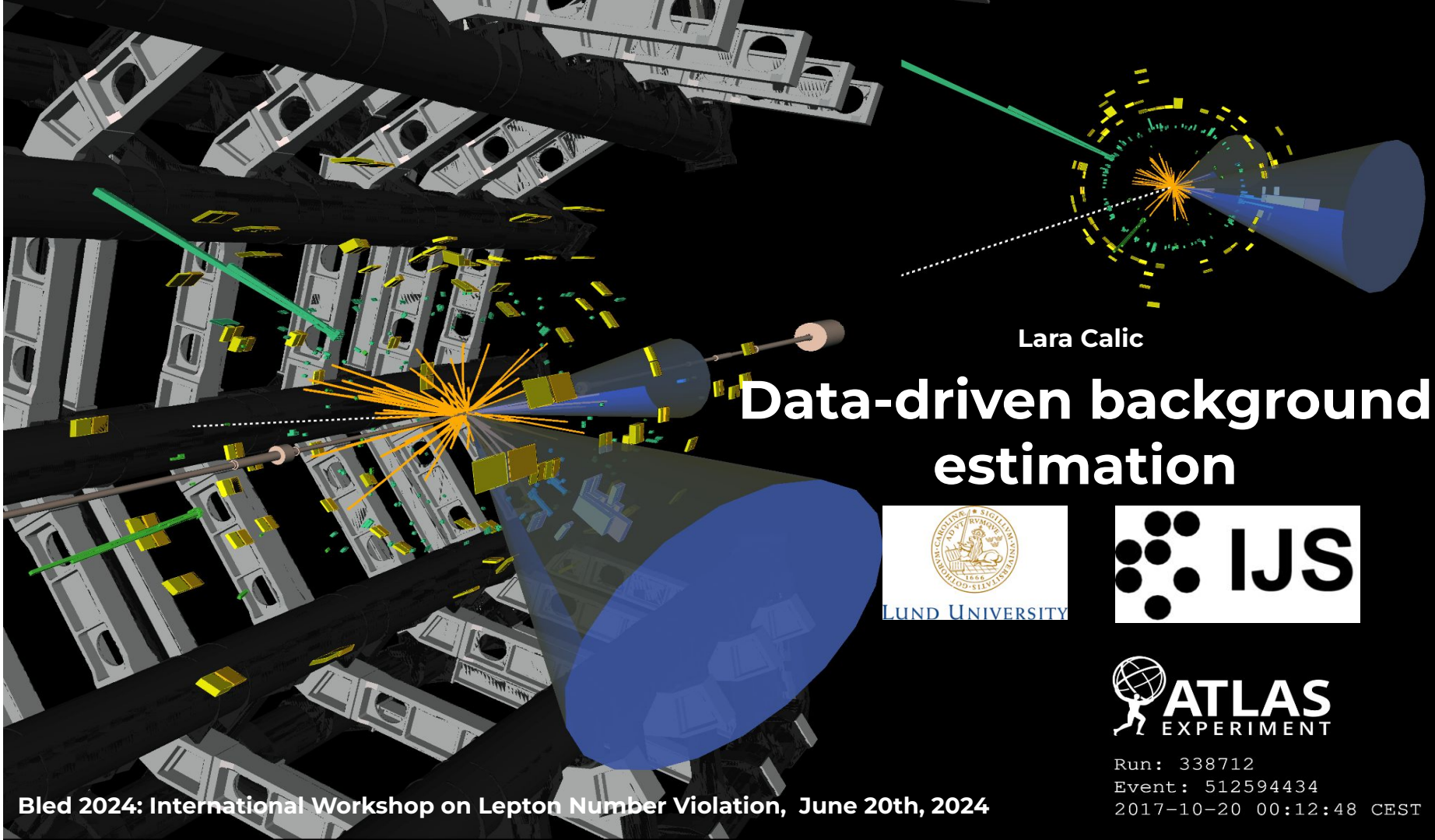


Lara Calic

Data-driven background estimation

Lund University

IJS

ATLAS EXPERIMENT

Run: 338712
Event: 512594434
2017-10-20 00:12:48 CEST

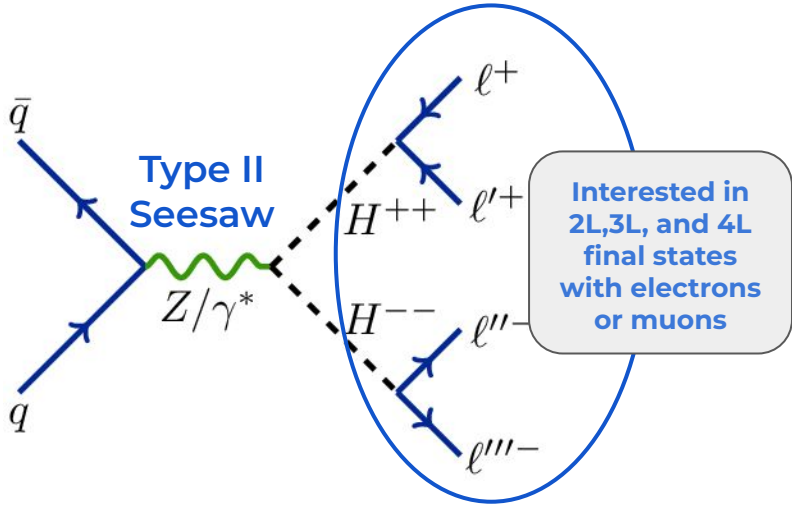
Bled 2024: International Workshop on Lepton Number Violation, June 20th, 2024

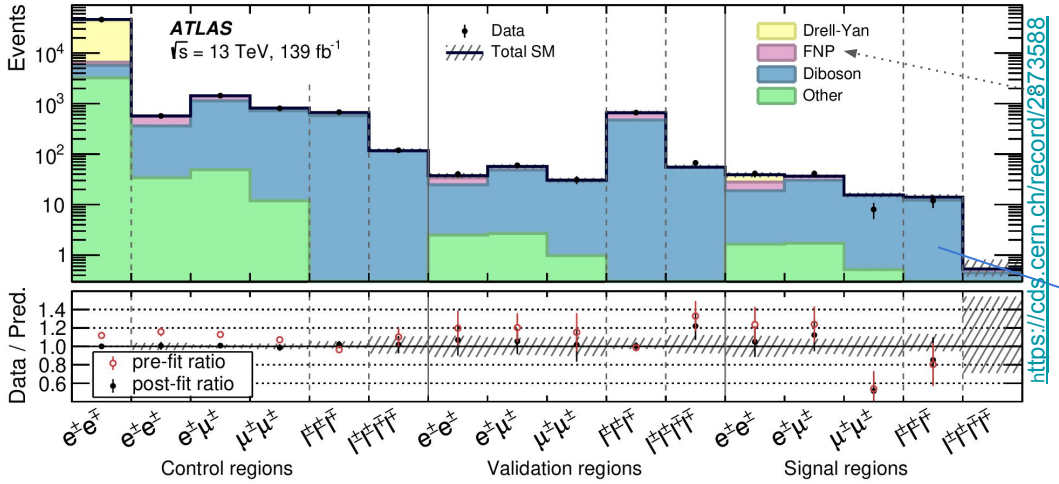
Background Estimations

An overview on the Exotics analyses

* **vacuum expectation** value of the left-right spontaneous symmetry breaking

- The dominant production mechanism of the doubly charged Higgs (DCH) boson: ***Drell-Yan mechanism***
- This search focuses on small values v_Δ*(decays into a pair of same charge leptons, irrespectively of flavour combination)
- ***Lepton Flavour Violation (LFV)*** is allowed by this model

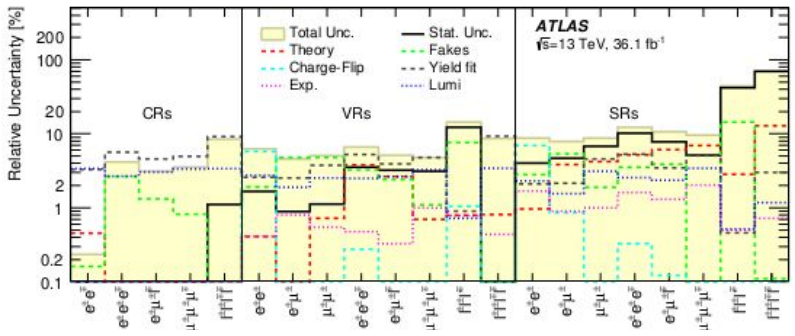




The normalisation factors of the dominant backgrounds, Drell-Yan and diboson processes, are extracted from the final binned maximum-likelihood fit of the distribution in all control and signal regions

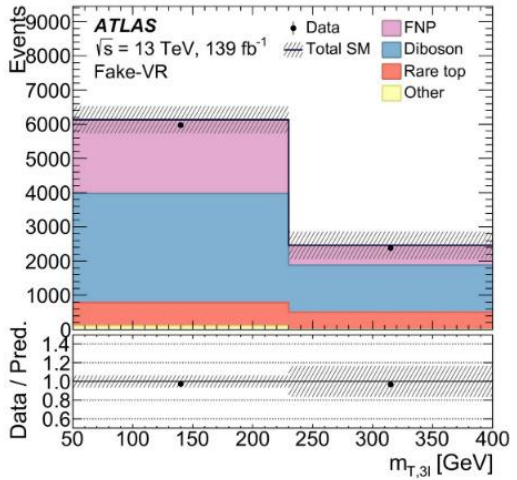
Charge - flip

- Significant source of background from misreconstructed objects (Reducible backgrounds)
- Incorrectly reconstructed jets
- Non-prompt leptons from meson decay within jets
- Electron-photon conversion
- Usually, modeled relatively poorly in Monte Carlo → strong motivation to use data-driven approaches



https://cds.cern.ch/record/2643902/files/CERN-THESIS-2018-196.pdf

FAKES



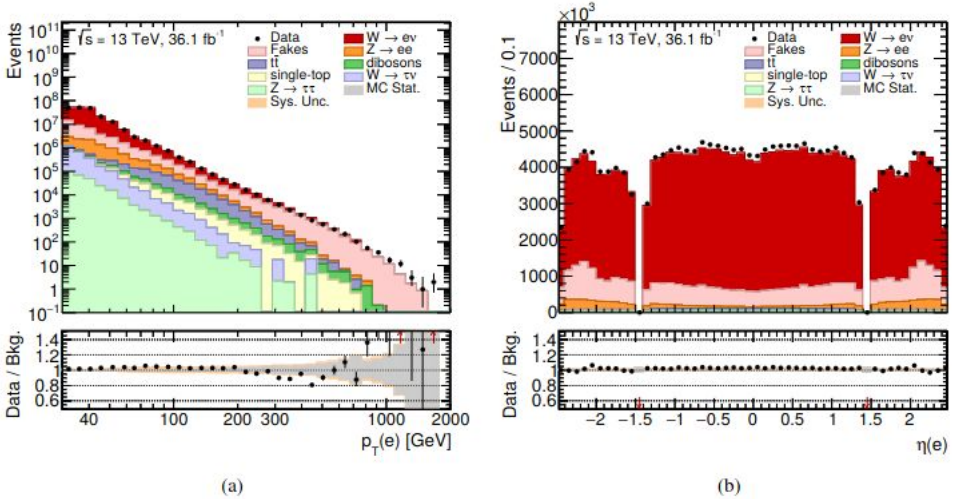
https://cds.cern.ch/record/2730768

Background prediction

- **Prompt SM backgrounds**
  - Estimated using simulated samples
  - Discard MC events with fake leptons to avoid overlap with data-driven estimations

- **Electron charge misidentification**
  - Main background in the electron channel due to **charge-flip** from bremsstrahlung
  - Stiff tracks: incorrect charge, correctly matched track (high momenta, few hits)
  - ***Estimated with MC where charge-flip probabilities are corrected to match the data***

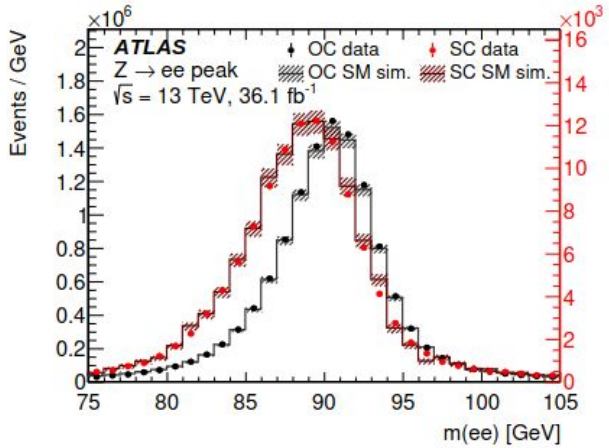
https://cds.cern.ch/record/2643902/files/CERN-THESIS-2018-196.pdf



Validation of Charge Misidentification

di-electron mass distributions

https://cds.cern.ch/record/2873588



*OC/SC - opposite/same charge pairs

- The misidentification of charge of prompt electrons (charge-flip) is responsible for the largest background in electron channel
- Three regions: the main region and two sidebands orthogonal to the main region
- Side-bands regions are used to estimate the non-Z background and subtract it from the main region

Event type	main region	side-bands
opposite-sign	$\lvert m(ee) - m_{OS}(Z)\rvert < 14$ GeV	14 GeV $< \lvert m(ee) - m_{OS}(Z)\rvert < 28$ GeV
same-sign	$\lvert m(ee) - m_{SS}(Z)\rvert < 15.8$ GeV	15.8 GeV $< \lvert m(ee) - m_{SS}(Z)\rvert < 31.6$ GeV

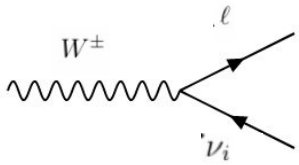
Comparison between OC/SC data (filled markers) with Monte Carlo simulation (solid line) after applying corrections for charge misidentification

Fake and non-prompt backgrounds

Background Modelling

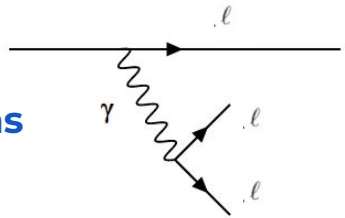
- 2 types of fake leptons:
  - **Misidentified hadrons** (wrongly classified as leptons by reconstruction algorithms)
  - **Non-Prompt leptons** (actual leptons that are not produced at the primary vertex)
    - Includes leptons from decays of long-lived hadrons (e.g., b-quarks, c-quarks) or photon conversion
    - Excludes prompt leptons produced at the primary vertex or from short-lived particles indistinguishable from it

Real leptons

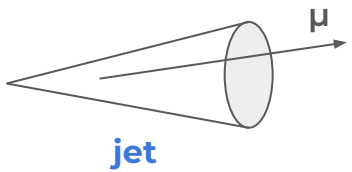


Prompt leptons coming from W and Z bosons

Fake leptons



Conversion leptons from photon radiations



Heavy/light flavour fakes from jets

Matrix and Fake Method

Matrix Method

- The core equation that maps the number of events with a real or fake lepton onto the number of events with a tight or loose lepton

a) Matrix Method (MM) expression for one lepton:

$$\begin{pmatrix} N^t \\ N^l \end{pmatrix} = \begin{pmatrix} r & f \\ 1-r & 1-f \end{pmatrix} \begin{pmatrix} N_r \\ N_f \end{pmatrix}.$$

N_t – N_{tight}, N_l – N_{loose}

r – efficiency of real leptons

N_r – N_{real}, N_f – N_{fake}

N^t_f – N^{tight}_{fake}

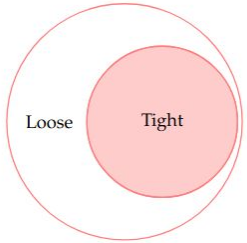
b) Solving for N_f and N_r :

$$\frac{1}{f-r}\begin{pmatrix} f-1 & f \\ 1-r & -r \end{pmatrix} \begin{pmatrix} N^t \\ N^l \end{pmatrix} = \begin{pmatrix} N_r \\ N_f \end{pmatrix}$$

c) Expression for Fake events in Tight Selection:

$$N^t_f = f N_f = \frac{f}{f-r}((1-r)N^t - rN^l).$$

f - fake lepton efficiency



Loose selection (orthogonal to the nominal selection)
Tight selection (same as nominal selection)

It can be generalized for Multiple Leptons as:

$$N^{j_1,j_2,\ldots,j_M} = \sum_{i_1\in\{r,f\}} \sum_{i_2\in\{r,f\}} \cdots \sum_{i_M\in\{r,f\}} \left(\prod_{k=1}^{M} \epsilon^{j_k}_{i_k}\right) N_{i_1,i_2,\ldots,i_M}, \quad \epsilon^{j_k}_{i_k} = \begin{cases} r_k, & \text{if } j_k = t, i_k = r \\ f_k, & \text{if } j_k = t, i_k = f \\ (1-r_k), & \text{if } j_k = l, i_k = r \\ (1-f_k), & \text{if } j_k = l, i_k = f \end{cases}$$

Fake Factor Method

- If we start with the inversion of the matrix :

$$\begin{pmatrix} N_{RR} \\ N_{RF} \\ N_{FR} \\ N_{FF} \end{pmatrix} = \frac{1}{(r-f)^2} \begin{pmatrix} (1-f)^2 & (f-1)f & f(f-1) & f^2 \\ (f-1)(1-r) & (1-f)r & f(1-r) & -rf \\ (r-1)(1-f) & (1-r)f & r(1-f) & -rf \\ (1-r)^2 & (r-1)r & r(r-1) & r^2 \end{pmatrix} \begin{pmatrix} N_{TT} \\ N_{TL'} \\ N_{L'T} \\ N_{L'L'} \end{pmatrix}$$

* connects the number of real and fake objects to the number of tight and loose objects

R = REAL LEPTONS, F = FAKE LEPTONS

- Approximation for the matrix (real rate r=1, and parameters A, B, C are given by this expression):

Total contribution of fake leptons to the nominal selection:

$$N_{TT}^{\text{fakes}} = AN_{TT} + B(N_{TL'} + N_{L'T}) + CN_{L'L'}$$

$$A = \alpha[2rf(f-1)(1-r) + f^2(1-r)^2],\ B = \alpha(1-f)fr^2, \text{ and } C = -\alpha r^2 f^2$$

Contamination of the fakes:

$$N_{TT}^{\text{fakes}} = (F(N_{TL'} + N_{L'T}) - F^2 N_{L'L'})_{\text{data}} - (F(N_{TL'} + N_{L'T}) - F^2 N_{L'L'})_{N_{RR} \text{ from MC}}$$

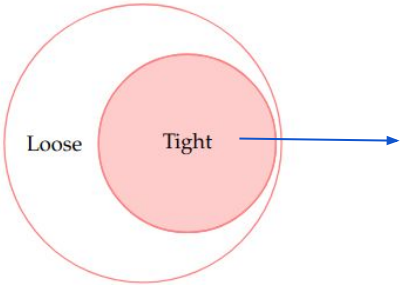
FAKE FACTOR

$$F = \frac{f}{1-f}$$

This expression in commonly used in our analysis, and shows that Fake Factor is not a constant (it should be evaluated for each lepton separately)

* f - fake rate, T, L' - tight/loose objects

Fake Factor method



Loose region: events with loose selection criteria
Tight region: events with tight selection criteria

$$F = \frac{f}{1-f} = \frac{N^{\text{fake}}_{\text{tight}}}{N^{\text{fake}}_{\text{loose-tight}}} = \frac{N^{\text{fake}}_{\text{num}}}{N^{\text{fake}}_{\text{den}}}$$

Fake Factor (FF): measured by observing the ratio of objects in the numerator to those in the denominator within a fake-enriched sample

Focused on this formula

$$N^{\text{dilepton}}_{\text{fakes}} = \left[\sum_{TL} F_2 + \sum_{LT} F_1 - \sum_{LL} F_1 F_2\right]_{\text{data}} - \left[\sum_{TL} F_2 + \sum_{LT} F_1 - \sum_{LL} F_1 F_2\right]_{\text{prompt simulation}}$$

https://cds.cern.ch/record/2730768

What is the advantage of using FF method?

Eliminates the need to obtain the real rate directly from data, instead using prompt-only Monte Carlo for evaluation

Fake rate:

$$f = \frac{N_{\text{fake, pass}}}{N_{\text{fake, total}}}$$

$N_{\text{fake (real),pass}}$ - number of fake (real) leptons passing the selection criteria

Real rate:

$$r = \frac{N_{\text{real, pass}}}{N_{\text{real, total}}}$$

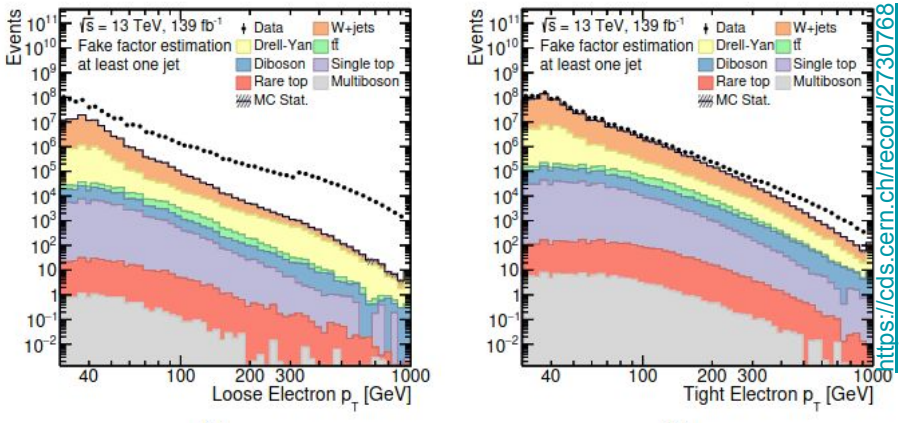
$N_{\text{fake (real), total}}$ - total number of (fake) real leptons

https://www.sciencedirect.com/science/article/abs/pii/S0168900223003662

Electron Fake Factor Measurements

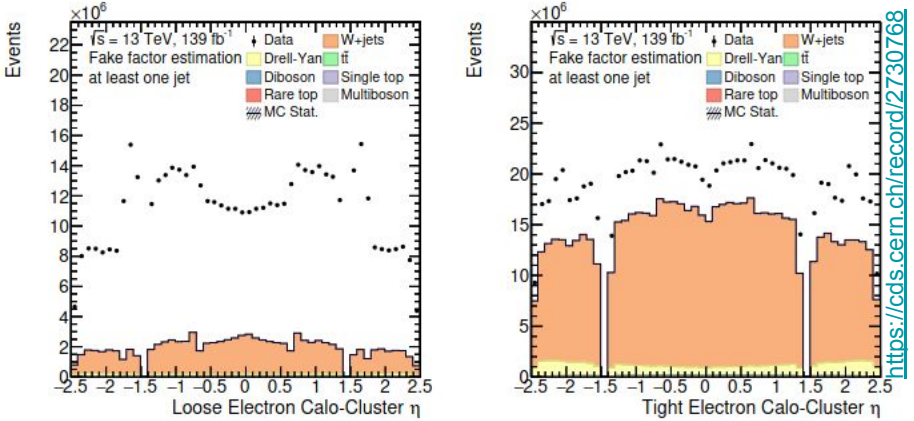
Loose and Tight electrons in the fake-enriched region

Distribution of p_T for loose and tight electrons



- **Loose distribution:** Shows a broader distribution (indicating a mix of true and fake electrons)
- **Tight distribution:** Shows a narrower distribution, indicating a higher likelihood of true electrons.

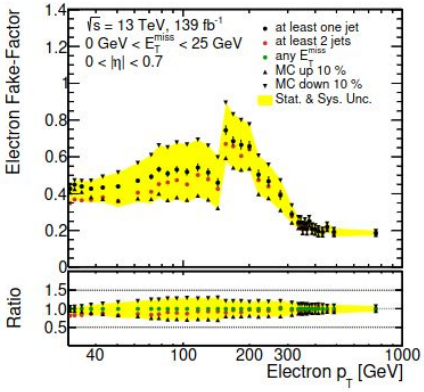
Distribution of calorimeter cluster η for loose and tight electrons



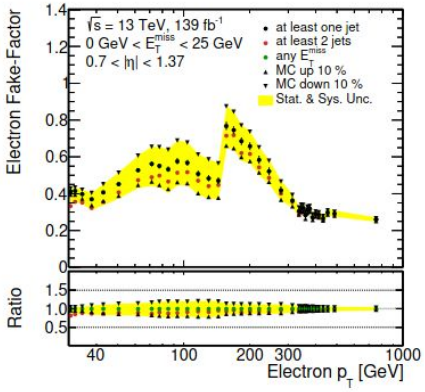
- **Loose distribution:** Shows how electrons that pass loose criteria are distributed in the detector, often with wider coverage (more fake electrons)
- **Tight distribution:** Shows a more concentrated distribution, indicating regions where tight criteria are effective in identifying true electrons

p_T dependence of Fake Factor on the variations of the bins

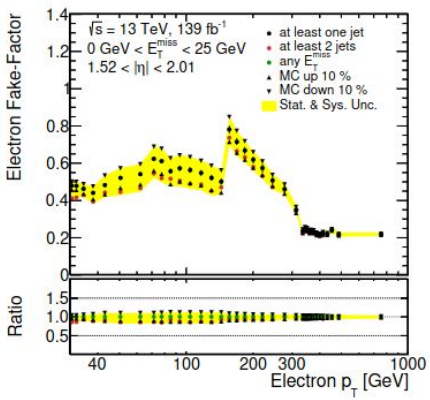
1st bin



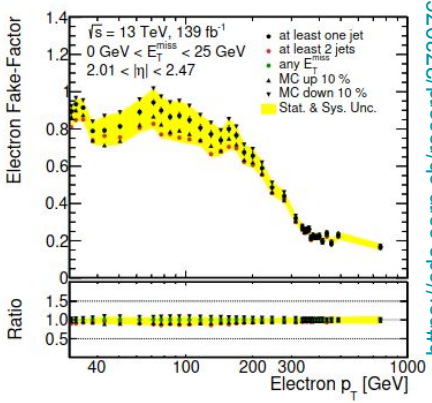
2nd bin



3rd bin



the last bin



https://cds.cern.ch/record/2730768

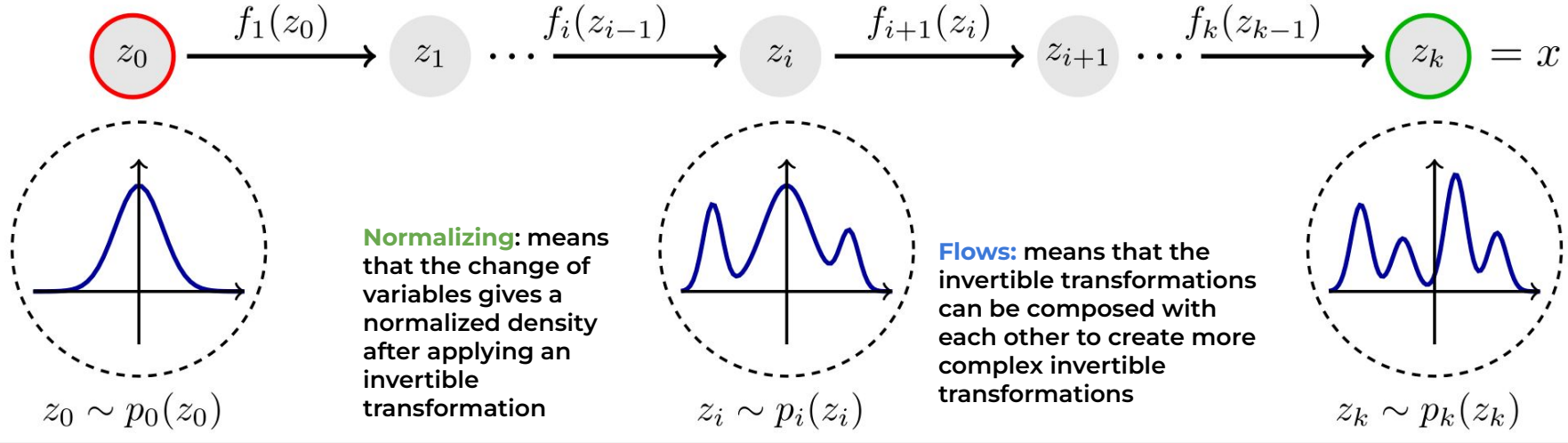
- Each plot represents a **different systematic variation** applied to the fake factor estimation
- The yellow shaded areas represent the systematic and statistical uncertainties
- The black dots with error bars indicate the measured fake factor for each p_T bin

Impact of the systematic variations

Variation	Purpose
Flipped requirement on the number of jets	fake composition
Removing E_T^{miss} requirement	fake composition
MC scaled up by 10%	MC modelling, cross-section and luminosity
MC scaled down by 10%	MC modelling, cross-section and luminosity

Machine Learning Solution

Simplify idea of Normalising flows (NF)



Normalizing: means that the change of variables gives a normalized density after applying an invertible transformation

Flows: means that the invertible transformations can be composed with each other to create more complex invertible transformations

Base distribution

Target distribution

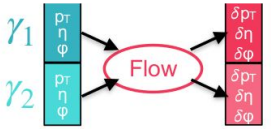
Normalising flow direction

https://github.com/janosh/tikz/blob/main/assets/normalizing-flow/

Flow - chart on our idea

*MC sample (background, ttbar)
* Fake Factor calculation
* Closure test

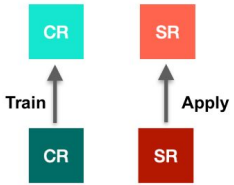
Test and validate a fake factor method on MC samples



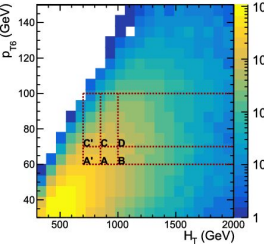
Train your model on data

* Probability Density Function
* Training on Normalising Flows

New method that can be widely used in all ATLAS physics analyses



Test and validate a method on synthetic data



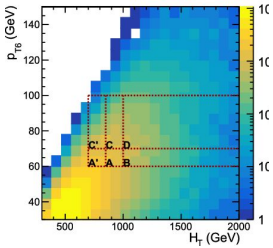
*DataToys sample
*Gaussian distribution (with fakes and reals)
* Closure test

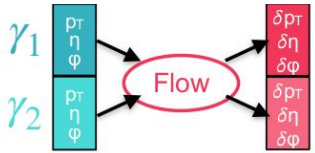
bbbbackground estimation[Roguljic, Hartman]

Gain with the Machine Learning solution

Replace the 2D p_T, eta dependent fake factors with a multi-dimensional ML-based solution

Increase the statistics of the anti-Tight regions with generative models to reduce the statistical uncertainty in fake background estimations





Machine Learning solution

- **Base Distribution**: Start with a simple base distribution (in our case Gaussian).
- **Transformations:** Apply a series of invertible transformations to map this base distribution to the complex target distribution observed in the real data.
- **Density Estimation:** Use the transformed distribution to estimate the probability density of observed events, enabling precise background modeling.

This is what we want to test it

Training process: Train the normalizing flow on both Monte Carlo (MC) simulations and real data

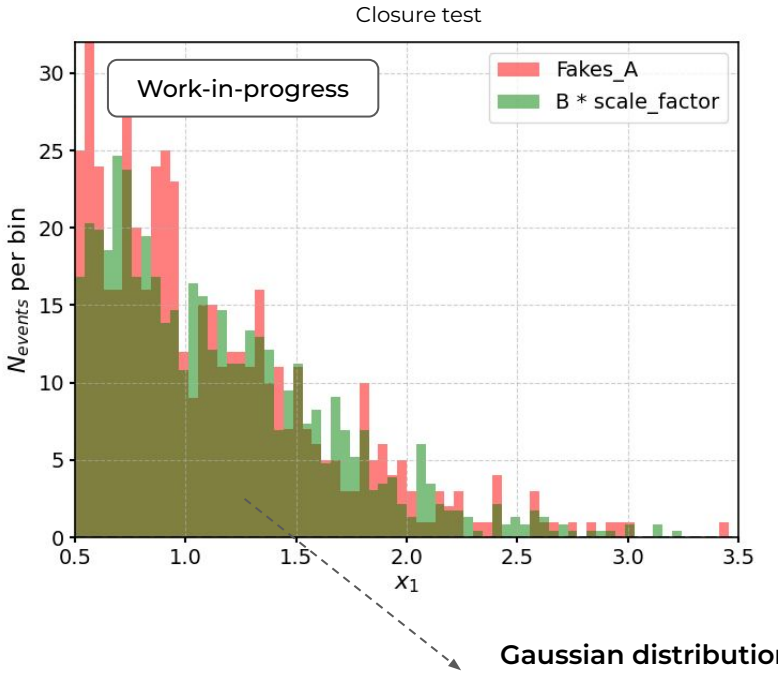
Apply the **trained normalizing flow** model to **estimate the background** (with the systematic variations)

Use the trained model to **generate samples with more variables** (e.g., p_T, η, E_T^{miss}) for a complex analysis

How are we going to deal with negative MC weights

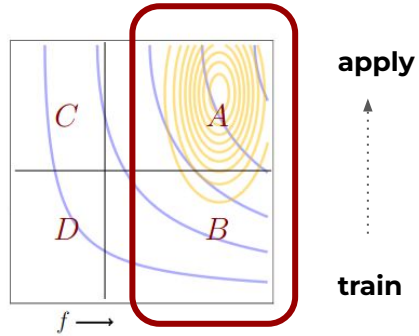
DataToySample

* T - tight objects
!T - anti-tight objects



- Performing a closer test in fakes-enriched region using the same lepton classification
- X_1 - feature (variable) of choice
- **Scale factor** - ratio between number of fakes in region C/D
- Background in the signal region (CR) can be predicted from the C,D,B regions (VR regions)

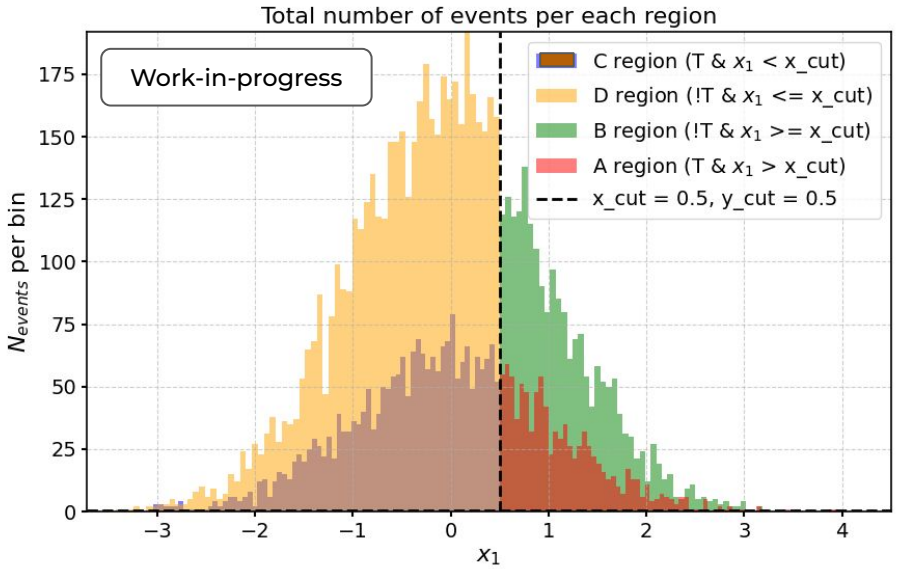
$$N_A = \frac{N_B N_C}{N_D},$$



Gaussian distributions where we made certain assumptions (including 50% fakes and 50% reals)

DataToySample

* T - tight objects
!T - anti-tight objects



Chosen 2 independent variables (X_1,X_2)

Applied lepton classification criteria (tight and loose objects)

Validation of fake factor method: number of data points falling into each of the four ABCD regions based on the given thresholds

Additional simple tests of our method: verify the calculations by comparing the estimated number of fake events in region A with the actual number of fake events

Conclusion

Accurate background estimation is crucial for making precise measurements in HEP

Data-driven technique that gives us more insight on systematic uncertainties

Fake Factor method enhanced by machine learning are key tools in achieving accurate background estimates

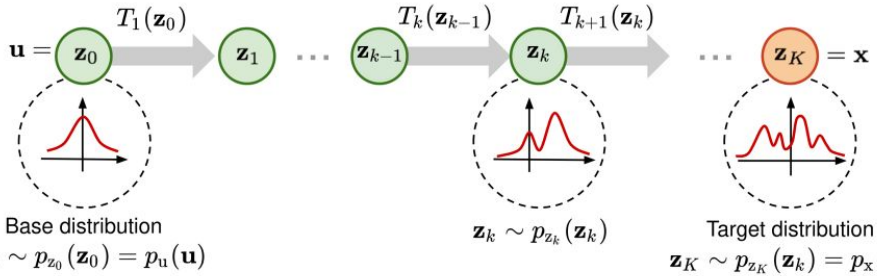
BACKUP SLIDES

NF key concepts

Forward and inverse directions

- **Forward direction:** $z_k = T_k(z_{k-1})$ for $k = 1, \ldots, K$ with $z_0 = u$ (infer)
- **Inverse direction:** $z_{k-1} = T_k^{-1}(z_k)$ for $k = K, \ldots, 1$ with $z_K = x$ (train)
- The log-determinant of a flow is

$$\log|\det J_T(z_0)| = \log\left|\prod_{k=1}^{K} \det J_{T_k}(z_{k-1})\right| = \sum_{k=1}^{K} \log|\det J_{T_k}(z_{k-1})|$$



- Similar to autoencoder: forward mode ⇔ decoder, backward mode ⇔ encoder

More details in paper [J. Gavranovic, B. Kersevan]

In Normalising Flow model mapping between Z and X is given by:

$$f_\theta : \mathbb{R}^n \to \mathbb{R}^n$$

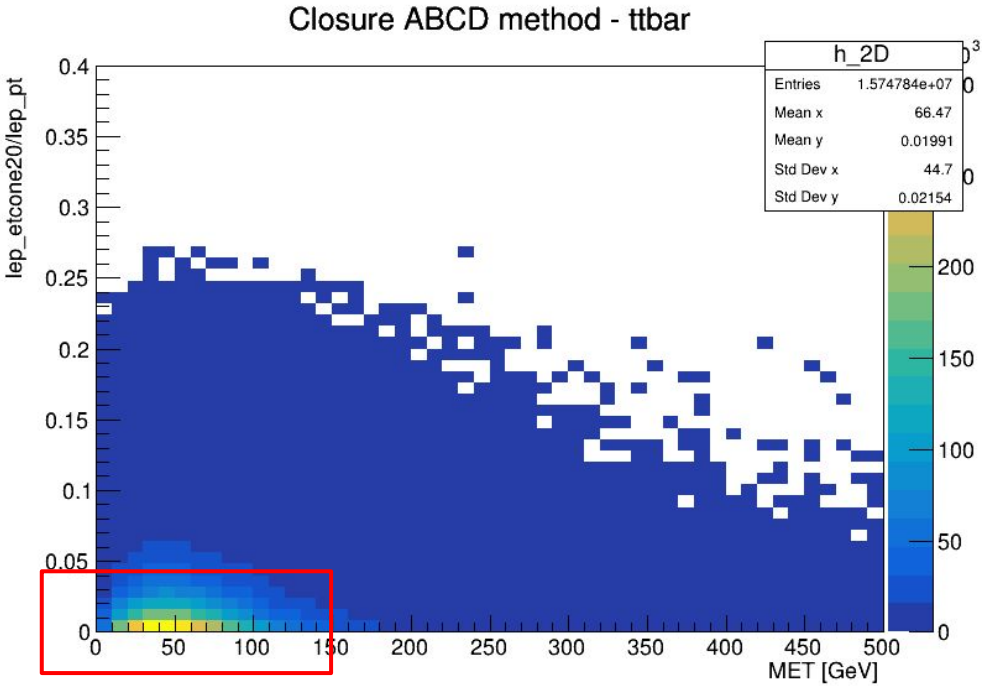
It is deterministic and invertible such that:

$$X = f_\theta(Z)$$
$$Z = f_\theta^{-1}(X)$$

Using the change of variable, the likelihood is given by:

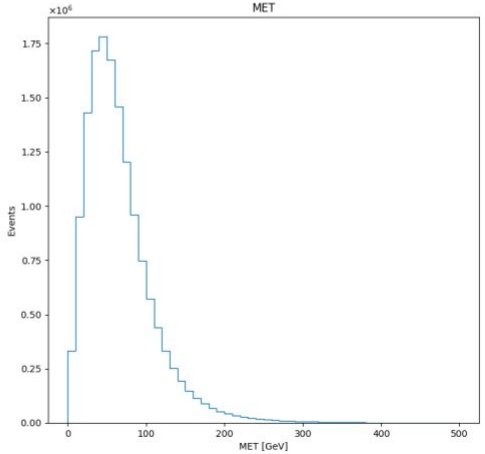
$$p_X(x;\theta) = p_Z(f_\theta^{-1}(x))\left|\det\left(\frac{\partial f_\theta^{-1}(x)}{\partial x}\right)\right|$$

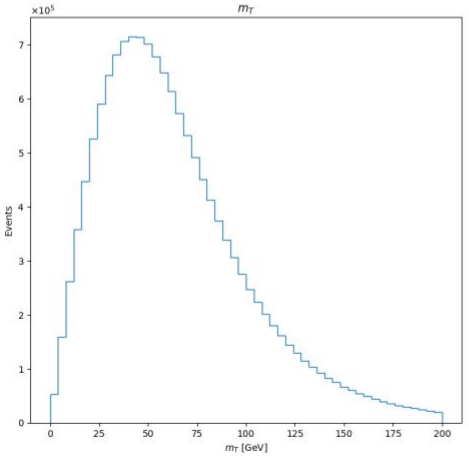
Closure test on MC sample



Samples: ATLAS Open Data

MC Sample: ttbar - electron identification



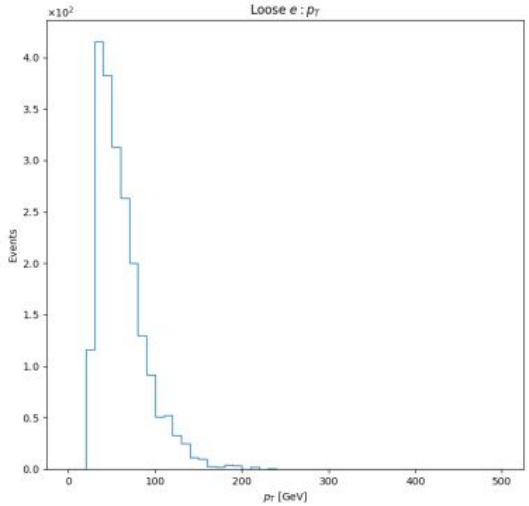


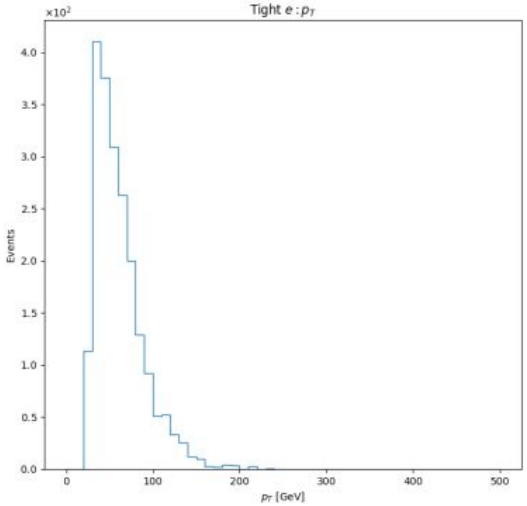
`mc_410000.ttbar_lep.1lep.root`

Samples: ATLAS Open Data

MC Sample: ttbar - electron identification

pT distribution (Loose vs Tight selection)

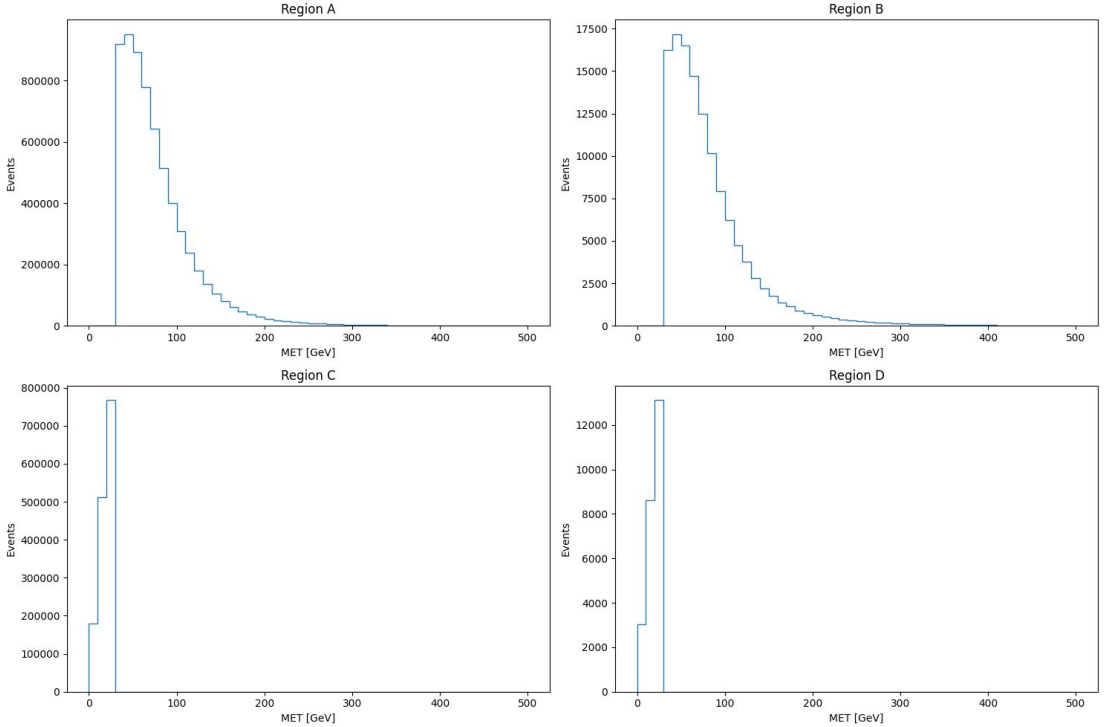




`mc_410000.ttbar_lep.1lep.root`

Samples: ATLAS Open Data

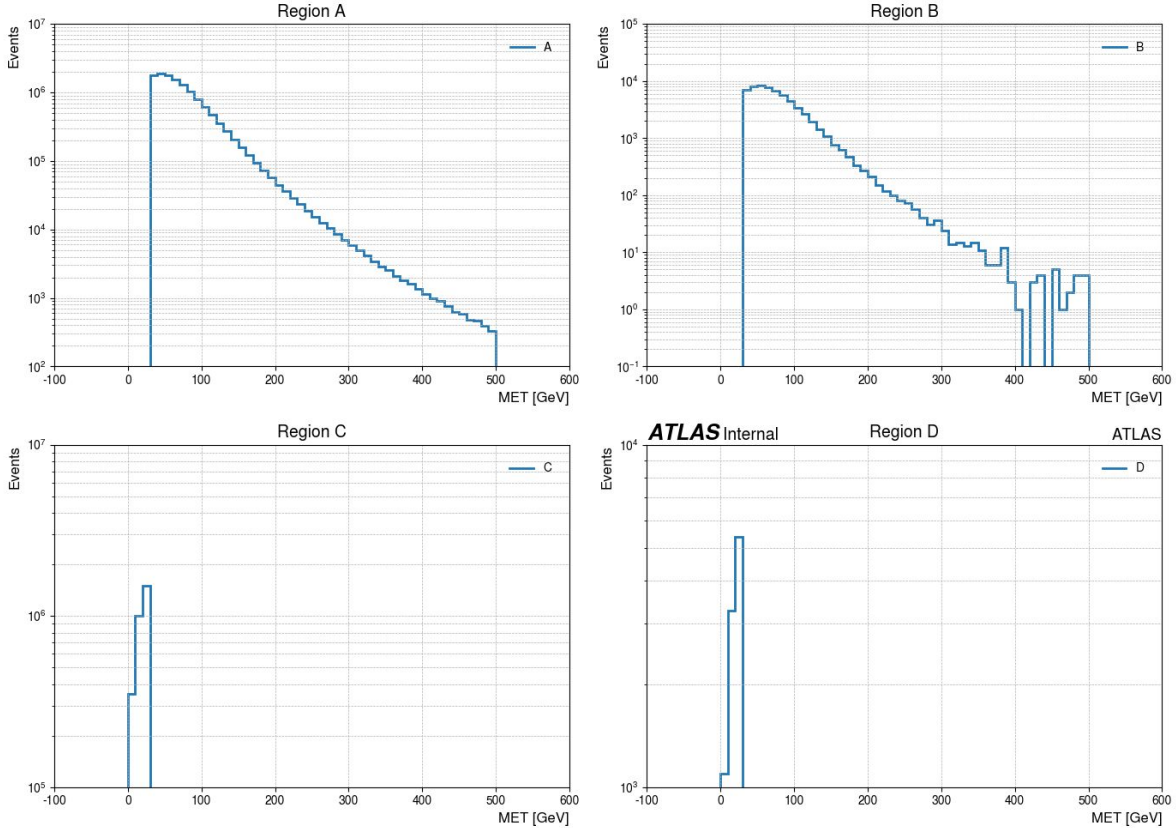
MC Sample: ttbar



Samples: ATLAS Open Data

`mc_410000.ttbar_lep.1lep.root`

MET distribution in Control and Validation regions

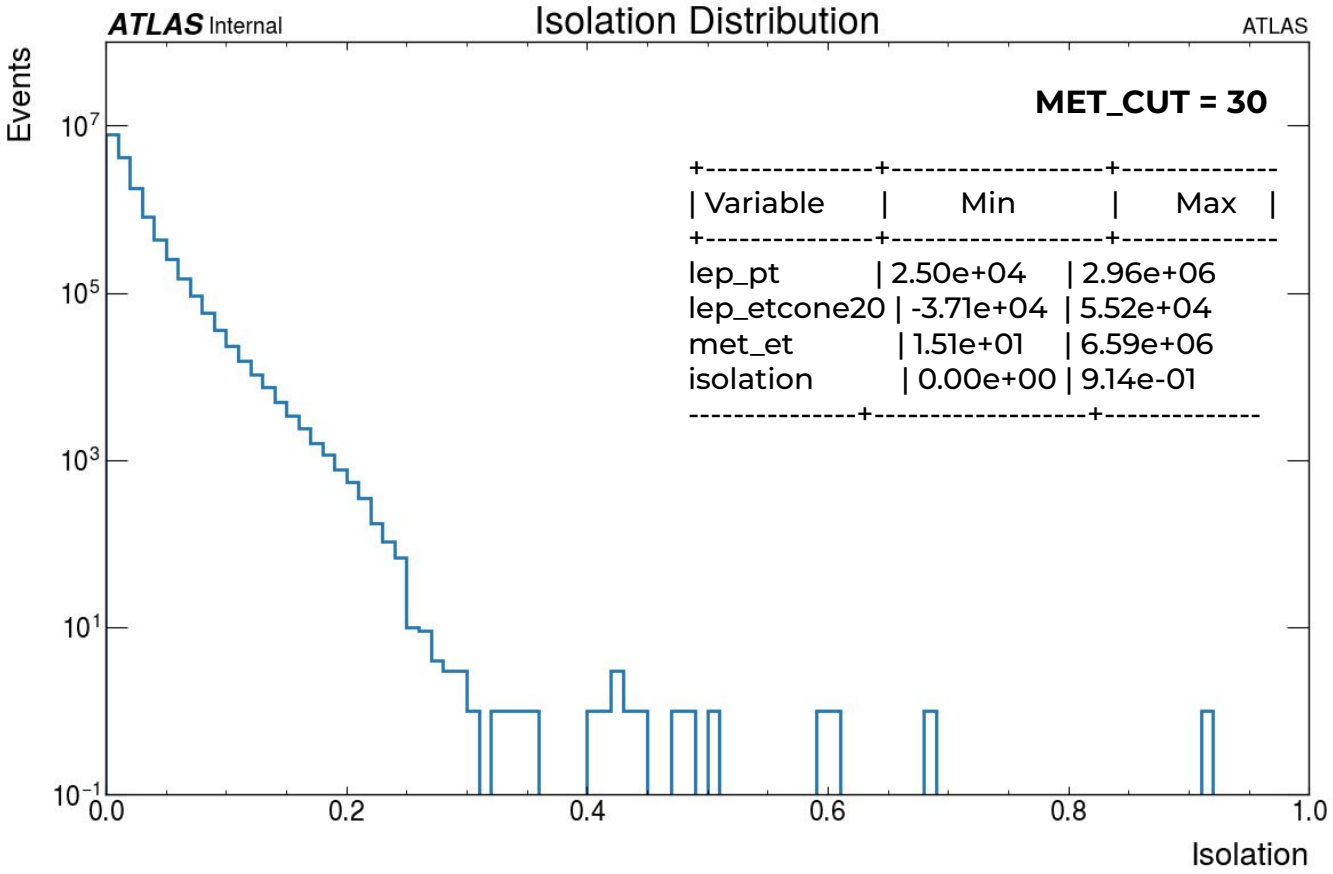


*Log scale

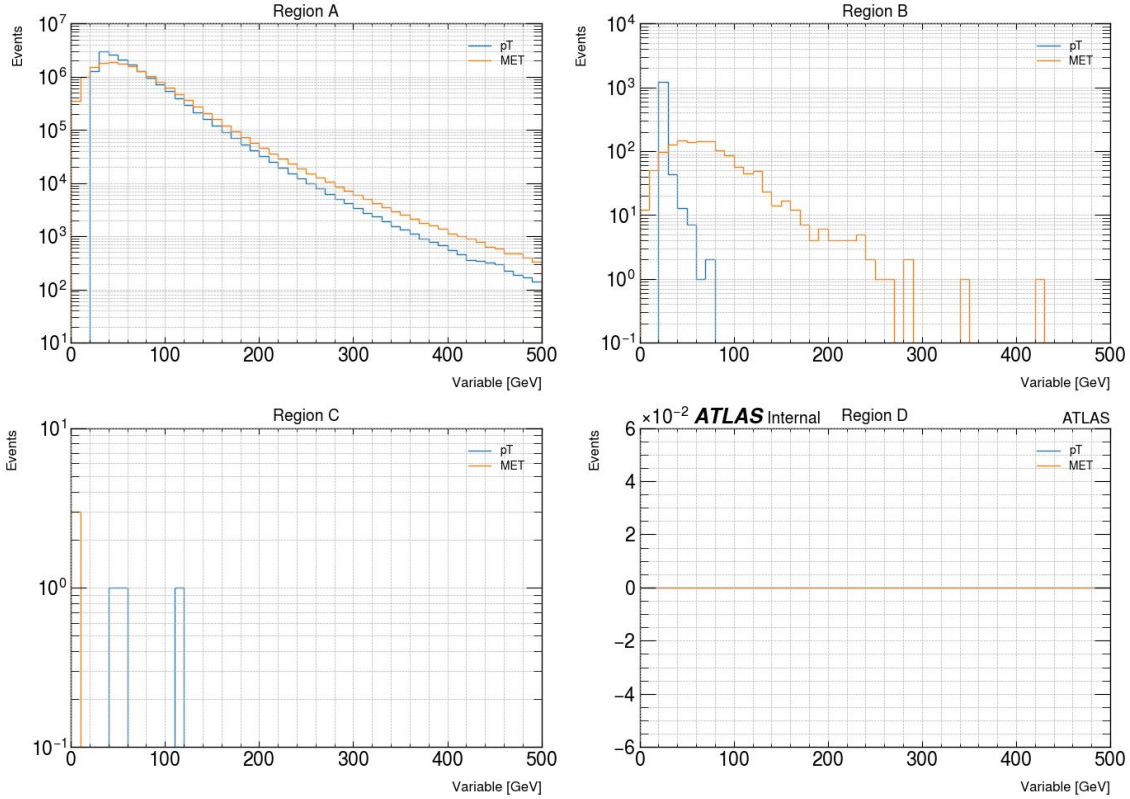
Samples: ATLAS Open Data

`mc_410000.ttbar_lep.1lep.root`

mc_410000.ttbar_lep.1lep.root



P_T, MET distribution in Control and Validation regions

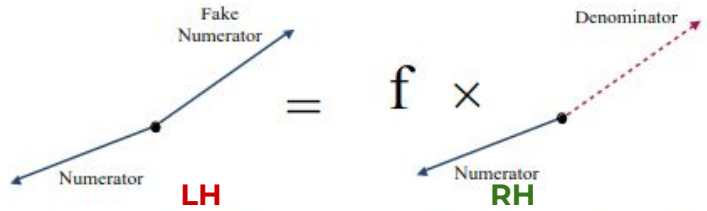


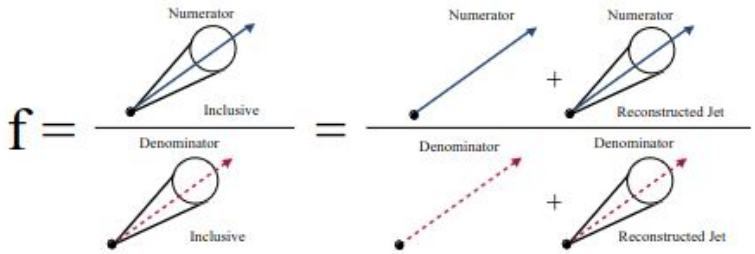
Samples: ATLAS Open Data

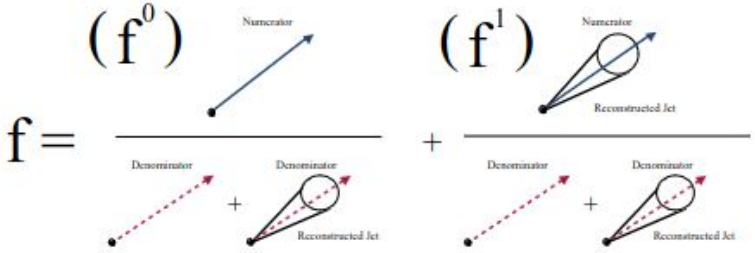
`mc_410000.ttbar_lep.1lep.root`

Comparison of Fake Factor components

Jet veto application







Left hand-side = represents the background (SR)

Right hand-side = represents the fake factor modeling of the background: W +jet (CR)

f_0 - extrapolates to fake background (there is no overlapping jets)

f_1 - extrapolates to fake background (but includes overlapping jets)

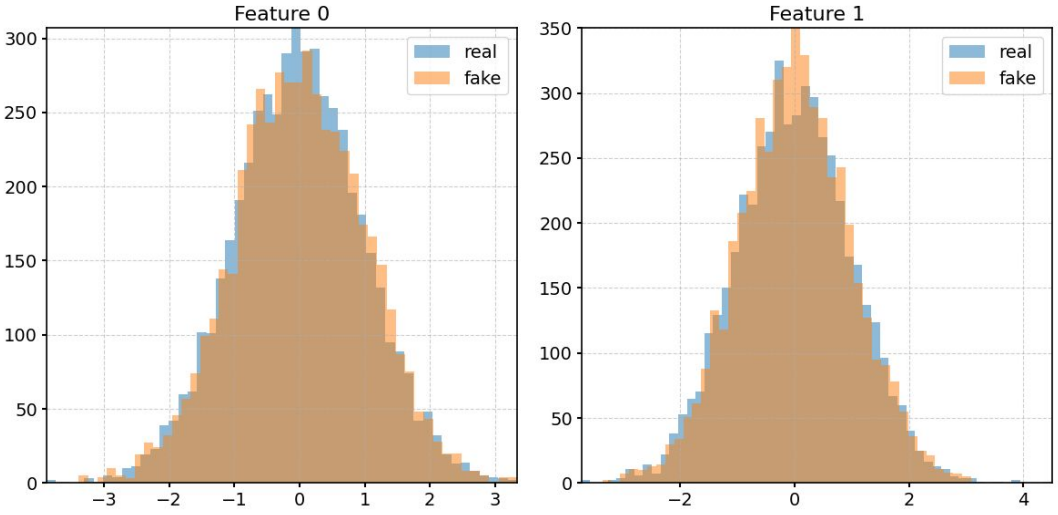
Key relation that ties the:

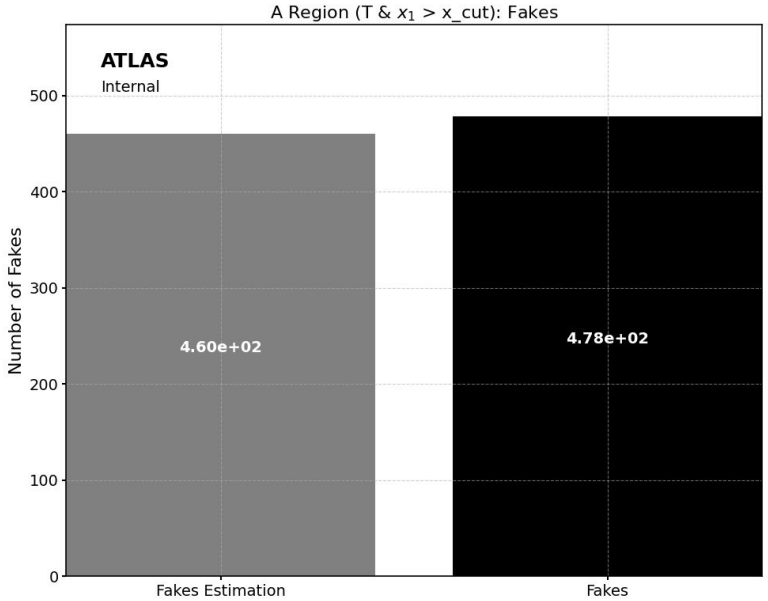
$$N_f^t = f N_f$$

fake factor rate between signal and control region

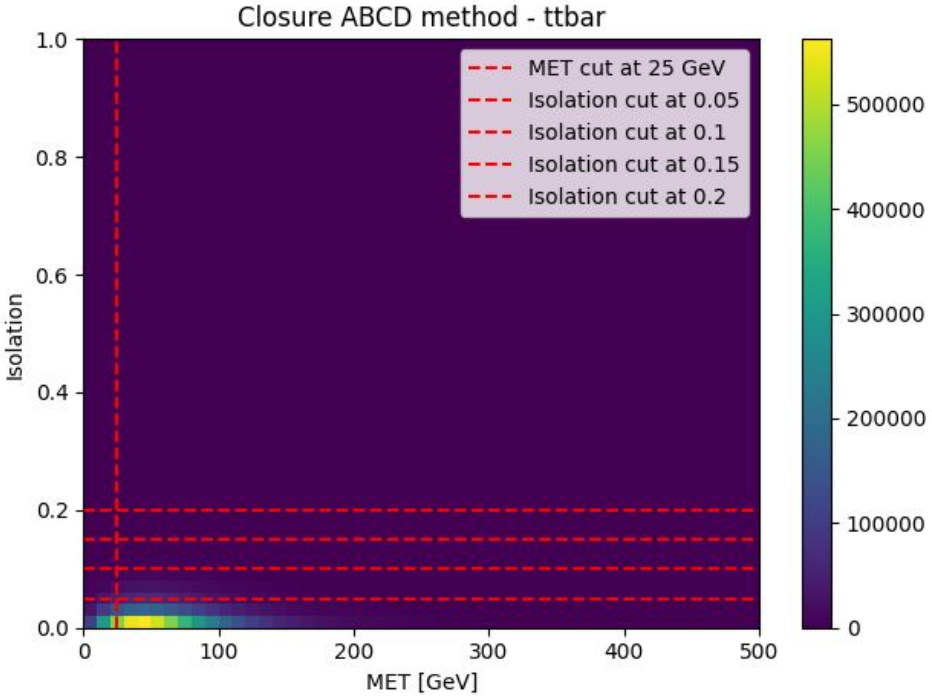
*numerator = events in the signal region
*denominator = events in the control region

DataToySample

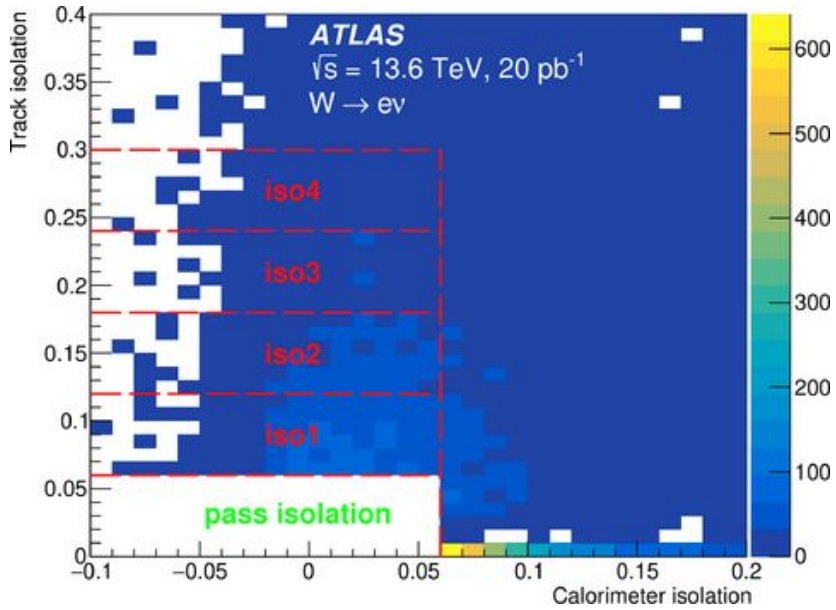




”Slicing approach”



Multijet background



Similar approach already exists

Outline

1. Background estimations in Exotics analysis
2. Significance of the statistics
3. Fake Factor method
4. Machine Learning solution
5. Summary and importance of Fake Factor (FF) method

References

[1] https://www.sciencedirect.com/science/article/abs/pii/S0168900223003662

[2] https://arxiv.org/pdf/2007.14400

[3] https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2018-34/

[4] https://cds.cern.ch/record/2805679/files/CERN-THESIS-2021-306.pdf

[5] https://cds.cern.ch/record/2873588/files/CERN-THESIS-2023-183.pdf

[6] https://cds.cern.ch/record/2730768

[7] https://cds.cern.ch/record/2873588?ln=en

[8] https://arxiv.org/abs/2003.13913

[9] https://www.physics.mcgill.ca/xhep/en/resources/thesis/2023_Beier_MSc_Atlas_Fakes.pdf

8] https://arxiv.org/abs/2003.13913

[9]https://www.physics.mcgill.ca/xhep/en/resources/thesis/2023_Beier_MSc_Atlas_Fakes.pdf

[10] https://atlas-glance.cern.ch/atlas/analysis/papers/details?ref_code=EXOT-2018-34

[11]https://atlas-glance.cern.ch/atlas/analysis/papers/details?ref_code=EXOT-2020-02

[12] https://inspirehep.net/literature/2770237

[13] https://inspirehep.net/literature/2710627

[14]https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2020-28/