ML-enabled End-to-End Tracking Reconstruction and Trigger Detection

Supported by DOE grant DE-SC0019518

Yu Sun, PI
CEO of Sunrise Technology Inc.
yu.sun@sunriseaitech.com
DOE SBIR NP Exchange Meeting, August/17/2023

About Sunrise Technology Inc.

- Founded in 2017
- Located in an incubator at Stony Brook University, Long Island, NY
- Team: three full-time employees, a part-time consulting scientist, and several graduate interns.
- Developing advanced AI/machine learning technology for autonomous systems, such as scientific experiments decision-making engines and education platforms.
- Projects

1) ML-based slow orbit feedback control, deployed at BNL NSLS-II in July 2023

2) Autonomous driving education toolkit

3) Collaborated with CFN at BNL to use machine learning method for x-ray scattering image classification

4) Particle Collision Triggering

SBIR Phase II Objectives

- SBIR Phase II award
  - Title “High Performance FPGA-based Embedded System for Decision Making in Scientific Environments”
  - Co-funded by NP and ASCR
  - End Year 3

- Ultimate Goal
  - Design real-time AI-enabled DAQ trigger algorithms applied to the very high-rate data streams from detectors.
  - Play a central role in sPhenix and future EIC detectors running under trigger systems and in-situ streaming analysis for event selections.

- Phase II Technical Objectives
  - Designing Graph Neural Networks for High-Speed Physics Event Triggers.
  - Collaborate with sPhenix team to integrate the algorithms to sPhenix experiment and reaches the target of 15Khz data acquisition rate.

- Phase II Commercialization Objective
  - Manufacture smart embedded system to facilitate real-time data collection for experiment and facility control

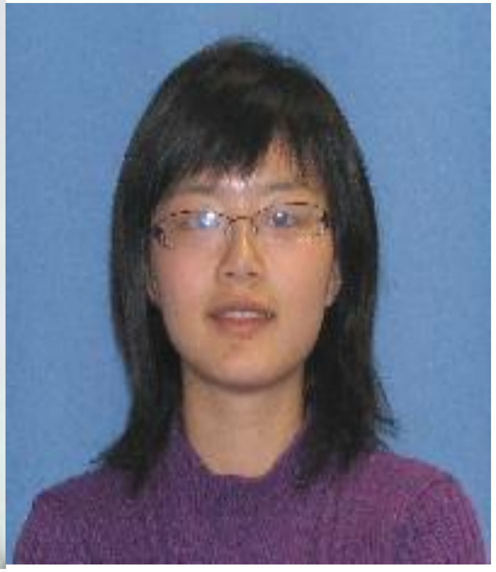
Team on this project

Yu Sun, PI

Giorgian Borca-Tasciuc

Kevin Mahon

Tingting Xuan



Yimin Zhu

Collaborators

- Dr. Ming Xiong Liu, Dr. Cameron Dean, LANL
- Dr. Jin Huang, Dr. Zhaozhong Shi, BNL

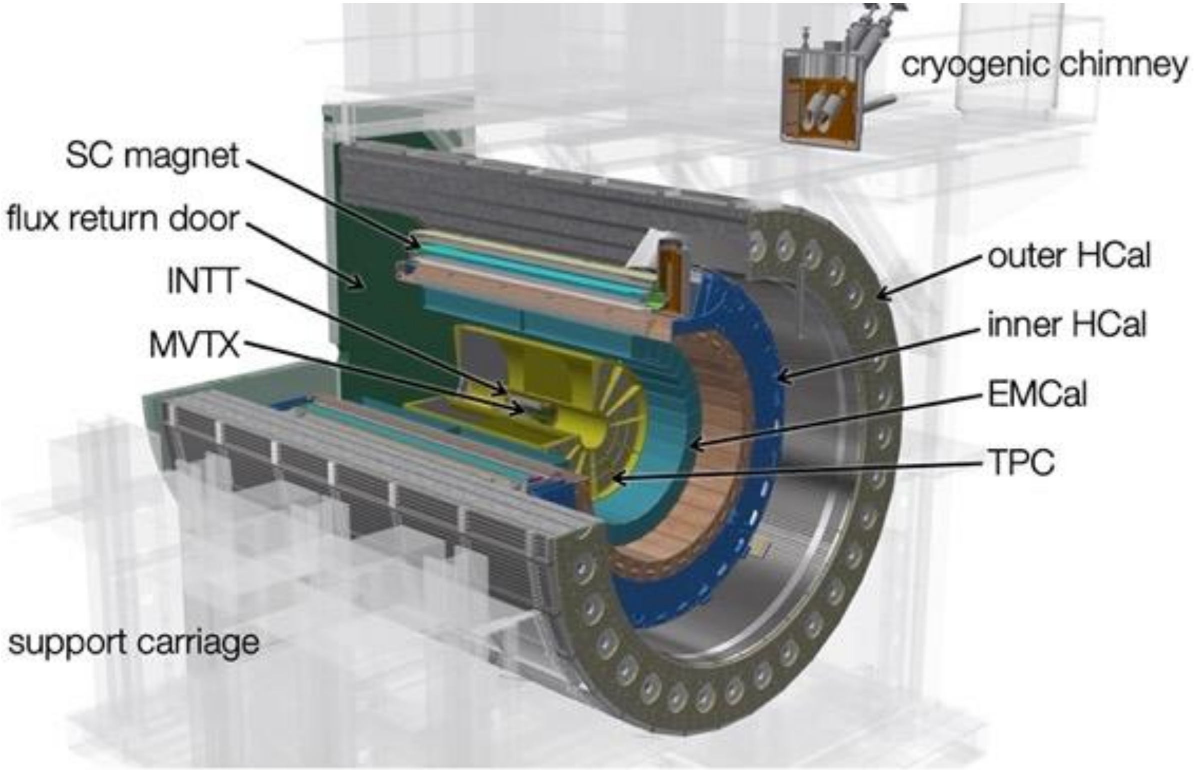
Motivation

- The readout challenge
  - Raw data Speed and Volume >> Hardware bandwidth/Storage Capacity
    Only a small fraction of data will be recorded to tape
- Trigger events are very rare, ~0.1% probability at RHIC
  - RHIC collision rate is several MHz, sPHENIX readout 15 kHz

  -Without an effective trigger algorithm, experiments must use random event taking.

  - With the same level of recall, AI-based trigger will significantly improve the detector efficiency.
- Integrate the AI-based trigger system into the sPHENIX experiment for p+p run in 2024
- Potential future deployment on Electron-Ion Collider (EIC)

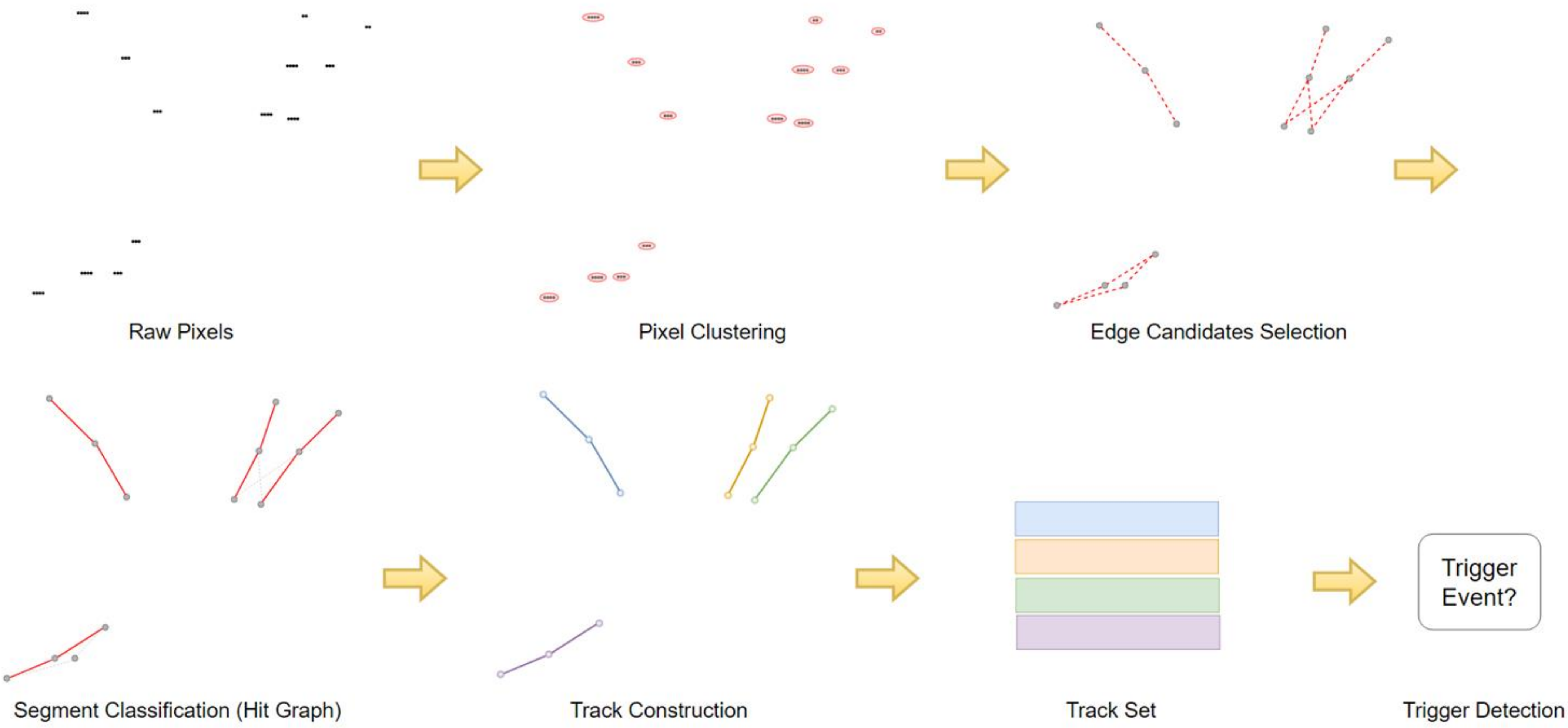
sPHENIX experiment



sPHENIX experiment under construction at RHIC:

- Running period 2023-2025
- ~4m long, ~3m high, 1000 tons
- 15kHz trigger rate
- 3 MVTX layers and 2 INTT layers - detectors capable of streamed readout

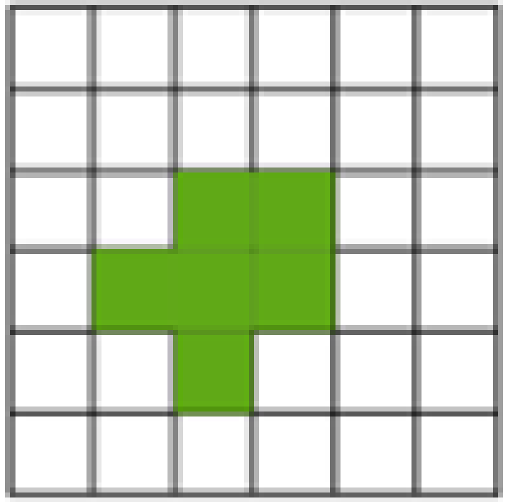
ML Solution Overview



Pixels $\mapsto$ Hits

From Pixels to Hits - Clustering

- Clustering is done by solving a spanning forest problem
- There is an edge between pixels that are adjacent to each other
- Mean of all pixels in a cluster is taken as the hit location
- Most time-consuming portion, we are developing a sparse CNN to perform faster clustering

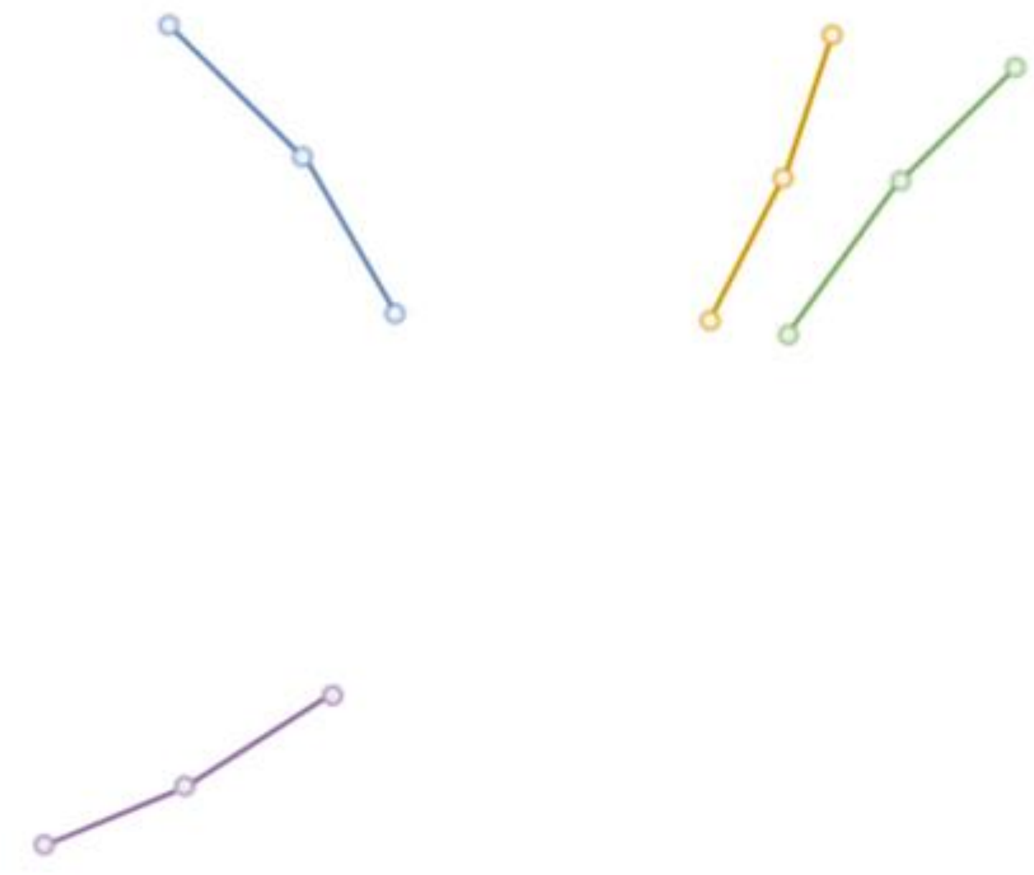


Pixels on Detector

Hits $\mapsto$ Tracks

From Hits to Tracks

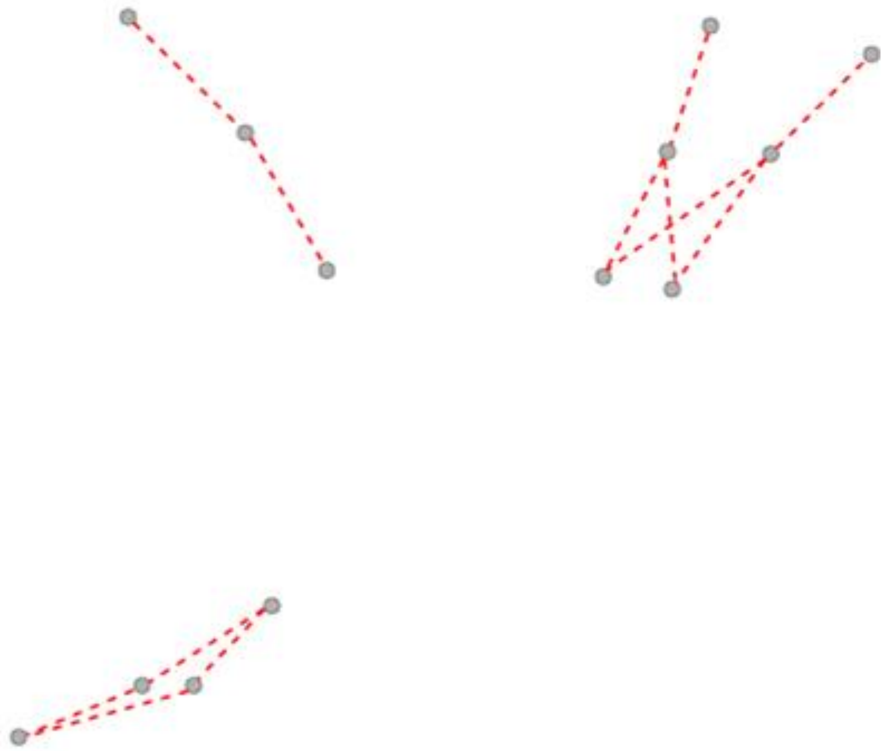
- Once we have hits, we want to group hits that came from the same particle into a track
- This will be solved by treating the problem as an edge classification problem
- Out of the N^2 possible edges between the hits, we want to know the true edges.



Track Construction

Edge Candidate Selection

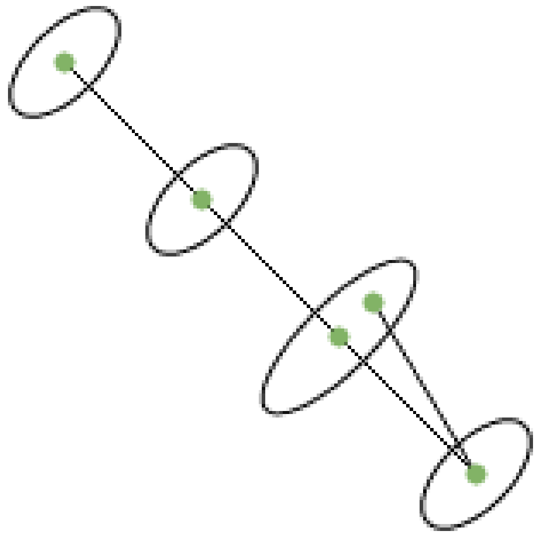
- Not all of the N^2 possible edges are plausible - we can eliminate a lot of edges from the get-go
- We can use some basic geometric constraints on the cylindrical coordinates of the hits
  - $|\Delta\varphi/\Delta r| <= \text{PHI\_SLOPE\_MAX}$
  - $|z_0| <= \text{Z\_ORIGIN\_MAX}$
  - $z_0 = z_1 - r\cdot(\Delta z/\Delta r)$
- The geometric constraints determine much of the latency and will play a vital role in further reducing the FPGA latency.

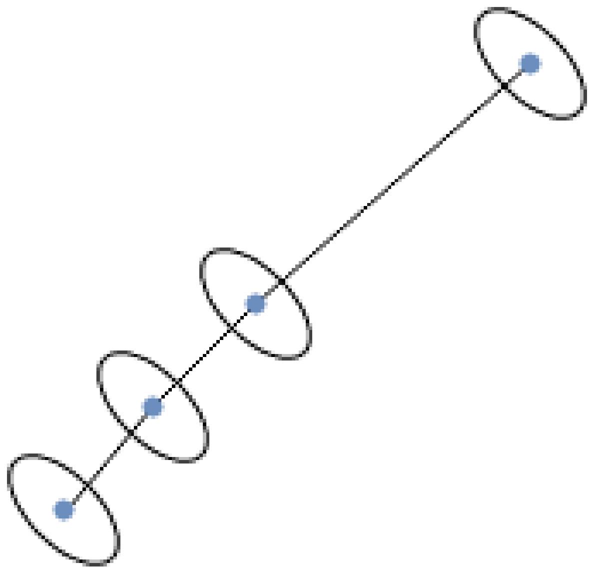


Edge Candidates Selection

Track Construction

- Once edge classification is performed, a track is constructed by finding the connected components





Track Construction Performance

	2022	2023
Accuracy	**96.30%**	92.07%
Precision	84.55%	**92.54%**
Recall	83.25%	**97.97%**
F1	83.89%	**95.18%**
Latency	**17.92µs**	**3.1725µs**

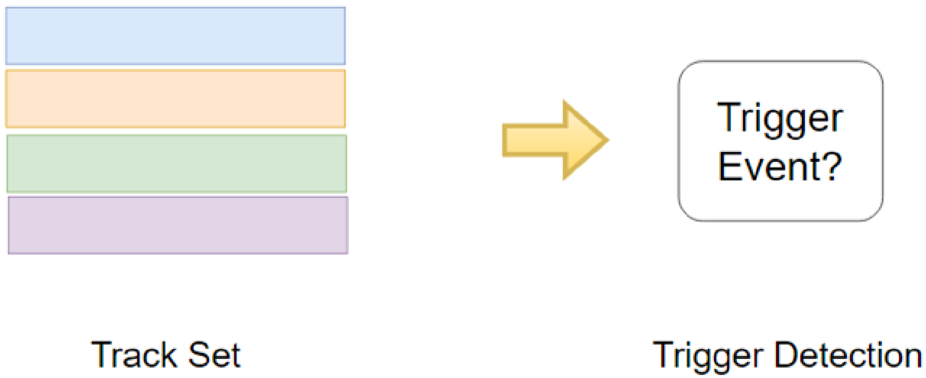
Software of Year 3 is much more hardware aware than that of Year 2!

- 1 iteration on hits generation instead of 4 iterations
- Hidden layer of MLP is reduced from 1024 to 8
- Much more constraints on geometry to select edge candidate

Tracks $\mapsto$ Trigger Label

From Tracks to Trigger

- After creating the tracks, we have a set of tracks
- We want to know whether the event that created these tracks was a trigger event
- A *trigger event* is an event in which we had a $D_0 \mapsto (\pi^+, K^-)$ or $D_0 \mapsto (\pi^-, K^+)$ decay



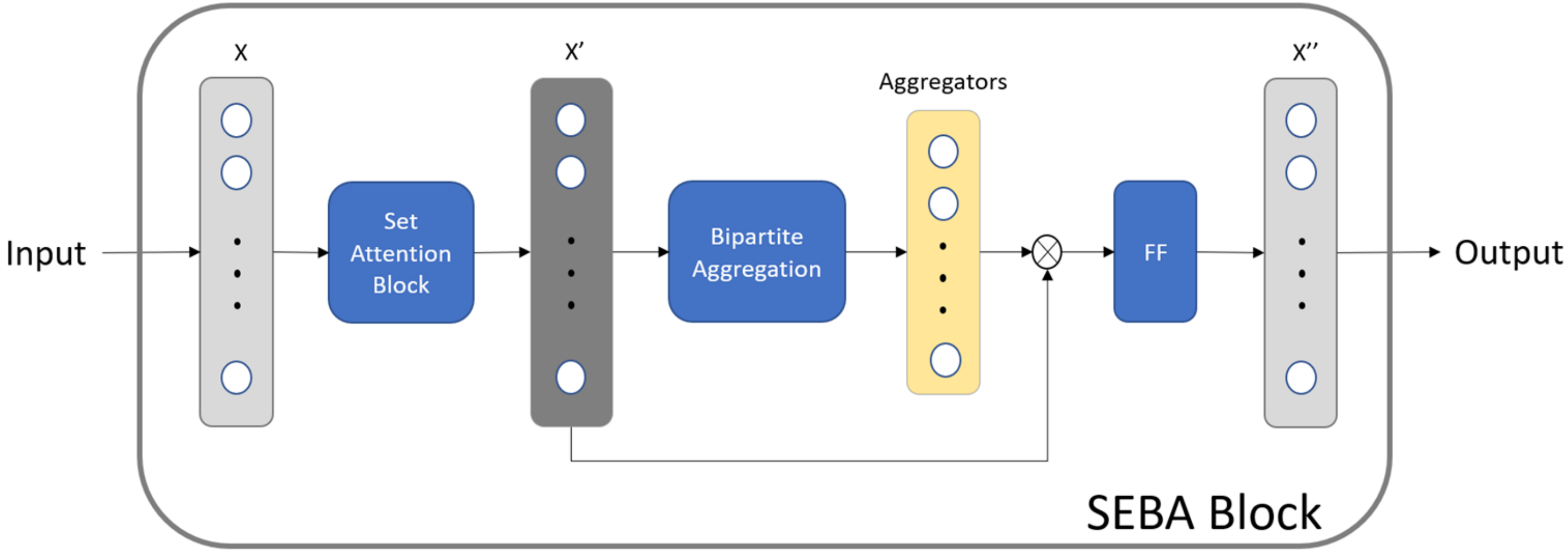
What needs to be modeled?

- $D_0 \mapsto (\pi^+, K^-)$ or $D_0 \mapsto (\pi^-, K^+)$
- Considering the problem from a high level perspective, we need to consider:
  - Track-to-track Interactions: Do these pair of tracks form a (π^+, K^-) or (π^+, K^-) pair?
  - Track-to-global Interactions: Where is the origin of this track?
  - Global-to-Track Interactions: Incorporate information about the origin of this track into the track embeddings

Architecture

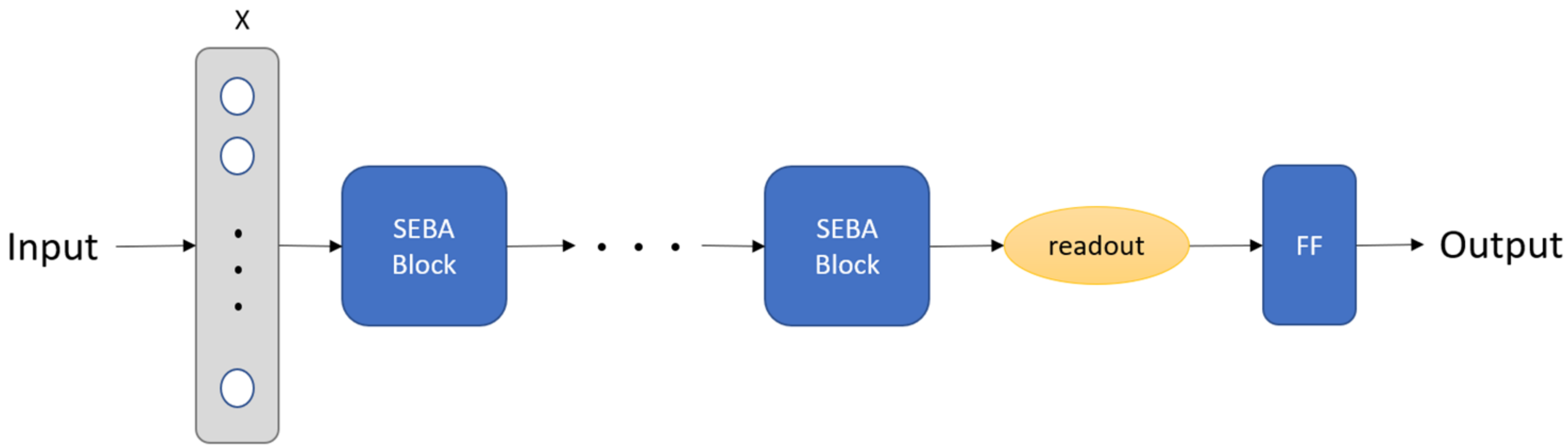
- Previous considerations motivate the following block.
  - Set Encoder: Track-to-Track interactions
  - Bipartite Aggregation: Track-to-Global and Global-to-Track interactions

SEBA (Set Encoder with Bipartite Graph Affinity)



Architecture

- Stack multiple SEBA Blocks
- Use Bipartite Aggregation with single aggregator to generate event embedding
- MLP on event embedding to predict Trigger Event



Physics Knowledge Added

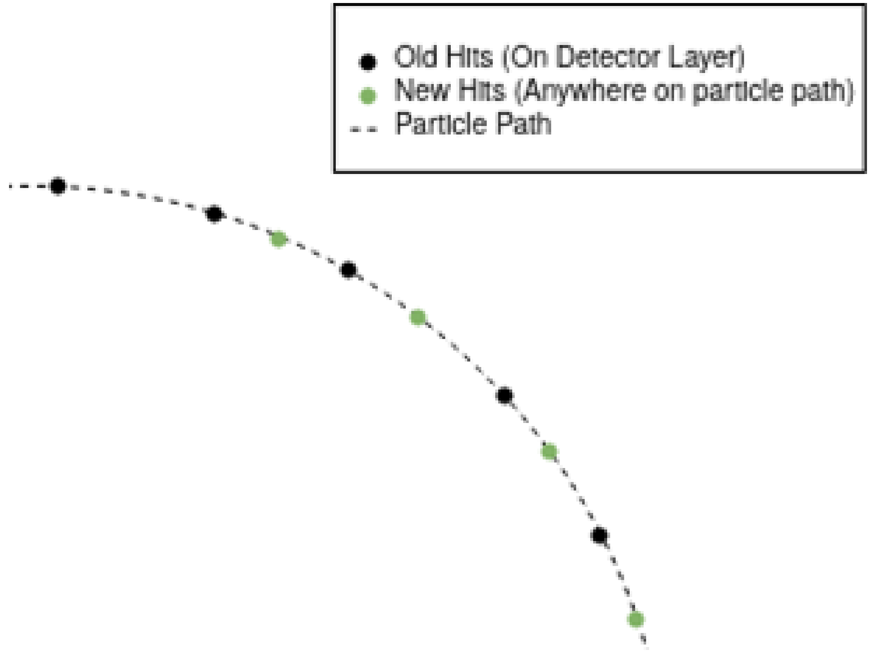
- Track given to trigger classifier has the following features:
  - (x, y, z) location of hit on each layer
  - Length segment between each layer
  - Angle formed by segments
  - Estimated radius of circle fit to hits
  - Estimated center of circle fit to hits
  - Estimated transverse momentum of track
- **Estimated radius and center** provided ~10% increase in accuracy

Multi-Task Learning to Improve model performance

- Several modifications to standard training process in order to improve the performance and robustness of our trigger algorithm
  - Data augmentation: We perturb hits off the detector layers while keeping it on the particle path

  - Track embeddings used predict whether two tracks come from the same parent

$$\mathscr{L} = L_{CE}(\text{trigger}_{pred}, \text{trigger}_{true}) + L_{CE}(A_{pred}, A_{true})$$



Trigger Prediction Performance

Data	Year	Metric	Result
Predicted Tracks	**2023**	**Accuracy**	**85.6%**
GT Tracks	**2023**	**Accuracy**	**90.22%**
GT Tracks	2023	Precision	86.35%
GT Tracks	2023	Recall	95.41%
Predicted Tracks	**2022**	**Accuracy**	**84.01%**
GT Tracks	**2022**	**Accuracy**	**87.5%**

Conclusion, Accomplishments and Milestone

- ML models have shown steady increases in performance on the triggering problem
- Incorporating physics knowledge has contributed to large performance improvement in trigger prediction
- Challenges remain in adapting the ML algorithm to the real-world latency and data availability constraints

Conclusion, Accomplishments and Milestone

1. Tingting Xuan, Yimin Zhu, Giorgian Borca-Tasciuc, Ming Xiong Liu, Yu Sun, Cameron Dean, Yasser Corrales Morales, Zhaozhong Shi and Dantong Yu. End-To-End Pipeline for Trigger Detection on Hit and Track Graphs, **The Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23),** Collocated with AAAI-23, February 7-14, 2023, Washington, DC.

2. Tingting Xuan, Giorgian Borca-Tasciuc, Yimin Zhu, Yu Sun, Cameron Dean, Zhaozhong Shi, and Dantong Yu. Trigger Detection for the sPHENIX Experiment via Bipartite Graph Networks with Set Transformer. **In Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022**, Grenoble, France, September 19-23, 2022.

Future Work

- Modifying algorithms to deal with pile-up
- Work on simplifying algorithms and reducing data quantity to meet latency challenges
    - Initial study of latency-accuracy tradeoff showed we could reduce edge quantity (critical for FPGA implementation) at the tracking stage by 60% with minimal loss in final trigger accuracy
- Ensure trigger algorithm works in explainable and robust way
    - Initial study has shown model prefers to drop non-trigger tracks without affecting event label and prefers to perturb hits as to not affect the track radius

	$d\phi_{max}$	dz_{max}	accuracy	Maximum Edge Candidates
0	0.025005	102.000000	0.885895	1030.0
1	0.014881	16.000000	0.885360	548.0
2	0.011599	155.000000	0.884555	638.0
3	0.026555	113.000000	0.884320	1077.0
4	0.024582	178.000000	0.883860	1022.0
5	0.010320	48.000000	0.882630	556.0
6	0.012193	14.220353	0.881850	463.0
7	0.030000	200.000000	NaN	1171.0

Test model with real sPhenix experimental data!!! (end of 2023 expected)

Acknowledgement

- Thank DOE Office of Science, Dr. Michelle Shinn for funding this project, and every contributor for working on this project!