LIME Run-2 energy and z MVA regressions

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

Reconstruction & analysis meeting, 30 March 2023

Recap of the method

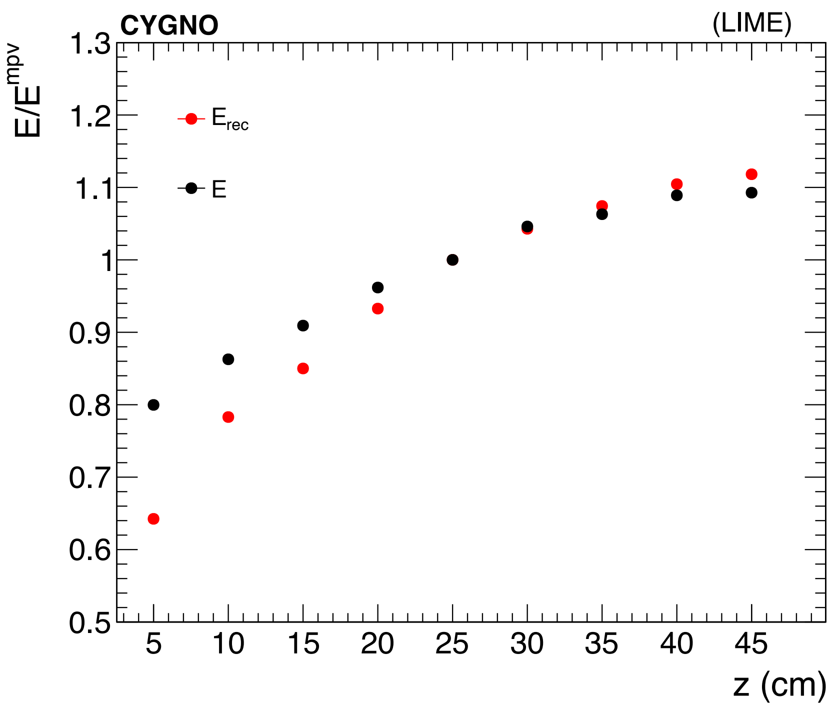
- General principle is to derive a best estimate of the dependent variable (in our case the ***true cluster energy***, or the ***Z position*** of the interaction) given a set of **measured variables** (measured light, position in XY, cluster shape parameters, etc)
  - One objective is to correct the 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

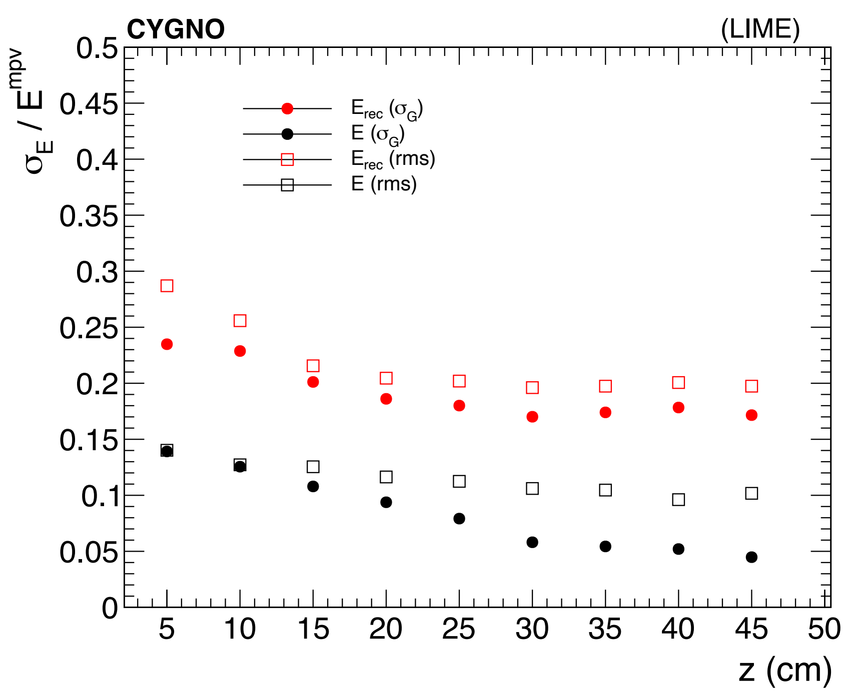
Optimisation with LNF data

- Since we want to predict E_{true} , we need a training sample which contains the correlation of E_{true} with $\vec{\theta}$, i.e. we need a sample with different values of E_{true}, with E_{true} known
  - With the MC simulation, easy: one can do a sample flat in E_{true} within the range of interest
  - With data, we have one clean candle, ^{55}Fe, but it is only 1 fixed energy ($5.9\,\text{keV}$)
    - More points added with the multi source: K_α and K_β lines for many sources
    - This is something we don't have (and maybe NEVER have) in LIME / CYGNO
- BIG limitation of the LNF data:
  - While with ^{55}Fe we had a "flat z_{true}" sample, with the z scan, multi source was at fixed z (25 cm)
  - i.e. with this sample no hope to correct for saturation effect for a wide range of E
- Still, largely improved the resolution because it flattens E response vs x-y

LNF

$E/E_{z=25\text{cm}}$





LNF

σ_E/E

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

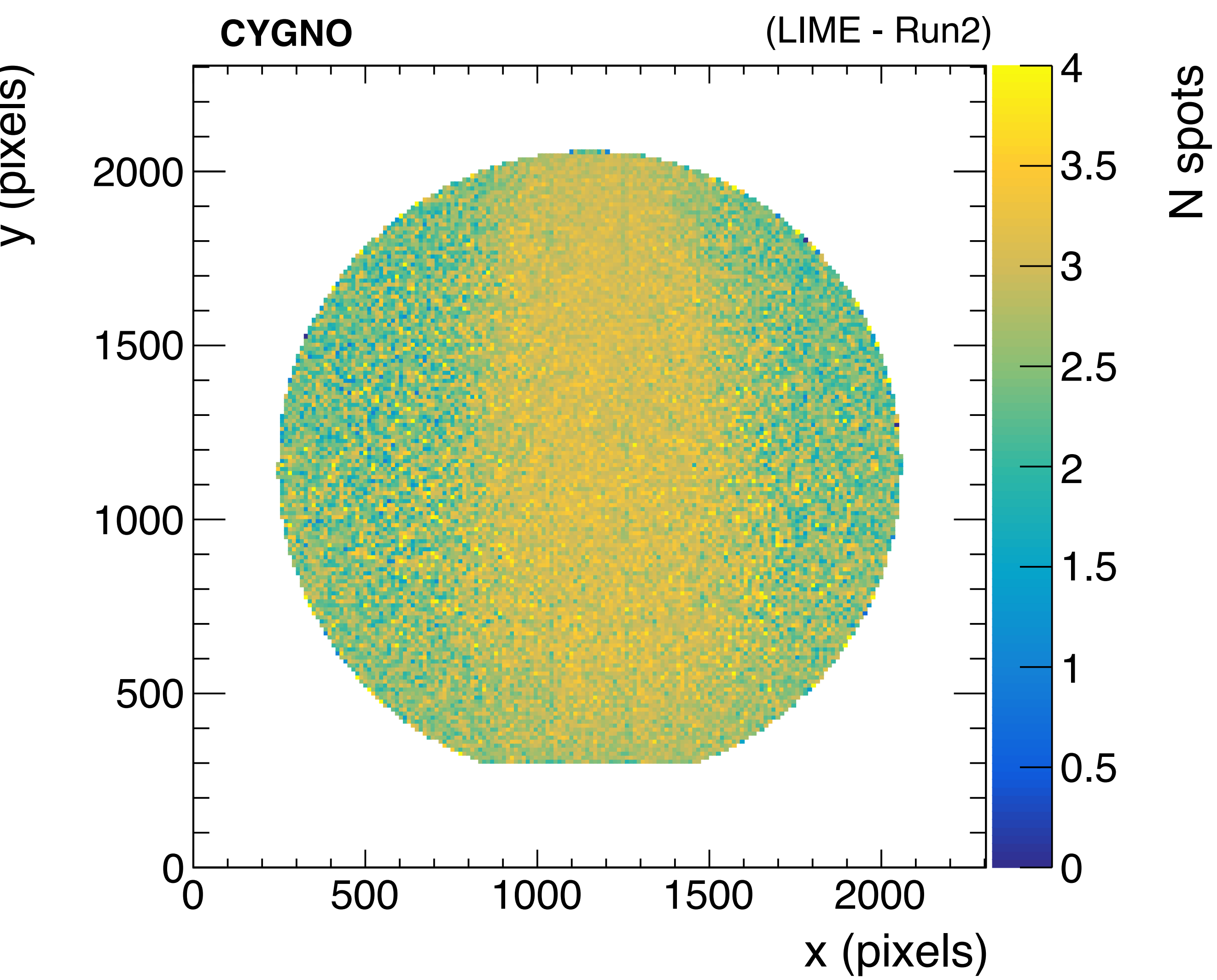
- Used the 2D [$E_{\rm true}$, $z_{\rm true}$] scan with ^{55}Fe source taken Feb 22nd. Each point has 400 events

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

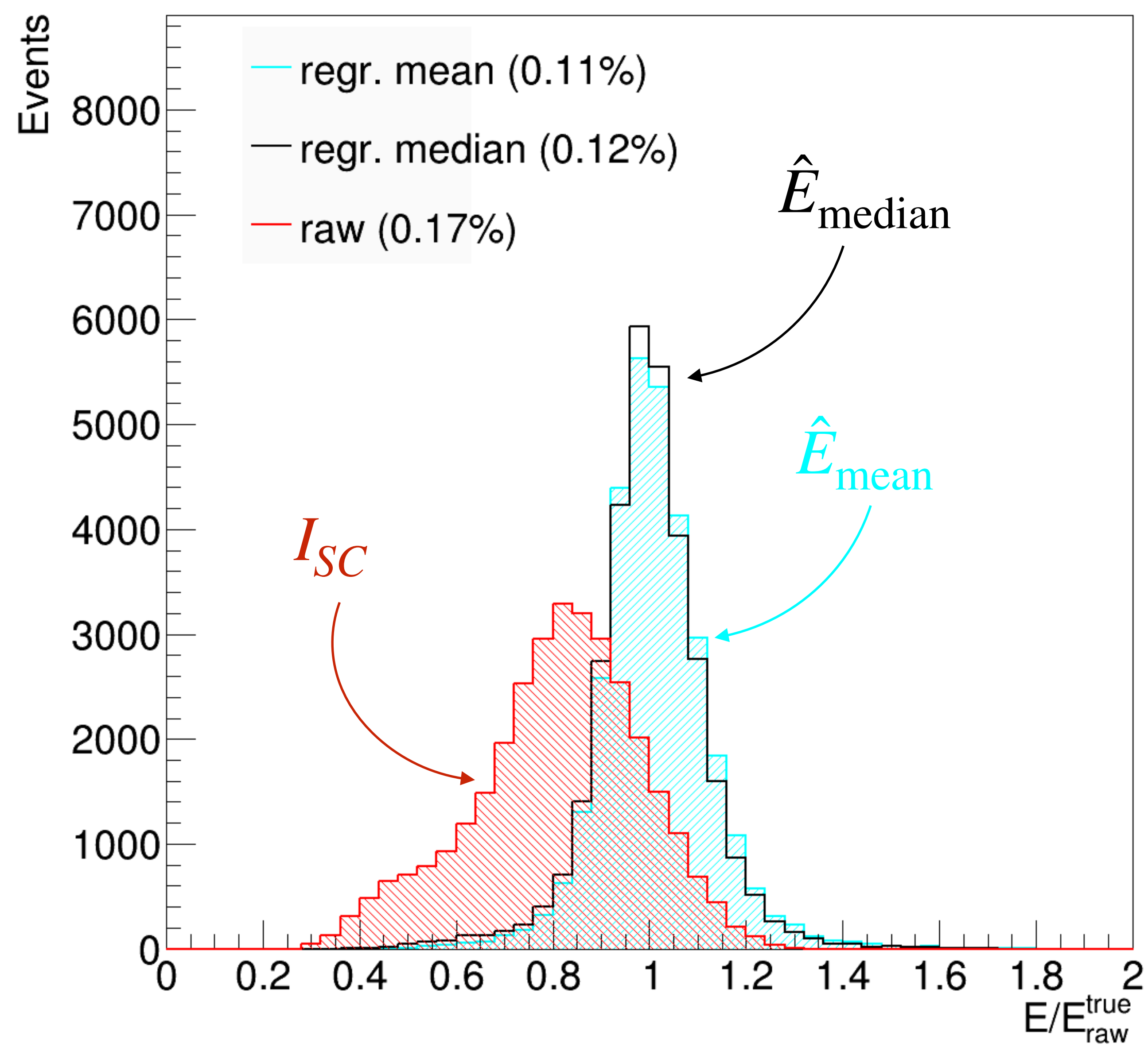
- Set of variables used for energy regression:
  - $\vec{\theta} = [I_{SC}, \delta, I_{\rm rms}, x, y, \sigma_T, \text{width}]$
  - Model: Gradient Boost Regression (GBR) with a Boost Decision Trees algorithm
  - Model parameters: max_depth=3, min_samples_split=6, min_samples_leaf=7, learning_rate=0.1, 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 $E_{\rm true}$ (this is our $\hat{E}$)
    - **Quantile regressions**: a given quantile of the output distribution matches $E_{\rm true}$:
      - 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:
  - $I_{SC} > 10^3$, $I_{\rm rms} > 8$: 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



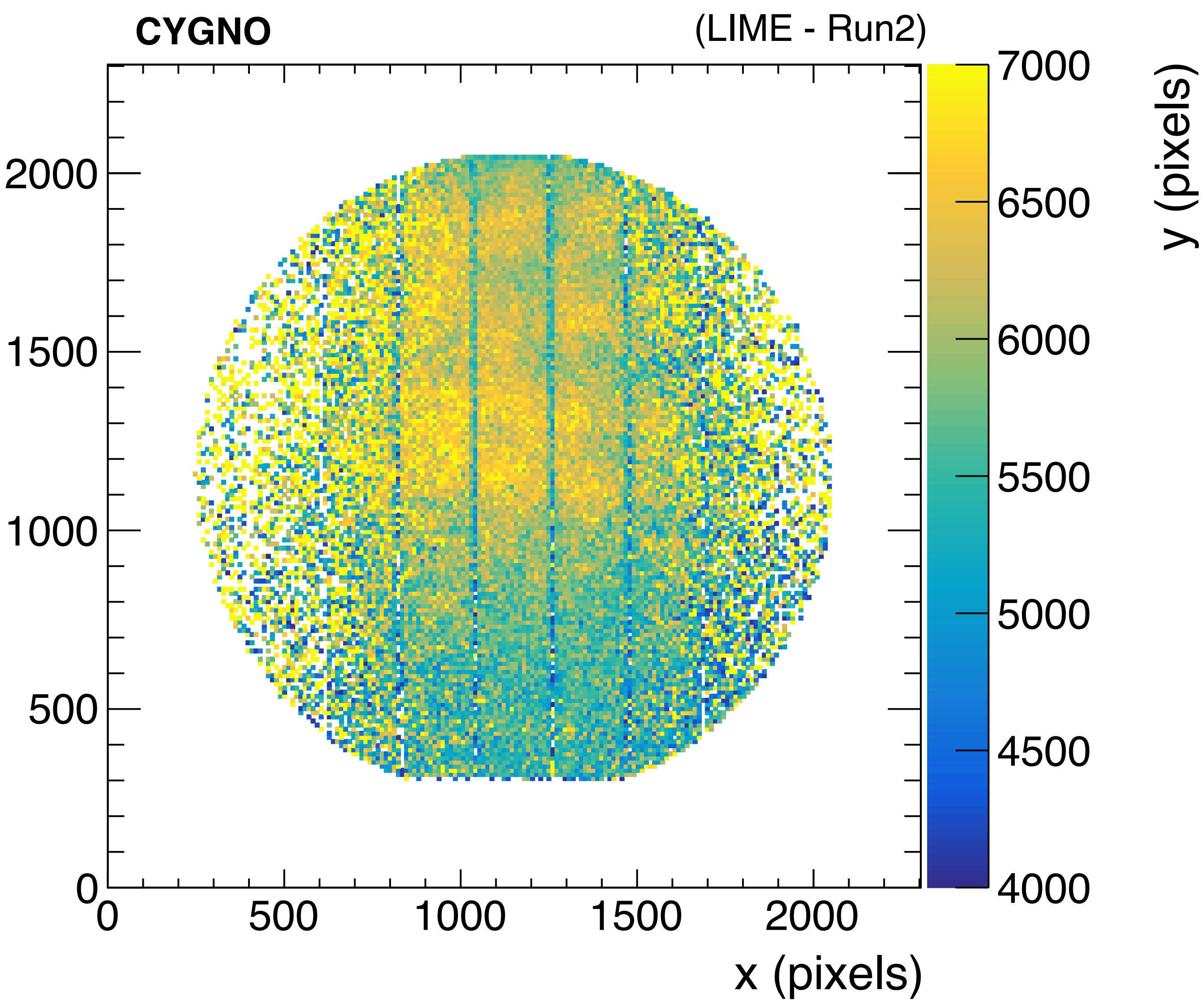
Number of clusters / image passing the selection



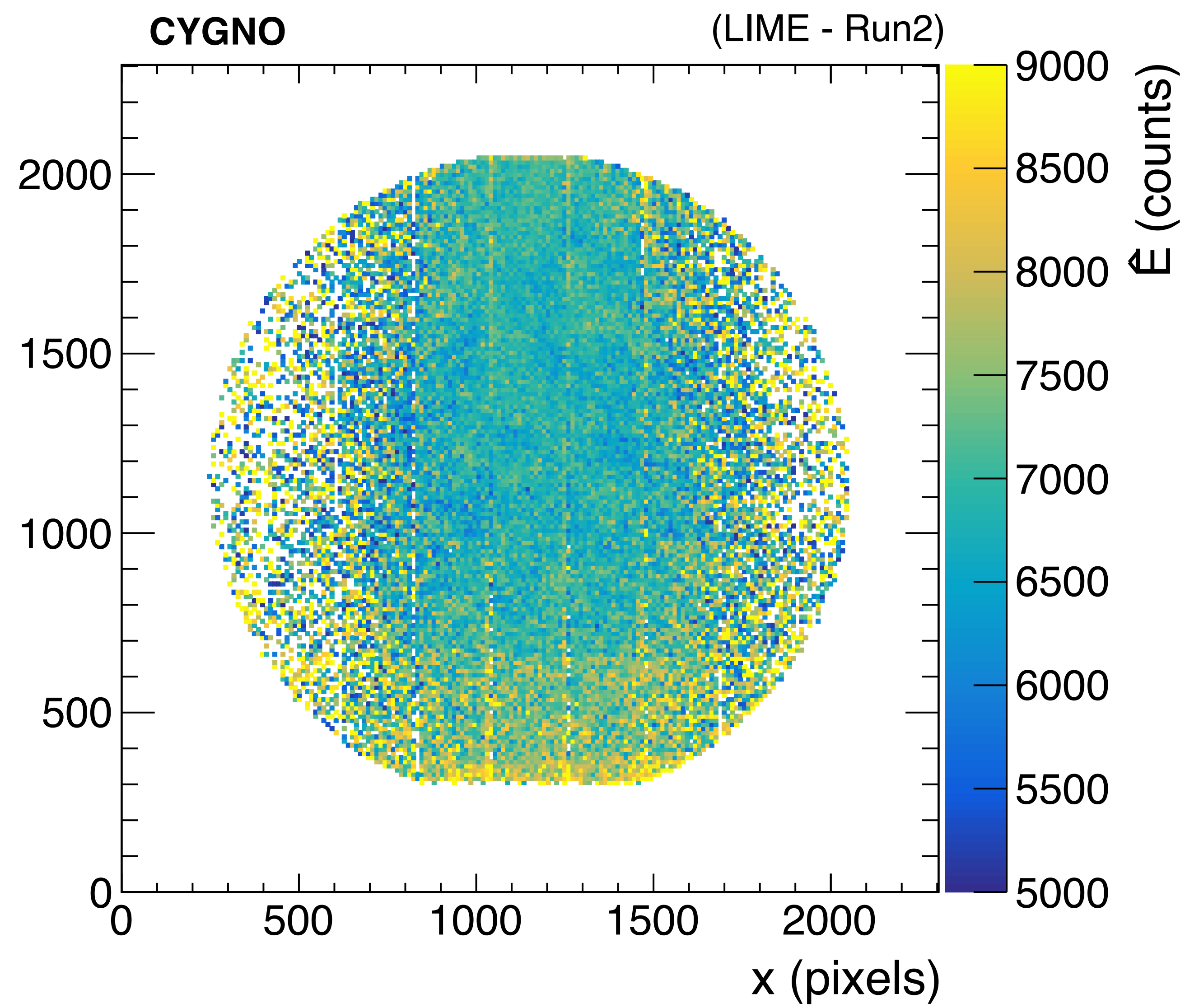
All clusters, at any E_{true} (i.e. HV_{GEM1}) and any z_{true}

Correction vs x,y

Raw I_{SC}



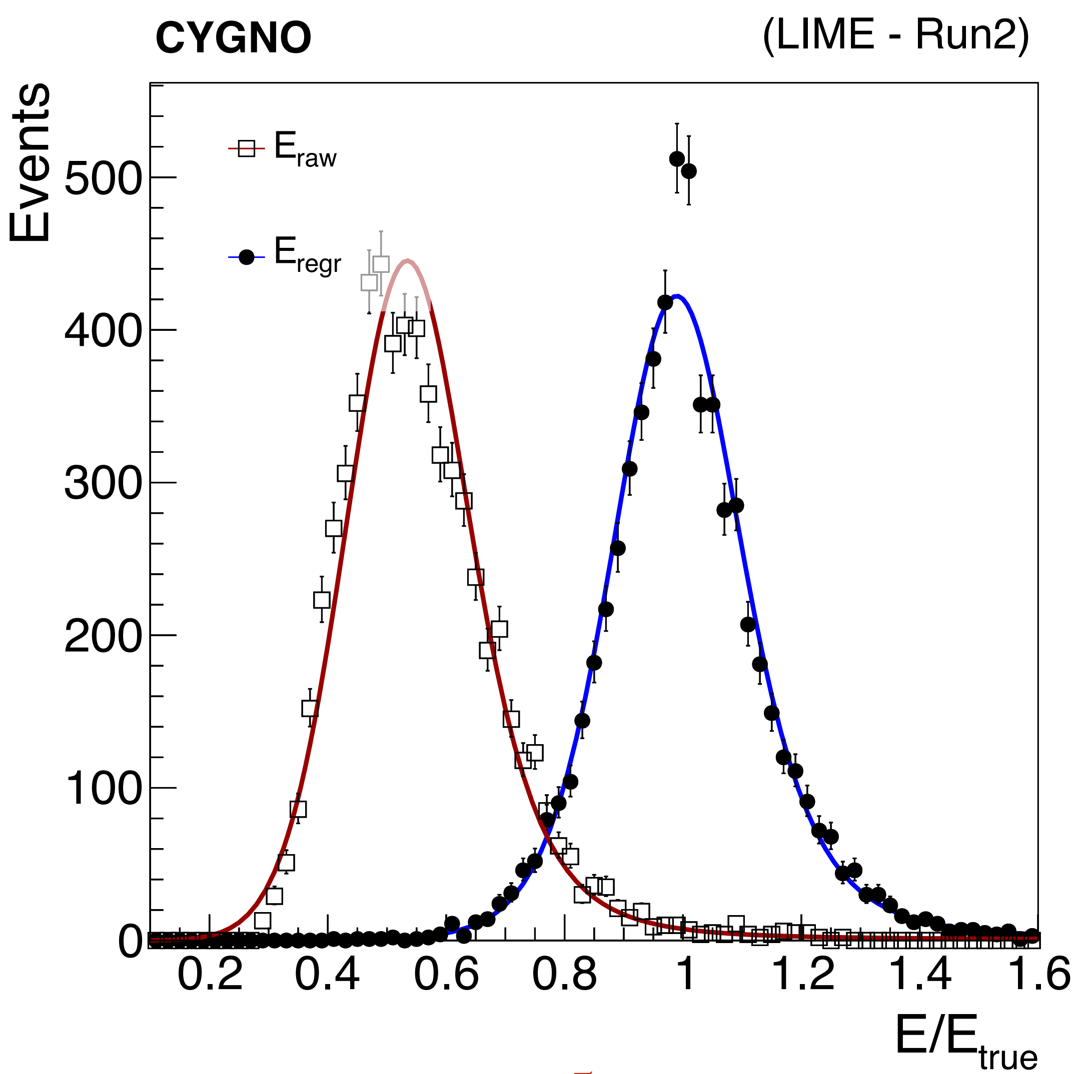
Median regression $\hat{E}_{\mathrm{median}}$



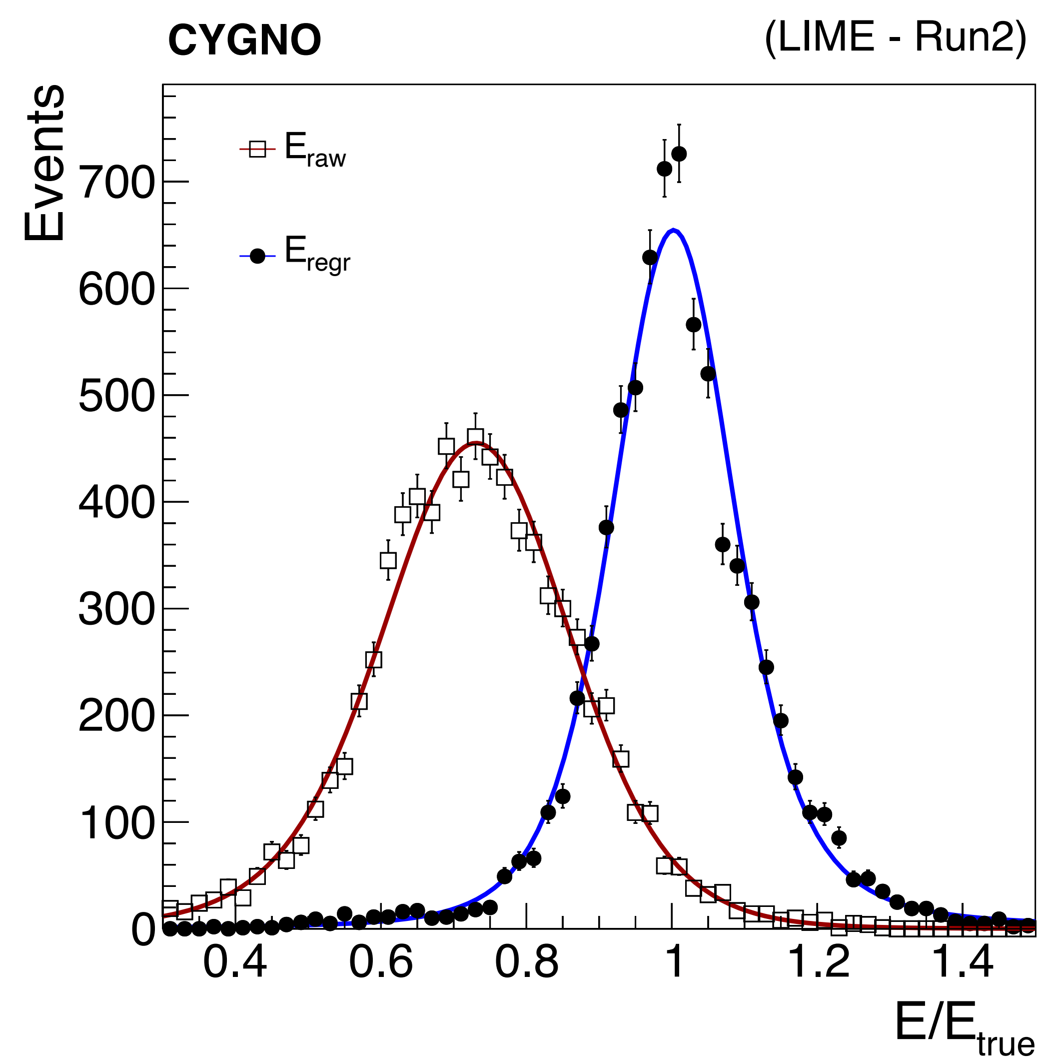
- Z-scale in the plots rescaled by the mean of the $\hat{E}$ distribution for a fair comparison
- Regression flattens the energy response in x-y, very visible close to the GEM sector boundaries
  - Some step for y<600 to be understood
- $\hat{E}_{\mathrm{mean}}$ similar, but a bit worse around the boundaries

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

- Fit $I_{SC} \equiv E_{\text{raw}}$ and $\hat{E} \equiv E_{\text{regr}}$ 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)



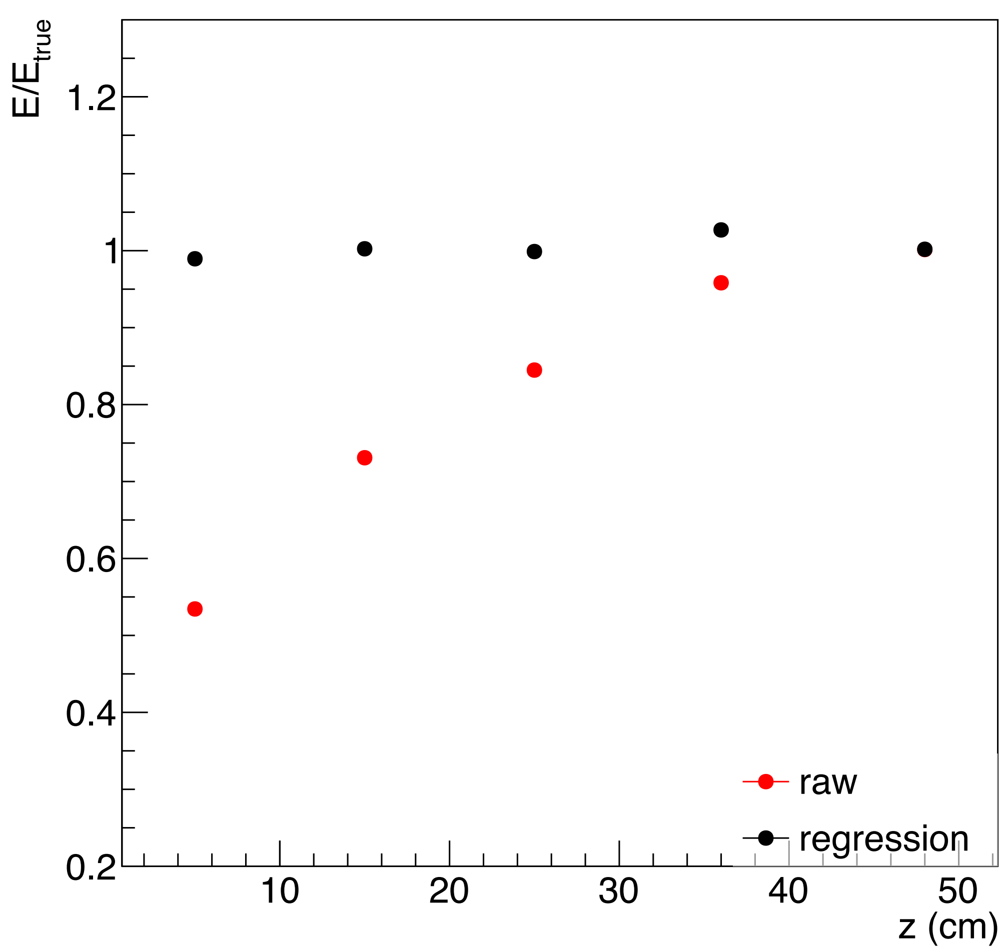
$z_{\text{true}} = 5\,\text{cm}$

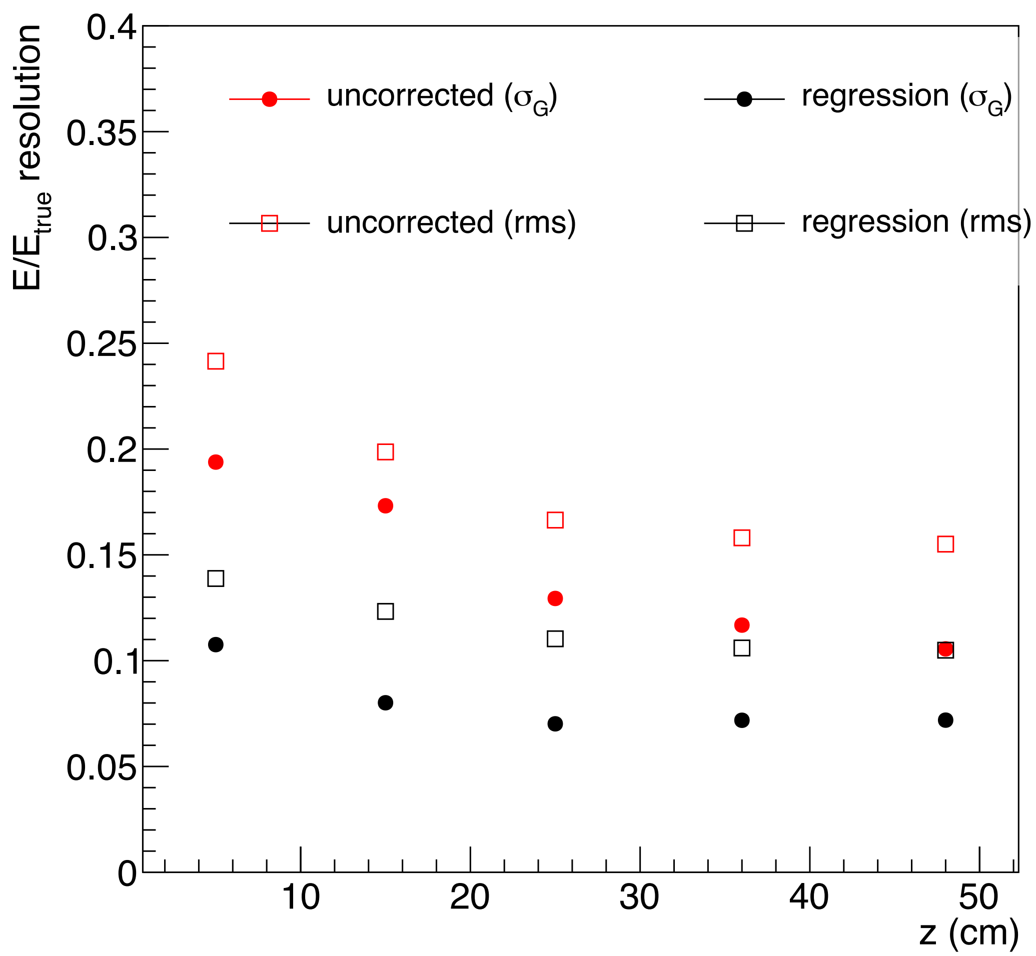


$z_{\text{true}} = 15\,\text{cm}$

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

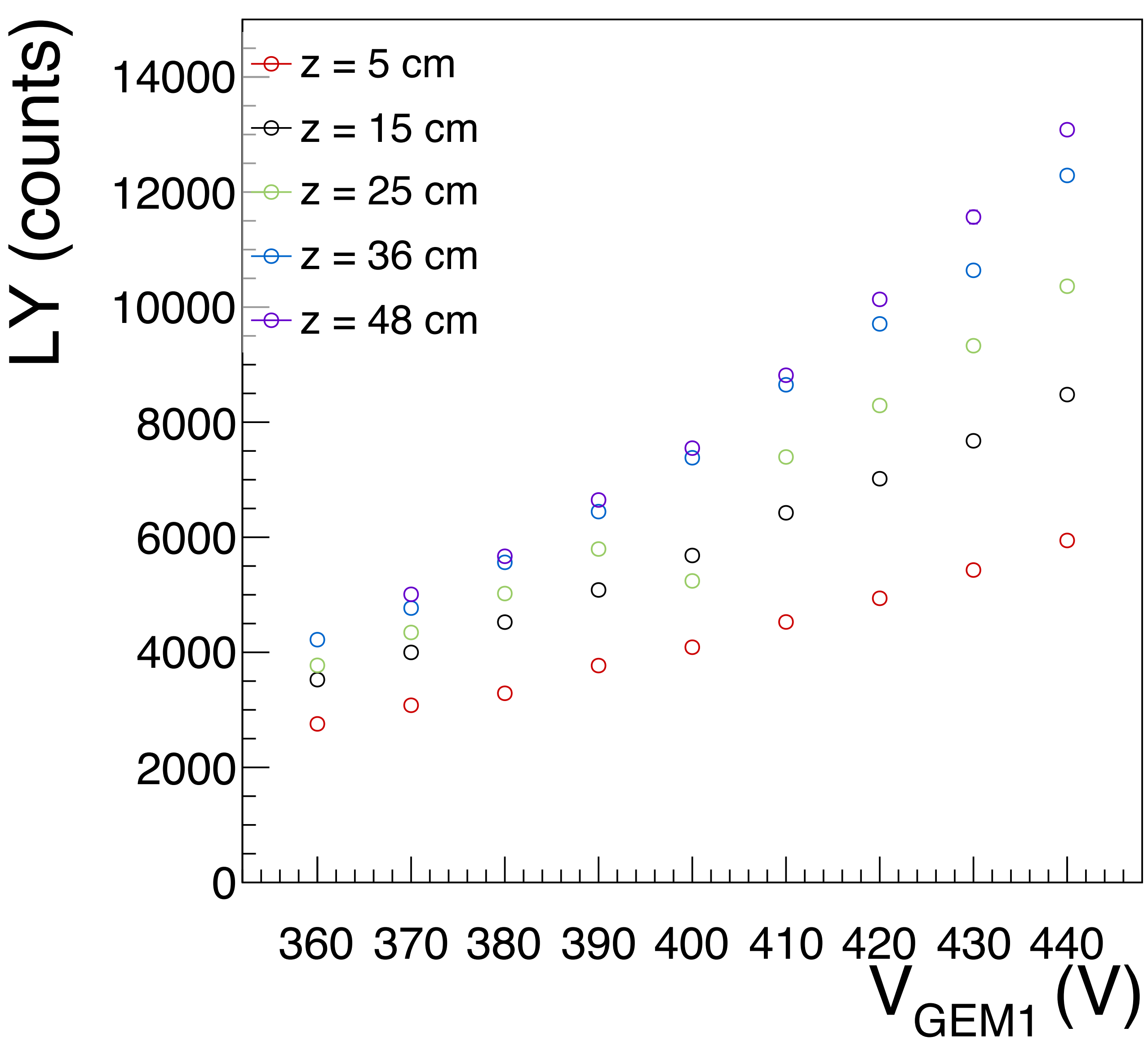
- Raw LY varies by a factor 2 for z in [5,48] cm, as known
- Corrected $\hat{E}$ (here median, but similar for mean) almost flat
- Energy resolution improved at any z
  - Estimate **11% improvement (in quadrature) at z=48 cm**, i.e. the contribution from the non-z dependence
  - 19% improvement at z=5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

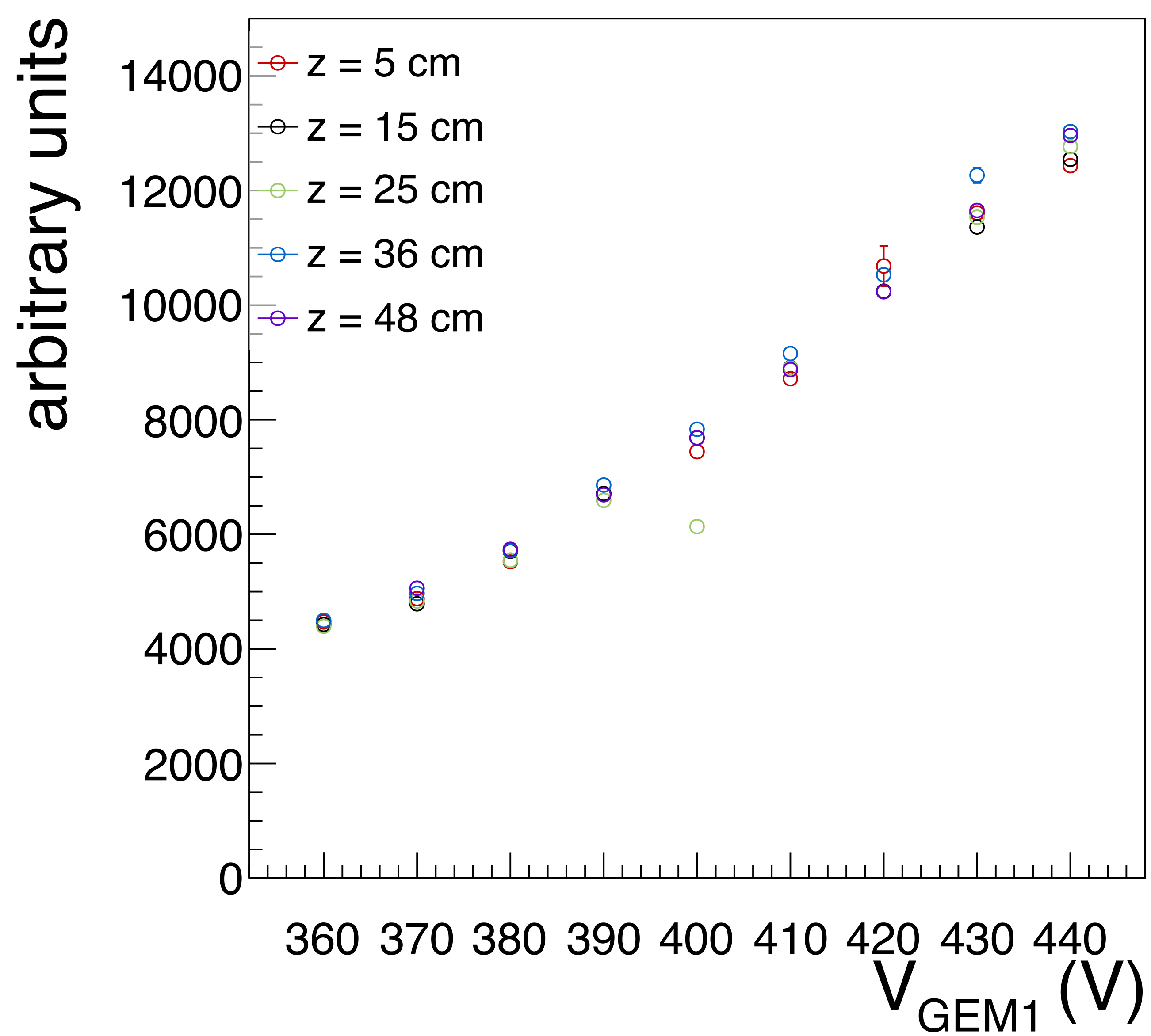




Closure test: response

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

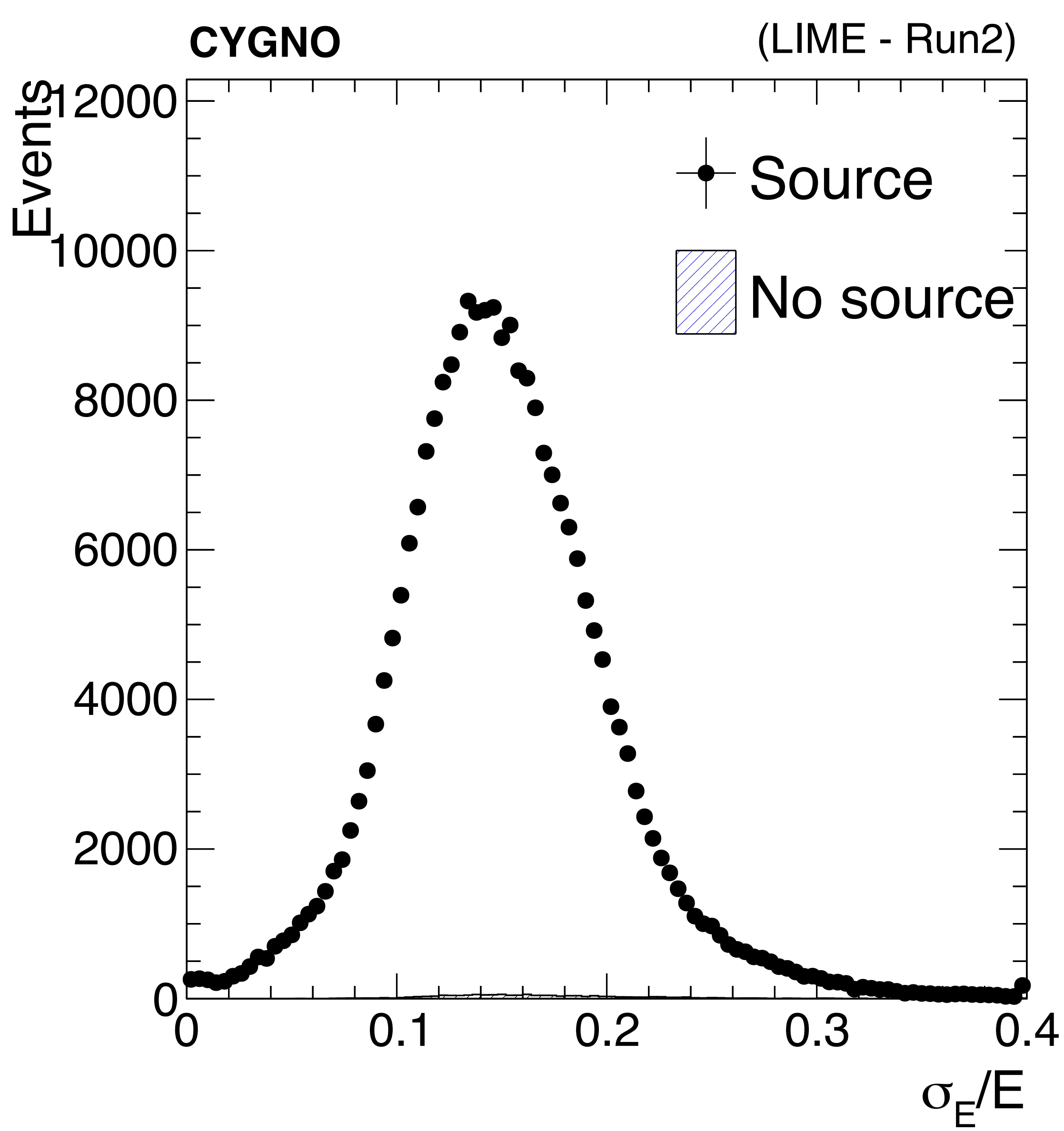




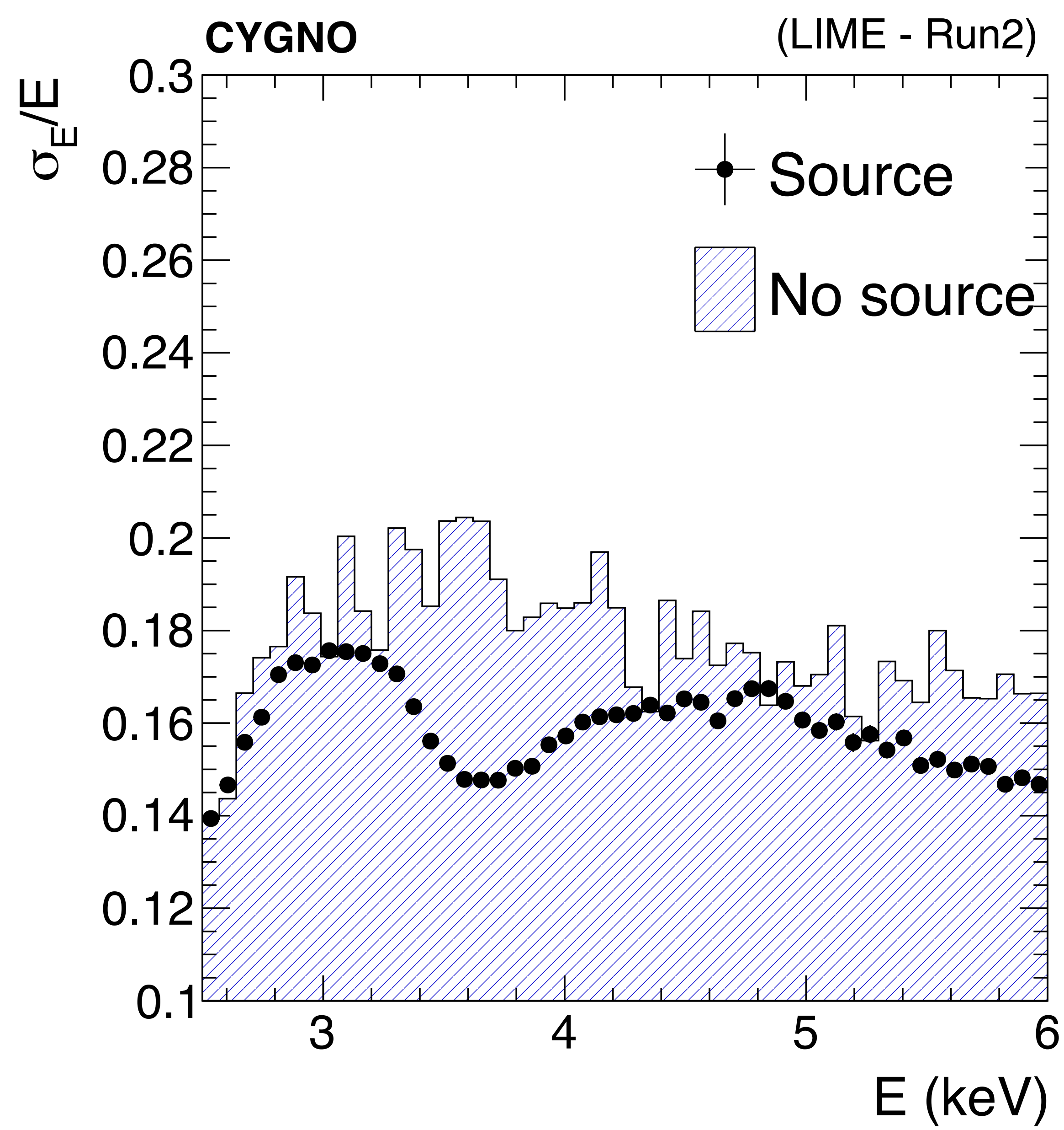
The correction of saturation holds at any (mocked up) E_{true}

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

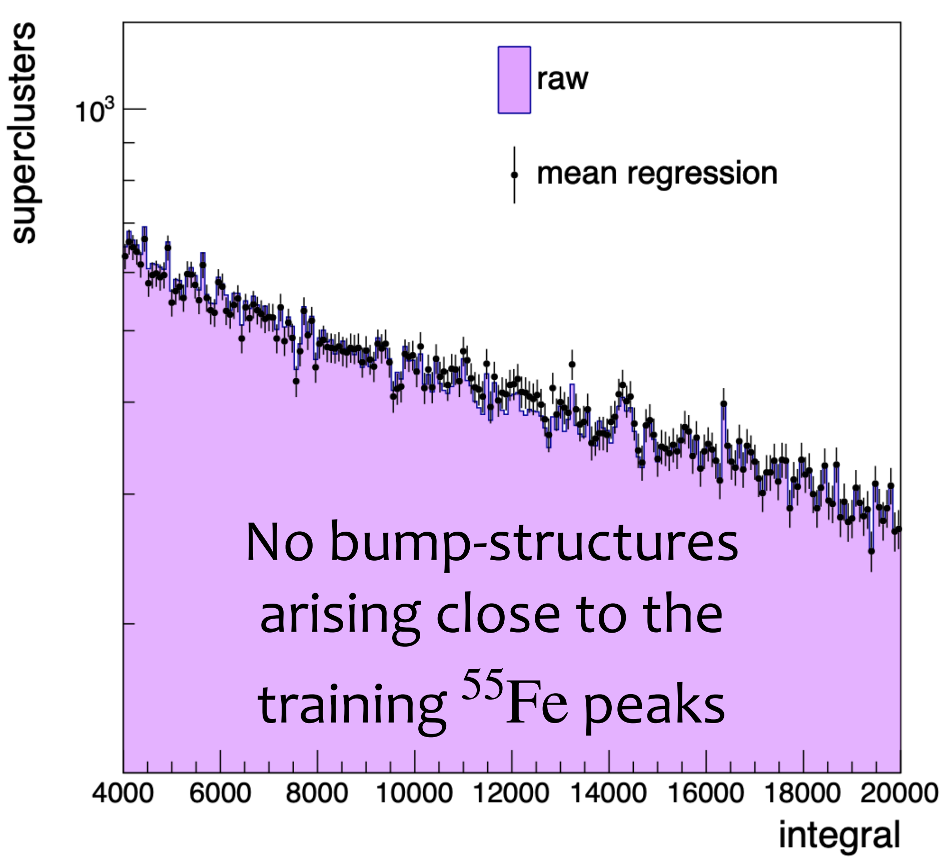


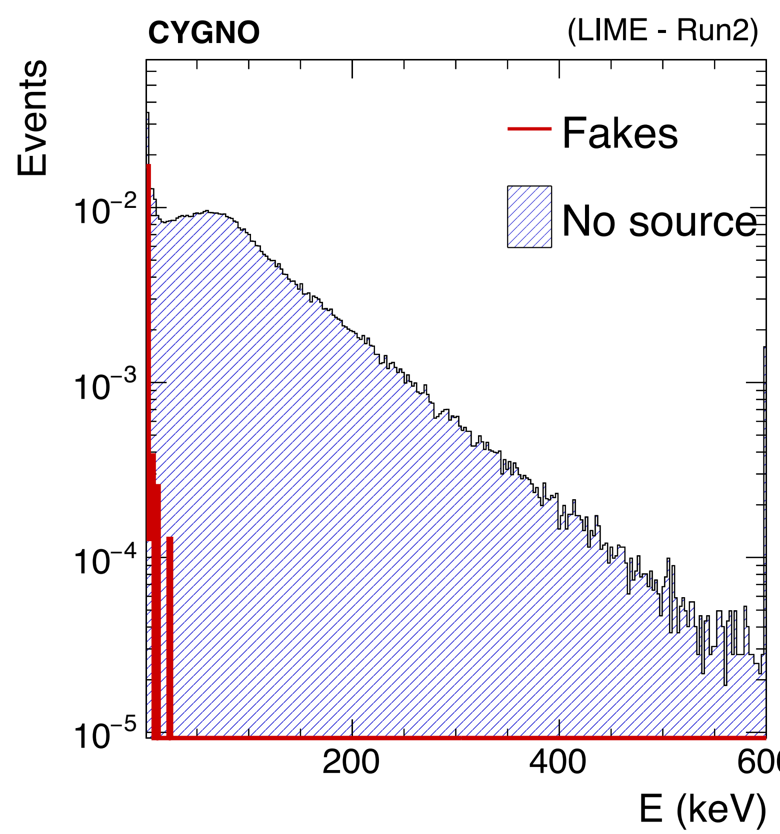
There are 11 steps in E_{true}, but resolution mixes

Need to disentangle different z's

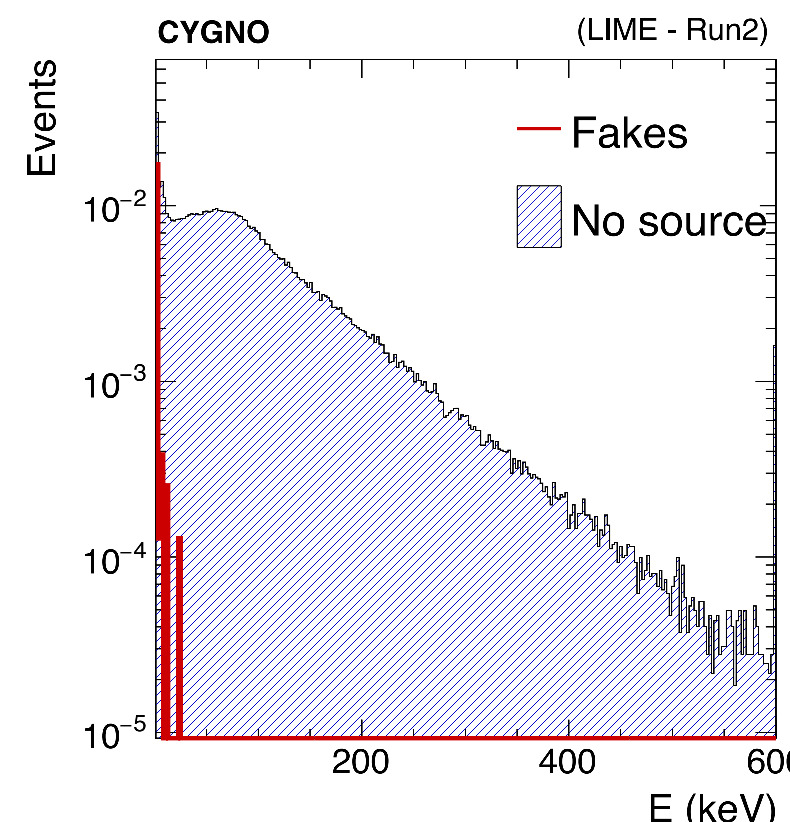
Validation: $\hat{E}$ in no-source data

- 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}

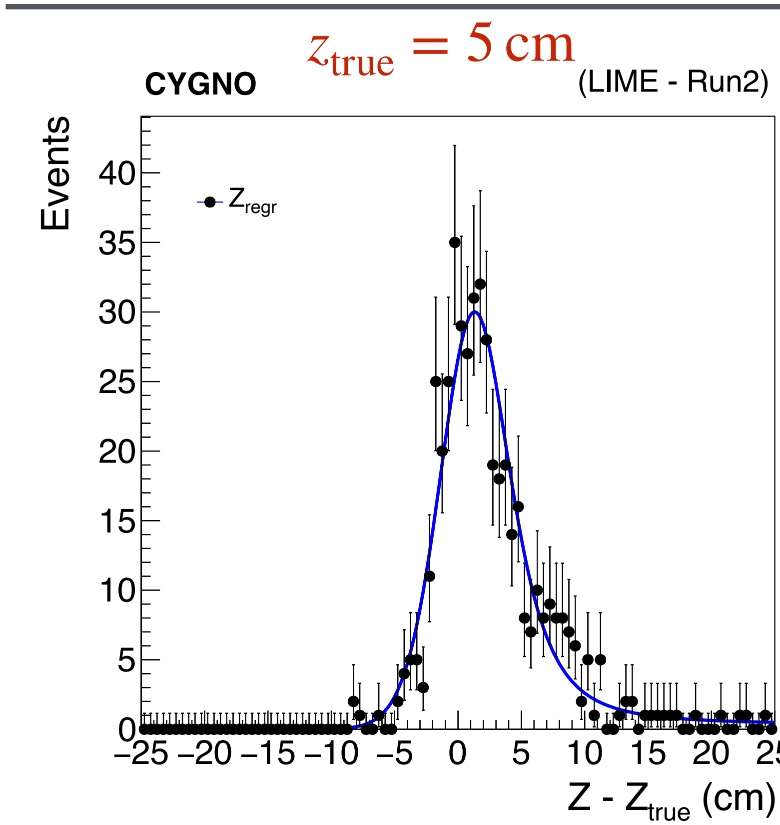


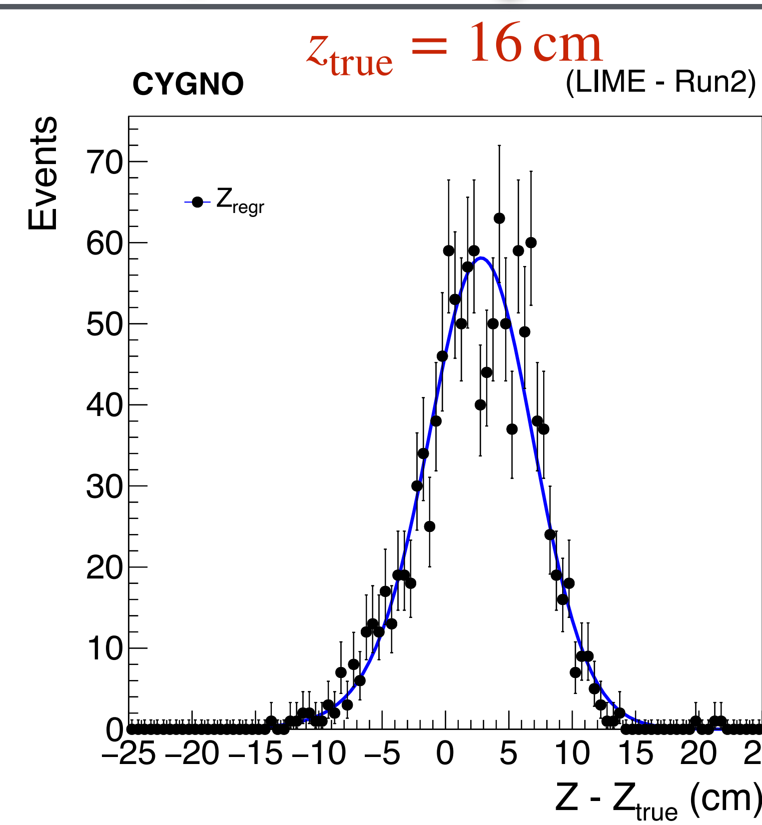
E from $\hat{E}_{50\%}$

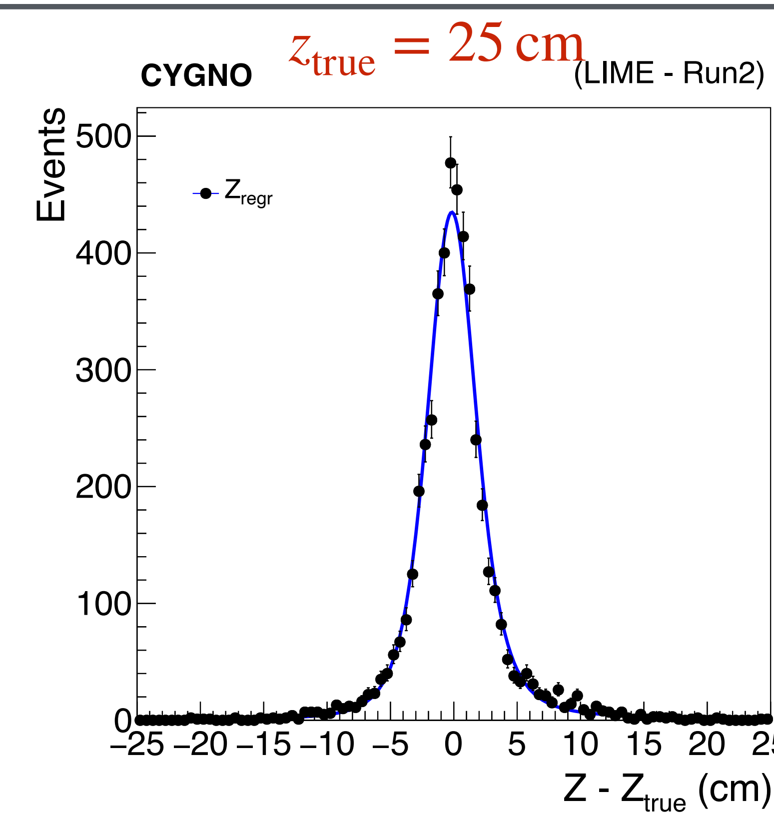
Z regression

- 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 $\pm 0.5\ \mathrm{cm}$ uncertainty (conservative)
  - In addition, the collimation of the source adds another $\Delta_z^{\mathrm{collim.}} \approx 8\ \mathrm{mm}$ to the z_{true} of the interaction
  - ☞ for "internal" z positions, smear the true value by a Gaussian with $\sigma_z = 1\ \mathrm{cm}$
  - To avoid border effects, for $z = 5, 48\ \mathrm{cm}$ 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

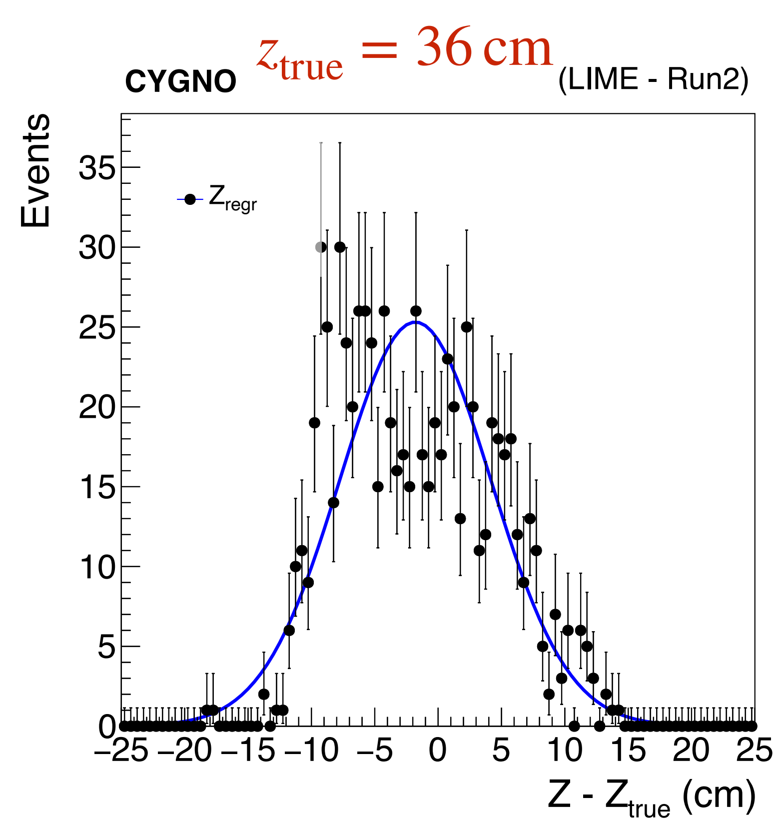
Output of $\hat{z}$-regression: bias

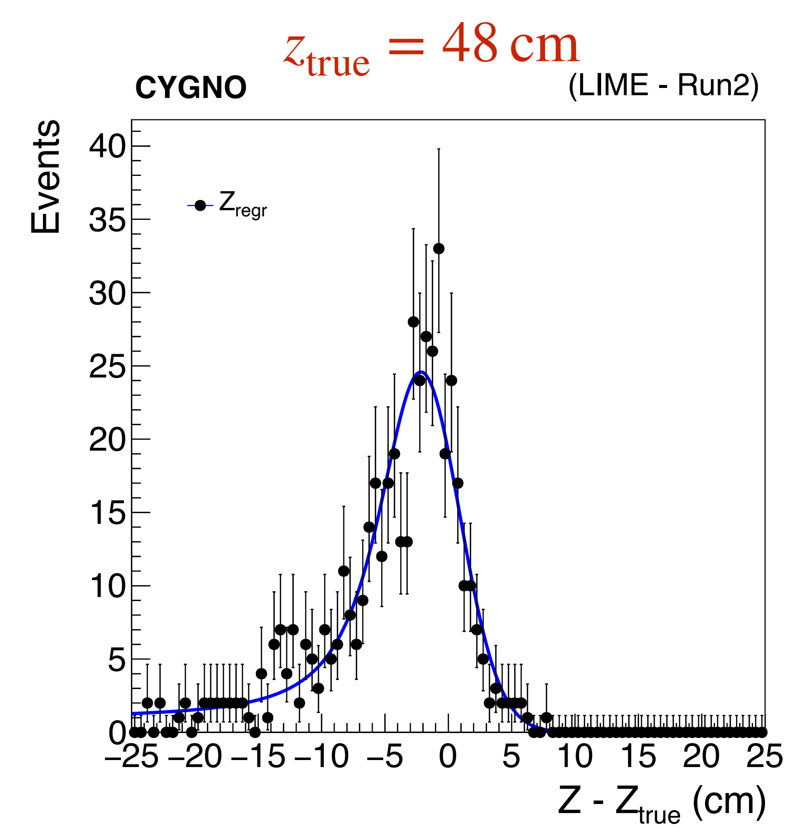






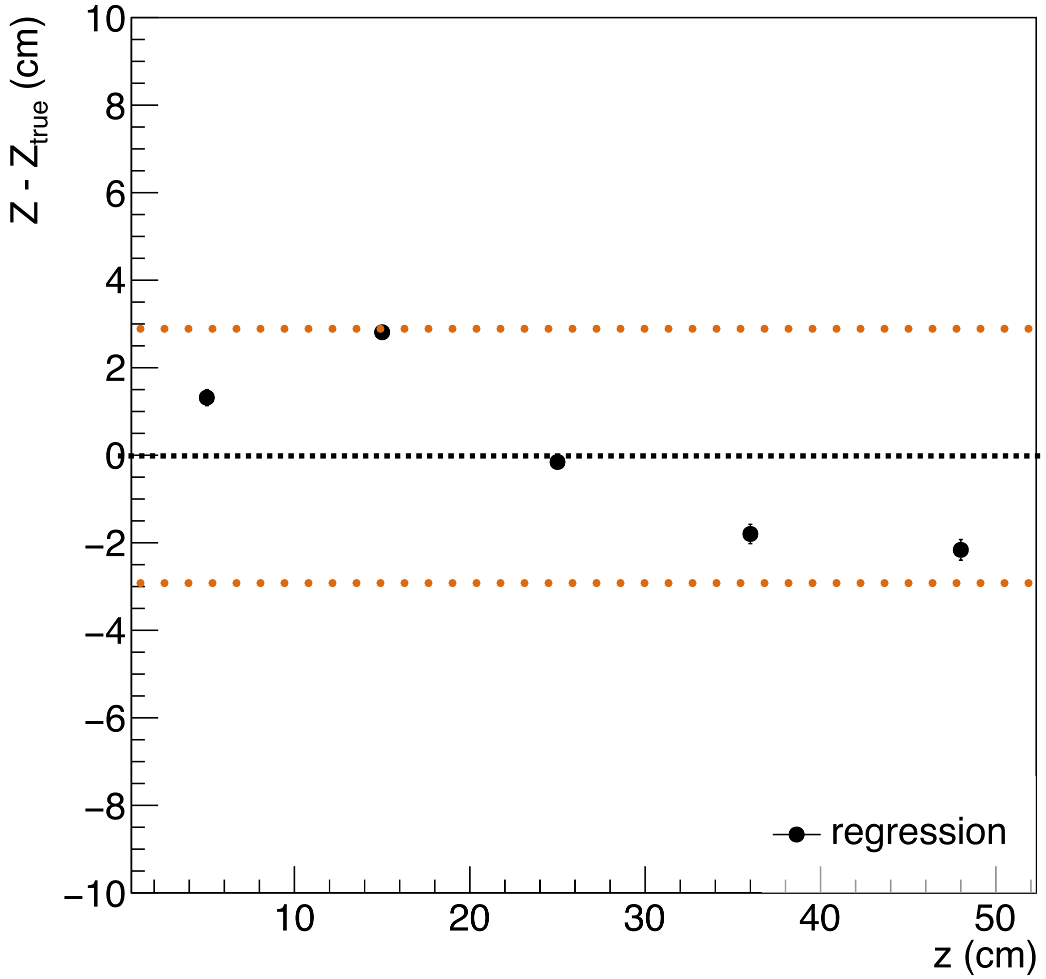
- Output at center: **no bias,** $\sigma_z \approx 2\,\text{cm}$
- Output at extrema: **small bias (1-2 cm)**, understandable because cannot predict out of detector, $\sigma_z \approx 3\,\text{cm}$
- 3-4 cm bias in the intermediate positions, to be understood

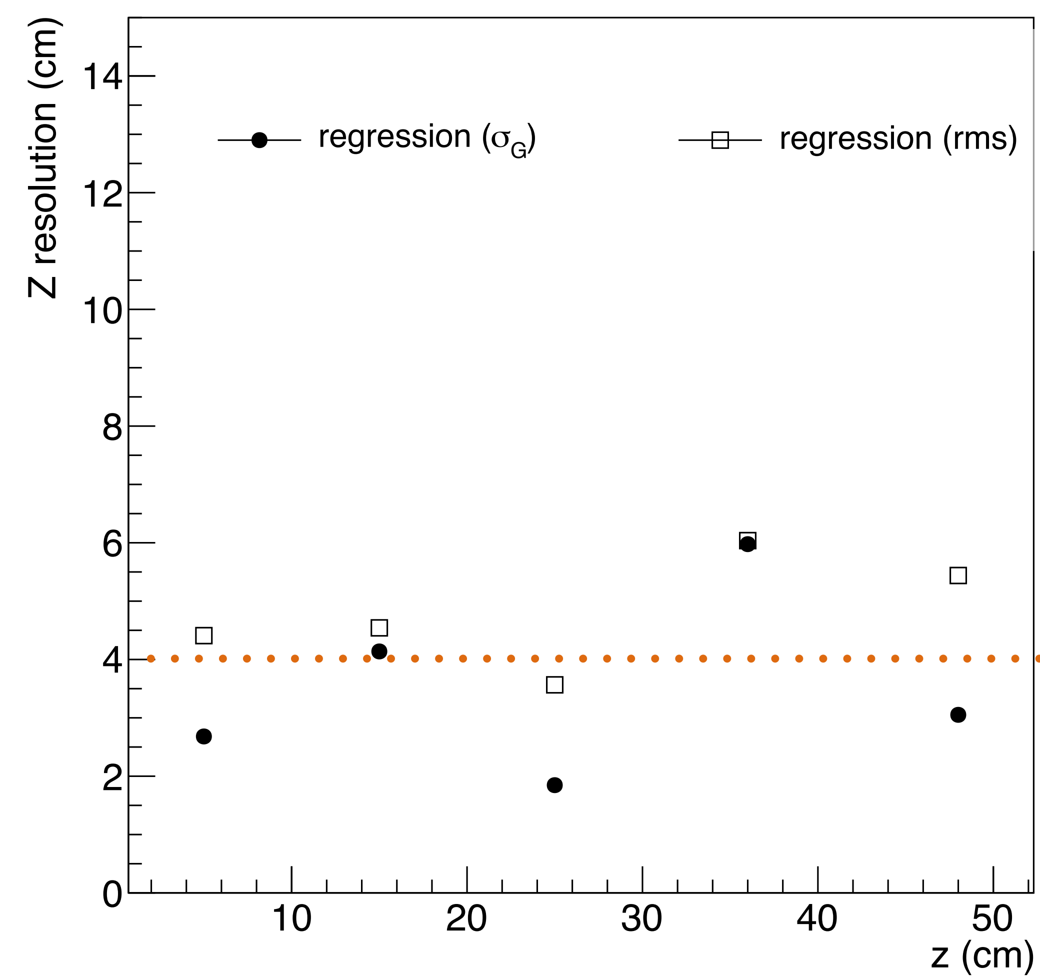




Output of $\hat{Z}$-regression

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





Conclusions

- Energy and Z MVA regressions trained on the 2D [z; HV] scans using ^{55}Fe source mimicking different energy equivalent to a LY of ERs in ~[2-6] keV at HV=440 V
  - Results for energy seems good in terms of correction for x-y non-uniformities (like the LNF one)
  - Also big improvement in terms of correction from saturation
    - This sensitivity wrt the LNF one comes from having multiple "energy"-equivalent points at a multiple z values, allowing a good model fit of the $E = f(E_{\text{true}}, z_{\text{true}} \,|\, \vec{\theta})$ likelihood function
  - Small bias at any energy, and **resolution around 10% at any z or E**
  - **Cluster-by-cluster energy estimate consistent with the predictions**
  - Limitations in the applicability:
    - Restricted to the phase space of the training, mostly: short tracks with an energy deposit similar to the 6 keV ERs.
      - The bias outside the training phase space could be estimated with MC
    - Could be different in ERs and NRs (again, MC can shade *some* light)
  - Validation: Z regression trained and shows reasonable prediction, but biases for intermediate points to be further investigated. In any case **Z bias < 3 cm and $\sigma_z \approx 4\,\mathrm{cm}$**
- The estimated energy and z from the regressions are computed and stored in trees copied on the cloud for ANY run of Run2.
  - Can be attached to all other variables of the trees as "friend" tree

The End