Multi-Scale Transformer Encoder for Di-Tau Invariant Mass Reconstruction

Valentina Camagni

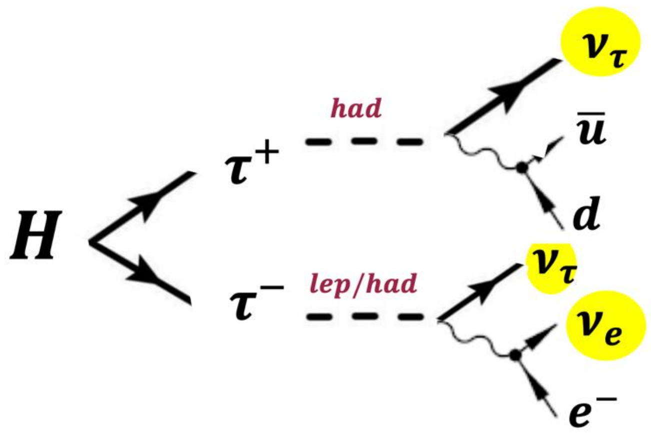
on behalf of the CMS collaboration

EPS-HEP 2025

July 9, 2025

Introduction (1)

Reconstructing the di-tau invariant mass ($m_{\tau\tau}$) is crucial for precise SM measurements and BSM searches.
Neutrinos from tau decays escape detection, worsening mass resolution and complicating resonance identification.



⟹ The CMS experiment currently employs the **Secondary Vertex Fit (SVFit)** algorithm [1]

✕ High computational time ⟹ New strategies based on deep learning **Tau Pair Mass Transformer (TPMT)**

Aim

Reconstruct the four-momentum of each τ lepton prior to its decay, in order to recover the kinematics of the parent particle and accurately estimate the invariant mass

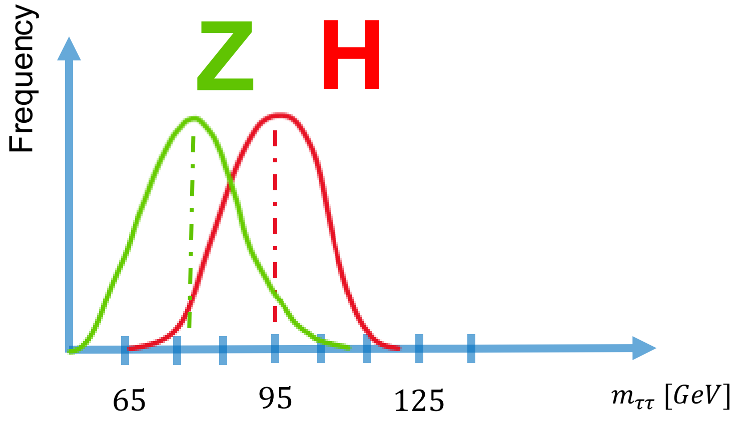
Introduction (2)

measured from visible decay products only

$\boldsymbol{\tau_1}$ p_T^{RECO}

$\boldsymbol{\tau_2}$ p_T^{RECO}

$m_{\tau\tau}^{RECO}$



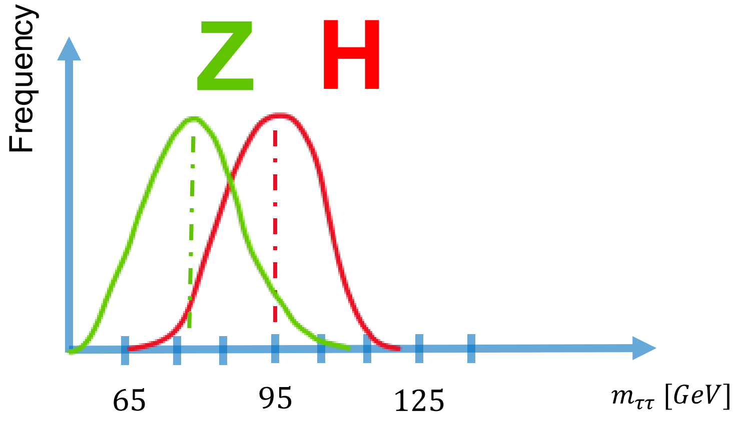
Introduction (2)

measured from visible decay products only

τ_1 p_T^{RECO} $\Rightarrow$ CORRECTION

τ_2 p_T^{RECO} $\Rightarrow$ CORRECTION

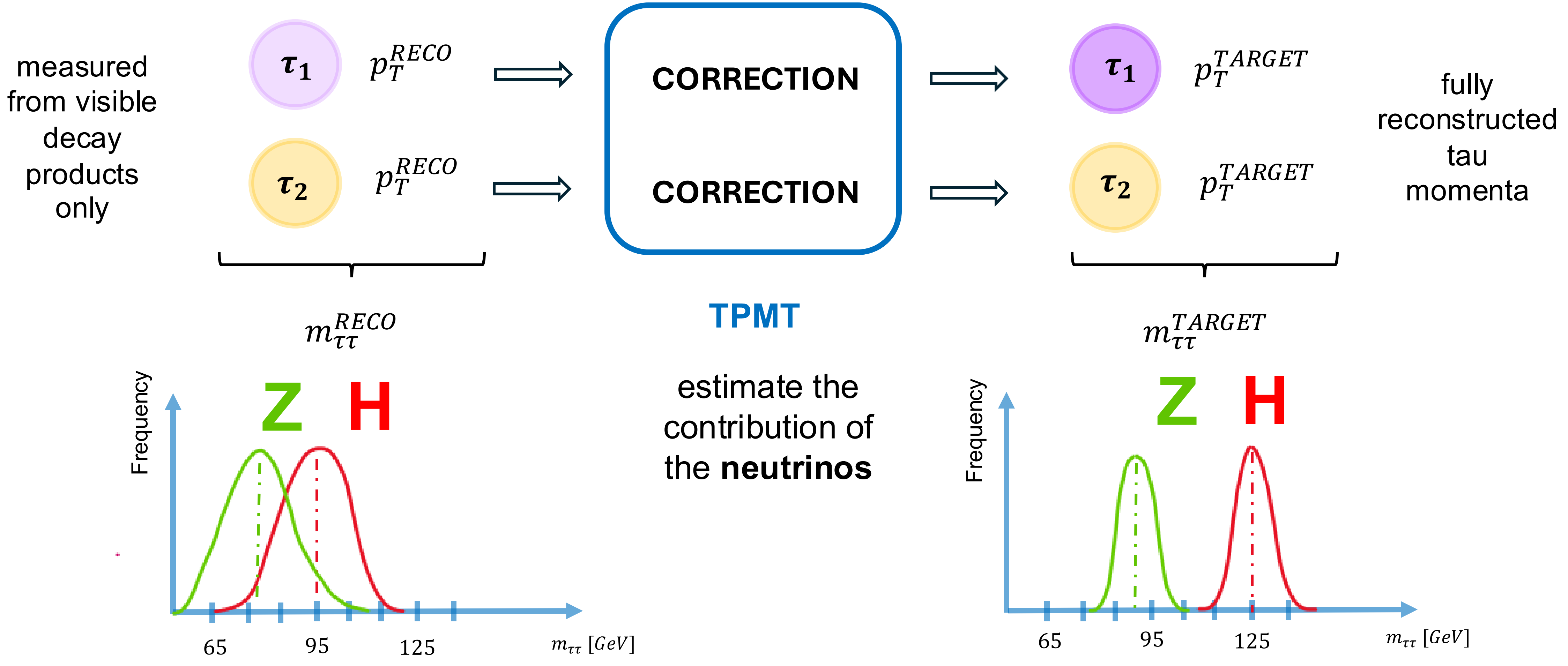
$m_{\tau\tau}^{RECO}$



TPMT

estimate the contribution of the **neutrinos**

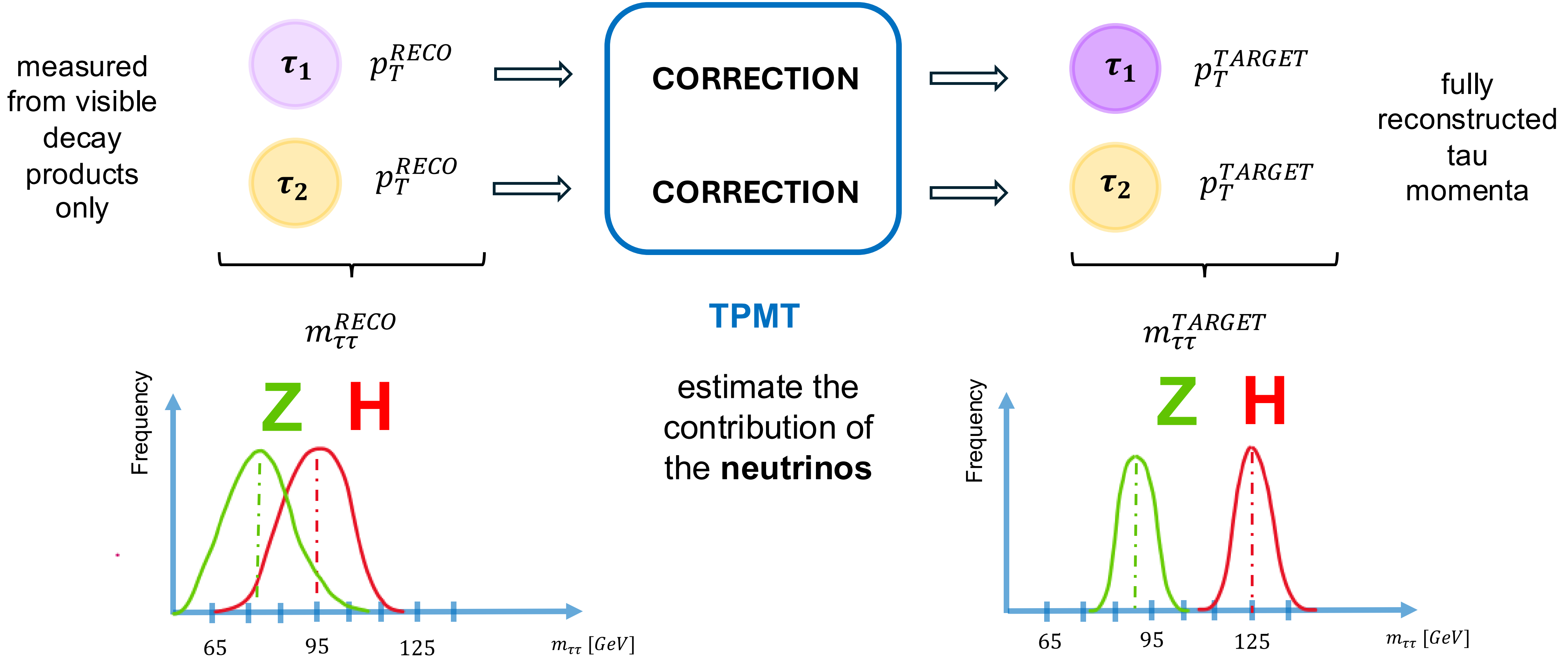
Introduction (2)



Introduction (2)

Assumption

- Uses η and ϕ from reconstructed taus
  → *Collinear approximation* valid for taus with $p_T^{RECO} > 20\ GeV$



First Strategy: Flat-mass samples

Goal: improve the reconstruction of the di-tau invariant mass by correcting the visible p_T of tau candidates for the momentum carried away by neutrinos, *without biasing the model toward any specific mass value*

Training configuration:

- Events from $\boldsymbol{X} \rightarrow \boldsymbol{\tau\tau}$ decays, generated via gluon-gluon fusion and vector boson fusion
- The invariant mass of the parent particle was drawn from a flat distribution in the range 30–300 GeV
- Includes both hadronic ($\tau_h\tau_h$) and semileptonic tau decays ($\tau_h\tau_\mu$)

Evaluation configuration:

After training, the model was tested on realistic resonant processes to assess performance:

- **Higgs boson production** ($H \rightarrow \tau\tau$ with $m_{\tau\tau} = 125\ GeV$)
- **Drell–Yan** events ($Z \rightarrow \tau\tau$ with $m_{\tau\tau} \approx 91\ GeV$)

Motivation

Using flat-mass training allows the model to learn from kinematics alone, avoiding dependence on Z/H mass peaks and minimizing sculpting effects

TPMT inputs

Feature importance analysis with **Random Forest** on the full set of taus, jets, MET variables to maximize H / Z classification

[2], [3]

Taus

$\boldsymbol{\tau_h \tau_h}$

H/Z/SUSY/X: 2 genuine τ_h (opposite charge, $p_T^{RECO} > 20\ GeV$)
$\boldsymbol{t\bar{t}}$**:** 2 fake τ_h that match ($\Delta R < 0.4$) to top daughters (b or W decay products)

$\boldsymbol{\tau_h \tau_\mu}$

H/Z/SUSY/X: 1 genuine τ_h + 1 e/μ from tau decay (opposite charge, $p_T^{RECO} > 20\ GeV$)
$\boldsymbol{t\bar{t}}$**:** 1 fake τ_h + 1 e/μ not from tau decay, both that match ($\Delta R < 0.4$) to top daughters (b or W decay products)

TauProd

Up to 10 decay products from the selected tau pair are included, sorted by p_T. In the semi-leptonic case, only the τ_h decay products are used. Zero-padding ensures fixed input size.

Jets

Up to 3 leading jets (with $\Delta \mathrm{R} > 0.4$ from the selected taus) are provided as input. If fewer than 3 are found, zero-padding is applied.

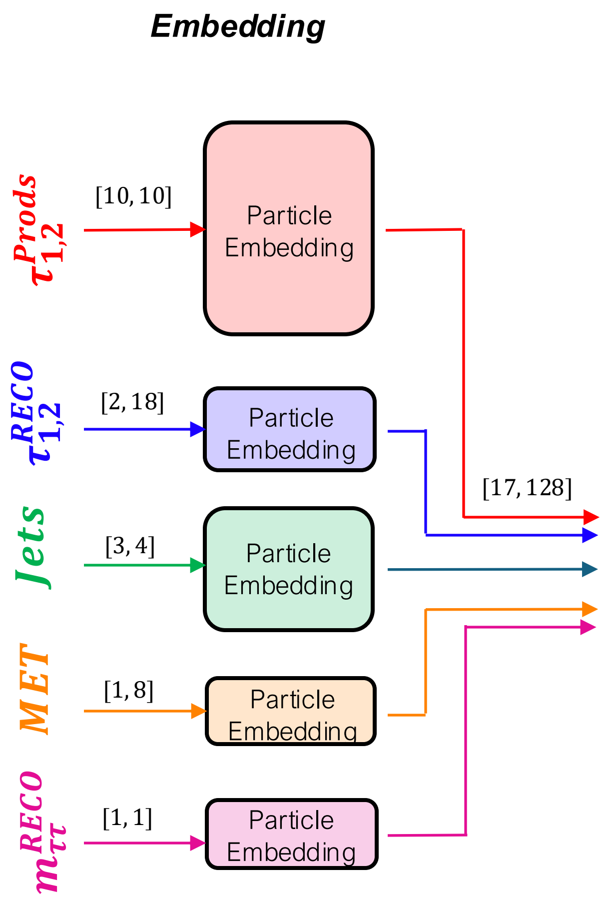
MET

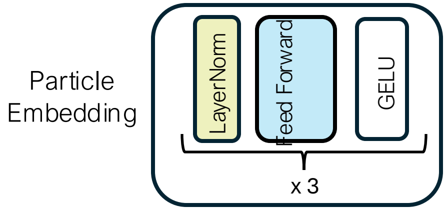
Particle Flow MET is included as a global feature to account for undetected neutrinos.

$\boldsymbol{m_{\tau\tau}^{RECO}}$

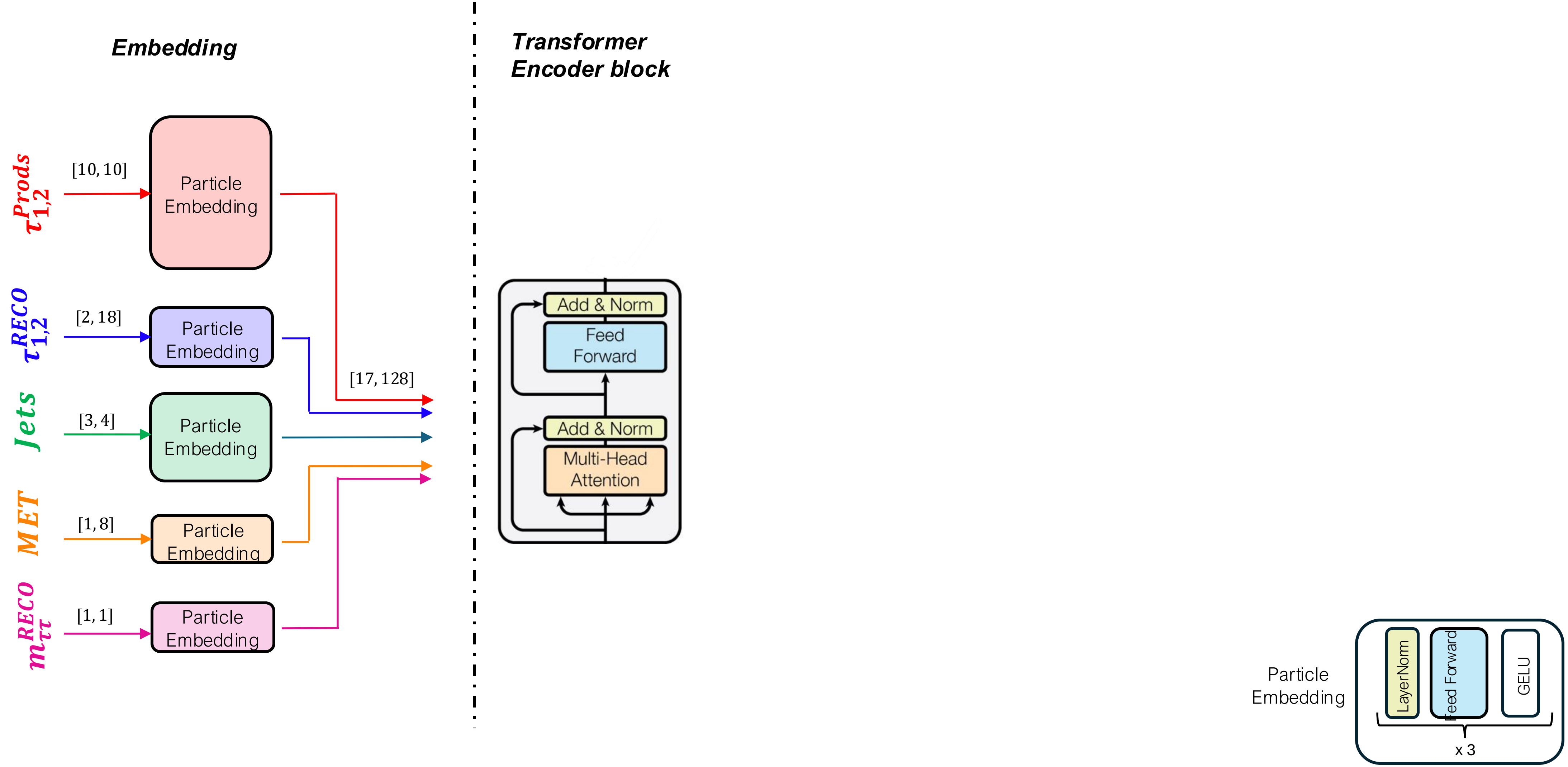
A scalar input with the visible di-tau mass.

TPMT Architecture [4]

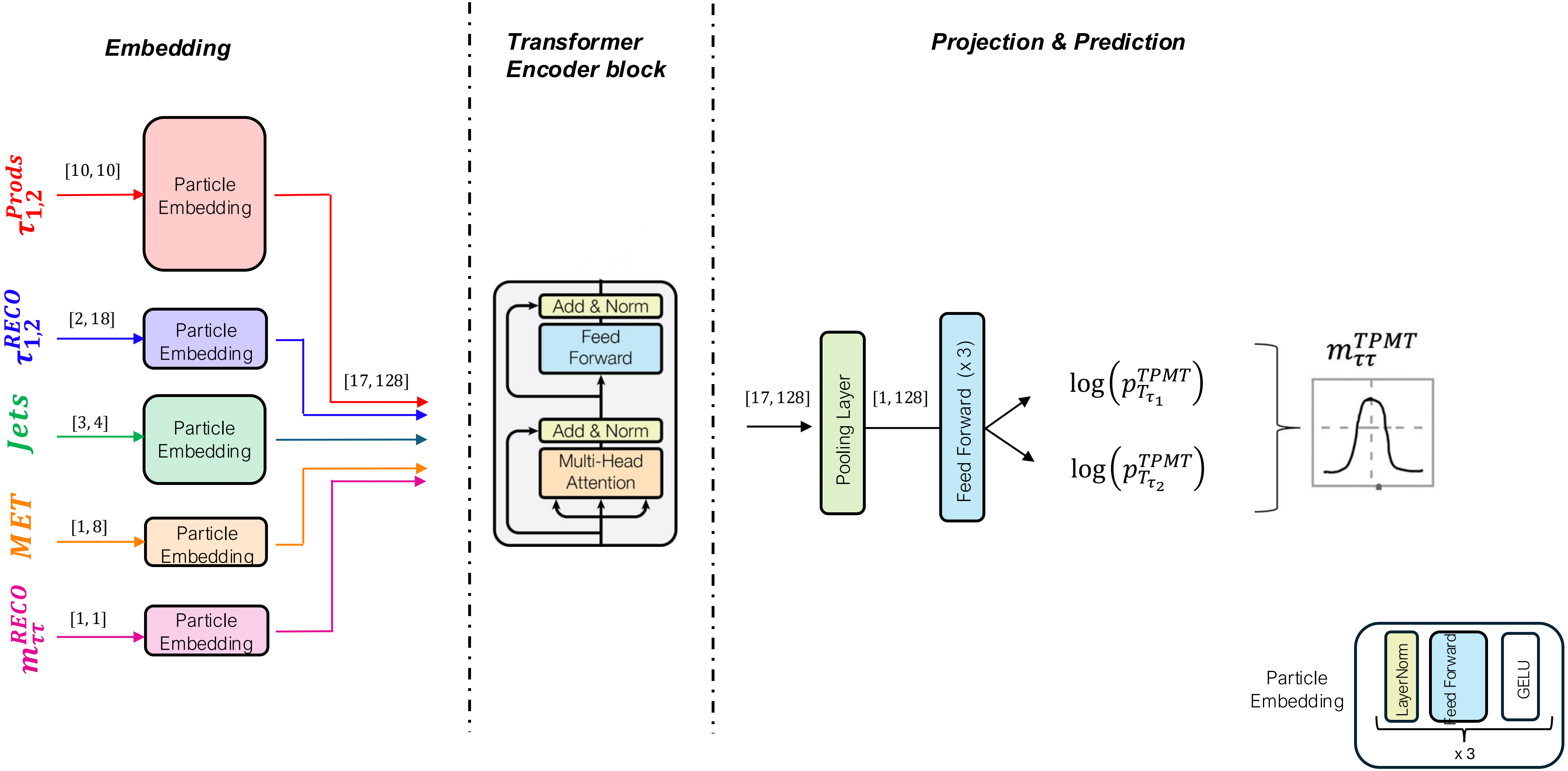




TPMT Architecture [4]



TPMT Architecture [4]



Training Hyperparameters

- ***Batch size***: 1024
- **Initial learning rate**: 1×10^{-4}
  Reduced to 1×10^{-6} using *ReduceLROnPlateau*
- **N° Encoder blocks**: 1
- **N° Attention heads**: 8
- **d_model** (Embedding dimension): 512
- **EarlyStopping patience**: 15 epochs
- **ReduceLROnPlateau patience**: 10 epochs
- **Optimizer**: Adam
- **Trainable parameters**: ~ 800,000
- **Hardware**: NVIDIA Tesla T4 (16GB)
- **Inference time:** $10^{-3} s$ per event
  (for SVFit, $\mathcal{O}$(s) per event)
- **GPU memory usage**: ~ 50%

$\boldsymbol{Flat\ Mass\ Trainings}$

Pairtype	*Total Training Events*
$\tau_h \tau_h$	~ 800 k

VBF: 400k, GGF: 400k

Loss Design

To guide the network toward physically meaningful predictions, the loss function is the weighted sum of two terms, both based on the MAE:

$$\mathcal{L}_{\text{total}} = \lambda_\tau \cdot \mathcal{L}_\tau + \lambda_{m_{\tau\tau}} \cdot \mathcal{L}_{m_{\tau\tau}}$$

- **Tau p_T loss ($\mathcal{L}_\tau$):**

  $MAE(p_T^{TPMT}, p_T^{TARGET})$ for the two taus

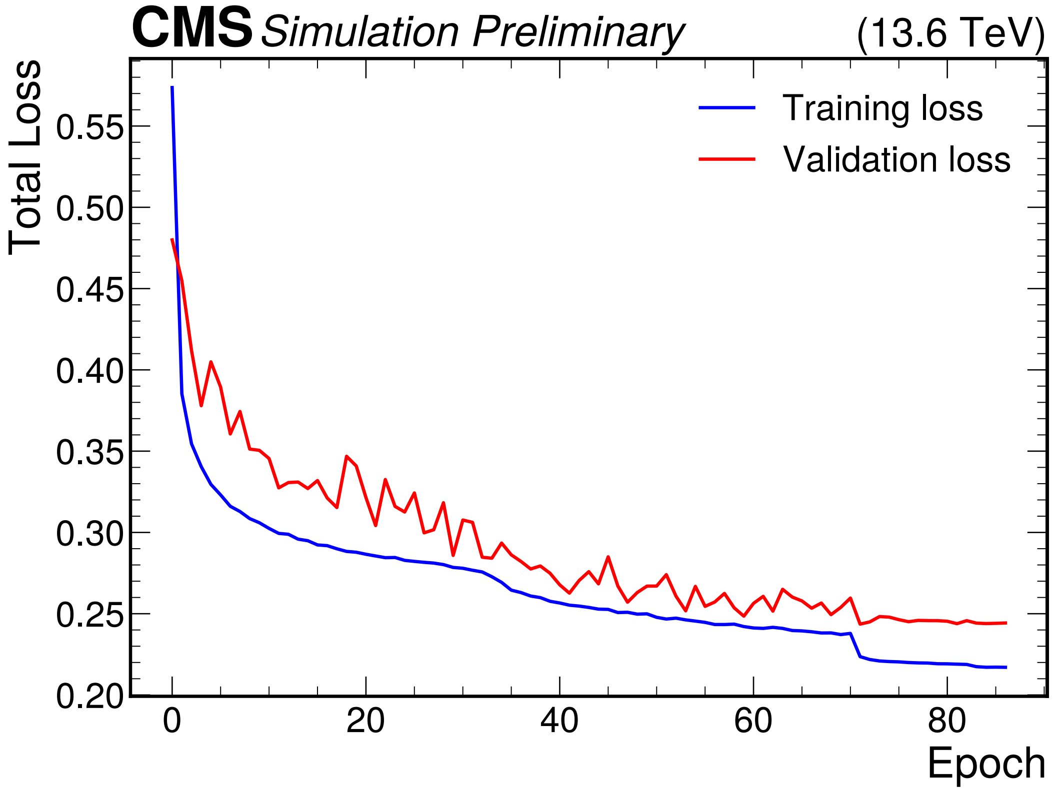
- **Invariant Mass Loss ($\mathcal{L}_{m_{\tau\tau}}$):**

  $$MAE(m_{\tau\tau}^{TPMT}, m_{\tau\tau}^{TARGET})$$

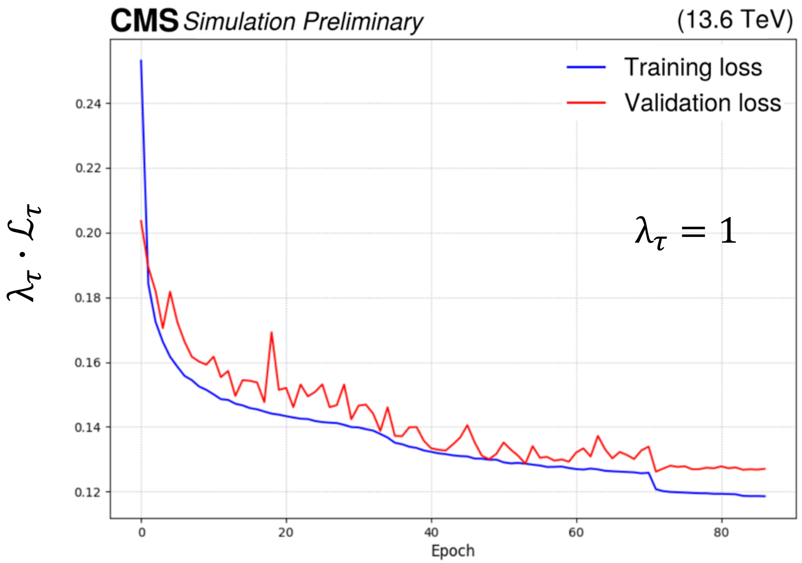
While $\mathcal{L}_\tau$ ensures per-object accuracy, small errors in p_T can cause large deviations in $m_{\tau\tau}$ due to its non-linear dependence on kinematics: $\mathcal{L}_{m_{\tau\tau}}$ term helps correct for this and encourages physically consistent predictions

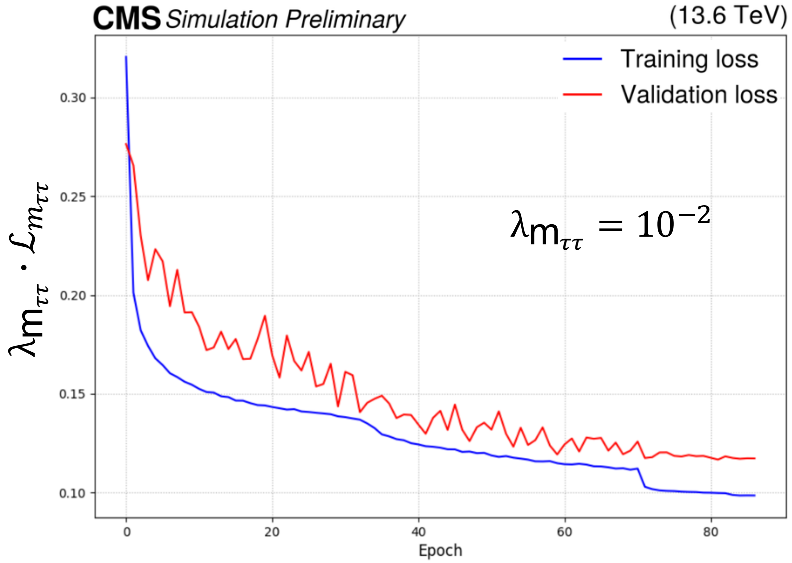
$\lambda_\tau = 1$
$\lambda_{m_{\tau\tau}} = 10^{-2}$

chosen to balance the two contributions to the same order of magnitude during training



- Rapid decrease in early epochs
- Learning rate drops visible (triggered by ReduceLROnPlateau)
- Losses plateau around epoch ~ 75 ($lr = 10^{-6}$)
- Training vs validation:
  good agreement
  → good generalization





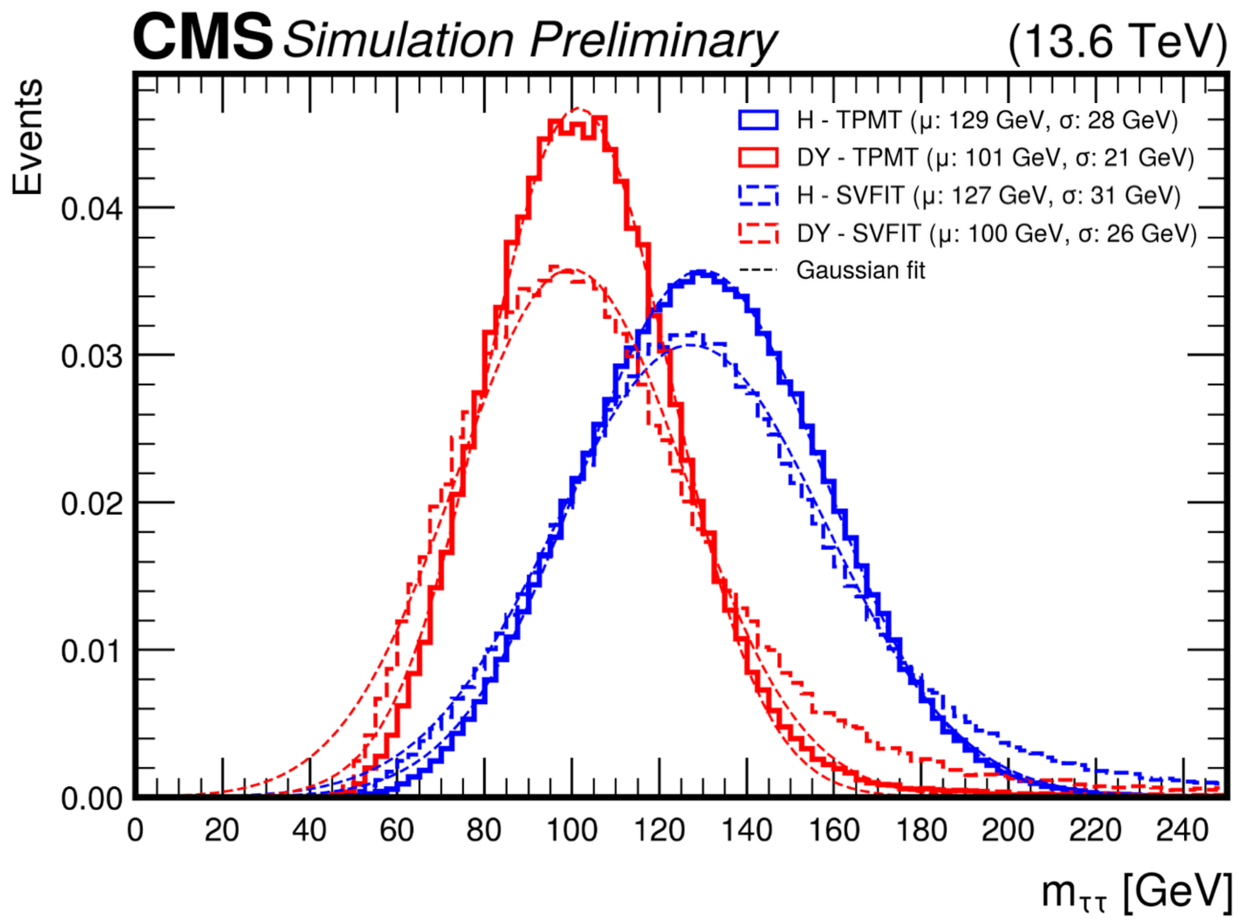
- Convergence still under study
- Loss components follow similar trends

↓

smooth and stable total loss

Results

$m_{\tau\tau}^{TPMT}$ distribution from Flat Mass Training



TPMT vs SVFit comparison on test samples:

- $H \rightarrow \tau\tau$ and $Z \rightarrow \tau\tau$

Solid lines: TPMT prediction from predicted $\log(p_T^{TARGET})$ of the two taus

Dashed lines: SVFit $m_{\tau\tau}$ reconstruction + Gaussian fits

TPMT matches SVFit resolution on $H \rightarrow \tau\tau$ and $Z \rightarrow \tau\tau$ (fully hadronic),
with $> 1000x$ faster inference (ms vs $s/event$)

Second Strategy: Resonant-mass samples

Specialized model, tailored for $H \to \tau\tau$, not general-purpose τ reconstruction

Training and evalution configuration:

- **SM-only:**
  - $DY \to \tau\tau \ (m_{\tau\tau} > 50\ GeV),$
  - $H \ \to \tau\tau \ (m_{\tau\tau} = 125\ GeV),$
  - $t\bar{t} \to (W^{+}b)(W^{-}\bar{b})$
- **SM+BSM:**
  SM-only processes
  \+ SUSY 2HDM signals
  (**ggH**/**bbH**, $m_{\tau\tau} = 350\ GeV$)

Motivation

Resonant samples to better reflect the conditions expected at application time, where a resonance at a fixed mass is assumed

Second Strategy: Resonant-mass samples

Specialized model, tailored for $H \rightarrow \tau\tau$, not general-purpose τ reconstruction

Training and evalution configuration:

- **SM-only:**
  - $DY \rightarrow \tau\tau\ (m_{\tau\tau} > 50\ GeV)$,
  - $H \rightarrow \tau\tau\ (m_{\tau\tau} = 125\ GeV)$,
  - $t\bar{t} \rightarrow (W^+b)(W^-\bar{b})$
- **SM+BSM:**
  SM-only processes
  + SUSY 2HDM signals
  (**ggH**/**bbH**, $m_{\tau\tau} = 350\ GeV$)

! The choice of the 350 GeV benchmark is not driven by any specific theoretical motivation

Motivation

Resonant samples to better reflect the conditions expected at application time, where a resonance at a fixed mass is assumed

Channel-specific trainings for better adaptation to physics scenario and decay mode

Second Strategy: Resonant-mass samples

Specialized model, tailored for $H \to \tau\tau$, not general-purpose τ reconstruction

Training and evalution configuration:

- **SM-only:**
  - $DY \to \tau\tau\ (m_{\tau\tau} > 50\ GeV)$,
  - $H \to \tau\tau\ (m_{\tau\tau} = 125\ GeV)$,
    - ➢ $p_T^{TARGET} = p_T^{MC}$
  - $t\bar{t} \to (W^+b)(W^-\bar{b})$
    - ➢ used as **fake-τ** background
    - ➢ $p_T^{TARGET} = p_T^{RECO}$

enables the model to preserve the original p_T^{RECO} input for fake candidates, ensuring consistent treatment across genuine and misidentified taus

- **SM+BSM:**

  SM-only processes
  + SUSY 2HDM signals (ggH/bbH, $m_{\tau\tau} = 350\ GeV$)

! The choice of the 350 GeV benchmark is not driven by any specific theoretical motivation

Motivation

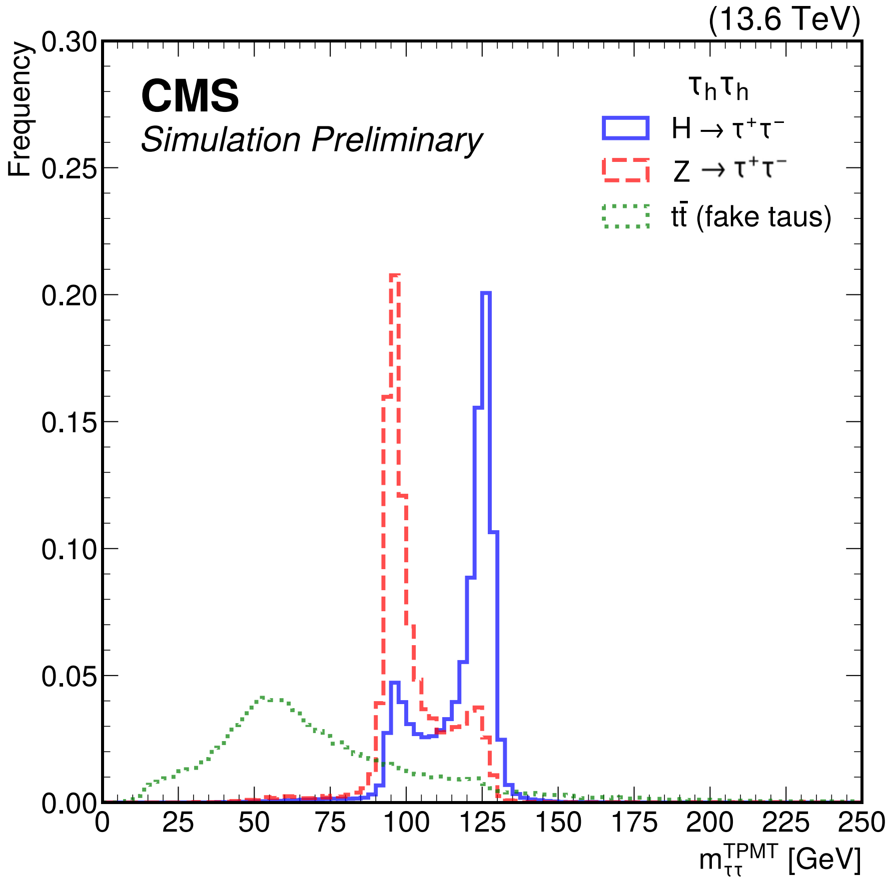
Resonant samples to better reflect the conditions expected at application time, where a resonance at a fixed mass is assumed

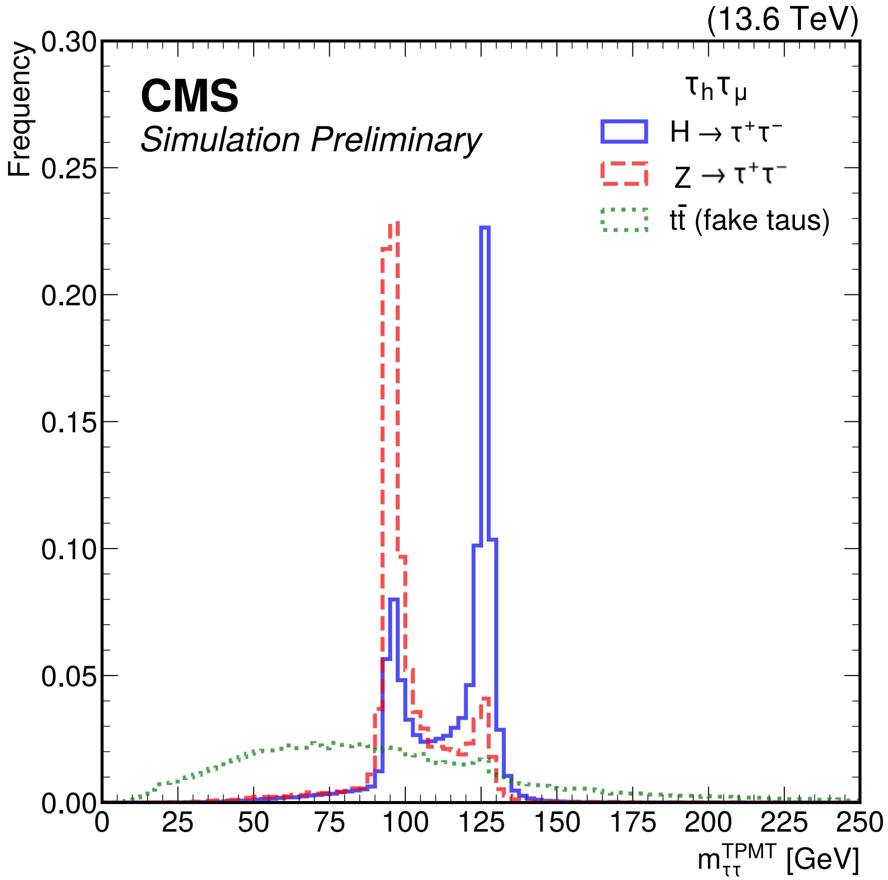
Channel-specific trainings for better adaptation to physics scenario and decay mode

Results

$m_{\tau\tau}^{TPMT}$ **distribution from** ***SM-only scenario***
Resonance Mass Training

Differences across final states → under study





TPMT shows lower Z/$t\bar{t}$ contamination in a signal-specific mass window designed to retain 90% signal efficiency

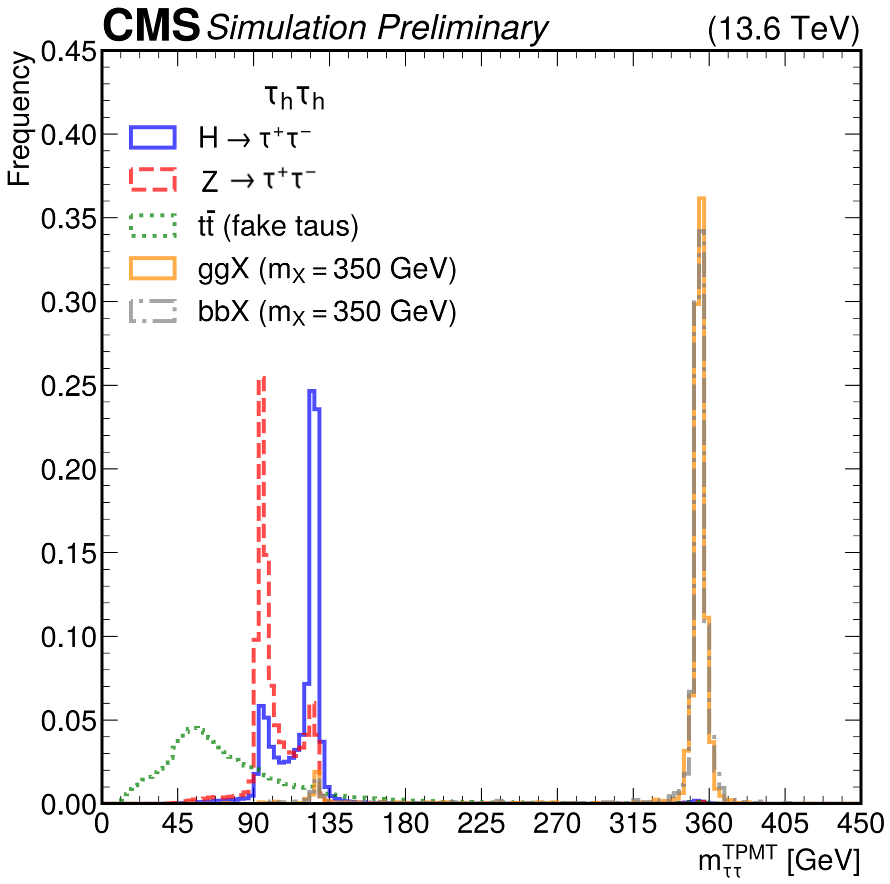
Two-mode behavior:
H and Z each show primary + secondary peaks
→ Model favors either ~125 or ~ 91 GeV, reflecting learned resonance structure

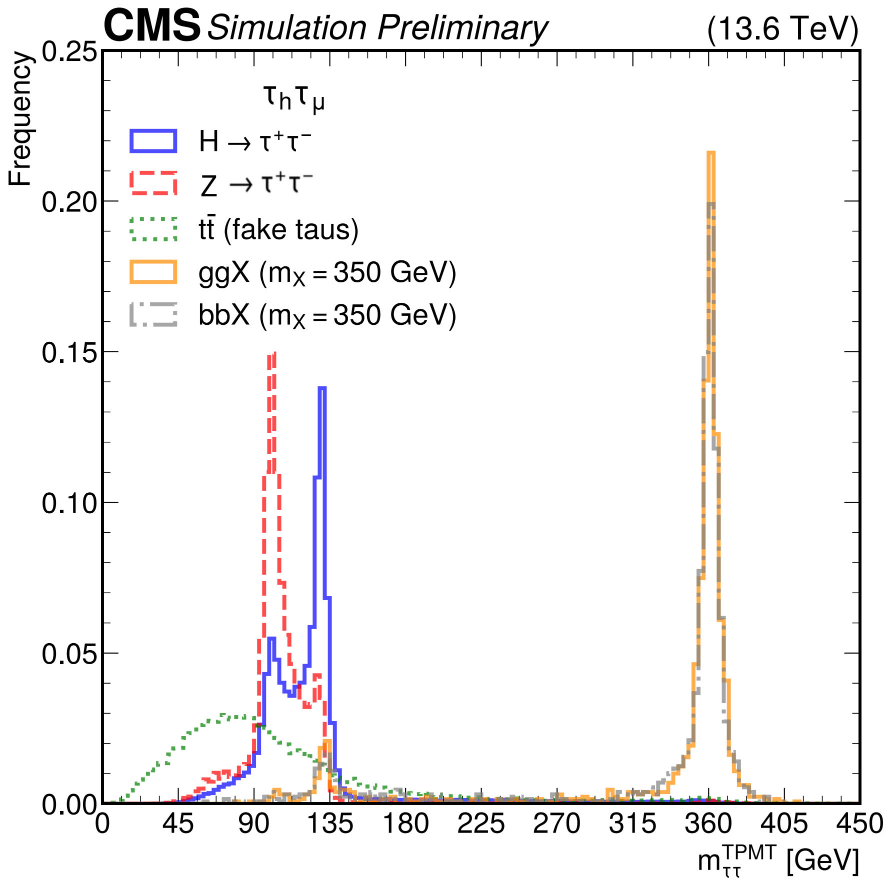
Results

$m_{\tau\tau}^{TPMT}$ distribution from Resonance Mass Training

SM+BSM scenario

Differences across final states → under study





Z vs misreco H (~ 91 GeV):

Features overlap → hard to separate (Random Forest AUC = 0.60)

Two-mode behavior:
H and Z each show primary + secondary peaks
→ Model favors either ~125 or ~ 91 GeV, reflecting learned resonance structure

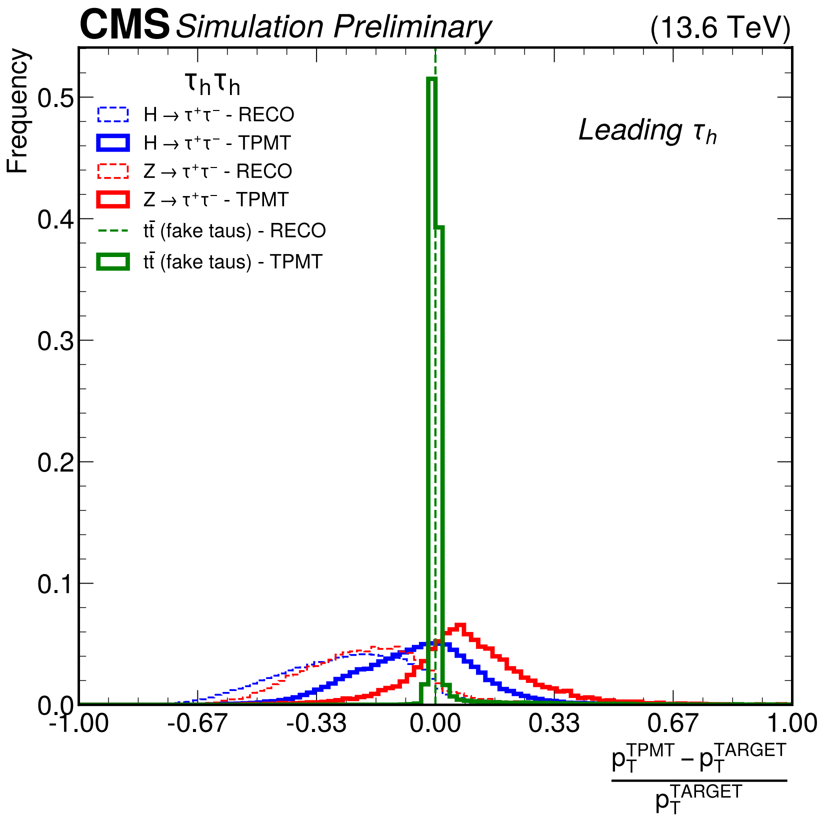
Results

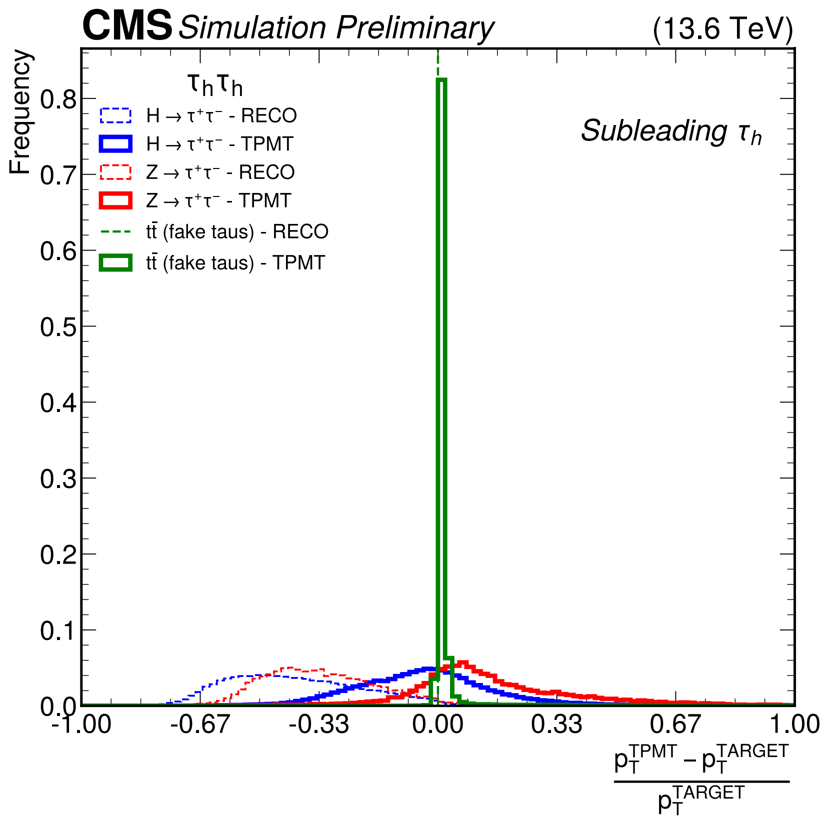
p_T distribution from Resonance Mass Training

SM-only scenario

$\tau_h\tau_h$

Leading vs subleading τ: differences still under study





TPMT predictions
↓
better match to target vs. visible p_T^{RECO}

in $t\bar{t}$ (fakes):

- $p_T^{TARGET} = p_T^{RECO}$
- model learns to keep them unchanged

Broader distributions for H and Z
↓
residual bias remains

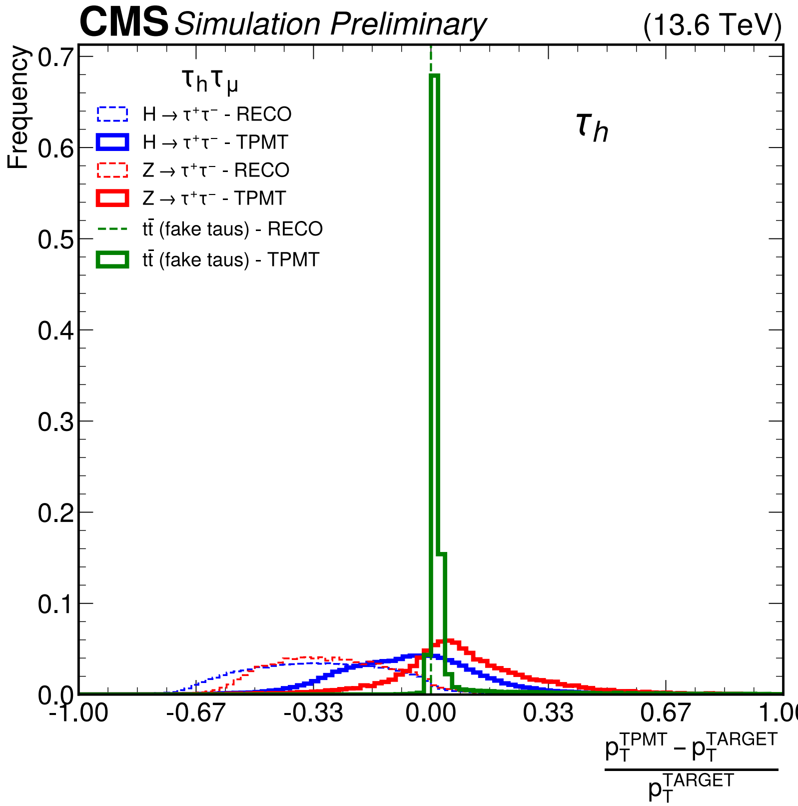
Results

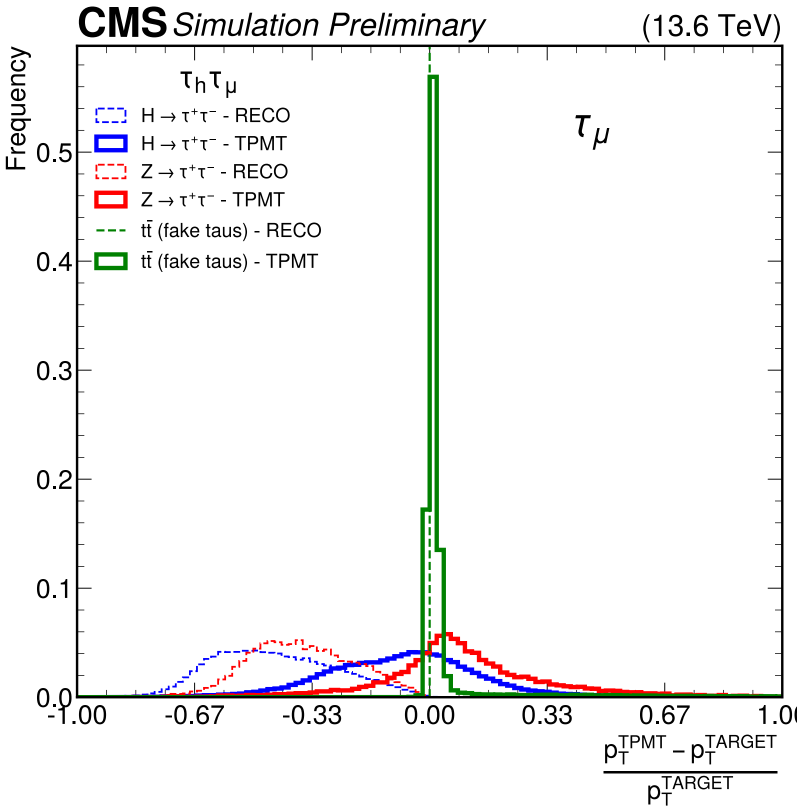
$\tau_h\tau_\mu$

p_T distribution from Resonance Mass Training

SM-only scenario

τ_h and τ_μ: differences still under study





TPMT predictions
↓
better match to target vs. visible p_T^{RECO}

in $t\bar{t}$ (fakes):
- $p_T^{TARGET} = p_T^{RECO}$
- model learns to keep them unchanged

Broader distributions for H and Z
↓
residual bias remains

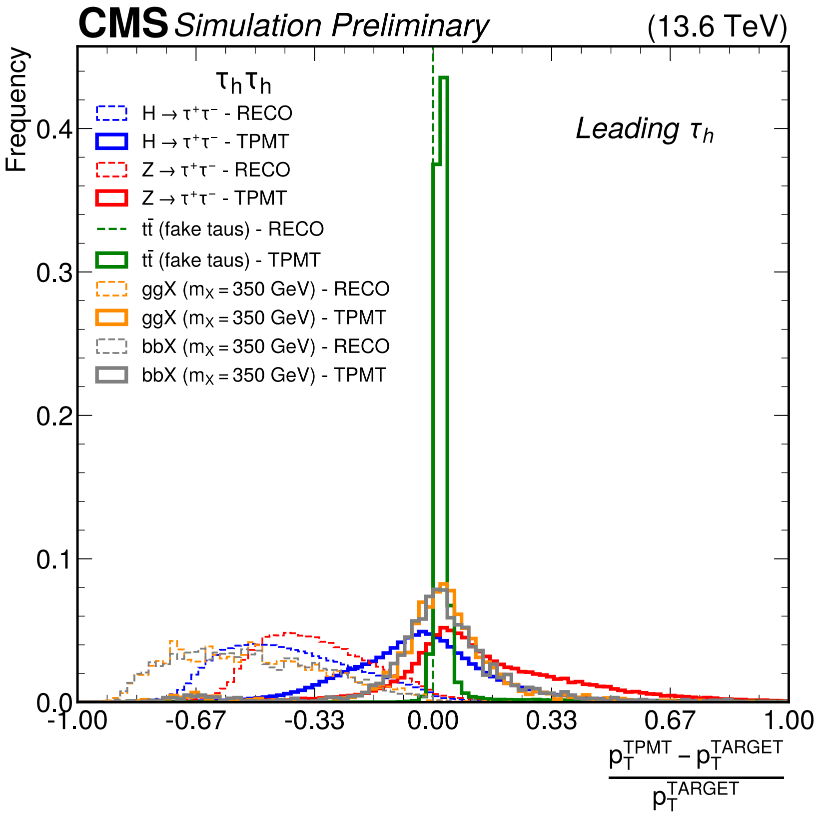
Results

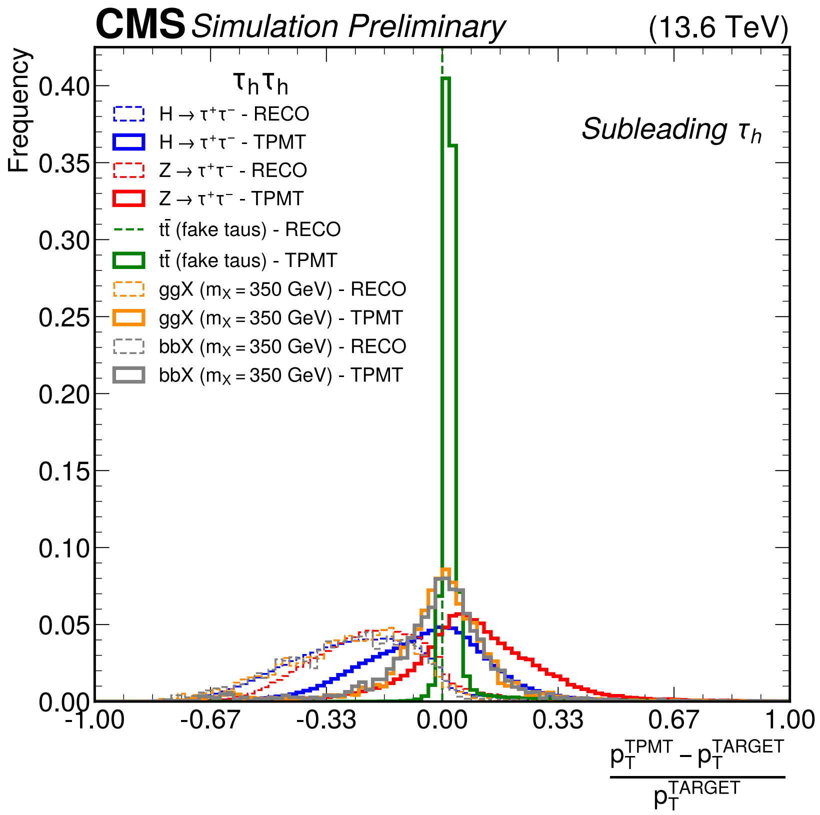
p_T distribution from Resonance Mass Training

SM+BSM scenario

$\tau_h\tau_h$

Leading vs subleading τ: differences still under study





TPMT predictions
↓
better match to target vs. visible p_T^{RECO}

in $t\bar{t}$ (fakes):
- $p_T^{TARGET} = p_T^{RECO}$
- model learns to keep them unchanged

Broader distributions for H and Z
↓
residual bias remains

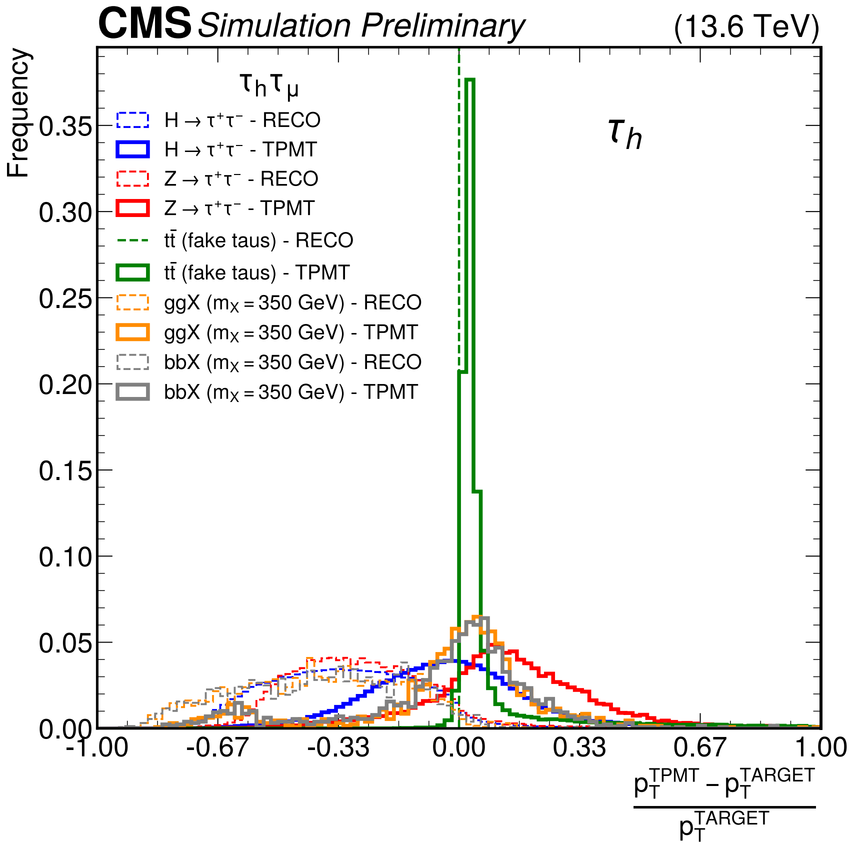
Results

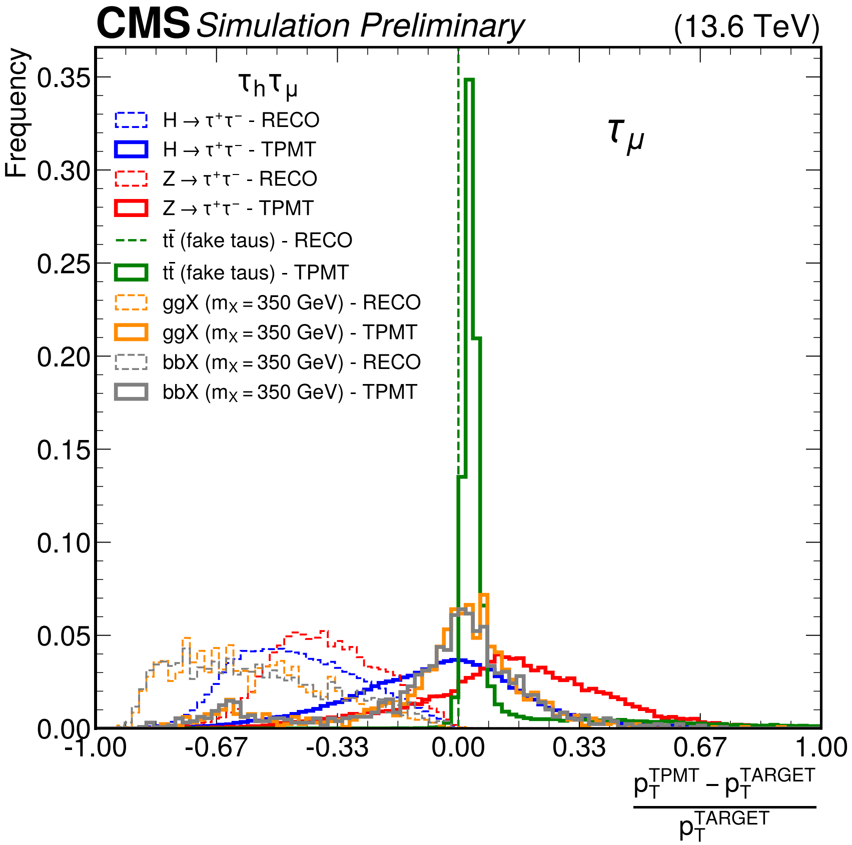
p_T distribution from Resonance Mass Training

SM+BSM scenario

$\tau_h \tau_\mu$

τ_h and τ_μ: differences still under study





TPMT predictions
↓
better match to target vs. visible p_T^{RECO}

in $t\bar{t}$ (fakes):

- $p_T^{TARGET} = p_T^{RECO}$
- model learns to keep them unchanged

Broader distributions for H and Z
↓
residual bias remains

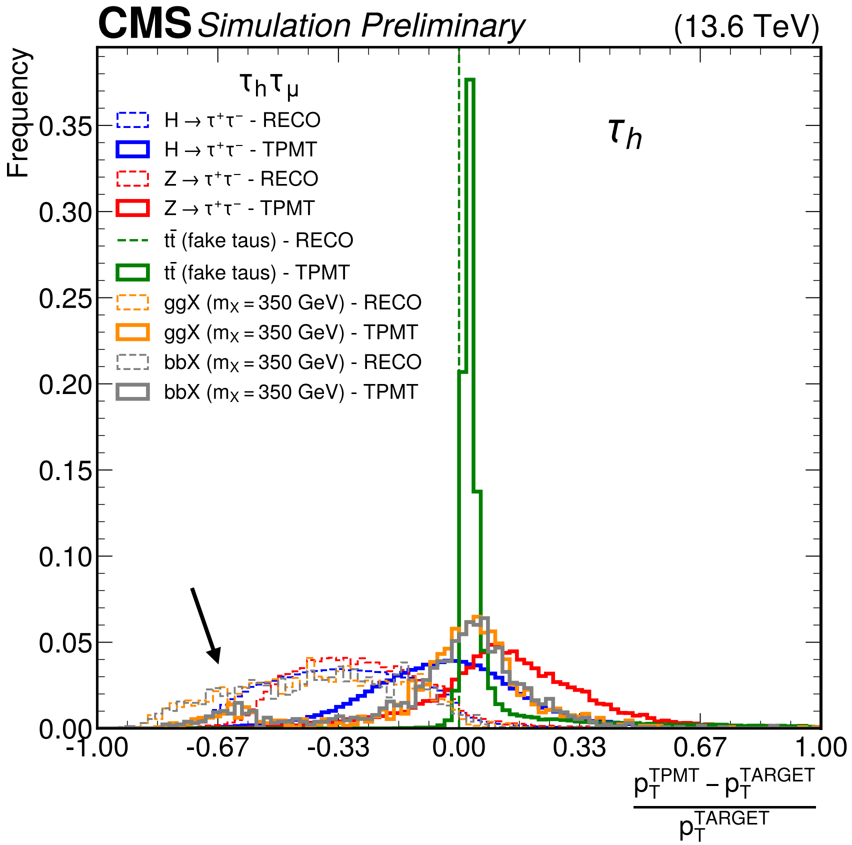
Results

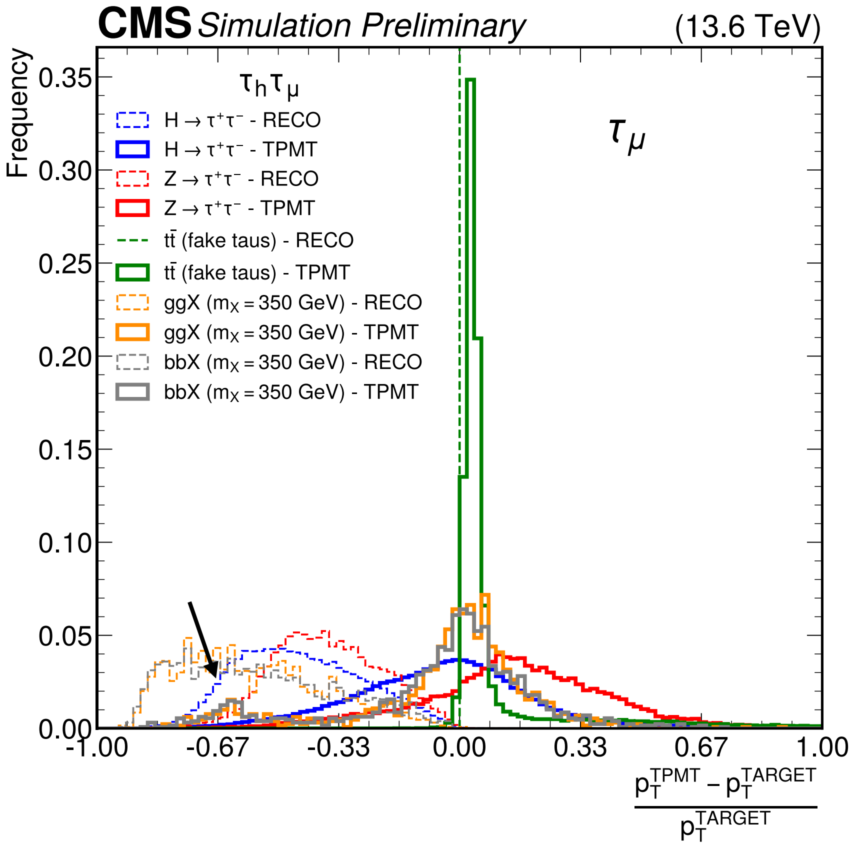
p_T distribution from Resonance Mass Training

SM+BSM scenario

$\tau_h \tau_\mu$

τ_h and τ_μ: differences still under study





This structure reflects a systematic underestimation of p_T, causing SUSY events to be misreconstructed near 125 GeV

TPMT predictions
↓
better match to target vs. visible p_T^{RECO}

in $t\bar{t}$ (fakes):
- $p_T^{TARGET} = p_T^{RECO}$
- model learns to keep them unchanged

Broader distributions for H and Z
↓
residual bias remains

Conclusions & Outlook

- Explored a Transformer-based approach for di-τ identification and kinematic reconstruction in SM and BSM scenarios

- The model shows promising performance, but further refinement is needed.

Key Observations:

- Binary-like behavior in mass reconstruction, especially H vs Z separation

- Limited smooth interpolation across overlapping mass regions.

Future developments:

- Introduce a parametric variant (mass hypothesis as input) to improve flexibility

- Focus on better generalization across mass spectra and transition regions

Thank you for the attention

References

[1] 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.

[2] Qu, Huilin, and Loukas Gouskos. "Jet tagging via particle clouds." *Physical Review D* 101.5 (2020): 056019.

[3] CMS Collaboration. "Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS. " *Journal of Instrumentation* 11.01 (2016): P01019.

[4] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017)

[5] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector." *Journal of Instrumentation* 12.10 (2017): P10003.

[6] CMS Collaboration, “Identification of hadronic tau lepton decays using a deep neural network,” *Journal of Instrumentation*, 17 (2022): P07023