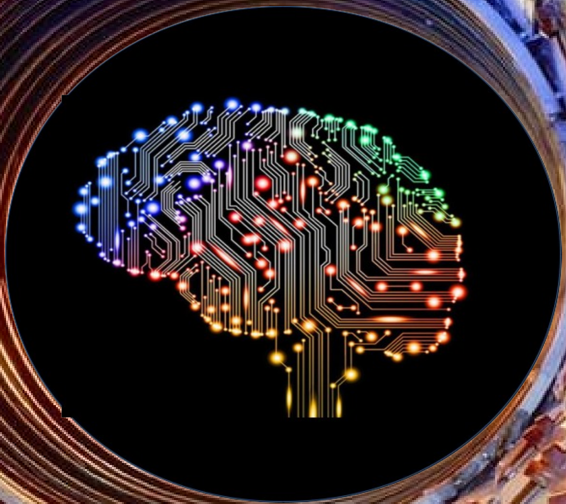
New computing paradigms in High Energy Physics

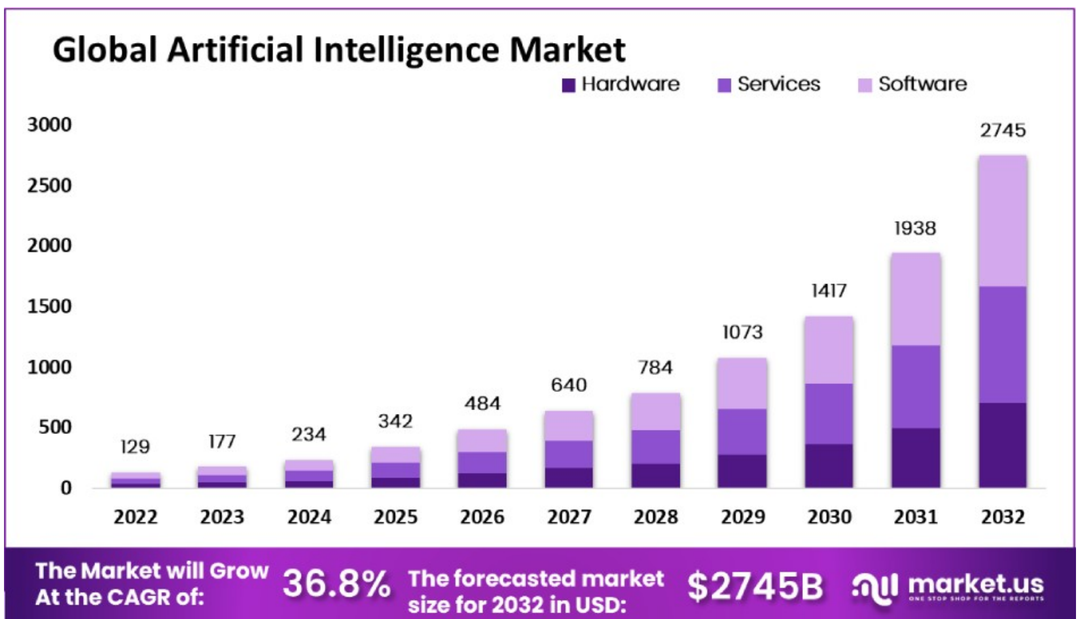
Third COMCHA School
University of Oviedo

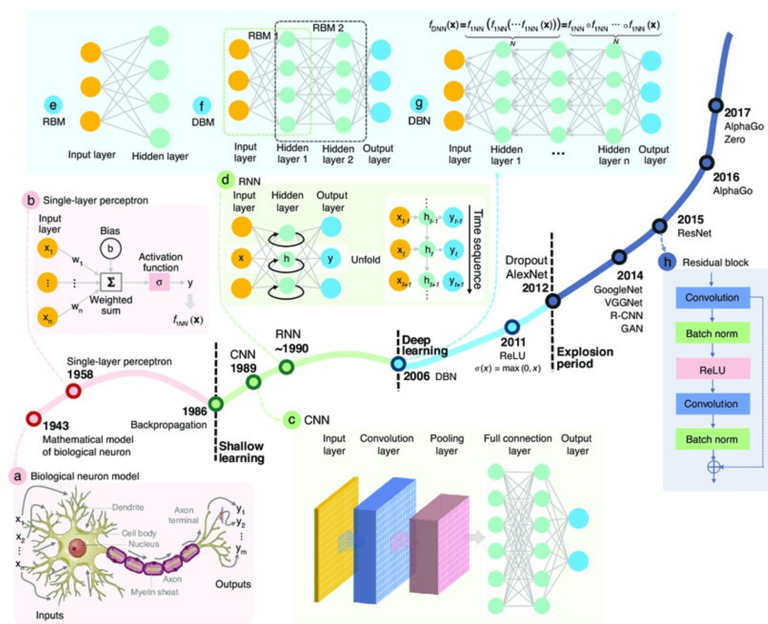
Pablo Martínez Ruiz del Árbol



A complex, rapidly growing ecosystem

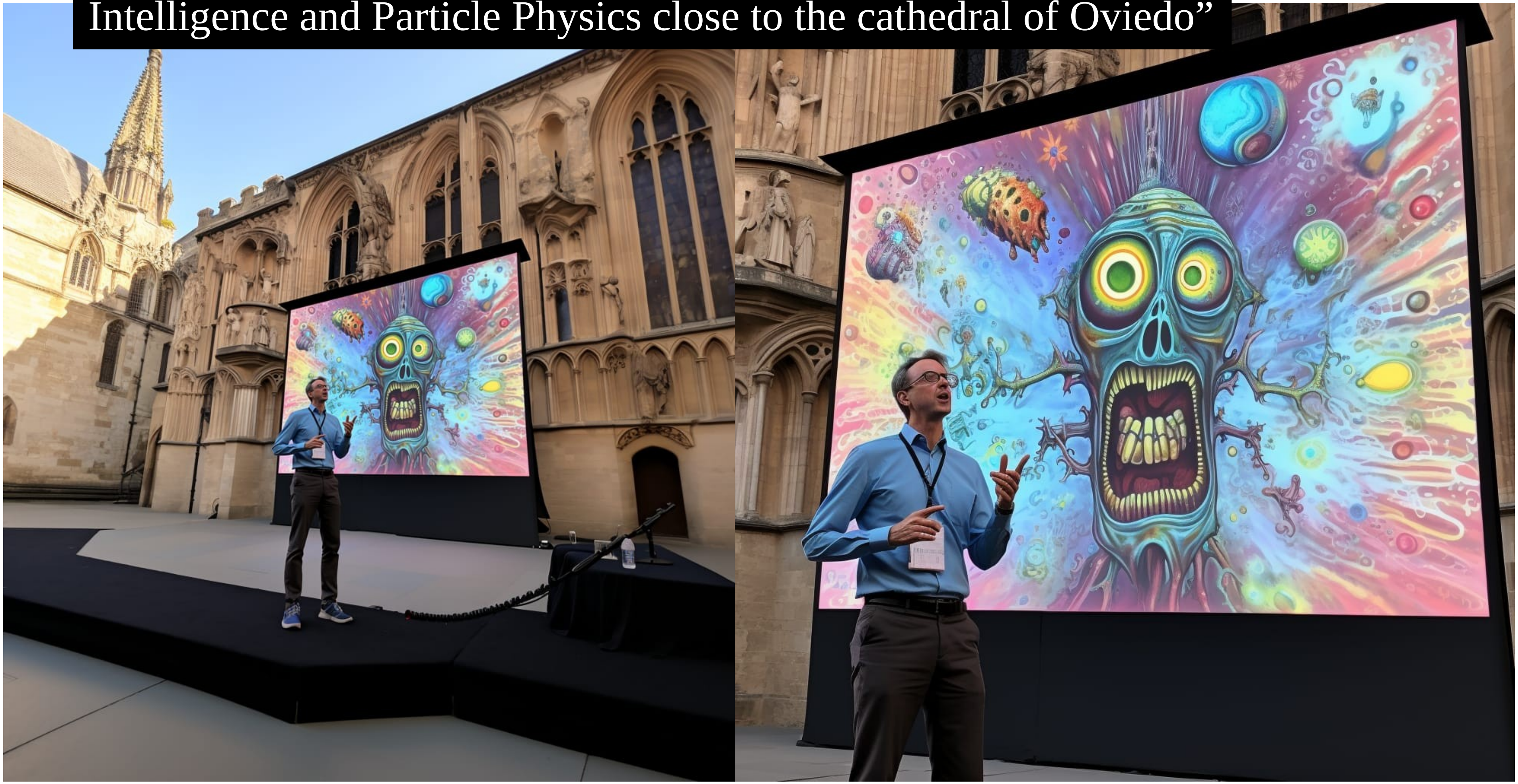
- New computing technologies are quickly emerging and evolving all over the world
- Progress on both hardware and software makes this growing heteregenous and complex
  - More hardware availability or better price leads to new software developments
  - New algorithms and solutions encourage the fight for producing more advanced hardware
- The scale of the new developments is also incredibly fast (especially in software)
  - In most cases your cutting-edge algorithm is obsolete in a matter of a few months





But never give up your brain

Midjourney prompt: “Nerd guy giving a talk on Artificial Intelligence and Particle Physics close to the cathedral of Oviedo”



Don’t ever forget that success will come by using technology **wisely** not blindly!!!

Use of new computing technologies in HEP

- Complex and rich environment of new techniques with two clear accelerators:
  - New hardware architectures (GPU, FPGA, TPU, Quantum computers)
  - Developments in AI, Deep Learning, Quantum computer, others
- Some of these techniques are being used in almost all aspects of the LHC experiments.

Simulation

- **Fast MC generation**
- **Particle/Matter interaction**
- **Detector Response**

Reco

- **Trigger**
- **Reconstruction**
- **DQM (Anomaly detection)**

Analysis

- **Event Classification**
- **Parameter Estimation**

Detector design

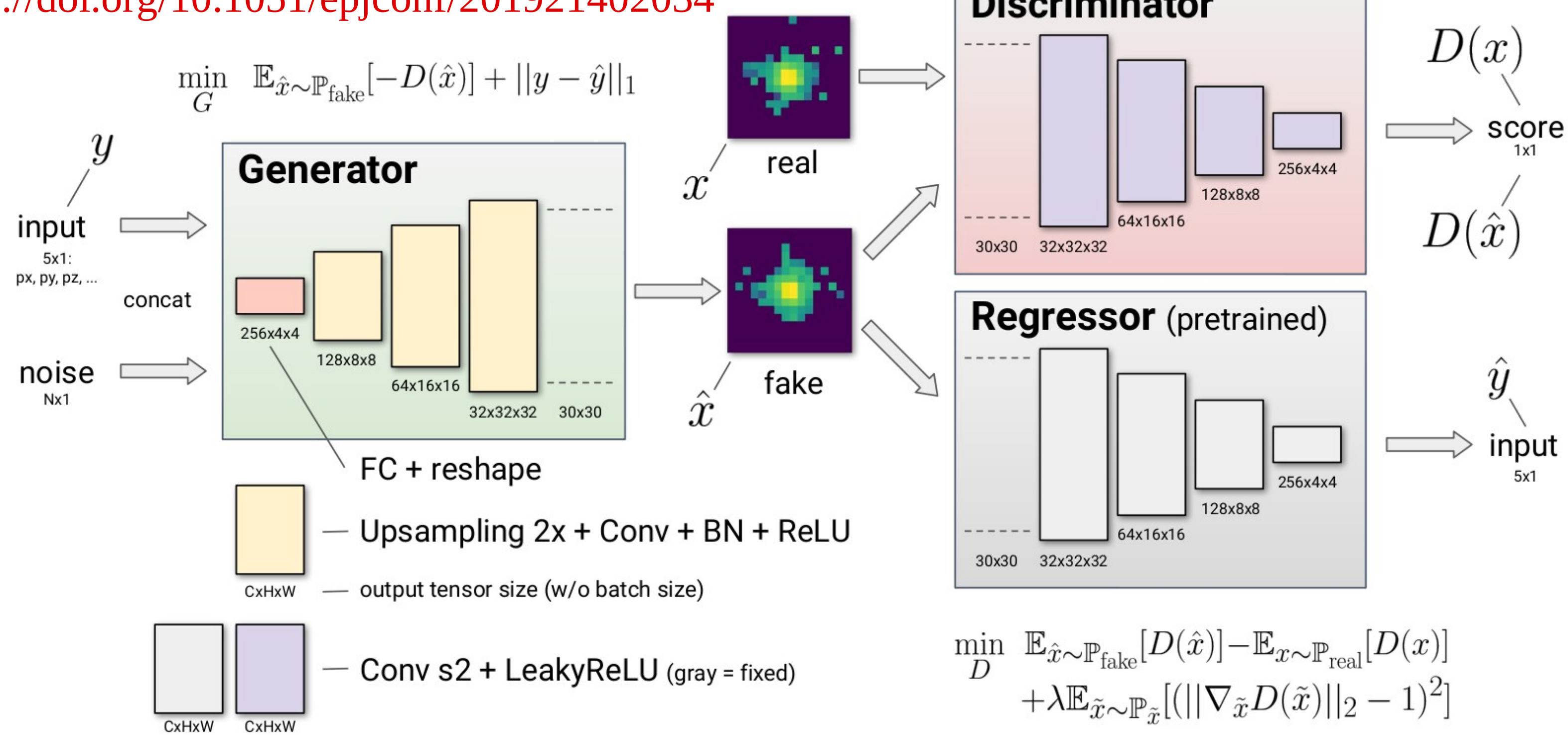
- **Optimization**

Fast simulation

GAN architectures for LHCb Fast Simulation

- This machinery aims at simulating the electron interactions in the ECAL of LHCb.
  - The system simulates the energy deposition in a 30x30 matrix of ECAL cells.
- A Wasserstein Generative-Adversarial-Neural Network is used as learning scheme
  - A regressor block is added in order to predict the momentum of the incoming particle.

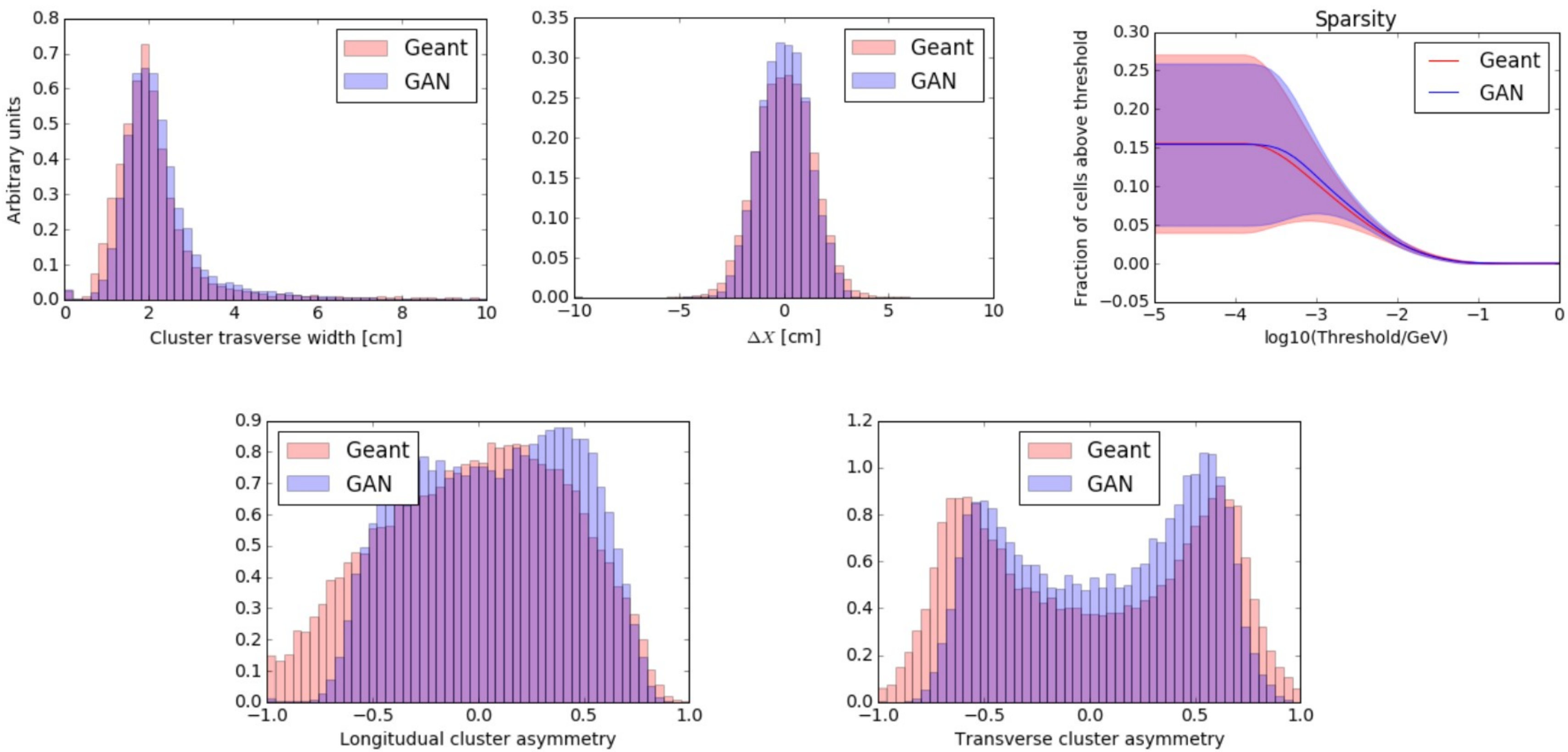
https://doi.org/10.1051/epjconf/201921402034



GAN architectures for LHCb Fast Simulation

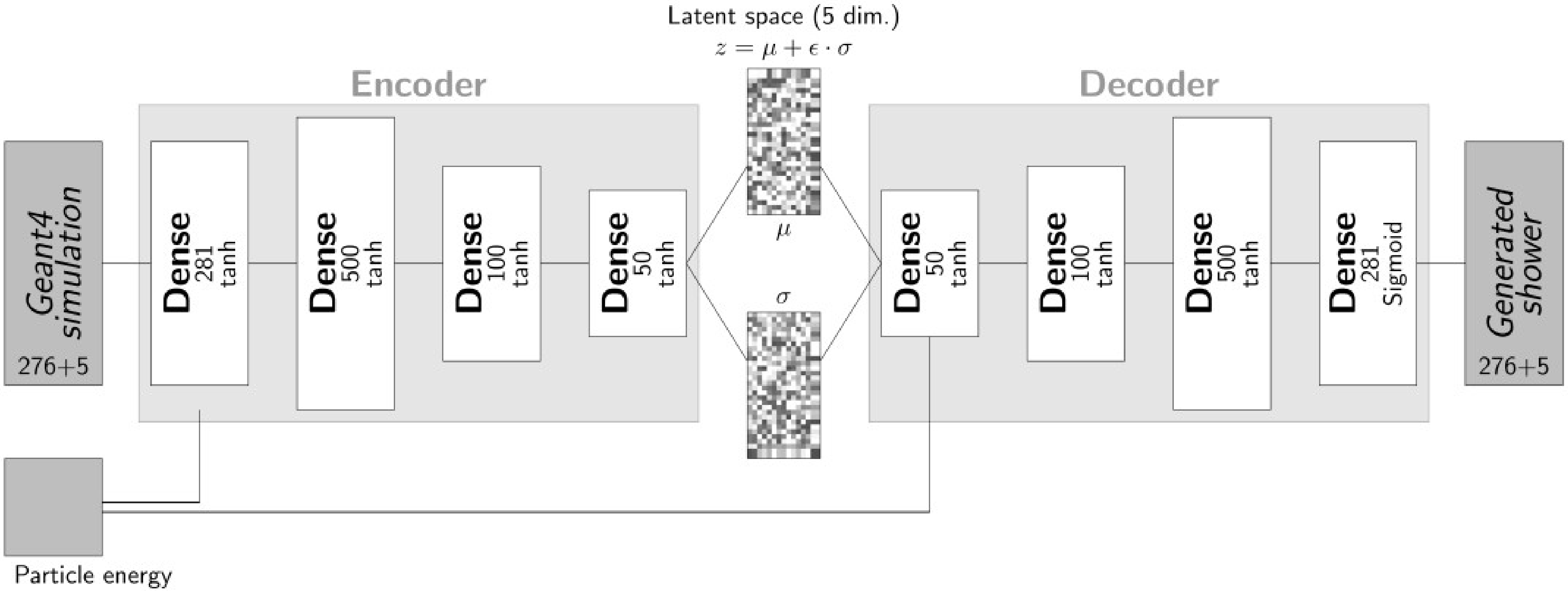
IFCA
Instituto de Física de Cantabria

- The GAN is trained with detailed GEANT4-based simulations
  - A total of 50000 events for the training + 10000 events for the test datasets
- A reasonable agreement between GEANT4 and the GAN is found for the main features
- The speed up in the generation is x10000 with respect to the detailed GEANT4



VAE architectures for ATLAS Fast Simulation

- ATLAS has also studied the simulation of the ECAL showering for photons.
  - Two algorithms: GAN model and a Variational Auto-Encoders (VAE)
- The target (as for the LHCb case) is to generate the energy deposition in a block of cells.
  - A total of 266 ECAL cells are considered from the different ECAL layers.



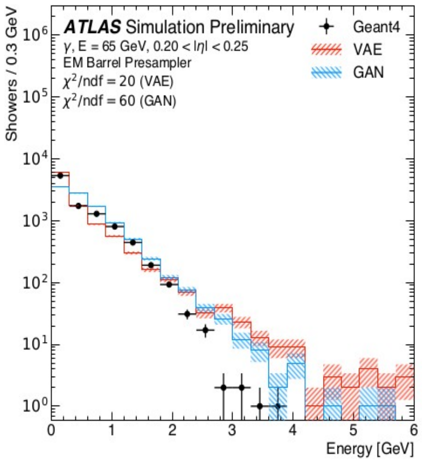
10.1088/1742-6596/1525/1/012077

VAE architectures for ATLAS Fast Simulation

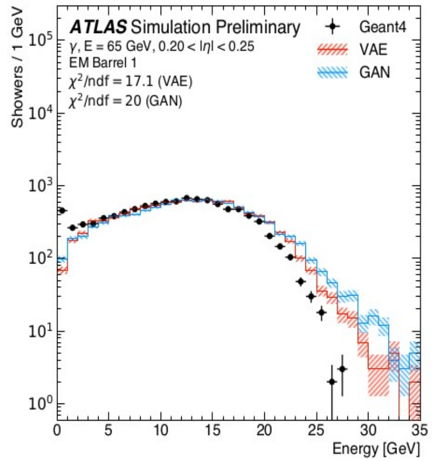
Instituto de Física de Cantabria

- The system is trained using a detailed GEANT4-based dataset with 90000 events.
  - Divided in 9 blocks of 10000 with 9 different incident energies.
  - Only one region of the calorimeter is taken into account (fixed phi and eta)
- The agreement between the VAE and the Geant4 is reasonable good
  - But still far to be used for precision measurements
  - The GAN approach (not explained here) seems to have a better performance.

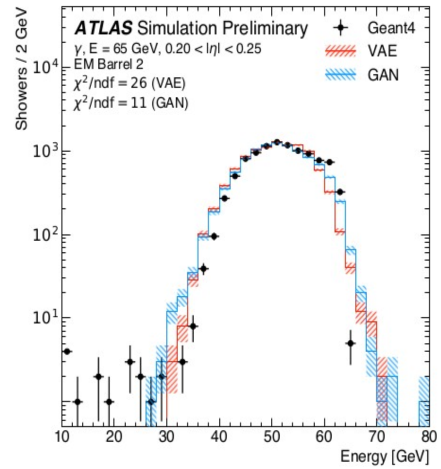
Presampler



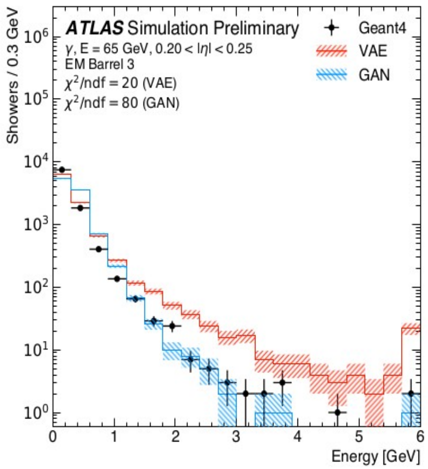
Front layer



Middle layer

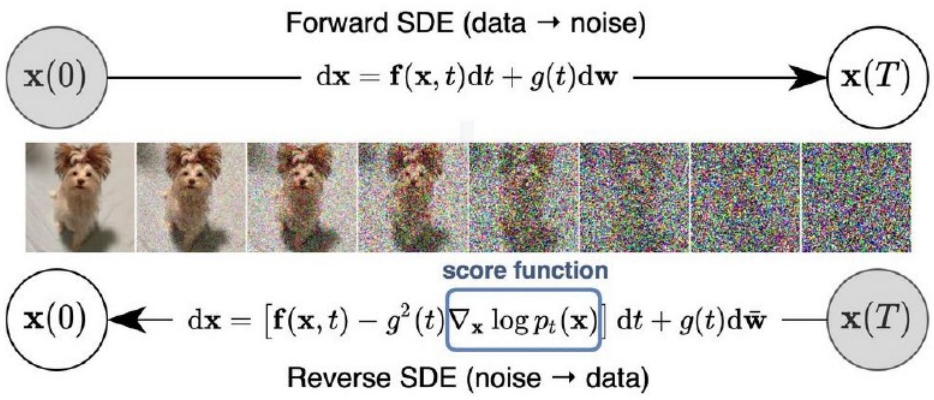


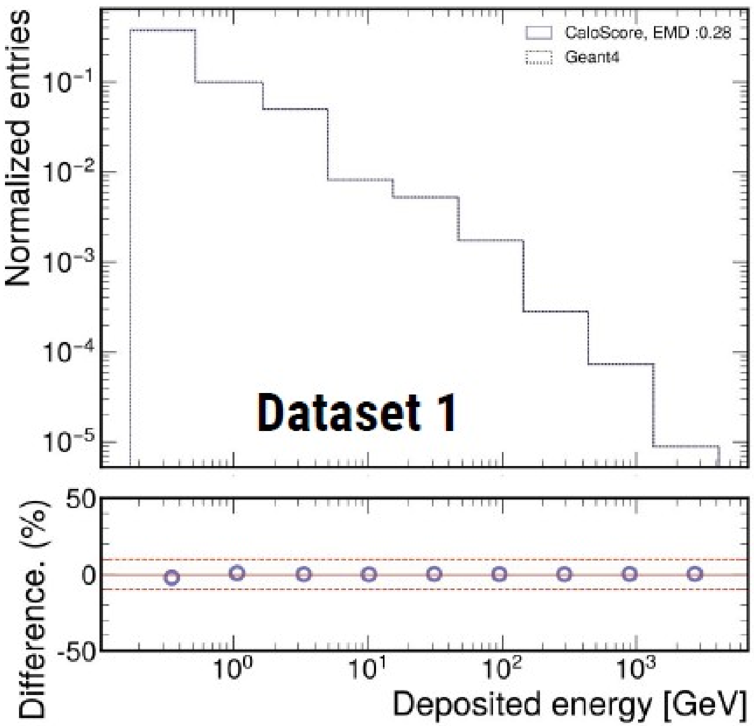
Back layer



Diffusion models for ATLAS ECAL

- Diffusion models were used in the ATLAS dataset of the CALO challenge (2023)
- These models have a non-equilibrium thermodinamycs inspiration. Two phases:
  - Diffusion phase where noise is added to a train-real image
  - Denoising phase where the noise is removed in steps to recover the image
- Diffusion models are outperforming GANs and VAE
- Although they are more CPU consuming algorithms



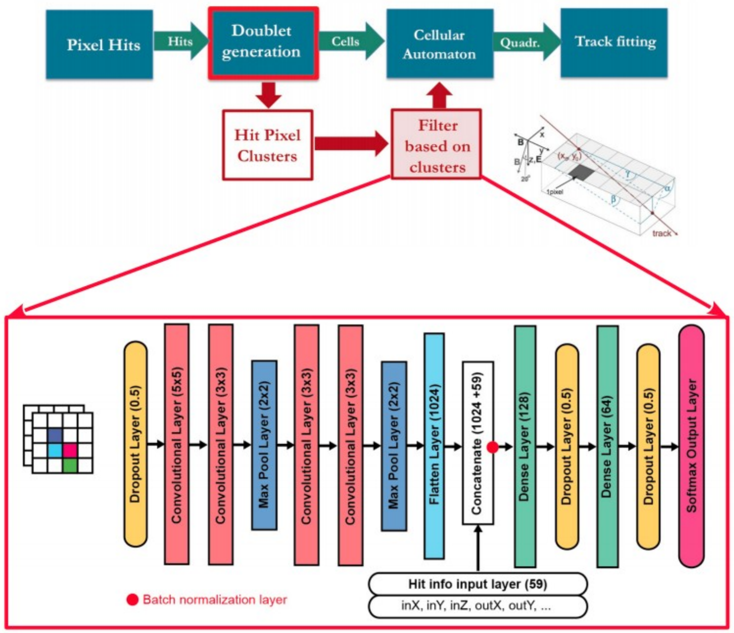


Calo challenge Summary

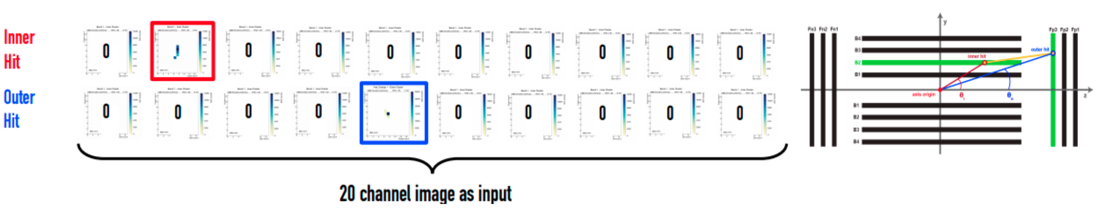
Trigger

CNN for track seed filtering at the CMS HLT

- Tracks reconstructed with the pixel detector are used online for fast tracking and vertexing
- This is a challenge for Phase2 where the PU is expected to scale up to 200
- CMS is devising a full, parallelizable HLT RECO running on GPUs and using CAs
  - Still there is a bottleneck on the number of “doublets” that will be further processed
  - A CNN has been proposed to filter these seeds as a classification problem (Valid or not)
  - Hits represented as 16x16 pixel pads images with colors proportional to deposited charge



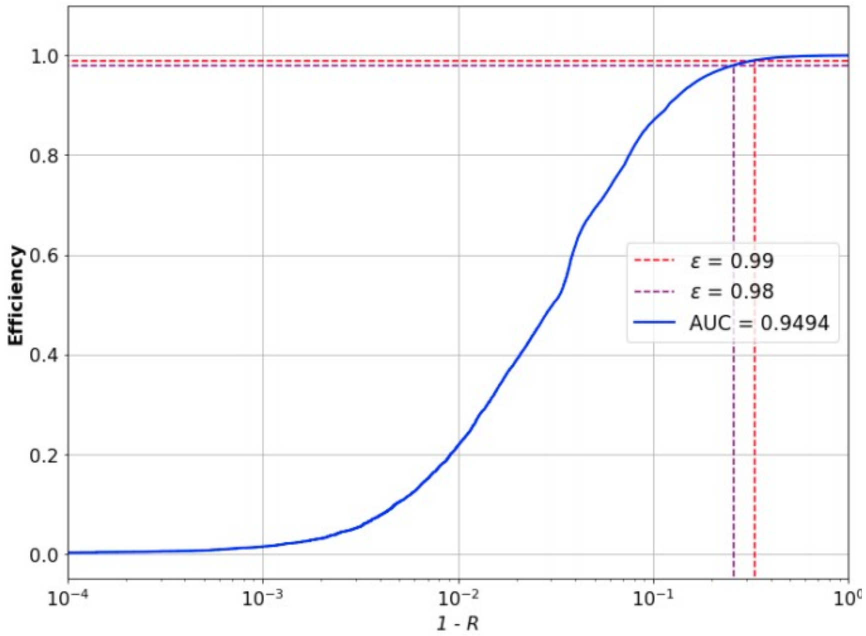
- Images are combined in 20 channels/levels
- Accounting for the different inner/outer layers

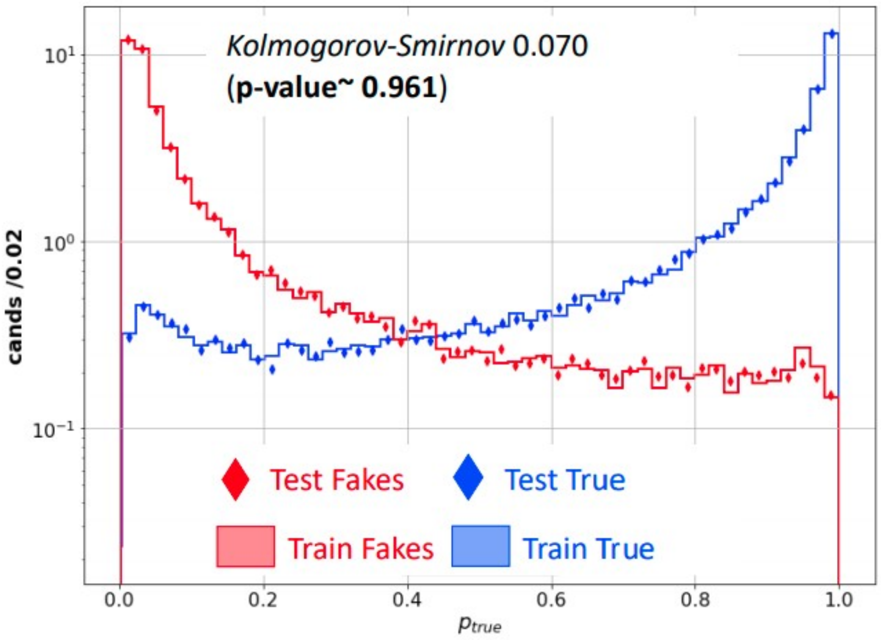


https://indico.cern.ch/event/819693/contributions/3438504/attachments/1858975/3054502/Patatrack_DiFlorio_CMSCalcolo.pdf

CNN for track seed filtering at the CMS HLT

- Test have been donde with a training on O(10^7) doublets from RECO simulation
  - Obtained with only O(100) events
  - True doublets are those where the hits can be matched to a same GEN particle
- The system retains about 99% of efficiency while 2/3 of the fake doublets are rejected

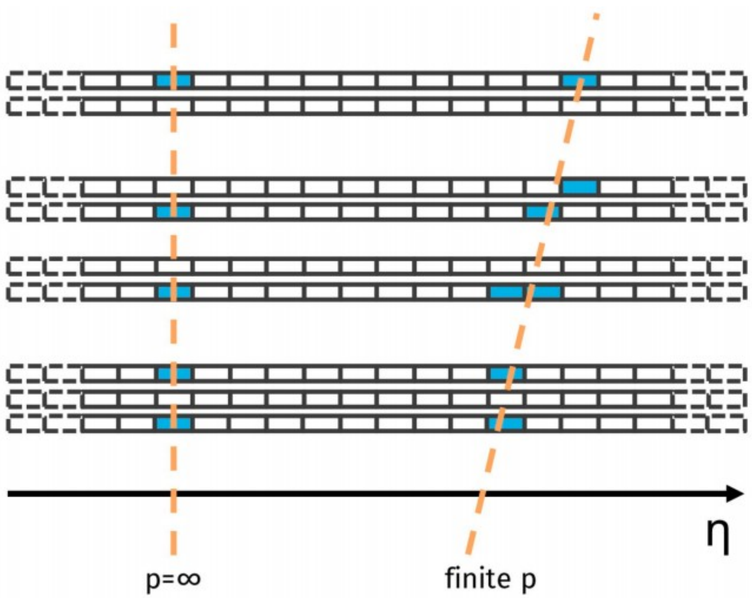


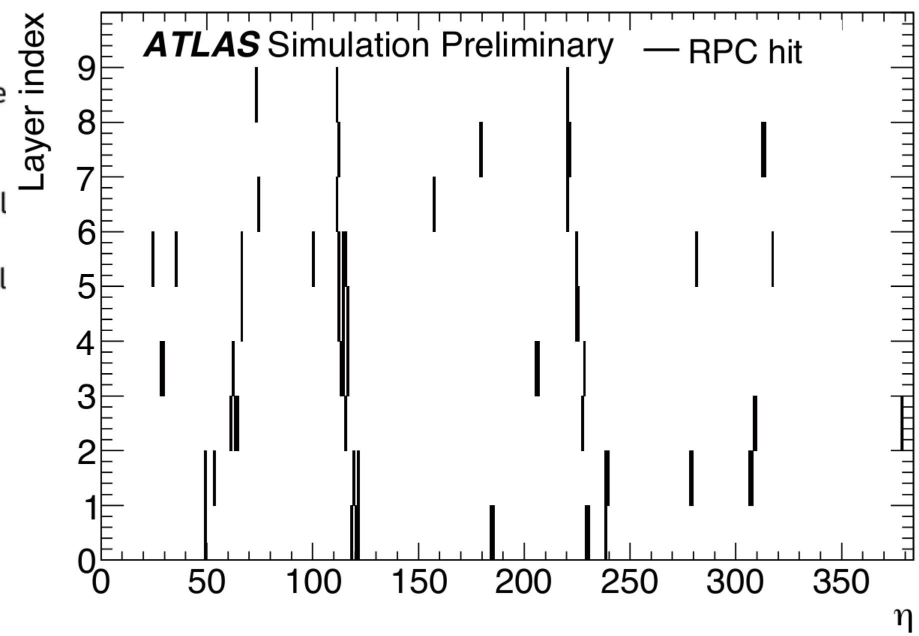


ATLAS Phase-II RPC Muon Barrel Trigger

- The ATLAS collaboration is working on a CNN running on a FPGA for the muon trigger.
- Events are interpreted and treated as images that are further fed into a CNN.
  - The RPC hits are represented as eta Vs. layer maps in the RPCs
  - Image size is 384 bins in eta x 9 RPC stations
- The CNN performs a regression to 5D space [$p_T^{leading}$, $\eta^{leading}$, $p_T^{leading}$, $\eta^{leading}$, # muons]

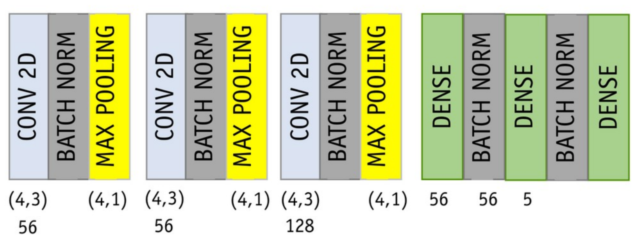
ATLAS-L0-MUON-PUBLIC

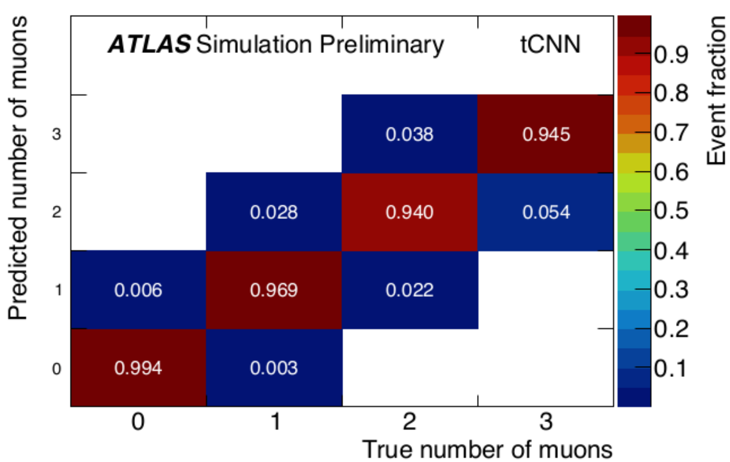


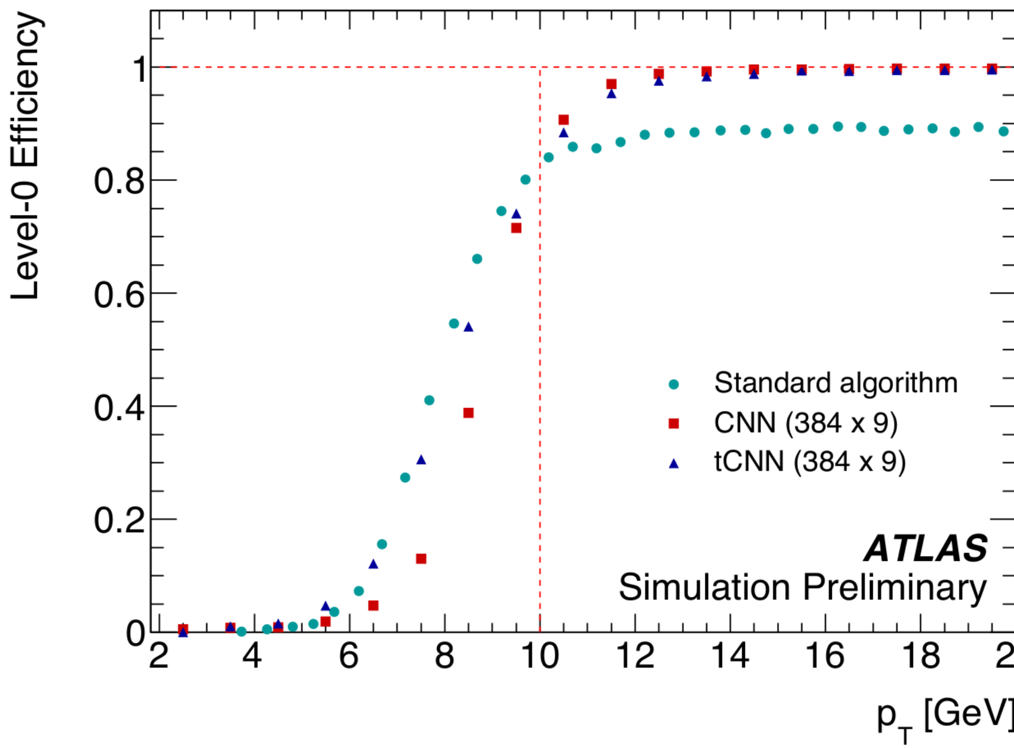


ATLAS Phase-II RPC Muon Barrel Trigger

- The size and architecture of the CNN has to match the characteristics of the FPGA
- In order to reduce memory consumption a “Ternary CNN” is proposed
  - Weights and activations can only take {-1, 0, 1} values instead of floating point.
  - Memory is reduced by a factor 16 thanks to this procedure
- The network outperforms by ~10% the classical algoritm in terms of efficiency.





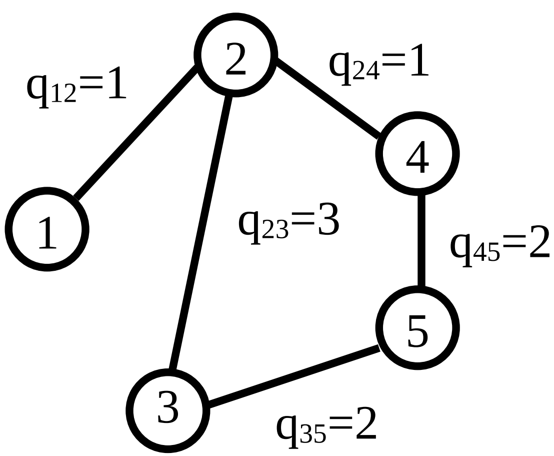


Reconstruction/ Identification

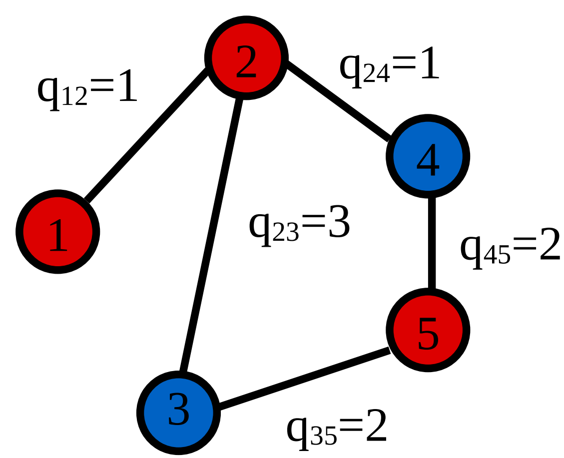
Vertex reconstruction in Quantum Computing

- Vertex Reconstruction is the process of clustering tracks into a set of vertices
- This problem is combinatorial in nature: consider a problem with 2 true vertices.
  - Which track combinations minimize their relative distances and maximize to the others?
  - Actually this problem can be seen as a well-known problem in Graph Theory: Max-Cut
- Given a graph with nodes and weighted edges → assign labels (red or blue) to the nodes in such a way that the sum of the weights crossing from one group to the other is maximal
- Encoding the solutions as vectors x = [0, 1, 0, 1, 0] then need to maximize:

$$F = \sum_i \sum_j x_i Q_{ij} (1 - x_j)$$



→



https://doi.org/10.1051/epjconf/202227409002

Vertex reconstruction in Quantum Computing

- Consider tracks 3D points as a fully connected graph with weights equal to their distance
- Finding the assignment of tracks to vertices equals to finding two groups that maximize their mutual distance to each other → which is precisely a Max-Cut problem
- Most Quantum Computers can implement an Ising Hamiltonian of the form

$$H(\sigma) = -\sum_{\langle i\ j \rangle} J_{ij}\sigma_i\sigma_j - \mu \sum_j h_j \sigma_j \longrightarrow F = -\sum_i \sum_j x_i Q_{ij}(1 - x_j)$$

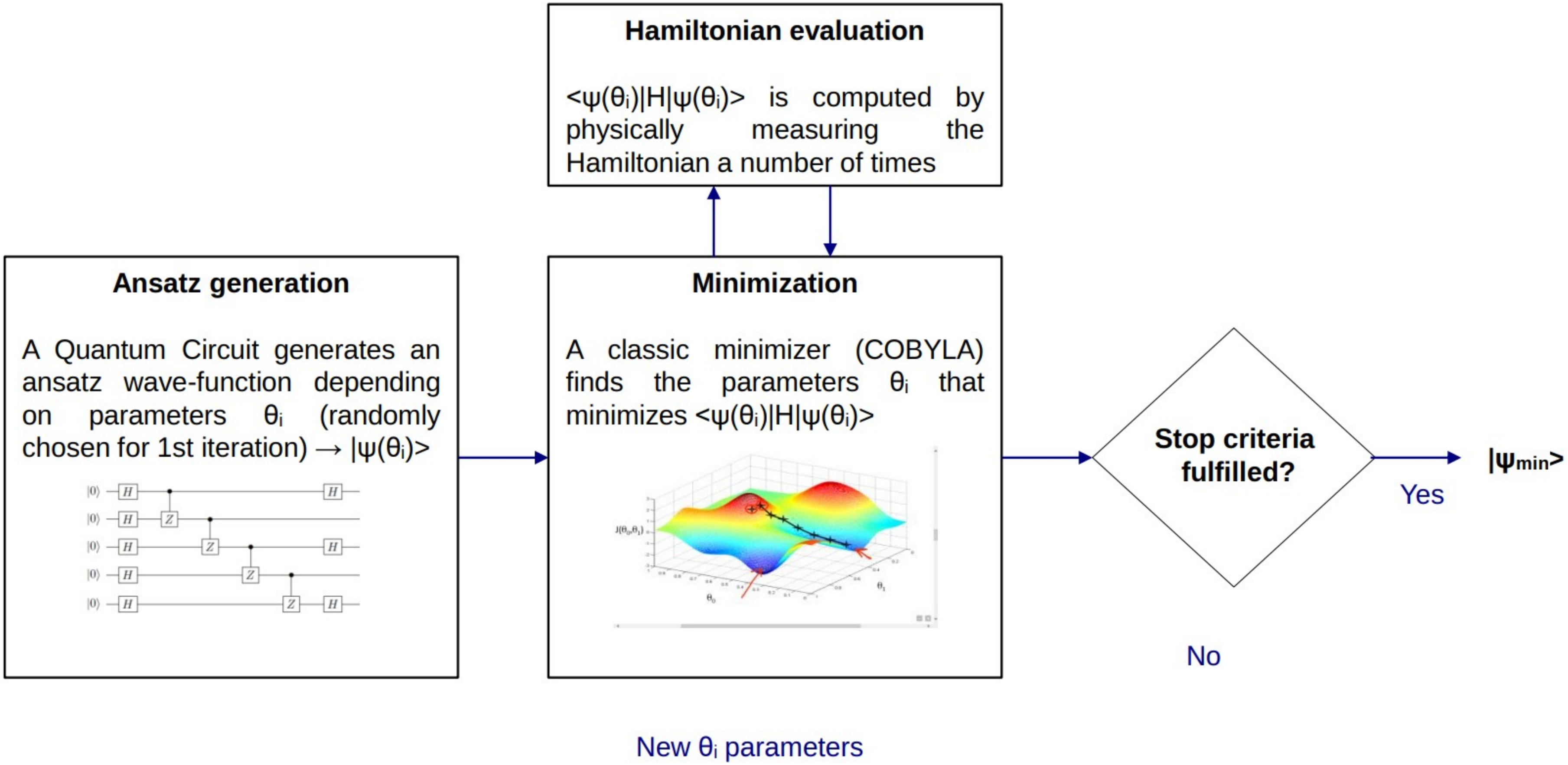
- The group assignment can be encoded in a set of quantum bits A = [q_1, q_2, q_3, …, q_N]
- The x_j operator is defined to be 1/2(1 + σ_j) with value 0 or 1 when applied to q_j
- The quantum state for which this Hamiltonian is minimum is the solution to the problem

https://doi.org/10.1051/epjconf/202227409002

The VQE algorithm

Instituto de Física de Cantabria

- The Variational Quantum Eigensolver is a hybrid classic-quantum algorithm
  - Aiming at finding the multi-qbit state that minimiez a given Hamiltonian

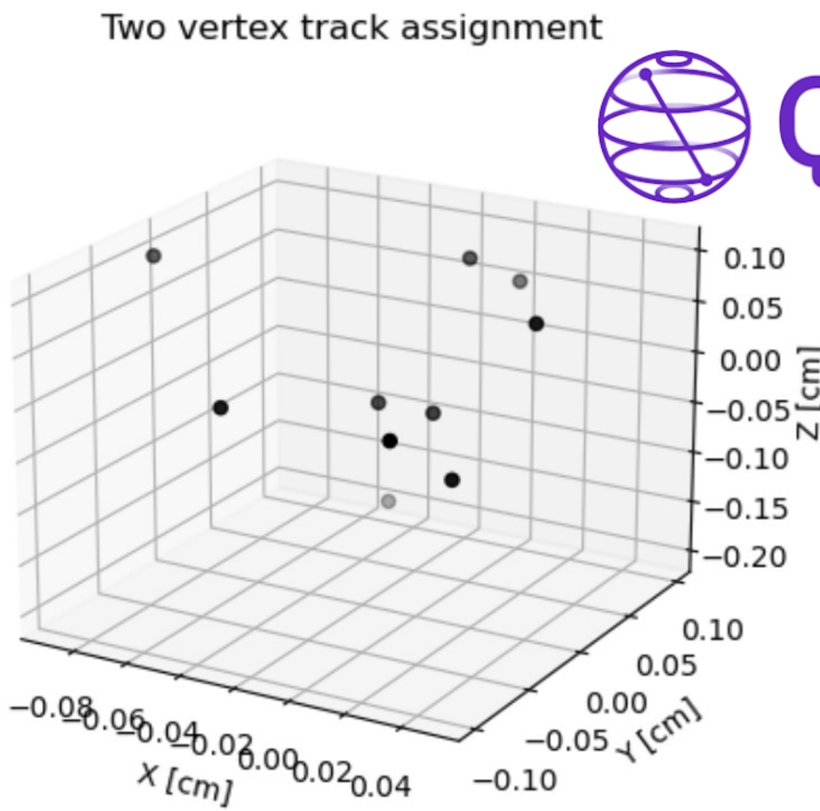


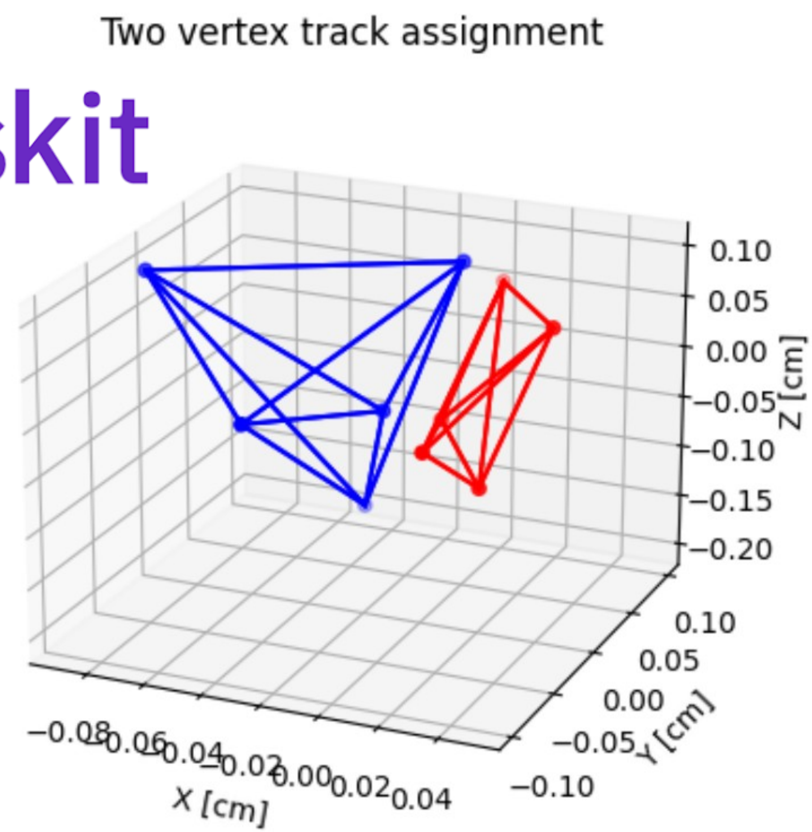
https://doi.org/10.1051/epjconf/202227409002

Vertex Reconstruction: results

Instituto de Física de Cantabria

- Algorithm implemented in the simulation framework of IBM Qiskit
- Reconstruction efficiency above 90% for vertices statistically separated below 1 mm

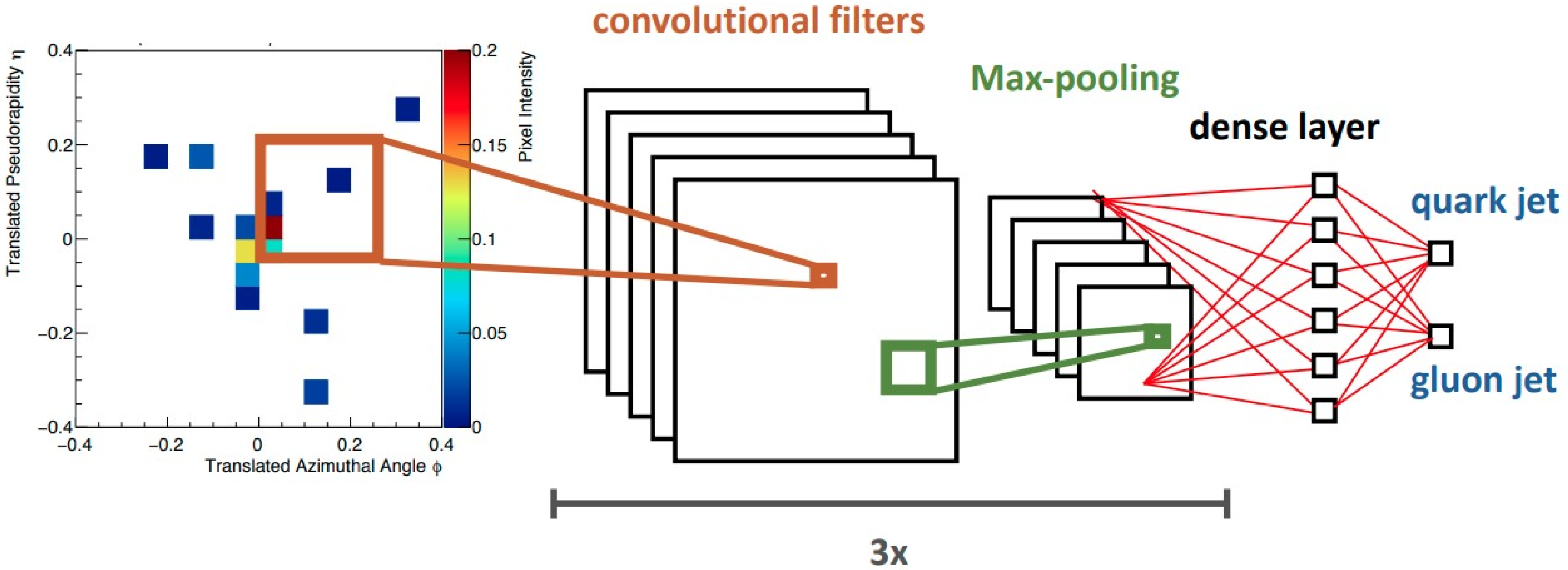




https://doi.org/10.1051/epjconf/202227409002

ATLAS quark-gluon discriminator with CNN

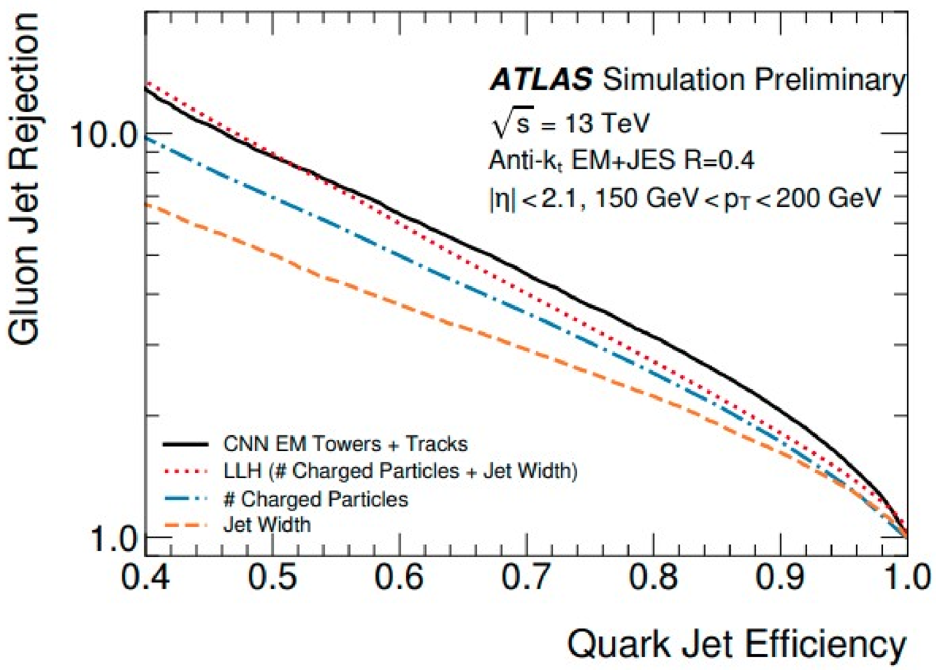
- ATLAS has also explored Convolutional Neural Networks to learn jet substructure
- Jet constituents are represented in eta – phi images with 16x16 binning
  - Tracks and tower or topocluster information are represented in different images
  - The color is proportional to the pt of the constituent (and then normalized)
  - The best performance is found when combining the track + tower/topocluster input

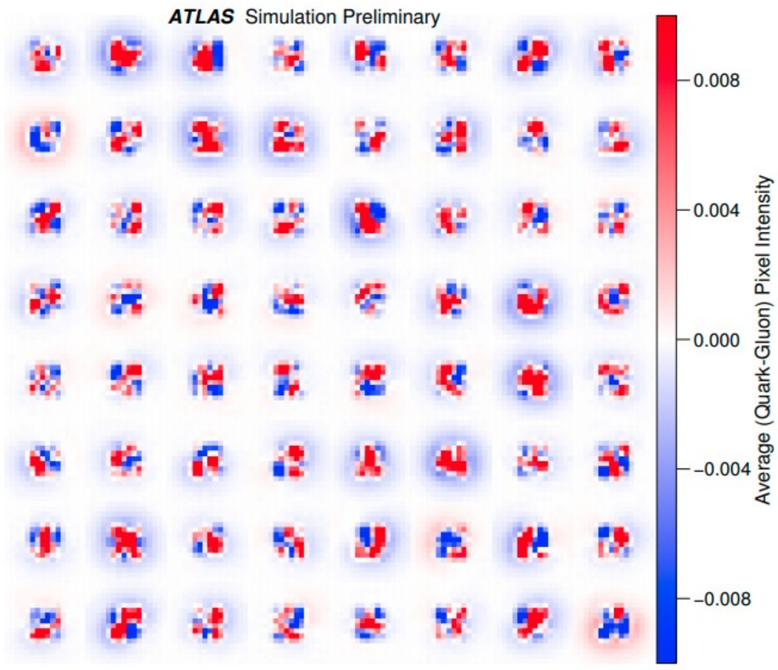


ATL-PHYS-PUB-2017-017

ATLAS quark-gluon discriminator with CNN

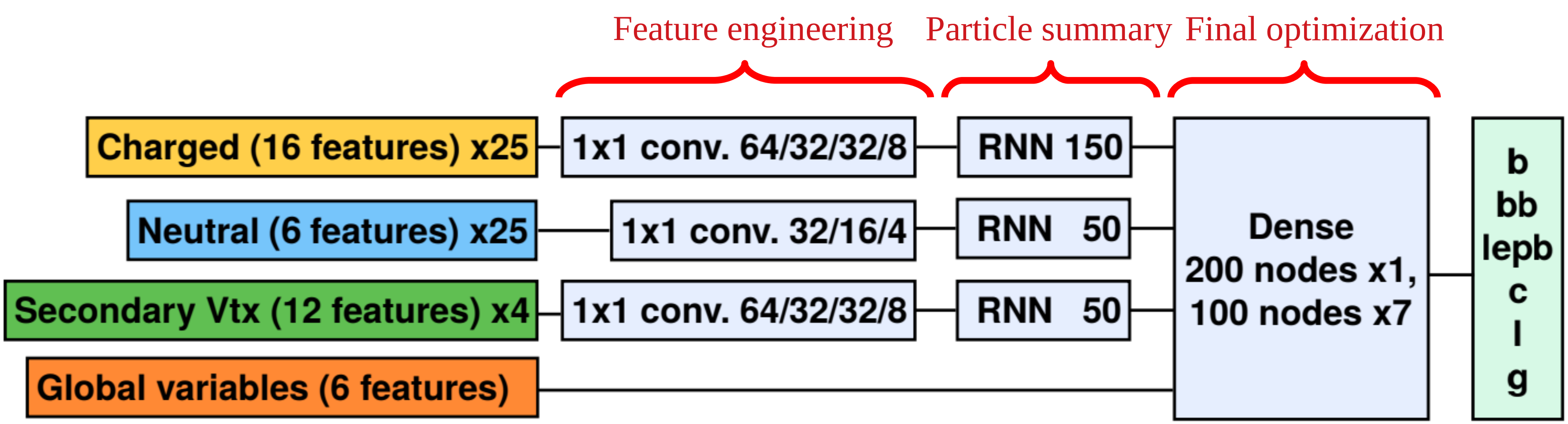
- The network is trained with 2 fragmentation models (Pythia8 and Herwig++) + GEANT4
  - The train dataset is composed of about 224000 images and the test about 56000
- Much better performance than the likelihood based quark-gluon discriminator
- Explainability of the tagger functioning can be also studied by looking at the filters
  - The average jet/quark images are convoluted with the filters and compared





DeepJet tagging in CMS

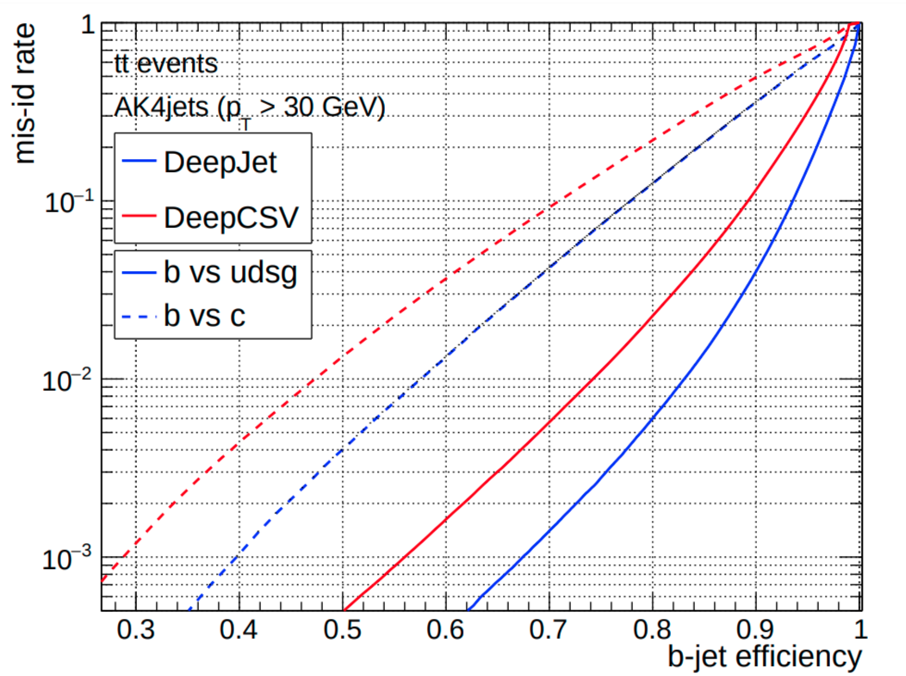
- CMS has devised a system to perform jet tagging combining CNN, RNN and Dense Layers.
- The network uses 4 levels of features:
  - Charged particles: 16 features per particle x 25 charged particles
  - Neutral particles: 6 features x 25 neutral particles
  - Secondary vertex: 12 features x 4 secondary vertex
  - Global variables: 6 features (number of vertices, jet pt, eta, etc.).
- The CNN creates features per particle while the RNN (LSTM) summarizes sequentially

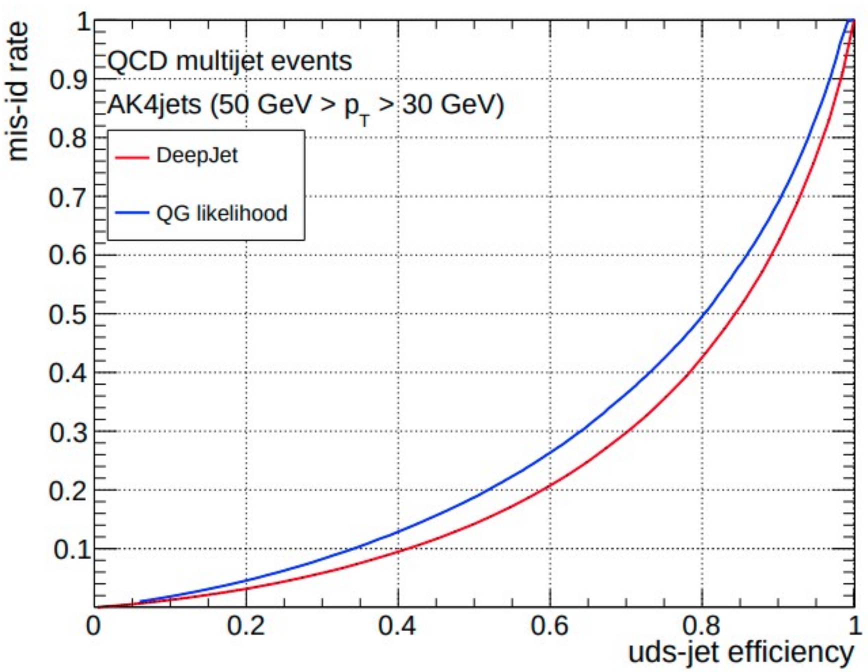


https://arxiv.org/pdf/2008.10519.pdf

DeepJet tagging in CMS

- The algorithm is trained with 130 million jets coming from simulated QCD and ttbar.
- The performance is compared to the CMS DeepCSV algorithm based on a fully dense ANN.
  - DeepJet outperforms by ~12% the b-tagging efficiency for 0.001 misidentification rate.
- Also the performance is compared to the likelihood-based quark-gluon discriminator.
  - DeepJet outperforms by ~10% the quark-gluon discriminator for 0.3 misidentificate rate.

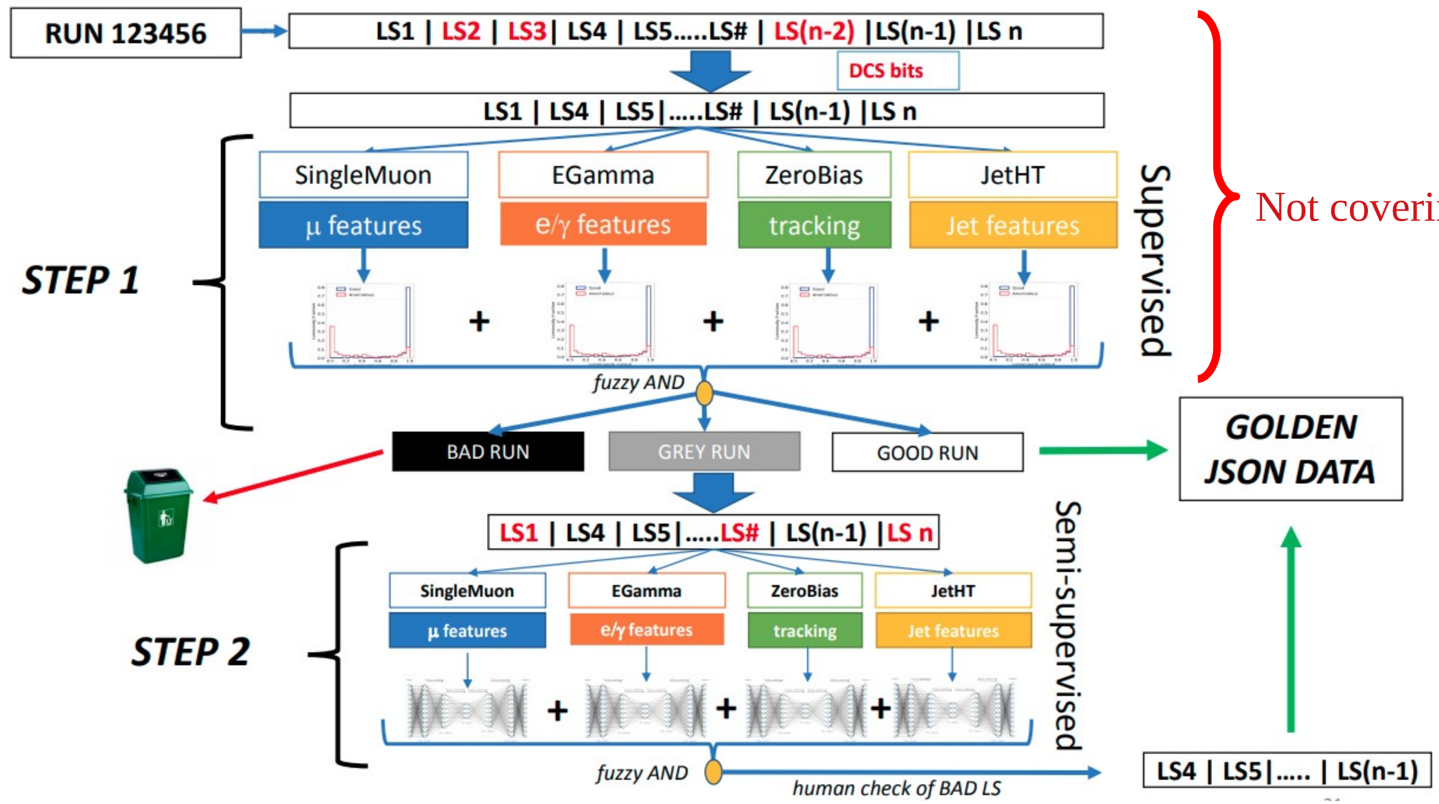




DQM

Anomaly Detection in CMS Data Certification

- Data Certification: subsystem experts assign a quality flag to runs and lumisections
  - Tedious and time consuming task (for example rarely DC really occurs per lumisection)
- CMS setting up a 2-step DC procedure combining supervised & unsupervised ML methods

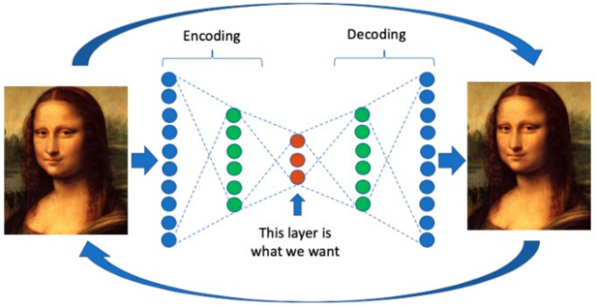


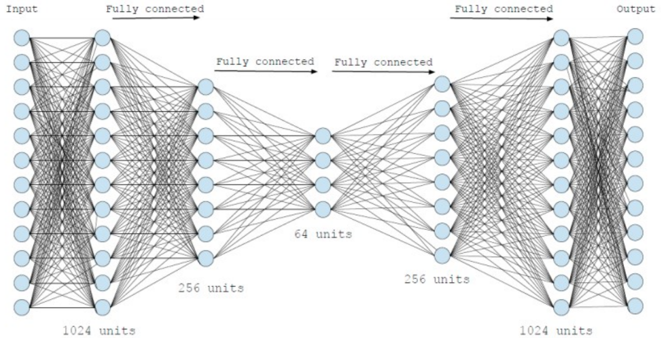
Not covering this here

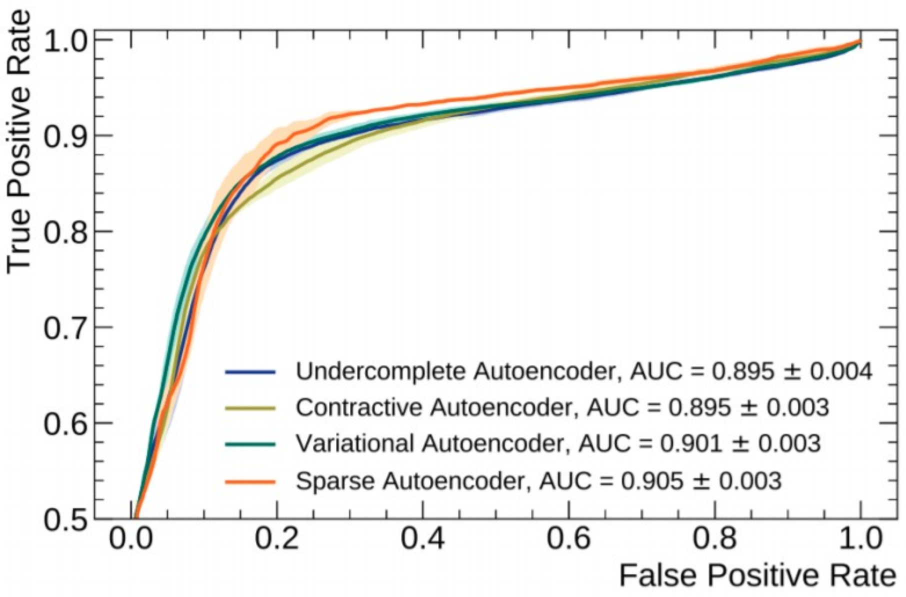
10.1088/1742-6596/1085/4/042015

Anomaly Detection in CMS Data Certification

- The second step uses Variational Auto-Encoders to assign the quality flags
  - No need of BAD data for training (good since fortunately most of the data is GOOD)
  - The source of a given anomaly can be traced back (interpretability of results)
- The VAE learns to compress and uncompress the internal structure of the GOOD data
  - This process does not work for anomalies resulting in an output very different to the input
- The input to the autoencoder are the 5-quantile + mean + RMS of key histograms







Detector Design Optimization

Differential programming (quick definition)

- A new paradigm in which a computer program/function can be differentiated
- This is achieved by using automatic differentiation usually exploiting the chain rule

```
def myTargetFunction(x):

    x1, dx1dx = funcA(x)
    y, dydx1 = funcB(x1)
    return y, dydx1*dx1dx
```

```
def funcA(x):

    x1 = x*x
    dx1dx = 2*x
    return x1, dx1dx
```

```
def funcB(x):

    y = 1.0 / (1.0 + x)
    dydx1 = -1.0/(1.0 + x)**2
    return y, dydx1
```

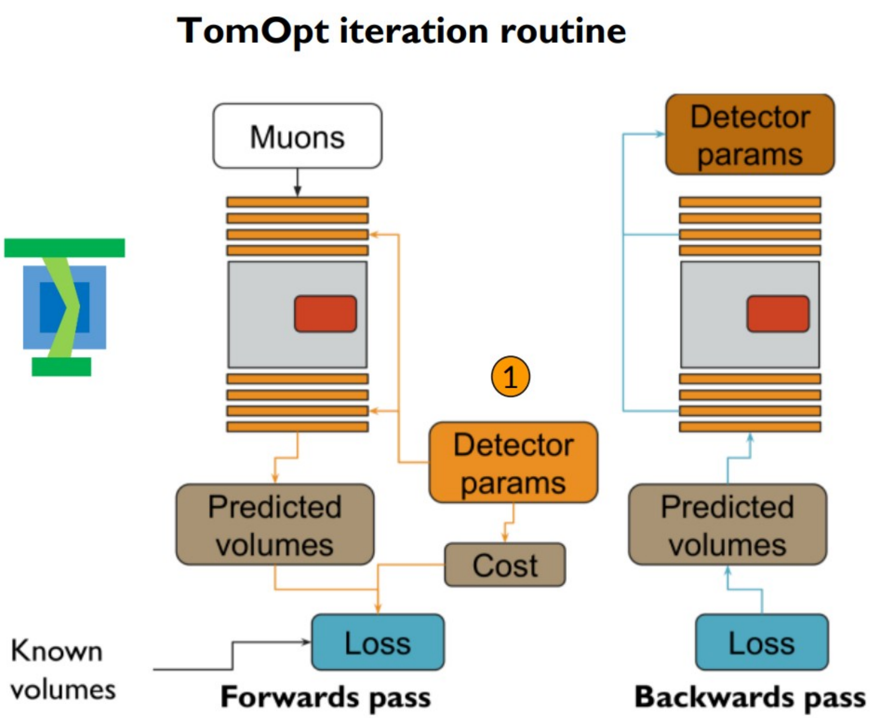
- This technique allows to quickly and efficienctly estimate gradients of complex functions
- It is possible to minimize complicated loss functions using the gradient and SGD
- In a very simplistic way you can see this as a generalization of the backpropagation method
  - But applied on generic functions and not on simple structures such as neurons

https://doi.org/10.48550/arXiv.2309.14027

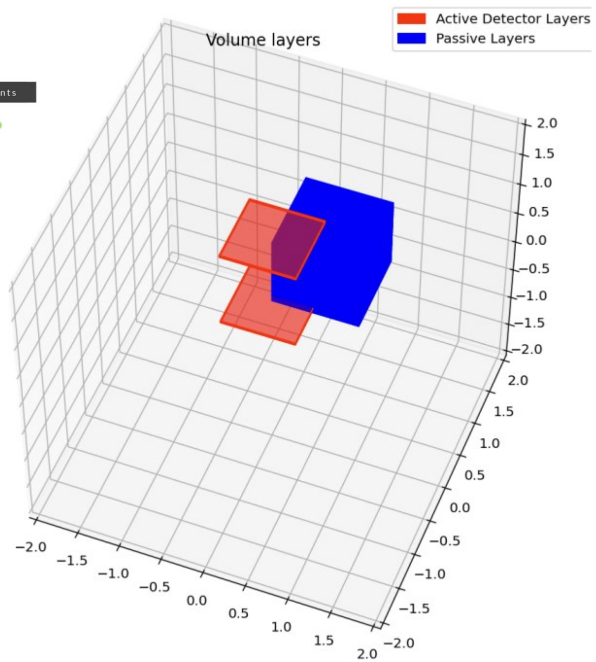
Differential programming for detector design

Instituto de Física de Cantabria

- Optimal design/configuration of a particle detector can be estimated using DP
- The objective function should contain metrics about all important parameters in the design:
  - Performance (efficiencies, resolutions, etc), cost, constraints in the system
- These ideas are being exploited to produce optimal design for a muography experiment



MODE
Machine-learning Optimized Design of Experiments



https://doi.org/10.48550/arXiv.2309.14027

Conclusions

- New computing paradigms are emerging, growing and changing the world as we know it
- In particular at the LHC and other HEP experiments they are starting to have a strong impact
- A large plethora of different algorithms are being used at different places of the experiments
- A few examples have been shown on Generation, Trigger, Reco/Identification and DQM
- A set of different algorithms discussed but be aware that many new algorithms are coming
  - Graph Neural Networks (tracking), Autoregressive networks, ...
- Also some steps are being given in the direction of improving explainability of the systems
- Large gain and in some cases impressive results
  - But remember that usually in the talks only the successful examples are shown :-)
  - The large gain usually comes with a large effort in understanding the details

Thanks for your attention

Enjoy the school and enjoy Oviedo!