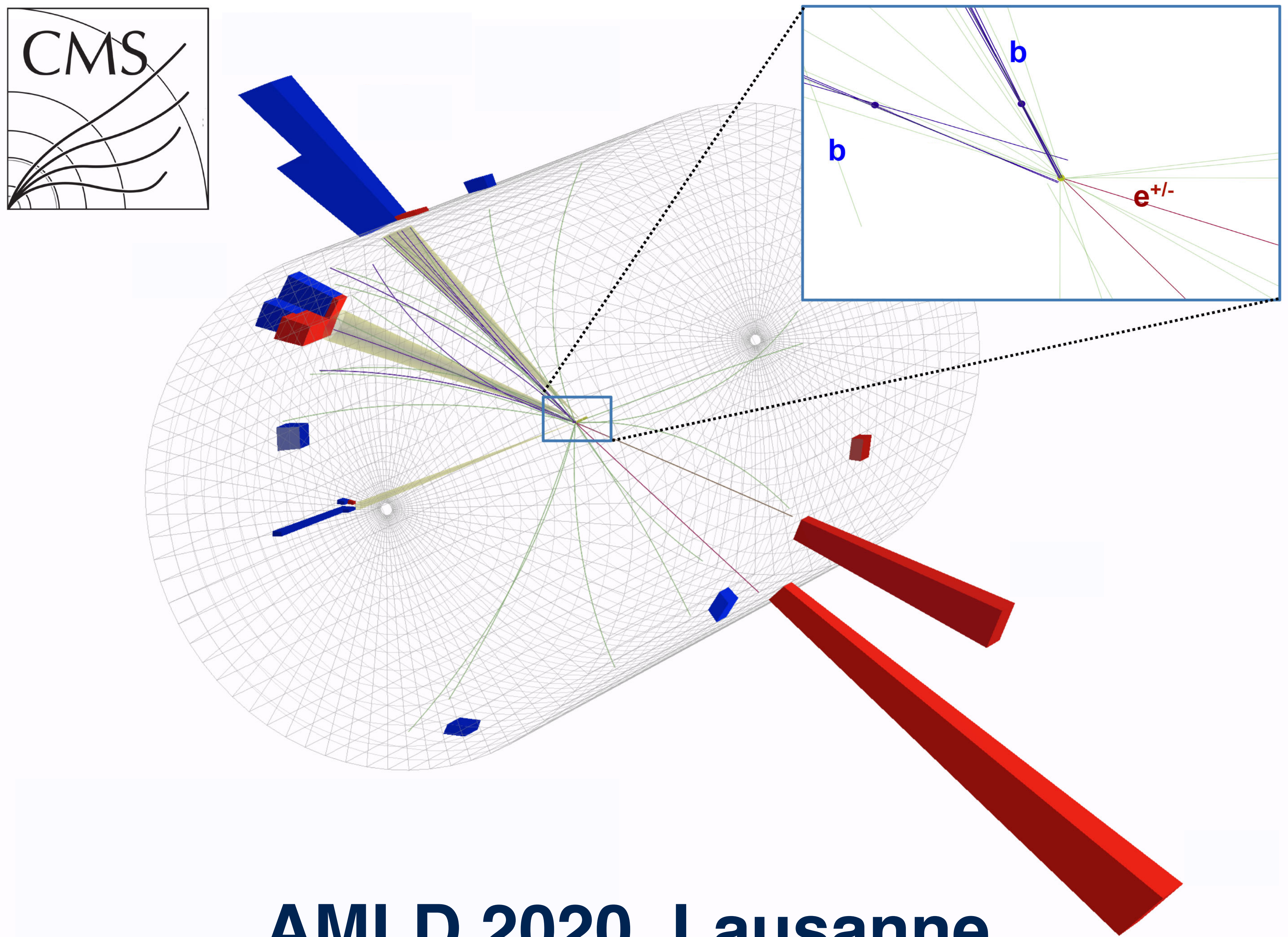
b-jet energy regression for the CMS experiment

Nadya Chernyavskaya - ETH Zurich

on behalf of the CMS collaboration



AMLD 2020, Lausanne

THE LARGE HADRON COLLIDER

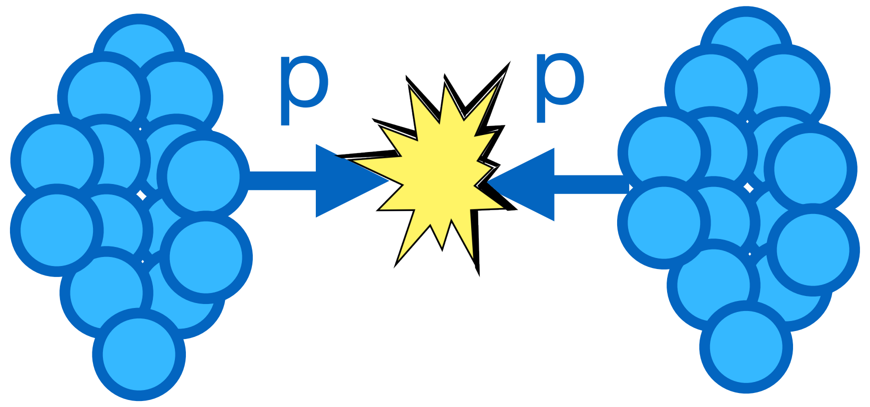
proton-proton collider @ 13 TeV center-of-mass energy
4 interaction points (detectors)

Photo : Maximilian Brice CERN

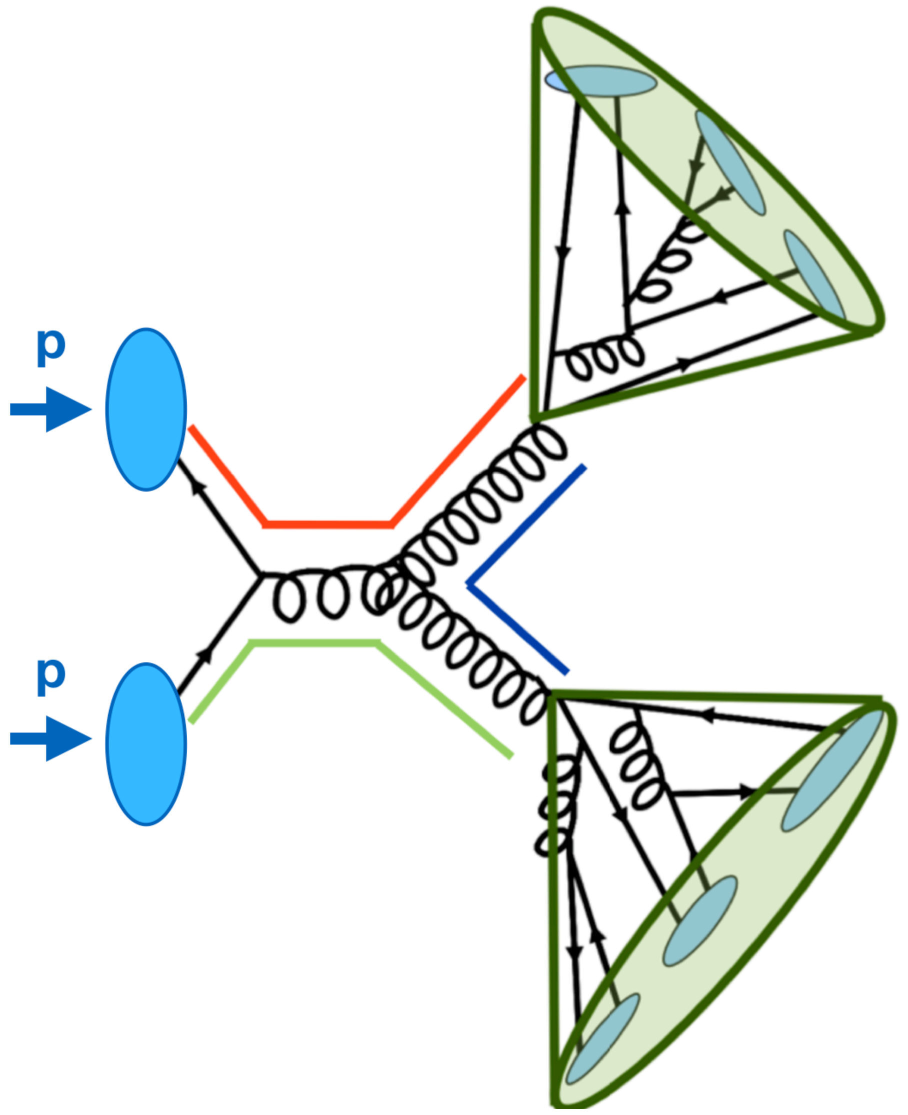

THE LARGE HADRON COLLIDER
Giant 'cameras' recording particles emerging after pp collision
LHCb
SUISSE
FRANCE
CMS
ALICE
CMS
Total weight 14,000 tonnes
15 m
40 million collisions / sec
p
LHC 27 km
p
proton-proton collider @ 13 TeV center-of-mass energy
4 interaction points (detectors)
Photo : Maximilian Brice CERN

Physics at LHC

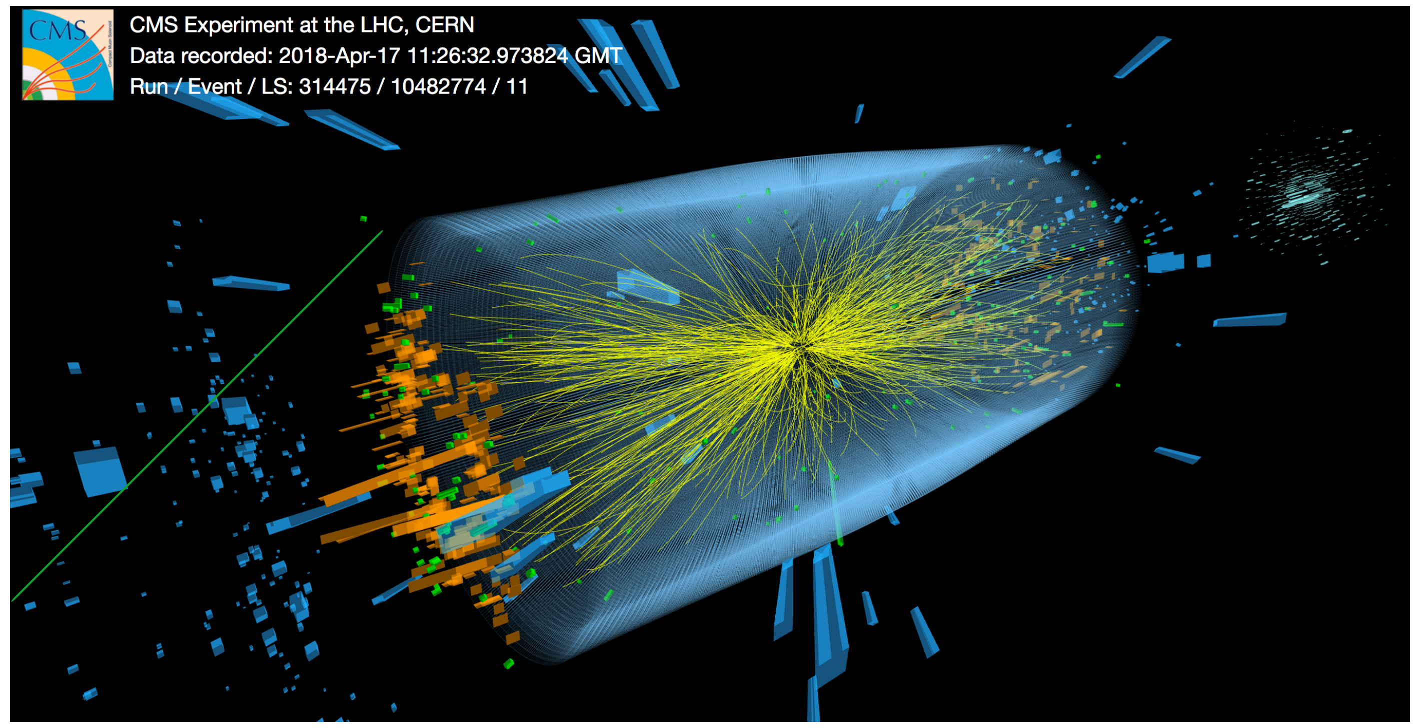
Protons collide



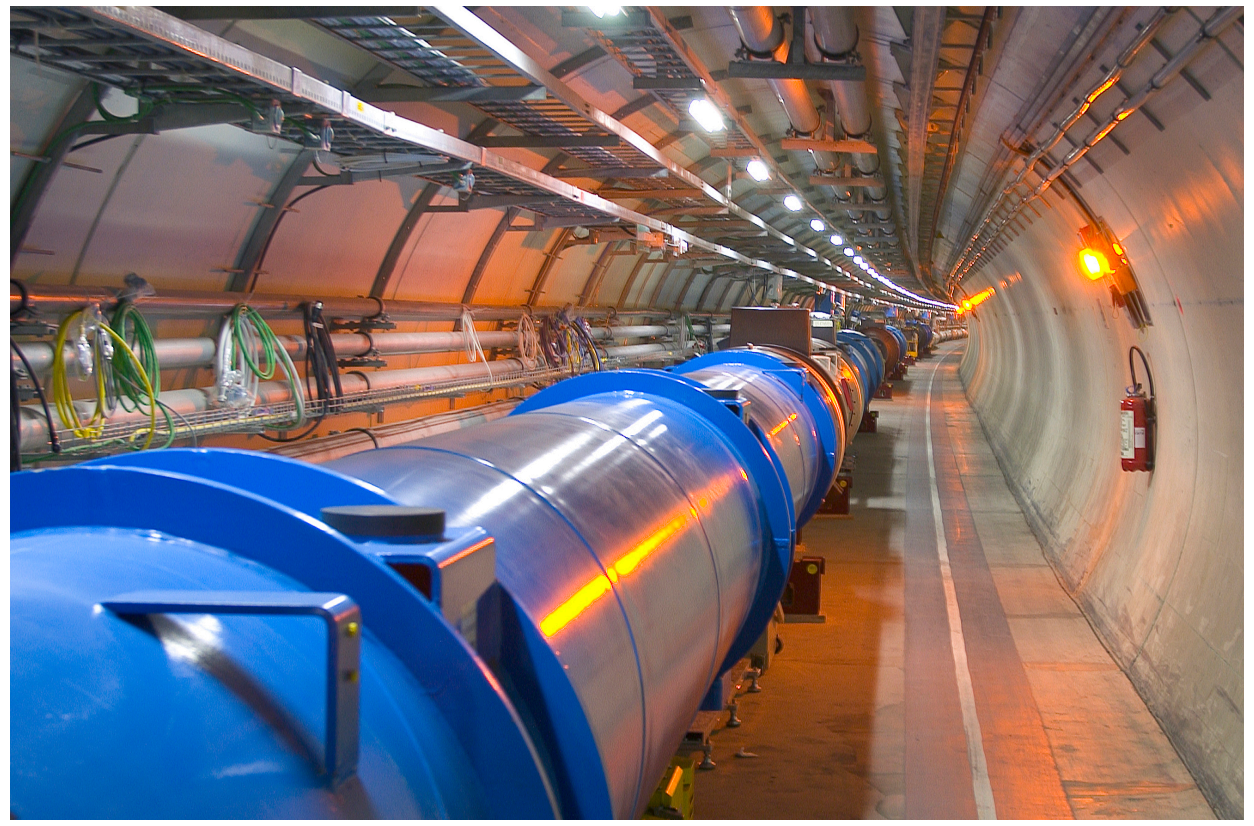
Underlying physics



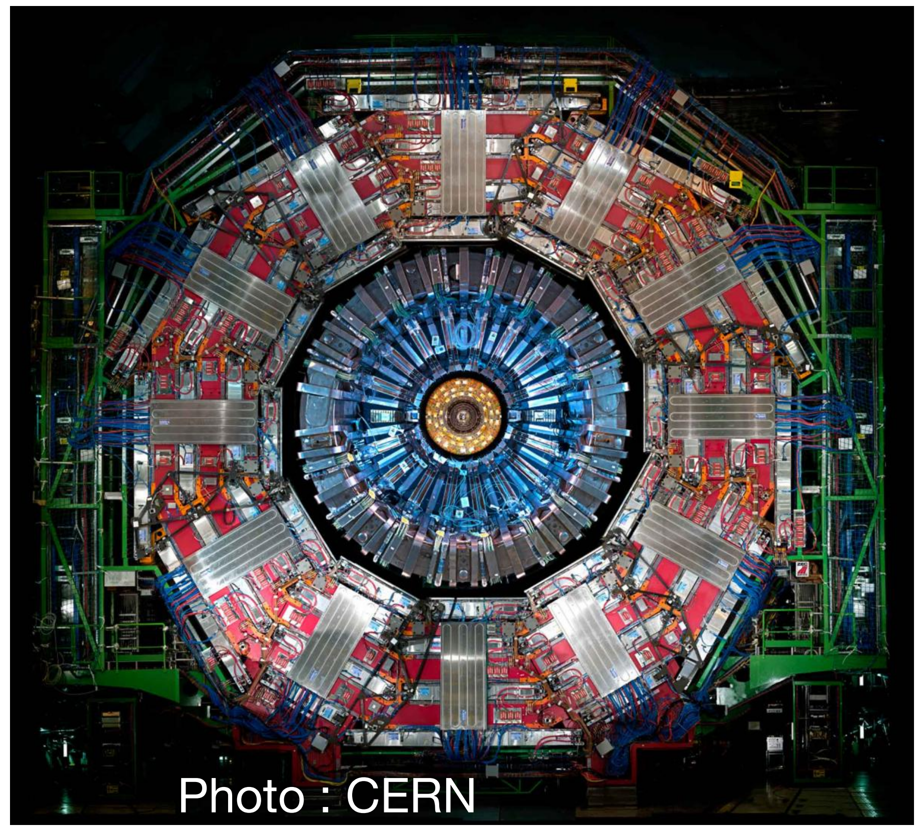
Detection of particle produced & underlying physics extraction



LHC tunnel



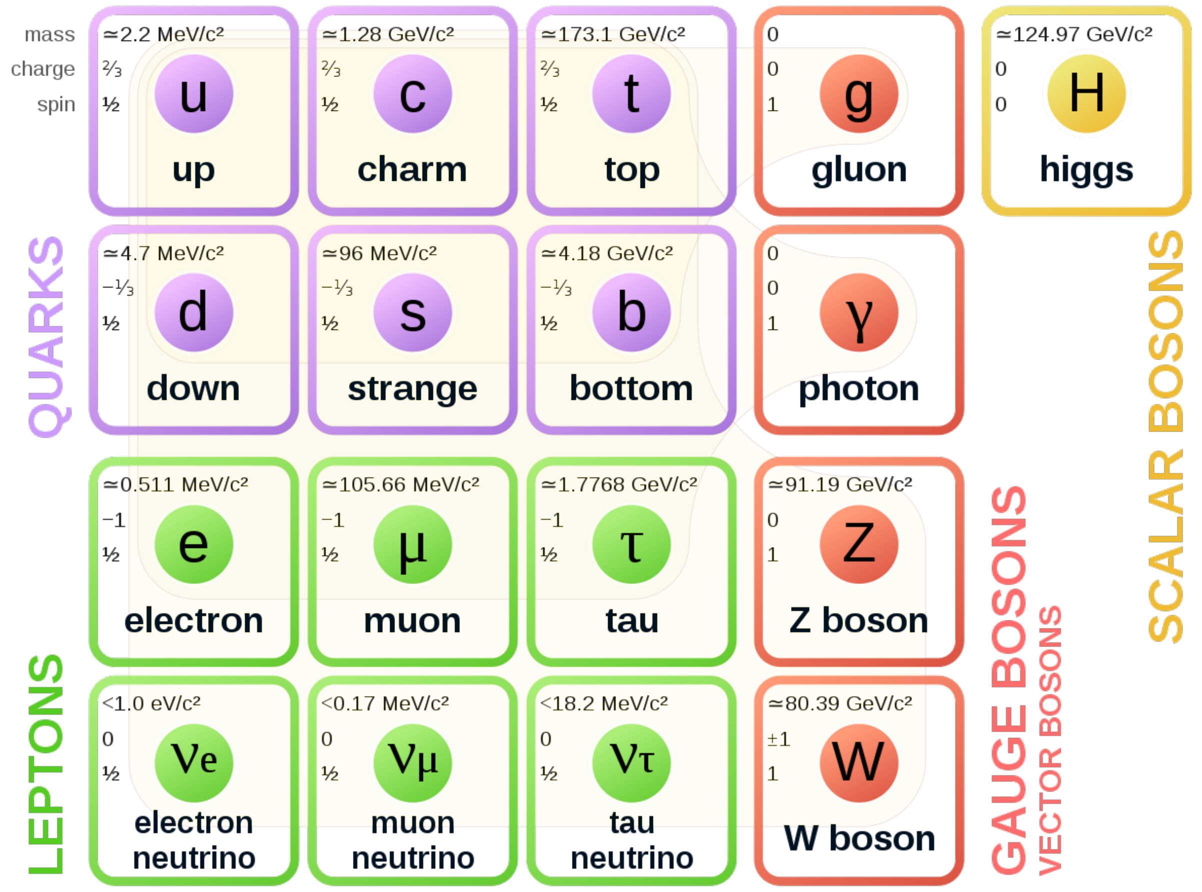
CMS detector



CMS detector

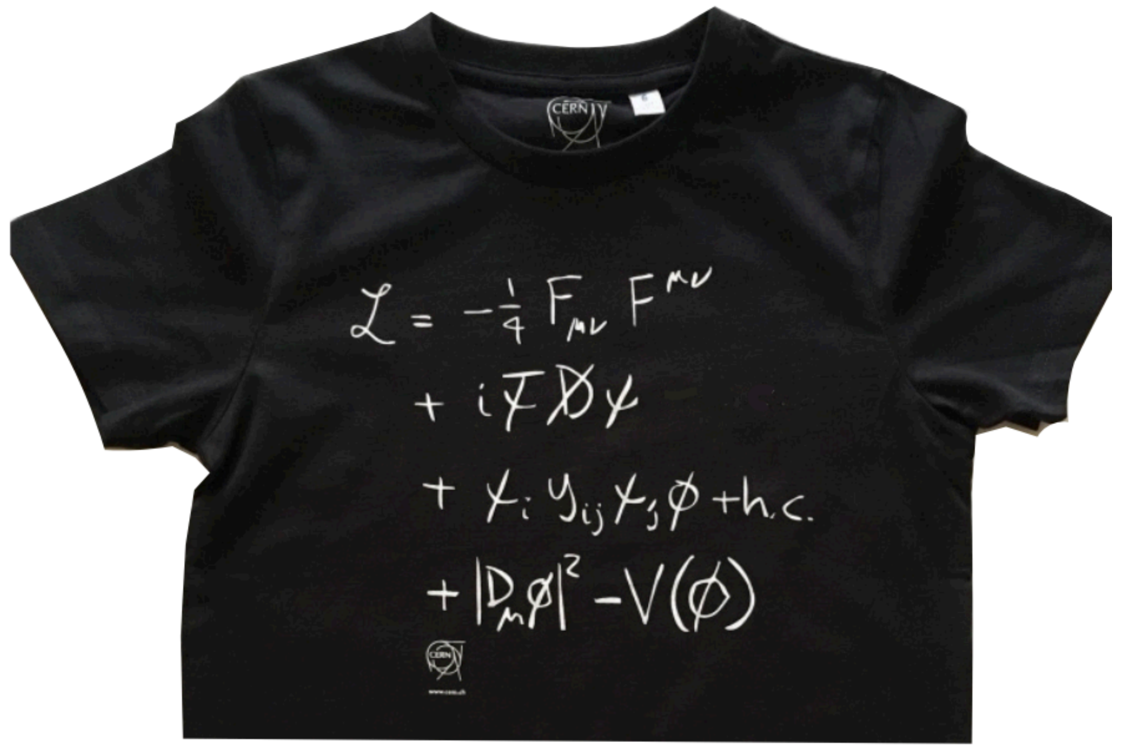
Underlying physics governed by Standard Model (and beyond?)

Particles

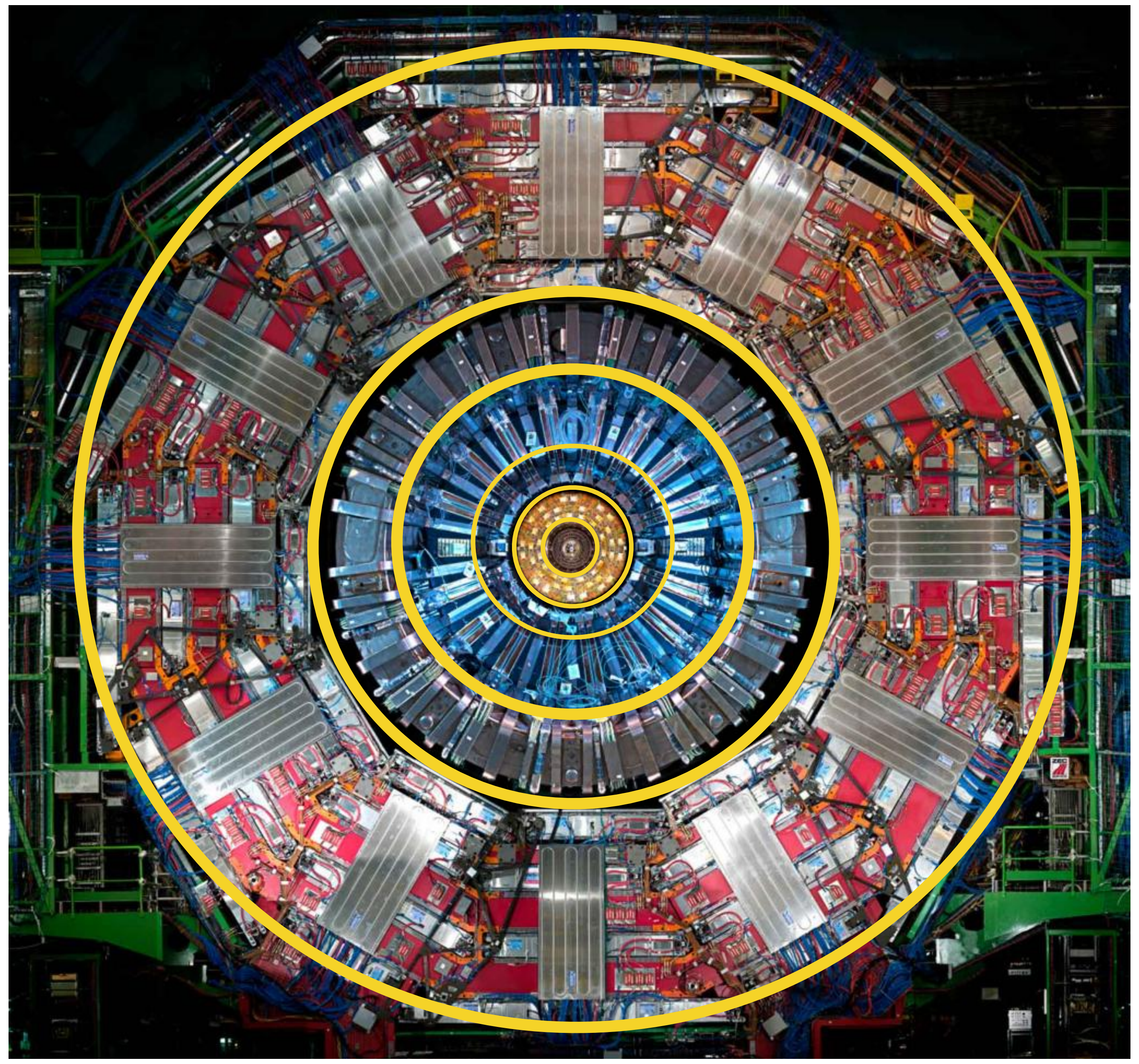


+

Interactions



CMS detects created particles

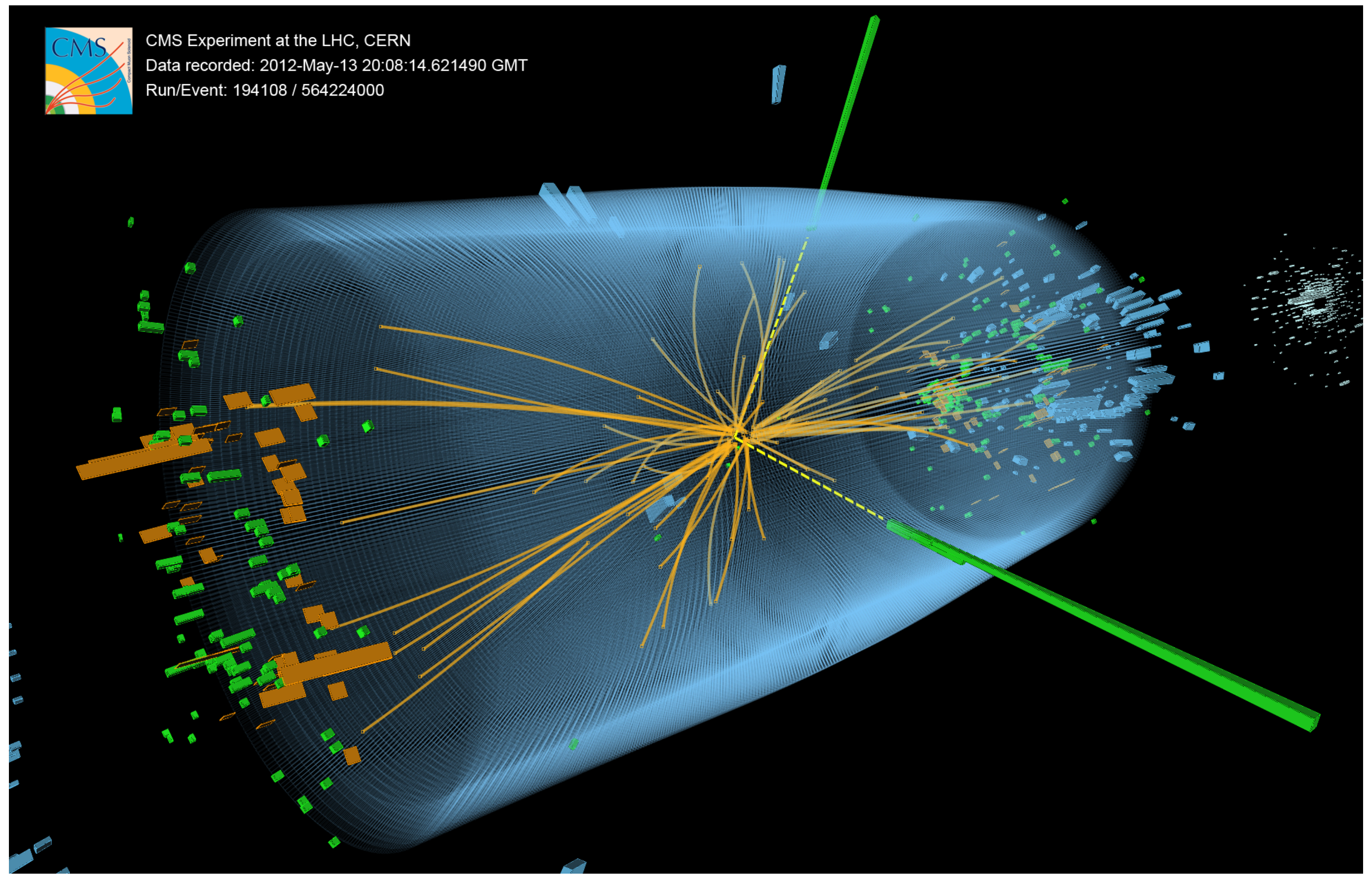


- Detector has **onion** structure, and is hermetic
- Consists of several subdetectors to detect different particles

Reconstruction of the events

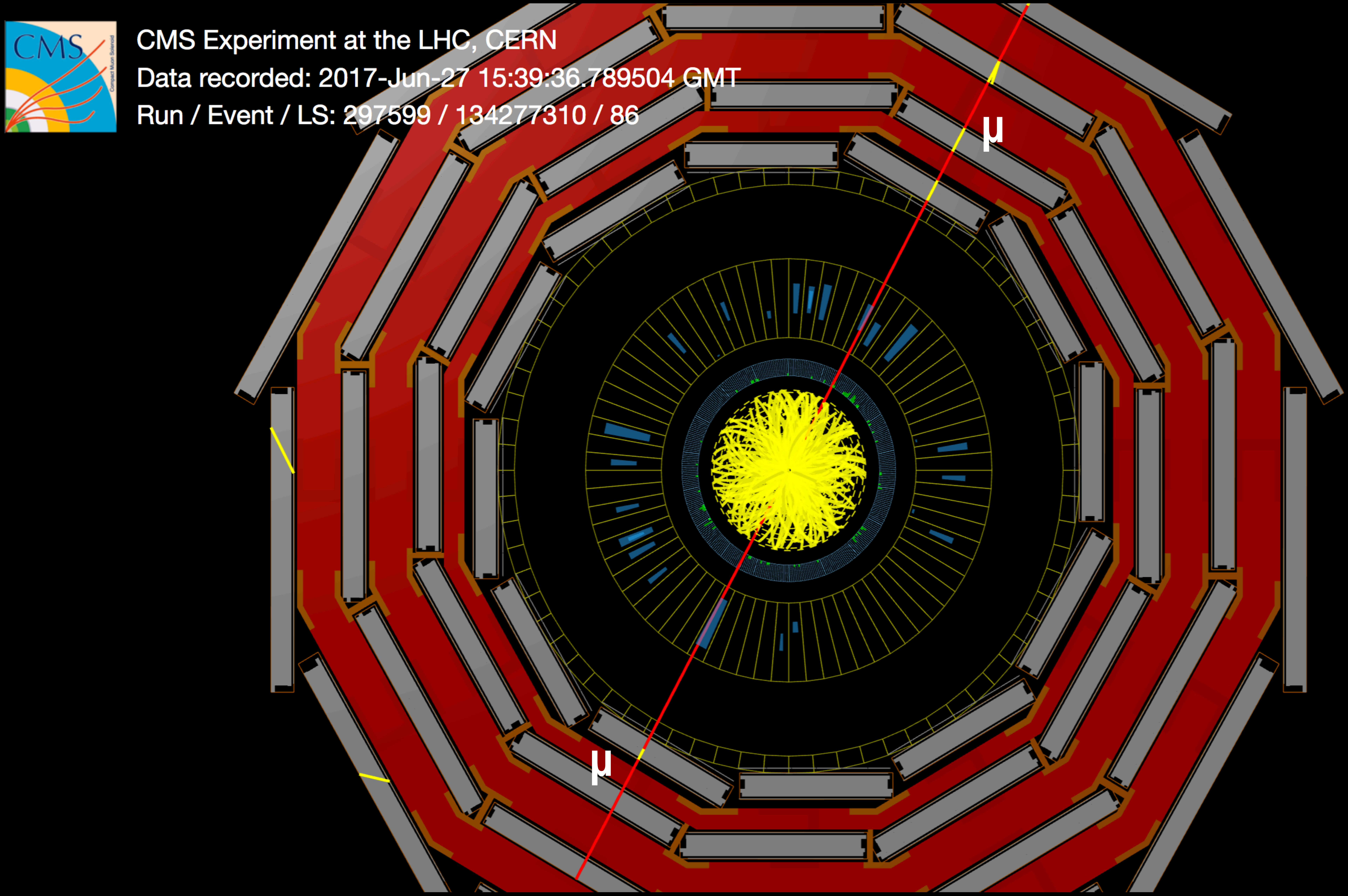
- The goal of the **event reconstruction** is to assign each energy deposit to individual particles
- From the reconstructed particles we can reconstruct full event kinematics and **infer** what **underlying physics process** led to such final state in the detector

From mess of particle hits



to a Nobel Prize

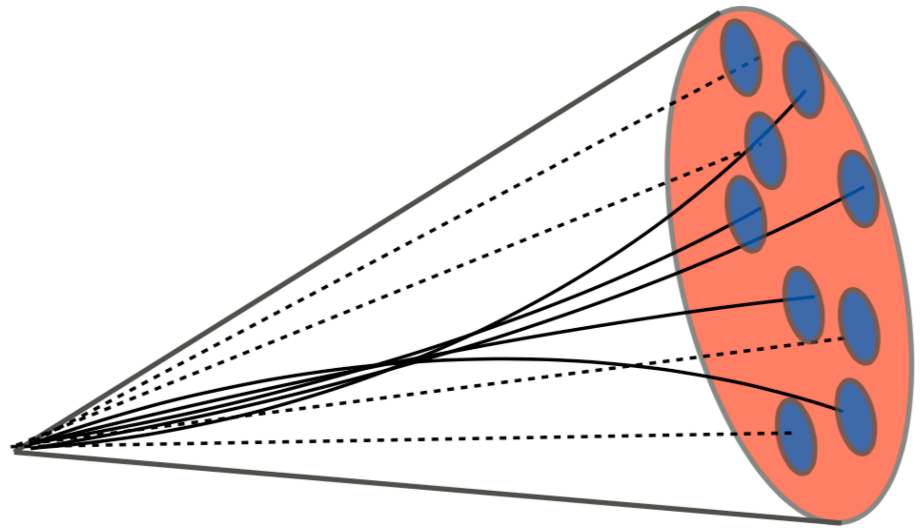
Discovery of the Higgs boson (2012)

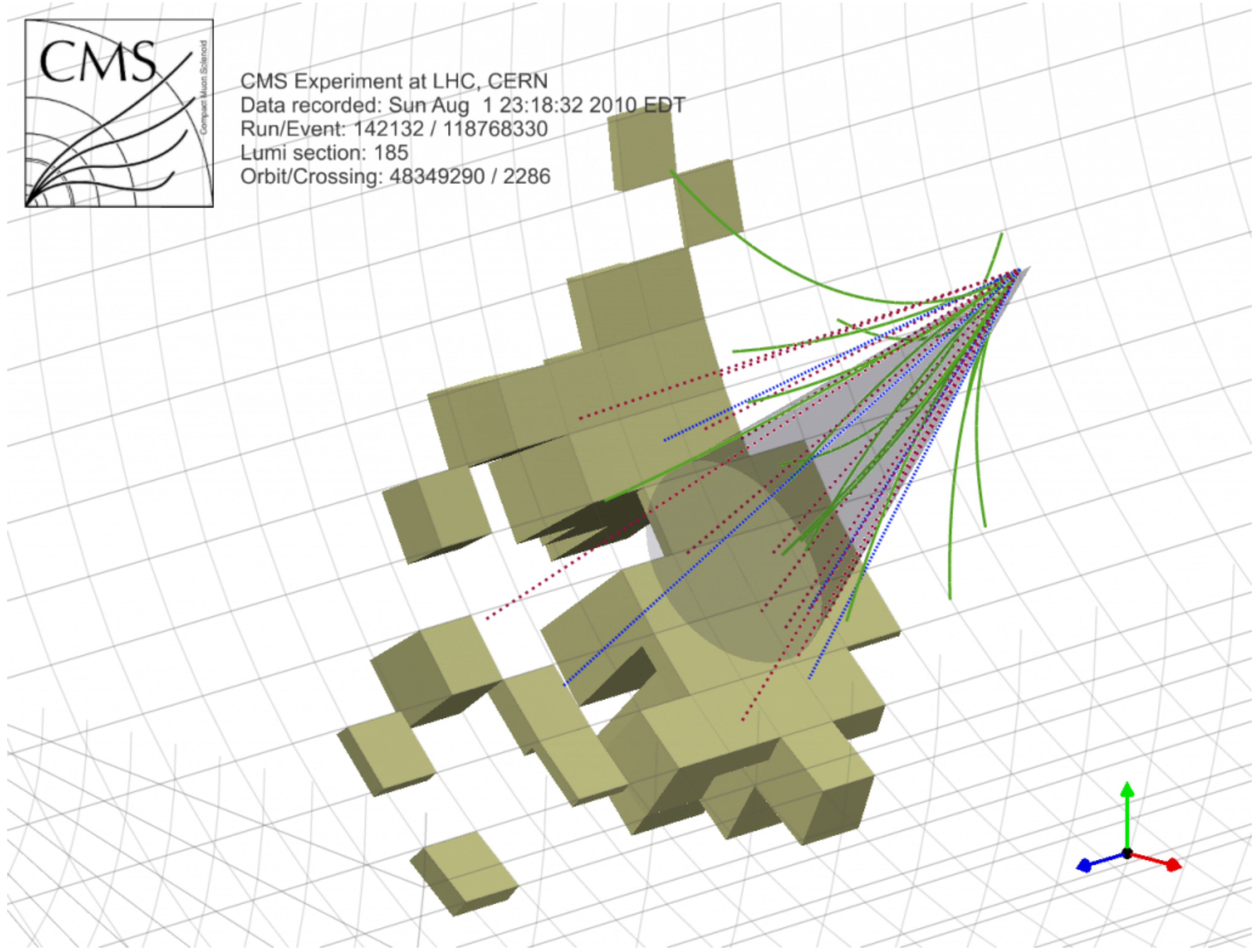

CMS Experiment at the LHC, CERN
Data recorded: 2017-Jun-27 15:39:36.789504 GMT
Run / Event / LS: 297599 / 134277310 / 86
μ
μ

Jet reconstruction

Some particles are harder to reconstruct than others :

- Quarks and gluons cannot exist as free particles and when produced they create sprays of tens of particles called **jets**





- Jets can be reconstructed from energy deposits and tracks

b jets and problem formulation

Jets arising from b quarks (**b jets**) are *challenging to reconstruct because*:

- b jets often decay to a final state with a **neutrino**, a particle with such a feeble interaction that it **leaves the detector undetected**
- Originate from secondary displaced vertex
- b jets tend to spread radially over a wider area than other light jets. This often **leads to a leakage of energy** outside of the jet clustering region

These properties of b-jets lead to **an underestimation of the b jet energy and degradation of its resolution.** However, **b jets are important** for many LHC physics analysis.

The better we can reconstruct the b-jet energy and estimate their resolution, the more sensitive we are to interesting physics!

Idea : implement a multidimensional regression to infer the true b-jet energy from the reconstructed detector information

b-jet energy regression in CMS :

- Implemented in a Deep Neural Network
- Trained on a large set of 10^8 MC simulated b jets
- Developed to **improve resolution of b jets based on their composition and properties**
- Improvement brought by this regression **helped to reach the milestone observation of Higgs decay to bottom quarks H → bb**

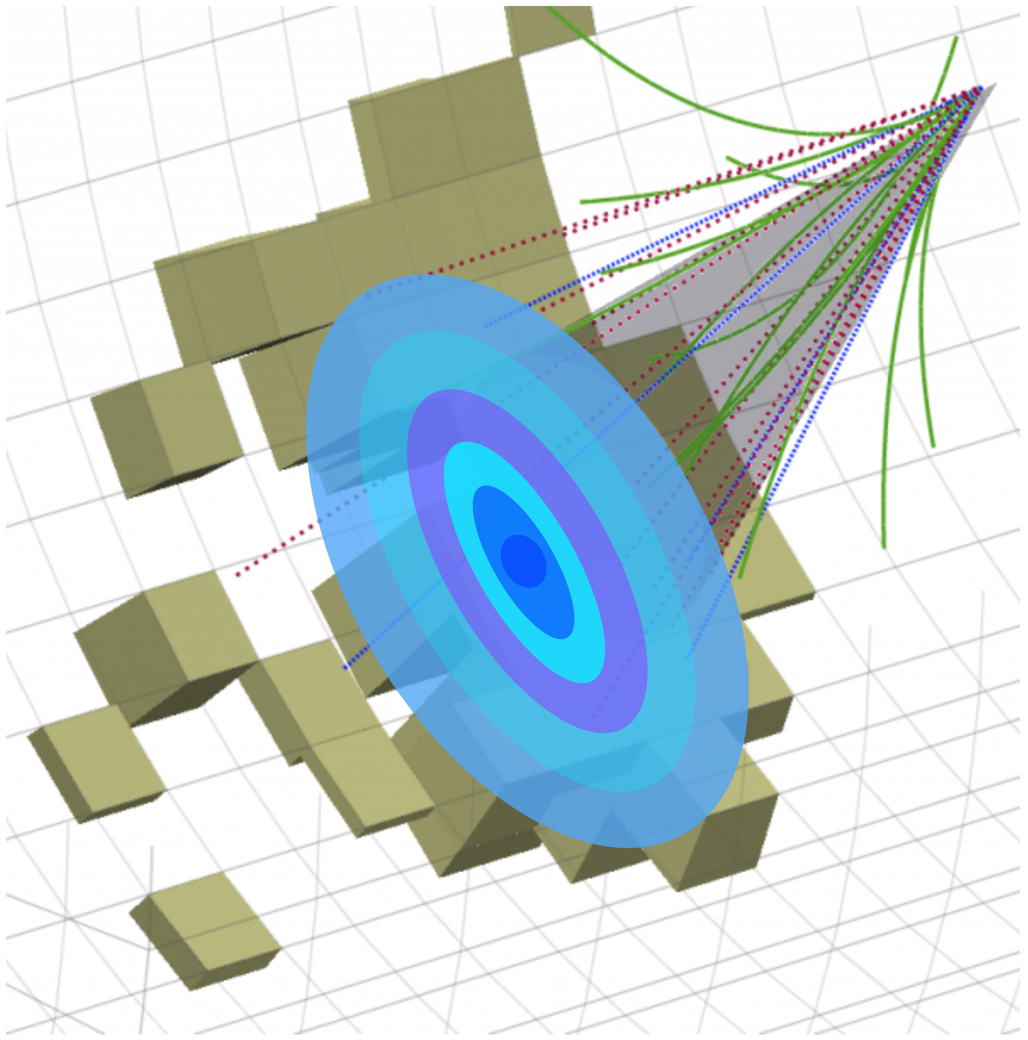
Higgs boson decay to bottom quarks

Phys. Rev. Lett. 121 (2018) 121801

b-jet energy regression in CMS

Multidimensional regression : infer true b-jet energy from the reconstructed detector information.

MC generate 100 M b jets and pass them through detector simulation.



- **DNN inputs** (simulated MC jets passed through detector simulation) :
  - Combine information about jet's :
    - reconstructed kinematics
    - constituents : tracks, secondary vertices, and individual energy deposits reconstructed by the different subdetectors
    - composition and jet shapes : energy fractions carried by constituents (electrons, photons, charged hadrons, neutral hadrons, muons)
- **DNN target** (simulated MC jets original energy):
  - MC truth b-jet energy with included 'missing energy' from the undetected neutrinos divided by the detector reconstructed energy p_T^{gen}/p_T^{reco}

Loss function for DNN regression

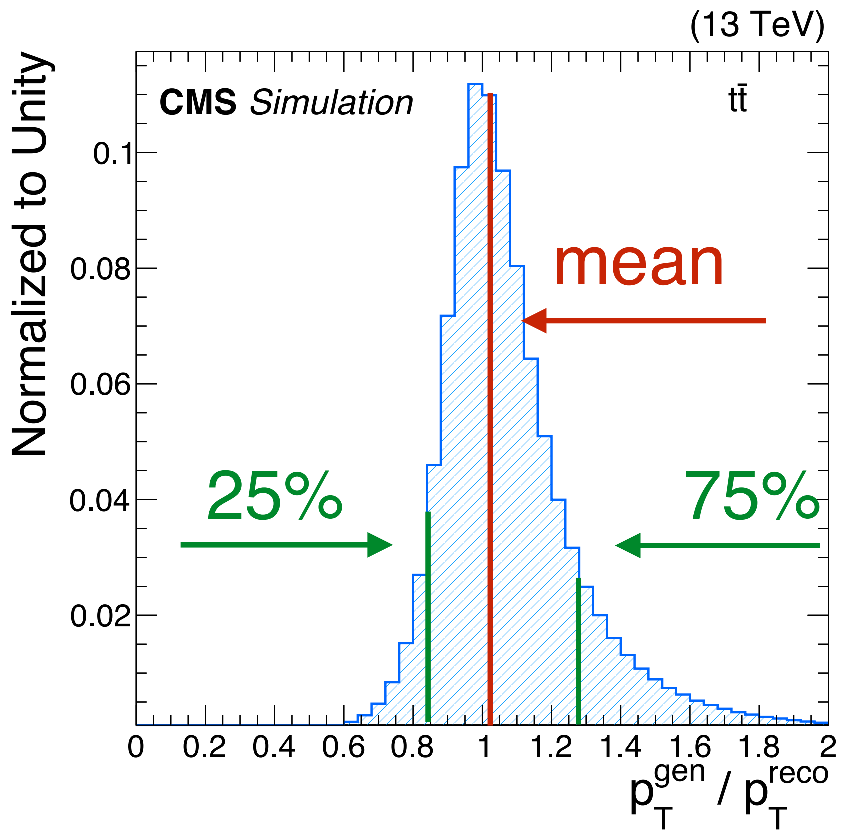
- Regression task : **energy correction** to improve resolution and provide a **jet resolution estimator per-jet**
- Regression target $y = \dfrac{p_T^{gen}}{p_T^{reco}}$, mean estimator $\hat{y}$, $z = y - \hat{y}$
- To get energy correction we use the Huber loss :

$$H_\delta(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| < \delta; \\ \delta \cdot |z| - \frac{1}{2}\delta^2, & \text{otherwise,} \end{cases}$$

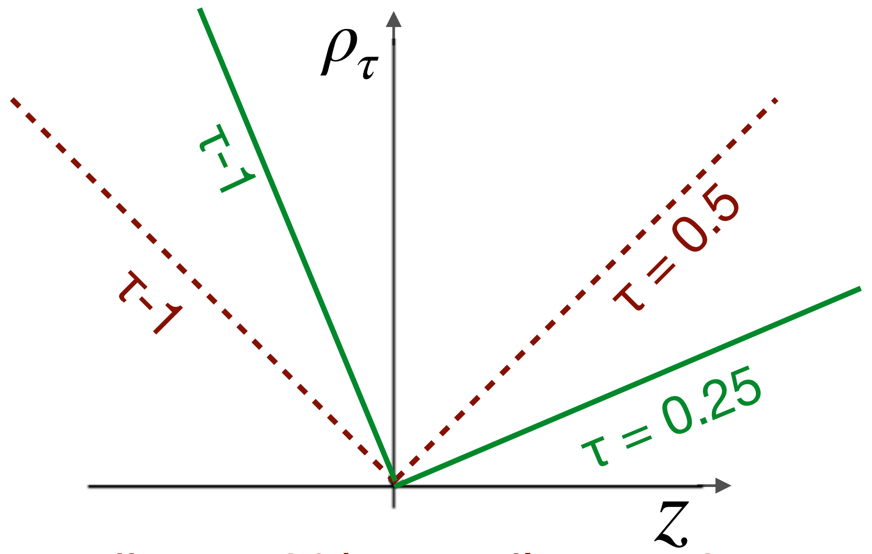
- As resolution estimator use two quantile loss functions for 25% and 75% quantiles, τ - quantile :

$$\rho_\tau(z) = \begin{cases} \tau \cdot z, & \text{if } z > 0; \\ (\tau - 1) \cdot z, & \text{otherwise,} \end{cases}$$

Resolution distribution



Quantile loss $\rho_\tau(z)$



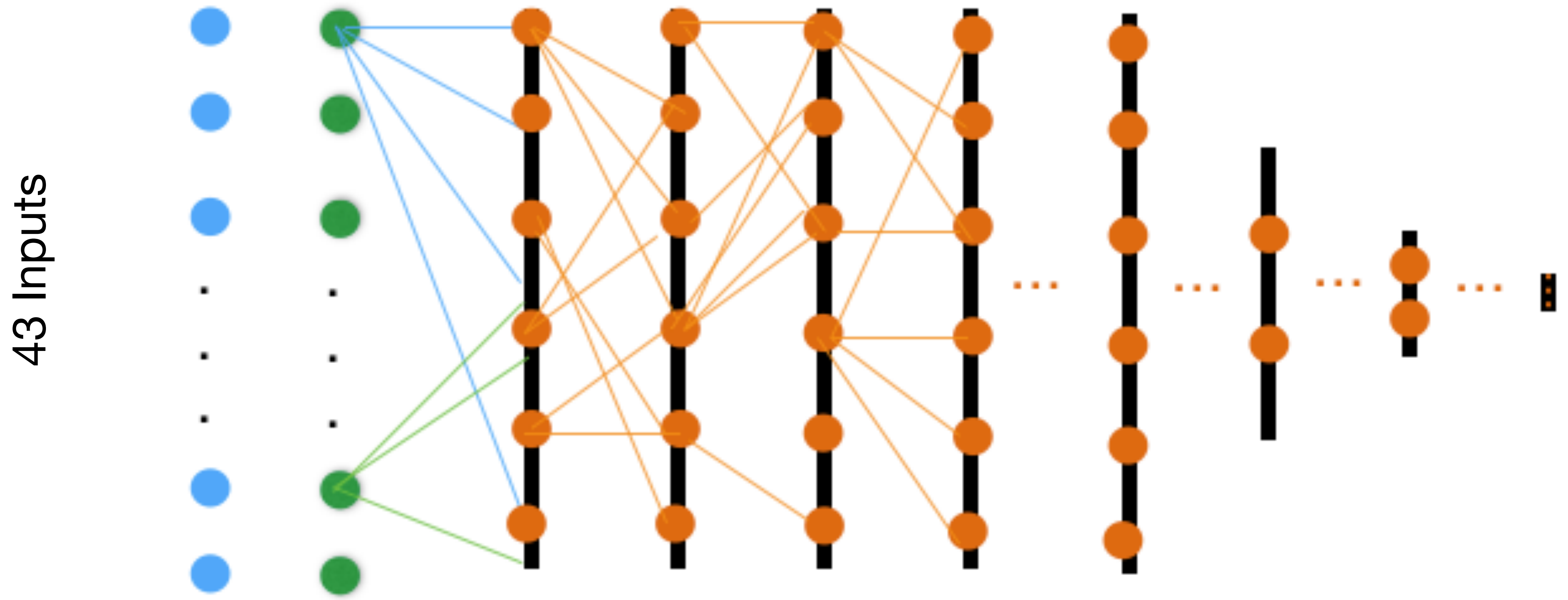
median - 50% quantile : τ = 0.5
25% quantile : τ = 0.25

Joint loss function for correction (Huber) and resolution (quantiles) :

$$\boxed{Loss = H_1(y - \hat{y}(x)) + \rho_{0.25}(y - \hat{y}_{25\%}(x)) + \rho_{0.75}(y - \hat{y}_{75\%}(x))}$$

DNN architecture

DNN architecture : Feed-forward fully connected DNN



3 Outputs

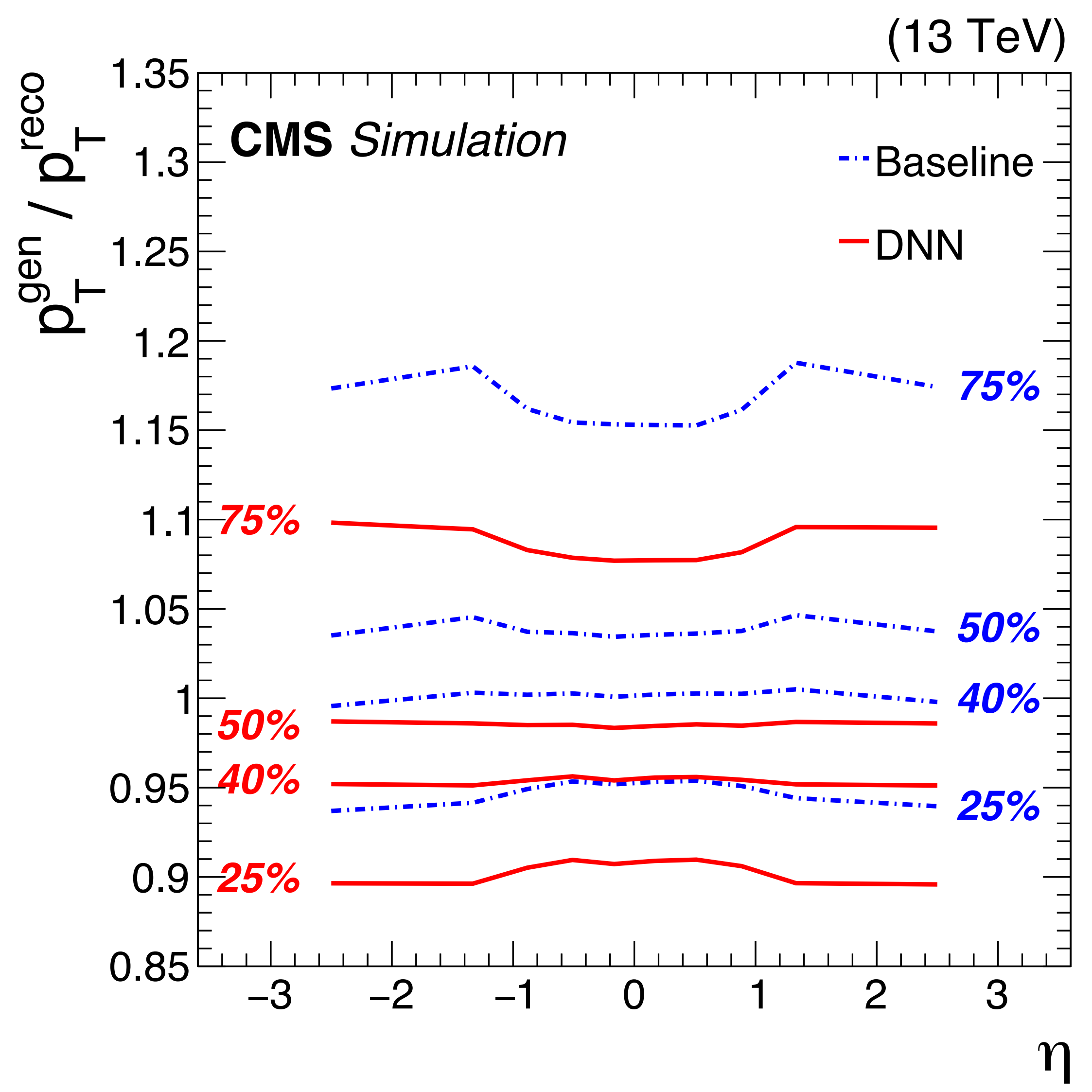
- $\hat{y}$ - correction
- $\hat{y}_{25\%}$ - 25% quantile
- $\hat{y}_{75\%}$ - 75% quantile

resolution estimator $\hat{s} = \dfrac{\hat{y}_{75\%} - \hat{y}_{25\%}}{2}$

- DNN is implemented in Keras with TensorFlow backend
- Back-propagation using stochastic gradient descent with Adam optimizer
- Hyperparameters and architectures were optimized using randomized grid search
- 6 layers with # neurons : [1024, 1024, 1024, 512, 256, 128]
- The network was trained on a single NVIDIA GeForce GTX 1080 Ti

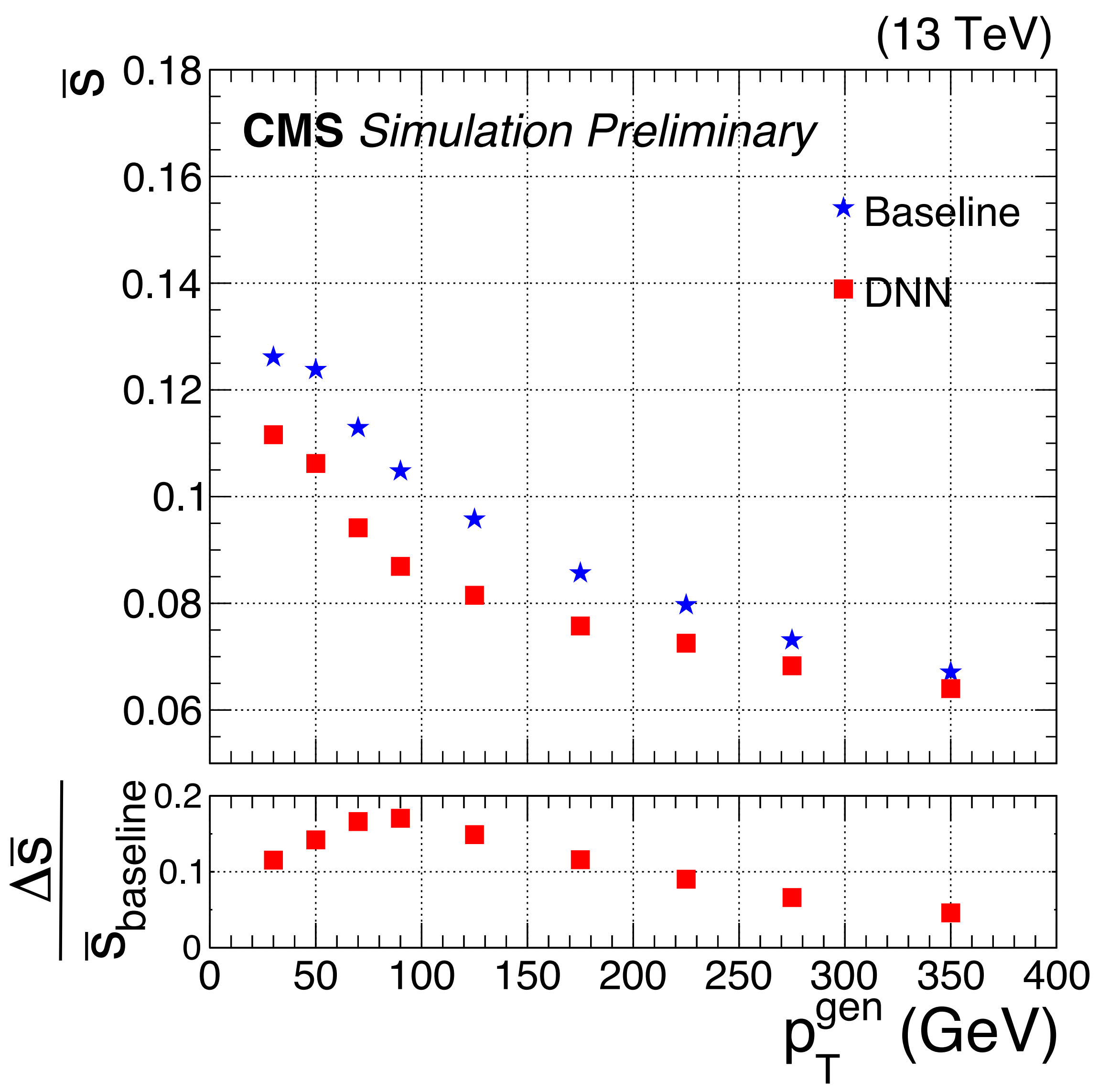
- Evaluate b-jet energy scale p_T^{gen}/p_T^{reco} after the application of the regression correction as a function of jet kinematics (quantiles 25%, 40%, 50%, 75%)

- Compare to before-regression p_T^{gen}/p_T^{reco}
  - narrower distributions
  - flatter response



Quantify relative resolution improvement:

- Relative resolution estimated as $\bar{s} = \dfrac{s}{q_{40\%}} = \dfrac{q_{75\%} - q_{25\%}}{2q_{40\%}}$
- After regression **per-jet** relative resolution is improved by **~13%**
- Very similar performance achieved for b jets arising from different physics processes



Resolution estimator

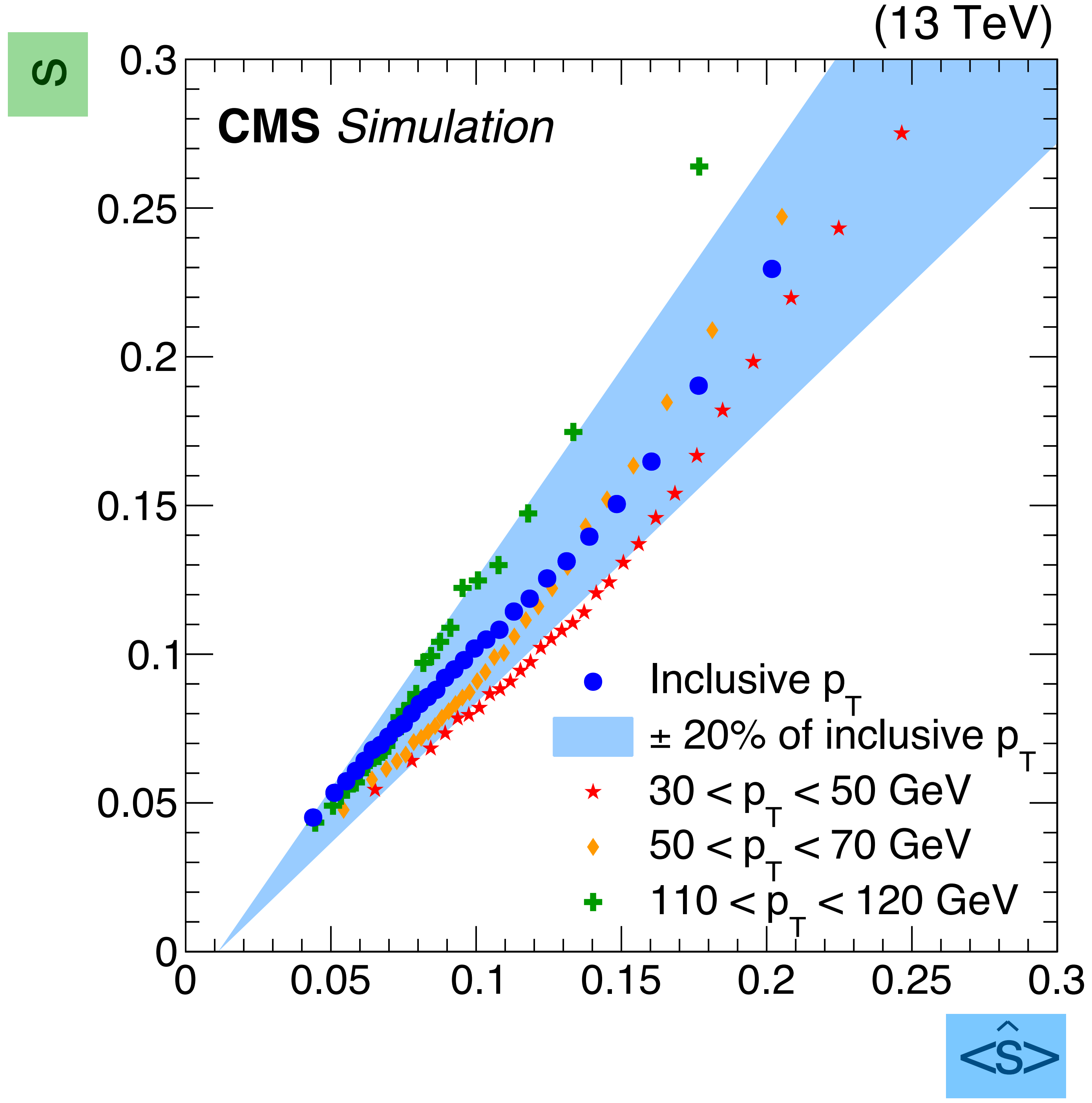
- **Knowledge of jet resolution on a jet-by-jet basis** can be exploited in physics analyses searching for resonant production of b jet pairs to **increase their sensitivity**
- Therefore, it is **important** that the resolution estimator provided as an output by our DNN *correctly represents* jet resolution

Check:

- Split the sample of jets into several equidistant quantiles of jet resolution estimator $\hat{s}$
- In each bin quantify the jet resolution $s = \frac{q_{75\%} - q_{25\%}}{2}$ using MC truth information
- Check if the two correspond to each other
- Repeat the same test in the bins of jet momentum p_T

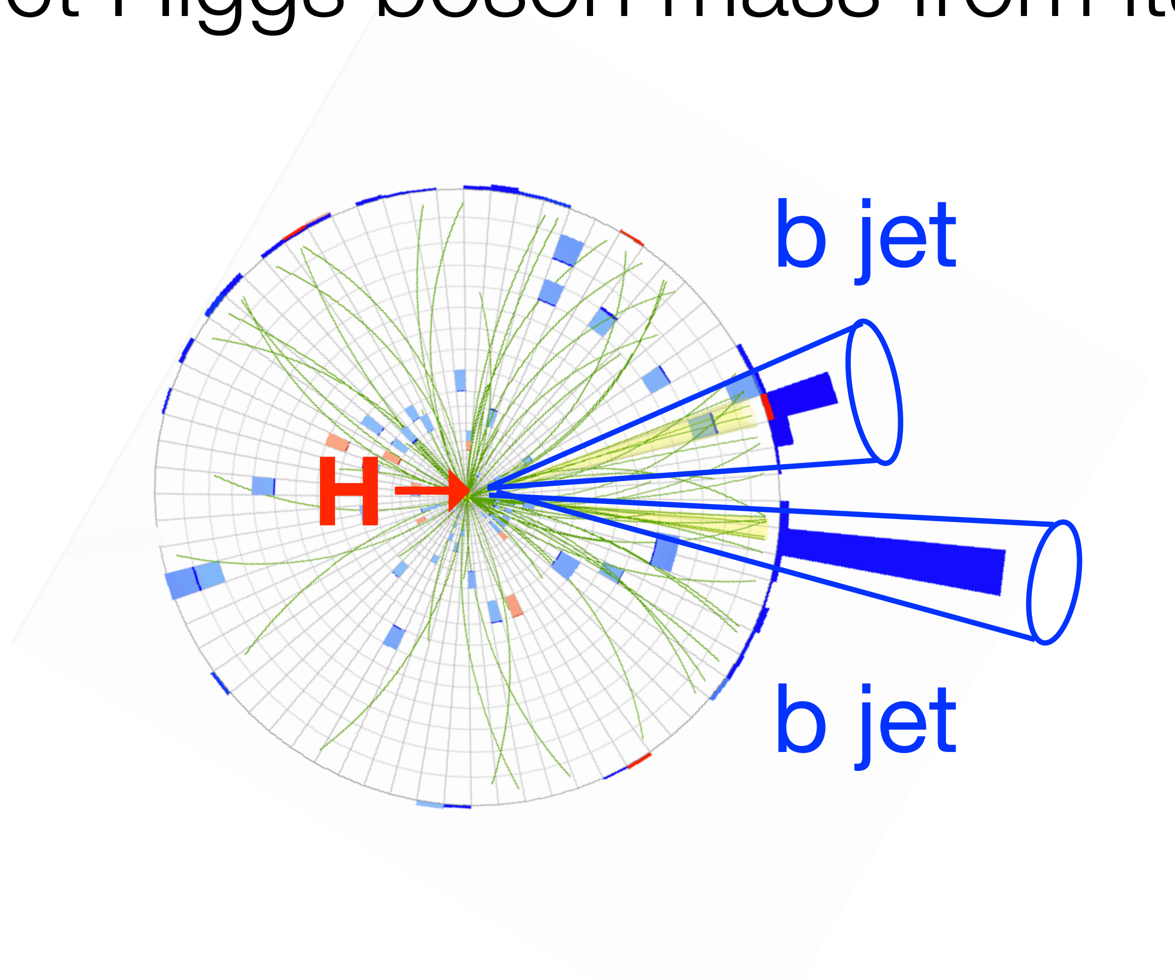
Linear dependence confirms that our resolution estimator $\hat{s}$ correctly represents the jet resolution s

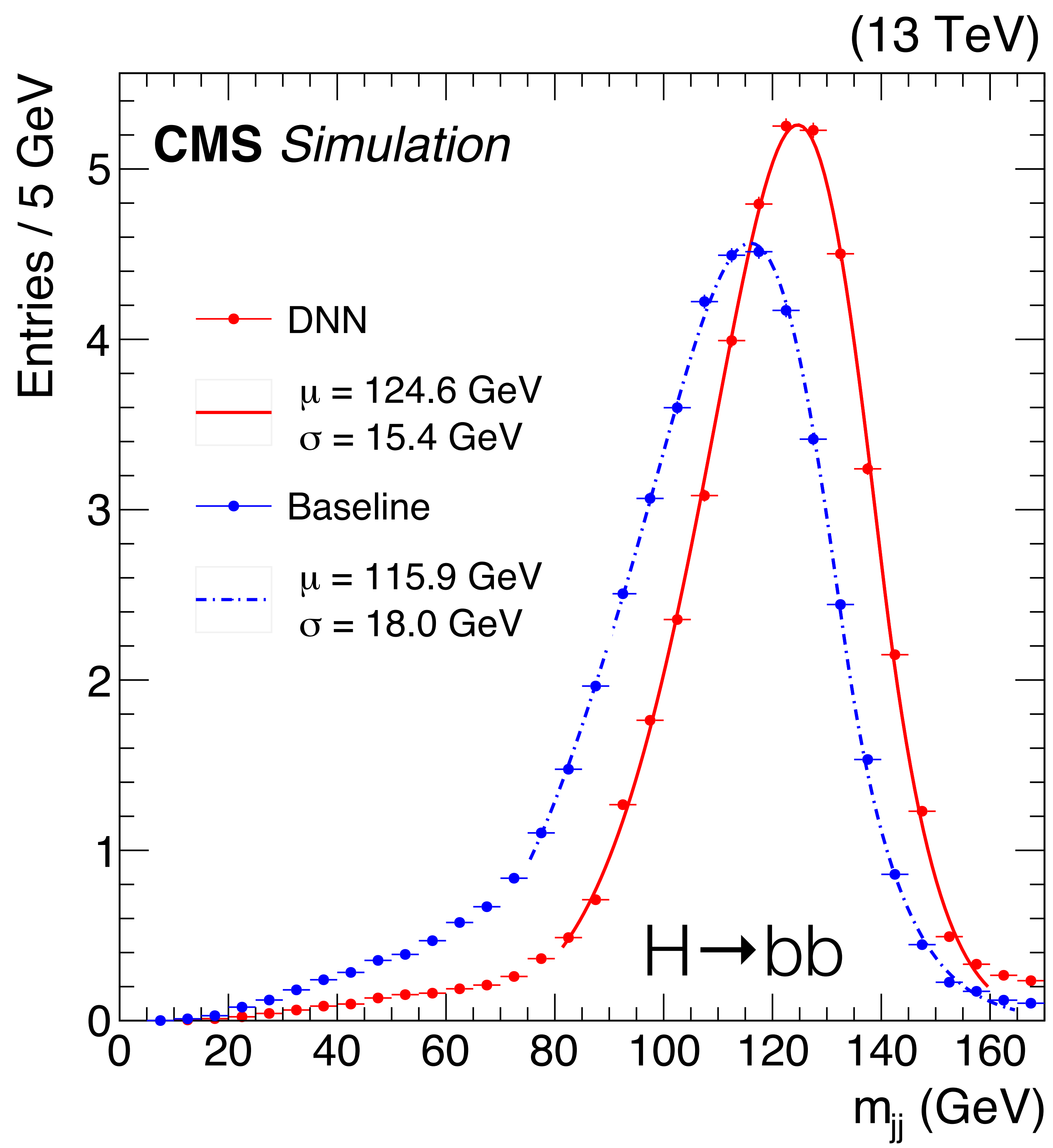
Correlations between s and $\hat{s}$



Dijet resolution improvement

- Many physics analyses use mass of a particle decaying to two b jets as a discriminating variable for signal extraction
  - e.g. reconstruct Higgs boson mass from its decay to b jets : H→bb





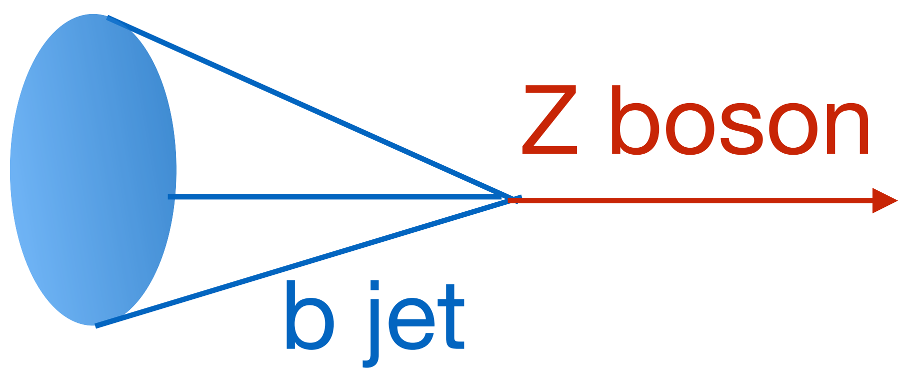
- Resolution improvement for dijet invariant mass is larger than for a single jet

Significant improvement that helped to reach observation of Higgs decay to bottom quarks

20% improvement in dijet mass resolution

Validation on data

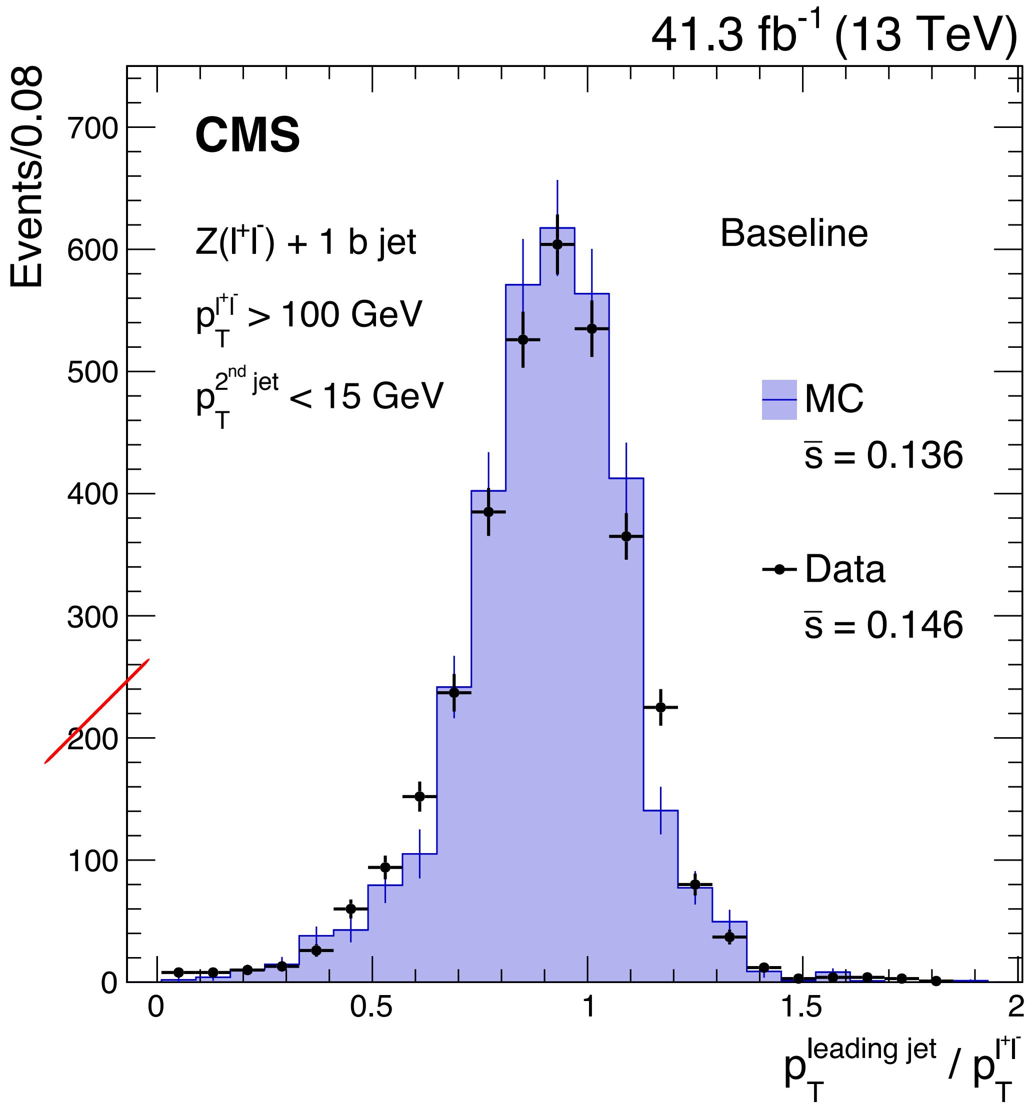
- Can this good performance be **transferred** from MC to the **domain of LHC data** ?
- Select a high purity sample of events with a well reconstructed Z boson (leptonic decay) and b jet in data
- In such an event topology the Z boson and b jet are produced **back-to-back** and the better the b-jet resolution, the narrower the p_T balance distribution is



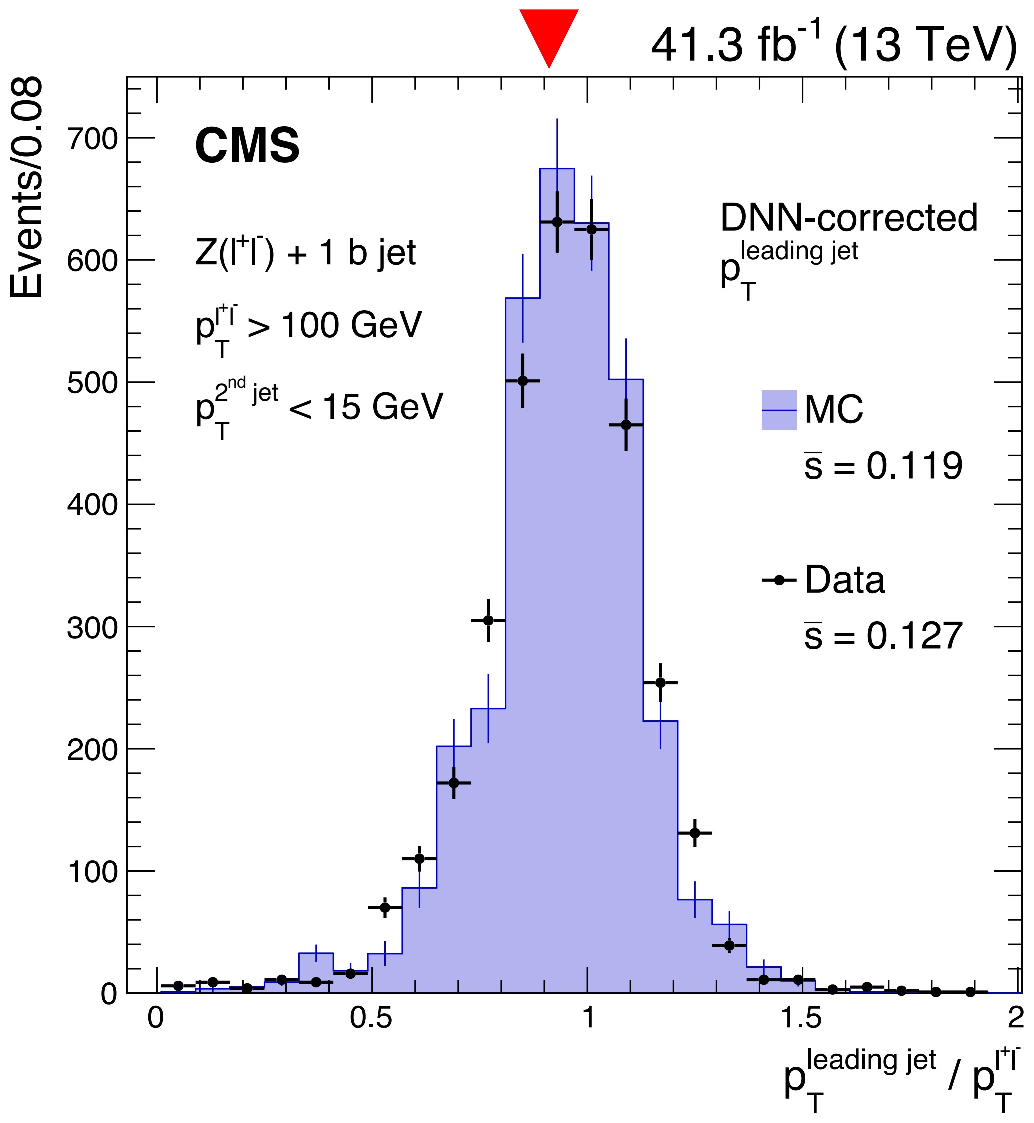
- Performance in data evaluated with p_T balance $= \dfrac{p_T^{b\ jet}}{p_T^{Z}}$

- **Resolution improvement** is consistent for MC and data and is **13 %**

Resolution improvement achieved in MC is successfully transferred to the data domain!



b-jet regression

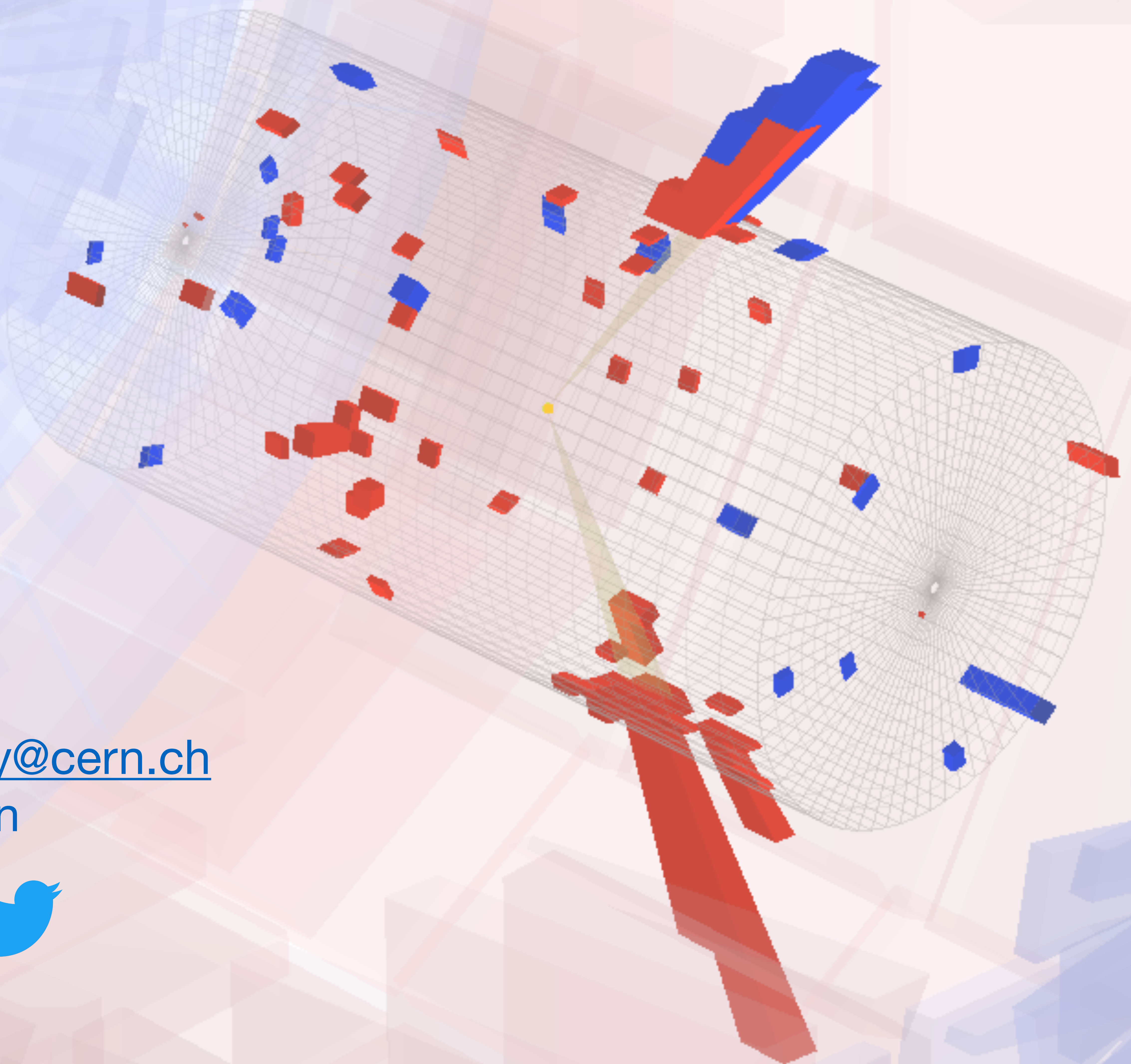


Summary

- We developed DNN based b-jet energy regression for the CMS experiment
- b-jet regression was trained using jet structure and composition information, and outputs energy correction and jet resolution estimator
- The technique was **validated on data** recorded by CMS at the LHC
- The regression was successfully applied to reach the **observation of Higgs boson decay to bottom quarks** ***Phys. Rev. Lett. 121 (2018) 121801***
- Paper focusing on this regression is submitted to Computing and Software for Big Science, ***CMS-HIG-18-027*** **and** ***arXiv-1912.06046***

CMS Experiment at LHC, CERN
Data recorded: Tue May 5 11:05:27 2015 CEST
Run/Event: 243484 / 35552557
Lumi section: 50
Orbit/Crossing: 12904927 / 208

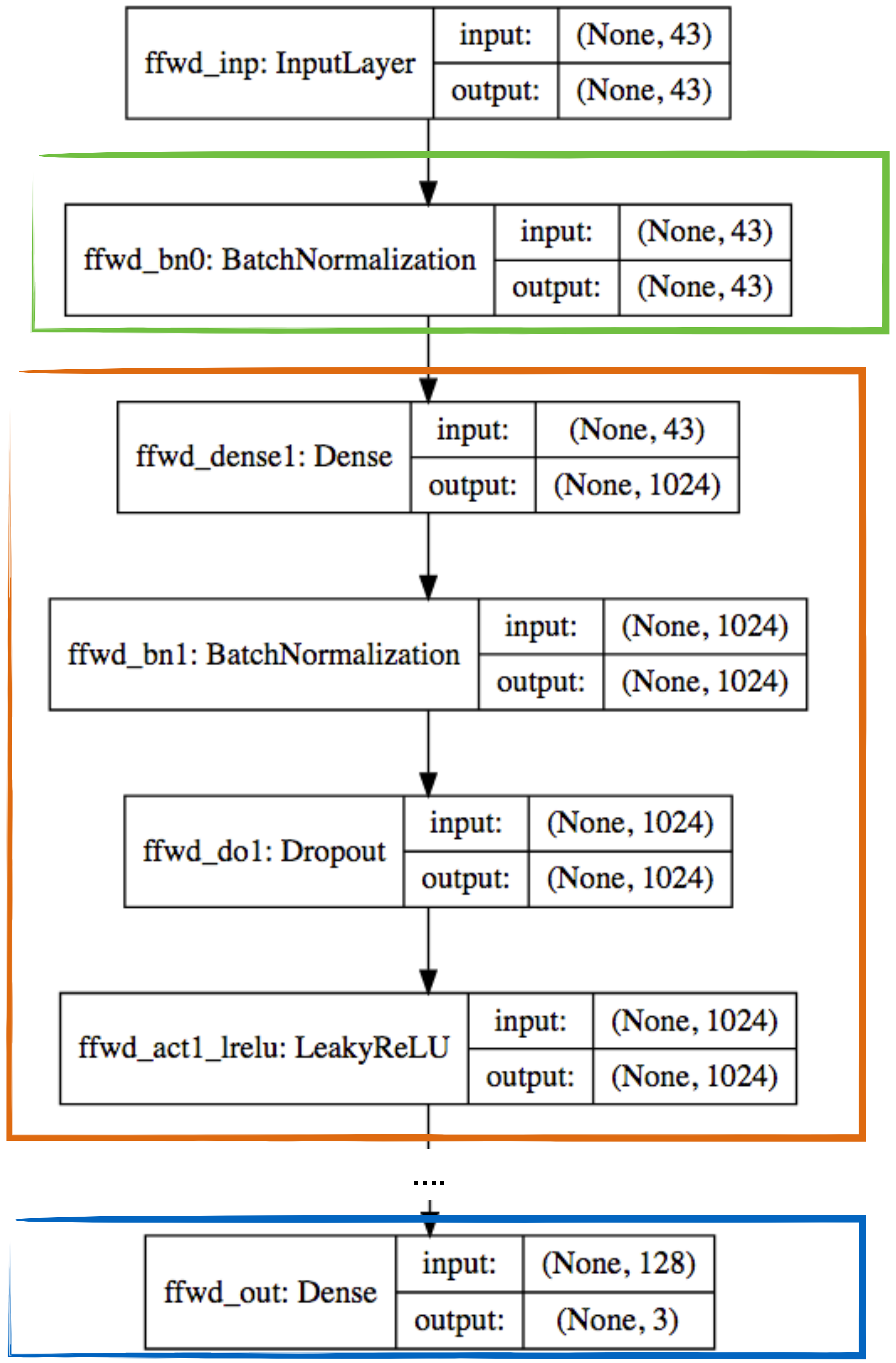
Thank you!



nadezda.chernyavskay@cern.ch
@NadyaChern

Additional Material

DNN architecture



DNN architecture : Feed-forward fully connected NN

- Input layer
- Batch normalization → internal data standardization

- Each hidden layer has 4 operations :
  - Linear transformation
  - Batch normalization
  - Dropout
  - Non-linear activation function
    - Leaky ReLU activation with α = 0.2

....

- Output : target is standardized (to zero-mean unit-variance)