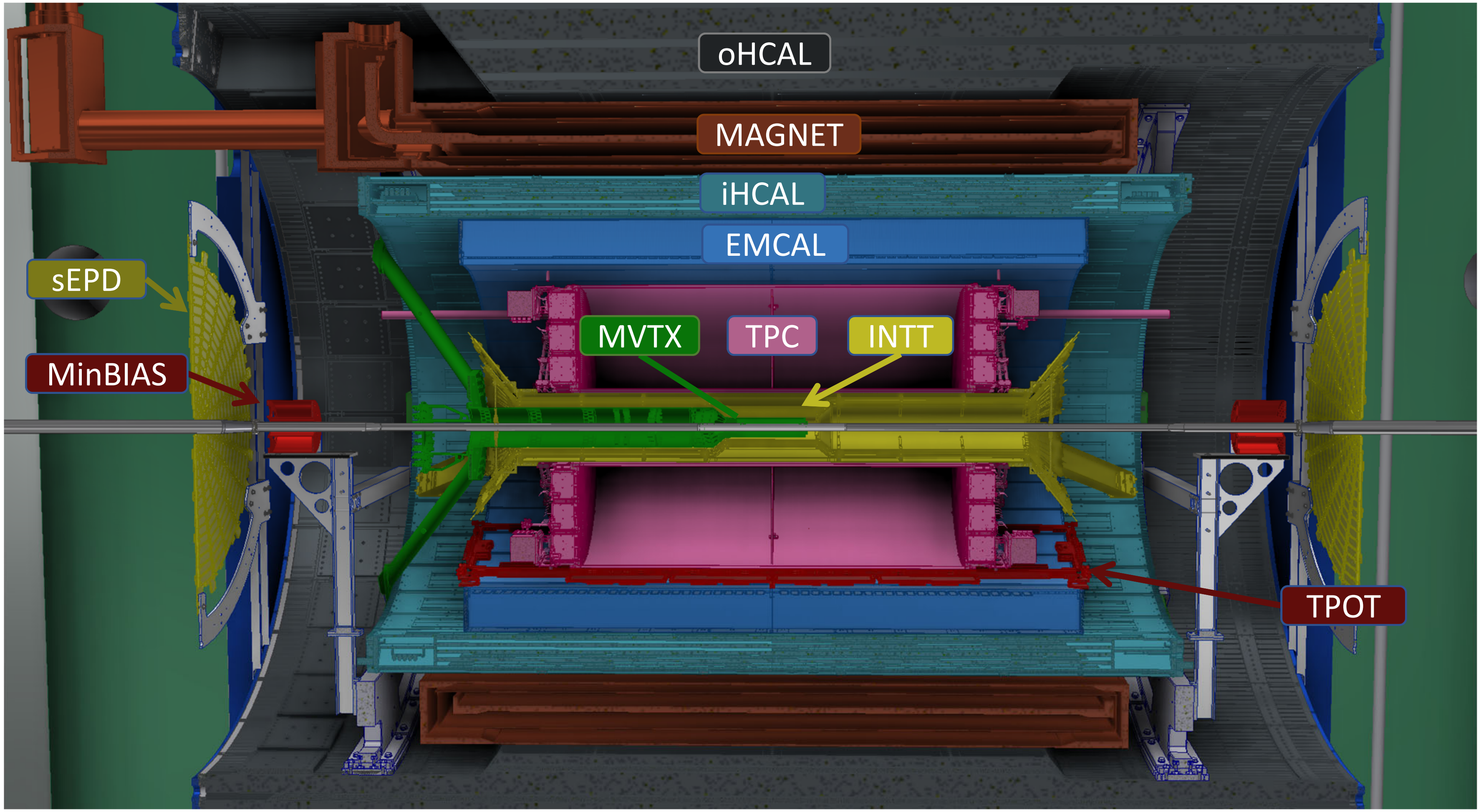
Fast Data Processing
&
Autonomous Detector Control
---- for sPHENIX and future EIC detectors

Huan Zhong Huang
Department of Physics and Astronomy, UCLA
for the sPHENIX Collaboration and the FastML Team

SQM 2024, Strasbourg, France

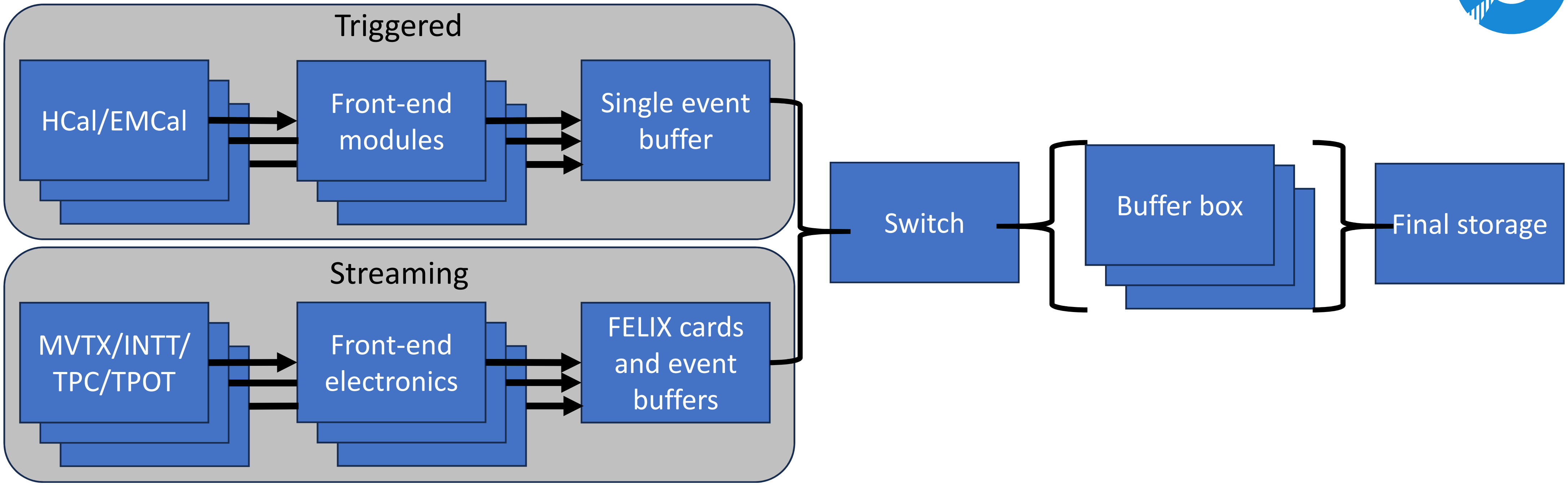
Our playground for p+p collisions

ZDCs on either side of IR



ZDCs on either side of IR

sPHENIX Readout Scheme



- RHIC pp collision rate is 3 MHz
- sPHENIX calorimeter DAQ max. rate is 15 kHz
  - Limits sPHENIX to recording ~0.5% of triggered proton-proton collisions
- Trackers are all streaming readout (SRO) capable
  - TPC dominates data rate, can't save all streamed data
  - 10% trigger-enhanced SRO increases open HF MB rate ~300 kHz

DOE AI/ML program

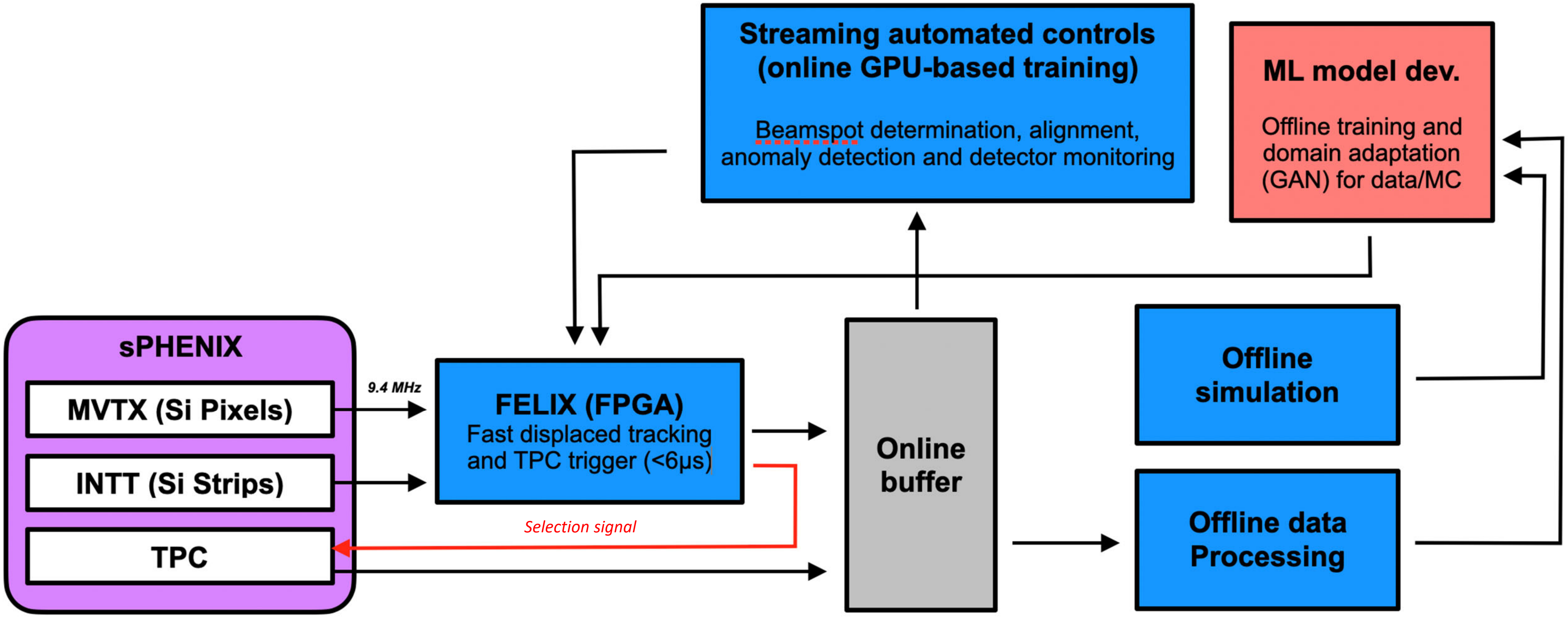
sPHENIX

Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors

A proposal submitted to the DOE Office of Science
April 30, 2021

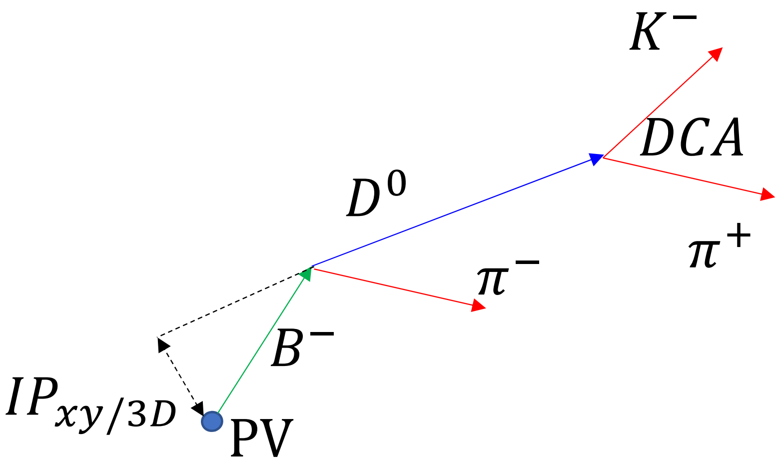
- **Stream MVTX and INTT Data to AI/ML Branch and Determine if reconstructable HF Topology is present;**
- **If Yes, Send tag downstream to enable Tracking Detector Readout**
- **Allows us to sample almost 100% of p+p collisions for rare HF physics**

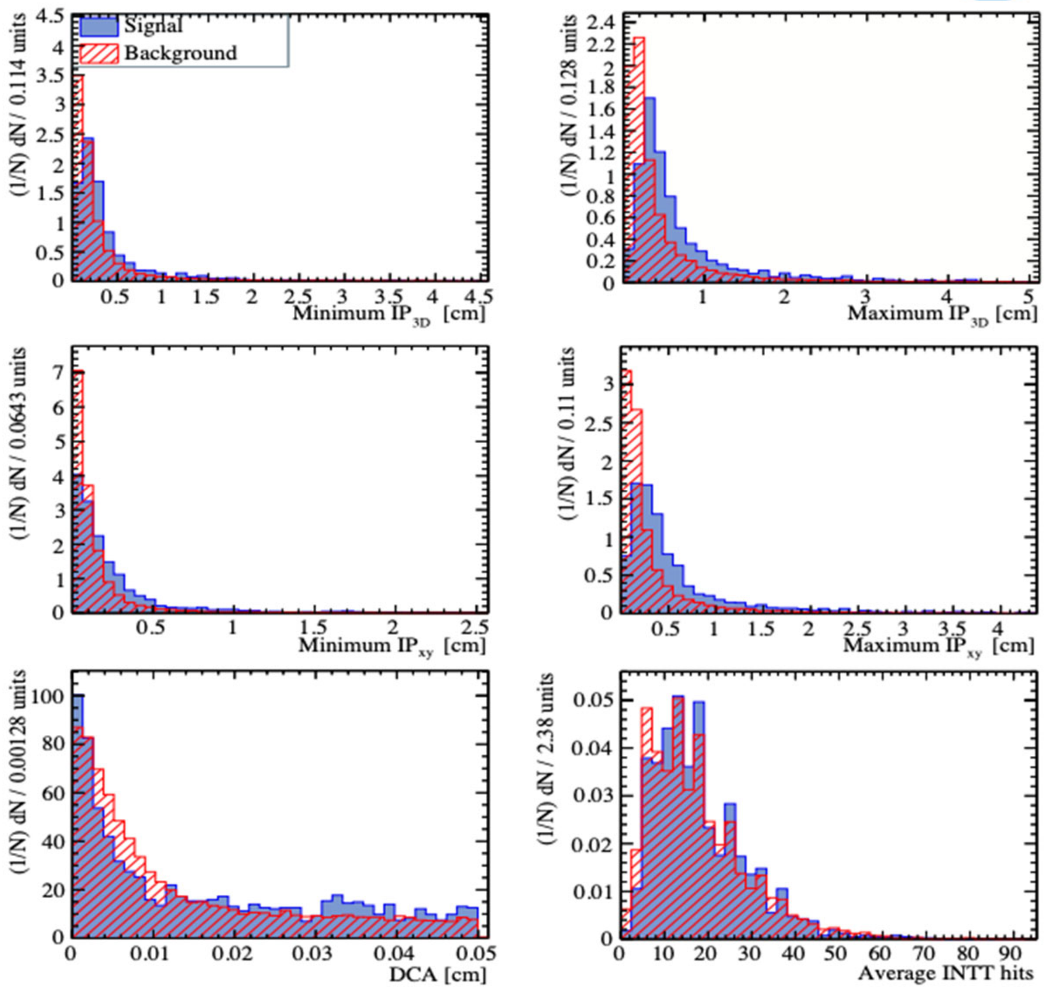
Block for AI/ML based decision making



AI HF selections

- Question: Can ML do better for selecting HF decays over conventional selections?
- Challenges: Decision time, Must run online, in FPGA. Hence variables must be "simple"





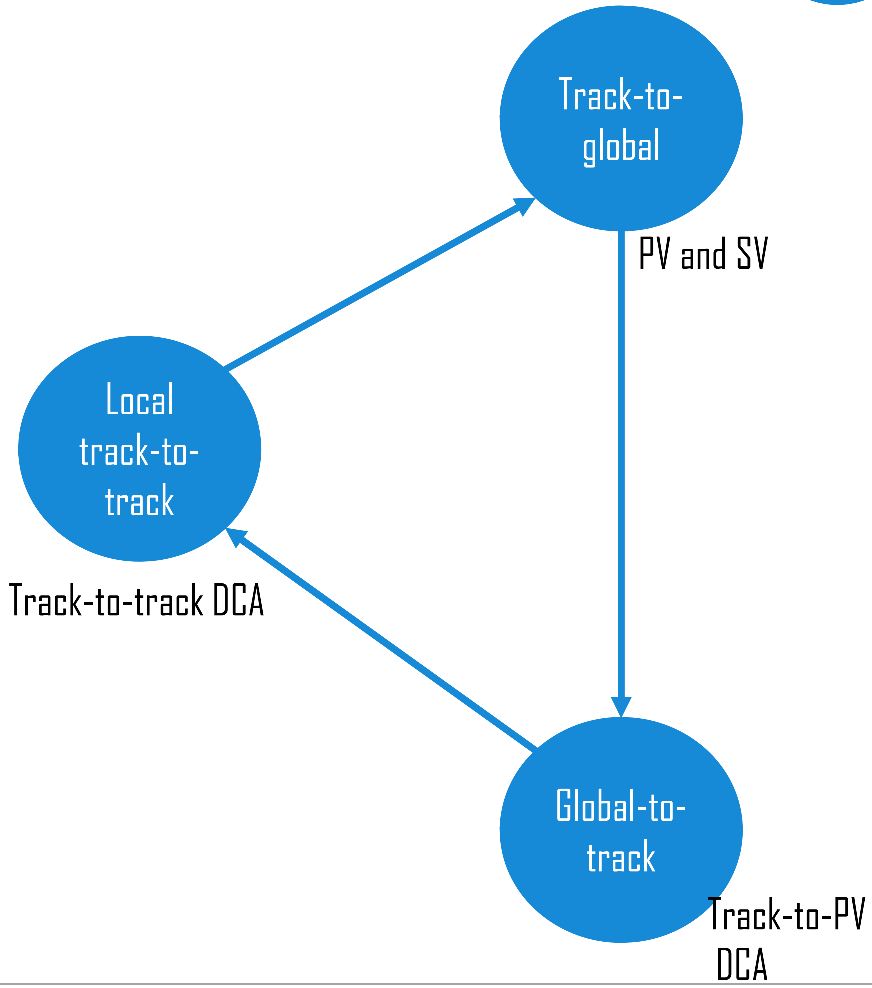
Constructing ML algorithms

- Developed algorithms as Graph Neural Networks (GNN)
- Advantageous over Convolutional Neural Networks (CNN) by adding edge information
- Algorithms deployed at several points on FPGAs:

1. Data decoding – Conventional logic
2. Hit clustering – Conventional logic
3. Fast tracking – Machine learning
4. Topological separation of HF signal from background – Machine learning

Feedback algorithms

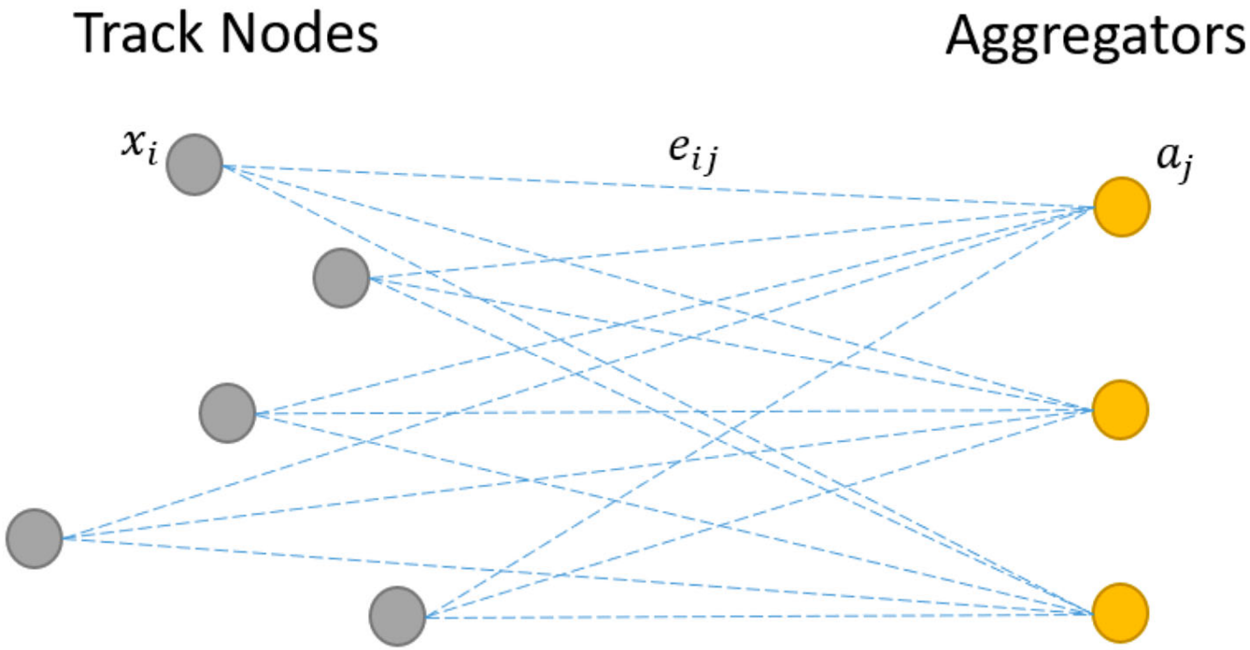
- Tracking algorithms developed using simulated signal and background events in the MVTX and INTT
- Used these models to feed into physics selection models to select interesting events
  - Models are bi-directional, local information is passed to global and global information is passed back to local to refine
- Initial trainings and models are developed on GPU
  - NVIDIA Titan RTX, A5000, and A6000
  - Will take the model and convert it to IP block for FPGA deployment
  - Models developed with PyTorch and PyTorch Geometric



Tagging with machine learning

Graph Neural Net design

- Track node input vectors
  1. 5 hits (MVTX + INTT)
  2. Length of each segment: $L = |\overrightarrow{x_{i+1}} - \overrightarrow{x_i}|$
  3. Angle between segments
  4. Total length of segments
- Aggregators
  1. Primary vertex
  2. Secondary vertex
- Current ML tracklet algorithm has
  - Accuracy > 91% for building tracks
  - Area under receiver-operating characteristic curve (AUC) > 97% liken to "probability of combining the correct track elements compared to incorrect elements" – random chance is 50%
  - Purity and rejection studies are underway

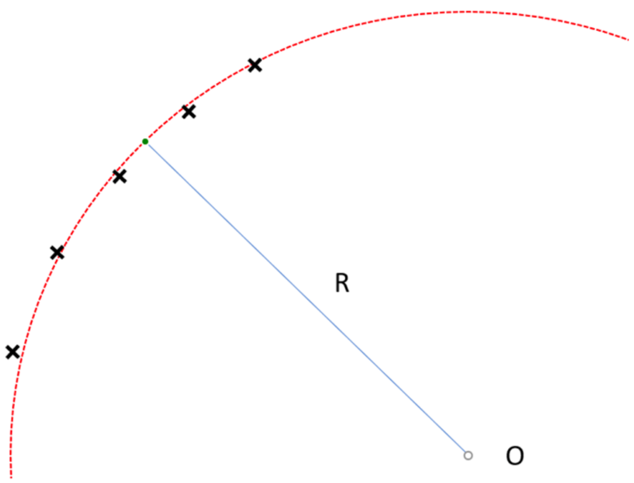


$e_{ij} = s_{ij}x_i$ is track-aggregator messages
s_{ij} is the weight

ECML PKDD 2022, Sub 1256

pT estimation

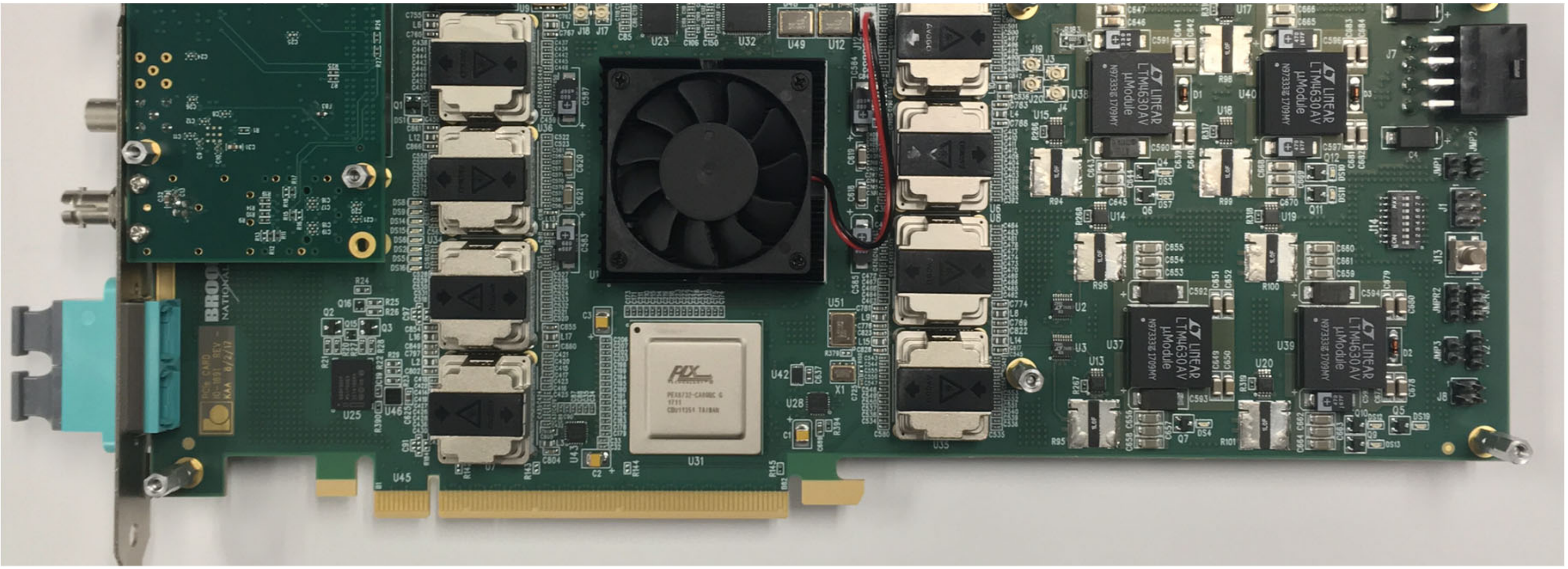
- A feed-forward neural net is used to predict the pT
- Uses least-squares method to estimate track radius
- ~15% improvement in tracking with pT estimation



Model	with LS-radius			without radius		
	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	300,802	84.17%	90.61%	300,418	69.80%	76.25%
GarNet	284,210	90.14%	96.56%	284,066	75.06%	82.03%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	355,042	**92.18%**	**97.68%**	354,786	**76.45%**	**83.61%**

Hidden dim	LS		MLP	
	Accuracy	AUC	Accuracy	AUC
32	91.52%	97.33%	91.48%	97.31%
64	92.18%	97.68%	92.23%	97.73%
128	**92.44%**	**97.82%**	**92.49%**	**97.86%**

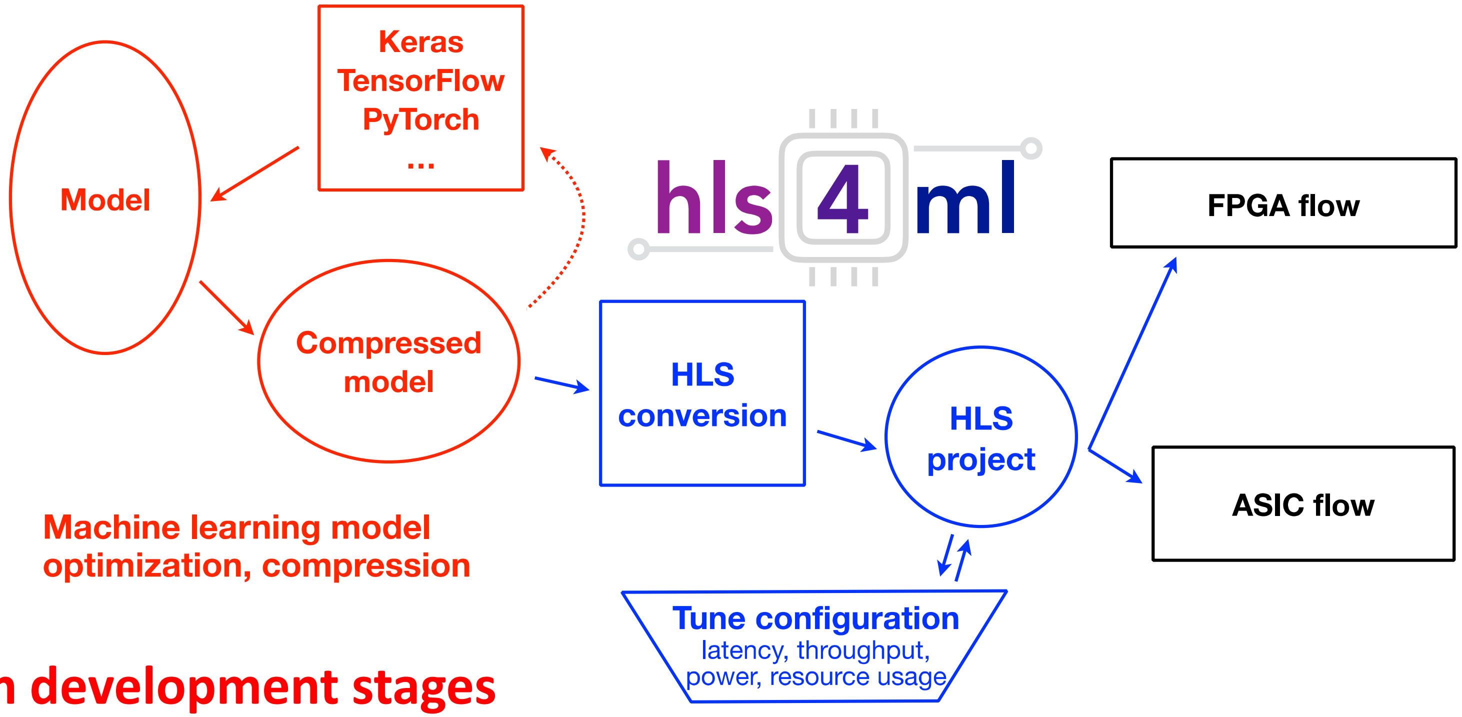
Hardware design



- Decision hardware is currently a BNL-712 FELIX board
  - Same as deployed at sPHENIX for ease of integration
  - Team can successfully transfer data from BNL-712 to KC-705 evaluation board
- Ongoing work on reducing resource usage

From Development to Firmware Implementation

- Algorithms must have low latency and resource use
- *hls4ml* translates NN algorithms into high level synthesis
- Also generates IP cores for easy implementation



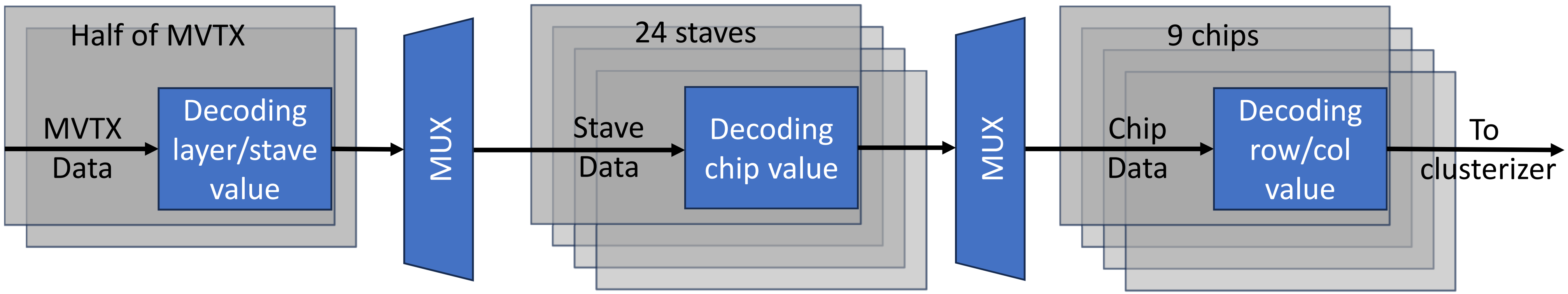
Red – typical ML algorithm development stages

Blue – HLS conversion to IP

Black – typical implementation onto chips

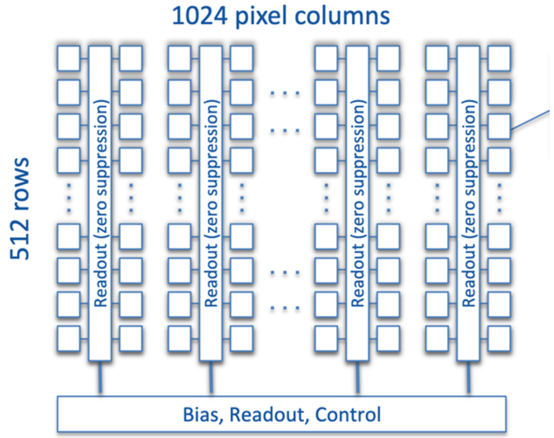
Decoding for MVTX

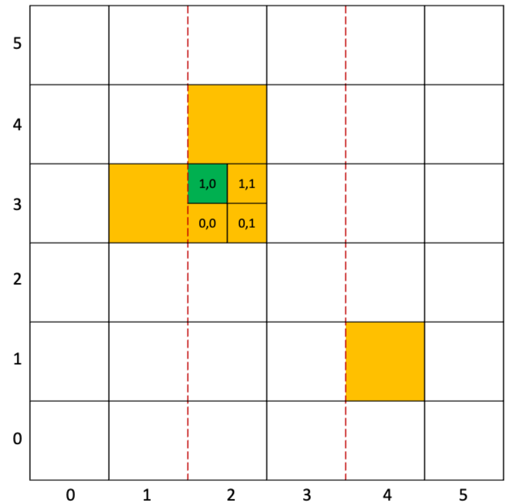
- Entire decision making must be performed in roughly 10 μs to allow recording of TPC hit
  - Parallelization of complex tasks in necessary to achieve this
- MVTX alone consists of 432 pixel chips with > 500k pixels / chip
  - 48 staves x 9 chips / stave
- Luckily, occupancy is low, ~ 20 hits / chip / collision for proton-proton collisions
- Each chip's information is sent to its own decoder to find active pixels



Clustering of MVTX pixels

- ALPIDE reads data out in double columns from 0 to 1023
  - Decoded hits thus arrive double column-by-double column
- Clusters can be assembled as they arrive
  - No hits in the next columns three adjacent pixels means cluster is ready to be sent out
- After finding pixel with centroid, pixel can be divided into grids to improve resolution using only 2 more bits
- Can get 13.5 μm cluster resolution at the global level from 31 bits
  - 6 bits to define layer and sensor number
  - 4 bits to define chip number on the sensor
  - 21 bits for cluster position on chip (9 for row, 10 for column, 2 for quadrant)
- After changing to global cluster position, detector layout has become abstracted

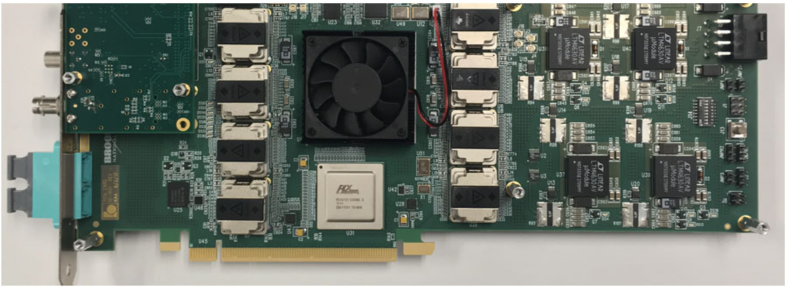




Putting it all together

- Tracking GNN has been synthesis and benchmarked on Alveo U280 accelerator card

Look up tables	14.9% (194k)
Flip flops	8.2% (214k)
Block RAM	20.2% (406)
Digital signal processing	5.4% (488)



Felix board will be used

- Processing time is undergoing rapid improvements
  - 380 μs in August 2023
  - ~9 μs in May 2024
- Second stage of the algorithm uses tracks to construct secondary vertices, a signature of particle decays

Secondary vertex finding with sim. $D^0 \rightarrow K^-\pi^+$ signal and random background for 1% sig. to bkg. tuning

Bkg. track rejection	Signal eff.	Sample purity*
90%	72.5%	7.25%
95%	48.9%	9.78%
99%	15.0%	15.0%

* % of final events with signal you're looking for

Status and Outlook

We expect to deploy the AI/ML decision module in the summer of 2024

The tracking detectors’ AI/ML aided stream readout will greatly benefit the sPHENIX scientific program for rare particles in 2024 p+p run

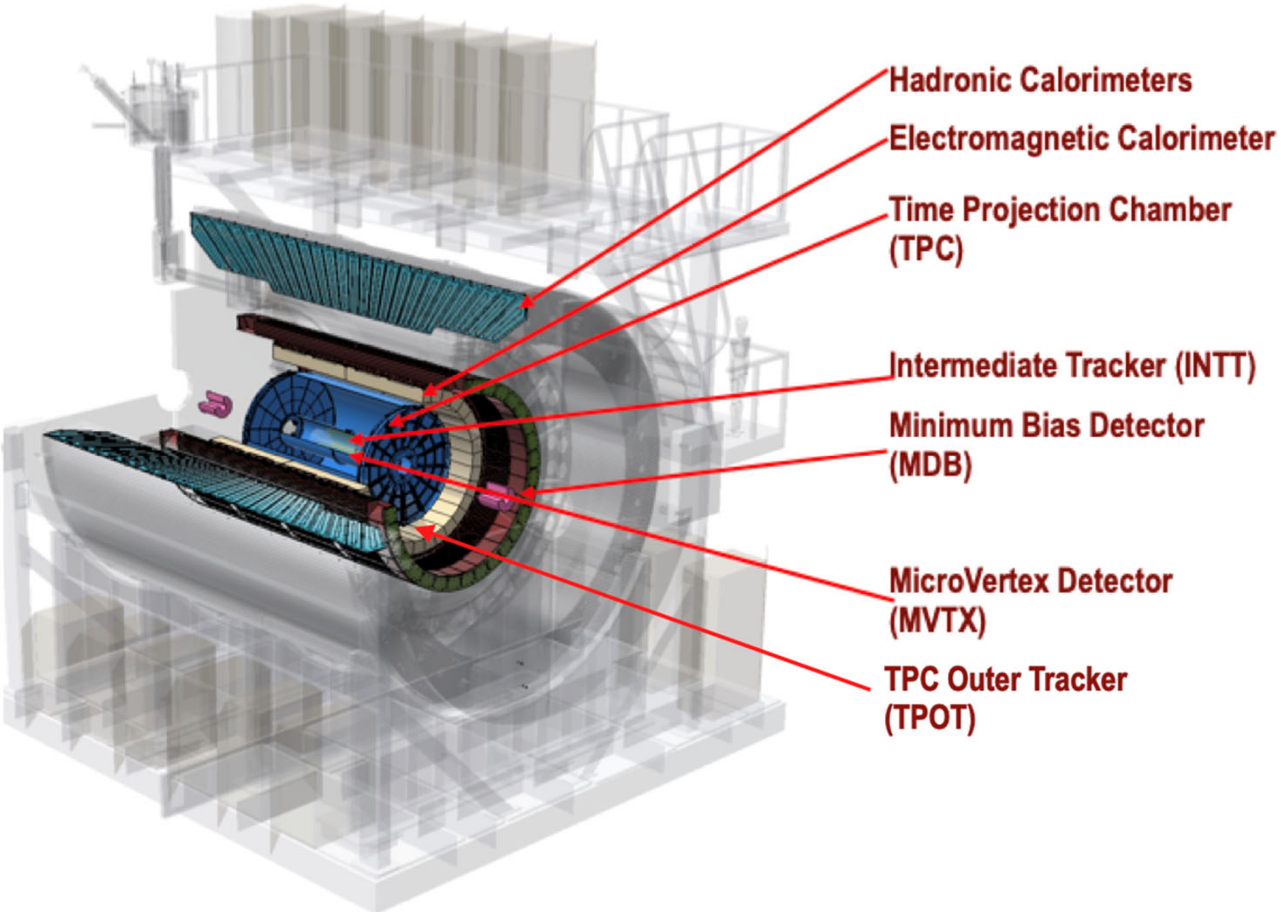
The FastML Team will extend the development of the project for future EIC

Backup

The FastML Team

- Cross-discipline group of computer scientists, engineers and physicists
- Formed in 2020 from DE-FOA-0002490
- Consists of groups from
  - Los Alamos National Laboratory
  - Massachusetts Inst. of Technology
  - New Jersey Institute of Technology
  - Fermilab
  - Oak Ridge National Laboratory
  - Stony Brook
  - Georgia Institute of Technology
  - University of North Texas
  - Central China Normal University

sPHENIX



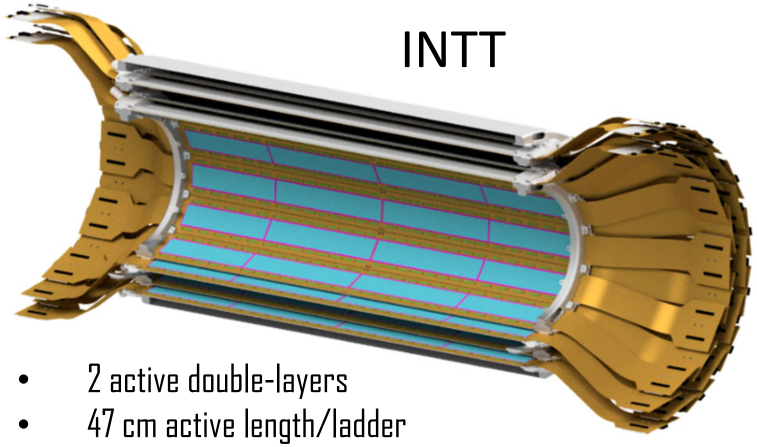
First run year	2023
$\sqrt{s_{NN}}$ [GeV]	200
Trigger Rate [kHz]	15
Magnetic Field [T]	1.4
First active point [cm]	2.5
Outer radius [cm]	270
$\lvert\eta\rvert$	$\leqslant 1.1$
$\lvert z_{vtx}\rvert$ [cm]	10
N(AuAu) collisions*	1.43×10^{11}

* In 3 years of running

Tracking at sPHENIX

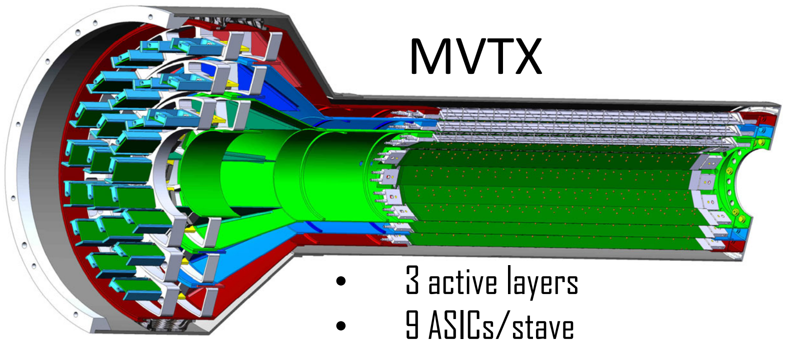
- Tracking consists of 3 sub-detectors:
  - Pixel Vertex Detector (MVTX)
  - Intermediate Silicon Tracker (INTT)
  - Time Projection Chamber (TPC)
- MVTX and INTT are both capable of streaming readout
- Combined tracking to r = 10.3 cm

INTT



- 2 active double-layers
- 47 cm active length/ladder
- Silicon strip detector

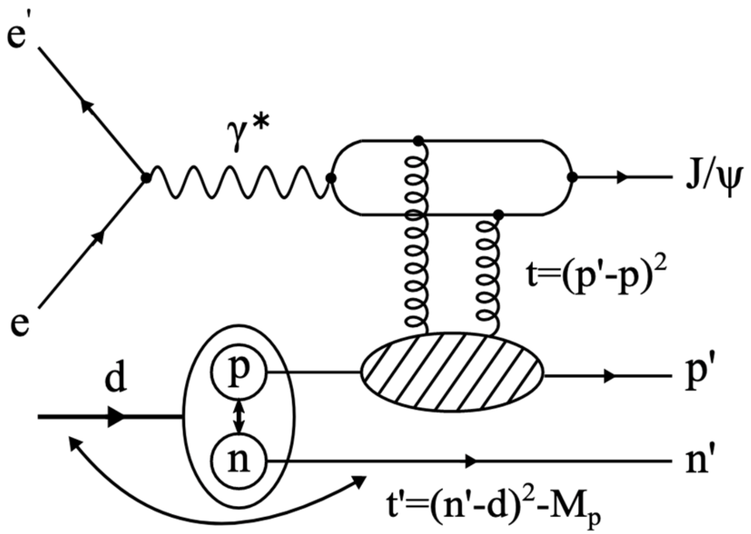
MVTX

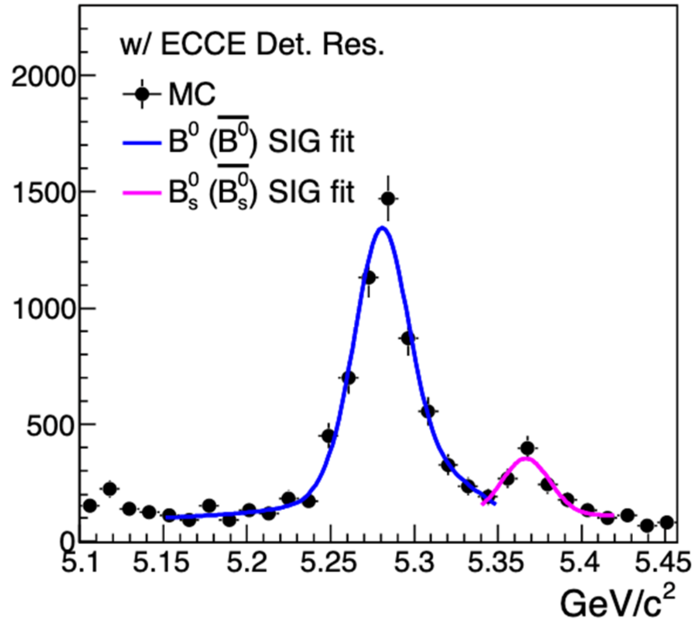


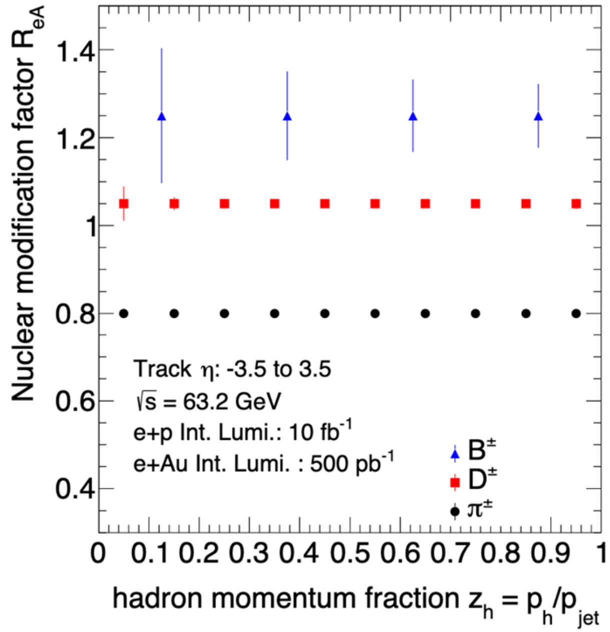
- 3 active layers
- 9 ASICs/stave
- 27 cm active length/stave
- Pixel detector

Heavy flavor at the EIC

- Why?
  - Main HF production is through photon-gluon processes
  - Good probe of gluon parton distribution function







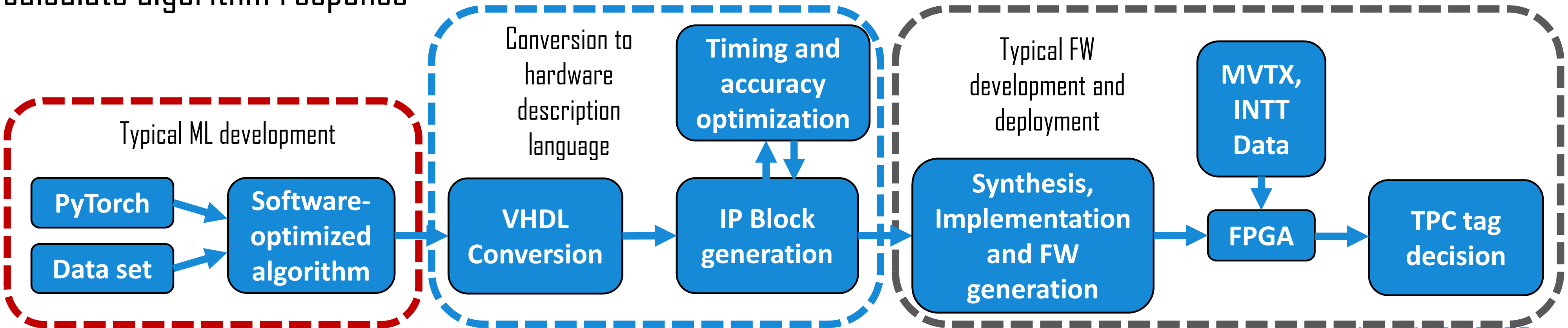
arXiv.2207.10632
arXiv:2103.05419

Development of Tagging with machine learning

- Algorithms must have low latency and resource use
- hls4ml translates NN algorithms into high level synthesis
- Also generates IP cores for easy implementation
- Rest of firmware can be built around IP core to calculate algorithm response

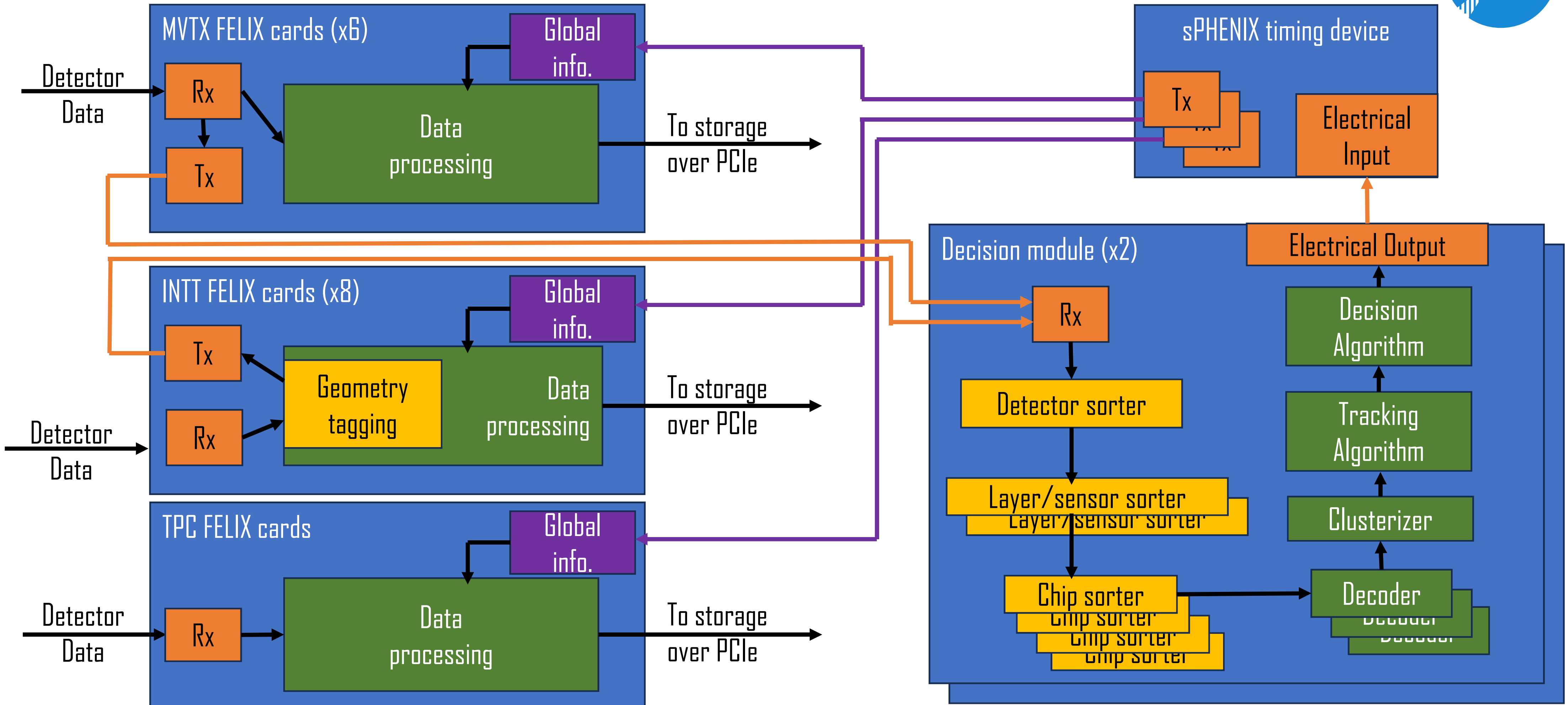

Server for algorithm conversion and FW generation


FELIX card (712) on server for FW testing

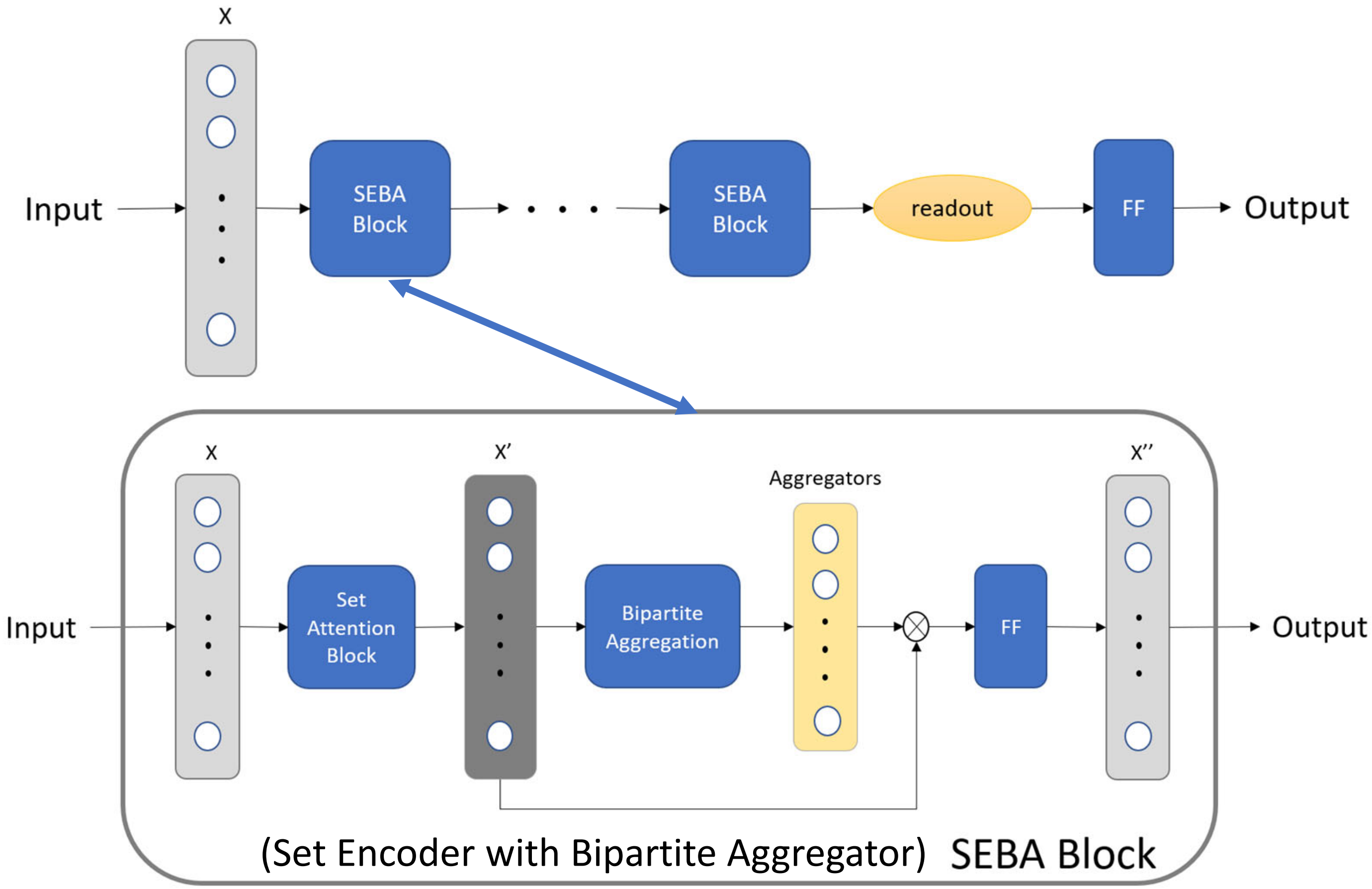


arXiv 2103.05579

Overview of Data Process/Readout Scheme



GNNs with set transformers



The cycle

1. Track information is initially defined
2. This is relayed to all primary and secondary vertex information
3. Weights are assigned to each link
4. The PV and SV information go through a feedforward NN
5. This updates the track information