How many bits is a jet? Tokenizing LHC data for foundation models

Jeff Krupa, Sam Klein, Matthias Vigl, Michael Kagan

TREASURE Meeting

4 November 2026

SLAC NATIONAL ACCELERATOR LABORATORY

TREASURE

Showing results from 2610.13876 and 2609.31862v1

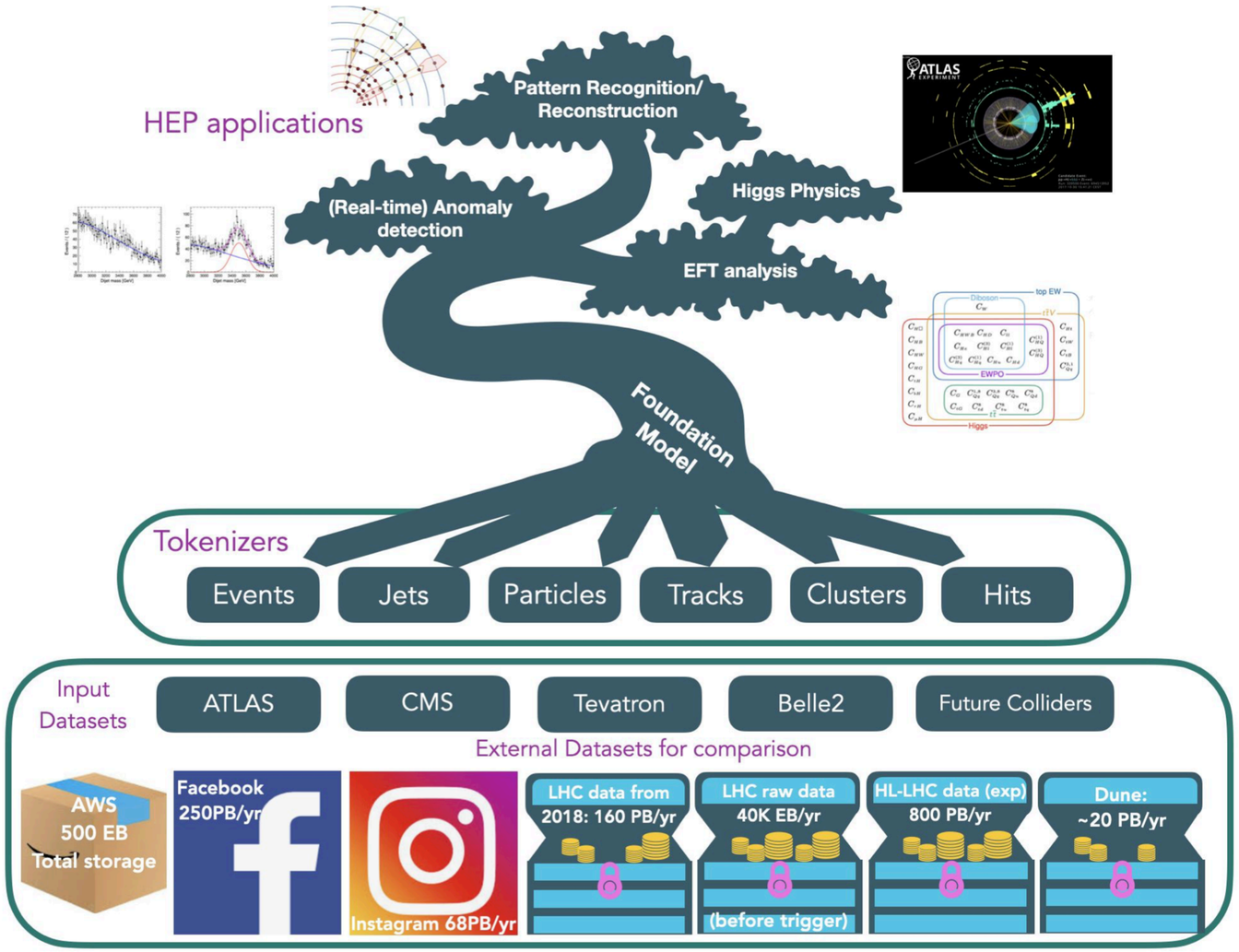
Precision Tokenization for Multi-Modal Data in High Energy Physics

Jeffrey Krupa*[a], Samuel Klein*[a], Nilotpal Kakati*[b], Daniel Murnane[c], Baran Hashemi[d], Eilam Gross[b], Lukas Heinrich[c], Michael Kagan[a]

Prompting Particle Physics: Tokenized Multi-modal Foundation Models for Combinatorially Many Tasks

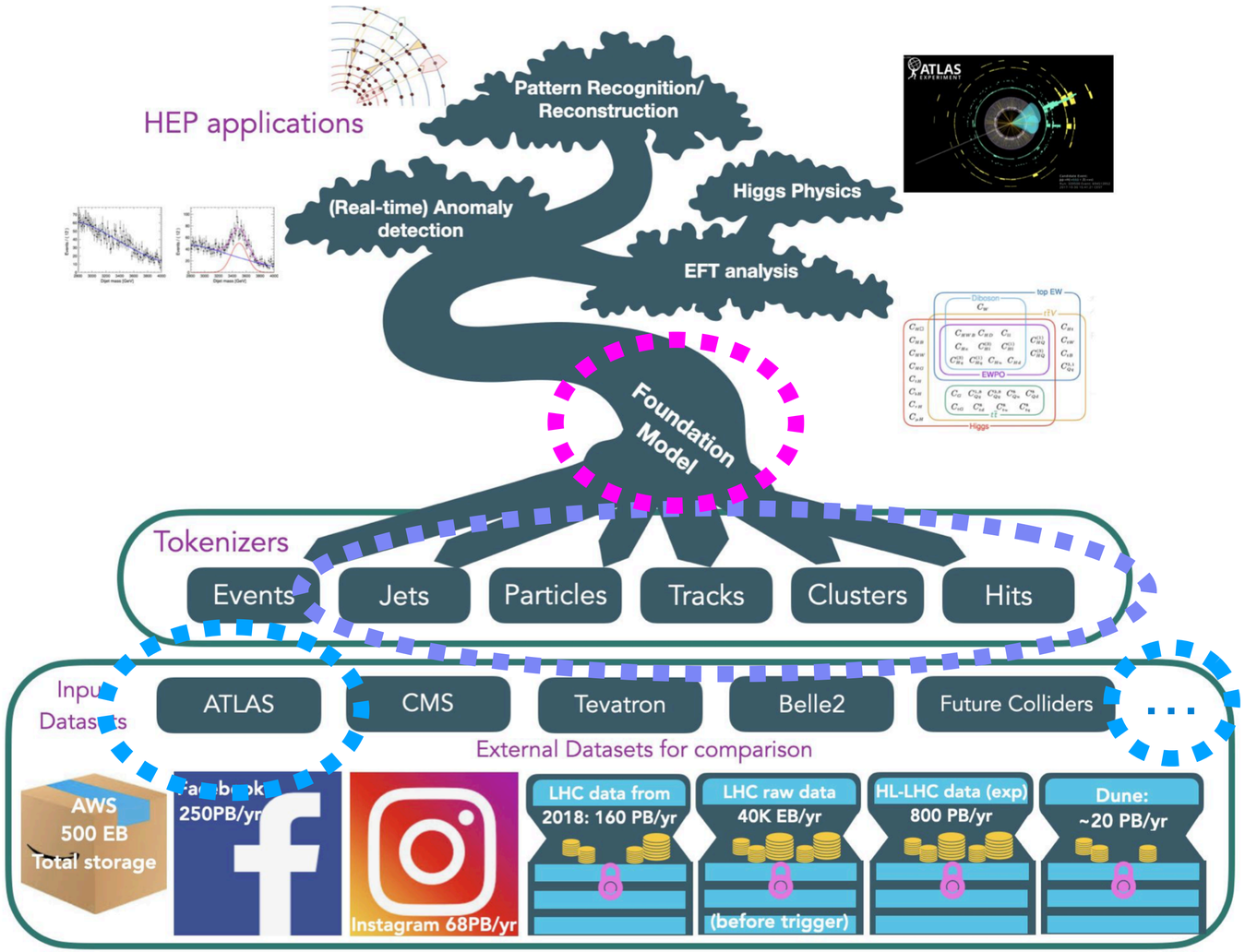
Nilotpal Kakati[1,*], Daniel Murnane[2,3,*], Baran Hashemi[4], Samuel Klein[5], Jeffrey Krupa[5], Eilam Gross[1], Lukas Heinrich[6], Michael Kagan[5]

Building Towards Foundation Models





Building Towards Foundation Models



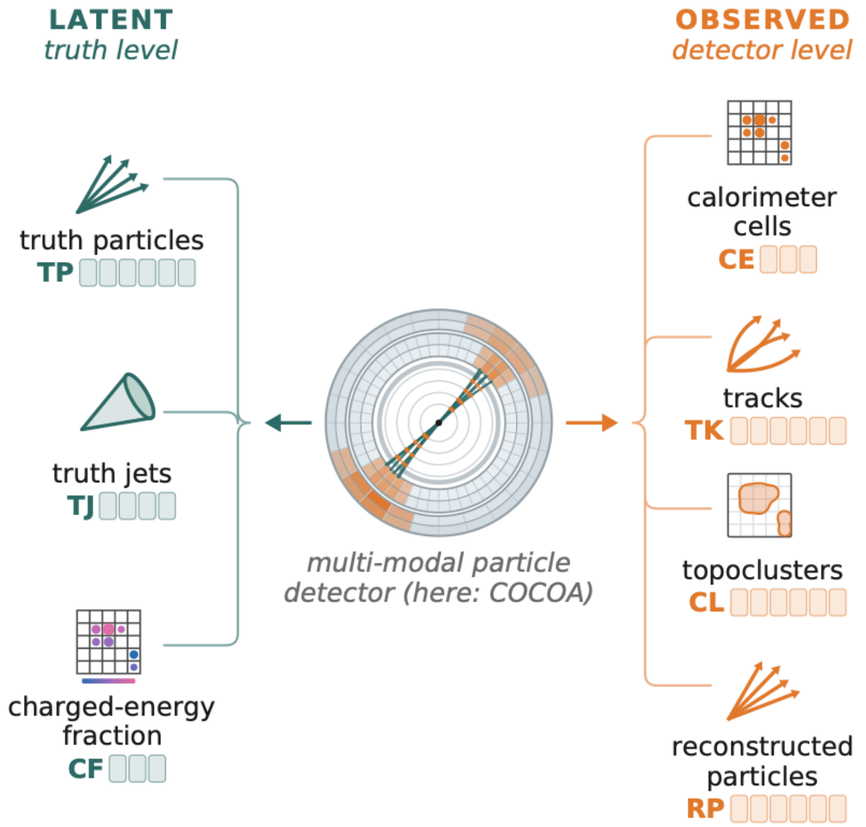


This talk presents our latest results on **multi-modal tokenization** on **ATLAS+open data** and progress on HEP **FMs**

2609.31862v1

Is Prompting All We Need?

FM mapping between modalities



Energy deposits

Paths of charged particles

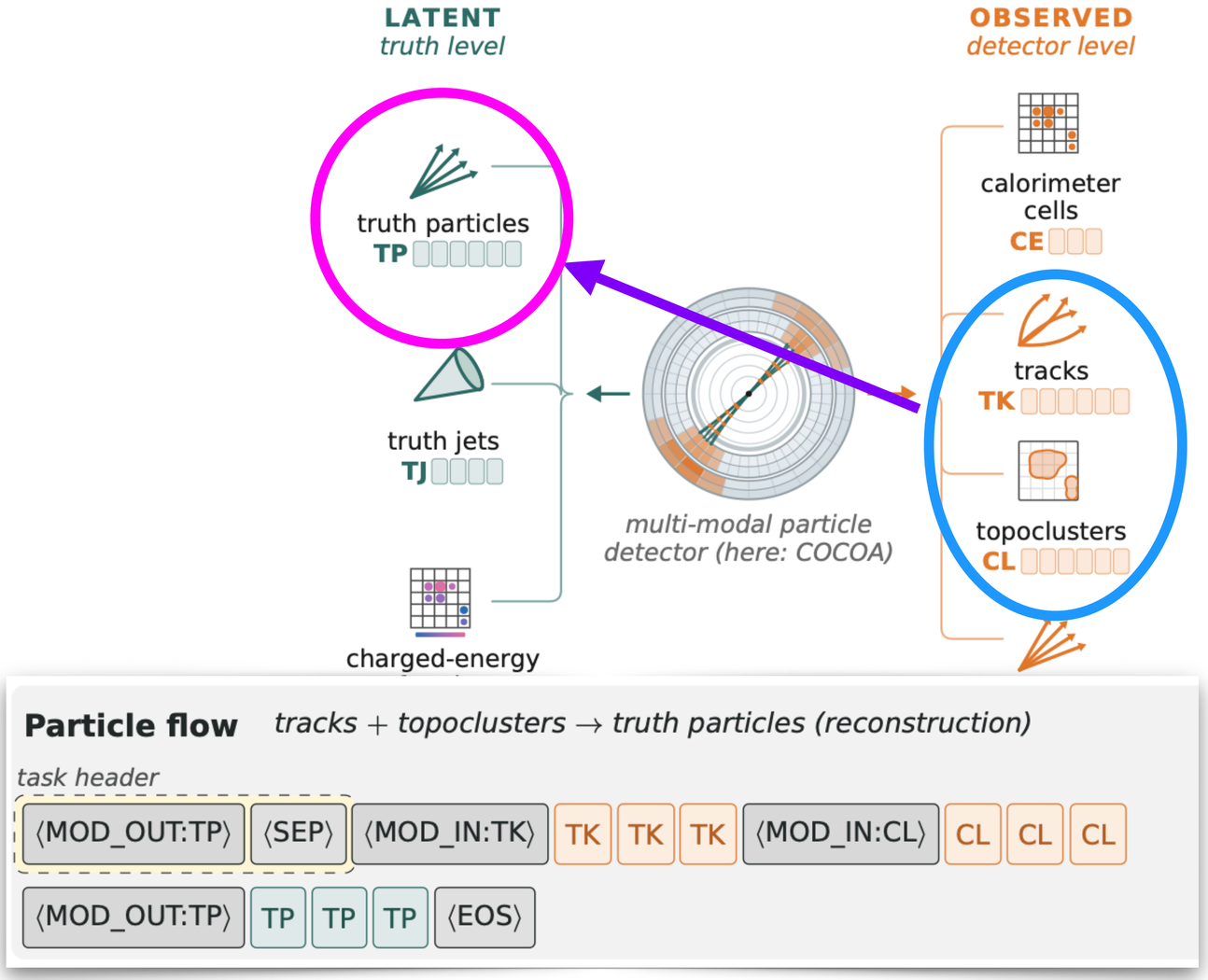
Groups of neighboring cells

Combination of tracks+topoclusters

2609.31862v1

Is Prompting All We Need?

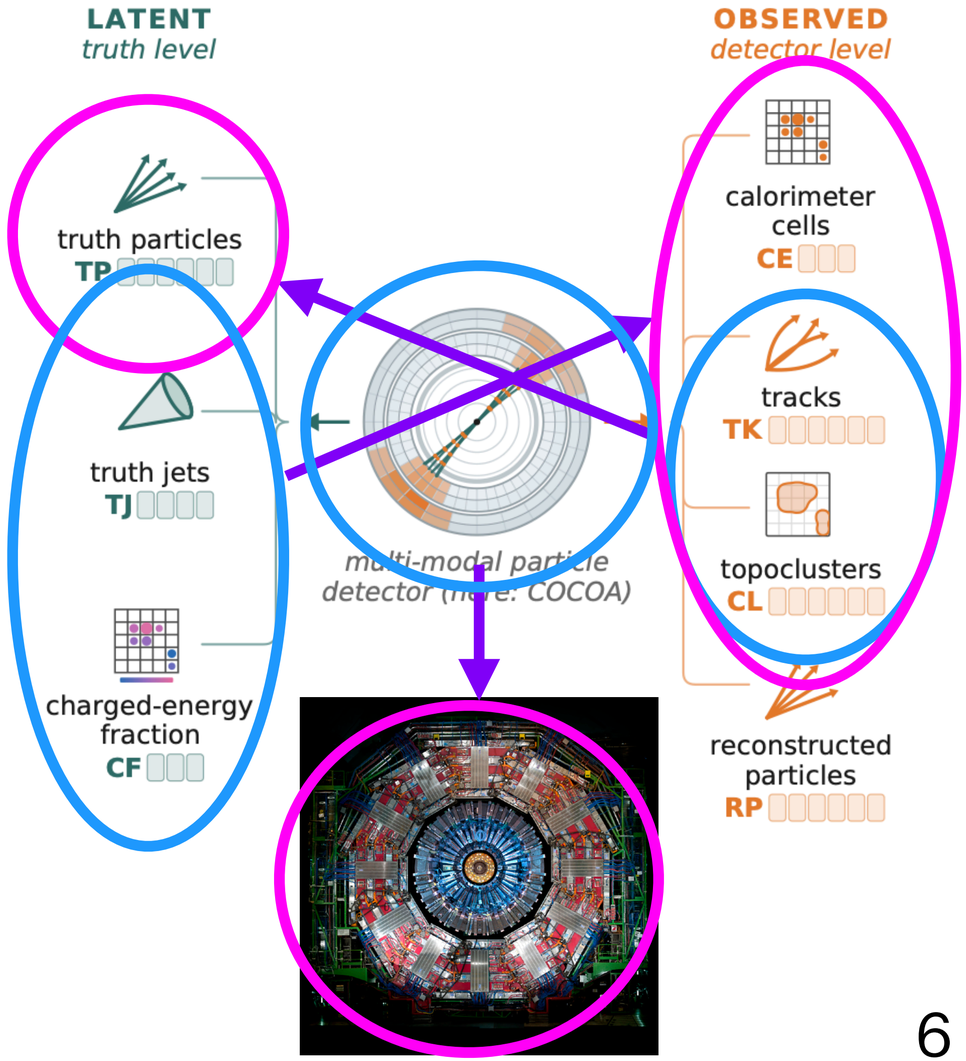
FM mapping between modalities



This is the particle flow prompt: "**given track tokens and topoclusters tokens**, predict truth particles"
one of 1302 possible directions!

Why Tokenize?

Tokens turn every modality into the same kind of object



Any modality in, any out using a shared vocabulary

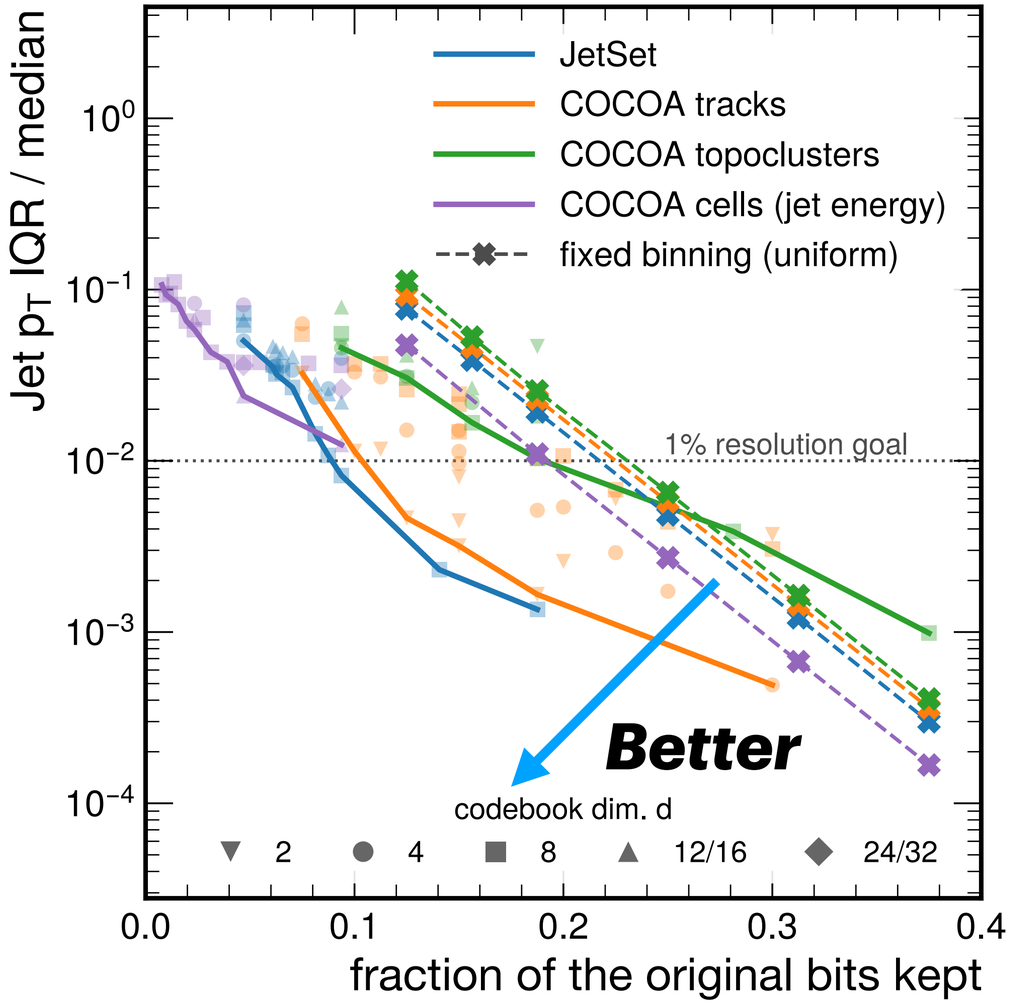
Loss reduced to **cross entropy for all tasks**

Compact representation: at what cost?

Use **language model machinery** off the shelf

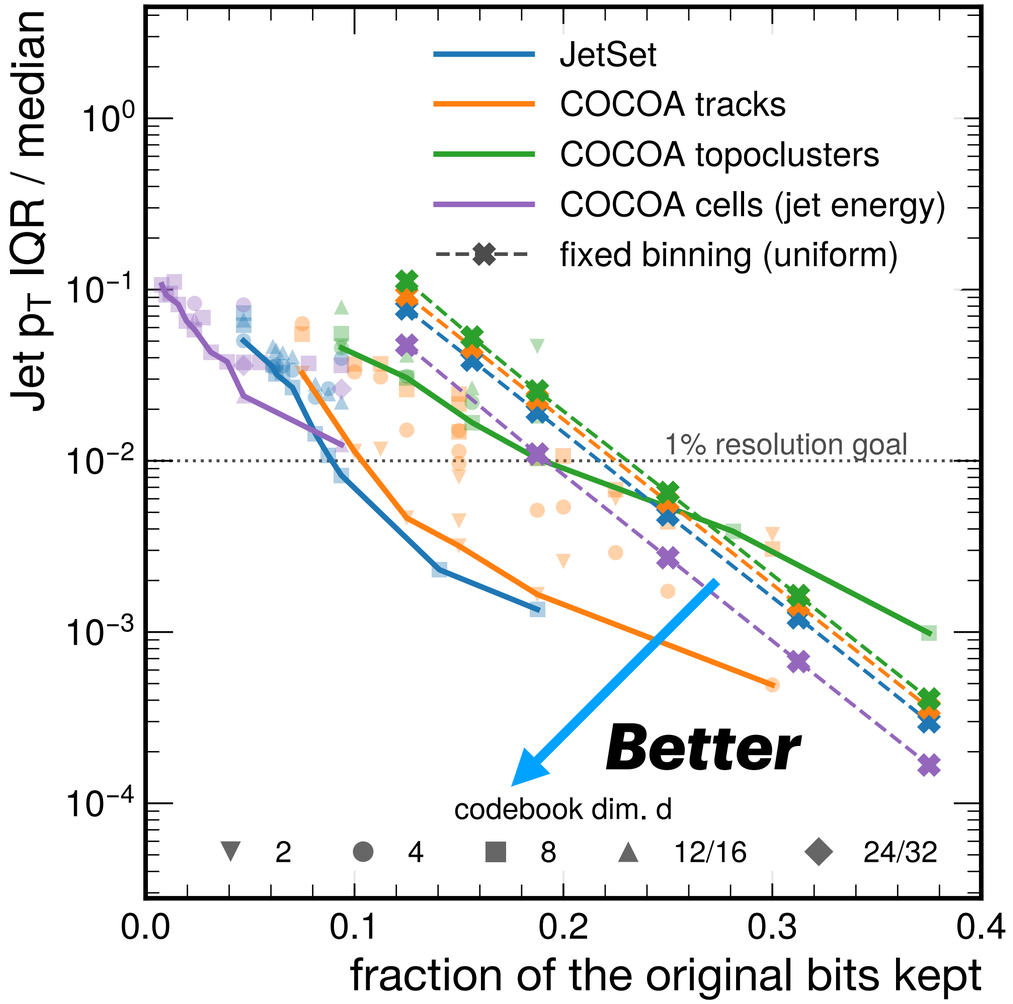
2610.13876

Understanding the Tradeoffs



In brief: learned tokens achieve 1% pT resolution at ***4-11x compression*** *across 4 modalities,* ***using up to 2.4x fewer bits than fixed binning***

Understanding the Tradeoffs



In brief: learned tokens achieve 1% pT resolution at ***4-11x compression*** *across 4 modalities,* ***using up to 2.4x fewer bits than fixed binning***

Also: we evaluate tokenization downstream by evaluating tokens on regression, particle flow, and flavor tagging

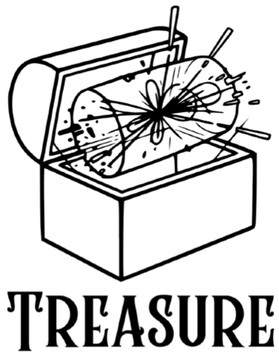
Landscape

Early FM efforts like masked particle modeling, followed by OmniJet-α, tokenized particles (mapping each input to a vector) with difficulty controlling fidelity

Continuous targets can outperform tokens on masking tasks *at current scale*. How much physics do tokens actually keep, and at what compression?

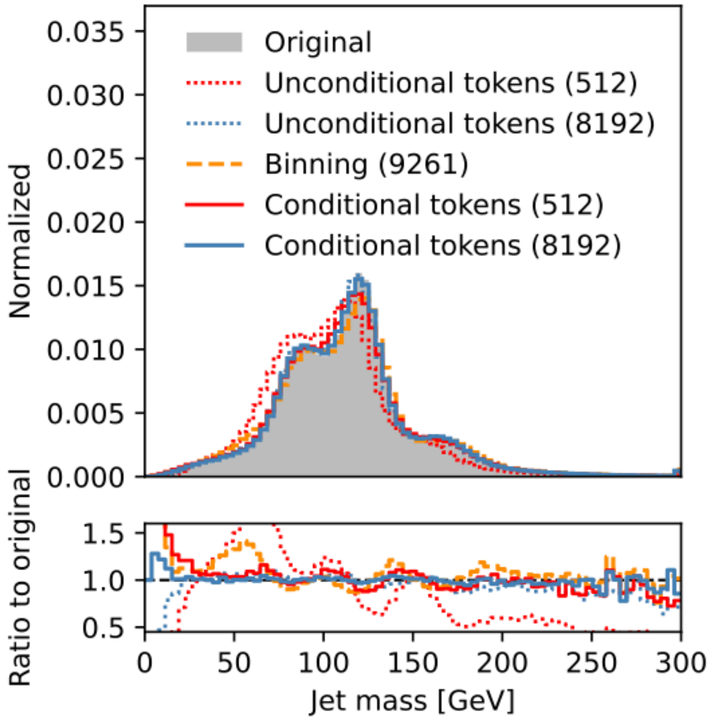
Since then tokenization has been expanded to showers, tracking, and high-granularity calorimeters

TREASURE aims to systematize this



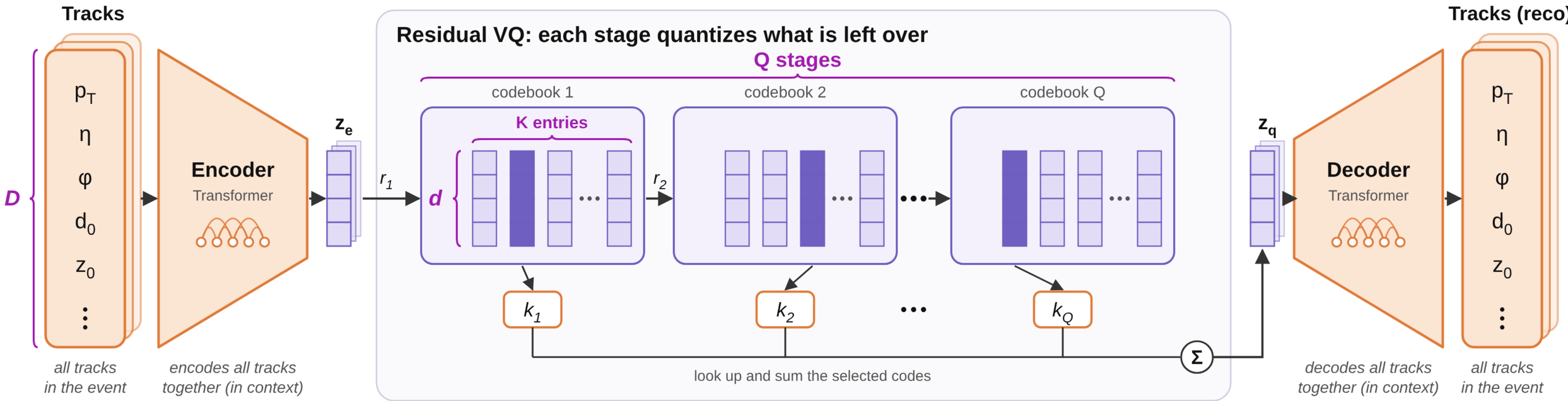
Fidelity depends on codebook size + design

OmniJet-α



Residual VQ

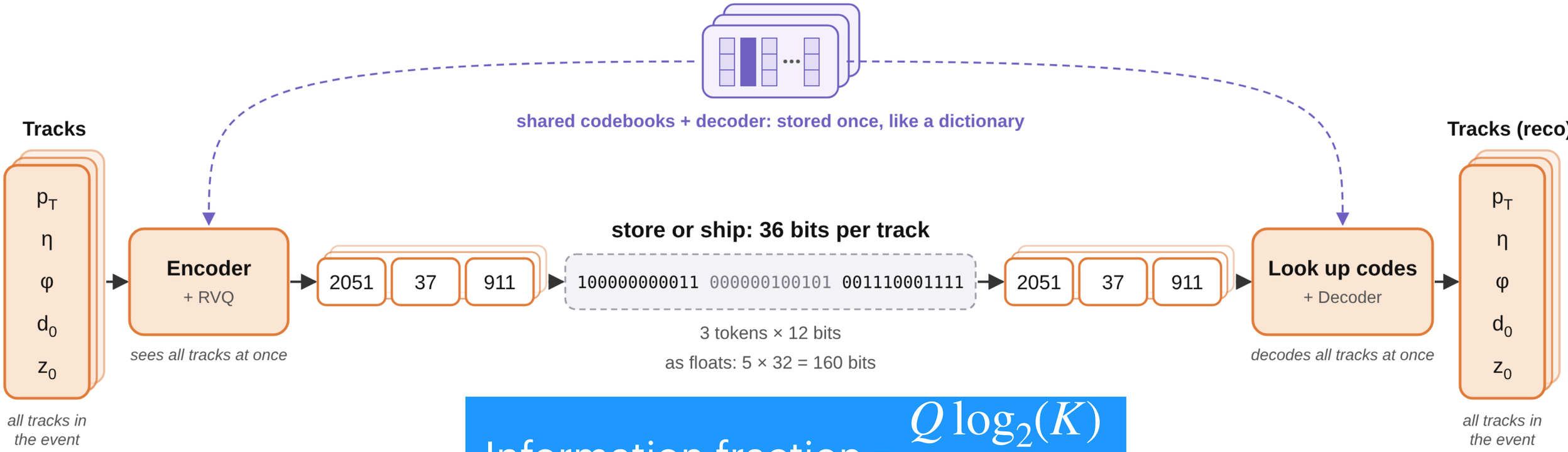
Q quantizer stages, K codebook entries per quantizer, codebook dimension d



Early observation: transformers "in context" beat MLP "no context" models as encoders.

Compression

1. Continuous: D floats = $32D$ bits
2. Q indices of K codes = $Q \log_2(K)$ bits



$$\text{Information fraction} = \frac{Q \log_2(K)}{32D}$$

Here Q=3, K=2^{12}, D=5 → info frac = 0.22

Datasets

JetSet: fullsim ATLAS Run 3 official open data

ATLAS $t\bar{t}$ simulation for ML-based jet flavour tagging (JetSet)

ATLAS collaboration

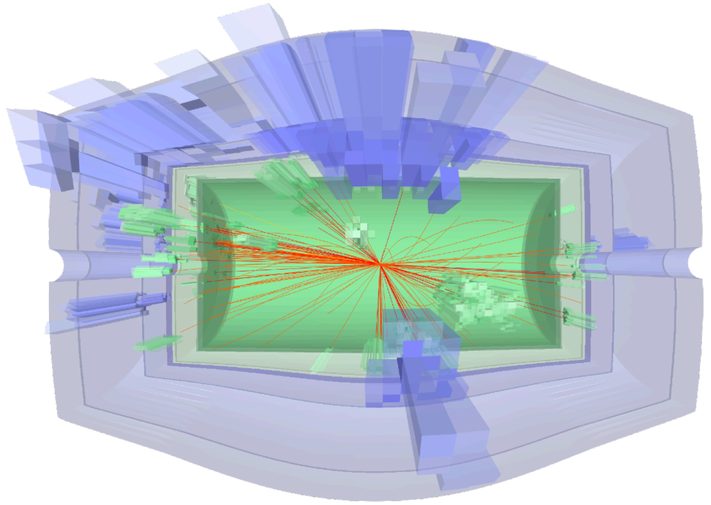
Cite as: ATLAS collaboration (2025). ATLAS $t\bar{t}$ simulation for ML-based jet flavour tagging (JetSet). CERN Open Data Portal.

There are 7 publications referring to these data

Dataset | Derived | Simulated | Datascience | ATLAS | 13.6TeV | pp | CERN-LHC

Modality	Dimension D
Tracks	20

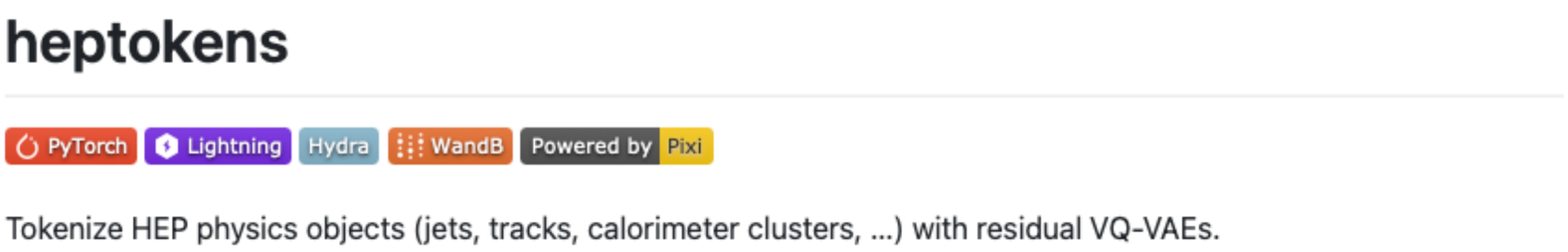
COCOA: multi-modal open data



Modality	Dimension D
Cells^	64
Tracks*	5
Topoclusters	8
Truth particles	5

^we tokenize cells as images (backup)
**fastsim*

heptokens Tokenization Library



heptokens: an experiment-agnostic tokenization library used across experiments in the context of the TREASURE project (ATLAS, CMS, Belle II)

pip installable: pip install heptokens

Data loaders, models, callbacks, utilities

Emphasis on reproducibility and scalability using industry standards Pixi/ Hydra/Snakemake

Metrics

Does the encoder-RVQ-decoder trip preserve the data?

1. **Jet reconstruction**: Can jets from tokenized tracks reproduce the jet from untokenized tracks? Goal = sub-1%
2. **Feature reconstruction**: normalized Mean Absolute Error (MAE) over inputs

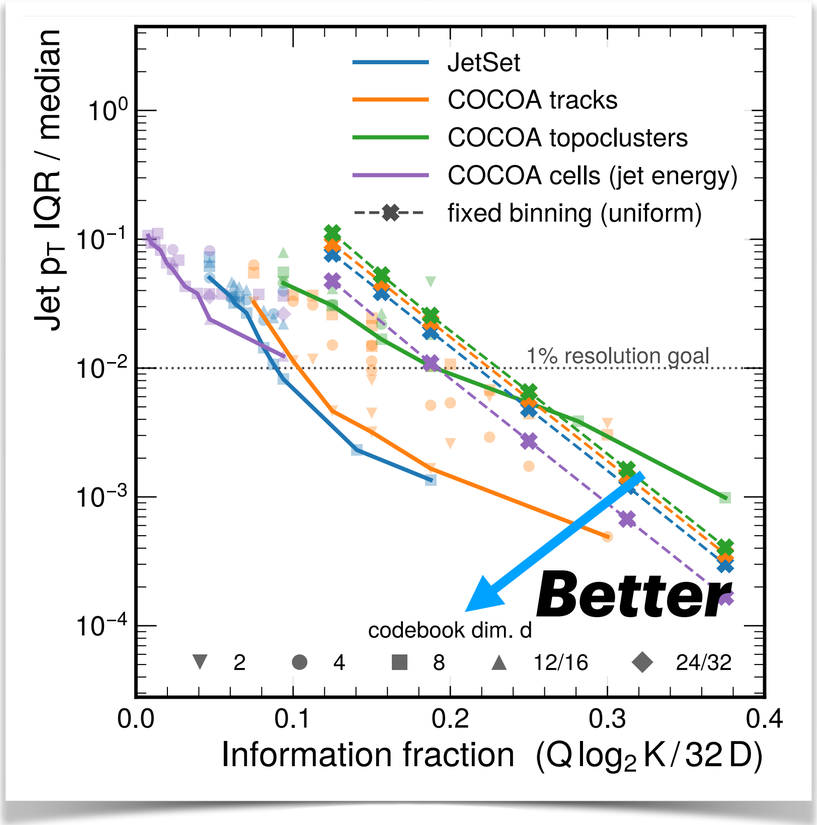
How effective are the tokenized representations in downstream tasks?

3. Jet pT regression
4. Particle flow
5. Flavor tagging

Results: jet pT

Mapping the Pareto frontier of HEP compression through tokenization

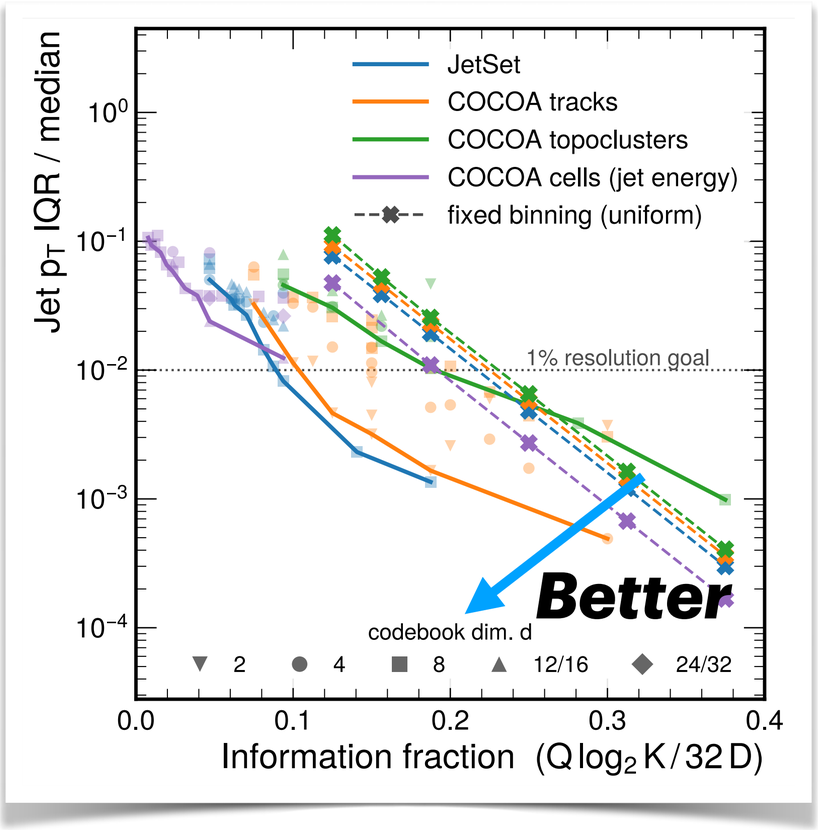
Learned approach is up to 2.4x more efficient than simple fixed binning



Results: jet pT

Mapping the Pareto frontier of HEP compression through tokenization

Learned approach is up to 2.4x more efficient than simple fixed binning



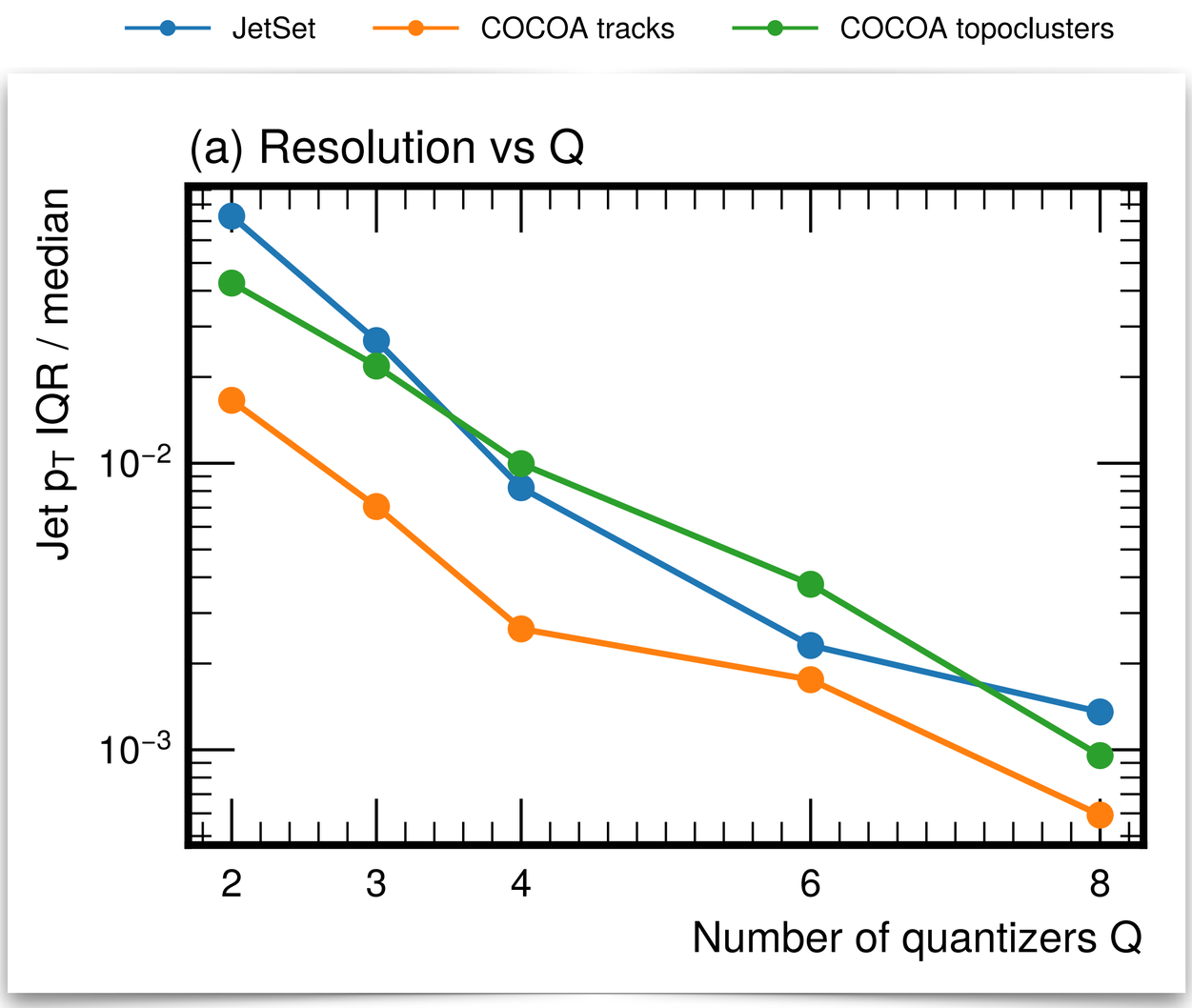
To achieve 1% jet resolution with good MAE, compression varies between ~4-11x

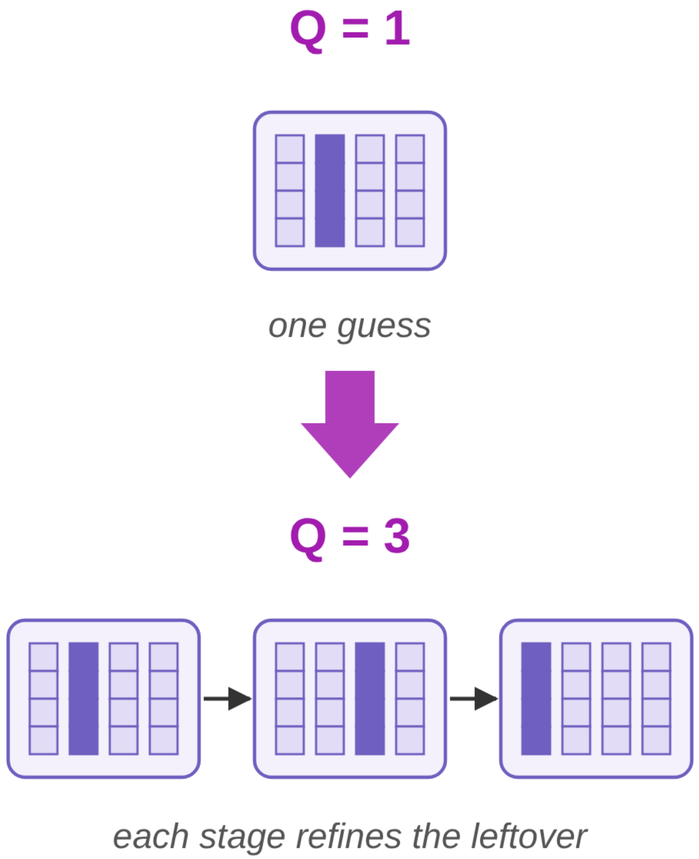
Modality	Bits	Compr.*	Res.	Util.
Jet constituents (JetSet)	60	10.7×	0.82%	77%
Tracks (COCOA)	36	4.4×	0.68%	71%
Topoclusters (COCOA)	72	3.6×	0.39%	99%
Calorimeter cells (COCOA)	192	10.7×	1.24%†	88%

* Compression = 1/info fraction

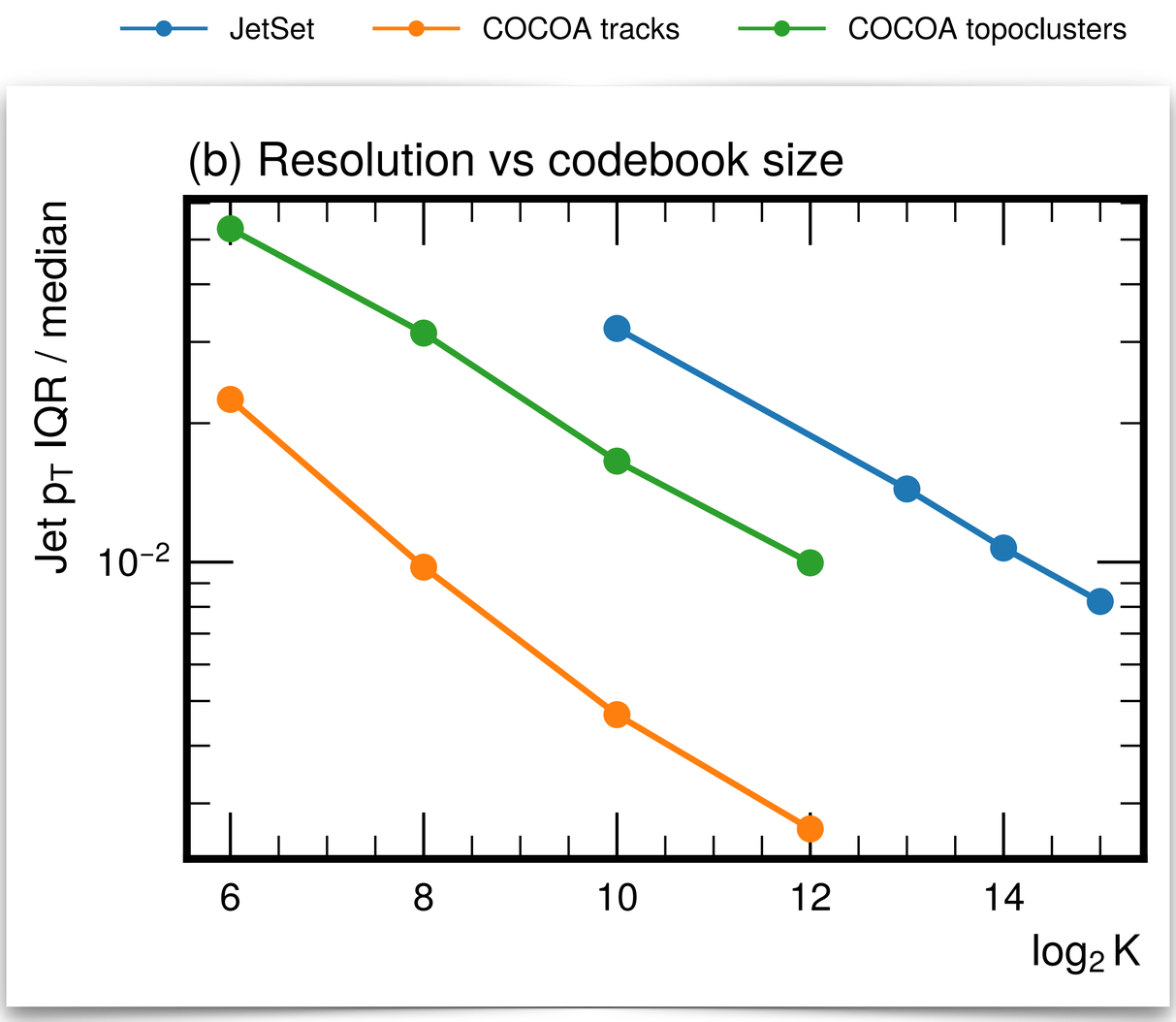
† best resolution achieved

Design Rules





Design Rules

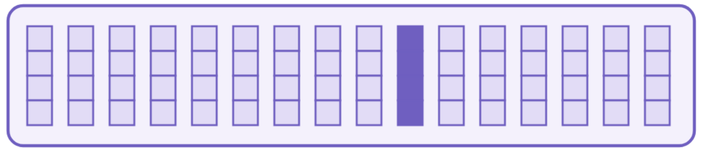


small K



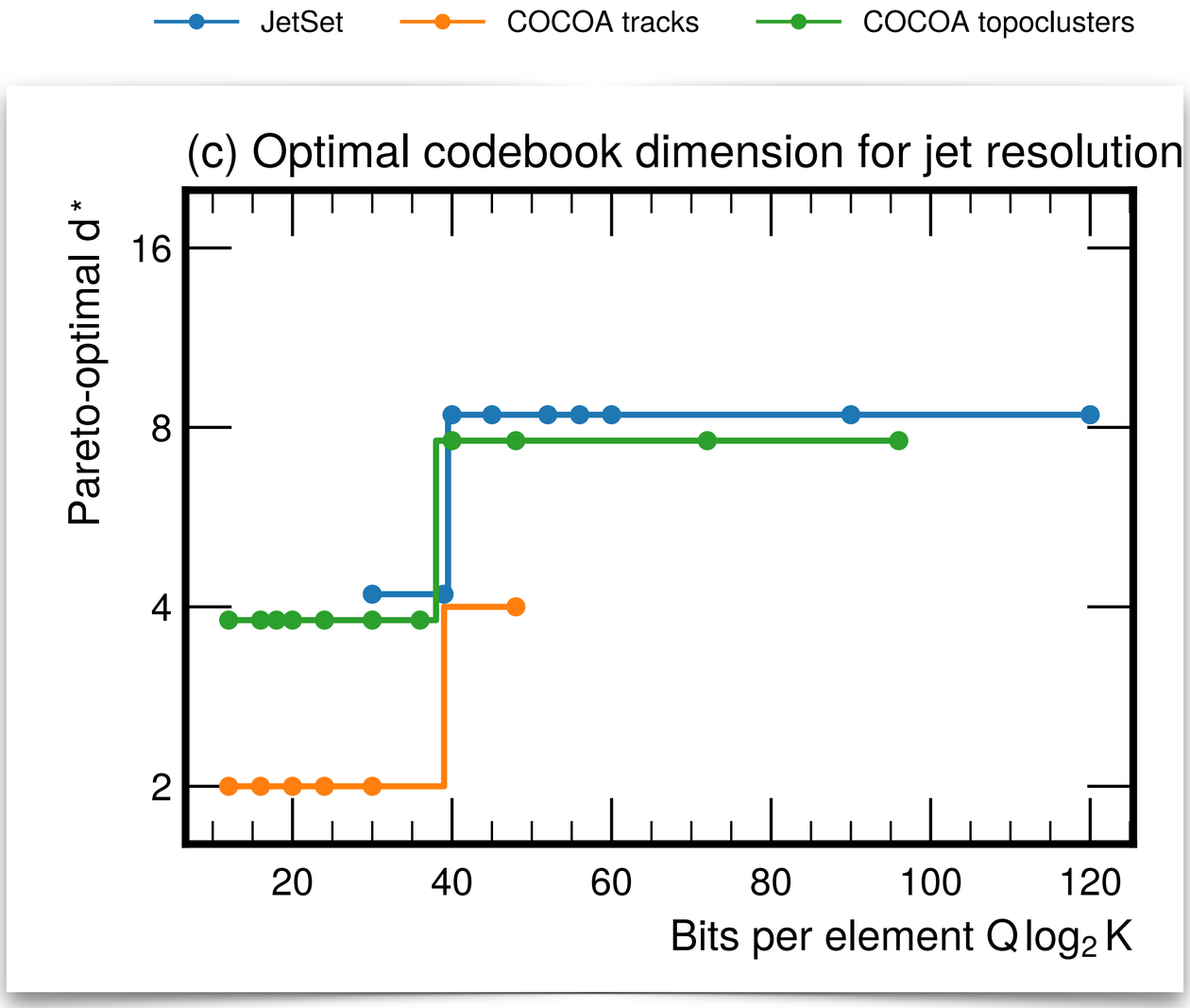
few codes to choose from

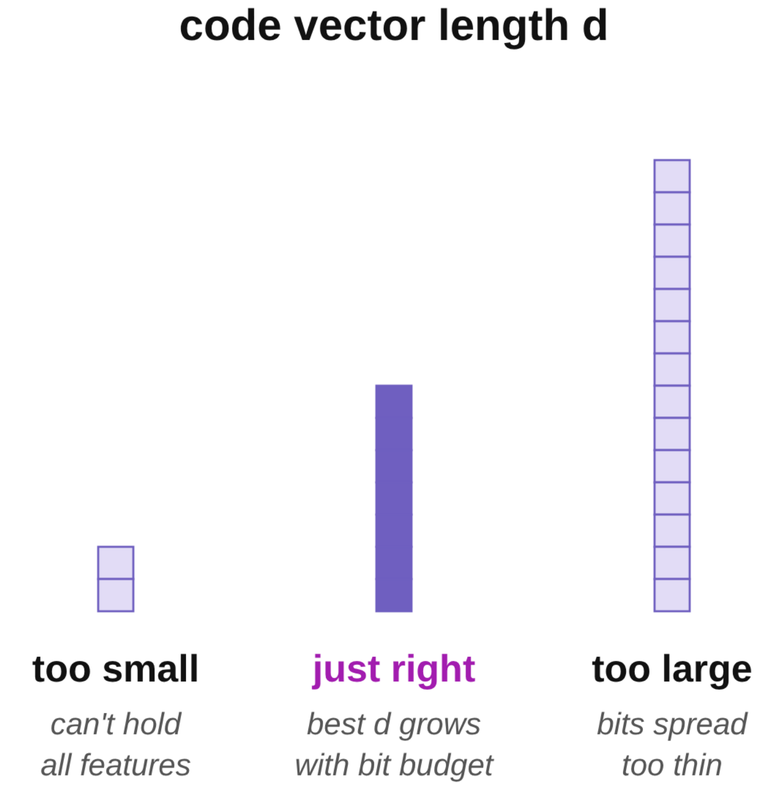
large K



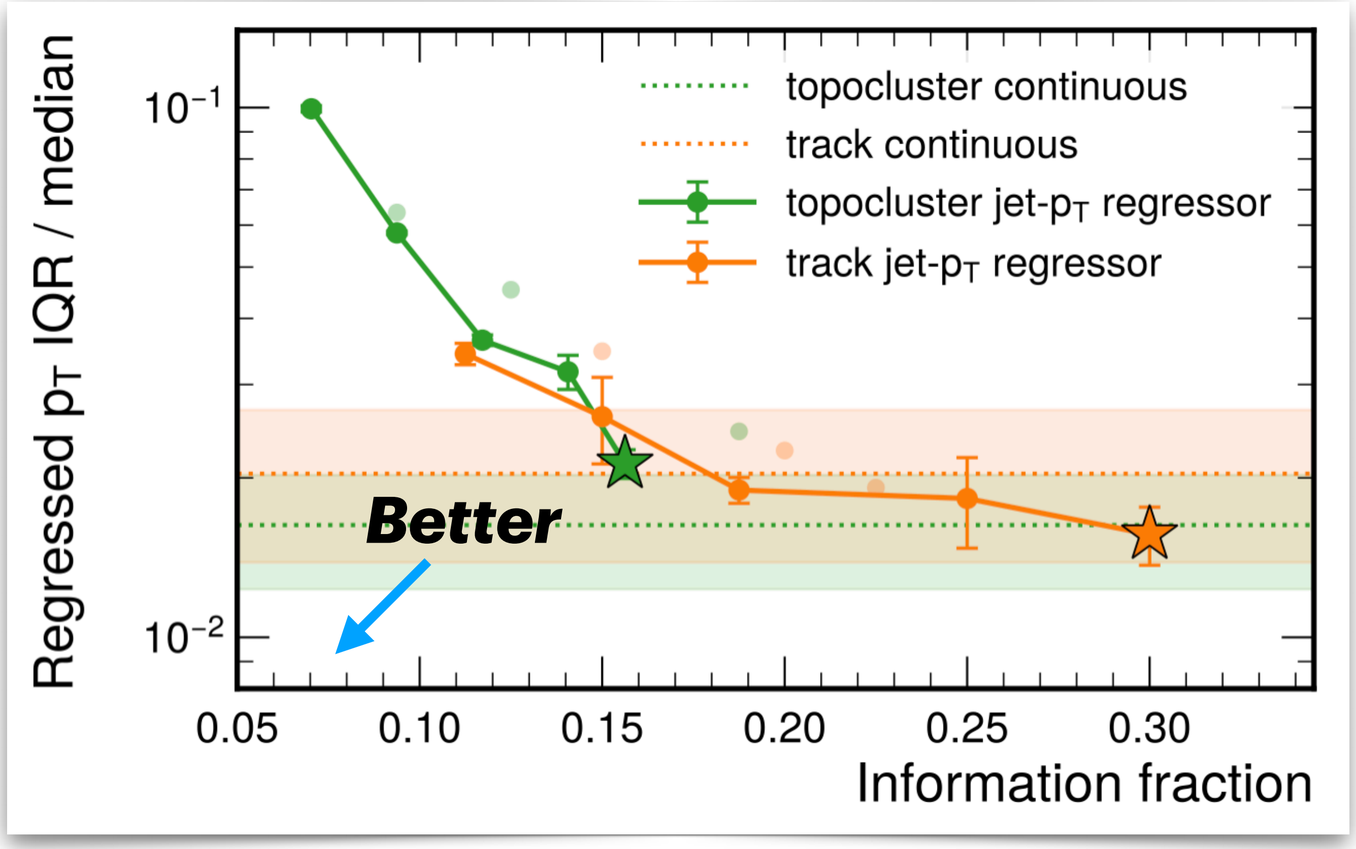
finer grid

Design Rules



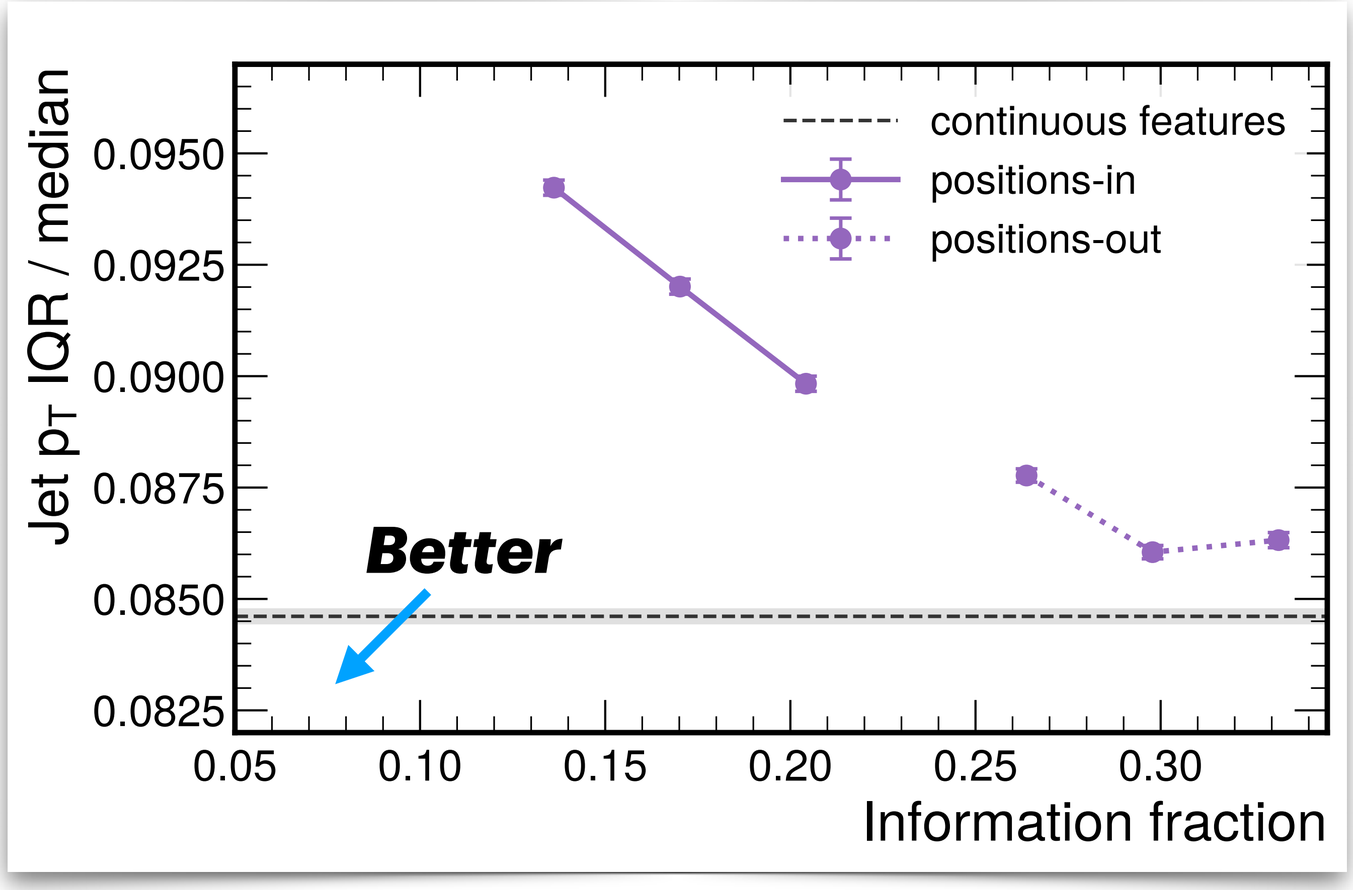


Downstream Costs 1/3



Regression ***predicting jet pT from track/topocluster tokens*** *reaches a regression trained on continuous inputs* ***within uncertainties***

Downstream Costs 2/3



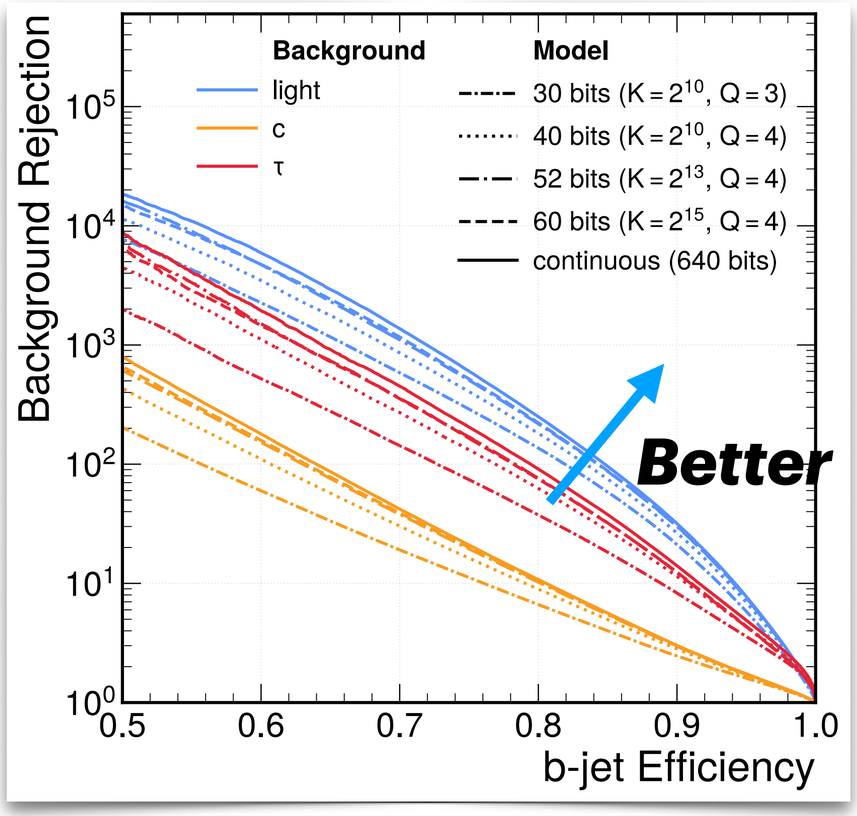
positions-out: η, ϕ binned outside the RVQ, see backup

Particle flow model ***predicting truth particles tokens from track+topocluster tokens*** *reaches a particle flow model trained on continuous inputs* ***within ~2%*** *(rel)*

Downstream Costs 3/3

Tagging b-jets with track tokens

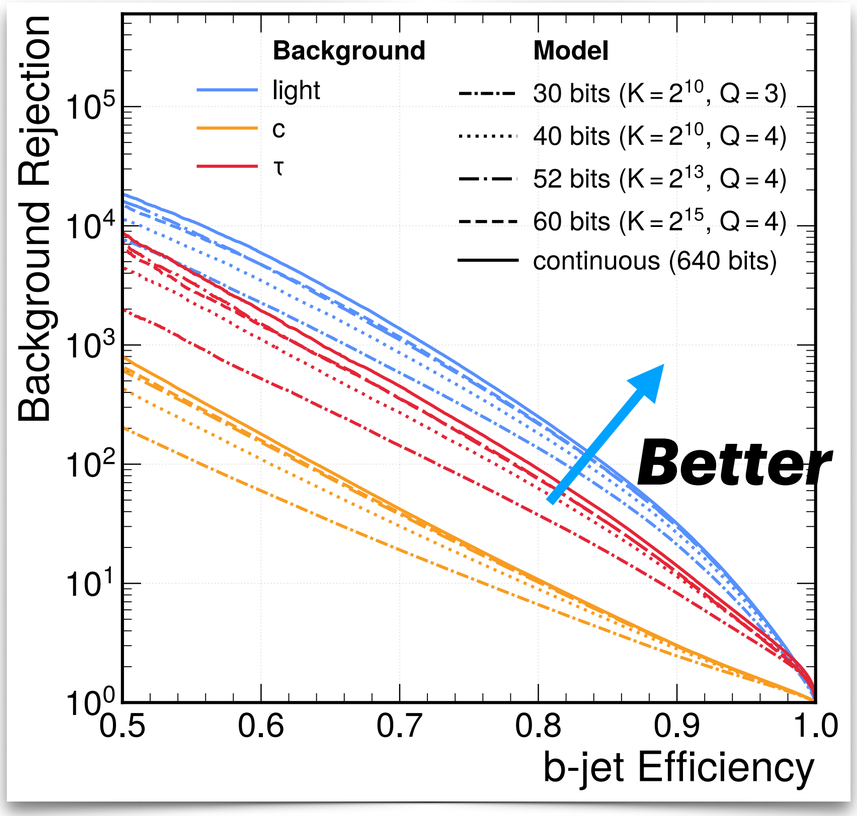
A classifier model predicting truth flavor from JetSet track tokens ***matches continuous baseline AUC at per-mille level*** *at 60 bits...*



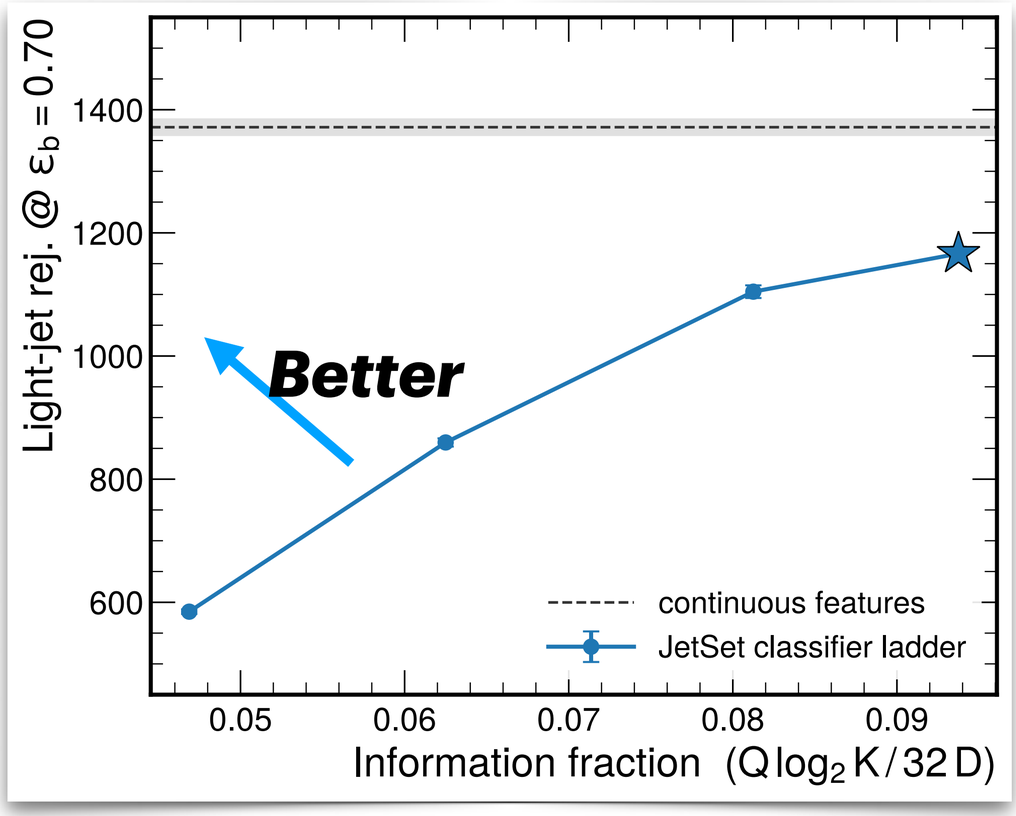
Downstream Costs 3/3

Tagging b-jets with track tokens

A classifier model predicting truth flavor from JetSet track tokens ***matches continuous baseline AUC at per-mille level*** *at 60 bits…*



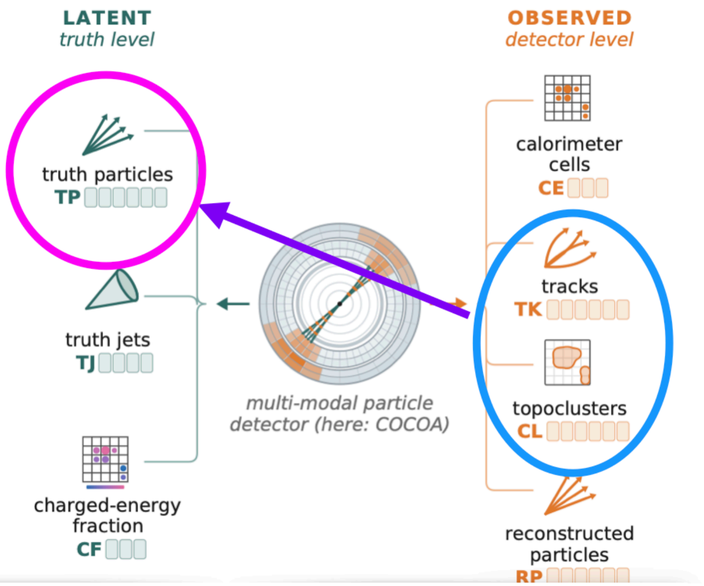
…with 85% of continuous-feature background rejection retained

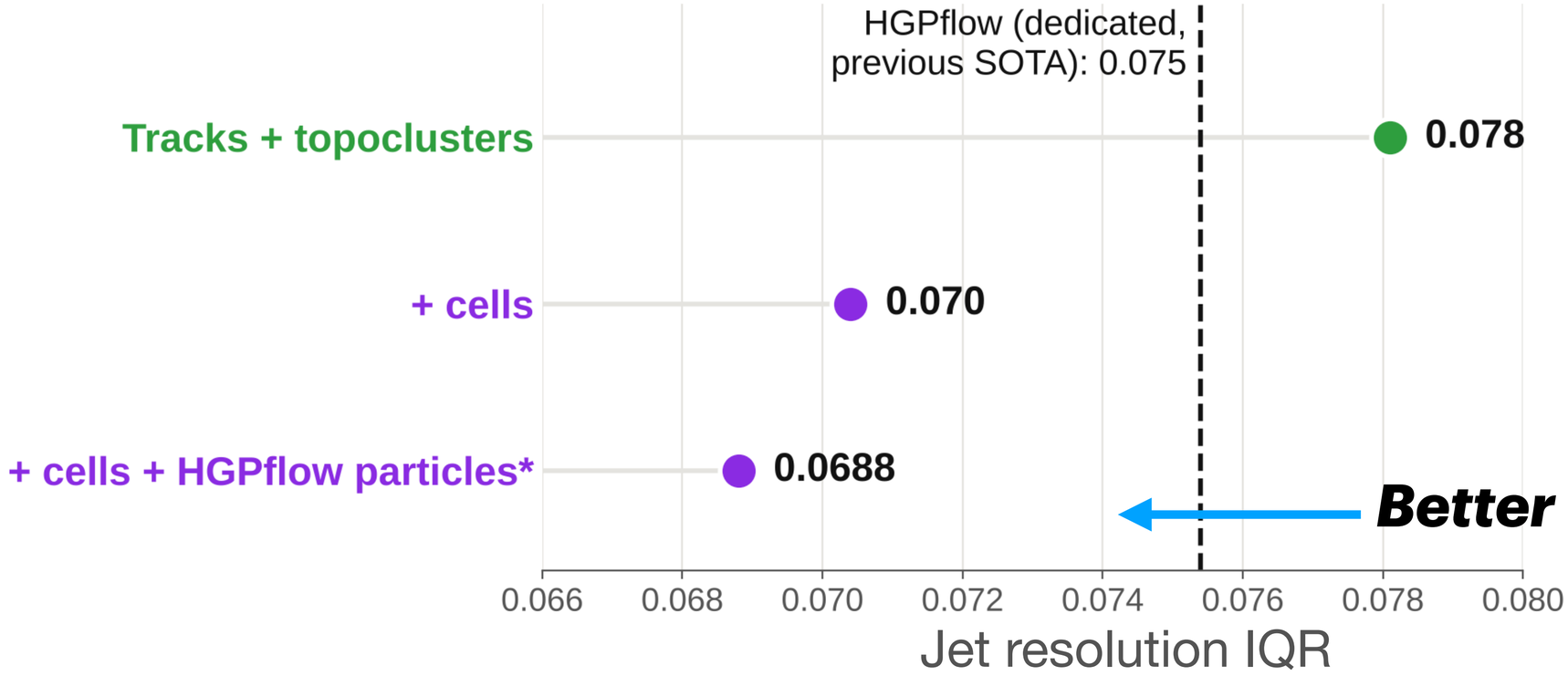


2609.31862v1

HEP Foundation Models

FMs built on tokens can deliver new precision and flexibility



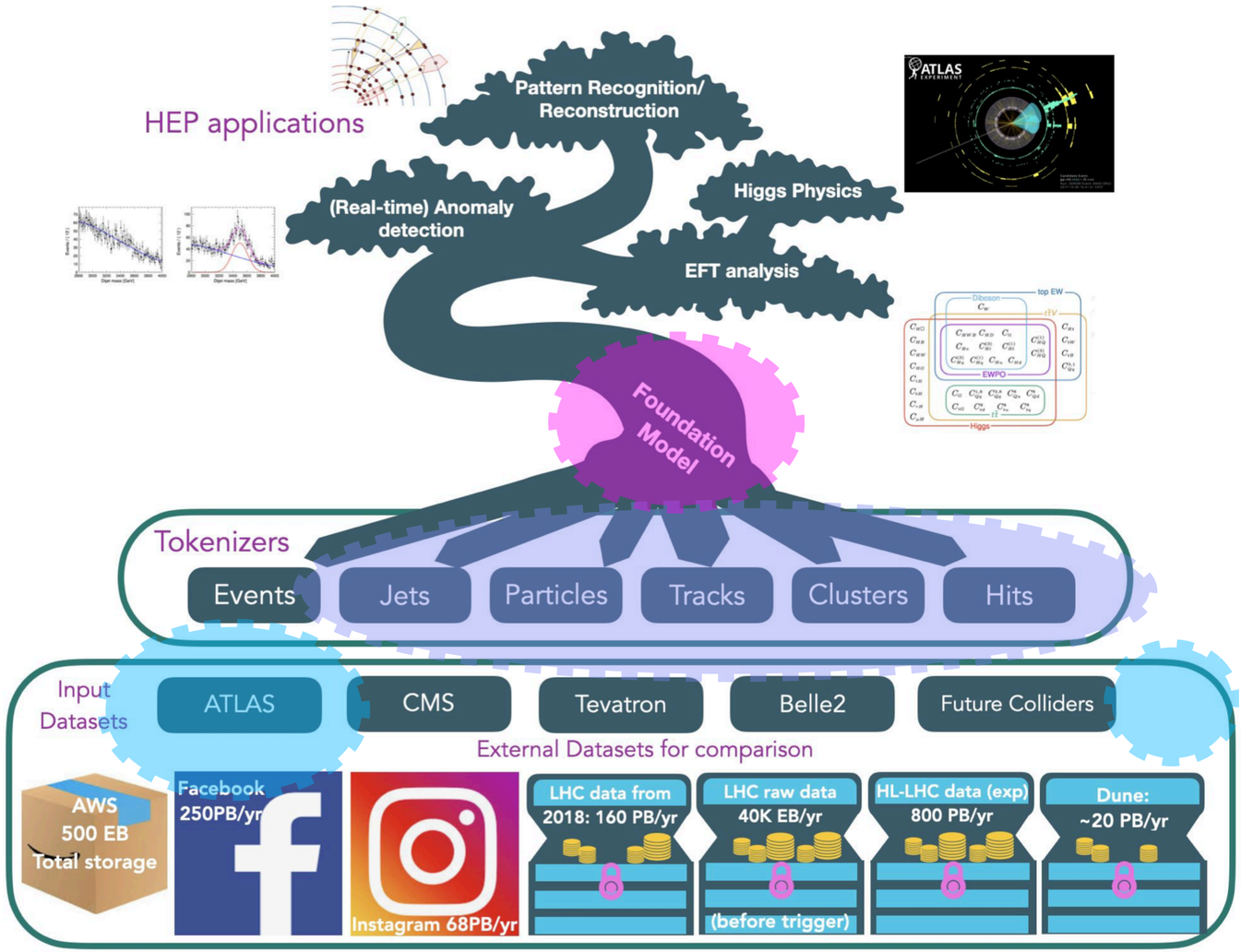


**Never seen in training. Same weights, different prompt: adding inputs a dedicated algorithm can't use beats the previous state of the art.*

Conclusion: how many bits is a jet?

- Learned RVQ tokens reach sub-percent jet pT resolution at **~4–11× compression**, with up to 2.4× fewer bits than fixed binning.
- **Design rules**: use several residual stages, large codebooks, d matched to the bit budget, and positions binned outside the RVQ when they dominate.
- Downstream cost is small: regression matches continuous inputs, particle **flow is within ~2%**, and flavor tagging keeps **85% of light rejection** at 60 bits.
- Tokens enable **prompt-based foundation models** at a cost we now know

Outlook

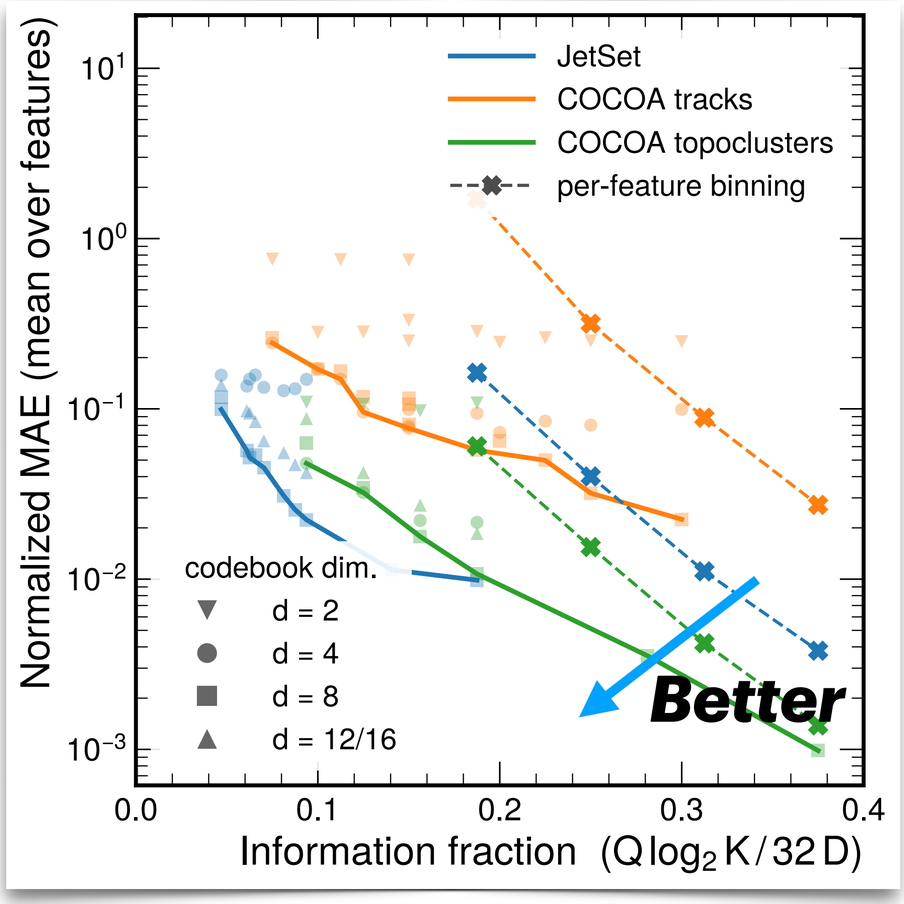


- We now have a good handle on tokenization cost
- Need to converge on tokenization schemes across modalities between/ across experiments
- Understand scale, interoperability
- Always more data to look at!

Backup

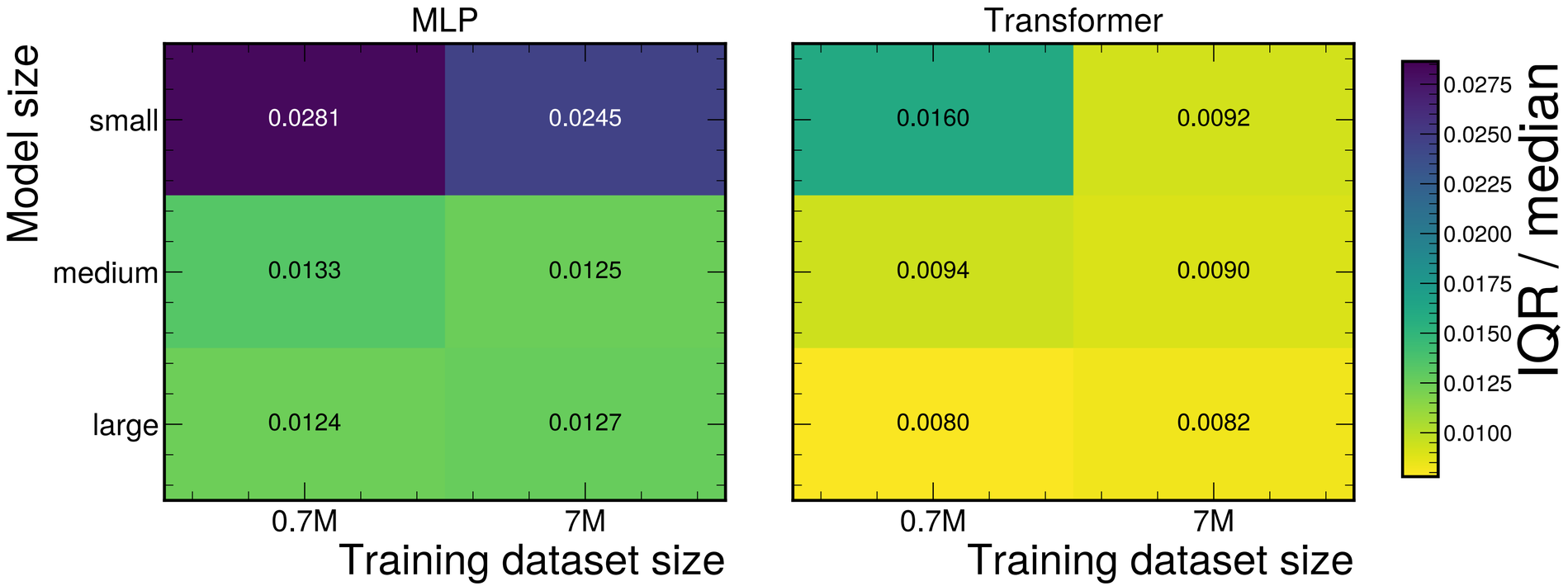
Results: Feature Reconstruction

Jet pT only probes kinematic reconstruction. MAE considers every feature equally, including other parameters like impact parameters



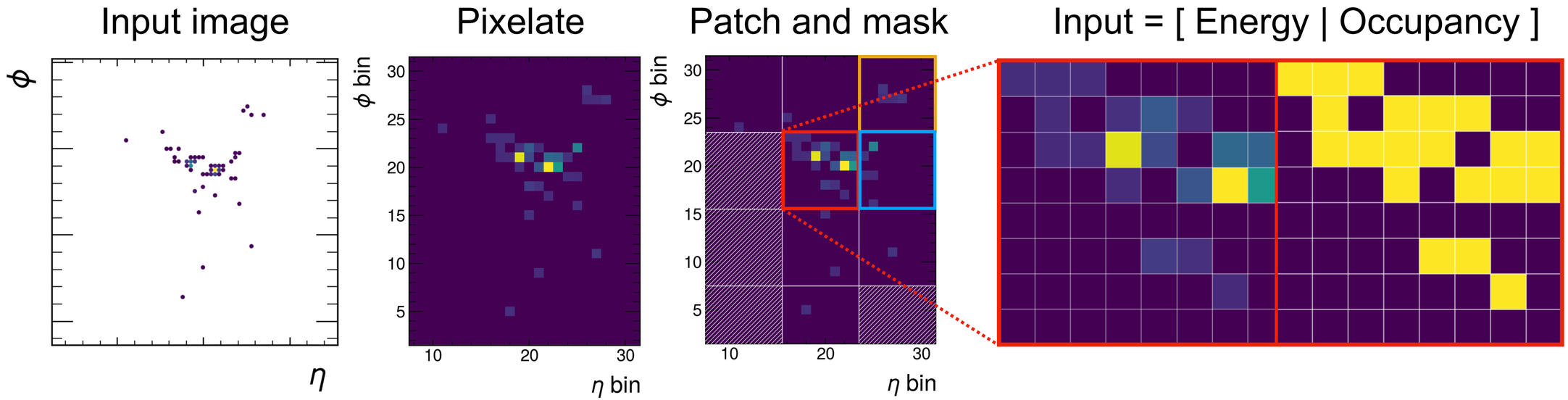
10^{-2} means 1% of the feature's typical spread

In-Context vs No-Context Encoding



Methods: Dealing With Images

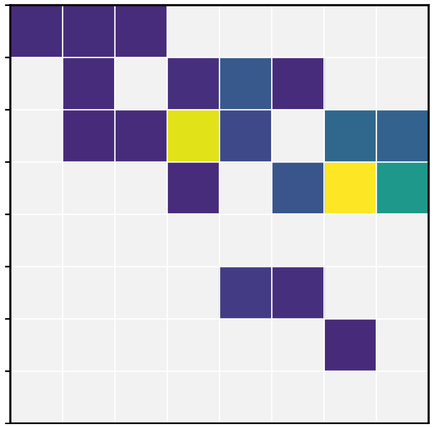
Calorimeter cells are images. We cut them into 8x8 patches and tokenize each patch.



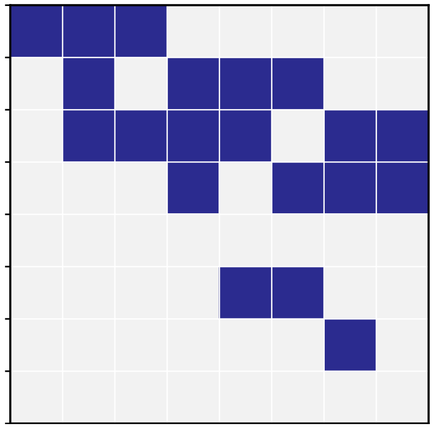
The tokenizer learns **in which** cells energy is deposited, and **how much** energy is deposited (more details).

Cell Training and Inference Details

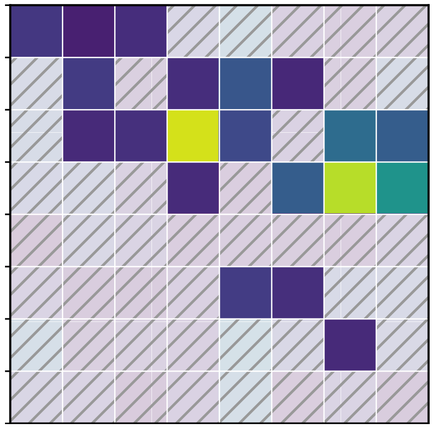
a) Training



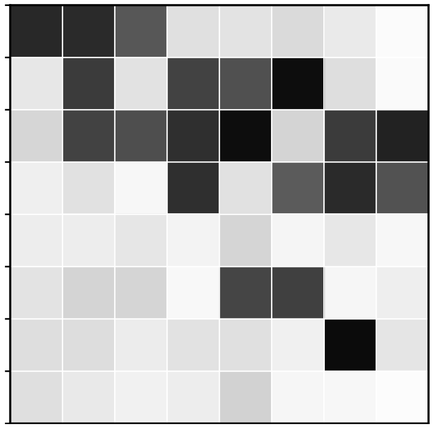
Energy target



Occupancy target

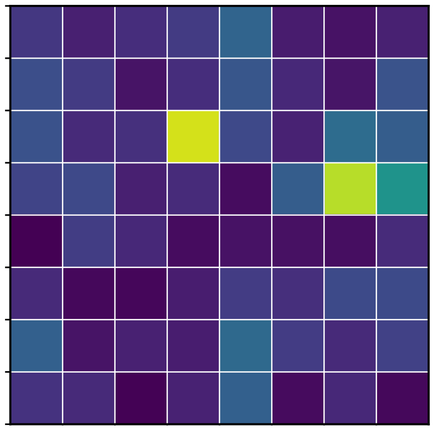


Energy prediction loss
(truth occupied pixels)

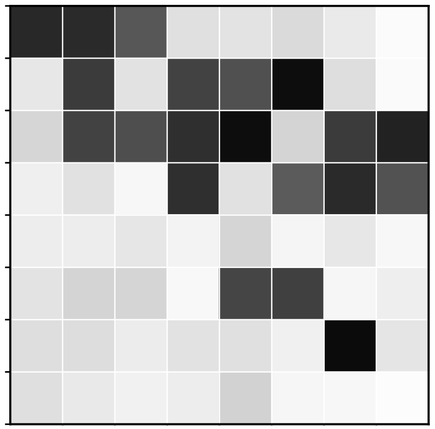


Occupancy prediction loss
(all pixels)

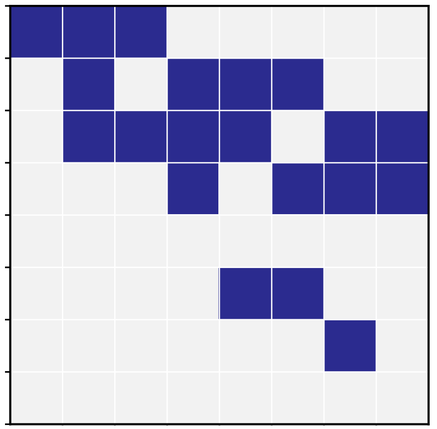
b) Inference



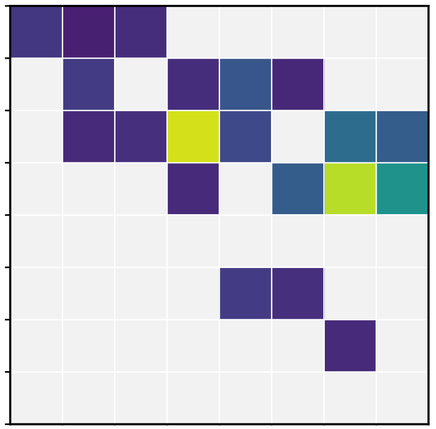
Energy prediction $\hat{E}$



Occupancy prediction $\hat{o}$



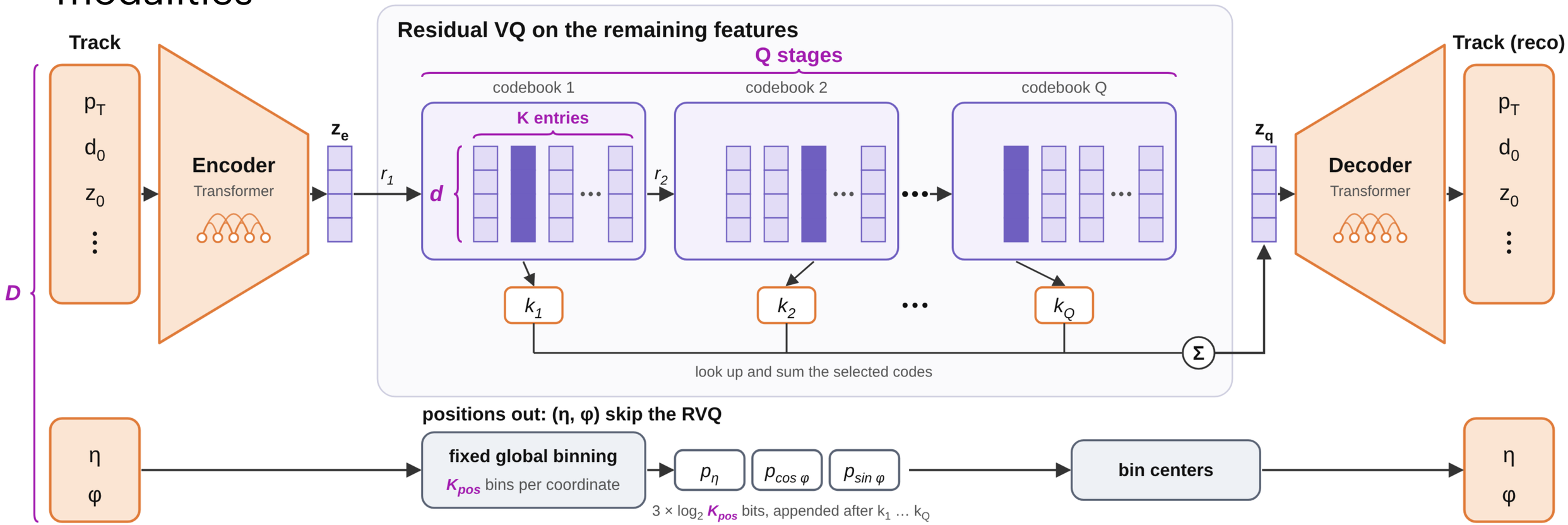
Keep pixels with $\hat{o} > 1/2$



Reconstruction $\hat{E}\left[\hat{o} > 1/2\right]$

Positions-in vs Positions-out

Removing positions from RVQ and instead using a fixed binning encodes them at a known resolution and provides a position vocabulary across modalities



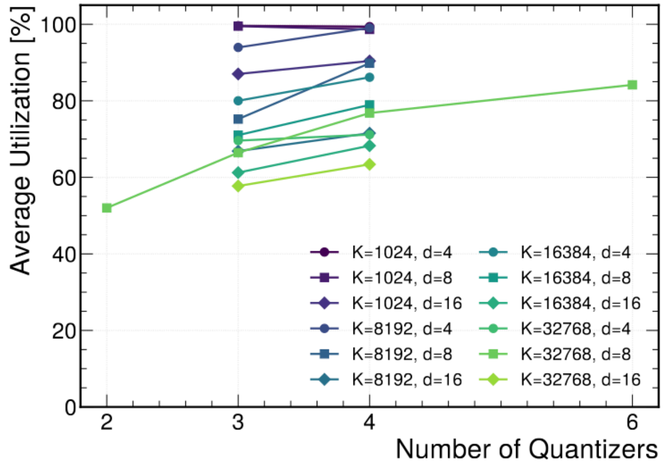
Background Rejection

Background rejection at b-jet efficiency 0.5 and 0.7 for various tokenized models

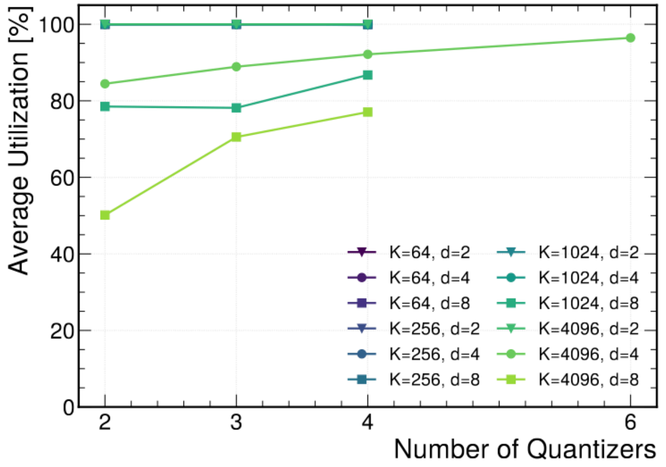
		Continuous	Tokenized			
	Bits/constituent	640	60	52	40	30
	Compression	—	10.7×	12.3×	16.0×	21.3×
$\varepsilon_b = 0.50$	light	$18\,500 \pm 650$	$15\,050 \pm 490$	$16\,090 \pm 540$	$11\,410 \pm 320$	$7\,780 \pm 170$
	c	791 ± 14	659 ± 10	611 ± 9	434 ± 6	204 ± 2
	τ	$8\,680 \pm 800$	$6\,360 \pm 510$	$7\,340 \pm 630$	$4\,480 \pm 300$	$2\,000 \pm 90$
$\varepsilon_b = 0.70$	light	$1\,370 \pm 10$	$1\,170 \pm 10$	$1\,100 \pm 10$	860 ± 7	585 ± 4
	c	42.0 ± 0.2	39.0 ± 0.2	38.0 ± 0.2	30.2 ± 0.1	19.1 ± 0.1
	τ	451 ± 10	354 ± 7	358 ± 7	271 ± 4	143 ± 2

Codebook Utilization

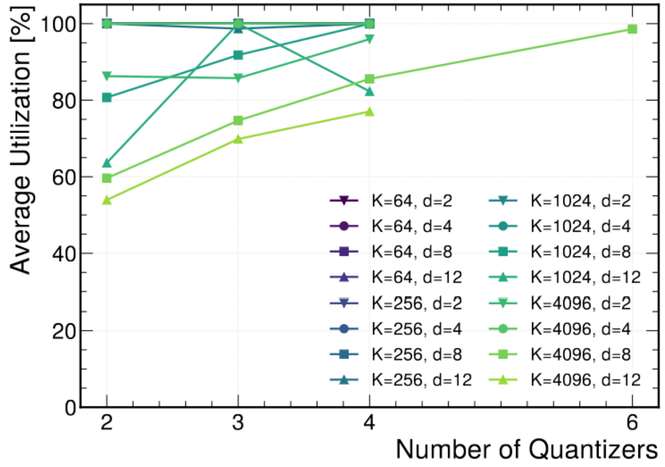
70%+ codebook utilization for the models chosen here



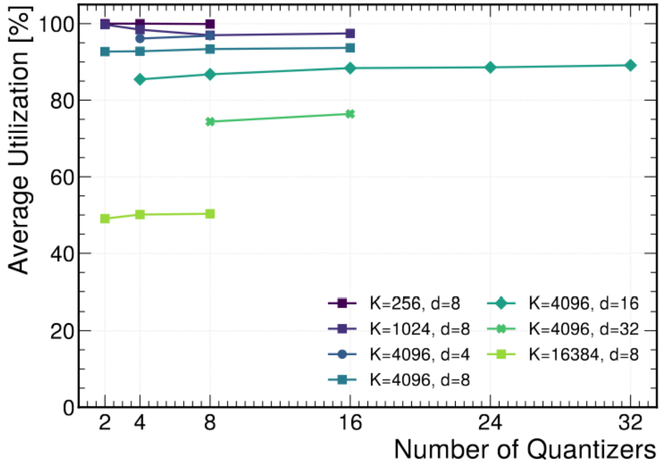
(a) JetSet tracks.



(b) COCOA tracks.



(c) COCOA topoclusters.



(d) COCOA calorimeter cells.

Quantizer Stages are Learning

tSNE embedding of the quantizer vector values

