Update on Multivariate energy regression

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

CYGNO reco and analysis meeting,
28 October 2021

Remainder of the method

- Multivariate regression is a method to exploit the dependency of the CYGNO camera(s) light yield on several variables (x-y position, diffusion vs z-distance, etc) simultaneously to improve energy resolution

- Input variables are used to train the regression using the Gradient Boost Regression (based on a BDT in scikit-learn) are X,Y position of the reconstructed cluster and several cluster shapes

- ^{55}Fe Data is used, because of the well known energy peak

- GBR target is cluster integral/peak value (to have a variable centered at 1

- Two alternative metrics are tested:

  - mean squared errors

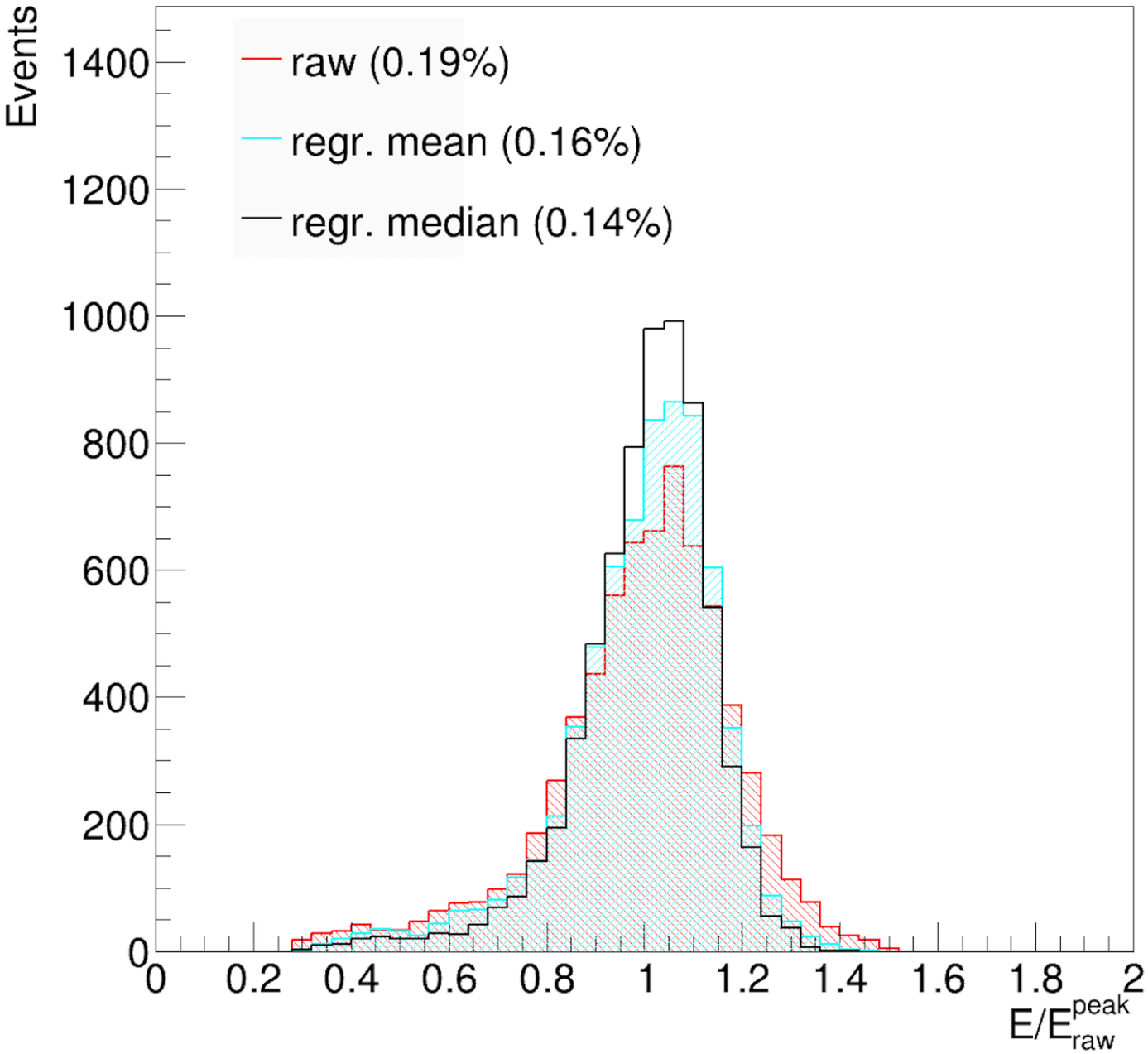
  - 50% quantile (median), and 5% and 95% quantiles

    - 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

Inclusive Results

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



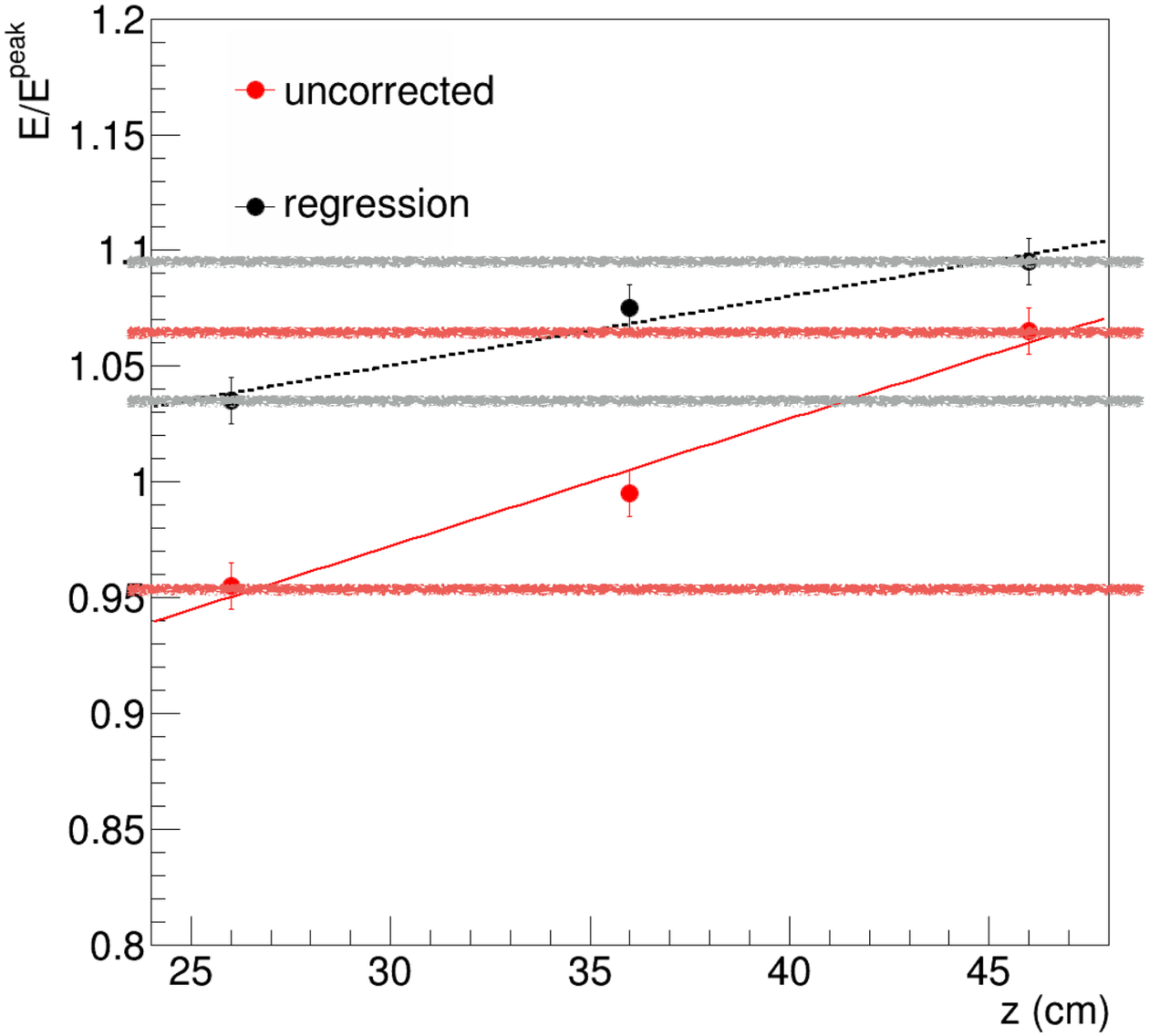
Regression gives significant improvement to the energy resolution

- 10% for mean regression and 14% for quantile regression in quadrature

It is necessary to check the robustness vs E

- check with SIM
- check with multi-E X-ray source data

Saturation correction?



Energy response using RAW energy estimate varies 11% from 25 to 45cm

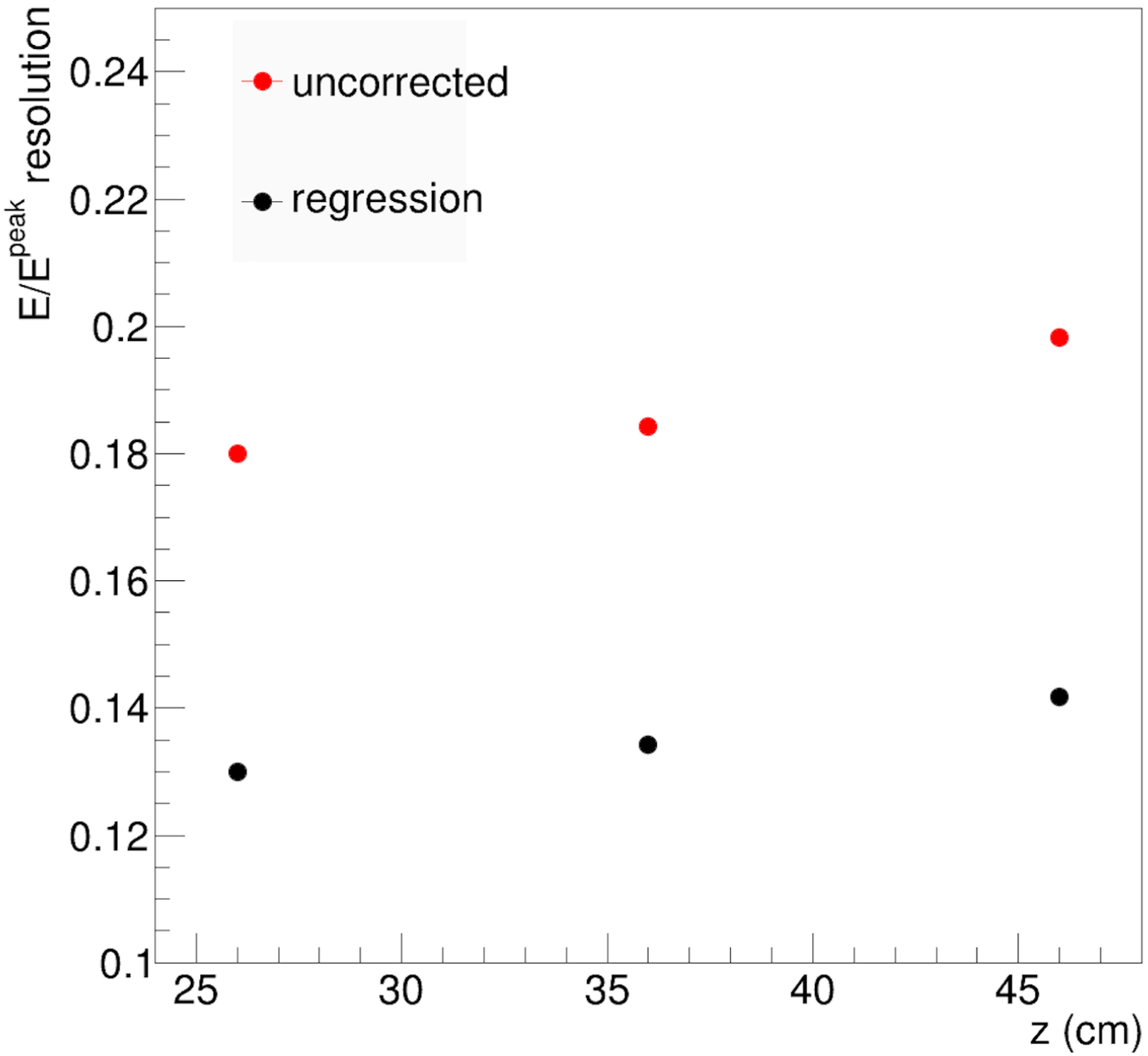
- two possible competing effects: saturation (main) and transport efficiency + diffusion

Dependency reduced to 6% with regression

- need to validate the method with more points in Z (April ^{55}Fe data)

Caveat: interplay with simulation of saturation! (check data-MC again...)

Resolution vs Z



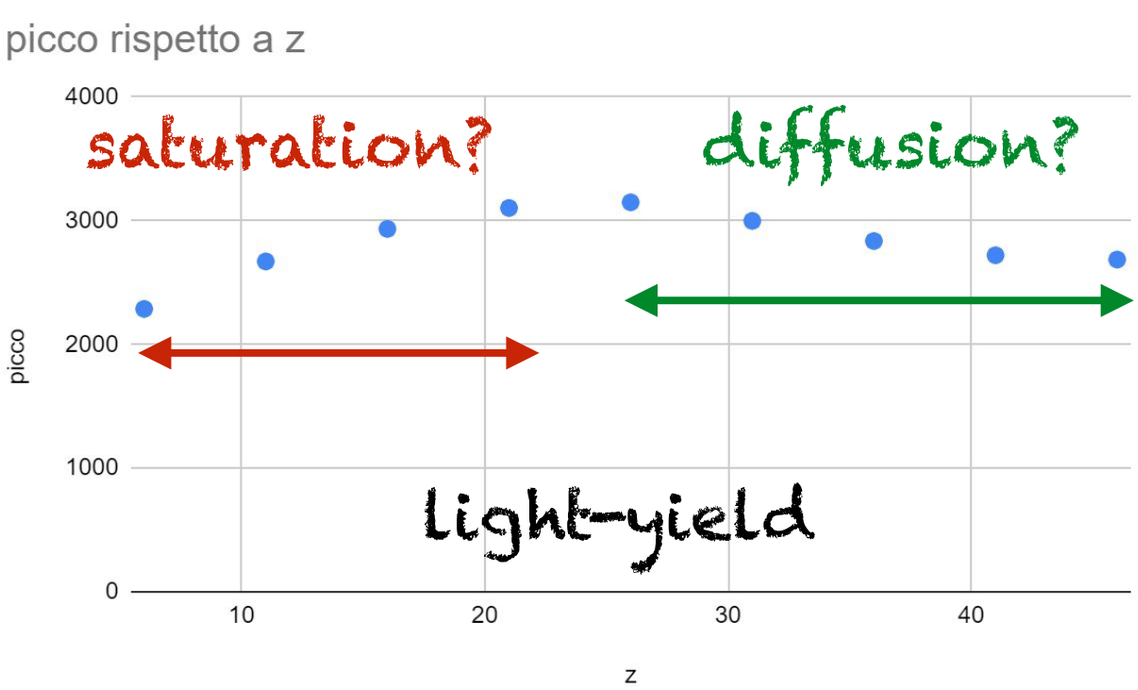
Improvement in resolution substantial at any z

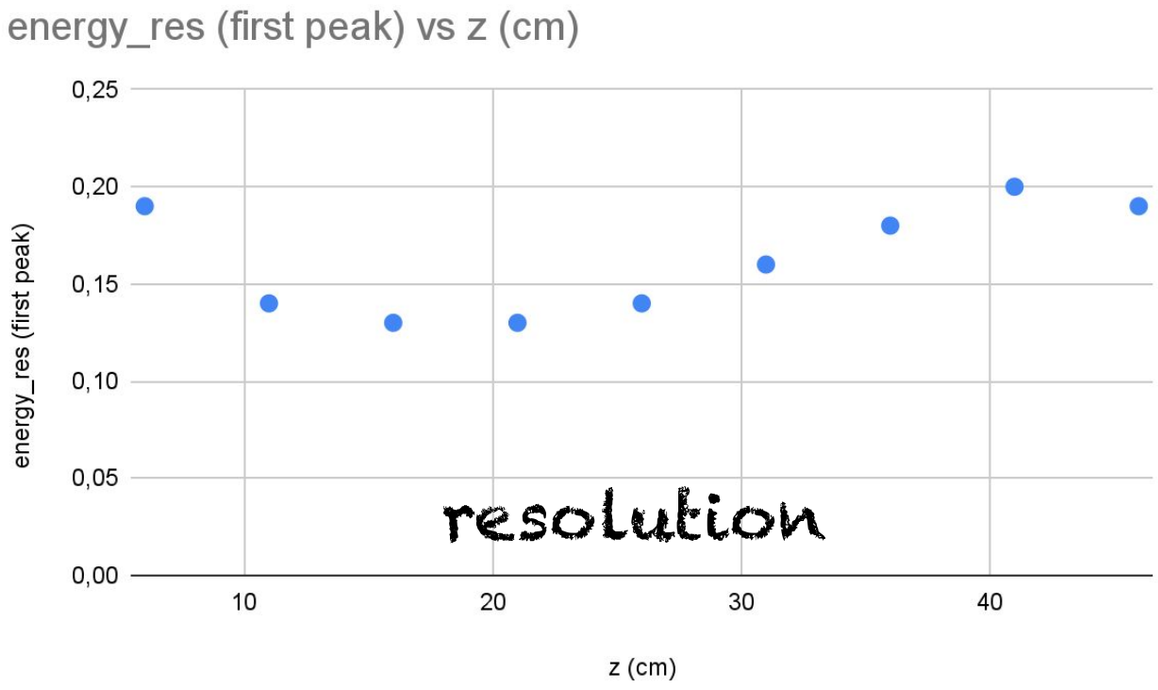
April LIME data

Reconstructed all the runs with 55Fe. Will show only the ones with V_{GEM1}=440 V

Since the exposure is 1/5 of the July data (10ms vs 50ms) because data were taken without the std DAQ, so no limitations on exposure time, bkg from cosmics is much reduced (by 1/5).

These runs cover a wider range for Z: 6.5 / 11 / 16 / 21 / 26 / 31 / 36 / 41 / 46 cm. Results from Donatella Tozzi's analysis shows this response behavior:





Regression re-training

The regression is re-trained on this data for consistency, using as training sample data with source at all Z.

Added variables that in principle are correlated to possible sources of saturation, to correct for it (e.g. # pixels over thresholds / # total pixels of the clusters <-> local energy density in a GEM hole)

The inclusive improvement in resolution seems similar to the one shown in July data. Try to evaluate the performances more quantitatively:

1. Fit the light yield distribution with a Crystal Ball

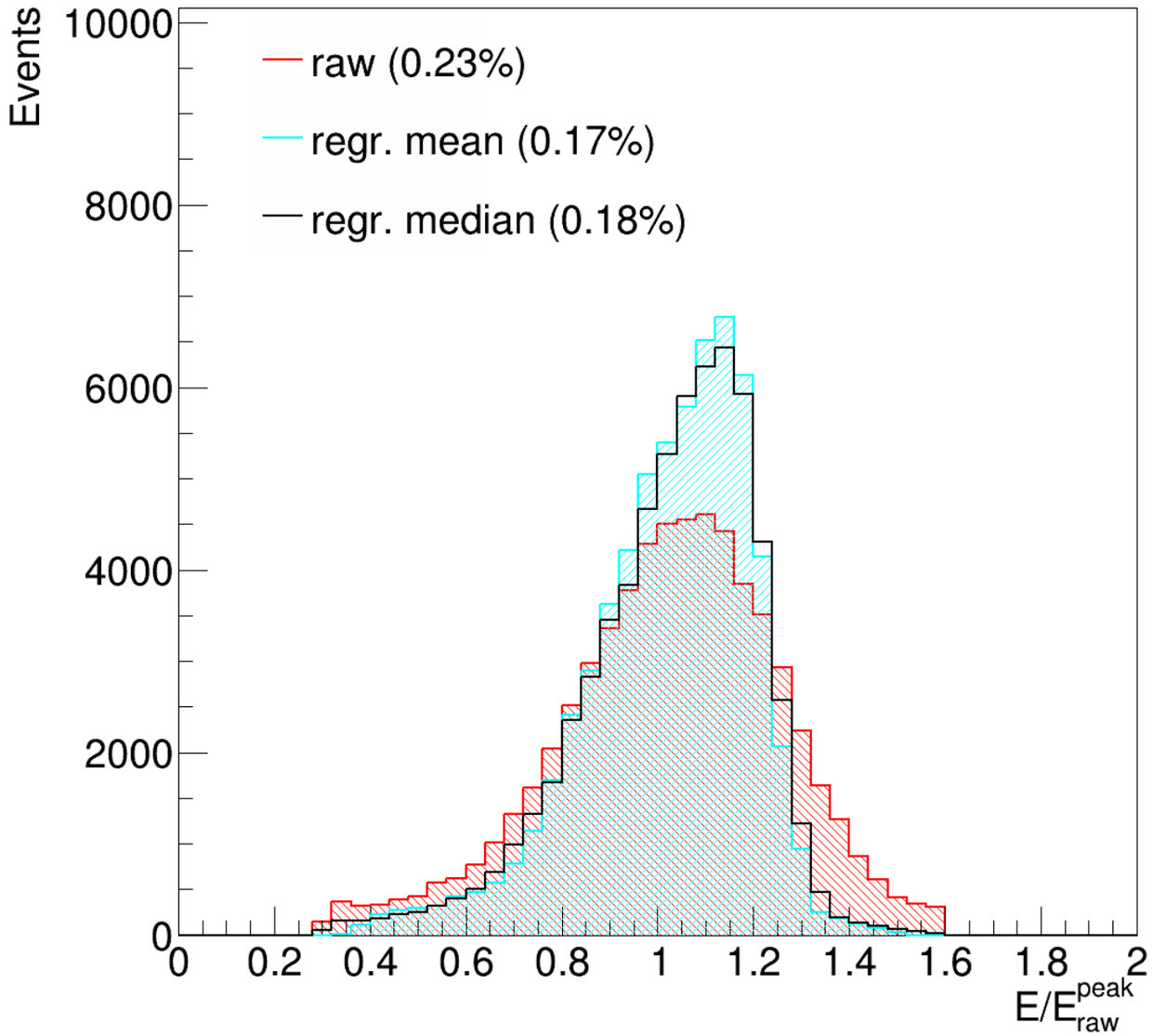
2. estimate the response as the peak position

3. estimate the resolution both with the σ of the Gaussian core (optimistic) and with the full RMS (including tails)

- the hypothesis is that the core is driven by the intrinsic resolution, and the tail by the effect of saturation

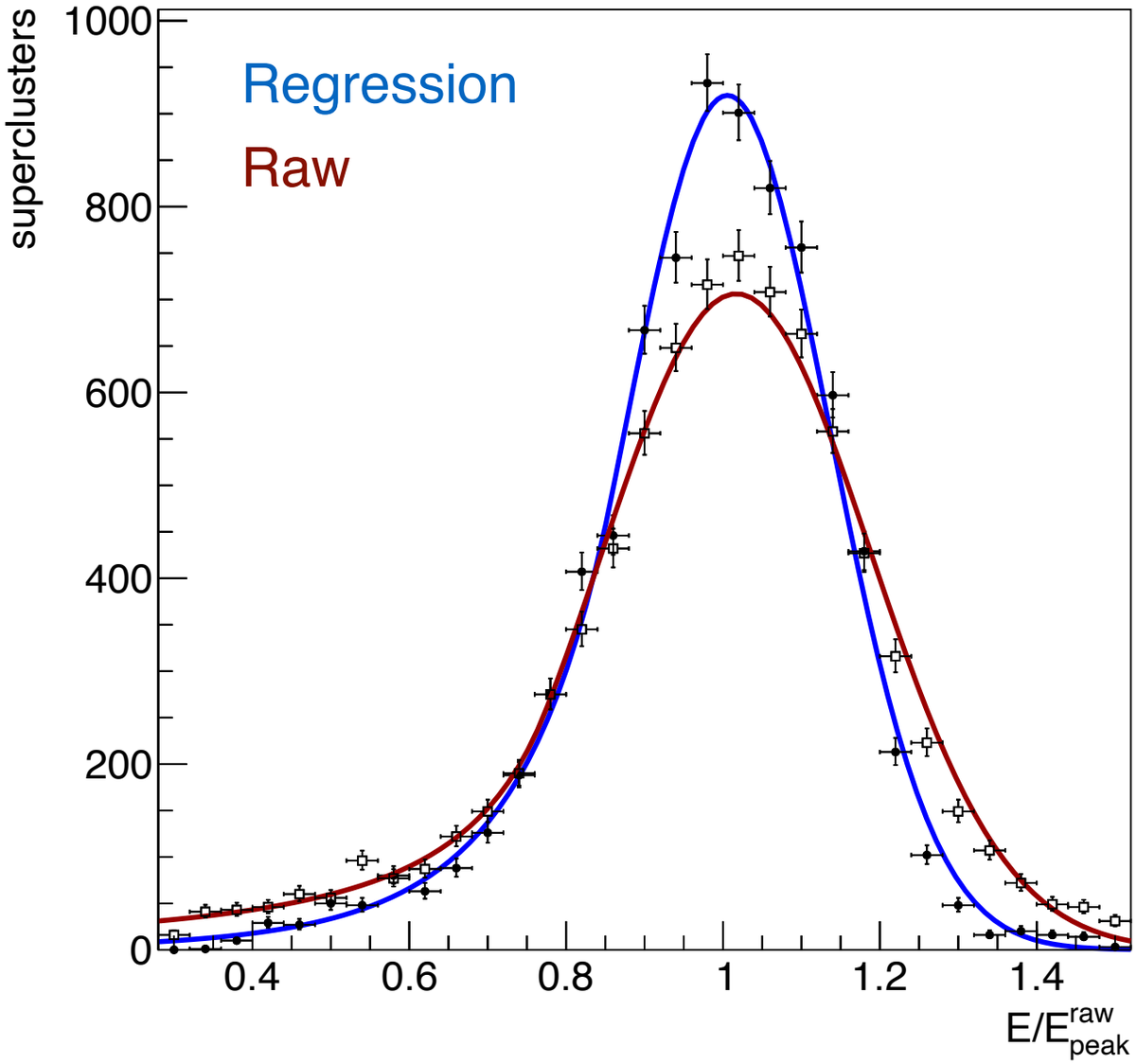
Inclusive results

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.

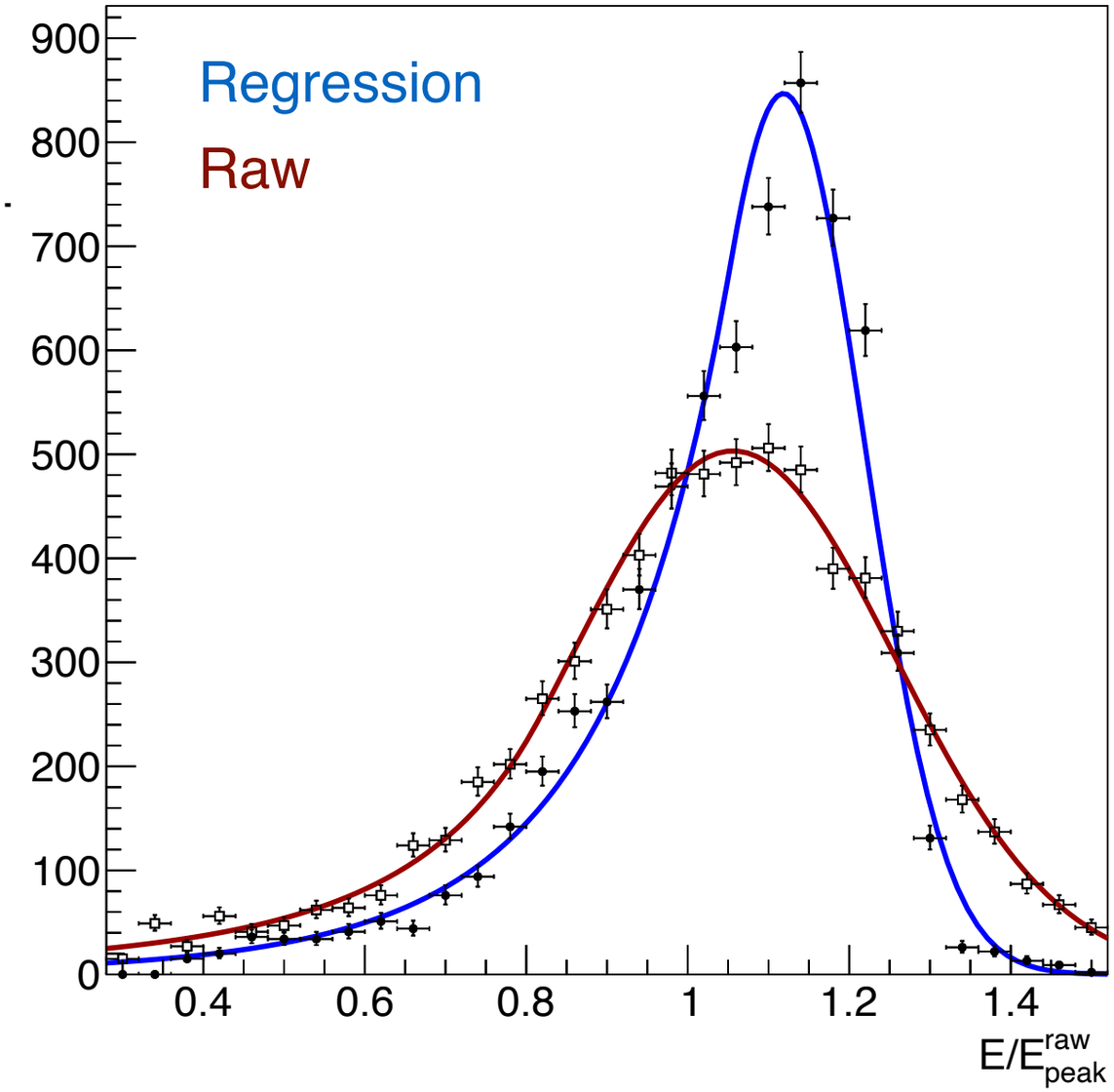


Examples at different z's

z = 11 cm

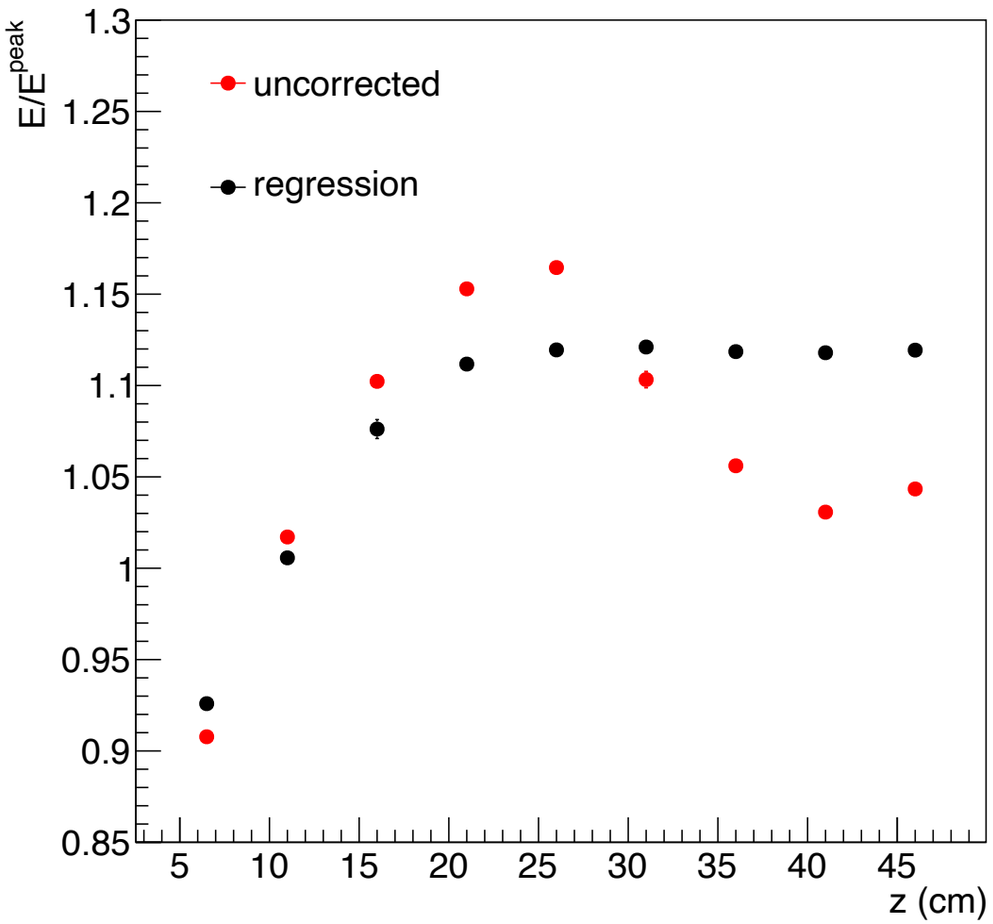


z = 36 cm



Results vs Z

light yield peak

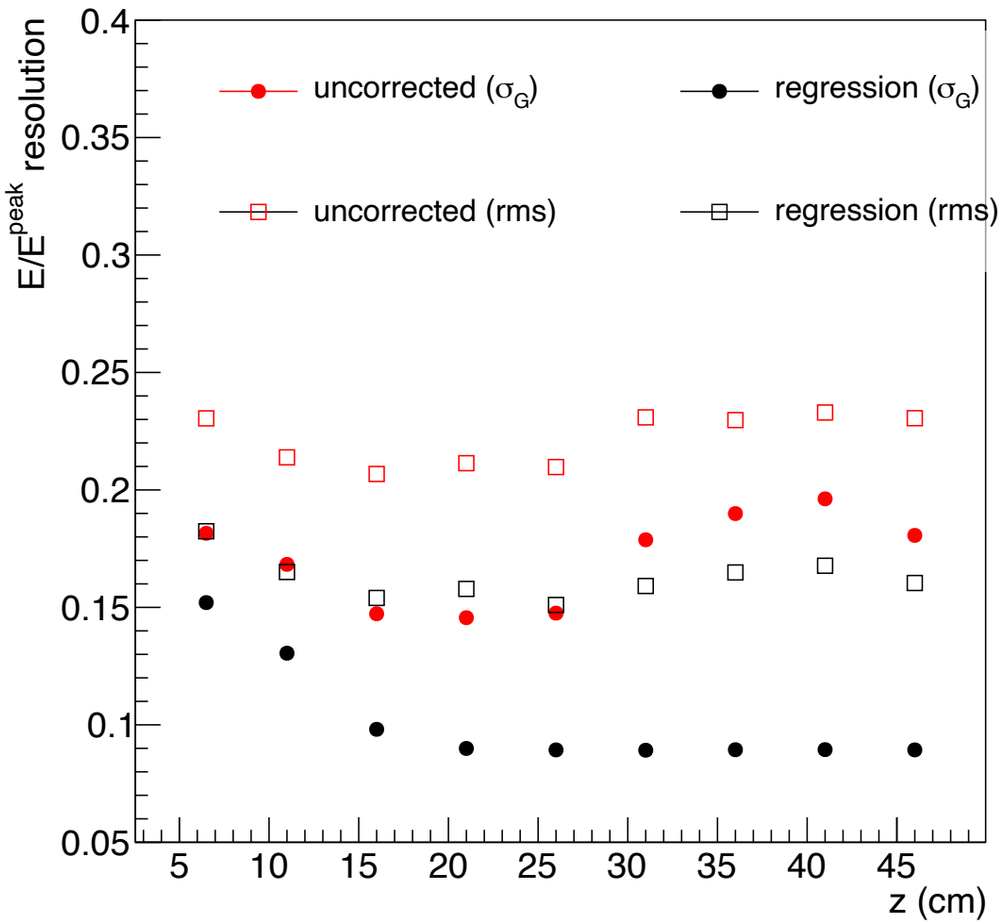


Regression does NOT correct (yet) for saturation

=> look for more sensitive variables

Regression cures the variation vs z when there is not saturation

light yield resolution



Resolution significantly improved everywhere

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

Conclusions

- Multivariate regression gives a large improvement in energy resolution for clusters from 5.9 keV photons

- Regression can give a resolution estimate as well which can be used for categorization, or on a per-cluster basis

- Using the April data it seems that regression can improve the energy resolution of 6 keV ele-recoils everywhere when there is NOT saturation

  - energy resolution maybe can be improved to better than 10%

- for low z's, saturation dominates and with these input variables it does NOT correct for it (studies ongoing...)

- Writing a module to calculate regressed energy on top of the ntuples ("post-processing" reconstruction step) from stored GBR likelihood configuration

  * needed to make some extensive validations and closure tests