Conclusion

'It seems very pretty,' she said when she had finished it,
'but it's rather hard to understand! [...]
Somehow it seems to fill my head with ideas —
only I don't exactly know what they are!'

Lewis Carroll, Through the Looking-Glass

The operation of the LHC since its switching on in 2010 has been remarkable. The Higgs boson was discovered in 2012, previous exclusion limits in searches for new physics have been significantly extended, and a great number of Standard Model processes have been measured to high precision.

This thesis has introduced two analyses of the complete 8 TeV dataset performed by the author. Both analyses focus on final states with large jet multiplicities, but they take different, complementary approaches. The first analysis looks for events with 8, 9, 10 or even more jets, since these could indicate the presence of new physics. Relying on a powerful data-driven technique to determine the dominant Standard Model backgrounds, the search is able to exclude supersymmetric gluinos with masses below 1.1 TeV in a simplified model where they decay via $\tilde{g} \rightarrow t + \bar{t} + \tilde{\chi}_1^0$. The second analysis focuses on events with jet multiplicities of 4 or more, performing the first measurement of their cross section at $\sqrt{s} = 8$ TeV. The measurement is performed differentially in several variables which provide discrimination between different theoretical predictions. Both the SUSY multi-jet search and the SM measurement have been published in refereed journals in slightly updated versions [1, 2].

The LHC will turn on again in 2015 (only some months from the time of writing), and start taking data at 13 TeV. No one knows what kind of phenomena may occur at such high energies, and although searches for new physics such as the one presented in this thesis have been unsuccessful so far, this may very well change in the near future. Simplified versions of SUSY have been highly constrained by LHC searches, but as discussed in the opening chapter of this thesis, more general SUSY models

M. Crispín Ortuzar, *High Jet Multiplicity Physics at the LHC*,
Springer Theses, DOI 10.1007/978-3-319-43461-2

remain alive. Complementing the searches with measurements such as the one presented in this thesis allows the theoretical community to make progress in parallel with the calculation of the SM backgrounds. This is important both to deepen our understanding of the physics of the Standard Model, and to improve the sensitivity of new physics searches and the robustness of future discoveries.

This feedback process between searches, measurements and theory is at the heart of LHC physics, and guarantees that at the end of its lifetime, no matter what Nature is hiding at high energies, the experiment will have been a success.

References

1. ATLAS Collaboration. (2013). Search for new phenomena in final states with large jet multiplicities and missing transverse momentum at $\sqrt{s} = 8\,\text{TeV}$ proton-proton collisions using the ATLAS experiment. *Journal of High Energy Physics*, *2013*(10), 1–50.
2. ATLAS Collaboration. (2015). Measurement of four-jet differential cross sections in $\sqrt{s} = 8\,\text{TeV}$ proton-proton collisions using the ATLAS detector. *Journal of High Energy Physics*, *2015*(12), 1–76.

Appendix A
Trigger Efficiencies

Trigger efficiency curves allow one to study the fraction of events which fire a particular trigger as a function of a certain variable x. The inefficiencies are introduced by differences between the objects and decisions made at the different stages of the trigger and at the offline level. The sample of events used for the test must be carefully chosen so as not to introduce a bias. One option to measure the efficiency of a trigger A is to select a set of events which fired an orthogonal trigger B, in which case the efficiency of A, $\varepsilon(A)$, is simply the bin-by-bin ratio

$$\varepsilon(A, x) = \frac{\mathcal{H}(x)^{\text{Events triggered by A and B}}}{\mathcal{H}(x)^{\text{Events triggered by B}}}, \tag{A.1}$$

where $\mathcal{H}(x)^{\alpha}$ is the histogram of the variable x for events selected according to α.

It is also possible to use a sample triggered by a looser trigger C, such that $\varepsilon(C|A) = 1$, and then, according to Bayes' theorem,

$$\varepsilon(A) = \frac{\varepsilon(A|C)\,\varepsilon(C)}{\varepsilon(C|A)} = \varepsilon(A|C)\,\varepsilon(C), \tag{A.2}$$

where $\varepsilon(C)$ is implicit in the event selection, and $\varepsilon(A|C)$ can also be obtained as

$$\varepsilon(A|C, x) = \frac{\mathcal{H}(x)^{\text{Events triggered by A and C}}}{\mathcal{H}(x)^{\text{Events triggered by C}}}. \tag{A.3}$$

A certain trigger C is said to be *looser* than a given trigger A if its cuts are less restrictive than the cuts implicit in A; for example, C could have the same configuration as A but demand a lower p_{T} from a particular object.

The calculation of the efficiency is more delicate when the triggers are prescaled, since the correlation between the events that pass the reference and the test triggers is lost. Fortunately, the response from jet triggers can be easily emulated, as all the information about the trigger objects needed (L1 cells, L2 clusters and Event Filter jets) is kept. The efficiency of a prescaled jet trigger can then be calculated as long as

M. Crispín Ortuzar, *High Jet Multiplicity Physics at the LHC*,
Springer Theses, DOI 10.1007/978-3-319-43461-2

there exist an unbiased sample of events triggered by either an orthogonal trigger or a looser trigger, any of which could be prescaled as well. The prescale of the trigger of interest is effectively removed by reconstructing the trigger response for each of the events that passed the (prescaled or not) reference trigger. The prescale of the reference trigger is unavoidable, but the only effect it will have on the final result is to reduce the total luminosity.

In this thesis, efficiencies are typically computed bin-by-bin in the p_{T} of the Nth jet for an N-jet trigger, or in the minimum separation dR_{min}^{4j} between the jets, defined in Sect. 4.3.2 on p. 110. Using a bin-by-bin notation, the efficiency $\varepsilon^i(A)$ for a given trigger A in bin i can also be written as

$$\varepsilon_i(A) = \frac{N_i^{\text{Events triggered by A and B}}}{N_i^{\text{Events triggered by B}}}, \tag{A.4}$$

where the denominator is the number of events passing the reference trigger B in bin i, and the numerator is the number of events passing both the reference trigger B and the trigger of interest A in bin i. This defines a 'differential' efficiency curve, since the efficiency is calculated in each bin in an exclusive manner. To better illustrate certain trends in the turn-on curves, the 'integral' efficiency is also studied in most cases. The integral efficiency is defined as:

$$\varepsilon_i^{integ}(A) = \frac{\sum_{j=i}^{\infty} N_j^{\text{Events triggered by A and B}}}{\sum_{k=i}^{\infty} N_k^{\text{Events triggered by B}}}. \tag{A.5}$$

The integral efficiency informs of the *global*, inclusive efficiency of the trigger for a particular set of cuts, rather than for an exclusive bin.

In the figure legends, efficiencies will be indicated as 'Nominal trigger wrt. reference trigger'. The data taken by the reference trigger are not corrected for prescales.

Appendix B
Deriving Variable-Width Binnings

The goal of this method is to derive a binning that yields an approximately flat purity (as opposed to steeply falling, as would be the case for a p_{T} spectrum with constant binning) with an average of $P = 80\ \%$. The stability distributions are used as a cross-check. Purity and stability are defined in Sect. 4.5.

Let i be the index used to enumerate reco-level quantities. Let r^i represent the bin centres along the reco axis. The first step consists on projecting the contents of each reco bin such that they can be studied as individual one-dimensional histograms. The standard deviation of each of these distributions is then plotted versus r_i, as an estimate of the resolution of the variable.

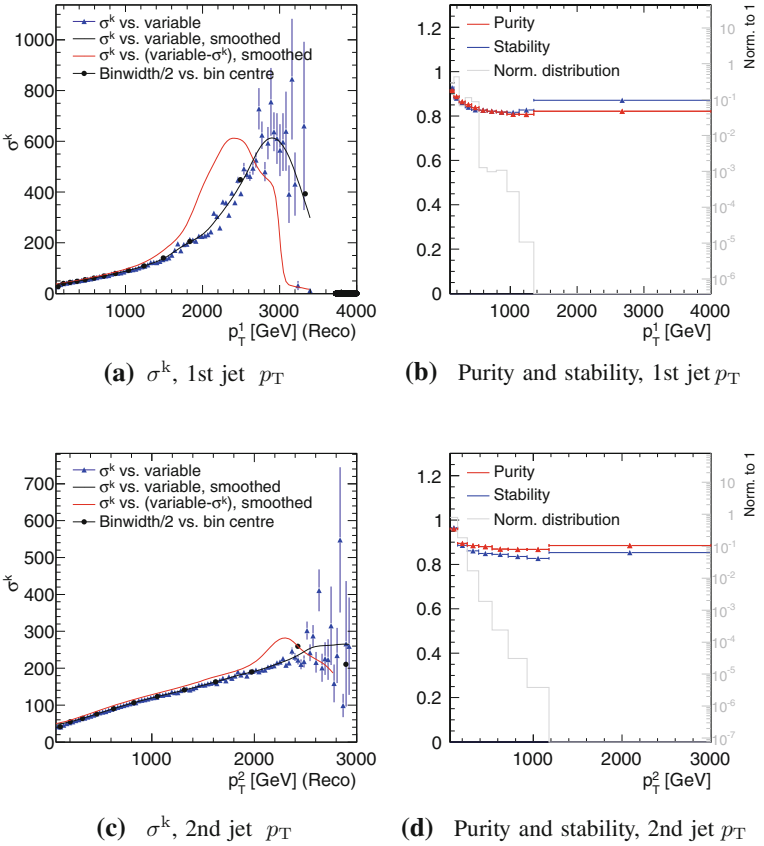
The resolution at a particular value of r_i ($\sigma_i^{\mathrm{Resol}}$) is in principle taken as the targeted width of the bin centred at r_i. The distribution of $\sigma_i^{\mathrm{Resol}}$ as a function of r_i therefore shows how the bin width changes as a function of r_i. However, what is needed is a discrete set of non-overlapping, successive bins, so the next step is to obtain a smooth distribution and sample from it.

The scatter distribution of $\sigma_i^{\mathrm{Resol}}$ as a function of r_i is smoothed by calculating a non-parametric Gaussian kernel regression. The regression curve is formed by a succession of points, each of which is computed as the weighted average of all the $\sigma_i^{\mathrm{Resol}}$. For each point contributing to the average, the weight is determined from a Gaussian centred in r_i and evaluated at said point. The standard deviation (or 'bandwidth') of the Gaussian is fixed for each of the variables studied, and its value is chosen so as to be sensitive to the shape of the distribution of $\sigma_i^{\mathrm{Resol}}$, while reducing the statistical fluctuations of the data. The left plots in Fig. B.1 show the $\sigma_i^{\mathrm{Resol}}$ distributions (blue triangles) and the corresponding regressions (black line) for the leading and sub-leading jet p_{T}.[1]

The binning is obtained by sampling the $\sigma_i^{\mathrm{Resol}}$ curve. In order to simplify the sampling algorithm, the distribution used is actually $\sigma_i^{\mathrm{Resol}}$ versus $r_i - \sigma_i^{\mathrm{Resol}}$, defining the curve $C(r_i)$. In other words, rather than studying the bin widths as a function of

[1] The distributions are weighted by a factor k, for reasons that will be explained later.

M. Crispín Ortuzar, *High Jet Multiplicity Physics at the LHC*,
Springer Theses, DOI 10.1007/978-3-319-43461-2



(a) σ^{k}, 1st jet p_{T} **(b)** Purity and stability, 1st jet p_{T}

(c) σ^{k}, 2nd jet p_{T} **(d)** Purity and stability, 2nd jet p_{T}

Fig. B.1 σ_i^{k} distribution, purity and stability for $\Sigma_{p_{\mathrm{T}}^{\mathrm{central}}}$ and $\Delta Y_{ij}^{\mathrm{max}}$. In the σ_i^{k} distributions (*left*), the *blue triangles* correspond to the truth σ_i^{k} as calculated from PYTHIA for the central value of each of the reco bins, r_i; the *black curve* is the regression of the σ_i^{k} versus the central value of each of the reco bins; the *red curve* is the regression of the σ_i^{k} versus $r_i - \sigma^{\mathrm{k}}$ (this is the curve used to do the bin sampling); and the *black points* correspond to the centres and widths of the bins proposed. The *grey histograms* (*right*) correspond to the resulting variable distributions

the bin centres, we study them as a function of the lower bin edges. The algorithm then proceeds by iterating over the data and creating new bins in the following way:

1. The lower edge of the first bin is set manually.
2. The width of the first bin is $2 \times \sigma_1^{\mathrm{Resol}}$—i.e., instead of evaluating the fitted curve at r_i, the first raw value of $\sigma_i^{\mathrm{Resol}}$ is used. This is done because the regression curves tend to be higher than the data at the lower end of the spectrum, giving artificially large bin widths. This condition forces the first bin to be small. Figure B.1 shows that it is safe to do this for the momentum variables, as the statistical uncertainty is smallest in that region.

3. The upper edge of the first bin is then renamed as the lower edge of the second bin.
4. The bin width of the second bin is calculated as $2 \times C(b_{low})$, where b_{low} is the lower edge of the bin.
5. The new upper edge turns into the lower edge of the third bin, and so on.
6. The process finishes when the upper edge is bigger than a maximum r_i set manually.

Bins are then merged if:

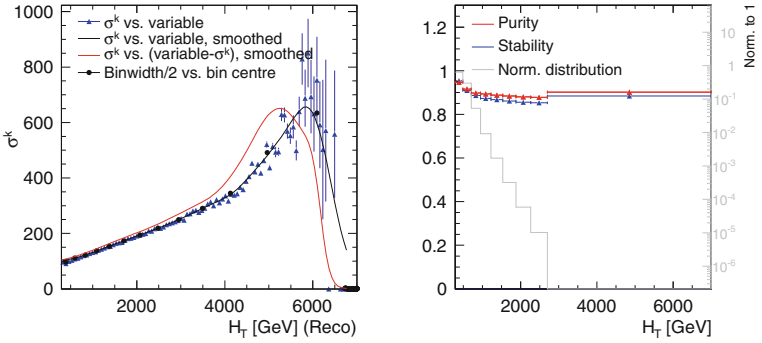
- The width is smaller than a certain number set by hand.
- The number of Monte Carlo events equivalent to 20.3 fb^{-1} is smaller than 100.
- The purity is smaller than 70 %.

Taking σ_i^{Resol} as the starting point to derive the binnings results in purities between 60 and 70 %. In order to increase this value to roughly 80 %, we multiply the values of σ_i^{Resol} by a constant. We define, for convenience, $\sigma_i^{\text{k}} = k \times \sigma_i^{\text{Resol}}$, where k is the multiplicative factor used to increase the purity.

Figure B.1 shows the σ_i^{k} curves, purity and stability for the leading and subleading jet p_{T}. The proposed bin centres, before applying the constraints on statistics or minimum purity, correspond to the black circles. They are obtained by sampling the red curve (σ_i^{k} versus $r_i - \sigma_i^{\text{k}}$) following the method described above, and are found to agree with the black curve (σ_i^{k} versus r_i), as expected. Figs. B.2 and B.3 show equivalent results for other momentum variables. Table B.1 shows the values of x_{min}, x_{max} and bandwidth used to derive the binning.

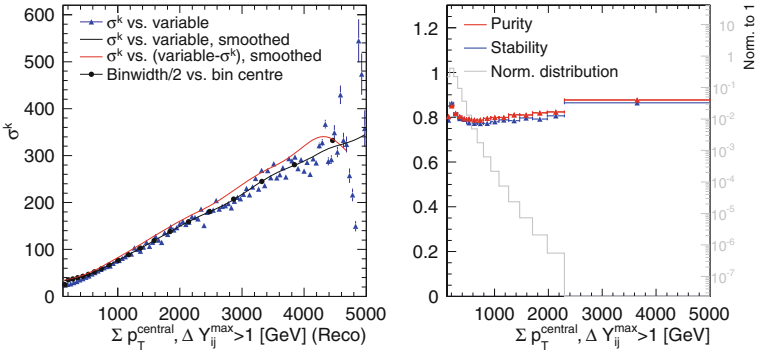
Table B.1 Parameters used to derive the variable-width bins for the momentum and mass variables

Variable	x_{min}	x_{max}	Bandwidth	k factor
p_{T}^1	100 GeV	4000 GeV	300 GeV	2.0
p_{T}^2	64 GeV	3000 GeV	300 GeV	3.0
p_{T}^3	64 GeV	2000 GeV	200 GeV	3.0
p_{T}^4	64 GeV	1500 GeV	200 GeV	2.0
H_{T}	292 GeV	7000 GeV	700 GeV	3.0
M_{jjjj}	100 GeV	7000 GeV	2000 GeV	1.3
$M_{jj}^{\text{min}}/M_{jjjj}$	0	0.4	0.05	1.5
$\Sigma_{p_{\text{T}}^{\text{central}}}, \Delta Y_{ij}^{\text{max}} > 1$	128 GeV	5000 GeV	700 GeV	1.0
$\Sigma_{p_{\text{T}}^{\text{central}}}, \Delta Y_{ij}^{\text{max}} > 2$	128 GeV	5000 GeV	700 GeV	1.3
$\Sigma_{p_{\text{T}}^{\text{central}}}, \Delta Y_{ij}^{\text{max}} > 3$	128 GeV	5000 GeV	700 GeV	1.5
$\Sigma_{p_{\text{T}}^{\text{central}}}, \Delta Y_{ij}^{\text{max}} > 4$	128 GeV	5000 GeV	700 GeV	1.7



(a) σ^{k}, H_T **(b)** Purity and stability, H_T

Fig. B.2 σ_i^{k} distribution, purity and stability for H_T. In the σ_i^{k} distributions (*left*), the *blue triangles* correspond to the truth σ_i^{k} as calculated from PYTHIA for the central value of each of the reco bins, r_i; the *black curve* is the regression of the σ_i^{k} versus the central value of each of the reco bins; the *red curve* is the regression of the σ_i^{k} versus $r_i - \sigma_i^{\mathrm{k}}$ (this is the curve used to do the bin sampling); and the *black points* correspond to the centres of the bins proposed. The *grey histograms* (*right*) correspond to the resulting variable distributions



(a) σ^{k}, $\Sigma p_{\mathrm{T}}^{\mathrm{central}}\ \Delta Y_{ij}^{\mathrm{max}} > 1$ **(b)** Purity $\Sigma p_{\mathrm{T}}^{\mathrm{central}}\ \Delta Y_{ij}^{\mathrm{max}} > 1$

Fig. B.3 σ_i^{k} distribution, purity and stability for $\Sigma_{p_{\mathrm{T}}^{\mathrm{central}}}$, with $\Delta Y_{ij}^{\mathrm{max}} > 1$. In the σ_i^{k} distributions (*left*), the *blue triangles* correspond to the truth σ_i^{k} as calculated from PYTHIA for the central value of each of the reco bins, r_i; the *black curve* is the regression of the σ_i^{k} versus the central value of each of the reco bins; the *red curve* is the regression of the σ_i^{k} versus $r_i - \sigma^{\mathrm{k}}$ (this is the curve used to do the bin sampling); and the *black points* correspond to the centres of the bins proposed. The *grey histograms* (*right*) correspond to the resulting variable distributions

About the Author

Mireia Crispín Ortuzar is a Research Fellow at Memorial Sloan-Kettering Cancer Center (MSKCC) in New York, and a Title A Fellow at Trinity College, University of Cambridge. She holds a DPhil (PhD) in Particle Physics from the University of Oxford, and undergraduate degrees in Physics and Music from the University of Valencia and the Valencia Higher Conservatory of Music, respectively.

Between 2011 and 2016 Mireia was a member of the European Laboratory for Nuclear Research in Geneva (CERN), where she worked on searches for new supersymmetric candidates at the Large Hadron Collider (LHC). She also led the measurement of the cross section of multi-jet events at 8 TeV, and studied the expected performance of the ATLAS detector after future upgrades. Previously she had worked on gravitational-wave detection techniques at the LIGO experiment. Her current research at MSKCC focuses on the development of predictive computational models to personalise cancer treatment.

For her work she has received the Winton Prize (2014), Perkins Prize (2012) and Foley-Bejar Scholarship (2011) by the University of Oxford; the Telefonica Award by the British-Spanish Society (2014); the National Prize for Academic Excellence by the Spanish Ministry of Education (2014); the Prize for Extraordinary Performance in Physics by the University of Valencia (2011); and the Arquimedes Award for the Rapprochement of Cultures by the Spanish Ministry of Education (2011), among others.

Mireia is also interested in current affairs, science communication and supporting women in STEM. She co-founded the Madariaga Series at the University of Oxford (2014), helped organise the 1st Conference for Undergraduate Women in Physics in the UK (2015), and co-leads the Tri-Institutional Science and Education Policy Association in New York. In her free time Mireia also enjoys choral singing, and has performed at Carnegie Hall in New York, the Royal Albert Hall in London and St Mark's Cathedral in Venice, among others.

M. Crispín Ortuzar, *High Jet Multiplicity Physics at the LHC*,
Springer Theses, DOI 10.1007/978-3-319-43461-2