MASKED-TOKEN PREDICTION FOR ANOMALY DETECTION AT THE LHC

Ambre Visive

Nikhef & University of Amsterdam

Worked realized in collaboration with Roberto R de Austri, Polina Moskvitina, Sascha Caron and Clara Nellist

Nikhef

Universiteit van Amsterdam

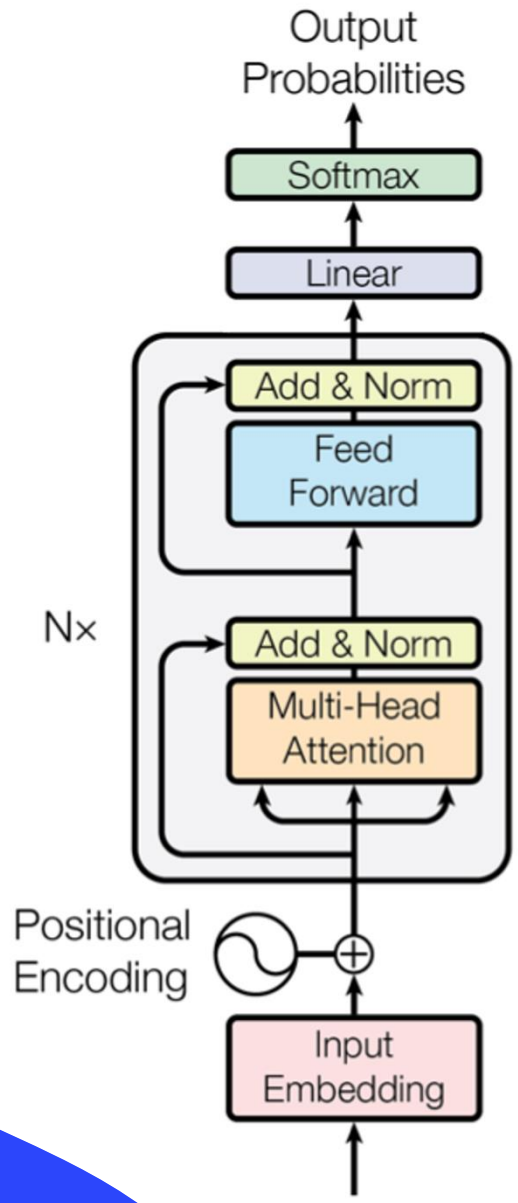
TABLE OF CONTENT

- Large-Language Model & Tokenization
- Dataset & Input Data
- Reflections on the Global Strategy
- Tokenization Strategies
- Results & Outlooks
- Conclusion

LARGE-LANGUAGE MODEL & TOKENIZATION

LLMs, Tokens, and Other Things We Should Probably Define

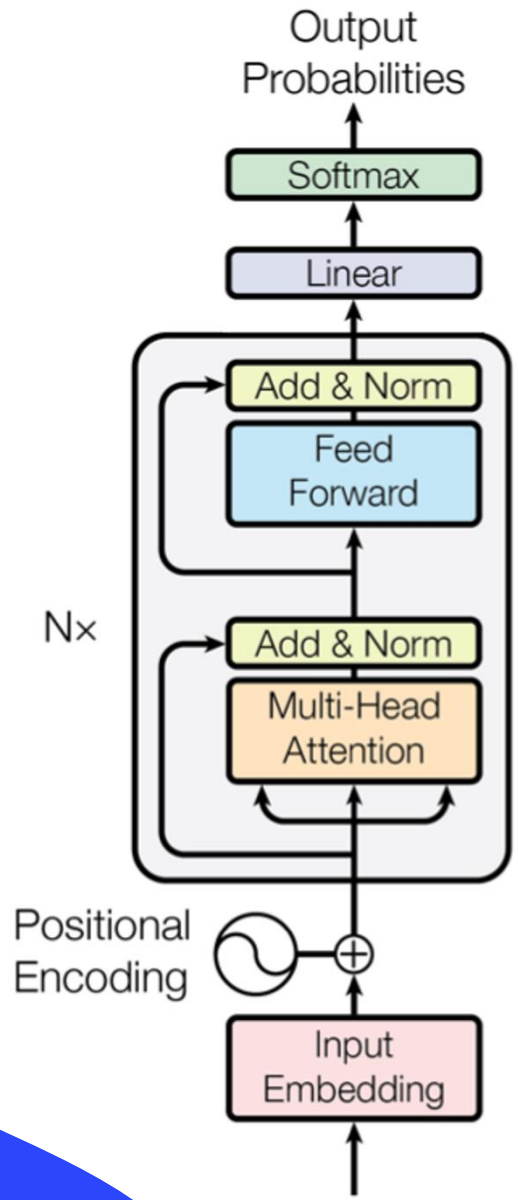
LARGE LANGUAGE MODEL



Modified from [1]

Transformer model that uses tokens as input

LARGE LANGUAGE MODEL



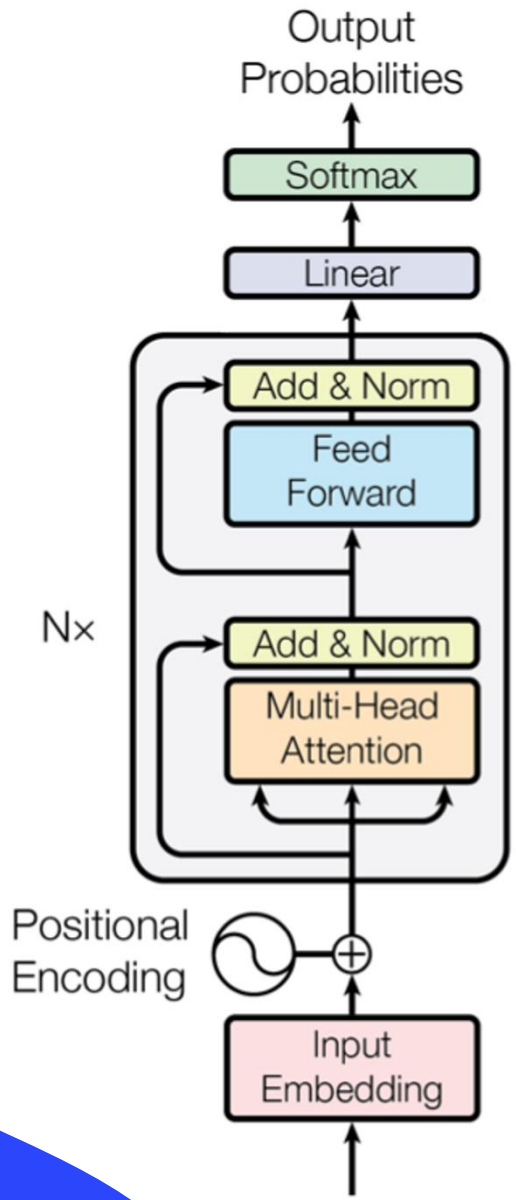
Modified from [1]

Transformer model that uses tokens as input

Very good at learning the data distribution

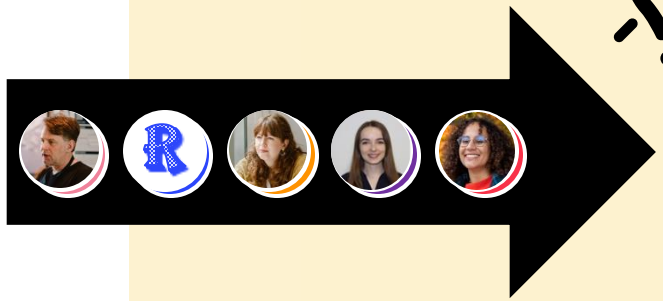
Modified from [1]

LARGE LANGUAGE MODEL



Transformer model that uses tokens as input

Very good at learning the data distribution



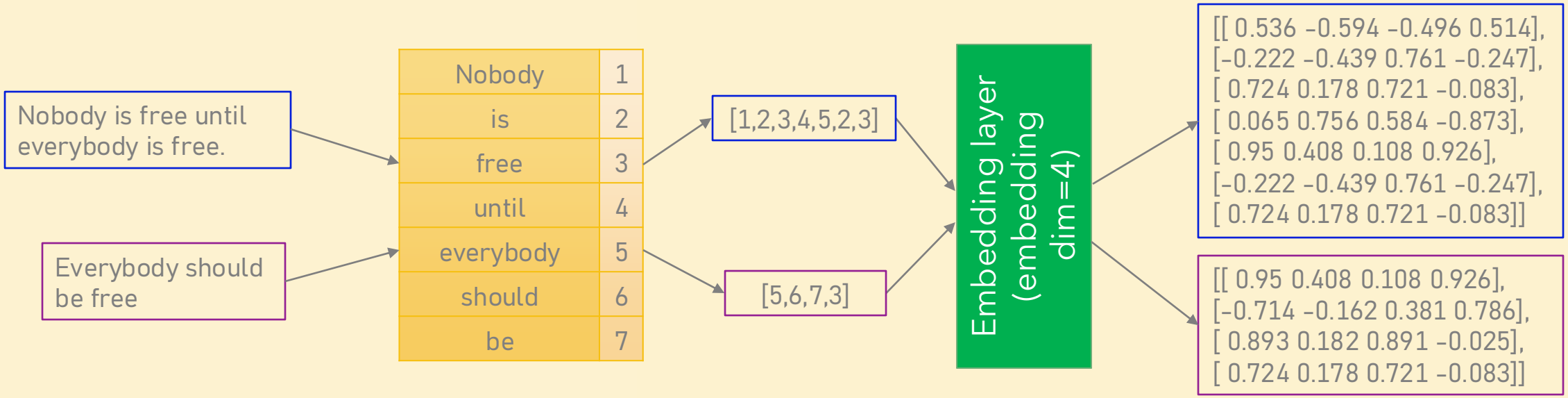
Build the encoder part of a simple LLM-like model and make it learn the distribution of background data so it can flag anomalous events

TOKENIZATION

Key process of LLMs

Whole idea = convert complicated data into a representation that the model can understand **(= token)**

Model knows that there is a limited number of possible tokens (Think dictionary → limited number of words)

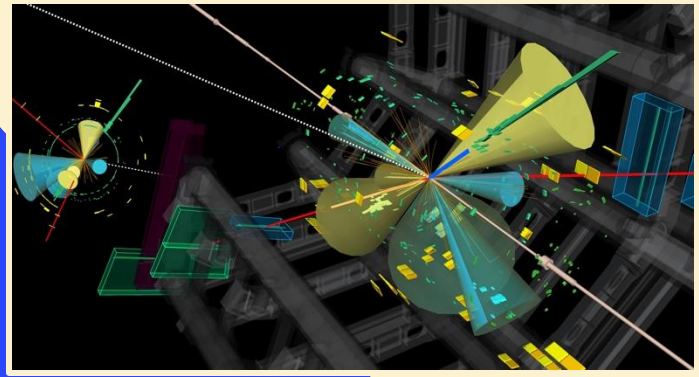


Thanks to this process, the model can process complicated data which have meaning to us.

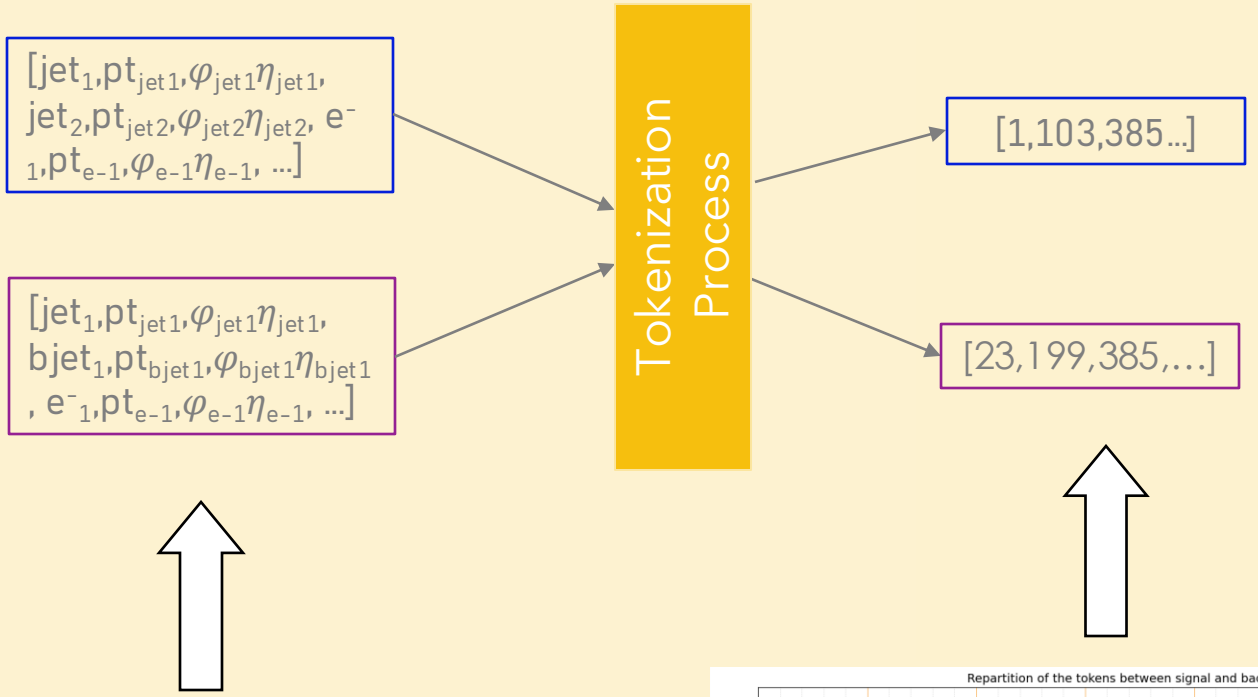
TOKENIZATION

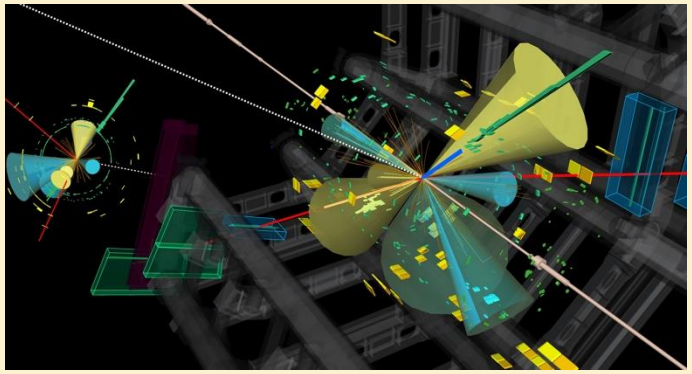
$[jet_1, pt_{jet1}, \varphi_{jet1} \eta_{jet1}, jet_2, pt_{jet2}, \varphi_{jet2} \eta_{jet2}, e^-_1, pt_{e-1}, \varphi_{e-1} \eta_{e-1}, \ldots]$

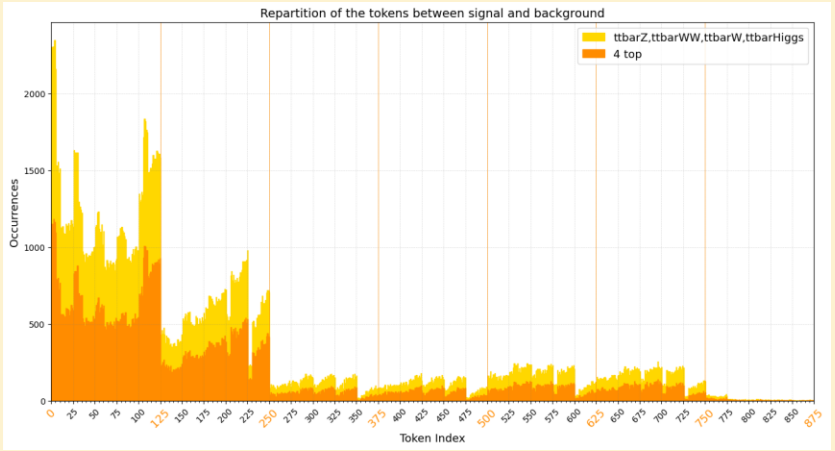
$[jet_1, pt_{jet1}, \varphi_{jet1} \eta_{jet1}, bjet_1, pt_{bjet1}, \varphi_{bjet1} \eta_{bjet1}, e^-_1, pt_{e-1}, \varphi_{e-1} \eta_{e-1}, \ldots]$



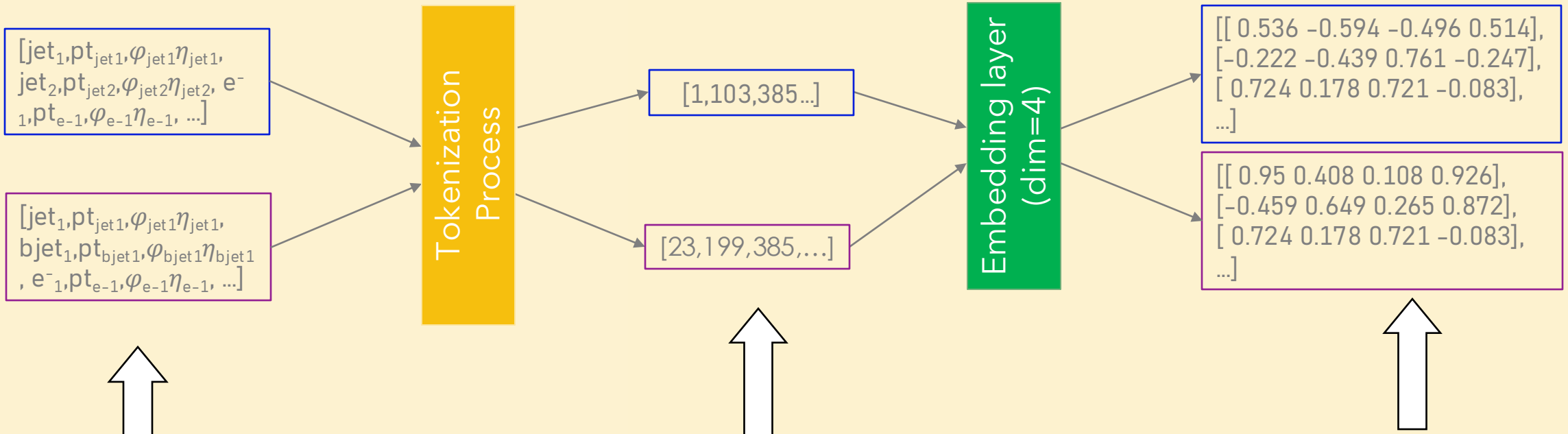
TOKENIZATION

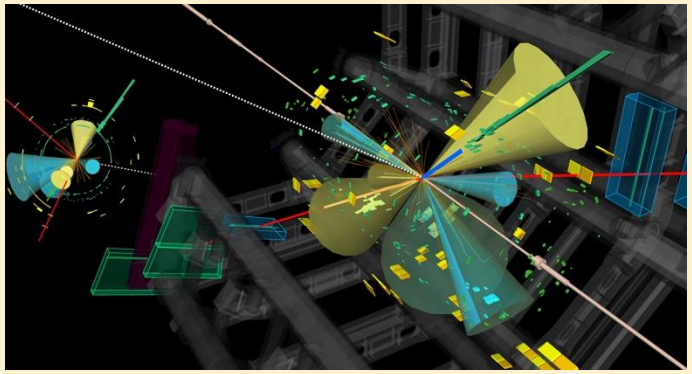


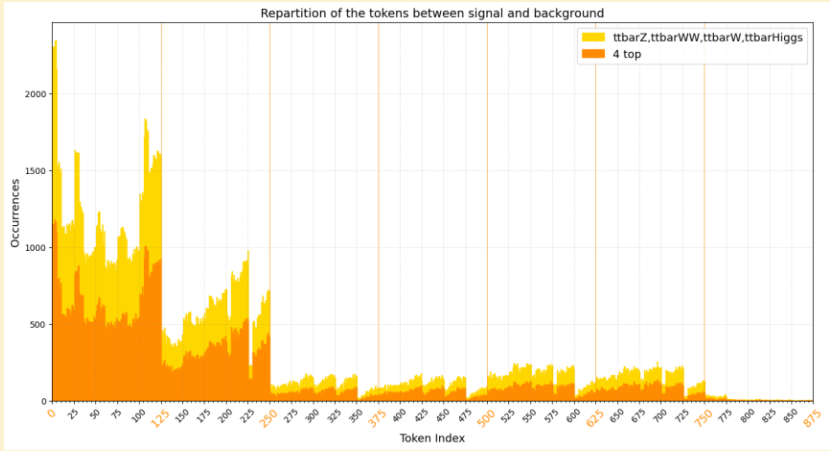


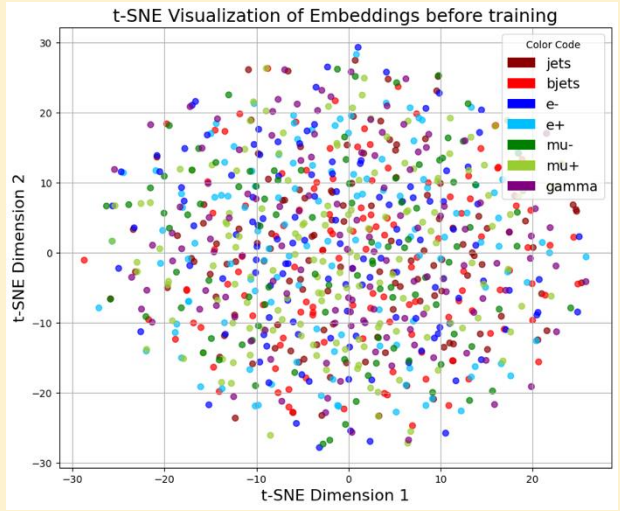


TOKENIZATION

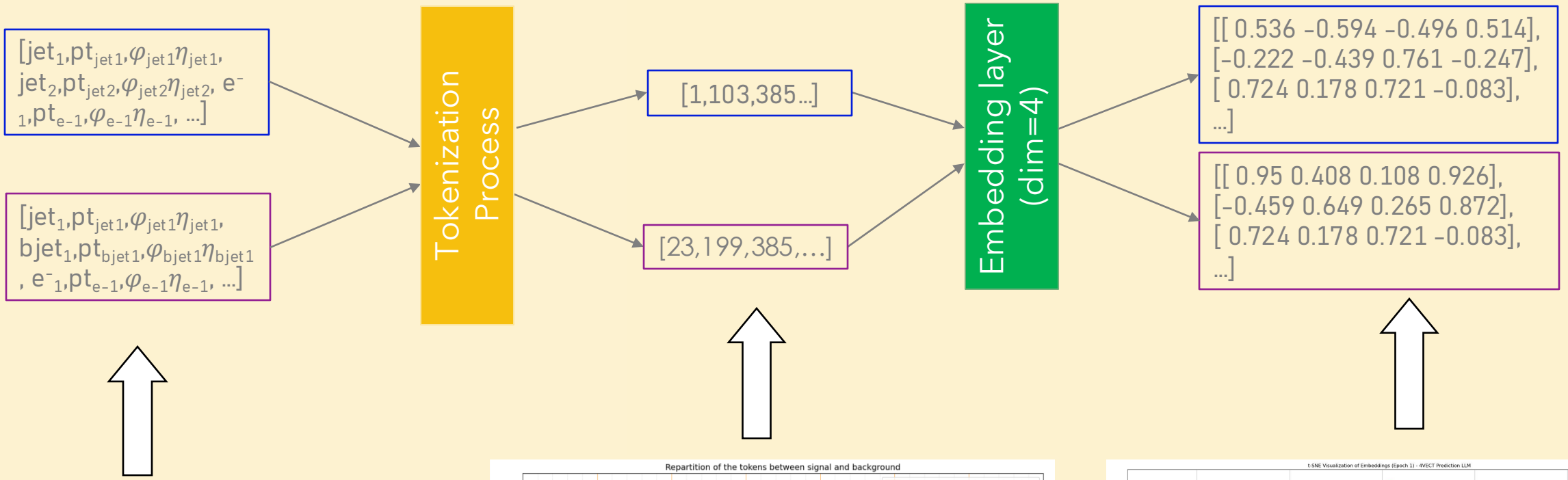


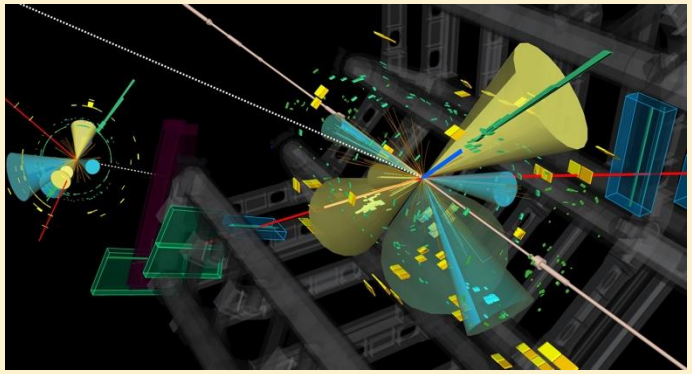


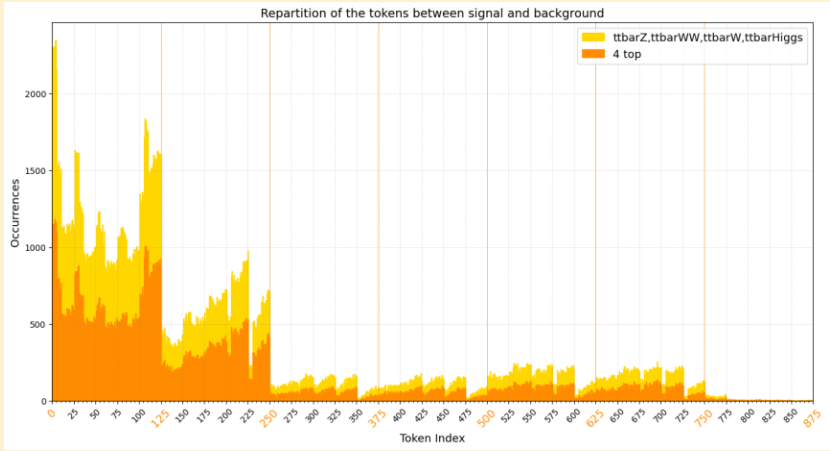


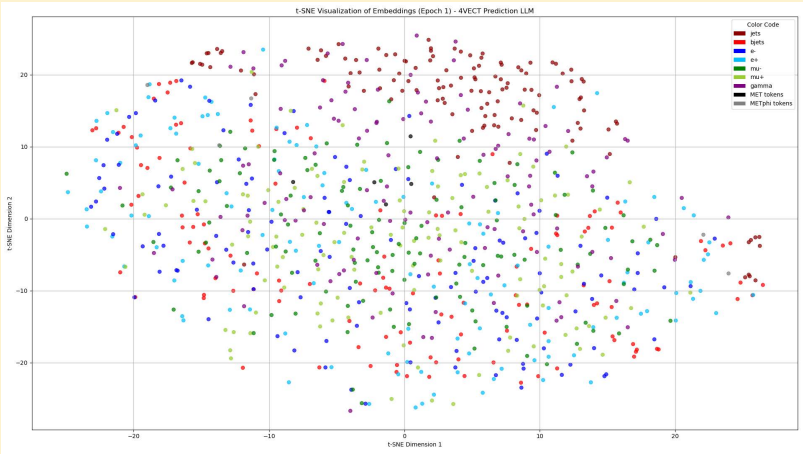


TOKENIZATION





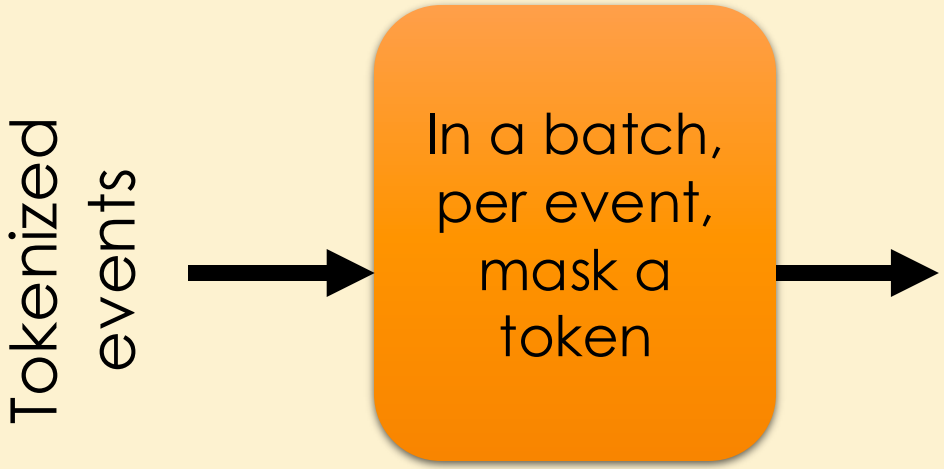




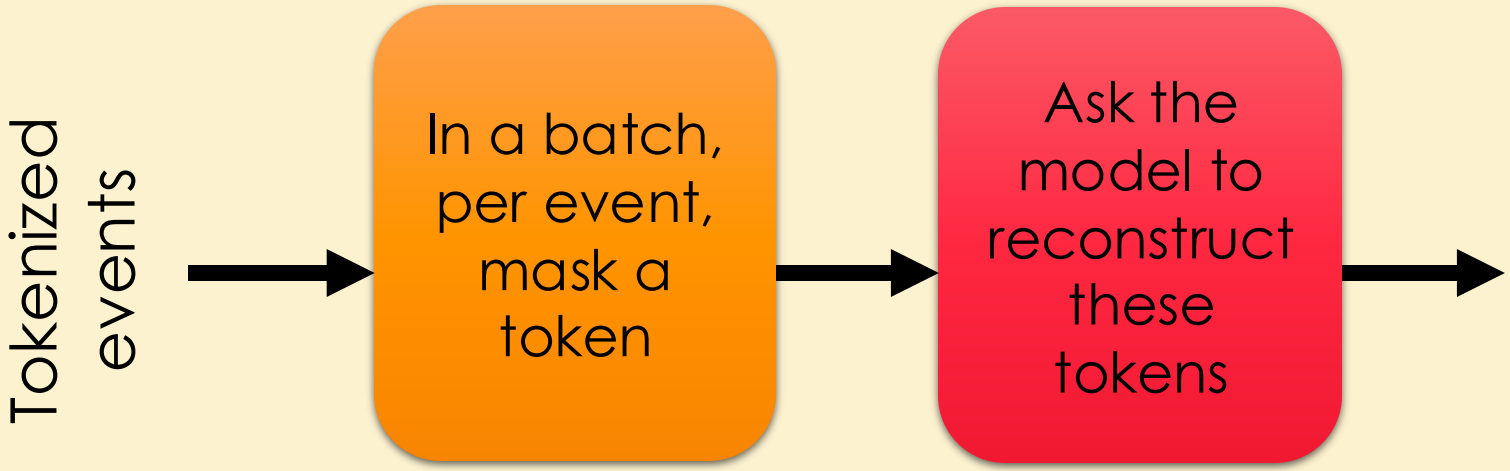
TRAINING PROCESS

Tokenized events →

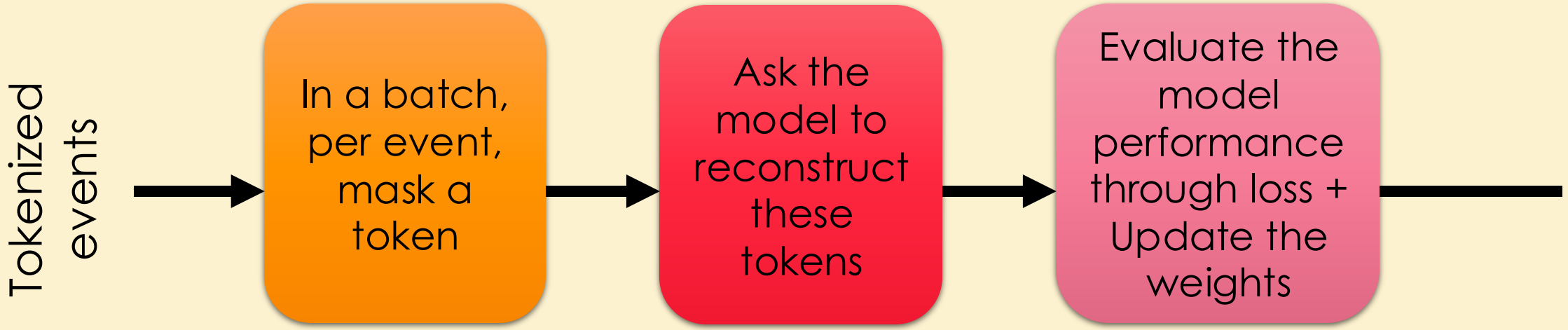
TRAINING PROCESS



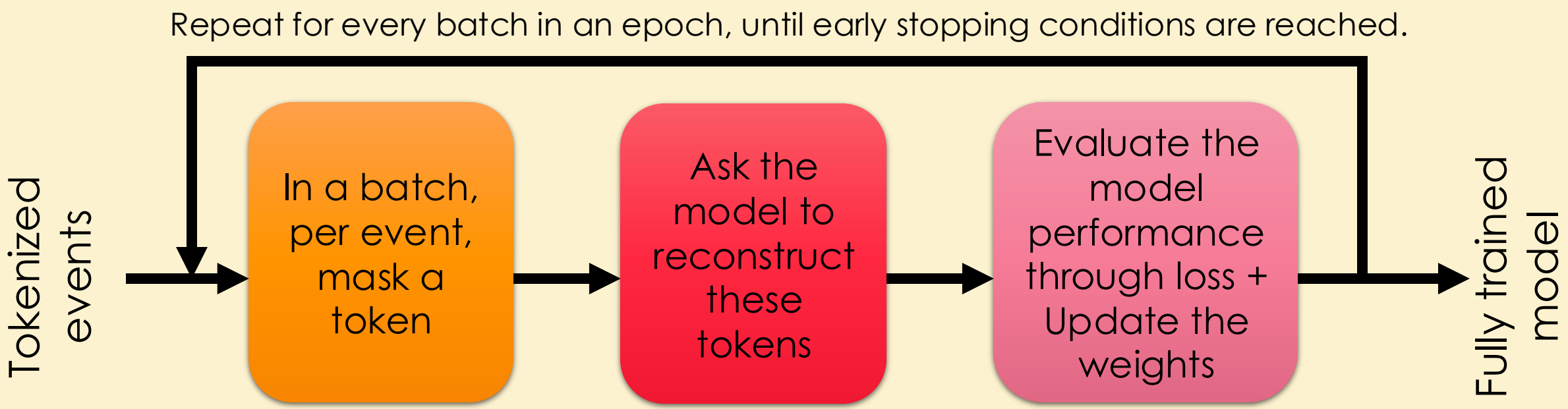
TRAINING PROCESS



TRAINING PROCESS



TRAINING PROCESS



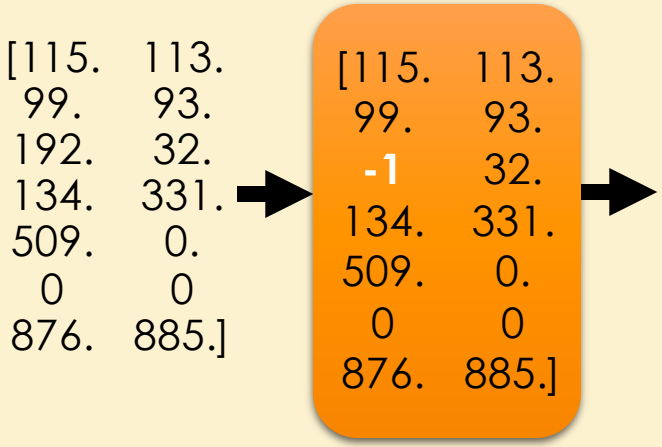
TRAINING PROCESS

Example for a batch of 1 event

[115. 113.
99. 93.
192. 32.
134. 331. ➡
509. 0.
0 0
876. 885.]

TRAINING PROCESS

Example for a batch of 1 event

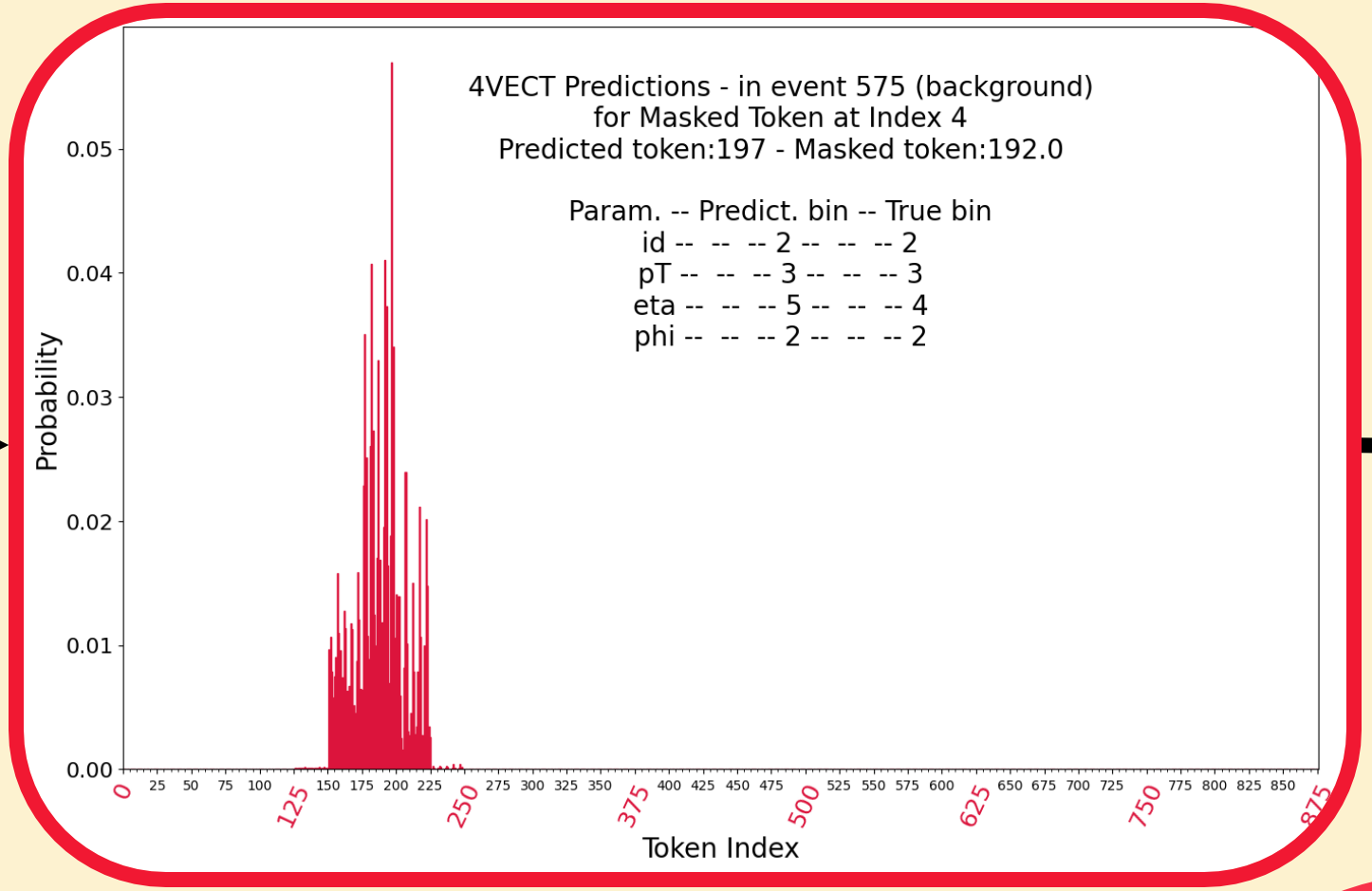


TRAINING PROCESS

Example for a batch of 1 event

[115. 113.
99. 93.
192. 32.
134. 331.
509. 0.
0 0
876. 885.]

[115. 113.
99. 93.
-1 32.
134. 331.
509. 0.
0 0
876. 885.]

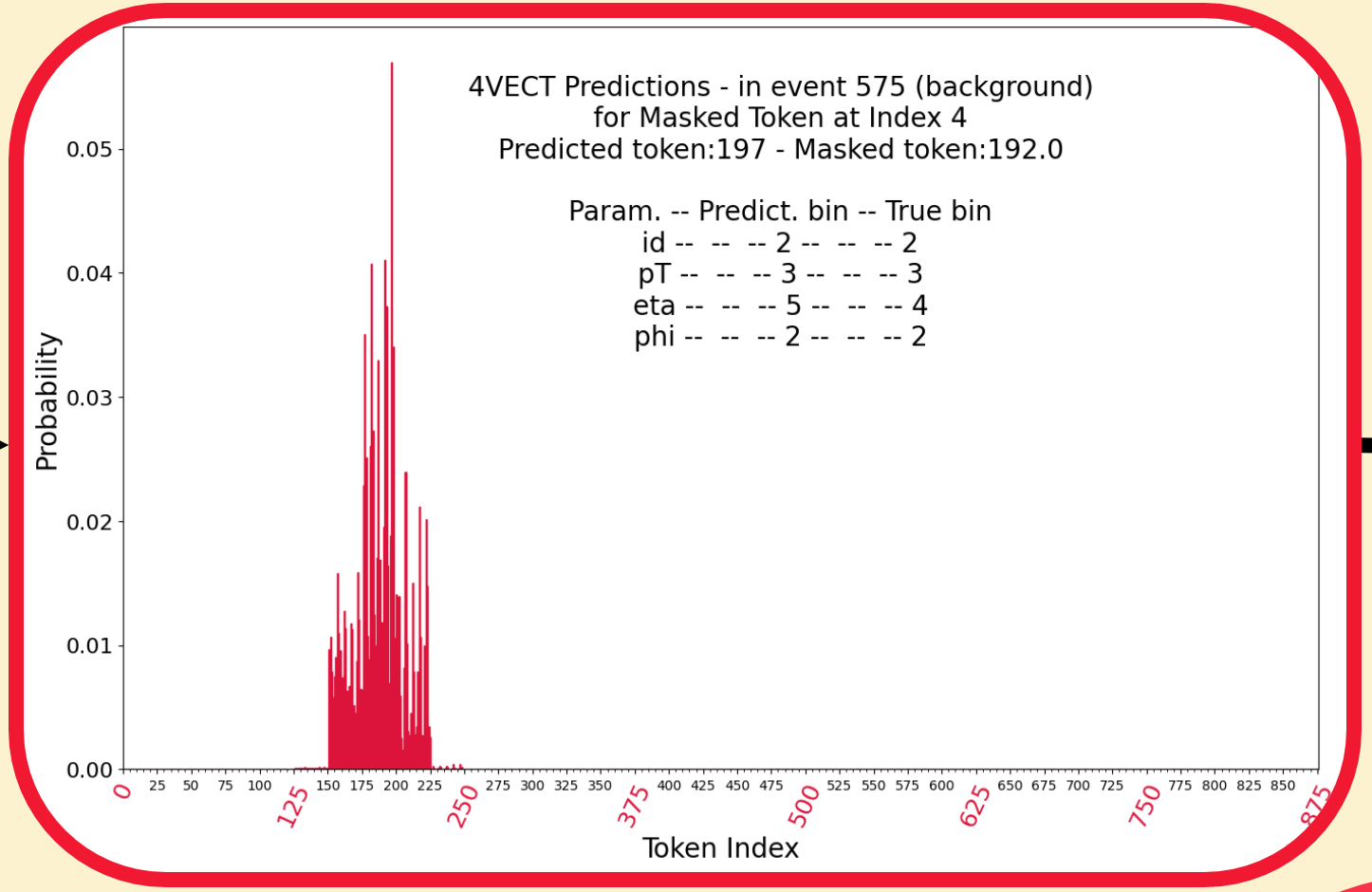


TRAINING PROCESS

Example for a batch of 1 event

[115. 113.
99. 93.
192. 32.
134. 331.
509. 0.
0 0
876. 885.]

[115. 113.
99. 93.
-1 32.
134. 331.
509. 0.
0 0
876. 885.]

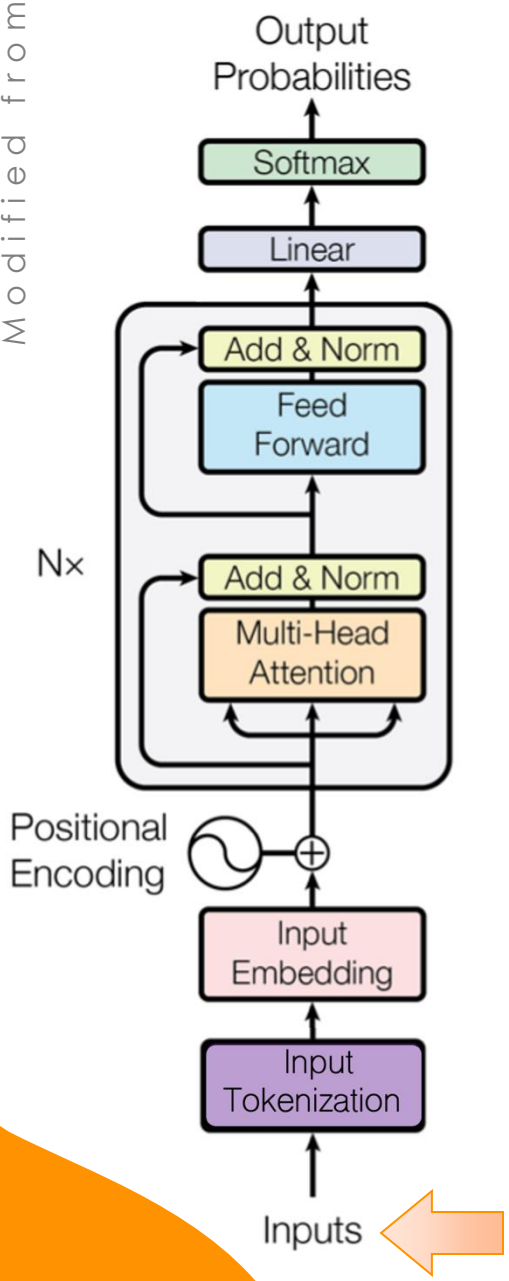


Loss = Sparse Categor. Cross-Entropy (**pred_proba**, true_token)

DATASET & INPUT DATA

Which cats should we use to train our model?

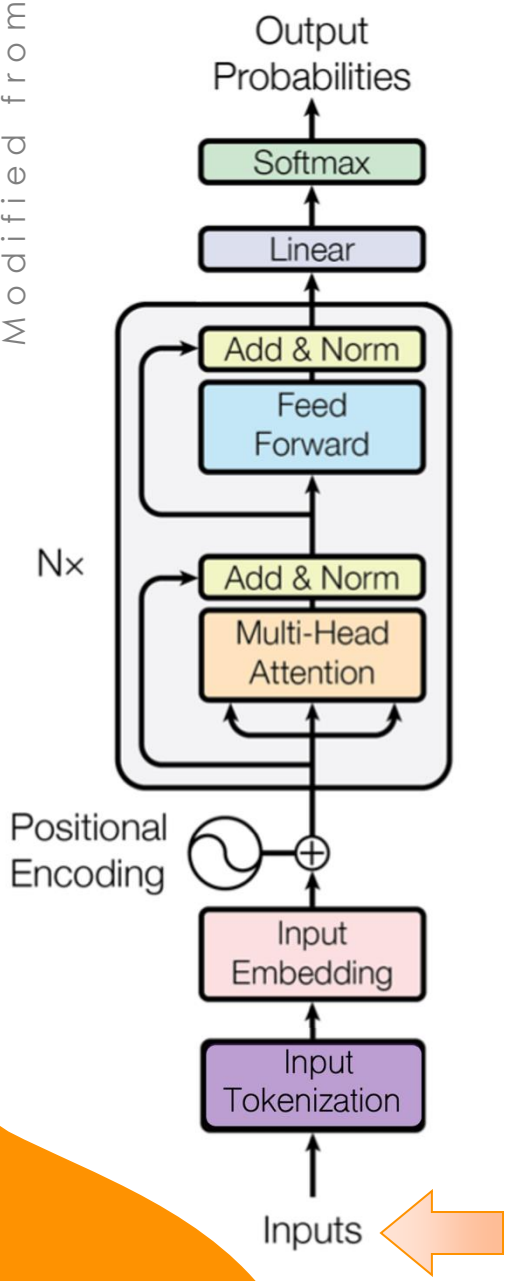
Modified from [1]



INPUT OF THE MODEL

- a sequence of tokens → can represent a particle-type (and charge), its pt, its φ and its η, the MET, METφ of the event…

Modified from [1]



INPUT OF THE MODEL

actually 120k of them

- **a sequence of tokens** → can represent a particle-type (and charge), its pt, its φ and its η, the MET, METφ of the event…

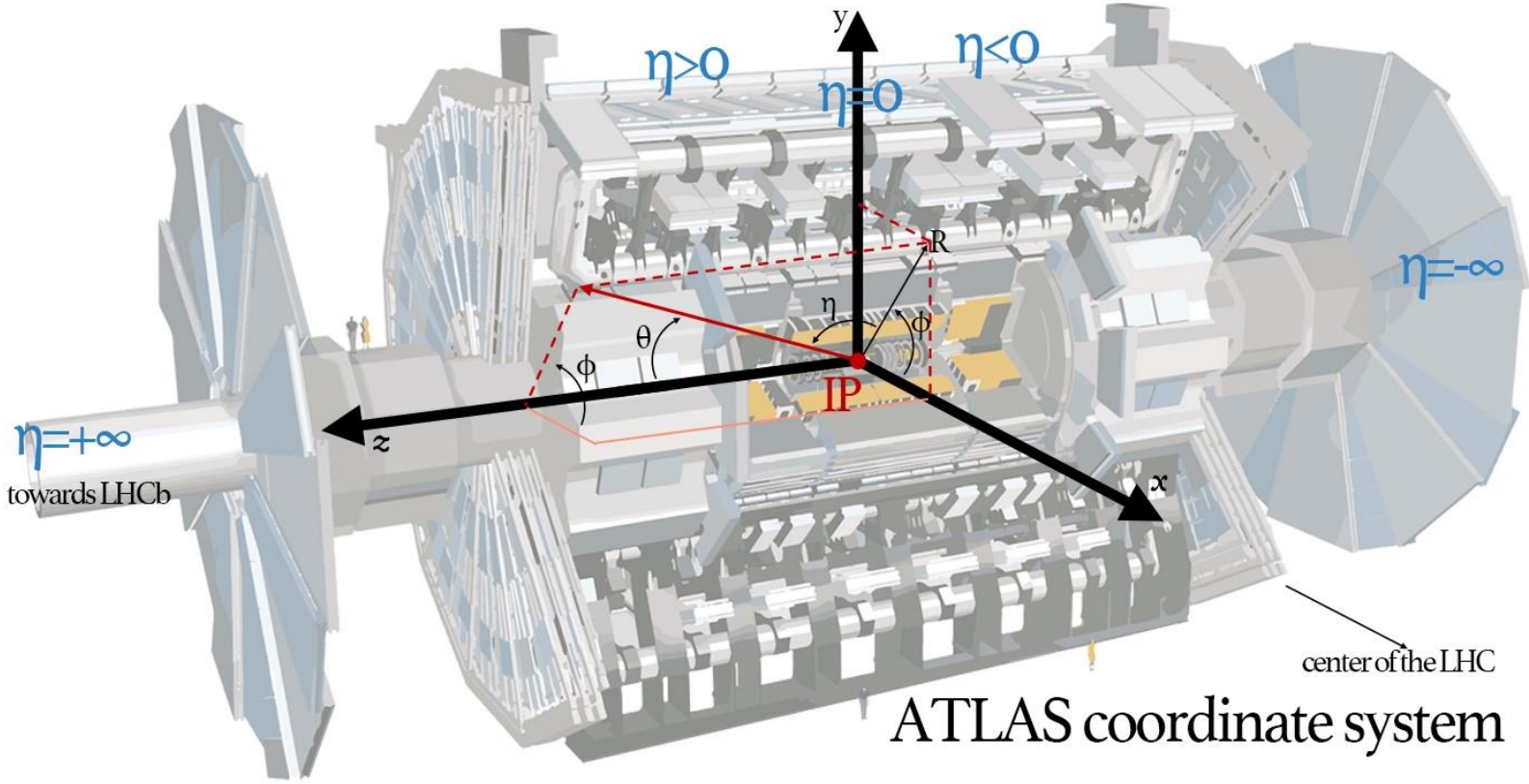
"simulated ATLAS data"

- **SM "background" dataset of the Dark Machines Challenge**[3]
  - ttW, ttWW, ttZ, ttH processes
  - events are labeled and p_t ordered by particle-type[4]

(QUICK) TYPOLOGY

MET = (magnitude of) missing transverse energy

MET_{φ} = azimuthal angle of MET



Modified from [2]

4-VECT = (energy, p_t, φ, η)

p_t = transverse mometum

4-VECT = (id*, p_t, φ, η)

id = particle-type + its charge

REFLECTIONS ON THE GLOBAL STRATEGY

How to build an anomaly detector that will take us to Stockholm

OUR IDEA

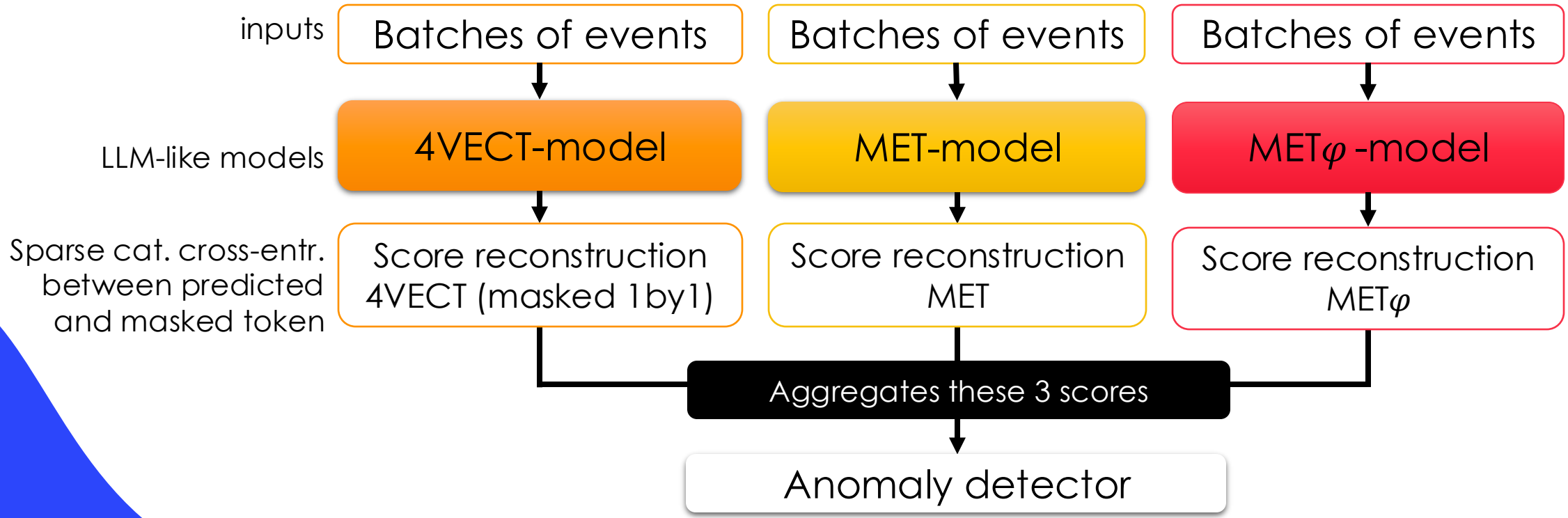
4VECT-model

MET-model

METφ-model

During Training

OUR IDEA



At Inference

OUR IDEA

→ REFLEXIONS & CHOICES

4VECT-model

- Should MET/METφ be included in the events?
- How to tokenize the 4-vectors?

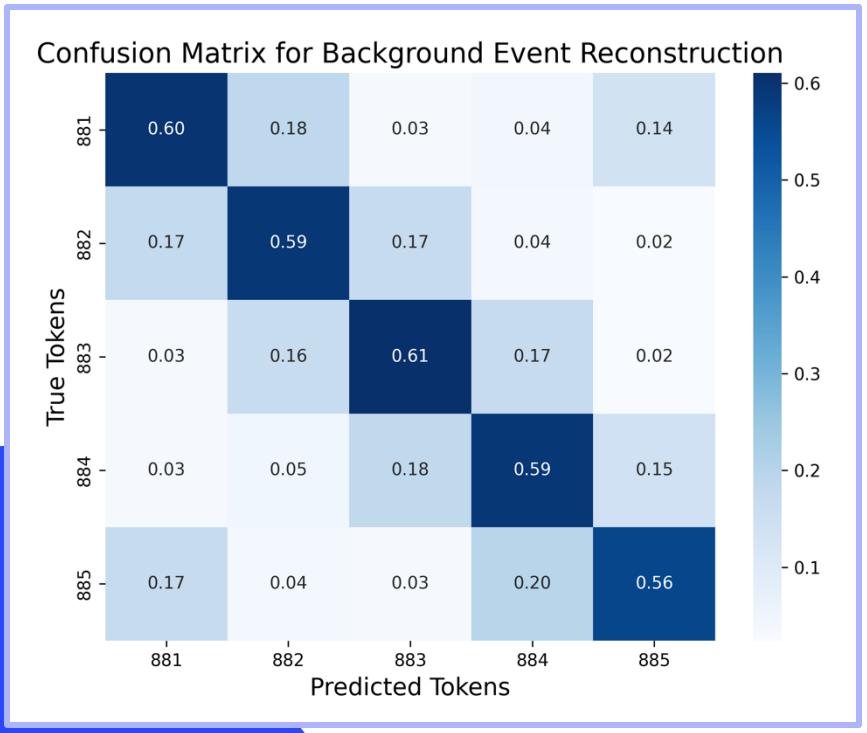
MET-model

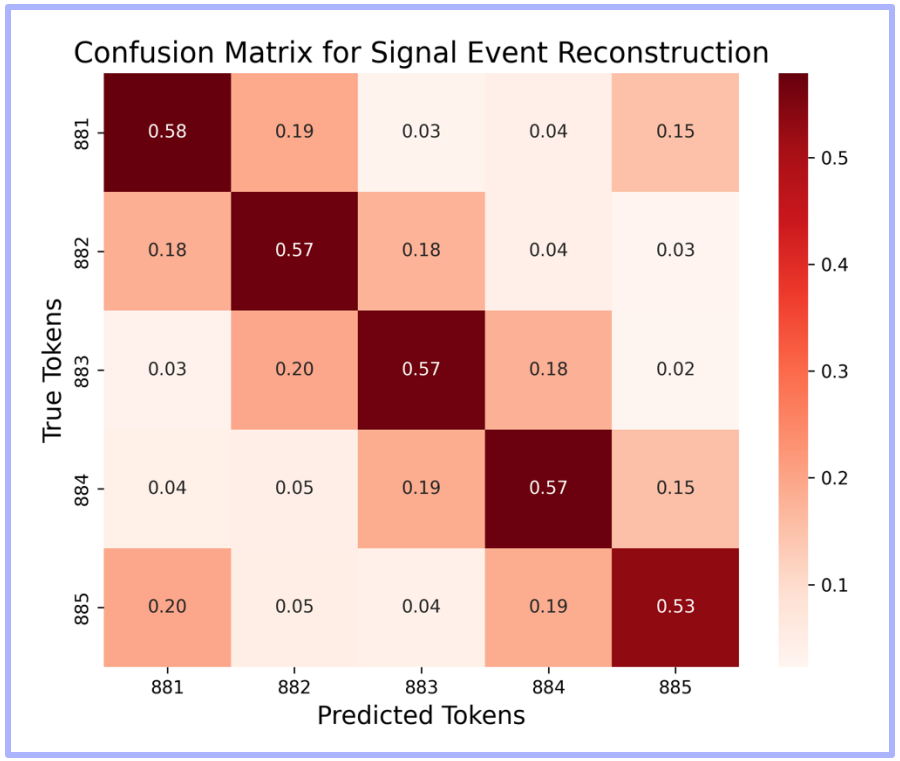
- Should METφ be included in the events?
- How to tokenize the MET?

METφ -model

- How to tokenize the METφ?

METφ MODEL





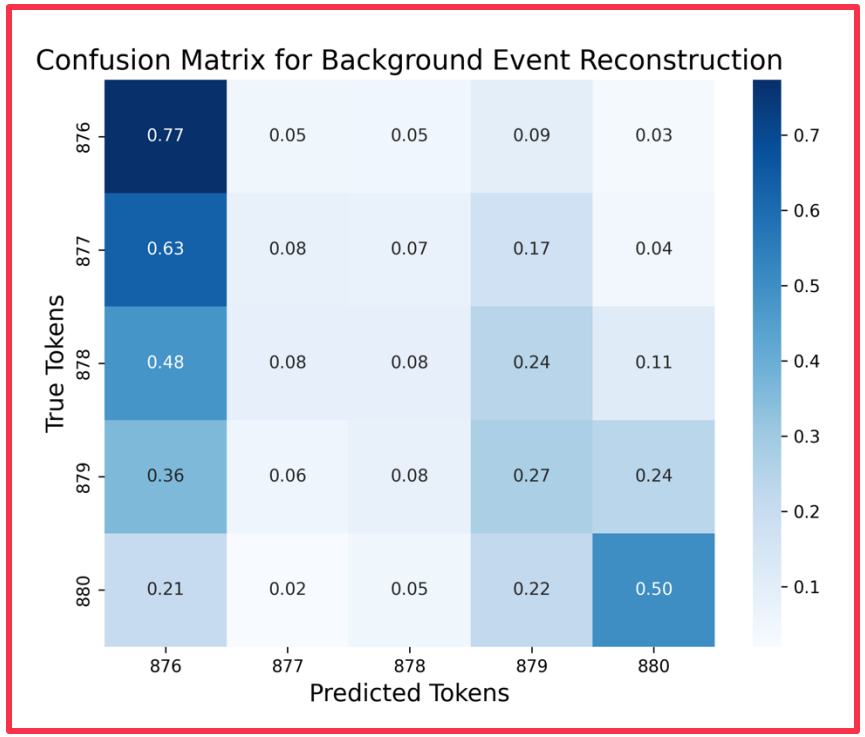
num epochs	dropout rate	ROC-AUC
6	0.1	0.5191
7	0.05	0.5207
7	0.05	0.5200
7	0.05	0.5202
7	0.05	0.5201
8	0.1	0.5191
9	0.1	0.5195
11	0.1	0.5187
11	0.1	0.5203
11	0.05	0.5198
12	0.05	0.5200
14	0.05	0.5197

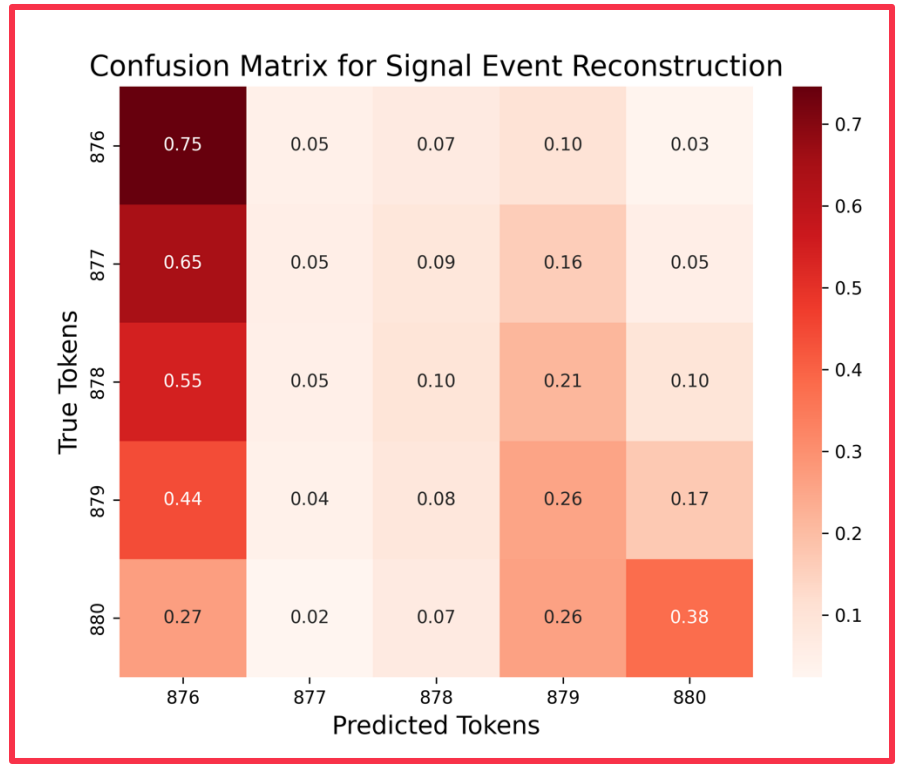
ROC curves calculated at inference for different models

MET MODEL

num epochs	dropout rate	ROC-AUC
6	0.1	0.5474
6	0.1	0.5475
6	0.05	0.5478
6	0.05	0.5468
14	0.1	0.5438
14	0.05	0.5456
28	0.05	0.5451
31	0.05	0.5445
35	0.05	0.5467

ROC curves calculated at inference for different models



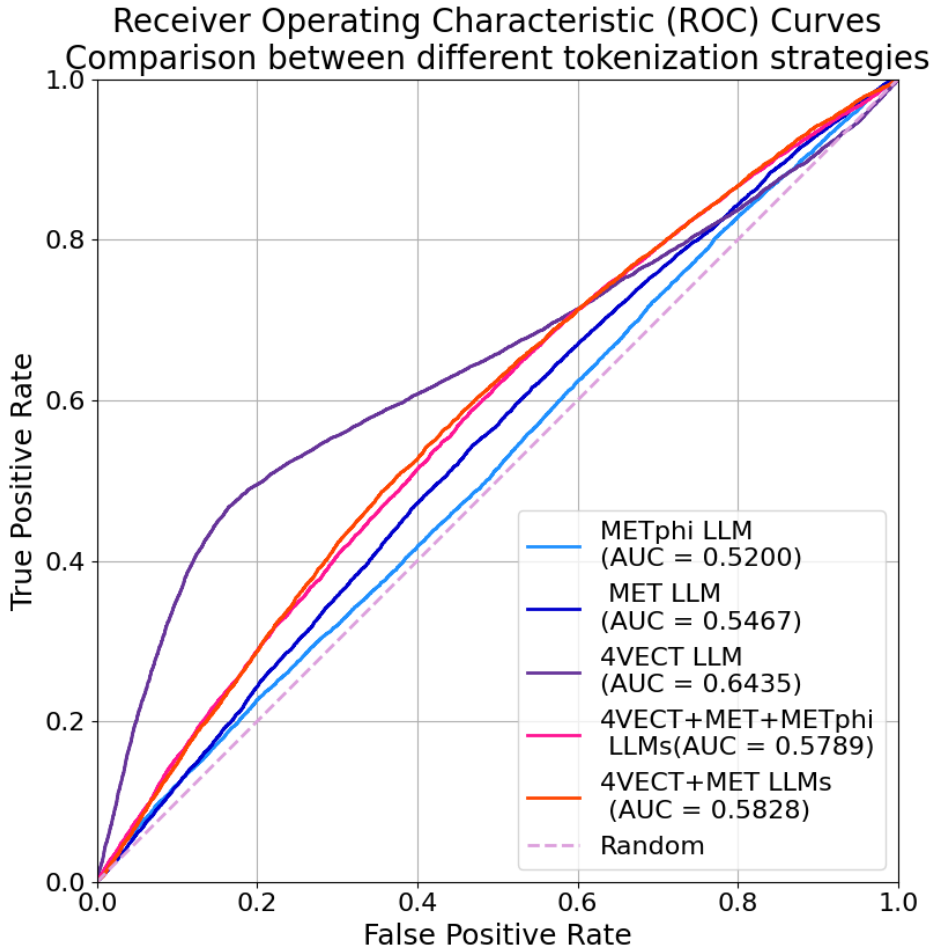


OUR IDEA

REFLECTIONS & CHOICES

Should abandon the MET and METφ specific models

Not performing well on their own → diminish the performance of the "global" model



TOKENIZATION STRATEGIES

Slicing collisions like a good comté: a token at a time

OVERVIEW
OF OUR TOKENIZATION STRATEGIES

- VQVAE
- GNN with VQ layer
- 8 “simple” (by-hand) tokenization

Should we add MET, METφ tokens?

OVERVIEW
OF OUR TOKENIZATION STRATEGIES

Should we add MET, METφ tokens?

- VQVAE
- GNN with VQ layer
- 8 “simple” (by-hand) tokenization → by **binning** the data

OVERVIEW
OF OUR TOKENIZATION STRATEGIES

Should we add MET, METφ tokens?

- VQVAE
- GNN with VQ layer
- 8 “simple” (by-hand) tokenization

by **binning** the data

- into how many bins?
- how are bin edges defined?

VQVAE

inspired by [6]

Input sequence

Encoder Network

Continuous Latent Representation

Vector Quantization

Discrete Latent Representation (codebook indices)

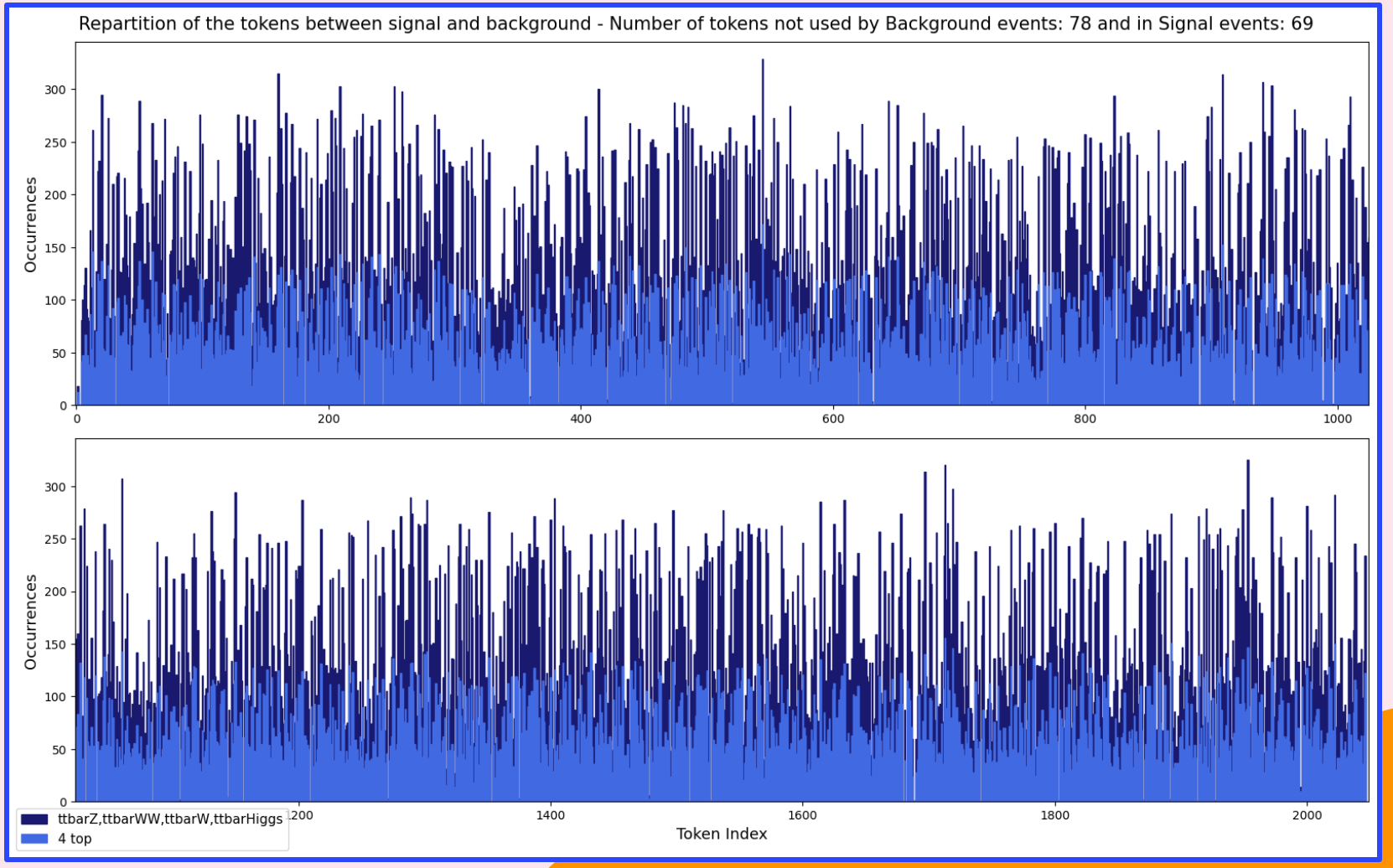
Decoder Network

Reconstructed sequence

[5]

Tokens for our model

MET+METφ included in the sequence as a token



VQVAE

inspired by [6]

Input sequence

↓

Encoder Network

↓

Continuous Latent Representation

↓

Vector Quantization

↓

Discrete Latent Representation (codebook indices)

↓

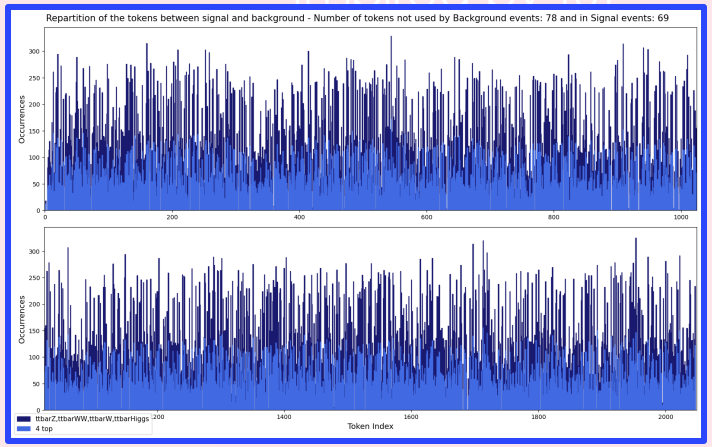
Decoder Network

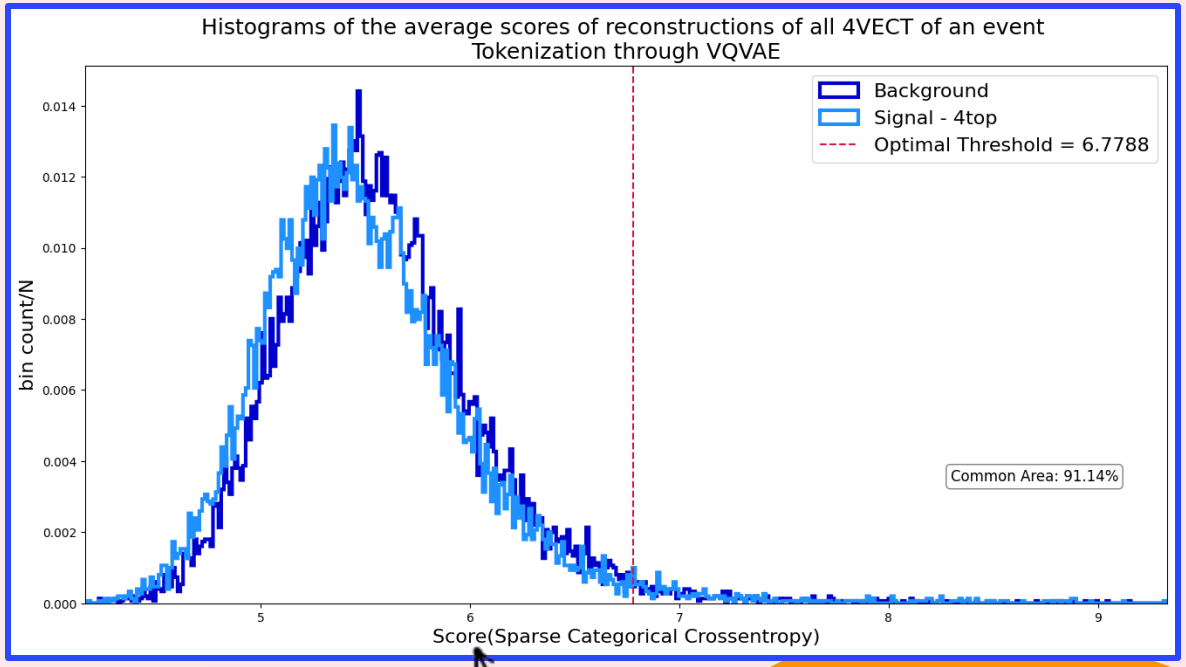
↓

Reconstructed sequence

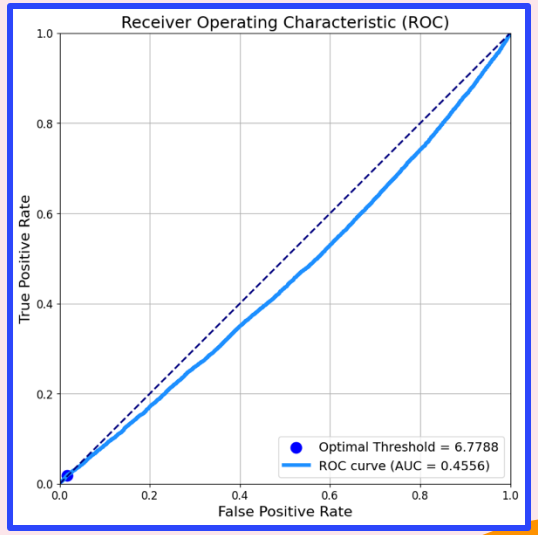
[5]

Tokens for our model





= How good is the token reconstruction



GNN

Input Graphs (nodes = 4VECT, edges = inv. Mass)

3 GATConv* Layers

Continuous Latent Representation

Vector Quantization

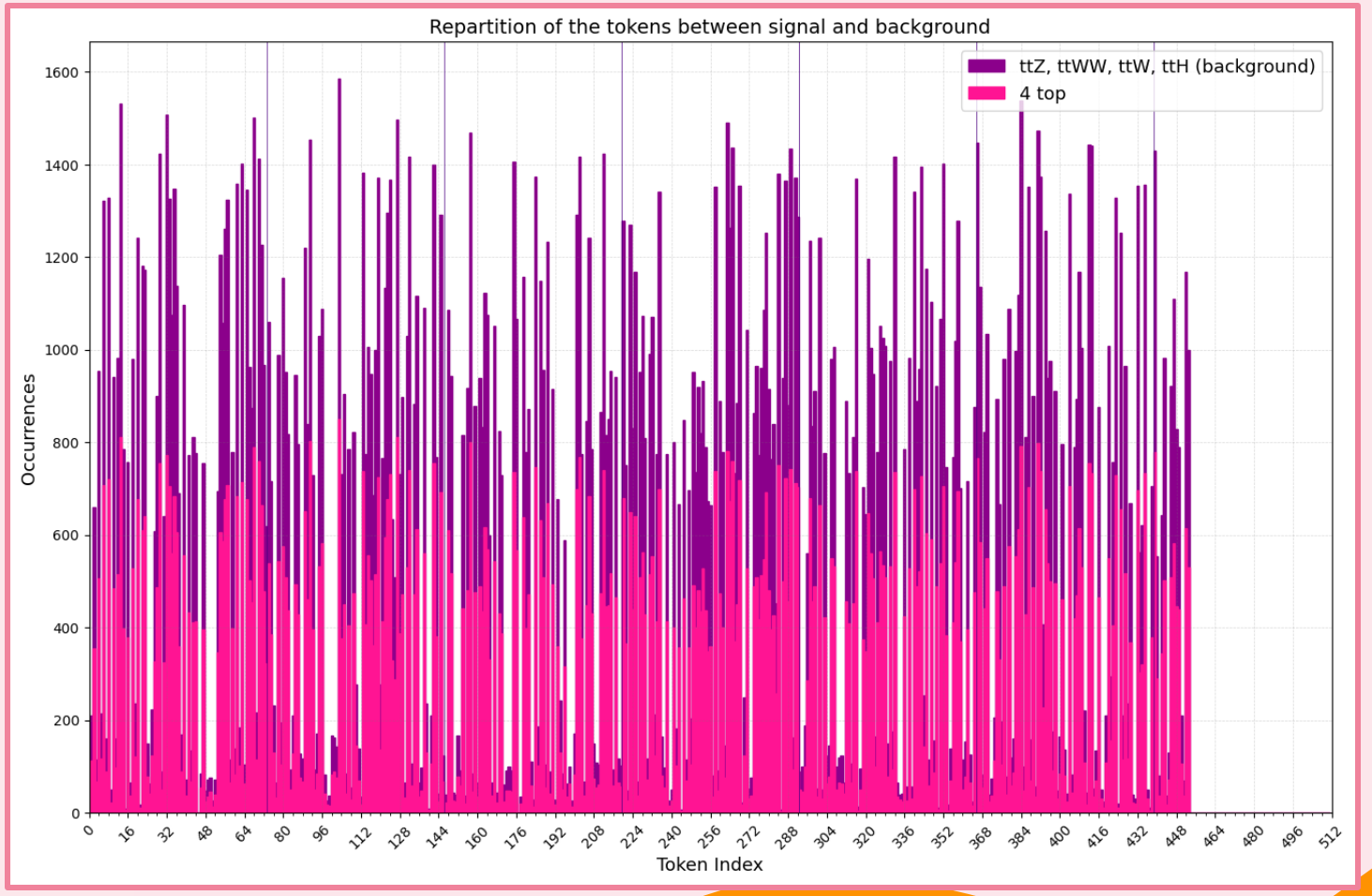
Discrete Latent Representation (codebook indices)

Classification Head

Output (2 classes)

Tokens for our model

codebook_size = 512 { used tokens in test dataset: 452
used tokens in train dataset: 453



*Graph Attention Convolution (layer from Graph Attention Network Architecture [7])

GNN

Input Graphs (nodes = 4VECT, edges = inv. Mass)

3 GATConv* Layers

Continuous Latent Representation

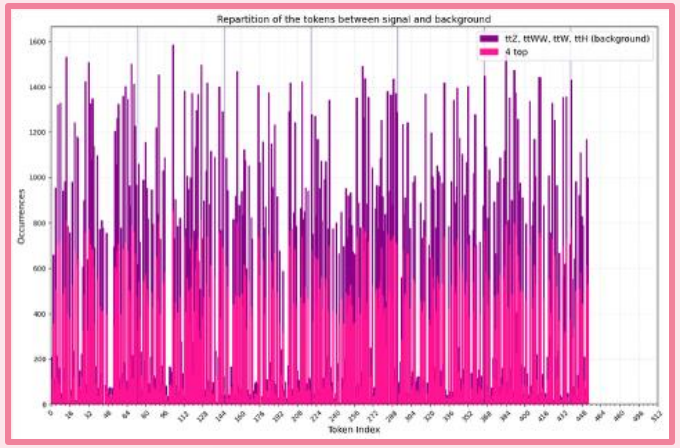
Vector Quantization

Discrete Latent Representation (codebook indices)

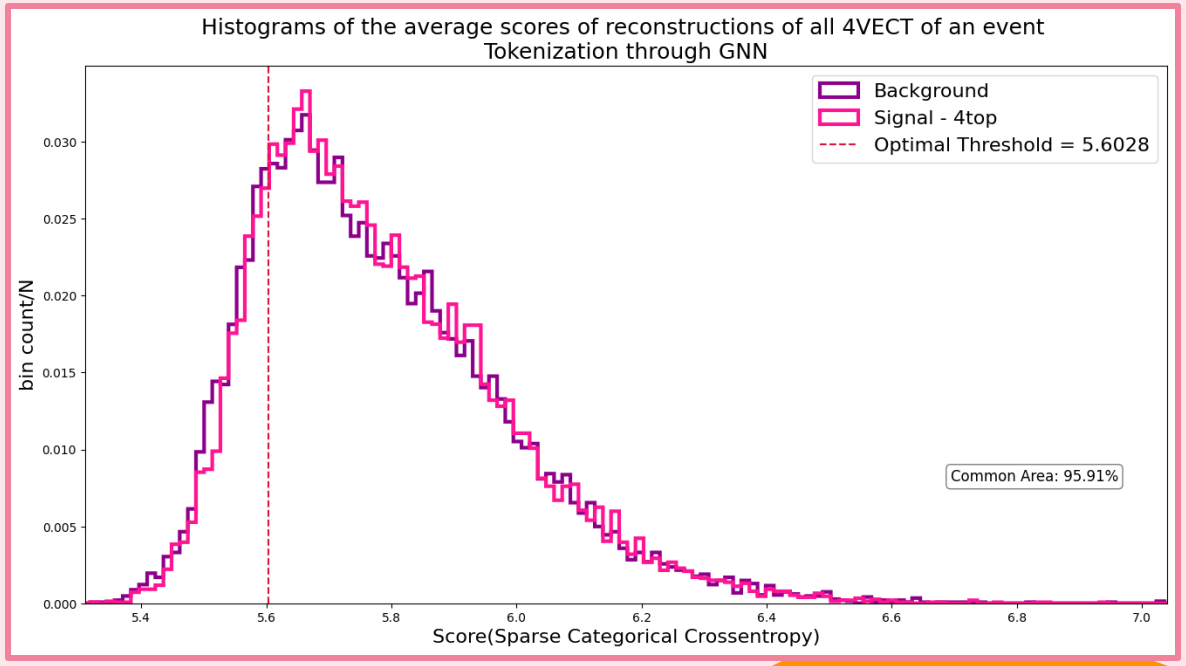
Classification Head

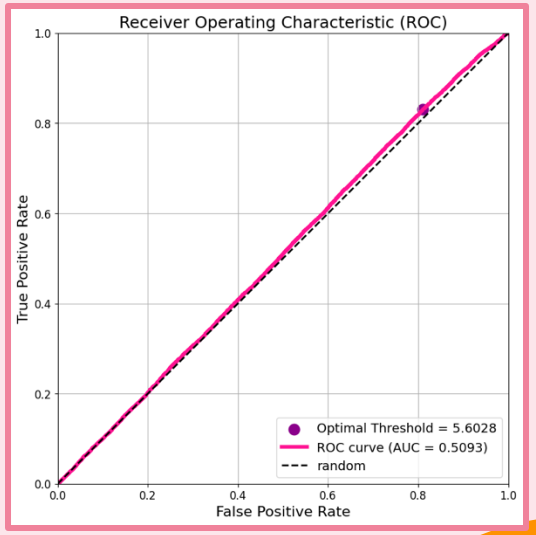
Output (2 classes)

Tokens for our model



- Codebook size: 512
- Num tokens used in test: 452
- Num tokens used in train: 453

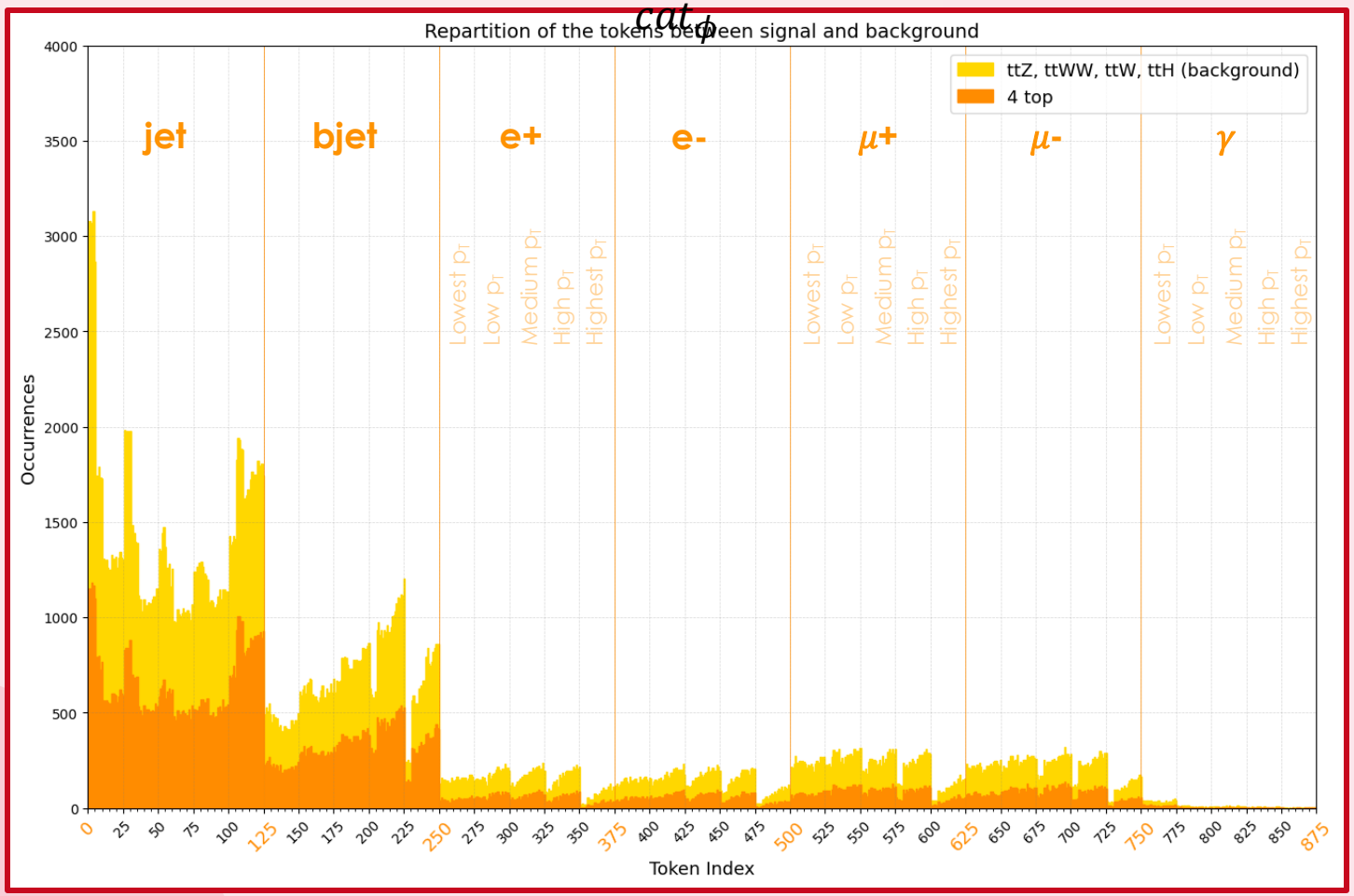




*Graph Attention Convolution (layer from Graph Attention Network Architecture [7])

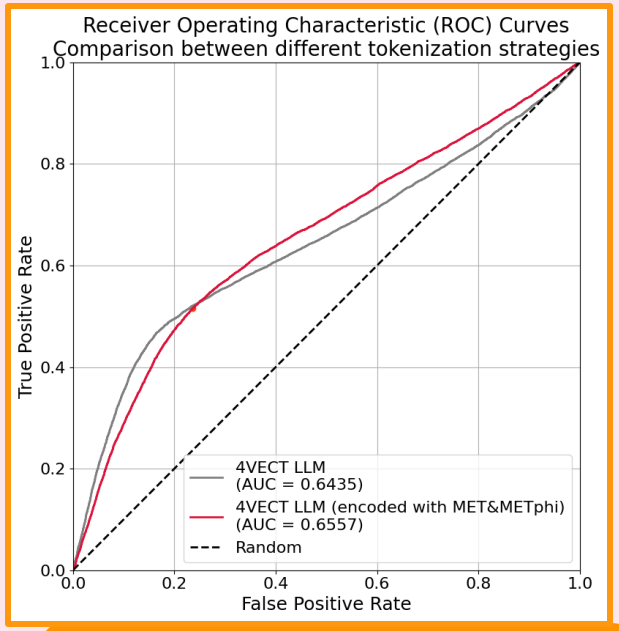
BEST TOKENIZATION SO FAR…

$$token_{4VECT} = (cat_{id} - 1) \times 125 + (cat_{pt} - 1) \times 25 + (cat_{\eta} - 1) \times 5 + cat_{\phi}$$



- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ
- 1 bin per id

- 5 bins of 20% of MET
- 5 geometrical bins of METφ

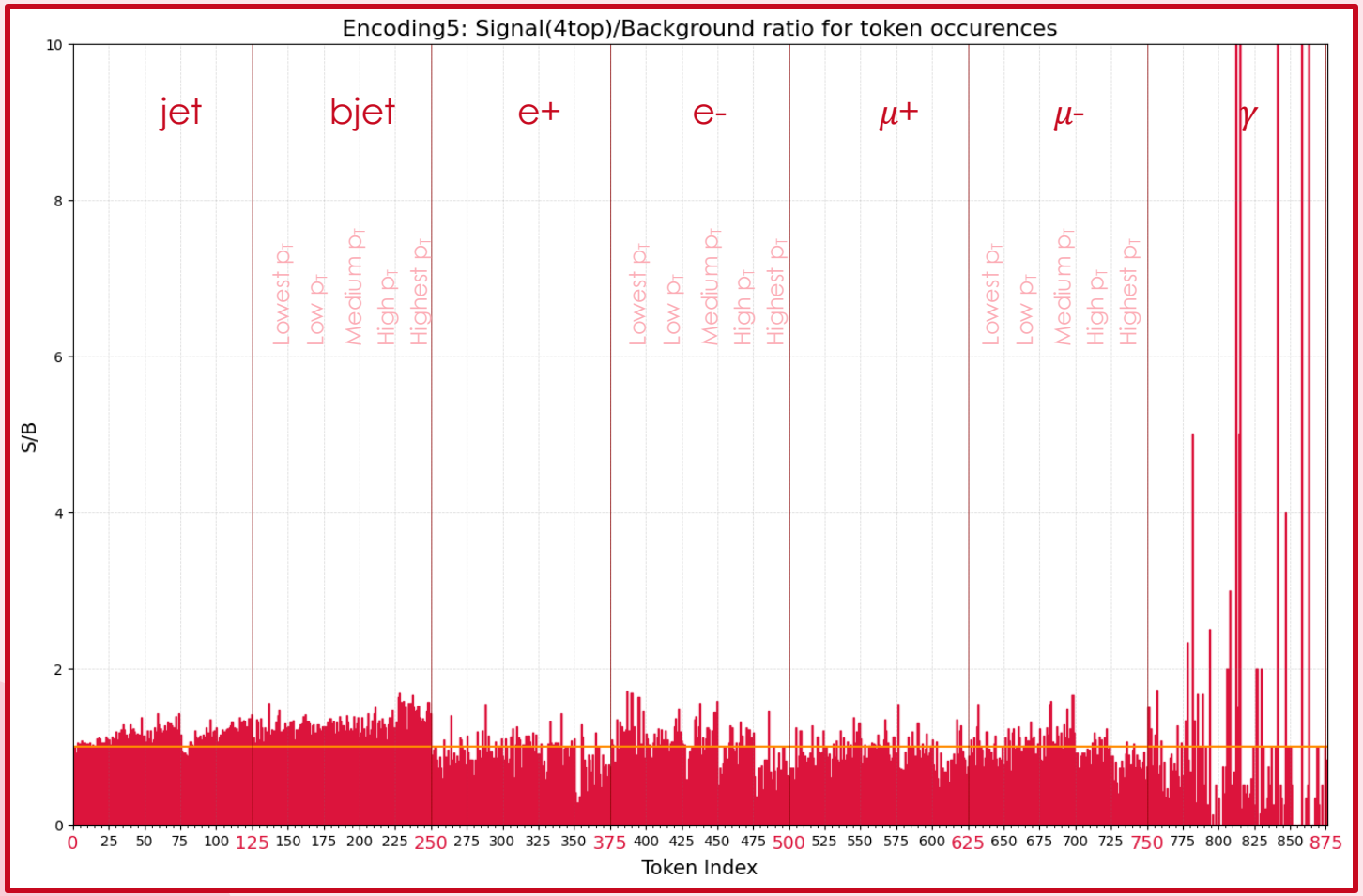


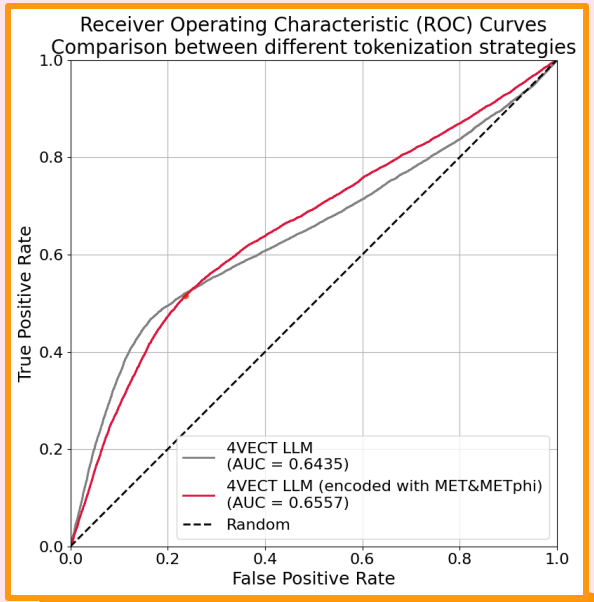
MET and METφ tokens are included at the end of the sequence

BEST TOKENIZATION
SO FAR…

- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ
- 1 bin per id
- 5 bins of 20% of MET
- 5 geometrical bins of METφ

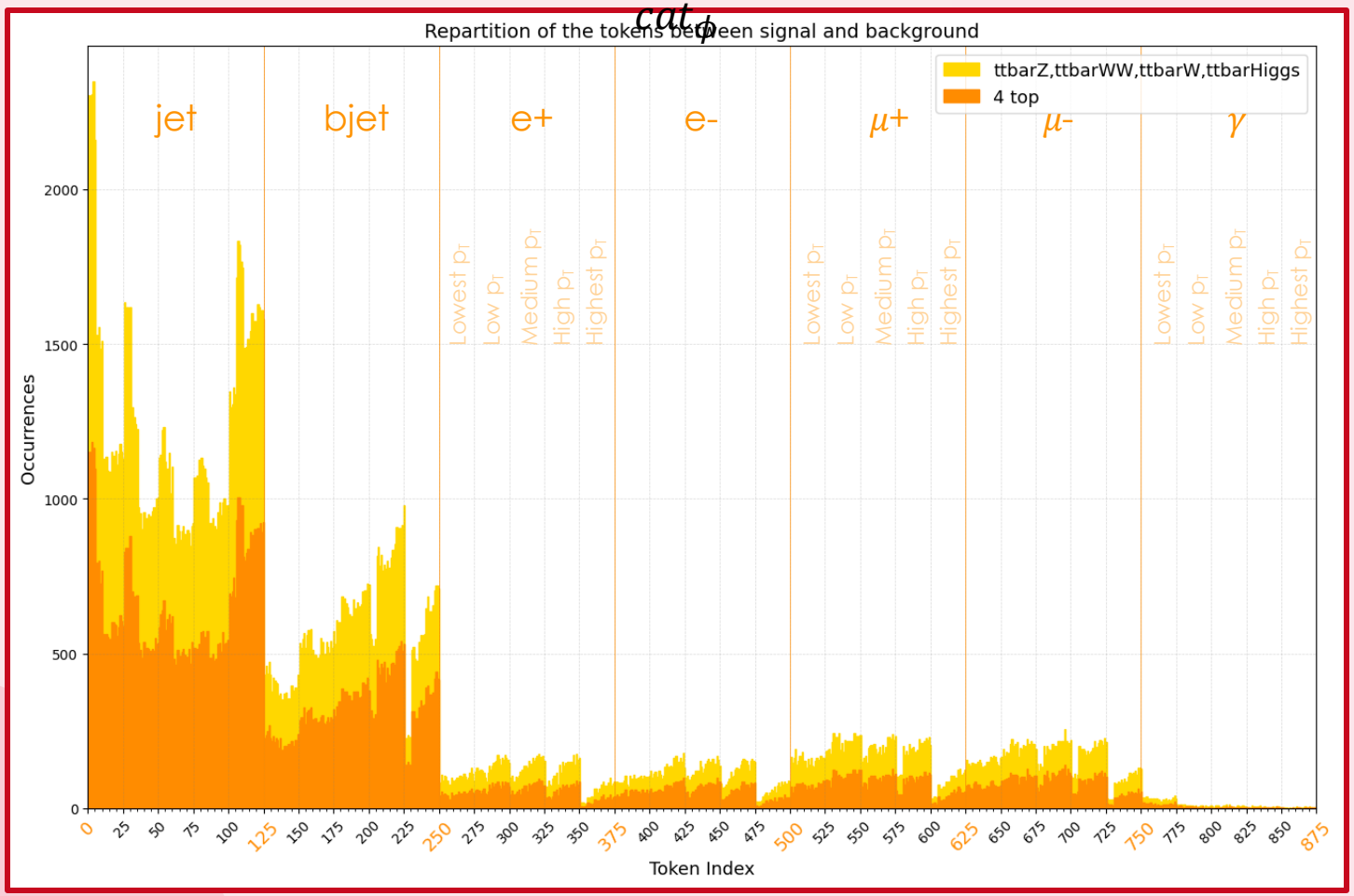
MET and METφ tokens are included at the end of the sequence





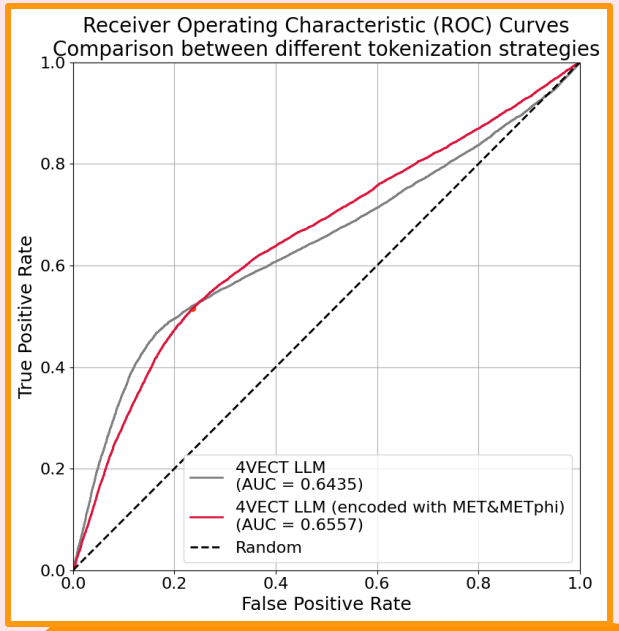
BEST TOKENIZATION SO FAR…

$$token_{4VECT} = (cat_{id} - 1) \times 125 + (cat_{pt} - 1) \times 25 + (cat_{\eta} - 1) \times 5 + cat_{\phi}$$



- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ
- 1 bin per id

- 5 bins of 20% of MET
- 5 geometrical bins of METφ



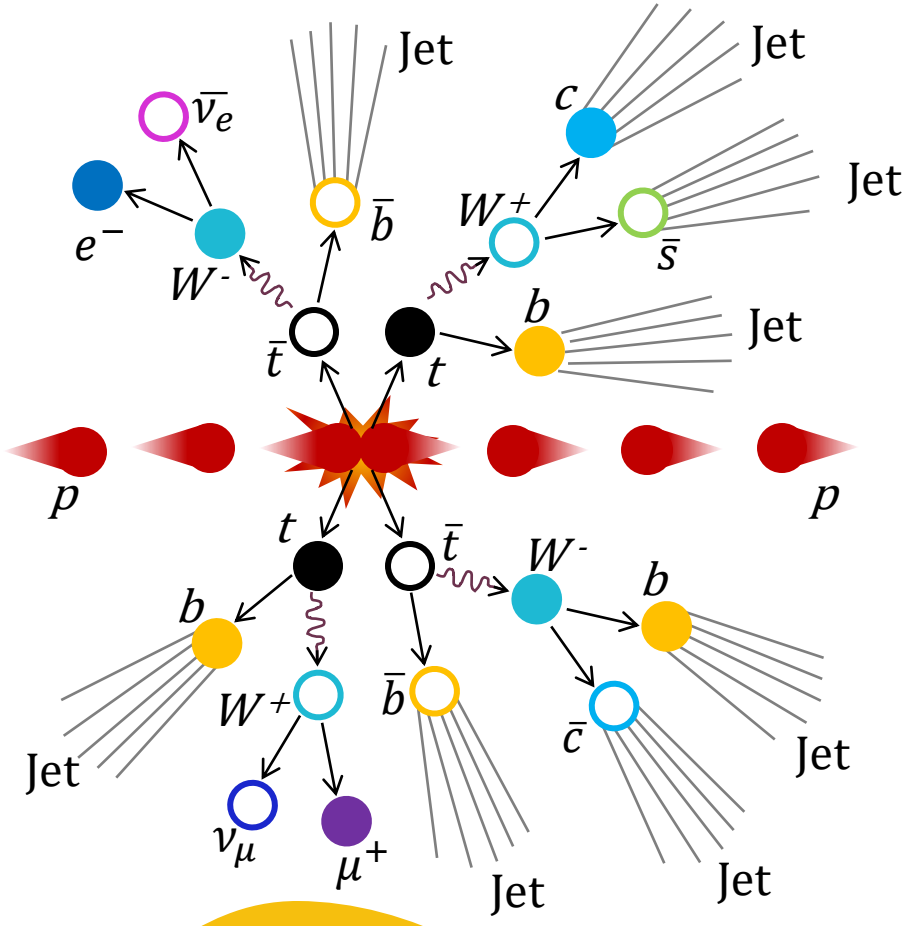
MET and METφ tokens are included at the end of the sequence

RESULTS & OUTLOOKS

Finding the needle in the datastack

4-TOP EVENTS

since $t \rightarrow W + b$ and $|V_{tb}|^2 \sim 1$
Signature: 0-4 leptons and 4-12 jets



Example of a 4-top signature

t t W W

Almost the same signature (but 2 b-jets less 😱)

t t W

since $t \rightarrow W + b$

Quite similar still (2 bjets less at least)

t t Z

Z decays 80% into jets or leptonically

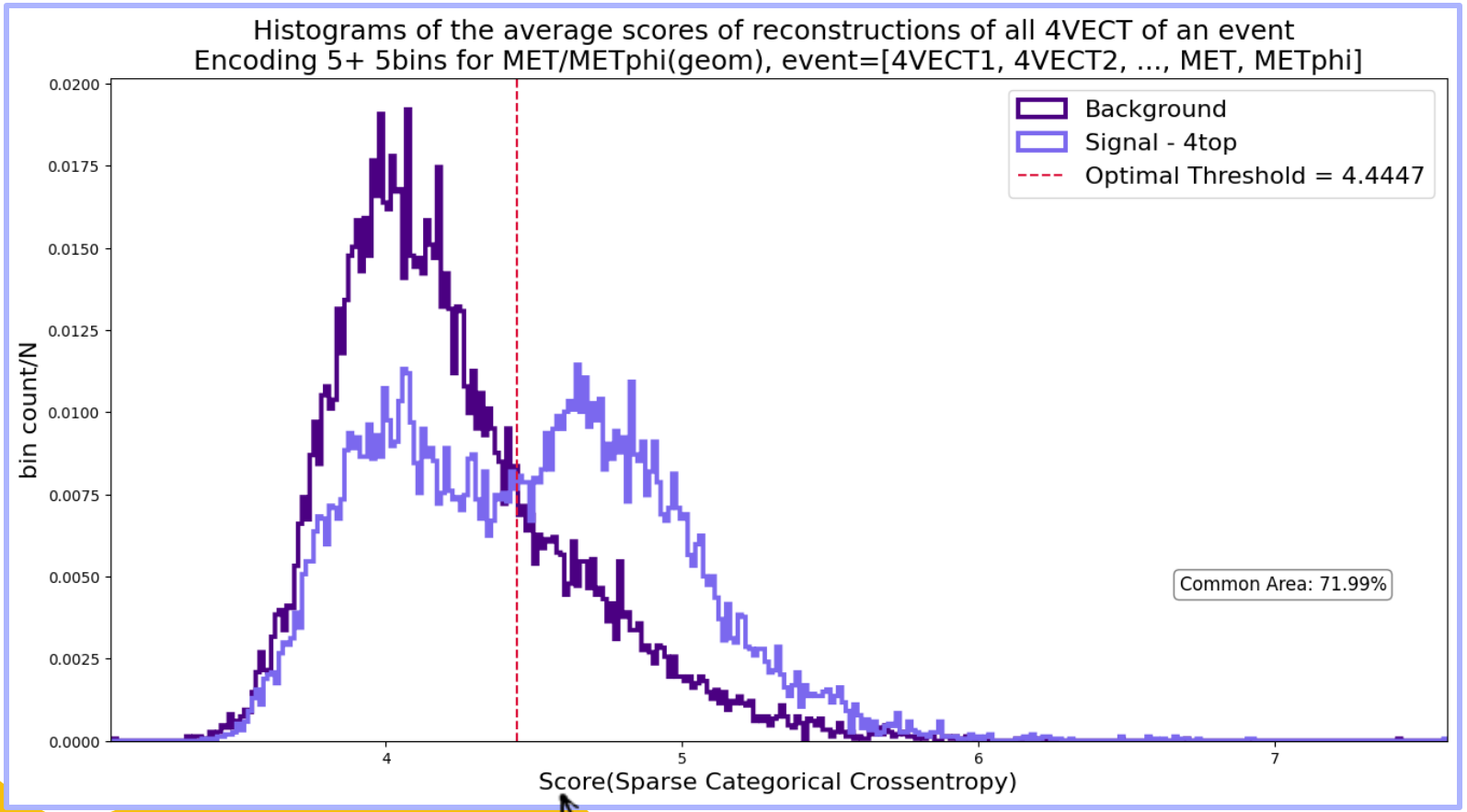
Can have a very similar signature to W (so not good 😩)

t t H

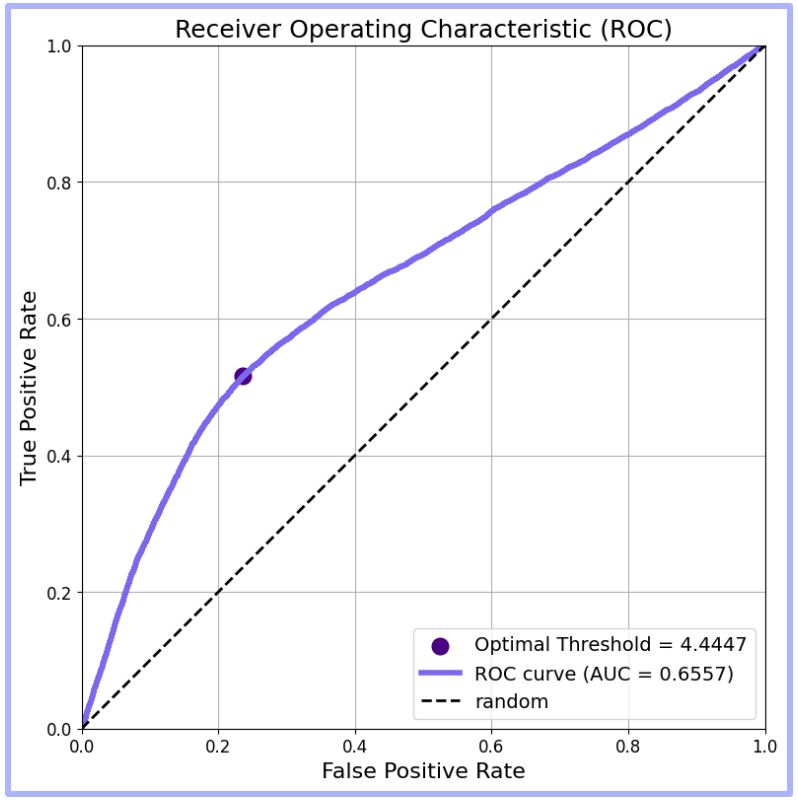
H can decay into 2 jets or W-pair or Z-pair

Can have a very similar signature (still same issues 😤)

APPLIED TO RARE EVENTS



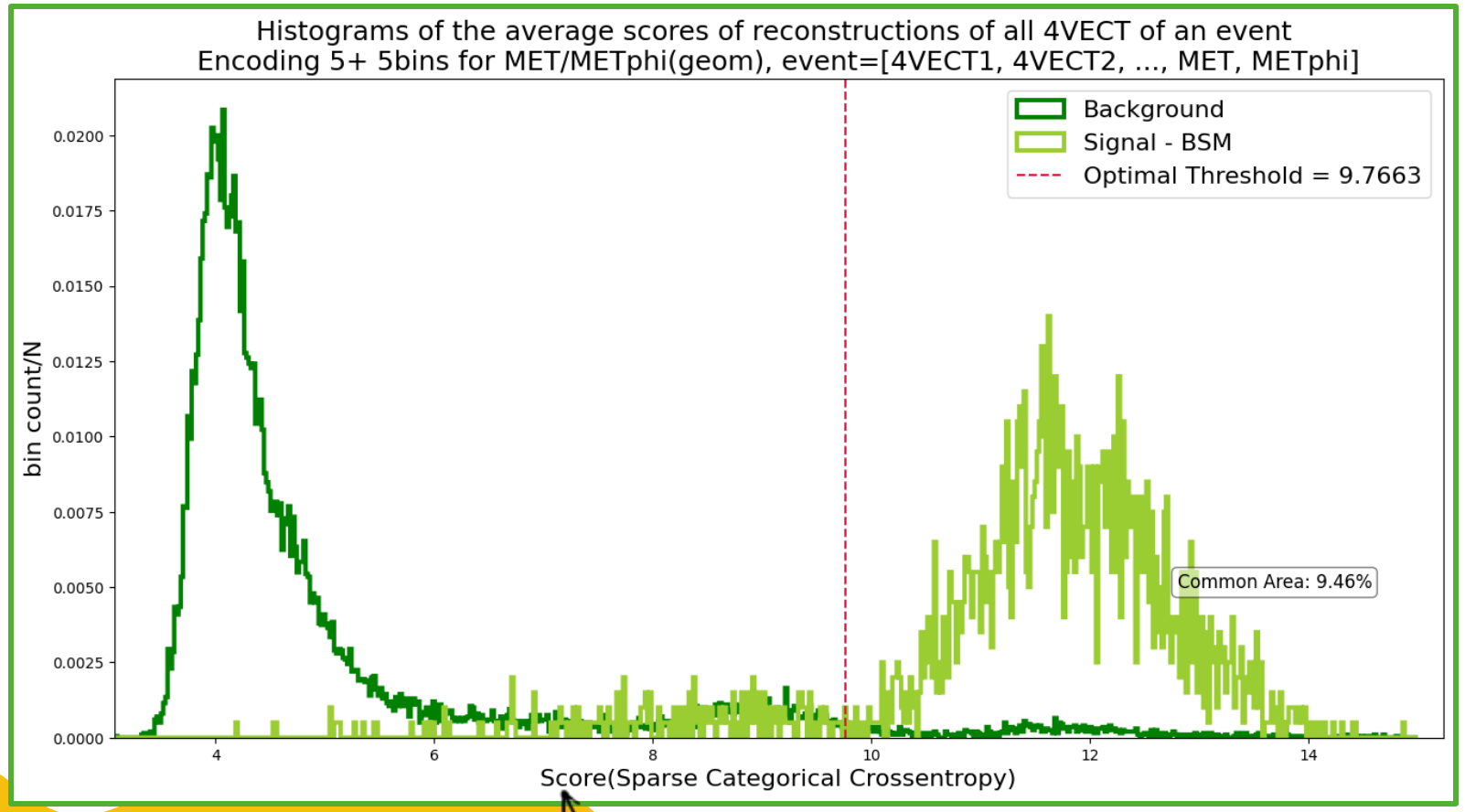
= How good is the token reconstruction



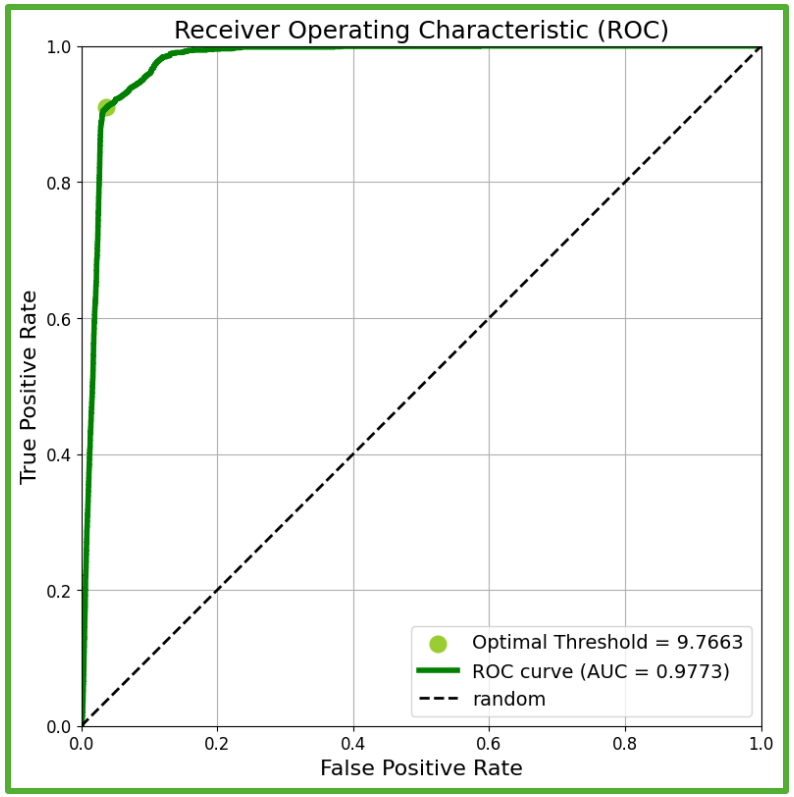
- **Model is trained on background events only**
- Trained over 21 epochs
- With batch of 512 events

APPLIED TO BSM EVENTS

SUSY gluino-gluino production (gluino decaying into a tt pair + MET)



= How good is the token reconstruction

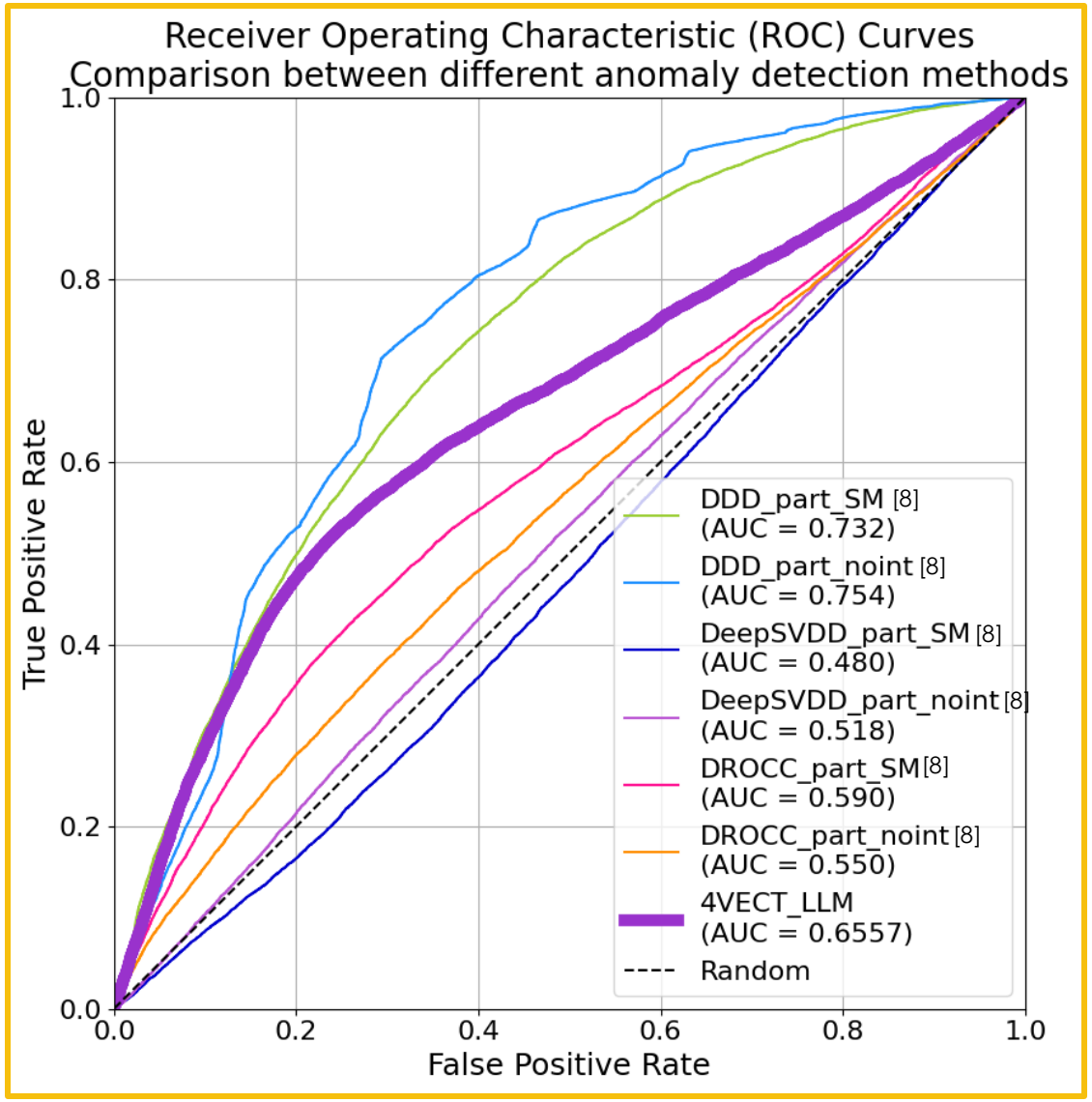


- **Model is trained on background events only**
- Trained over 21 epochs
- With batch of 512 events

OUTLOOKS

- Not a fine-tuned model yet
- Still investigating best tokenization method
- However, method sounds promising

Stay tuned, paper will be published soon



CONCLUSION

From anomaly detection to foundation models: tokenization will reshape how we search for the unknown.

THANK YOU FOR LISTENING

Ambre Visive
Nikhef & University of Amsterdam
ambre.visive@cern.ch

Nikhef

UNIVERSITEIT VAN AMSTERDAM

REFERENCES

[1] Vaswani, A., Shazeer, N. et al. (2023). Attention Is All You Need. arXiv preprint arXiv:*1706.03762*.

[2] ATLAS detector schematics. Atlas experiment website. *https://atlas.cern/Discover/Detector*

[3] Aarrestad, T., van Beekveld, M., Bona, M. et al. (2022). The Dark Machines Anomaly Score Challenge: Benchmark data and model independent event classification for the Large Hadron collider. SciPost Physics, 12(1). *https://doi.org/10.21468/scipostphys.12.1.043.*

[4] Builtjes, L., Caron, S., Moskvitina P. et al. (2025). Attention to the strengths of physical interactions: Transformer and graph-based event classification for particle physics experiments. arXiv preprint arXiv:*2211.05143*.

[5] Van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2018). Neural Discrete Representation Learning. arXiv preprint arXiv:*1711.00937*.

[6] Birk, J., Hallin, A., & Kasieczka, G. (2024). OmniJet-α: the first cross-task foundation model for particle physics. Machine Learning Science and Technology, 5(3), 035031. *https://doi.org/10.1088/2632-2153/ad66ad*

[7] Veličković, P., Cucurull, G., Casanova, et al. (2018). Graph attention networks. . arXiv preprint arXiv:*1710.10903*

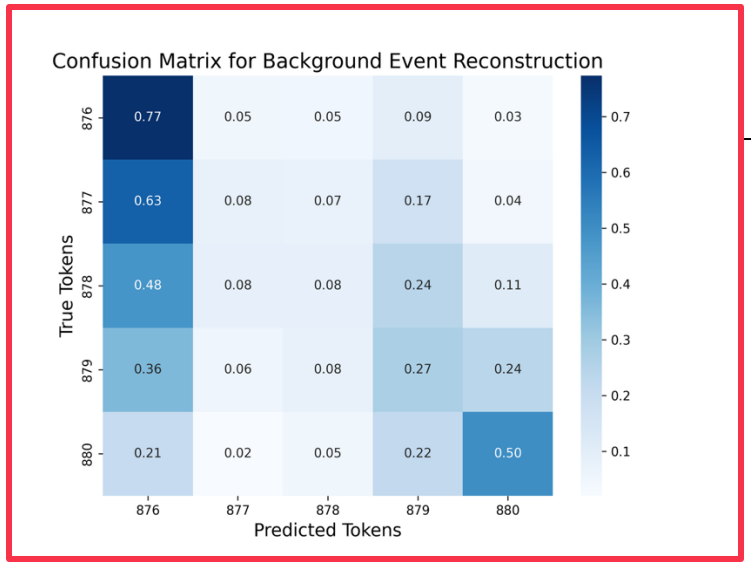
[8] Caron, S., Garcia Navarro, J., Moreno Llacer M. et al. (2025). Universal anomaly detection at the LHC: transforming optimal classifiers and the DDD method. 85(4). *https://doi.org/10.1140/epjc/s10052-025-14087-z.*

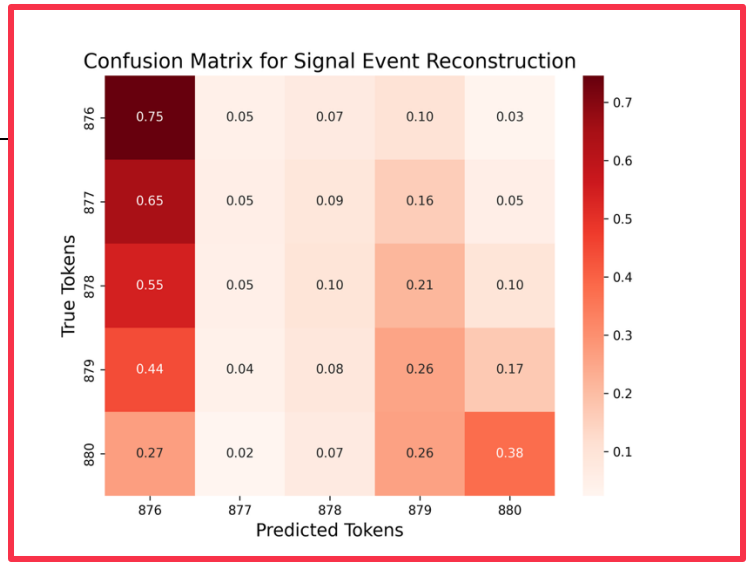
BACK-UPS

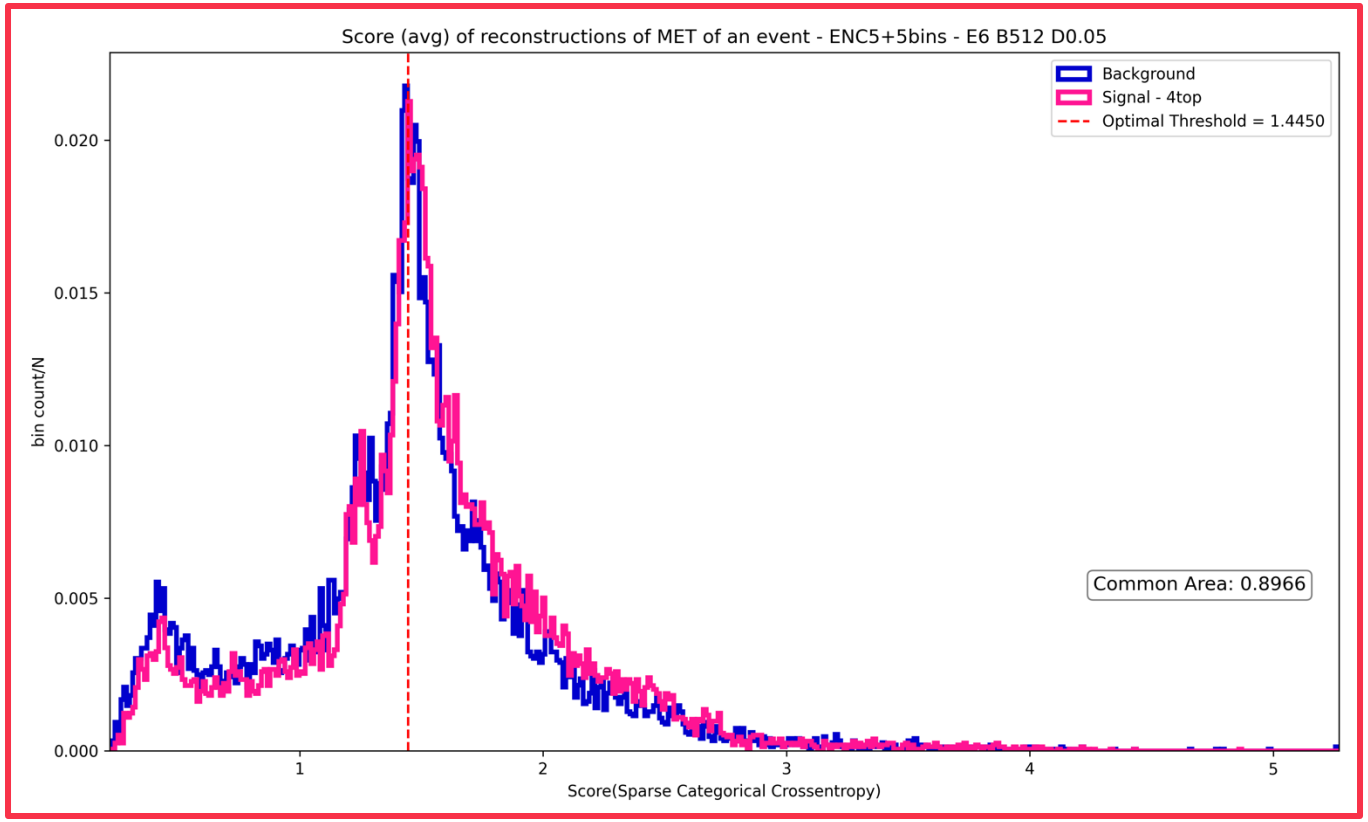
MET MODEL

ROC curves calculated at inference for different model

num epochs	dropout rate	ROC-AUC
6	0.1	0.5474
6	0.1	0.5475
6	0.05	0.5478
6	0.05	0.5468
14	0.1	0.5438
14	0.05	0.5456
28	0.05	0.5451
31	0.05	0.5445
35	0.05	0.5467



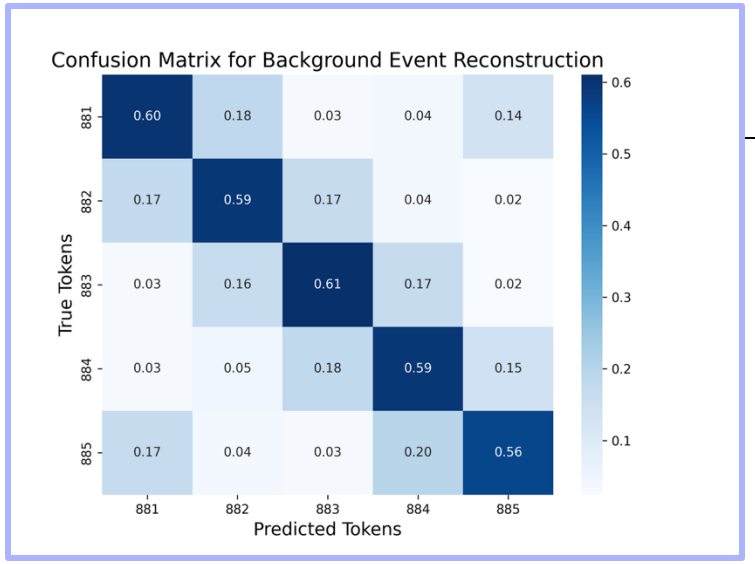


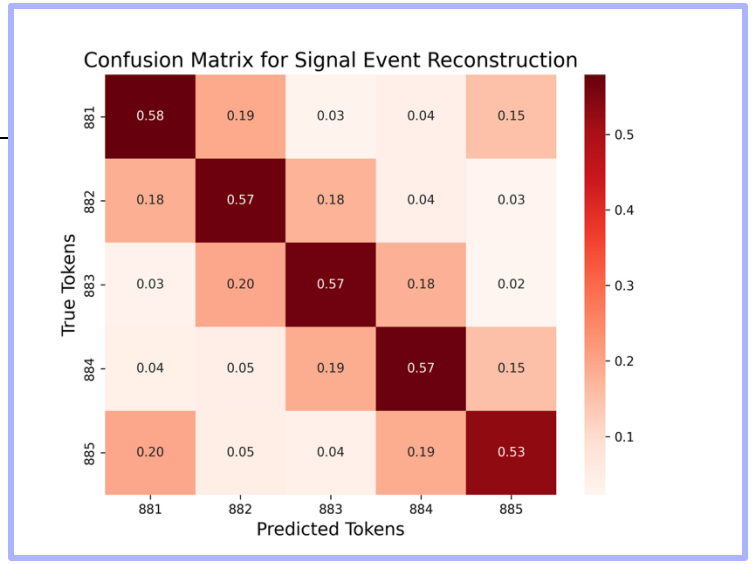


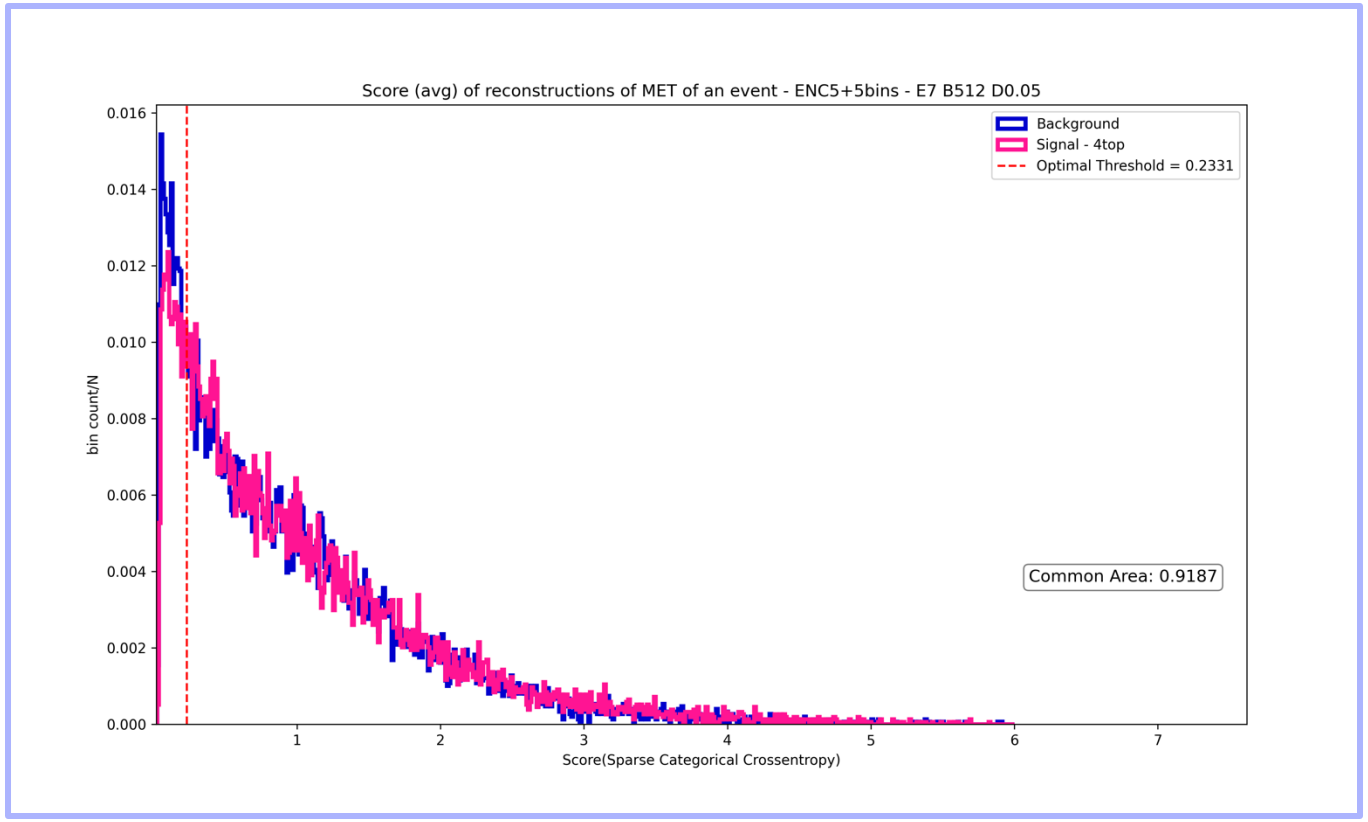
METφ MODEL

ROC curves calculated at inference for different model

num epochs	dropout rate	ROC-AUC
6	0.1	0.5191
7	0.05	0.5207
7	0.05	0.5200
7	0.05	0.5202
7	0.05	0.5201
8	0.1	0.5191
9	0.1	0.5195
11	0.1	0.5187
11	0.1	0.5203
11	0.05	0.5198
12	0.05	0.5200
14	0.05	0.5197







VQVAE

inspired by [6]

Input sequence

Encoder Network

Continuous Latent Representation

Vector Quantization

Discrete Latent Representation (codebook indices)

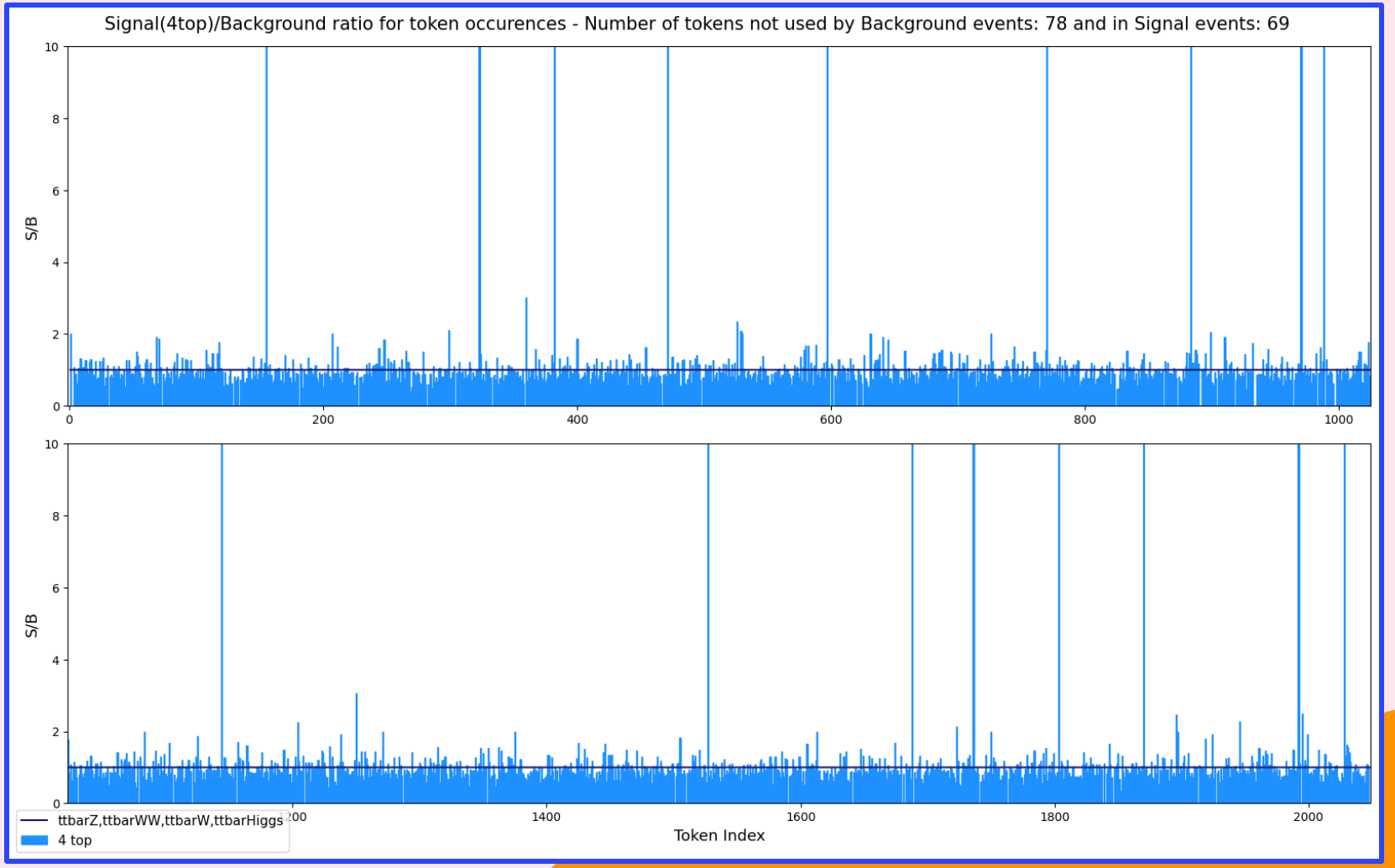
Decoder Network

Reconstructed sequence

[5]

Tokens for our model

MET+METφ included in the sequence as a token



GNN

Input Graphs (nodes = 4VECT, edges = inv. Mass)

3 GATConv* Layers

Continuous Latent Representation

Vector Quantization

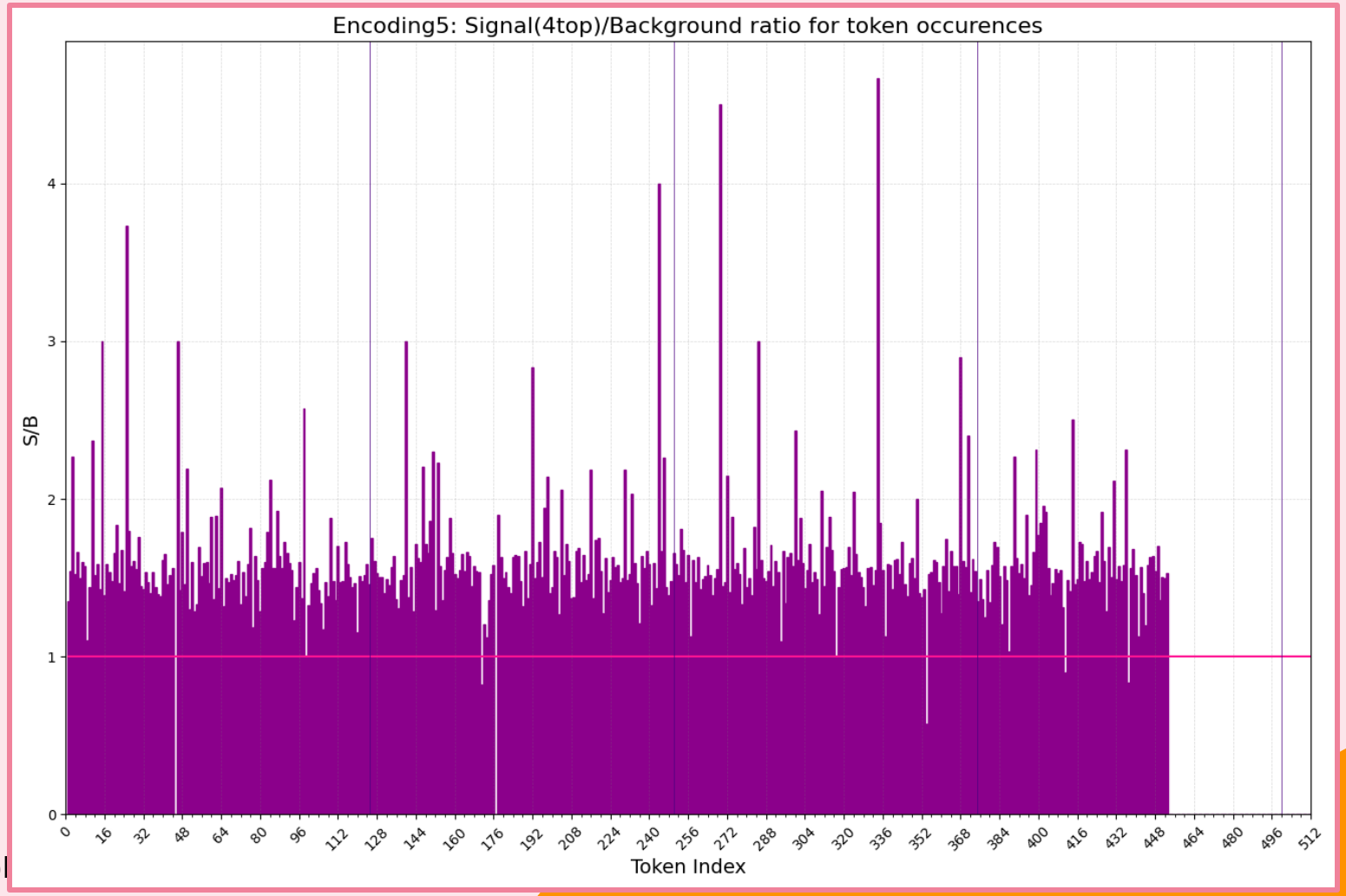
Discrete Latent Representation (codebook indices)

Classification Head

Output (2 classes)

Tokens for our model

codebook_size = 512 { used tokens in test dataset: 452; used tokens in train dataset: 453



*Grapl

GNN

Input Graphs (nodes = 4VECT, edges = inv. Mass)

3 GATConv* Layers

Continuous Latent Representation

Vector Quantization

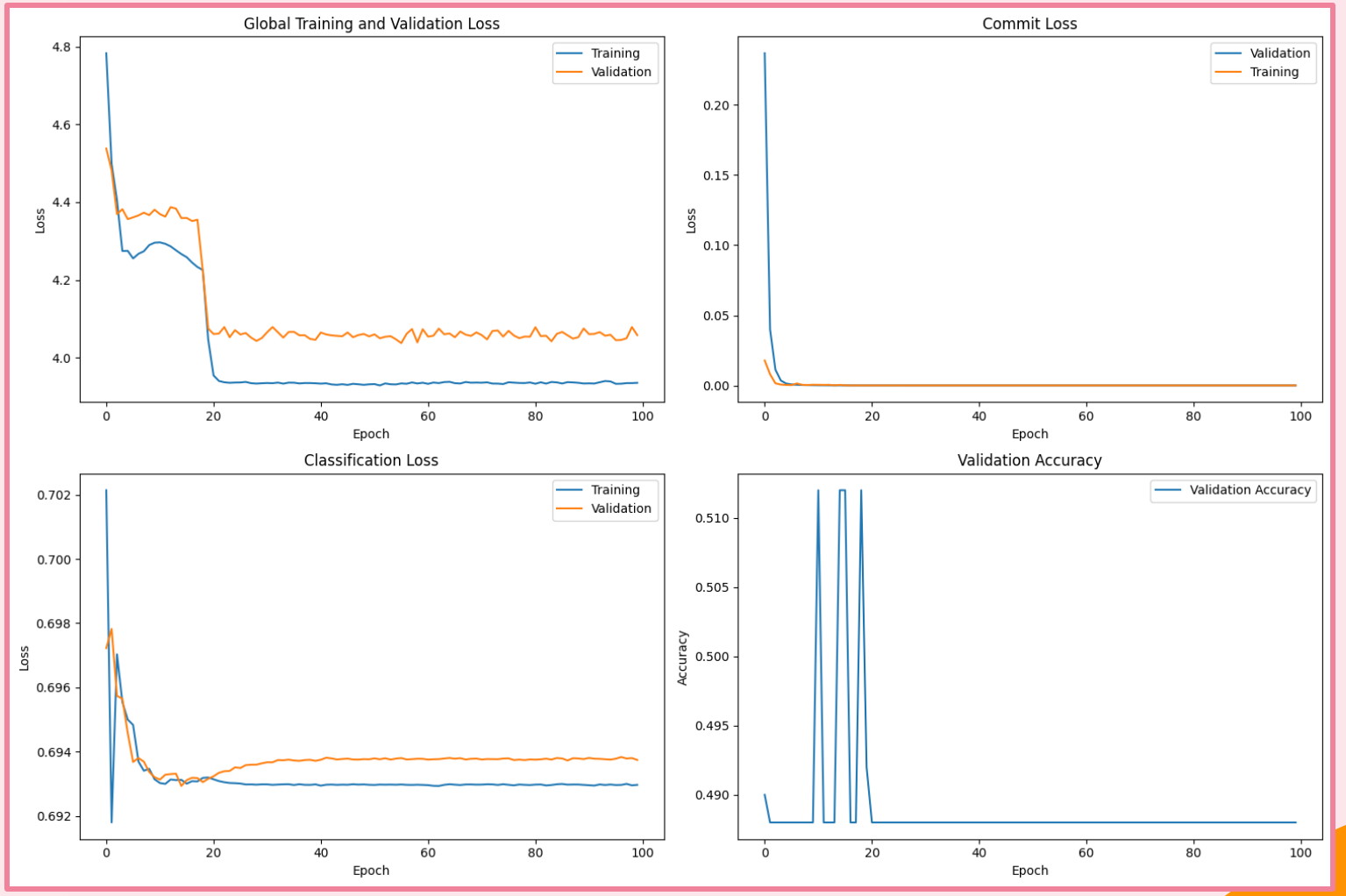
Discrete Latent Representation (codebook indices)

Classification Head

Output (2 classes)

Tokens for our model

Training of the GNN



*Graph Attention Convolution (layer from Graph Attention Network Architecture [7])

GNN

Input Graphs (nodes = 4VECT, edges = inv. Mass)

↓

3 GATConv* Layers

↓

Continuous Latent Representation

↓

Vector Quantization

↓

Discrete Latent Representation (codebook indices)

→ Tokens for our model

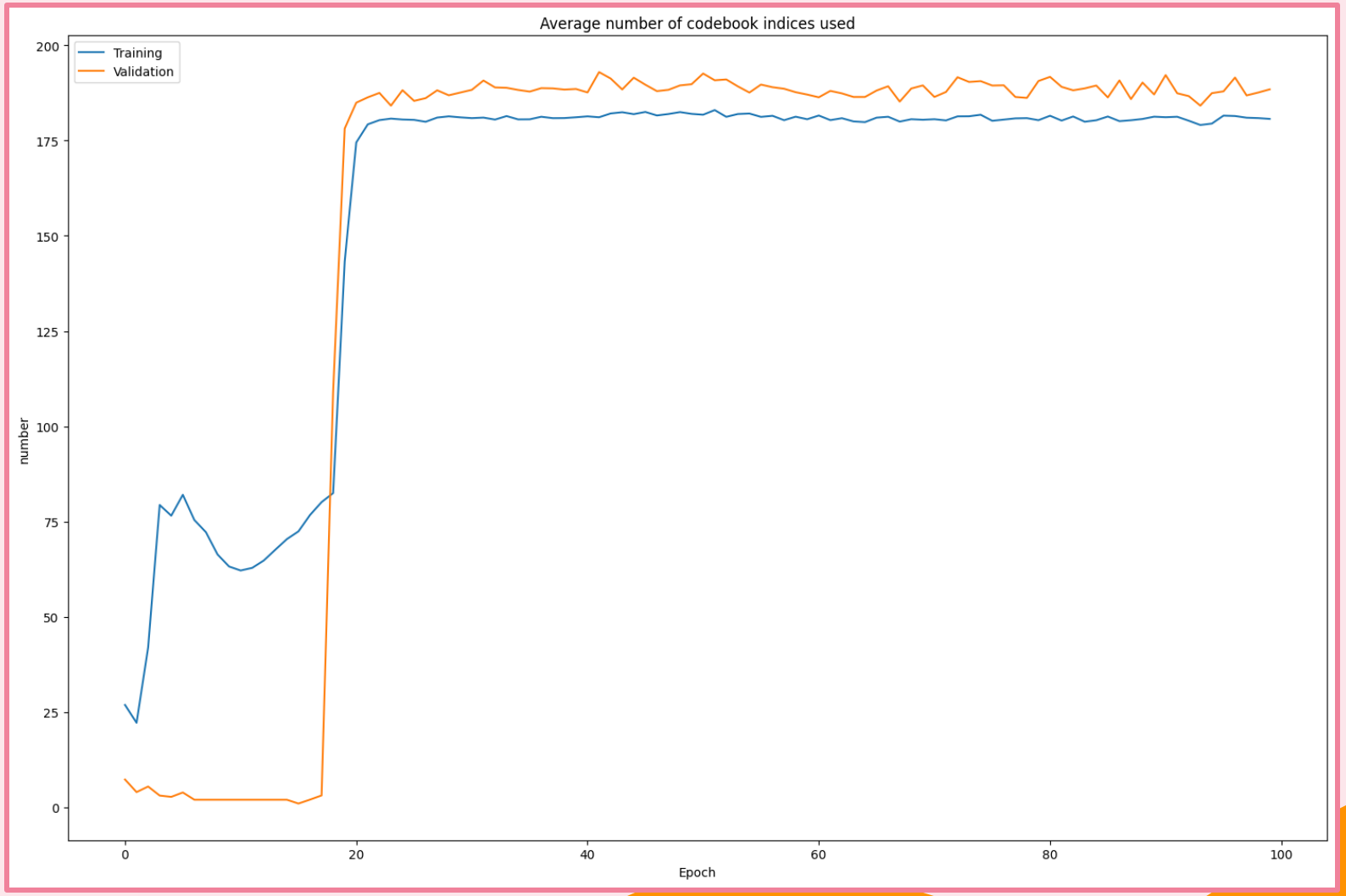
↓

Classification Head

↓

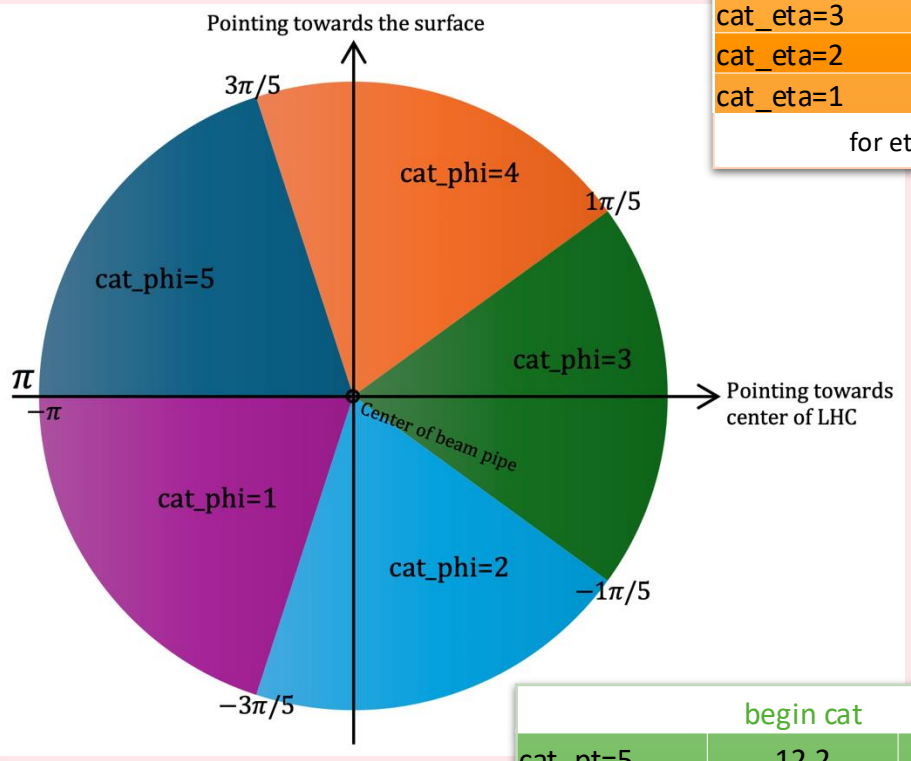
Output (2 classes)

Training of the GNN



*Graph Attention Convolution (layer from Graph Attention Network Architecture [7])

BEST TOKENIZATION
SO FAR…



	begin cat	end cat	description
cat_eta=5	0	0.35	in the transverse plan
cat_eta=4	0.35	0.7	slightly forward particle
cat_eta=3	0.7	1.07	forward particle
cat_eta=2	1.07	1.6	very forward particle
cat_eta=1	1.6	+inf	most forward particle
for eta in absolute values			around 20% in each bin

	description
cat_id=1	jet
cat_id=2	bjet
cat_id=3	positron
cat_id=4	electron
cat_id=5	antimuon
cat_id=6	muon
cat_id=7	gamma

	begin cat	end cat	description
cat_pt=5	12.2		very high pt
cat_pt=4	11.58	12.2	high pt
cat_pt=3	11.12	11.58	medium pt
cat_pt=2	10.66	11.12	low pt
cat_pt=1	0	10.66	very low pt
			around 20% in each bin

- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ
- 1 bin per id
- 5 bins of 20% of MET
- 5 geometrical bins of METφ

MET and METφ tokens are included at the end of the sequence

$$token_{4VECT} = (cat_{id} - 1) \times 125 + (cat_{pt} - 1) \times 25 + (cat_{\eta} - 1) \times 5 + cat_{\phi}$$

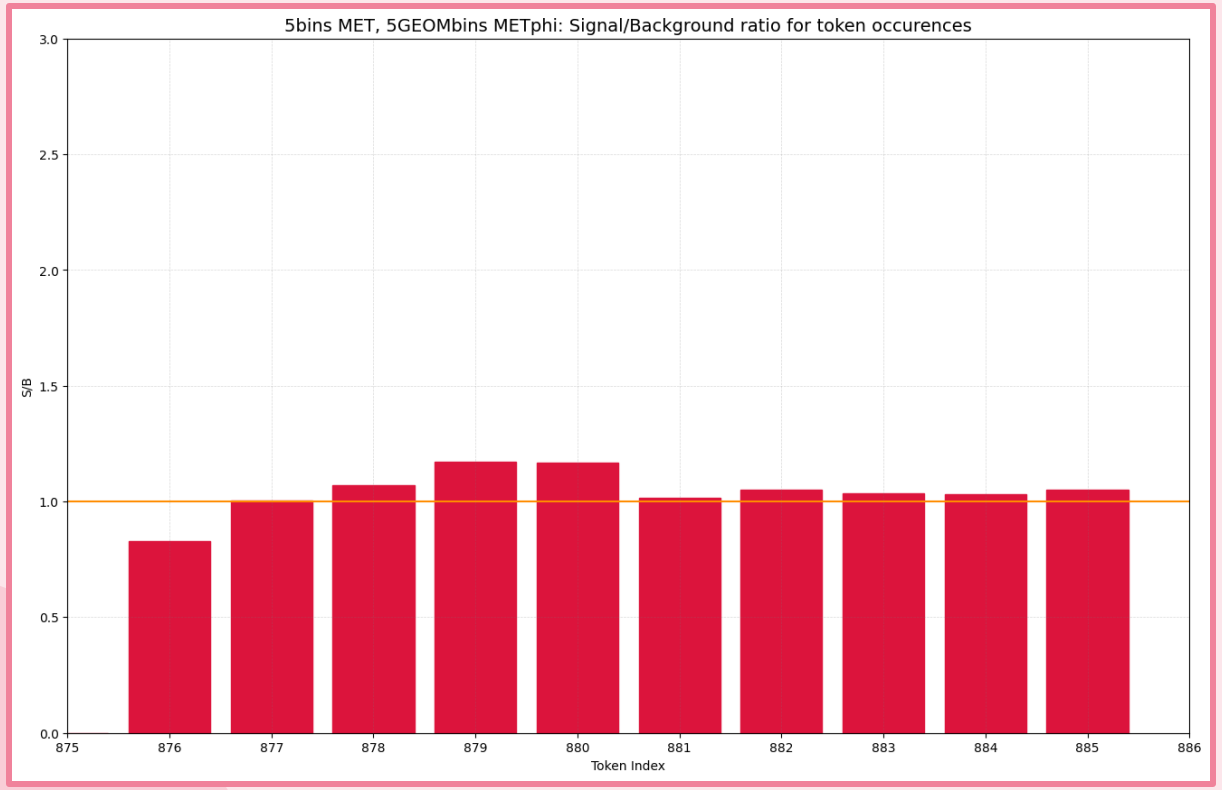
BEST TOKENIZATION

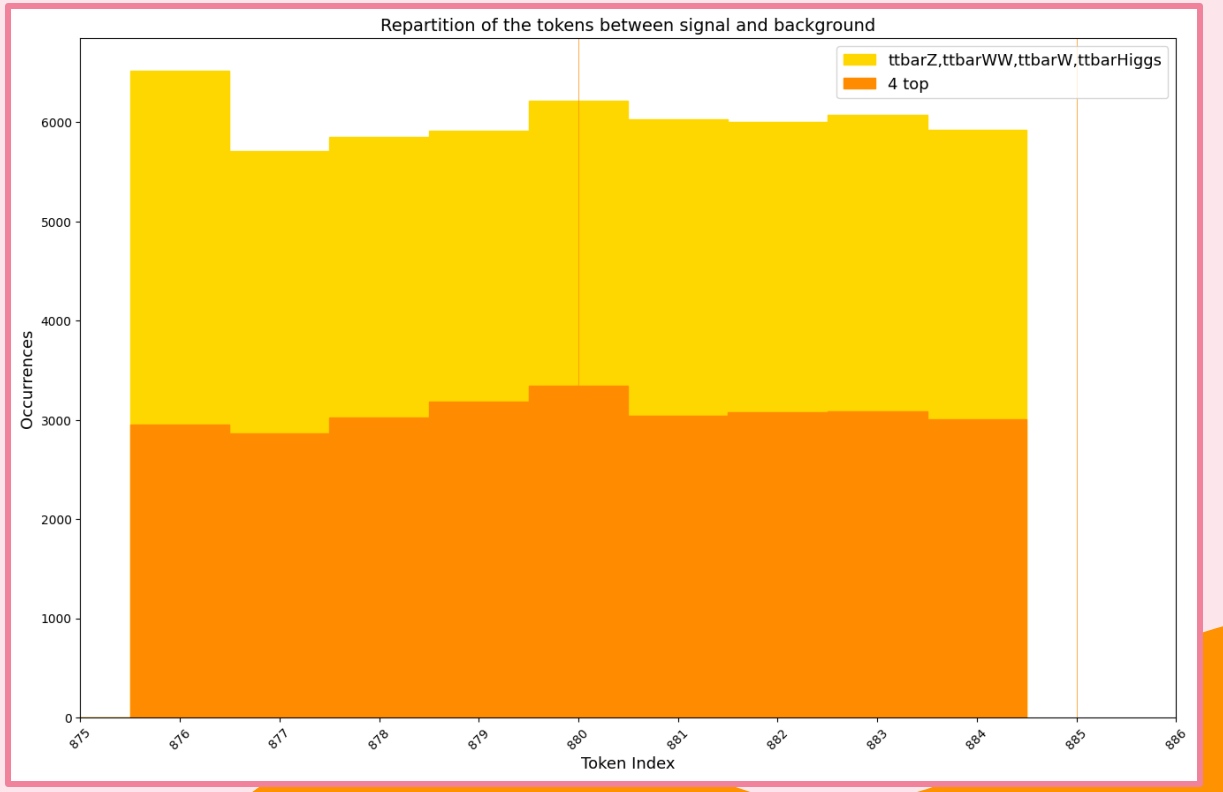
SO FAR…

Focus on MET/METφ bins

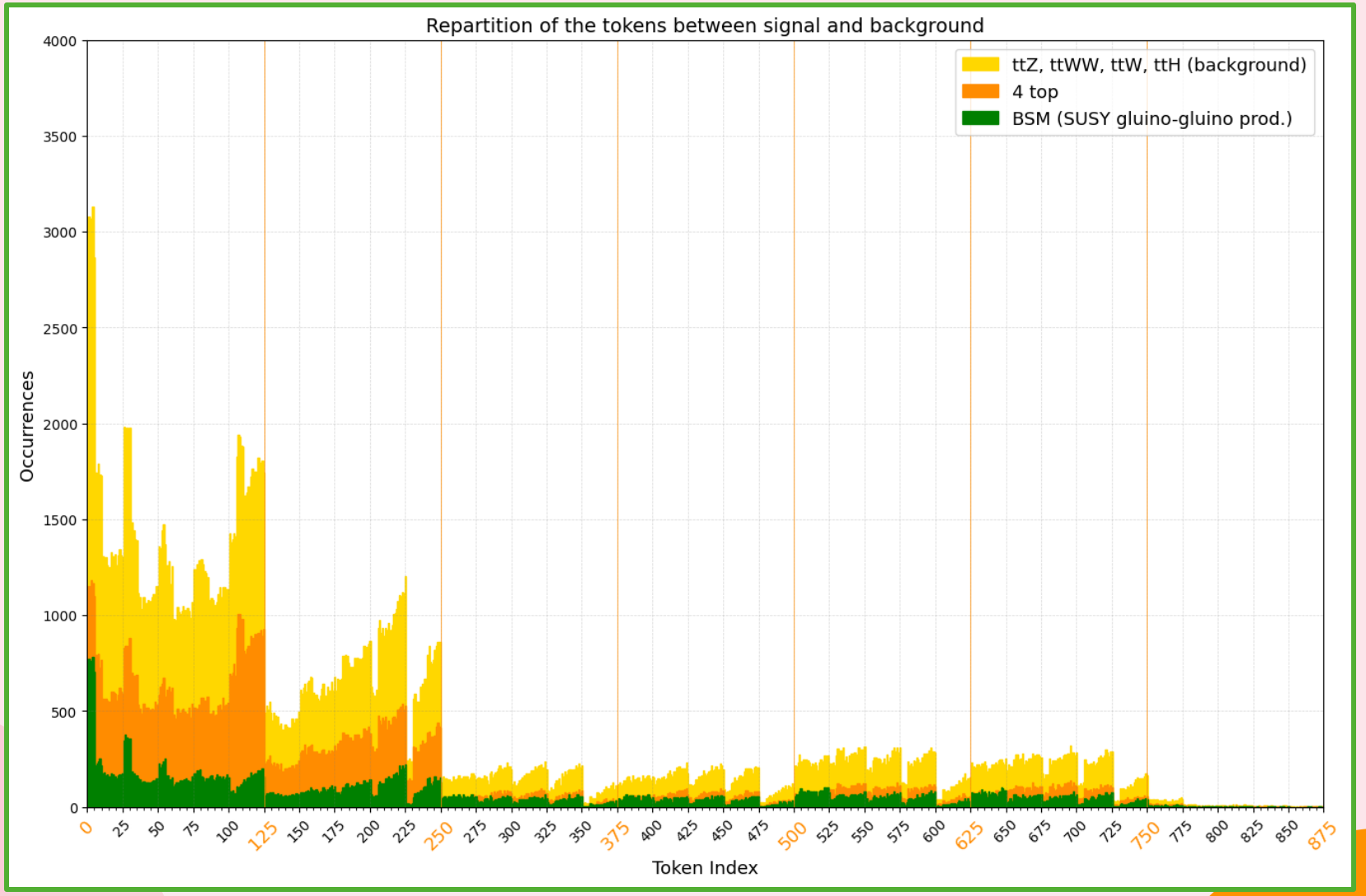
- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ

- 1 bin per id
- 5 bins of 20% of MET
- 5 geometrical bins of METφ



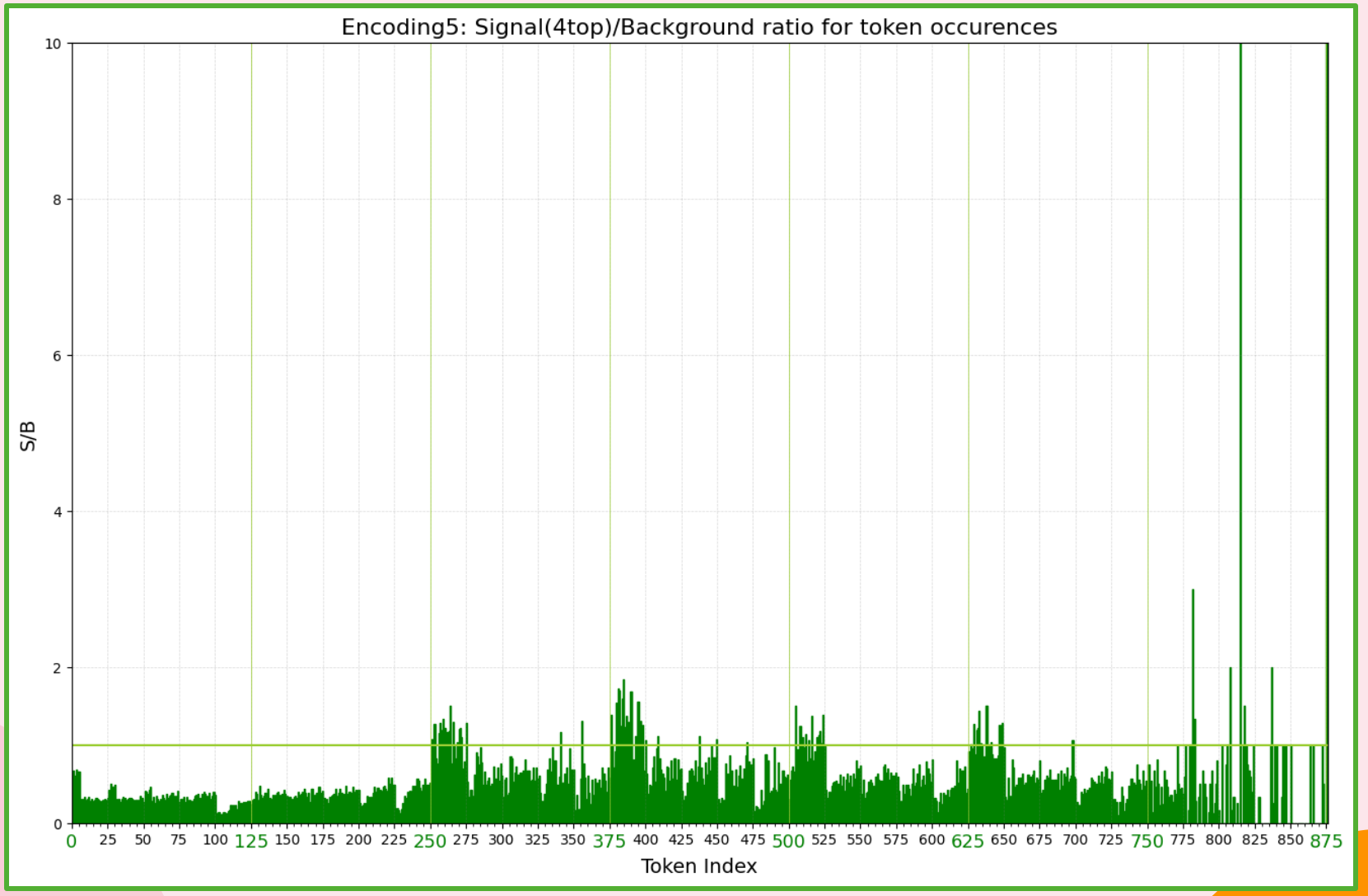


BEST TOKENIZATION
SO FAR…



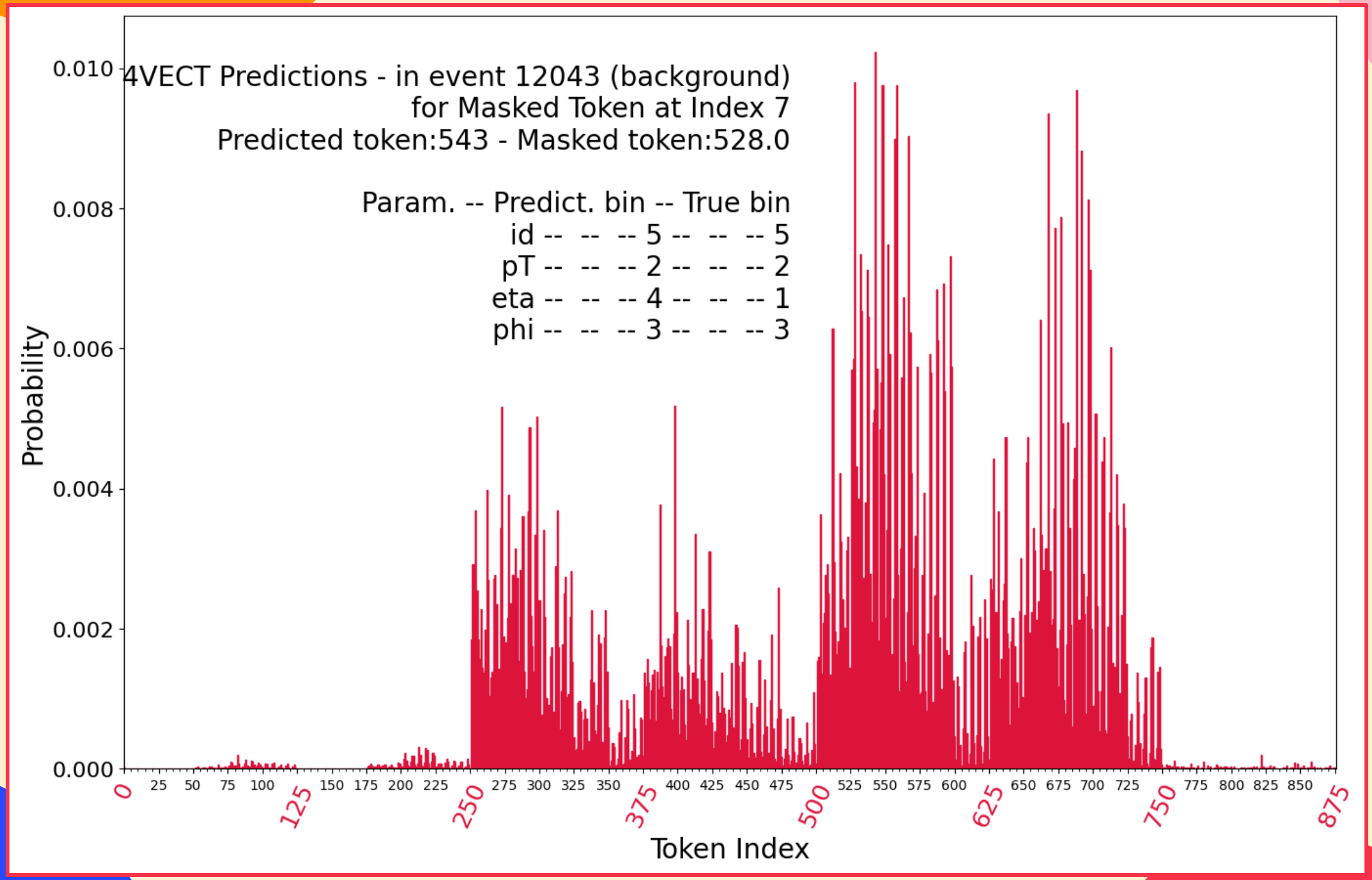
- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ
- 1 bin per id
- 5 bins of 20% of MET
- 5 geometrical bins of METφ

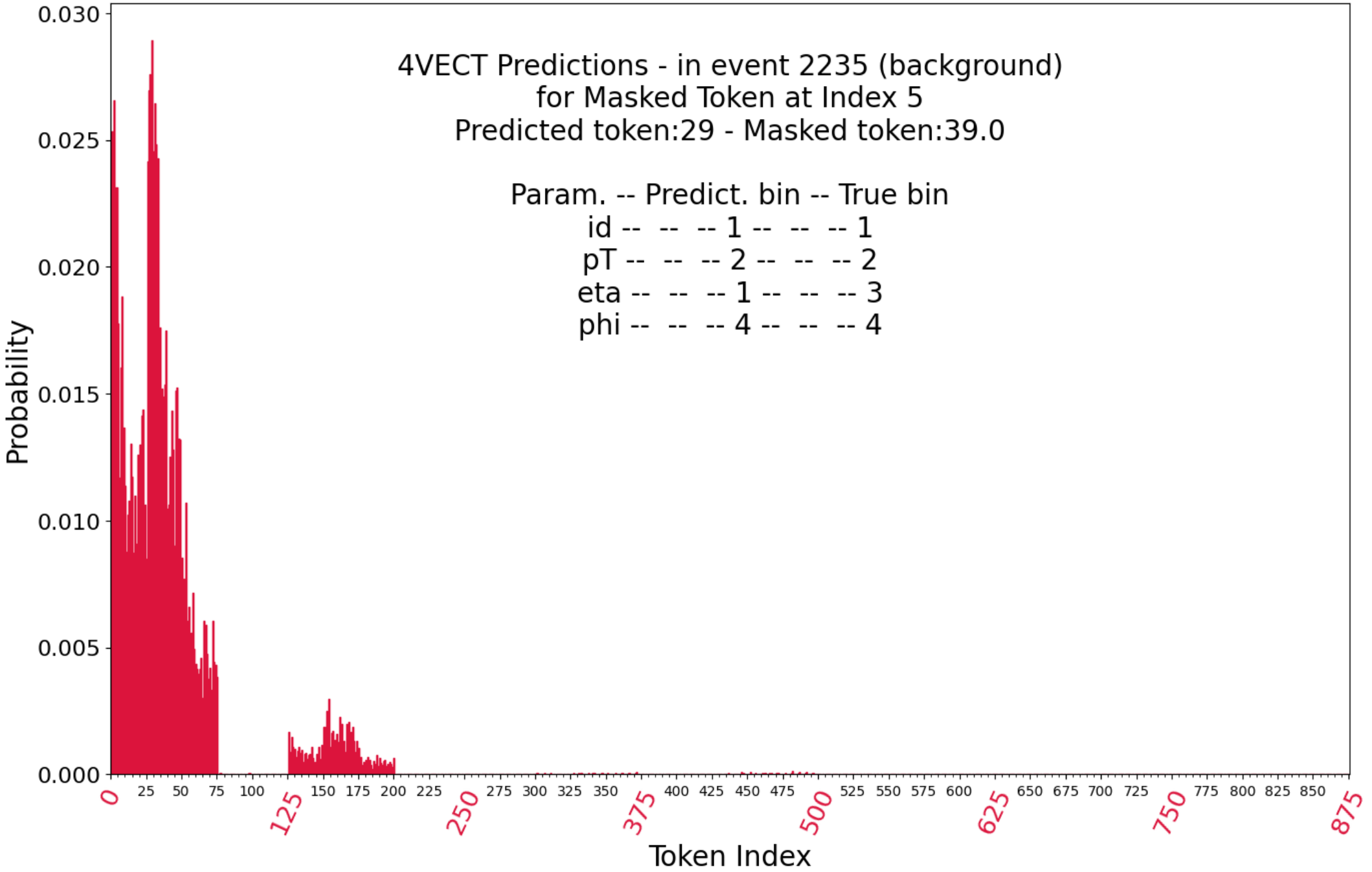
BEST TOKENIZATION
SO FAR…

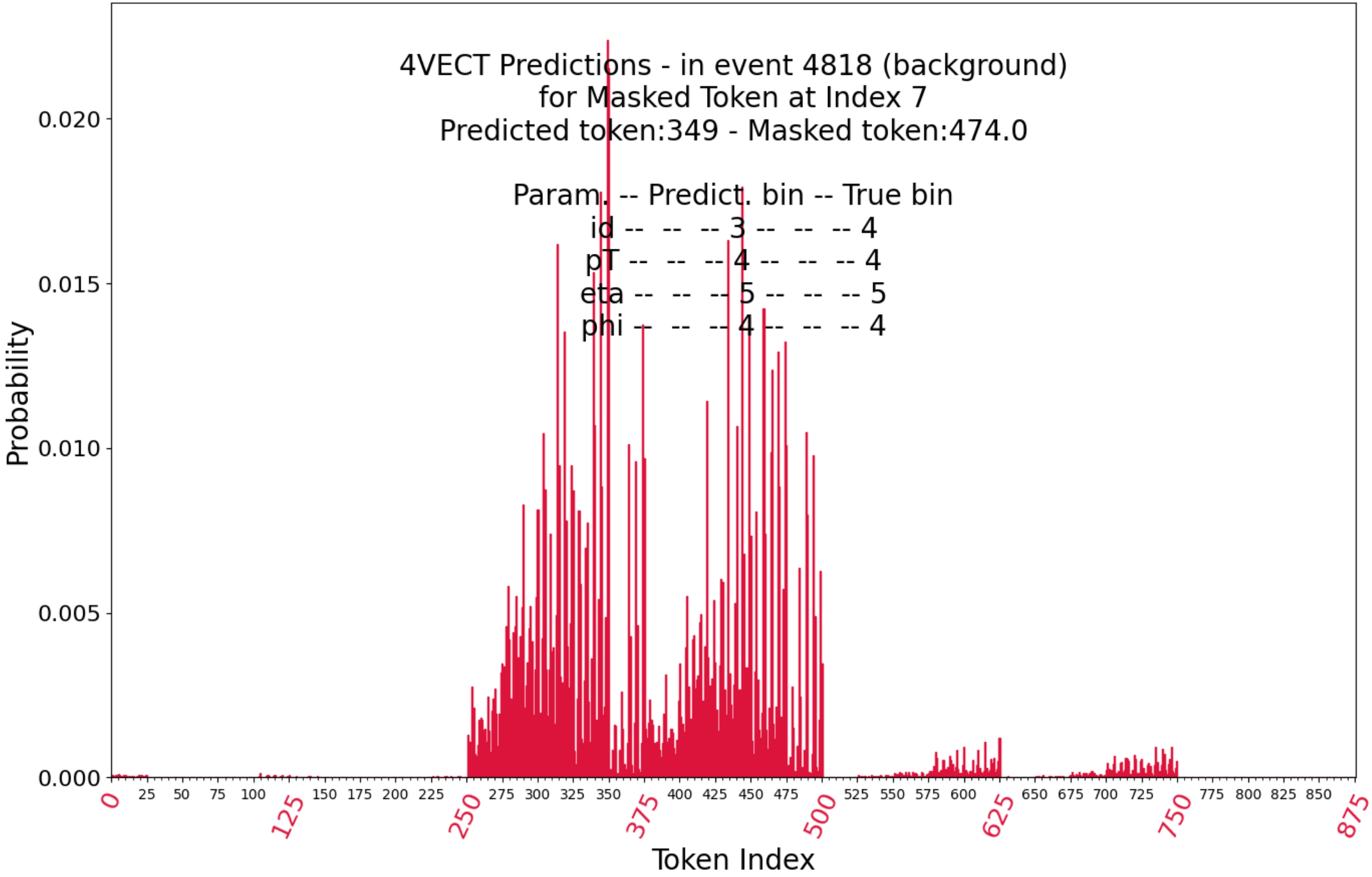


- 5 bins of 20% of p_T
- 5 bins of 20% of $|\eta|$
- 5 geometrical bins of φ
- 1 bin per id
- 5 bins of 20% of MET
- 5 geometrical bins of METφ

Focus on BSM events (SUSY gluino-gluino production)


4VECT Predictions - in event 12043 (background)
for Masked Token at Index 7
Predicted token:543 - Masked token:528.0
Param. -- Predict. bin -- True bin
id -- -- -- 5 -- -- -- 5
pT -- -- -- 2 -- -- -- 2
eta -- -- -- 4 -- -- -- 1
phi -- -- -- 3 -- -- -- 3
Probability
0.010
0.008
0.006
0.004
0.002
0.000
0 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875
Token Index


4VECT Predictions - in event 2235 (background)
for Masked Token at Index 5
Predicted token:29 - Masked token:39.0
Param. -- Predict. bin -- True bin
id -- -- -- 1 -- -- -- 1
pT -- -- -- 2 -- -- -- 2
eta -- -- -- 1 -- -- -- 3
phi -- -- -- 4 -- -- -- 4
Probability
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875
Token Index


4VECT Predictions - in event 4818 (background)
for Masked Token at Index 7
Predicted token:349 - Masked token:474.0
Param. -- Predict. bin -- True bin
id -- -- -- 3 -- -- -- 4
pT -- -- -- 4 -- -- -- 4
eta -- -- -- 5 -- -- -- 5
phi -- -- -- 4 -- -- -- 4
Probability
0.020
0.015
0.010
0.005
0.000
0 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875
Token Index