PROCEEDINGS OF SCIENCE

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

Valentina Camagni[a,b] **for the CMS collaboration**

[a] *INFN, Sezione di Milano-Bicocca, Milano, Italy*

[b] *Dipartimento di Fisica, Università degli Studi di Milano-Bicocca Milano, Italy*

E-mail: valentina.camagni@cern.ch

A transformer-based encoder is proposed for the offline reconstruction of tau-lepton decays, with the goal of improving the estimation of the di-tau invariant mass. The model performs a regression of the transverse momentum of each tau candidate to recover the contribution carried away by neutrinos in the decay, while keeping η and ϕ fixed in accordance with the collinear approximation valid in the targeted phase space. The self-attention mechanism enables the network to combine information from the visible decay products together with event-level observables when inferring the tau kinematics. The performance is evaluated on samples with di-tau final states from Standard Model and beyond-the-Standard-Model processes and compared with the FastMTT algorithm, commonly used as a reference for di-tau mass reconstruction. The transformer-based method shows potential as an alternative technique for offline di-tau mass reconstruction at the CMS experiment.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

ISSN 1824-8039. Published by SISSA Medialab.

PoS(EPS-HEP2025)075

1. Introduction

Reconstructing the invariant mass of tau-lepton pairs ($m_{\tau\tau}$) plays a central role in precision Standard Model (SM) measurements and in searches for new resonances decaying to $\tau\tau$, as motivated in many beyond-the-Standard-Model scenarios (BSM).
The presence of neutrinos in tau decays leads to missing transverse momentum, which degrades the mass resolution and complicates the identification of resonant structures.

The CMS experiment currently relies on the Secondary Vertex Fit (SVFit) algorithm [1], a likelihood-based method that combines the visible decay products with the event's missing transverse energy (MET) to estimate the di-tau four-momentum. While SVFit provides good accuracy, its computational cost becomes significant in large-scale analyses or when evaluating systematic variations. The FastMTT algorithm [2] achieves a ∼100× speed-up by adopting analytically solvable approximations, which enable rapid likelihood evaluation but reduce flexibility across different kinematic regimes.

To investigate complementary approaches, this work explores the use of deep-learning techniques. The proposed method uses a transformer encoder [3] that regresses the transverse momentum of each tau candidate while keeping η and ϕ fixed, under the collinear approximation, to reconstruct the full tau kinematics. The predicted p_T can be used directly in an analysis, for instance in object selections, while the reconstructed di-tau mass is mainly used as a figure of merit to assess the quality of the regression. In the following, this architecture will be referred to as the Tau Pair Mass Transformer (TPMT). Two training strategies are considered:

- **Flat-mass setup:** to learn tau-kinematic features independently of any specific resonance, a generic spin-0 state is simulated and used to train the TPMT. Tau pairs are isotropically produced via gluon-gluon fusion (ggH) and vector-boson fusion (VBF), with a uniform invariant-mass spectrum between 30 and 300 GeV. The performance is evaluated on resonant processes: Higgs-boson (H) production via ggH ($m_H = 125$ GeV) and Drell–Yan (DY) events in the region of the Z boson.

- **Resonance-mass setup:** training and evaluation use resonant samples in two configurations:

  - **SM setup:** DY events with $m_{\tau\tau} > 50$ GeV and H production via ggH with $m_H = 125$ GeV are used as sources of *genuine* taus, defined as reconstructed HPS [4] candidates matched to a generator-level tau lepton and identified with PNet [5]. Top-quark pair production ($t\bar{t}$) is included to represent *misidentified* taus, i.e. reconstructed HPS+PNet τ candidates originating from jets or leptons that are misidentified as taus and therefore not matched to a generator-level tau lepton.
  - **SM+BSM setup:** the same samples are supplemented with a benchmark BSM signal consisting of a heavy neutral Higgs boson (H/A) with a mass of 350 GeV, motivated by supersymmetric (SUSY) Two-Higgs-Doublet Model (2HDM) scenarios. The H/A state is generated via ggF and b-associated production. This mass point is also included in the training and is used to evaluate the method's performance on non-SM resonances.

In the long term, this approach may be extended to parametrised frameworks capable of accommodating multiple signal masses within a unified model.

2. Pre-selection criteria and input features

This work considers two di-tau final states, i.e. pair types: the fully hadronic channel ($\tau_h\tau_h$) and the semi-leptonic muonic channel ($\tau_h\tau_\mu$). A consistent pre-selection is applied across all samples: in the $\tau_h\tau_h$ channel, events are required to contain two genuine hadronic tau candidates (τ_h) with opposite electric charge and $p_T > 20$ GeV; in the $\tau_h\tau_\mu$ channel, one genuine hadronic tau and one genuine muon from a tau decay (τ_μ) are required, also with opposite electric charge and $p_T > 20$ GeV. For $t\bar{t}$ events, misidentified taus are selected by requiring reconstructed tau candidates within $\Delta R < 0.4$ of top-quark decay products and not matched to generator-level taus.

The TPMT model uses a compact set of physics-motivated input features describing the kinematics, identification scores, decay-mode information, and track-level properties of reconstructed taus; the kinematic and particle-type properties of their visible decay products; the four-momentum of jets (restricted to those with $\Delta R > 0.4$ from the selected tau candidates); and the magnitude, direction, and covariance of the MET. This feature set is derived from a Random Forest–based feature-importance study; although classification-oriented, the most discriminating variables also provide the kinematic information needed for the regression.

3. TPMT architecture and training strategy

The TPMT is a transformer-based model designed to infer the kinematics of the two tau leptons from heterogeneous event-level inputs, including reconstructed taus, jets, MET, visible mass, and visible decay products. A schematic overview of the model architecture is shown in Figure 1. Each object type is encoded through a dedicated embedding MLP and processed by a transformer encoder with multi-head self-attention. An attention-pooling layer produces a global event representation, which is passed to two regression heads predicting $\log(p_T)$ for the two tau candidates. Generator-level p_T is used as target for genuine taus, while misidentified taus are assigned their reconstructed p_T, ensuring stable behaviour in these cases. The predicted kinematics are combined with the reconstructed η, ϕ, and mass to compute the TPMT di-tau mass.

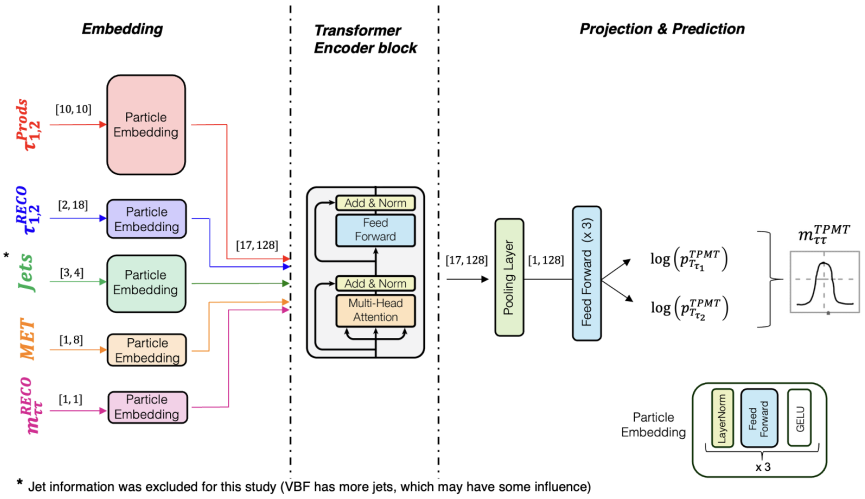


Figure 1: Schematic overview of the TPMT architecture.

Separate trainings are performed for each configuration (flat-mass and resonance-mass) and for the pair types defined within each of them. The flat-mass setup is trained only for the fully hadronic channel ($\tau_h\tau_h$), whereas the resonance-mass setup includes both the $\tau_h\tau_h$ and $\tau_h\tau_\mu$ pair types. Overall, the resonance-mass training uses about 3M simulated events, split evenly between the two di-tau final states, while the flat-mass setup relies on roughly 1M fully hadronic $\tau_h\tau_h$ events.

The loss function combines a per-tau p_T regression term and a MAE term on the reconstructed di-tau mass,

$$\mathcal{L}_{m_{\tau\tau}} = \mathrm{MAE}(m_{\tau\tau}^{\mathrm{TPMT}}, m_{\tau\tau}^{\mathrm{MC}}),$$

leading to a total loss

$$\mathcal{L} = \lambda_\tau \mathcal{L}_\tau + \lambda_{m_{\tau\tau}} \mathcal{L}_{m_{\tau\tau}}.$$

The mass term provides a global kinematic constraint that stabilises the regression and improves the discrimination between DY and Higgs processes.

4. Results

The performance of the TPMT model is evaluated separately for the two training strategies; for the flat-mass setup, trained exclusively on the $\tau_h\tau_h$ final state, the TPMT performance is compared with both SVFit and FastMTT. Figure 2 shows the reconstructed $m_{\tau\tau}$ distributions for Higgs ($H \to \tau\tau$) and Drell–Yan ($Z \to \tau\tau$) test samples. TPMT achieves a mass resolution comparable to SVFit, while offering a reduction of the per-event inference time from $O(1)$ s to $O(10^{-3})$ s, enabling fast and scalable event processing. Compared to FastMTT, TPMT and SVFit peak at consistent masses, while FastMTT shows a clear shift toward lower values.

In the resonance-mass setup, the comparison is restricted to FastMTT, focusing on the signal-over-background discrimination strategy described in Section 4.1. The corresponding reconstructed mass distributions for each setup and pair type are shown in Fig. 3, while Fig. 4 reports the corresponding relative residuals, illustrating the quality of the p_T correction for the two taus.

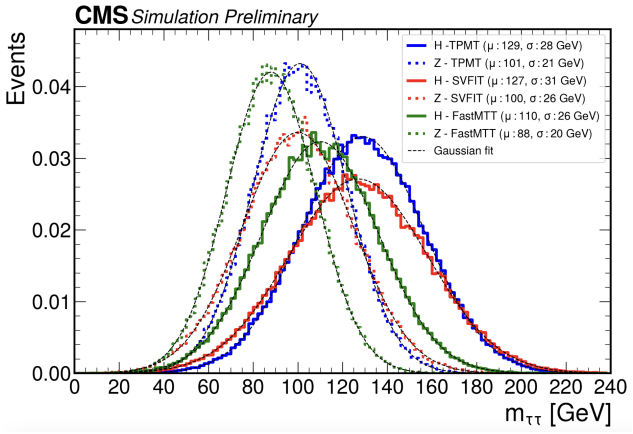
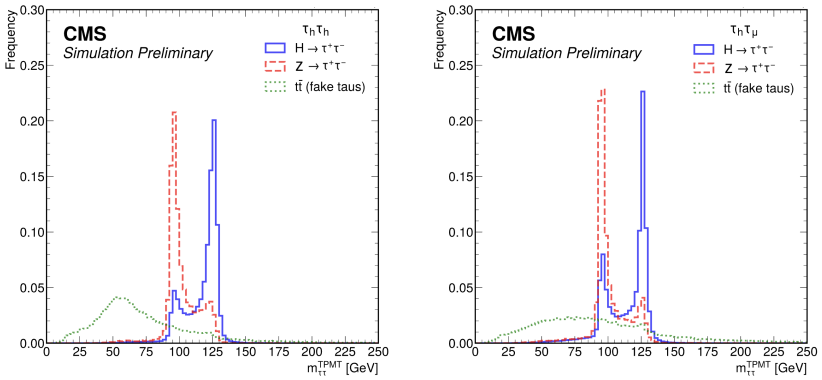
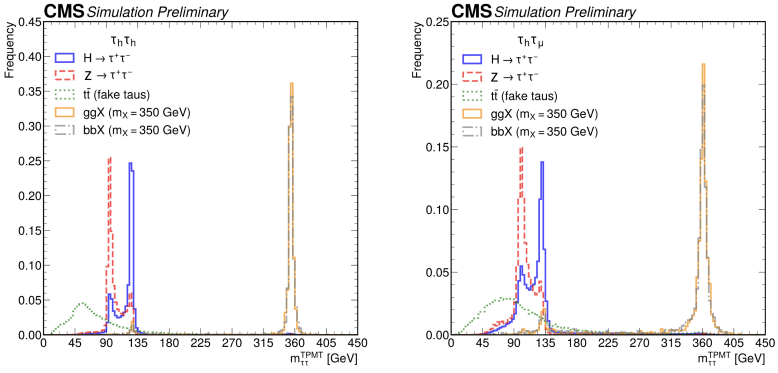


Figure 2: Reconstructed di-tau mass distributions for $H \to \tau\tau$ and $Z \to \tau\tau$ in the fully hadronic channel after TPMT flat-mass training. Solid curves show the H sample (blue: TPMT, green: FastMTT, red: SVFit), while dashed curves correspond to Z. Gaussian fits are overlaid with parameters reported in the legend.



(a) SM-only setup



(b) SM+BSM setup

Figure 3: Invariant mass distributions of the di-tau system after the resonance mass training in the SM-only (a) and SM+BSM (b) scenarios for $\tau_h\tau_h$ (left) and $\tau_h\tau_\mu$ (right) pair types.

In regions where Higgs and Drell–Yan events become kinematically similar, the regression exhibits secondary structures in the reconstructed mass spectrum. These arise from the overlap present in the training samples and divert part of the signal into masses away from the main peak, thereby altering the mass window required to achieve a fixed signal efficiency.

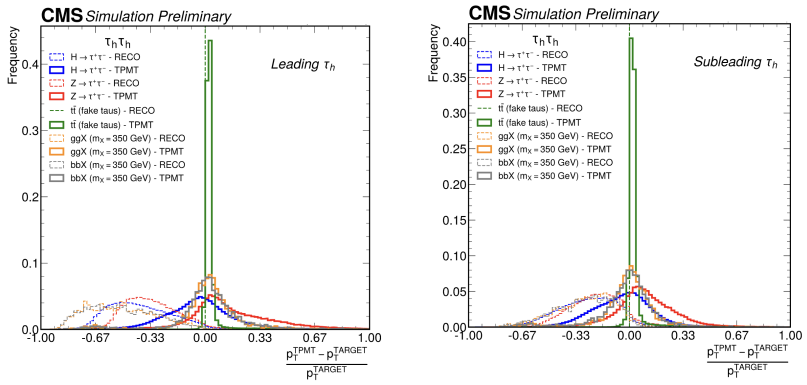
4.1 Signal efficiency and background rejection strategy

The discrimination performance of TPMT and FastMTT is evaluated by defining, for each algorithm, a signal window from a Gaussian fit to the reconstructed $H \to \tau\tau$ peak. Each window is chosen to retain a fixed signal efficiency (90%, or 80% in the BSM $\tau_\mu\tau_h$ case) and is determined separately for the two algorithms. Background rejection is then measured as the fraction of background events inside the corresponding window. The SM setup includes DY and $t\bar{t}$, while the BSM setup also includes a 350 GeV SUSY benchmark.

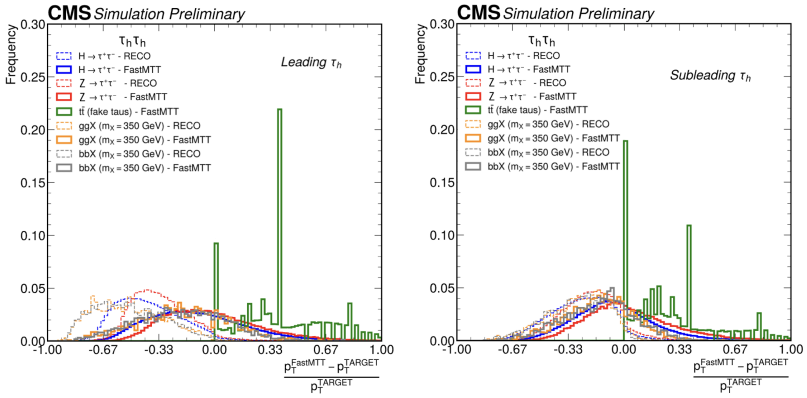
The results reported in Table 1 show that, across all scenarios and pair types, TPMT retains fewer DY and $t\bar{t}$ events than FastMTT at fixed signal efficiency, reflecting its narrower mass windows and better resolution. An exception occurs for the SUSY sample in the BSM $\tau_h\tau_\mu$ configuration, where a residual structure around 125 GeV increases the SUSY yield inside the TPMT window.

Table 1: Signal efficiency, selection window, and retained background percentage for each training, pair type, and algorithm.

Scenario	Pair Type	Algorithm	Eff. (%)	Window	Z (%)	$t\bar{t}$ (%)	SUSY (%)
SM	$\tau_h\tau_h$	TPMT	90	(97, 142)	57.31	14.97	–
		FastMTT	90	(73, 152)	83.12	52.81	–
	$\tau_h\tau_\mu$	TPMT	90	(94, 146)	78.63	29.98	–
		FastMTT	90	(78, 172)	95.66	49.58	–
BSM	$\tau_h\tau_h$	TPMT	90	(95, 140)	65.96	16.63	3.78
		FastMTT	90	(75, 153)	79.82	47.54	4.32
	$\tau_h\tau_\mu$	TPMT	80	(96, 144)	78.33	25.71	8.90
		FastMTT	80	(89, 161)	83.73	37.60	4.61



(a)



(b)

Figure 4: Relative residuals $\left(\frac{p_T^{\mathrm{TPMT}} - p_T^{\mathrm{TARGET}}}{p_T^{\mathrm{TARGET}}}\right)$ for the SM+BSM training scenario for the $\tau_h\tau_h$ pair type. (a) Dashed lines correspond to using the visible p_T, while solid lines use the TPMT prediction. (b) Dashed lines correspond to using the visible p_T, while solid lines use FastMTT. The RECO (fake tau) contribution is not shown, since the network is designed to leave fake candidates unchanged, resulting in a trivial residual distribution peaked at zero.

5. Conclusions and outlook

A transformer-based model for di-tau mass reconstruction has been presented. Two training strategies were explored. The flat-mass setup provides a model-independent solution applicable across analyses without assumptions on the production mechanism. The resonance-mass setup instead leverages physics-informed hypotheses—such as the kinematics of Z, H, or HH production—which allow the model to exploit information beyond the tau decay products. In this regime, full-event observables, including jets, supply useful context and help refine the proxy for the missing transverse momentum.

For the flat-mass training, TPMT achieves a modest improvement over SVFit in mass resolution while reducing the inference time from seconds to milliseconds. In the resonance-mass setup, TPMT reaches a latency comparable to FastMTT and provides higher background rejection at fixed signal efficiency, indicating improved separation power in challenging regions. In this context, the background is defined as all non-Higgs processes producing $H \to \tau\tau$ final states, including Drell–Yan, $t\bar{t}$, and the SUSY sample used to probe BSM-like kinematic configurations.
The reconstructed mass nevertheless shows a binary behaviour when Higgs and Drell–Yan kinematics overlap, producing secondary peaks that reflect the structure of the input features.

Future developments will focus on parametrised extensions of the model, enabling the resonance mass (or other process-dependent information) to be included as input. This is expected to improve interpolation across mass hypotheses and enhance robustness in regions where multiple processes overlap.

References

[1] L. Bianchini, J. Conway, E. K. Friis and C. Veelken, *Reconstruction of the Higgs mass in $H \to \tau\tau$ events by dynamical likelihood techniques*, *J. Phys. Conf. Ser.* **513** (2014) 022035, doi:10.1088/1742-6596/513/2/022035.

[2] W. Matyszkiewicz and A. Kalinowski, *Tau-pair invariant mass estimation using maximum likelihood and the collinear approximation*, Acta Phys. Polon. B Proc. Suppl. **18** (2025) 5-A21, doi:10.5506/APhysPolBSupp.18.5-A21.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Attention is All You Need*, Adv. Neural Inf. Process. Syst. **30** (2017).

[4] CMS Collaboration, "Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS," *Journal of Instrumentation* **11** (2016) P01019.

[5] H. Qu and L. Gouskos, "Jet tagging via particle clouds," *Phys. Rev. D* **101** (2020) 056019.