Multi-scale cross attention transformer encoder for τ lepton pair invariant mass reconstruction

Valentina Camagni[12]

Simone Gennai[2], Pietro Govoni[12], Francesco Brivio[12], Paolo Dini[12]

1 Università Milano Bicocca

2 INFN - Milano Bicocca

AI –INFN – User Forum

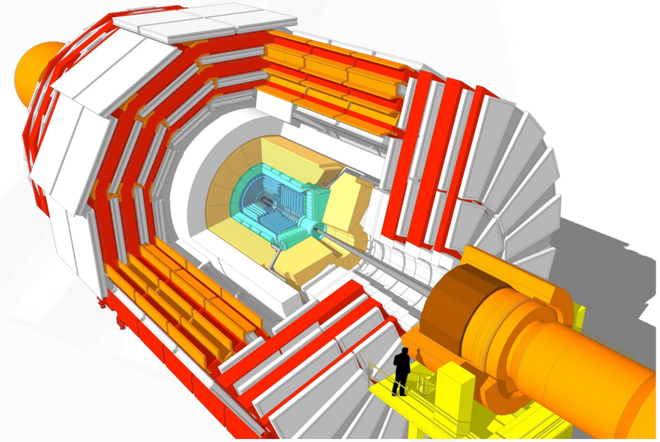
June 11/12, 2024

The Higgs Boson as a probe for new physics

- The latest highlight in the success story of the Standard Model (SM) is the **Higgs boson (H)**, discovered in 2012 at the LHC by the collaborations of ATLAS and CMS experiments

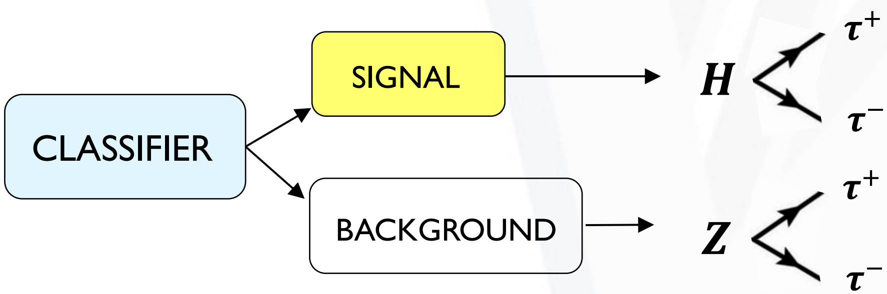
- The precise characterization of properties and couplings of H is now of utmost importance, since deviations from SM predictions may point to physics beyond the SM (BSM)

- One of the most promising method to directly probe the H self-coupling is via the study of H pair production (HH) in the $b\bar{b}\tau^+\tau^-$ decay channel



CMS detector [1]

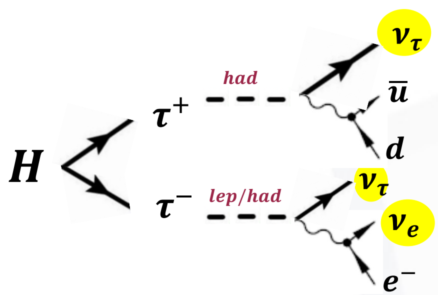
However, identifying the $H \rightarrow \tau^+\tau^-$ signal is challenging due to the presence of irreducible background



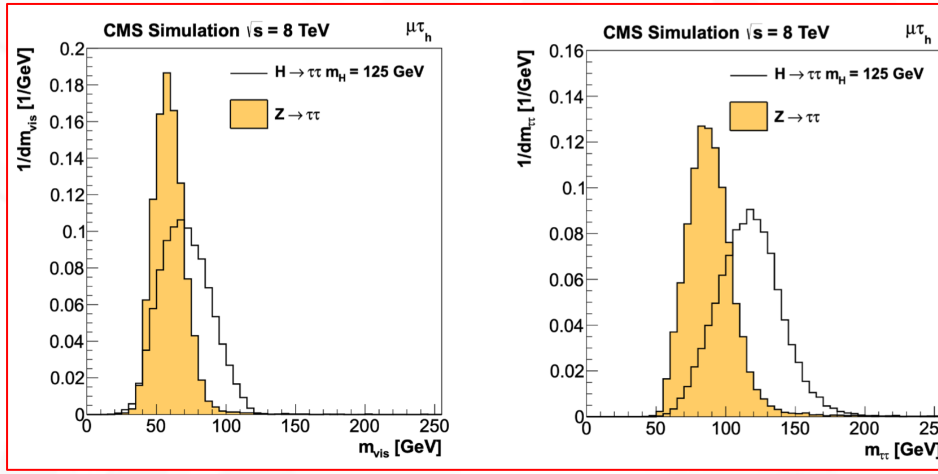
Fundamental discriminant variable:

the invariant mass of the di-τ system

Analysis of interest
di-τ invariant mass reconstruction



The presence of neutrinos from tau decay prevent the full reconstruction of the di-tau system invariant mass, allowing only the reconstruction of the visible tau-decay products ($m_{\tau\tau}^{VIS}$) whose low resolution doesn't help in the signal discrimination task



SVFit algorithm [2]

- Improves the $m_{\tau\tau}$ resolution only marginally
- High computational time

⇩

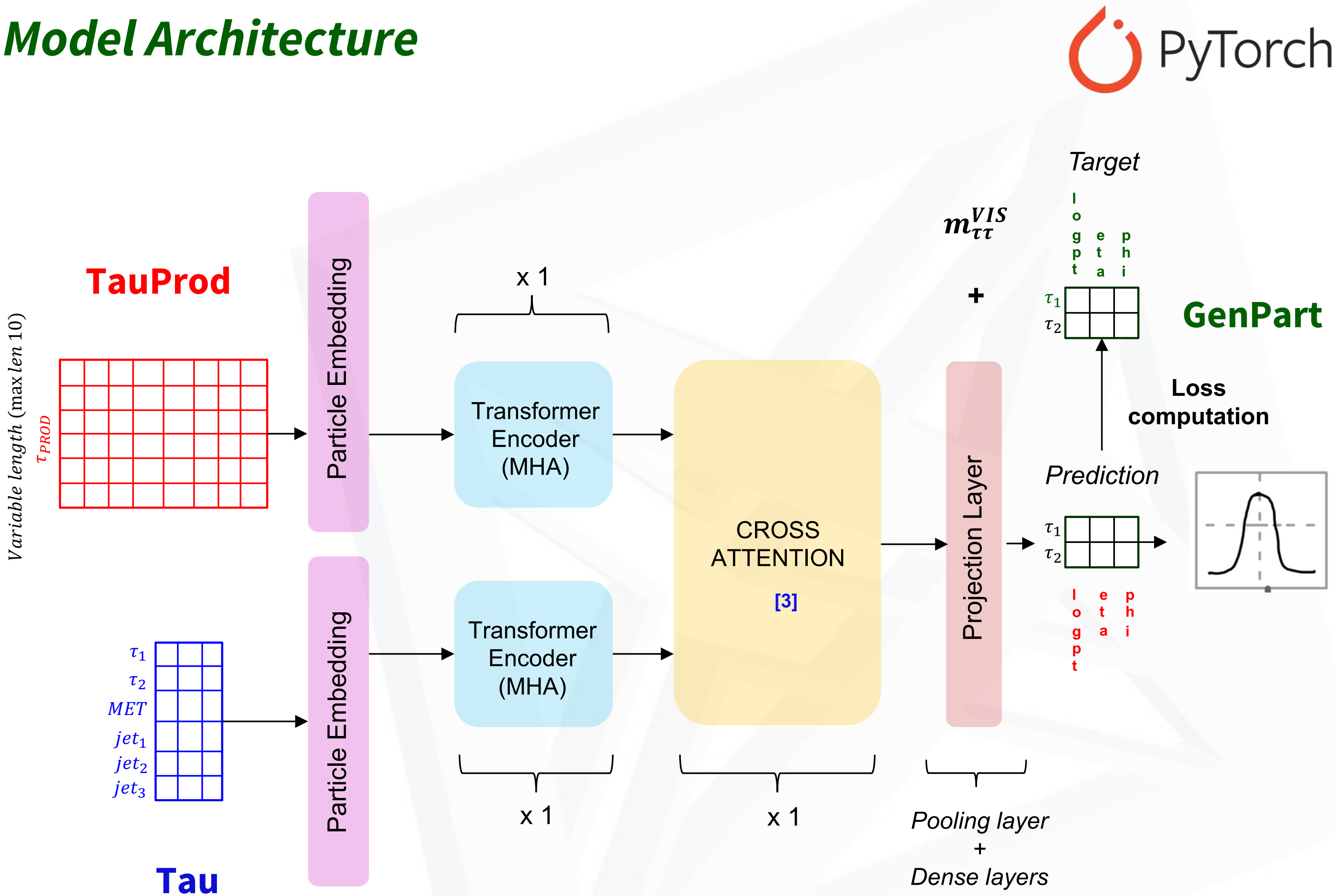
Tau Pair Mass Transformer
TPMT

Objective: Reconstruct the four-momentum of each τ particle before decay to accurately estimate the invariant mass and retrieve the kinematics of the parent particle

GOAL

Understand the model functionality on $H \rightarrow \tau^+\tau^-$ and $Z \rightarrow \tau^+\tau^-$
and considering only taus that decay hadronically so far

Model Architecture



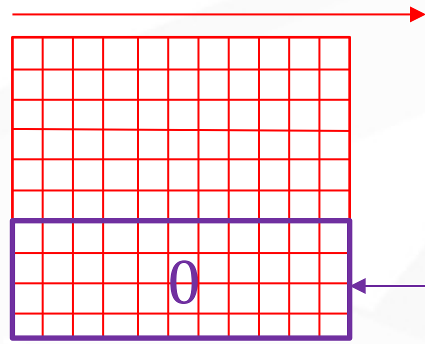
Training time: $\sim 2\ min\ per\ epoch$ ($\sim 40\%$ GPU Tesla T4 usage)
Inference time: $\sim 2{\times}10^{-3}\ s\ per\ event$
Number of parameters: $\sim 1\ M$

Input features (1)

① **TauProd**

Decay products of the two taus

Shape: $(10, 12)$



If an event has less than 10 tau products

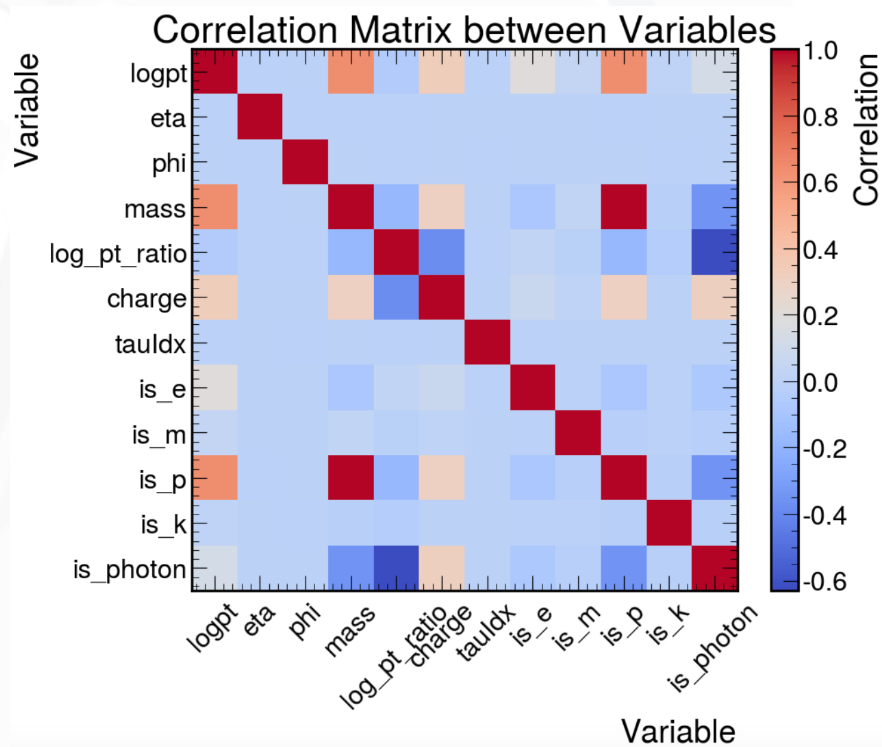
PADDING

*

$$logp_t$$
$$\eta$$
$$\phi$$
$$m$$
$$log\left(\frac{p_T}{p_{T(\tau)}}\right)$$
$$charge$$

$$\boxed{\begin{array}{c} tauIdx \\ is\_electron \\ is\_muon \\ is\_pion \\ is\_kaon \\ is\_photon \end{array}}$$

Categorical variables from particle ID



Input features (2)

② **Tau**

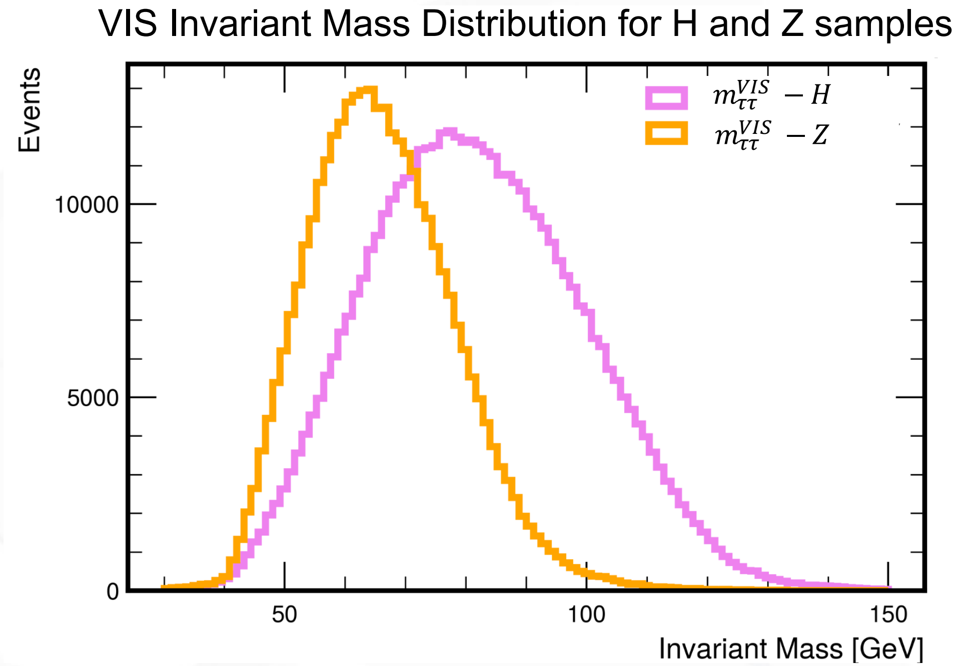
Shape: $(6, 3)$

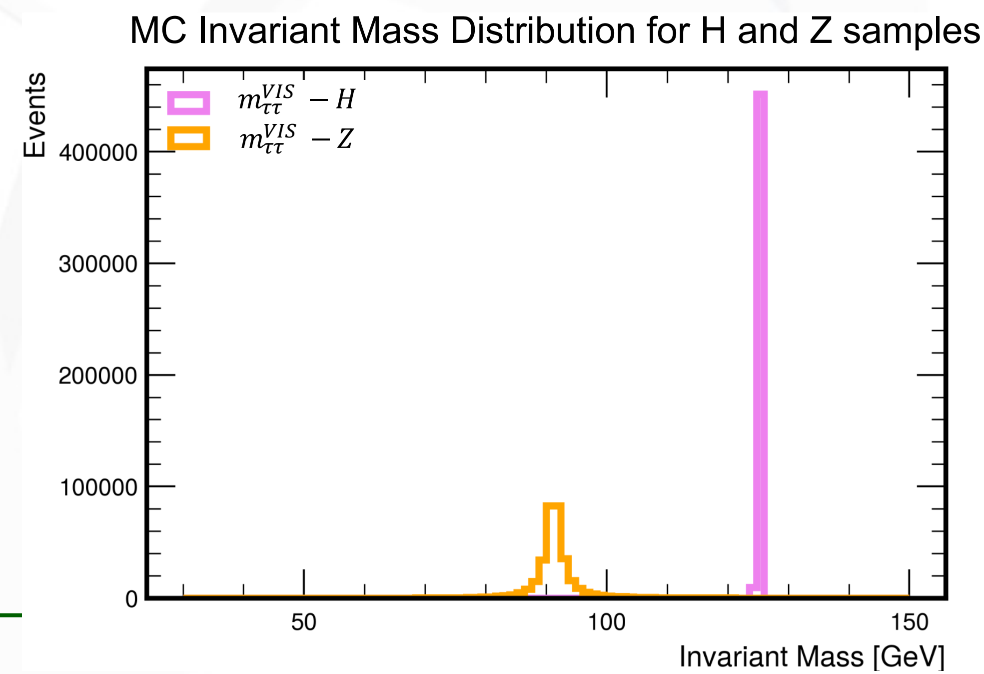
	$log p_T$	η	ϕ
τ_1			
τ_2			
MET			
jet_1			
jet_2			
jet_3			

GenPart

Shape: $(2, 3)$

	$log p_T$	η	ϕ
τ_1			
τ_2			





mass from gen taus only
(no ISR/FSR considered)

Pre-processing steps

TAU SELECTION

At least 2 taus
- Gen matched
- Hadronic decay
- $p_T \geq 20$ GeV

JETS SELECTION

First 3 leading jets with $\Delta R(jet, tau) > 0.4$

(minimum p_T: 10 GeV)

VARIABLE ENCODING & FEATURE ENGENEERING

- Definition of new variables
- Order TauProd with respect to their p_T and padding with $\max\_len = 10$

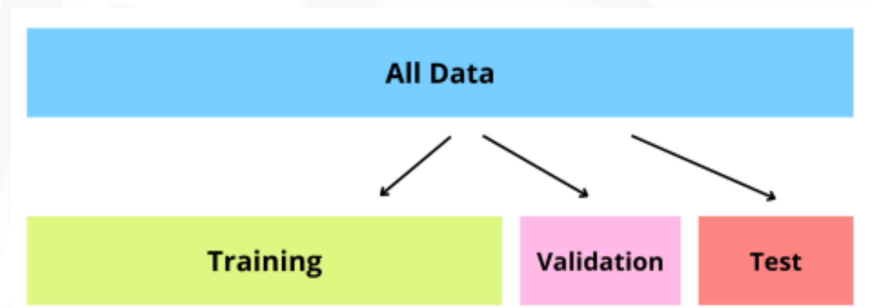
SPLIT IN TRAIN, TEST AND VALIDATION

H

Train size = 320 000
Validation size = 40 000
Test size = 40 000

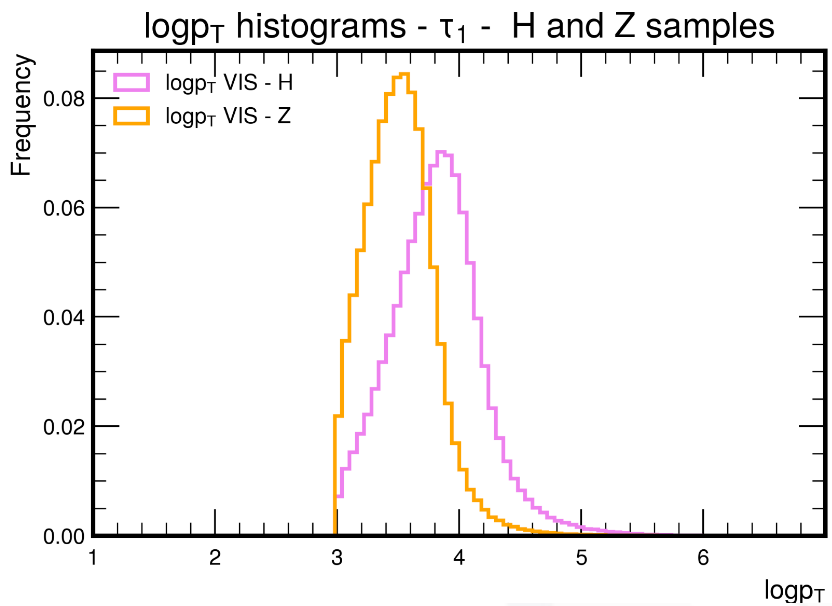
Z

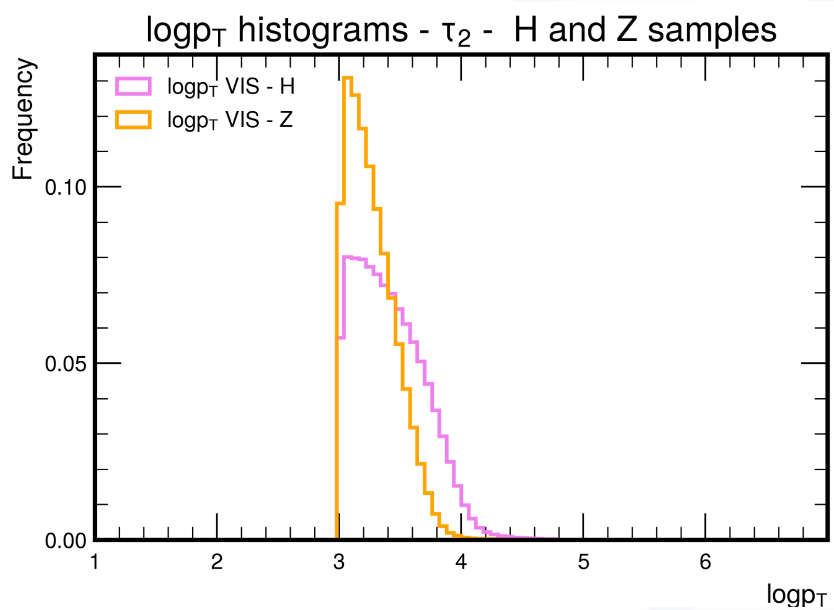
Train size = 256 000
Validation size = 32 000
Test size = 32 000



80% 10% 10%

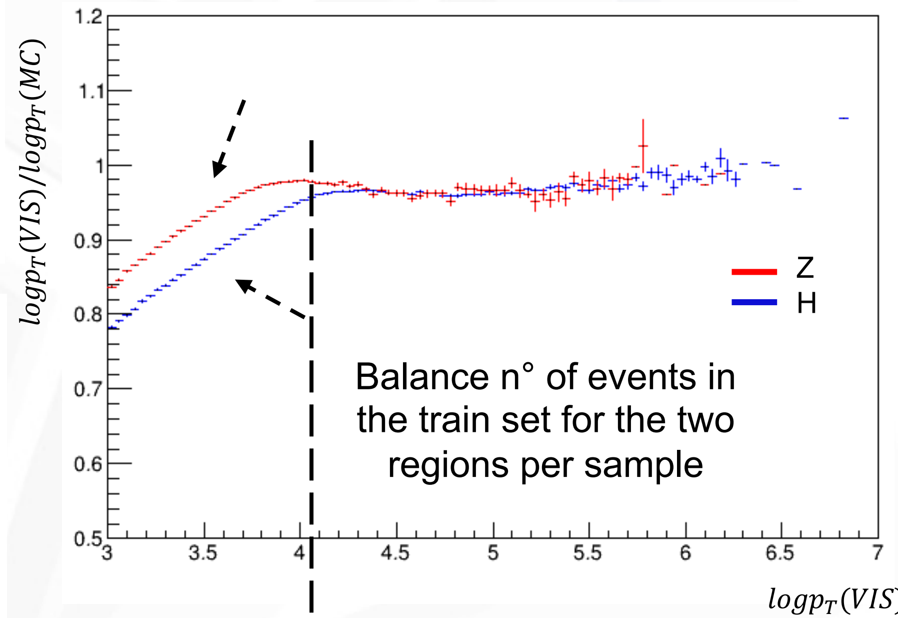
$logp_T$ histograms





There is a $logp_T$ transition region: for $logp_T < 4$, H and Z taus' $logp_T$ need a different scale factor

RATIO $logp_T^{VIS}$ and $logp_T^{MC}$ as a function of $logp_T^{VIS}$



***H*: 160 000**	***H*: 160 000**
***Z*: 128 000**	***Z*: 128 000**

Loss function

1. $Mean\ between\ MAE_{logp_T}, MAE_{\eta}, MAE_{\phi}$ $for\ the\ two\ taus$

+

2. $MAE\ between\ M_{\tau\tau}^{TPMT}\ and\ M_{\tau\tau}^{MC}$ (7% of the total loss)

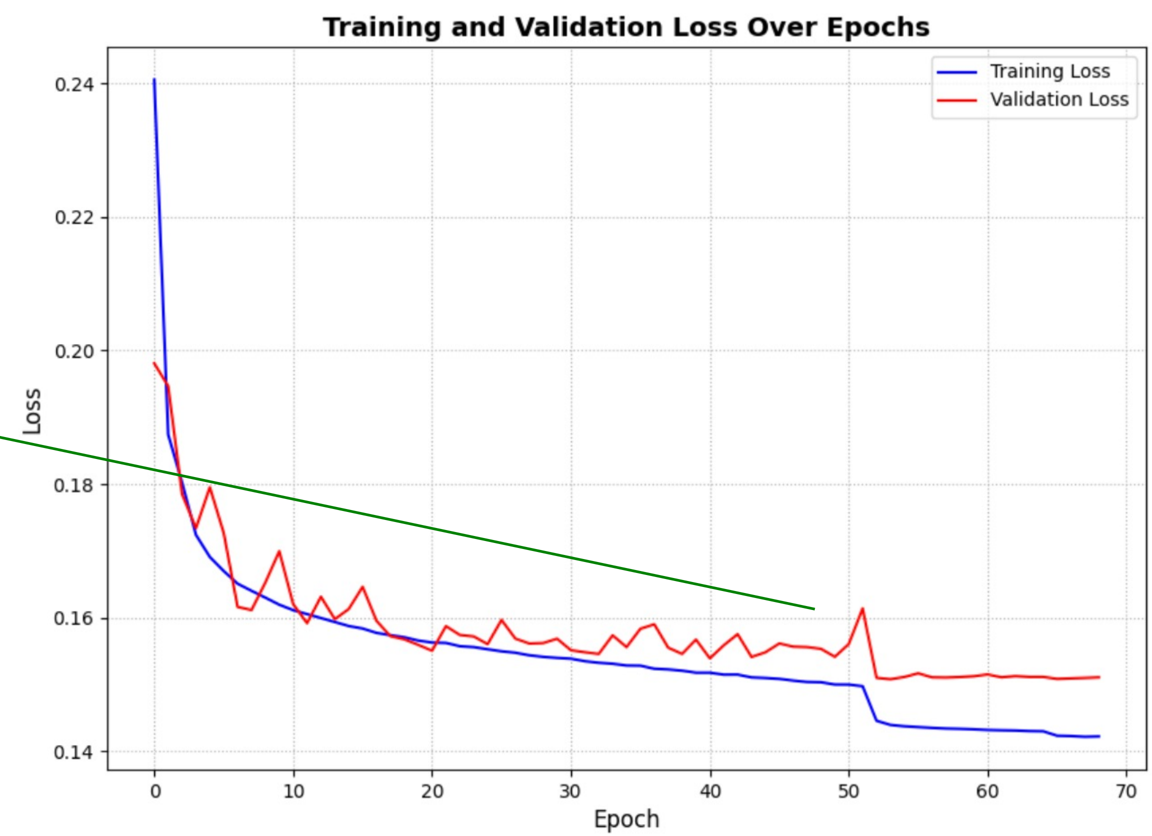
Hyper-parameters

Batch size = 128
Start learning rate = 10^{-4}
ReduceLROnPlateau (Patience = 10 epochs)
End epoch = 68
Early stopping (Patience = 15 epochs)

Mean Absolute Error

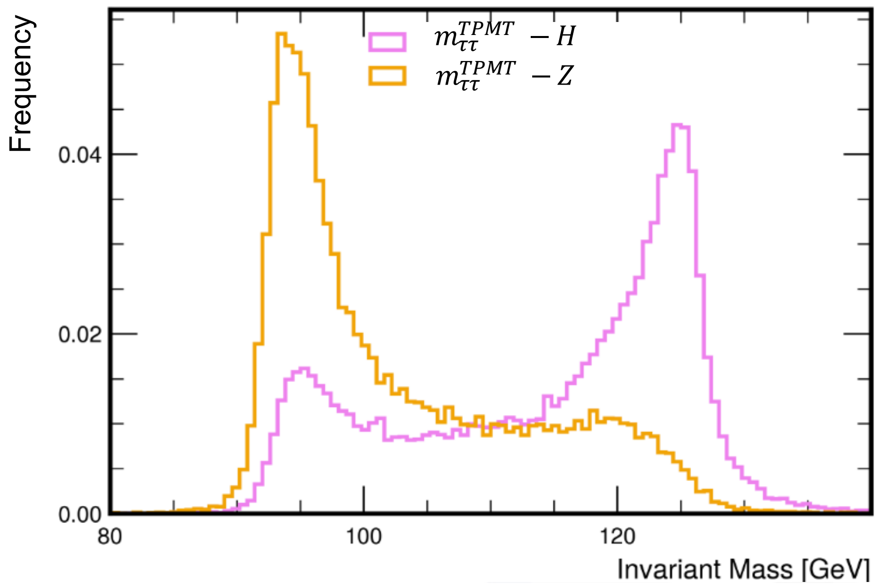
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Discontinuities due to learning rate decrease



Results $M_{\tau\tau}$

TPMT Invariant Mass Distribution for H and Z samples

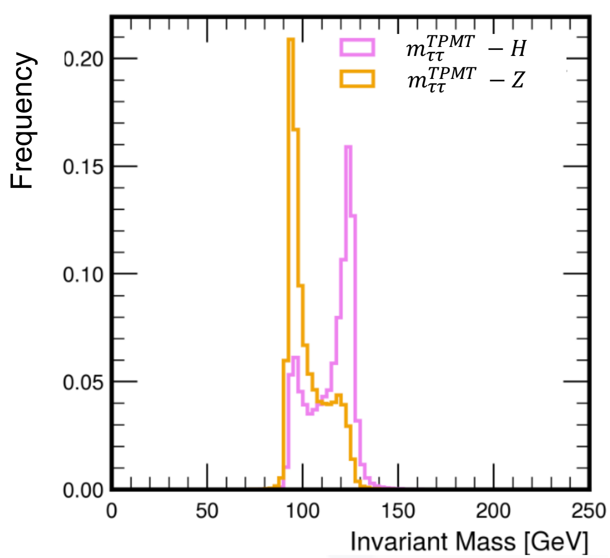


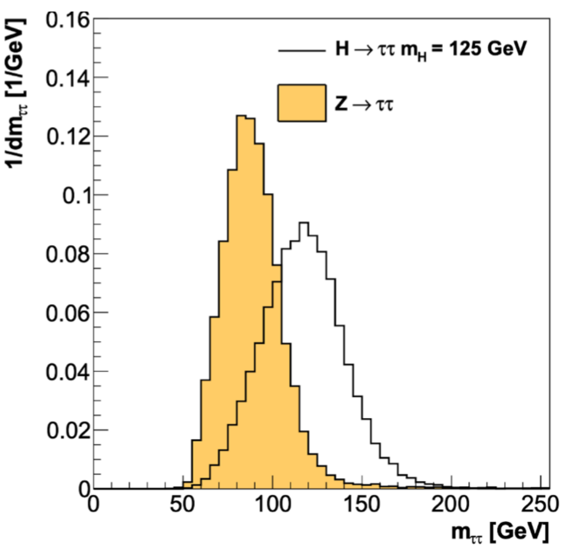
A high percentage of H is wrongly reconstructed compared to Z (even if numerically more represented)

⇩

Not related to η and ϕ since they remain unchanged (no need to predict them!)

Still better separation than SVFit !

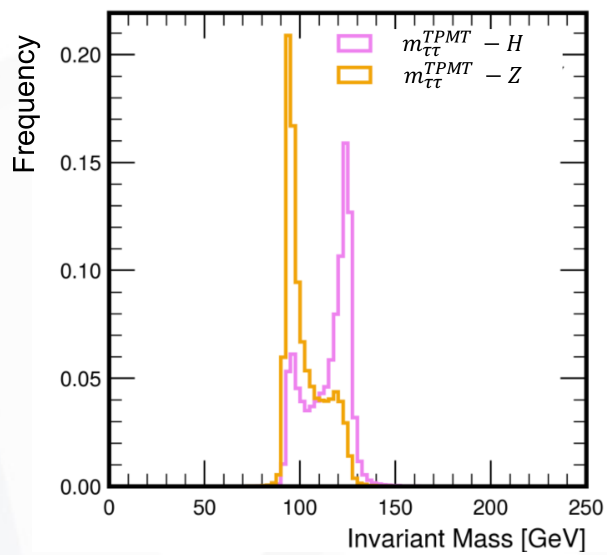




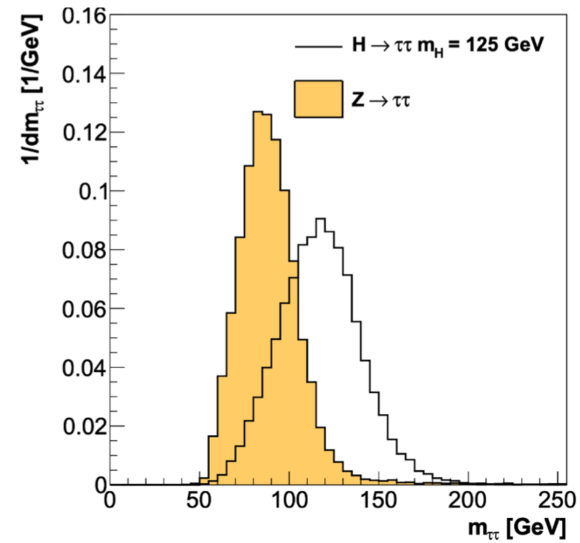
Conclusions

- **Transformer →** competitive approach for $m_{\tau\tau}$ reconstruction
- Better $m_{\tau\tau}$ resolution
- Computation time of $O(10^{-3}s)$ per event VS $O(1s)$ per event of SVFit

TPMT $m_{\tau\tau}$ for H and Z samples



SVFIT $m_{\tau\tau}$ for H and Z samples



Next steps

- ***Improve model training with flat mass samples and leptonically decaying taus***
- ***Model improvements***
  - Increase model depth
  - Analyze the learned information through the examination of the attention maps
  - Optimization through better initialization

References

[1] Castillo, Luis Roberto Flores. *The Search and Discovery of the Higgs Boson: A brief introduction to particle physics*. Morgan & Claypool Publishers, 2015.

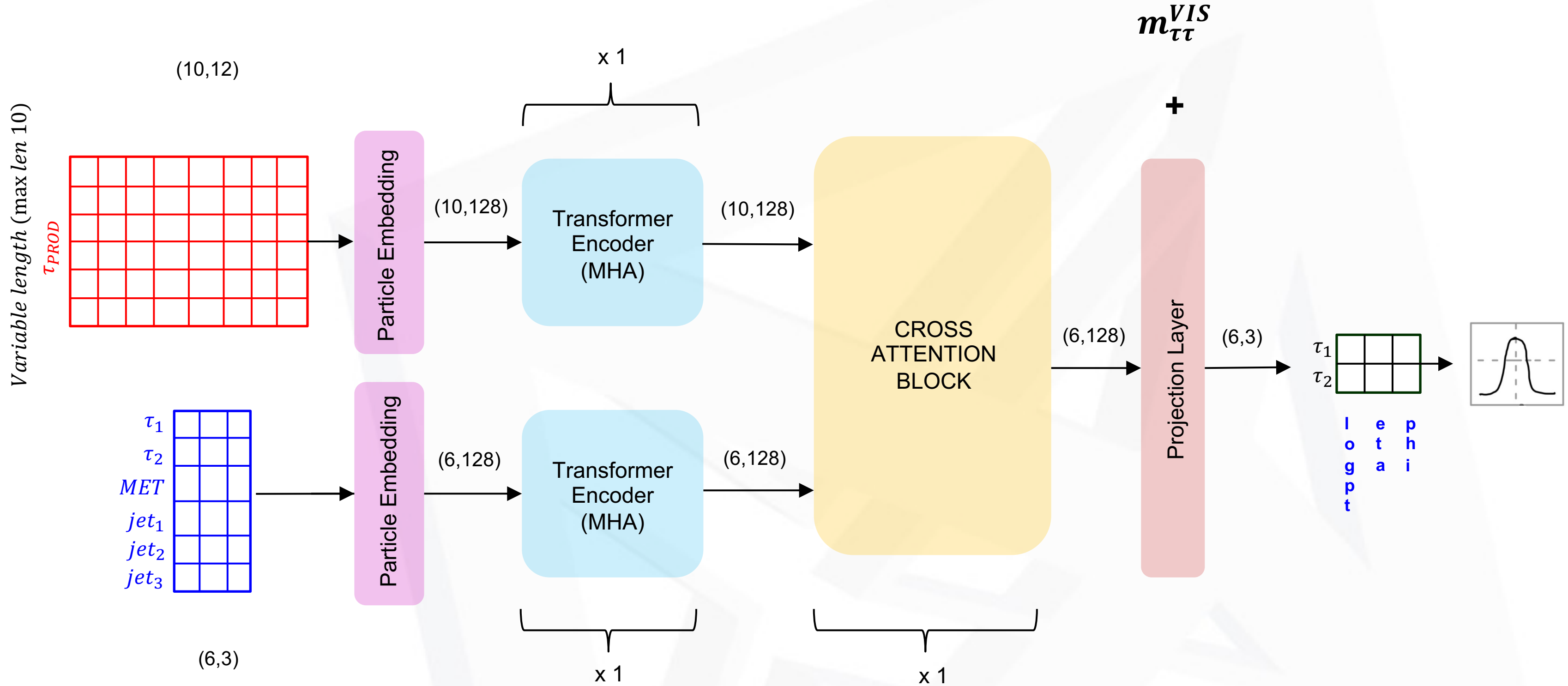
[2] Bianchini, Lorenzo, et al. "Reconstruction of the Higgs mass in H→ ττ events by dynamical likelihood techniques." *Journal of Physics: Conference Series*. Vol. 513. No. 2. IOP Publishing, 2014.

[3] Hammad, A., S. Moretti, and M. Nojiri. "Multi-scale cross-attention transformer encoder for event classification." *arXiv preprint arXiv:2401.00452* (2023).

Thank you for the attention

BACKUP

Model Architecture - Details

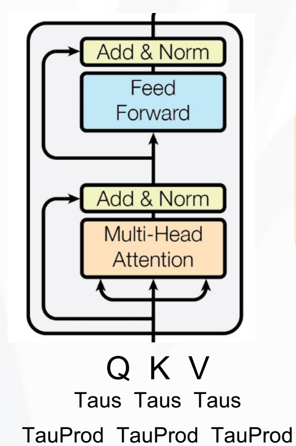


Particle Embedding

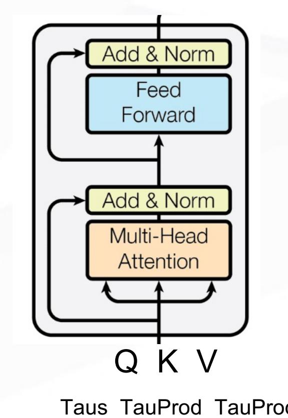
particle embedding of a dimension $d = 128$, encoded from the input particle features using a 3-layer MLP with $(128, 512, 128)$ nodes each layer with GELU nonlinearity, and LN is used in between for normalization

Transformer Encoder (MHA)

Padding Mask passed to key_padding_mask parameter of Multi-Head-Attention to mask padded entries



CROSS ATTENTION BLOCK



Projection Layer

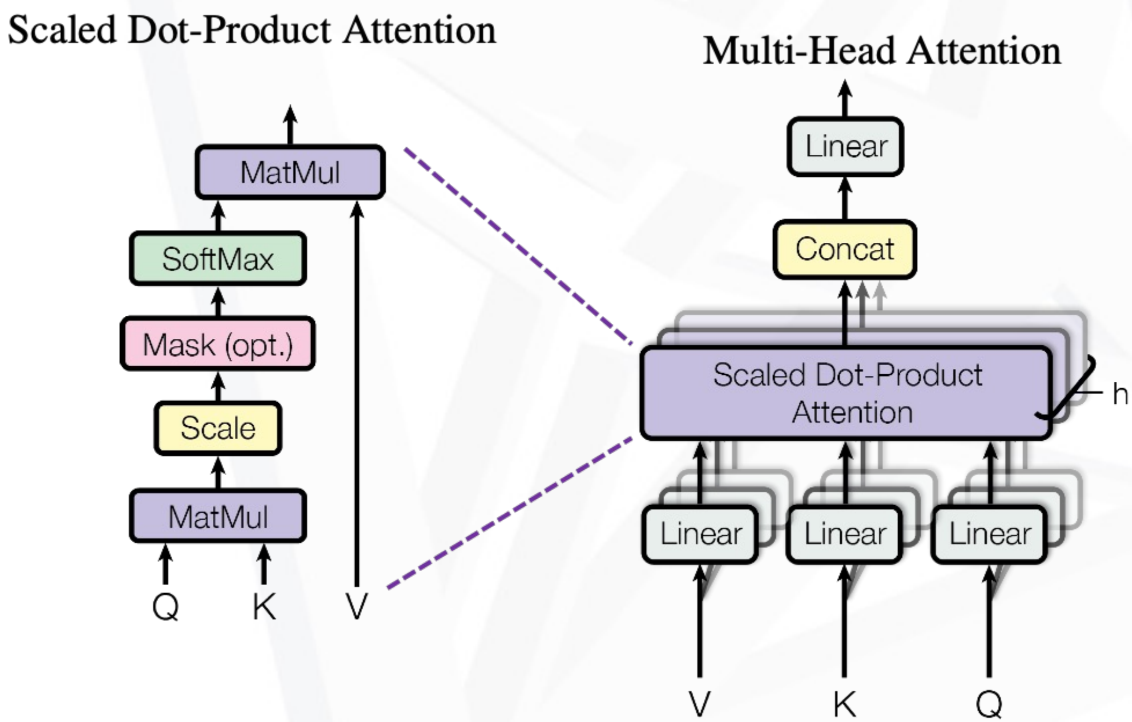
- AdaptiveAvgPool1D
- Concat pooled output with normalized $m_{\tau\tau}^{VIS}$
- Serie of linear layers

Scaled Dot - Product

$$Multihead(Q,K,V) = Concat(head_1, head_2, ..., head_h)W^O$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Self-Attention

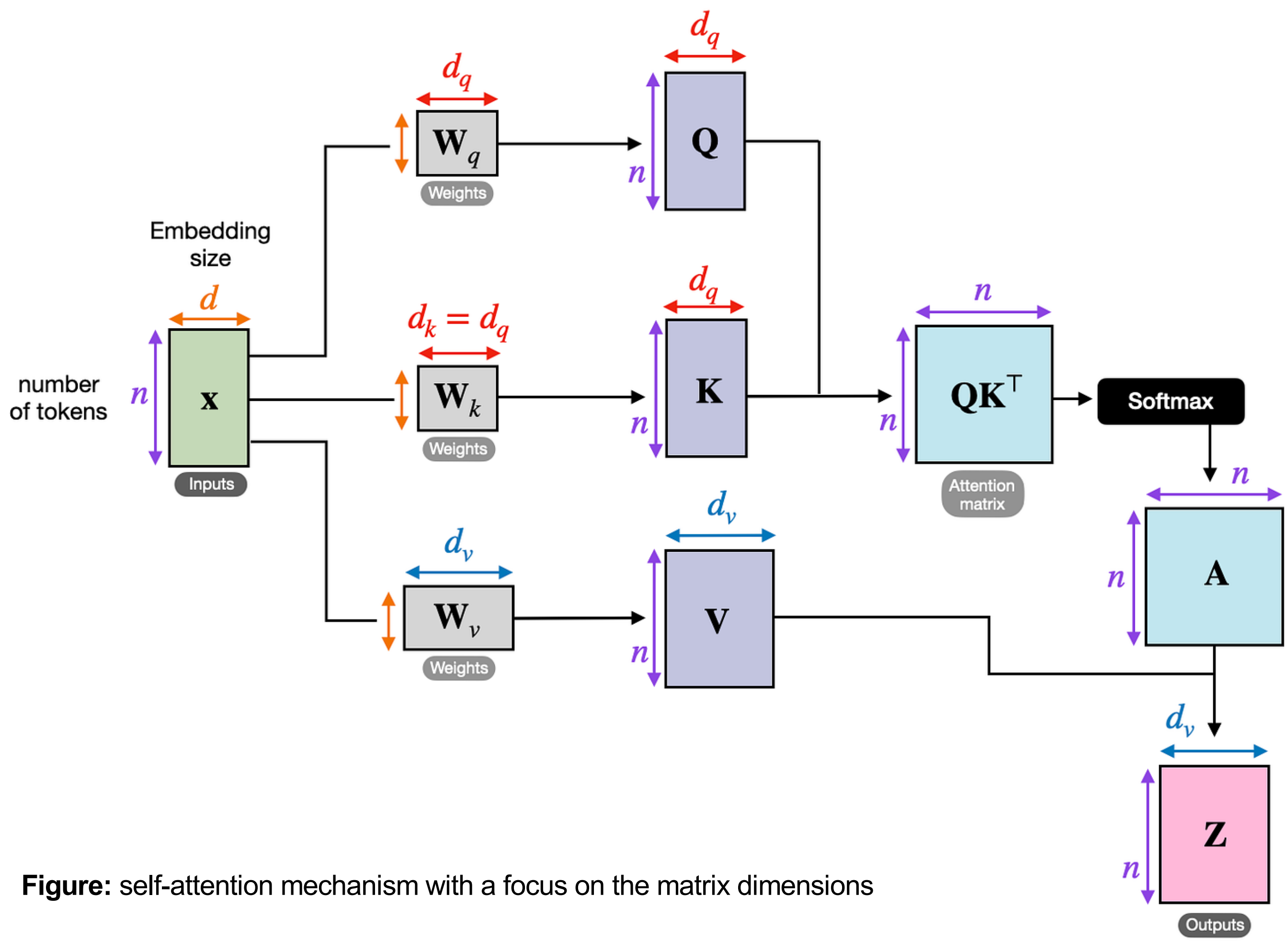


Figure: self-attention mechanism with a focus on the matrix dimensions

Cross-Attention

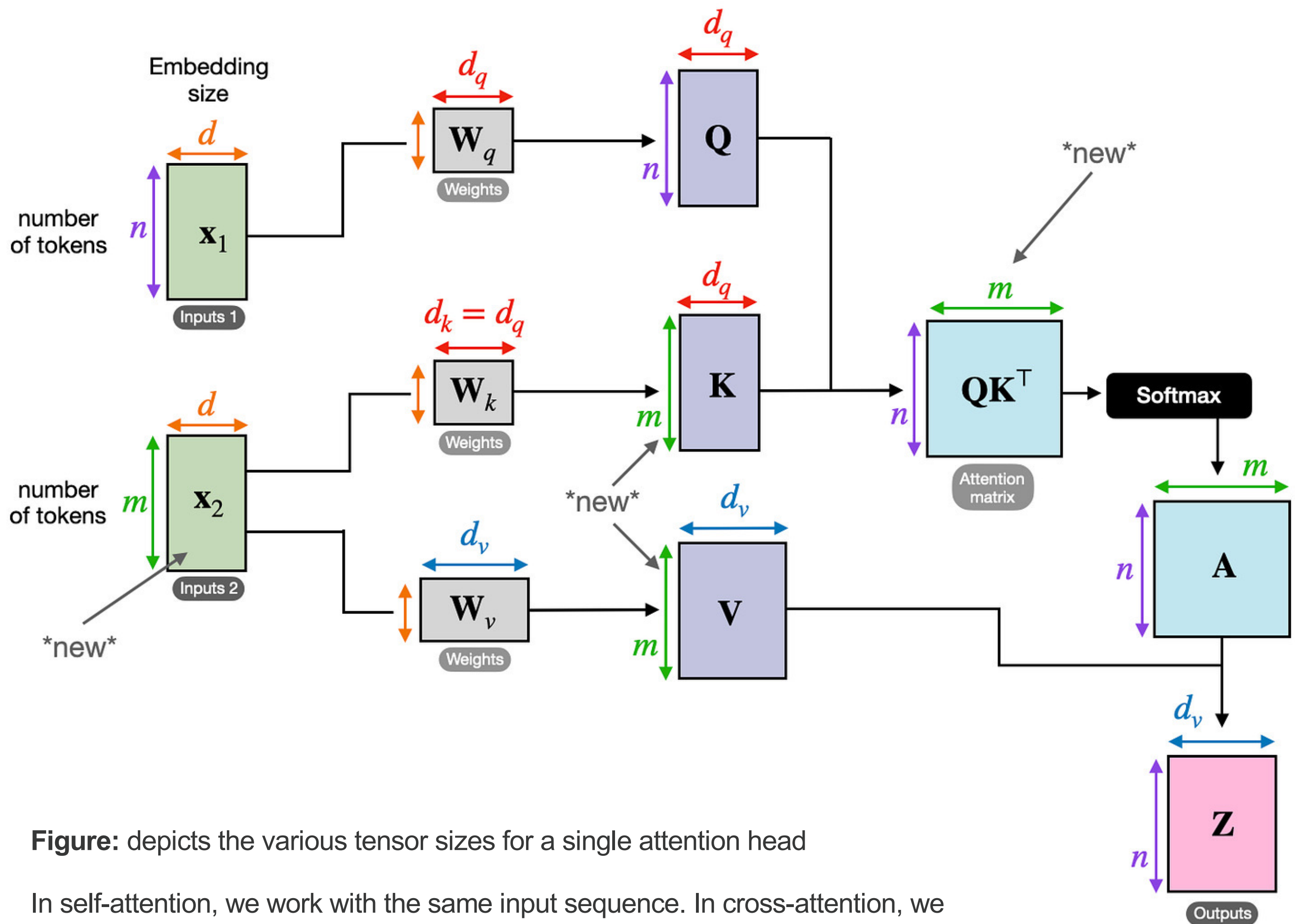


Figure: depicts the various tensor sizes for a single attention head

In self-attention, we work with the same input sequence. In cross-attention, we mix or combine two *different* input sequences. In the case of the original transformer architecture, that's the sequence returned by the encoder module and the input sequence being processed by the decoder part on the right. The two input sequences and can have different numbers of elements. However, their embedding dimensions must match.

arxiv

Multi-scale cross-attention transformer encoder for event classification

A. Hammad[a], S. Moretti[b,c] and M. Nojiri[a,d,e]

[a]Theory Center, IPNS, KEK, 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan.
[b]School of Physics and Astronomy, University of Southampton, Highfield, Southampton, UK.
[c]Department of Physics & Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden.
[d]The Graduate University of Advanced Studies (Sokendai), 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan
[e]Kavli IPMU (WPI), University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan

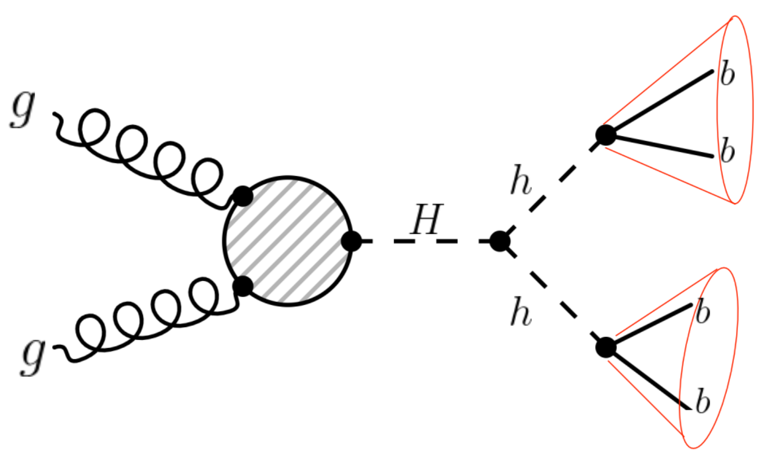
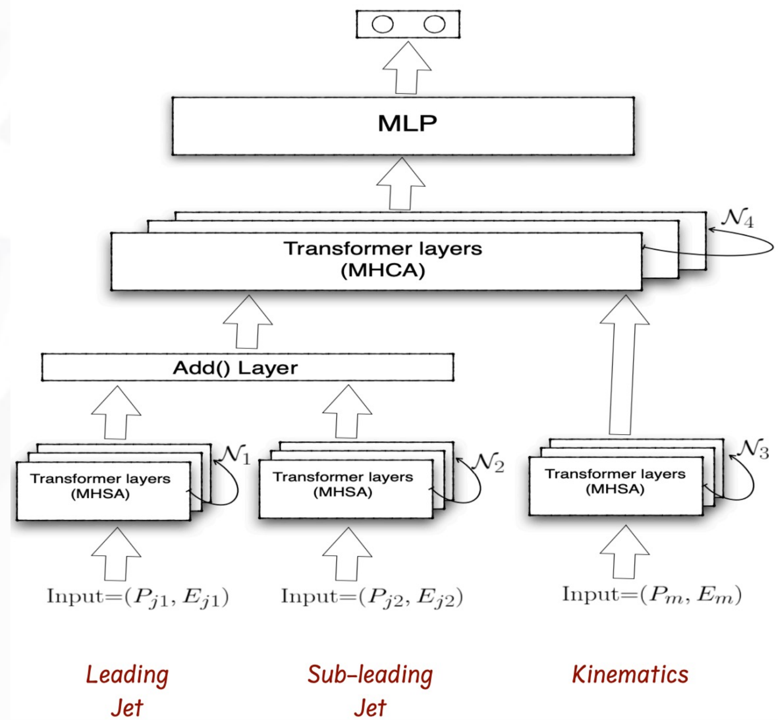


Figure 2: Feynman diagram for the signal process.



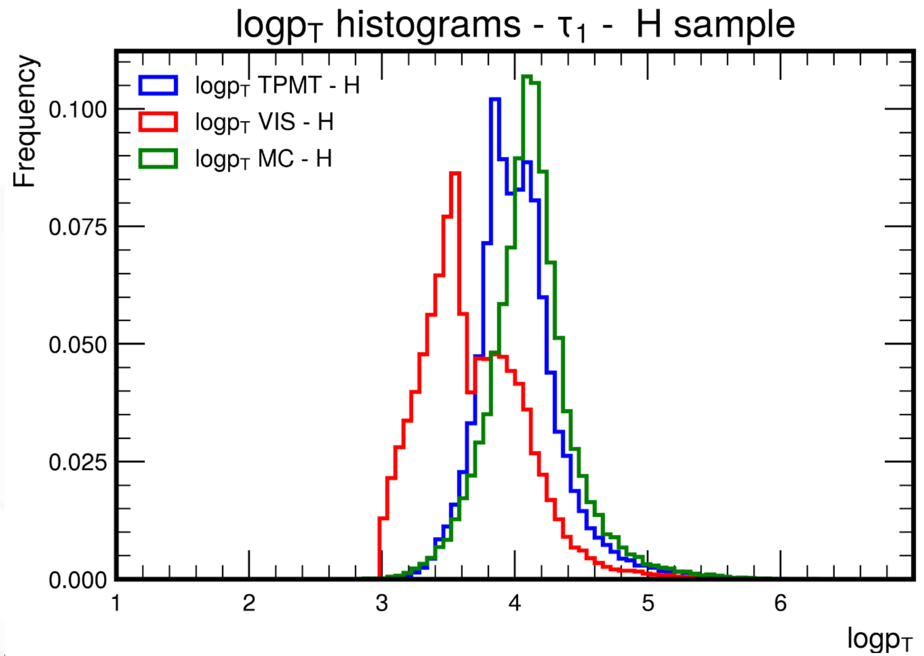
Results
$logp_T$ histograms

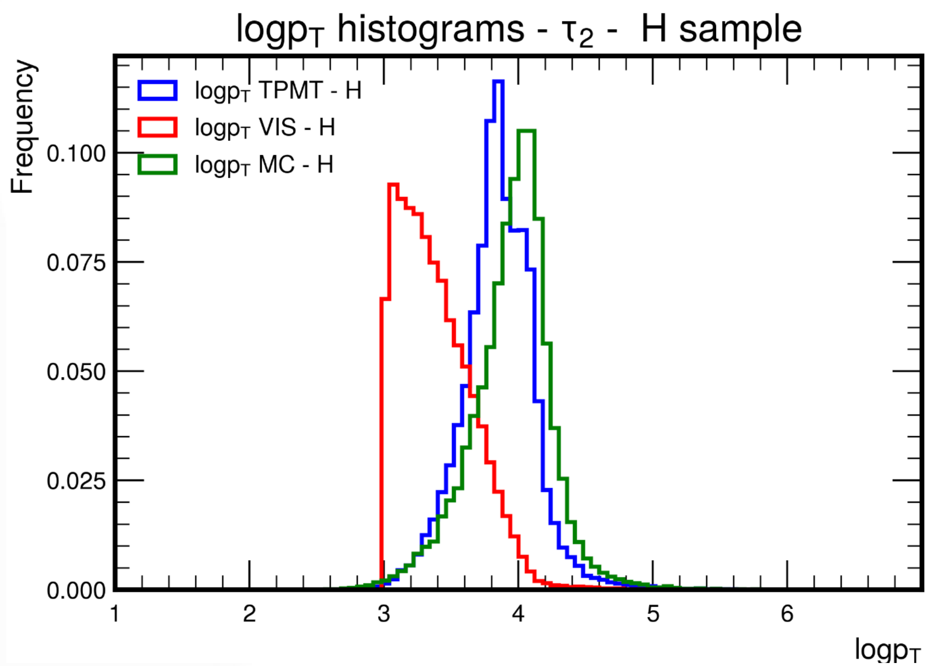
Inference on H test set

τ_1: $MAE = 0.153$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

τ_2: $MAE = 0.165$





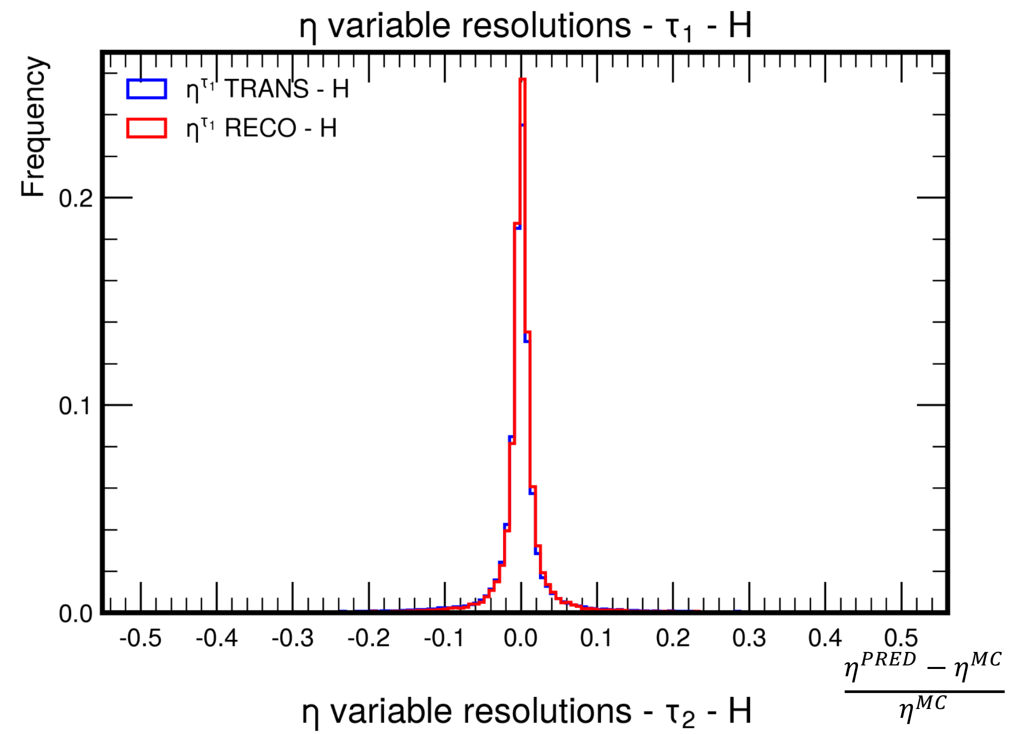
Results
η resolutions

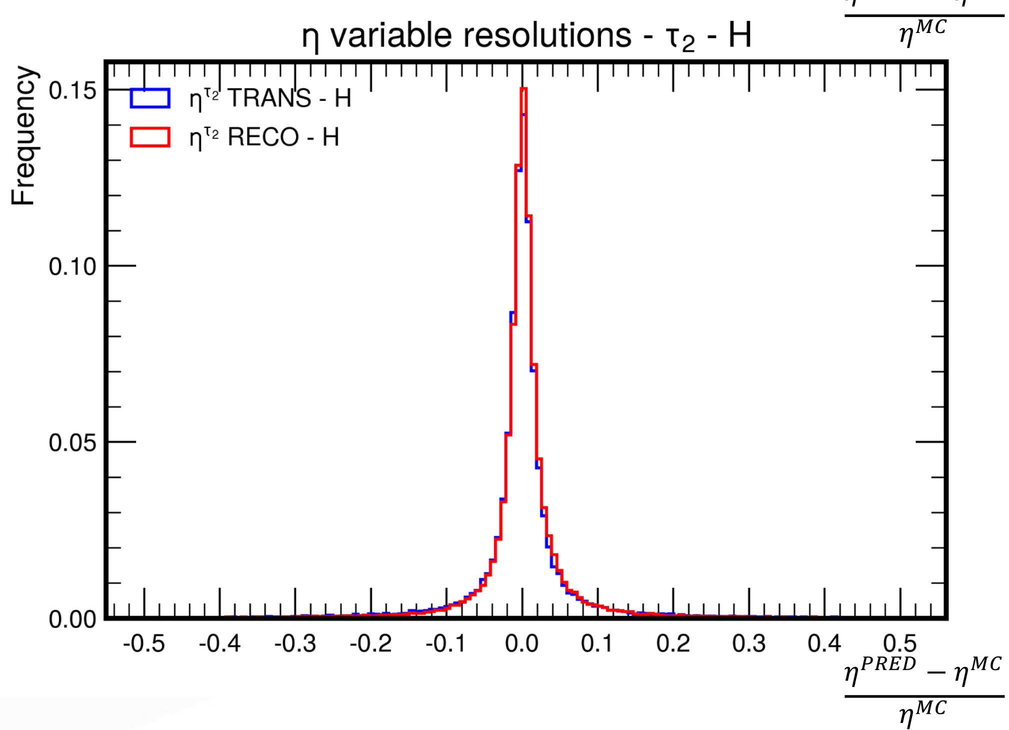
Inference on H test set

τ_1: $MAE = 0.017$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

τ_2: $MAE = 0.023$





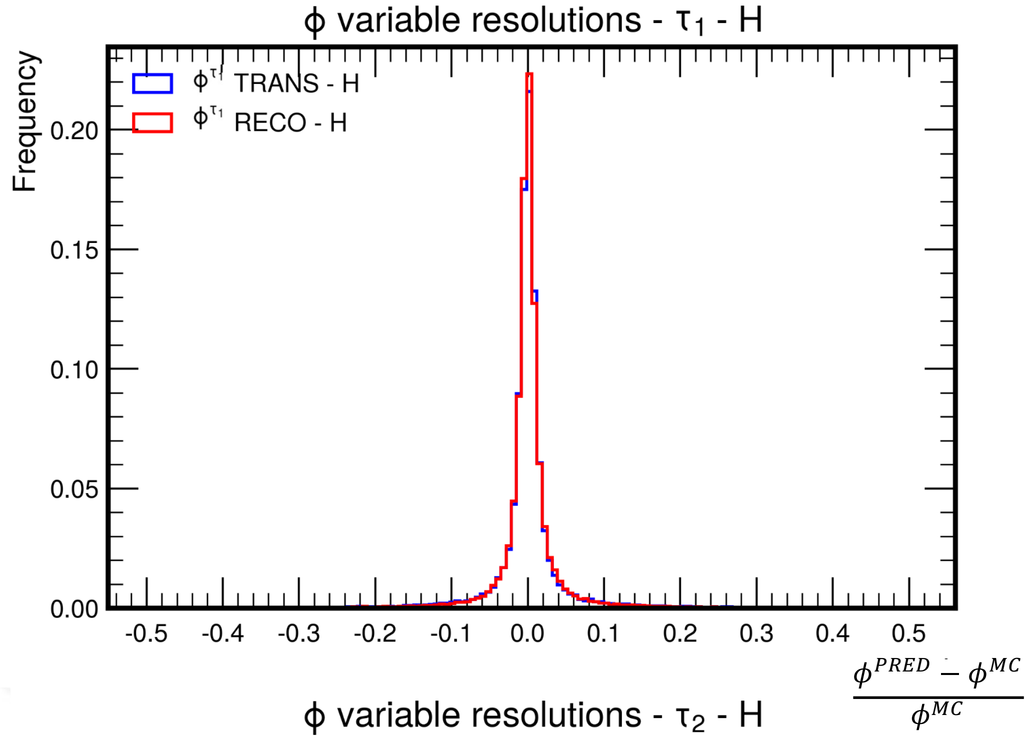
Results
ϕ resolutions

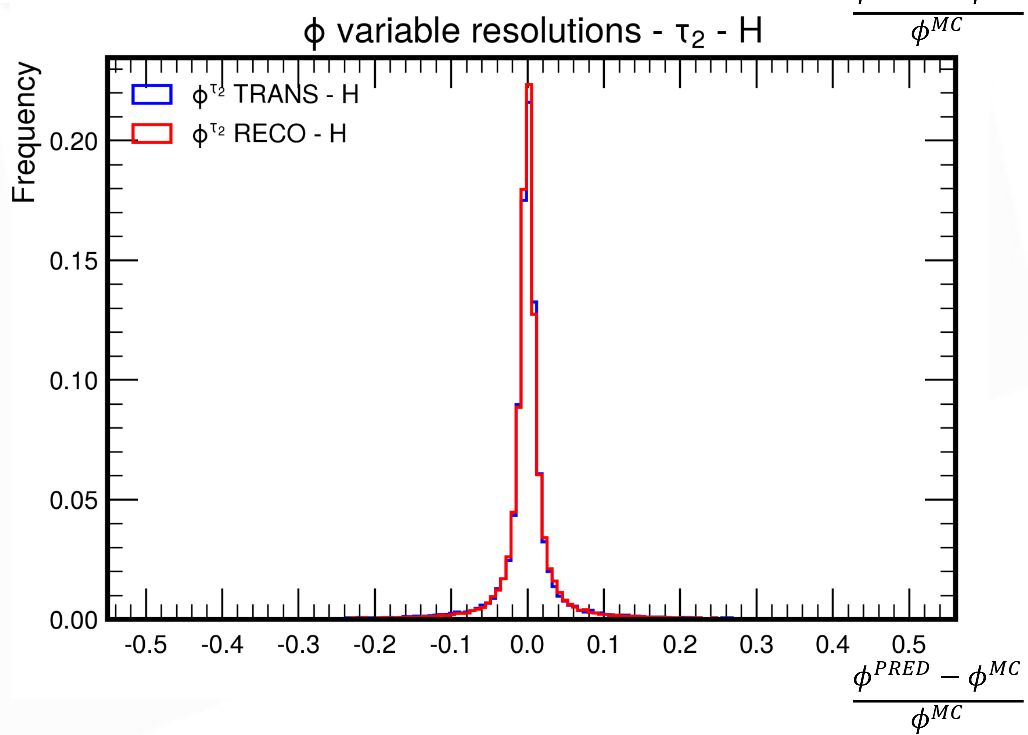
Inference on H test set

τ_1: $MAE = 0.031$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

τ_2: $MAE = 0.047$





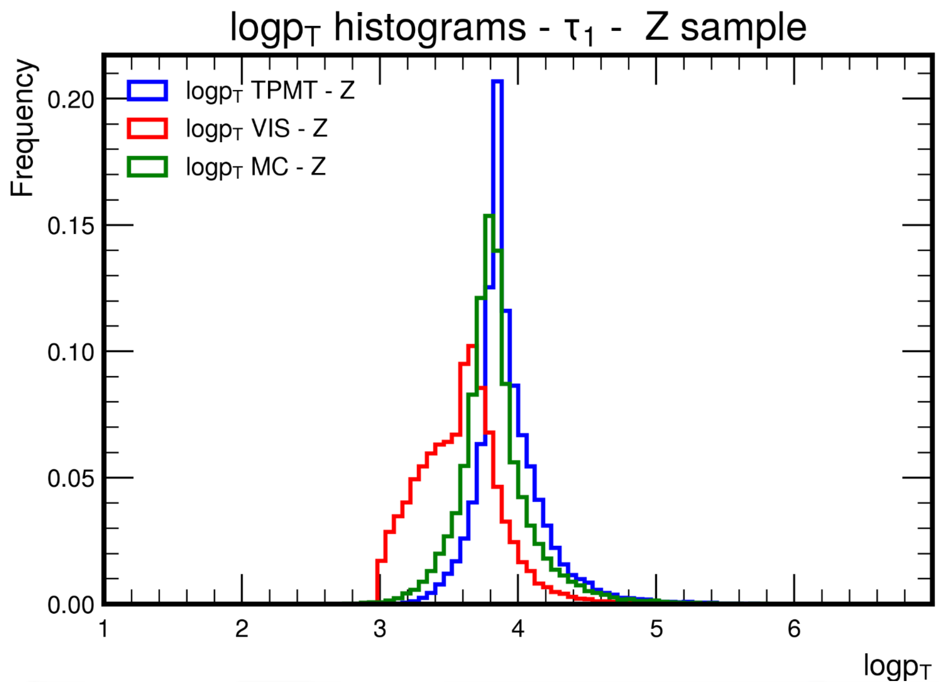
Results
$logp_T$ histograms

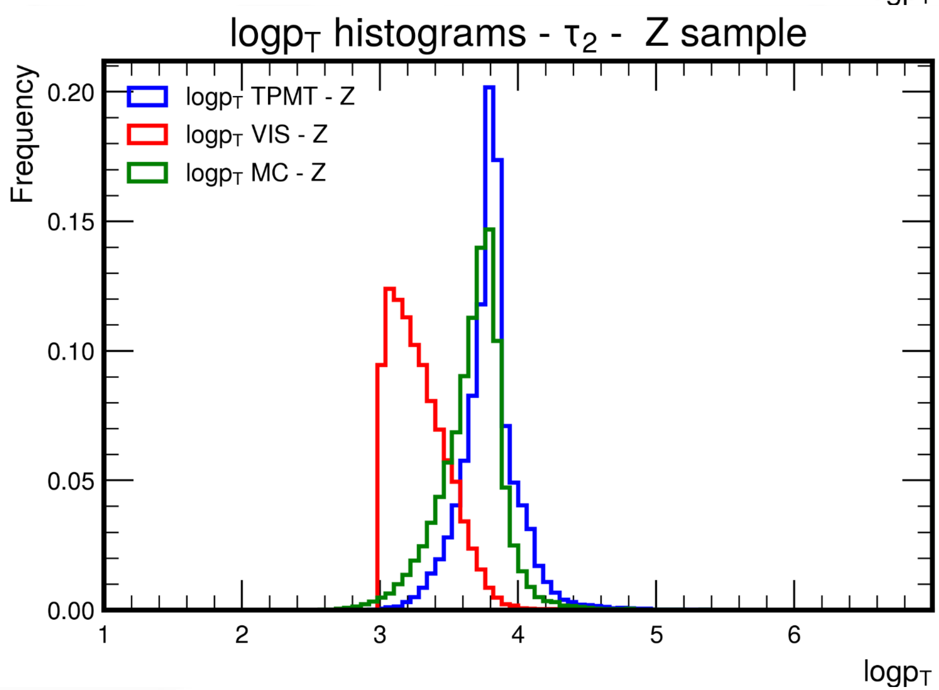
Inference on DY test set

τ_1: $MAE = 0.139$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

τ_2: $MAE = 0.160$





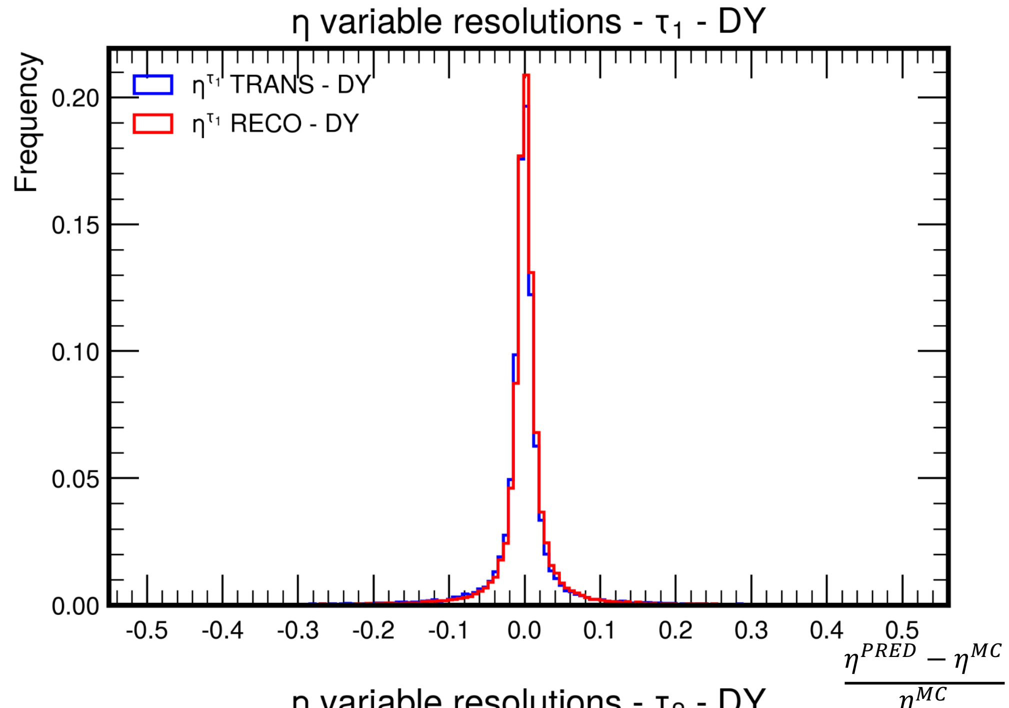
Results
η resolutions

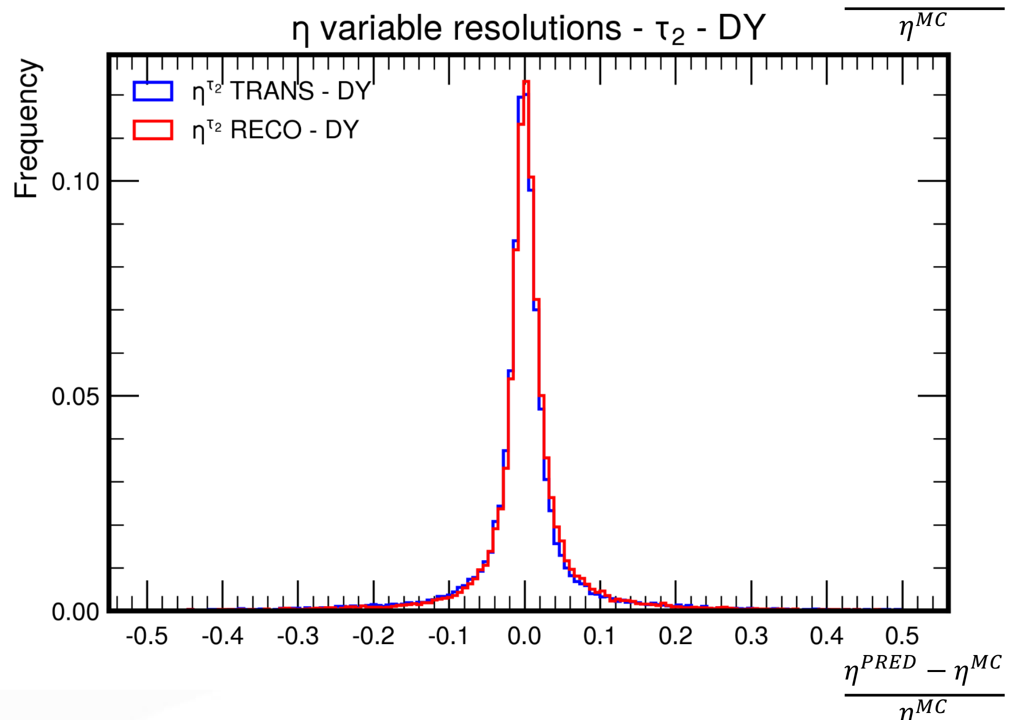
Inference on DY test set

τ_1: $MAE = 0.018$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

τ_2: $MAE = 0.025$





Results
ϕ resolutions

Inference on DY test set

τ_1: $MAE = 0.036$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

τ_2: $MAE = 0.051$