Accelerating Discovery at the Large Hadron Collider

Elham E Khoda

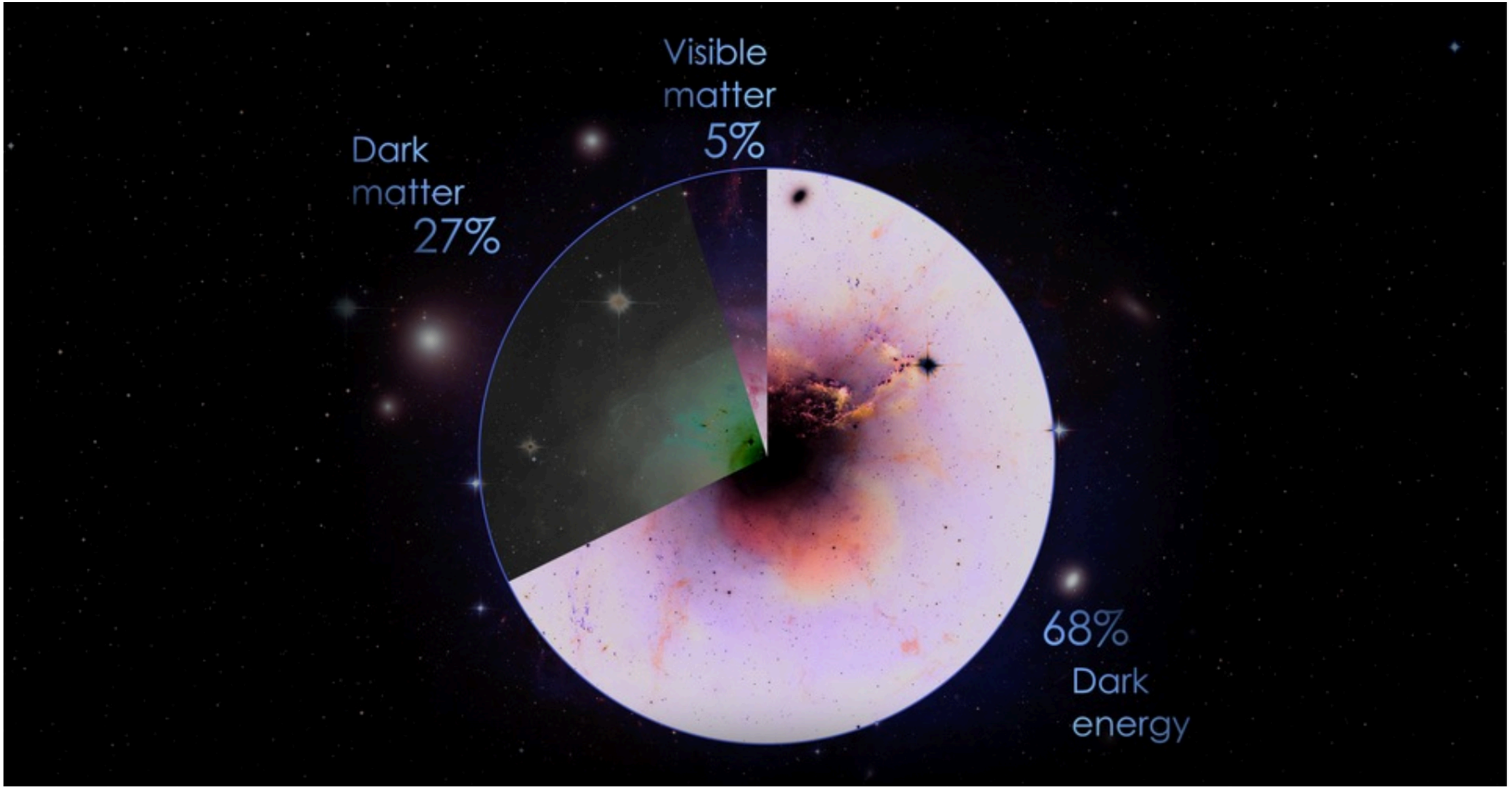
Departmental Colloquium, Physics and Astronomy
University of Hawaii, Manoa

March 27, 2024

ATLAS EXPERIMENT

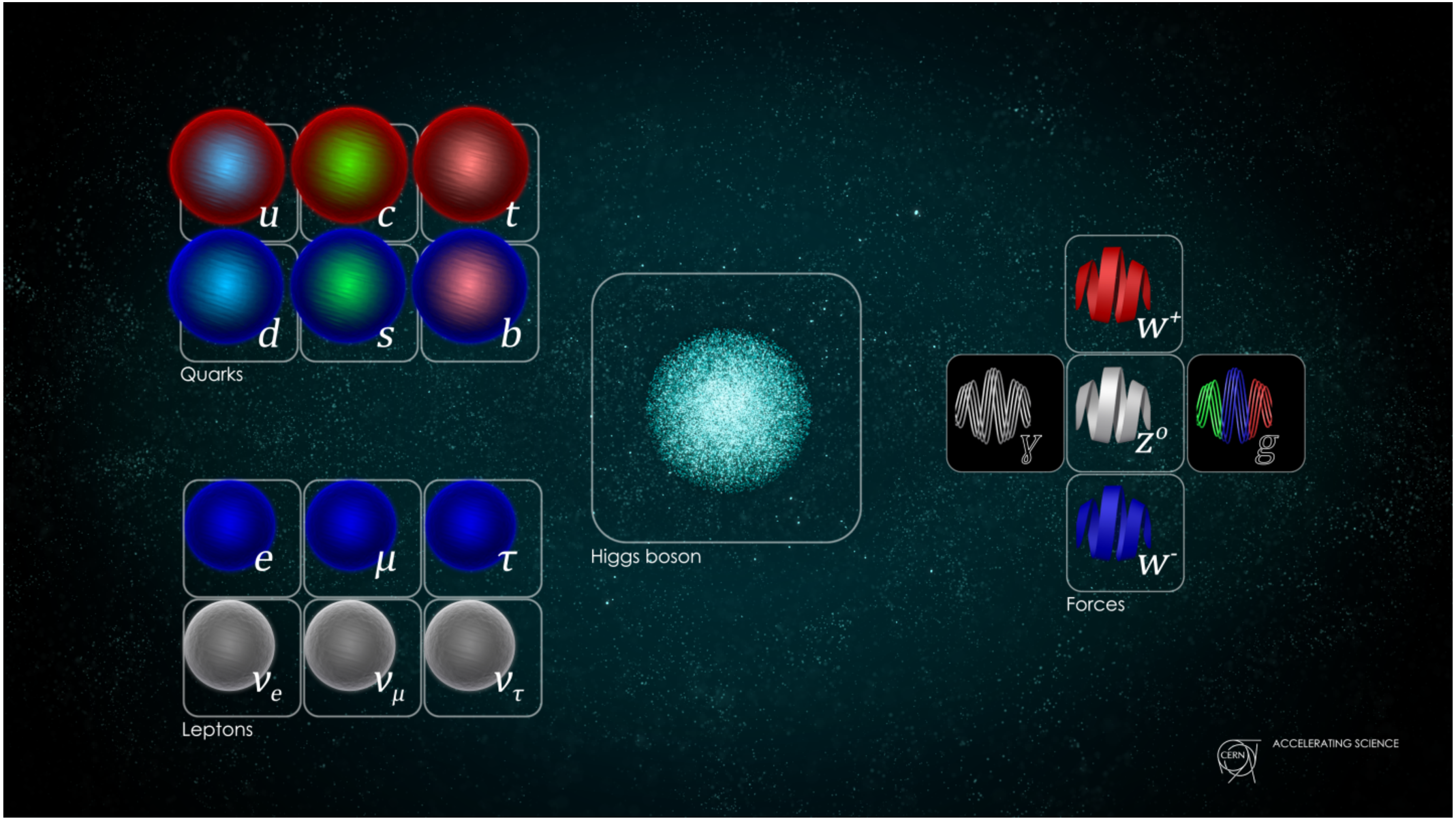
a3d3 Accelerated AI Algorithms for Data-Driven Discovery

Our Universe



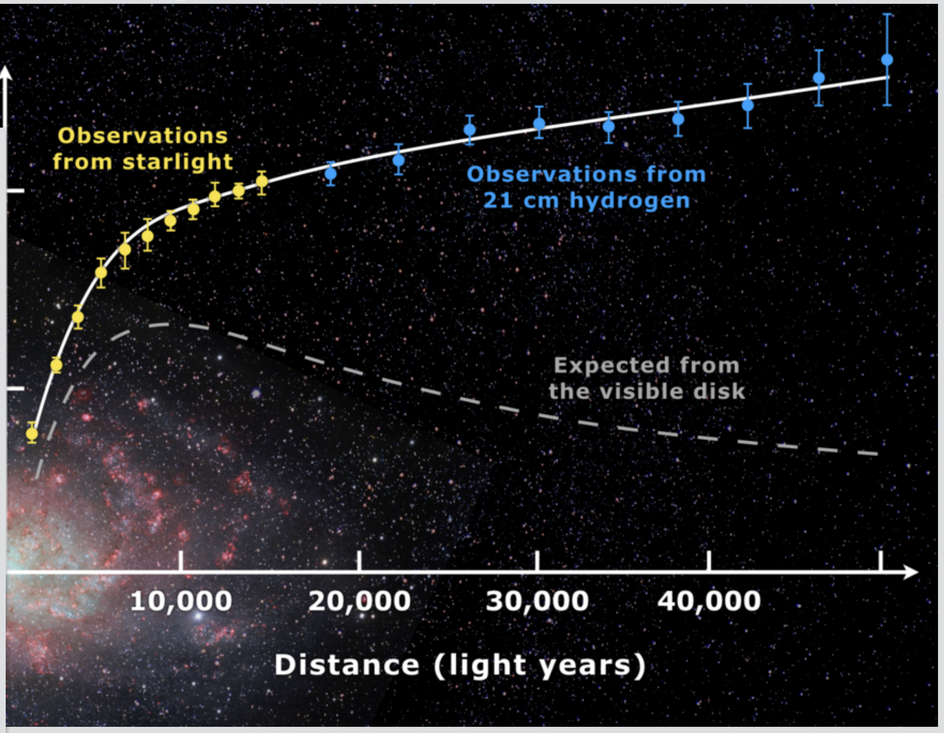
Credit: NASA, Planck

Standard Model describes the visible matter

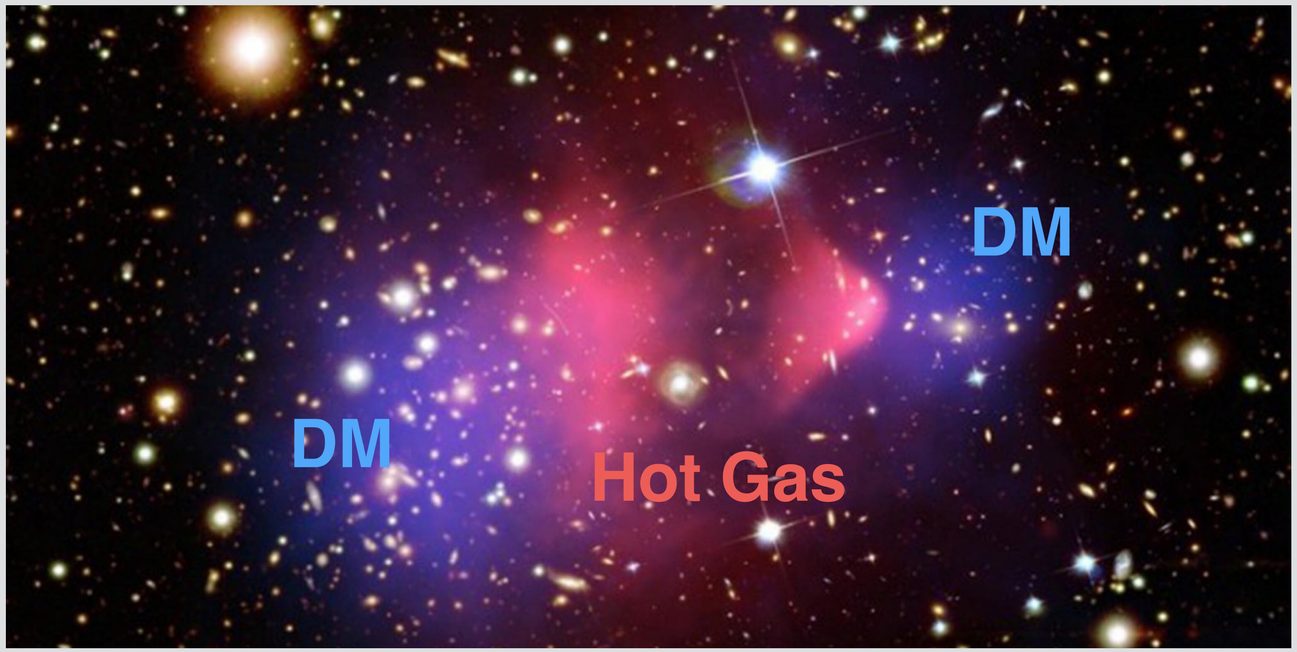


Dark Matter?

Galaxy Rotation Curve



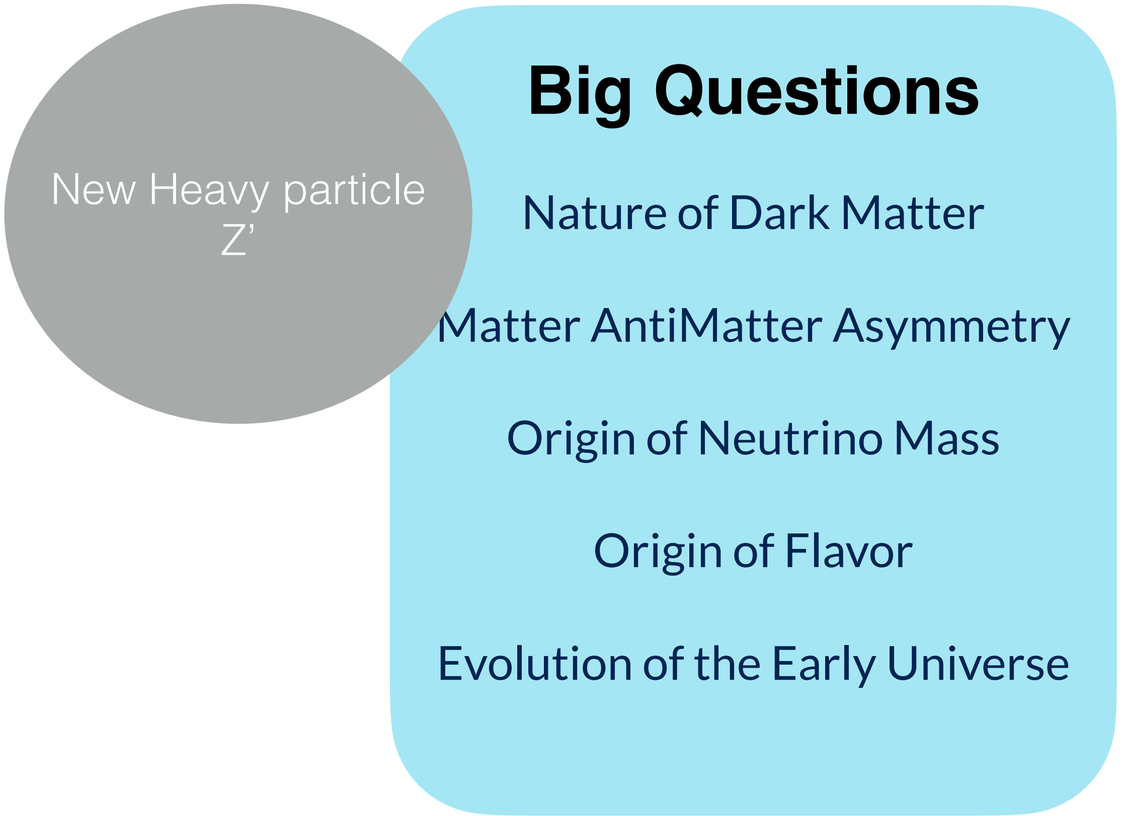
The Bullet Cluster



Credit: NASA, Planck

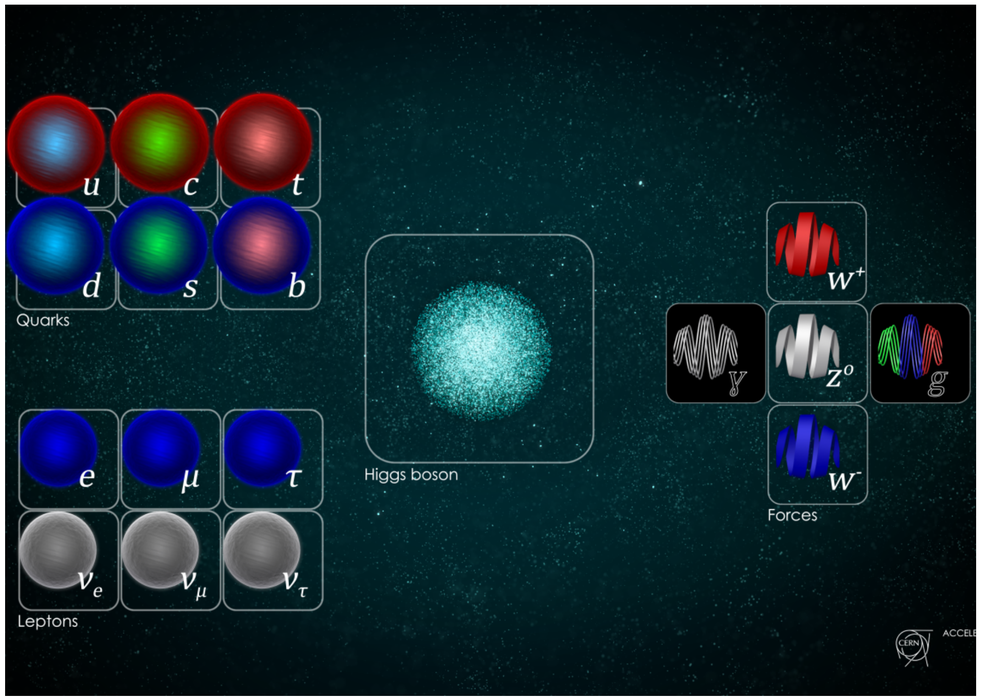
Physics Beyond the Standard Model?

Many other shortcomings in the SM!



Mediators to the Dark Sector

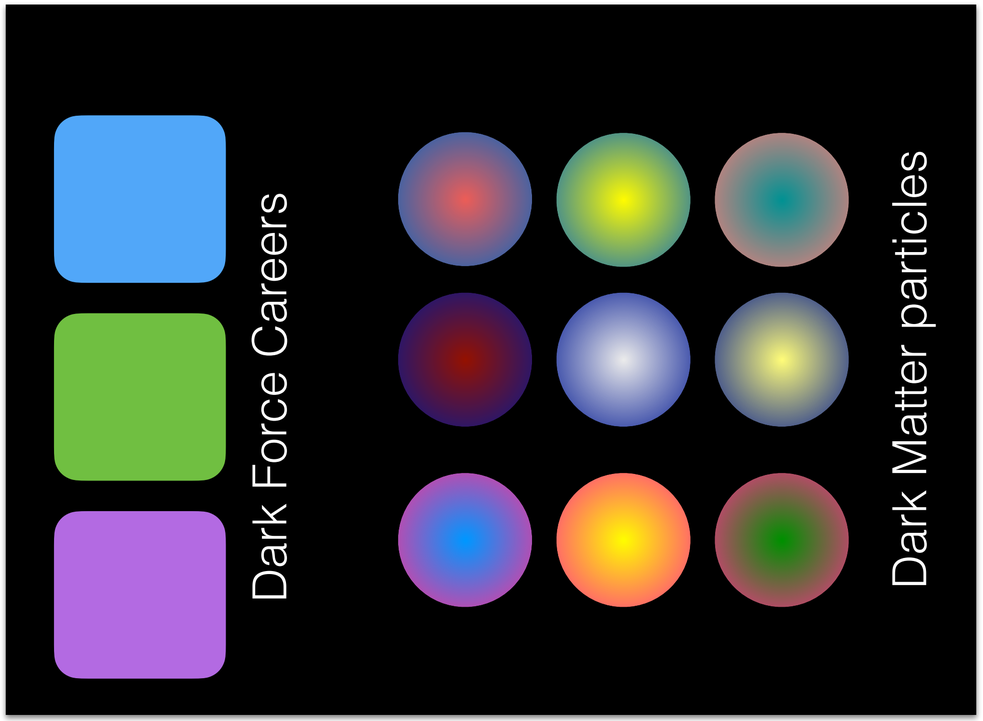
Standard Model



Mediators
?

Z'
Z' boson

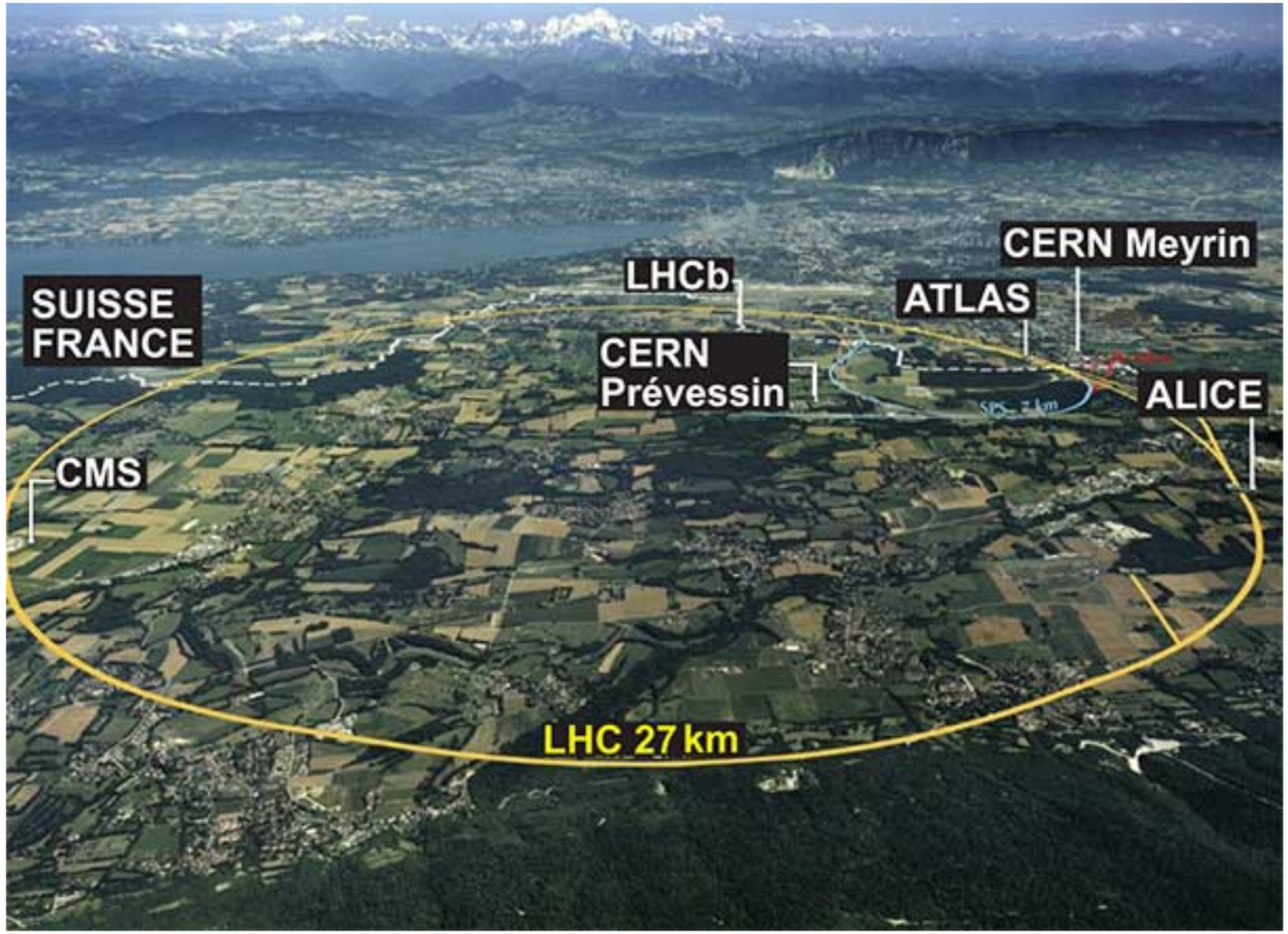
Dark Sector



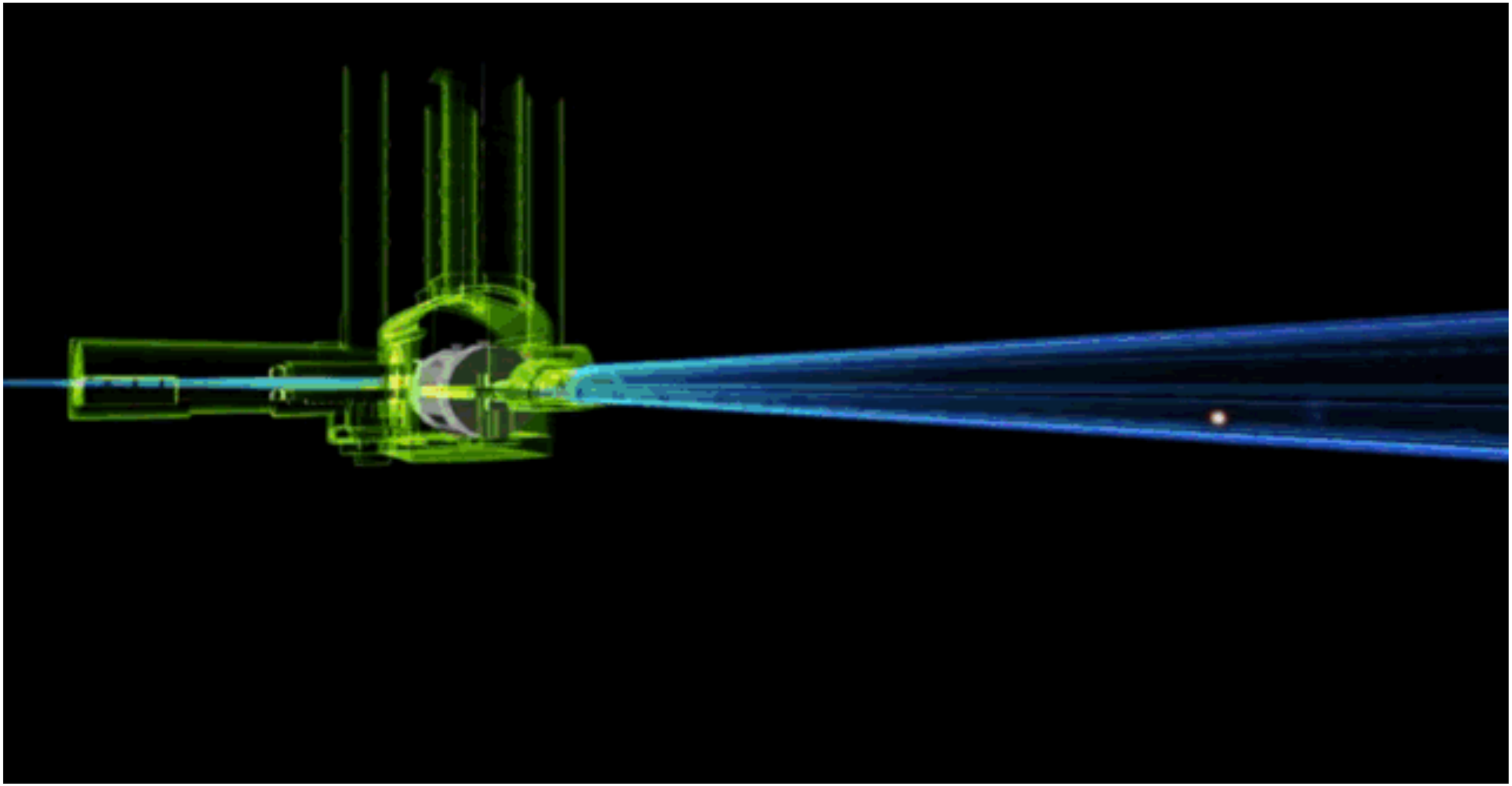
How to look for Dark Matter at a Collider Experiment?

Large Hadron Collider (LHC) at CERN

- World's largest and most powerful particle collider
- Collides protons (most of the time) bunches (~10^{11} protons in a bunch) spaced by 25 ns

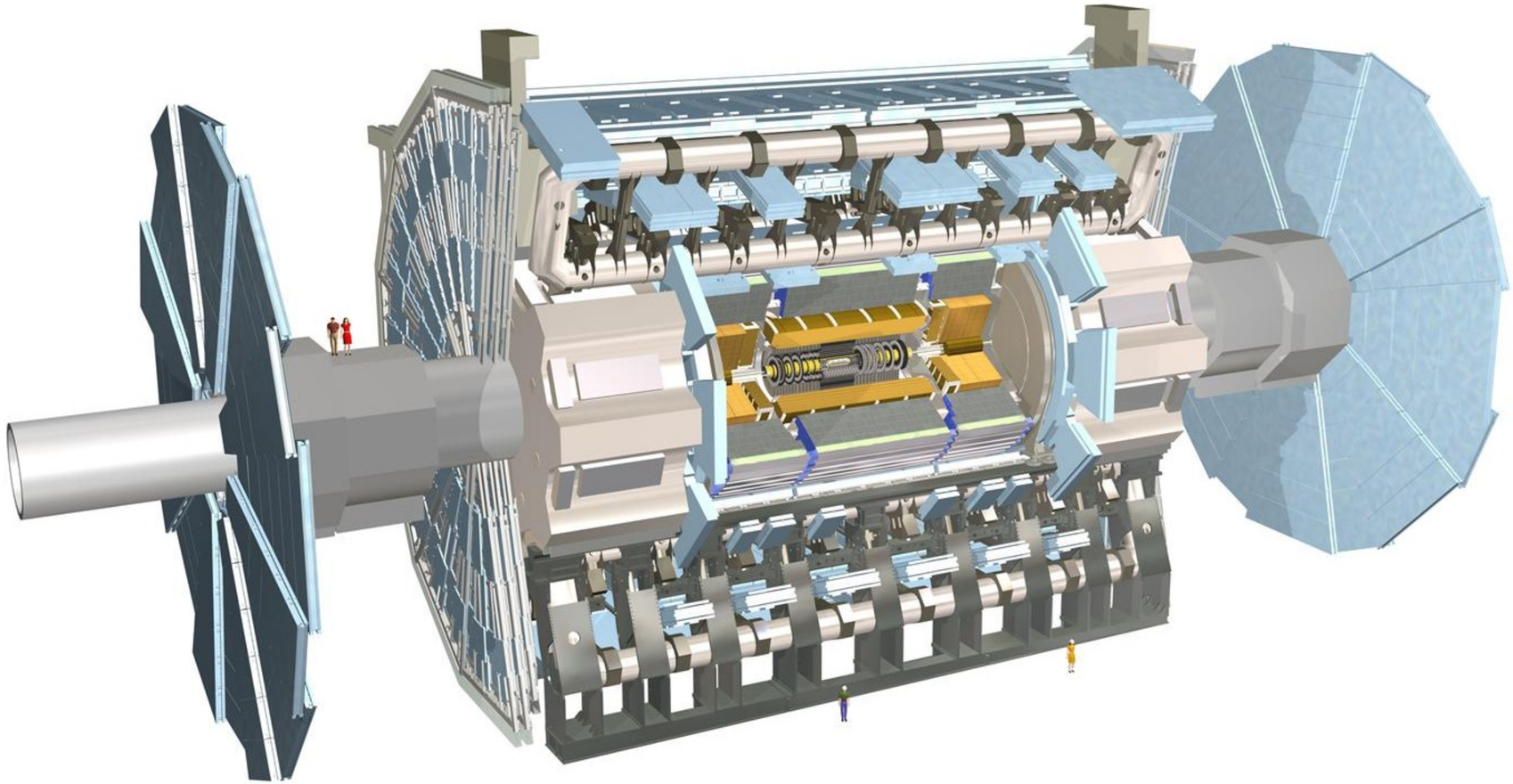


What happens when two proton beams collide



The ATLAS Experiment

General purpose detector → studies a wide range of physics

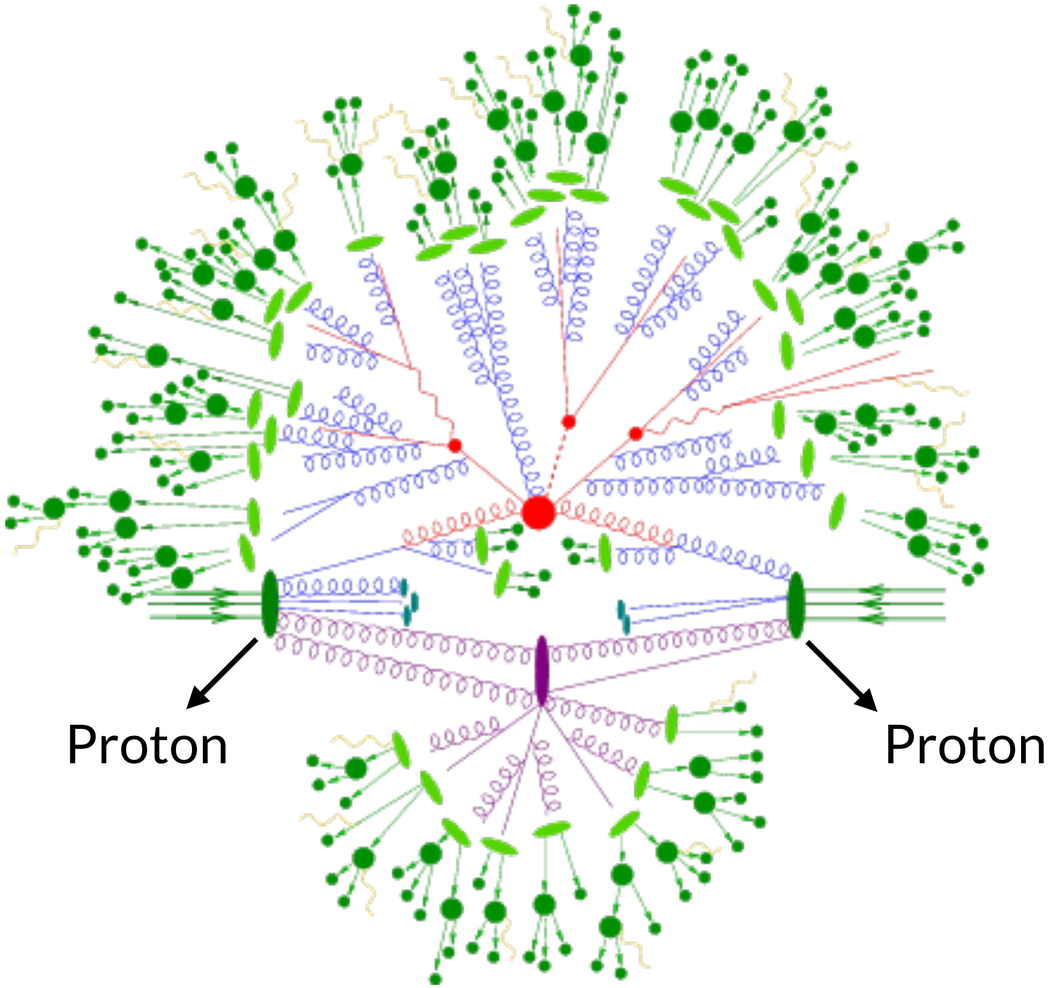


Proton collisions at the LHC

Proton is composite particles

3 valence quarks and a sea of gluons

Single p-p collision



With in ~ femto meter distance

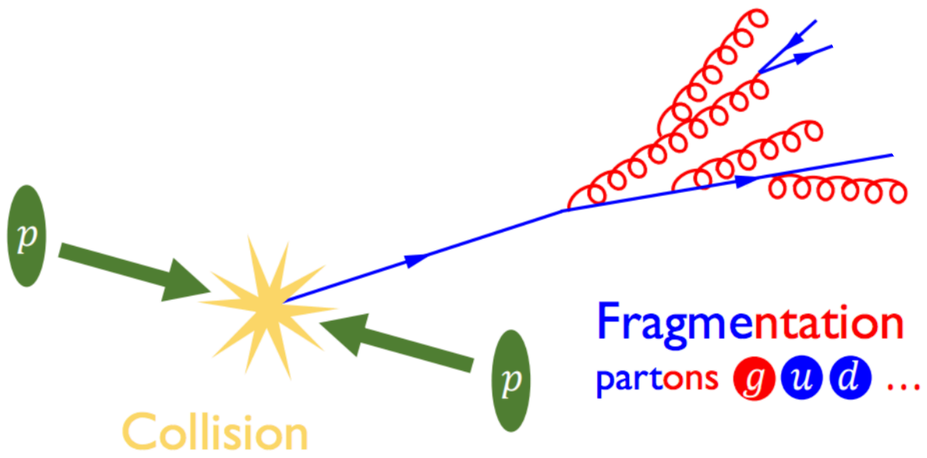
Jets

Due to QCD confinement we do not see quarks in isolation

→ *only exists in confinement of a hadron*

Parton Shower

Cascade of gluons



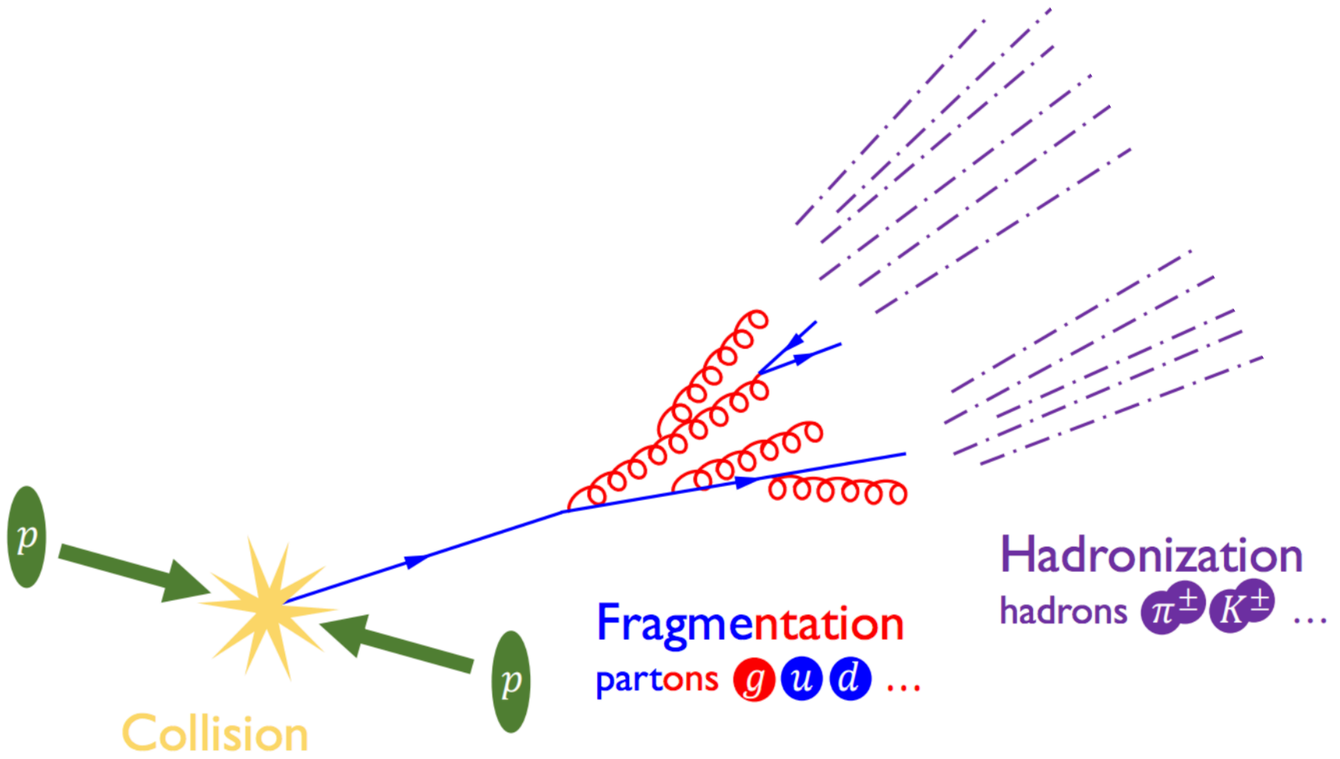
Jets

Due to QCD confinement we do not see quarks in isolation

→ *only exists in confinement of a hadron*

Parton Shower
Cascade of gluons

Jets:
Collection of particles



Jets

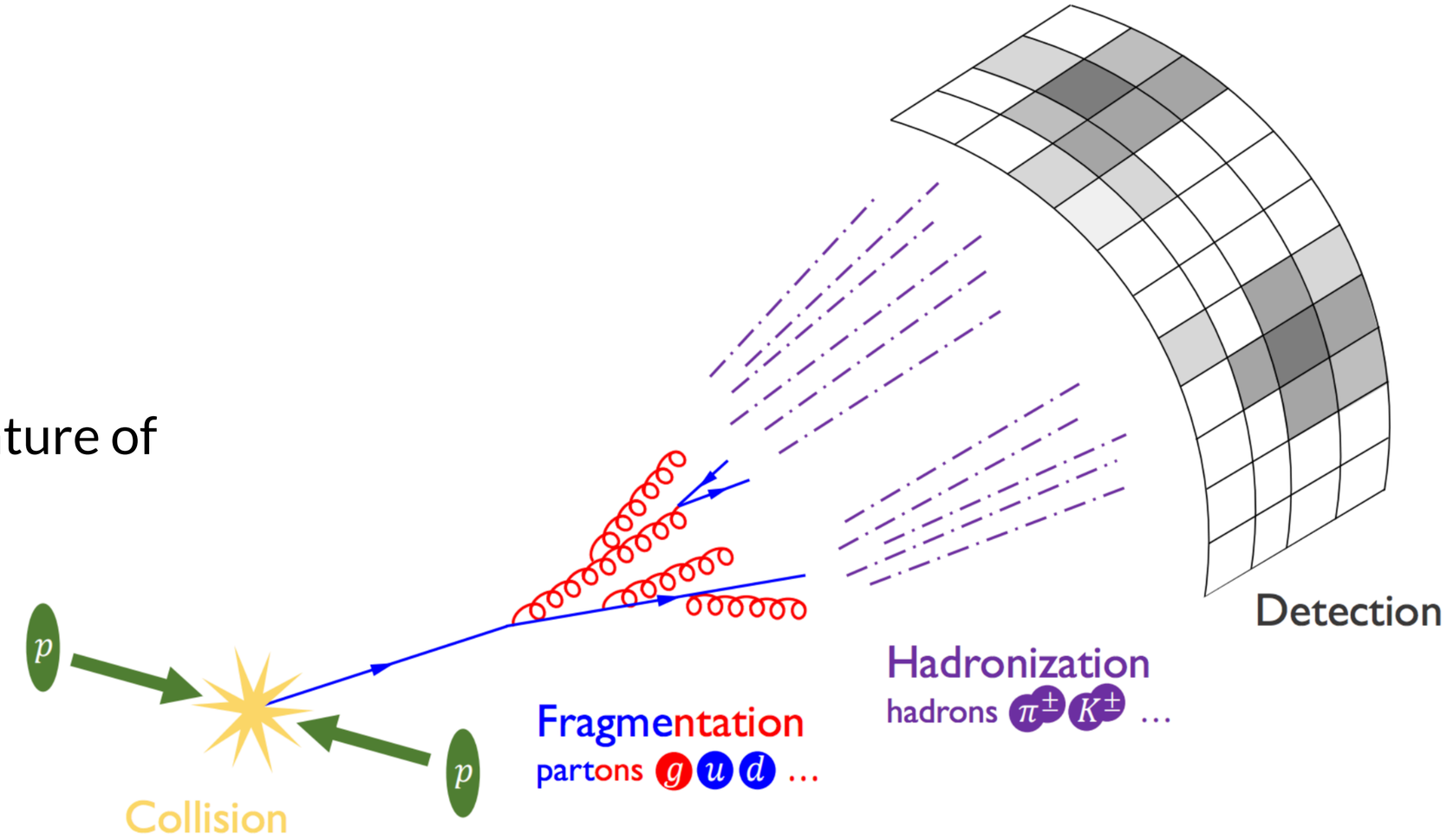
Due to QCD confinement we do not see quarks in isolation

→ *only exists in confinement of a hadron*

Parton Shower
Cascade of gluons

Jets:
Collection of particles

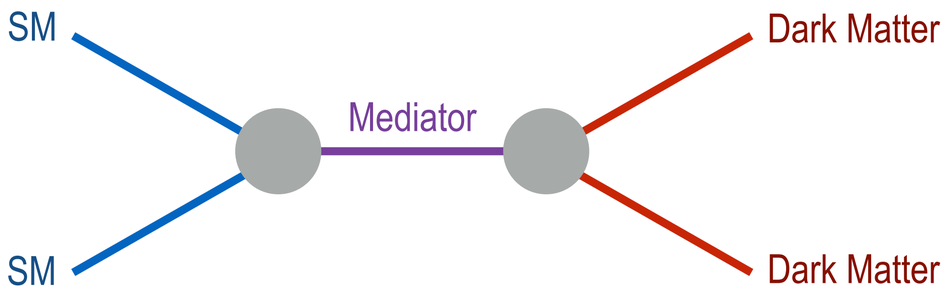
Jets are experimental signature of quarks and gluons



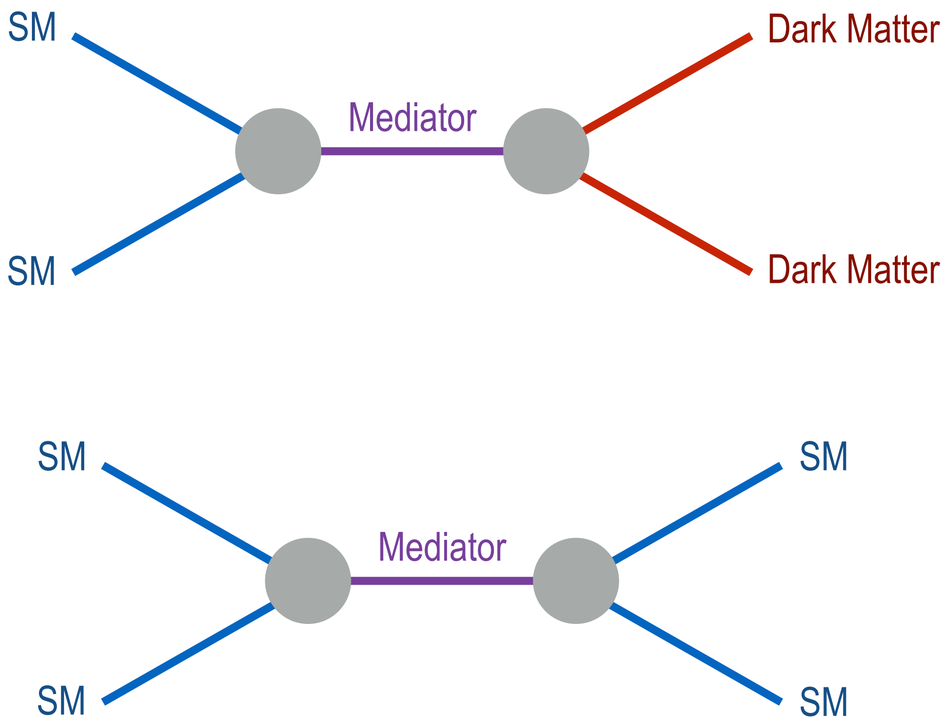
How to look for Dark Matter at ~~a Collider~~ Experiment?

the ATLAS

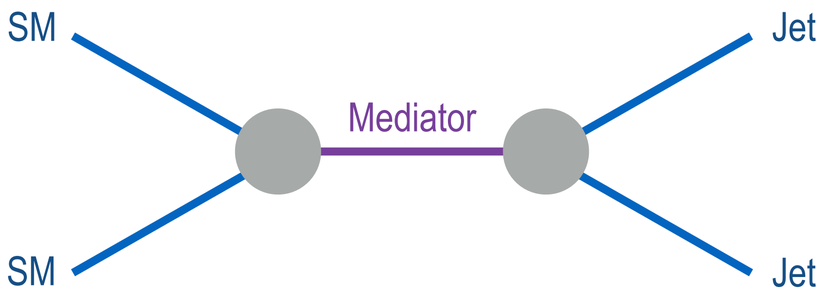
Dark Matter in ATLAS

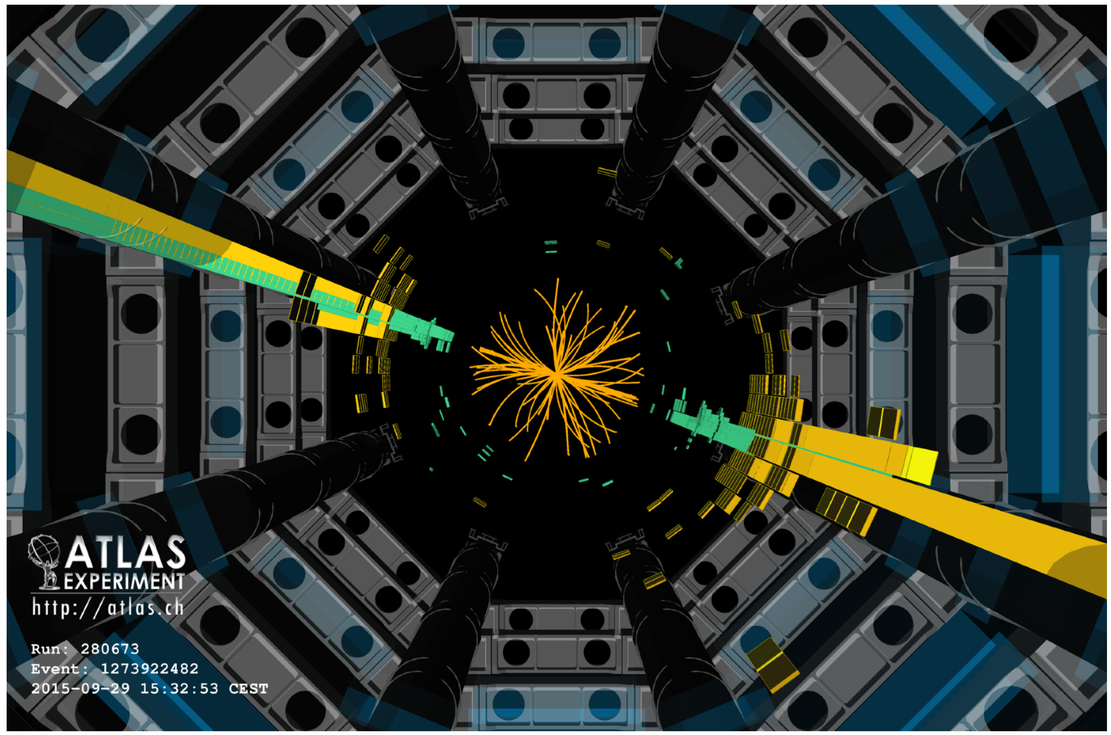


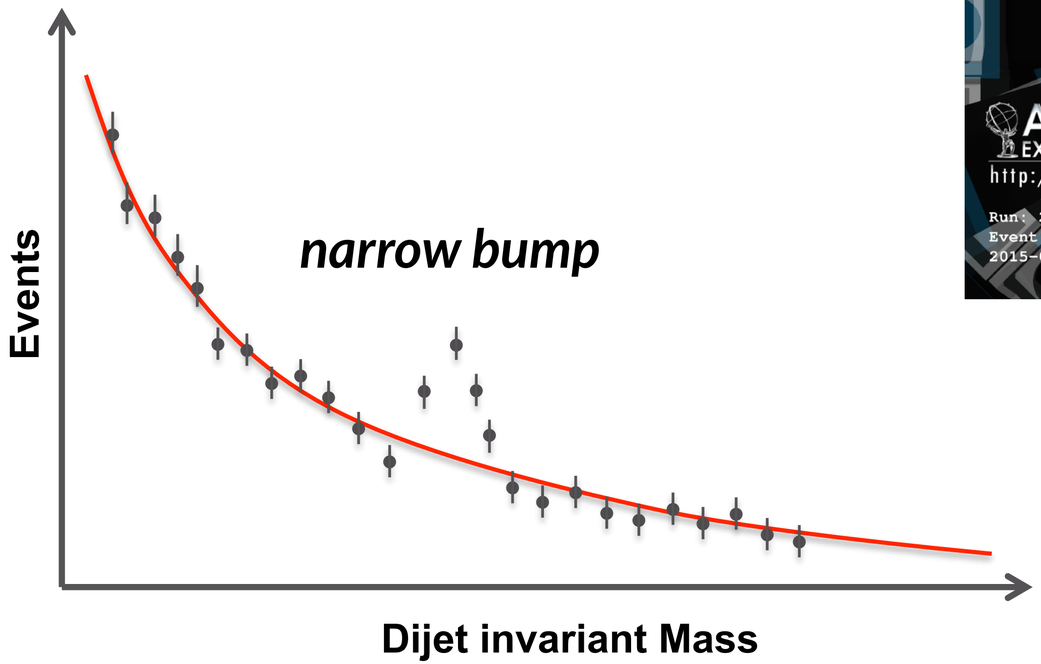
Direct Mediator Search



Direct Mediator Searches in ATLAS

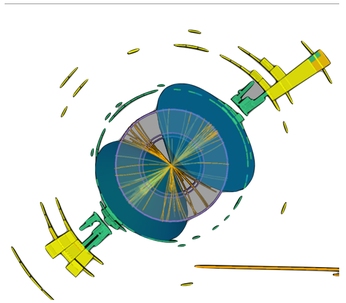


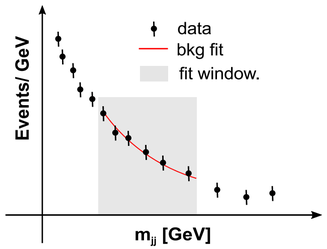


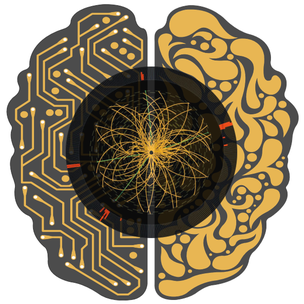


Overview of the talk

Offline







Online

New Physics Search: Model-dependent

- ➡ Example: top-antitop resonance search
- ➡ Highlights challenges → ML-based solutions

Anomaly Detection

- ➡ Resonant anomaly detection with Generative algorithms
- ➡ Future prospects

Fast Machine Learning Inference at the trigger

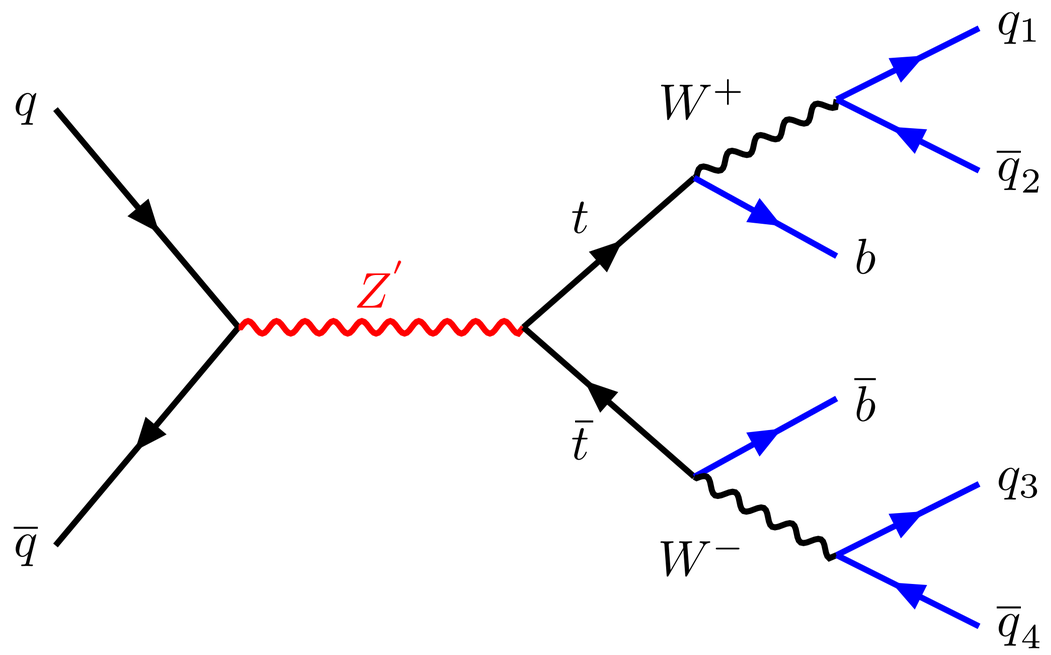
- ➡ Anomaly detection at the low-level trigger
- ➡ Phase-II trigger upgrade
- ➡ Future prospects

Top anti-top Quark Final States

Two major decay channels

1. All quarks → all-hadronic (46%)
2. Leptons + quarks → leptons + jets (~30% with e & μ

ATLAS $t\bar{t}$ resonance Search
JHEP 10 (2020) 061

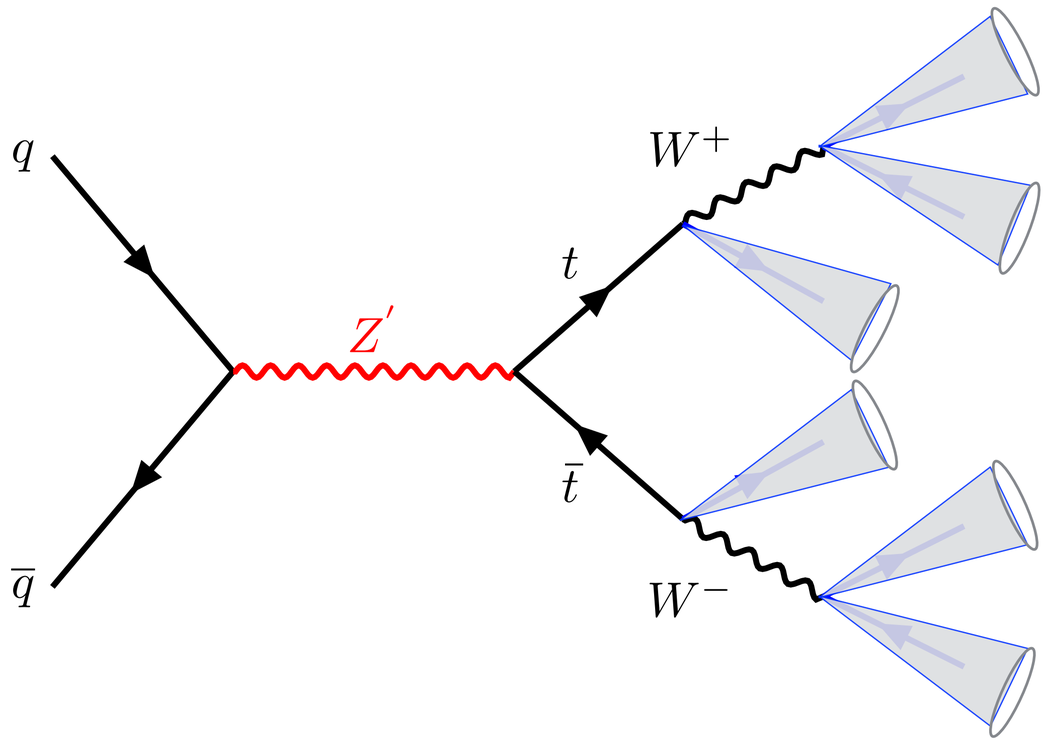


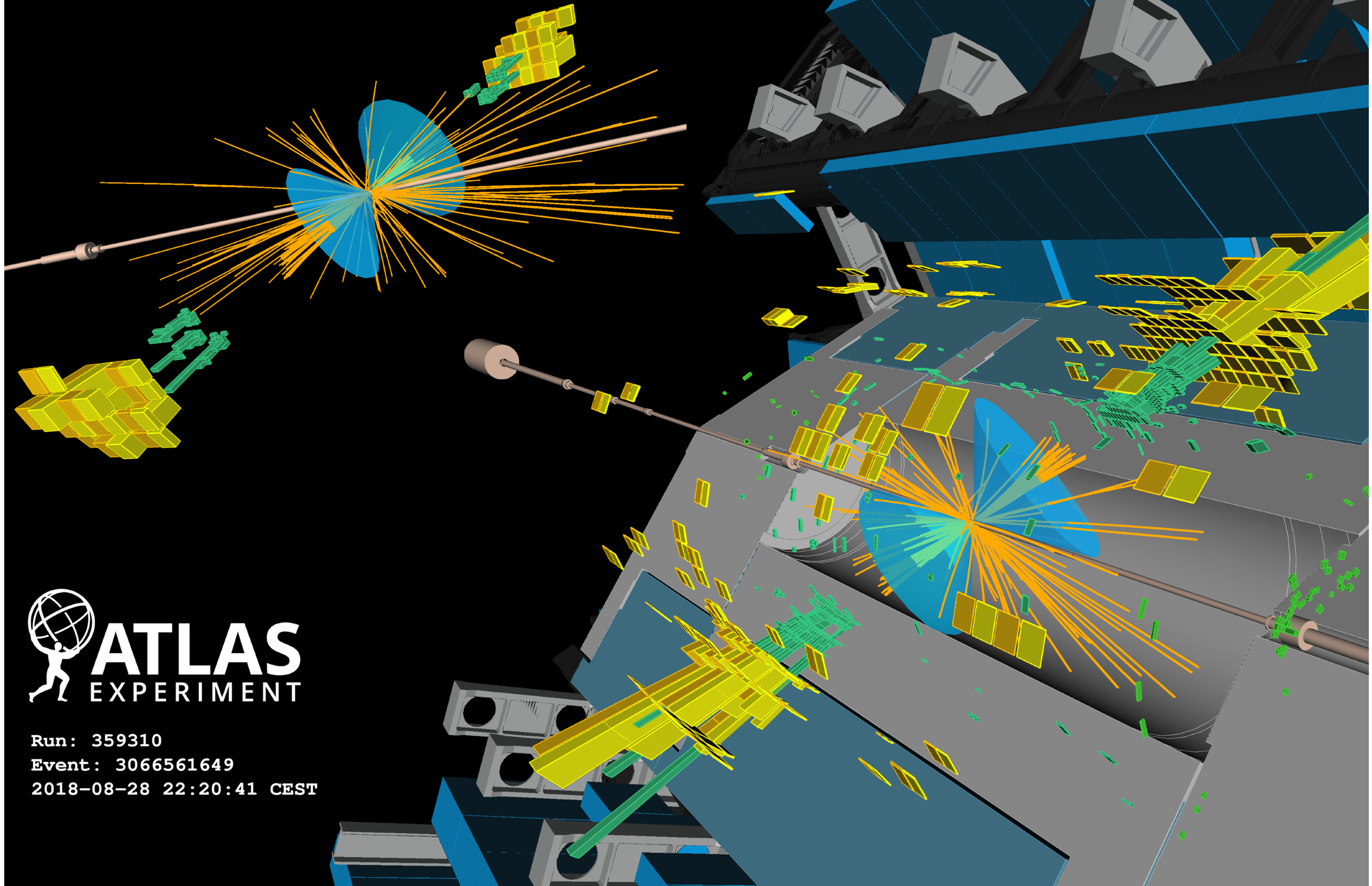
All-hadronic final state

All-hadronic final state

1. All quarks → all-hadronic (46%)

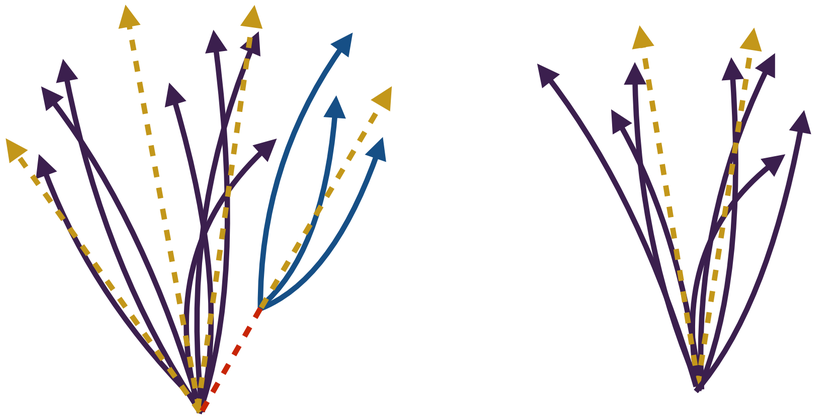
ATLAS $t\bar{t}$ resonance Search
JHEP 10 (2020) 061



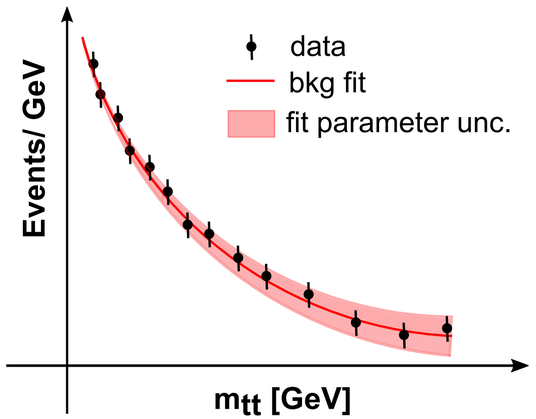

ATLAS
EXPERIMENT
Run: 359310
Event: 3066561649
2018-08-28 22:20:41 CEST

A few Challenges

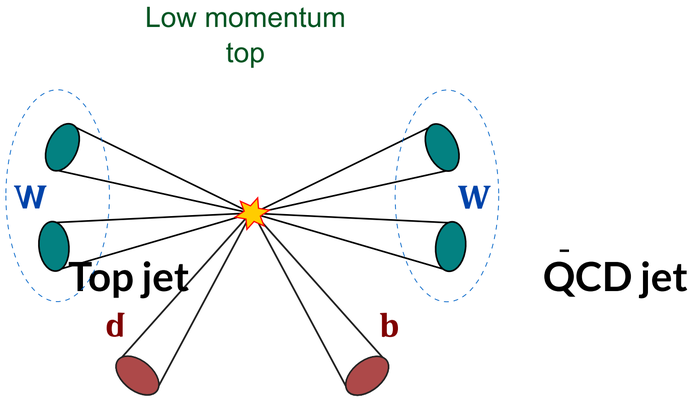
Identifying signals from background



Estimating backgrounds



Reconstructing the events

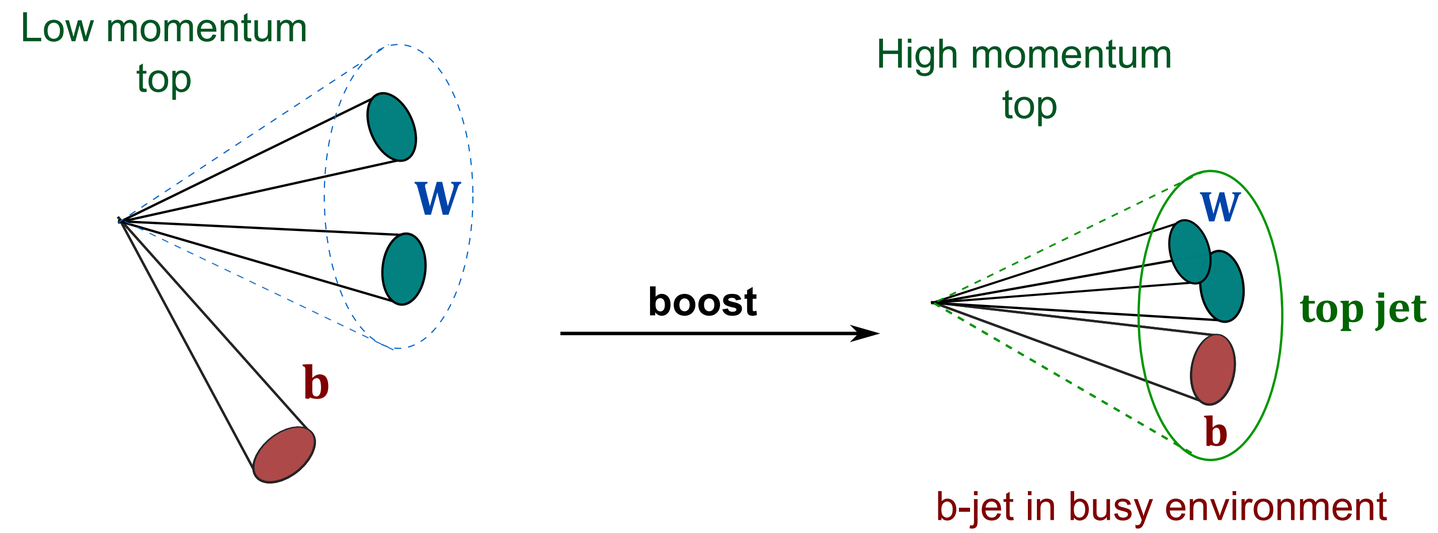


topology of **resolved** analysis

Combinatorial challenge!

High pT top quark decay

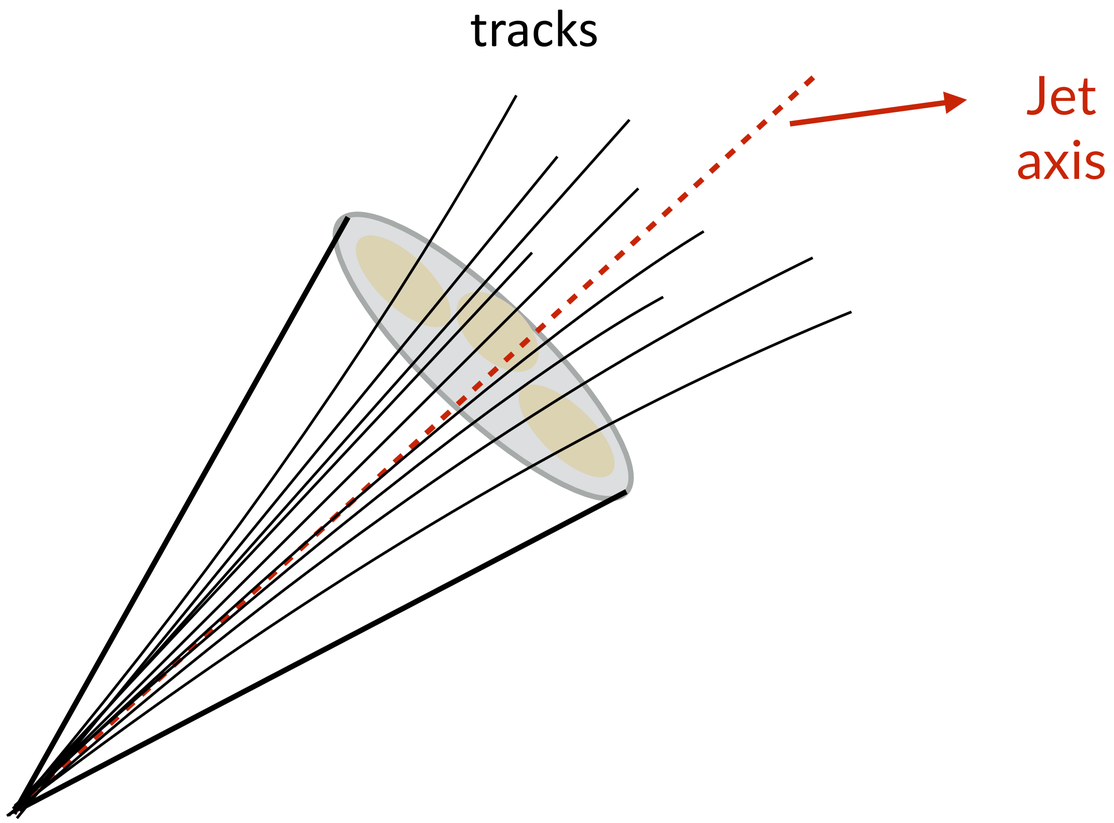
Decay products of a high momentum top quark get collimated along the top quark momentum

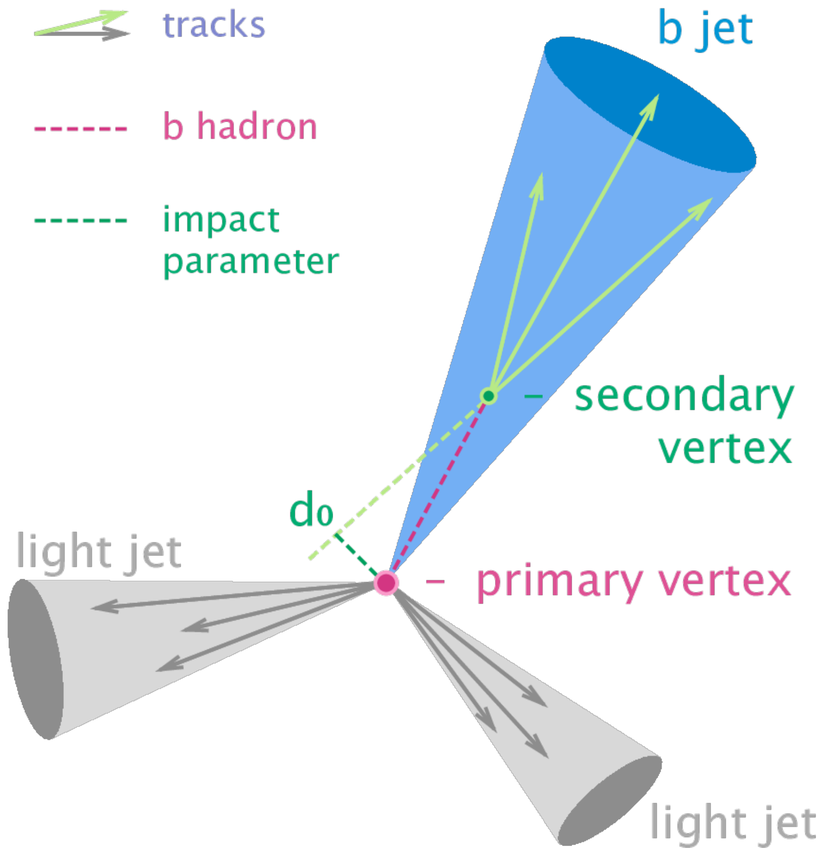


topology of **resolved** analysis

topology of **boosted** analysis

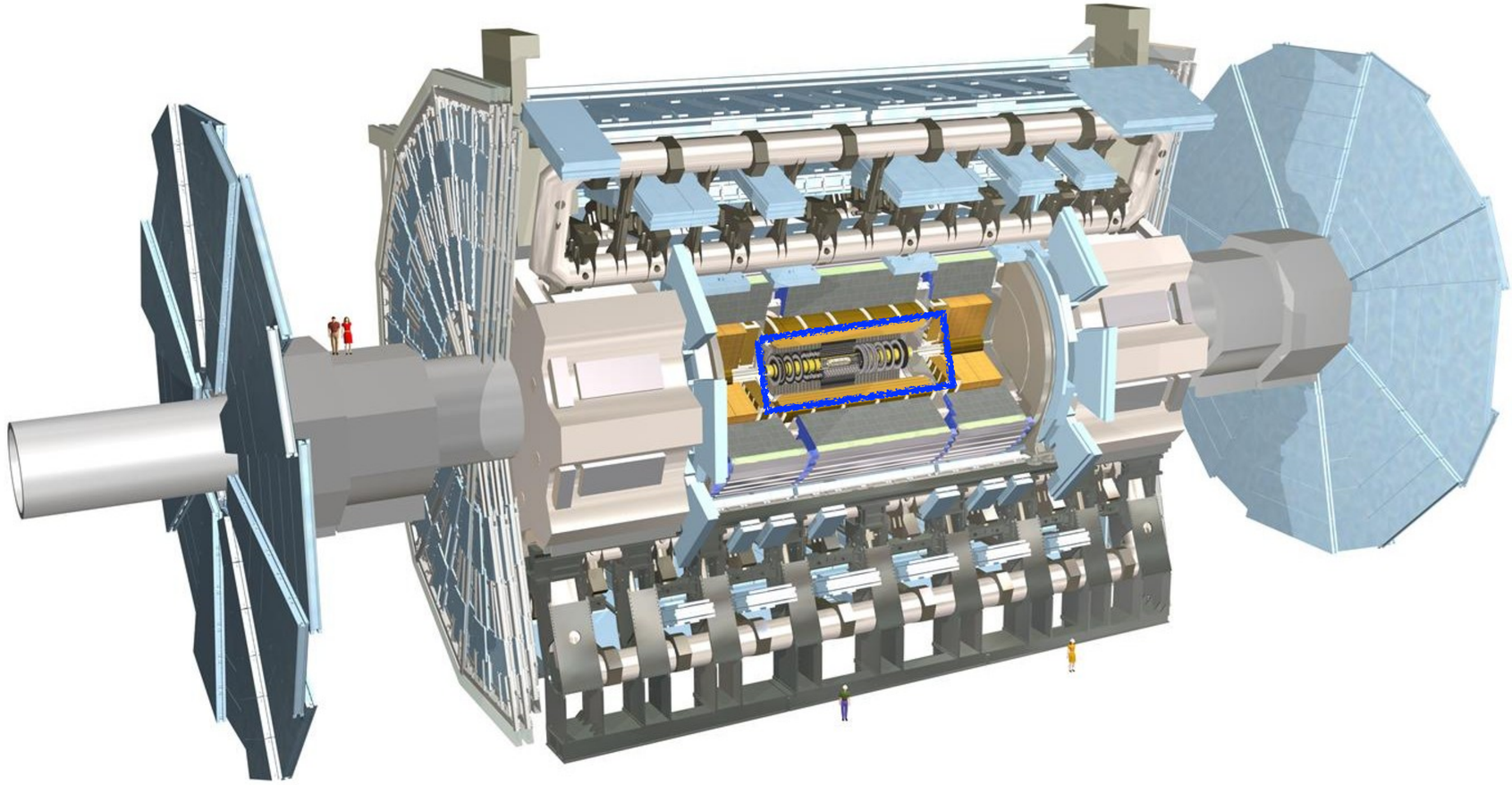
Goal is to identify signal jets (coming from top-quark) from the other QCD jets (background)





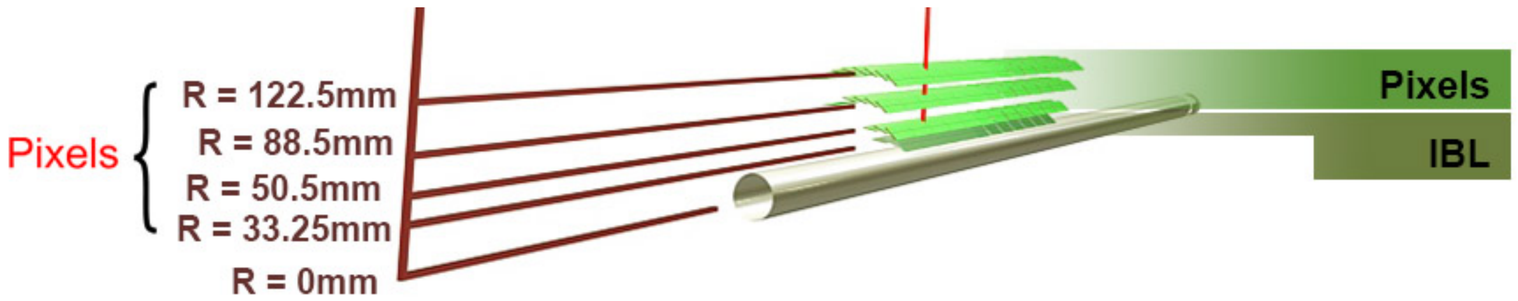
b-tagging plays an important role in Identifying the jets

The ATLAS Experiment

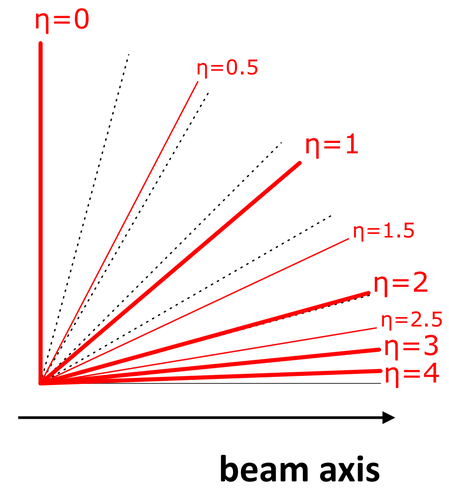


ATLAS Pixel Detector

- Total coverage |η| < 2.5
- Inside 2T solenoid field



$$\eta = -\ln\tan(\frac{\theta}{2})$$



Pixel: 3 layers, 3 disks, 50 × 400μm^2

IBL: 1 barrel layer, 50 × 250μm^2

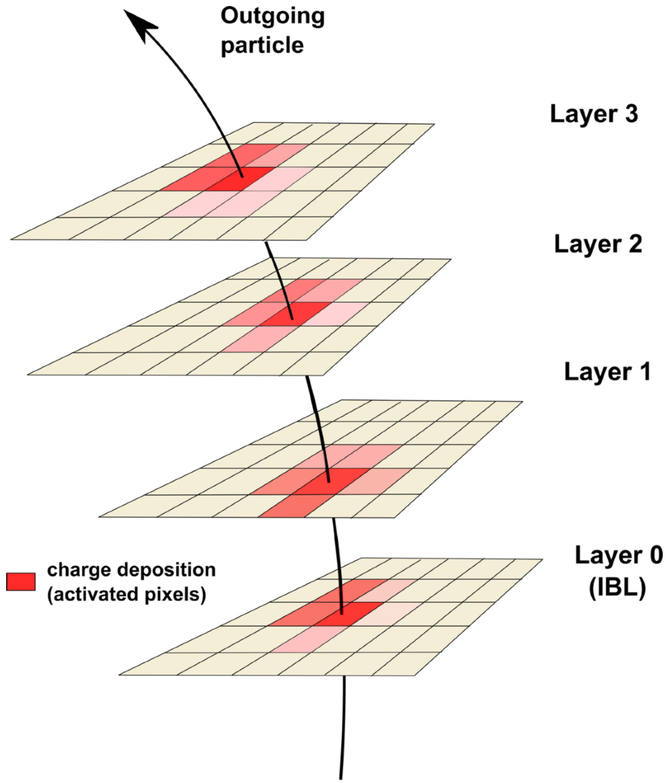
Tracking in Dense Environment

Dense Environment: average separation between highly collimated tracks is comparable to the granularity of individual sensors

EK, PoS LHCP2019 (2019) 009

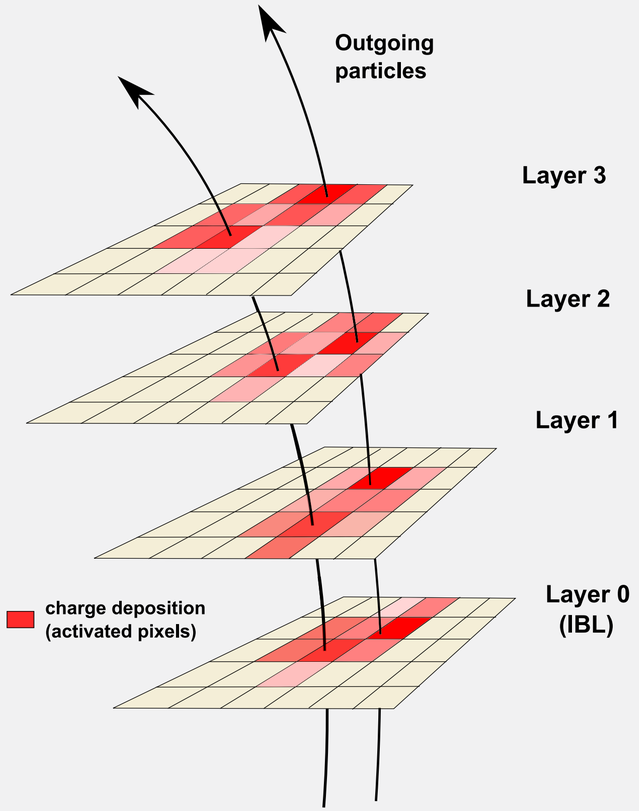
Ex: Cores of highly energetic hadronic top jets

Isolated track

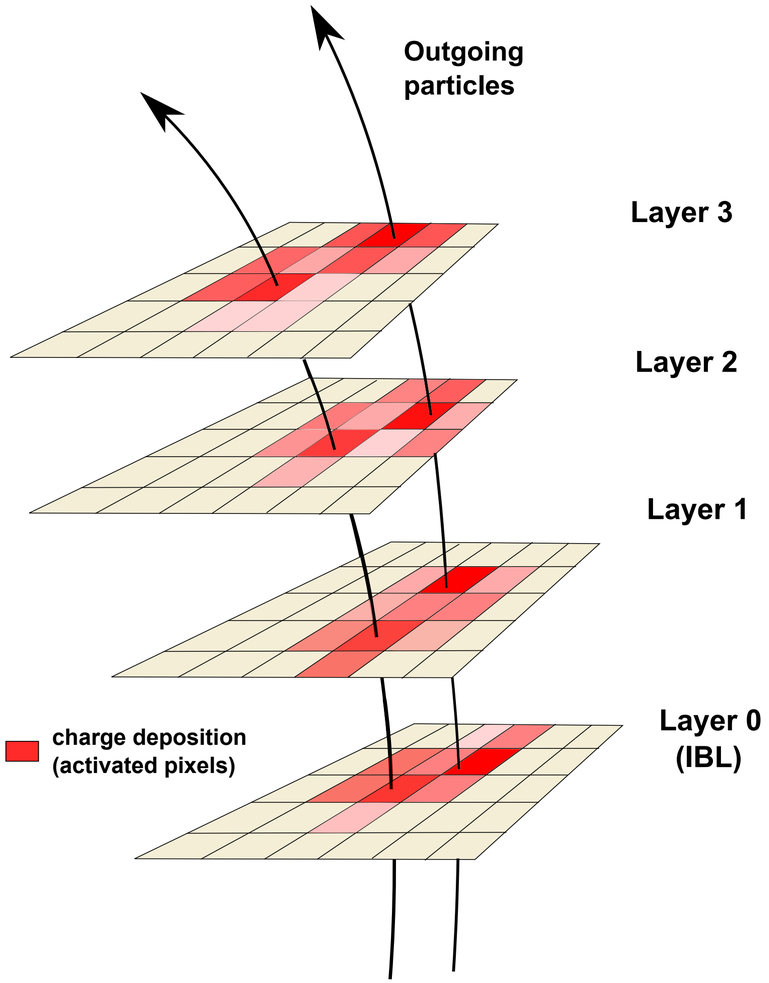


Merged clusters

Tracks are too close to each-other



It is important to split the merged hits



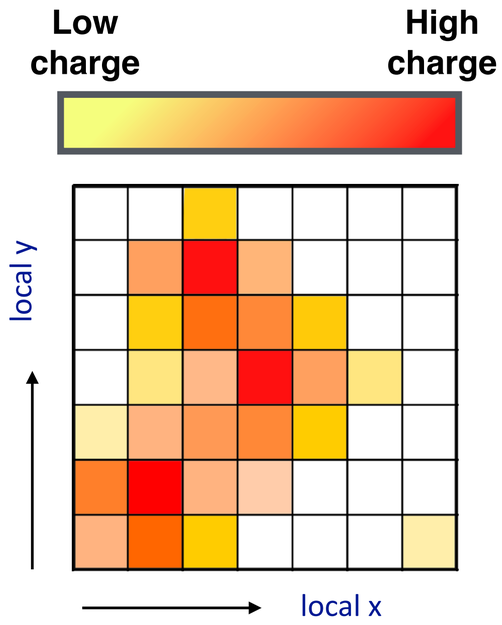
EK, PoS LHCP2019 (2019) 009

- All possible track candidates are scored
- **Shared clusters** penalizes the track score

- **Tracks with low score are not fitted and stored**
  → loss of track reconstruction efficiency

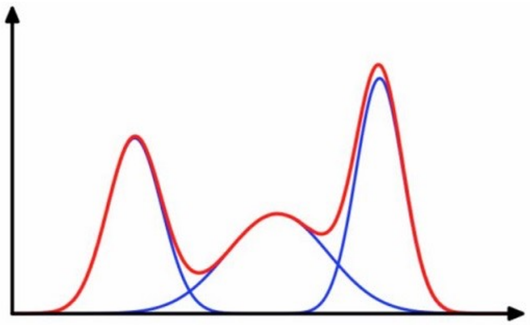
Mixture Density Network

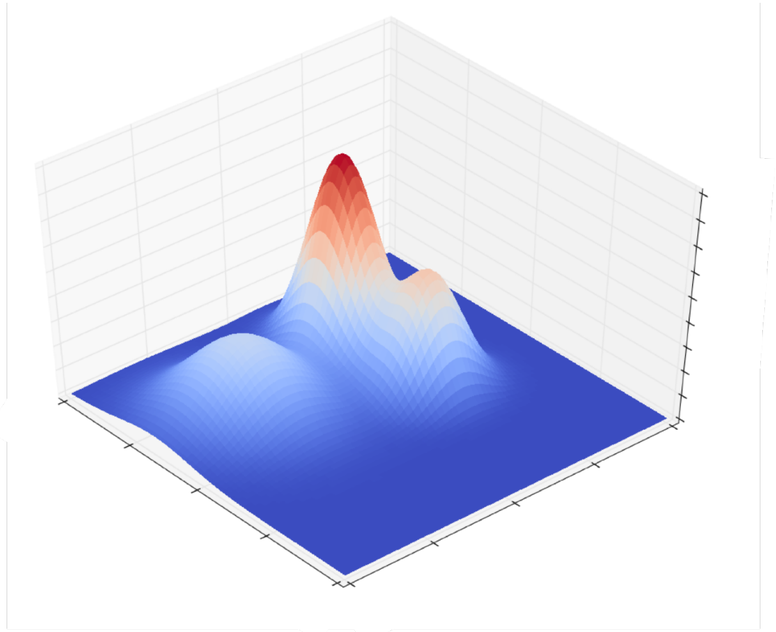
EK, PoS LHCP2019 (2019) 009



Approximating
$p(\text{hit position} \mid \text{input})$

with **Gaussian Mixture model**

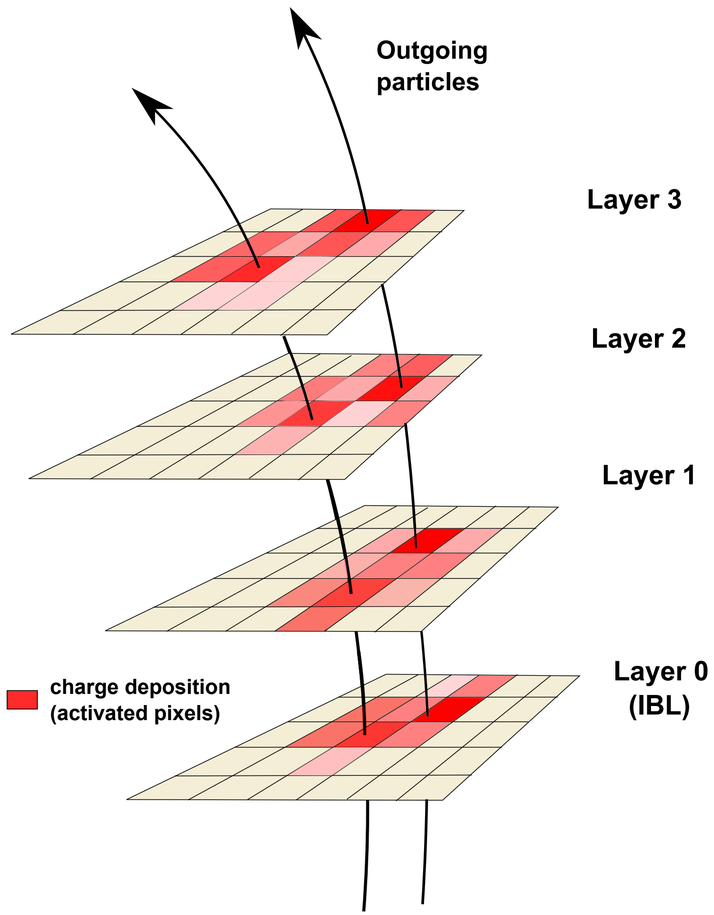




Extract
mean and **std**
of each Gaussian component

MDN in in ATLAS Run-3

MDN is deployed as the baseline algorithm in Run-3

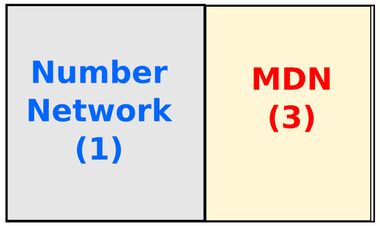


EK, PoS LHCP2019 (2019) 009

ATLAS, tracking paper arXiv:2308.09471

Run 3 ATLAS tracking
4 Neural Networks

Improved track efficiency and b-tagging



Number of hits (Classifier)

Mixture Density Network (Estimates hit position and uncertainty)

Currently in production and running in the experimental data taking

Identifying top jets

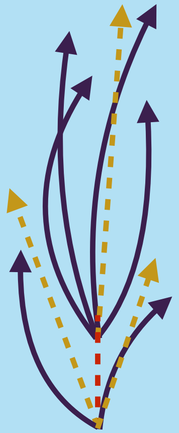
Machine Learning plays an important role

First Run-2 ATLAS analysis
with
Deep Neural Network-based
top tagging method

b / c jets

Jets from
b- or c-quark

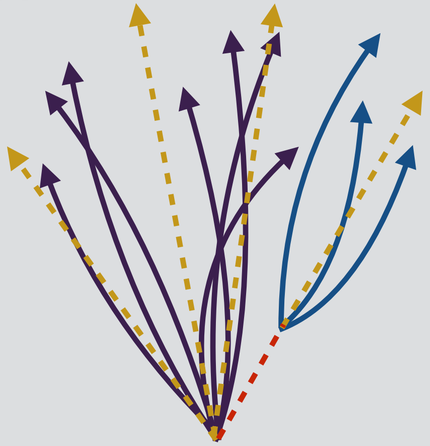
→ Displaces
vertex



top jets

Jets coming from
top-quark

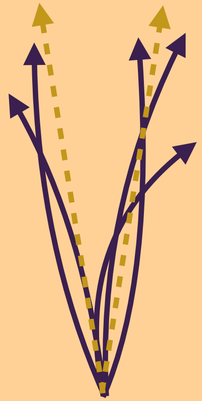
→ 3-prong
structure



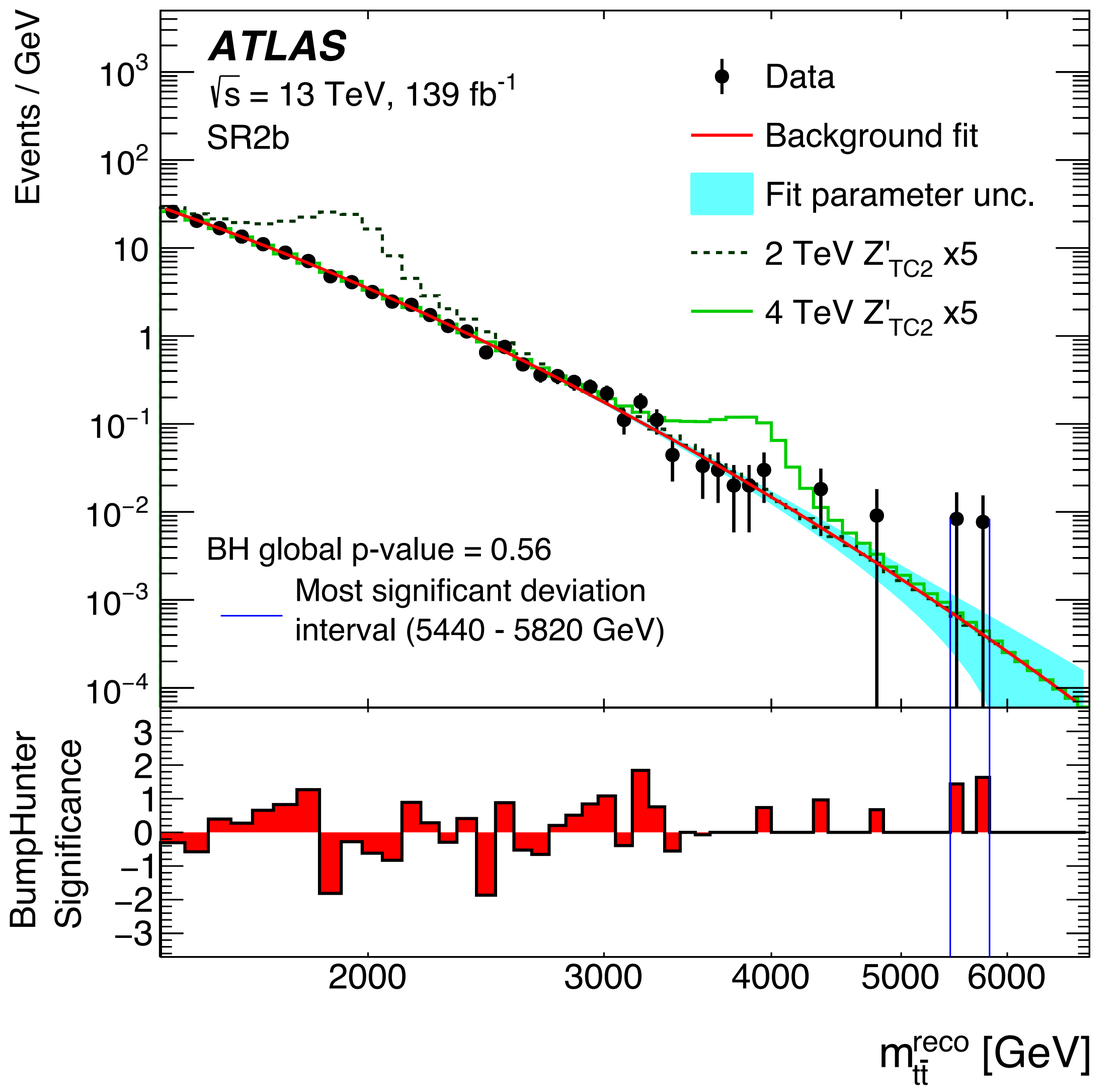
light jets

Jets from
u/d/s-quarks

→ no specific
structure



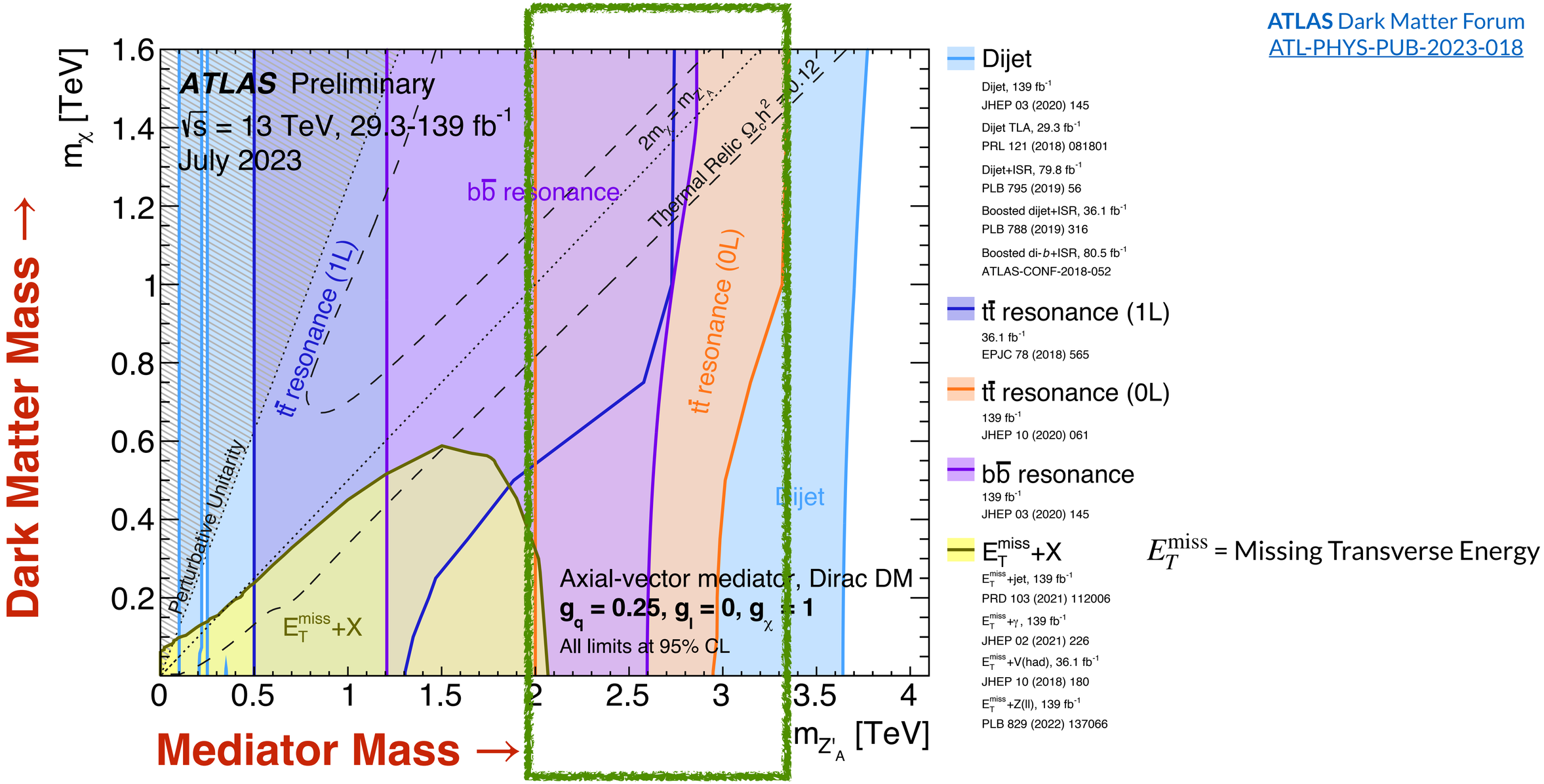
Searching in the data



ATLAS $t\bar{t}$ resonance Search
JHEP 10 (2020) 061

No evidence of
new physics in the data (Full Run 2)

Dark Matter Interpretation



ATLAS Dark Matter Forum
ATL-PHYS-PUB-2023-018

E_T^{miss} = Missing Transverse Energy

Model-dependent Search

Most sensitive approach for a particular new physics scenario

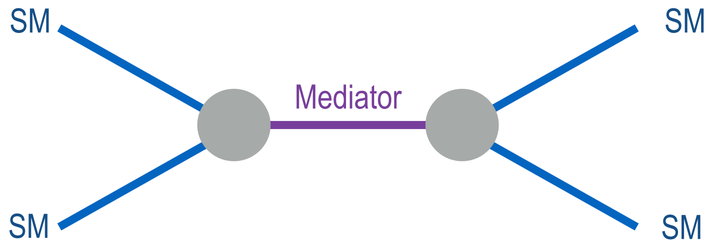
→ Unlikely to probe a different scenario

New Heavy particle Z'

Big Questions

- Nature of Dark Matter
- Matter AntiMatter Asymmetry
- Origin of Neutrino Mass
- Origin of Flavor
- Evolution of the Early Universe

Physics Scenario
Signal model

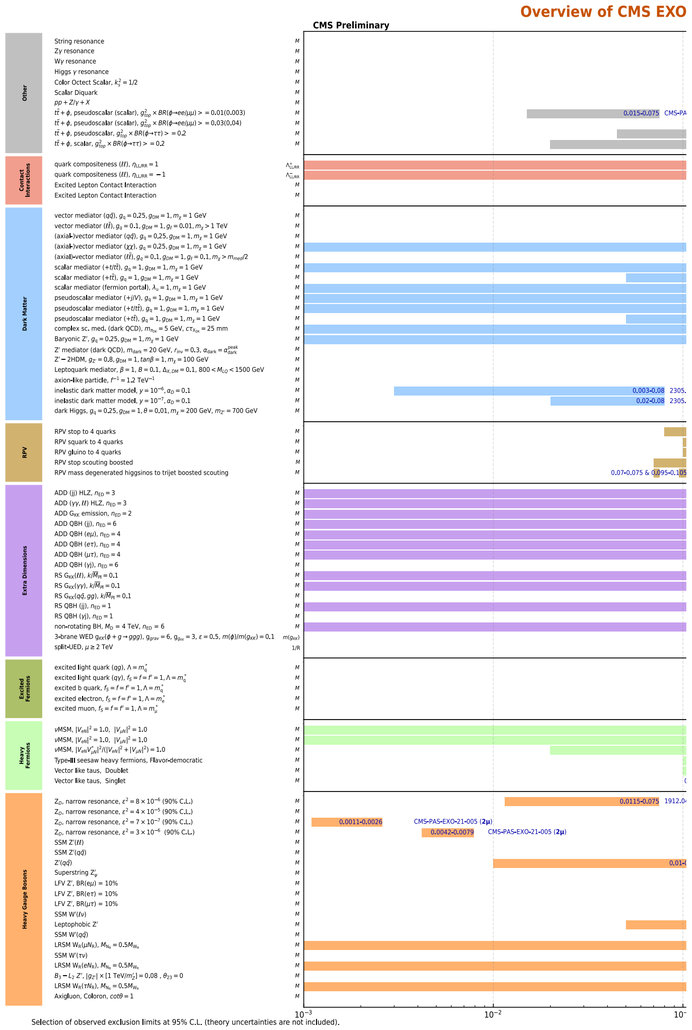


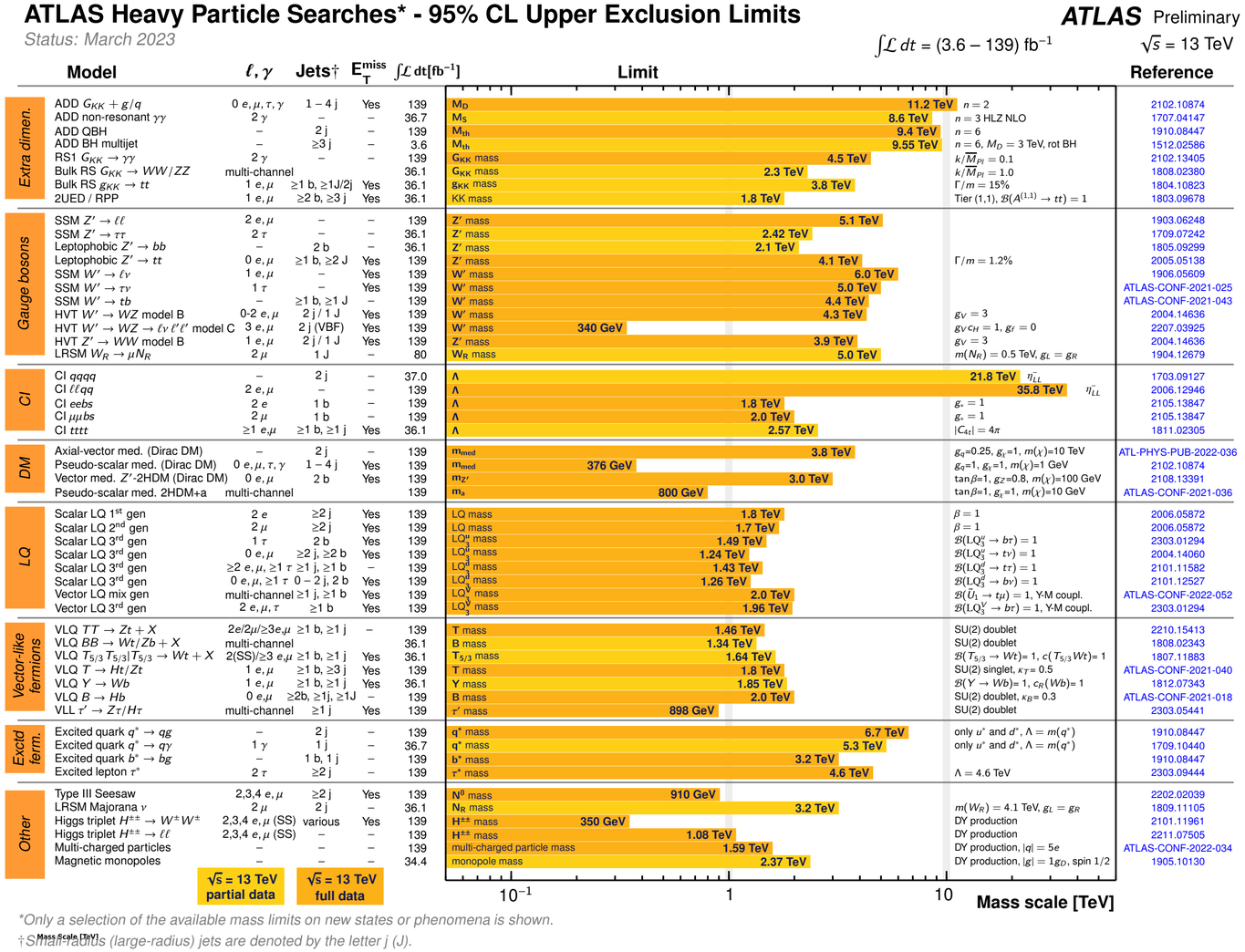
Maximixe Sensitivity

Rich Search program in ATLAS

Dozens of searches in ATLAS and CMS

→ **No sign of BSM physics!!**





May be looking at the wrong place?

Model Independent Search

In many cases we don't even know how the BSM physics would look like!

Change the initial question?

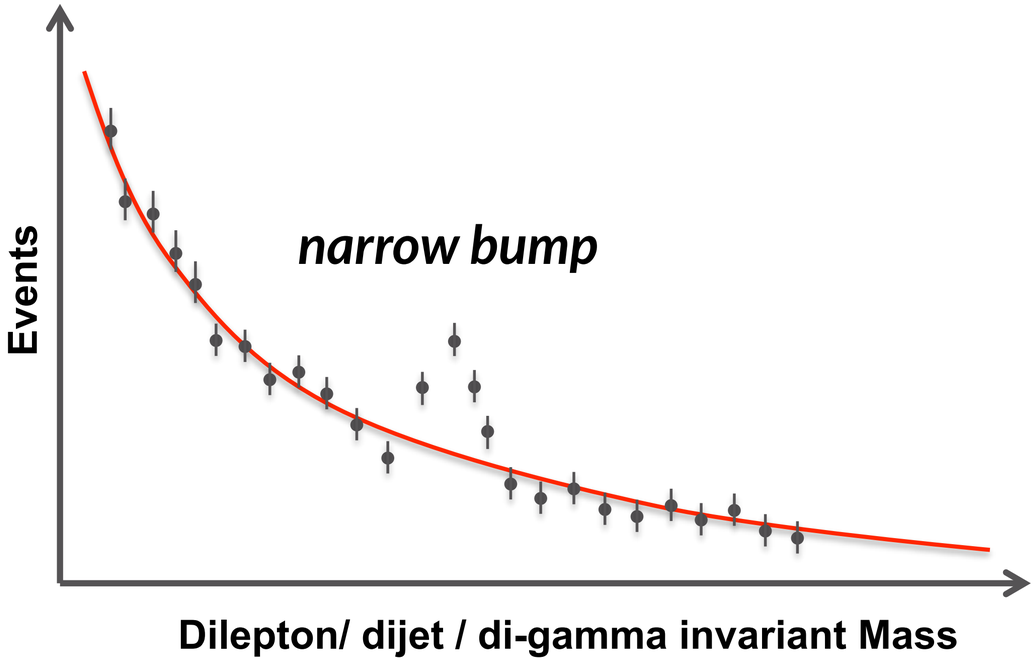
Does this event look like a certain BSM Model?

Does this event look like the Standard Model?

Anomaly detection

Identify data with features that appear inconsistent with those of the majority of the dataset

For HEP application: Interested in an ensemble of events rather than a single outlier



Resonant Anomaly Detection Search

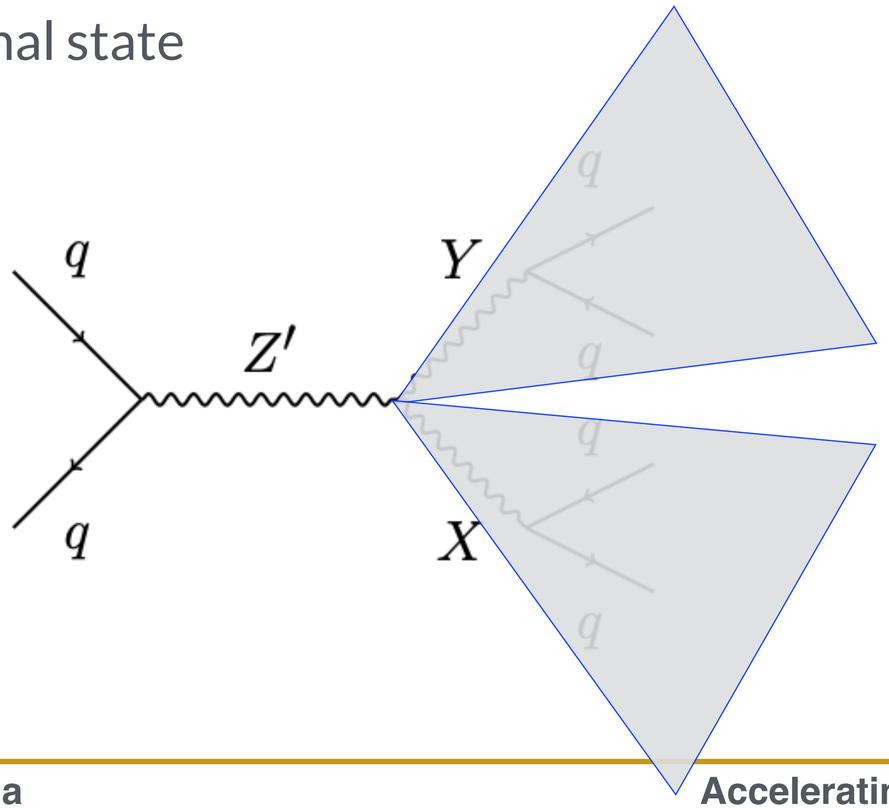
Assumption:

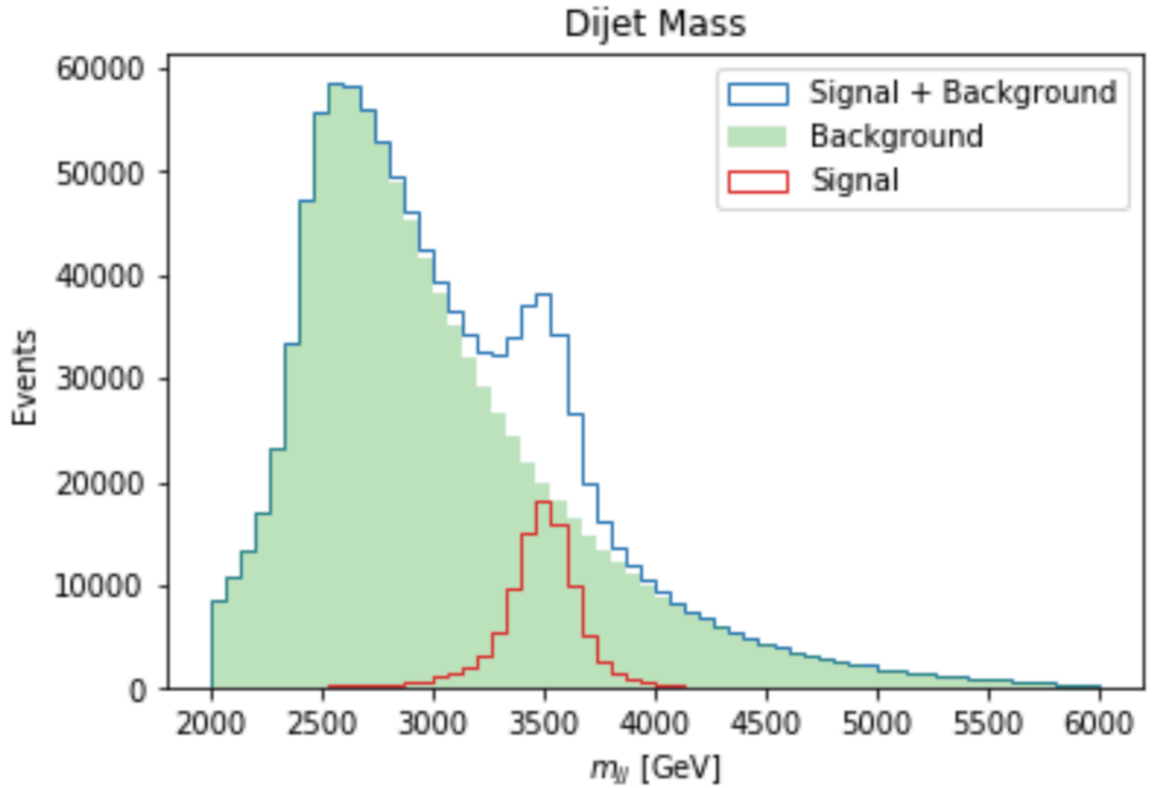
Signal is localized at least in one of the feature spaces (x)

$p_{\text{signal}}(x)/p_{\text{background}}(x)$ is high

Example:

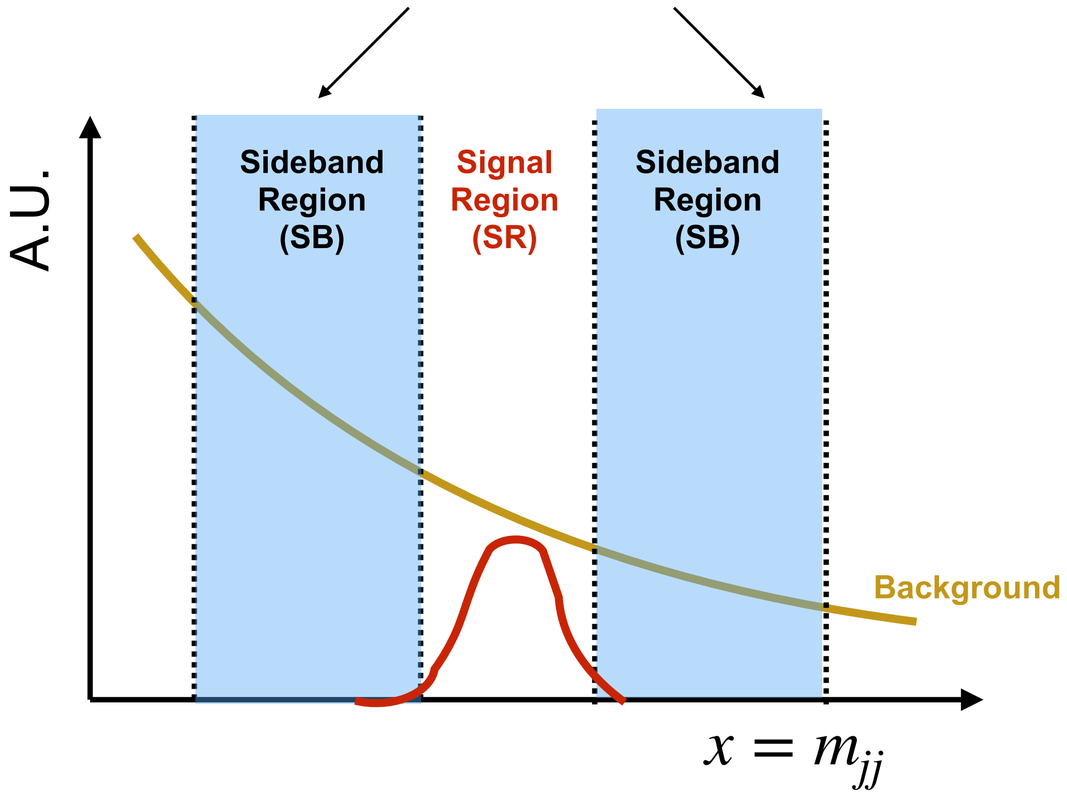
2 jet final state



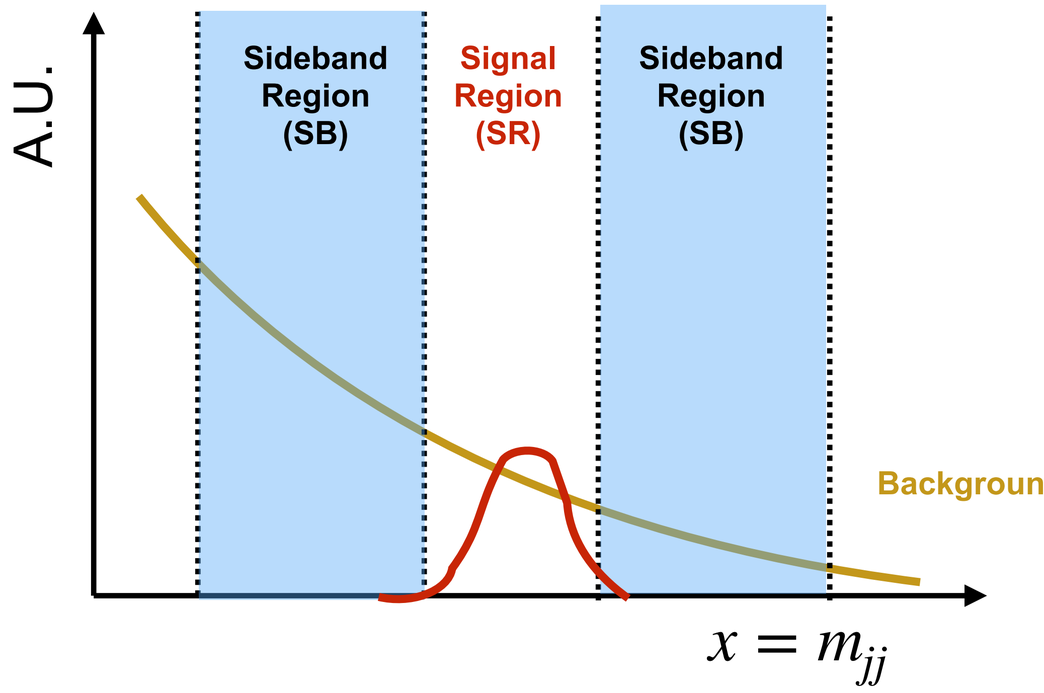


Increase Signal Sensitivity

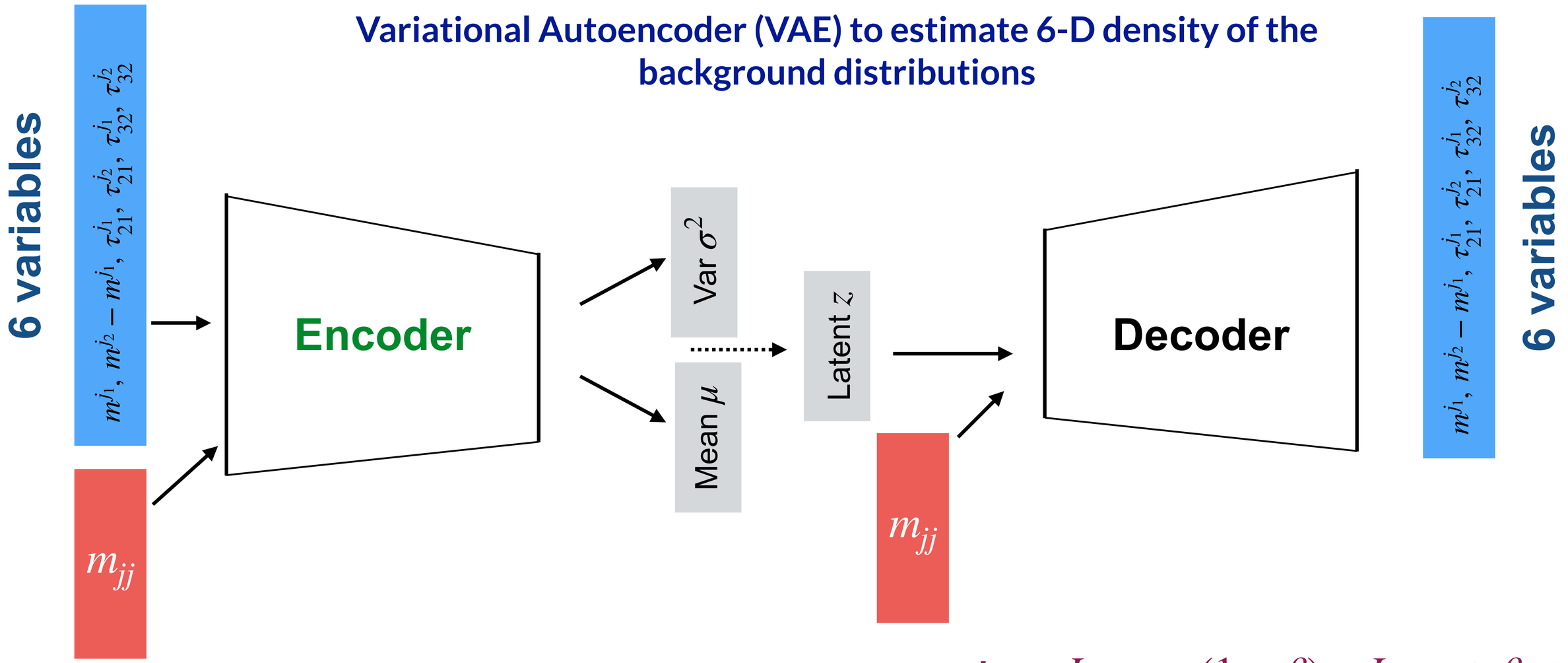
Estimate background
with Generative algorithms



Reduce background keeping as
much signal as possible



VAE as density estimator

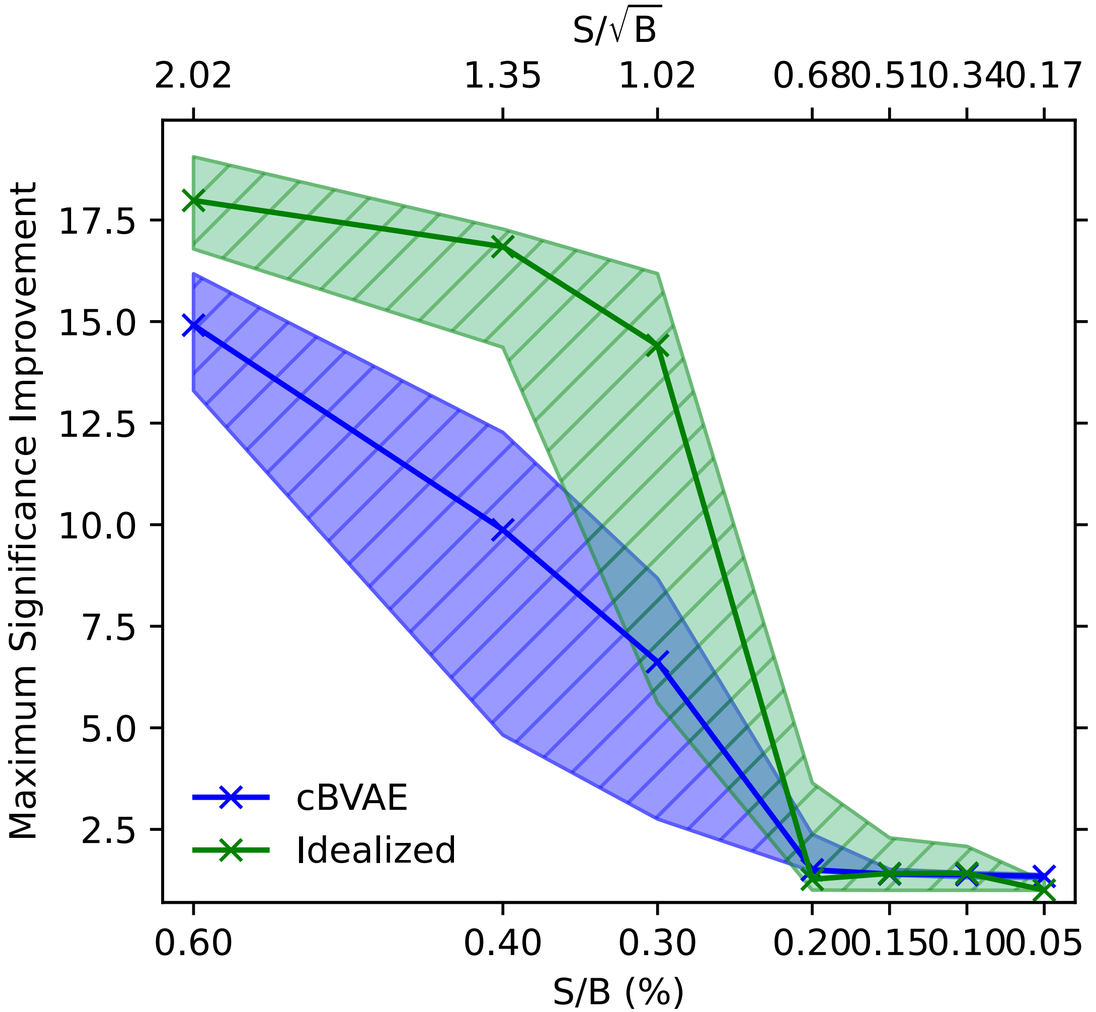


Loss: $L_{VAE} = (1-\beta) \times L_{MSE} + \beta \times KL$

Significance improvement

Significance improvement is close to fully supervised method

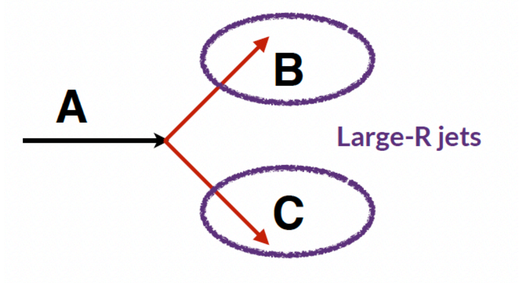
EK, ML4Jets 2022

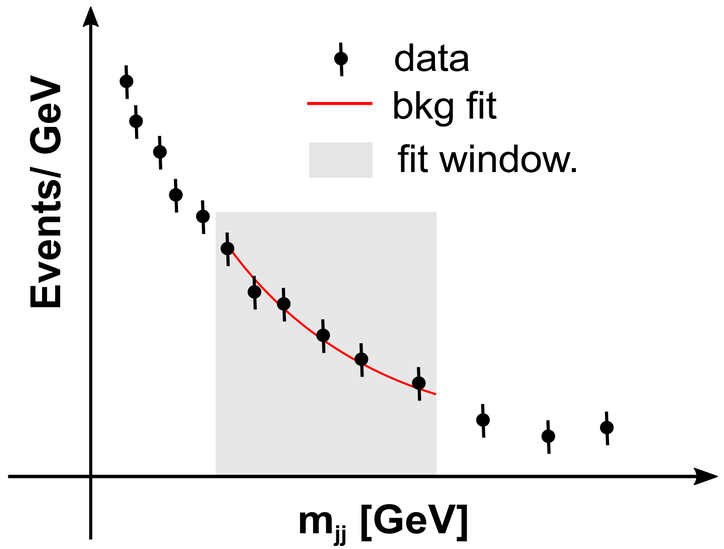


How to use it in an ATLAS analysis?

- Bring this method to ATLAS analysis: di-jet resonance search
- Generic A → B C final state

- Compare contrast with the Anomaly detection methods → exploit the complementarity

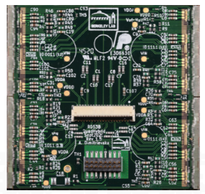




But are we storing all these anomalous events?

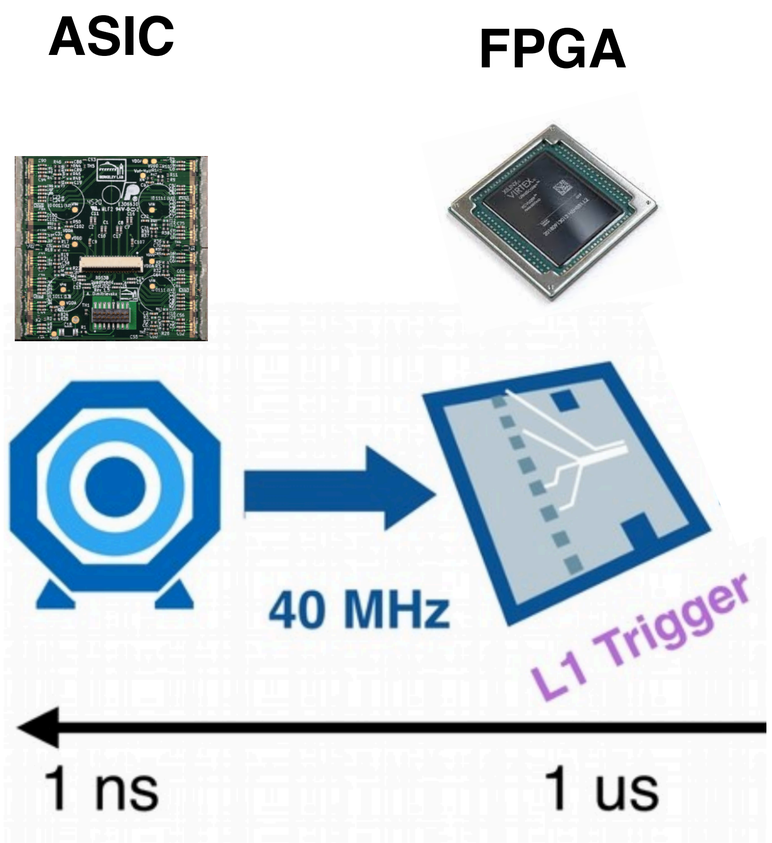
ATLAS Run-3 Data Processing

ASIC

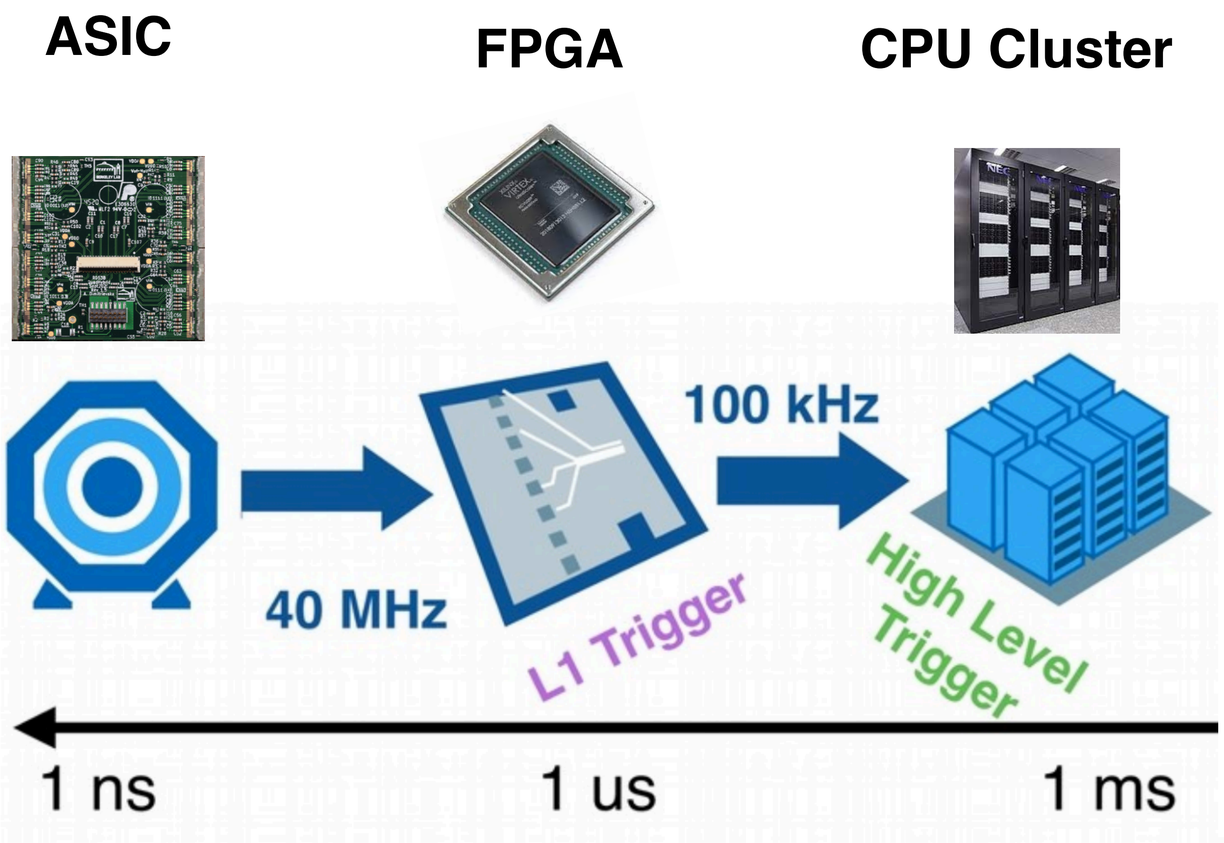


1 ns

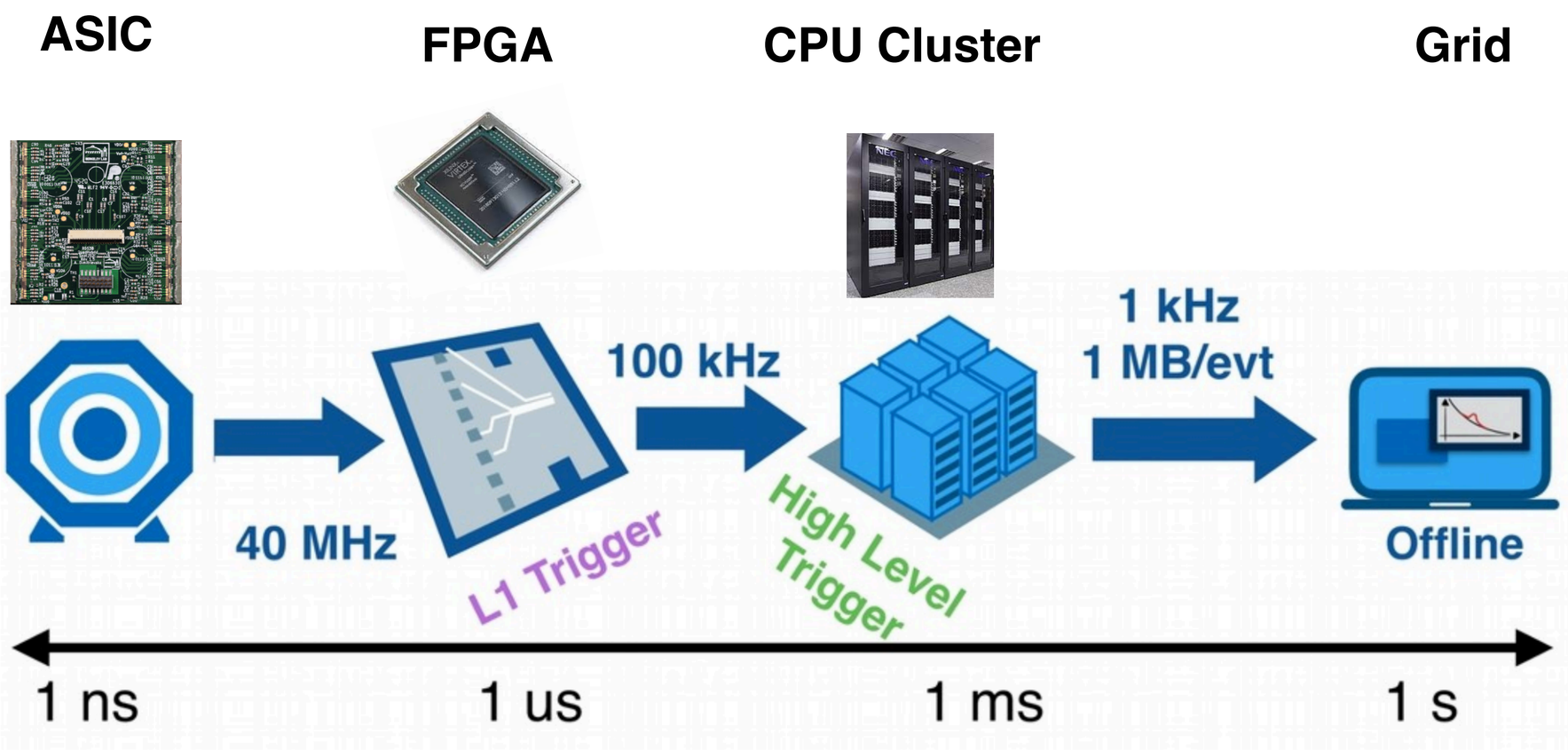
ATLAS Run-3 Data Processing



ATLAS Run-3 Data Processing

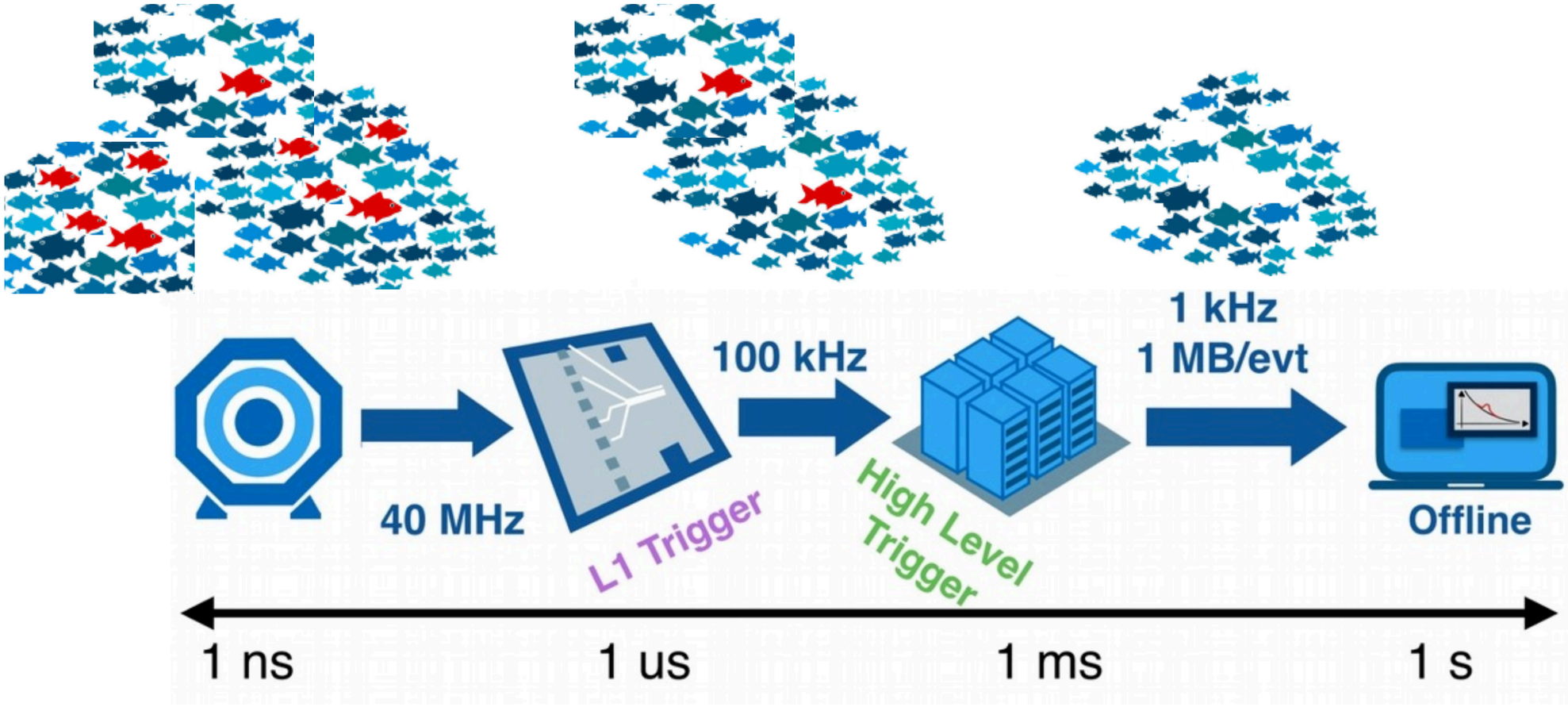


ATLAS Run-3 Data Processing

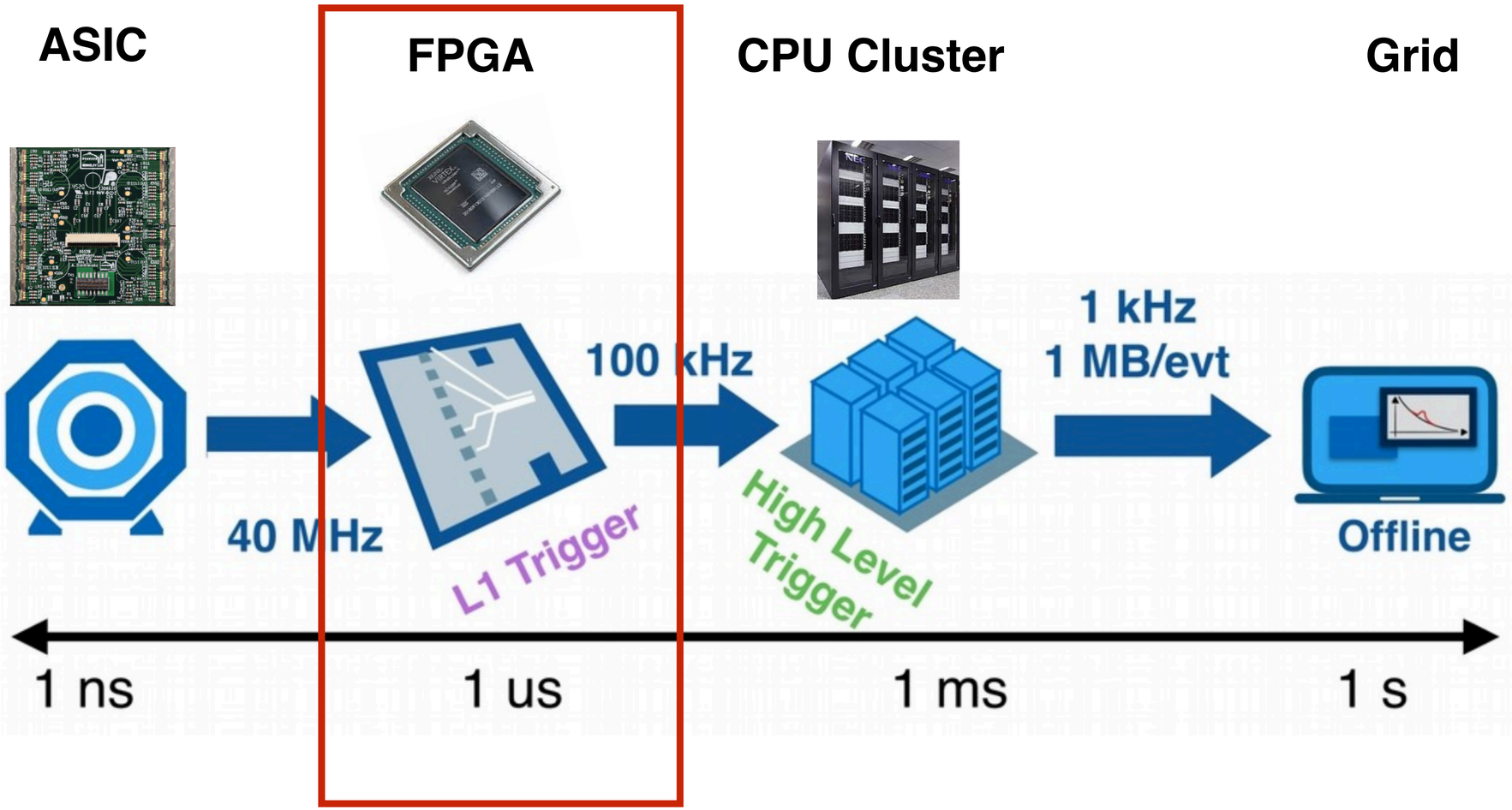


Are we storing them?

- New physics is clearly very good at hiding from us
- Depending on anomaly, we could have none left in recorded data



ATLAS Run-3 Data Processing



Start more ML algorithms here

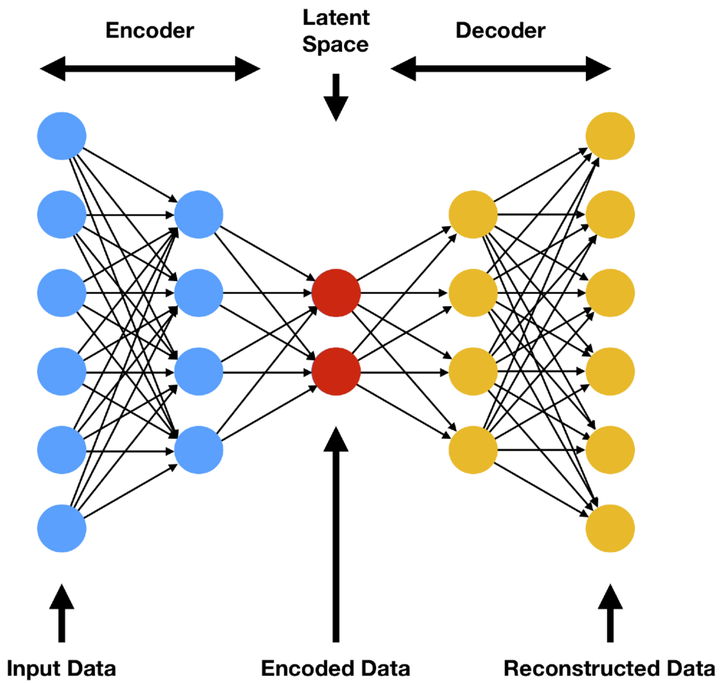
Can we do Anomaly Detection in the Trigger?

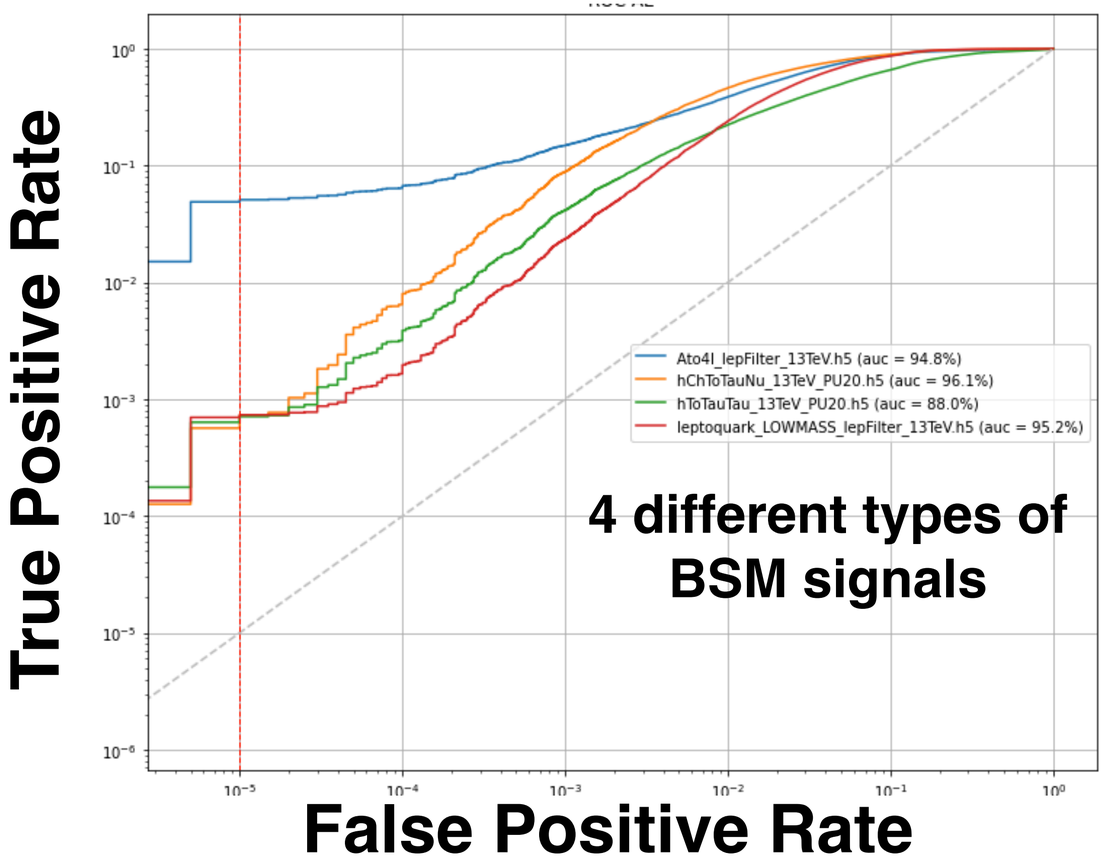
Yes

One Possible Method: Autoencoder

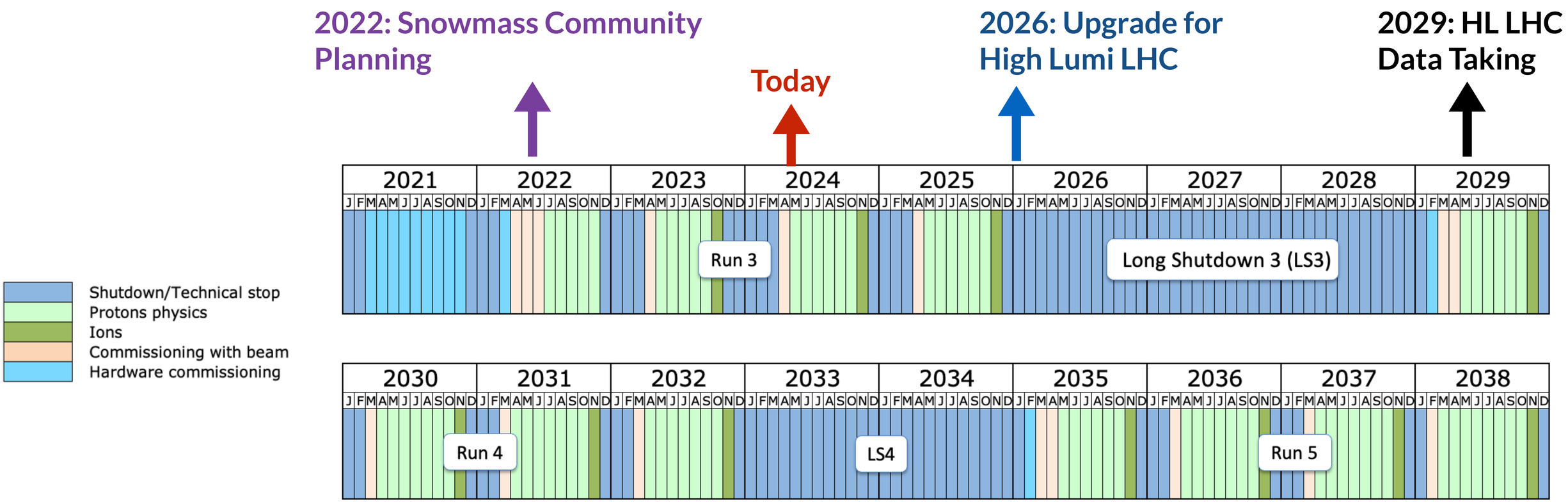
Autoencoders or Variational Autoencoders

- Reconstruction loss between input and output could be used as anomaly score
- CMS is already using this idea in their Run-3 trigger





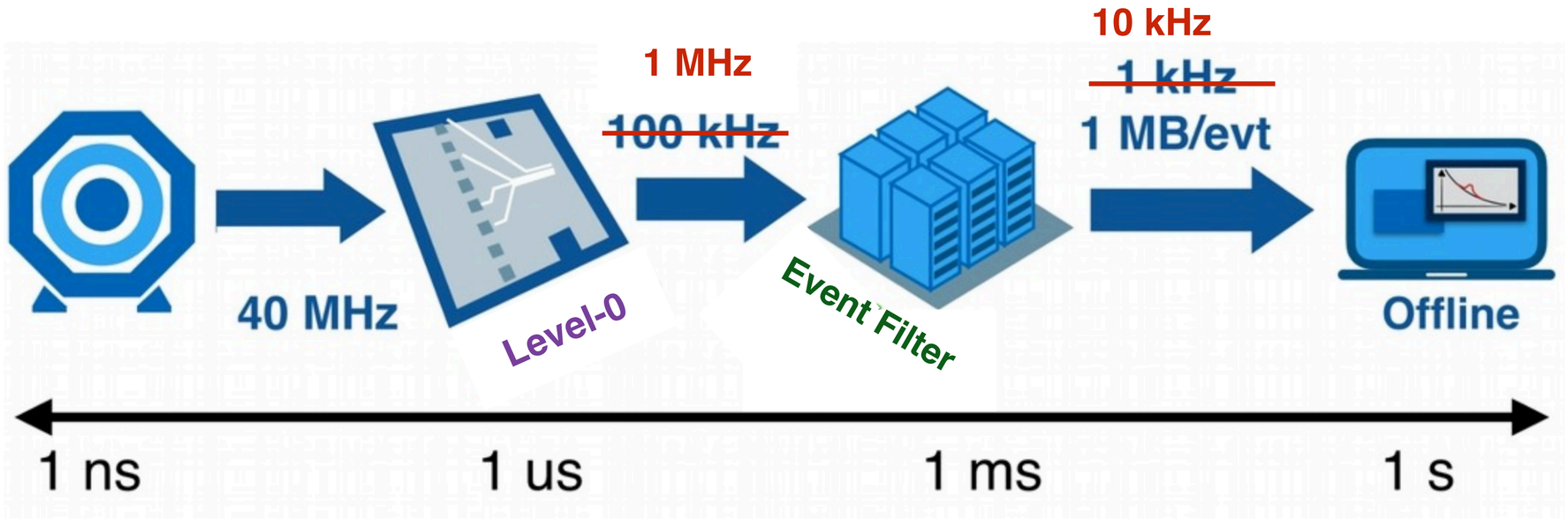
Future of the LHC



High Luminosity LHC (HL-LHC) *starts in 2029*

- 4x increase in average number of collision

ATLAS HL-LHC Data Processing: Online



ATLAS detector upgrade:
Many subsystems will be upgraded to be compatible with high occupancy / trigger rates

Upgrade in Detector Readout
- *10x faster data collection*
- Better hardware in Level-0 and EventFilter

Level-0 Hardware Trigger

hls4ml: A software interface for implementing Neural Networks on an FPGA

ML tools (python-based)

(Q) Keras + TensorFlow

(Q) ONNX

PYTORCH

hls4ml

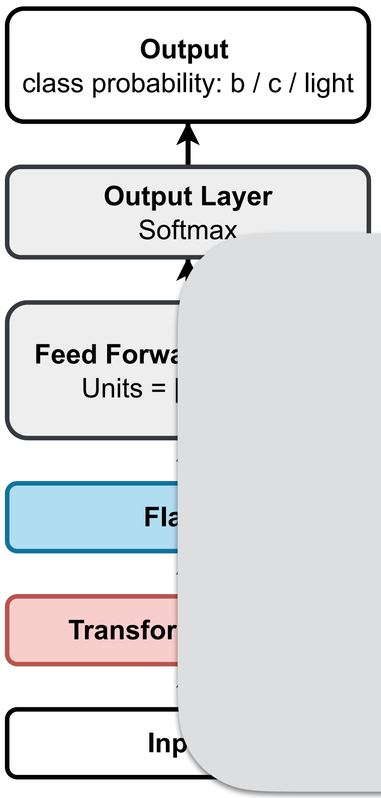
Converts into hardware synthesizeable Instructions (VHDL/Verilog)

- Algorithm processing platform to evaluate ML algorithms

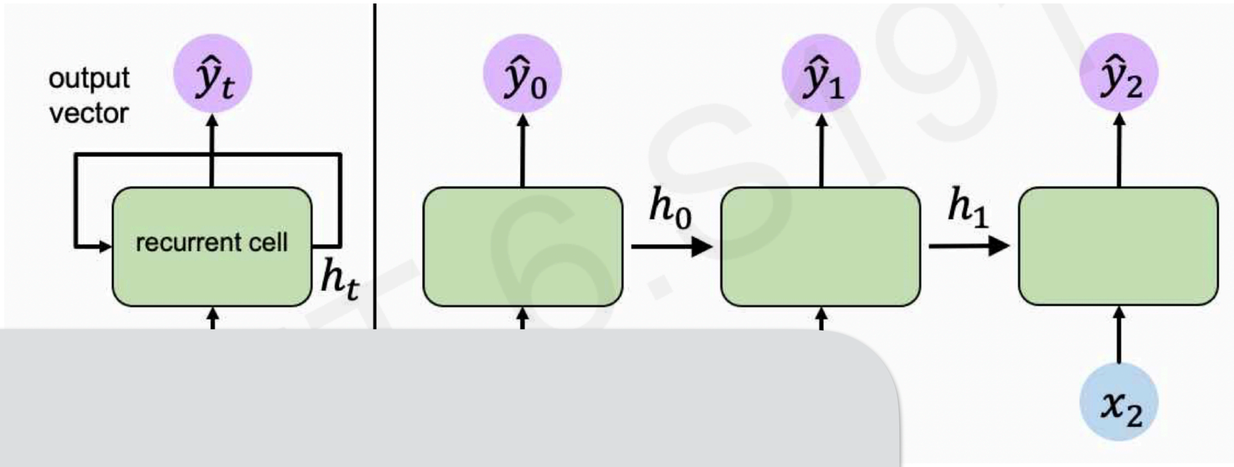
E. Zhiang, **EK**, et al, A3D3 High throughput workshop 2023

ML Inference with FPGA

Tranformers



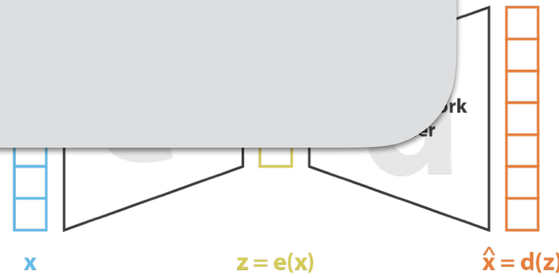
Recurrent Neural Networks



al, Mach. Learn.: Sci. Tec. 4 025004

More on these in my Seminar tomorrow

l Autoencoders



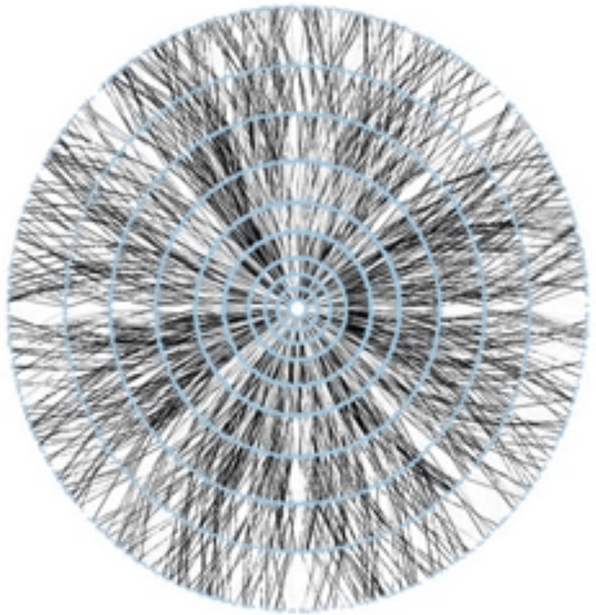
$$\text{loss} = \|x - \hat{x}\|^2 = \|x - d(z)\|^2 = \|x - d(e(x))\|^2$$

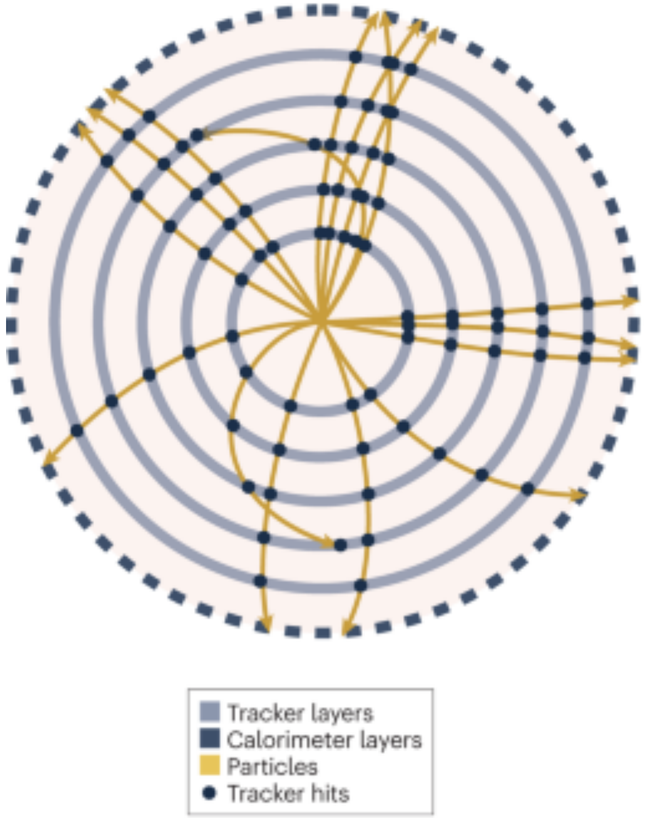
EK et al, FastML@ICCAD 2023
arXiv 2402.04274

EK et al, NeurIPS 2023, arXiv 2402.01047

Event Filter

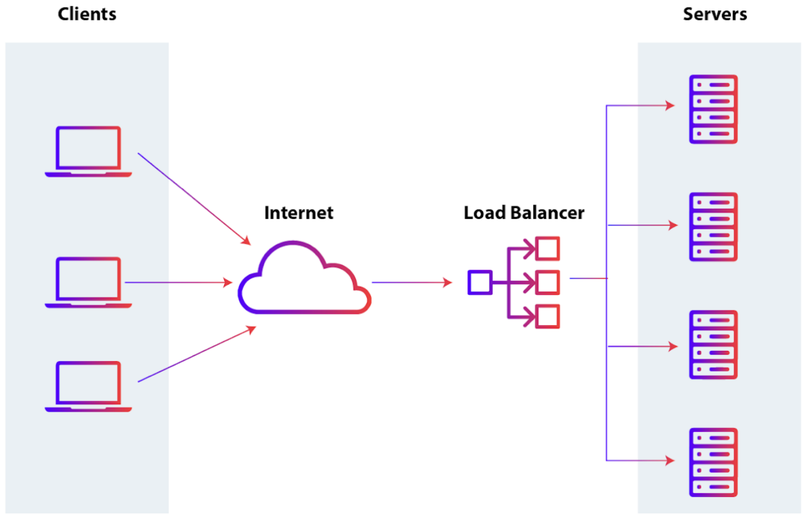
On-going efforts of putting these algorithms into GPU and FPGAs





As a Service

No need to have a local GPU



H. Zhao, X. Ju, .. **EK**, et al, ACAT 2023

FAIR Universe ML Challenge

Uncertainties play a crucial role in particle physics
How to train an uncertainty-aware ML model?

Active field of research
- Need a large dataset
- Novel metric

Creating a ML challenge to drive the research in this direction

Pilot Challenge in ACAT 2024 (currently ongoing)

Website: https://fair-universe.lbl.gov/

R. Chakkappai, W Bhimji, .. **EK**, et al
AI and Uncertainty Workshop 2023
ACAT 2024

Broader Impact beyond HEP

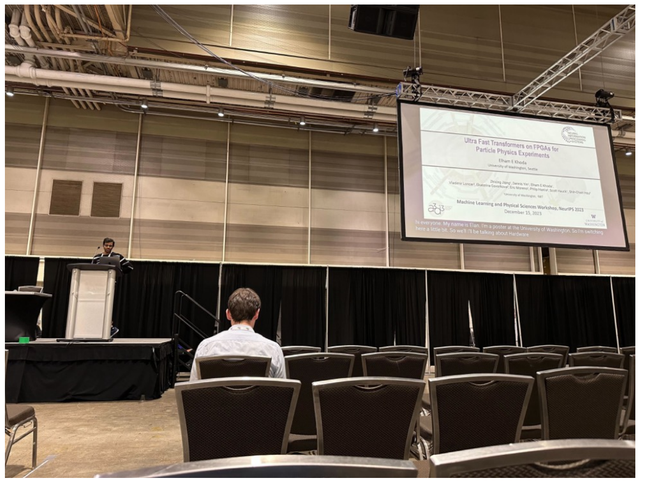
ML and Data Science Training

- Snowmass 2022 tutorial on real-time AI
- Created the US_ATLAS ML program in 2022
- Snowmass White papers on Future e+e- and muon collider sensitivity

Mentorship and Equity

- Part of the Postbacc selection committee
- Member of A3D3 Equity and Career committee

Interdisciplinary Research

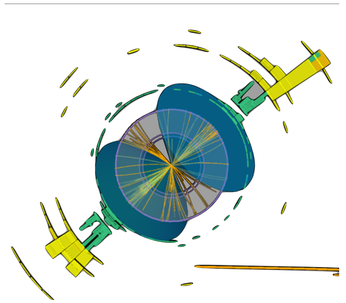


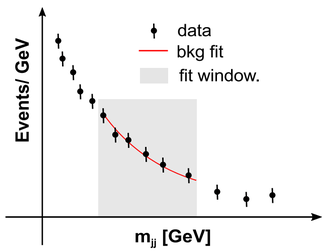
NeurIPS 2023 spotlight talk on Transformers on FPGA

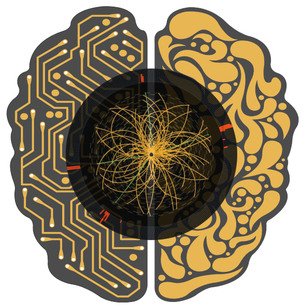
- Fast inference of Neural population dynamics
- Fast ML co-processor coordinator

Summary

Offline







Online

New Physics Search: Model-dependent

- ➡ Most sensitive way to search for theory-driven BSM physics
- ➡ Optimized the DNN-based top-tagging to improve sensitivity by 2x
- ➡ No sign of new physics yet!

Anomaly Detection

- ➡ Sensitive to several types of BSM physics
- ➡ Developed a Generative (VAE & GAN) algorithm-based method for anomaly detection
- ➡ Bring my method in ATLAS

Fast Machine Learning Inference at the trigger

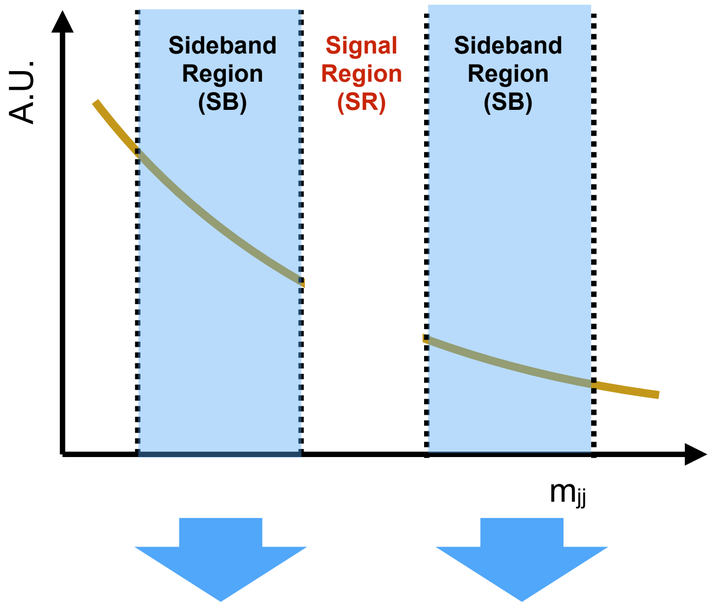
- ➡ Demonstrated ability to deploy ML algorithms on FPGA
- ➡ Enable new physics via anomaly detection trigger
- ➡ Success of future physics program at LHC & HL-LHC will require innovation throughout

Many thanks to my Collaborators!

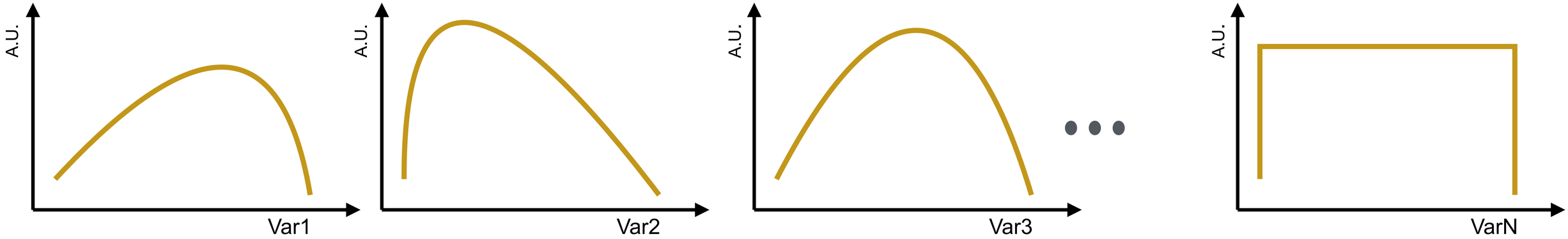
Thank You!

Extra Slides

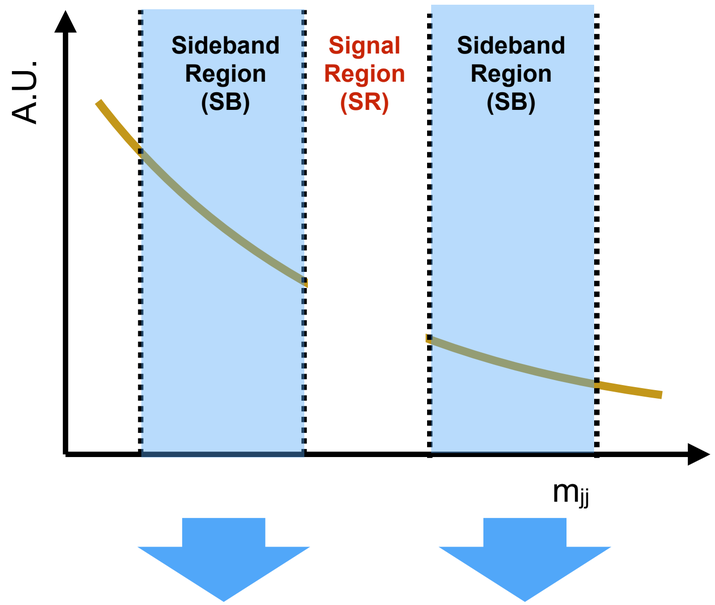
Predict the background in the signal region



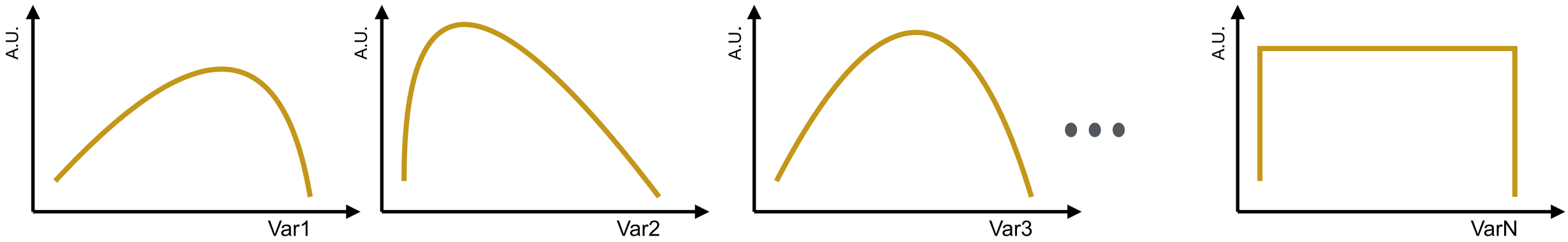
Model the multiple observables in the sideband regions



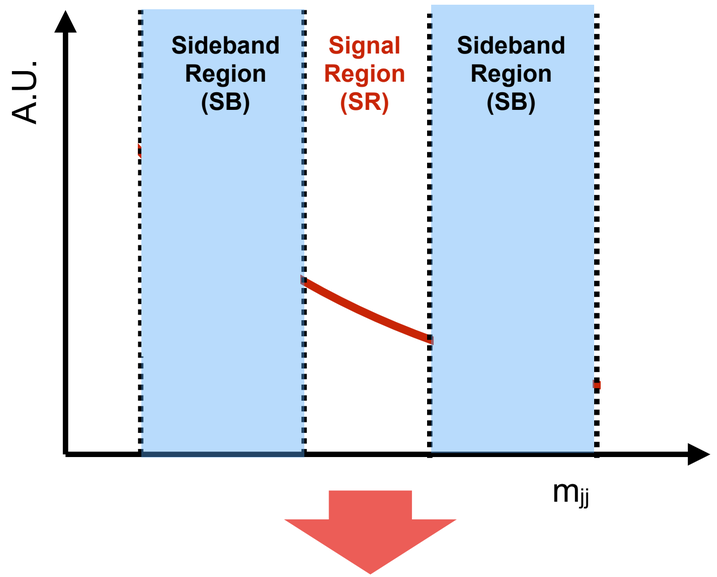
My Strategy to estimate the density



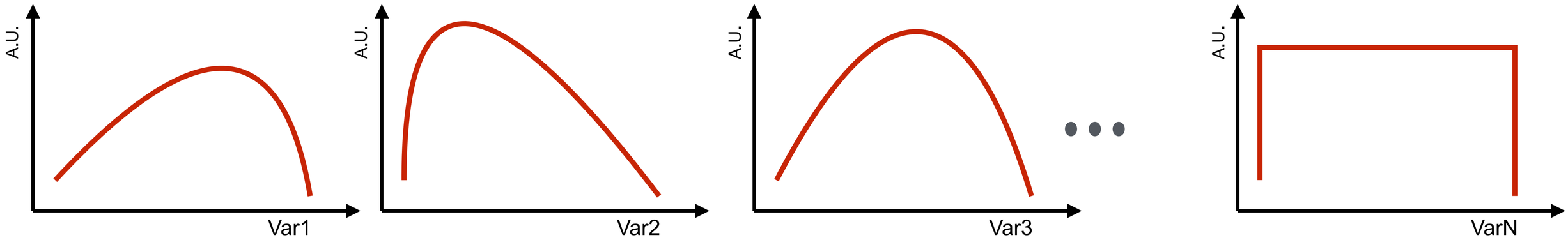
Generative models (Generative Adversarial Network or Variational Autoencoder) to estimate the densities



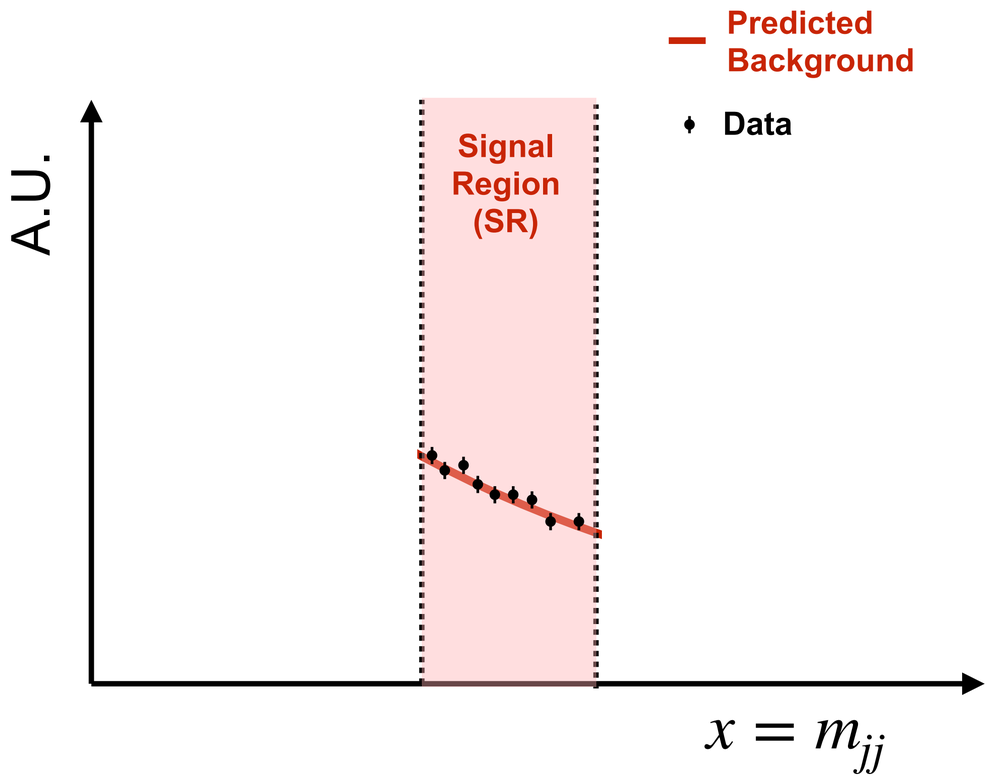
Model background in the side-band



Predict them in the signal region

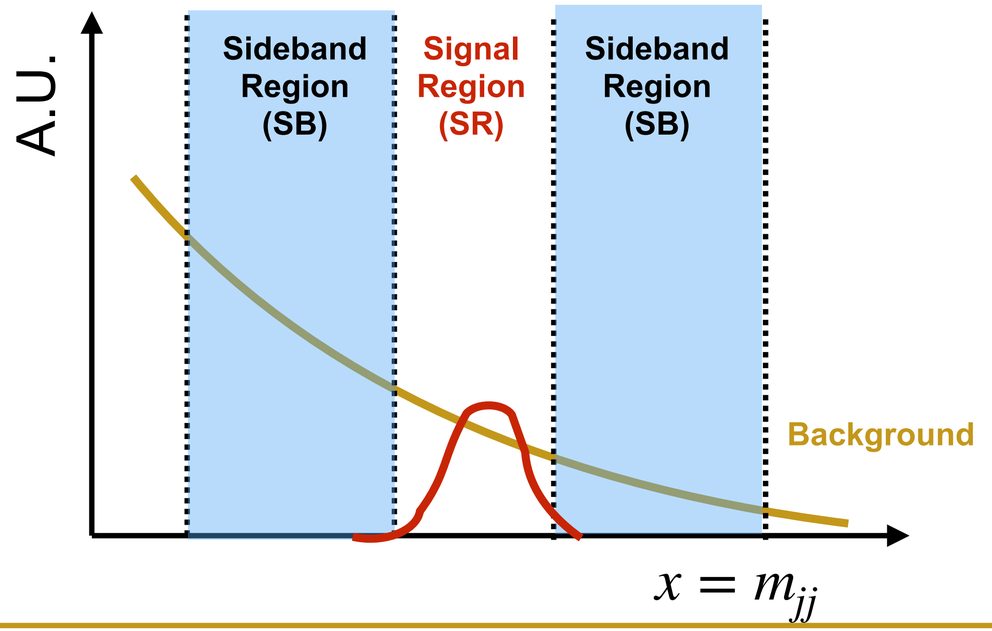


Compare Data and Prediction

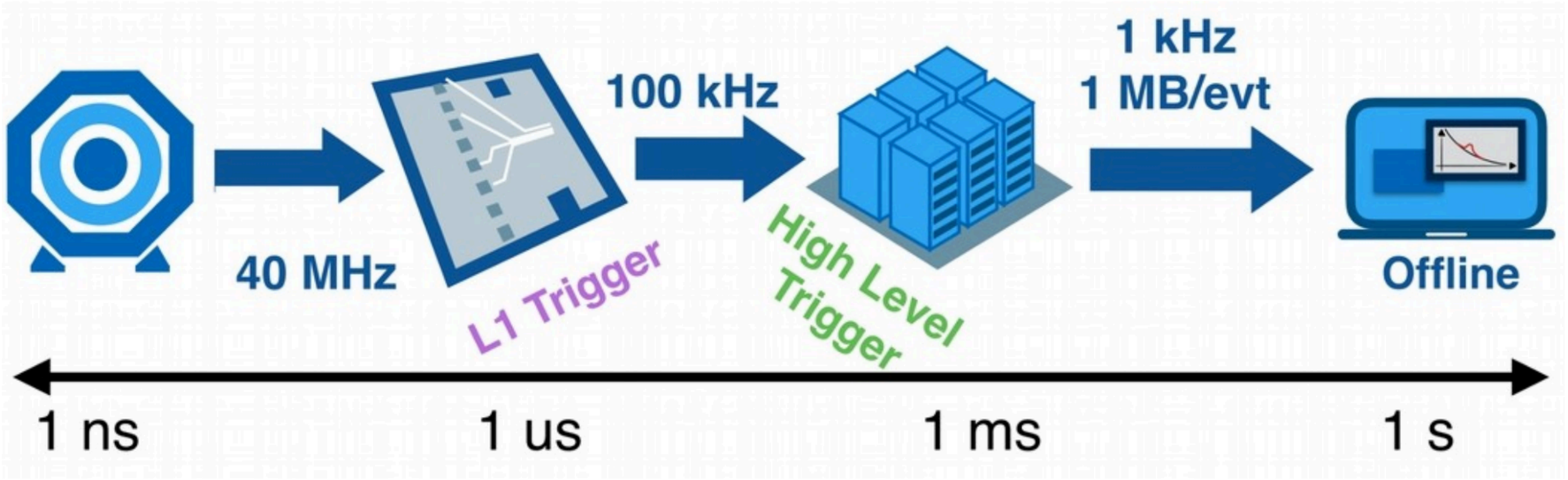


Compare

Example:
ML classifier trained to differentiate signal-rich data vs background



ATLAS Run-3 Data Processing



L1 Trigger (hardware: FPGAs) – *O(μs) hard latency*

- Typically coarse selections are applied

High Level Trigger (software: CPUs) – *O(100 ms) soft latency*

- More complex algorithms (full detector information available), some BDTs and DNNs used

Offline (software: CPUs)

- Full event reconstruction, bulk of machine learning usage in ATLAS/CMS

ML in the Event Filter

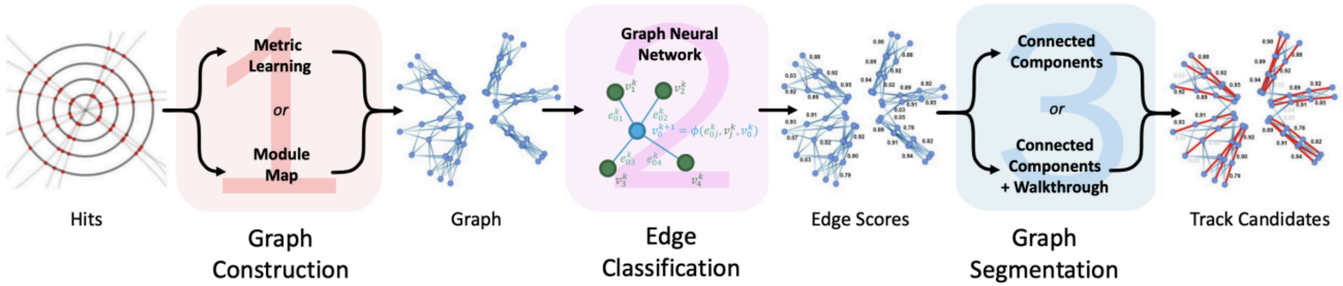
Event Filter:

- Future high-level software trigged for ATLAS
- Computing infrastructure could include GPUs and FPGAs along with CPUs

Promising Graph NN-based track finding algorithm

- Good performance with HL-LHC simulations
- Can be accelerated on different coprocessors

On-going efforts of putting these algorithms into GPU and FPGAs

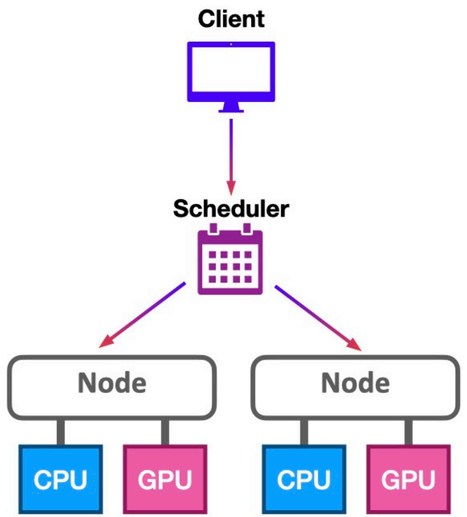


Heterogeneous Computing Model

- Support for different type of hardware is becoming important

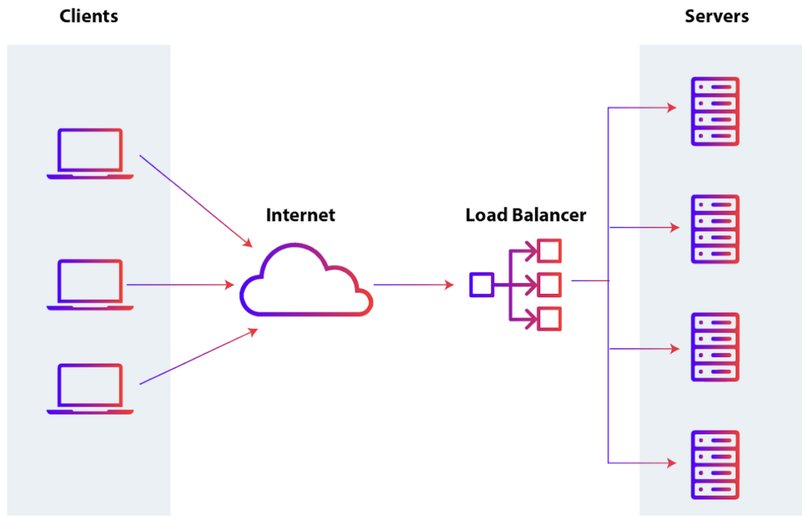
Direct Connection

CPUs and GPUs are connected



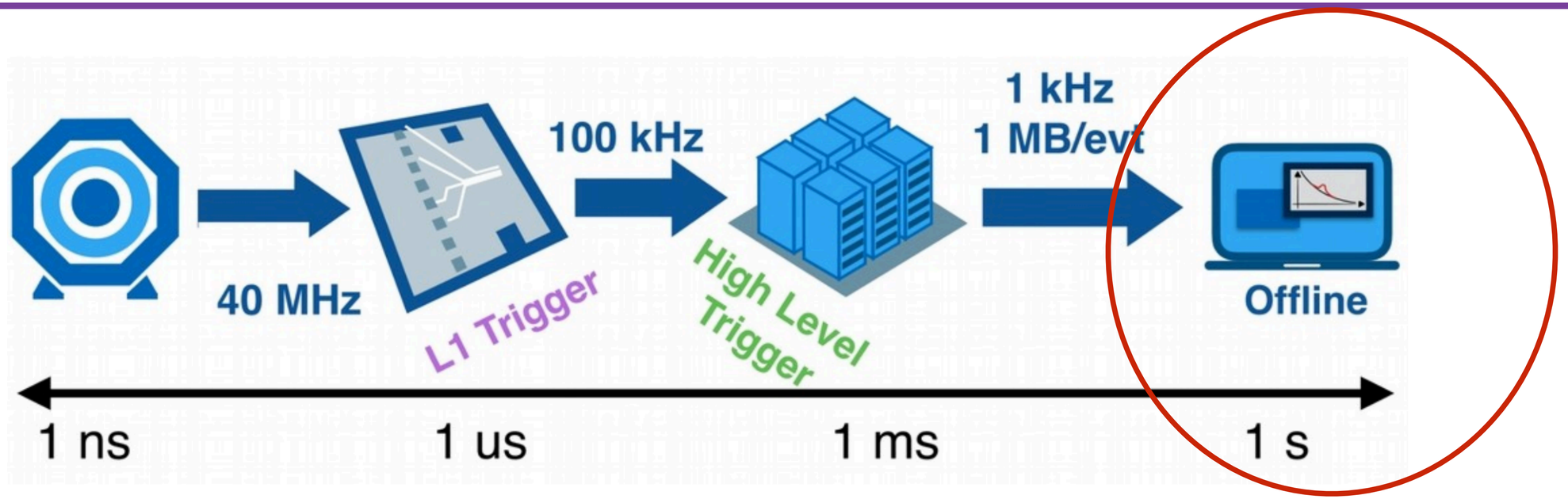
As a Service

No need to have a local GPU



- Simple support for mixed hardware
- Scaleable
- Throughput optimization for multiple-core

ATLAS Run-3 Data Processing

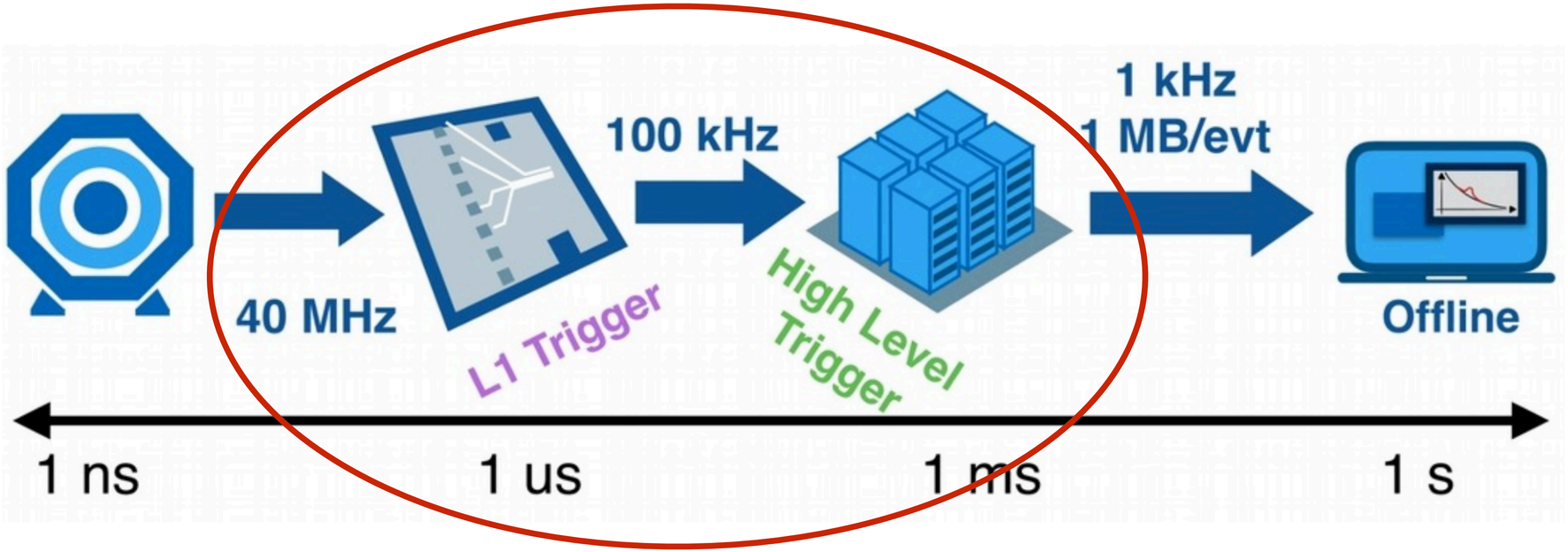


- All the physics searches and measurements

- Active R&D
  - Further improvements driven by more complicated algorithms
  - Usage of ML is growing over time

Usage of GPUs will be beneficial for the future LHC Runs (after 2026)

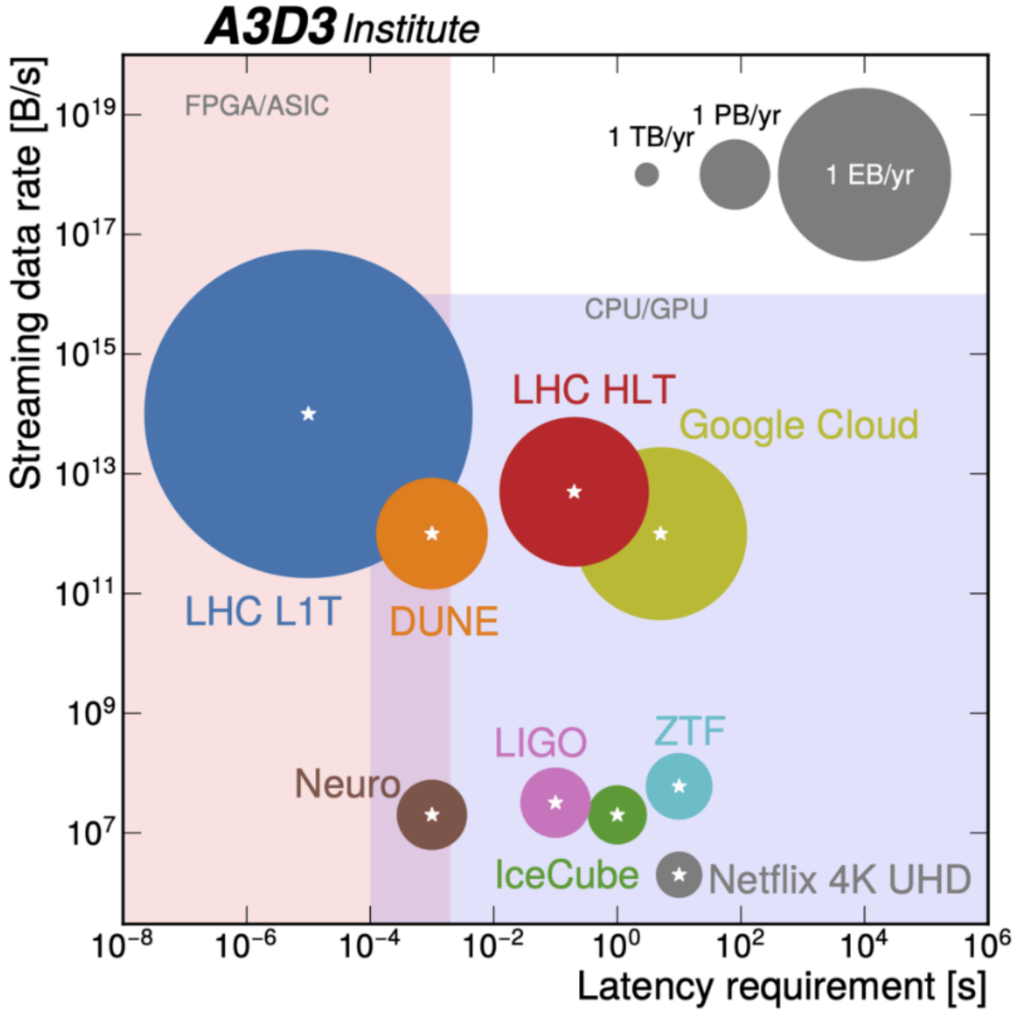
ATLAS Run-3 Data Processing



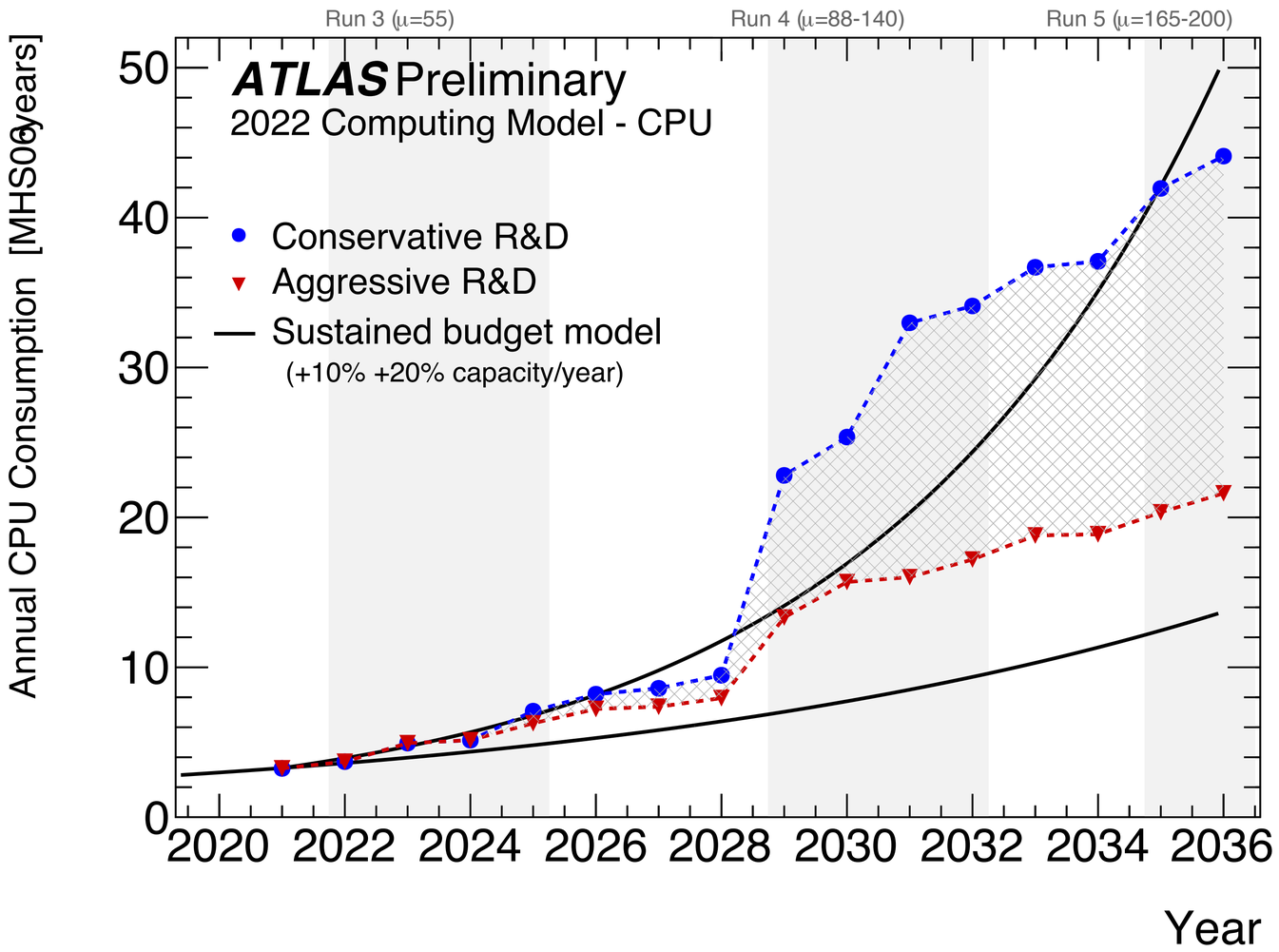
- ML has potential to improve physics performance in the trigger system
- **Strict latency requirements:** μs (ms) for **Level-0** (**Event Filter**)

  For **Level-0** trigger → we need to run ML on FPGAs

Low-latency High Throughput applications



Critical Computing Challenges

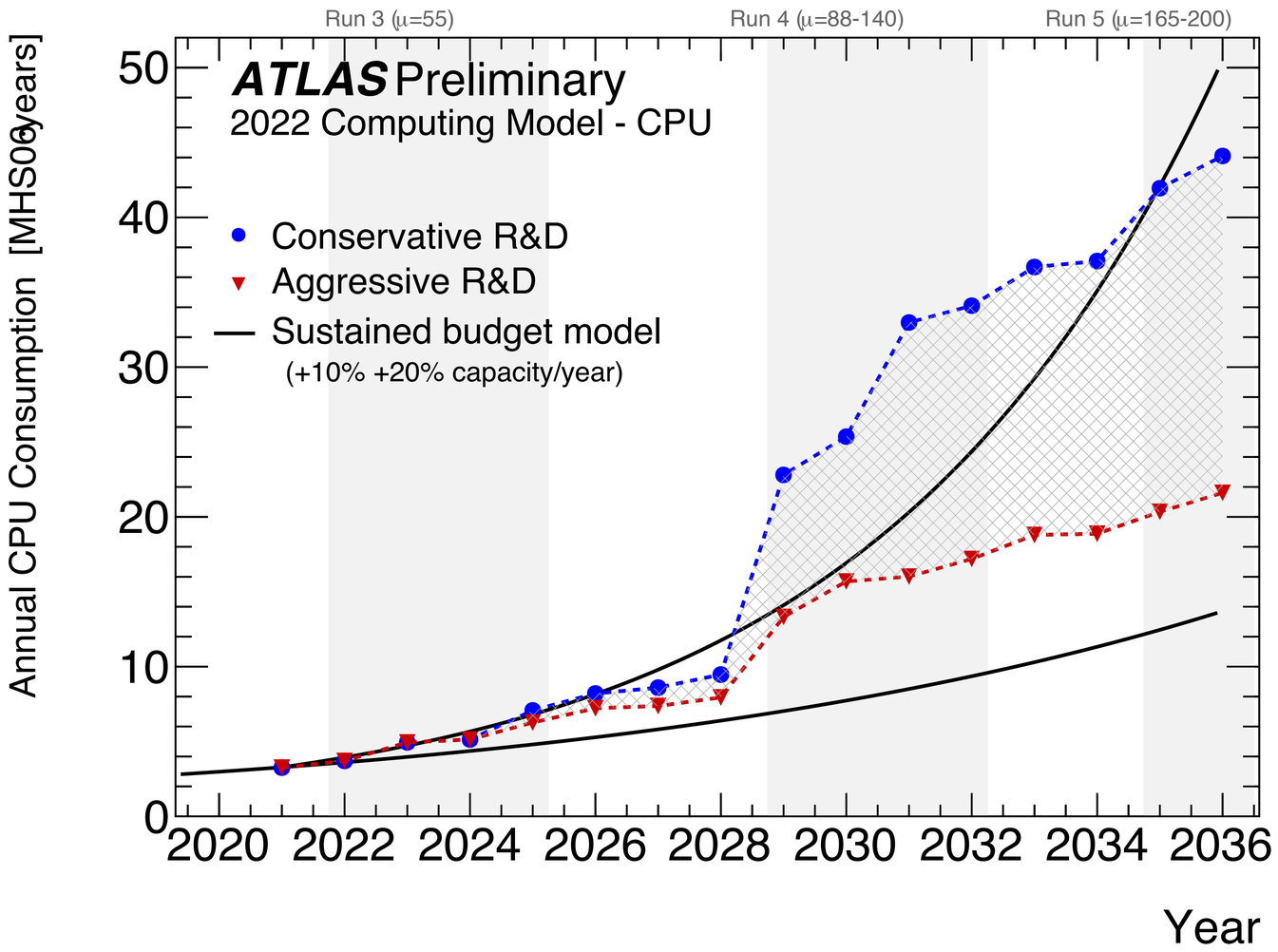


To preserve current physics:

- 4 times the current data taking rate

Lacking sufficient budget to sustain required computing

Critical Computing Challenges



To preserve current physics:

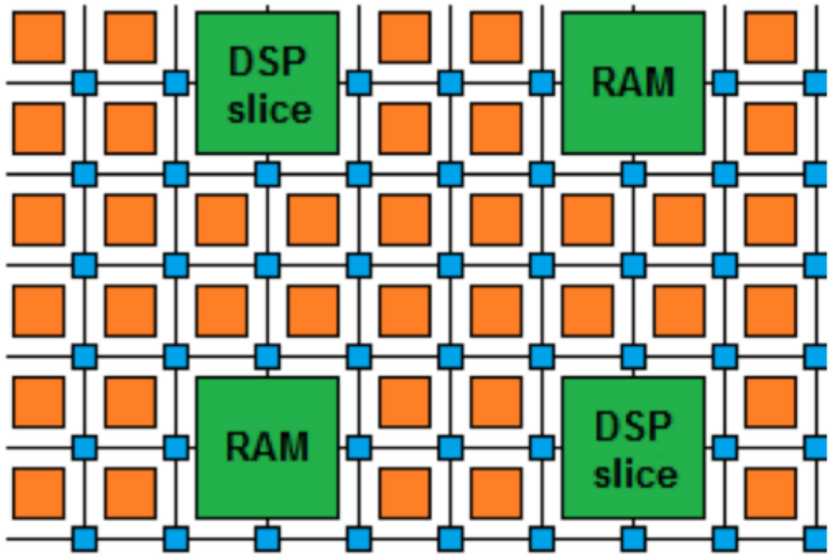
- 4 times the current data taking rate

Lacking sufficient budget to sustain required computing

What is an FPGA?

Field **P**rogrammable **G**ate **A**rrays (FPGAs) are reprogrammable integrated circuits

- Contain many different building blocks (‘resources’) which are connected together as you desire
- Originally popular for prototyping ASICs, but now also for high performance computing



Building blocks:

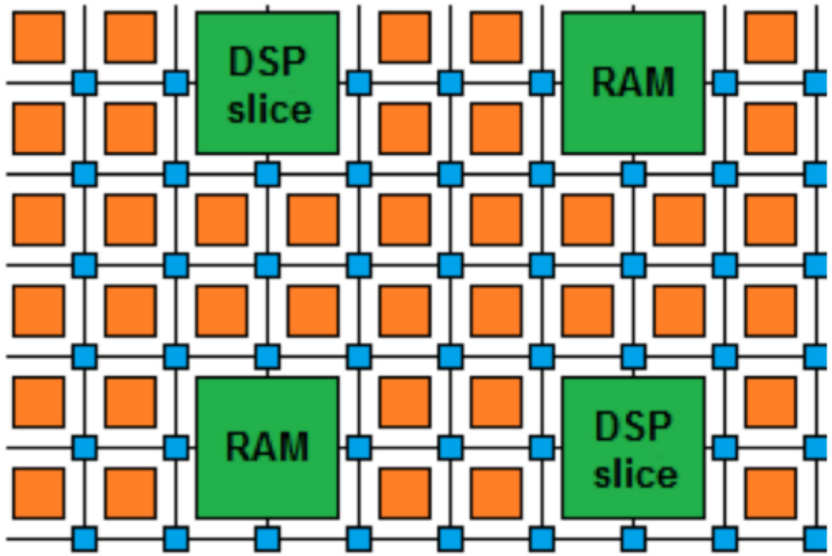
- **Multiplier units (DSPs)** [arithmetic]
- **Look Up Tables (LUTs)** [logic]
- **Flip-flops (FFs)** [registers]
- **Block RAMs (BRAMs)** [memory]

What is an FPGA?

- Run at high frequency - *O(100 MHz)*
  - Can compute outputs in O(ns)

- Low-level Hardware Description Language for programming
  Verilog/VHDL

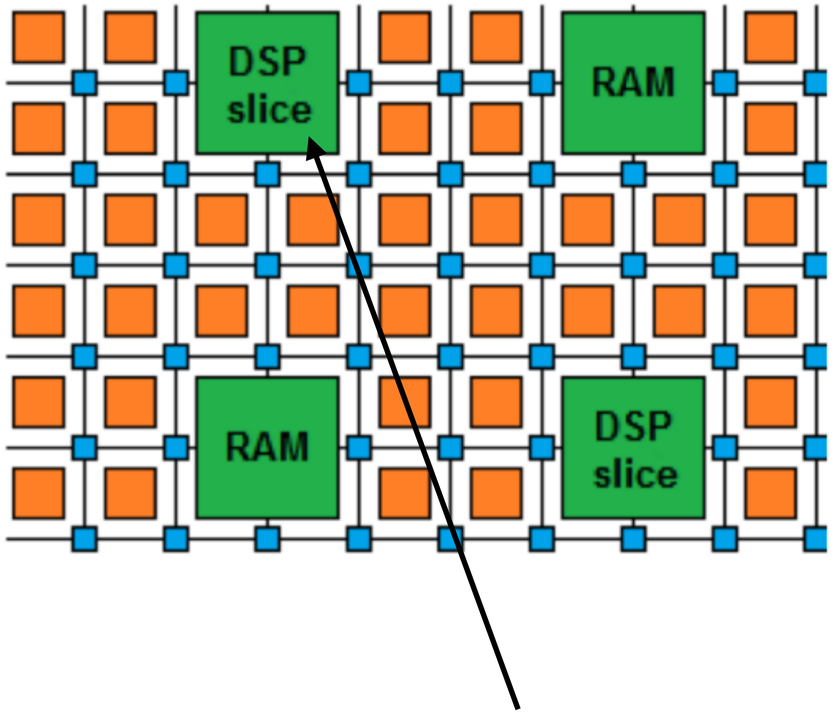
- Possible to translate C/C++ → Verilog/VHDL using High **Level Synthesis (HLS)** tools



Building blocks:
- **Multiplier units (DSPs)** [arithmetic]
- **Look Up Tables (LUTs)** [logic]
- **Flip-flops (FFs)** [registers]
- **Block RAMs (BRAMs)** [memory]

What is an FPGA?

- **DSPs** (Digital Signal Processor) are specialized units for multiplication and arithmetic
- DSPs are often the most scarce for NNs
- Faster and more efficient than using LUTs for these types of operations

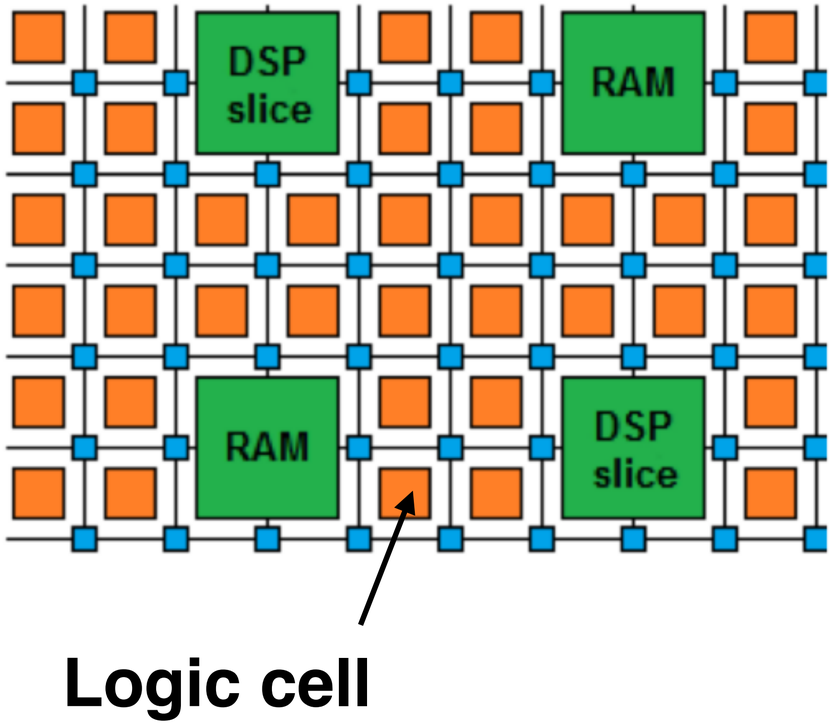


Building blocks:
- **Multiplier units (DSPs)** [arithmetic]
- **Look Up Tables (LUTs)** [logic]
- **Flip-flops (FFs)** [registers]
- **Block RAMs (BRAMs)** [memory]

What is an FPGA?

- **Logic cells / Look Up Tables** perform arbitrary functional operations on small bit-width inputs (2-6)
  - boolean, arithmetic
  - small memories

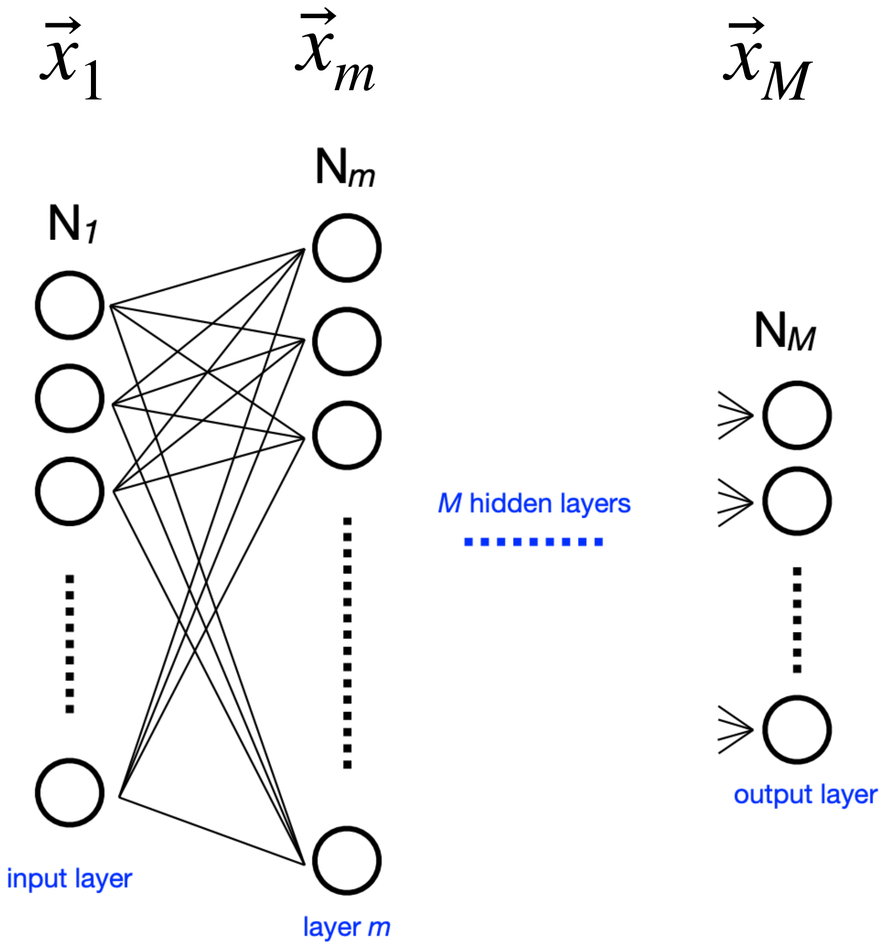
- **Flip-Flops** register data in time with the clock pulse



Building blocks:

– **Multiplier units (DSPs)** [arithmetic]
– **Look Up Tables (LUTs)** [logic]
– **Flip-flops (FFs)** [registers]
– **Block RAMs (BRAMs)** [memory]

Inference on an FPGA



$$\vec{x}_m = g_m\left(W_{m,m-1}\vec{x}_{m-1} + \vec{b}_m\right)$$

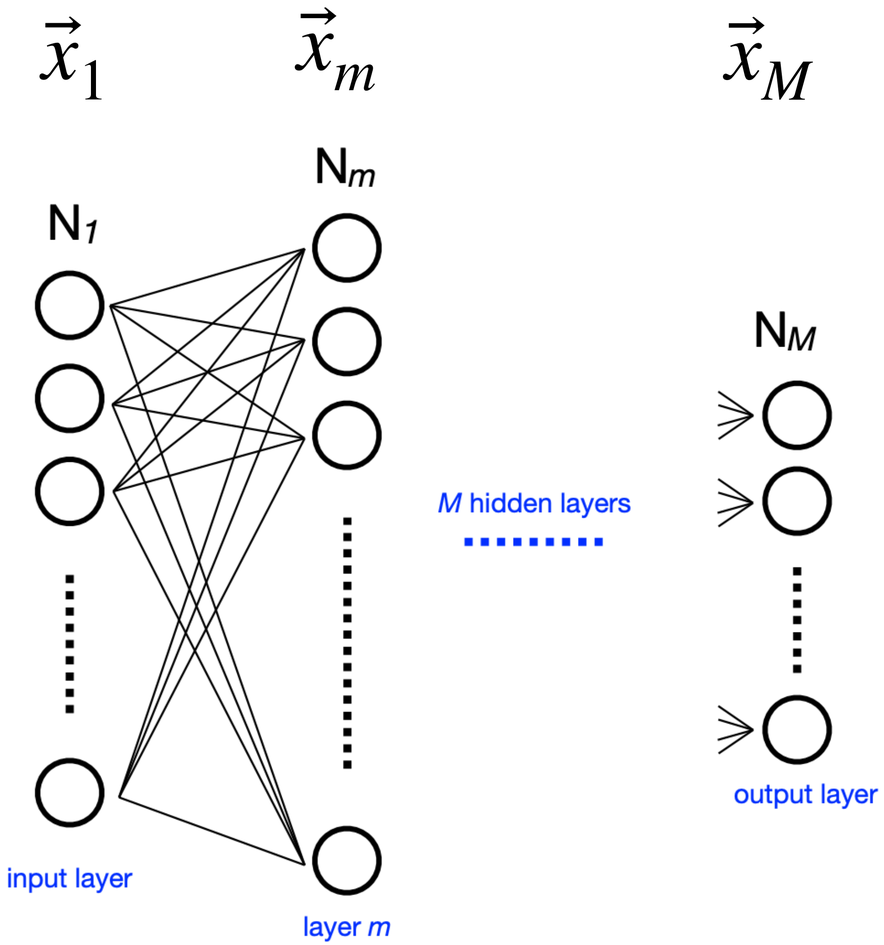
multiplication

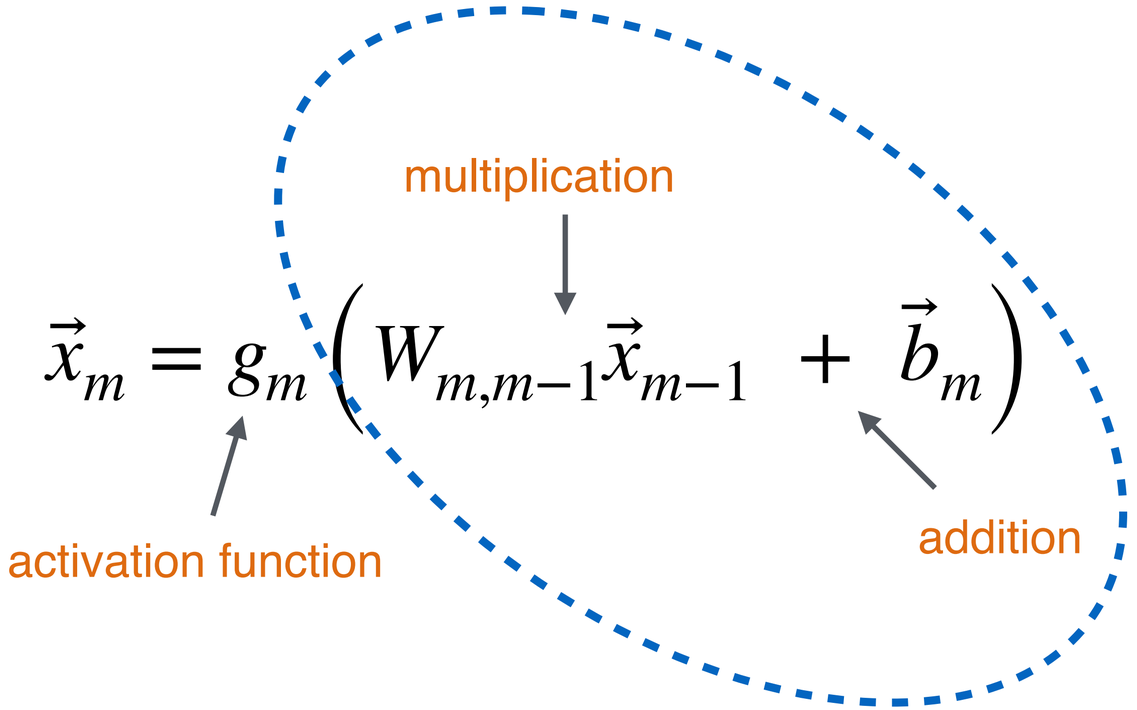
activation function

addition

Credit: Dylan Rankin

Inference on an FPGA



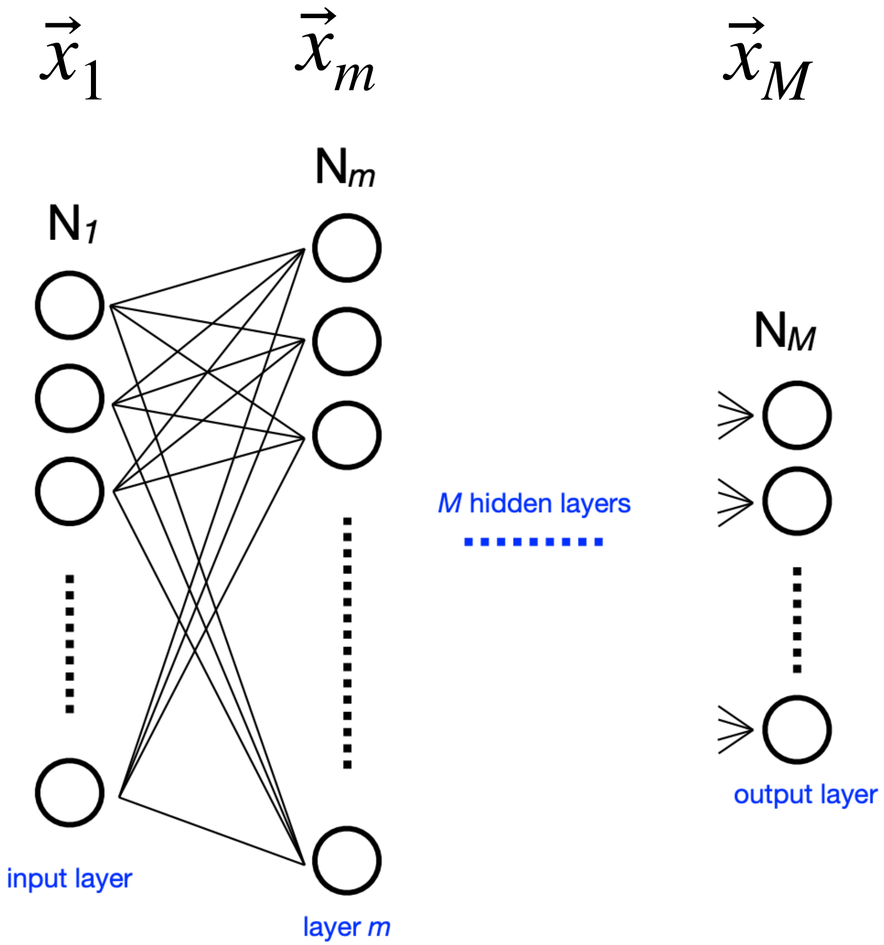


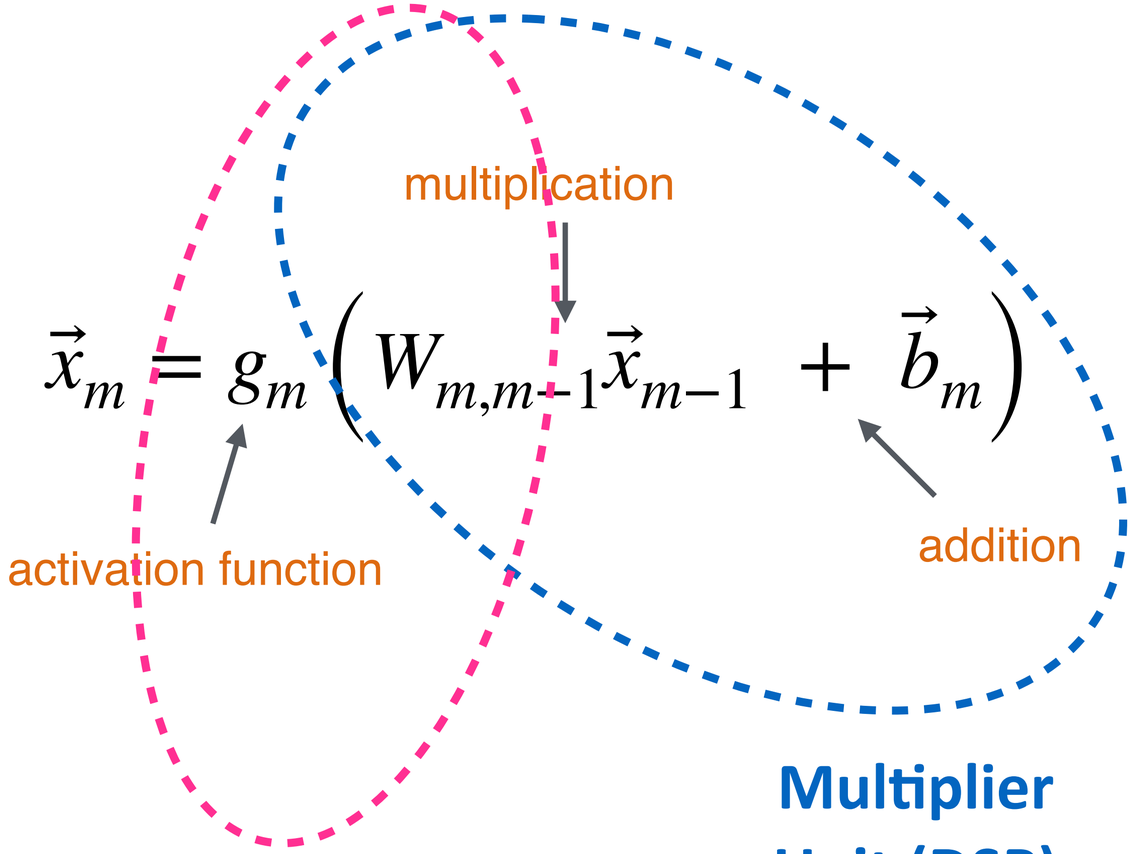
Multiplier Unit (DSP)

up to ~6k parallel operation (VU9P)

Credit: Dylan Rankin

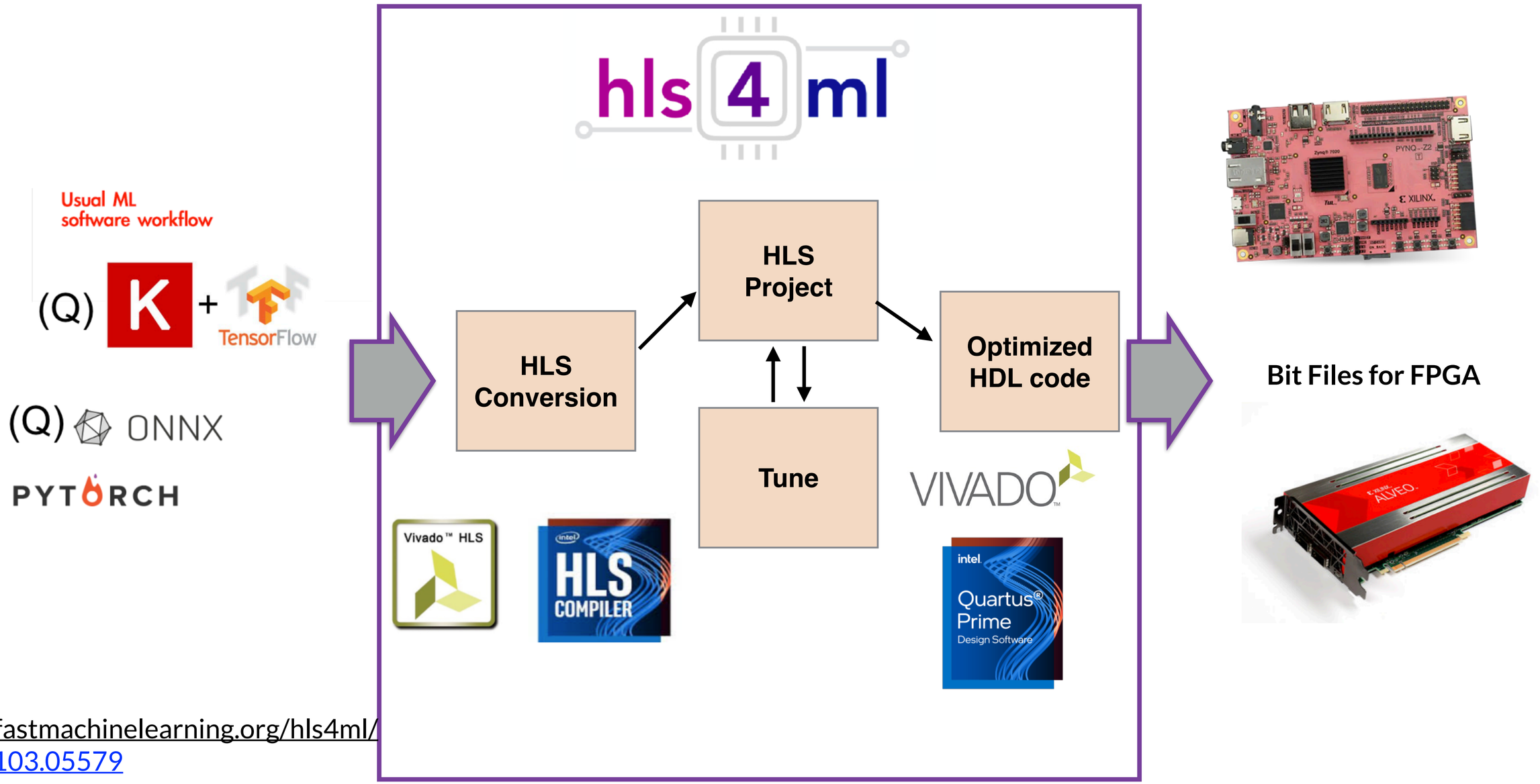
Inference on an FPGA





Credit: Dylan Rankin

High Level Synthesis for Machine Learning (hls4ml)



https://fastmachinelearning.org/hls4ml/
arXiv:2103.05579

High Level Synthesis with Machine Learning (hls4ml)

https://fastmachinelearning.org/hls4ml/
arXiv:2103.05579

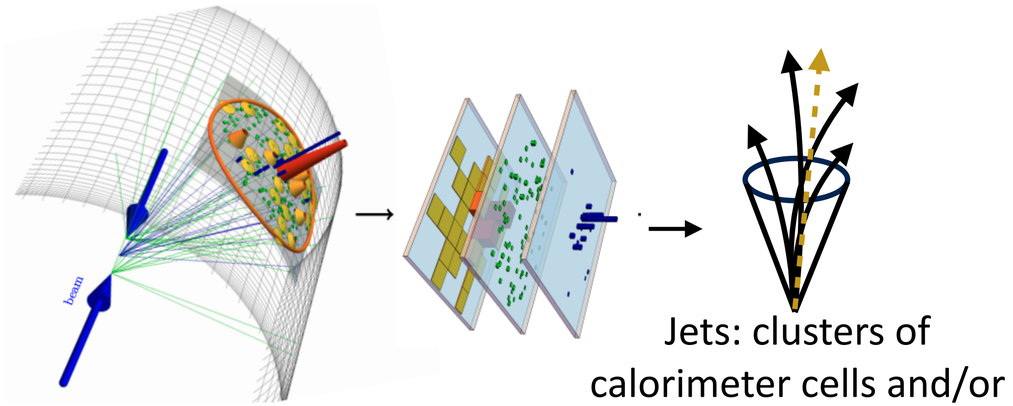
A software interface for implementing Neural Networks on an FPAG

- Supports many common layer like DNN, CNN, etc
- Recursive Neural Networks were not implemented until late 2022

RNN-based algorithms could be deployed at the Level-0 trigger
Example:
- Tau-particle identification
- Missing Transverse Energy reconstruction

Particle Jet Classification at the LHC

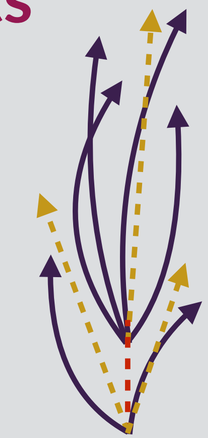
Jets are experimental signature of quarks and gluons



Jets: clusters of calorimeter cells and/or tracks

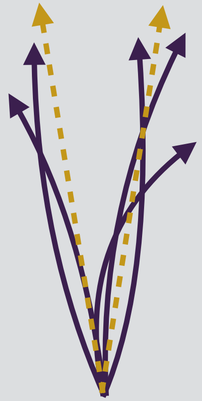
b or c jets

- Jets coming from b- or c-quark
- Displaces vertex



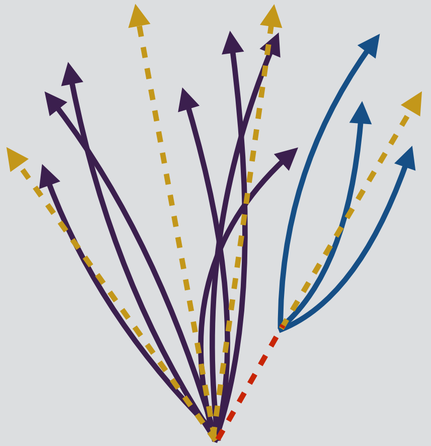
light-jets

- Jets coming from u/d/s-quarks



top jet

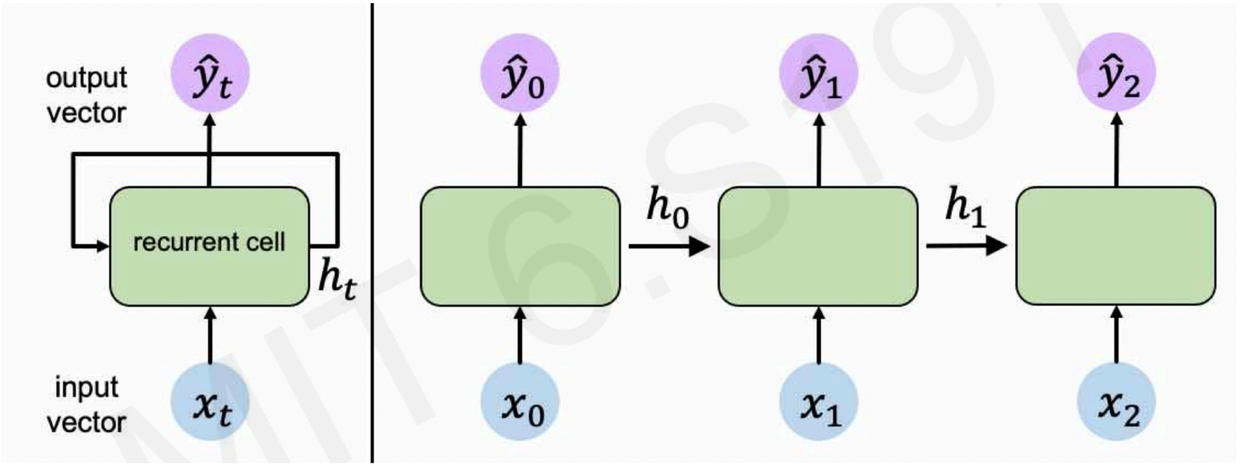
- Jets coming from top-quark
- 3-prong structure



Recurrent Neural Network (RNN)

Recurrent Neural Networks

- Designed to work with sequential data
  - Text, audio, video, strokes, etc
- RNNs have a state, h_t, that is updated at each time step as the sequence is processed
- Recurrence relation at every time step



$$\hat{y} = f(x_t, h_{t-1})$$

Output Input past memory

$$h_t = f_W(x_t, h_{t-1})$$

cell state

Function with weights W

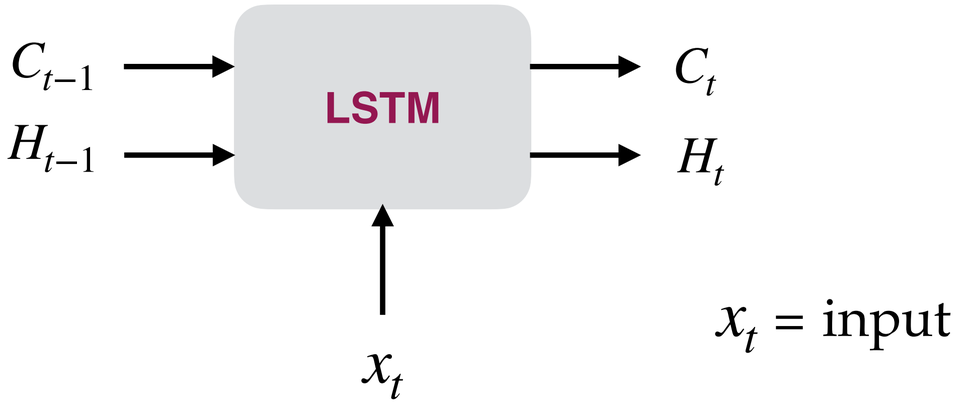
Input

old state

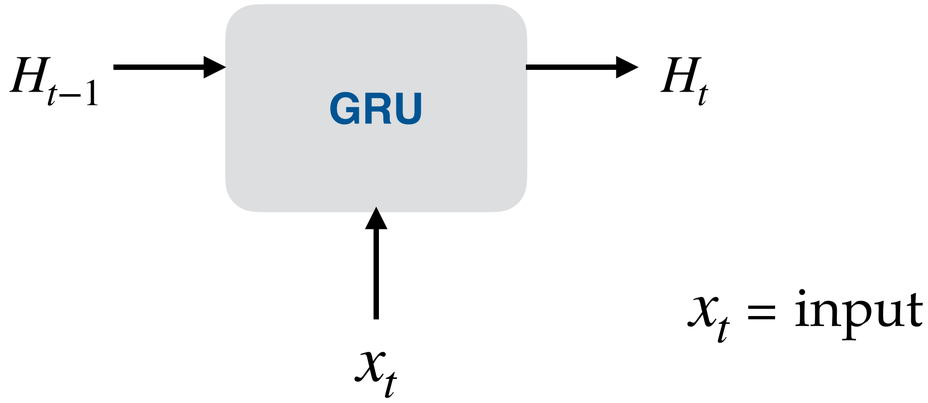
Implementation of RNN models:

- LSTM (Long Short-Term Memory)
- GRU (Gated Recurrent Unit)

LSTM vs GRU



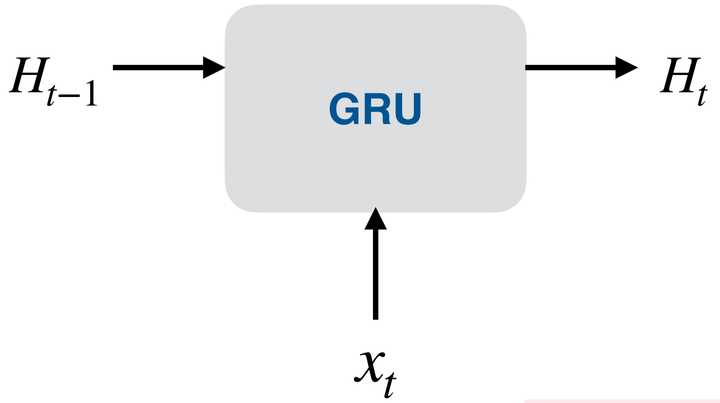
- **3 gates:** Input, Output, Forget
- **2 States:** Cell state (C_t) and Hidden state (H_t)

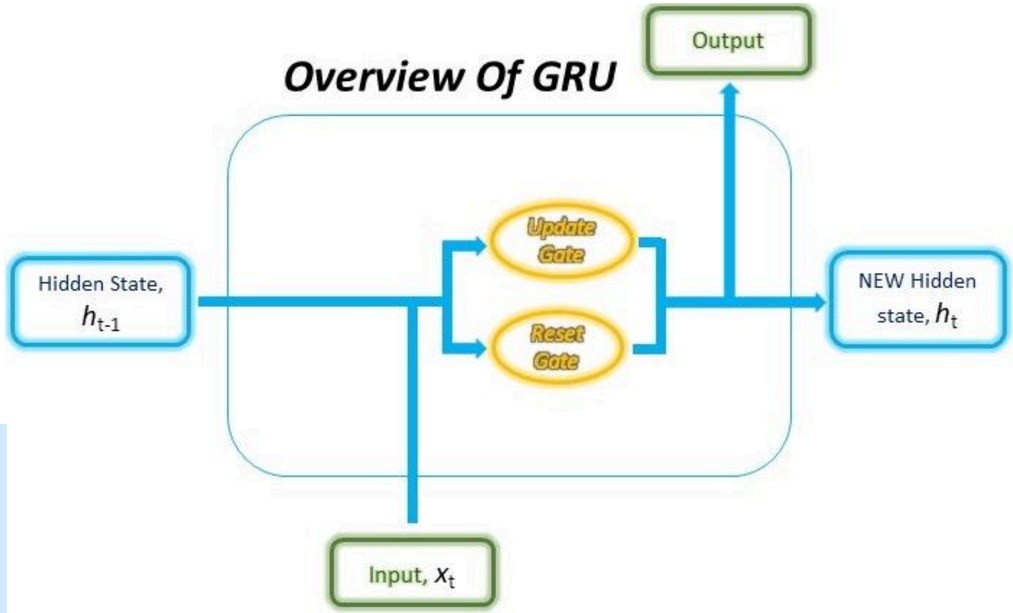


- **2 gates:** Update and Reset
- **Single** Hidden state (H_t)

- Less number of matrix multiplications
- Faster to train

Gated Recurrent Unit (GRU)





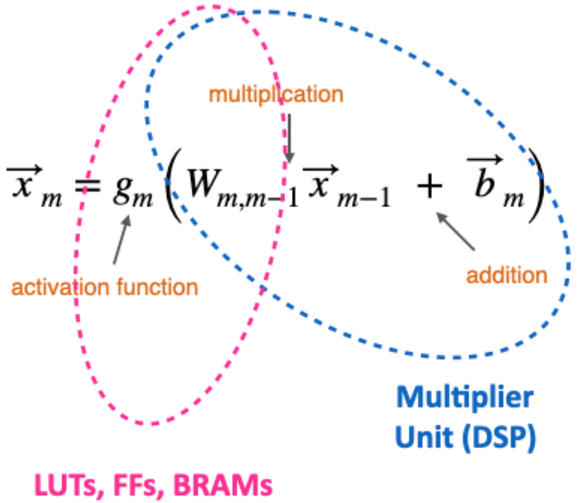
Dense **Dense**

Reset: $r_t = \sigma\ (W_{xr} \cdot x_t + b_r + W_{hr} \cdot h_{t-1})$

Update: $u_t = \sigma\ (W_{xu} \cdot x_t + b_u + W_{hu} \cdot h_{t-1})$

Candidate hidden state: $\tilde{h}_t = \tanh\ (W_{xh} \cdot x_t + b_h + (r_t \odot h_{t-1}) \cdot W_{hh})$

Output hidden state: $h_t = u_t \odot h_{t-1} + (1 - u_t) \cdot \tilde{h}_t$



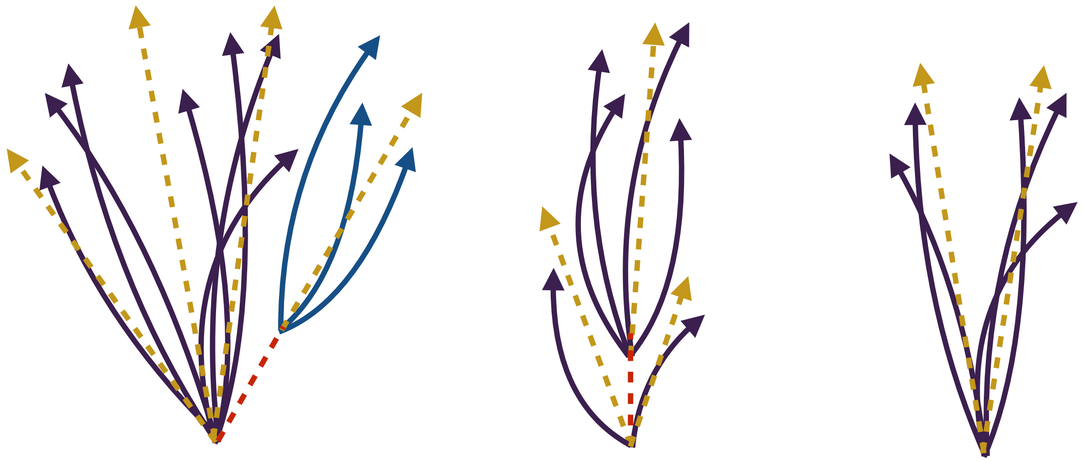
Benchmark Examples

Three benchmark cases

1. Binary classifier: ~4k parameters
   Identify top-quarks

2. 3-class classifier: ~60k parameters
   Classify b / c / light jets

3. 5-class classifier: ~130k parameters
   QuickDraw dataset: differentiate between Bees, Butterflies, Mosquitos, Snails, Ants

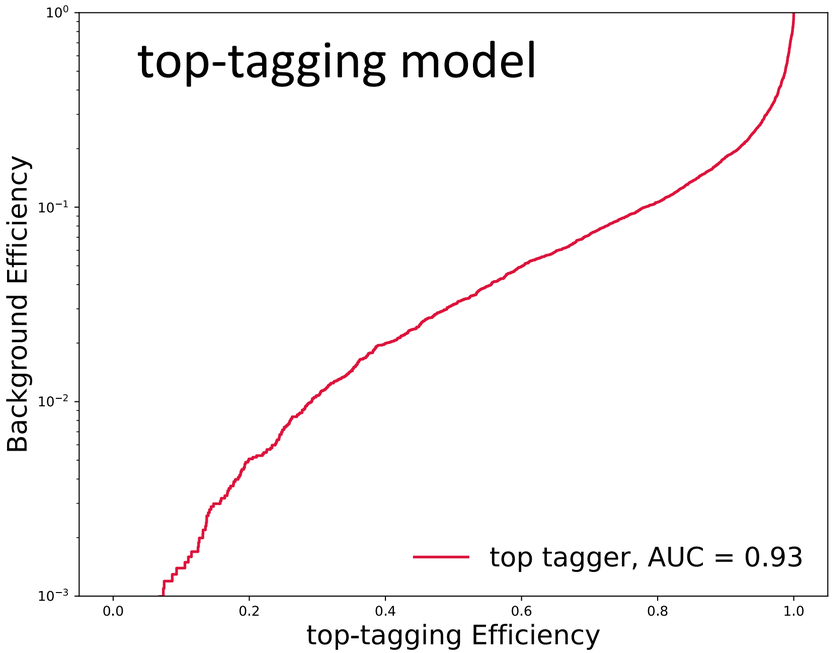


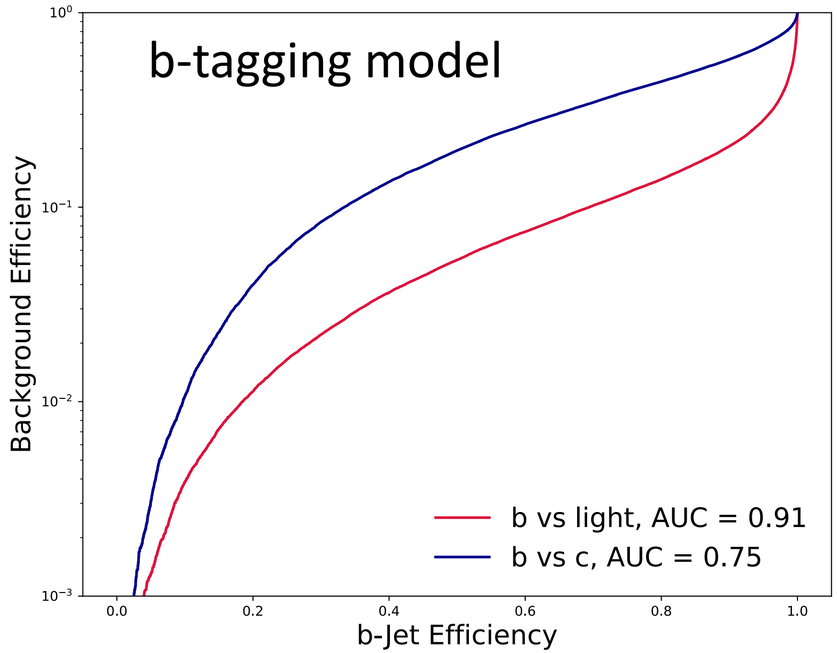
top b/c light

QuickDraw dataset

Model Performance: ROC

- All the benchmark models are trained using **Keras + TensorFlow**
- Weights and biases are represented by 32 bit floating point numbers





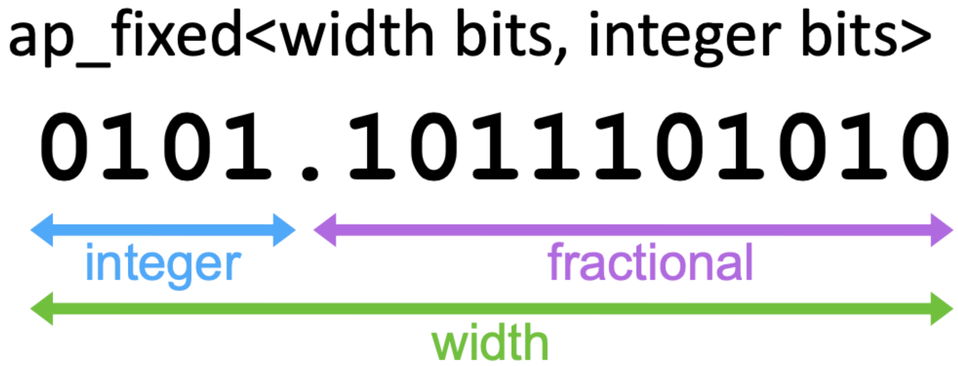
Quantization

Quantization – Reducing the bit precision used for NN arithmetic

Why this is necessary?

- Floating-point operations (32 bit numbers) on an FPGA consumes large resources
- Not necessary to do it for desired performance

- **hls4ml uses fixed-point representation for all computations**
  - Operations are integer ops, but we can represent fractional values

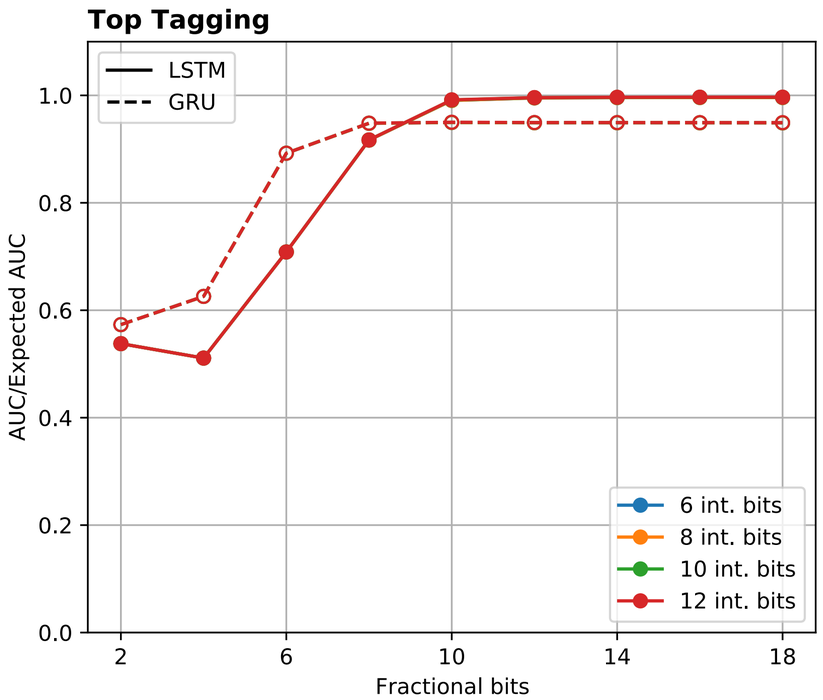


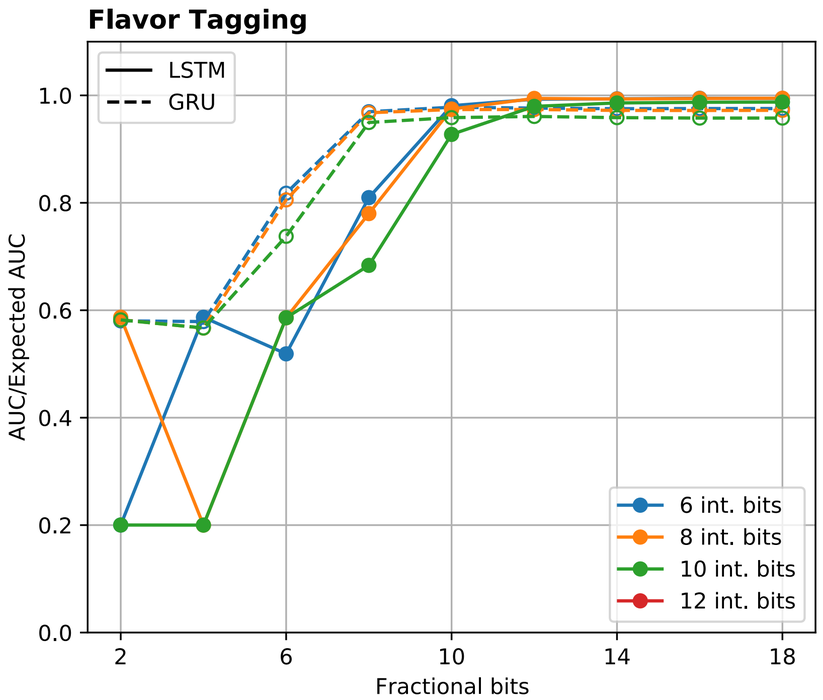
AUC after HLS conversion

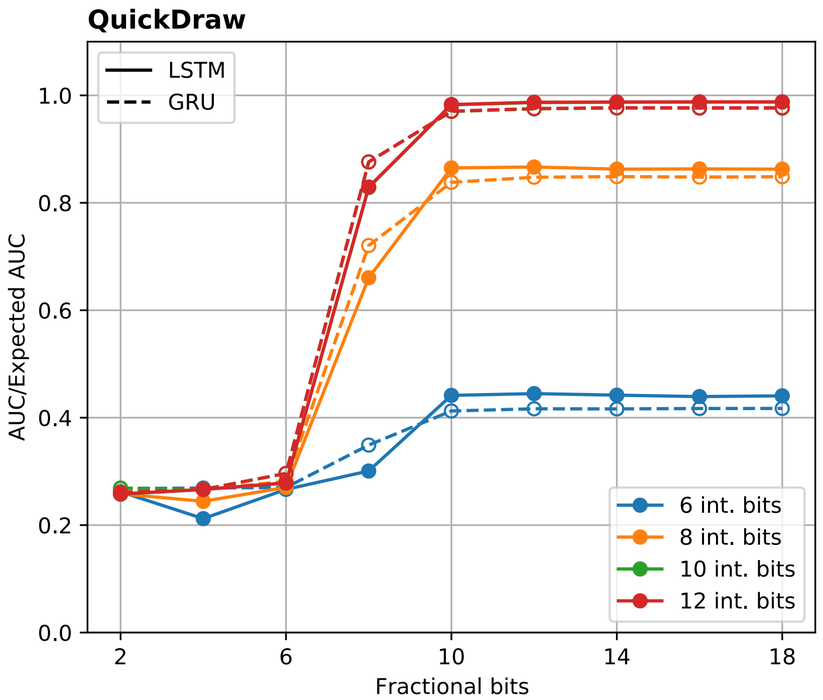
Mach. Learn.: Sci. Tec. 4 025004

$$\text{Relative AUC} = \frac{\text{AUC}_{\text{HLS}}}{\text{AUC}_{\text{Keras}}}$$

- Post-training quantized **LSTM models** (with optimal precision) performs similar to the floating-point models
- Small performance degradation (< 5%) in the **GRU models** after quantization

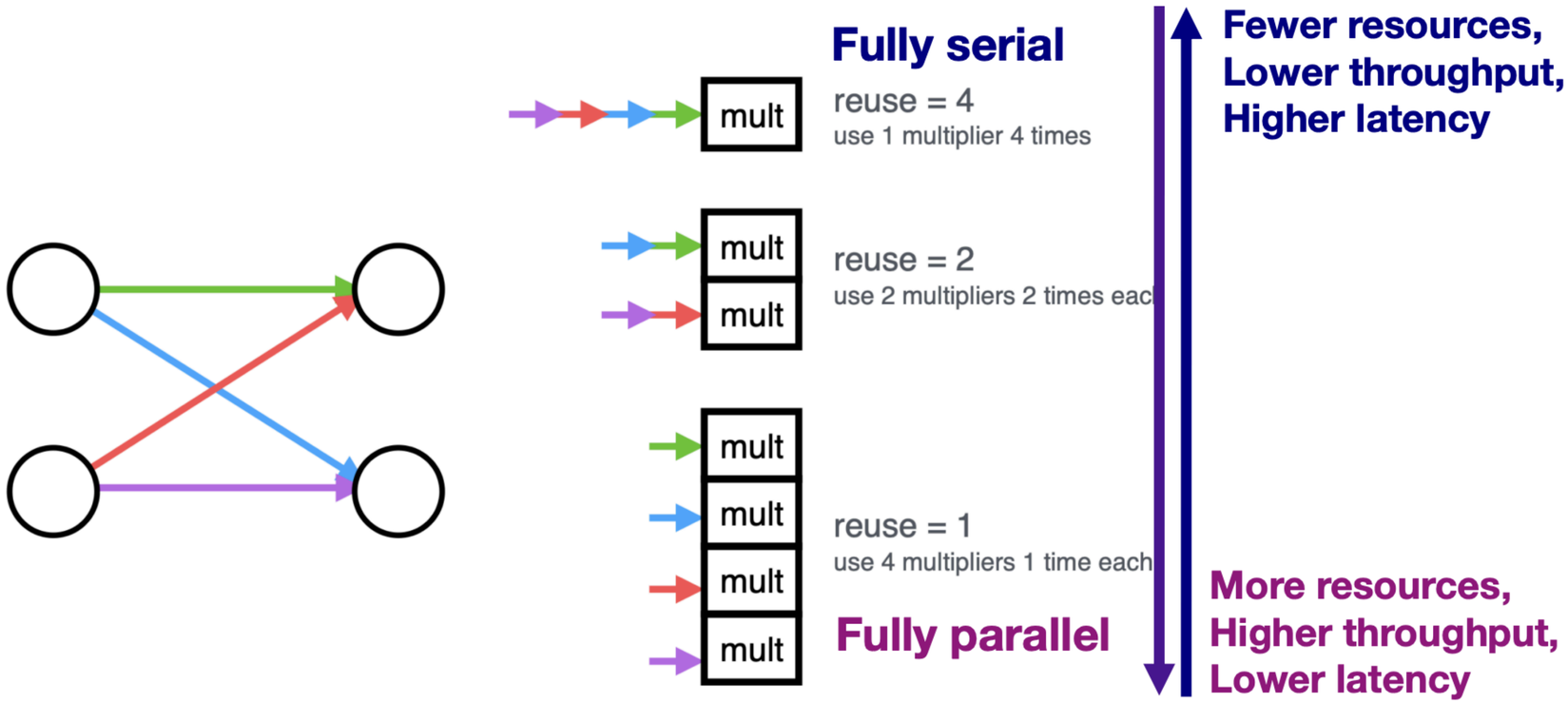






Parallelization

- Trade-off between latency and FPGA resource usage determined by the parallelization of the calculations in each layer
- Configure the "reuse factor" = number of times a multiplier is used to do a computation

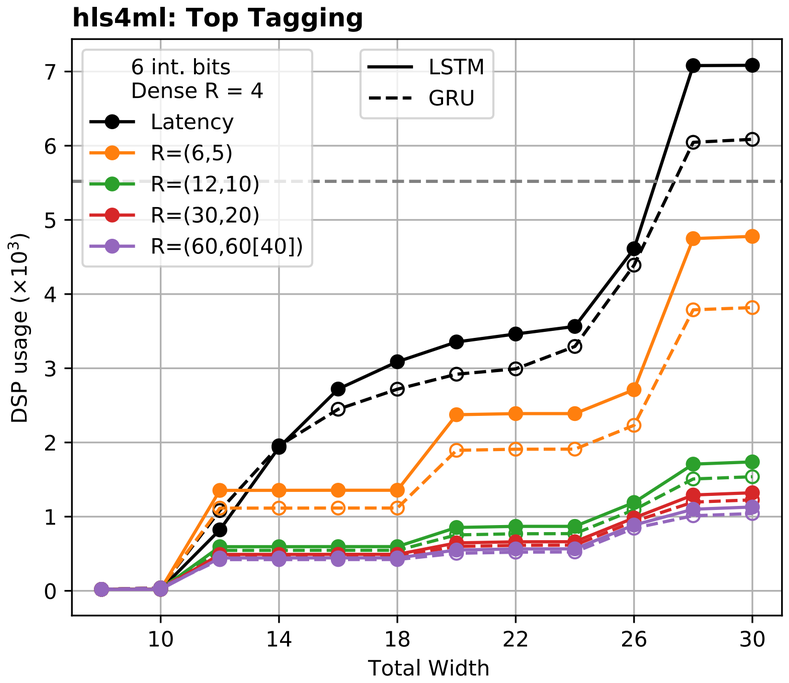


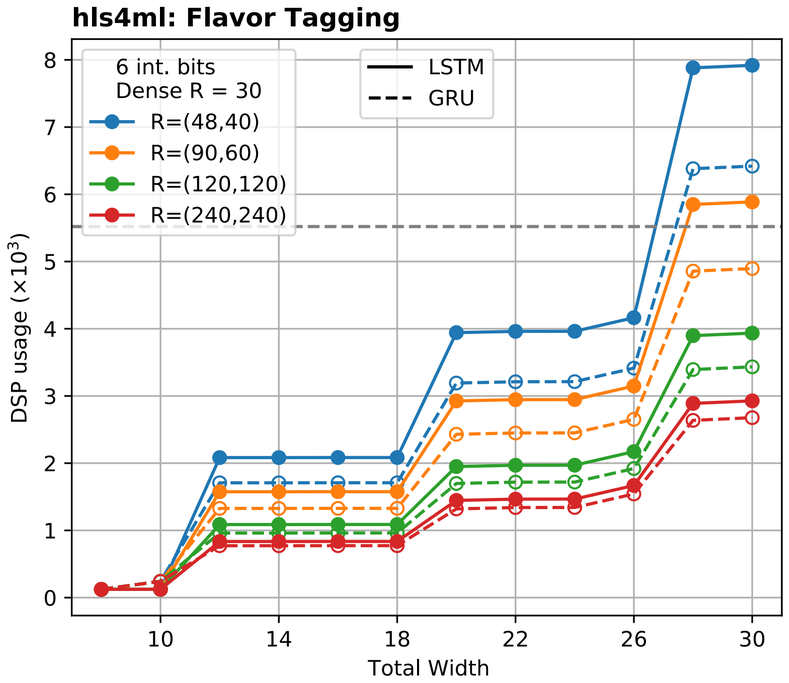
HLS Synthesis: DSP Usage

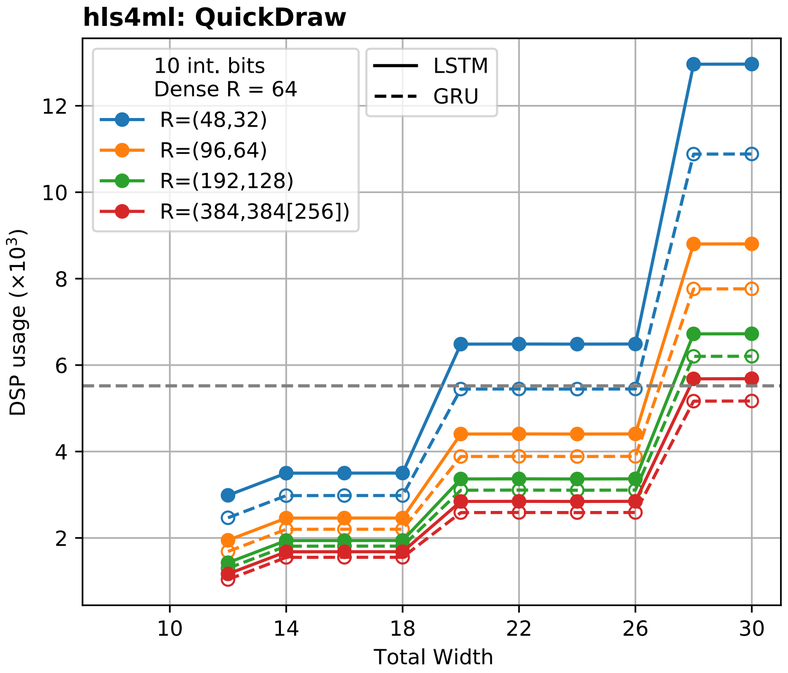
Mach. Learn.: Sci. Tec. 4 025004

- **DSP usage** as a function of **Total bit width** after HLS synthesis
- The Jumps correspond to DSP input width

Synthesized using Xilinx Kintex UltraScale FPGA
FPGA part: xcku115-flvb2104-2-i







Who am I

I am a postdoc at the University of Washington at Seattle (since March 2021)
*Working with Shih-Chieh Hsu for **ATLAS** and **NSF A3D3 Institute***

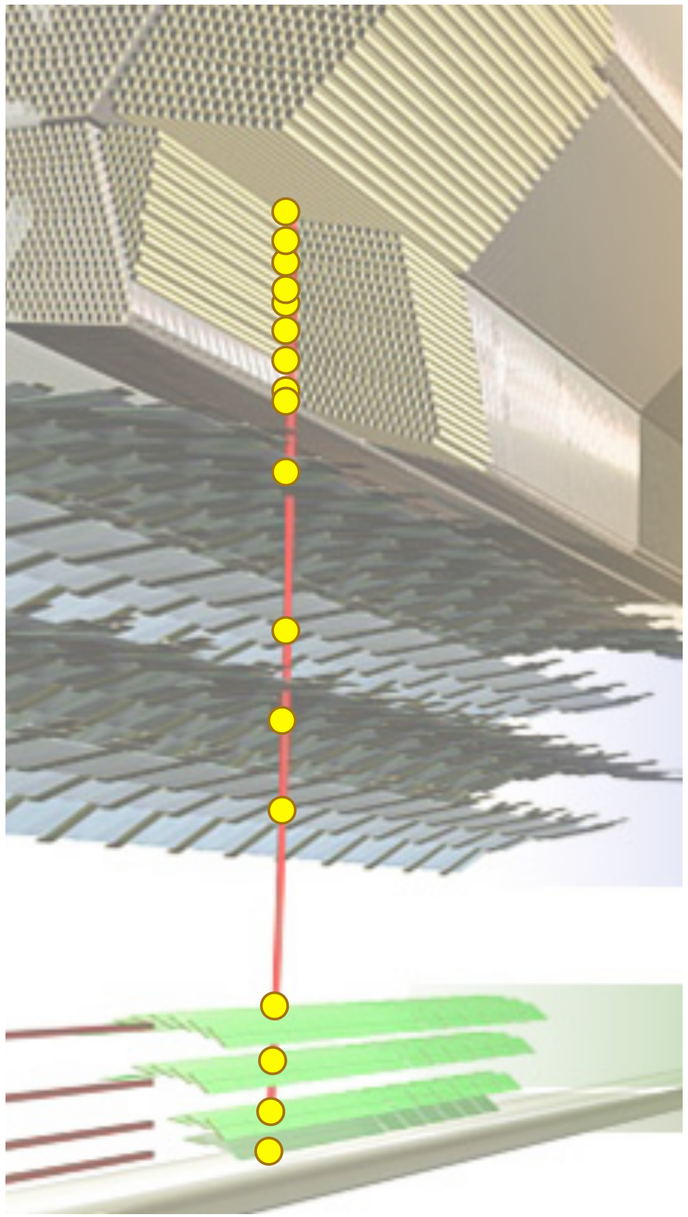
I am visiting LBNL as a postdoc affiliate (March 2022 - Now)

Short Bio:
March 2021, Ph.D.
University of British Columbia, Vancouver, Canada

Thesis: Searches for new high-mass resonances in top-antitop and di-electron final states using the ATLAS detector

ID Track Reconstruction

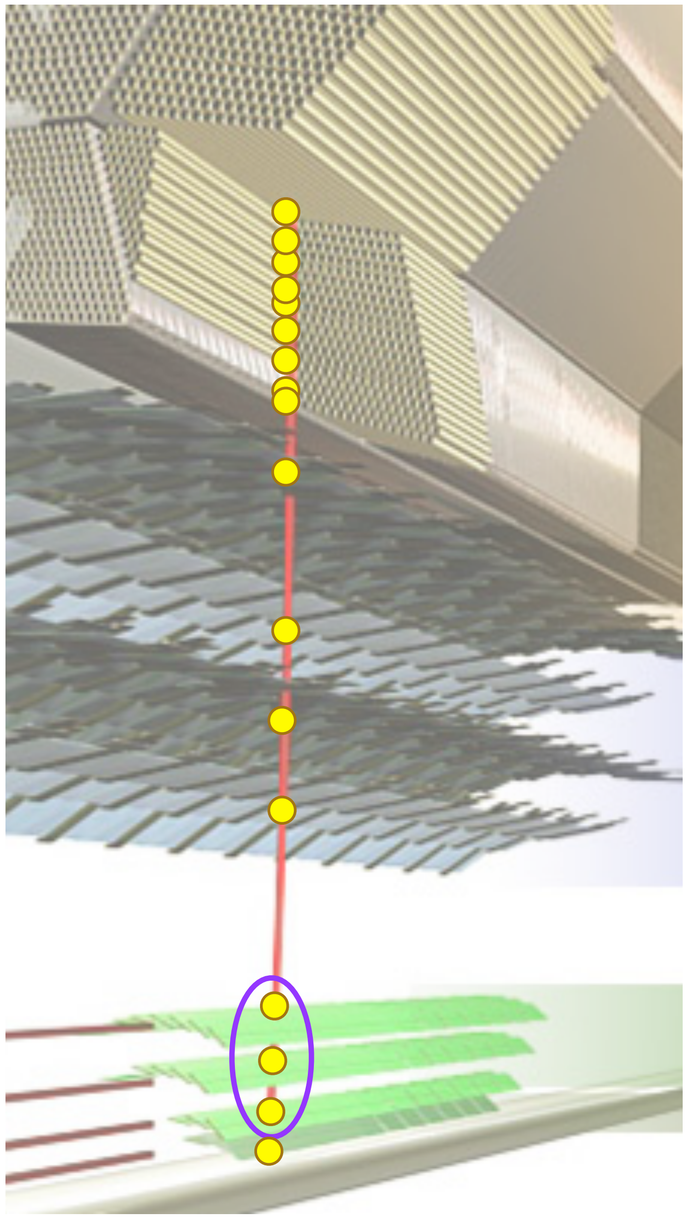
Design Credit: G. Gaycken



Build Clusters

ID Track Reconstruction

Design Credit: G. Gaycken

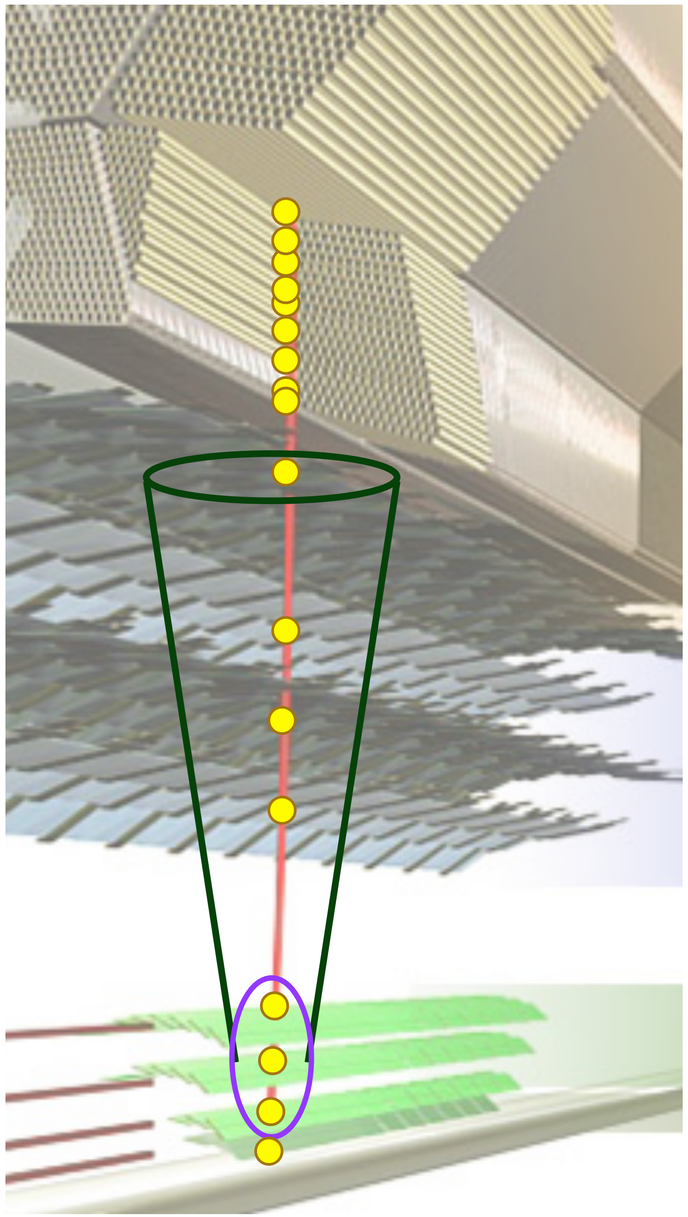


Build Clusters

Make track seeds

ID Track Reconstruction

Design Credit: G. Gaycken



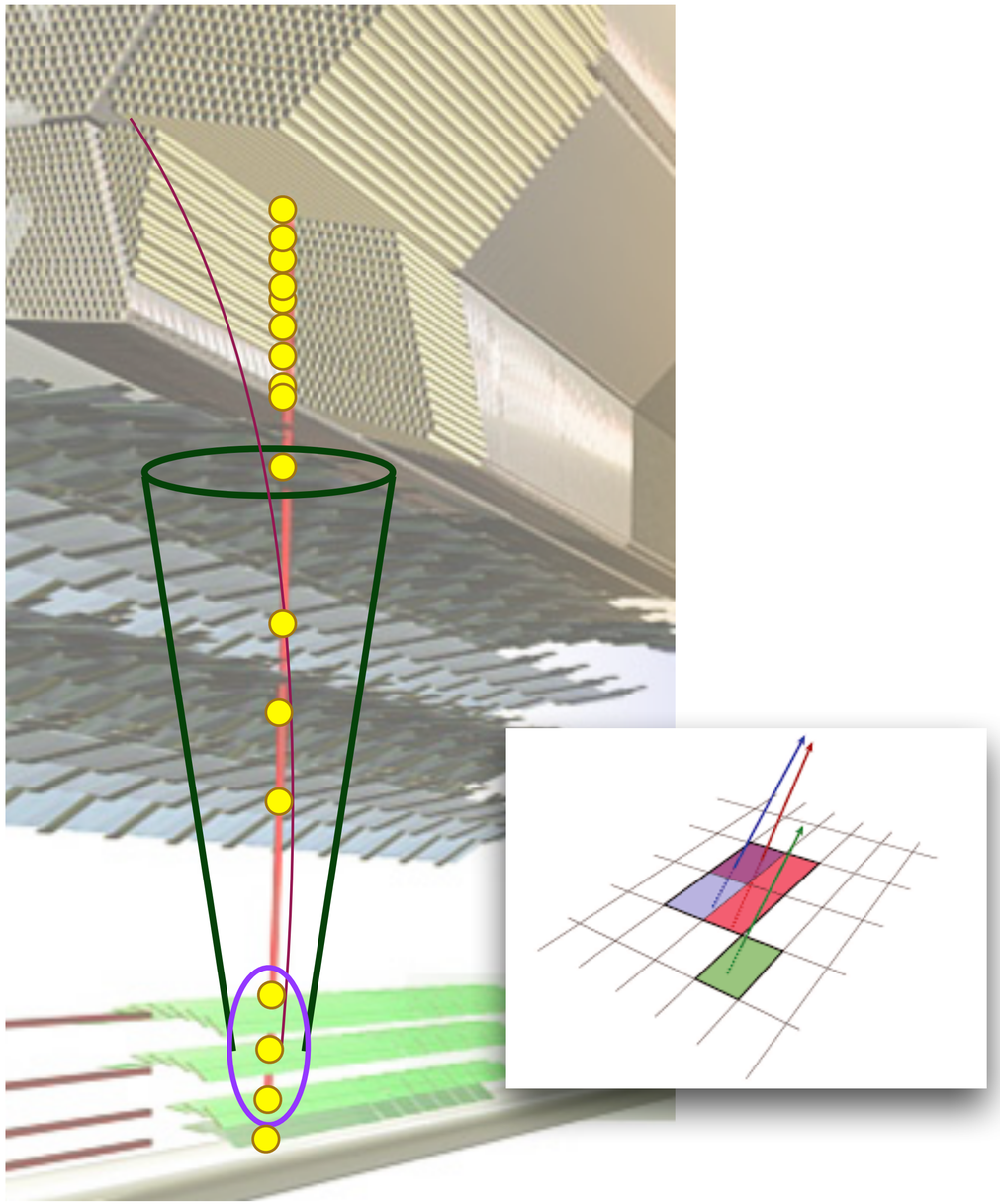
Build Clusters

Make track seeds

Find tracks

ID Track Reconstruction

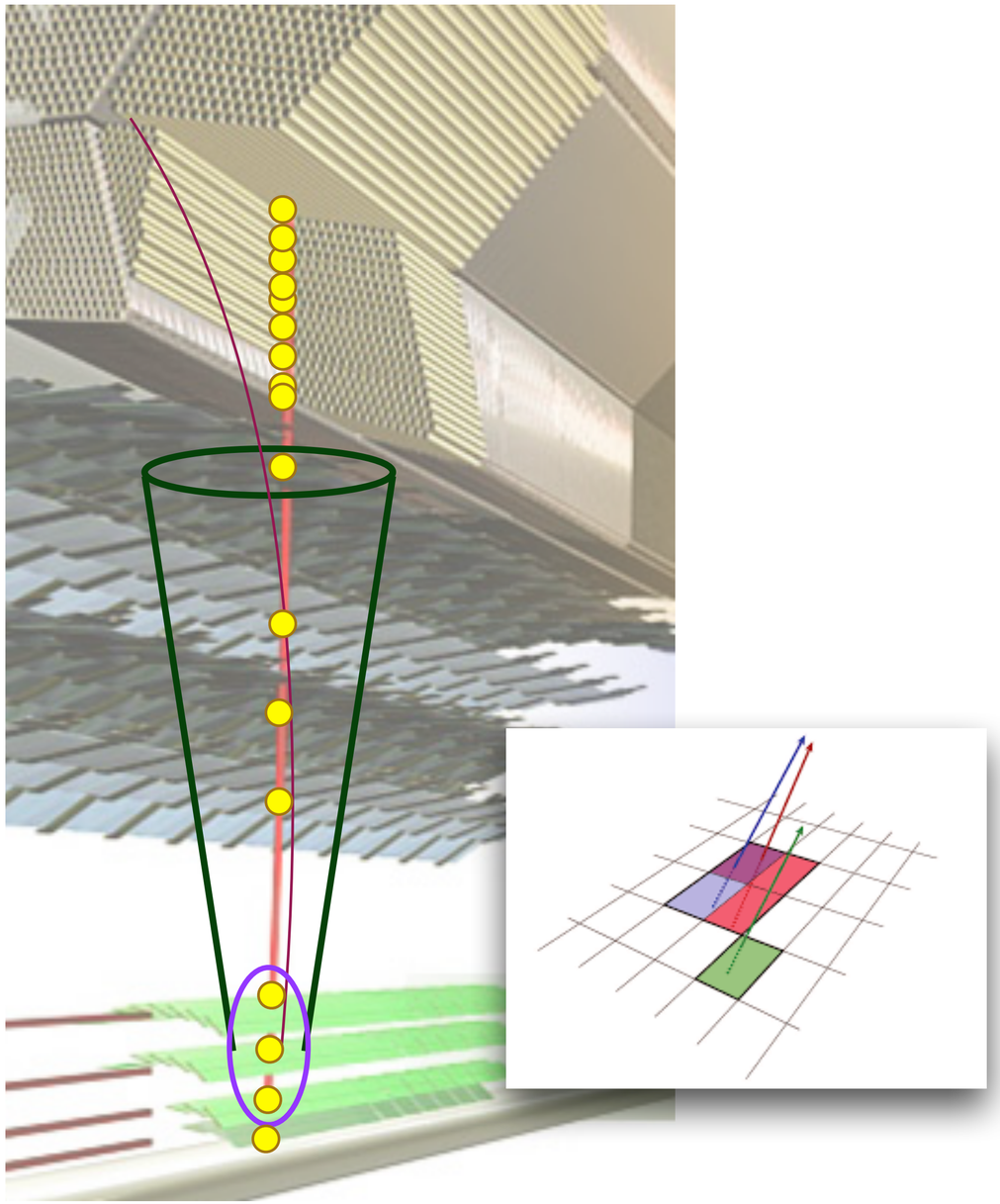
Design Credit: G. Gaycken



- Build Clusters
- Make track seeds
- Find tracks
- Handle merged Pixel clusters (NN)

ID Track Reconstruction

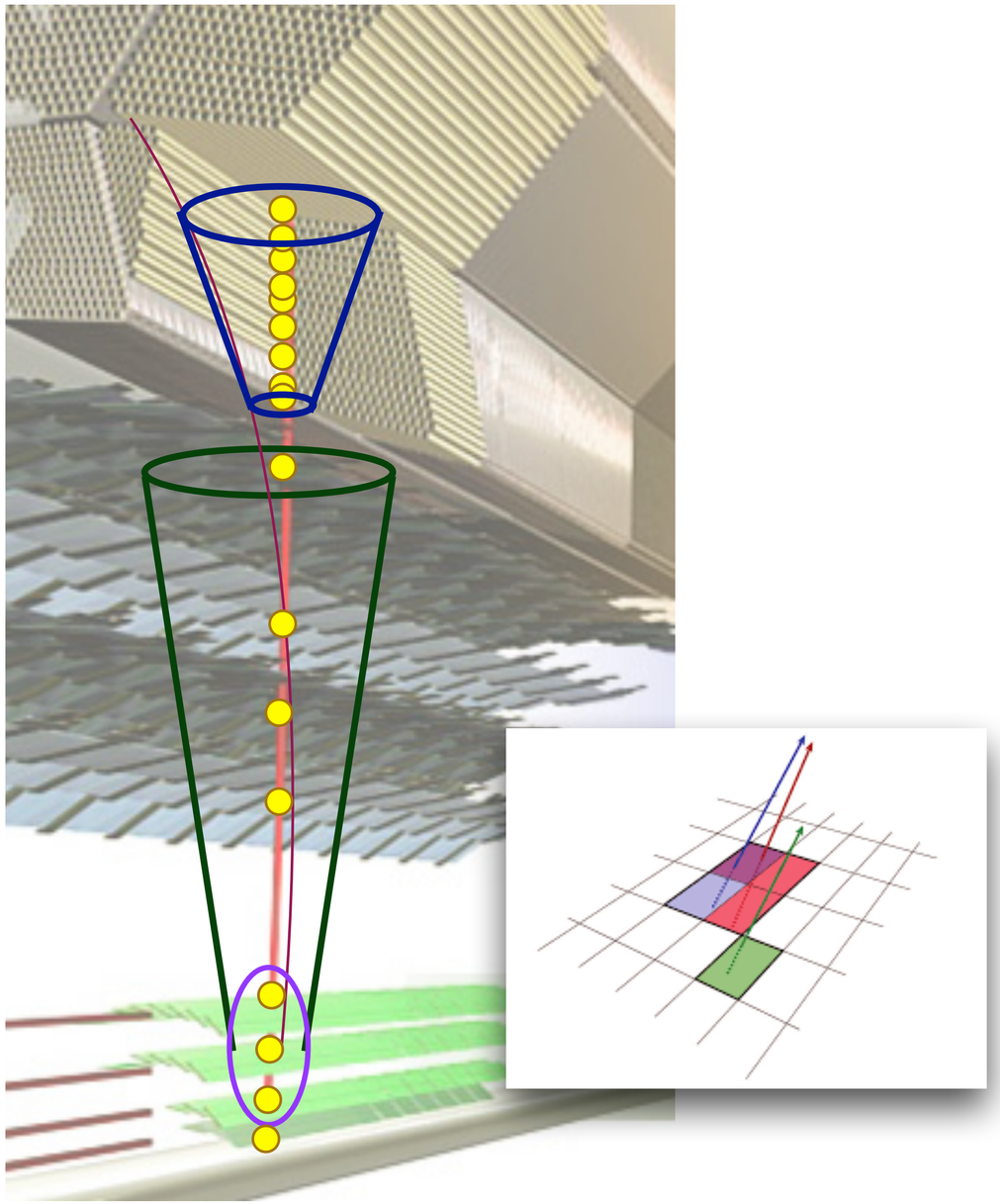
Design Credit: G. Gaycken



- Build Clusters
- Make track seeds
- Find tracks
- Handle merged Pixel clusters (NN)
- Resolve ambiguities & fit track

ID Track Reconstruction

Design Credit: G. Gaycken



Build Clusters

Make track seeds

Find tracks

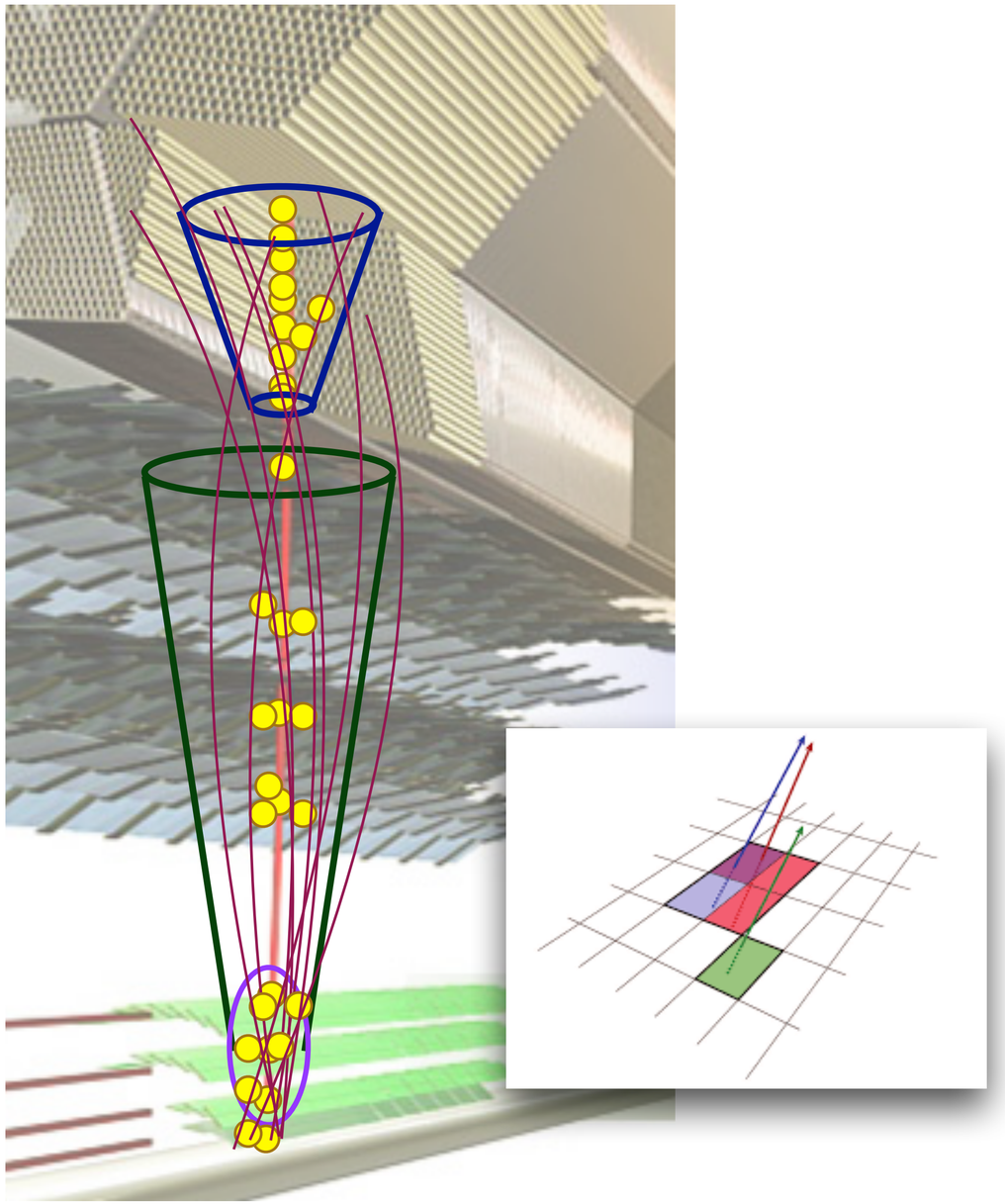
Handle merged Pixel clusters (NN)

Resolve ambiguities & fit track

Extend to TRT

ID Track Reconstruction

Design Credit: G. Gaycken



Build Clusters

Make track seeds

Find tracks

Handle merged Pixel clusters (NN)

Resolve ambiguities & fit track

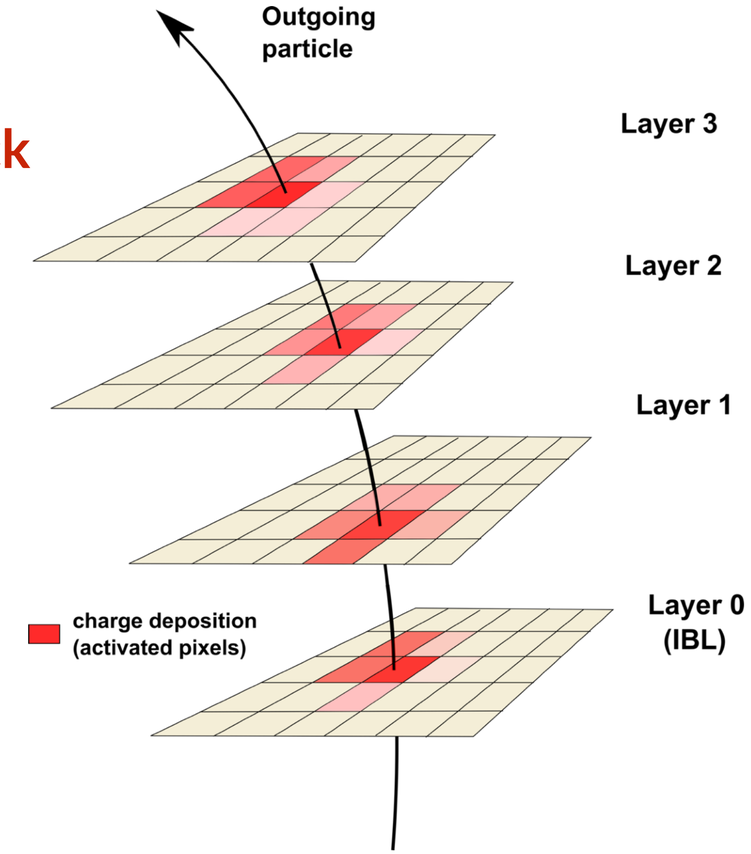
Extend to TRT

Tracking in Dense Environment

Dense Environment: average separation between highly collimated tracks is comparable to the granularity of individual sensors

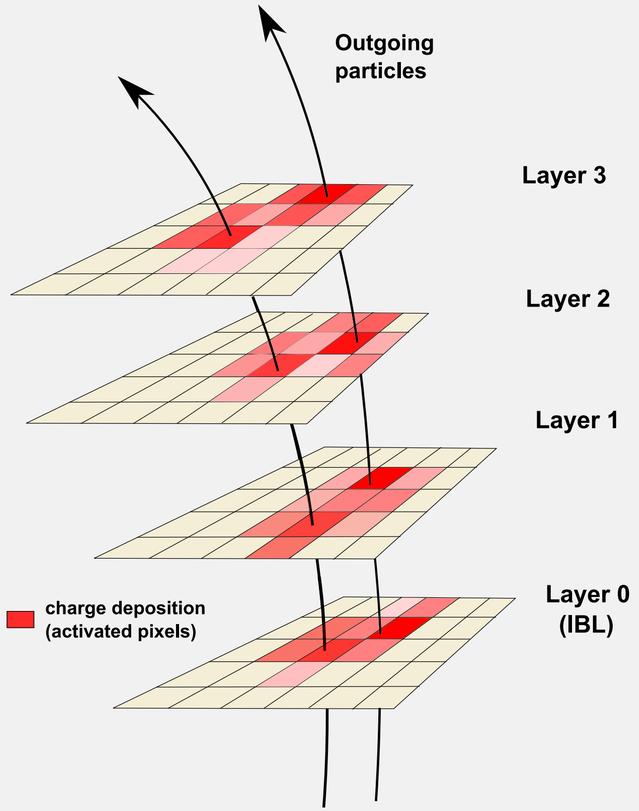
Ex: Cores of highly energetic hadronic top jets, or tau

Isolated track

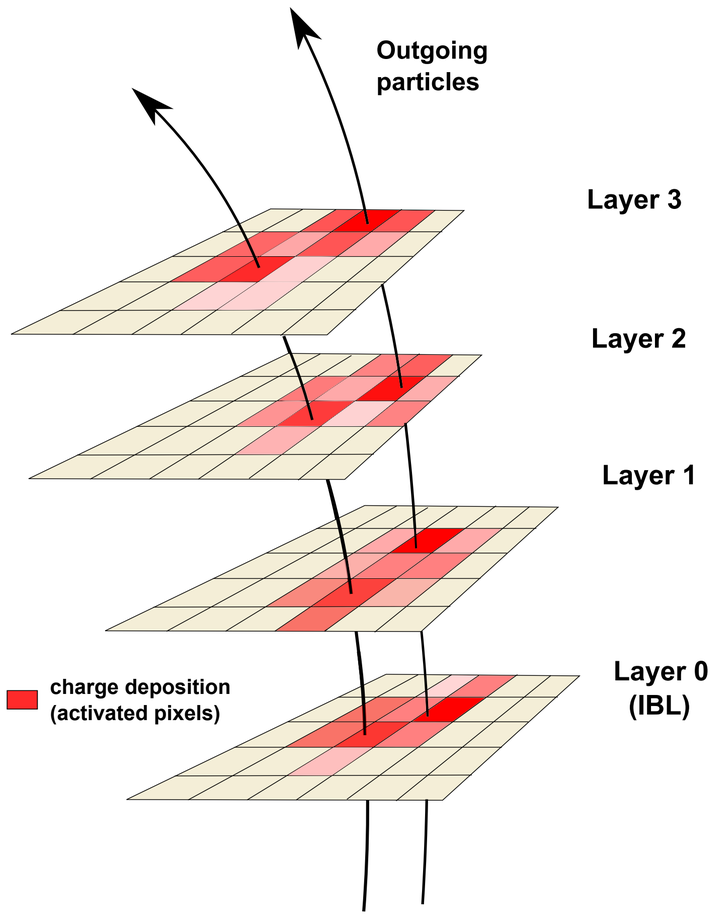


Merged clusters

Tracks are too close to each-other



Merged Pixel Cluster splitting with MDN

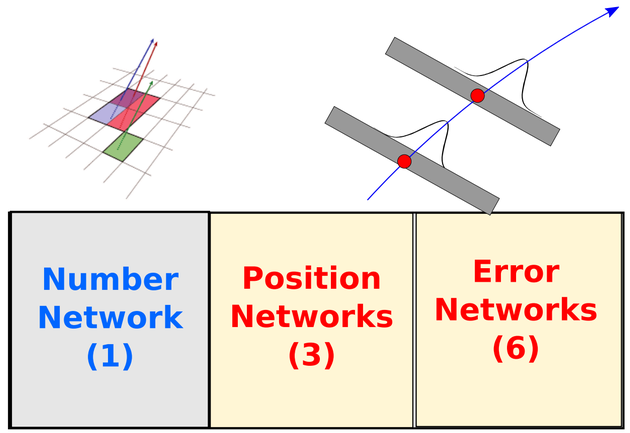


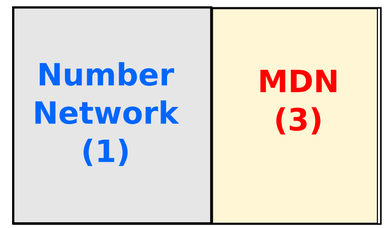
- All tracks are scored in the ambiguity solving stage
- **Shared clusters** penalizes the track score

- Tracks with low score are not fitted and stored
  - ➡ loss of track reconstruction efficiency

Run 2
10 Neural Networks

Run 3
4 Neural Networks

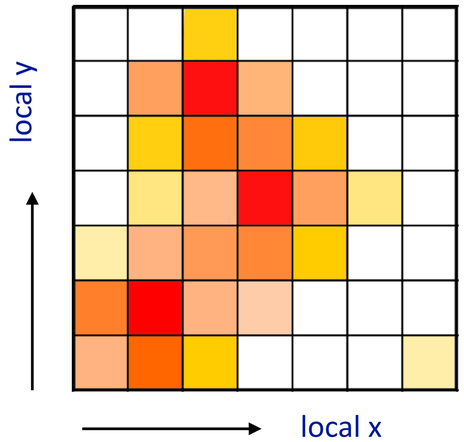




Number of hits (Classifier) | Hit position (Regression) | Hit uncertainty (Classifier)

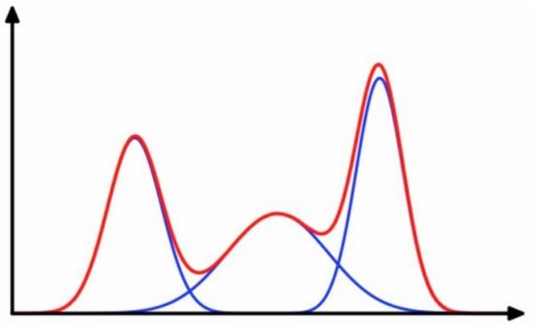
Number of hits (Classifier) | Mixture Density Network (Estimates hit position and uncertainty)

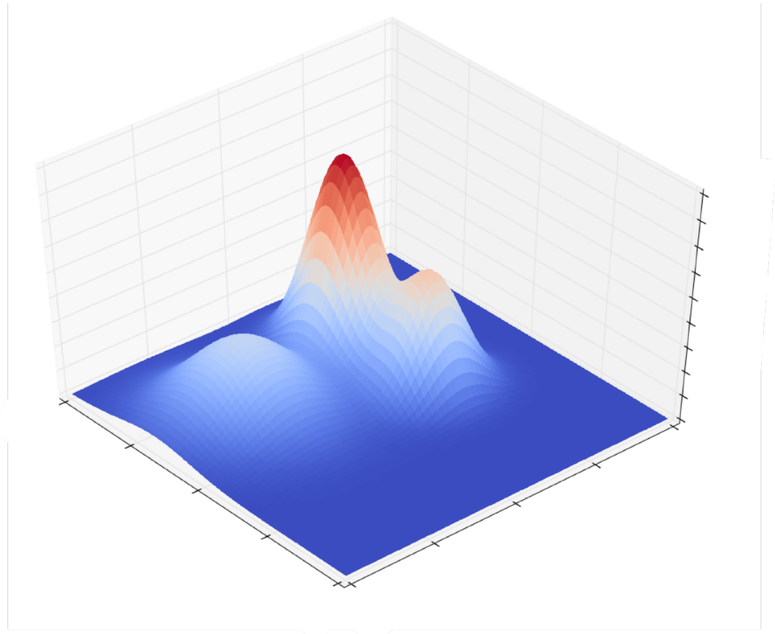
Mixture Density Network



Approximating
$p(\text{hit position} \mid \text{input})$

with **Gaussian Mixture model**





Extract
mean and **std**
of each Gaussian component

Standard NN vs MDN

Input feature vector $\mathbf{x} = \{x_1, \ldots x_d\}$
Target vector $\mathbf{t} = \{t_1, \ldots t_c\}$

Neural Network Training

A training example: $\{\mathbf{x}^q, \mathbf{t}^q\}$
Training: Learns the underlying data generator

$$p(\mathbf{t}, \mathbf{x}) = p(\mathbf{t}|\mathbf{x})p(\mathbf{x})$$

Conventional Least Square Approach

Minimize

$$E(\mathbf{w}) = \frac{1}{2}\sum_{q=1}^{n}\sum_{k=1}^{c}\left[f_k(\mathbf{x}^q; \mathbf{w}) - t_k^q\right]^2$$

f_k = Network Function

=

Equivalent Approach: Maximum Likelihood

Assume:

$$p(\mathbf{t}|\mathbf{x}) = \prod_{k=1}^{c} p(t_k|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{c}\{f_k(\mathbf{x}; \mathbf{w}) - t_k\}^2\right]$$

σ = global variance

Maximize: $$\mathcal{L} = \prod_{q=1}^{n} p(\mathbf{t}^q, \mathbf{x}^q) = \prod_{q=1}^{n} p(\mathbf{t}^q|\mathbf{x}^q)p(\mathbf{x}^q)$$

Standard NN vs MDN

Input feature vector $\mathbf{x} = \{x_1, \ldots x_d\}$
Target vector $\mathbf{t} = \{t_1, \ldots t_c\}$

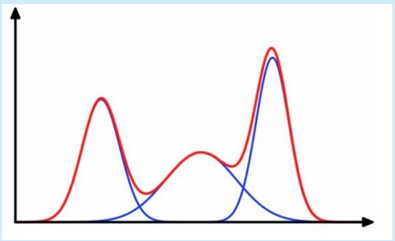
Neural Network Training

A training example: $\{\mathbf{x}^q, \mathbf{t}^q\}$
Training: Learns the underlying data generator

$$p(\mathbf{t}, \mathbf{x}) = p(\mathbf{t}|\mathbf{x})p(\mathbf{x})$$

Mixture Density Network

Assume: $p(\mathbf{t}|\mathbf{x}) = \sum_{i=1}^{m} \alpha_i(\mathbf{x})\phi_i(\mathbf{t}|\mathbf{x})$ Conditional distribution = Mixture Model



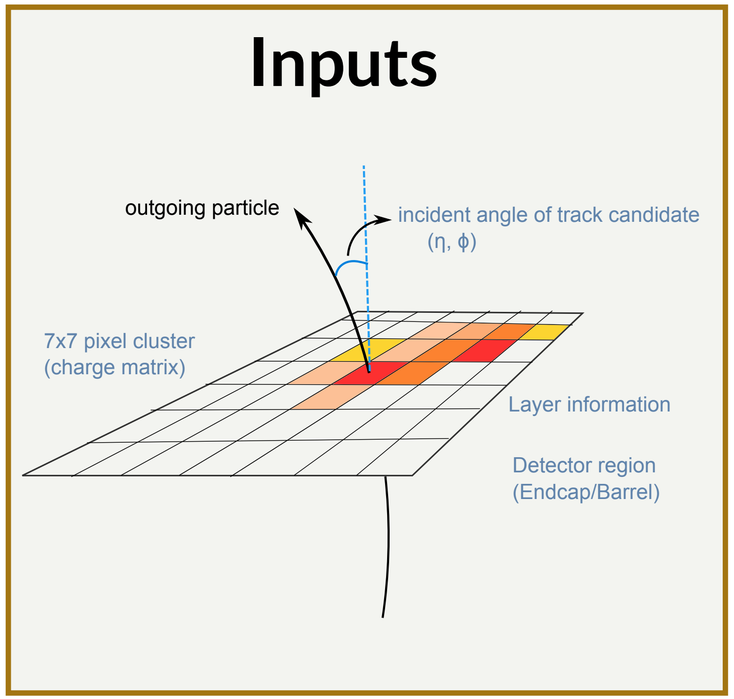
i^{th} Kernel Density: $\phi_i(\mathbf{t}|\mathbf{x}) = \sqrt{\frac{\beta_i(\mathbf{x})}{2\pi}} \exp\left[-\frac{\beta_i(\mathbf{x})}{2}\|\mathbf{t} - \mu_i(\mathbf{x})\|^2\right]$ β_i = precision = 1 / covariance

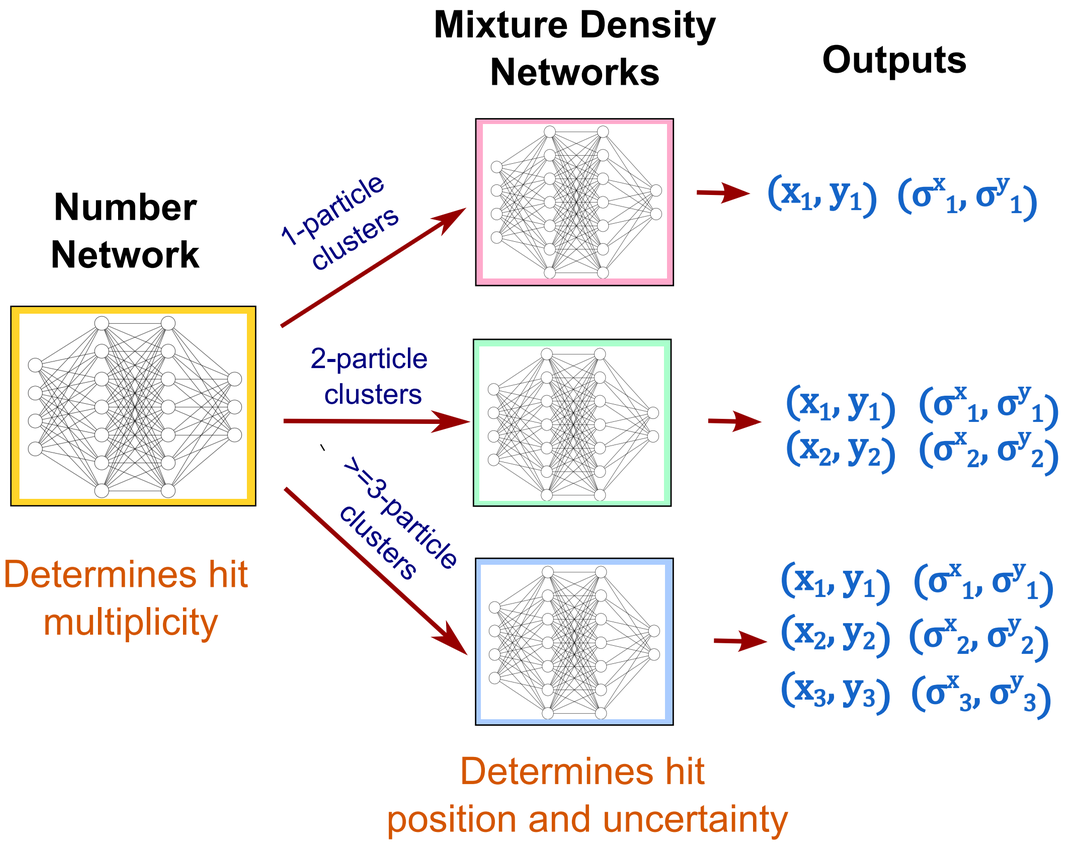
Maximize: $\mathcal{L} = \prod_{q=1}^{n} p(\mathbf{t}^q, \mathbf{x}^q) = \prod_{q=1}^{n} p(\mathbf{t}^q|\mathbf{x}^q)p(\mathbf{x}^q)$

Pixel cluster splitting with NNs

To split the merged clusters, MDN is currently used in Run-3

- 1 **Number network** (3-class classifier)
- 3 **Mixture Density Networks** (MDN)

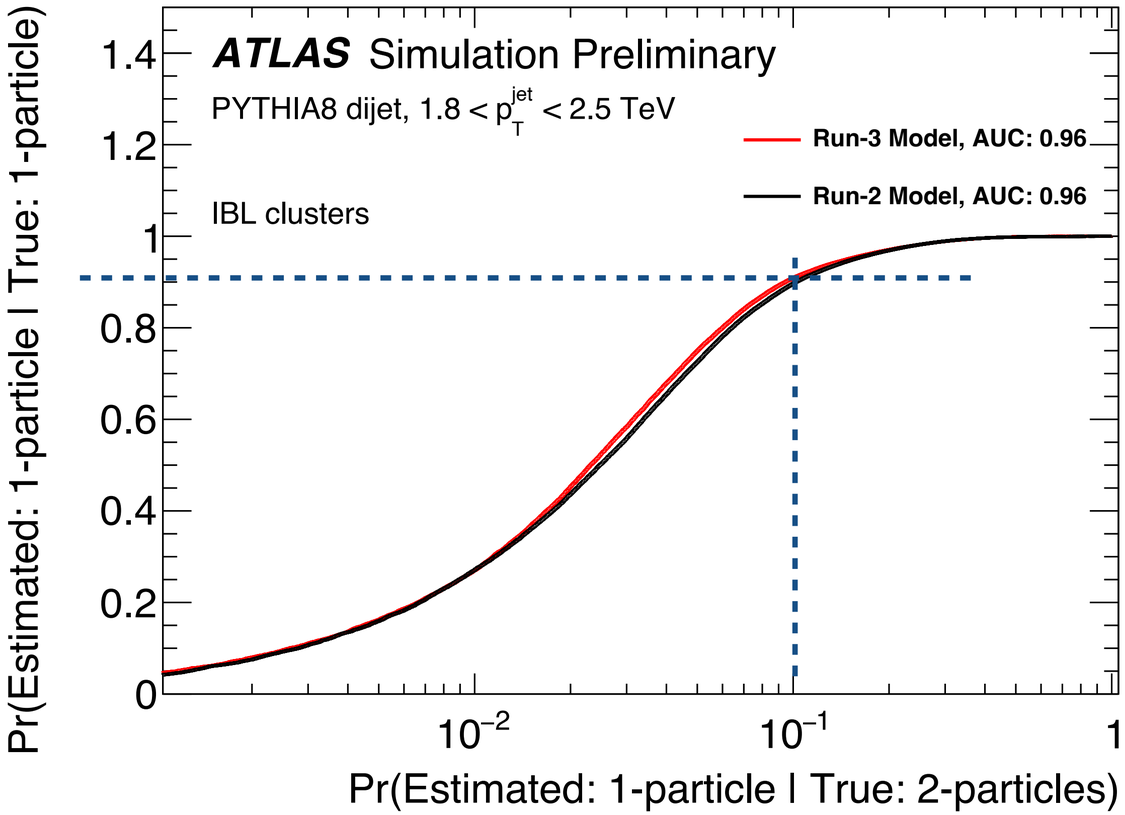




Number NN ROC Curves

IDTR-2021-001

1 particle efficiency
vs
2 particles faking 1 particle



MDN: 1-particle clusters

PoS (LHCP2019) 009

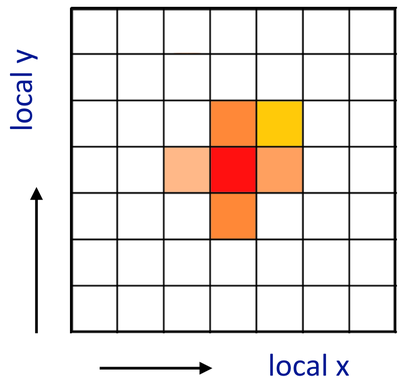
IDTR-2021-001

Position estimation

$$\text{residual} = x(y)_{\text{pred}} - x(y)_{\text{true}}$$

Error estimation

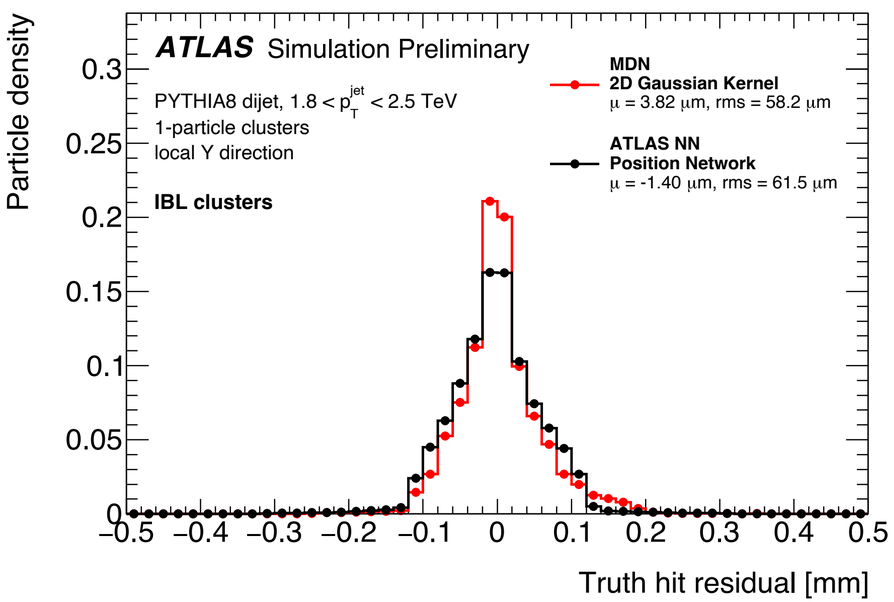
$$\text{pull} = \frac{\text{residual}}{\sigma_{x(y),\text{pred}}}$$

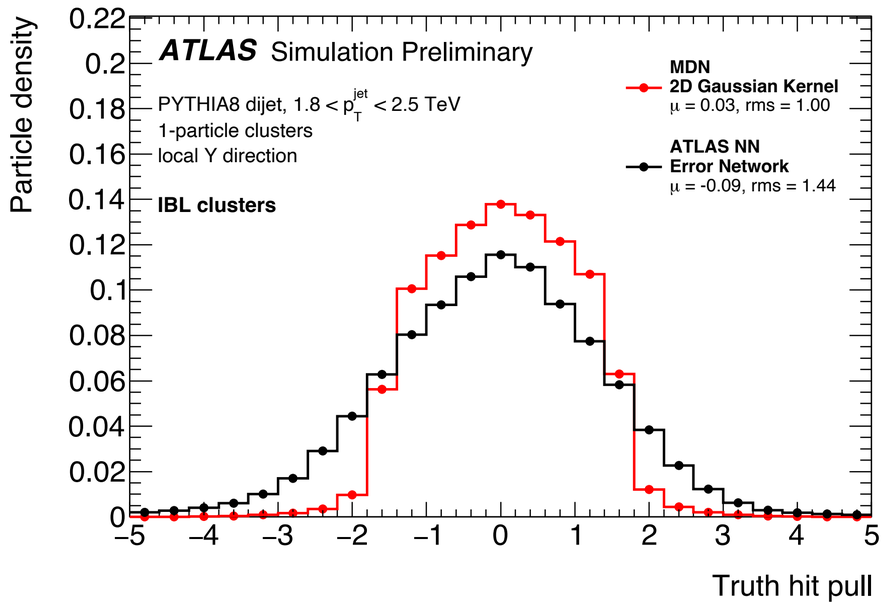


MDN residuals: large peak at zero → more accurate

MDN Pulls: $\sim \mathcal{N}(0,1)$

IBL 1-particle clusters: Y-direction





MDN: 3-particle clusters

PoS (LHCP2019) 009

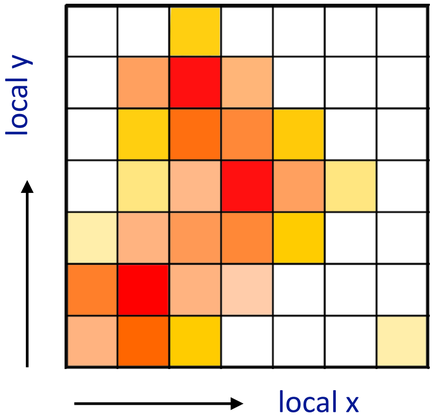
IDTR-2021-001

Position estimation

$$\text{residual} = x(y)_{\text{pred}} - x(y)_{\text{true}}$$

Error estimation

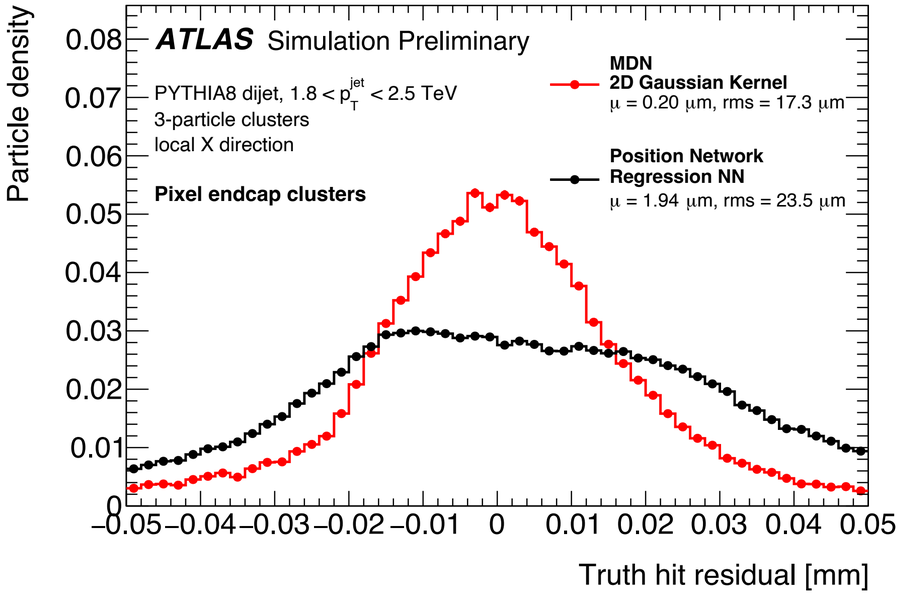
$$\text{pull} = \frac{\text{residual}}{\sigma_{x(y),\text{pred}}}$$

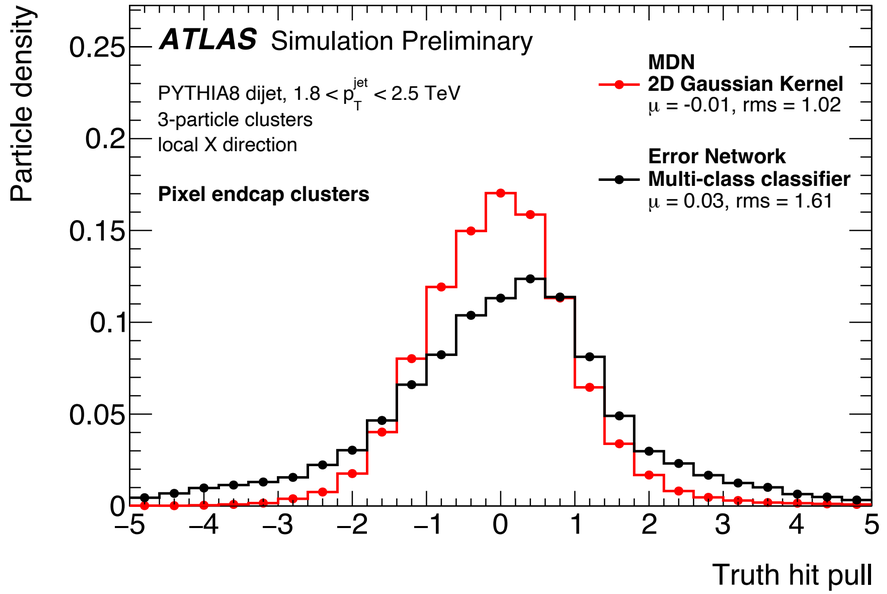


MDN residuals: large peak at zero → more accurate

MDN Pulls: $\sim \mathcal{N}(0,1)$

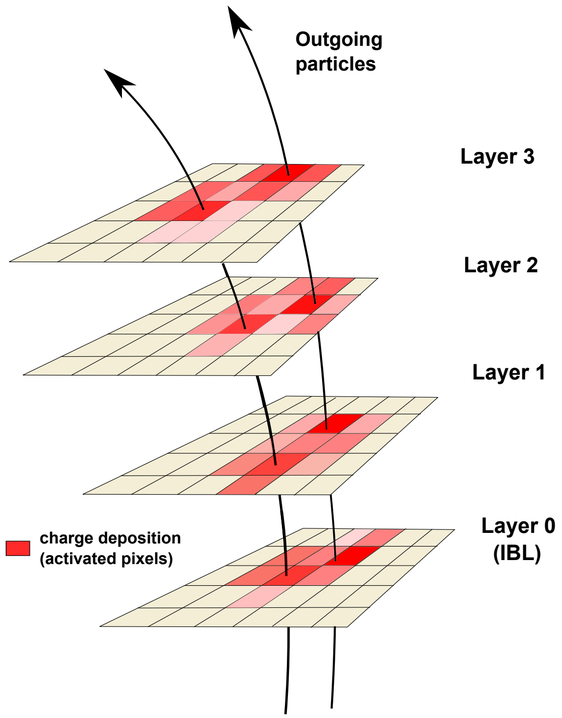
Endcap 3-particle clusters: X-direction

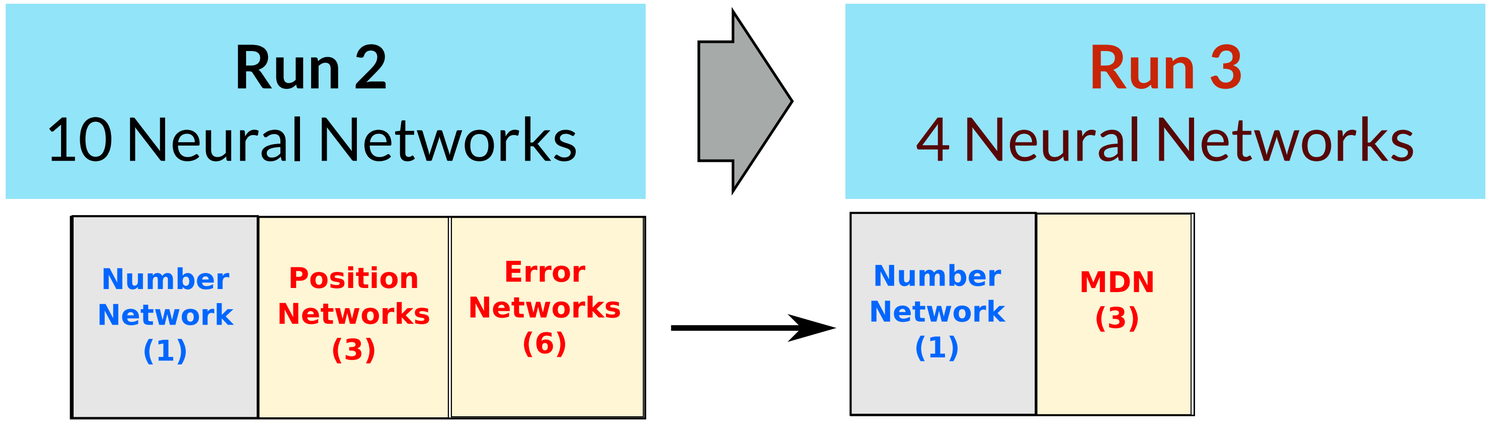




Summary so far ..

Merged charge cluster splitting using Neural Networks





MDNs learn the probability density function of the hit position

- Estimate both hit position and uncertainty
- Approximates the density with mixture of Gaussians

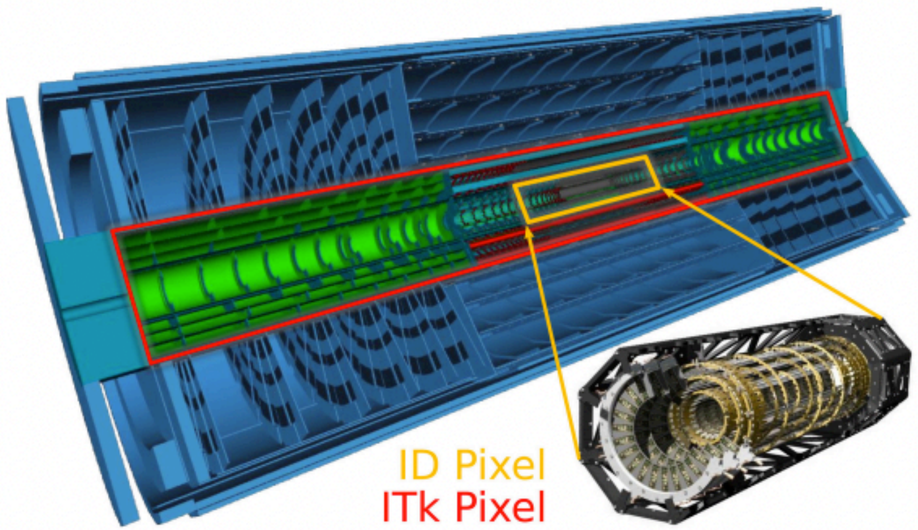
ATLAS Upgrade: Run 4 and beyond

- ATLAS Inner tracking detector will be replaced after 2026
- New pixel pitch will be much smaller

  $50 \times 400 \mu m^2 \rightarrow 50 \times 50\ \mu m^2$ and $25 \times 100\ \mu m^2$

Starting to develop new NN-based algorithm for future tracking

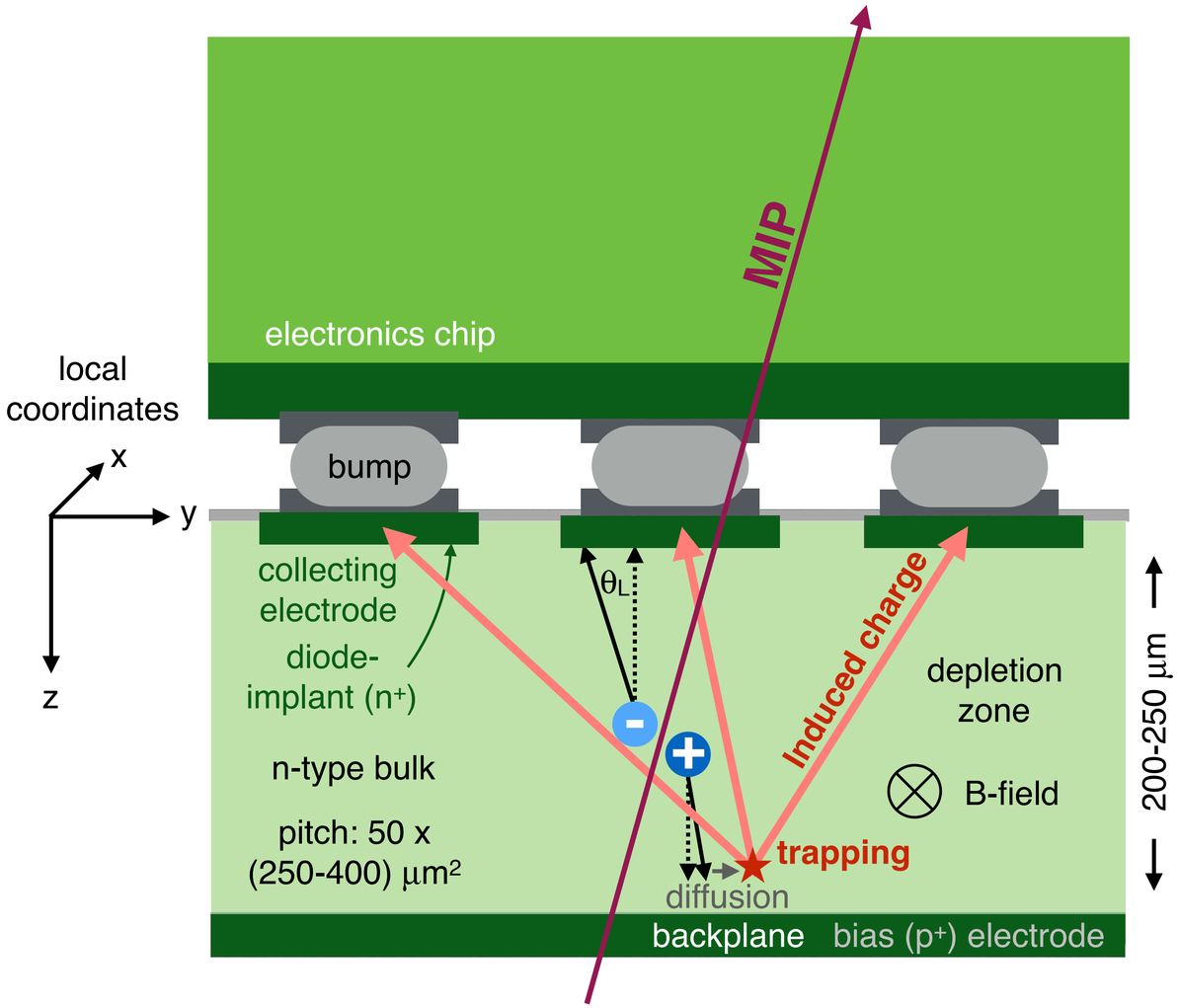
Future ATLAS tracking detector: ITk



Radiation Damage Digitizer* in Pixel Sensors

- Radiation damage deforms the drift of the charged carries towards the electrodes
- Their path gets deflected by magnetic field (Lorentz angle) and diffusion
- Due to radiation damage they could be trapped and induce/screen a fraction of their charge (Ramo potential)
- These effects will be modeled in the run-3 simulation, but haven't been so far

IDET-2020-01, arXiv:2106.09287, accepted by Jinst



JINST 14 P06012

***Digitization happens after simulated charge deposition and before space point reconstruction**

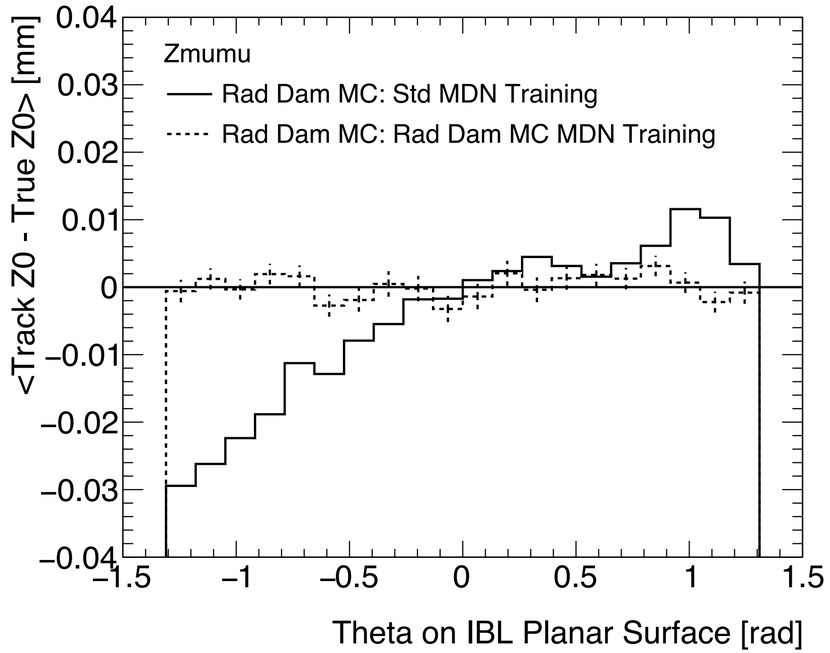
Summary and Outlook

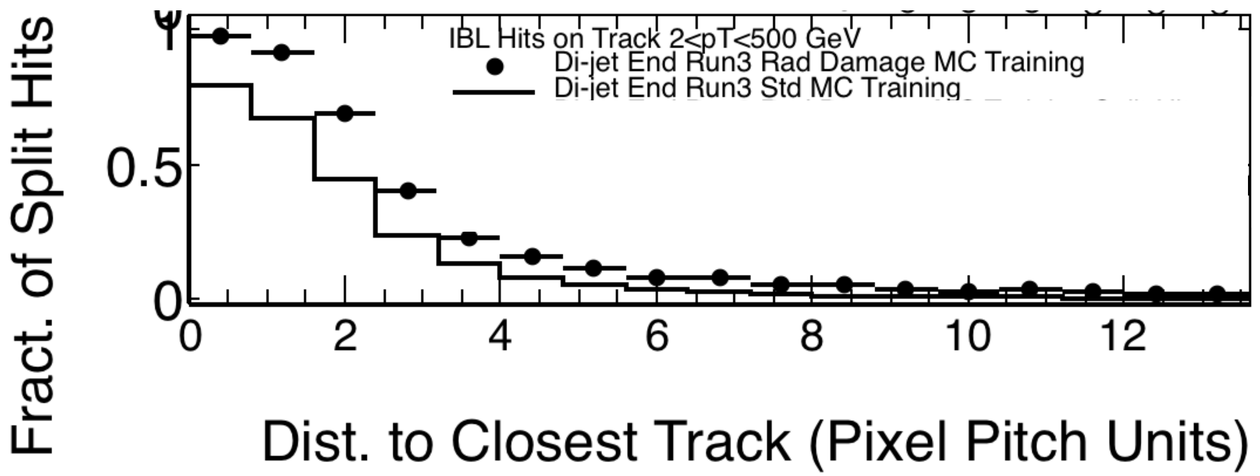
- Reduction of charge loss (due to radiation damage) affects pixel split rate

- Performance of retrained NNs with radiation damage MC looks promising
  - Studied performance with end of Run-3 condition

Next

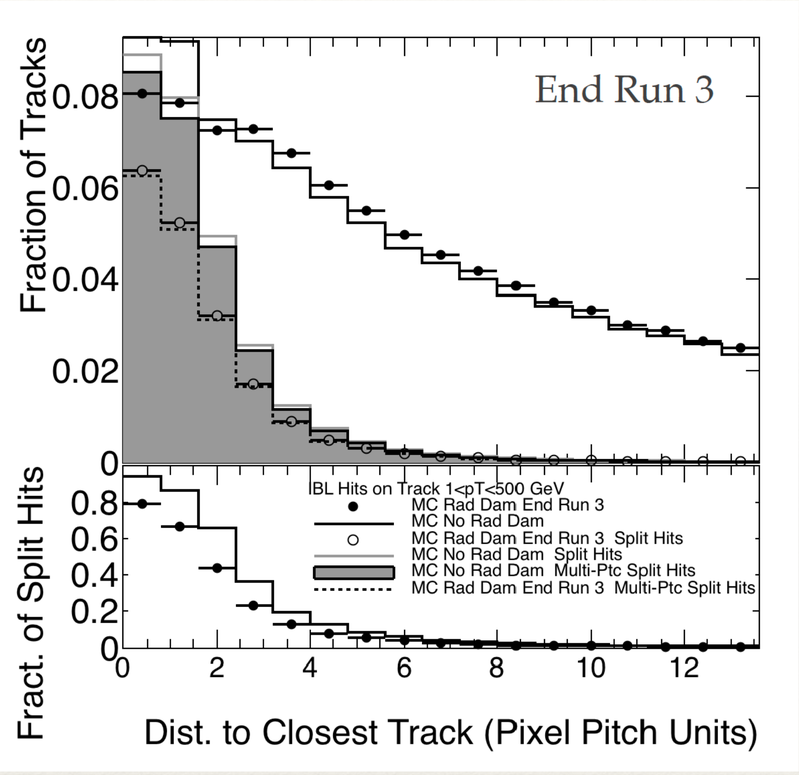
- Studies with early and mid Run-3 conditions
- We intend to use the training with rad-damage MC for Run3 with expected condition for mid year 2022-2024

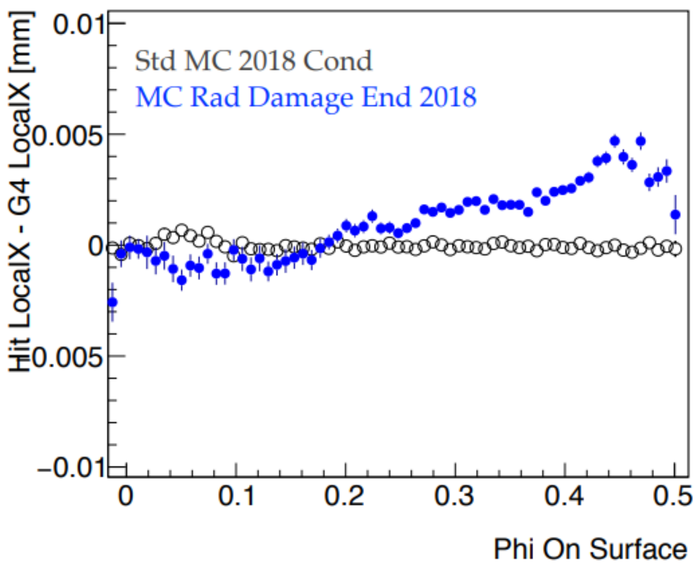


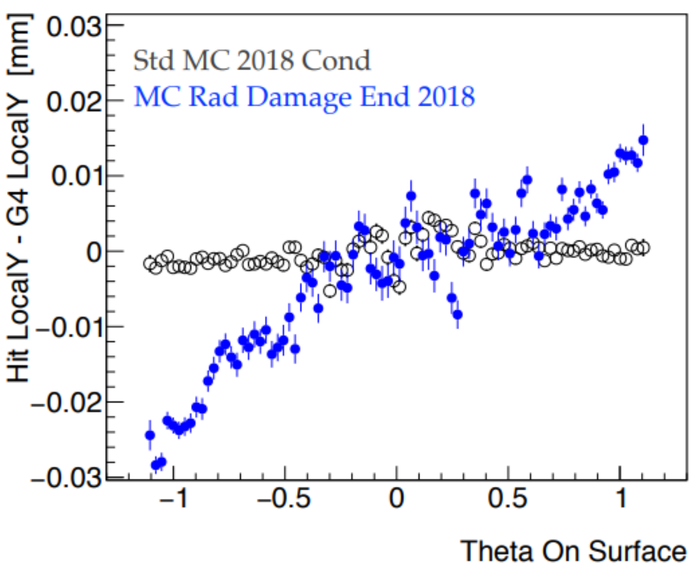


Effect of Radiation Damage

- **MC Rad Damage predicts significant loss of rate of pixel split hits with increasing fluen Si.**
- **Residual asymmetry as a function of incident angle**



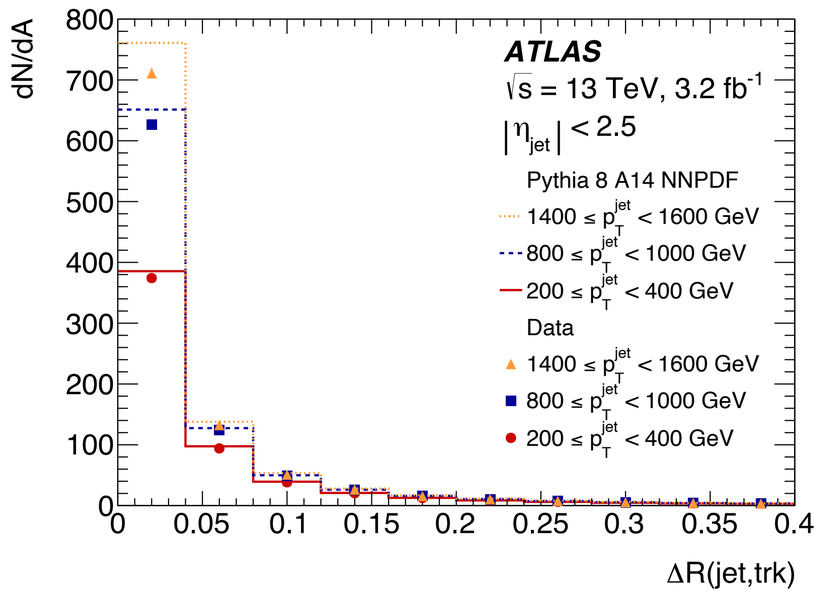


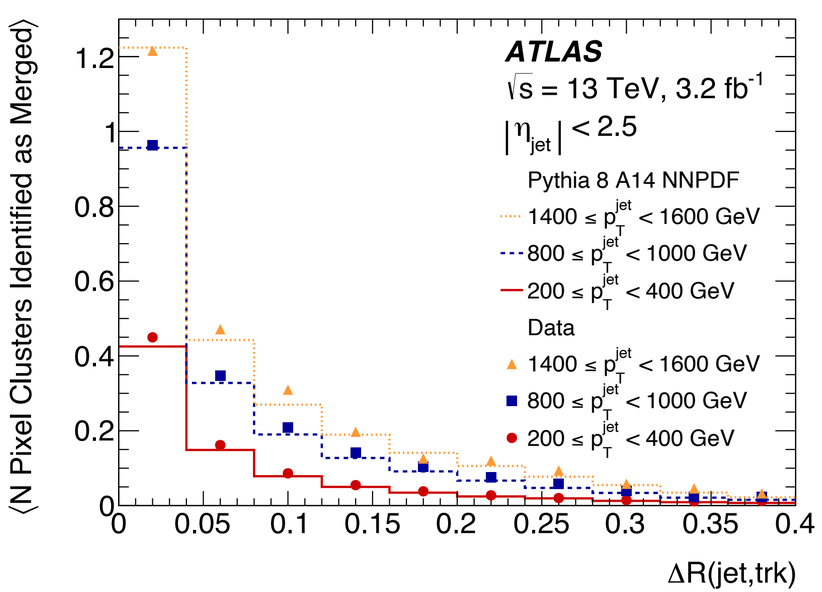


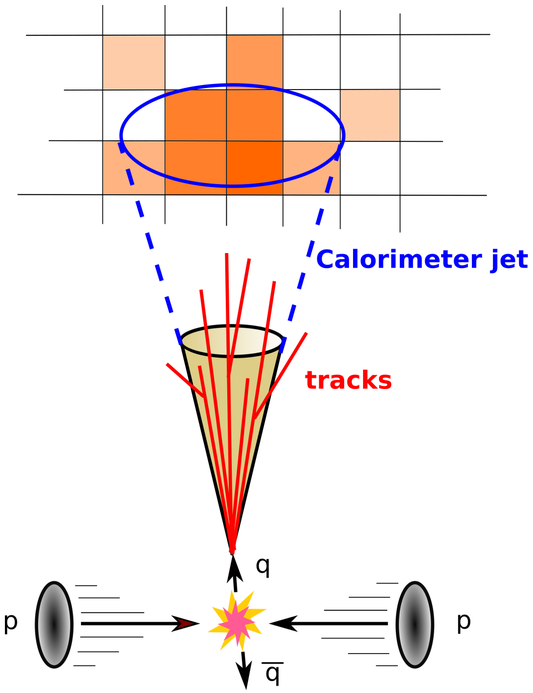
Tracking in Dense Environment

Dense Environment: average separation between highly collimated tracks is comparable to the granularity of individual sensors

Ex: Cores of highly energetic hadronic top jets, or tau







Most tracks are located within an **angular distance of 0.05** from the jet axis

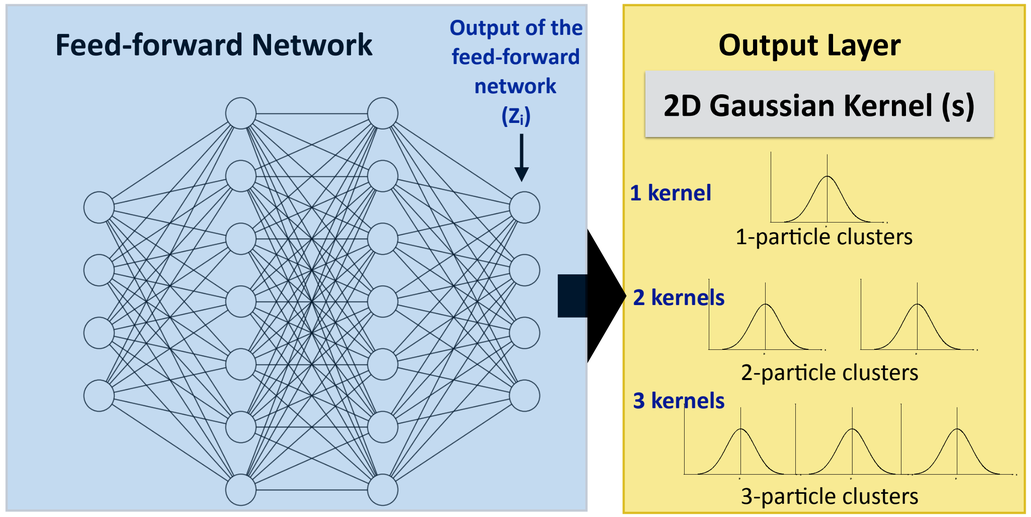
Number of pixel clusters identified as merged

Eur. Phys. J. C 77 (2017) 673

MDN: Architecture

Input

Same as the tracking NNs



$$p(\mathbf{t}|\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \cdot \phi_i(\mu, \beta)$$

Software

Keras + TensorFlow

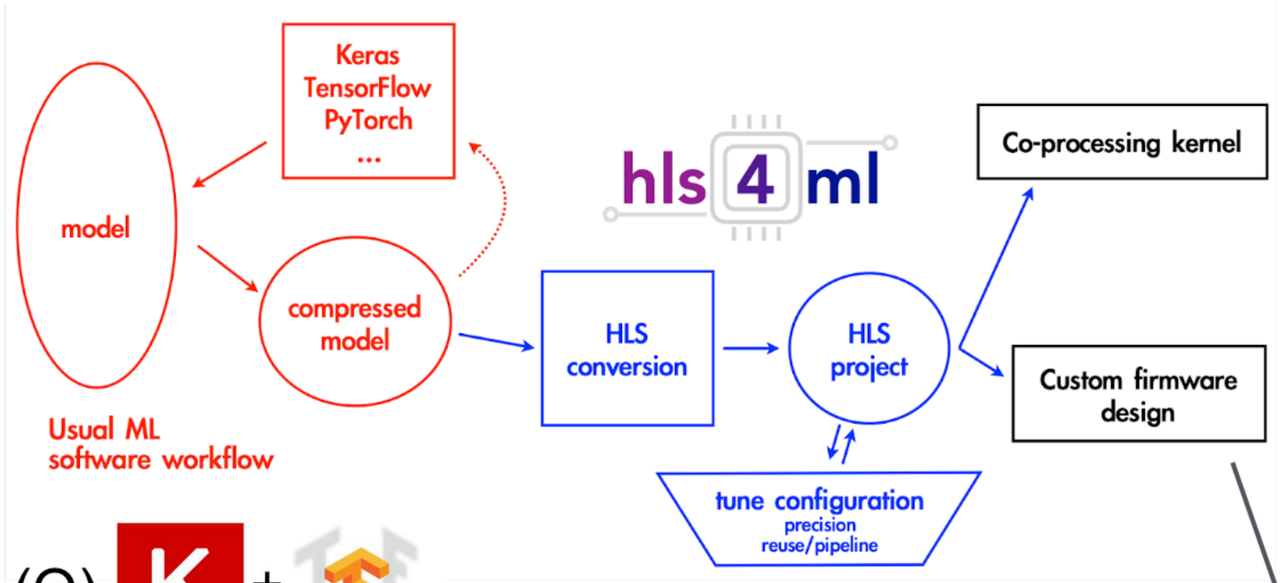
Keras + theano

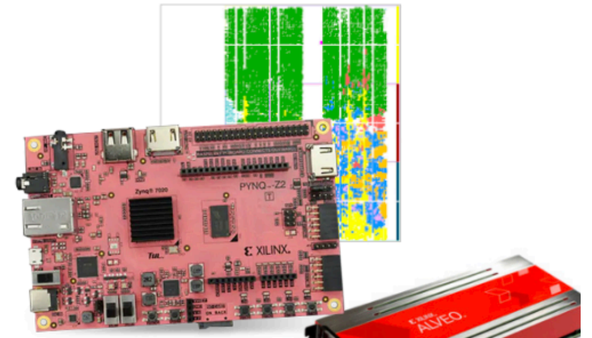
Table 1: Table summarizes the network structure and the hyperparameters. The input and output layer size is denoted with parenthesis.

Hyperparamaters	MDN (1 particle)	MDN (2 particles)	MDN (3 particles)
Structure	(60)-100-50-50-(1-2-2)	(60)-100-80-50-(2-4-4)	(60)-100-80-50-(3-6-6)
Activation	ReLu	ReLu	ReLu
Output Layer	1 GMM	2 GMM	3 GMM
Output activation	(softmax-linear-absolute)	(softmax-linear-absolute)	(softmax-linear-absolute)
Learning rate	0.0001	0.0001	0.0001
L2 regularizer	0.0001	0.0001	0.0001
Batch Size	100	100	100
Gradient Clipping	clipnorm = 1	clipnorm = 1	clipnorm = 1
Loss Function	MDN loss	MDN loss	MDN loss

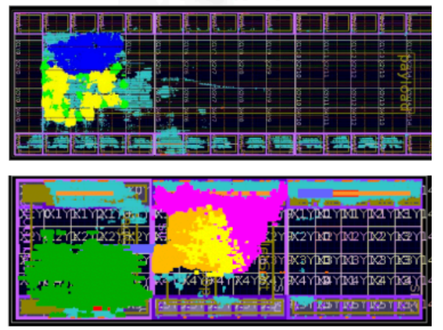
PoS (LHCP2019) 009

High Level Synthesis with Machine Learning (hls4ml)



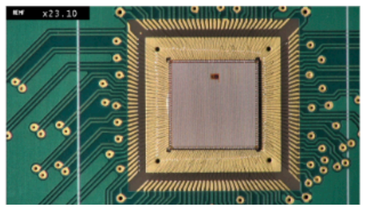


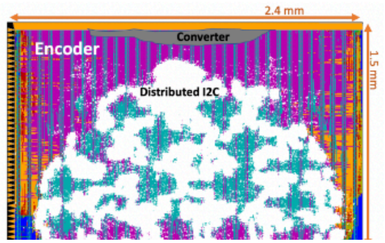
VivadoAccelerator
for Xilinx accelerators / SoCs



Custom design
(talk 1, 2)

ASICs (talk 1, 2)





(Q) K + TensorFlow

(Q) ONNX

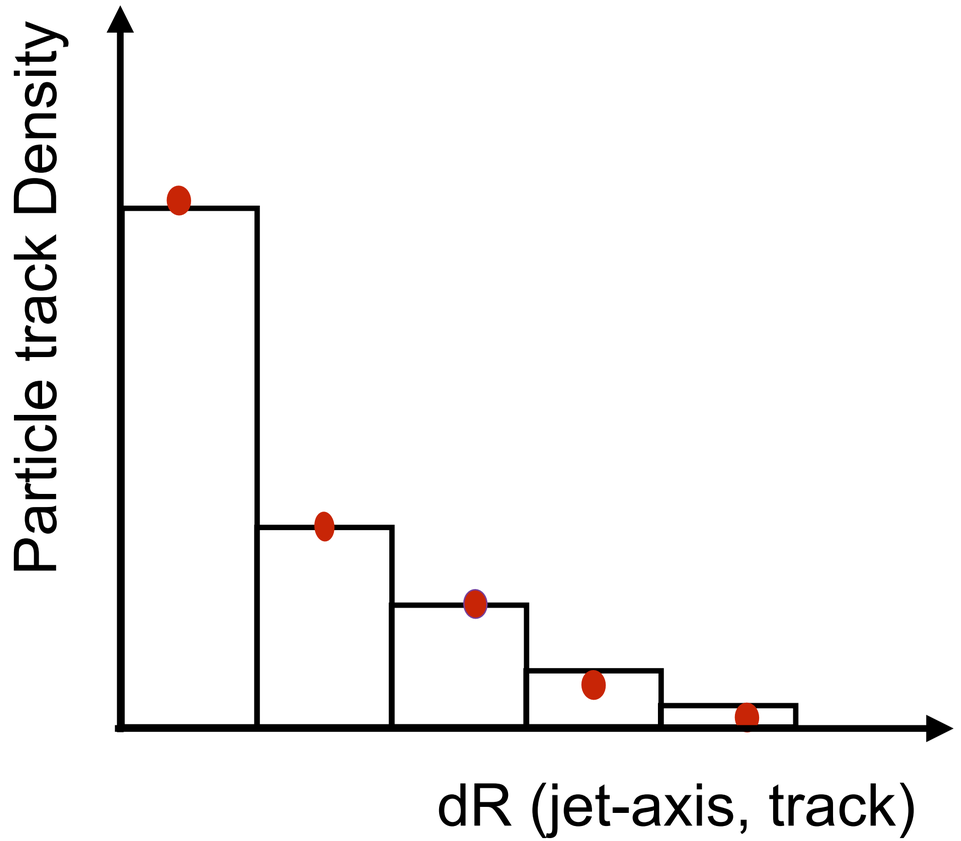
PYTORCH

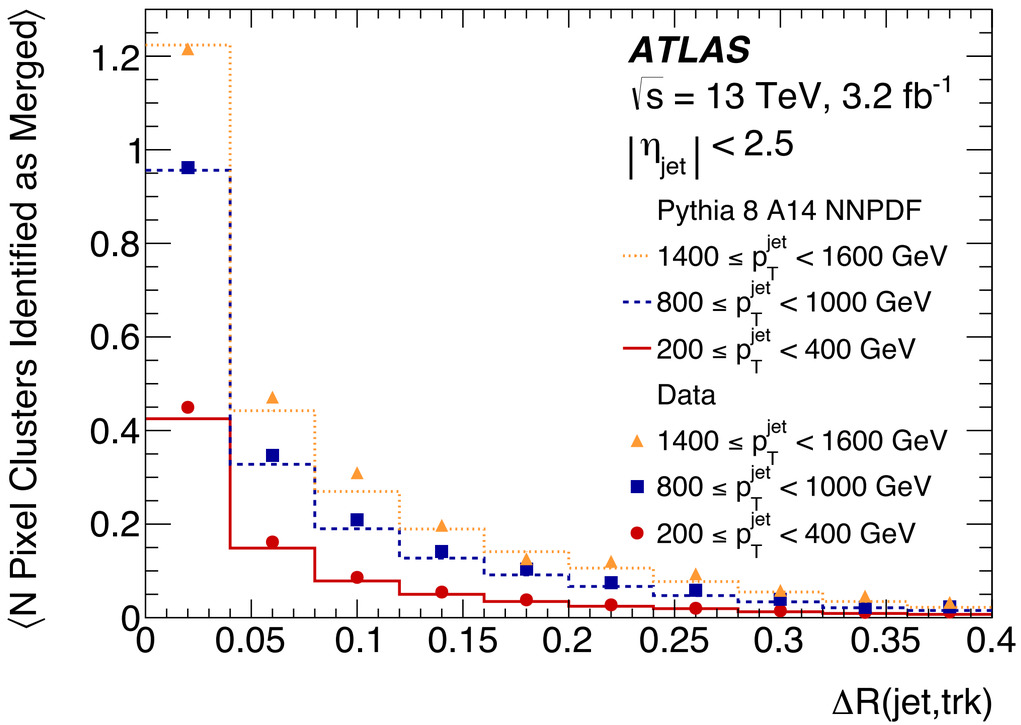
Vivado™ HLS

HLS COMPILER

Mentor®
Catapult
Coming soon

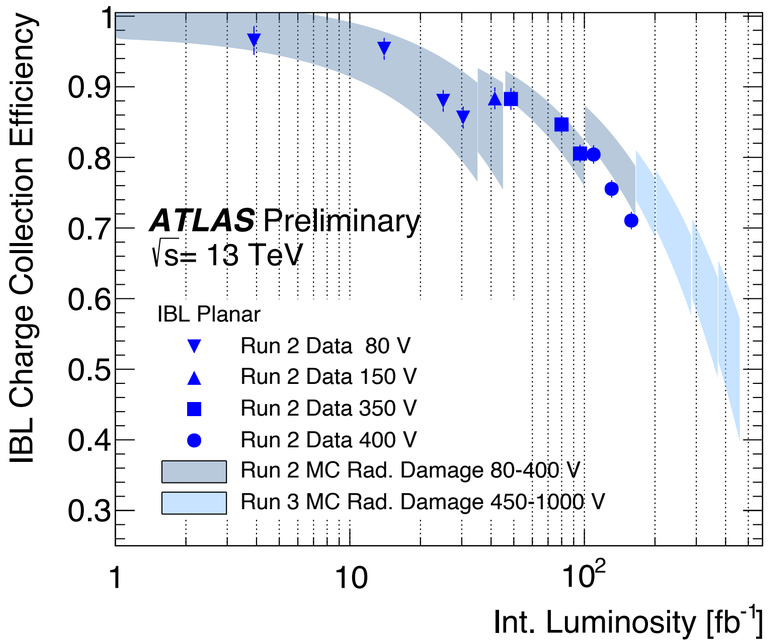
Particle density: dense environment

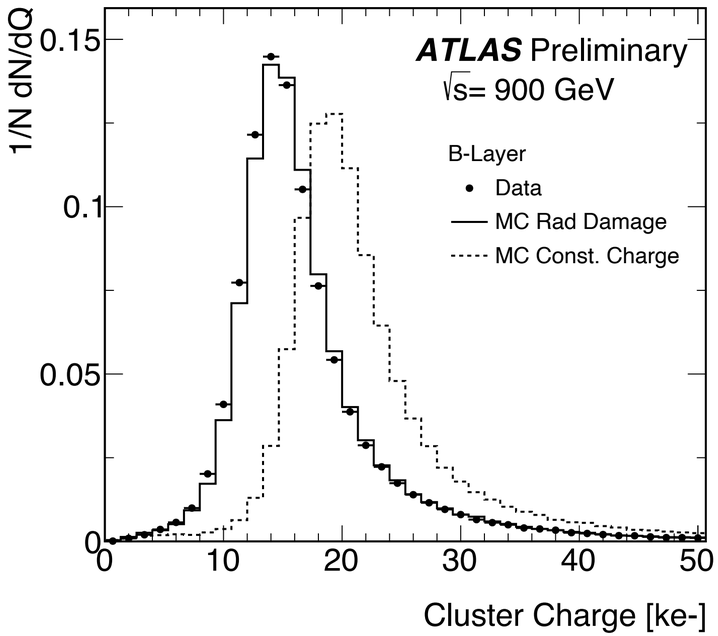


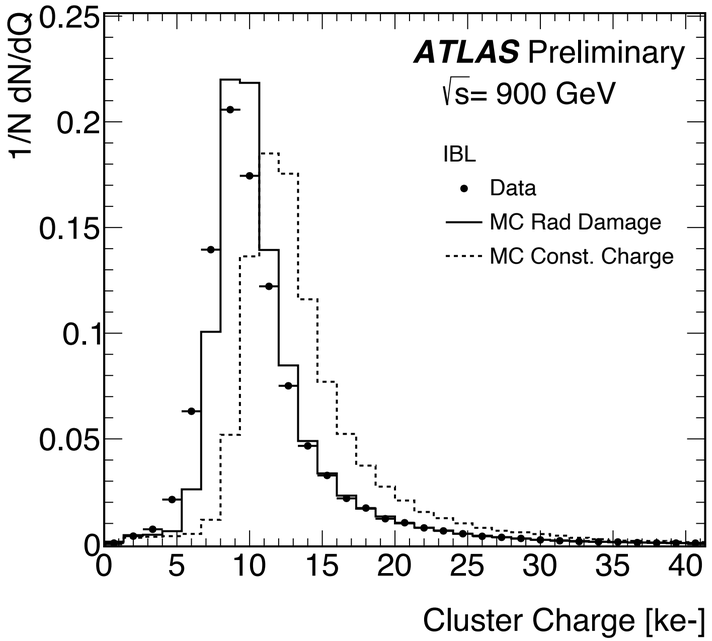


Radiation Damage: IBL and B-layer charge loss

- Charge collection is reduced by ~30% at the end of Run-2
- IBL & B Layer charge loss is expected to be 40-50% by the end of Run 3
- Radiation damage MC represents a unique tool to understand these effects





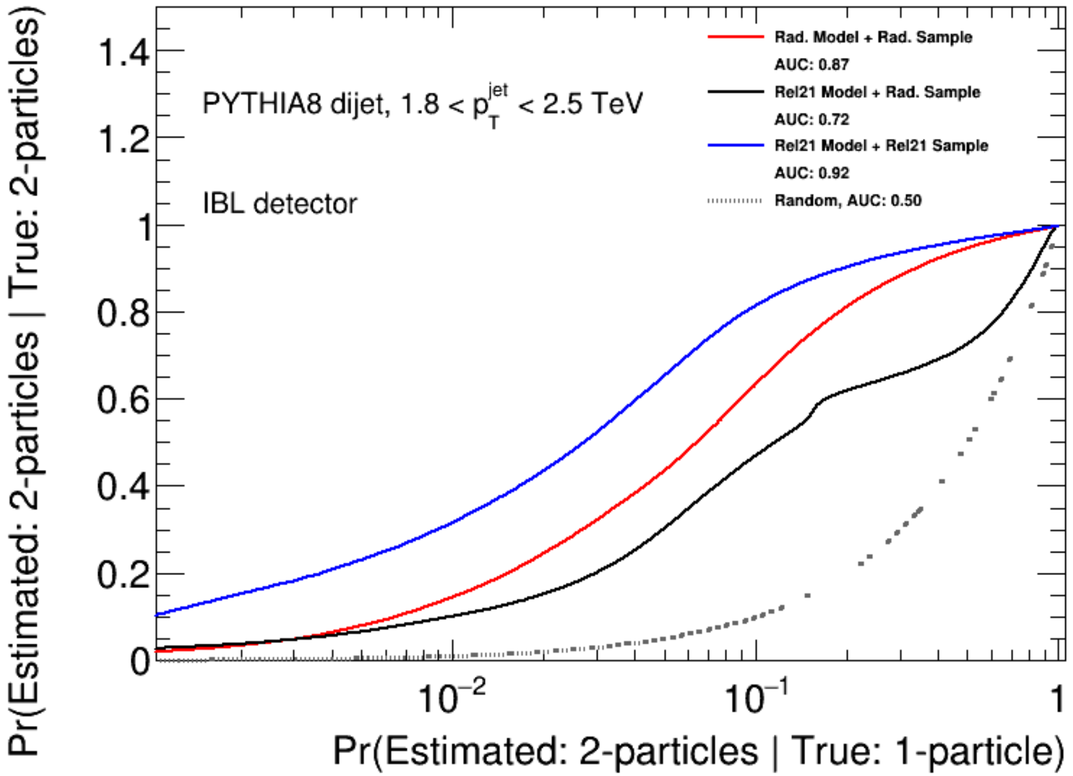


ATL-PHYS-PUB-2022-033

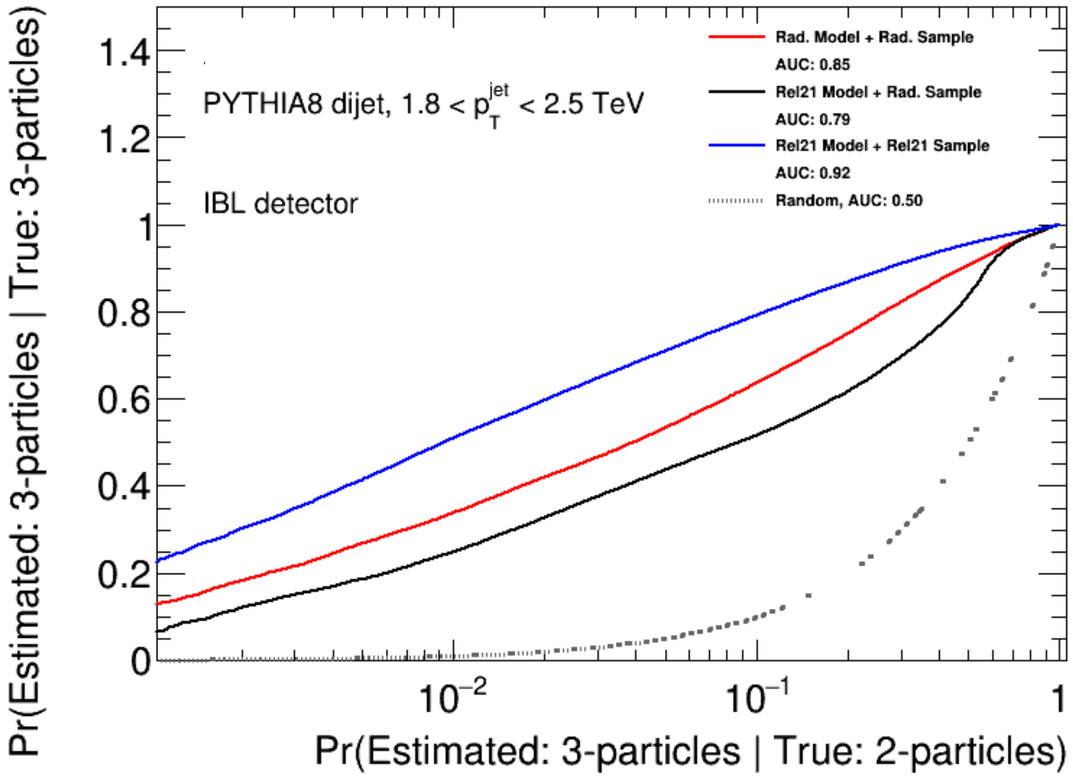
Radiation Damage-aware number NN

- Number NN is retrained with Pythia di-jet (JZ7) MC including radiation damage effect
- Performance improved after retraining the number NN

2 vs 1 ROC Curve

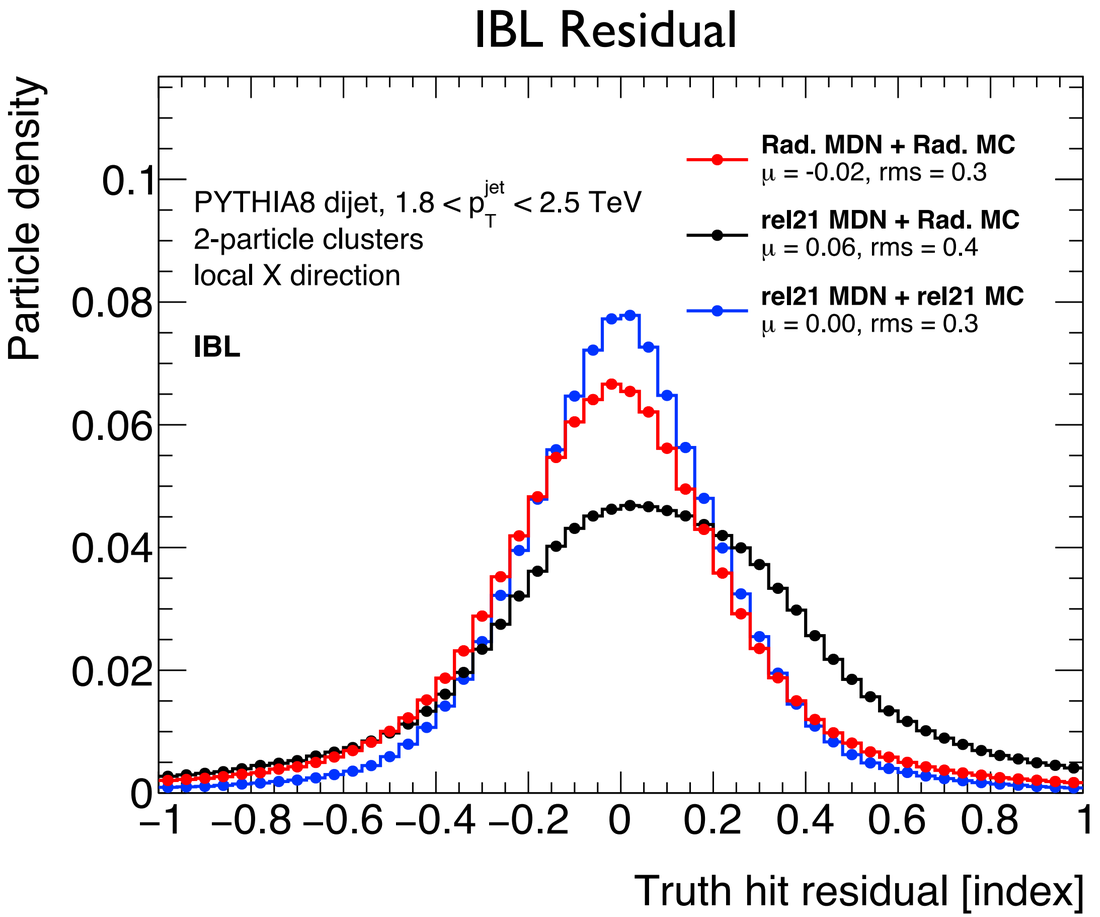


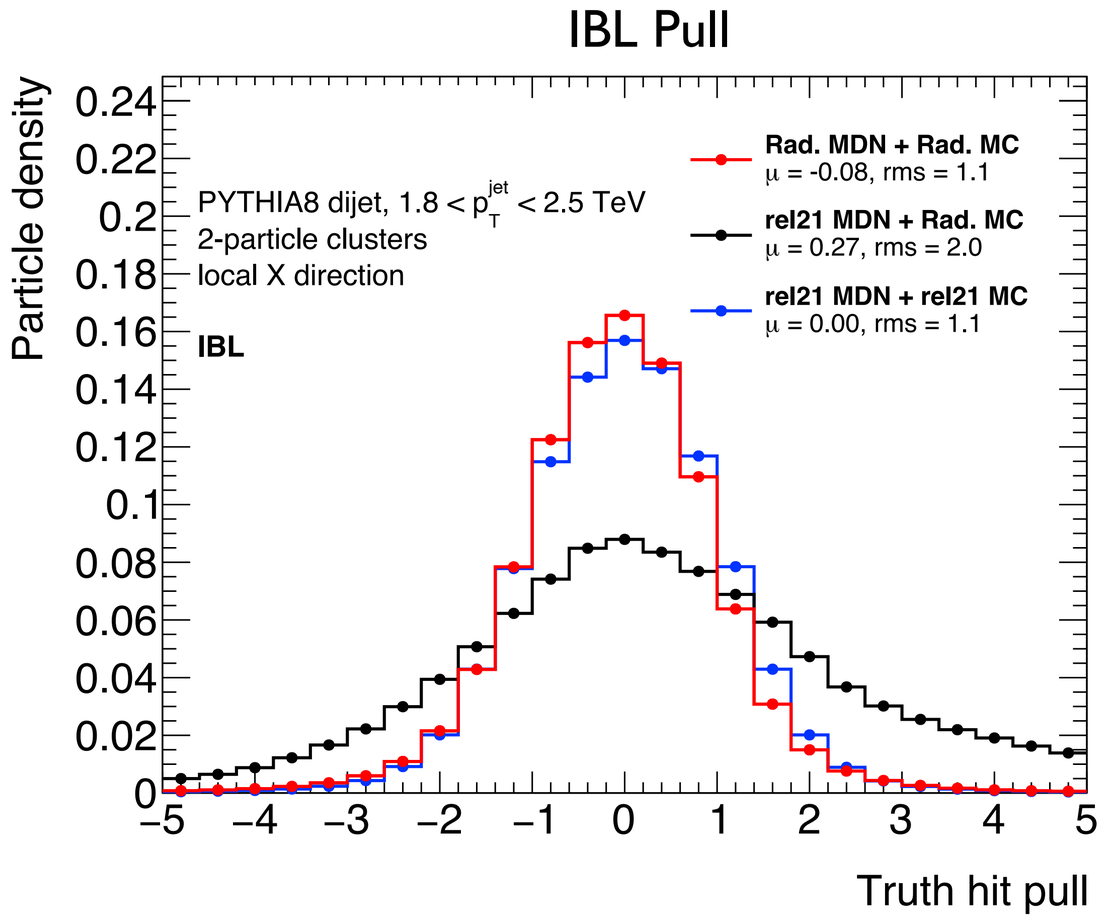
3 vs 2 ROC Curve



Radiation Damage-aware MDN

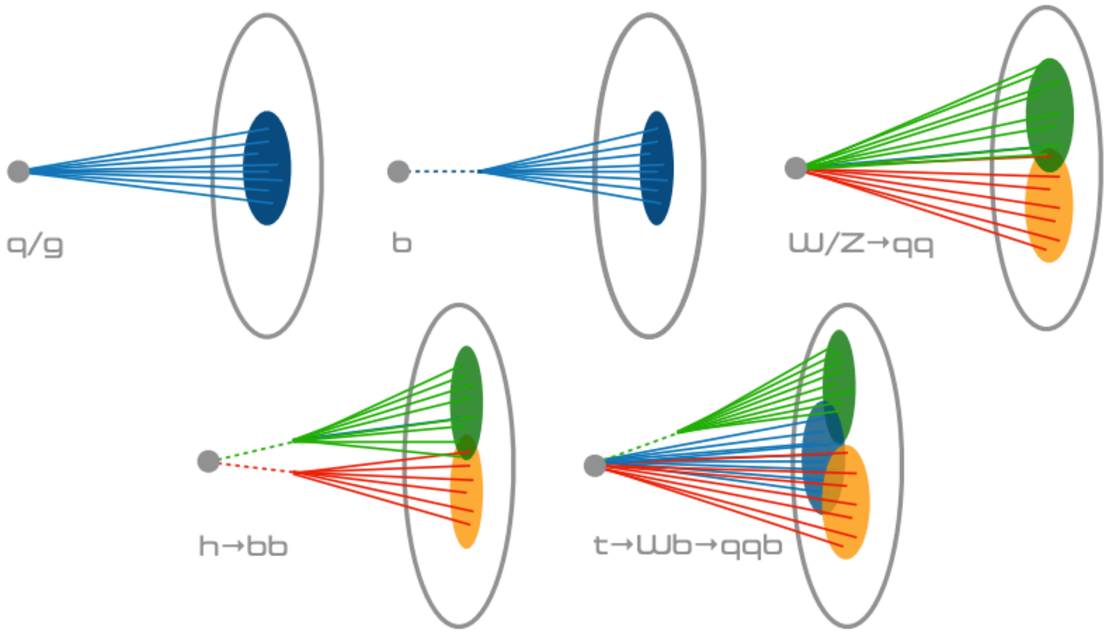
- All three MDNs are retrained with Pythia di-jet (JZ7) MC including radiation damage effects
- Retrained model is closer to the rel21 model (trained with standard MC)



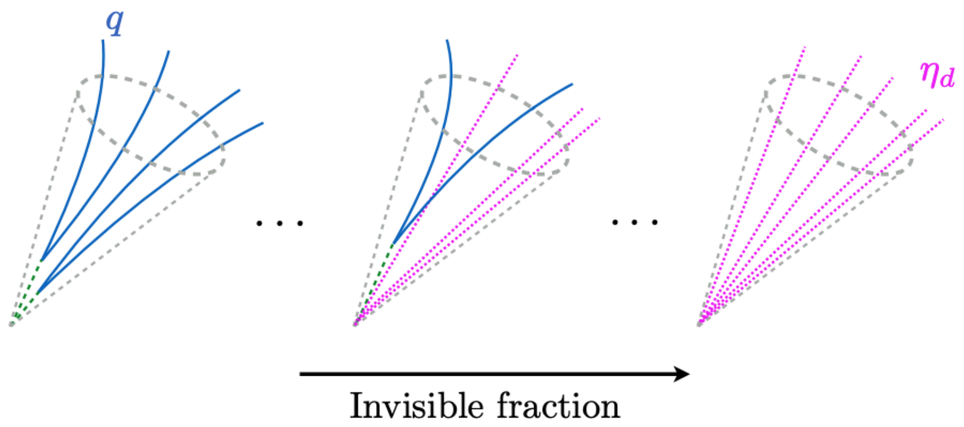


Structures within jets

QCD jets
Could have different substructure



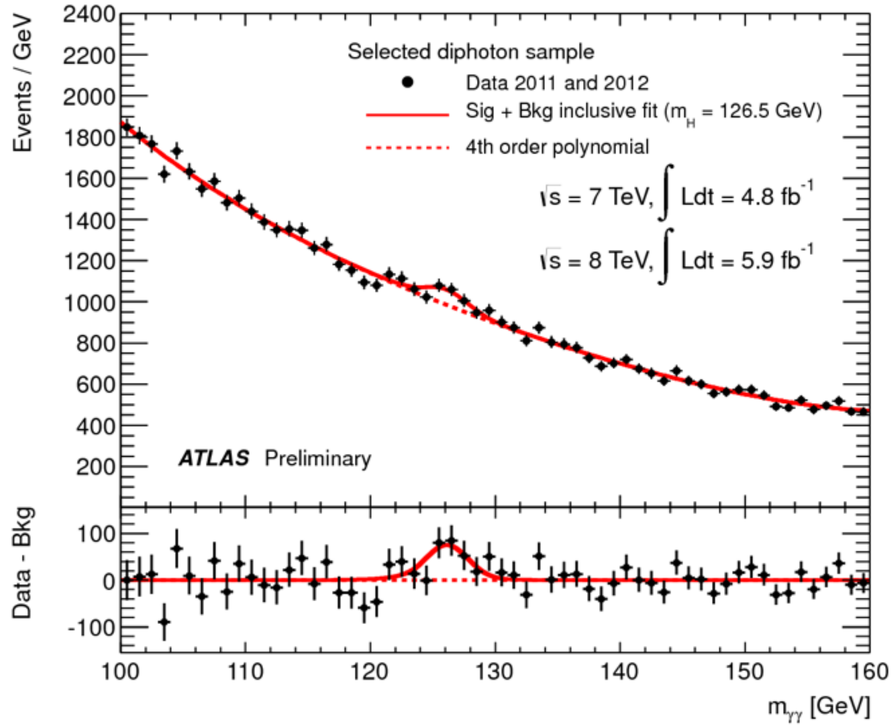
Semi-visible jets
Contains dark-hadrons

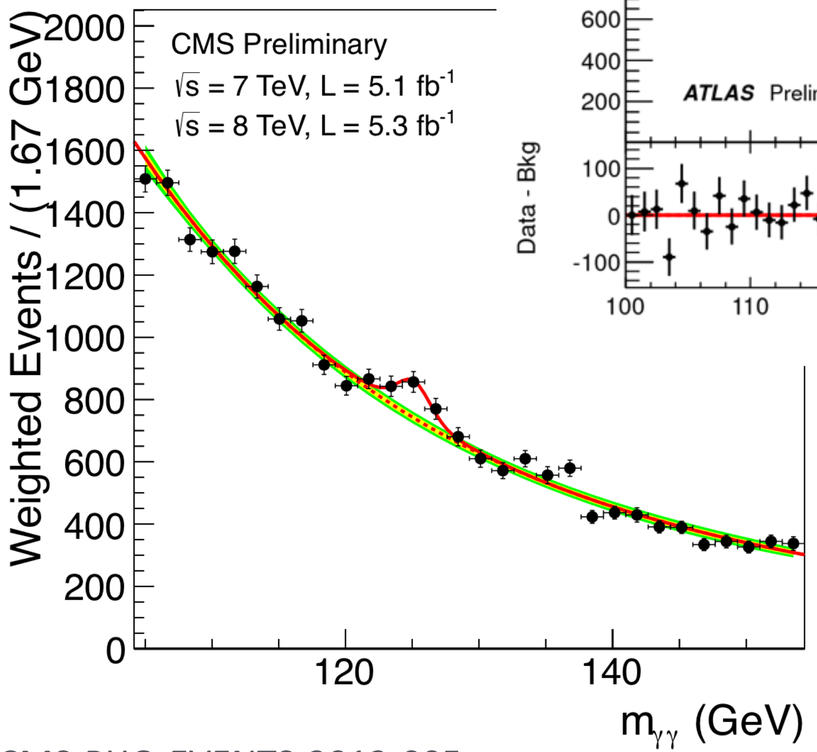


July 4th, 2012: CERN Auditorium

Major Breakthrough at the LHC: Higgs Boson

ATLAS-PHOTO-2018-020





CMS-PHO-EVENTS-2012-005

The Higgs Boson: Not Just a Triumph for Science

by James Clark Ross
6 July 2012

Featured image: Peter Higgs, in tears, on 4th July 2012, the day the discovery of the Higgs boson was announced to the world.

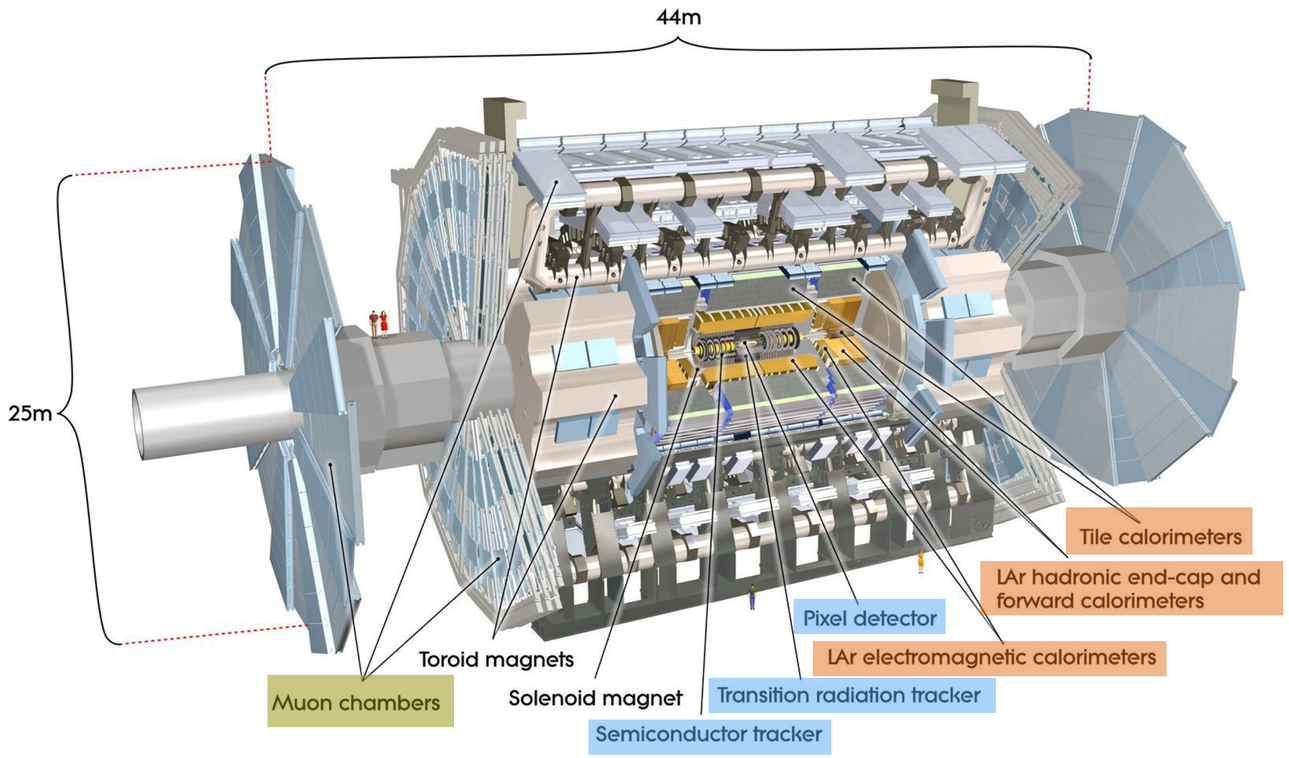
The ATLAS Experiment

General purpose detector

Muon Spectrometer:
Four different detector technology

Calorimeter:
Electromagnetic (Liquid Argon), Hadronic (Liquid Argon (endcap) & Tile (barrel))

Solenoid Magnet: 2.0 T



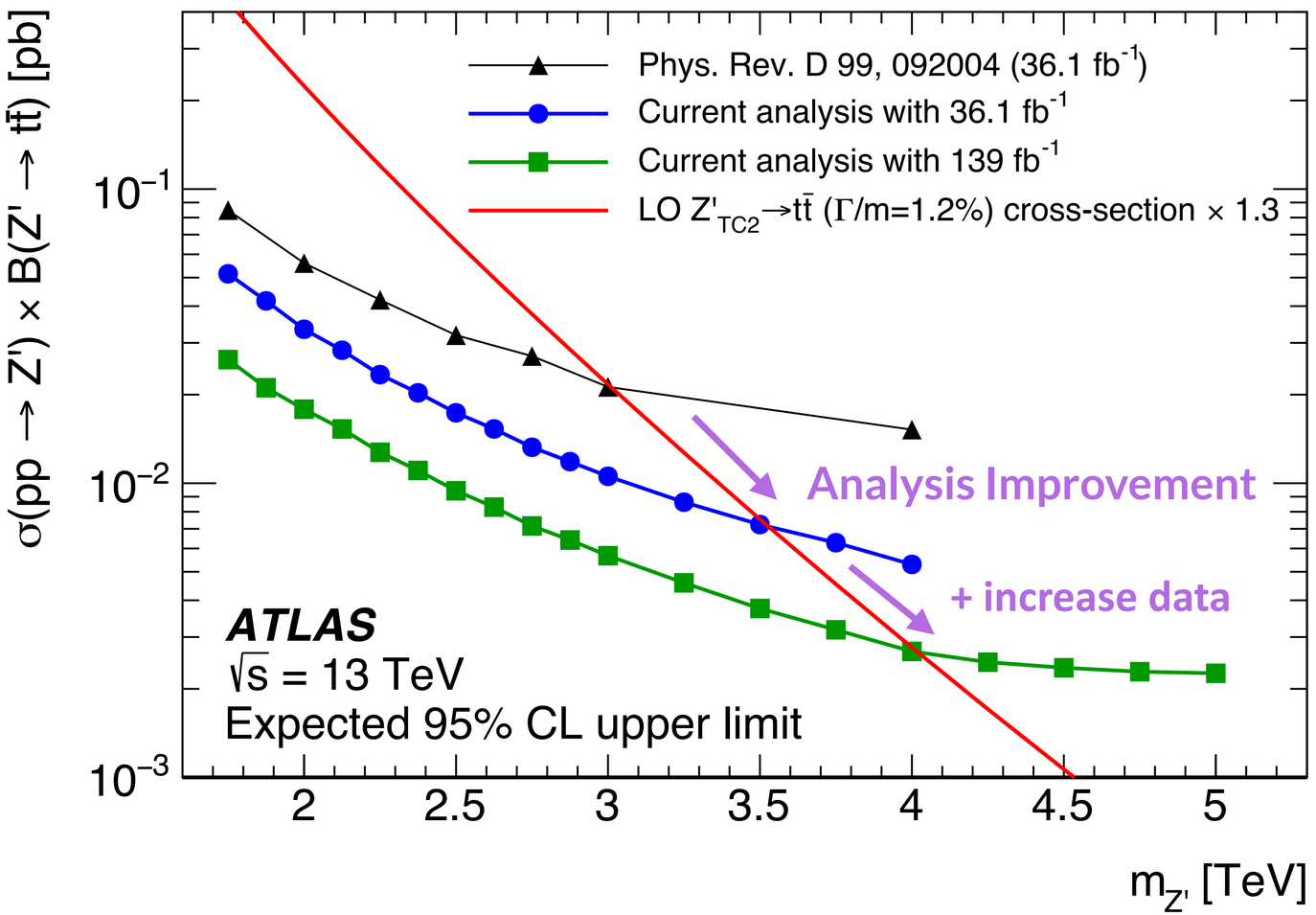
Inner Detector:
Three different detector technology

1. Silicon Pixel
2. Silicon Strip
3. Straw Tubes: Transition Radiation Tracker (TRT)


LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
LHC 27 km

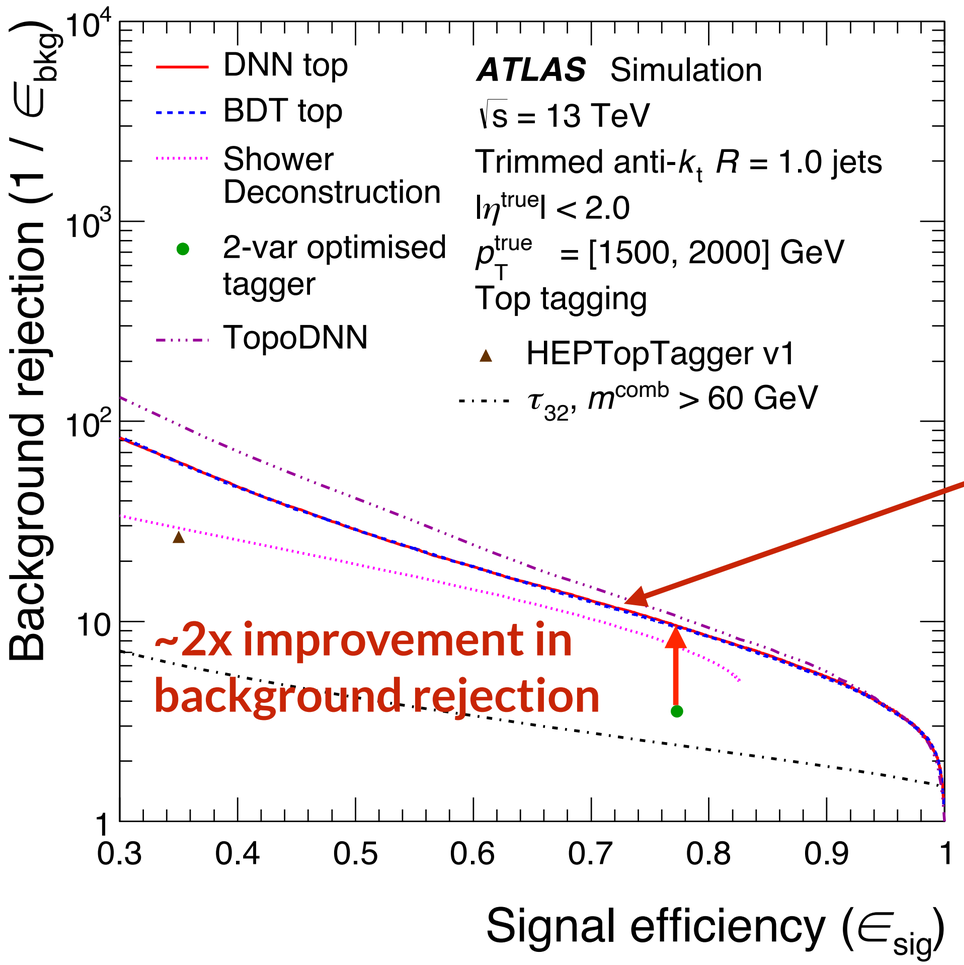

2020/2/5 13:51

Significant improvements in the limit



ML-driven improvements in top-tagging

Deep Neural Network (DNN)-based top-tagger improved the analysis



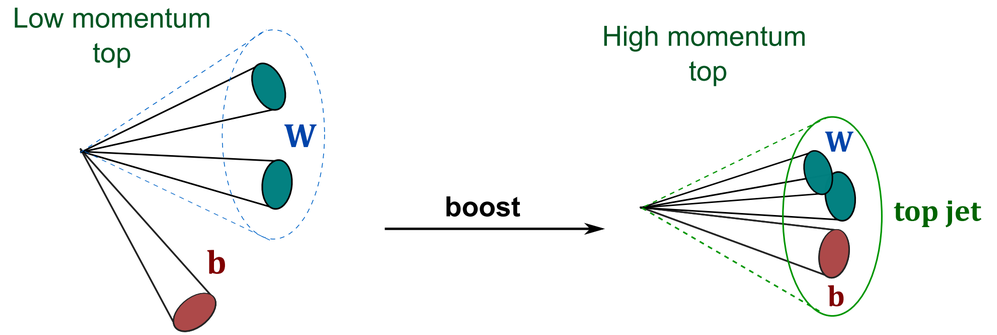
DNN Top-tagger

Old tagger (tau_{32}, mass)

Used in the partial Run-2 analysis

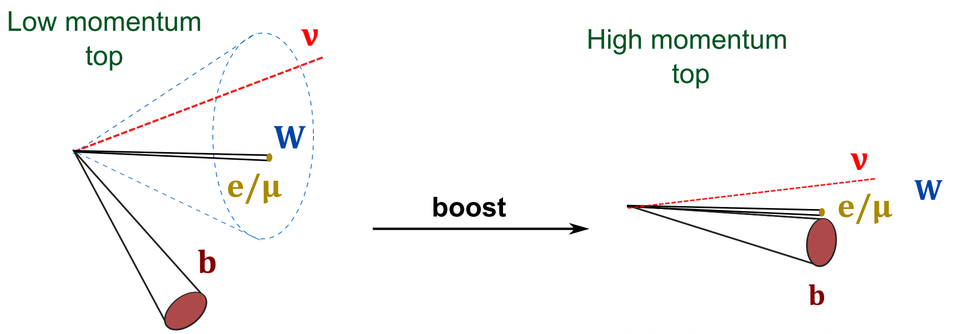
High pT top quark decay

Decay products of a high momentum top quark get collimated along the top quark momentum



topology of **resolved** analysis

topology of **boosted** analysis



topology of **resolved** analysis

topology of **boosted** analysis

Goal is to identify signal jets (coming from top-quark) from the other QCD jets (background)

Resonant Anomaly Detection Search

Assumption:

Signal is localized at least in one of the feature spaces (x)

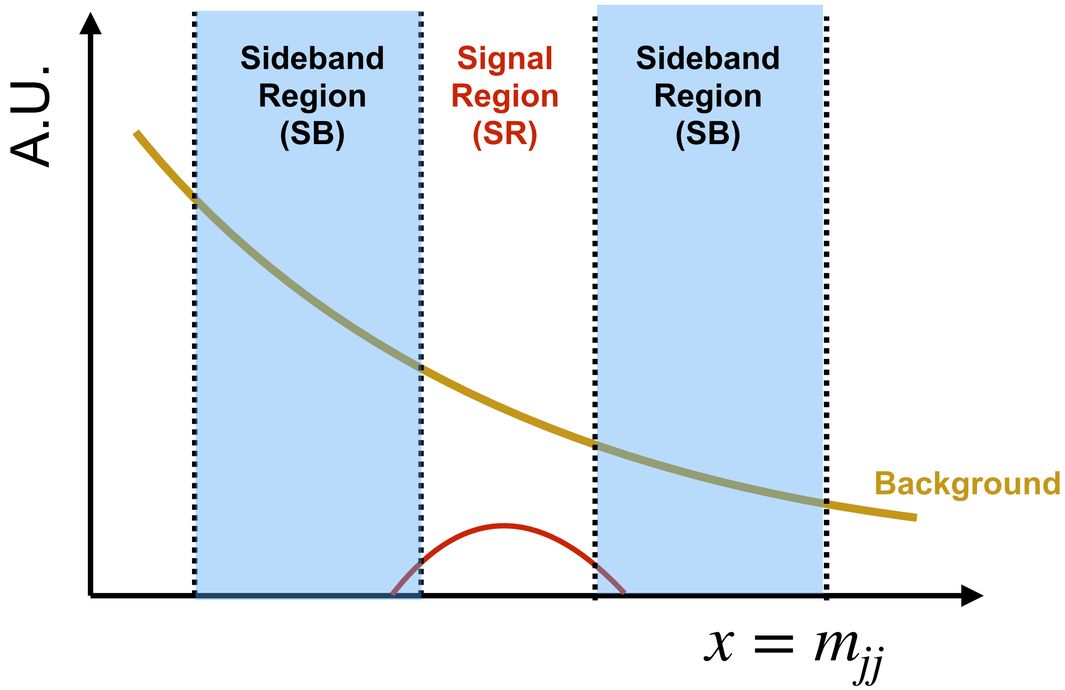
$p_{\text{signal}}(x)/p_{\text{background}}(x)$ is high

Often assumed to be some combined invariant mass

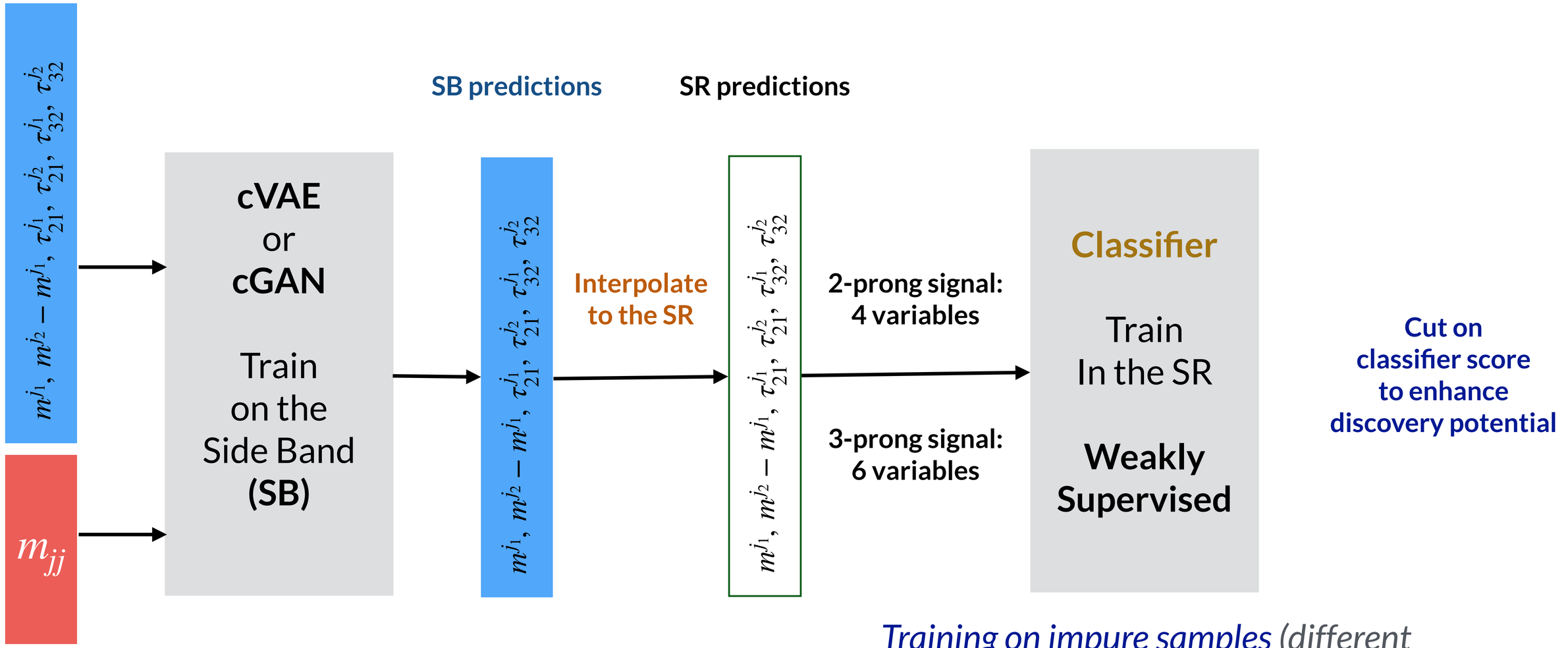
- Expected to appear as a bump

Standard BumpHunt-like search

- Define the signal and control regions using variable x
- Use side-bands to learn the background distribution in the signal region
- Compare predicted background with data in the signal region



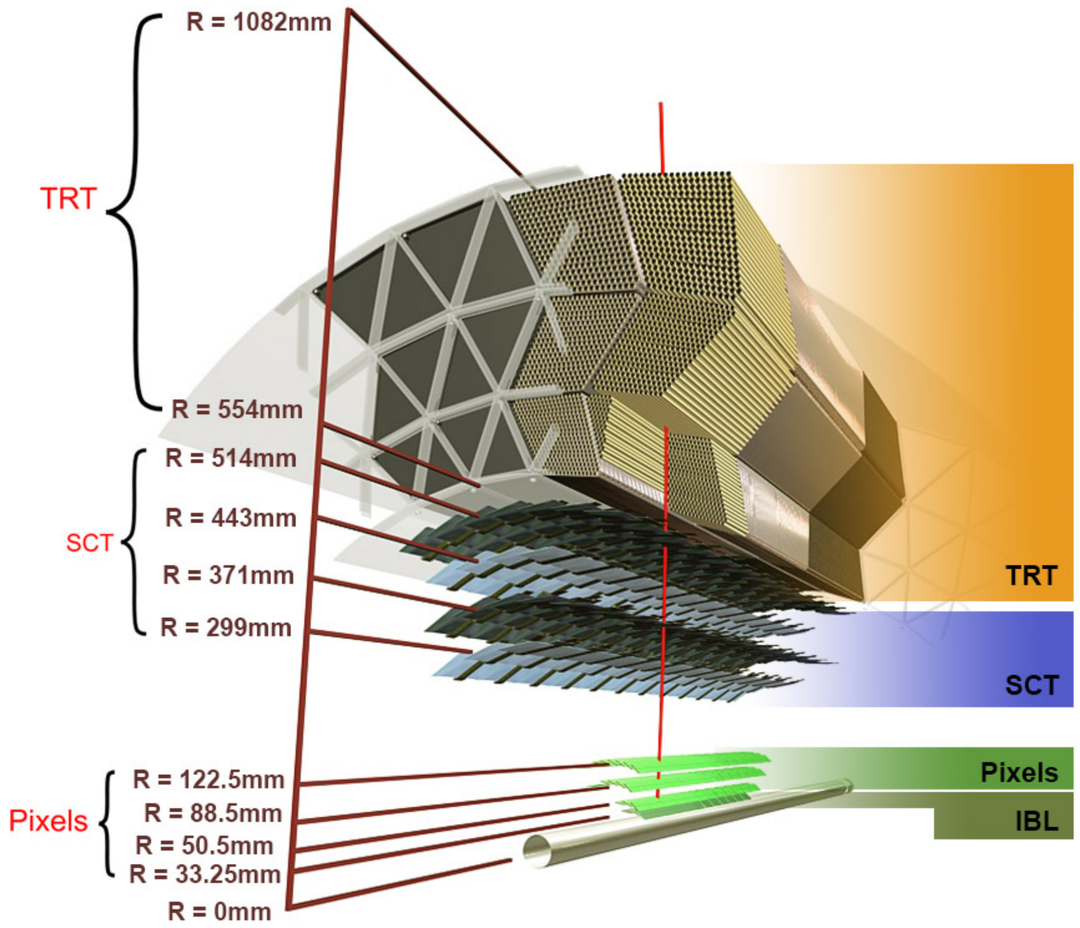
Let's use Generative Models



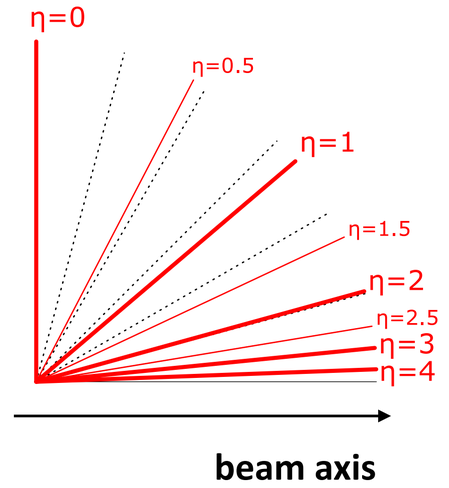
Training on impure samples (different admixture of S and B) is asymptotically equivalent of training on pure samples

ATLAS Inner Detector (ID)

- The total ID coverage |η| < 2.5 with radius of 1.1m
- ID is inside 2T solenoid field



$$\eta = -\ln\tan(\frac{\theta}{2})$$



TRT:
Xe (Ar) filled straw tubes
(= 4 mm), O(30) crossed straws per track.

SCT:
4 layers of double sided strip
9 disks, 80 μm× ~ 6 cm,

Pixel: 3 layers, 3 disks, 50 × 400μm^2

IBL: 1 barrel layer, 50 × 250μm^2