EVENT TOKENIZATION AND NEXT-TOKEN PREDICTION FOR ANOMALY DETECTION AT THE LHC

Ambre Visive[1,2], Roberto Ruiz de Austri[3], Clara Nellist[1,2], Sascha Caron[1,4], Polina Moskvitina[1,4]

1. Nikhef, Amsterdam, Netherlands
2. University of Amsterdam, Amsterdam, Netherlands
3. Valencia University, IFIC, Valencia, Spain
4. Radboud University, Nijmegen, Netherlands

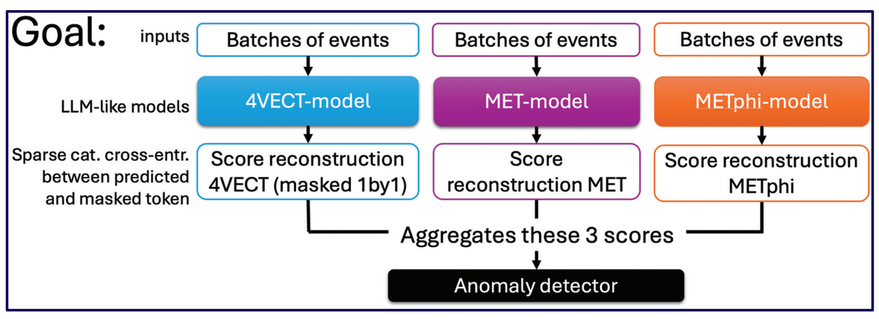
An LLM-like Model

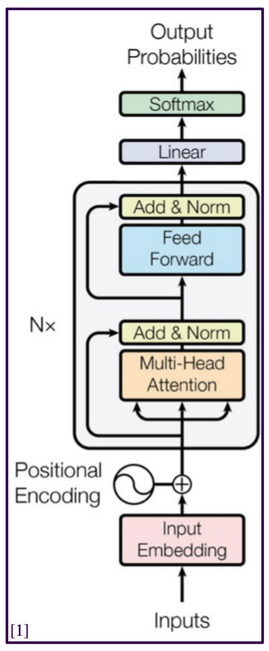
-**We would like to build an LLM-like model that correctly reconstructs a «hidden» particle from an event**.

-We are using **encoder-only transformers**, that use (sequences of) tokens as input(s) and have ~260 trainable parameters.

-The input of the model is a sequence of particles: an event.

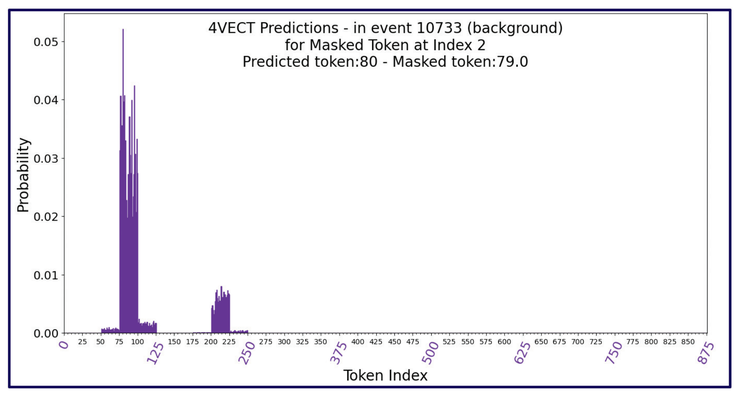
view the embedding process of our model





-During training and inference, we mask one particle from that event to do a next-token-like prediction.

-The output of our model is a two-dimensional tensor containing for each particle of the input event, the probabilities for each token to represent that particle.



Tokenization strategy

-The initial data set (Dark Machines Collaboration [2]), contains for each event and for up to 18 particles of that event, the type of the particle including its charge, its p_t, φ and η, and the missing transverse energy of the event (MET) and its azimutal angle(METphi). The events are 0-padded so all events are the same length.

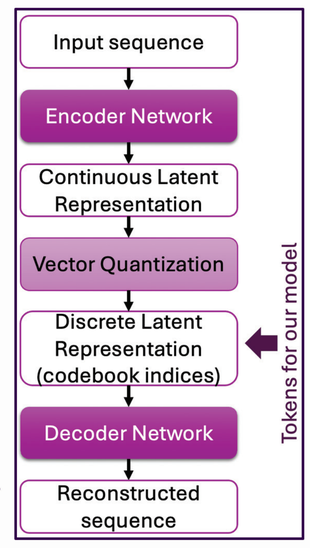
-Since our model needs tokens as input, the particle physics dataset had to be tokenized and the 0-pads are masked.

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

-We tried several approaches to tokenize the events by hand, including the tokenization of the 4-vectors of a particle. The example above is obtained by categorizing each property of a particle in 5 bins.

-We are currently trying to do a tokenization with VQ-VAE to see if it improves the performace of our model.

-Special attention is also given to the selection of information given to the model (MET, METphi, number of jets, of leptons ...).

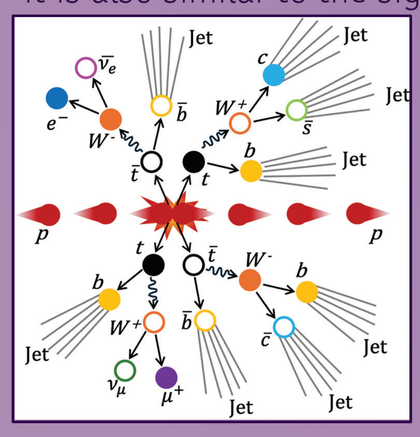


MORE ABOUT CONTEXT

4-top, ttW, ttWW, ttZ, ttH

-**4-top-quark events decay into 4 to 12 jets and 0 to 4 charged leptons**. Its signature shares similarities with a ttW or a ttWW event.

-It is also similar to the signature of a ttZ event, since Z bosons decay mostly into jets or leptonically.



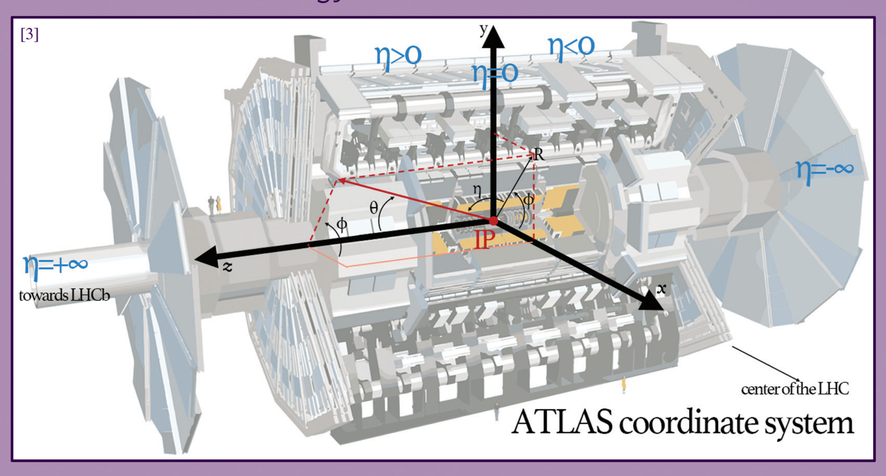
-Lastly, it is similar to the signature of a ttH event, as the Higgs boson can decay into 2 jets, a W- or a Z-pair.

-**ttW, ttWW, ttZ & ttH will be the background events**.

ATLAS detector

-ATLAS is one of the four detectors of the LHC at CERN.

-The dataset, from the Dark Machines Collaboration [2], is from **simulated proton-proton collisions** inside the ATLAS detector (as it was during RUN 2) and at the center of mass energy of 13 TeV.

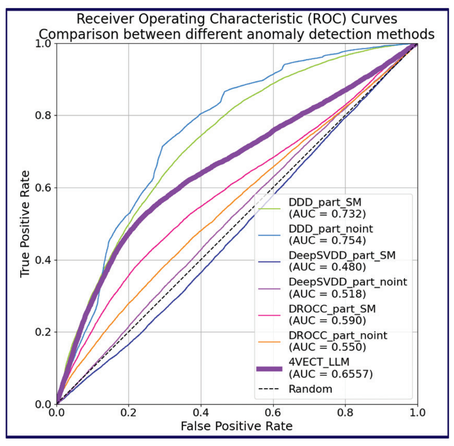


Inference and Anomaly detection

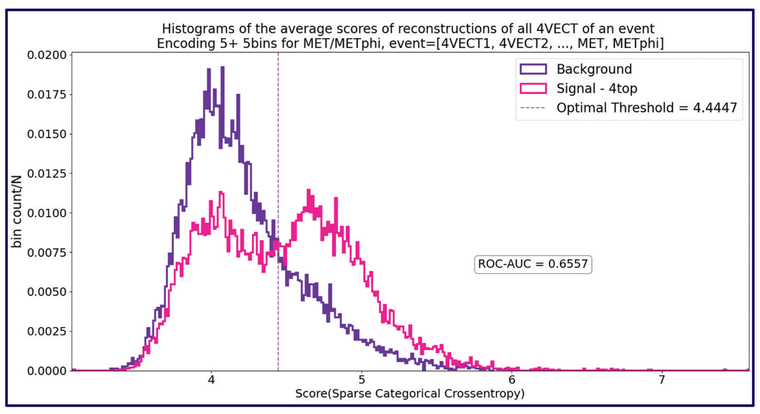
-The language model is trained exclusively on background-only events, where the objective is to reconstruct a deliberately masked particle within each event. This task forces the model to learn the underlying structure of standard model processes, effectively capturing the correlations and distributions typical of background physics.

-At inference, the trained model is applied to events from the control region (CR), which includes both background and a small admixture of signal-like events. For each event, the model predicts the masked particle and assigns a reconstruction score that reflects how consistent the predicted particle is with background-like behavior.

ROC



HISTO



-An anomaly threshold is defined by selecting a score that best separates CR background from CR signal, based on their reconstruction score distributions.

-Finally, we apply the model to events from the signal region (SR). Events with reconstruction scores that fall below the anomaly threshold are classified as anomalies, and are treated as signal

Nikhef

UNIVERSITEIT VAN AMSTERDAM

References:
[1] modified from arXiv:1706.03762
[2] modified from https://atlas.cern/Discover/Detector
[3] arXiv:2211.05143v2

EVENT TOKENIZATION AND NEXT-TOKEN PREDICTION FOR ANOMALY DETECTION AT THE LHC

Ambre Visive[1,2], Roberto Ruiz de Austri[3], Clara Nellist[1,2], Sascha Caron[1,4], Polina Moskvitina[1,4]

1. Nikhef, Amsterdam, Netherlands
2. University of Amsterdam, Amsterdam, Netherlands
3. Valencia University, IFIC, Valencia, Spain
4. Radboud University, Nijmegen, Netherlands

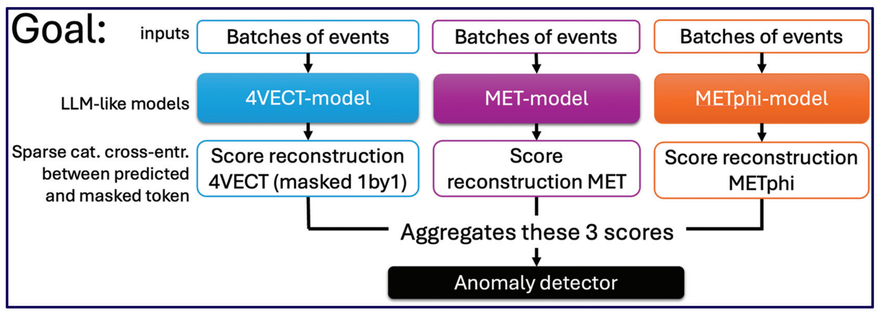
An LLM-like Model

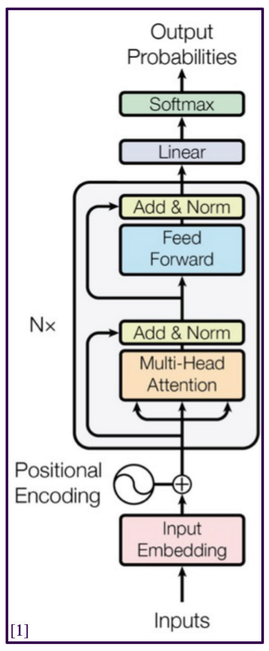
-**We would like to build an LLM-like model that correctly reconstructs a «hidden» particle from an event**.

-We are using **encoder-only transformers**, that use (sequences of) tokens as input(s) and have ~260 trainable parameters.

-The input of the model is a sequence of particles: an event.

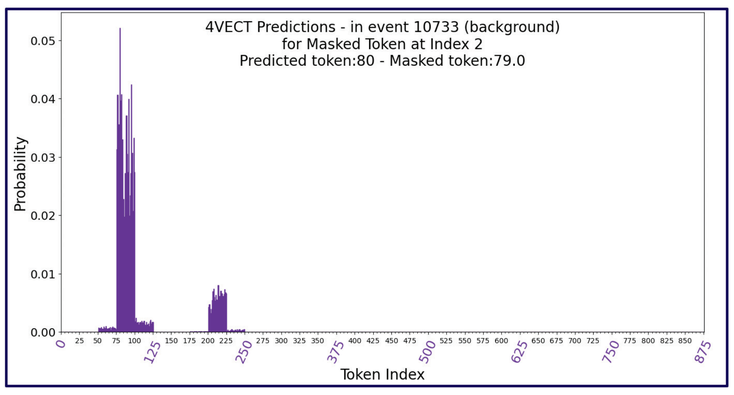
view the embedding process of our model





-During training and inference, we mask one particle from that event to do a next-token-like prediction.

-The output of our model is a two-dimensional tensor containing for each particle of the input event, the probabilities for each token to represent that particle.



Tokenization strategy

-The initial data set (Dark Machines Collaboration [2]), contains for each event and for up to 18 particles of that event, the type of the particle including its charge, its p_t, φ and η, and the missing transverse energy of the event (MET) and its azimutal angle(METphi). The events are 0-padded so all events are the same length.

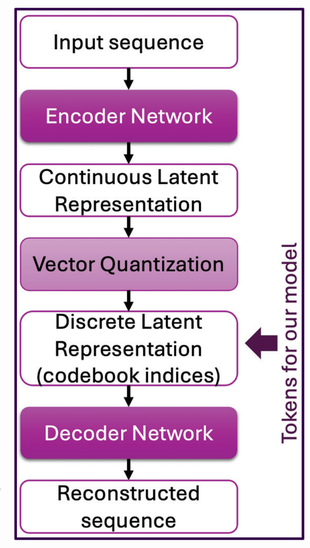
-Since our model needs tokens as input, the particle physics dataset had to be tokenized and the 0-pads are masked.

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

-We tried several approaches to tokenize the events by hand, including the tokenization of the 4-vectors of a particle. The example above is obtained by categorizing each property of a particle in 5 bins.

-We are currently trying to do a tokenization with VQ-VAE to see if it improves the performace of our model.

-Special attention is also given to the selection of information given to the model (MET, METphi, number of jets, of leptons ...).

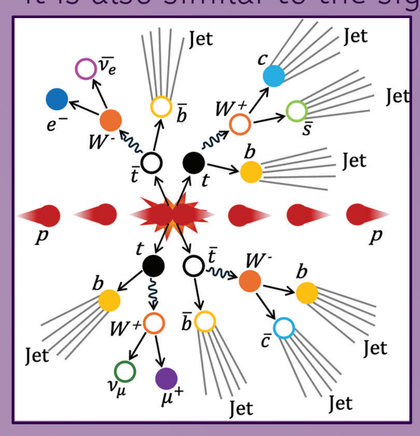


MORE ABOUT CONTEXT

4-top, ttW, ttWW, ttZ, ttH

-**4-top-quark events decay into 4 to 12 jets and 0 to 4 charged leptons**. Its signature shares similarities with a ttW or a ttWW event.

-It is also similar to the signature of a ttZ event, since Z bosons decay mostly into jets or leptonically.



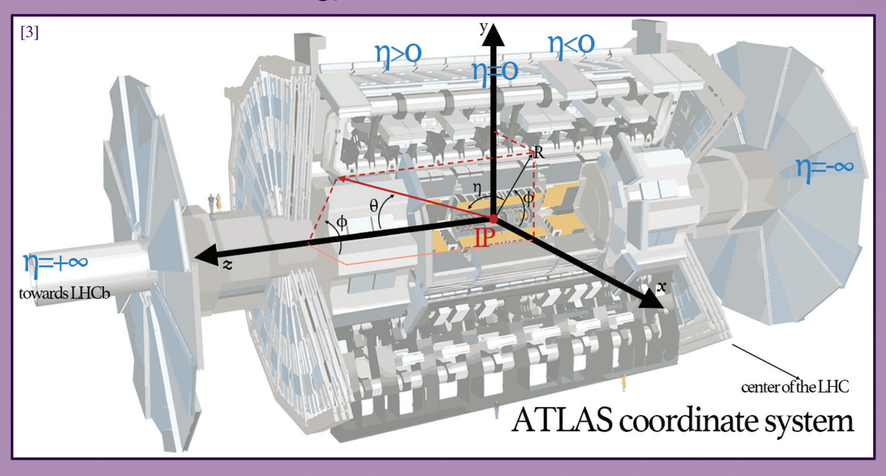
-Lastly, it is similar to the signature of a ttH event, as the Higgs boson can decay into 2 jets, a W- or a Z-pair.

-**ttW, ttWW, ttZ & ttH will be the background events**.

ATLAS detector

-ATLAS is one of the four detectors of the LHC at CERN.

-The dataset, from the Dark Machines Collaboration [2], is from **simulated proton-proton collisions** inside the ATLAS detector (as it was during RUN 2) and at the center of mass energy of 13 TeV.

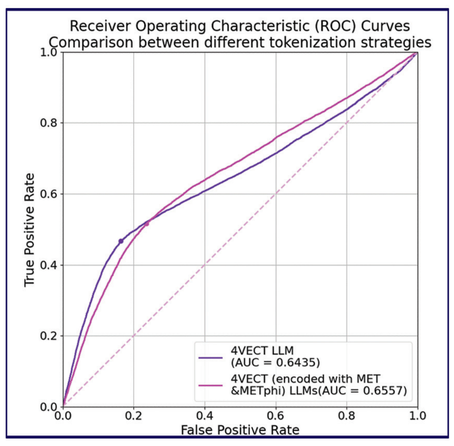


Inference and Anomaly detection

-The language model is trained exclusively on background-only events, where the objective is to reconstruct a deliberately masked particle within each event. This task forces the model to learn the underlying structure of standard model processes, effectively capturing the correlations and distributions typical of background physics.

-At inference, the trained model is applied to events from the control region (CR), which includes both background and a small admixture of signal-like events. For each event, the model predicts the masked particle and assigns a reconstruction score that reflects how consistent the predicted particle is with background-like behavior.

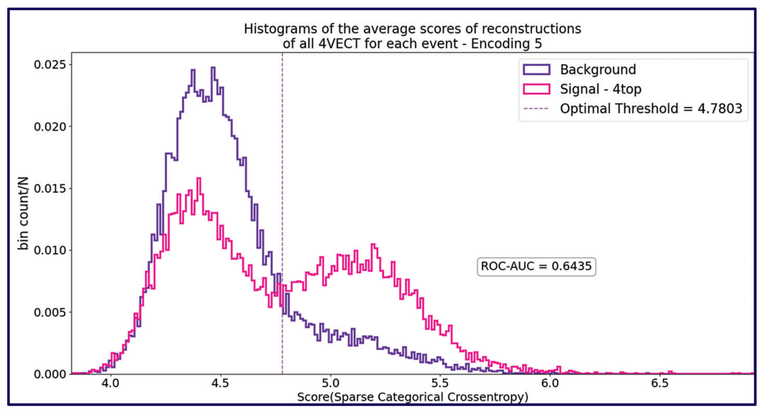
ROC



-An anomaly threshold is defined by selecting a score that best separates CR background from CR signal, based on their reconstruction score distributions.

-Finally, we apply the model to events from the signal region (SR). Events with reconstruction scores that fall below the anomaly threshold are classified as anomalies, and are treated as signal

HISTO



References:
[1] modified from arXiv:1706.03762
[2] modified from https://atlas.cern/Discover/Detector
[3] arXiv:2211.05143v2

EVENT TOKENIZATION AND NEXT-TOKEN PREDICTION FOR ANOMALY DETECTION AT THE LHC

Ambre Visive[1,2], Roberto Ruiz de Austri[3], Clara Nellist[1,2], Sascha Caron[1,4], Polina Moskvitina[1,4]

1. Nikhef, Amsterdam, Netherlands
2. University of Amsterdam, Amsterdam, Netherlands
3. Valencia University, IFIC, Valencia, Spain
4. Radboud University, Nijmegen, Netherlands

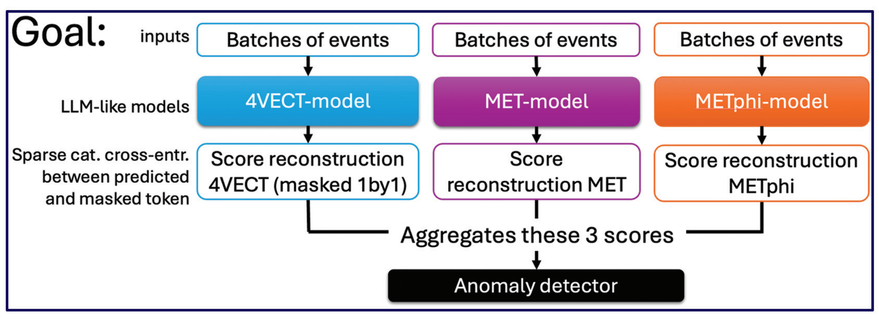
An LLM-like Model

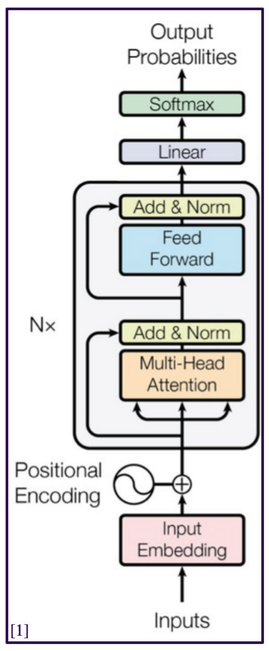
-**We would like to build an LLM-like model that correctly reconstructs a «hidden» particle from an event**.

-We are using **encoder-only transformers**, that use (sequences of) tokens as input(s) and have ~260 trainable parameters.

-The input of the model is a sequence of particles: an event.

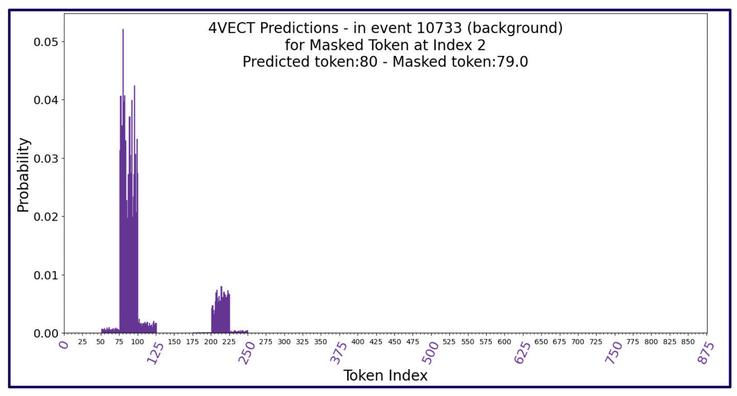
view the embedding process of our model





-During training and inference, we mask one particle from that event to do a next-token-like prediction.

-The output of our model is a two-dimensional tensor containing for each particle of the input event, the probabilities for each token to represent that particle.



Tokenization strategy

-The initial data set (Dark Machines Collaboration [2]), contains for each event and for up to 18 particles of that event, the type of the particle including its charge, its p_t, φ and η, and the missing transverse energy of the event (MET) and its azimutal angle(METphi). The events are 0-padded so all events are the same length.

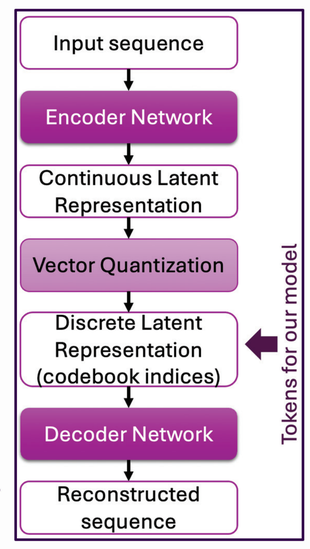
-Since our model needs tokens as input, the particle physics dataset had to be tokenized and the 0-pads are masked.

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

-We tried several approaches to tokenize the events by hand, including the tokenization of the 4-vectors of a particle. The example above is obtained by categorizing each property of a particle in 5 bins.

-We are currently trying to do a tokenization with VQ-VAE to see if it improves the performace of our model.

-Special attention is also given to the selection of information given to the model (MET, METphi, number of jets, of leptons ...).

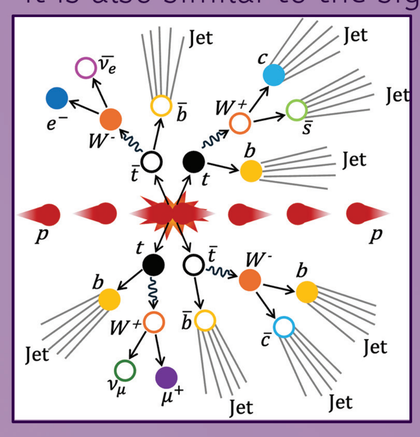


MORE ABOUT CONTEXT

4-top, ttW, ttWW, ttZ, ttH

-**4-top-quark events decay into 4 to 12 jets and 0 to 4 charged leptons**. Its signature shares similarities with a ttW or a ttWW event.

-It is also similar to the signature of a ttZ event, since Z bosons decay mostly into jets or leptonically.



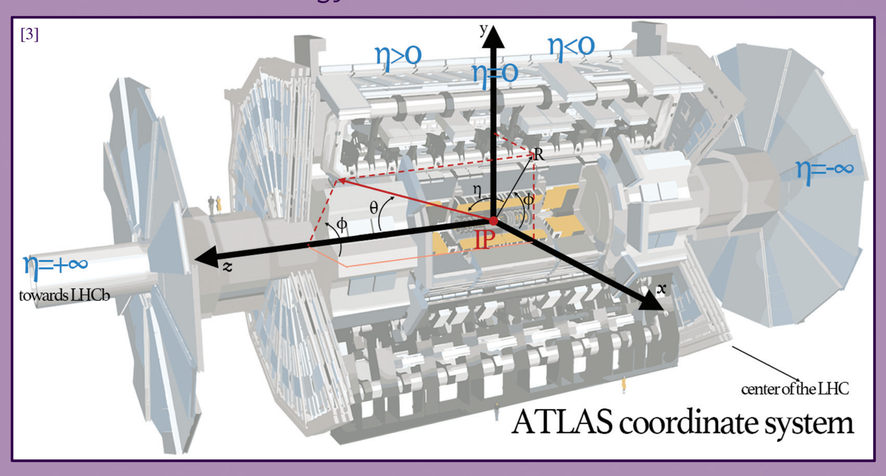
-Lastly, it is similar to the signature of a ttH event, as the Higgs boson can decay into 2 jets, a W- or a Z-pair.

-**ttW, ttWW, ttZ & ttH will be the background events**.

ATLAS detector

-ATLAS is one of the four detectors of the LHC at CERN.

-The dataset, from the Dark Machines Collaboration [2], is from **simulated proton-proton collisions** inside the ATLAS detector (as it was during RUN 2) and at the center of mass energy of 13 TeV.

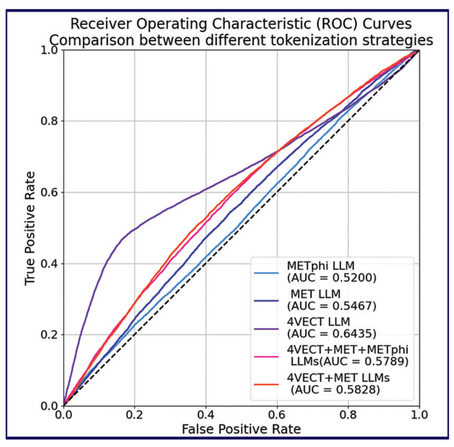


Inference and Anomaly detection

-The language model is trained exclusively on background-only events, where the objective is to reconstruct a deliberately masked particle within each event. This task forces the model to learn the underlying structure of standard model processes, effectively capturing the correlations and distributions typical of background physics.

-At inference, the trained model is applied to events from the control region (CR), which includes both background and a small admixture of signal-like events. For each event, the model predicts the masked particle and assigns a reconstruction score that reflects how consistent the predicted particle is with background-like behavior.

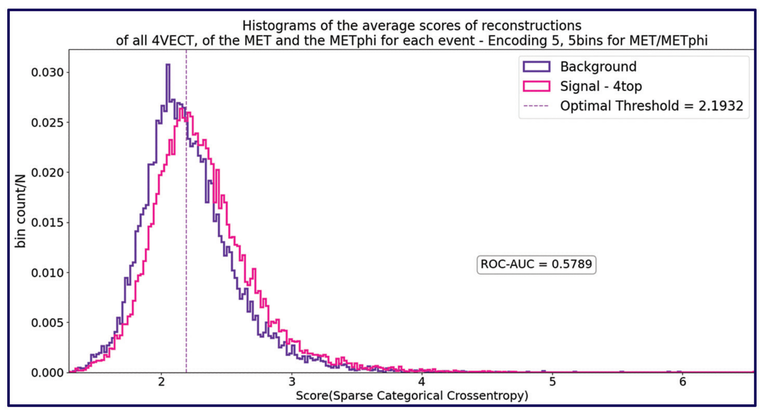
ROC



-An anomaly threshold is defined by selecting a score that best separates CR background from CR signal, based on their reconstruction score distributions.

-Finally, we apply the model to events from the signal region (SR). Events with reconstruction scores that fall below the anomaly threshold are classified as anomalies, and are treated as signal

HISTO



References:
[1] modified from arXiv:1706.03762
[2] modified from https://atlas.cern/Discover/Detector
[3] arXiv:2211.05143v2