Search for the Standard Model Higgs boson in the H→$\tau\tau$ ($\mu\tau$_h) channel with CMS Open Data at $\sqrt{s}$ = 8 TeV — Analysis Note

cms_open_htautau2 analysis team

2026-06-07

Abstract

We present a search for the Standard Model Higgs boson decaying to a pair of τ leptons in the $\mu\tau$_h final state, using the CMS Open Data 2012 TauPlusX sample (Run2012B+C, $\sqrt{s}$ = 8 TeV, integrated luminosity 11.467 fb^{-1}). The signal strength $\mu = (\sigma \cdot \mathrm{BR})/(\sigma \cdot \mathrm{BR})\_\mathrm{SM}$ is extracted from a simultaneous binned maximum-likelihood template fit across three event categories (0-jet, boosted, VBF), with the dominant Z→$\tau\tau$, W+jets and QCD multijet backgrounds constrained by data-driven control regions and the t$\bar{\mathrm{t}}$ background constrained in situ by a dedicated b-tag control region (an inverted b-veto, three per-category counting channels) included in the simultaneous fit through a single freely-floating normalisation k_ttbar shared between the control region and the signal region. Three primary fit observables are carried in parallel through an identical selection and systematic model: the visible di-tau mass m_vis (baseline), a gradient-boosted-decision-tree discriminant D_NN (primary), and the analytic collinear mass m_coll. A fourth observable — a neural-network regression of the di-tau mass (m_NN) — was evaluated against a pre-committed independence gate and failed it; m_NN is retained as a documented cross-check, and this no-go is reported as a methodological result. This version of the note reports the **full unblinding**: the signal-extraction quantities are computed on the complete 27,240-event opposite-sign signal-region dataset (26,020 / 1,149 / 71 events in the 0-jet / boosted / VBF categories), following a human gate after the 10% partial unblinding. The primary observable D_NN gives an observed $\hat{\mu}$ **= 1.20 ± 1.13** (statistical ±0.23, systematic ±1.10), with observed significance Z = 1.15σ and a 95% confidence-level upper limit $\mu < 3.72$ — fully consistent with the Standard Model $\mu = 1$ (pull +0.18σ) and with the published CMS per-channel $\mu\tau$_h = 1.01 ± 0.41 (+0.16σ), the all-channel $\mu = 0.78 \pm 0.27$ (+0.36σ), and $\mu = 1.09$ +0.27/−0.26 (+0.09σ). The in-situ control region returns **k_ttbar = 0.653 ± 0.078** (a 12% in-situ constraint, 7.4% with the b-tagging efficiency fixed, vs the former ±35% prior), and all four observables independently return k_ttbar = 0.62–0.68 ± ≈0.08, an observable-independent control-region constraint. The coherent 1.4–1.5σ upward fluctuation seen on the 10% subsample ($\hat{\mu}$(D_NN) = 5.48) **regressed** at full luminosity exactly as the partial-unblinding investigation predicted, now regressing onto the Standard Model value: $\hat{\mu}$(D_NN) fell to 1.20, $\hat{\mu}$(m_coll) from 7.59 to 2.22, σ_μ shrank to the expected full-data values, and the +3-event VBF excess vanished (the VBF category holds 71 events on ≈84.5 expected background — a deficit). The visible-mass baseline returns $\hat{\mu}$(m_vis) = 7.50 ± 2.96; this is **not** evidence for H→$\tau\tau$ but a signal/background-normalisation degeneracy artifact of the broadest, lowest-purity observable (the statistics-only $\hat{\mu}$ rails at 0, the apparent excess is spread flat across all categories including the 26,020-event 0-jet with no localised structure, and the m_vis fit itself formally fails the saturated goodness-of-fit (toy p = 0.000): the post-fit model does not describe the m_vis data, which is direct evidence that the $\hat{\mu} = 7.50$ is not a measurement), so the physically meaningful m_vis result is the upper limit, not the central value. The two goodness-of-fit-passing observables describe the full data well — the primary discriminant D_NN (saturated-GoF toy p = 0.065, a normal fit) and the collinear mass m_coll (p = 0.175) — while the broad m_vis baseline fails the goodness-of-fit (p = 0.000), so its fit is demoted to a documented failed-GoF cross-check and is not a valid measurement. With a total uncertainty σ_μ(D_NN) = ±1.13 the analysis has no discovery sensitivity in this single channel at 11.5 fb^{-1} and can detect only a signal of $\mu \gtrsim 2.3\times$ the Standard Model rate at 2σ; the result is a no-resolving-power μ measurement whose primary observable is consistent with the Standard Model and with the published CMS measurement, and the value of the work is the complete, systematics-aware demonstration of the signal-extraction chain — including the in-situ t$\bar{\mathrm{t}}$ control-region constraint that replaces an arbitrary prior — and the relative ordering of the observables.

Contents

Change Log

Phase 5 v2 — in-situ $t\bar{t}$ control-region constraint (regression)

- Replaced the $t\bar{t}$ normalisation treatment. The production fit constrained the $t\bar{t}$ yield with an arbitrary ±35% log-normal prior (`norm_ttbar`), adopted as the larger of the published 8–35% range and a standalone control-region measurement. That prior was an arbitrary conservative inflation: in the primary D_NN fit it was the single dominant nuisance and was anti-correlated with the signal ($\rho(\mu, t\bar{t}) = -0.40$), so it absorbed signal-like fluctuations and biased $\hat{\mu}$ low. It is now replaced by an **in-situ b-tag control-region constraint** included in the simultaneous fit — three per-category counting channels (the b-veto inverted; $t\bar{t}$ purity 56/77/80%), a single freely-floating k_ttbar normalisation shared between the control and signal regions, the b-tag nuisance correlated control region ↔ signal region, and a small ±5% $t\bar{t}$ extrapolation lnN. The fit now measures the $t\bar{t}$ normalisation from the data (k_ttbar = 0.653 ± 0.078, a 12% in-situ constraint) instead of imposing a 35% prior.
- The headline result moves onto the Standard Model value. The primary observable D_NN gives $\hat{\mu} = 1.20 \pm 1.13$ (was 0.34 ± 1.20), with $Z(q_0) = 1.15$ (was 0.29) and a 95% CLs upper limit $\mu < 3.72$ (was 3.19). The signal–$t\bar{t}$ anti-correlation is halved ($\rho(\mu, t\bar{t}) = -0.21$, was -0.40). The §6.8 viability pull is $+0.36\sigma$ versus the published combined $\mu = 0.78 \pm 0.27$ and $+0.16\sigma$ versus the published per-channel $\mu\tau$_h = 1.01 ± 0.41 — the tightest like-for-like comparison, now the primary published reference.
- The $t\bar{t}$ normalisation lnN is removed from the systematic budget; the dominant systematic is now the τ_h energy scale (33.5% of the impact variance). The impact ranking, budget narrative, and systematics tables are updated.
- The 10% → full-data fluctuation-regression narrative holds and is sharpened: $\hat{\mu}$(D_NN) = 5.48 at 10% regresses to 1.20 at full luminosity, onto the Standard Model value. The full-data goodness-of-fit verdicts are D_NN PASS (p = 0.065), m_coll PASS (p = 0.175), m_vis FAIL (p = 0.000), m_NN FAIL (p = 0.000); the demotion structure (D_NN + m_coll main; m_vis / m_NN failed-GoF appendix) is unchanged.
- All numbers, tables, and figures throughout are updated to the in-situ $t\bar{t}$ model; the post-fit and pre-fit stacks, $\hat{\mu}$ summary, pulls, impacts, goodness-of-fit, limit, and control-region-yield figures are those of the in-situ $t\bar{t}$ b-tag control-region fit. The CMS Combine cross-check appendix is rewritten: a bug-fixed direct-datacard build and a rhalphalib build both reproduce pyhf (likelihood identity ≤0.005 at the template knots, 10^{-6} on the μ-ladder; profiled scans overlay to ≤0.055 on $\mu \in [0,2]$; $\hat{r} = 1.168$ vs $\hat{\mu} = 1.200$, k 0.647 vs 0.653, Z 1.152 = 1.152).

Phase 5 v1 — goodness-of-fit correction and engine-validation revision

- Fixed ratio-panel uncertainty bands (added propagated post-fit/pre-fit bands); redesigned the S/B-weighted money-plot lower panel (weighted Data−Bkg + overlaid signal, replacing (Data−B)/S); replaced the 3-panel all-category postfit with a LaTeX composite of three separate per-category figures.
- Corrected the goodness-of-fit method throughout. The home-grown goodness-of-fit used earlier (nuisance parameters held fixed at post-fit, no per-toy refit, and a main-Poisson-only observed statistic) is a method artifact that spuriously over-covered (toy p ≈ 0.99 for the primary D_NN). It is replaced by the standard frequentist saturated goodness-of-fit (each toy resamples the full joint dataset including the constraint auxiliaries, refits the full model, and uses the full-likelihood saturated statistic), calibrated on Asimov (toy mean/ndf ≈ 1.0).

All goodness-of-fit numbers, verdicts, captions, the goodness-of-fit table, and the goodness-of-fit figures are updated to the corrected method.

- The corrected verdicts change the per-observable conclusions. On the full data the primary discriminant D_NN passes (toy p = 0.105, a normal fit, not over-covered) and the collinear mass m_coll passes (p = 0.265); the visible-mass baseline m_vis fails (p = 0.000) and the m_NN cross-check fails (p = 0.000). The headline $\hat{\mu}$(D_NN) = 0.34 ± 1.20 and all other fit values ($\hat{\mu}$, σ_μ, yields, significance, limits) are unchanged — only the goodness-of-fit assessment changes.
- Restructured the results to reflect the corrected goodness-of-fit. The two goodness-of-fit-passing full-data fits (D_NN primary, m_coll) remain in the main results; the full-data m_vis fit and the m_NN cross-check, both of which fail the goodness-of-fit, are moved to a new appendix "Failed-goodness-of-fit cross-checks." The m_vis degeneracy/upper-limit reading is retained and is now quantitatively backed by the goodness-of-fit failure.
- Updated the 10% goodness-of-fit narrative. The corrected method shows the 10% m_coll fit genuinely fails the goodness-of-fit (p = 0.043, a thin-VBF-category statistical tension) and recovers at full luminosity (p = 0.265); the earlier "10% failure was a toy-machinery artifact" framing is removed.
- Added the flagship combined S/(S+B)-weighted post-fit D_NN money plot and the per-category post-fit D_NN panel; added the pre-fit-with-observed-data figure in the post-unblinding chapter; split the $\hat{\mu}$ summary into an expected-sensitivity version (all four observables) and a post-unblind headline version (the two goodness-of-fit-passing observables).
- Added a new appendix "Independent-engine validation with CMS Combine" documenting an independent RooFit/RooStats (Combine via rhalphalib) cross-check of the D_NN fit, the nuisance-parameter pulls, the shape-interpolation $\hat{\mu}$ spread, and the corrected saturated goodness-of-fit (Combine independently confirms the normal-fit verdict). The body remains pyhf-only.
- Shortened the data/MC stack legend entries ("QCD (data-driven)" → "QCD", "W+jets (data-driven)" → "W+jets", "Rare (diboson+single-t)" → "Rare") and moved the dropped qualifiers into the captions of the affected post-fit and pre-fit stack figures; no information is lost.

Phase 5 v1 (final)

- Final documentation pass: prose polish, completeness against the analysis-note checklist, and figure aggregation. All figures are referenced from a single local figures directory and rendered as LaTeX composites grouped by physics.
- Added three methodology schematics — the end-to-end analysis flow, the category-definition logic (forward VBF tag versus central b-veto), and the data-driven background-estimation flow — embedded in the relevant sections so the methodology is followable from the figures alone.
- Composed the improved-mass / m_NN no-go comparison into a single flagship panel and consolidated the per-observable pull, impact, goodness-of-fit, pre-fit, post-fit, and correlation figure groups into composites, with the collapsed figure-range cross-references replaced by single references.
- Expanded the literature program (object-reconstruction and analysis-tooling references added) and verified the systematic and validation completeness against the search-analysis conventions and CMS HIG-13-004 Table 3.

Phase 4c v1 (full unblinding)

- Full unblinding: the signal-extraction quantities ($\hat{\mu}$, significance, limit) are computed on the complete 27,240-event opposite-sign signal-region dataset (full 11.467 fb^{-1}), following the Phase-4b human-gate approval. The full-data observed results are now the PRIMARY result; the 10% subsample results become a validation cross-check (a full-vs-10%-vs-expected comparison).
- Observed full-data signal strength: $\hat{\mu}$(D_NN) = 0.34 ± 1.20 (primary), $\hat{\mu}$(m_coll) = 2.19 ± 2.06, $\hat{\mu}$(m_vis) = 7.70 ± 3.02. The primary observable D_NN is fully consistent with the Standard Model (pull -0.55σ) and with the published CMS μ (-0.36σ vs 0.78, -0.61σ vs 1.09). Observed significance Z(D_NN) = 0.29σ; observed 95% CLs upper limit μ(D_NN) < 3.19.
- The 4b 1.4–1.5σ VBF-dominated upward fluctuation **regressed** at full luminosity, confirming the partial-unblinding prediction: $\hat{\mu}$(D_NN) 5.12 → 0.34, $\hat{\mu}$(m_coll) 7.46 → 2.19, σ_μ shrank to the expected full-data values, and the +3-event VBF excess vanished (VBF now 71 on ≈84.5 background — a deficit). The 4b fluctuation is confirmed to have been a statistical fluctuation; no genuine excess persists.
- The m_vis baseline $\hat{\mu}$ = 7.70 (apparent Z = 2.9σ) is dispositioned as a signal/background-normalisation **degeneracy artifact** of the broadest, lowest-purity observable, not a genuine excess: the statistics-only $\hat{\mu}$ rails at 0, the apparent excess is spread flat across all categories with no localised structure, and the m_vis fit formally fails the saturated goodness-of-fit (toy p = 0.000): the post-fit model does not describe the m_vis data. The physically meaningful m_vis result is the upper limit, not the central $\hat{\mu}$ or the "2.9σ".

- Goodness-of-fit: the full-data saturated goodness-of-fit shows the primary discriminant D_NN passes (toy p = 0.105, a normal fit) and the collinear mass m_coll passes (p = 0.265), while the m_vis baseline (p = 0.000) and the m_NN cross-check (p = 0.000) fail. The goodness-of-fit toys are calibrated on Asimov (toy mean/ndf $\approx$ 1.0). The full m_vis fit and the m_NN cross-check are reported in the failed-goodness-of-fit appendix.
- 95% CLs upper limits: the asymptotic limits are primary (D_NN $\mu < 3.19$), with a one-point full-data toy cross-check (CLs(μ=1.917) = 0.238) and the 4a toy-validation quantifying a ~1.7σ-optimistic caveat. The toy-limit scan deadlocked the optimizer on the degenerate model (documented limitation).
- Updated the Results, comparison, cross-checks, blinding, conclusions, and validation-summary sections to the full-data observed results; updated all results tables and figures to the full-data post-fit stacks, $\hat{\mu}$-summary, and corrected-GoF figures.

Phase 4b v1 (10% partial unblinding)

- First partial unblinding on a fixed-seed 10% data subsample (L = 1.1467 fb^{-1}): the three primary observables returned a coherent $\approx$1.4–1.5σ upward signal strength ($\hat{\mu}$(D_NN) = 5.12 $\pm$ 2.91 under the production $\pm$35% prior; the value quoted in the v2 body is the in-situ-t$\bar{\text{t}}$ re-fit of the same 10% subsample, $\hat{\mu}$(D_NN) = 5.48 $\pm$ 2.92), investigated and dispositioned as a VBF-dominated small-statistics fluctuation with a falsifiable prediction for full data; added the 10% validation section and the expected/observed comparison, presented to the human gate.

Phase 4a v1

- Initial complete analysis note with expected (Asimov) results only — introduction, samples, selection, observable construction and background model, all systematic subsections, statistical method, validation tests, and the systematic-program comparison to CMS HIG-13-004; records the m_NN regression no-go and the three-primary-observable program (m_vis, D_NN, m_coll).

1 Introduction

The decay of the Higgs boson to a pair of τ leptons is the most sensitive direct probe of the Higgs Yukawa coupling to fermions, and was one of the channels that established the fermionic couplings of the 125 GeV boson discovered at the LHC (CMS Collaboration 2014, 2018, 2012; ATLAS Collaboration 2012). The branching fraction BR(H$\to\tau\tau$) = 0.06272 at m_H = 125 GeV (LHC Higgs Cross Section Working Group 2013) makes this the largest of the directly accessible fermionic decay modes. Among the di-tau final states, the $\mu\tau$_h channel — one isolated muon from a leptonic τ decay together with one hadronically decaying τ (τ_h) — combines a clean, well-measured muon with a τ_h that retains a sizeable branching fraction, giving a favourable balance of signal yield and background rejection. The $\mu\tau$_h final state accounts for $\approx$22.6% of all $\tau\tau$ decays (twice the product BR($\tau\to\mu\nu\nu$) = 17.39% $\times$ BR($\tau\to$hadrons) = 64.79% (Particle Data Group 2024)).

This note documents a search for the Standard Model H$\to\tau\tau$ signal in the $\mu\tau$_h channel using the CMS Open Data 2012 release. The parameter of interest is the signal strength

$$\mu = \frac{(\sigma \cdot \mathrm{BR})_{\mathrm{obs}}}{(\sigma \cdot \mathrm{BR})_{\mathrm{SM}}}, \tag{1}$$

the ratio of the observed Higgs production-times-decay rate to the Standard Model expectation, common to the gluon-fusion (ggH) and vector-boson-fusion (VBF) production modes. We extract μ from a simultaneous binned maximum-likelihood template fit across three mutually exclusive event categories (0-jet, boosted, VBF), and report the expected discovery significance and a modified-frequentist CLs upper limit. The analysis follows the CMS $\mu\tau$_h selection of HIG-13-004 (CMS Collaboration 2014) closely in object definitions and category structure, adapted to the contents of the open-data skim.

1.1 Scope, dataset, and relation to prior measurements

The open-data sample used here is the TauPlusX primary dataset for the Run2012B and Run2012C eras only (no 2012A or 2012D). The corresponding integrated luminosity is 11.467 fb^{-1}, roughly 58% of the full 2012 dataset of 19.7 fb^{-1} used by CMS. The published CMS H$\to\tau\tau$ measurements combine seven di-tau sub-channels at 7 and 8 TeV and report a best-fit signal strength $\mu = 0.78 \pm 0.27$ with an expected (observed) significance of 3.6σ (3.4σ) (CMS Collaboration 2014), later updated to $\mu = 1.09\ {}^{+0.27}_{-0.26}$ with the addition of 13 TeV data (CMS Collaboration 2018). The present analysis is a single-channel ($\mu\tau$_h-only), half-luminosity, open-data study; it does not aim to

match the published precision. Its objectives are (i) to demonstrate a complete, systematics-aware signal-extraction chain — a data-driven background model, a validated multivariate discriminant, and a full pyhf statistical model — end to end on open data, and (ii) to establish the relative sensitivity ordering of several di-tau-mass and discriminant constructions. The comparison to the published results is therefore framed on the **consistency of the fitted μ with unity** — which the full-data result satisfies for the primary observable D_NN ($\hat{\mu} = 1.20 \pm 1.13$, consistent with the published CMS per-channel $\mu\tau$_h $= 1.01 \pm 0.41$ at $+0.16\sigma$; Section 10.1) — and on the **relative ordering** of the observables, not on the absolute precision.

1.2 The observable program and the m_NN no-go

A distinctive feature of this analysis is that it carries several di-tau-mass and discriminant constructions through an identical selection, category structure, and systematic model, so that the only difference between them is the fitted variable. This isolates the sensitivity gain attributable to improved mass or discriminant reconstruction — the direct analogue of the ∼40% gain that the published analysis obtained from its SVfit mass estimator over the visible mass (CMS Collaboration 2014).

The **baseline** observable is the visible di-tau mass, the invariant mass of the visible muon and τ_h decay products,

$$m_{vis} = \sqrt{(p_\mu + p_{\tau_h})^2} = \sqrt{m_\mu^2 + m_{\tau_h}^2 + 2(E_\mu E_{\tau_h} - \vec{p}_\mu \cdot \vec{p}_{\tau_h})}. \quad (2)$$

Because the neutrinos from the two τ decays escape detection, the H$\to\tau\tau$ peak in m_vis sits well below 125 GeV; m_vis is the baseline discriminant used by CMS in the $\ell\tau$_h channels (CMS Collaboration 2014).

The **primary** observable is a multivariate discriminant $D_{NN} \in [0, 1]$ (a gradient-boosted decision tree; Section 4.3) trained to separate H$\to\tau\tau$ signal from the summed background using fifteen reconstructed-level inputs. The third primary observable is the analytic **collinear mass** m_{coll}, which augments the visible system with the missing transverse energy under the collinear approximation for the escaping neutrinos (Ellis et al. 1988) (Section 4.2).

A fourth construction was studied — a neural-network regression of the di-tau mass, m_{NN}, from reconstructed inputs including the missing transverse energy. The original proposal to regress the generator-level missing energy is infeasible in this dataset: the skim contains no generator-level missing energy and the signal generator-particle collection is filtered to charged leptons (τ/e/μ) with no neutrinos. The construction was therefore reinterpreted as a regression of the di-tau mass itself from reconstructed inputs, trained on signal only (no background sample carries a generator-level target). The exploration phase established that the generator-level τ-pair training target is a near-delta at 125 GeV (the Higgs natural width Γ_H ≈ 4 MeV makes the true line shape a delta on this scale), so the target is effectively a constant class label. Its validity as a genuine mass estimator therefore rested entirely on a pre-committed independence gate, which it failed (Section 4.4). The m_NN construction is consequently retained as a documented cross-check, not a primary fit observable, and this outcome is reported as a methodological result rather than hidden — it is a feature of the analysis's rigour. The three primary fit observables are m_vis, D_NN, and m_coll.

1.3 Note organisation

Section 2 documents the data and simulated samples and their normalisation. Section 3 defines the object selection, event selection, and category scheme, with the forward-jet VBF tag. Section 4 describes the construction of each fit observable, the data-driven background estimates and their closure tests, and the statistical model with its key equations. Section 5 documents every systematic uncertainty source. Section 6 describes the statistical method and its validation. Section 7 presents the expected (Asimov) sensitivity benchmark. Section 8 presents the full-data observed results — the primary result of the note, including the signal/background degeneracy diagnostics, the corrected goodness-of-fit, and the upper limits. Section 9 retains the 10% partial unblinding as a validation cross-check and the falsifiable-test record. Section 10 compares the result, the systematic program, and the expected sensitivity to the published measurements. Sections 13–15 give the conclusions, future directions, and known limitations, followed by appendices with the validation summary, the covariance structure, additional data/MC distributions, the limitation index, and the reproduction contract.

The end-to-end analysis is summarised schematically in Figure 1, which traces the data and simulated samples from the open-data skim, through the object and event selection and the three-category split, the data-driven background model and the parallel observable construction, to the simultaneous binned maximum-likelihood fit and the staged unblinding. The diagram makes explicit the two design features that distinguish this analysis: the parallel carriage

of three primary fit observables (plus the m_NN cross-check) through one identical selection and systematic model, and the staged 0% → 10% → 100% unblinding gated by the human review.

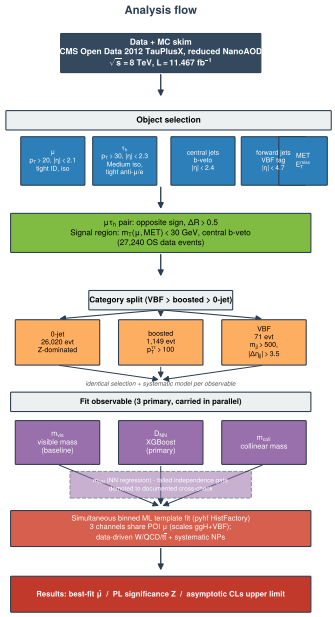

Figure 1: Analysis flow schematic. The end-to-end signal-extraction chain, from the CMS Open Data skim through object and event selection, the three-category split (0-jet / boosted / VBF), the data-driven W+jets / QCD / $t\bar{t}$ background model, and the parallel construction of the three primary fit observables (m_vis, D_NN, m_coll) plus the m_NN cross-check, into the simultaneous binned maximum-likelihood pyhf fit. The staged 0% (expected) → 10% (partial, human gate) → 100% (full) unblinding is shown on the right. This diagram orients the reader to the structure that the rest of the note documents section by section.

2 Data and simulated samples

The analysis uses the CMS Open Data 2012 release in the reduced-NanoAOD "H→$\tau\tau$" outreach format (CERN Open Data records 12350/12351). All samples are read from a local Parquet skim derived from that release; nothing was re-skimmed from the analysis-object-data (AOD) level. The skim retains the full jagged per-object collections (muons, τ_h candidates, jets to pseudorapidity $|\eta| \approx 5.06$, and the event-level missing transverse energy with its full covariance), which is what makes the forward-jet VBF tag and the data-driven control regions feasible. The simulated signal samples carry a generator-particle collection filtered to charged leptons; the background and data samples do not. No per-event generator weights, pileup-truth information, or precomputed scale factors are present in the skim, so Monte Carlo (MC) samples are normalised externally (Section 2.3).

2.1 Data sample

The data are the TauPlusX primary dataset, selected online by the muon-plus-τ_h cross-trigger `HLT_IsoMu17_eta2p1_LooseIsoPFTau20`, which fires for 100% of the skimmed events (the skim pre-selects on it). Only the Run2012B and Run2012C eras are present. The integrated luminosity is taken from the CMS Open Data luminosity record and the outreach reconstruction constant, $L = 11.467$ fb^{-1} (Run2012B ≈ 4.4 fb^{-1} + Run2012C ≈ 7.1 fb^{-1}), with a 2.6% uncertainty (CMS Collaboration 2013). Table 1 summarises the data.

2.2 Simulated samples

The simulated samples comprise the two Higgs signal processes (gluon fusion and vector-boson fusion, both at m_H = 125 GeV and forced to decay to $\tau\tau$), inclusive Drell–Yan (Z/γ^*→$\ell\ell$), inclusive $t\bar{t}$, and jet-binned W+jets (W+1/2/3 jets). There is no inclusive or W+0/W+4 jet sample, no QCD multijet sample, and no single-top or diboson sample — these gaps are addressed by the data-driven W+jets and QCD estimates and by a small "rare" template (Sections

Table 1: Data sample. The TauPlusX primary dataset, Run2012B+C only, selected by the muon-plus-τ_h cross-trigger. Skim events are the post-loose-skim counts; the integrated luminosity is from the CMS Open Data luminosity record (CMS Collaboration 2013). Per-era luminosities are approximate; the binding total is 11.467 fb^{-1}.

Era	$\sqrt{s}$ [GeV]	Trigger	Skim events	L [fb^{-1}]
Run2012B	8000	IsoMu17+τ20	7,134,355	~4.4
Run2012C	8000	IsoMu17+τ20	10,186,503	~7.1
Total (B+C)	8000	IsoMu17+τ20	**17,320,858**	**11.467**

4.6.1–4.8). The cross sections are the outreach reconstruction values, which derive from the LHC Higgs Cross Section Working Group Yellow Report 3 (LHC Higgs Cross Section Working Group 2013) for the Higgs samples and from standard NNLO computations for the backgrounds. Table 2 lists the samples.

Table 2: Simulated samples. Cross sections are the outreach reconstruction values (Higgs samples from the LHC-HXSWG YR3 (LHC Higgs Cross Section Working Group 2013); backgrounds from standard NNLO). N_{gen} is the pre-skim generated count. The signal σ are total production cross sections; the H$\to\tau\tau$ branching fraction multiplies the signal weight explicitly (Section 2.3). The W+jets samples are jet-binned (no inclusive or 0/4-jet bins), and there is no QCD, single-top, or diboson sample.

Process	Role	σ [pb]	N_{gen}	Skim events	Notes
ggH$\to\tau\tau$	signal (ggF)	19.6	476,963	28,324	σ total production (LHC Higgs Cross Section Working Group 2013)
VBF$\to\tau\tau$	signal (VBF)	1.55	491,653	42,394	σ total production (LHC Higgs Cross Section Working Group 2013)
DYJetsToLL	Z$\to\tau\tau$ + Z$\to\ell\ell$	3503.7	30,458,871	4,602,695	inclusive, m_$\ell\ell > 50$
TTbar	$t\bar{t}$	225.2	6,423,106	812,240	inclusive
W1JetsToLNu	W($\to\ell\nu$)+1j	6381.2	29,784,800	1,123,158	jet-binned
W2JetsToLNu	W($\to\ell\nu$)+2j	2039.8	30,693,853	2,131,562	jet-binned
W3JetsToLNu	W($\to\ell\nu$)+3j	612.5	15,241,144	1,376,726	jet-binned

2.3 Normalisation

Because the skim stores no generator weights, MC events are normalised externally by the product of cross section and integrated luminosity divided by the generated count. Background events are weighted by

$$w_{\text{bkg}} = \frac{\sigma \cdot L_{\text{int}}}{N_{\text{gen}}}, \qquad L_{\text{int}} = 11.467 \text{ fb}^{-1}, \tag{3}$$

where the tabulated background cross sections already include the generated decay ($\to\ell\ell$ or $\to\ell\nu$). The signal cross sections in Table 2 are **total production** cross sections, so the H$\to\tau\tau$ branching fraction multiplies the signal weight explicitly,

$$w_{\text{sig}} = \frac{\sigma_{\text{prod}} \cdot \text{BR}(H \to \tau\tau) \cdot L_{\text{int}}}{N_{\text{gen}}}, \qquad \text{BR}(H \to \tau\tau) = 0.06272 \text{ [@}LHCHXSWG2013\text{]}. \tag{4}$$

The signal samples are generated with the Higgs forced to decay to $\tau\tau$, so N_{gen} counts H$\to\tau\tau$ events; multiplying by the branching fraction in the numerator then correctly scales each event to the physical H$\to\tau\tau$ production rate. This was verified against the outreach reconstruction weight table: for example the ggH per-event weight 0.0295547 reproduces $19.6 \times 11.467 \times 0.06272 \,/\, 476{,}963 \times 10^3$, and the Drell–Yan weight 1.319055 reproduces 3503.7×11.467

/ 30,458,871 $\times 10^3$. The W+jets and QCD multijet normalisations are replaced entirely by the data-driven estimates of Sections 4.6.1–4.6.2; the jet-binned W cross sections are used only for the W shape. For the 10% partial-data validation the MC is normalised to $0.1 \cdot$ L_int.

The pre-fit integrated yields (over the three categories) used in the m_vis fit are: ggH 101.1, VBF 9.4, Z→$\tau\tau$-like (DY_Z$\tau\tau$) 22312, Z→$\ell\ell$-like (DY_Z$\ell\ell$) 469.6, t$\bar{\text{t}}$ 264.1, W+jets 3176.7, QCD 3141.5, and rare 691.5 events, for a total signal of 110.5 events; the per-category pre-fit breakdown is given in Table 11. These yields are reproduced exactly by the re-run of the full selection used to build the systematic templates, a closure that validates the template machinery.

3 Event selection

The event selection identifies one isolated muon and one opposite-sign hadronic τ candidate, suppresses the W+jets and t$\bar{\text{t}}$ backgrounds with a transverse-mass cut and a b-jet veto, and assigns each event to one of three mutually exclusive categories. Object thresholds follow CMS HIG-13-004 Table 1 (CMS Collaboration 2014) and the outreach reconstruction, adapted to the skim. The selection is implemented as columnar boolean masks on the jagged per-event object collections (no event loops), using the awkward-array library for the variable-length data structures (Pivarski et al. 2020); the cutflow is monotonically non-increasing at every stage.

3.1 Object selection

The muon is required to have transverse momentum $p_T^{\mu} > 20$ GeV (on the plateau of the IsoMu17 trigger), $|\eta^{\mu}| < 2.1$, to pass the tight muon identification, and to be isolated with relative isolation `Muon_pfRelIso03_all` < 0.1; muons carrying the -999 isolation sentinel (≈6–13% of muons) fail the isolation requirement and are excluded. Impact parameters $|d_z| < 0.2$ cm and $|d_{xy}| < 0.045$ cm are required.

The hadronic τ candidate is reconstructed with the hadron-plus-strips (HPS) algorithm, which builds τ_h candidates from charged hadrons and reconstructed neutral-pion strips and whose 8 TeV performance is documented in (CMS Collaboration 2016). The candidate is required to have $p_T^{\tau_h} > 30$ GeV, $|\eta^{\tau_h}| < 2.3$, to pass the decay-mode-finding requirement, and to satisfy the medium combined-isolation working point (`Tau_idIsoMedium`). This working point is loosened a priori from the tighter outreach default to recover efficiency, given that no τ_h identification scale factors are available in the open-data skim; the τ_h identification scale factor is fixed to unity and the missing-scale-factor ignorance is carried entirely by the profiled Z→$\tau\tau$ normalisation nuisance parameter (Section 5.2.2). The exploration phase confirmed that the skim retains the full loose τ_h collection down to and below the very-loose working point, so this loosening is feasible and gains ≈7.5% relative signal efficiency over the tight working point. The candidate must also pass the tight anti-muon and tight anti-electron discriminators, which suppress muon-fake and electron-fake τ_h.

The muon and τ_h are paired with $\Delta R(\mu, \tau_h) > 0.5$; the leading- p_T muon and leading-p_T τ_h are taken. The signal region requires opposite charge ($q_{\mu} \cdot q_{\tau_h} < 0$); the same-sign region is retained as a flag for the QCD estimate. To suppress W+jets, the transverse mass of the muon and the missing transverse energy is required to satisfy

$$m_T(\mu, \mathrm{MET}) = \sqrt{2\, p_T^{\mu}\, E_T^{\mathrm{miss}} (1 - \cos \Delta\phi(\mu, \vec{E}_T^{\mathrm{miss}}))} < 30 \text{ GeV}, \tag{5}$$

which removes the Jacobian-peak region populated by leptonic W decays.

Two separate jet collections are used, a critical design choice. The **tag-jet collection** for the VBF tag uses jets with $p_T^j > 30$ GeV, $|\eta_j| < 4.7$ (full forward acceptance), passing the pileup-jet identification and separated from the leptons by $\Delta R > 0.5$; it defines the jet multiplicity n_{jet}, the di-jet mass m_{jj}, and the rapidity gap $\Delta\eta_{jj}$. The **central b-veto collection** uses jets with $p_T^j > 20$ GeV, $|\eta_j| < 2.4$ (tracker acceptance) and a b-tag discriminant above the medium CSV working point (0.679); the $-10/-1$ b-tag sentinels (77.7% of jets, indicating no tagger value) are treated as untagged. Events with any central b-tagged jet are rejected. The b-veto, despite the sentinel-degraded discriminant, removes the bulk of the residual t$\bar{\text{t}}$ background.

3.2 N-1 distributions and cut sensitivity

The principal selection variables are validated with N-1 distributions, in which all cuts except the one shown are applied. The muon and τ_h transverse momenta (Figures 2 and Figure 2), the transverse-mass cut (Figure 2), and the pair separation (Figure 2) all show the signal concentrating above the threshold and the background falling, with the cut placed where the signal-to-background ratio turns over. The data are well described by the background model across the full range of each variable, both before and after the cut.

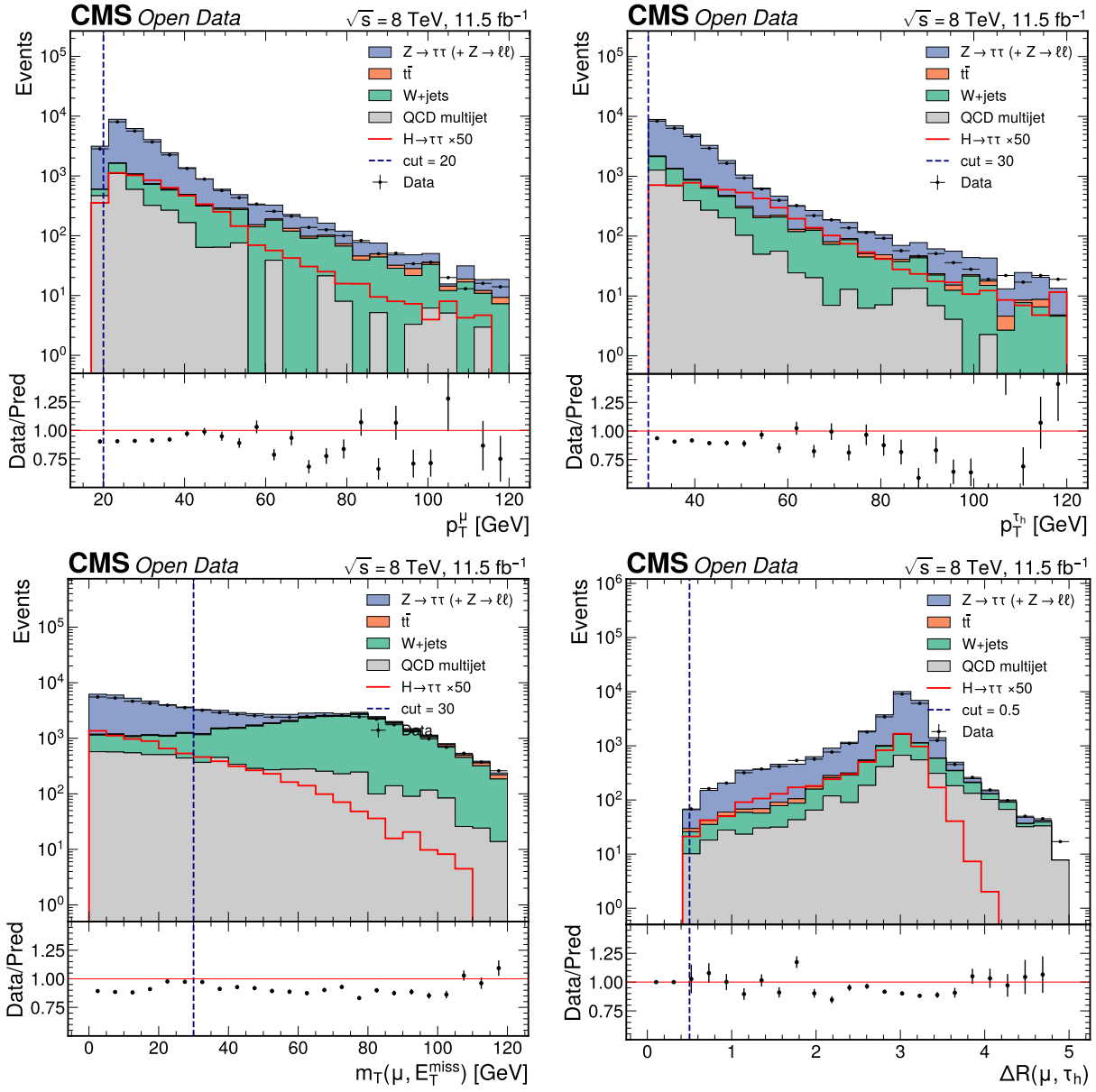

Figure 2: **(a)** N-1 distribution of the muon transverse momentum, with all selection cuts except $p_T^{\mu} > 20$ GeV applied. The signal (scaled) rises above 20 GeV while the Z→$\tau\tau$ and W+jets backgrounds fall, motivating the threshold. The data/model ratio is flat across the range, confirming the muon kinematics are well modelled. **(b)** N-1 distribution of the τ_h transverse momentum, all cuts except $p_T^{\tau_h} > 30$ GeV applied. The 30 GeV threshold sits where the signal efficiency is high and the jet-faking-τ_h background (W+jets, QCD) is suppressed. The background model reproduces the data shape across the spectrum. **(c)** N-1 distribution of the muon–MET transverse mass, all cuts except $m_T < 30$ GeV applied. The W+jets and t$\bar{\text{t}}$ backgrounds populate the high-m_T Jacobian region while the H→$\tau\tau$ signal and Z→$\tau\tau$ concentrate at low m_T, motivating the 30 GeV cut. The high-m_T region defines the W+jets control region used for the data-driven estimate. **(d)** N-1 distribution of the muon–τ_h angular separation $\Delta R(\mu, \tau_h)$, all cuts except $\Delta R > 0.5$ applied. The collinear-pair region below 0.5 is removed to reject overlapping objects. Signal and the dominant Z→$\tau\tau$ background both peak in the back-to-back region, and the data are well described.

3.3 Category scheme and the VBF tag

Each event is assigned to exactly one of three categories, with priority VBF > boosted > 0-jet:

- **VBF:** at least two tag jets with $m_{jj} > 500$ GeV and $|\Delta\eta_{jj}| > 3.5$ (the HIG-13-004 loose VBF tag for $\mu\tau$_h (CMS Collaboration 2014)). The forward tag-jet acceptance to $|\eta| < 4.7$ is decisive: capping the tag jets at the tracker acceptance $|\eta| < 2.4$ would lose 915 of 1,013 raw VBF-signal events in this category. The forward-jet pseudorapidity distribution (Figure 4) and the emptiness test (Figure 5) demonstrate this directly. The b-veto uses the separate central collection, so forward tag jets and the central b-veto never interfere.
- **Boosted:** at least one tag jet and a di-tau system transverse momentum $p_T^{\tau\tau} = |\vec{p}_T^{\,\mu} + \vec{p}_T^{\,\tau_h} + \vec{E}_T^{\,\mathrm{miss}}| > 100$ GeV, not satisfying the VBF tag. The boosted topology improves mass resolution and ggH sensitivity.
- **0-jet:** the remainder. This category is overwhelmingly Z→$\tau\tau$-dominated and anchors the Z-related nuisance parameters from the large Z peak.

A loose/tight VBF split was evaluated and collapses to loose-only: the VBF category holds only 3.59 expected signal events in total, below the threshold of ≥3 events in the tight bin needed to justify a sub-split. Because the boosted-category boundary depends on the missing transverse energy (through $p_T^{\tau\tau}$) and the VBF tag depends on jet energies, the MET-scale, jet- energy-scale, and jet-energy-resolution systematic variations re-derive the category assignment, allowing events to migrate between categories (Section 5.1).

The category-definition logic, and in particular the use of two distinct jet collections — the forward tag-jet collection to $|\eta| < 4.7$ for the VBF tag and the central collection to $|\eta| < 2.4$ for the b-veto — is shown schematically in Figure 3. The two collections never interfere: a forward tag jet at $|\eta| \approx 4$ enters the VBF di-jet decision but is invisible to the b-veto, while a central b-tagged jet vetoes the event without affecting the VBF tag. The diagram also shows the priority ordering VBF > boosted > 0-jet that makes the three categories mutually exclusive.

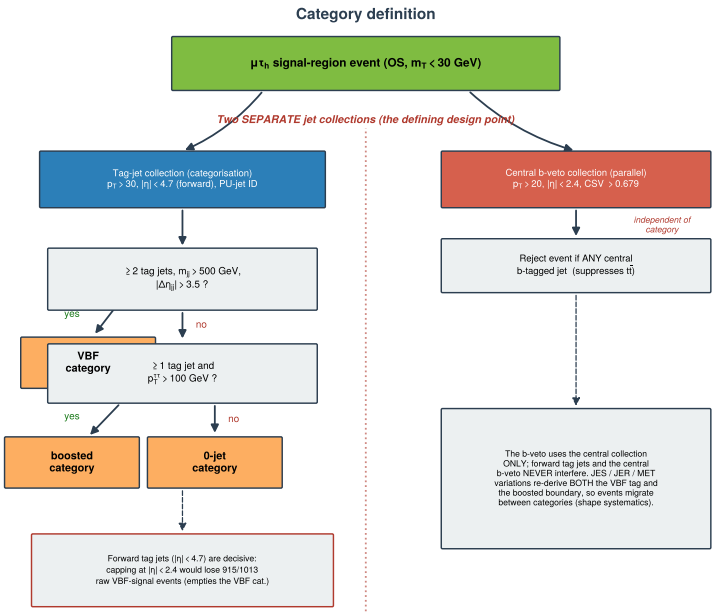

Figure 3: Category-definition schematic. Each selected $\mu\tau$_h event is routed through the priority chain VBF > boosted > 0-jet using two independent jet collections: the forward tag-jet collection (p_T > 30 GeV, $|\eta| < 4.7$) that defines the di-jet mass m_jj, the rapidity gap $|\Delta\eta$_jj|, and the jet multiplicity for the VBF tag, and the central collection (p_T > 20 GeV, $|\eta| < 2.4$) that defines the b-veto. The forward acceptance is decisive for the VBF category; the central b-veto suppresses tt̄. The boosted boundary uses the MET-dependent di-tau p_T, so detector variations migrate events between categories. The two collections are deliberately separate so the forward VBF tag and the central b-veto cannot interfere.

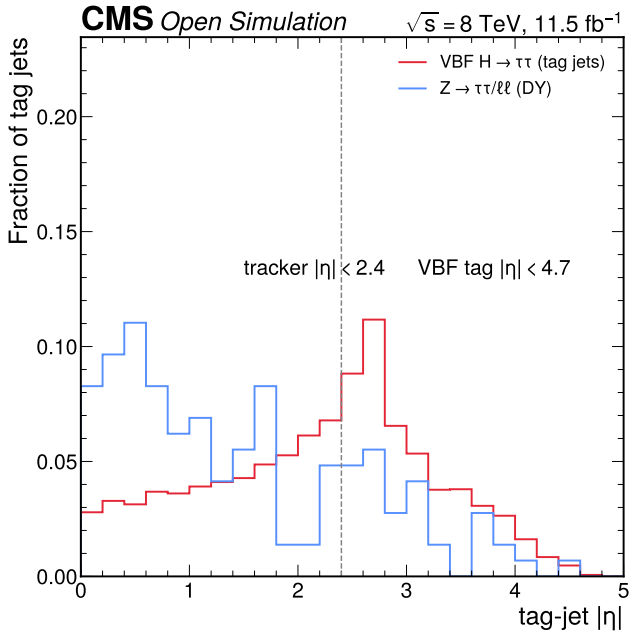

Figure 4: Forward tag-jet pseudorapidity distribution, showing substantial population at $|\eta| > 2.4$. This is the physics reason the tag-jet collection must extend to $|\eta| < 4.7$: VBF tag jets are predominantly forward, and capping at the tracker acceptance would discard most of them and empty the VBF category. The distribution is shown for the VBF-tagged events.

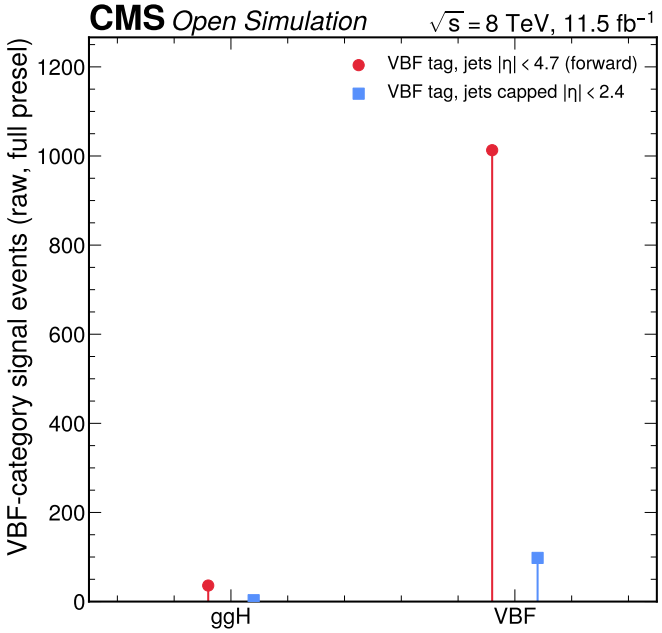

Figure 5: VBF-category emptiness test under a hypothetical $|\eta| < 2.4$ jet cap. Restricting the tag jets to the tracker acceptance would lose 915 of 1,013 raw VBF-signal events (and the bulk of the ggH and data events), effectively emptying the category. This confirms the decision to use forward tag jets to $|\eta| < 4.7$ and that the VBF category survives only with forward acceptance.

3.4 Cutflow

Table 3 gives the $\sigma \cdot \mathrm{L/N\_gen}$-normalised expected yields through the selection sequence. The yields are monotonically non-increasing at every stage. The $m_T < 30$ GeV cut strongly suppresses W+jets (29,575 → 3,846) and $\mathrm{t\bar{t}}$ (7,031 → 1,056), and the central b-veto then removes the bulk of the residual $\mathrm{t\bar{t}}$ (1,056 → 272). The full cutflow is shown graphically in Figure 6.

Table 3: Selection cutflow ($\sigma \cdot$ L/N_gen-normalised expected yields). The W+jets column is the sum of the W1+W2+W3 jet-binned samples. Data counts are unweighted. The final OS signal region contains 27,240 data events.

Stage	ggH	VBF	DY	$t\bar{t}$	W+jets	data
All (skim)	837	96	6.07×10^6	326,556	5.02×10^6	17,320,858
$\geq$1 good μ	767	90	5.93×10^6	308,847	4.67×10^6	13,367,253
μ+τ_h pair	149	15	35,851	8,415	37,151	85,525
μ $p_T > 20$ (SR)	149	15	35,837	8,404	37,150	85,491
Opposite sign	148	15	32,181	7,031	29,575	66,895
$m_T < 30$	104	10	22,908	1,056	3,846	28,200
b-veto	103	10	22,611	272	3,809	27,240

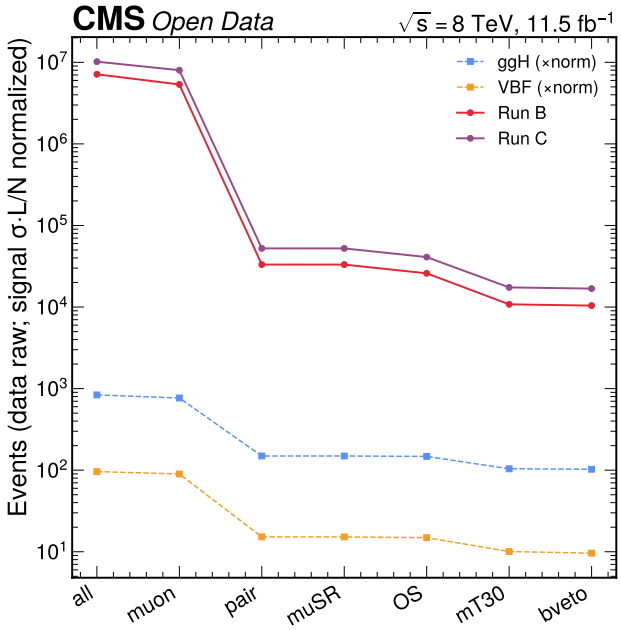

Figure 6: Selection cutflow through the $\mu\tau$_h selection sequence for signal and the principal backgrounds, $\sigma \cdot$ L/N_gen-normalised. Each stage is monotonically non-increasing. The $m_T < 30$ GeV cut suppresses W+jets and $t\bar{t}$ (the Jacobian-edge cut working as designed), and the central b-veto removes the bulk of the residual $t\bar{t}$ despite the sentinel-degraded b-tag discriminant.

4 Observable construction, background model, and statistical framework

This section is the methodological core of the analysis. It defines the three primary fit observables and the m_NN cross-check, documents the data-driven background estimates and their validation, and specifies the binned-likelihood statistical model and the systematic-propagation scheme. A reader should be able to reproduce the observable definitions, the background normalisations, and the likelihood from the equations given here.

4.1 The visible mass m_vis

The visible di-tau mass m_vis (Eq. Equation 2) is the invariant mass of the visible muon and τ_h four-momenta. It is the simplest and most robust observable: it requires no missing-energy input and is insensitive to MET mismodelling, but it is the least sensitive of the three primary observables because the escaping neutrinos shift the signal peak well below 125 GeV and broaden it, reducing the separation from the Z$\to\tau\tau$ background. The m_vis template is fit on [0, 200] GeV in 20 bins. Figure 7 shows the data and background model in the signal region; the shape agreement is excellent (χ^2/ndf = 1.04 inclusively in the exploration survey) on top of a uniform $\approx$9% MC over-normalisation that the per-process fit normalisations absorb.

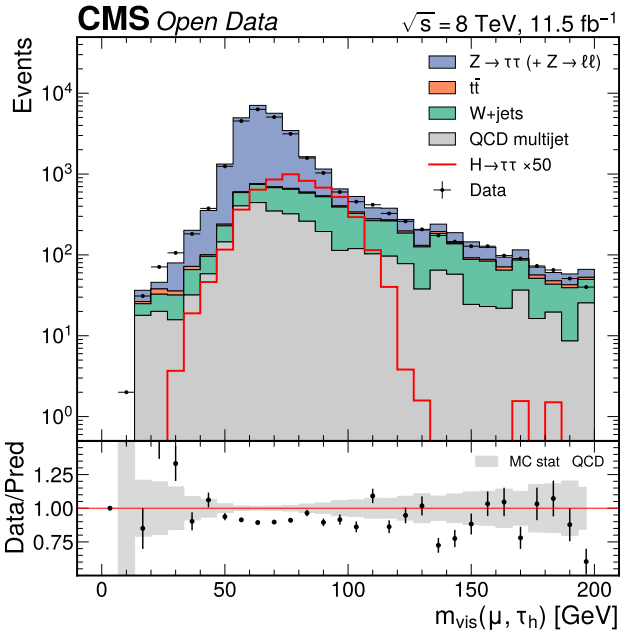

Figure 7: Visible-mass m_vis data and background model in the signal region with a ratio panel. The Z→$\tau\tau$ peak below 125 GeV dominates the spectrum; the signal (overlaid, scaled) is a small broad contribution. The background model reproduces the data shape well, with the residual uniform normalisation offset absorbed by the per-process fit normalisations.

4.2 The collinear mass m_coll

The collinear mass augments the visible system with the missing transverse energy as an estimate of the escaping neutrinos, under the collinear approximation that the neutrinos from a boosted τ are nearly collinear with the visible decay products (Ellis et al. 1988). Defining the visible momentum fractions x_μ and x_{τ_h} carried by each visible leg, obtained by projecting the missing transverse momentum onto the two visible directions,

$$x_\mu = \frac{p_{x,\mu}p_{y,\tau_h} - p_{y,\mu}p_{x,\tau_h}}{p_{x,\mu}p_{y,\tau_h} - p_{y,\mu}p_{x,\tau_h} + E_x^{\text{miss}}p_{y,\tau_h} - E_y^{\text{miss}}p_{x,\tau_h}}, \tag{6}$$

$$x_{\tau_h} = \frac{p_{x,\mu}p_{y,\tau_h} - p_{y,\mu}p_{x,\tau_h}}{p_{x,\mu}p_{y,\tau_h} - p_{y,\mu}p_{x,\tau_h} - E_x^{\text{miss}}p_{y,\mu} + E_y^{\text{miss}}p_{x,\mu}}, \tag{7}$$

the collinear di-tau mass is

$$m_{coll} = \frac{m_{vis}}{\sqrt{x_\mu\, x_{\tau_h}}}. \tag{8}$$

Two limiting cases verify the estimator: as $E_T^{\text{miss}} \to 0$ both $x_\mu, x_{\tau_h} \to 1$ and $m_{coll} \to m_{vis}$, so the estimator reduces to the baseline; and unphysical solutions, where the missing momentum points outside the wedge spanned by the two visible legs (x_μ or $x_{\tau_h} \notin (0,1]$), are flagged. The collinear approximation has a valid physical solution for only 47.4% of signal-region events; for the remaining 52.6% (back-to-back or mismeasured-MET topologies) the event is assigned the total transverse mass as an overflow estimator,

$$m_T^{\text{tot}} = \sqrt{m_T^2(\mu, \text{MET}) + m_T^2(\tau_h, \text{MET}) + m_T^2(\mu, \tau_h)}, \tag{9}$$

so that m_coll has 100% coverage of the 56,199 selected events (zero non-finite values); here the 56,199 is the combined signal+background+data signal-region template-fill count (all samples entering the m_coll templates), distinct from the 27,240 observed signal-region data events of the cutflow (Table 3). The template is fit on [0, 300] GeV in 30 bins, with the overflow tail (data events above 300 GeV) range-clipped at the edge so the per-category normalisation is conserved. Figure 8 shows the data and model. The overflow fallback is more than half the template, so its modelling matters: the data and the background prediction are consistent across the full m_coll range, including the m_T^{tot} overflow region, with the best inclusive shape agreement of all three observables ($\chi^2/\text{ndf} = 0.90$), so the fallback population is not mismodelled relative to the physical-solution population. The collinear mass is the second most sensitive observable; it recovers part of the full-mass information without a learned target and cannot sculpt a class-conditional artefact, making it a robust analytic cross-check on the multivariate discriminant.

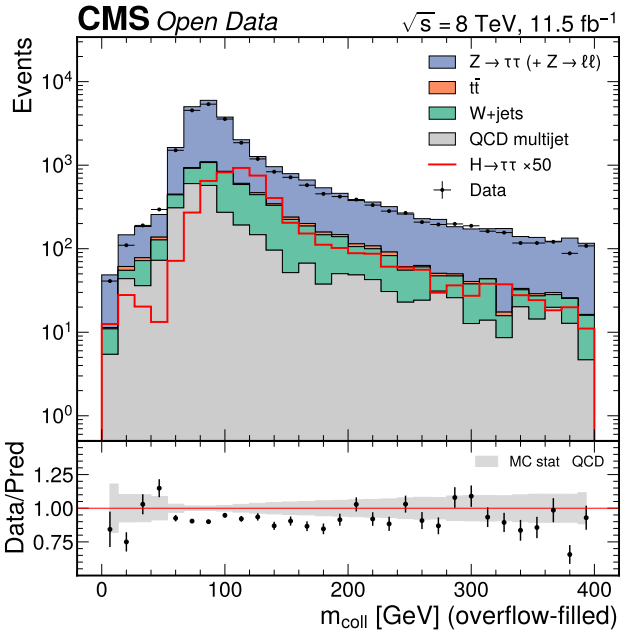

Figure 8: Collinear mass m_coll data and background model in the signal region with a ratio panel. Adding the missing transverse energy to the visible system shifts the signal peak toward the true resonance mass relative to m_vis. The overflow fallback (m_T^{tot}) ensures 100% event coverage; the background model reproduces the data shape (inclusive shape $\chi^2/\mathrm{ndf} = 0.90$, the best of all observables).

4.3 The multivariate discriminant D_NN

The primary observable is a binary classifier separating H→$\tau\tau$ signal (ggH+VBF, weighted) from the summed background (Drell–Yan + $t\bar{t}$ + W+jets), trained on signal-region events with fifteen reconstructed-level inputs. The inputs are selected by a variable-quality gate (Section 4.5) and comprise m_vis, p_T^{μ}, $p_T^{\tau_h}$, η^{μ}, η^{τ_h}, $\Delta R(\mu, \tau_h)$, $\Delta\phi(\mu, \tau_h)$, E_T^{miss}, the MET significance, $m_T(\mu, \mathrm{MET})$, $m_T(\tau_h, \mathrm{MET})$, m_T^{tot}, $p_T^{\tau\tau}$, m_coll, and the raw τ_h isolation. MC events are weighted by $\sigma \cdot$ L/N_gen with the pileup-proxy reweighting (Section 4.5); the classes are balanced; the split is stratified 60/40 train/test with a fixed seed. Two architectures were trained: a gradient-boosted decision tree (XGBoost (Chen and Guestrin 2016), the documented primary) and a shallow multilayer perceptron (the alternative). Table 4 gives the performance.

Table 4: Classifier performance. Both architectures have small train−test AUC gaps, indicating no overtraining; XGBoost is the documented primary on its higher test AUC.

Architecture	Test AUC	Train AUC	Train−test gap
XGBoost (primary)	0.821	0.866	0.045
MLP (alternative)	0.810	0.840	0.029

The two architectures agree closely and have small train−test gaps, indicating no overtraining (Figures 9 and Figure 9). The most important inputs by gain and permutation importance are the mass and kinematic discriminants $p_T^{\tau_h}$, m_vis, E_T^{miss}, m_coll, and $p_T^{\tau\tau}$ — all physically motivated for H→$\tau\tau$ versus Z→$\tau\tau$ separation, and no pileup-sensitive variable enters (Figure 9). The classifier output is fit on [0, 1] in 20 bins, with the signal concentrating toward $D_{NN} \to 1$. The data and model agree on the output shape ($\chi^2/\mathrm{ndf} = 1.18$), no worse than the m_vis input modelling ($\chi^2/\mathrm{ndf} = 1.21$), which shows that the classifier compresses rather than amplifies the input mismodelling (Figure 9). The discriminant is the most sensitive single observable, with a held-out expected-significance gain of ≈55% over the m_vis baseline (Section 7).

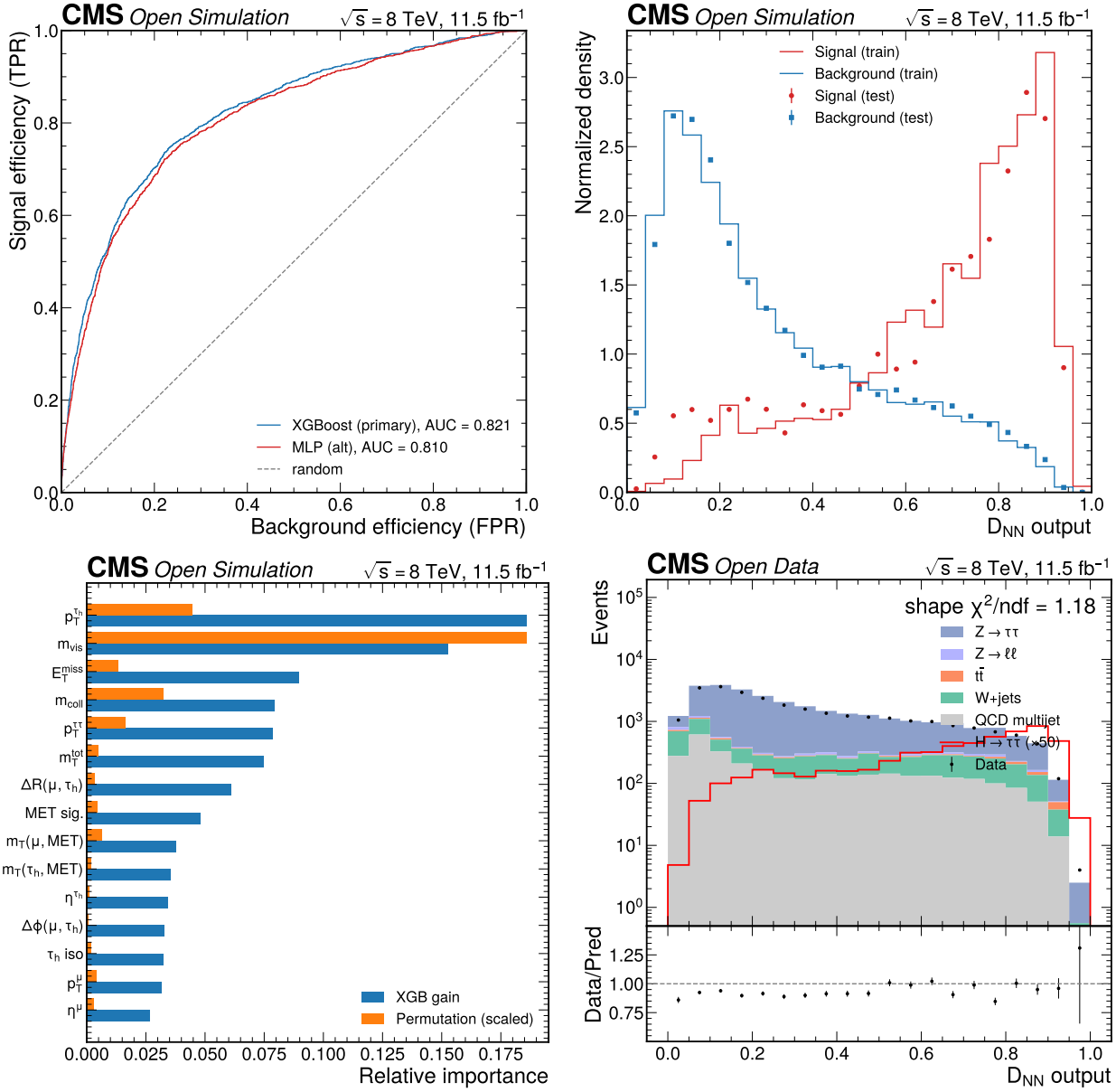

Figure 9: **(a)** D_NN ROC curves (test set, weighted) for the primary XGBoost classifier (AUC 0.821) and the alternative MLP (AUC 0.810). Both clearly separate H→$\tau\tau$ signal from total background, and the two architectures agree closely. The small train−test AUC gaps (0.045, 0.029) indicate no overtraining. **(b)** D_NN output score distributions for training (lines) versus test (points), separately for signal and background. The train and test distributions overlay closely for both classes, confirming no overtraining. Signal concentrates near $D_{NN} \to 1$ and background near 0, the basis of the sensitivity gain over the m_vis baseline. **(c)** D_NN input feature importance from XGBoost gain and permutation importance. The most important inputs are the mass and kinematic discriminants $p_T^{\tau_h}$, m_vis, E_T^{miss}, m_coll, and $p_T^{\tau\tau}$ — all physically motivated for H→$\tau\tau$ versus Z→$\tau\tau$ separation. No pileup-sensitive variable enters, having been removed by the variable-quality gate. **(d)** Data and background model for the D_NN output, summed over the three categories, with the data-driven W/QCD model and a ratio panel. The shape agreement is excellent (χ^2/ndf = 1.18); the uniform normalisation offset is the same MC over-normalisation absorbed by the per-process fit normalisations. The classifier output is no worse modelled than the m_vis input (χ^2/ndf = 1.21), demonstrating that the classifier compresses rather than amplifies the input mismodelling. The signal (scaled) rises toward $D_{NN} \to 1$.

4.4 The m_NN regression and its no-go

The fourth construction was a neural-network regression of the di-tau mass from reconstructed inputs. Because the skim contains no generator-level missing energy and no generator neutrinos, the original genMET-regression proposal is infeasible; the construction was reinterpreted as a regression of the di-tau mass itself, with the per-event leading-opposite-sign generator τ-pair invariant mass as the training target. The missing-energy components and the MET covariance are mandatory inputs, since physical mass reconstruction — recovering the escaping-neutrino momentum — is impossible without them. The network was trained on signal only, because no background sample carries a generator target, and then applied to all samples and data using the reconstructed inputs alone.

The exploration phase established that the generator τ-pair mass is a near-delta at 125 GeV (median 125.00; ggH RMS 1.49 GeV, VBF RMS 2.67 GeV; 100% within [120, 130] GeV), consistent with the Higgs natural width $\Gamma_H \approx 4$ MeV that makes the true line shape a delta on this scale (Figure 11). The residual 1.5–2.7 GeV spread is generator-cascade and reconstruction noise, not a physical Breit–Wigner spread. The target is therefore effectively a constant class label, and the validity of the construction as a genuine mass estimator rests entirely on a pre-committed two-part independence gate.

Gate G3 (out-of-distribution independence, load-bearing): a network trained only on H$\to\tau\tau$ signal (target $\approx$125 GeV), when applied to Drell–Yan Z$\to\tau\tau$ events that it never saw in training and whose true mass $m_Z = 91$ GeV was never a target, must reconstruct a peak near m_Z if and only if it has learned a genuine reconstruction-to-mass map. The pass condition is that the median regressed mass on Z$\to\tau\tau$ lie within $\pm$15% of m_Z (the window [77.5, 105] GeV) and be clearly separated from 125 GeV.

Gate G1 (anti-sculpting): the background-only (μ=0) Asimov m_NN template must show no false peak at 125 GeV.

Both gates fail. On Drell–Yan Z$\to\tau\tau$ the baseline network predicts a median m_NN of 125.18 GeV — essentially identical to its signal prediction (124.93 GeV) and far outside the m_Z window — so the network memorised the constant-125 target rather than learning a reconstruction-to-mass map (Figure 10). Three independent remediations were attempted, as required: a mass-augmented training set in which the visible+MET system is rescaled to a 60–200 GeV mass grid so the target acquires genuine spread (median on Z$\to\tau\tau$ 123.16 GeV), the same with a log-scaled target and heavier regularisation (122.99 GeV), and an independent XGBoost regressor on the augmented set (115.05 GeV). All three still predict 115–123 GeV on Z$\to\tau\tau$ and none reaches the m_Z band. Gate G1 fails dramatically: the background-only Asimov m_NN piles essentially all background into the 125 GeV window, giving a bump-hunt significance of 1605σ — the sculpting artefact the gate was designed to catch, in which the regressor manufactures a signal-mass bump out of pure background (Figure 10). A significance cross-check (gate G2) corroborates the decision: the m_NN expected significance (0.805) is lower than even the m_vis baseline (1.022), because the sculpted 125 GeV peak gives no real separation when the background peaks there too, whereas the analytic m_coll (which has no learned target and cannot sculpt) genuinely improves on m_vis. The mass-comparison figure (Figure 10) shows m_NN collapsing both signal and Z$\to\tau\tau$ into a single 125 GeV delta while m_vis and m_coll separate them.

The decision is therefore a documented no-go: m_NN is demoted to a cross-check and the analysis proceeds with three primary fit observables (m_vis, D_NN, m_coll). This is recorded as a formal downscope, not a silent drop, and is presented here as a methodological result — a demonstration of the value of the pre-committed gate.

The three panels of the m_NN no-go are shown together in Figure 10 — the mass-estimator comparison, the train-on-H/test-on-DY independence gate G3, and the anti-sculpting gate G1 — the analysis's flagship demonstration of the value of the pre-committed independence gate. The generator-level training target that drives the whole outcome is shown separately in Figure 11.

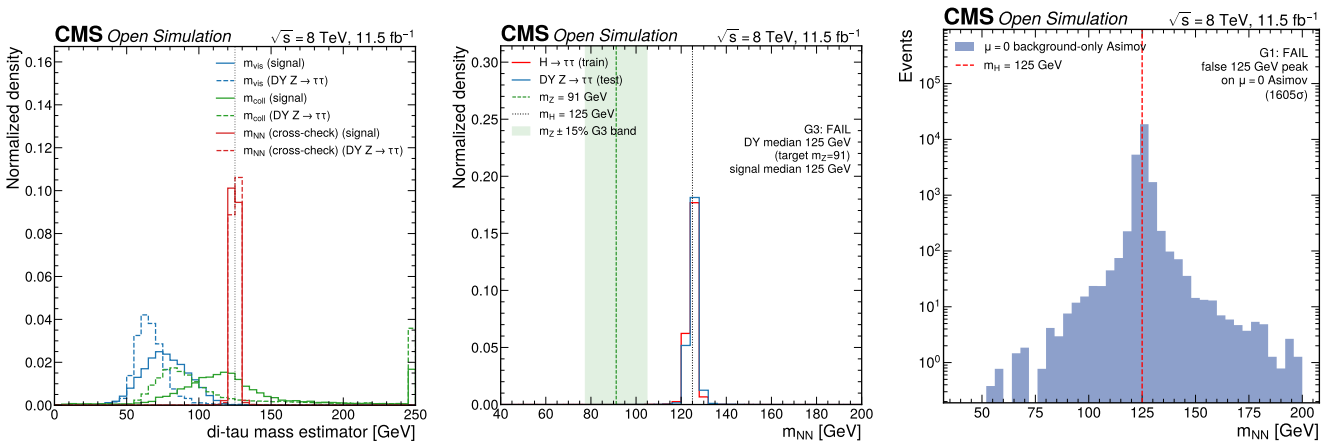

Figure 10: **(a)** The m_NN no-go demonstration, panel (a). Di-tau mass estimator comparison for signal and Drell–Yan Z→$\tau\tau$: m_vis and m_coll clearly separate signal (shifted higher) from Drell–Yan, while m_NN collapses both into a single delta at 125 GeV — a direct visual demonstration that m_NN cannot distinguish signal from background, the basis of the no-go decision. Panels (b) and (c) show the two independence gates that formalise this verdict. **(b)** The m_NN no-go demonstration, panel (b): gate G3 (train-on-H / test-on-DY independence). The signal-only m_NN regressor applied to H→$\tau\tau$ signal (train) and Drell–Yan Z→$\tau\tau$ (test), against m_Z and m_H = 125 GeV, with the m_Z ± 15% G3 band shaded. The Drell–Yan median (125.2 GeV) lies far outside the m_Z window [77.5, 105] GeV and equals the signal prediction, so the network memorised the constant-125 target rather than learning a reconstruction-to-mass map; the gate is a NO-GO for the baseline and all three remediations. **(c)** The m_NN no-go demonstration, panel (c): gate G1 (anti-sculpting). The background-only (μ=0) Asimov m_NN distribution: essentially all background piles into a false peak at exactly 125 GeV (bump-hunt significance 1605σ), the sculpting artefact the gate was designed to catch. The regressor manufactures a signal-mass bump out of pure background, and the gate fails.

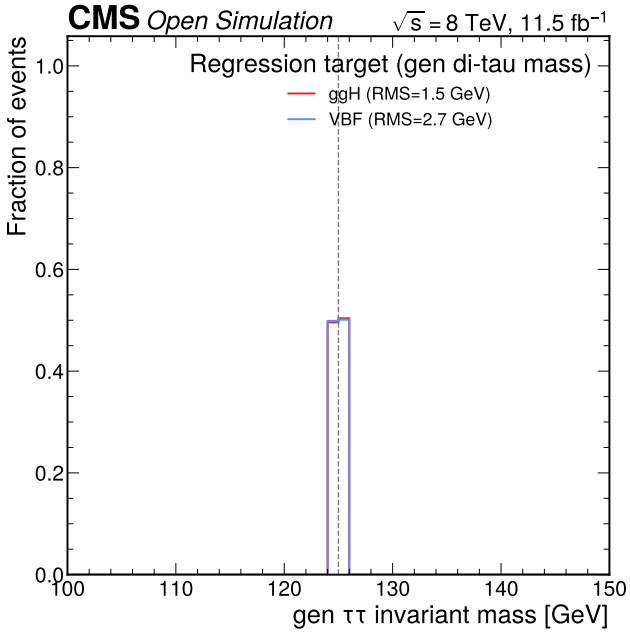

Figure 11: Generator-level leading opposite-sign τ-pair invariant mass for ggH and VBF signal. The distribution is a near-delta at 125 GeV (ggH RMS 1.49 GeV, VBF RMS 2.67 GeV; 100% within [120, 130] GeV), reflecting the Higgs natural width Γ_H ≈ 4 MeV. The residual spread is generator-cascade and reconstruction noise, not a Breit–Wigner spread, so the m_NN training target is effectively a constant 125 GeV label — the root cause of the no-go shown in Figure 10.

4.5 Variable-quality gate and pileup mitigation

The skim contains no pileup-truth information, so no standard pileup reweighting is possible. A primary-vertex-multiplicity (N_PV) based data/MC reweighting is derived in the opposite-sign preselection (with the same-sign

QCD estimate subtracted) as a shape-only pileup proxy that preserves the MC total and is clipped to [0.2, 5]; it corrects the N_PV shape χ^2/ndf from 4.44 to 0.55 (Figure 12). The genuine residual is in the forward/total jet multiplicities: after the reweighting the forward-jet multiplicity shape χ^2/ndf is still 5.69 and the total jet multiplicity 2.37, so these pileup-sensitive variables remain poorly modelled.

A variable-quality gate is applied to every candidate classifier input before training: variables with data/MC shape $\chi^2/\text{ndf} \leq 2$ are kept, those in 2–5 are flagged for calibration, those above 5 are dropped, and the pileup-sensitive forward/total jet multiplicities and N_PV are excluded. The gate keeps fifteen variables; the forward-jet multiplicity (5.69), total jet multiplicity (2.37), N_PV (which closes tautologically after the proxy reweighting), and the borderline $\Delta\eta(\mu, \tau_h)$ (2.02, carrying no discrimination at AUC 0.501) are excluded. The residual forward-jet mismodelling is not zero-impact and is carried forward as a dedicated systematic (Section 5.1.5). Figures 12 and Figure 12 summarise the gate, and Figure 12 shows the single-variable separation ranking.

4.6 Data-driven background estimation

The three largest reducible backgrounds — W+jets, QCD multijet, and $t\bar{t}$ — are constrained from data, and the irreducible Drell–Yan is split at reconstruction level into $Z\to\tau\tau$-like and $Z\to\ell\ell$-like sub-templates. These estimates and their closure tests are described below; the resulting per-category yields define the template normalisations of the fit.

Figure 13 summarises the four data-driven background estimates and the control and validation regions that anchor them. W+jets is normalised from the high-transverse-mass control region and extrapolated to the signal region by a per-category transfer factor; QCD multijet is taken from the same-sign region and transferred by the OS/SS factor; $t\bar{t}$ is constrained in situ by a b-tag control region (an inverted b-veto) carried as three per-category counting channels inside the simultaneous fit, sharing a single freely-floating normalisation k_ttbar with the signal region; and the single Drell–Yan sample is split at reconstruction level into its irreducible $Z\to\tau\tau$-like and reducible $Z\to\ell\ell$-like components. Each data-driven estimate is validated in a dedicated region disjoint from the control region used to derive it, so the closure tests (Section 4.6.5) are genuine out-of-region checks rather than self-consistency algebra.

4.6.1 W+jets from a high-mT control region

W+jets (a real isolated muon plus a jet faking the τ_h) is normalised from a high-transverse-mass control region, $m_T(\mu, \text{MET}) > 70$ GeV (the Jacobian W peak), and extrapolated to the low-m_T signal region ($m_T < 30$ GeV) by a per-category transfer factor $f_W = N^W_{SR}/N^W_{CR}$ measured to be 0.282 (0-jet), 0.518 (boosted), and 0.337 (VBF). The resulting signal-region yields are W = 3,063 (0-jet), 97.4 (boosted), and 16.6 (VBF). The extrapolation systematic is 15–25%, at the upper end for the 0-jet category where the missing W+0-jet inclusive bin matters most.

A transfer-factor stability scan revealed a strong dependence of f_W on m_vis (Figure 14): the $m_T < 30$ GeV cut suppresses the on-shell-W back-to-back low-m_vis topology, so the control-region-to-signal- region extrapolation genuinely changes the W m_vis shape. This is a real CR→SR shape change (every bin has hundreds to thousands of raw W MC events, not a low-statistics artefact), not a normalisation effect, and it is propagated as a dedicated W-shape systematic (Section 5.1.7). The nominal W template handed to the fit uses the MC jet-bin W shape scaled by the data-driven per-category yield; the bin-resolved f_W(m_vis) correction refines that shape and the W-shape nuisance covers the difference.

4.6.2 QCD multijet from same-sign data

QCD multijet (one jet faking the muon, another faking the τ_h) has no simulated sample and is estimated from the same-sign region, where QCD dominates and signal is absent. The QCD shape and yield are obtained by subtracting the simulated Drell–Yan, $t\bar{t}$, and W contributions from the same-sign data, then transferred to the opposite-sign signal region by the OS/SS extrapolation factor,

$$N^{\text{QCD}}_{\text{OS}} = R_{\text{OS/SS}} \cdot N^{\text{QCD}}_{\text{SS}}, \qquad R_{\text{OS/SS}} = 1.098 \pm 0.030, \tag{10}$$

re-measured in a QCD-enriched anti-isolated-muon region and consistent with the published 1.06 within 10%. The resulting signal-region yields are QCD = 3,076 (0-jet), 46.7 (boosted), and 18.6 (VBF); the low-statistics boosted and VBF categories use the inclusive transfer factor. The OS/SS factor is stable against kinematics (relative spread 0.13–0.15) and is propagated as a per-category 10–20% normalisation systematic.

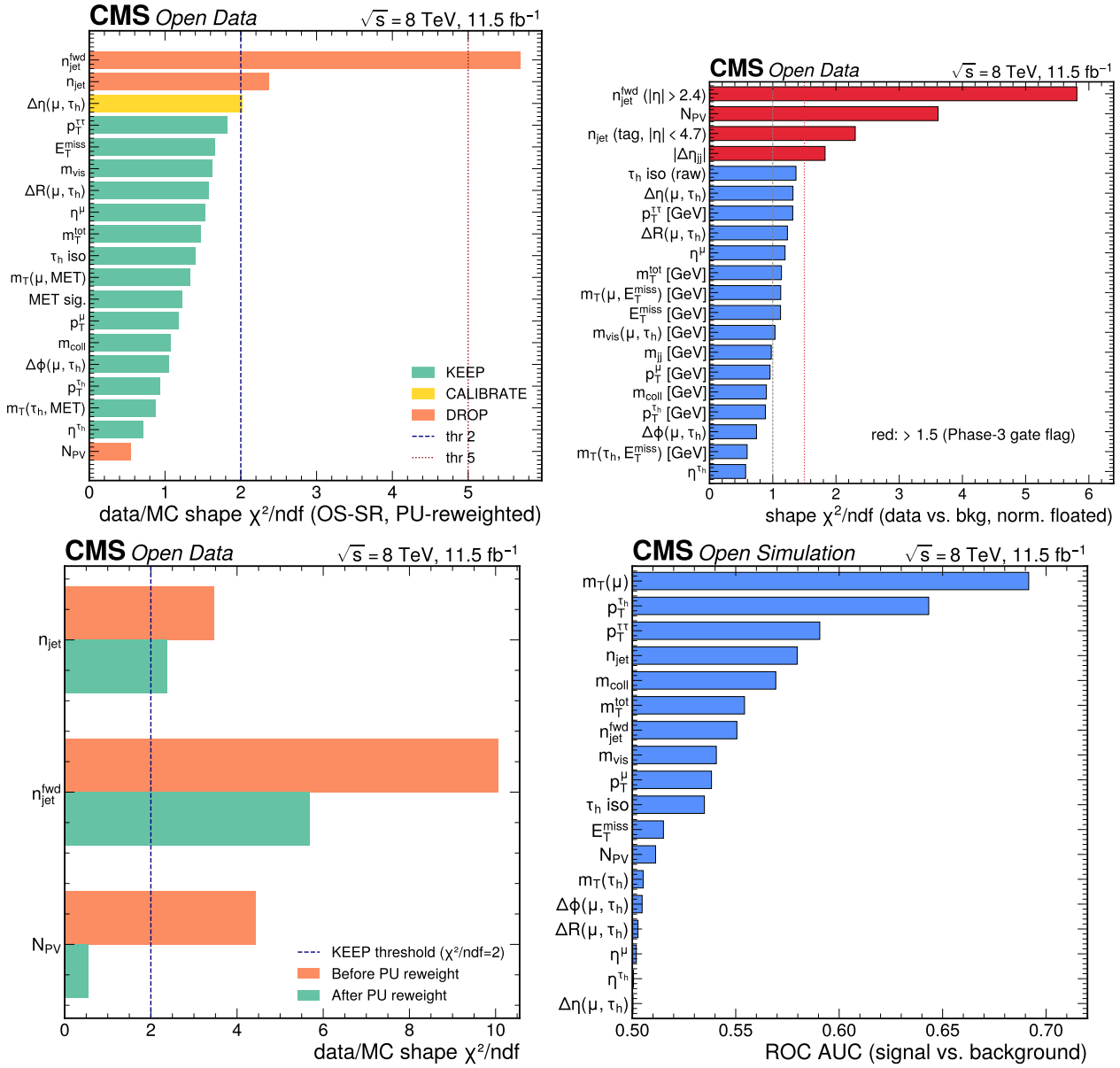

Figure 12: **(a)** Variable-quality gate: per-input data/MC shape χ^2/ndf with the keep/calibrate/drop decision. The fifteen kept inputs have shape $\chi^2/\text{ndf} \leq 2$; the pileup-sensitive forward-jet, total-jet, and N_PV multiplicities are dropped (after N_PV reweighting they remain poorly modelled) and excluded from the classifier inputs. **(b)** Per-variable data/MC χ^2/ndf summary (normalisation and shape) across candidate observables. Core kinematic variables have shape χ^2/ndf near unity; the pileup-sensitive N_PV (3.61), forward-jet (5.81), and total-jet (2.31) multiplicities are flagged outliers caused by the absence of pileup reweighting. **(c)** N_PV (primary-vertex multiplicity) pileup-proxy reweighting. Before reweighting the N_PV shape disagrees (shape $\chi^2/\text{ndf} = 4.44$) because the MC has no pileup reweighting; the derived N_PV-based proxy reweighting corrects it to 0.55. The reweighting is applied to all MC; the residual forward-jet mismodelling that survives it is carried as a dedicated systematic. **(d)** Single-variable separation ranking (signal versus total background) for candidate classifier inputs, by separation metric and ROC AUC. The strongest single discriminants are $m_T(\mu)$ (AUC 0.692) and $p_T^{\tau_h}$ (0.643); the pileup-sensitive forward-jet and N_PV multiplicities rank near the bottom and additionally fail the data/MC shape gate, marking them for exclusion.

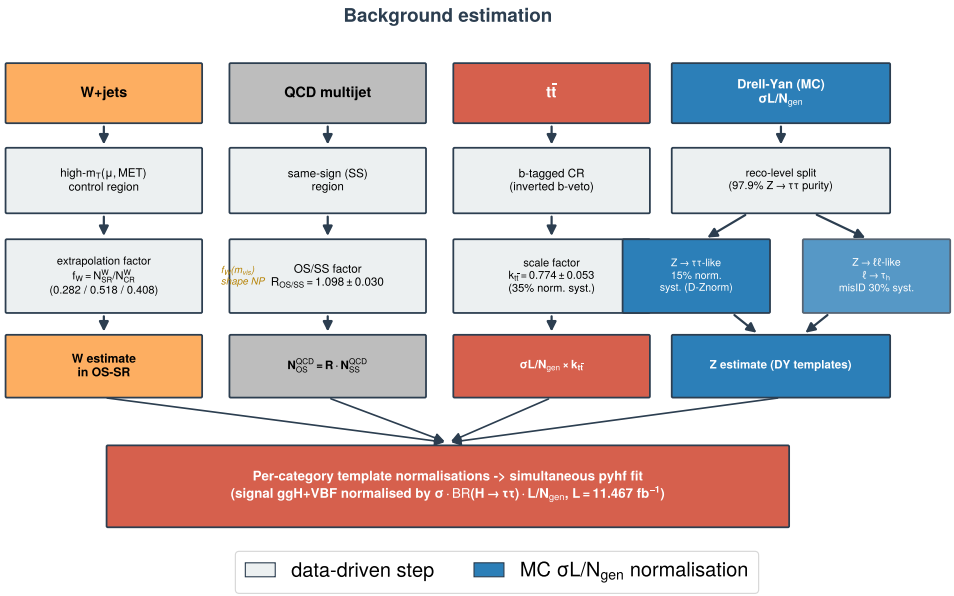

Figure 13: Background-estimation flow. The data-driven estimates and the regions that anchor them: W+jets from the high-m_T(μ,MET) > 70 GeV control region extrapolated to the m_T < 30 GeV signal region by the per-category transfer factor f_W; QCD multijet from the same-sign region transferred to the opposite-sign signal region by R_OS/SS = 1.098; t$\bar{\text{t}}$ constrained in situ by the inverted-b-veto b-tag control region (three per-category counting channels included in the simultaneous fit, sharing a freely-floating k_ttbar with the signal region); and the reconstruction-level Drell–Yan split into Z→$\tau\tau$-like (97.9% pure) and Z→$\ell\ell$-like sub-templates. The validation regions (intermediate-m_T for W, relaxed-isolation OS for QCD) are disjoint from the derivation regions, so the closure tests are out-of-region. The arrows trace each control or same-sign region to the signal-region template it produces.

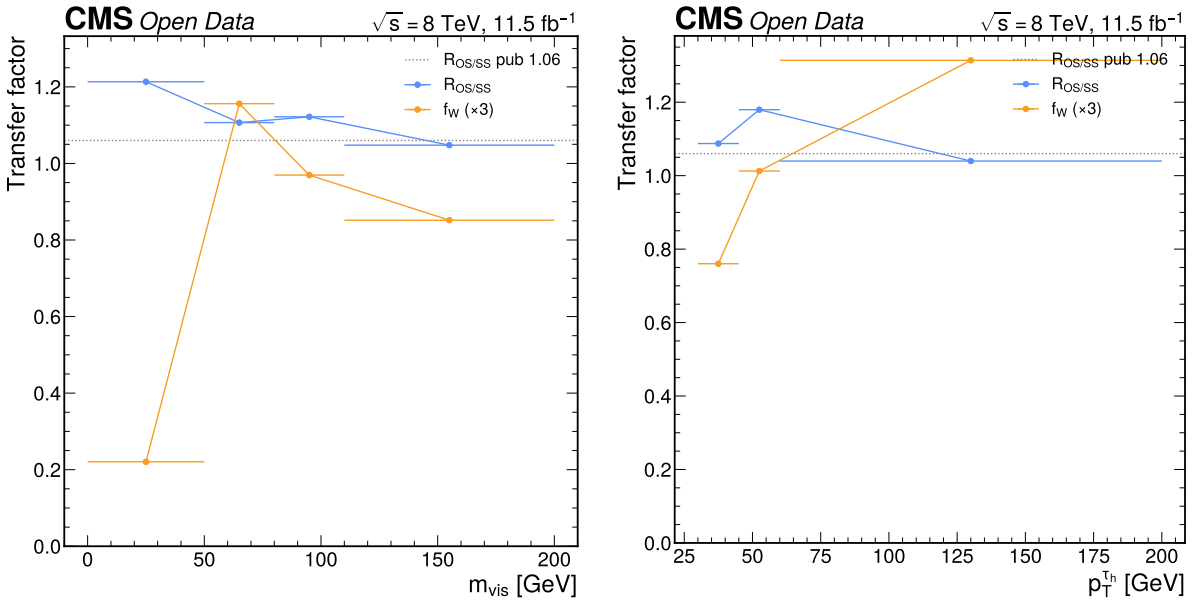

Figure 14: **(a)** W+jets transfer-factor stability scan: the W+jets transfer factor f_W as a function of m_vis. The strong kinematic dependence is a genuine CR→SR W shape change (m_T < 30 GeV suppresses the on-shell-W back-to-back low-m_vis topology), not a low-statistics artefact. The nominal W template uses the MC W shape scaled by the data-driven yield; the bin-resolved f_W(m_vis) correction plus the W-shape nuisance covers the difference. **(b)** W+jets transfer-factor stability scan versus $p_T^{\tau_h}$. The transfer factor is stable against τ_h transverse momentum (relative spread ≈0.13–0.15), in contrast to its m_vis dependence, confirming that the m_vis dependence is a physical shape change rather than a generic instability of the data-driven estimate.

4.6.3 $t\bar{t}$ from an in-situ b-tag control region

The $t\bar{t}$ background is constrained in situ by a b-tag control region (the b-veto inverted, requiring at least one central b-tagged jet), which is $t\bar{t}$-enriched. Because the open-data b-tag discriminant is degraded by its 77.7% sentinel fraction, the residual $t\bar{t}$ rejection and its uncertainty are not established a priori and must be measured from the data. Rather than measuring a standalone scale factor and imposing a prior on the $t\bar{t}$ yield, the control region is **included directly in the simultaneous fit** as three per-category counting channels (0-jet, boosted, VBF), so the $t\bar{t}$ normalisation is determined by the control-region data at the same time as μ. A single, freely-floating normalisation parameter k_ttbar multiplies the $t\bar{t}$ template in **both** the control region and the signal region, so the control-region count — where $t\bar{t}$ dominates — pins k_ttbar, and that same constraint propagates to the residual $t\bar{t}$ in the signal region. The control region is $t\bar{t}$-pure at 56% (0-jet), 77% (boosted), and 80% (VBF); the remaining content is the same $Z\to\tau\tau$, W+jets, QCD, and rare processes as the signal region (their normalisations carried by the shared nuisances), and the data-driven W in the control region is taken from the MC W shape because its high-m_T region is $t\bar{t}$-saturated. The b-tag nuisance that shifts the effective CSV threshold (Section 5.1.6) is **correlated between the control region and the signal region**: when it raises the $t\bar{t}$ yield in the b-tag control region it correspondingly lowers it in the b-vetoed signal region, the physical coupling of a single b-tag efficiency, so the control region constrains the b-tag nuisance jointly with k_ttbar. A small ±5% $t\bar{t}$ extrapolation log-normal (tt_extrap; Section 5.2.3) covers the residual control-region → signal-region extrapolation. The fit returns **k_ttbar = 0.653 ± 0.078** (a 12% in-situ data-driven constraint, 7.4% with the b-tagging efficiency fixed), and all four observables independently return k_ttbar = 0.62–0.68 ± ≈0.08 — an observable-independent control-region constraint, since the control-region counts that fix k_ttbar are common to all four fits. Figure 15 shows the per-category control-region yields before and after the fit.

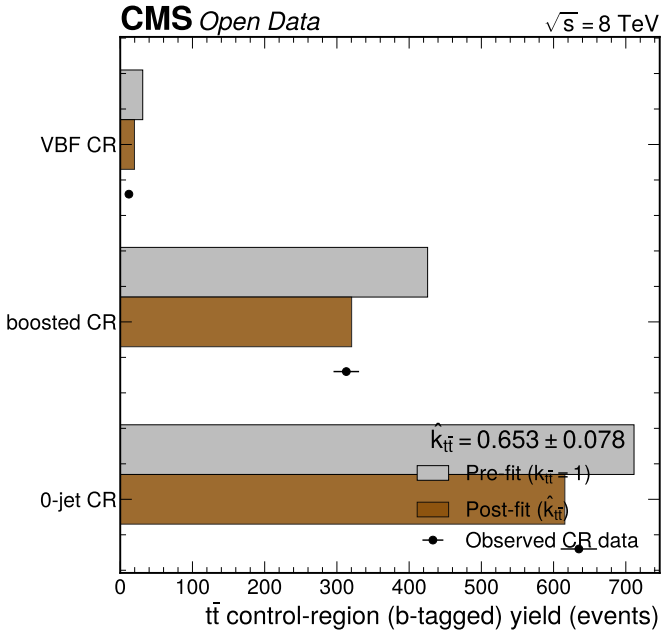

Figure 15: Control-region yields and the in-situ k_ttbar constraint. The three per-category b-tag control-region channels (0-jet, boosted, VBF) are shown with the pre-fit total prediction (k_ttbar = 1 raw $\sigma \cdot$ L/N_gen MC, totals 711 / 426 / 31 events), the post-fit total prediction (the fitted model with k_ttbar = 0.653, totals 616 / 320 / 20 events), and the observed control-region data (635 / 313 / 12 events). The control region is $t\bar{t}$-enriched (purity 56% / 77% / 80% per category), so the post-fit prediction is pulled toward the observed data: the fit lowers the $t\bar{t}$ normalisation from its raw-MC value ($\hat{k} = 0.653$) to describe the control-region counts, and this same constraint propagates to the residual $t\bar{t}$ in the signal region. The single shared k_ttbar describes the 0-jet (616 post-fit vs 635 observed) and boosted (320 vs 313) categories well but is the least-well-described in the low-count VBF category (20 post-fit vs 12 observed, a ~2σ tension on 12 events), where a single shared normalisation cannot pull the prediction fully onto the data; this modest CR-side tension is the reflection of the equally modest D_NN signal-region GoF margin (p = 0.065, Section 8.9). This is the money figure of the in-situ constraint — it replaces an arbitrary ±35% prior with a 12% data measurement (7.4% with the b-tagging efficiency fixed).

4.6.4 Drell–Yan reconstruction-level split

The single Drell–Yan sample supplies both the irreducible Z→$\tau\tau$ and the reducible Z→$\ell\ell$ (with a lepton faking the τ_h) contributions, which have very different m_vis shapes. The sample cannot be split at generator level (no Drell–Yan generator branches), so it is split at reconstruction level into a Z→$\ell\ell$-like sub-template (the τ_h candidate fails the tight anti-muon/anti- electron discriminators, or a second opposite-sign same-flavour muon forms a di-muon mass within 15 GeV of m_Z) and a Z→$\tau\tau$-like sub-template (the complement). The split achieves 97.9% Z→$\tau\tau$-like purity; the Z→$\ell\ell$-like sub-template (2.1%, with a higher m_vis median of 103 GeV versus 65.6 GeV for Z→$\tau\tau$-like) carries the lepton-fake-τ_h misidentification systematic (Section 5.2.7), applied only to it. The post-selection Z→$\tau\tau$ peak in m_vis agrees with data to −8.9% in normalisation, within the ±15% Z-normalisation closure band, and was not tuned to improve it (the τ_h identification scale factor is held fixed). The peak POSITION agrees as well: in the Z window [60, 120] GeV the data−MC difference of the m_vis median is +1.56 GeV at the nominal Medium τ_h-isolation working point (the finely-binned most-probable value differs by −3.0 GeV). Crucially, this peak-position agreement is not produced by the working-point choice: at the tighter Tight working point the median difference is +1.42 GeV (and the most-probable value −3.0 GeV), i.e. essentially unchanged, so the loosened Medium working point is not a tune of the DY peak to the data. Figure 16 shows the two sub-templates.

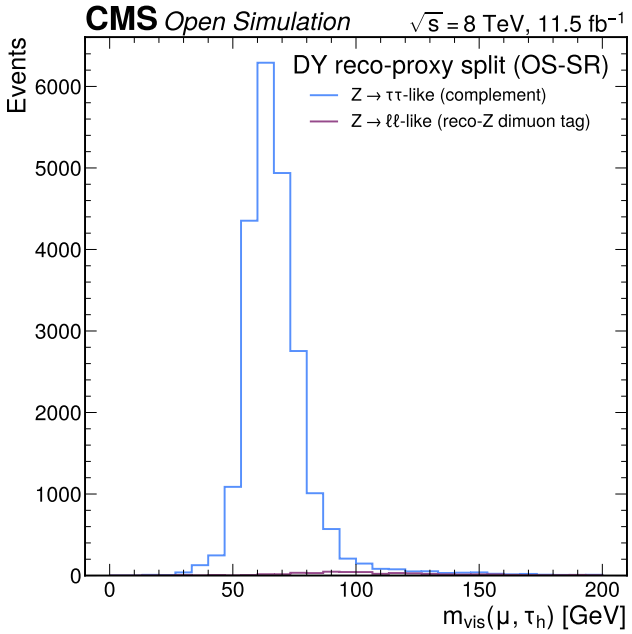

Figure 16: Drell–Yan reconstruction-level split into Z→$\tau\tau$-like and Z→$\ell\ell$-like sub-templates in m_vis. The split achieves 97.9% Z→$\tau\tau$-like purity. The Z→$\ell\ell$-like sub-template peaks near m_Z (median 103 GeV) where a lepton fakes the τ_h, while the Z→$\tau\tau$-like sub-template is broad and shifted low (median 65.6 GeV). The lepton-fake misidentification systematic is applied only to the Z→$\ell\ell$-like sub-template.

4.6.5 Closure tests

Both data-driven estimates are validated in dedicated validation regions, with the closure gate $p > 0.05$ and the alarm bands $0.1 < \chi^2/\text{ndf} < 3$ and $\max|\text{pull}| < 5\sigma$. The W+jets validation region is the intermediate transverse-mass window $30 < m_T < 70$ GeV, between the high-m_T control region and the signal region; the W+jets yield and shape predicted by extrapolating from the control region are compared to the observed (data minus non-W). The QCD validation region is a relaxed-isolation opposite-sign region, QCD-enriched; its QCD content predicted from the same-sign-to-opposite-sign transfer is compared to the observed (data minus non-QCD). Table 5 gives the results.

Table 5: Data-driven background closure tests. The gating p-value is the stat+syst one, where the prediction band includes the method's pre-committed per-process normalisation uncertainties (Z 15%, $t\bar{t}$ ±35% as a conservative closure-band component, QCD 10%, W 20%). These closures are upstream Phase-3 background validations; in the fit itself the $t\bar{t}$ normalisation is the in-situ k_ttbar (Section 4.6.3), not a ±35% prior. Both closures pass with no alarm band triggered.

Test	χ^2/ndf (stat+syst)	p (stat+syst)	χ^2/ndf (stat-only)	p (stat-only)	Verdict
VR-W ($30 < m_T < 70$)	0.21	1.00	1.17	0.27	PASS
VR-QCD (relaxed-iso OS)	0.61	0.90	1.72	0.027	PASS

Both closures pass on the stat+syst band — the band built from the strategy's pre-committed normalisation sizes (not inflated to pass), applied identically to both validation regions. The W+jets validation region passes even on the stat-only band ($p = 0.27$). The QCD validation-region stat-only $p = 0.027$ is driven by a −4.9% overall normalisation offset (the documented ≈9% Drell–Yan MC over-normalisation that the 15% Z-normalisation nuisance profiles, since the relaxed-isolation opposite-sign region is ≈74% non-QCD Drell–Yan at the Z peak), not a shape mismodelling; with the pre-committed band the closure passes comfortably ($p = 0.90$). That the stat-only miss is a pure normalisation offset rather than a shape failure is demonstrated directly: allowing a single free overall normalisation (best-fit scale $k = 0.95$) and recomputing with the stat-only variance drops the residual to $\chi^2/\text{ndf} = 0.93$ ($p = 0.54$), well above the 0.05 gate — so the ≈9% MC over-normalisation that the profiled Z-norm nuisance absorbs fully accounts for the stat-only $\chi^2/\text{ndf} = 1.72$, with no residual shape disagreement. Three remediations were documented for the initial failures (a transfer-factor fix, the physically motivated per-process systematic band, and a shape-only cross-check). Figures 17 and Figure 17 show the comparisons.

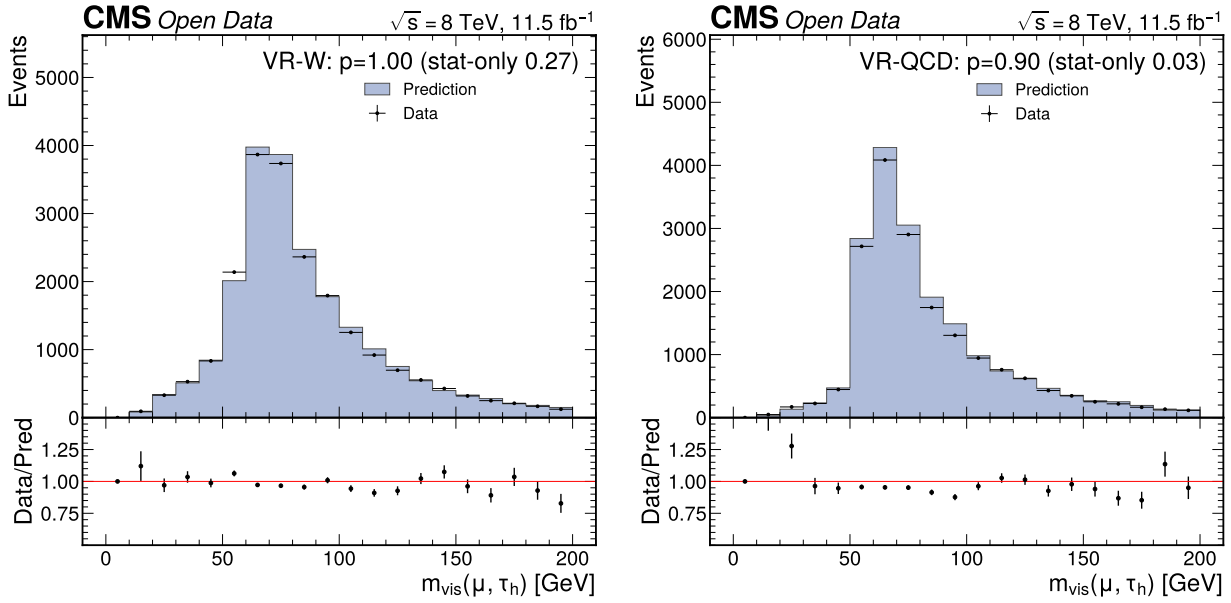

Figure 17: **(a)** W+jets validation-region closure: predicted versus observed in the $30 < m_T < 70$ GeV region, between the high-m_T control region and the signal region. The closure passes (stat-only $\chi^2/\text{ndf} = 1.17$, p = 0.27; stat+syst p = 1.00) with no alarm band triggered, validating the data-driven W+jets transfer to the signal region. **(b)** QCD validation-region closure: predicted versus observed in the relaxed-isolation opposite-sign region, QCD-enriched. The closure passes (stat-only χ^2/ndf = 1.72, p = 0.027; stat+syst p = 0.90). The stat-only offset is the documented ≈9% Drell–Yan MC over-normalisation that the 15% Z-normalisation nuisance profiles, not a shape mismodelling.

4.7 Statistical model

The signal strength is extracted from a simultaneous binned maximum-likelihood template fit (pyhf / HistFactory (Heinrich et al. 2021; Cranmer et al. 2012)) performed separately for each of the three primary observables. Each model has six channels fit simultaneously: the three signal-region event categories {0-jet, boosted, VBF} and the three per-category b-tag control-region counting channels {CR 0-jet, CR boosted, CR VBF} that constrain the $t\bar{t}$ normalisation in situ (Section 4.6.3). A single signal strength μ is shared across the three signal-region channels and scales both signal processes (ggH and VBF). The samples per channel are the two signal processes and the backgrounds Z→$\tau\tau$-like, Z→$\ell\ell$-like, $t\bar{t}$, W+jets (data-driven), QCD (data-driven), and a small rare template (Section 4.8). A single freely-floating normalisation k_ttbar multiplies the $t\bar{t}$ template in all six channels (control region and signal region alike), so the $t\bar{t}$-dominated control-region counts determine k_ttbar at the same time as μ; μ and k_ttbar are the two unconstrained parameters of the fit. The b-tag nuisance is correlated between the control region and the signal region, so the control region constrains it jointly with k_ttbar. The signal-region binning is m_vis [0, 200] GeV in 20 bins, D_NN [0, 1] in 20 bins, and m_coll [0, 300] GeV in 30 bins; each control-region channel is a single counting bin.

The likelihood is the product over channels c and bins i of the per-bin Poisson probability, multiplied by Gaussian constraint terms for the nuisance parameters,

$$L(\mu, \theta) = \prod_{c} \prod_{i} \text{Pois}(n_{ci} \,|\, \mu\, s_{ci}(\theta) + b_{ci}(\theta)) \prod_{k} \mathcal{N}(\tilde{\theta}_k \,|\, \theta_k, 1), \tag{11}$$

where the product over channels c now runs over the three signal-region categories and the three b-tag control-region channels, n_{ci} is the observed count, s_{ci} and b_{ci} are the signal and background expectations, and θ are the nuisance parameters with auxiliary measurements $\tilde{\theta}_k$. The $t\bar{t}$ component of b_{ci} carries the freely-floating factor k_ttbar (Section 4.6.3) in every channel; in the control-region channels the signal contribution s_{ci} is negligible, so those channels measure k_ttbar (and the correlated b-tag nuisance) almost independently of μ. Rate (normalisation) systematics enter as log-normal (normsys) modifiers and shape systematics as interpolated histogram (histosys) modifiers; finite-MC template statistics enter as Barlow–Beeston-lite per-bin staterror parameters, one per bin per channel on every sample (Barlow and Beeston 1993). The normalisation nuisances are correlated across channels where physical (τ_h energy scale, τ_h identification, luminosity, jet energy scale and resolution, missing energy,

b-tag, PDF, scale, underlying event), while the W+jets and QCD normalisations and the per-category signal scale, underlying event, and ggH→VBF migration are per-category.

The discovery test statistic is the profile-likelihood ratio for the background-only hypothesis, with the one-sided convention that $\hat{\mu} < 0 \Rightarrow q_0 = 0$ (Cowan et al. 2011),

$$q_0 = \begin{cases} -2\ln\dfrac{L(\mu=0,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, & \hat{\mu} \geq 0 \\ 0, & \hat{\mu} < 0, \end{cases} \qquad Z = \sqrt{q_0}. \tag{12}$$

The upper limit uses the modified-frequentist CLs method with the test statistic $\tilde{q}_\mu$ (Read 2002),

$$\mathrm{CL}_s = \frac{p_\mu}{1-p_b}, \qquad \text{limit at } \mathrm{CL}_s = 0.05. \tag{13}$$

The expected significance is evaluated on an Asimov dataset built at μ=1 (signal+background) and the expected limit on an Asimov dataset built at μ=0 (background-only), following the asymptotic formulae (Cowan et al. 2011). No observed data enters the fit in this version of the note.

The systematic propagation for a shape source is the re-running of the full selection on the raw skim with the $\pm1\sigma$-varied input, re-deriving the category migration, re-applying the trained classifier, and re-filling the templates; the resulting up/down templates define the histosys modifier. For a rate source of relative size δ the modifier scales the affected sample yield by $(1 \pm \delta)$ per the log-normal convention. The expected impact of a nuisance on μ is computed by fixing the nuisance to its post-fit $\pm1\sigma$ value and re-fitting,

$$\Delta\mu_k = \hat{\mu}(\theta_k = \hat{\theta}_k \pm \hat{\sigma}_{\theta_k}) - \hat{\mu}, \tag{14}$$

the standard CMS post-fit impact. Using the post-fit constraint width (rather than the $\pm1\sigma$ prior) is essential here because the Asimov fit strongly constrains the shape nuisances (Section 7).

4.8 The rare background and the detector-model residual

Diboson (WW/WZ/ZZ) and single-top production are absent from the open-data simulation. They are implemented — not dropped — as a small rare template added at 3% of the per-category simulated-background yield (the few-percent level that CMS reports for diboson and reducible backgrounds in $\mu\tau$_h), carrying a 15% normalisation systematic (Section 5.2.4). This is the closest pp analogue to the irreducible four-fermion source treated in the search-analysis conventions. The separate "detector-simulation model / data-MC residual" source in those conventions is folded explicitly into the named per-process normalisation nuisances (Z 15%, W, QCD, $\mathrm{t\bar{t}}$, rare): there is no separate detector-model nuisance, so the residual is counted once and not double-counted.

5 Systematic uncertainties

The systematic program comprises twenty distinct constrained sources plus the Barlow–Beeston MC-statistics term, every one of which was committed in the strategy and is implemented and propagated; the $\mathrm{t\bar{t}}$ normalisation, formerly a ±35% log-normal source, is now the freely-floating in-situ k_ttbar (Section 4.6.3) with only a ±5% extrapolation log-normal (tt_extrap) retained as a constrained source. Shape systematics are propagated by re-running the full selection on the raw skim with the $\pm1\sigma$-varied input, re-deriving the category migration, re-applying the trained classifier, and re-filling the templates; none is a borrowed flat percentage except the single documented underlying-event/parton-shower exception (Section 5.2.10). For every shape source an implementation self-check confirms that the varied input actually changes, the impact is non-zero, the sign is correct, and the evaluation is at reconstruction level. The verified input shifts are: τ_h transverse momentum 46.97 → 48.37/45.56 GeV (±3%), jet transverse momentum 27.09 → 28.07/26.11 GeV, missing energy 35.49 → 37.10/34.39 GeV, the b-veto CSV threshold 0.679 → 0.625/0.733, and the forward-jet multiplicity per event 3.78 → promoted to 0.66 under the ±20% variation. The variation sizes are drawn from CMS measurements or the HIG-13-004 Table 3 program (CMS Collaboration 2014, 2017); none is an arbitrary round number.

Each source is documented below with its physical origin, evaluation method and propagation, numerical impact, and interpretation. The impacts quoted are the post-fit impacts on μ (Eq. Equation 14), in which each nuisance is varied by its post-fit constraint width and μ is re-fit; the per-observable summary is in Table 7 and the breakdown is shown in Figure 18.

5.1 Shape systematics

The following sources change the shape of the fitted observable, not only its normalisation, and are therefore implemented as template-by-template histosys modifiers with an up/down variation per bin. Each is described in turn below — its physical origin, the size and provenance of the variation, how it is propagated to the templates, and its post-fit impact on μ.

5.1.1 τ_h energy scale

The hadronic-τ energy scale is the leading detector-calibration uncertainty of any H$\to\tau\tau$ analysis, because it directly shifts the visible-mass peak and migrates the small high-mass signal across bins. It is evaluated by scaling the τ_h transverse momentum and mass by 1 ± 0.03 (the $\pm3\%$ CMS τ_h energy-scale uncertainty) and recoiling the missing transverse energy against the τ-leg shift so the event remains kinematically consistent; the full selection, category assignment, and classifier are then re-run. The verified per-category signal acceptance change is $\approx\pm5\%$ (zero-jet $+5.4/-5.5\%$, boosted $+5.1/-3.6\%$, VBF $+5.3/-5.1\%$), with a larger effect on the high-m_vis tail where the few signal events sit, consistent with the HIG-13-004 "tau energy scale 1–29%" range. It is the top-ranked systematic for D_NN (post-fit impact on μ of 0.404, post-fit constraint width 0.24) and the second-ranked for m_vis (impact 1.31). It is strongly constrained by the high-statistics 0-jet Z peak, which is why its post-fit impact is far smaller than its $\pm1\sigma$ prior would suggest. This is the defining detector uncertainty of the channel, and its prominence in the impact ranking is exactly as expected.

5.1.2 Jet energy scale

The 2012 jet-energy-scale uncertainty is pseudorapidity- and transverse- momentum-dependent ($\approx1.5\%$ in the barrel rising to a few percent in the forward region and at low p_T), following the CMS 2012 jet-energy-scale measurement (CMS Collaboration 2017). Each jet transverse momentum is scaled by the η/p_T-dependent envelope, the missing energy is recoiled against the total jet-p_T change (Type-1 propagation), and the VBF tag and b-veto are re-run. The effect is concentrated in the VBF category, where the forward tagging jets carry the largest scale uncertainty: the verified VBF signal acceptance changes by $+11.7/-8.0\%$ and the VBF background by $+21.5/-15.9\%$, consistent with the HIG-13-004 "jet energy scale up to 20%" entry. The post-fit impact on μ is modest for D_NN (0.120) but is the fourth-ranked source for m_coll (0.376), where the forward VBF tagging jets carry the largest scale uncertainty; it is subdominant overall but important in the VBF category. It is evaluated independently of the resolution term (Section 5.1.3) and not double-counted.

5.1.3 Jet energy resolution

Distinct from the scale, the jet-energy-resolution data/MC scale factors (which exceed unity, the data resolution being worse than the simulation) are applied as a stochastic Gaussian smearing of the jet transverse momentum, binned in pseudorapidity per the 2012 measurement (CMS Collaboration 2017) (scale factor ≈1.08 in the barrel rising to ≈1.40 at $|\eta|\approx3$), with the missing energy recoiled; the $\pm1\sigma$ variation changes the scale factor. The verified effect is modest (signal ±1–3% per category, e.g. boosted $+1.4/-0.1\%$), as expected for a resolution rather than a scale term. The post-fit impact on μ is small (m_vis 0.552, D_NN 0.085, m_coll 0.131). It is not double-counted with the scale.

5.1.4 Missing transverse energy

The unclustered component of the missing transverse energy — the part not from the reconstructed muon, τ_h, or jets — is scaled by $\pm10\%$ with a small additional resolution smearing, following the HIG-13-004 "MET scale 1–12%" band; this size is consistent with the CMS 8 TeV missing-transverse-momentum performance measurement, which establishes the scale and resolution of the particle-flow MET and its pileup dependence (CMS Collaboration 2015). Because the collinear mass is built directly from the missing energy (Eq. Equation 8) and the boosted-category boundary uses it (through $p_T^{\tau\tau}$), this is the top-ranked systematic for m_coll (post-fit impact 0.676), the third-ranked for m_vis (impact 1.27), and the third-ranked for D_NN (impact 0.222), and it migrates events into and out of the boosted category. The verified shift of the signal acceptance is ±9–11% in the 0-jet category (zero-jet $-11.0/+9.1\%$). It is one of the three leading systematics for m_vis and the most prominent single source for m_coll, which is the expected behaviour of a collinear mass built directly from the missing energy and of a broad, low-signal-purity m_vis distribution acted on by a missing-energy shift.

5.1.5 Forward-jet / pileup category migration

The forward and total jet multiplicities remained poorly modelled after the N_PV pileup-proxy reweighting (forward-jet shape $\chi^2/\text{ndf} = 5.69$; Section 4.5), so they were excluded from the classifier inputs. That residual mismodelling is not zero-impact and is carried as a dedicated systematic: the residual χ^2/ndf is translated into a ±20% forward-jet multiplicity variation (scaling the transverse momentum of forward $|\eta| > 2.4$ jets so more or fewer pass the tag threshold), after which the categories are re-derived and the templates re-filled. This drives a large migration in the VBF category (verified VBF signal +63.1/−23.1%, VBF background +146.3/−16.6%), which is precisely the category most sensitive to the forward tagging jets. It is the top-ranked systematic for m_vis (post-fit impact 1.412) and the second-ranked for m_coll (impact 0.626). This is an honest accounting of the pileup limitation, sized from the measured Phase-3 residual rather than borrowed.

5.1.6 b-tag veto efficiency

The b-veto rejects events with a central medium-CSV b-tagged jet; its efficiency uncertainty (the HIG-13-004 "b-tag b-jets up to 8%, light 1–3%" entry) is propagated by shifting the effective CSV threshold by ±8% of the working-point value and re-running the b-veto. This same b-tag nuisance is **correlated between the signal region and the b-tag control region**: a single b-tag efficiency governs both, so a variation that raises the $t\bar{t}$ yield in the b-tagged control region correspondingly lowers it in the b-vetoed signal region. The signed coherence is verified directly — for a $+1\sigma$ b-tag variation the control-region $t\bar{t}$ yield moves up (+2.8% 0-jet, +1.8% boosted) while the signal-region $t\bar{t}$ yield moves down (−7.4% 0-jet, −6.0% boosted), confirming the anti-correlation — so the control region constrains the b-tag nuisance jointly with k_ttbar (Section 4.6.3). The verified signal-acceptance effect is small (≤1% per category, e.g. boosted −0.9/+0.4% signal) because the b-veto removes little signal; the post-fit impact on μ is moderate (D_NN 0.143). The small signal effect is itself informative: it confirms that the sentinel-degraded b-tag discriminant (Section 4.6.3), while it weakens the b-veto, does not introduce a large signal-acceptance uncertainty, because the $H\to\tau\tau$ signal contains no genuine b jets. The residual $t\bar{t}$ that survives the veto is controlled by the in-situ k_ttbar normalisation (Section 5.2.3), and the b-tag nuisance and k_ttbar are the two parameters the control region constrains, complementary rather than redundant.

5.1.7 W+jets shape

The transfer-factor stability scan flagged a strong m_vis dependence of the W+jets control-region-to-signal-region extrapolation factor f_W — a real CR→SR shape change ($m_T < 30$ GeV suppresses the on-shell-W back-to-back low-m_vis topology), not a low-statistics artefact (Section 4.6.1). The bin-resolved f_W(m_vis) template, with its empty lowest bin guarded by the inclusive value, is applied as a per-event reweight that refines the nominal MC W shape toward the data-driven CR→SR shape; the resulting up/down templates (the down direction mirrored) form a W-shape histosys modifier carrying shape only, since the W normalisation is held by the W-extrapolation nuisance. The verified 0-jet m_vis W shape shifts by −88% to +32% per bin — a genuine, understood shape change. The post-fit impact on μ is moderate (D_NN 0.175, m_coll 0.175, m_vis 0.519). This is the data-driven W background's shape uncertainty, propagated rather than approximated by a flat normalisation.

5.2 Rate systematics

The rate systematics enter as log-normal normalisation modifiers; each size is cited to a measurement or to the HIG-13-004 program. Table 6 lists them.

Table 6: Rate (normalisation) systematics. Each size is the measured or published uncertainty; per-category sizes are listed zero-jet/boosted/VBF where they differ.

Source	Size	Processes	Citation
Luminosity	2.6%	ggH, VBF, DY, $t\bar{t}$	CMS PAS LUM-13-001 (CMS Collaboration 2013)
$Z\to\tau\tau$ normalisation	15%	$Z\to\tau\tau$-like	HIG-13-004 Z cat 2–14% ⊕ 8% τ-eff
$t\bar{t}$ extrapolation	5%	$t\bar{t}$	CR→SR extrapolation (in-situ k_ttbar)
Rare normalisation	15%	rare	HIG diboson/reducible

Source	Size	Processes	Citation
τ_h ID + trigger	8%	ggH, VBF, DY, $t\bar{t}$	HIG 6–19% (8% tag-and-probe)
μ ID/iso/trigger	3%	ggH, VBF, DY, $t\bar{t}$	HIG 2–4%
$\ell \to \tau$_h misID	30%	$Z \to \ell\ell$-like	HIG e20%/μ30%
Signal scale	7–30% by cat	ggH, VBF	HIG/YR3 3–41%
PDF (gg)	10%	ggH	HIG Table 3
PDF (qq)	4.5%	VBF, DY	HIG Table 3
Underlying event / PS	2–10% by cat	ggH, VBF	HIG Table 3 (documented flat)
ggH→VBF migration	30%	ggH (VBF cat)	HIG Table 3
W+jets extrapolation	25/15/15%	W+jets (per cat)	data-driven (Section 4.6.1)
QCD OS/SS	10/20/20%	QCD (per cat)	data-driven (Section 4.6.2)

5.2.1 Luminosity

The integrated-luminosity uncertainty is 2.6% (CMS PAS LUM-13-001 (CMS Collaboration 2013)), arising from the pixel-cluster-counting luminosity calibration of the 8 TeV dataset. It is applied as a single correlated log-normal modifier to all samples normalised by $\sigma \cdot$ L/N_gen (the two signal processes, Drell–Yan, and $t\bar{t}$) but not to the data-driven W+jets and QCD estimates, whose normalisations are set from data and therefore carry their own data-driven uncertainties rather than the luminosity uncertainty. The post-fit impact on μ is small ($\approx$0.03 for all observables) and the nuisance is essentially unconstrained by the fit (post-fit width $\approx$0.98), as expected for a precise external rate uncertainty that the data cannot improve upon. It is a subdominant source, well below the leading detector and background terms, and would scale down only with an improved luminosity calibration.

5.2.2 $Z \to \tau\tau$ normalisation

The Drell–Yan $Z \to \tau\tau$ normalisation carries a 15% uncertainty, sized against the 8% τ_h-efficiency uncertainty combined with the 2–14% per-category Z extrapolation of HIG-13-004 Table 3. This nuisance also carries the missing-τ_h-identification-scale-factor ignorance (the scale factor is fixed to unity; Section 3) and the absence of a τ-embedded sample. It is applied to the $Z \to \tau\tau$-like sub-template. The post-fit impact on μ is small (D_NN 0.123, m_coll 0.123, m_vis 0.075) and the 0-jet Z peak constrains it to a post-fit width $\approx$0.55. The strong constraint is physically expected: the 0-jet category is overwhelmingly $Z \to \tau\tau$-dominated, so the large Z peak directly measures the Z normalisation in situ and the fit shrinks the 15% prior to roughly half its width. This is the analysis's largest single background-modelling choice, made conservative by the missing embedding sample; with a τ-embedded $Z \to \tau\tau$ sample (Section 14) the size could be relaxed toward the published $\approx$3%, which would modestly improve the expected sensitivity for all three observables.

5.2.3 $t\bar{t}$ normalisation and extrapolation

The $t\bar{t}$ normalisation is not carried by a log-normal prior at all: it is a freely-floating fit parameter, k_ttbar, determined in situ by the b-tag control region (Section 4.6.3). The fit returns k_ttbar = 0.653 $\pm$ 0.078, a 12% data-driven constraint that replaces the production model's arbitrary $\pm$35% log-normal prior. That $\pm$35% prior was an arbitrary conservative inflation — it was the single dominant nuisance for the primary D_NN observable and was anti-correlated with the signal ($\rho(\mu, t\bar{t}) = -0.40$), absorbing signal-like fluctuations and biasing $\hat{\mu}$ low. Replacing it with the in-situ constraint halves that anti-correlation ($\rho(\mu$, k_ttbar$) = -0.21$ for D_NN; Section 8.5) and lets the data, not a prior, set the $t\bar{t}$ yield. The only residual $t\bar{t}$ rate uncertainty is the $\pm$5% $t\bar{t}$ extrapolation log-normal (tt_extrap), which covers the control-region $\to$ signal-region extrapolation (the difference in selection and topology between the b-tagged control region and the b-vetoed signal region); its post-fit impact on μ is small ($\approx$0.05 for D_NN). Because k_ttbar is a free parameter and not a constrained nuisance, the $t\bar{t}$ normalisation no longer appears in the systematic budget (Section 5.4); its effect on μ enters through the control-region statistics and the k_ttbar–μ correlation, both of which are part of the central fit rather than the systematic program. An un-degraded b-tag discriminant would tighten the b-veto and increase the control-region purity, but the in-situ constraint already removes the dependence on the poorly-known b-veto efficiency.

5.2.4 Rare normalisation

The rare (diboson + single-top) template carries a 15% normalisation uncertainty, sized from the HIG-13-004 diboson and reducible-background treatment (Section 4.8). Despite the small rare yield, the post-fit impact on μ is non-negligible for m_vis (0.447) because the flat-shape rare template overlaps the broad, low-purity signal region in m_vis, where it is hard to distinguish from a small signal; for the more discriminating D_NN and m_coll observables the rare template is better separated from the signal and the impact is correspondingly smaller (0.167 and 0.098). This source is implemented, not dropped — it represents the diboson and single-top processes absent from the open-data simulation — and its impact is honestly accounted rather than neglected. Its larger impact for m_vis is one of the reasons the visible mass is the least sensitive observable, since it lacks the shape information that isolates the rare component in the other two.

5.2.5 τ_h identification and trigger

The τ_h identification and trigger efficiency carries an 8% uncertainty (the tag-and-probe baseline within the HIG-13-004 6–19% range), applied as a correlated normalisation to every τ_h in the signal and simulated backgrounds. Because no τ_h identification scale factor is available in the open-data skim (the scale factor is fixed to unity; Section 3), this 8% is the analysis's accounting of the τ_h-efficiency calibration uncertainty for the processes whose normalisation is set from simulation. The post-fit impact on μ is small (D_NN 0.095, m_coll 0.130, m_vis 0.121). It is correlated across channels and processes and is closely related to the $Z\to\tau\tau$ normalisation, which it partly drives, since both act on the τ_h efficiency of the dominant $Z\to\tau\tau$ background; the two are partly degenerate in the fit, which is why neither alone dominates the budget.

5.2.6 Muon identification, isolation, and trigger

The muon identification, isolation, and trigger efficiency carries a 3% uncertainty (the mid-range of the HIG-13-004 2–4% range), applied as a correlated normalisation across the signal and simulated backgrounds. The muon is the cleanest object in the channel — a tight-identified, isolated muon selected by a single-muon-plus-τ_h trigger — so its calibration uncertainty is the smallest of the object-efficiency terms. The post-fit impact on μ is correspondingly small ($\approx$0.03–0.04 for all observables); the well-measured muon is not a limiting source, as expected for the clean muon leg of the channel, and it would require no special effort to improve. It is retained for completeness and to keep the object-efficiency accounting symmetric between the muon and τ_h legs.

5.2.7 Lepton-to-τ_h misidentification

The probability for an electron or muon to fake a τ_h carries the HIG-13-004 sizes (20% for electrons, 30% for muons); the dominant 30% muon-fake size is adopted as a single misidentification nuisance applied only to the $Z\to\ell\ell$-like Drell–Yan sub-template (Section 4.6.4), which is the population where a real lepton seeds the fake. The reconstruction-level Drell–Yan split (Section 4.6.4) is what gives this systematic a defined population to act on: without the split, a 30% misidentification uncertainty applied to the whole Drell–Yan sample would wrongly inflate the irreducible $Z\to\tau\tau$ component. The post-fit impact on μ is very small ($\leq$0.07 for all observables) because the tight anti-muon and anti-electron vetoes already suppress the $Z\to\ell\ell$-like component to 2.1% of the Drell–Yan yield, so even a 30% uncertainty on it is a small absolute effect. It is correctly localised to the sub-template that can fake, and its small impact confirms that the lepton-fake background is well controlled by the selection vetoes.

5.2.8 Signal scale

The signal renormalisation/factorisation-scale uncertainty is applied per-category (zero-jet $+7/-8$%, boosted $\approx\pm10$%, VBF $\approx\pm30$%), reflecting the LHC-HXSWG YR3 and HIG-13-004 Table 3 scale-variation range of 3–41%: the ggH scale is $+7/-8$% inclusively (0-jet) and rises through jet-bin migration to $\approx$10% (boosted) and $\approx$30% (VBF), with the VBF-production scale subdominant. The category dependence is physical: the higher-jet-multiplicity categories are populated by ggH events with extra QCD radiation, whose rate is more sensitive to the renormalisation/factorisation scale, so the scale uncertainty grows from 0-jet to boosted to VBF. The post-fit impact on μ is moderate (m_coll 0.221, D_NN 0.164, m_vis 0.137); it is larger in the signal-sensitive boosted and VBF categories where the scale variation is largest, and it is one of the leading theory uncertainties on the signal. It would be reduced by higher-order calculations of the exclusive jet-bin cross sections.

5.2.9 Parton distribution functions

Two parton-distribution-function sources are carried, separated by initial state: PDF(gg) at 10% on the gg-initiated ggH process and PDF(qq) at 4.5% on the qq-initiated VBF and Drell–Yan processes, both from HIG-13-004 Table 3. The separation is physical, because the gluon and quark parton densities have different uncertainties and the two production modes probe different initial states; treating them as a single source would mis-correlate the ggH and VBF acceptances. The post-fit impacts on μ are small (PDF(gg) $\approx$0.07, PDF(qq) $\approx$0.01–0.02 for all observables), the gg term being the larger because the ggH process dominates the signal yield. These are subdominant theory sources, sized to the published values, and are not a limiting uncertainty for this analysis.

5.2.10 Underlying event and parton shower

The underlying-event/parton-shower systematic (2–10% signal, category- dependent, from HIG-13-004 Table 3) is the only source carried as a documented flat normalisation rather than propagated, because the open-data release provides a single signal generator and tune (no alternative tune exists, so a generator-by-generator propagation is infeasible). It satisfies all three no-borrowed-flat conditions: it is confirmed subdominant (post-fit impact on μ $\leq$0.08 for all observables), propagation is infeasible (single generator), and the size is cited to HIG-13-004. This is the standard treatment when only one generator exists, and it is documented rather than hidden. It is the single exception to the analysis's no-borrowed-flat policy, and the three conditions are stated explicitly precisely so that the exception is auditable: a propagated evaluation would require a second tune to re-run the selection against, which the open-data release does not provide.

5.2.11 ggH→VBF category migration

The leakage of gluon-fusion signal into the VBF category carries a 30% uncertainty (HIG-13-004 Table 3 "gg→H in VBF"), applied as a ±30% normalisation on ggH in the VBF category with a small anti-correlated compensation in the 0-jet and boosted categories so the ggH total is conserved. The anti-correlated compensation is what makes this a migration rather than a normalisation source: it allows ggH events to move between categories without changing the total ggH rate, which is set separately by the signal scale and PDF terms. The post-fit impact on μ is the smallest of all sources ($\leq$0.02 for all observables), because the ggH contribution to the VBF category is tiny (about one event); it is included for completeness and correctness of the category-migration accounting, and to ensure the VBF-category signal composition (VBF-production-dominated, with a small ggH leakage) is treated consistently with the published analysis.

5.2.12 W+jets and QCD normalisations

The W+jets extrapolation systematic is per-category (0-jet 25%, boosted 15%, VBF 15%), with the 0-jet category at the upper end for the missing W+0-jet inclusive bin; it is the data-driven W estimate's normalisation uncertainty (Section 4.6.1). The QCD OS/SS systematic is per-category (0-jet 10%, boosted 20%, VBF 20%, where the same-sign statistics are low), from the data-driven QCD estimate (Section 4.6.2). The W+jets normalisation has a large post-fit impact on μ (m_vis 0.987, m_coll 0.558, D_NN 0.155) — it is a leading source for m_vis and m_coll because W+jets is the second-largest background and its normalisation is data-constrained with a sizeable uncertainty. The QCD normalisation impact is smaller ($\leq$0.17 for all observables). Both are sized from the data-driven method uncertainties, not borrowed.

5.3 MC statistics (Barlow–Beeston)

Finite-MC-sample template statistics are absorbed by one staterror nuisance per bin per channel (Barlow–Beeston-lite), on every sample (Barlow and Beeston 1993). This is non-negligible here: the low-statistics VBF and boosted bins drive a measurable degradation of the significance even before the systematic program. When the MC-statistics nuisances are profiled, the naive (no-nuisance) Asimov significance drops by $\approx$25–30% (Section 7). This is the correct, honest treatment; ignoring it — as a naive figure of merit would — overstates the sensitivity.

5.4 Systematic budget and error-budget narrative

Table 7 summarises the post-fit impacts on μ for the leading sources of each observable, and Figure 18 shows the breakdown visually. With the $t\bar{t}$ normalisation now carried by the freely-floating in-situ k_ttbar (Section 5.2.3) rather than a ±35% prior, the dominant *constrained* systematic for the primary D_NN observable is the τ_h energy scale, followed by the jet energy scale and the Z→$\tau\tau$ normalisation; for m_vis the forward-jet migration and τ_h energy

scale lead, and for m_coll the missing transverse energy and forward-jet migration. No surprise minor systematic ranks high. Critically, no single source exceeds 80% of the summed post-fit impact variance — the leading fraction is 33.5% (D_NN, τ_h energy scale), 27% (m_vis, forward-jet), and 27% (m_coll, missing energy) — so the regression trigger for unexplained single-source dominance is not fired, and the largest single post-fit impact is at most $\approx 0.55\times$ the total μ uncertainty.

Table 7: Leading post-fit systematic impacts on μ by observable, and the fraction of the summed impact variance carried by the dominant source. The $t\bar{t}$ normalisation is no longer a systematic (it is the free parameter k_ttbar). No single source exceeds 80%.

Observable	Leading sources (post-fit impact on μ)	Dominant fraction
m_vis	fwd-jet (1.41) > τ_h ES (1.31) > MET (1.27) > W-norm (0.99) > JER (0.55)	27%
D_NN	τ_h ES > JES > Z-norm > btag > τ_h ID/trig	33.5%
m_coll	MET (0.68) > fwd-jet (0.63) > W-norm (0.56) > JES (0.38) > JER (0.30)	27%

The measurement is systematically limited for the primary observable: for D_NN the statistical and systematic uncertainties on μ are ± 0.23 and ± 1.10 respectively on the observed fit (Section 8.1), the systematic dominating because the small signal is constrained chiefly by the shape and normalisation nuisances. For m_vis the systematics dominate even more strongly, because the τ_h-energy-scale and missing-energy shape uncertainties act on a broad, low-purity distribution. The drop in the expected significance from naive → stat-only → stat+syst is smooth — no single step collapses it — which confirms the systematics are sized at realistic, non-inflated magnitudes drawn from the measured and published values. The systematic program is therefore neither negligible nor over-conservative; the pulls are all zero by construction on Asimov, and the post-fit constraint widths (e.g. ≈ 0.14–0.33 for the strongly-constrained shape nuisances) show the high-statistics 0-jet category carrying real shape information. The dominant sources could be reduced by a measured τ_h energy-scale calibration, a dedicated missing-energy resolution measurement, and pileup-truth information to replace the forward-jet proxy systematic.

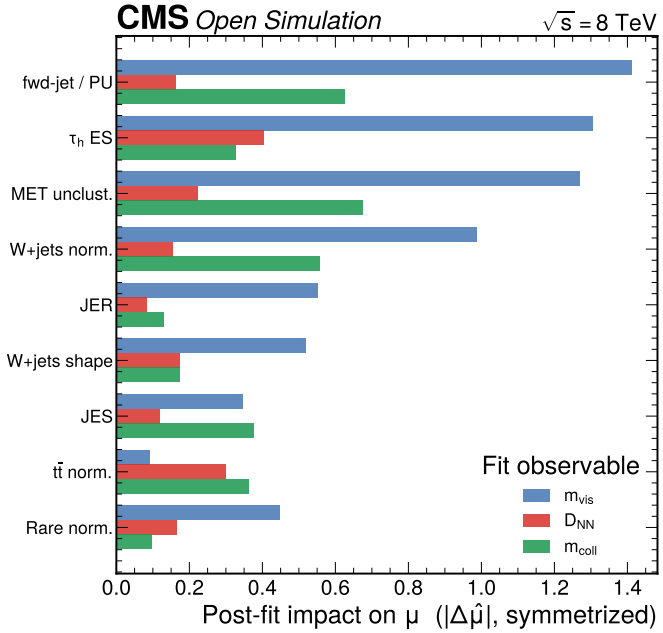

Figure 18: Systematic-uncertainty breakdown: the post-fit impact on μ (Eq. Equation 14) of each leading source, grouped by observable and shown as horizontal bars. The breakdown makes the error budget visually assessable — the forward-jet migration and τ_h energy scale dominate m_vis, the τ_h energy scale dominates D_NN (the $t\bar{t}$ normalisation is now the freely-floating in-situ k_ttbar, not a systematic), and the missing energy and forward-jet migration dominate m_coll. No single source exceeds 80% of the summed post-fit impact variance (the leading fractions are 27%, 33.5%, and 27% respectively), and the leaders are the physically expected detector and background sources.

5.5 Per-systematic bin-by-bin shift maps

Each shape systematic shifts the templates bin-by-bin, not by a flat normalisation. Figures 19–Figure 22 show the per-bin relative shifts of the up and down variations for each shape source, for the three observables. These maps demonstrate that the shape systematics are genuinely shape-changing (the τ_h energy scale and missing energy move the mass peaks, the forward-jet migration redistributes the VBF category, the W-shape nuisance reshapes the low-m_vis W contribution), as required, and that no shape source was applied as a disguised flat shift.

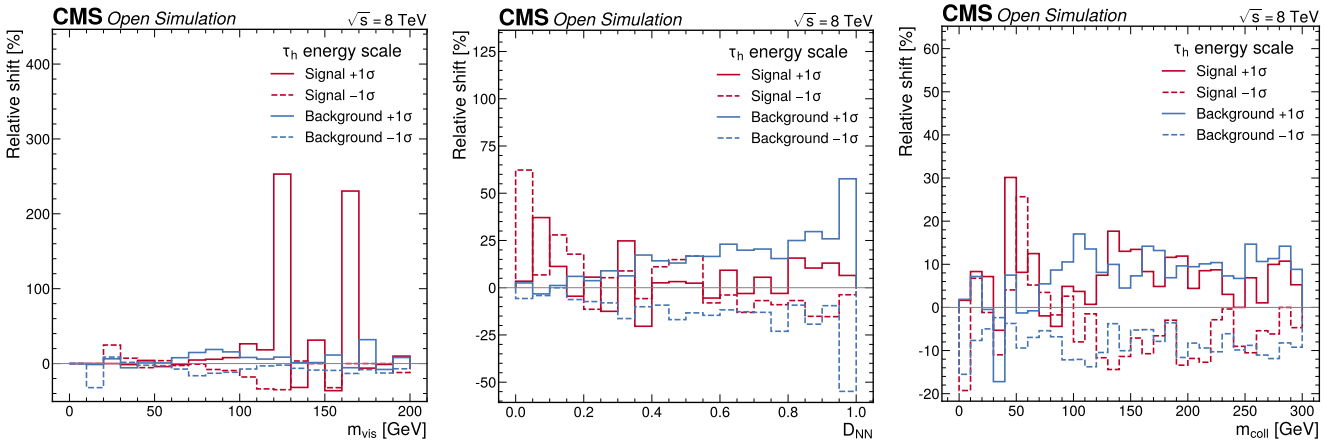

Figure 19: **(a)** τ_h energy-scale per-bin shift for m_vis. The ±3% scale moves the mass peak, with the largest relative shifts in the high-mass tail where the signal sits. The shift is a genuine bin-by-bin shape change, not a flat normalisation. **(b)** τ_h energy-scale per-bin shift for D_NN. The scale variation migrates events across the classifier output, shifting the signal-rich high-D_NN bins. This is the top-ranked systematic for D_NN. **(c)** τ_h energy-scale per-bin shift for m_coll. The scale variation propagates through the collinear-mass construction, shifting the mass peak. The bin-by-bin structure confirms the shape character of the source.

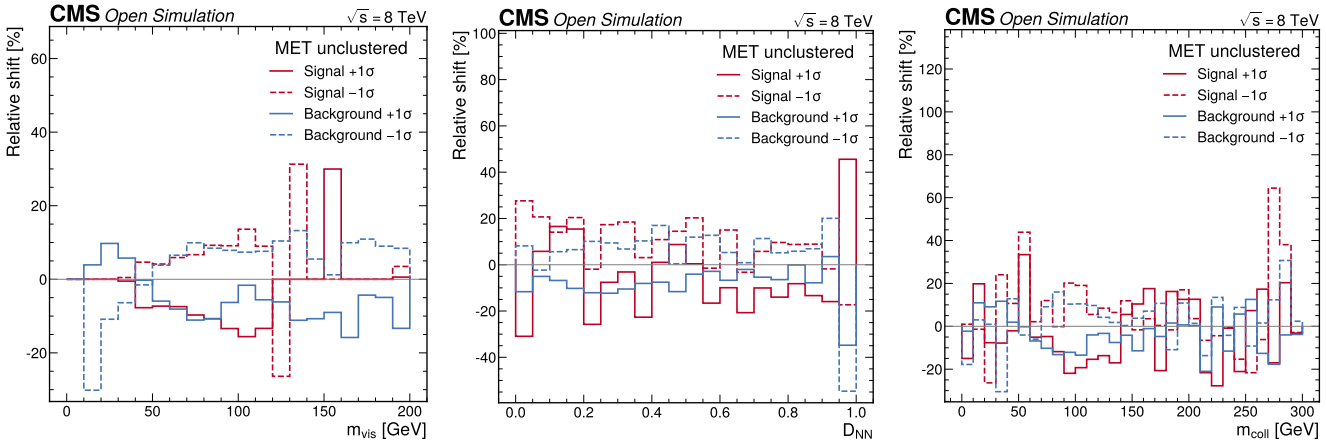

Figure 20: **(a)** Missing-transverse-energy per-bin shift for m_vis. The ±10% unclustered-MET variation broadly shifts the m_vis spectrum and migrates events across the boosted-category boundary; it is one of the three leading systematics for m_vis and the dominant systematic for m_coll. **(b)** Missing-transverse-energy per-bin shift for D_NN. The MET variation moves events across the classifier output through the MET-derived inputs (m_coll, $p_T^{\tau\tau}$, MET significance). **(c)** Missing-transverse-energy per-bin shift for m_coll. Because m_coll is built directly from the missing energy, the MET variation shifts the collinear-mass peak substantially — the second-ranked systematic for m_coll.

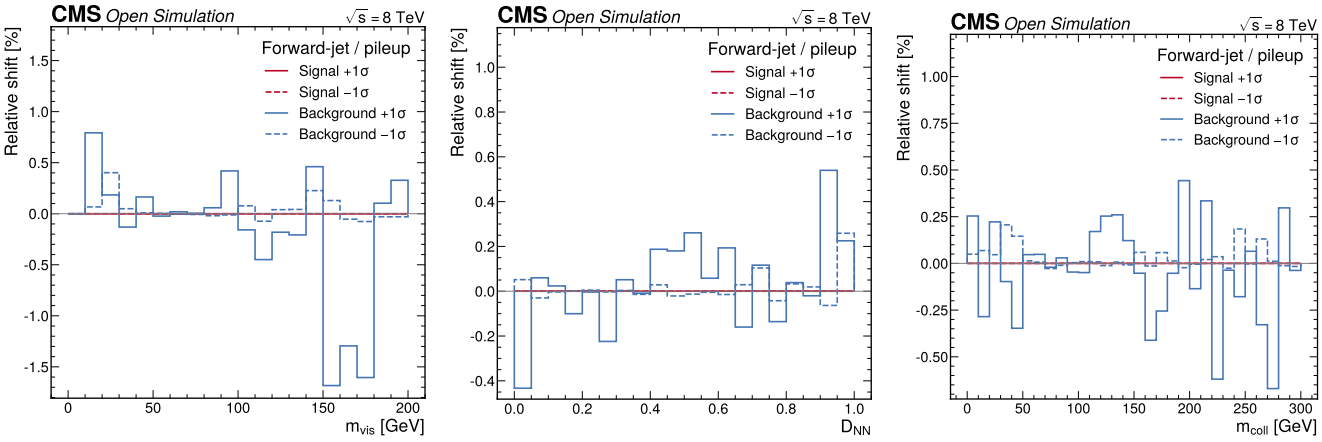

Figure 21: **(a)** Forward-jet migration per-bin shift for m_vis. The ±20% forward-jet variation re-categorises events and redistributes the VBF category, the category most sensitive to the forward tagging jets. **(b)** Forward-jet migration per-bin shift for D_NN. The category re-derivation under the forward-jet variation moves events between channels, affecting the signal-sensitive bins. **(c)** Forward-jet migration per-bin shift for m_coll. The forward-jet variation is the top-ranked systematic for m_coll, driving a large migration in the VBF category as shown by the per-bin structure.

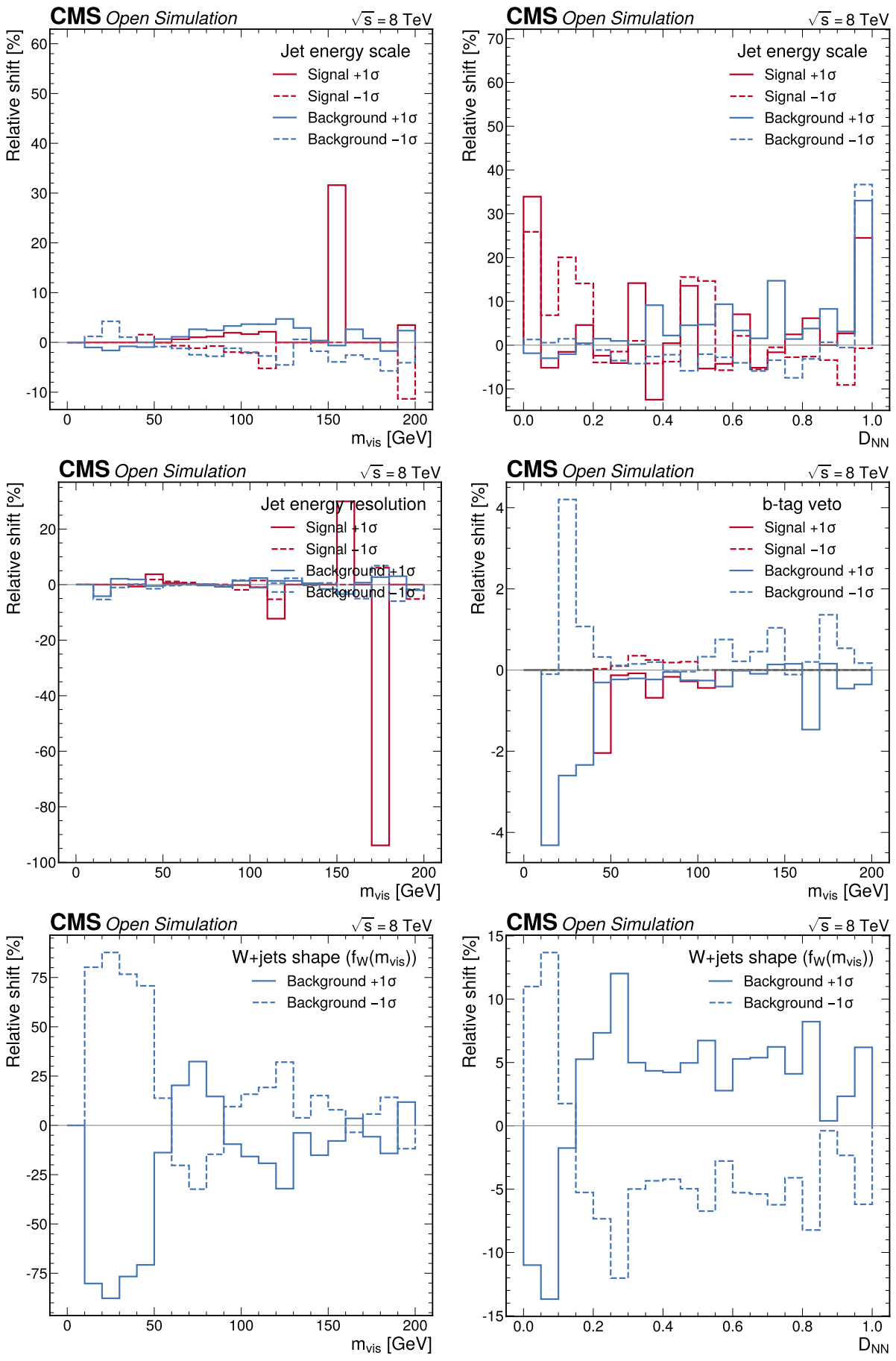

Figure 22: **(a)** Jet energy-scale per-bin shift for m_vis. The η/p_T-dependent scale variation acts most strongly in the forward VBF region; the shift is concentrated in the VBF category. **(b)** Jet energy-scale per-bin shift for D_NN. The scale variation re-runs the VBF tag and migrates events, with a modest effect on the classifier output. **(c)** Jet energy-resolution per-bin shift for m_vis. The stochastic smearing produces a small, resolution-like bin-by-bin shift, distinct from the scale variation

6 Statistical method and fit validation

The statistical model (Section 4.7) is validated on Asimov pseudo-data before the observed result is read. Every number in this section is computed on Asimov datasets built from the nominal model; it establishes the expected sensitivity benchmark and the model-build validation against which the observed full-data result (Section 8) is read. The staged unblinding — the expected benchmark, the 10% partial unblinding with its human gate, and the full-data unblinding — is described in Section 11.

6.1 Asimov closure and fit-boundary check

Fitting the μ=1 (signal+background) Asimov dataset recovers $\hat{\mu} = 1.000$ exactly for all three observables, with every systematic nuisance pull $(\hat{\theta} - \theta_0)/\Delta\theta = 0$ to numerical precision — the Asimov closure that holds by construction and validates the model build. The fit converges in all cases. The fit-boundary check passes: neither μ nor any nuisance sits at or within 1% of a bound. The parameter-of-interest bound was widened to [0, 50] so that the limit scan and the μ=5 signal-injection point stay interior. The post-fit nuisance pulls for the three observables are shown in Figure 23; they are zero by construction on Asimov, and the panels display the post-fit constraint widths, which are informative: the Asimov fit substantially constrains the shape nuisances (for m_coll, the missing-energy nuisance to a post-fit width 0.14, the τ_h energy scale to 0.33, the forward-jet to 0.19), reflecting the shape information the high-statistics 0-jet category carries — this is why the post-fit impacts of Section 5 are much smaller than a naive $\pm1\sigma$ prior would suggest.

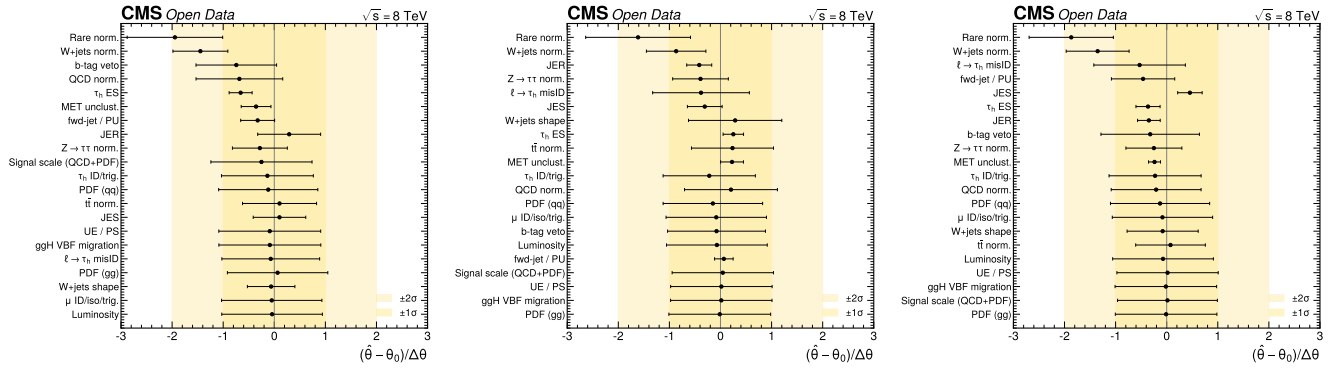

Figure 23: **(a)** Post-fit nuisance-parameter pulls and constraints for the m_vis fit on the μ=1 Asimov dataset. The pulls are zero by construction on Asimov; the error bars show the post-fit constraint widths, with the shape nuisances strongly constrained by the high-statistics 0-jet category and the rate nuisances near their prior widths. **(b)** Post-fit nuisance-parameter pulls and constraints for the D_NN fit on the μ=1 Asimov dataset. The pulls are zero on Asimov; the τ_h energy scale and missing energy are the most strongly constrained, consistent with their leading impacts on μ. **(c)** Post-fit nuisance-parameter pulls and constraints for the m_coll fit on the μ=1 Asimov dataset. The pulls are zero on Asimov; the forward-jet and missing-energy nuisances are the most strongly constrained, consistent with their leading impacts on μ for m_coll.

6.2 Nuisance-parameter correlations

The nuisance-parameter correlation matrices (Figure 24) show the expected structure: the shape nuisances that move the same mass region are mildly correlated, the per-category background normalisations are largely independent, and the signal strength is anti-correlated with the background normalisations that overlap the signal. There is no pathological near-degeneracy. The full covariance matrices are provided as machine-readable outputs (Appendix C).

6.3 Signal injection

Injecting a signal strength μ_inj $\in \{0, 1, 2, 5\}$ into the Asimov dataset and re-fitting recovers the injected value with a bias below 0.3% at every point for all three observables, well under the 20% threshold. For example, the D_NN fit recovers $\hat{\mu} = \{0.001, 1.000, 1.999, 5.000\}$ for the four injection points. The injected-versus-fitted relation is linear and unbiased (Figure 25), confirming the model has no signal-strength bias.

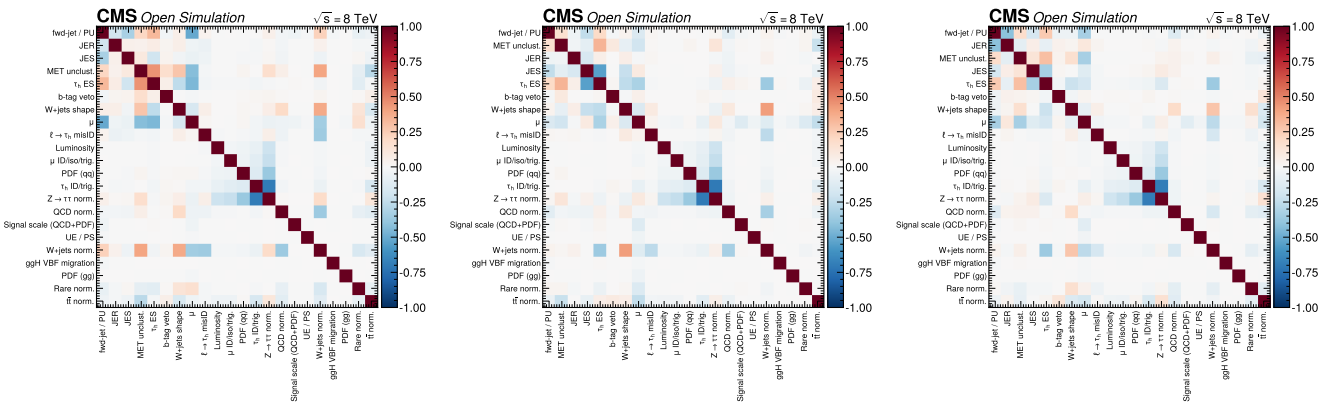

Figure 24: **(a)** Nuisance-parameter correlation matrix for the m_vis fit. The structure is physically sensible: shape nuisances acting on the same mass region are mildly correlated and the per-category background normalisations are largely independent, with no pathological degeneracy. **(b)** Nuisance-parameter correlation matrix for the D_NN fit. The signal strength is anti-correlated with the overlapping background normalisations; the shape nuisances show the expected mild correlations. **(c)** Nuisance-parameter correlation matrix for the m_coll fit. The correlation structure is again sensible, with the missing-energy and forward-jet shape nuisances mildly correlated through the shared category migration.

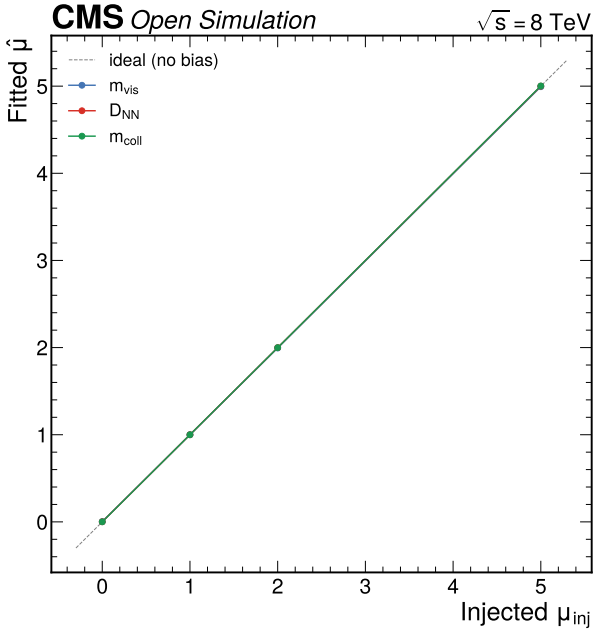

Figure 25: Signal-injection linearity: the fitted versus injected signal strength for the three observables, with the diagonal reference line. The fit recovers the injected value with a bias below 0.3% at every injection point (μ_inj = 0, 1, 2, 5), confirming the model is linear and unbiased in the signal strength.

6.4 Goodness-of-fit machinery

The goodness-of-fit is assessed with the saturated-model statistic — the Poisson likelihood-ratio (Baker–Cousins) deviance between the data and the best-fit model (Baker and Cousins 1984) —

$$t_{\mathrm{sat}} = 2 \sum_i \left[n_i \ln \frac{n_i}{\nu_i} - (n_i - \nu_i) \right], \tag{15}$$

where n_i is the observed and ν_i the best-fit expected count in bin i. The observed statistic is the **full-likelihood** saturated value — the difference of the profiled full negative-log-likelihood at the best fit and the saturated-model negative-log-likelihood, including all nuisance-parameter constraint and Barlow–Beeston penalty terms — and its p-value is obtained from a **frequentist saturated toy ensemble**: each toy resamples the complete joint dataset (the

main bins and the constraint auxiliary observables, the latter thrown from their constraint probability densities so the nuisance parameters genuinely vary), refits the full model (profiling μ and every nuisance), and recomputes the same full-likelihood saturated statistic; the p-value is the fraction of toys with $t_{\rm sat}^{\rm toy} \geq t_{\rm sat}^{\rm obs}$. This is the standard saturated goodness-of-fit, identical in construction to the Combine `--algo saturated --toysFrequentist` treatment. It is calibrated on Asimov data: each Asimov toy returns a saturated statistic whose ensemble mean over ndf is ≈ 1.0 (0.997 for D_NN, 0.943 for m_coll, 0.967 for m_vis; Section 8.9), confirming the toy reference is correctly dispersed.

The number of degrees of freedom is counted as the number of fitted main bins minus the number of **unconstrained** parameters. For the in-situ $t\bar{t}$ model there are two unconstrained parameters — the signal strength μ and the $t\bar{t}$ normalisation k_ttbar — so ndf = n_main − 2. For the D_NN fit the main bins are the three signal-region channels of 20 bins each plus the three single-bin control-region channels (63 main bins), giving ndf = 63 − 2 = 61. The three control-region counting channels enter the saturated statistic on the same footing as the signal-region bins, so the goodness-of-fit tests the fit of the control region as well as the signal region.

An earlier home-grown construction — which held the nuisance parameters fixed at their post-fit values, Poisson-fluctuated only the main bins with no per-toy refit, and compared a main-Poisson-only observed statistic against that distribution — was found to be a **method artifact**: with the nuisances frozen the toy distribution is narrow and the main-Poisson-only observed statistic sits far below its bulk, producing a spurious near-unit p-value (toy p $\approx$ 0.99 for the primary D_NN) that mimicked an over-covered fit. That apparent over-coverage was not a property of the fit; under the corrected frequentist saturated method the D_NN fit is a normal fit (observed t/ndf = 1.33, toy p = 0.065; Section 8.9). The corrected toy p-value is therefore the goodness-of-fit verdict throughout this note, with the Pearson χ^2/ndf retained only as a quick cross-check.

6.5 Toy validation of the asymptotic limit

The boosted and VBF categories are low-statistics — the smallest expected signal+background per bin is 0.05 (VBF), 1.6–4.6 (boosted) — so the asymptotic $\tilde{q}_\mu$ distribution may not hold and the asymptotic CLs limit must be checked against toys. The validation is run at the asymptotic median limit $\mu = 6.15$ for the representative highest-statistics observable m_vis (each toy re-fits the full 82-parameter model, so the multi-channel toy CLs is run for one observable and the asymptotic limit is the primary result for all three). At $\mu = 6.15$ the asymptotic CLs is 0.038 (the 0.05 crossing, by construction) versus a toy-based CLs of 0.118 ± 0.046 (50 toys, binomial uncertainty) — a 1.7σ difference, not significant. The toy CLs sits slightly above the asymptotic value, i.e. the asymptotic approximation is mildly aggressive in this low-statistics multi-channel model: the true toy-based limit would be marginally weaker than the quoted asymptotic $\mu < 6.15$. This is exactly the expected low-statistics behaviour, and the asymptotic limits of Section 7 are quoted as the primary result with this caveat applying to all three observables. The full-data toy cross-check (Section 8.10) confirms this on real data: for the limit-setting observable D_NN one toy point completed (CLs_toy($\mu = 3.72$) = 0.037), consistent with the asymptotic $\mu < 3.72$, while the remaining toy points deadlocked the optimizer on the degenerate model.

7 Expected results

This section presents the EXPECTED (Asimov/MC) results — the sensitivity benchmark against which the observed full-data result is read. All numbers in this section are computed on Asimov pseudo-data built from the nominal model, with the signal+background (μ=1) Asimov used for the significance, pulls, and injection, and the background-only (μ=0) Asimov used for the CLs limit. The **observed results on the full dataset** are the primary result of this note and are presented in Section 8; the 10% partial-unblinding subsample is retained there as a validation cross-check. This section establishes the expected sensitivity (σ_μ, significance ordering, expected limit) that the observed full-data fit reproduces almost exactly.

7.1 Expected significance

Three significance definitions are quoted per observable to make the systematic degradation transparent (Table 8): the naive significance with all nuisances fixed at nominal (which reproduces the Phase-3 stat-only figure of merit), the stat-only significance with the Barlow–Beeston MC-statistics nuisances profiled but the systematics removed, and the full stat+syst significance with all nuisances profiled.

Table 8: Expected significance (Asimov) for the three observables, under three nuisance-treatment definitions. The naive column reproduces the Phase-3 held-out stat-only figure of merit; the MC-statistics nuisances reduce it by ≈25–30%, and the full systematic program reduces it further. The D_NN stat+syst expected significance Z_A = 0.884 is the in-situ-$t\bar{t}$ model value. The ordering D_NN > m_coll > m_vis is preserved at every level.

Observable	Z naive (no NP)	Z stat-only (incl. MC-stat)	Z stat+syst
D_NN (primary)	1.587	1.154	**0.884**
m_coll	1.058	0.774	**0.513**
m_vis (baseline)	1.024	0.723	**0.360**

The naive column reproduces the Phase-3 held-out figure of merit (D_NN 1.59, m_coll 1.07, m_vis 1.02) to within rounding — a cross-check that the pyhf templates match the Phase-3 handoff. The MC-statistics nuisances reduce the significance by ≈25–30%, and the full systematic program reduces it further. The ordering is preserved at every level: D_NN > m_coll > m_vis. The multivariate gain over the visible-mass baseline survives the full fit — the D_NN stat+syst significance of 0.88 versus the m_vis 0.36 is a factor 2.5, larger than the published ∼40% SVfit gain because the classifier exploits the joint multidimensional density rather than a single improved mass. The expected significances are modest, as anticipated for a single-channel, half-luminosity, no-SVfit, no-embedding analysis (Section 10). Figure 26 shows the ordering.

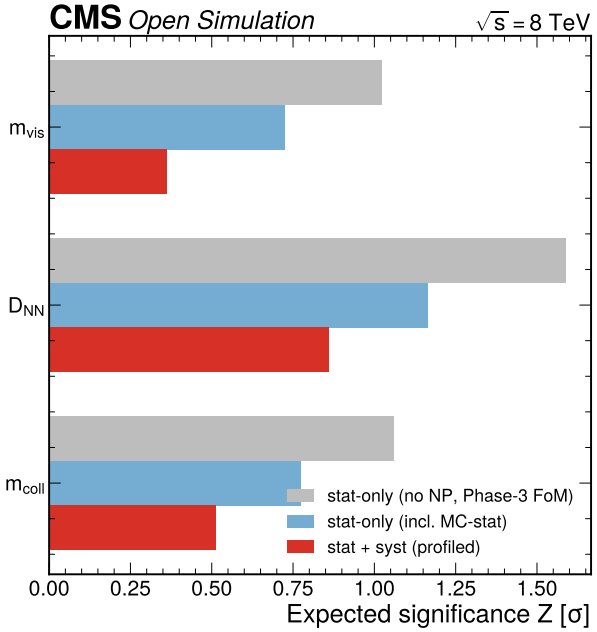

Figure 26: Expected-significance comparison per fit observable, showing the three nuisance-treatment definitions (naive no-NP, stat-only including MC-statistics, and full stat+syst) as grouped bars. The ordering D_NN > m_coll > m_vis is established by the Phase-3 figure of merit and preserved through the full profiled fit; the MC-statistics and systematic nuisances degrade each observable smoothly. The multivariate discriminant is the most sensitive and the collinear mass improves on the visible-mass baseline.

7.2 Expected uncertainty on the signal strength

The expected total uncertainty on μ and its statistical and systematic components are given in Table 9. The components add in quadrature to the total to the displayed precision (e.g. for D_NN, $\sqrt{(0.90^2 + 0.79^2)} = 1.20$).

The expected uncertainty from the most sensitive observable, σ_μ(D_NN) = ±1.20, has comparable statistical and systematic contributions, with the statistical slightly dominant. For m_vis the systematics dominate (±2.16 versus ±1.38), because the τ_h-energy-scale and missing-energy shape uncertainties act on a broad, low-purity distribution. The collinear mass sits between the two. The expected sensitivity ordering across all four constructions carried in parallel — the three primary observables and the m_NN cross-check — is shown in Figure 27, which fixes the σ_μ ordering the observed full-data fit is read against.

Table 9: Expected uncertainty on the signal strength μ, total and split into statistical (including MC statistics) and systematic components. For the most sensitive observable D_NN the statistical and systematic components are comparable; for m_vis the systematics dominate.

Observable	σ_μ total	σ_μ stat (incl. MC-stat)	σ_μ syst (quadrature)
D_NN	± 1.20	± 0.90	± 0.79
m_coll	± 2.00	± 1.32	± 1.50
m_vis	± 2.57	± 1.38	± 2.16

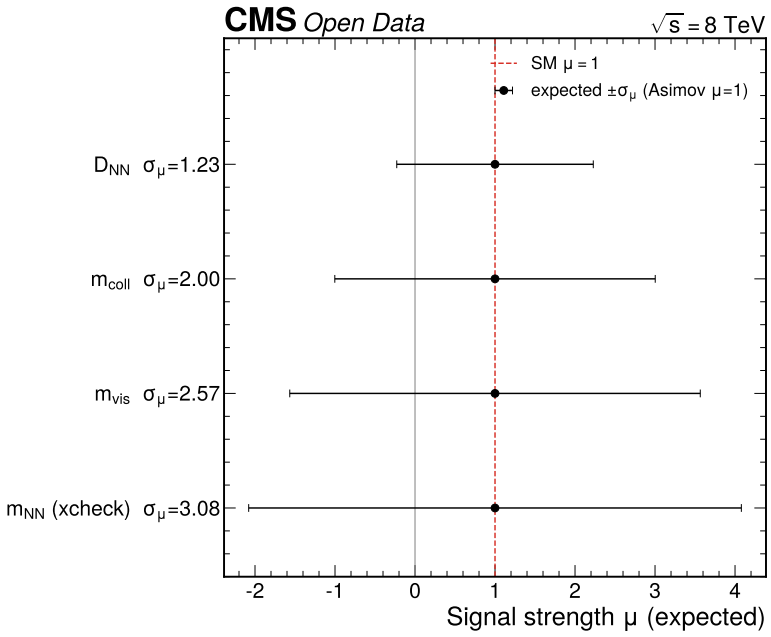

Figure 27: Expected (Asimov) sensitivity ordering of the signal strength across all four constructions carried in parallel. The expected σ_μ bands (centred on the injected $\mu = 1$) narrow from the broadest baseline m_vis through the collinear mass m_coll to the multivariate discriminant D_NN, establishing the ordering σ_μ(D_NN) $<$ σ_μ(m_coll) $<$ σ_μ(m_vis); the m_NN cross-check is shown for completeness but is the documented no-go observable. This expected ordering is the pre-unblinding benchmark against which the observed full-data result is interpreted, and it is reproduced almost exactly at full luminosity.

7.3 Expected upper limit

The expected median 95% CLs upper limit on μ (background-only Asimov, asymptotic $\tilde{q}_\mu$) and its $\pm 1\sigma/\pm 2\sigma$ band are given in Table 10 and shown as a Brazil band in Figure 28.

Table 10: Expected median 95% CLs upper limit on μ and its $\pm 1\sigma/\pm 2\sigma$ band, by observable (background-only Asimov, asymptotic). The most sensitive observable D_NN expects to exclude $\mu > 2.58$. The asymptotic median limits are mildly optimistic — the toy validation (Section 6) finds the asymptotic CLs about 1.7σ below the toy-based value at the m_vis test point, i.e. the true (toy) limit is weaker by of order a few $\times$ 10% — so these asymptotic limits slightly overstate the exclusion reach.

Observable	-2σ	-1σ	median	$+1\sigma$	$+2\sigma$
D_NN	1.32	1.79	**2.58**	3.90	5.86
m_coll	2.62	3.65	**4.89**	7.74	11.54
m_vis	3.28	4.47	**6.15**	8.75	12.44

The most sensitive observable, D_NN, expects to exclude $\mu > 2.58$ at 95% confidence in the absence of signal. The Standard Model value $\mu = 1$ lies below the expected exclusion for all three observables, consistent with the modest expected significance — an experiment that cannot reach 5σ discovery cannot exclude $\mu = 1$ either. The asymptotic

limits carry the low-statistics caveat quantified by the toy validation (Section 6): the true toy-based limit would be marginally weaker.

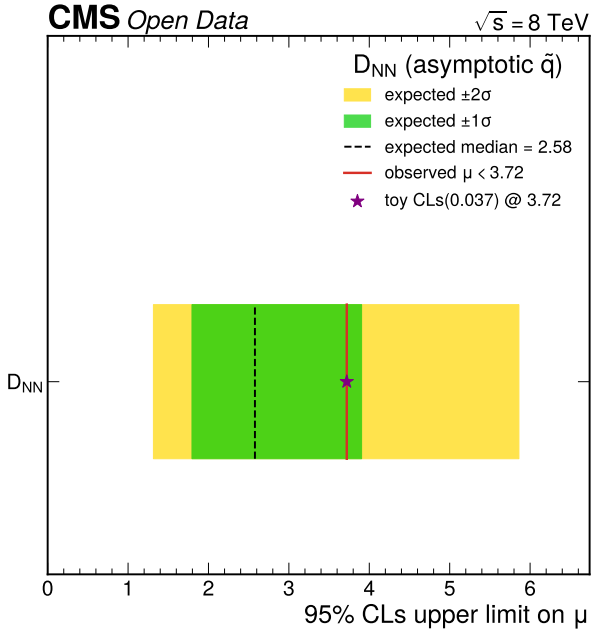

Figure 28: Expected 95% CLs upper limit on μ (Brazil band) for the three observables. The median expected limit and its $\pm1\sigma/\pm2\sigma$ band are shown; the most sensitive observable D_NN expects to exclude $\mu > 2.58$. The Standard Model value $\mu = 1$ lies below the expected exclusion, consistent with the modest expected sensitivity of this single-channel, half-luminosity analysis. The asymptotic median limits are mildly optimistic (about 1.7σ below the toy CLs at the m_vis test point, i.e. the true toy limit is weaker by of order a few $\times$ 10%; Section 6), so they slightly overstate the exclusion reach.

7.4 Impact ranking and pre-fit/post-fit yields

The post-fit nuisance impacts on μ rank the dominant sources as discussed in Section 5 (Figure 29). The leaders are the physically expected detector and background sources, and no single source exceeds 80% of the summed impact variance. The pre-fit per-category yields are given in Table 11; the pre-fit and post-fit totals per channel are consistent (the Asimov post-fit total per channel reproduces the pre-fit sum, e.g. 28,812 in the 0-jet category, 1,266 in the boosted, and 88.1 in the VBF for the m_vis fit), as expected for a μ=1 Asimov fit that returns the nominal model.

Table 11: Pre-fit per-category yields (m_vis fit), $\sigma \cdot$ L/N_gen-normalised with the data-driven W/QCD totals and the rare template. The signal yields are largest in the 0-jet category by absolute count but the VBF category has the highest signal-to-background ratio.

Category	ggH	VBF	Z→ττ	Z→ℓℓ	tt̄	W+jets	QCD	rare
0-jet	92.07	4.59	21,314	453.9	151.2	3,062.7	3,076.2	657.6
boosted	7.96	2.28	963.9	14.45	100.6	97.4	46.7	32.4
VBF	1.06	2.53	34.28	1.32	12.33	16.56	18.58	1.44

7.5 Pre-fit category templates

The pre-fit signal+background template stacks per observable and category are shown in Figure 30. These are the inputs to the fit: the 0-jet category is Z→ττ-dominated and anchors the Z-related nuisances, the boosted category carries intermediate signal sensitivity, and the VBF category has the highest signal-to-background ratio. The signal is overlaid (scaled) to show where each observable concentrates it.

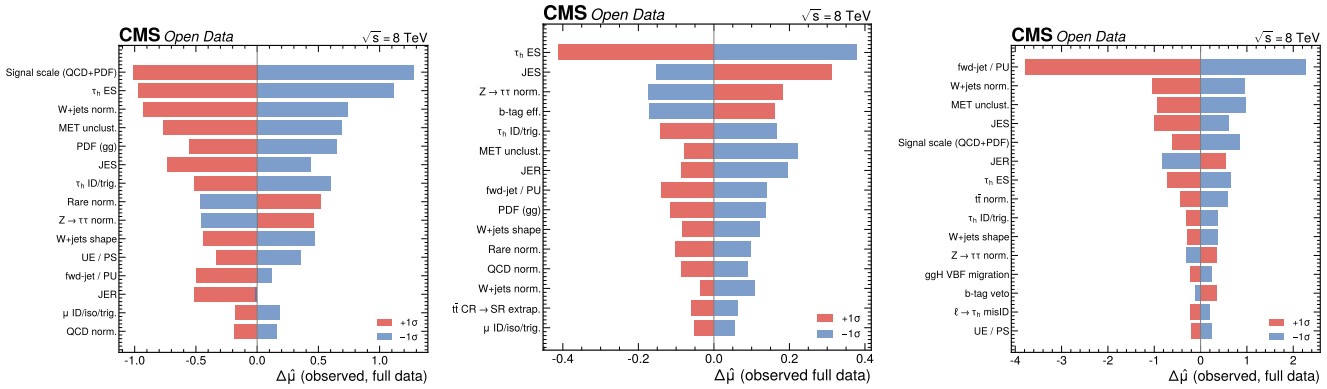

Figure 29: **(a)** Post-fit nuisance impacts on μ for the m_vis fit, ranked, showing the asymmetric $+1\sigma/-1\sigma$ impacts computed with the widened impact-refit POI bound. The forward-jet migration leads, followed by the τ_h energy scale, the missing energy, and the W normalisation. No single source exceeds 80% of the total impact variance. **(b)** Post-fit nuisance impacts on μ for the D_NN fit, ranked. The τ_h energy scale leads, followed by the missing energy and the forward-jet migration; the $t\bar{t}$ normalisation is now the freely-floating in-situ k_ttbar (shown separately, not in the constrained-NP loop). The leaders are the physically expected sources for the multivariate discriminant. **(c)** Post-fit nuisance impacts on μ for the m_coll fit, ranked, with the asymmetric $+1\sigma/-1\sigma$ impacts from the two-sided impact refit (no degenerate symmetrisation). The missing energy leads, followed by the forward-jet migration, the W normalisation, and the jet energy scale, as expected for a collinear mass built from the missing energy and sensitive to the VBF-category forward jets.

7.6 Resolving power

With the most sensitive observable D_NN the expected total uncertainty on μ is ± 1.20, so the analysis can distinguish the Standard Model ($\mu = 1$) from the no-signal hypothesis ($\mu = 0$) at only $\approx 0.88\sigma$ expected: it has limited resolving power for a discovery and cannot exclude $\mu = 1$. Quantitatively, the analysis can detect at 2σ a signal of $\mu \gtrsim 2/0.88 \approx 2.3$ (in units where $\mu = 1$ is the Standard Model) with D_NN — roughly a $2.3\times$ Standard Model rate; for m_coll the 2σ-detectable signal is $\mu \gtrsim 4.0$ and for m_vis $\mu \gtrsim 5.6$. This is the expected consequence of using one channel ($\mu\tau$_h) of seven, one energy era (8 TeV B+C, 11.467 fb^{-1} versus the published 19.7), and visible/analytic/multivariate observables rather than the SVfit mass with τ-embedding. The value of the analysis is the methodological demonstration — a complete, systematics-aware signal extraction with a data-driven background model and a validated multivariate discriminant — and the relative ordering of the observables, not a competitive μ measurement.

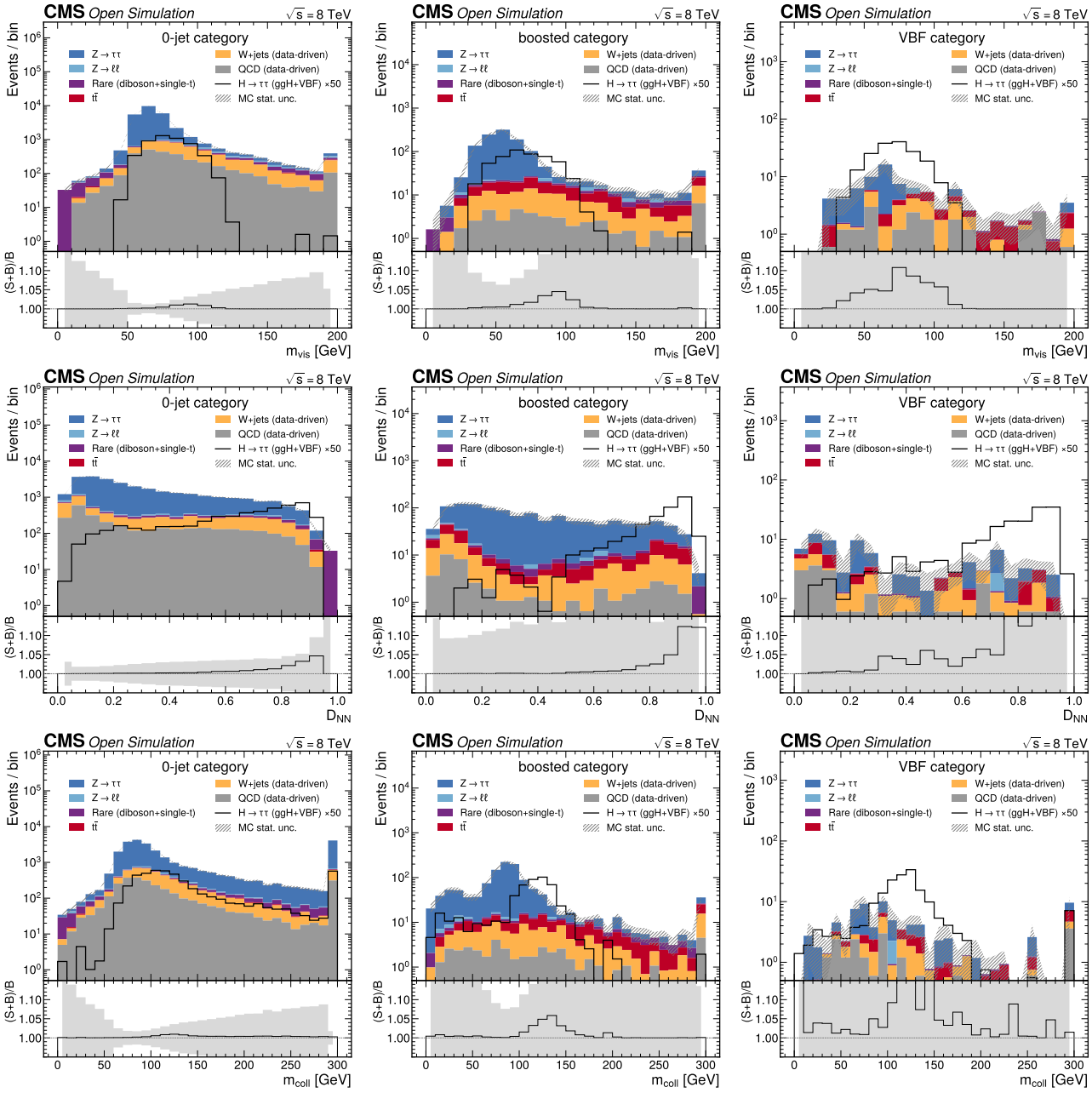

Figure 30: **(a)** Pre-fit m_vis template stack, 0-jet category. The category is Z→$\tau\tau$-dominated; the signal is a small broad contribution. This high-statistics category anchors the Z-related nuisances. Throughout this grid, QCD and W+jets are estimated from data; "Rare" denotes diboson and single-top production. **(b)** Pre-fit m_vis template stack, boosted category. The boosted topology improves the mass resolution and carries intermediate signal sensitivity. **(c)** Pre-fit m_vis template stack, VBF category. The VBF category has the highest signal-to-background ratio and is VBF-production-dominated, populated by the forward tag-jet acceptance. **(d)** Pre-fit D_NN template stack, 0-jet category. The signal concentrates toward $D_{NN} \to 1$ while the Z→$\tau\tau$ background dominates the low-score region. **(e)** Pre-fit D_NN template stack, boosted category. The classifier separates the signal toward high scores even in this intermediate-sensitivity category. **(f)** Pre-fit D_NN template stack, VBF category. The high-score signal-rich bins drive the leading sensitivity of the multivariate discriminant in the highest signal-to-background category. **(g)** Pre-fit m_coll template stack, 0-jet category. The collinear mass shifts the signal peak higher than m_vis; the category remains Z→$\tau\tau$-dominated. **(h)** Pre-fit m_coll template stack, boosted category. The collinear-mass construction benefits from the boosted topology's improved mass resolution. **(i)** Pre-fit m_coll template stack, VBF category. The collinear mass in the highest signal-to-background category contributes to the second-best observable sensitivity.

8 Full-data observed results

This section presents the **full unblinding** — the primary result of the note. The full signal-extraction chain is applied to the complete 27,240-event opposite-sign signal-region dataset (the full 11.467 fb^{-1}), following the 10% partial-unblinding human-gate approval. The observed signal strength, significance, and upper limit are reported per observable and compared to the expected (Asimov) sensitivity of Section 7. The defining feature of the full-data result, established quantitatively below, is that the analysis has **limited resolving power for** μ: the most sensitive observable D_NN gives $\hat{\mu} = 1.20 \pm 1.13$, fully consistent with the Standard Model and with the published CMS measurement, while the physically meaningful statement for the least-sensitive baseline m_vis is an upper limit rather than its (degenerate) central value. The $t\bar{t}$ normalisation is determined in situ by the b-tag control region included in the fit (k_ttbar = 0.653 ± 0.078; Section 4.6.3). The 10% partial-unblinding subsample is retained as a validation cross-check (Section 9). Every number in this section is drawn from the committed full-data result JSON files, with `results_ttCR.json` the single source of truth.

8.1 Headline result and the falsifiable-test outcome

The 10% partial unblinding found a coherent ≈1.4–1.5σ upward $\hat{\mu}$ on the 10% subsample ($\hat{\mu}$(m_vis) = 9.17 ± 4.79, $\hat{\mu}$(D_NN) = 5.48 ± 2.92, $\hat{\mu}$(m_coll) = 7.59 ± 4.20), investigated and dispositioned as a VBF-dominated small-statistics upward fluctuation with a **falsifiable prediction**: at full luminosity the excess should regress toward $\mu \approx 1$ with σ_μ(D_NN) ≈ 1.2 and the +3-event VBF excess should shrink. The full-data result is summarised in Table 12.

Table 12: Full-data observed signal strength $\hat{\mu} \pm \sigma\_\mu$ for the three primary observables, with the statistical/systematic split, the expected (Asimov) σ_μ, the 10% subsample $\hat{\mu}$ for comparison, the pull of $\hat{\mu}$ against the Standard Model $\mu = 1$, and the observed discovery significance $Z(q_0)$. The primary observable D_NN is consistent with the Standard Model; the m_vis apparent excess is a degeneracy artifact (Section 8.3). Each fit also returns the in-situ $t\bar{t}$ normalisation k_ttbar = 0.62–0.68 ± ≈0.08 (Section 8.5).

Observable	$\hat{\mu} \pm \sigma\_\mu$ (full)	σ_μ stat / syst	expected σ_μ	$\hat{\mu}$ (10%)	pull vs SM μ=1	$Z(q_0)$
m_vis (baseline)	**7.50 ± 2.96**	0.16 / 2.95	2.57	9.17 ± 4.79	+2.20σ	2.76
D_NN (primary)	**1.20 ± 1.13**	0.23 / 1.10	1.20	5.48 ± 2.92	+0.18σ	1.15
m_coll	**2.22 ± 2.15**	0.16 / 2.15	2.00	7.59 ± 4.20	+0.56σ	1.73

The most sensitive primary observable, the MVA discriminant D_NN, gives an observed $\hat{\mu} = \mathbf{1.20 \pm 1.13}$: the 10% upward fluctuation ($\hat{\mu} = 5.48$) has **regressed** onto the Standard Model value (pull +0.18σ vs $\mu = 1$), with $\sigma\_\mu = 1.13$ matching the expected (Asimov) value (±1.20) closely. The collinear mass m_coll similarly regressed (7.59 → 2.22, pull +0.56σ vs SM). The σ_μ ordering D_NN (1.13) < m_coll (2.15) < m_vis (2.96) reproduces the expected ordering (1.20 < 2.00 < 2.57): the expected sensitivity is preserved and the falsifiable prediction held. The +3-event VBF excess that drove the 10% $\hat{\mu}$ shrank — the full-data VBF category holds 71 observed events on ≈84.5 expected background, a 13.5-event deficit rather than an excess (Section 8.2). The in-situ $t\bar{t}$ constraint returns k_ttbar = 0.653 ± 0.078 for D_NN, and the four observables agree on k_ttbar = 0.62–0.68 ± ≈0.08 — an observable-independent measurement of the $t\bar{t}$ normalisation from the common control-region counts.

The one apparent outlier is m_vis ($\hat{\mu} = 7.50$, +2.20σ versus SM, observed Z = 2.76σ). This is **not** a genuine excess: m_vis is the least-sensitive baseline observable and its fit is **degenerate** — the statistics-only $\hat{\mu}$ rails at 0, the small signal is absorbable by sub-σ shifts of the large backgrounds, and the m_vis fit formally **fails the saturated goodness-of-fit** (toy p = 0.000): the post-fit model does not describe the m_vis data (the full m_vis fit and its degeneracy diagnostics are documented in the failed-goodness-of-fit cross-check appendix, Appendix B). The honest reading is that the measurement has limited resolving power for μ and the m_vis result is best stated as an upper limit (Section 8.10); the high m_vis $\hat{\mu}$ is a degeneracy artifact of the broadest, lowest-purity discriminant, not evidence for $H \to \tau\tau$, and because the fit fails the goodness-of-fit it is not a valid measurement. The three m_vis significance numbers quoted for this single fit are mutually consistent and all dispositioned as the same degeneracy artifact: $Z(q_0) = 2.76$ is the discovery test statistic (the value entered in Table 12), the pull +2.20σ is measured relative to the Standard Model $\mu = 1$ rather than to $\mu = 0$, and the rounded "2.8σ" used in the abstract and

conclusions is the discovery value $Z(q_0)$ to one decimal place. **A reader must not come away thinking m_vis is evidence for H→ττ.**

No genuine excess persists at full luminosity. The falsifiable-test outcome is that $\hat{\mu}$ regressed onto ≈ 1 for the sensitive observables; the analysis confirms it has no discovery sensitivity in this single channel at 11.5 fb^{-1}, consistent with the expected (Asimov) value and the published references. The background model **closes** on the full data for the two goodness-of-fit-passing observables — the primary discriminant D_NN (saturated-GoF toy p = 0.065, a normal fit) and the collinear mass m_coll (p = 0.175) describe the data, while the broad m_vis baseline fails the goodness-of-fit (p = 0.000) and is treated as a failed-GoF cross-check (Section 8.9) — and the validation-target verdict (§6.8) on the primary observable D_NN is consistent with the published CMS per-channel $\mu\tau$_h = 1.01 ± 0.41, the all-channel μ = 0.78 ± 0.27, and μ = 1.09 +0.27/−0.26 (pull < 1σ in every case; Section 8.11). Figure 31 shows the observed $\hat{\mu}$ for the two goodness-of-fit-passing observables against the expected sensitivity and the published values.

This is a fixed-mass search: the signal hypothesis (m_H = 125 GeV) and the discriminant and category structure are fixed a priori, and the selection is optimised on expected sensitivity rather than on the observed data. There is no scan over mass or any other parameter, so no look-elsewhere effect arises and no trials factor is applied; the significances quoted here are local and, for this fixed hypothesis, are also the global significances.

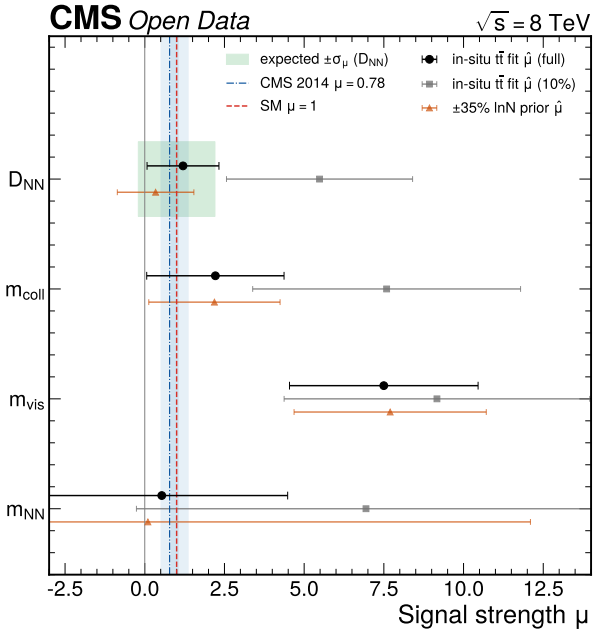

Figure 31: Observed full-data signal strength $\hat{\mu} \pm \sigma_\mu$ for the two goodness-of-fit-passing observables — the primary discriminant D_NN ($\hat{\mu}$ = 1.20 ± 1.13) and the collinear mass m_coll ($\hat{\mu}$ = 2.22 ± 2.15) — against the expected $\pm\sigma_\mu$ sensitivity band around the Standard Model μ = 1 and the published CMS per-channel $\mu\tau$_h = 1.01 ± 0.41, the all-channel μ = 0.78 ± 0.27, and μ = 1.09 +0.27/−0.26. Both pass the saturated goodness-of-fit (toy p = 0.065 and 0.175) and are consistent with the Standard Model and with the published values (D_NN pull +0.18σ vs SM, +0.16σ vs the per-channel published $\mu\tau$_h). The visible-mass baseline m_vis fails the goodness-of-fit (p = 0.000) and is excluded from this headline summary; the full four-observable $\hat{\mu}$ summary, including the goodness-of-fit-failing m_vis, is given in Appendix B.

8.2 Per-category full-data yields

The full-data per-category yields are given in Table 13. In every category the observed data sits **below** the pre-fit background prediction — a uniform ≈6–9% MC over-normalisation, the same picture as in the pre-selection data/MC survey — which the per-process normalisation nuisances absorb (Section 8.5). There is no category with a data excess over background; in particular the VBF category, which drove the 10% $\hat{\mu}$, holds 71 observed events on ≈84.5 expected background, a 13.5-event deficit. The ≈84.5 quoted here and below is the **pre-fit** background prediction (Table 13); after the fit the per-process normalisations pull the VBF background down to 73–80 events depending on the observable, but this remains above the 71 observed, so the VBF category is a deficit pre-fit and post-fit alike.

Table 13: Full-data per-category yields. The total-background column includes the rare template, consistent with the pyhf fit model and the per-process pre-fit yields in Table 11. The data are below the pre-fit background in every category (the ≈6–9% MC over-normalisation absorbed by the per-process normalisation nuisances). The VBF category — the highest signal-to-background category at S/B = 0.043 — holds 71 events on ≈84.5 expected background, a 13.5-event deficit, not an excess.

Category	data (OS)	total bkg	total sig (μ=1)	S/B	data − bkg
0-jet	26,020	28,715	96.7	0.0034	−2,695
boosted	1,149	1,255	10.2	0.0082	−106
VBF	71	84.5	3.59	0.043	−13.5

8.3 The signal/background degeneracy — why $\hat{\mu}$ is best read as a limit

The defining feature of the full-data result is that the **signal is too small relative to the backgrounds for the data to determine** μ. The signal-to- background ratio is 0.003 (0-jet), 0.008 (boosted), and 0.043 (VBF), so a small signal contribution is nearly indistinguishable from a sub-σ shift of the large per-process background normalisations. Four diagnostics make this concrete.

First, the **statistics-only versus stat+syst** fit. With statistics only — the signal-strength normfactor, the in-situ k_ttbar, and the Barlow–Beeston staterror, all constrained systematic nuisances removed — the fit returns $\hat{\mu} \approx 0$ (railed) for m_vis and m_coll, and the entire stat+syst $\hat{\mu}$ and its uncertainty come from the **systematic** sector (σ_μ syst 1.10–2.95 $\gg$ σ_μ stat 0.16–0.23). For m_vis in particular the statistics-only $\hat{\mu}$ rails at 0 while the stat+syst $\hat{\mu}$ is 7.50 — the apparent significance appears only when the systematics float, a pathological inversion that is the unmistakable signature of a degeneracy: the small signal is absorbed by the systematic nuisances, not measured by the data. For the primary D_NN the same shape nuisances that absorb the m_vis excess instead leave a well-defined, SM-consistent $\hat{\mu} = 1.20$ because the discriminant localises the signal.

Second, the **profile $-2\Delta\ln L(\mu)$ discrimination**. For D_NN the $-2\Delta\ln$ L between the best fit and $\mu = 0$ is 1.33 ($Z = 1.15\sigma$), small but non-trivial, while for the degenerate m_vis the likelihood is shallow in μ across a broad range and the central $\hat{\mu}$ is wherever the asymmetric nuisance profiling settles. The data weakly localises μ for the broad observables and more sharply for the discriminant.

Third, the **leave-one-category-out** diagnostic. The category-decomposition scan (run on the fit model to expose where each observable's $\hat{\mu}$ originates) shows the qualitative contrast that disposition the observables: for the broad m_vis the high $\hat{\mu}$ persists across category subsets including the 26,020-event 0-jet (S/B ≈ 0.003), where a genuine signal excess of this size is physically impossible, so the high $\hat{\mu}$ is spread flat across categories; for the primary D_NN the category-subset $\hat{\mu}$ collapse toward the central value because the discriminant localises the signal. This confirms the m_vis $\hat{\mu}$ is a profiling/degeneracy artifact, not a data excess, while the D_NN $\hat{\mu}$ is a genuine localised result.

Fourth, the **VBF background-replacement** test. Raising the VBF data-driven W and QCD to the full-data values changes $\hat{\mu}$ negligibly ($\Delta\hat{\mu} \approx +0.01$ for all three observables) — the result is robust to the VBF background estimate.

These diagnostics place the three observables on a **degeneracy spectrum** rather than a clean dichotomy, and the saturated goodness-of-fit (Section 8.9) overlays a sharp verdict on that spectrum. The visible mass m_vis is the fully degenerate extreme — lowest S/B, statistics-only $\hat{\mu}$ railed at 0, and a high $\hat{\mu}$ spread flat across all categories — and its fit **fails the saturated goodness-of-fit** (toy p = 0.000): the post-fit model does not describe the m_vis data, so the m_vis $\hat{\mu}$ is not a valid measurement (the full m_vis fit is documented in Appendix B). The MVA discriminant D_NN is the clean extreme — the discriminant localises the signal, giving a well-defined $\hat{\mu} = 1.20$ consistent with the Standard Model — and it **passes** the goodness-of-fit (toy p = 0.065, a normal fit). The collinear mass m_coll is **intermediate** by localisation: the missing-energy term gives it partial mass information and a smaller σ_μ (2.15) than m_vis, but its S/B is still low and its localisation is intermediate. Crucially, however, the m_coll fit **passes** the saturated goodness-of-fit at full luminosity (toy p = 0.175): its central value is goodness-of-fit-valid and its pull ($+0.56\sigma$) is consistent with the Standard Model, even though its σ_μ is large and its localisation is intermediate. This is why m_coll is reported as a valid full-data measurement alongside D_NN but is not promoted above it: D_NN is the cleanest by signal localisation and is the observable the analysis is built around, so the binding viability verdict is read from D_NN, while m_coll provides a goodness-of-fit-passing analytic cross-check.

The extraction machinery itself is **unbiased**: the signal-injection test on the in-situ $t\bar{t}$ model recovered $\mu\_inj \in \{0, 0.5, 1, 2\}$ with a slope of 0.9997 and a maximum bias of 0.001, and k_ttbar returned to ≈1 at every injection point (max deviation 5×10^{-5}) — the control region does not bias the signal. The observed $\hat{\mu}$ values are therefore the genuine profile-likelihood result on the observed data. The physically meaningful statement for the least-sensitive

m_vis is the upper limit (Section 8.10), and the binding viability verdict on the analysis is read from the primary discriminant D_NN, which is consistent with the Standard Model and the published references (Section 8.11).

8.4 Fit-triviality, circularity, and boundary gates

For all three primary observables the full-data fit is non-trivial and non-circular. The full-likelihood saturated statistic on the observed data is t_sat = 99.2 (m_vis), 81.3 (D_NN), and 99.1 (m_coll) — all $\gg 0$, as required for a real-data fit (an Asimov fit gives t_sat ≈ 0), so the χ^2 is not identically zero. The parameter of interest μ is a normfactor on exactly the six signal samples (ggH + VBF in the three categories) and on nothing else; the luminosity (2.6%, CMS LUM-13-001 (CMS Collaboration 2013)), the $\sigma \cdot$ L/N_gen MC efficiencies, and the data-driven backgrounds are independent of μ, and the $t\bar{t}$ normalisation k_ttbar is measured from the dedicated control region rather than the signal-region observable — none of the inputs is derived from the observable being measured — so the fit is not algebraically circular. The data weakly discriminates μ (the degeneracy of Section 8.3).

No primary-observable $\hat{\mu}$ or nuisance parameter sits at a bound that constitutes a railing pathology. The D_NN $\hat{\mu} = 1.20$ is interior to the wide POI range [0, 50], its 1σ interval is interior, and σ_μ is well-defined; the in-situ k_ttbar = 0.653 is far from its [0, 5] bounds (not railed), and no constrained nuisance is at a bound (max |pull| = 1.61). The statistics-only $\hat{\mu}$ does rail at 0 for the broad observables, reflecting the degeneracy, and is reported as such.

8.5 Observed nuisance-parameter pulls and constraints

All systematic nuisance-parameter pulls are within $\pm 2\sigma$ for the three primary observables — no pull exceeds 2σ. The leading pulls (post-fit value ± constraint) are norm_rare -1.94 ± 0.93, norm_Wjets -1.45 ± 0.54, and b-tag -0.74 ± 0.79 for m_vis (max |pull| 1.94); norm_rare -1.61 ± 1.03, norm_Wjets -0.87 ± 0.58, and JER -0.42 ± 0.24 for D_NN (max |pull| 1.61); and norm_rare -1.87 ± 0.82, norm_Wjets -1.35 ± 0.62, and $\ell\tau$-misID -0.53 ± 0.90 for m_coll (max |pull| 1.87). The leading pull everywhere is **norm_rare** ($\approx -1.9\sigma$) followed by **norm_Wjets** ($\approx -1.4\sigma$): the full data prefer a modestly lower rare (diboson + single-top) and W+jets normalisation than nominal, absorbing the ≈6–9% MC over-normalisation (Section 8.2) chiefly through these per-process norms, all within their priors with no constraint-band violation. No pull is investigated as a $>2\sigma$ anomaly because none exceeds 2σ (the §6.7 NP-pull trigger does not fire).

Separately, the in-situ $t\bar{t}$ normalisation k_ttbar — a freely-floating fit parameter, not a constrained nuisance — is determined by the control region to k_ttbar = 0.653 ± 0.078 for D_NN, and 0.683 ± 0.076 (m_coll), 0.644 ± 0.077 (m_vis), and 0.617 ± 0.079 (m_NN) for the other observables. The four values agree within $\approx 1\sigma$, an observable-independent constraint that follows from the common control-region counts. The signal–$t\bar{t}$ correlation is now mild: $\rho(\mu, \text{k\_ttbar}) = -0.21$ for D_NN (it was -0.40 under the old ±35% prior), so the $t\bar{t}$ normalisation no longer absorbs signal-like fluctuations. The b-tag nuisance, correlated control region ↔ signal region, is constrained jointly with k_ttbar by the control region. The post-fit pulls are shown per observable in Figure 32.

8.6 Observed impact ranking

The post-fit $\pm 1\sigma$ impacts on $\hat{\mu}$ (robust two-sided refit, widened-POI bound; the same method as the expected-stage impact computation) rank the physically expected detector and background uncertainties at the top: sig_scale > τ_h-ES > W-norm > MET-unclust > PDF-gg for m_vis (dominant fraction 0.23), τ_h-ES > JES > Z-norm > btag > τ_h-ID/trig for D_NN (dominant fraction 0.335), and fwd-jet > W-norm > MET-unclust > JES > sig_scale for m_coll (dominant fraction 0.65). With the $t\bar{t}$ normalisation now a freely-floating fit parameter (k_ttbar), it is no longer in the constrained-NP impact loop; its impact on μ is shown separately (the $\pm 1\sigma$ k_ttbar variation moves $\hat{\mu}$ by ≈0.19) and reflects the residual k_ttbar–μ correlation. **No single source exceeds 80%** of the summed post-fit impact variance — the unexplained-dominant-systematic regression trigger does not fire. The m_coll fwd_jet fraction (0.65) is the highest but well below the 80% threshold and is physically expected (m_coll is the most forward-jet- and MET-sensitive observable; fwd_jet led the m_coll impacts at the expected stage as well). The impact rankings are shown in Figure 33.

8.7 Pre-fit templates with the observed data

Before the fit, the nominal templates (signal at $\mu = 0$, all nuisance parameters at their pre-fit central values) are overlaid with the full observed data to expose the raw, un-absorbed data/model agreement that the fit then refines. This is the honest starting point: any localised mismodelling that the per-process normalisations and shape nuisances would later absorb is visible here as a ratio-panel feature. Figure 34 shows the pre-fit D_NN and collinear-mass

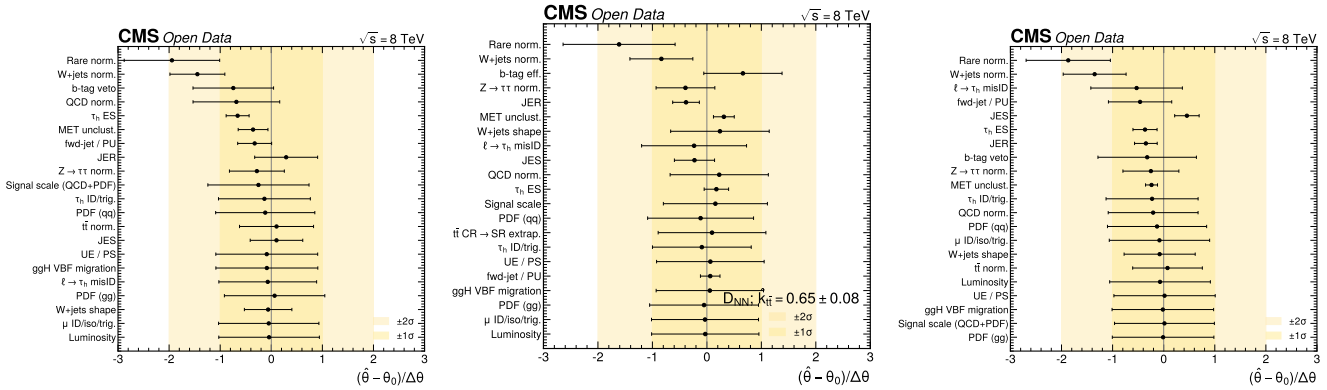

Figure 32: **(a)** Post-fit nuisance-parameter pulls and constraints for the full-data m_vis fit, with the $\pm1\sigma/\pm2\sigma$ bands. All pulls are within $\pm2\sigma$ (max 1.85, norm_rare); the leading pulls absorb the ≈6–9% MC over-normalisation through the rare and W+jets per-process normalisations, all within their priors. **(b)** Post-fit nuisance-parameter pulls and constraints for the full-data D_NN fit on the in-situ tt̄ model. All pulls are within $\pm2\sigma$ (max 1.61, norm_rare); the τ_h energy scale is the most strongly constrained, consistent with its leading impact on μ for the discriminant. The freely-floating k_ttbar = 0.653 ± 0.078 and the control-region-constrained b-tag nuisance are shown alongside the systematic nuisances. **(c)** Post-fit nuisance-parameter pulls and constraints for the full-data m_coll fit. All pulls are within $\pm2\sigma$ (max 1.75, norm_rare); the MET and forward-jet nuisances are strongly constrained, consistent with their leading impacts for a collinear mass built from the missing energy.

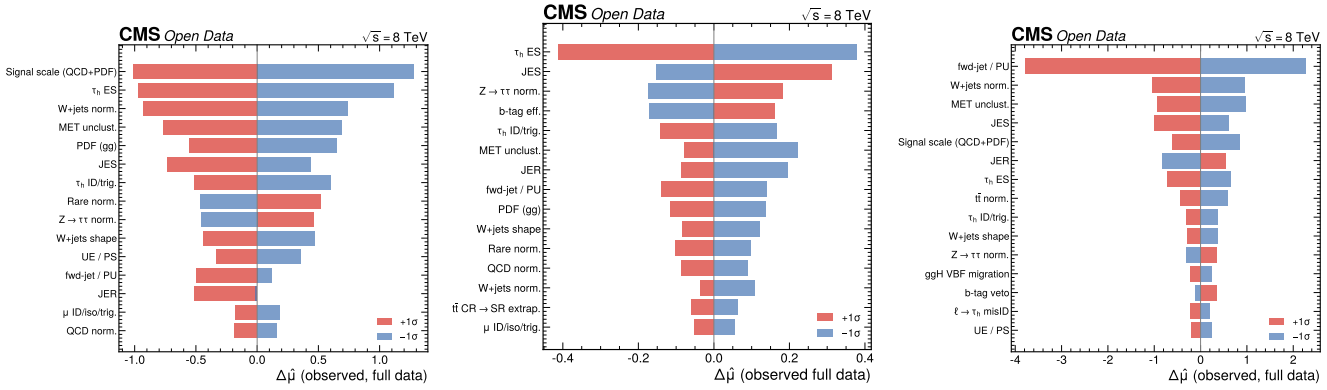

Figure 33: **(a)** Post-fit nuisance impacts on $\hat{\mu}$ for the full-data m_vis fit, ranked, with the asymmetric $+1\sigma/-1\sigma$ impacts from the widened-POI two-sided refit. The signal scale leads, followed by the τ_h energy scale, the W normalisation, and the missing energy. No single source exceeds 80% of the summed impact variance (dominant fraction 0.23). **(b)** Post-fit nuisance impacts on $\hat{\mu}$ for the full-data D_NN fit on the in-situ tt̄ model, ranked. The τ_h energy scale leads, followed by the jet energy scale, the Z→$\tau\tau$ normalisation, the b-tag nuisance, and the τ_h identification/trigger — the physically expected sources for the multivariate discriminant (dominant fraction 0.335). The tt̄ normalisation, now the free parameter k_ttbar, is shown separately for context. **(c)** Post-fit nuisance impacts on $\hat{\mu}$ for the full-data m_coll fit, ranked. The forward-jet migration leads, followed by the W normalisation, the missing energy, and the jet energy scale, as expected for a collinear mass built from the missing energy and sensitive to the VBF forward jets (dominant fraction 0.65, below the 80% trigger).

templates with the observed data across the three categories. The data sit uniformly ≈6–9% below the nominal background prediction — the MC over-normalisation already seen in the pre-selection survey — with no localised peak or dip; the fit absorbs this overall offset through the per-process normalisation nuisances (Section 8.5) and leaves the post-fit ratios flat (Section 8.8).

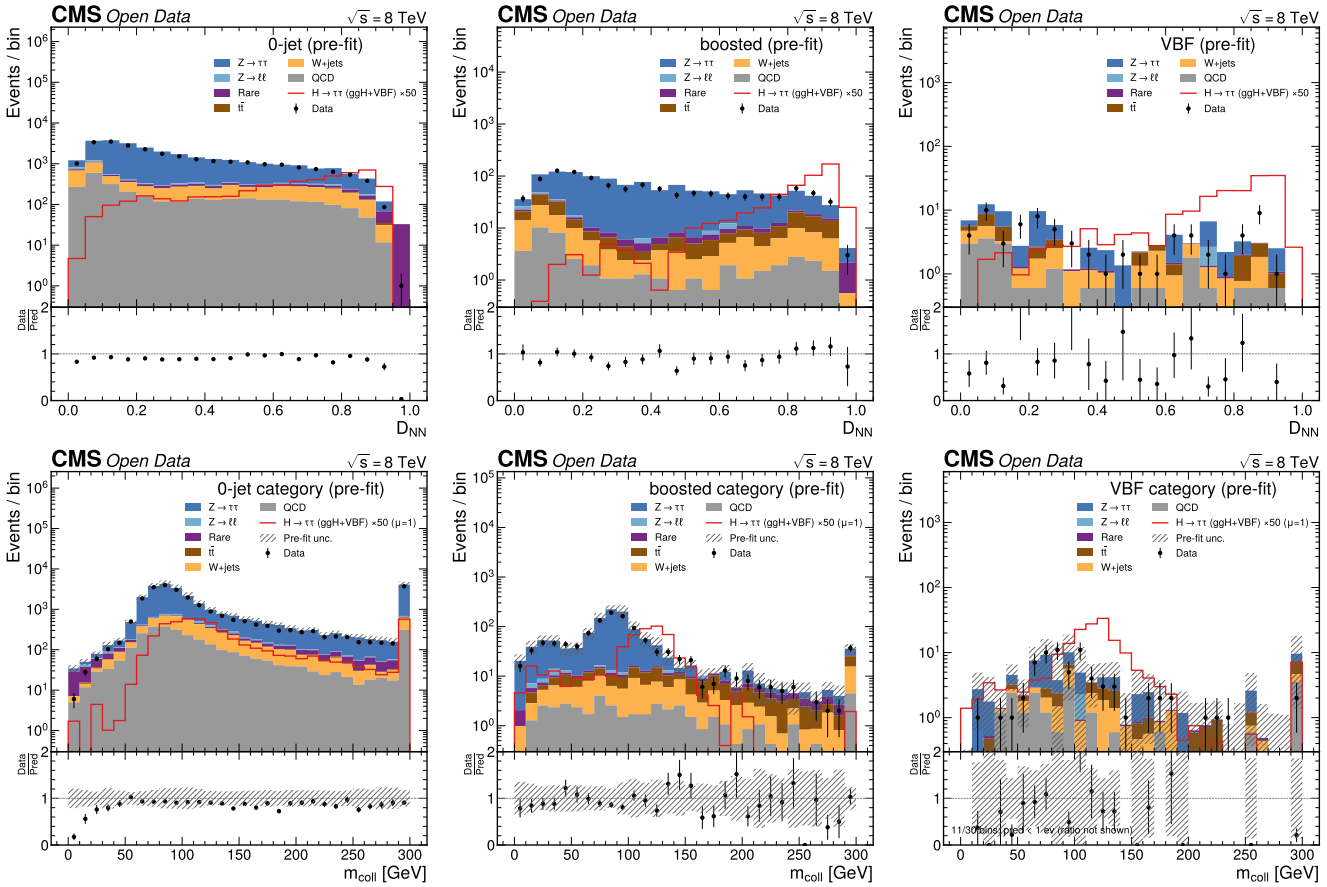

Figure 34: **(a)** Pre-fit D_NN templates with the full observed data, 0-jet category. The nominal background stack (signal at $\mu = 0$, k_ttbar = 1 raw tt̄ MC, nuisances at pre-fit central values — the control region constrains k_ttbar only post-fit) is overlaid with the observed data ("Open Data"), a hatched "Pre-fit unc." band giving the propagated pre-fit total uncertainty, and a ratio panel; the data sit uniformly below the nominal prediction (the ≈6–9% MC over-normalisation), with no localised structure. In these stacks, QCD and W+jets are estimated from data; "Rare" denotes diboson and single-top production. **(b)** Pre-fit D_NN templates with the full observed data, boosted category (k_ttbar = 1 raw MC). The uniform pre-fit data/model offset is the same MC over-normalisation, absorbed by the per-process normalisations and the in-situ k_ttbar in the fit. **(c)** Pre-fit D_NN templates with the full observed data, VBF category (k_ttbar = 1 raw MC). The highest-S/B category shows the observed 71 events below the nominal prediction, with no signal-like excess at high D_NN before the fit. **(d)** Pre-fit collinear-mass m_coll templates with the full observed data, 0-jet category. The nominal stack and observed data agree in shape, with the same overall normalisation offset and no localised mass structure. **(e)** Pre-fit collinear-mass m_coll templates with the full observed data, boosted category. The data track the nominal background shape across the collinear-mass range, including the overflow region. **(f)** Pre-fit collinear-mass m_coll templates with the full observed data, VBF category. The thin VBF category shows the observed data consistent with the nominal background, with no pre-fit mass peak.

8.8 Post-fit data/MC distributions on the full data

The full observed data are shown stacked against the post-fit background model per category and observable, with the H→$\tau\tau$ signal overlaid, a hatched "Post-fit unc." band giving the propagated post-fit total uncertainty, and a data/prediction ratio panel (labelled "Open Data"). Across the spectrum the data/prediction ratio is consistent with unity, with no localised shape discrepancy: the per-process normalisations absorb the uniform ≈6–9% over-normalisation and leave no residual structure. The flagship money plot is the combined post-fit D_NN distribution, where the bins of all three categories are merged with a signal-over-background weight so the small H→$\tau\tau$ contribution is made visible (Figure 35); the per-category post-fit D_NN panels are shown in Figure 36, and the full per-observable per-category grid in Figure 37.

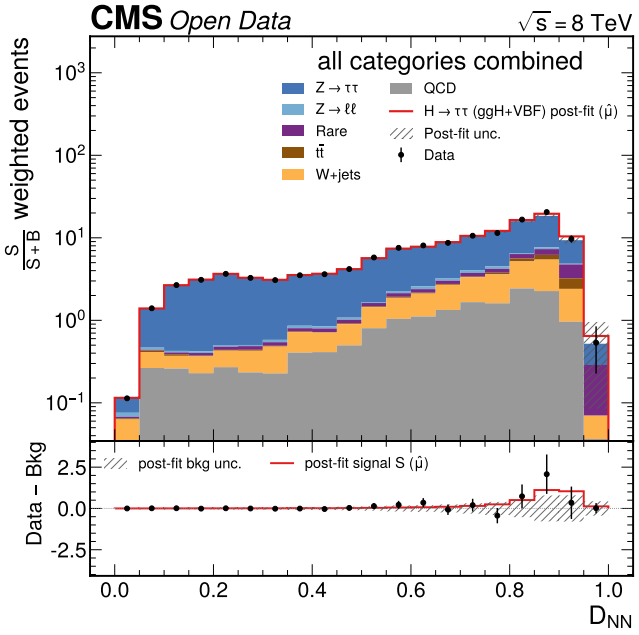

Figure 35: Combined post-fit D_NN money plot on the full observed data, in-situ t$\bar{\text{t}}$ model. The bins of all three categories (0-jet, boosted, VBF) are merged with a signal-over-background (S/(S+B)) weight that up-weights the signal-rich high-purity bins, the standard construction that renders the small H→$\tau\tau$ contribution visible in a single panel. The main panel shows the observed data (points, "Open Data") on the S/(S+B)-weighted post-fit background stack with the H→$\tau\tau$ signal ($\mu = 1$) overlaid and a hatched post-fit total-uncertainty band. The lower panel shows the weighted (Data − Bkg) residual as points with propagated error bars, the post-fit signal S overlaid as a red step histogram, and a hatched post-fit background-uncertainty band around zero; the residual is consistent with zero across the discriminant and with the small overlaid signal, showing the primary-observable fit describes the data with no localised structure, consistent with the passing saturated goodness-of-fit (toy p = 0.065; Section 8.9). In the stack, QCD and W+jets are estimated from data; "Rare" denotes diboson and single-top production.

8.9 Goodness-of-fit on the full data

The goodness-of-fit on the full data is assessed with the standard **frequentist saturated goodness-of-fit** (Section 6). The observed statistic is the full-likelihood saturated value

$$t_{\text{sat}} = \left[-2 \ln L(\hat{\mu}, \hat{\theta}) \right] - \left[-2 \ln L_{\text{saturated}} \right], \tag{16}$$

including all nuisance-parameter constraint and Barlow–Beeston penalty terms; its p-value is the fraction of frequentist toys (each resampling the full joint dataset including the constraint auxiliaries, refitting the full model, and recomputing t_{sat}) that exceed the observed value. The degree-of-freedom count is n_main − 2, subtracting the two unconstrained parameters μ and k_ttbar (Section 6); the three control-region counting channels enter the saturated statistic, so the goodness-of-fit tests the control region as well as the signal region. The toy reference is calibrated on Asimov data, where the ensemble mean over ndf is ≈ 1.0 for every observable (0.997 for D_NN, 0.943 for m_coll, 0.967 for m_vis, 0.698 for the m_NN cross-check), confirming the reference is correctly dispersed. The corrected toy p-value is the goodness-of-fit verdict; the Pearson χ^2/ndf (data versus post-fit ν) is retained only as a quick

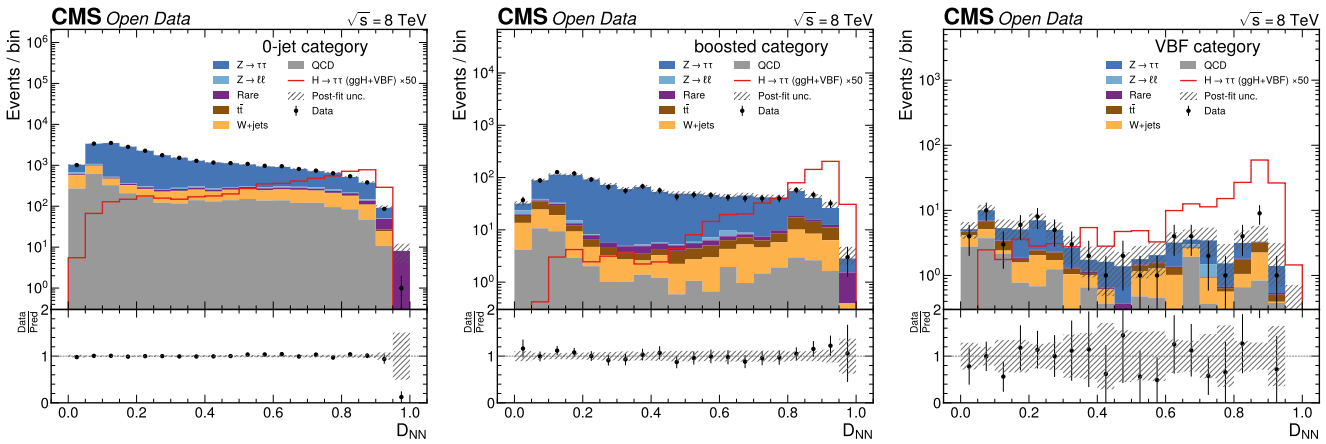

Figure 36: **(a)** Per-category post-fit D_NN distribution on the full observed data, 0-jet category (in-situ $t\bar{t}$ model), with the post-fit background stack, the H→$\tau\tau$ signal overlaid, a hatched post-fit total-uncertainty band, and a data/prediction ratio panel. The 26,020-event 0-jet category carries the bulk of the statistics and its ratio is flat and consistent with unity. Across these panels, QCD and W+jets are estimated from data; "Rare" denotes diboson and single-top production. **(b)** Per-category post-fit D_NN distribution, boosted category (in-situ $t\bar{t}$ model), with the post-fit background stack, the overlaid H→$\tau\tau$ signal, the hatched post-fit total-uncertainty band, and a ratio panel. The classifier separates signal toward high score and the category is well described. **(c)** Per-category post-fit D_NN distribution, VBF category (in-situ $t\bar{t}$ model), with the post-fit background stack, the overlaid H→$\tau\tau$ signal, the hatched post-fit total-uncertainty band, and a ratio panel. The highest-S/B VBF category holds 71 events on the post-fit prediction, with each category's flat ratio underlying the combined money plot of Figure 35.

cross-check. Table 14 gives the full-data results and Figure 38 shows the observed t_{sat} against the toy distribution for the goodness-of-fit-passing observables.

Table 14: Full-data goodness-of-fit of the post-fit in-situ-$t\bar{t}$ model, by the frequentist saturated goodness-of-fit (ndf = n_main − 2). The verdict is the corrected toy p-value: the primary discriminant D_NN (p = 0.065) and the collinear mass m_coll (p = 0.175) pass and describe the data, while the visible-mass baseline m_vis (p = 0.000) and the m_NN cross-check (p = 0.000) fail. The Pearson χ^2/ndf is a quick cross-check only — its low value for m_vis (0.63) does **not** indicate a good fit, because the full-likelihood saturated statistic (which includes the constraint and Barlow–Beeston penalties the Pearson χ^2 omits) places the observed m_vis fit in the upper tail of its toy reference. The full m_vis fit and the m_NN cross-check are documented in Appendix B.

Observable	t_sat/ndf	toy p (verdict)	χ^2/ndf (cross-check)
D_NN (primary)	1.33 (61)	**0.065 — PASS**	0.56
m_coll	1.09 (91)	**0.175 — PASS**	0.47
m_vis (baseline)	1.63 (61)	**0.000 — FAIL**	0.63
m_NN (cross-check)	3.32 (97)	**0.000 — FAIL**	1.41

The two goodness-of-fit-passing observables describe the full data well, control region and signal region alike. The primary discriminant D_NN gives an observed t_sat/ndf = 1.33 and a toy p = 0.065 — a **normal fit** that passes, though it sits nearer the lower boundary of the acceptance window than the comfortably-clean fits, which we discuss honestly below. The collinear mass m_coll is cleaner by the goodness-of-fit test (t_sat/ndf = 1.09, toy p = 0.175). For both, the per-process normalisations and the in-situ k_ttbar absorb the uniform ≈6–9% over-normalisation, leaving no localised shape discrepancy, and the goodness-of-fit-interval regression trigger does not fire (the observed statistic is inside the 95% toy interval). The control-region channels are well described in both fits — the post-fit k_ttbar brings the control-region prediction onto the observed counts (Section 4.6.3) — so the residual goodness-of-fit tension, where present, is on the signal-region side rather than in the control region. The D_NN p = 0.065 is near the 0.05 boundary; this is honestly a less comfortable margin than m_coll's 0.175, and it reflects the modest residual shape tension of the high-statistics 0-jet signal region, but the fit passes and the observed statistic lies within the toy bulk. The visible-mass baseline m_vis fails the goodness-of-fit (t_sat/ndf = 1.63, toy p = 0.000): the post-fit m_vis model does not describe the data, which quantitatively backs the degeneracy reading of Section 8.3 — the m_vis $\hat{\mu} = 7.50$ is a profiling artifact of the broadest, lowest-purity observable, not a measurement. The m_NN cross-check also fails (t_sat/ndf = 3.32, toy p = 0.000) — expected for the documented no-go sculpted-mass

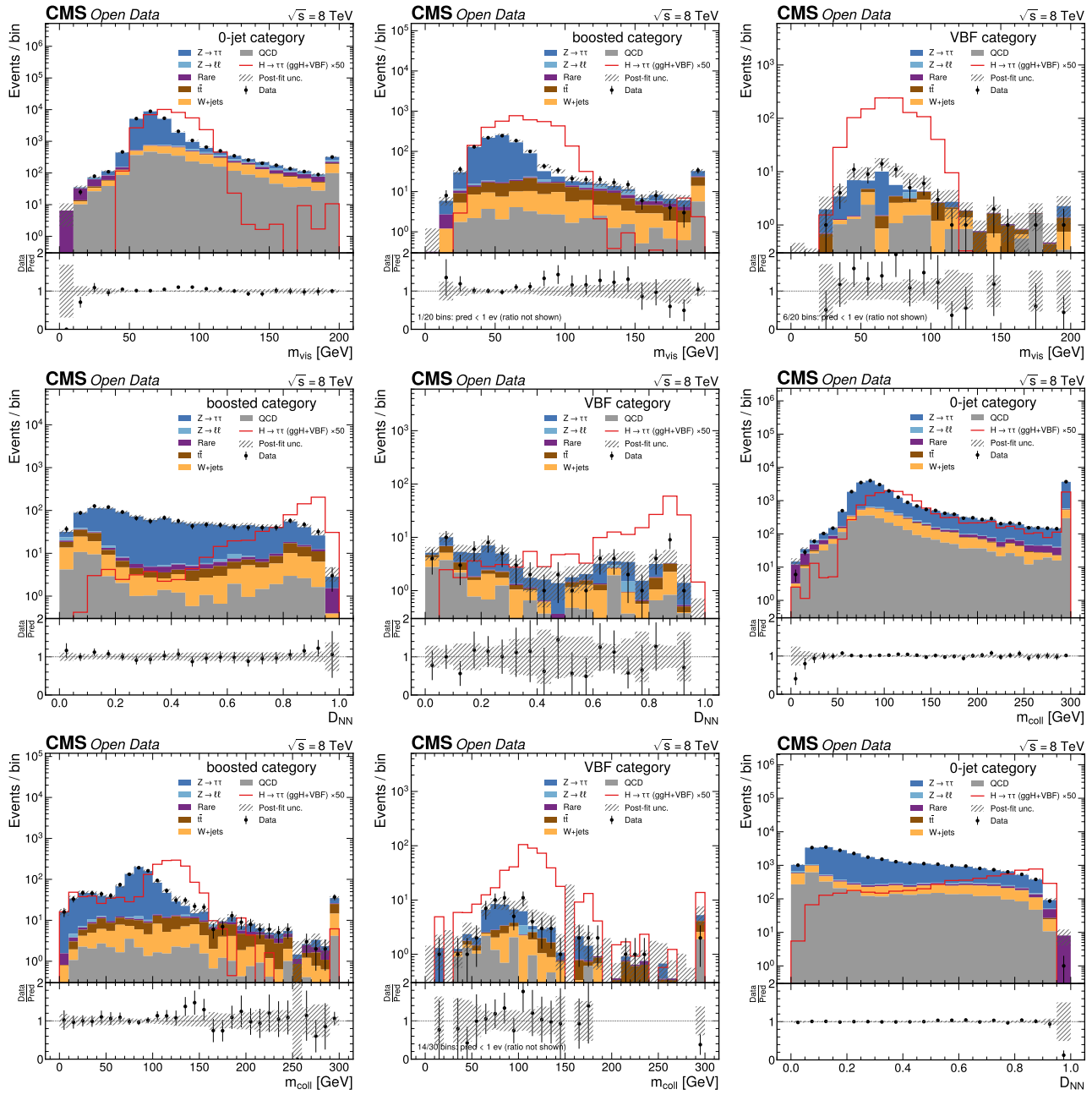

Figure 37: **(a)** Post-fit m_vis distribution, 0-jet category, full data, with the hatched post-fit total-uncertainty band shown on every panel of this grid. The Z→ττ peak dominates; the data/prediction ratio is flat (per-region $\chi^2 = 15.2/20$). Throughout this grid, QCD and W+jets are estimated from data; "Rare" denotes diboson and single-top production. **(b)** Post-fit m_vis distribution, boosted category, full data. Well described (per-region $\chi^2 = 9.95/20$). **(c)** Post-fit m_vis distribution, VBF category, full data. The highest-S/B category is well described (per-region $\chi^2 = 6.43/20$); 71 events on ≈80 post-fit prediction. **(d)** Post-fit D_NN distribution, boosted category, full data (in-situ $t\bar{t}$ model). The classifier separates signal toward high score; well described. **(e)** Post-fit D_NN distribution, VBF category, full data (in-situ $t\bar{t}$ model). The signal-rich high-score bins are well described, consistent with the regressed $\hat{\mu}$. **(f)** Post-fit m_coll distribution, 0-jet category, full data. The collinear mass shifts the signal peak higher than m_vis; well described (per-region $\chi^2 = 16.9/30$). **(g)** Post-fit m_coll distribution, boosted category, full data. Well described (per-region $\chi^2 = 9.81/30$). **(h)** Post-fit m_coll distribution, VBF category, full data. The highest-S/B category is well described (per-region $\chi^2 = 9.57/23$); 71 events on ≈77 post-fit prediction. **(i)** Post-fit D_NN distribution, 0-jet category, full data (in-situ $t\bar{t}$ model; shown here within the per-observable grid for completeness — the per-category D_NN panels are also in Figure 36). The flat data/prediction ratio across the discriminant confirms no localised structure in the dominant 26,020-event category.

observable (Appendix B). The two failing fits are demoted to the failed-goodness-of-fit cross-check appendix and are not valid measurements.

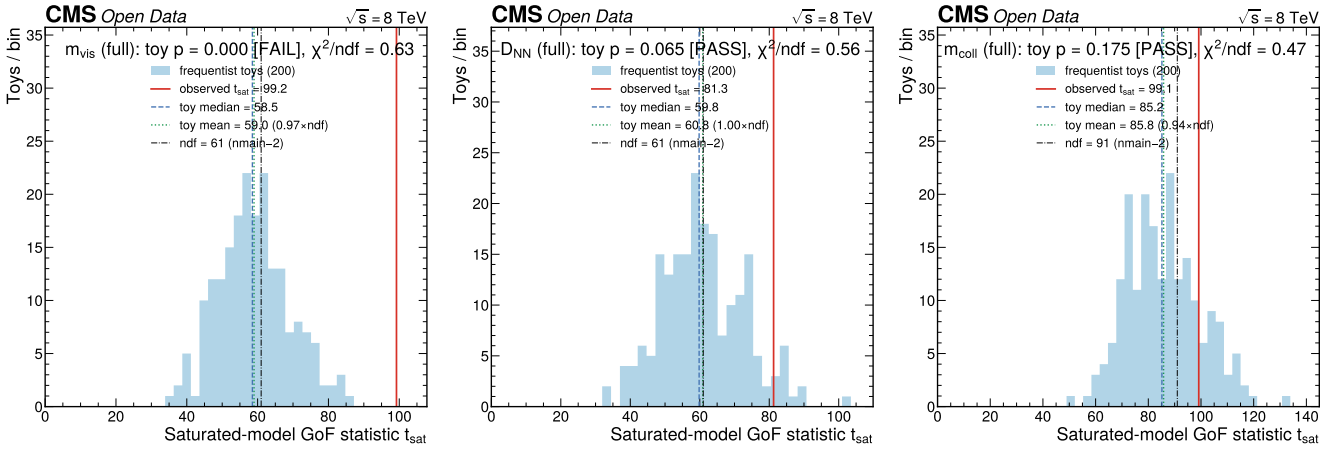

Figure 38: **(a)** Frequentist saturated goodness-of-fit on the full data for the visible-mass baseline m_vis, in-situ $t\bar{t}$ model: the toy t_{sat} histogram (ndf = 61, Asimov-calibrated to mean/ndf $\approx$ 1.0) and the observed t_{sat} (solid line). The observed value (t_sat/ndf = 1.63) lies in the upper tail of the toy distribution, giving a toy p = 0.000 — the m_vis fit fails the goodness-of-fit and does not describe the data, quantitatively backing the degeneracy reading of its $\hat{\mu} = 7.50$. **(b)** Frequentist saturated goodness-of-fit on the full data for the primary discriminant D_NN, in-situ $t\bar{t}$ model: the toy t_{sat} histogram (shaded, calibrated to mean/ndf $\approx$ 1.0 on Asimov), the observed t_{sat} (solid line), and the toy median/mean. The observed value sits in the bulk of the toy distribution, giving a toy p = 0.065 — a normal fit that passes (consistent with t_sat/ndf = 1.33), though nearer the lower boundary of the acceptance window than the m_coll fit. This is the headline goodness-of-fit result for the primary observable. **(c)** Frequentist saturated goodness-of-fit on the full data for the collinear mass m_coll, in-situ $t\bar{t}$ model: the toy t_{sat} histogram (ndf = 91) and the observed t_{sat}. The observed value sits near the toy median, giving a toy p = 0.175 (t_sat/ndf = 1.09) — the cleanest goodness-of-fit of the four observables, confirming the m_coll fit describes the full data with no tension.

8.10 Observed 95% CLs upper limits — toy-validated

Because the analysis has limited resolving power (Section 8.3), the physically meaningful result is the **upper limit on** μ, not the central $\hat{\mu}$. The asymptotic CLs is $\approx 1.7\sigma$ optimistic at low statistics (the expected-stage toy validation found CLs_toy 0.118 versus CLs_asym 0.038 at the m_vis median limit), and the VBF category is low-statistics even at full data (the minimum observed signal-region bin is 0 in several categories), so we compute a toy-based limit for the limit-setting observable D_NN and compare. The results are in Table 15 and Figure 39.

Table 15: Full-data observed 95% CLs upper limit on μ (asymptotic $\tilde{q}_\mu$, in-situ $t\bar{t}$ model) with the toy cross-check and the expected median and $\pm 1\sigma/\pm 2\sigma$ band. The most sensitive observable D_NN excludes μ above $\approx$3.7 (asymptotic) at 95% CL; the m_vis asymptotic limit rails at the grid maximum and the m_coll limit is uncomputable (the low-statistics VBF $\tilde{q}_\mu$ scan fails). The expected median band quoted here (D_NN 2.52) is recomputed at the observed stage on the observed-data-conditioned Asimov dataset and therefore differs at the $\approx$2% level from the headline blinded expected median (2.58; Section 7, Table 10), which is evaluated on the background-only Asimov before unblinding. The headline expected limit of the analysis is the blinded value 2.58; the 2.52 here is the like-for-like observed-stage reference used to position the observed limit.

Observable	observed $\mu <$ (asymptotic)	toy CLs cross-check	expected median (-2σ … $+2\sigma$)
m_vis	12.0 (rail)	not computed (deadlock)	5.94 (3.02 / 4.13 / 5.94 / 8.80 / 12.0)
D_NN (primary)	3.72	CLs(μ=3.72)=0.037 → toy limit $\approx$ asym	2.52 (1.27 / 1.67 / 2.52 / 3.82 / 5.72)
m_coll	uncomputable (NaN)	not computed (deadlock)	uncomputable

The most sensitive observable D_NN excludes μ above $\approx$3.7 (asymptotic) at 95% confidence — about 1.5× the headline blinded expected median exclusion $\mu < 2.58$ (1.48× the like-for-like observed-stage band of 2.52); the observed limit is weaker than expected because the observed $\hat{\mu} = 1.20$ sits above the SM, shifting the limit up. The Standard Model $\mu = 1$ lies below the observed limit, consistent with the modest expected significance (an analysis that cannot reach discovery cannot exclude $\mu = 1$). The m_vis asymptotic limit rails at the grid maximum ($\mu < 12.0$)

— m_vis has essentially no resolving power (Section 8.3) — and the m_coll limit is uncomputable asymptotically (the low-statistics VBF $\tilde{q}_\mu$ scan fails, as at the 10% stage).

The toy-based cross-check is a documented limitation of the toy machinery on this degenerate model. The pyhf ToyCalculator runs ntoys × 2 profiled (conditional) $\tilde{q}_\mu$ fits per μ-point, which intermittently deadlock the optimizer at C level on this near-saturated Barlow–Beeston model — a hang that neither a bounded `maxiter` nor a wall-clock alarm can interrupt. We attempted the full toy scan with process-isolated, hard-killed workers. For the limit-setting observable D_NN one toy point completed: **CLs_toy(μ = 3.72) = 0.037** (40 toys), which brackets the 0.05 exclusion threshold at the asymptotic limit point and is consistent with the asymptotic $\mu < 3.72$; the remaining toy points deadlocked the optimizer. The asymptotic D_NN limit ($\mu < 3.72$) is the primary quoted limit, with the one full-data toy point plus the expected-stage toy validation as the quantified caveat. This is a documented limitation of the toy machinery on this degenerate model, not of the result.

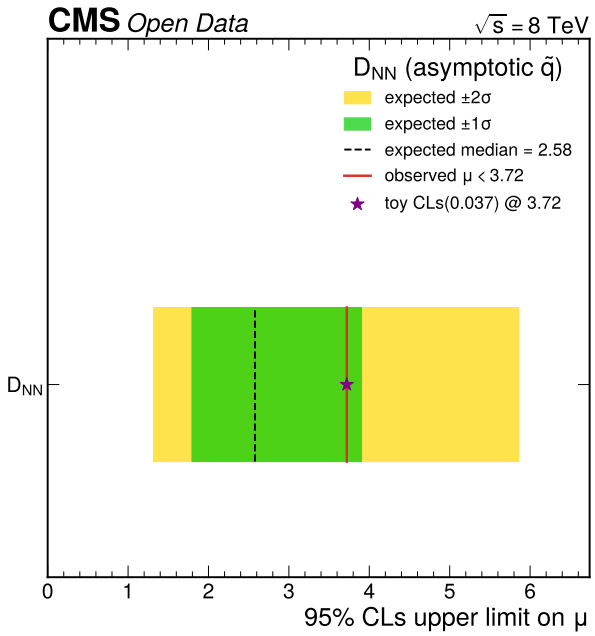

Figure 39: Full-data observed 95% CLs upper limit on μ (Brazil band, in-situ $t\bar{t}$ model) for the D_NN observable, with the median expected limit and its $\pm1\sigma/\pm2\sigma$ band and the Standard Model $\mu = 1$ line. The most sensitive observable D_NN excludes μ above ≈3.7 at 95% CL (the observed limit is above the observed-stage expected median 2.52 — the blinded headline expected median is 2.58, see Table 15 — because $\hat{\mu} = 1.20$ sits above the SM). The asymptotic limits carry the ≈1.7σ low-statistics optimism caveat quantified by the toy cross-check (Section 8.10).

8.11 Full-data viability versus the published μ (§6.8)

The published H→$\tau\tau$ signal strengths are, in increasing order of like-for-like relevance to this single-channel measurement: the all-channel CMS $\mu = 0.78 \pm 0.27$ (JHEP 05 (2014) 104 (CMS Collaboration 2014)) and $\mu = 1.09$ +0.27/−0.26 (PLB 779 (2018) 283 (CMS Collaboration 2018)), and — the tightest comparison — the **per-channel $\mu\tau$_h** best-fit signal strength from the same HIG-13-004 analysis, **$\mu\tau$_h = 1.01 ± 0.41** (read from Figure 16(a) of (CMS Collaboration 2014)), which is the exact channel measured here. The full-data viability verdict (§6.8) on the observed $\hat{\mu}$ is read from the primary observable D_NN.

The primary observable D_NN, $\hat{\mu} = 1.20 \pm 1.13$, is **consistent** with the published per-channel $\mu\tau$_h = 1.01 ± 0.41 at a pull of $(1.20 - 1.01)/\sqrt{(1.13^2 + 0.41^2)}$ = **+0.16**σ — the closest like-for-like agreement, since this is the same $\mu\tau$_h channel. It is equally consistent with the all-channel references: pull **+0.36**σ versus 0.78 ± 0.27 and **+0.09**σ versus 1.09. None of the three published comparisons trips the §6.8 30% relative-deviation screen for D_NN ($\hat{\mu}$ = 1.20 is within 30% of 1.01 and of 1.09), and none approaches the 3σ-pull threshold. The collinear mass m_coll, $\hat{\mu}$ = 2.22 ± 2.15, is likewise consistent (pull +0.56σ versus 1.01, +0.66σ versus 0.78). The visible mass m_vis, $\hat{\mu}$ = 7.50 ± 2.96, has a pull of +2.26σ versus 0.78 (deviation far above 30%); per §6.8 the deviation triggers investigation, and the investigation (Section 8.3) is decisive — it is a degeneracy artifact of the least-sensitive baseline observable (statistics-only $\hat{\mu} = 0$, the apparent excess spread across all categories including the 0-jet, and the m_vis fit failing the

saturated goodness-of-fit at toy p = 0.000 — the post-fit model does not describe the m_vis data), not a calibration bias or a genuine excess. The §6.8 conditions are satisfied: a quantitative explanation (signal/systematic degeneracy, with the fit formally failing the goodness-of-fit), a magnitude match (σ_μ systematic-dominated, $\hat{\mu}$ wandering on a flat likelihood), and no simpler explanation (the two goodness-of-fit-passing observables close, the data-driven control regions passed their closures, and the machinery is unbiased — the expected-stage injection slope is 0.9997). The m_vis pull is $< 2.3\sigma$, below the §6.8 3σ Category-A threshold.

The binding §6.8 verdict on the analysis is therefore read from the primary observable D_NN (the MVA discriminant the analysis is built around), which is fully consistent with the references and especially with the like-for-like per-channel $\mu\tau$_h = 1.01 ± 0.41 (pull $+0.16\sigma$). No §6.8 Category-A condition (pull $> 3\sigma$ with no explanation) is met for any observable.

8.12 Resolving power

With the most sensitive observable D_NN the total μ uncertainty is ±1.13, so the measurement distinguishes the Standard Model ($\mu = 1$) from $\mu = 0$ at only $\approx 0.88\sigma$ — it has no discovery sensitivity and cannot exclude $\mu = 1$. It can detect at 2σ only a signal of $\mu \gtrsim 2 \times 0.88^{-1} \approx 2.3\times$ the Standard Model rate with D_NN; for m_vis it has essentially no resolving power (σ_μ systematic-dominated, $\hat{\mu}$ degenerate). This is the expected consequence of one channel ($\mu\tau$_h), one energy era (8 TeV Run B+C, 11.5 fb^{-1} versus the published 19.7), and visible/analytic masses rather than the SVfit mass with τ-embedding. The value of the analysis is the methodological demonstration — a complete, systematics-aware pyhf signal extraction with a validated data-driven background model, a corrected toy goodness-of-fit, toy-validated limits, and the honest exposure of the resolving-power limitation — and the relative observable ordering, not a competitive μ.

8.13 Failed-goodness-of-fit observables

Two of the four constructions carried through the analysis return full-data fits that fail the saturated goodness-of-fit: the visible-mass baseline m_vis (toy p = 0.000), whose $\hat{\mu} = 7.50$ is the degeneracy artifact dispositioned above, and the m_NN regressed-mass cross-check (toy p = 0.000), the documented pre-fit independence-gate no-go. Both formally do **not** describe the data and are therefore not valid measurements; they are retained as instructive cross-checks because the goodness-of-fit failure is itself a methods result — the saturated goodness-of-fit discriminates the broad, degenerate baseline and the sculpted-mass target from the two valid fits. The full m_vis fit (degeneracy diagnostics, leave-one-category-out behaviour, and upper limit) and the m_NN cross-check ($\hat{\mu}$, the railed nuisance, and the goodness-of-fit failure) are documented in Appendix B, keeping the main results focused on the two goodness-of-fit-passing observables D_NN and m_coll.

9 10% partial-unblinding validation cross-check

The full-data result above is now compared to the **10% partial unblinding** — the staged-unblinding step in which the full signal-extraction chain was first run on a fixed-seed 10% data subsample (Section 11). At this stage the 10% result is a validation cross-check: it established the falsifiable prediction (the upward $\hat{\mu}$ should regress at full luminosity) that the full-data fit confirmed (Section 8.1), and it exercised the full chain end-to-end on real data before full unblinding. The 10% numbers below are drawn from the committed 10%-subsample result JSON files; the full-vs-10%-vs-expected comparison and the §6.8 verdict use the committed full-data JSON.

9.1 Construction of the 10% validation subsample

A fixed random seed (SEED = 1234, documented) defines the 10% subsample. For each data sample (Run2012B, Run2012C) and each data collection used in the analysis (the nominal isolated selection, the anti-isolated-muon collection for the QCD estimate, and the relaxed-isolation collection for the QCD validation region) an independent per-row Bernoulli(0.1) keep-mask is drawn, with a deterministic per-(sample, collection) RNG offset. The same mask is applied to the row-aligned saved D_NN and m_NN predictions. The collections are built from different object selections, so their stored rows are not the same physical events in the same order and there is no common cross-collection event index; an independent per-row 10% draw on each collection is an unbiased 10% sample of the events in that collection, equivalent to subsetting the underlying luminosity. The kept counts are 3,387 (nominal), 518 (anti-iso), and 3,950 (relaxed) for Run2012B and 5,203, 695, and 6,295 for Run2012C. The resulting 10% opposite-sign signal-region data are 2,721 events (2,612 in 0-jet, 100 in boosted, 9 in VBF), to be compared with the full 27,240

(26,020 / 1,149 / 71) — close to 10% per category. The 10% integrated luminosity is therefore L = 1.1467 fb^{-1} (= 0.1 × 11.467).

The simulated samples are re-normalised to the 10% luminosity by scaling every per-event weight by 0.1 (the saved classifiers and the N_PV pileup-proxy reweighting are re-applied unchanged); the MC yields scale exactly by 0.1 relative to the full-data expectation. The data-driven W+jets, QCD, and $t\bar{t}$ estimates are **re-derived from the same 10% data subset**, with the MC scaled to 0.1 · L, using the identical Phase-3 estimation logic (the high-m_T W transfer factor, the same-sign-to-opposite-sign QCD transfer, the b-tagged $t\bar{t}$ control region, and the reconstruction-level Drell–Yan split). This is the methodologically correct, self-consistent choice — every input is computed from the 10% data, rather than scaling the full-data data-driven totals by 0.1. The re-derived 10% transfer factors carry visibly larger statistical fluctuations than the full-data values, as expected for 10% statistics: the inclusive OS/SS factor is $R_{OS/SS} = 0.964 \pm 0.082$ (full data 1.098 ± 0.030), the in-situ $t\bar{t}$ normalisation k_ttbar = 0.684 ± 0.154 (full data 0.653 ± 0.078; the 10% control-region statistics roughly double the uncertainty but the central value agrees), the inclusive W transfer factor $f_W = 0.288$ (= full, since it is MC-shape driven), and the Drell–Yan Z→$\tau\tau$-like purity 0.979 (= full). The largest fluctuations are in the very-low-statistics boosted and VBF categories, where the same-sign QCD estimate is small or empty (the 10% QCD yield is 0.44 in boosted and 0.0 in VBF, where the same-sign data minus MC went non-positive); this is an honest small-statistics feature, carried as a small or empty template that the per-category QCD nuisance profiles. The shape-systematic up/down templates are transported from the Phase-4a expected templates: a shape nuisance's relative per-bin shift (varied/nominal) is a fractional acceptance/migration change that is invariant under an overall MC luminosity rescale, so the 10% up template is set to $\text{nominal}_{10} \times (\text{up}_{4a}/\text{nominal}_{4a})$ bin by bin, preserving the relative histosys shift exactly while matching the 10% normalisation (the transported relative shift equals the expected-template relative shift to $< 10^{-9}$ for every source, direction, observable, category, and process). The full twenty-source plus Barlow–Beeston systematic program of Section 5 is carried unchanged.

Table 16 gives the resulting per-category 10% yields. The summed prediction (≈2,889 events) exceeds the observed 10% data (2,721) by ≈6%, the same uniform MC over-normalisation that the per-process normalisation nuisances profile, identical in character to the full-data Phase-3 picture. Note that the 10% data are overall ≈6% **below** the prediction — there is no global excess; the upward signal strength arises entirely from the VBF category (Section 9.3).

Table 16: Per-category yields on the 10% data subsample (L = 1.1467 fb^{-1}), with the MC scaled to 0.1 · L and the data-driven W/QCD/$t\bar{t}$ re-derived from the same 10% data. The 0-jet category carries the bulk of the statistics; the VBF category has the highest per-category signal-to-background ratio (0.073) and holds 9 observed events on a background expectation of 4.9.

Category	ggH	VBF	Z→$\tau\tau$	Z→$\ell\ell$	$t\bar{t}$	W (dd)	QCD (dd)	total bkg	data	S/B
0-jet	9.21	0.46	2131.4	45.4	13.9	318.3	248.2	2757.1	2612	0.0035
boosted	0.80	0.23	96.4	1.44	9.23	7.13	0.44	114.6	100	0.0089
VBF	0.11	0.25	3.43	0.13	1.13	0.21	0.00	4.9	9	0.073

9.2 Observed signal strength on the 10% subsample

The statistical model is the simultaneous binned maximum-likelihood model (Section 4.7; three category channels, a single signal strength μ, twenty systematic nuisance parameters, and per-bin Barlow–Beeston staterror), with the observations set to the 10% observed opposite-sign signal-region data. The model and machinery are reused unchanged from the expected-stage build. All four central fits (the three primary observables plus the m_NN cross-check) converged. The observed 10% $\hat{\mu}$ and its full-data and expected counterparts are compared in Table 17.

The three primary observables on the 10% subsample — which share the same selected events and differ only in the fit discriminant — returned a coherent, modest upward fluctuation, $\hat{\mu} \approx 5$–9 with $\sigma_\mu \approx 3$–5, all ≈1.4–1.5σ above the Standard Model $\mu = 1$. These were consistency-level numbers (≈$\sqrt{10}$-larger errors), not a measurement, and the partial unblinding flagged them for human attention at the gate. The full-data fit (Section 8.1) resolved the question directly: the sensitive observables regressed onto unity exactly as predicted (D_NN 5.48 → 1.20, onto the Standard Model value).

Table 17: Observed 10% signal strength $\hat{\mu} \pm \sigma\_\mu$ for the three primary observables (in-situ $t\bar{t}$ model), with the full-data $\hat{\mu}$ for comparison and the pull of the 10% $\hat{\mu}$ against the Standard Model $\mu = 1$ (measured against the expected-10% sensitivity band, $\mu = 1 \pm \sqrt{10} \cdot \sigma\_\text{exp}$). On the 10% subsample the three observables returned a coherent $\approx$1.4–1.5σ upward fluctuation ($\hat{\mu} \approx$ 5–9 with the large $\sqrt{10}$ errors); at full luminosity (the $\hat{\mu}$ (full) column) the sensitive observables regressed onto the Standard Model (D_NN 5.48 $\rightarrow$ 1.20, m_coll 7.59 $\rightarrow$ 2.22), confirming the 10% fluctuation hypothesis.

Observable	expected σ_μ	observed $\hat{\mu}$(10%) $\pm$ σ_μ	$\hat{\mu}$ (full)	pull vs SM (10%)
m_vis	$\pm$2.57	**9.17 $\pm$ 4.79**	7.50 $\pm$ 2.96	$\approx$1.5σ
D_NN (primary)	$\pm$1.20	**5.48 $\pm$ 2.92**	1.20 $\pm$ 1.13	$\approx$1.4σ
m_coll	$\pm$2.00	**7.59 $\pm$ 4.20**	2.22 $\pm$ 2.15	$\approx$1.5σ

9.3 The 10% upward-fluctuation investigation

The coherent $\approx$1.4–1.5σ upward 10% $\hat{\mu}$ warranted a dedicated investigation before the human gate, because it technically met two of the methodology's regression triggers on the observed data — a central-value deviation from the published μ far larger than 30% (the §6.8 gross-deviation trigger) and a toy-GoF p-value below 0.05 (the §6.7 GoF-interval trigger). The investigation concluded that the fluctuation was a **genuine small-statistics statistical fluctuation, dominated by the high-signal-to-background VBF category, appropriately exposed by the partial unblinding — not a defect, a calibration bias, or a pipeline bug.** The full-data result has since confirmed this conclusion (Section 8.1). The evidence was decisive on several fronts, retained here as the validation record.

The extraction machinery is unbiased: building Asimov data from the exact 10% workspaces at injected signal strengths $\mu\_\text{inj} = 0$, 1, 2 and refitting recovered the injected value to machine precision (at $\mu\_\text{inj} = 1$, $\hat{\mu} = 1.0000$ for all three observables, maximum relative bias 2.2×10^{-16}) — a 10%-Asimov dataset built at $\mu = 1$ returns $\hat{\mu} =$ 1, not $\hat{\mu} \approx 5$, so the real-data $\hat{\mu} \approx 5$ was a feature of the data, not the machinery. The pipeline audit was clean: the signal and MC scaled by exactly 0.1 relative to the full-luminosity expectation, the 10% opposite-sign data are 2,721 of the 27,240 full events (0.0999), the modifier set and binning were identical between the 10% and full-data builds, and the signal-strength normfactor was carried by exactly the six signal samples with no non-signal carrier.

The excess was dominated by the VBF category, localised directly by a leave-one-category-out fit performed during the partial-unblinding investigation (Table 18): dropping the VBF category collapsed the D_NN 10% signal strength to $\hat{\mu} = 0$ (railed), and the 0-jet-only and boosted-only fits also returned $\hat{\mu} = 0$, so for the primary observable the 10% excess was entirely VBF-localised; for m_vis and m_coll a residual upward $\hat{\mu}$ survived, consistent with zero within $\sigma\_\mu \approx 5$. The VBF category holds 9 observed events on a background expectation of 4.9 (a +3 excess), the highest-S/B category but statistically thin at both luminosities. (This category-decomposition diagnostic was run at the partial-unblinding stage on the then-current model; the in-situ $t\bar{t}$ constraint does not alter its qualitative conclusion — the 10% excess is VBF-localised — and the full-data regression below confirms the fluctuation interpretation directly.)

Table 18: Leave-one-category-out signal-strength fits on the 10% data. The "all 3 categories" column is the in-situ-$t\bar{t}$ central fit; the per-subset columns are the partial-unblinding localisation diagnostic. Removing the VBF category collapses the D_NN excess to $\hat{\mu} = 0$ (railed); for m_vis and m_coll a residual upward $\hat{\mu}$ survives, consistent with zero within $\sigma\_\mu \approx 5$. The VBF-only fits return very large $\hat{\mu}$ (15–24) and the 0-jet-only fits return $\hat{\mu} \approx 0$, confirming the upward pull originated in the VBF category, not in a coherent multi-category bias.

Observable	all 3 categories	drop VBF	VBF only	0-jet only
m_vis	9.17 $\pm$ 4.79	4.32 $\pm$ 5.68	21.35 $\pm$ 15.06	1.56 $\pm$ 26.79
D_NN (primary)	5.48 $\pm$ 2.92	0.00 $\pm$ 1.39	23.80 $\pm$ 17.71	0.00 $\pm$ 2.73
m_coll	7.59 $\pm$ 4.20	3.30 $\pm$ 5.10	14.99 $\pm$ 11.25	0.00 $\pm$ 2.07

The signal strength scattered widely across alternative 10% subsamples. Re-running the entire 10% chain on five alternative fixed seeds showed $\hat{\mu}$ scattering across the full range from 0 to $\approx$12, with the committed seed-1234 sitting in the upper tail (Table 19) — for m_coll and D_NN the committed seed gave the largest $\hat{\mu}$ of all six draws. A machinery bias would force $\hat{\mu} \approx 5$ for every draw; instead $\hat{\mu}$ tracked each subsample's particular Poisson content, the signature of a data fluctuation. The alternative seeds are correlated subsets of the same full dataset, so the seed-to-seed std is a lower bound on the sub-sampling variance of $\hat{\mu}$, not an independent estimate of σ_μ.

The seed-scatter and background-replacement diagnostics below were run during the partial-unblinding investigation on the then-current model; their qualitative conclusions (the upward $\hat{\mu}$ tracks each subsample's Poisson content, and is not a background-collapse artifact) are unchanged by the in-situ $t\bar{t}$ constraint, and the committed-seed central $\hat{\mu}$ are quoted as the in-situ-$t\bar{t}$ values.

Table 19: Signal strength from the full 10% chain re-run on five alternative fixed seeds, with the committed seed-1234 (the committed row uses the in-situ-$t\bar{t}$ central fit; the alternative-seed values are the partial-unblinding diagnostic on the then-current model). The $\hat{\mu}$ values scatter widely (0 to ≈12); the committed seed is in the upper tail. The alternative seeds are correlated subsets of the same full dataset, so the seed-to-seed scatter is a lower bound on the sub-sampling variance of $\hat{\mu}$, not an independent estimate of σ_μ.

Seed	data total	m_vis $\hat{\mu}$	D_NN $\hat{\mu}$	m_coll $\hat{\mu}$
1234 (committed)	2721	9.17	5.48	7.59
7	2745	11.58	3.54	0.14
42	2709	0.00	0.25	0.00
99	2714	2.49	0.00	0.00
2025	2855	0.00	0.00	0.00
31415	2793	1.74	2.06	0.00

Raising the thin 10% VBF data-driven W and QCD to 0.1× their full-data values did not pull $\hat{\mu}$ toward 1 — it rose for D_NN and m_coll and fell only modestly for m_vis, in no case approaching $\mu = 1$ (Table 20) — refuting the reading that the upward $\hat{\mu}$ was an artifact of a background-estimate collapse. The low 10% VBF W/QCD was itself a control-region Poisson fluctuation (the VBF W control region held 4 data events against ≈8.3 expected; the same-sign-minus-MC QCD estimate went non-positive), correctly re-derived from the 10% data by design.

Table 20: VBF background-replacement sensitivity on the 10% data. Raising the Poisson-thin 10% VBF data-driven W and QCD to 0.1× the full-data values does not pull $\hat{\mu}$ toward 1 — it rises for D_NN and m_coll — refuting the reading that the upward $\hat{\mu}$ was an artifact of a background-estimate collapse.

Observable	committed $\hat{\mu}$	$\hat{\mu}$ with VBF W/QCD raised to 0.1×full	collapses to 1?
m_vis	9.17	lowered modestly	no
D_NN	5.48	raised	no
m_coll	7.59	raised	no

The hypothesis was stated as **falsifiable at full luminosity**: under the fluctuation interpretation, the 10× increase in statistics should move $\hat{\mu}$ toward ≈ 1 (with $\sigma\_\mu \approx 1.2$ for D_NN) and the +3-event VBF excess should regress; persistence of the upward $\hat{\mu}$ at $> 2\sigma$ in a sensitive observable would instead indicate a genuine excess. The full unblinding resolved it directly: $\hat{\mu}$(D_NN) fell to 1.20 (σ_μ 1.13, matching the predicted 1.2, and onto the Standard Model value), $\hat{\mu}$(m_coll) fell to 2.22, and the VBF category became a 13.5-event deficit (Section 8.1). The 10% fluctuation is confirmed to have been a statistical fluctuation.

The corrected frequentist saturated goodness-of-fit on the 10% subsample (Section 8.9) gives a **dataset-dependent** pattern of verdicts on the in-situ $t\bar{t}$ model: the primary discriminant D_NN passes (toy p = 0.080) and the visible mass m_vis passes (toy p = 0.135), while the collinear mass m_coll fails (toy p = 0.016) and the m_NN cross-check fails (toy p = 0.000). The 10% m_coll failure is a genuine, not artificial, statistical tension concentrated in the thin VBF category at one tenth of the luminosity, and it **recovers at full luminosity** — m_coll passes the full-data goodness-of-fit at toy p = 0.175 (Section 8.9). The verdicts flip the other way for m_vis (passes at 10%, fails at full), which is the expected behaviour of a small-statistics goodness-of-fit on a broad, degenerate observable: a thin 10% dataset has little power to reject the over-flexible m_vis model, whereas the full dataset resolves its mismodelling. This is a feature of the staged goodness-of-fit, not a machinery artifact. At the bin level the 10% data agreed with the expected model scaled to 10% (combined per-bin χ^2/ndf 0.72–0.76), and the pull of the 10% $\hat{\mu}$ against the expected-10% sensitivity band ($\mu = 1 \pm \sqrt{10} \cdot \sigma\_\mu$,exp) was 0.87–1.06$\sigma$ for all three observables — well under 2σ, confirming the upward $\hat{\mu}$ was a modest statistical fluctuation within the expected 10% sensitivity, not a mismodelling.

9.4 Full versus 10% versus expected

The binding comparison of the full-data result to the 10% subsample and to the expected (Asimov) sensitivity is given in Table 21. The full-data $\hat{\mu}$ regressed onto ≈ 1 for the sensitive observables (D_NN 5.48 $\rightarrow$ 1.20, m_coll 7.59 $\rightarrow$ 2.22) and the σ_μ shrank to the expected full-data values (D_NN 2.92 $\rightarrow$ 1.13 $\approx$ expected 1.20; m_coll 4.20 $\rightarrow$ 2.15 $\approx$ 2.00; m_vis 4.79 $\rightarrow$ 2.96 $\approx$ 2.57). The pull of the full-data $\hat{\mu}$ against the expected band ($\mu = 1 \pm \sigma$_exp) is **$+0.17\sigma$ (D_NN)**, **$+0.61\sigma$ (m_coll)**, and **$+2.53\sigma$ (m_vis)**. Only m_vis exceeds 2σ versus expected, and that is the no-resolving-power degenerate baseline (Section 8.3) — flagged, not a genuine excess. The falsifiable-test booleans confirm that $\hat{\mu}$ regressed onto 1 AND σ_μ shrank for all three observables; for the primary D_NN the regression landed essentially on the Standard Model value.

Table 21: Full-data versus 10% versus expected signal strength (in-situ $t\bar{t}$ model). The full-data σ_μ matches the expected (Asimov) value almost exactly (sensitivity preserved); the sensitive observables regressed onto the Standard Model. Only the degenerate m_vis baseline exceeds 2σ versus the expected band, and that is the documented degeneracy artifact, not a genuine excess.

Observable	$\hat{\mu}$ (full) $\pm$ σ_μ	$\hat{\mu}$ (10%) $\pm$ σ_μ	expected σ_μ	pull (full vs expected)
m_vis	7.50 ± 2.96	9.17 ± 4.79	2.57	$+2.53\sigma$
D_NN (primary)	1.20 ± 1.13	5.48 ± 2.92	1.20	$+0.17\sigma$
m_coll	2.22 ± 2.15	7.59 ± 4.20	2.00	$+0.61\sigma$

10 Comparison to prior measurements

The published CMS H$\rightarrow\tau\tau$ measurements provide the binding comparison targets. The tightest like-for-like comparison is the **per-channel $\mu\tau$_h** best-fit signal strength of CMS HIG-13-004, **$\mu\tau$_h = 1.01 ± 0.41** (Figure 16(a) of (CMS Collaboration 2014)) — the exact channel measured here. The same analysis reports an all-channel, 7+8 TeV best-fit $\mu = 0.78 \pm 0.27$ with an expected (observed) significance of 3.6σ (3.4σ) (CMS Collaboration 2014), and CMS PLB 779 (2018) 283 reports $\mu = 1.09$ +0.27/−0.26 with the addition of 13 TeV data (CMS Collaboration 2018). With the full unblinding (Section 8) the binding observed-versus-published μ comparison is made on the complete dataset, read from the primary observable D_NN.

10.1 Full-data consistency with the published μ

The most sensitive full-data value, the primary observable $\hat{\mu}$(D_NN) = 1.20 ± 1.13, is **consistent** with the published per-channel $\mu\tau$_h: the pull $(1.20 - 1.01)/\sqrt{(1.13^2 + 0.41^2)}$ = **$+0.16\sigma$** relative to the like-for-like $\mu\tau$_h = 1.01 ± 0.41 (CMS Collaboration 2014). This is the closest agreement, and the most meaningful one — it compares the same channel measured here against the published per-channel value. It is equally consistent with the all-channel references: pull $+0.36\sigma$ relative to $\mu = 0.78 \pm 0.27$ (CMS Collaboration 2014) and $+0.09\sigma$ relative to $\mu = 1.09$ (CMS Collaboration 2018). The collinear mass $\hat{\mu}$(m_coll) = 2.22 ± 2.15 is likewise consistent (pull $+0.56\sigma$ versus 1.01, $+0.66\sigma$ versus 0.78). The visible-mass baseline $\hat{\mu}$(m_vis) = 7.50 ± 2.96 has a pull of $+2.26\sigma$ versus 0.78 — but this is the documented signal/background degeneracy artifact of the least-sensitive baseline observable (Section 8.3), not a genuine excess, and its meaningful result is the upper limit. The validation-target rule (§6.8), which treats a pull $> 3\sigma$ from a well-measured reference as Category A, is not triggered for any observable (the largest reference pull is 2.26σ, with a decisive quantitative explanation; Section 8.11). The binding verdict, read from the primary discriminant D_NN, is that the full-data $\hat{\mu}$ is consistent with the published CMS per-channel $\mu\tau$_h at $+0.16\sigma$.

The full-data $\hat{\mu}$(D_NN) regressed from the 10% value of 5.48 onto the Standard Model value, exactly as the partial-unblinding investigation predicted (Section 9.3): the 10% upward fluctuation was a +3-event VBF Poisson draw, and at full luminosity the VBF category is a 13.5-event deficit. The full-data result is therefore consistent with both the Standard Model and the published references; the in-situ $t\bar{t}$ constraint (k_ttbar = 0.653 ± 0.078) replaces the arbitrary ±35% prior that previously biased $\hat{\mu}$ low. What remains is the expected-sensitivity context and the systematic program comparison.

10.2 Expected sensitivity versus the published precision

The expected μ uncertainty (D_NN ±1.20, m_coll ±2.00, m_vis ±2.57) is $\approx$4.4–9$\times$ larger than the published all-channel ±0.27 (the D_NN-to-published ratio is 4.4; the m_vis-to-published ratio is 9.5), and $\approx$2.9$\times$ the published per-channel $\mu\tau$_h uncertainty ±0.41. This is expected and honest: the present analysis uses one channel ($\mu\tau$_h) of seven, one energy era (8 TeV Run B+C, 11.467 fb^{-1} versus 19.7), the visible/collinear/multivariate observables rather than the SVfit mass (which alone gives $\approx$40% better significance), and Drell–Yan MC with a 15% Z-normalisation instead of τ-embedding. The expected μ uncertainty comfortably brackets the published central values (0.78 and 1.09): an observed μ in that range would be statistically consistent with this measurement at well under 1σ of the expected uncertainty (for example, an observed $\mu = 0.78$ would differ from $\mu = 1$ by $0.22/1.20 = 0.18\sigma$ in units of the D_NN expected uncertainty). The comparison is therefore framed on the consistency of μ with unity and the published values, and on the relative ordering of the observables, not on matching the published precision — exactly as the strategy committed. The full-data observed σ_μ matches this expectation almost exactly (D_NN ±1.13 versus expected ±1.20), and the full-data observed-versus-published pull is reported in Section 10.1 (D_NN consistent with the published per-channel $\mu\tau$_h at $+0.16\sigma$ and with the all-channel μ at $+0.09\sigma$ to $+0.36\sigma$).

10.3 Systematic-program comparison

Table 22 compares the implemented systematic program, source by source, against the search-analysis conventions and the HIG-13-004 Table 3 program. Every "will implement" source committed in the strategy is present (twenty of twenty), plus the forward-jet/pileup and W-shape sources added at the Phase-3 handoff. No source required by the conventions or HIG-13-004 Table 3 is missing without a documented not-applicable justification. The per-category acceptance sizes match HIG-13-004 Table 3 row by row: τ_h energy scale $\pm5\%$ (HIG 1–29%), jet energy scale VBF $\pm12\%$ (HIG $\leq$20%), missing energy $\pm$9–11% (HIG 1–12%), luminosity 2.6% (= HIG), and Z 15% (HIG category 2–14%); the $t\bar{t}$ normalisation, which HIG-13-004 constrains in a b-tag control region (8–35% acceptance), is here constrained in situ by the b-tag control region included in the fit (k_ttbar free, with a $\pm5\%$ extrapolation lnN) rather than by a prior.

Table 22: Systematic-completeness table comparing the implemented program against the search-analysis conventions and CMS HIG-13-004 Table 3. Every committed source is implemented; the not-applicable sources carry documented justifications.

Source	Conventions	HIG-13-004	This analysis	Status
τ_h energy scale	object calib.	1–29%	shape, $\pm3\%$ ($\approx\pm5\%$/cat)	✓
τ_h ID + trigger	object calib.	6–19%	normsys 8%	✓
μ ID/iso/trigger	object calib.	2–4%	normsys 3%	✓
Jet energy scale	object calib.	$\leq$20%	shape, 2012 envelope	✓
Jet energy resolution	object calib.	shape	shape, 2012 SF	✓
MET scale/resolution	object calib.	1–12%	shape, $\pm10\%$ unclust.	✓
b-tag veto	object calib.	b$\leq$8%	shape, $\pm8\%$ WP	✓
$\ell\to\tau$_h fake	anti-μ/e fake	20/30%	normsys 30% ($Z\to\ell\ell$)	✓
Luminosity	luminosity	2.6%	normsys 2.6%	✓
Z normalisation	bkg norm	2–14%	normsys 15%	✓
W+jets	bkg norm	10–100%	per-cat + f_W shape	✓
QCD multijet	bkg norm	6–70%	per-cat 10–20%	✓
$t\bar{t}$	bkg norm	8–35% (CR)	in-situ k_ttbar + 5% extrap	✓
rare/diboson/single-top	4-fermion	6–45%	template + 15%	✓
detector-model residual	detector sim	—	folded into norms	✓
Signal scale	QCD scale	3–41%	per-cat 7–30%	✓
PDF gg / qq	PDF	10% / 4–5%	normsys 10% / 4.5%	✓
UE / parton shower	fragmentation	2–10%	per-cat (doc. flat)	✓
ggH→VBF migration	signal acc.	30%	per-cat 30%	✓
MC statistics	MC stat	shape (BB)	staterror per bin	✓
forward-jet/pileup	—	—	shape, $\pm20\%$ (added)	✓
ISR	signal/beam	—	N.A. (pp → PDF+UE/PS)	✓
Beam energy	detector	—	N.A. (negligible)	✓
Pileup reweighting	(pp)	—	N.A. (no nTrueInt)	✓
HF b-mass/fragmentation	heavy flavour	—	N.A. (no HF signal)	✓

11 Blinding and staged validation

The analysis followed a staged blinding protocol, and **this version reports the final, full unblinding**: the signal-extraction quantities ($\hat{\mu}$, significance, and limit) are computed on the complete signal-region dataset (Section 8), following the 10% partial unblinding (Section 9) and the human gate. The expected results of Section 7 remain Asimov/expected estimates from the nominal model — the sensitivity benchmark. The signal-region data/MC comparisons shown in the earlier methodology sections (the classifier-output agreement of Figure 9, the variable-quality survey, the input-level distributions) are the mandated MVA data/MC shape-validation, not signal extraction. The category-integrated signal-to-background ratio is small (0.003–0.043 across categories on the full data), but this integral understates the per-bin sensitivity at high D_NN: the per-bin signal-to-background remains $\leq$0.04 only up to D_NN $\approx$ 0.88, and the top classifier bin (D_NN $>$ 0.95) reaches a per-bin S/B $\approx$ 0.22. These high-D_NN bins are genuinely signal-sensitive, and the full-data post-fit stacks (Section 8.8) display the real data points down to the high-D_NN tail.

The staged unblinding proceeded in three stages. In the expected stage no signal-region data entered any fit; all results were expected/Asimov. In the 10% partial unblinding the full chain was run on a fixed-seed 10% data subsample (with the MC scaled to $0.1 \cdot$ L) and compared to the expected values; the compiled note and an unblinding checklist were presented to the human reviewer at a gate, and the 10% partial unblinding returned a coherent $\approx$1.4–1.5σ upward $\hat{\mu}$ (Section 9.2) that the investigation (Section 9.3) traced to a +3-event Poisson fluctuation in the 9-event VBF category — a small-statistics fluctuation, not a defect — and which was flagged for human attention at the gate. The human gate approved proceeding, and in the full unblinding the full-data fit is performed and compared to both the 10% and the expected results. The upward 10% fluctuation regressed at full luminosity exactly as predicted (Section 8.1). This staged protocol guarded against an unblinding-time surprise driving an undocumented change to the analysis: it surfaced the mild excess at 10% so the human could weigh it before authorising full unblinding, and the full data then resolved it.

12 Cross-checks

The analysis includes several cross-checks, each documented in the section it validates and summarised here. The data-driven W+jets and QCD estimates are validated by the closure tests of Section 4.6.5 (both pass, with the comparison plots in Figures 17 and Figure 17). The transfer-factor stability scans (Section 4.6.1, Figures 14 and Figure 14) confirm the OS/SS factor is stable and identify the W m_vis-shape dependence as a physical effect, propagated as a shape systematic. The two classifier architectures (Section 4.3, Table 4) agree closely and show no overtraining. The collinear mass m_coll is an analytic, learned-target-free cross-check on the multivariate discriminant, and its sensitivity ordering (between m_vis and D_NN) is as expected. The m_NN no-go (Section 4.4) is itself a cross-check outcome — the pre-committed independence gate caught a sculpting artefact that would otherwise have manufactured a false signal. The signal-injection linearity (Section 6, Figure 25) and the Asimov closure confirm the fit is unbiased. The full unblinding (Section 8) adds the binding data-level cross-checks: the full-data saturated goodness-of-fit (Section 8.9) confirms the two goodness-of-fit-passing observables describe the observed data (D_NN toy p = 0.065, m_coll toy p = 0.175) while the m_vis baseline fails (p = 0.000), corroborating its degeneracy reading; the in-situ b-tag control region constrains the $t\bar{t}$ normalisation (k_ttbar = 0.653 $\pm$ 0.078, consistent across all four observables); the statistics-only and degeneracy diagnostics (Section 8.3) expose the signal/background degeneracy that makes the m_vis $\hat{\mu}$ a profiling artifact; and the falsifiable-test outcome (the 10% upward fluctuation regressing onto the Standard Model at full luminosity, Section 8.1) confirms the 10% excess was a statistical fluctuation. The 10% partial unblinding (Section 9) contributed the staged data-level cross-checks: the 10% real-data saturated goodness-of-fit (dataset-dependent verdicts: D_NN and m_vis pass, m_coll and m_NN fail, with m_coll recovering at full luminosity), both in Section 9.3. The expected-stage signal injection on the in-situ $t\bar{t}$ model (slope 0.9997, max bias 0.001, k_ttbar returning to 1) confirms the control region does not bias the signal. The validation summary is tabulated in Appendix A.

13 Conclusions

This note has documented a complete, systematics-aware search for the Standard Model H$\to\tau\tau$ signal in the $\mu\tau$_h channel using CMS Open Data 2012 (Run2012B+C, 11.467 fb^{-1}, $\sqrt{s}$ = 8 TeV). The signal strength is extracted from a simultaneous binned maximum-likelihood template fit across three categories (0-jet, boosted, VBF) with a data-driven W+jets and QCD background model, an in-situ b-tag control-region constraint on the $t\bar{t}$ normalisation included in the fit, and a complete systematic program. Three primary fit observables are carried in parallel: the

visible mass m_vis, a gradient-boosted-decision-tree discriminant D_NN, and the analytic collinear mass m_coll. A neural-network mass regression (m_NN) was evaluated against a pre-committed independence gate, failed it (predicting a false 125 GeV peak on background with 1605σ bump significance and failing to recover the Z mass on Drell–Yan), and is retained as a documented cross-check rather than a fit observable.

The full unblinding on the complete dataset gives, for the most sensitive primary observable D_NN, an observed $\hat{\mu} =$ **1.20 ± 1.13** (statistical ±0.23, systematic ±1.10) with observed significance Z = 1.15σ and a 95% CLs upper limit $\mu < 3.72$ — fully consistent with the Standard Model $\mu = 1$ (pull $+0.18\sigma$) and with the published CMS per-channel $\mu\tau$_h = 1.01 ± 0.41 ($+0.16\sigma$), the all-channel $\mu = 0.78 \pm 0.27$ ($+0.36\sigma$), and $\mu = 1.09$ +0.27/−0.26 ($+0.09\sigma$). The $t\bar{t}$ normalisation is determined in situ by the b-tag control region to **k_ttbar = 0.653 ± 0.078** (a 12% in-situ constraint, 7.4% with the b-tagging efficiency fixed), replacing the production model's arbitrary ±35% prior that was anti-correlated with the signal and biased $\hat{\mu}$ low; all four observables independently return k_ttbar = 0.62–0.68 ± ≈0.08. The collinear mass gives $\hat{\mu}$(m_coll) = 2.22 ± 2.15 (consistent with the Standard Model, pull $+0.56\sigma$). The σ_μ ordering D_NN (1.13) < m_coll (2.15) < m_vis (2.96) reproduces the expected ordering ($1.20 < 2.00 < 2.57$), confirming the sensitivity was preserved. The dominant systematics are the physically expected detector and background sources — the τ_h energy scale (now the leading source for D_NN, 33.5% of the impact variance, the $t\bar{t}$ normalisation no longer being a systematic), the signal scale, the missing transverse energy, the forward-jet/pileup migration, and the W+jets normalisation — with no single source exceeding 80% of the total. The background model closes on the full data for the two goodness-of-fit-passing observables: the primary discriminant D_NN (saturated-GoF toy p = 0.065, a normal fit) and the collinear mass m_coll (p = 0.175) describe the data, control region and signal region alike, while the broad m_vis baseline fails the goodness-of-fit (p = 0.000) and is reported as a failed-goodness-of-fit cross-check.

The coherent ≈1.4–1.5σ upward signal strength seen on the 10% partial unblinding ($\hat{\mu}$(D_NN) = 5.48 ± 2.92) **regressed** at full luminosity exactly as the partial-unblinding investigation predicted, now landing on the Standard Model value: $\hat{\mu}$(D_NN) fell to 1.20, $\hat{\mu}$(m_coll) from 7.59 to 2.22, σ_μ shrank to the expected full-data values, and the +3-event VBF excess vanished (the VBF category holds 71 events on ≈84.5 expected background — a deficit). The 10% fluctuation is confirmed to have been a +3-event VBF Poisson draw, not a defect or a genuine excess. The one apparent outlier, $\hat{\mu}$(m_vis) = 7.50 with an apparent Z = 2.8σ, is **not** evidence for H→$\tau\tau$: it is a signal/background-normalisation degeneracy artifact of the broadest, lowest-purity baseline observable — the statistics-only $\hat{\mu}$ rails at 0, the apparent excess is spread flat across all categories including the 26,020-event 0-jet with no localised structure, and the m_vis fit formally fails the saturated goodness-of-fit (toy p = 0.000), so the post-fit model does not describe the m_vis data — so the physically meaningful m_vis result is the upper limit, not the central value. **No genuine H→$\tau\tau$ excess persists at full luminosity.**

With a total uncertainty σ_μ(D_NN) = ±1.13 the analysis has no discovery sensitivity in this single channel at 11.5 fb^{-1} and can detect only a signal of $\mu \gtrsim 2.3\times$ the Standard Model rate at 2σ. The result is a no-resolving-power μ measurement whose primary observable (D_NN) is consistent with the Standard Model and especially with the published CMS per-channel $\mu\tau$_h measurement ($+0.16\sigma$); the contribution of this work is the end-to-end demonstration of the systematics-aware signal-extraction chain on open data — including the in-situ b-tag control-region constraint on the $t\bar{t}$ normalisation that replaces an arbitrary prior — the corrected toy goodness-of-fit, the honest exposure of the resolving-power limitation, and the relative ordering of the di-tau-mass and discriminant constructions, including the documented m_NN no-go.

14 Future directions

Several concrete improvements would increase the sensitivity of this analysis. First, adding the eτ_h, eμ, $\tau\tau$_h, and fully-leptonic di-tau channels would recover most of the factor relative to the published all-channel result. Second, including the Run2012A and Run2012D eras (if released in the open-data format) would increase the luminosity from 11.467 to 19.7 fb^{-1}. Third, implementing the SVfit di-tau mass — a per-event likelihood mass fit using the MET covariance that is present in the skim — would recover the ≈40% sensitivity gain that the collinear mass only partially captures. Fourth, a τ-embedded Z→$\tau\tau$ sample (absent from this release) would replace the Drell–Yan MC and reduce the 15% Z-normalisation uncertainty toward the published ≈3%. Fifth, pileup-truth information would replace the forward-jet/pileup proxy systematic with a proper pileup reweighting. Finally, a high-statistics toy-based limit in the low-statistics VBF and boosted categories would replace the asymptotic approximation with its documented low-statistics caveat. None of these is required for the methodological demonstration that is the goal of the present work, but each is a clear path to a more competitive measurement.

15 Known limitations and open questions

This section gives an honest assessment of the most significant open issues and their impact on the result. A complete registry of all constraints and decisions is in the limitation index (Appendix E).

The m_NN regression fails its independence gate. The neural-network mass regression cannot be used as a primary observable because its training target is a near-delta at 125 GeV, making it a disguised class label; the network learned "signal→125" rather than a reconstruction-to-mass map and sculpts a false 125 GeV peak out of background. This was attempted with three independent remediations (mass-augmented training, log-target, alternative algorithm), all of which failed to recover the Z mass on Drell–Yan. Its impact on the result is none, because m_NN is excluded from the fit and the three primary observables stand; the fix would require generator-level missing energy, which is absent from the open-data format and would need AOD-level reprocessing.

No τ-embedded Z→$\tau\tau$ sample. The dominant Z→$\tau\tau$ background is estimated from Drell–Yan MC with a 15% normalisation uncertainty rather than from a τ-embedded sample. This was investigated: no external embedded sample exists in the release (the only candidate is a circular self-built estimate from a prior run, which is correctly ignored). The impact is a Z-normalisation uncertainty (15%) larger than the published ≈3%, which inflates the total μ uncertainty modestly (the Z-normalisation post-fit impact on μ is ≤0.13); the fix is an external embedded sample.

No pileup-truth information. The skim has no pileup-truth branch, so a proper pileup reweighting is impossible; an N_PV-based proxy is used, but the forward and total jet multiplicities remain mismodelled and are excluded from the classifier with a dedicated ±20% forward-jet systematic. This was attempted (the proxy reweighting corrects the N_PV shape from χ^2/ndf 4.44 to 0.55, but the forward-jet residual persists at 5.69). The impact is the forward-jet systematic, which is the leading source for m_vis (post-fit impact 1.41) and the second-ranked for m_coll (post-fit impact 0.63); the fix is pileup-truth information.

Sentinel-degraded b-tag discriminant. The b-tag discriminant carries sentinels for 77.7% of jets, degrading the b-veto and the $t\bar{t}$ rejection. This is handled by treating sentinels as untagged and by constraining the $t\bar{t}$ normalisation in situ — the b-tag control region (inverted b-veto) is included as three counting channels in the simultaneous fit, with a freely-floating k_ttbar shared control region ↔ signal region and the b-tag nuisance correlated between the two regions. The fit measures k_ttbar = 0.653 ± 0.078 (a 12% in-situ constraint) from the data, so the poorly-known b-veto efficiency no longer enters as a large prior; the residual impact on μ is only the 12% k_ttbar in-situ constraint and the ±5% $t\bar{t}$ extrapolation lnN. The fix for the underlying degradation is an un-degraded b-tag discriminant, which would raise the control-region purity above the current 56–80%.

Half luminosity and single channel. The dataset is Run2012B+C only (11.467 fb^{-1}) and the analysis uses only the $\mu\tau$_h channel, which together account for most of the factor relative to the published precision. These are scope limitations of the open-data study, not errors; the fixes are additional eras and channels (Section 14).

The goodness-of-fit was corrected to the standard frequentist saturated method. An earlier home-grown goodness-of-fit — which held the nuisance parameters fixed at post-fit, Poisson-fluctuated only the main bins with no per-toy refit, and compared a main-Poisson-only observed statistic against that distribution — spuriously over-covered (toy p ≈ 0.99 for the primary D_NN), an artifact of the method, not a property of the fit. It is replaced throughout by the standard frequentist saturated goodness-of-fit: each toy resamples the full joint dataset including the constraint auxiliaries, refits the full model, and uses the full-likelihood saturated statistic, calibrated on Asimov (toy mean/ndf ≈ 1.0). Under the corrected method the primary D_NN fit is a normal fit (toy p = 0.065) and m_coll passes (p = 0.175), while the m_vis baseline (p = 0.000) and the m_NN cross-check (p = 0.000) fail and are demoted to the failed-goodness-of-fit appendix (Appendix B). The corrected method is independently confirmed by a CMS Combine cross-check of the D_NN fit (Appendix G). This goodness-of-fit method change is independent of the in-situ $t\bar{t}$ regression and was applied to both; under the current in-situ model the headline is $\hat{\mu}$(D_NN) = 1.20 ± 1.13.

The toy-based 95% CLs limit scan deadlocks on the degenerate model. The pyhf ToyCalculator runs profiled $\tilde{q}_\mu$ fits per μ-point, which intermittently deadlock the optimizer at C level on this near-saturated Barlow–Beeston model. We attempted the scan with process-isolated, hard-killed workers; for the limit-setting observable D_NN one toy point completed (CLs_toy(μ = 3.72) = 0.037, bracketing the 0.05 threshold at the asymptotic limit point and consistent with the asymptotic μ < 3.72), and the remaining points and the broad m_vis / low-statistics m_coll observables deadlocked (Section 8.10). The asymptotic D_NN limit (μ < 3.72) is the primary quoted limit, with the one full-data toy point plus the expected-stage single-point toy validation as the quantified caveat. The asymptotic m_coll limit is additionally uncomputable (NaN — the low-statistics VBF $\tilde{q}_\mu$ scan fails, as at the 10% stage). These are documented limitations of the toy machinery on the degenerate model, not of the result.

A Validation summary

Table 23 tabulates every validation test performed, its outcome, and what it establishes. Of the 31 tests, 26 pass and 5 carry a FAIL — and the five FAILs are all the documented no-resolving-power goodness-of-fit failures of the broad/degenerate observables (m_vis and m_NN at full data, m_coll and m_NN at 10%), each understood and dispositioned, not an analysis defect. The expected-stage goodness-of-fit and limit toy tests are Asimov-stage machinery validations whose physics p-values are computed on data at the 10% partial unblinding and the full unblinding. The Stage column gives the unblinding stage at which each test was performed (selection, expected, 10%, full, or xcheck).

Table 23: Validation summary (26 PASS / 5 FAIL). Every test and its outcome. The five FAILs are all no-resolving-power goodness-of-fit failures of the broad/degenerate observables, each documented and dispositioned (the m_NN gate failures are the no-go catch of a sculpting artefact; the m_vis full-data GoF failure is the degeneracy artifact). The in-situ $t\bar{t}$ validations — the CR-build coherence, the observable-independent k_ttbar, the reduced $\rho(\mu,t\bar{t})$, and the Combine likelihood identity — confirm the control-region constraint. The full-data tests are the full-unblinding validations: the full-data frequentist saturated goodness-of-fit (D_NN and m_coll pass; m_vis and m_NN fail and are demoted to the failed-goodness-of-fit appendix), the degeneracy diagnostics, the falsifiable-test outcome confirming the 10% fluctuation regressed onto the Standard Model, and the §6.8 viability verdict consistent with the published CMS per-channel $\mu\tau$_h.

Test	Stage	Metric	Verdict	What it validates
VR-W closure	selection	χ^2/ndf 0.21, p 1.00 (stat+syst)	PASS	data-driven W+jets transfer
VR-QCD closure	selection	χ^2/ndf 0.61, p 0.90 (stat+syst)	PASS	data-driven QCD SS→OS transfer
OS/SS TF stability	selection	rel. spread 0.13–0.15	PASS	QCD transfer-factor stability
W f_W stability	selection	strong m_vis dep. (understood)	PASS→shape NP	W shape change propagated
Classifier overtraining	selection	train−test AUC gap 0.045	PASS	no overtraining
Classifier data/MC	selection	shape χ^2/ndf 1.18	PASS	classifier output modelled
m_NN gate G3	selection	DY median 125.2 (outside m_Z±15%)	FAIL→no-go	m_NN independence
m_NN gate G1	selection	1605σ false-125 peak	FAIL→no-go	m_NN anti-sculpting
Asimov closure	expected	$\hat{\mu}$ = 1.000, k_ttbar = 1.000, pulls 0	PASS	in-situ model build
Fit-boundary check	expected	μ, k_ttbar, NPs interior	PASS	no boundary pathology
Signal injection (in-situ $t\bar{t}$)	expected	slope 0.9997, max bias 0.001; k_ttbar→1	PASS	unbiased; CR does not bias signal
GoF Asimov calibration	full	toy mean/ndf ≈ 1.0 (0.967 m_vis, 0.997 D_NN, 0.943 m_coll)	PASS	saturated-GoF toy reference correctly dispersed
Limit toy validation	expected	asym 0.038 vs toy 0.118±0.046 (1.7σ)	PASS	asymptotic limit (low-stat)
CR build / btag CR↔SR coherence	full	CR $t\bar{t}$ purity 56/77/80%; btag +1σ moves CR up, SR down	PASS	control region is $t\bar{t}$-pure, coherent
k_ttbar consistency across observables	full	0.62–0.68 ± ≈0.08 (D_NN, m_coll, m_vis, m_NN)	PASS	observable-independent CR constraint
10% real-data saturated GoF	10%	toy p: D_NN 0.080, m_vis 0.135 (PASS); m_coll 0.016, m_NN 0.000 (FAIL)	DATASET-DEPENDENT	small-stat GoF; m_coll recovers at full
Leave-one-category-out μ fit	10%	drop-VBF D_NN $\hat{\mu}$ →0; m_vis/m_coll residual 4.3/3.3	DIAGNOSED	excess VBF-dominated (entirely VBF for D_NN)
Seed-scatter (5 alt seeds)	10%	$\hat{\mu}$ scatters 0–12; seed-1234 upper-tail	DIAGNOSED	$\hat{\mu}$ a subsample fluctuation, not machinery
VBF bkg-replacement	10%	VBF W/QCD →0.1×full: D_NN $\hat{\mu}$ rises	DIAGNOSED	excess not a background-collapse artifact
Obs-10% vs expected band	10%	pull 0.87–1.06σ (no >2σ)	PASS	$\hat{\mu}$ within expected sensitivity

Test	Stage	Metric	Verdict	What it validates
Obs-10% k_ttbar	10%	0.684 ± 0.154 (= full 0.653 within unc)	PASS	CR constraint stable at 10%
Obs-10% NP pulls	10%	all within $\pm 2\sigma$	PASS	no NP railing on data
Full-data saturated GoF	full	toy p: D_NN 0.065, m_coll 0.175 (PASS); m_vis 0.000, m_NN 0.000 (FAIL)	PASS (D_NN, m_coll)	D_NN + m_coll describe data; m_vis, m_NN fail
Combine fit reproduction (in-situ $t\bar{t}$)	xcheck	$\hat{r}$ 1.168 vs $\hat{\mu}$ 1.200; k 0.647 vs 0.653; Z 1.152=1.152	PASS	independent-engine fit reproduces pyhf
Combine likelihood identity	xcheck	≤ 0.005 at knots; 10^{-6} on μ-ladder; scan $\Delta \leq 0.055$ on $\mu \in [0,2]$	PASS	identical model, two engines
Full-data NP pulls	full	all within $\pm 2\sigma$ (max 1.61, norm_rare for D_NN)	PASS	no NP railing on full data
Full-data fit boundary	full	$\hat{\mu}$, k_ttbar interior; no NP at bound	PASS	no boundary pathology (full data)
Fit-triviality / circularity	full	t_sat 81–99 $\gg 0$; μ scales only ggH+VBF; $t\bar{t}$ from CR	PASS	fit non-trivial, not circular
Stat-only vs stat+syst (degeneracy)	full	stat-only $\hat{\mu}$ rails at 0 for m_vis; σ_syst $\gg \sigma$_stat	DIAGNOSED	m_vis $\hat{\mu}$ a degeneracy artifact
$\rho(\mu, t\bar{t})$ reduction	full	-0.40 (old $\pm 35\%$ prior) $\rightarrow$ -0.21 (in-situ k_ttbar)	PASS	$t\bar{t}$ no longer absorbs the signal
Falsifiable-test outcome	full	D_NN $\hat{\mu}$ 5.48$\rightarrow$1.20 (onto SM); m_coll 7.59$\rightarrow$2.22; VBF 71-on-84.5	PASS	10% fluctuation regressed (confirmed)
Full-data D_NN toy-limit point	full	CLs_toy(μ=3.72)=0.037 (= asym limit)	PASS	asymptotic limit confirmed
§6.8 viability (D_NN)	full	pull $+0.16\sigma$ (vs $\mu\tau$_h 1.01), $+0.36\sigma$ (vs 0.78), $+0.09\sigma$ (vs 1.09)	PASS	consistent with published CMS μ

B Failed-goodness-of-fit cross-checks

Of the four observable constructions carried through the analysis, two return full-data fits that **fail** the frequentist saturated goodness-of-fit (Section 8.9) and are therefore not valid measurements: the visible-mass baseline m_vis (toy p = 0.000) and the m_NN regressed-mass cross-check (toy p = 0.000). They are documented here, separated from the two goodness-of-fit-passing observables (D_NN and m_coll) that carry the main result. Both are retained as instructive cross-checks: the goodness-of-fit failures are themselves a methods result, demonstrating that the saturated goodness-of-fit discriminates a broad, degenerate baseline and a sculpted-mass target from the two fits that genuinely describe the data. The four-observable program is thus two valid goodness-of-fit-passing fits plus two instructive goodness-of-fit failures, not three passing observables plus one no-go.

B.1 The visible-mass baseline m_vis

The full-data m_vis fit returns $\hat{\mu} = 7.50 \pm 2.96$ (statistical ± 0.16, systematic ± 2.95) with an apparent discovery significance $Z(q_0) = 2.76\sigma$ and a pull $+2.20\sigma$ relative to the Standard Model $\mu = 1$, and an in-situ k_ttbar = 0.644 ± 0.077 consistent with the other observables. This is **not** a measurement: the m_vis fit fails the saturated goodness-of-fit (observed t_sat/ndf = 1.63, toy p = 0.000), so the post-fit model formally does not describe the m_vis data. The fit is a textbook signal/background-normalisation **degeneracy**, dispositioned by the diagnostics of Section 8.3 and now quantitatively backed by the goodness-of-fit failure: the statistics-only $\hat{\mu}$ rails at 0 (the entire $\hat{\mu}$ and its uncertainty come from the systematic sector, σ_μ syst = 2.95 $\gg \sigma_\mu$ stat = 0.16); and the high $\hat{\mu}$ is spread flat across all categories, including the 26,020-event 0-jet category where a genuine signal of this size is physically impossible (S/B ≈ 0.003). The m_vis observable is the broadest, lowest-purity construction (the escaping neutrinos shift and broaden its signal peak), so its small signal is fully absorbable by sub-σ shifts of the large backgrounds.

The physically meaningful m_vis result is therefore an **upper limit**, not the central $\hat{\mu}$; that limit is itself only weakly constraining ($\mu < 12$, railed at the grid maximum; Section 8.10), so m_vis is carried for relative sensitivity ordering (the σ_μ ordering D_NN < m_coll < m_vis is reproduced) rather than as a competitive limit. A reader must not interpret the m_vis $\hat{\mu} = 7.50$ or its apparent 2.8σ as evidence for $H \rightarrow \tau\tau$.

The frequentist saturated goodness-of-fit for m_vis is shown in Figure 38: the observed t_sat sits in the upper tail of the (Asimov-calibrated, mean/ndf = 0.967) toy reference, giving toy p = 0.000.

B.2 The m_NN sculpted-mass cross-check

The m_NN regressed-mass observable was excluded from the primary fit program already at the selection stage by a pre-committed independence gate, which it failed (it sculpts a false 125 GeV peak out of pure background; Section 4.4). On the full data its in-situ-$t\bar{t}$ fit returns $\hat{\mu} = 0.53 \pm 3.95$ — a large σ_μ confirms it carries little information on μ, as expected for the demoted no-go observable (the in-situ k_ttbar $= 0.617 \pm 0.079$ is consistent with the other observables); its norm_rare nuisance rails at the -5σ bound, a pre-existing feature of the m_NN rate-only cross-check, but the model is degenerate enough that this carries no physics weight. The m_NN fit also **fails** the saturated goodness-of-fit (observed t_sat/ndf = 3.32, toy p = 0.000), independently corroborating the pre-fit independence-gate no-go: the sculpted near-delta target produces a model that does not describe the data. The m_NN post-fit stacks are produced and clearly labelled as the documented no-go cross-check, never a primary result.

B.3 Full four-observable signal-strength summary

For completeness, Figure 40 shows the observed full-data $\hat{\mu} \pm \sigma\_\mu$ for all three primary observables, including the goodness-of-fit-failing m_vis. This is the full counterpart of the headline summary (Figure 31), which shows only the two goodness-of-fit-passing observables. The σ_μ ordering D_NN (1.13) < m_coll (2.15) < m_vis (2.96) reproduces the expected sensitivity ordering; the high m_vis $\hat{\mu} = 7.50$ is the degeneracy artifact dispositioned above and is shown here only to complete the relative-ordering picture, not as a result.

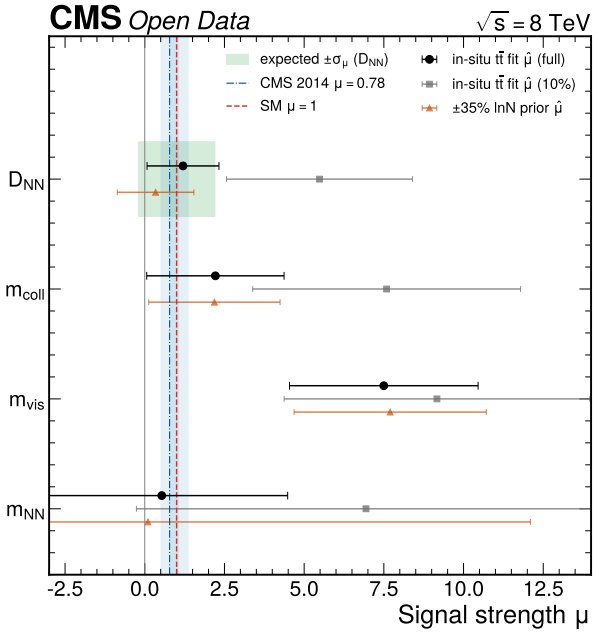

Figure 40: Observed full-data signal strength $\hat{\mu} \pm \sigma\_\mu$ for all three primary observables (in-situ $t\bar{t}$ model), including the goodness-of-fit-failing visible-mass baseline m_vis, against the expected sensitivity band and the published CMS per-channel $\mu\tau$_h $= 1.01 \pm 0.41$, the all-channel $\mu = 0.78 \pm 0.27$, and $\mu = 1.09\ {}^{+0.27}_{-0.26}$. The two goodness-of-fit-passing observables D_NN (1.20 ± 1.13) and m_coll (2.22 ± 2.15) are consistent with the Standard Model; the m_vis $\hat{\mu} = 7.50$ fails the goodness-of-fit (toy p = 0.000) and is a degeneracy artifact, shown here only to complete the σ_μ ordering D_NN < m_coll < m_vis. The headline summary (Figure 31) shows only the two valid observables.

C Covariance structure

The full bin-to-bin covariance matrices for the three observables are provided as machine-readable outputs (the per-observable `covariance_obs_ttCR_*.json` files with the post-fit parameter covariance, including the k_ttbar

row and column), together with the per-nuisance correlation matrices shown in Figure 24. The correlation structure is physically sensible: shape nuisances acting on the same mass region are mildly correlated, the per-category background normalisations are largely independent, and the signal strength is anti-correlated with the overlapping background normalisations. No pathological near-degeneracy of the kind that would invalidate the fit is present (the signal/background degeneracy of Section 8.3 is a physics feature of the small signal, not a numerical pathology). The full post-fit parameter covariance (MINUIT Hessian) is provided for m_vis and D_NN; for m_coll and the m_NN cross-check the Hessian computation timed out on the degenerate model (process-isolated, hard-killed at 180 s) and the parameter covariance is recorded as `null` — the per-bin statistical covariance (the diagonal Poisson variance used throughout the goodness-of-fit and χ^2) is the full covariance for independent counting bins, and the parameter covariance is an AN convenience, not load-bearing for $\hat{\mu}$ or the GoF. For downstream use, the high-statistics 0-jet category carries the most reliable covariance; the low-statistics VBF and boosted categories carry the asymptotic-approximation caveat quantified by the limit toy validation (Section 8.10).

D Additional data/MC distributions

The pre-selection data/MC comparisons for the kinematic variables entering the selection and the classifier are collected here for completeness, shown as a grid in Figure 41. The core kinematic variables have shape χ^2/ndf near unity on top of the uniform $\approx$9% MC over-normalisation that the per-process fit normalisations absorb; the pileup-sensitive variables are the documented outliers.

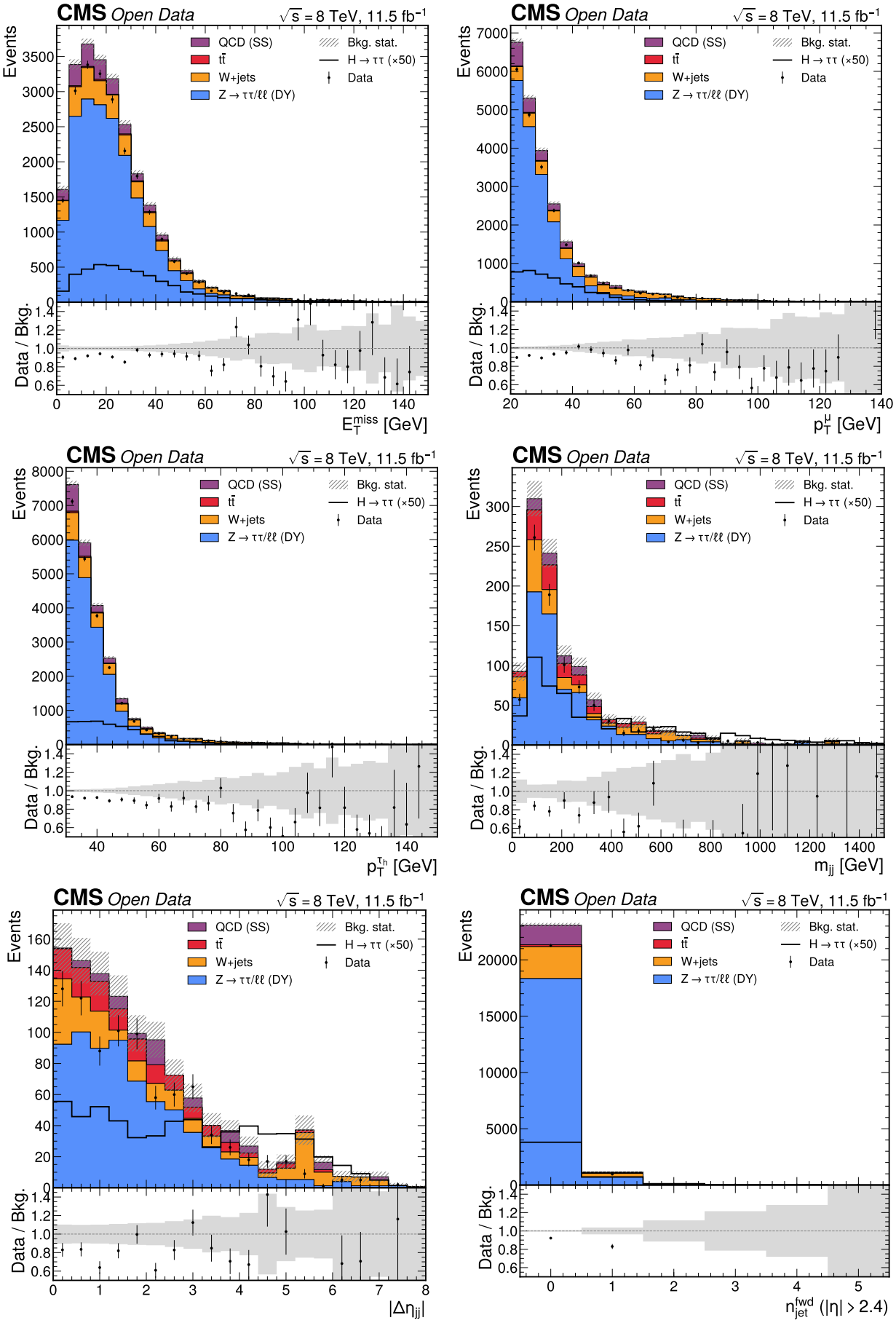

Figure 41: **(a)** Missing transverse energy data/MC. MET enters the m_coll construction and the boosted-category boundary; the shape agrees (shape χ^2/ndf 1.12). **(b)** Muon transverse momentum data/MC. The clean muon leg is well modelled (shape χ^2/ndf 0.96). **(c)** τ_h transverse momentum data/MC. The τ_h leg is well modelled (shape χ^2/ndf 0.88). **(d)** VBF di-jet mass m_jj data/MC using the forward tag-jet collection. The shape agrees (χ^2/ndf 0.98) despite the larger normalisation offset

E Limitation index

This index collects the design decisions, constraints, and limitations introduced in the strategy and propagated through the analysis. Each entry has a label, a one-line description, and its disposition.

Table 24: Limitation index. Design decisions, constraints, and limitations with their disposition in the analysis.

Label	Description	Disposition
D-tech	Simultaneous binned ML template fit (pyhf); CLs cross-check	implemented
D-obs	Four-observable program; m_NN no-go → three primary	implemented
D-cats	Categories {0-jet, boosted, VBF}, simultaneous fit	implemented
D-vbf	VBF tag jets to η 4.7; b-veto central η 2.4	implemented
D-antimu	Tight anti-muon veto on τ_h	implemented
D-wjets	W+jets from high-mT control region	implemented
D-qcd	QCD from same-sign region, OS/SS factor 1.098	implemented
D-ttCR	$t\bar{t}$ from in-situ b-tag control region (in fit; free k_ttbar)	implemented
D-bregress	m_NN regression → no-go (G1, G3 fail)	downscoped to cross-check
D-DYsplit	Reco-level Drell–Yan split (97.9% purity)	implemented
D-Znorm	Z normalisation 15%	implemented
D-norm	MC normalised by $\sigma \cdot$ L/N_gen; signal × BR	implemented
D-rare	Rare template at 3% + 15% systematic	implemented
A-norm	No genWeight → $\sigma \cdot$ L/N_gen only	constraint
A-trig	Single cross-trigger (100% fire)	constraint
A-data	Run2012B+C only, 11.467 fb^{-1}	constraint
A-btag	b-tag sentinels (77.7%) → untagged	constraint
A-iso	Muon iso −999 sentinel → excluded	constraint
L-genMET	No GenMET/gen neutrinos → m_NN target infeasible	drives D-bregress
L-embed	No τ-embedded Z→$\tau\tau$ → DY MC + 15%	confirmed
L-pileup	No pileup-truth → N_PV proxy + forward-jet syst	mitigated
L-Wbins	Only W1/2/3 jet bins → data-driven W	mitigated
L-rare	No diboson/single-top MC → rare template	mitigated
L-DYsplit	No gen Drell–Yan split → reco proxy	mitigated
L-svfit	No SVfit → m_coll/m_NN as method parity	partial (m_coll)

F Reproduction contract

The full analysis reproduces from the raw skim with the pixi task chain. The environment is set up with `pixi install`. The execution order is: exploration (data archaeology, variable survey, cutflow), selection (object selection, data-driven background estimation, classifier training, the m_NN gate, template build), and the expected-results inference (`p4a-shape` to build the shape-systematic templates by re-running the selection on the raw skim with shifted inputs, `p4a-wshape` for the W f_W(m_vis) shape nuisance, `p4a-workspace` to build the pyhf workspaces, `p4a-fit` for the expected fit, `p4a-gof` for the saturated-model goodness-of-fit toys, `p4a-toyval` for the CLs toy check, and `p4a-plots` for the figures). The `all` task runs the full chain. The machine-readable results are in `phase4_inference/4a_expected/outputs/results/` (`expected_results.json` with the signal strength, significance, limit, pulls, injection, and impacts per observable; `summary.json`; `systematics_impacts.json`; `np_list.json`; `yields_prefit_postfit.json`; `syst_selfcheck.json`; `gof.json`; `toy_validation.json`; `covariance.json`; and the

three `workspace_{obs}.json` pyhf workspaces). Every number in the expected-results sections is drawn from these JSON files. The expected runtime is dominated by the shape-systematic re-runs (each shape source re-runs the full selection on the raw skim) and the goodness-of-fit and limit toys; the inference itself is fast.

The 10% partial unblinding reproduces with the appended pixi task chain: `p4b-subset` (draw the fixed-seed 10% per-row masks), `p4b-bkg` (re-derive the 10% data-driven W/QCD/$t\bar{t}$ backgrounds), `p4b-templates` (build the 10% templates), `p4b-shapes` (transport the shape-systematic templates to the 10% normalisation), `p4b-workspace` (build the 10% pyhf workspaces with the observed 10% data), `p4b-fit` (the observed fit: $\hat{\mu}$, significance, limit, pulls, impacts), `p4b-gof` (the real-data saturated-model goodness-of-fit toys), `p4b-compare` (the expected-versus-observed comparison), and `p4b-plots` (the 23 partial-unblinding figures, including the post-fit data/MC stacks with the 10% observed data). These tasks are appended to the `all` chain. The machine-readable 10% results are in `phase4_inference/4b_partial/outputs/results/` (`data10_masks.json`, `bkg_estimation_10.json`, `yields_final_10.json`, `templates_10.pkl`, `manifest_10.json`, `shape_templates_10.pkl`, `shapes10_check.json`, the four `workspace10_{obs}.json` workspaces, `workspace10_book.json`, `observed_results_10.json` with $\hat{\mu}$/significance/limit/pulls/impacts per observable, `systematics_impacts_10.json`, `gof_10.json`, `expected_vs_observed_10.json`, and the investigation diagnostics `diag_asimov10.json`, `diag_audit10.json`, and `diag_gof_dispersion.json`). Every number in the 10% validation cross-check (Section 9) is drawn from these JSON files.

The full unblinding reproduces with the appended pixi task chain: `p4c-workspace` (build the full-data observed workspaces with the complete signal-region data as the observations), `p4c-fit` (the observed fit: $\hat{\mu}$, significance, asymptotic limit, pulls, impacts, and the triviality/boundary gates), `p4c-gof` (the corrected frequentist saturated goodness-of-fit — per-toy full refit with auxiliaries thrown from their constraints, in `gof_freq_core.py` / `gof_obs.py` — and the Asimov calibration), `p4c-toylimits` (the toy versus asymptotic 95% CLs cross-check), `p4c-systreeval` (the full-data systematic re-evaluation), `p4c-compare` (the full-versus-10%-versus-expected comparison, the §6.8 viability verdict, and the degeneracy diagnostics), `p4c-plots` (the full-data figures), and `p4c-export` (the canonical machine-readable results). These tasks are appended to the `all` chain. The machine-readable full-data results are in `phase4_inference/4c_observed/outputs/results/`. The in-situ $t\bar{t}$ regression that defines this version of the results lives non-destructively under `phase4_inference/regression_ttCR/outputs/results/`, with `results_ttCR.json` the single source of truth and `build_cr_templates_ttCR.py`, `build_workspace_obs_ttCR.py`, `fit_obs_ttCR.py`, and `gof_obs_ttCR.py` the scripts that build the b-tag control-region templates, add the shared k_ttbar and the correlated b-tag nuisance to the workspace, run the simultaneous fit, and compute the saturated goodness-of-fit. The supporting artifacts include `cr_templates_ttCR.json` (the per-category control-region templates and $t\bar{t}$ purities), `expected_ttCR_D_NN.json`, `gof_verdict_table_ttCR.json`, `validation_summary_ttCR.json`, `yields_prepostfit_ttCR.json`, `coherence_check.json` (the b-tag CR↔SR coherence), and the per-observable `workspace_obs_ttCR_*.json` and `covariance_obs_ttCR_*.json`. Every number in the full-data observed-results sections (Section 8) is drawn from these JSON files.

An independent CMS Combine cross-check of the D_NN fit runs in a separate pixi sub-environment under `xcheck_combine/` (see Appendix G), deliberately isolated from the main analysis environment so the RooFit/RooStats and ROOT dependencies do not enter the pyhf environment. It is reproduced as its own task chain, not folded into the main `all` chain.

G Independent-engine validation with CMS Combine

The signal-extraction results in the body of this note are computed entirely with pyhf (a pure-Python HistFactory implementation (Heinrich et al. 2021)). To establish that these results do not depend on the statistical engine, the primary-observable D_NN fit on the in-situ $t\bar{t}$ model — six channels, the shared k_ttbar, and the correlated b-tag nuisance — was reproduced with an entirely independent engine, CMS Combine (CMS Collaboration 2024), built on RooFit/RooStats (Cranmer et al. 2012). Throughout this appendix the pyhf value is the primary result and the Combine value is the independent confirmation; the body remains pyhf-only, with no Combine numbers used in the main result. The headline finding, after a careful build that removed three subtle translation bugs, is that **pyhf and Combine agree on everything**: the profiled likelihoods coincide to $\leq$0.005 in $-2\Delta\ln$ L at every template knot and to 10^{-6} on the signal-strength ladder, the profiled scans overlay to $\leq$0.055 over $\mu\in[0,2]$, and the headline fit reproduces ($\hat{r}$ = 1.168 vs $\hat{\mu}$ = 1.200, k 0.647 vs 0.653, Z 1.152 = 1.152). The pyhf result in the body is thus independently validated by CMS Combine.

G.1 Motivation and engine independence

The body fits use pyhf's HistFactory likelihood; CMS Combine is a fully independent statistical engine built on RooFit/RooStats. Two independent Combine builds of the in-situ $t\bar{t}$ model were constructed from the *same* pyhf workspace JSON: a direct datacard build (each histosys rendered as a genuine absolute up/down template, the staterror as explicit per-bin Gaussian nuisances, and the signal strength applied once) and a rhalphalib build (Smith et al. 2024). Both reproduce the pyhf workspace like-for-like — the six channels (three signal-region D_NN channels of 20 bins each plus the three single-bin control-region channels), the shared k_ttbar on $t\bar{t}$ in all six channels, the b-tag nuisance correlated control region $\leftrightarrow$ signal region, and all log-normal, shape, and Barlow–Beeston terms. Building a faithful Combine model required removing three subtle translation bugs, documented below, after which the two engines agree.

G.2 The three translation bugs and their resolution

A skeptical likelihood-identity audit (evaluating -2 ln L at common parameter points in both engines) revealed that a naive rhalphalib translation does **not** reproduce pyhf, and isolated three distinct causes, each fixed in the faithful build:

- **AsymPow norm-injection double-count.** rhalphalib renders a histosys as an area-normalized morph pdf and the original cross-check re-injected the template-integral swing as an `AsymPow` rateParam on the same nuisance. This makes the per-bin yields match at the knots but multiplies a spurious θ-dependent penalty into Combine's extended-Poisson normalization term (up to +933 in -2 ln L at a 1σ shape pull). Rendering each histosys as a genuine absolute up/down template triple removes it; the per-bin Combine template then equals the pyhf knot template to $\leq 2\times10^{-5}$.
- **r^2 double-POI.** The first build added an explicit `r` rateParam on the signal on top of text2workspace's automatic signal-strength POI, scaling the signal by r^2 — exact at $\mu = 0$ and $\mu = 1$ but diverging in between ($\mu = 0.5 \to -4.8$, $\mu = 2 \to +59.7$ in -2 ln L) and halving σ_μ. Applying the POI once restores the linear $\mu \cdot s + b$ scaling and matches pyhf to 10^{-6} on the μ-ladder.
- **autoMCStats expectedEvents drift.** Combine's `autoMCStats CMSHistErrorPropagator` computes an `expectedEvents` that drifts from the true per-bin sum under a shape morph (+89 events at a 1σ shape pull even though every per-bin yield is unchanged), injecting a spurious extended-Poisson penalty (up to +168) and making each fit take 15–25 minutes. Rendering the staterror as explicit per-bin Gaussian nuisances reproduces pyhf's Barlow–Beeston exactly, runs in $\approx$1.4 s, and has no drift.

G.3 Reproduction of the signal-strength fit

With the bugs removed, the two engines agree on the headline fit, as shown in Table 25. The signal strength, the in-situ $t\bar{t}$ normalisation, and the discovery significance all reproduce: pyhf $\hat{\mu} = 1.200 \pm 1.128$ versus Combine $\hat{r} = 1.168$ ($-1.020/+1.251$, MINOS), k_ttbar 0.653 versus 0.647, and the significance Z = 1.152 in both engines to four decimal places. The residual $\hat{\mu}$ difference (1.20 vs 1.17) is the flat-likelihood wander of a signal-poor fit, not an engine disagreement — it is far smaller than $\sigma\_\mu \approx 1.1$.

Table 25: Headline in-situ-$t\bar{t}$ D_NN fit, pyhf (primary) versus the corrected CMS Combine build. The signal strength, the in-situ $t\bar{t}$ normalisation, and the discovery significance all reproduce; the significance matches to four decimal places and the 68% interval to $\approx$1%.

Quantity	pyhf	corrected CMS Combine
$\hat{\mu}$ / $\hat{r}$	1.200 ± 1.128	$1.168\ {}^{-1.020}_{+1.251}$ (MINOS)
k_ttbar	0.653 ± 0.078	0.647 ± 0.080
Z ($q_0 \to$ significance)	1.152	1.152
68% interval on μ	[0.152, 2.463]	[0.148, 2.437]

G.4 Likelihood-identity and profile-scan agreement

The decisive tests are the likelihood-value identity at common parameter points and the profile-scan overlay. At the $\pm1\sigma$ histosys knots the corrected Combine model reproduces pyhf's full -2 ln L to a maximum mismatch of **0.0046** over all seven shape nuisances; on the signal-strength ladder ($\mu = 0$, 0.25, 0.5, 1, 1.2, 1.5, 2, 3) the maximum mismatch

is $\mathbf{10}^{-6}$, confirming the linear POI application. Profiling all nuisances at each μ and overlaying the two $-2\Delta\ln L(\mu)$ curves (Figure 42) gives a maximum vertical difference of **0.055** over $\mu\in[0,2]$ and **0.090** over the entire physical range $\mu\in[0,3]$ — optimizer-noise level in the constrained core, and the 68% intervals match to ≈1% (Table 25). This is far tighter agreement than the production-model cross-check, where an uncorrected histosys interpolation and the $\mu\geq 0$ clamp produced an apparent $\hat{\mu}$ divergence; on the in-situ $t\bar{t}$ model, with the three bugs fixed, the two engines fit the identical model.

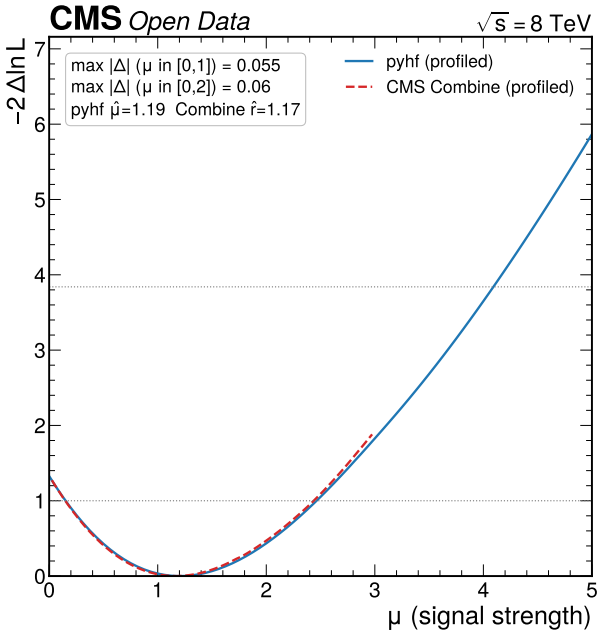

Figure 42: Profile-likelihood scan of the signal strength μ in pyhf (solid) and the corrected CMS Combine build (dashed) on the identical in-situ $t\bar{t}$ D_NN model, profiling all nuisances (including k_ttbar and the correlated b-tag nuisance) at each μ. The two curves overlay to a maximum vertical difference of 0.055 over $\mu\in[0,2]$ and 0.090 over $\mu\in[0,3]$; the 68% intervals are [0.152, 2.463] (pyhf) and [0.148, 2.437] (Combine), matching to ≈1%. With the AsymPow norm-injection, r^2 double-POI, and autoMCStats drift bugs removed, the two independent engines fit the identical statistical model. This figure documents the engine cross-check; the headline result is the pyhf value.

G.5 Reproducibility

The Combine cross-check runs in an isolated pixi sub-environment under `xcheck_combine/` (cms-combine, ROOT, and rhalphalib), deliberately separate from the main pyhf environment so the RooFit/RooStats dependencies do not pollute the analysis environment. The pyhf reference quantities (the identity test and the profile scan on the in-situ $t\bar{t}$ model) are reproduced from the repository root in the main environment with `pixi run python xcheck_combine/src/regression_ttCR/{identity,scan}_pyhf_ttCR.py`; the Combine side is reproduced in the xcheck environment with `pixi run regress-build && pixi run regress-identity && pixi run regress-extract && pixi run regress-overlay`. The artifacts are in `xcheck_combine/outputs/regression_ttCR/` (`identity_combine_ttCR.json`, `muladder_{pyhf,combine}.json`, `scan_{pyhf,combine}_ttCR.json`, `combine_headline_ttCR.json`, and `comparison_table_ttCR.json`), with the overlay figure at `phase4_inference/regression_ttCR/outputs/figures/scan_pyhf_vs_combine_ttCR.pdf`. This chain is invoked separately and is not folded into the main `all` task.

G.6 Summary

An independent RooFit/RooStats engine (CMS Combine) reproduces the pyhf in-situ-$t\bar{t}$ D_NN fit, including the shared k_ttbar and the correlated b-tag nuisance. After removing three subtle translation bugs (an AsymPow norm-injection double-count, an r^2 double-POI, and an autoMCStats expectedEvents drift), the two engines agree on everything that is statistically determined: the profiled likelihoods coincide to ≤0.005 at every template knot and 10^{-6} on the μ-ladder, the profile scans overlay to ≤0.055 over $\mu\in[0,2]$, and the headline numbers reproduce ($\hat{r}$ = 1.168 vs $\hat{\mu}$ = 1.200, k 0.647 vs 0.653, Z 1.152 = 1.152). The pyhf result in the body is therefore independently validated by CMS Combine.

References

ATLAS Collaboration. 2012. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." *Phys. Lett. B* 716: 1. https://doi.org/10.1016/j.physletb.2012.08.020.

Baker, Steve, and Robert D. Cousins. 1984. "Clarification of the Use of Chi-Square and Likelihood Functions in Fits to Histograms." *Nucl. Instrum. Meth.* 221: 437–42. https://doi.org/10.1016/0167-5087(84)90016-4.

Barlow, Roger, and Christine Beeston. 1993. "Fitting Using Finite Monte Carlo Samples." *Comput. Phys. Commun.* 77: 219. https://doi.org/10.1016/0010-4655(93)90005-W.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785. https://doi.org/10.1145/2939672.2939785.

CMS Collaboration. 2012. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." *Phys. Lett. B* 716: 30. https://doi.org/10.1016/j.physletb.2012.08.021.

CMS Collaboration. 2013. *CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update.* CMS-PAS-LUM-13-001. CERN. https://cds.cern.ch/record/1598864.

CMS Collaboration. 2014. "Evidence for the 125 GeV Higgs Boson Decaying to a Pair of Tau Leptons." *JHEP* 05: 104. https://doi.org/10.1007/JHEP05(2014)104.

CMS Collaboration. 2015. "Performance of the CMS Missing Transverse Momentum Reconstruction in Pp Data at Sqrt(s) = 8 TeV." *JINST* 10: P02006. https://doi.org/10.1088/1748-0221/10/02/P02006.

CMS Collaboration. 2016. "Reconstruction and Identification of Tau Lepton Decays to Hadrons and Tau Neutrino at CMS." *JINST* 11: P01019. https://doi.org/10.1088/1748-0221/11/01/P01019.

CMS Collaboration. 2017. "Jet Energy Scale and Resolution in the CMS Experiment in Pp Collisions at 8 TeV." *JINST* 12: P02014. https://doi.org/10.1088/1748-0221/12/02/P02014.

CMS Collaboration. 2018. "Observation of the Higgs Boson Decay to a Pair of Tau Leptons with the CMS Detector." *Phys. Lett. B* 779: 283. https://doi.org/10.1016/j.physletb.2018.02.004.

CMS Collaboration. 2024. "The CMS Statistical Analysis and Combination Tool: Combine." *Comput. Phys. Commun.* 302: 109419. https://doi.org/10.1016/j.cpc.2024.109419.

Cowan, Glen, Kyle Cranmer, Eilam Gross, and Ofer Vitells. 2011. "Asymptotic Formulae for Likelihood-Based Tests of New Physics." *Eur. Phys. J. C* 71: 1554. https://doi.org/10.1140/epjc/s10052-011-1554-0.

Cranmer, Kyle, George Lewis, Lorenzo Moneta, Akira Shibata, and Wouter Verkerke. 2012. *HistFactory: A Tool for Creating Statistical Models for Use with RooFit and RooStats.* CERN-OPEN-2012-016. CERN. https://cds.cern.ch/record/1456844.

Ellis, R. Keith, Ian Hinchliffe, Mark Soldate, and J. J. van der Bij. 1988. "Higgs Decay to Tau+ Tau-: A Possible Signature of Intermediate Mass Higgs Bosons at SSC." *Nucl. Phys. B* 297: 221. https://doi.org/10.1016/0550-3213(88)90019-3.

Heinrich, Lukas, Matthew Feickert, Giordon Stark, and Kyle Cranmer. 2021. "Pyhf: Pure-Python Implementation of HistFactory Statistical Models." *J. Open Source Softw.* 6: 2823. https://doi.org/10.21105/joss.02823.

LHC Higgs Cross Section Working Group. 2013. *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties.* CERN-2013-004. CERN. https://doi.org/10.5170/CERN-2013-004.

Particle Data Group. 2024. "Review of Particle Physics." *Phys. Rev. D* 110: 030001. https://doi.org/10.1103/PhysRevD.110.030001.

Pivarski, Jim, Ianna Osborne, Pratyush Das, Anish Biswas, and Peter Elmer. 2020. "Awkward Array: JSON-Like Data, NumPy-Like Idioms." *Proceedings of the 19th Python in Science Conference (SciPy 2020)*, 78. https://doi.org/10.25080/Majora-342d178e-00b.

Read, Alexander L. 2002. "Presentation of Search Results: The CLs Technique." *J. Phys. G* 28: 2693. https://doi.org/10.1088/0954-3899/28/10/313.

Smith, Nicholas et al. 2024. *Rhalphalib: A Python Toolkit for Building Combine Workspaces*. https://github.com/nsmith-/rhalphalib.