Data-driven fake-τ backgrounds: from binned fake factors to a learned, continuous fake factor

Validation in the type-II/III seesaw searches

Partikeldagarna 2026, Lund

17 September 2026

Lara Calic, Lund University

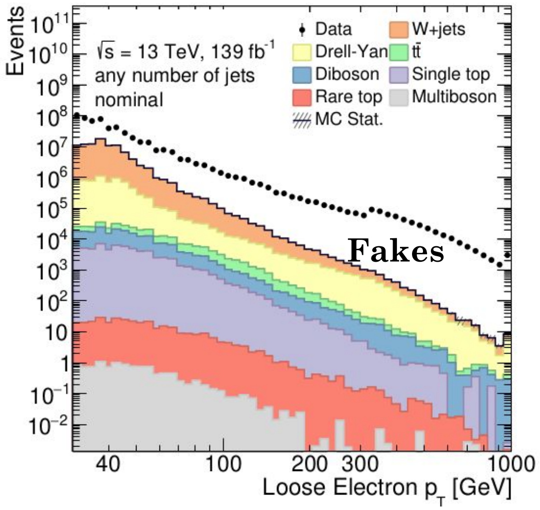
Motivation

- Many analyses in ATLAS suffer from from fakes or non-prompt objects

- The probability for fake to pass all the selection criteria depends:
    - Detector response
    - Physics processes
    - Reconstruction techniques → Motivation for **data-driven approaches!**
    - Event kinematics

- Different analyses require different analysis strategy

arXiv: [2211.07505]



What do we consider by term “fakes’?

- Generally, we deal with two different “types” of fakes:
    - **Non-prompt leptons**: ‘genuine’ leptons that do not originate from prompt hard interaction (heavy flavour decays, meson decays, ..)
    - **Misconstructed leptons**: reconstructed as leptons (but the underlying particle was different for e.g. jets reconstructed as electrons, electrons reconstructed as photons, …)

Simplified idea: How do we estimate fakes?

Goal: estimate the fake contribution in the SR* using measurements performed elsewhere.

Same goal for every method, they differ in **what** is measured, **how** it is transferred and **where** it is measured!

Where?

1. Object selection
Signal region enriched
with prompt object

What?

2. Measure the detector response
Prompt/Fake efficiencies
Template extraction
Fake/Transfer factors (FF/FT)

How?

3. Fake rate prediction
Event weights
Template fit
Fake estimation from FF/FT

Does it make sense?

4. Validation
Closure tests
Validation regions
Residual corrections

How well?

5. Systematic uncertainties
Extrapolation
Statistics
Detector description

***signal region (SR)** = region of phase space where we look for our signal
(defined by object requirements and kinematic cuts)

All four methods are centrally supported in ATLAS, the most optimal choice is **analysis dependent**

Method	STEP 2 **What is measured**	STEP 3 **How the SR yield is built**	**Typical use**
Matrix method	real and fake efficiencies ε_r, ε_f	invert the efficiency matrix $\to$ per-event weights	multi-lepton, low statistics
Fake factor	$F = \varepsilon_f/(1-\varepsilon_f)$ in a fake-enriched CR	reweight anti-ID data, event by event	**τ and lepton searches**
Template fit	shape of an ID or isolation variable	fit the fake normalisation in the SR	photons, jet $\to \gamma$
ABCD	yields in two uncorrelated variables	$N_D = N_B N_C / N_A$	inclusive, low statistics

Fake Factor method (for 2 leptons)

- **Data-driven technique** to estimate the contribution of fake leptons in the signal (tight) region
- Two sets of lepton identification criteria: **tight** and **loose** (without tight)
- The fake factor F is defined as the ratio:

$$F = \frac{f\, N_f^L}{(1-f)\, N_f^L} = \boxed{\frac{f}{1-f}}$$

- Estimated number of fake events in the signal region is:

	Tight	Loose
Loose	LT 1 fake lepton $F_1(N_{LT}^{\text{data}} - N_{LT}^{\text{MC, prompt}})$	LL 2 fake leptons $-F_1F_2(N_{LL}^{\text{data}} - N_{LL}^{\text{MC, prompt}})$
Tight	TT SR	TL 1 fake lepton $F_2(N_{TL}^{\text{data}} - N_{TL}^{\text{MC, prompt}})$

$$N_{TT}^{\text{fakes}} = \underbrace{F_1\left(N_{LT}^{\text{data}} - N_{LT}^{\text{MC, prompt}}\right)}_{\substack{\text{1 fake lepton (LT region)} \\ \text{fake factor } F_1\times(\text{data}-\text{prompt MC})}} + \underbrace{F_2\left(N_{TL}^{\text{data}} - N_{TL}^{\text{MC, prompt}}\right)}_{\substack{\text{1 fake lepton (TL region)} \\ \text{fake factor } F_2\times(\text{data}-\text{prompt MC})}} - \underbrace{F_1F_2\left(N_{LL}^{\text{data}} - N_{LL}^{\text{MC, prompt}}\right)}_{\substack{\text{2 fake leptons (LL region)} \\ \text{fake factor } F_1F_2\times(\text{data}-\text{prompt MC}),\ \text{subtracted}}}$$

Goal: Improve fake factor estimation using machine learning tools!

ML method: Continuous fake factor estimation

Why?

- Traditionally, fake factors are binned in p_T, $|\eta|$, and MET (some **limitations**: limited statistics per bin, interpolation issues, and difficulty extrapolating to more variables)

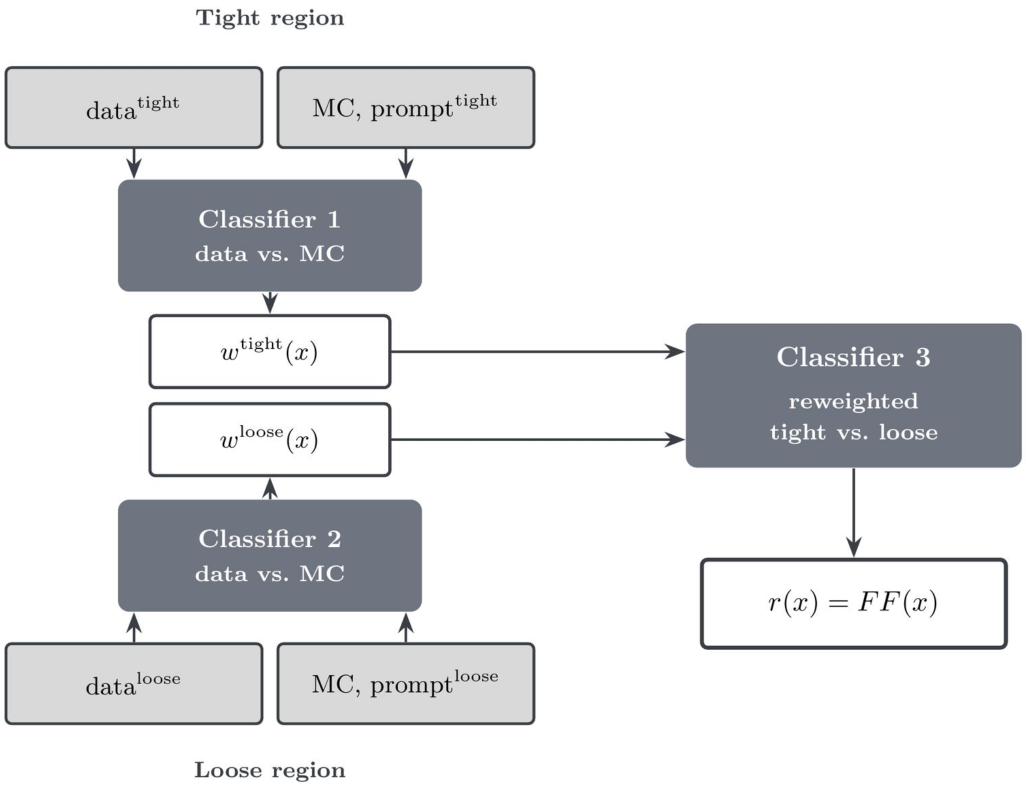
How?

- Idea: estimate the fake factor as an unbinned, continuous density ratio, learned directly from data

The density-ratio trick:

The fake factor is a ratio of prompt-subtracted densities in the tight and loose regions:

$$r(x) \equiv w(x) = \frac{p\left(x^{\text{tight}}_{\text{data}}\right) - p\left(x^{\text{tight}}_{\text{MC}}\right)}{p\left(x^{\text{loose}}_{\text{data}}\right) - p\left(x^{\text{loose}}_{\text{MC}}\right)}$$

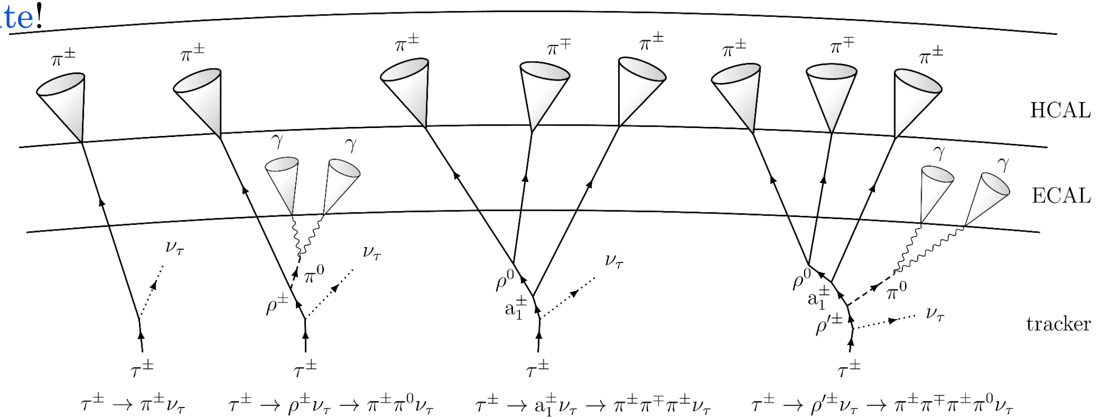


JHEP04(2026) 188, arXiv: [2511.06972]

- τ leptons decay before the detector ($c\tau \approx 87\ \mu m$) – they decay inside the beam pipe, before the first pixel layer
  - 65% hadronically → we only ever see $\boldsymbol{\tau}_{\textbf{had-vis}}$: 1 or 3 charged tracks and π^0s, inside a narrow cone
  - 35% leptonic → look like e/μ
  - Broadly speaking, there are 2 flavours of fake leptons:
    - **non-prompt real objects**
    - **misidentified**
  - Objects that we need to estimate are reconstructed jets that we identify as $\boldsymbol{\tau}_{\textbf{had-vis}}$

Source: Izaak Neutelings, https://tikz.net/tau_decay/

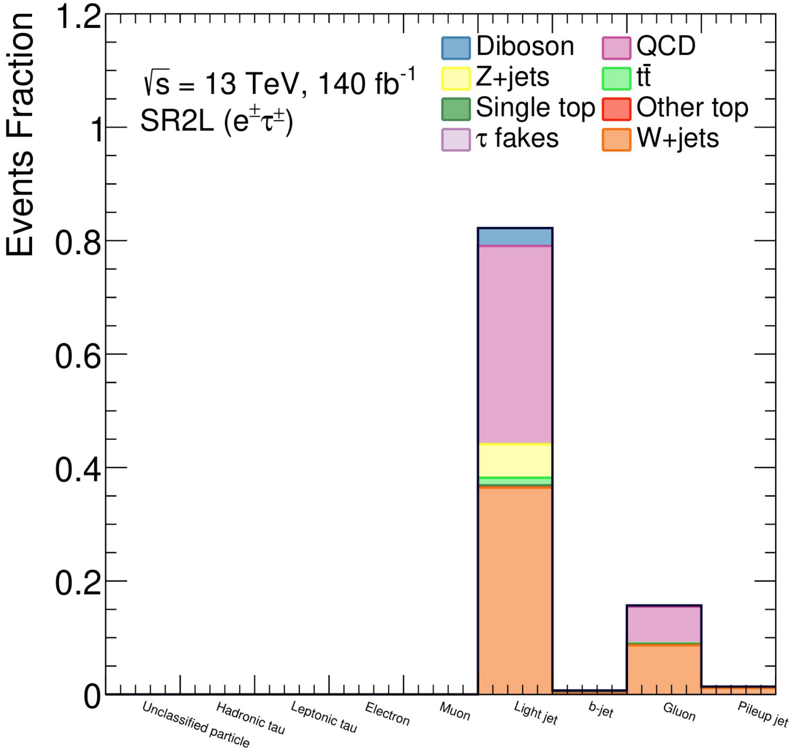
Every τ_{had} analysis needs a data-driven estimate!



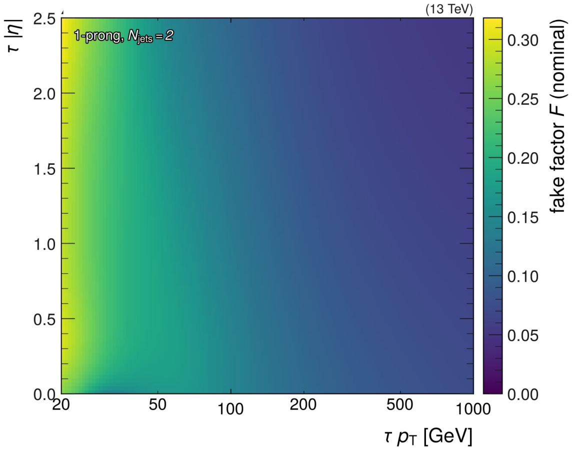
Source of τ_{had} candidates

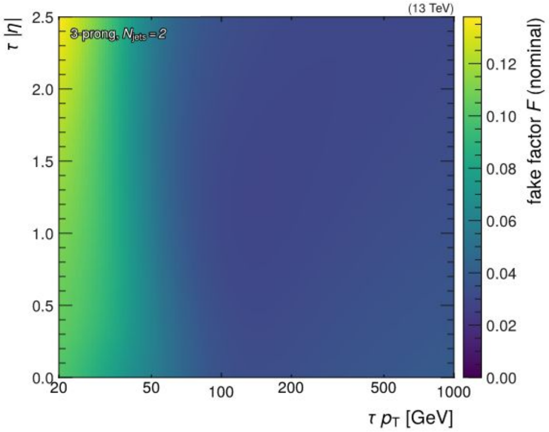
- Also, we can distinguish fakes into 4 different categories based on truth flavor:
  - light quark (q)
  - gluon (g)
  - b
  - pile -up (p)
- 80 % of τ_{had} candidates are light quark jets

Work in progress!



(τ) **Fake factor methods:** Learned continuous fake factor





Fake factor is a density ratio:

$$F = \frac{N^T_{\text{data}} - N^T_{\text{MC}}}{N^L_{\text{data}} - N^L_{\text{MC}}}$$

$$\boxed{F(x) = \frac{p^T(x)}{p^L(x)}}$$

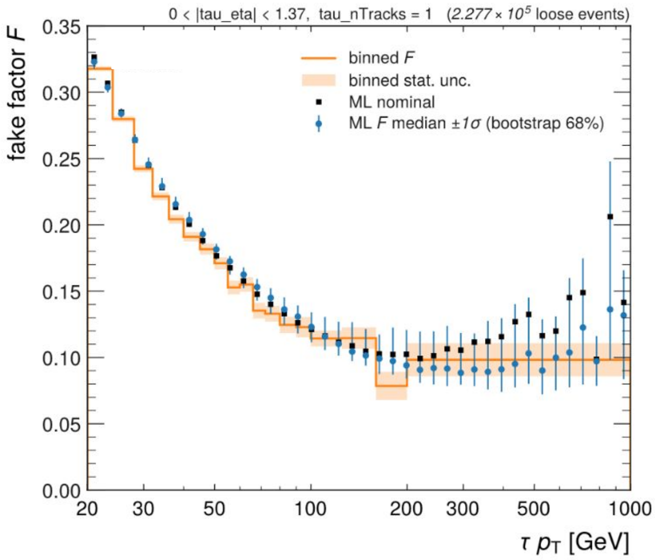
The idea: estimate the ratio with classifiers trained on data (one for subtraction and one for tight/loose ratio) → per event F(x) in 11 variables

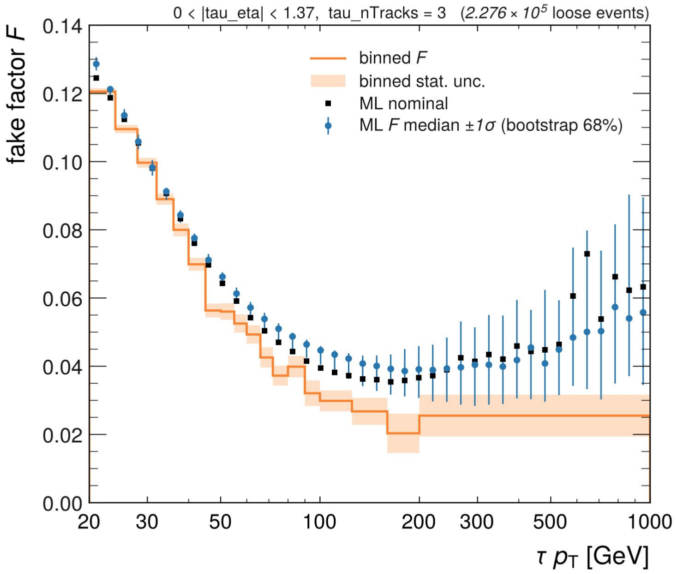
- More details: JHEP04(2026) 188, arXiv: [2511.06972]
- Statistical uncertainties come from a Poisson bootstrap
- Classifier based density ratio estimation, trained on data

Measured binned fake factors for taus (Tight)

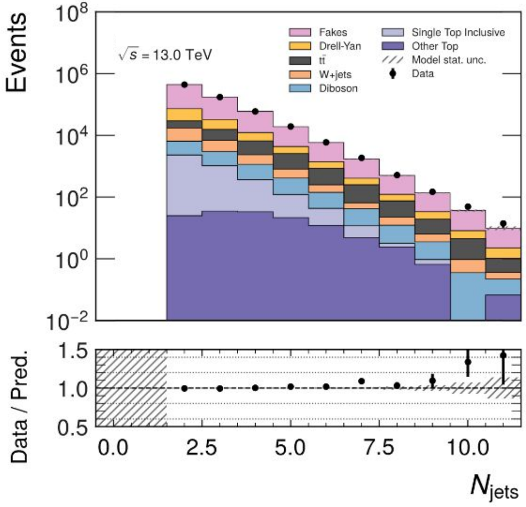
Typical CR: W ($\mu\nu$)+jets: trigger on the muon (τ candidate is trigger-unbiased)

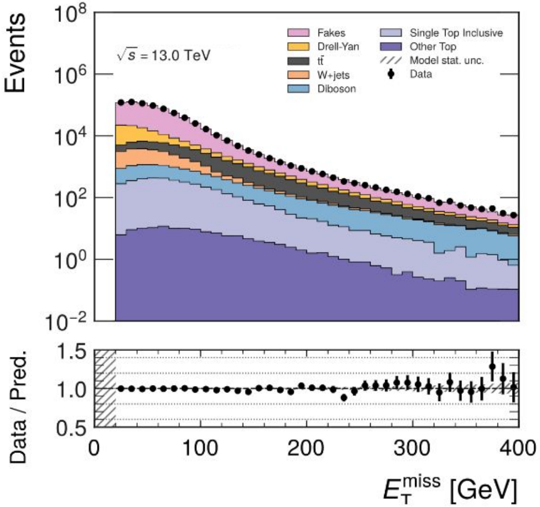
Work in progress!

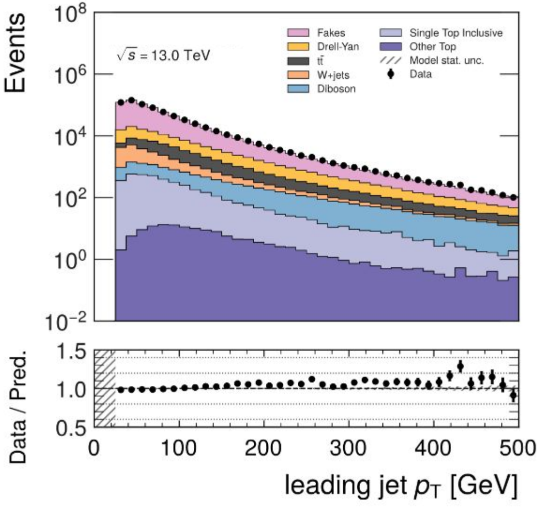




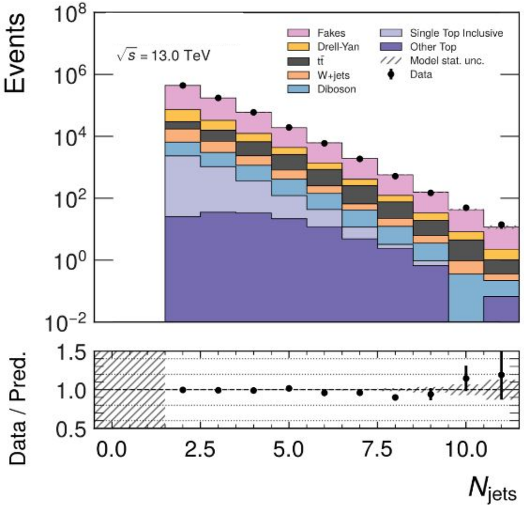
Work in progress!

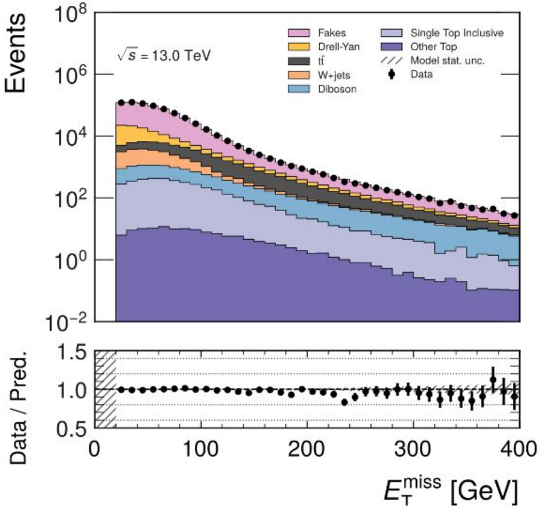


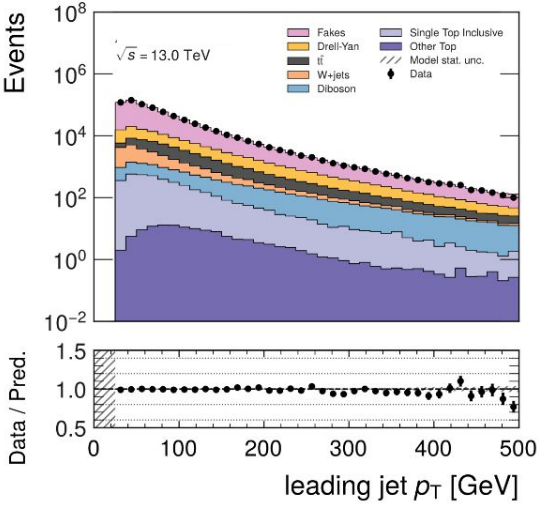




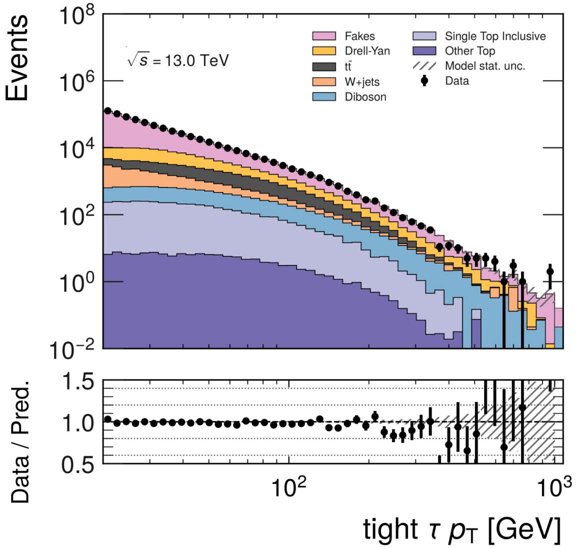
Work in progress!

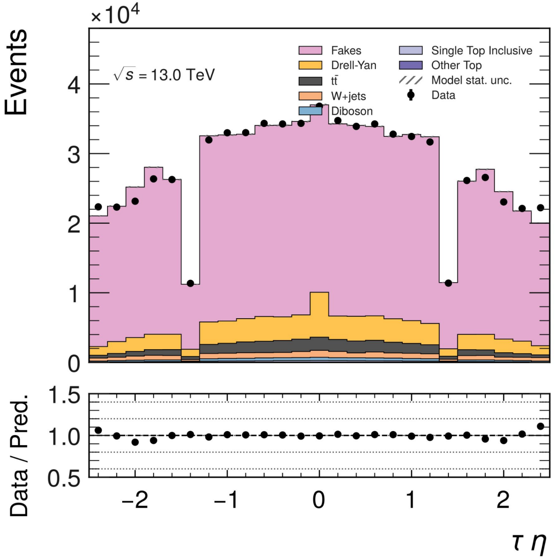


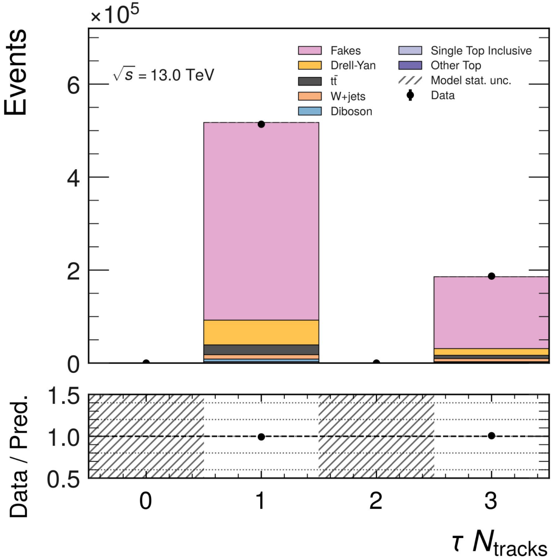




Work in progress!

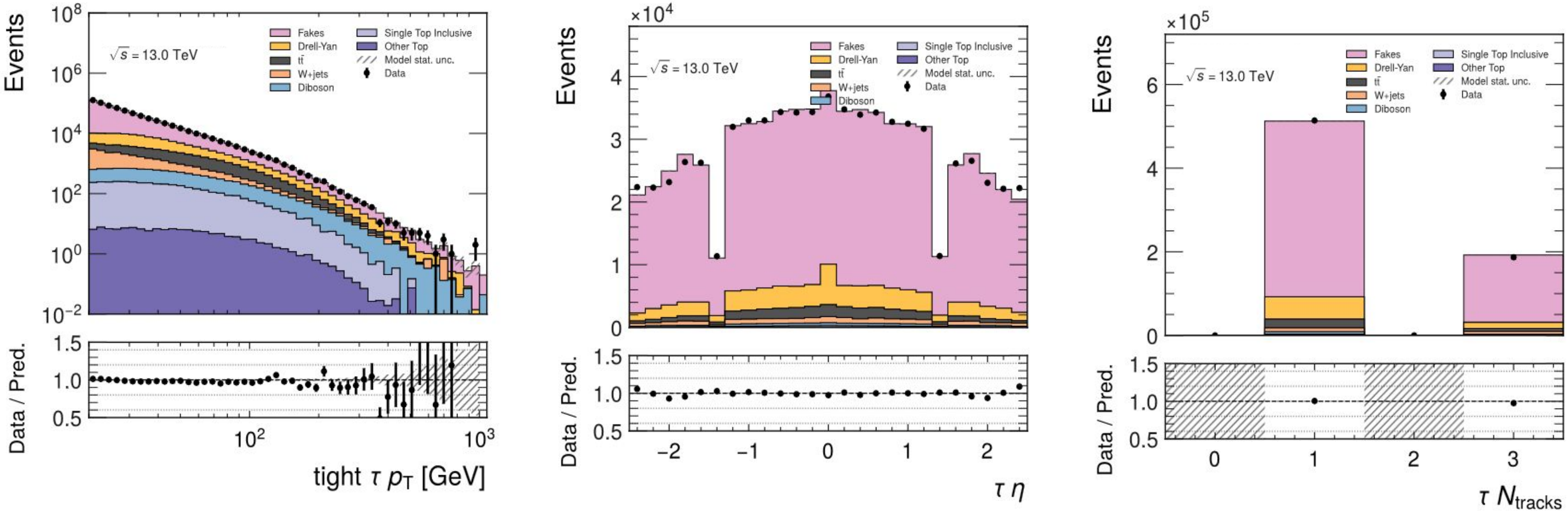






- Fake factors applied back to the anti-ID sample of the same region (prediction)

Work in progress!



- Fake factors applied back to the anti-ID sample of the same region (prediction)
- Prediction agrees with data, directly comparable to the binned closure shown earlier

Conclusions and Future work

- Fake taus cannot be taken from simulation → data-driven estimate is a must
- Binned fake factors measured in W(μν)+jets CR, 1- and 3-prong, two η ranges
- Closure test passes in tau p_T, η and number of tracks
- Main limitations: CR/SR composition mismatch and coarse binning at high p_T
- Learned fake factor: unbinned, per-event, 11 variables, trained on data

Future work:

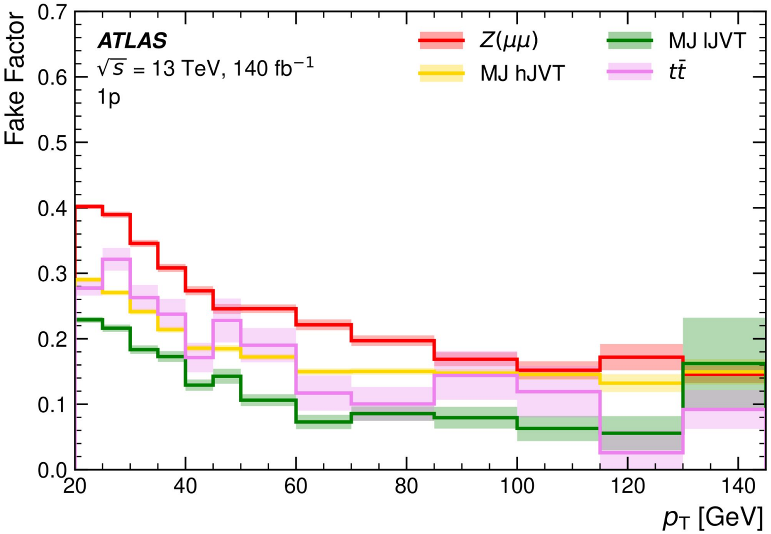
- Validate learned fake factor in type-II/III seesaw regions
- Compare binned vs. UFF vs. learned on the same regions
- Combine UFF composition handling with the continuous estimate
- Full systematics: composition, extrapolation, training variations

Backup slides

1. **Composition:**

arxiv:2502.04156v2

- The fake factor is different for each fake factor type light quark (q), gluon (g), b and pile up (p) jets fake τ at different rates
- We cannot measure per type fake factors directly (every region in a phase space is a mixture)
- The "ideal" CR would have same q,g,b, p mixture as the SR
  (not an issue task to do → need for dedicated CR per SR)
- Composition mismatch is the dominant systematics



2. **Binning:**

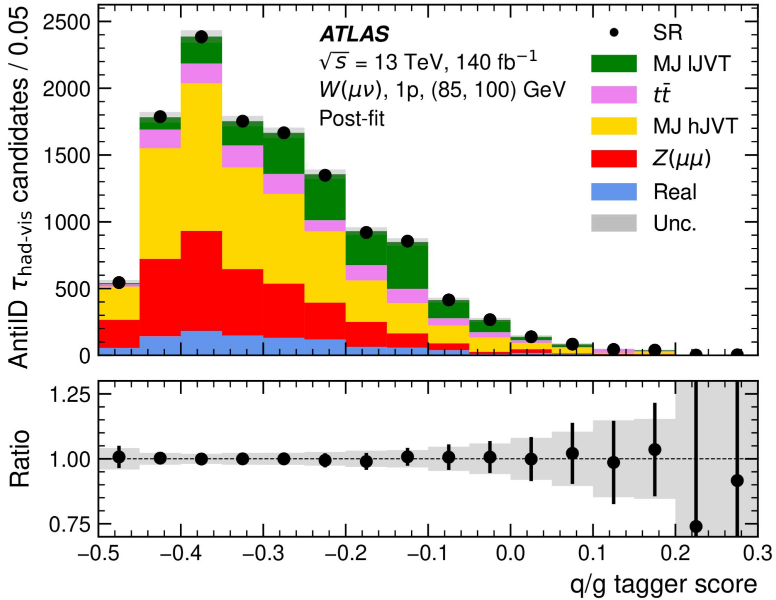
- p_T ,η and prongness is 3D projection depending on E_T^{miss}, N_{jets}
- At high p_T the parameterization degenerates into one wide bin

Fix: Universal Fake Factor Method

$$F_{\mathrm{UFF}} = \underbrace{\mu_Z\, F_{Z(\mu\mu)}}_{q} + \underbrace{\mu_{t\bar{t}}\, F_{t\bar{t}}}_{b} + \underbrace{\mu_{\mathrm{MJ}_h}\, F_{\mathrm{MJ}_h}}_{g} + \underbrace{\mu_{\mathrm{MJ}_l}\, F_{\mathrm{MJ}_l}}_{\text{pile-up}}$$

- **Idea**: 4 predefined fake enriched regions (measured once)
- **Origin of weights**: template fit of a quark/gluon tagger score to anti-ID candidates of your signal region
- The fit determines the *composition*, and not fake factors
- An analysis provides **one histogram**: no dedicated, perfectly matched CR needed
- It solves **1**. and it does not do anything with **2**.

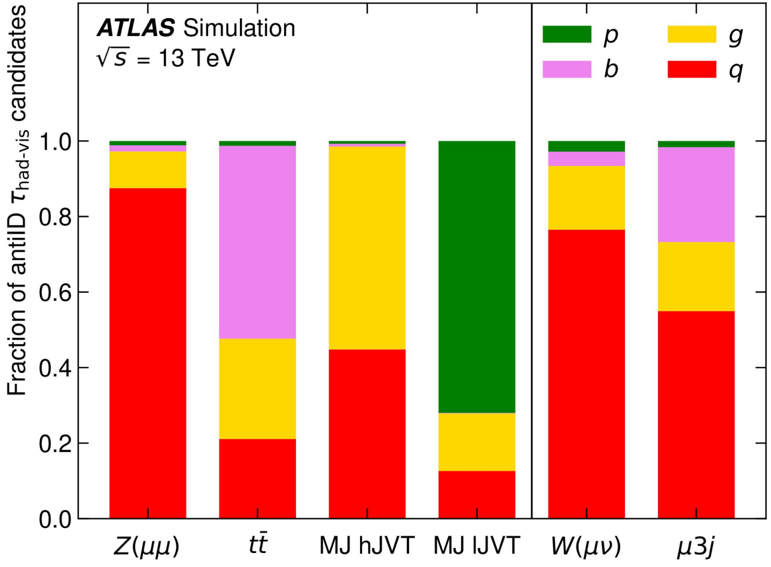
Still binned!



arxiv:2502.04156v2

UFF method: Composition problem

arXiv:2502.04156



Data-driven estimation is more than needed!

**So, we do not want to simulate fakes "better", we measure them.*

- Misidentification probability depends on **soft fragmentation, π^0 reconstruction** and **hadronic response of detector** (the least well modelled "part" of simulation)
- We need MC statistics scale as following:
  - N_{jets} x (mis-ID rate)$^{-1}$:
- As shown on the left plot, fake composition *changes* from one region to another
- The fake factor is different for **each fake factor type**
  - Light quark (q), gluon (g), b and pile up (p) jets fake τ at different rates

Method 1: Matrix Method

$$\underbrace{\begin{pmatrix} N^t \\ N^{\bar{t}} \end{pmatrix}}_{\text{counted in data}} = \underbrace{\begin{pmatrix} \varepsilon_r & \varepsilon_f \\ 1-\varepsilon_r & 1-\varepsilon_f \end{pmatrix}}_{\text{measured elsewhere}} \underbrace{\begin{pmatrix} N^b_r \\ N^b_f \end{pmatrix}}_{\text{unknown}}$$

$$N^b_f = \frac{(1-\varepsilon_r)\,N^t - \varepsilon_r\,N^{\bar{t}}}{\varepsilon_f - \varepsilon_r}$$

$$N^t_f = \varepsilon_f\,N^b_f$$

N^t events passing the *tight* selection

$N^{\bar{t}}$ events in the *anti-tight* sample

ε_r probability that a loose *prompt* lepton also passes tight

ε_f probability that a loose *fake* also passes tight

N^b_r prompt events in the loose (baseline) sample

N^b_f fake events in the loose (baseline) sample

- Needs both efficiencies: one from tag-and-probe, and another from a fake enriched CR
- Generalises to N objects at the cost of a 2^N x 2^N matrix
- Two variants:
  - **Asymptotic Matrix method**
  - **Poisson likelihood Matrix Method**

In high statistics limit, all methods agree → for *τ_had* ***all uncertainties comes from fakes!***

Method 2: Fake factor method

Fake factor is **simplified matrix method** - we define loose (not tight) and tight regions that are orthogonal to each other

Define two lepton definitions:

a) Tight: nominal signal selection
b) Loose: Relaxed identification, isolation cuts
c) Strictly loose: tight not loose region, used for fake Estimation

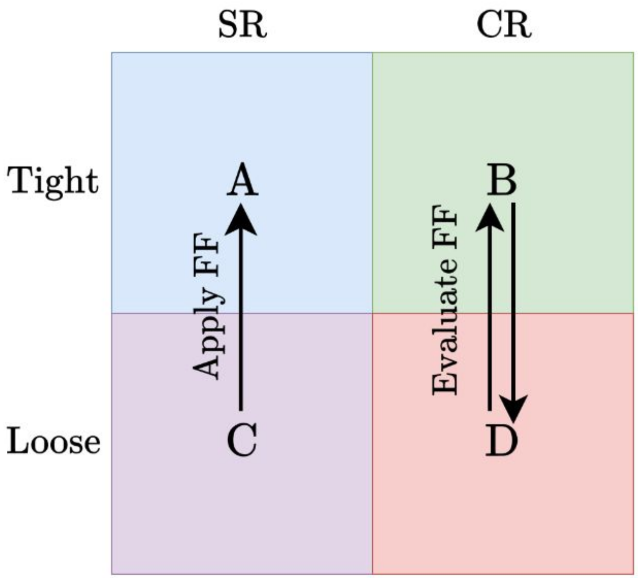
Traditionally, binned in **p_T, η** and **prongness**

Fake background estimated using the Fake factor method

- Using single lepton regions
- Fake estimation is applied in fake enriched region

$$N^{\text{fake}} = \sum_{i=1}^{N_{\text{SB}}^{\text{data}}} (-1)^{N_{L,i}+1} \prod_{l=1}^{N_{L,i}} F_l - \sum_{i=1}^{N_{\text{SB}}^{\text{MC}}} (-1)^{N_{L,i}+1} \prod_{l=1}^{N_{L,i}} F_l$$

$$\stackrel{l=2}{=} \left[\sum_{TL} F_2 + \sum_{LT} F_1 - \sum_{LL} F_1 F_2\right]_{\text{data}} - \left[\sum_{TL} F_2 + \sum_{LT} F_1 - \sum_{LL} F_1 F_2\right]_{\text{prompt}}$$



Definition

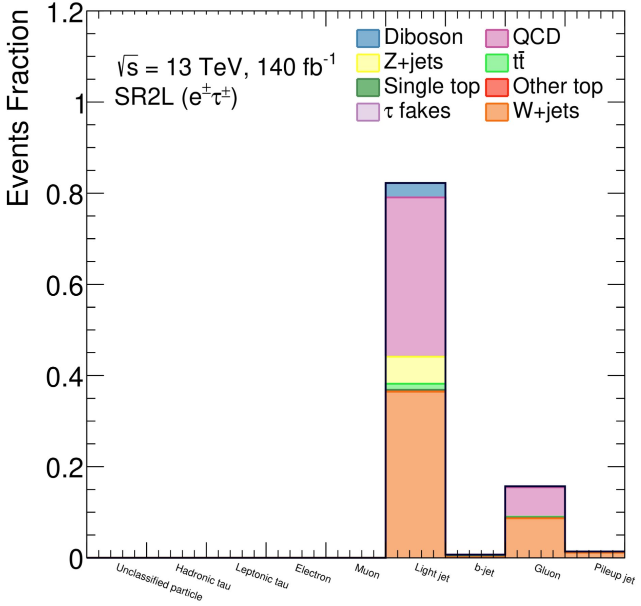
- **Reconstruction**: τ_had-vis are seeded by anti k_T (R =0.4 jets with p_T > 10 GeV, $|\eta|$ < 2.5 with 1 or 3 associated tracks with $|q|$= 1)

- **Identification**: NN jets discriminant based on associated tracks and calorimeter clusters
  - RNN (for Run 2)
  - GNTau (for Run 3)
  - With loose, medium and tight working point

- **Truth definition**: reconstructed candidate is
  - *real* if it is matched within $\Delta R < 0.2$ to hadronic τ
  - *fake* if it fails
- Usually, we apply e (μ) veto to remove lepton → τ candidates
  → in practice, we consider ***fake≡jet*** $\rightarrow \tau_{had\text{-}vis}$

- Also, we can distinguish fakes into 4 different categories based on truth flavor:
  - **light quark (q)**
  - **gluon (g)**
  - **b**
  - **pile -up (p)**

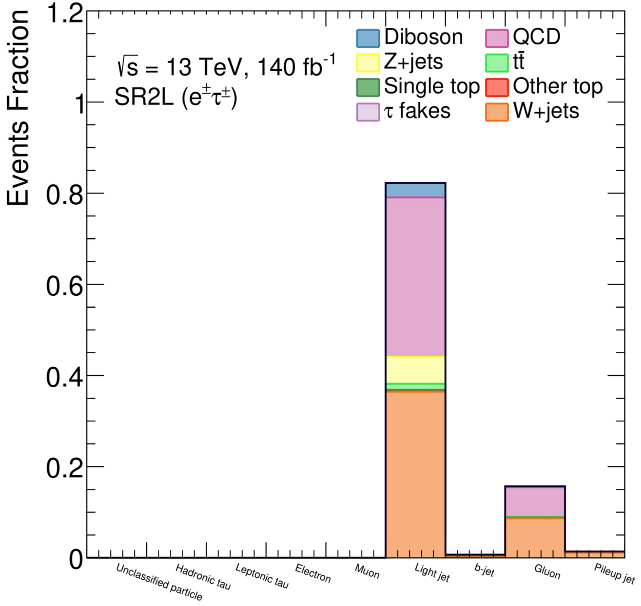
Work in progress!



- **Prong** = a charged track from the τ decay. Hadronic τ decays come in two flavours:
  - **1-prong**: one charged hadron (π± or K±), possibly with neutral pions, plus the neutrino. Example $\tau^- \to \pi^- \nu$, $\tau^- \to \pi^- \pi^0 \nu$. About 3 out of 4 hadronic τ decays.
  - **3-prong**: three charged hadrons, possibly with neutral pions, plus the neutrino. Example $\tau^- \to \pi^- \pi^+ \pi^- \nu$. About 1 in 4 hadronic τ decays

- In both cases the tracks are collimated in a narrow cone and the total charge is ±1, which is why the reconstruction asks for exactly 1 or 3 tracks with |q| = 1.

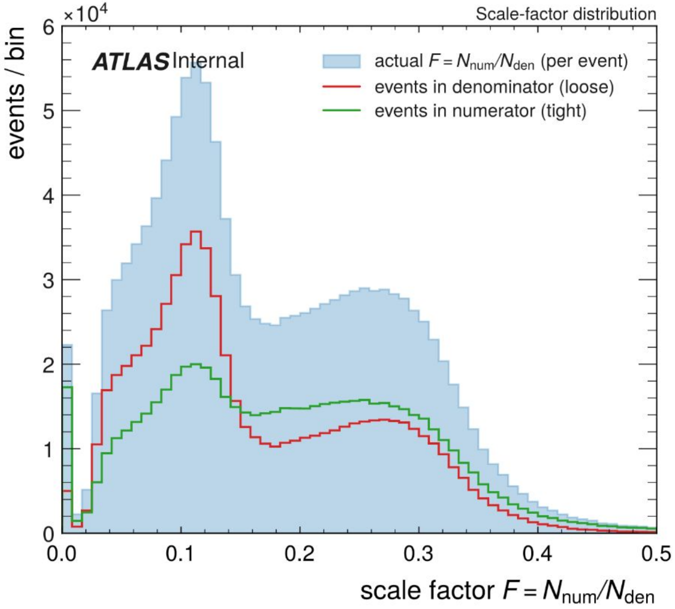
Work in progress!



Source: Exotics Plenary: data-driven background estimation methods, talk by Jan

Application to fake taus in the W+jets region

- Trained on the **same events** as the binned measurement: the OS $W(\mu\nu)$+jets region (data + MC).
- **11 input features**:
  - τ: p_T, $|\eta|$, ϕ, prongness,
  - event: N_{jets}, E_T^{miss}, ϕ^{miss}, E_T^{miss} significance,
  - leading jet: p_T, η, ϕ.
- Output: a **continuous fake factor per event**.
- The 1-prong ($F \approx 0.25$) and 3-prong ($F \approx 0.10$) fake factors appear automatically, each spread by the kinematic dependence.

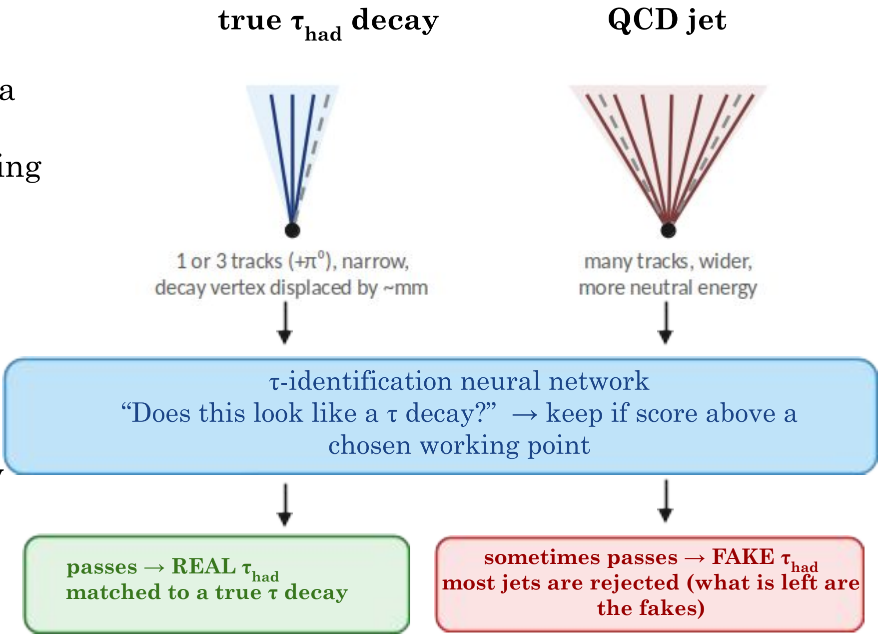


distribution of per-event fake factors in the W+jets region

17 / 23

What is a "fake τ"? — a jet that the τ identification chooses

- **A τ_{had} candidate starts as a jet**
  - Take a narrow jet with 1 or 3 charged tracks and ask a neural network: does this look like a τ decay or like a QCD jet? Candidates above a chosen threshold ("working point") are our τ_{had}
- **Real vs fake (in simulation):** real if it matches a true hadronic τ decay, fake otherwise
- **In this talk, fake = jet → τ_{had}**
  - Electrons and muons can also fake a τ, but they are removed by dedicated vetoes and estimated separately
- **Not all jets fake equally**
  - Quark, gluon, b and pile-up jets have different track multiplicities and shapes, so their fake rates differ (this is what makes fake estimation hard)



How big is it? In a typical signal region ~80% of τ_{had} candidates are jets → mostly light-quark jets from W+jets and QCD

Poisson bootstrap for uncertainties

- Statistical uncertainty on $F(\boldsymbol{x})$ is estimated with a Poisson bootstrap vectorized inside the model.
- Each event weight i receives, in each of K channels a multiplier:

$$\nu_i^{(k)} \sim \text{Poisson}(1) \quad \text{for } k = 1, 2, \ldots, K-1 \quad \text{and} \quad \nu_i^{(0)} = 1\,. \tag{1}$$

- This is computed deterministically by hashing the event number together with k, so the same event always receives the same multipliers.
- Channel 0 is the nominal fit, channels $1, 2, \ldots, K-1$ are trained on resamples of the full (data + MC) samples.
- A network has K heads (one ensemble member per channel), evaluated in a single vectorized forward pass.
- Every loss term is a $\nu^{(k)}$ weighted mean (the K channels are K independent estimators trained simultaneously).
- The uncertainty band is the 16th/84th percentiles of channels $1, 2, \ldots, K-1$.