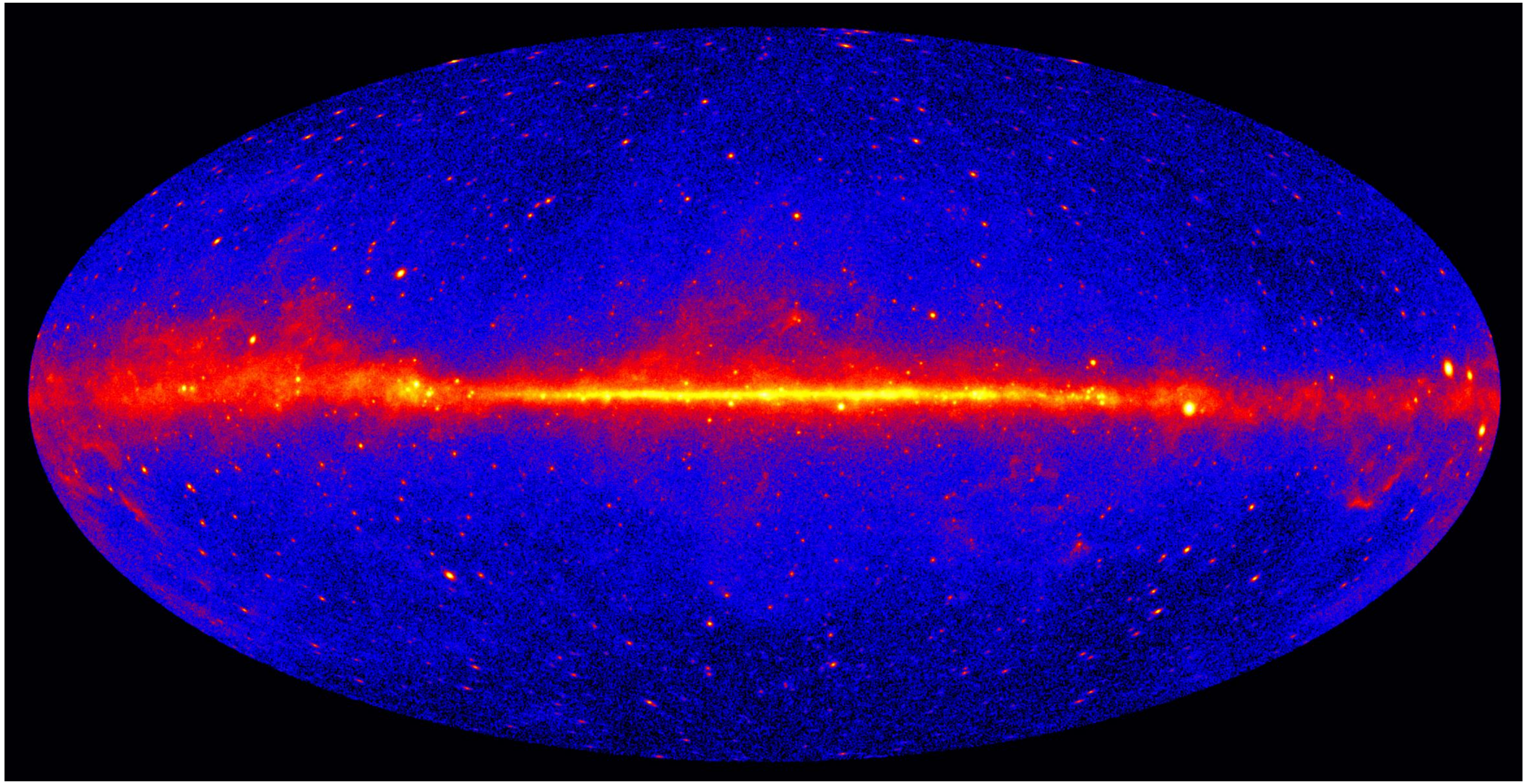
Using Deep Learning to Search for Fermi-LAT Point Sources

*S. Bhattacharyya**, C. Oetelaar , R. Austri, G. Johanesson*, S. Caron, G. Zaharijas, B. Panes

Results Shown in the LAT Catalog Meeting (June, 2022)

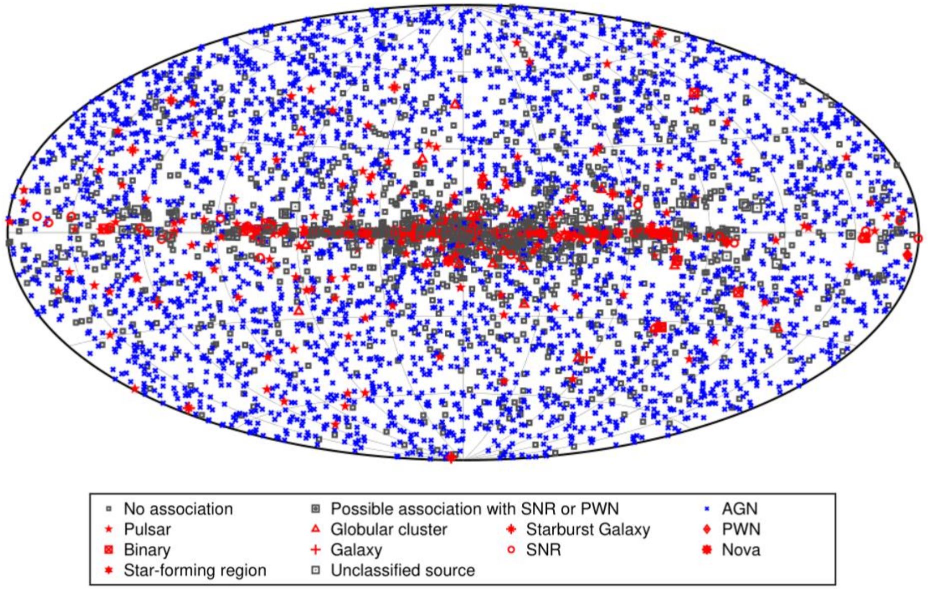
21.07.2022

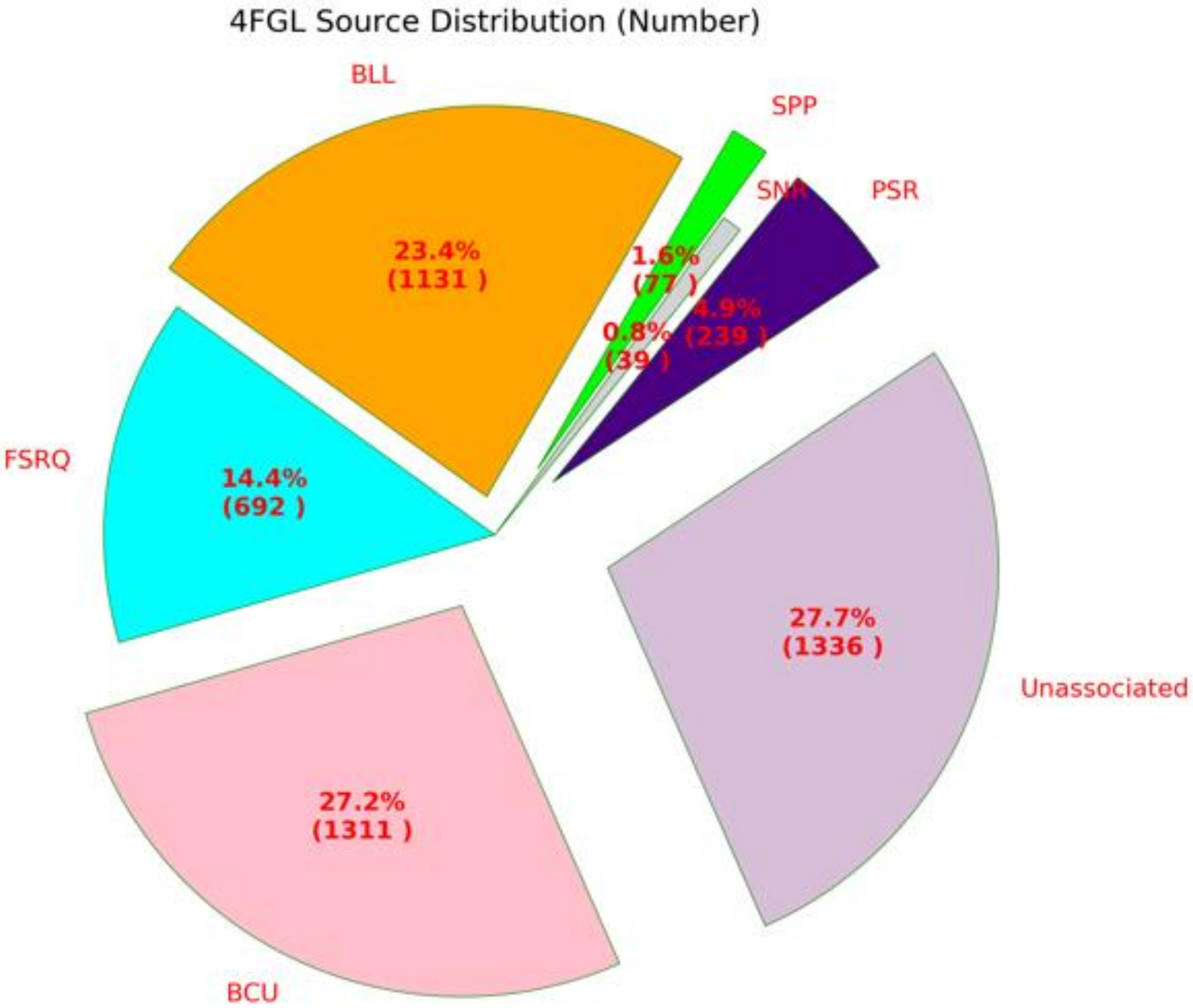
Detect & Localize Point Sources



Objective: Given such a γ-ray map, can a neural network detect and find the precise location of point sources?

4FGL Catalog; 8 years of Data, 5064 Sources





Ref: Fermi-LAT, 4FGL
ApJS 247, 33 (2020)

4FGL-DR2: Brief Info

- Incremental version of 4FGL catalog.
- Based on 10 years of data, ranging from 50 MeV to 1 TeV.
- Data analysis scheme is identical to 4FGL.
- Dataset consists of > 5700 sources. More than 3200 identified/associated sources are of active galaxies of 'Blazar' class and ~250 are pulsars.
- Sources are tested with 3 types of spectral models
    - Power-Law (PL)
    - Log Parabola (LP)
    - Power-law with exponential cut-off (PLEC)

[4FGL-DR2; 2005.11208]

Training Data Generation: Supervised Machine Learning

- To learn a mapping from input to output based on example input-output pairs. 'Supervised Learning'
- Create set of full sky simulations (sky-maps) with source properties based on the distribution in 4FGL-DR2.
  - Include BLLacs, FSRQs, PWN/SPP/SNR (LP distribution) and PSR (PLEC distribution).
- Consider yearly photon data over the 10 year period [2008-2018].
- The full data analysis pipeline is a two step process. Localization and Classification.
  - Use variability of the blazars as another information. For Classification purpose only.

Training Data Generation: Using Full Detector Potential

- To generate robust training data we exploit fully the detector potential.
  - Treat front and back γ-ray events separately with appropriate IRFs.
    - Photons that convert in front section have better angular resolution.
  - Bin the photon counts in 6 different energy bins starting from 300 MeV to 1 TeV.

  - The spatial resolution of the sky-maps increases with increasing energy
    - Per-photon angular resolution $\sim 5^{\circ}$ at 100 MeV, improving to 0.8° at 1 GeV and $0.1^{\circ} \geq 20$ GeV [LAT].

Robust Training Data

- We are using raw photon data to precisely localize and classify point sources using deep learning.
  - 'Proof of Principle': already published using a simpler dataset (AutoSourceID: A&A, arxiv: 2103.11068).
  - 2 Source classes (AGN, PSR) and same resolution for all energy bins.

- Develop a robust data analysis pipeline:
  - Will help us to understand source detection possibility using our method by comparing with DR2 Catalog.

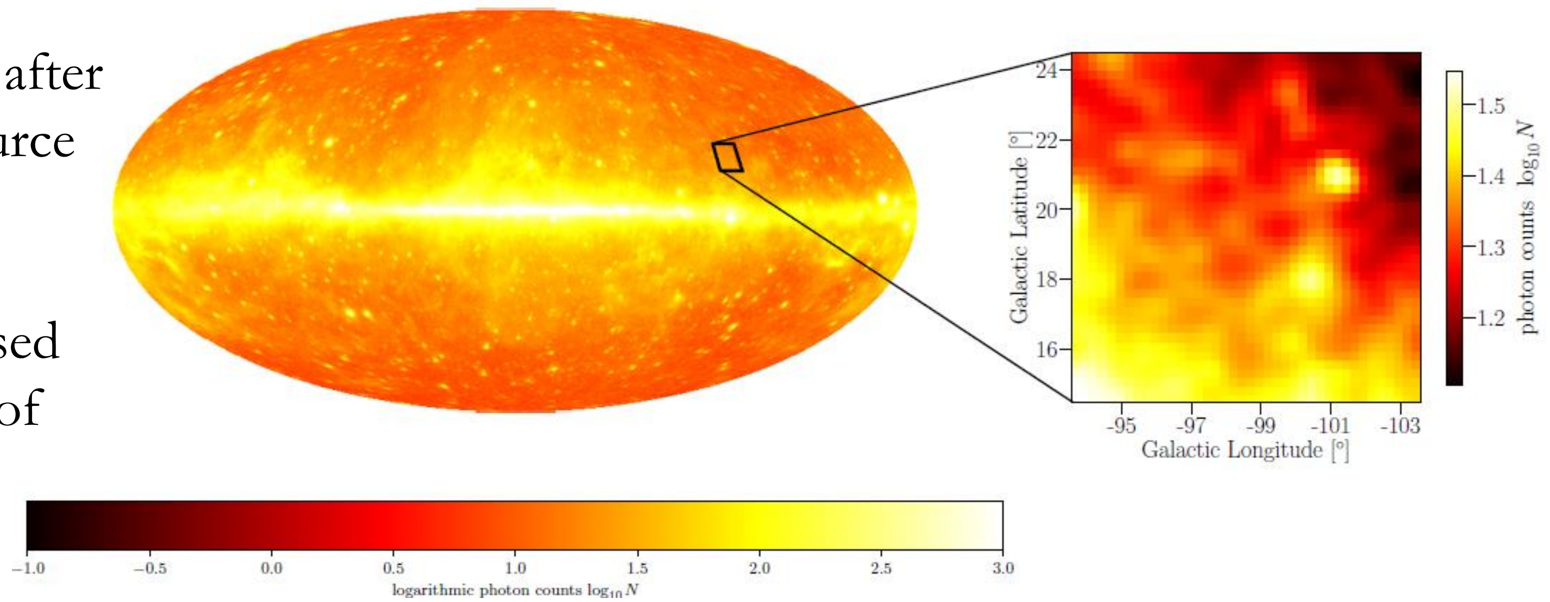
Localization and Classification: Pipeline

- Convolution of Specified source model, raw photon counts with IRF. [Fermitools]

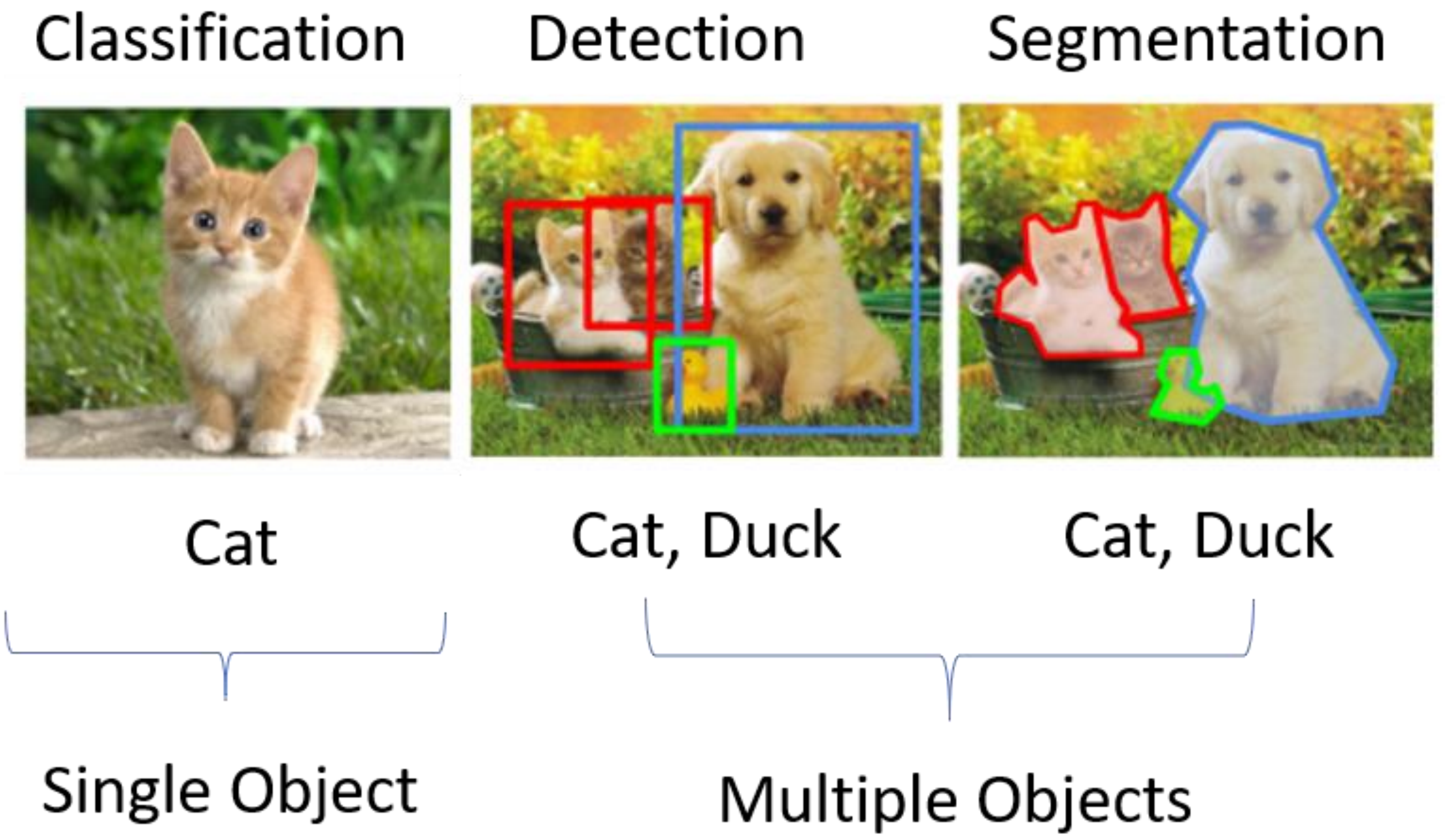
- Separate network for localization and classification.

- Split the sky into $10^\circ \times 10^\circ$ patches and after localization cut $1^\circ \times 1^\circ$ patch around source for classification.

- Random patches (locations of sky) are used for training data. Reduces the possibility of localization network 'learning' the background rather sources.
  - Also tested different background models.

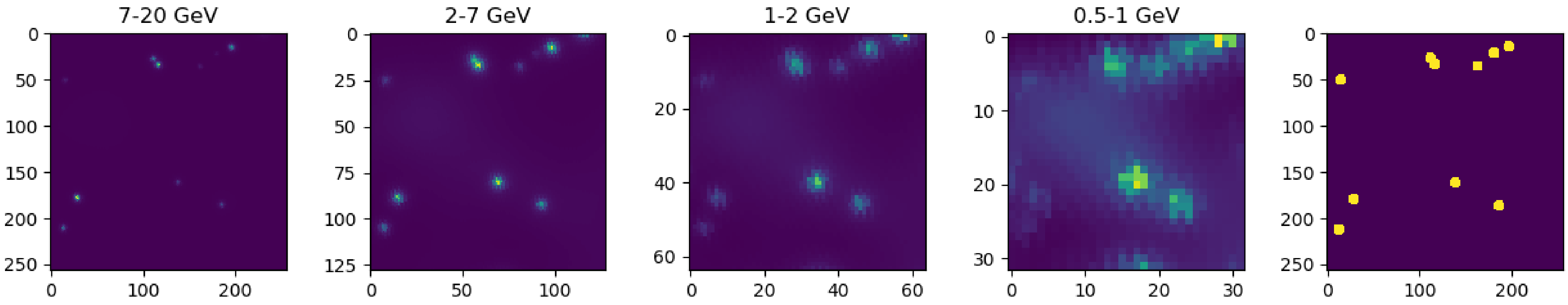


Types of Computer Vision Tasks (Preliminary)



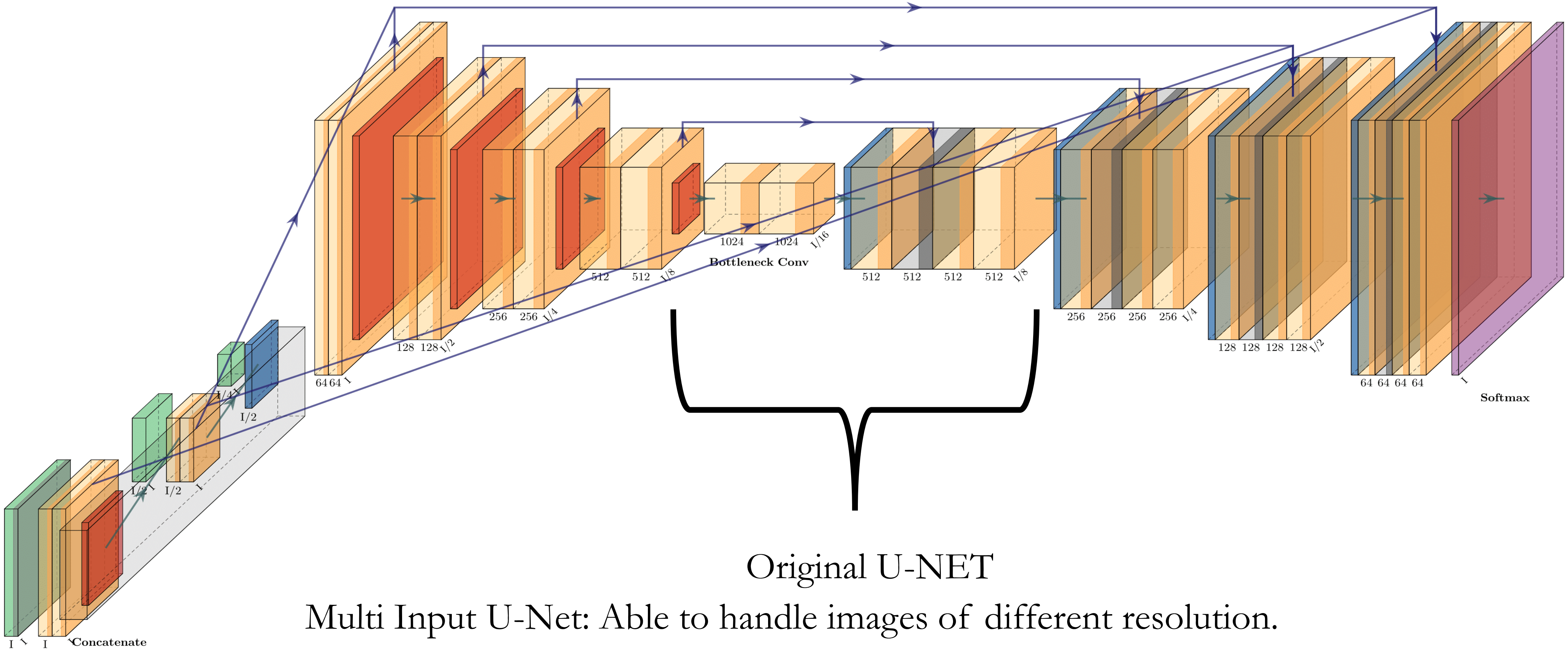
Training Data & Localization Network

- Images of full sky data in 6 energy bins [0.3 GeV - 1 TeV].



- Step1: Implement U-Net like algorithm. Segmentation task.
  - Each pixel is assigned with a label score (≈1, pixel belongs to region around sources, ≈0, otherwise).
- Step2: Apply K-Means algorithm
  - Group the pixels in a cluster and center of cluster is source location. (Lon, Lat)

Multi-Input U-NET Structure



Multi Input U-Net: Able to handle images of different resolution.

Produces a binary mask (1: Source, 0: Rest), Same resolution as the highest resolution input.

Performance Evaluation: Metrics

General performance metric in Deep Learning:

Purity or Precision $= \frac{TP}{TP+FP}$; Completeness or Recall $= \frac{TP}{TP+FN}$

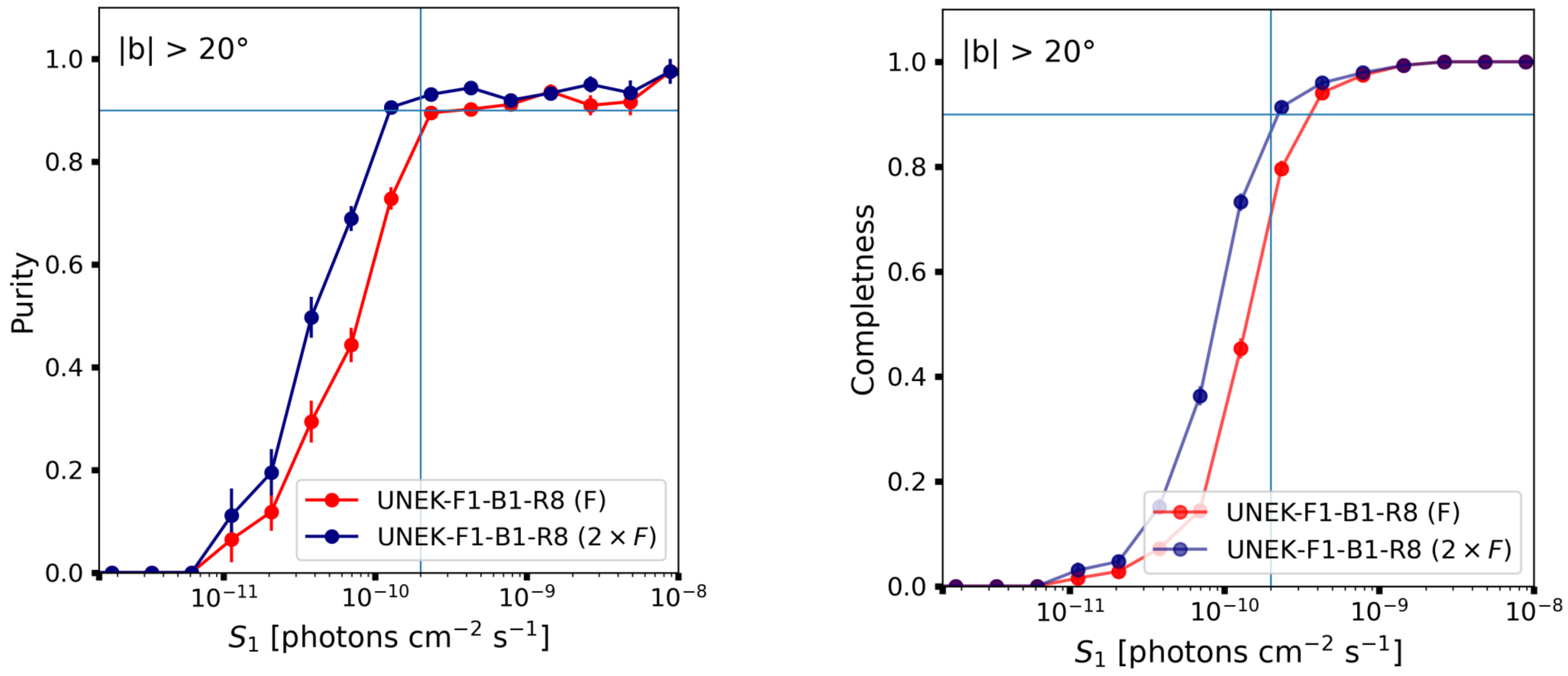
TP: True Positive; Network Identifies a point source present in simulation.

FP: False Positive; Network falsely identifies a point source not present in simulation.

FN: False Negative; Network fails to identify a point source present in the simulation.

How do the precision and recall change as a function of photon flux?

Performance Evaluation on Simulated Data



Comparison of network performance with Front Only (F) and 2 times Front Data. (2F)

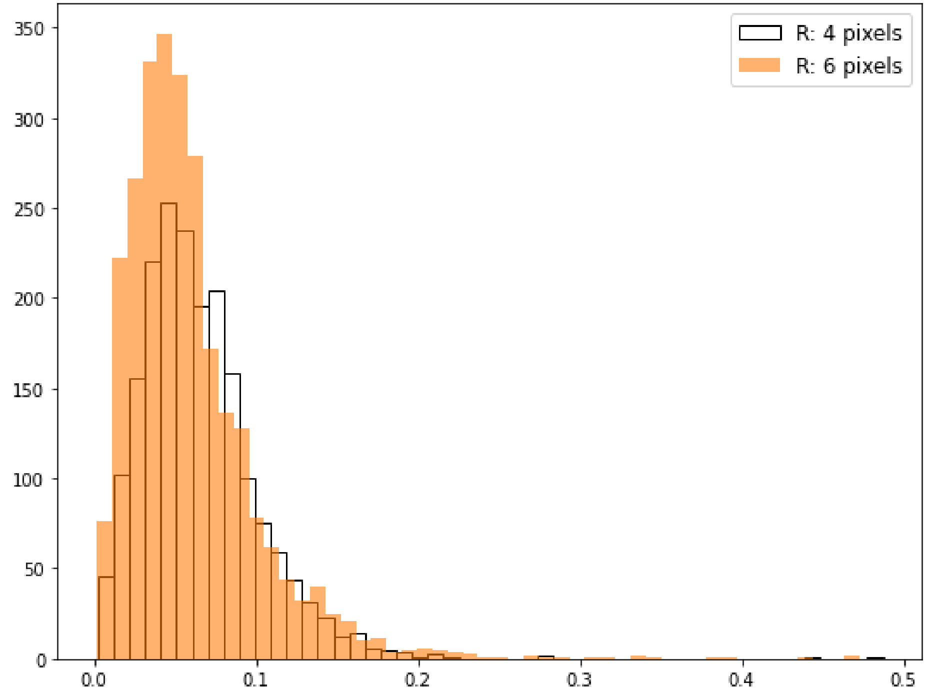
Vertical Blue Line: LAT 4FGL catalog threshold. 2×10^{-12} erg cm^{-2}s^{-1}; [from 4FGL paper]

Assuming power law with index -2, photon flux: 2×10^{-10} photons cm^{-2} s^{-1} above 100 MeV.

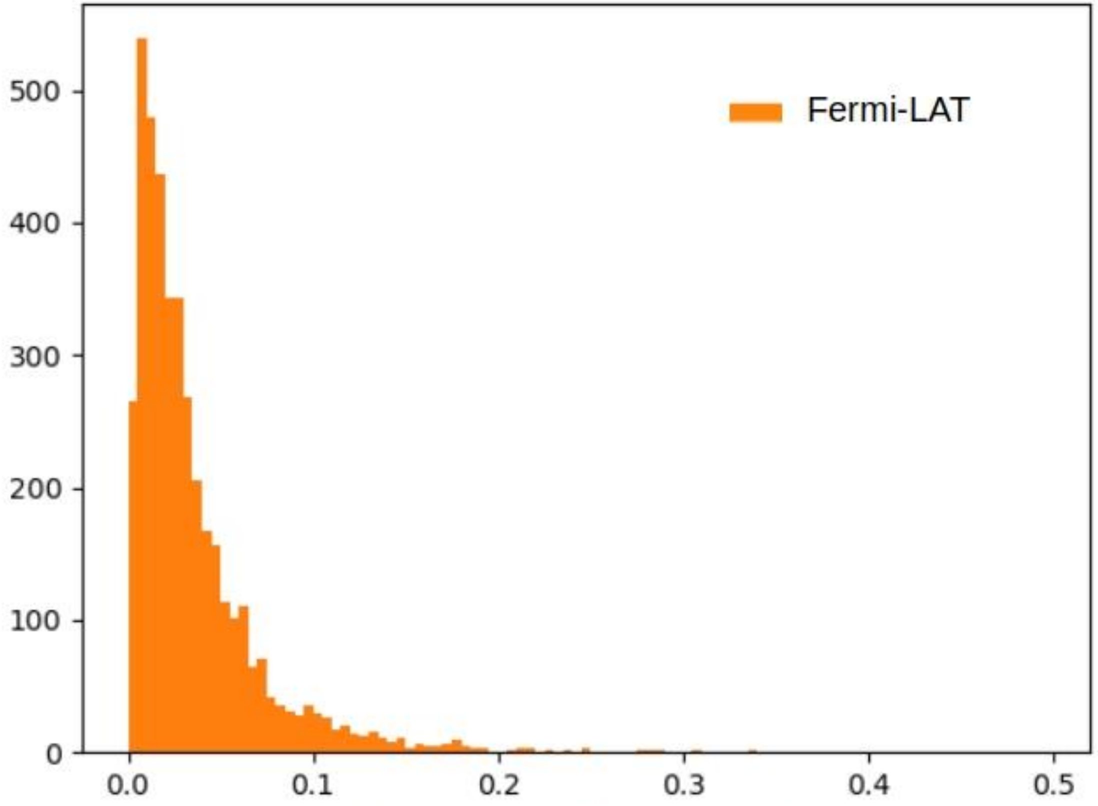
Localization Algorithm Performance:

Accurate Location Prediction: Our Results are based only on γ-ray Data.

Comparison with Original 4FGL Catalog ; Source and Associated Source



Angular Distance (deg.)



Angular Distance (deg.)

Performance on the Real Data:

- Create count cubes (‘gtbin’) of same size for different energy bins from the detected photons (‘Front’).
- Use the best model (based on the performance on the simulated/training data) to generate location of source centers.
    - Number of detected sources depends on the threshold (binary classification of source and background).
    - For all the results shown here, threshold is set to 0.2.

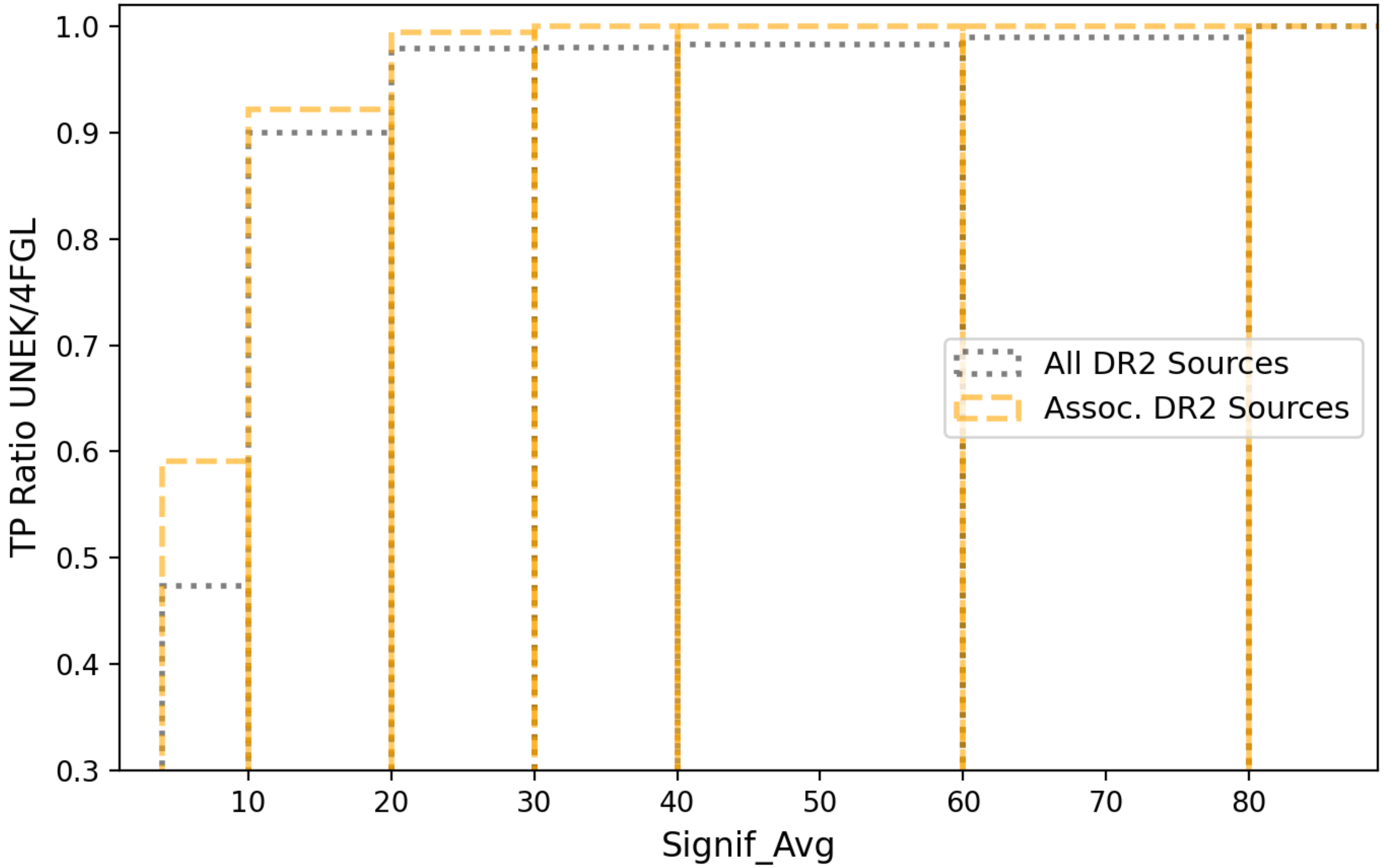
Performance on the Real Data:

- Create count cubes ('gtbin') of same size for different energy bins from the detected photons ('Front').
- Use the best model (based on the performance on the simulated/training data) to generate location of source centers.
    - Number of detected sources depends on the threshold (binary classification of source and background).
    - For all the results shown here, threshold is set to 0.2.

- Compare predicted location lists (Latitude, Longitude) with DR2 catalog locations ('GLAT', 'GLON').
    - Keep the nearest neighbor within 0.5° : True Positive.
        - Association distance is under discussion. Stable results within 0.3° for high significance sources.
    - Iterative search: If a source is associated once, it is removed from the predicted source list.

Performance on the Real Data: 'Signif_ Avg'

- Ratio of True Positives from UNEK and DR2 catalog are shown for different significance of 4FGL source detection.
- Comparison with the 'Associated' list: All sources above significance 40 were detected.
- $\sigma < 10$; the detection ratio drops down to ~59% for associated list.



Performance on the Real Data: Latitude Dependence

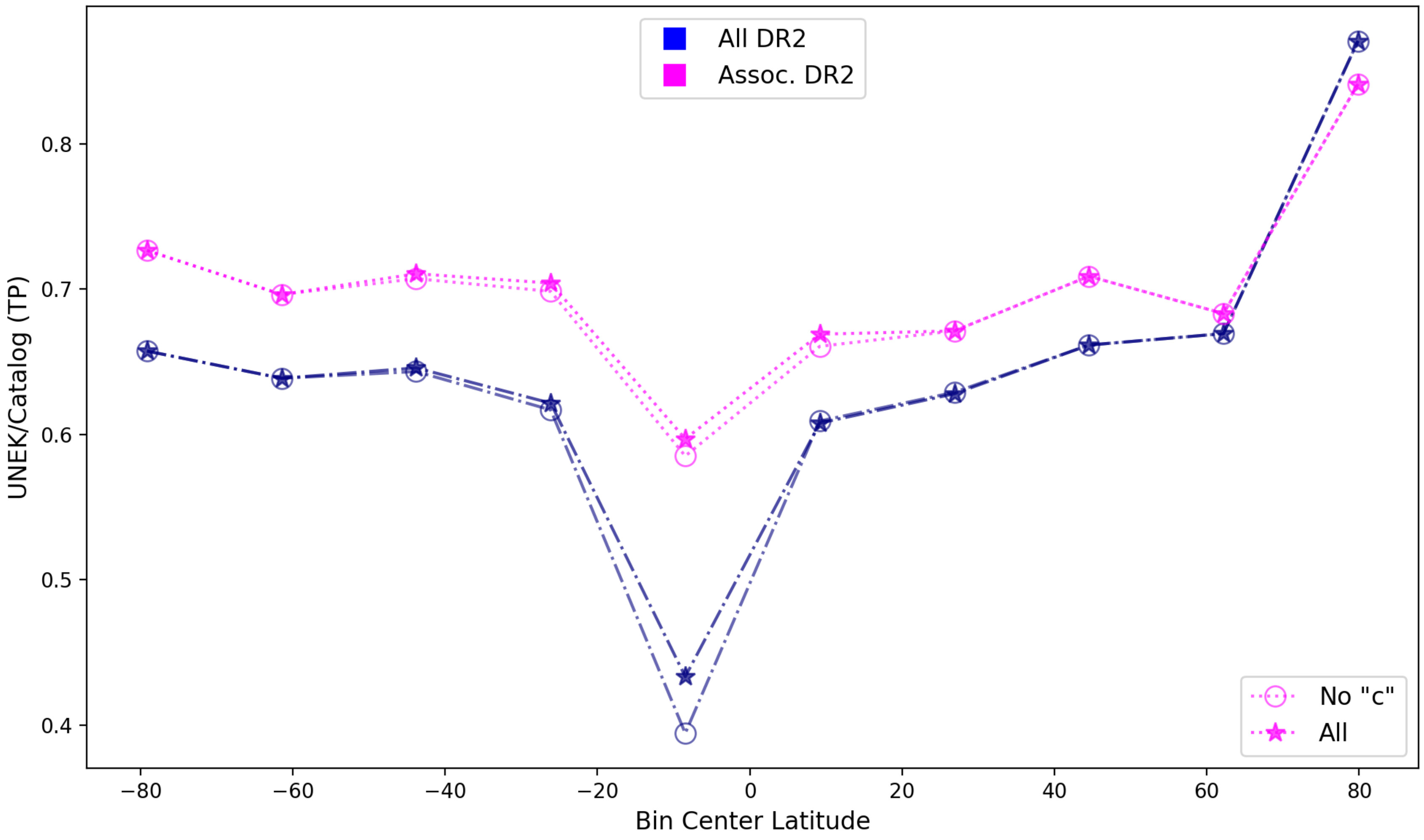
- True Positive ratio at different Latitudes for full DR2 catalog and 'Associated' catalog.

- We also check the effect of removing the 'c' sources

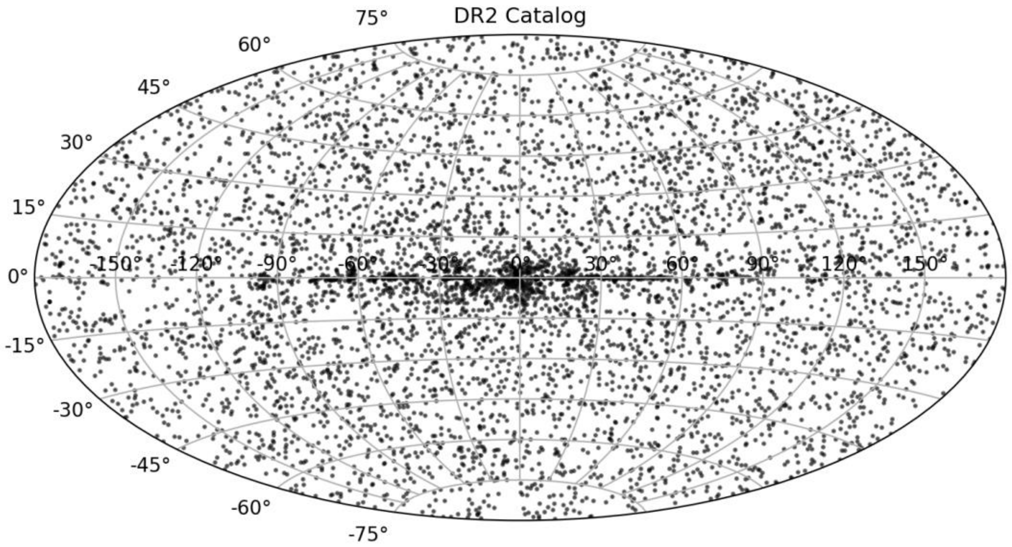
  - Sources coincident with interstellar clump.

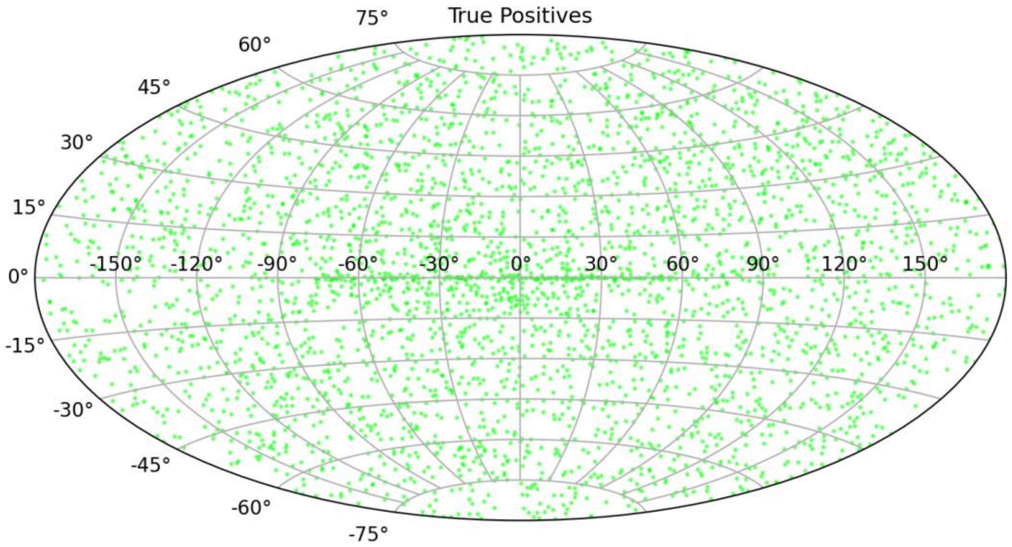
  - 200 'c' sources in full DR2 catalog.

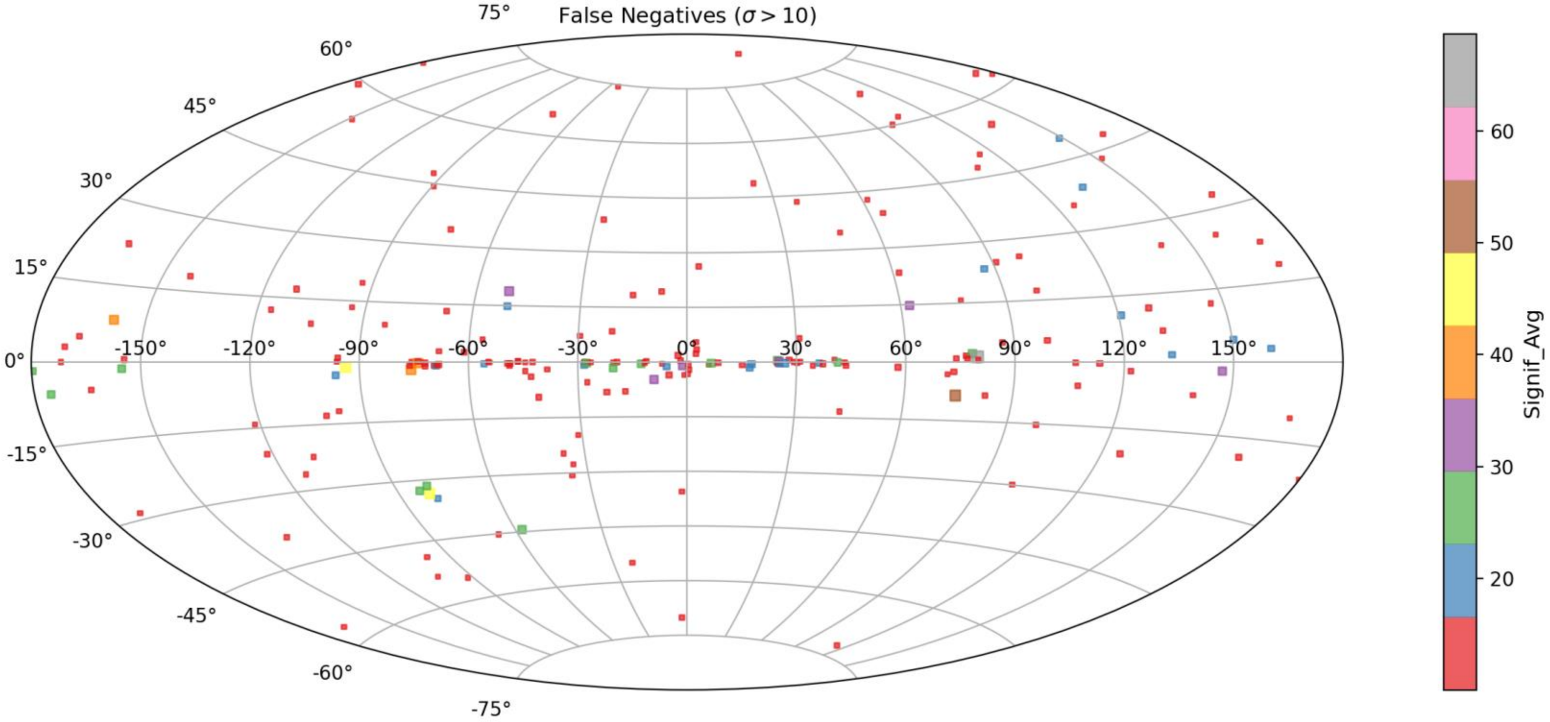
- Association rate drops significantly near galactic plane.



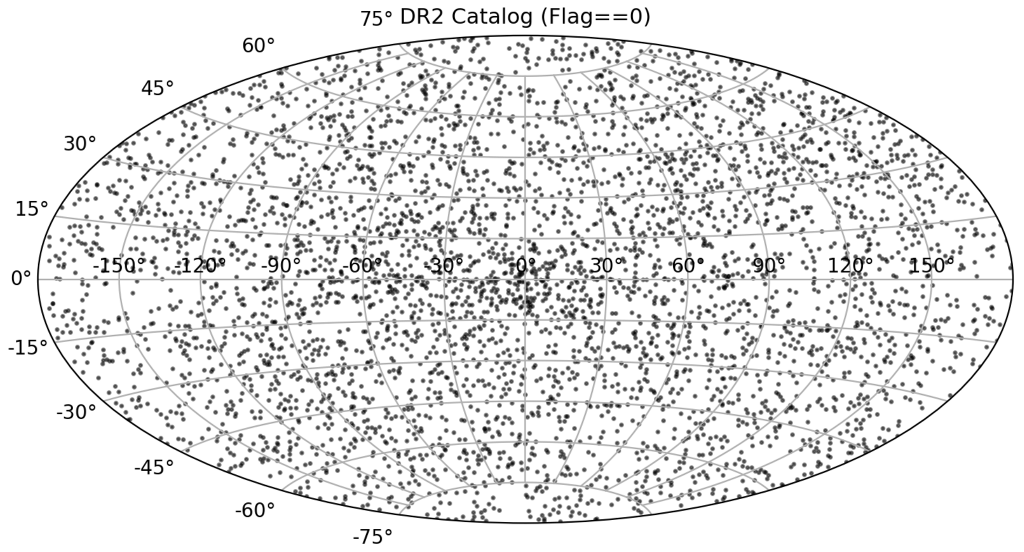
DR2 Catalog: True Positives (using UNEK): False Negatives ($\sigma > 10$):

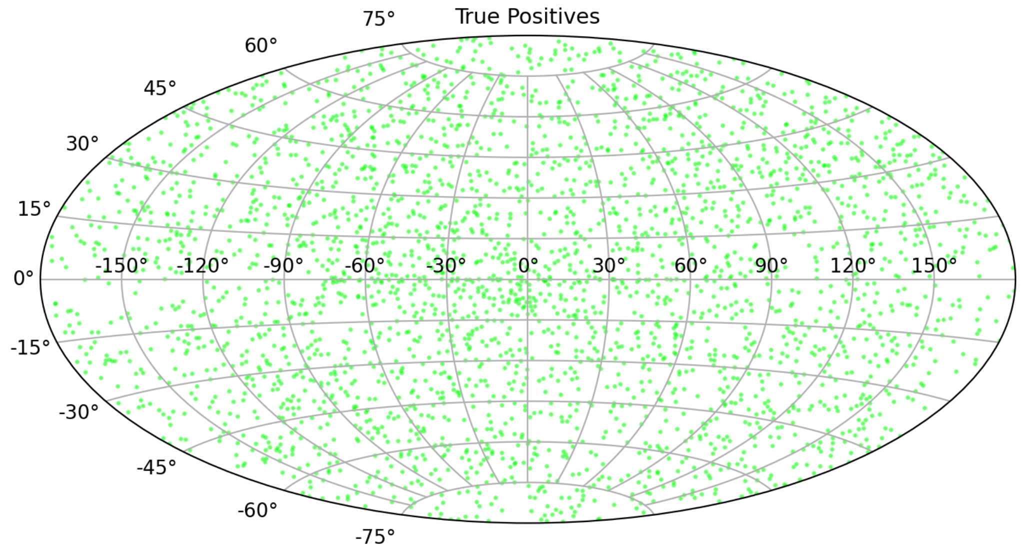


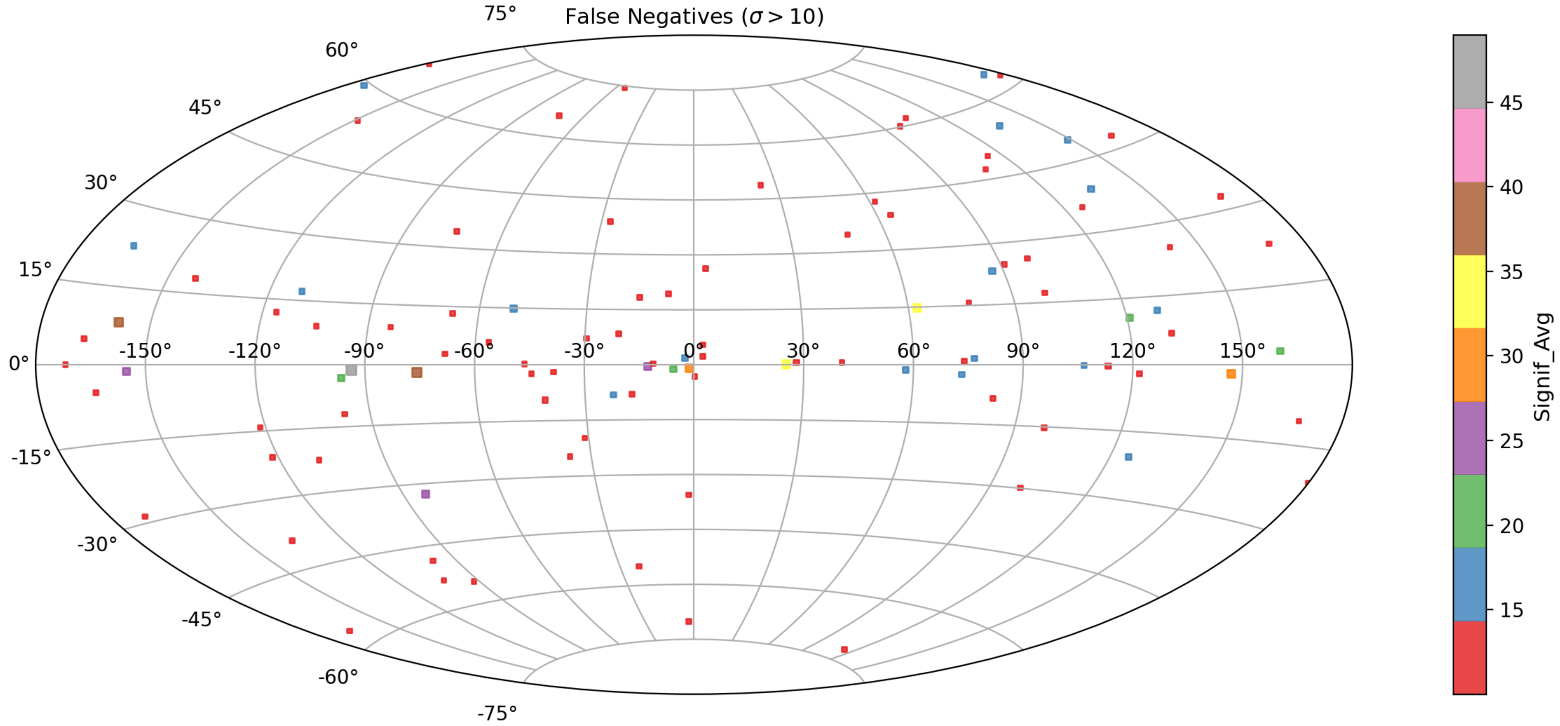




DR2 Catalog (Flag==0): True Positives (using UNEK): False Negatives ($\sigma > 10$):







Summary:

- Develop an automatic gamma-ray data analysis pipeline (only using gamma-ray photon data) for source detection, localization using Deep Neural Network.
    - Results shown here are before classification results. (Ongoing)
- Exploit full detector potential & various source properties to simulate realistic representation of the γ-ray sky.
    - Include various sources and also yearly data with variability information.
- List of detected and localized sources (UNEK) were compared with DR2 catalog.
    - Beyond $\sigma > 10$, association ratio is 90% onwards.
    - Below $\sigma < 10$, association ratio drops to 48% for full catalog; 59% for associated catalog.
- Total number of detected sources with threshold 0.2: ~9200. Possibility of multi-wavelength association?

Summary:

- Develop an automatic gamma-ray data analysis pipeline (only using gamma-ray photon data) for source detection, localization using Deep Neural Network.
    - Results shown here are before classification results. (Ongoing)
- Exploit full detector potential & various source properties to simulate realistic representation of the γ-ray sky.
    - Include various sources and also yearly data with variability information.
- List of detected and localized sources (UNEK) were compared with DR2 catalog.
    - Beyond $\sigma > 10$, association ratio is 90% onwards.
    - Below $\sigma < 10$, association ratio drops to 48% for full catalog; 59% for associated catalog.
- Total number of detected sources with threshold 0.2: ~8014. Possibility of multi-wavelength association?

Summary: Coming Soon..

Snapshot of a Comparison Table/New Catalog

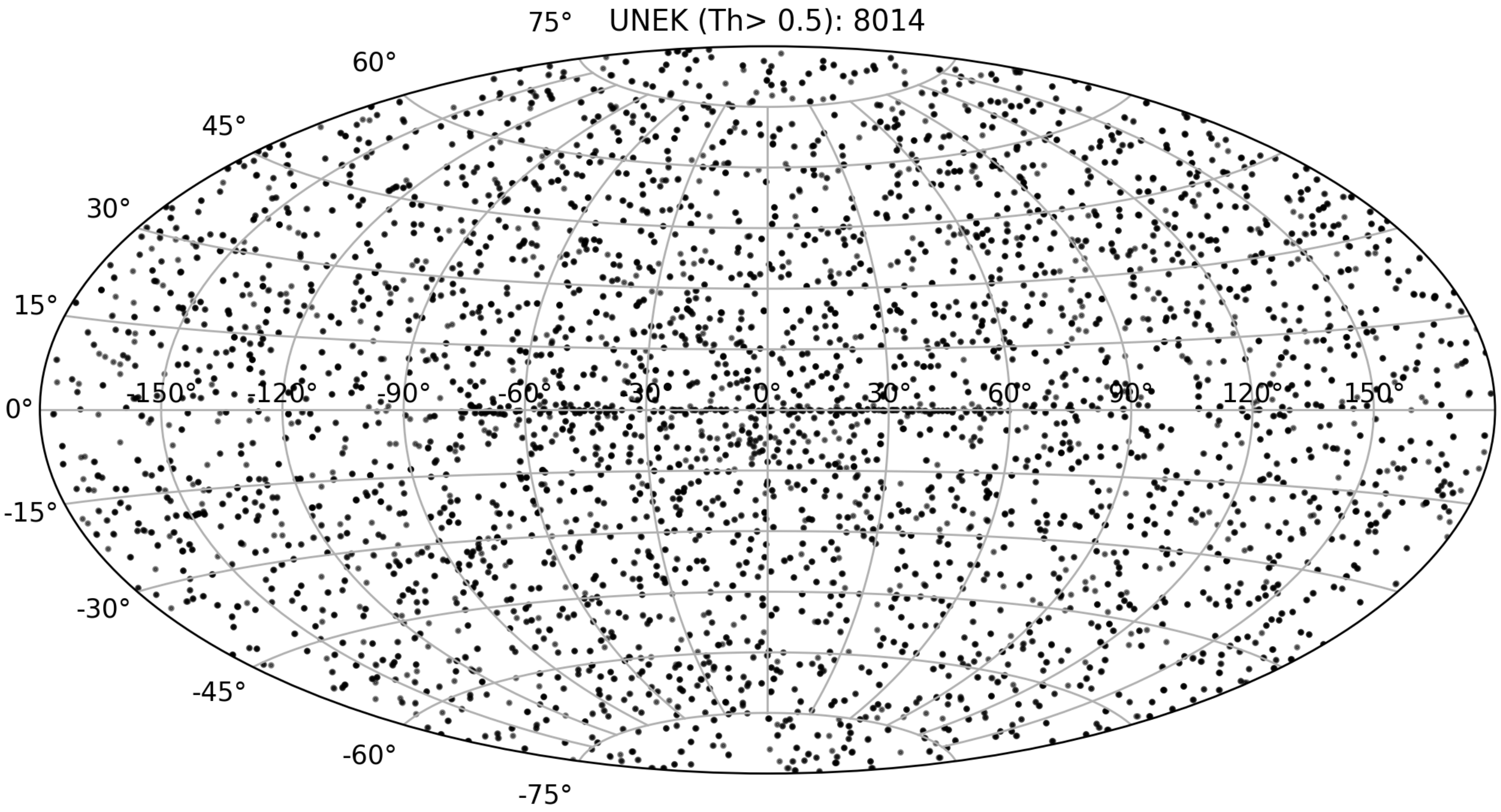
	UNEK			4FGL			
	Probability	LON	LAT	GLAT	GLON	Name	Class
0	0.999870	97.394623	-87.90771	-87.969360	97.574394	4FGL J0047.5-2517	sbg
1	0.999958	142.454575	-87.65952	-87.680092	141.986465	4FGL J0054.7-2455	bll
2	0.999935	68.030205	-86.34456	-86.349205	68.049530	4FGL J0038.2-2459	fsrq
3	0.999683	170.338394	-85.87928	-85.847473	170.384888	4FGL J0104.8-2416	fsrq
4	0.277088	25.296101	81.41749	NaN	NaN	NaN	NaN

'LAT', 'LON': Predicted source location from our algorithm.

'Probability': with 0.2 threshold for background and source pixel classification.

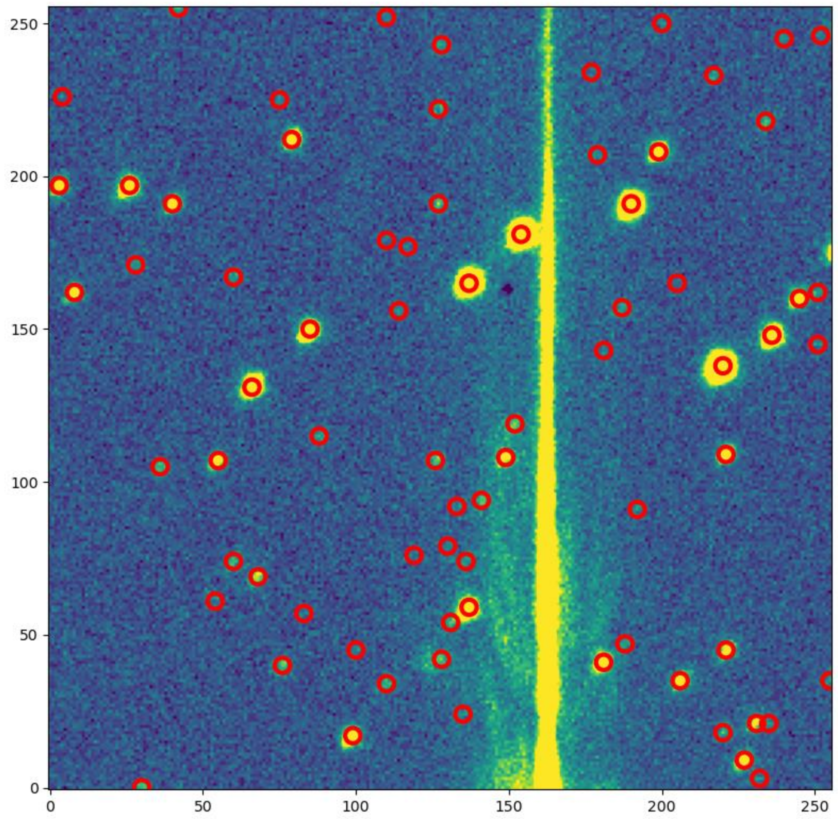
Once we obtain the classification results, we will add a column with 'Fake' tag.

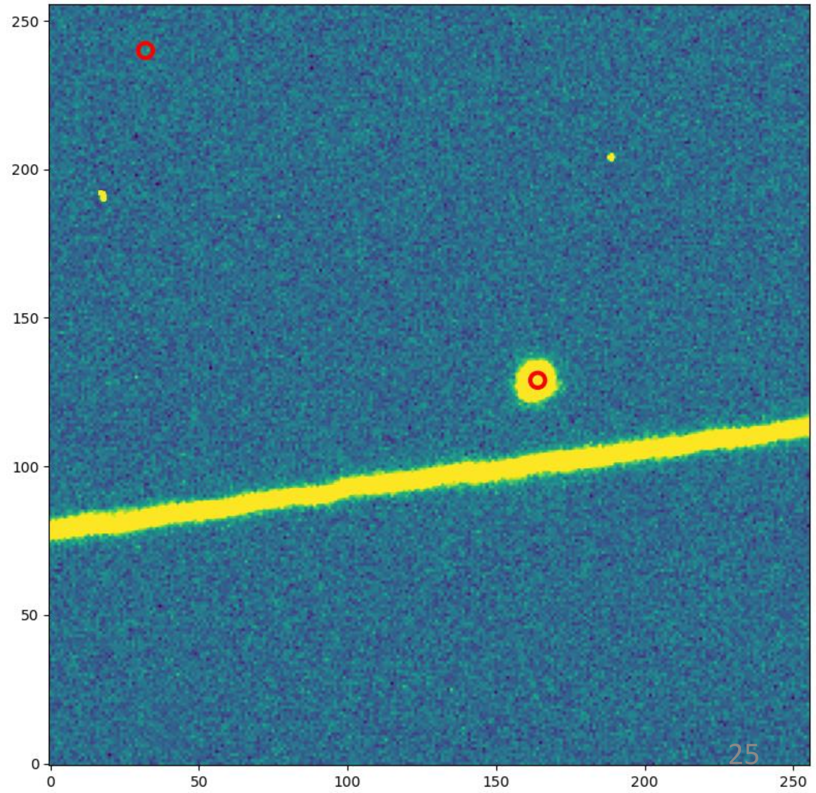
Summary: Predicted Sources: UNEK



Application on Optical Images

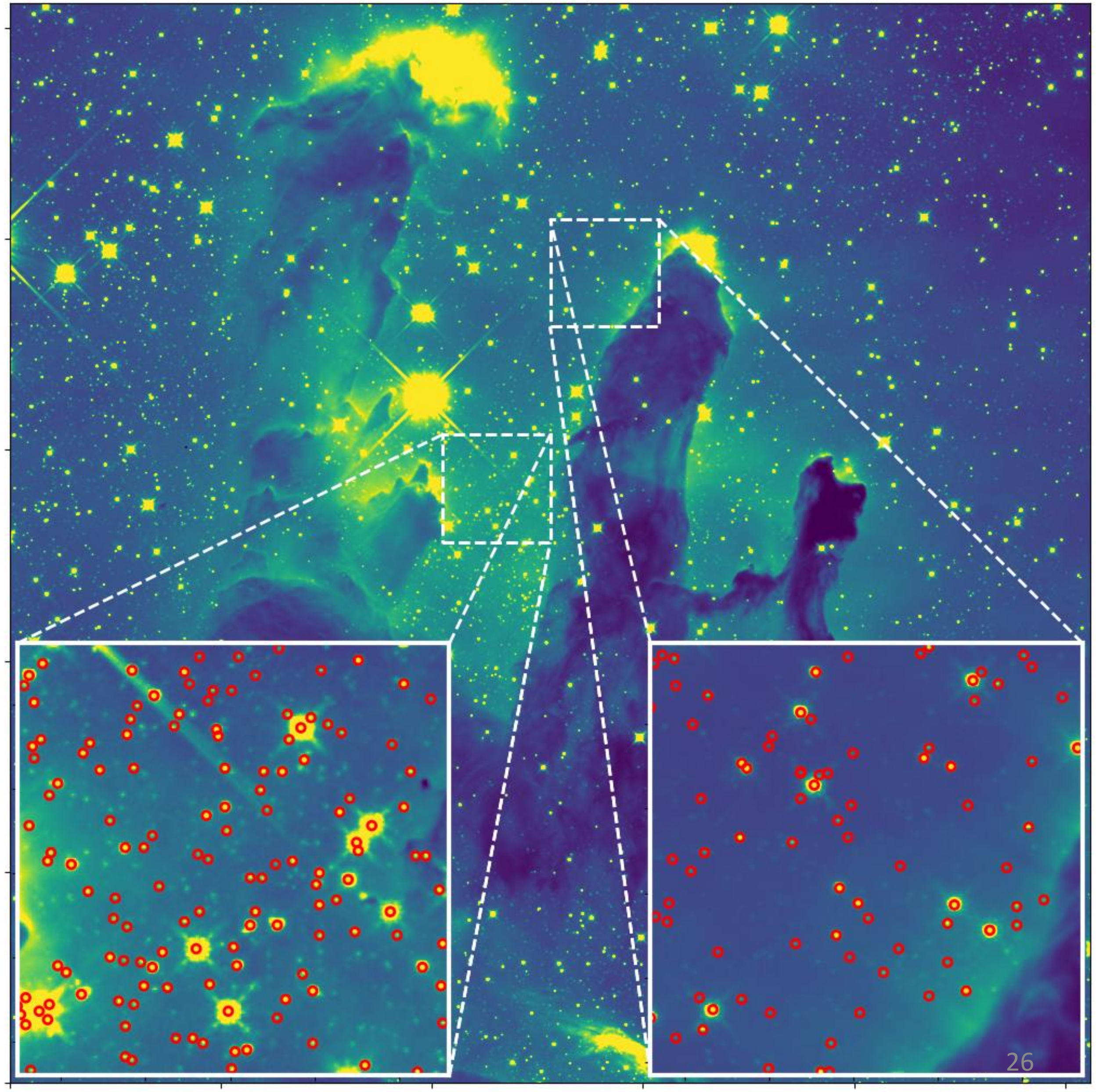
- Work in Collaboration with Optical & ML group in Netherlands. Published in A&A (arXiv: 2202.00489)
- Performance far superior than state of the art source detector.
- Automatically reject satellite motion, flares, cosmic-rays.





Application on Optical Images: Example of Transfer Learning

- Model trained on MeerLicht data, tested against Hubble Telescope Data.
- Without any fine tuning, the model already recovers many sources.
- Robustness of the model: Different backgrounds, different PSFs. 0.11 arc seconds for HST to 2-3 arc seconds for MeerLicht.



Backup

Localization Algorithm Performance:

How accurate the performance is ?

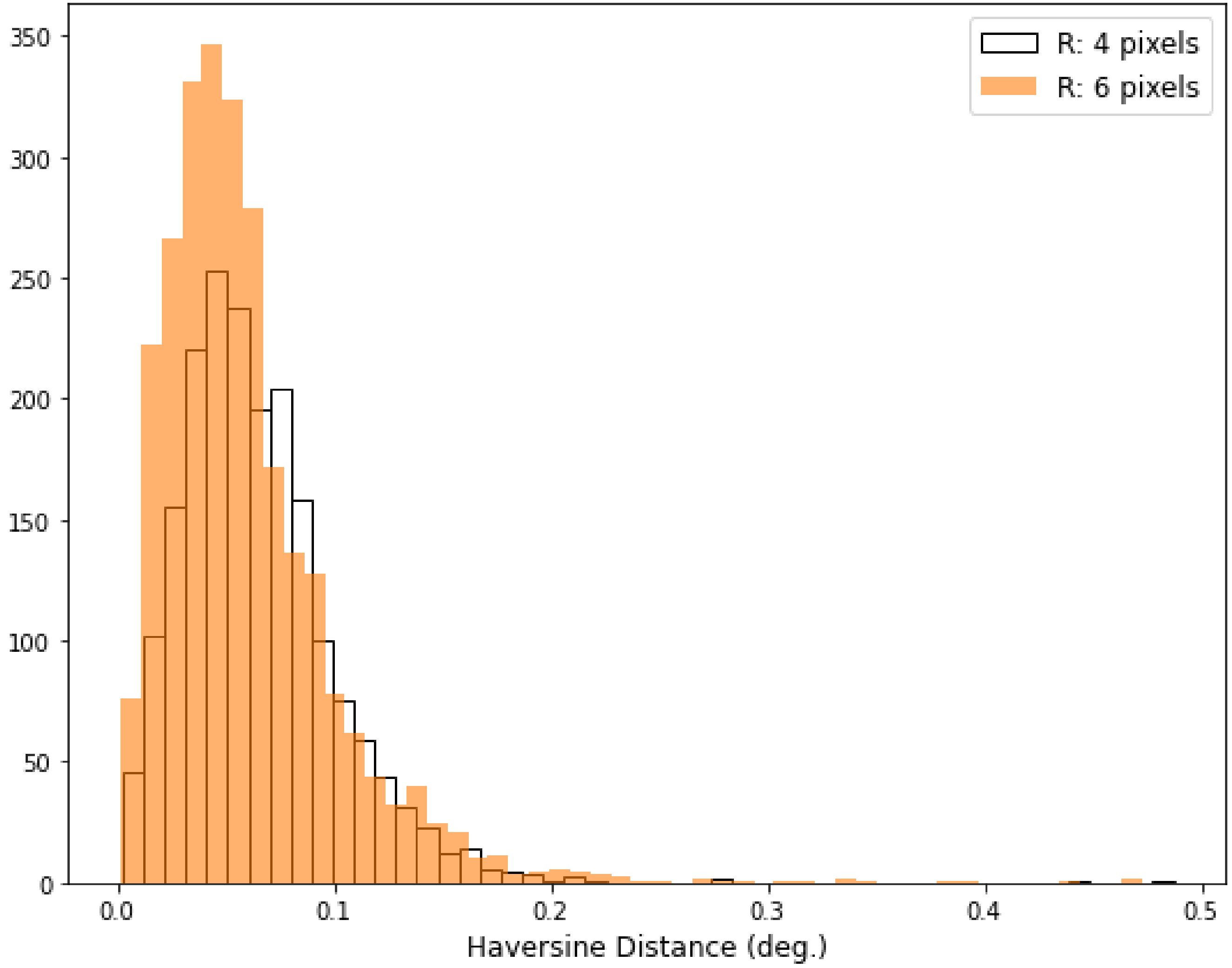
Calculate Haversine Distance between True location and Predicted location.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

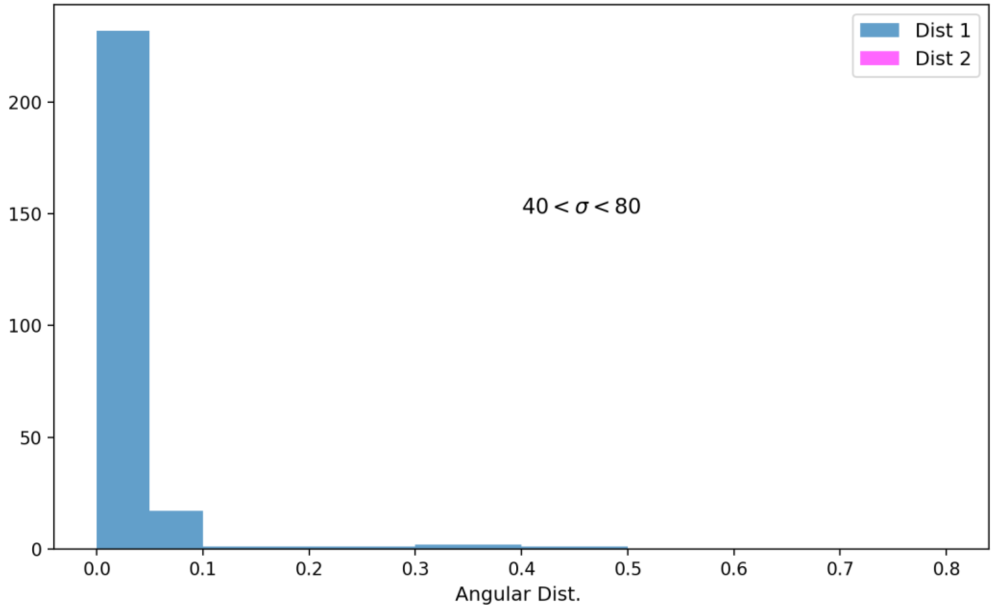
λ_1, λ_2 Longitudes of point 1, 2

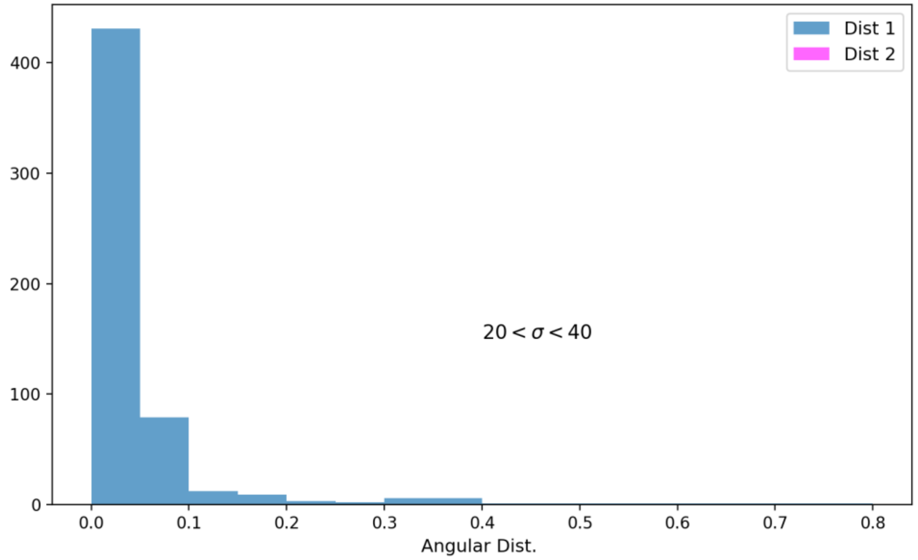
ϕ_1, ϕ_2 Latitudes of point 1, 2

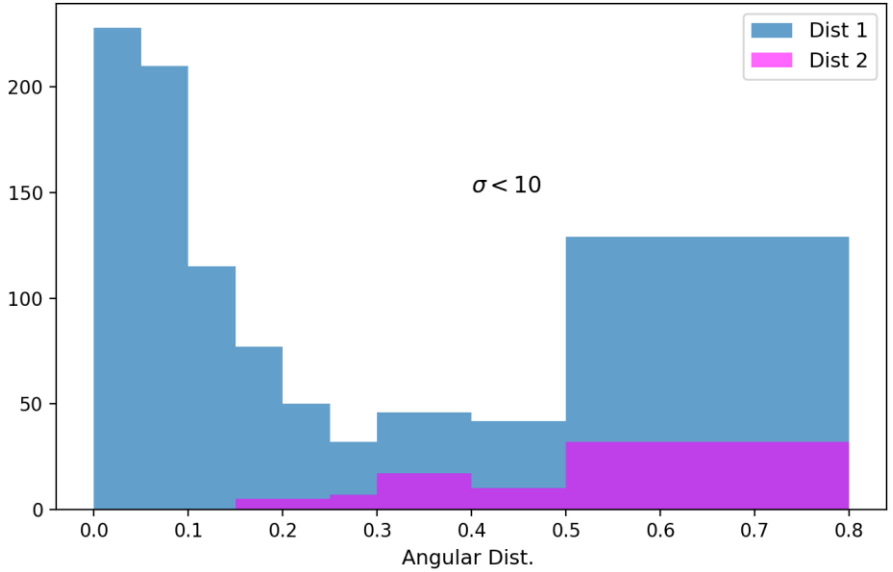
Calculated using Astropy Module.



Nearest Neighbor Distance

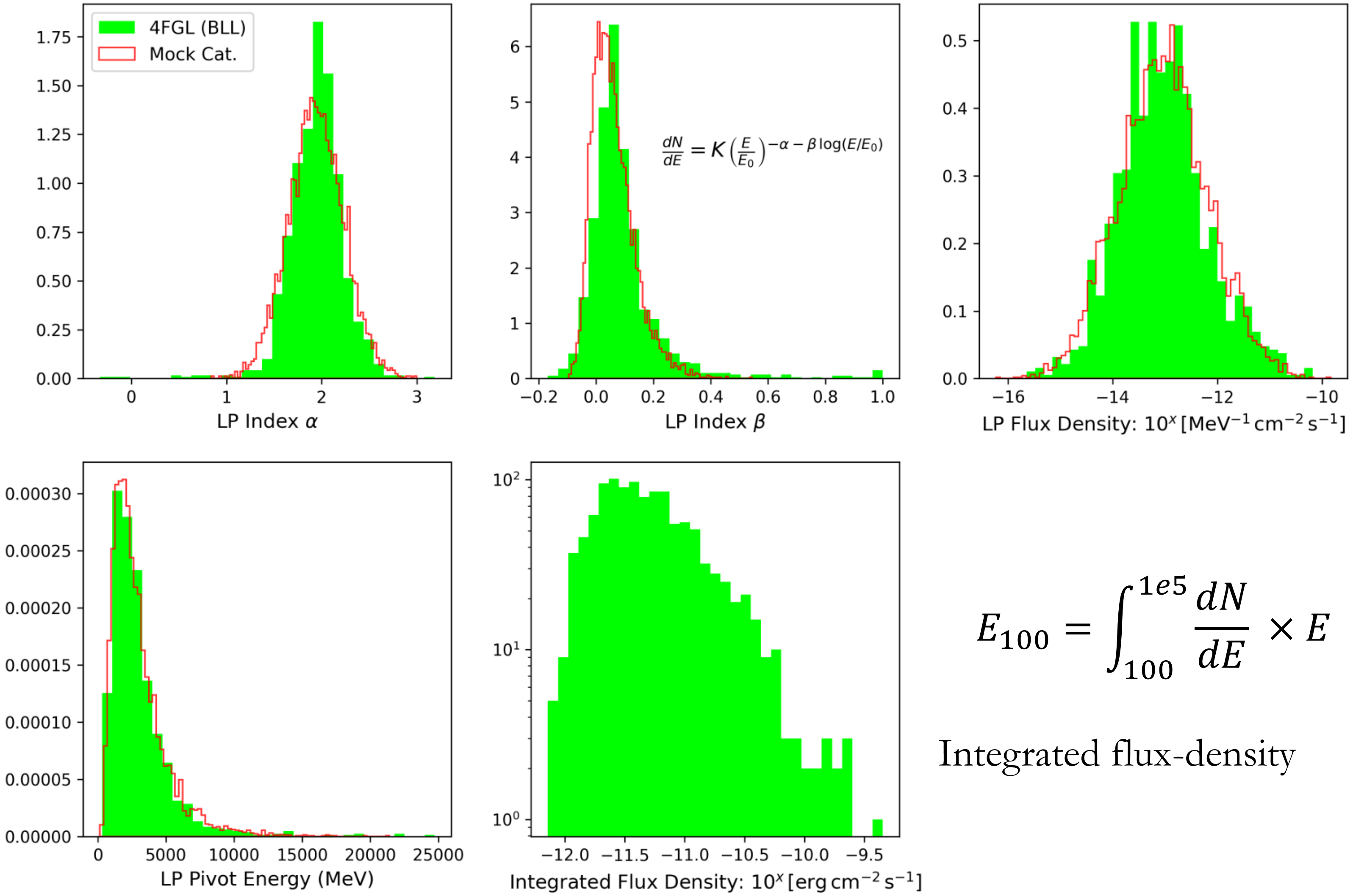






Simulation: Mock Catalog Generation

- Spectral shape:
  - Log Parabola
  - $\frac{dN}{dE} = K\left(\frac{E}{E_0}\right)^{-\alpha-\beta\log\left(\frac{E}{E_0}\right)}$
  - AGNs (BLLac, FSRQ, PWN, SPP)
- Distribution in Sky:
  - BLLac, FSRQ : Uniformly distributed over the whole sky.
  - PSR, PWN/SPP : Uniform distribution in longitude
  - Latitude distribution peaks at the plane.

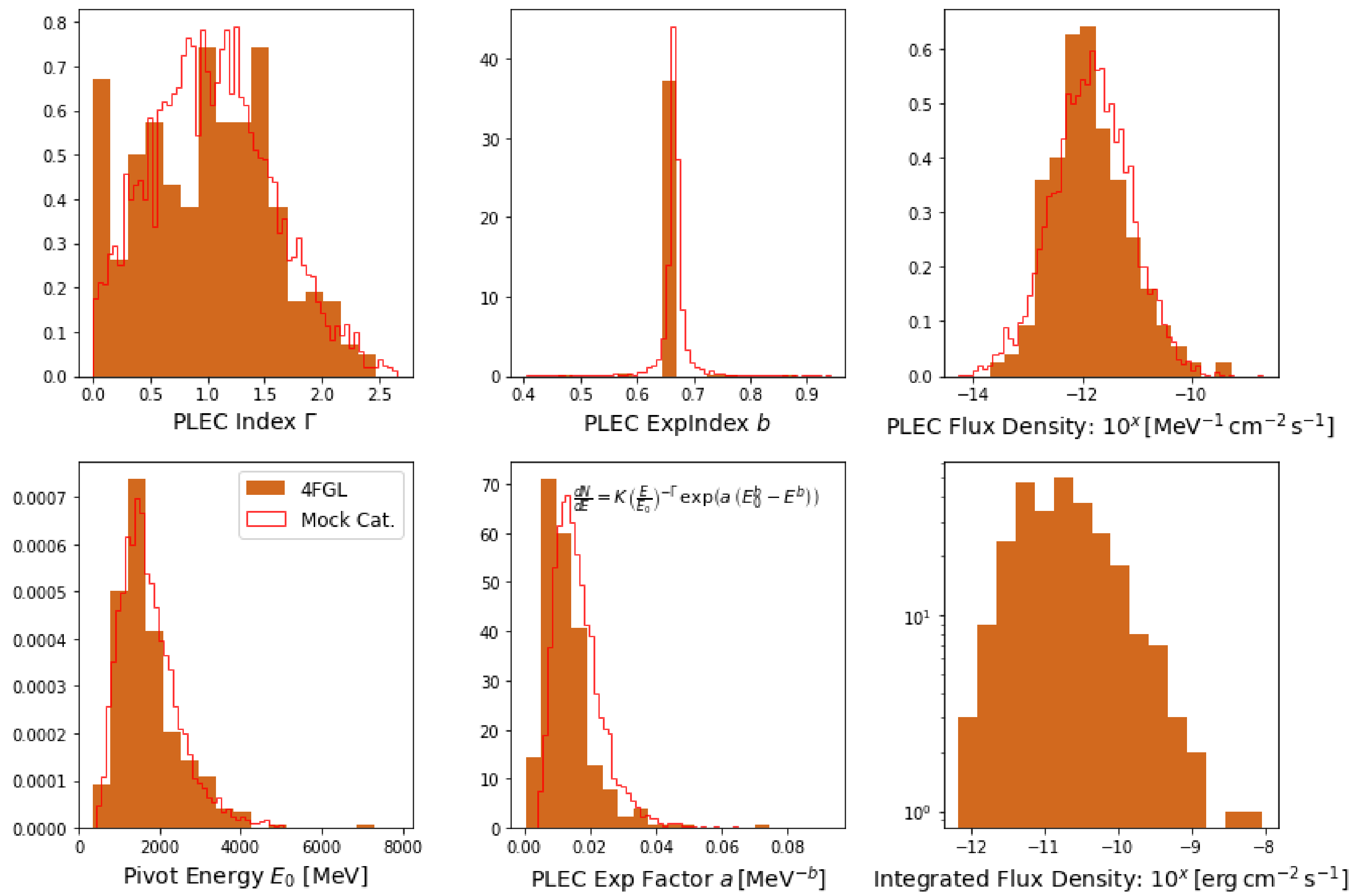


$$E_{100} = \int_{100}^{1e5} \frac{dN}{dE} \times E$$

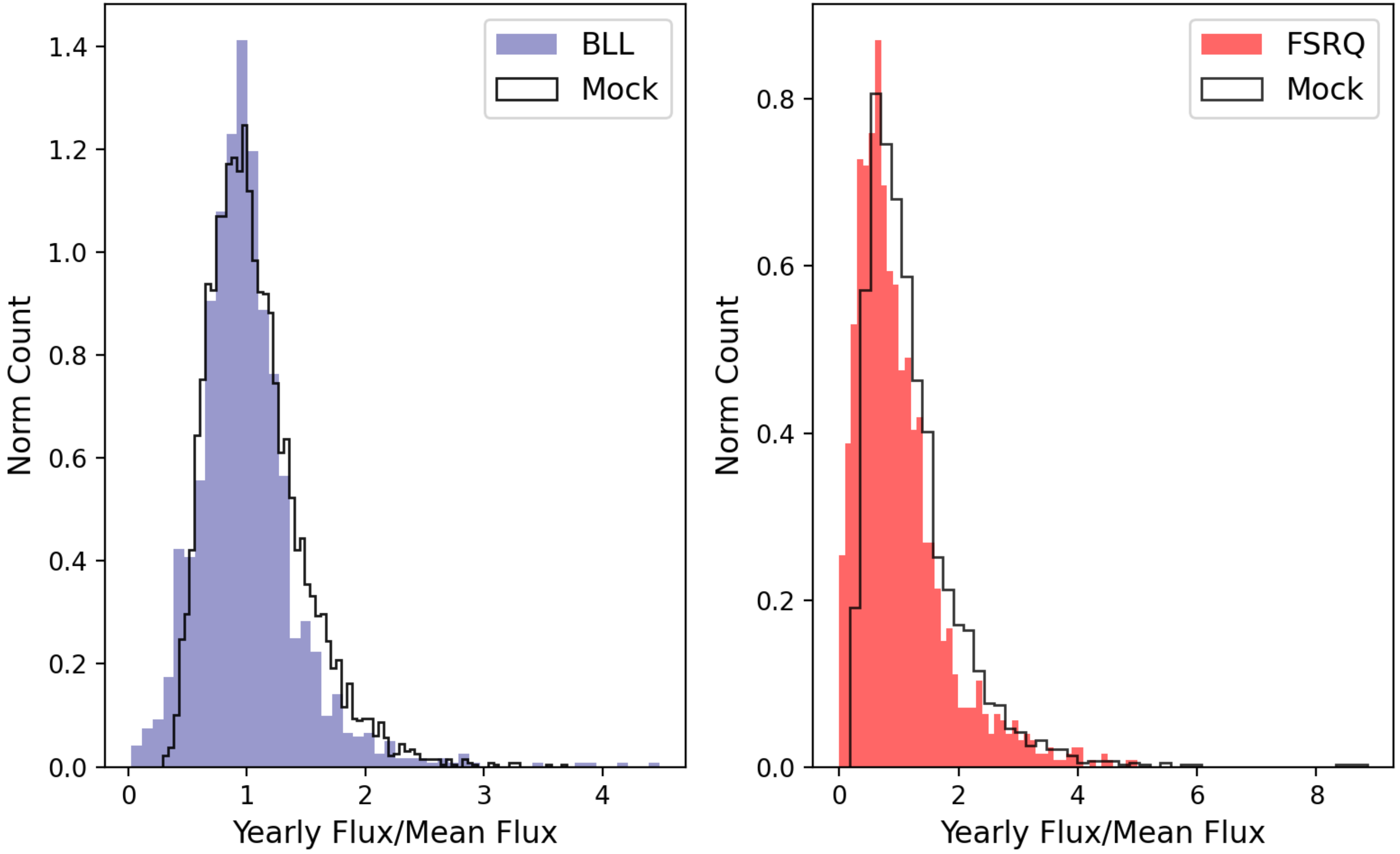
Integrated flux-density

Simulation: Mock Catalog Generation

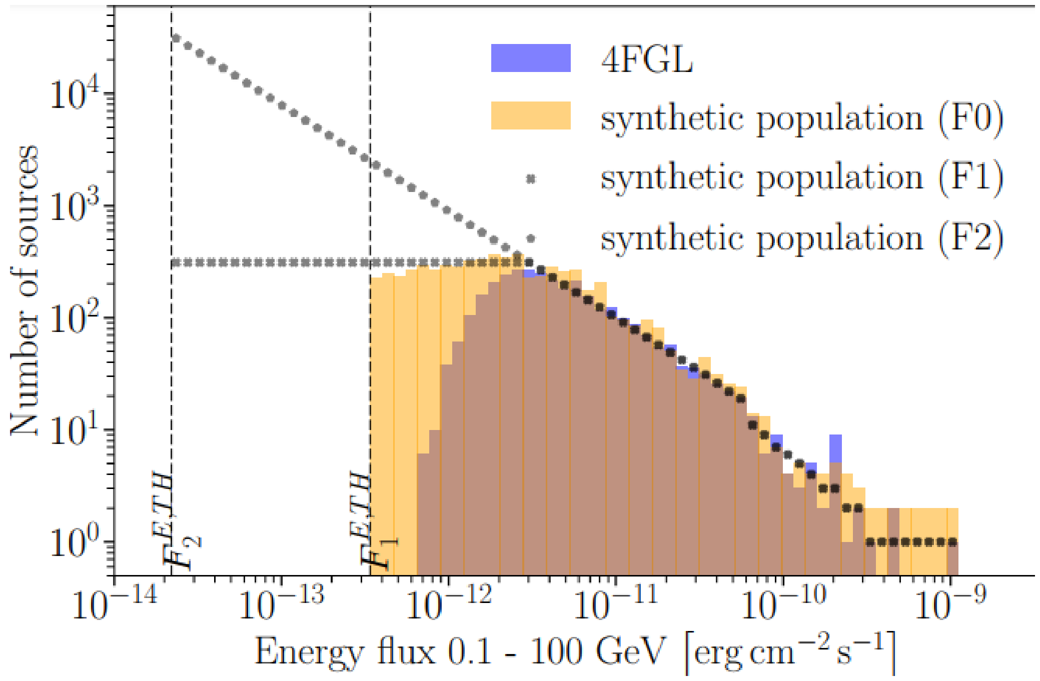
- Spectral shape:
  - PLEC
  - $\frac{dN}{dE} = K\left(\frac{E}{E_0}\right)^{-\Gamma} \exp(a\,(E_0^b\ - E^b))$; PSR.
- Distribution in Sky:
  - PSR: Double Gaussian for LAT.



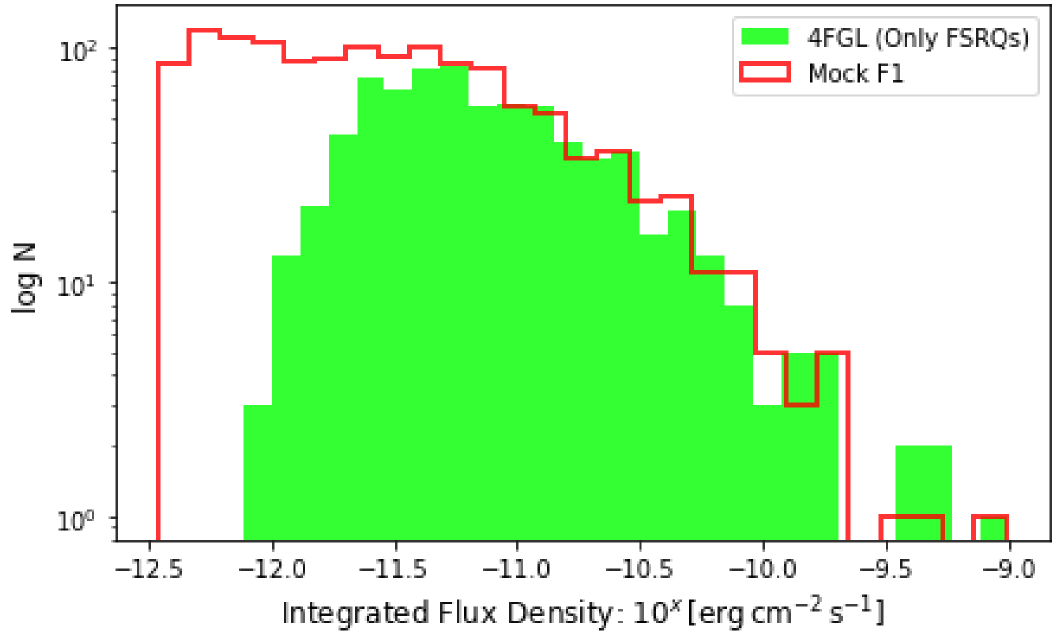
Variability of Blazars:

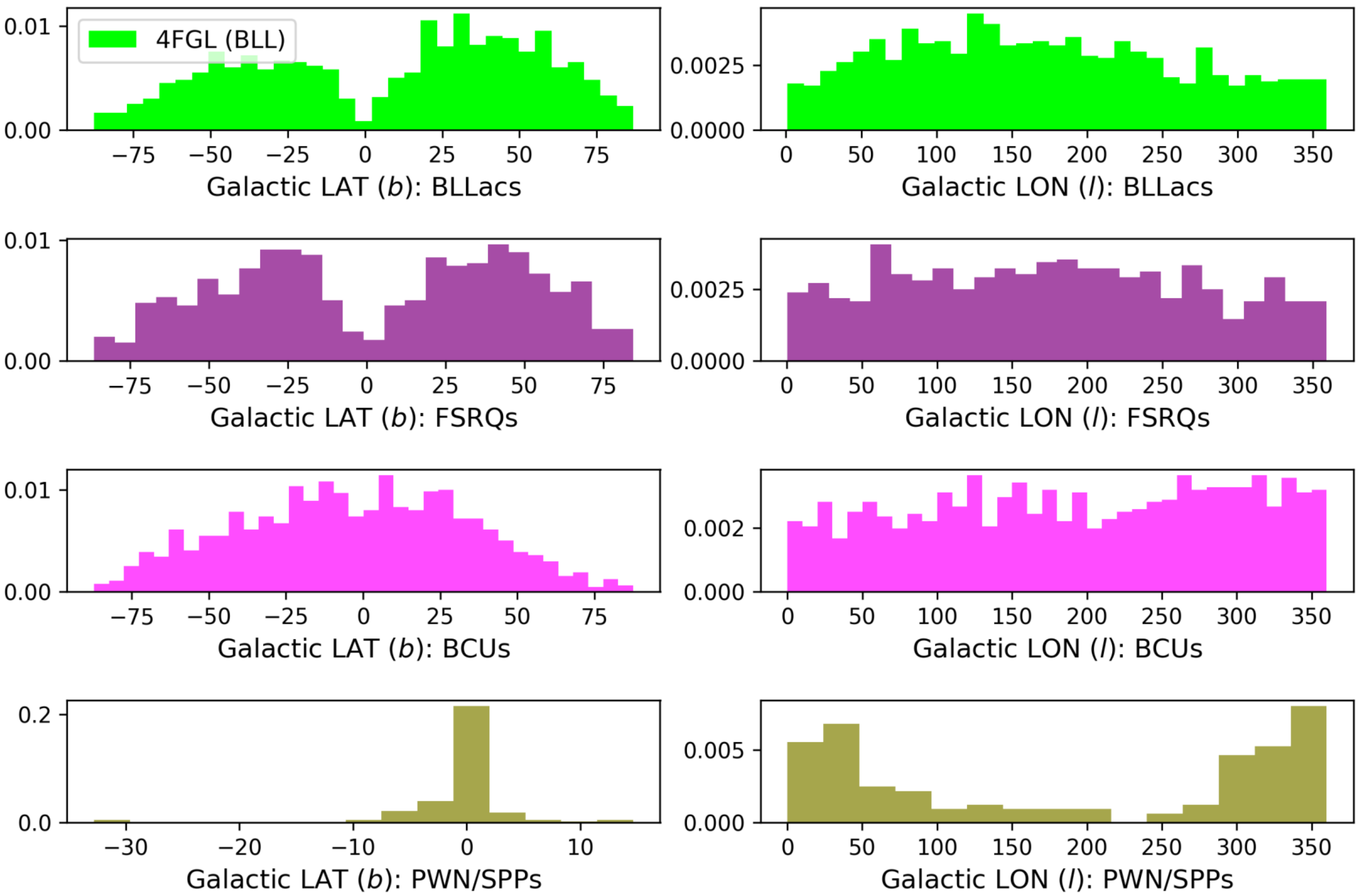


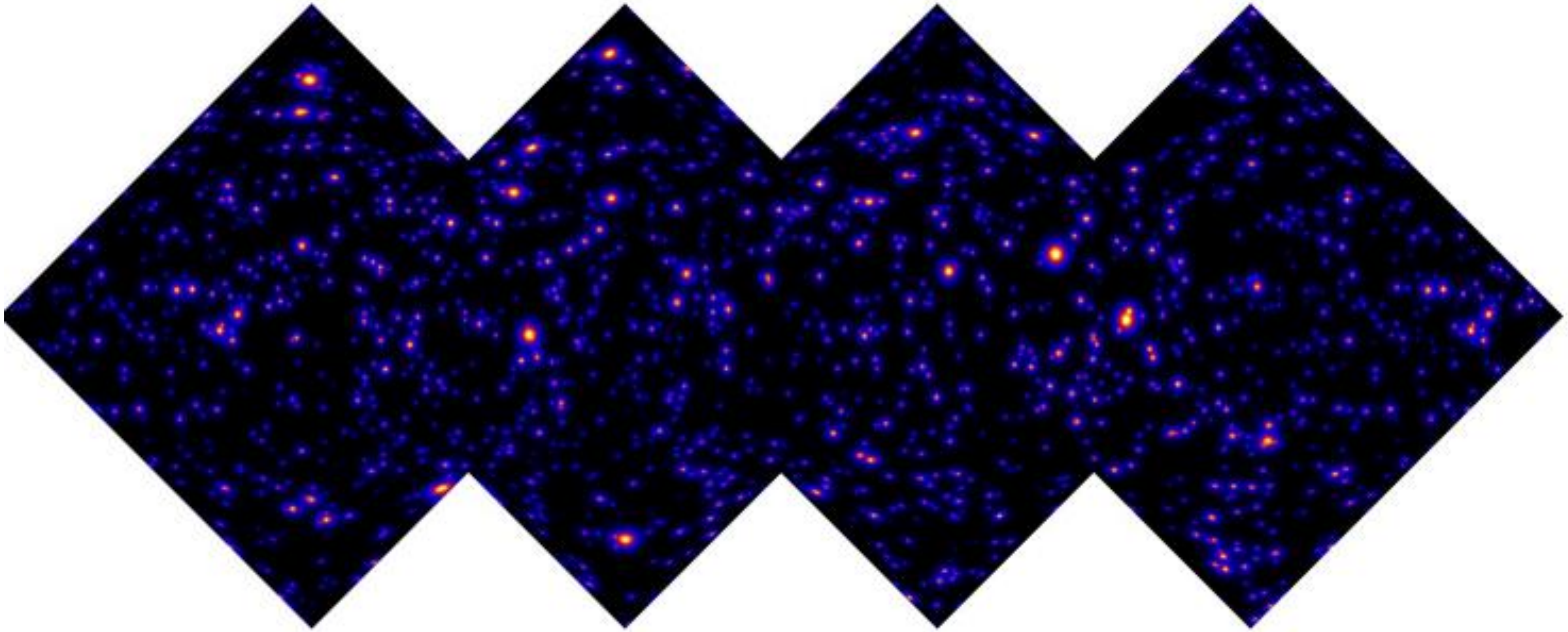
F1-F0 Data Old



F1- Data New



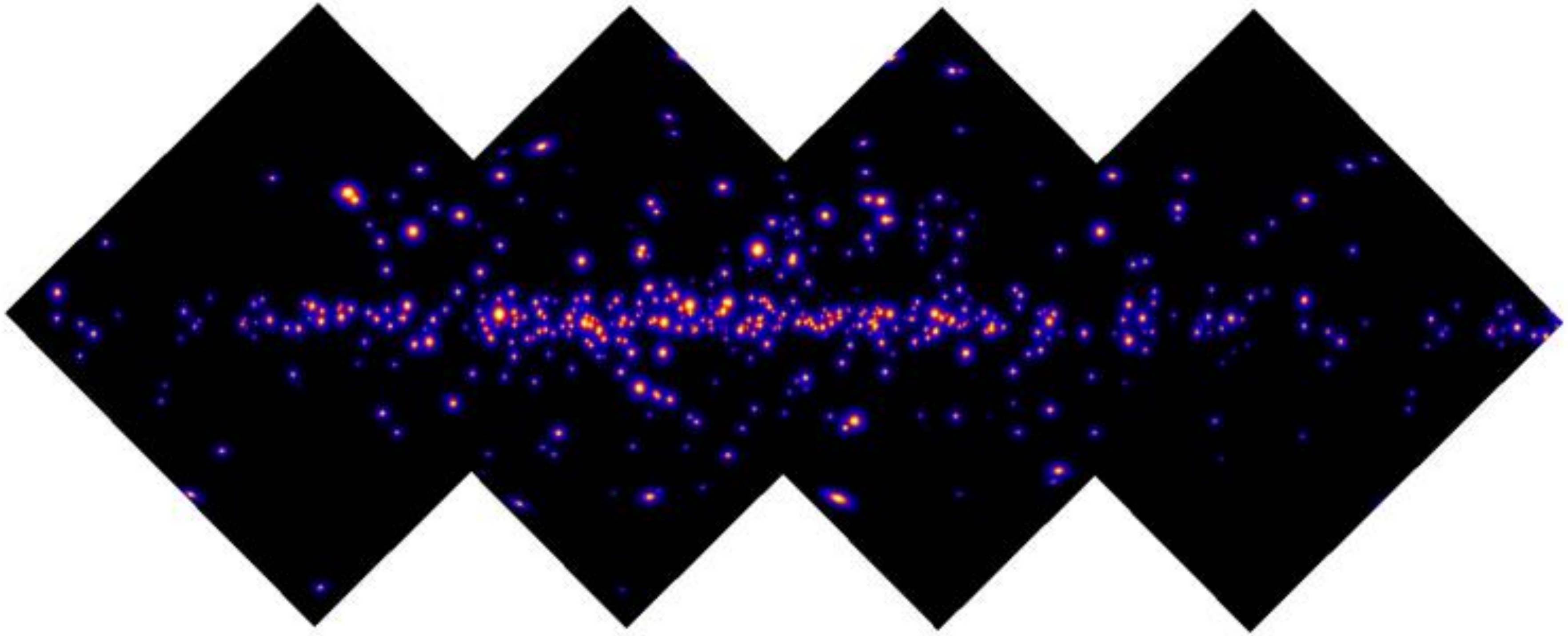

4FGL (BLL)
Galactic LAT (b): BLLacs
Galactic LON (l): BLLacs
Galactic LAT (b): FSRQs
Galactic LON (l): FSRQs
Galactic LAT (b): BCUs
Galactic LON (l): BCUs
Galactic LAT (b): PWN/SPPs
Galactic LON (l): PWN/SPPs



Simulated Mock Catalog for 'BLLac' Class
Energy Bin 1-2 GeV, Year 2017-2018,
No. of. Sources: 1618

‘PSR’; 2017-2018 (1-2 GeV)

No. of. Sources : 468



Classification Network: Training Strategy

- Dealing with class-imbalance data-set. Training data dominated by BLL and FSRQs.
    - Use augmentation; Weighted Loss.

- To incorporate variability of sources effectively we use 4D data structure (6, 6, 10, 6).

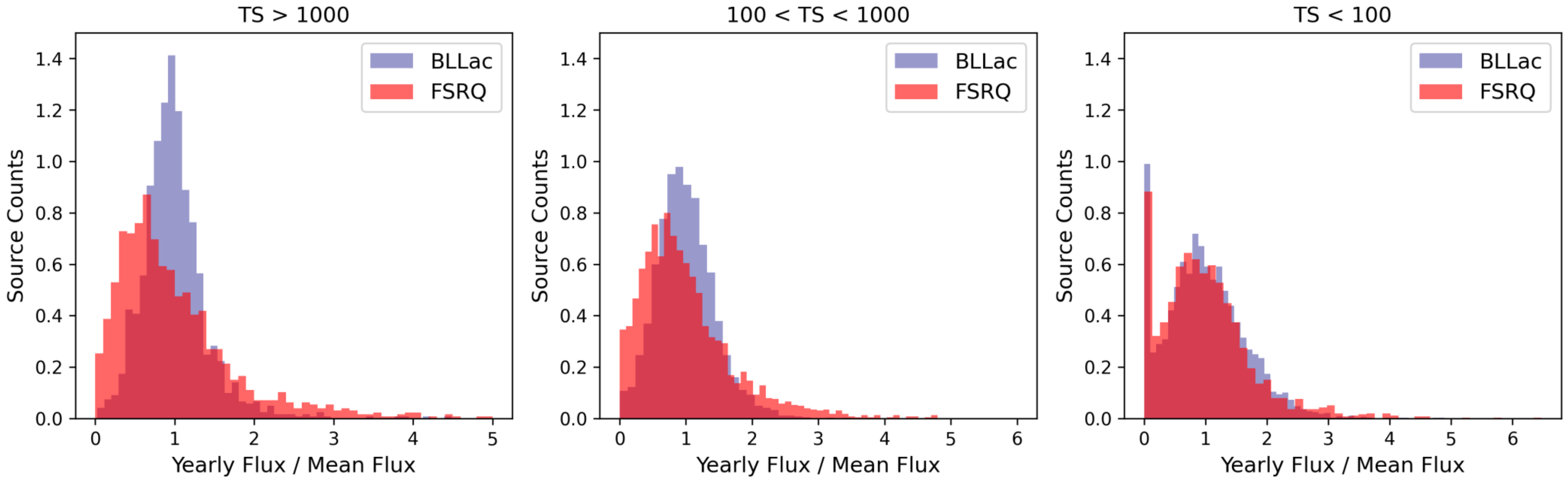
  (Width, Height, Year, Energy)

- Arrange yearly data with decreasing counts. Max count on top.

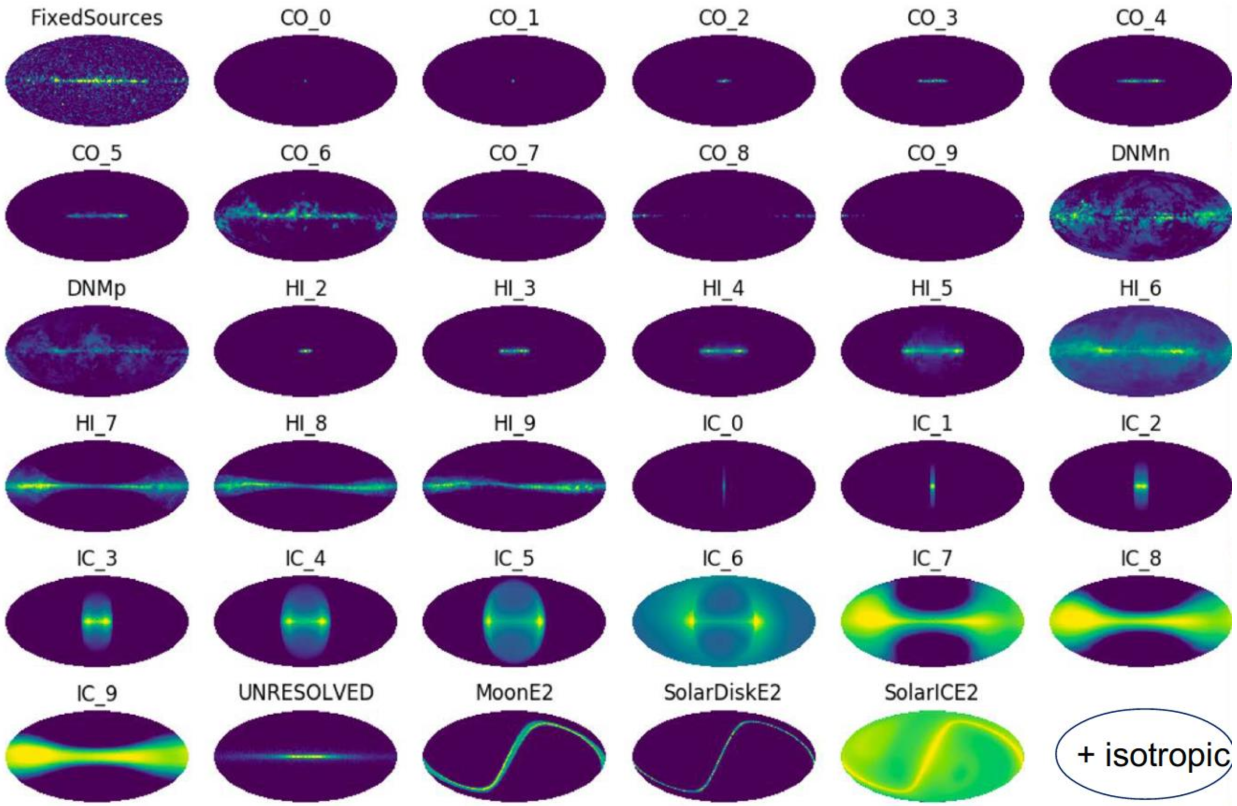
- Mean of 10 years data subtracted from the yearly counts. Variability component dominates.

Training Data Generation: Variability of Blazars

- Variability: Considered Yearly Flux/Mean Flux (10 years)
- All Blazars are likely variable, but fainter sources contain large statistical uncertainty.
- Consider yearly photon measurement and sum over 10 years to get total photon count.



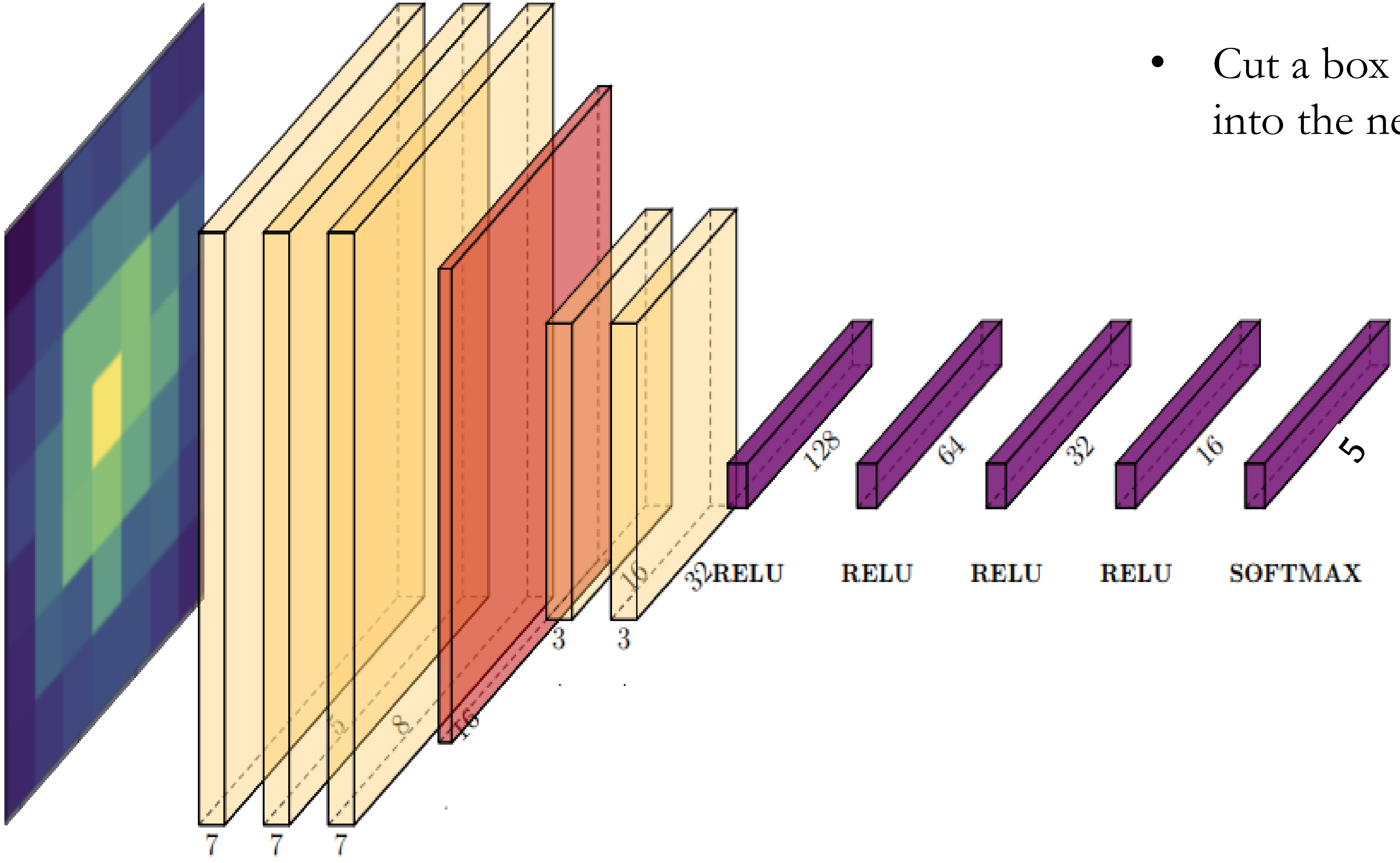
Traditional Method



[ICRC 2019, I. Moskalenko, G. Johannesson; GALPROP]

- Build a model of interstellar emission (IEM) using various templates.
- Find sources using maximum likelihood.
- Test Statistic $TS = 2\log\frac{L}{L_0}$, how significantly a source emerges from the background.
- For 4FGL catalog, requirement is set as $TS > 25$

Classification Network



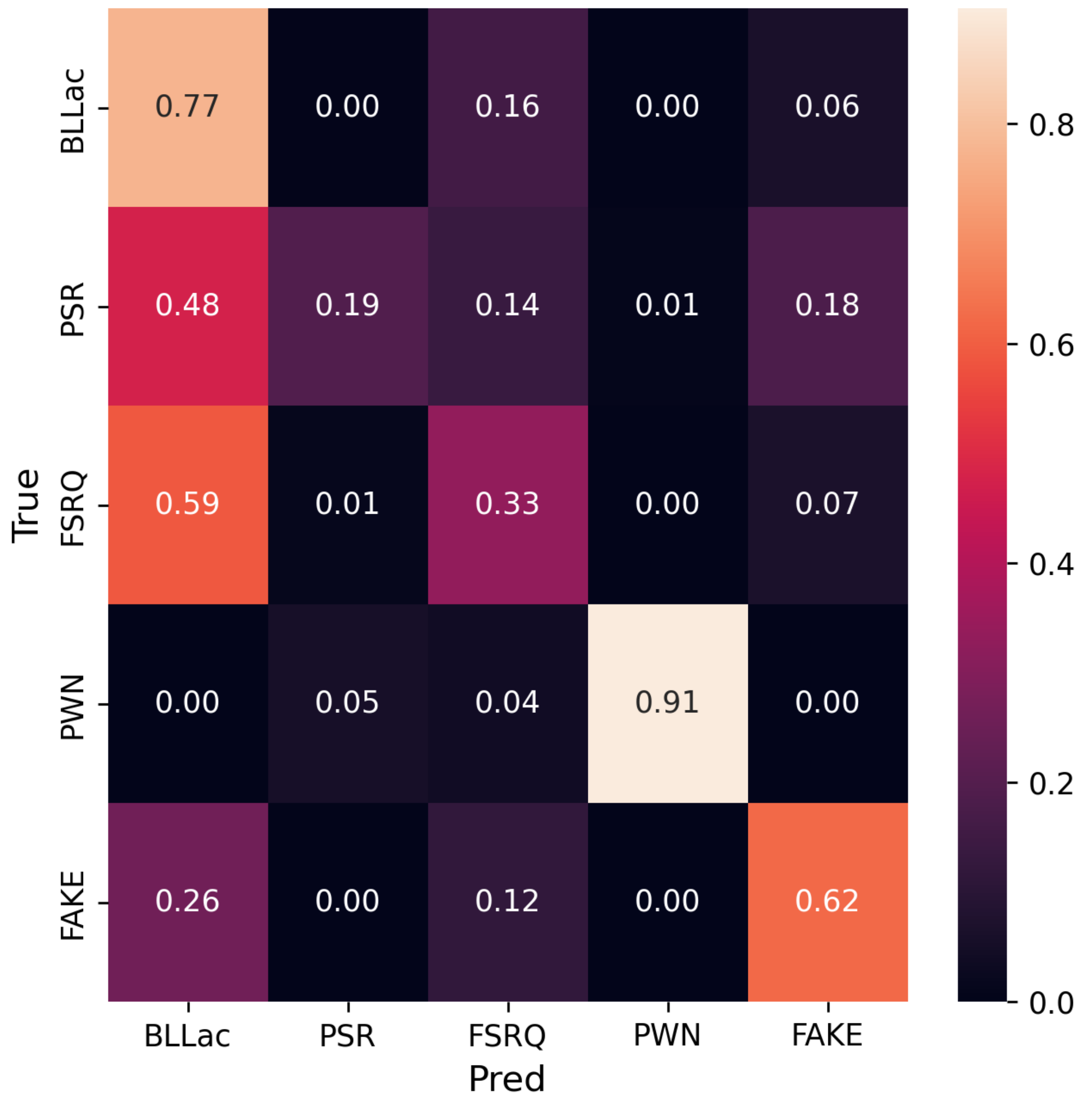
- Localized sources are then acting as inputs for a separate classification network.
- Cut a box around the predicted location and feed into the network.

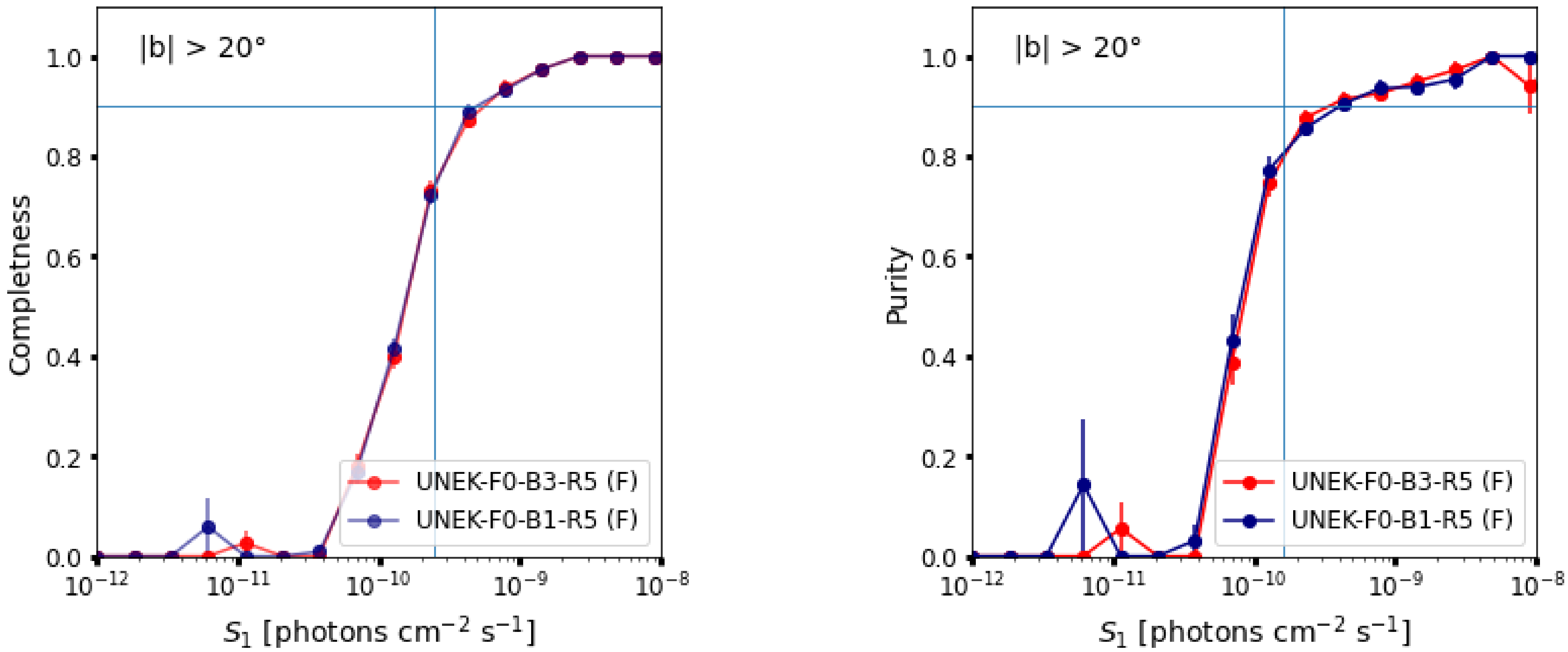
- Classification Network is a 3D CNN.
- Input shape (W, H, 10, 6).
- 10 years, 6 energy bins.
- Class-imbalance problem

Classification Performance (Under Progress)

Performance improvement

- with 4D data structure and 3D Convolution. Treating time and energy separately.
- with arranging max count on top considering yearly photon counts.
- with subtracting the mean (of 10 year counts) from yearly data.
  - Still in progress with some fine tuning of the network.




|b| > 20°
1.0
0.8
0.6
0.4
0.2
0.0
Completness
10^{-12}
10^{-11}
10^{-10}
10^{-9}
10^{-8}
S_1 [photons cm^{-2} s^{-1}]
UNEK-F0-B3-R5 (F)
UNEK-F0-B1-R5 (F)
|b| > 20°
Purity
UNEK-F0-B3-R5 (F)
UNEK-F0-B1-R5 (F)
S_1 [photons cm^{-2} s^{-1}]