



Search for Charged Higgs Bosons Decaying to a Neutral Higgs Boson with the ATLAS Detector at CERN

Adwait Meppurath

Supervisor: doc. Dr. André Sopczak

Master Thesis Presentation

September 29, 2025

UNIVERSITÉ
Clermont Auvergne

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

technische universität
dortmund

1. INTRODUCTION

- Theory, motivation and setting the premise

2. METHODOLOGY

- Quick overview of methods of analysis and tools used

3. ML DIAGNOSTICS

- Summary of classifier performance

4. DATA-DRIVEN FAKE ESTIMATION

- Results of template-fits to estimate fake-leptons in the background

5. ANALYSIS SENSITIVITY

- Summary plot of sensitivity of the analysis towards a potential signal

6. CONCLUSION AND OUTLOOK

- Overview of achieved results and immediate future extensions to this work

Introduction

Two-Higgs Doublet Model

- One of the simplest and most compelling extensions to the Higgs sector – cornerstone of the SUSY framework [1]
- 2HDM type II (the focus of this study) is the Higgs sector of the MSSM [1]
- 2HDM could potentially provide additional sources of CP violation in the weak interaction and potentially contribute to the baryon asymmetry in the universe [2], [3]
- Strong motivations to search for H^\pm under the 2HDM type II model.

Introduces a second scalar complex doublet to EW Lagrangian. Resulting in potential,

$$V(\Phi_1, \Phi_2) = m_{11}^2 \Phi_1^\dagger \Phi_1 + m_{22}^2 \Phi_2^\dagger \Phi_2 - [m_{12}^2 \Phi_1^\dagger \Phi_2 + h.c.] + \frac{1}{2} \lambda_1 (\Phi_1^\dagger \Phi_1)^2 + \frac{1}{2} \lambda_2 (\Phi_2^\dagger \Phi_2)^2 + \\ + \lambda_3 (\Phi_1^\dagger \Phi_1) (\Phi_2^\dagger \Phi_2) + \lambda_4 (\Phi_1^\dagger \Phi_2) (\Phi_2^\dagger \Phi_1) + [\frac{1}{2} \lambda_5 (\Phi_1^\dagger \Phi_2)^2 + h.c.] \quad (1)$$

Expansion around VEVs,

$$\Phi_{1,2} = \begin{pmatrix} \Phi_{1,2}^+ \\ (v_{1,2} + h_{1,2} + i\eta_{1,2})/\sqrt{2} \end{pmatrix} \quad (2)$$

8 degrees of freedom, **5 physical particles after EWSB.**

Two-Higgs Doublet Model

1. Two CP-even scalars – an “SM-like” h and a heavy scalar H
2. A CP-odd pseudoscalar A
3. Two charged Higgs bosons H^\pm

→

The focus of this study

Two-Higgs Doublet Model Type II

Four distinct **types** of 2HDM after Z_2 is imposed on the Lagrangian to avoid FCNCs. For **type II** (focus of this study), the Yukawa sector is,

$$\mathcal{L}_{H^\pm} = -H^+ \left(\frac{\sqrt{2}V_{ud}}{v} \bar{u}(m_u(\cot\beta)P_L - m_d(\tan\beta)P_R)d - \frac{\sqrt{2}m_l}{v}(\tan\beta)\bar{\nu}l_R \right) + h.c, \quad (3)$$

where $\tan\beta = \frac{v_2}{v_1}$ **and** $v = \sqrt{v_1^2 + v_2^2} = 246 \text{ GeV}$

\Rightarrow Couplings are proportional to the fermion masses. Decays to t, b, τ dominate, but the chosen parameter space ($\tan\beta = 20$) ensures $H^\pm \rightarrow W^\pm h$ **dominates** over $H^\pm \rightarrow tb$ [4]

Feynman Diagram : tbH^+ Production and $2l_{SS}1\tau$ Final State

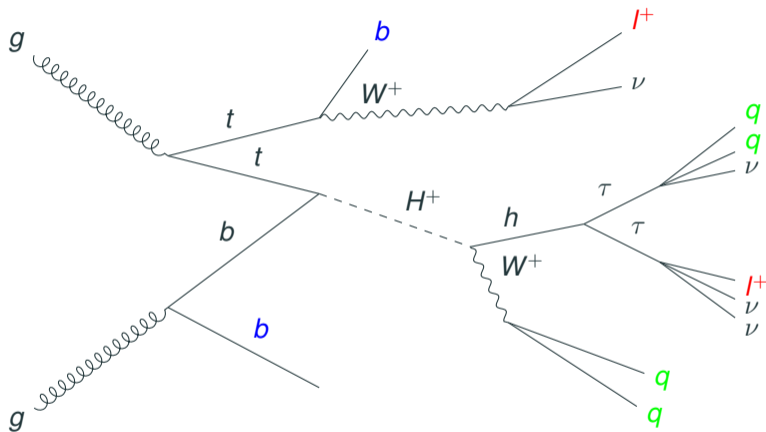


Figure 1: Associative H^+ production with t and b and subsequent decay to the $2l_{SS}1\tau_{had}$ final state

- Second complex scalar doublet added to EW Lagrangian
- EWSB now results in 5 physical particles, two of them are charged Higgs
- Decay channels of charged Higgs are identified for type II 2HDM and appropriate parameter space is chosen to study $H^+ \rightarrow W^+ h$.
- Complex, multi-jet final state with missing p_T requires good signal-background separation

Methodology

Fundamental goal : Test signal hypothesis with profile-likelihood fit

Fundamental goal : Test signal hypothesis with profile-likelihood fit

Statistical model : MC samples of SM backgrounds and multiple signal hypotheses characterized by the H^+ mass.

Fundamental goal : Test signal hypothesis with profile-likelihood fit

Statistical model : MC samples of SM backgrounds and multiple signal hypotheses characterized by the H^\pm mass.

What's done in this study : Expected sensitivity towards a potential signal **before** the model is tested on recorded ATLAS data (Blinded fit). Estimating signal strength in the background-only hypothesis.

Fundamental goal : Test signal hypothesis with profile-likelihood fit

Statistical model : MC samples of SM backgrounds and multiple signal hypotheses characterized by the H^\pm mass.

What's done in this study : Expected sensitivity towards a potential signal **before** the model is tested on recorded ATLAS data (Blinded fit). Estimating signal strength in the background-only hypothesis.

What's covered here : Producing signal MC samples, Custom features in datasets, Optimal classifier, Data-driven fake-lepton estimation

Monte Carlo Production with ATLAS Athena : Signal (tbH^+)

- Signal MC samples are produced with MadGraph5AMC@NLO, and showering is handled by Pythia8. PDF Base Fragment used : NNPDF30NLO in the 4 Flavour Scheme
- The BSM Model used for MC production is 2HDMTypeII
- In total 14 signal samples are used in this analysis, each characterized by the hypothesized H^+ mass.
- $H^+ \rightarrow Wh$ and $h \rightarrow \tau\tau$ are forced
- Events are filtered with cuts on light lepton p_T (20 GeV and 8 GeV)
- All NTuples used in this analysis are produced with TopCPToolkit v3 maintained by the Tau + X umbrella of the LPX Exotics Working Group at ATLAS.

Monte Carlo Production with ATLAS Athena : Background

Sample Name	DSIDs	Sample Name	DSIDs
$t\bar{t}$	410470	tZ	410560
$threeTop$	304014	$ttWW$	410081
Z -jets	700322, 700324, 700325, 700335, 700336, 700337	$fourTop$	412043
W -jets	700338, 700339, 700340, 700349, 700341, ..., 700348 (12 total)	WtZ	410408
VV	700589, 700591, 700592, 700593, 700594, 700603, 700604, 700605	$ttWZ$	500463
$SingleTop$	410659, 410644, 410645	$ttHH$	500460
VH	346645, 346646	$ttWH$	500461
ttZ	410276, 410277, 410278	$tt\gamma$	500800, 504554
ttW	700168	$V\gamma$	700398, 700399, 700400, 700401, ..., 700404 (7 total)
ttH	346343, 346344, 346345	$ttZZ$	500462
		tH	545796

- FastFrames is used to define custom features in the dataset, make pre-selections and define object qualities (for e, μ, τ and jets)
- Scale-based approach :
 - Two separate (low-scale and high-scale) XGBoost models are trained on the **Signal Region (SR)**
 - Low-Scale : 250,350,400GeV ; High-Scale : 500-3000GeV (11 datasets)
- Background contributions are mainly expected through fake leptons, hence
 - CRs are defined for fake leptons using **reconstructed variables**.
 - Simultaneous CR+SR fit (SR blinded) to recorded ATLAS data (Run 2) to obtain **norm factors** for fakes
- Expected limit on cross-section is obtained by performing a profile-likelihood fit for the background-only hypothesis on the signal probability distribution for each hypothesized mass point of H^+

Section Summary

- Statistical model for hypothesis testing is built with MC datasets of signal hypotheses and SM backgrounds
- A scale-based ML approach is employed for optimal S-B separation over the entire mass range.
- Data driven **template fits** are performed to correctly estimate fake contribution in the SR.
- Profile likelihood fit is performed on the signal probability variable, so that **better separation \Rightarrow stronger limits**

ML Diagnostics

XGB Training Curves : Low-Scale

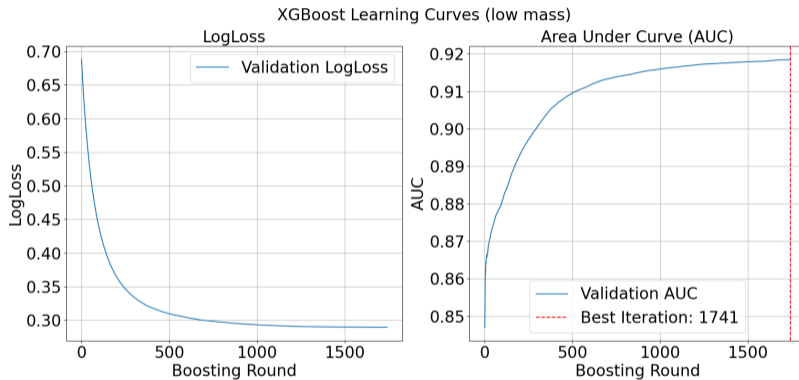


Figure 2: Training curves for XGB-low

XGB Training Curves : High-Scale

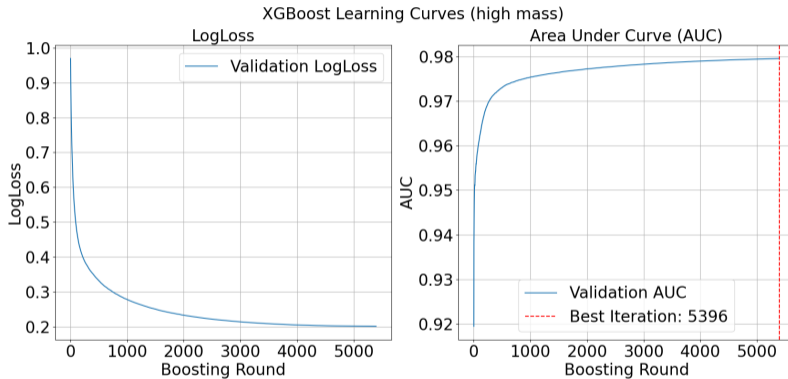


Figure 3: Training curves for XGB-high

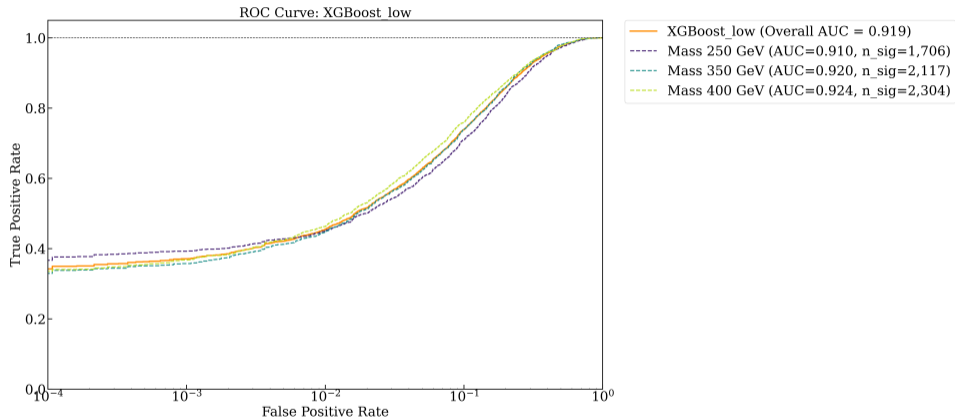


Figure 4: ROC for XGB-low

ROC : High-Scale

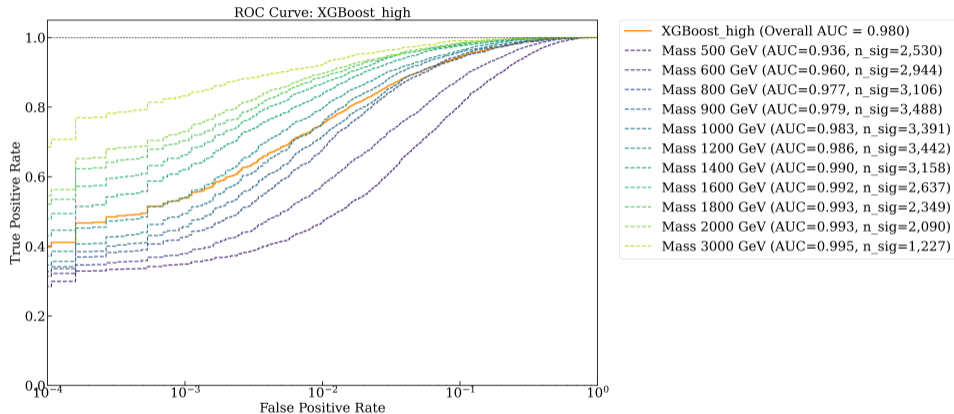


Figure 5: ROC for XGB-high

Precision, Recall and F1

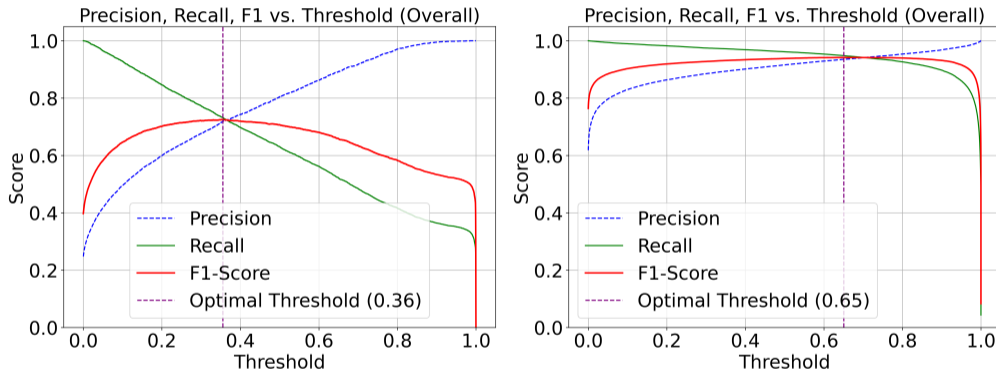


Figure 6: Optimal F1-based thresholds for XGB-low and XGB-high

Feature Importance Ranking

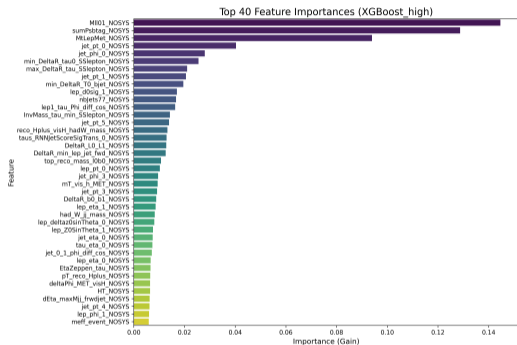
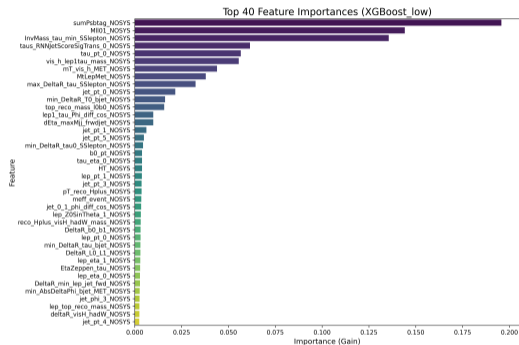


Figure 7: Feature importance ranking for XGB-low and XGB-high. Apart from the top two features, the ranking is quite distinct. Indicates scale-dependence of feature importance

Model Hyperparameters

- Both model hyperparameters were separately optimized with Optuna
- The hyperparameters are significantly different for both models, indicating mass-scale dependence of optimal model performance

Low-scale Model

n_estimators	= 6000,
learning_rate	= 0.0109,
max_depth	= 10,
subsample	= 0.6656,
colsample_bytree	= 0.8735,
gamma	= 0.0070,
reg_lambda	= 0.0095,
alpha	= 0.2164

High-scale Model

n_estimators	= 6000,
learning_rate	= 0.0157,
max_depth	= 8,
subsample	= 0.8577,
colsample_bytree	= 0.6873,
gamma	= 3.43e-05,
reg_lambda	= 3.95e-07,
alpha	= 0.0003

- Both ML models are able to classify S/B quite well (high AUC values).
- Scale-based approach could be beneficial as the optimized model structure for both models are different.
- The F1-based optimal threshold could be used to obtain/remove region with high signal purity.

Data-driven Fake Estimation

Both the signal and background samples are **re-classified** into either **prompt** or **non-prompt** samples, which fall under one of the following categories :

1. **Fake e** : EXACTLY one fake electron. Norm : $\lambda(e_f)$
2. **Fake μ** : EXACTLY one fake muon. Norm : $\lambda(\mu_f)$
3. **Fake $l + \tau$** : EXACTLY one fake lepton and fake tau. Norm : $\lambda(multi - fake)$
4. **Double Fake**: Both light leptons are fake. Norm : $\lambda(multi - fake)$
5. **Fake τ** : EXACTLY one fake tauon. Norm : $\lambda(\tau_f)$
6. **Triple Fake**: All three leptons are fake. Norm : $\lambda(\tau_f)$

Each event can only be part of ONE of these categories

SR after Re-classification

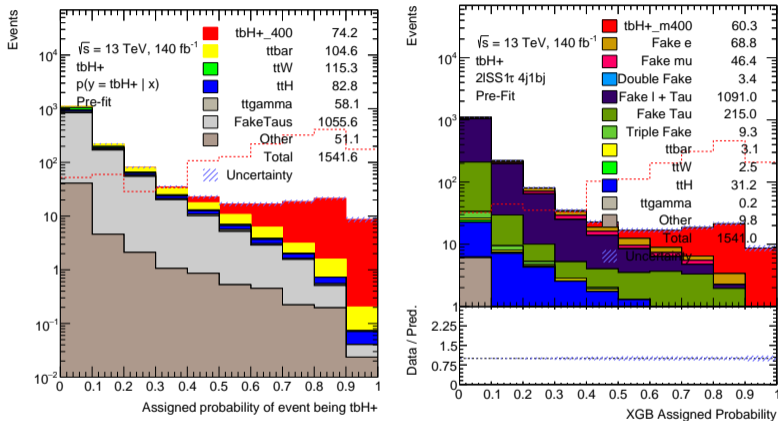


Figure 8: Re-classification of samples into prompt and non-prompt categories. Total number of events remain the same

Control Region : Fake τ and Triple-Fake

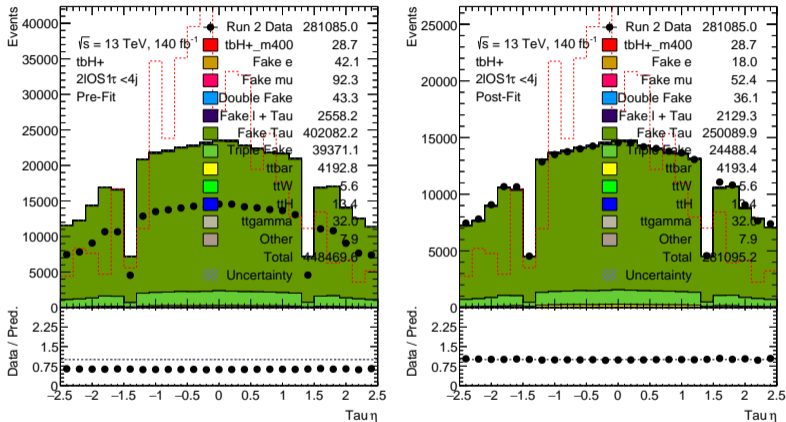


Figure 9: Pre-fit and Post-fit η_τ distribution in the CR for *Fake τ* and *Triple Fake*; Norm-Factor $\lambda(\tau_f)$

Control Region : Fake $l + \tau$ and Double-Fake

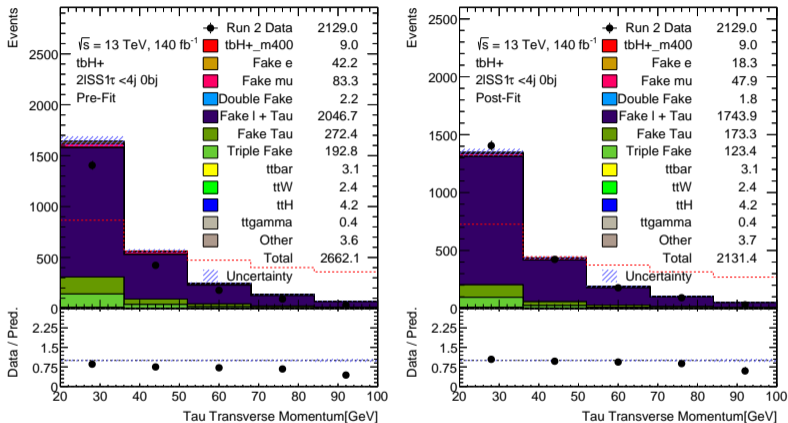


Figure 10: Pre-fit and Post-fit $p_{T,\tau}$ distribution in the CR for *Fake $l + \tau$* and *Double Fake*; Norm-Factor $\lambda(multi - fake)$

Control Region : Fake e

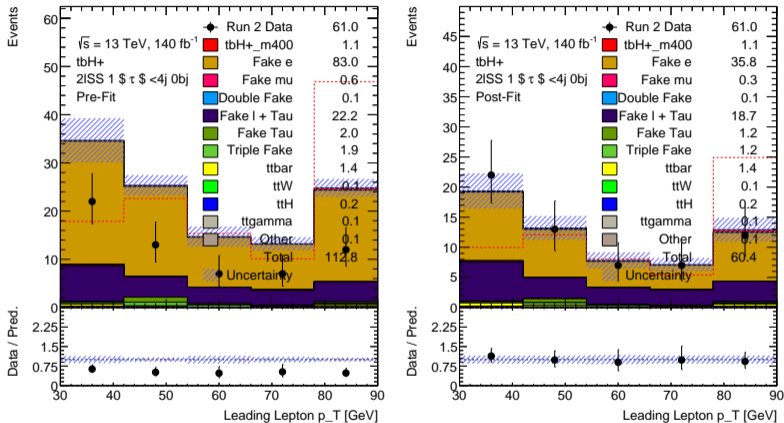


Figure 11: Pre-fit and Post-fit lepton p_T distribution in the CR for *Fake e* ; Norm-Factor $\lambda(e_f)$

Control Region : Fake μ

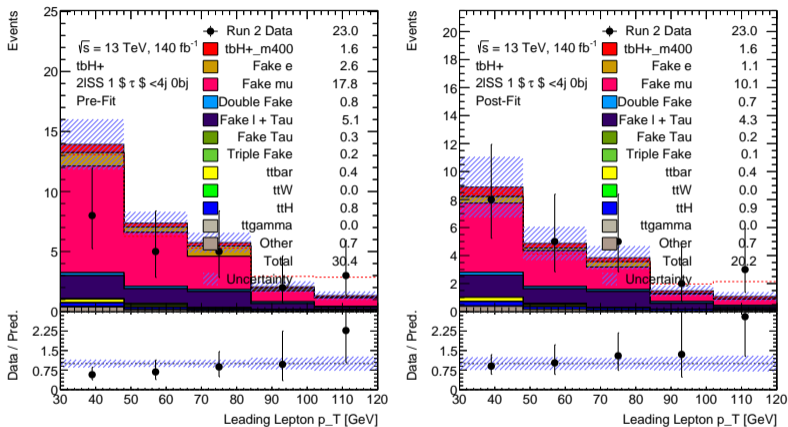


Figure 12: Pre-fit and Post-fit lepton p_T distribution in the CR for *Fake μ* ; Norm-Factor $\lambda(\mu_f)$

Correlation and Estimated Norms

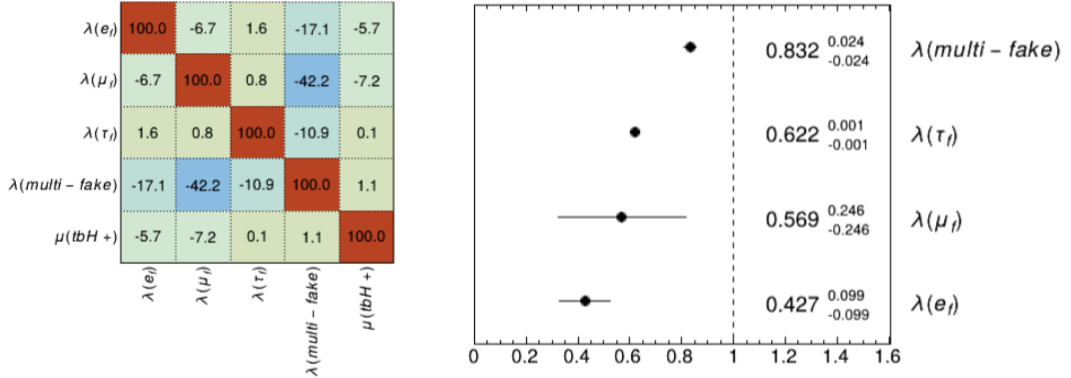


Figure 13: The correlation matrix between norm-factors and the estimated norm-factor after template fit (for $m_{H^+} = 400$ GeV)

- Possible fake contribution categories identified for $2l_{SS}1\tau$ final state
- Dedicated, orthogonal CRs created for 4 norm-factors
- Norm factors estimated after fit to recorded ATLAS data are stable and show minimal correlation with each other (except for $\lambda(\mu_f)$)

Sensitivity Analysis

Estimated Upper Limit on Cross-section

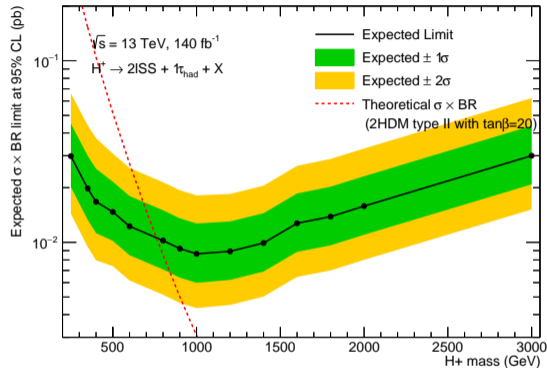


Figure 14: Expected 95% CL upper limit on the cross-section for all mass hypotheses. The solid black line shows the median expected limit, with the green and yellow bands representing the $\pm 1\sigma$ and $\pm 2\sigma$ uncertainties. The red dotted line is the theoretical prediction for the signal.

Previous Results : Rel. 21 vs Rel. 22

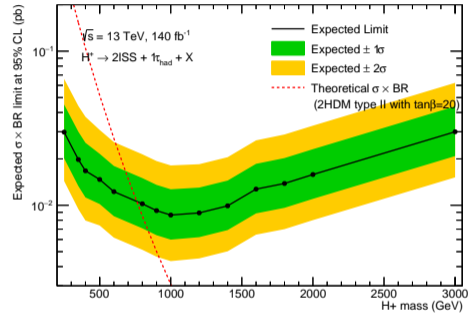
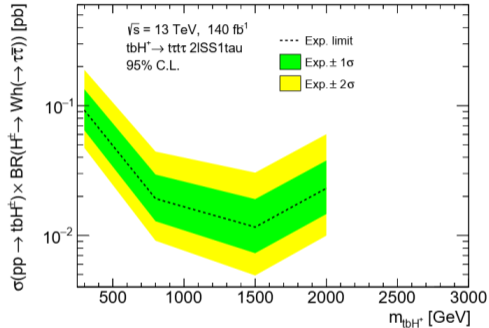


Figure 15: Comparison of expected upper limits on cross section between previous analysis in Rel. 21 [5] and this analysis (Rel. 22)

Previous Results : CMS Observed Limits

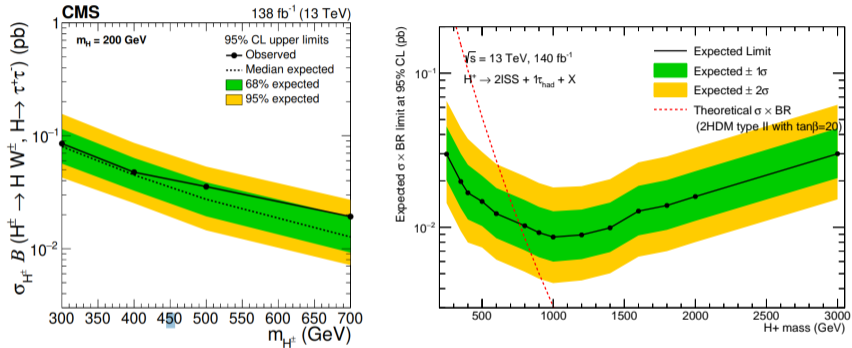


Figure 16: Comparison of observed upper limit on cross-section from CMS [6] and expected upper limit estimated in this analysis







Conclusion and Outlook

- Full analysis chain conducted on 14 signal hypotheses of tbH^+ (2HDM type II; $\tan\beta = 20$) from MC production to expected upper limit on cross-section.
- 70+ custom feature definitions added to the dataset, mostly specific to $2l_{SS}1\tau$ final state
- Two optimized XGB models trained separately for S-B discrimination
- Data-driven norm-factors for fake leptons determined with dedicated CRs
- Expected upper limits on cross-section point towards the **most sensitive exclusion of the charged Higgs till date**
- Internal Note : (**ANA-HMBS-2024-42 : H^+ to Wh , h to $\tau\tau$**)

Next Steps

- Include systematic uncertainties
- Unblind signal region
- Estimate observed upper limits
- Aiming for publication

References

-  H. Haber and G. Kane, “The search for supersymmetry: Probing physics beyond the standard model,” *Physics Reports*, vol. 117, no. 2, pp. 75–263, 1985.
-  A. Tranberg and B. Wu, “Cold Electroweak Baryogenesis in the Two Higgs-Doublet Model,” *JHEP*, vol. 07, p. 087, 2012.
-  S. Iguro and Y. Omura, “The direct CP violation in a general two Higgs doublet model,” *JHEP*, vol. 08, p. 098, 2019.
-  C. Degrande, R. Frederix, V. Hirschi, M. Ubiali, M. Wiesemann, and M. Zaro, “Accurate predictions for charged Higgs production: Closing the $m_{H^\pm} \sim m_t$ window,” *Phys. Lett. B*, vol. 772, pp. 87–92, 2017.
-  M. Rames, “Search for $tbh^+(\tau\tau)$ with performance optimisation for signal and background separation using machine learning with atlas data,” Master’s thesis, CTU, 2023.
-  CMS Collaboration, “Search for a charged higgs boson decaying into a heavy neutral higgs boson and a w boson in proton-proton collisions at $\sqrt{s} = 13$ tev,” *Journal of High Energy Physics*, vol. 2023, no. 9.

Thank you for your attention!

Backup Slides

Table 1: DSIDs and hypothesized H^+ mass points

DSID	H^+ Mass (GeV)	DSID	H^+ Mass (GeV)
512185	250	567613	900
512186	3000	567614	1000
512187	800	567615	1200
567608	350	567616	1400
567609	400	567617	1600
567610	500	567618	1800
567611	600	567619	2000

Object Definition & Event Selection

WP: wpSet0

Electron: el_select_TightLH_Loose_VarRad_NOSYS && el_AmbiguityType == 0 &&
abs(el_eta) <= 2.5

Muon: mu_select_Medium_NonIso_NOSYS && mu_select_Loose_Tight_VarRad_NOSYS &&
abs(mu_eta) <= 2.5

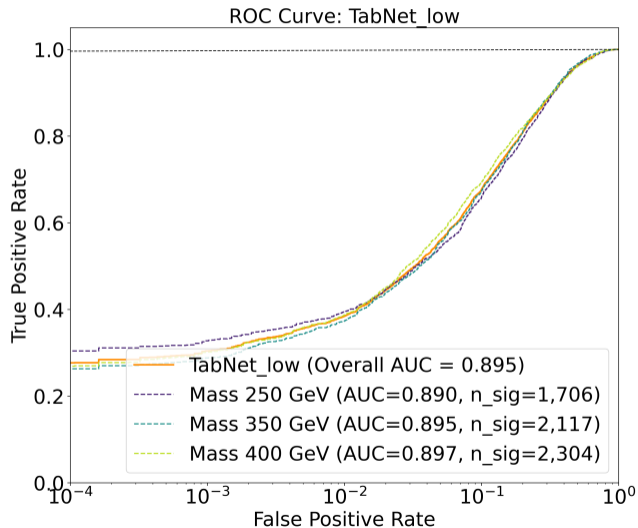
Tau:

Jet: abs(jet_eta) < 2.5

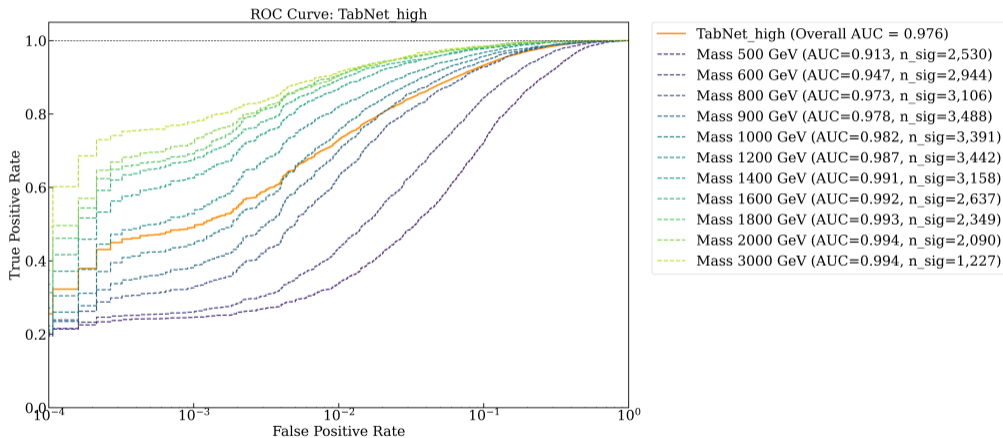
FastFrames: Exactly 2 Leptons, 1 Tau, At least 1 Jet, Leading Lepton $p_T > 25\text{GeV}$,
Sub-leading Lepton $p_T > 10\text{GeV}$

TRExFitter: Same-Sign Condition on Leptons, At least 1 b-jet ($\text{nbJets77} > 0$), At least 4
jets – $2l_{SS}1\tau4j1b$

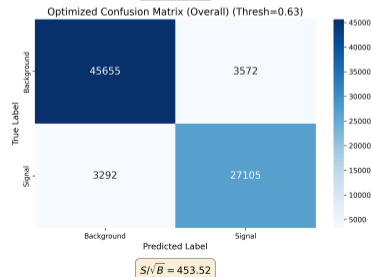
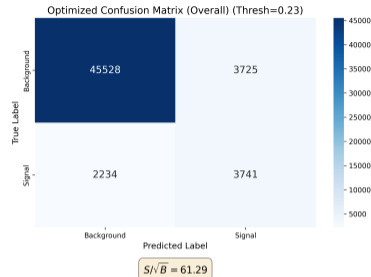
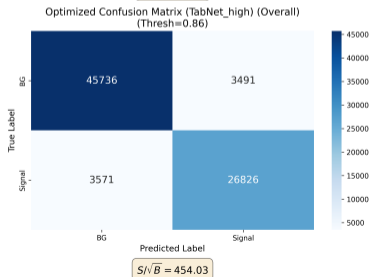
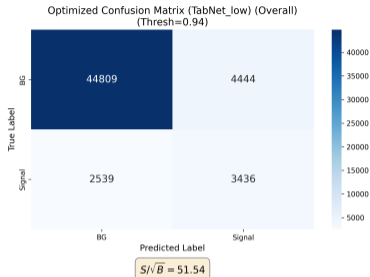
TabNet ROC : Low-Scale



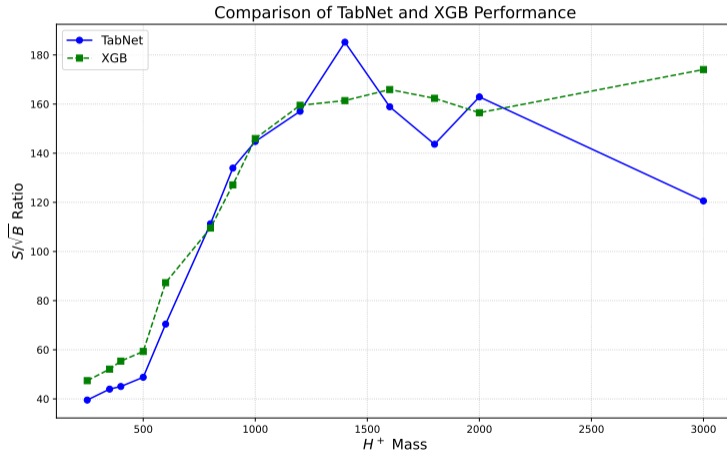
TabNet ROC : High-Scale



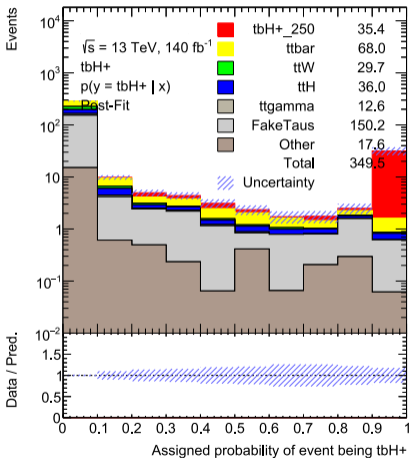
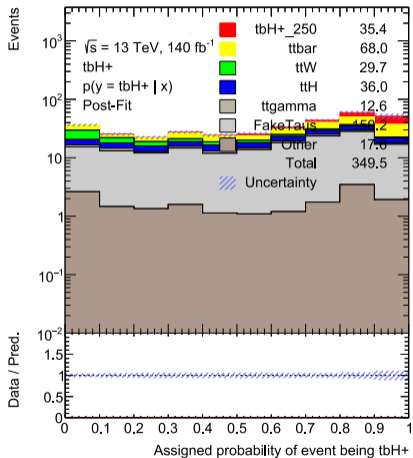
Confusion Matrices



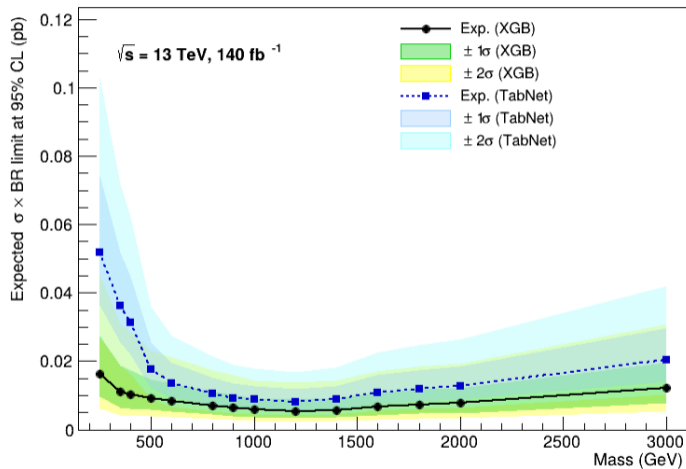
S/\sqrt{B} Comparison



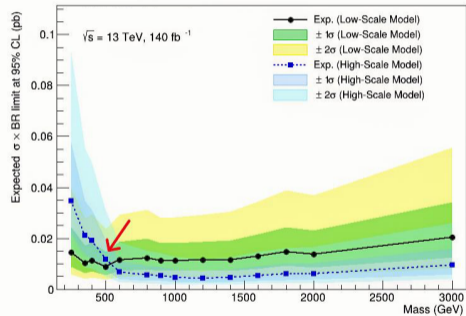
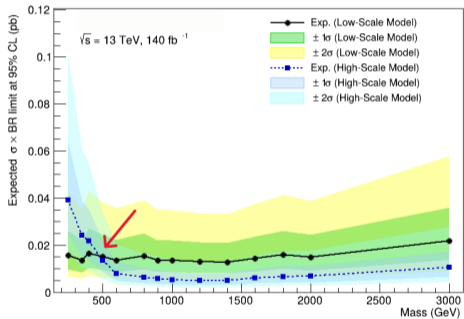
Probability Distribution : TabNet v XGB



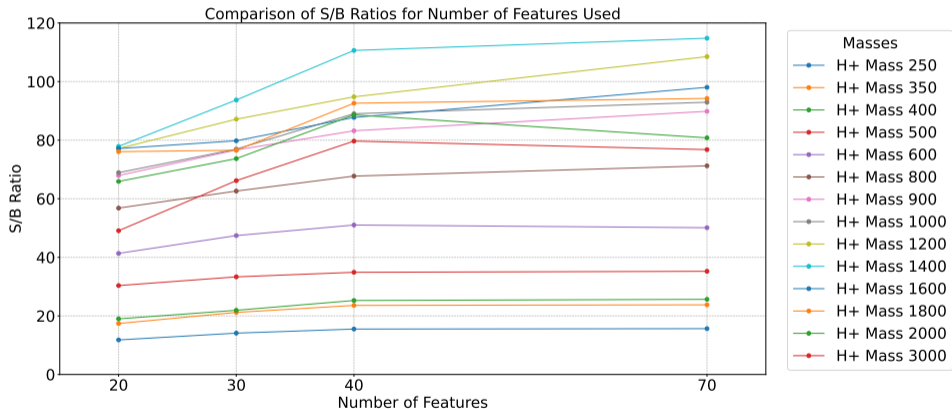
Limit Estimates : TabNet v XGB



Optuna Optimization : XGB-low and XGB-high



N Features Usage



Fake Categories

Fake Category	Selection Logic
Fake e	<code>((lep_flav_0_NOSYS == 0 && XXX_OnlyLeadFake) (lep_flav_1_NOSYS == 0 && XXX_OnlySubLFake))) && HadTau_truth_NOSYS</code>
Fake μ	<code>((lep_flav_0_NOSYS == 1 && XXX_OnlyLeadFake) (lep_flav_1_NOSYS == 1 && XXX_OnlySubLFake))) && HadTau_truth_NOSYS</code>
Double Fake	<code>!XXX_Prompt_0 && !XXX_Prompt_1 && HadTau_truth_NOSYS</code>
Fake $/ + \tau$	<code>(XXX_OnlyLeadFake XXX_OnlySubLFake) && !HadTau_truth_NOSYS</code>
Fake τ	<code>(!HadTau_truth_NOSYS && XXX_BothPrompt)</code>
Triple Fake	<code>!XXX_Prompt_0 && !XXX_Prompt_1 && !HadTau_truth_NOSYS</code>

Table 2: Boolean flags used to define different categories of fake leptons and taus

Flag Name	Boolean Logic Definition
XXX_Prompt_0	<code>(leps_IFFtype_0_NOSYS == 2 leps_IFFtype_0_NOSYS == 4 leps_IFFtype_0_NOSYS == 7)</code>
XXX_Prompt_1	<code>(leps_IFFtype_1_NOSYS == 2 leps_IFFtype_1_NOSYS == 4 leps_IFFtype_1_NOSYS == 7)</code>
XXX_BothPrompt	<code>(XXX_Prompt_0 && XXX_Prompt_1)</code>
XXX_OnlyLeadFake	<code>(!XXX_Prompt_0 && XXX_Prompt_1)</code>
XXX_OnlySubLFake	<code>(!XXX_Prompt_1 && XXX_Prompt_0)</code>

Table 3: Prompt recognition using IFFtool-based variables¹

¹[TruthClassification repo](#)