



NATIONAL RESEARCH
UNIVERSITY

Using Machine Learning to Speed Up and Improve Detector R&D

4th Inter-experimental Machine Learning Workshop, 21 October 2020

Alexey Boldyrev¹, Denis Derkach¹, Pavel Fakanov¹,
Leonid Matyushin¹, Fedor Ratnikov^{1,2}, Andrey Shevelev¹

See also:

[A. Boldyrev et al 2020 JINST 15 C09030](#)
and [arXiv:2003.05118 \[physics.ins-det\]](#)

The idea

To choose optimal design of the detector:

- Cover entire R&D cycle, whenever possible
- Define a metric
- Realise each step of the R&D cycle from first principles (or use computationally cheap yet reliable alternative)
- Build a pipeline on top of them
- Evaluate the importance of each step of R&D cycle

The parameters in scope

We can solve the problem how to arrange sensitive elements (modules) of the detector in effective and generalised way if we know the following:

- Accurately simulated responses of given modules technology
- Cost of that technology
- Metrics obtained from reference physics processes within required pileup conditions

We aim to optimize physics performance metrics / overall cost

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

Surrogate modelling with Gaussian process

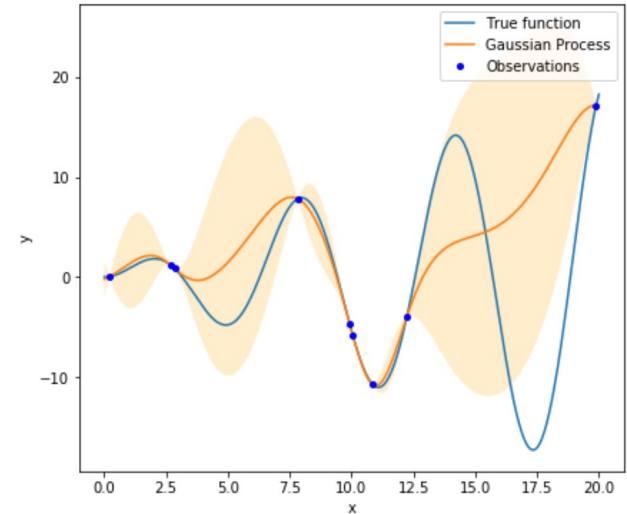
Gaussian process regression is commonly used approach in the surrogate modelling. The main idea: each point in the fitted space is sourced from Gaussian distribution. We thus are able to produce prediction for the next point.

Pros.:

- Predictions include variance

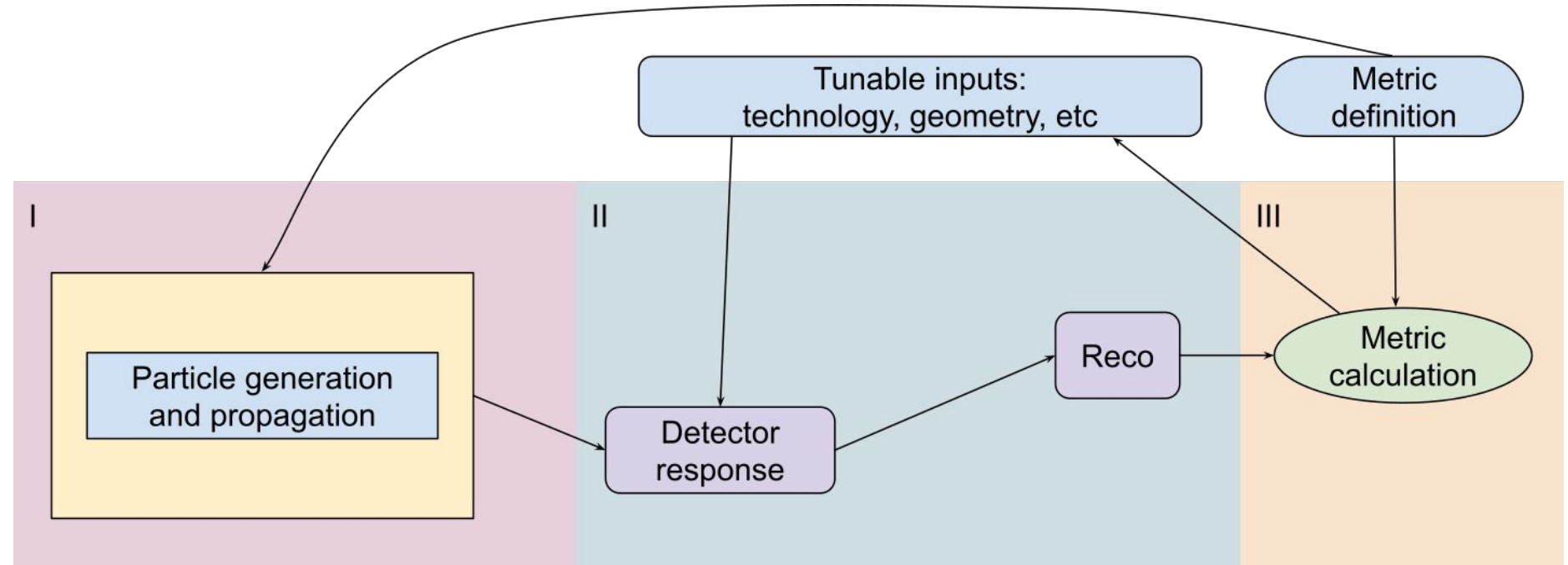
Cons.:

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



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

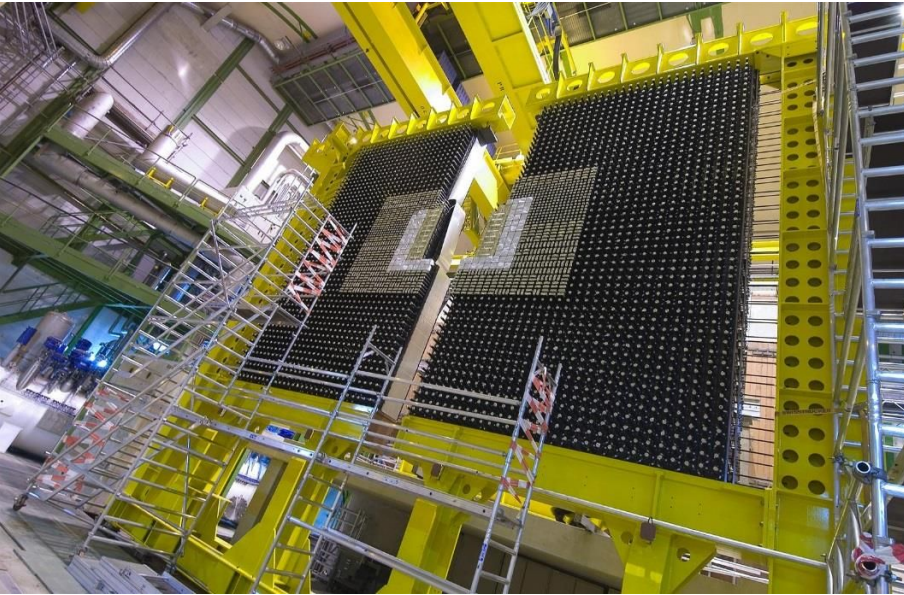
The pipeline



Optimisation cycle itself does not depend on the modules technology & arrangement, reconstruction, metric, etc.

LHCb ECAL

Current configuration



Size: 7.8x6.3x0.5 m

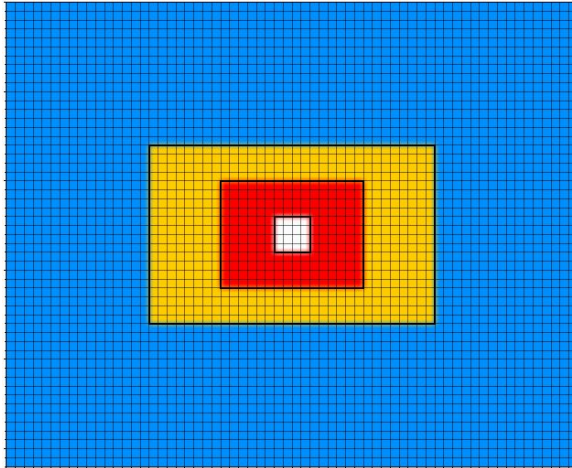





Module size 12x12 cm²

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

Future LHCb ECAL

Starting from current configuration

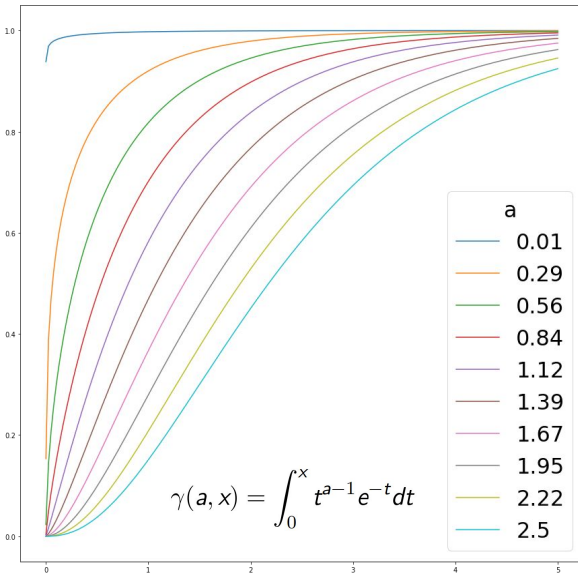


Module type	# of modules
 (inner): 3x3 cells (4.04x4.04 cm ² each)	176 (1536 ch.)
 (middle): 2x2 cells (6.06x6.06 cm ² each)	448 (1792 ch.)
 (outer): single cell (12.12x12.12 cm ²)	2688 (2688 ch.)

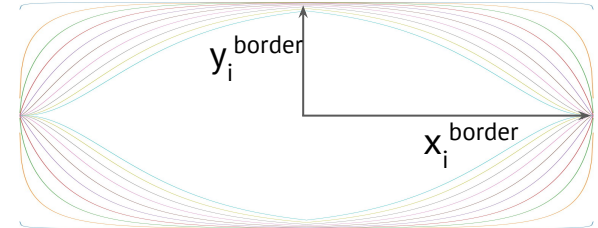
- What is the best configuration for given modules (fix cost) in terms of given physics metric?
- What is the best way to arrange a certain number of new modules?

Defining region border

To describe the borders between regions of modules of same type we choose incomplete gamma functions of real variable x (Young tableaux is plan B) :

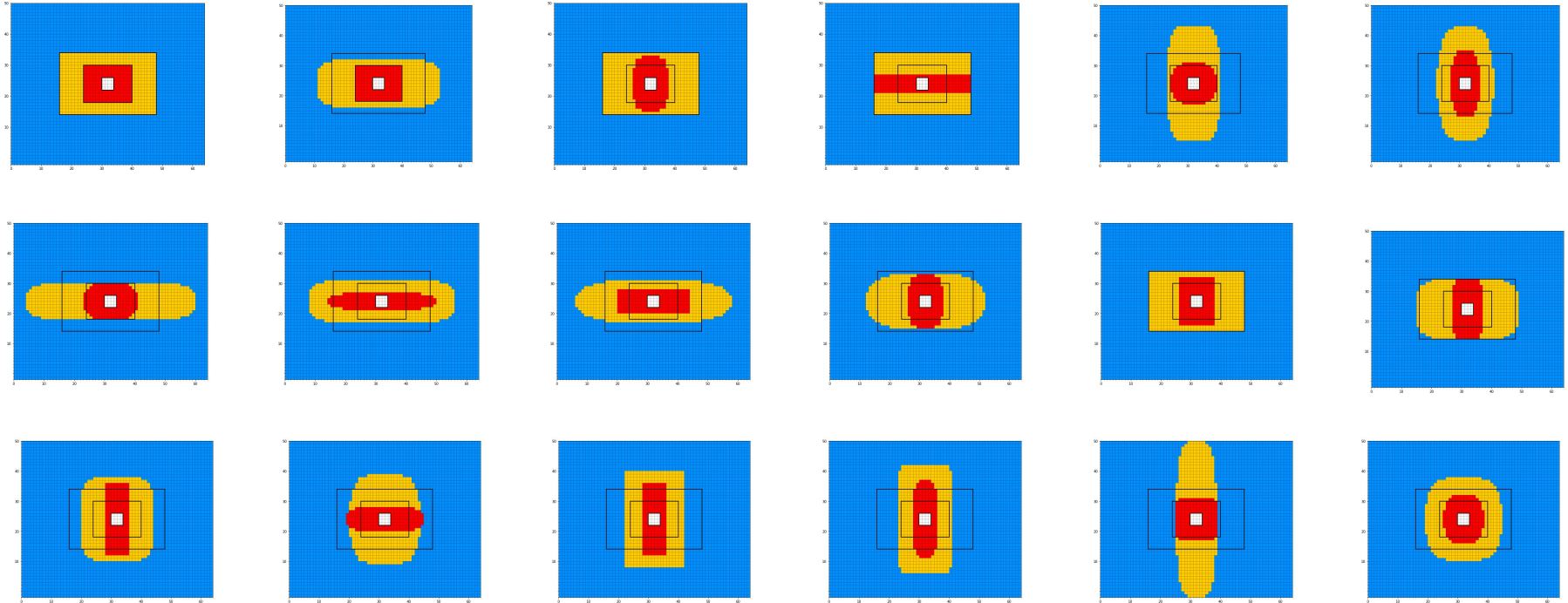


- We consider ECAL to be symmetrical over X and Y axes. Therefore two reflections of such border function are needed:



- Border function is sampled on discrete space of modules
- For current LHCb ECAL-like configurations we have 2 borders of 3 parameters each
- There are 31712 non-trivial arrangements

LHCb ECAL-like configurations



...

Total: 702 configurations with non-overlapping borders.

Producing samples & responses in G4

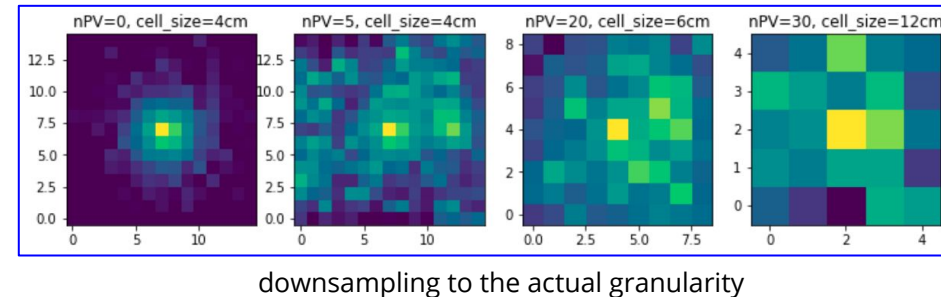
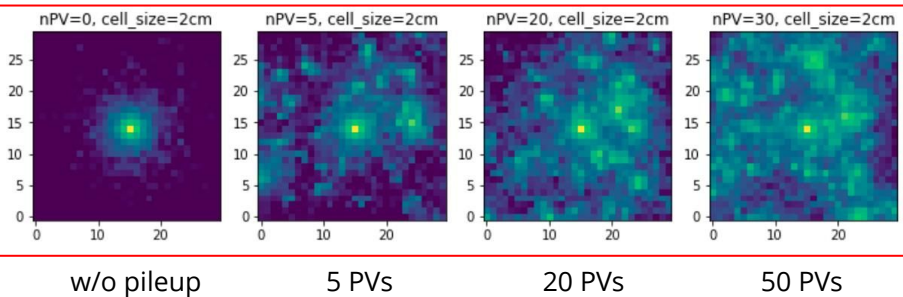
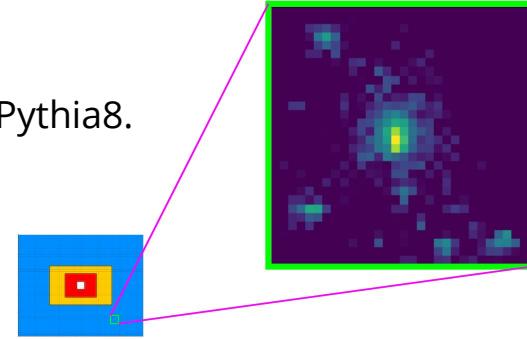
Signal sample: $B_s^0 \rightarrow J/\psi(\rightarrow \mu^+\mu^-)\pi^0(\rightarrow \gamma\gamma)$ Signal events are generated using Pythia8.

Background sample: LHCb Upgrade Minimum Bias sample

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

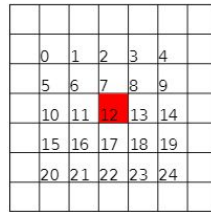
- Record type, momentum, hit position and time at the front of the ECAL
- Perform Geant4 standalone simulation of clusters in $N*N(*66)$ cells(*layers) using the momentum & type as input

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

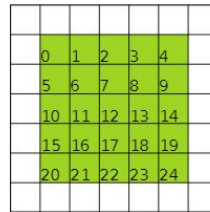


ML-based reconstruction

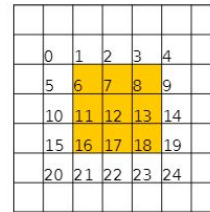
Then we're looking for the cell contained maximum energy deposit (seed).



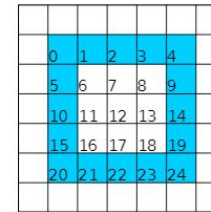
$$E_{seed} \text{ or } E_{seed}^2$$



$$\sum_i E_i \text{ or } (\sum_i E_i)^2$$



$$\sum_i E_i \text{ or } (\sum_i E_i)^2$$



$$(\sum_i E_i)^2$$

Two regressors allows us to reconstruct the π^0 :

- XGBoost for Energy reconstruction
- XGBoost for Spatial reconstruction

Chosen performance metric is the width of the $m_{inv}(\pi^0)$ fit.

Features:

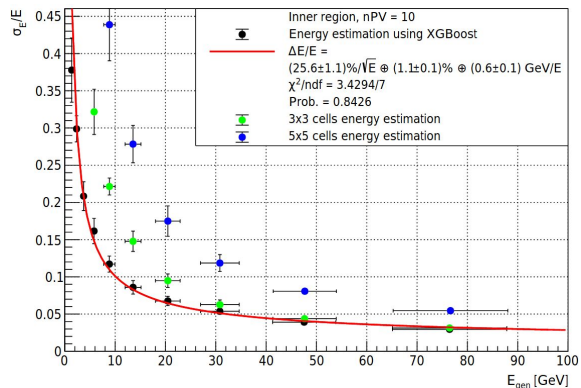
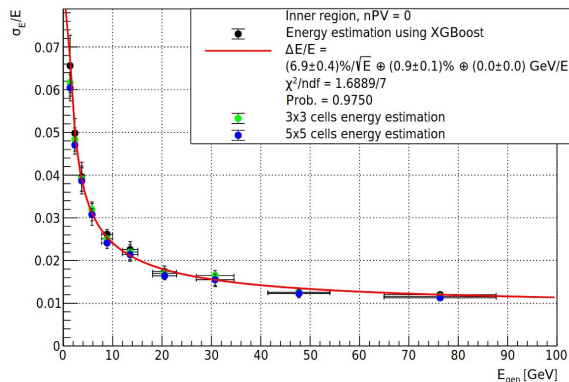
- Energy deposits of 25 cells around seed cell (2*25 in case of Z-split modules)
- Barycenter
- Time information
- Sums, squared sums, rings, etc. of energy deposits



Inner validation of the ML-based reconstruction

Since we have two independent regressors for Energy and Spatial reconstruction, we are able to validate them by calculating the metric using ML Energy reconstruction and ideal spatial resolution (position for MC), and vice versa.

This identified that energy resolution dominates to the metric with increased pile-up:



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

At increased pile-up we're forced to add more channels

Accounting possible options

At the moment we have:

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

How to rule them all?

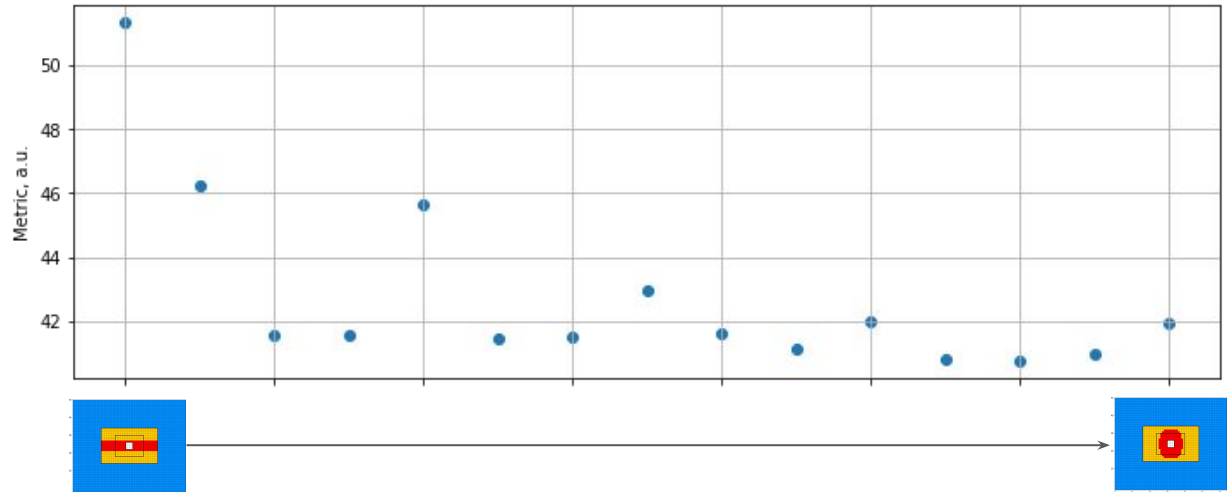


Bayesian optimization (again)

... the answer is:
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 EI
3. Evaluate suggested point via real physical simulation
4. Add point to history
5. Repeat



Conclusions

- The R&D process requires time consuming computation steps to evaluate physics performance for different detector techniques and configurations.
- 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