Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Smart pixels with data reduction at source: possible applications for linear e^+e^- colliders

Jennet Dickinson and Lindsey Gray

International Workshop on Future Linear Colliders

May 17, 2023

jennetd@fnal.gov

The future of particle physics at colliders

- Detectors with high precision 😀

  Smaller pixels, timing capability

- New & upgraded colliders 😀

  Higher luminosity, higher energy

The future of particle physics at colliders

- Detectors with high precision 😀 ⟶ more readout channels 😕

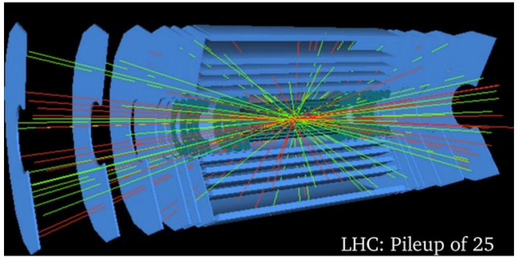
  Smaller pixels, timing capability

- New & upgraded colliders 😀 ⟶ more hits, esp. backgrounds 😕

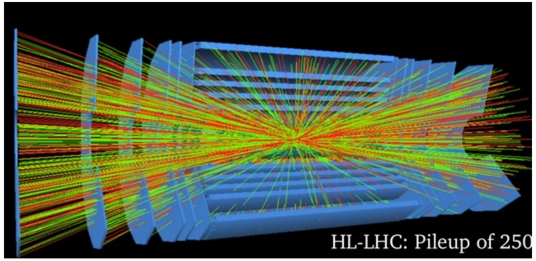
  Higher luminosity, higher energy

hh

Higher luminosity ⟶ pileup ⟶ limitations on trigger & amount of recorded data



⟶



Solution!
smart pixels

physics-motivated data reduction with AI on-ASIC

The future of particle physics at colliders

- Detectors with high precision 😀 ⟶ more readout channels 😕

  Smaller pixels, timing capability

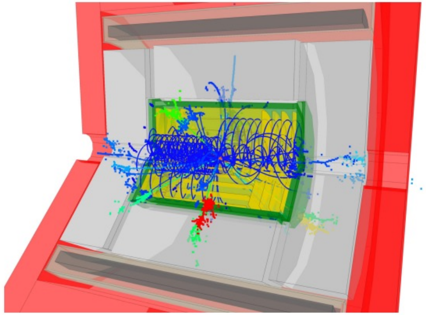
- New & upgraded colliders 😀 ⟶ more hits, esp. backgrounds 😕

  Higher luminosity, higher energy

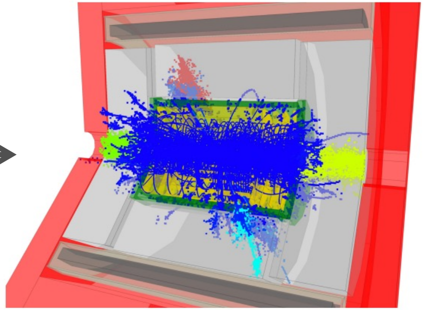
hh

ee

Higher energy ⟶ more beam background ⟶ limitations on detector design



⟶



Solution?
smart pixels

physics-motivated data reduction with AI on-ASIC

Fermilab

Outline

- Smart pixels in the context of the HL-LHC
  - Filtering
  - Featurization
- Applications at linear e^+e^- colliders

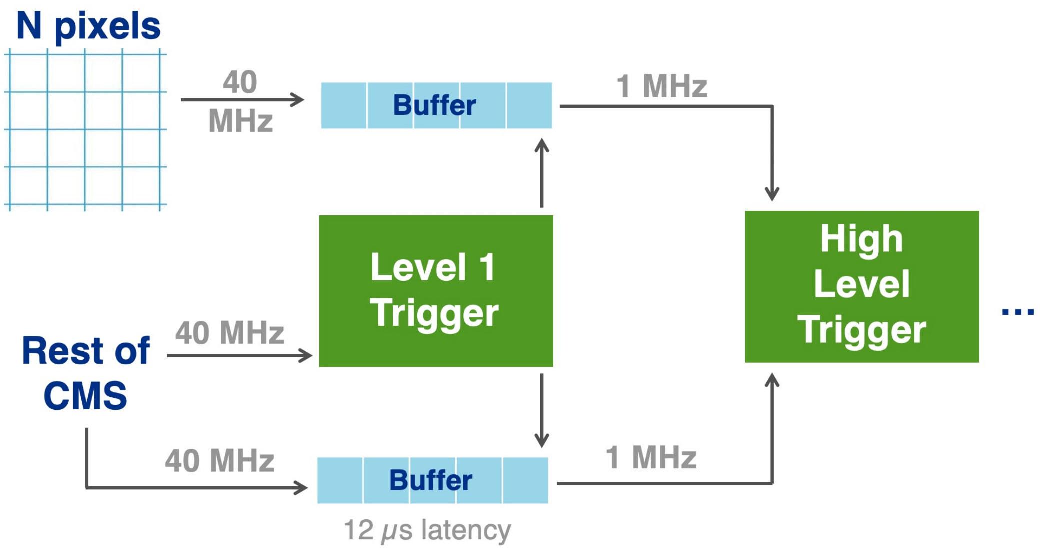
Pixel readout chain: CMS at HL-LHC

- Detector is an array of N pixels

  100 x 25 μm pitch

  100 μm thick sensor

- Pixel data sits in buffer until L1 decision is made
- Passed to HLT at 1 MHz



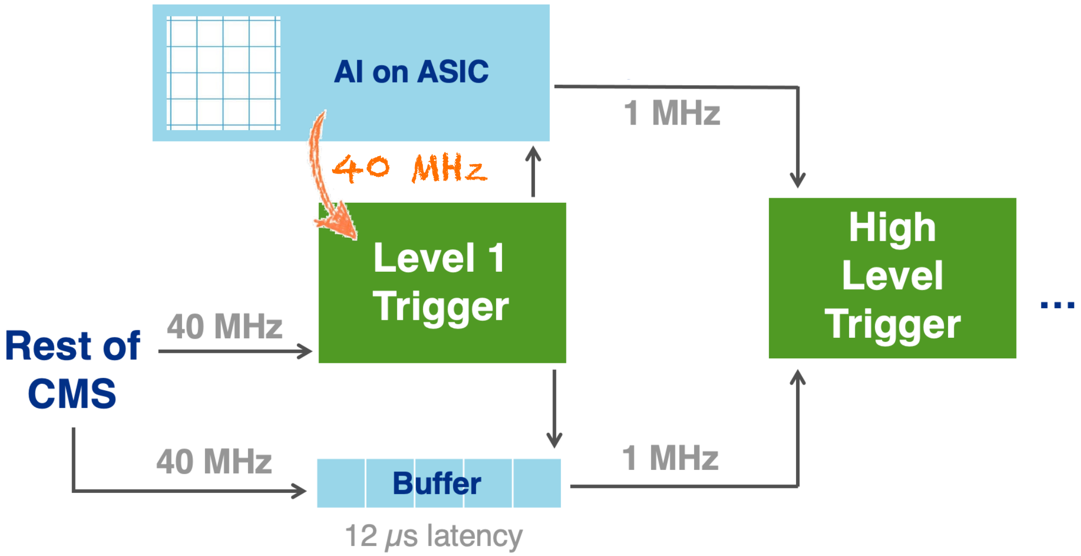
Pixel readout chain: our futuristic CMS detector

- Detector is an array of **4N** pixels

  50 x 12.5 μm pitch

  100 μm thick sensor

- **Pixel data is passed to L1 trigger at 40 MHz**
- Passed to HLT at 1 MHz



Can we use smart pixels to transfer 4-160x more data?

Charged particle signatures in our futuristic detector

- State-of-the-art dataset for developing algorithms for implementation on-ASIC (link)

  Initial conditions = fitted track params from CMS Run 2 data

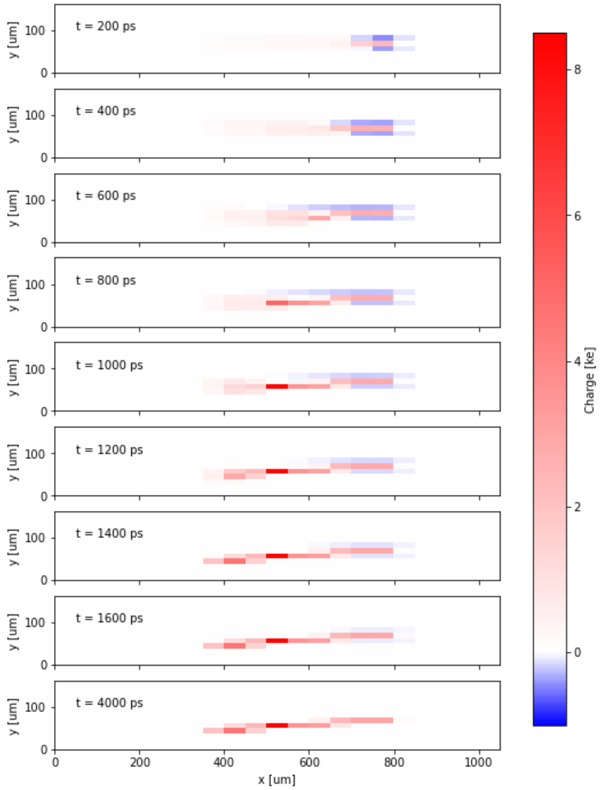
- Simulated pions traversing a 21x13 array of pixels

  50x12.5 μm pitch, 100 μm thickness

  Located at radius of 30 mm

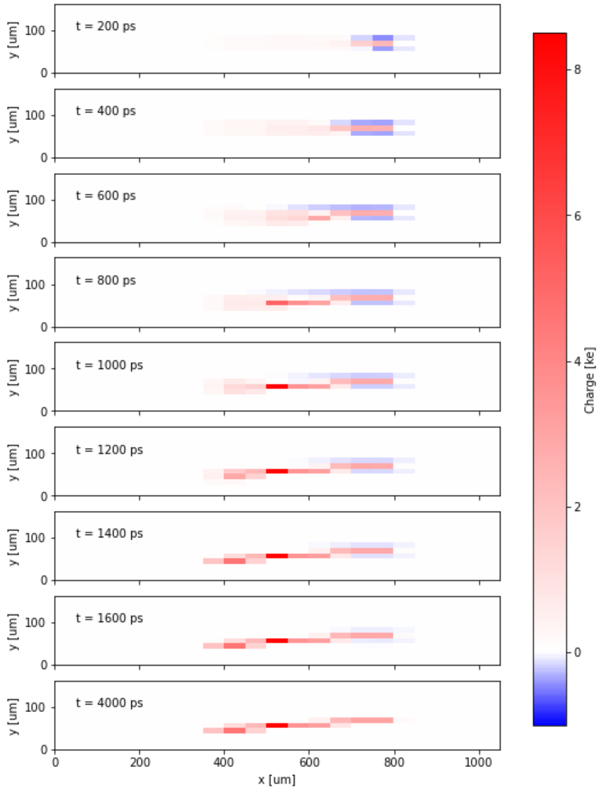
  3.8 T magnetic field

  20 time steps of 200 picoseconds (4 ns total)



Charged particle signatures in our futuristic detector

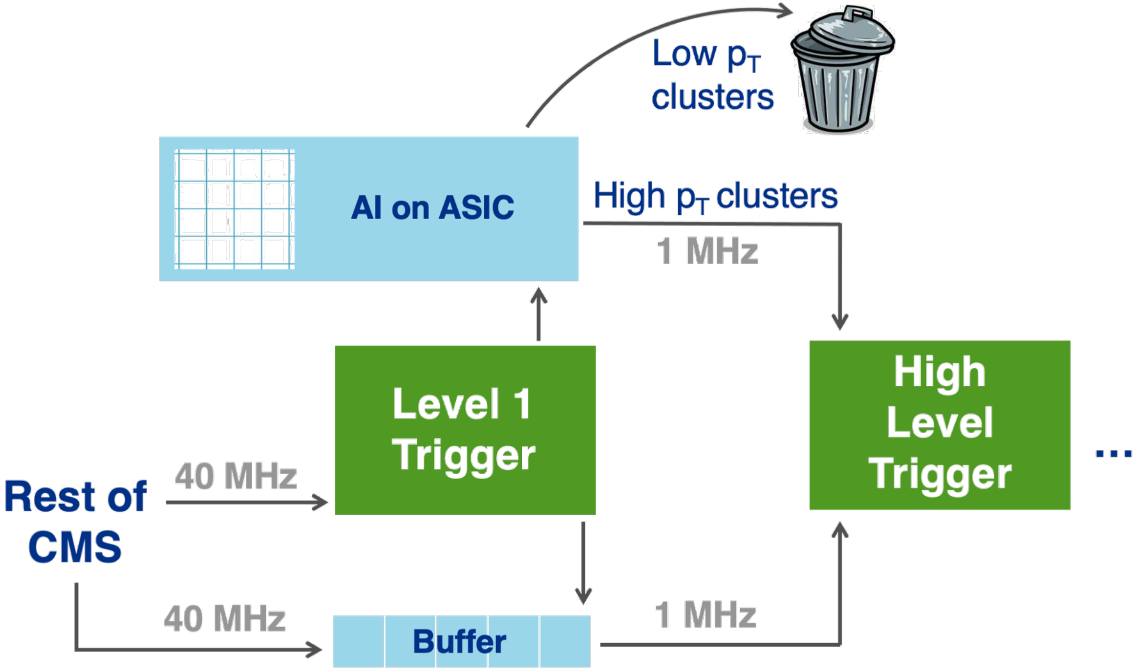
- State-of-the-art dataset for developing algorithms for implementation on-ASIC (link)
  - Initial conditions = fitted track params from CMS Run 2 data
- Simulated pions traversing a 21x13 array of pixels
  - 50x12.5 μm pitch, 100 μm thickness
  - Located at radius of 30 mm
  - 3.8 T magnetic field
  - 20 time steps of 200 picoseconds (4 ns total)
- For e^+e^- :
  - Simulation with electrons needed
  - Is signal collected over 4 ns fast enough?



Data filtering on-ASIC

- Select and read out clusters created by particles with high transverse momentum (p_T)
- More than 95% of CMS pixel hits have $p_T < 2$ GeV

Potential for large reduction at HL-LHC



Fermilab

Classification based on particle p_T

- p_T ~ radius of curvature, correlated with

  Magnetic field strength (B)

  Position of the hit in the bending direction (y_0)

  Angle in the bending plane of B (β)

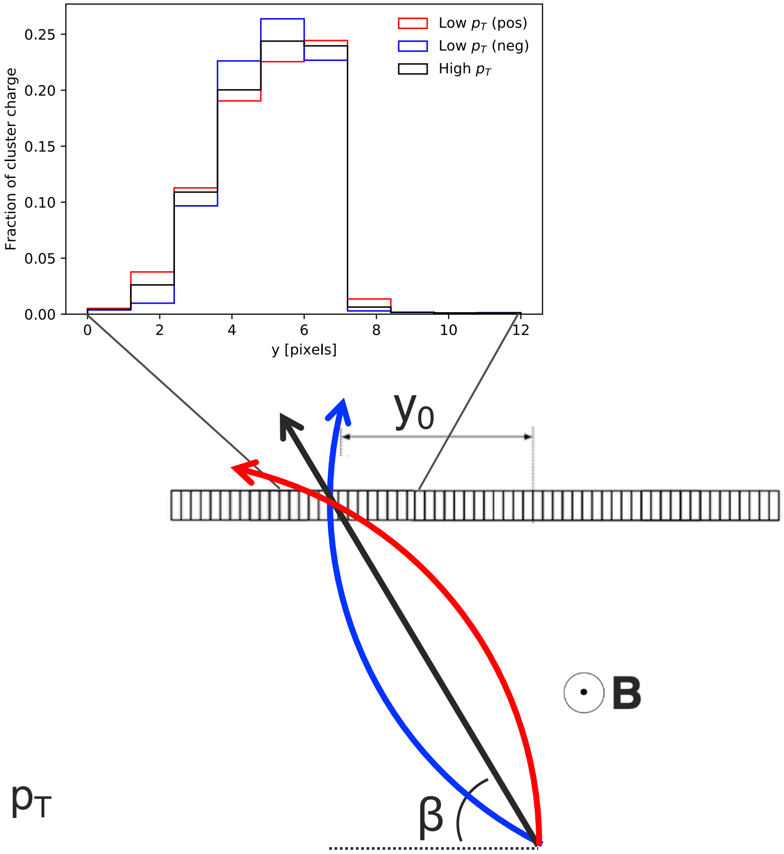
  Sign of the charge

- Train a classifier to select clusters with $p_T > 200$ MeV

  Input data: cluster image projected onto y-axis

- Three classes:

  Low p_T negative charge, low p_T positive charge, high p_T



Fermilab

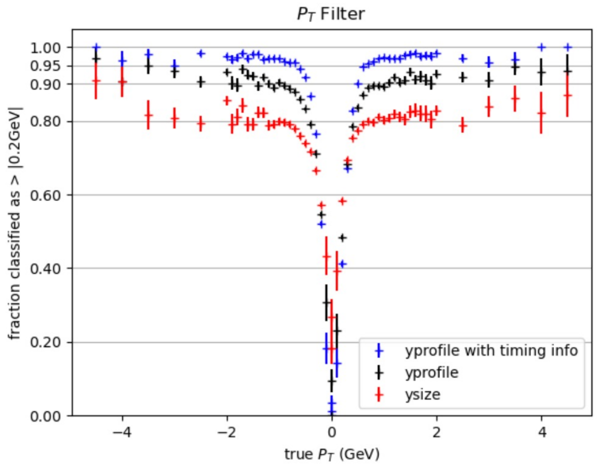
Performance of the DNN p_T filter

- Full precision network:

  Projected cluster size only (1)

  Projected cluster shape (2) (selected as baseline)

  Timing information (3) promises 5-10% efficiency gain



Model	Data Reduction	False Positives	False Negatives
	How much do we discard?	How much of what we keep is $p_T < 2$ GeV?	How much of what we discard is $p_T > 2$ GeV?
Model 1	77.9 %	13.4%	20.7 %
Model 2	65.8 %	10.9%	13.4 %
Model 3	57.4 %	8.9 %	8.1%

Model 4: Spiking neural network is a work in progress

Fermilab

Implementation in 28nm CMOS

- Network quantization

  Input charge distribution binned for 2 bit ADC

  Quantization-aware training in Tensorflow/Keras

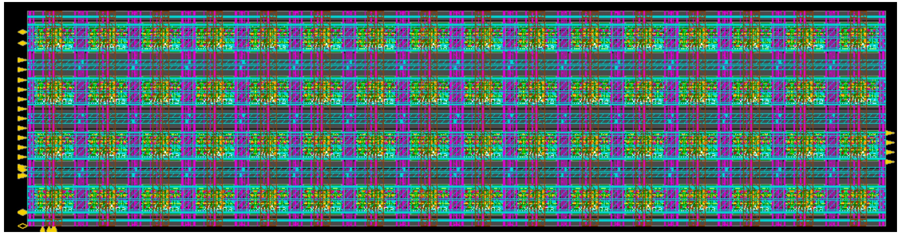
  hls4ml translates model into hardware specification for high-level synthesis (HLS)

  Siemens Catapult HLS to generate RTL implementation

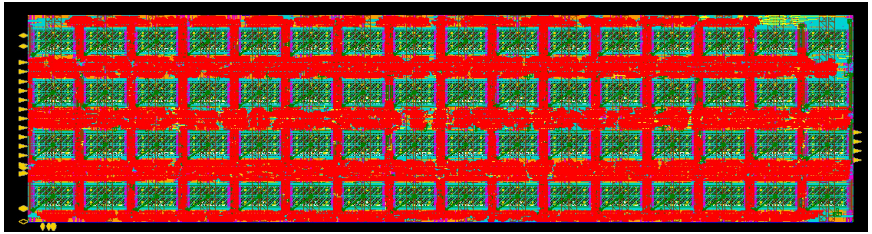
- Fully reprogrammable NN weights

ADC output	Charge interval $[e^-]$
00	< 400
01	$400 - 1600$
10	$1600 - 2400$
11	> 2400

Floorplan with analog pixels with power and bias grid



Red = classifier algorithm



Data featurization on-ASIC

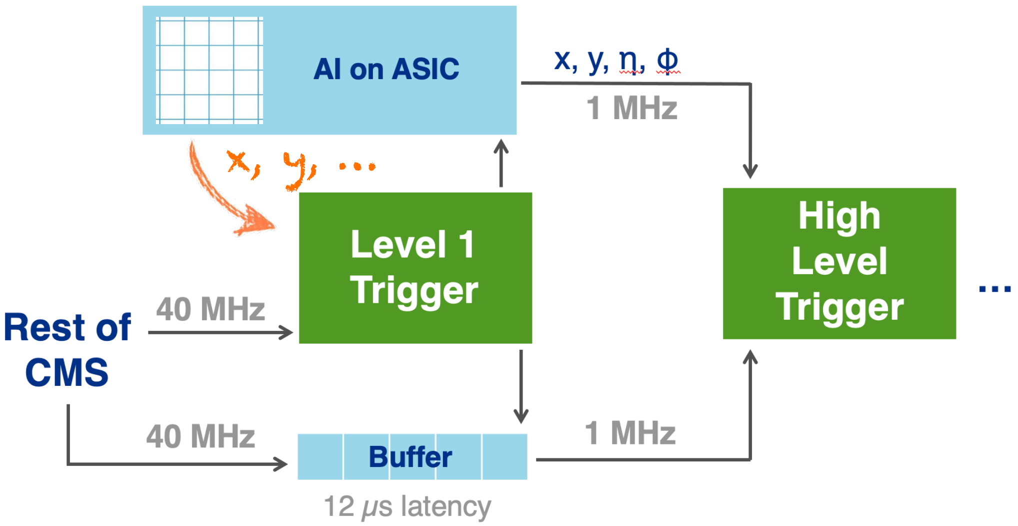
- Train an algorithm to extract properties of the incident particle

  Read this out instead of raw data

- A form of **compression**

  Technically lossy, but fully preserves physics information

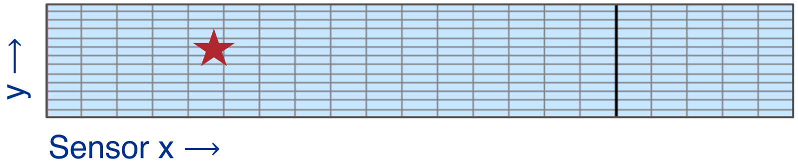
  Size of readout depends on number of clusters, but not on cluster size



- Mixture density network gives us both central value and **meaningful uncertainty** on the measured quantity

  Predicts the parameters of a likelihood distribution

Predicting hit position & uncertainty



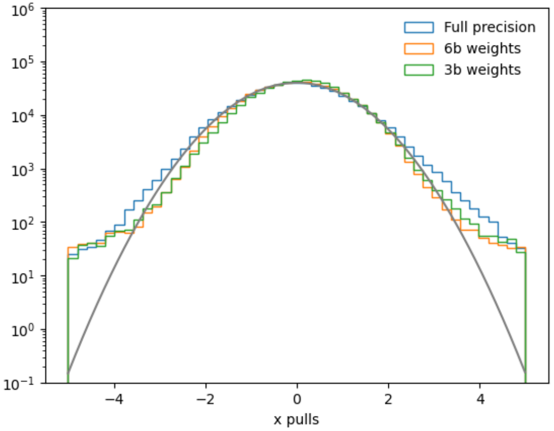
- Gaussian loss function (predict μ, σ)
- Simple networks very performant

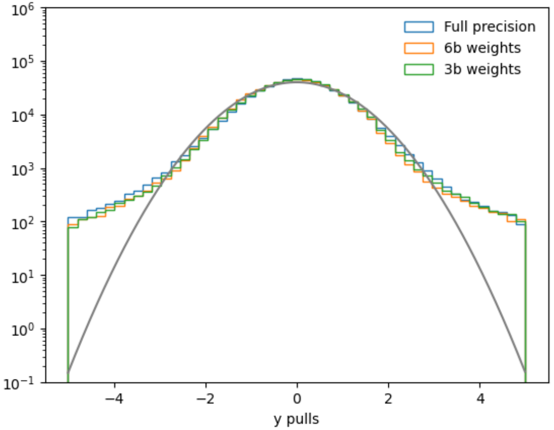
  Single layer NN, 10^6 training clusters

  Quantization aware training with QKeras

- Training input = cluster shape projected onto relevant axis

  Negligible correlation between x, y





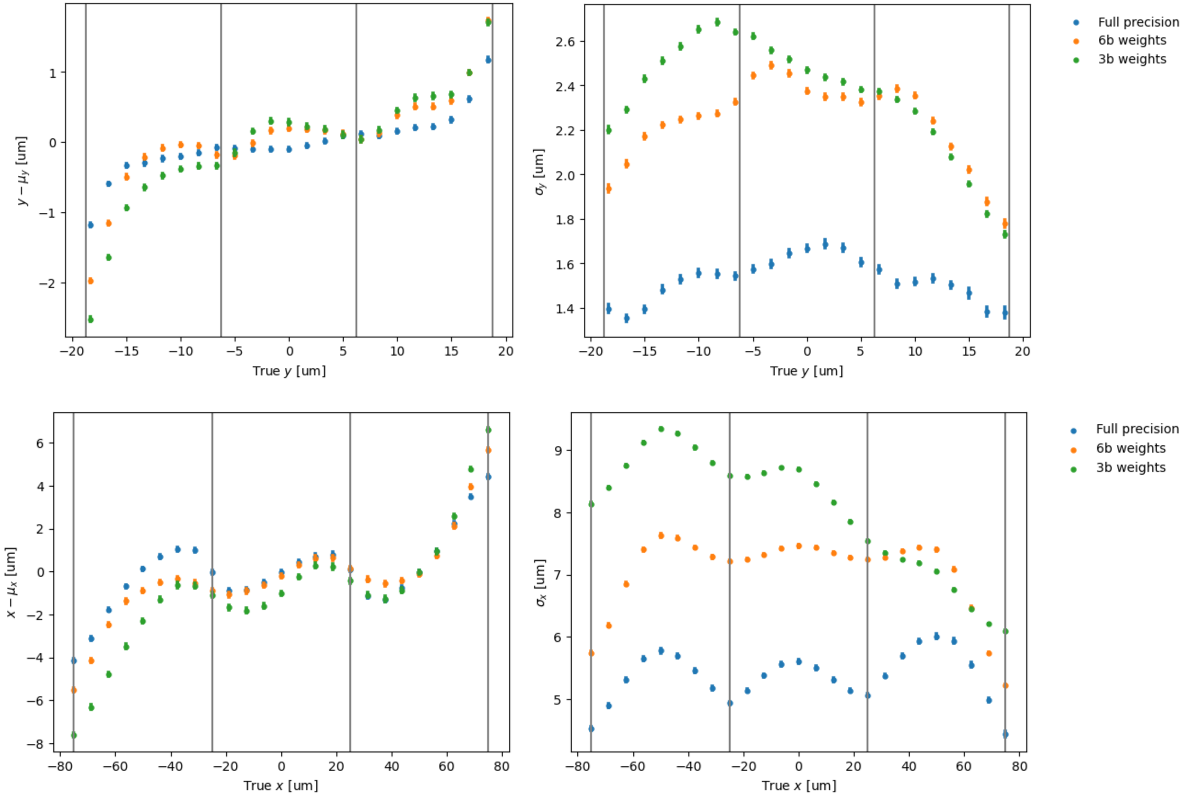
Predicting hit position & uncertainty

- Top row: y, bottom row: x
- Left column: residuals
- Right column: uncertainty

  Low precision weights ⟶ larger uncertainty

- Models shown use cluster shape input only

  Addition of timing information improves performance



Predicting angles & uncertainty

- Beta distribution loss function (params a, b)

  Mode and variance are functions of a, b

- More complex networks compared to x, y

  3 NN layers, 10^6 training clusters

- Training input = cluster shape

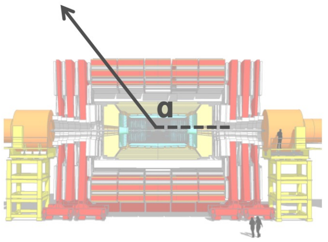
  Full 2D image (orange) and projected (blue)

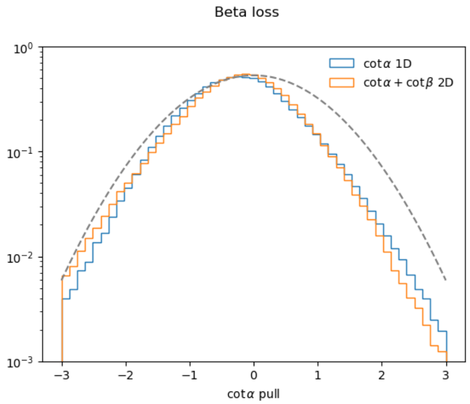
  Negligible correlation between α, β

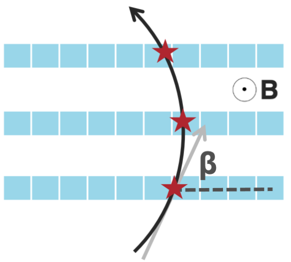
- Predict absolute value of cotangent

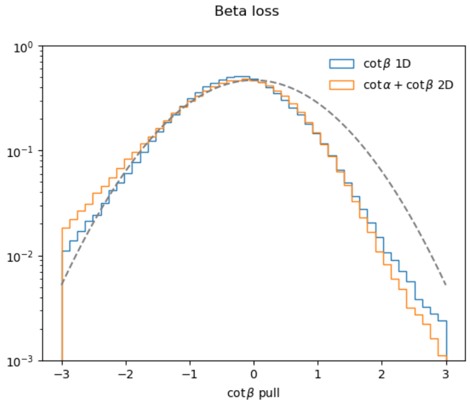
  Approximately linear in projected cluster size

  Sign of the angle requires timing information









Fermilab

Predicting angles & uncertainty

- Single-pixel clusters impact prediction at low cot α

  Pixel pitch is “large” in x

  Best guess = center of the pixel ⟶ some large residuals

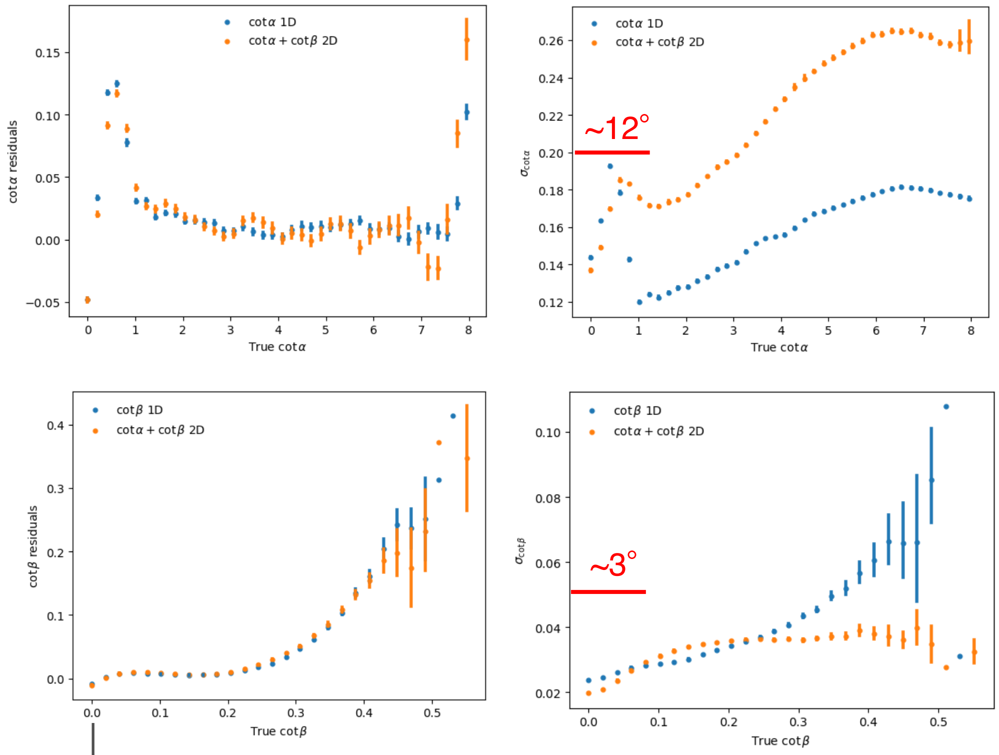
- Dataset is only populated to ±30° from normal in β

  Expect to reduce bias by expanding dataset

- Studies ongoing

  Convolutional NN ⟶ less area

  Quantization to be explored



Normal incidence

Fermilab

Angles & their uncertainties

- More complex final states ⟶ more hits ⟶ more hit combinations for track seeding

  Computationally very expensive and slow 😕

Fermilab

Angles & their uncertainties

- More complex final states ⟶ more hits ⟶ more hit combinations for track seeding

  Computationally very expensive and slow 😕

- Predicted angle + uncertainty gives a cone where you can expect a hit in the next layer, **reducing combinatorics**

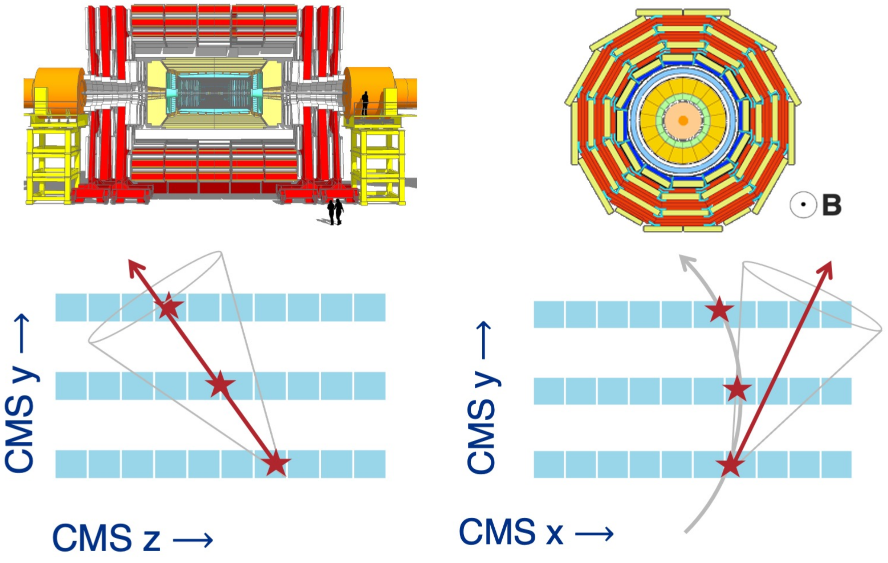
  Small uncertainty ⟶ small cone

  No timing ⟶ |α|, |β| only ⟶ 2 cones

- Fast tracking and vertexing

  Very valuable for hh, e+e- and $\mu\mu$!

  At HL-LHC: makes pixel trigger feasible?

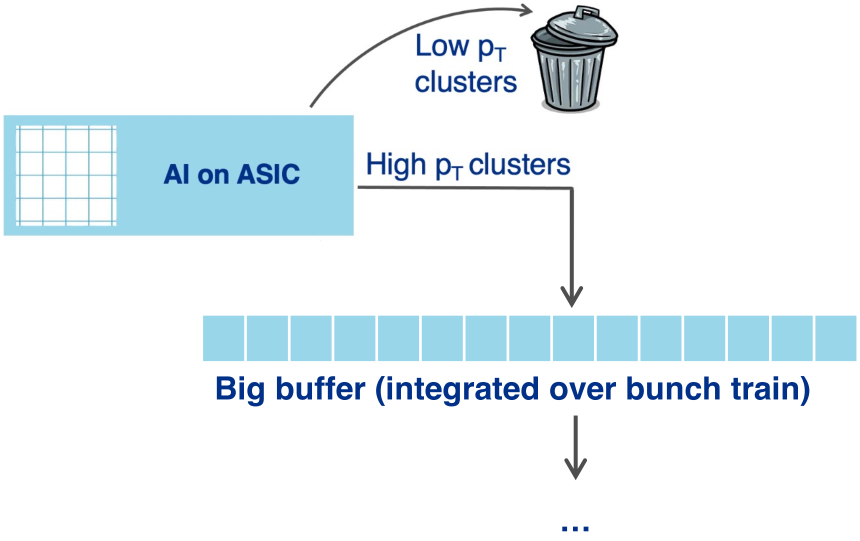


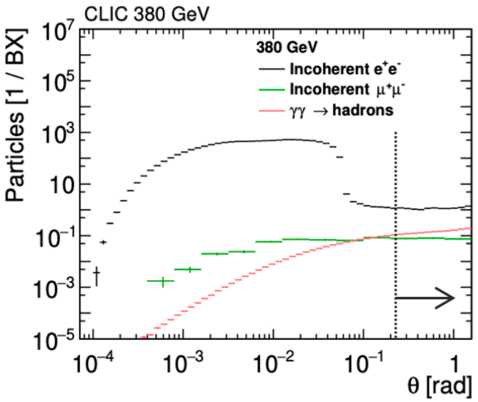
Fermilab

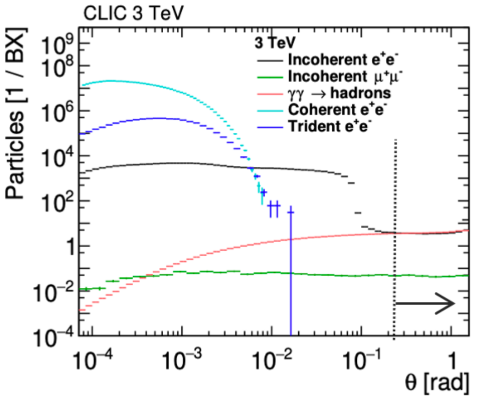
e^+e^- collisions: on-ASIC filtering

- Reject hits from beam backgrounds corresponding to spiraling e^+e^- from e.g. incoherent pair production

Train a classifier to remove hits from tracks with low p_T, small θ







e^+e^- collisions: on-ASIC featurization

- Read out incident particle properties and uncertainties instead of pixel array
- Size of stored pixel data is **independent of geometry**

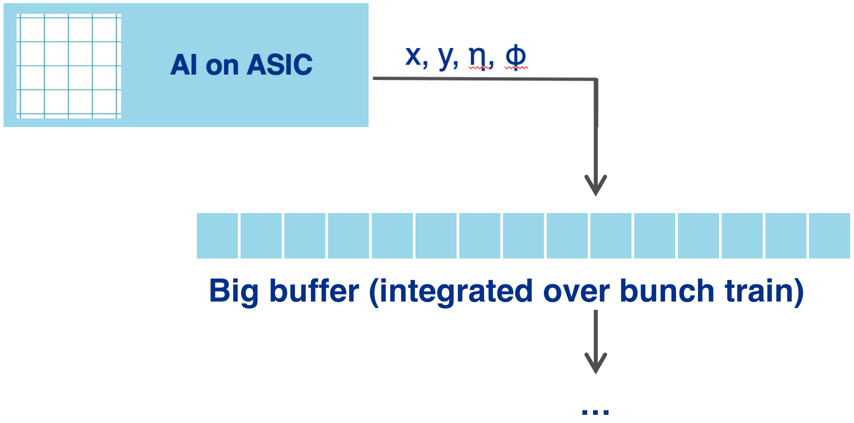
  Decouples buffer size from choice of pixel pitch, sensor thickness

- Good use case for **MAPS with 3D integration**

  MAPS ⟶ small pixels, thin sensors

  MAPS shift information to the periphery

  Additional layer on top would provide area to do the NN regression



Fermilab

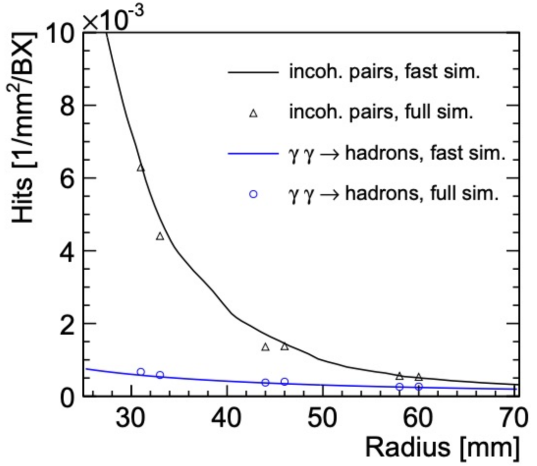
e^+e^- collisions with on-ASIC data reduction

- **Critical design constraints** to keep beam backgrounds out of acceptance could be **relaxed**

  Innermost radius of vertex detector (right), θ acceptance, pixel pitch

- **Reprogramming NN weights for different √s** ensures the most effective filtering

  Adjust NN weights to change p_T and θ thresholds or compensate for radiation damage effects



- Possibility to **reduce thermal load** (material budget) by reading out less data

  Spiking neural networks are especially energy-efficient

- Can significantly **increase the rate**

  Could consider a low-energy machine that might require a trigger

Smart pixels: summary

- AI on-chip has great potential to **reduce data rates to manageable levels**
  - First implementation of the p_T filtering looks very promising!
  - Feature extraction for x, y, α, β underway
- Plan to leverage **emerging technologies** to improve energy efficiency, accuracy
- Co-design with focus on preserving **information that is useful for physics**
  - For e^+e^- this reaches all the way down to accelerator level
  - Smart pixels would provide **more flexibility** in experimental design at linear e^+e^- machines

The smart pixels team

Alice Bean, Doug Berry, Manuel Blanco Valentin, Jennet Dickinson, Giuseppe Di Guglielmo, Karri DiPetrillo, Farah Fahim, Lindsey Gray, Jim Hirschauer, Shruti R. Kulkarni, Ron Lipton, Petar Maksimovic, Corrinne Mills, Benjamin Parpillon, Gauri Pradhan, Morris Swartz, Nhan Tran, Jieun Yoo, Aaron Young

KU THE UNIVERSITY OF KANSAS

Fermilab

OAK RIDGE National Laboratory

JOHNS HOPKINS UNIVERSITY

UIC UNIVERSITY OF ILLINOIS CHICAGO

THE UNIVERSITY OF CHICAGO

Backup material

Pixel detectors at the LHC

- Highest data rates in HEP!

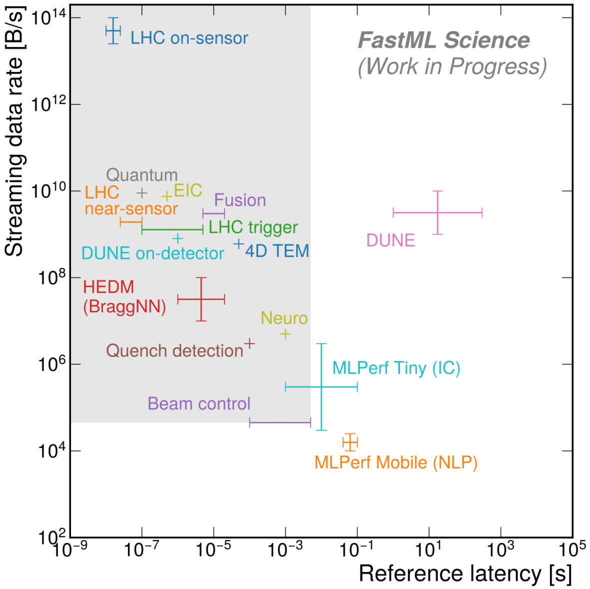
  Only read out for triggered events

  Measure charged particle tracks and vertices

- And getting higher…

  Next generation detectors promise better resolution (position & angle), precision timing

  More information, but also more data



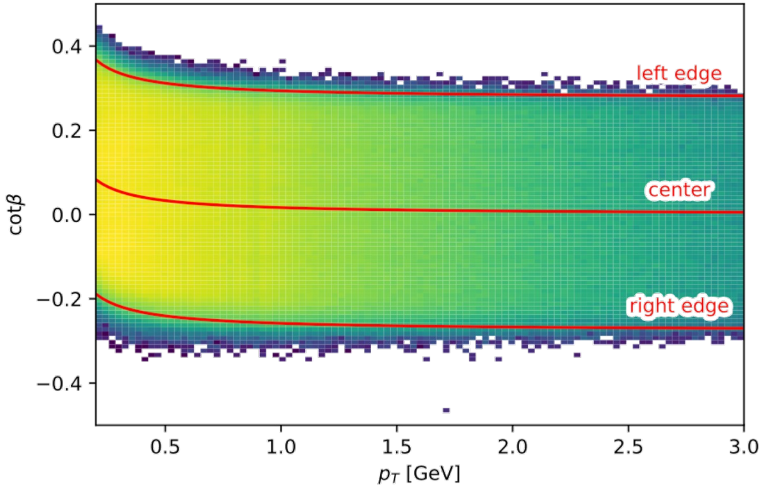
[2207.07958](https://arxiv.org/abs/2207.07958)

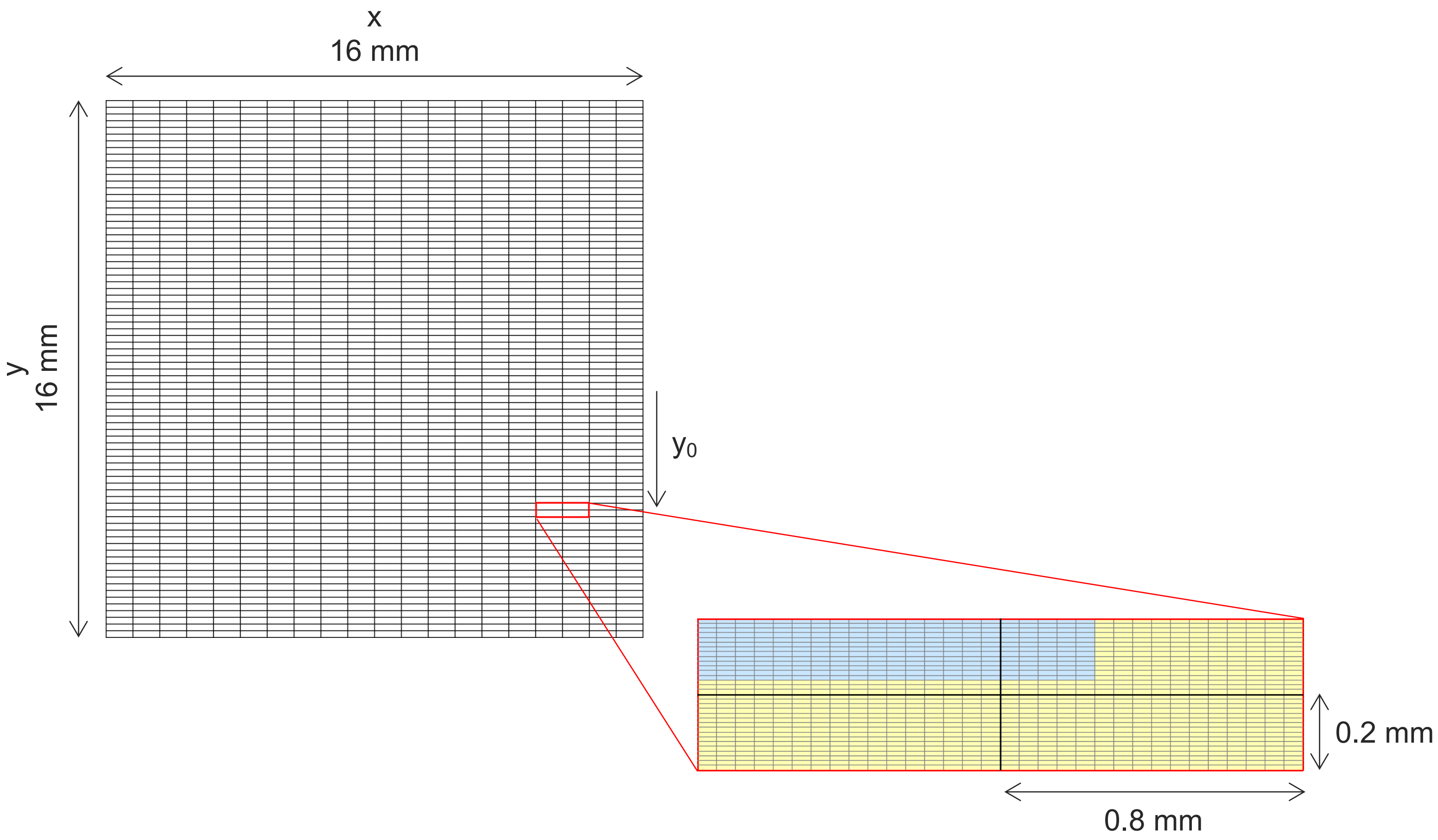
What would we gain if we could analyze it all? Some aspirational targets:

- ***Higgs self-coupling*** *:* 5x increase in the low-m_{hh} spectrum from b-jet triggers.
- ***WIMP dark matter*** : 50x rate for low-p_T / disappearing tracks / long-lived particles.
- ***New capabilities for high-rate, soft objects*** : e.g. dark sector BSM, B-physics, and more!

Incident angle, p_T, and y_0

- Positive charge only

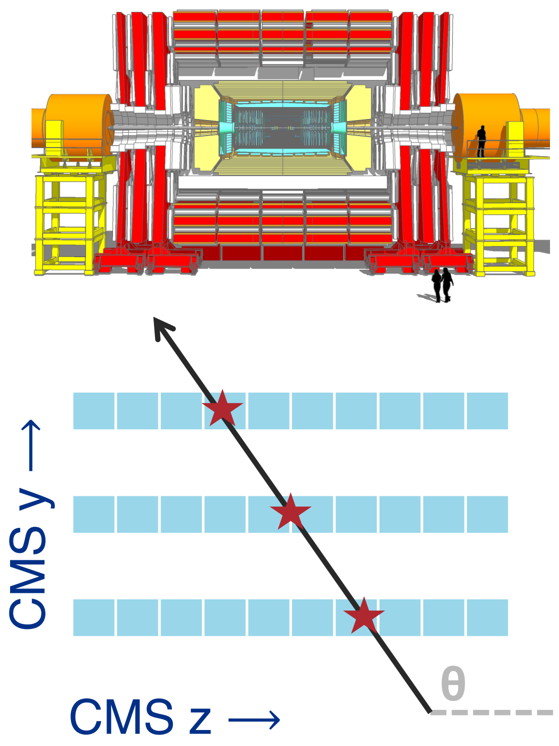


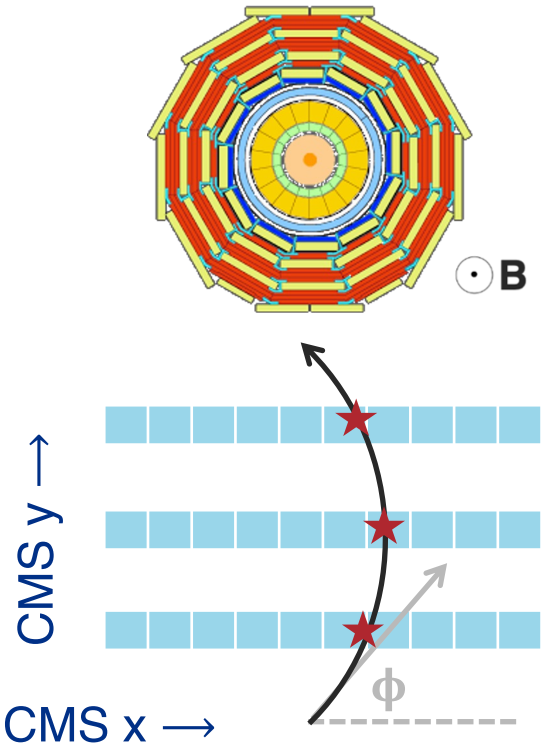


x
16 mm
y
16 mm
y_0
0.2 mm
0.8 mm

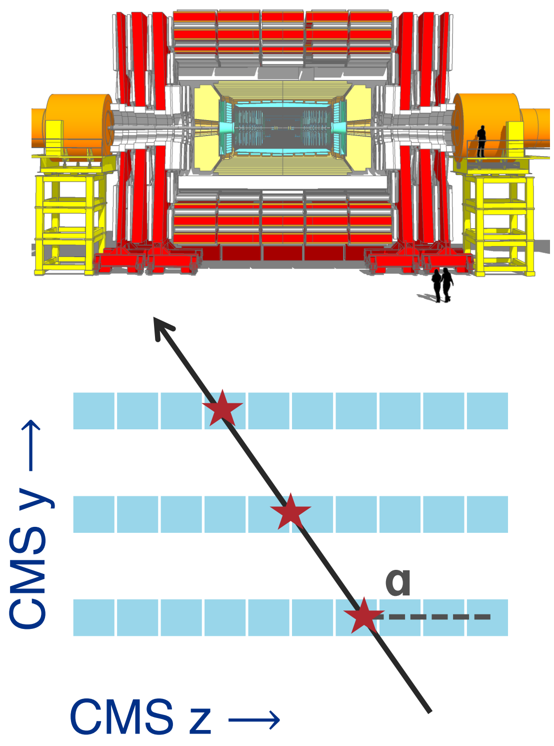
Track diagrams

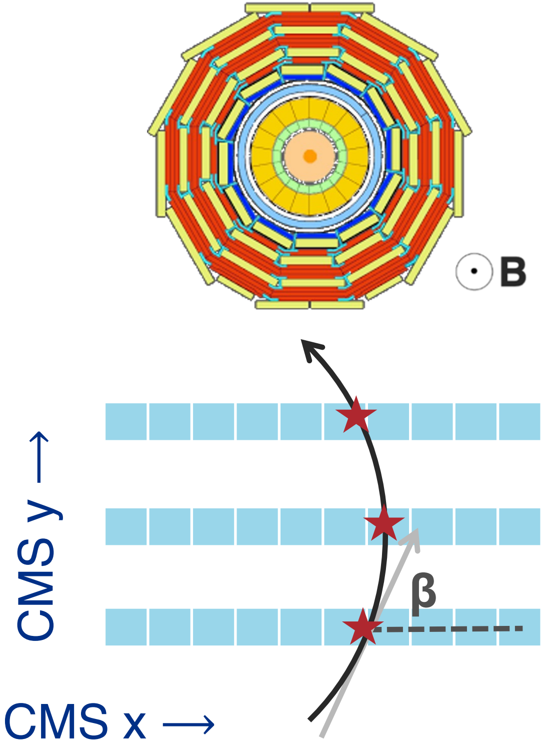
- Combined with other sensor layers for 3D tracking



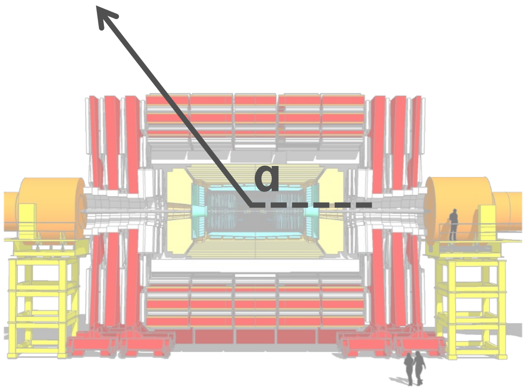


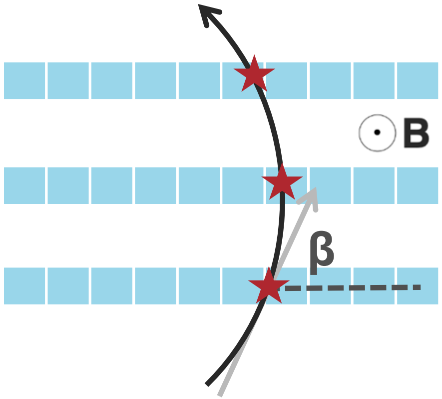
Track diagrams: angles





Track diagrams: angles





Track diagrams: angles (2)