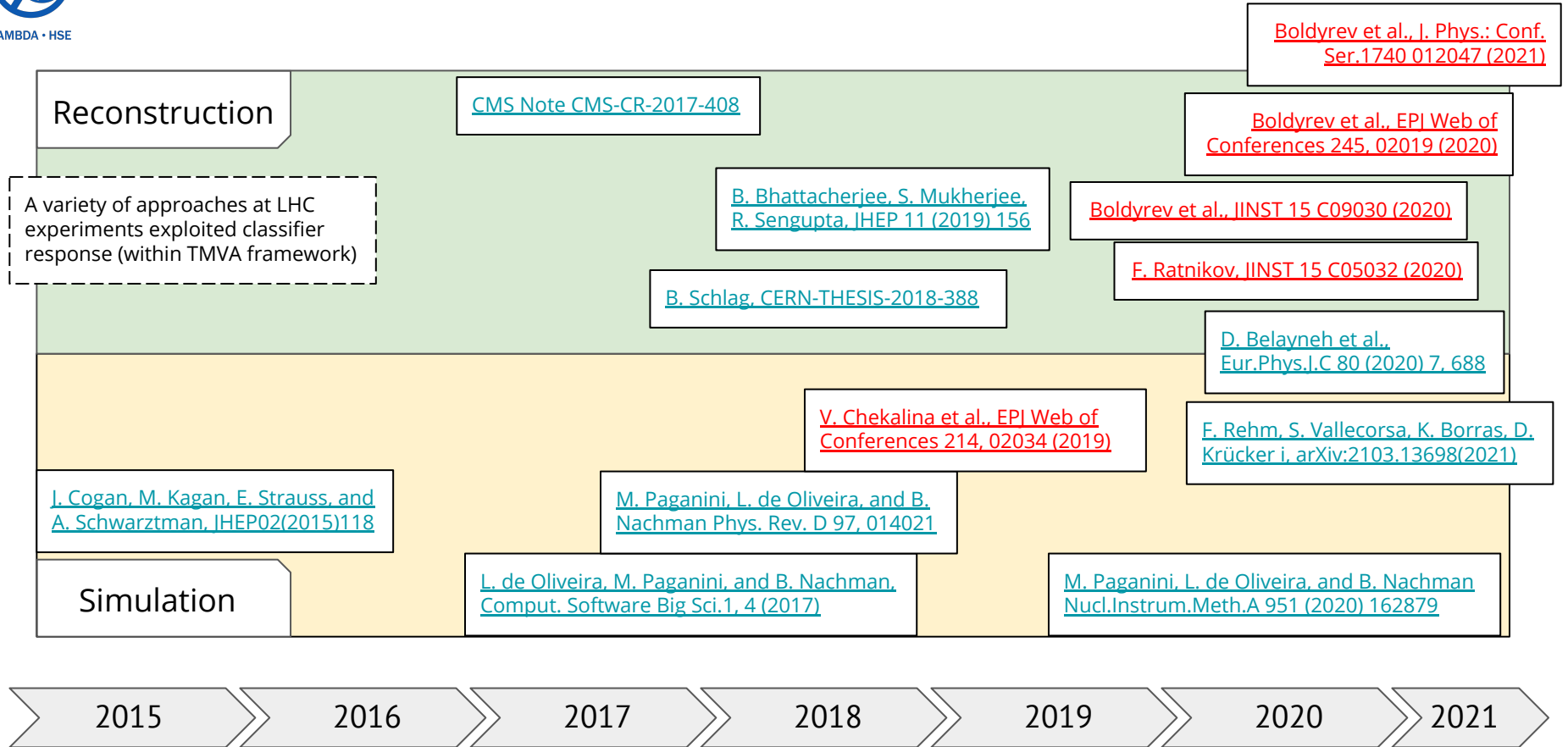


First MODE Workshop on Differentiable Programming, 7 September 2021

Alexey Boldyrev¹, Denis Derkach¹, Pavel Fakanov¹,
Fedor Ratnikov^{1,2}, Andrey Shevelev¹

See also:
[1] [JINST 15 C05032 \(2020\)](#)
[2] [JINST 15 C09030 \(2020\)](#)
[3] [EPI Web of Conferences 245, 02019 \(2020\)](#)
[4] [J. Phys.: Conf. Ser.1740 012047 \(2021\)](#)

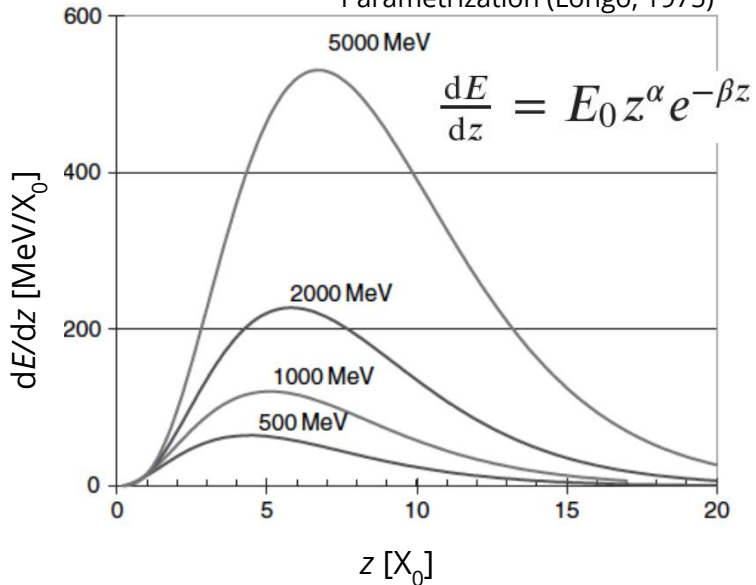
ML in calorimetry



Calorimetry in a nutshell

Longitudinal EM shower profile

Parametrization (Longo, 1975)

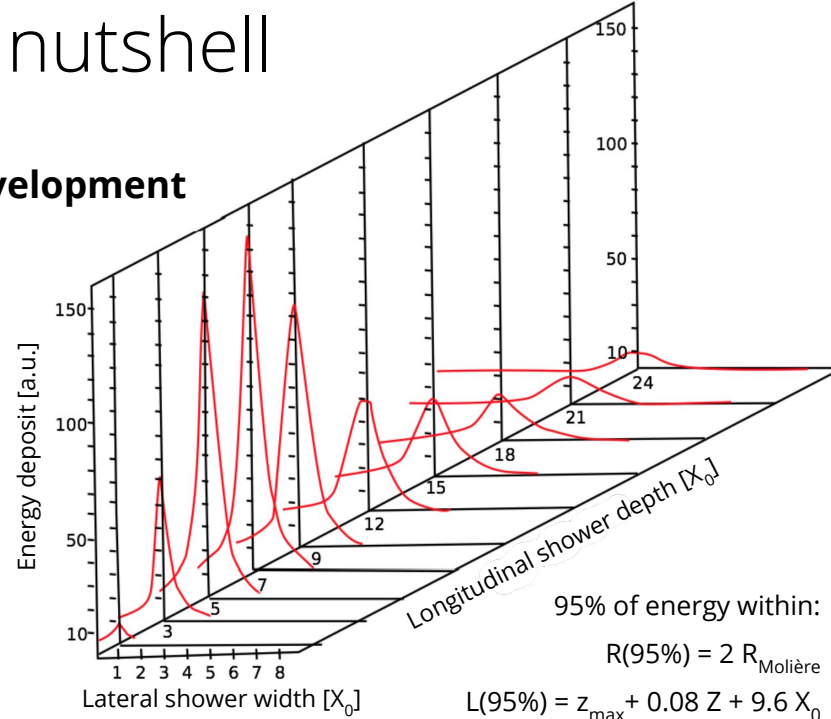


Differences between showers induced by γ & e

$$z_{\max} = \frac{\alpha - 1}{\beta} = \ln\left(\frac{E_0}{E_C}\right) + C$$

$$C_\gamma = -0.5, C_e = -1.0$$

EM Shower development



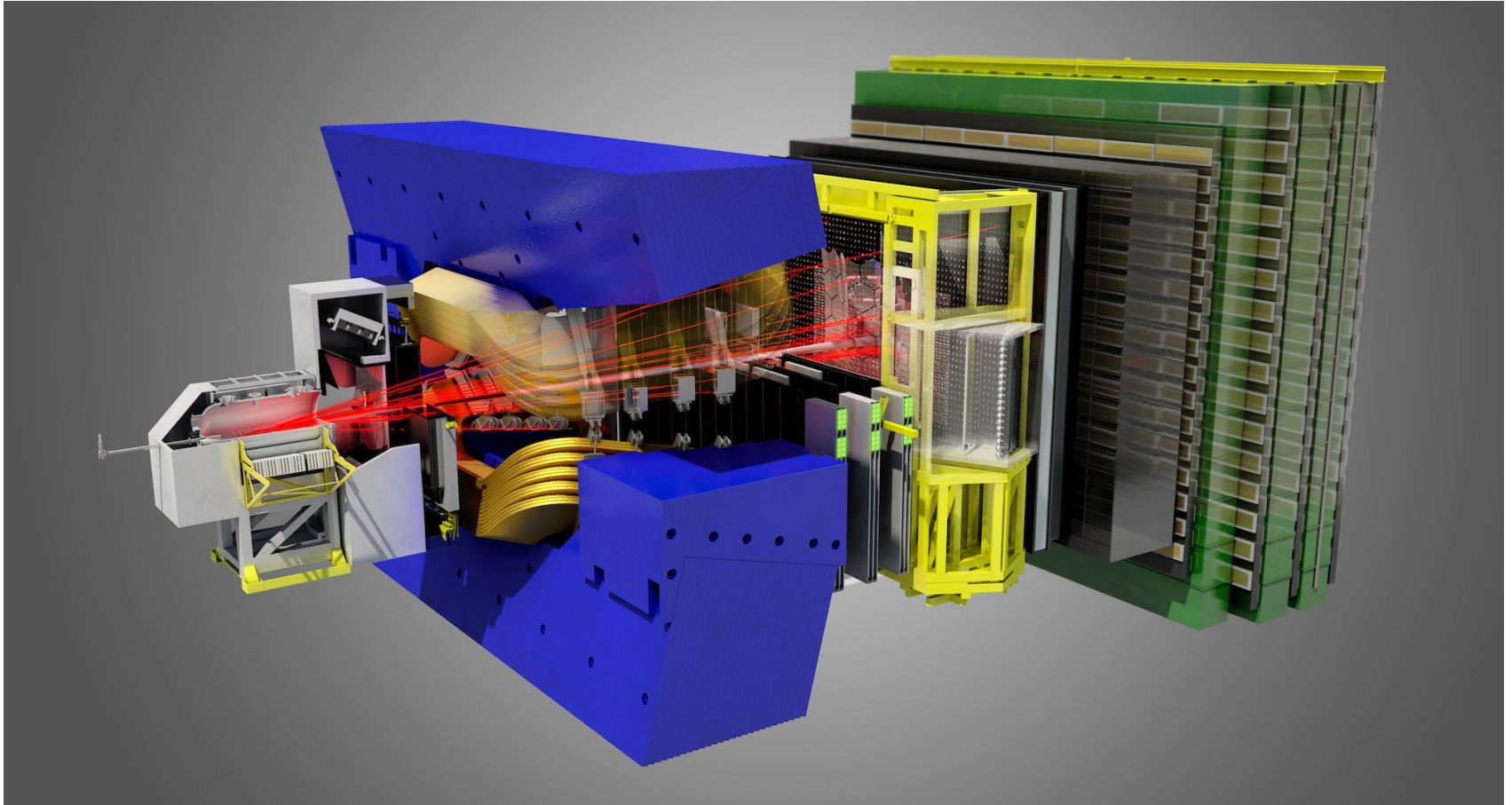
Energy resolution

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

Timing resolution

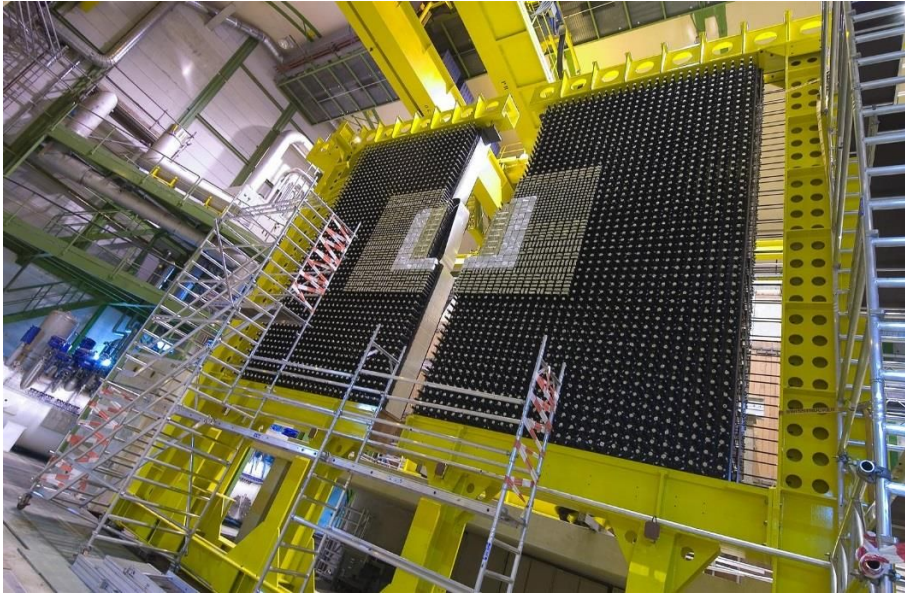
$$\sigma_t = A/\sqrt{E} \oplus B$$

LHCb detector



LHCb ECAL

Current configuration

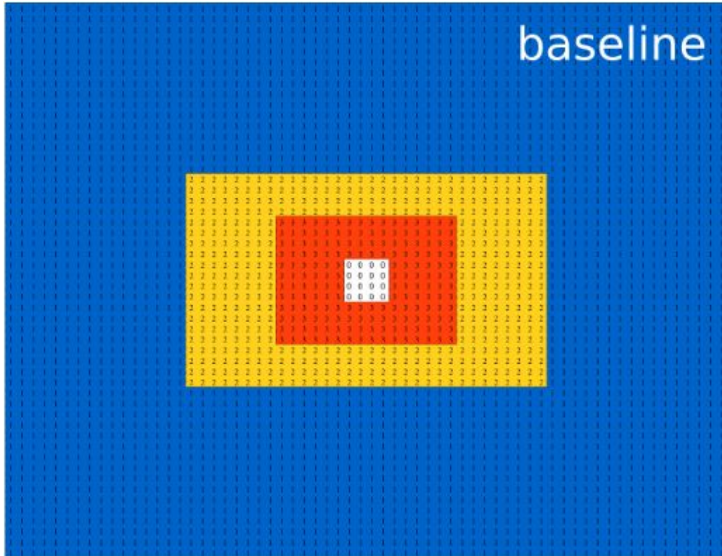





Size: 7.8x6.3x0.5 m



Module size 12x12 cm²

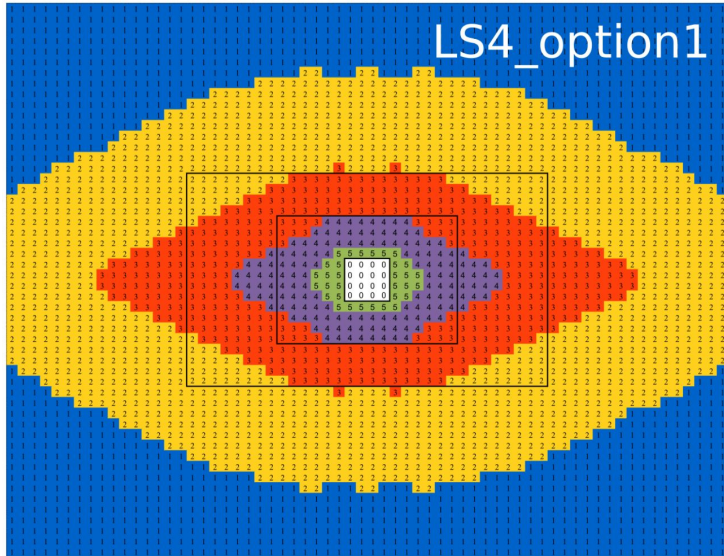
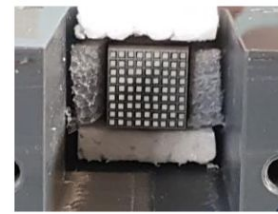
176 inner modules: 9 cells with size 4x4 cm²
448 middle modules: 4 cells with size 6x6 cm²
2688 outer modules: 1 cell with size 12x12 cm²



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.)

Starting from current configuration

Future LHCb ECAL



Reuse of current “Shashlik” modules

- 1 : Outer region, cell size = $12.12 \times 12.12 \text{ cm}^2$
- 2 : Middle region, cell size = $6.06 \times 6.06 \text{ cm}^2$
- 3 : Inner region, cell size = $4.04 \times 4.04 \text{ cm}^2$

New “Spacal” modules

- 4 : cell size = $3.03 \times 3.03 \text{ cm}^2$
- 5 : cell size = $1.515 \times 1.515 \text{ cm}^2$
(+ longitudinal split)



Questions for future ECAL:

- What is the best configuration for given modules (fix cost) in terms of given physics metric?
- What is the best way to arrange a certain number of new modules?



Producing samples & responses in Geant4

Reference physics

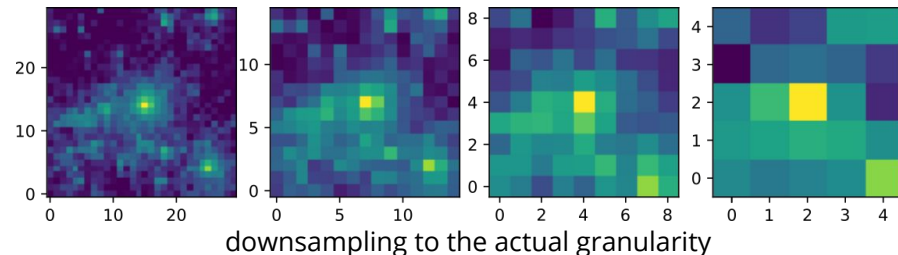
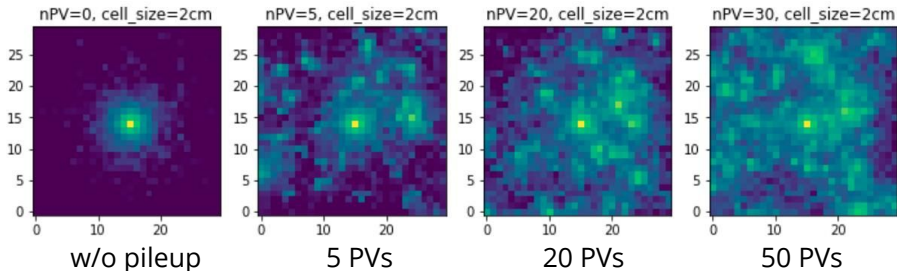
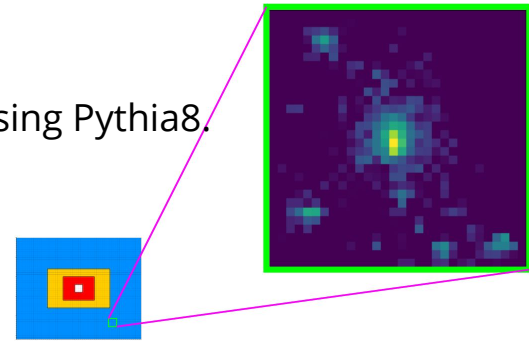
sample: $B_s^0 \rightarrow J/\psi(\rightarrow \mu^+\mu^-)\pi^0(\rightarrow \gamma\gamma)$ Signal events are generated using Pythia8.

Background sample: LHCb Upgrade MC Minimum Bias sample

We consider background contributions from $\gamma, \pi^+, \pi^-, e^-, e^+, n, p$
For each of the signal/background particle we:

- Record type, momentum, hit position and time at the front of the ECAL
- Perform Geant4 standalone simulation of clusters in $N*N(*66)$ cells(*layers) using the momentum & type as input

Thus, we have the **library** of the mapping of particle (px, py, pz, type) and its electromagnetic cluster.
In future, there is possibility to use GAN-generated library.



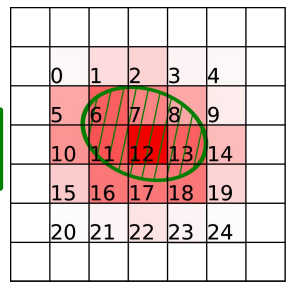
Energy and Time information for ML reco

- MC particle is propagated to ECAL front plane using common LHCb frameworks
 - Hit position, momentum, time and type recorded
- MC time is corrected by angle to compare simultaneous events in the calorimeter
- Simulated Geant4 response is the 5x5(x2) array of cells
 - Used as base features for Spatial and Energy regressors
 - Time regressor uses weighted energy deposits

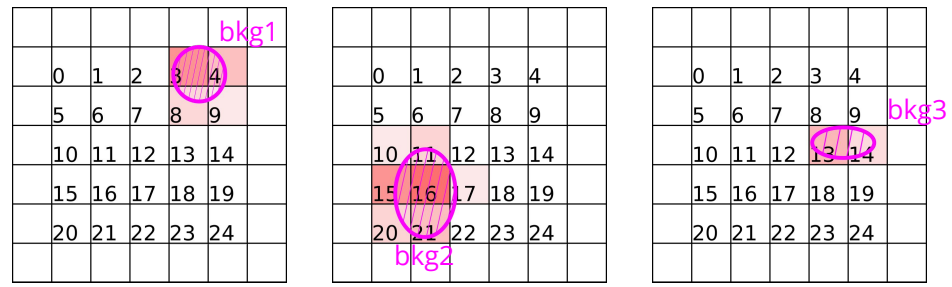
Time information: approach

Suppose that there are 3 background contributions in the 5x5 cells vicinity of the seed cell and $t^{bkg1} < t^{sig} < t^{bkg2} < t^{bkg3}$.

Signal energy deposits and shower spot

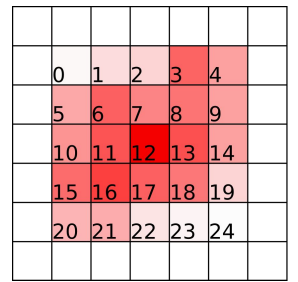


Background energy deposits and showers' spots



Resulting energy deposits:

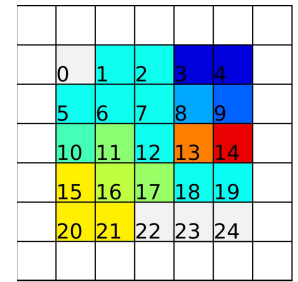
$$E_i = E_i^{sig} + E_i^{bkg1} + E_i^{bkg2} + E_i^{bkg3}$$



(Used as raw features for position and energy regressors)

Resulting time of cells:

$$t_i = \frac{\sum_{i,j} t_i^j E_i^j}{\sum_{i,j} E_i^j}$$



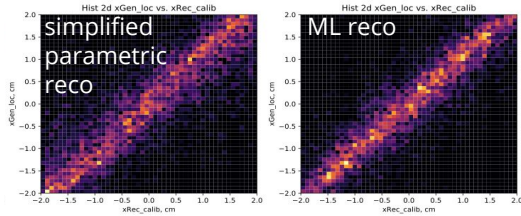
Calculated cell time

MC time t^{bkg1} t^{sig} t^{bkg2} t^{bkg3}

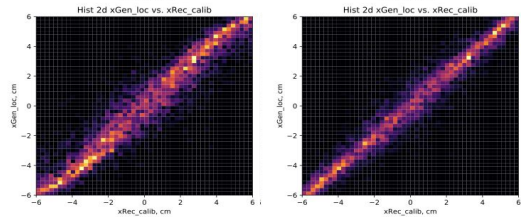
Spatial, Energy & Time reconstruction

Spatial reconstruction

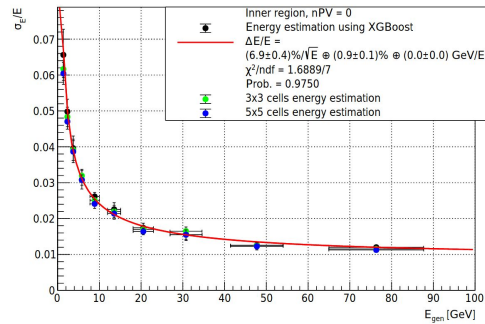
module
4x4 cm²



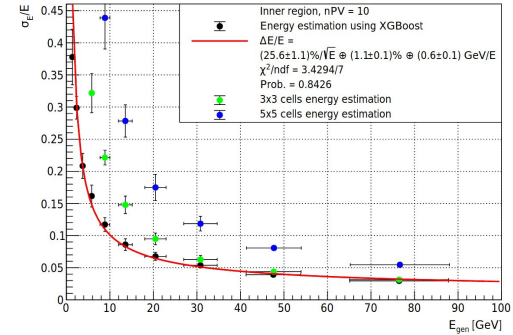
module
12x12 cm²



Energy reconstruction

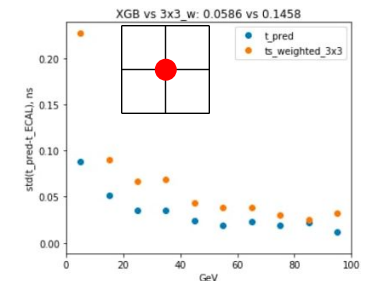
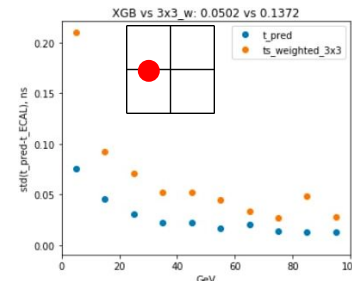
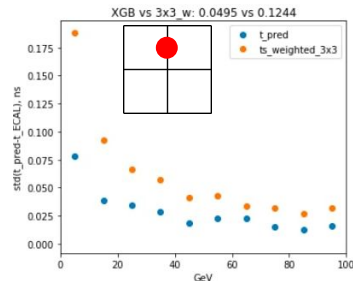
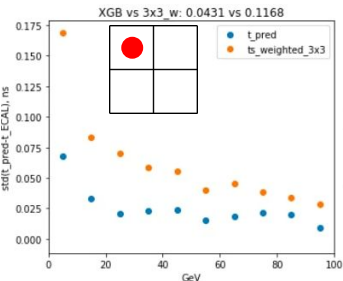


W/o pile-up the energy resolution is consistent with LHCb ECAL design



At **increased pile-up** ML reco still shows meaningful estimation

Time reconstruction

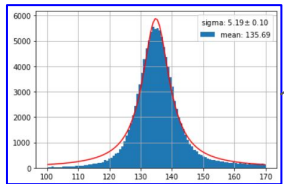
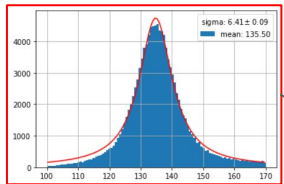
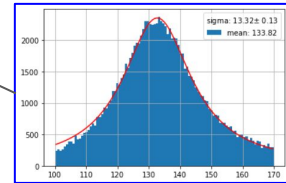
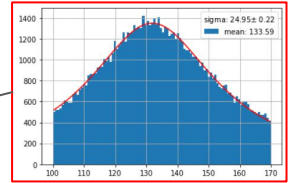
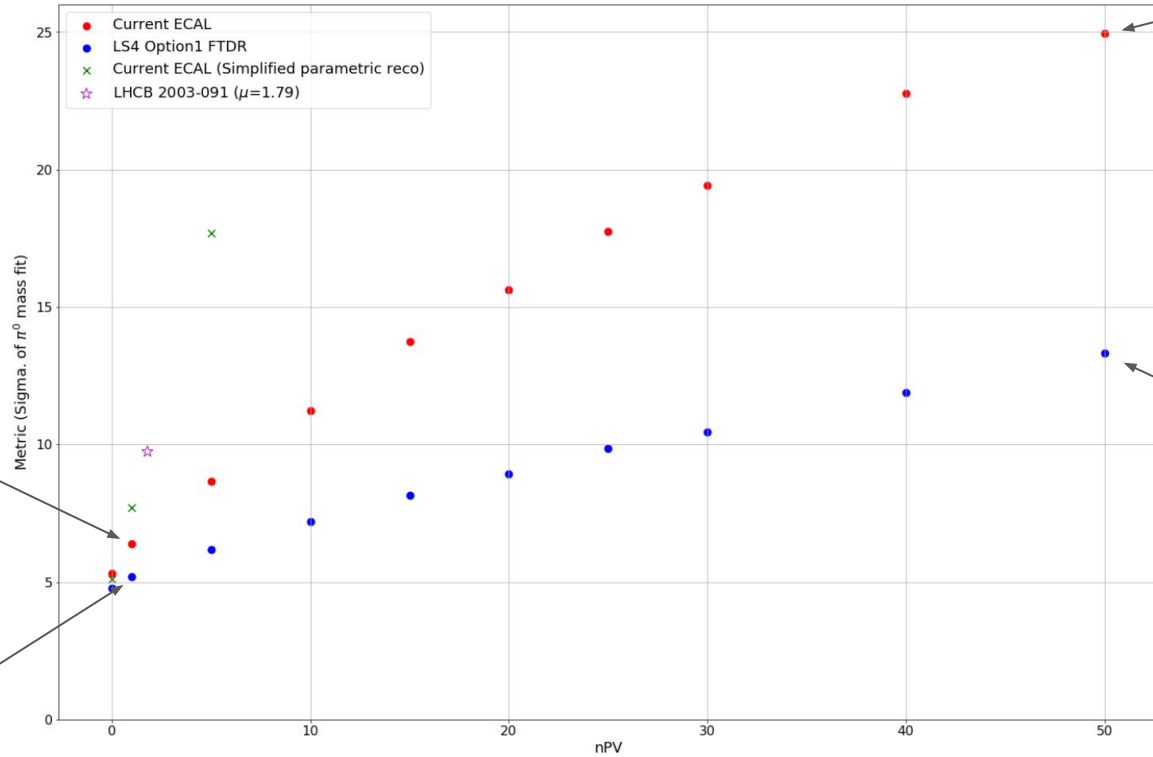




Reconstructed π^0 width

Two photons reconstructed:

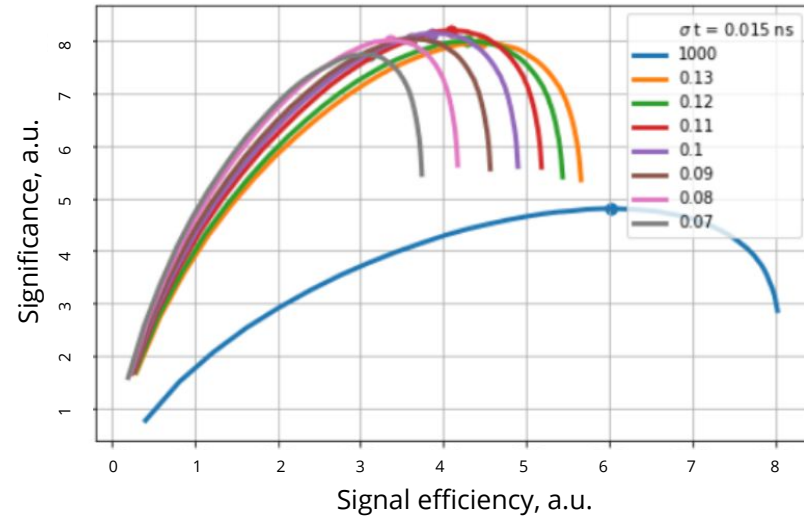
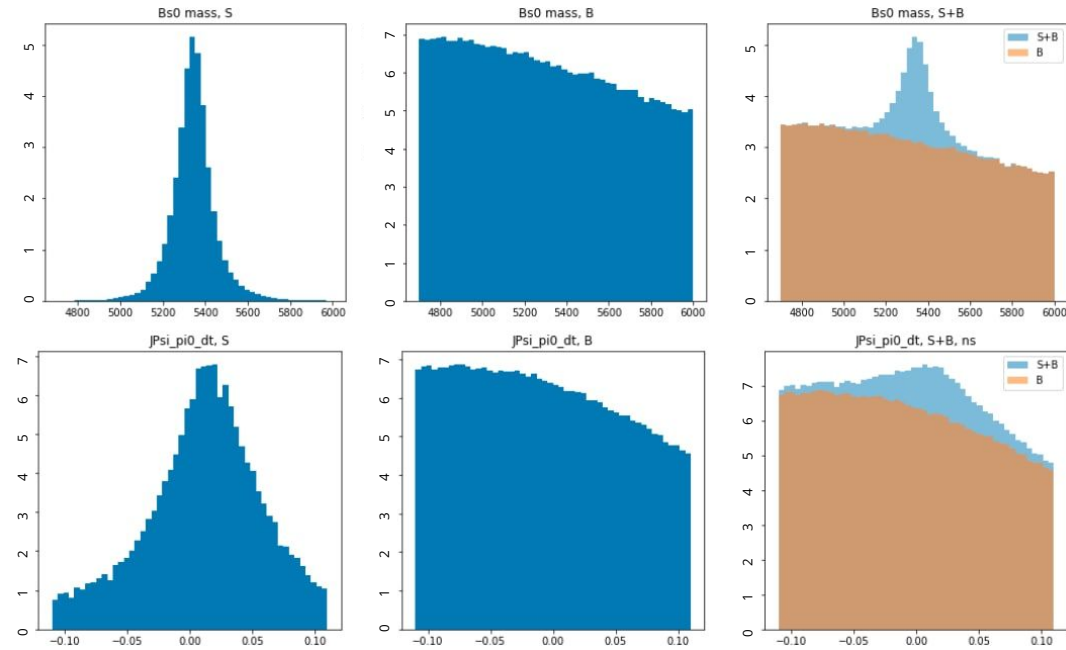
- Energy from regressor
- Position from regressor
- No timing



B_s reconstruction

$J/\psi \pi^0$ candidates reconstructed using reco'd π^0 and MC J/ψ .
Time regressor is used to define time window.

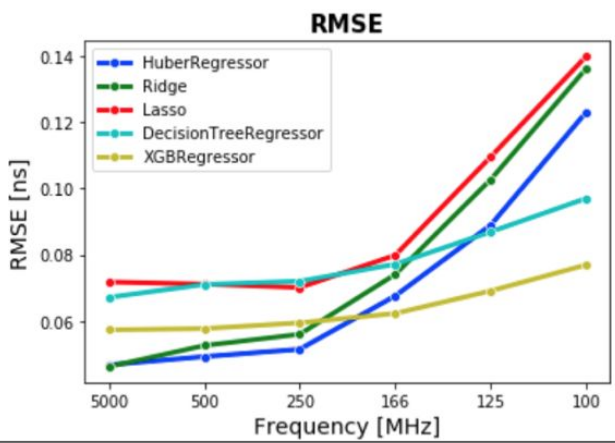
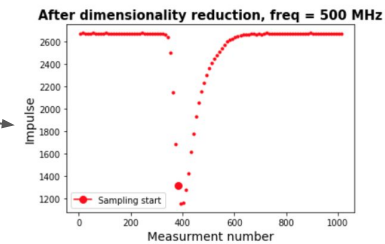
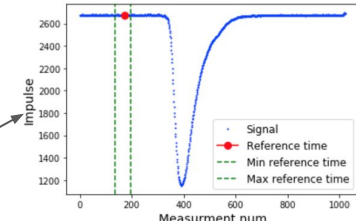
For the selected time window, B_s mass window is optimised by finding the maximum of significance.



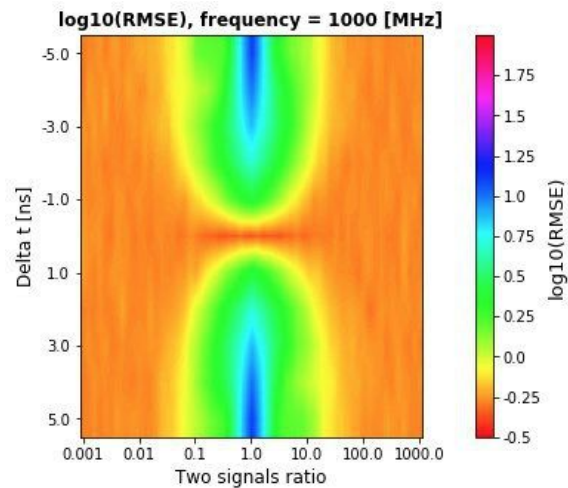


Time reco for two signals discrimination

- Use signal readout data for the prototype exposed on the **test beam**
- Use scintillator signal as a reference time
- Emulate different sampling rates by selecting readout points
- Use different regressors to verify consistency of the result
- Train regressor to extract timing for the bigger signal in presence of the second signal



Two signals discriminating



Accounting possible options

At the moment we have:

- Thousands of configurations
- Module technology options
- Longitudinal segmentation option
- Timing information

How to rule them all?



Black-box optimisation problem

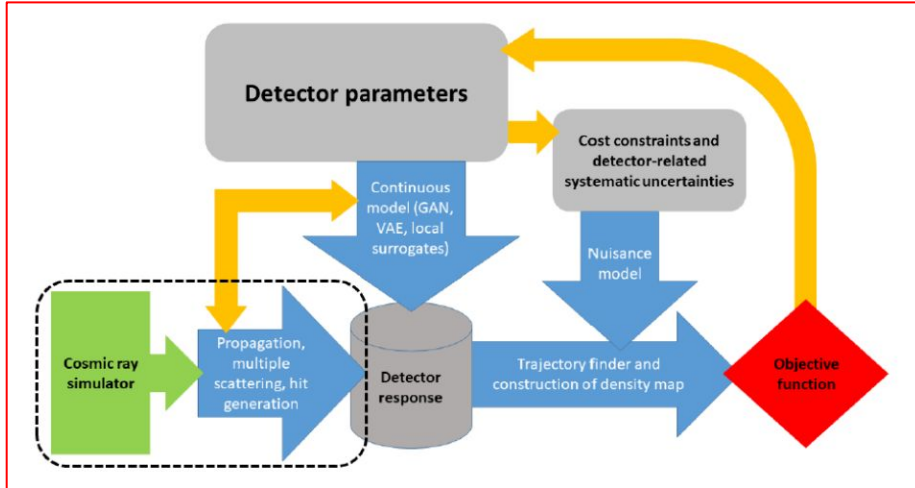
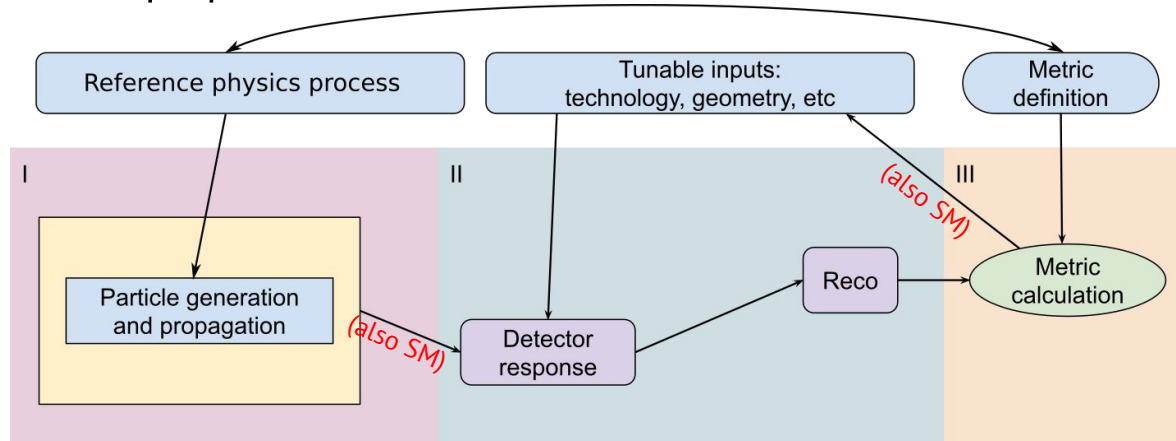
We need to have the number of calls of the function to be optimised as low as possible.

Two main ingredients:

- Surrogate model
 - approximates the true function
 - cheap to evaluate
 - in general, any regression can be chosen, with preference to that returning variance of prediction
- Acquisition function
 - estimates profit for optimisation
 - uses surrogate model

The pipelines

The conceptual layouts of the pipelines are consistent each other (apart from Reco for ECAL).



Main limitations for differentiability of the metric as a function of tunable inputs for **current realisation** of the ECAL ML pipeline are from:

- Geant4 simulation of the detector response
- XGBoost model (but NN can be differentiable)
- Piecewise-constant function of modules position

Surrogate modelling with Gaussian process

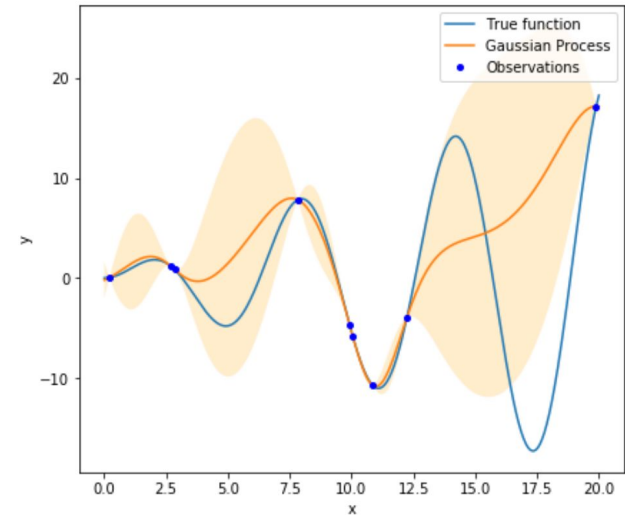
Gaussian process regression is commonly used approach in the surrogate modelling. The main idea: each point in the fitted space is sourced from Gaussian distribution. We thus are able to produce prediction for the next point.

Pros.:

- Predictions include variance

Cons.:

- Computationally expensive, $O(n^3)$

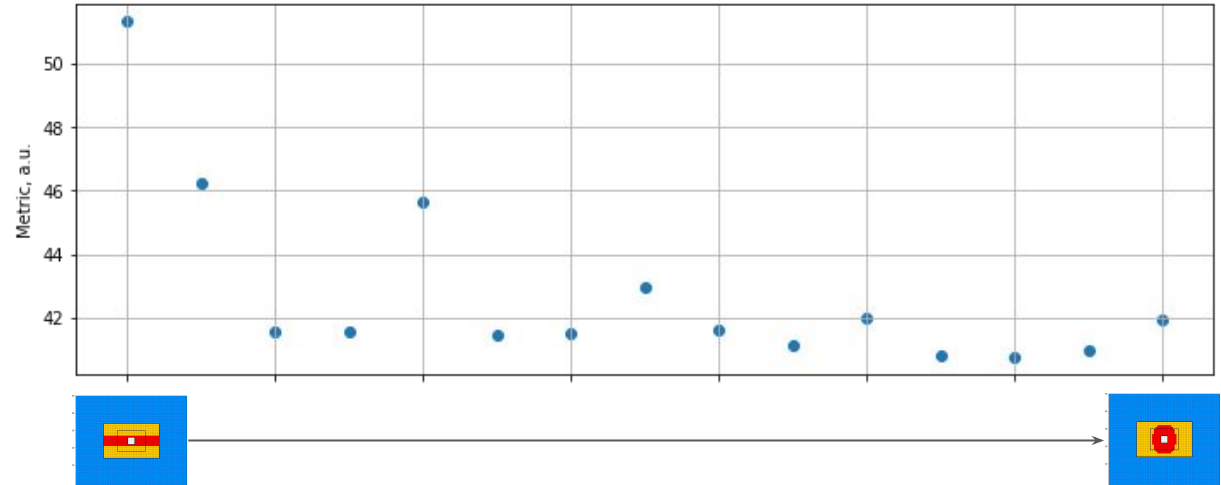


Optimisation cycle

The possible answer how to account all the options of the R&D cycle is to use Bayesian Optimization with Gaussian Processes

The full optimization cycle will look as follows:

1. Construct surrogate model over known history
2. Find the maxima of Expected Improvement algorithm
3. Evaluate suggested point via real physical simulation
4. Add point to history
5. Repeat



More information in A. Filatov's [talk @ICPPA meeting](#) and [proceedings](#) (SHiP shield optimization).

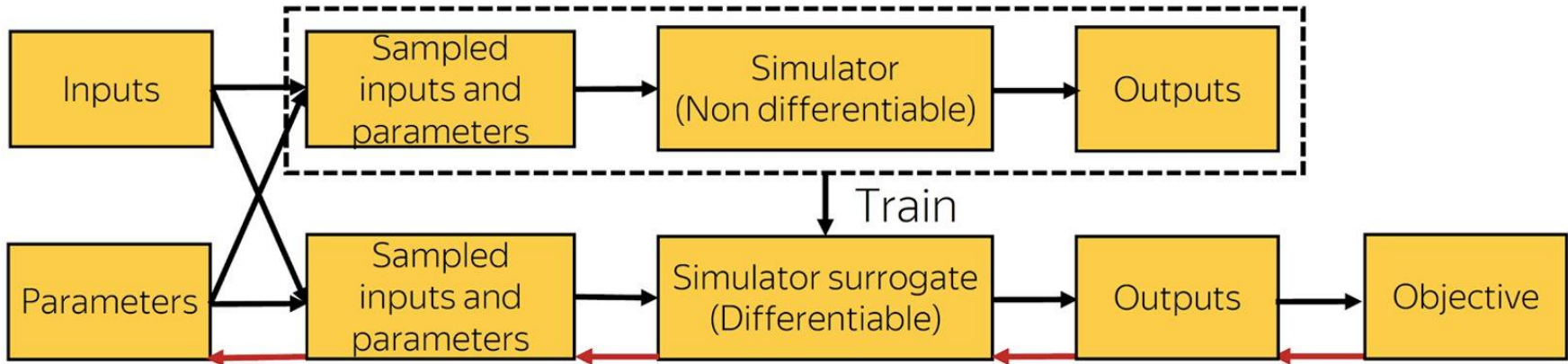
Black-box optimization with Local Generative Surrogates

LAMBDA · HSE

NN-based alternative to Bayesian optimization with Gaussian Processes.

Let's approximate a stochastic black-box with a local generative surrogate.

This allows computing gradients of the objective w.r.t. parameters of the black-box.



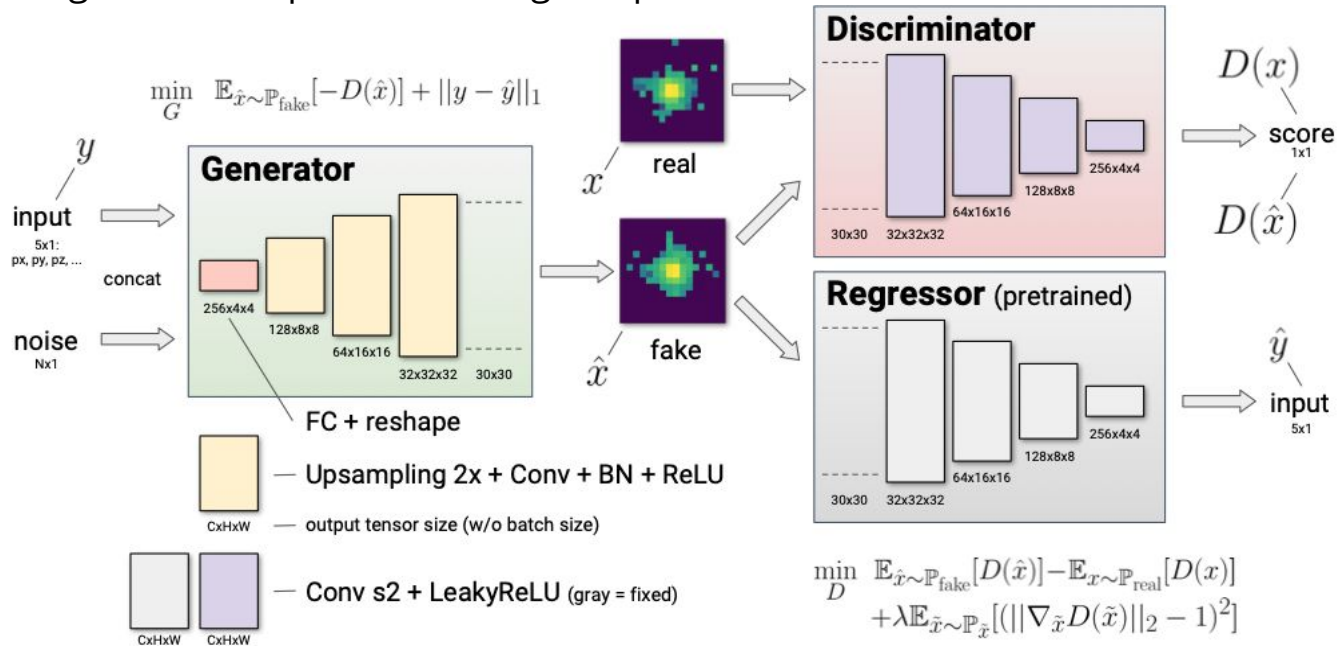
Shirobokov S., Belavin V., Kagan M, AU, Baydin A., NeurIPS'20 paper, [arXiv:2002.04632 \[cs.LG\]](https://arxiv.org/abs/2002.04632)

Conclusions

- The R&D process requires time consuming computation steps to evaluate physics performance for different detector techniques and configurations.
- ML reco is consistent with common reconstruction and don't need to be fine-tuned.
- Surrogate ML models may be used for most steps that are necessary for evaluating quality of different solutions. Such models are automatically trained on available datasets and provide possibility to consistently estimate the resulting physics performance.
- Using automatic training speeds up the turnover for the performance studies and ensures consistency and uniformity of obtained results.

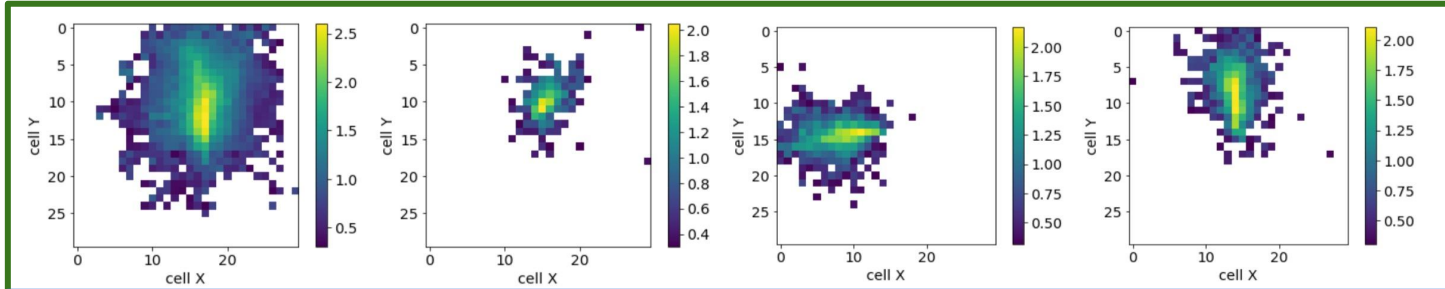
Backup slides

- Collect GEANT responses for the calorimeter technology of track parameters in standalone setup
- Train conditional generative model on simulated data
- Use the model to generate response for the given particle

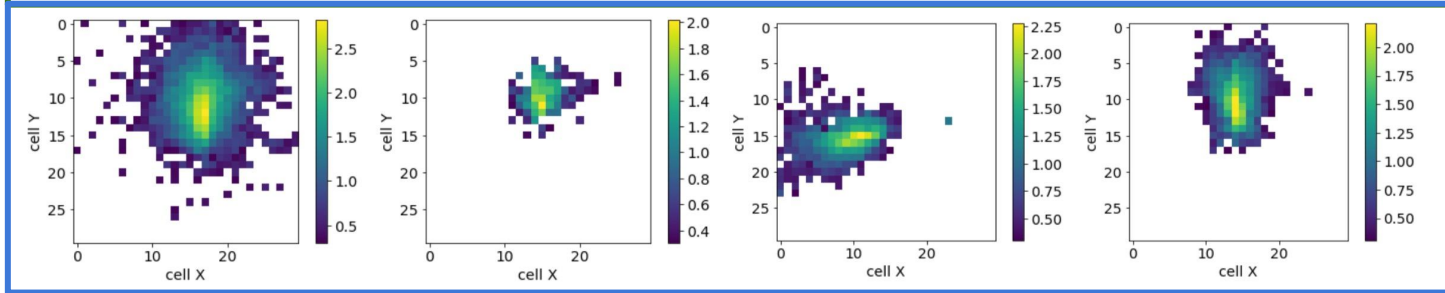


- Collect GEANT responses for the calorimeter technology of track parameters in standalone setup
- Train conditional generative model on simulated data
- Use the model to generate response for the given particle

GEANT4



GAN



(a)

$E_0 = 63.7 \text{ GeV}$

(b)

$E_0 = 6.5 \text{ GeV}$

(c)

$E_0 = 15.6 \text{ GeV}$

(d)

$E_0 = 15.9 \text{ GeV}$