Pre-thesis internship

Study of hadronic interactions in calorimeter with high granularity

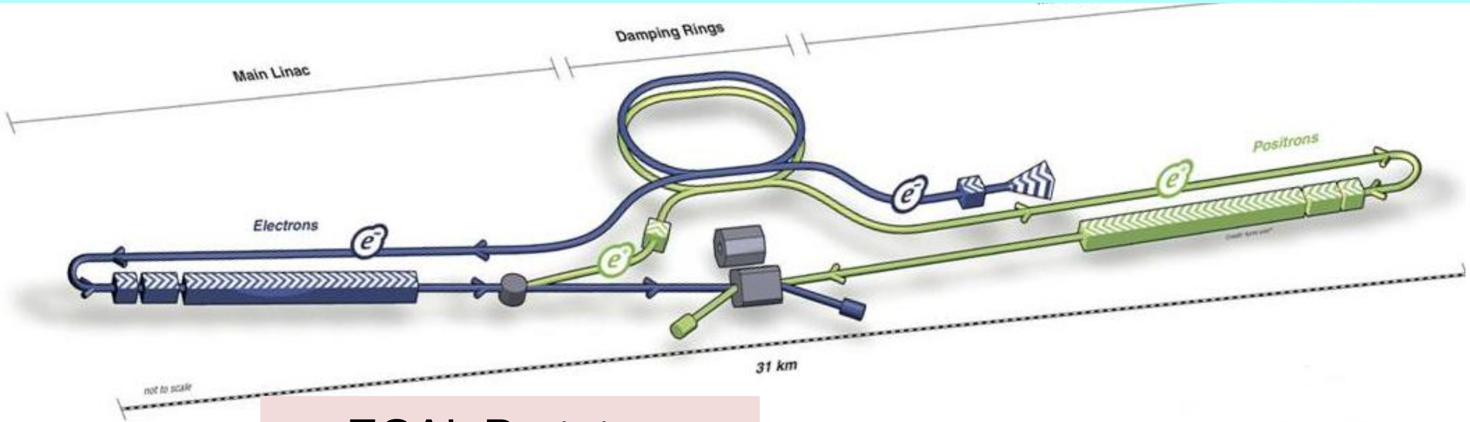
The author :
Yaroslav Nikolaiko

Supervision :
Roman Poeschl and **Naomi van der Kolk**,
LAL, ILC group-CNRS/IN2P3

Content

- Introduction : Future Linear Collider and Particle Flow
- Machine learning as tool for PF Alhorithms improvement
- Hadronic Interaction in Geant4
- Trivial classification of interactions
- Summary

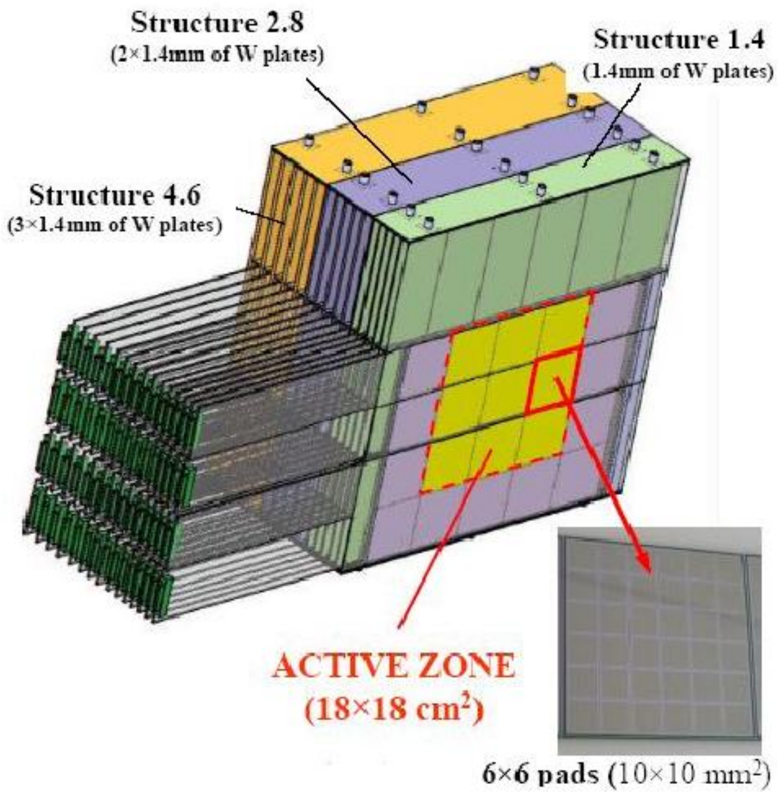
Introduction



ILC

0.5-1 TeV $e^{+}e^{-}$ collider

ECAL Prototype



- Precise study of SM parameters and beyond.
- Requires excellent energy resolution to separate multiple jet events.

- 30 layers sandwich structure (Si – active, W- absorber materials)
- Total depth – 24 X_o or 1 λ_i
- Followed by HCAL with 48 longitudinal samples

Particle Flow

Particle Flow

Energy resolution = Software + Hardware

Software → **Confusion term** $\approx 2\%$

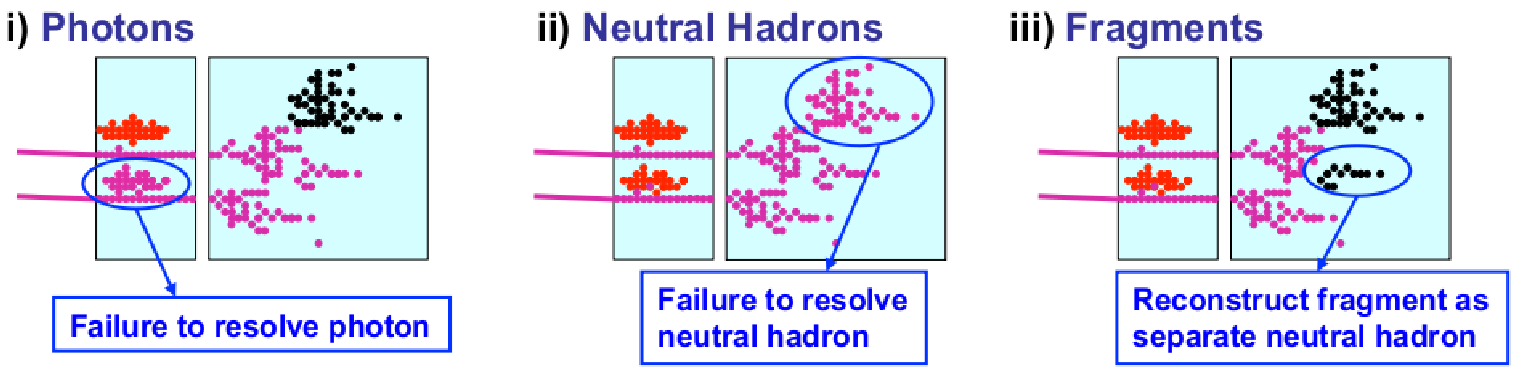
Hardware → Calorimeter resolution $\approx 2\%$

Particle Flow

Energy resolution = Software + Hardware

Confusion term $\approx 2\%$

Calorimeter resolution $\approx 2\%$



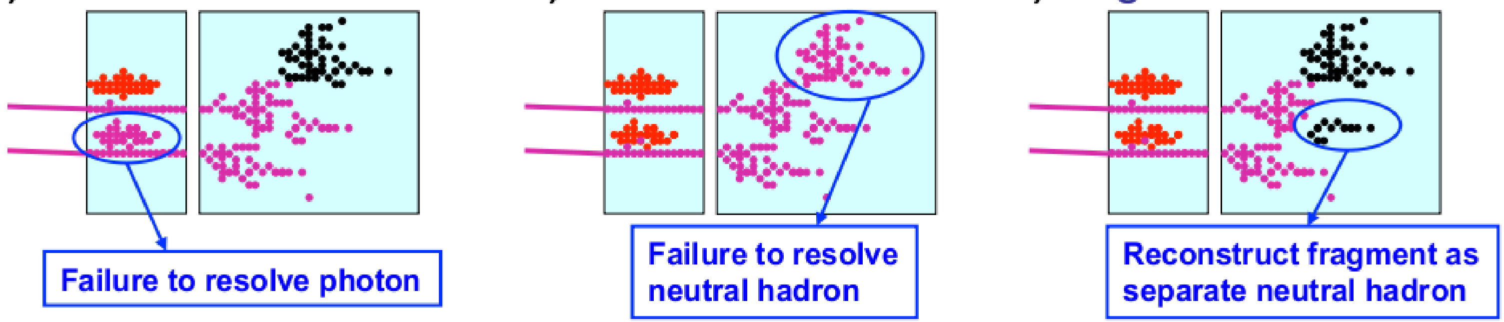
Particle Flow

Energy resolution = Software + Hardware

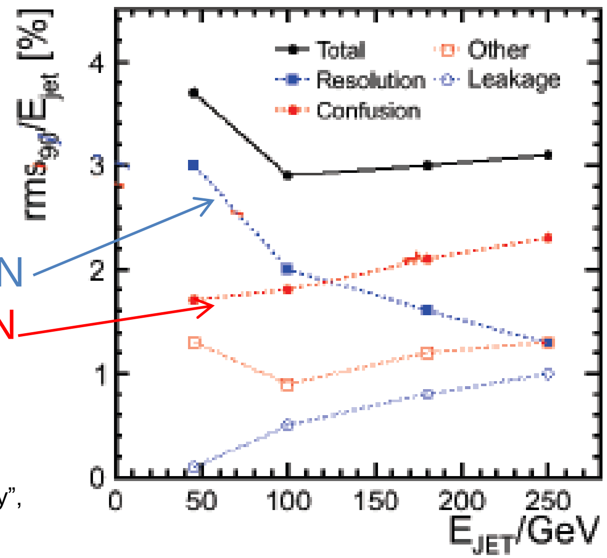
Software → **Confusion term** $\approx 2\%$

Hardware → Calorimeter resolution $\approx 2\%$

i) **Photons** ii) **Neutral Hadrons** iii) **Fragments**



- Contribution to resolution
  - Low energy jets: RESOLUTION
  - High energy jets: CONFUSION



from M. Thomson,"Particle Flow Calorimetry", Mainz, February 2013

PandoraPFA performance

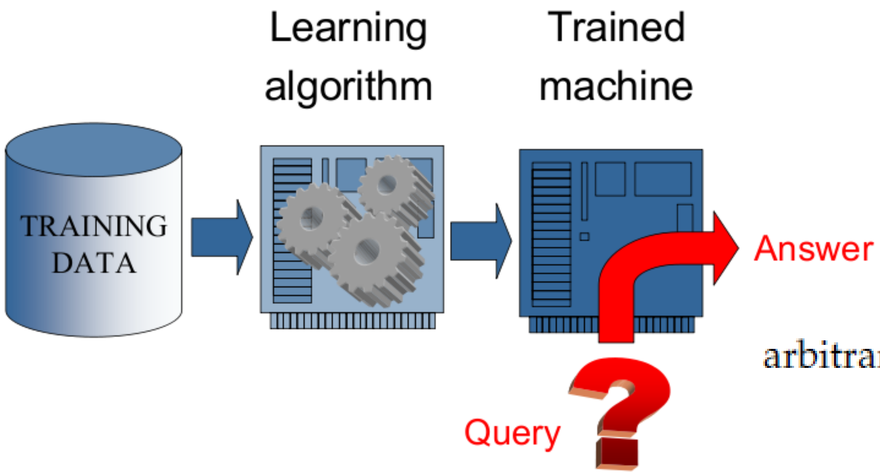
$\sigma_E / E_j \approx 4\%$ (45_GeV)

Our goal – remove confusion term

$\sigma_E / E_j \approx 2\%$

Supervised and Semi-Supervised Machine Learning

Machine learning concept



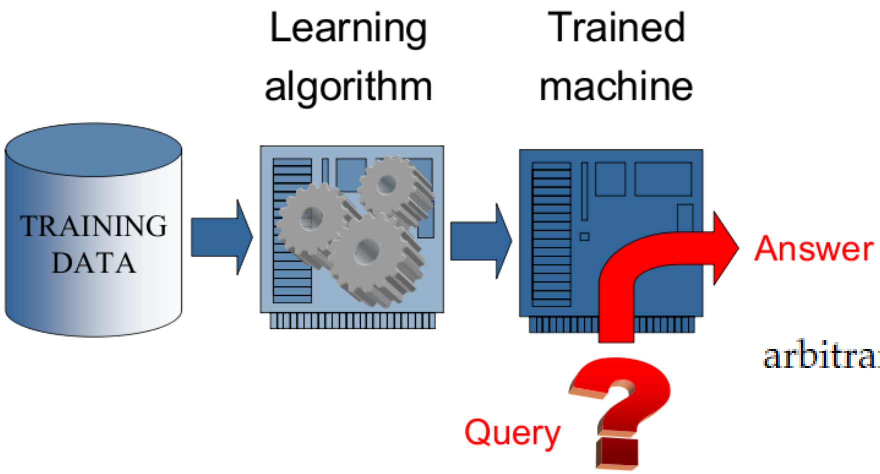
The goal of supervised learning is to infer a function $g : \mathcal{X} \to \mathcal{Y}$ from a data set

$$\mathcal{D} = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

The quality of g on an arbitrary pair $(\mathrm{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is measured by an *error* or *loss* function $L(g, (\mathrm{x}, y))$

Supervised and Semi-Supervised Machine Learning

Machine learning concept



The goal of supervised learning is to infer a function $g : \mathcal{X} \rightarrow \mathcal{Y}$ from a data set

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

The quality of g on an arbitrary pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is measured by an *error* or *loss* function $L(g, (x, y))$

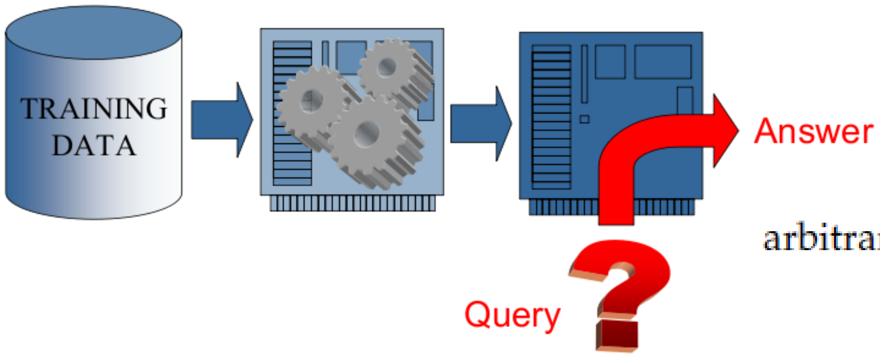
Train Dataset

9720 pixels

Raw Data

Supervised and Semi-Supervised Machine Learning

Machine learning concept



The goal of supervised learning is to infer a function $g : \mathcal{X} \to \mathcal{Y}$ from a data set

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

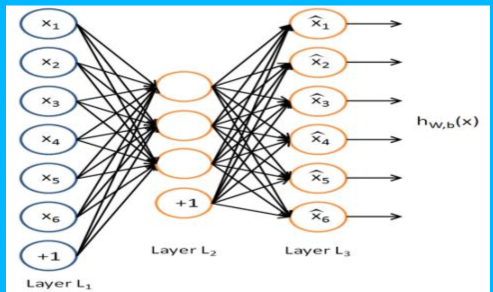
The quality of g on an arbitrary pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is measured by an *error* or *loss* function $L(g, (x, y))$

Train Dataset

9720 pixels

Raw Data

Deep Learning

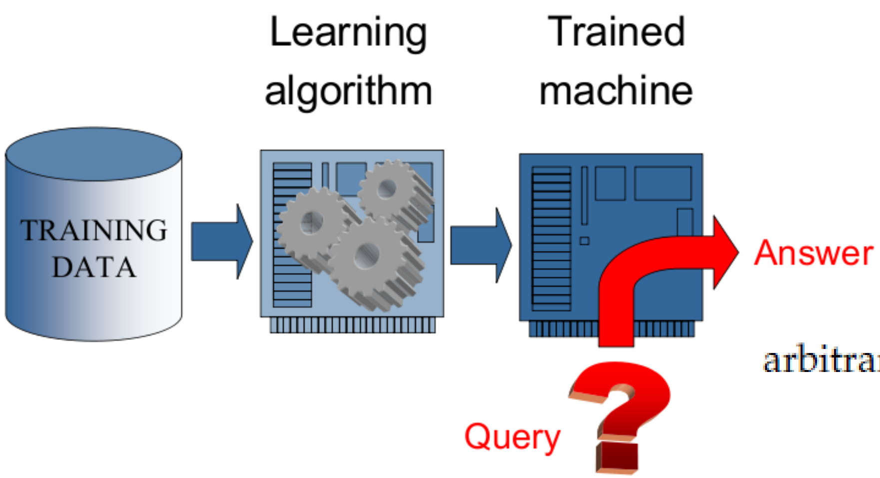


Auto-encoders, RBM, etc

Automatic Features Extraction

Supervised and Semi-Supervised Machine Learning

Machine learning concept



The goal of supervised learning is to infer a function $g : \mathcal{X} \rightarrow \mathcal{Y}$ from a data set

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

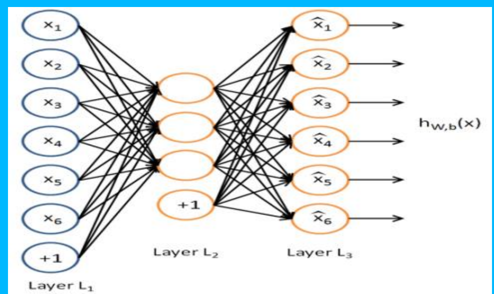
The quality of g on an arbitrary pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is measured by an *error* or *loss* function $L(g, (x, y))$

Train Dataset

9720 pixels

Raw Data

Deep Learning



Auto-encoders, RBM, etc

Automatic Features Extraction

9720 → 100-500 filters

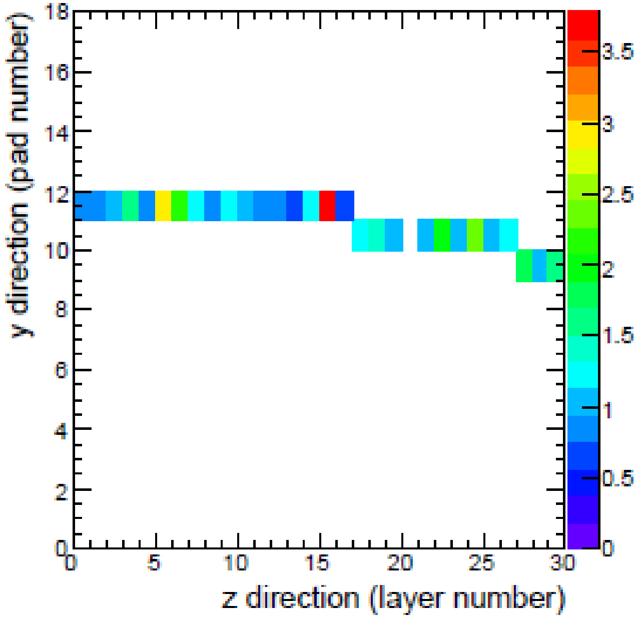
$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Labels (from Geant4 Simulation)

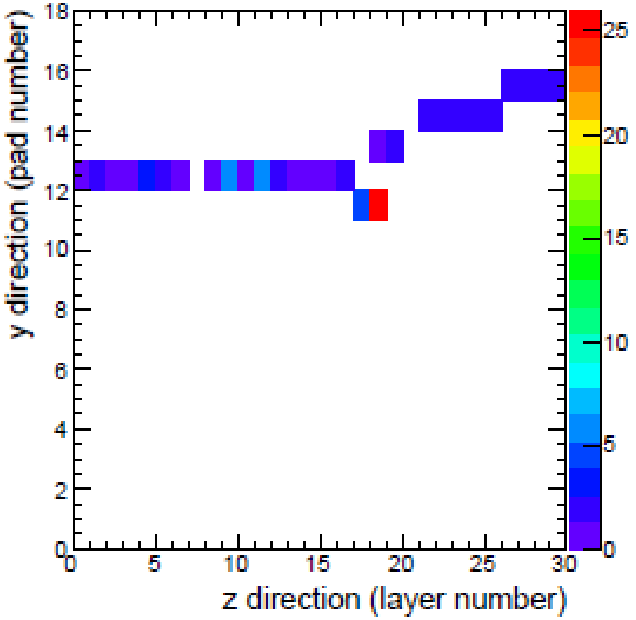
Classification

MultiBoost
and others

Test-Beam Events Interaction Classification



“Elastic”

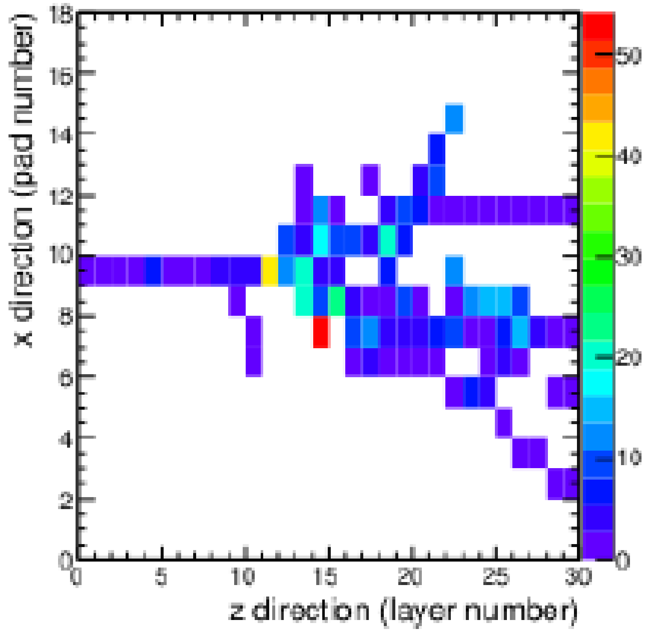


“Point-like”

Performance of ECAL physical prototype

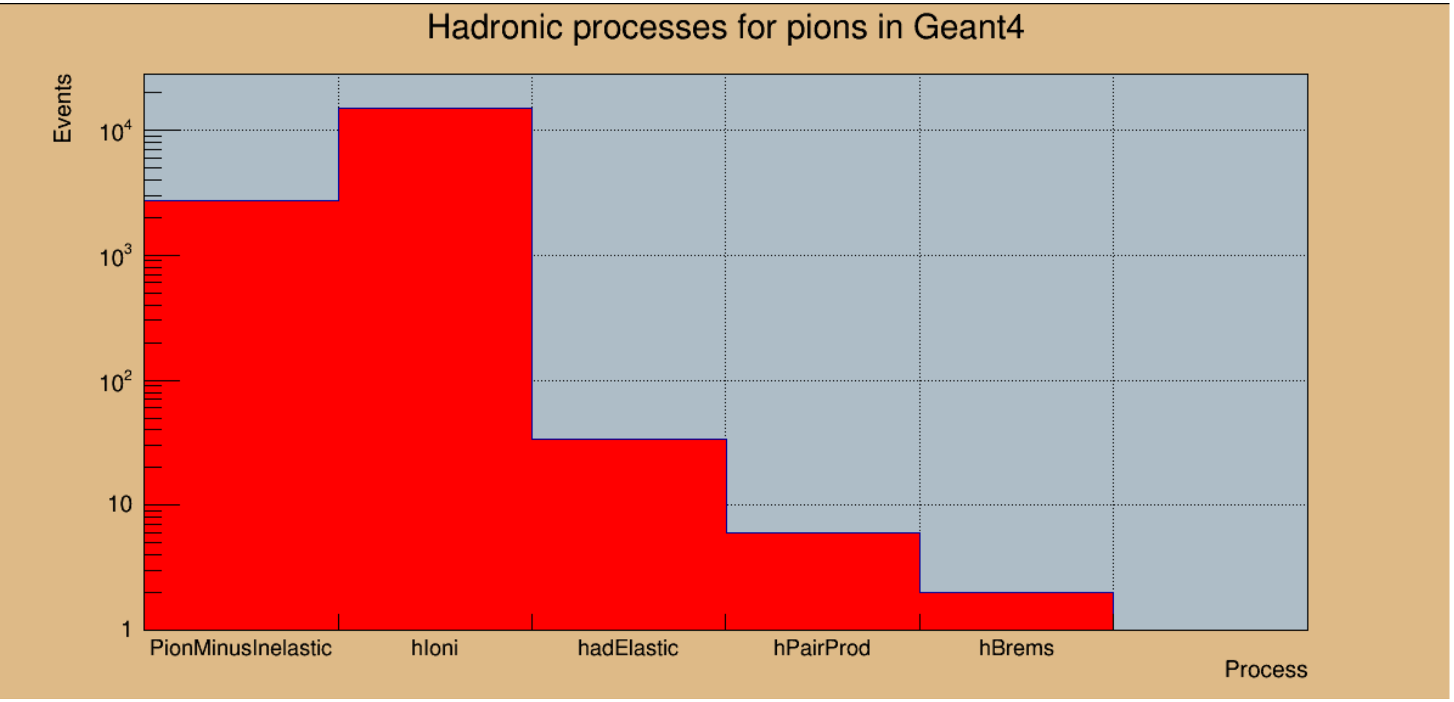
- Pions with energy 2-10 GeV

“FireBall”



Geant4(simulation toolkit) Set-Up

Ionization (MIP) | **Elastic** | **Inelastic** | **Pair Production**



- π^{-} 10 GeV
- QGSP_BERT physics list
  - ✓ quark-gluon string model of protons, neutrons, kaons and nuclei
  - ✓ Bertini cascade + parameterized model for low energy
- Full Detector geometry simulation (Mokka software)

Geant4 Set-Up

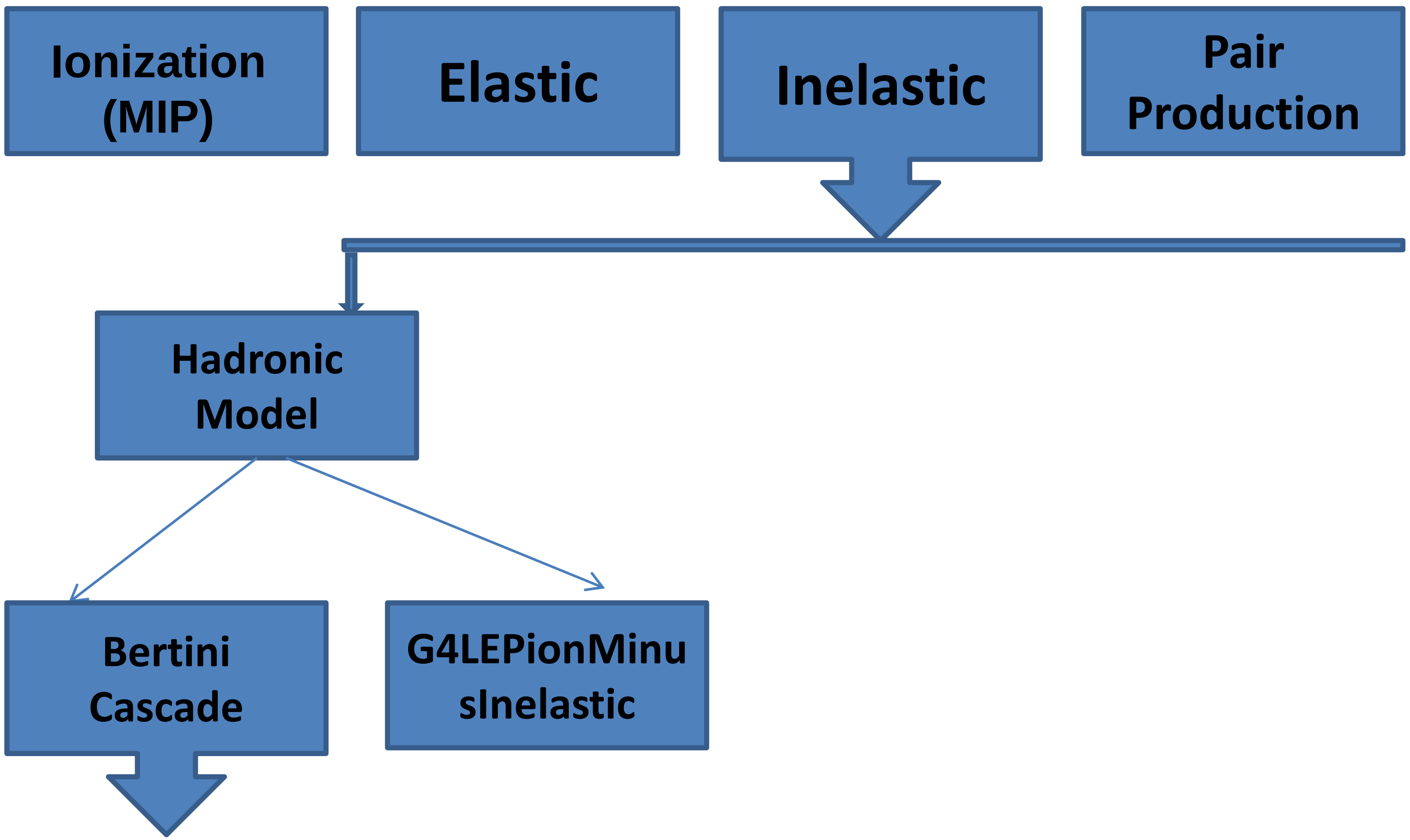
Ionization (MIP)

Elastic

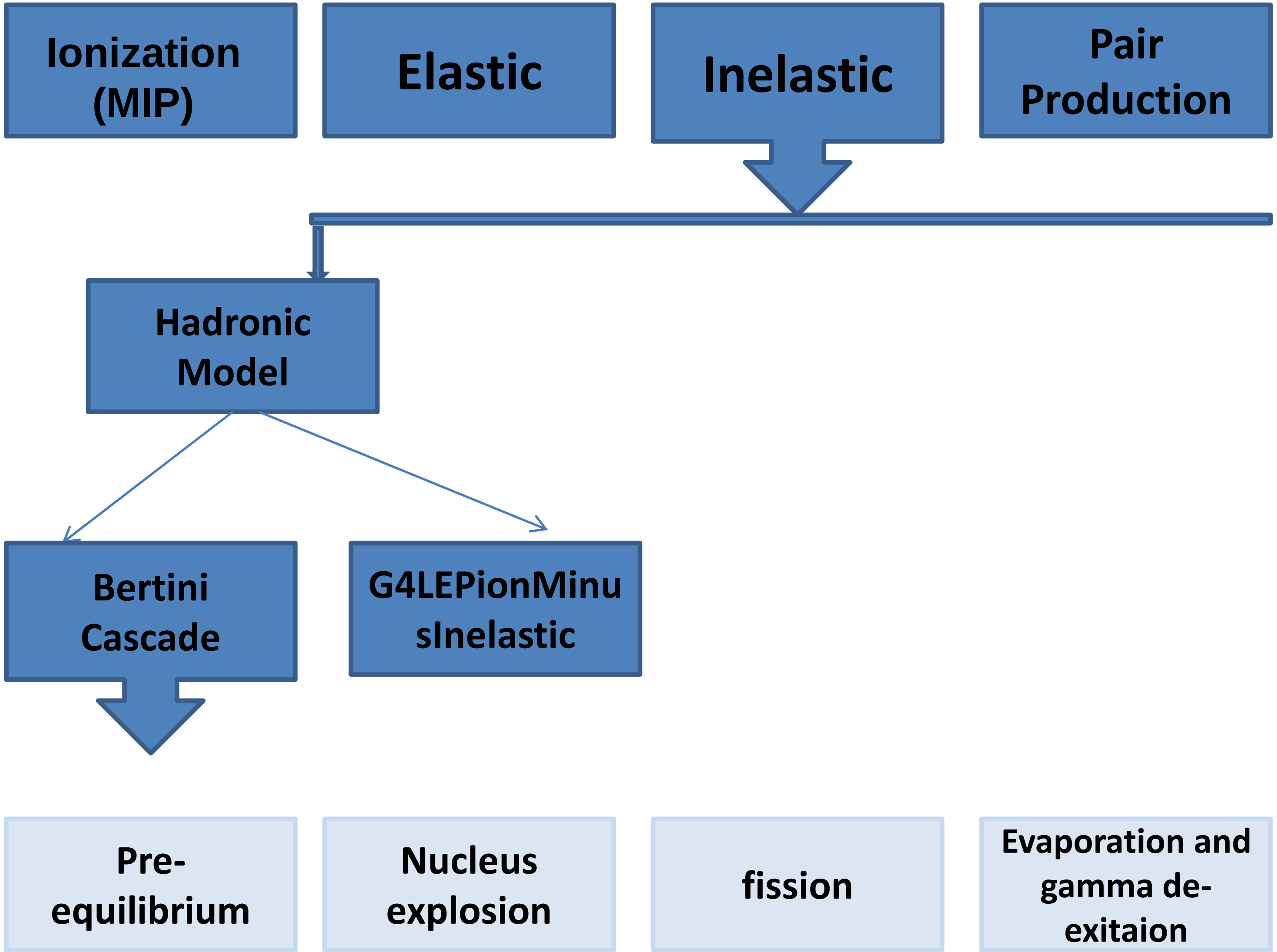
Inelastic

Pair Production

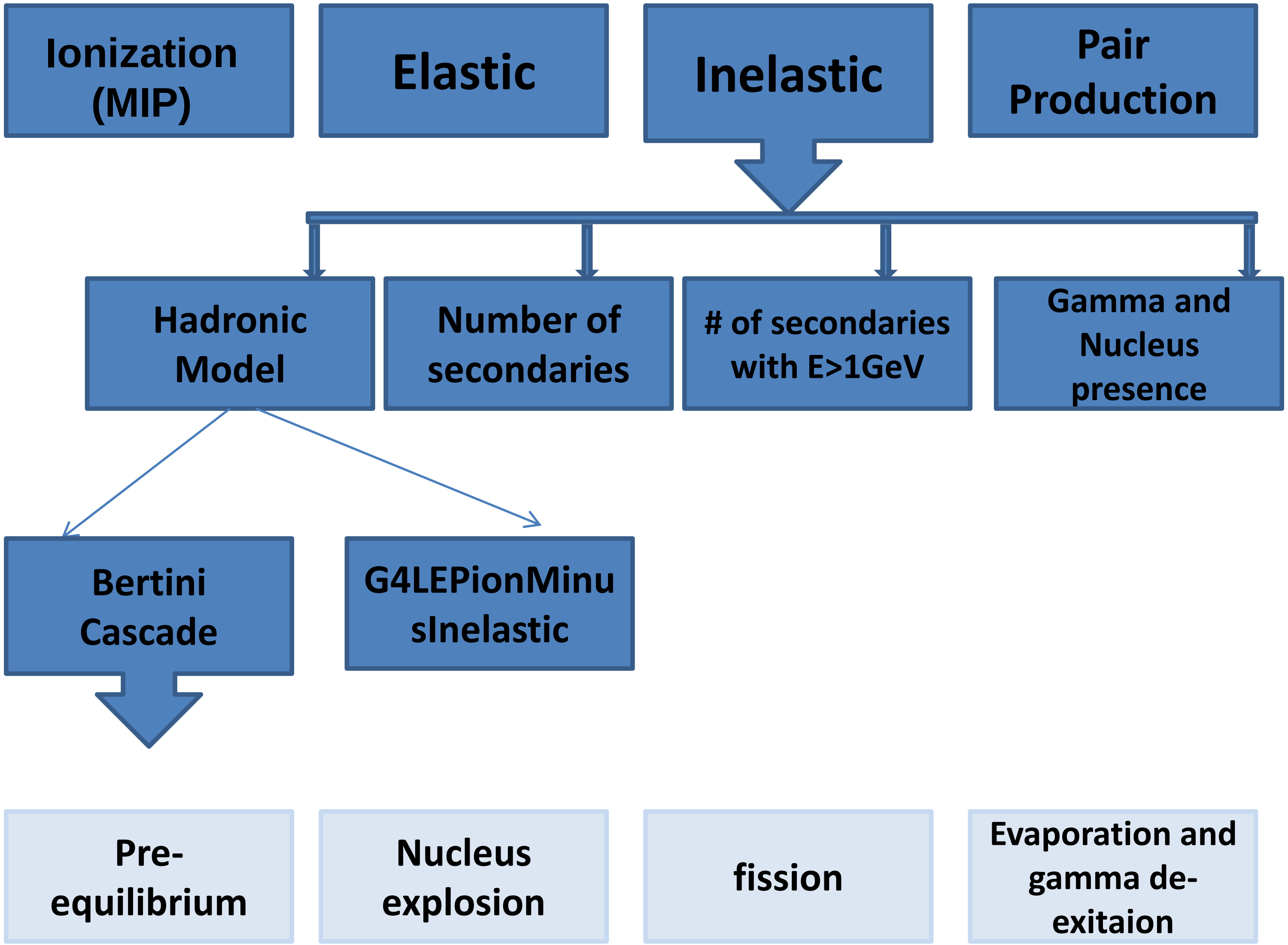
Geant4 Set-Up



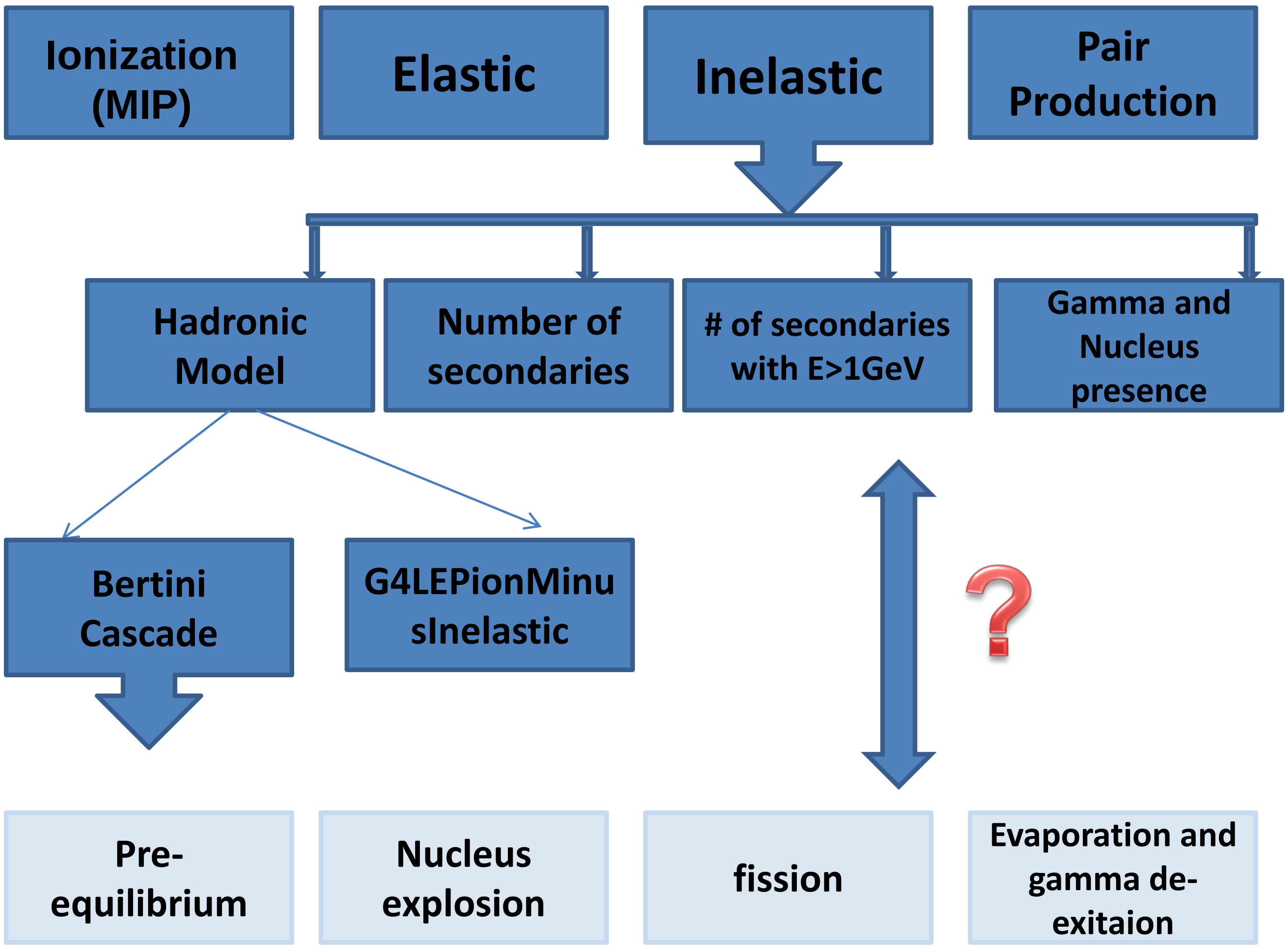
Geant4 Set-Up



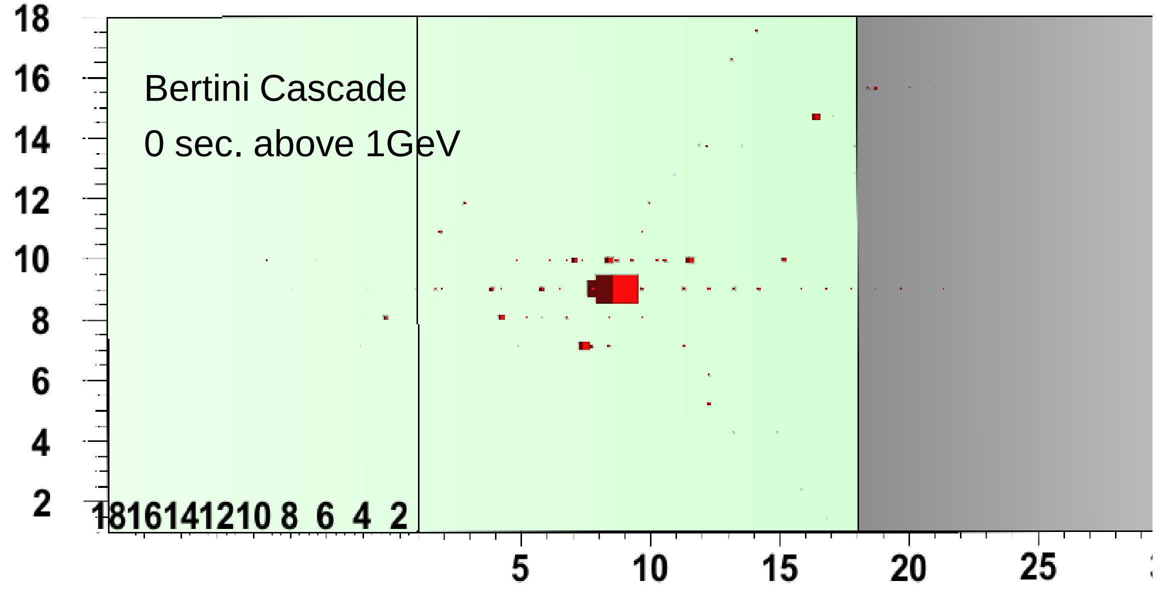
Geant4 Set-Up

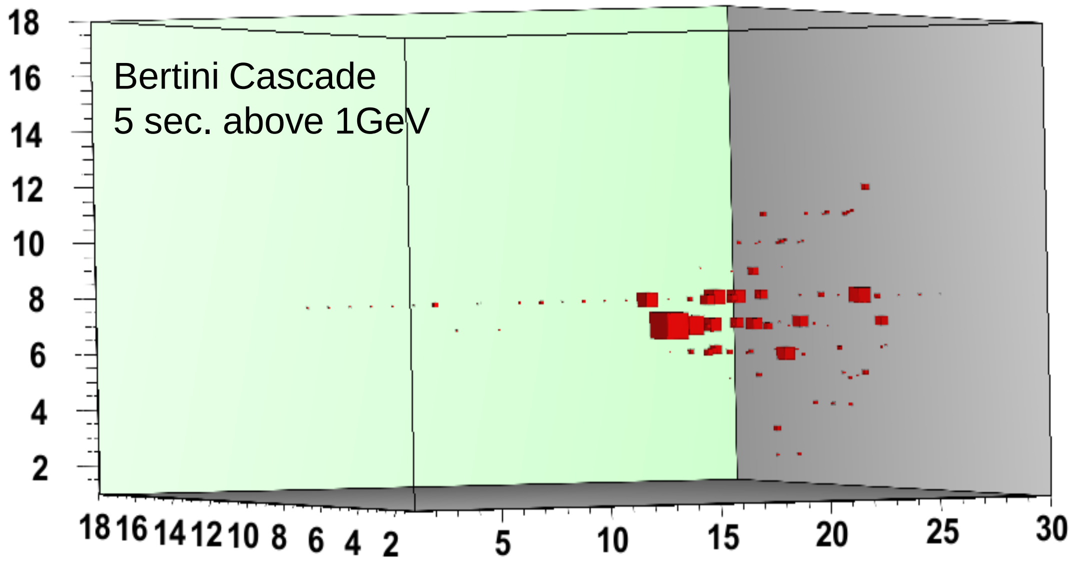


Geant4 Set-Up

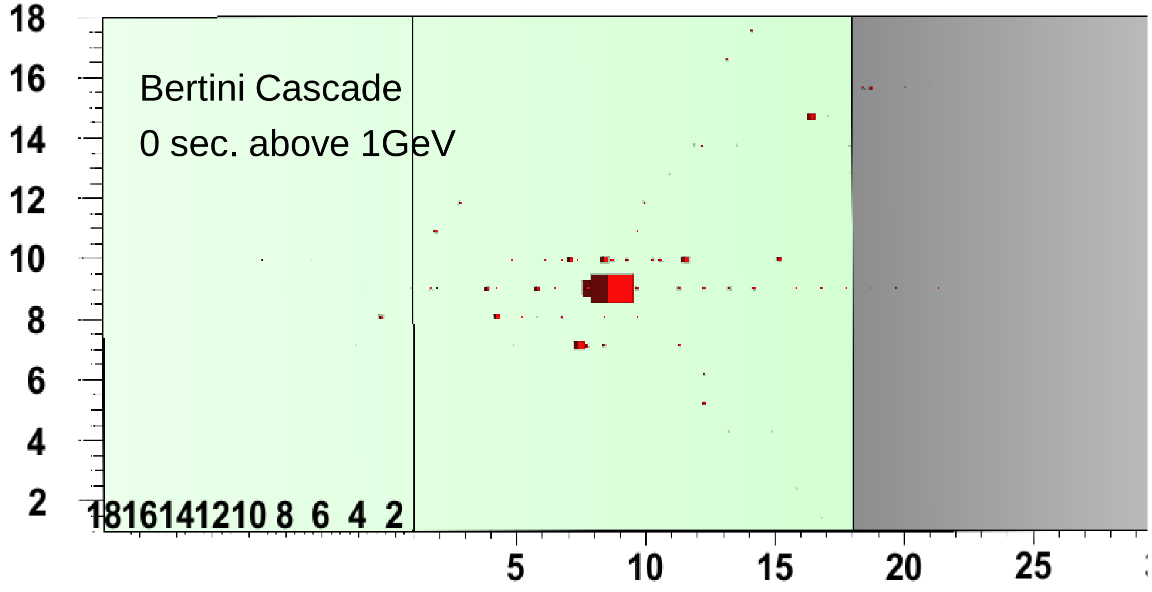


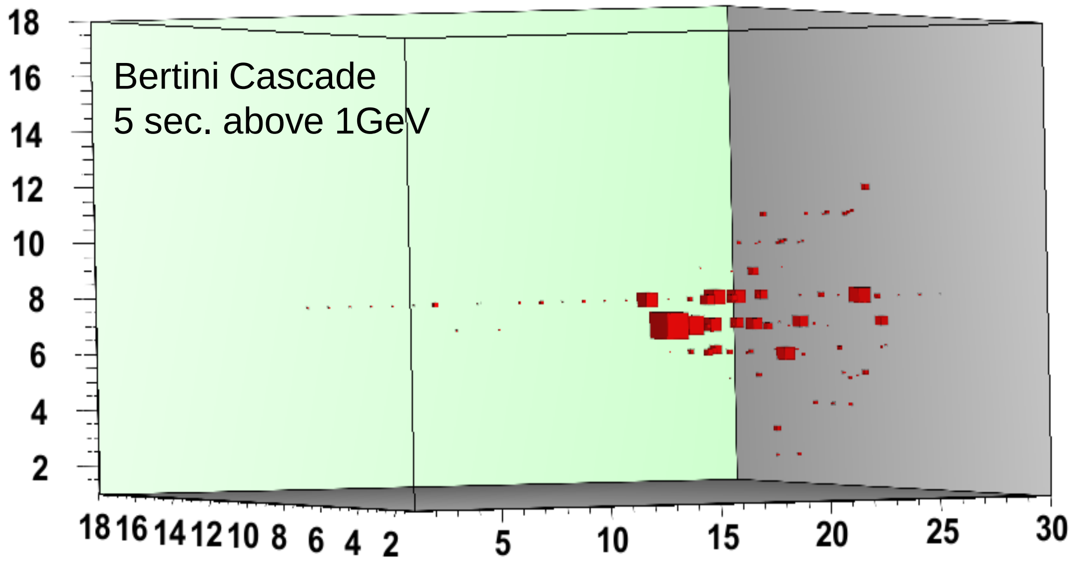
Events examples

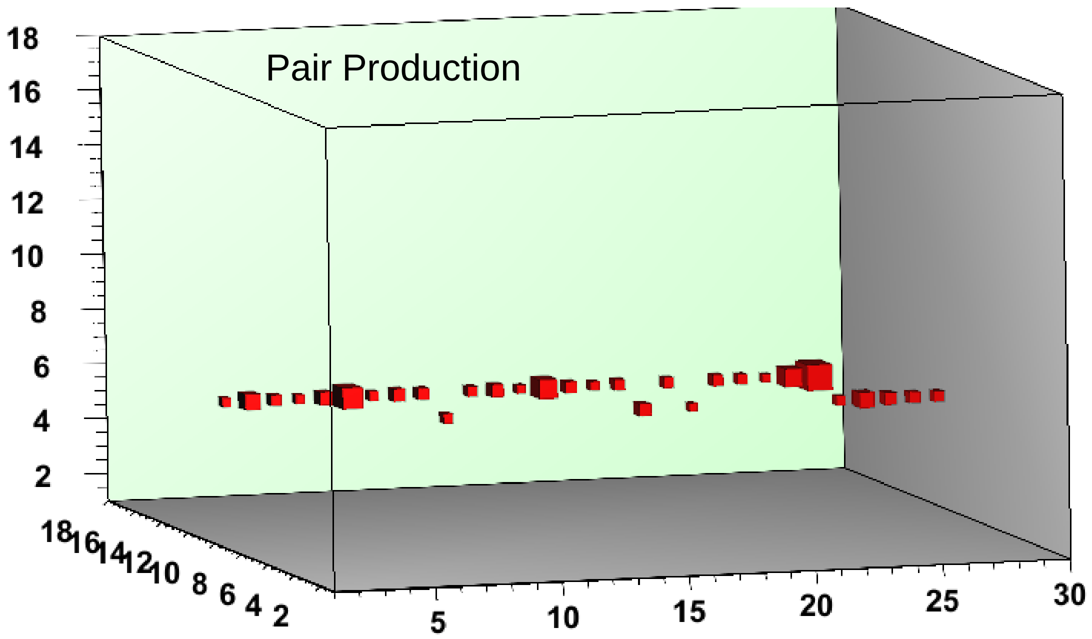




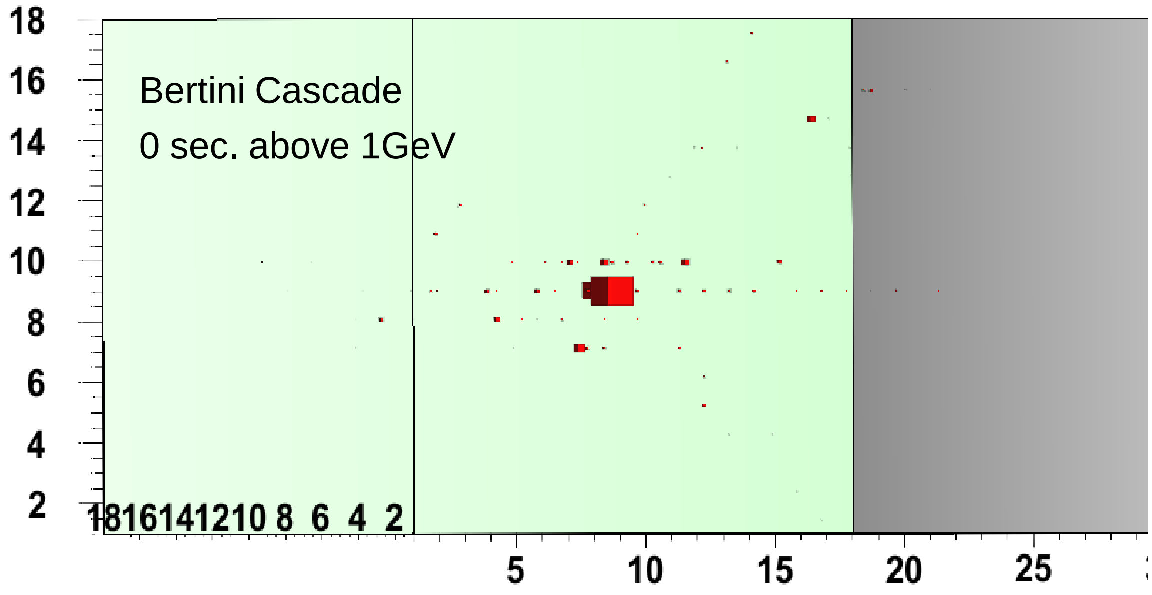
Events examples

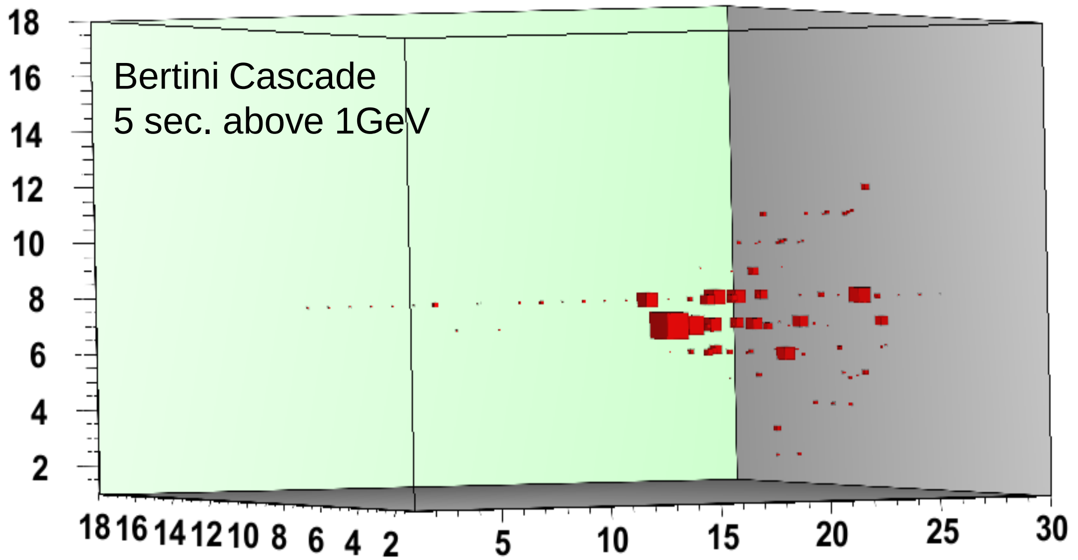


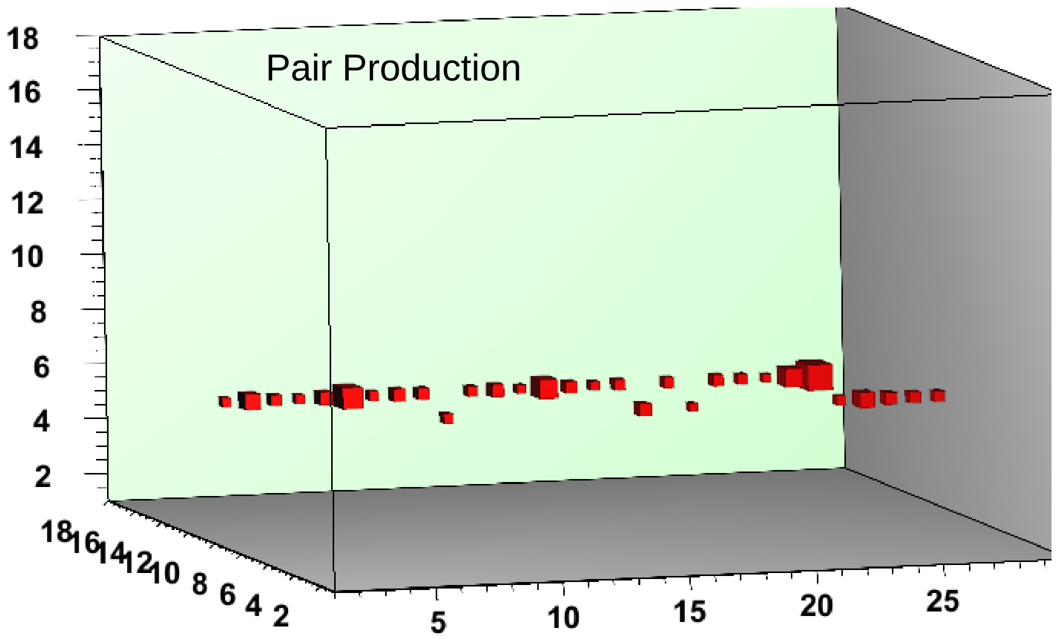


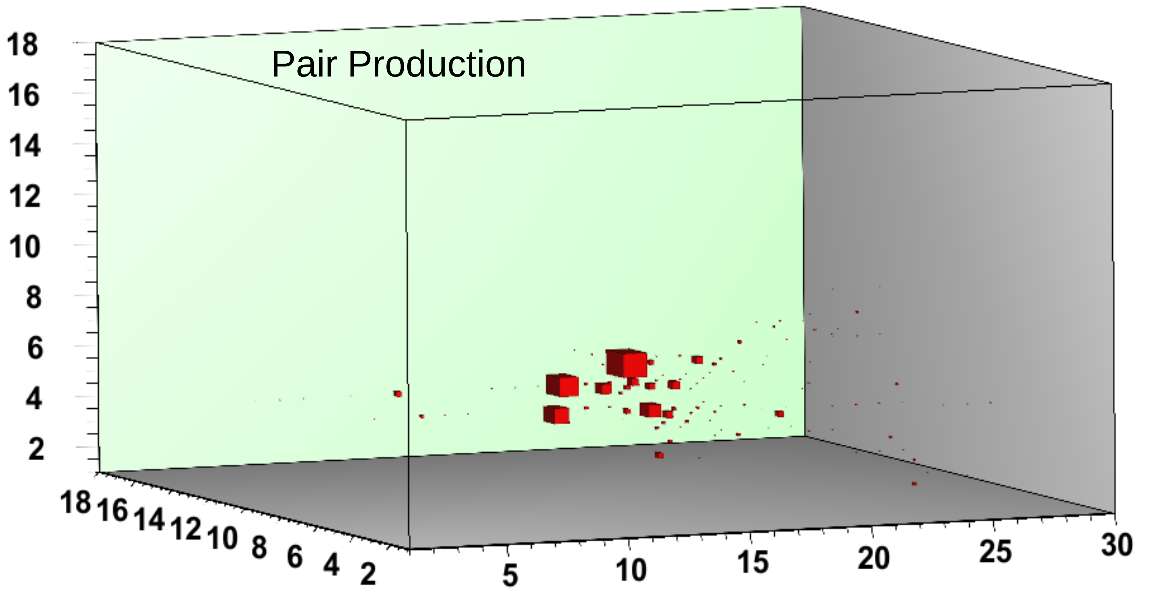


Events examples

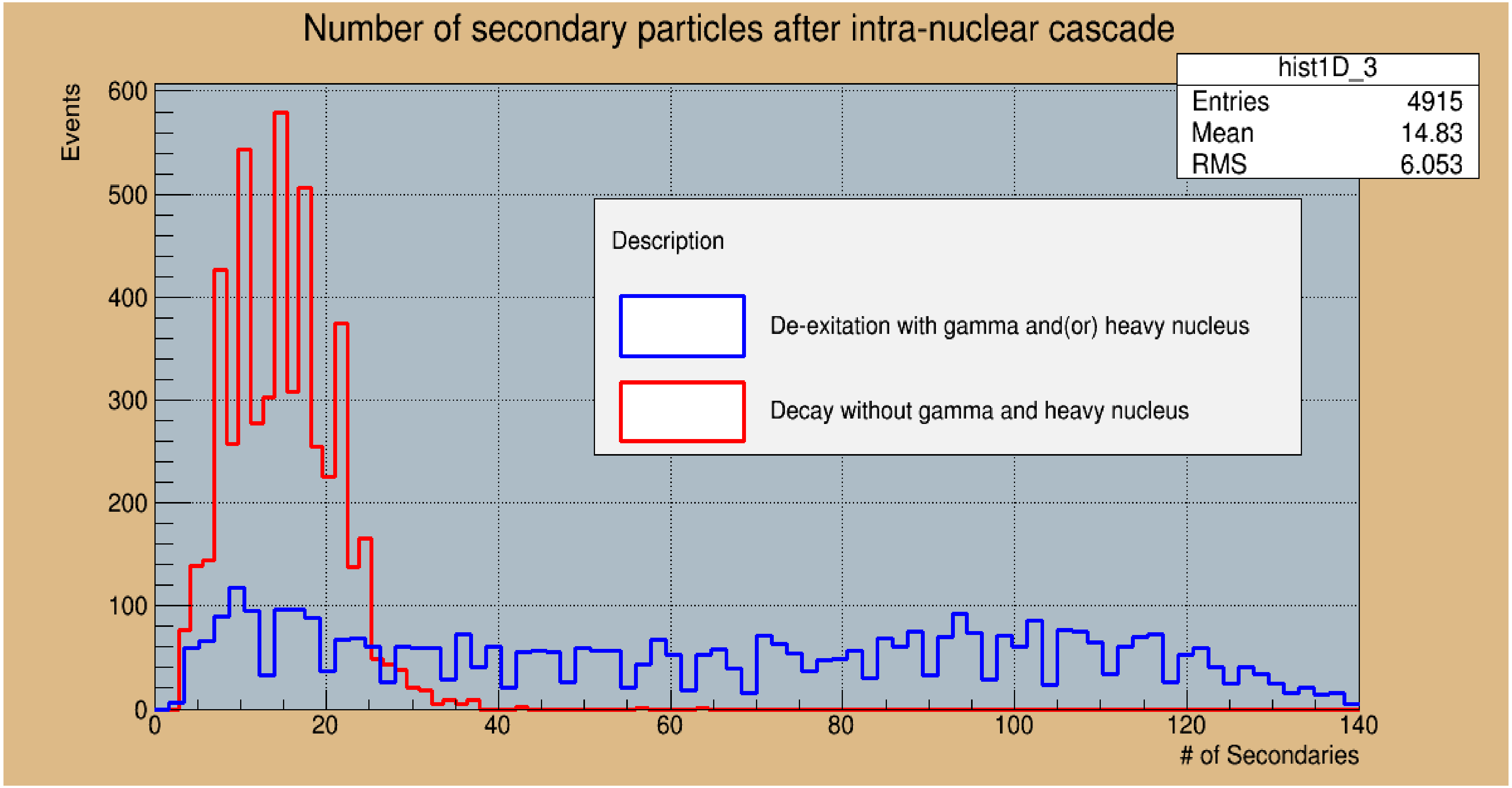




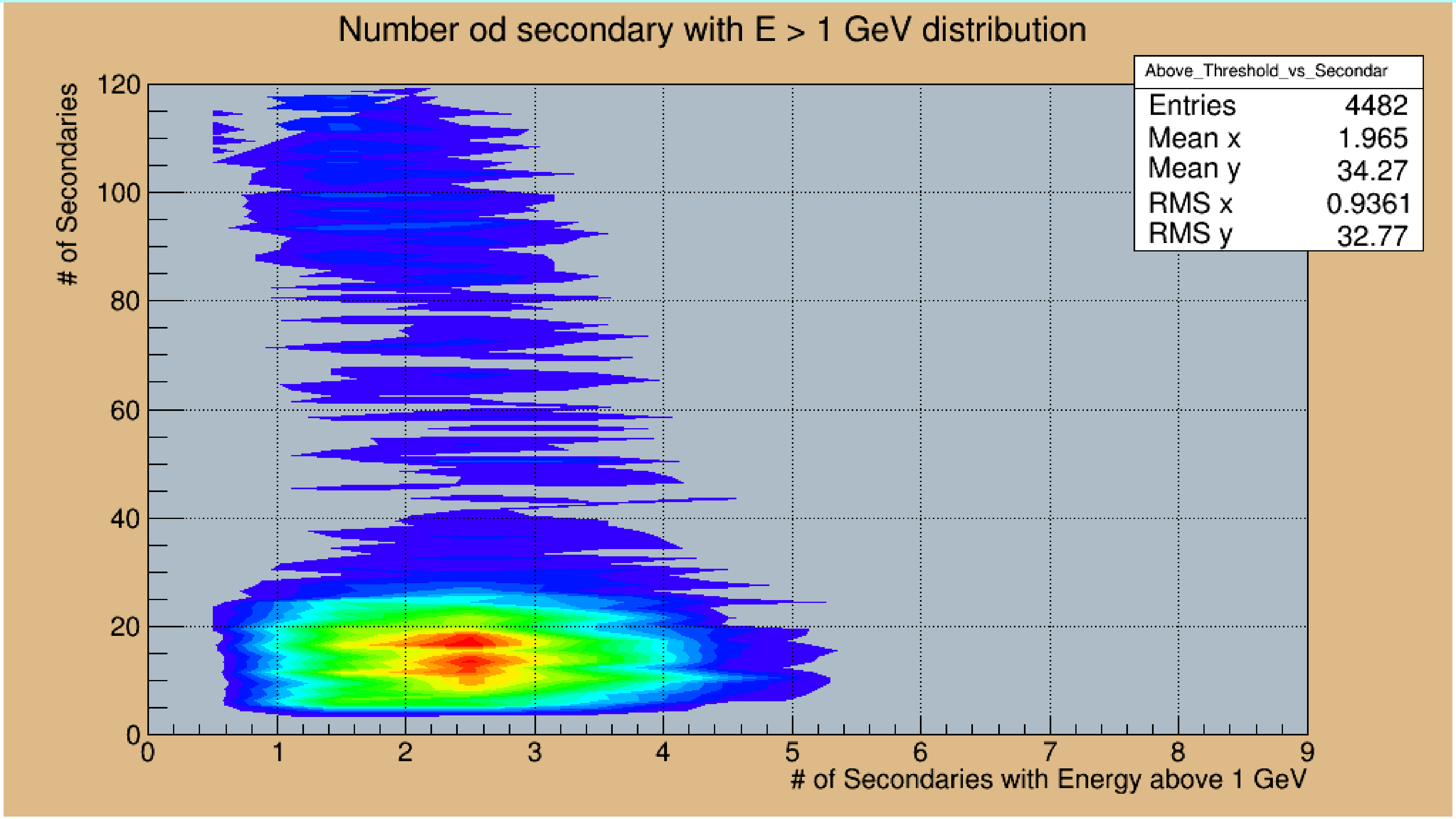




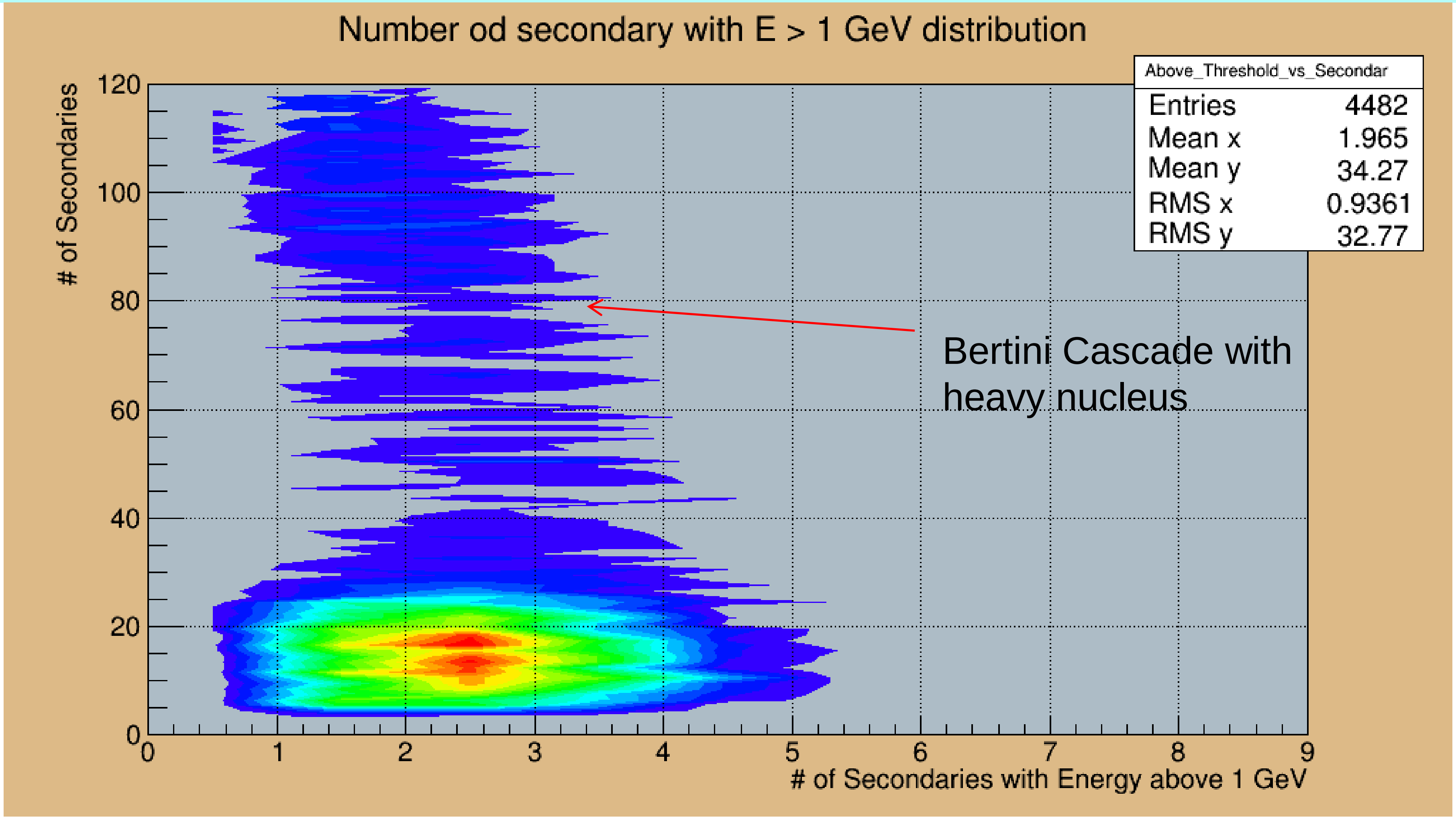
Relations between labels



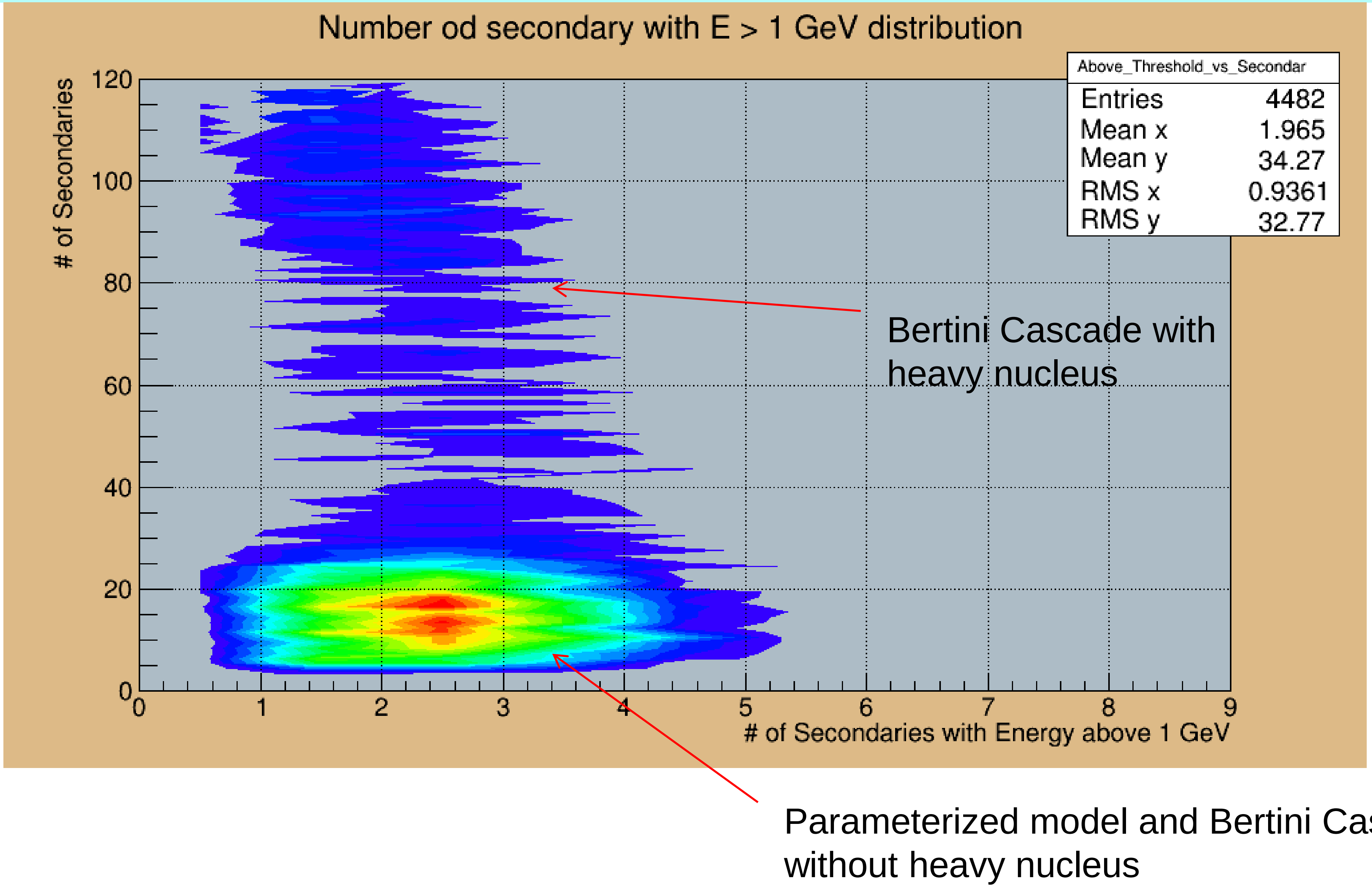
Relations between labels



Relations between labels



Relations between labels



MultiBoost Performance

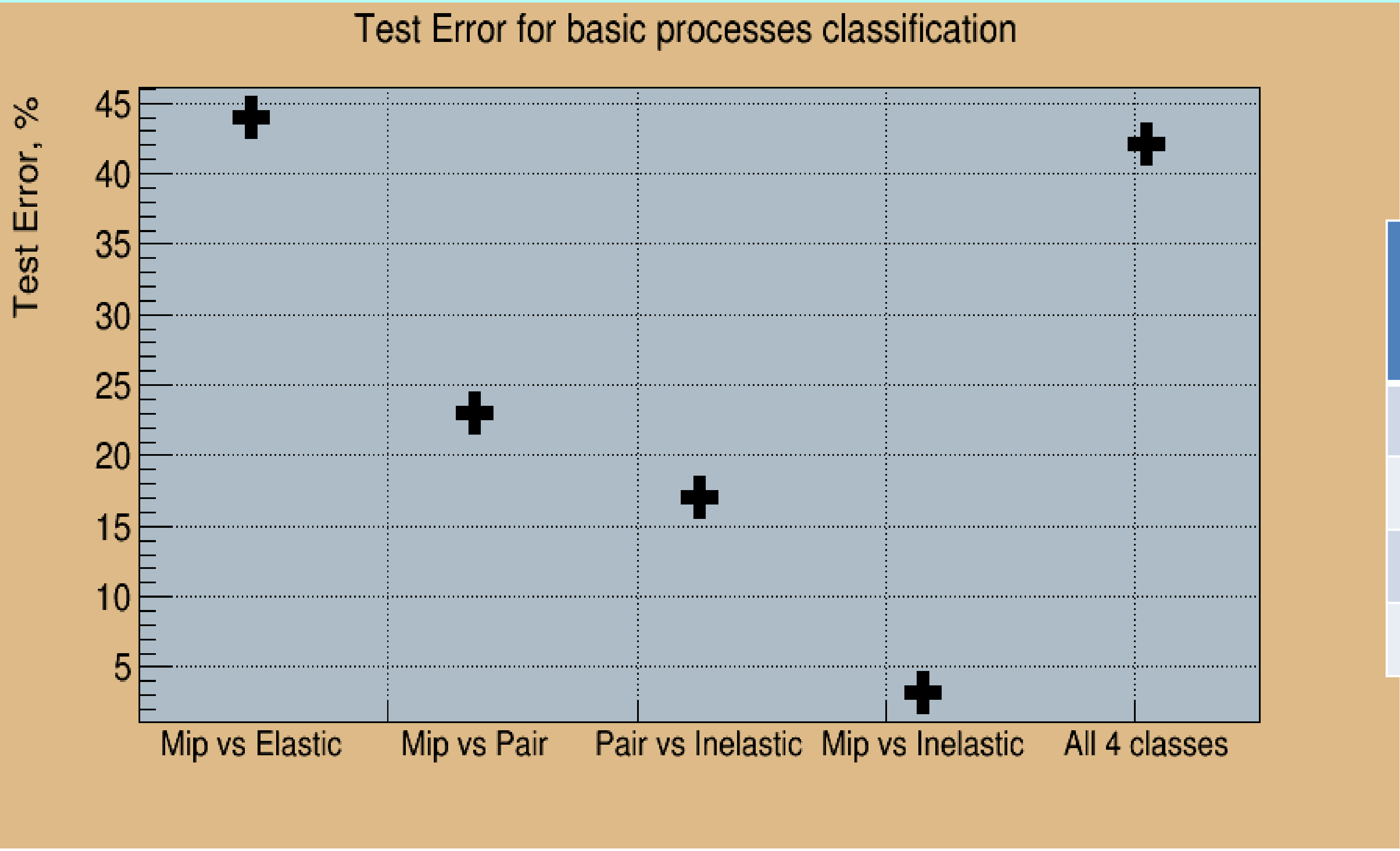
- Lerner type – Tree Learner
- BaseLearner – Single Stump Learner
- Number of Leaves – 2
- Number of iterations 100-300

MultiBoost Performance

- Lerner type – Tree Learner
- BaseLearner – Single Stump Learner
- Number of Leaves – 2
- Number of iterations 100-300

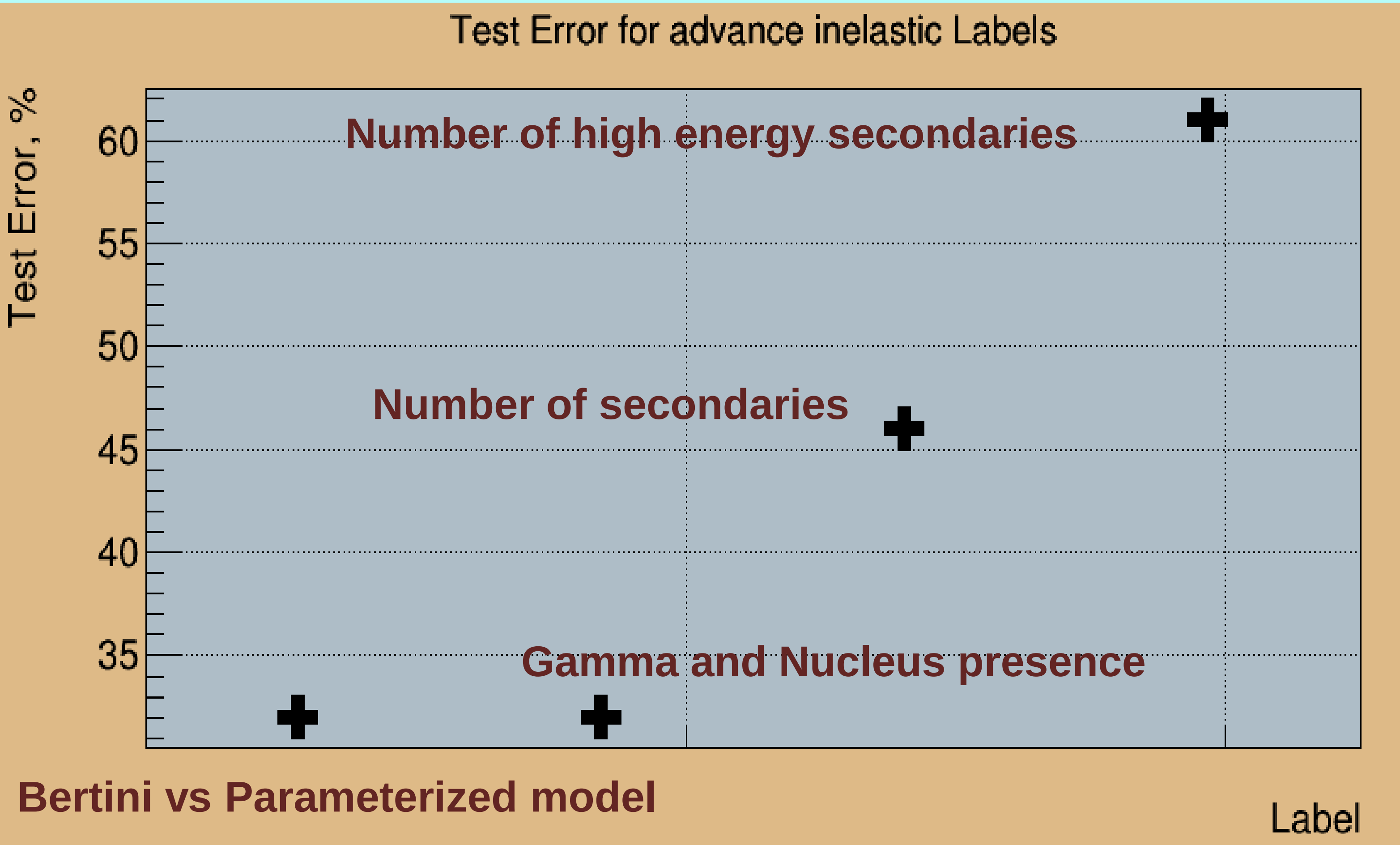
Manual Features – different combinations of longitudinal profile, energy per layer, radius per layer, Hough transform, etc

MultiBoost Performance



	Train set # of events	Test set # of events
Mip	3739	935
Pair	1760	440
Inelastic	3465	867
Elastic	2788	698

MultiBoost Performance



Bertini vs Parameterized model

	Train set	Test set
Bertini	6565	1633
G4LEP	7429	1866
Gamma & Nucleus	6416	1602
No Gamma & Nucleus	7578	1897
# sec. 0-20	7398	1888
# sec. 20-40	2326	521
# sec. 40-70	1182	316
# sec.70-100	1527	384
# sec.>100	1561	390
#>1GeV. 0	576	153
#>1GeV. 1	4430	1169
#>1GeV. 2	5446	1316
#>1GeV. 3	2815	675
#>1GeV. >3	727	186

Conclusions & Outlook

- Architecture of Geant4 Kernel was explored .
- The basic labels which corresponds to hadronic physical processes for 10 GeV pions was produces with QGSP_BERT geant4 physics list
- More advanced investigation of pions Inelastic interactions with detector media was studied.
- Supervised learning set-up was tested for hadrons interaction classification and produced benchmarks for further study with deep learning technique. The results is satisfactory with basic classes(physical processes), however in advance inelastic interaction classification results is not significant.
- Further study will imply usage of deep deep learning for existing labels and creation of new ones, more significant.
- After interaction classification problem and precise tune of pipeline, semi-supervised learning will be applied for the study of multiple events and optimization of existing PFA with the aim to remove confusion term and improve jet energy resolution.

Gratitude

- The author is thankful to ILC and AppStat team for their help, advises and friendly company. And especial gratitude to **Naomi van der Kolk, Balazs Kegl, Thibault Frisson, Philip Bambade, Franck Dubard**.

Backup Slides

e01 refers to the *one-error*

$$\hat{R}_{\mathrm{I}}\left(\mathbf{f}^{(t)},\mathbf{X},\mathbf{Y}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{y_{i,\ell_{\mathbf{f}^{(t)}}(\mathbf{x}_i)}<0\right\}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\max_{\ell:y_{i,\ell}<0}f_{\ell}^{(t)}(\mathbf{x}_i)\geq\max_{\ell:y_{i,\ell}>0}f_{\ell}^{(t)}(\mathbf{x}_i)\right\}.$$

where

$$\ell_{\mathbf{f}^{(t)}}(\mathbf{x}_i)=\arg\max_{\ell'}f_{\ell'}^{(t)}(\mathbf{x}_i) \tag{4}$$

is the single label predicted by $\mathbf{f}^{(t)}$.[3] If only one label is positive (single-label multi-class), the error condition is equivalent to $\ell_{\mathbf{f}^{(t)}}(\mathbf{x}_i)\neq\ell(\mathbf{x}_i)$ where $\ell(\mathbf{x}_i)$ is the correct single label of $\mathbf{x}_i$. If more than one labels are positive (multi-label), it is sufficient for a good prediction if the predicted label (4) is *one* of the positive labels.

Backup Slides

$\text{AdaBoost}(\mathcal{D}, \text{Base}(\cdot,\cdot), T)$

$\mathbf{w}^{(1)} \leftarrow (1/n, \ldots, 1/n)$ ▷ *initial weights*

for $t \leftarrow 1$ **to** T

$\quad h^{(t)} \leftarrow \text{Base}(\mathcal{D}, \mathbf{w}^{(t)})$ ▷ *base classifier*

$\quad \epsilon^{(t)} \leftarrow \sum_{i=1}^{n} w_i^{(t)} \mathbb{I}\left\{h^{(t)}(\mathbf{x}_i) \neq y_i\right\}$ ▷ *weighted error of the base classifier*

$\quad \alpha^{(t)} \leftarrow \frac{1}{2} \ln\left(\frac{1-\epsilon^{(t)}}{\epsilon^{(t)}}\right)$ ▷ *coefficient of the base classifier*

$\quad$ **for** $i \leftarrow 1$ **to** n ▷ *re-weighting the training points*

$\quad\quad$ **if** $h^{(t)}(\mathbf{x}_i) \neq y_i$ **then** ▷ *error*

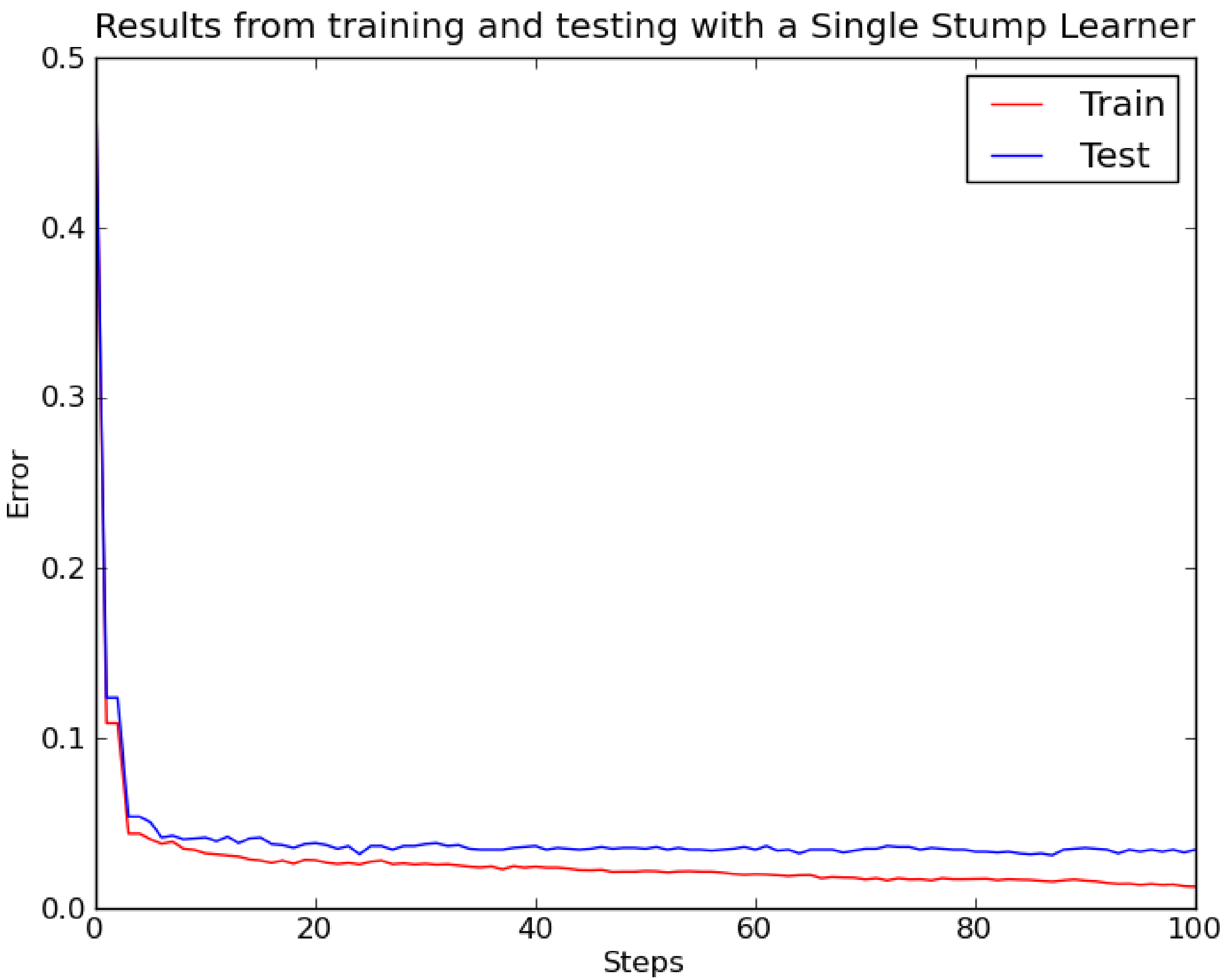
$\quad\quad\quad w_i^{(t+1)} \leftarrow \frac{w_i^{(t)}}{2\epsilon^{(t)}}$ ▷ *weight increases*

$\quad\quad$ **else** ▷ *correct classification*

$\quad\quad\quad w_i^{(t+1)} \leftarrow \frac{w_i^{(t)}}{2(1-\epsilon^{(t)})}$ ▷ *weight decreases*

return $f^{(T)}(\cdot) = \sum_{t=1}^{T} \alpha^{(t)} h^{(t)}(\cdot)$ ▷ *weighted "vote" of base classifiers*

Backup Slides
Mip vs inelastic



Backup Slides
Mip vs Elastic