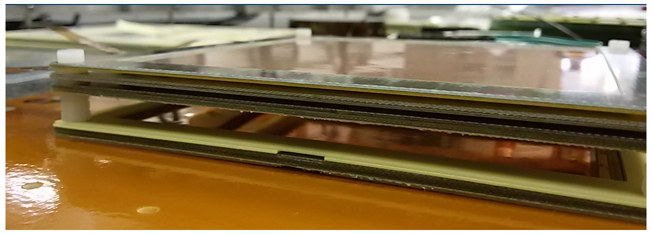
LIME clustering and energy response

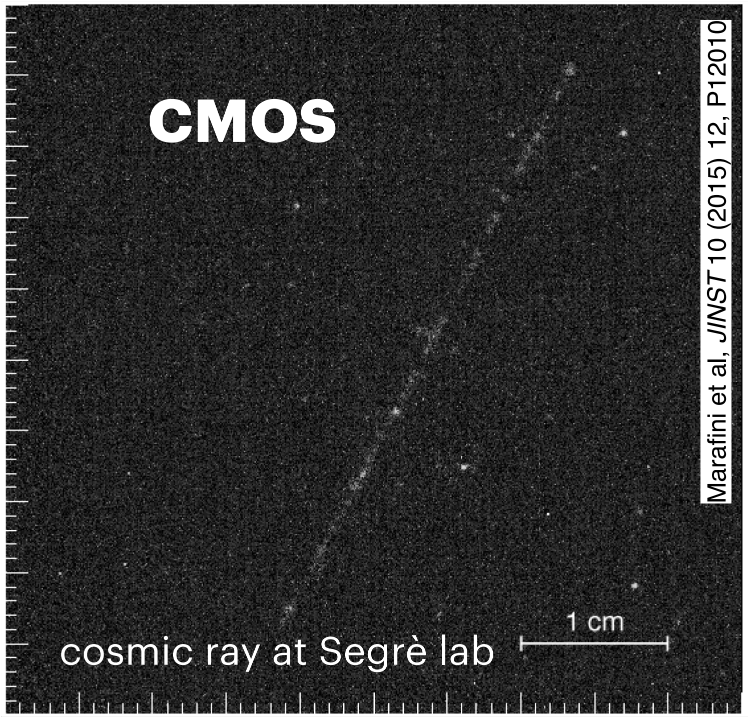
E. Di Marco

CYGNO collaboration meeting,
20 December 2021

Detector evolution

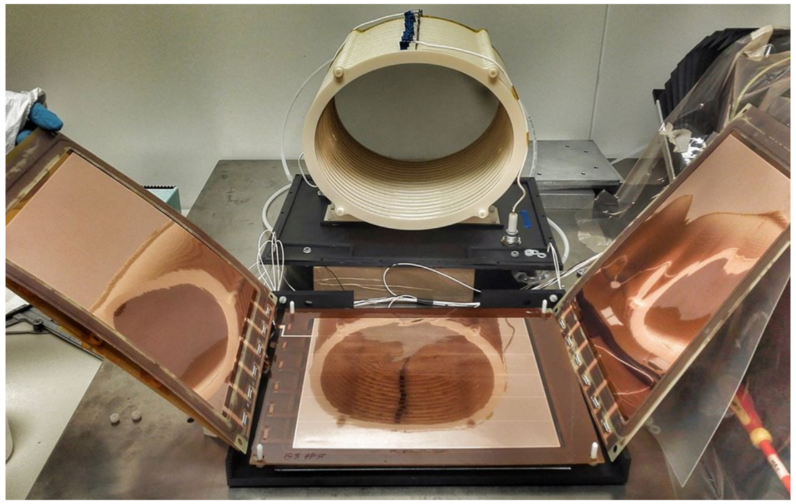
ORANGE:
10 x 10cm^2
1 cm sensitive gap

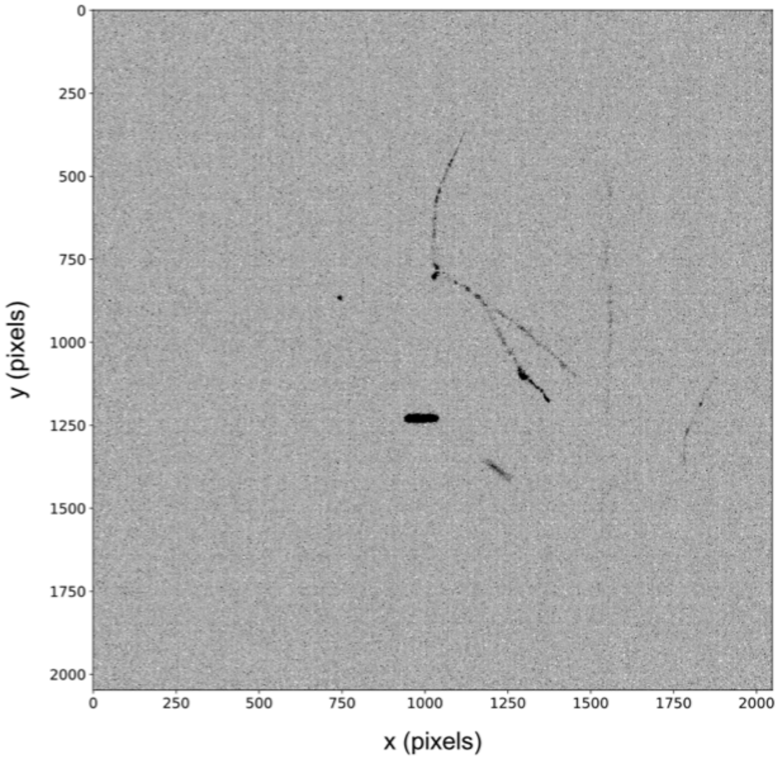




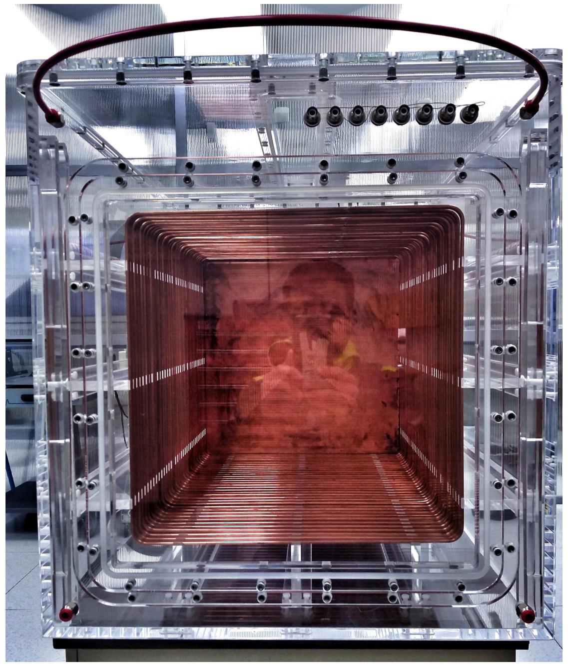
Marafini et al, *JINST* 10 (2015) 12, P12010

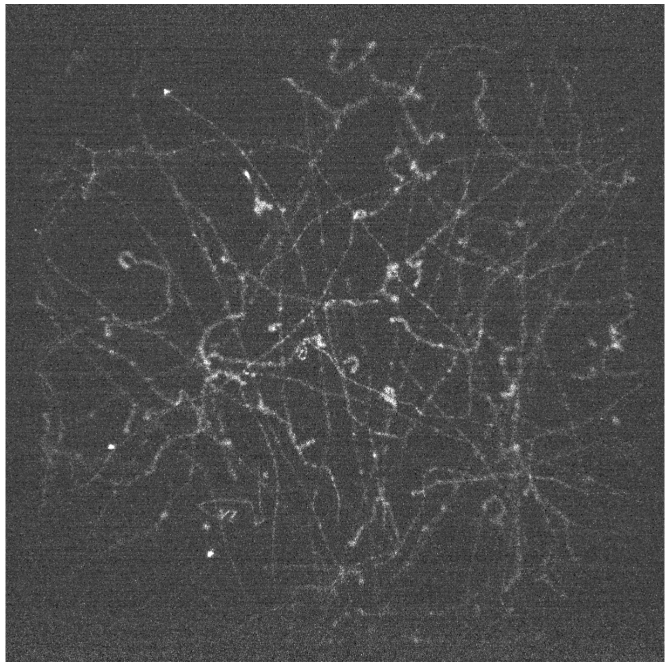
LEMON:
500 cm^2
20 cm sensitive gap





LIME:
1000 cm^2
50 cm sensitive gap





Every season needs its clothes

ORANGE:
10 x 10cm²
1 cm sensitive gap

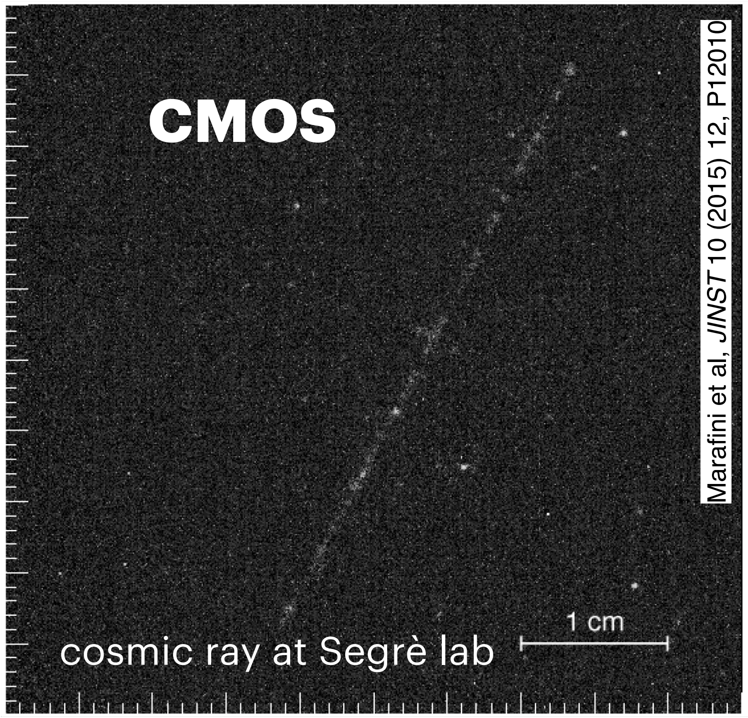
bkg occupancy:
≤ 1 track / event

LEMON:
500 cm²
20 cm sensitive gap

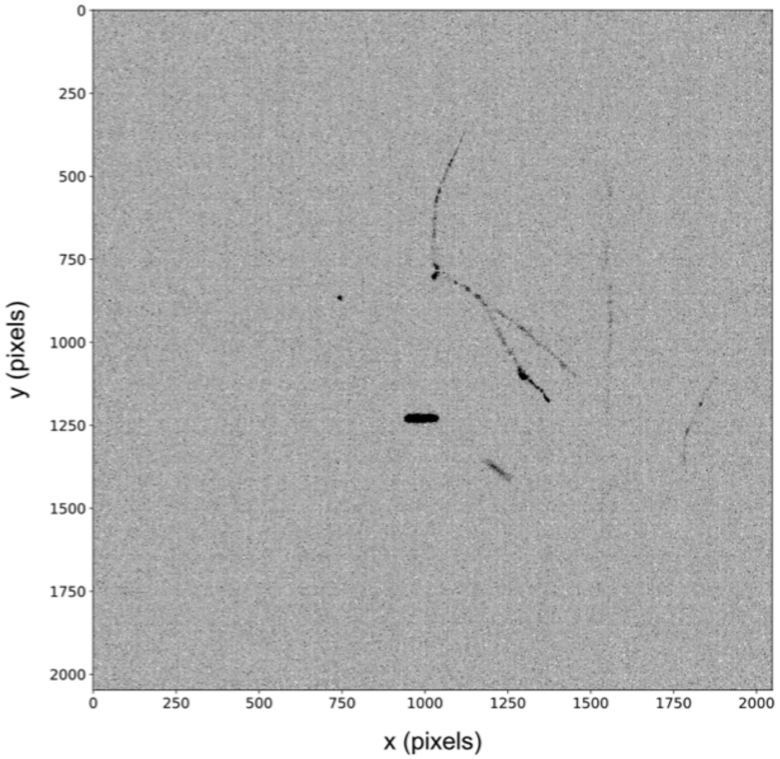
bkg occupancy:
1 ≤ tracks/event ≤ 10

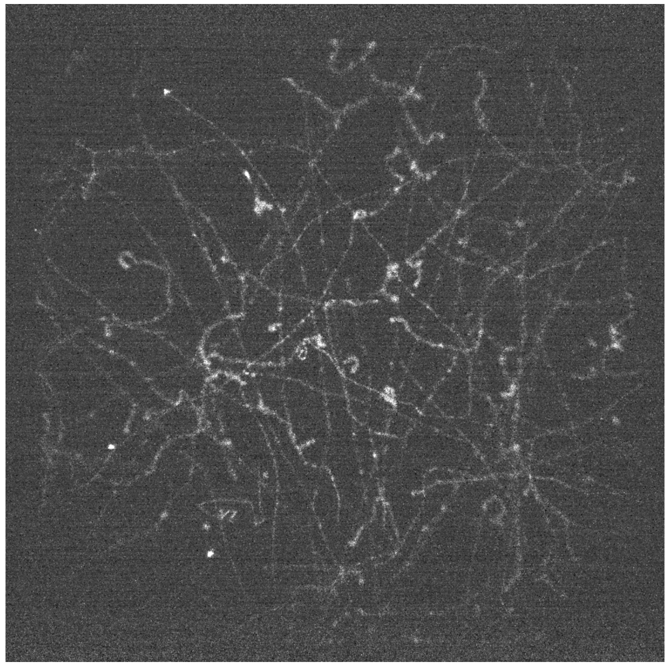
LIME:
1000 cm²
50 cm sensitive gap

bkg occupancy:
10 ≤ tracks/event ≤ 50
with overlaps



Marafini et al, *JINST* 10 (2015) 12, P12010





Clothes used so far

ORANGE:
10 x 10cm^2
1 cm sensitive gap

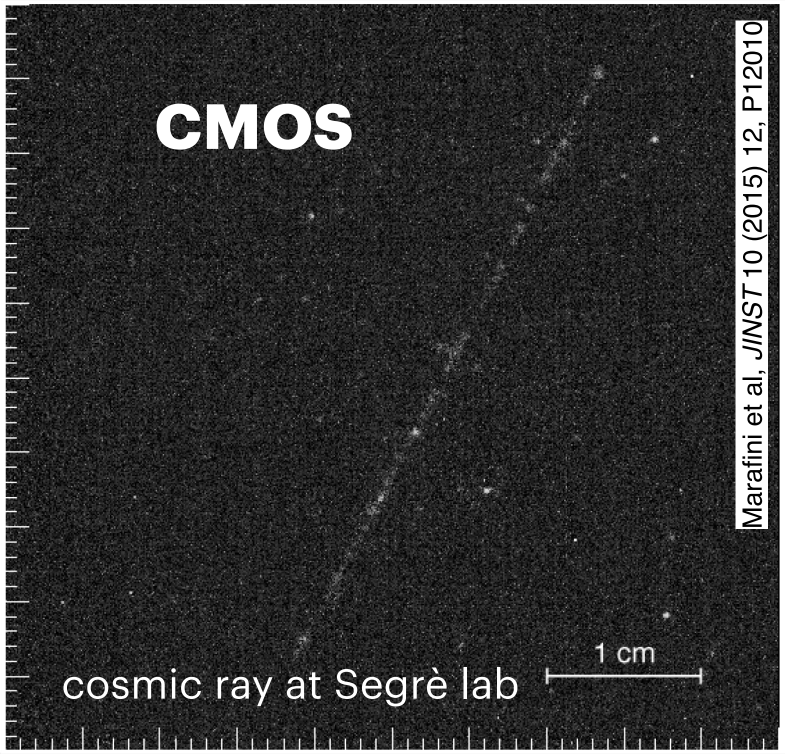
NNC, DBSCAN

LEMON:
500 cm^2
20 cm sensitive gap

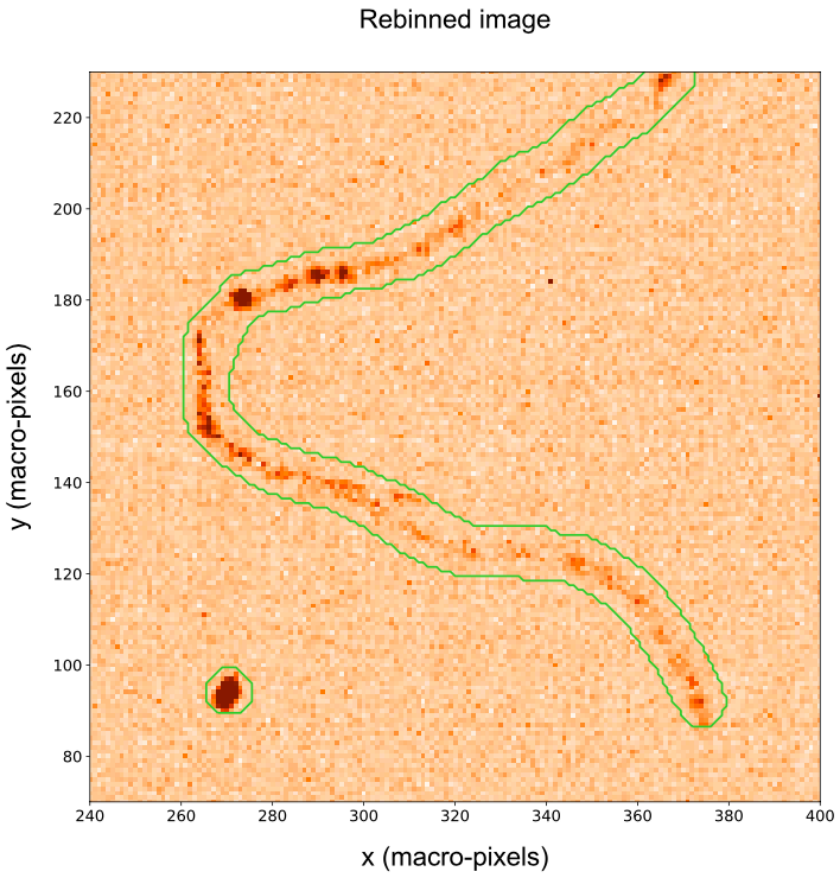
Geodesic Active Contours (GAC) to reconstruct both long and short tracks

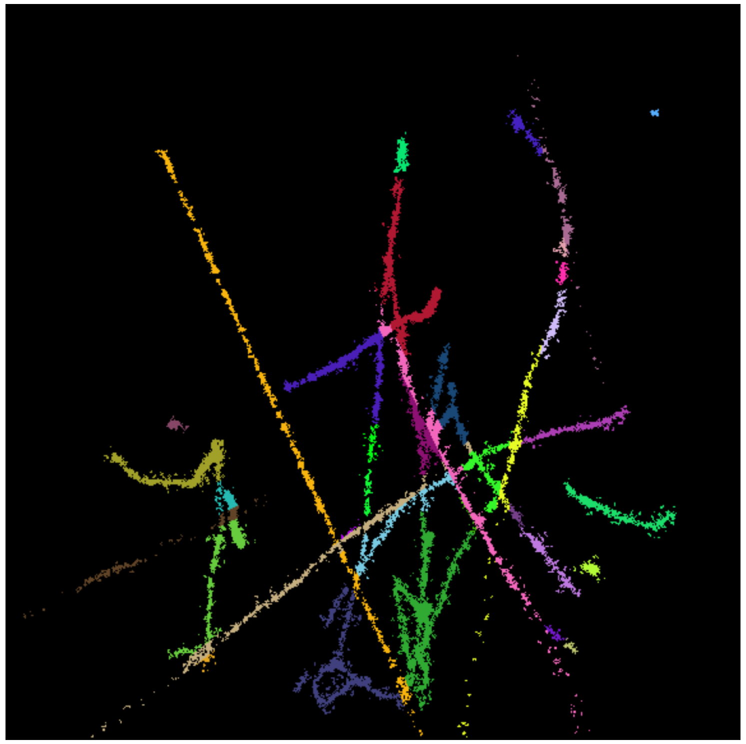
LIME:
1000 cm^2
50 cm sensitive gap

directional DBSCAN for the long and overlapping tracks and **DBSCAN** for the remaining



Marafini et al, *JINST* 10 (2015) 12, P12010





LIME data at LNF

Frascati is **notoriously** a radioactive place and exposed to a continuous shower of cosmic rays => *in any data taken so far, we have this background overlapped*

Occupancy depends on the volume (fixed), but also with the exposure:

- data taken with the DAQ has a minimum exposure of **200 ms => o(50 tracks/event)**
- exposure can be reduced with data taken by hand with HOKAWO. We took some data with **50ms => o(10 tracks / event)**

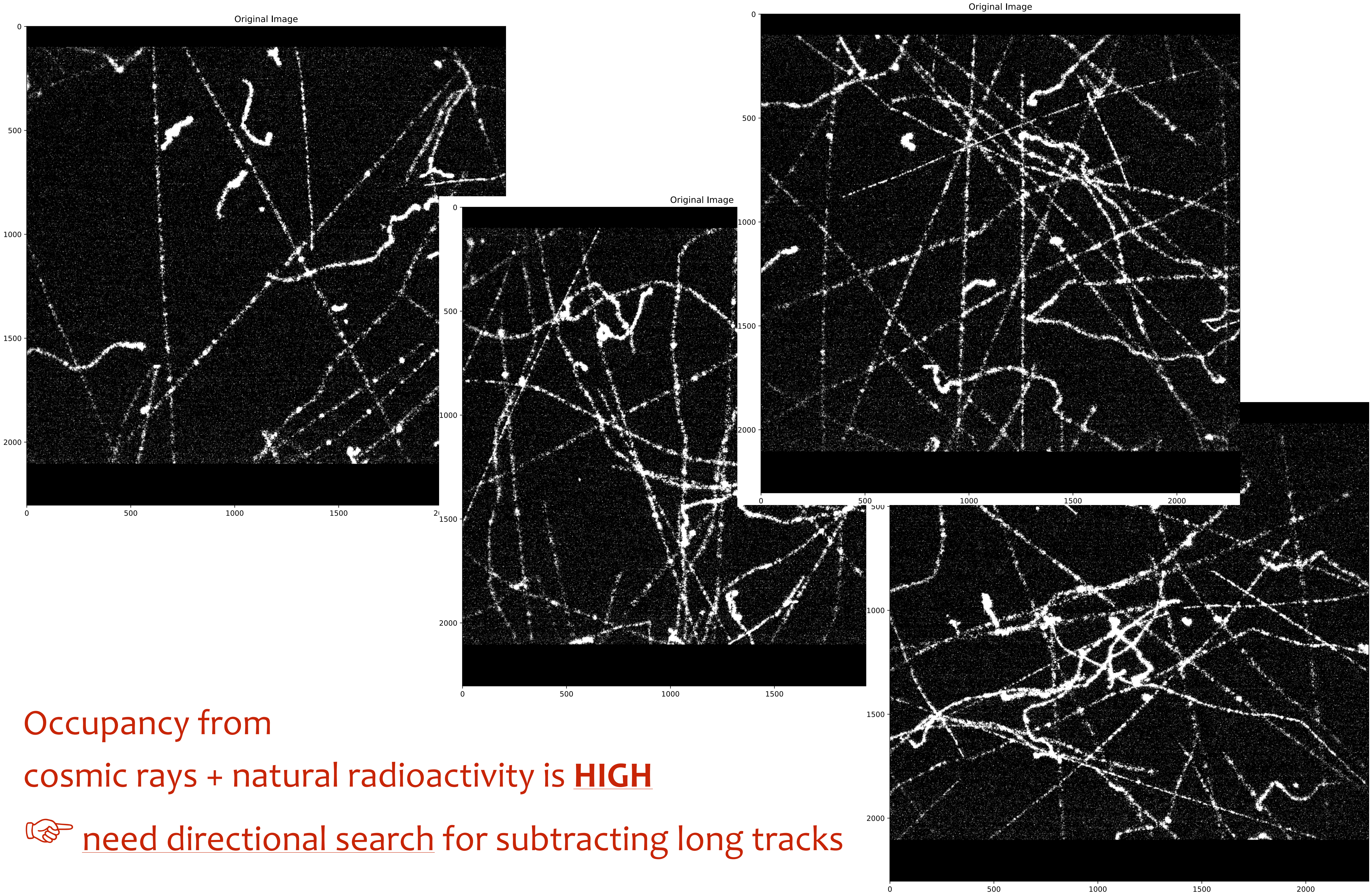
This talk focuses on results based on both types of data: they used the same clustering, geometrical and response corrections and analysis method.

Some parameters, though, are fine-tuned for 50ms or 200ms exposure.

BTW, what is next season, and which clothes we have to prepare?

LIME will go ***under Gran Sasso*** *soon*, so probably the occupancy will be **<2 tracks/event** => something naive and simple, as NNC or DBSCAN will be sufficient: back to the origin

Typical occupancy



Occupancy from

cosmic rays + natural radioactivity is **HIGH**

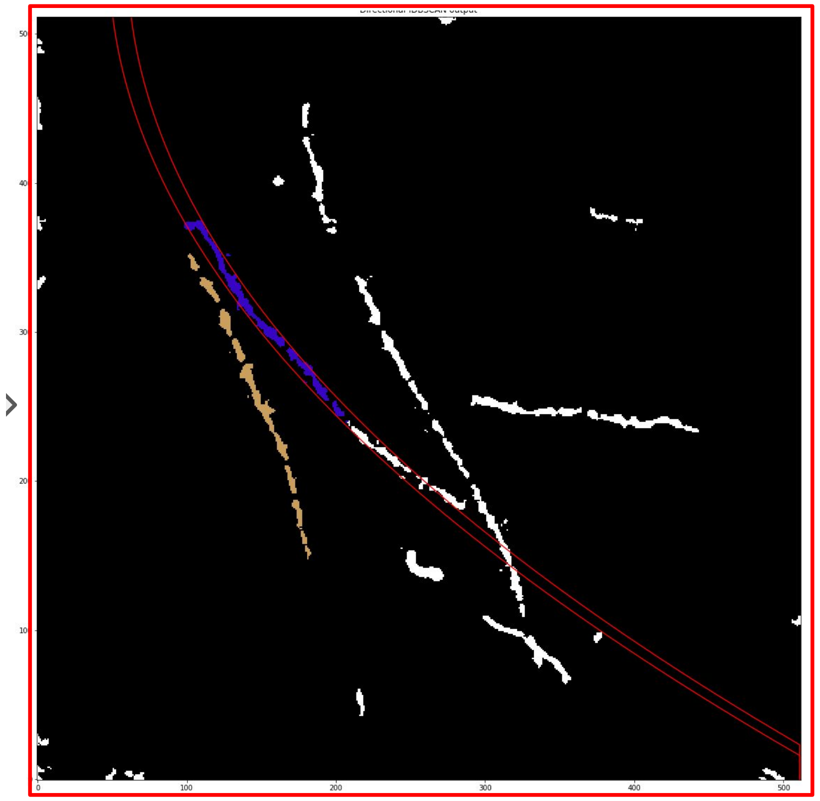
☞ need directional search for subtracting long tracks

Directional tracking

For this I. Pains has developed a clustering that search for patterns compatible with polynomials (line or 3rd order polynomial). Links to presentations here and here.

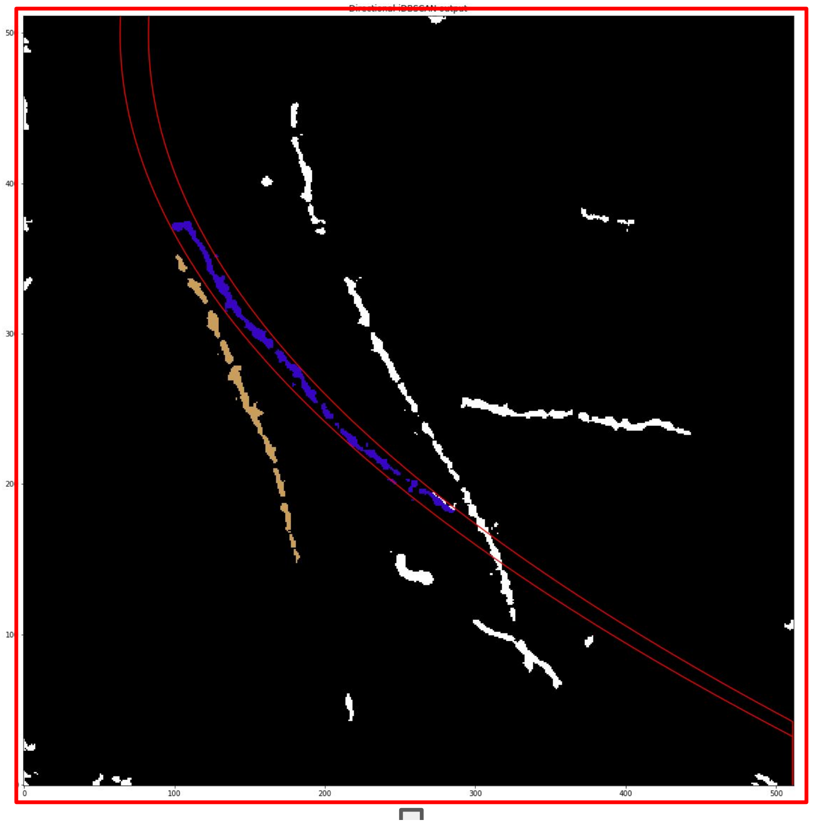
Reminder of the method:

- starts with DBSCAN with a short radius
- tests if starting from these clusters, one can find other clustered points compatible with a polynomial
    - the polynomial is fitted iteratively until points are added to the supercluster



The next one needed more steps to finish

I. Pains



Customization for high pileup

- As soon as the occupancy increases, everything gets merged when the “seed” cluster is in a crowded region

- it is slow, because of the many fits/seed done

- 3-rd order polynomial sometimes not sufficient, but fitting with higher order can get crazy soon

ATTEMPTs explored to improve:

1. Use “isolated” seeds to start directional search, i.e. with the miniminum $I = \sum_{i}^{\Delta R=200} A_i$. ($A_i$ = i—pixel amplitude). If >1 has I=0, then sort by the best linear fit X^2.

2. Use Bernstein polynomials to approximate the curve, to improve stability

3. Each pixel is “weighted” proportionally to its intensity to improve the contrast

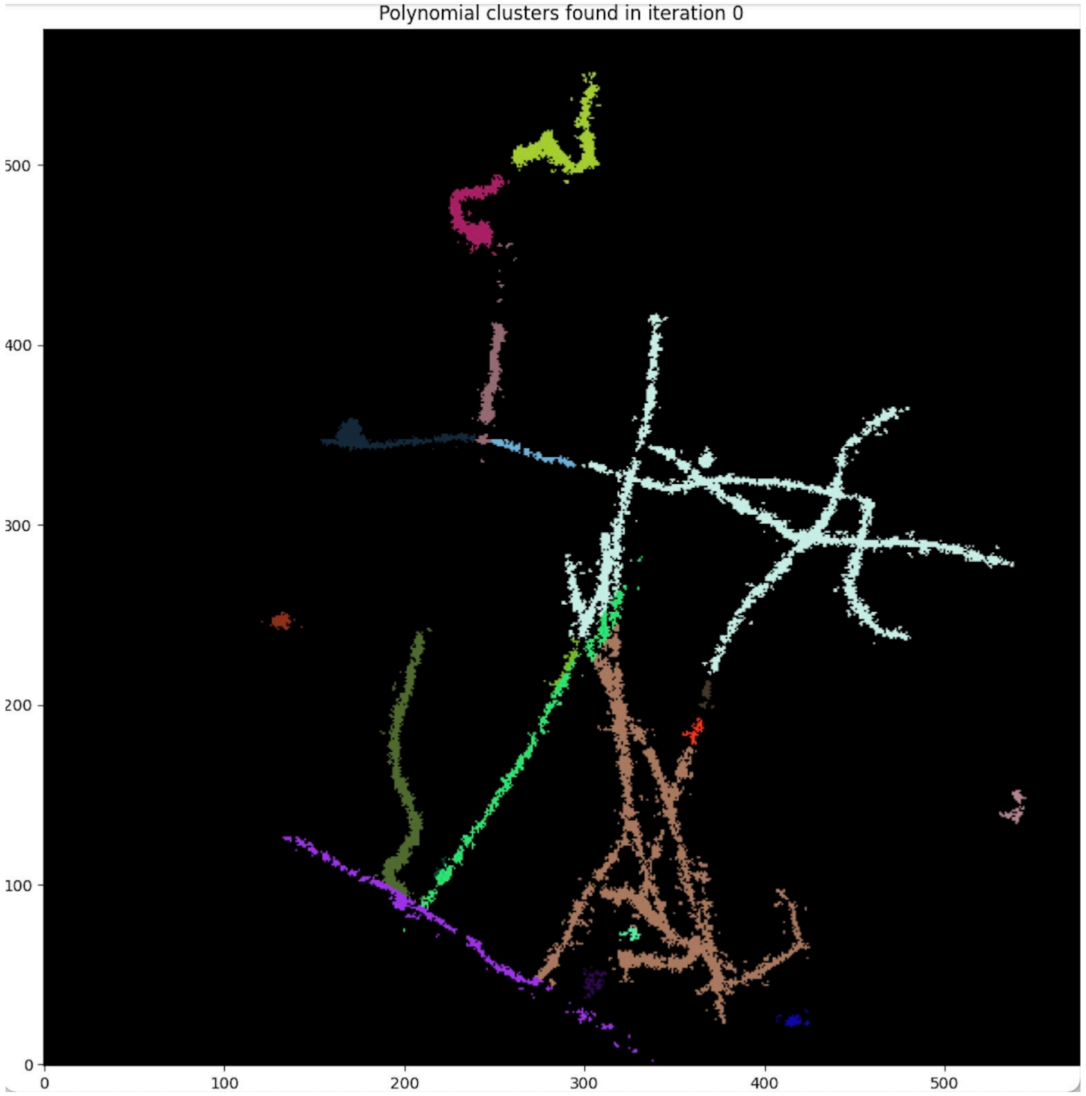
4. The remaining clusters are done without fitting, with naive DBSCAN, only if they are isolated by directional clusters

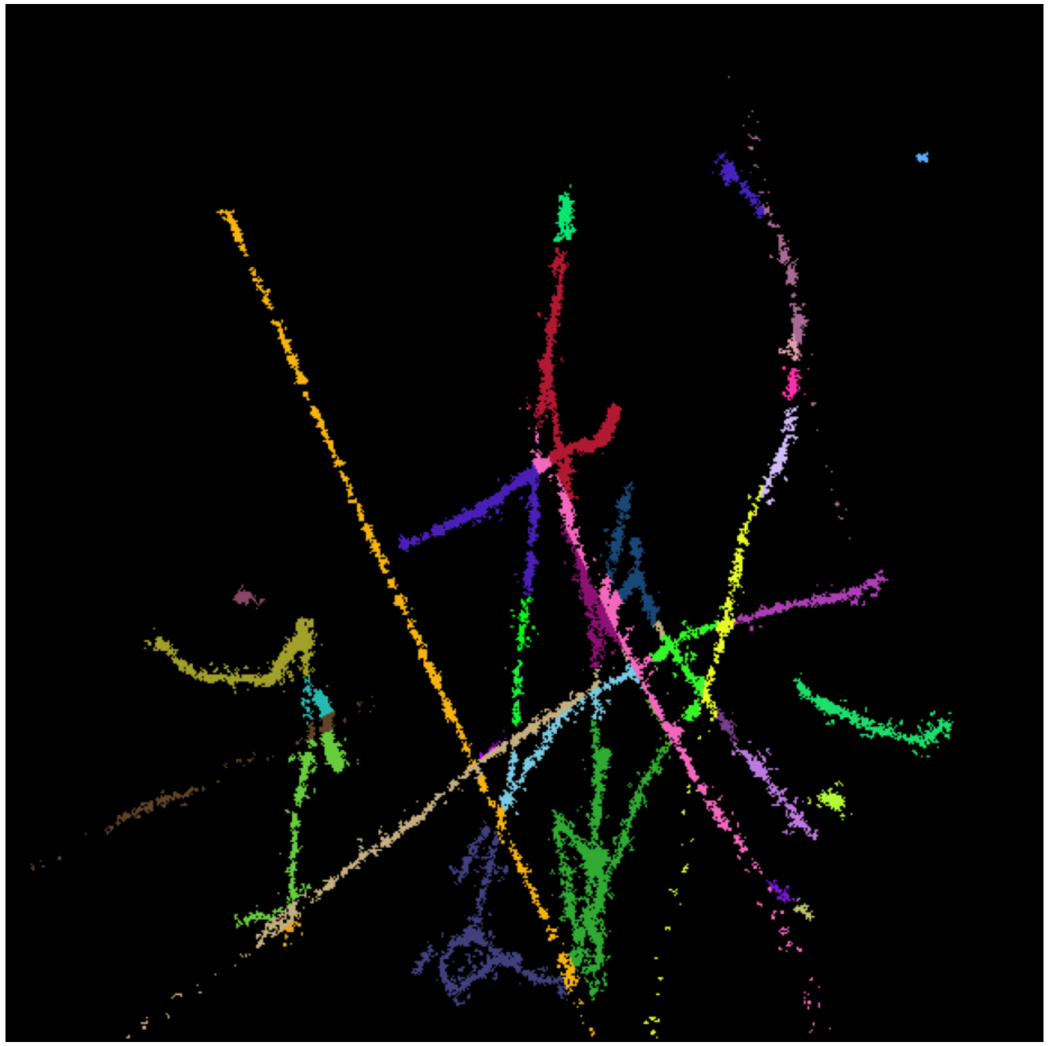
The product is a merged collection of “SUPERCLUSTERS” which contains both long and short clusters

Examples: 200ms exposure

These are the typical images taken with the minimum camera aperture allowed with the CYGNO DAQ

- N.B. This is after a lot of tuning of parameters (isolation definition, clustering metric, etc.)

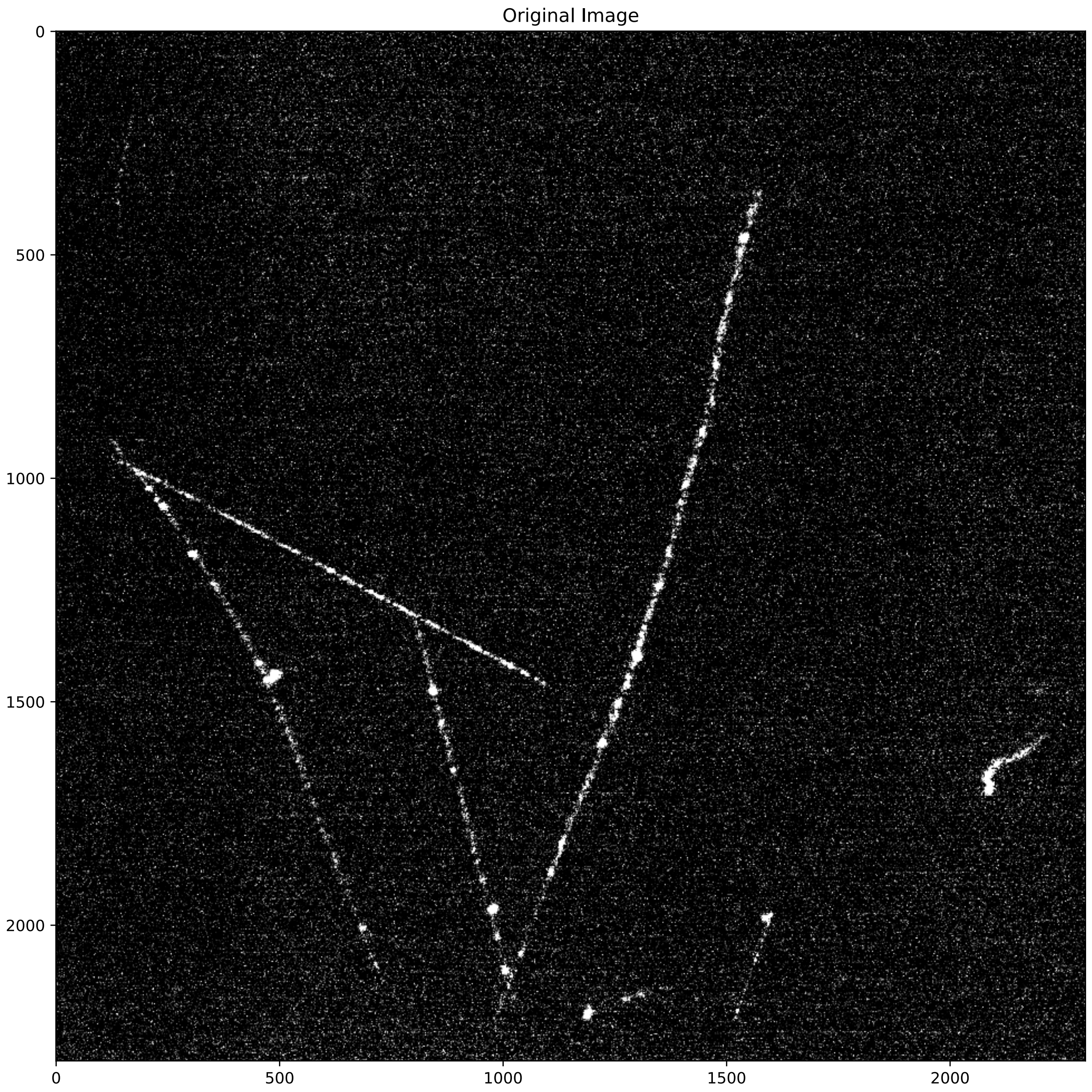


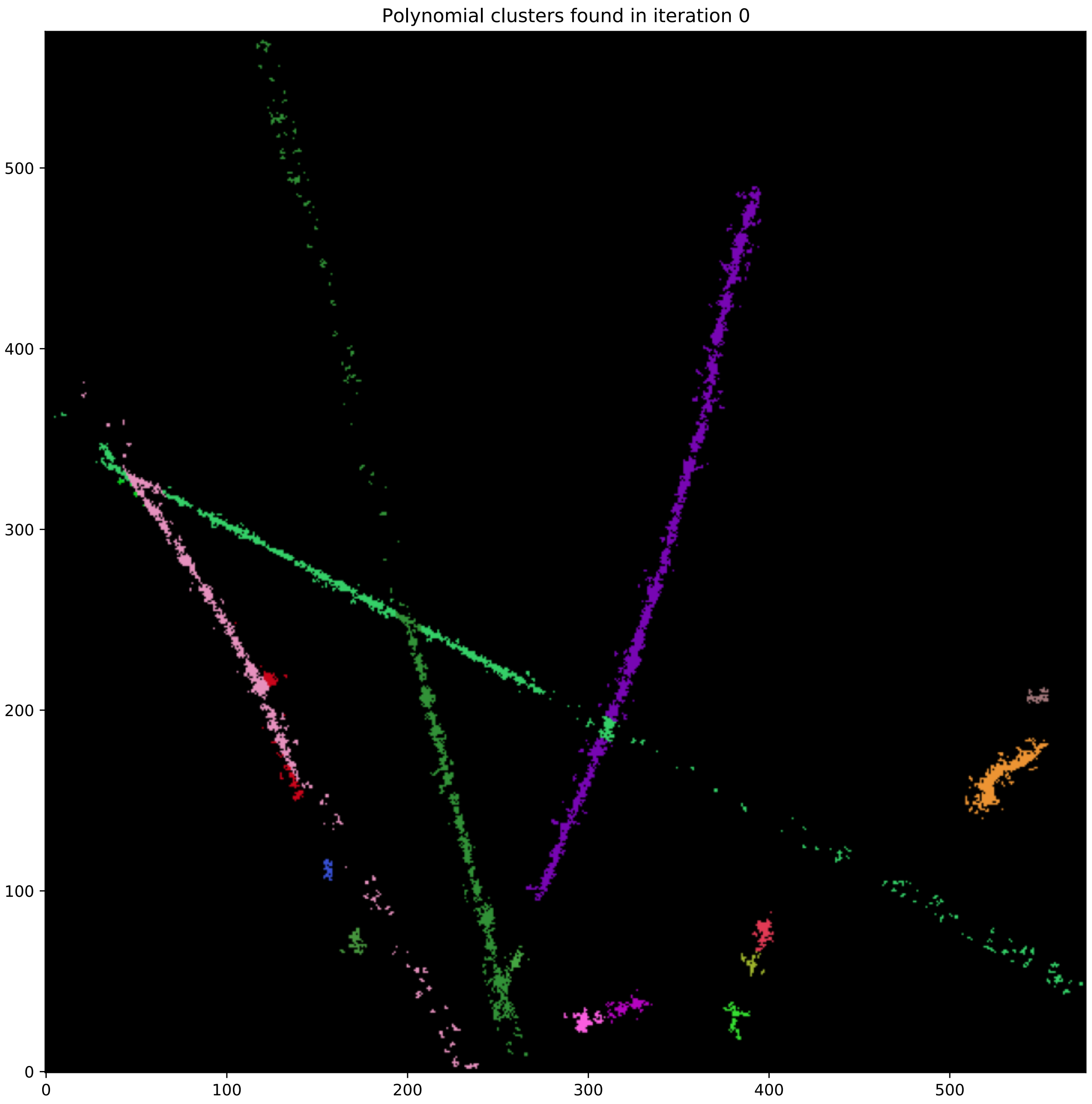


The eagerness of the directional is on purpose exaggerated because it is better to eat some piece of another track that leave a disjoint piece around (signal fake!)

Examples: 50ms exposure

Data taken with HOKAWO, without the DAQ

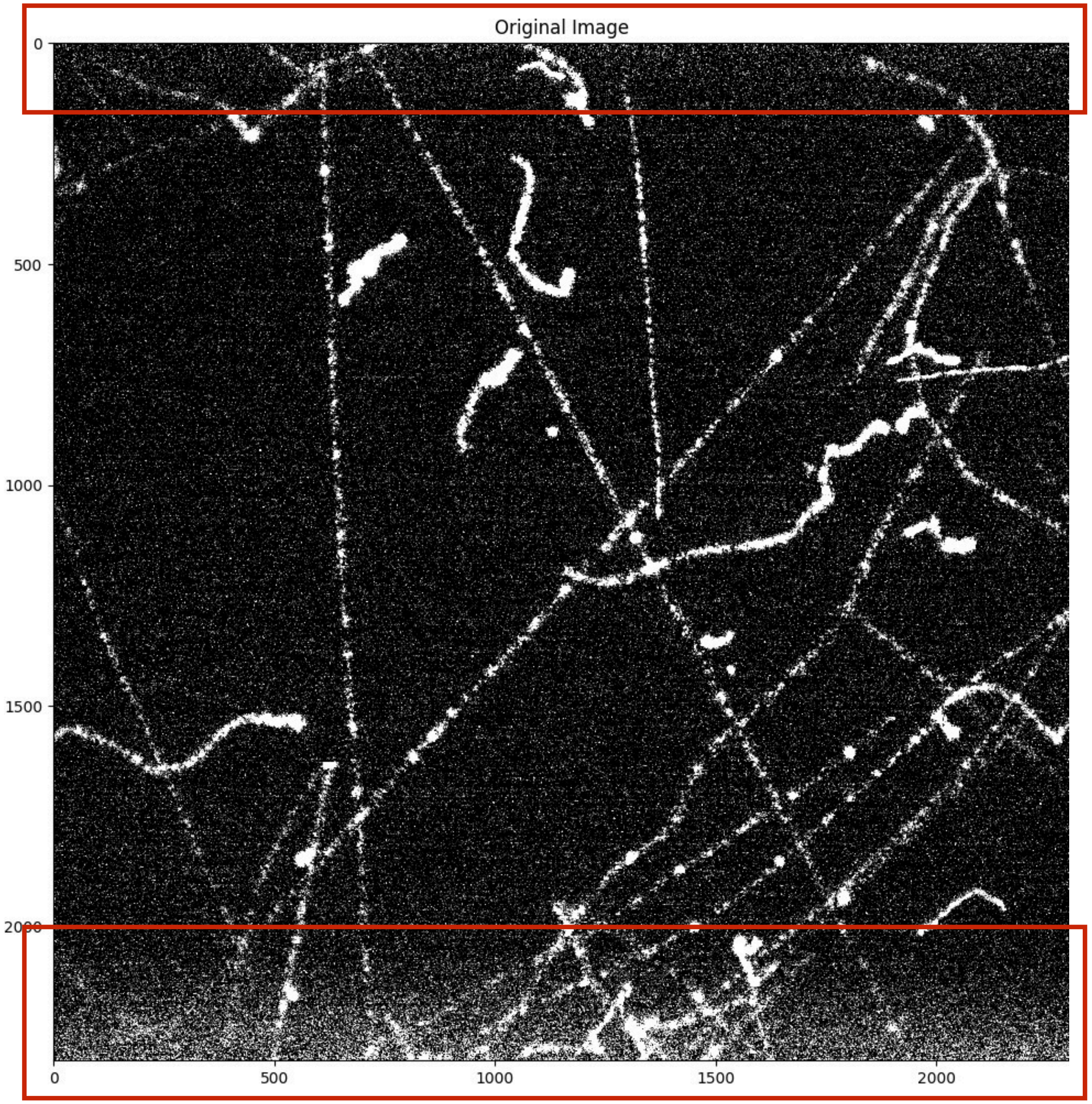

Original Image


Polynomial clusters found in iteration 0

One step back: noise subtraction

Before clustering, there is all the usual chain of noise filtering:

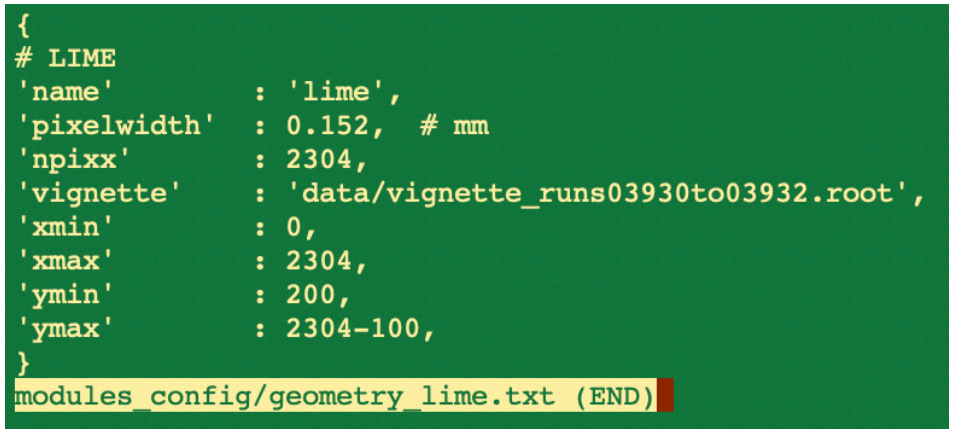
- pedestal subtraction + zero suppression (pixel-wise) + neighbor filtering + median filtering + acceptance cuts



Bottom and top strips of the sensor hot after pixel-by-pixel baseline subtraction

For now, cut away the strips:

- "acceptance" can be set in modules_config/geometry_lime.txt

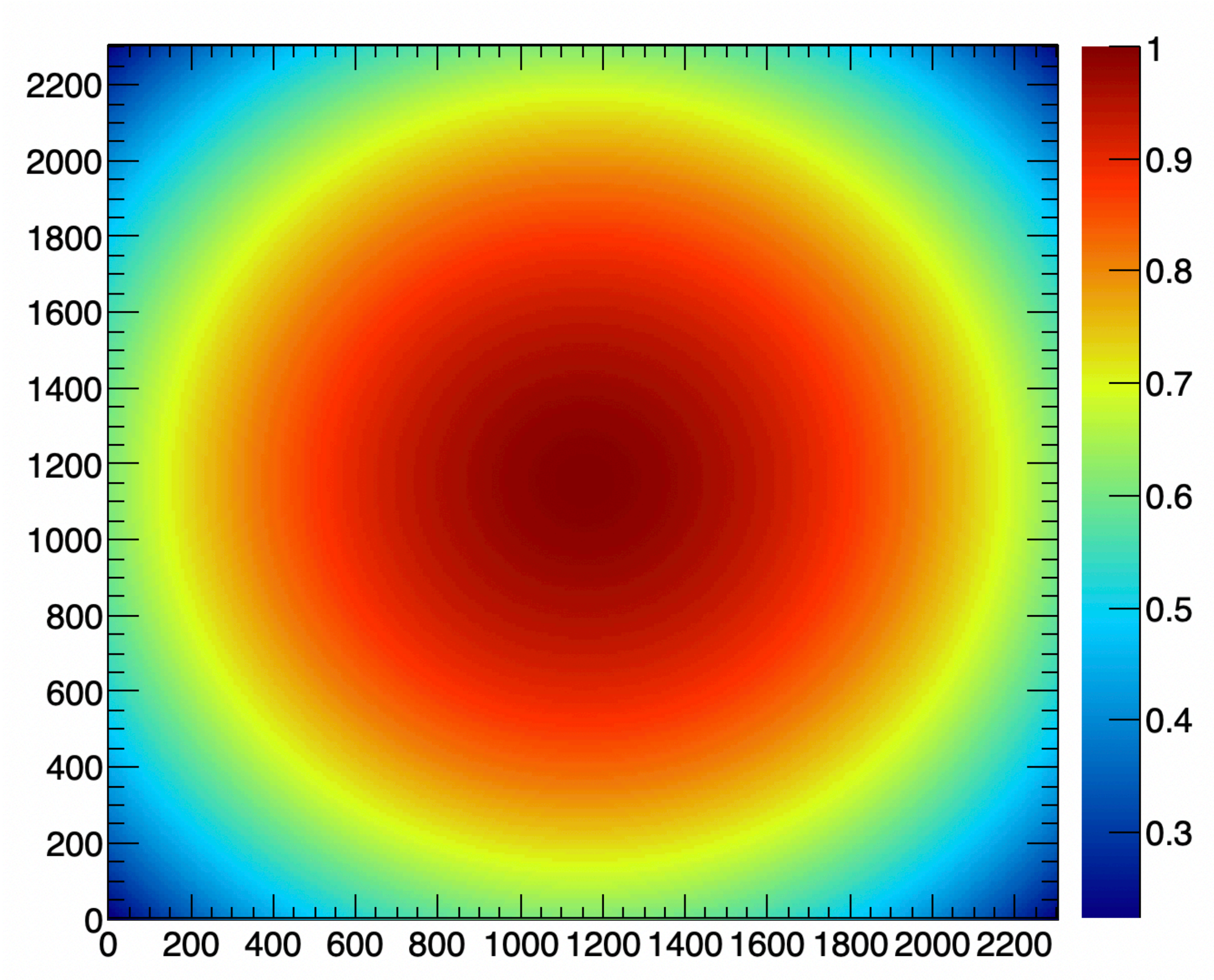


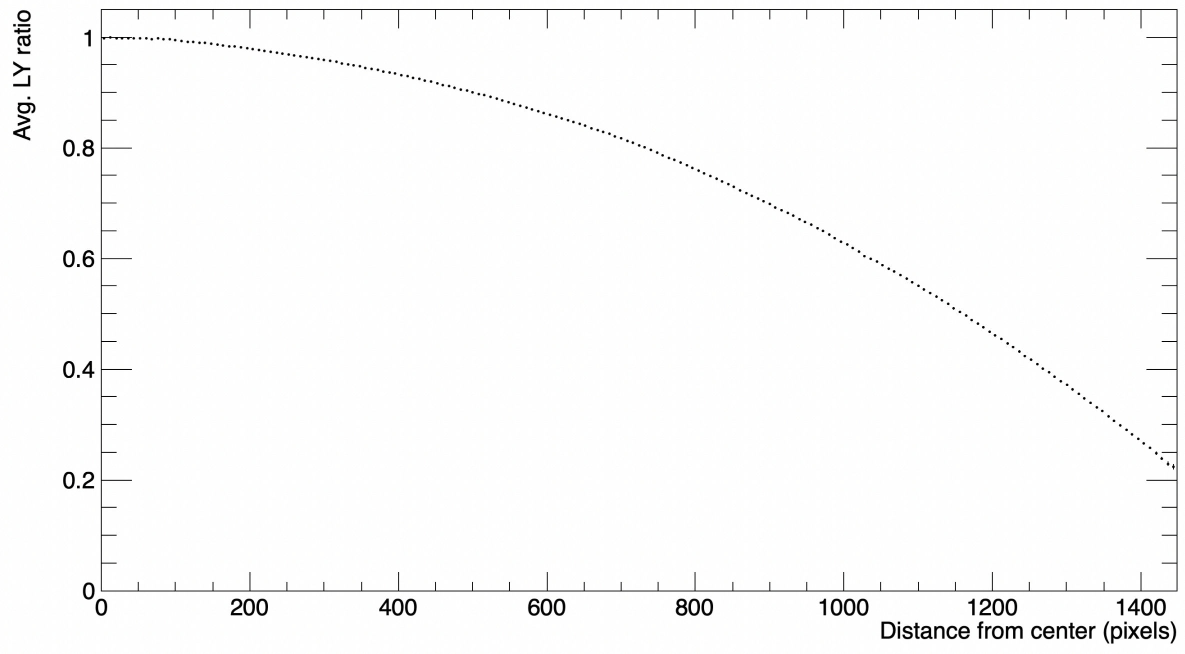
```
{
# LIME
'name'          : 'lime',
'pixelwidth'    : 0.152,   # mm
'npixx'         : 2304,
'vignette'      : 'data/vignette_runs03930to03932.root',
'xmin'          : 0,
'xmax'          : 2304,
'ymin'          : 200,
'ymax'          : 2304-100,
}
modules_config/geometry_lime.txt (END)
```

Vignetting corrections

Once the noise is subtracted, the response of each pixel is corrected with the inverse of the pure-optical vignetting map

- map obtained with white pictures => correct the main optical effect, independent on all other LIME geometrical non-uniformities





a big effect:
light yield (LY) down to 20%
in the corners wrt the center

- unavoidable effect: **the correction amplifies the noise in the low LY regions**, so expect a worse energy resolution far from the center

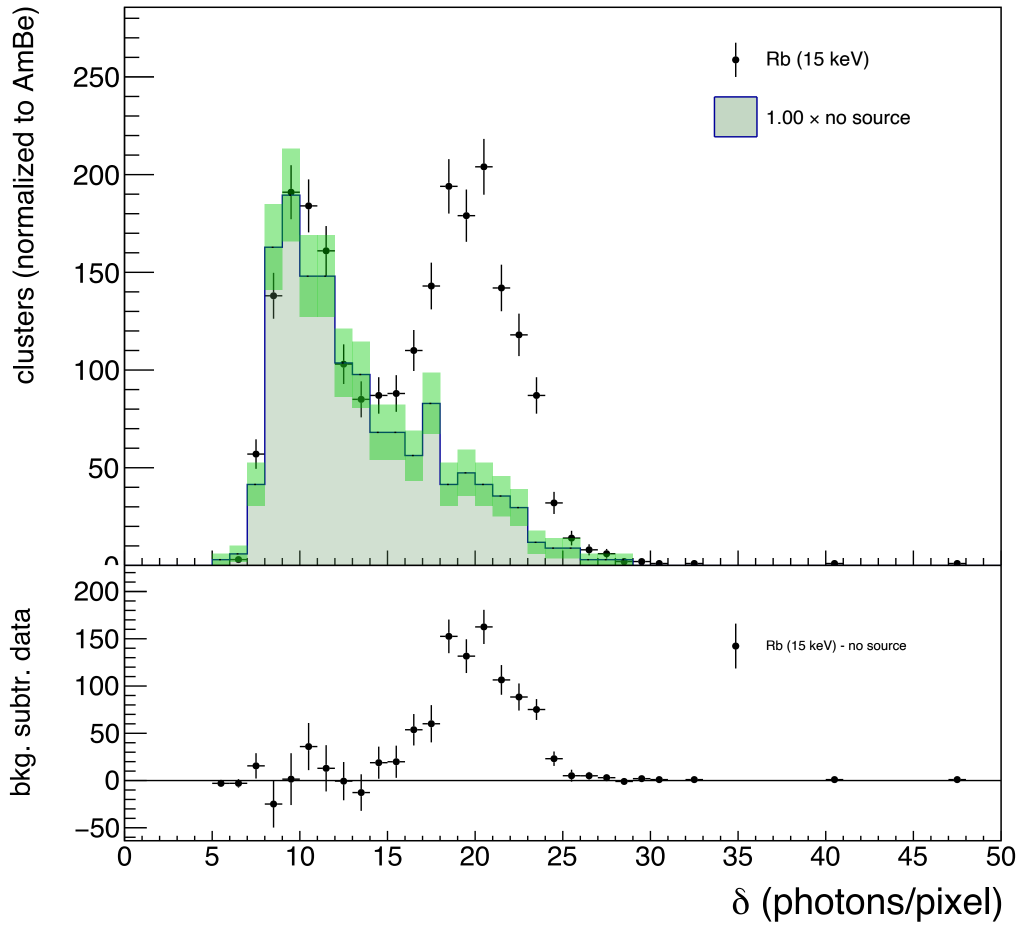
Cluster properties

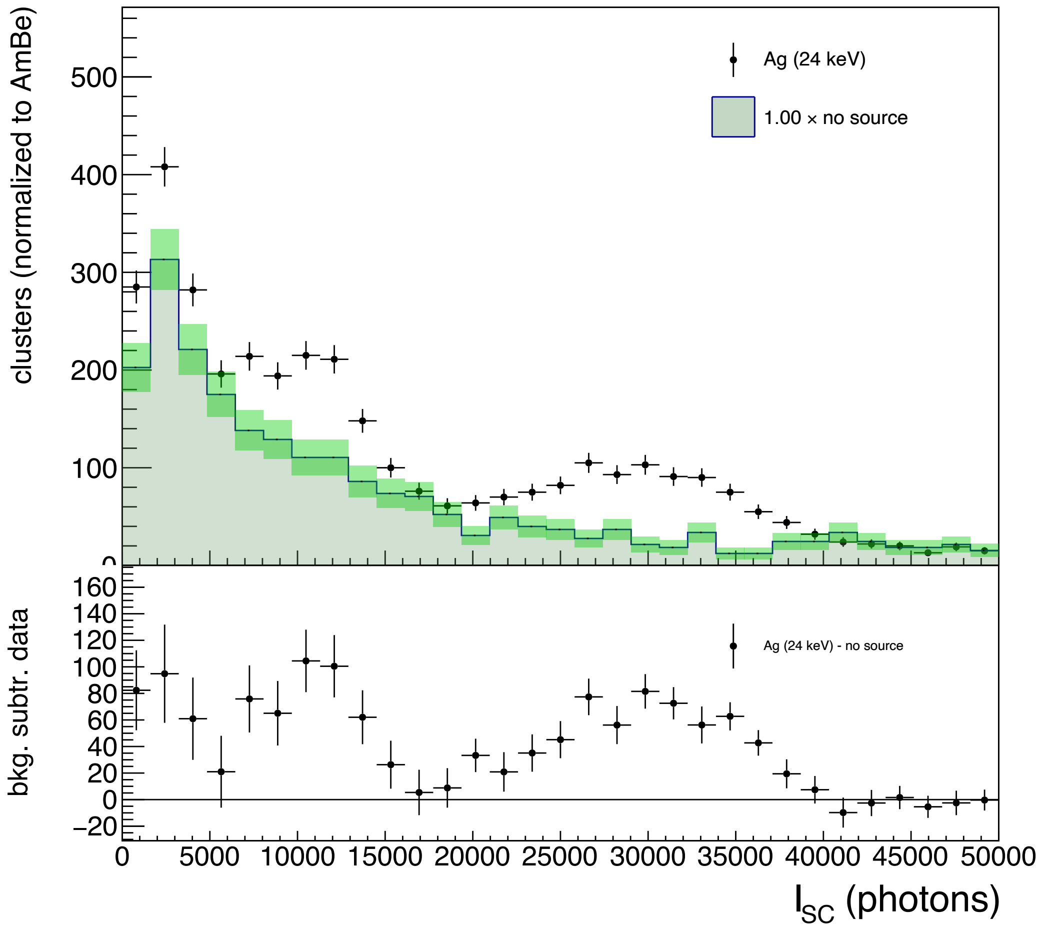
Once the superclusters are done, cluster shapes and properties can be computed and stored in the ntuples as plain floats.

Examples: length, width, row energy (in counts), transverse and longitudinal RMS and Gaussian widths, curved path length, etc.

Possibility to save all the pixels belonging to a cluster for furhter studies

N.B. this is independent on clustering technique: MODULARITY !





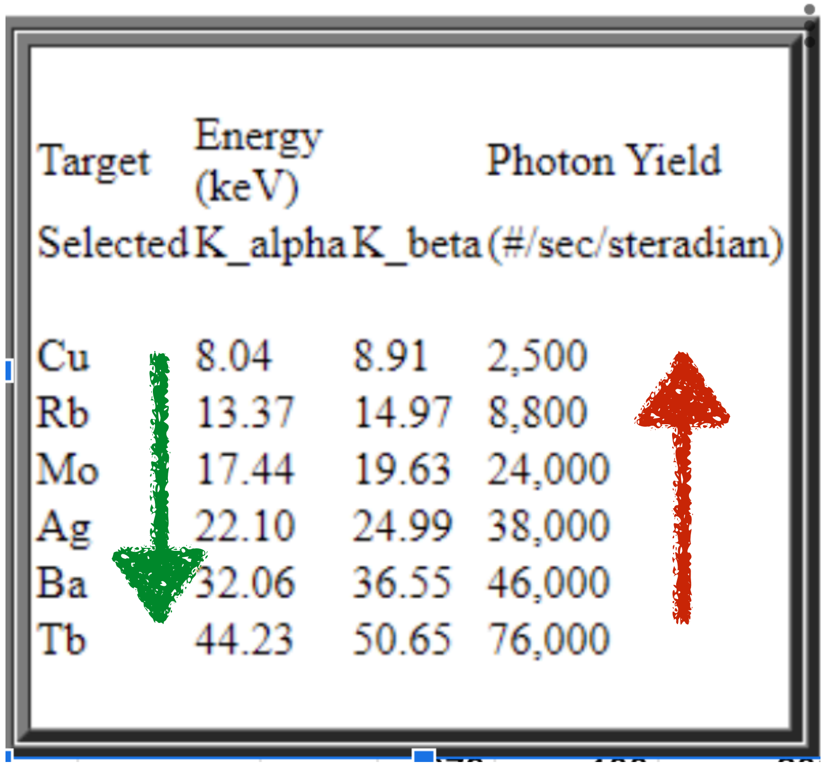
Let’s use these data: energy response linearity

X-rays sources

Data taken with a variable X-rays source: ^{241}Am source impinging different materials produce lines at characteristic energy lines

Note that:

- 1) X-rays yield lowers a lot when lowering energy (8 keV yield is 3% than 50 keV)
- 2) absorption by the LIME teflon window lower energies



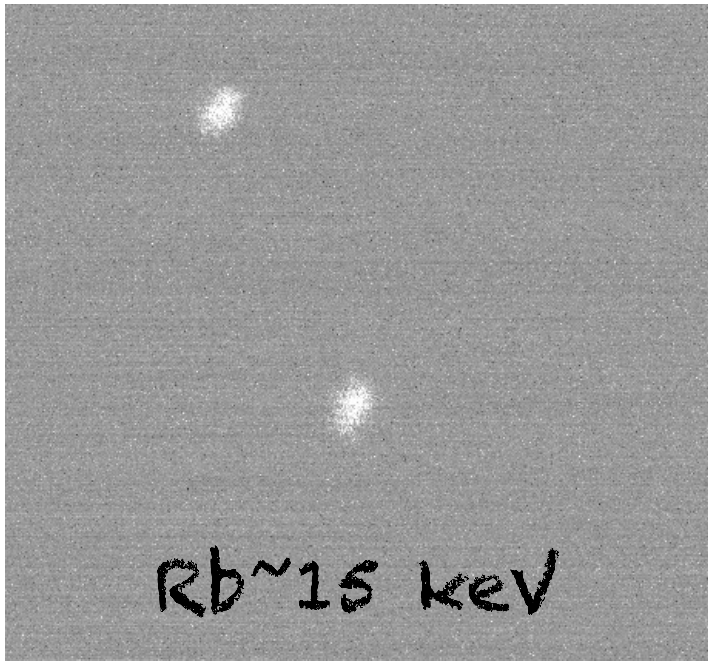
Target	Energy (keV)		Photon Yield
Selected	K_alpha	K_beta	(#/sec/steradian)
Cu	8.04	8.91	2,500
Rb	13.37	14.97	8,800
Mo	17.44	19.63	24,000
Ag	22.10	24.99	38,000
Ba	32.06	36.55	46,000
Tb	44.23	50.65	76,000

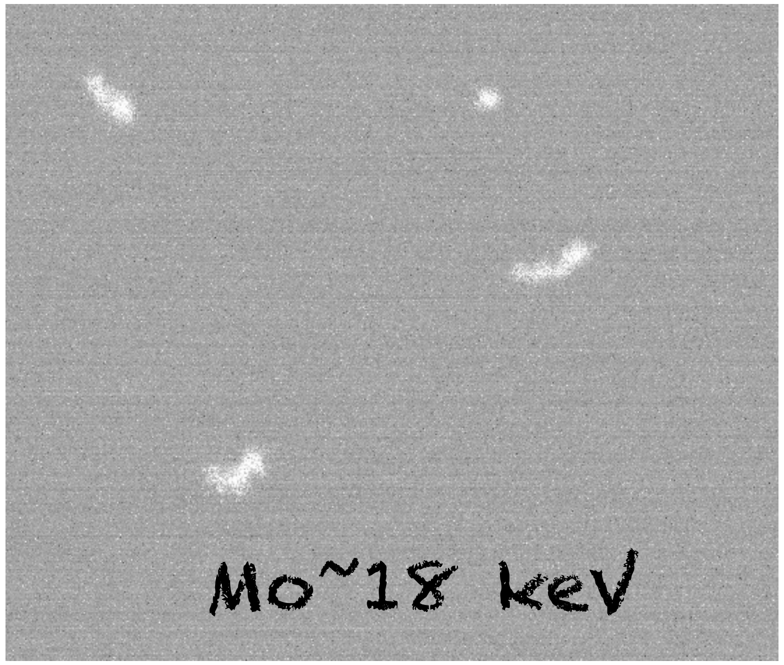
=> we need a lot of data to get a peak at lower energies

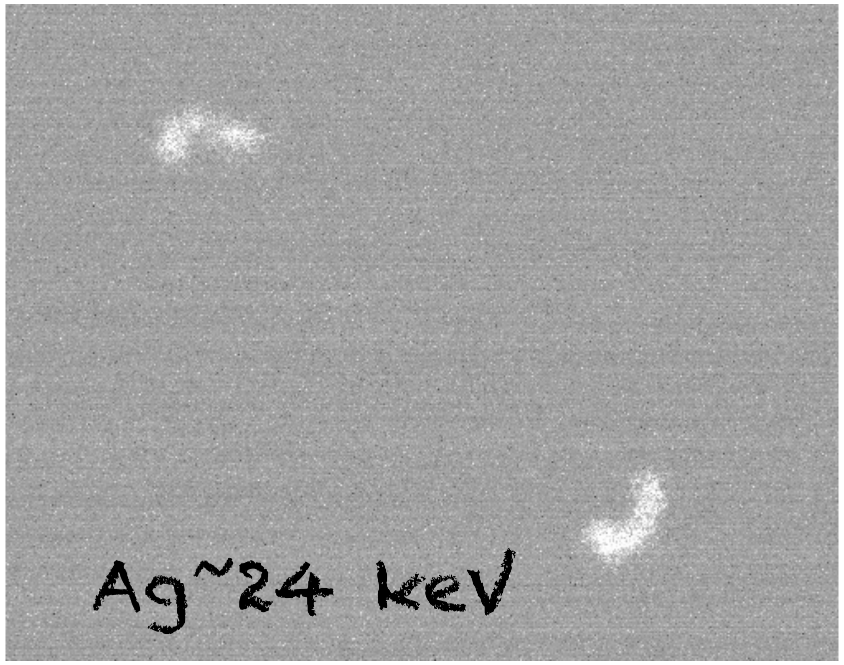
People at LNF took a lot of data with multiple energy sources and detector configurations, so we can use this data to study **the linearity of LIME energy response in different conditions.**

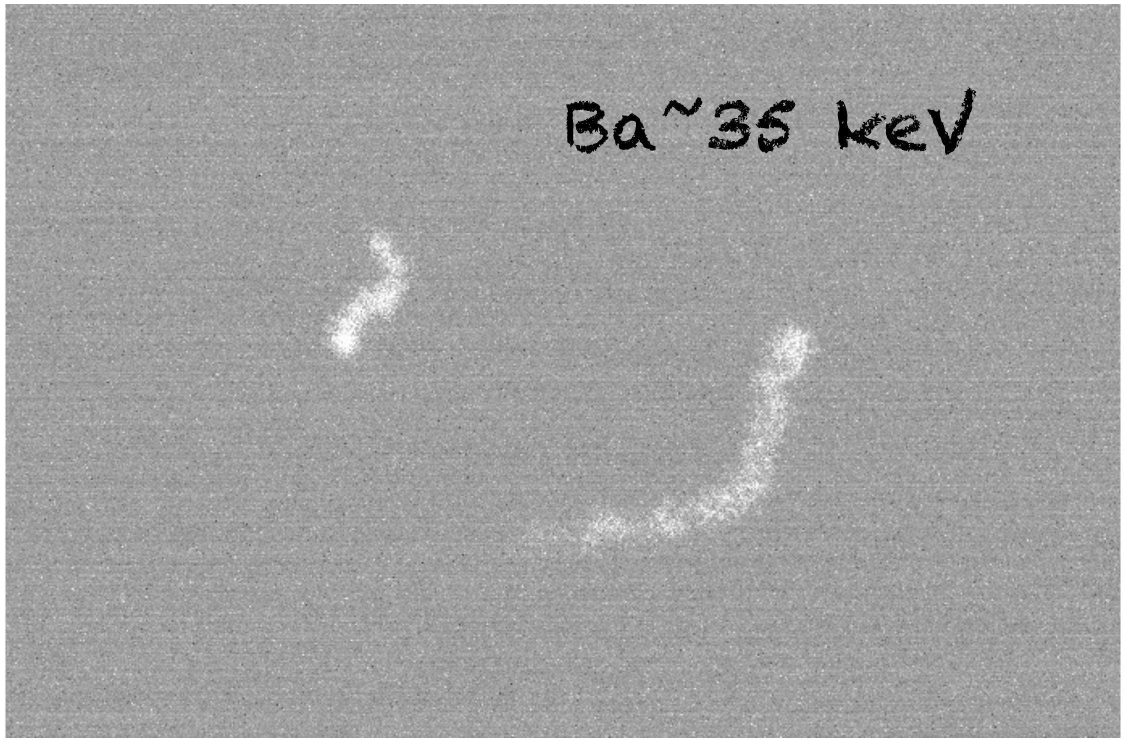
Raw event displays

selection detail: track length cut a bit relaxed for higher energies

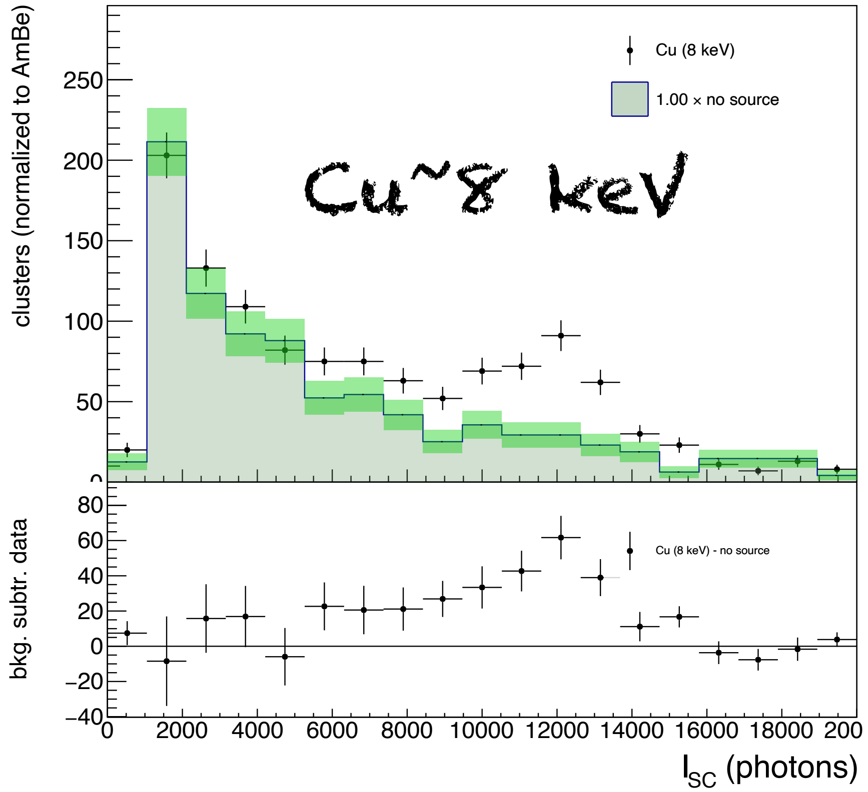


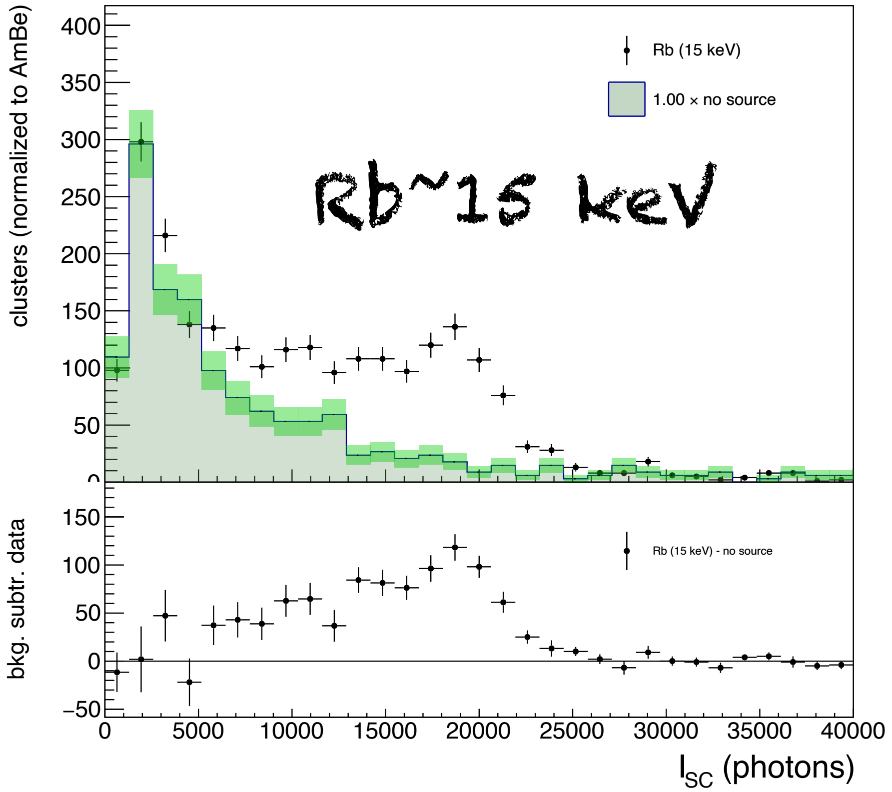


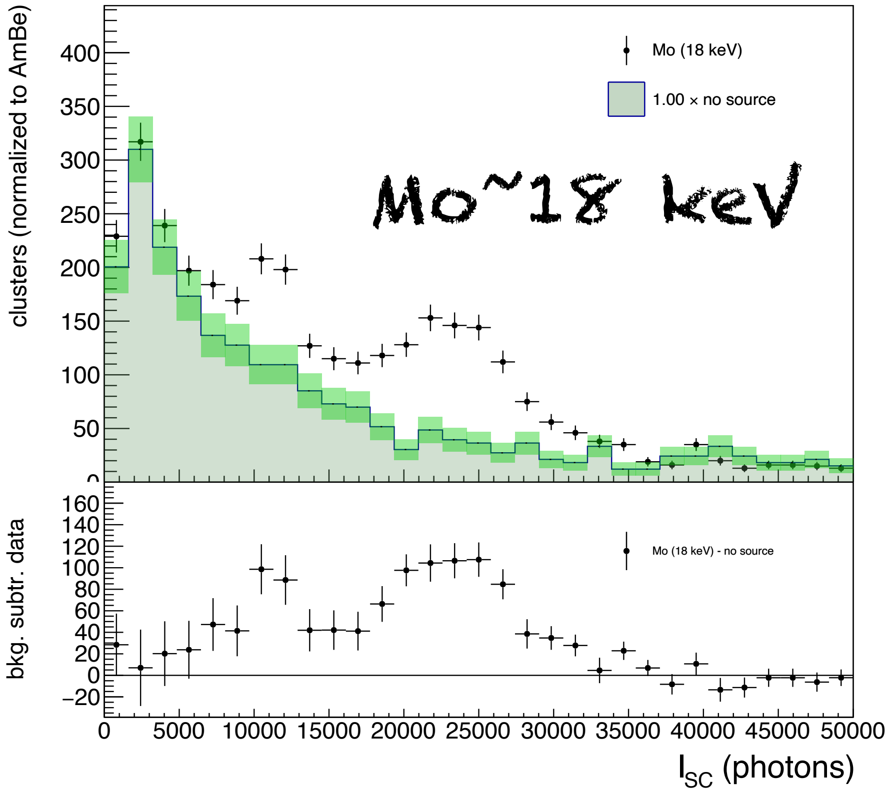


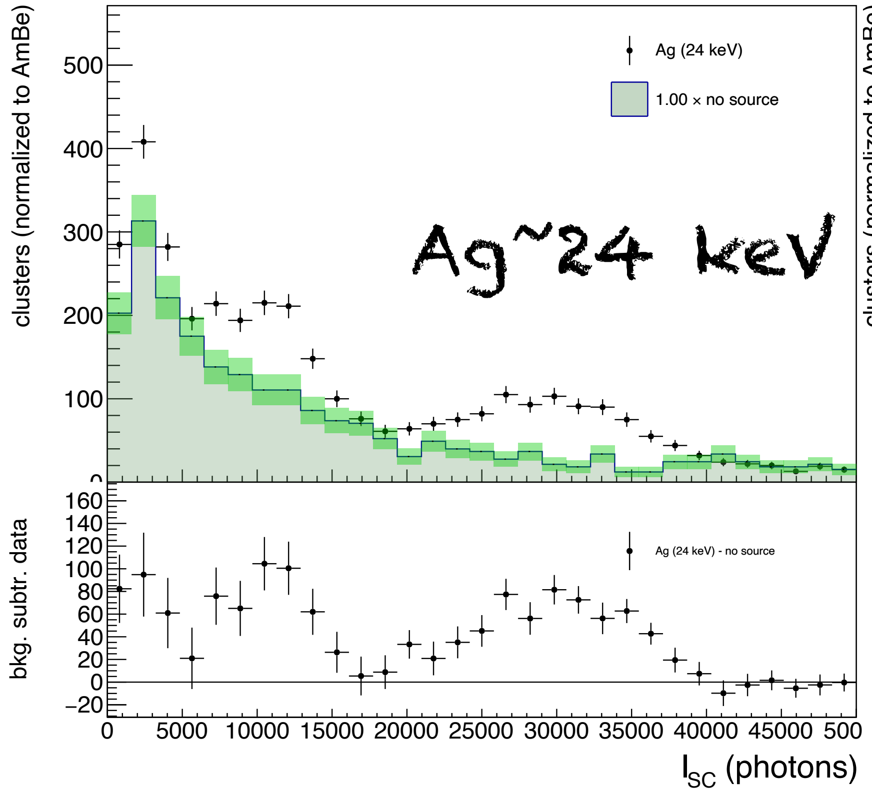


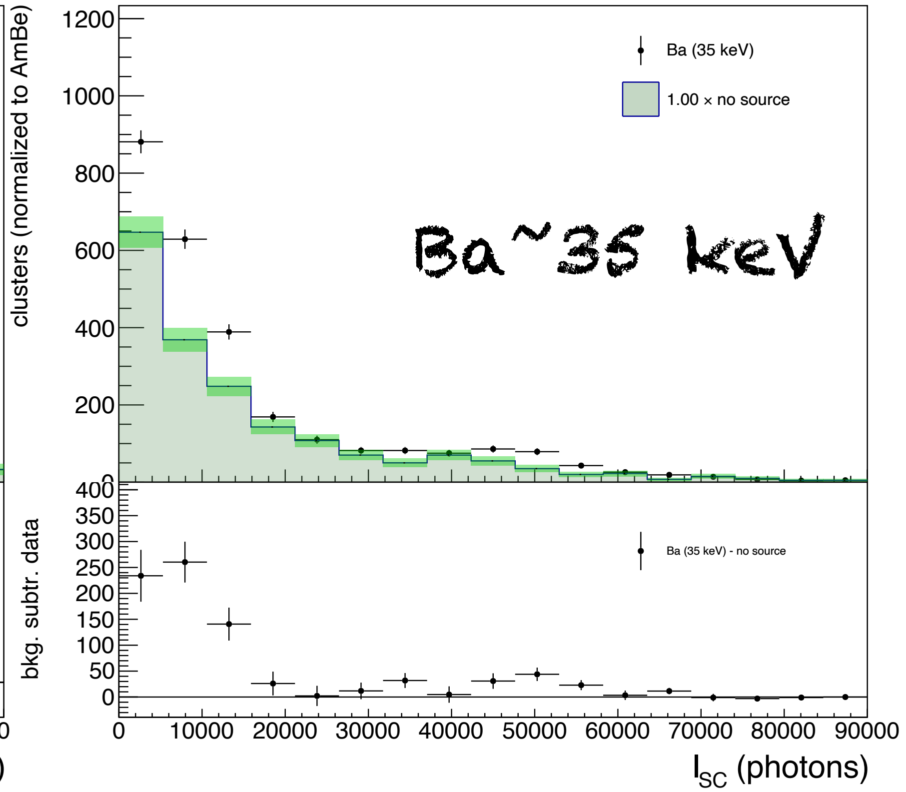
Raw energy spectra

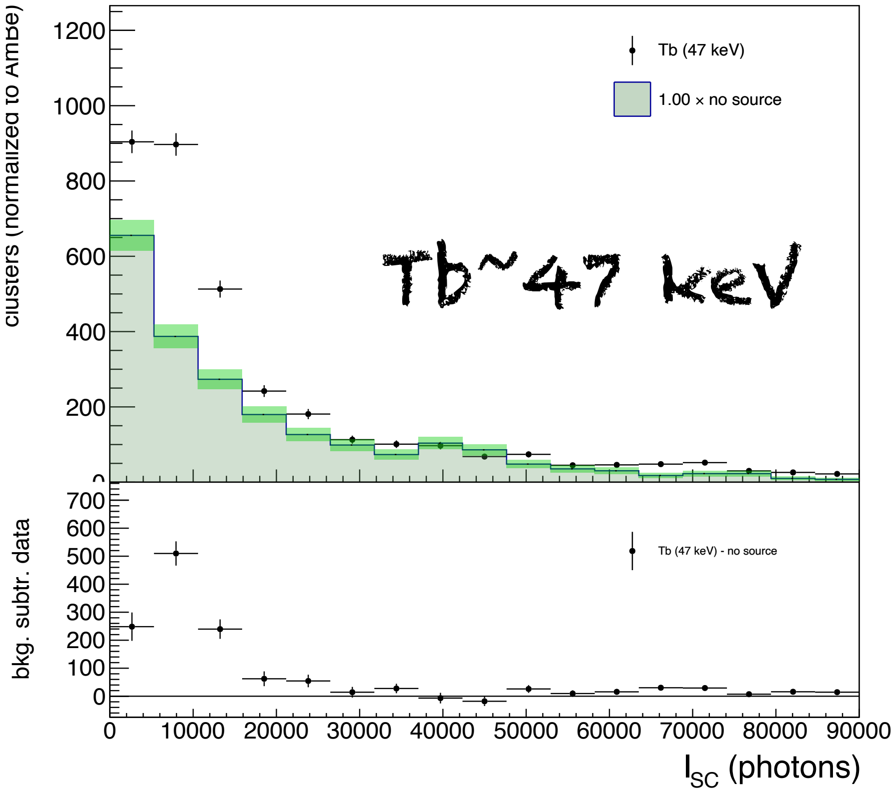








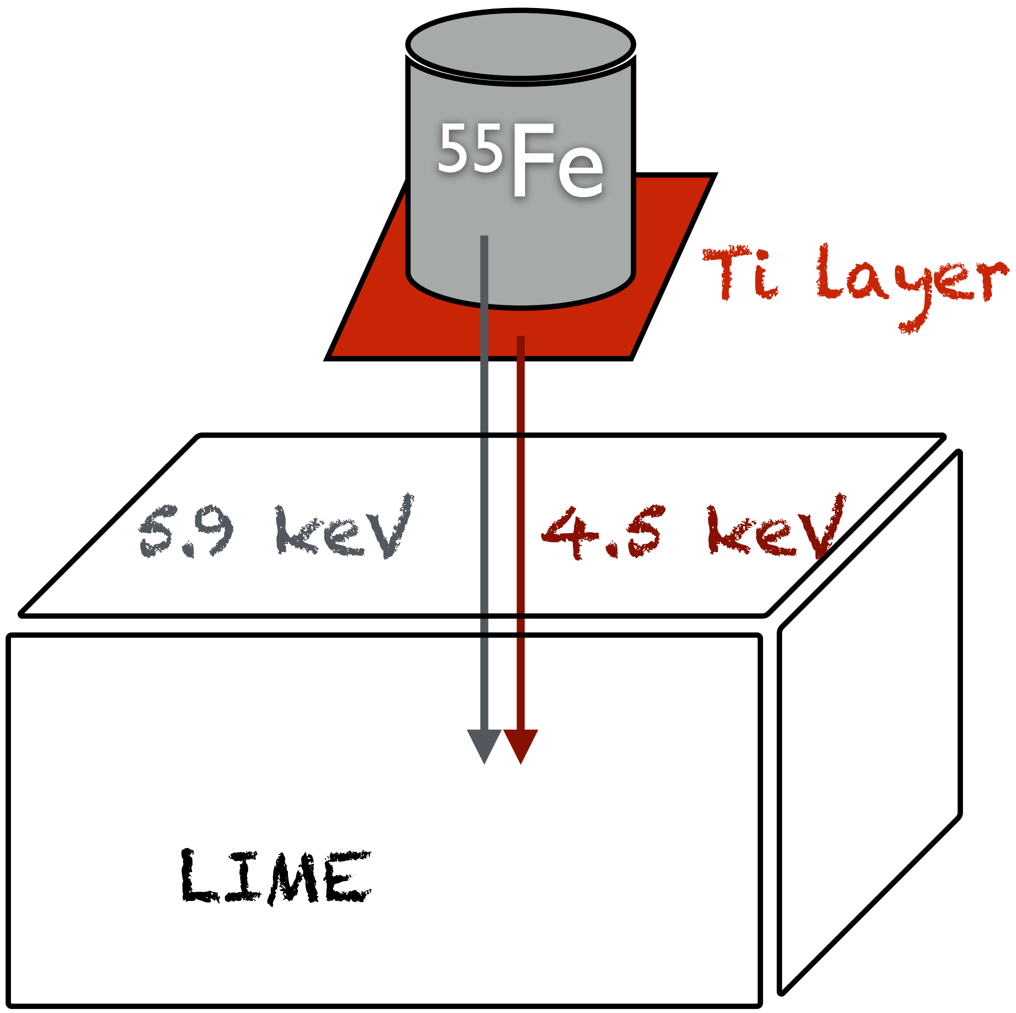




Titanium: another brick in the wall

Ti excited with 5.9 keV ^{55}Fe expected to emit 4.5 keV photons. First experimental setup: a thin layer in “penetration” mode

Data was taken later in “reflection” mode (see later on)



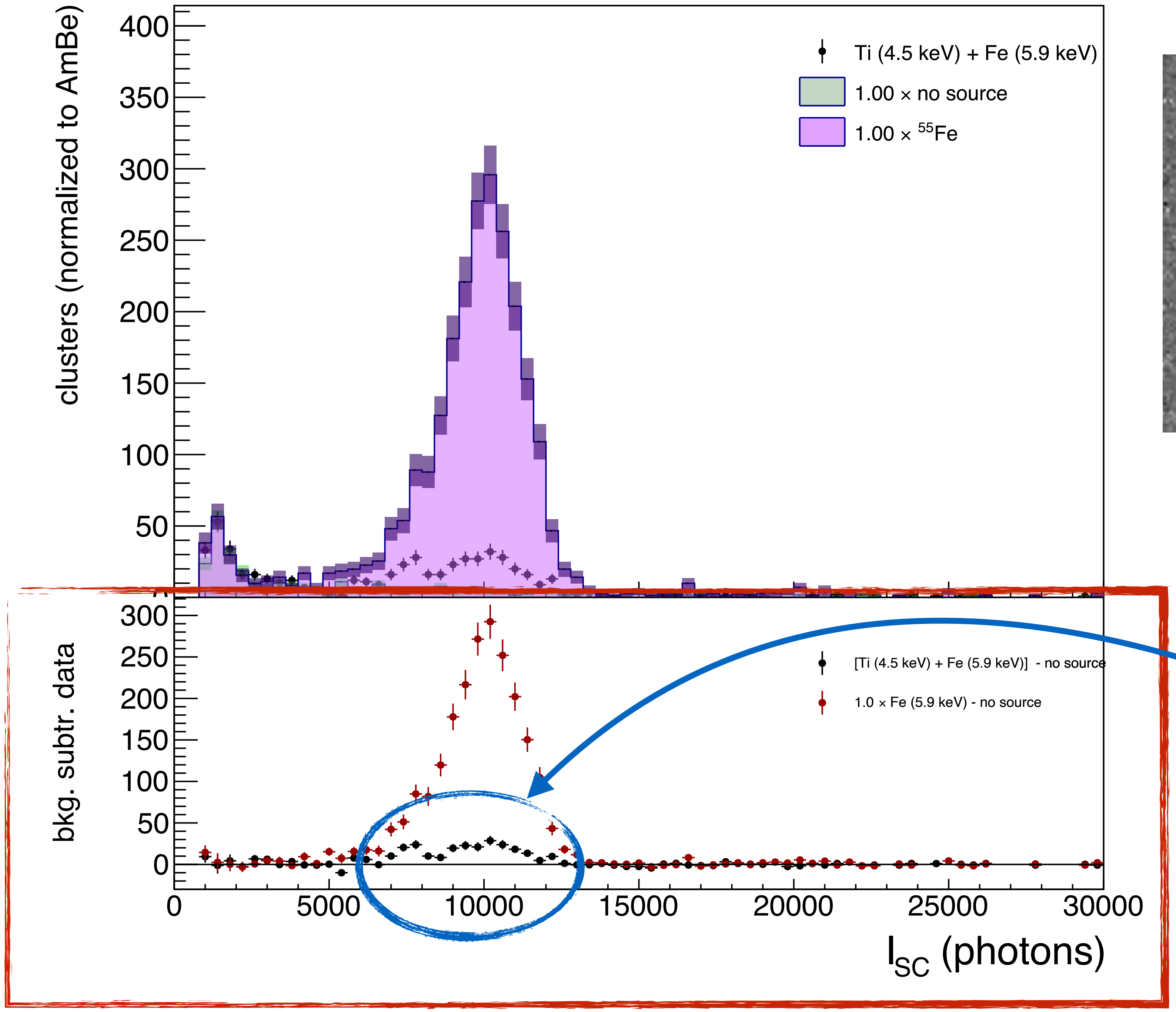
Expect to see inside LIME:

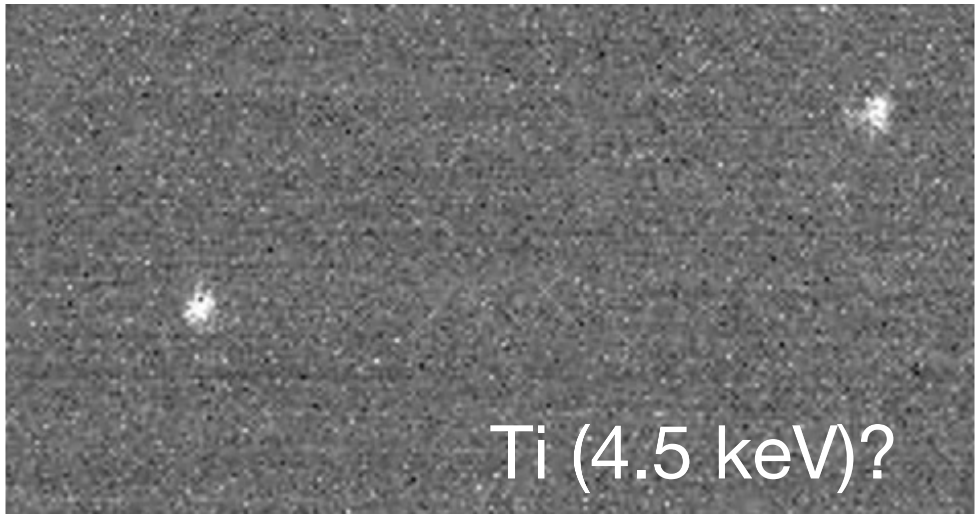
- the fraction of 5.9 keV X-rays not absorbed by Ti and teflon window
- a (smaller) fraction of 4.5 keV X-rays not absorbed by teflon window

i.e. a double peak

Titanium data after selection

As usual compare data with Fe-only, Fe+Ti, bkg-only. Subtract bkg-only normalized to exposure time.





roughly what we expect,
where we expect

LIME energy linearity results

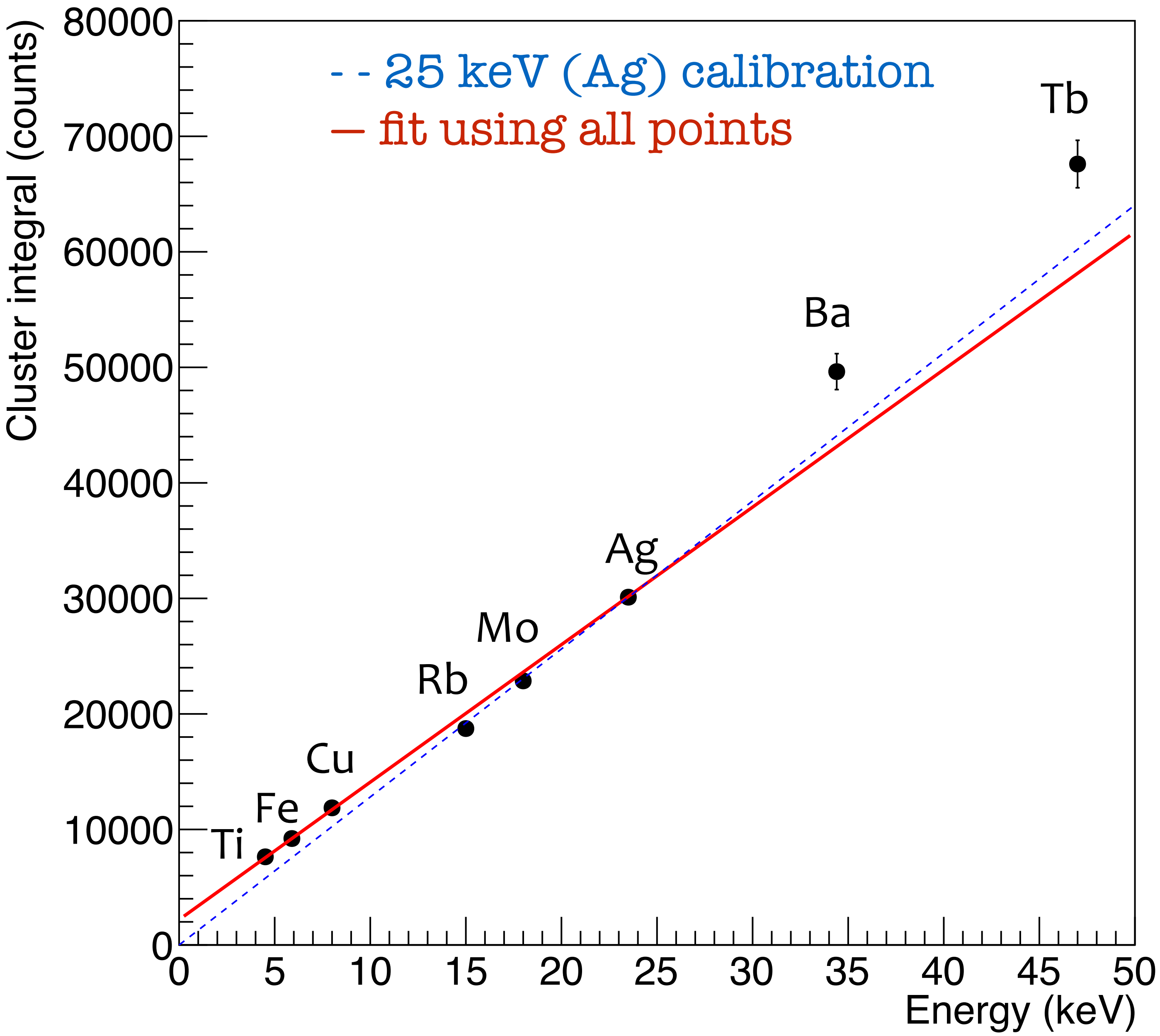
Bkg is subtracted from no-source data, resulting spectrum fitted with a Gaussian.

Other bumps are seen, but used only the expected one

N.B. These are roughly at the Cu "line", but indeed Am source can excite the Cu inside LIME

Last two points affected by large SYSTEMATIC error from bkg subtraction

energy response

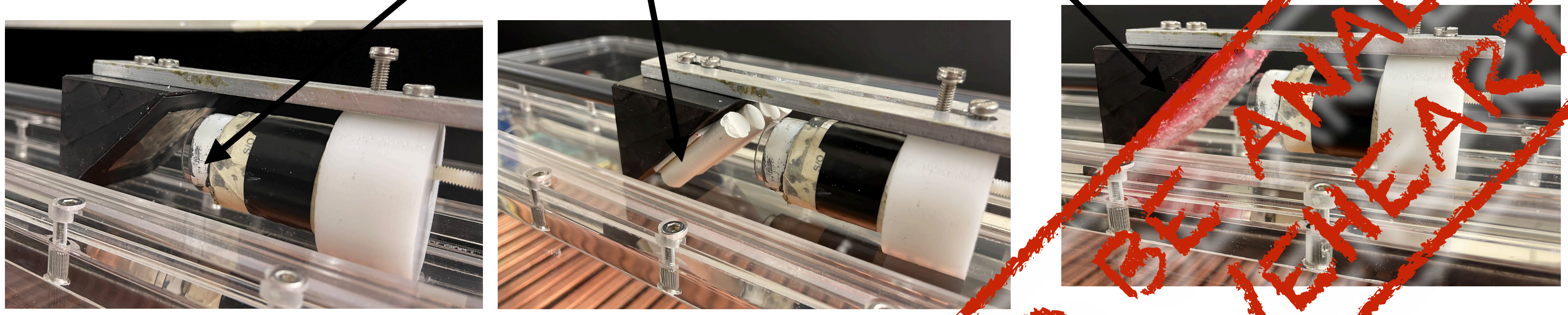


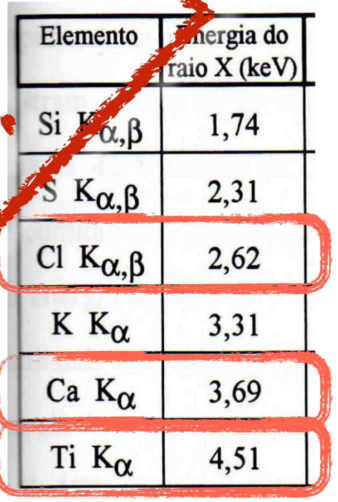
Physics interest is towards lower energies: can we go lower than 4.5 keV?

Going furhter down…

Suggestion from Cristina to use different materials excited by 5.9 keV X-rays from ^{55}Fe to produce low(er) energy X-rays

We tried Titanium, gypsum (Ca), salt (Cl)



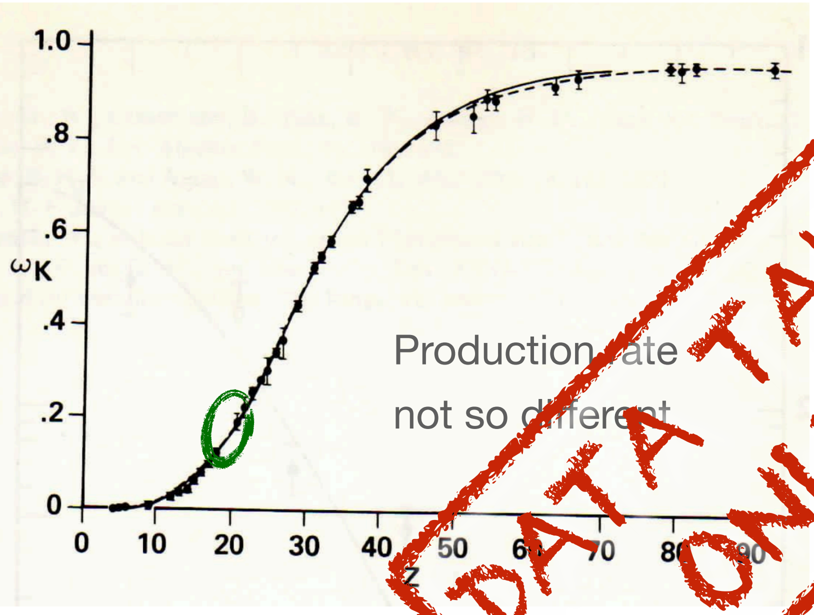


Elemento	Energia do raio X (keV)
Si $K_{\alpha,\beta}$	1,74
S $K_{\alpha,\beta}$	2,31
Cl $K_{\alpha,\beta}$	2,62
K K_{α}	3,31
Ca K_{α}	3,69
Ti K_{α}	4,51

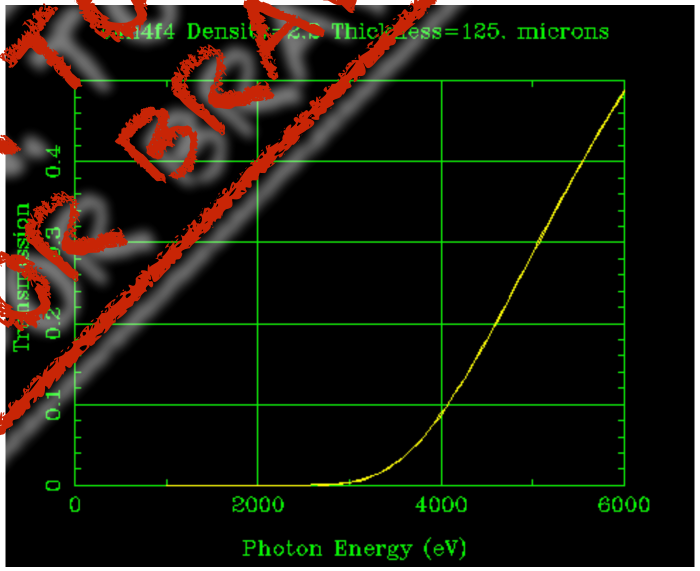
z=17

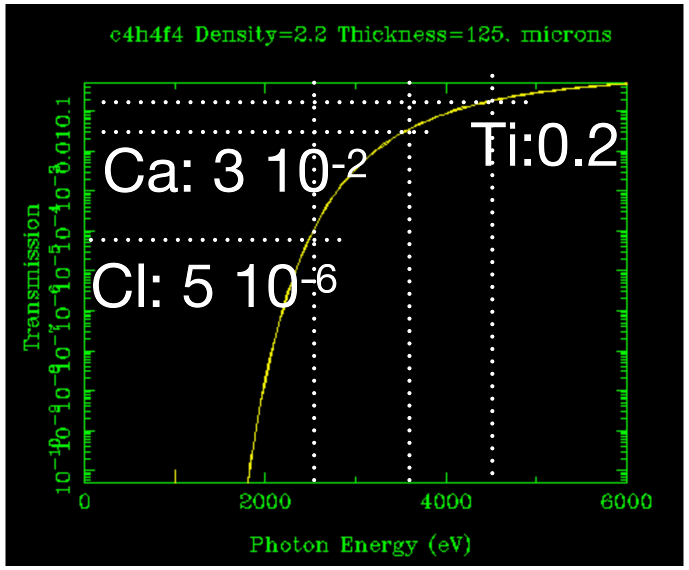
z=20

z=22



Production rate not so different





Very different probability of entering the 125

DATA TAKEN, TO BE ANALYZED:

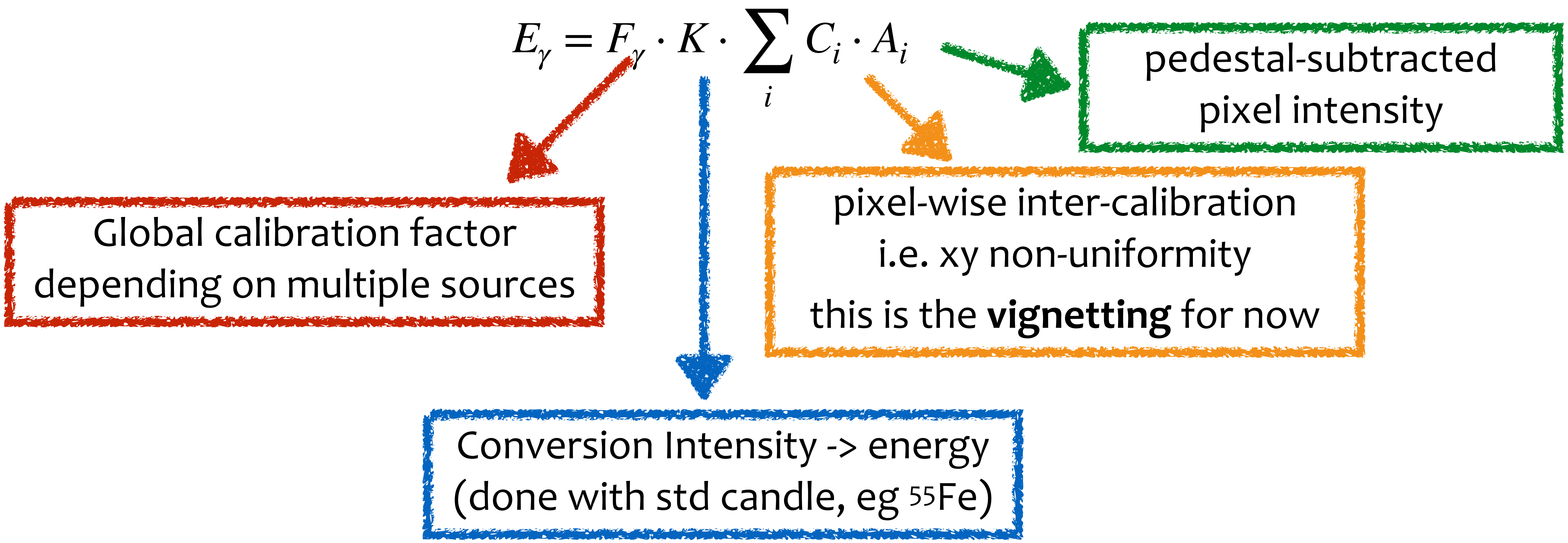
ONLY FOR BRAVEHEARTS!

Davide, Roberto, Luigi took a lot of data with "45degree" reflection from material with the trolley built by Roberto

Energy corrections

From raw intensity to energy

Once the supercluster is reconstructed, its energy is:



$$E_\gamma = F_\gamma \cdot K \cdot \sum_i C_i \cdot A_i$$

pedestal-subtracted
pixel intensity

Global calibration factor
depending on multiple sources

pixel-wise inter-calibration
i.e. xy non-uniformity
this is the **vignetting** for now

Conversion Intensity -> energy
(done with std candle, eg ^{55}Fe)

- K is a global constant, computed as response to 5.9 keV averaged on x,y and z (GEM distance)

- C_i can account also drift field non-uniformities, but decided to keep it robust and simple: vignetting only

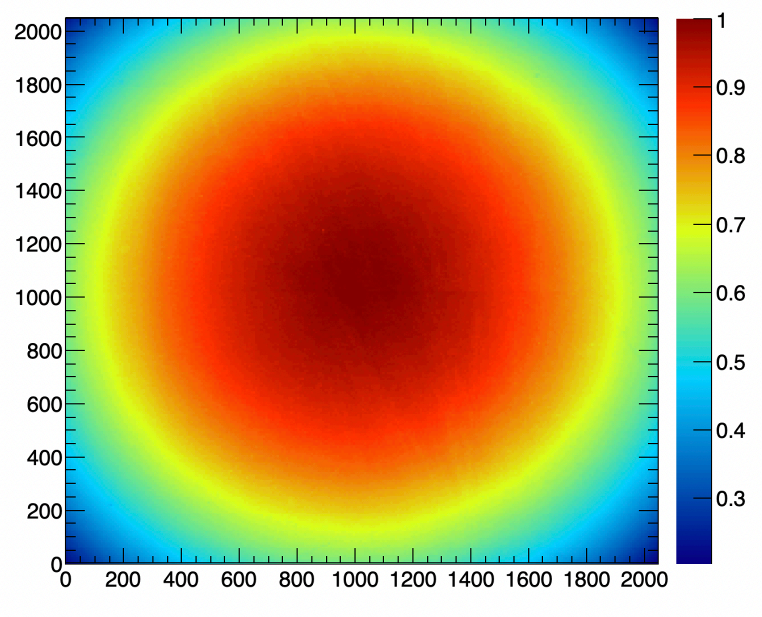
- F_γ can be computed on top of reconstructed clusters and it is discussed in the following

Corrections F_γ

To our knowledge, there are two main sources of light-yield non-uniformities, depending on either x-y (transverse projection) or z (distance wrt GEM plane):

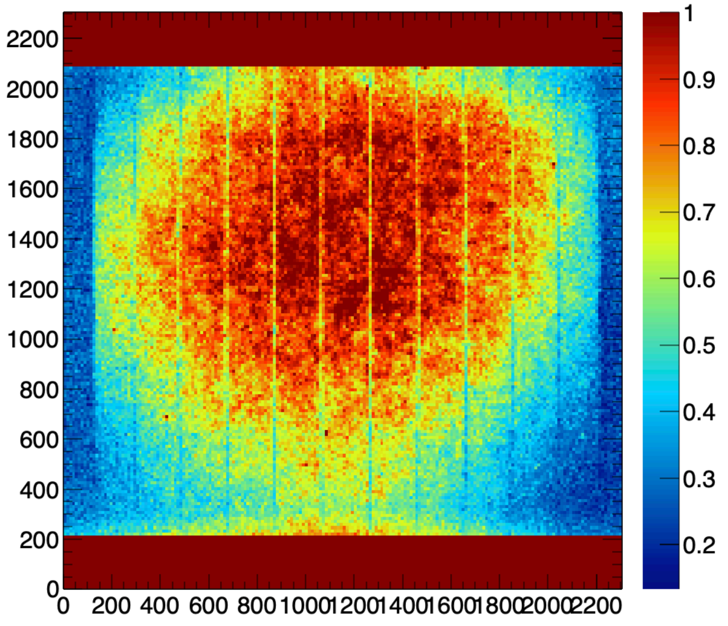
1. there is a LY pattern **F(x,y)** different than simple “radial” function caused by the vignetting

Vignetting-ONLY

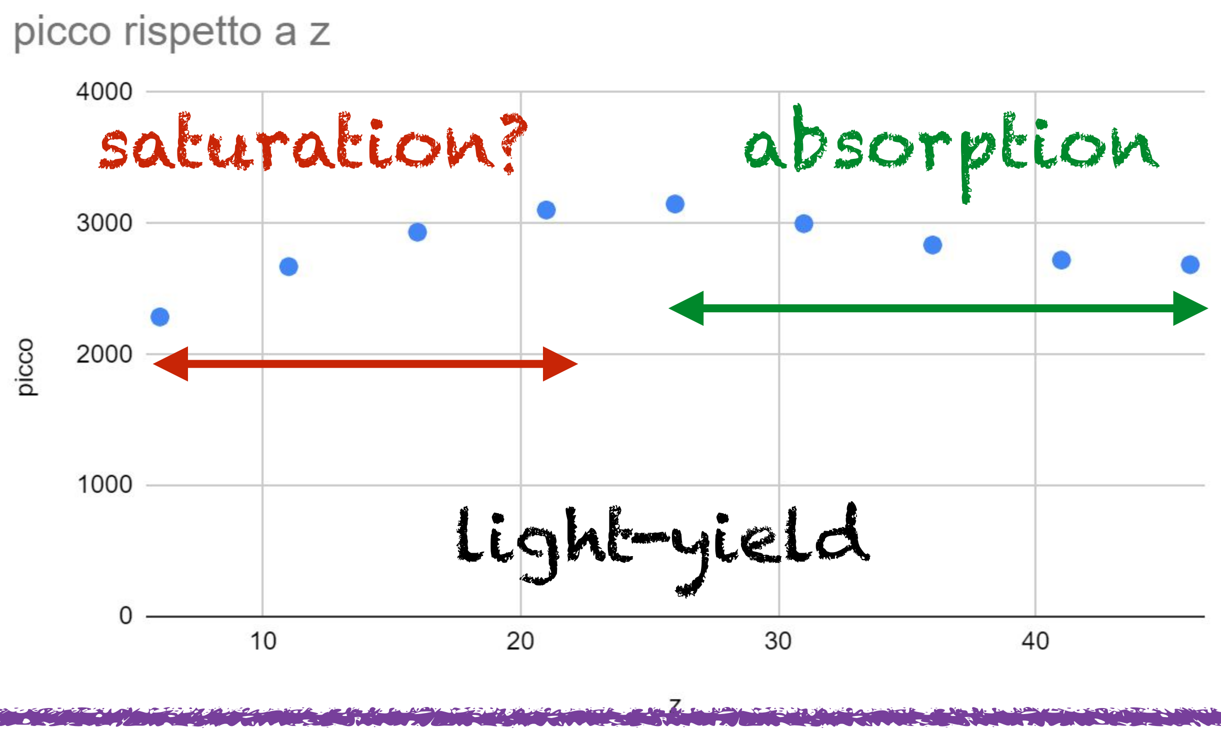


$\neq$

Long-exposure image



2. Saturation and diffusion: **F(z)**



saturation?

absorption

light-yield

F_γ with a multivariate regression

- General principle is to derive a best estimate of the dependent variable (in this case the true cluster energy) given a set of independent variables (position, cluster shape parameters, etc)
- Davide’s empiric correction was an energy correction using the projection of the energy scale onto 1 variable (density δ)
- In an event classification problem this is like using the projected likelihood in several variables (which is fully optimal as long as the correlations between variables are not relevant)
- In a classification problem one can use a multidimensional probability density, Boosted Decision Tree, or Neural Net to take into account the correlations
- We can do the same for multivariate regression
- This can easily correct **F(x,y)**, but the hope is that cluster shapes can be sensitive also to **F(z)** through correlations

Selection and sample

Ideally this should be done on SIM. Target would be E_{true}/E_{reco} (or its full PDF, not just an estimator of it, as its mean)

PRO of the SIM: Can be trained on both ERs and NRs (our signals) of whatever energy / condition / prototype

CON of the SIM: Sensitive to data-SIM disagreement of ANY of the regression inputs. At this stage we haven’t a reliable, extensive, data-SIM comparison in LIME

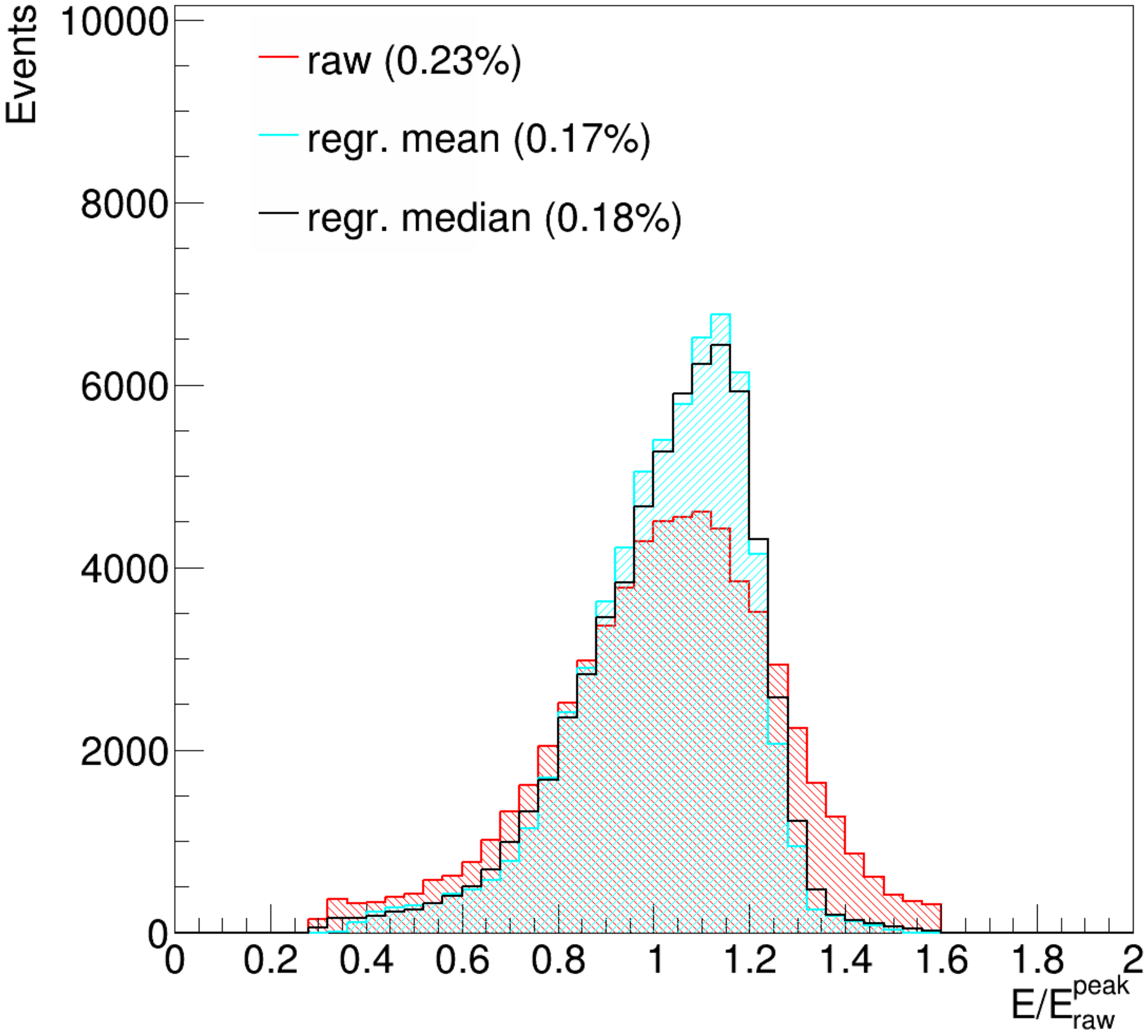
- keep in mind for the next future (needs a comparison of ALL the variables)

So right now train on DATA, ^{55}Fe, for which we have a sample with high statistics and high S/B ratio. Target is the known energy (5.9 keV, in raw pixel counts), normalized to the peak position

- Data taken with Z in the range [5-45] cm used
- Selection: length<100 pix ; width/length>0.6 ; 0.3<integral/9000<1.7 (cut away fake clusters and merged spots), R<900 pixels (avoid highly vignetted region)

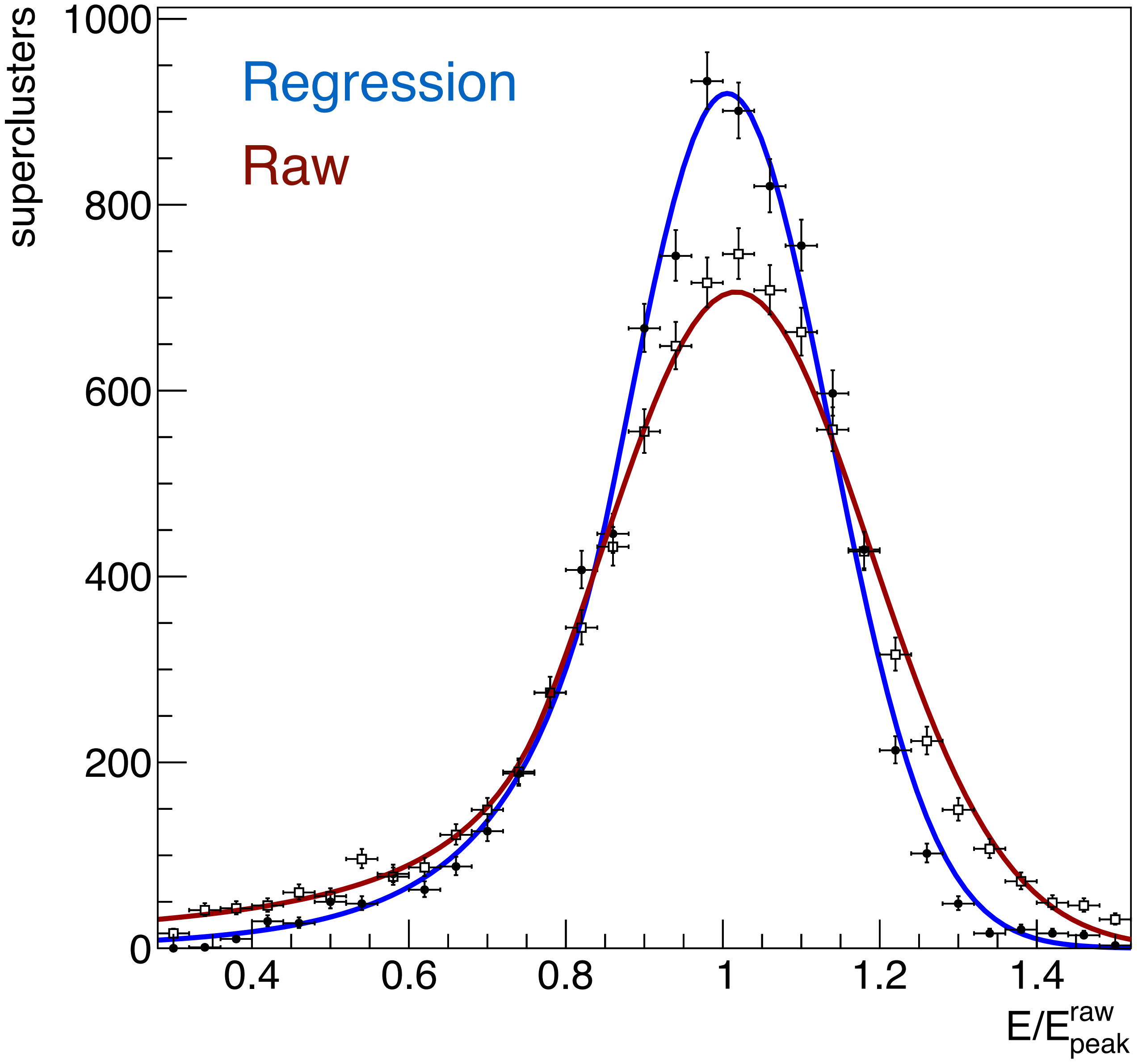
Inclusive results

N.B. Raw resolution worse than July data because it includes data with z(source-GEM) < 15cm where saturation is happening smearing the energy response.

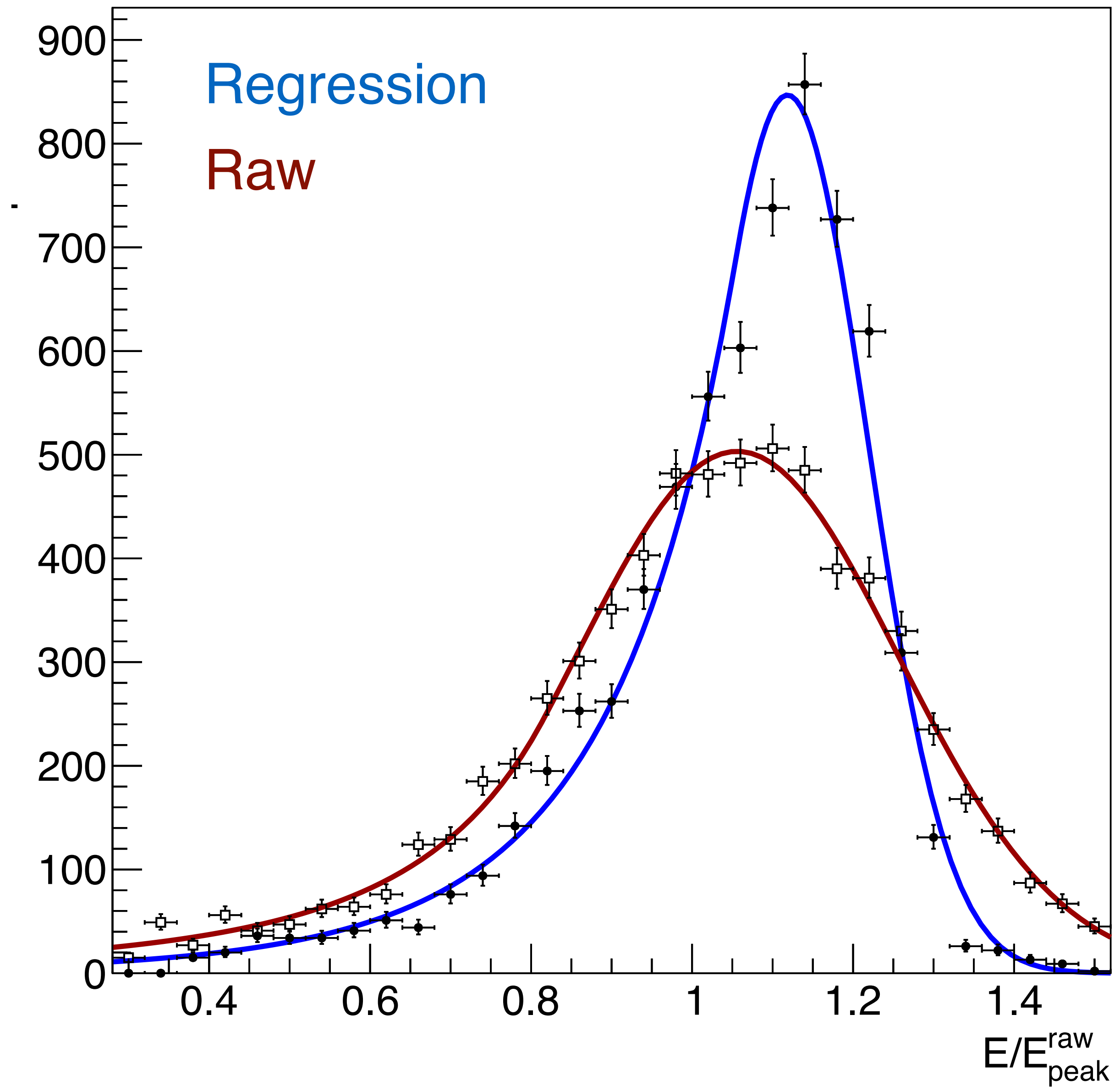


Examples at different z’s

z = 11 cm

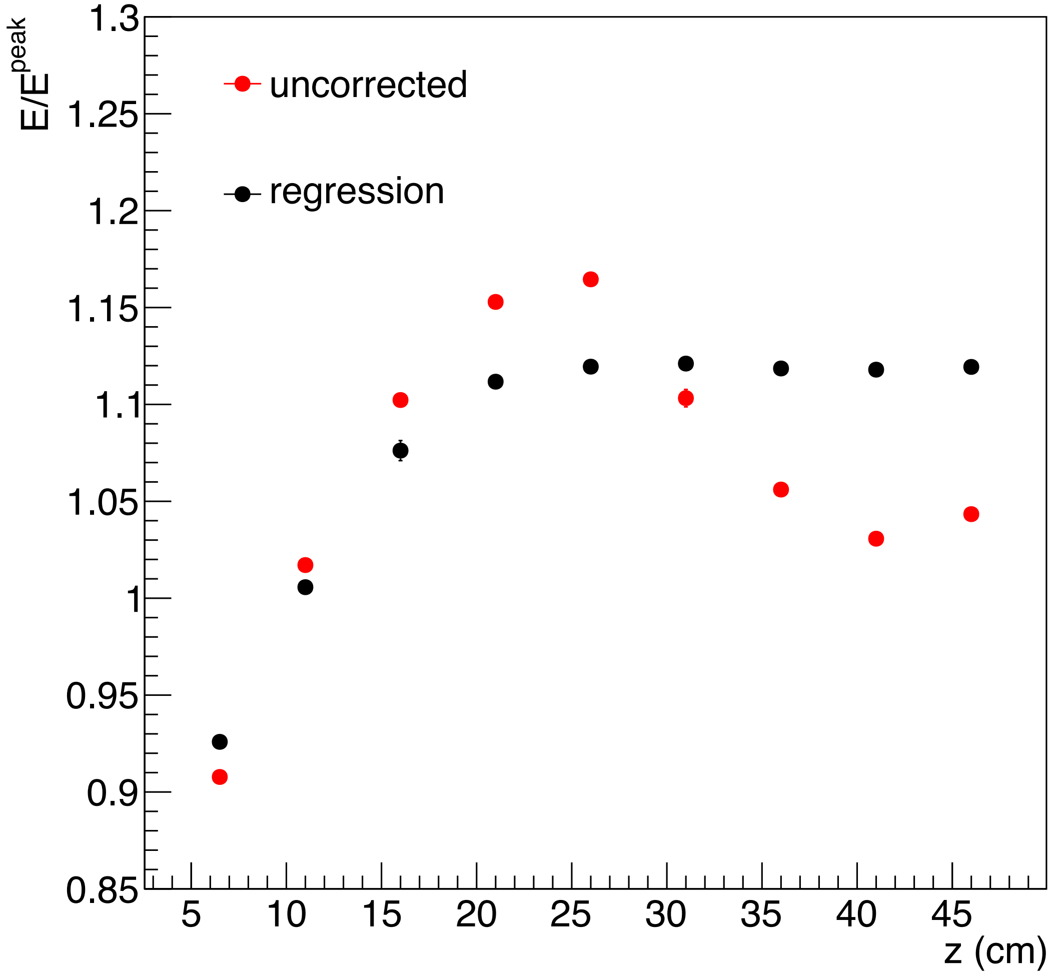


z = 36 cm

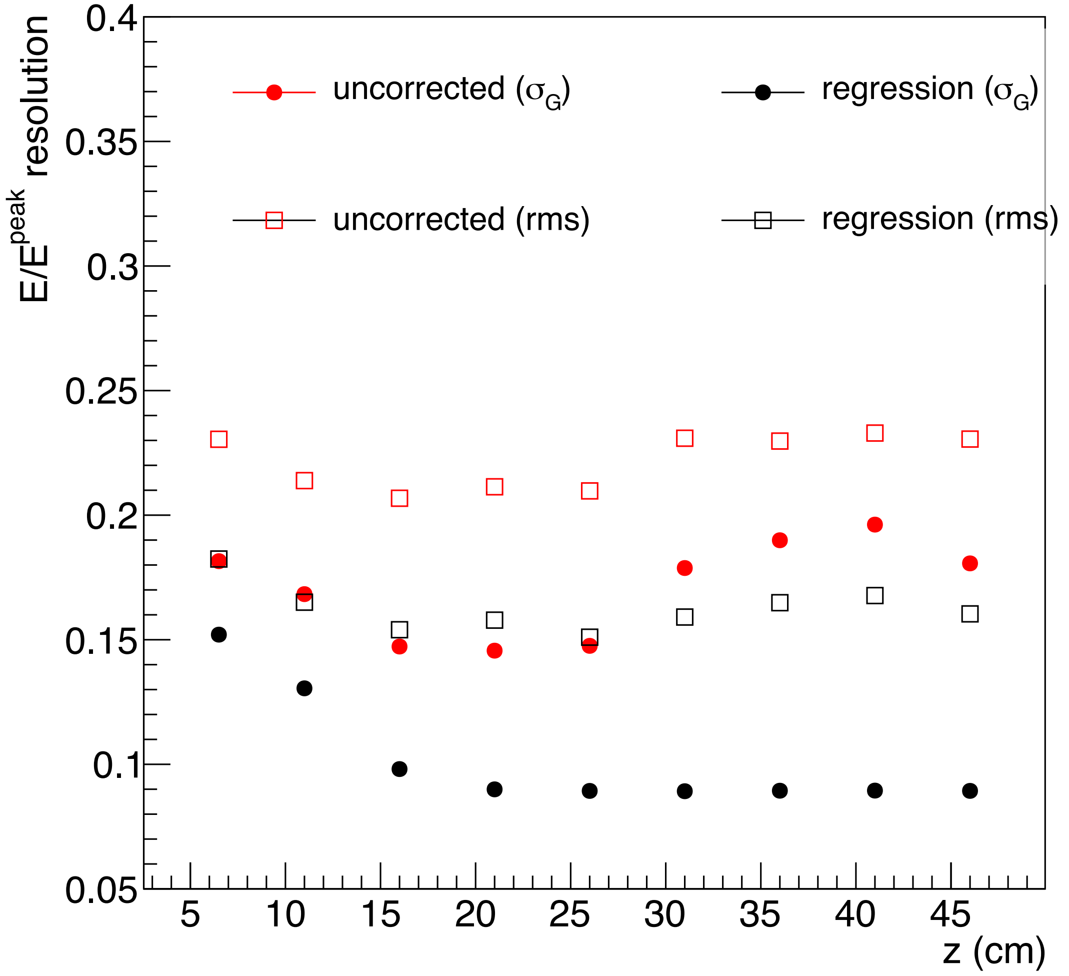


Results vs Z

light yield peak



light yield resolution



Regression does NOT correct (yet) for saturation

=> look for more sensitive variables

Regression cures the variation vs z when there is not saturation

Resolution significantly improved everywhere

Core Gaussian resolution can be better than 10% (if no saturation)

Conclusions

- We have now in our wardrobe (*github*) clothes (*clusterings*) for many Terrestrial seasons, on the surface, and under the surface

- We have analyzed most of the data taken with LIME so far at LNF, which is the worst situation in terms of backgrounds, to get results on efficiency and energy response in a wide range of energies

- The energy linearity in response to X-rays is reasonable in the range [4.5 - 50] keV

- raw energy resolution is about 15%, but MVA regression can improve it up to 7%, even if it doesn't correct yet for the saturation

  After the MVA regression, the saturation introduce a non linearity of max 20% for the closest Z tested with ^{55}Fe source (5 cm from the GEM)

- the same A-Z analysis should be performed on simulation to validate it (and to understand the origins of response differences)

- when the simulation can reproduce the performance observed in data under many aspects (energy, cluster shapes, etc), we can trust it to make physics projections (e.g. NR vs ER with sophisticated techniques) based on SIM.

- **We should write a paper summarizing LIME detector performance with LNF data.**

Backup

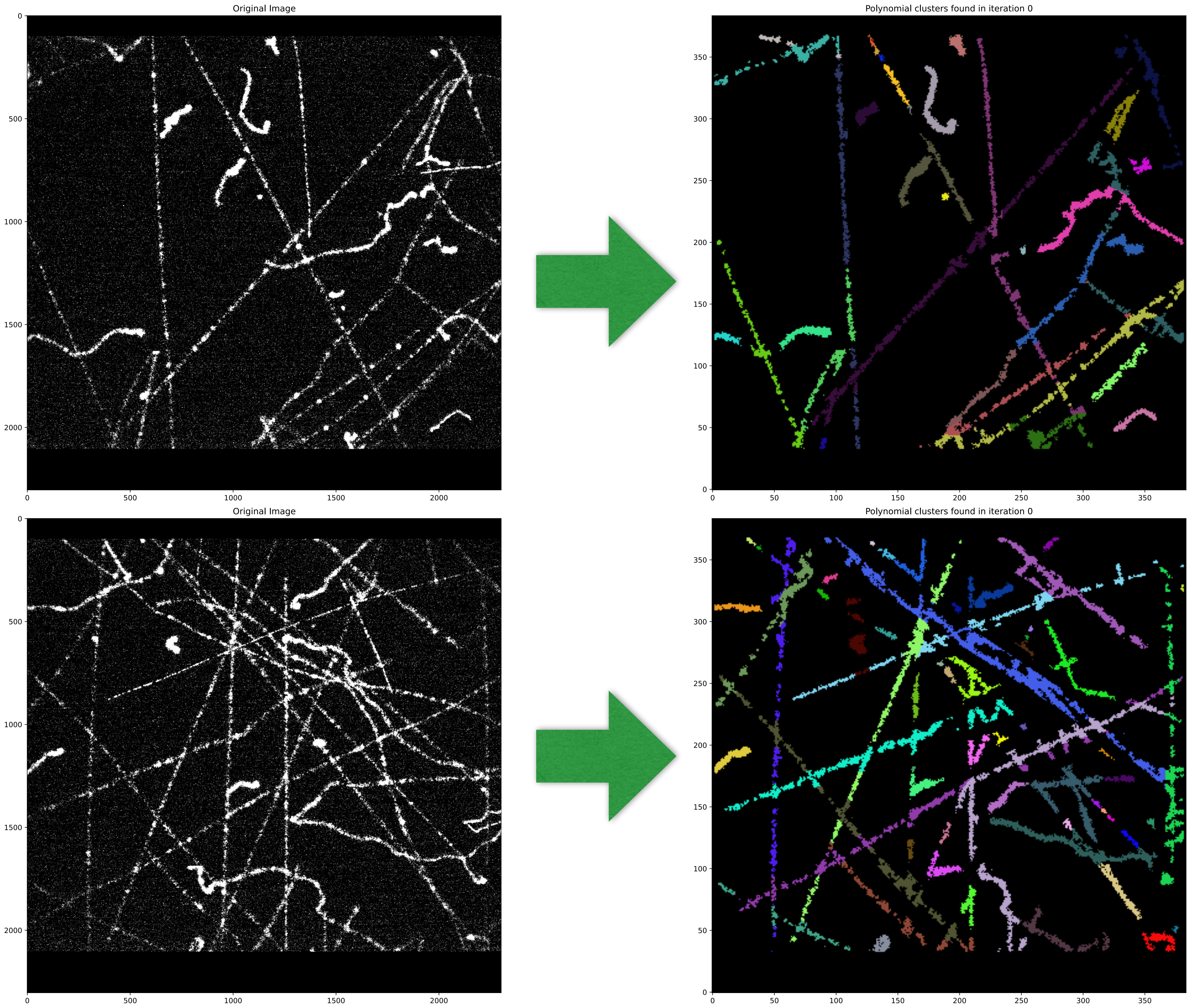
Pileup removal strategy

Developments in “autumn21” GIT branch (UNSTABLE!)

Starting from I. Pains directional clustering (3D “weighted” version, i.e. each pixel has weight = #photons)

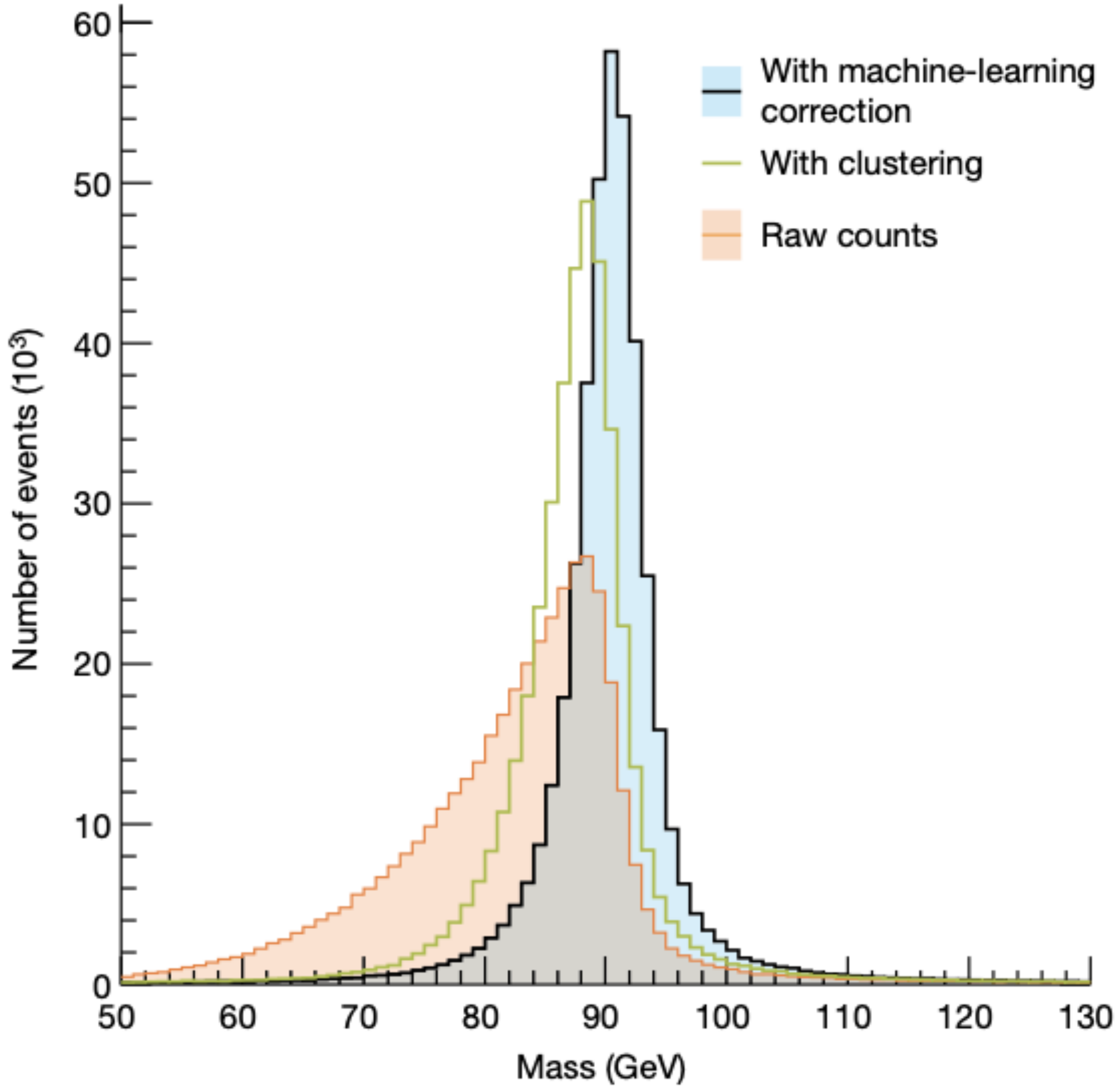
- Done the minimal to run and achieve something reasonable:
- rebinned image x6 (x4 would be better to resolve overlaps, but too slow with this pileup)
- improved fits for the directional tracking
- tightened the isolation requirement when looking for “signal” small clusters after the long tracks have been reconstructed
    - preferred smaller efficiency to the risk of getting unclustered pieces of long tracks
    - residual subtraction will be done by the statistical analysis

Output super-clusters



Regression used in CMS-ECAL

We used it extensively to correct the energy response of the ECAL in CMS wrt many effects (local containment, pileup dependency, etc)



$Z \to e^+e^-$ invariant mass

[1] 10.1088/1748-0221/10/06/P06005

[2] 10.1088/1748-0221/10/08/P08010

MVA implementation

- Input variables are used to train a multivariate regression using the Gradient Boost Regression (based on a BDT in scikit-learn).

- GBR target is integral/9000 (to have a variable centered at 1)

    - normalization also helps in reducing the phase space of the target variable when training with variable energy clusters

- The loss function are:

    - mean squared errors

    - 50% quantile (median), and 5% and 95% quantiles

        - 50% quantile gives the central prediction, the other two give per-cluster energy resolution estimates (+ and - asymmetric errors)

- Detailed training options to be further optimized

Data used for MVA regression

July 30th runs with ^{55}Fe with V_{GEM1} = 440 V at different Z values: 46, 36, 26, 6cm

- here focusing on Energy, but a dedicated regression can target Z (exploit dependency of cluster shapes - through diffusion, mainly) to give a per-cluster Z estimate
  - a straightforward way to make a 3D reconstruction. Need more Z points to test this (e.g. data taken in April and analyzed by Donatella)
- About 8000 clusters used, 90% for training, 10% for testing
- A thought for the future:
  - data with the source moved on all 4 sides of LIME would help covering uniformly the XY plane with spots