Classification of Fermi-LAT sources with machine learning and dataset shifts

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Dmitry Malyshev

RICAP, 23.09.2024 – 27.09.2024

Fermi Gamma-ray Space Telescope

DFG Deutsche Forschungsgemeinschaft German Research Foundation

ECAP ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Plan

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- Dataset shifts (difference in the distribution of associated and unassociated sources)
  - Covariate shift
  - Prior shift
- Data selection and definition of classes
- Covariate shift model of gamma-ray sources
- Prior shift model of gamma-ray sources
- Reconciling covariate and prior shift approaches

Survey bias

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- United states 1936 elections, Landon vs Roosevelt
  - Literary Digest postal survey of 2.4 million people: Landon wins
  - Gallup survey of 50 thousand people: Roosevelt wins
- Roosevelt won – why?
  - Isn't more data = better results
  - Sampling bias – distribution of Literary Digest readers is different from the distribution of US voters
- Not the only case, US elections 1948, Dewey vs Truman
  - Chicago Tribune used results of a Gallup phone survey
  - Phones were mostly owned by wealthy people at that time

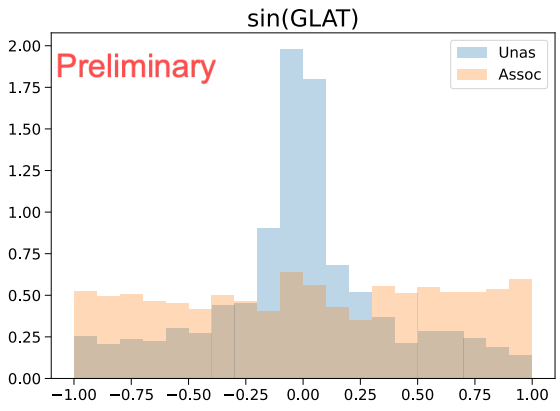
1948 president-elect (not Dewey)

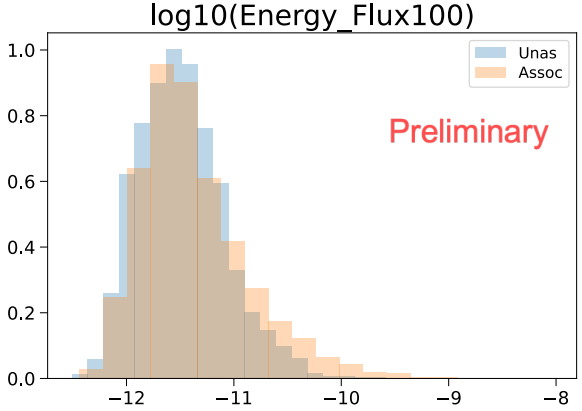
Wikimedia commons

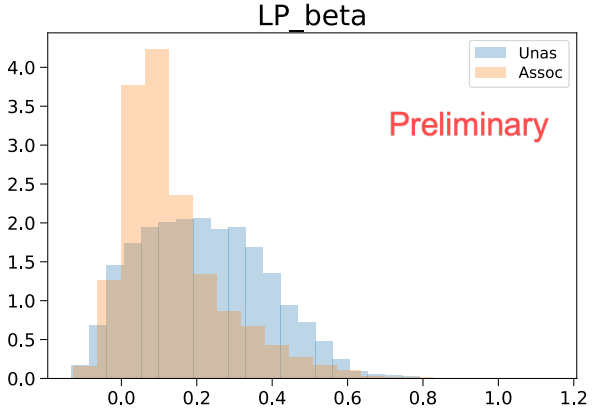
Classification of Fermi-LAT sources

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- Big problem: about 1/3 of Fermi-LAT sources are unassociated
  - Use associated sources to train machine learning (ML) algorithms to probabilistically determine classes of unassociated sources?
- Analogy with elections?
  - Survey = associated sources
  - Election = unassociated sources







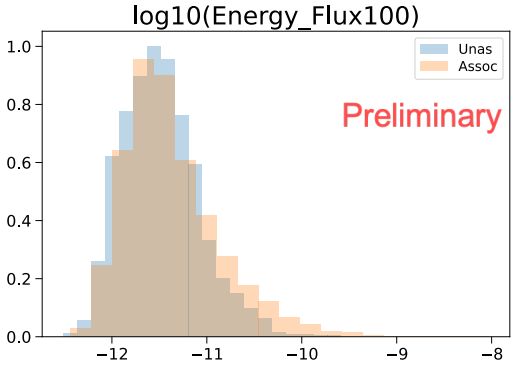
- Are we making the same mistake as in the biased election surveys?
- Does it affect the results of ML classification of unassociated sources?

Dataset shifts

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- Standard ML assumption $P_{train}(X, k) = P_{target}(X, k)$
  - X – input features, k – output features, e.g., classes in classification
- Dataset shift: $P_{train}(X, k) \neq P_{target}(X, k)$
- There are two special cases of dataset shifts

  Moreno-Torres et al, Pattern Recognit. 45 (2012)
- Covariate shift:
  - $P(X, k) = P(k|X)\, P(X)$
  - $P_{train}(X) \neq P_{target}(X)$ but we assume that $P_{train}(k|X) = P_{target}(k|X)$
  - In this case the difference in distributions of associated and unassociated sources affects all classes proportionally, i.e., conditional class probabilities $P(k|X)$ are not affected (by the covariate shift assumption)



- Prior shift:
  - $P(X, k) = P(X|k)\, P(k)$
  - $P_{train}(k) \neq P_{target}(k)$ but we assume that $P_{train}(X|k) = P_{target}(X|k)$

Data selection and classification algorithm

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

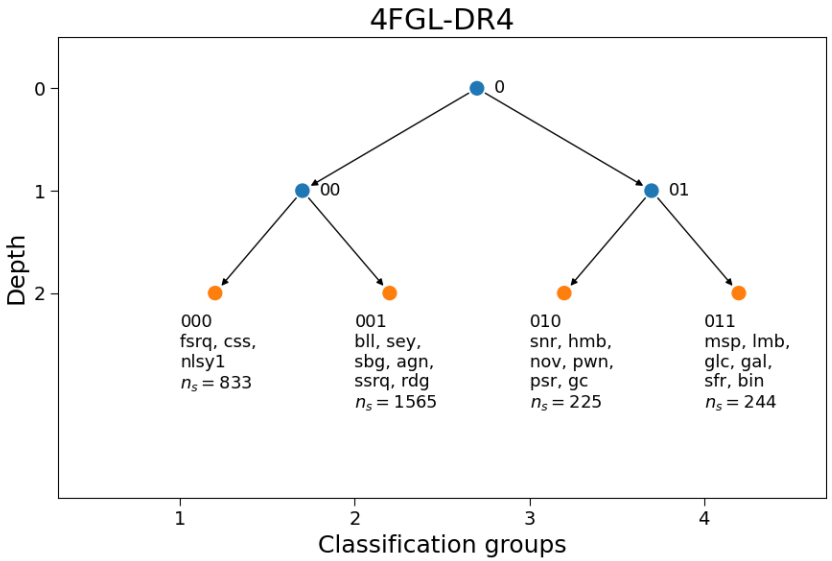
- We use 4FGL-DR4 (v34) catalog
- Features (no coordinate features):
  - Covariate shift model (7 features): 'log10(Energy_Flux100)', 'log10(Unc_Energy_Flux100)', 'log10(Signif_Avg)', 'LP_index1000MeV', 'LP_beta', 'LP_SigCurv', 'log10(Variability_Index)'
  - Prior shift model (3 features): 'log10(Energy_Flux100)', 'LP_beta', 'log10Epeak'
- Four classes (no bcu or spp sources) determined using hierarchical class division with Gaussian mixture models (Malyshev&Bhat 2023)

  fsrq+: fsrq, nlsy1, css

  bll+: bll, sey, sbg, agn, ssrq, rdg

  psr+: snr, hmb, nov, pwn, psr, gc

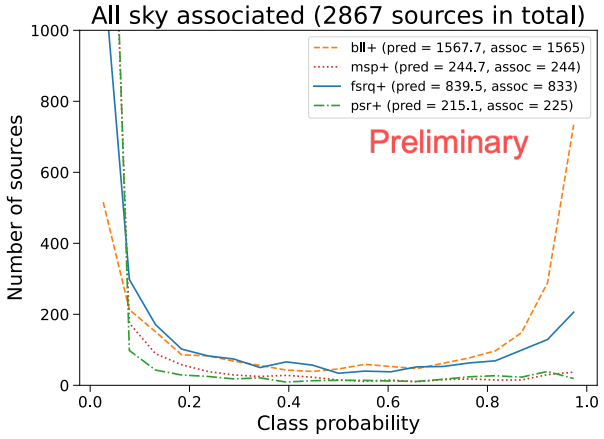
  msp+: msp, lmb, glc, gal, sfr, bin
- Classification: random forest

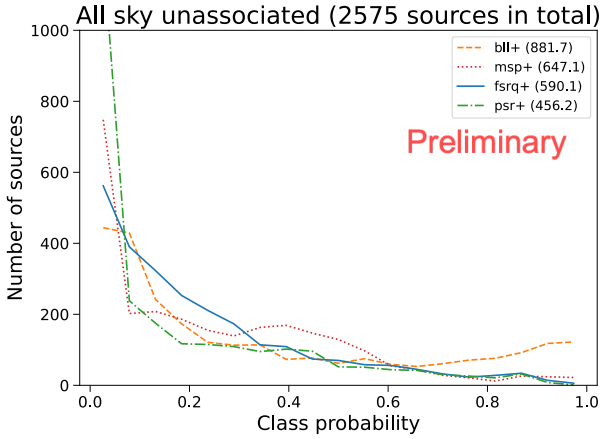


Covariate shift model

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- Use classification of associated sources to predict classes for unassociated sources
- Here are the histograms of the distributions of class probabilities for associated (left) and unassociated (right) sources
  - We use 70/30% split into training and testing samples and repeat the split randomly until each associated source falls at least 5 times in testing samples. The class probabilities for associated sources are obtained by averaging over testing samples





- We see that for the associated sources there is a reasonable separation of sources, but for the unassociated sources there are few sources with p > 0.5

Prior shift model

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- In this model, the main assumption is that the distributions of sources are the same for associated and unassociated sources in each class

$$p_{\mathrm{unas}}(x|k) = p_{\mathrm{assoc}}(x|k)$$

where x are input features and k are classes.

- The model for the unassociated sources is then

$$p_{\mathrm{unas}}(x) = \sum_k p_{\mathrm{assoc}}(x|k)\pi_k$$

where π_k is the frequency of class k.

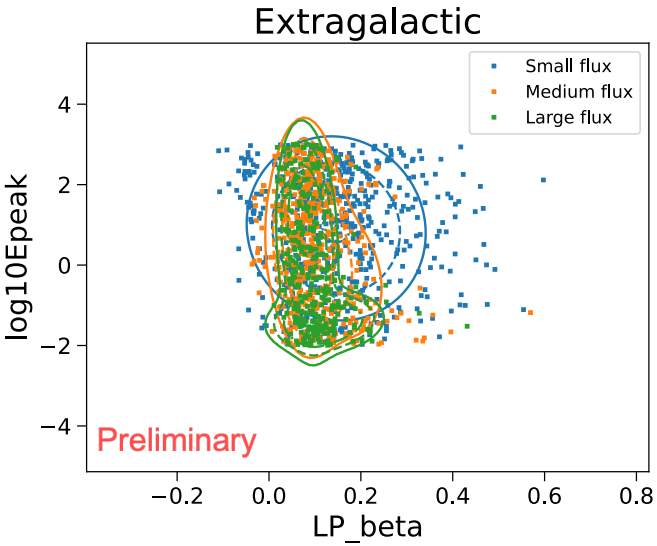
- The coefficients π_k are determined from the fit of the model to the data, i.e., the distribution of unassociated sources, by maximizing the unbinned Poisson log likelihood

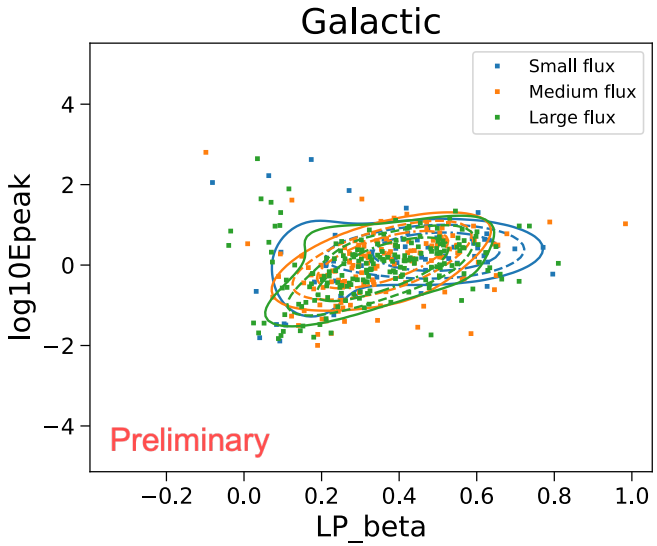
$$\log L(p_k) = \sum_{i \in \mathrm{unas}} \log(p_{\mathrm{unas}}(x_i)) - N_{\mathrm{unas}} \int p_{\mathrm{unas}}(x)dx$$

Dependence of distributions on source flux

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- One of the caveats of the prior shift model is that the distributions of associated sources in a class *k* and a possible distribution of class-*k* sources among the unassociated sources is not necessarily the same
  - For instance, the distribution of extragalactic sources changes as a function of flux, while the distribution of Galactic sources does not depend strongly on the flux

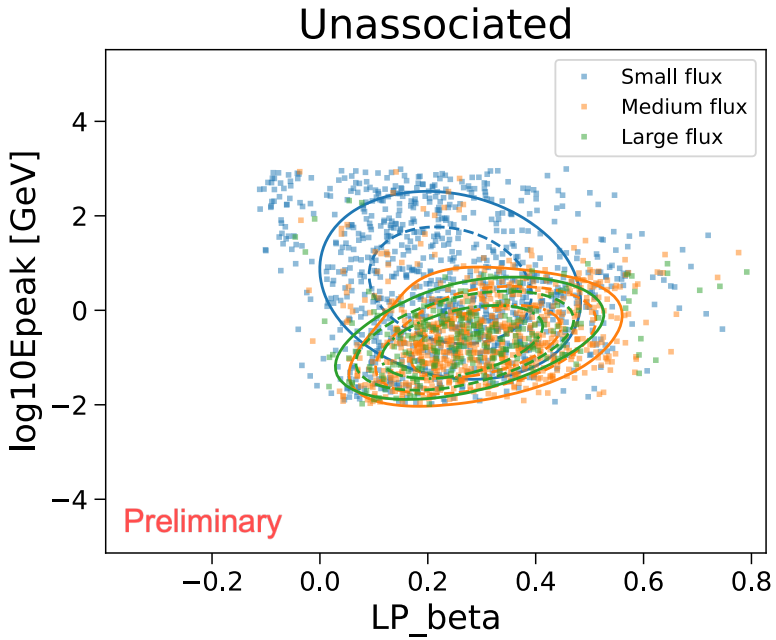




Distribution of unassociated sources

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- The distribution of unassociated sources is dominated by Galactic sources at medium and high fluxes and by extragalactic sources at small fluxes.
- In order to be able to fit the distribution of unassociated sources, one needs to suppress the contribution of high-flux extragalactic sources and enhance the contribution of low-flux extragalactic sources
- This can be done with flux-dependent prior shift model



Flux-dependent prior shift model

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- The model is a relatively simple generalization of the standard prior shift model

$$p_{\text{unas}}(x) = \sum_k p_{\text{assoc}}(x|k)\pi_k(x)$$

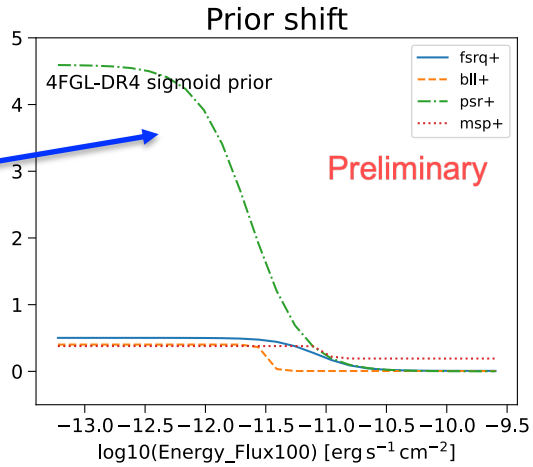
where $\pi_k(x) = \pi_k(\log10(\text{Energy\_Flux100}))$ modeled as a sigmoid function plus a constant

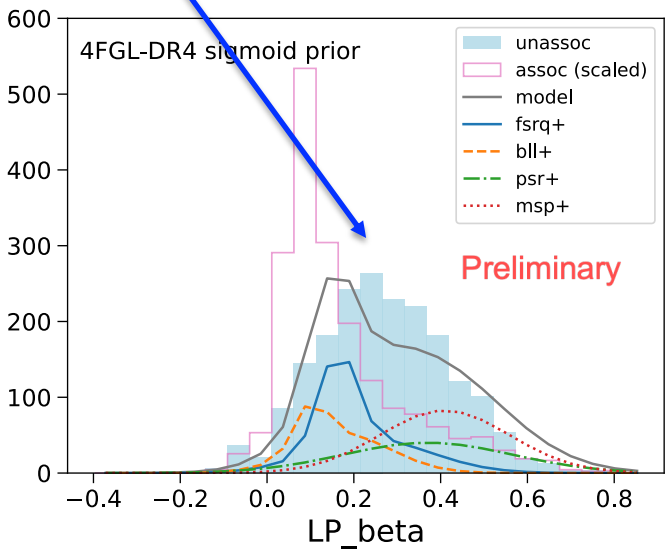
$$\sigma(x) = \frac{a}{1 + e^{(x-b)/c}} + d$$

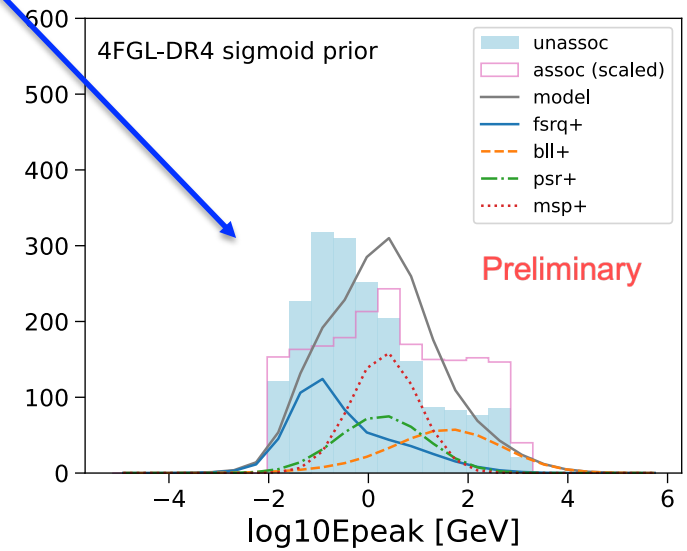
- NB: $\pi_k(x)$ are not probabilities anymore, but they are normalized with an overall factor to ensure that $p_{\text{unas}}(x)$ is a PDF.

Flux-dependent prior shift

- Prior shifts enhance contribution of low-flux sources
  - Looks unphysical for the psr+ class
- The model deviates from the distribution of unassociated sources at intermediate beta values and small Epeak values.







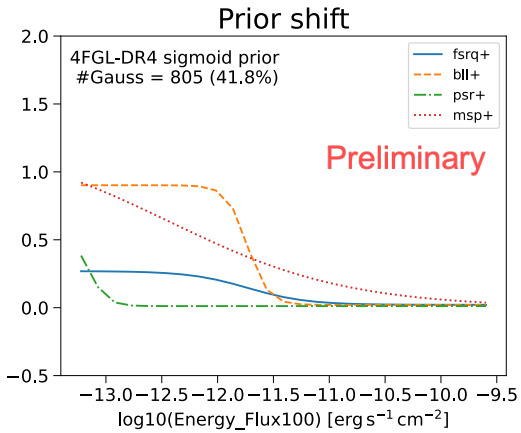
- A new component is needed?

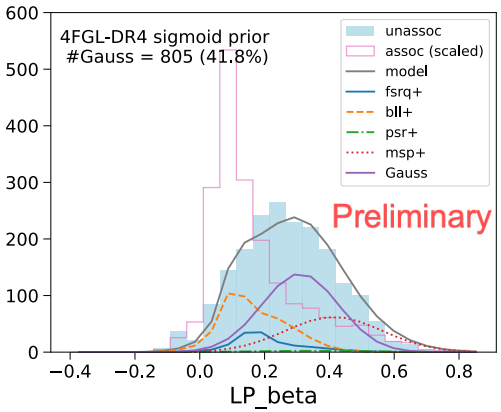
Flux-dependent prior shift + Gaussian

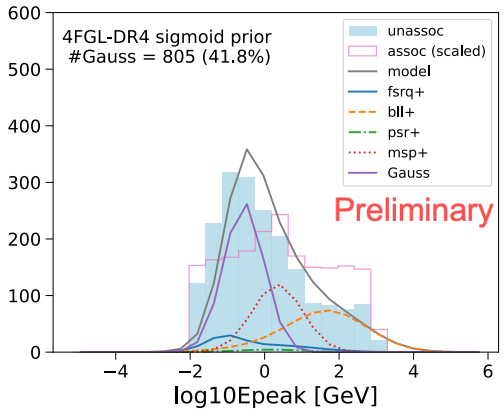
ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- We add a Gaussian in the three input features
  - 7 parameters: normalization, means and sigmas in the 3 coordinates

- The model looks more reasonable and fits the data better
  - However, it predicts that about 40% of unassociated sources belong to a new component modeled by the Gaussian



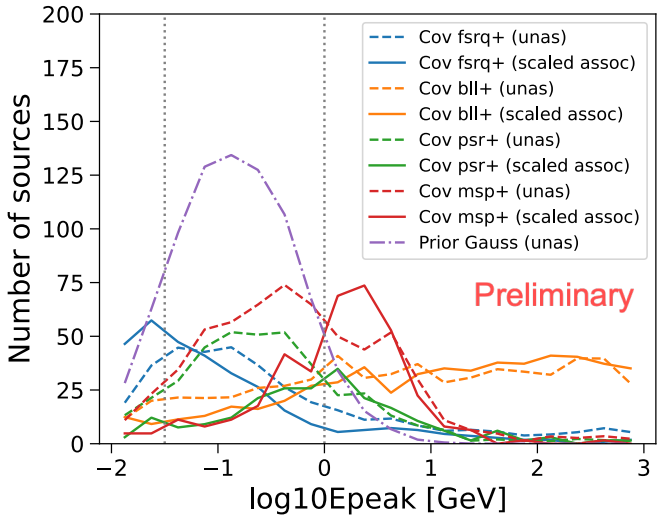


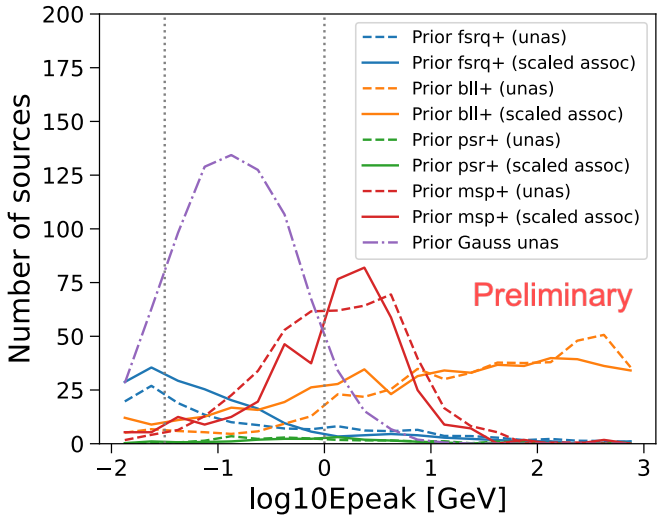


Predicted distributions of unassoc sources

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- On the left (right) we compare the prediction for the covariate shift (prior shift) model with the distributions of associated sources scaled to match the number of unassociated sources with Epeak > 1 GeV or Epeak < 30 MeV
  - In both cases we also show the Gaussian component from the prior shift model
  - Covariate shift predictions agree with the distributions of associated sources outside of the Gaussian component domain

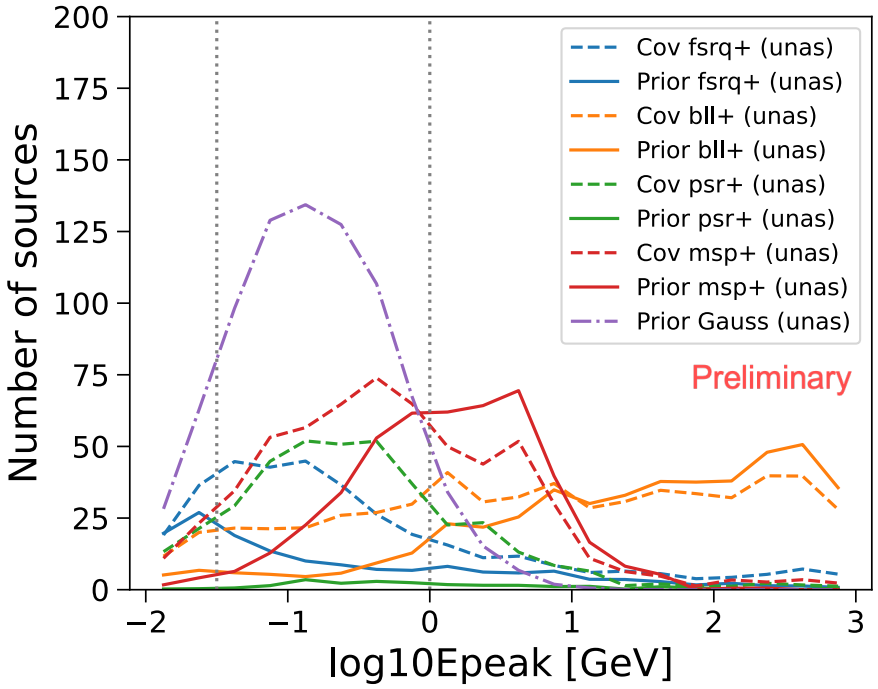




Covariate vs prior shift models

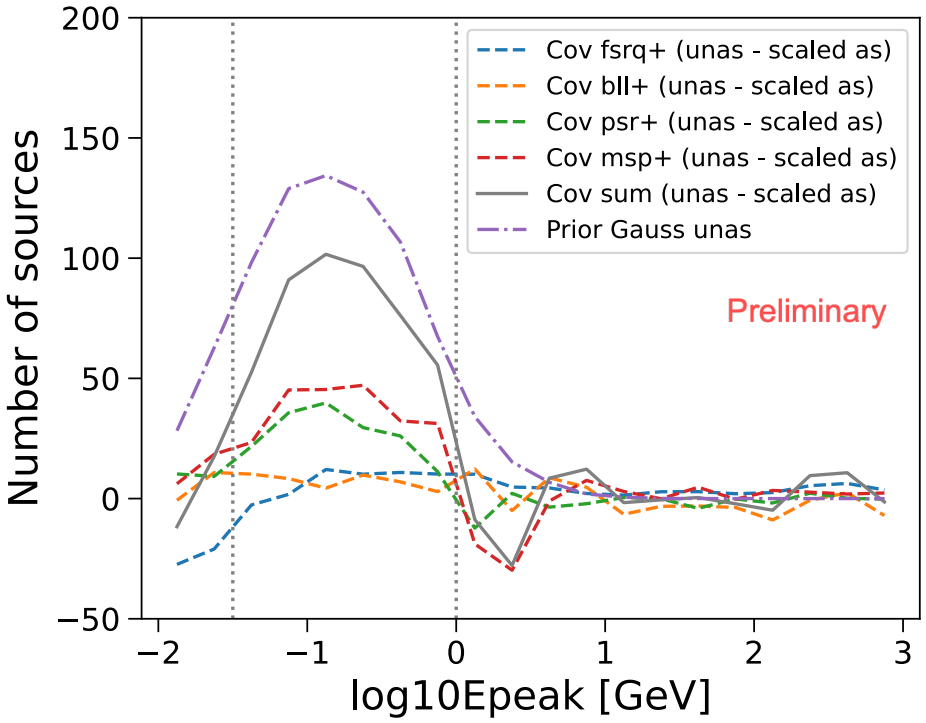
ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- Below we compare the prediction for the contributions of the 4 classes for the covariate and prior shift models
  - The predictions disagree mostly in the area, where there is a significant contribution from the Gaussian component in the prior shift model



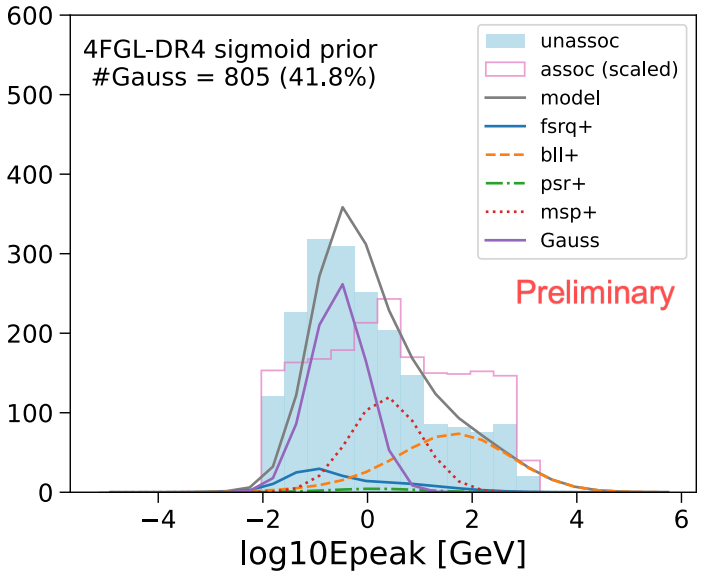
Residual predictions

- We subtract the scaled distributions of associated sources from the predictions for the unassociated sources in the covariate shift case
  - The sum of the differences is similar to the Gaussian component in the prior shift model!



Conclusions

- There seems to be an evidence for a new component in the distribution of unassociated sources
  - NB. The discussion of soft Galactic unassociated (SGU) sources in the 4FGL-DR3 and 4FGL-DR4 catalog papers (2201.11184, 2307.12546)
- Covariate and prior shift ML models generally agree with each other, but the covariate shift model tries to accommodate a possible new component by scaling proportionally the existing four classes in the area, where the new (Gaussian) component is present in the prior shift model
- The nature of this component is not clear:
  - Mismodeling of diffuse emission?
  - Sub-population of an existing class, e.g., pulsars or MSPs?
  - A “new” population of sources, e.g., young star clusters (Peron+ 2024)



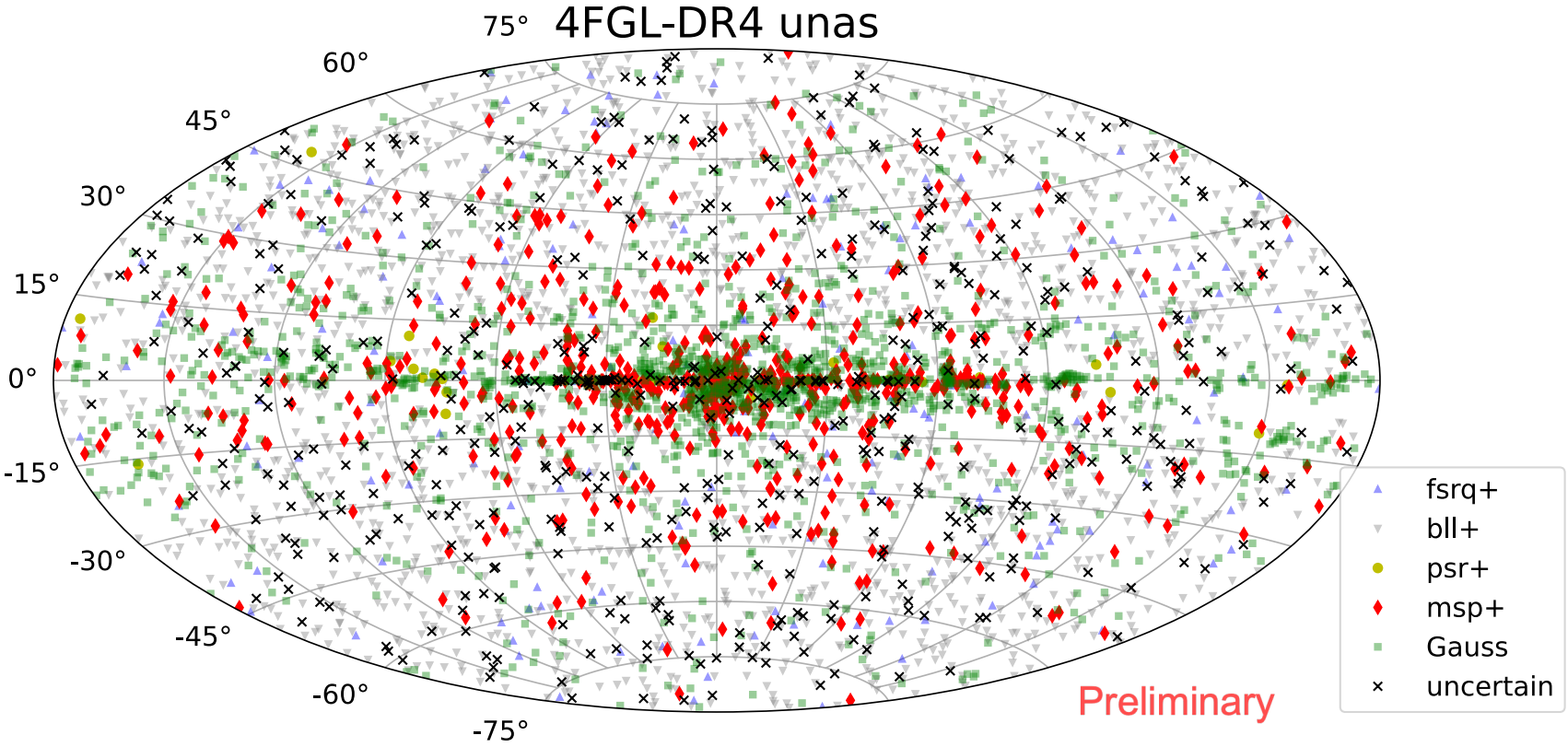
Bonus slides

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

Distribution of source candidates

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- Class candidates have reasonable distribution on the sky, bll+ and fsrq+ sources are ~ isotropic, psr+ candidates are close to the Galactic plane, msp+ candidates have a broad distribution around the Galactic plane.
  - Gaussian sources have a similar distribution to msp+ sources – local?



Distribution of fluxes

ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

- The fluxes of sources in the Gaussian component are not very small
  - They are generally larger than unassociated sources attributed to bll+ or fsrq+ sources and comparable to fluxes of msp+ candidates
  - Also an argument in favor of local nature?