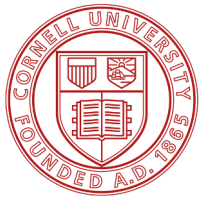
In-pixel machine learning in the presence of silicon detector noise

By: **Ben Weiss** on behalf of the Smartpixels group
Cornell University – bw539@cornell.edu

CPAD – 21 October 2026





UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN











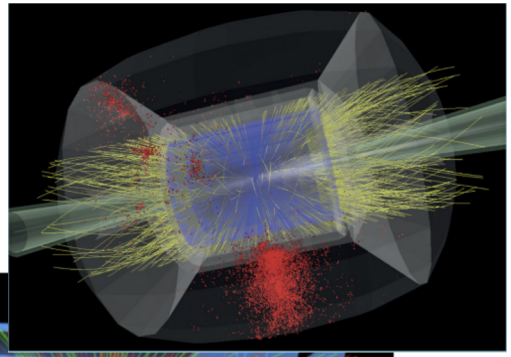




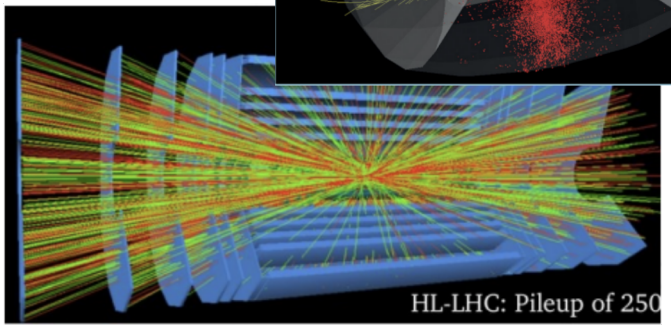
Pixel Granularity VS Data-rate

- Si-pixel tracking in high-luminosity colliders requires ultra-high resolution that strains bandwidth
- **Intelligent front-ends** relieve the burden on triggers downstream by beginning reconstruction at the source

Muon Collider:



FCC-hh:



FCC-hh: The Hadron Collider

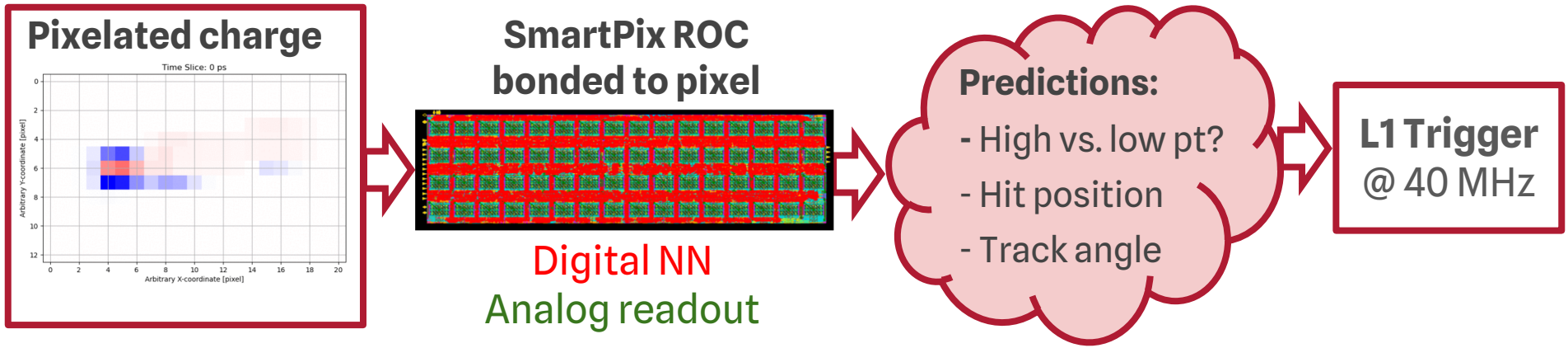
Future Circular Collider Conceptual Design Report Volume 3

Clearly, more sophisticated trigger algorithms, like isolation for muons or longitudinal segmentation for the calorimetry are needed to keep the trigger rates at acceptable levels. This essentially means that today's offline algorithms have to be migrated to the trigger.

Enter Smartpixels!

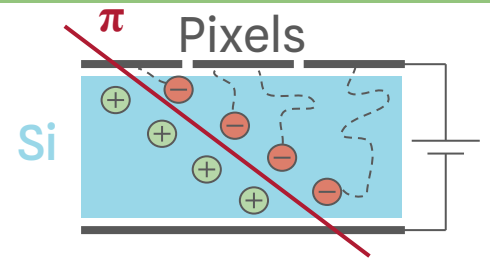
Smartpixels goal: embed small NNs on Si-pixel ASIC readout chips that perform pre-trigger data refinement at the collision rate.

- p_T filtering and regression of track geometry from a single Si-layer
  - **Advantage**: track parameter readout does not scale with granularity!
- Extreme-edge implementation targeting CMS Phase 3 conditions

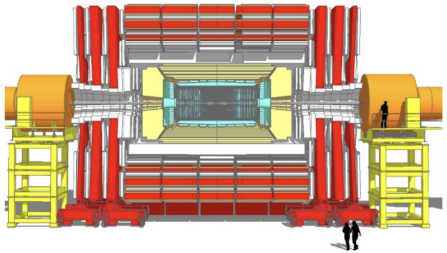


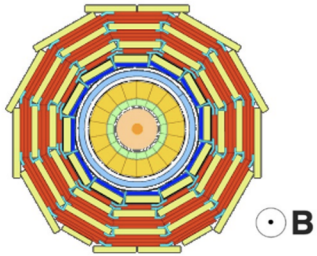
Smartpixels Datasets

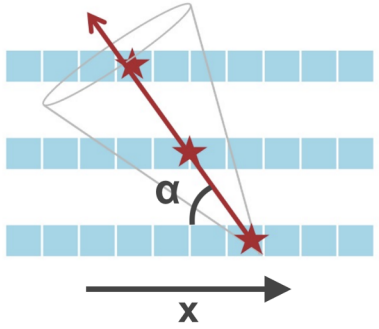
Incident pion kinematics taken from real CMS data.
Sensor charge collection simulated with **PixelAV**.

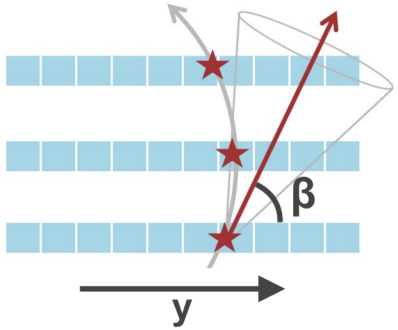


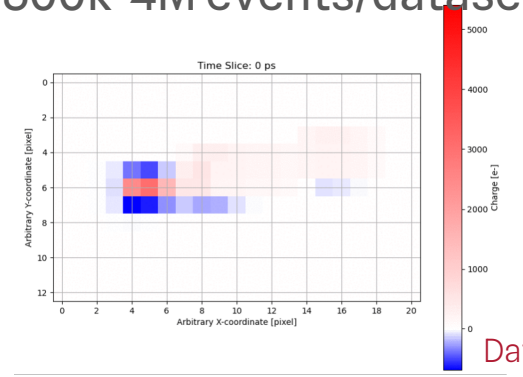
- 16x16 pixel array
- **50μm x 12.5μm** pixel pitch
- 20 timesteps (200ps appart)
- Subset available on **Zenodo**
- 500k-4M events/dataset











Dataset production work by Danush Shekar, Morris Swartz

Noise in a Si-pixel frontend

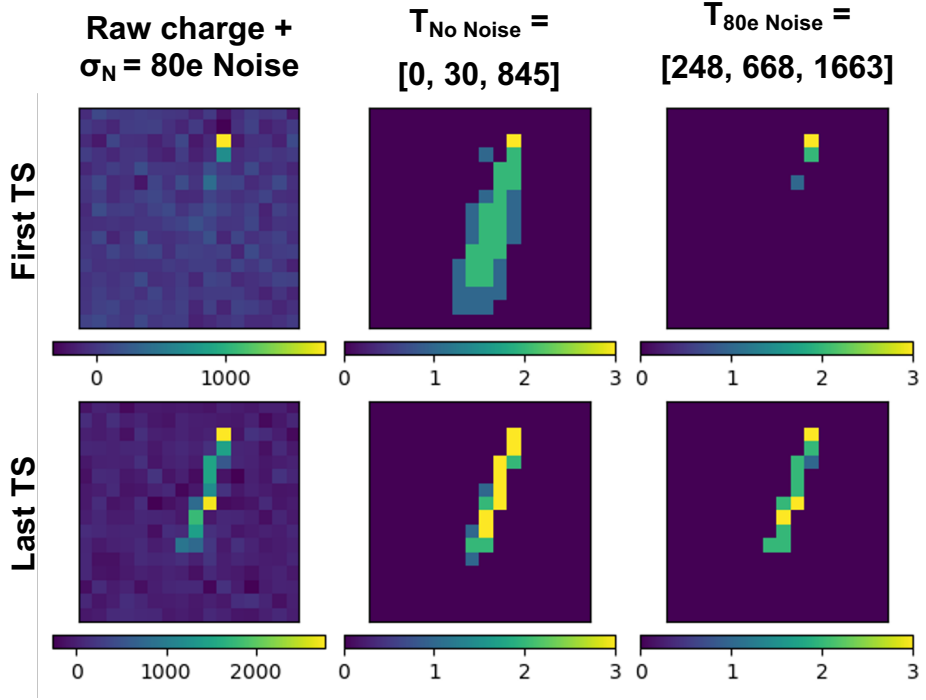
Si-pixel noise sources include thermal excitations, leakage current, irradiation...

- **This study assumes noise sampled from gaussian with μ=0, $\sigma=\sigma_N$**
- Detailed simulation of FE response presented in Harshul Gupta's talk
- Simulated irradiation studied in SmartPixels filter.
- Standard denoising approach: impose a 3-5σ_N threshold

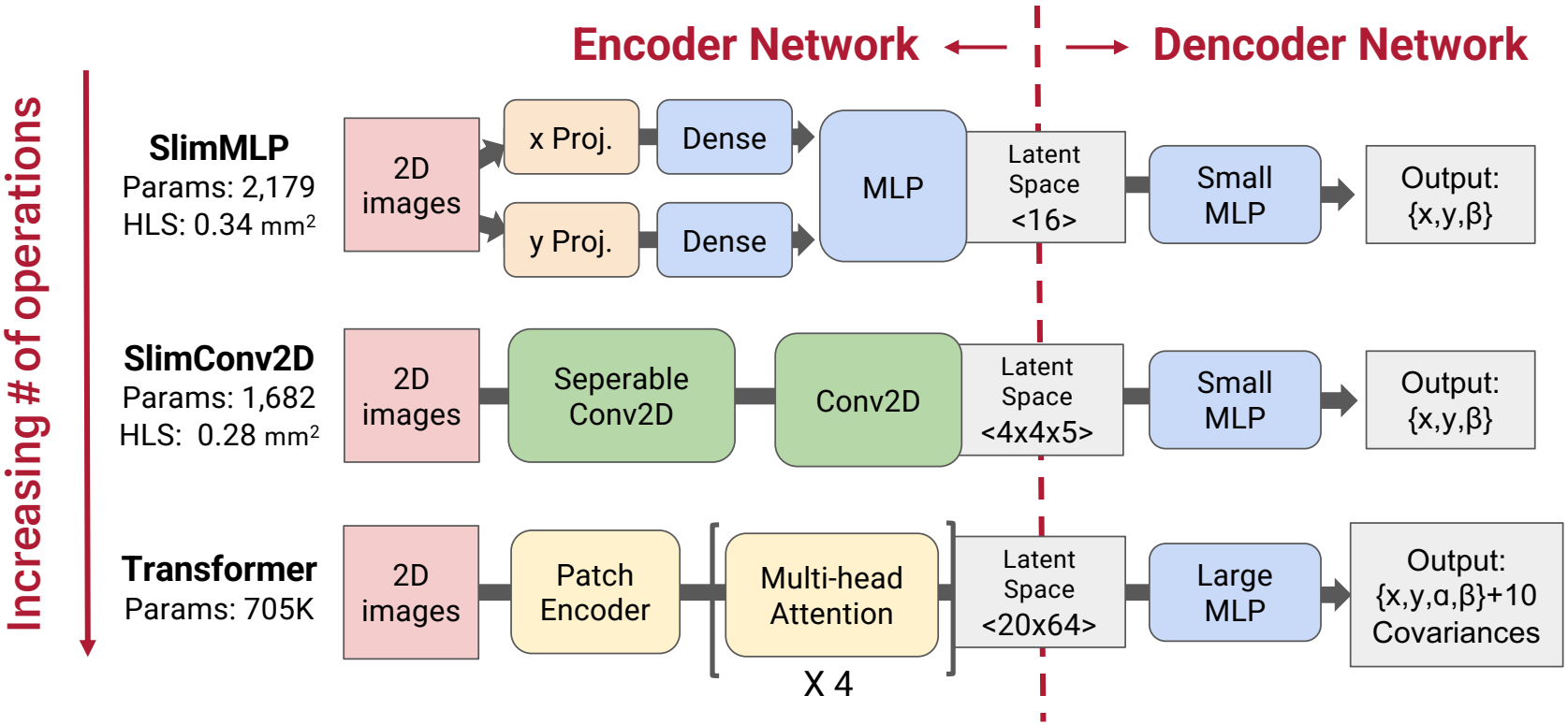
Noise estimated in https://arxiv.org/pdf/2510.07485

Estimated noise contributions

σ_{Th}	ENC	ENC w/ 0.5 GRad	Total σ_N
45e	40e	55e	~78e



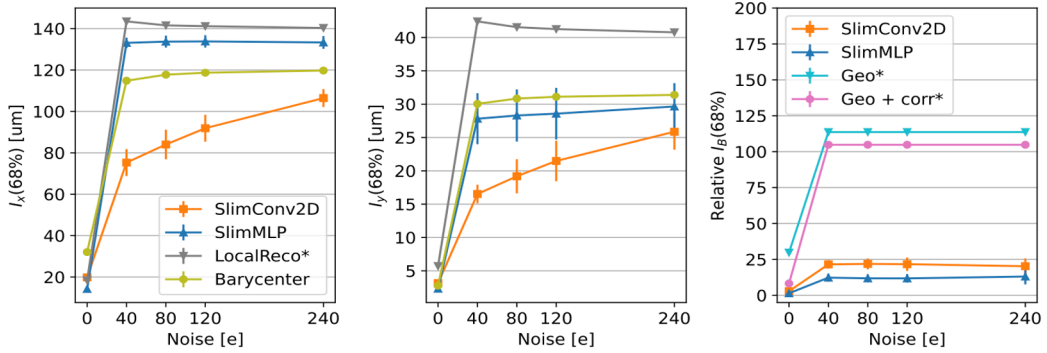
3 regression algorithms considered



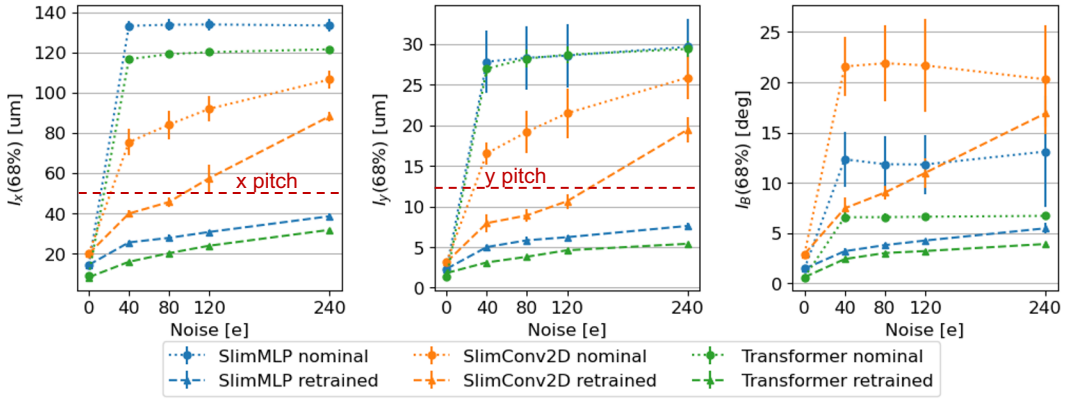
Regression in the presence of noise

- Base deployable network performance competes with non-ML techniques but both fail with a small amount of injected noise
- Retraining with noise recovers performance but barely reaching sub-pixel accuracy
- Networks here use ADC thresholds assuming no noise… Can we optimize these?

Base performance



Retraining with noise

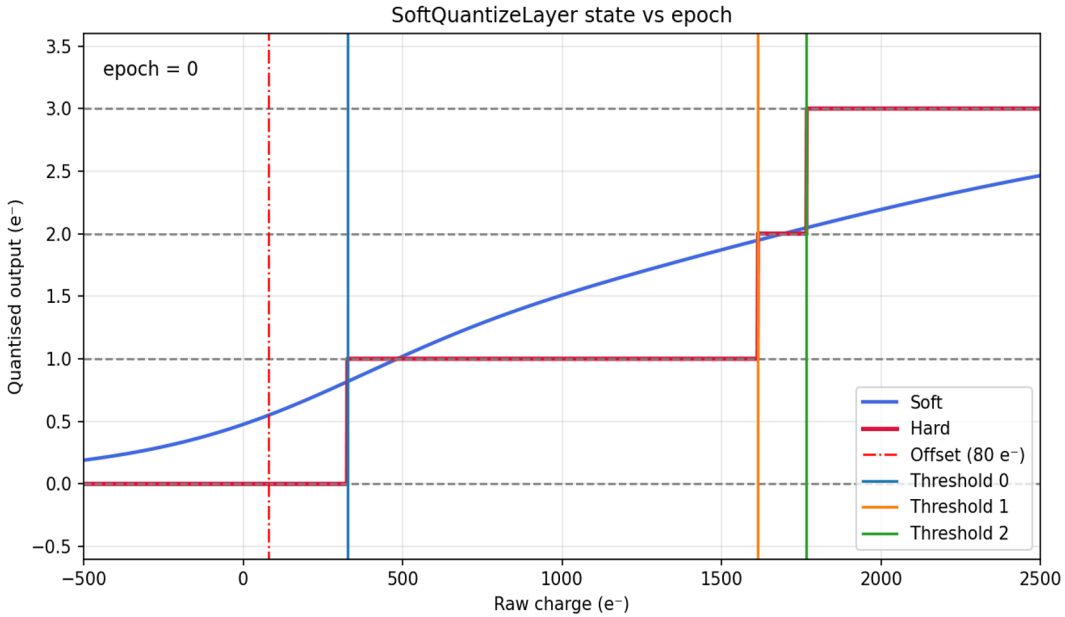


Soft Quantization: Optimize FE in training

Choice of 2-bit ADC thresholds matters! Treat these as **trainable parameters.**

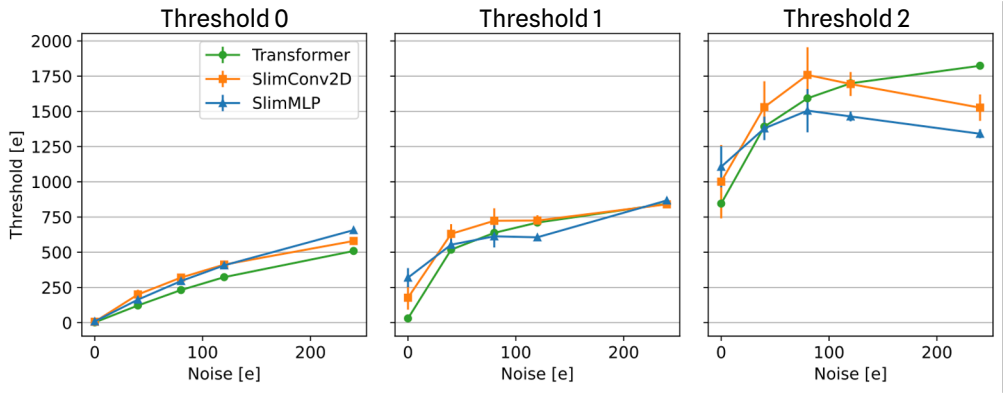
- First network layer: Smooth step function with trainable step location.
- Annealed slowly for optimization stability

$$D_{\text{soft}}(x) := \sum_{j=0}^{2^B-1} \sigma\left(\frac{k \cdot (T_j - x)}{\tau_j}\right)$$

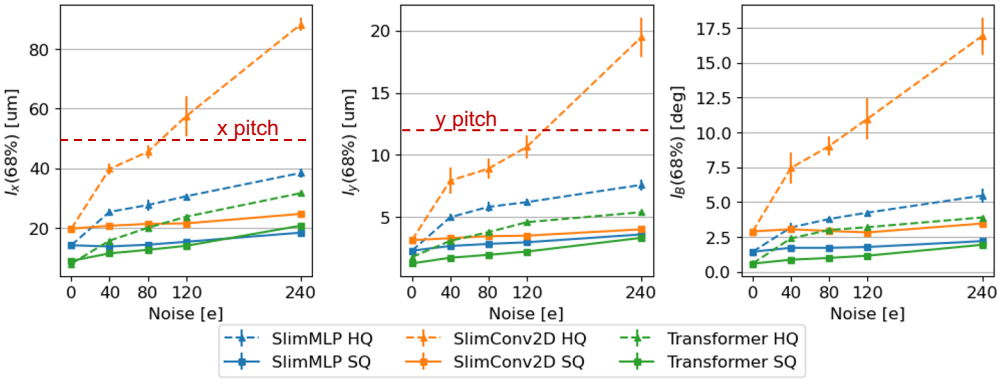


SQ Technique and animation by Arghya Das https://arxiv.org/abs/2602.15946v3

SQ thresholds and regression results



SQ improves regression



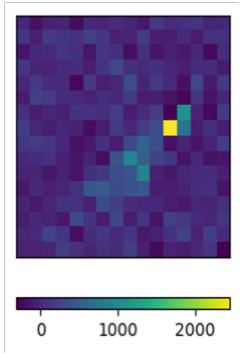
Soft Quantization naturally sets the lowest threshold to roughly 3xσ_N

- SQ drastically improves regression performance
- Can we go further?
  - Can charge information bellow σ_N be reintroduced by **denoising**?

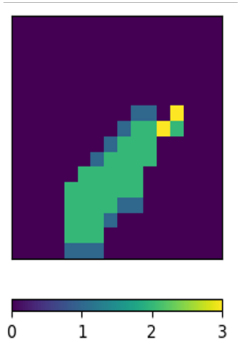
In pixel denoising

Denoising Goal: recover the cluster shape according to σ_N = 0 thresholds.

Raw charge collected +noise



Denoise & digitize

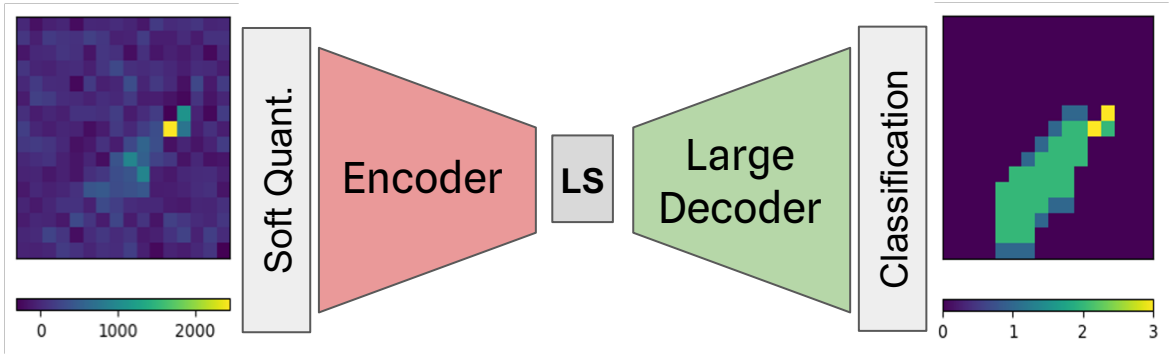


Denoised 2-bit image w/ Th(σ_N =0e): [0e, 30e, 845e]

In pixel denoising – SQAutoEncoder

1. Train each encoder to denoise using an **AutoEncoder** architecture

Raw charge collected + σ_N noise

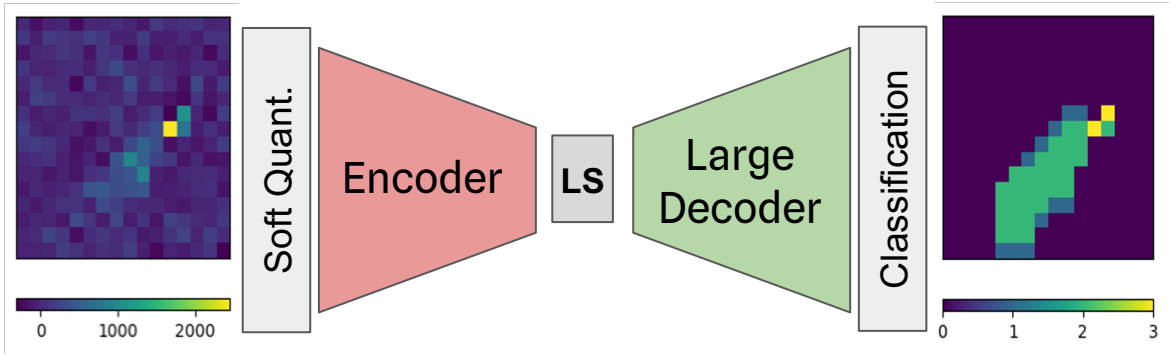


Denoised 2-bit image w/ Th(σ_N =0e): [0e, 30e, 845e]

In pixel denoising – SQAutoEncoder

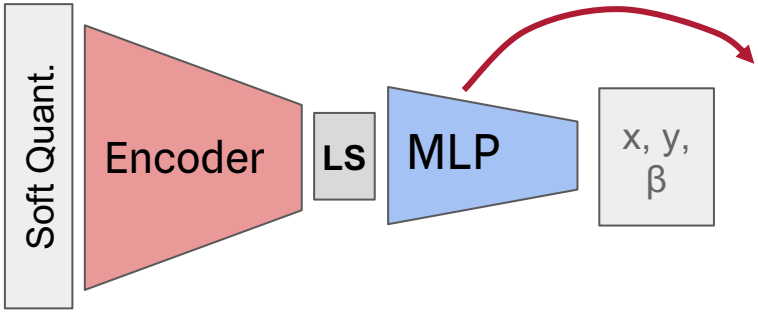
1. Train each encoder to denoise using an **AutoEncoder** architecture

Raw charge collected +noise



Denoised 2-bit image w/ Th(σ_N =0e): [0e, 30e, 845e]

2. Freeze the encoder and retrain the regression MLP

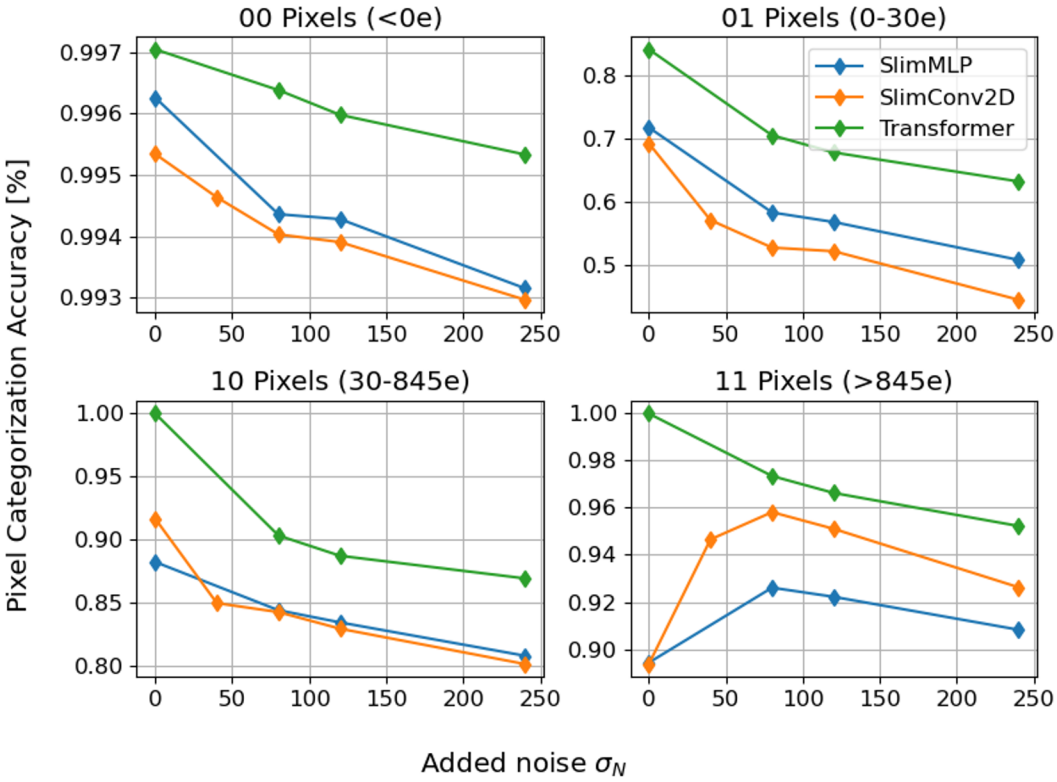


The MLP now operates on a LS rich with charge shape information below σ_N!

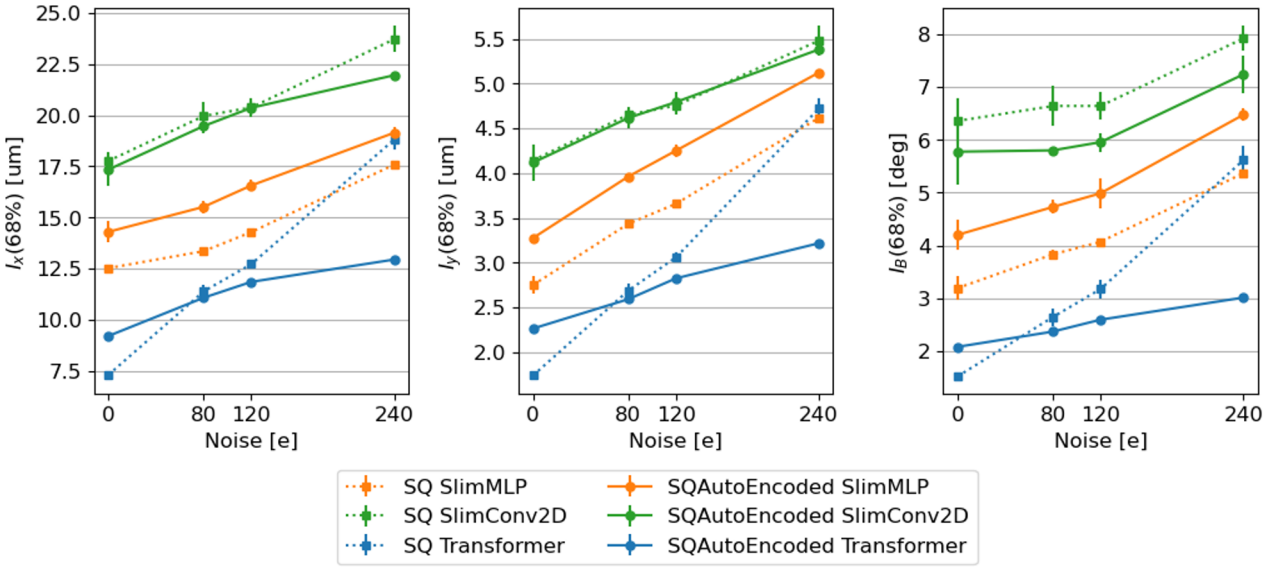
SQAutoEncoder Pixel Categorization accuracy

All SQ AutoEncoder architectures achieve **high denoised pixel categorization accuracy**, even well bellow σ_N

- Note: 90% of denoised pixels are 00
- Comparison to accuracy with standard digitization scheme in backup!



SQAutoEncoder: Denoising and regression



So, can we go further?

- A little bit, but only with a more complex on-chip algorithm

- SlimMLP performance degrades with SQAutoEncoder
  - Denoising may shift the learning objective too far from the regression task
  - Future work: multi-objective training

Conclusions

Si-Pixel detector noise degrades on-ASIC ML track parameter regression, but performance is recoverable with ML tricks.

- Intelligent front-ends offers a route towards tracking in HL environments.
- Even a simple noise model on a Si-AFE destroys track geometry predictions.
- Retraining the on-chip network improves performance. Optimizing ADC thresholds in the training loop helps even more.
- It is possible to further improve regression using a denoising autoencoder to enrich some network layers with information below the noise floor.
    - Currently this technique only works with a more sophisticated network on-chip.

The Smartpixels group!

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, Petra 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, Carissa Kumar, Emily Pan, Rachel Kovach-Fuentes, Aidan Nicholas, Eliza Howard, Eric You

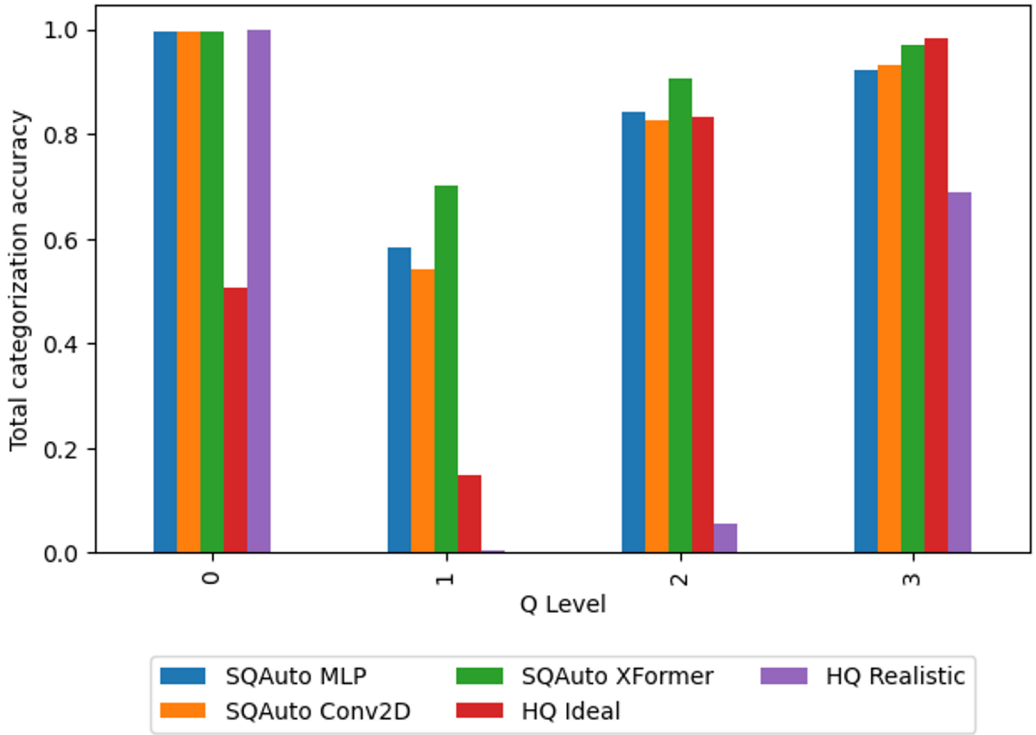
University of Colorado Boulder: Jannicke Pearkes, Ricardo Silvestre

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

SQAutoEncoders (MLP, Conv2D, XFormer)



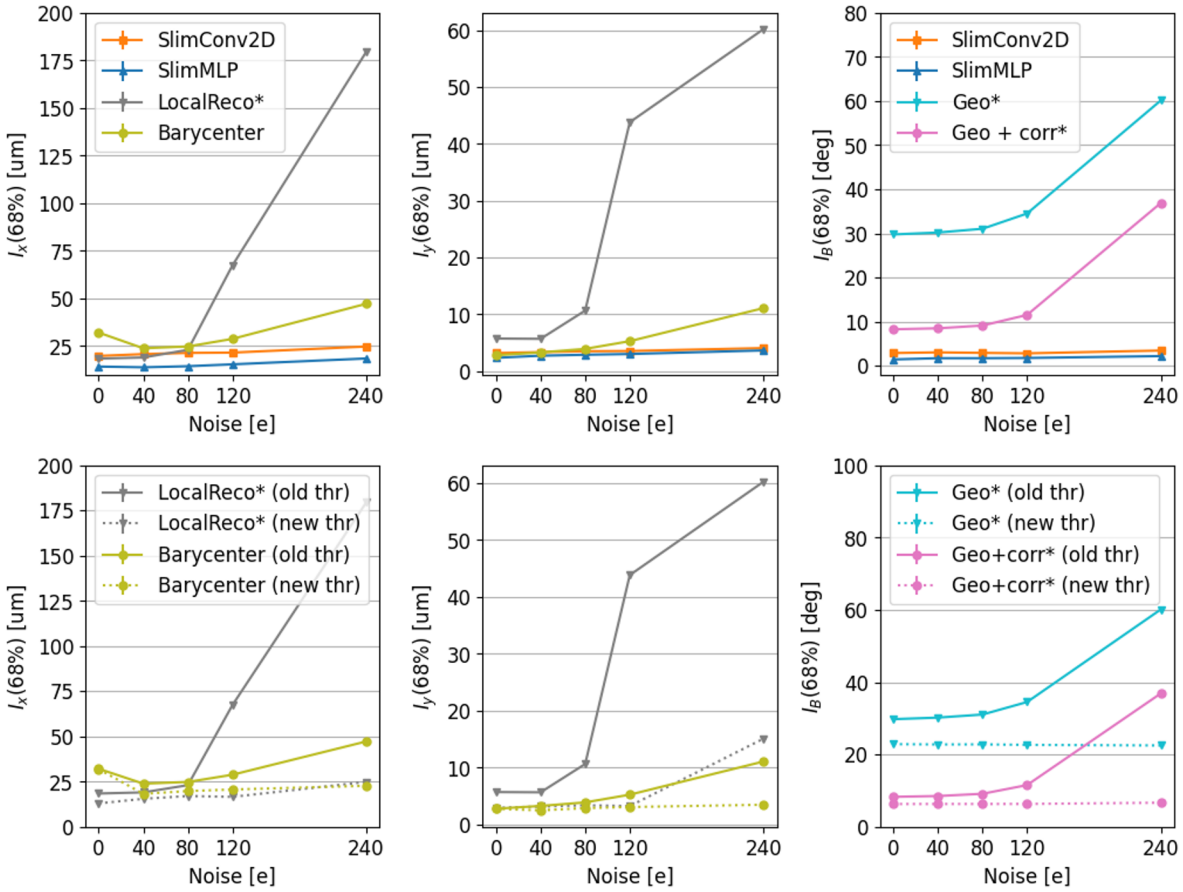
Given a cluster with 80e gauss. noise added, how well can each model predict denosied pixel values

- Average pixel categorization accuracy
- Denoised pixel values defined by Xformer optimized SQ thresholds with no added noise:
  - T_{Ideal} = [0e, 30e, 845e]
  - T_{Real} = [231e, 637e, 1592e]\

Non-ML performance

Using 80e thresholds -> (but SQ ML models)

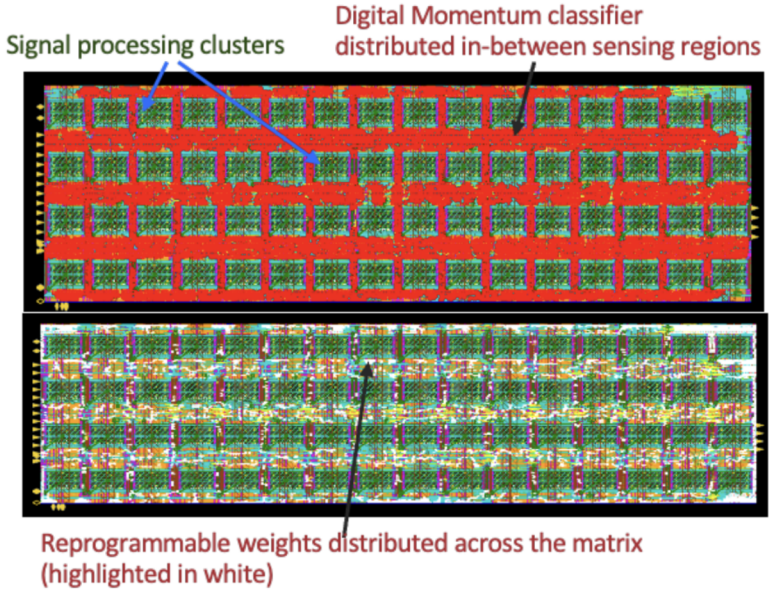
Optimized thresholds -> by noise

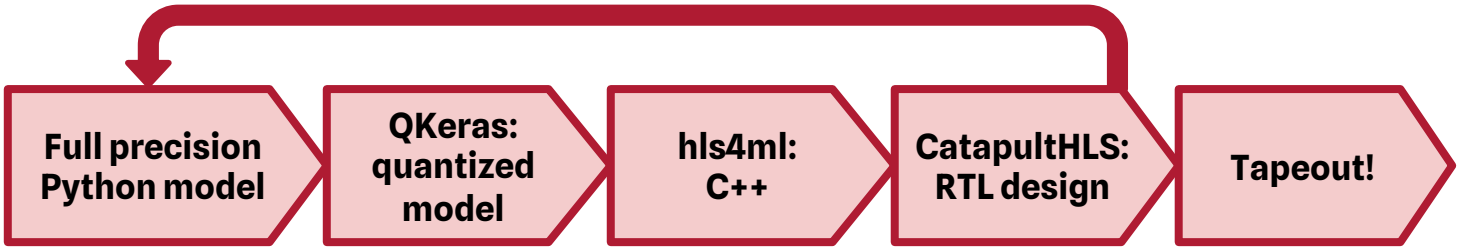


ASIC synthesis

Established pipeline to translate quantized python level model to RTL design

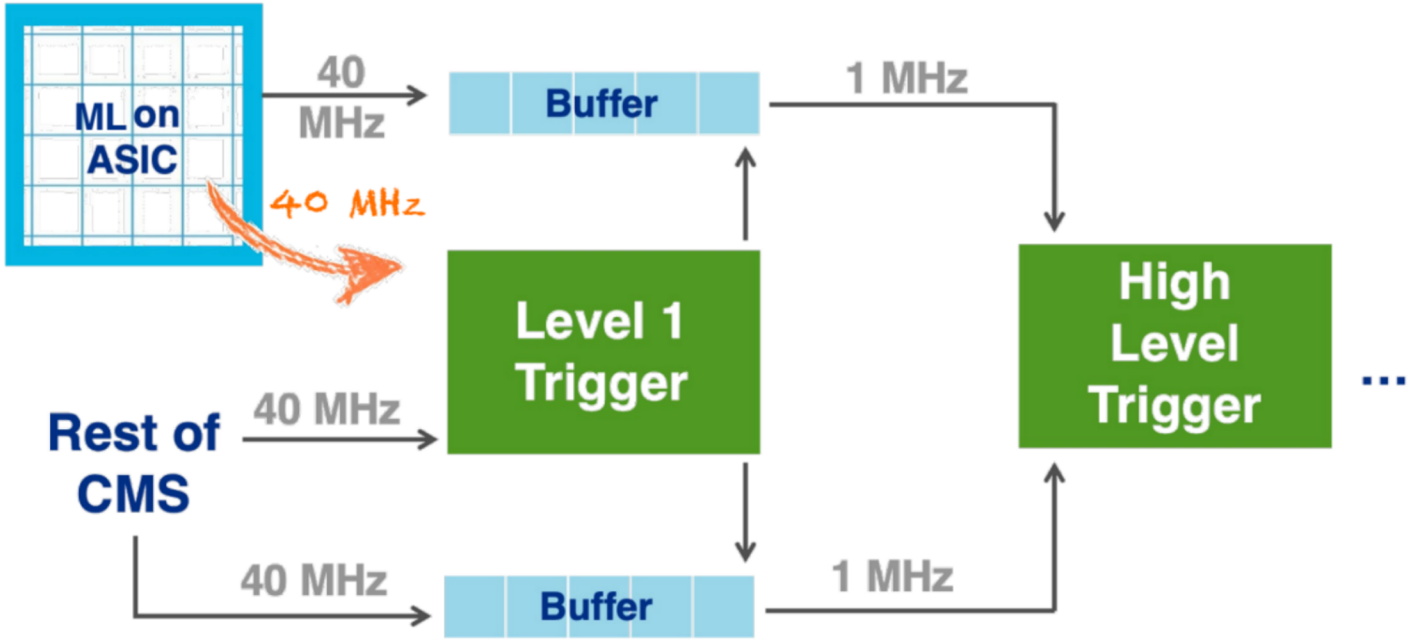
- hls4ml converts QKeras model to C++
- Catapult HLS converts C++ to RTL
- Estimates model footprint (mm^2), power consumption
- Feedback for python optmization





Credit: Giuseppe Di Guglielmo

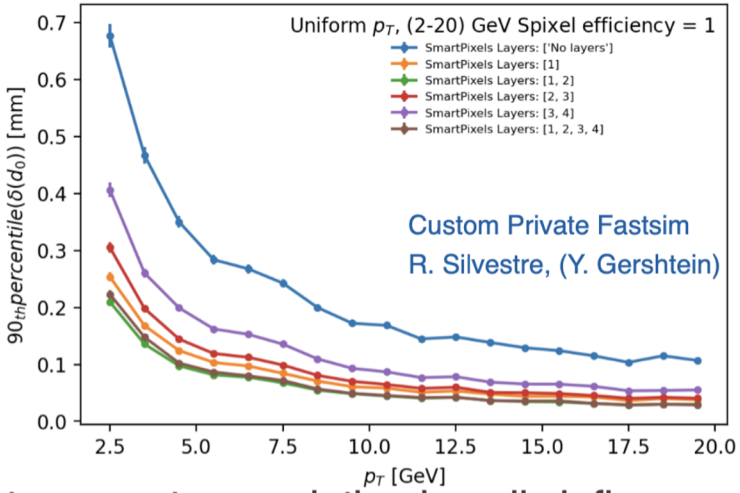
Readout chain: a futuristic smart pixels detector



Use ML to perform physics-motivated data reduction on-ASIC

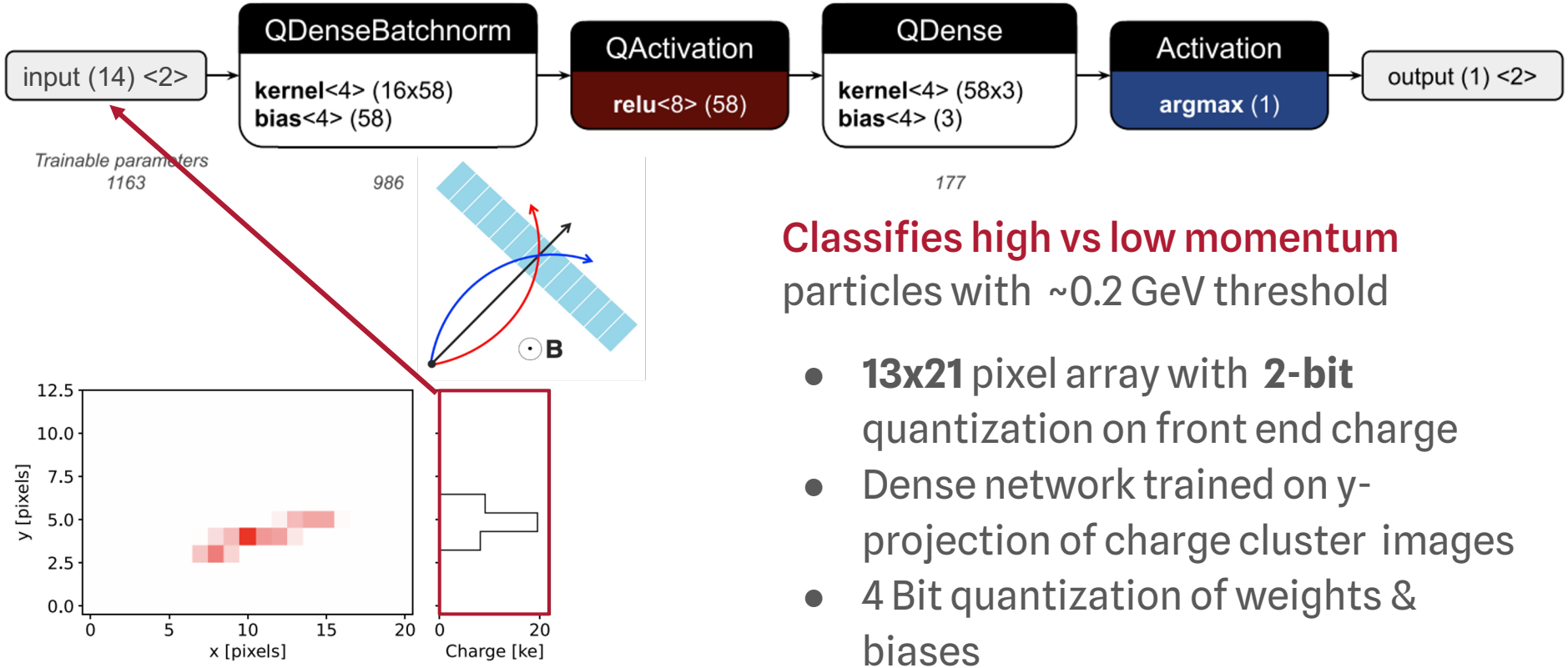
Fermilab

An Aside on Transverse Impact Parameters



- Transverse impact parameter resolution heavily influenced by pixel cluster precision and multiple scattering
  - LHC detector trackers are *heavy*
- Without innermost measurements:
  - extrapolated track fit has enormous errors at low pT
  - Physics improvements only at large lifetimes / high transverse boost
  - *Poor precision* for physically interesting phase space

Filtering algorithm



Classifies high vs low momentum particles with ~0.2 GeV threshold

- **13x21** pixel array with **2-bit** quantization on front end charge
- Dense network trained on y-projection of charge cluster images
- 4 Bit quantization of weights & biases

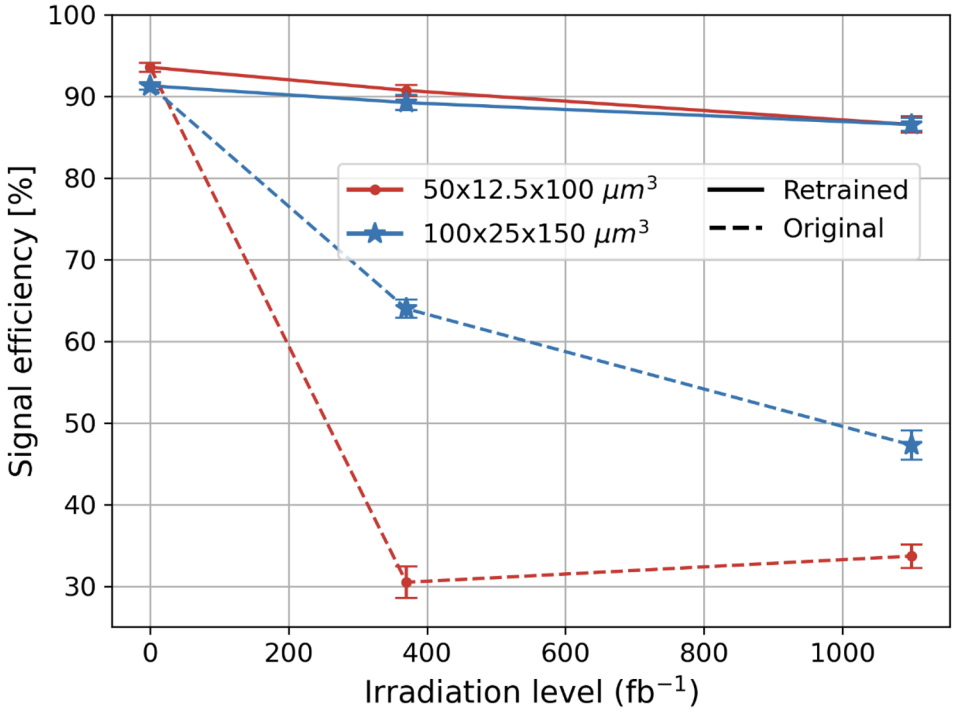
Credit: Jieun Yoo et al 2024. https://doi.org/10.1088/2632-2153/ad6a00

Filtering algorithm performance

Multiple geometries (0.2 GeV thresh.):

Sensor geometry [um^3]	Bias voltage [V]	Signal efficiency %	Data reduction %
50 X 10 X 100	100	95.4 ± 0.5	33.1 ± 1.0
50 X 12.5 X 100	100	93.9 ± 0.5	33.1 ± 0.9
50 X 15 X 100	100	93.3 ± 0.5	30.7 ± 0.9
50 X 20 X 100	100	91.2 ± 0.9	28.4 ± 0.9
50 X 25 X 100	100	88.3 ± 0.7	27.3 ± 0.8
100 X 25 X 100	100	88.6 ± 0.9	26.9 ± 1.0
100 X 25 X 150	175	91.9 ± 0.7	29.7 ± 1.0

Retrainable to endure irradiation:



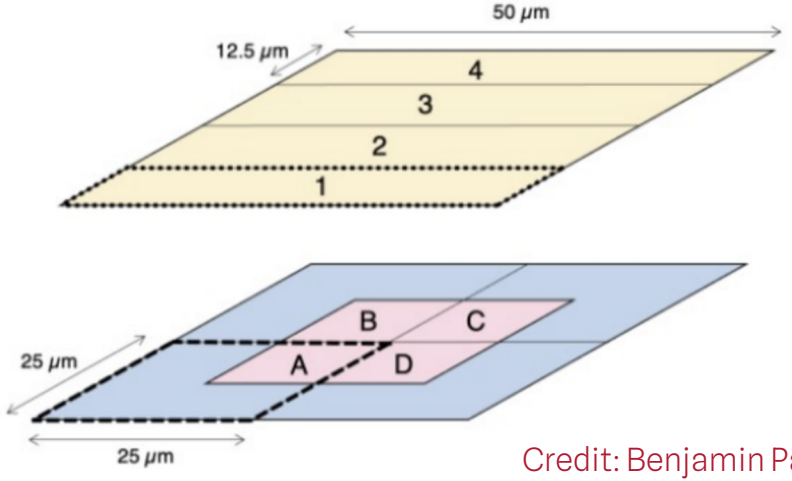
Credit: Danush Shekar, Jennet Dickinson

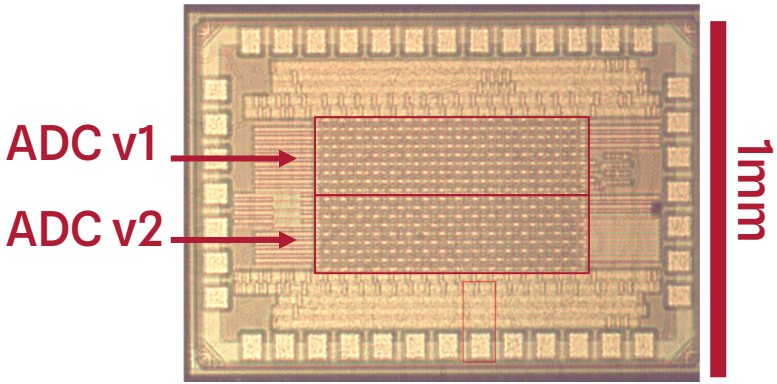
New Prototype SmartPixels ROIC

Newest Smartpixels **TSMC 28 nm** chip hosts per-pixel charge injection and programmable network for **classifier**.

- 2 sets of 8x32 pixel arrays (2 ADC vers.)
- ADC+DNN consumes ~6 μW/pixel @ 40 MS/s

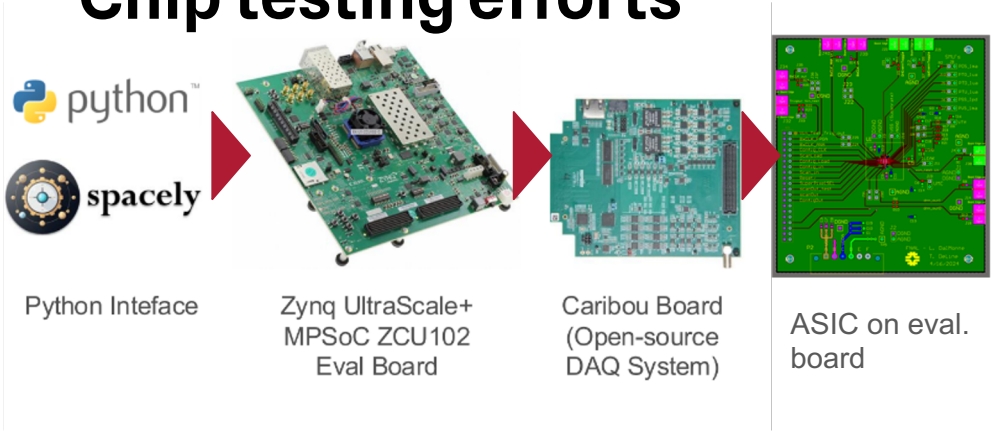
Spec.	Value	Units
Parameters	4652	w+b
DNN+ADC area	~0.2	mm^2
Total area	~1.6	mm^2
Clock frequency	40	MHz
Latency	50	ns





Credit: Benjamin Parpillon, Anthony Badea

Chip testing efforts

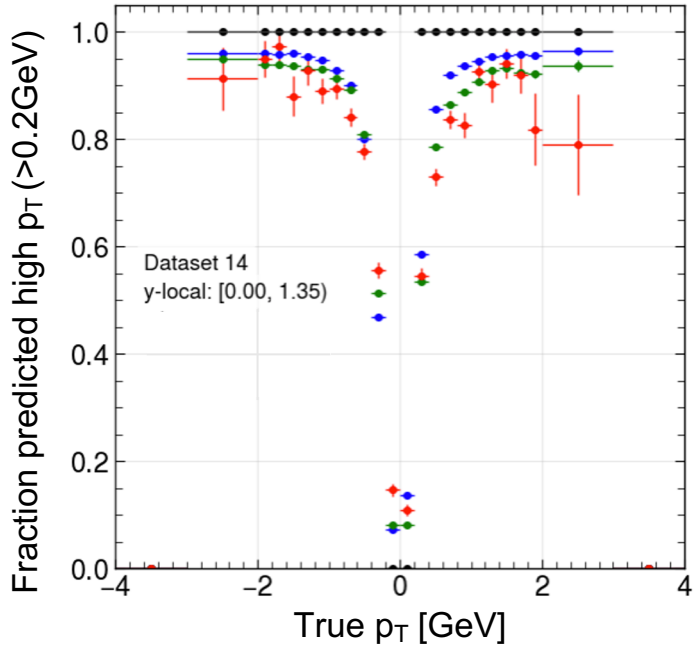


FPGA based setup to communicate with the smartpixels ROIC

- Spacely+Caribou+Peary

Test routines developed to measure:

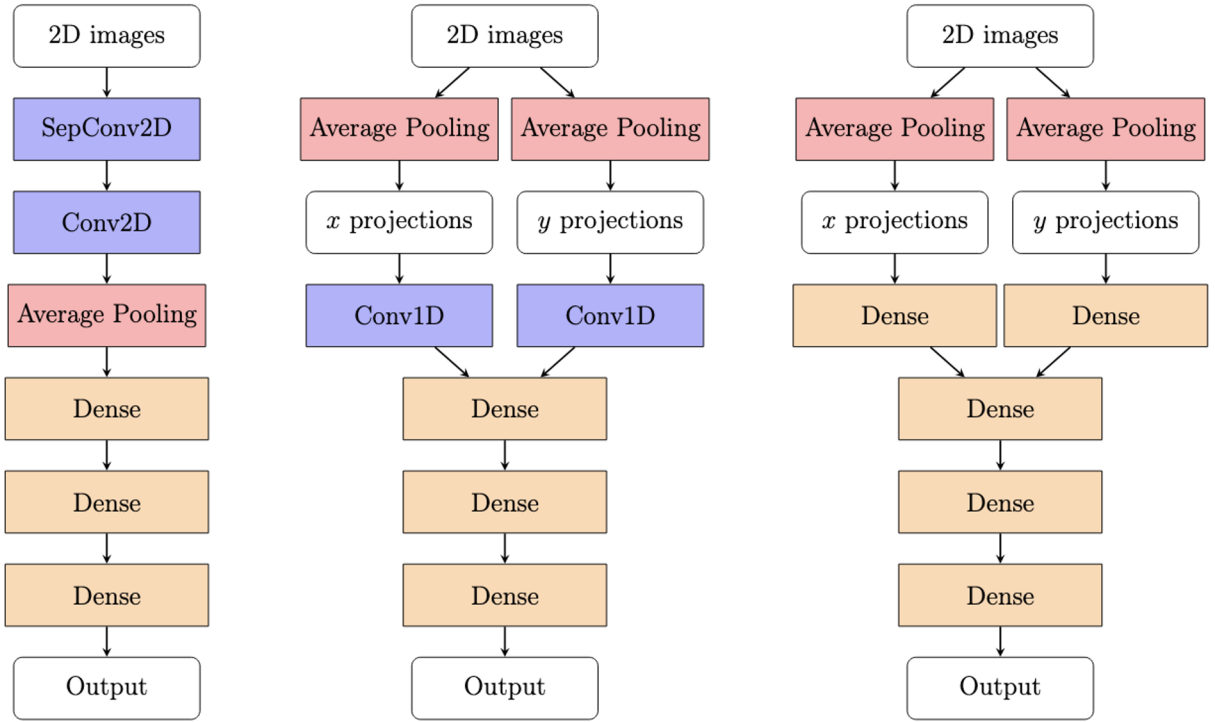
- Power consumption
- Pixel S-curves
- DNN programing/ predictions on chip



Credit: Benjamin Parpillon, Anthony Badea

Regression Alg. design space exploration

Reducing # of operations

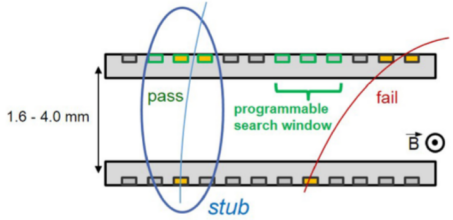


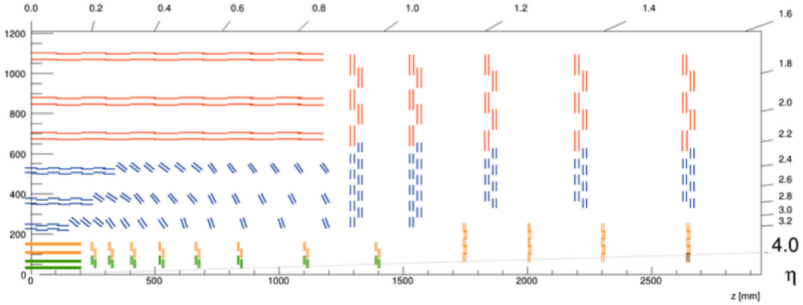
Current development to reduce model size for hardware implementation

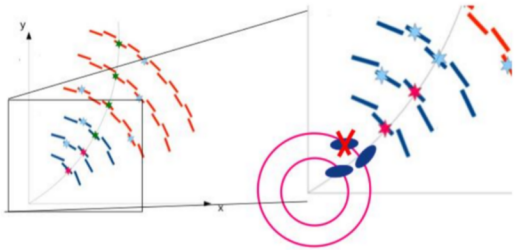
- On ASIC, generally: Conv2D>Conv1D>Dense

ASIC Synthesis (45nm)	
Clock Period	5 ns
Area Estimate	1.4 mm^2
Buffer Area	0.0017 mm^2
Inverter Area	0.055 mm^2
Logic Area	0.80 mm^2
Sequential Area	0.52 mm^2
Latency	27 μs

Using smartpixels data in a detector system







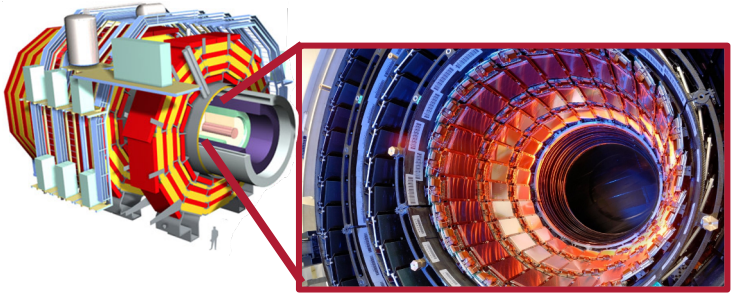
- We get roughly 2-3 degree resolution in the azimuthal direction
  - "Inside-out" tracking will not work very well, combinatorics too high
  - However we can use the outer tracker tracks to identify regions of interest and then the angular reconstruction from smartpixels to match the extrapolated track
- With more processing on detector we could try to find "pixel seeds" instead
  - More clean and pure, but this would require on detector track finding to deal with combinatorics

Regression ROIC s

- **Algo #1: “conv2d model”**
  - I/O
    - 13x21x2 inputs, bit-width 4 (fixed<4,1>)
    - 14 outputs fixed<8,2>
  - HLS area **0.81 mm²** (Syn 1.23 mm², PnR 1.46 mm²)
- **Algo #2: “conv1d model”**
  - I/O
    - 13x21x2 inputs fixed<4,1>, bit-width 4 (fixed<4,1>)
    - 14 outputs fixed<8,2>
  - HLS area **0.39 mm²** (Syn 0.6 mm², PnR 0.72 mm²)
- **Algo #3: “MLP model”**
  - I/O
    - 16x16x2 inputs, fixed<16,6>
    - 3 outputs fixed<16,6> (x and y + angle)
  - HLS area **0.52 mm²** (Syn 0.78 mm², PnR 0.94 mm²)

Data rates at the LHC

*Very rough estimate, YouTube streams ~60B hrs/year



≈

YouTube

Compact Muon Solenoid (CMS) Pixel detector

~Raw O(TB/s)

YouTube's total average streaming data rate*

~O(TB/s)

Triggers



