Summer Student Report

Background composition measurement in the photon-jet resonance search with the ATLAS experiment

Fran Cassinese Parisi[1]

Supervisor: Francisco Alonso

August 2023

Abstract

In this Summer Student project, a data-driven technique known as ABCD method is implemented on 2018 proton-proton collision data from the ATLAS Detector of the LHC. The method uses two orthogonal variables, defined through photon identification and isolation criteria. The purity of prompt photons with respect to fake photons stemming from misidentified high-energy jets is estimated with this method, as well as its dependence on the photon candidate transverse momentum and pseudorapidity. Monte Carlo simulated signal events are also used for correction factors known as signal leakage fraction, whose dependence on the same kinematic variables is also studied.

1. Introduction

1.1 Prompt Photons as a Background for New Physics

High energy proton-proton collisions resulting in a **photon and a jet with large transverse momenta** provide a powerful way of searching for new physics. Of particular interest is the search for **exotic high-mass photon-jet resonances**. There are many proposed new physics processes capable of producing a bump over the continuum SM background, including production of excited quarks and quantum black holes.

The main background is the **prompt photon production**, either through **Compton scattering of a quark and a gluon** (Figure 1) or through **quark-antiquark annihilation** (Figure 2). Additionally events with a photon and a jet can arise from multi-jet production (fragmentation).

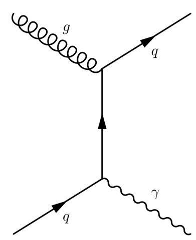

Figure 1: *Quark and gluon Compton scattering*

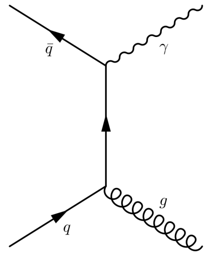

Figure 2: *Quark-antiquark annihilation*

[1]fcassine@cern.ch

Fake photons

Some objects can leave an electromagnetic trace indistinguishable from prompt photons. When jets dominantly fluctuate to a π^0 that carries most of the momentum and later decay into two collimated photons, they can be misidentified as photons[2]. These photon candidate objects stemming from multi-jet events are therefore called **fake photons**, and constitute the main background for prompt photon production.

Since Monte Carlo simulations are not reliable for these fake photon objects[6], a **data-driven** method is implemented for estimating the contribution of fake photons to the prompt photon background. Simulated MC samples are still used, but only for prompt photons (signal), i.e. the method used does not rely on fake photon samples from MC simulations.

1.2 The ABCD Method[5][6][7]

The ABCD method is a data-driven technique that consists of applying cuts in two variables to divide the data set into four regions. Region A is expected to contain mostly **signal** events, while regions B, C and D will be enriched in background events, and are taken as **control** regions. The two variables used for this method will be variables that are already being utilised in photon identification.

Applying this method, it is possible to compute the number of signal events in region A, N_A^{sig}, as:

$$N_A^{sig} = N_A - R^{bg} \cdot (N_B - f_B \cdot N_A^{sig}) \cdot \frac{N_C - f_C N_A^{sig}}{N_D - f_D N_A^{sig}} \tag{1}$$

Where N_A, N_B,..., are the total number of events in each of the four defined regions. R^{bg} is defined as:

$$R^{bg} = \frac{N_A^{bg} \cdot N_D^{bg}}{N_B^{bg} \cdot N_C^{bg}}, \tag{2}$$

with N_K^{bg} being the number of background events in each region. The only assumption of the ABCD method is that the two variables are **uncorrelated for background events**. This equates to setting $R^{bg} = 1$.

The factors f_B, f_C and f_D are called **signal leakage fractions**, and they are introduced to take into account the possibility of having some of the signal events in the control regions. They are defined for region K as:

$$f_K = \frac{N_K^{sig,MC}}{N_A^{sig,MC}}, \tag{3}$$

where the MC superscript indicates that these fractions are computed using the Monte Carlo samples.

Photon Identification

The first variable is a binary identification criteria. In order to reduce hadronic background, some requirements are applied variables describing the electromagnetic showers using information from the electromagnetic and hadronic calorimeter. This involves the application of cuts in the distributions of the so-called **Shower Shapes** (Figure 3), and the photon candidates satisfying the criteria are known as **tight photons**.

s. For photons not meeting the tight criteria, several **loose** regions are defined:

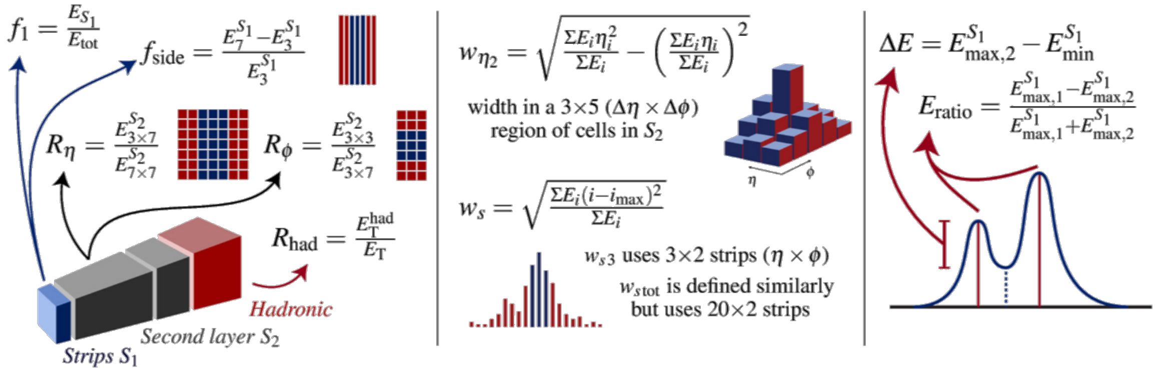

Figure 3: *Definitions of some of the standard Shower Shape variables used in Photon Identification*

- *loose' 1*: Dropping cuts on F_{side} from Tight, but failing Tight
- *loose' 2*: Dropping cuts on F_{side} and w_{s3} from Tight, but failing Tight
- *loose' 3*: Dropping cuts on F_{side}, w_{s3} and ΔE from Tight, but failing Tight
- *loose' 4*: Dropping cuts on F_{side}, w_{s3}, ΔE and E_{ratio} from Tight, but failing Tight
- *loose' 5*: Dropping cuts on F_{side}, w_{s3}, ΔE, E_{ratio} and $w_{s,tot}$ from Tight, but failing Tight

The current analysis is done using *loose' 4* as the region for **non-tight** photons.

Photon Isolation

The first variable chosen for the ABCD method is an **isolation variable**. While prompt photons are only correlated to activity from the primary interaction vertex, jets faking photons are expected to have some hadronic or electromagnetic activity around, whose energy is deposited in the calorimeters.

The absence of this surrounding energy depositions is called **isolation**[3], and it can be quantified by establishing a **cone** of radius R in (η, ϕ) space, centered in the photon cluster barycenter and adding the calorimetric energy inside. The choice of $R = 0.4$ leads to E_T^{cone40}.

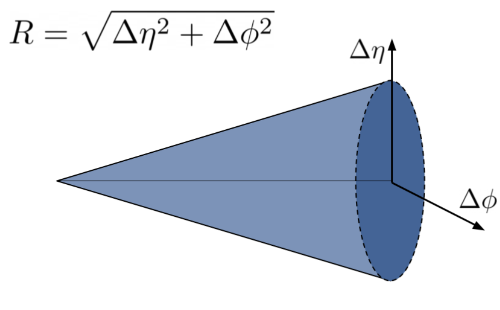

Figure 4: *Isolation cone definition*

2. Data and MC Samples

For this project, the data is taken from pp collisions at $\sqrt{s} = 13\,\text{TeV}$ from 2018, and the simulated signal is taken from the MC20e campaign (SHERPA).

2.1 Event Selection

A preselection was performed on both real and simulated data to keep only events with:

- At least one photon $n > 0$
- Leading photon momentum $p_T > 150\,\text{GeV}$
- Leading photon pseudorapidiy $|\eta| < 2.37$
- Events and outside of the barrel-endcap crack: $|\eta| < 1.37$ or $|\eta| > 1.52$

2.2 Binning

Since tight criteria is defined differently for **unconverted** and **converted** photons (i.e., photons which transform to an e^-e^+ pair at least once inside the detector), they are treated separately[2]. To study the background composition dependence on p_T and η, 10 p_T bins and 4 η bins were established. The binning strategy is:

$$p_T \,:\, [150, 175, 200, 250, 300, 350, 400, 500, 600, 800, 1000]$$
$$|\eta| \,:\, [0, 0.6, 1.2, 1.37]\ [1.52, 2.37]$$

2.3 Isolation distributions

The distribution of isolation variable E_T^{cone40}, as well as a p_T-independent alternative were studied by making a histogram in each p_T-η-conv bin. The range chosen for E_T^{cone40} was $[-10, 60]$, and 100 bins with the same width were established.

Figure 5 shows an example for the first p_T and $|\eta|$ bin, and for unconverted photons. The data events were divided according to the ID variable, i.e. tight and non-tight (NT) events. For the Monte Carlo events, only tight samples are shown. An additional histogram shows the addition of the non-tight data and the tight simulated data.

As can be seen from Figure 5, both real and simulated events show a peak around $E_T^{cone40} = 0$ for their tight-passing events. Non tight data shows a peak at $E_T^{cone40} \sim 10GeV$, and the fact that the low-values tail indicates the presence of background in the signal region.

Data normalisation

Since the isolation and identification variables are assumed to be uncorrelated for background events, there should not be a distinction between tight and non-tight distributions for non-isolated (that is, high E_T^{cone40}) events. Therefore, non-tight data is **normalised** to the tight data such as to make the tails match. For this project, the choice was to multiply the non-tight histogram such that the integral for $E_T^{cone40} > 10\,GeV$ matched the tight histogram. This was done separately in each p_T-η-conv bin. However, this procedure might be improved by having a **dynamical lower limit** for the tail integral, since the peaks of the tight and non-tight distributions change with p_T and η.

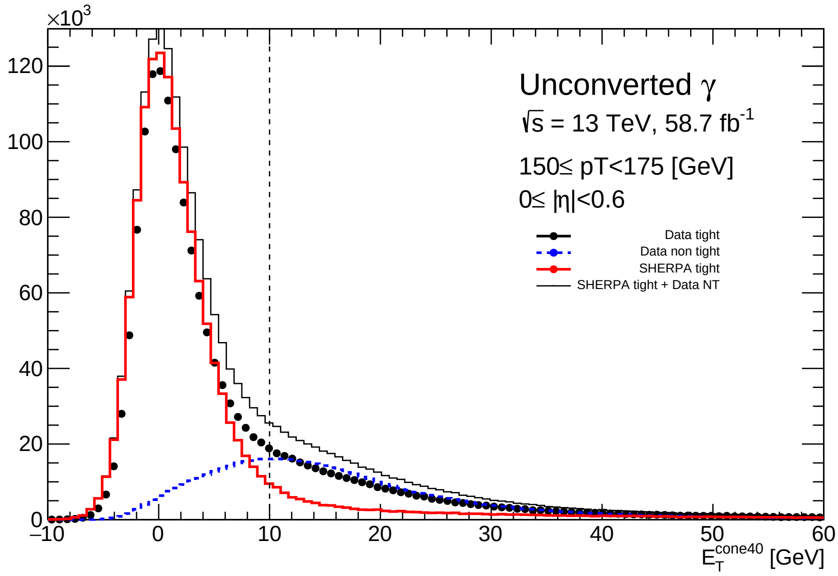

Figure 5: *Isolation variable E_T^{cone40} distributions. Data samples were divided into tight and non-tight distributions, and the normalization procedure in the text was applied. Only tight distribution is shown for MC samples. A final histogram of tight simulated events + non-tight data events is shown.*

3. Analysis and Results

3.1 ABCD regions

Taking the photon ID and isolation variables defined previously, the A, B, C and D regions are defined as:

- **Region A** (Signal): tight photons, $E_T^{cone40} < a \times p_T + b$
- **Region B** (Control): tight photons, $E_T^{cone40} > a \times p_T + b + GAP$
- **Region C** (Control): non-tight photons, $E_T^{cone40} < a \times p_T + b$
- **Region D** (Control): non-tight photons, $E_T^{cone40} > a \times p_T + b + GAP$

Regions B and D are defined with a "gap" from region A to have well separated background-control and signal regions and minimise migrations across the borders [6]. Parameters a, b, GAP were initially set to 0.0042, $4.8\,GeV$ and $2\,GeV$ respectively. The same technique was later repeated with parameters set to 0.0022, $2.45\,GeV$ and $4.5\,GeV$.

An example of the distributions corresponding to each region is shown for the data samples in Figure 6. The total tight and non-tight distributions are also shown. As can be seen, the use of a gap generates a clear separation with no migration between isolated and non-isolated regions, even when the isolation cut depends on p_T. The non-tight regions C and D are normalised with the same factor introduced in the previous section. Region B shows an excess in events for $E_T^{cone40} < 10$ with respect to region D.

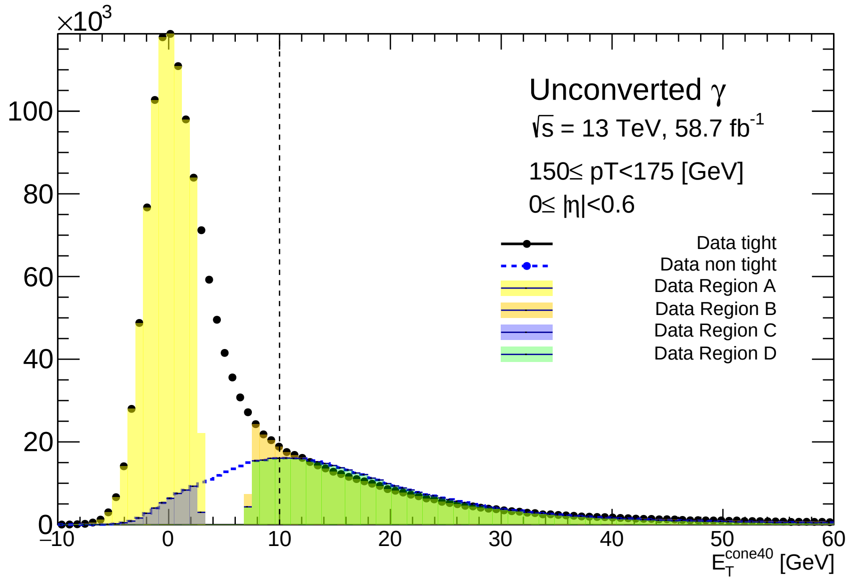

Figure 6: *Isolation variable E_T^{cone40} distributions. Data samples were divided into tight and non-tight distributions, and the normalisation procedure in the text was applied.*

A 2D histogram showing both variables simultaneously is also presented in Figure 7, this time using $E_T^{cone40} - 0.0022 \times p_T$ to show the precise boundaries of each region.

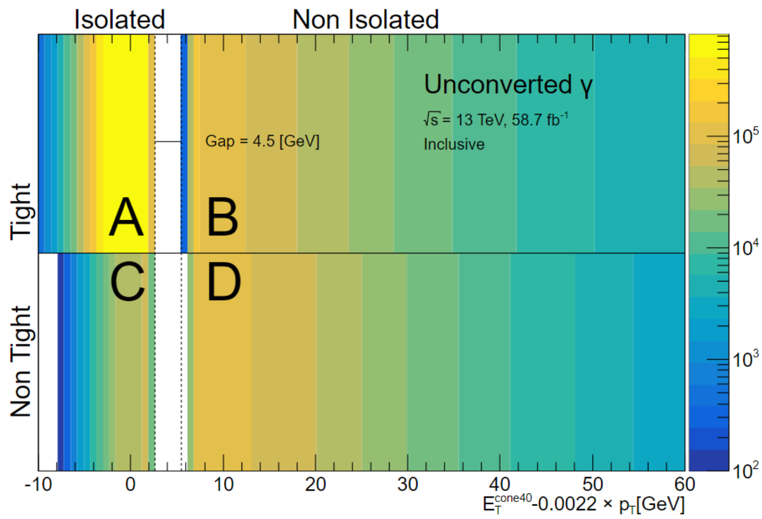

Figure 7: *2D (inclusive) histogram showing both variables involved in ABCD method and the four regions.*

3.2 Signal Leakage Fractions

By dividing the MC signal samples in the same regions defined above, the signal leakage fractions are computed. Statistical uncertainties are included by first using sum of weights squared on the histograms, computing $N_A^{sig,MC}$ and $N_K^{sig,MC}$ through integrals with errors, and propagating the uncertainties for f_K.

Figure 8 shows the signal leakage fractions computed, as a function of p_T and in different η bins. Signal contamination in region B appears to be slightly decreasing with p_T for some η bins, and constant for others, it being always the region with the highest signal contamination. Regions C and D can be decreasing or constant with p_T, again depending on the η bins. As it is expected, region D is always the less contaminated with signal events, since it corresponds to non-tight and non-isolated events.

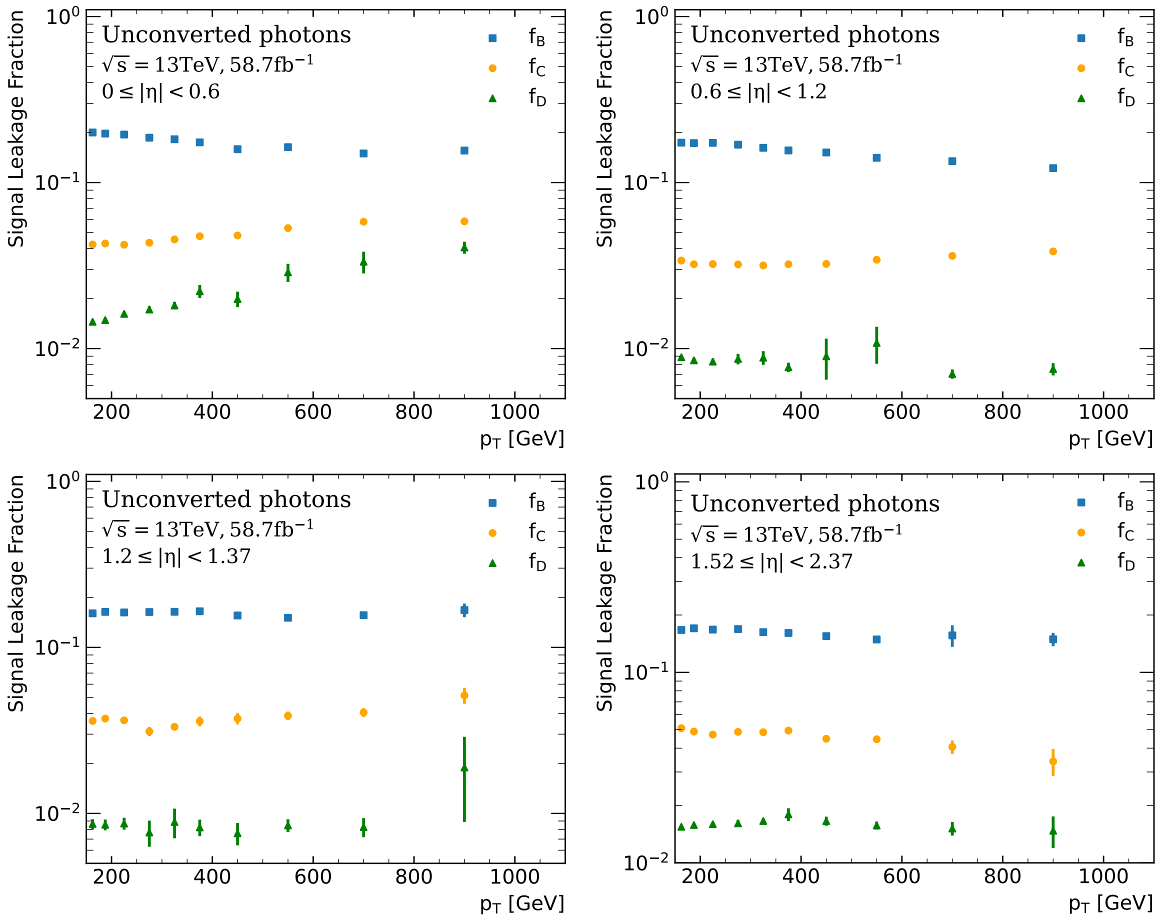

Figure 8: *Signal leakage fractions computed with MC samples, shown as functions of p_T and in different η bins.*

3.3 Signal purity

Purity is defined as the fraction of signal events to the total events in region A, N_A^{sig}/N_A, and was first computed by neglecting the signal leakage fractions, which reduces equation (1) to the simpler form:

$$N_A^{sig} = N_A - N_B \frac{N_C}{N_D} \quad (4)$$

Figure 9 shows the variation of purity with p_T and η. Statistical uncertainty is included for the purity, which was obtained through propagation of the statistical errors regarding the histograms and their integrals.

Purity was later computed using the complete equation (1). Figure 10 shows such purities.

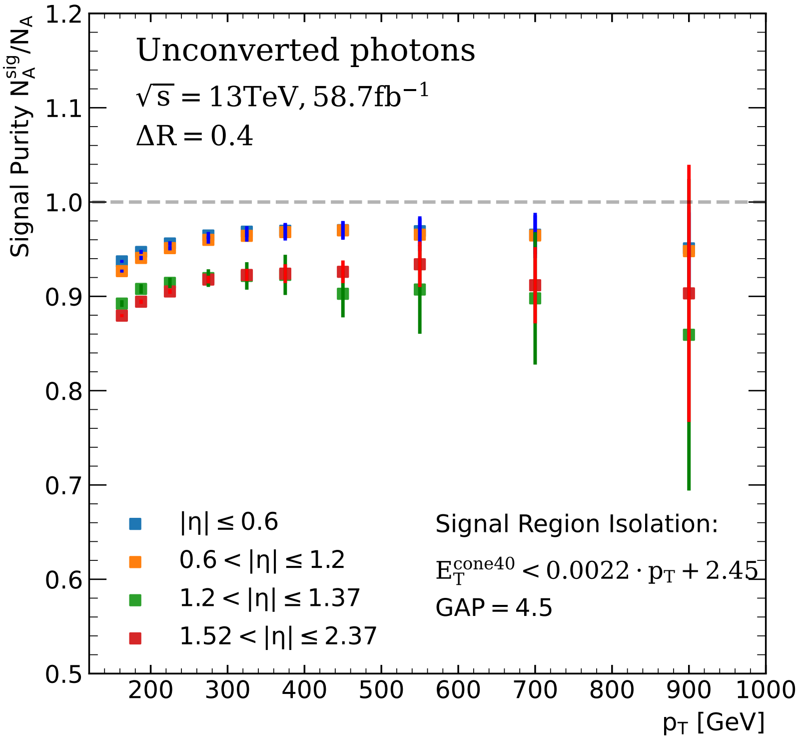

Figure 9: *Signal purity computed by neglecting the signal leakage fractions.*

Purities obtained from both equations show the same increasing behaviour from low p_T to about $\sim 300\,GeV$. After this value the purity starts decreasing. Purities computed without signal leakage show lower values than those computed with them, but uncertainties in the latter method are bigger. In particular, events in the $|\eta|$: $[1.2, 1.37]$ bin have abnormally big uncertainties. Some of the purities computed using the signal leakage fractions had values bigger than one, which is not mathematically prohibited in the method. However, for a correct interpretation of the purities, all values bigger than unity were set to one[6].

4. Possible future lines of work

Firstly, **systematic errors** should be taken into account, which can be done by varying the E_T^{cone40} limit for the region definition as well as repeating the analysis for different *loose'* regions for non-tight [2]. Furthermore, the analysis should be repeated for *converted photons*, and a comparison ought to be done between purities in converted and unconverted photons. Lastly, the assumption of an uncorrelated background should be tested, which can be done by further subdividing the ABCD regions [5].

As an exploration of a new technique, the ABCD method could be replaced with a **binned template fit** [7] on the isolation variable distribution.

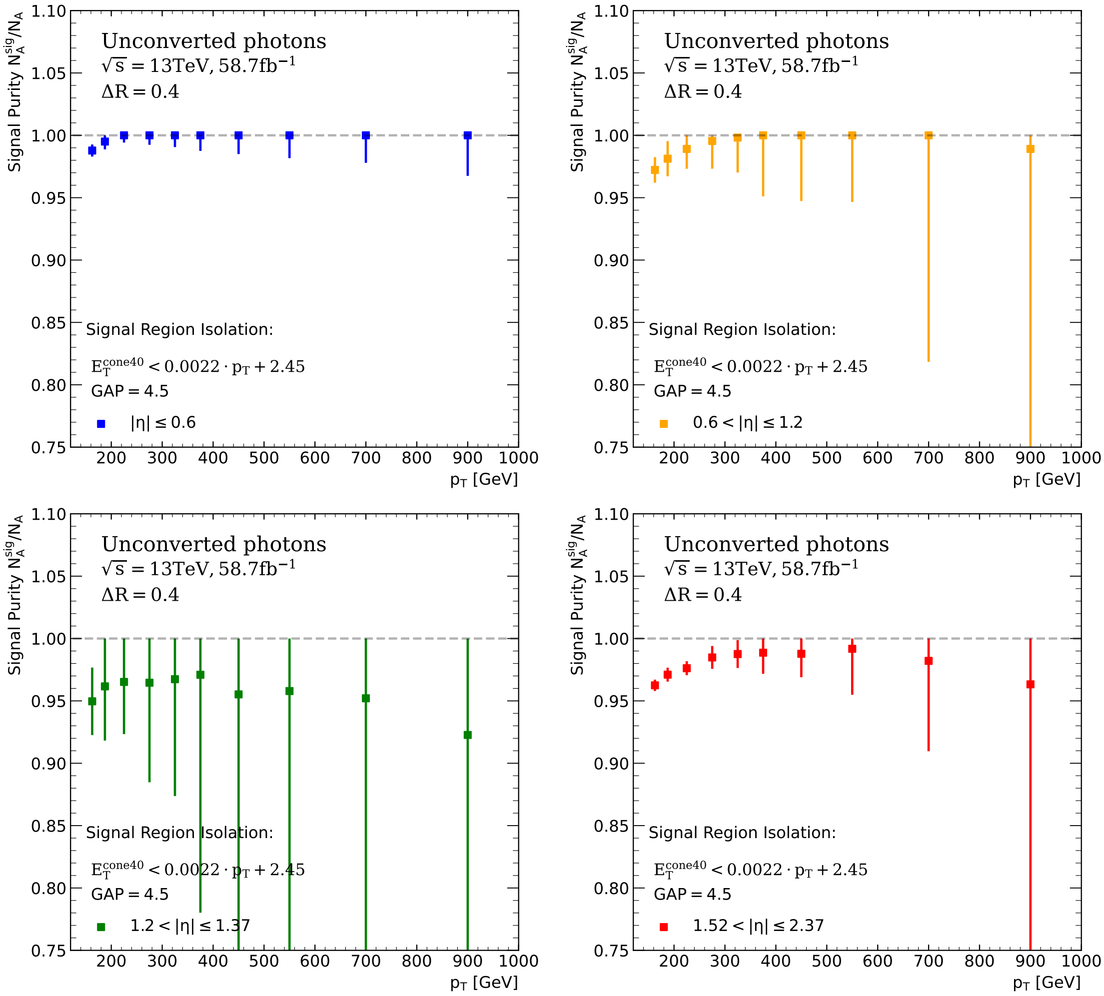

Figure 10: *Signal purities computed using the signal leakage fractions. Purities bigger that 1 were set equal to 1 for a correct interpretation of signal purity.*

Aknowledgements

I want to thank CERN for giving me the possibility of coming here and participate in a research project. Staying here has been absolutely amazing because of the work of NMS Summer Student Programme coordinators. In particular, I want to thank the donors who allow Non Member States students to participate in the Summer Programme. I also want to thank my supervisor Francisco Alonso, who devised an interesting project for me, both challenging and achievable. I leave CERN with a feeling of accomplishment that only clear results and nice plots can give. Francisco made time to help me whenever I needed it, and gave me advice not only for my project but also for living at CERN. María Teresa Dova and Hernán Wahlberg from Universidad Nacional de La Plata also gave me their support and advice during my stay.

Last but not least, I want to thank all the friends I made here at CERN. For eight weeks we all worked here and shared the same experience. We travelled together, had fun together, did science together and supported each other all the way through, and I could not have asked for better company. I am definitely sure we will meet again, probably sooner than later. In the meantime, I can only say that I will miss you all.

References

[1] M. Tripiana, *Medida de la sección eficaz de producción de fotones directos aislados en colisiones pp a* $\sqrt{s}$*=7 TeV en el experimento ATLAS.* **La Plata U., 2012.**

[2] The ATLAS Collaboration, *Inclusive-photon production and its dependence on photon isolation in pp collisions at* $\sqrt{s}$ *= 13 TeV using 139 fb*$^{-}1$ *of ATLAS data.* **CERN-EP-2022-247arXiv:2302.00510**

[3] The ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015-2017 LHC proton-proton collision data.* **JINST 14 (2019) P12006arXiv:1908.00005**

[4] The ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016.* **Eur. Phys. J. C 79 (2019) 205arXiv:1810.05087**

[5] The ATLAS Collaboration, F. Alonso, F. Arduh, M. T. Dova, J. Hoya, G. Orellana, H. Wahlberg, *Search for SUSY signatures in events with photons, missing transverse energy and jets using 139 fb*$^{-1}$. **Eur. Phys. J. C 79 (2019) 205 ATL-COM-PHYS-2018-1420**

[6] D. Camarero , C. Glasman , J. Terrón, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016.* **ATL-COM-PHYS-2021-383**

[7] J-D. Bossio, L. Carminati, A. Cueto, Y. Okumura, R. Piegaia, K. Sekhon,4 K. Terashi, M. Villaplana, K. Sekhon, R. Edgar, D. Amidei, *JDM* γ*+jet : Search for new physics in photon+jet final states at 13 TeV (2015+2016).* **EXOT-2016-26**