Fermilab

May 20th 2026

Smartpixels: In-pixel AI for on-sensor data filtering

Benjamin Parpillon
FEE2026

smartpixels

U.S. DEPARTMENT of ENERGY
Fermi National Accelerator Laboratory is managed by FermiForward for the U.S. Department of Energy Office of Science

FERMI NATIONAL ACCELERATOR LABORATORY

Agenda

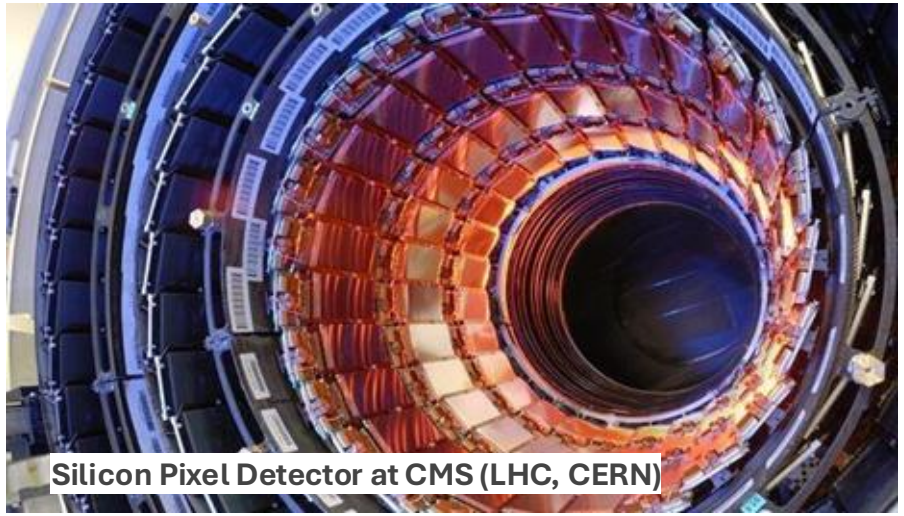
- Physics Motivation
- Filtering Machine Learning
- Front-End Implementation
- Test Result
- Status And Next Steps

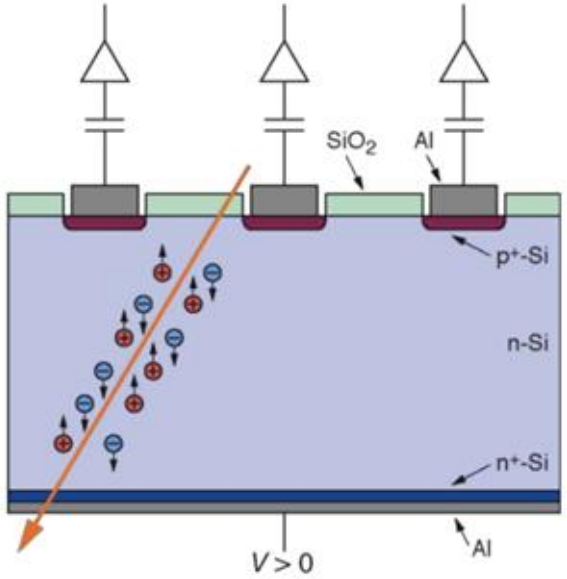
01

Physics Motivation

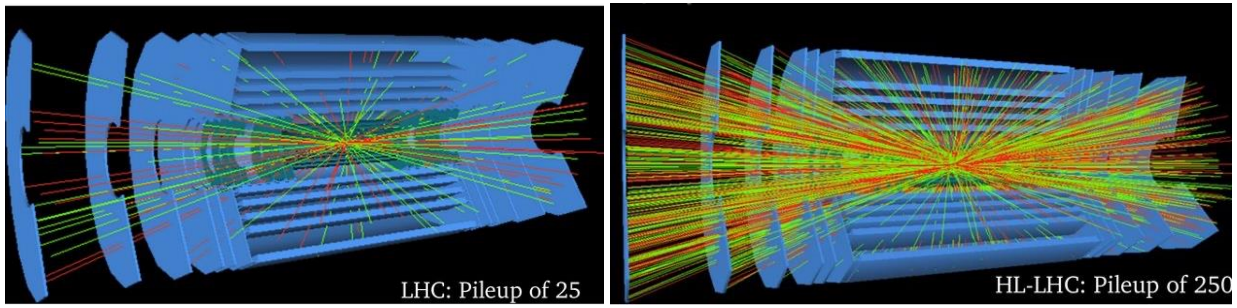
Silicon Pixel Detectors

- Experiments at colliders typically have a silicon pixel detector at the center
  - Concentric rings tiled with sensors
- Silicon sensors are depleted of charge carriers by high voltage
- When a charged particle from a collision passes through, it creates e/h pairs
- Charge is read out and transferred off-detector
  - Charge cluster information is used for physics analysis offline


Silicon Pixel Detector at CMS (LHC, CERN)



Case Study



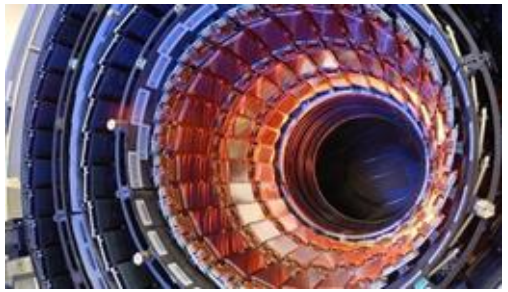
- > 90% of detector data originates from silicon pixels
- Pixel detector data rates continue to scale aggressively
- In 2018, CMS saw ~40 simultaneous pp collisions
  - The High Luminosity LHC will increase this to ~200
  - 5 times improved luminosity (radiation)
  - 7 times higher interaction rate (~3Ghit/cm2)

Technology	CMS Phase-2	Exploratory R&D smartpixels
Pixel size	**100x25**	**50x12.5 µm²**
Pixels #	157.6k	0.63M
Trigger rate	**750KHz**	**40MHz**
Readout data rate	1-4 links @ 1.28Gbps	Photonic link @ 30-100 Gbps
Radiation tolerance	**500Mrad**	**1Grad**
Power density	**1 W /cm²**	**1 W /cm²**

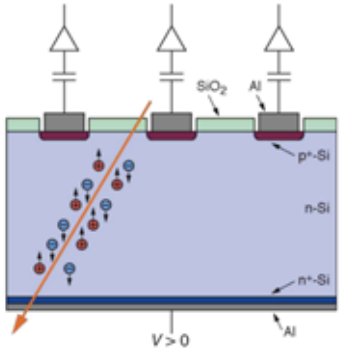
FERMI NATIONAL ACCELERATOR LABORATORY

High Level Proposed System

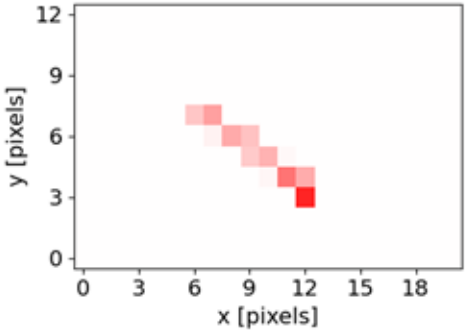
1. Collision



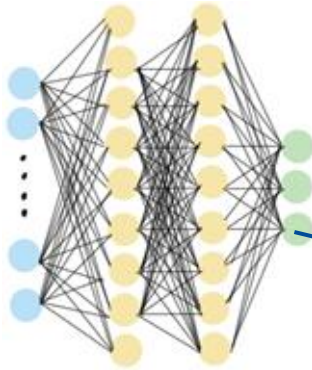
2. Sensor



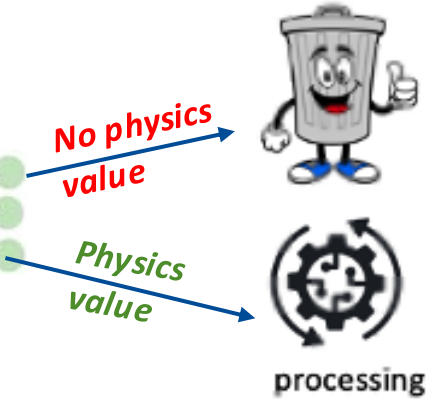
3. Cluster



4. ML model



5. Decision



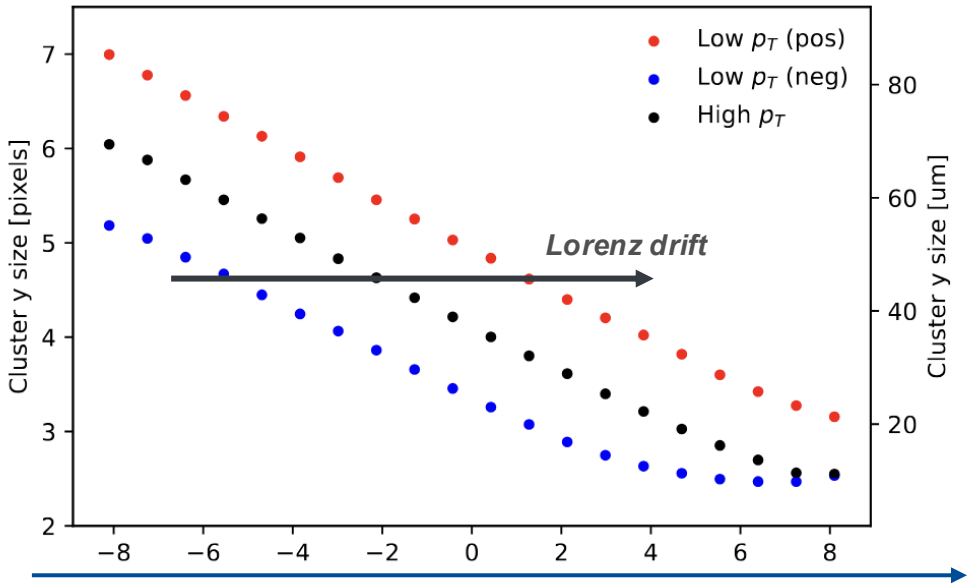
AI on chip

02

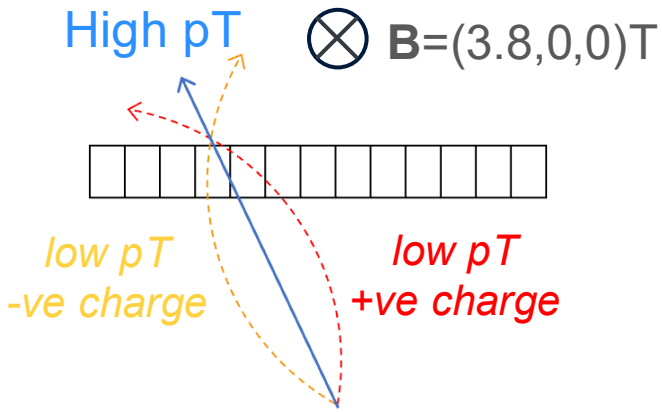
Filtering Machine Learning

Clusters Properties

Incident charge particle → deposits cluster of charge in pixel array
https://iopscience.iop.org/article/10.1088/2632-2153/ad6a00

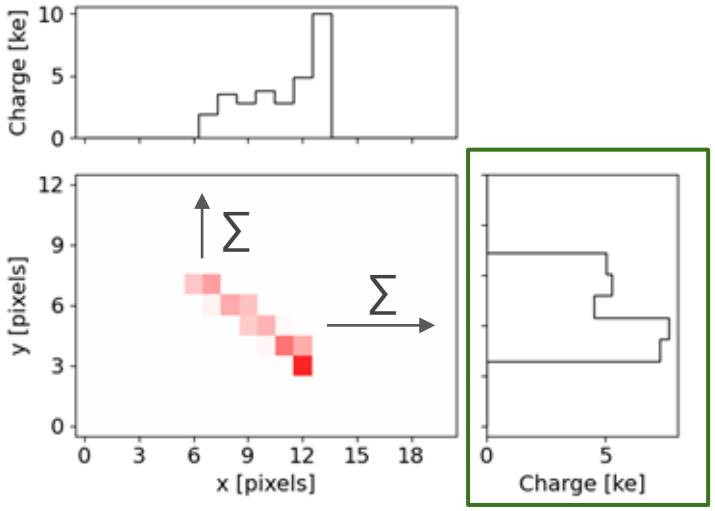


*y-direction - along the bending plane of magnetic field

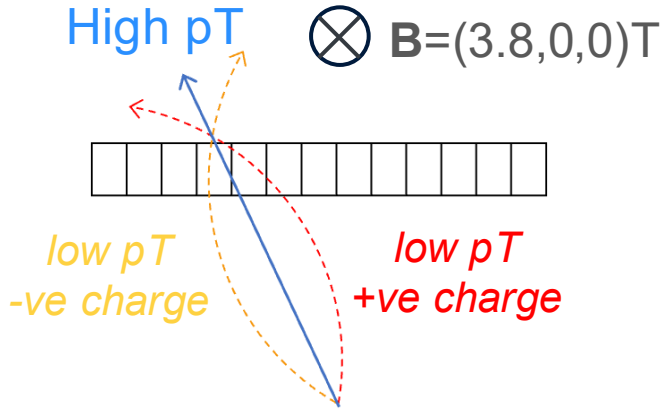


- ❑ Cluster information across the y-direction can be used to infer the momentum of the track
- ❑ Predicting pT from cluster morphology is not easily reducible to LUT logic

Filtering Neural Network

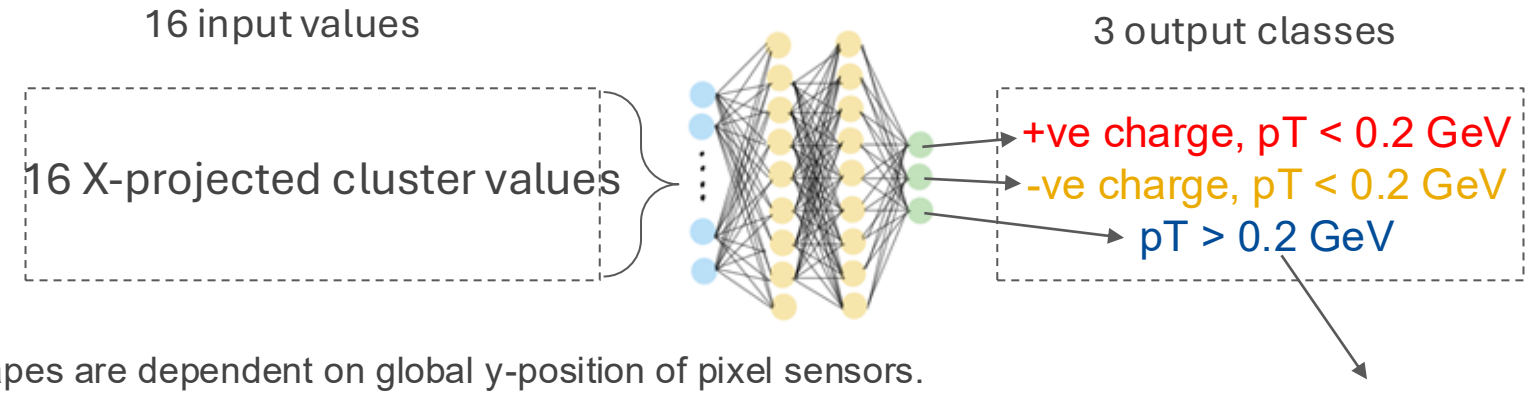


Sample event display at 4 ns



Quantity of interest:
Cluster profile along Y (sum of charge values over pixel rows after 4 ns)
Correlated with:

- y-position
- incident particle's pT



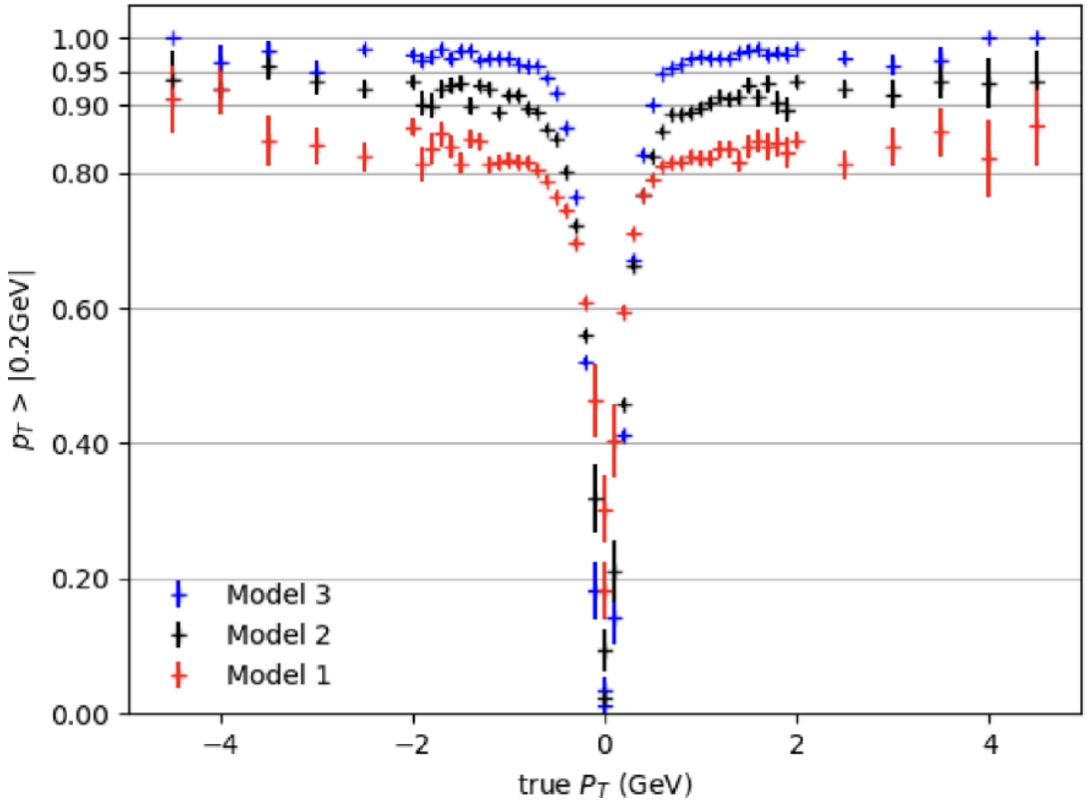
*The cluster shapes are dependent on global y-position of pixel sensors.

Separate weights and biases are trained for different y-local regions

First Step: p_T Filter at Full Precision

- Train fully-connected, feed-forward network
- Test selection of input features
  1. Cluster size (in bending direction) only
  2. Projected cluster shape, integrated over 4 ns
  3. Eight temporal "snapshots" of cluster shape in 200 ps intervals, showing evolution of cluster
- Key metrics:
  - **Efficiency** (tracks with true $p_T > 2$ GeV correctly identified as high p_T)
  - Background **rejection** (tracks with true $p_T < 2$ GeV correctly identified as low p_T)
  - Data **reduction** (all tracks classified as low p_T)

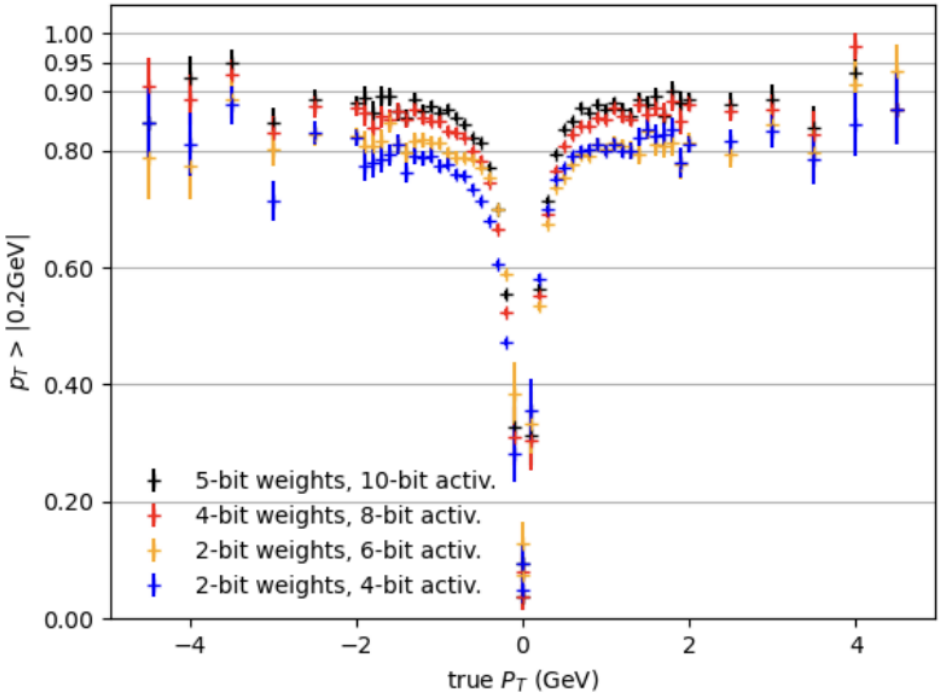
Model	Sig. efficiency*	Bkg. rejection*
Model 1	84.8%	26.6%
Model 2	93.3%	25.1%
Model 3	97.6%	21.7%



Model 2 selected as benchmark for compromise between performance and simplicity

** Produced in 2024 with preliminary dataset*

Digitization And Simulated Performance



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

Table 3: Mapping between 2-bit ADC output and collected charge.

- *Charge digitalization: 2-bit "flash" ADC*
- *Weight/activation quantized to 4-bit and 8-bit respectively*

$$\textbf{Data reduction} = 0.4 \times \textbf{bkg reduction} + 0.6 \times \textbf{untracked data reduction}$$

- Untracked data reduction (single-pixel hits (noise), loopers) estimated between 61to 100%
- Bkg. reduction > 20%
- Total data reduction estimated between 54 to 75%

https://iopscience.iop.org/article/10.1088/2632-2153/ad6a00

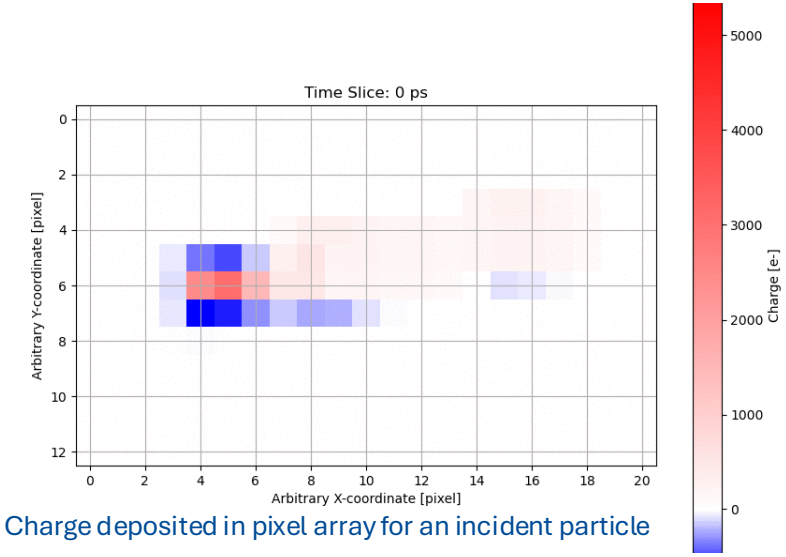
FERMI NATIONAL ACCELERATOR LABORATORY

Training Data: Simulation

- 50 x 12.5 μm^2 pitch pixels (100 μm thickness)
- Dataset extracted for 16x16 pixel array.
- Electric field maps of pixel sensor are generated using TCAD simulation.
- Field maps are imported into PixelAV
- Provide realistic detector response
- 200ps time slice
- Public dataset !

A detailed simulation tool for silicon detectors

★ Datasets now made available publicly on zenodo!
https://zenodo.org/records/17180303
https://zenodo.org/records/18472791



Charge deposited in pixel array for an incident particle

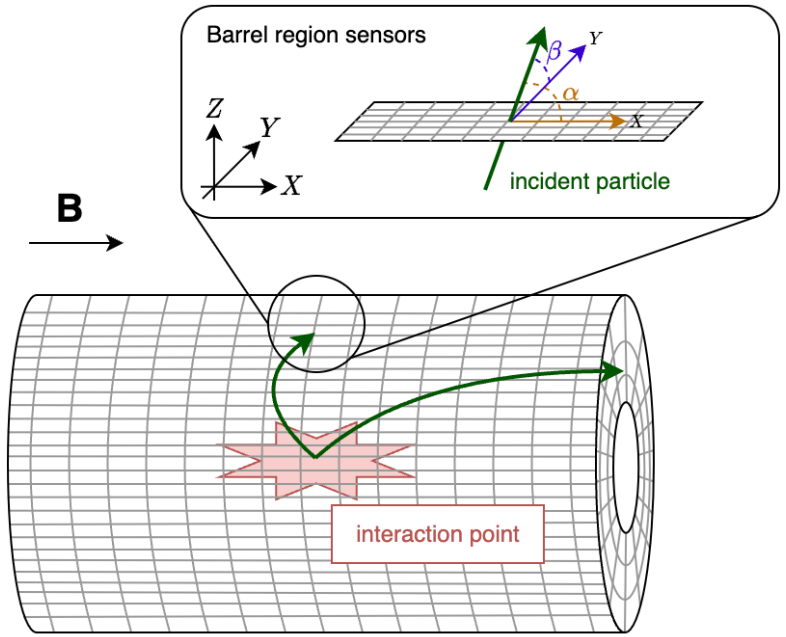
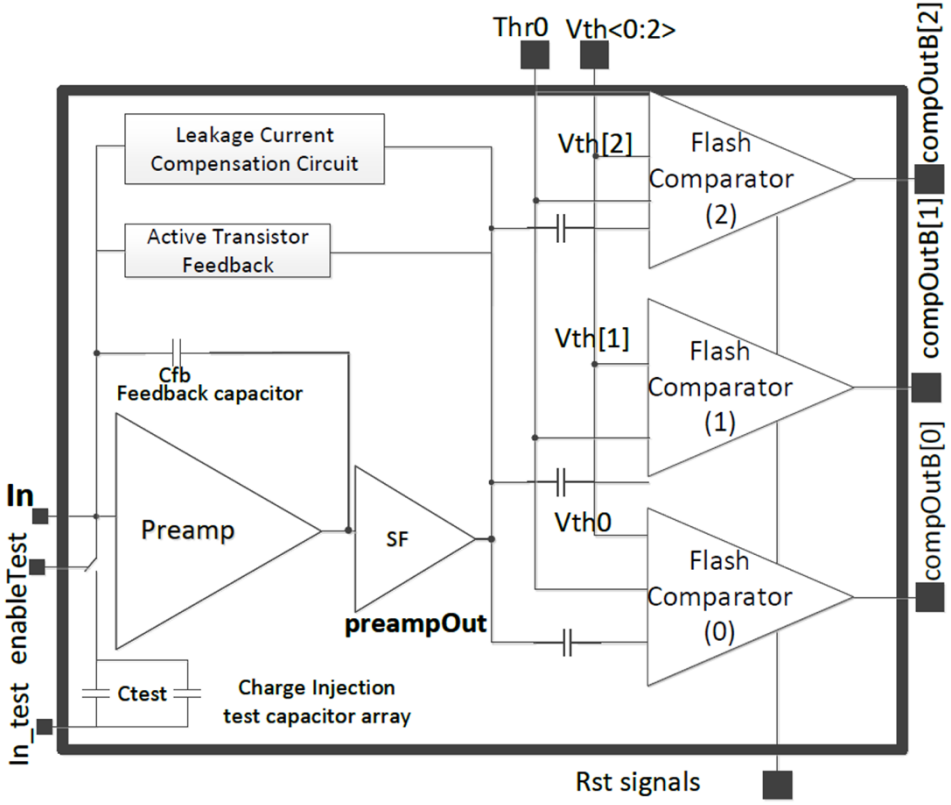


image credits: Danush Shekar

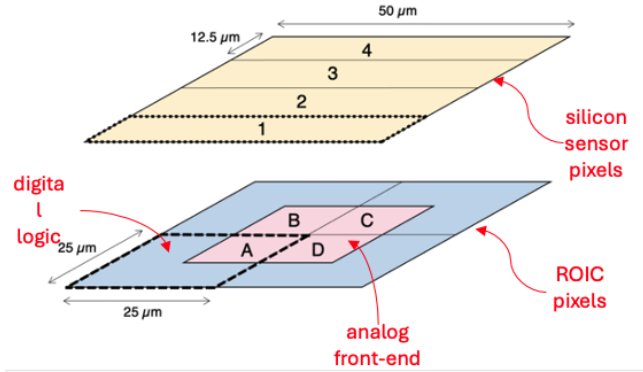
03

Front-End Implementation

Front-End Pixel Architecture: Synchronous ADC



Detail design here



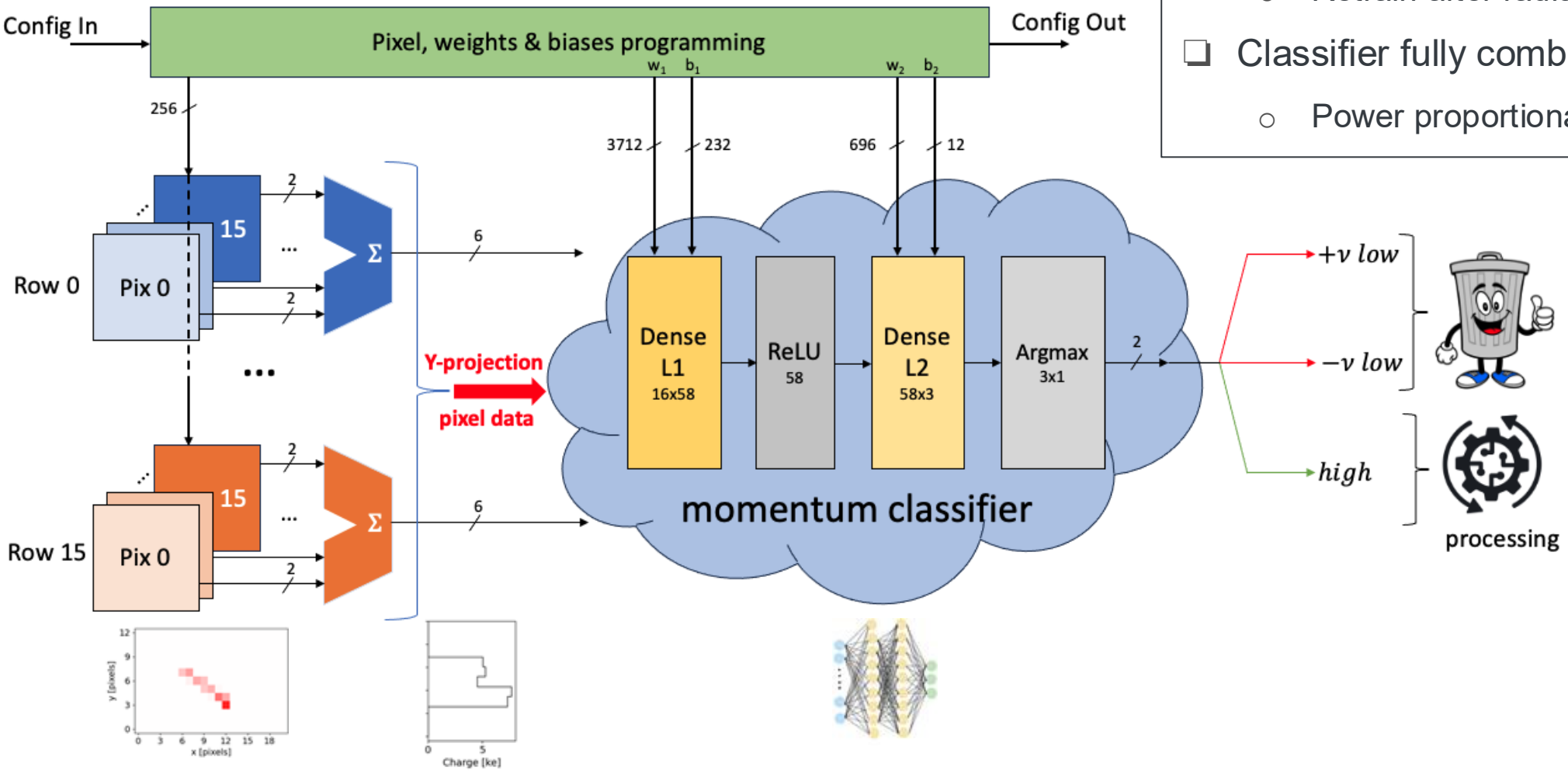
- 25μm × 25μm pixel sizes
- <1W/cm^2 Power density
- CSA architecture inspired by FCP130 (not Krummenacher)

  Comparison study here
- AC coupled 40MSPS in-pixel 2-bit flash ADC
  - Auto-Zero (AZ) in every pixel for threshold correction
  - Insensitive to pile-up
  - 2 architectures tested
- **Global** charge injection site with **programmable** capacitive array
- No sensor bonded to this prototype

Implementable Network

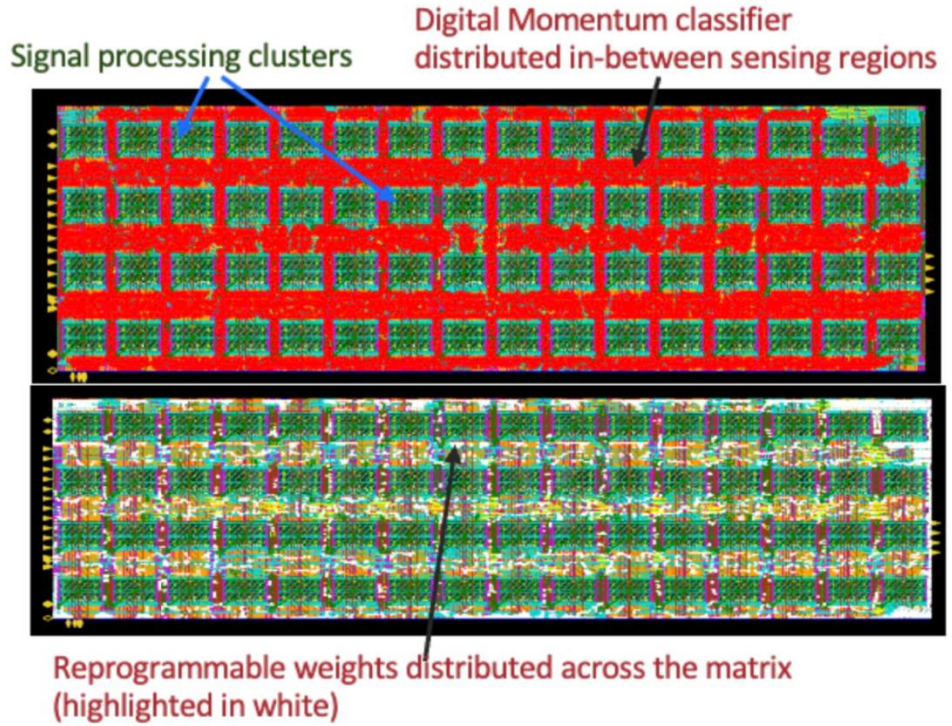
Key Take-Away:

- ❑ Reprogrammable Network
  - Region specific training
  - Improve resilience against faulty pixels
  - Retrain after radiation damage
- ❑ Classifier fully combinatorial
  - Power proportional to hit density



First *Smartpixels* Tapeout With Network

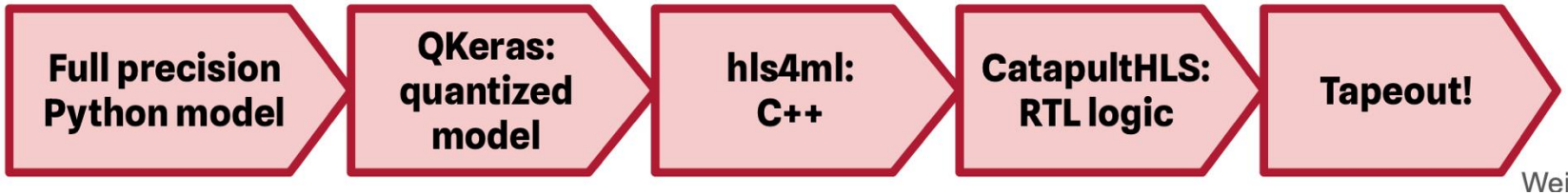
ASIC synthesis



Established pipeline to translate quantized python level model to RTL design

- hls4ml converts QKeras model to C++
- Catapult HLS converts C++ to RTL
- Estimates model footprint (mm^2), power consumption

Balancing performance, area, power



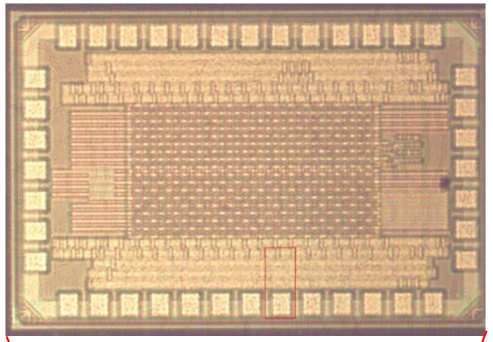
Weiss – FastML 13

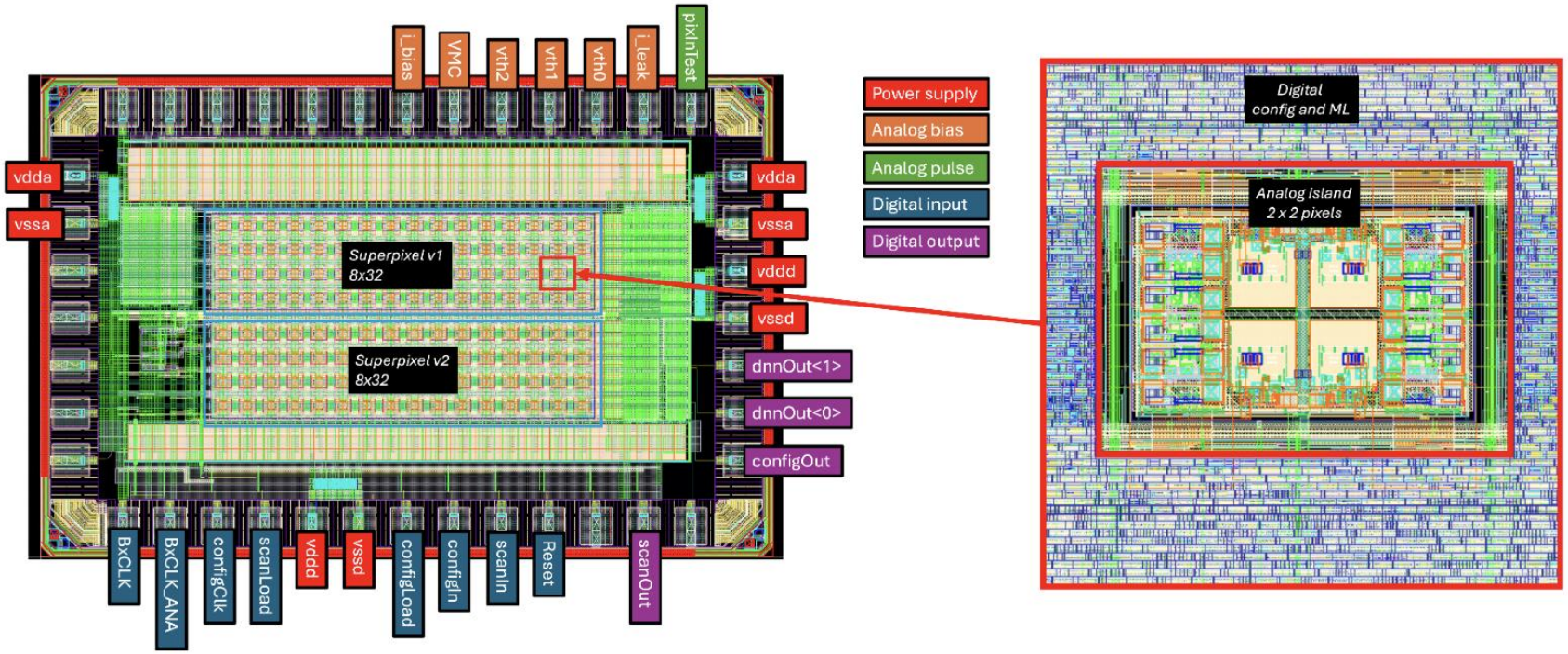
04

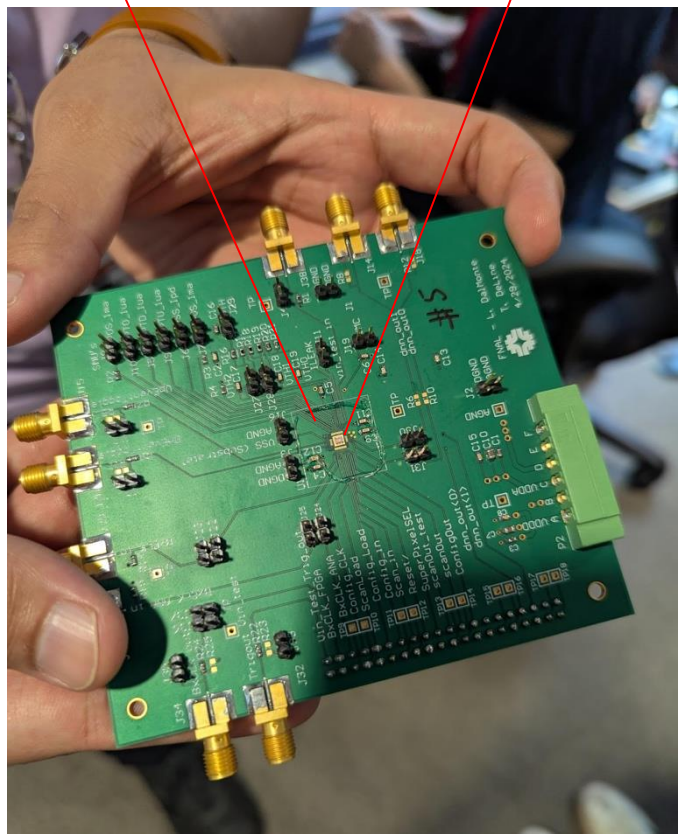
Test Result

First Prototype

- 1.5mm^2 die – tsmc28
- *Pixel Analog Front End Prototype v1 with* ***digital ML Filter***
- two16x16 (sensor side) superpixel variants
- Mixed signal implementation but still analog on top

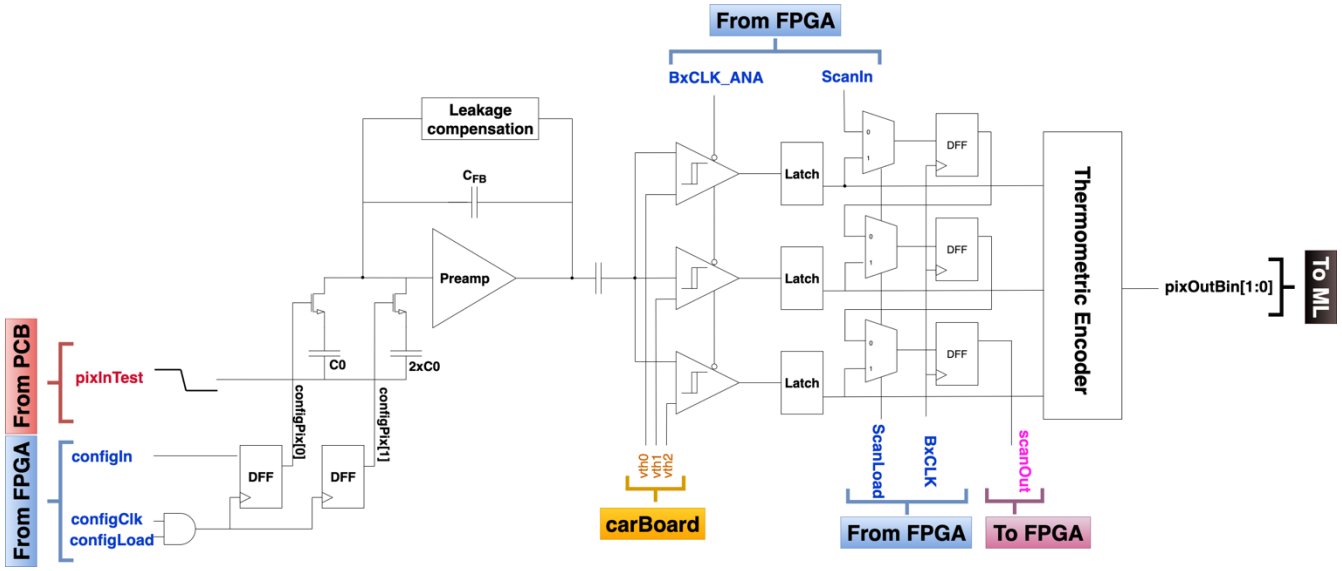


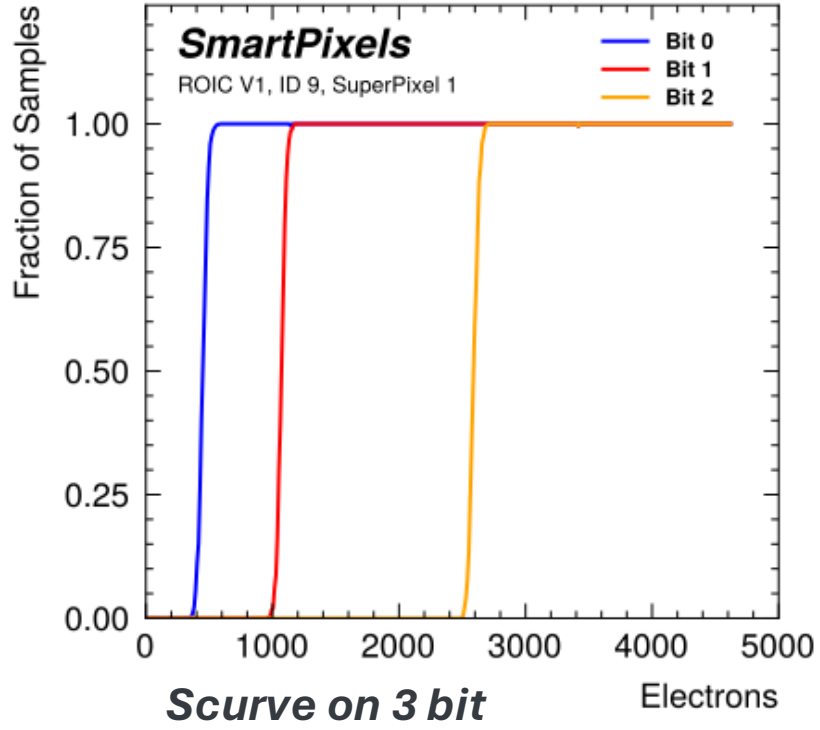




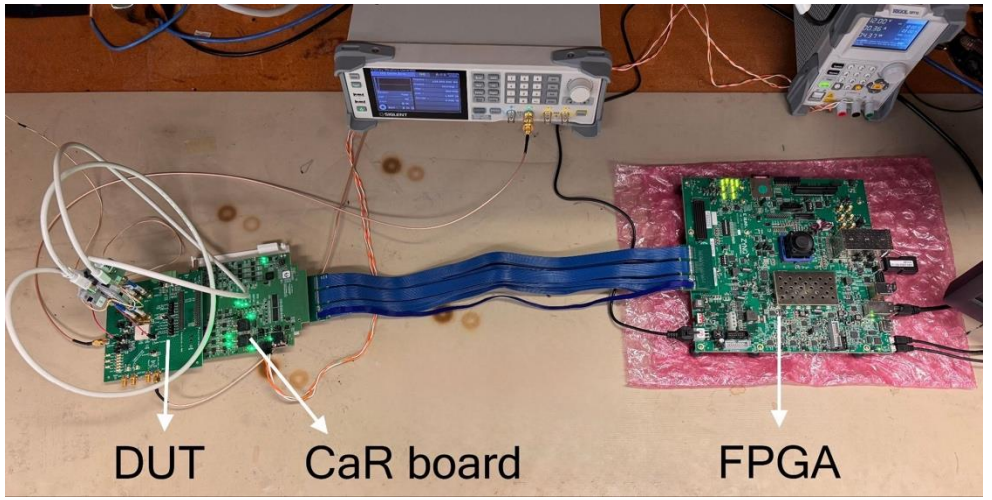
ASIC Test Stand

https://arxiv.org/abs/2510.07485





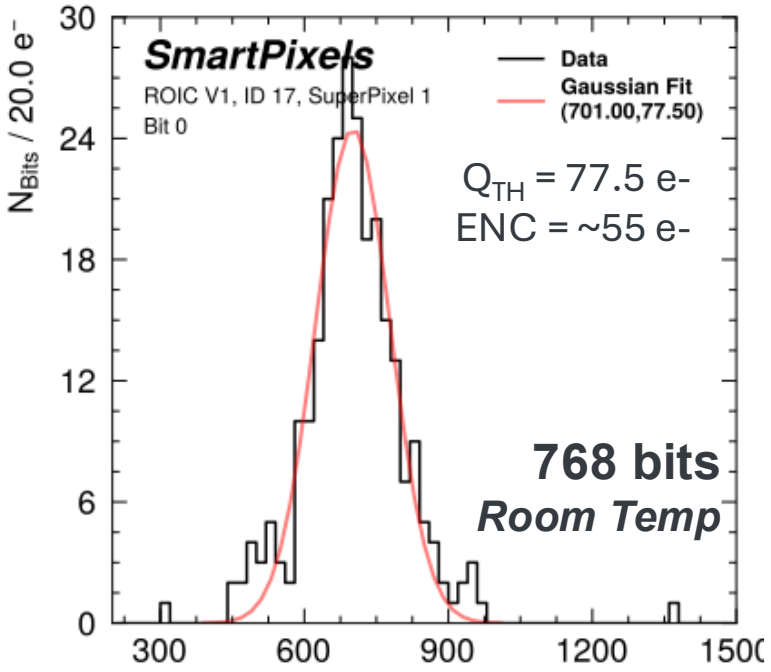
- Lightweight test stand duplicated at 3 universities
- Pixel interface with DAQ and FPGA validated
- Three comparators implement 2-bit flash ADC
- Per-pixel S-curves extracted across matrix
- Threshold and ENC characterization



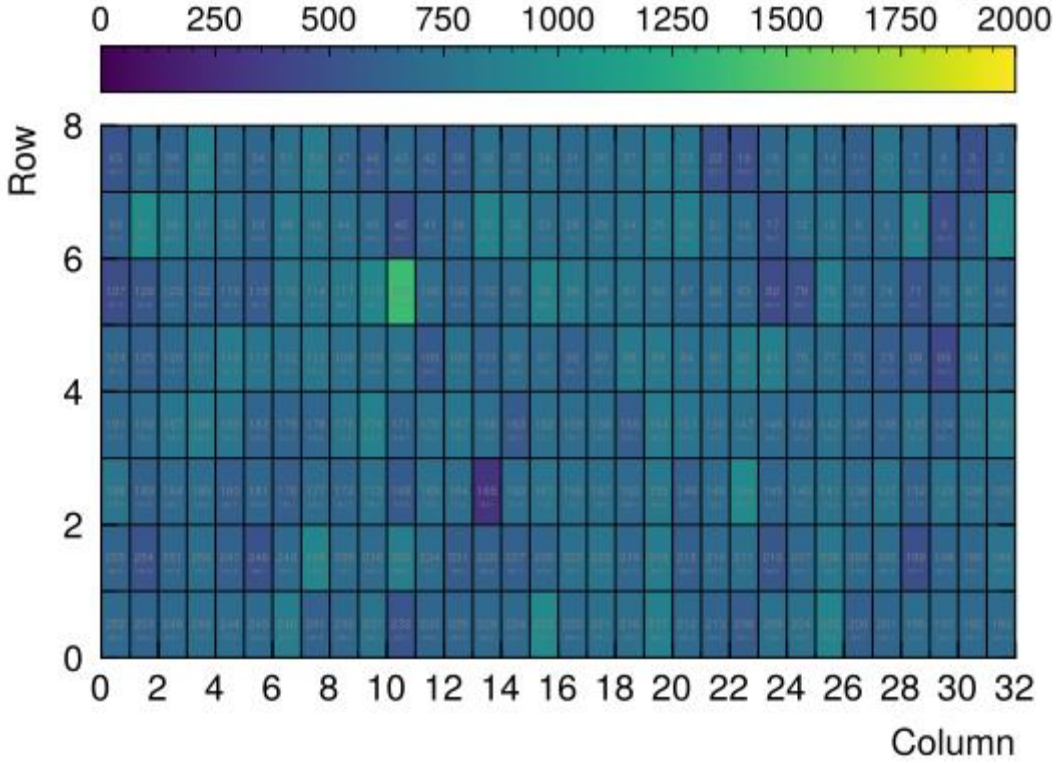
FERMI NATIONAL ACCELERATOR LABORATORY

Threshold Dispersion Characterization

Threshold Dispersion



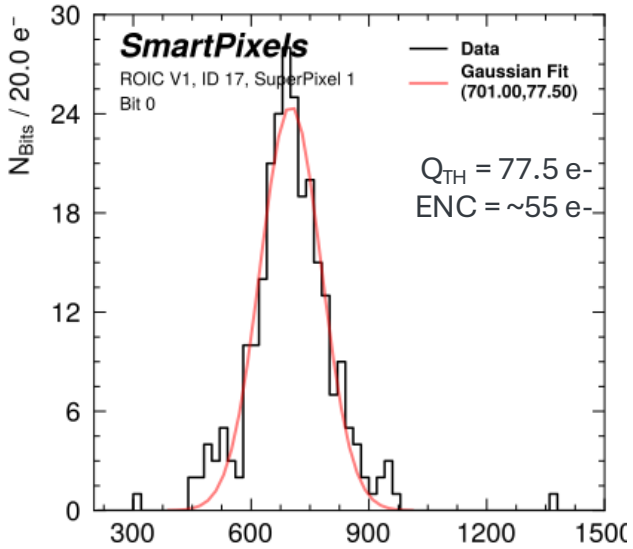
Threshold heat map



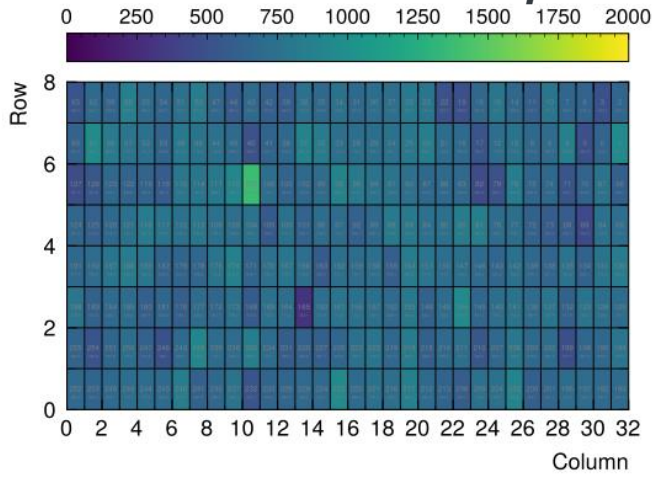
- ❑ Measured charge dispersion and ENC larger than expected from simulation
- ❑ Total charge dispersion $\sqrt{ENC^2 + Q_{TH}^2} = 95e -$
- ❑ Observed significantly larger spatial variation than expected from intrinsic front-end mismatch alone

Characterization Non-Idealities Identified

Threshold Dispersion

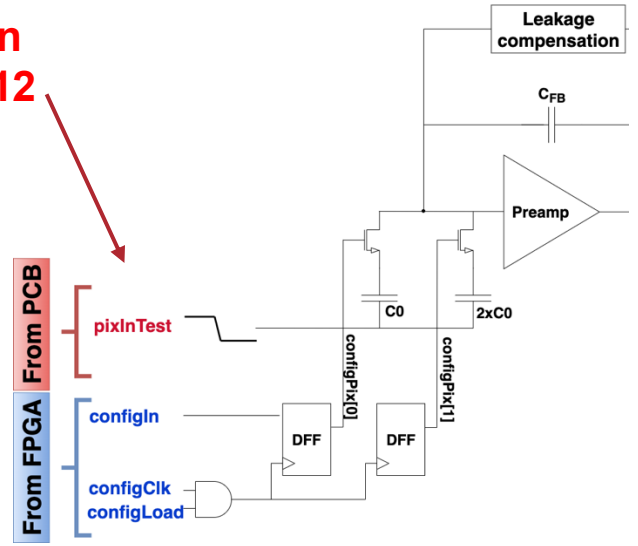


Threshold heat map



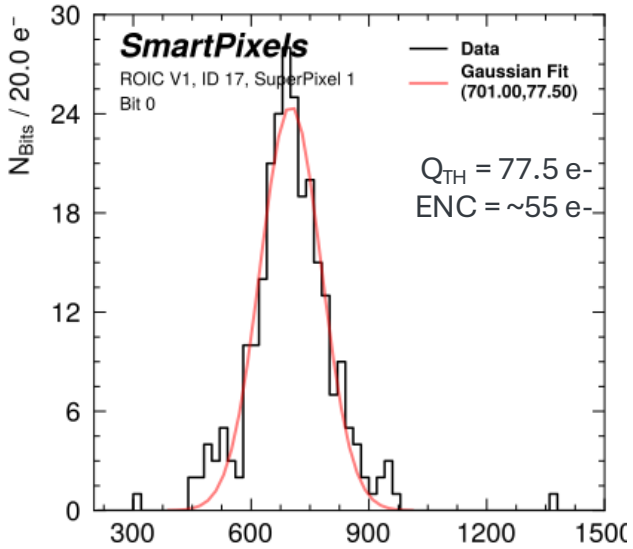
1. Global injection network effect
   - Distributed RC degradation on global injection grid
   - Spatially varying injected charge amplitude
   - Artificial shift in extracted thresholds

Global Injection connected to 512 pixels

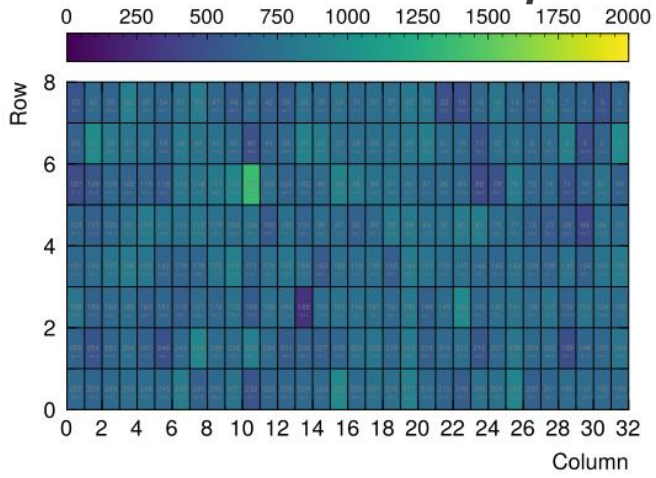


Characterization Non-Idealities Identified

Threshold Dispersion



Threshold heat map



1. Global injection network effect
   - Distributed RC degradation on global injection grid
   - Spatially varying injected charge amplitude
   - Artificial shift in extracted thresholds

2. Threshold-bias sensitivity
   - Shared VTH bias grid affected by high leakage (>uA)
   - Leakage creates systematic offsets across both matrices

✓ Spatial pattern likely reflects combined injection + bias-grid effects

✓ Effect dominated by characterization infrastructure

✓ Fix implemented in future ASIC iteration

Key question:
How much do these front-end non-idealities affect ML inference?

Measured Classification Performance

Comparison of software, quantized, and measured ASIC inference

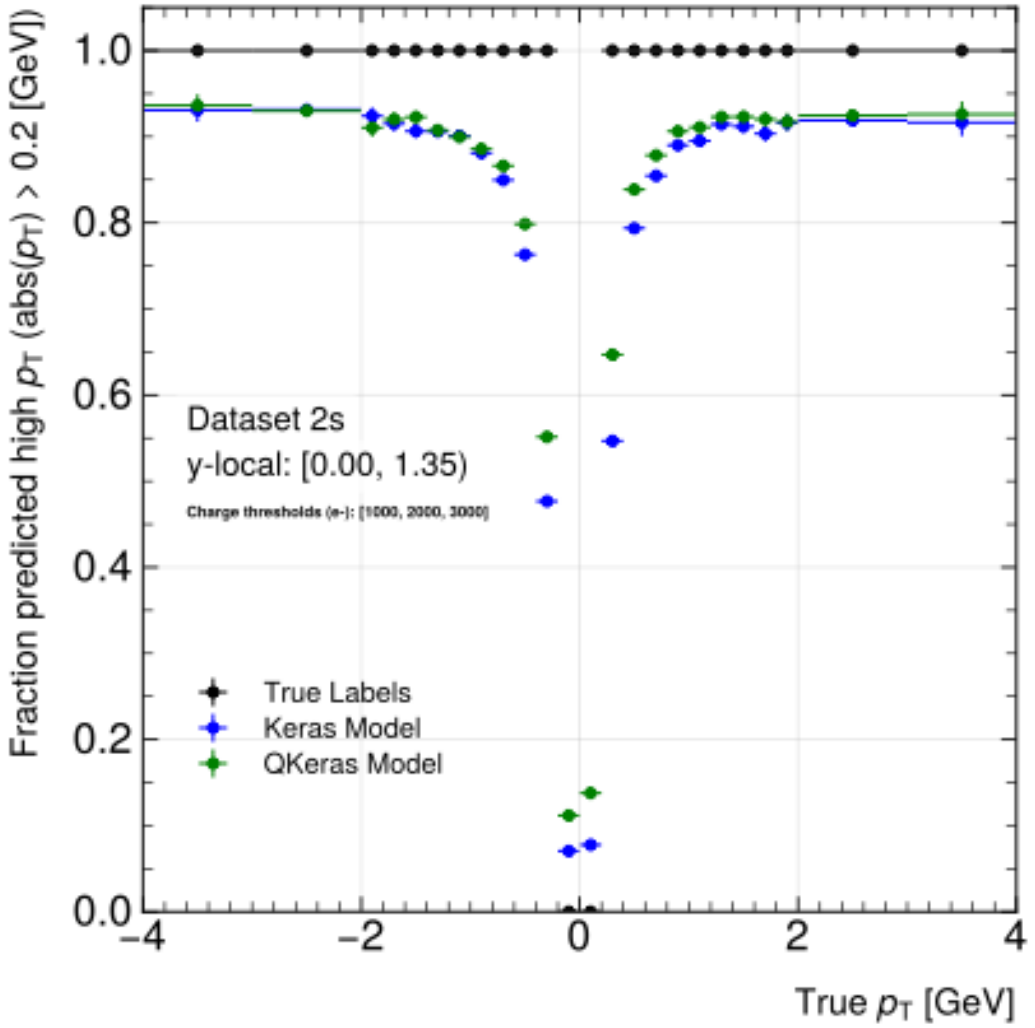
- On-chip classifier evaluated using measured front-end response
- Region-specific weights trained for y-local range [0.0, 1.35) mm

	Sig. efficiency	Bkg rejection
Full precision model	**~92**	**42**
Quantized model	**~92**	**40**

Key metrics:

Signal Efficiency (tracks with true $p_T > 2$ GeV correctly identified as high p_T)

Bkg rejection (tracks with true $p_T < 2$ GeV correctly identified as low p_T)

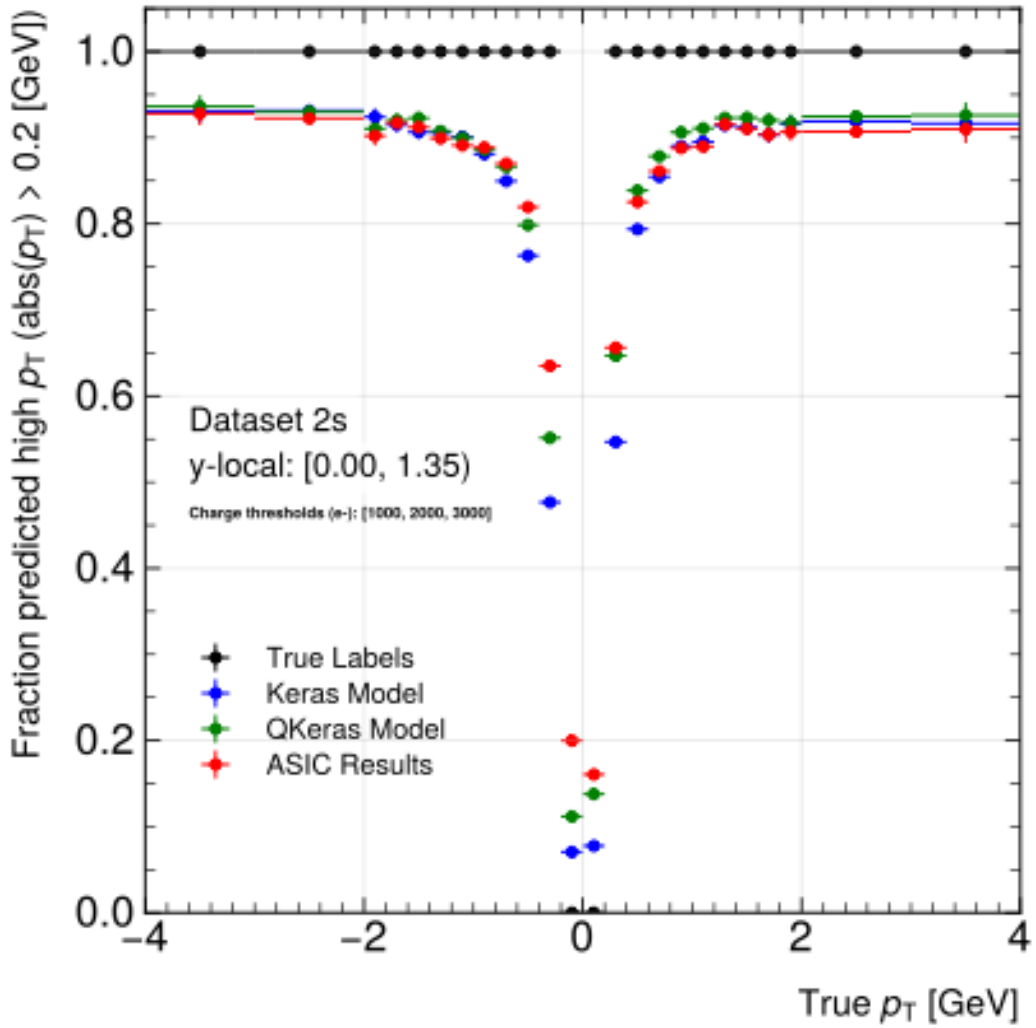


Measured Classification Performance

Comparison of software, quantized, and measured ASIC inference

- On-chip classifier evaluated using measured front-end response
- Region-specific weights trained for y-local range [0.0, 1.35) mm
- 2% loss of performance with threshold set at 10σ

	Sig. efficiency	Bkg rejection
Full precision model	**~92**	**43**
Quantized model	**~92**	**38**
ASIC @1000e-	**~90**	**36**



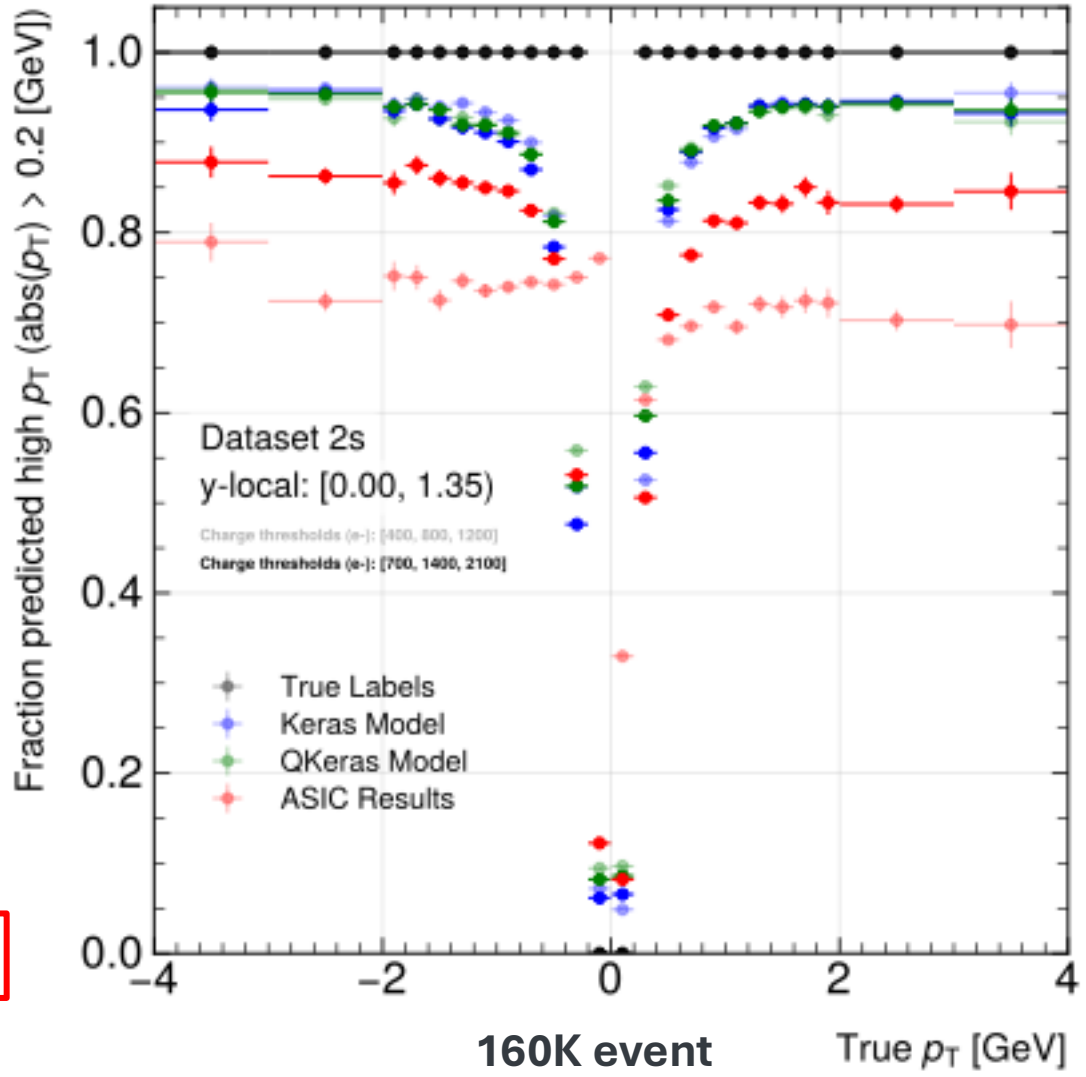
Measured Classification Performance

Comparison of software, quantized, and measured ASIC inference

- On-chip classifier evaluated using measured front-end response
- Region-specific weights trained for y-local range [0.0, 1.35) mm
- 10% loss of performance with threshold set at 7σ
- 23% loss of performance with threshold set at 7σ

	Sig. efficiency	Bkg rejection
Full precision model	**~94**	**42**
Quantized model	**~94**	**40**
ASIC @ 700e-	**~84**	**44**
ASIC @ 400e-	~71	33

Noise level directly impact the signal efficiency

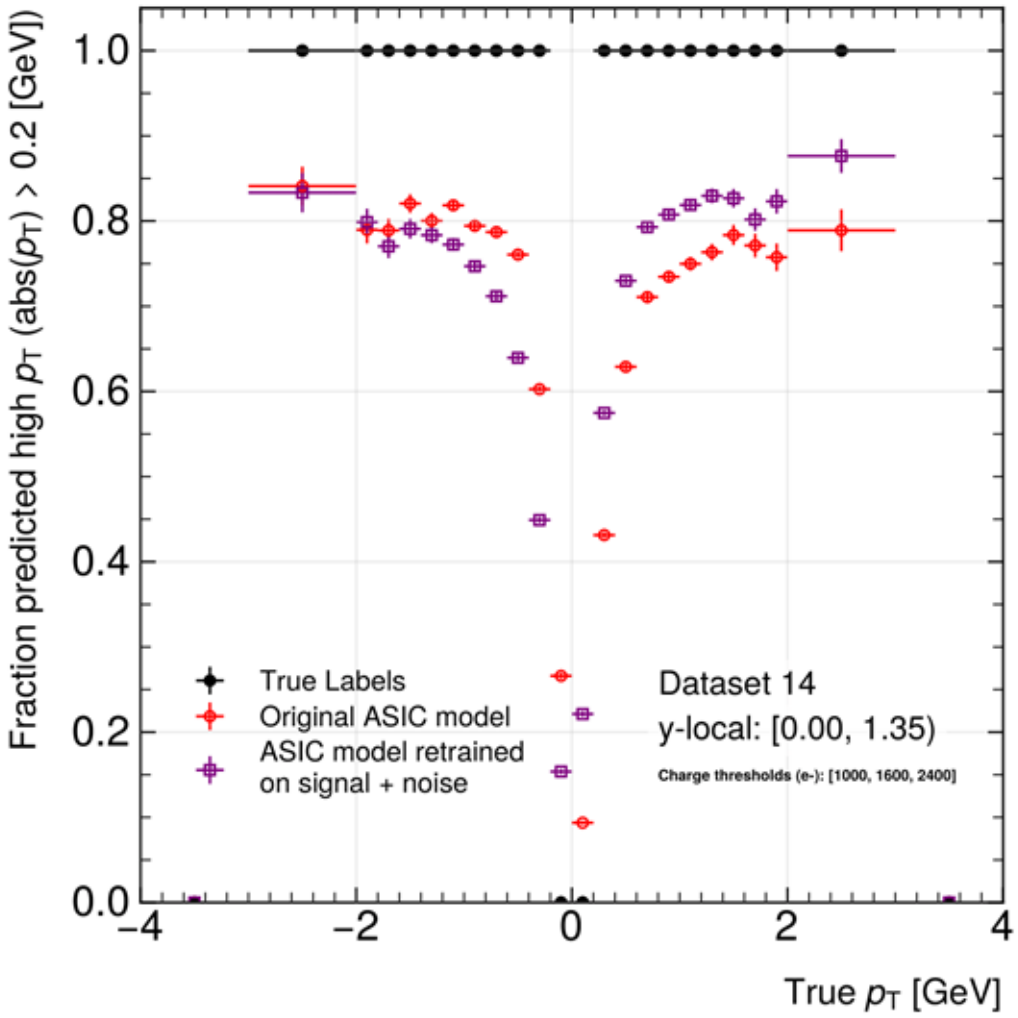


Promises On Noise Retraining

smartpixel CPAD 2025

- The model was re-trained with noise-injected simulation data
- Retraining with realistic noise recovers substantial efficiency
- Same Hardware, updated weights and biases
- Flow is complex and mostly manual pipeline is a work in progress

Training Model *In ASIC @1000e-*	Sig. efficiency	bkg rejection
signal	78.91 %	45.42 %
signal & noise	87.64 %	45.33 %



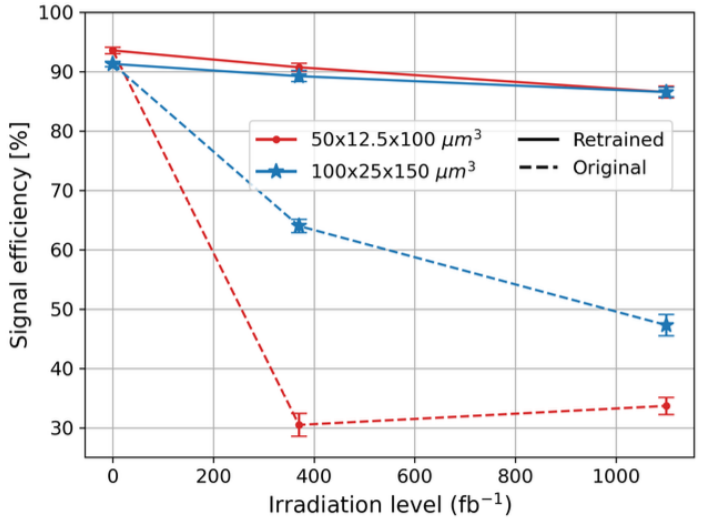
Performance at Room Temperature

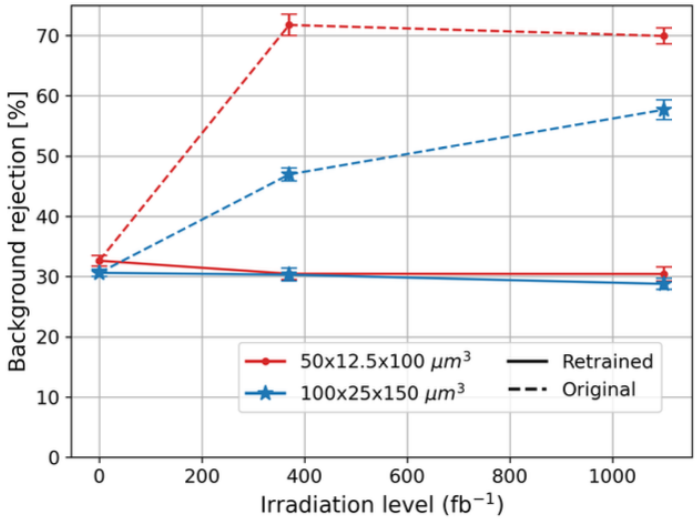
Specification	Target	Measured	Units
Analog power	< 5	3.1	*uW/pixel*
Digital power	300	243	*uW*
Power density	<1	0.65	W/cm^2
ENC rms ***(no sensor)***	40	55 [2] 30 to 70	*e-*
Threshold dispersion Q_{TH}	< 40	70 [1,2]	*e-*
Total Charge Dispersion $\sqrt{ENC^2 + Q_{TH}^2}$	<60	95 [1,2]	*e-*
Signal efficiency	> 90	71 to 90	*%*
Background rejection ***Low pT from tracked dataset***	>20	39	*%*
Data reduction ***All low pT event***	> 50		*%*

1. Leakage on bias grid
2. Global Charge injection

Incoming Test: Radiation Damage Correction

- HL-LHC will reach unprecedented radiation levels, highest in barrel pixel layer 1, 3 cm from collisions
- Generate datasets with simulated radiation damage (models from PixelAV)
  - 3.3 x 10^{15} n_{eq}/cm^2 (370 fb^{-1} (Run 3) at Layer 1)
  - 1.0 x 10^{16} n_{eq}/cm^2 (1100 fb^{-1} (half of HL-LHC) at Layer 1)
- Retraining (new weights, same structure) recovers most performance for both benchmark and HL-LHC sensor geometry, and will be possible for an eventual detector





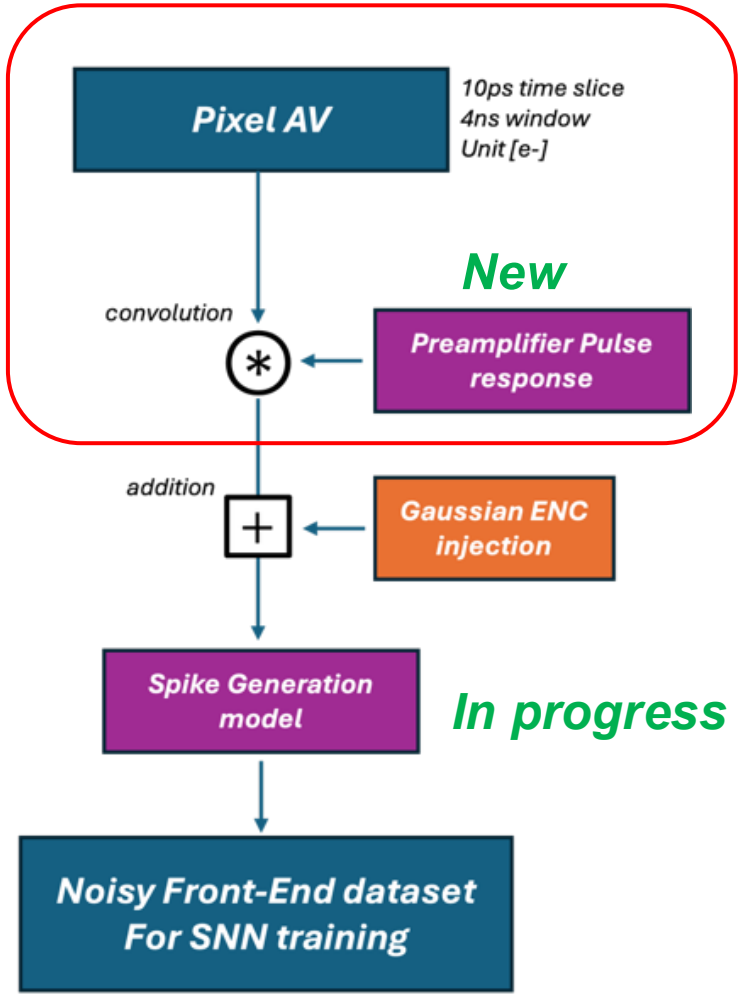
https://arxiv.org/abs/2510.06588
Under review NIM-A

- Next step: leverage leakage current injection to mimic effect of radiation and retrain the ASIC to recover performance

05

Status And Next Steps

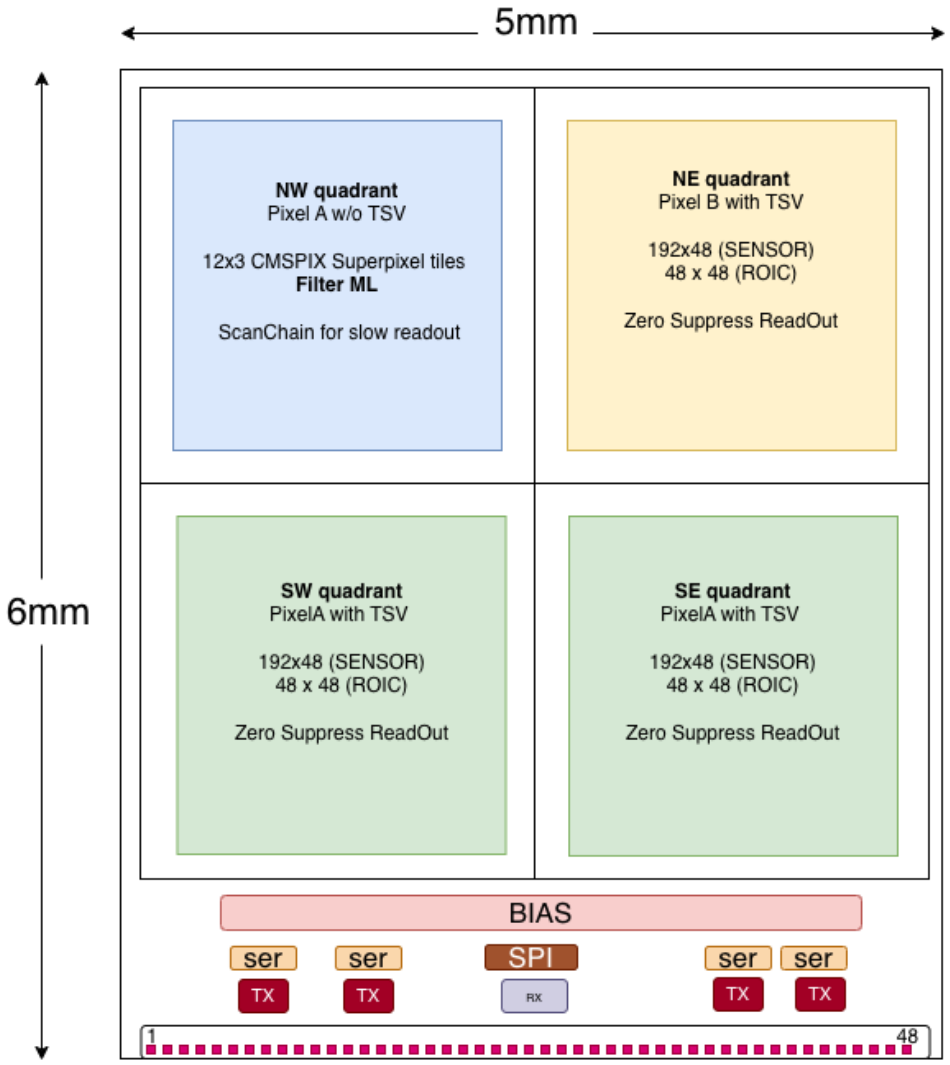
Future Dataset Productions



Envisioned dataset production flow

- Current filtering network is mostly time-agnostic
- Regression and neuromorphic ML need **accurate** timing modeling
- We are upgrading our dataset flow
  - CSA model **ready** and predicts rise time and saturation behavior correctly
  - ADC and spike generation timing modeling is ***a work in progress***
  - ENC noise injection is ***next*** to train our ML with realistic ASIC conditions

VIZARD 2026



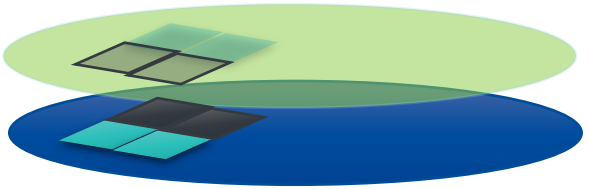
Status

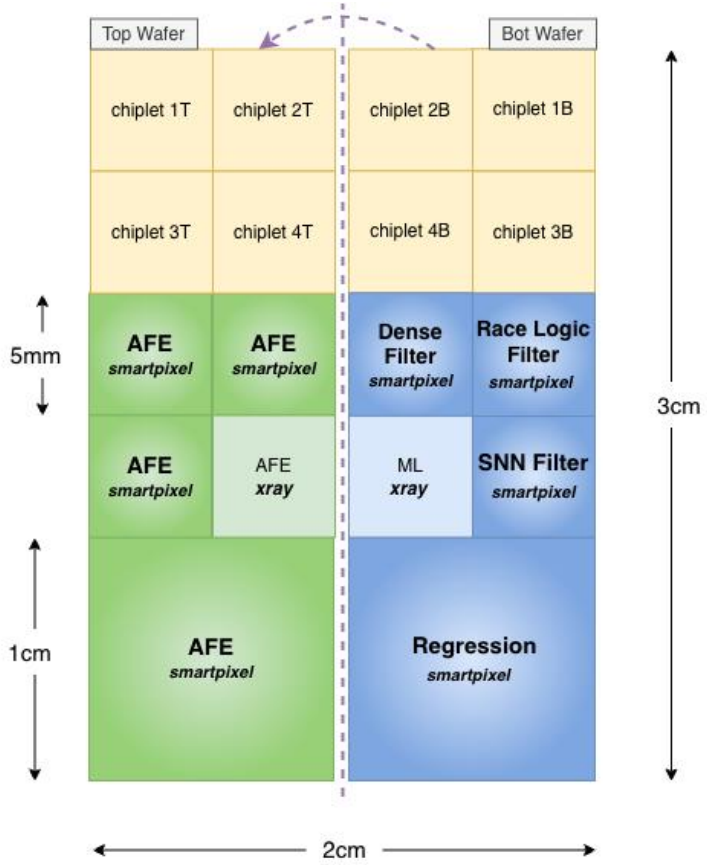
- ❑ ***Tapeout**: End 2026*
- ❑ ***Engineer run with CERN***
- ❑ ***Chip delivery**: Early 2027*

Overview

- ❑ Two 30mm2 chips
- ❑ VIZARD_v0a - backside sensor connection (TSVs)
- ❑ VIZARD_v0b – frontside sensor connection (Bump bond)
- ❑ Improved and novel AFE architecture
- ❑ Zero suppress readout on three 192x48 pixel matrices quadrant
- ❑ Implementation of 36 digital filter ML (NW quadrant)
- ❑ SPI interface for configuration and slow readout
- ❑ High Speed RX/TX interface (1.28Gbps)

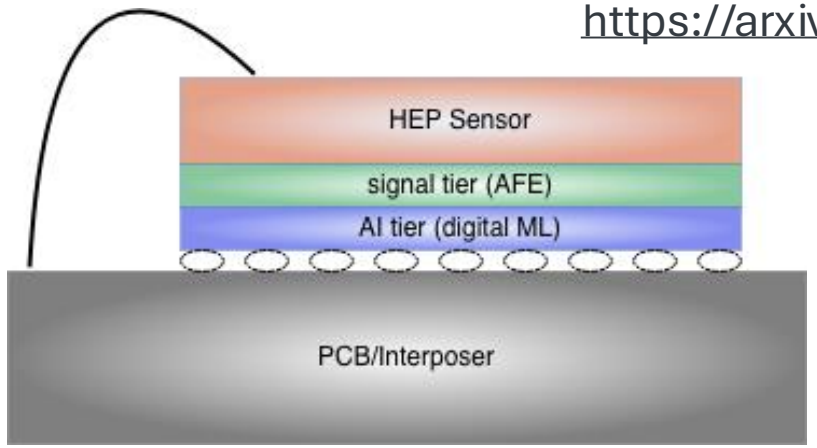
3D chip tapeout





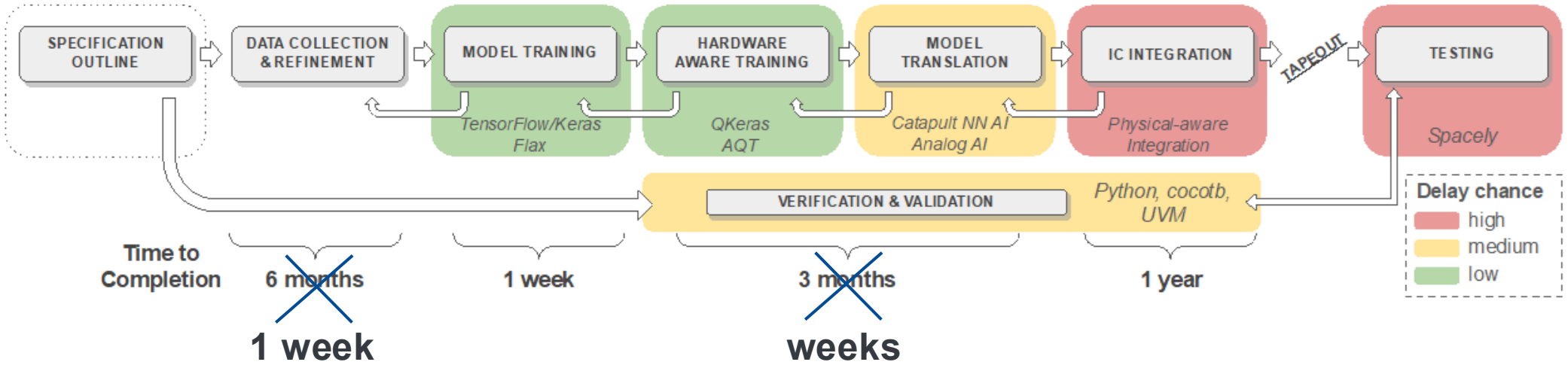
- 3D integration planned in January 2027
- 8 chiplets design dedicated for ***smartpixels***
  - Improved and novel AFE architecture
  - Smarter digital auto zero cycle control
  - Explore novel Filter ML architectures (race logic, SNN)
  - Regression algorithms for real-time physics feature extraction
- Fabrication and test planned for 2027-2028

https://arxiv.org/abs/2602.15946



3D detector product

Codesign Flow



- Dataset production accelerated dramatically
- Silicon Realistic Dataset including AFE timing and noise modeling
- ML-to-ASIC flow improved Reduced HLS4M
  - Tight integration of HLS4ML with siemens catapult
  - We produced with siemens a more automatized flow that works for ASIC
- Hardware iteration remains the bottleneck

A chip result is worth a thousand simulations

The *Smartpixels* Team

Fermi National Accelerator Laboratory: Abhijith Gandrakota, Benjamin Parpillon, Chinar Syal, Douglas Berry, Farah Fahim, Gauri Pradhan, Giuseppe Di Guglielmo, James Hirschauer, Jennet Dickinson, Lindsey Gray, Nhan Tran, Ron Lipton

Cornell University: Jennet Dickinson, Ben Weiss

Johns Hopkins University: Dahai Wen, Morris Swartz, Petar Maksimovic

Northeastern University: Nick Manganelli

Northwestern University: Manuel Blanco Valentin

Oak Ridge National Laboratory: Aaron Young, Shruti R. Kulkarni

Purdue University: Mia Liu, Arghya Das

University of Chicago: Karri DiPetrillo, Anthony Badea, Carissa Kumar, Emily Pan, Rachel Kovach-Fuentes, Aidan Nicholas, Eliza Howard, Eric You

University of Colorado Boulder: Jannicke Pearkes, Ricardo Silvestre

University of Illinois Chicago: Corrinne Mills, Danush Shekar, Jieun Yoo, Mohammad Abrar Wadud

University of Illinois Urbana-Champaign: Mark S. Neubauer, David Jiang

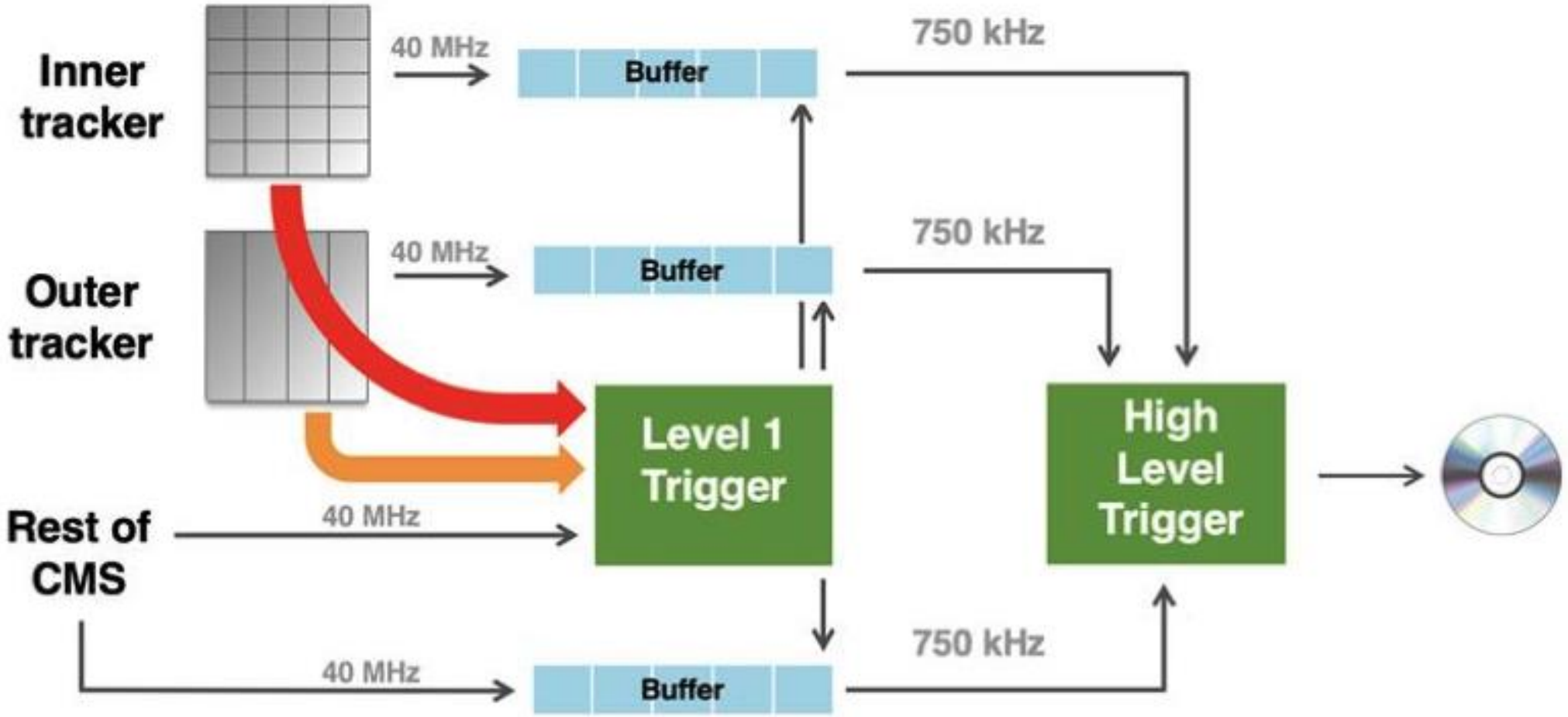
University of Kansas: Alice Bean

FERMI NATIONAL ACCELERATOR LABORATORY

06

Backup

Smartpixel In The Trigger



Filtering neural network: performance metrics

Physics-relevant quantities defined to infer the model's accuracy to select tracks with pT > 2 GeV, and quantify the fraction of rejected data:

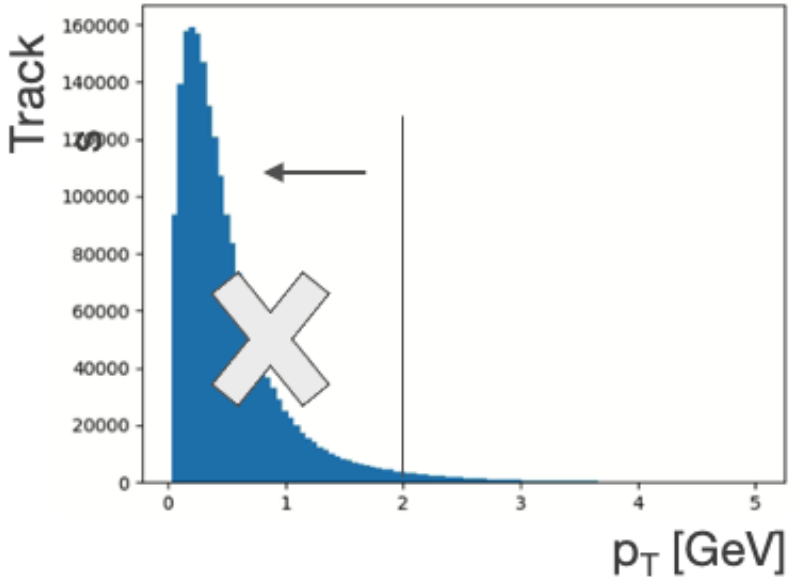
(2 GeV threshold as it is relevant for physics analysis)

$$\text{Signal efficiency} = \frac{\text{Num. of clusters classified as pT} > 2\text{ GeV}}{\text{Num. of clusters with pT} > 2\text{ GeV}}$$

Target signal efficiency: > 90%

$$\text{Data reduction} = \frac{\text{Num. of clusters classified as pT} < 2\text{ GeV}}{\text{Total num. of clusters}}$$

Target data reduction: > 50%



★ Signal efficiency is not to be confused with hit-efficiency (ratio of measured to true/expected number of hits when particles traverse through the detector)

Data reduction

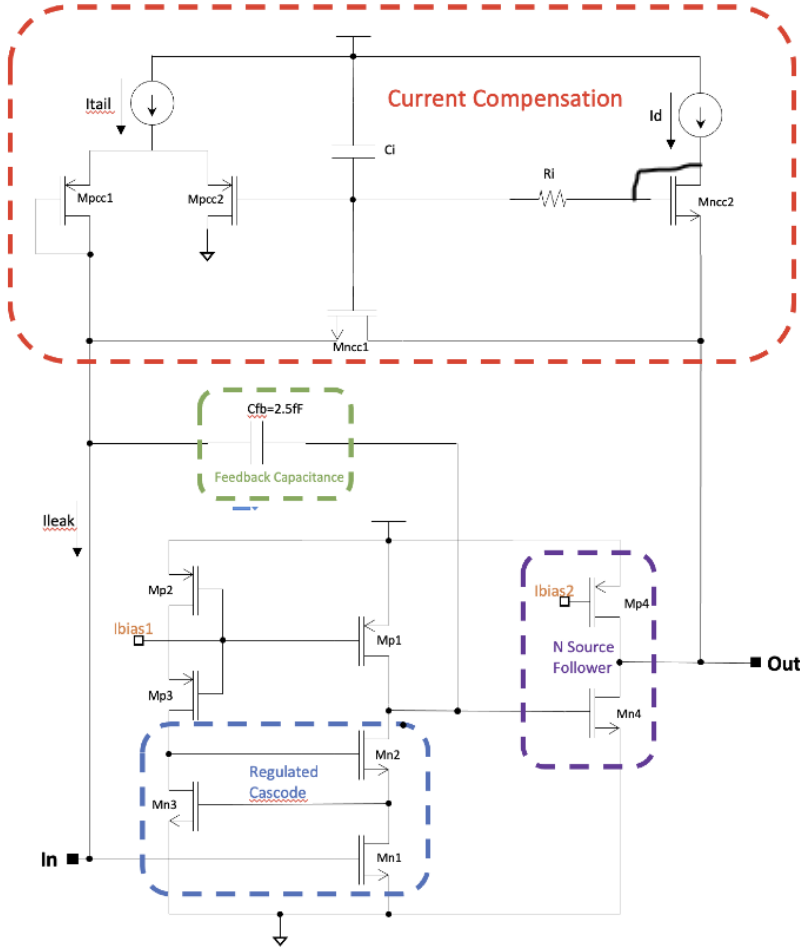
Dataset breakdown seen in the detector

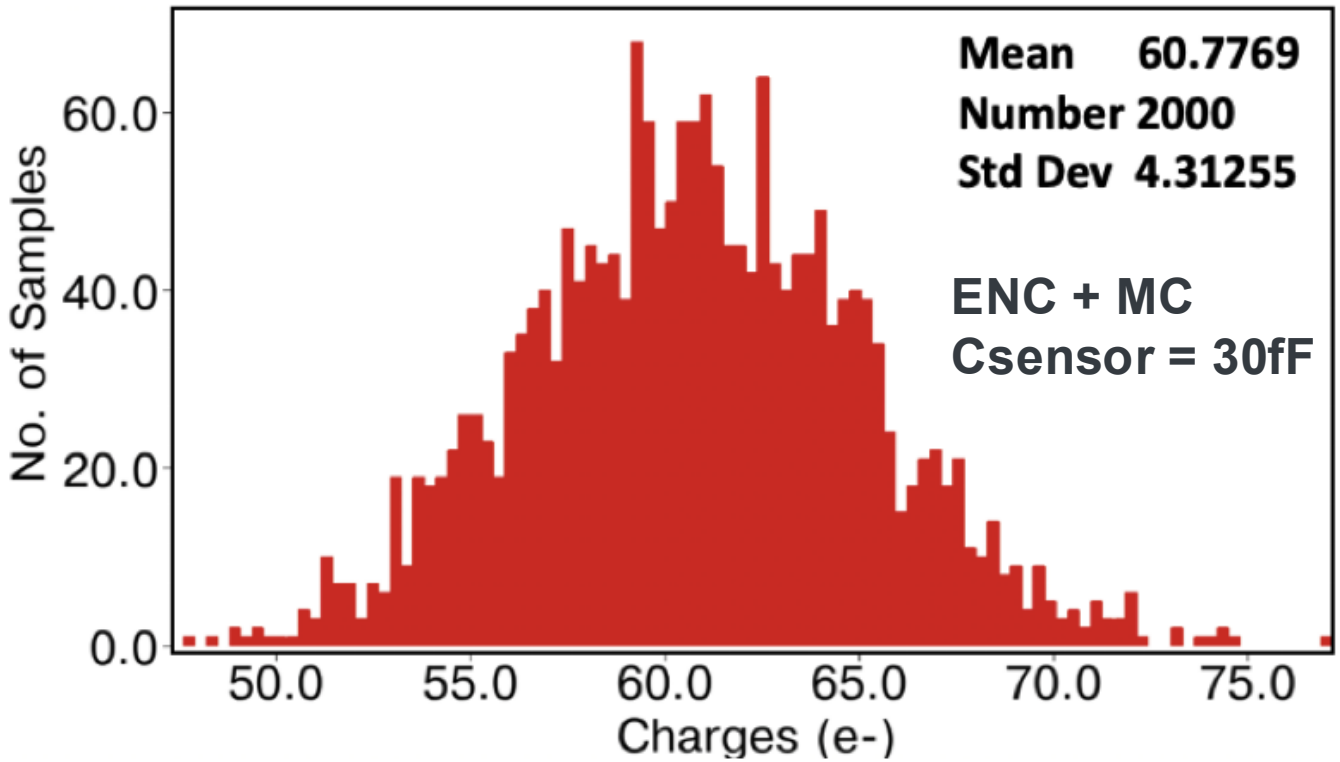
	Fraction of dataset	Rejection rate
Simulated tracks	40%	37.6 ±1.0%
Multi-pixel untracked	55%	63.2 ±1.1%
Single pixels	5%	100%

Table 2: Breakdown of the total dataset seen by the detector and the rejection rate achieved on each subsample.

https://arxiv.org/abs/2310.02474

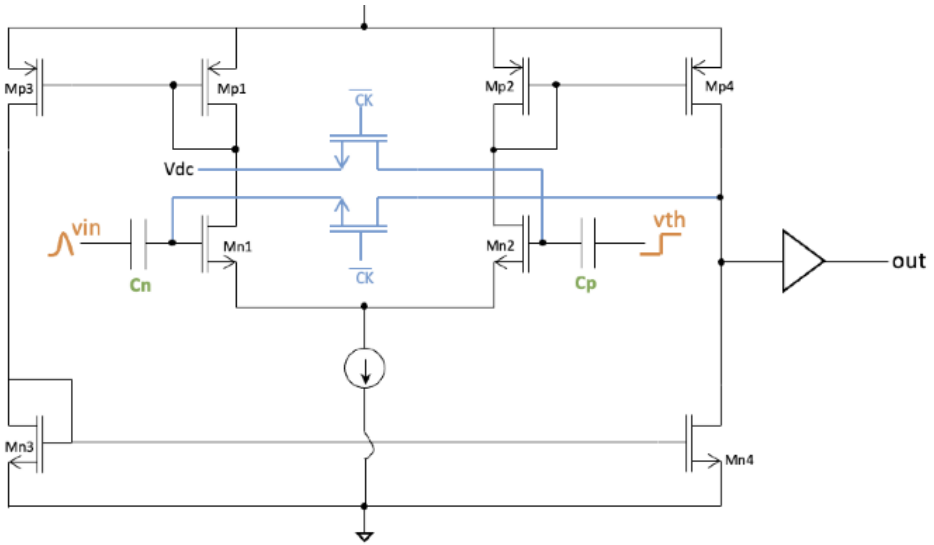
CSA schematic

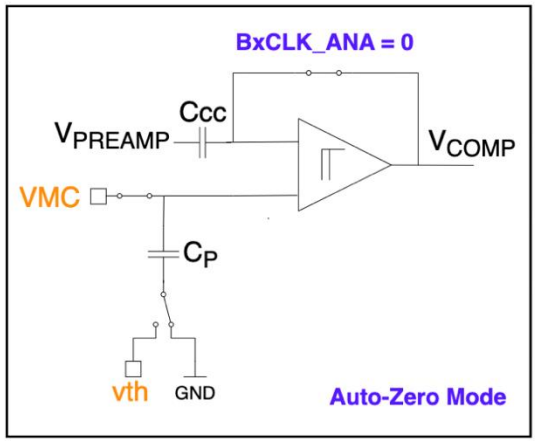


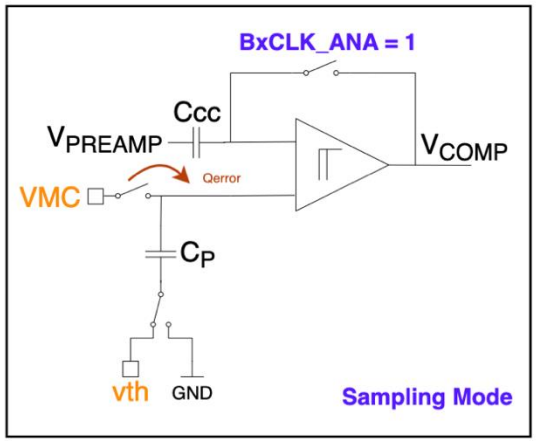


ENC simulation with 30fF sensor capacitor

Comparator Schematic and AZ Modes







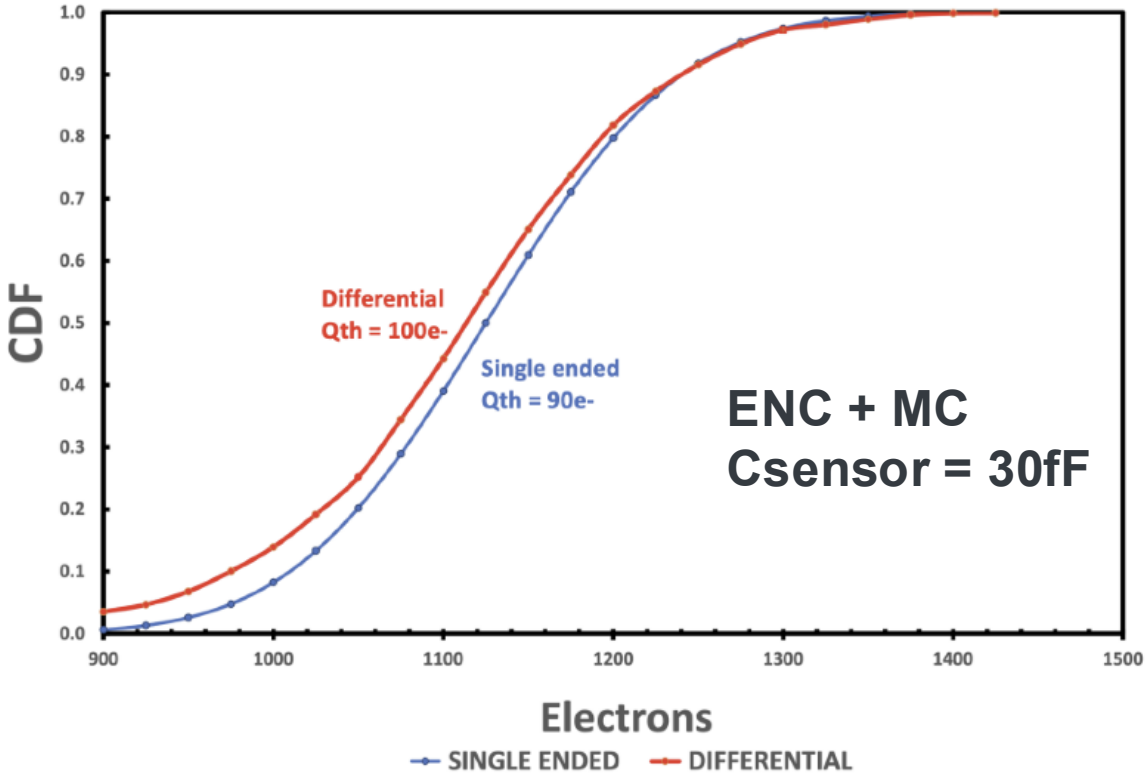
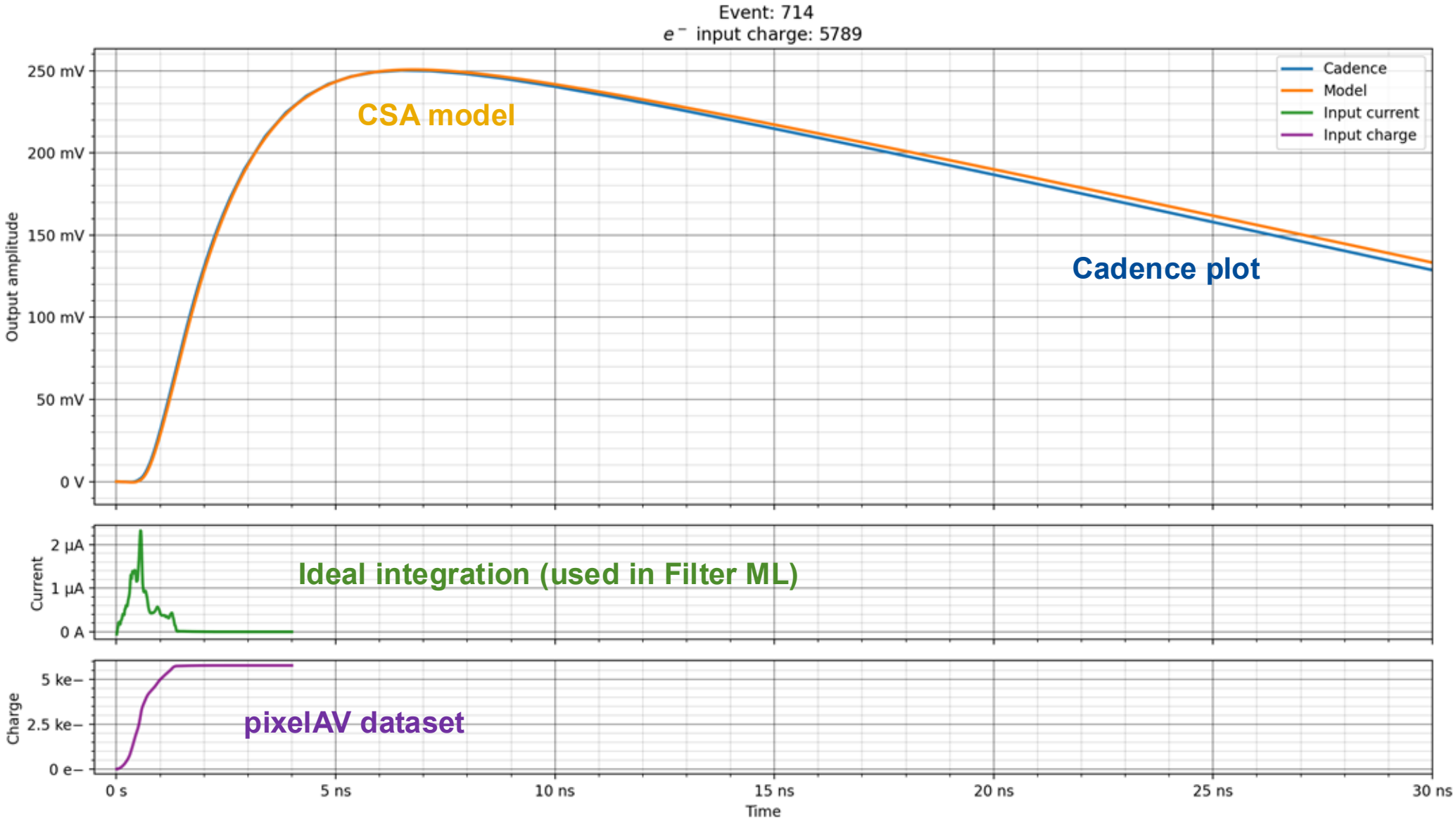


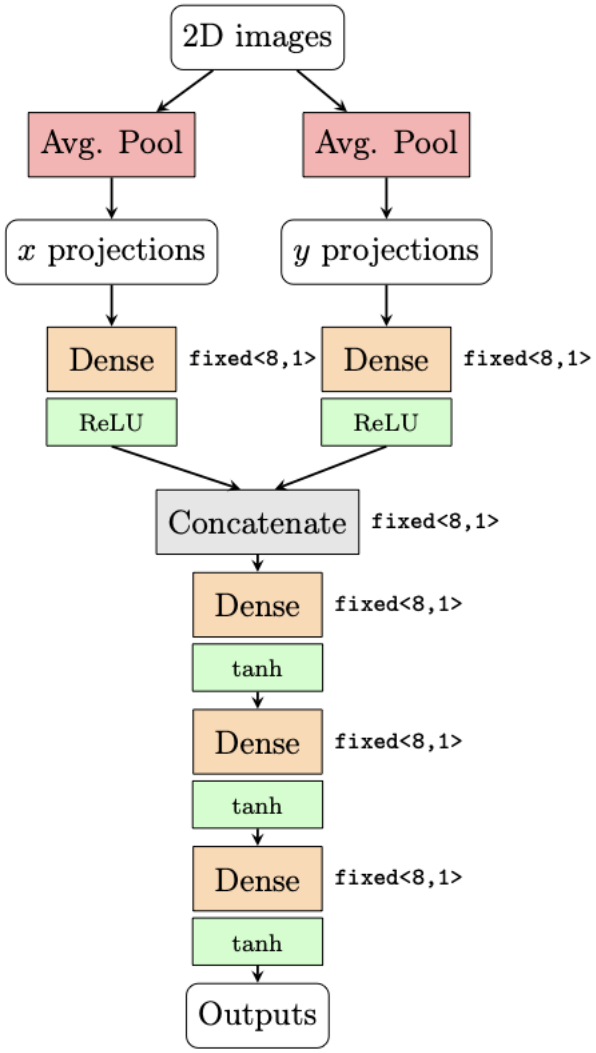
Figure 8: Cumulative Distribution Function (CDF) of both single ended and differential comparator structures after 300 Monte-Carlo runs

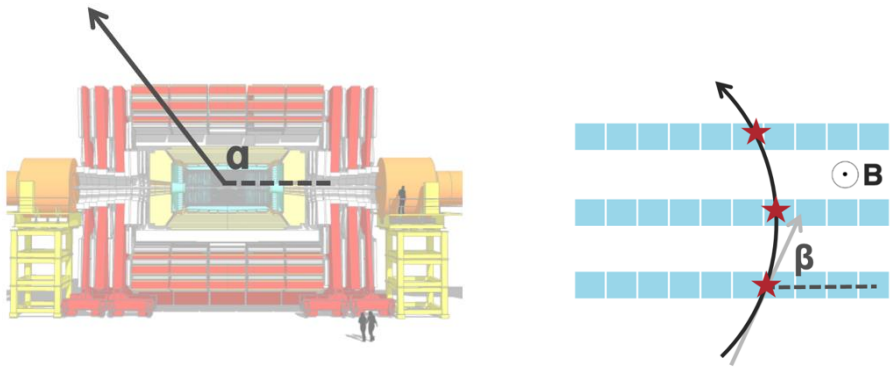
Future Dataset productions



CNN approach - Feature Regression Network

- Treat the charge deposited in the pixel array as 2D image
- **Convolutional layers** are powerful for image processing
  - Each 200 ps time slice is treated as a channel
- Predict and read out the traversing particle's **hit position** (x, y) and **incidence angle** (cot α)



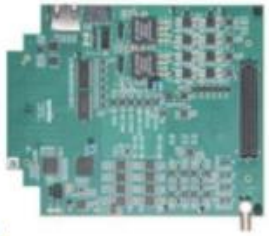


https://arxiv.org/abs/2602.15946

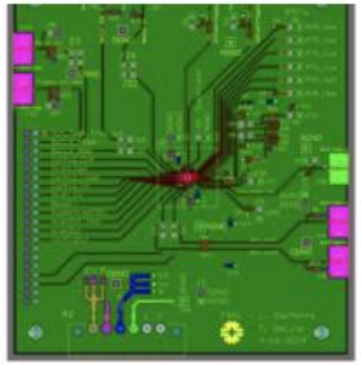
Test Setup Flow

Python Inteface

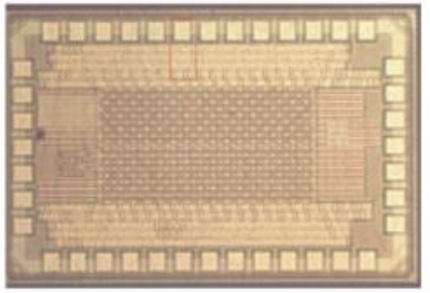
Zynq UltraScale+ MPSoC ZCU102 Eval Board



Caribou Board (Open-source DAQ System)



Custom ASIC board



Wire bonded ASIC