Front-end neural network filtering implemented in a silicon pixel detector

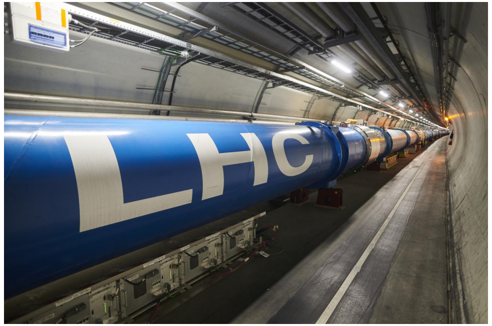
CPAD Workshop
November 8, 2023

Alice Bean, Douglas Berry, Manuel Blanco Valentin, Jennet Dickinson, Giuseppe Di Guglielmo, Karri DiPetrillo, Farah Fahim, Lindsey Gray, James Hirschauer, Shruti R. Kulkarni, Ron Lipton, Petar Maksimovic, Corrinne Mills, Mark S. Neubauer, Benjamin Parpillon, Gauri Pradhan, Morris Swartz, Chinar Syal, Nhan Tran, Dahai Wen, **Jieun Yoo**, Aaron Young

Fermi National Accelerator Laboratory, Johns Hopkins University, Northwestern University, Oak Ridge National Laboratory, The University of Chicago, University of Illinois Chicago, University of Illinois Urbana-Champaign, University of Kansas

Motivation

- LHC upgrade requires technologies to deal with an increase in luminosity, pileup, & data, in a high radiation-environment
- LHC pp collisions occur at 40MHz, are selected by a trigger to read out events ~ 1MHz
- Currently, events with new physics only in the pixel data are not selected at all
- AI *embedded* on a chip can be used to filter data at the source, enabling data reduction AND taking advantage of pixel information to enable new physics measurements and searches

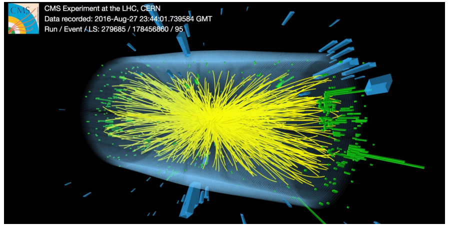


LHC Luminosity

- LHC design 10^{34} cm^{-2} s^{-1}
- LHC Runs 2/3: 2 x LHC
- HL-LHC: 5 to 7 x LHC

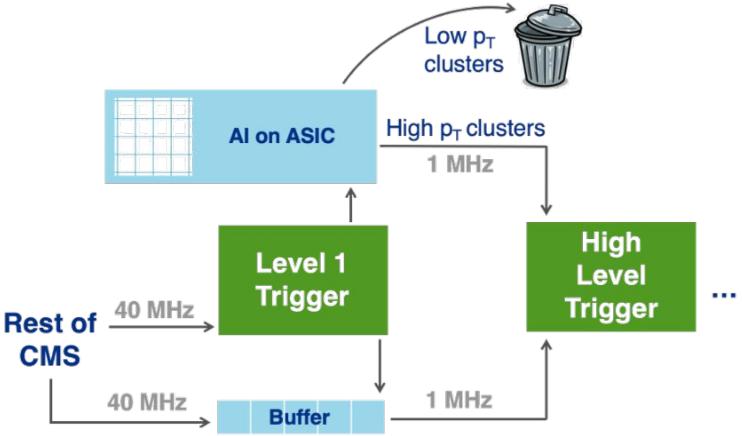
To Learn More:

- CPAD 2022: J. Dickinson, Smart pixels with data reduction at source
- CPAD 2023: B. Parpillon, Readout IC for future Phase III high luminosity upgrade of the large Hadron collider
- ICAD 2023: G. Di Guglielmo, Smart pixel sensors: towards on-sensor filtering of pixel clusters with deep learning



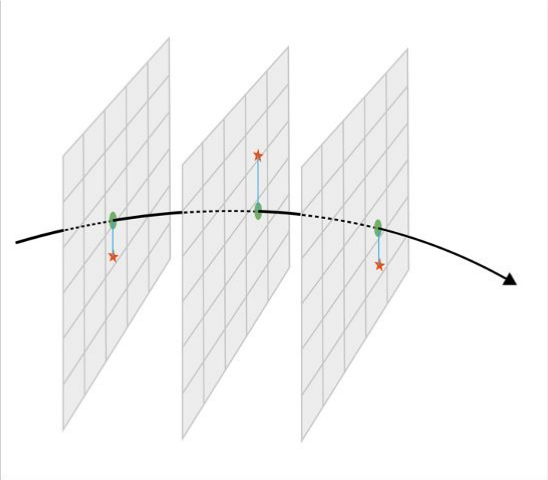
Data reduction

- Data reduction through
  - **Filtering** through removing low p_T clusters
  - **Featurization** through converting raw data to physics information
- Combination of approaches can reduce data rate enough to use pixel information at Level 1

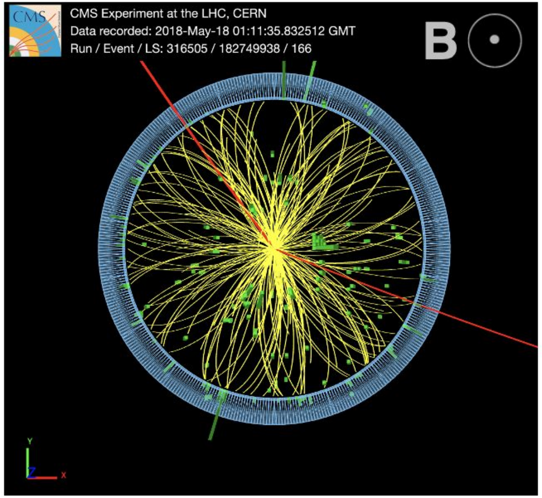


Particle tracks

- Reconstructing vertices is critical
- Connecting the dots between charge collected in different pixel layers creates a particle track
- Solenoid magnet immerses the pixel detector in a B-field, causing tracks to curve

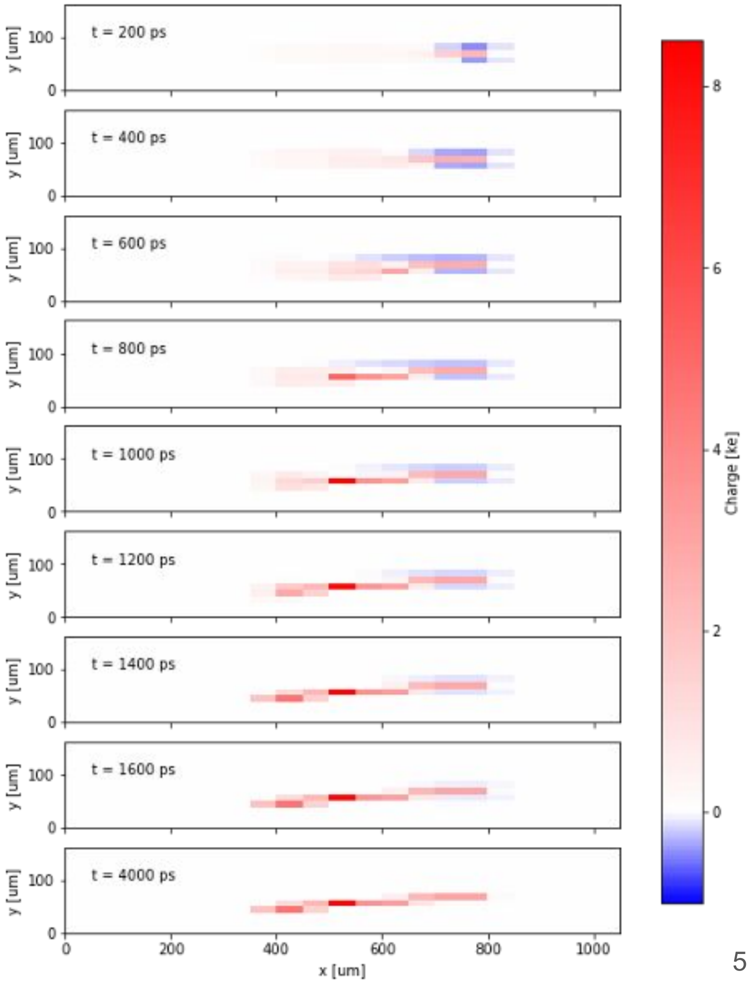


Very curved → low momentum
Almost straight → high momentum



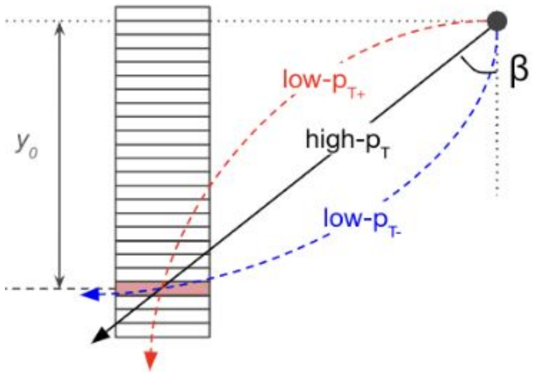
Simulated dataset (link)

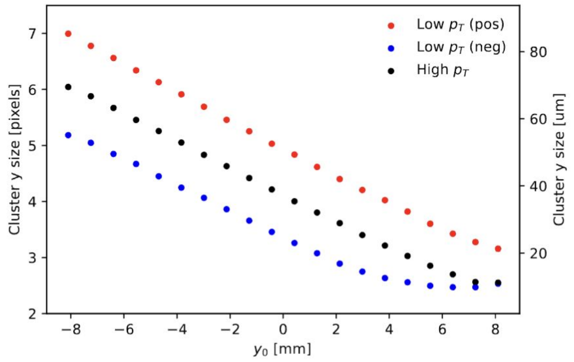
- Simulated charge deposition from pions
    - Initial conditions = fitted tracks from CMS
    - For a range of hit positions, incident angles
- Assume a futuristic pixel detector
    - 21x13 array of pixels
    - 50x12.5 µm pitch, 100 µm thickness
    - Located at radius of 30 mm
    - 3.8 T magnetic field
    - Time steps of 200 picoseconds



ML Inputs: y-position

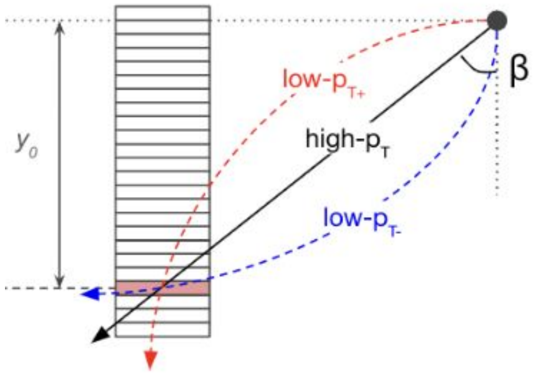
- The shape of the cluster is strongly correlated with its y-position (its azimuthal position with respect to the center of the sensor)
- Cluster y-size vs. y-position shows clear correlation between size & position
    - Decrease in cluster size from left to right is due to Lorentz drift
    - The final model chosen uses y-profile (not y-size) due to the former's better performance



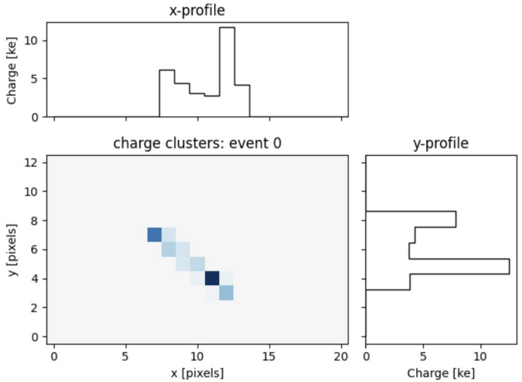


ML Inputs: y-profile

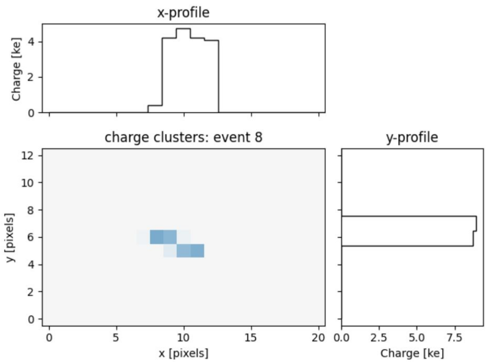
- We use ML due to complicated pulse shapes, and drift & induced currents
- y-profile (sum over pixel rows) projects the cluster shape on the y-axis and is sensitive to the incident angle β and thus the particle's p_T
- x-profile (sum over pixel columns) is parallel to B, and uncorrelated with p_T



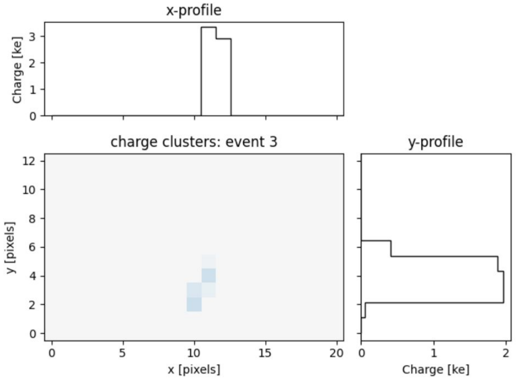
High p_T cluster



Low p_T positively charged cluster



Low p_T negatively charged cluster

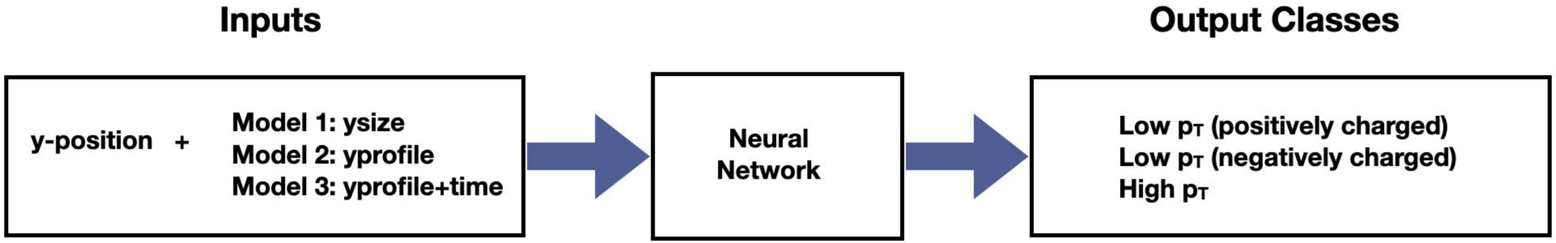


Classification Goals

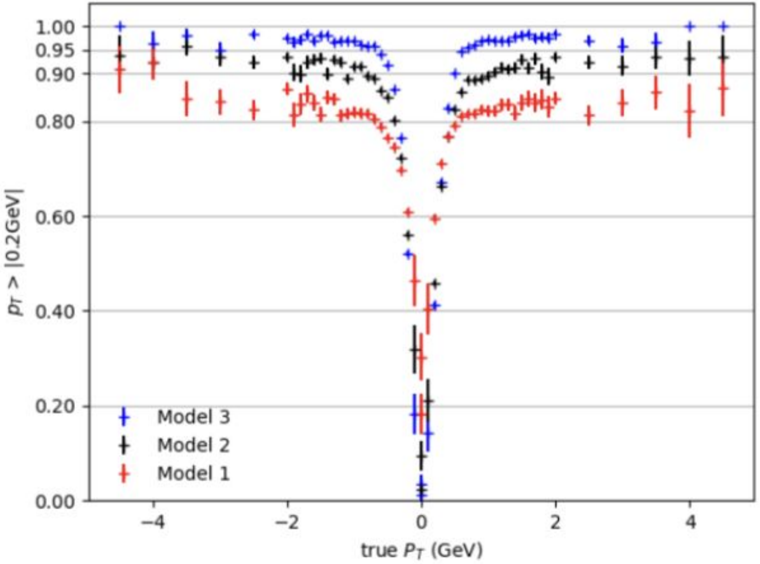
- Keep as many high p_T clusters as possible for physics
- Decrease data bandwidth

Baseline full precision model

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 flatten (Flatten)           (None, 14)                0

 dense (Dense)               (None, 128)               1920

 dense_1 (Dense)             (None, 3)                 387

=================================================================
Total params: 2,307
Trainable params: 2,307
Non-trainable params: 0
```



Metrics



$$Signal\ Eff. = \frac{\#\ clusters\ classified\ as\ high\ p_T}{\#\ clusters > 2\ GeV}$$

$$Bkg.\ Rej. = \frac{\#\ clusters\ classified\ as\ low\ p_T}{\#\ clusters < 2\ GeV}$$

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 was chosen for implementation

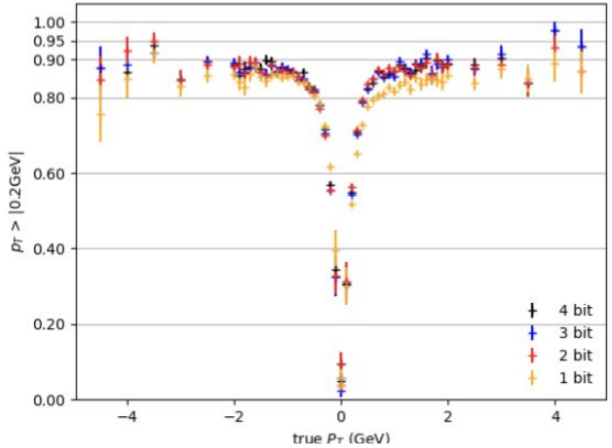
Data Reduction: Estimate 54.4% ~ 75.4%

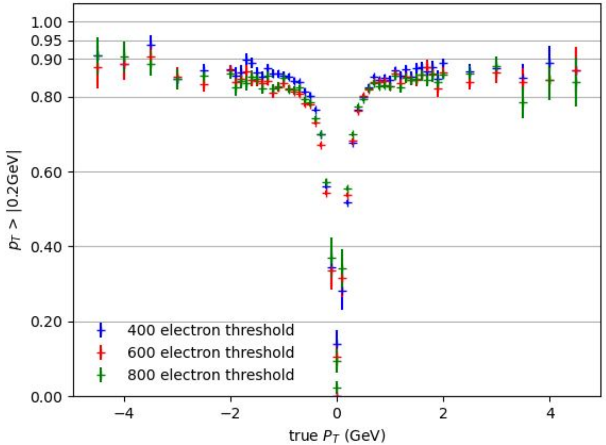
	Fraction of dataset	Rejection rate
Simulated tracks	40%	36.3%
Multi-pixel untracked	55%	61.9%
Single pixels	5%	100%

- Current detector only rejects single pixels; We can vastly improve on this!
- We reject 36.3% of simulated clusters (40% of dataset), 61.9% of multi-pixel untracked clusters (55% of dataset), and all single pixels (5% of dataset), giving a lower bound data reduction rate of 54.4%
- If we reject all untracked clusters, get an upper bound data reduction rate of 74.5%
- Since data readout is proportional to number of pixels in a cluster, if we reweight clusters by number of pixels, we reduce data by 54.4 ~ 75.4%

Model Quantization

- y-profile: 2-bit quantization chosen
- y-position quantized to 6-bit
- QKeras library for quantization-aware training
- Also, 400 e- electron threshold chosen

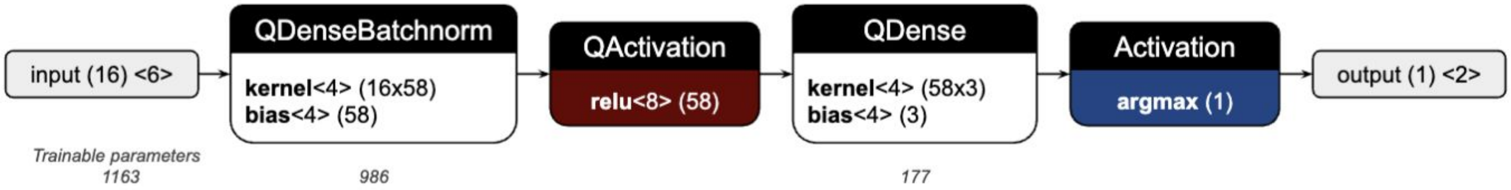




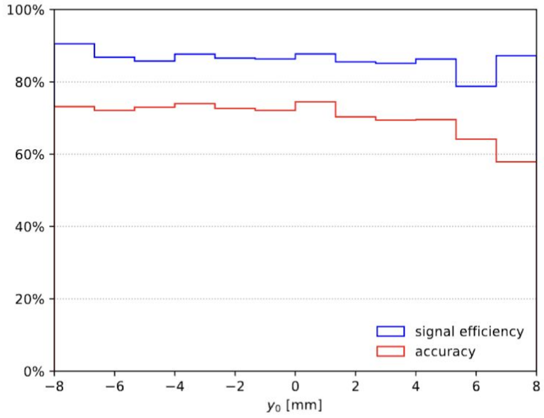
qkeras

On-chip implementation

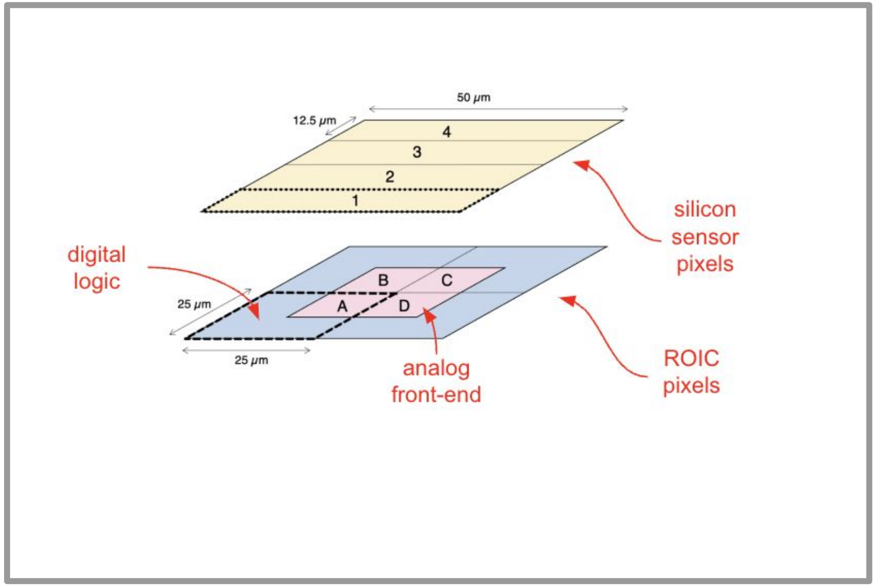
- Design space optimization

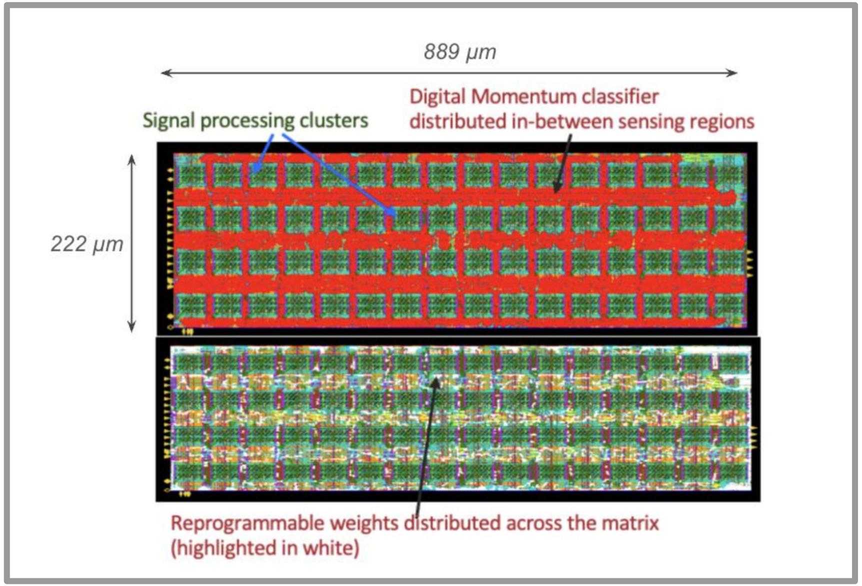


- Region specific implementation
  - 13 locally customizable (reprogrammable weights) neural networks implemented directly in the front-end
  - Reconfigurable weights so we can adapt to changing detector conditions



ROIC chip



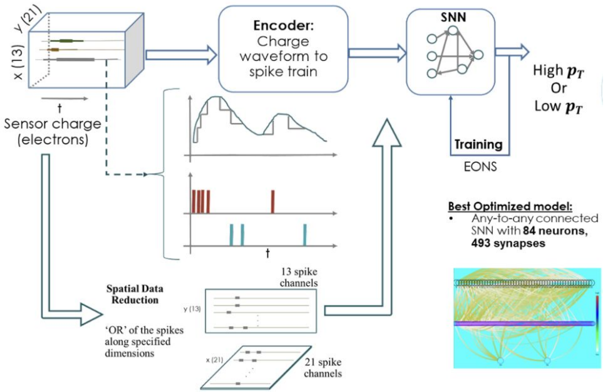


- 4 analog frontends, surrounded by a digital region
- Simulation: 13 x 21; Chip: 16 x 16

- Design expected to operate at < 300 µW
- Area < 0.2mm^2

Future Directions

- Hardware implementation
  - Tapeout expected by the end of this year
  - CPAD talk : Benjamin Parpillon: [Readout IC for future Phase III high luminosity upgrade of the large Hadron collider]
- Ongoing work
  - Studies on untracked clusters
  - Neuromorphic Approach with SNN
  - Regression studies: Train an algorithm to extract properties (positions, angles, and errors); expect further 5x improvement in data reduction!
  - Applications to other colliders (we're holding a workshop this Dec. Contact us for more info.!)
- Eventually enable improved AI performance through the ability to share data across layers (e.g., use photonic links)
- For more info!
  - Check out our preprint: https://arxiv.org/abs/2310.02474



Fig, by Shruti R. Kulkarni

Shruti R. Kulkarni et. al., On-Sensor Data Filtering using Neuromorphic Computing for High Energy Physics Experiments, ICONS '23: Proceedings of the 2023 International Conference on Neuromorphic Systems, Aug. 2023. https://dl.acm.org/doi/10.1145/3589737.3605976