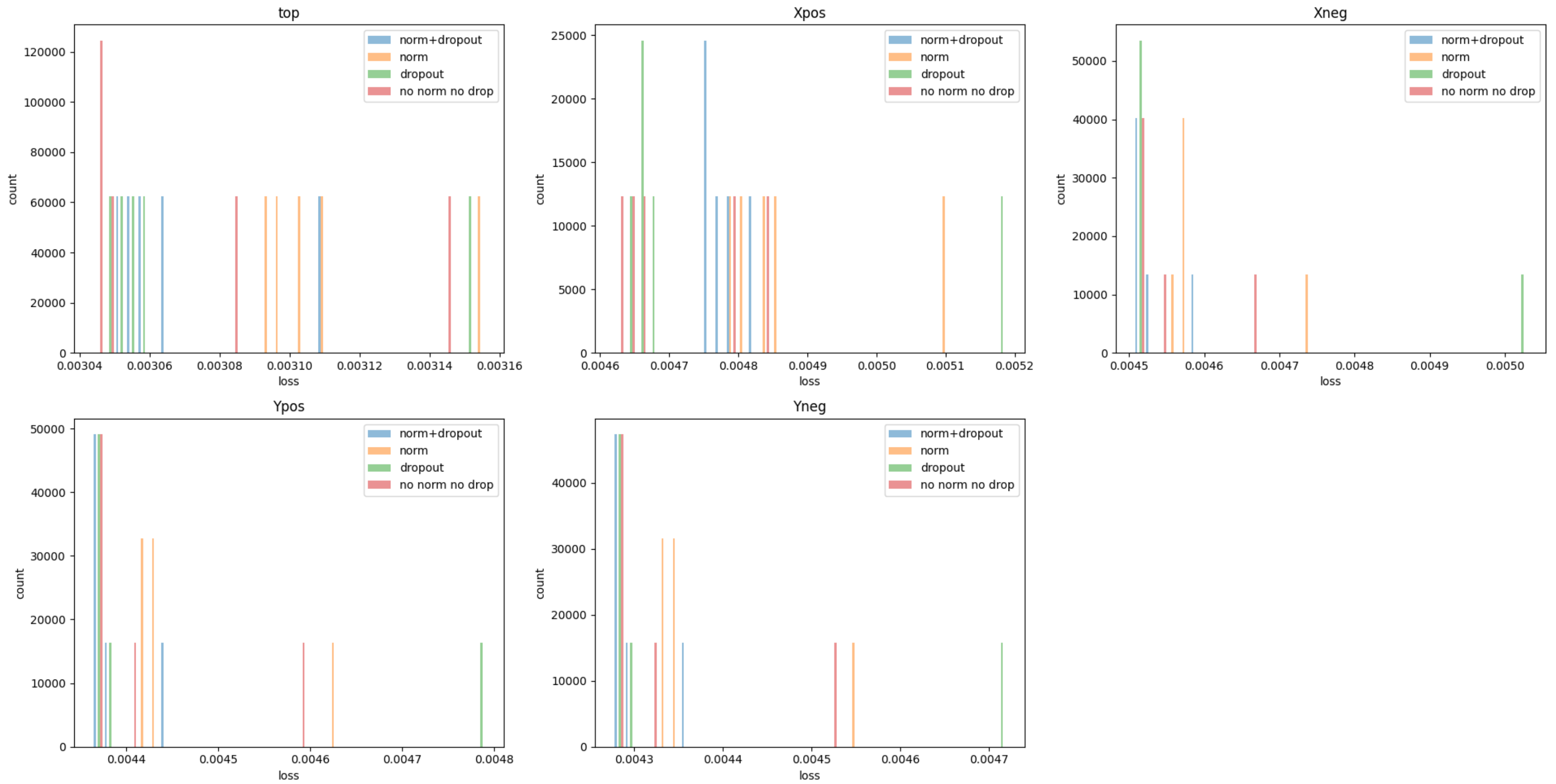
NN Training and results

These histograms have been subdivided based on the number of layers.



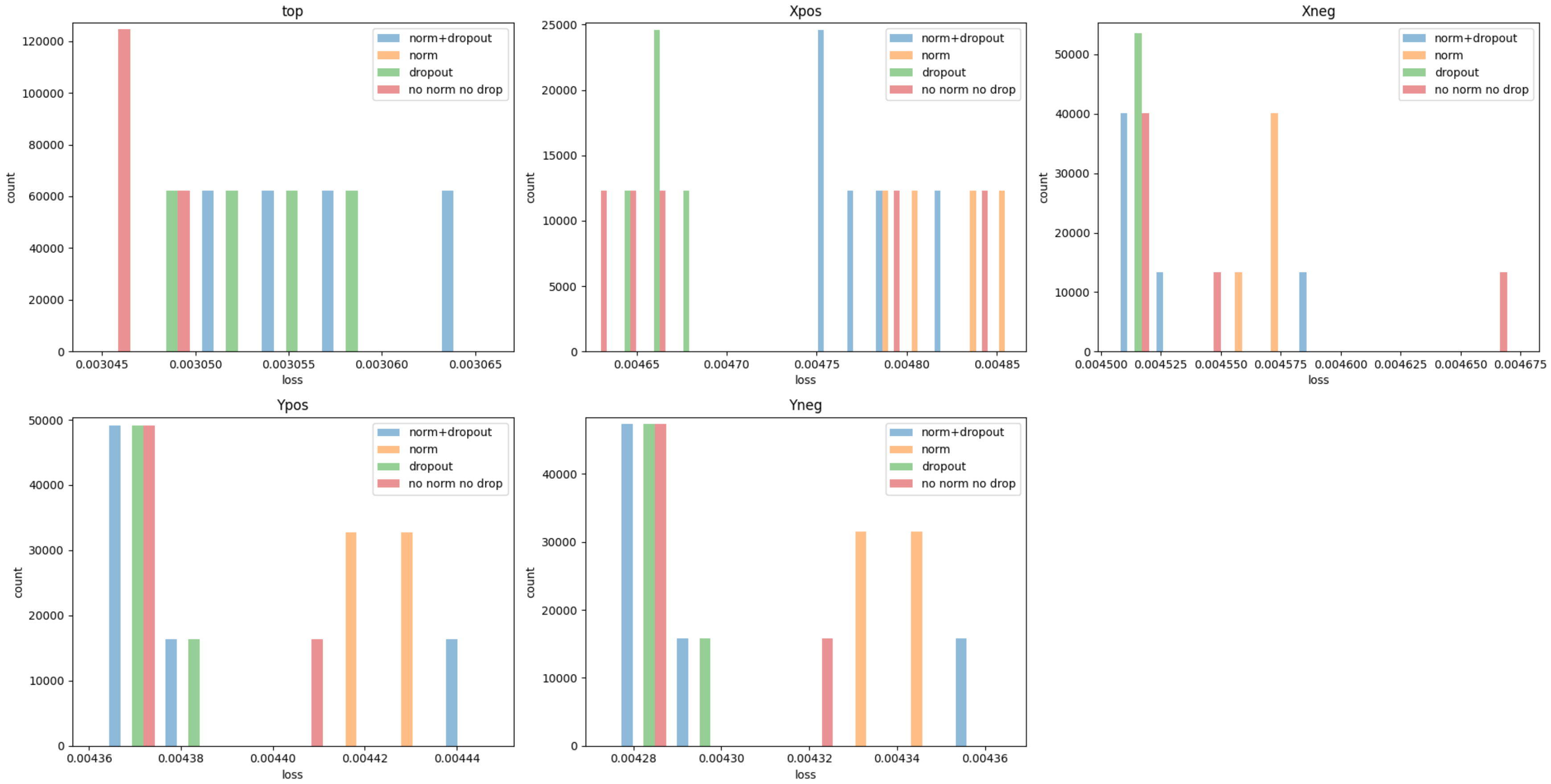
NN Training and results

The same histograms have been plotted only for models with 1 layer



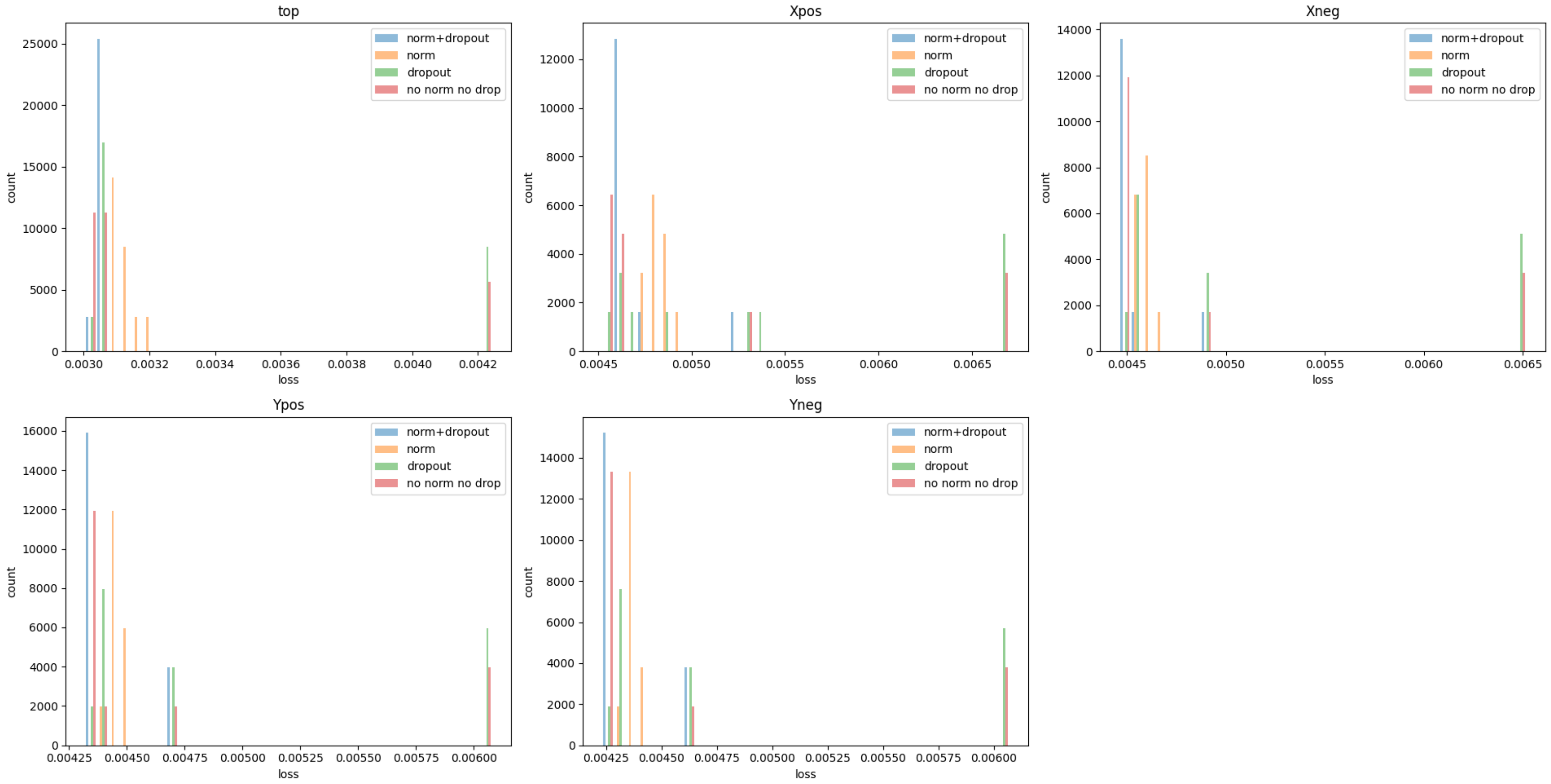
NN Training and results

The same histograms have been plotted only for models with 1 layer



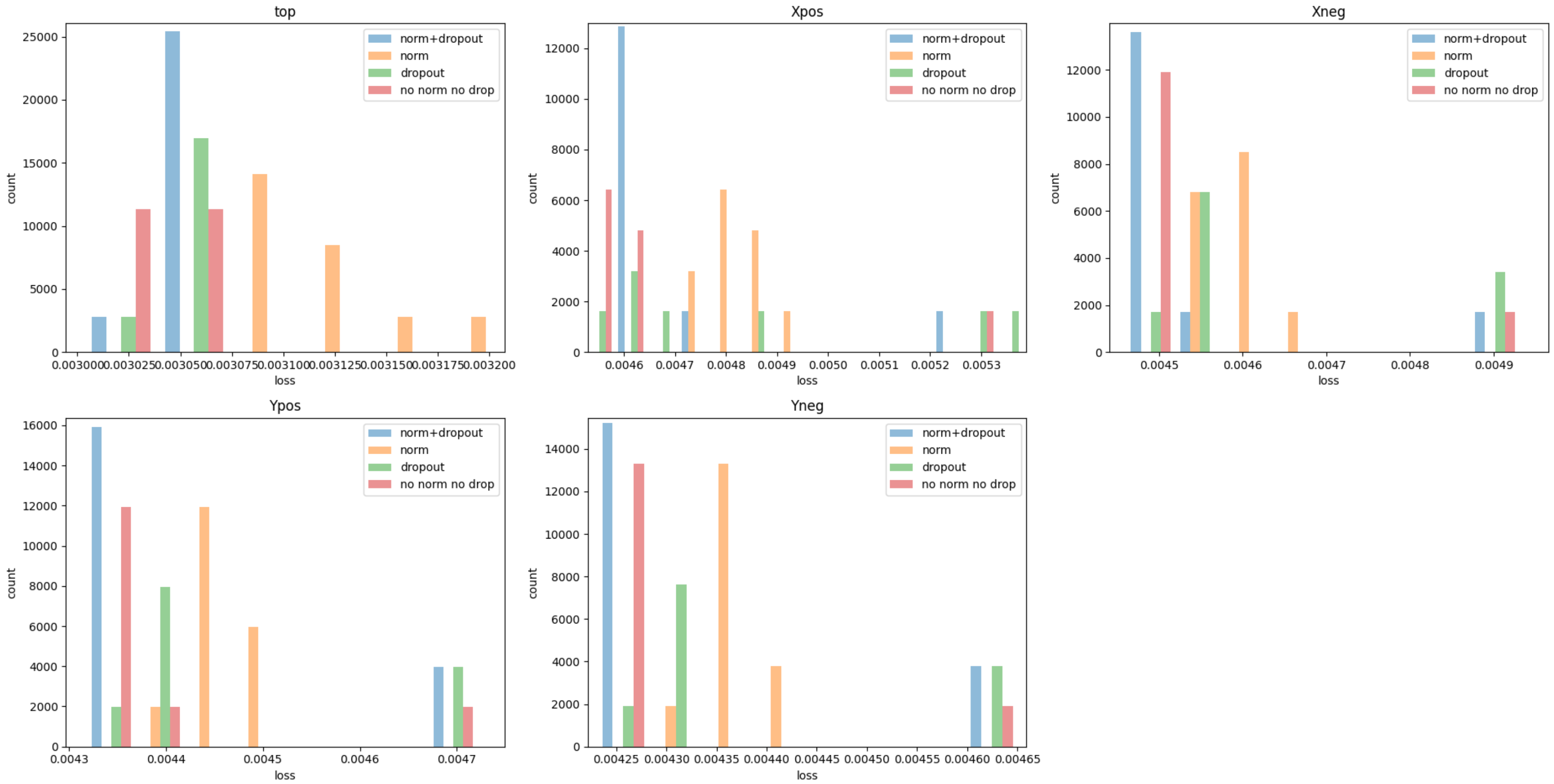
NN Training and results

The same histograms have been plotted only for models with 2 layer



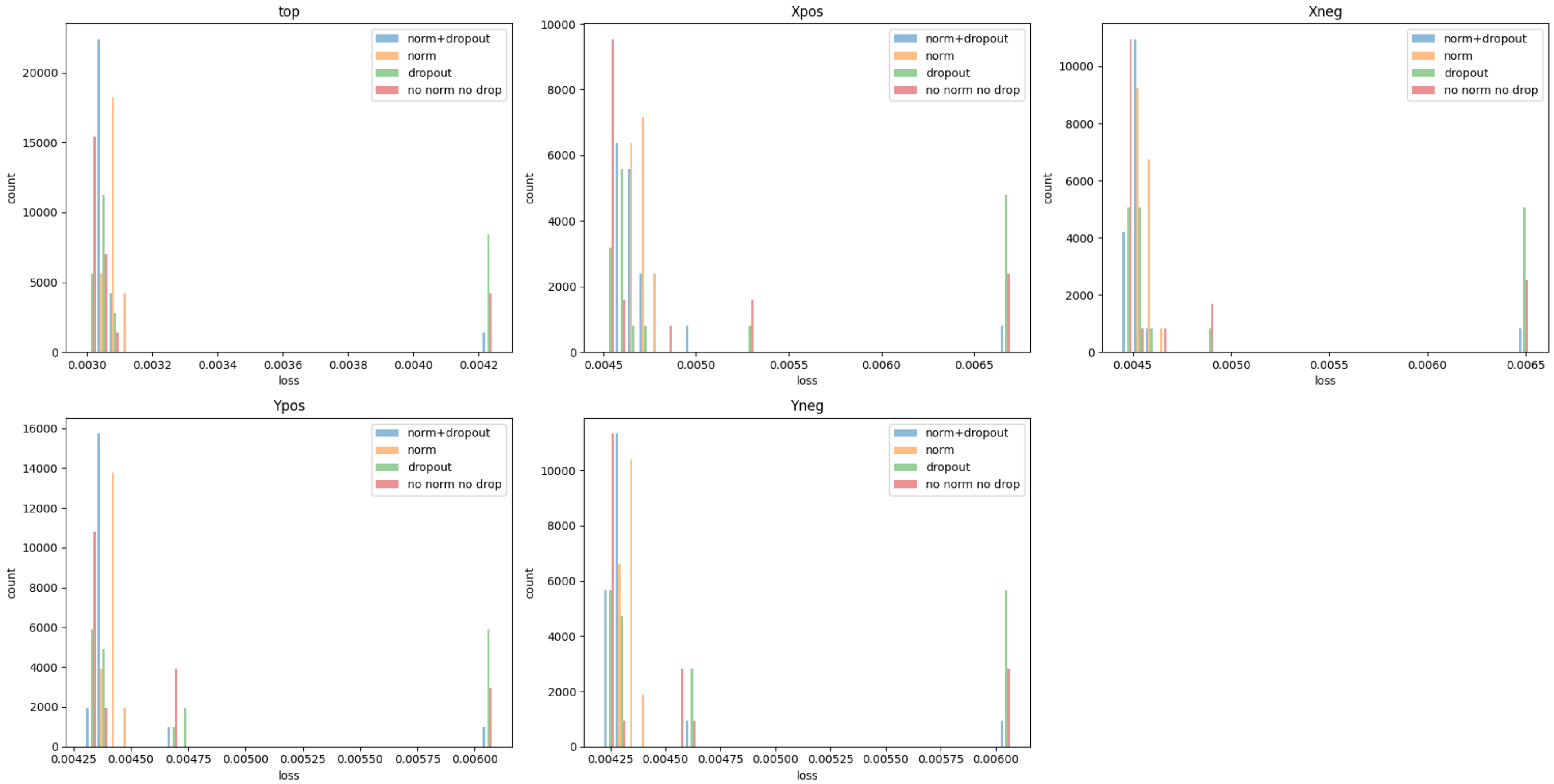
NN Training and results

The same histograms have been plotted only for models with 2 layer



NN Training and results

The same histograms have been plotted only for models with 3 layer



NN Training and results

The same histograms have been plotted only for models with 3 layer