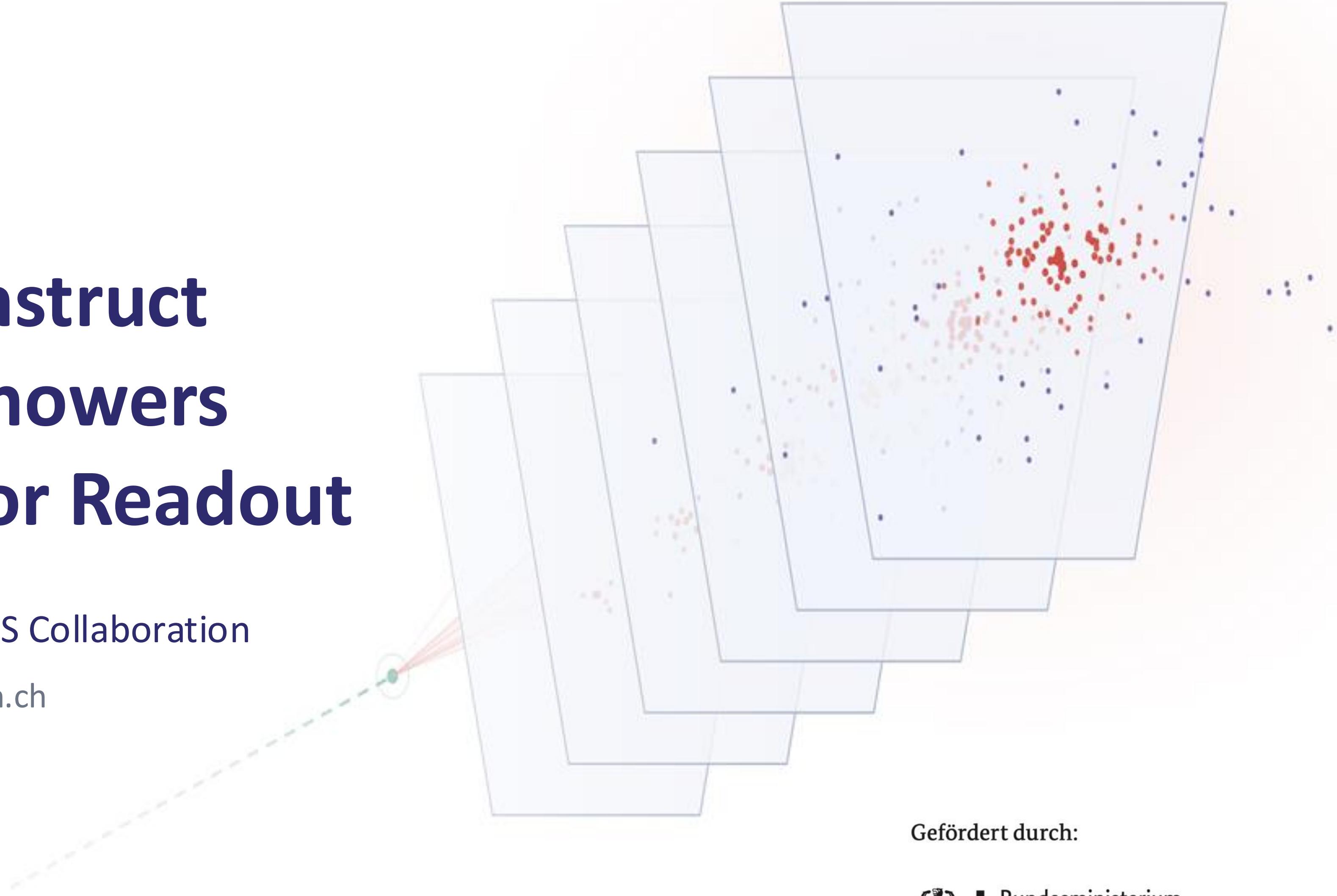
Learning to Reconstruct Muon Detector Showers from Raw Detector Readout

Asu Güvenli · on behalf of the CMS Collaboration

ML4Jets 2026 · ayse.asu.guvenli@cern.ch



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

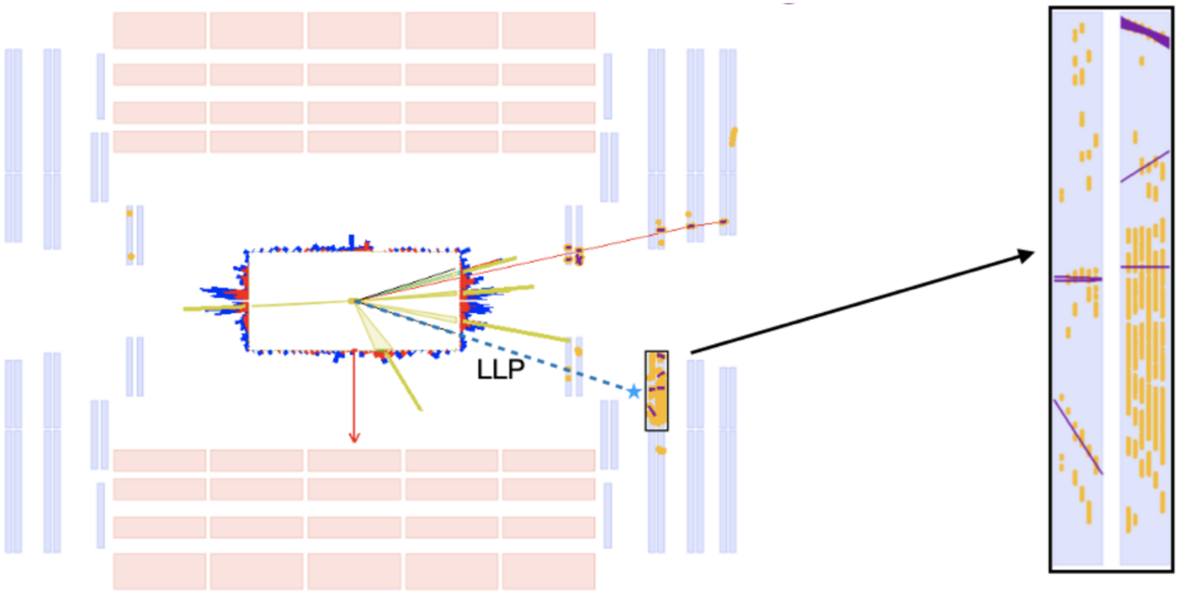
Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Muon detector showers

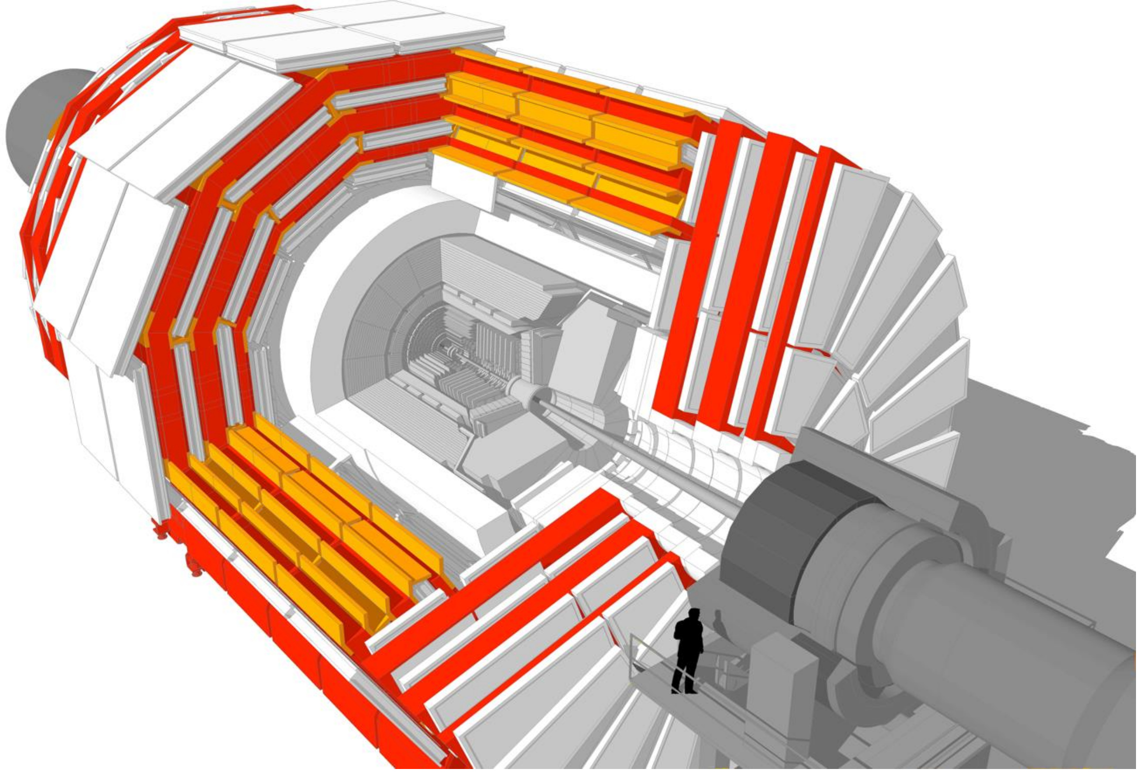
Long-lived particles decaying inside the endcap muon system

A **long-lived particle** (**LLP**) can decay in the steel of the CMS return yoke, and the shower is sampled by the gas layers of the **cathode strip chambers** (**CSC**s).

The **reconstructed hits** (**RecHits**) are clustered into a **muon detector shower** (**MDS**): at least 50 hits per cluster.



An LLP shower in the endcap muon detectors, next to a prompt muon signature



The CMS muon system, with the chambers interleaved with the iron return yoke [1]

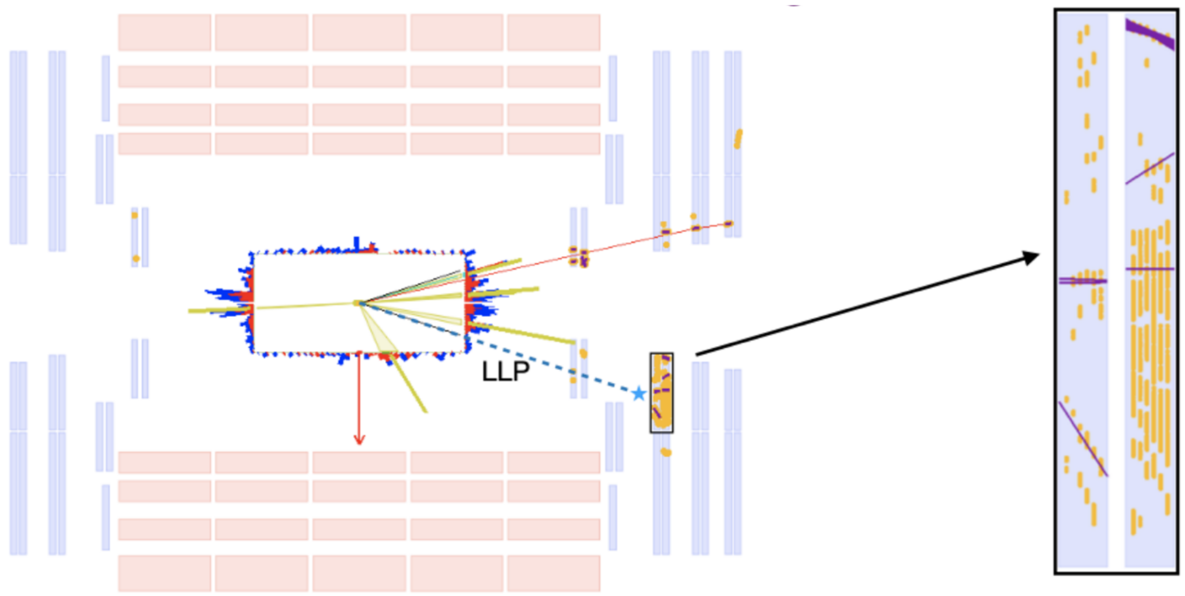
[1] CMS Open Data Workshop lessons, CMS DPOA group

Muon detector showers

Long-lived particles decaying inside the endcap muon system

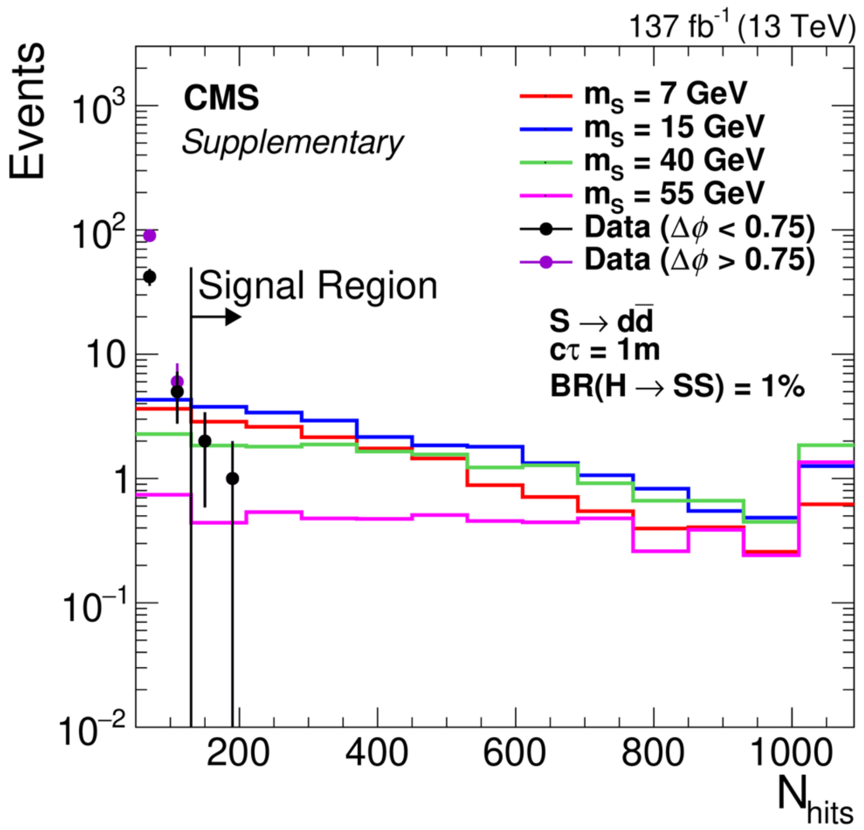
A **long-lived particle** (**LLP**) can decay in the steel of the CMS return yoke, and the shower is sampled by the gas layers of the **cathode strip chambers** (**CSC**s).

The **reconstructed hits** (**RecHits**) are clustered into a **muon detector shower** (**MDS**): at least 50 hits per cluster.



An LLP shower in the endcap muon detectors, next to a prompt muon signature

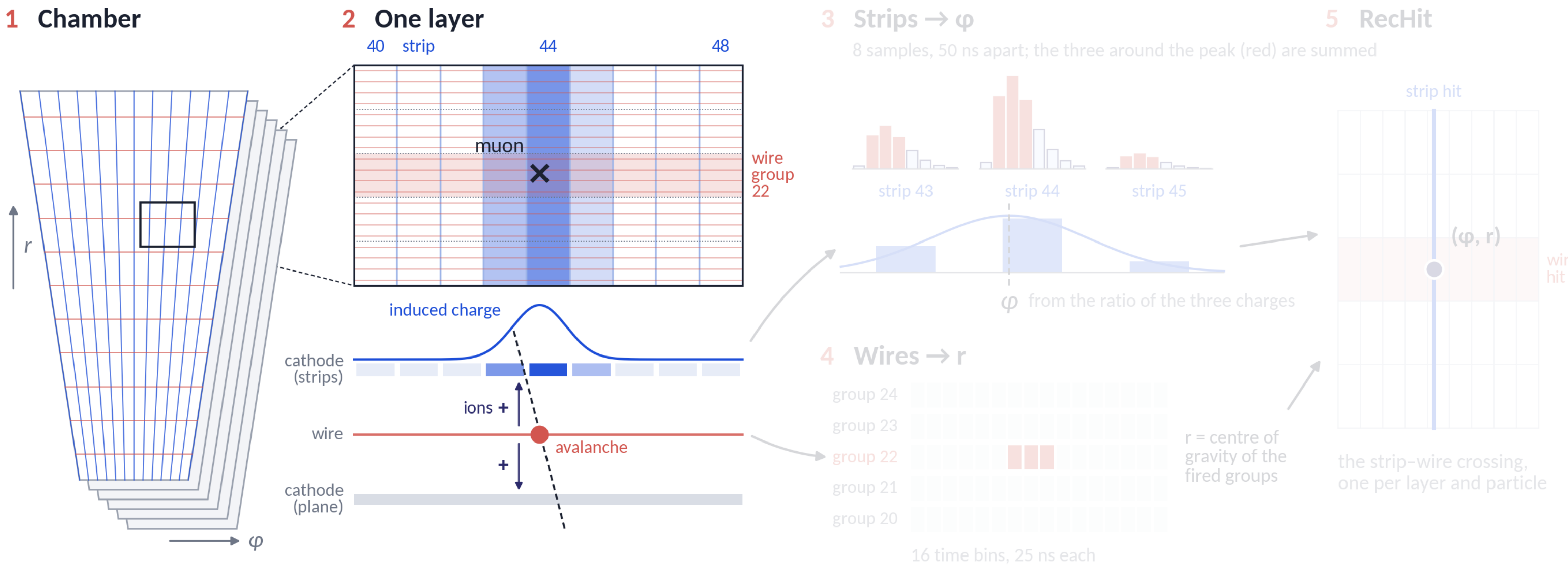
Of the cluster variables the search selects on, the **hit multiplicity N_{hits}** is the strongest [2]. It is counted in dense showers, a regime the CSC reconstruction was not designed for.



N_{hits} in the cluster, signal and data, EXO-20-015 [2]

How a CSC hit is made

Cathode strips measure ϕ, anode wire groups measure r [3,4]

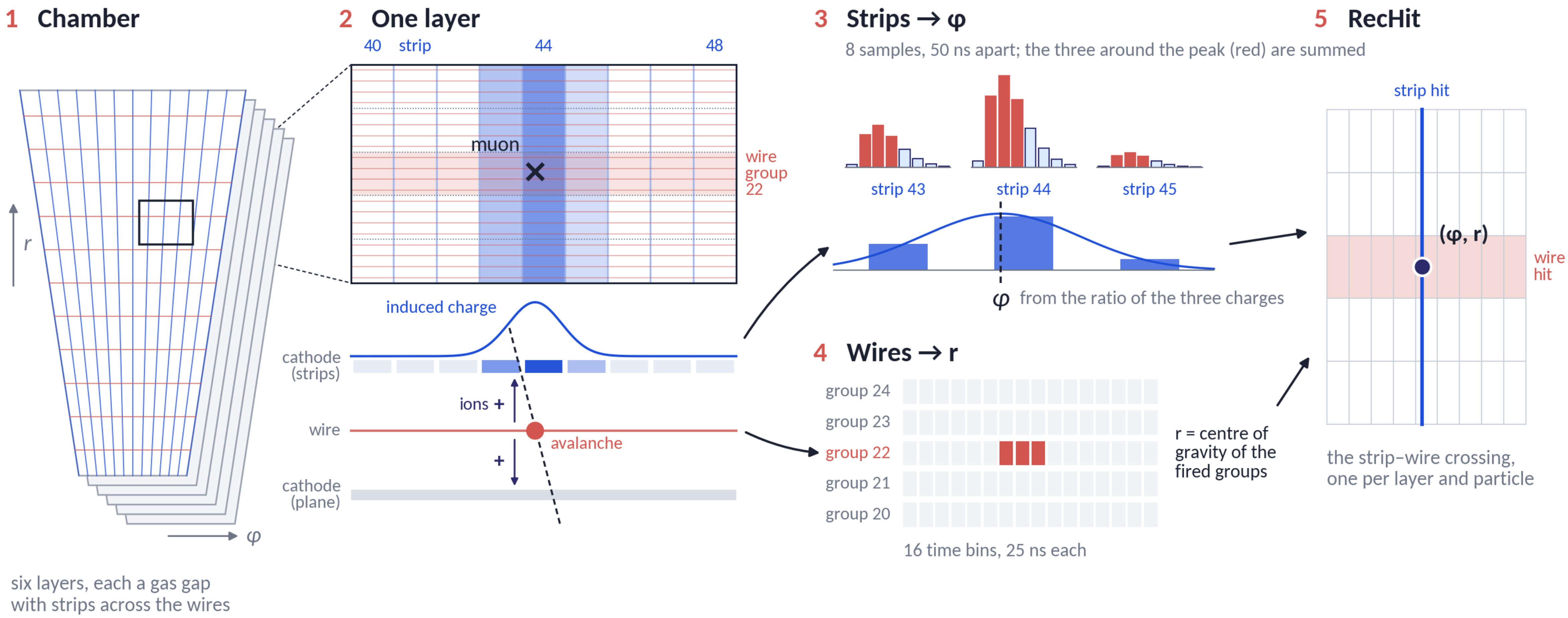


six layers, each a gas gap with strips across the wires

Each strip cluster and each wire-group cluster is taken as one particle; strip–wire pairs become the RecHit candidates.

How a CSC hit is made

Cathode strips measure ϕ, anode wire groups measure r [3,4]



Each strip cluster and each wire-group cluster is taken as one particle; strip–wire pairs become the RecHit candidates.

Ghost hits

A hit is a strip–wire crossing: N_s fired strips and N_w fired wire groups give $N_S \times N_W$ candidates

A

One particle

one fired strip, one fired wire

B

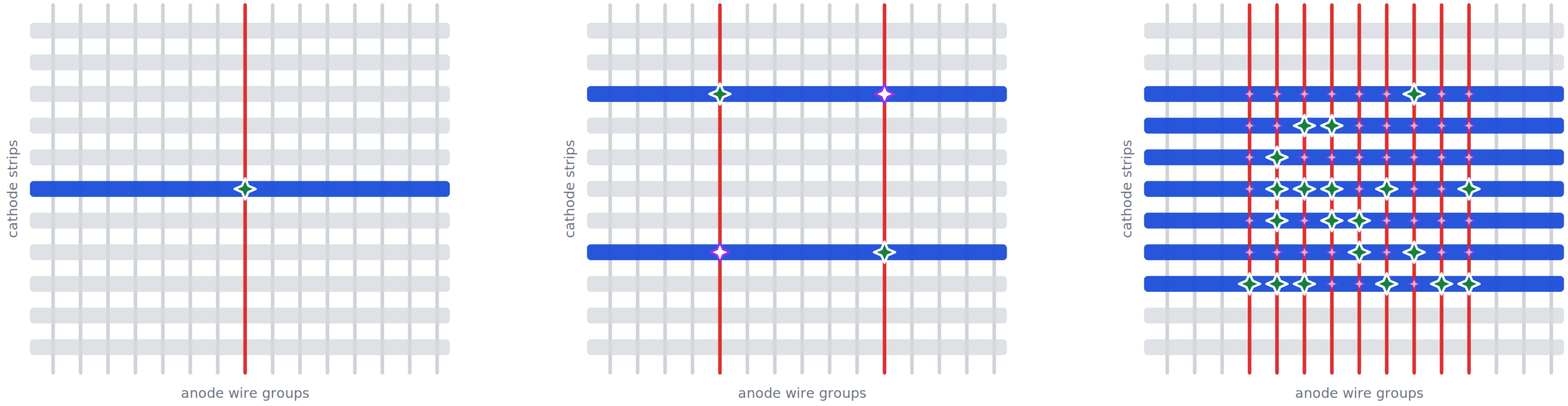
Two particles

the pairing is already ambiguous

C

A shower

hundreds of channels fire at once



1 × 1 = 1 candidate, 1 is real

2 × 2 = 4 candidates, 2 are real

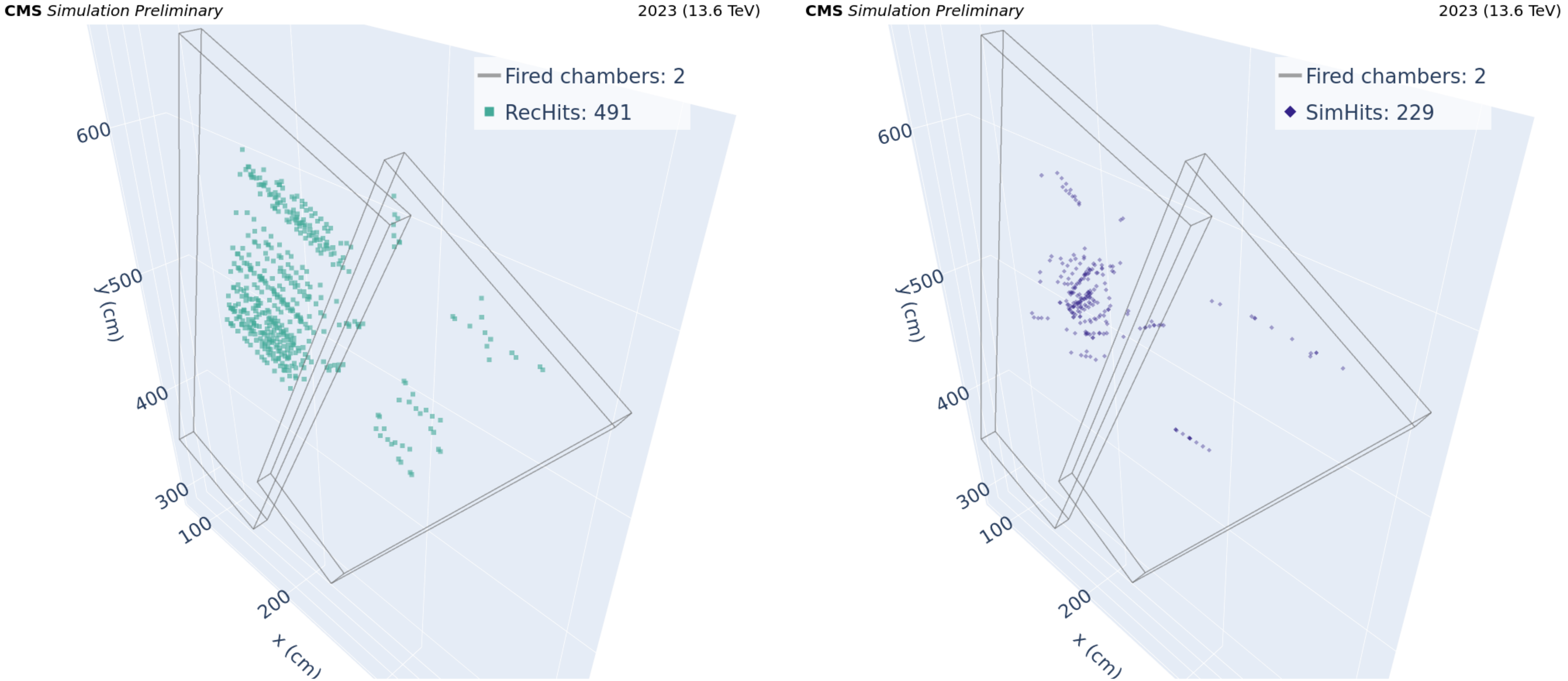
7 × 9 = 63 candidates, 20 are real

fired cathode strip · fired anode wire group · true hit · ghost candidate

Charge matching removes part of them.

Ghost hits

A hit is a strip–wire crossing: N_s fired strips and N_w fired wire groups give $N_S \times N_W$ candidates



A simulated LLP shower in two chambers: 491 RecHits for 229 Geant4 SimHits. The excess comes from strip–wire crossings that pair signals of different particles.

Inputs and target

One token per fired channel, carrying its full time series

Strip tokens

8 ADC (Analog-to-Digital Converter) samples × 50 ns

the full pulse shape of charge deposit

Wire-group tokens

16 occupancy bins × 25 ns

the timing measurement

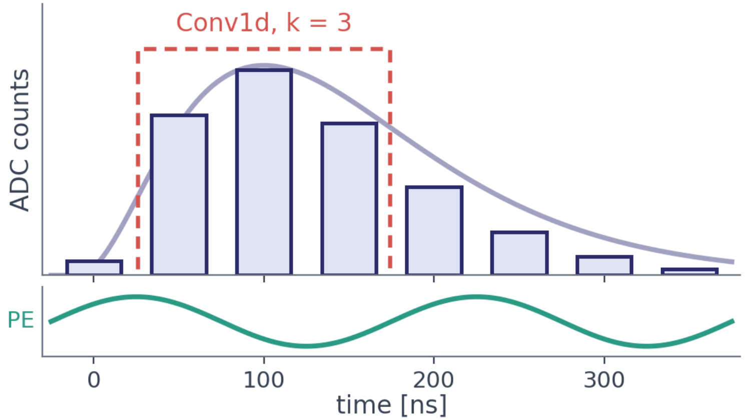
Geometry, per token

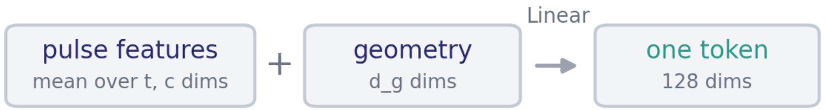
begin and end in (x, y, z) coordinates

plus the DetID (which chamber and layer) and the strip or wire-group number

Target - SimHits

Geant4 SimHits with position, timing, and deposited energy

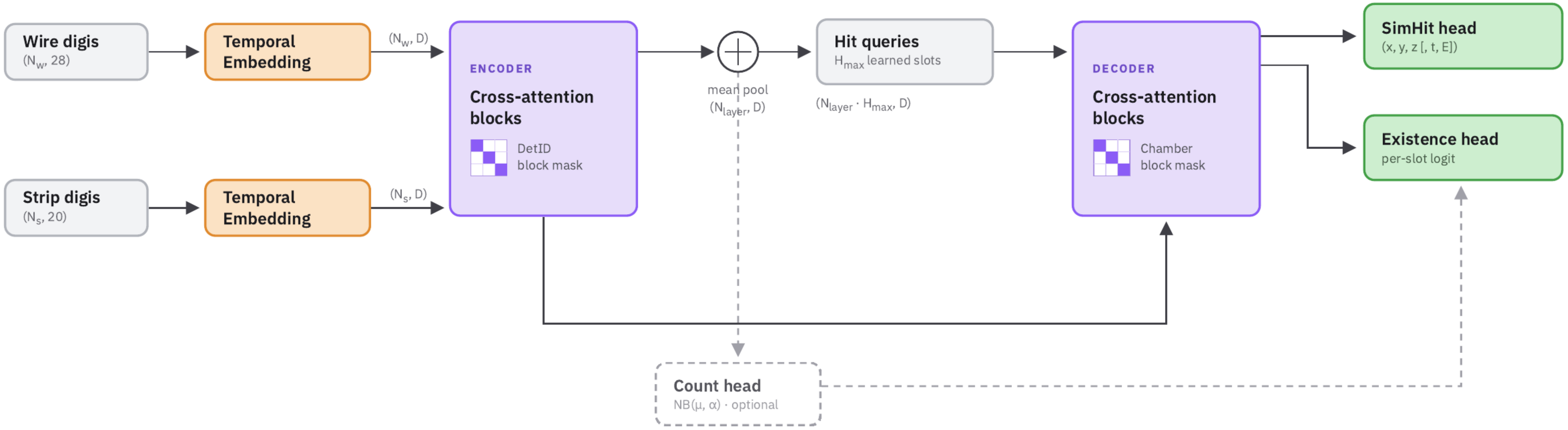




Both token types carry a time series. Each bin is lifted to c channels, given a sinusoidal encoding in ns, passed through a small Conv1d stack along the time axis and mean-pooled, then merged with the geometry projection.

Architecture, block form

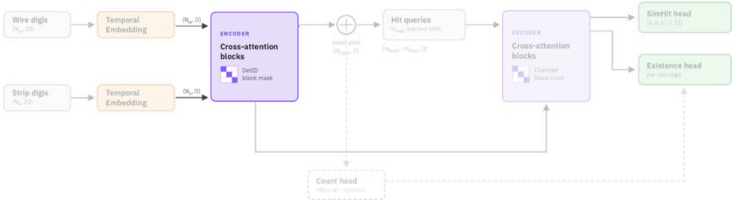
Geometry-aware cross-attention transformer [5,6], 1.34 M parameters



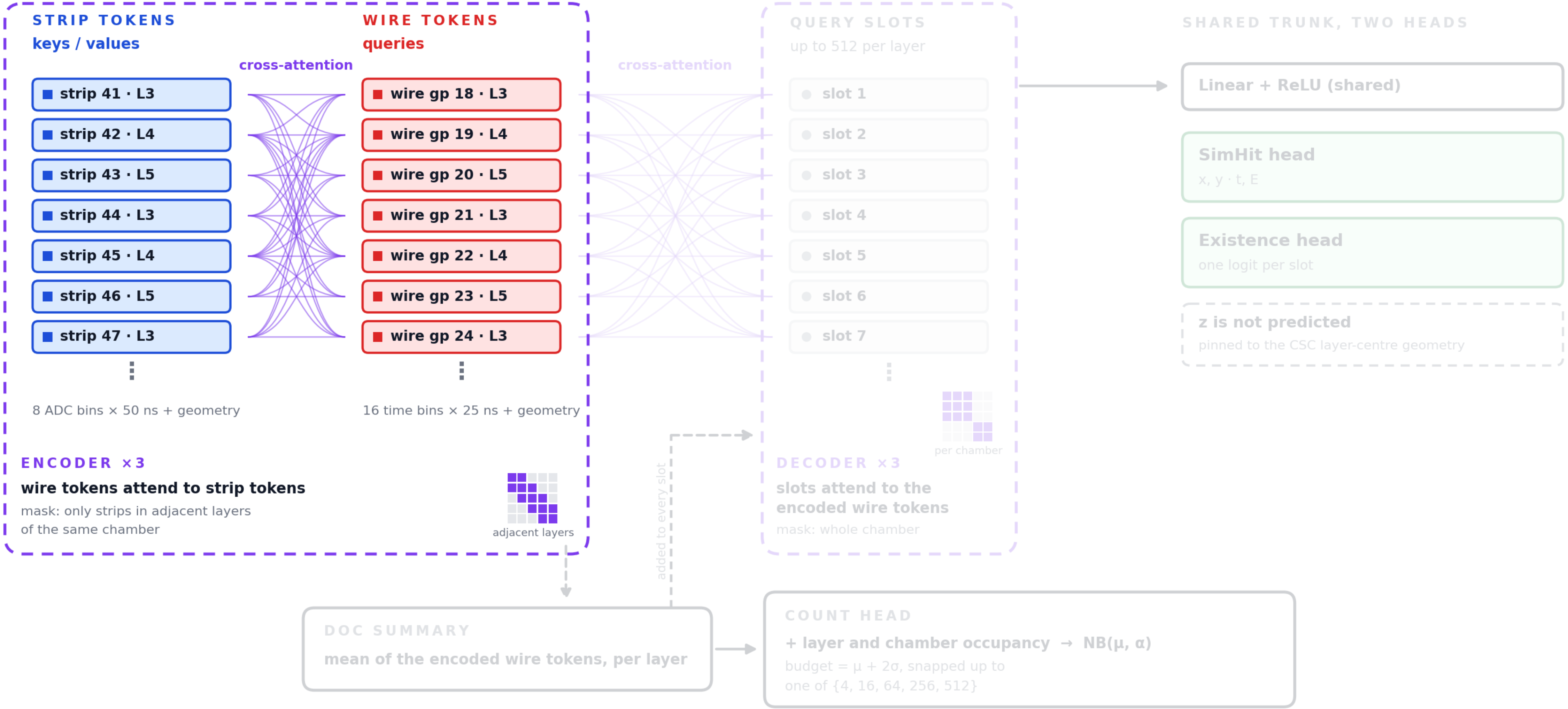
Two token streams, one encoder mixing them inside a readout neighbourhood, one decoder of query slots per layer, and three heads: existence, SimHit and a per-layer count. The attention masks are block-sparse, implemented with FlexAttention [7].
The next three slides walk through the same diagram one stage at a time.

[5] Attention is all you need · [6] DETR · [7] FlexAttention

Encoder: wires and strips

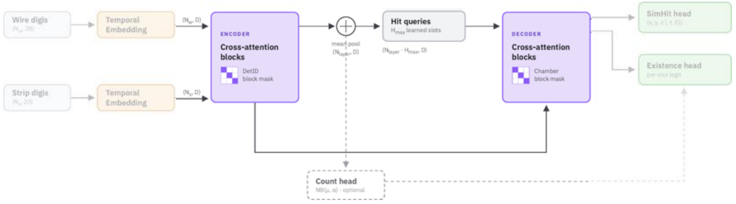


Wire tokens are the queries, strip tokens the keys and values

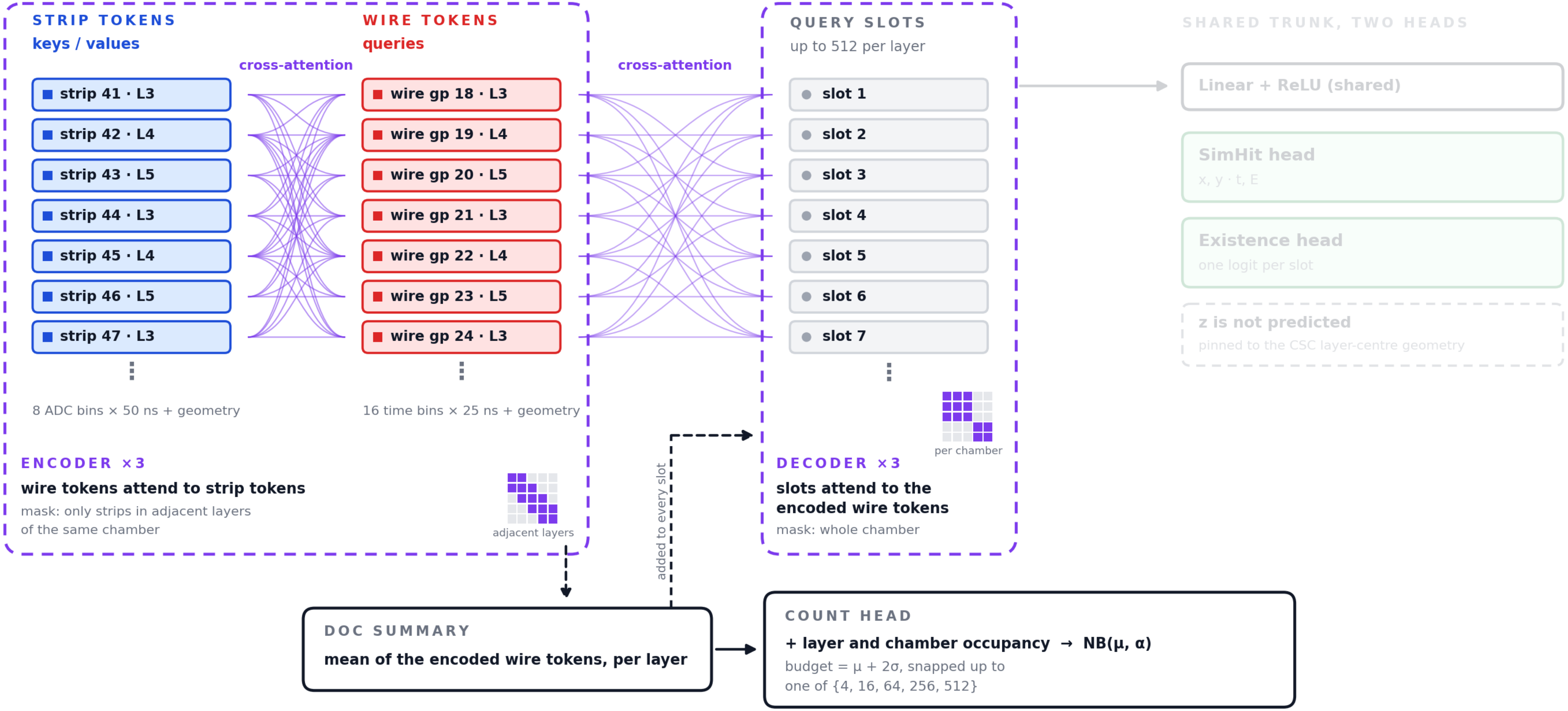


1.34 M parameters · embedding 128 · 3 encoder + 3 decoder blocks · 8 heads

Decoder: hit queries

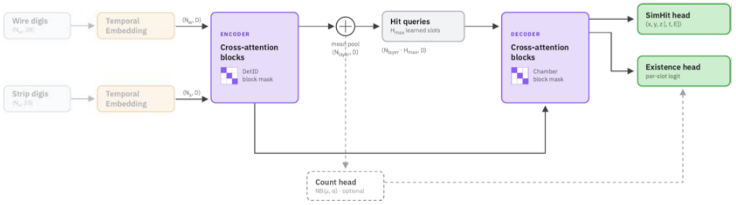


A learned embedding per slot, plus a summary of the layer, attending to the encoded wire tokens

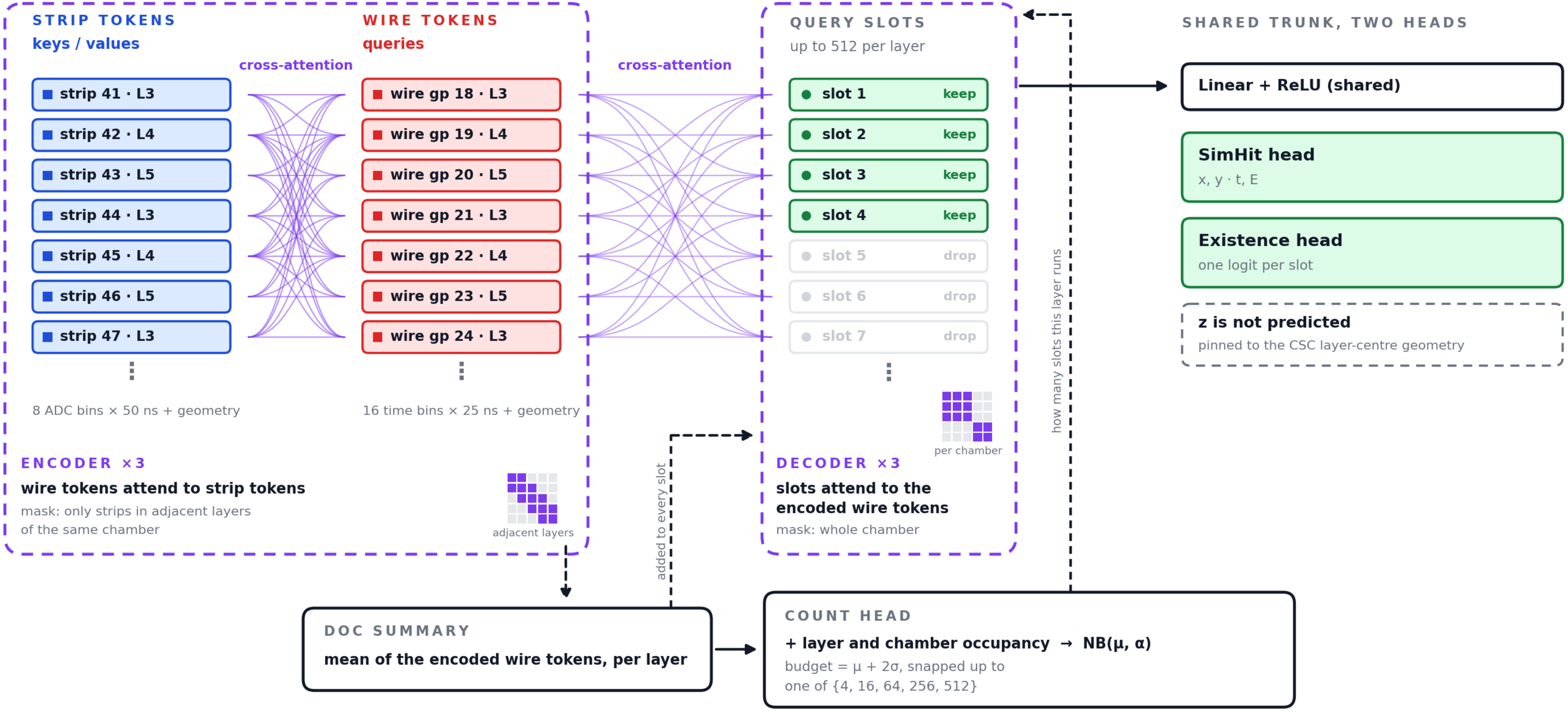


1.34 M parameters · embedding 128 · 3 encoder + 3 decoder blocks · 8 heads

Heads, and counting the hits



A Negative-Binomial count head sets how many slots each layer actually runs



1.34 M parameters · embedding 128 · 3 encoder + 3 decoder blocks · 8 heads

Samples

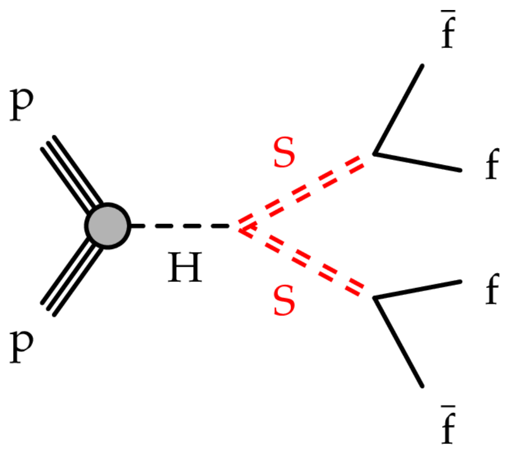
2023 detector conditions, 13.6 TeV, powheg + pythia8, Geant4 [8]

Signal Twin-Higgs H → SS → 4b [9], with $m_{\mathrm{H}} = [125, 1000]$ GeV, $m_{\mathrm{S}} = [15, 50]$ GeV, $c\tau_{\mathrm{S}} = 1$ m, decaying inside the CSCs.

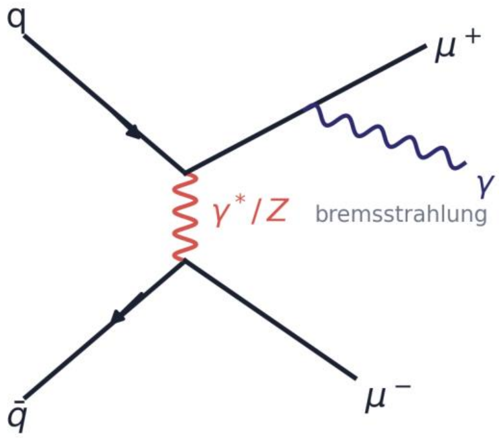
Background-like Drell-Yan → 2μ, prompt-muon bremsstrahlung showers, $\mathrm{M}_{\mu\mu}$ from 50 GeV to above 6 TeV.

Training set DBSCAN clusters [10] on MDS RecHits with at least 50 hits: ≈260k training, ≈32k validation, ≈32k test, about 62 % signal clusters.

Prompt-muon showers are in the training mixture, so one model covers the dense and the sparse regime.



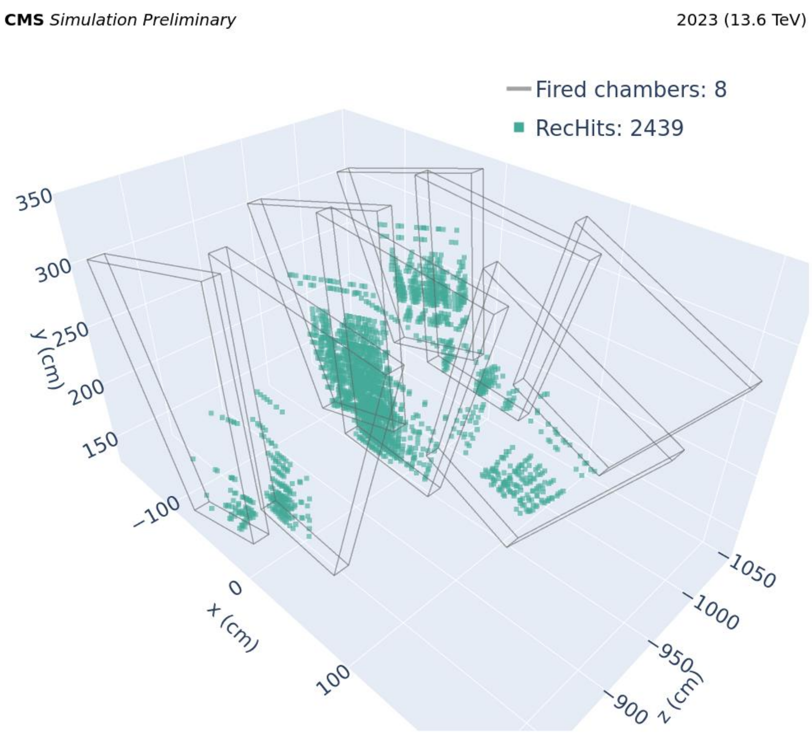
signal, EXO-25-005 [9]



background-like

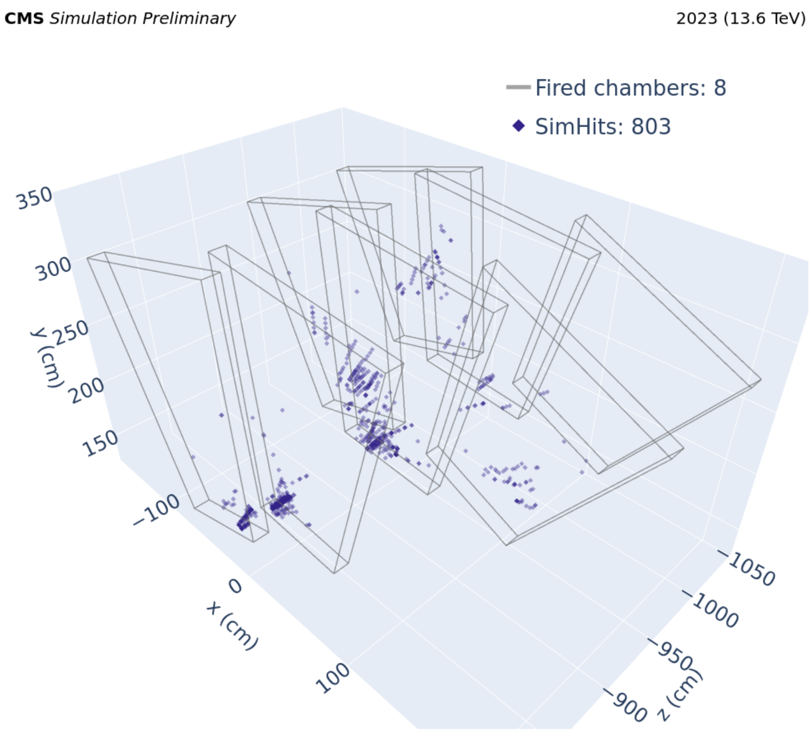
Event display: a dense LLP shower

The same cluster across 8 fired chambers



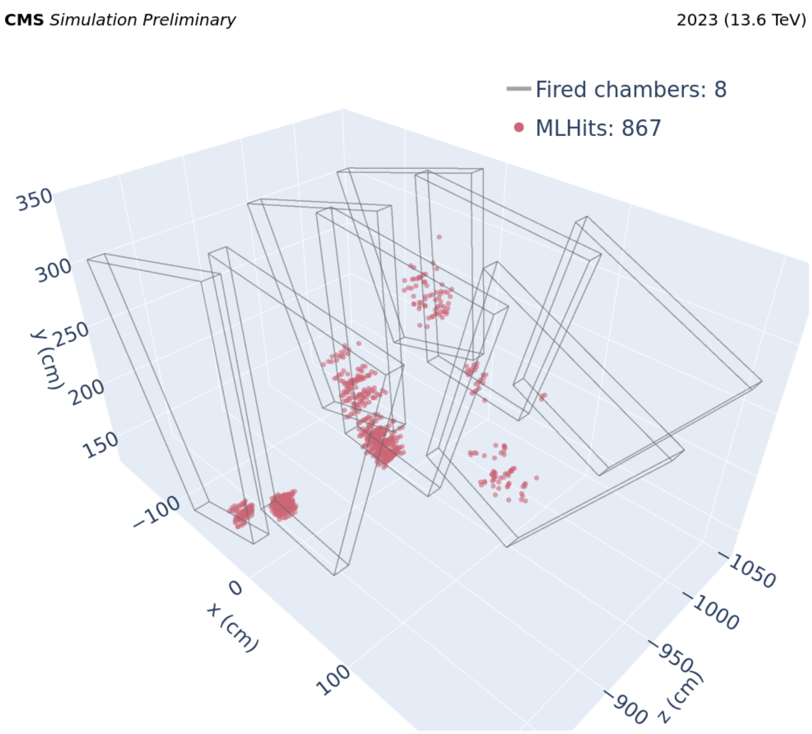
RecHit

2439



SimHit

803

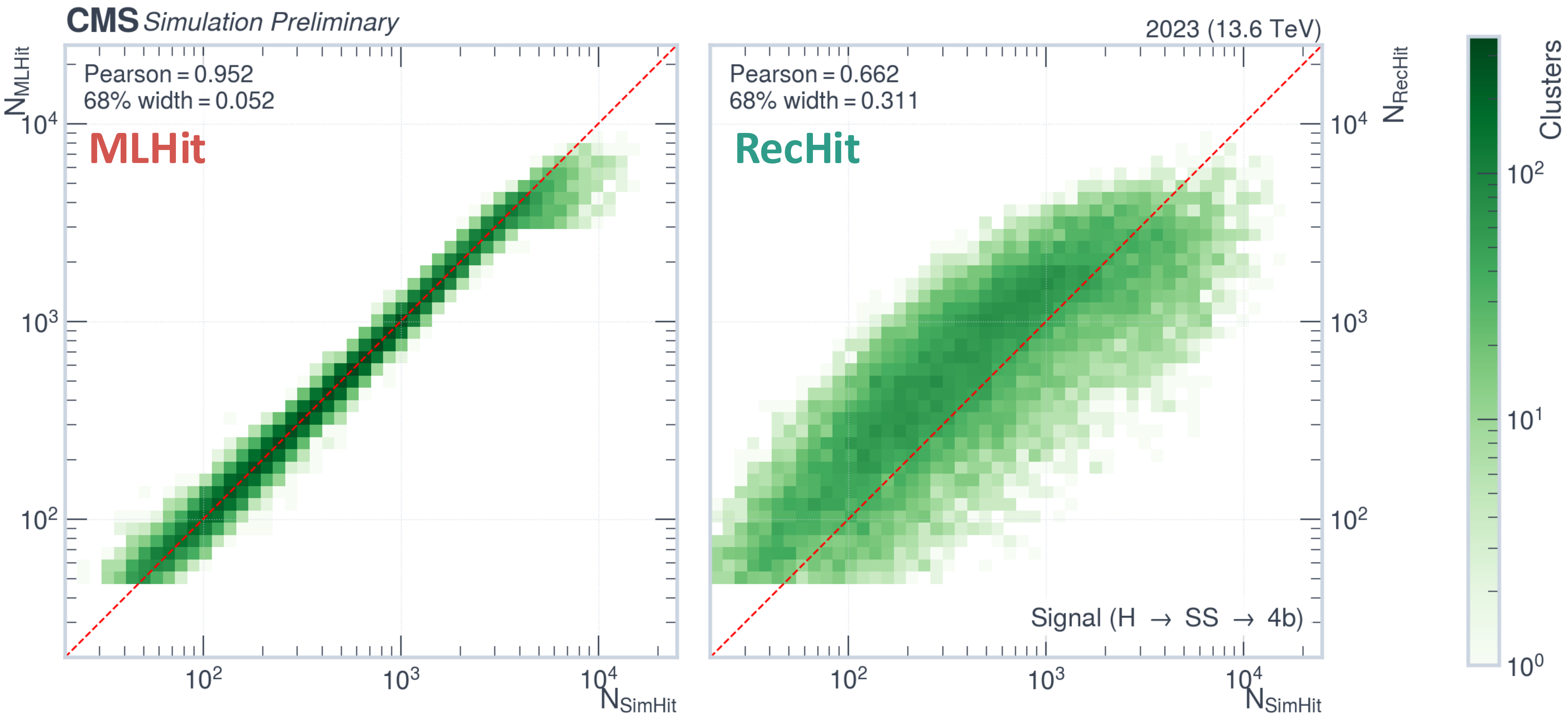


MLHit

867

Hit multiplicity

Reconstructed against true cluster multiplicity, LLP signal



0.95

$R_{Pearson}$ MLHit

0.66

$R_{Pearson}$ RecHit

68 % width

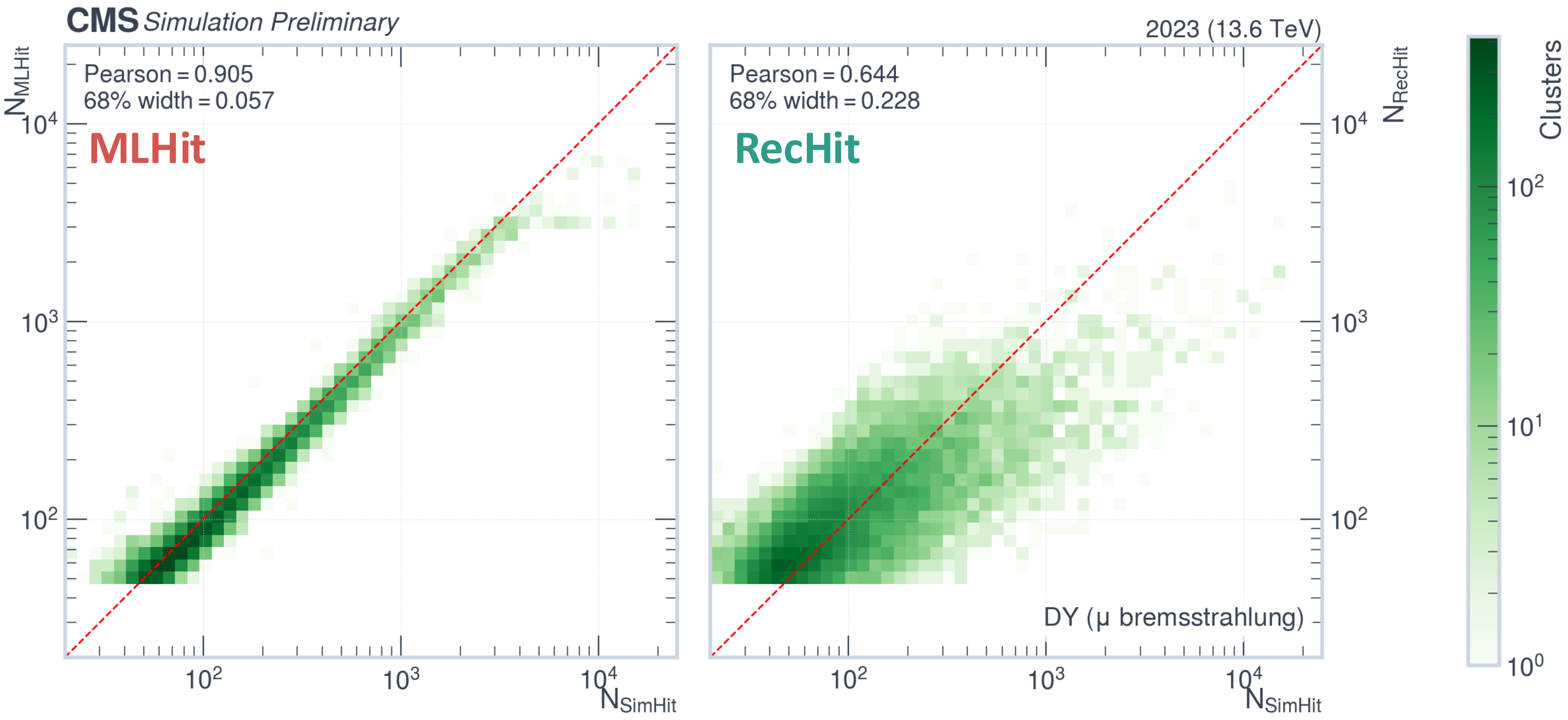
0.05 against 0.31

The RecHit count is biased and broad. The MLHit count follows the diagonal and turns over only where the 512-slot per-layer budget runs out.

*Half-width of the central 68% interval (16th to 84th percentile) of $\log_{10}(N_{Hit}/N_{SimHit})$

Prompt-muon showers

The same model, no retuning: Drell-Yan muon bremsstrahlung



0.91

$R_{Pearson}$ MLHit

0.64

$R_{Pearson}$ RecHit

68 % width

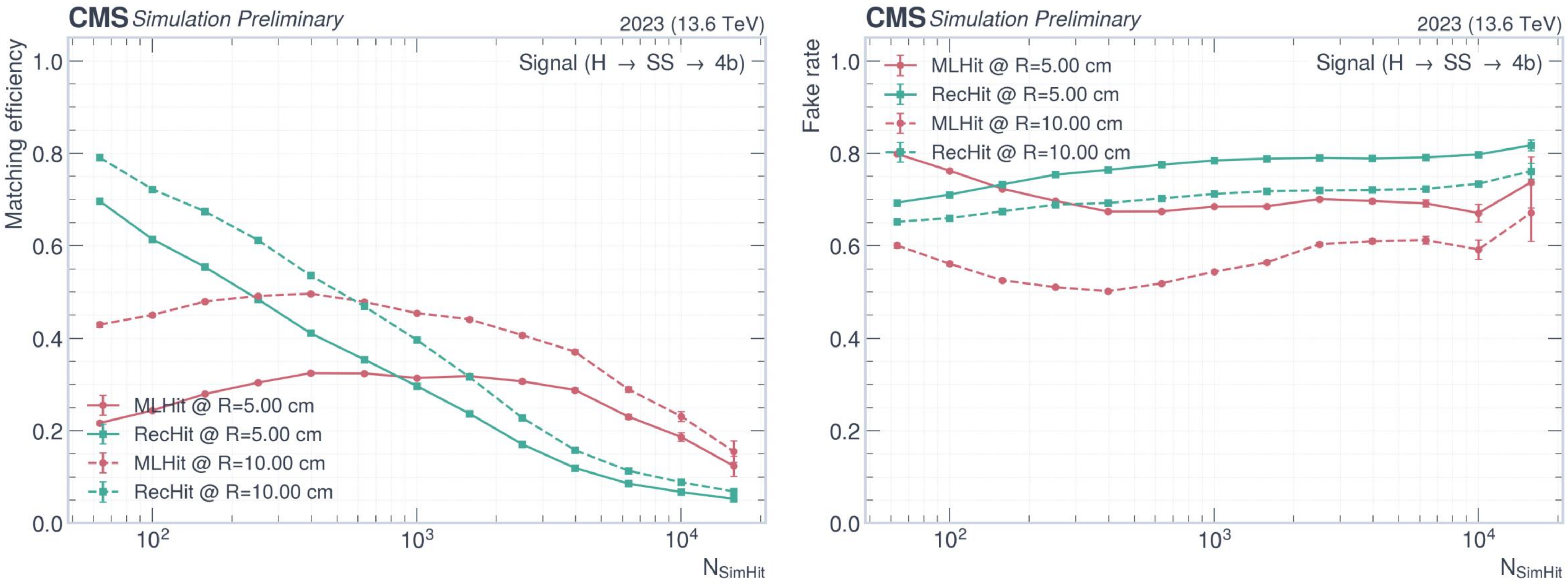
0.06 against 0.23

These clusters come from muon bremsstrahlung rather than an LLP decay, and they are part of the training mixture.

*Half-width of the central 68% interval (16th to 84th percentile) of $\log_{10}(N_{Hit}/N_{SimHit})$

Efficiency and fake rate

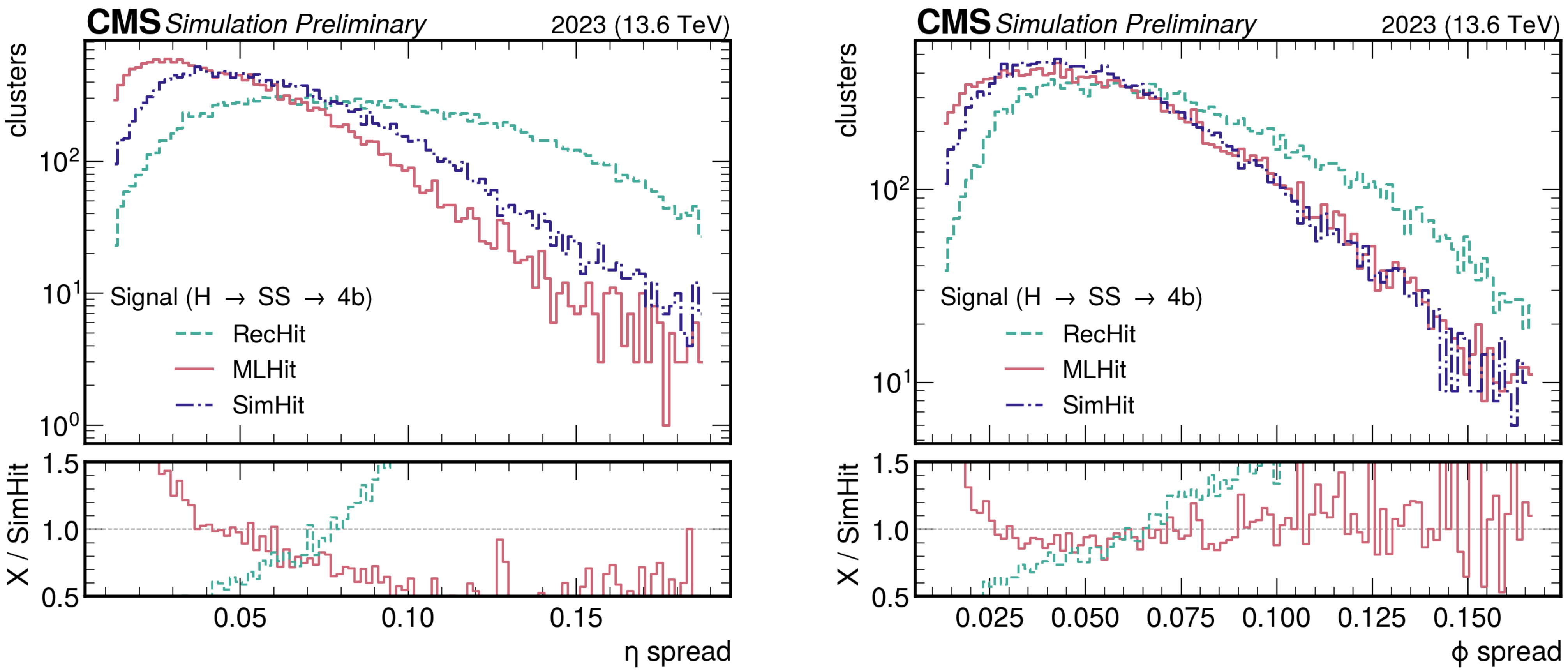
Hit matching against true cluster multiplicity, LLP signal



RecHits are better below N ≈ 800. Above that there are fewer RecHits per layer than SimHits, so both the efficiency and the fake rate favour the MLHits.

Cluster shape

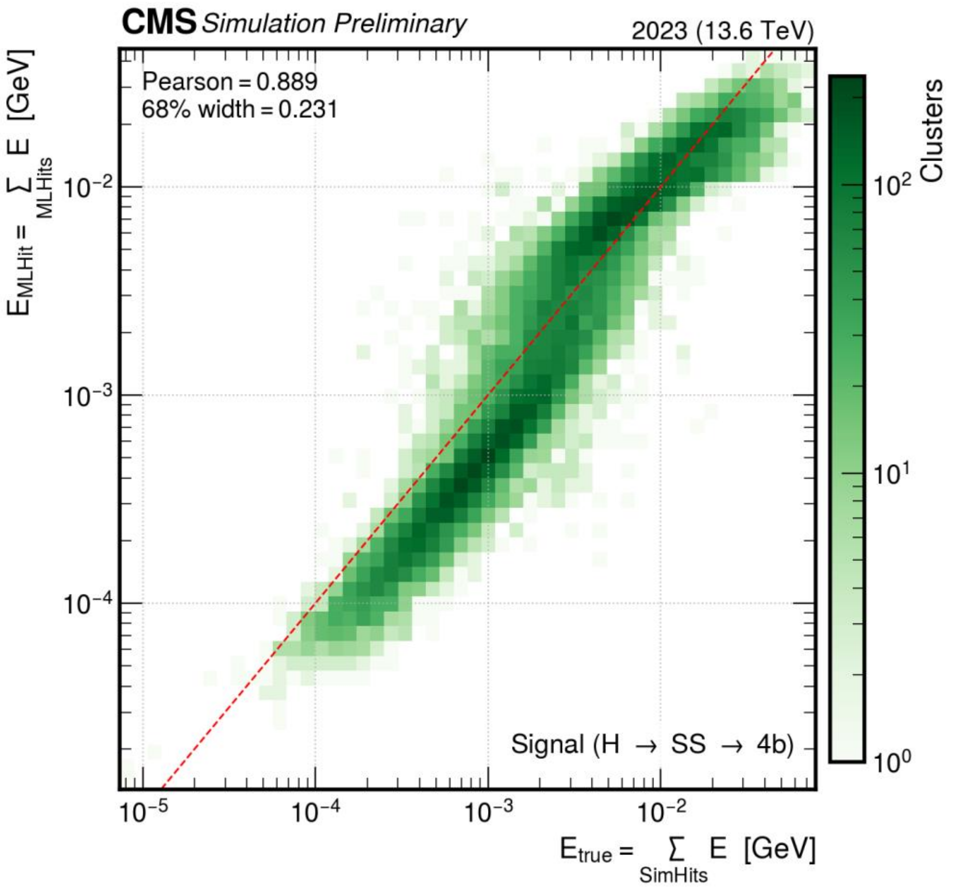
η and ϕ spread, rebuilt from each hit collection, LLP signal



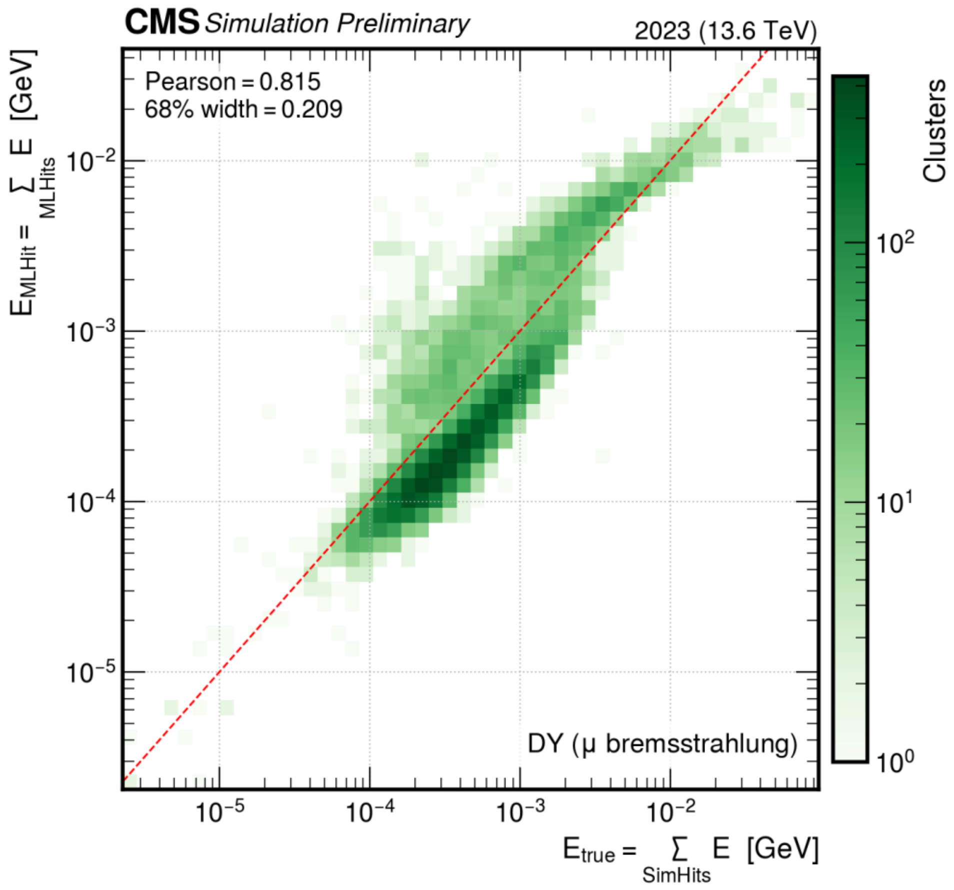
Both spreads enter the search selection. The MLHit ϕ spread sits on the SimHits and the η spread follows it closely; the RecHit distributions are broader in both.

Deposited energy

Summed cluster energy, from the SimHit head; no RecHit baseline exists



LLP signal



Prompt-muon showers

Correlation with the true summed deposited energy. The RecHits carry no calibrated energy, so this is an observable the search does not have today.

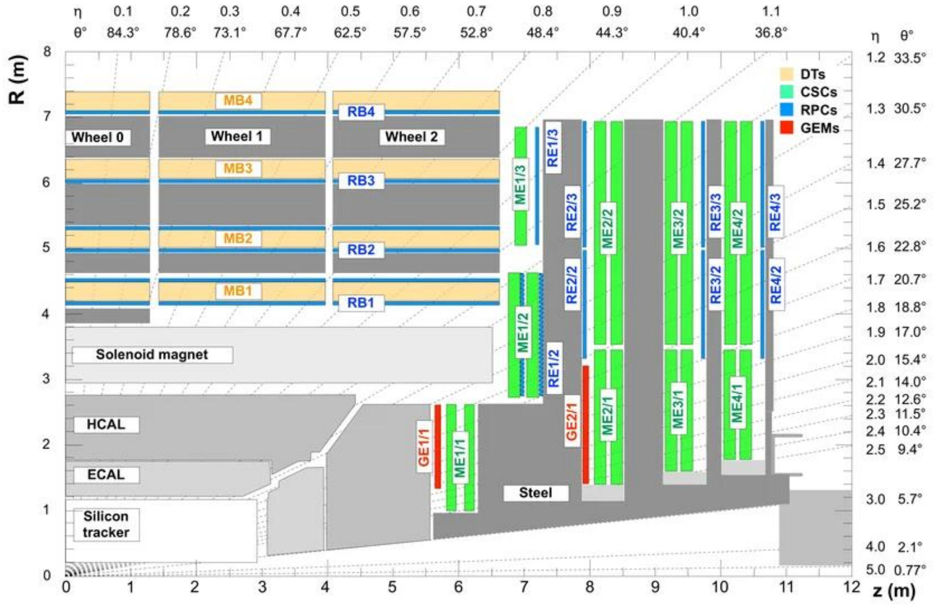
*Half-width of the central 68% interval (16th to 84th percentile) of $\log_{10}(N_{Hit}/N_{SimHit})$

Next steps

All in progress

Into the barrel

Extension to the drift tubes (DTs), for MDS coverage beyond the endcaps.



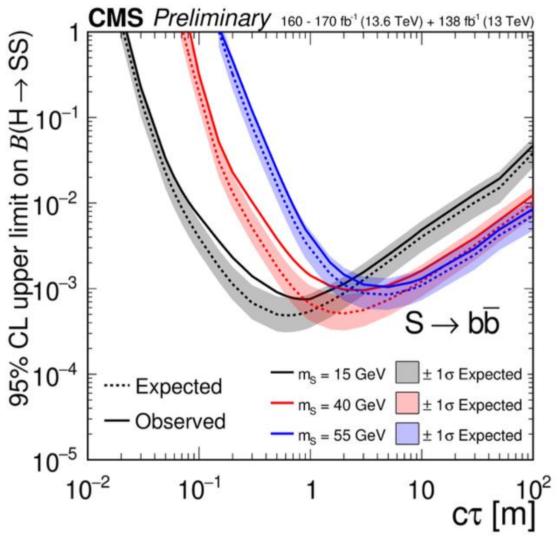
Quadrant of the CMS muon system: DTs in the barrel, CSCs in the endcaps [11]

High-pT muons

The same hits build CSC segments and standalone muons, so the method is being studied where showering degrades muon reconstruction.

Search sensitivity

What all of this does to the single-cluster category of the MDS search.



MDS search: limits on B(H → SS) against cτ [9]

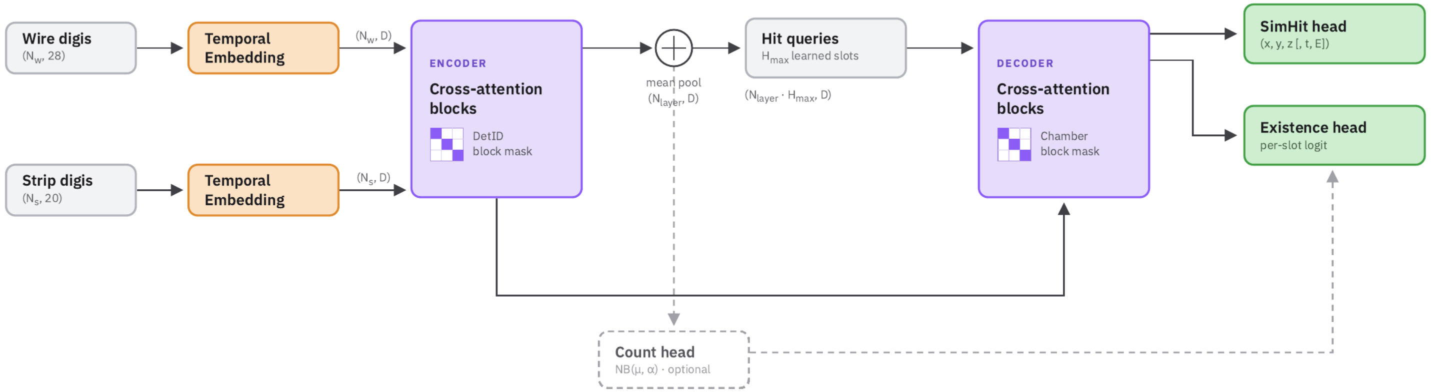
Summary

Learning CSC hit reconstruction from the raw readout

A cross-attention transformer over the raw CSC strip and wire signals, trained against Geant4 SimHits with no per-hit labels.

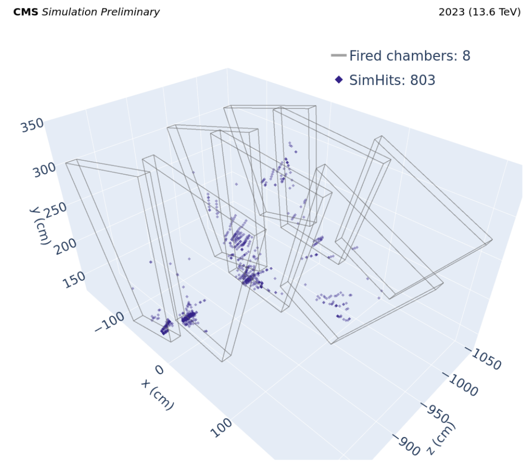
It recovers the main observable the search cuts on, the cluster hit multiplicity, together with the matching efficiency, the shower shape and timing, and a deposited-energy sum that the RecHits do not provide.

Full details in the [CMS Detector Performance Summary][12]



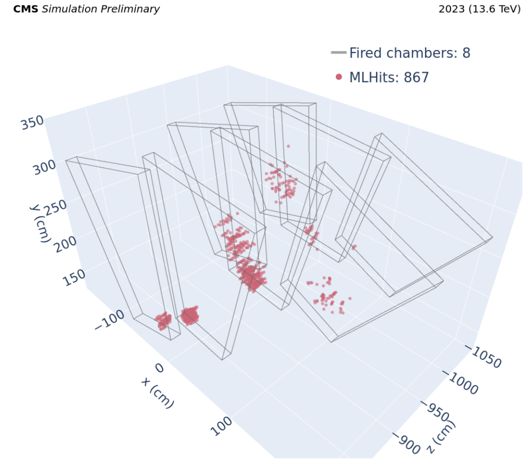
SimHit

803 hits



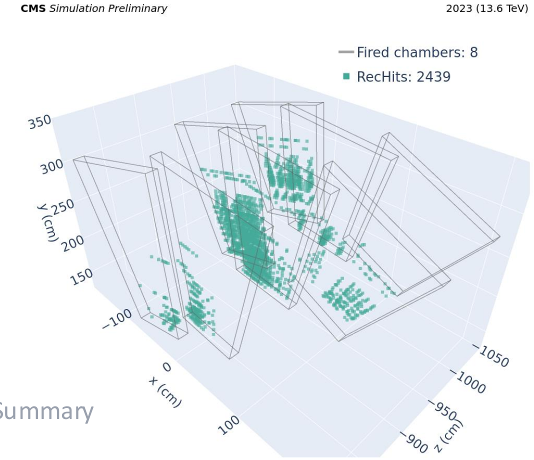
MLHit

867 hits



RecHit

2439 hits



Backup

References, the loss and its ingredients, training, attention masks, temporal encoding, extra event displays, timing

References

[1] CMS Data Preservation & Open Access group, "CMS Open Data Workshop lessons", CERN, CC-BY-4.0

[2] CMS Collaboration, "Search for long-lived particles decaying in the CMS muon detectors", EXO-20-015, Phys. Rev. D 108 (2023) 032007

[3] V. Khachatryan et al., CSC local reconstruction, EPJ Web Conf. 214 (2019) 02014 (CHEP 2018)

[4] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with pp collisions at 13 TeV", JINST 13 (2018) P06015

[5] A. Vaswani et al., "Attention Is All You Need", arXiv:1706.03762

[6] N. Carion et al., "End-to-End Object Detection with Transformers", arXiv:2005.12872

[7] J. Dong et al., "Flex Attention: A Programming Model for Generating Optimized Attention Kernels", arXiv:2412.05496

[8] S. Agostinelli et al., "Geant4: a simulation toolkit", NIM A 506 (2003) 250

[9] CMS Collaboration, "Search for long-lived particles decaying in the CMS muon detectors", CMS-PAS-EXO-25-005

[10] M. Ester et al., "A density-based algorithm for discovering clusters", KDD 1996 (DBSCAN)

[11] D. Abbaneo et al., "Studies on the upgrade of the muon system in the forward region of the CMS experiment at LHC with GEMs", JINST 9 (2014) C01053

[12] CMS Collaboration, "Cross-Attention-Based Reconstruction of Muon Detector Showers in the CMS Endcap Muon Detectors from Low-Level Inputs", CMS Detector Performance Summary, 2026

[13] T.-Y. Lin et al., "Focal Loss for Dense Object Detection", arXiv:1708.02002

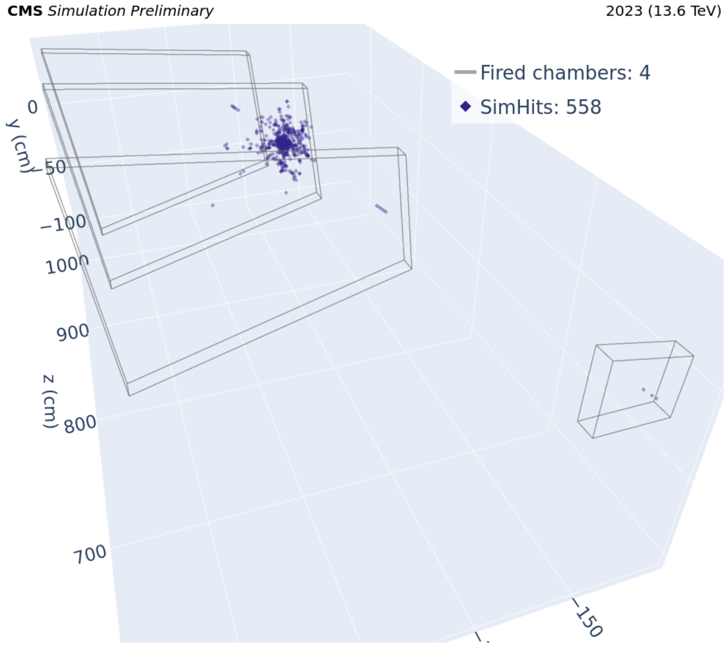
[14] M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport", NeurIPS 2013

[15] P.-E. Sarlin et al., "SuperGlue: Learning Feature Matching with Graph Neural Networks", arXiv:1911.11763 (dustbin formulation)

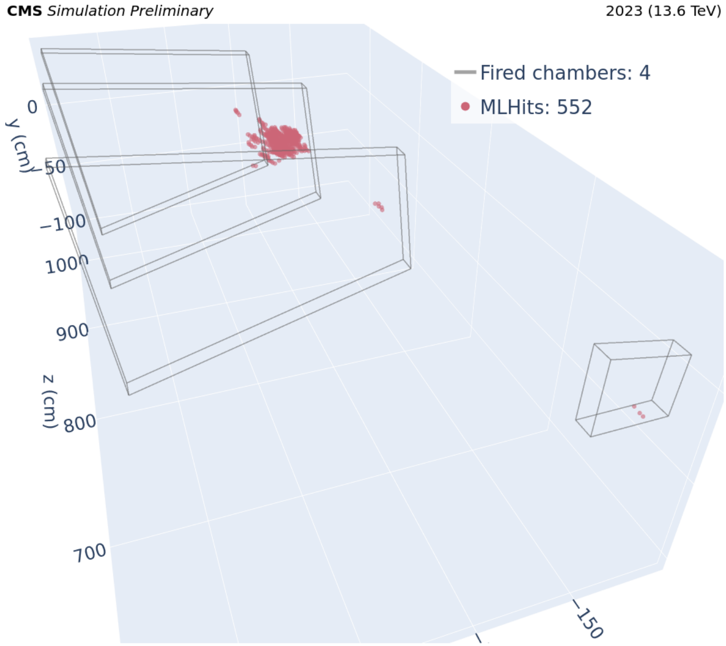
[16] Z. Cao et al., "Learning to Rank: From Pairwise Approach to Listwise Approach", ICML 2007 (ListNet)

Prompt-muon shower cluster

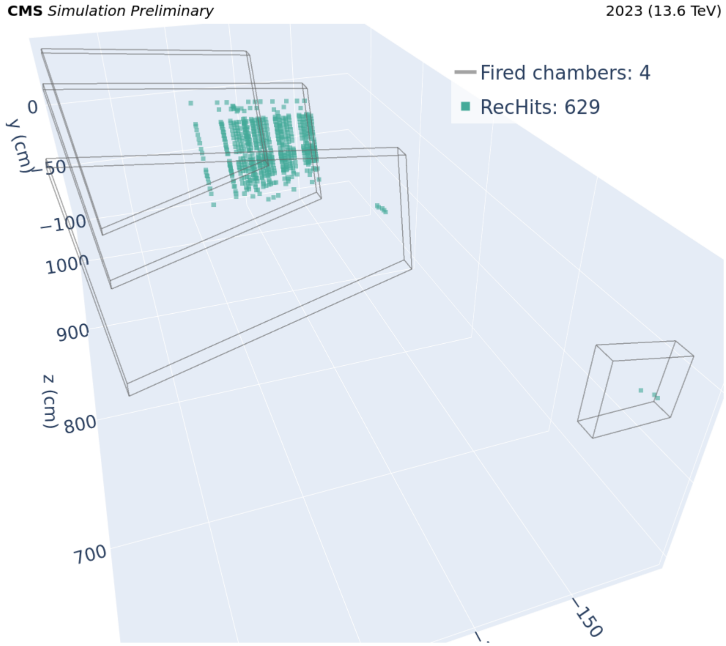
Four fired chambers



SimHit 558



MLHit 552

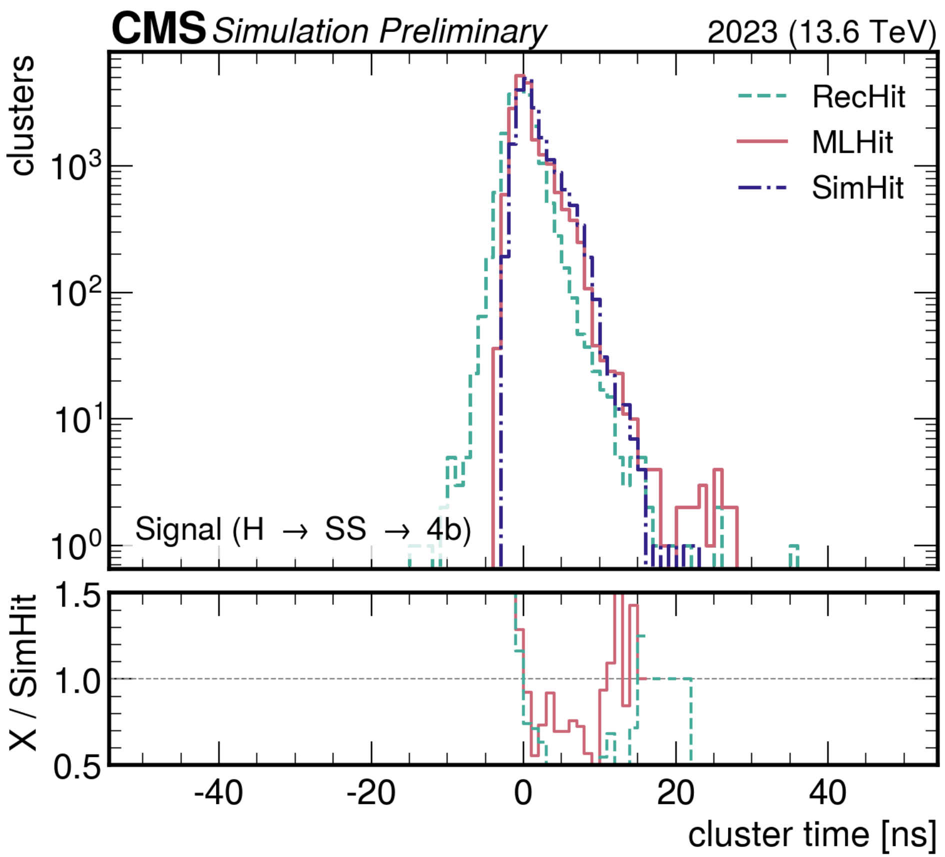


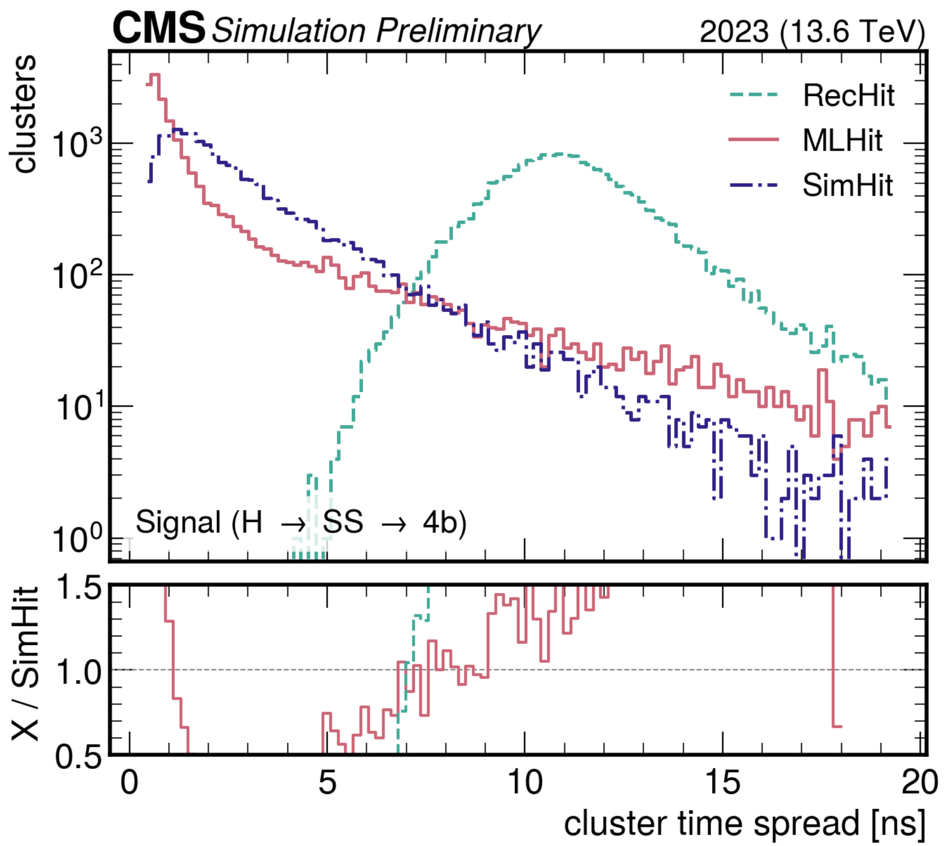
RecHit 629

A prompt-muon shower rather than an LLP decay: 552 MLHits against 558 true deposits, 629 RecHits.

Timing

Cluster time and time spread, rebuilt from each hit collection, LLP signal

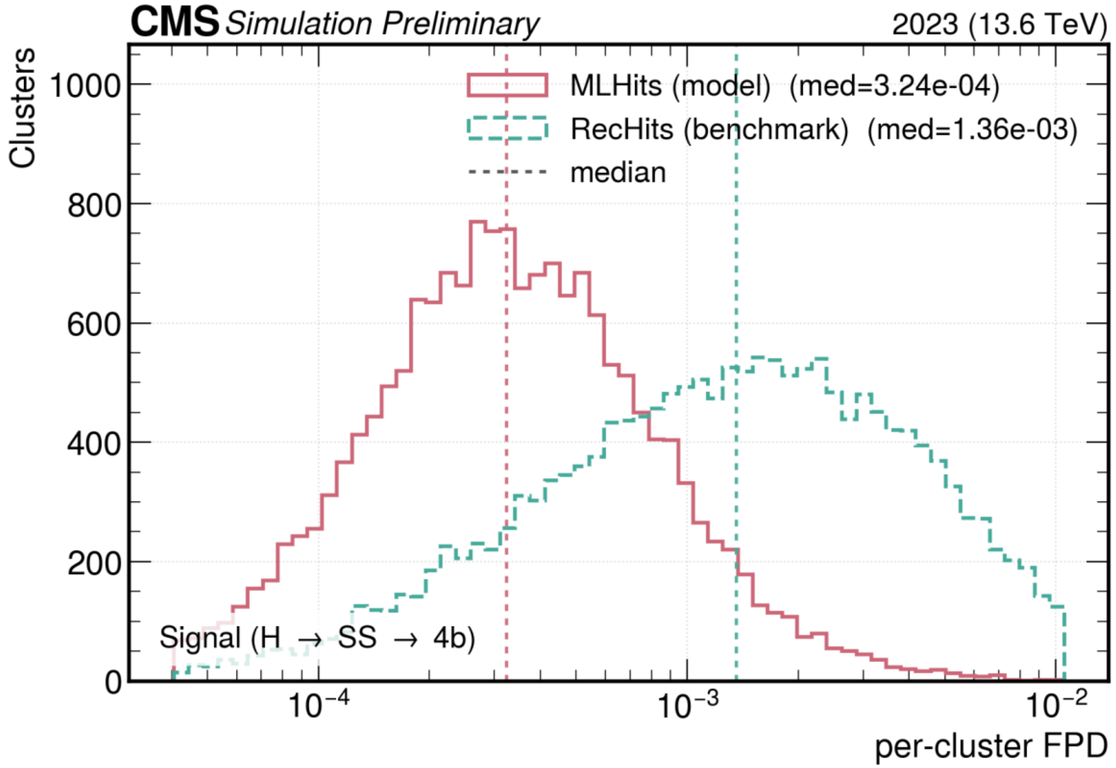




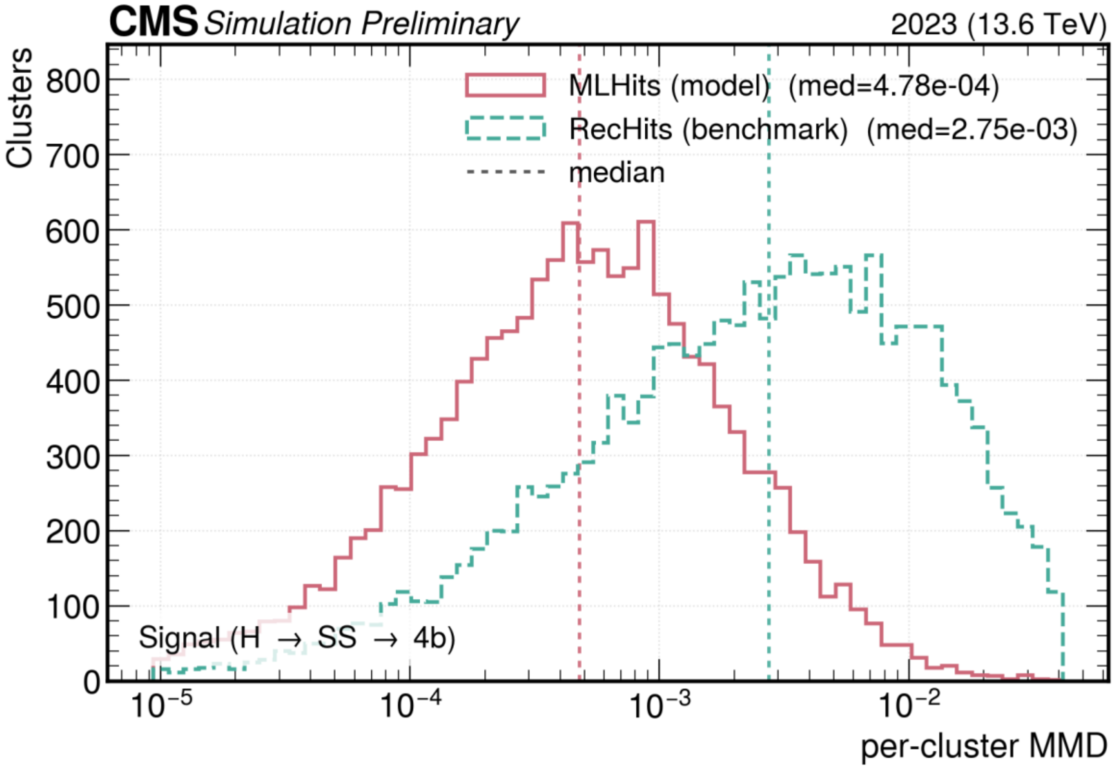
Cluster time is the mean of the hit times in the cluster, and it is the handle the search uses against beam-induced background.

Distribution-level agreement

FPD and MMD² between reconstructed and true hit clusters, LLP signal



FPD 3.2×10^{-4} against 1.4×10^{-3}



MMD²

Loss (DP note)

Set prediction with no per-hit labels: the version behind the results shown

Assignment

Greedy mutual-nearest matching on transverse distance, up to four rounds. z is pinned by geometry and the matching carries no gradient to the heads.

Geometry

Squared transverse distance on the matched pairs; time and energy enter the matching cost as tie-breakers.

$$\mathcal{L} = \lambda_{ex}\,\mathcal{L}_{focal} + \lambda_{pos}\,\mathcal{L}_{pos} + \lambda_t\,\mathcal{L}_t + \lambda_E\,\mathcal{L}_E + \lambda_N\,\mathcal{L}_{count}$$

Existence and order

Focal loss [13] on a binary per-slot target: a slot either holds a matched hit or it does not.

Count

A Negative-Binomial fit to the number of true hits in each layer. At inference the K best slots per layer are kept, with K taken from that fit.

Every term is reweighted per layer, $\propto (\mathrm{N}_{\mathrm{truth}} + 1)^{-p}$, so the dense layers do not dominate the sparse ones.

[13] Focal loss

Loss (current)

What changed after the DP note, in the dashed boxes

Assignment

Sinkhorn optimal transport [14] spreads each true hit over nearby slots, with a dustbin column [15] for slots that match nothing.

was: greedy mutual-nearest matching

Geometry

Distances weighted by those assignment weights, plus each true hit's distance to its nearest slot, so no deposit is ignored.

was: distance on the matched pairs

$$\mathcal{L} = \boxed{\lambda_{\text{geo}}\,\mathcal{L}_{\text{geo}}} + \boxed{\lambda_{\text{ex}}\,(\mathcal{L}_{\text{focal}} + 0.1\,\mathcal{L}_{\text{ListNet}})} + \lambda_t\,\mathcal{L}_t + \lambda_E\,\mathcal{L}_E + \lambda_N\,\mathcal{L}_{\text{count}}$$

Existence and order

Focal loss [13] against the assigned mass rather than yes or no, plus ListNet [16] on the slot scores, since only the top-K are kept.

was: focal on a binary target, no ranking term

Count

A Negative-Binomial fit to the number of true hits in each layer. At inference the K best slots per layer are kept, with K taken from that fit.

The results shown were produced with the previous version.

[13] Focal loss · [14] Sinkhorn · [15] SuperGlue dustbin · [16] ListNet

Training and inference

The configuration behind the results shown

Samples

258 248 training, 32 281 validation, 32 282 test clusters, about 62 % signal.

Model

1.34 M parameters. Embedding 128, three encoder and three decoder blocks, 8 heads, 512 query slots per layer, MLP width 512, dropout 0.1.

Schedule

30 epochs, 10 h training time on 4 × Tesla V100-SXM2 32 GB, one process per GPU.

Loss weights

λ_geo 5.0, λ_ex 0.1, λ_t 0.01, λ_E 0.01, λ_count 0.1; focal $\alpha = 0.75$, $\gamma = 2.0$; per-layer reweighting exponent 0.3; time and energy tie-breakers at 0.1.

Batching

Dynamic token budget rather than a fixed batch: about 1200 strip tokens per step, ≈ 3.4 clusters per step and ≈ 18 900 steps per epoch. Full precision.

Inference

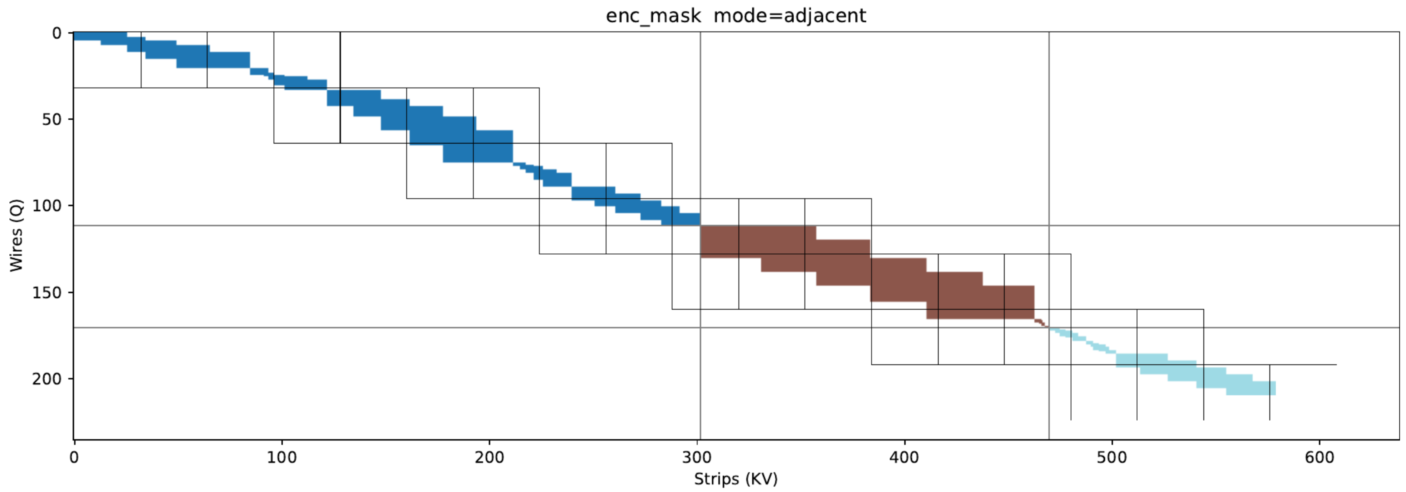
6.9 ms per event on a V100 at batch 1.

Learning rate

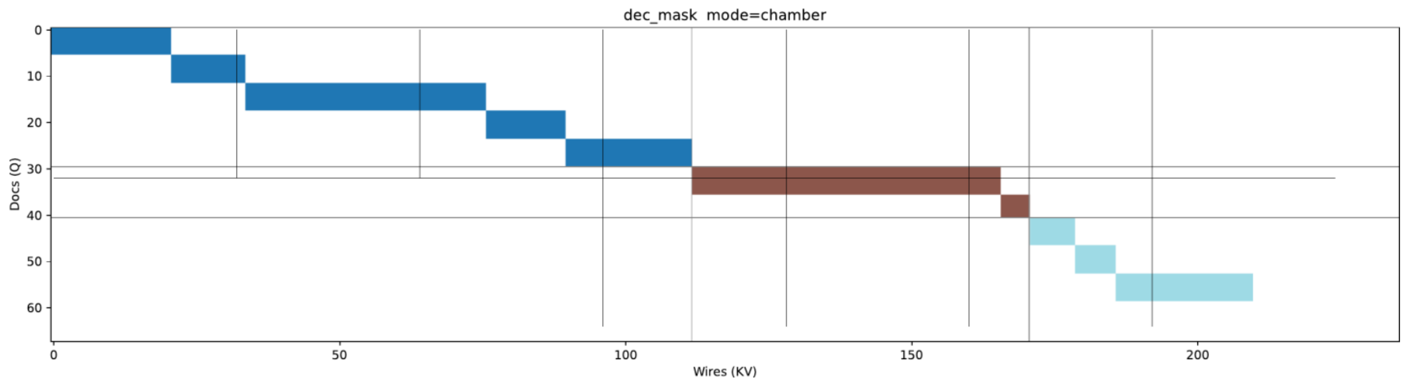
AdamW, 1×10^{-6}, with a 1×10^{-5} delta on the newly added parameters.

Attention masks

Physical adjacency written into the attention pattern



Encoder: strips in adjacent layers of the same chamber

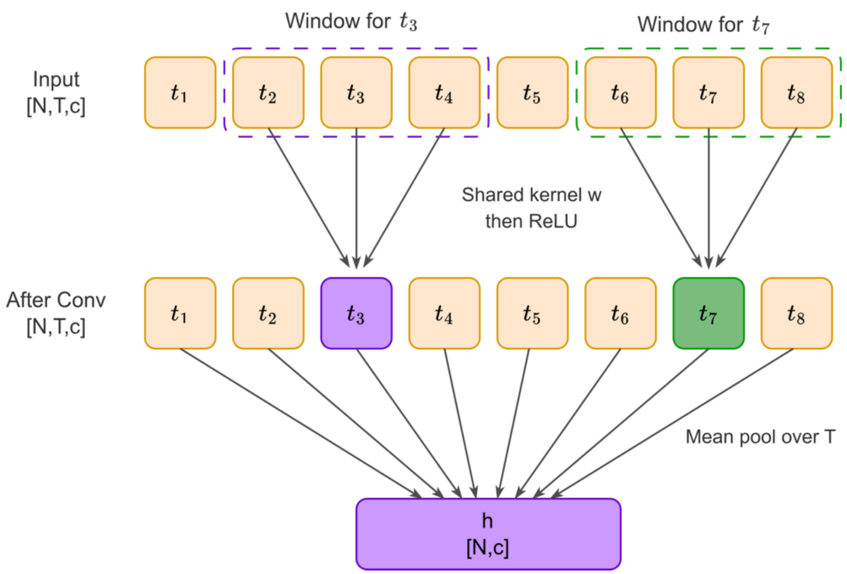


Decoder: the whole chamber

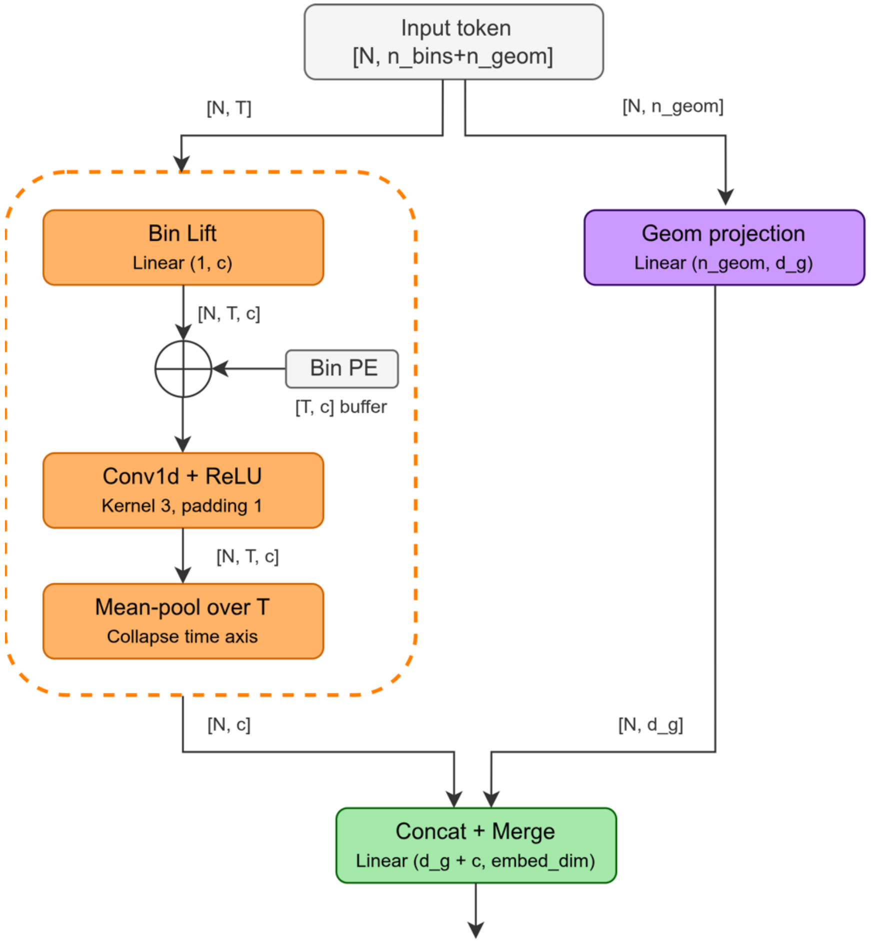
Blue marks the pairs a token may attend to. In the encoder a wire group sees only strips in adjacent layers of its own chamber; in the decoder every query slot sees the whole chamber.

Temporal Encoding

How each token's time series becomes part of its embedding



the shared kernel sliding over the time bins



the same thing as a block diagram

Lift. Each time bin value goes to c channels through one Linear(1, c), shared over the bins.

Time. A sinusoidal encoding of the bin time in ns is added, so 50 ns strip bins and 25 ns wire bins sit on one time axis.

Shape. Conv1d (kernel 3) + ReLU along the time axis: the same kernel reads the rise, peak and tail wherever the pulse sits.

Pool. The mean over the T bins gives a c-dim pulse feature, independent of T (8 for strips, 16 for wires).

Merge. Geometry is projected to d_g, concatenated with the pulse feature, and a Linear gives the 128-dim token.