ORBIT: Online Real-time Bandwidth reduction via Information Tokens

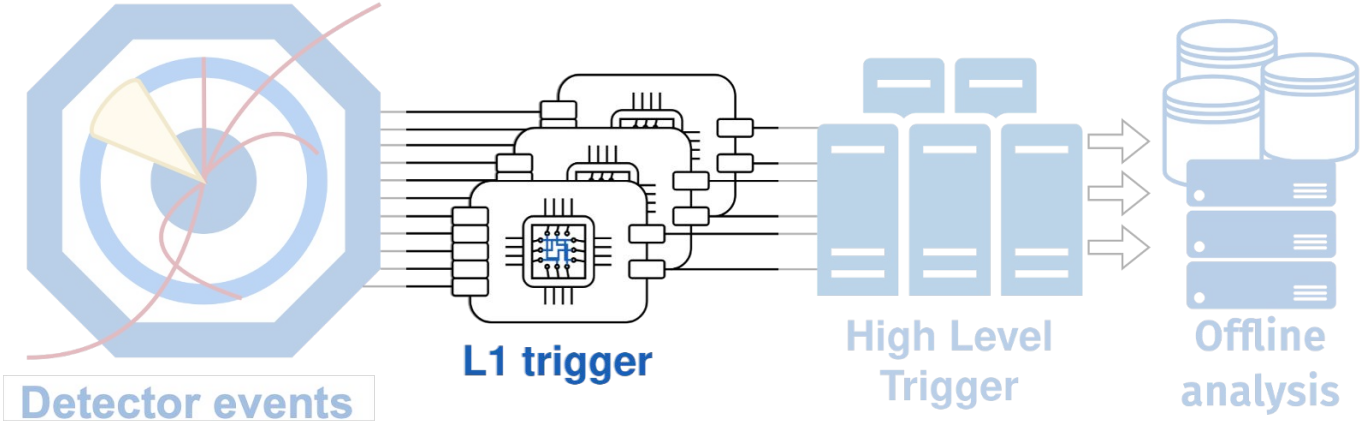
Younes Elberkennou, Philipp Wagner,
Philip Ploner, Francesco Proietti, Donatella Genovese,
Maurizio Pierini, Giovanni Petrucciani, Jennifer Ngadiuba,
Sabrina Giorgetti and Thea Klaeboe Aarrestad

17.08.2026, ML4Jets 2026

The HL-LHC Data Problem

ETH zürich

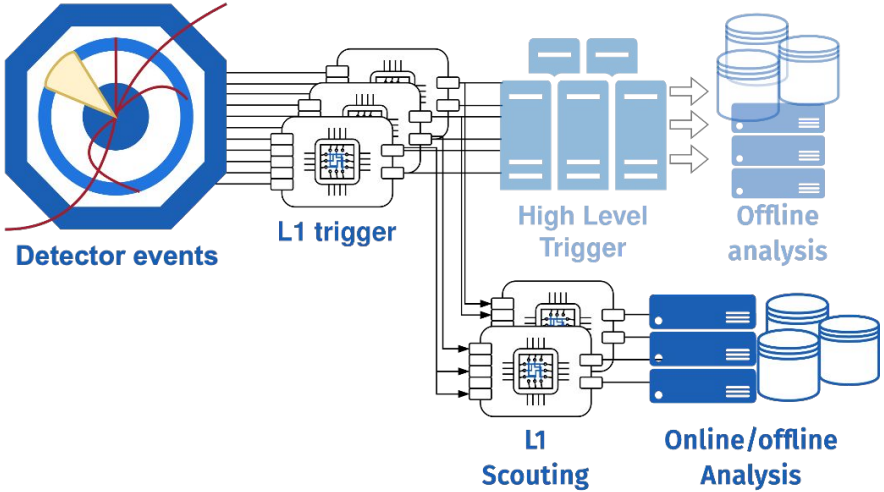
- **40 MHz bunch-crossing rate** → O(10 Tb/s) of raw detector data
- Phase-2 Trigger at CMS discards aggressively: L1 trigger accepts at less than 750 kHz, the High-Level Trigger less than 7.5 kHz
- Fixed trigger thresholds bias what's kept: new physics may be thrown away!



CMS Phase-2 L1 Scouting

ETH zürich

- Phase-2 L1 trigger will perform offline-like particle-flow (PF) reconstruction. Sufficient quality for physics analysis!
- **40 MHz Data Scouting** reads out L1T reconstructed physics objects for every bunch crossing
- Access to **every single collision event**, but at a lower resolution



L1 scouting system: data requirements

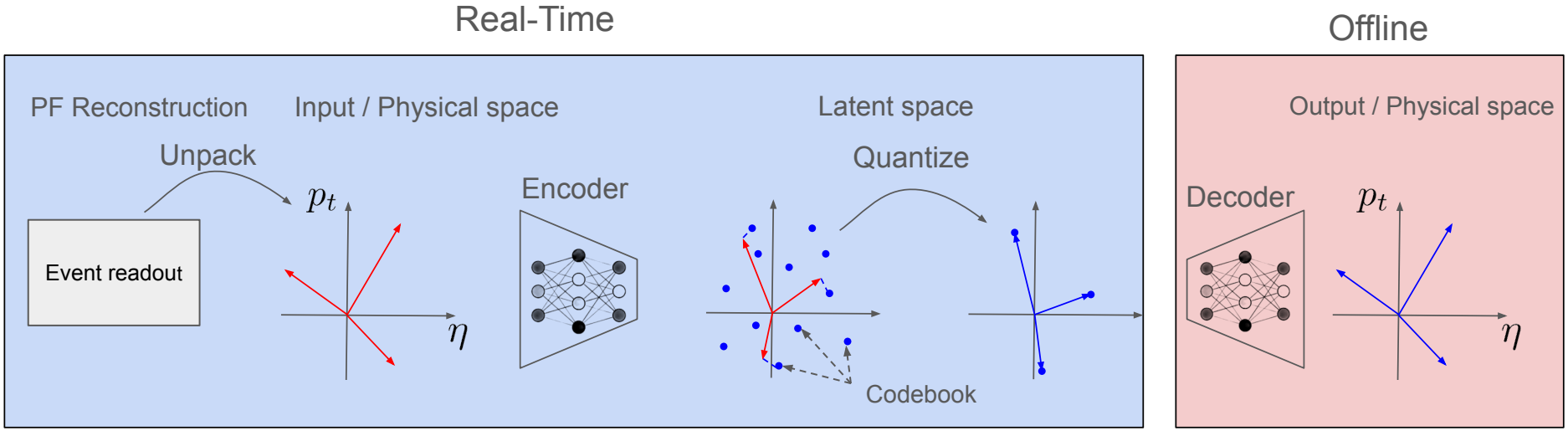
ETH zürich

Scenario	Event size	Selection	Compression	Throughput	Dataset/year
L1 Trigger (baseline)	4 kB	1	1	120 GB/s	888 PB
L1 Scouting, with PU subtraction	0.4 kB	1	1/2.4	5 GB/s	37.0 PB
L1 Scouting, no PU subtraction	4 kB	1/3	1/8	5 GB/s	37.0 PB

- Total L1 Trigger throughput is 120 GB/s, **must be reduced** to the scouting bandwidth of 5 GB/s: 24x reduction for all PF particles, 2.4x for PU subtracted.
- Pileup subtraction might reduce sensitivity to New Physics (e.g soft signatures)
- How to achieve data reduction without significant degradation?
  - Bandwidth allocation problem: Neural nets are good at learning efficient representations!

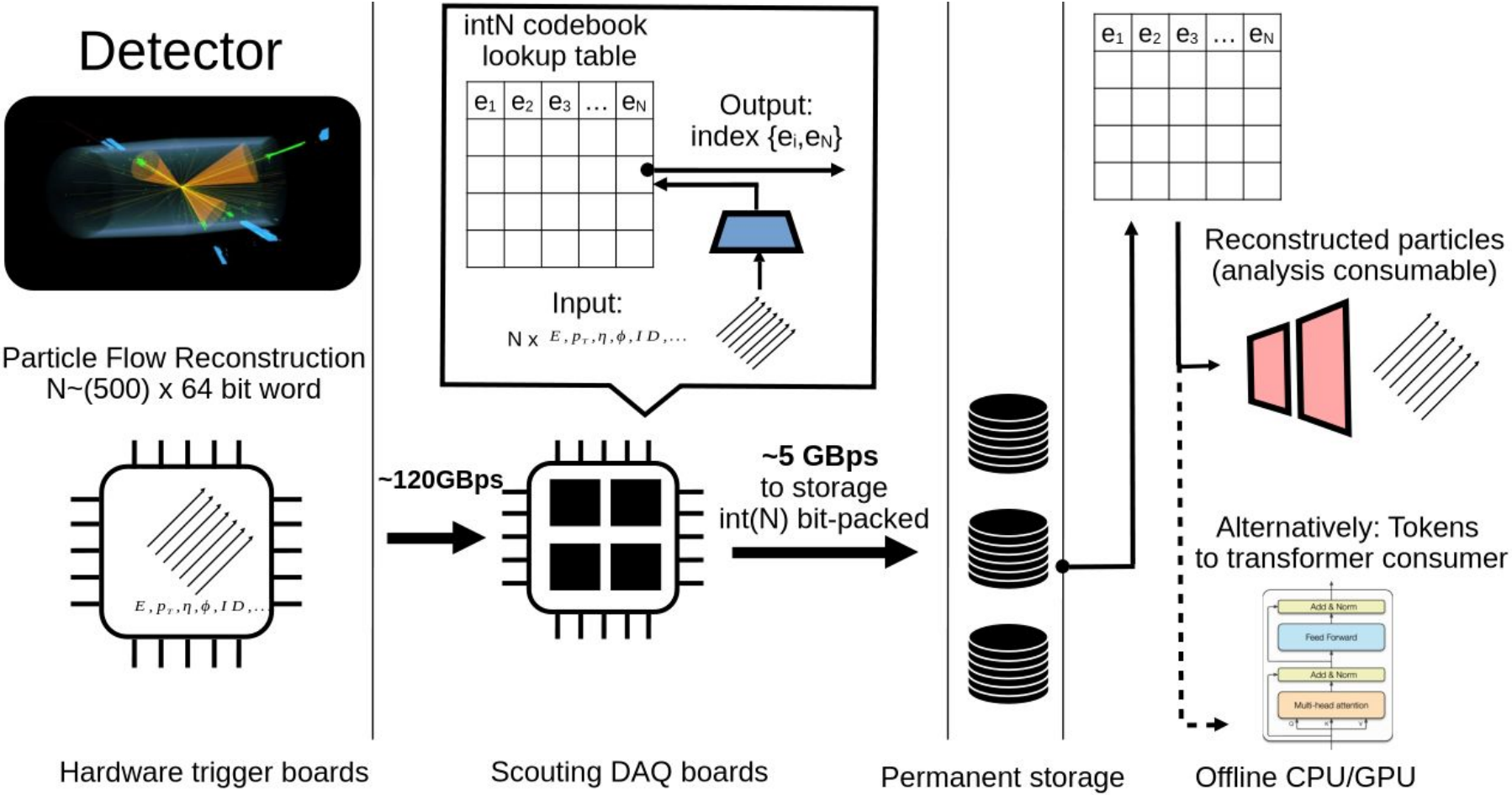
Our Approach: Neural tokenization

- Each particle becomes a token - each event becomes sequence of integers
- Map vectors in latent space onto a discrete, finite codebook of tokens
- Store an index as int(N), reducing number of bits needed to store an event
- Decode codebook for downstream applications



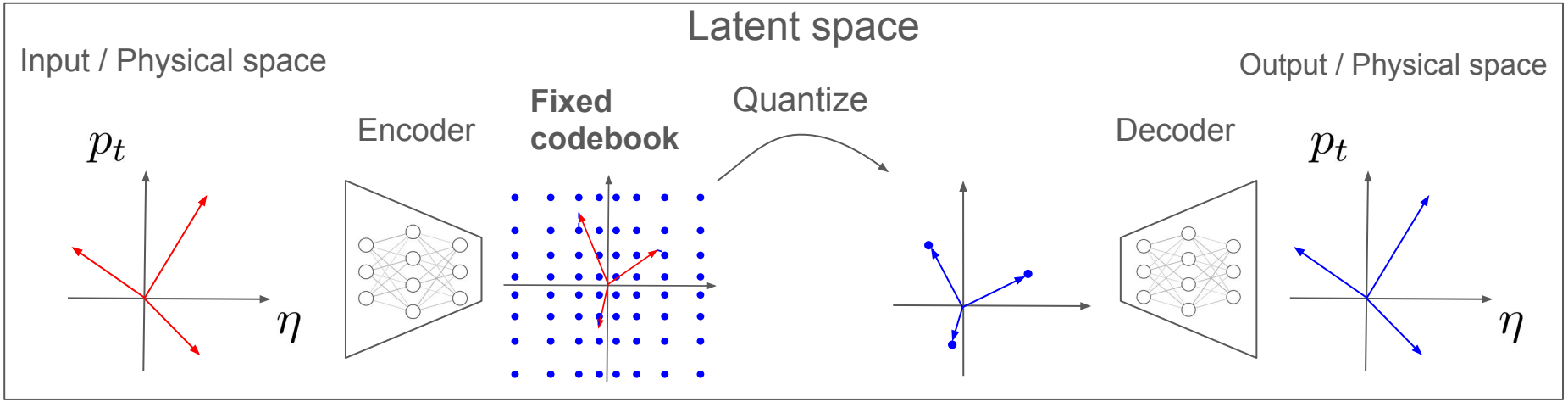
ORBIT: Online Real-time Bandwidth reduction

ETH zürich



Quantization Toolbox I:
Finite Scalar Quantization [2309.15505]

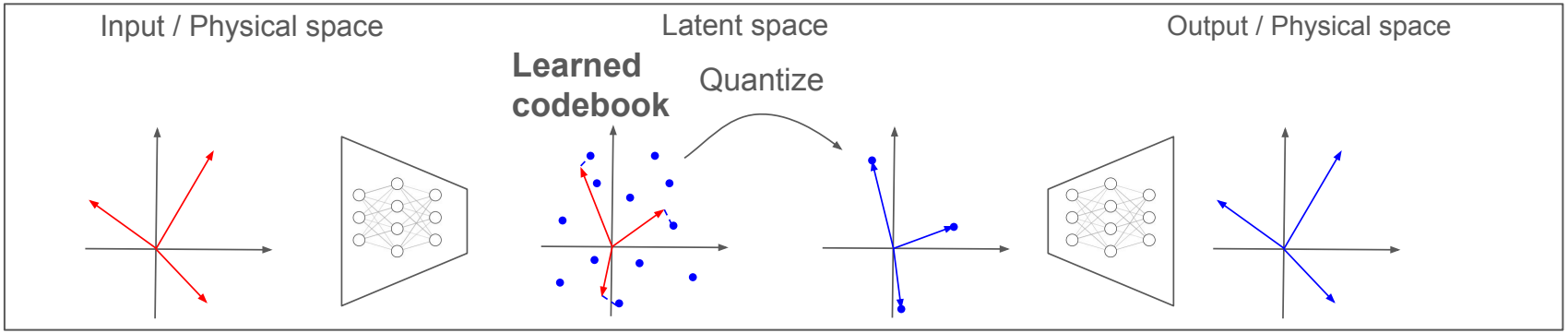
- Each latent dimension is quantized independently onto a **fixed scalar grid**
- Implicit codebook of size $K = m^d$ (d = latent dimension, m = granularity)
- In our case, apply the operation $\hat{z} = \text{round}\left(\frac{m-1}{2}\tanh(z)\right)$ for each dimension



Quantization Toolbox II: VQ-VAE [1711.00937]

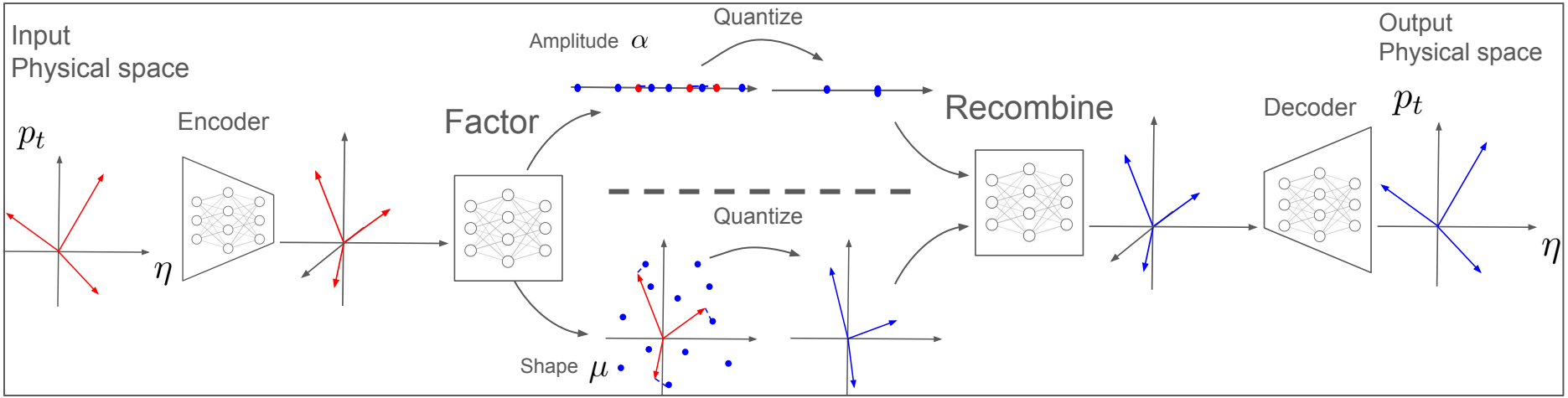
ETH zürich

- **Learned codebook:** encoder output $z(x)$ is snapped to the nearest of K learned embeddings $e_1, \ldots, e_K$
- Quantization is non-differentiable: deal with quantization step during backpropagation using straight-through estimator/rotation trick



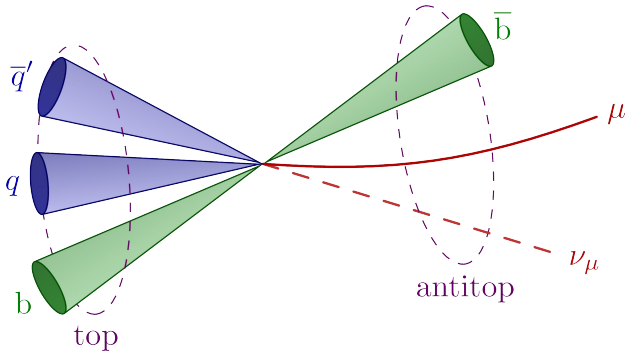
Quantization Toolbox III: PHAEDRA [2602.03915]

- PHAEDRA splits latent vectors into a shape component (direction) and a gain component (magnitude)
- Each branch is quantized separately, using its own codebook
- Hope: shape and amplitude learned separately, leading to a more accurate total reconstruction, especially in tails

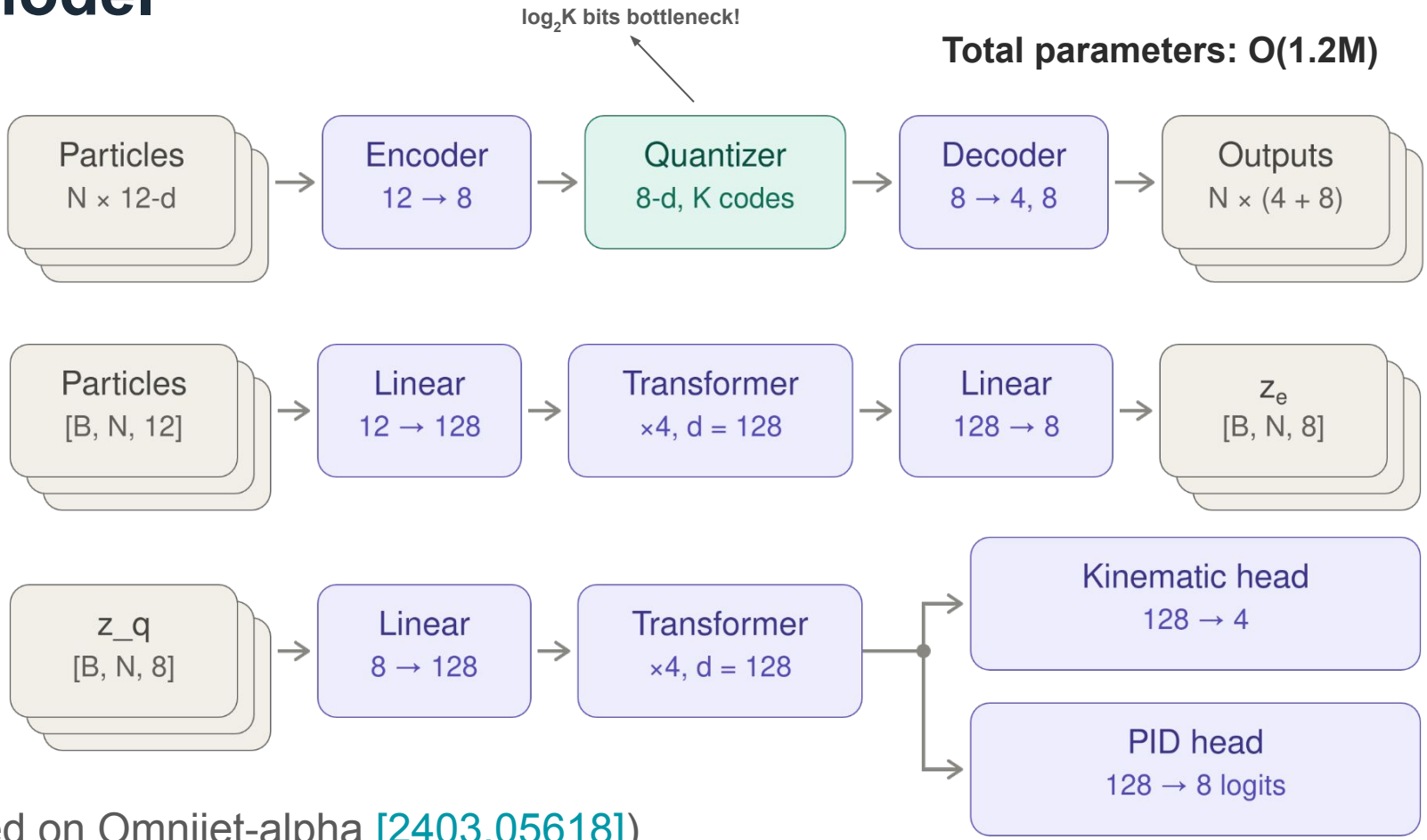


Data & Setup

- **Training is performed on two datasets (Collide-2V, L1 view):**
  - $t\bar{t}$ inclusive (high multiplicity, jets/leptons/MET simultaneously), useful because this has high object multiplicity and diversity, so tokenizer learns to deal with complicated events
  - SM-mixture cocktail ($t\bar{t}$ + QCD + V+jets + VV, constant proportions), 100k events, exposes the tokenizer to realistic conditions it might see at L1T
- **Test:** $gg \to H \to b\bar{b}, t\bar{t}$, SM cocktail, ZeroBias (10k events)
- **Per-particle features:** $\log(p_T) - 1.8,\ \cos\varphi,\ \sin\varphi,\ \eta/3,\ \mathrm{PID}$ (one-hot, 8 classes)
- **Currently:** 128 pileup-subtracted particles, expect O(500) PF candidates in L1T

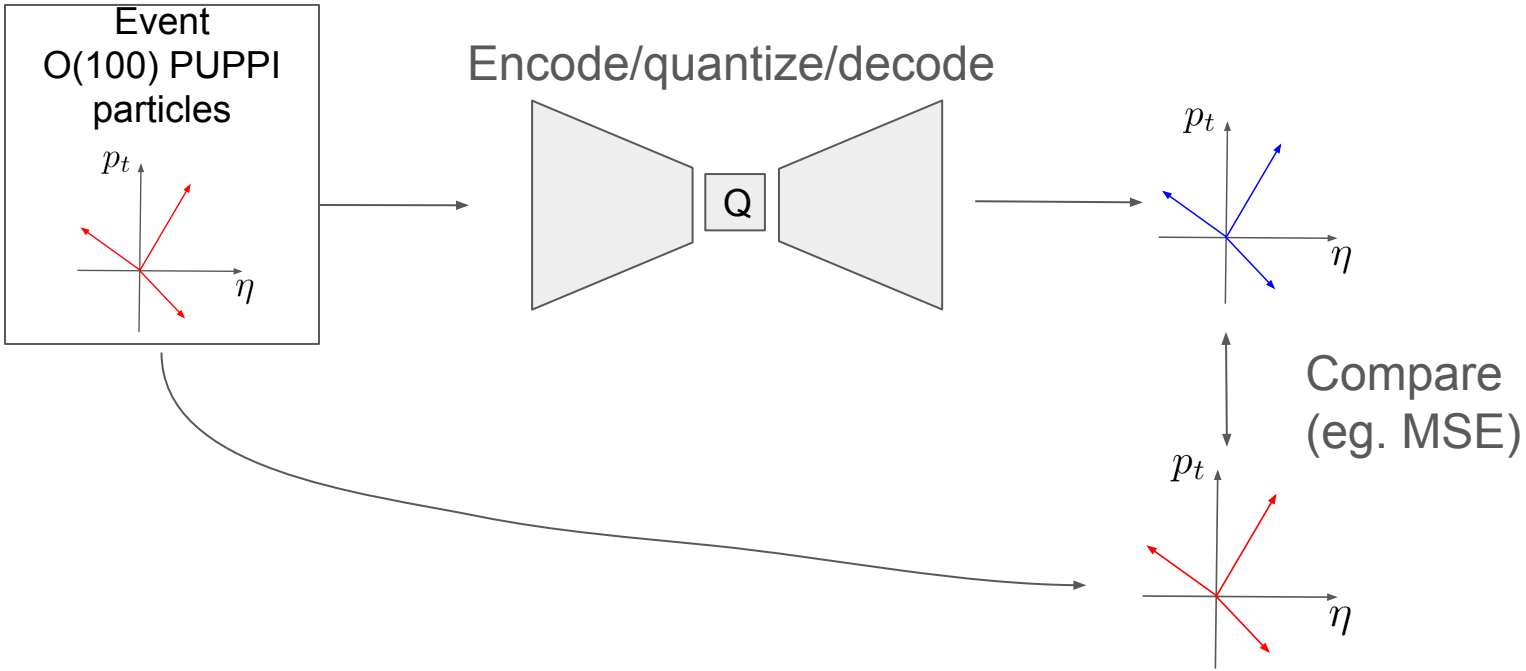


Model



(based on Omnijet-alpha [2403.05618])

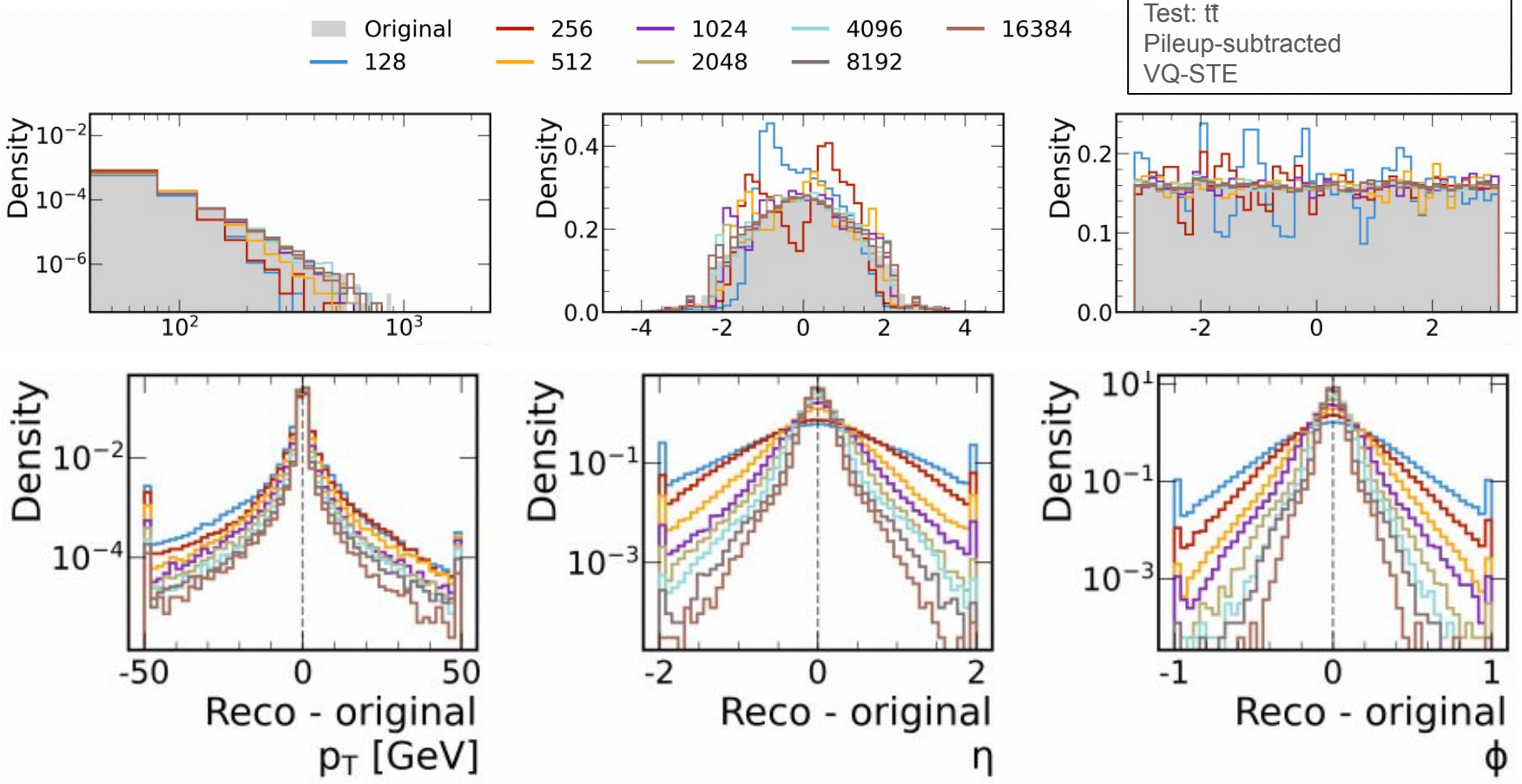
Reconstruction fidelity



Reconstruction fidelity scan

ETH zürich

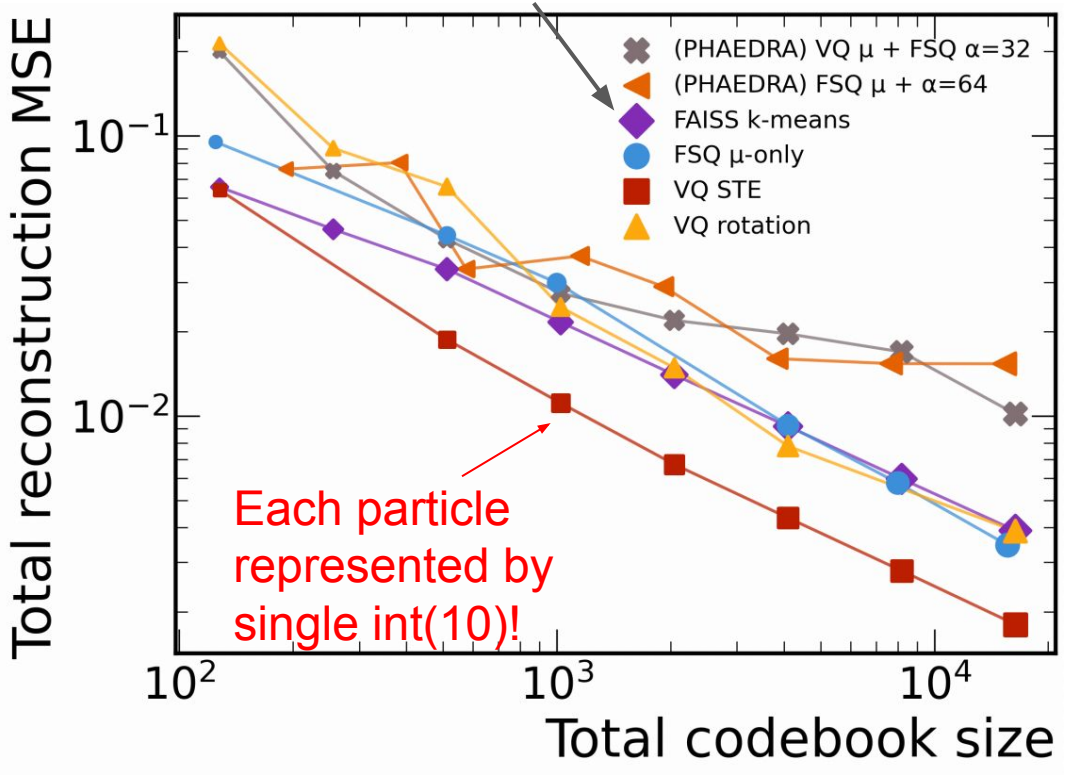
Train: tt̄
Test: tt̄
Pileup-subtracted
VQ-STE



Reconstruction fidelity comparison

ETH zürich

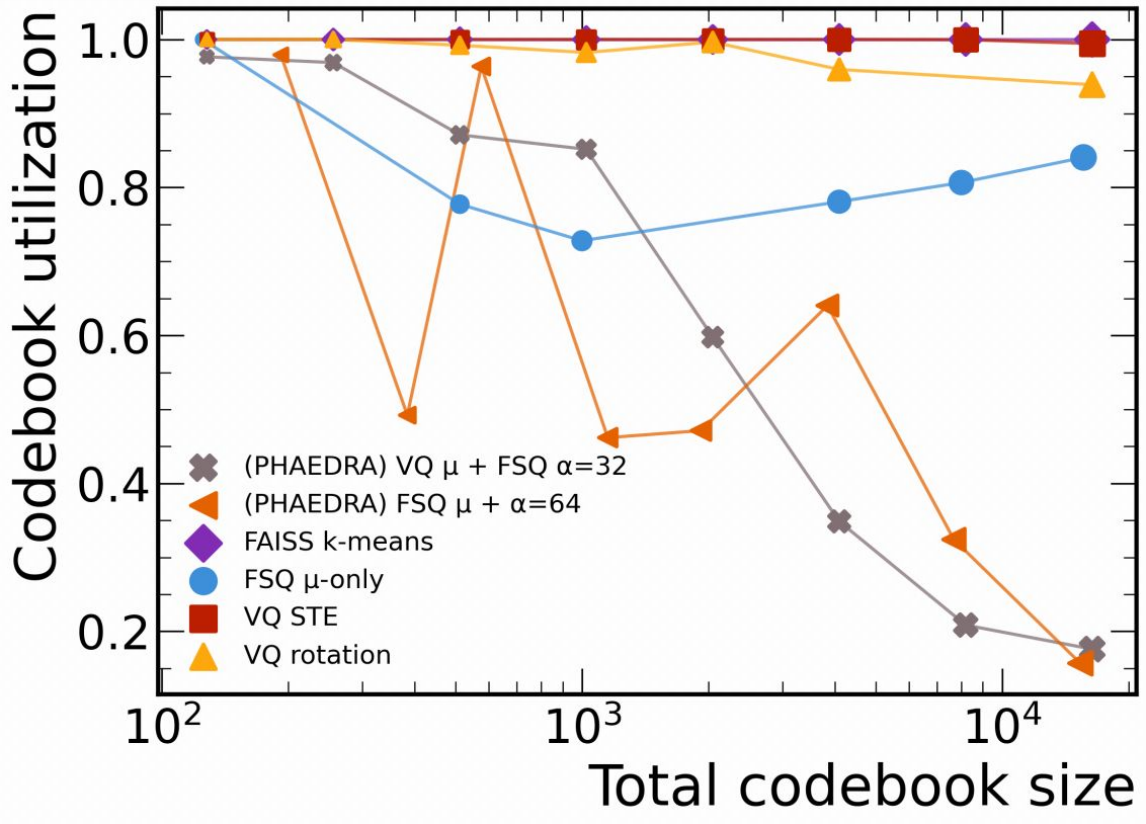
Baseline: K-means clustering in physical space



Train: $t\bar{t}$
Test: $t\bar{t}$
Pileup-subtracted

Codebook size ∝ data type! E.g. codebook size 1024→int(10)

Codebook usage comparison



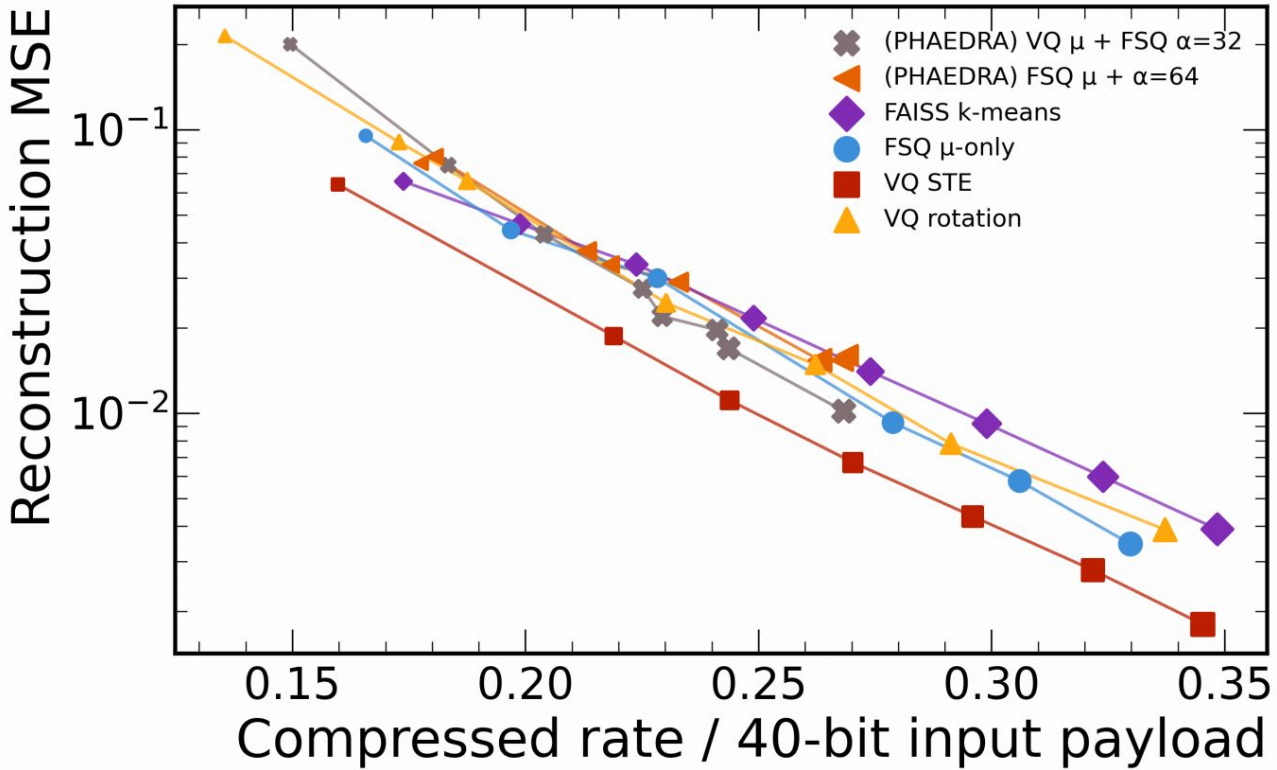
Train: $t\bar{t}$
Test: $t\bar{t}$
Pileup-subtracted

Codebook utilization is the portion of codes active in any event of the test set.

Bandwidth requirement not only a function of dictionary size, but also of the code frequency distribution. This varies significantly.

Reconstruction fidelity vs. compression

ETH zürich



Train: $t\bar{t}$
Test: $t\bar{t}$
Pileup-subtracted

Including utilization, VQ-VAE STE still most efficient

Compression factor (conversion to ROOT storage format)

Train: $t\bar{t}$
Model: VQ-STE, 4096 codes
Pileup-subtracted
1K events

Tokenization yields ~2.5-3x compression

Compression: LZMA Level 9

	Raw	EDM	nanoAOD
Sample	B (%)	B (%)	B (%)
Minimum bias, plain	23.0 (100.0)	27.2 (118.2)	19.1 (82.9)
Minimum bias, tokenized	10.3 (44.6)	9.3 (40.4)	7.6 (33.2)
$H \to b\bar{b}$, plain	126.4 (100.0)	113.7 (90.0)	98.1 (77.6)
$H \to b\bar{b}$, tokenized	41.2 (32.6)	44.8 (35.5)	41.3 (32.6)
$t\bar{t}$, plain	207.2 (100.0)	179.8 (86.8)	159.1 (76.8)
$t\bar{t}$, tokenized	65.5 (31.6)	72.0 (34.7)	67.1 (32.4)

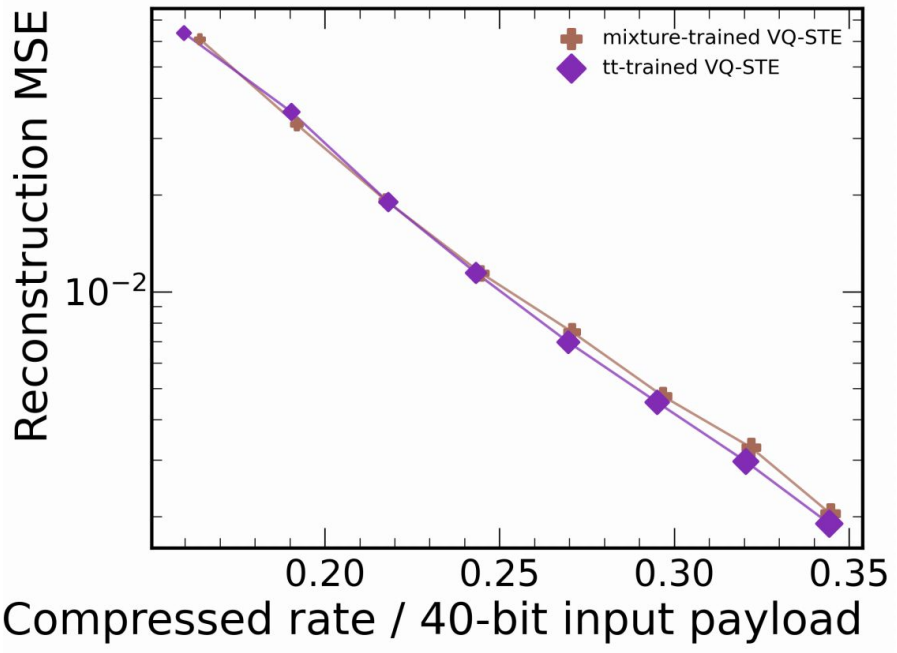
Before tokenization

After tokenization

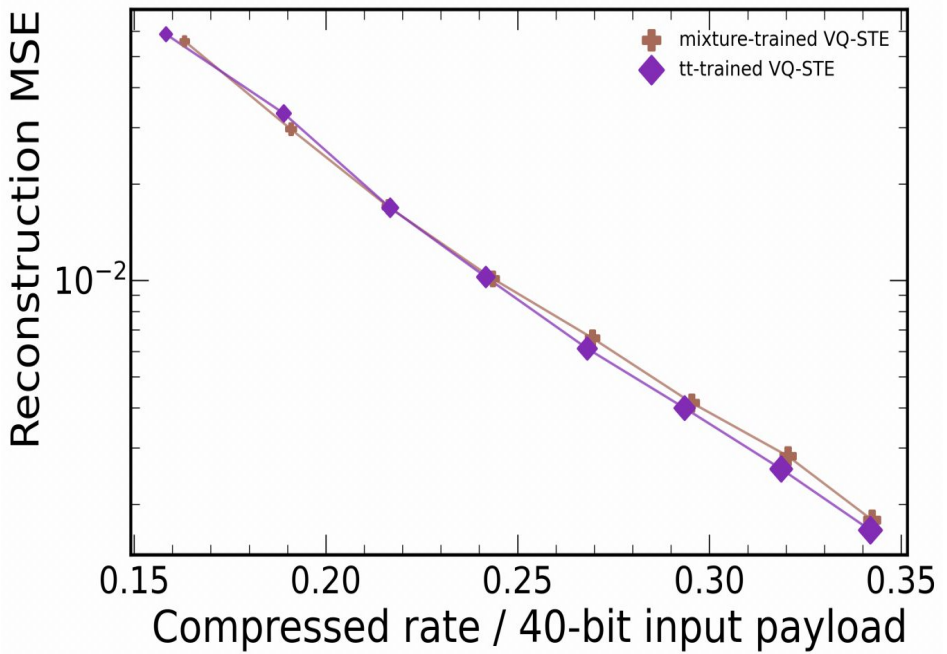
Ongoing comparison of training dataset

ETH zürich

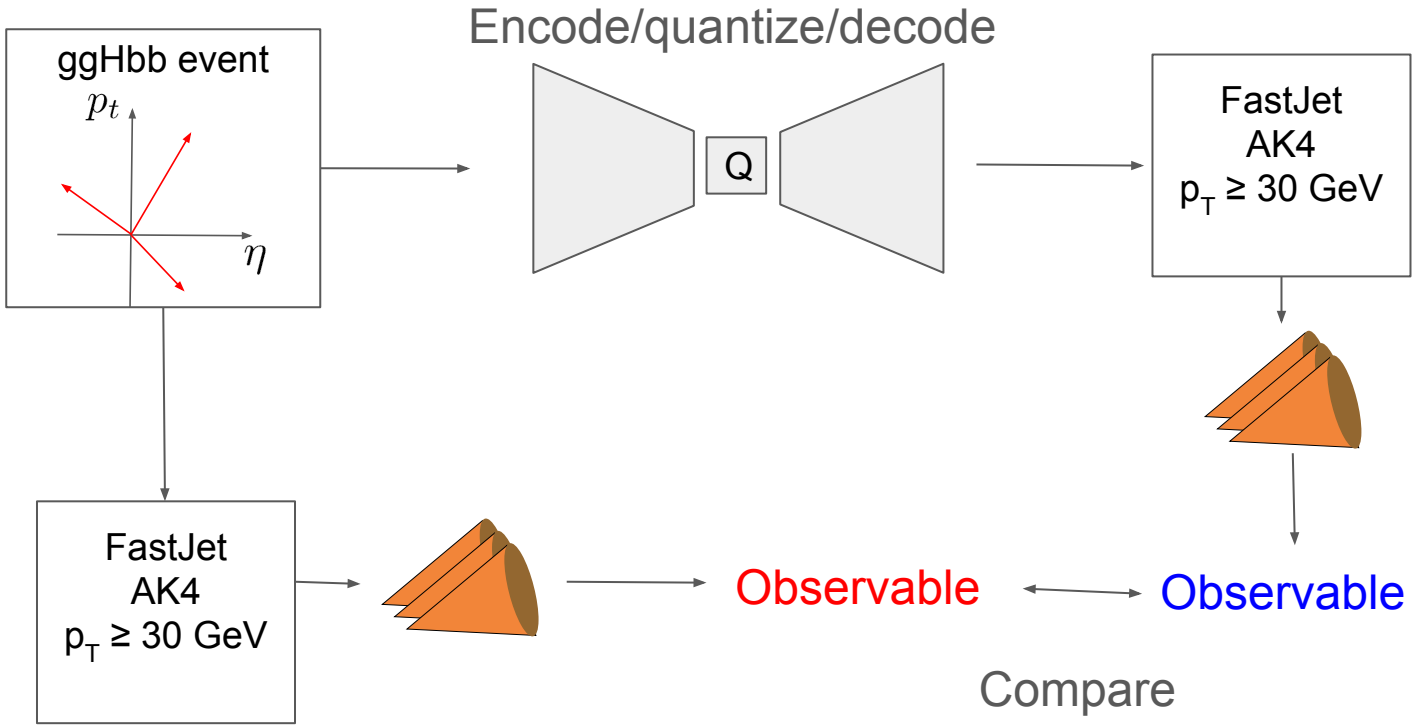
Test = SM mixture
Model = VQ-STE
Pileup-subtracted



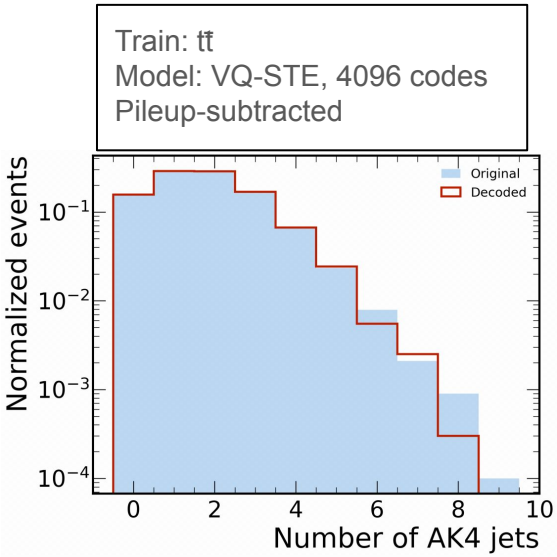
Test = ggHbb
Model = VQ-STE
Pileup-subtracted

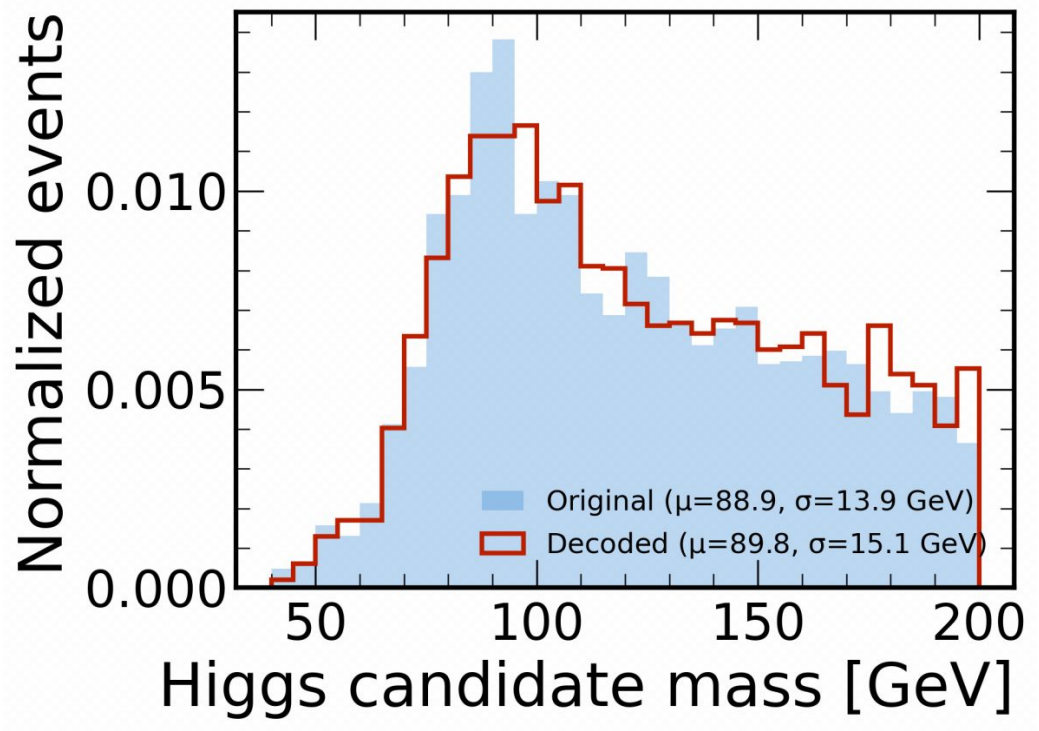


Downstream physics observables



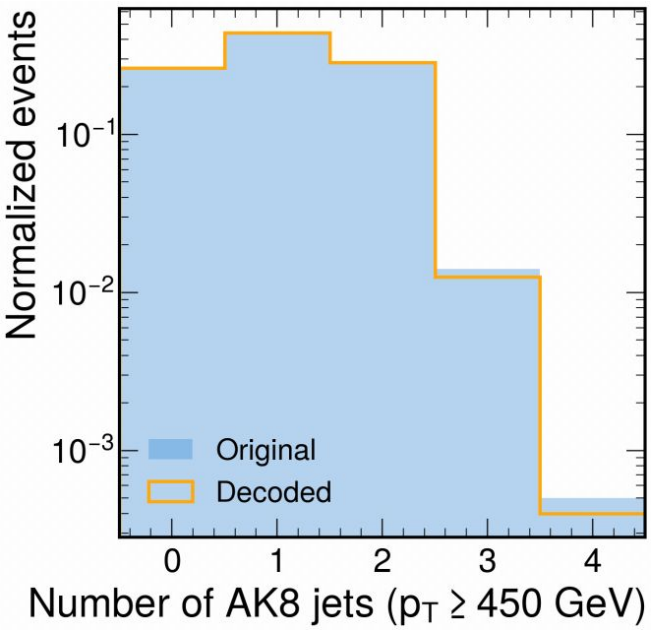
Downstream physics performance: Higgs mass reconstruction from gg→H→bb

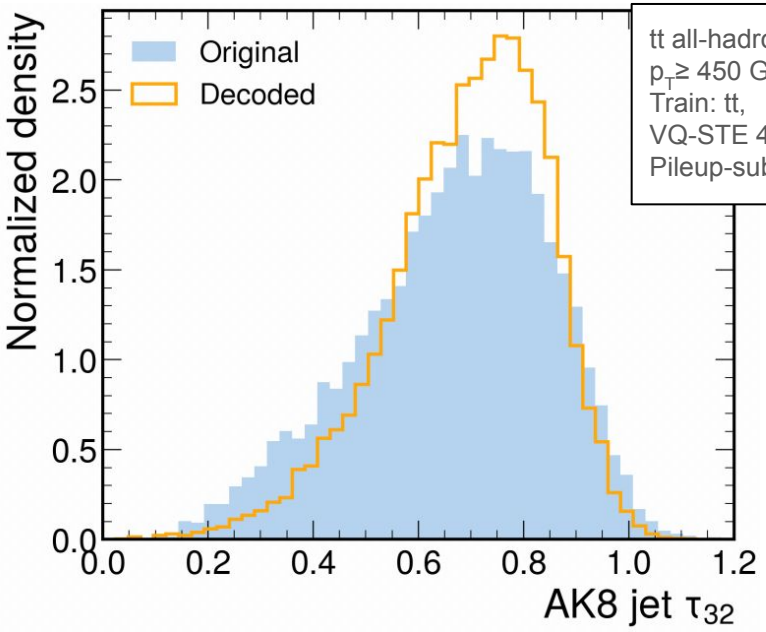




Downstream physics performance: τ_{32}

Quantizing particle directions hits hardest on observables built from spatial distribution of constituents (soft-drop mass, n-subjettiness, energy correlation functions).





Conclusions and outlook

ETH zürich

- Plain VQ-VAE is most efficient for compression all-around
- Still need higher compression factor for full PF
- Reconstruction not perfect (especially for large radius jet observables)

Next steps

- Scale up to O(500) particles per event (before Level-1 PU subtraction)
- Improve reconstruction in distribution tails
- Test on other signals / observables
- Timing studies on GPUs (need to operate real-time in scouting system)

Extra slides

Gradient update: STE vs. Rotation Trick

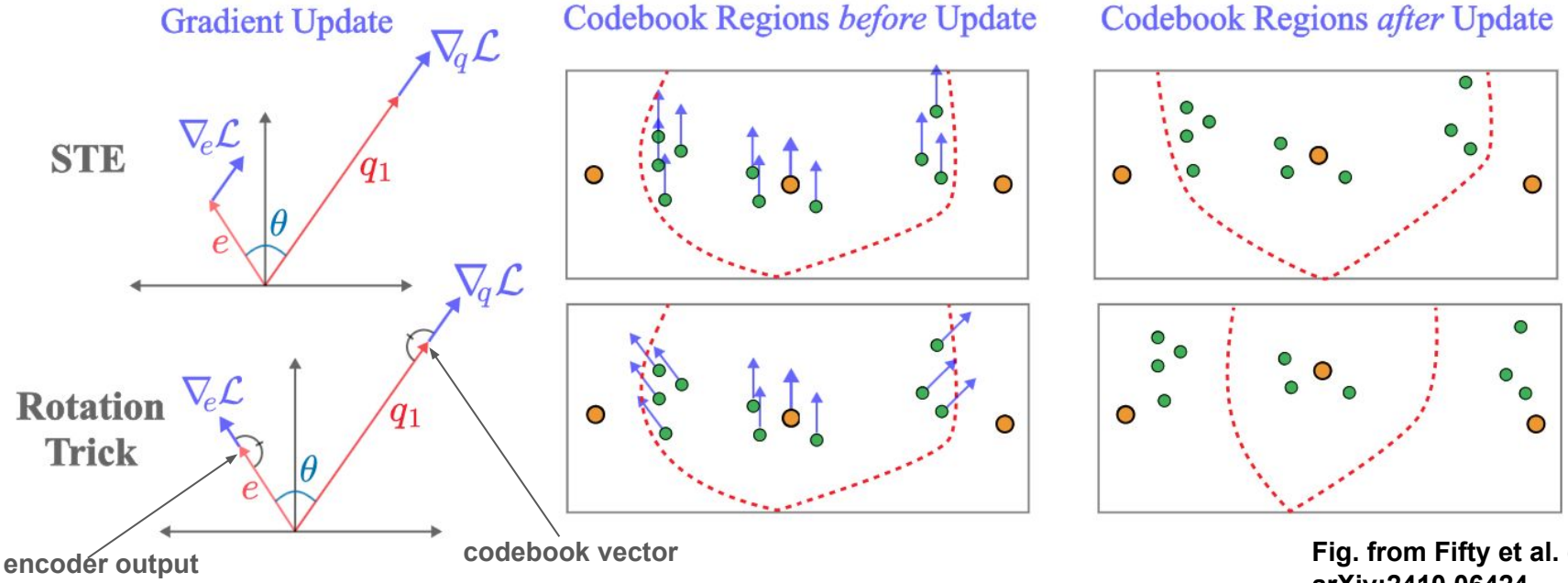


Fig. from Fifty et al. (2024), arXiv:2410.06424

- Using STE $\nabla_e \mathcal{L}$ equals $\nabla_{q_1} \mathcal{L}$ during backpropagation
- Using rotation trick angle between q_1 and $\nabla_{q_1} \mathcal{L}$ equals angle between e and $\nabla_e \mathcal{L}$
  - Pushes encoder outputs in the same codebook region in different directions, may lead to improved codebook utilization and reconstruction

Loss function

- FSQ uses the MSE as a loss function: $\mathcal{L}_{\text{FSQ}} = \|x - \hat{x}\|_2^2$ (x input features, $\hat{x}$ reconstructed features)
- VQ-VAEs additionally learns the codebook, so we add a term pushing the codebook entries e_k towards the encoder output $z(x)$, commitment term prevents encoder output from diverging

sg: stop-gradient, gradients don't flow through this term

$$\mathcal{L}_{\text{VQ}} = \underbrace{\|x - \hat{x}\|_2^2}_{\text{reconstruction}} + \alpha\underbrace{\|\text{sg}[z(x)] - e_k\|_2^2}_{\text{codebook update}} + \beta\underbrace{\|z(x) - \text{sg}[e_k]\|_2^2}_{\text{commitment}}, \quad k = \arg\min_j \|z(x) - e_j\|_2$$

This selects the codebook entry closest to the encoder output

- PHAEDRA uses the same loss as FSQ/VQ, but added per-branch

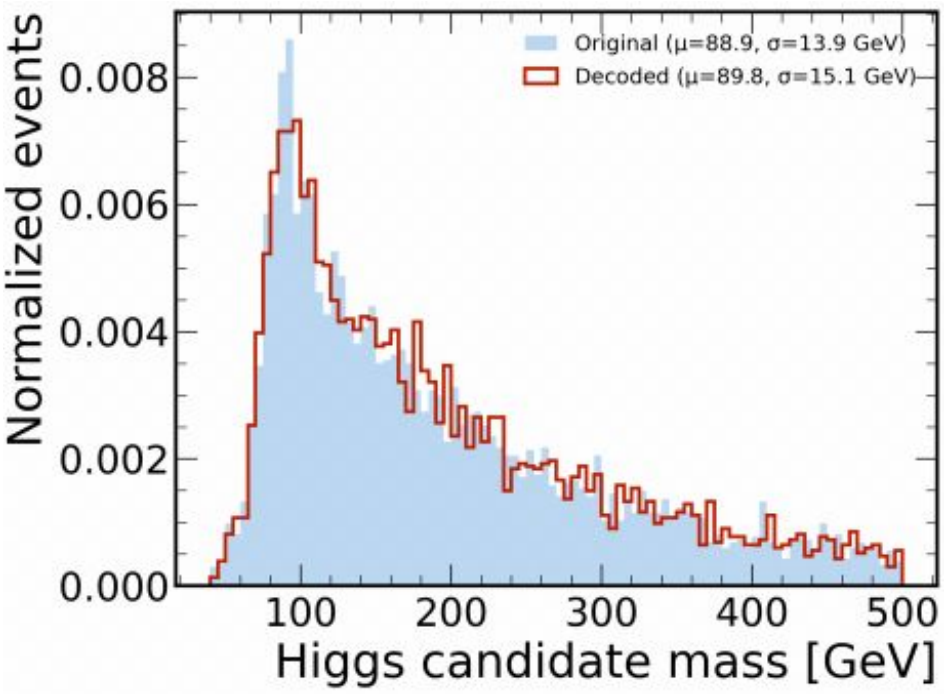
Result: Reached compression (full event)

ETH zürich

Train: $t\bar{t}$
Model: VQ-STE, 4096 codes
Full event

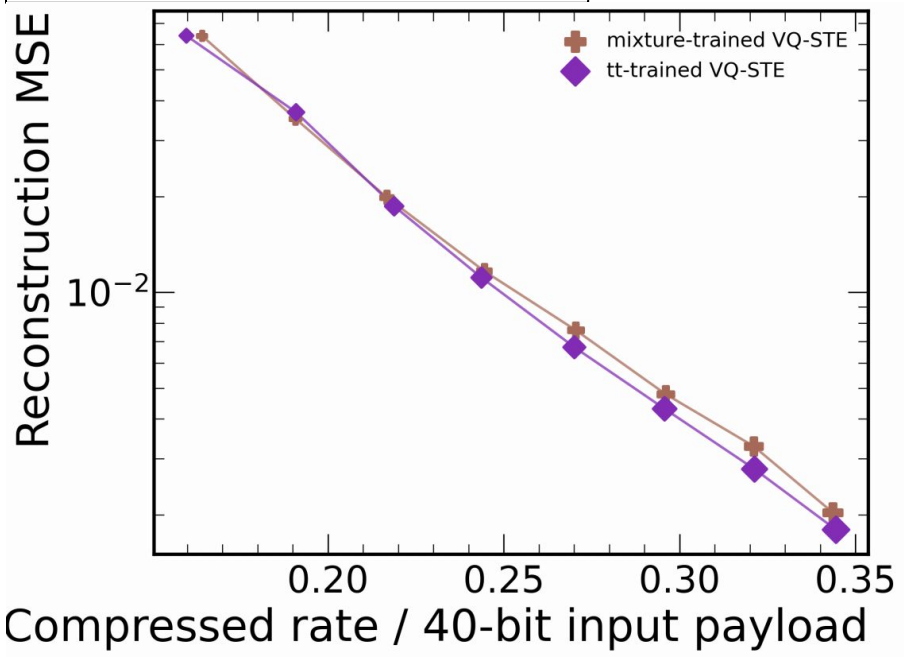
Sample	$\langle N \rangle$/evt	Raw	Standalone	EDM	nanoAOD
		B (%)	B (%)	B (%)	B (%)
Minimum bias, plain	500.0	2504.8 (100.0)	1796.0 (71.7)	1674.1 (66.8)	1618.8 (64.6)
Minimum bias, tokenized	500.0	754.5 (30.1)	754.6 (30.1)	816.8 (32.6)	802.5 (32.0)
$H \to b\bar{b}$, plain	500.0	2504.7 (100.0)	1810.1 (72.3)	1689.3 (67.4)	1632.0 (65.2)
$H \to b\bar{b}$, tokenized	500.0	754.5 (30.1)	754.6 (30.1)	817.1 (32.6)	802.9 (32.1)
$t\bar{t}$, plain	500.0	2504.7 (100.0)	1825.5 (72.9)	1702.8 (68.0)	1645.2 (65.7)
$t\bar{t}$, tokenized	500.0	754.5 (30.1)	754.6 (30.1)	817.1 (32.6)	802.8 (32.1)

Higgs candidate mass, extended histogram



Train: $t\bar{t}$
Model: VQ-STE, 4096 codes
Pileup-subtracted

Test = tt
Model = VQ-STE
Pileup-subtracted



Codebook size to bitwidth

Codebook size	bits	relative size (vs. 14 (p_t) + 12 (η) + 11 (ϕ) + 3 (PID) = 40 bits)
128	7	0.18
256	8	0.20
512	9	0.23
1024	10	0.25
2048	11	0.28
4096	12	0.30
8192	13	0.33
16384	14	0.35

SM Mixture composition

Family	Included processes	Fraction per subprocess
QCD	`QCD_HT50tobb`, `QCD_HT50toInf`	12.5%
$t\bar{t}$	Hadronic, leptonic, semileptonic	8.33%
V + jets	`WJetsToLNu`, `WJetsToQQ`, `DYJetsToLL`, `ZJetsTobb`, `ZJetsTocc`, `ZJetsToQQ`, `ZJetsTovv`	3.57%
Diboson (VV)	WW, WZ, and ZZ, each with hadronic, leptonic, and semileptonic samples	2.78%

PF-level bit assignments

Common info for all candidate types				
Bits	Field	Bits	Format	Notes
13-0	p_T	14	unsigned int, LSB = 0.25 GeV	ap_ufixed(14,12) in GeV
25-14	η	12	signed int, LSB = $\pi/720 = \frac{1}{4}$ deg	
36-26	ϕ	11	signed int, LSB = $\pi/720 = \frac{1}{4}$ deg	
39-37	PID	3	see below	

Charged candidates specific info				
Bits	Field	Bits	Format	Notes
49-40	z_0	10	signed int, LSB = 0.5 mm	
57-50	d_{xy}	8	signed int	to be defined
60-58	quality	3	tbd	tbd
63-61	*unassigned*	3		

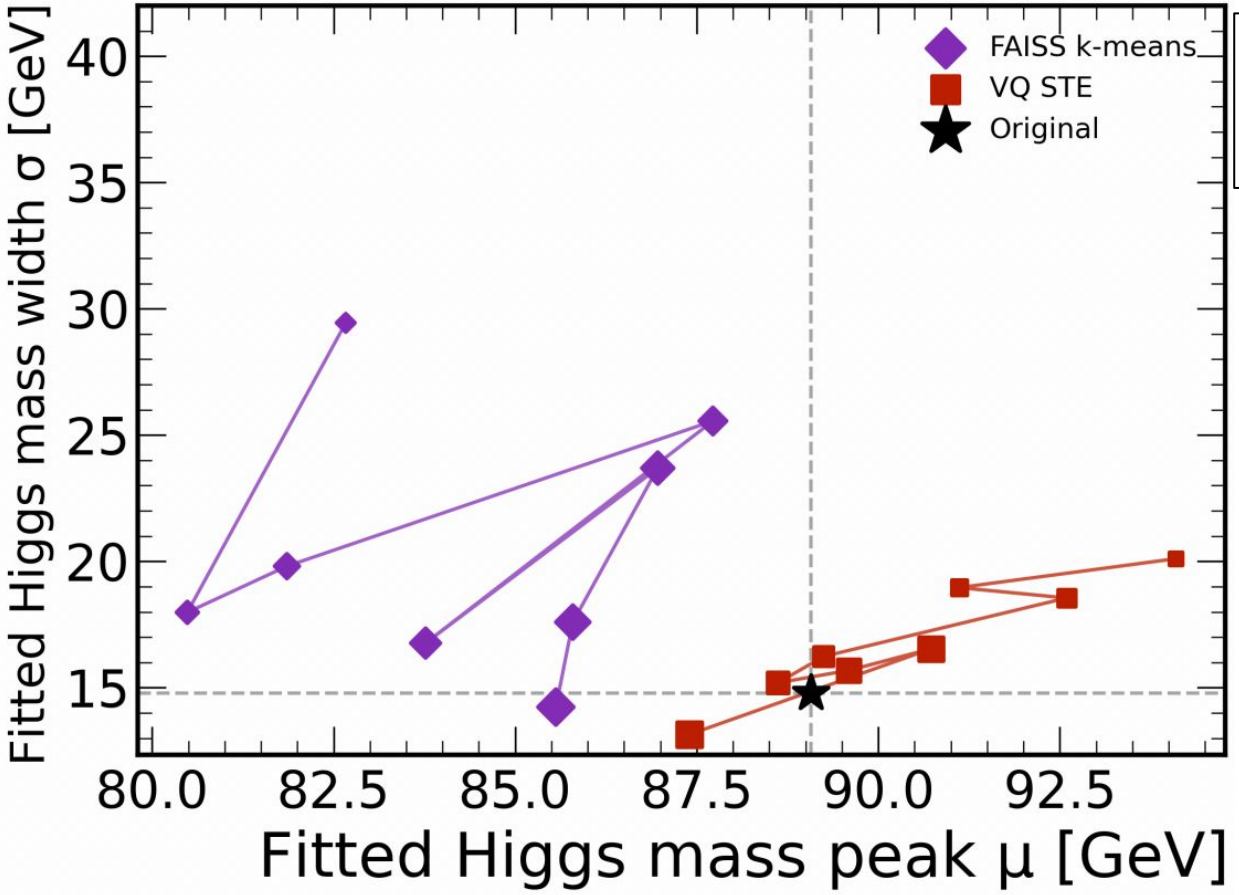
Neutral candidates specific info				
Bits	Field	Bits	Format	Notes
49-40	w_{Puppi}	10	unsigned int	to be defined
55-50	e/γ id	6	unsigned int	tdb
63-56	*unassigned*	8		

PID definitions

Val	Binary	Particle		PDG ID
0	000	h^0	neutral hadron	130 (K_L^0)
1	001	γ	photon	22
2	010	h^-	hadron of charge −1	−211 (π^-)
3	011	h^+	hadron of charge +1	+211 (π^+)
4	100	e^-	electron	+11
5	101	e^+	positron	−11
6	110	μ^-	muon	+13
7	111	μ^+	anti-muon	−13

Downstream physics performance

ETH zürich

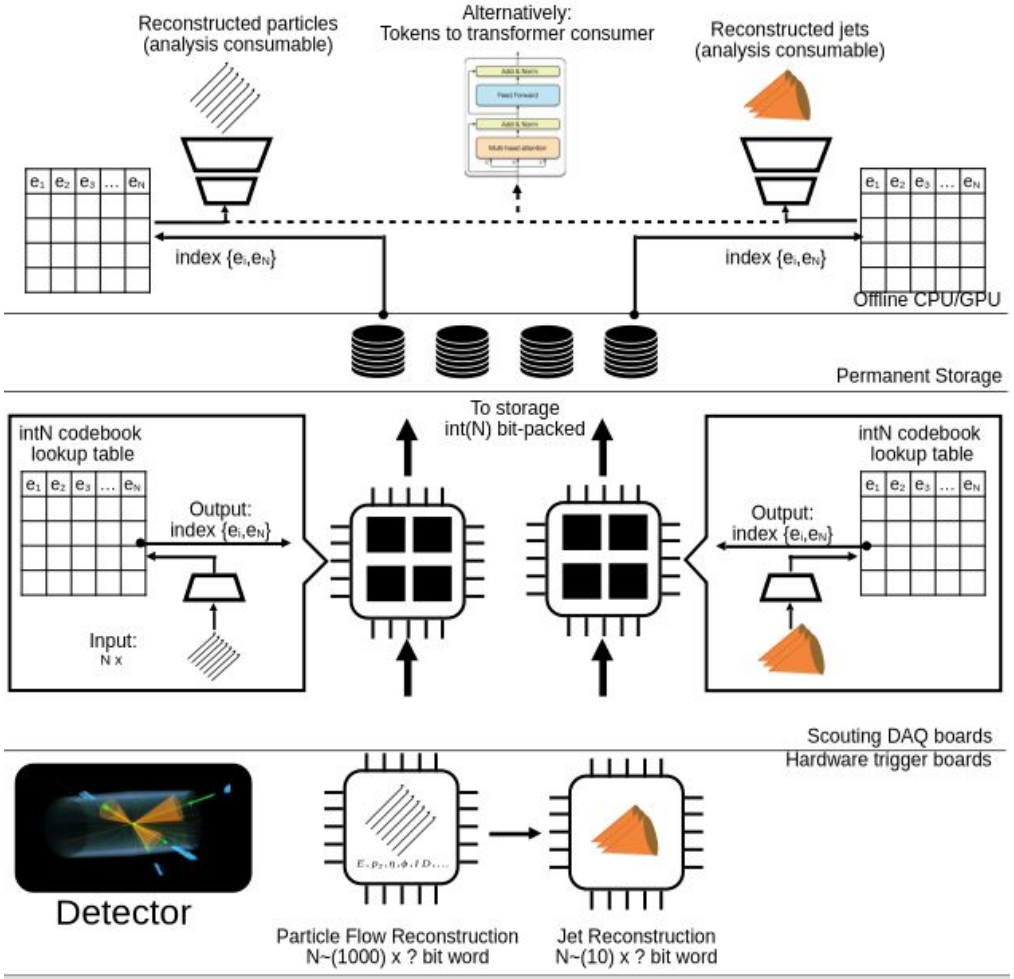


Train: tt
Test: ggHbb
Crystal ball fit

Architectural details

ETH zürich

- Transformer encoder/decoder, 4 blocks, 8 attention heads, pre-norm placement, GELU activations, MLP expansion factor 4, attention across all particles in event
- ~1.2M params
- FSQ: 3-dim latent space with 5-35 levels per dimensions
- VQ-VAE: Codebook sizes varying from 128-16384
- Split quantizer: 5-dim latent space with split morphology [128] (gain), [2,3,3,5] (shape)
- AdamW optimizer, learning rate schedule (5e-4 -> 1e-3 -> 3e-4), weight decay 10e-2


Reconstructed particles
(analysis consumable)
Alternatively:
Tokens to transformer consumer
Reconstructed jets
(analysis consumable)
index {e1,eN}
index {e1,eN}
Offline CPU/GPU
Permanent Storage
To storage
int(N) bit-packed
intN codebook
lookup table
Output:
index {e1,eN}
Input:
N x
intN codebook
lookup table
Output:
index {e1,eN}
Scouting DAQ boards
Hardware trigger boards
Detector
Particle Flow Reconstruction
N~(1000) x ? bit word
Jet Reconstruction
N~(10) x ? bit word