ML-assisted versatile approach to Calorimeter R&D

Alexey Boldyrev
on behalf of the LHCb Calorimeter Upgrade group

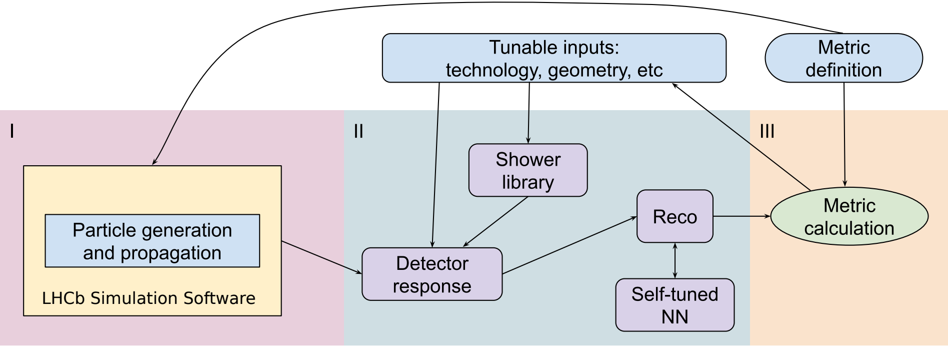
Lambda Lab @ National Research University
Higher School of Economics (Moscow)
The Yandex School of Data Analysis

Instrumentation for Colliding Beam Physics (INSTR'20)
Novosibirsk, Russia, 24–28 Feb 2020

Outline

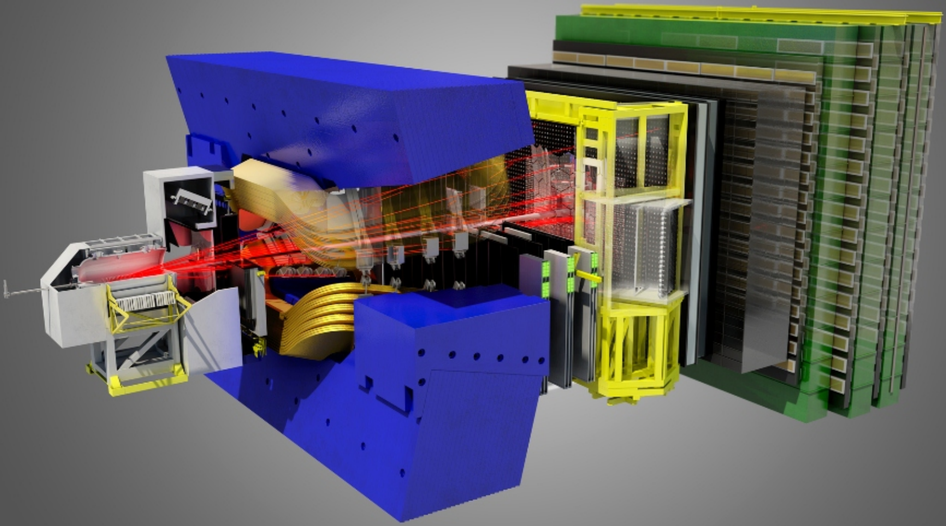
- Strategy
- LHCb Electromagnetic Calorimeter
- Samples and data preparation
- Simulation of the calorimeter response
- Results
  - Spatial reconstruction
  - Energy reconstruction
- Conclusions

Strategy & The pipeline



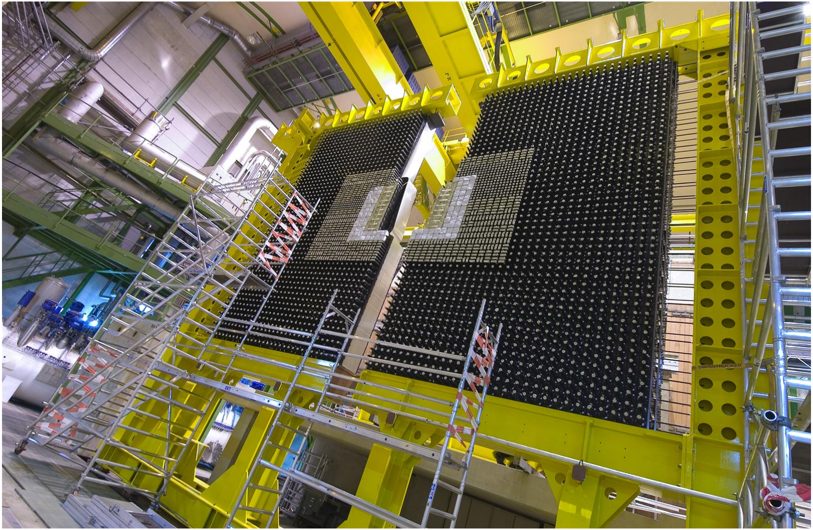
- Entire optimisation pipeline of calorimeter R&D is covered
- Optimisation cycle does not depend on the modules technology & arrangement, reconstruction, metric, etc.

LHCb detector

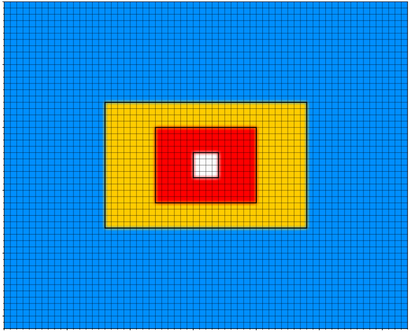


LHCb Electromagnetic Calorimeter (ECAL)

LAMBDA · HSE



Modules of LHCb ECAL



Module type	# of modules
(inner): 3x3 cells (4.04x4.04 cm^2 each)	176 (1536 ch.)
(middle): 2x2 cells (6.06x6.06 cm^2 each)	448 (1792 ch.)
(outer): single cell (12.12x12.12 cm^2)	2688 (2688 ch.)

Other differences:

- σ_E/E
- Molière radius

Samples and data preparation

Signal sample: $B_s^0 \to J/\psi(\to \mu^+\mu^-)\pi^0(\to \gamma\gamma)$

- Signal events are generated using PYTHIA8 with default LHCb tunings
- Signal photons to beginning of the ECAL using LHCb Simulation Software

Background sample: LHCb MC Minimum Bias sample for 14 TeV

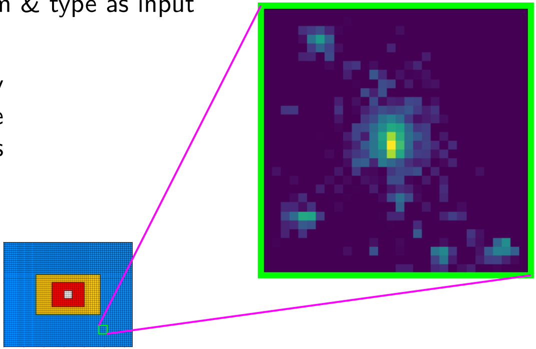
- Propagate all background particles to beginning of the ECAL

Background composition & cluster generation

We consider background contributions from γ, π^+, π^-, e^-, e^+, n, p ($>$ 92% of total background which reach the ECAL incl. secondary particles). For each of background particles we:

- Record momentum, type, hit position at the front of the ECAL
- Perform GEANT4 standalone simulation of clusters in 30x30 cells using the momentum & type as input

Thus, we have the library of the mapping of particle (*px, py, pz, type*) and its electromagnetic cluster.

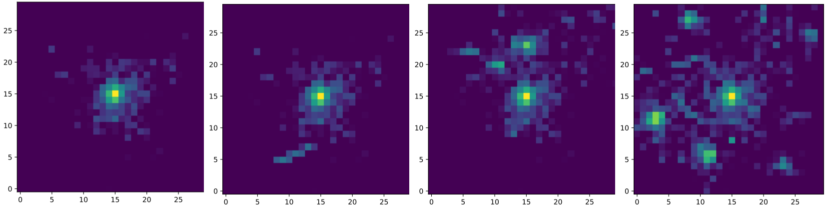


Electromagnetic clusters vs pile-up

Signal clusters were similarly produced using signal sample.

By stacking events which have multiple primary vertices (nPVs) arbitrary pile-up can be simulated.

The same signal cluster surrounded by the background clusters for different pile-up conditions:



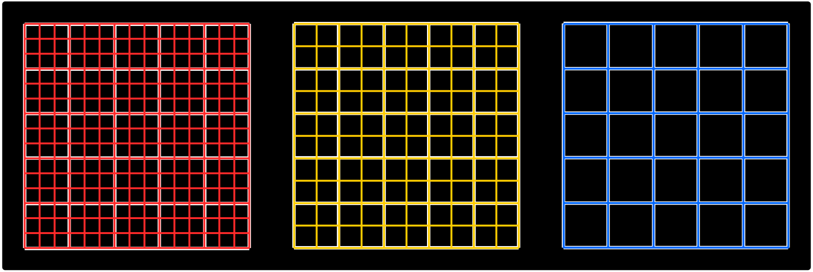
without pile-up | nPV = 5 | nPV = 20 | nPV = 50

ECAL-like simulation using GEANT4

Dataset of simple GEANT4 simulation of ECAL-like modules

- 66 layers of 2 mm absorber + 4 mm scintillator
- 5 x 5 modules
- each module is split over 6 x 6 cells

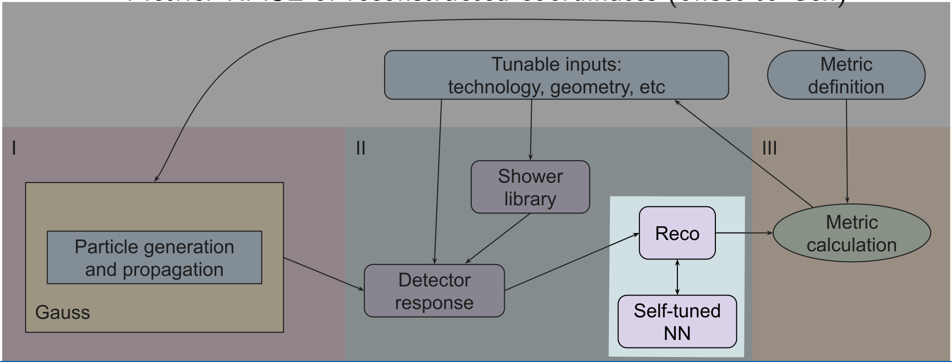
Thus, it can emulate all types of present ECAL cells:



Spatial reconstruction

For each type of ECAL modules:

- Cluster position determination (using its barycentre over 5x5 cells)
- Obtaining S-curve (Gen – Reco coordinates relationship)
- S-curve calibration using parametric and ML approaches
- Metric: RMSE of reconstructed coordinates (offset to Gen)

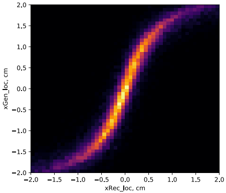


Approaches to S-curve calibration

S-curve was calibrated using:

- **Parametric approach**: $x = a \cdot arcsinh(b \cdot x_{rec})$

  Parameters were obtained using random search with 1000 points in the range (0.01, 100) for each parameter.



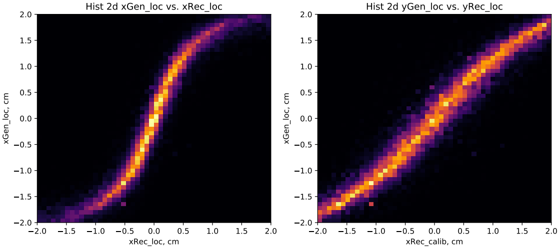
- **Machine Learning approach** (XGBoost[1]) with features:
  - Maximum energy deposit cell (seed)
  - Each cell energy deposit in 5x5 cells around seed
  - Sum of energy deposits in *3x3*
  - Sum of energy deposits in *5x5*

  The hyperparameters of XGBoost *colsample_bytree, gamma, max_depth* and *min_child_weight* were optimised using BayesSearchCV within the ranges (1, 20), (1, 10), (0.1, 0.9) and (0.3, 0.7).

[1] XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754

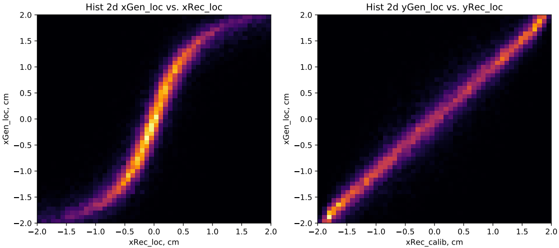
S-curve calibration: inner modules, nPV $= 0$

LAMBDA · HSE



$1.15 \cdot \text{arcsinh}(2.07 \cdot x)$

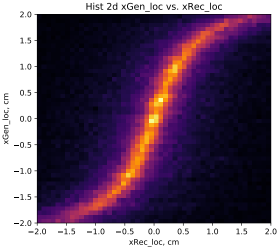
RMSE:
0.86 to 0.52

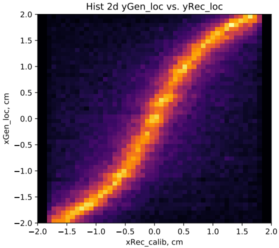


XGBoost

RMSE:
0.86 to 0.28

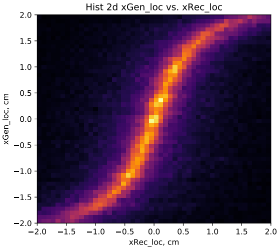
S-curve calibration: inner modules, nPV = 10

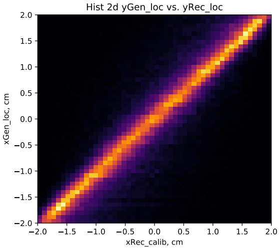




$1.13 \cdot \text{arcsinh}(1.96 \cdot x)$

RMSE:
1.39 to 1.38



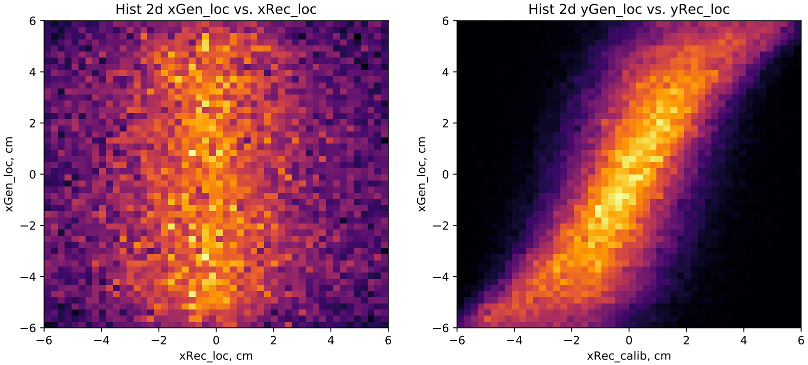


XGBoost

RMSE:
1.39 to 0.57

S-curve calibration: outer modules, nPV = 50

Outer region, cell size = 12.12 cm
nPV = 50



Left: uncalibrated, RMSE: 4.88 ± 0.05 cm
Right: calibrated with XGBoost, RMSE: 2.65 ± 0.02 cm

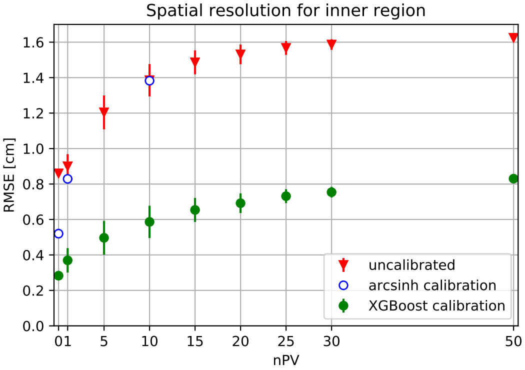
Spatial resolution results

Inner region, cell size = 4.04 cm

Spatial resolution is saturated for nPV > 15

Parametric approach is acceptable for spatial reconstruction without background

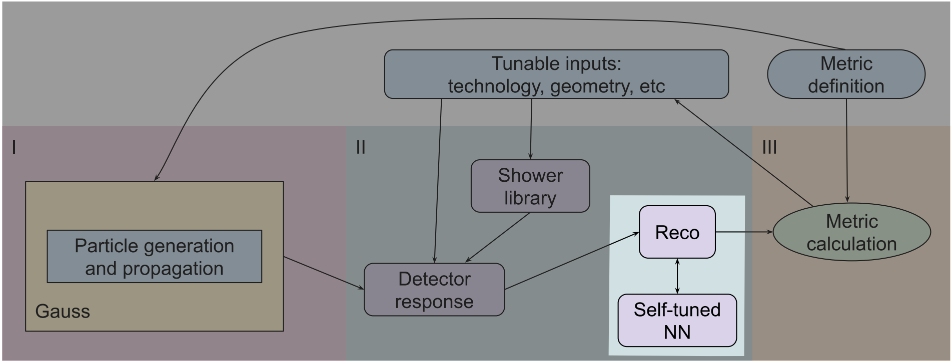
XGBoost shows good performance for entire range of pile-up



Energy reconstruction

For each type of ECAL modules:

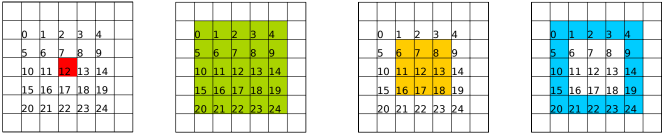
- Train a regressor to minimise reconstructed energy of signal clusters
- Metric: RMSE σ_E/E



Features for the regressor

Besides the basic features from the observables:
barycentre position of the cluster, particle incident angles

We used cell energy deposits around maximum energy deposit (seed cell):



E_{seed} or E_{seed}^2 | $\sum_{i} E_i$ or $(\sum_{i} E_i)^2$ | $\sum_{i} E_i$ or $(\sum_{i} E_i)^2$ | $(\sum_{i} E_i)^2$

Energy resolution results

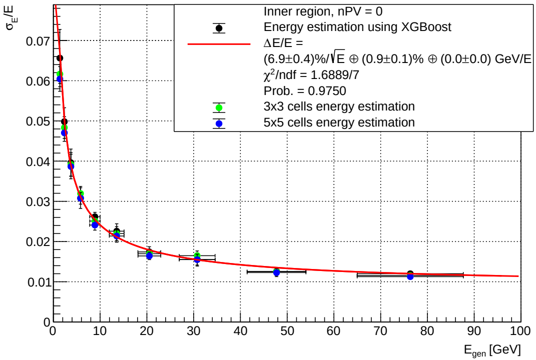
Inner region, cell size = 4.04 cm, **without pile-up**

$$\frac{\sigma_{reco}}{E_{reco}} = \frac{a}{\sqrt{E_{gen}}} \oplus b \oplus \frac{c}{E_{gen}}$$

a/b/c – stochastic/constant/noise terms

Energy resolution is consistent with LHCb ECAL design

XGBoost shows the same performance as for energy estimation using total energy in 3x3 or 5x5 cells



Energy resolution results

LAMBDA • HSE

Inner region, cell size 4.04 cm, **nPV = 10**

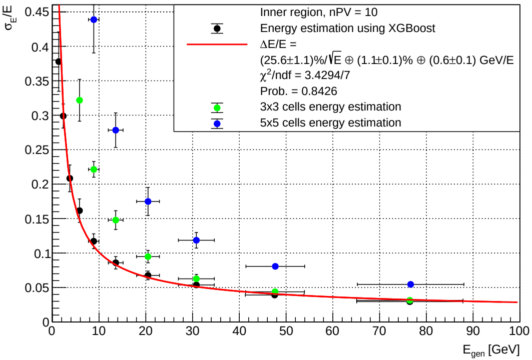
$$\frac{\sigma_{reco}}{E_{reco}} = \frac{a}{\sqrt{E_{gen}}} \oplus b \oplus \frac{c}{E_{gen}}$$

a/b/c – stochastic/constant/noise terms

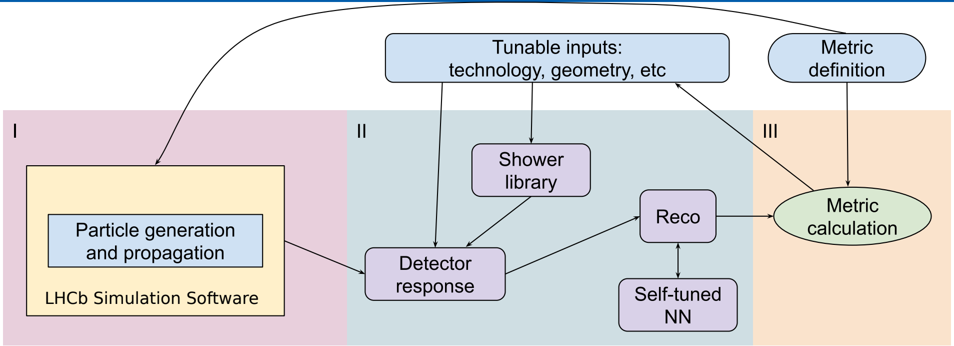
Energy resolution for energy estimated using total energy in 3x3 and 5x5 cells are tend to split at increased pile-up

XGBoost demonstrates better energy resolution compared to energy estimation using total energy in 3x3 cells at energies < 30 GeV

Energy resolution of $26\%/\sqrt{E}$ is unsufficient for physics measurements. The LHCb is studying the possibility of considering time information to mitigate pile-up.



Pipeline global optimisation



- The pipeline is designed in a such way that parameters of it parts can be considered as tunable inputs
- As a result, entire calorimeter can be optimised for given performance metric using any optimisation technique (Random search, Bayesian optimisation, Gradient methods)

Conclusions

- We present the pipeline which can optimise a calorimeter by evaluation of chosen performance metric for any interesting ECAL module technology and configuration, and run optimisation on top of them
- Machine Learning approaches inside the pipeline substitute fine tuning of the parameters at simulation and reconstruction steps of the calorimeter R&D
- The pipeline is able to avoid most CPU-intensive parts of calorimeter full simulation while using the GEANT4 clustering
- The pipeline provides physics performance for arbitrary pile-up conditions
- Spatial and energy resolutions for high pile-up were presented using **current LHCb electromagnetic calorimeter configuration** as an example

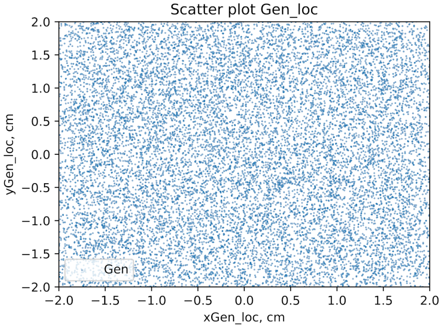
Backup slides

Background composition at ECAL face

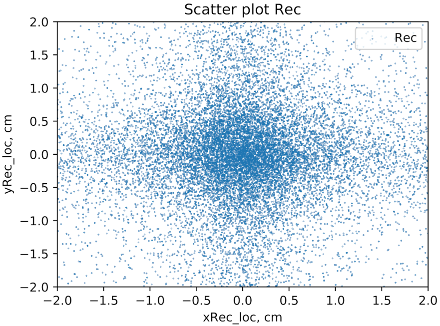
Bkg. particle	% of Total bkg.
γ	23.50
π^+	14.72
π^-	14.17
e^-	12.94
e^+	12.01
n	8.05
‘ p	7.07
K^+	1.28

$>$ 92% of background at z_{ECAL} is from γ, π^+, π^-, e^-, e^+, n, p.

Cluster position determination



Generated hit position (across inner-type cells)

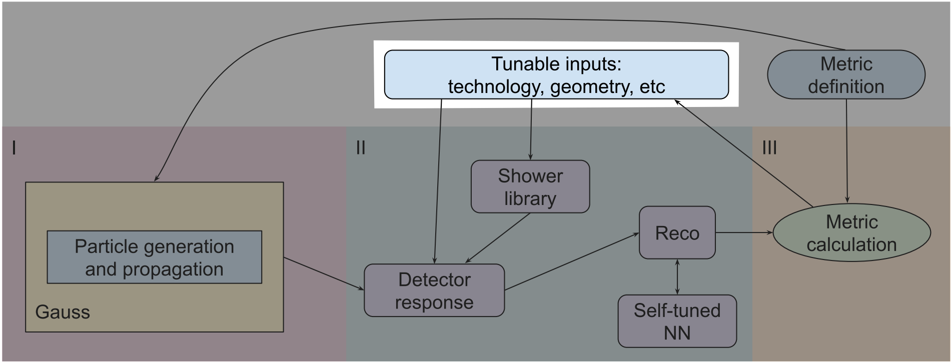


Reconstructed hit position using clusters' barycentre over 5x5 cells

Tunable input: geometry

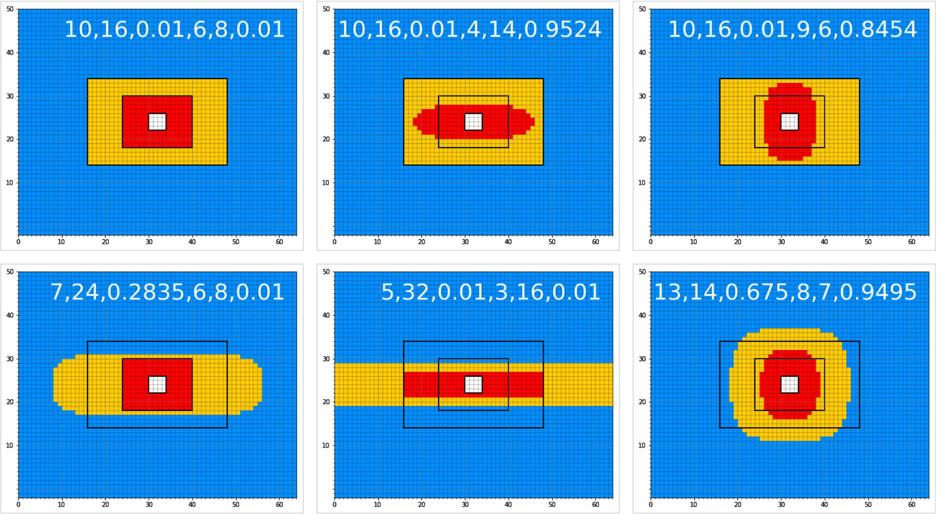
Two options:

- Parametric (convex borders for three regions)
- Custom (arbitrary arrangement of 12.12x12.12 cm^2 modules)



Geometry: examples of parametrically specified configurations

LAMBDA · HSE



A feature engineering

We also tried some combinations of the features:

$$E_{seed} \;/\; (\sum_{i} E_i)^2$$

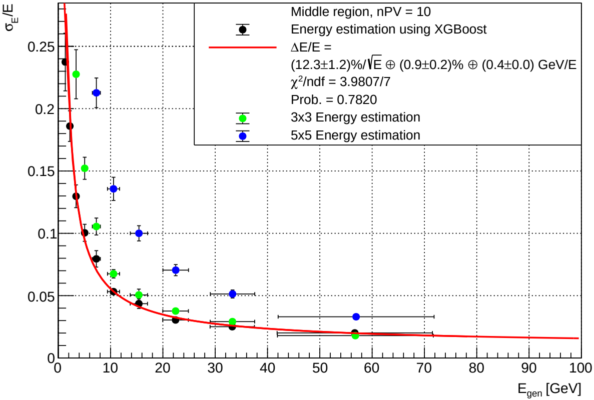
$$E_{seed}^2 \;/\; \sum_{i} E_i$$

$$(\sum_{i} E_i)^2 \;/\; (\sum_{i} E_i)^2$$

etc.

Energy resolution: XGBoost vs. 3x3 vs. 5x5

Middle region, cell size = 6.06 cm
nPV = 10



Energy resolution: XGBoost vs. 3x3 vs. 5x5

Outer region, cell size = 12.12 cm
nPV = 10