

Machine Learning driven improvements for the high-mass MSSM search in the di-tau final state with CMS

Motivation

BDT for Jet Misidentification Rate

Future Directions

Summary

References

[Irene Andreou](#)¹, David Colling, Lucas Russell, Klitos Savva, George Uttley, Daniel Winterbottom

¹irene.andreou@cern.ch

IOP Joint APP and HEPP Annual Conference 2025
April 8th, 2025

Motivation

BDT for Jet
Misidentification
Rate

Future
Directions

Summary

References

① Motivation

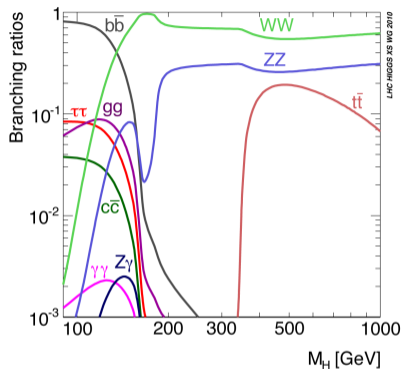
② BDT for Jet Misidentification Rate

③ Future Directions

④ Summary

$$\mathcal{L}_{\text{Yukawa}}^\tau = -m_\tau \bar{\tau} \tau - \frac{m_\tau}{v} H \bar{\tau} \tau$$

- High branching ratio [1]
- Many **Beyond Standard Model** scenarios predict the existence of additional Higgs bosons
- Additional Higgs bosons often have enhanced couplings to down-type fermions, such as taus, especially in models like the Type II MSSM [2]



Model	Up-type quarks	Down-type quarks	Leptons
Type I	Φ_2	Φ_2	Φ_2
Type II	Φ_2	Φ_1	Φ_1
Type X	Φ_2	Φ_2	Φ_1
Type Y	Φ_2	Φ_1	Φ_2

- Dominant backgrounds: **Jets misidentified as hadronic taus ($jet \rightarrow \tau_h$)** and genuine di- τ pairs
- Smaller backgrounds (diboson, $t\bar{t}$ and Drell-Yan) are estimated from simulations
- $Jet \rightarrow \tau_h$ backgrounds are estimated from a data-driven method: **Fake Factor method (F_F)**

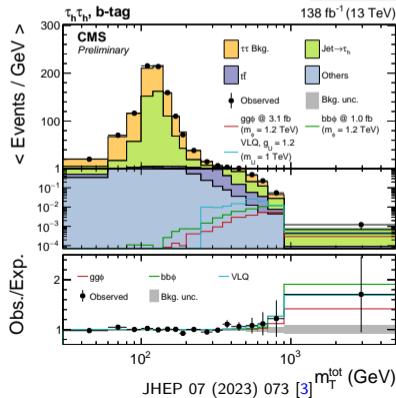
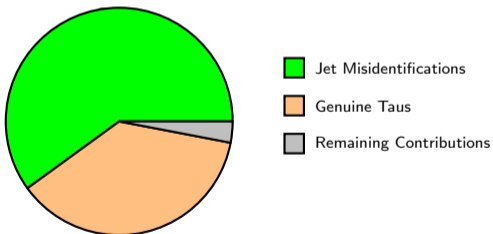
Motivation

BDT for Jet Misidentification Rate

Future Directions

Summary

References



Motivation

BDT for Jet
Misidentification
Rate

Future
Directions

Summary

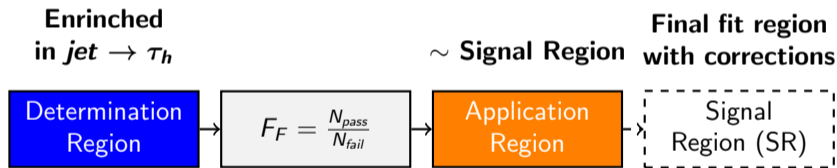
References

- Jet $\rightarrow \tau_h$ backgrounds are difficult to model in MC due to:
 - Small probability of a jet being misidentified as a $\tau_h \Rightarrow$ need high-statistics simulations
 - Poor description of jet $\rightarrow \tau_h$ misidentification rate

- Jet $\rightarrow \tau_h$ backgrounds are difficult to model in MC due to:
 - Small probability of a jet being misidentified as a $\tau_h \Rightarrow$ need high-statistics simulations
 - Poor description of jet $\rightarrow \tau_h$ misidentification rate
- Use a data-driven method: F_F method
 - **Determination Region (DR)**: Jet $\rightarrow \tau_h$ enriched control region (orthogonal to signal region)
 - **Application Region (AR)**: Same as signal region, but τ_h passes the loosest and fails the nominal tau identification score

$$F_F = \frac{N_{pass}}{N_{fail}} \text{ measured in DR and applied in AR}$$

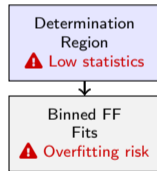
- Jet $\rightarrow \tau_h$ backgrounds are difficult to model in MC due to:
 - Small probability of a jet being misidentified as a $\tau_h \Rightarrow$ need high-statistics simulations
 - Poor description of jet $\rightarrow \tau_h$ misidentification rate
- Use a data-driven method: F_F method



- The fake factor is fitted in each bin of a few selected variables independently
- To account for residual mismodelling, **non-closure and extrapolation corrections** are applied to improve agreement in the SR

- **Method Limitations:**

- The number of parameters is fundamentally limited by the available statistics in the determination region
- In low-statistics bins, fake factor estimates can suffer from **large statistical fluctuations**, leading to poor modeling and potential overfitting
- Run 2 BSM $H \rightarrow \tau\tau$ analysis relied on a highly granular setup, requiring $\mathcal{O}(100)$ **fits and closure corrections** [3]

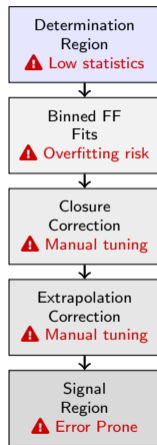


- **Method Limitations:**

- The number of parameters is fundamentally limited by the available statistics in the determination region
- In low-statistics bins, fake factor estimates can suffer from **large statistical fluctuations**, leading to poor modeling and potential overfitting
- Run 2 BSM $H \rightarrow \tau\tau$ analysis relied on a highly granular setup, requiring **$\mathcal{O}(100)$ fits and closure corrections** [3]

- **Workflow complexity:**

- The classical approach often involves **multiple reweighting stages**, each introducing additional sources of uncertainty and complexity
- These stages can make the overall workflow difficult to validate and maintain, increasing the risk of implementation errors or inconsistencies across regions and channels

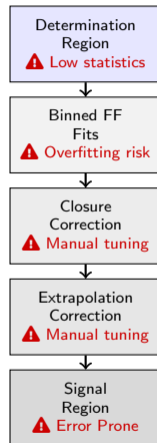


- **Method Limitations:**

- The number of parameters is fundamentally limited by the available statistics in the determination region
- In low-statistics bins, fake factor estimates can suffer from **large statistical fluctuations**, leading to poor modeling and potential overfitting
- Run 2 BSM $H \rightarrow \tau\tau$ analysis relied on a highly granular setup, requiring **$\mathcal{O}(100)$ fits and closure corrections** [3]

- **Workflow complexity:**

- The classical approach often involves **multiple reweighting stages**, each introducing additional sources of uncertainty and complexity
- These stages can make the overall workflow difficult to validate and maintain, increasing the risk of implementation errors or inconsistencies across regions and channels



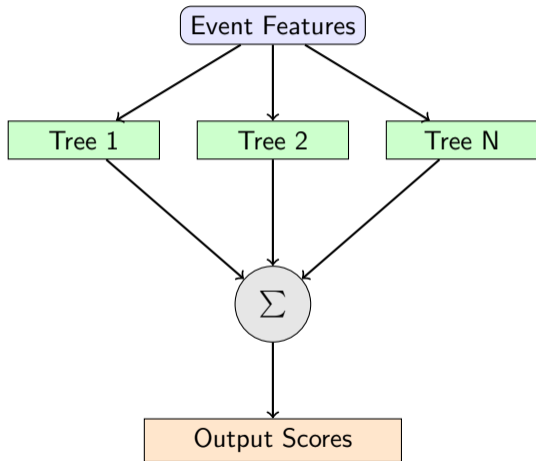
We propose a **machine learning reweighting approach** to overcome these issues

BDT Reweighter:

- Trained as a **classifier** to distinguish **isolated** vs. **anti-isolated** τ_h candidates
- **Input features** describe each τ_h candidate (e.g. $p_T, \eta, \phi \dots$)
- A **set of decision trees** is trained sequentially to classify events
- Each tree improves on the errors of the previous ones
- All tree outputs are summed to produce the final BDT score
- The BDT output is used to compute a reweighting factor per event:

$$\text{reweight} = \frac{p_{\text{isolated}}}{p_{\text{anti-isolated}}}$$

isolated: Pass Medium ID score
anti-isolated: Fail Medium and Pass VVLoose ID score



IMPERIAL The Solution: Machine Learning Reweighting

Motivation

BDT for Jet
Misidentification
Rate

Future
Directions

Summary

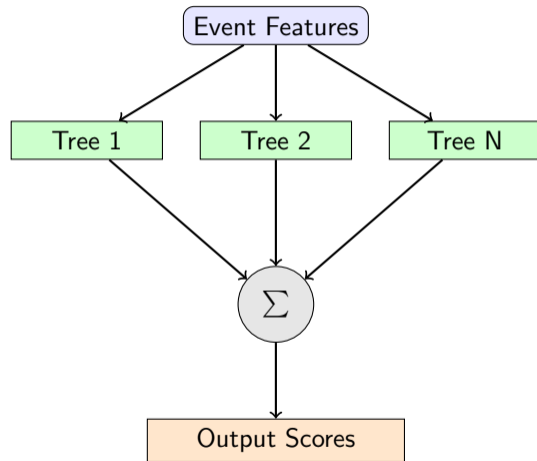
References

Advantages over traditional F_F :

- No binning needed — can use many variables
- Better modelling across years, channels, processes
- No need for separate sideband/extrapolation corrections
- Can simultaneously model various regions to improve statistics

$$\text{reweight} = \frac{p_{\text{isolated}}}{p_{\text{anti-isolated}}}$$

isolated: Pass Medium ID score
anti-isolated: Fail Medium and Pass VVLoose ID score



IMPERIAL BDT Reweighting

Motivation

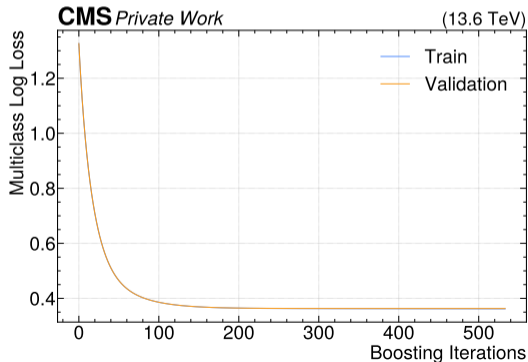
BDT for Jet
Misidentification
Rate

Future
Directions

Summary

References

- Early Run 3 (2022-2023) data were used to construct training, testing, and validation sets
- A BDT is trained to distinguish genuine vs. misidentified τ_h with information distinguishing between leading and sub-leading objects
- Local variables are only used for the corresponding τ_h
- The input variables include the decay mode τ_h , p_T , η , ϕ , and the ratio $p_T^{\text{jet}}/p_T^{\tau_h}$ of the τ_h candidate



IMPERIAL BDT Reweighting - Results

Motivation

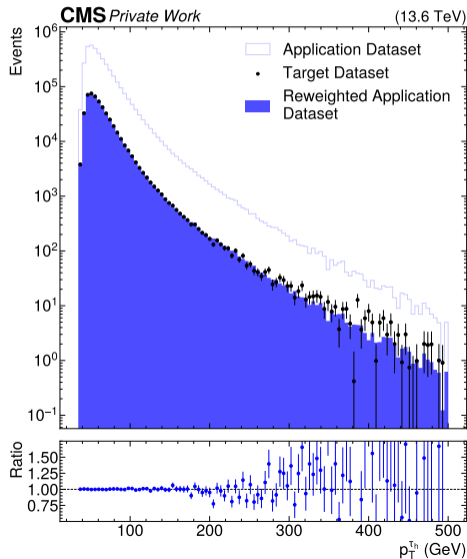
BDT for Jet
Misidentification
Rate

Future
Directions

Summary

References

- Control plots have been produced for all training input variables and key di-tau kinematic observables.
- Good closure observed in most bins



IMPERIAL BDT Reweighting - Uncertainties

Motivation

BDT for Jet
Misidentification
Rate

Future
Directions

Summary

References

Main sources of systematic uncertainty:

- 1 **Non-closure uncertainty:** Uncertainty accounting for non-closures in variables not used in the training
- 2 **Extrapolation uncertainty:** Uncertainty related to the fact that training and application datasets explore a different phase-space (same-sign \rightarrow opposite-sign)

 **Non-closure uncertainty**



 **Extrapolation uncertainty**

IMPERIAL BDT Reweighting - Extrapolation Uncertainty

Motivation

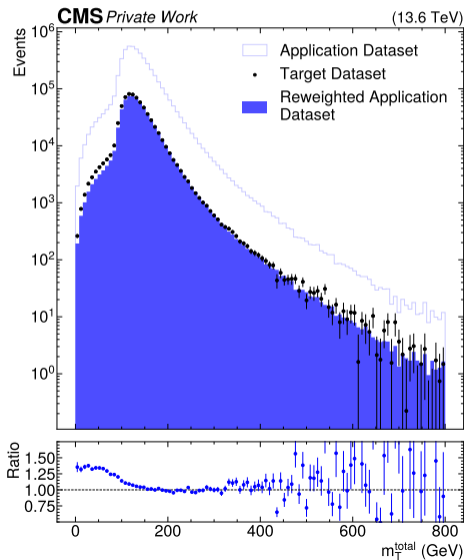
BDT for Jet
Misidentification
Rate

Future
Directions

Summary

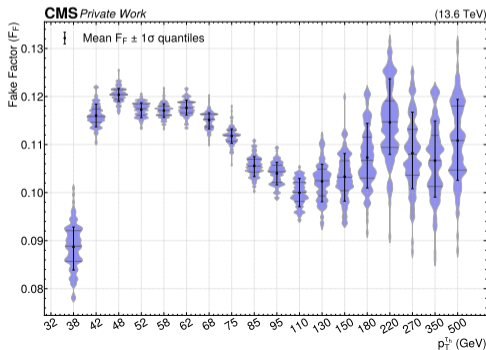
References

- The total transverse mass, m_T^{total} , is the key discriminant in the analysis (strong separation power S/B).
- This plot demonstrates that the reweighting provides a good modelling of m_T^{total} across most of the spectrum in the AR.



IMPERIAL BDT Reweighting - Statistical Uncertainty

- Employed bootstrapping to get an estimate of the statistical variation of the FFs calculated from the **optimised** BDT
- Multiple resampled datasets were generated by randomly sampling with replacement from the original training set
- The spread in fake factors across these resampled datasets provides an estimate of the statistical uncertainty



- The BDT reweighting method can be extended to other misidentification rates:
 - **Electron** $\rightarrow \tau_h$ misidentification rate
 - **Muon** $\rightarrow \tau_h$ misidentification rate
- These backgrounds are also poorly modeled in simulation and would benefit from a data-driven ML approach.
- In the long term, this method can contribute to:
 - ① Improved estimation of misidentified backgrounds in $\tau_h\tau_h$ analyses
 - ② More accurate derivation of τ_h **identification efficiency** scale factors
 - ③ Improved **energy scale corrections**, especially in regions dominated by misidentified backgrounds

A **unified ML-based framework** could standardize misidentification corrections across all τ_h misidentification sources

Motivation

BDT for Jet
Misidentification
Rate

Future
Directions

Summary

References

- Developed a **machine learning reweighting approach** using a BDT to model jet $\rightarrow \tau_h$ misidentification
- Introduced systematic approaches for the determination of statistical and systematic uncertainties
- Framework is **extendable to other misidentification rates** (e.g., $e \rightarrow \tau_h$, $\mu \rightarrow \tau_h$)
- Enables more accurate τ_h identification and energy scale corrections, with potential for wider ML-based improvements in new physics searches

Thank you!

Motivation

BDT for Jet
Misidentification
Rate

Future
Directions

Summary

References

Backup

References

- [1] LHC Higgs Cross Section Working Group (2012-2013). URL: <https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections>.
- [2] G. C. Branco et al. “Theory and phenomenology of two-Higgs-doublet models”. In: *Phys. Rept.* 516 (2012), pp. 1–102. DOI: 10.1016/j.physrep.2012.02.002. arXiv: 1106.0034 [hep-ph].
- [3] CMS Collaboration. “Searches for additional Higgs bosons and for vector leptoquarks in $\tau\tau$ final states in proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *JHEP* 07 (2023), p. 073. DOI: 10.1007/JHEP07(2023)073. arXiv: 2208.02717 [hep-ex].