Brookhaven National Laboratory

U.S. DEPARTMENT of ENERGY

Tracking our Treasures

Punit Sharma

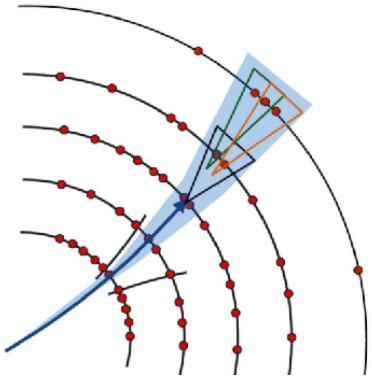
April 28, 2026

@BrookhavenLab

Overview

Some overview of the tracking

- Dataset Generation
- Pipeline Overview



Low Level (Level 4?) data not available in the ATLAS Open Data!

ACTS + ODD

ACTS- A Common Tracking Software

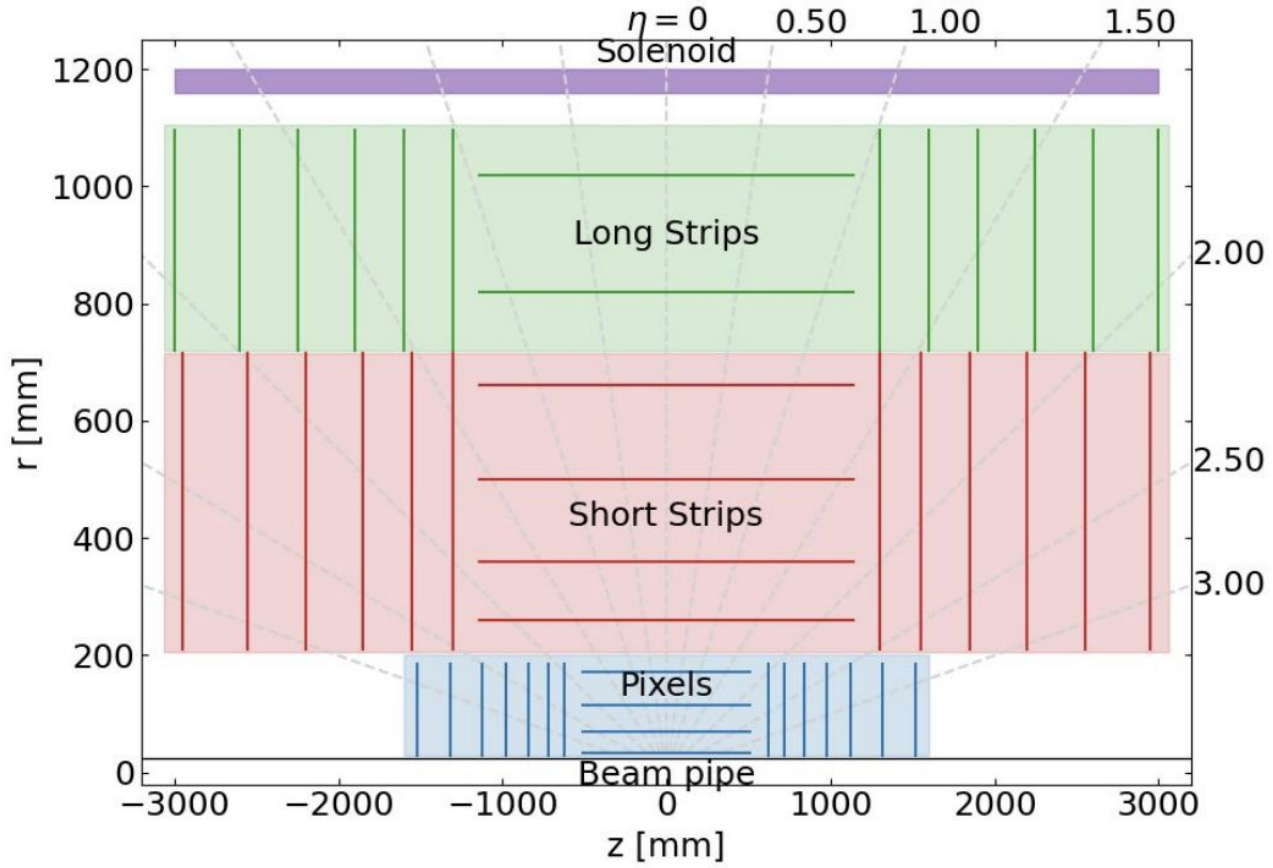
- ACTS provides algorithms for clusterization, track finding, track fitting, vertexing, …
- Benefits of being experiment-independent: knowledge transfer, common solutions for various domain specific problems
- Already used in production by various experiments

ODD- Open Data Detector …

- provides a template (HL-)LHC style particle detector for algorithm research and development
- An evolution of the TrackML detector with more realistic material budget and full simulation support
- Full silicon prototype inner detector based on DD4hep
- The detector description does not specify digitization
- Used in Acts for validation and performance studies
- For our studies only the silicon part was important

ODD layout

- Pixel
  - 2D + time
  - Resolution: 15 μm spatial, 25 mm time
  - 4 barrel layers
  - 7 endcap disks
- Short Strip
  - 2D
  - Resolution: 43 μm / 1.2 mm spatial
  - 4 barrel layers
  - 6 endcap disks
- Long Strip
  - 1D, two sided, with stereo angle
  - Resolution: 72 μm spatial
  - 2 barrel layers
  - 6 endcap disks



Collider ML dataset

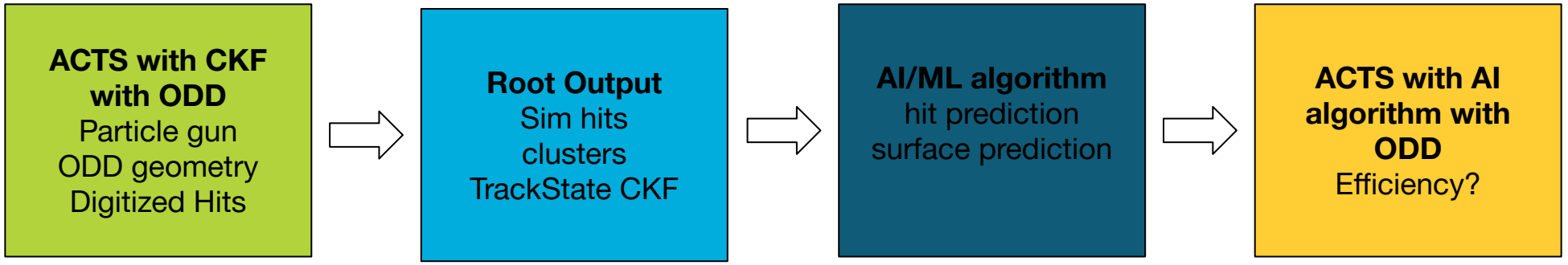
Pipeline

Simulation

	ID	Process Label	Process Type	Comments
$<\mu>=0,200$	1	ttbar	$pp \to t\bar{t}$	Top quark production; important for SM tests and complex final states.
$<\mu>=200$	2	zee	$pp \to Z \to e^+e^-$	Drell-Yan process; detector calibration and electroweak probe.
$<\mu>=200$	3	zmumu	$pp \to Z \to \mu^+\mu^-$	Drell-Yan process; detector calibration and tracking probe.
$<\mu>=0,200$	4	dihiggs	$gg \to HH$	Gluon fusion di-Higgs; benchmark for high-Luminosity studies.
	5	diphoton	$pp \to \gamma\gamma$	Diphoton production; loop-induced QCD/EW process.
	6	multijet	$pp \to$ jets	QCD production; dominant background for hadronic signatures.
$<\mu>=0,200$	7	ggf	$pp \to H$	Gluon-gluon fusion; main Higgs production channel.
	8	susy	$pp \to \tilde{g}\tilde{g}$	GMSB gluino production; $\tilde{\chi}_1^0$ NLSP provides displaced vertex benchmark.
$<\mu>=0$	9	zprime	$pp \to Z'_{SSM}$	Heavy resonance; standard candle for bump hunting.
$<\mu>=0$	10	hiddenvalley	$pp \to Z' \to$ dark	Dark QCD sector; semi-visible jets for anomaly detection.

Slide stolen from ACTS workshop

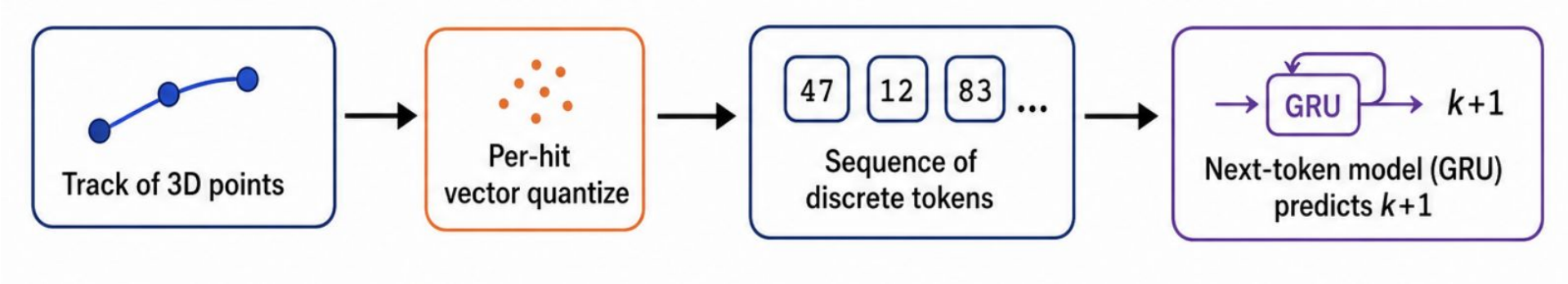
Dataset generation



ColliderML Arxiv

What is our problem statement?

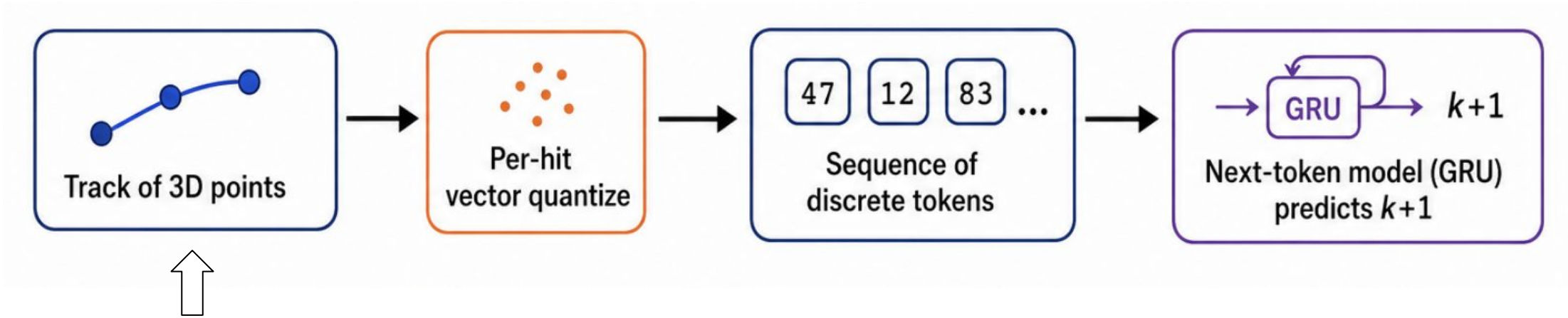
The first order problem that I started looking into was if I can tokenize the tracks and use the token prediction as the way of track extrapolation.



Some things to think about
1. Detector is big, ~1000*1000*3000
2. How much resolution do we want with tokenization?
3. IS predicting next token enough?

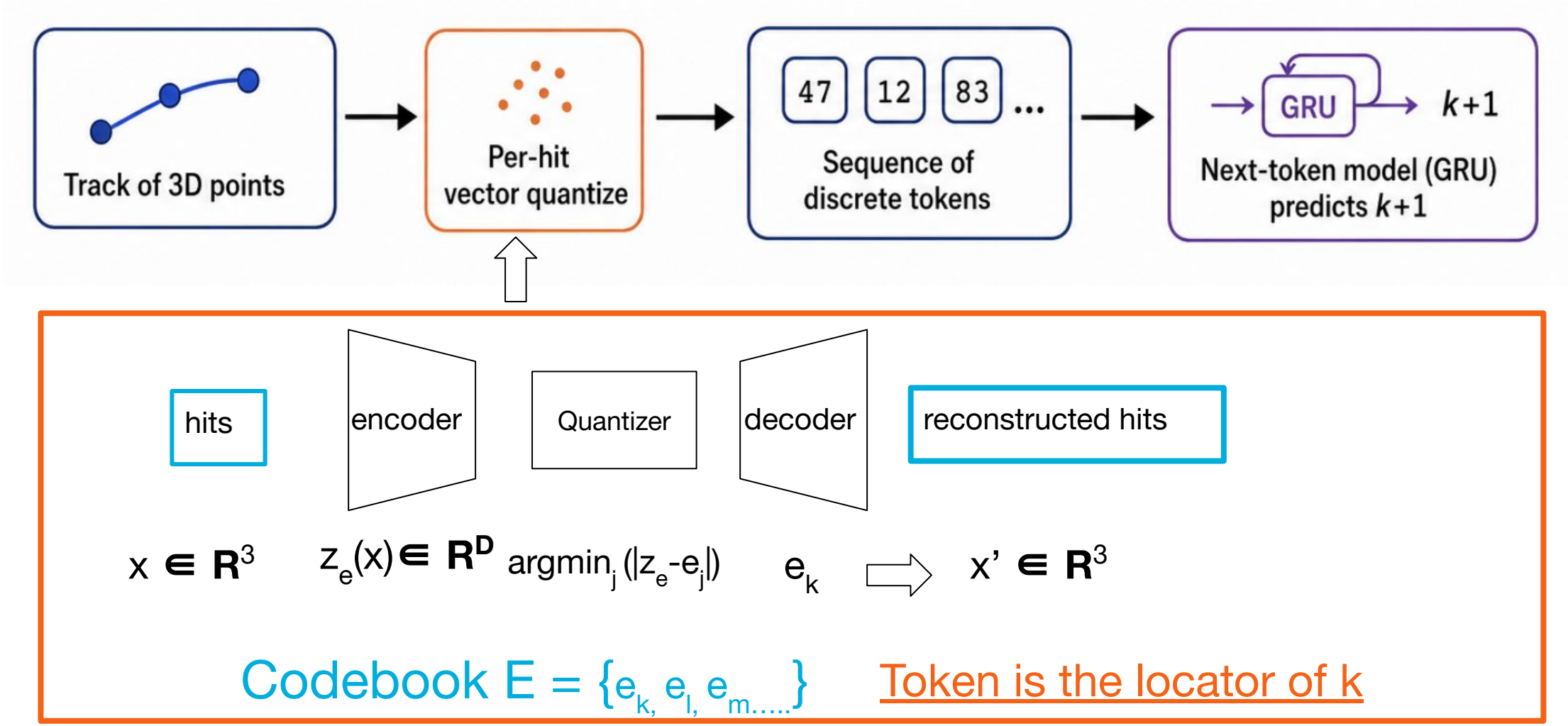
A question to ask,
Let's say I want 10mm resolution on Z prediction, what would be the codebook size?

Preprocessing

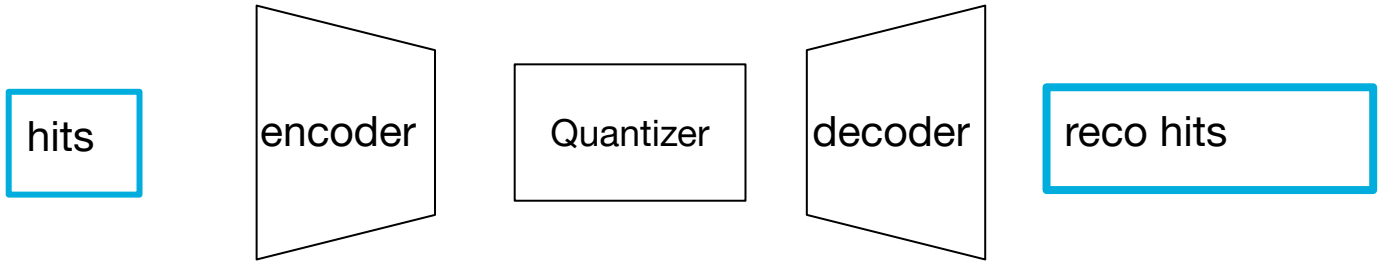


! Preprocessing of the dataset is needed.
Plot some tracks to see if they make sense.
Normalizations, splitting in test/train etc.

VQ-VAE and Tokenization



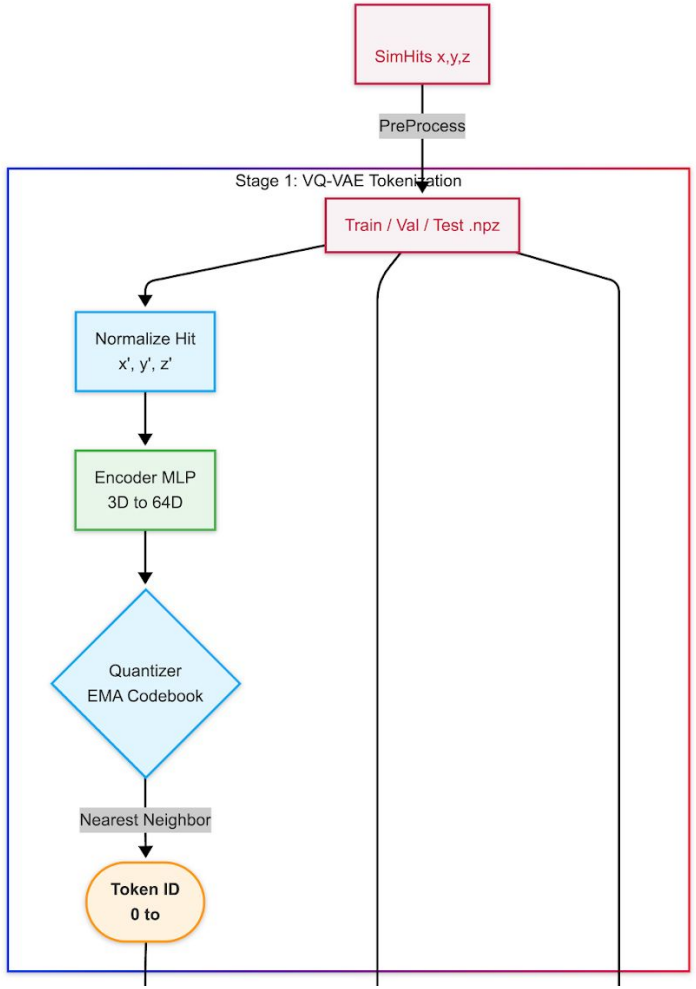
VQ-VAE and Tokenization



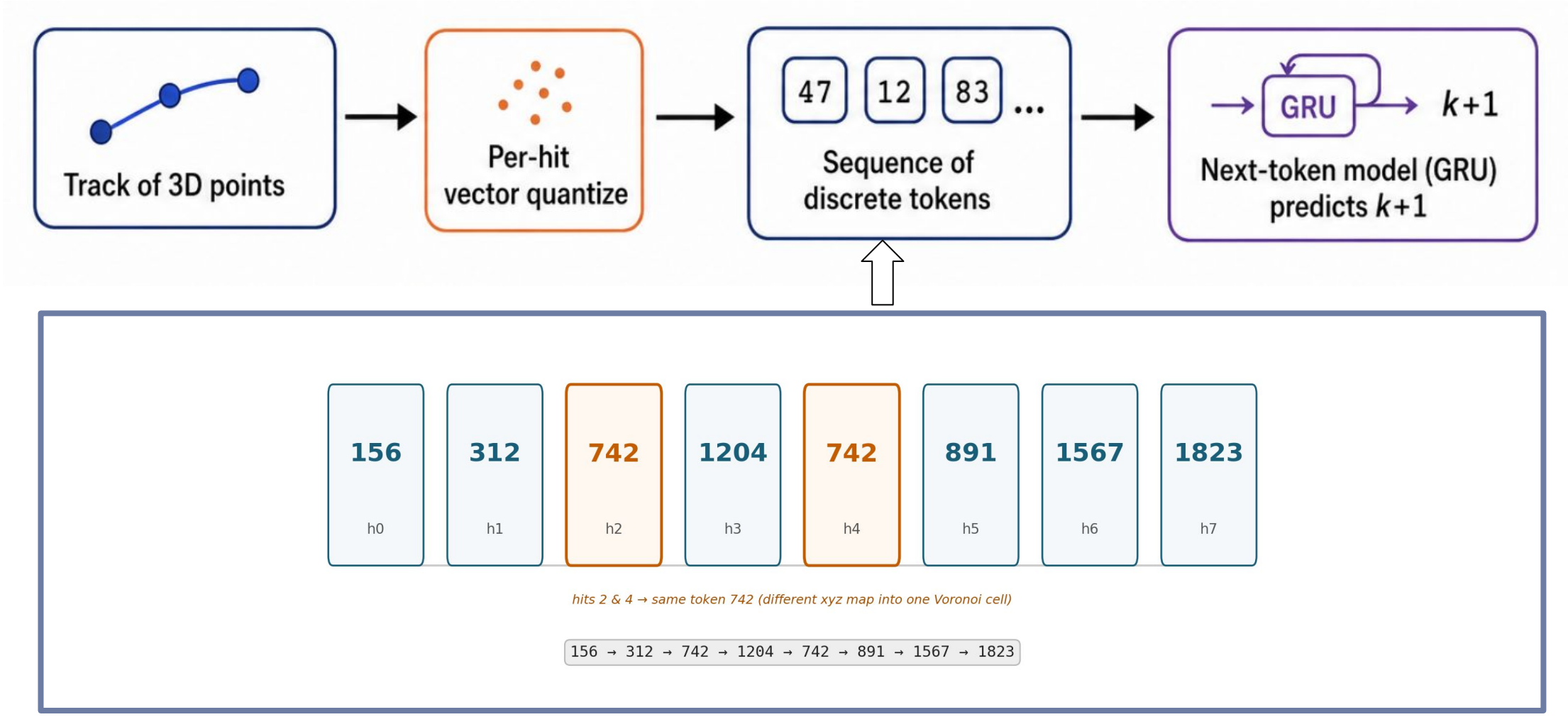
$x \in \mathbf{R}^3$ $z_e(x) \in \mathbf{R}^D$ $\text{argmin}_j (|z_e - e_j|)$ e_k ⇨ $x' \in \mathbf{R}^3$

Codebook is $\{e_k, e_l, e_m.....\}$

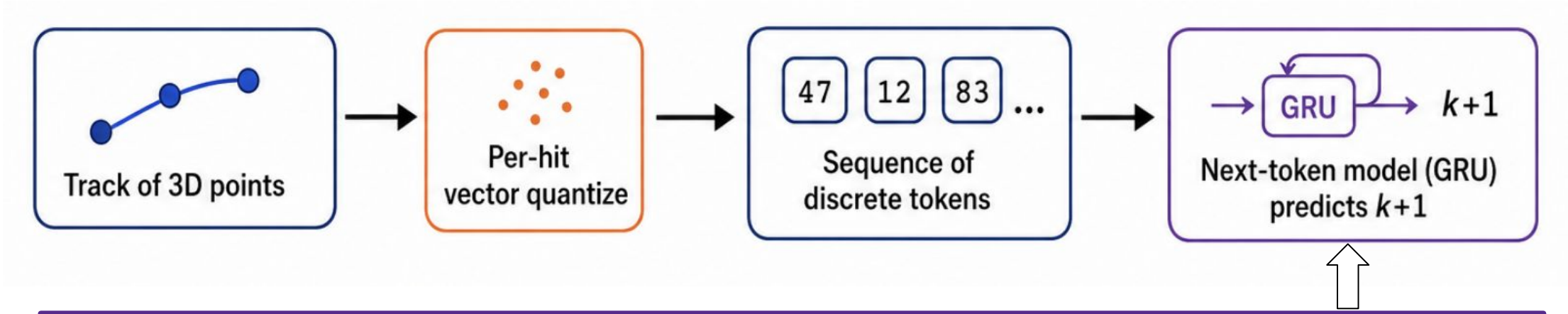
In my case
D= 64,
codebook size of 2048

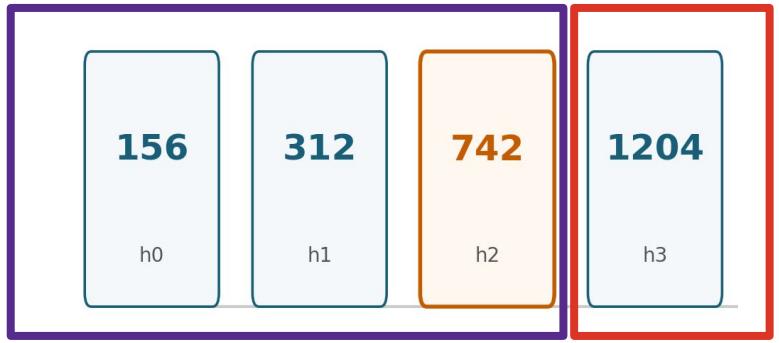


Tracks to Tokens



Prediction of Next Token (Ideal)

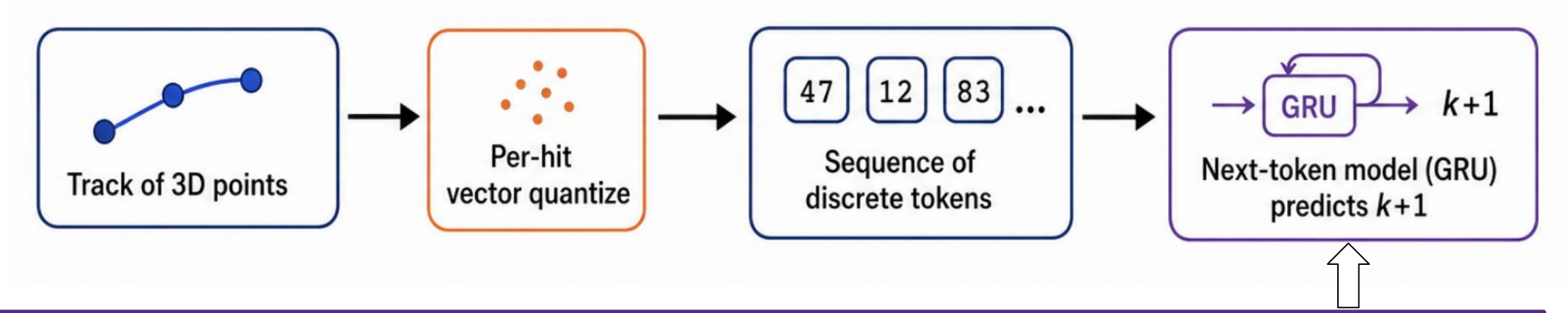




Take a sequence of 3 tokens and predict the next

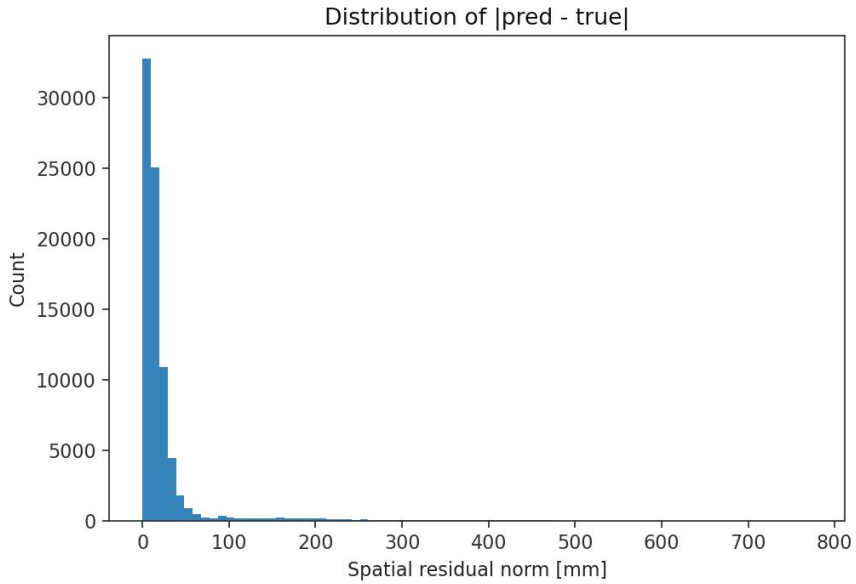
A rolling window used to create a forecasting like model

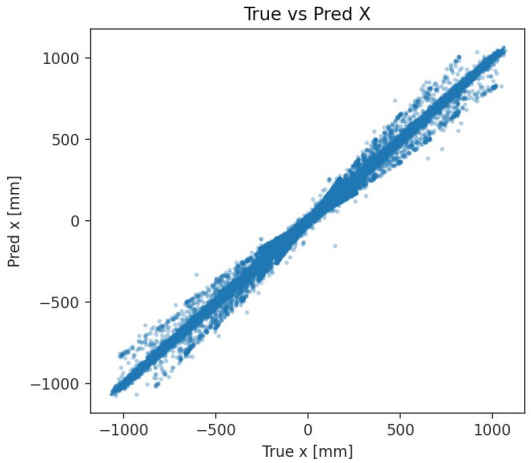
Is Prediction of Next Token enough?

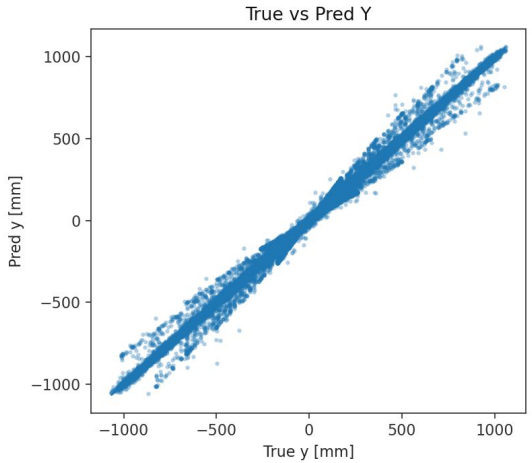


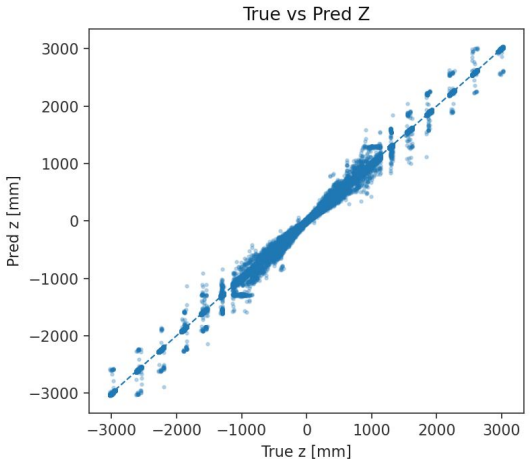
I found the answer to this question as no! Most likely the model + dataset I had isn't sufficient for this.
I had to use a concatenation of tokens with hits and layer information to predict the next token, hit and layer.

Efficiency





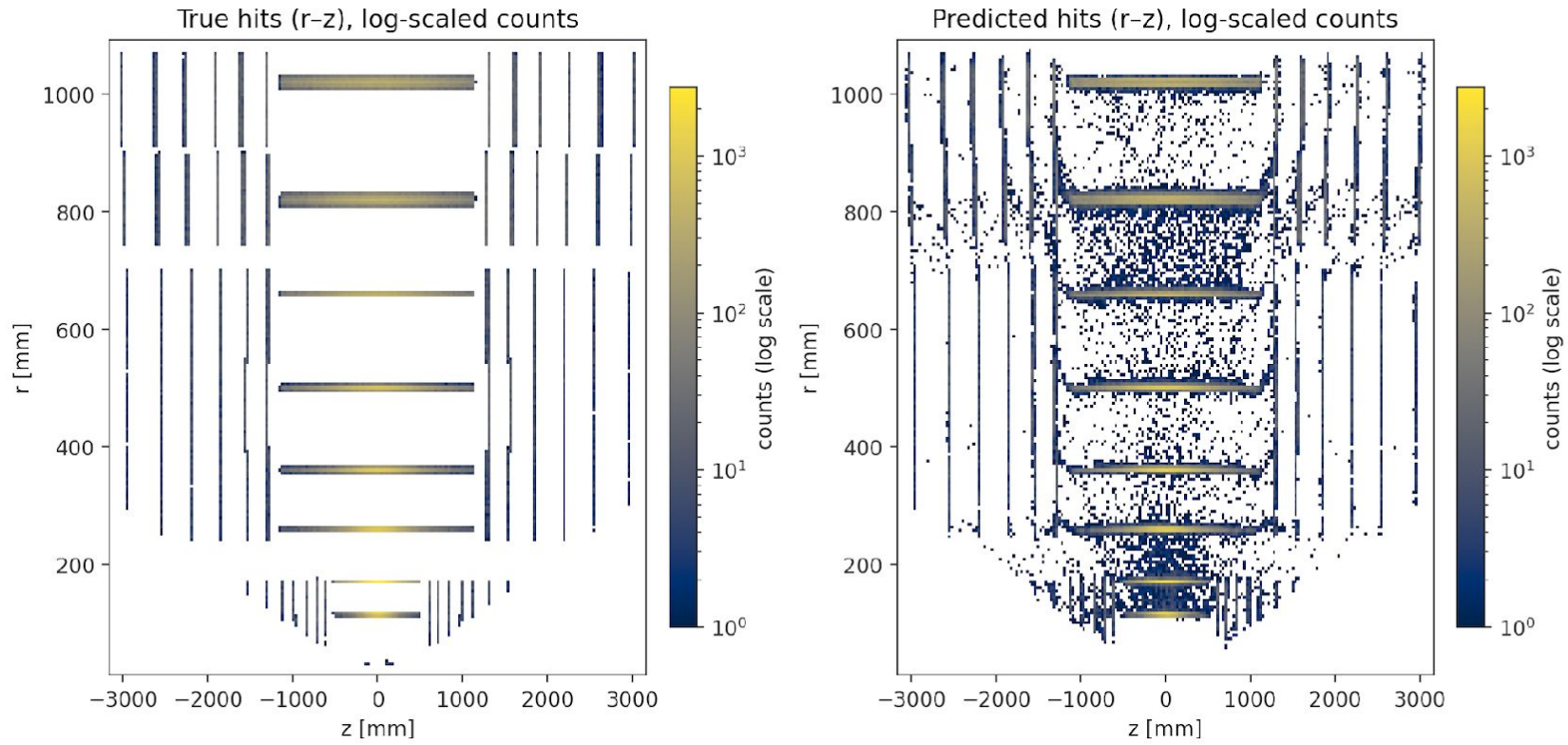




- 100mm uncertainty on predicted hit.

Brookhaven National Laboratory

Early Results



- 100mm uncertainty on predicted hit.

Brookhaven National Laboratory

Summary and Next steps

- It’s an interesting way of thinking the same problem.
- How would we add the ODD to the dataformat we discussed yesterday?
- Many use-cases, first one being using it with CKF?
- Are just having tokenization of the hits enough? Most likely not, we need track states on surface to be tokenized as well?
- Many other SOTA models can be used to fancy tracking tasks on the tokenized dataset.