Low transverse momentum Drell Yan resolution using nanoAOD(plus)

Resolution limits & Analisys framework

Autor: BSc. David Guiérrez Menéndez
Supervisor: Dr. Armando Bermudez Martinez

Project Report

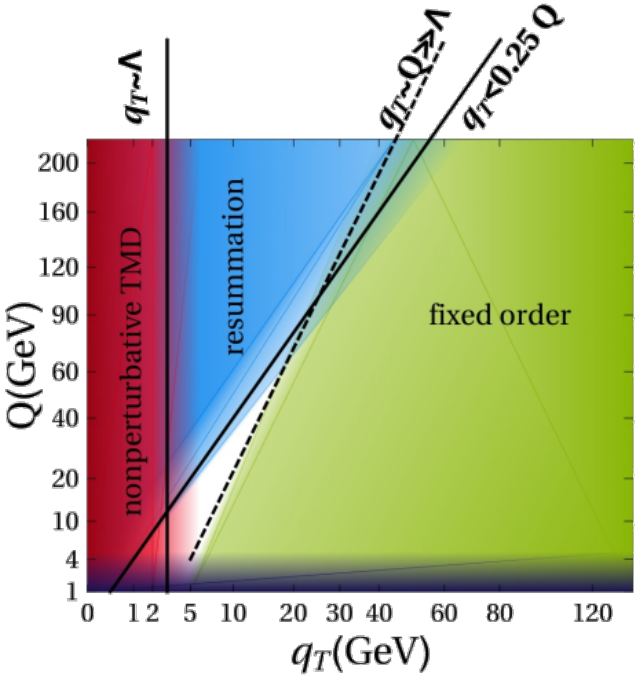
Summer Program 2022 - CMS Group - DESY

Hamburg
September 2022

Contents

1 Introduction

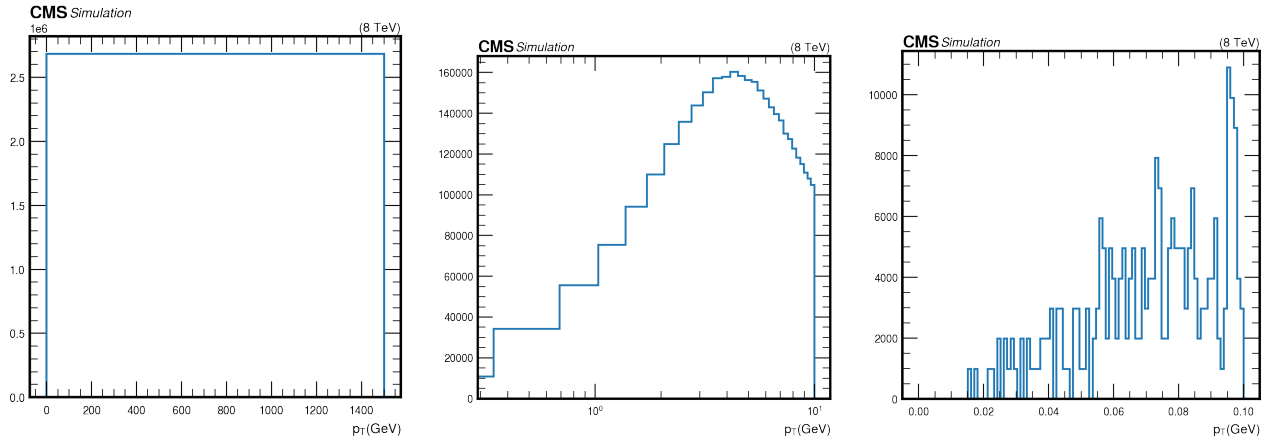
The annihilation of a quark-antiquark pair into two oppositely charged leptons, through the exchange of a virtual photon or a Z boson in the s channel, is known as the Drell–Yan (Z/γ^*) process. The perturbative theoretical derivation of the matrix element is available up to next- to-next-to-leading order (NNLO) in perturbative quantum chromodynamics (QCD) with next- to-leading order electroweak corrections, and a precise measurement can add valuable information on the nonperturbative part of the process, including parton distribution functions (PDFs). [CMS-Collaboration, 2020]



As is shown in the previous figure, the low p_T region is only accessible by Soft-Gluon perturbative Resummation and nonperturbative contributions, e.g. Transverse Momentum Dependent Parton Distribution Functions (TDM PDFs or TDMs).

Also, a better description of the p_T spectra of the vector bosons is important in order to improve precision in the W boson mass, since it holds mayor influence on the transverse mass, or momentum, of simulated leptons used to extract the m_W. [Bermudez Martinez et al., 2019]

In order to increase the precision in analysis and predictions, a greater resolution is needed in the low p_T region, but detectors are not perfect, and a finite resolution imposes over the desire of absolute truth. An example of this is shown in the next figure where increasing resolution (left to right) does not always means better, understandable, results.



When analyzing data related to these events, a proper resolution must be defined and measurements grouped accordingly, but knowing how this transformation affects the usability of said data is crucial for success.

1.1 Objectives

In this path, one of the project's goals was the study of these effects on multiple observables, and the relations between those variations.

Another motivation for this project was the opportunity to work with the nanoAOD(plus) format, exploit his capabilities and validate his necessity in modern analyses.

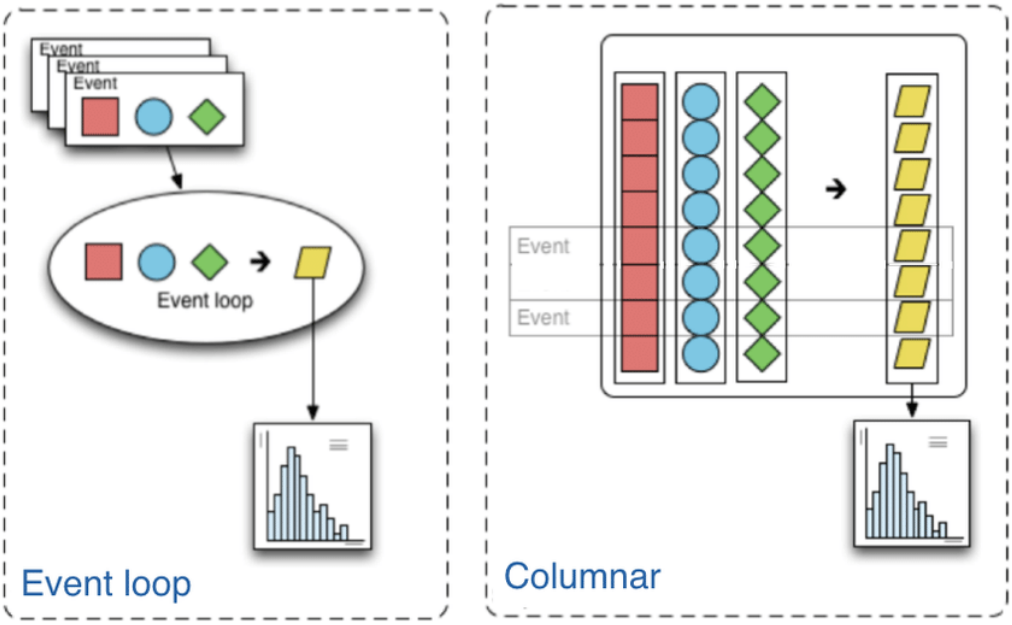
And last but not least, popularize the importance of Python's scientific software stack in the creation of flexible and powerful analyses that take advantage of both the columnar data layout and modern computing capabilities.

2 It's all about the Data

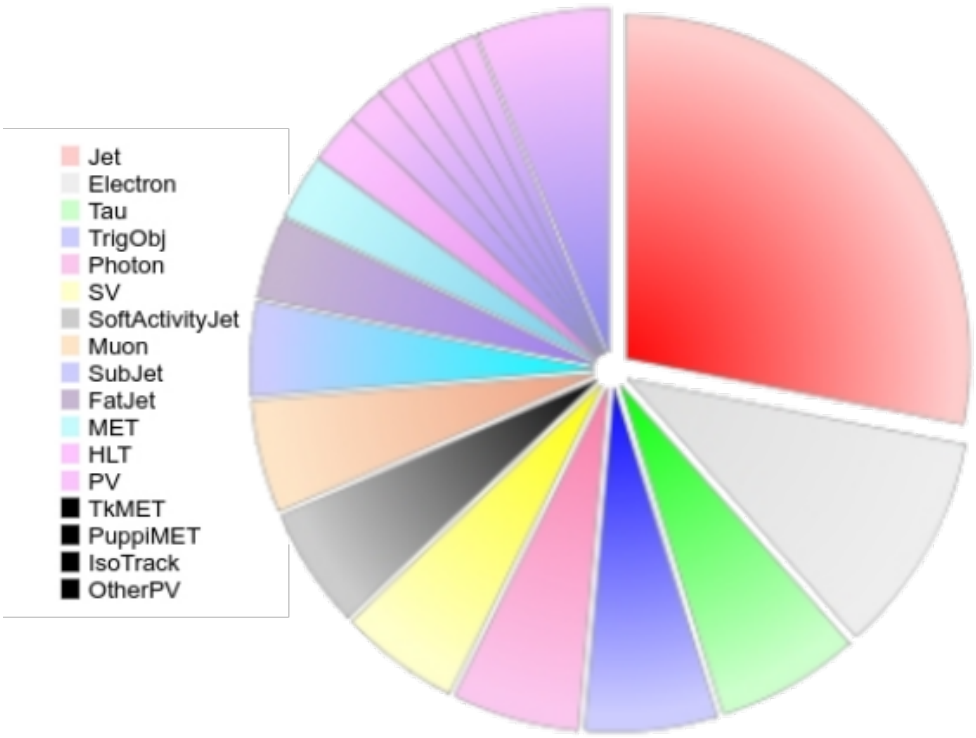
Accessing data from physics experiments must be a painless process, as long you have the right permissions, and the regularization of those datasets is an important part of the process. Multiple formats are used in the CMS data available for research, depending of the amount of information needed and the level of post-processing accepted, the most advanced format in the pipeline can provide significant performance and storage advantages.

This time, the nanoAOD(plus) format was chose, a plain ROOT::TTree with data from Run 1 designed to bring the closest information possible to suit a large number of analyses created for Run 2 in his nanoAOD format, the "(plus)" means that additional variables are available when needed. [Rizzi et al., 2019]

As it has been mentioned, the nanoAOD(plus) format specifies a single ROOT::TTree with branches corresponding to observables [Stäger et al., 2019]. This allows an approach that exploits the natural layout of the data: columnar analysis and data manipulation. As opposite to the traditional event loop based analysis, selection, transformation, alignment and accumulation can be performed column-wise, unlocking the benefits of dealing with contiguous data of the same type (something that modern computers love).



Unlike his "brother", nanoAOD(plus) data can't be obtained from miniAOD, only available for Run 2, so the code has to be written to extract CMSSW events and mimic, when possible, the nanoAOD behavior. The result is a highly compliant and optimized dataset, capable of participate with dignity on "the machinery of science".



All the trust was deposited on the "orange" slice of the chart above, the Muons; allowing a clean identification of candidate events for Drell-Yan. This data in particular is originally generated by MC event generators, and is subjected to a simulation of the CMS detector in order to reproduce the effects of the detection apparatus and the reconstruction algorithms. The MC generated events and the muons reconstructed by the detector are utilized as counterparts for the analyses.

2.1 Event Selection

The main criteria for the DY candidates is the muon component of the events, a $\mu\mu^-$ pair are required to meet a series of cuts in order to select relevant data for the reconstruction of bosons, our subject for study.

At different levels of the analysis, a variety of cuts are applied to the data, for example: only the two strongest muons (in the sense of high p_T) are used, some other cuts are summarized next:

Generator level:

- $abs(pdgId) = 13$
- $status = 1$

Detector level:

- $isolation < 0.12 \quad [dR = 0.4]$
- $PV\_npvsGood >= 1$
- $TightID = True$
- $abs(\eta) < 2.4$
- $p_T > 8.0$
- $dxy < 0.2$

Dimuon system:

- $mass > 50$
- $y < 2.1$

2.2 Quality Metric

The proposed way of measure when the obtained distribution of an observable is no longer useful is trough the **purity**, which is the ratio of events generated in certain interval of the spectra that is reconstructed in the same interval and doesn't migrate to another as result of the resolution of the detector. Consistent results can be achieved by requiring a certain minimum for the purity in correspondence to the variable properties. [Abercrombie et al., 2019]

3 Analysis Framework

In the duration of the project, several tools where created for the extraction, filtration, and manipulation of the data. As is has been mentioned, the programming language Python was used to code said tools using multiple scientific, and general propose, libraries to assemble a framework capable of preserve information in a columnar format and committed to maintain flexible, reliable and powerful.

The use of interactive environments is a fast and reliable way of testing and developing ideas while being able to verify partial results, for this reason the majority of the code is in the form of Jupyter Notebooks, a front end for interactive python workflows that is fast, accessible and capable of cover a wide range of use cases: from educational notebooks to high performance workflows.

```
gen_system = read_df('gen_system')
det_system = read_df('det_system')

int_system = pd.merge(
    gen_system, det_system,
    on='event', how='inner',
    suffixes=('_gen', '_det')
)
```

[153] ✓ 0.1s

	mass		pt		rapidity	
	det	gen	det	gen	det	gen
event						
28463207	109.314659	108.167908	17.642130	16.186666	1.697918	1.688509
28463230	98.615936	98.090302	59.900818	59.842503	0.830399	0.836784
28463231	90.314606	91.091751	24.067490	24.160704	1.204804	1.200580
28463283	89.236366	90.488365	11.803128	10.795930	1.576683	1.574080
28463287	89.593163	90.260475	29.587938	30.215862	0.668830	0.668196
...	...	...	...	...	...	...
22373336	84.231834	84.932930	7.233629	7.646422	0.409831	0.418316
22373380	88.487328	90.627190	7.166288	6.977647	0.364980	0.384418
22373416	93.731140	94.724464	62.573864	63.466602	0.519642	0.518527
22373419	92.751572	91.939011	20.776375	19.984383	0.069082	0.064090
22373458	90.789886	91.580132	12.432115	11.874322	0.122079	0.120800

2684395 rows × 6 columns

Figure 3.1: Overview of a dataset.

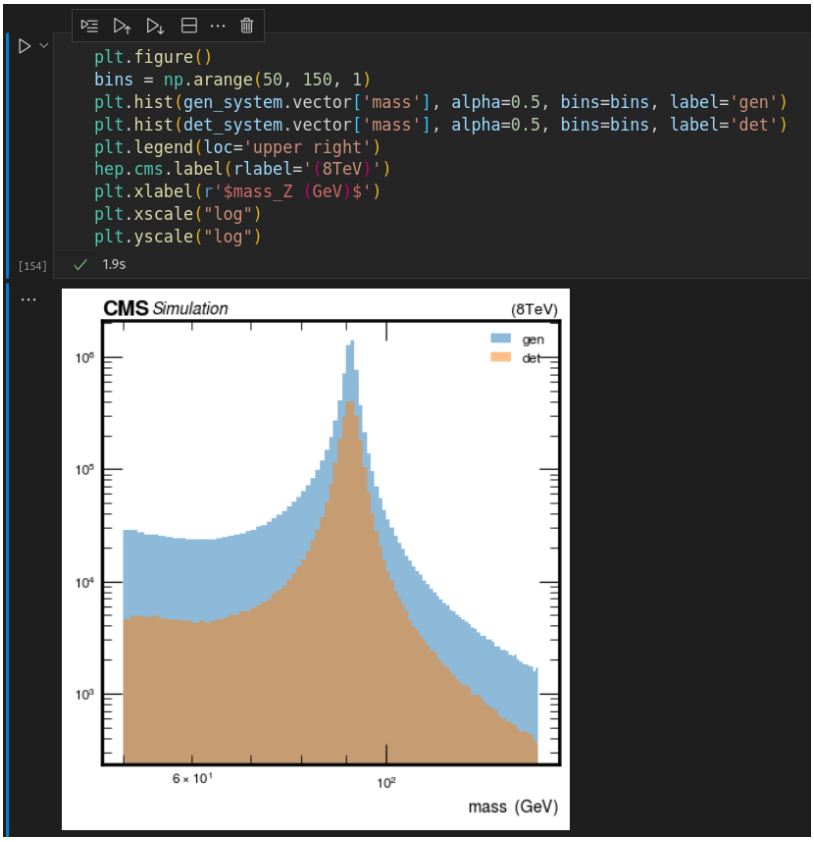


Figure 3.2: Integrated plots, useful for quick histograms.

Scalability is a key aspect for analysis, the ability to distribute the load in high performance or hight throughput clusters can improve the quality of the analysis and save time. This option was also tested successfully but it wasn't needed due to the memory efficiency of the data format.

4 Results

Bellow is presented a plot of the purity of each p_T bin by mass interval, this distinction in the mass dimension is interesting because reveals the detector's capabilities of reconstructing accurate events.

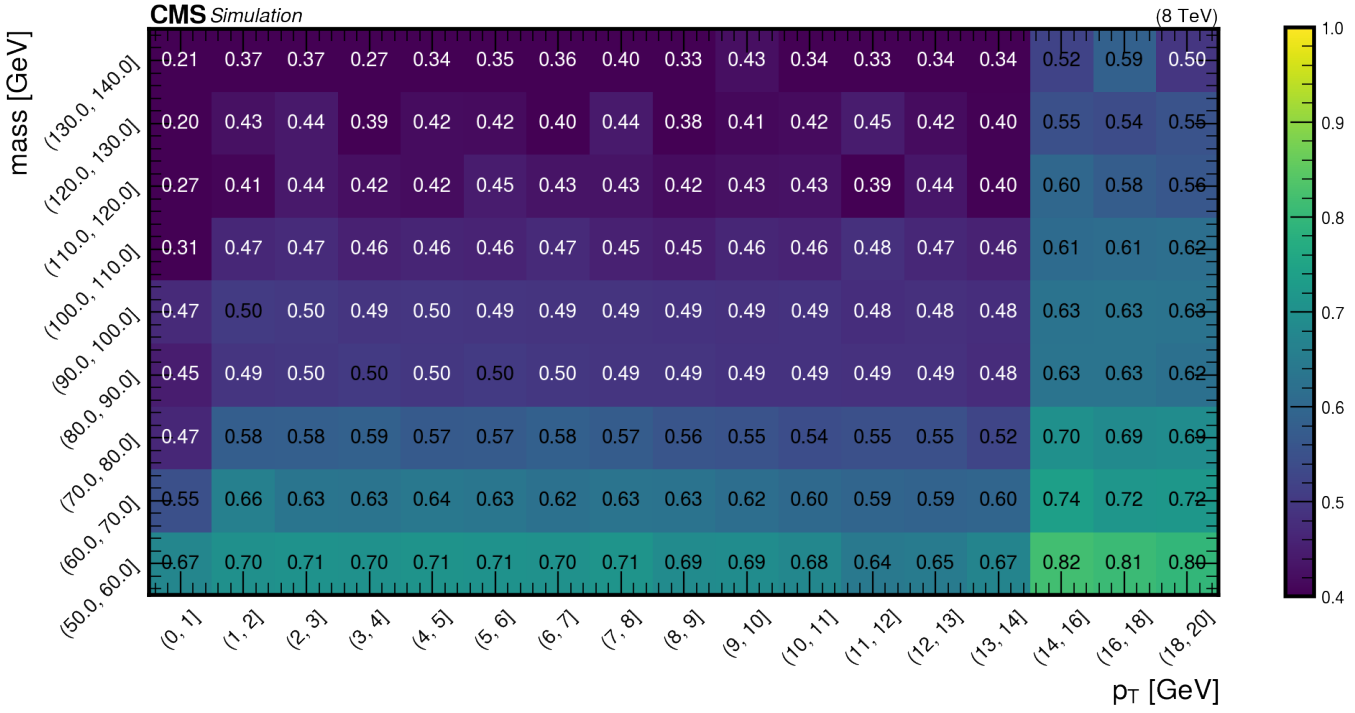


Figure 4.1: Purity of p_T bins by mass interval.

The observed behavior indicates that in the low p_T regions, where the efforts are focused, is expected to obtain good results as long as the mass region is also low.

In an effort to push harder the resolution of the p_T spectra, a second result involves only the first bin, but reflects the behavior with increasing size of said bin.

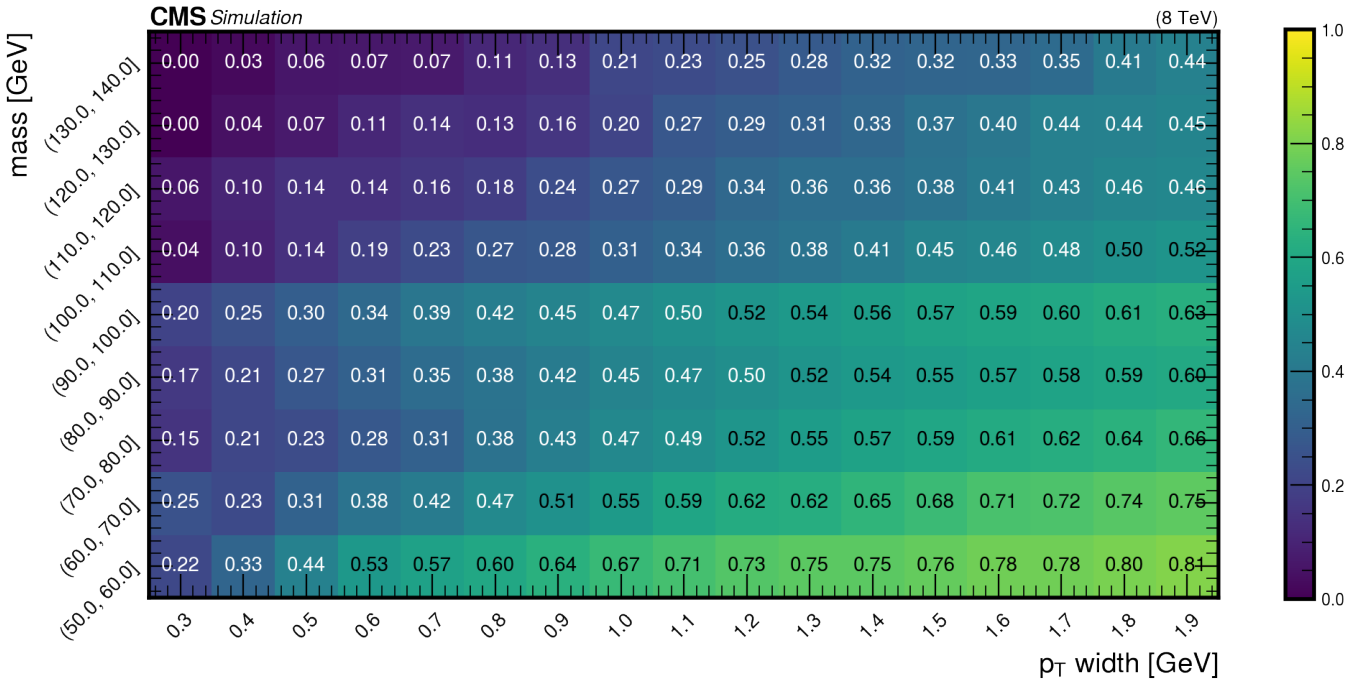


Figure 4.2: Purity of first p_T bin by mass interval.

A similar result is shown, reinforcing prior indications while diving deeper into the limits of a good p_T spectra. Most common analyses are conducted with a bin size of 1 GeV, even increasing bin sizes at larger p_T regions, but from the plot can be concluded that a bin size of 0.6 GeV in the lower region can deliver consistent results while increasing the resolution.

Showing good looking metrics is not enough to ensure quality in future analyses, we want to se the pretty histograms... their are an acquired taste, so the chosen way was to apply an unfolding to the data, also with a python package, and because the comparison is performed on the same data it should match perfectly.

As a rule of thumb, using significantly more events at the generator level and a grater resolution at the detector level is advised, and this test was performed using over 6 million generated events with a bin size of 1 GeV versus 1.5 million at detector level with a bin size of 0.5 GeV.

The unfolded distribution matches the "true" from the MC generator, proving that the resolutions used are safe. Future verifications can be made on real data from the detector to extract more useful information where thing get harder, and nobody knows the truth.

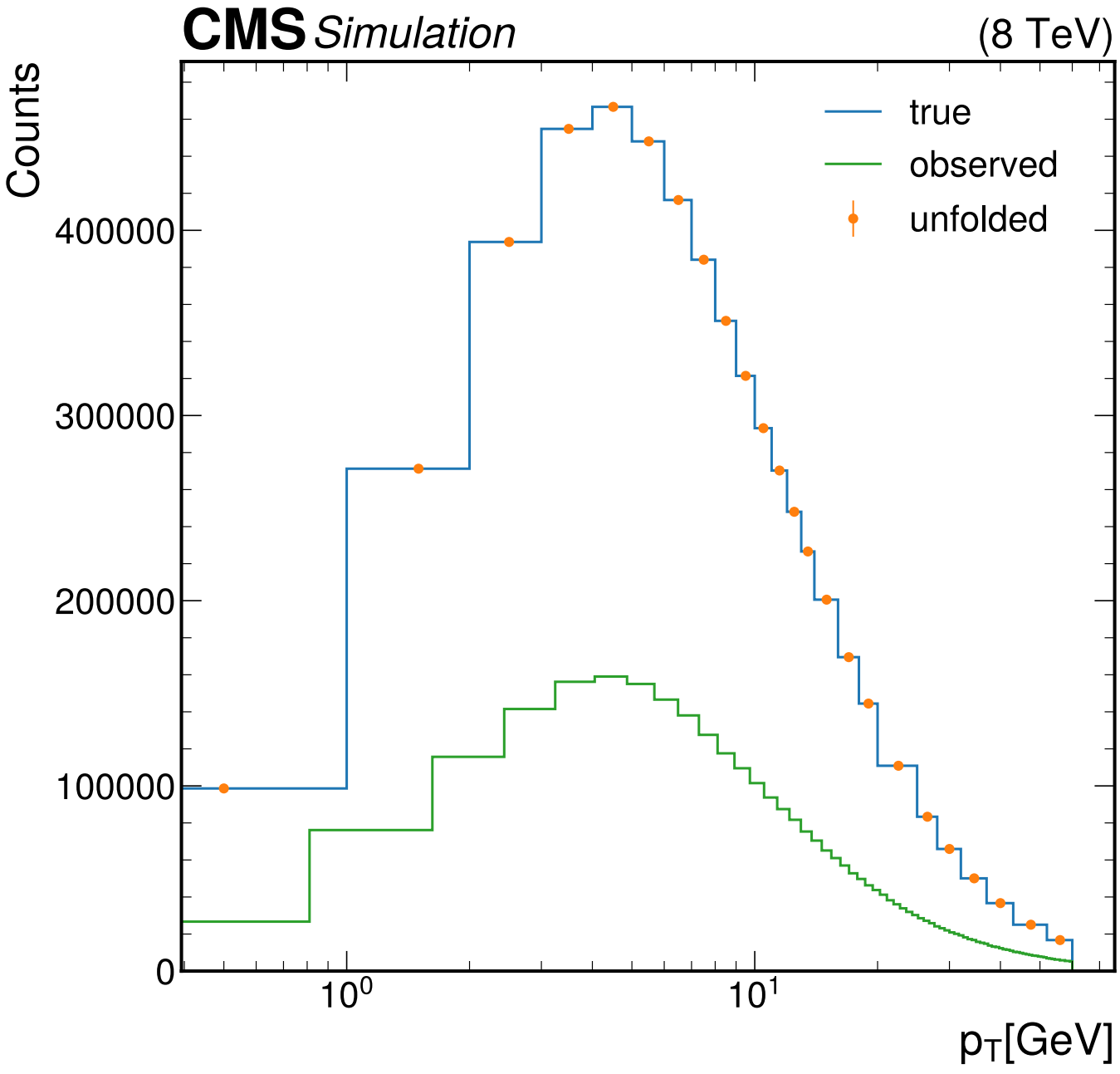


Figure 4.3: Unfolding of p_T spectra.

5 Conclusions

It started slow, then I gain a bit of confidence and it became a smother process, event after event I was incorporating new knowledge; sometimes I had to start over, it happens, but at the end, all the hard work paid off. This was one of the most productive, and fun, summers of my life, where I met an amazing place with amazing people in it. My project also went well, but I already had 10 pages about that ;)

Bibliography

[Abercrombie et al., 2019] Abercrombie, D., Apyan, A., Brandt, S., Demiragli, Z., Gómez-Ceballos, G., Hsu, D., Hu, M., Iiyama, Y., Kovalskyi, D., Klute, M., Lawhorn, J., Marini, A., Narayanan, S., Niu, X., Paus, C., Spiropulu, M., Wang, C., and S, X. (2019). Measurements of differential z boson production cross sections in pp collisions at snn=13 tev. *CERN*.

[Bermudez Martinez et al., 2019] Bermudez Martinez, A., Connor, P. L. S., Dominguez Damiani, D., Estevez Banos, L. I., Hautmann, F., Jung, H., Lidrych, J., Schmitz, M., Taheri Monfared, S., Wang, Q., and Žlebčík, R. (2019). Production of z bosons in the parton branching method. *Phys. Rev. D*, 100:074027.

[CMS-Collaboration, 2020] CMS-Collaboration (2020). Differential measurements of the drell–yan process in the muon channel in ppb collisions at snn = 8.16 tev.

[Rizzi et al., 2019] Rizzi, A., Petrucciani, G., and Peruzzi, M. (2019). A further reduction in cms event data for analysis: the nanoaod format. *EPJ Web of Conferences*, 214:06021.

[Stäger et al., 2019] Stäger, F., Geiser, A., and Metwally, J. (2019). nanoaod(plus) validation from comparison to 2010 mumonitor and muonia open data examples.

A Workflow Diagram