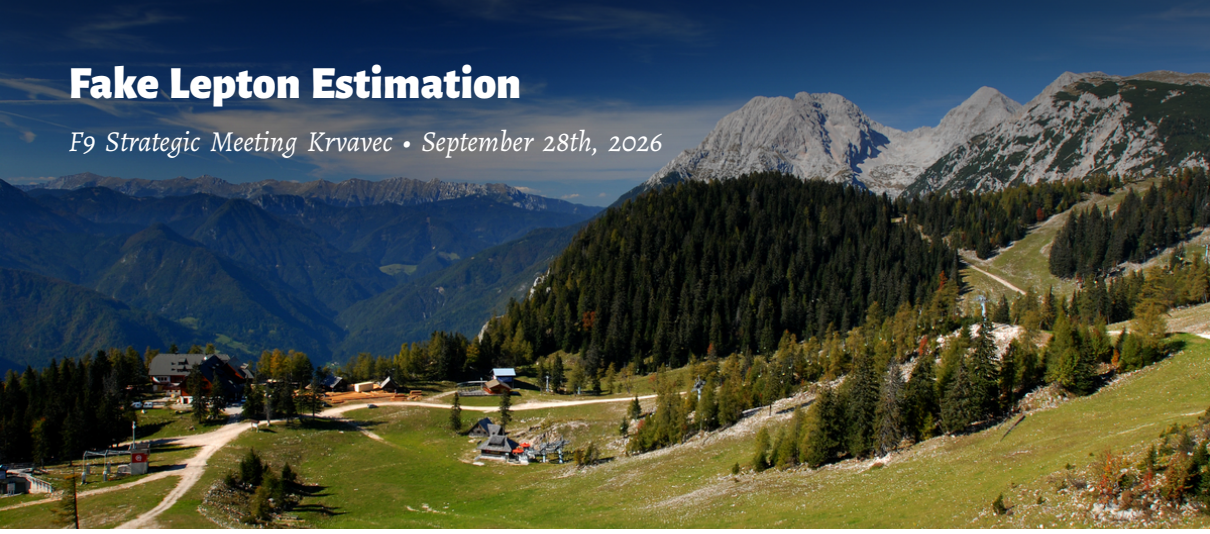
Fake Lepton Estimation

F9 Strategic Meeting Krvavec • September 28th, 2026



Jan Gavranovič

Jožef Stefan Institute and University of Ljubljana

Institut "Jožef Stefan" Ljubljana, Slovenija

ATLAS EXPERIMENT

Why fake leptons, and why data-driven

- Any analysis that selects leptons (e, μ, τ_{had}) also selects objects that only *look* like prompt leptons: **fake (misidentified or non-prompt) leptons**.
- In many final states (multilepton, same-sign, low lepton p_{T}, ...) they are **a dominant background**.
- **MC cannot model fakes reliably** (tails of jet fragmentation and detector response, enormous MC statistics needed) $\rightarrow$ data-driven estimation is mandatory.
- Example in this talk: **electrons**, faked by sources with **different fake rates**:
  - light-flavour jets misidentified as electrons,
  - photon conversions ($\gamma \rightarrow e^+e^-$),
  - non-prompt electrons from semileptonic heavy-flavour (b, c) decays.
- ATLAS uses a family of data-driven techniques (fake factors, matrix method, template fits, ABCD, ...) $\rightarrow$ choice depends on the analysis.

This talk: fake-factor (FF) method for electrons $\rightarrow$ F as a density ratio $\rightarrow$ ML estimator.

The fake-factor method

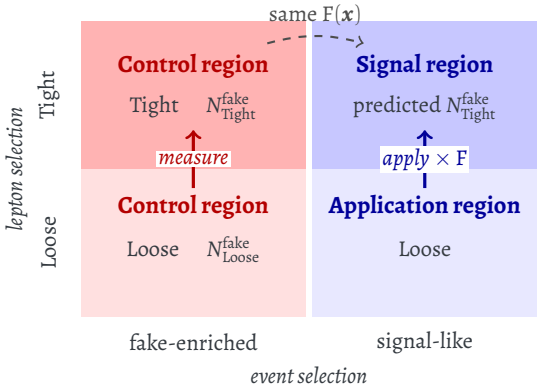
- Split by **lepton selection**: **Tight** (analysis ID + isolation) and **Loose** (relaxed, fail Tight).
- **Measured** in a fake-enriched **control region (CR)**, real leptons subtracted with MC:

$$\mathrm{F} = \left.\frac{N^{\text{data}}_{\text{Tight}} - N^{\text{real}}_{\text{Tight}}}{N^{\text{data}}_{\text{Loose}} - N^{\text{real}}_{\text{Loose}}}\right|_{\text{CR}}.$$

- **Applied** to the Loose events of the SR (Tight SR data never used):

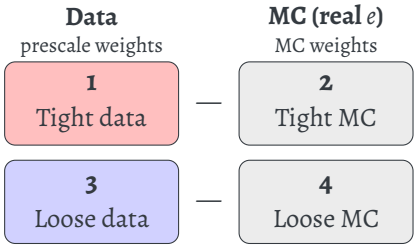
$$N^{\text{fake}}_{\text{Tight}} = \mathrm{F} \times \left.\left(N^{\text{data}}_{\text{Loose}} - N^{\text{real}}_{\text{Loose}}\right)\right|_{\text{SR}}.$$

- Assumption: **same** F **in CR and SR at fixed** $\boldsymbol{x}$ $\rightarrow$ parametrized in $\boldsymbol{x}$ (e.g. $p_{\mathrm{T}} \times |\eta| \times E^{\text{miss}}_{\mathrm{T}}$).



Example: electron fakes in a fake-enriched region

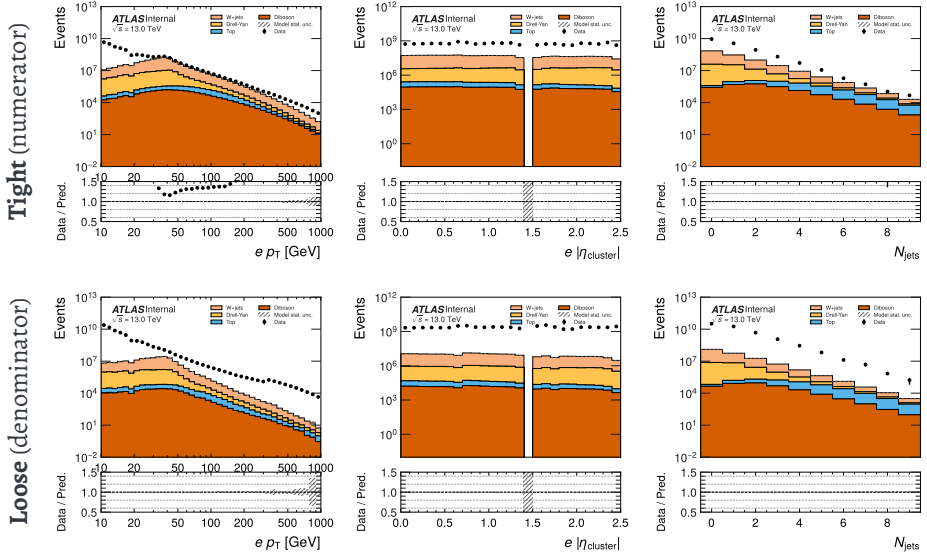
- Events collected with **prescaled electron triggers**.
- Price: every data event carries its **prescale as a weight**, from 1 up to 10^5 → **weighted** densities.
- b**-jet veto** suppresses $t\bar{t}$ and heavy-flavour decays.
- **Real electrons** (e.g. $W(e\nu)$+jets, $Z(ee)$+jets) in both subregions are subtracted with MC → **four samples**.
- Baseline: **binned** F in $p_{\mathrm{T}} \times |\eta| \times E_{\mathrm{T}}^{\mathrm{miss}}$ → reference for the ML estimate.



$$\mathrm{F} = \left.\frac{N_1 - N_2}{N_3 - N_4}\right|_{\mathrm{CR}}$$

Dataset	Trigger	electron p_{T}	electron $\lvert\eta\rvert$	$N_{b\text{-jet}}$	F binning
Full Run 2	prescaled electron	$\geqslant$ 10 GeV	$<$ 2.47, excl. 1.37–1.52	$= 0$	$p_{\mathrm{T}} \times \lvert\eta\rvert \times E_{\mathrm{T}}^{\mathrm{miss}}$

Tight and loose distributions



From binned to continuous fake factors

- The fake factor is a **ratio of real-subtracted densities** in the *Tight* and *Loose* regions:

$$\mathrm{F} = \frac{N_f^{\mathrm{T}}}{N_f^{\mathrm{L}}} = \frac{N_{\mathrm{data}}^{\mathrm{T}} - N_{\mathrm{MC}}^{\mathrm{T}}}{N_{\mathrm{data}}^{\mathrm{L}} - N_{\mathrm{MC}}^{\mathrm{L}}} \quad \longrightarrow \quad \mathrm{F}(\boldsymbol{x}) = \frac{N_f^{\mathrm{T}}(\boldsymbol{x})}{N_f^{\mathrm{L}}(\boldsymbol{x})} \, .$$

- Traditionally binned in a few variables (for electrons p_{T}, $|\eta|$, $E_{\mathrm{T}}^{\mathrm{miss}}$): **binning choice** (bias vs. variance), statistics per bin, empty or **negative bins** after the subtraction, jumps at bin edges, and **at most a few dimensions**.
- **Idea: estimate the density ratio with neural networks trained directly on data** ("data-based inference") [JHEP 04 (2026) 188]: a **continuous, unbinned, per-event** fake factor in a high-dimensional feature space.
- Event weights (MC modeling, prescales) and the real-electron subtraction are handled **natively in the training → one network, one training**.

The building blocks

- **Count density** $n(\boldsymbol{x})$: expected events per unit volume of $\boldsymbol{x}$.
- **Weighted density** $D(\boldsymbol{x})$: the same as $n(\boldsymbol{x})$ with weights.
- **Factorization**: count density times conditional mean weight:

$$D(\boldsymbol{x}) = n(\boldsymbol{x})\,\bar{w}(\boldsymbol{x})\,, \qquad \bar{w}(\boldsymbol{x}) = \mathbb{E}[w \mid \boldsymbol{x}] \qquad \left(\text{in one bin: } \textstyle\sum_i w_i = N\bar{w}\right).$$

- **Classifier ratio**: train a classifier on sample a ($y = 1$) vs. sample b ($y = 0$) with binary cross-entropy and *no weights*, its logit q_{ab} gives:

$$\ell(q, y) = -y \log \sigma(q) - (1 - y) \log\big(1 - \sigma(q)\big) \quad \Longrightarrow \quad e^{q_{ab}(\boldsymbol{x})} = \frac{n_a(\boldsymbol{x})}{n_b(\boldsymbol{x})}\,.$$

- **Mean weight regression**: least squares on the events of sample a,

$$\ell(\bar{w}, w) = (\bar{w} - w)^2 \quad \Longrightarrow \quad \bar{w}_a(\boldsymbol{x}) = \mathbb{E}[w \mid \boldsymbol{x}, a]\,,$$

  negative weights only enter inside a square $\rightarrow$ safe.

- No weights in n, the sign lives in $\bar{w} \rightarrow$ signed weights never enter a classifier loss.

Decomposition of the fake factor

- The goal is to express F using only ratios, via the identity $u - v = u(1 - v/u)$.
- Divide numerator and denominator by their own weighted data density:

$$\mathrm{F}(\boldsymbol{x}) = \frac{D_1(\boldsymbol{x}) - D_2(\boldsymbol{x})}{D_3(\boldsymbol{x}) - D_4(\boldsymbol{x})} = \frac{D_1(\boldsymbol{x})\,(1 - D_2(\boldsymbol{x})/D_1(\boldsymbol{x}))}{D_3(\boldsymbol{x})\,(1 - D_4(\boldsymbol{x})/D_3(\boldsymbol{x}))} = r_{13}\frac{1 - 1/r_{12}}{1 - 1/r_{34}}\,,$$

  where every ratio has the same form,

$$r_{ab}(\boldsymbol{x}) = \frac{D_a(\boldsymbol{x})}{D_b(\boldsymbol{x})}\,, \quad \text{so} \quad r_{13} = \frac{D_1}{D_3} \quad \text{and} \quad r_{12} = \frac{D_1}{D_2} \quad \text{and} \quad r_{34} = \frac{D_3}{D_4}\,.$$

- We can now factorize each weighted density using:
  - $D(\boldsymbol{x}) = n(\boldsymbol{x})\bar{w}(\boldsymbol{x})$ to get $D_a = n_a\bar{w}_a$,
  - $e^{q_{ab}(\boldsymbol{x})} = n_a(\boldsymbol{x})/n_b(\boldsymbol{x})$ to get the three classifier heads $e^{q_{12}} = n_1/n_2$, $e^{q_{34}} = n_3/n_4$ and $e^{q_{13}} = n_1/n_3$,
- *The final result is a decomposition into seven parts:*

$$r_{ab}(\boldsymbol{x}) = e^{q_{ab}(\boldsymbol{x})}\,\frac{\bar{w}_a(\boldsymbol{x})}{\bar{w}_b(\boldsymbol{x})} \quad \text{for} \quad (a, b) \in \{(1, 2),\ (3, 4),\ (1, 3)\}\,.$$

Poisson bootstrap for uncertainties

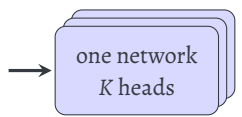
- Statistical uncertainty on $\mathrm{F}(\boldsymbol{x})$ from a **Poisson bootstrap vectorized inside the model**.
- Each event i gets a multiplier per channel k: $p_{0,i} = 1$ (nominal) and $p_{k,i} \sim \mathrm{Poisson}(1)$ for $k \geqslant 1$, fixed by hashing the event number with k.
- The multipliers turn each channel into a **statistically independent pseudo-dataset**.
- Every loss term is a $p_{k,i}$-weighted mean of a per-event quantity v_i over a sample S:

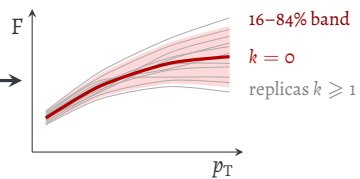
$$\langle v \rangle_k^S = \frac{\sum_{i \in S} p_{k,i}\, v_i}{\sum_{i \in S} p_{k,i}} .$$

- The loss averages the K channels $\rightarrow$ K **independent estimators** (K heads) trained in **one forward and backward pass** $\rightarrow$ band = 16th/84th percentiles over $k \geqslant 1$.

multipliers $p_{k,i}$

event	k=0	k=1	k=2	$\cdots$
1	1	2	0	$\cdots$
2	1	0	1	$\cdots$
3	1	1	3	$\cdots$





The loss, term by term

Seven heads per channel k: logits q_{12}, q_{34}, q_{13} and mean weights $\bar{w}_1, \ldots, \bar{w}_4$.

- **Classifier**, one per ratio $(a, b) \in \{(1,2), (3,4), (1,3)\}$: binary cross-entropy ℓ_{ab} on $S_{ab} = a \cup b$, **no event weights**:

$$C_{ab} = \left\langle \ell_{ab} \right\rangle_k^{S_{ab}} \quad \Longrightarrow \quad e^{q_{ab}} = \frac{n_a}{n_b} \,.$$

- **Mean weight**, one per sample $a = 1, \ldots, 4$: relative squared error:

$$R_a = \left\langle \left(\frac{\bar{w}_a - w}{\mathrm{sg}[\bar{w}_a]} \right)^2 \right\rangle_k^a \quad \Longrightarrow \quad \bar{w}_a = \mathbb{E}[w \,|\, \boldsymbol{x}, a] \,.$$

- **Positivity**, one per subtraction $(1,2), (3,4)$: penalize $r_{ab} < 1$, i.e. more MC than data:

$$P_{ab} = \gamma \left\langle \max(0, -\log r_{ab})^2 \right\rangle_k^{S_{ab}} \quad \Longrightarrow \quad r_{ab} \geqslant 1 \;\; \text{(no negative fake yield)} \,.$$

The total loss

- **Total loss**: add the terms per channel, average over the K bootstrap channels,

$$L = \frac{1}{K} \sum_{k=0}^{K-1} \Bigg[\underbrace{\sum_{(a,b)} C_{ab,k}}_{\text{ratios } n_a/n_b} + \underbrace{\sum_{a=1}^{4} R_{a,k}}_{\text{mean weights}} + \underbrace{\sum_{(1,2),\,(3,4)} P_{ab,k}}_{\text{positive subtraction}} \Bigg] ,$$

  one backward pass trains all seven heads in all K channels together.

- **After training** the heads give the fake factor, channel by channel:

$$r_{ab} = e^{q_{ab}} \frac{\bar{w}_a}{\bar{w}_b} \quad \longrightarrow \quad \mathrm{F} = r_{13} \frac{1 - 1/r_{12}}{1 - 1/r_{34}} ,$$

  channel $k = 0$ is the nominal F, channels $k \geqslant 1$ give the band.

Applying F needs loose data only

- Applying F to the real-subtracted loose sample gives the **predicted tight-fake density**:

$$P(\boldsymbol{x}) = \mathrm{F}(\boldsymbol{x})\,(D_3(\boldsymbol{x}) - D_4(\boldsymbol{x}))\,.$$

- Factor the **loose subtraction** with the same identity as before, $u - v = u(1 - v/u)$:

$$D_3 - D_4 = D_3(1 - 1/r_{34}).$$

- Insert the decomposition of F and **cancel the loose factor**:

$$P = r_{13}\,\frac{1 - 1/r_{12}}{1 - 1/r_{34}} \times D_3(1 - 1/r_{34}) = D_3\,r_{13}\,(1 - 1/r_{12})\,.$$

- The only density left is D_3, so at event level this is the **loose data** sample weighted by $w_i\,r_{13}(\boldsymbol{x}_i)\,\big(1 - 1/r_{12}(\boldsymbol{x}_i)\big)$: **no MC events are needed** (*data-driven*).

From three logs to F

- The **three logs** are put together from the heads:

$$v_{ab} = \log r_{ab} = q_{ab} + \log \bar{w}_a - \log \bar{w}_b \,.$$

- Each factor $1 - 1/r_{ab}$ of the decomposition is the **fake share**:

$$\mathrm{F} = r_{13}\,\frac{1 - 1/r_{12}}{1 - 1/r_{34}}\,, \qquad 1 - \frac{1}{r_{ab}} = 1 - e^{-v_{ab}} = -\,\mathrm{expm1}(-v_{ab})\,,$$

  so F is evaluated as

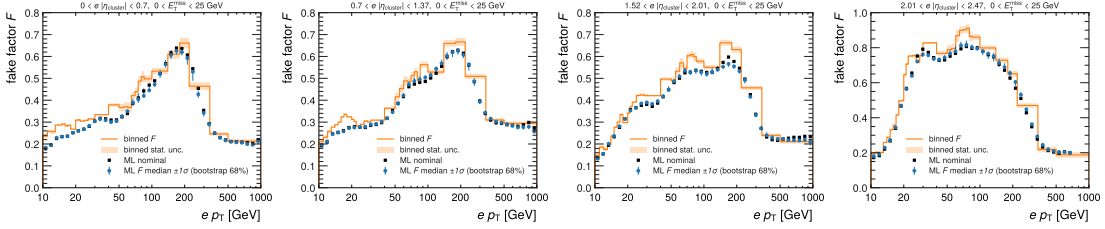
$$\mathrm{F} = e^{v_{13}} \cdot \frac{-\,\mathrm{expm1}(-v_{12})}{\max\big(-\,\mathrm{expm1}(-v_{34}),\ \varepsilon\big)}\,.$$

- Applied to loose data: **the denominator cancels**, so nothing divides by the small, noisy subtraction and ε is never needed.

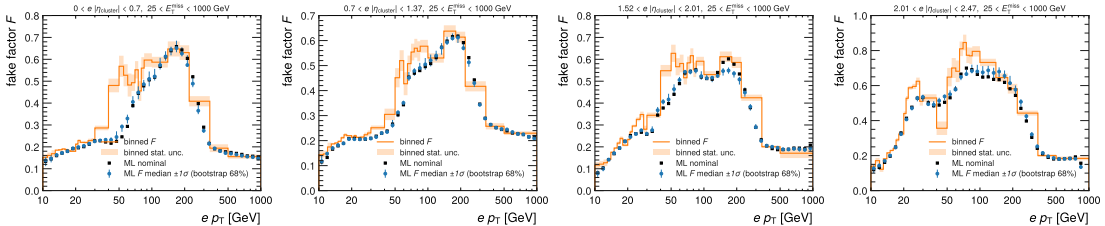
F vs p_T in slices of $|\eta|$ and E_T^{miss}

$0 < |\eta| < 0.7$ | $0.7 < |\eta| < 1.37$ | $1.52 < |\eta| < 2.01$ | $2.01 < |\eta| < 2.47$

$E_T^{miss} < 25$ GeV

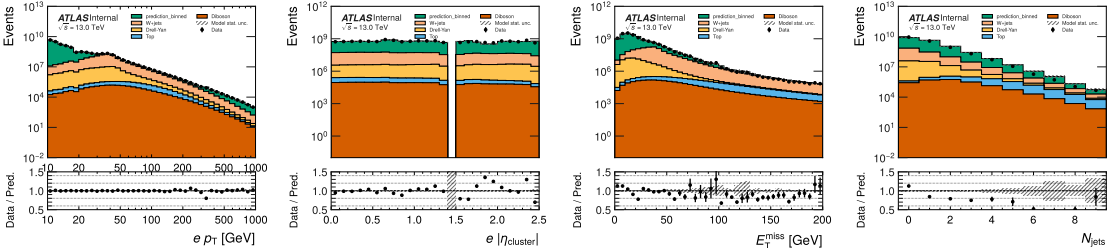


$E_T^{miss} > 25$ GeV

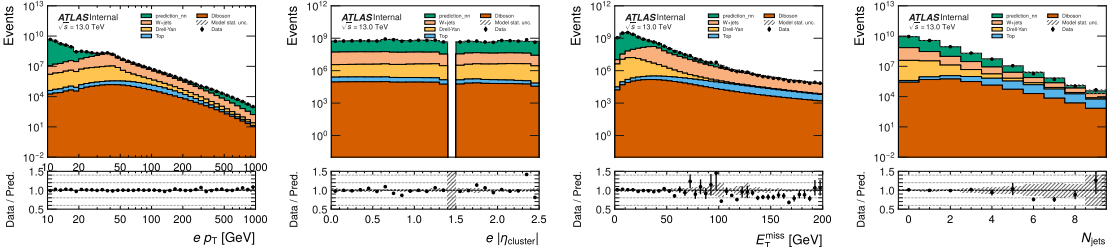


Closure: binned vs ML

Binned F



ML F



Conclusion

- Fake-factor method with a **continuous, unbinned, per-event** F, shown for electron fakes in a fake-enriched region.
- **One network, one training**: seven heads on a shared backbone give F directly, instead of three separately trained classifiers.
- Weighted densities factorize as $D = n\,\bar{w}$: classifiers learn count ratios, regressions learn mean weights $\rightarrow$ **no signed weight ever enters a classifier loss**.
- Uncertainties come for free from the vectorized **Poisson bootstrap**, in the same forward and backward pass.
- The closure is good in p_{T}, $|\eta|$, $E_{\mathrm{T}}^{\mathrm{miss}}$ and N_{jets}, and matches the binned fake factor.
- **Code:** `gitlab.cern.ch/ijs-f9-ljubljana/nnsf`.
- **Questions and feedback are very welcome.**