University Milano Bicocca
Physics and Astronomy Doctoral School (39th cycle)

Development of high-throughput machine learning techniques on FPGAs

Valentina Camagni

Tutor: Pietro Govoni[12]
Supervisors: Simone Gennai [2], Federico De Guio [12]

End of the year seminar

September 20, 2024

FPGAs for CMS Level-1 Trigger

Improved triggering with full detector view:
Trigger decision includes calorimeters, muons & tracker (~5us latency)

→ L1Rate = **750 kHz**
→ Latency = **12.5 us latency**
→ Bandwidth: ~ **50 Tb/s**
(1.8 Tb/s in Phase I)

CMS Phase II Level-1 Trigger system intends to perform precise physics selection using a global event reconstruction already at hardware level

FPGAs

- ✓ low-latency processing
- ✓ ability to handle highly parallel tasks
- ✓ reconfigurable nature allows for customization to meet specific requirements
- ✓ superior performance for real-time data processing, with lower power consumption

Deploying ML on FPGAs
→
New trigger algorithms

Challenges

- meet the stringent latency requirements (μs)
- FPGA resources are limited: ML models need to be compressed and optimized through *quantization* and *pruning*
- Model optimization: tools like ***hls4ml***, which facilitate high-level synthesis.

Planned activities

1 Starting from the Master Thesis work implement a DNN for the $\mathrm{di}-\tau$ mass regression to replace SVFit algorithm in all Run III analyses

Particle Transformer for τ lepton pair invariant mass reconstruction for the $HH \rightarrow b\bar{b}\tau^{+}\tau^{-}$ CMS analysis

Tau Pair Mass Transformer
TPMT

- ✓ Tau costituents
- ✗ b-jets information

2 Model distillation optimized for Phase-II implementation on FPGAs. Incorporating invariant mass information could lower the tau trigger threshold, currently set at 40 GeV, thereby recovering the corresponding phase space

3 Level-1 Trigger Scouting on soft taus.
Improvement of the trigger acceptance of tau leptons, specifically extending the coverage towards lower pT

As CERN Doctoral student

PhD courses, Workshops and Schools

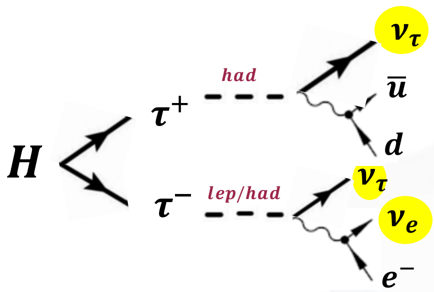
- ✓ Introduction to FPGAs *(November 2023)*
- ✓ ML@L1 Trigger Workshop at CERN *(December 2023)*
- ✓ 6th Inter-experiment Machine Learning Workshop
  + poster presentation *(February 2024)*
- ✓ Mandatory interdisciplinary courses:
  1. Communicating research in the era of social media
  2. Productivity tool for (young) researchers
  3. Surfing the academic job marketing
- ✓ Tutor activity for Laboratory II *(March-June 2024)*
- ✓ AI-INFN 1° User Form (talk) *(June 2024)*
- X Internal courses:
  Deep Learning for Physicists (**to attend**)
  Physics at Colliders (**to attend**)
  Particle Physics II (**ongoing**)
- ~ AI-PHY school *(October 2024)*

Best presentation award - 109th SIF Conference

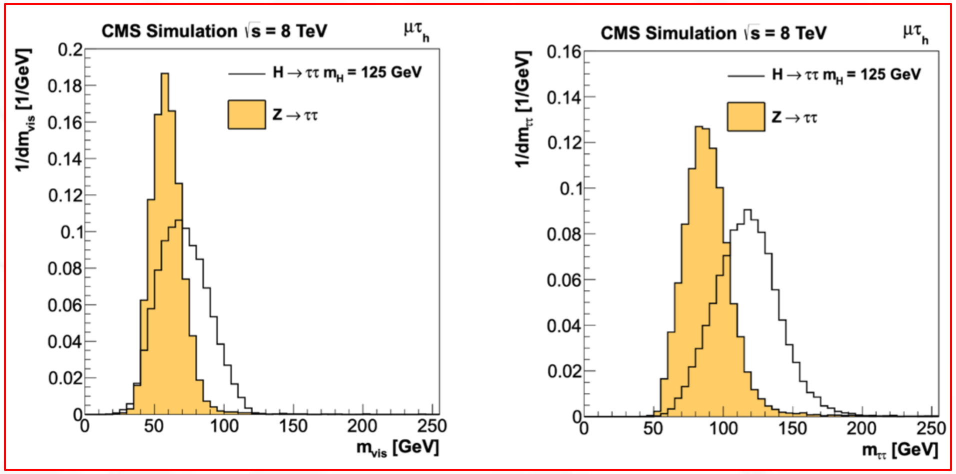
Article publication on ***Nuovo Cimento*** Journal

Open Access

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

- 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

<u>1° GOAL</u>

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

Input features

1 **TauProd**

Taus' decay products

Shape: $(10, 12) = (num\_tauprods\ ,\ num\_features)$

padding if an event has less than 10 tau products

$logp_t$
η
ϕ
m
$log\left(\frac{p_T}{p_{T(\tau)}}\right)$
$charge$
$tauIdx$

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

Categorical variables from particle ID

2 **Tau**

Shape: $(6, 3) = (num\_particles\ ,\ num\_features)$

τ_1	$logp_T$
τ_2	η
MET	ϕ
jet_1	
jet_2	
jet_3	

GenPart

Not the full 4-momentum since eta and phi does not change

Shape: $(2, 1) = (num\_particles\ ,\ num\_features)$

τ_1	$logp_T$
τ_2	

+ di-τ invariant mass at generator level $m_{\tau\tau}^{GEN}$

Pre-processing steps

Data sets **GluGluHToTauTau_M125** **DYJetsToLL_M-50-madgraphMLM**

TAU SELECTION

At least 2 taus

- Gen matched
- Hadronic decay
- $p_T \geq 20$ GeV

JETS SELECTION

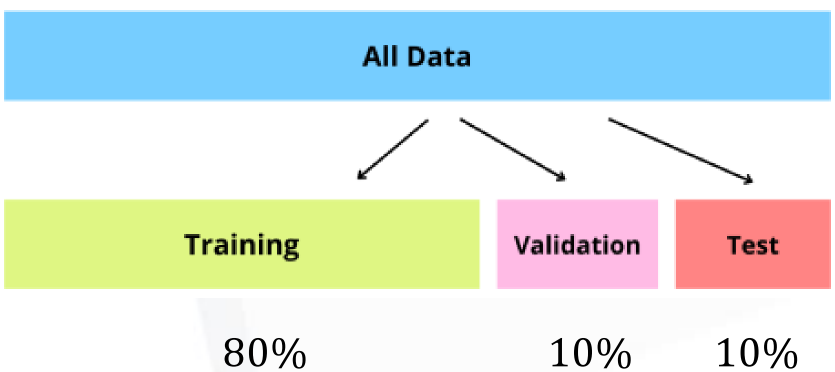
First 3 leading jets with $\Delta R(jet, tau) > 0.4$

(minimum p_T: $10\ \mathrm{GeV}$)

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



Model Architecture

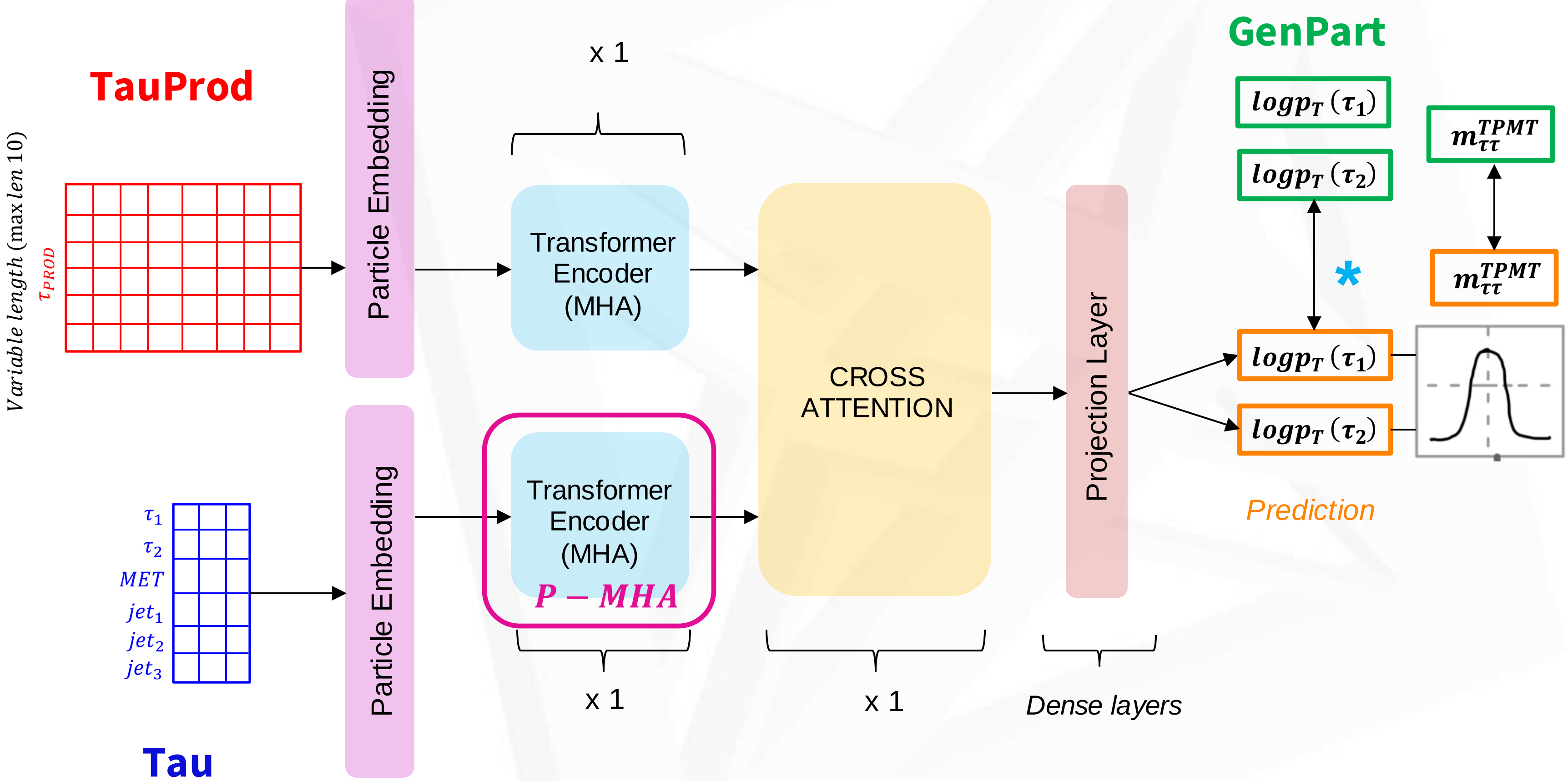
PyTorch

Loss function

① $Mean\ between\ MAE_{logp_T}\ for\ the\ two\ taus$

② $MAE\ between\ m_{\tau\tau}^{TRANS}\ and\ m_{\tau\tau}^{MC}$ $(7\%\ of\ the\ total\ loss)$

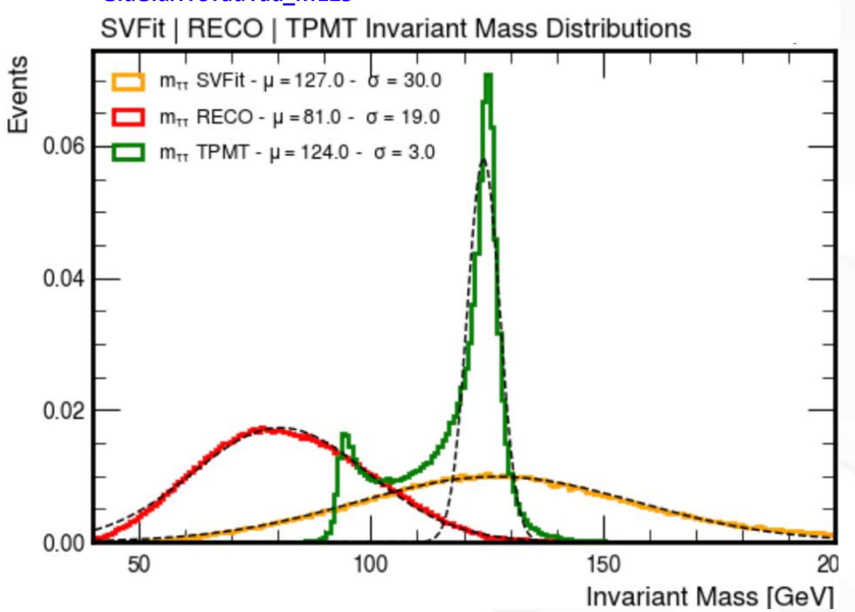
$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$



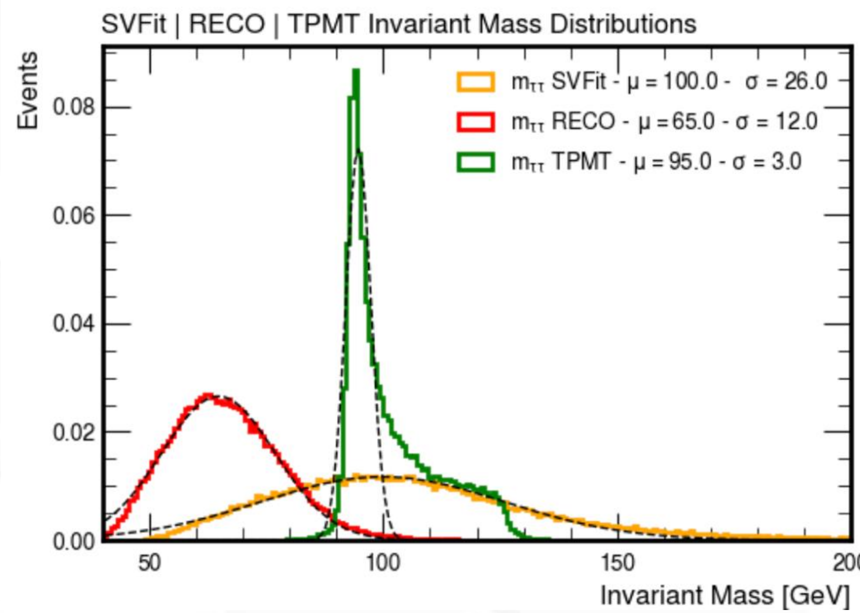
Training time: $\sim 1.5\ min\ per\ epoch$
Inference time: $\sim 2 \times 10^{-3}\ s\ per\ event$
Number of parameters: $\sim 0.5\ M$

$m_{\tau\tau}$ results

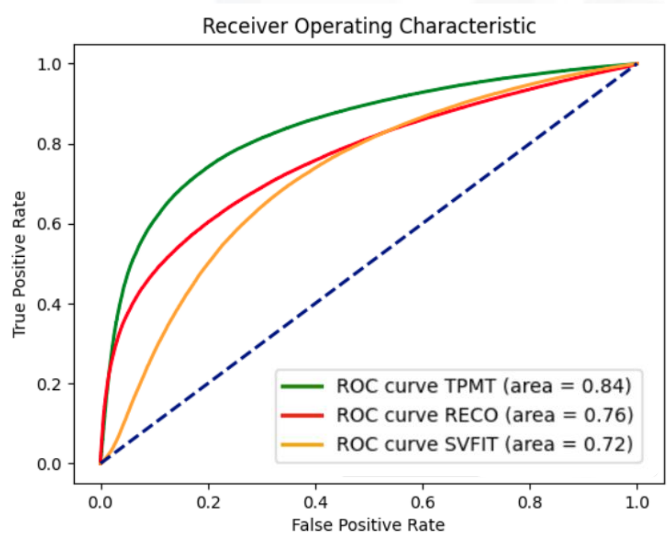
GluGluHToTauTau_M125



DYJetsToLL_M-50-madgraphMLM



AUC as evaluation metric



Train ratio

$H : Z = 2 : 1$

Best training:

$AUC\ score$ of 0.84

Preliminary considerations

- ✓ AUC suggests that TPMT has a better separation capability
- ✗ The wrong peak is slightly higher for H than for DY (due to the different response)
- ✓ Training time: $1.5\ min\ per\ epoch\ (\sim 80\ epochs)$
  Inference time: $2 \cdot 10^{-3}\ s$
- ✗ Inference on any other resonance would have worked worse
  (if not added in the train set composition)

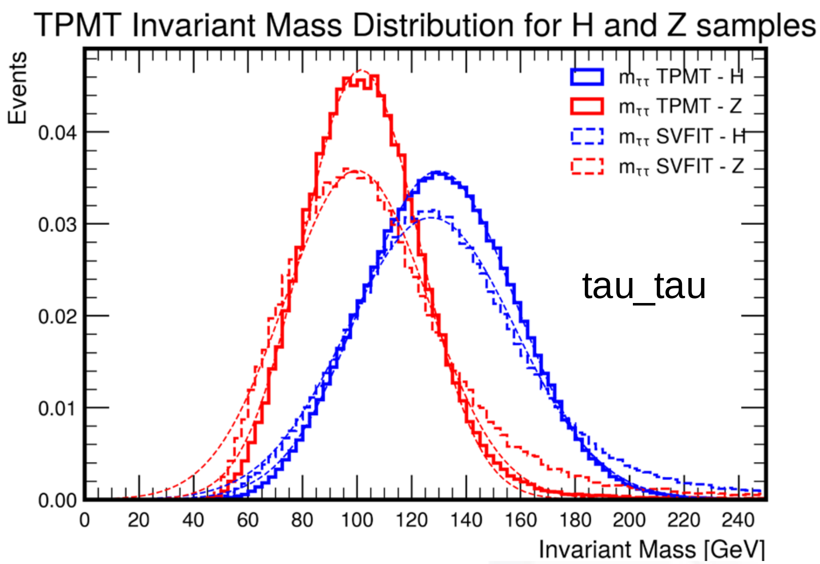
⇩

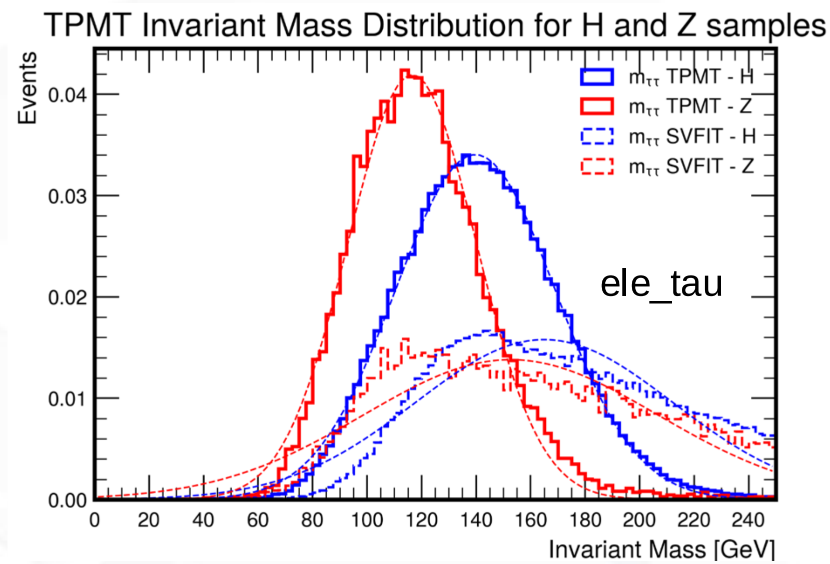
Training on flat mass samples
GluGlutoXto2Tau_M-30to300
VBFtoXto2Tau_M-30to300
and inference on H and Z samples

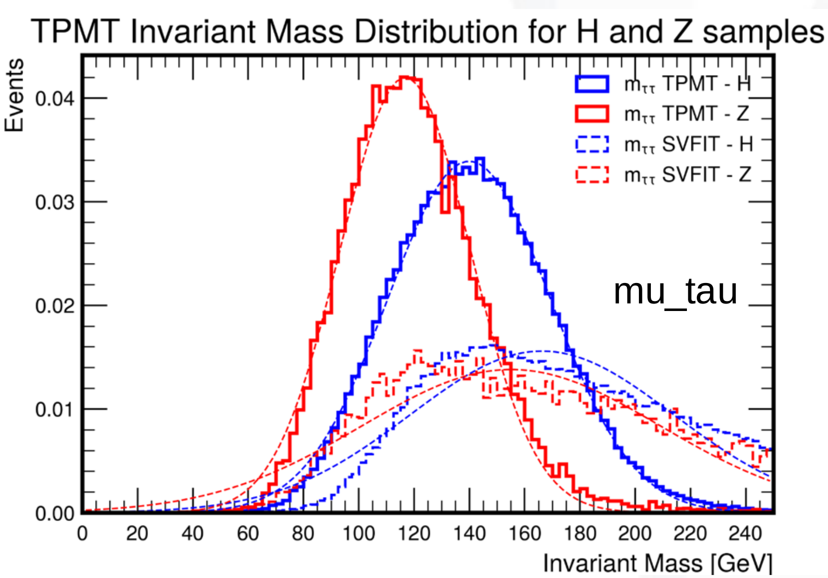
! No more jet information

! Leptonic decaying taus in addition to hadronic ones $(\tau_h + l\ (e, \mu))$

Overall training on ggF sample – tau_tau, ele_tau, mu_tau



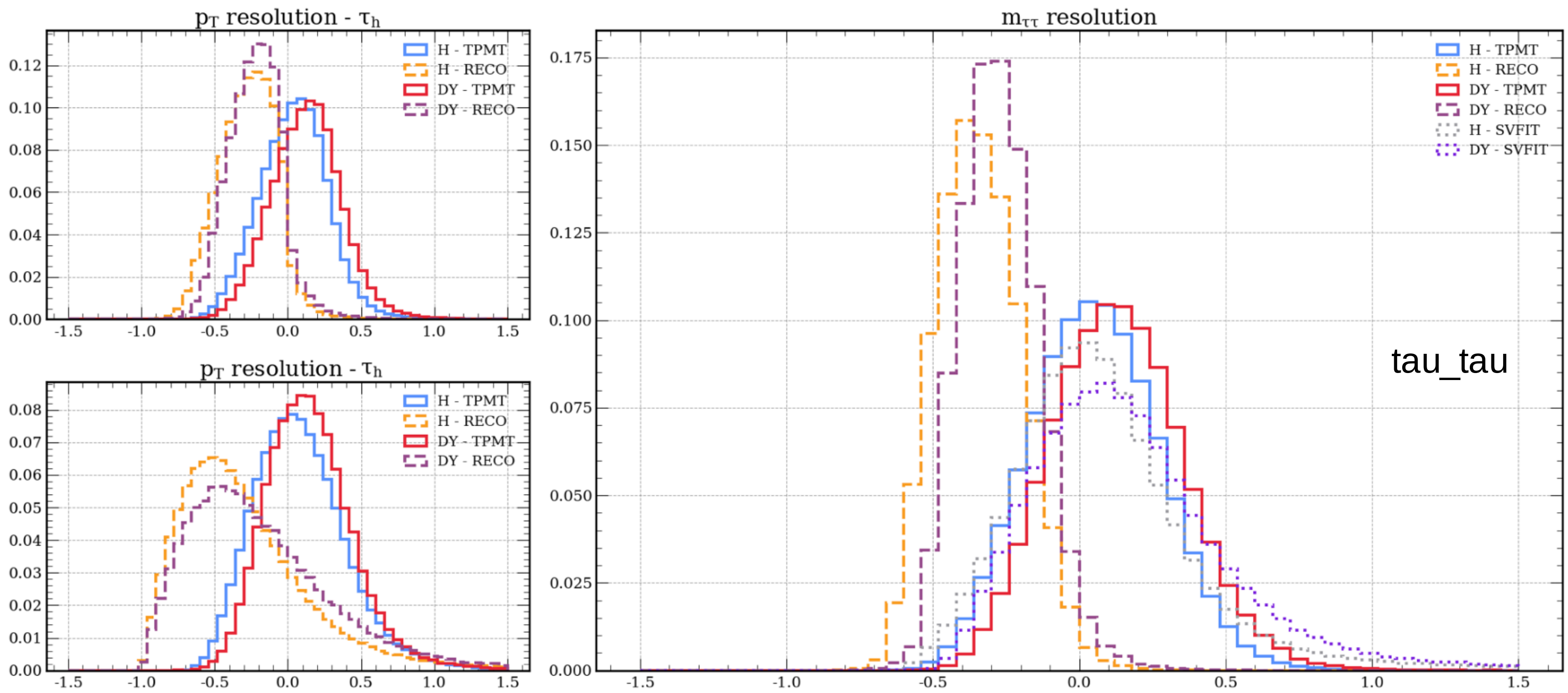




Fit Type		Mean	Std
$m_{\tau\tau}^{SVFit}$ – H	tau_tau	127.26	30.79
$m_{\tau\tau}^{TPMT}$ – H	tau_tau	129.81	28.08
$m_{\tau\tau}^{SVFit}$ – DY	tau_tau	99.83	26.28
$m_{\tau\tau}^{TPMT}$ – DY	tau_tau	101.47	21.18
$m_{\tau\tau}^{SVFit}$ – H	ele_tau	164.98	47.23
$m_{\tau\tau}^{TPMT}$ – H	ele_tau	139.72	30.22
$m_{\tau\tau}^{SVFit}$ – DY	ele_tau	152.69	53.85
$m_{\tau\tau}^{TPMT}$ – DY	ele_tau	116.53	23.47
$m_{\tau\tau}^{SVFit}$ – H	mu_tau	166.28	47.63
$m_{\tau\tau}^{TPMT}$ – H	mu_tau	139.91	29.51
$m_{\tau\tau}^{SVFit}$ – DY	mu_tau	155.12	53.94
$m_{\tau\tau}^{TPMT}$ – DY	mu_tau	116.81	23.50

$p_T^{\tau 1}$, $p_T^{\tau 2}$, $m_{\tau\tau}$ resolution results for tau_tau pairType

H & ***Z***



RECO

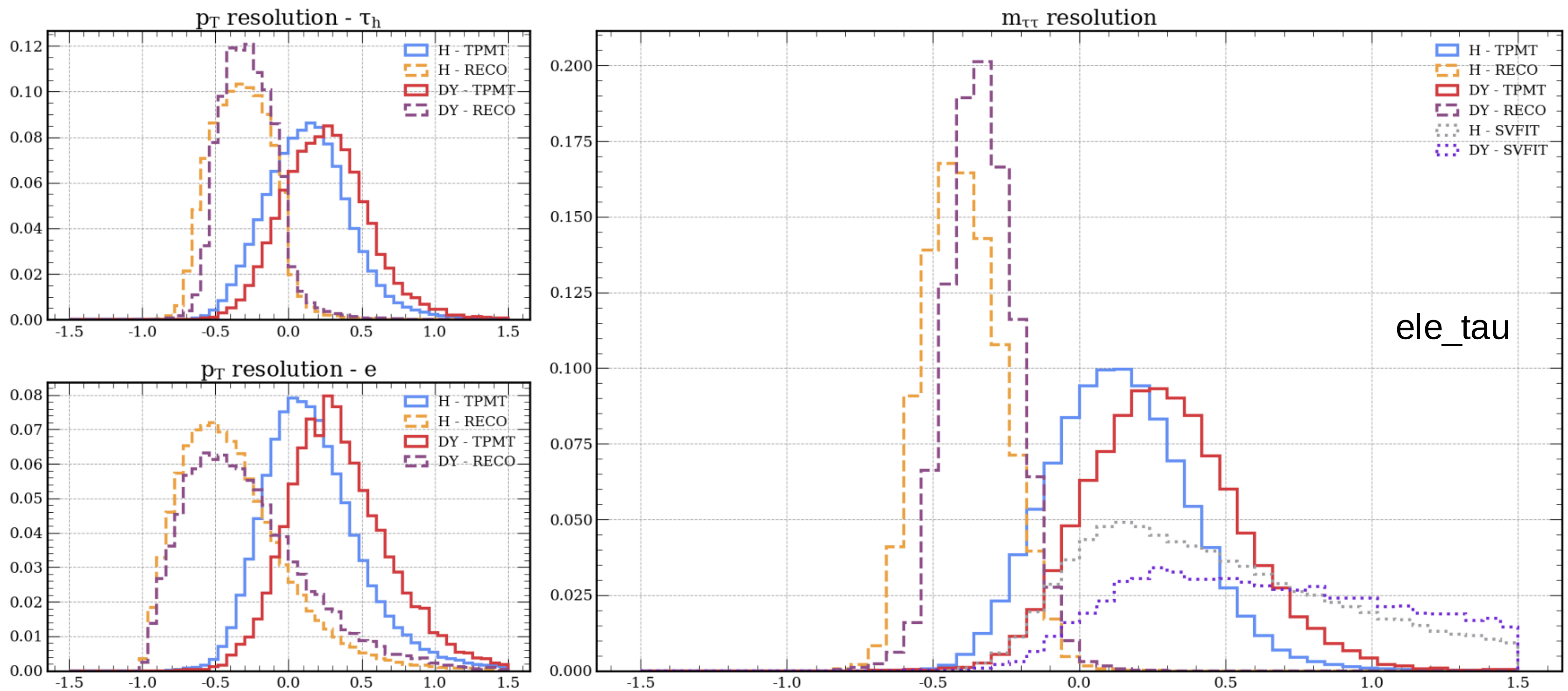
Fit Type	Mean	Std
p_T τ_1 - H	-0.26	0.2
p_T τ_2 - H	-0.42	0.35
$m_{\tau\tau}$ - H	-0.35	0.15
p_T τ_1 - DY	-0.22	0.18
p_T τ_2 - DY	-0.33	0.41
$m_{\tau\tau}$ - DY	-0.28	0.13

TPMT

Fit Type	Mean	Std
p_T τ_1 - H	0.04	0.23
p_T τ_2 - H	0.04	0.35
$m_{\tau\tau}$ - H	0.04	0.3
p_T τ_1 - DY	0.13	0.23
p_T τ_2 - DY	0.12	0.28
$m_{\tau\tau}$ - DY	0.12	0.23

$p_T^{\tau 1}$, $p_T^{\tau 2}$, $m_{\tau\tau}$ resolution results for ele_tau pairType

H & ***Z***



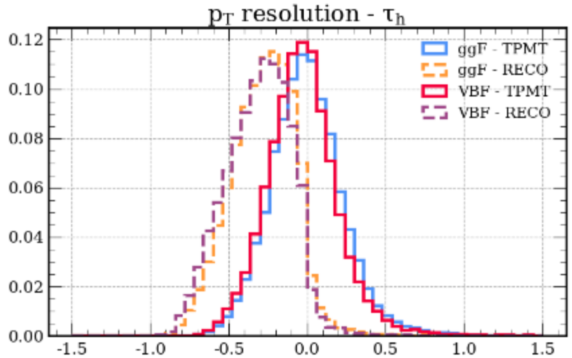
RECO

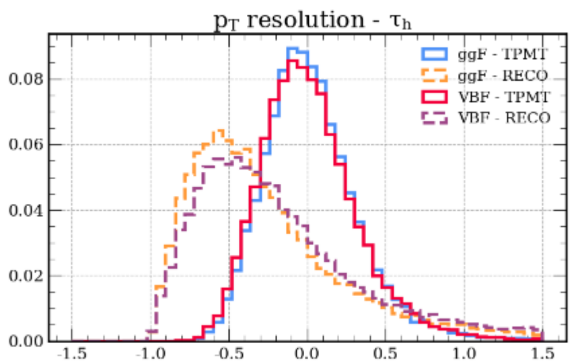
Fit Type	Mean	Std
$p_T\ \tau_1$ - H	-0.26	0.2
$p_T\ \tau_2$ - H	-0.42	0.35
$m_{\tau\tau}$ - H	-0.35	0.15
$p_T\ \tau_1$ - DY	-0.22	0.18
$p_T\ \tau_2$ - DY	-0.33	0.41
$m_{\tau\tau}$ - DY	-0.28	0.13

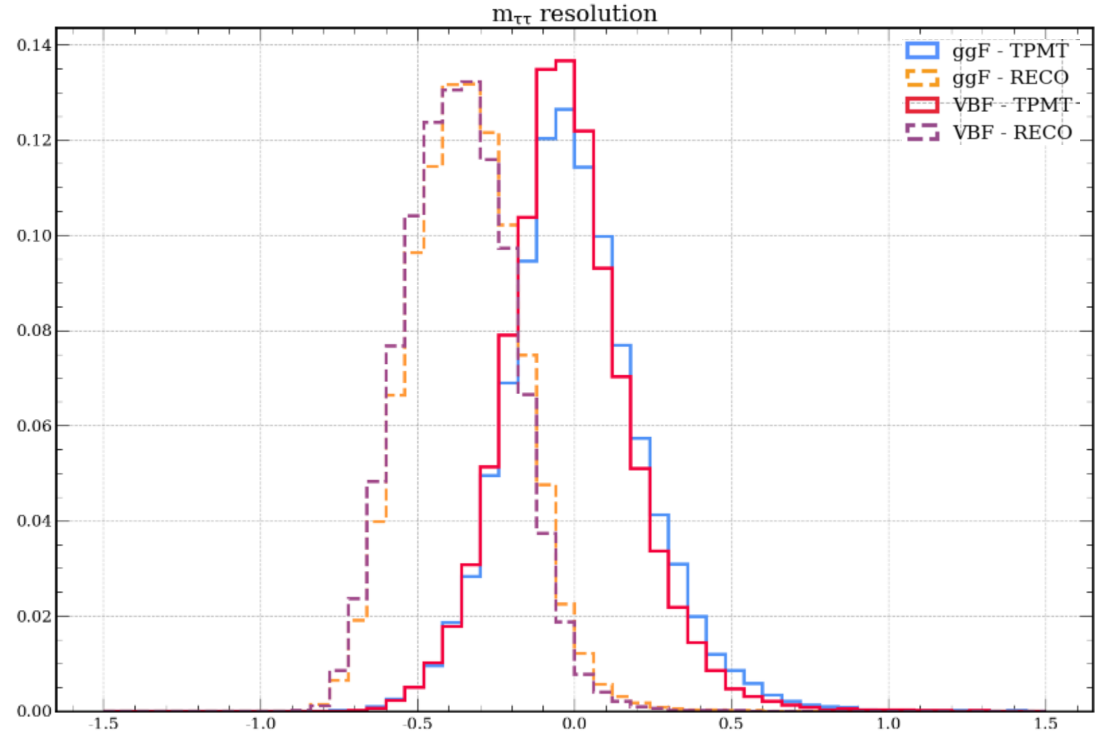
TPMT

Fit Type	Mean	Std
$p_T\ \tau_1$ - H	0.04	0.23
$p_T\ \tau_2$ - H	0.04	0.35
$m_{\tau\tau}$ - H	0.04	0.3
$p_T\ \tau_1$ - DY	0.13	0.23
$p_T\ \tau_2$ - DY	0.12	0.28
$m_{\tau\tau}$ - DY	0.12	0.23

$p_T^{\tau 1}, p_T^{\tau 2}, m_{\tau\tau}$ resolution results for tau_tau pairType *ggF & VBF*







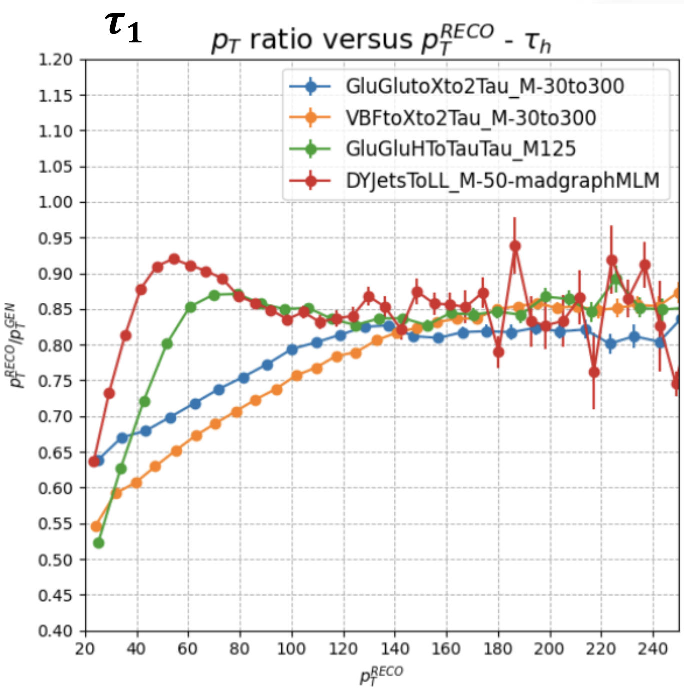
RECO

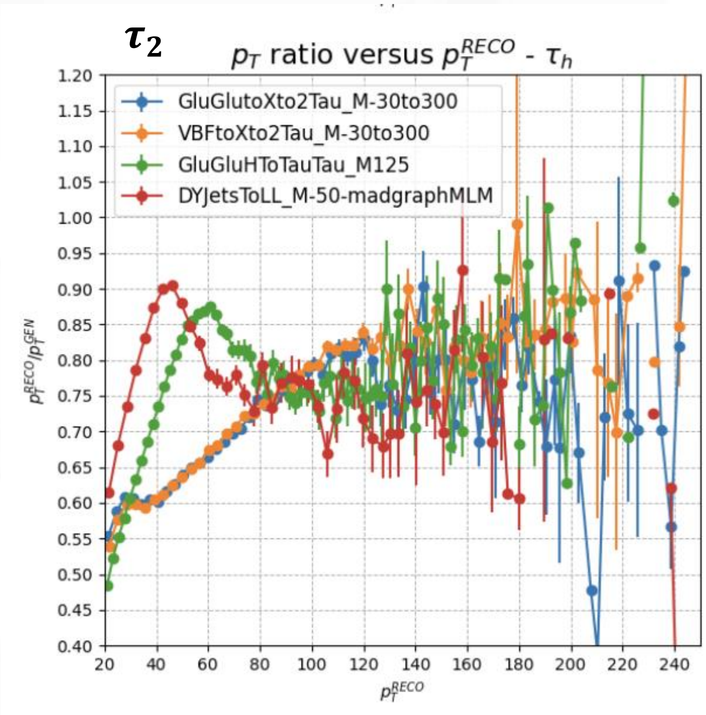
Fit Type	Mean	Std
p_T τ_1 - ggF	-0.26	0.2
p_T τ_2 - ggF	-0.44	0.35
$m_{\tau\tau}$ - ggF	-0.35	0.18
p_T τ_1 - VBF	-0.29	0.21
p_T τ_2 - VBF	-0.37	0.39
$m_{\tau\tau}$ - VBF	-0.37	0.18

TPMT

Fit Type	Mean	Std
p_T τ_1 - ggF	-0.02	0.21
p_T τ_2 - ggF	-0.02	0.26
$m_{\tau\tau}$ - ggF	-0.02	0.19
p_T τ_1 - VBF	-0.04	0.2
p_T τ_2 - VBF	-0.03	0.28
$m_{\tau\tau}$ - VBF	-0.04	0.18

p_T ratio versus p_T^{RECO} for tau_tau pairType





- **Different response between resonances and flat mass samples**
- **More differences between H and Z compared to ggF and VBF responses**

Due to convolution of tau resolution and p_T^{GEN} distribution

Studying new training strategies

Conclusions

- **Training on H and DY**
  - TPMT behavies as a classifier
  - Good mass resolution but strong dependent on the training samples

- **Training on ggF sample**
  - Resolution and fits much worst, still better than SVFit but suboptimal

For optimal training, it is essential to include samples that reflect the true underlying distributions of the events whose mass we aim to estimate, rather than using flat distributions that can lead to suboptimal performance

Future plans

- **Add a loss term regarding MET** $\mathcal{L}_{MET} = |MET_{observed} - (p_T^{neutrinos})|$
- **Train TPMT with the TauProd matrix divided by the two taus**

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

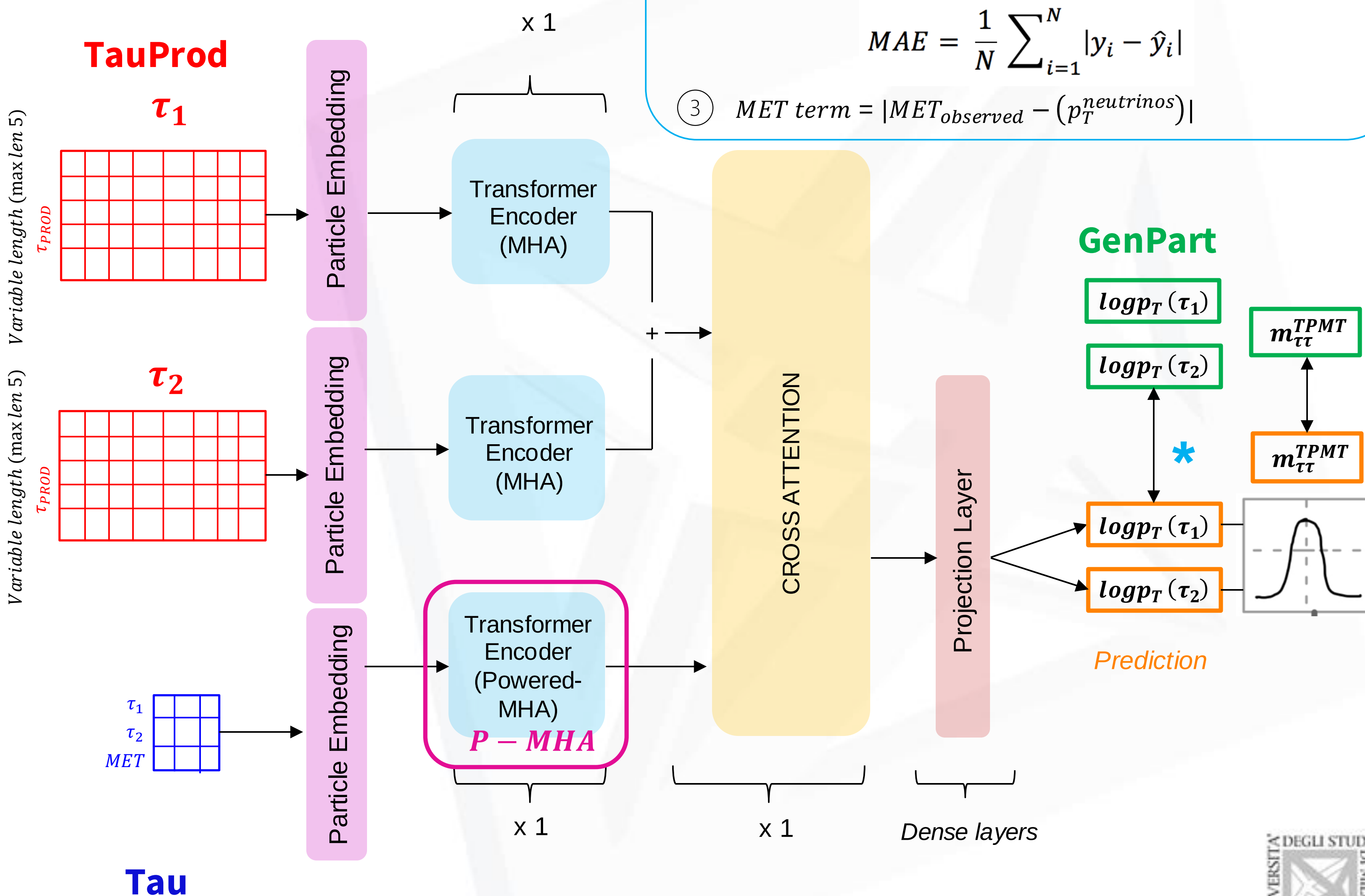
New Model Architecture

Loss function

1. $Mean\ between\ MAE_{logp_T}\ for\ the\ two\ taus$
2. $MAE\ between\ m_{\tau\tau}^{TRANS}\ and\ m_{\tau\tau}^{MC}$ $(7\%\ of\ the\ total\ loss)$

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

3. $MET\ term = |MET_{observed} - (p_T^{neutrinos})|$

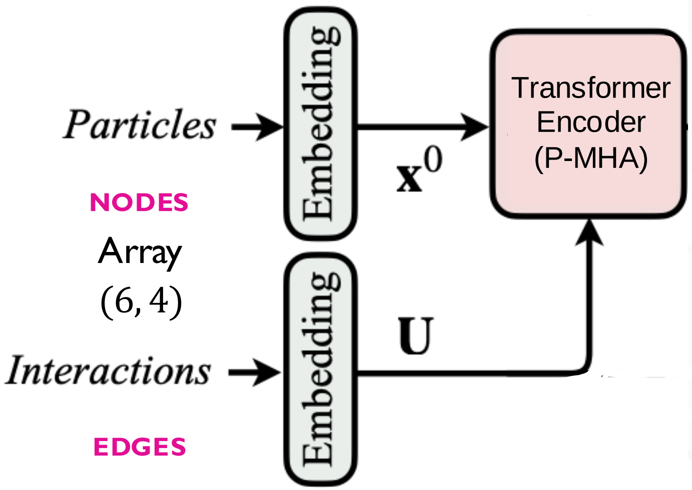


Number of parameters: $\sim 0.9\ M$

Thank you for your attention!

BACKUP

Powered MHA

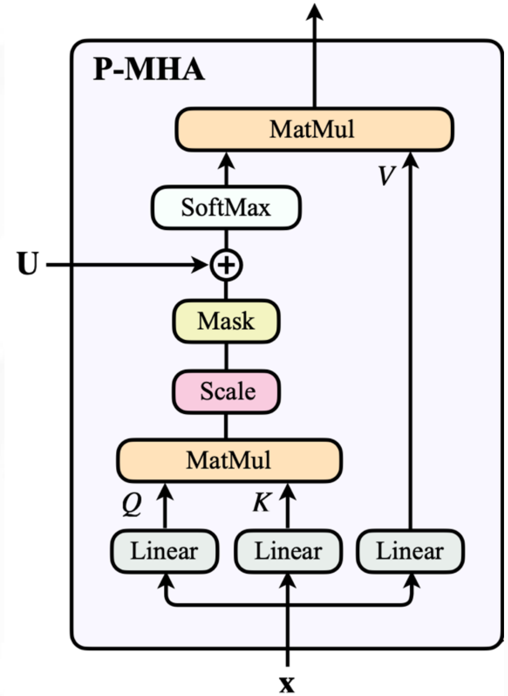


NODES

Array $(6, 4)$

EDGES

Array $(6, 6, 4)$



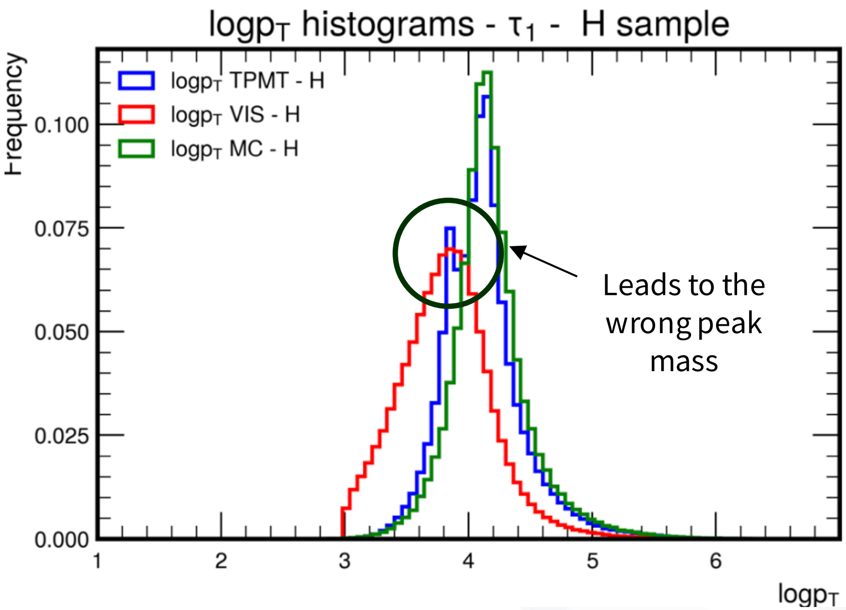
4 pairwise features $\boxed{(\Delta, k_T, z, m^2)}$ from [Particle Transformer for jet tagging]

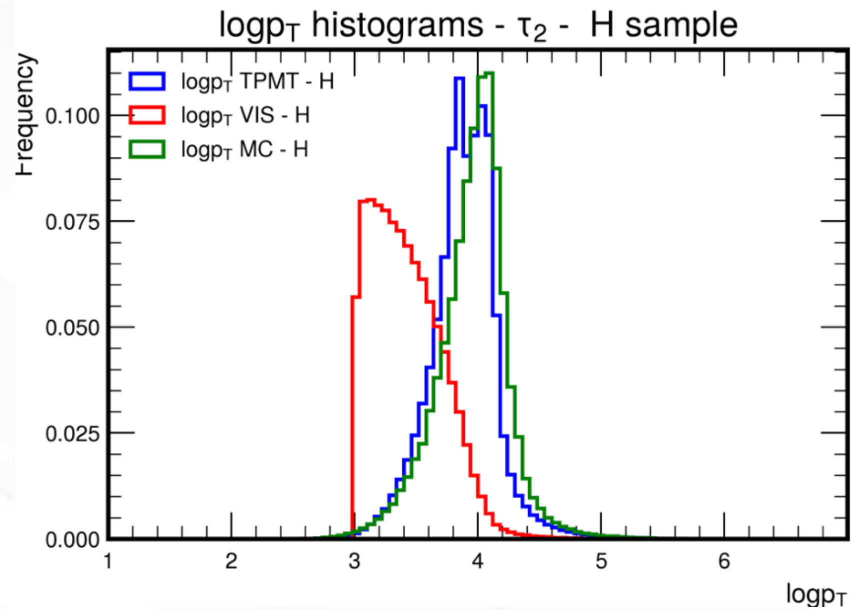
$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2},$$
$$k_\mathrm{T} = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})\Delta,$$
$$z = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})/(p_{\mathrm{T},a} + p_{\mathrm{T},b}),$$
$$m^2 = (E_a + E_b)^2 - \|\mathbf{p}_a + \mathbf{p}_b\|^2,$$

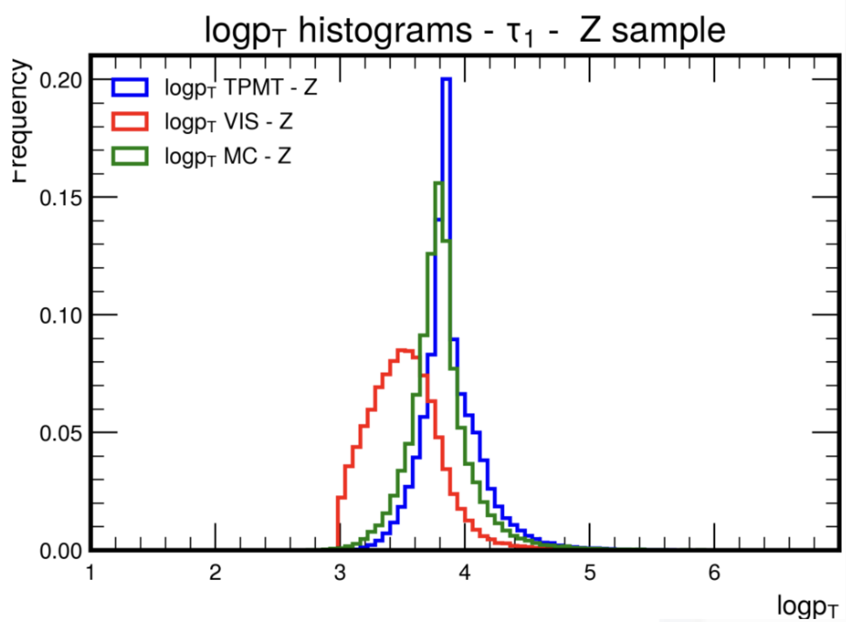
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

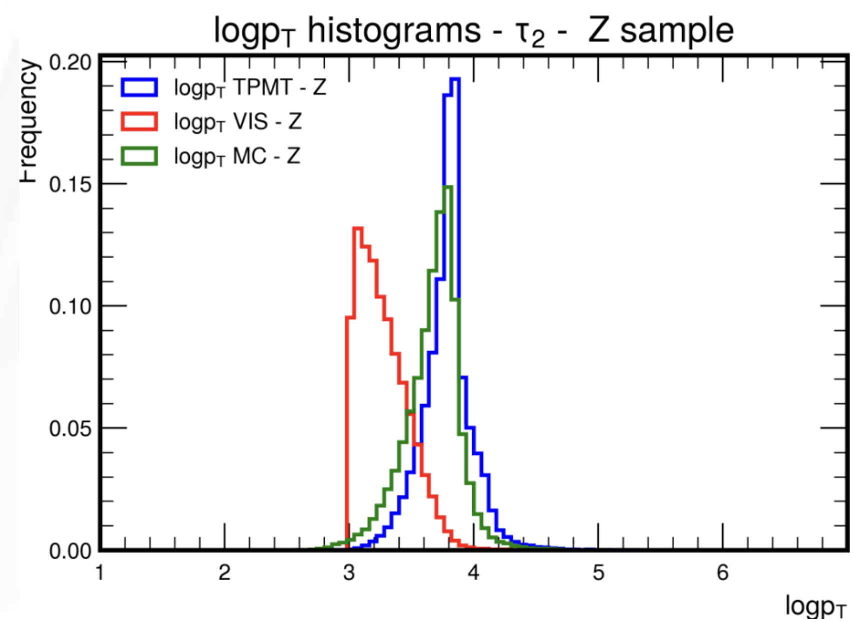
$logp_T^{\tau_1}, logp_T^{\tau_2}$ results

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

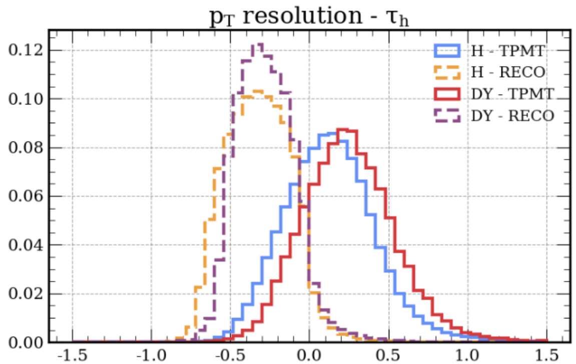


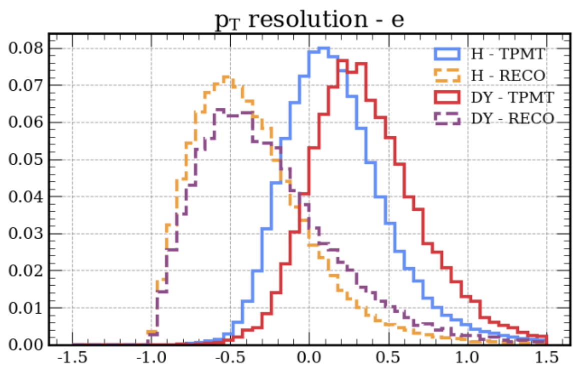


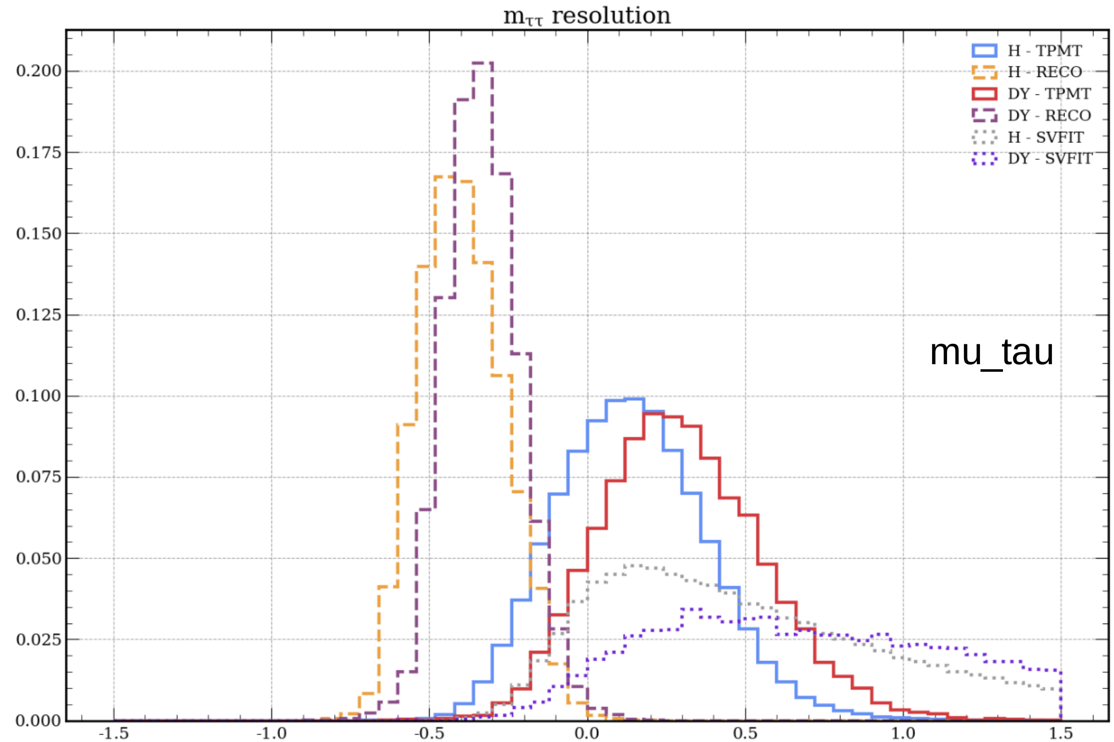




$p_T^{\tau 1}$, $p_T^{\tau 2}$, $m_{\tau\tau}$ resolution results for mu_tau pairType







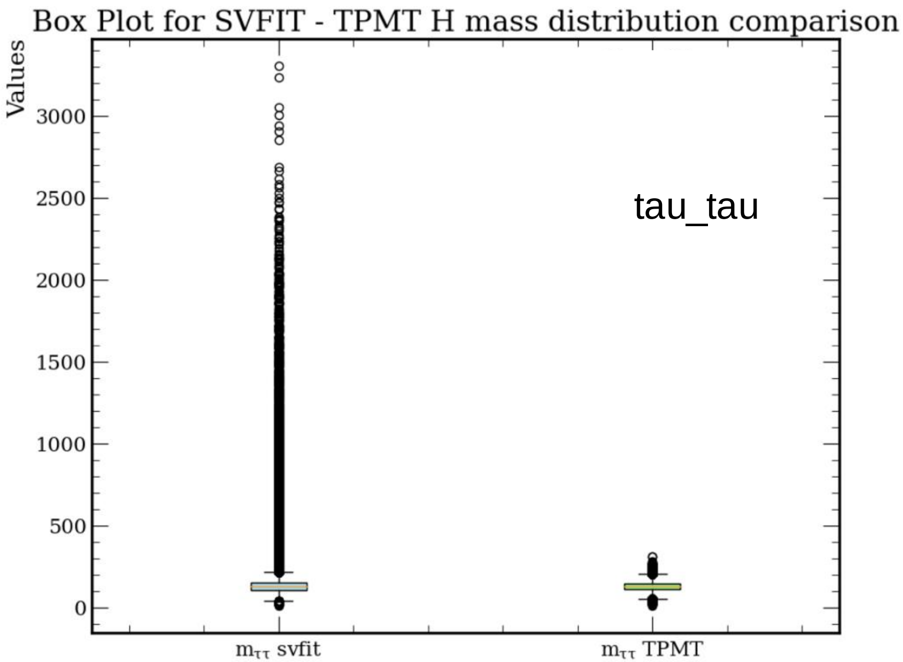
RECO

Fit Type	Mean	Std
p_T τ_1 - H	-0.26	0.2
p_T τ_2 - H	-0.42	0.35
$m_{\tau\tau}$ - H	-0.35	0.15
p_T τ_1 - DY	-0.22	0.18
p_T τ_2 - DY	-0.33	0.41
$m_{\tau\tau}$ - DY	-0.28	0.13

TPMT

Fit Type	Mean	Std
p_T τ_1 - H	0.04	0.23
p_T τ_2 - H	0.04	0.35
$m_{\tau\tau}$ - H	0.04	0.3
p_T τ_1 - DY	0.13	0.23
p_T τ_2 - DY	0.12	0.28
$m_{\tau\tau}$ - DY	0.12	0.23

$m_{\tau\tau}^{H}, m_{\tau\tau}^{Z}$ quartiles



Distribution	Q1	Q2	Q3
SVFIT - H	109.19	130.00	153.64
TPMT - H	111.37	130.18	149.17
SVFIT - DY	85.31	103.24	124.69
TPMT - DY	88.09	102.31	117.06

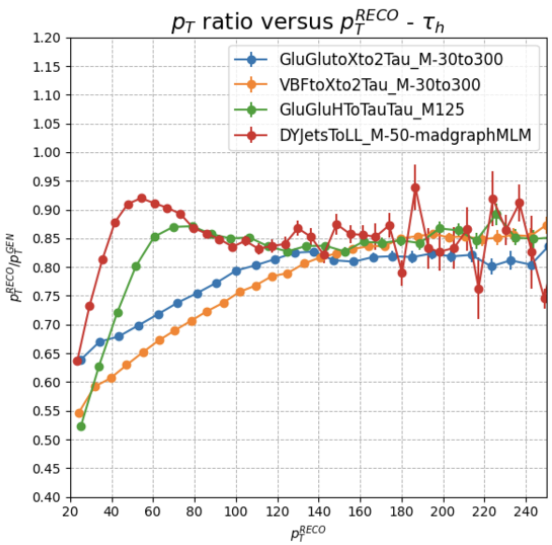
ele_tau

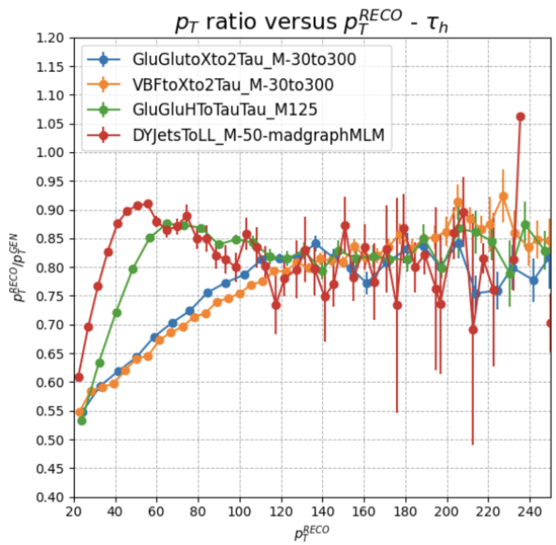
Distribution	Q1	Q2	Q3
SVFIT - H	147.15	191.63	270.05
TPMT - H	120.89	140.39	160.59
SVFIT - DY	132.44	184.51	279.27
TPMT - DY	101.99	117.80	134.57

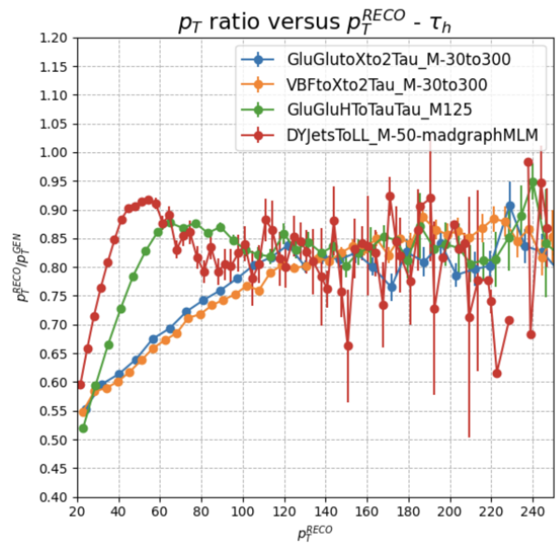
mu_tau

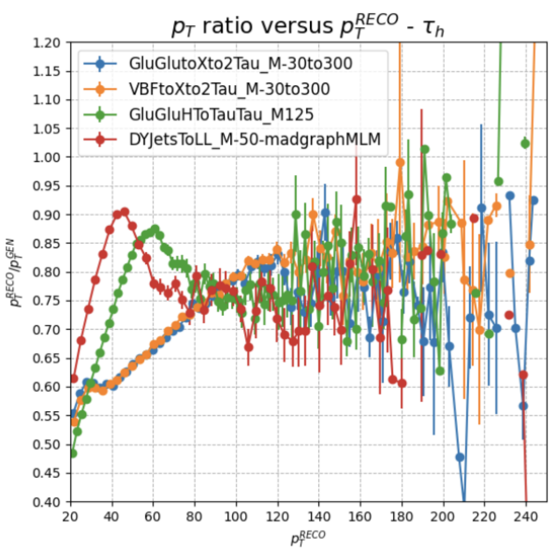
Distribution	Q1	Q2	Q3
SVFIT - H	148.48	193.65	272.63
TPMT - H	120.80	140.67	160.89
SVFIT - DY	134.61	186.57	279.32
TPMT - DY	102.45	117.99	134.84

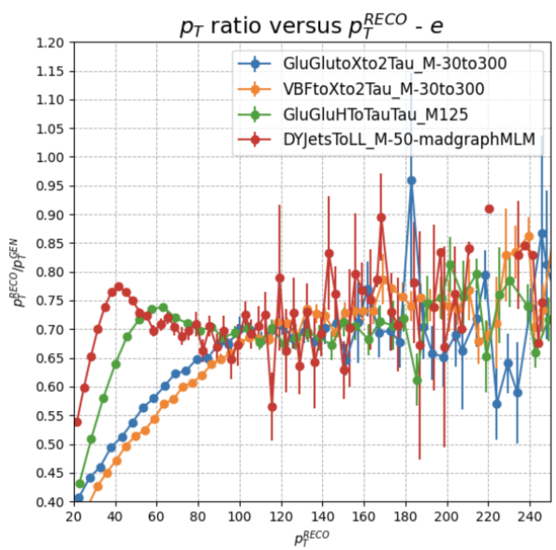
p_T ratio versus p_T^{RECO} for all pairTypes

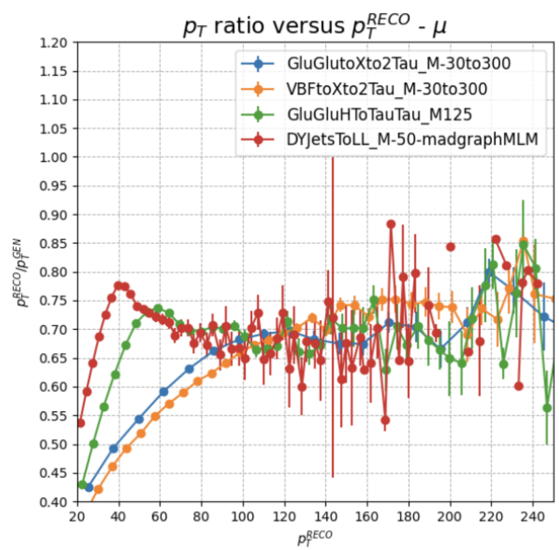




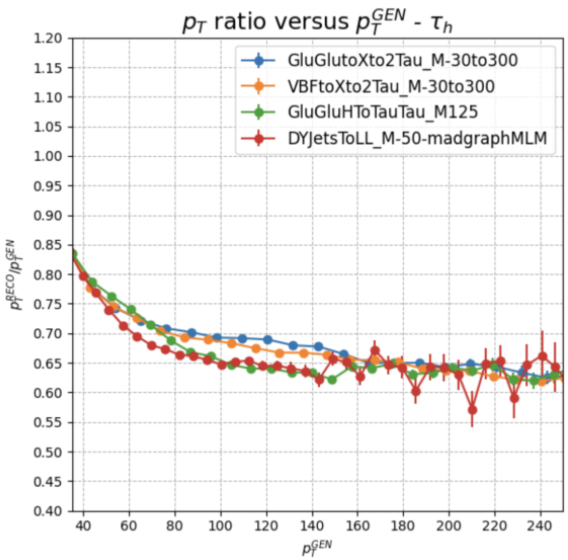


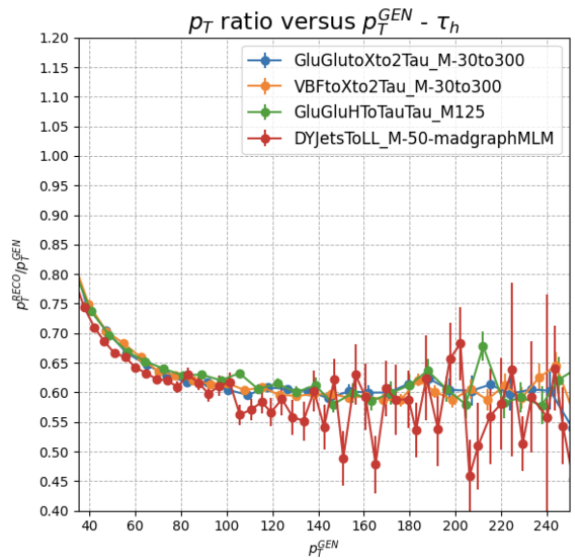


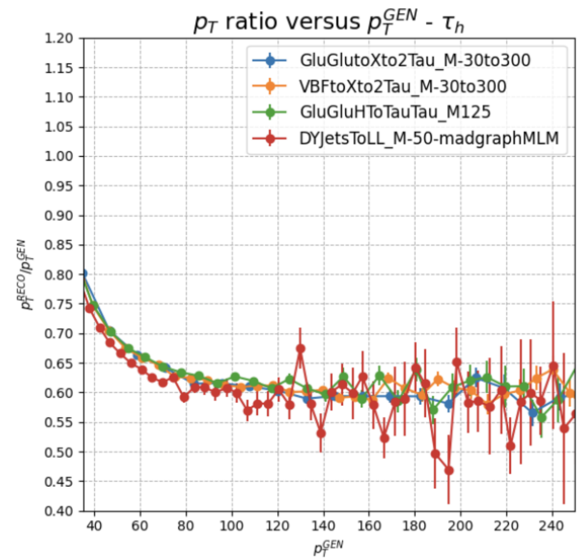


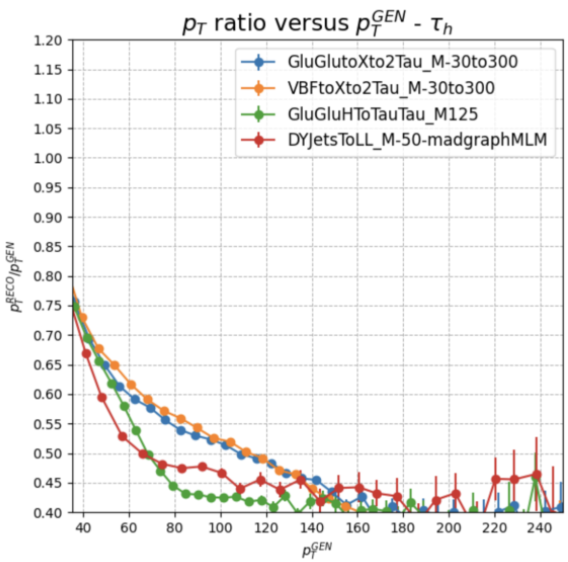


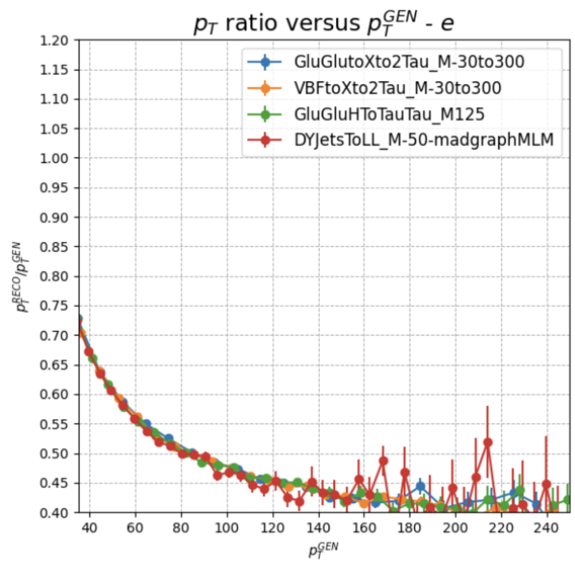
p_T ratio versus p_T^{RECO} for all pairTypes

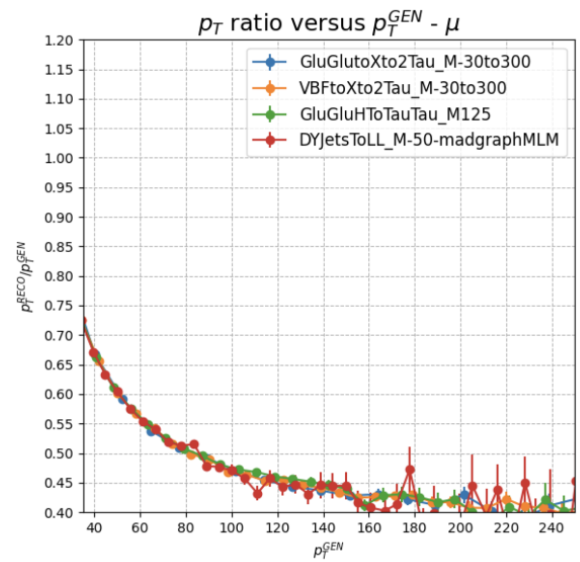


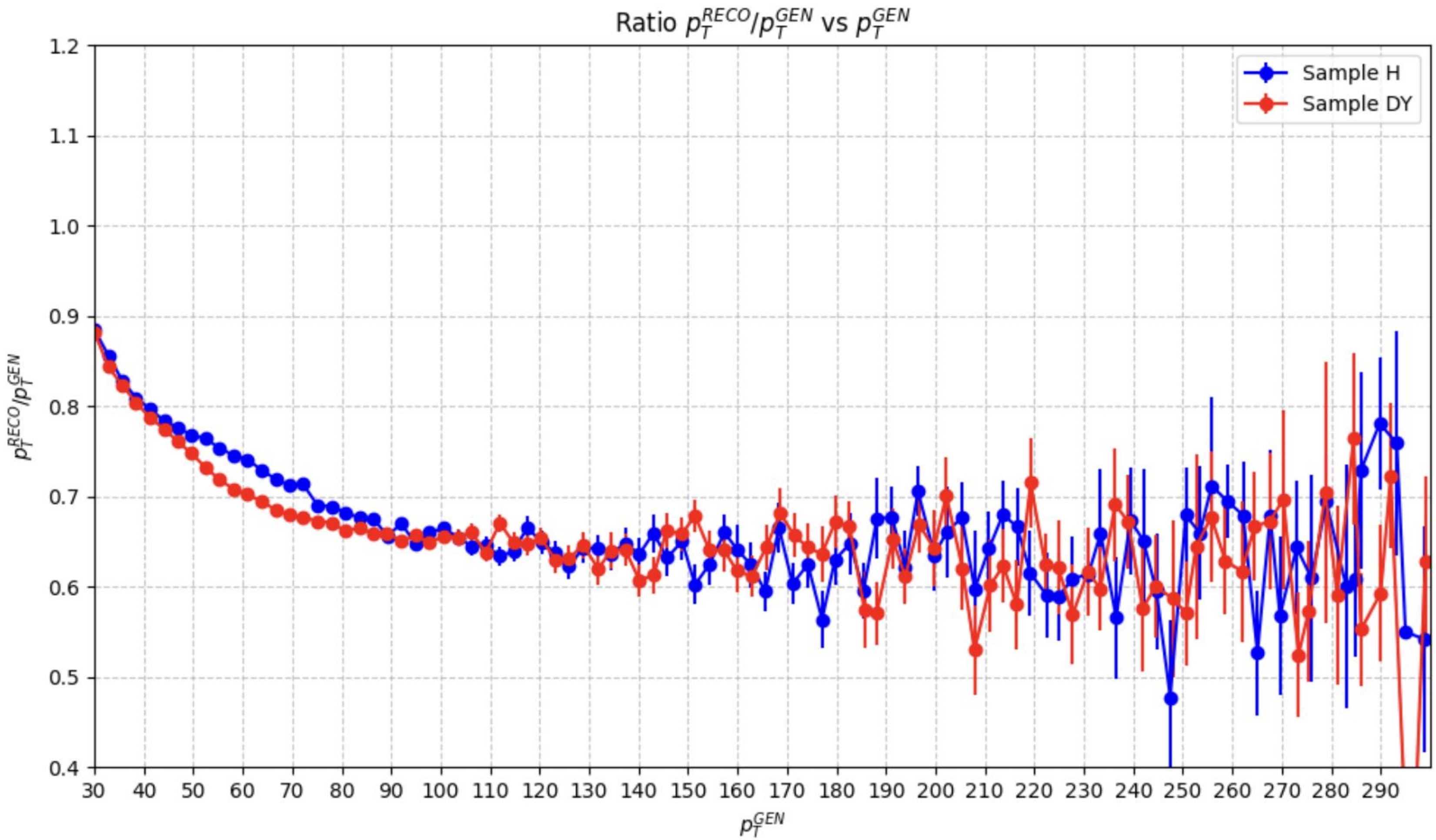








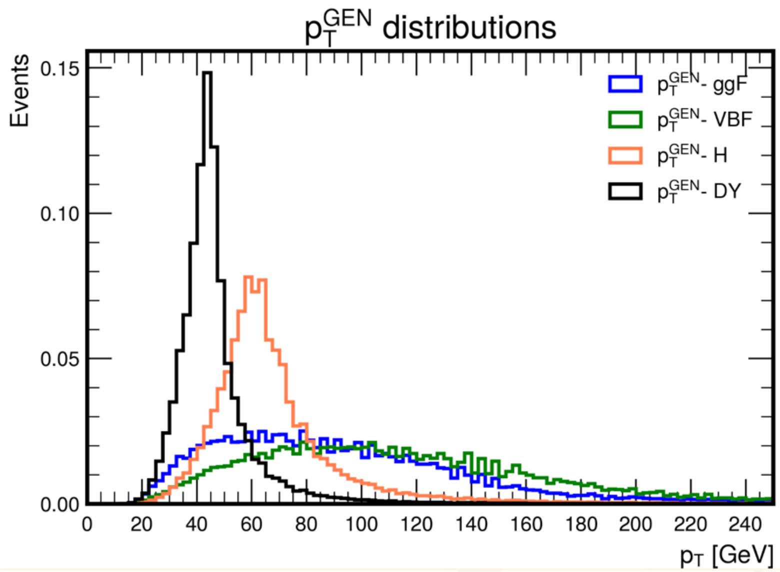


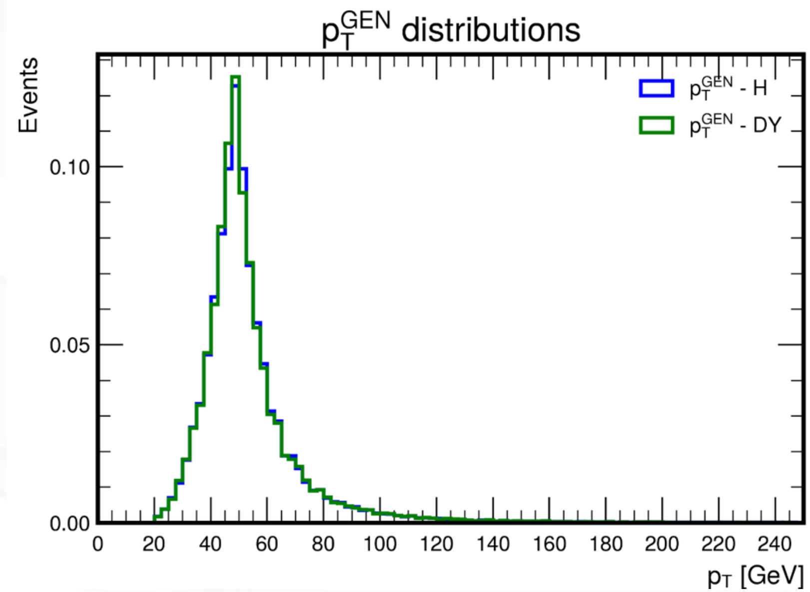

Ratio p_T^{RECO}/p_T^{GEN} vs p_T^{GEN}
Sample H
Sample DY
p_T^{RECO}/p_T^{GEN}
p_T^{GEN}
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290

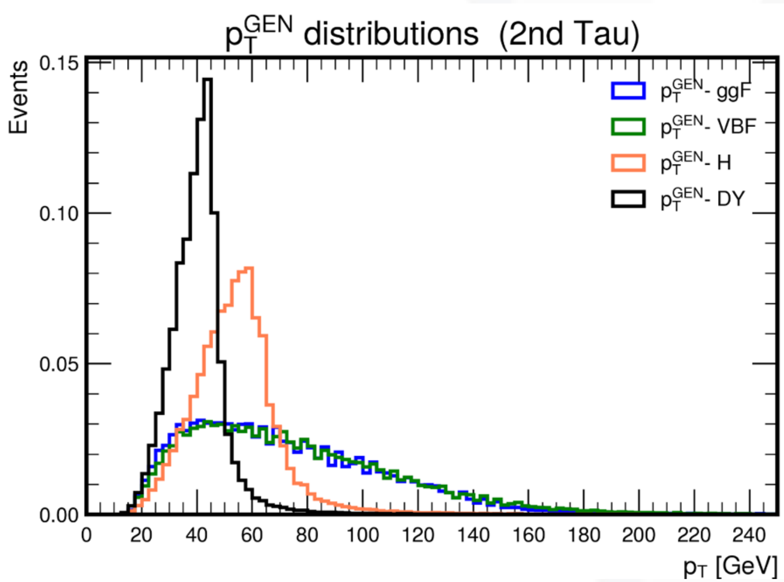
p_T^{GEN} distributions

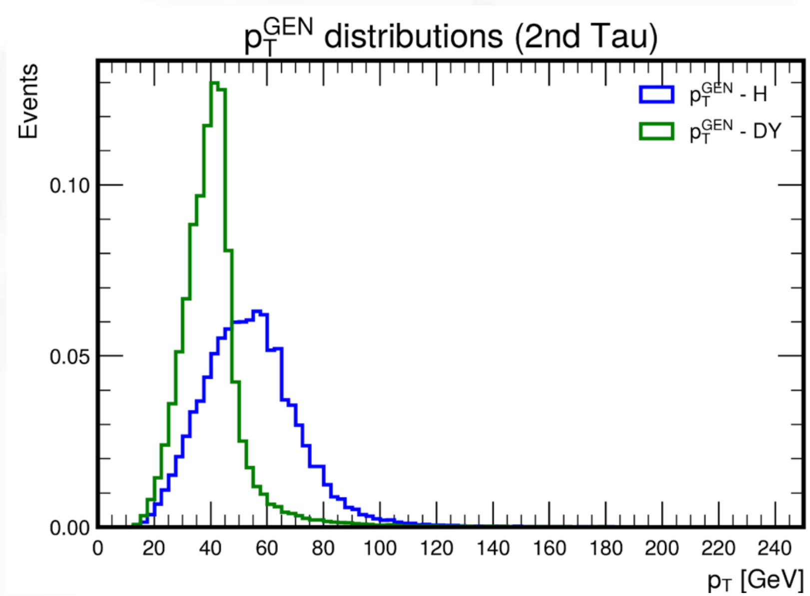
Before resampling on the first tau

After resampling on the first tau







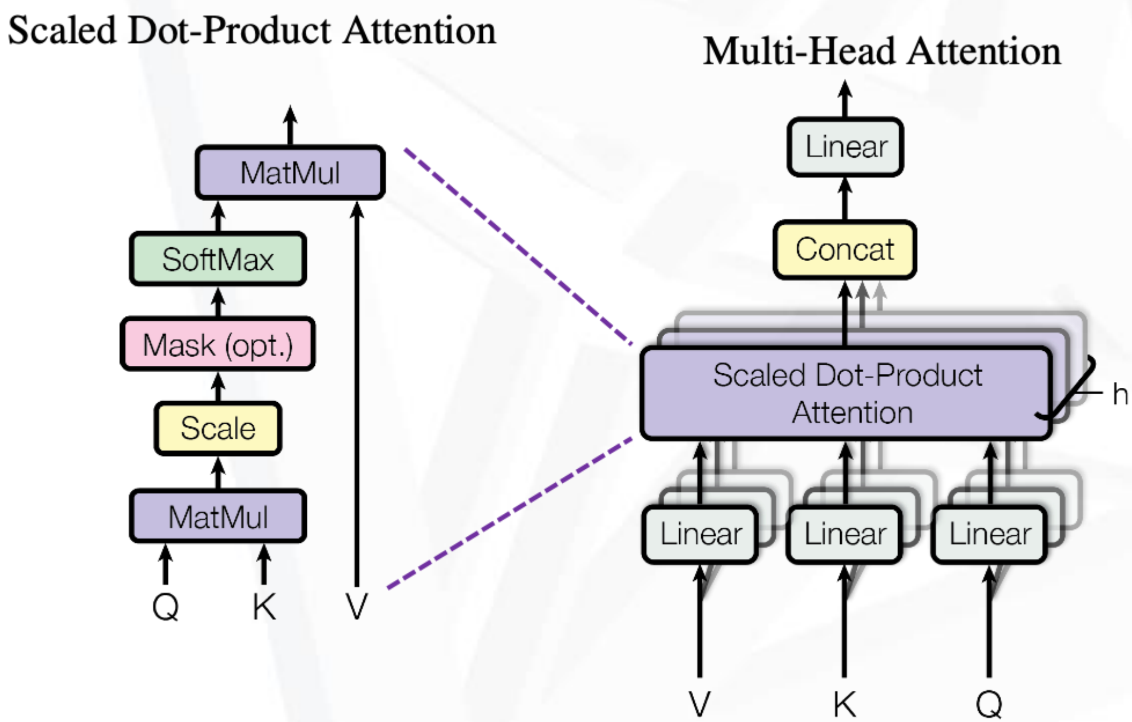


Scaled Dot - Product

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

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

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



Self-Attention

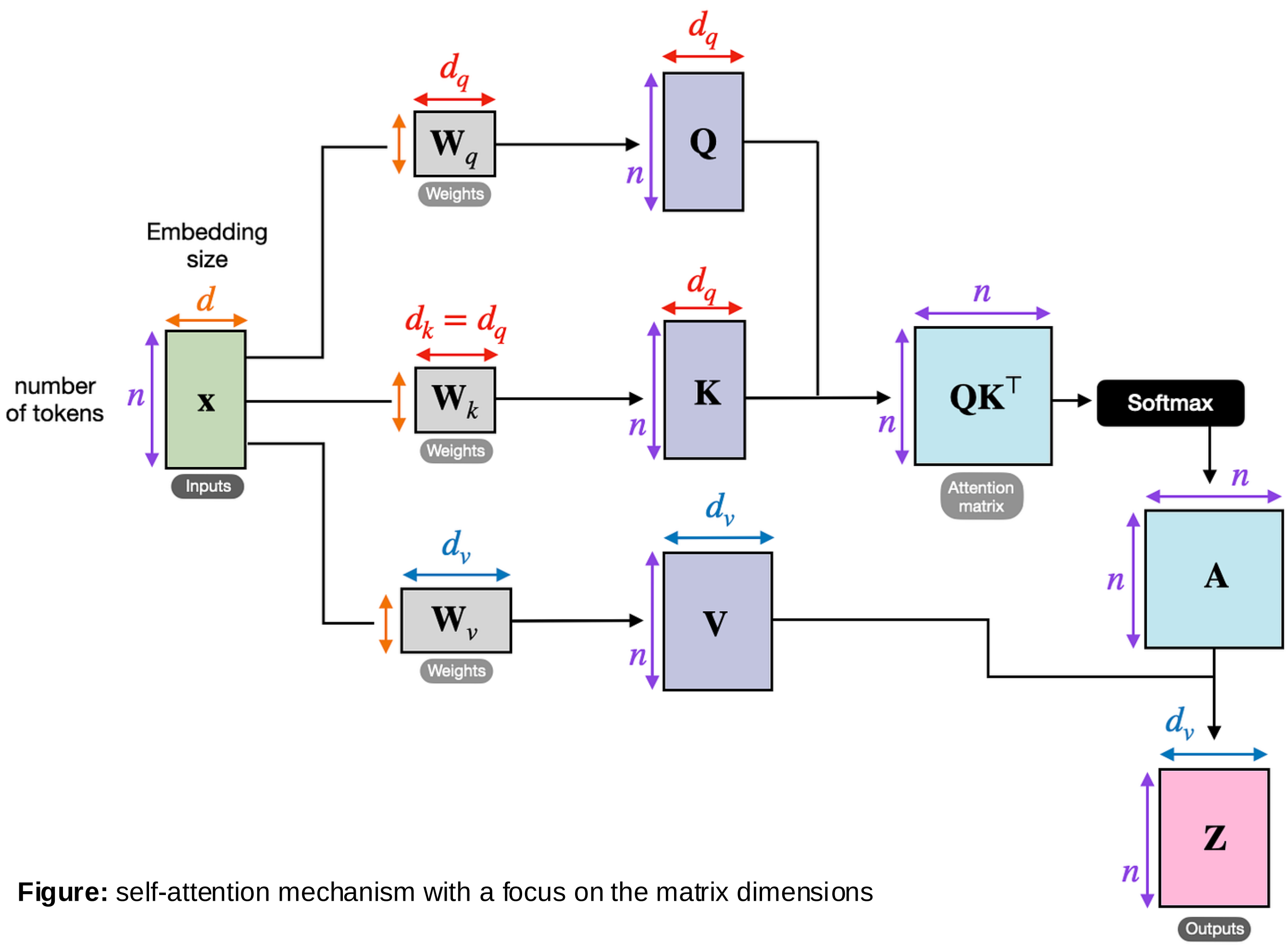


Figure: self-attention mechanism with a focus on the matrix dimensions

Cross-Attention

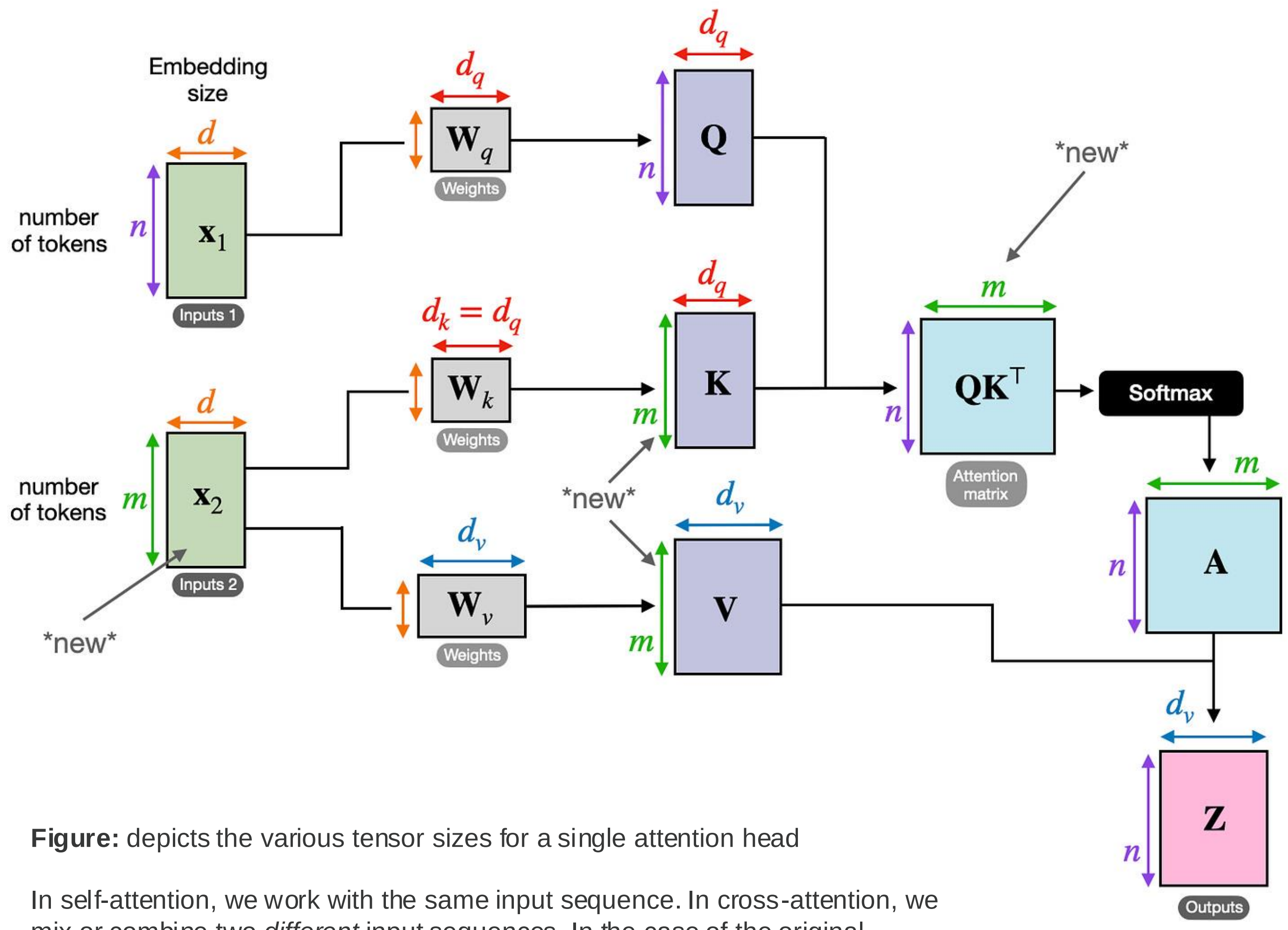


Figure: depicts the various tensor sizes for a single attention head

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

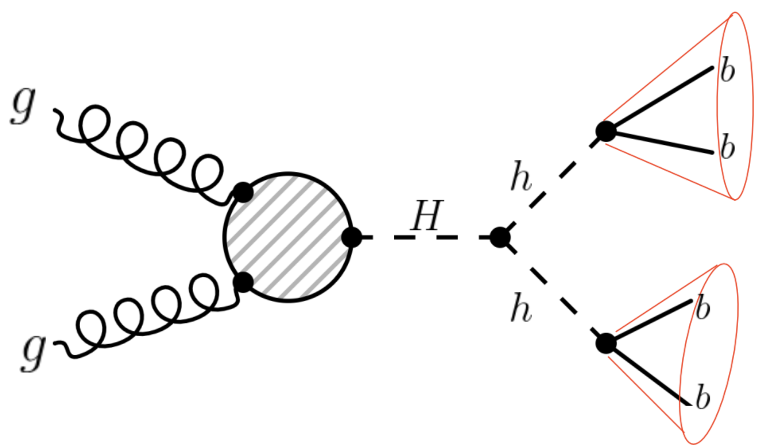


Figure 2: Feynman diagram for the signal process.