Foundation Models | Autoregressive Generation

SPADE

SPlit And DElay Embeddings for Autoregressive High-Granularity Calorimeter Simulation

ML4JETS 2026 · 16 SEPTEMBER 2026 · VIENNA

AUTHORS

Joschka Birk, Frank Gaede, Anna Hallin, Gregor Kasieczka, Martina Mozzanica, **Henning Rose**

henning.rose@uni-hamburg.de | arXiv:2606.11304 | github.com/uhh-pd-ml/SPADE

UHH

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

DESY.

From Text to Tensors

How an LLM reads text.

"**The** quick fox"

text input

From Text to Tensors

How an LLM reads text.

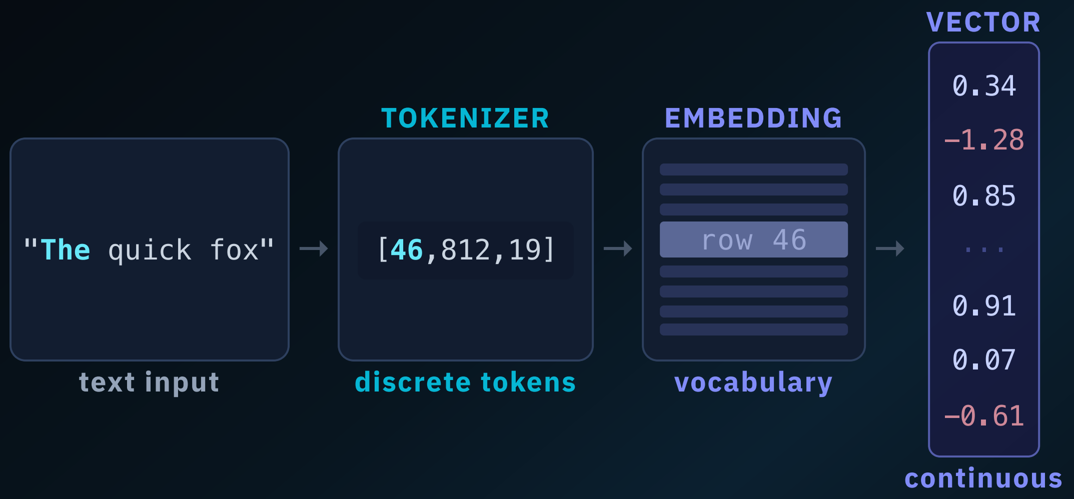
TOKENIZER

"The quick fox" → [46,812,19]

text input — discrete tokens

From Text to Tensors

How an LLM reads text.

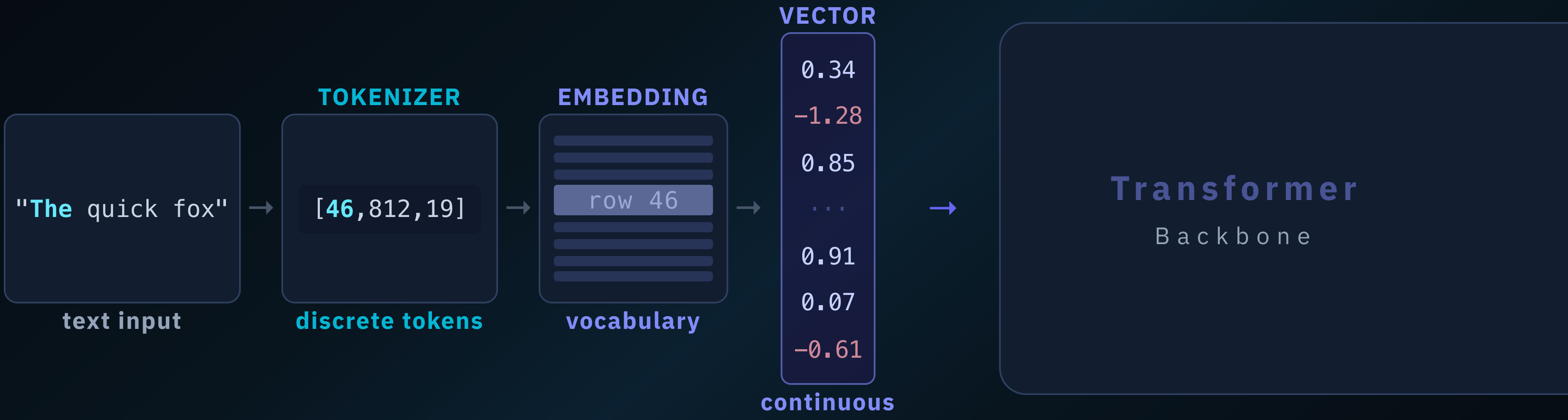


From Text to Tensors

How an LLM reads text.

From Text to Tensors

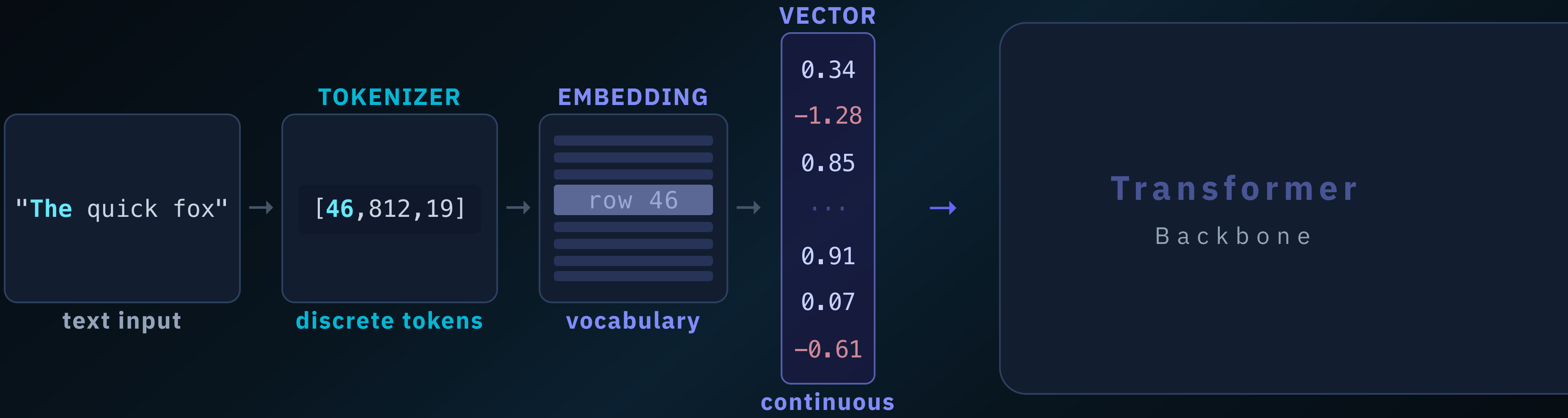
How an LLM reads text.



Text: a finite vocabulary of subword tokens, so every token is one lookup.

From Text to Tensors

How an LLM reads text.



Text: a finite vocabulary of subword tokens, so every token is one lookup.

Physics: part discrete, part continuous. One token per hit needs a huge vocabulary or a learned codebook.

Dataset & Geometry

SPLIT & DELAY

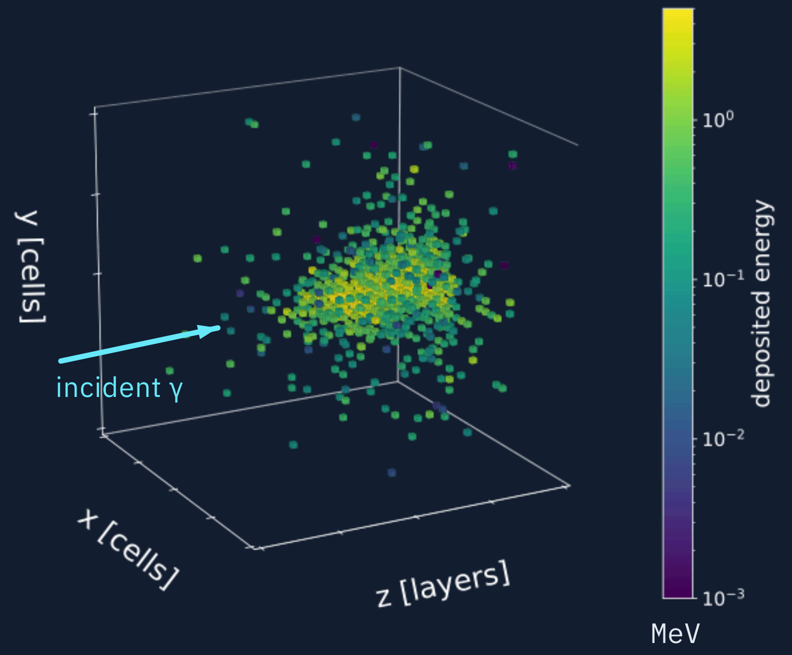
γ-showers [1]

High-granularity ILD ECAL (simplified) [2]

A 30×30×30 grid of sensor cells

Incident energy range:	10–100 GeV
Hits per shower:	100–1700
Structure:	4D, sparse

[1] Buhmann et al. 2005.05334 (2020)
[2] ILD Collaboration 2003.01116 (2020)



Dataset & Geometry

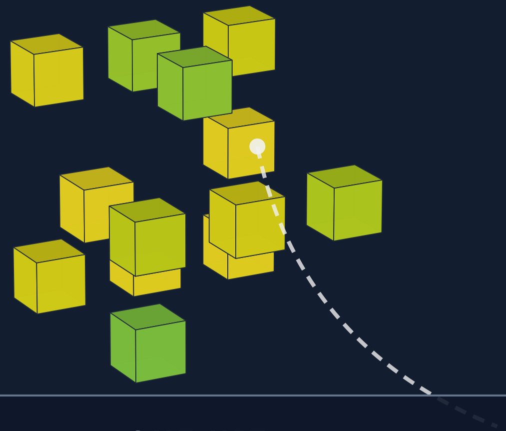
γ-showers [1]

High-granularity ILD ECAL (simplified) [2]

A 30×30×30 grid of sensor cells

Incident energy range:	10–100 GeV
Hits per shower:	100–1700
Structure:	4D, sparse

[1] Buhmann et al. 2005.05334 (2020)
[2] ILD Collaboration 2003.01116 (2020)



ONE HIT

[z: 5, x: 14, y: 22, E: 2.1 MeV]

Splitting the Vocabulary

△ Granularity Bottleneck

COMBINED VOCABULARY

One token per cell. The size is a product:

$$N_{vocab} = N_x \times N_y \times N_z$$

VOCABULARY SIZE

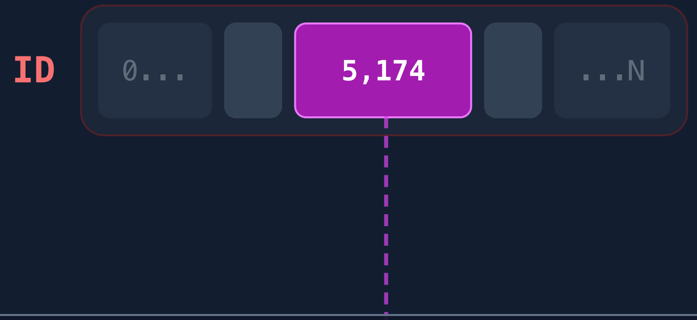
30×30×30 grid	27,000
Real ILD ECAL [2]	~30 M

[2] ILD Collaboration 2003.01116 (2020)

COMBINED VOCABULARY

1 CELL = 1 TOKEN ID

$$ID = x + y \cdot N_x + z \cdot N_x \cdot N_y$$



ONE HIT

[z: 5, x: 14, y: 22, E: 2.1 MeV]

Splitting the Vocabulary

SPLIT & DELAY

Split Vocabularies

ONE PER COORDINATE

One vocabulary each for z, x and y. E stays continuous. The size is a sum:

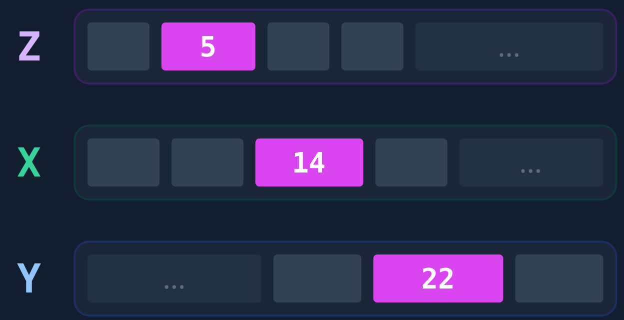
$$N_{vocab} = N_x + N_y + N_z$$

VOCABULARY SIZE

30×30×30 grid	**90** (vs 27,000)
Real ILD ECAL [2]	**hundreds** (vs ~30 M)

[2] ILD Collaboration 2003.01116 (2020)

SPLIT VOCABULARIES



ONE HIT

[z: 5, x: 14, y: 22, E: 2.1 MeV]

Delaying the Streams

NO DELAY: FOUR PARALLEL STREAMS

Z	z_1	z_2	z_3
X	x_1	x_2	x_3
Y	y_1	y_2	y_3
E	E_1	E_2	E_3

all four emitted at the same step → sampled independently

Parallel: $P(z) \cdot P(x) \cdot P(y) \cdot P(E)$ each given the previous hits

The four features of a hit are sampled independently, so the correlations within a hit are lost.

Delaying the Streams

SPLIT-AND-DELAY: EACH STREAM ONE STEP BEHIND THE ONE ABOVE

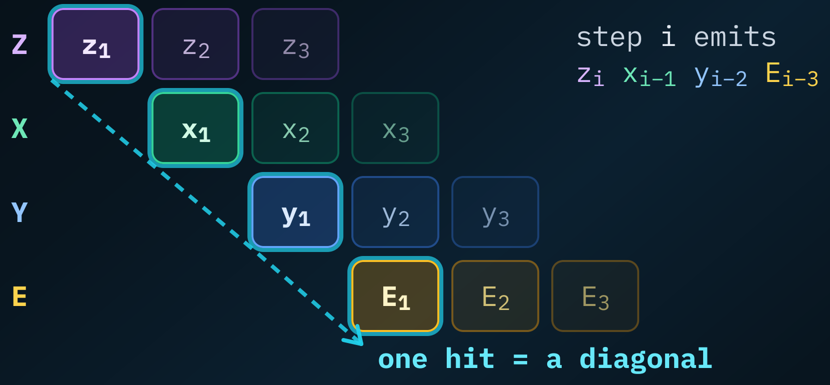
Z	z_1	z_2	z_3			
X		x_1	x_2	x_3		
Y			y_1	y_2	y_3	
E				E_1	E_2	E_3

step i emits
z_i x_{i-1} y_{i-2} E_{i-3}

Split-and-Delay: $P(z)\cdot P(x|z)\cdot P(y|z,x)\cdot P(E|z,x,y)$ each also given the previous hits
Exact chain rule within each hit, still one step per hit. **One token per feature would need 4× the steps.**

Delaying the Streams

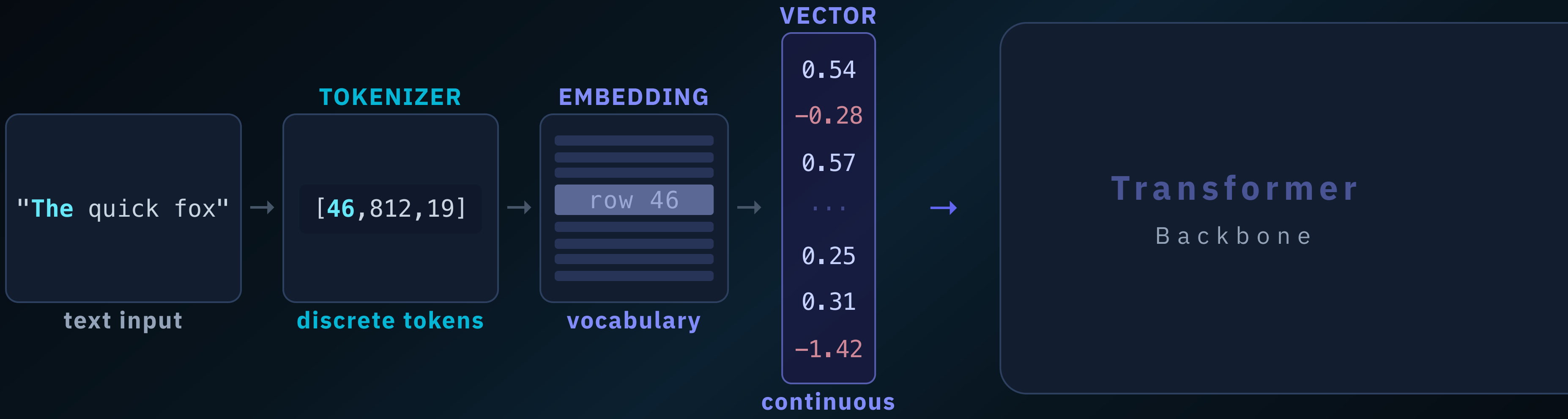
SPLIT-AND-DELAY: EACH STREAM ONE STEP BEHIND THE ONE ABOVE



Split-and-Delay: $P(z) \cdot P(x|z) \cdot P(y|z,x) \cdot P(E|z,x,y)$ each also given the previous hits

Exact chain rule within each hit, still one step per hit. **One token per feature would need 4× the steps.**

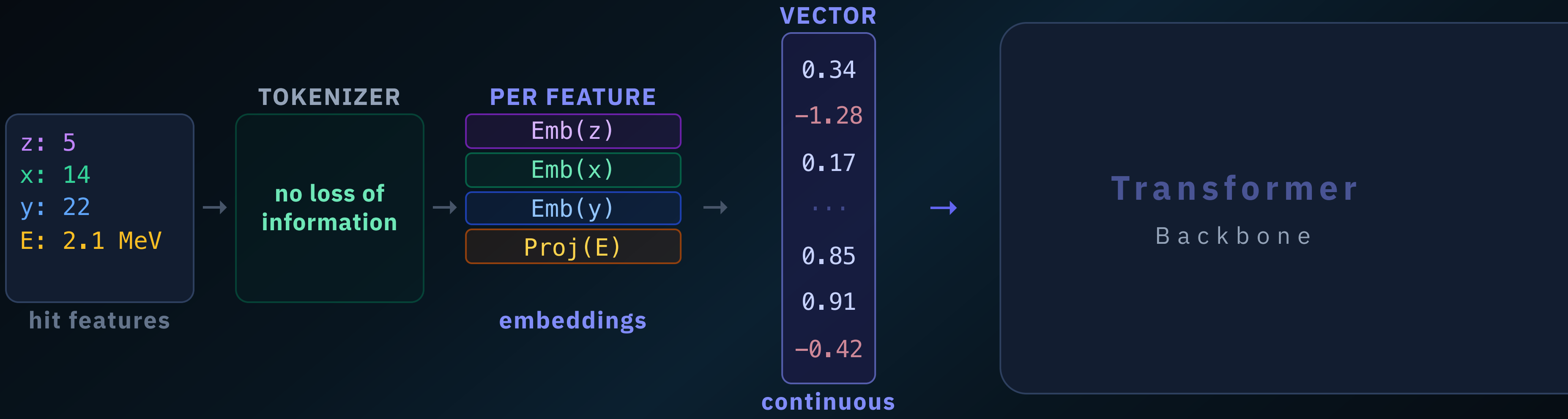
Embedding



Text: a finite vocabulary of subword tokens, so every token is one lookup.

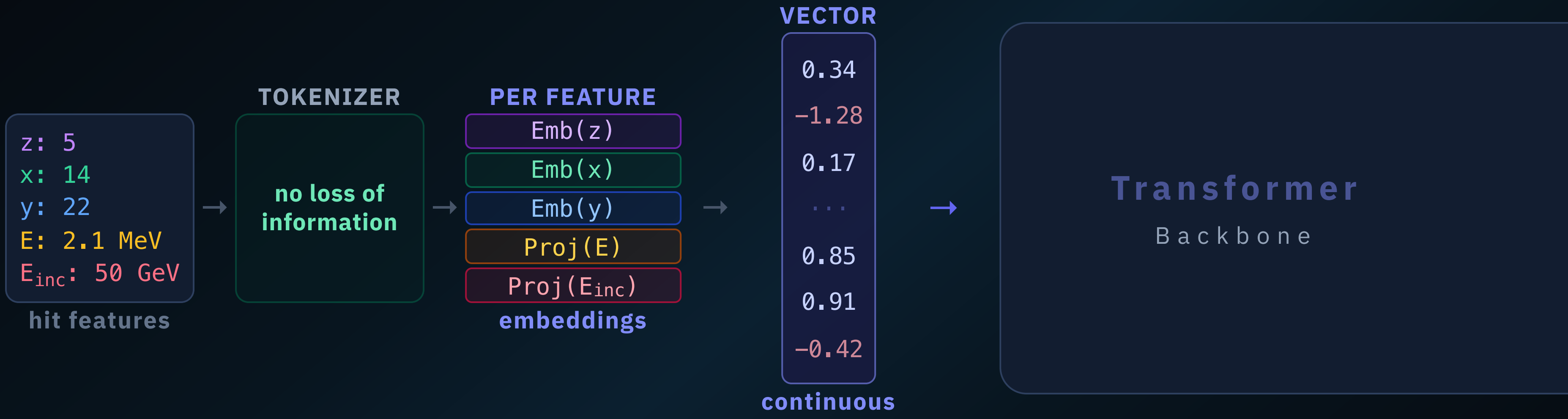
Embedding

SPADE
SPLIT & DELAY



Physics: one lookup per cell index, one projection per energy, **no codebook**.

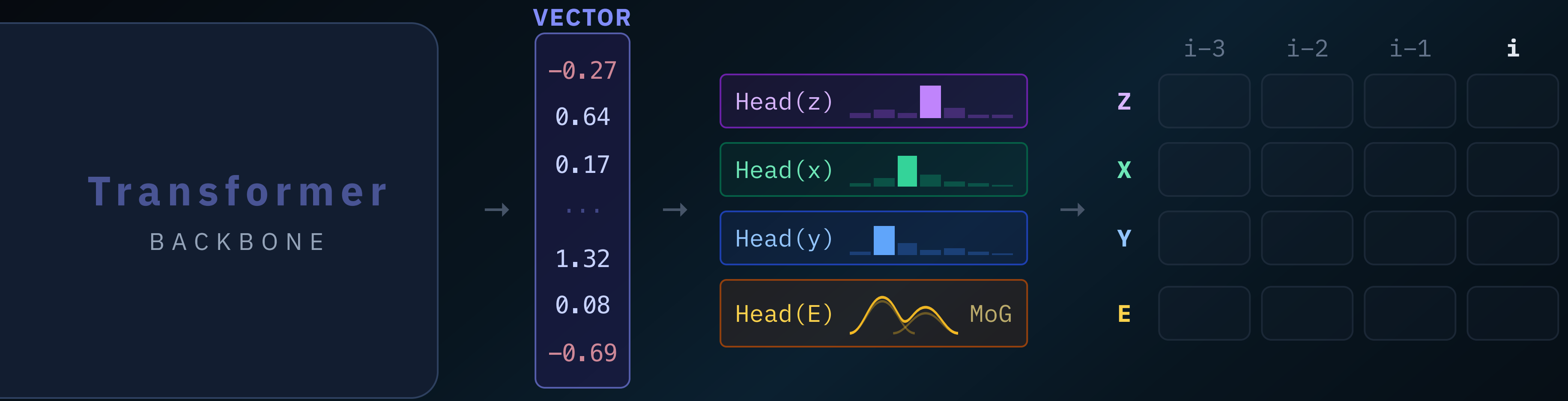
Embedding



Physics: one lookup per cell index, one projection per energy, **no codebook**.

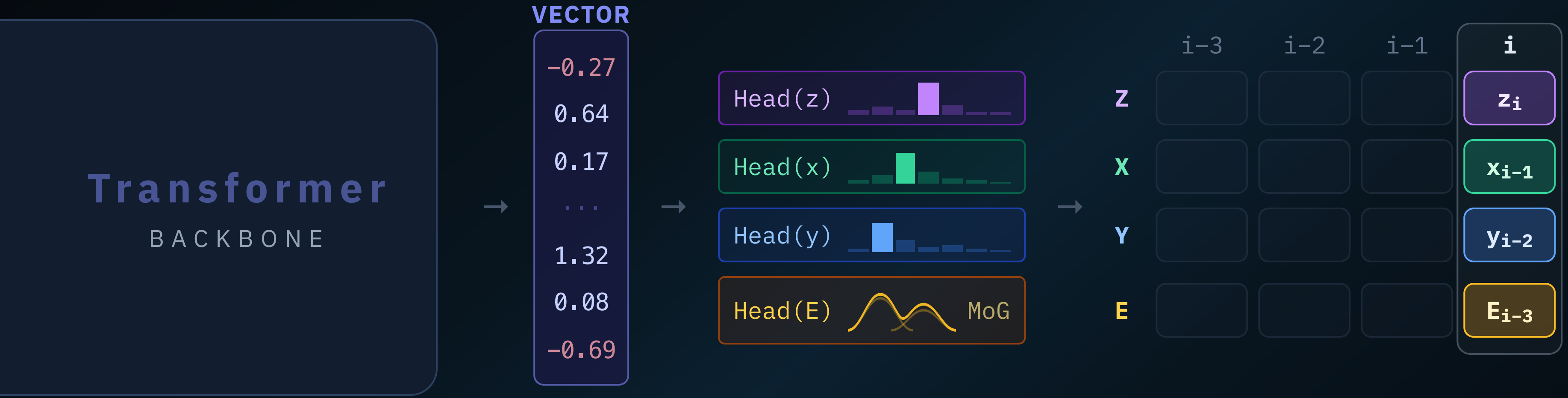
One Generation Step

S P D E
SPLIT & DELAY



One hidden state, four heads: categorical for z, x and y, a mixture of Gaussians for E.

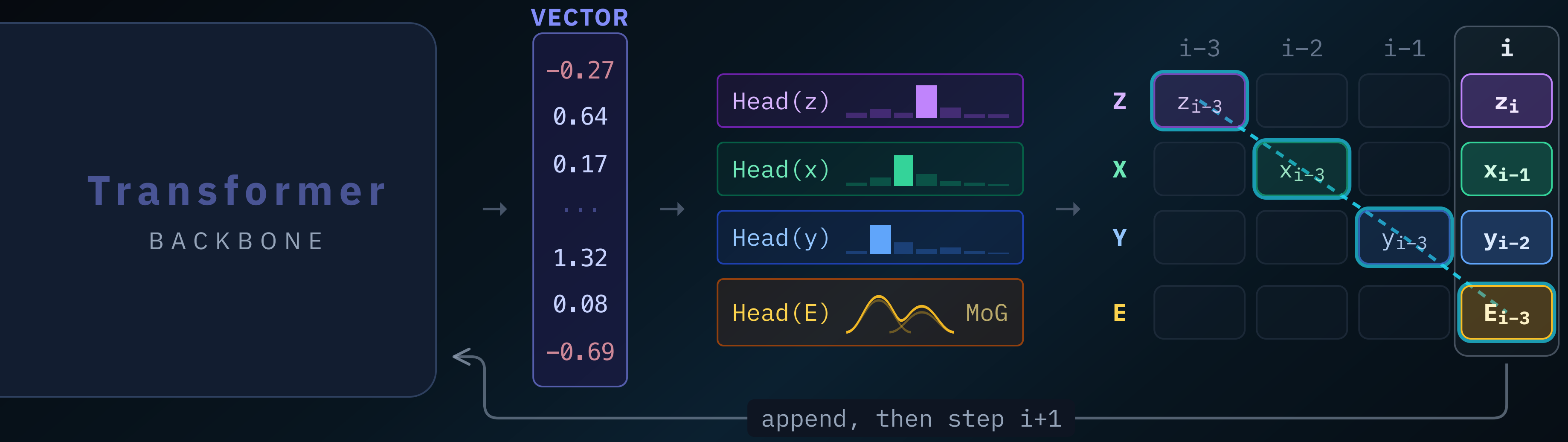
One Generation Step



One column, four hits. Each value emitted in this step belongs to a different hit.

One Generation Step

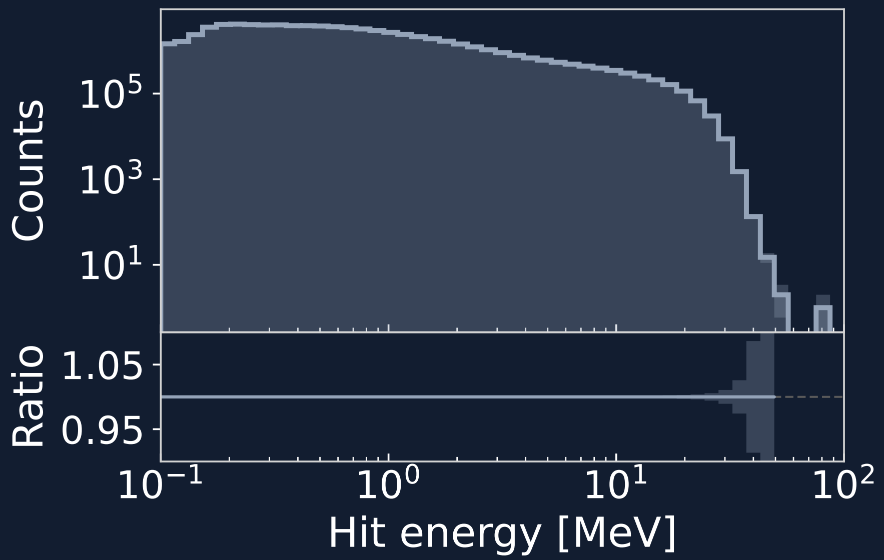
SPLIT & DELAY



One hit completes per step. A shower of N hits takes roughly N steps instead of 4N.

Energy Fidelity: Hit Spectrum

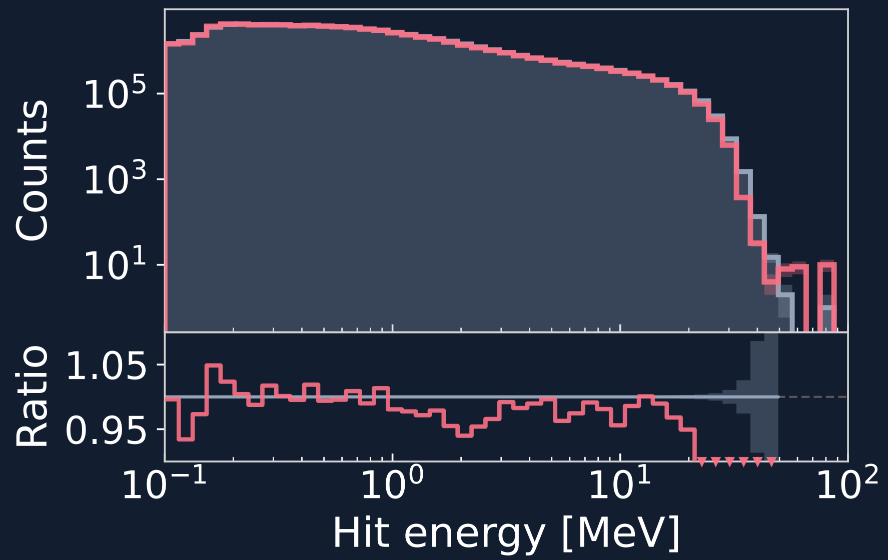
SPADE
SPLIT & DELAY



Geant4 TRUTH

Full detector simulation. All ratios are relative to it.

Energy Fidelity: Hit Spectrum



[3] Birk et al. 2501.05534 (2025)

Geant4 TRUTH

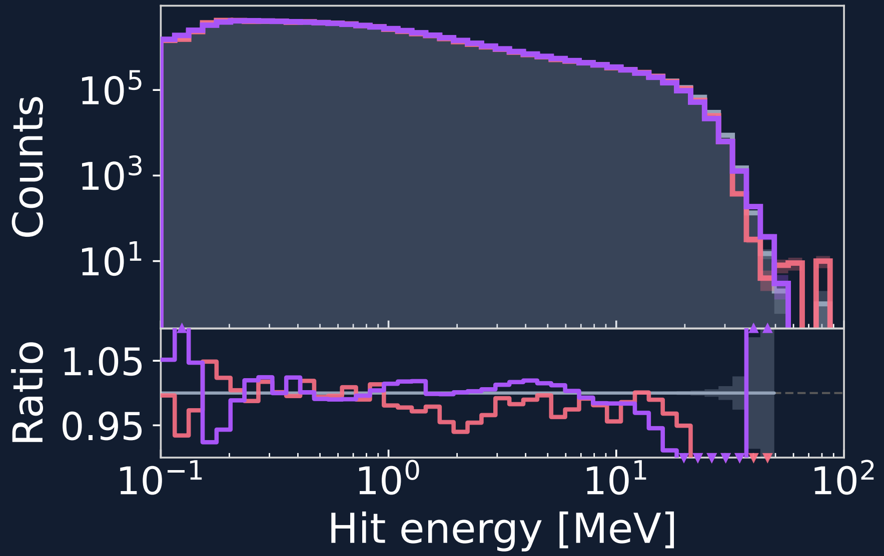
Full detector simulation. All ratios are relative to it.

OmniJet-α_C [3] TOKENIZED

Autoregressive like SPADE, but a **VQ-VAE codebook** turns each hit into a discrete token first.

Energy Fidelity: Hit Spectrum

SPADE
SPLIT & DELAY



[3] Birk et al. 2501.05534 (2025)

[4] Buss et al. 2601.11716 (2026)

Geant4 TRUTH

Full detector simulation. All ratios are relative to it.

OmniJet-α_C [3] TOKENIZED

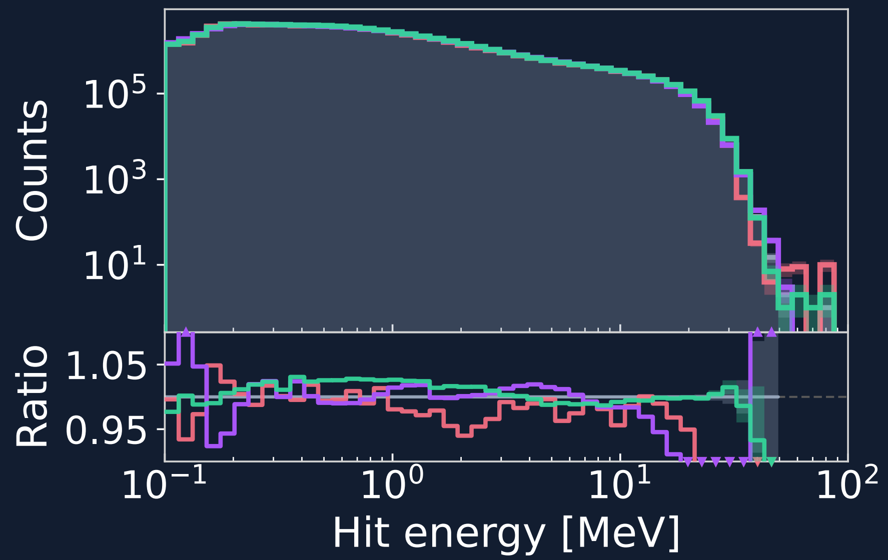
Autoregressive like SPADE, but a **VQ-VAE codebook** turns each hit into a discrete token first.

AllShowers [4] STATE OF THE ART

Conditional flow matching on the continuous hits, without tokens.

Energy Fidelity: Hit Spectrum

S P D E
SPLIT & DELAY



[3] Birk et al. 2501.05534 (2025)

[4] Buss et al. 2601.11716 (2026)

Geant4 TRUTH

Full detector simulation. All ratios are relative to it.

OmniJet-α_C [3] TOKENIZED

Autoregressive like SPADE, but a **VQ-VAE codebook** turns each hit into a discrete token first.

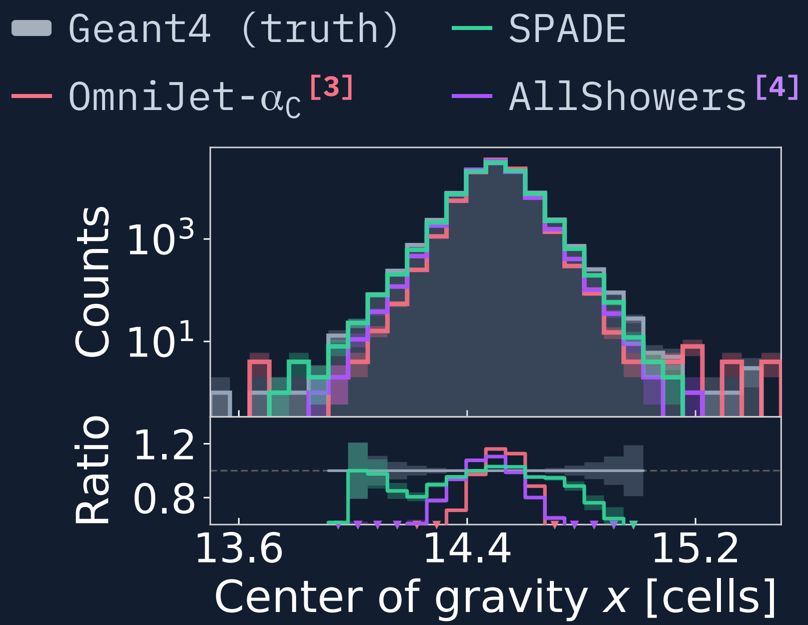
AllShowers [4] STATE OF THE ART

Conditional flow matching on the continuous hits, without tokens.

SPADE THIS WORK

Autoregressive with **no codebook**. Its energy head is continuous, never quantized.

Spatial Fidelity: Center of Gravity



[3] Birk et al. 2501.05534 (2025)

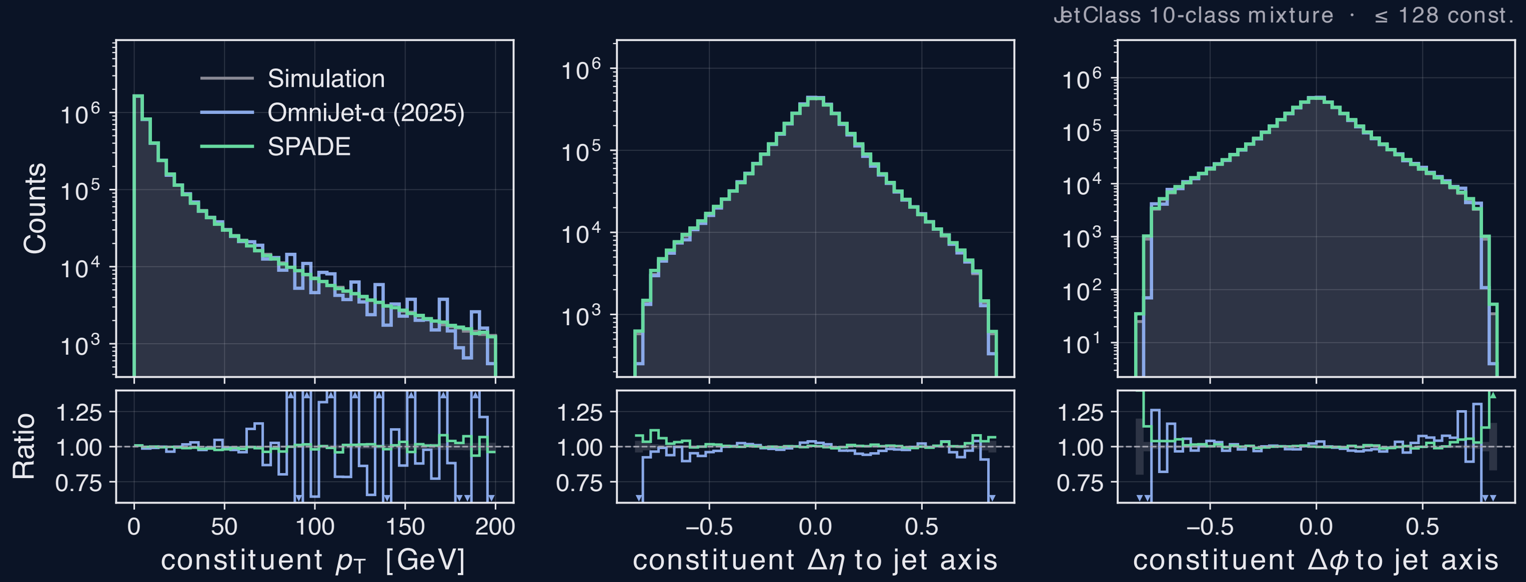
[4] Buss et al. 2601.11716 (2026)

RATIO TO GEANT4

- **Core:** SPADE within a few percent, OmniJet-α_C and AllShowers about 10 to 15% high
- **Tails:** all three fall below Geant4, SPADE follows it furthest out

Jets: Kinematics

SPLIT & DELAY

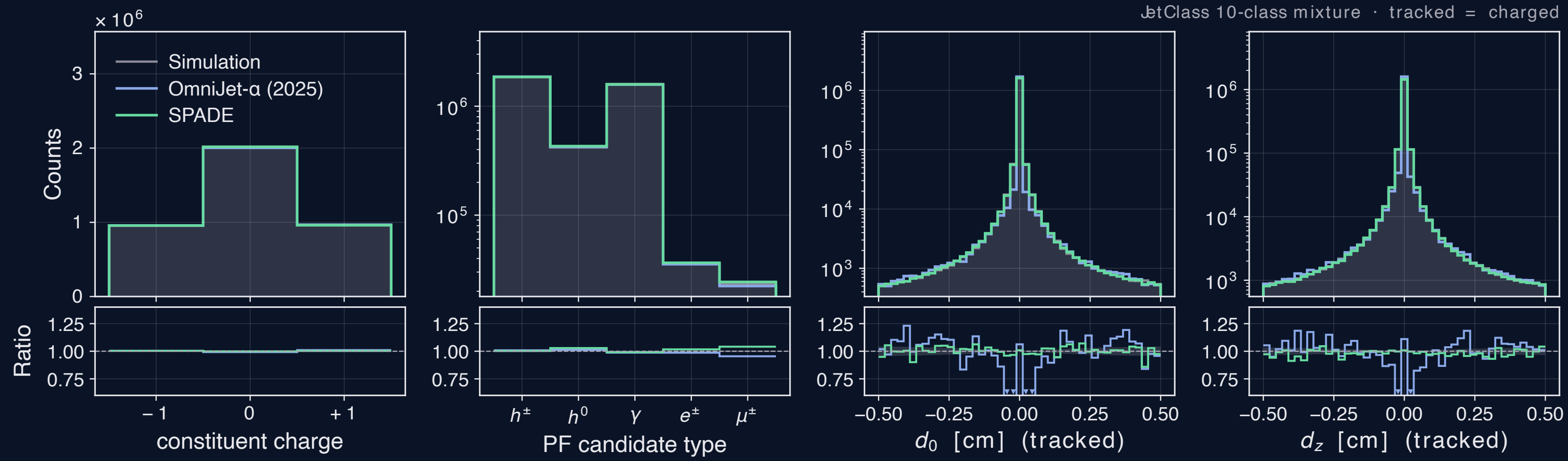


Preliminary. Same method as for the showers, with jet constituents in place of hits.

Baseline in blue: OmniJet-α (2025), a different model from the OmniJet-α_C on the shower slides.

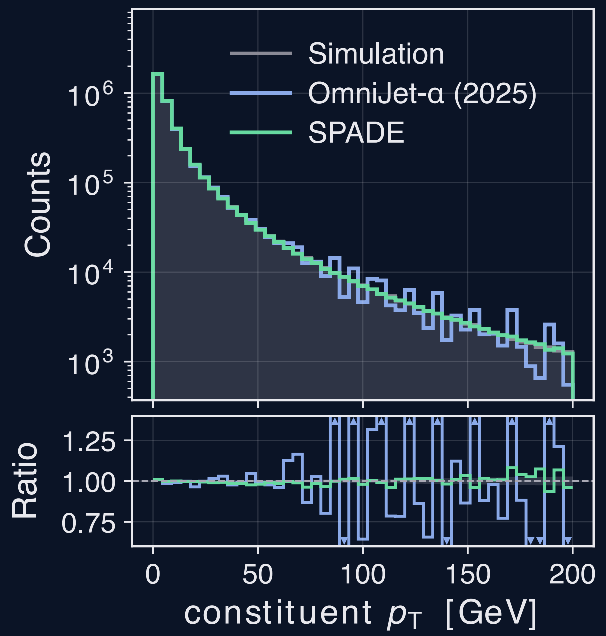
Jets: Beyond Kinematics

SPADE
SPLIT & DELAY



Preliminary. Discrete features get their own small vocabularies, continuous ones a continuous head, as for the showers.

Conclusion



SPLIT

Each feature gets its own small vocabulary.

DELAY

Every feature is predicted with increasing delays, and the sequence stays short.