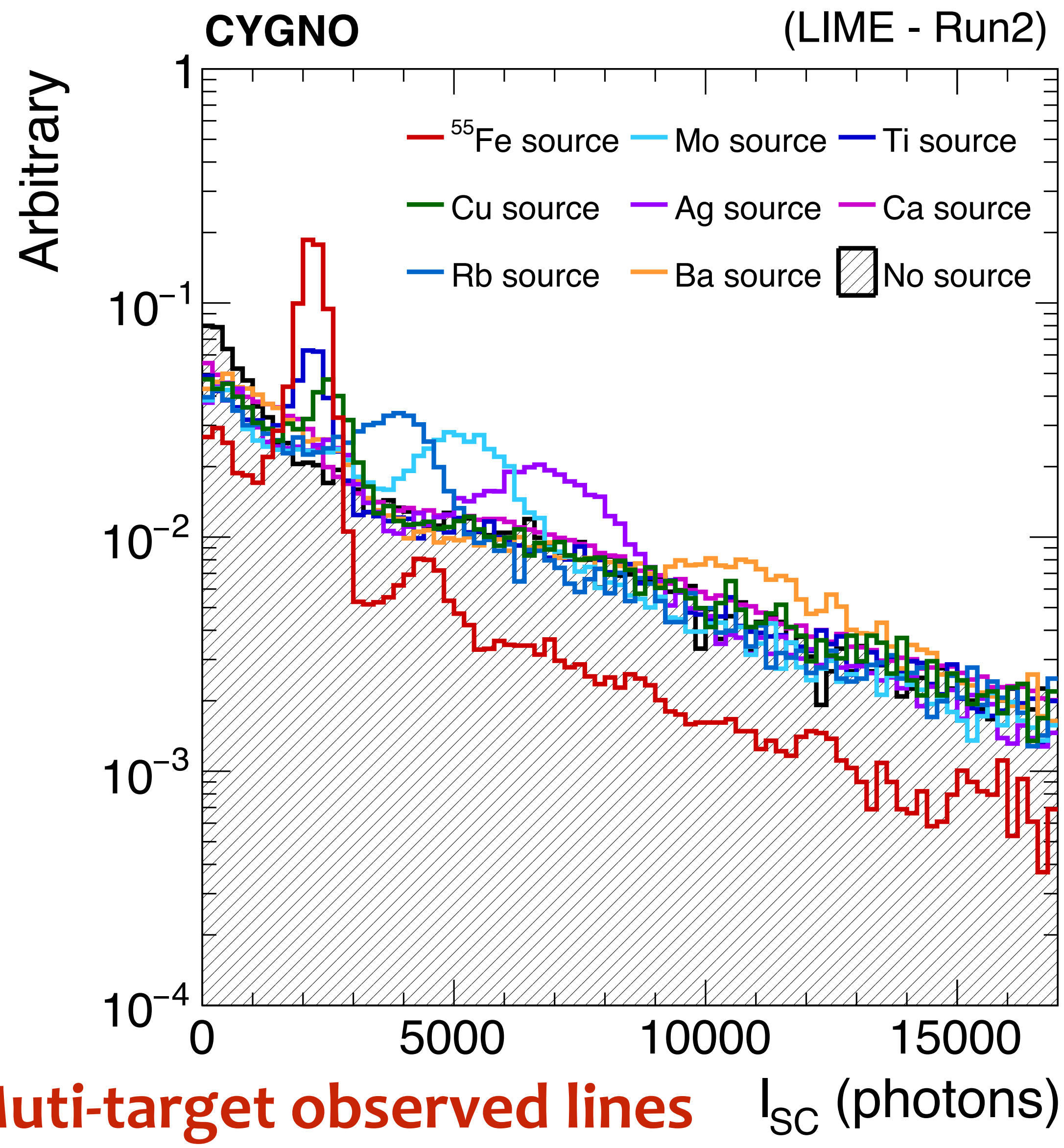
LIME energy response, corrections, and efficiency

G. Cavoto, E. Di Marco, D. Pinci

Analysis meeting, Coimbra, 7 June 2023

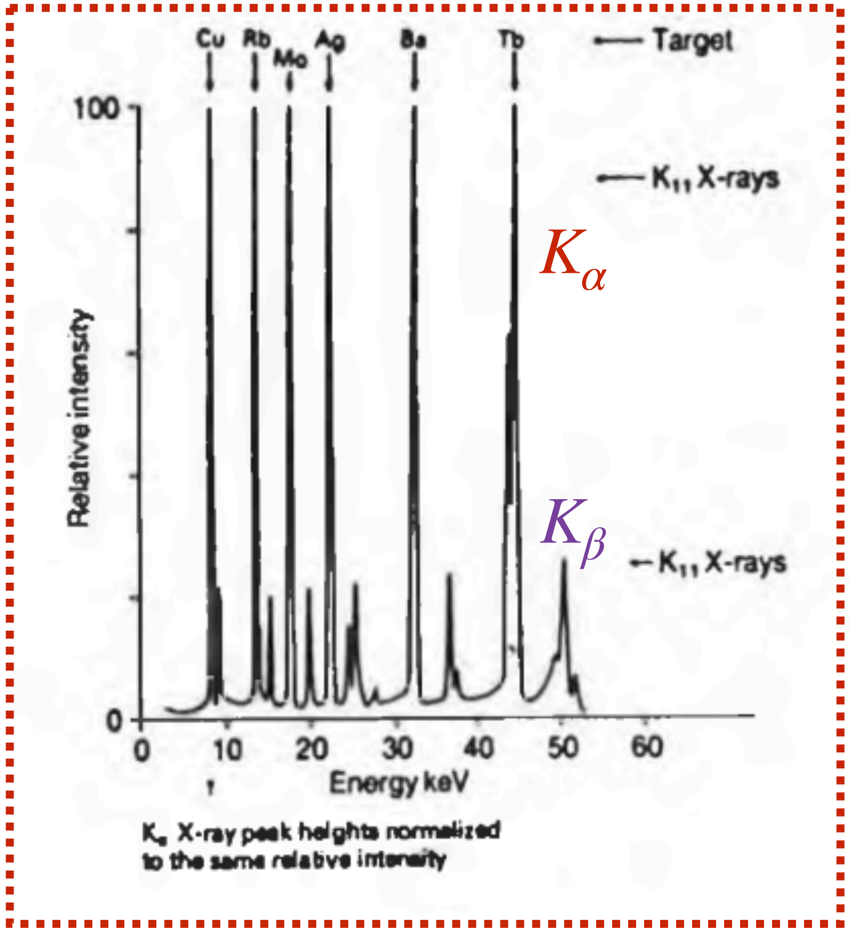
Energy response: the candles

- We have studied the absolute energy response of LIME with multiple sources
  - Give Photons => Energy [keV] absolute calibration
    - @LNF: ^{55}Fe with large activity (115 MBq) => high precision calibration at E=5.9 keV
    - @LNGS: multi-target X-ray source => used to test the linearity of the $LY = f(E)$ in the range [3.7 - 36.6] keV



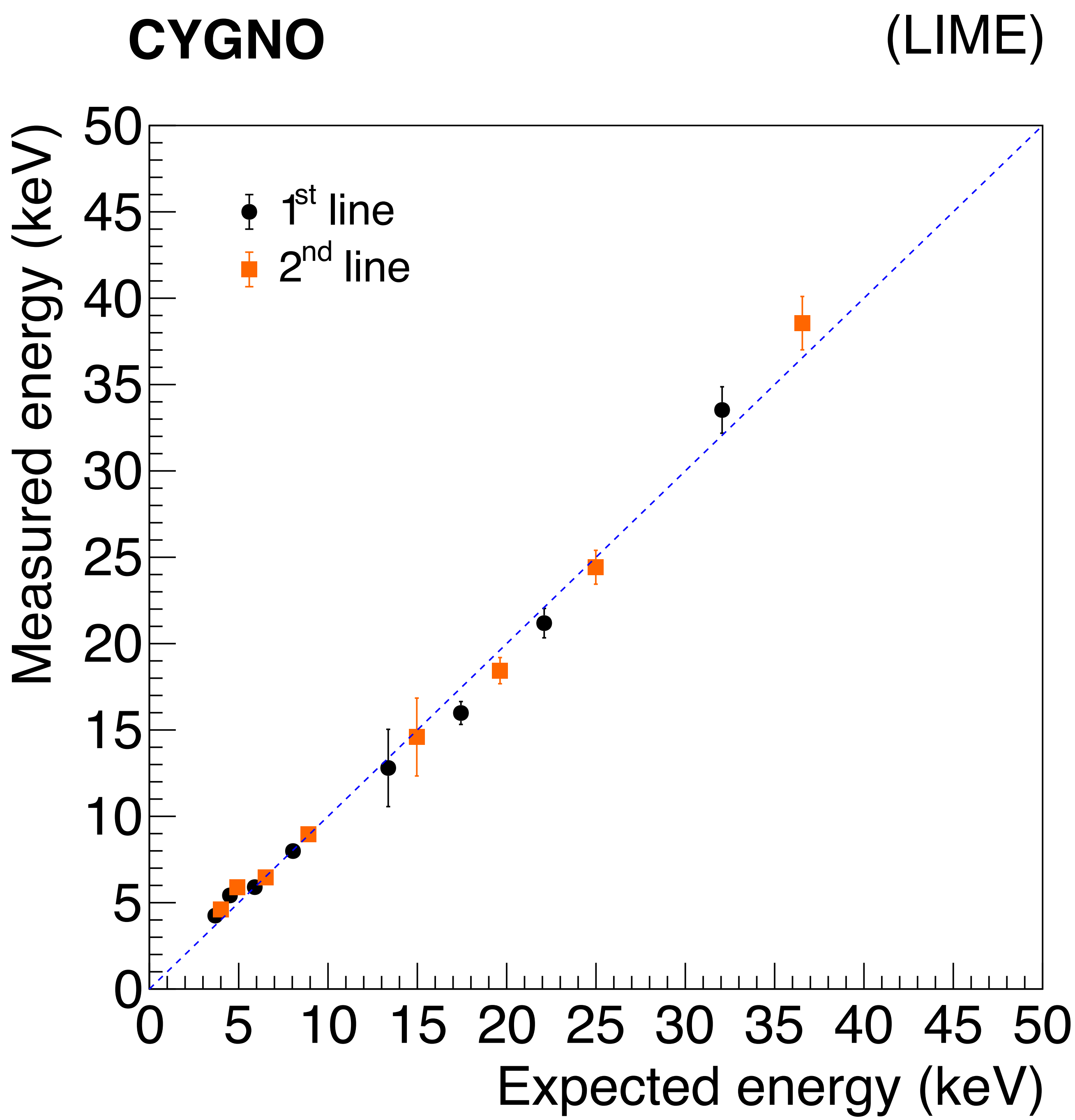
Fe & Muti-target observed lines

Muti-target expected lines



Energy linearity

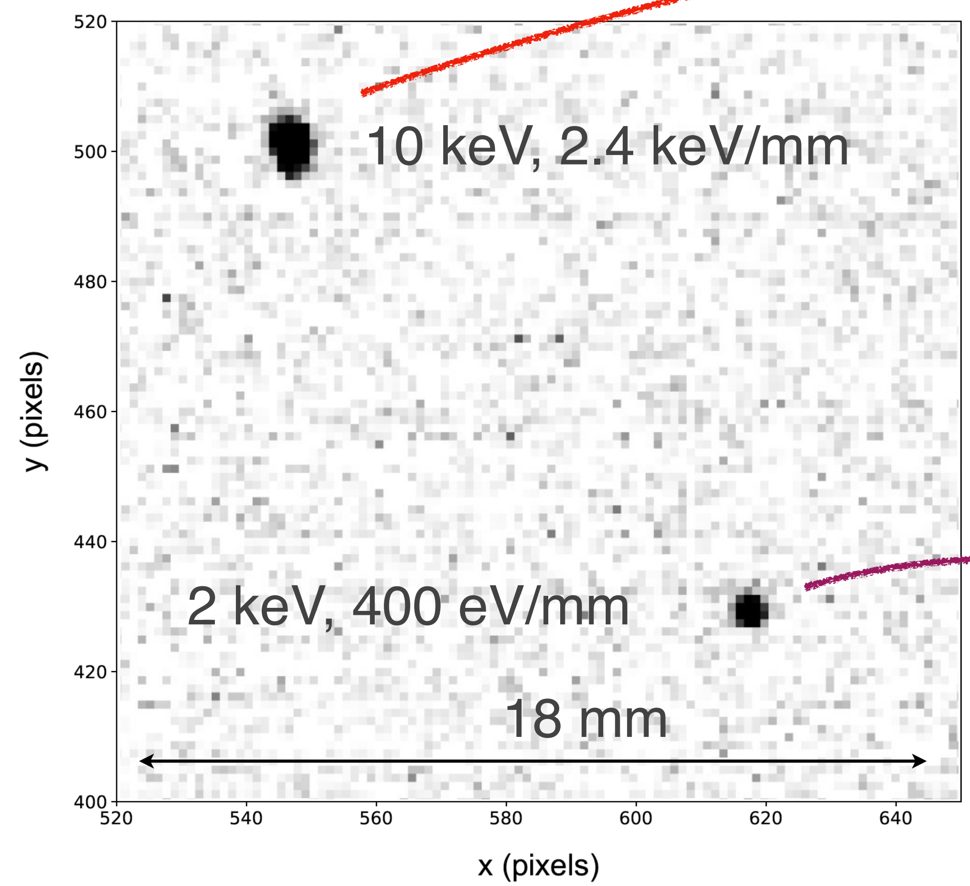
- Response for low-energy X-rays fully contained in the active volume reasonably linear
- Cannot repeat at LNGS (no multi-target source), probably we don't need it



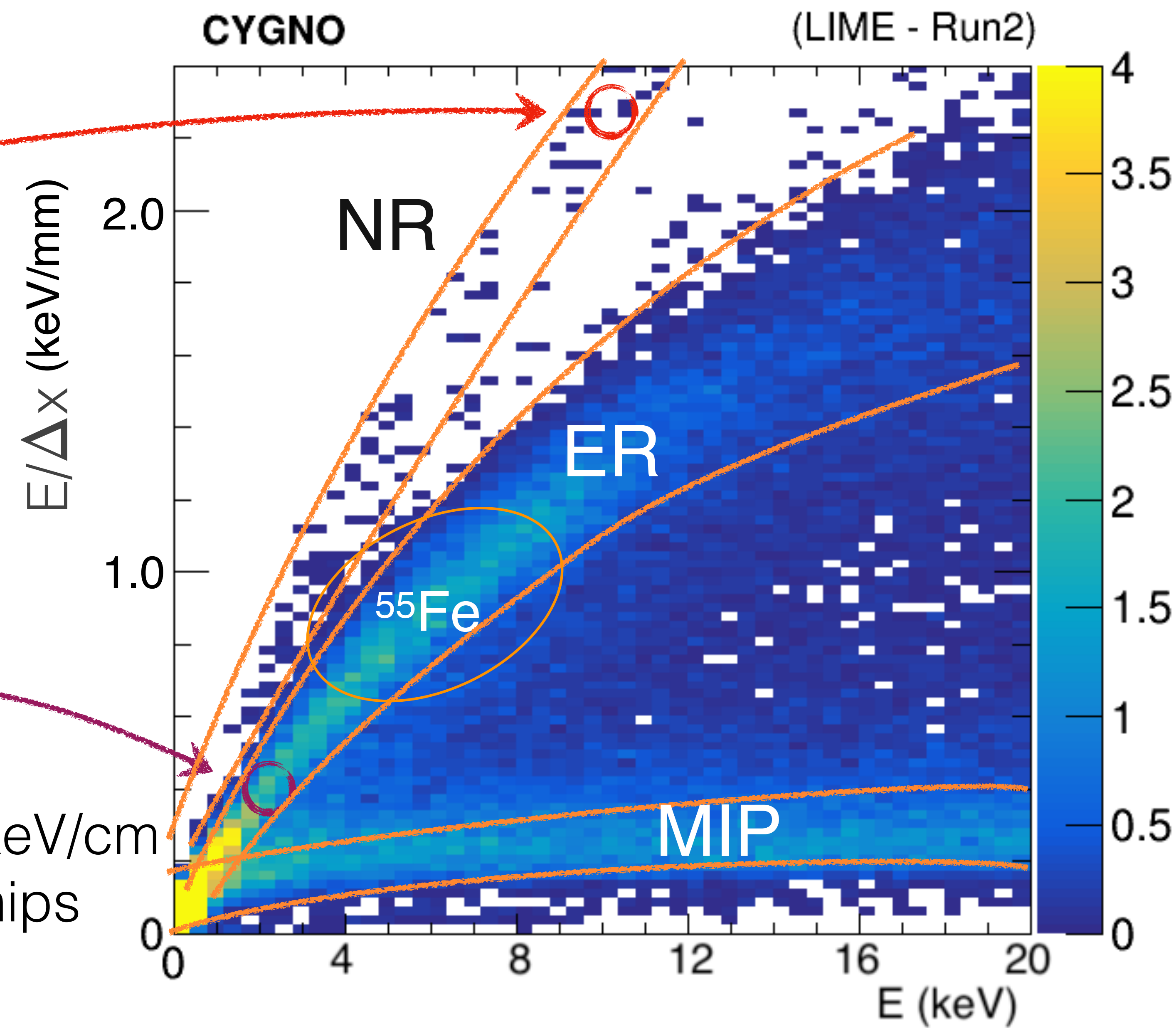
X-rays: not only LY - 1

- While waiting for the revamped gas system and the AmBe source to produce neutrons and a nuclear recoils calibration dataset, we can check the speculation on the "ER-band" below with LNF X-rays at different E
   - Remember: Δx for most of the multi-target X-rays is constant, because dominated by $\sigma_{\text{diffusion}}$, so $E/\Delta x$ increases with E (apart for Ba at E~30 keV)

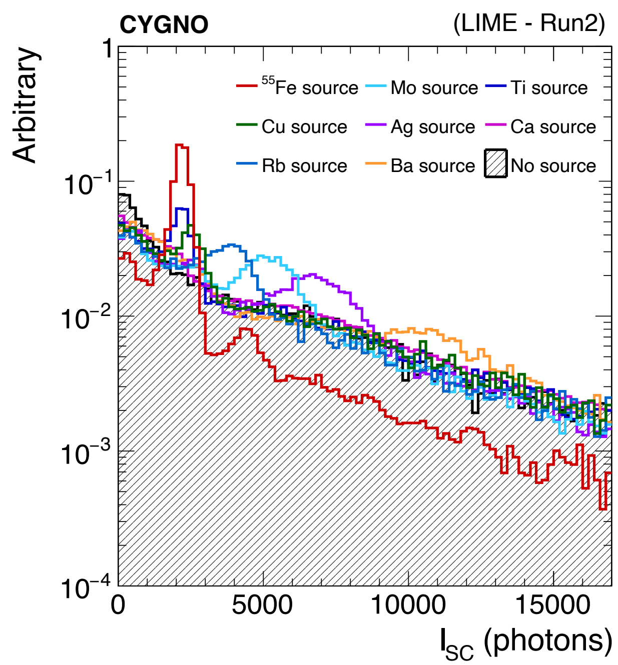
E/Δx variable can be used to have a first look to discrimination power



2.3 keV/cm mips

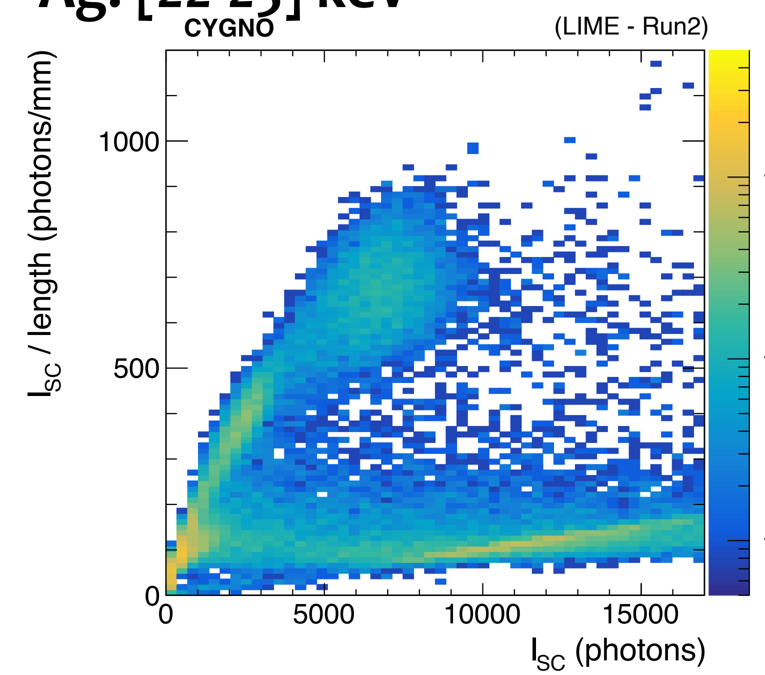


X-rays: not only LY - 1

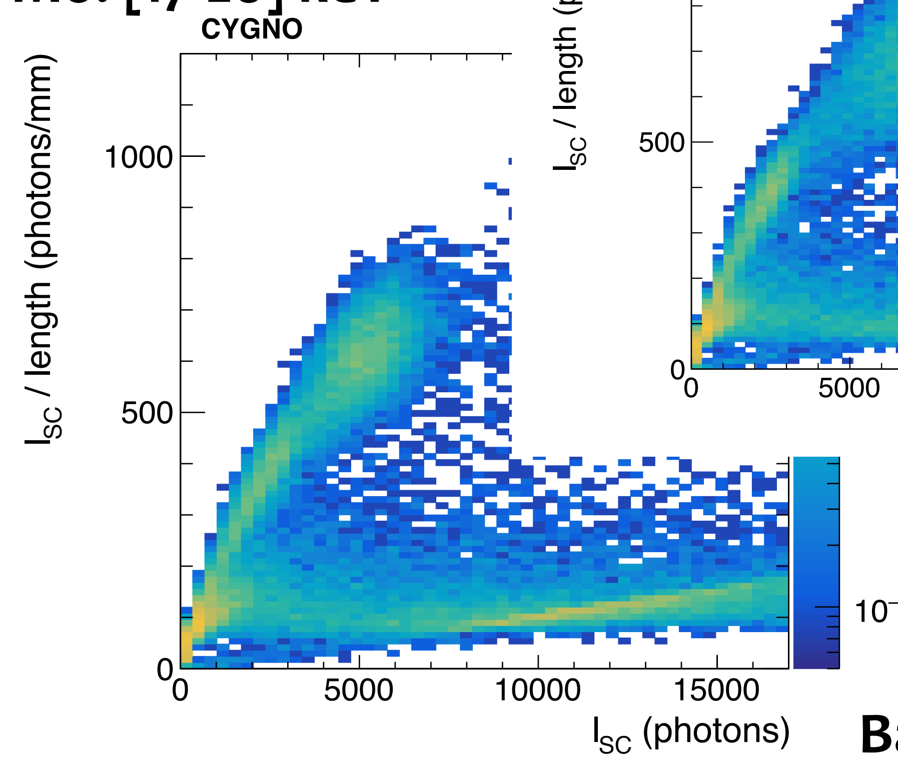


This proves that the ER band
is really ER-populated

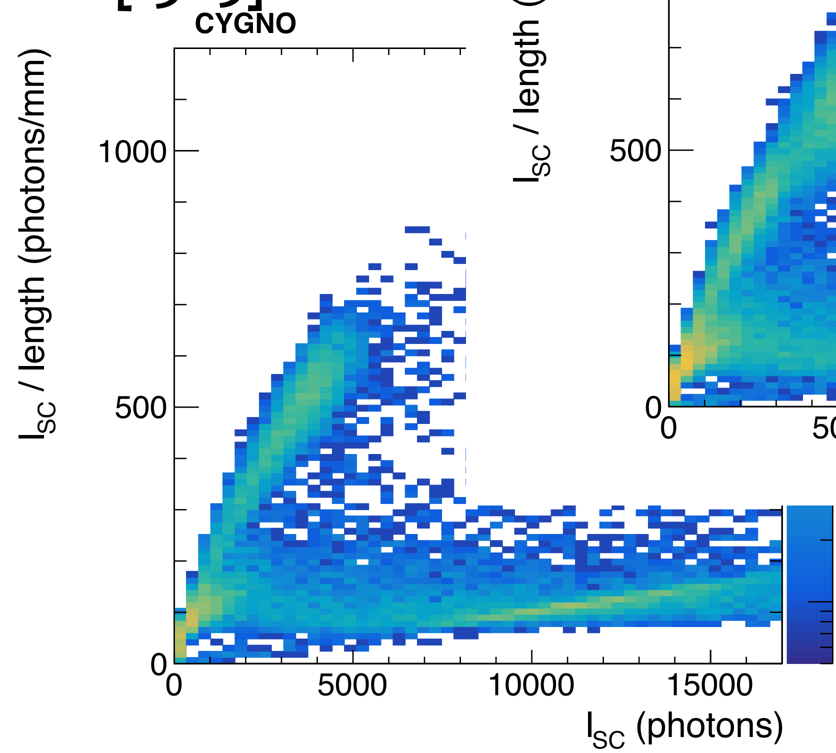
Ag: [22-25] keV



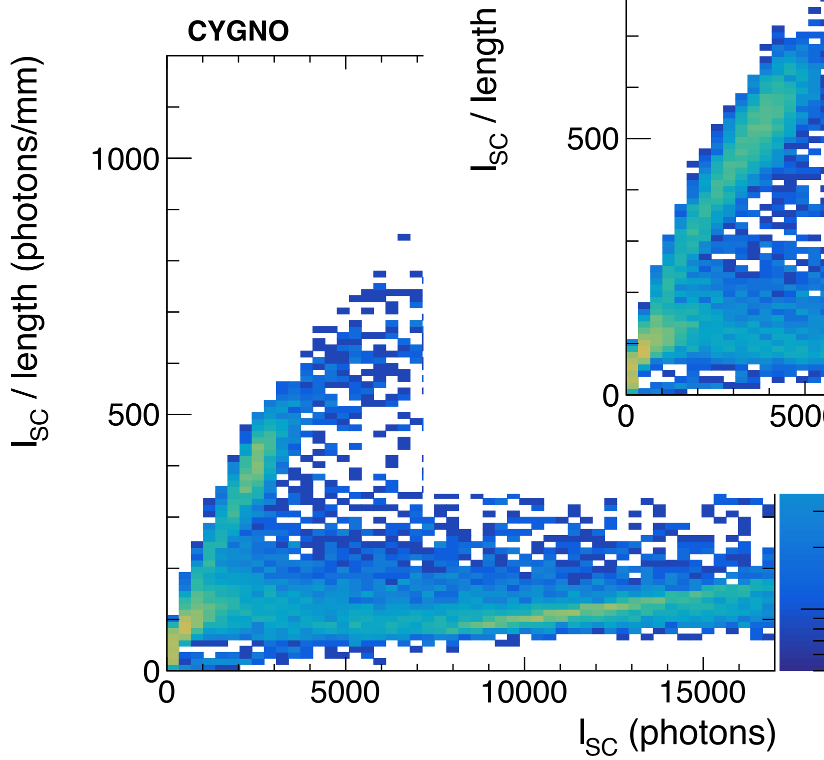
Mo: [17-20] keV



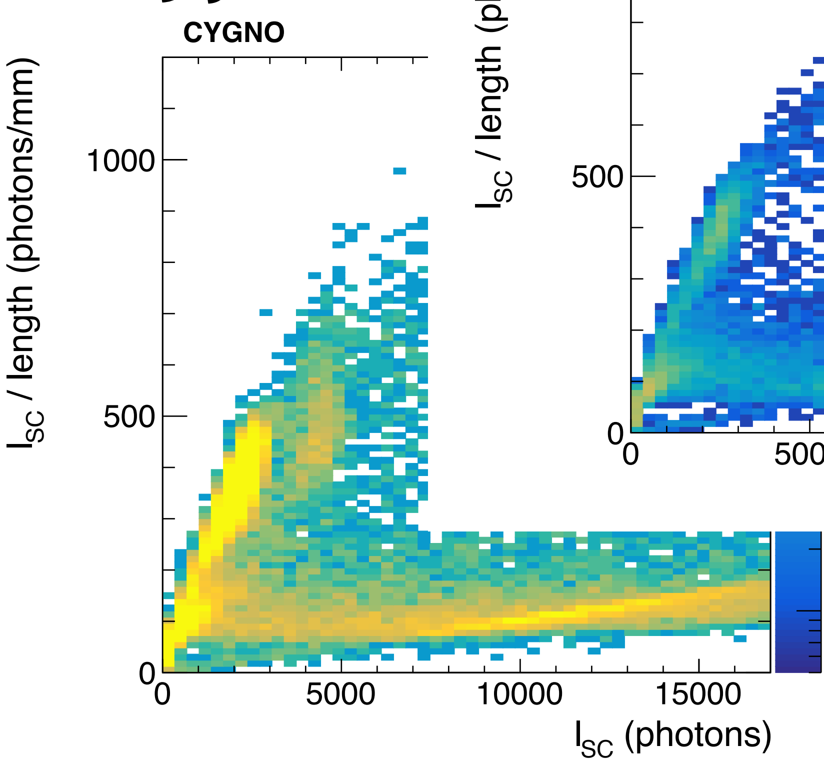
Rb: [13-15] keV



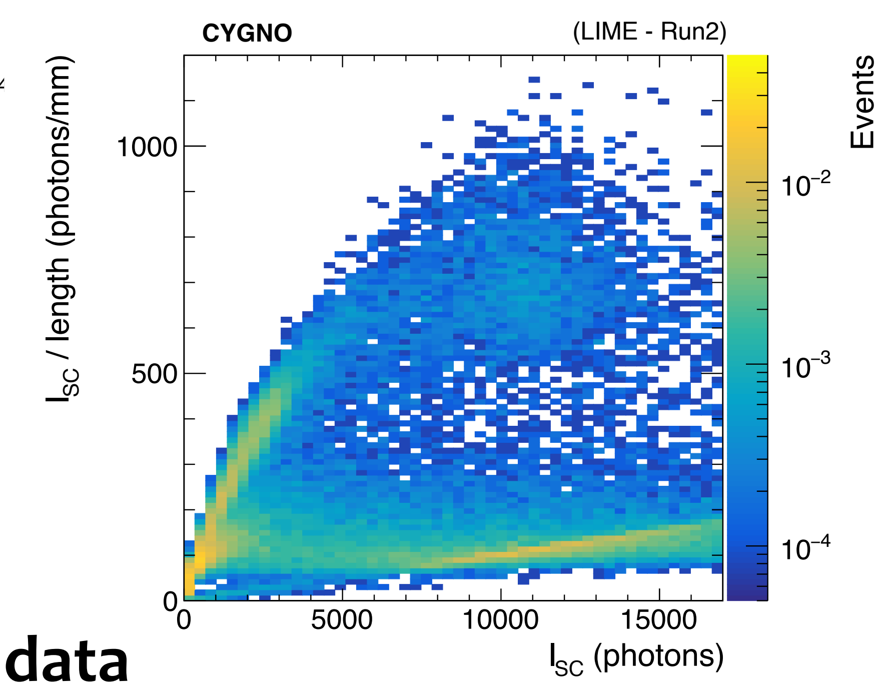
Cu: [8-9] keV



^{55}Fe: 5.9 keV



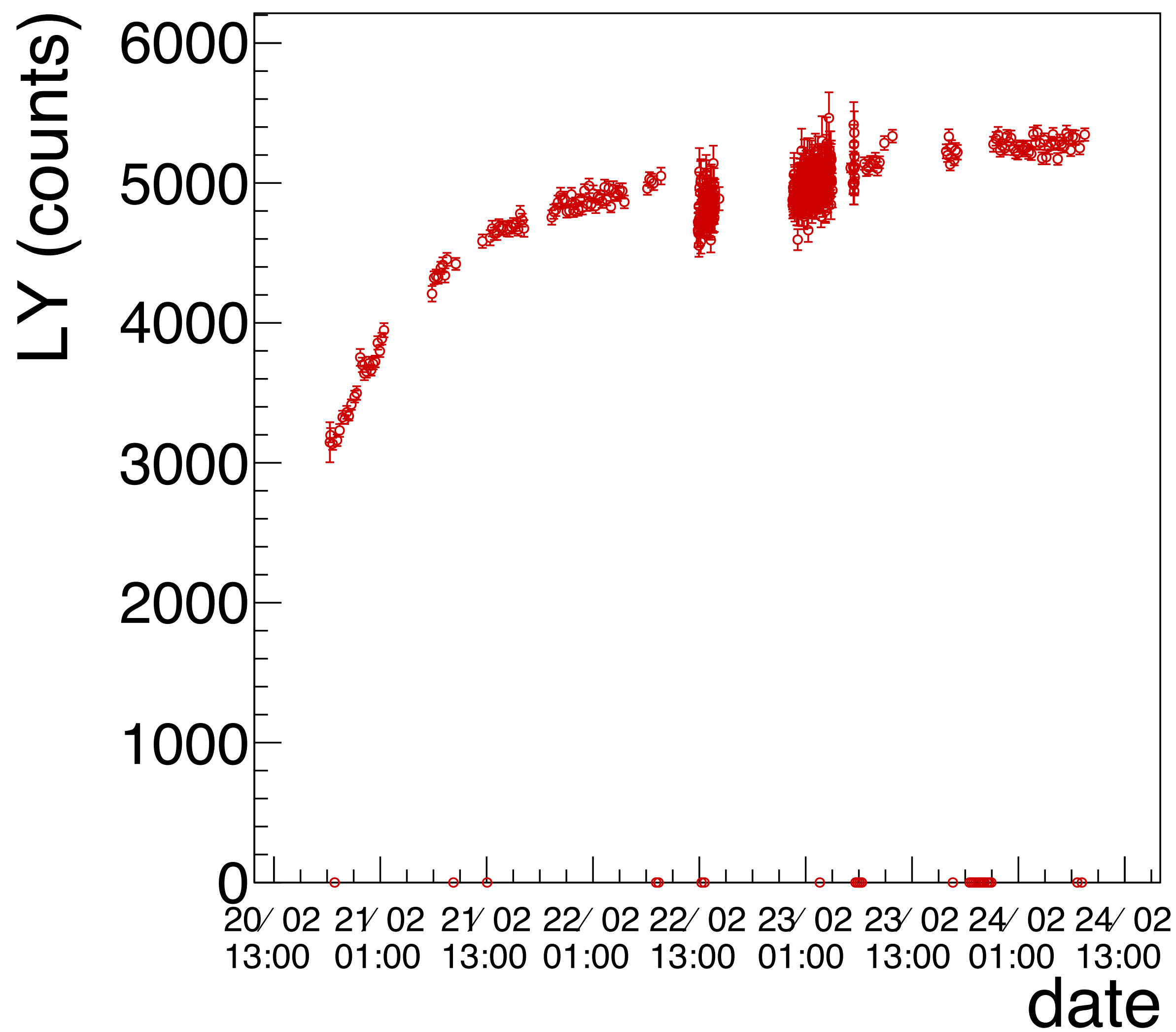
Ba: [32-37] keV



For the NR "band", need AmBe data

X-rays: not only LY - 2

- We constantly monitor LY response with a lower activity ^{55}Fe in the daily calibrations
  - Vs Z (saturation)
  - Vs HV (gain)
  - Vs time (stability)
  - Vs gas conditions: P, T (gain)
  - => See Rita Roque's talk at this meeting !



- Also useful to monitor:
  - Efficiency (# of clusters)
  - Noises
  - Hot-spots
  - => See Rita Antonietti's talk at this meeting !

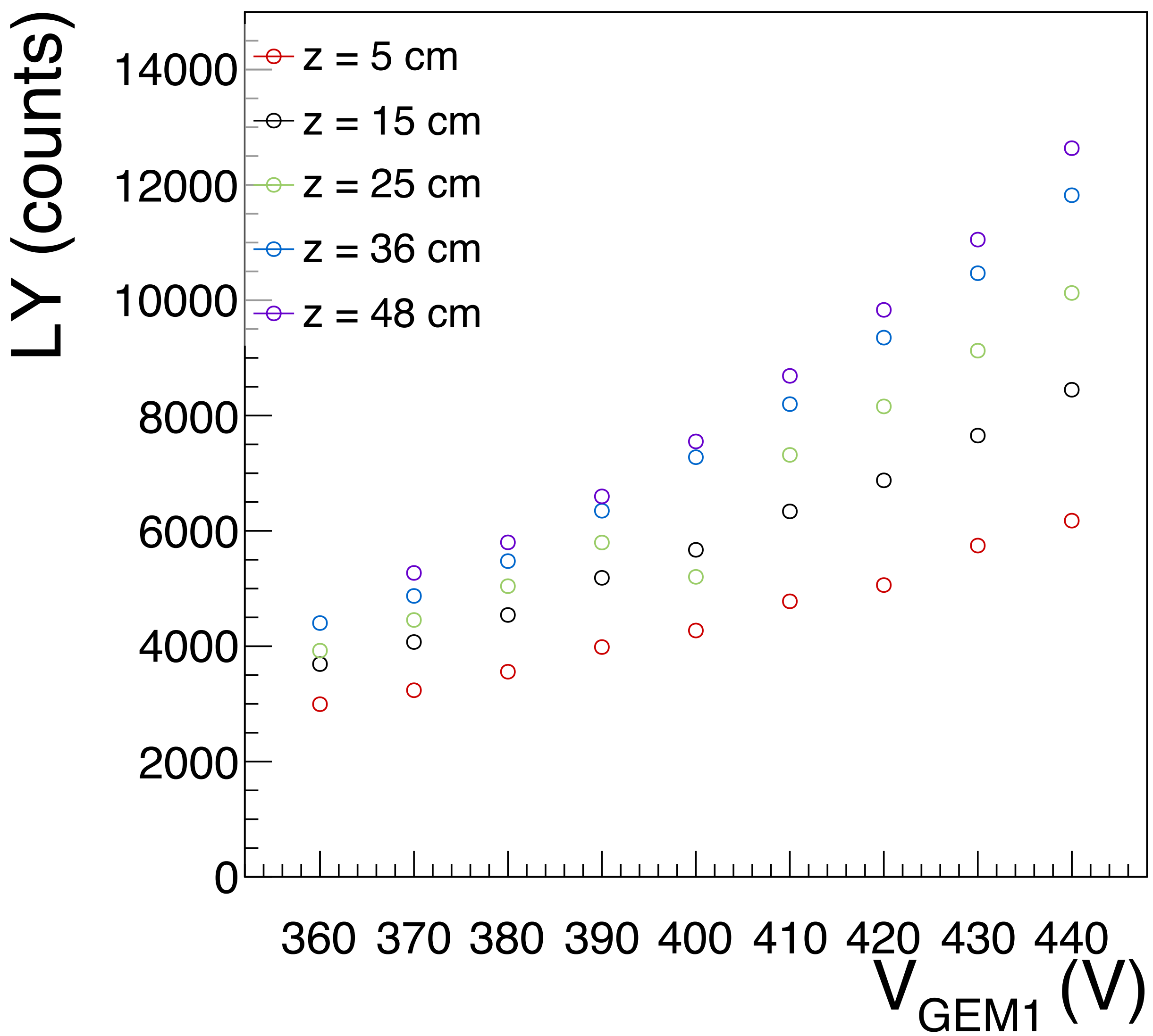
- Check the general conditions of the detector
  - Can be used in automatic Data Quality Monitoring (DQM)
    - See I. A. Costa's talk at this meeting !

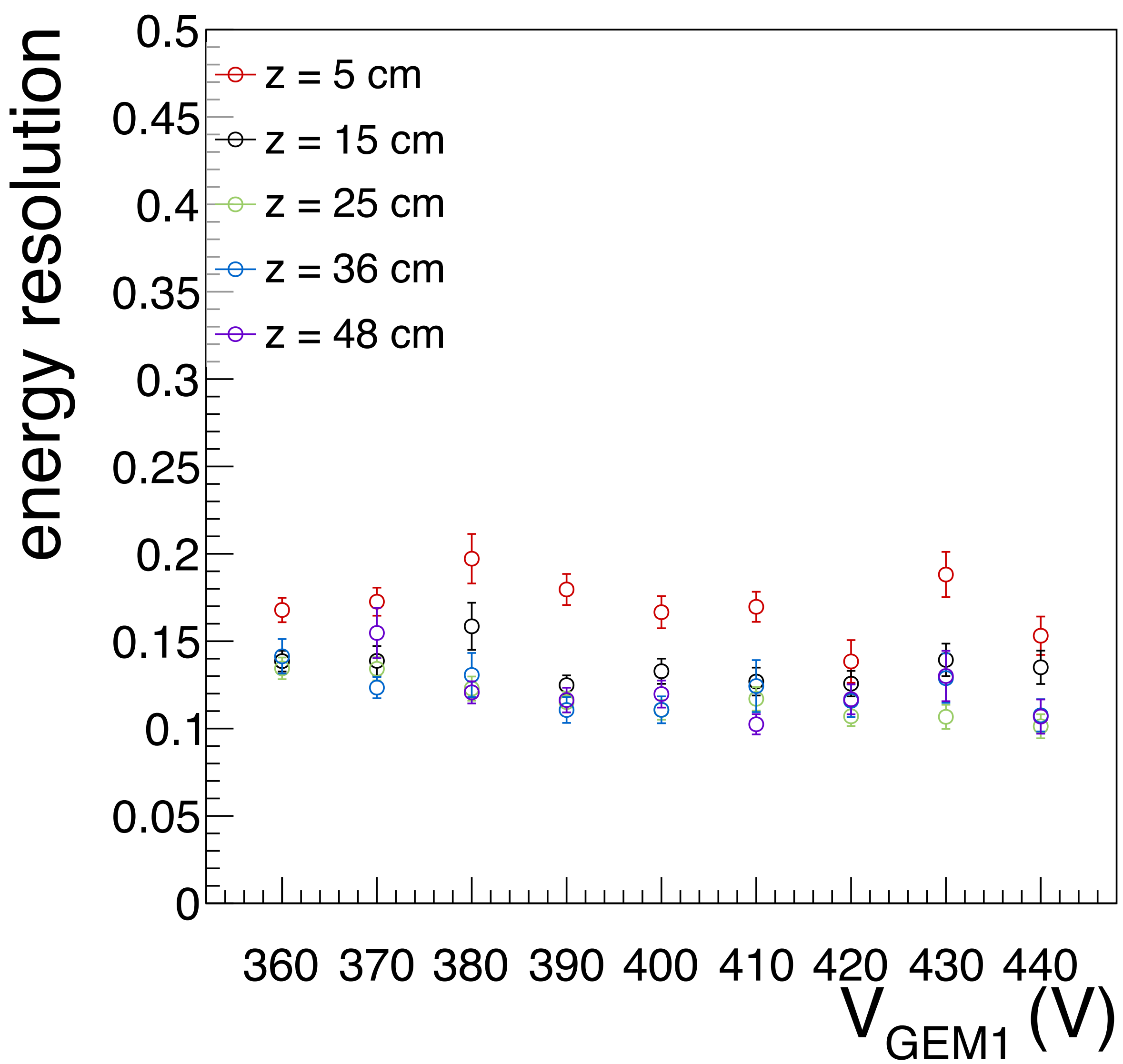
LIME Run 2 calibrations

- Regular calibration data taking with ^{55}Fe with HV=420 V or also HV=440 V
  - HV = 420 V → safer value for detector stability
  - HV = 440 V → point at highest LY that was running stably
    - ⇒ took our “golden” Run2 no-source datataset in this configuration
    - At HV=440 V: $LY_{z-max} \approx 1.4 \times 10^4$
  - We also took a 2D scan [HV - z]
    - Useful to train energy regression mimicking (a bit) variable energy from a fixed 5.9 keV source
    - Useful to study in great detail the gain variations vs detector conditions

Light Yield vs V_{GEM1}

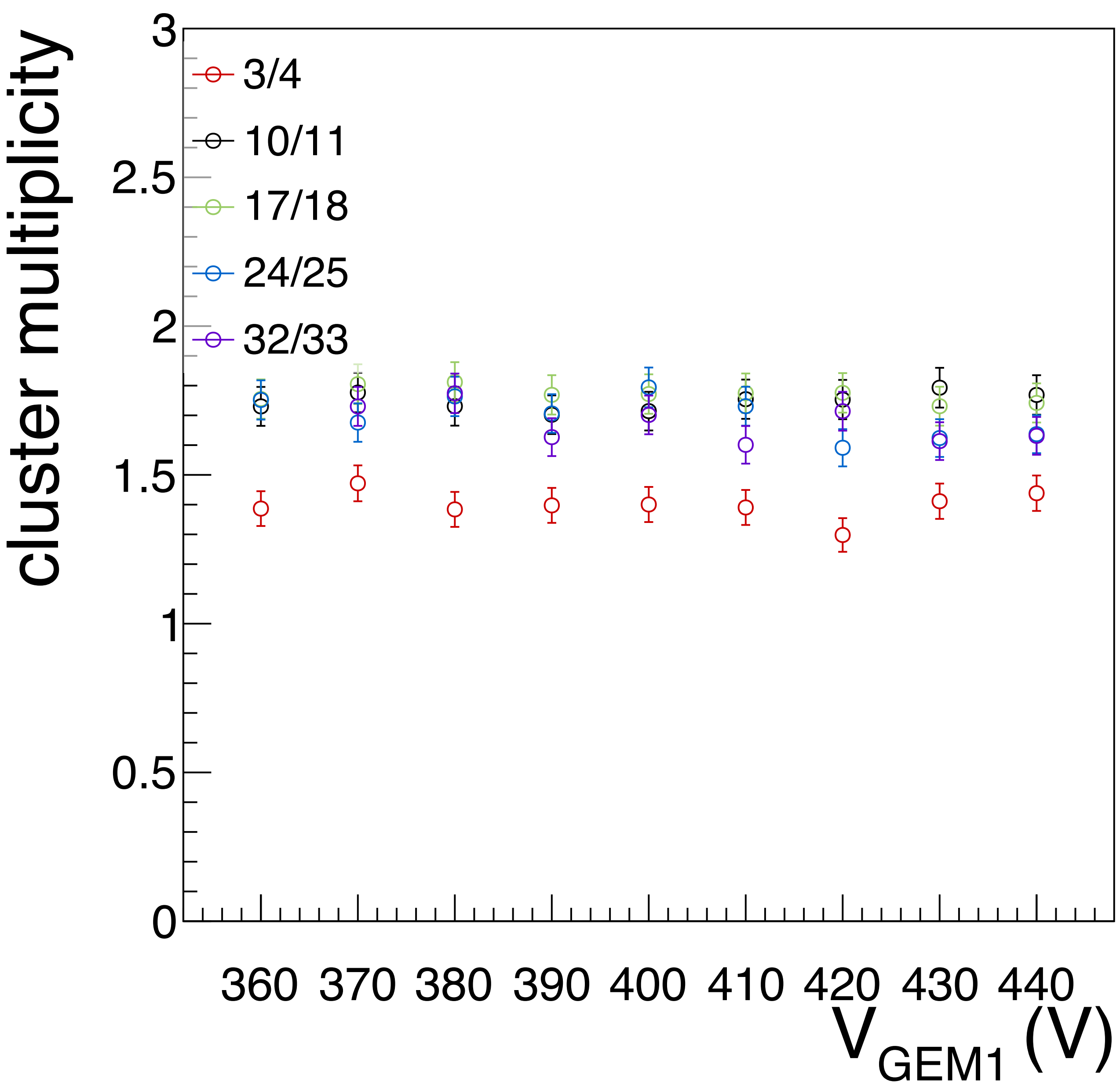
- LY increases with V_{GEM1}. From 360→440 V increases x 2.8 (2.0) at z=48cm (5cm)
  - i.e. saturation makes a difference also on the derivative vs VGEM1
- LY reduced x 2 from z=48 cm to z=5 cm even at the largest HV=440 V by saturation
- Energy resolution around 12 - 15% (apart for z=5 cm)





Efficiency

- Efficiency of reconstructing clusters down to small energies
  - For absolute efficiency one needs the MC (true number of X-rays interactions in the gas)
  - Still, can look at “turn-on” of efficiency wrt one reference point (eg. z=48 cm, HV=440V)



- Average number / image of clusters:
  - within R=800 pixels from the center
  - with L<5 cm
  - loosely compatible with a round spot

~ constant down to 360 V

For z=5 cm, efficiency ~ 80% despite LY~50%

Points below 360 V existing, to be analysed

Run-2 energy regression

- General principle is to derive a best estimate of the dependent variable (in our case the ***true cluster energy***, or the ***Z position*** of the interaction) given a set of **measured variables** (measured light, position in XY, cluster shape parameters, etc)
  - One objective is to correct the saturation effect, which depends on Z
  - A similar objective is determine Z (for 3D reco, fiducialization, etc.)
  - Main handle can be the cluster shape, which through diffusion have a transverse size $\sigma_T \propto \sqrt{z}$
    - e.g. $\eta = \sigma_T / A_T$ used with BTF electrons gives 20% precision. Rita Roques' Linear regression gives a $\sigma_z \approx 6\,\mathrm{cm}$
- But the light response (and the estimated $\hat{z}$) depends not only on z_{true}, but simultaneously on many quantities, $(\vec{\theta})$, which are in general correlated
- ☞ Use this dependence, and also the correlation information, to make a model to predict the true energy E_{true} (and z_{true}) as a function of the measured cluster shapes: $\hat{E} = f(\vec{\theta})$, and $\hat{z} = g(\vec{\theta}')$
  - Given that the saturation is the main effect that we want to solve, and this depends on z_{true}:
    - the two sets of variables $\vec{\theta}$ and $\vec{\theta}'$ have a lare overlap ($\vec{\theta}$ contains also I_{SC}, $\vec{\theta}'$ don't)
    - the training can be mostly the same
- The MVA regression is a way to make this inference in n-dimensions
  - Useful because the cluster shapes depend also e.g. on residual x-y position of the cluster (residual vignetting, optical distortion, electric field non-uniformity…)
- 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

LNGS Run-2 regression dataset

- At LNGS we have for now only the ^{55}Fe source, so fixed energy
  - We can still vary z as uniformly as we want, and we took data for $z = \{5, 15, 25, 36, 48\}$ cm
  - We mocked up variable E_{true} varying HV_{GEM1} in $[360 - 440]\ V$ range in steps of $10V$
    - In terms of LY is a variation by a factor ~3. Assuming 440V = 5.9 keV => $E_{\text{true}} \in [2.0 - 5.9]\,\text{keV}$
  - With this 2D scan $[E_{\text{true}}, z_{\text{true}}]$ we can correct for $\hat{E}$ saturation for a range of E_{true}
- BIG limitation(s):

1. The interactions are still the ones of **fixed** $E = 5.9\,\text{keV}$ X-ray, i.e. some cluster shapes which for physics depend on E_{true} are not representative of real X-rays of variable E_{true}
   - We are mocking up variable E_{true} only changing the LY by changing the GEM gain
     - Obvious example: track-length. To make the model more general, don't use track-length proportional variables.
       - When applying it, we can only apply to short tracks, or cluster-by-cluster segments of the track (but it requires running it during the reconstruction, not post-reco)
2. The interactions are for X-rays, it **might be not applicable to other kinds of interactions** (eg. NRs)
   - This is probably only 2nd order effect: since the main target is correct for saturation and x-y non-uniformities, and the main sensitivity comes from diffusion, and so by transverse cluster dimension, it might be similar for any type of interaction
3. The source illuminate only the central strip of the detector in x. In the future can think of inclinate the source to populate more the detector?

Training dataset

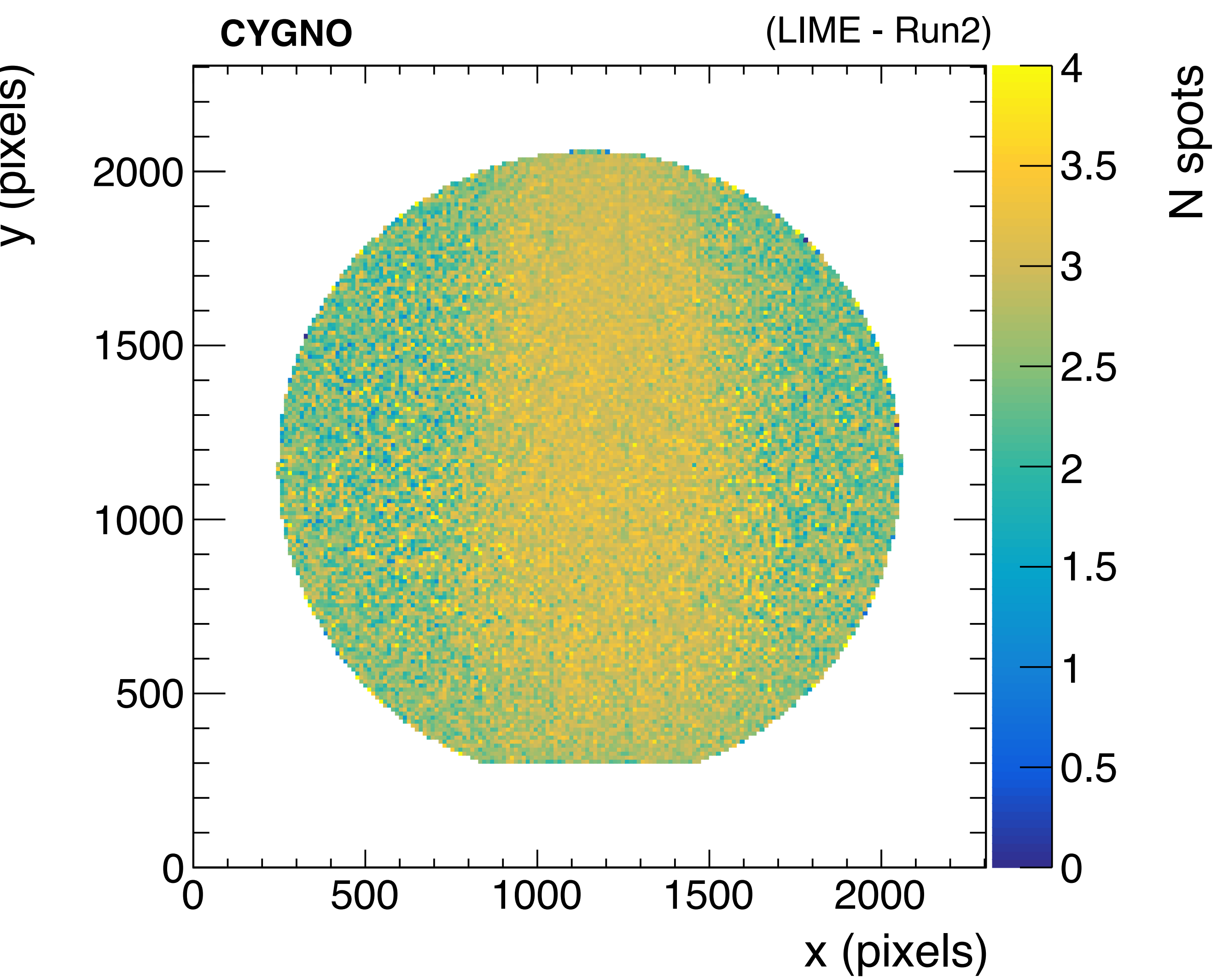
- Used the 2D $[E_{\text{true}}, z_{\text{true}}]$ scan with ^{55}Fe source taken Feb 22nd. Each point has 400 events

22-02 16:02 — to — 22-02 23:25	Scan VGEM 1	Yes	20	///	9352-9446
22-02 23:23 — to — 23-02 09:40	LY vs time	Yes	20	420	9447-9710
22-02 09:40 — to — 23-02 13:00	Scan VGEM 1	Yes	20	///	9711-9753

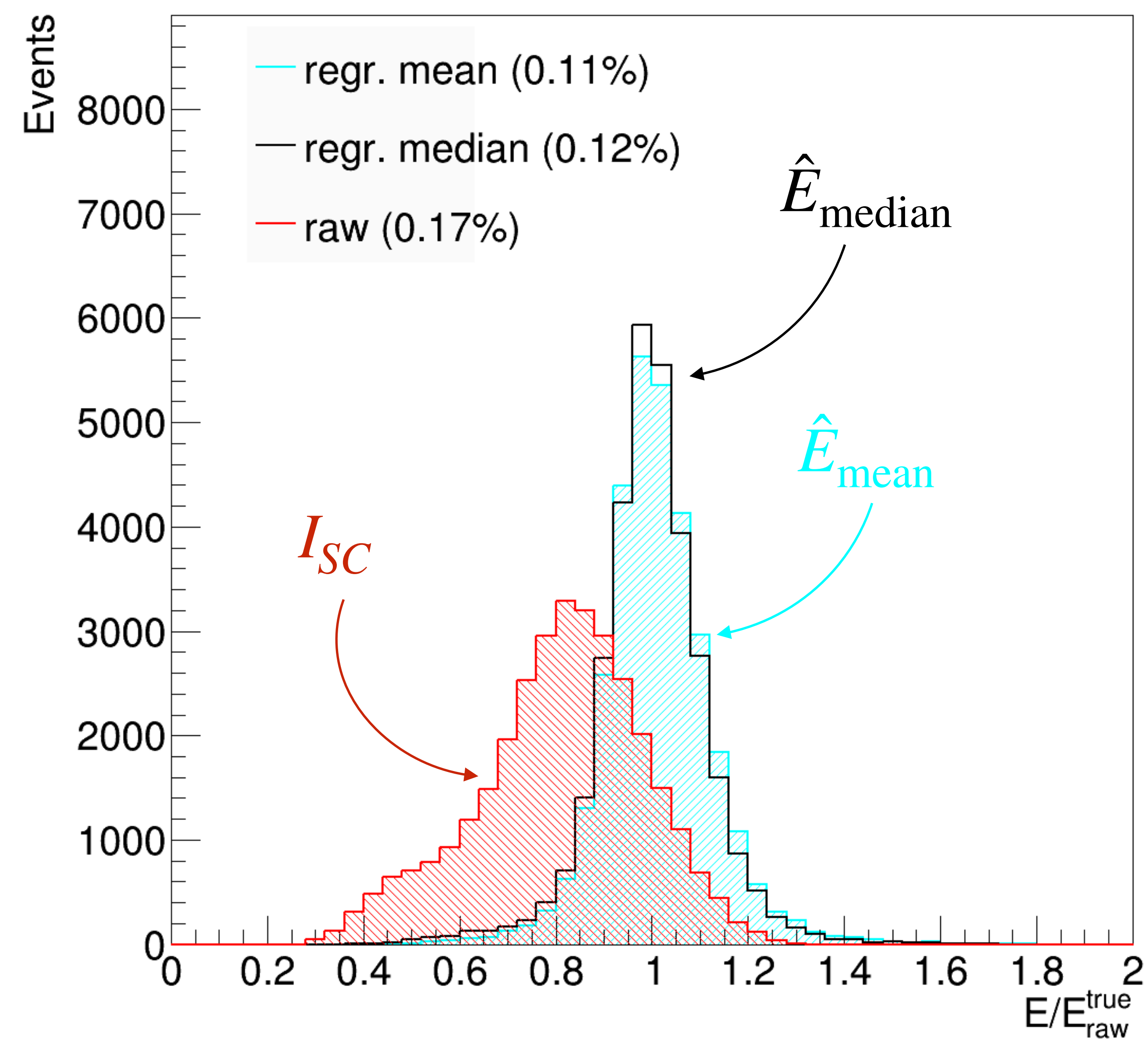
- Set of variables used for energy regression:
  - $\vec{\theta} = [I_{SC}, \delta, I_{\text{rms}}, x, y, \sigma_T, \text{width}]$
  - Model: Gradient Boost Regression (GBR) with a Boost Decision Trees algorithm
  - Model parameters: max_depth=3, min_samples_split=6, min_samples_leaf=7, learning_rate=0.1, n_estimators=500
  - Target: peak of the $I_{SC}^{z=48\text{ cm}}$ (supposed un-saturated) distribution
    - **Mean regression:** the mean of the output distribution matches E_{true} (this is our $\hat{E}$)
    - **Quantile regressions**: a given quantile of the output distribution matches E_{true}:
      - Quantiles trained: 50% (i.e. the median => this is our alternative $\hat{E}$)
      - 5% and 95% quantiles: useful because for each cluster we have an estimate of energy uncertainty a la Minos
- Selection:
  - $I_{SC} > 10^3, I_{\text{rms}} > 8$: suppress the fake clusters
  - $\sigma_T \gtrsim 300\,\mu m$: suppress the interactions in the CMOS
  - $R < 900$ pix: suppress the bad S/N regions (in any case, the source illuminates only the central strip)

Inclusive results

- For x<700 and x>1700 not many interactions to train (this is also a limit of applicability), while in y we have many events



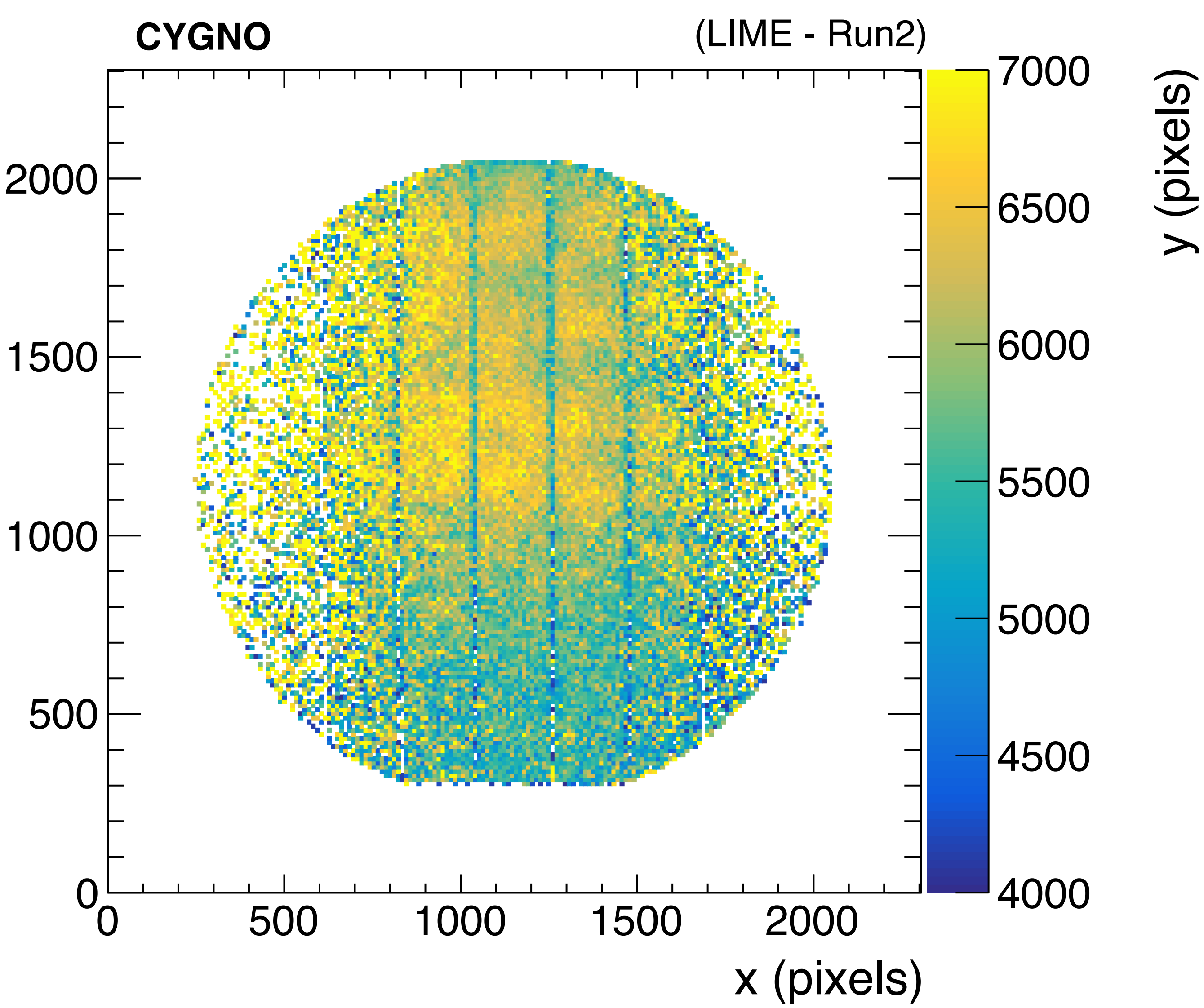
Number of clusters / image passing the selection



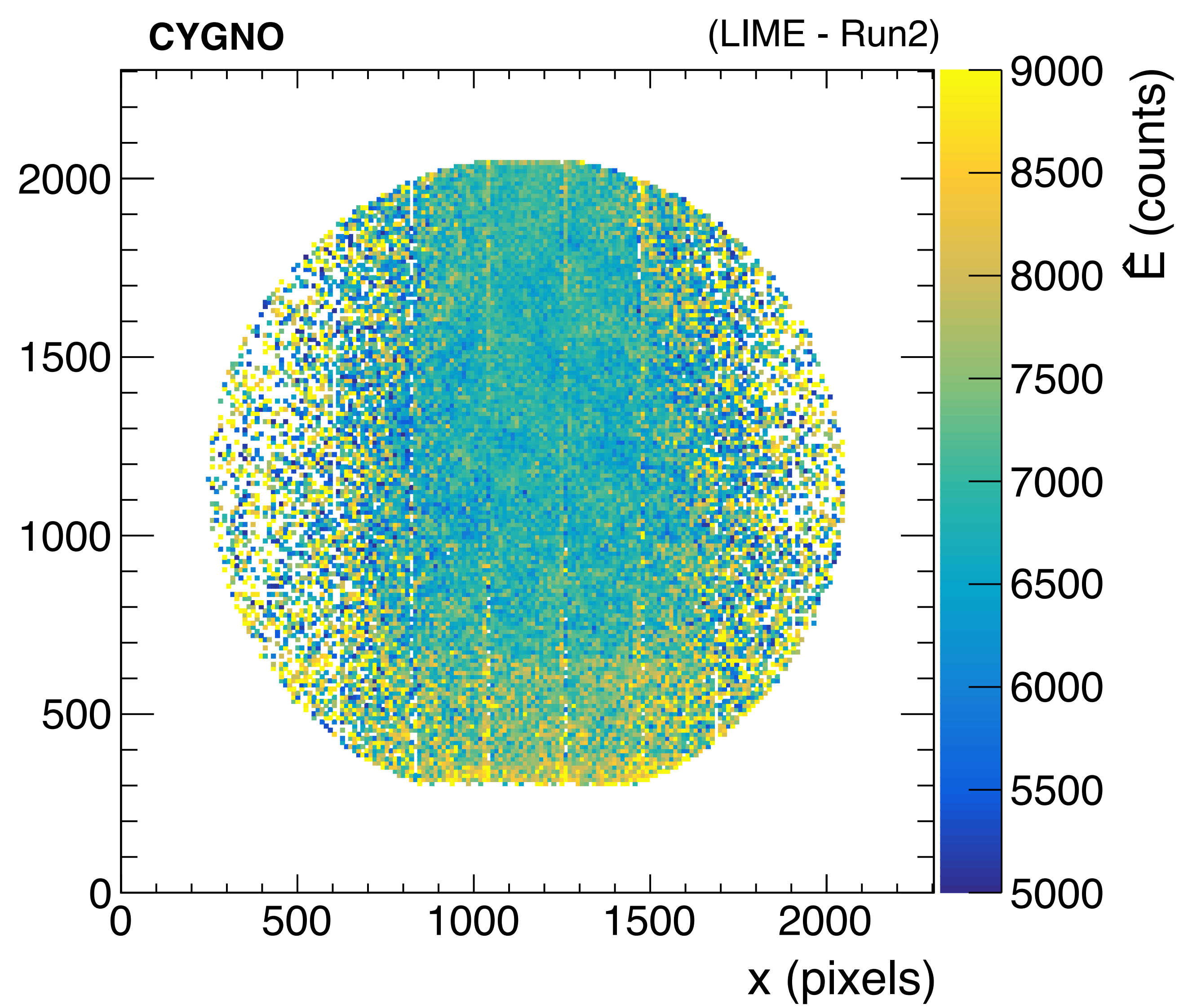
All clusters, at any E_{true} (i.e. HV_{GEM1}) and any z_{true}

Correction vs x, y

Raw I_{SC}



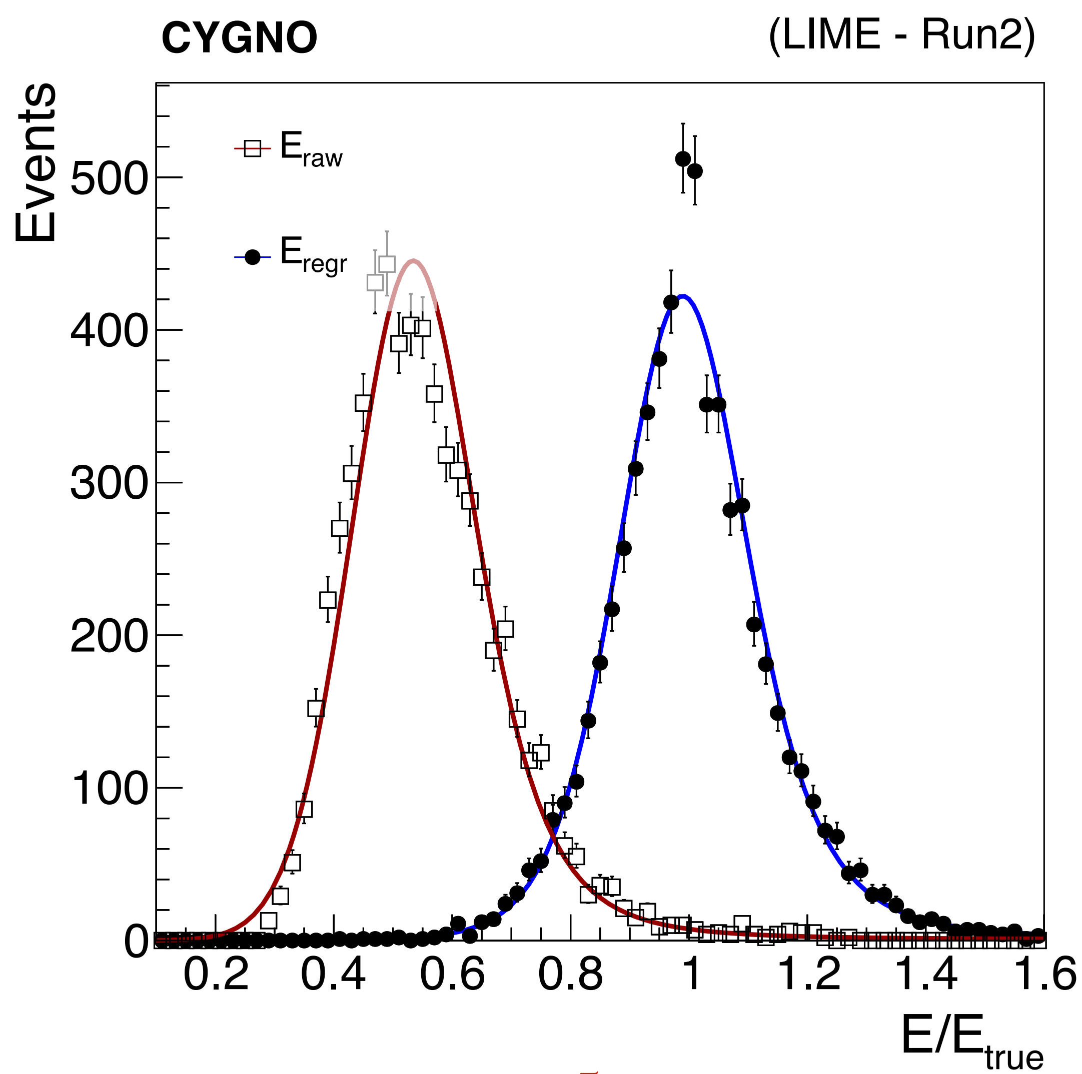
Median regression $\hat{E}_{\text{median}}$



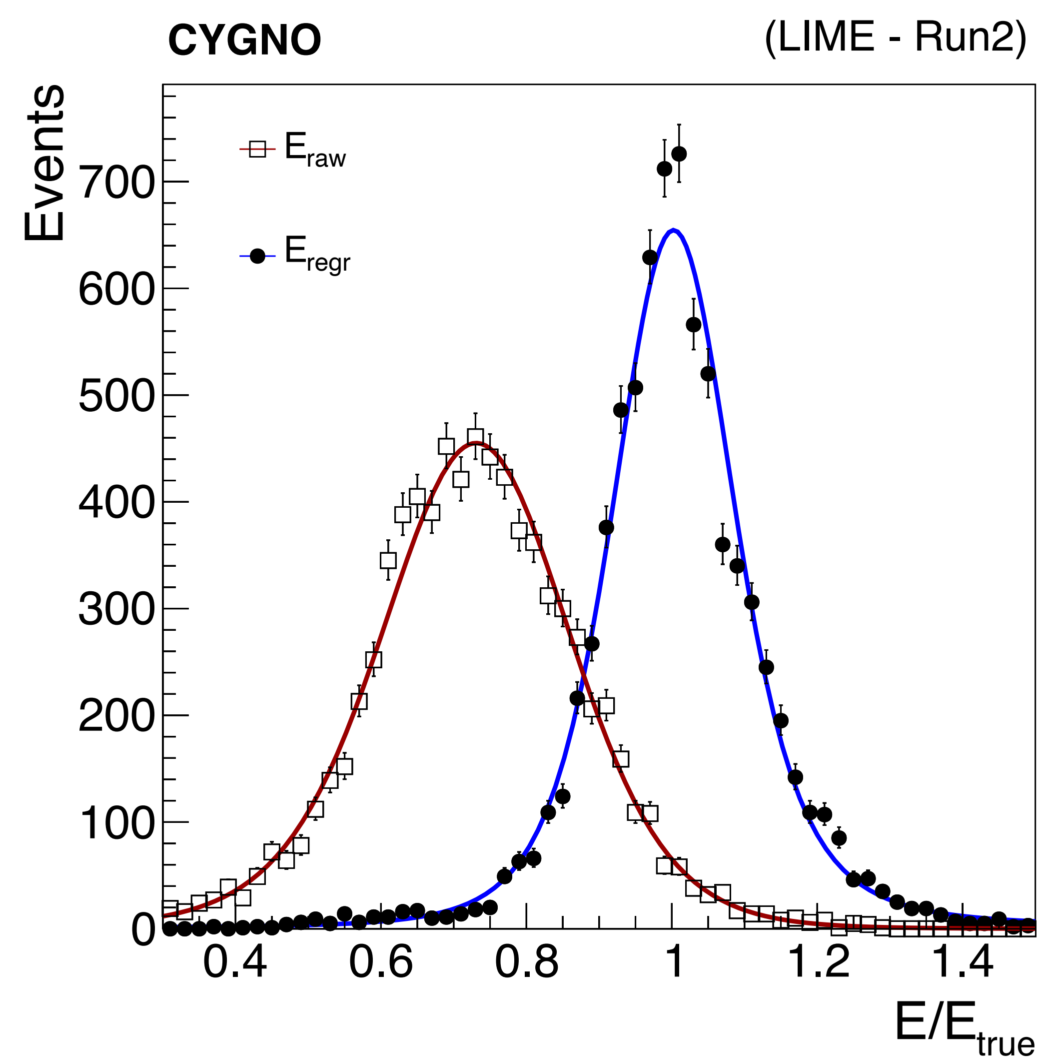
- Z-scale in the plots rescaled by the mean of the $\hat{E}$ distribution for a fair comparison
- Regression flattens the energy response in x-y, very visible close to the GEM sector boundaries
  - Some step for y<600 to be understood
- $\hat{E}_{\text{mean}}$ similar, but a bit worse around the boundaries

$\hat{E}$ vs z_{true}: distributions

- Fit $I_{SC} \equiv E_{raw}$ and $\hat{E} \equiv E_{regr}$ with a Cruijff function at different z_{true} to estimate response and energy resolution
  - The corrected energy $\hat{E}$ is more symmetric, at any z_{true}, as expected
  - Fits to be improved, but a starting point
  - Normalised to E_{true}, i.e. the peak value at 48 cm (least saturated)



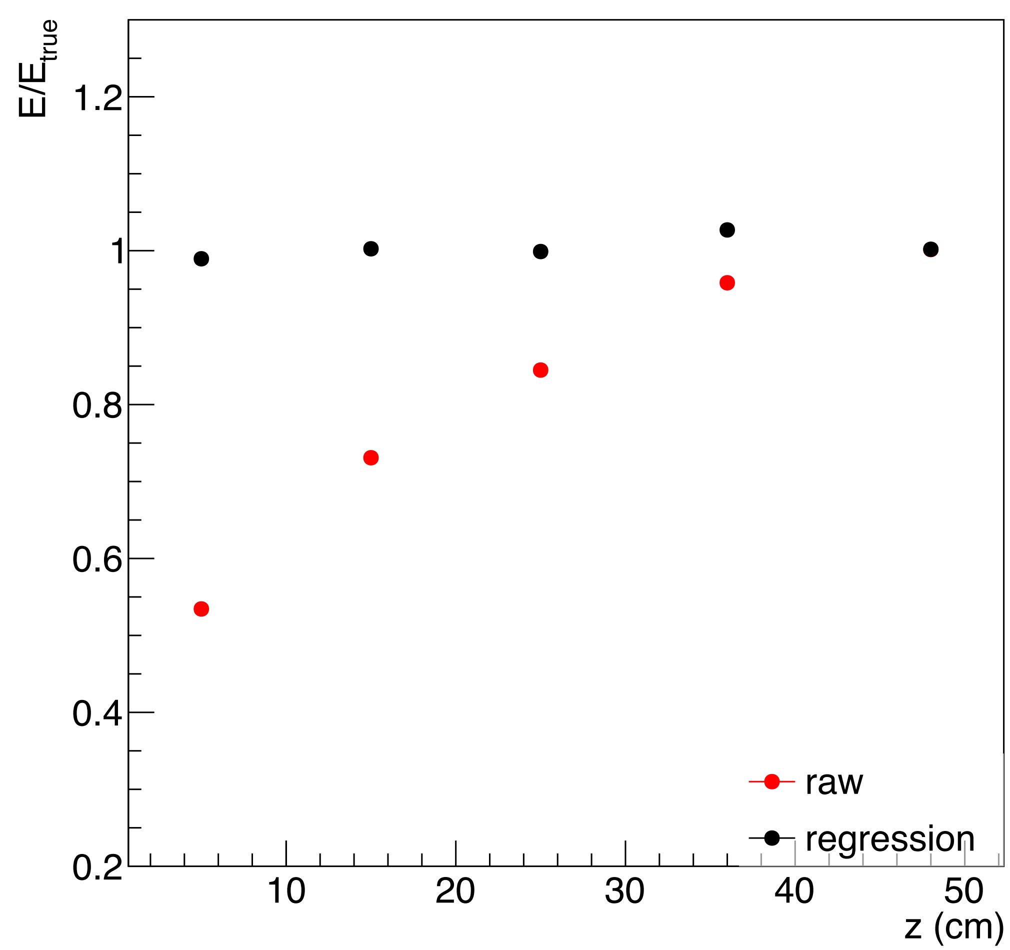
$z_{true} = 5$ cm

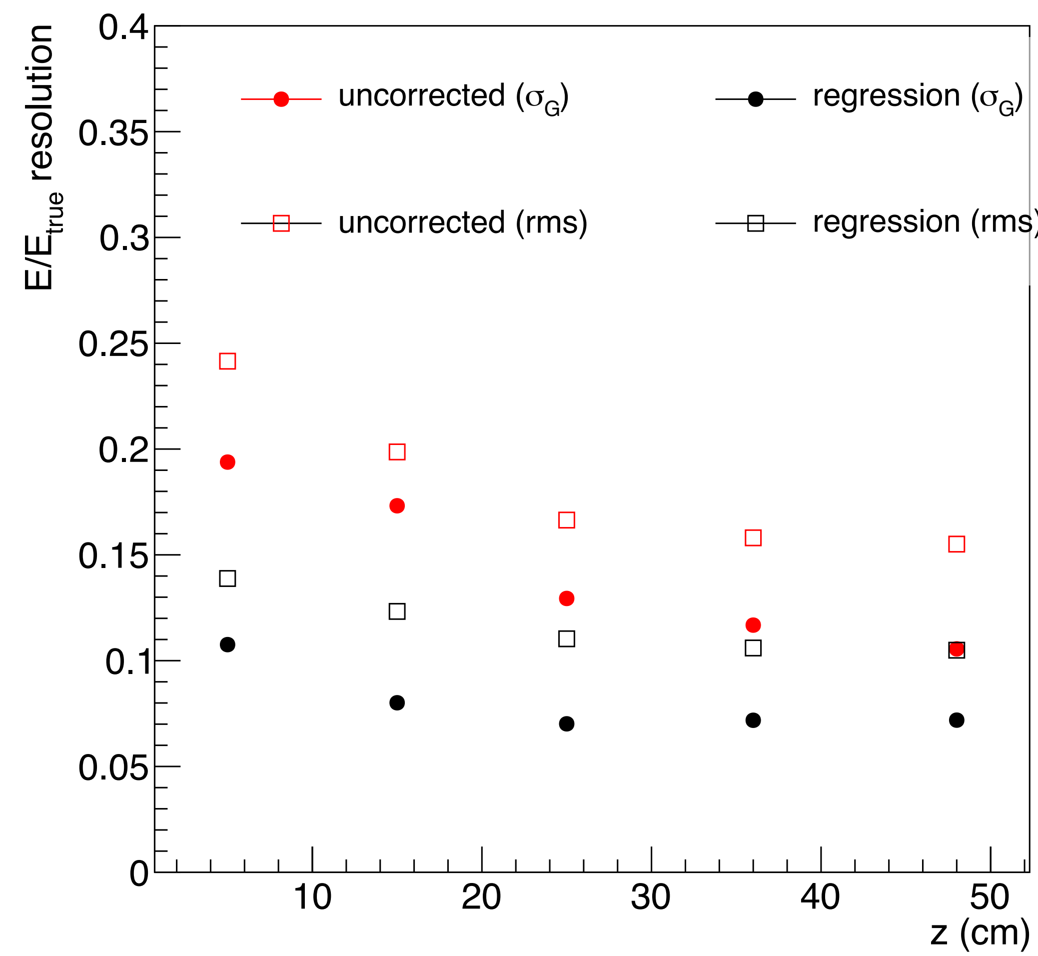


$z_{true} = 15$ cm

$\hat{E}$ vs z_{true}: response and resolution

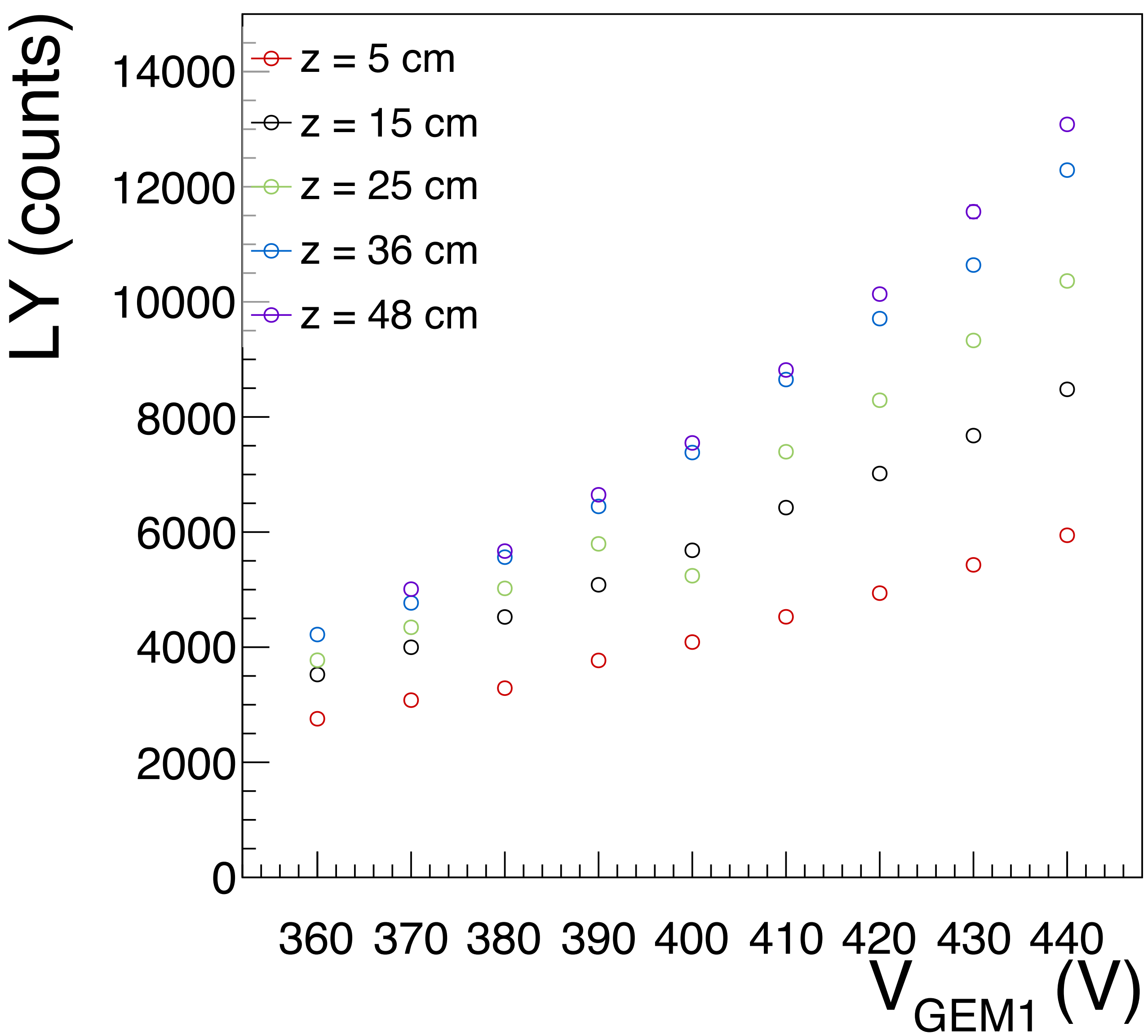
- Raw LY varies by a factor 2 for z in [5,48] cm, as known
- Corrected $\hat{E}$ (here median, but similar for mean) almost flat
- Energy resolution improved at any z
  - Estimate **11% improvement (in quadrature) at z=48 cm**, i.e. the contribution from the non-z dependence
  - 19% improvement at z=5 cm, so naively 15% **contribution from the z-correction**

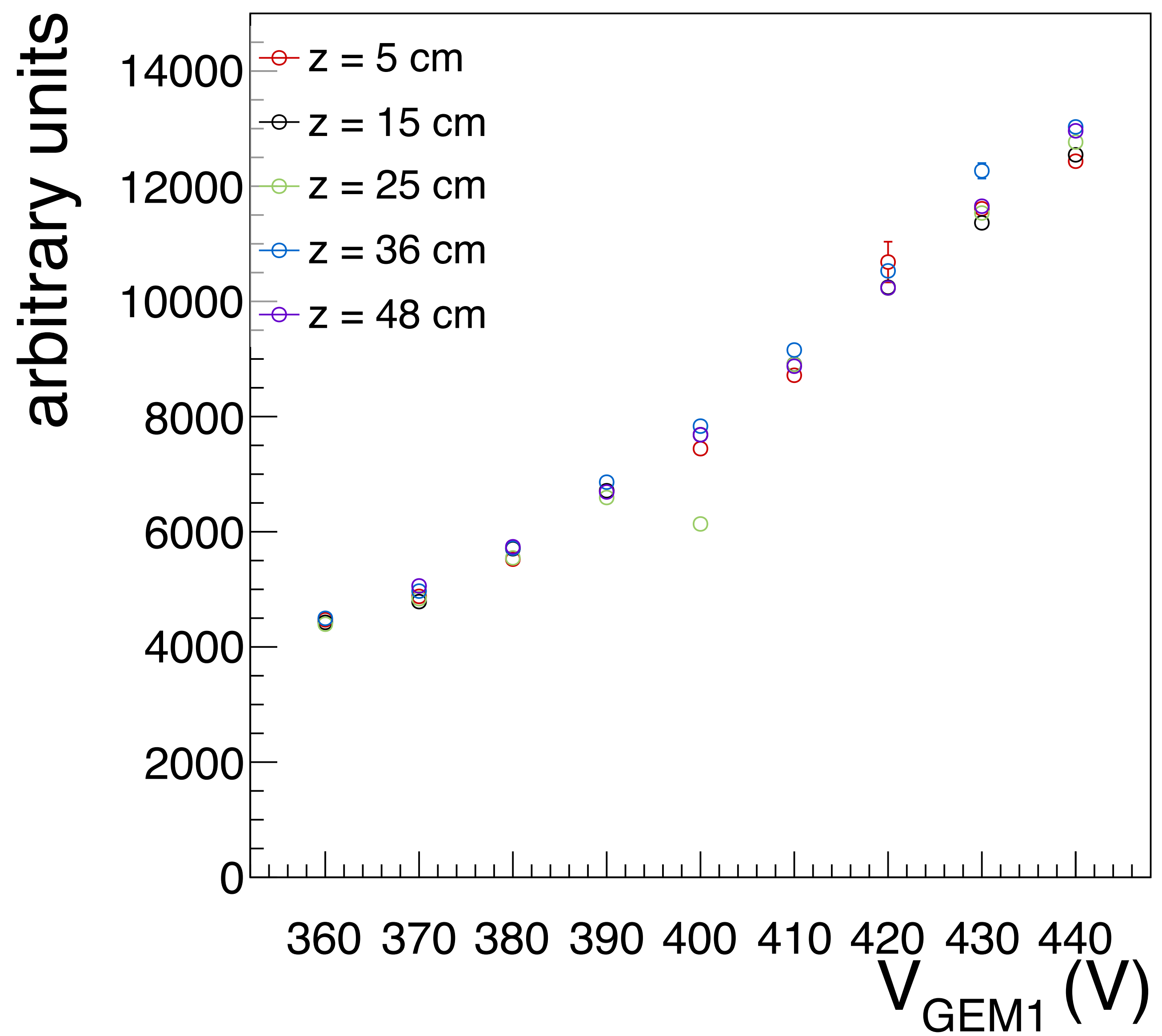




Closure test: response

- Using the ~half of the 2D scan dataset not used for training the regressions
  - Strange jump at $HV_{\mathrm{GEM1}} = 400V$ and $z = 25$ cm to be checked (even before regression)

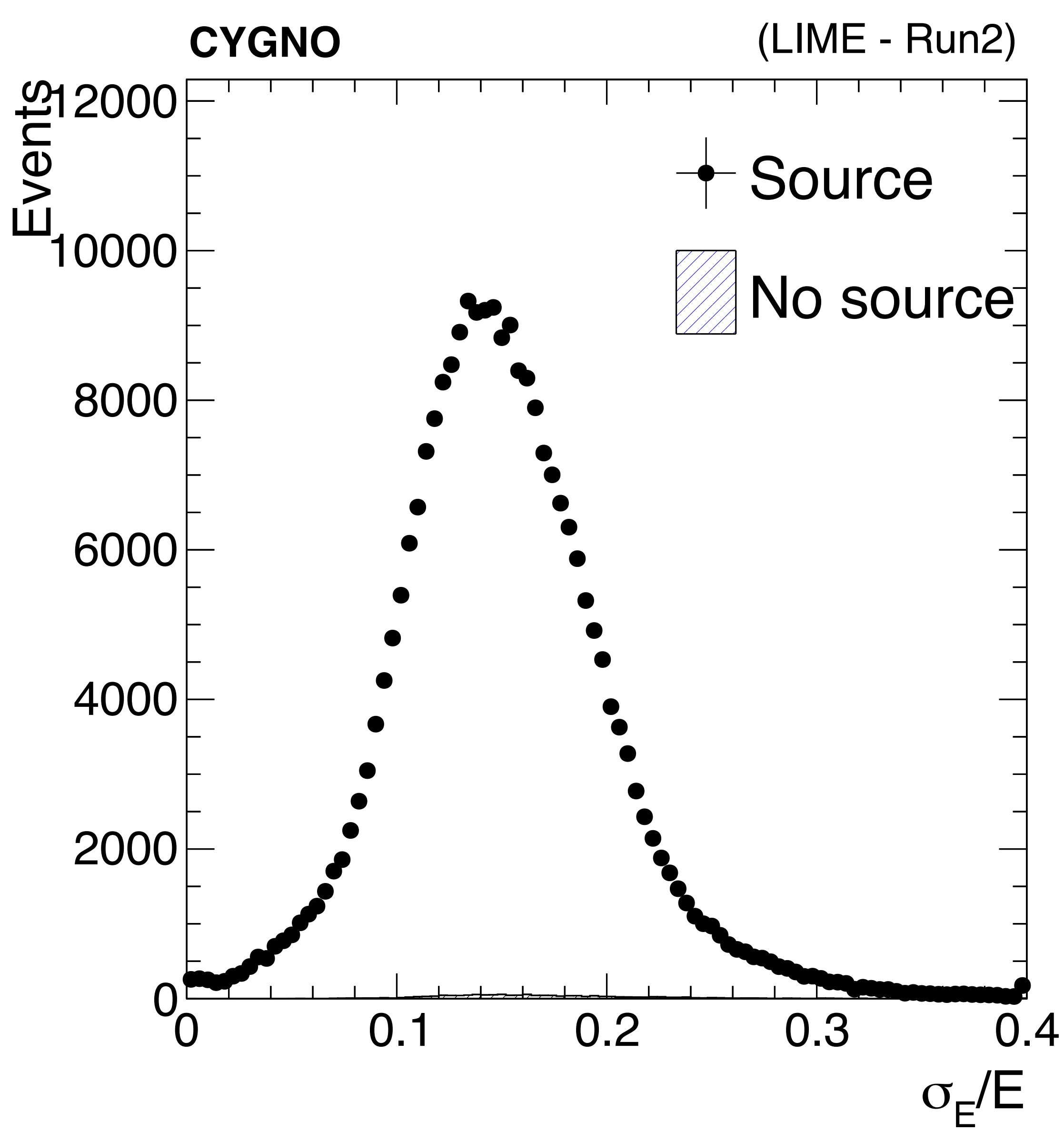




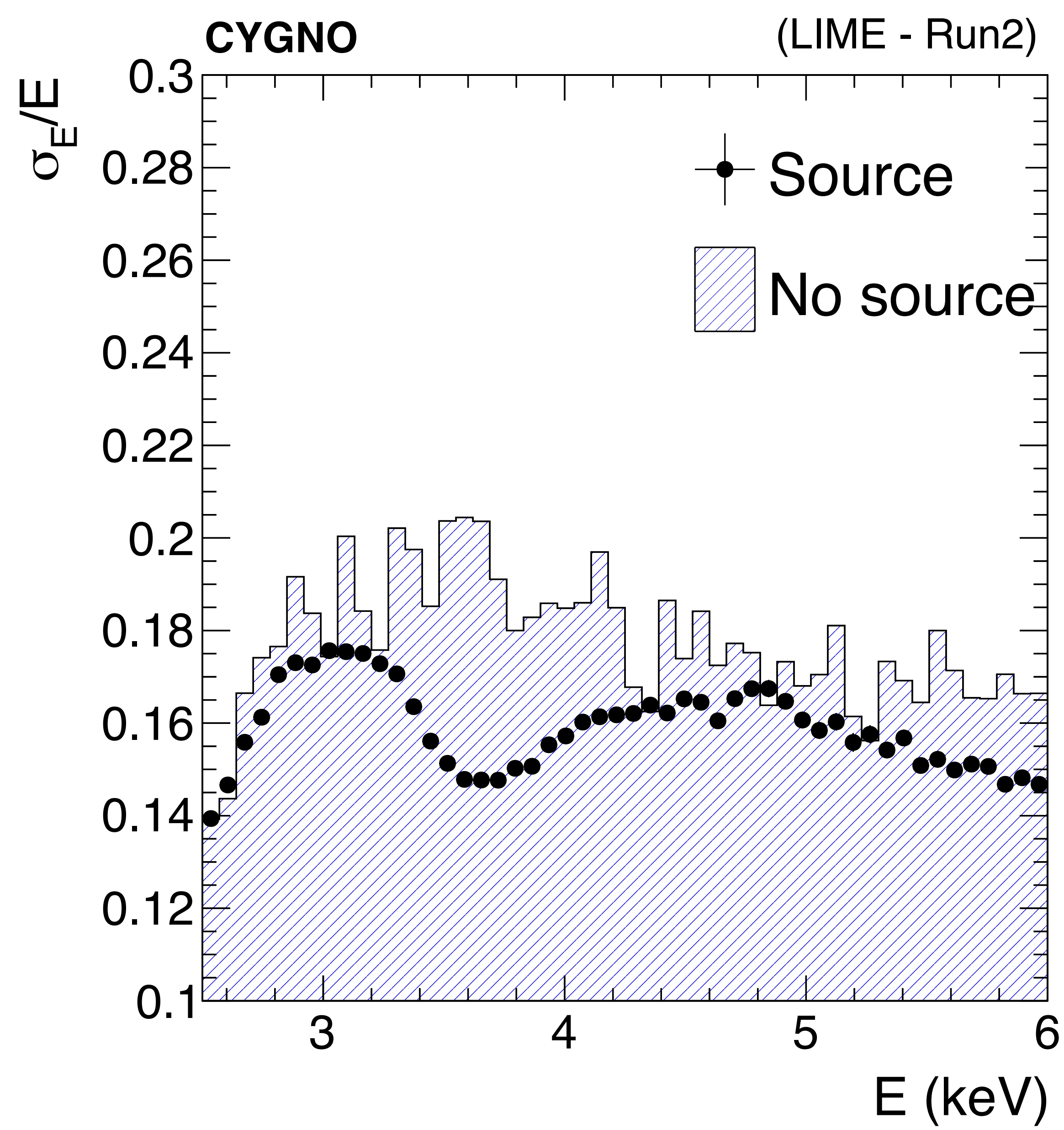
The correction of saturation holds at any (mocked up) E_{true}

Cluster resolution estimate

- From the quantile regression we have the per-cluster energy resolution estimate
  - Could be used to make categories of best-measured clusters, or just to exclude worst-measured ones



Inclusive, at all energies / z

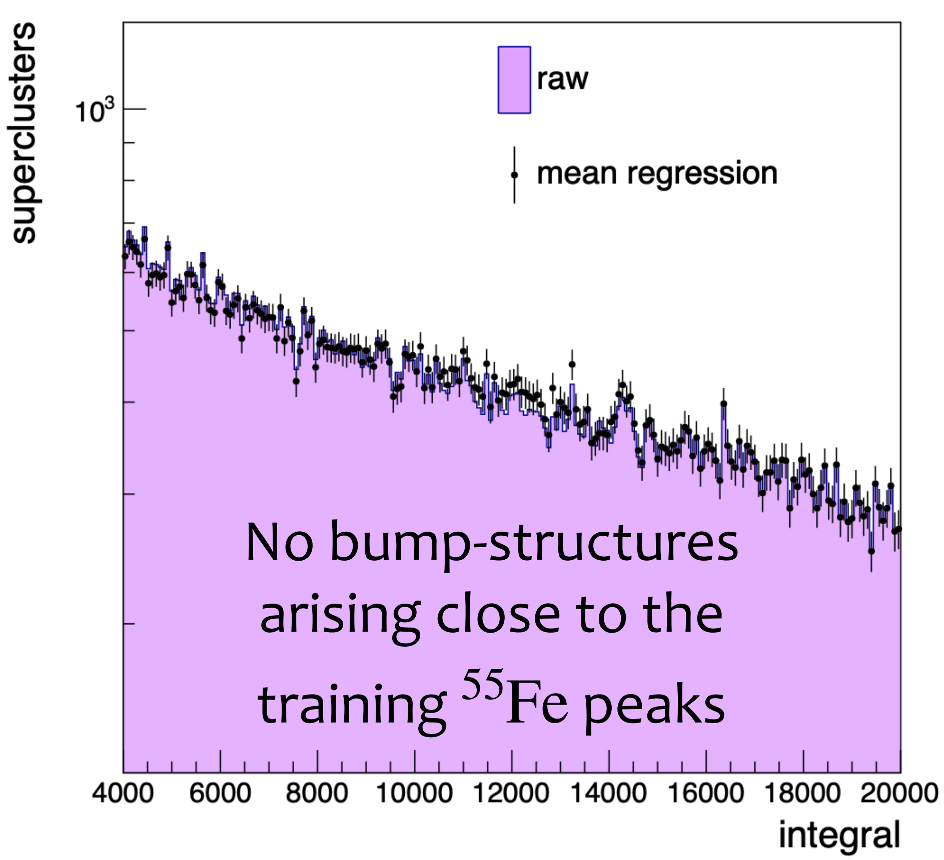


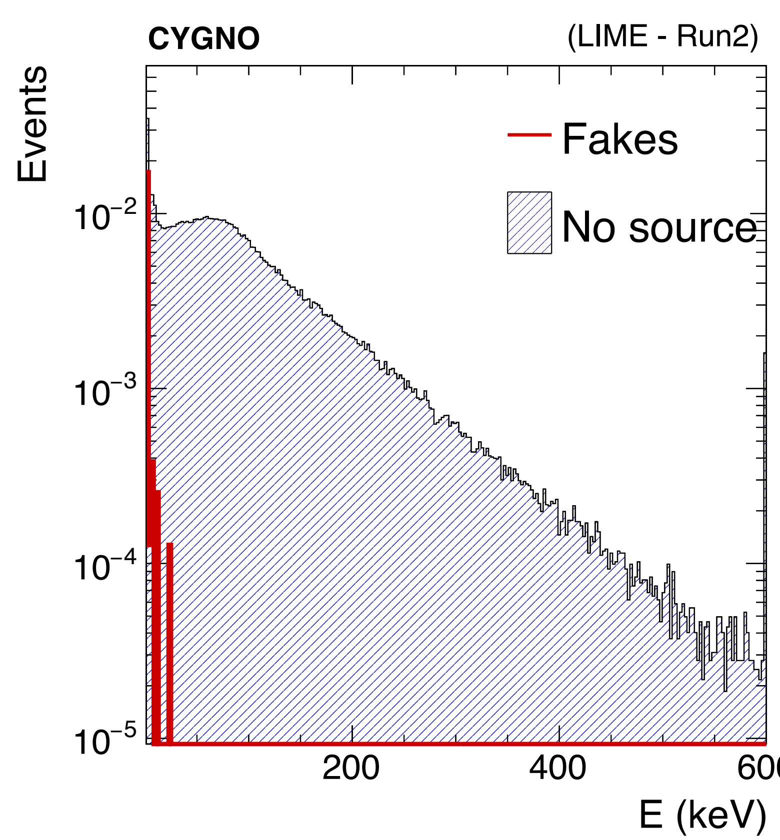
There are 11 steps in E_{true}, but resolution mixes

Need to disentangle different z's

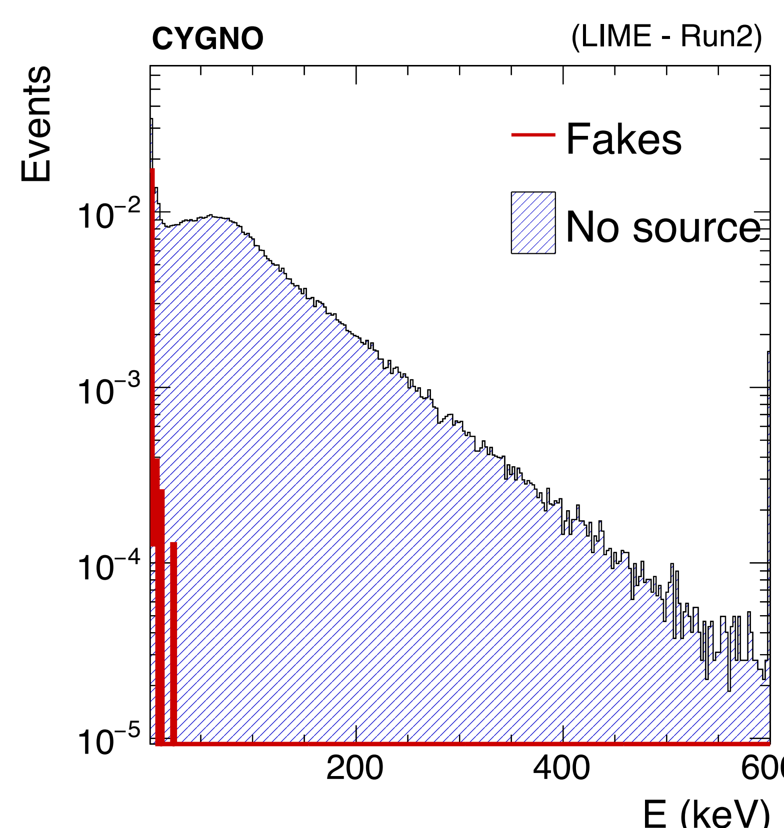
Validation: $\hat{E}$ in no-source data

- Computation of the 4 types of regression energy $\hat{E}_{\mathrm{mean}}, \hat{E}_{50\%}, \hat{E}_{5\%}, \hat{E}_{95\%}$ very fast.
  - Computed it for all the Run-2 Runs ("friend" ROOT trees, that can be attached to the RECO ones copied to cloud). Details in the wiki page here.
  - Will use $\hat{E}_{50\%}$ as example of regression energy estimate
  - **N.B. since the model is not linear, it is safer not to extrapolate (i.e. compute) the output outside the phase space of the training**
    - ☞ for any cluster not passing the cuts used to define the training dataset $\hat{E} \equiv I_{SC}$





E from I_{SC}



E from $\hat{E}_{50\%}$

Where to find the data with E_{regr}

- Refer to the GitHub reconstruction wiki page for Central Trees productions here
  - Main ROOT trees ⇒ look at the "*Winter23*" entry in the table "Central RECO productions"
  - Friend ROOT trees ⇒ look at the "*Winter23-pp1*" entry in the table "Central Post-Processing productions"

- The content of the ROOT friend trees are 4 additional variables / cluster:
  - *sc_regr_integral* ⇒ the corrected energy with the "mean regression"
  - *sc_qregr_integral* ⇒ the corrected energy with the "median regression" (50% quantile)
  - *sc_qregr_up_integral* ⇒ the corrected energy (50% quantile) + 1σ
  - *sc_qregr_dn_integral* ⇒ the corrected energy (50% quantile) - 1σ

- There is one friend tree / main tree. They can be joined together as explained here

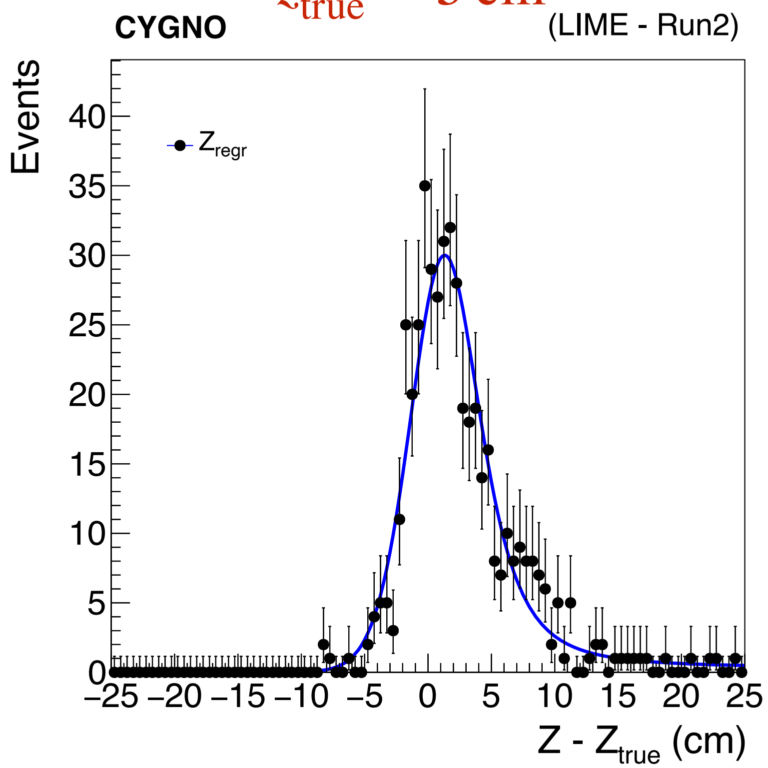
- What if I need to compute these variables on new runs?
  - For the Run-2 data training, the tensorflow files are committed, and can be run with the "postprocessing" tool as in the example here in GitHub

Z regression

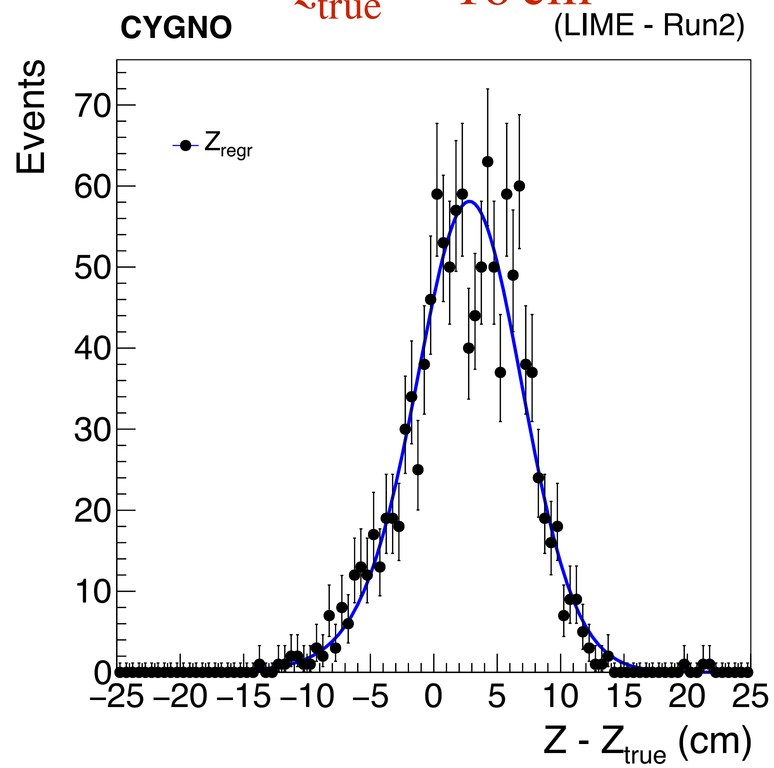
- As a validation of the energy regression, train a regression with the same model, same variables (apart I_{SC}: $\vec{\theta}' = \vec{\theta} - I_{SC}$)
    - Since regression seems to be able to correct the saturation, it must predict z as well
    - Not a surprise, see R. Roque's presentation, or the LEMON BTF paper
- Data used: the same dataset of the 2D scans used for energy regression, with the same selection
- Target: z_{true}
    - The z of the source is known with $\pm 0.5\,\text{cm}$ uncertainty (conservative)
    - In addition, the collimation of the source adds another $\Delta_z^{\text{collim.}} \approx 8\,\text{mm}$ to the z_{true} of the interaction
    - ☞ for "internal" z positions, smear the true value by a Gaussian with $\sigma_z = 1\,\text{cm}$
    - To avoid border effects, for $z = 5, 48\,\text{cm}$ make a domain continuation, at least in the [0-5] cm and [48-50] cm
        - Spread the first point as uniform distribution in [0-5.5]cm, and same for 48 cm

Output of $\hat{z}$-regression: bias

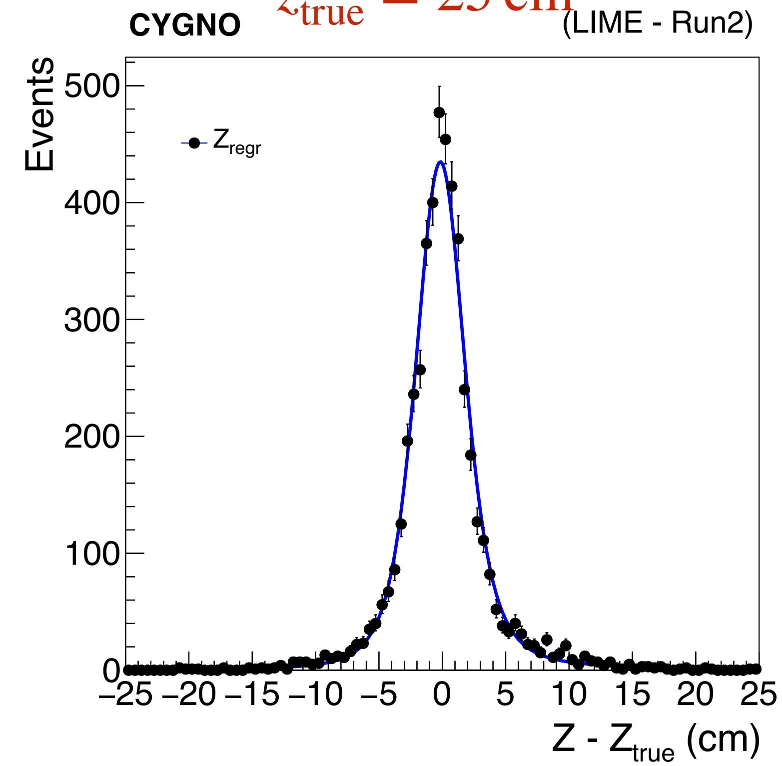
$z_{\text{true}} = 5\,\text{cm}$



$z_{\text{true}} = 16\,\text{cm}$

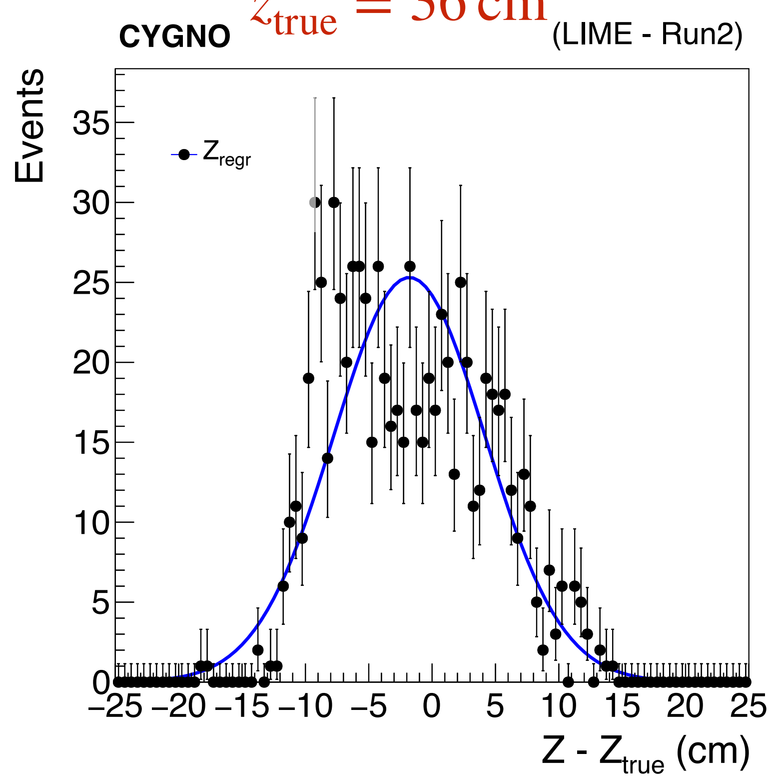


$z_{\text{true}} = 25\,\text{cm}$

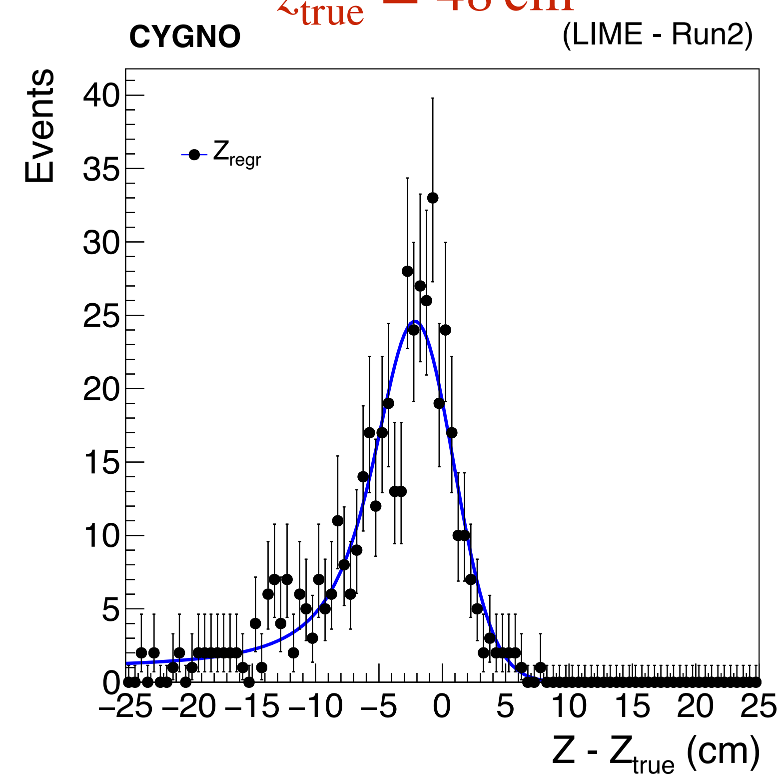


- Output at center: **no bias,** $\sigma_z \approx 2\,\text{cm}$
- Output at extrema: **small bias (1-2 cm)**, understandable because cannot predict out of detector, $\sigma_z \approx 3\,\text{cm}$
- 3-4 cm bias in the intermediate positions, to be understood

$z_{\text{true}} = 36\,\text{cm}$

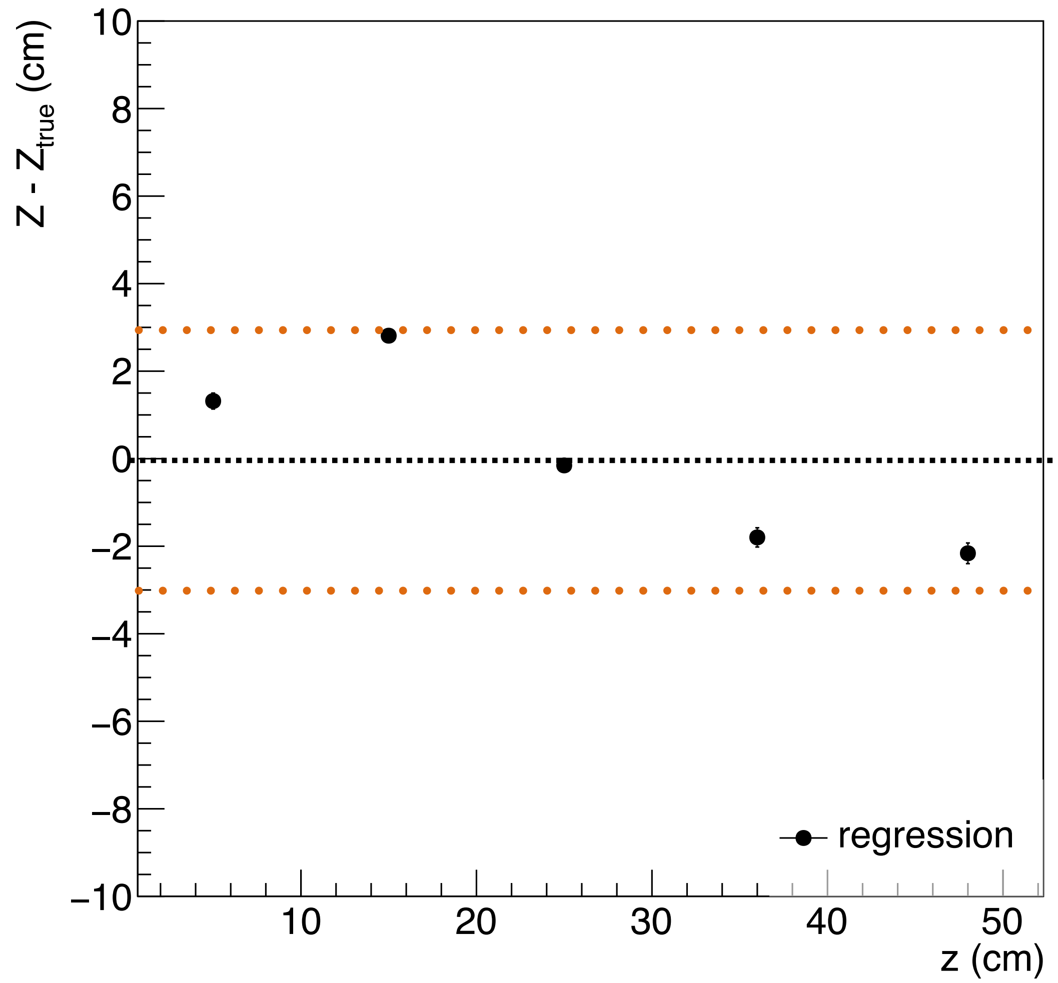


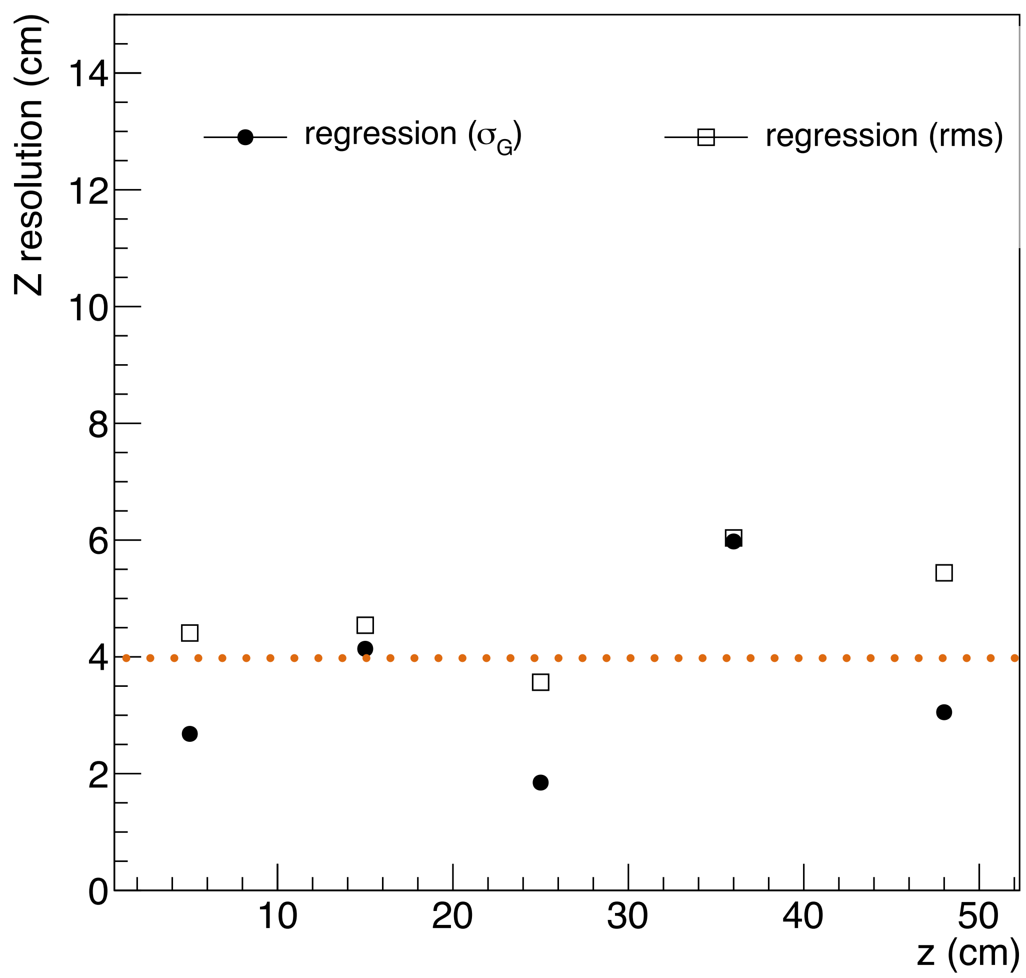
$z_{\text{true}} = 48\,\text{cm}$



Output of $\hat{z}$-regression

- In any case, bias within $\Delta z = \pm 3\,\mathrm{cm}$
- Resolution $\sigma_z \approx 4\,\mathrm{cm}$





- Where to find data with z_{regr}? In the same friend trees with E_{regr}.
- But I think that this estimate it is overtrained and biased towards the z=25 cm point.
    - ⇒ **Use with care!**

Conclusions

- We are using the ^{55}Fe source as regular calibration & monitoring candle for LIME at LNGS during Run-2 and beyond
- We used it also to derive energy cluster corrections & to derive a z-estimate, but:
  - the application is limited to tracks which are X-ray spot-like, i.e. not applicable to longer tracks or maybe to tracks which have a dE/dx very different from an ER
  - Without a variable X-ray source and AmBe source a way out is a reliable MC simulation
    1. Tune MC simulation to reproduce the main input variables (cluster shapes vs E, saturation)
    2. Train on a MC with flat x-y-E distribution
    3. Derive a residual data/MC correction of the absolute scale (possibly with a 2^{nd} adversarial NN regression)
  - this would solve an issue: right now the regression corrects for some effects that are in the SIM (e.g. saturation) and some that are not (non-uniformities)
- When a correction is stable enough, we can think of injecting it in the reconstruction itself, but I think that the post-processing is more flexible (much faster and allows to redo multiple trainings with just 1 full-reco)
  - Can be put in the automation as a 2^{nd} step when one run is reconstructed

The End