Toward Domain-Invariant Tokenization for Jet Foundation Models

Giovanni Ottaviano, Chris Scheulen, Tobias Golling

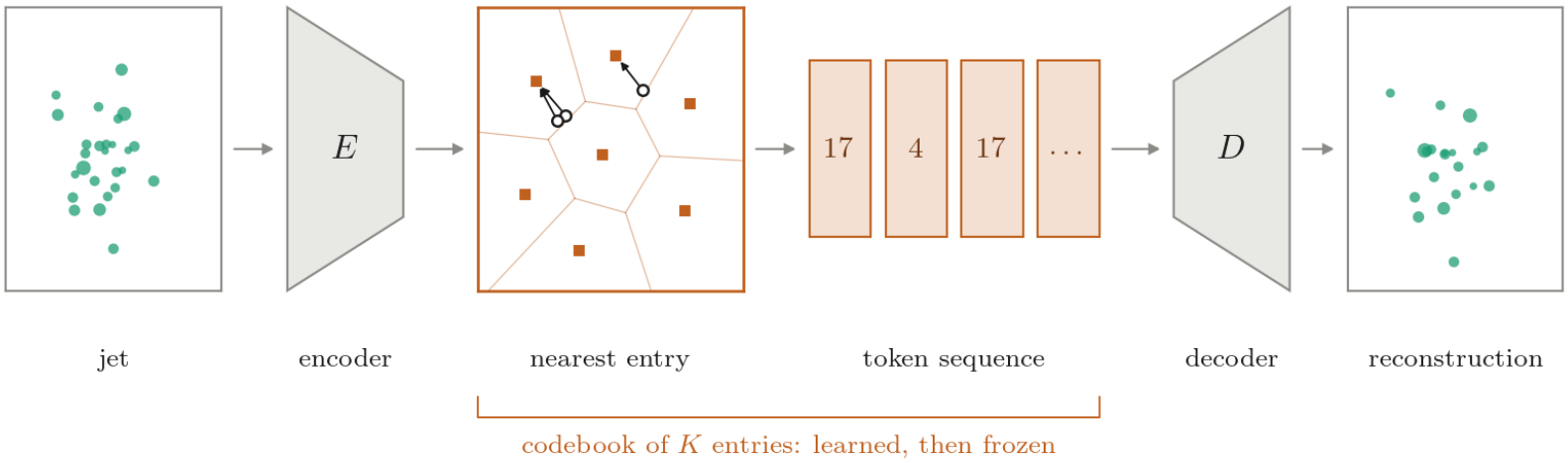
SCIENCES SORBONNE UNIVERSITÉ

AIPHY

UNIVERSITÉ DE GENÈVE
FACULTY OF SCIENCE
Department of Nuclear and Particle Physics

Why discretize and what it costs

Foundation models for physics: one backbone, many systems. Two routes: continuous representations or discrete tokens



The price: q always returns *a* code, never "none of these fit", and what the bottleneck discards is lost to every downstream task

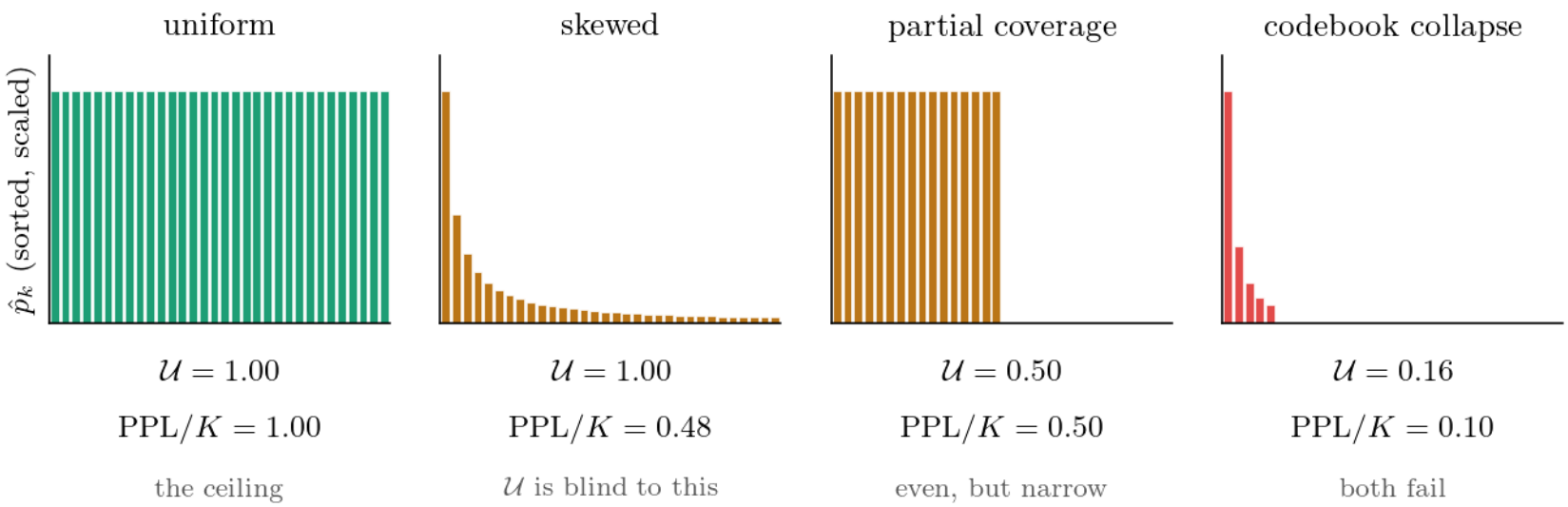
This talk

How do we check the vocabulary is doing its job, before training anything on top of it, and on physics it has never seen?

Evaluating the discrete bottleneck

Setting: nearest-neighbour quantizer q over a codebook $\mathcal{C}$ of size K $\rightarrow$ empirical code usage $\hat{p}$ over a stream of latents

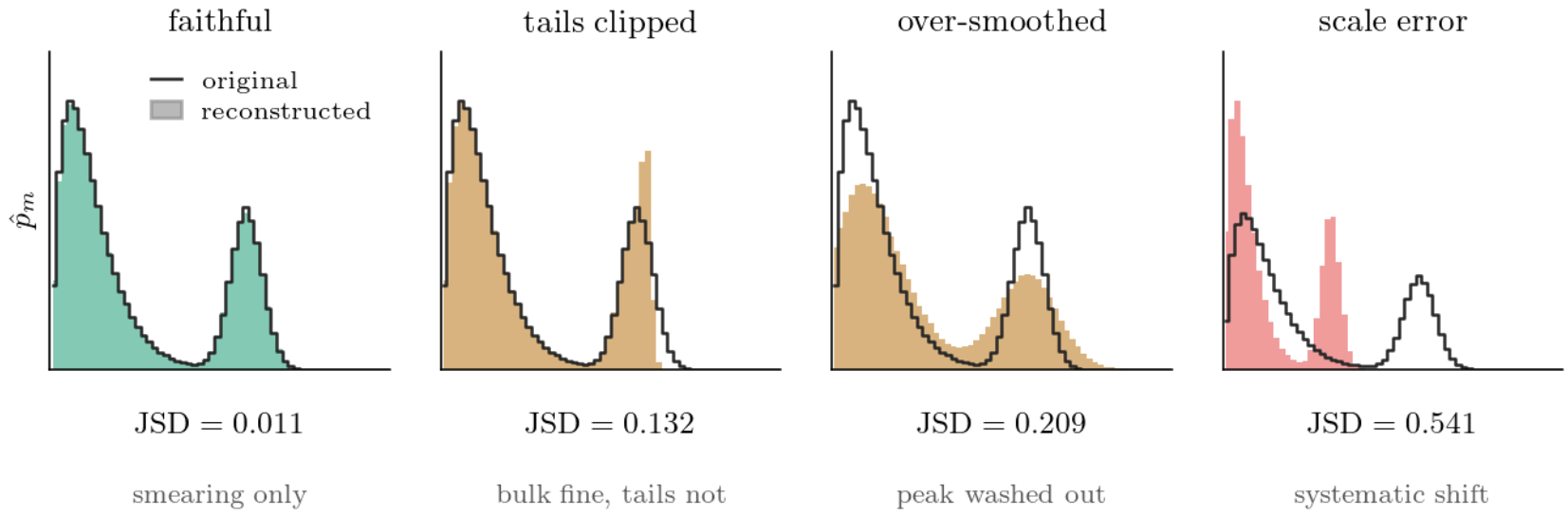
- Utilization $\mathcal{U} \in [0, 1]$: fraction of codes ever used $\rightarrow$ *coverage*
- Normalized perplexity $\mathrm{PPL}/K \in [0, 1]$: effective vocabulary size $\rightarrow$ *evenness*



Distributional reconstruction error

Setting: input x, reconstruction $\hat{x} = D(q(E(x)))$, physical quantity $\mathcal{O}$ $\rightarrow$ histograms $\hat{p}$, $\hat{q}$ on shared bins

Goal: does the tokenizer preserve the *statistics* of $\mathcal{O}$, not just its point-wise values?



Jensen–Shannon distance

Symmetric in $\hat{p}, \hat{q}$ · bounded: $\mathrm{JSD} \in [0, 1]$ · a true metric

Baseline: the OmniJet-α tuple

Setup: Same transformers-based VQ-VAE used for Omnijet-α FM

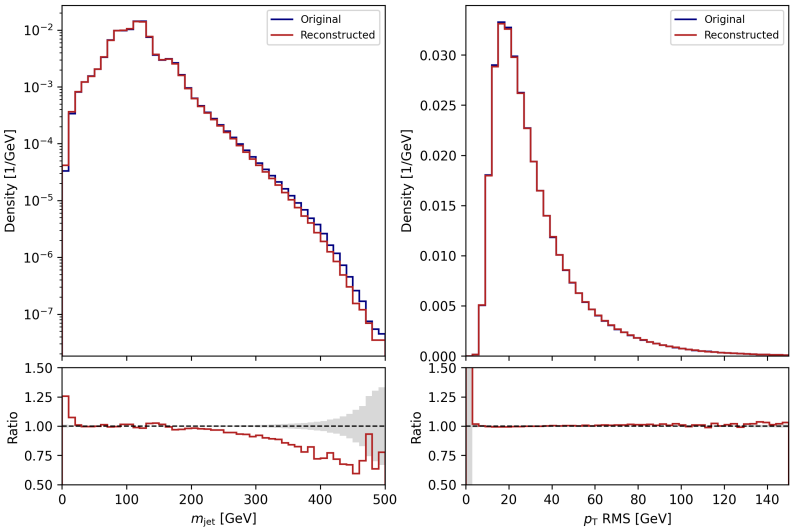
- Codebook: trained on JetClass, then frozen
- Selection: $p_{T\,\mathrm{jet}} \in [500, 1000]$ GeV, $\Delta R < 0.8$
- Normalization: applied unchanged to each dataset

Dataset	$\mathcal{U}$ (%)	PPL/K	jets
JetClass	86.99	0.748	20M
Aspen	87.05	0.623	$\sim$24M
Rodem	87.51	0.693	$\sim$1.8M

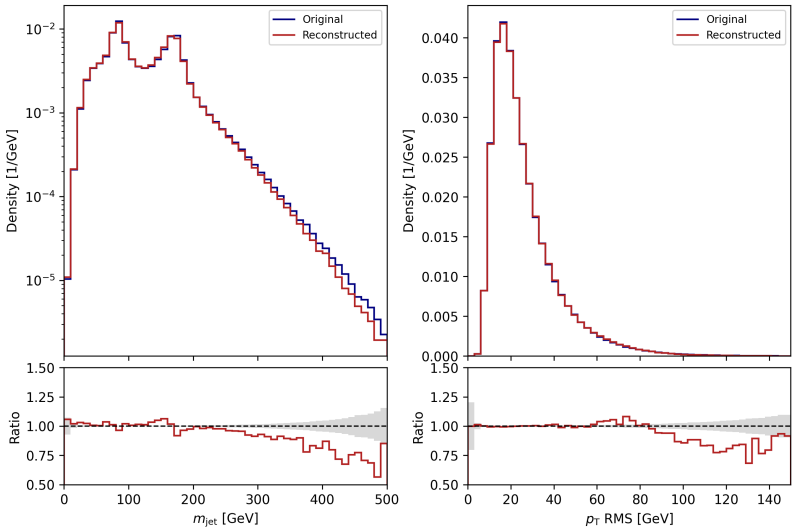
Coverage is flat across all three, evenness is not

Baseline: reconstructed distributions

Same frozen codebook and same round trip to compare original against reconstructed (in-domain vs. off-domain case).



JetClass: in-domain reference



Rodem: off-domain case

Reconstruction fidelity is a separated from codebook coverage, and Jensen-Shannon distance (JSD) is the tool we use to investigate it.

The jet-relative tuple

One token, three relative coordinates:

$$\Big(\underbrace{\log z}_{z = p_{T,i}/p_{T,\mathrm{jet}}} , \underbrace{\Delta\eta,\, \Delta\varphi}_{\text{w.r.t. the jet axis}} \Big) \qquad \text{vs.} \qquad \big(\log p_{T,i},\, \Delta\eta,\, \Delta\varphi\big)$$

Each entry is measured against the jet itself $\Rightarrow$ the tuple describes *what splitting happened*, not where the jet sits

- Which variable ($p_{T,i} \to z$): $p_{T,i}$ inherits the jet momentum, and where a jet's p_T sits is decided by *sample selection* (generation cut, skim). The ratio z is set by fragmentation, and QCD radiation is approximately scale invariant $\Rightarrow$ z is the near-universal entry
- Which transform ($z \to \log z$): among transforms of z, log spreads codebook occupancy far more evenly (ex. the median at 37.8% of the p01–p99 range against 13.2% for $\sqrt{z}$)[1]

[1] **p01, p99: the 1st and 99th percentiles of the distribution.**

Putting the energy scale back

The tuple is dimensionless $\Rightarrow$ the jet energy scale must be included in a different way

$$p_{T,i} = e^{\log z_i} \cdot p_{T,\mathrm{jet}} \qquad \text{(use jet-}p_T\text{ to re-encode } \textit{constituents'}\ p_{T,i}\text{)}$$

	Product quantization	Conditioning
Scale lives	inside the token stream	outside the vocabulary
Token	$(i_{\mathrm{const}}, i_{\mathrm{scale}})$	i_{const} only
Vocabulary	$K_c \times K_s$	K_c
Pros	self-contained sequence, scale is itself generable	zero rate cost; structure only codebook; easy off-envelope extrapolation
Cons	larger vocabulary; untrained embedding for off-envelope, $\mathcal{U}$ collapses	must enter encoder *and* decoder, then be ablated for side-channel collapse

Results: Jet-relative coordinates

- Same codebook recipe, same frozen-transfer protocol
- Same selection: $p_{T\,\mathrm{jet}} \in [500, 1000]$ GeV, $\Delta R < 0.8$
- Only the coordinates change: $\log p_{T,i} \to \log z$

Dataset	$\Delta\mathcal{U}$	$\mathcal{U}$ (%)	$\Delta\,\mathrm{PPL}/K$	PPL/K
JetClass	+10.4	**97.39**	+0.071	**0.819**
Aspen	+10.4	**97.42**	+0.239	**0.862**
Rodem	+9.8	**97.35**	+0.159	**0.852**

Coverage gains ∼10 percentage points everywhere, and evenness improves most off-domain → Aspen and Rodem now exceed JetClass, despite the codebook being fitted on it.

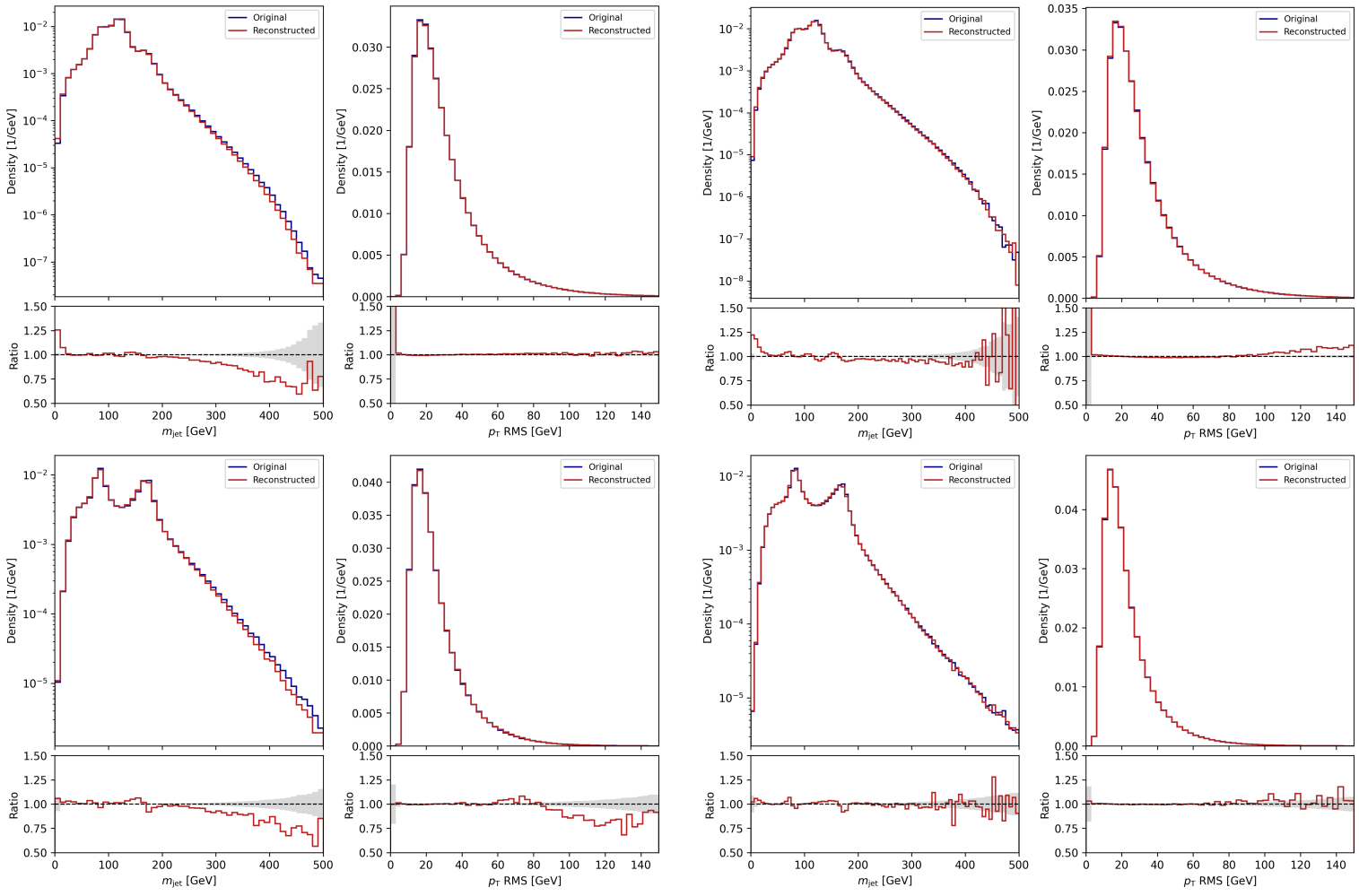
Reconstructed statistics: coordinates vs. domain

Baseline

Jet-relative

JetClass

Rodem



Across: the coordinate change. Down: in-domain to off-domain.

Summary

- The tokenizer is frozen before the backbone $\rightarrow$ whatever it discards is unavailable on every downstream task
- Measured with $\mathcal{U}$, PPL/K and JSD on reconstructed physical statistics (necessary, but not sufficient)
- Jet-relative coordinates: z removes a sample-selection variable (set only by physical arguments), log spreads occupancy far more evenly than z
- Scale re-enters through the constituent: product quantization or conditioning

Thank you!

ottaviano@lpsm.paris — Poster 50

slides & links

Backup

The Kullback–Leibler divergence

Definition: for p, q on $\{1, \dots, K\}$ with $p \ll q$,

$$\mathrm{KL}(p \,\|\, q) = \sum_{k=1}^{K} p_k \log_2 \frac{p_k}{q_k} = \underbrace{\mathbb{E}_{k\sim p}\left[\log_2 \frac{p_k}{q_k}\right]}_{\text{log-likelihood ratio under } p}$$

Reading: expected number of extra bits needed to encode samples from p using a code optimized for q

- Non-negative: $\mathrm{KL}(p\|q) \geqslant 0$, with equality iff $p = q$ (Jensen)
- Not a metric: $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$, and the triangle inequality fails
- Unbounded: $\mathrm{KL}(p\|q) = +\infty$ if $q_k = 0 < p_k$ for some k

Entropy decomposition

$$\mathrm{KL}(p \,\|\, q) = \underbrace{-\sum_k p_k \log_2 q_k}_{\text{cross-entropy } H(p,q)} - \underbrace{\Big(-\sum_k p_k \log_2 p_k\Big)}_{\text{entropy } H(p)}$$

Formulation - evaluating the discrete bottleneck

Setting: encoder E, codebook $\mathcal{C} = \{\beta_1, \ldots, \beta_K\}$, nearest-neighbour quantizer

$$q(\boldsymbol{z}) = \arg\min_{1 \leqslant k \leqslant K} \|\boldsymbol{z} - \beta_k\|_2, \qquad \boldsymbol{z} \in \mathbb{R}^d$$

Goal: quantify how well the K codes are exploited on a stream of latents $(\boldsymbol{z}_i)_{1 \leqslant i \leqslant N}$

- Empirical code distribution: $\hat{p}_k = \dfrac{1}{N} \sum_{i=1}^{N} 1\{q(\boldsymbol{z}_i) = k\}$
- Failure mode: index collapse, i.e. $\hat{p}$ concentrated on $\ll K$ codes

Codebook metrics

$$\underbrace{\mathcal{U}(\hat{p}) = \frac{1}{K} \sum_{k=1}^{K} 1\{\hat{p}_k > 0\}}_{\text{utilization} \in [0,1]} \qquad \underbrace{\mathrm{PPL}(\hat{p}) = 2^{\left(-\sum_k \hat{p}_k \log_2 \hat{p}_k\right)}}_{\text{perplexity} \in [1,K]}$$

Formulation - distributional reconstruction error

Setting: input x, reconstruction $\hat{x} = D(q(E(x)))$, physical quantity $\mathcal{O}$

Histograms on shared bins $B_1, \ldots, B_M$:

$$\hat{p}_m = \frac{1}{N} \sum_{i=1}^{N} 1\{\mathcal{O}(x)_i \in B_m\}, \qquad \hat{q}_m = \frac{1}{N} \sum_{i=1}^{N} 1\{\mathcal{O}(\hat{x})_i \in B_m\}$$

Goal: does the tokenizer preserve the *statistics* of $\mathcal{O}$, not just its pointwise values?

- Mixture: $\bar{m} = \frac{1}{2}(\hat{p} + \hat{q})$, with $\bar{m}_k \geqslant \frac{1}{2}\hat{p}_k$ and $\bar{m}_k \geqslant \frac{1}{2}\hat{q}_k$ for all k
- Why not KL: asymmetric and $= +\infty$ as soon as $\hat{q}_k = 0 < \hat{p}_k$

Jensen–Shannon distance

$$\mathrm{JSD}(\hat{p}, \hat{q}) = \underbrace{\sqrt{\tfrac{1}{2}\mathrm{KL}(\hat{p} \,\|\, \bar{m}) + \tfrac{1}{2}\mathrm{KL}(\hat{q} \,\|\, \bar{m})}}_{\text{symmetric, bounded: } \mathrm{JSD} \in [0,1]}$$

Adding the scale to the model

Product quantization: two indices, two embedding tables:

$$\texttt{emb} = \underbrace{E_c(i_{\text{const}})}_{K_c \times d} + \underbrace{E_s(i_{\text{scale}})}_{K_s \times d} \qquad \text{or} \qquad \texttt{emb} = W\left[\, E_c(i_{\text{const}})\,;\, E_s(i_{\text{scale}})\,\right]$$

These are the same function class: $[a; b]W = aW_1 + bW_2 \Rightarrow$ a dimension budget, not a modelling decision

- Sum: keeps d fixed, ties both tables to it
- Concat: splits d between the two, makes the provenance explicit

Conditioning: same sum, but the scale term is *continuous*:

$$\texttt{emb} = E_c(i_{\text{const}}) + \underbrace{g(\log p_{T,i})}_{\text{no cbook, no bins}} \qquad \text{or} \qquad h = \texttt{model}(\texttt{tokens},\ \texttt{cond} = \log p_{T,i})$$

Datasets specifics

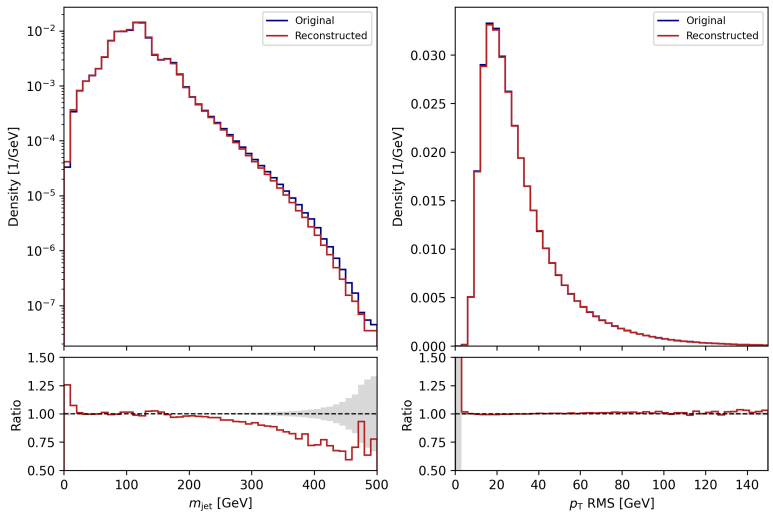
	Origin	Detector	R	Constituents	Jet p_T [GeV]
JetClass	MG5 + Pythia	CMS (Delphes)	0.8	PF	500–1000
Aspen	CMS 2016 data	CMS	0.8	PF, PUPPI	$\geqslant 300$
JetSet	Powheg + Pythia	ATLAS (Geant4)	0.4	tracks	20–500
Rodem	MG5 + Pythia	ATLAS (Delphes)	1.0	PF	450–1200

- JetClass: the reference benchmark (100M train / 5M val / 20M test), 10 classes, $|\eta| < 2$. No pileup, no measurement uncertainty.
- Aspen: CMS open data ($\sim$178M unlabelled jets), 16.4 fb^{-1}. Trigger bias near turn-on and detector artefacts (visible geometry at $|\eta| \simeq 1.6$, dead cells)
- JetSet: ATLAS open data, $\sim$199M jets from $t\bar{t}$ at $\sqrt{s} = 13.6$ TeV. The only full Geant4 simulation, and the only one built from charged tracks alone (no neutral component)
- Rodem: large-radius jets, hence $\sim$50.5 constituents/jet, and no soft-p_T threshold, the spectrum runs smooth to the bottom, with 5.09% below 0.5 GeV

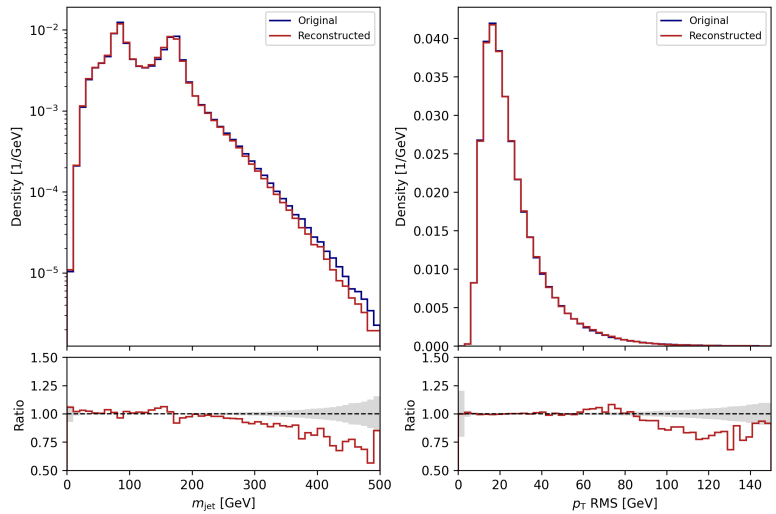
Full results: Baseline

Setup: single JetClass-trained codebook, applied frozen to all four datasets ($p_{T,\mathrm{jet}} \in [500, 1000]$ GeV, $\Delta R < 0.8$, same input normalization)

Dataset	$\mathcal{U}$ (%)	PPL/K	jets
JetClass	86.99	0.748	20M
Aspen	87.05	0.623	∼24M
JetSet	68.55	0.333	∼25M
Rodem	87.51	0.693	∼3M



JetClass: in-domain reference

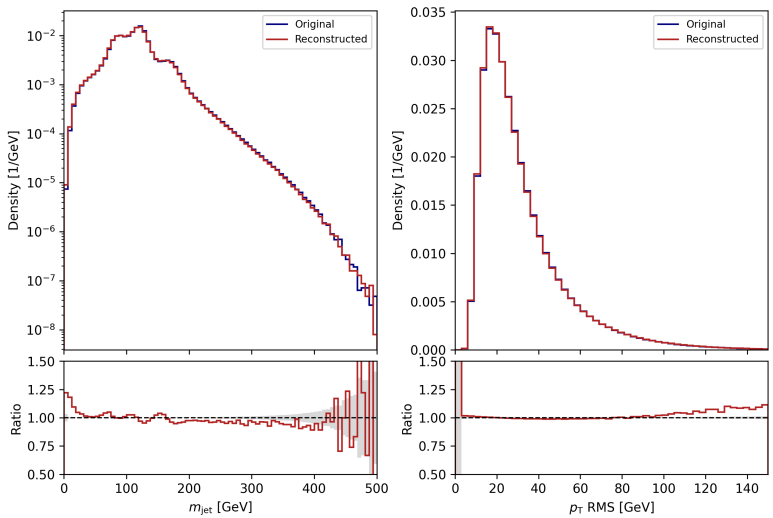


Rodem: original vs. reconstructed

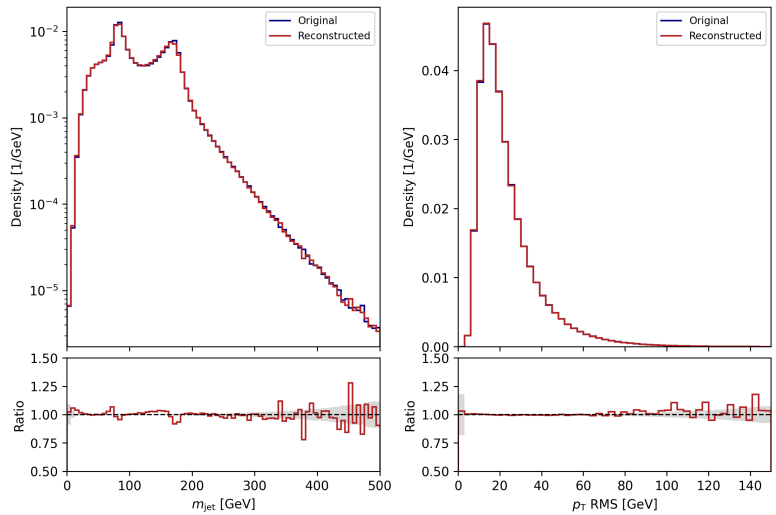
Full results: the jet-relative tuple

Same setup, same frozen JetClass codebook, only the coordinates change ($p_{T,\mathrm{jet}} \in [500, 1000]$ GeV, $\Delta R < 0.8$)

Dataset	$\mathcal{U}$ (%)	Δ	PPL/K	Δ
JetClass	97.39	**+10.4**	0.819	**+0.071**
Aspen	97.42	**+10.4**	0.862	**+0.239**
JetSet	68.76	+0.2	0.367	+0.034
Rodem	97.35	**+9.8**	0.852	**+0.159**



JetClass: in-domain reference



Rodem: original vs. reconstructed