# Future Smart Detectors

## Intelligent Pixel Detectors: Towards a Radiation Hard ASIC with On-Chip Machine Learning in 28nm CMOS

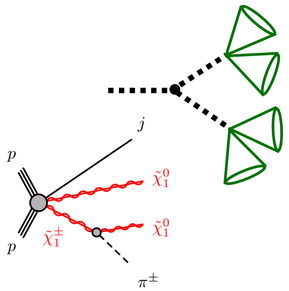

**Anthony Badea**

Enrico Fermi Institute, UChicago

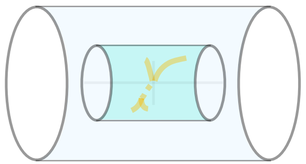
20 July 2024

ICHEP, Prague

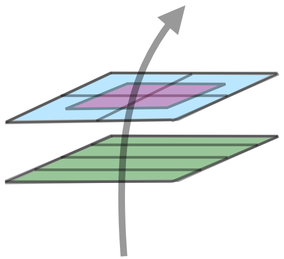
# Outline




Physics Motivation



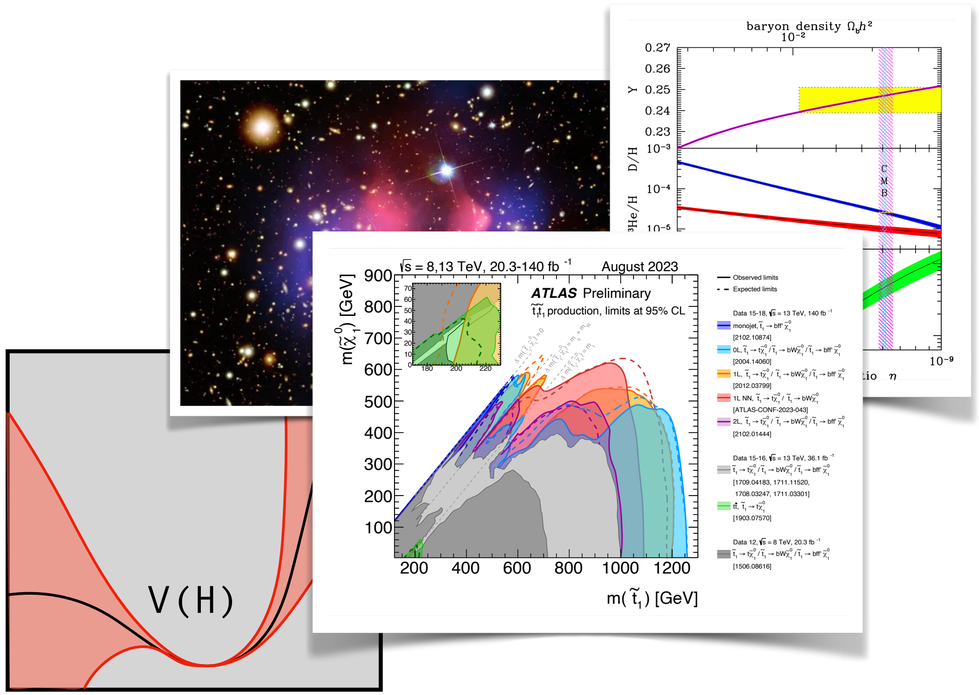
Real Time Tracking Challenges



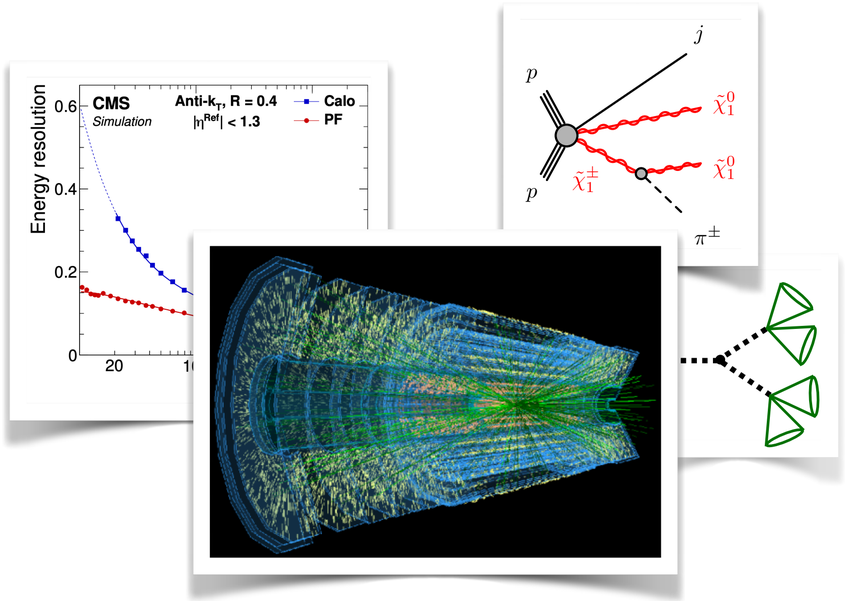
Track Classification in 28nm

# Motivation

We know there is physics to discover




At particle colliders, tracking is crucial



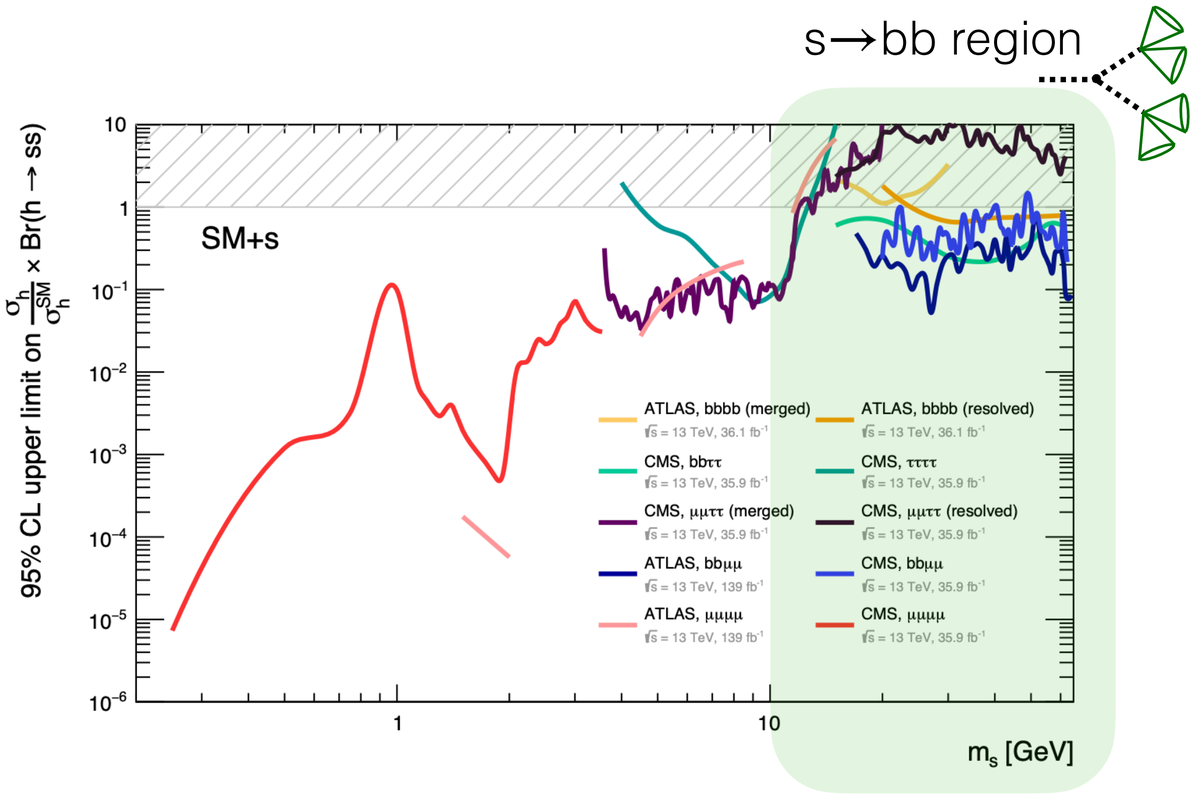
# Illustrative Example

1312.4992, 2111.12751, 2109.03294

Mixed Higgs-scalar scenarios lead to many soft b-quarks (like di-Higgs).

- Overwhelmed by QCD background
- Swap cross section for lepton trigger from associated W/Z

  → limited sensitivity

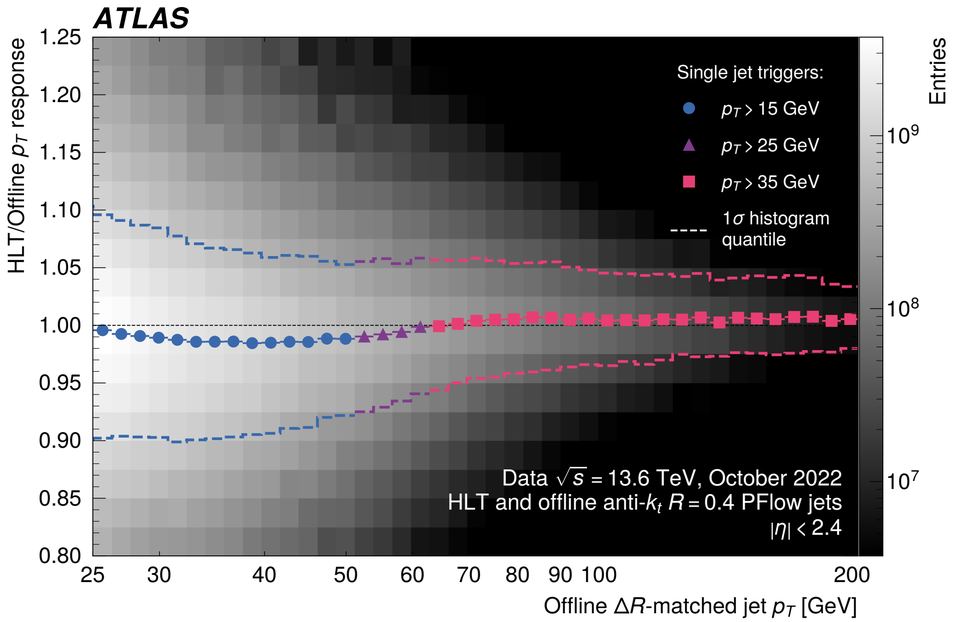


# Just Scout It?

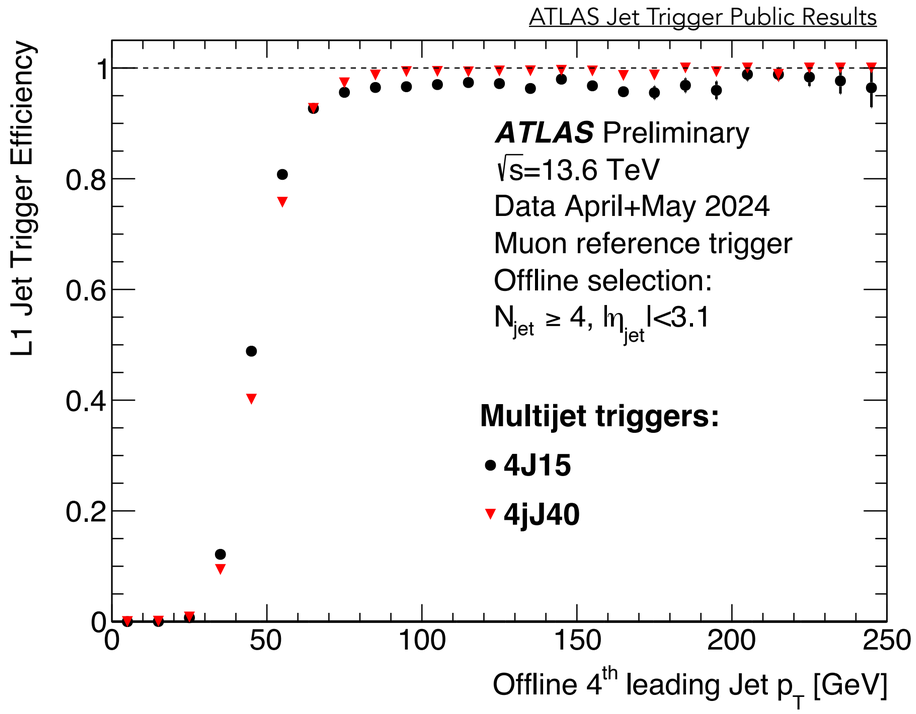
2401.06630

Perhaps use scouting to alleviate this?
HLT resolution within 10% of offline.

Ex. CMS scouting for RPV Higgsino's

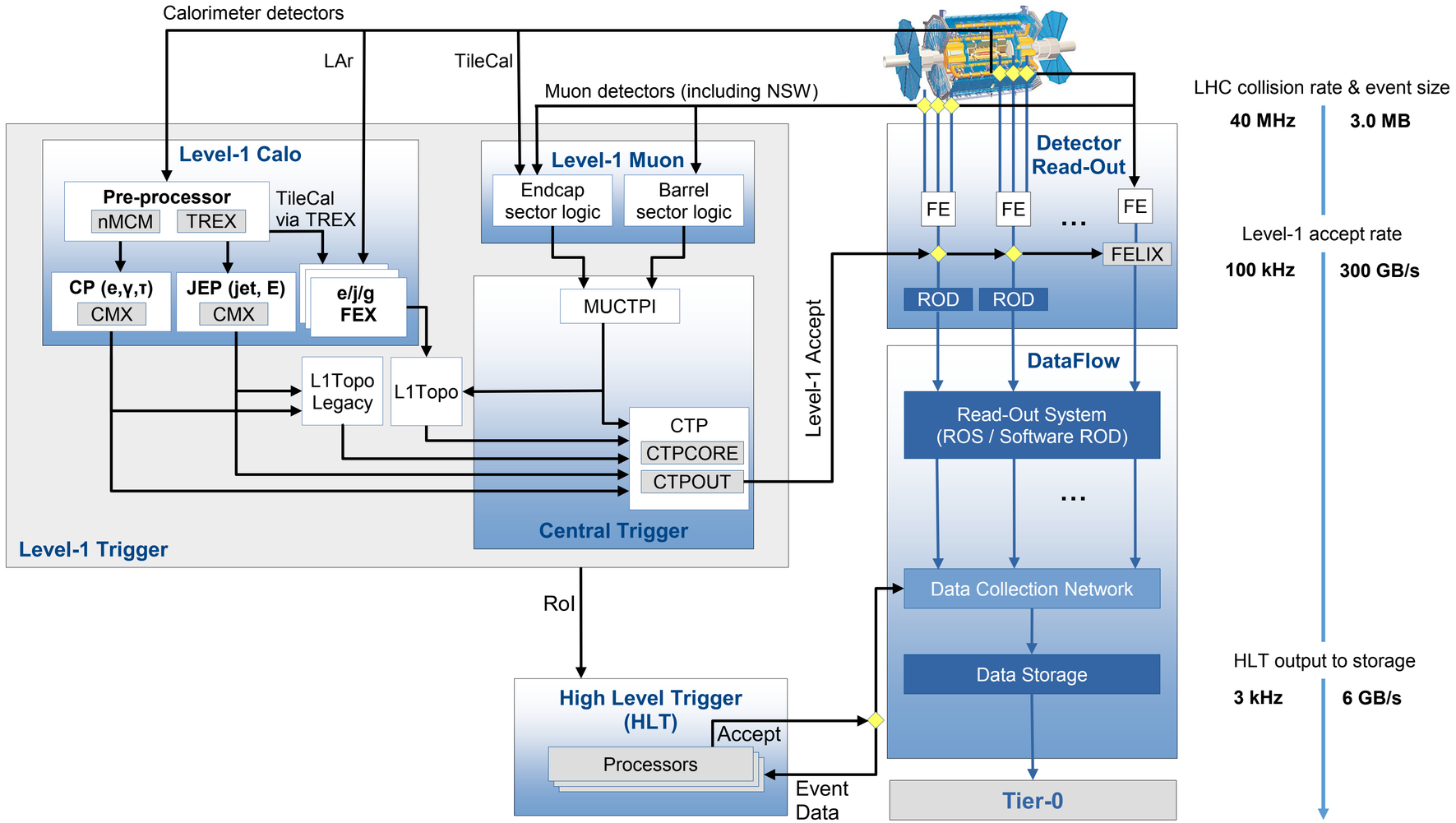



Limited by the L1 (hardware) trigger!



# The Limitation

2401.06630




LHC collision rate & event size

**40 MHz** **3.0 MB**

Level-1 accept rate

**100 kHz** **300 GB/s**

HLT output to storage

**3 kHz** **6 GB/s**

Current technology cannot handle O(PB/sec) data rates so we trigger.

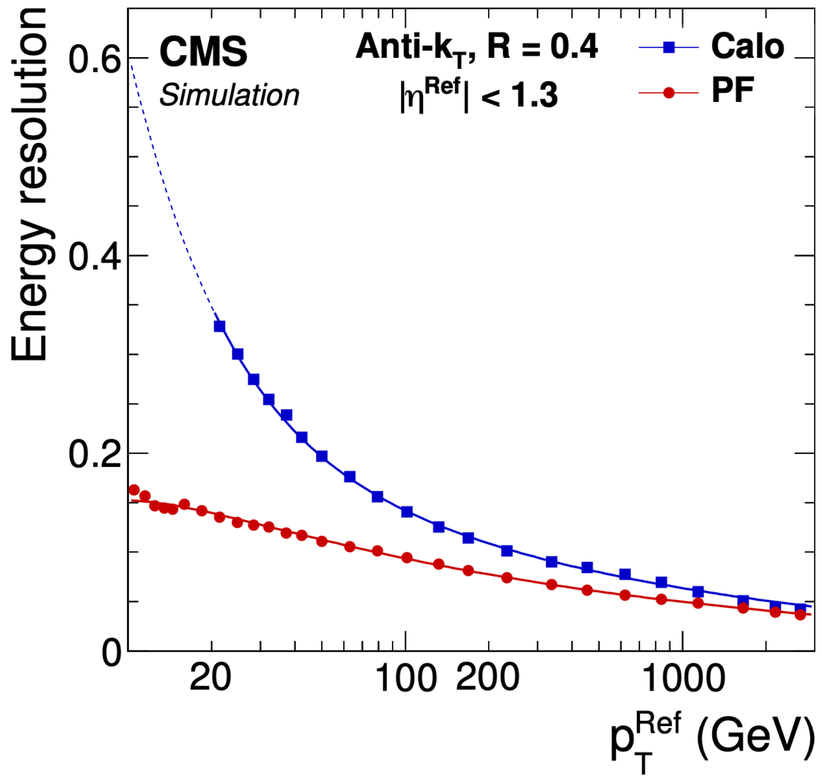
This decision is made without tracking information.

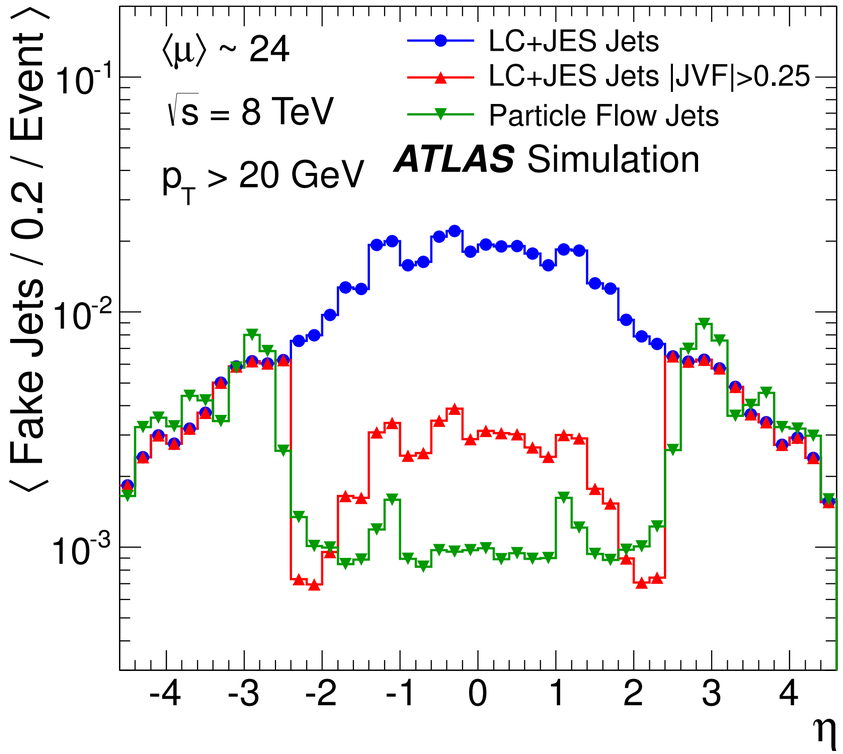
CMS is similar numbers but will have outer tracker information (strips) for HL-LHC

# Consequences

1706.04965, 1703.10485

Lesson already learned adding tracking into jet reconstruction via particle flow (PFlow) algorithm → resolution and fake rejection improve



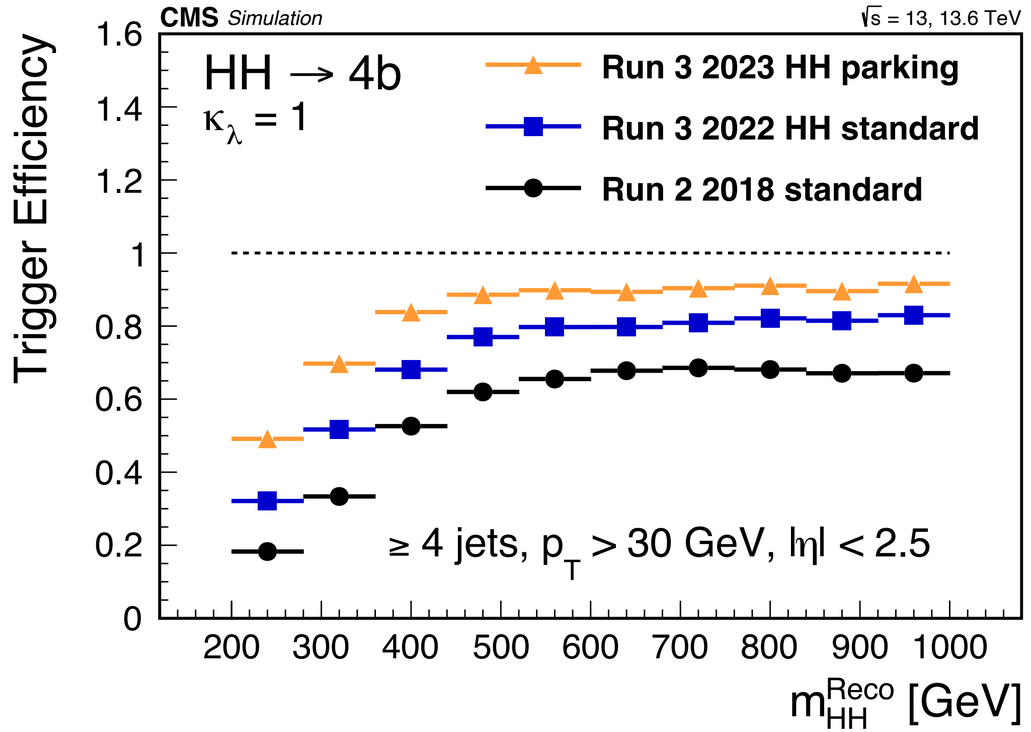
# Just Park It?

2403.16134

After HLT save a higher data rate to a delayed stream, consume less online resources

Shortcomings:

- acceptance limited because of L1 trigger
- No way we can park O(PB/s) of data



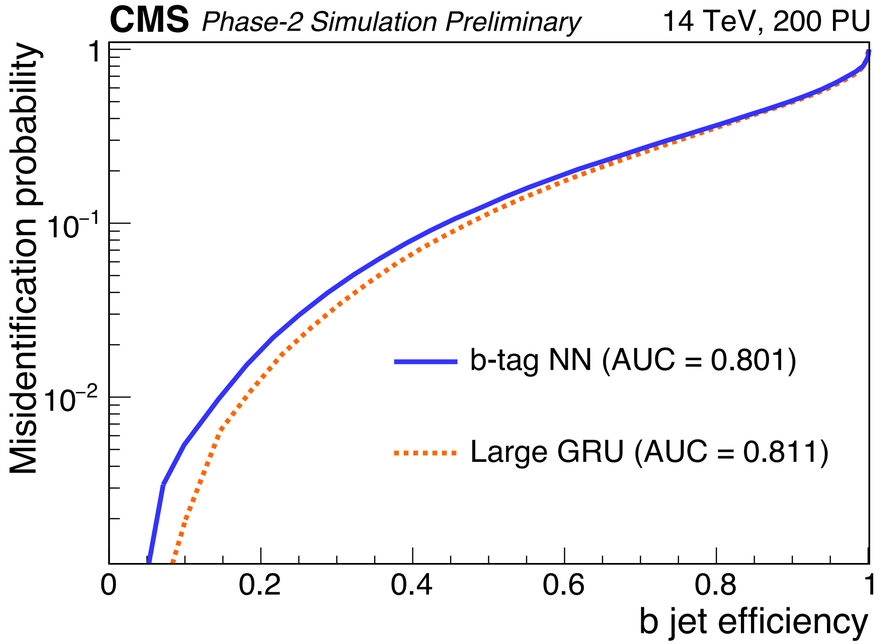
# Motivation

2010.13557, CERN-CMS-DP-2022-021, 2306.09738

## Real time tracking is crucial

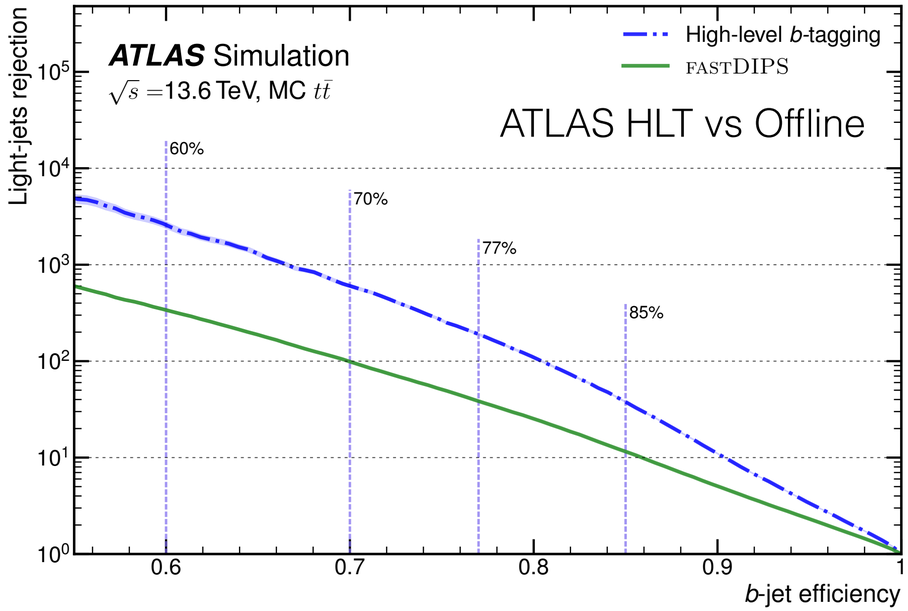
### CMS Phase-2 Outer Tracker (OT)

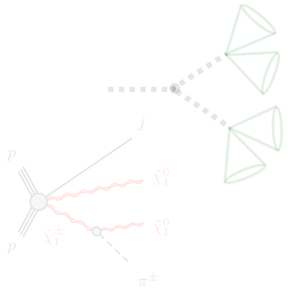
OT hits point back to displaced vertex



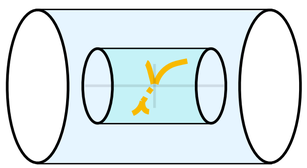

### >10x Improvement with Full Tracker

Not easy to implement online

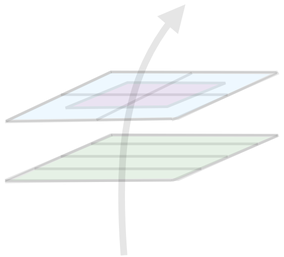




Physics Motivation



# Real Time Tracking Challenges



Track Classification in 28nm

# Technical Challenges

RD53 Collaboration, IEEE 10182033

To utilize tracking information, need readout chips capable of handling the physics conditions

Benchmark is the RD53 conditions for HL-LHC. Strict constraints so trigger rate at 1 MHz / 750 kHz

Talk by Flavio Loddo

| | ATLAS/CMS |
|---|---|
| Chip size | 20x21mm$^2$/21.6x18.6mm$^2$ |
| Pixel size | 50x50 µm$^2$ |
| Hit rate | 3 GHz/cm$^2$ |
| Trigger rate | 1 MHz/750kHz |
| Trigger latency | 12.5 us |
| Min. threshold | 600 e- |
| Radiation tolerance | 500 Mrad @-15C |
| Power | < 1W/cm$^2$ |

# Data Reduction at the Source

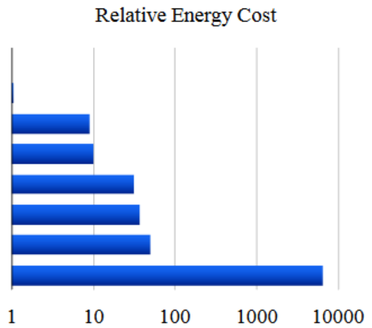
Horowitz in 45nm, UIUC Lectures, Nhan Tran

Aim to achieve tracking at 40 MHz within strict conditions

Key insight: moving data is expensive, doing computation cheaper → perform data reduction at the source could

**Cost of Operations**

| Operation | Energy [pJ] | Relative Cost |
|---|---|---|
| 32 bit int ADD | 0.1 | 1 |
| 32 bit float ADD | 0.9 | 9 |
| 32 bit Register File | 1 | 10 |
| 32 bit int MULT | 3.1 | 31 |
| 32 bit float MULT | 3.7 | 37 |
| 32 bit SRAM Cache | 5 | 50 |
| **32 bit DRAM Memory** | **640** | **6400** |




Mark Horowitz. Energy table for 45nm process, Stanford VLSI wiki
via Han et al., Learning both Weights and Connections for Efficient Neural Networks

# Machine Learning on Chip

Question: Can ML on chip perform effective data reduction at the source?

Key challenges:

- operate on silicon within power, space, timing constraints
- deliver the required data reduction to meet bandwidth requirements

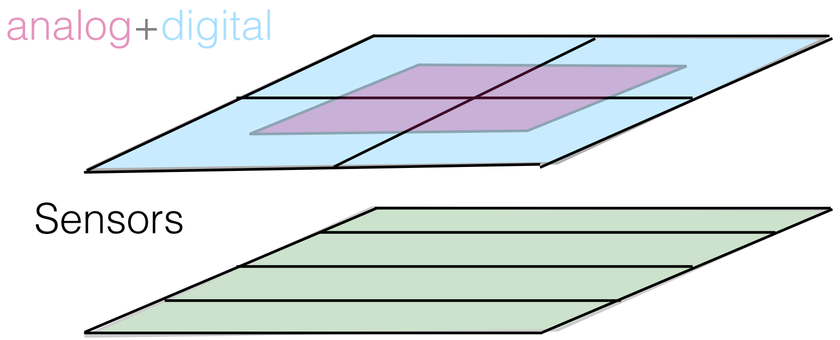



Diagram of pixel sensors + readout chip

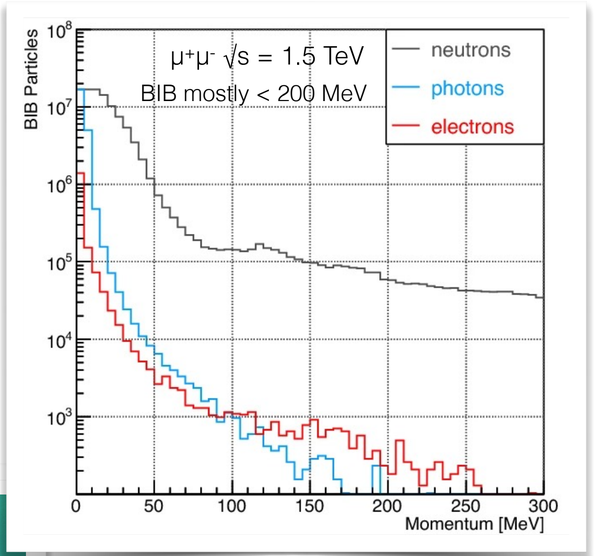
# *Broader Applications*

10.1007, 1808.02969, 1809.10213

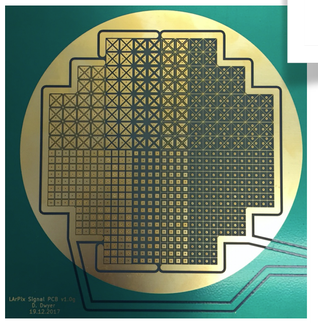
Could imagine distributed ML across detectors more generally … *Smart Detectors*

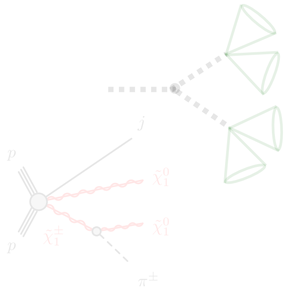
Examples:

- Beam-induced background at a MuC
- Dual readout calorimeters
- Ultra-high granularity sampling calorimeters
- Pixel LArTPCs

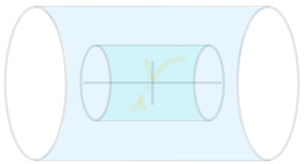



LArPix TPC Facing Side

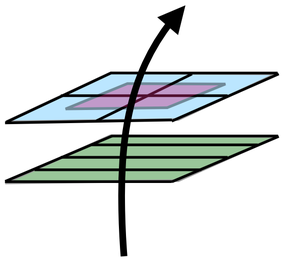




Physics Motivation



Real Time Tracking Challenges

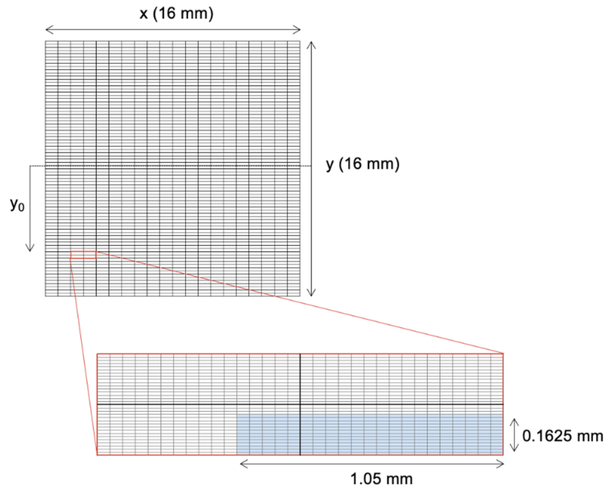


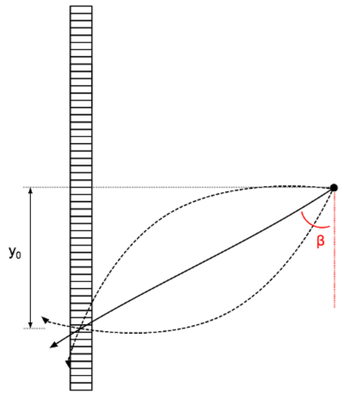
Track Classification in 28nm

# Physics Setup

CMS-NOTE-2002-027, 2310.02474

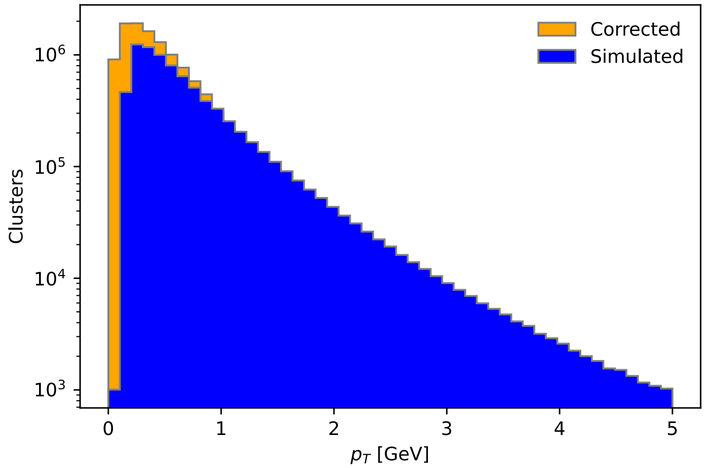
## Particle passing through a 100 μm thick single layer of silicon with small pitch 12.5 x 50 μm² pixels




Non-exhaustive list of important details:

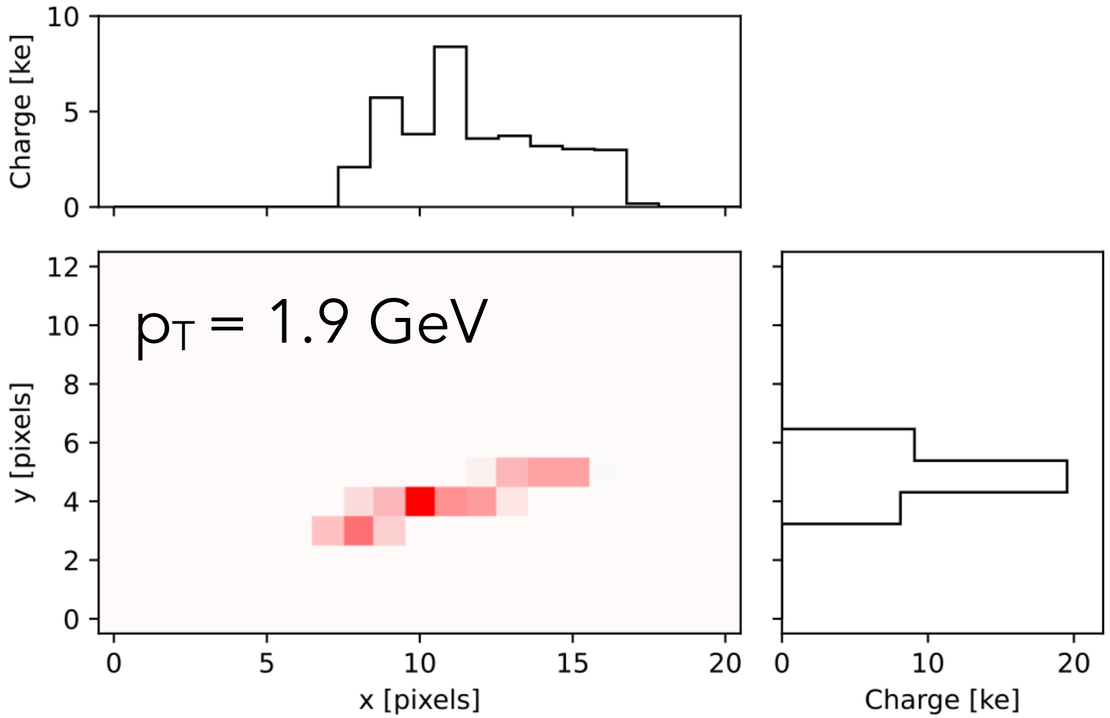
- Tracked data taken from CMS with $p_T$ up to ~5 GeV.
- Untracked data not included and includes CMS acceptances
- PixelAV silicon simulation used
- Single 100μm thick layer of silicon with 12.5x50 μm² pixels
- Overall sensor area 16x16 mm²
- Bias voltage of -100 V
- Simulation assumes only pions input
- Charge deposition recorded every 200ps
- Sitting on a cylinder of radius 30 mm
- 3.8 TeV B-field parallel to x-coordinate

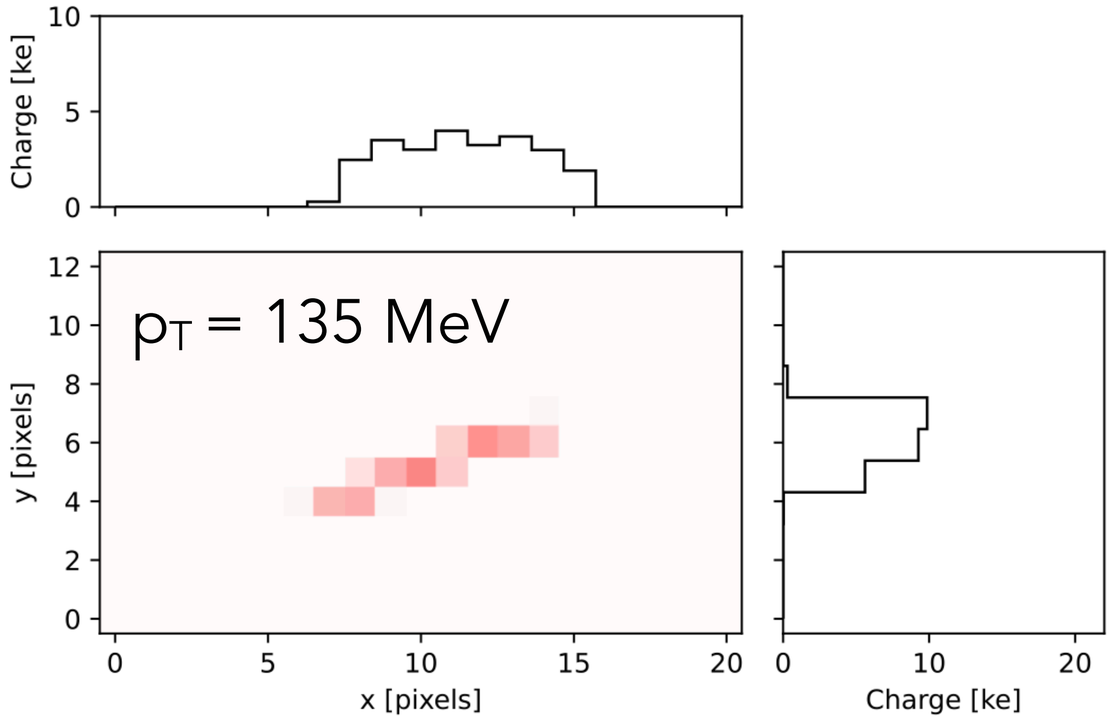


# Example Tracks

2310.02474

Can see visually high (left) vs low (right) pT tracks
bending in B-Field and different cluster shapes



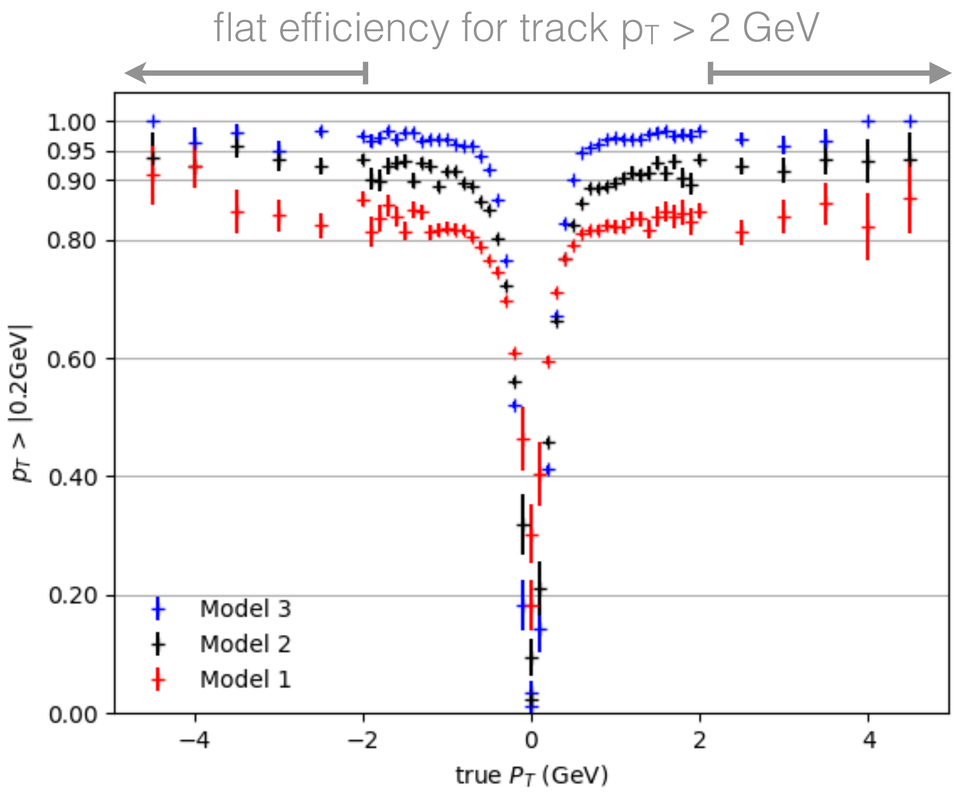
# Classification Network

2010.13557, 2310.02474, 2312.11676, CatapultAI

Flat signal efficiency for track $p_T > 2$ GeV. Data reduction of 57.1 - 75%.

Non-exhaustive list of important details:

- Input: y-profile of charge, no timing
- Output: predict if $p_T > 200$ MeV
- QKeras quantized 2 Layer DNN
- Translated to silicon with CatapultAI
- Power consumption 300μW



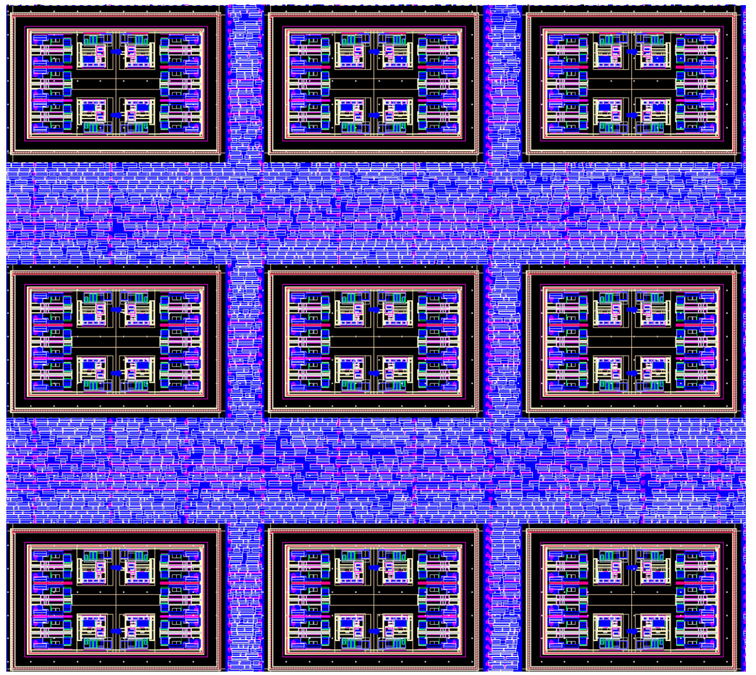
# Chip Tape-out

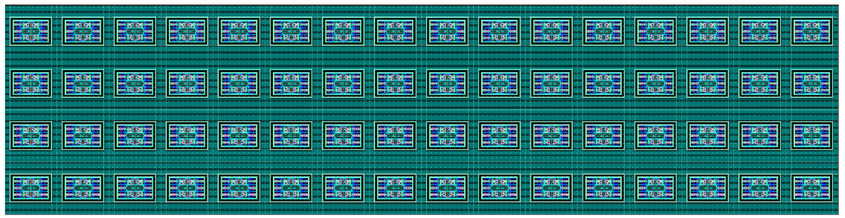
2406.14860

28nm CMOS demonstrated to be radiation tolerant by CERN. First tape-out in new technology node at Fermilab

Layout:

- 2x2 pixel analog islands (within black boxes) surrounded by digital with DNN and test interface (purple space)
- Taped-out as super pixels (16x4) corresponding to 32x8 physical pixels

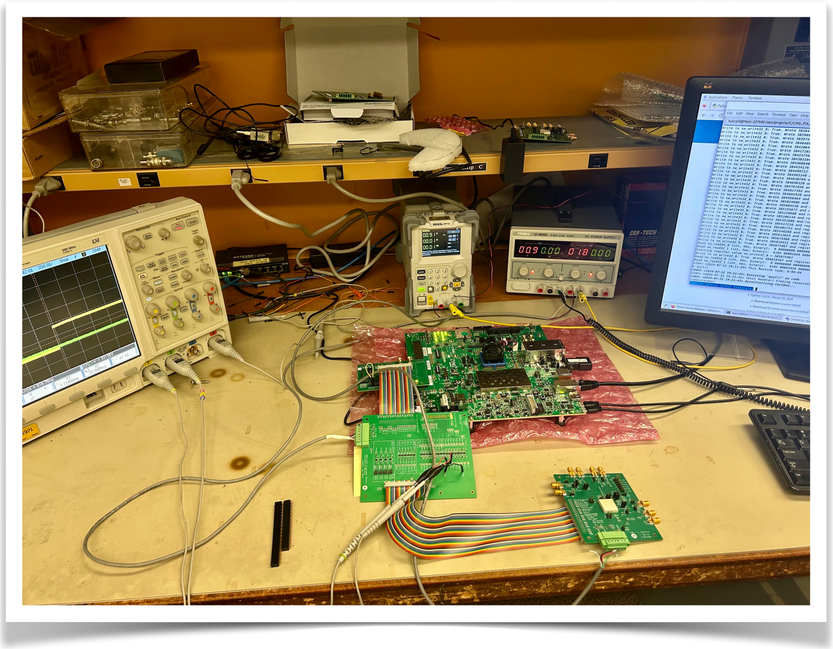




# First Test: Check for Timing Violations

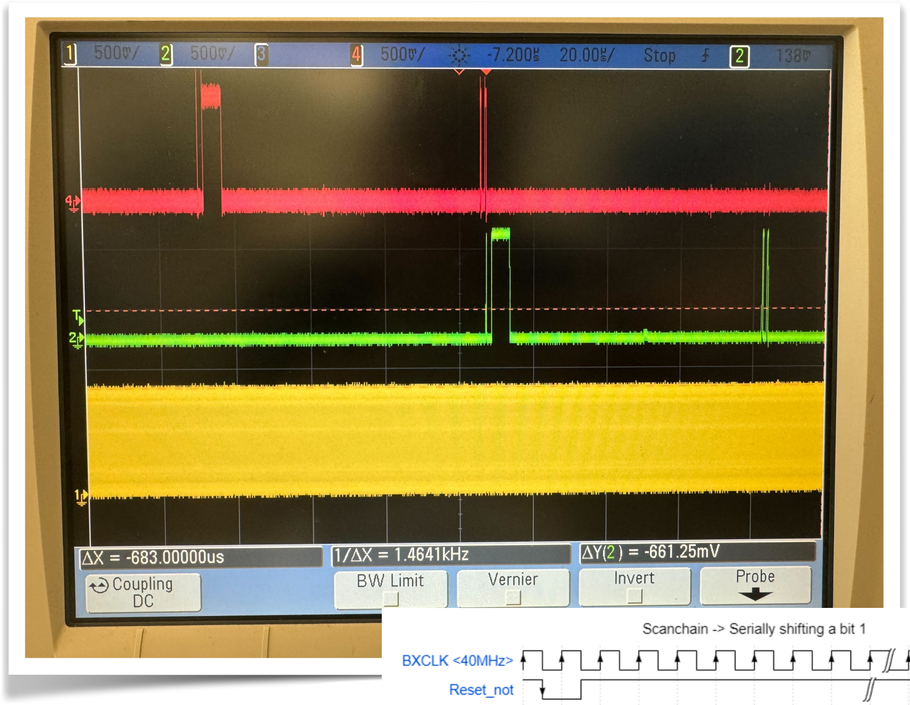
## Loopback test to check for timing violation

when a signal does not propagate through the circuits within the required time constraints, may cause unstable circuit behavior



## No violations are seen

Now working on high statistics pattern pulsing to test the analog+DNN



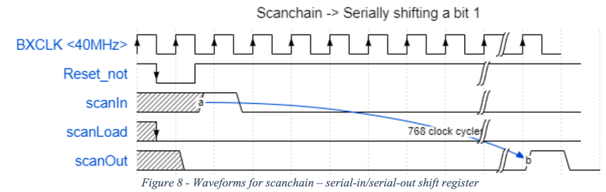




*Figure 8 - Waveforms for scanchain – serial-in/serial-out shift register*

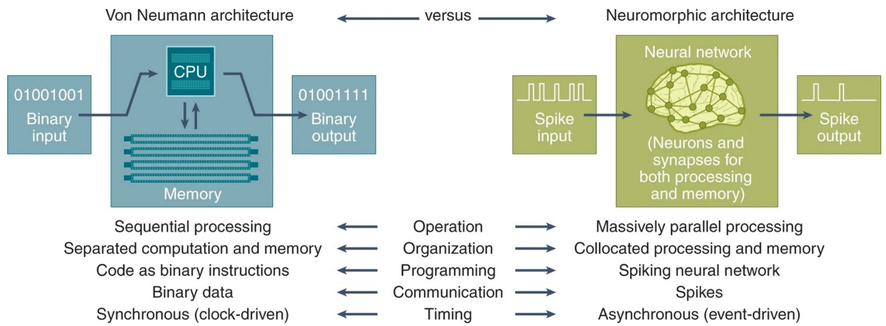
# Next Steps

Lots of work to do but lets assume everything works and mention exciting future directions

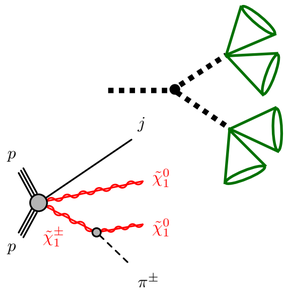
- Build a bigger chip to bond to a real sensor and examine in a test beam
- Analog NN: real edge computing, reduce digitization (1 ADC for entire chip as opposed to 1 per pixel)

Fermilab Test Beam Facility
Not our telescope but hopefully one day


Opportunities for neuromorphic computing algorithms and applications
Classical (~digital) vs Neuromorphic Computing (~analog)

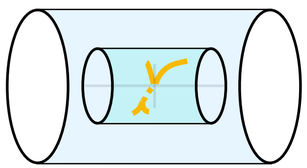


# Summary



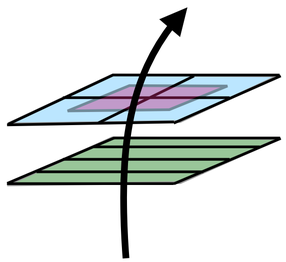

## Physics Motivation

*To study the electroweak symmetry breaking and search for new physics we need real time tracking*



## Real Time Tracking Challenges

*Data reduction at the source with ML on chip is an exciting R&D avenue to achieve real time tracking*



## Track Classification in 28nm

*Developed a classification network to predict track momentum and taped it out in 28nm CMOS. First steps toward creating the new technology*

# *Thank you from our team*

**Fermi National Accelerator Laboratory:** Abhijith Gandrakota, Benjamin Parpillon, Chinar Syal, Douglas Berry, Gauri Pradhan, Giuseppe Di Guglielmo, James Hirschauer, Jennet Dickinson, Lindsey Gray, Nhan Tran, Ron Lipton, Farah Fahim (lead)
**Johns Hopkins University:** Dahai Wen, Morris Swartz, Petar Maksimovic
**Northwestern University**: Manuel Blanco Valentin
**Oak Ridge National Laboratory**: Aaron Young, Shruti R. Kulkarni
**University of Chicago:** Karri DiPetrillo, Anthony Badea, Carissa Kumar, Emily Pan, Rachel Kovach-Fuentes, Aidan Nicholas, Eliza Howard, Eric You
**University of Illinois Chicago:** Corrinne Mills, Danush Shekar, Jieun Yoo, Mohammad Abrar Wadud
**University of Illinois Urbana-Champaign**: Mark S. Neubauer, David Jiang
**University of Kansas**: Alice Bean
**Purdue University:** Mia Liu, Arghya Das

smartpixels


Jieun Yoo


Danush Shekar


Rachel Fuentes


Eliza Howard


Carissa Kumar


Aidan Nicholas


David Jiang


Arghya Das


Shiqi Kuang

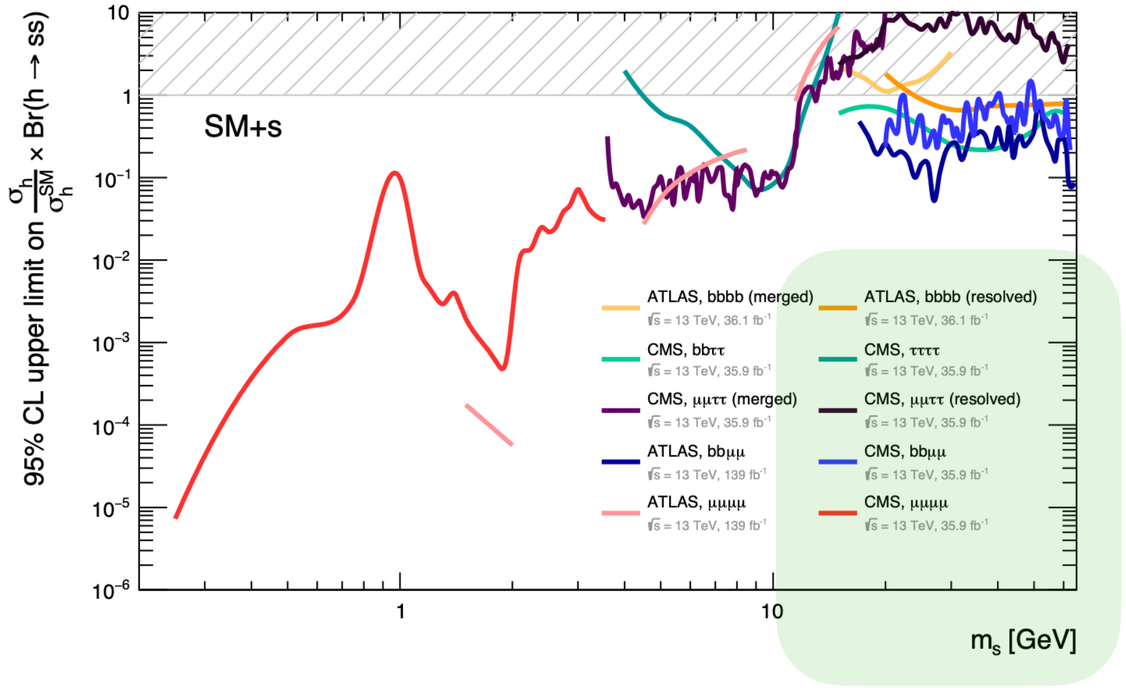
# BACKUP

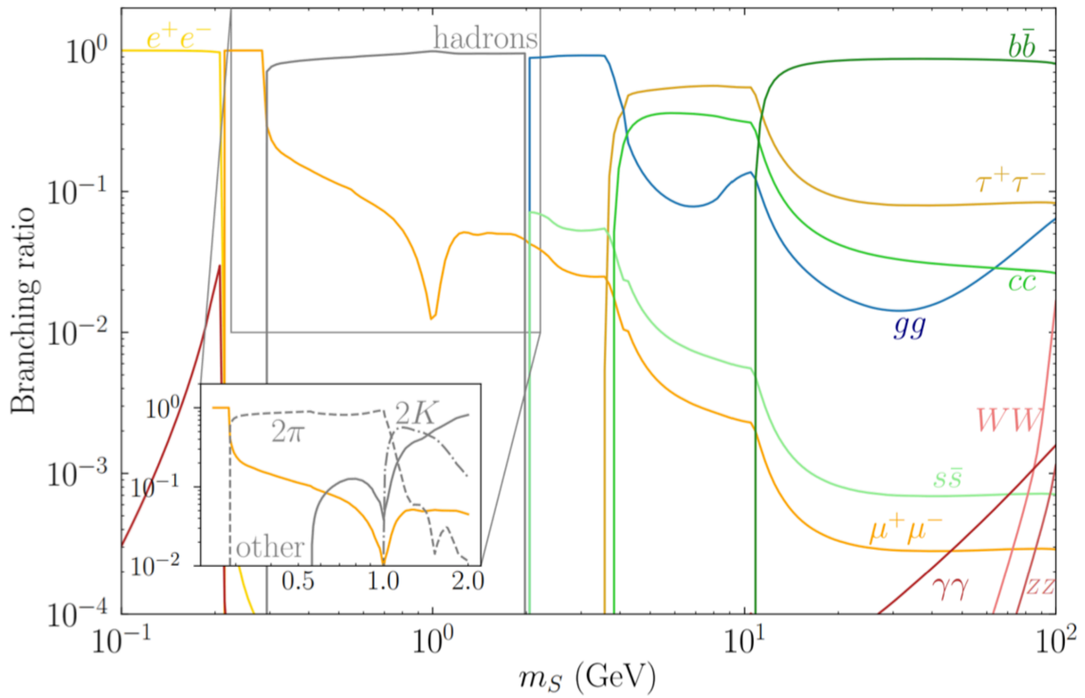
# Physics Scenarios

# SM+s Scenario

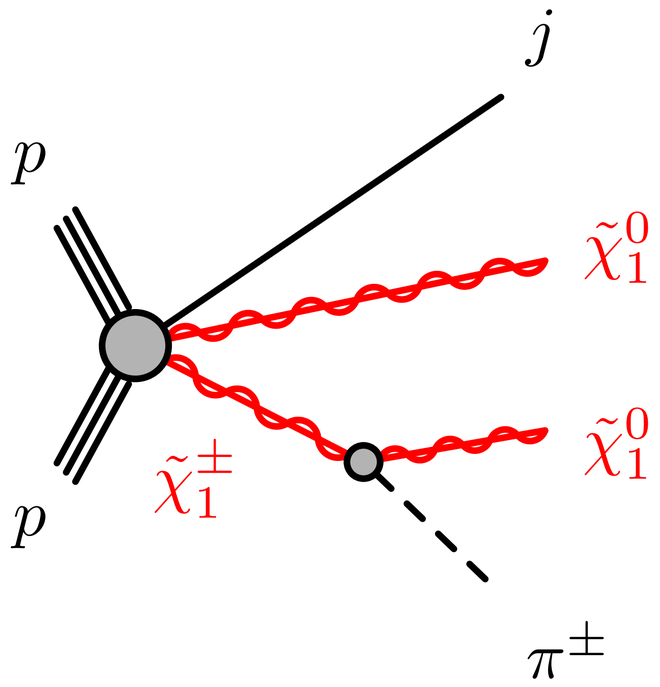
1312.4992, 2111.12751, 2109.03294

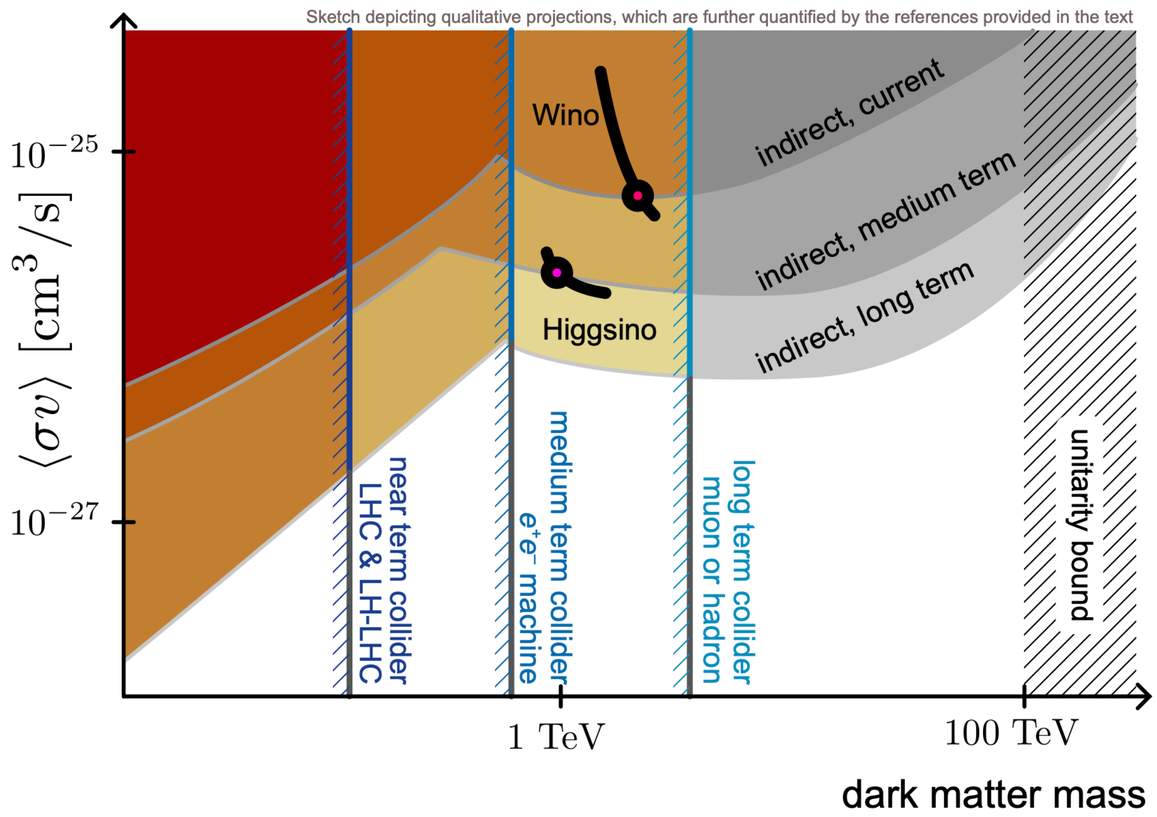
Mixed Higgs-scalar scenarios lead to many soft b-quarks (like di-Higgs). Large QCD background overwhelms data rate, so swap cross section for leptonic trigger from associated W/Z. This makes analysis possible but limits sensitivity.



# Displaced Track DM Scenario

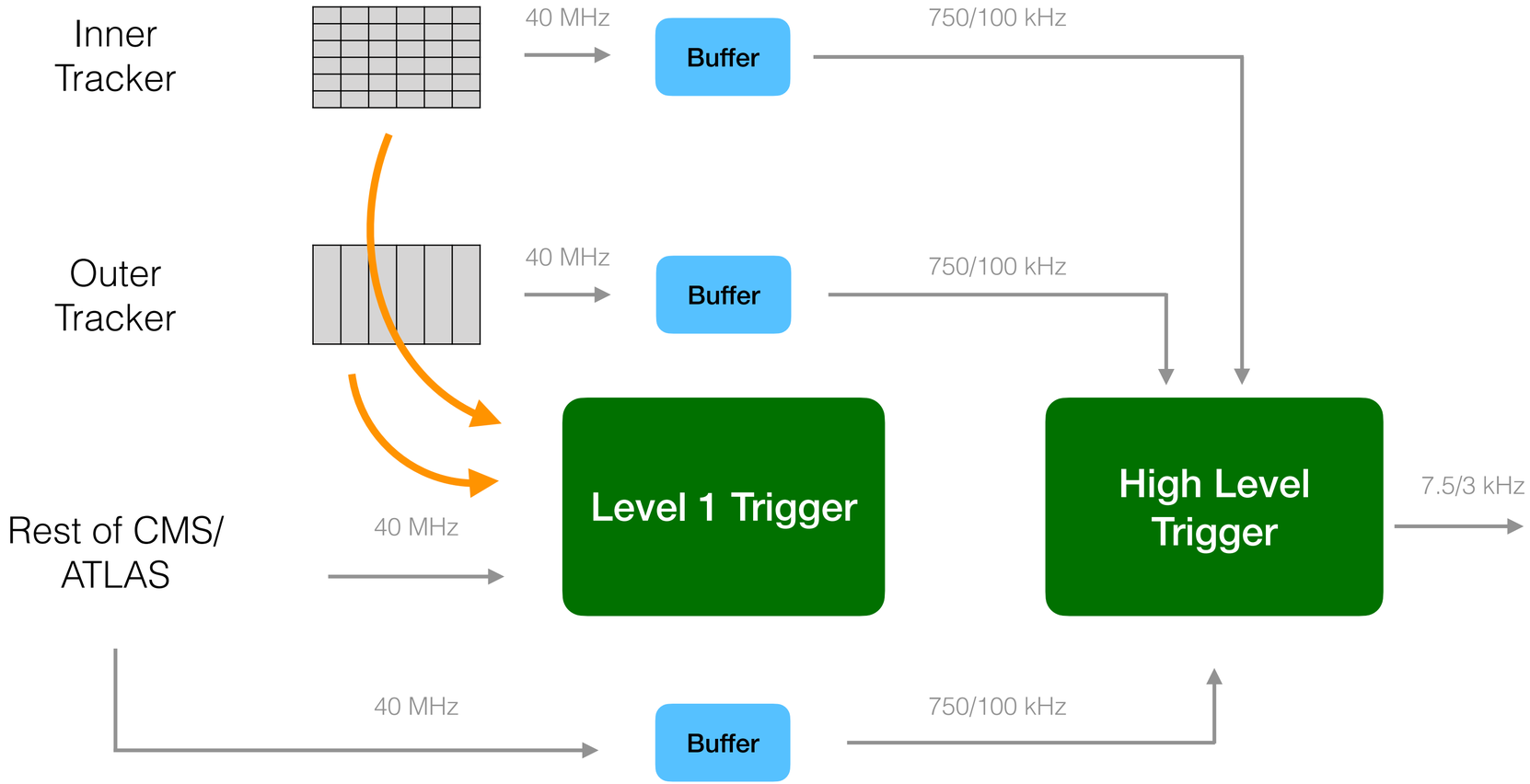


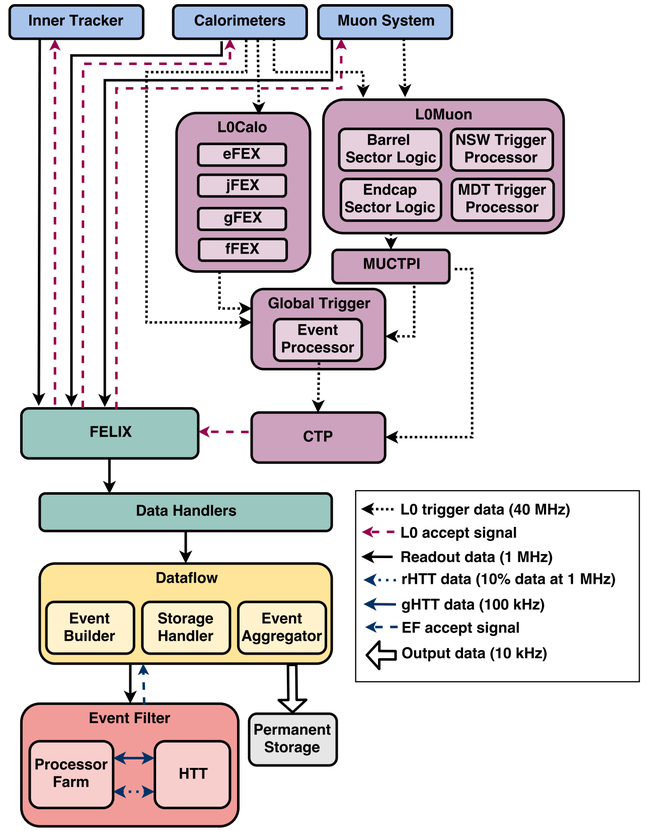



# Detector Considerations

# CMS/ATLAS Trigger Schemes

2209.15519, CERN-LHCC-2017-020



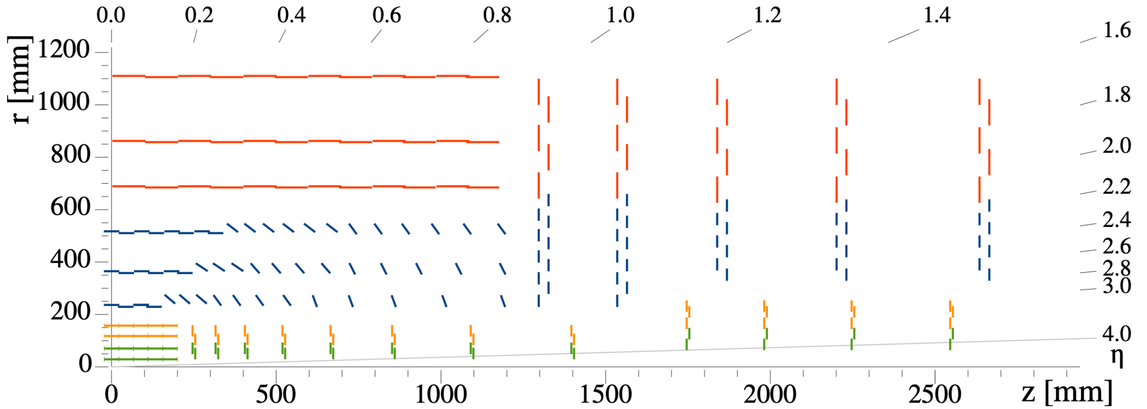



# Tracking

2010.13557, CERN-CMS-DP-2022-021

$$\left(\frac{\sigma_{p_\mathrm{T}}}{p_\mathrm{T}}\right) \sim \frac{p_\mathrm{T}}{BL^2}\frac{\sigma_\mathrm{point}}{\sqrt{N}}$$

CMS Phase 2



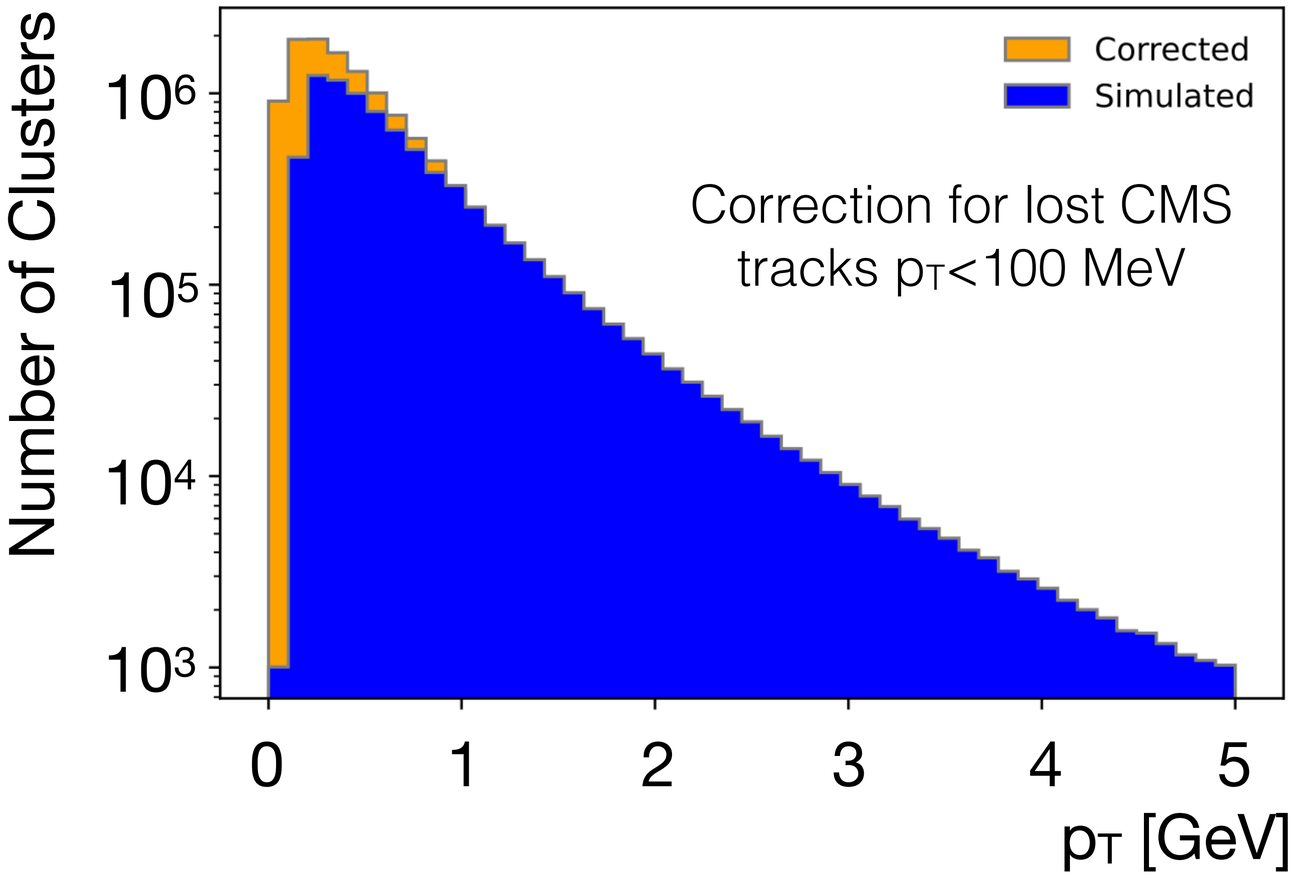

ATLAS Phase 2

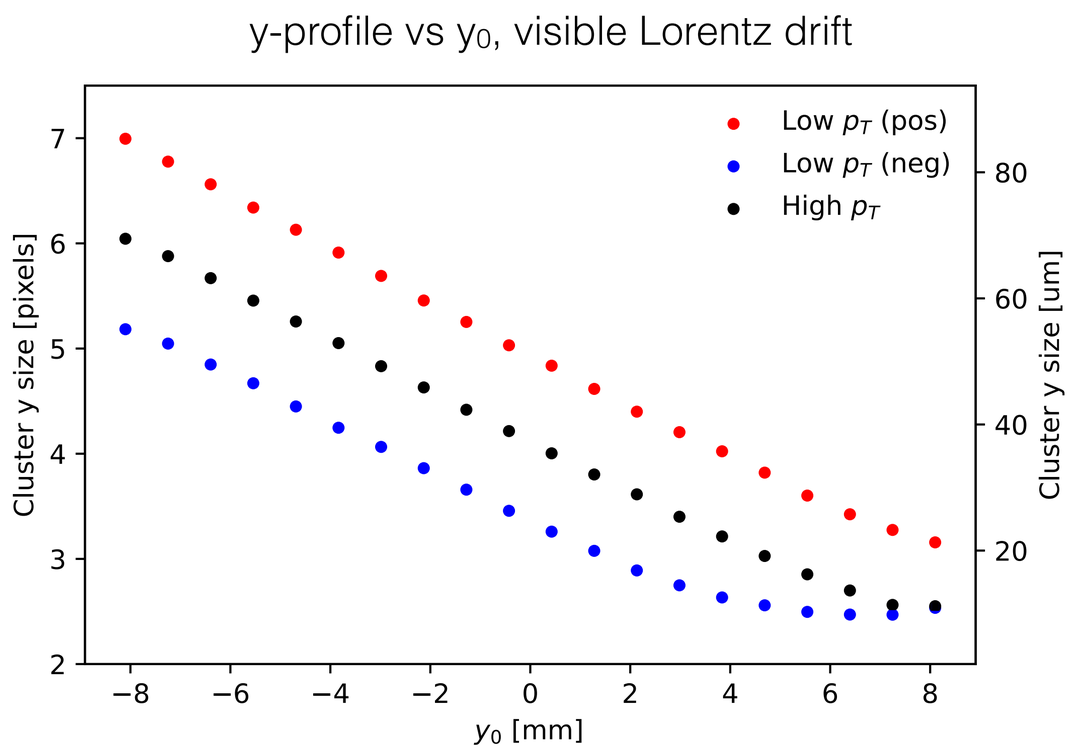
B-meson decay length

$$\langle L \rangle \sim 100\ \mathrm{mm} \times \left(\frac{E}{\mathrm{TeV}}\right)$$

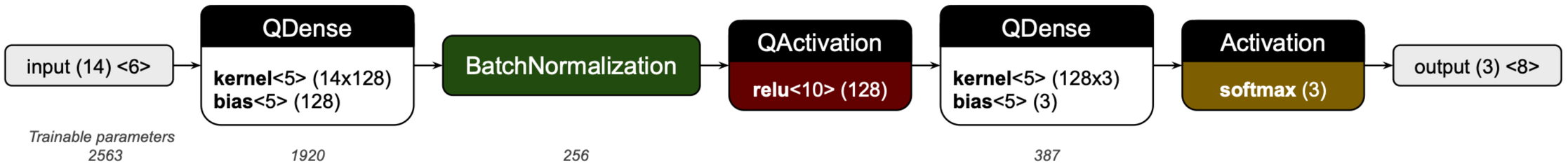
# Classification Network on 28nm v1 chip

# Training Data

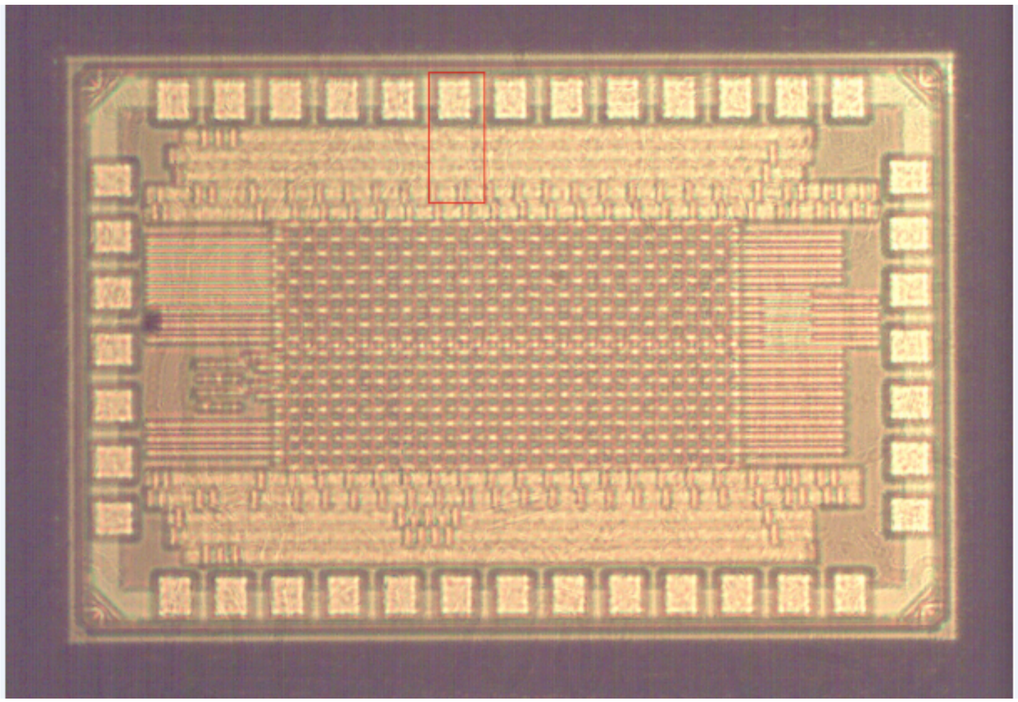


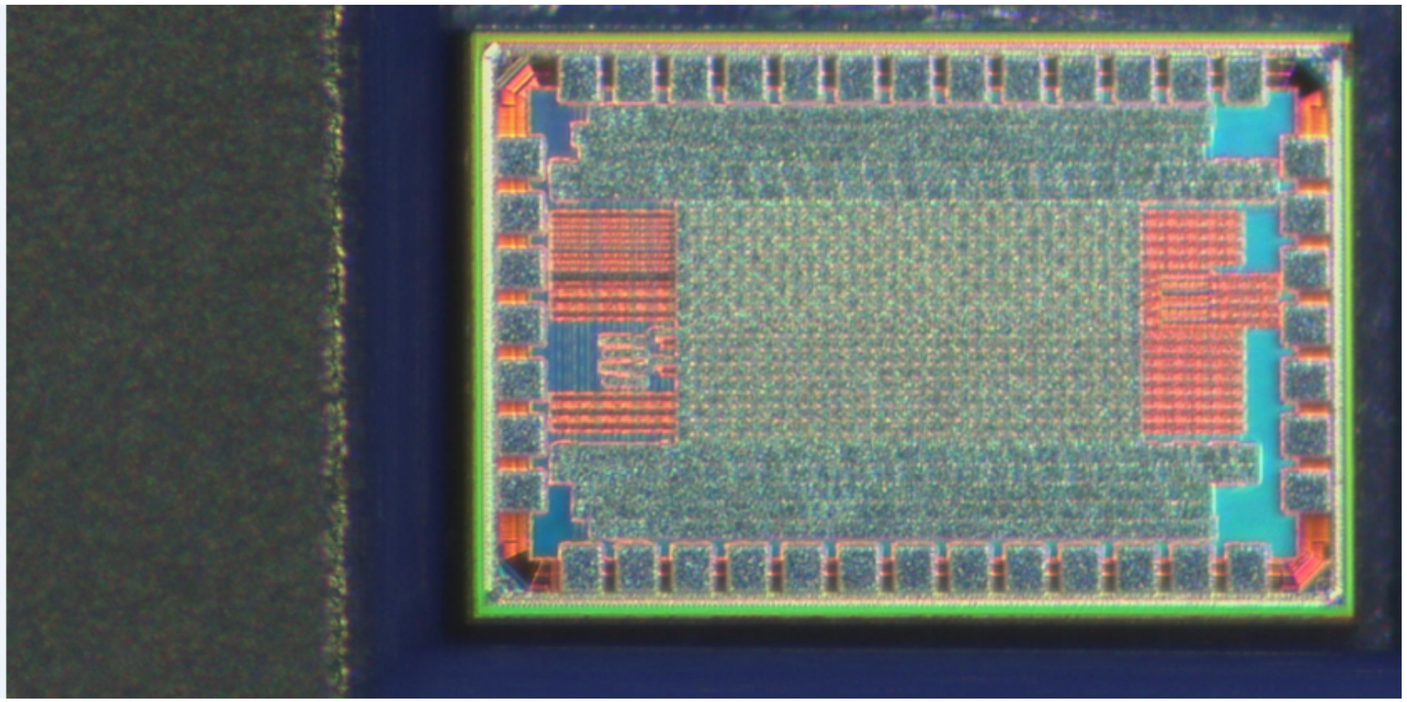



# Neural Network Design



# Chip Photos





# Chip Tape-out Details

2406.14860

Technical:

- Analog island with amplifier and 2-bit ADC
- Digital logic surrounding with DNN inside (translated with CatapultAI)
- 28nm CMOS process from TSMC with Muse

Logistics:

- Received chip back last month June ’24
- ~\$14k/mm$^2$, 1.5 mm$^2$ tape-out ~ \$30k

# Test Bench

2406.15181

② SoC FPGA

Peary server running on PS
Custom IP running on PL

③ Mezzanine

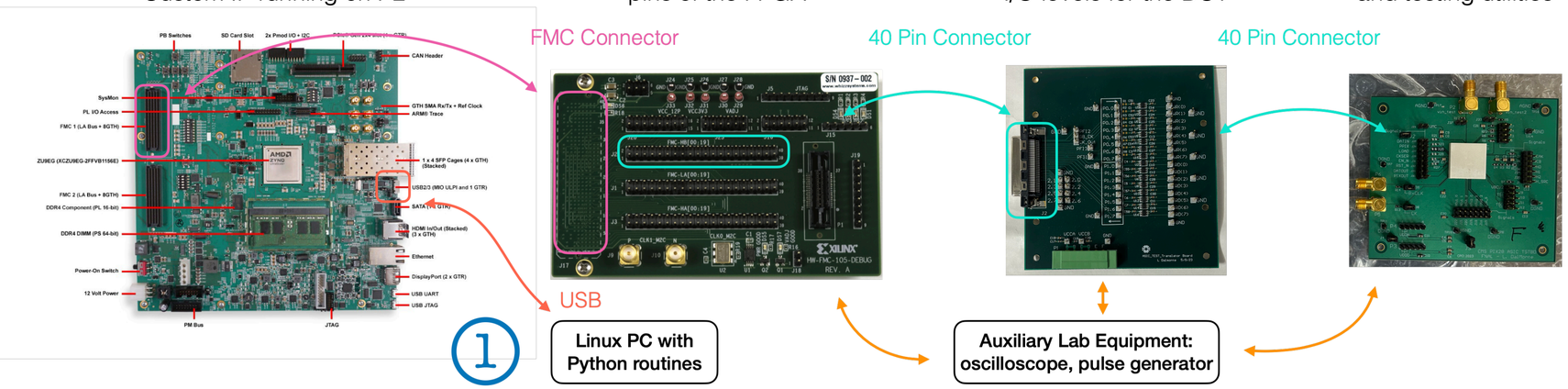
I/O interface to the FMC pins of the FPGA

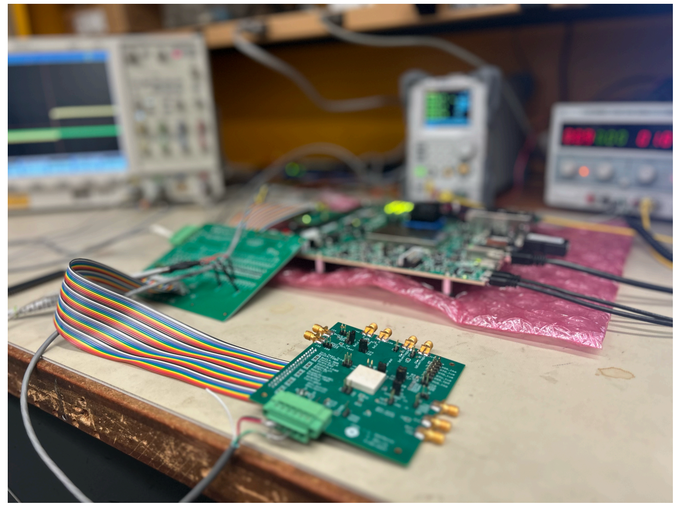
④ Level Shifter

PCB to shift the signals to I/O levels for the DUT

⑤ DUT

PCB housing the ASIC and testing utilities



Test bench that tests the core functionality of the chip. Reusable for future tape-outs. Builds upon Spacely (Adam Quinn)