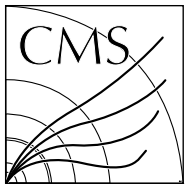
Available on CMS information server

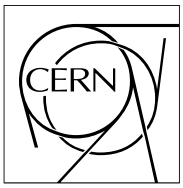
CMS AN-2008/034



The Compact Muon Solenoid Experiment

Analysis Note

The content of this note is intended for CMS internal use and distribution only



12 June 2008

Search Strategies for mSUGRA in the Muons + Jets + MET Final State

Philipp Biallass, Thomas Hebbeker, Carsten Hof, Markus Merschmeyer, Arnd Meyer, Holger Pieta, Daniel Teyssier, Clemens Zeidler

Abstract

Searches for new phenomena will be one of the primary goals of the LHC experiments once data taking commences. In particular, low mass supersymmetry might be found with modest amounts of integrated luminosity. We review the prospects for the discovery of mSUGRA with the early CMS data in final states containing a charged lepton, multiple jets, and missing transverse energy. Three types of analyses are envisioned: a traditional "cut based" search, a multivariate analysis technique utilizing boosted decision trees, and a model independent search for new physics. We compare the performance of the three analyses for several mSUGRA benchmark points, and discuss their benefits and drawbacks.

1 Introduction

Supersymmetry (SUSY) [1] predicts the existence of a new particle for each Standard Model (SM) particle, differing by half a unit in spin but otherwise sharing the same quantum numbers. In a supergravity scenario, the gravitino will acquire mass through the spontaneous breaking of local SUSY. The SUSY-breaking is then communicated to the so called observable sector so that, in particular, the gluino acquires its mass [2]. For the study presented here, we use a scenario of minimal supergravity (mSUGRA) with conserved R-parity. This leaves five model parameters: the universal soft breaking mass parameter for all scalars at the unification scale, m_0; the common gaugino mass, $m_{1/2}$; a universal trilinear coupling, A_0; the ratio of the vacuum expectation values of the two neutral Higgs fields, $\tan\beta$; and the sign of the higgsino mixing mass parameter, $sgn(\mu)$.

From current searches at the Fermilab Tevatron $p\bar{p}$ collider we know that the masses of gluinos and squarks have to be above ~ 310 GeV and ~ 380 GeV, respectively [3]. The sensitivity beyond these limits at CMS has been studied in great detail [4]. It has been found that one of the most promising signatures for early discoveries at the LHC comprises a single high transverse momentum isolated lepton (electron or muon), several jets, and large missing transverse energy $\not\!\!E_T$. In the present note, we compare the strength of three different approaches to such an analysis: first, the traditional way of tuning a series of one-dimensional selection criteria on a simulation of SM backgrounds and signal. Second, we investigate the gains from exploiting multidimensional correlations by using multivariate analysis techniques, in our example boosted decision trees (BDT). Finally, we present the sensitivity of a model independent search for new physics to the specific case of mSUGRA considered here.

The three types of analysis presented differ vastly in their sensitivity to detector understanding and simulation of signals and backgrounds. At one extreme, BDT's make full use of practically the entire event information both for signal and background; at the other end, the model independent search does not rely at all on signal simulation. The obviously different sensitivity to systematic uncertainties and physics biases of the different approaches should allow us ultimately to make optimal use of the data and to avoid false discoveries, while keeping maximal sensitivity to new physics.

Several characteristic mSUGRA parameter sets are commonly used at CMS to benchmark the sensitivity of different search channels (see Fig. 1). In this note, we study in particular several "low mass" (LM) points, which - if realized in nature - are expected to be discovered at the LHC with a relative low amount of integrated luminosity:

	m_0	$m_{1/2}$	$\tan\beta$	$sgn(\mu)$	A_0	σ (LO) [pb]	σ (NLO) [pb]
LM1	60	250	10	+	0	45.9	61.1
LM2	185	350	35	+	0	7.80	10.5
LM4	210	285	10	+	0	20.5	27.7

The total cross sections including all SUSY channels have been calculated with Prospino 2 [5] using CTEQ6L1 and CTEQ6M PDF's [6] for the leading order and next-to-leading order cross sections, respectively. The point LM1 is close to the Tevatron exclusion region. Point LM2 has a relatively large value of $\tan\beta$, leading to a larger number of tau leptons in the final state. For point LM4, the decay of the $\tilde{\chi}_2^0$ into on-shell Z's is characteristic. For all three points, squark and gluino masses are below ~ 700 GeV, leading to relatively large total cross sections at 14 TeV center of mass energy. Also, $\tilde{q}\tilde{g}$ production is dominant, contributing about half of the total cross section.

The results are mostly presented for an assumed integrated luminosity of 1 fb^{-1} at a pp center of mass energy of 14 TeV. The Monte Carlo samples are from CSA07 [7], and use the 100 pb^{-1} scenario for detector alignment and calibration. The bulk of the CSA07 MC production corresponds to an integrated luminosity of 1 fb^{-1}, thus MC statistics are not negligible.

2 Common Aspects

2.1 Monte Carlo samples

The different mSUGRA LM samples are generated using the SoftSusy [8] program to calculate the mass spectrum, and the SUSY-Hit [9] package to calculate the decays. Finally the event generator Pythia [10] is used for the simulation of the parton shower. For each LM point defined above 100k events have been produced.

The different SM backgrounds are split in three "soups". The soup "Chowder" contains the W+jets, Z+jets and $t\bar{t}$+jets backgrounds, and uses the Alpgen [11] generator (together with Pythia for parton showering), which is more appropriate (than e.g. Pythia) to simulate the high transverse momentum jets accompanying W and Z bosons or

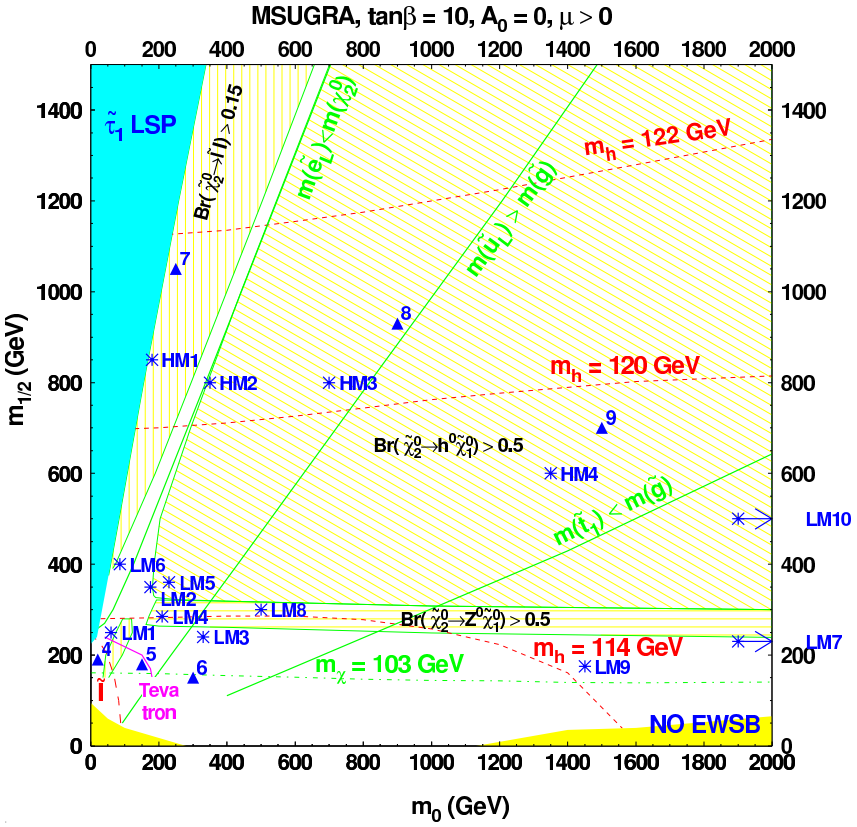


Figure 1: *Plane of mSUGRA parameters m_0 and $m_{1/2}$, with the CMS benchmark points LMx (low mass) and HMx (high mass) indicated [4].*

top quarks. The lepton enriched QCD, bottomonia and charmonia events, simulated by Pythia, are contained in the "Stew" soup. Finally, the "Gumbo" soup contains the QCD and photons+jets backgrounds, and minimum bias events. In the cut based and BDT analyses, where we focus on the μ + $\not\!\!E_T$ + jets final states, the analyses are performed directly on the optimized HLT-muon samples [12]. The three soups in the muon stream contain more than five million events, while the total number of generated events is more than 100 million in the CSA07 exercise. The MUSiC analysis ran both on electron and muon streams.

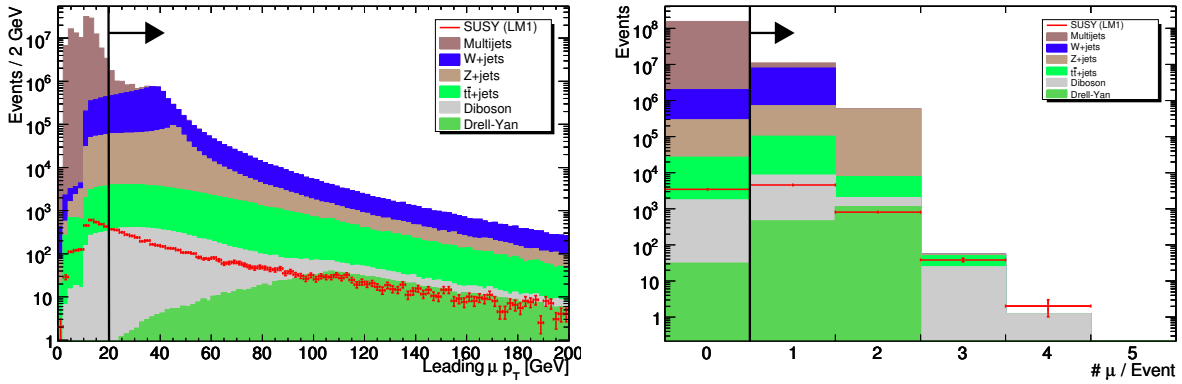
All events are simulated utilizing the full CMS detector simulation and event reconstruction framework CMSSW [13]. Using grid tools, the event selection, data reduction and analysis has been performed within the CMSSW framework in version 1.6.7. We use private code running locally on the federated Tier-2 Aachen/DESY for the last steps of the analyses.

3 Cut Based Analysis

3.1 Online selection

The online selection is critical for data taking under LHC conditions. The trigger consists of two steps, the Level 1 (L1) followed by the HLT (High Level Trigger) [14]. The muon stream samples used here are an admixture of different simulated trigger conditions. The trigger menu for such a HLT muon stream is, as described in [12]: HLT1MuonIso, HLT2MuonIso, HLT2MuonNonIso, HLT2MuonSameSign, HLTXMuonsJets, HLTXElectronMuon and HLTXElectronMuonRelaxed. We have chosen to focus on the subset containing the single and di-muon triggers, with and without isolation requirement. They correspond to the following technical trigger bits:

- HLT1MuonIso, single isolated μ with p_T threshold at 11 GeV
- HLT2MuonIso, double isolated μ, both p_T to greater than 3 GeV
- HLT2MuonNonIso, double relaxed μ, both p_T to greater than 3 GeV



(a) Transverse momentum p_T of the leading muon.

(b) Number of muons per event.

Figure 2: *Transverse momentum and number of muons for the expected signal and the dominant SM backgrounds.*

This specific trigger menu should be robust, even at the startup of the machine. These muon triggers present rates around 20 Hz for the HLT1MuonIso and 12 Hz for the HLT2MuonIso. The efficiency of such a trigger menu, after all cuts applied that will be described in the following paragraphs, is 93% for the LM1 point.

3.2 Offline selection

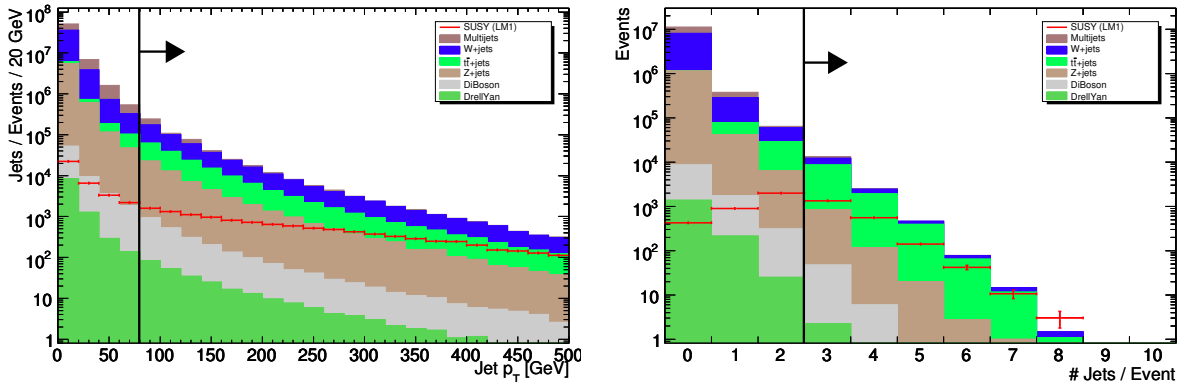
In the present analyses we are focusing on final states containing at least one muon, large missing transverse energy and several high transverse momentum jets. The muon signature is chosen due to its robustness against instrumental backgrounds and the Aachen expertise from detector construction and commissioning [15].

Lepton identification The muon object provides the trigger for the topology under study. Additional offline criteria are applied in order to refine the quality of the selected muons. We require at least 12 hits in the tracker and the muon system together, a normalized χ^2/N_{DOF} lower than 3 and $|\eta| < 2.1$. The two first conditions ensure that the muon is properly reconstructed, the latter restricts the acceptance to muons visible by the trigger. Additionally the sum of the transverse momenta of all tracks in a $\Delta R = 0.3$[1)] cone around the muon (excluding the muon track itself) has to be less than 6 GeV, mainly to reject non-isolated muons from heavy flavor decays. In Figure 2(a) the p_T distribution of all muons fulfilling these criteria in events passing the trigger is shown; only muons with $p_T \geq 20$ GeV are selected in order to match or exceed the expected trigger thresholds. Figure 2(b) shows the number of muons after all of the above cuts.

Preselection In R-parity conserving supersymmetry two neutralinos, which have a mass of the order of 100 GeV for the investigated LM points, will escape detection. Thus we require at least a missing transverse energy of 100 GeV at the preselection level.
To ensure properly reconstructed jets and a reasonably precise jet energy scale, only jets with $|\eta| \leq 2.5$ are considered. The iterative cone algorithm with a cone size of 0.5 is used and L2-L3 jet energy scale corrections are applied. A minimum of 10% of the jet energy is required to be reconstructed in the hadron calorimeter, in order to reject electrons and photons from the collection of jets.
The gluinos dominantly decay into a squark-quark-pair. Due to the relative small mass difference between gluino and squark, we expect an average jet arising from this decay to have a momentum of the order of 100 GeV. The squark is expected to decay into a quark and another sparticle, resulting in a hard jet of the order of a few hundred GeV, as the mass splitting between squarks and lighter sparticles is rather large. In total the dominant $\tilde{q}\tilde{q}$, $\tilde{q}\tilde{g}$ or $\tilde{g}\tilde{g}$ production should therefore lead to two hard and several softer jets. Consequently, we require at least three jets with $p_T \geq 80$ GeV (see Figure 3).

In the QCD multijet background or other processes without highly energetic invisible particles, missing transverse energy is expected to be caused by misreconstructed jets or muons. In order to suppress this background we restrict the angle between $\not\!\!E_T$and jets. The same motivation holds for poorly reconstructed muons or highly boosted W-bosons which may lead to a muon parallel to "fake" $\not\!\!E_T$.

[1)] $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$



(a) Transverse momentum p_T of all jets.

(b) Number of jets per event.

Figure 3: *Jet transverse momentum and number of jets per event for signal and background events.*

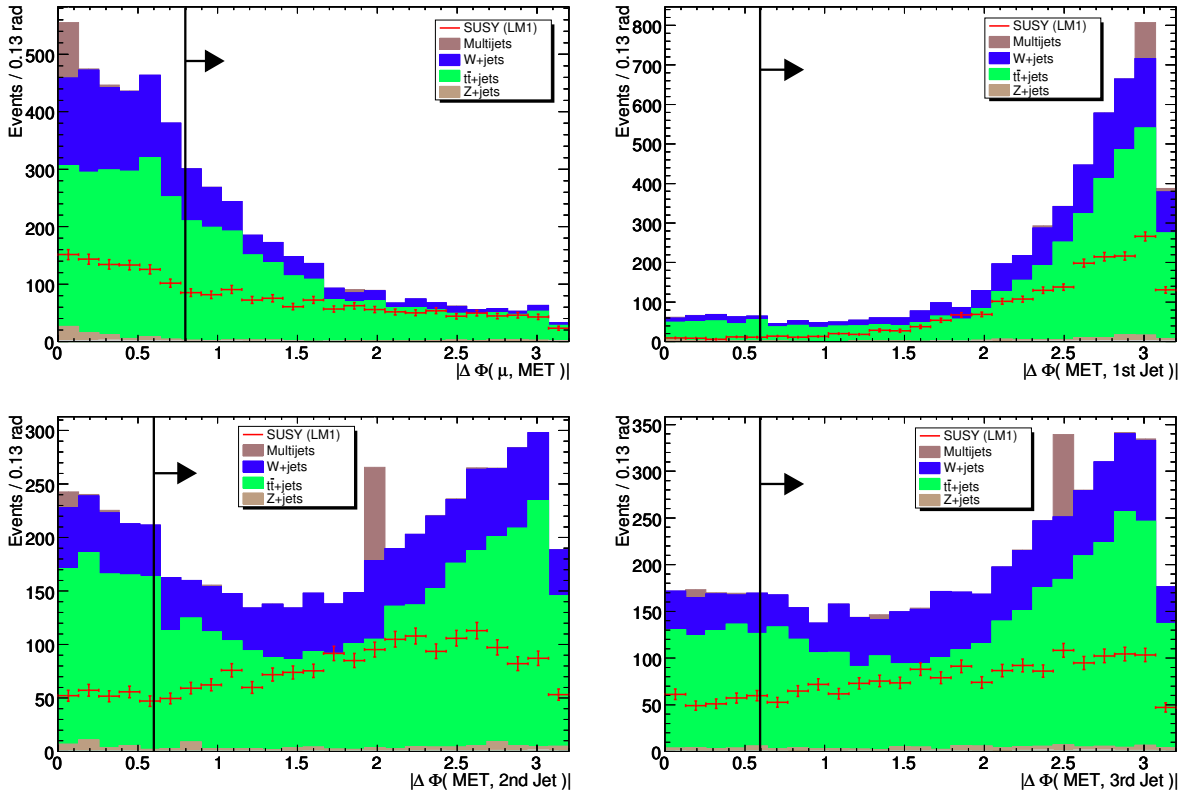


Figure 4: *Angular difference between the leading muon and the transverse missing energy as well as the angular difference between the transverse energy and the three leading jets.*

Optimization strategy During the following steps the optimization is performed on LM1, but the resulting cuts are applied to the other LM points. It turns out that there is no gain in optimization of the cuts with respect to a certain LM point.

To get a quick and reasonable estimate of the expected significance, we calculate N_S/σ_B, with the error on the background σ_B as the quadratic sum of the systematic uncertainties (see section 3.3) and the Poisson uncertainty $\sigma_P = \sqrt{N_B}$. Using this estimator, we optimize (in this order) the angular differences $|\Delta\phi(\mu, \not{E}_T)|$ and $|\Delta\phi(\not{E}_T, jet_{1,2,3})|$. Another reoptimization is performed on each of these cuts while keeping the other three fixed, starting over each time a significant change occurred (see Figure 4).

In the next step we, in a way like above, look at the jet momenta p_T^{1stjet}, p_T^{2ndjet} and the missing transverse energy $\not{E}_T$ (see Figure 5 and 6).

Finally, each of the above cuts is re-evaluated after the application of all other cuts.

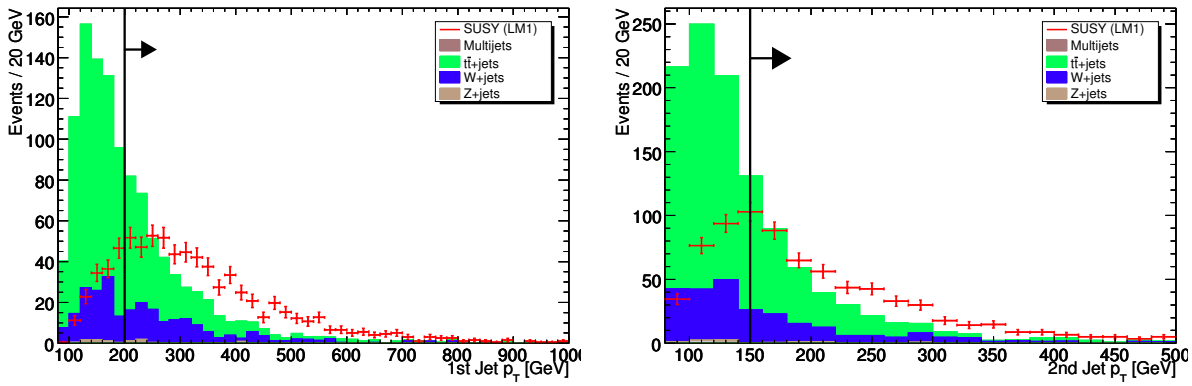


Figure 5: *Transverse momentum of two leading jets.*

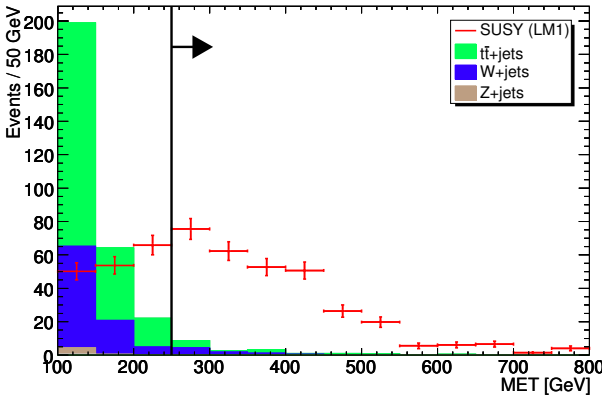


Figure 6: *Missing transverse energy.*

With this procedure the following cuts are determined:

- $|\Delta\phi(\mu, \not{E}_T)| \geq 0.8$
- $|\Delta\phi(\not{E}_T, jet_{1,2,3})| \geq 0.6$
- $p_T^{1stjet} \geq 200$ GeV
- $p_T^{2ndjet} \geq 150$ GeV
- $\not{E}_T \geq 250$ GeV

3.3 Systematic uncertainties

The following systematic uncertainties are considered. It should be noted that the objective of the present analysis is a comparison of the sensitivities of different analysis techniques. A complete discussion of systematic uncertainties and their derivation based on data is beyond the scope of this note.

Luminosity 5%, fully correlated for all samples.

Cross sections 10% (including uncertainties of parton density functions), assumed to be fully correlated for samples of one type (for example all jet multiplicity bins of the $t\bar{t}$ samples).

Jet energy scale To estimate the effect of the uncertainty of the jet energy, the energy of all jets is shifted by a certain percentage. The energy change of the jets with $p_T > 10$ GeV is then propagated to the $\not{E}_T$. The variation is performed twice, once by $\pm 5\%$ and once by $\pm 10\%$ of the jet energy. Thus, one obtains four deviations of the number of events passing all cuts: $\Delta N^{\pm 5\%}$ and $\Delta N^{\pm 10\%}$. With these deviations we calculate the mean uncertainty of the number of events:

$$\Delta N_{JES} = 1/4 \cdot (|\Delta N^{+5\%}| + |\Delta N^{-5\%}| + |\Delta N^{+10\%}|/2 + |\Delta N^{-10\%}|/2). \quad (1)$$

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	8.3	8.6	1.7	2.1	–
Z+jets	0.095	1.2	1.0	0.28	–
$t\bar{t}$	9.2	9.2	1.9	5.0	–
Diboson	0.25	0.44	0.19	0.097	–
Drell-Yan	0.013	0.058	0.043	0.014	–
QCD	0.0	9.8	6.0	2.2	–
Background sum	18	29	6.6	9.3	–
LM1	316	–	13	64	33
LM2	97	–	3.1	16	14
LM4	203	–	7.5	39	24

Table 1: *Numbers of events after all cuts with uncertainties scaled to an integrated luminosity of* 1 fb^{-1}. *The column labeled UL contains numbers of events after implementing an upper limit in samples with zero selected events, as described in the text.*

Depending on the dataset this uncertainty is between 5% and 50%. Datasets without any events passing the cuts are assigned a default 20% uncertainty. These uncertainties are assumed to be correlated for all samples.

In addition a *statistical* uncertainty $w \cdot \sqrt{N}$ is assigned to each event, depending on the event weight w and the total number of events N of the simulated dataset.

3.4 Results

The results include the following Standard Model background samples:

- single boson production W and Z
- diboson production WW, WZ and ZZ
- top quark pair production $t\bar{t}$
- QCD multijets with $\hat{p}_T > 300$ GeV (transverse momentum of the hard interaction in its rest frame)

Samples which contain zero events after the application of all cuts are conservatively set to a Poisson 68% upper limit of 1.15 events before scaling them to an integrated luminosity of 1 fb^{-1}. The estimation of an upper limit is most important for the QCD multijet background, since most other samples have been simulated with statistics corresponding to an integrated luminosity of 1 fb^{-1}.

For the W+0 jets MC sample, where insufficient MC statistics is available and no events are selected, we do not apply an upper limit: In the W+1 jet sample with sufficient statistics, no events are selected, and we can therefore conclude that the W+0 jet background is negligible.

As one can see in Table 2, there are no QCD events left after the angular correlation cuts. The following three hard cuts on jets and $\not{E}_T$ are expected to further supress this background by several orders of magnitude. Thus we assume QCD background to deliver a negligible contribution to the total background. Nevertheless we conservatively apply an upper limit on all QCD samples with $\hat{p}_T \geq 300$ GeV: In theses samples the sum of transverse momenta $\sum p_T$ of the selected objects is of the order of the $\hat{p}_T$ of the hard interaction. Ultimately, one would estimate the QCD background from real data, as e.g. outlined in the MUSiC analysis section.

The selected numbers of events are shown in Table 1. The dominant backgrounds remaining after the optimized selection are W+jets and $t\bar{t}$. The QCD background contributes mainly due to the lack of MC statistics. The expected statistical significance of the SUSY signals based on Poisson probabilities[2)] (thus neglecting systematic uncertainties and MC statistics) are also given in Table 1, and indicate the discovery reach with 1 fb^{-1} of integrated luminosity.

[2)] $P_{\text{Poisson}} = \int_{N_S}^{\infty} Poisson(N_B)$

Cut	Background sample								Signal sample		
	W+ 0 jets	W+jets	Z+jets	$t\bar{t}$	Diboson	DY	QCD	All BG	LM1	LM2	LM4
NoCuts	$5.07e+07$	$1.48e+07$	$6.58e+06$	$8.28e+05$	$1.08e+05$	$1.75e+03$	$8.03e+13$	$8.03e+13$	$6.11e+04$	$1.05e+04$	$2.77e+04$
HLT	$6.67e+06$	$2.22e+06$	$1.44e+06$	$1.24e+05$	$1.07e+04$	$1.65e+03$	$2.21e+08$	$1.05e+07$	$8.88e+03$	$1.33e+03$	$3.33e+03$
Efficiency	13%	15%	22%	15%	9.9%	94%	0.0%	0.0%	15%	13%	12%
μ acceptance	$5.39e+06$	$1.85e+06$	$1.19e+06$	$1.04e+05$	$9.20e+03$	$1.63e+03$	$4.17e+06$	$8.56e+06$	$6.08e+03$	$9.77e+02$	$2.89e+03$
Efficiency	81%	84%	83%	84%	86%	99%	1.9%	82%	69%	73%	87%
μ identification	$5.33e+06$	$1.82e+06$	$1.18e+06$	$9.86e+04$	$9.04e+03$	$1.62e+03$	$3.17e+06$	$8.44e+06$	$5.42e+03$	$8.44e+02$	$2.61e+03$
Efficiency	99%	98%	99%	95%	98%	99%	76%	99%	89%	86%	91%
3 Jets, $p_T \geq 80$ GeV	$0.00e+00$	$4.13e+03$	$9.27e+02$	$1.01e+04$	$5.13e+01$	$2.74e+00$	$1.28e+03$	$1.53e+04$	$2.09e+03$	$4.42e+02$	$1.39e+03$
Efficiency	0.0%	0.23%	0.078%	10%	0.57%	0.17%	0.04%	0.18%	39%	52%	53%
$\not\!\!E_T \geq 100$ GeV	$0.00e+00$	$1.36e+03$	$9.19e+01$	$3.43e+03$	$1.79e+01$	$3.90e-01$	$1.13e+02$	$4.93e+03$	$1.91e+03$	$4.18e+02$	$1.24e+03$
Efficiency	0.0%	33%	9.9%	34%	35%	14%	8.8%	32%	91%	94%	89%
Angular correlations	$0.00e+00$	$2.47e+02$	$8.43e+00$	$8.62e+02$	$3.56e+00$	$1.43e-01$	$0.00e+00$	$1.12e+03$	$7.76e+02$	$1.80e+02$	$5.04e+02$
Efficiency	0.0%	18%	9.2%	25%	20%	37%	0.0%	23%	41%	43%	41%
1st Jet $p_T \geq 200$ GeV	$0.00e+00$	$1.31e+02$	$5.02e+00$	$3.10e+02$	$1.74e+00$	$1.22e-01$	$0.00e+00$	$4.47e+02$	$6.24e+02$	$1.57e+02$	$4.21e+02$
Efficiency	0.0%	53%	60%	36%	49%	85%	0.0%	40%	80%	88%	84%
2nd Jet $p_T \geq 150$ GeV	$0.00e+00$	$9.42e+01$	$4.26e+00$	$2.04e+02$	$9.11e-01$	$6.78e-02$	$0.00e+00$	$3.03e+02$	$4.85e+02$	$1.29e+02$	$3.26e+02$
Efficiency	0.0%	72%	85%	66%	52%	56%	0.0%	68%	78%	82%	77%
$\not\!\!E_T \geq 250$ GeV	$0.00e+00$	$8.33e+00$	$9.49e-02$	$9.16e+00$	$2.48e-01$	$1.25e-02$	$0.00e+00$	$1.79e+01$	$3.16e+02$	$9.73e+01$	$2.03e+02$
Efficiency	0.0%	8.8%	2.2%	4.5%	27%	18%	0.0%	5.9%	65%	75%	62%

Table 2: *Expected number of events at* $1\ fb^{-1}$ *after each cut. Efficiencies are relative to the number of events passing the previous cut.*

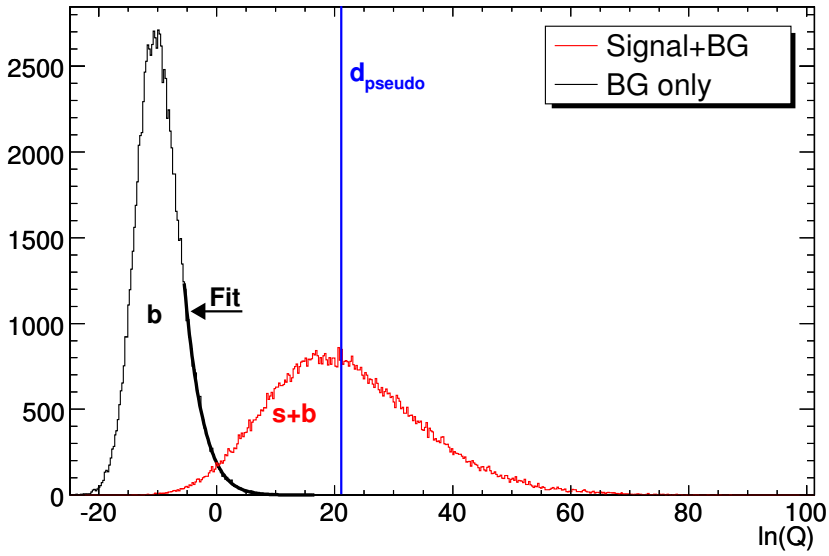


Figure 7: *$ln(Q)$ distribution of the background-only hypothesis(b) and the signal+background hypothesis ($s+b$), respectively. The separation of both distributions determines the discrimination power between the two hypotheses. An $ln(Q)$ for data (d) close to the background only curve states an agreement with the background-only hypothesis. On the opposite a large $ln(Q)$ for data (d) close to the $s+b$ curve would point to some new physics not consistent with the background only hypothesis.*

3.5 Statistical interpretation

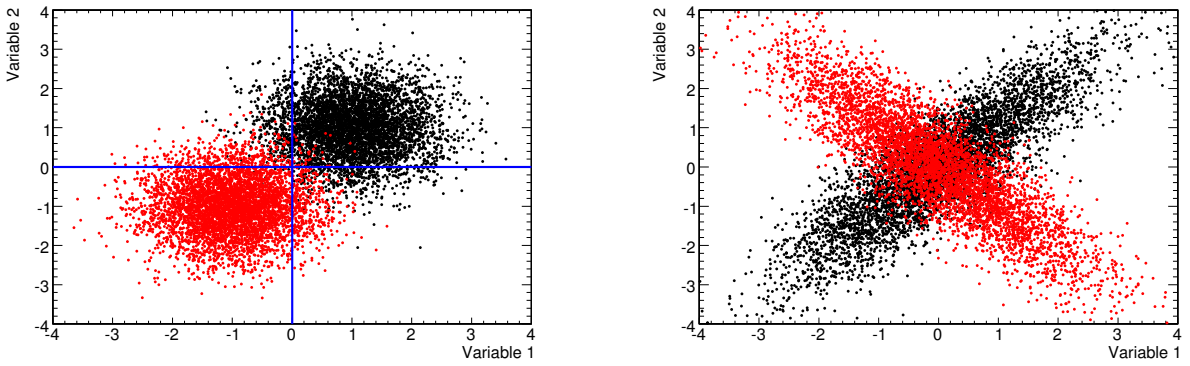
In oder to determine the expected significance of an observation, a modified frequentist approach [16, 17] is applied. If the expected number of signal events s and the number of background events b are known without any systematic uncertainties, one can calculate, for a measured number of events d, the log-likelihood-ratio $\ln(Q)$. Here Q is the probability of the signal+background hypothesis to yield the measured number of events, divided by the probability of the background-only hypothesis, assuming Poisson distributions for both hypotheses:

$$Q = \frac{(s+b)^d \cdot e^{-(s+b)}}{b^d \cdot e^{-b}} \tag{2}$$

The higher the value of $\ln(Q)$, the more *signal-like* the measured data are.

In case of systematic uncertainties on the expected numbers of events, this approach has to be extended: Pseudo-experiments are performed, in which the expected number of background events b and signal+background events $s+b$ are varied according to their statistical and systematic uncertainties, while taking correlations into account. A bunch of pseudo-experiments is performed assuming d to be background like, i.e. $d = Poisson(b)$ resulting in the expected background-only $\ln(Q)$ distribution (see b-curve in Figure 7). In addition pseudo-experiments are generated in which d is diced according to expected number of signal+background events, i.e. $d = Poisson(s+b)$ (see $s+b$-curve in Figure 7). The integral of both distribution is normalized to one. Finally, these two distributions have to be compared to the $\ln(Q)$ value of the measured events d_{meas}. Data are considered incompatible with the background-only hypothesis, if the $\ln(Q)$ value of the measured events d_{meas} is considerably larger than the $\ln(Q)$ values of the background only distribution, hence the probability to find a background only pseudo-experiment with a larger $\ln(Q)$ is sufficiently small. This probability, commonly denoted as $1 - CL_B$, corresponds to the fraction of the b-curve in Figure. 7 at the right side of the d-line. A probability of smaller than $2.8 \cdot 10^{-7}$, which corresponds to a Gaussian 5σ single-sided deviation, is commonly declared as a discovery. Since this is a MC only study without any measured data, the expected discovery potential for SUSY can be determined assuming that the data correspond to the median $\ln(Q)$ of the signal plus background curve (see d_{Pseudo}-line in Figure. 7).
In order to test the sensitivity for a 5σ deviation from the background only hypothesis, this approach would need at least 10^9 pseudo-experiments to have sufficient statistics in the region of interest. This can easily take days of computing time. Therefore we fit a Gaussian to the tail of the background-only distribution and extrapolate it to regions with very low probability close to the signal plus background distribution (see thick line in Figure 7). This extrapolation is proved to be prone to statistical fluctuations and tend to overestimate the significance, thus values significantly larger than 5σ have to be taken with a grain of salt. The calculated probability CL_B being a p-value can be translated via a one-sided Gaussian into standard deviations.

Table 3 summarizes the expected significances for the three investigated LM points. Both LM1 and LM4 are in the discovery reach within the first 1 fb^{-1} or earlier. It can be seen that the increase of the significance with the



(a) Usable correlations (b) Correlations lost

Figure 8: *Distribution of a toy signal (red) and background (black) with and without usable correlations. Trivial cuts are shown in the usable case.*

addition of more data does not follow the expectation of a search dominated purely by statistical uncertainties. This is due to the impact of systematic uncertainties, but also due to the limited MC statistics. The benchmark point LM2 requires a larger amount of integrated luminosity for discovery, reduced systematic uncertainties, or alternative search channels (see e.g. section 5).

LM-Point	10 pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	3.3	4.9	7.1	10
LM2	1.9	2.6	3.2	3.9
LM4	2.6	3.9	4.8	7

Table 3: *Modified Frequentist significances expected for the different SUSY benchmark points and different amounts of integrated luminosity.*

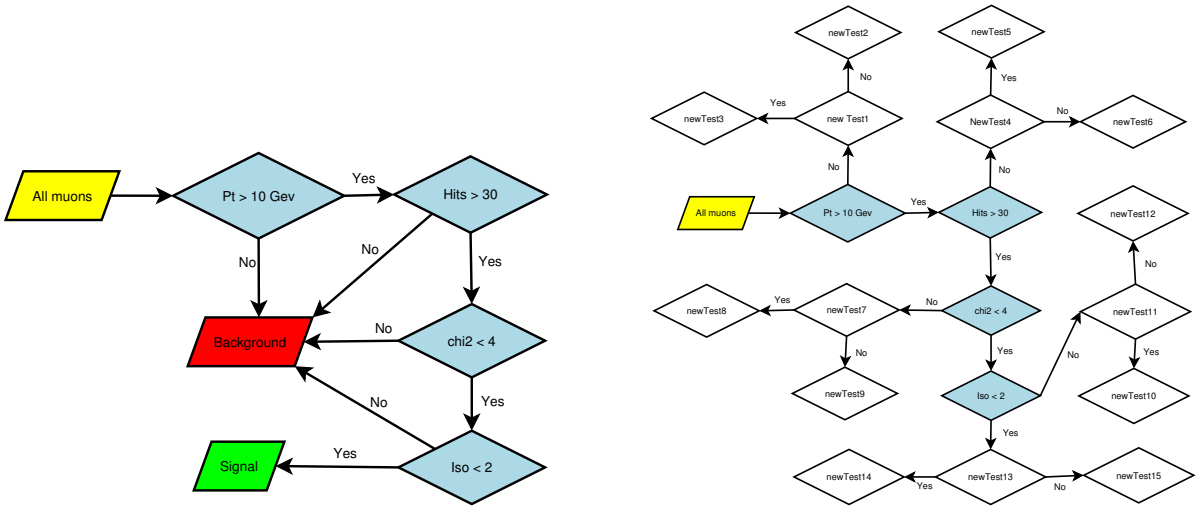
4 Multivariate Analysis Technique: Boosted Decision Trees

Analyses based on one dimensional cuts, like the one presented in the previous section, are straightforward, robust, and easily enjoy acceptance in the community. However, they have their limitations. In a situation like in Figure 8(a) the shown cuts are quite easy to determine. In Figure 8(b) the variables are correlated differently for signal and background separated, but this correlation is lost for signal and background together and can not be exploited using the chosen variables with one dimensional cuts. In this artificial example it is rather simple to define two new variables honoring the correlations, effectively rotating Figure 8(b) by $45°$. Confronted with real world problems, this will quickly become excessively complicated, as there is usually a rather high number of possible variables, resulting in many combinations to be tested for correlations. The situation is even worse if there are correlations between three or more variables. In addition all correlations would have to be reinvestigated for different model parameters.

Illustrating an exemplary cut based analysis as a decision tree (see Figure 9(a)), one can see an event being classified as signal, if and only if it passes all cuts. Mathematically speaking, in the space defined by all variables, exactly one region is selected, whose borders are perpendicular to the variable axes. Ideally this region is located and sized such that the separation of signal and background is maximized. However, depending on the physics involved, the best region may be irregularly shaped and being separated in numerous unconnected parts. One way to get near this best region is to extend the simple decision tree of Figure 9(a) to a full fledged one like shown in Figure 9(b). The leaves of this tree can then be classified as either signal or background, thus selecting more than one region. These regions are still rectangular, but can approximate any shape, if they are small and numerous enough.
In this analysis the ROOT [18] based multivariate analysis framework TMVA [19] in version 3.8.14 is used.

4.1 Training

The number of possible trees with all combinations of cut variables and values is basically unlimited, so it is obviously impossible to search all trees to find the best separating one. Thus the so called *growing* of the tree is done in a rather straightforward way: Using all events, the variable and the corresponding cut value resulting in the best separation is determined. The events are then split into two subgroups, one containing all events which pass



(a) Conventional analysis

(b) Extended analysis

Figure 9: *One dimensional cuts represented as a decision tree.*

this cut and the second containing all events which do not pass. For each of these two subgroups the procedure is repeated, resulting in now four subgroups. This splitting is then repeated recursively until the number of events in a subgroup drops under a certain number[3] N_0, then the splitting of this subgroup is stopped.

This is done for two reasons:

- Avoid overtraining: If the splitting is performed until the events in each leaf are either signal or background, the separation is perfect on the training sample, but on any other sample it is most likely deteriorated.
- Manageable tree size: A sample of N_{training} events will lead to a tree with at most $2^{N_{\text{training}}}$ splitting nodes.

In TMVA N_0 is set to

$$\max(20, N_{\text{training}}/(N^2_{\text{variables}}/10)) \tag{3}$$

which leads to a maximum size of $10N^2_{\text{variables}}$ splitting nodes and a sufficient chance for every variable to be used for splitting.

The best separating variable and its value is determined by trying each variable at $N_{\text{Cuts}} = 20$ values, equidistant between the minimum and maximum of this variable in all events in this node, and calculating the weighted average (weighted with the number of events in the nodes) of a *separation index* for both daughter nodes. Here the *Gini index* $p \cdot (1-p)$ with *purity* $p = N_S/(N_S + N_B)$ is used, which is at its maximum for perfectly mixed samples ($p = 0.5$) and drops to zero for samples which consist of signal or background only. Then the variable-value pair with the smallest separation index is used.

4.2 Boosting

A single decision tree grown like above is a rather weak classifier, as it is quite likely affected by statistical fluctuations of the sample. Furthermore it yields a binary yes-no decision whether it considers an event to be signal or background, and gives no information how signal- or background-like the event might be.

One way to overcome these drawbacks is *boosting*, which can be applied to most MVA classifiers. A common boosting algorithm is *AdaBoost* (adaptive boosting), where the weights of events misclassified by one decision tree are multiplied by a boost weight α, before an additional tree is grown using the modified sample. The boost weight is determined from the misclassification rate ϵ_{err} of the previous tree:

$$\alpha = \frac{1 - \epsilon_{err}}{\epsilon_{err}} \tag{4}$$

[3] Actually the splitting is not performed if the number of events in one daughter node drops below this number.

Thus events misclassified by an otherwise well performing tree have the highest impact during the growing of the next tree. The final response of this *forest* of decision trees is then calculated via

$$response = \sum \ln(\alpha_i) h_i \tag{5}$$

with the individual responses h_i of each tree, which is -1 for background and 1 for signal. The combined response will then be somewhere between -1 and 1, showing the signal-likeness the forest assigns to this event.

4.3 Pruning

A tree will quite likely contain splitting nodes, especially near its leaves, which are statistically insignificant. These nodes will increase the size of the tree without a significant performance gain and may lead to overtraining, as their cut values are just caused by statistical fluctuations. So they can be removed, if their gain in performance is beneath a certain threshold, depending on the number of events left. It is recommendable to first grow a tree to its maximum size and then cut it recursively back, starting at the leaves, because a cut may seem insignificant on its own, but lead to improved performance in subsequent cuts[4)].

Experience so far for the present application shows no performance increase on independent test samples, thus we avoided the additional work for tuning the pruning threshold. But this has to be further investigated.

4.4 Workflow

Two partially different approaches are followed:

Approach 1 Utilizing $\not\!\!E_T$, the highest p_T (leading) muon, and the momentum p_T of the three leading jets.

Approach 2 Utilizing $\not\!\!E_T$, the leading muon p_T, the leading jet p_T, and the number of jets and muons.

In both cases the following cleaning cuts are used:

Muons

- $p_T \geq 20$ GeV
- $\sum p_T^{\text{tracks}} \leq 6$ GeV in a $\Delta R = 0.3$ cone
- $\chi^2/N_{\text{DOF}} \leq 3$
- $N_{Hits} \geq 12$
- $|\eta| \leq 2.1$

Jets

- $p_T \geq 50$ GeV
- $|\eta| \leq 2.5$
- At least 10% hadronic energy

In both approaches the following attributes are used to train the BDTs:

Muons

- transverse momentum p_T
- Pseudorapidity $|\eta|$
- Isolation: $\sum p_T^{tracks}$, $\sum E_T^{calo}$, N_{Tracks} ($\Delta R \leq 0.3$ cone)
- Calorimeter compatibility

Jets

[4)] This will e.g. be the case in a situation like in Figure 8(b).

- transverse energy E_T
- Pseudorapidity $|\eta|$

$\not\!\!E_T$

- missing transverse energy $\not\!\!E_T$
- scalar sum of all transverse energy $\sum |E_T|$

Muon and jet

- invariant mass
- $|\Delta\phi|$

Muon and $\not\!\!E_T$, jet and $\not\!\!E_T$

- $|\Delta\phi|$

If no pruning is performed, the BDTs are most likely overtrained on the set of events used for training, resulting in a unphysical high separation power. Thus the events used for training and validation must be statistically independent. Consequently we separate the simulated events in two sets: One used for training, one for validation.

On the first third of the MC samples one set of BDTs is trained for each of the three LM points and both approaches, resulting in set of six BDTs. The last two thirds of the MC samples are then used to determine the minimal response, which yields the best significance (see Figure 10), and to calculate the significance. The optimal cut on the BDT response is determined by optimizing N_S/σ_B as described in section 3.2. In analogy to the cut-based analysis the significance is determined using the modified frequentist approach including systematic uncertainties as explained in section 3.5.

4.5 Results

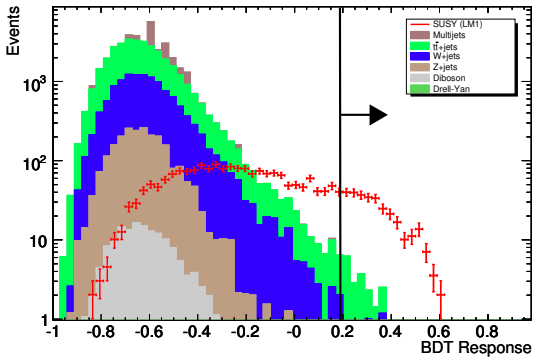
As in the previous cut based analysis we include the following Standard Model background samples:

- single boson production W and Z,
- diboson production (WW, ZZ, WZ)
- top quark pair production $t\bar{t}$ and
- QCD multijets with $\hat{p}_T > 300\ GeV$ (transverse momentum of the hard interaction in its rest frame).

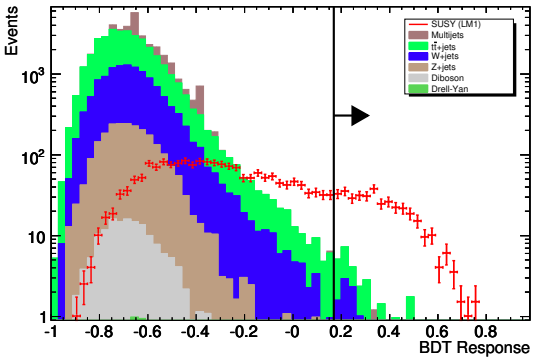
Samples with zero events passing all cuts are forced to contain 1.15 events (Poisson 68% upper limit) before normalization to an integrated luminosity of 1 fb^{-1} (see section 3 for details). The systematic uncertainties are taken from section 3.3, except for the jet energy scale uncertainty, which is determined again. For the results of the six BDTs see Tables 4 to 9. All six BDTs, for the two approaches and trained on the three different SUSY benchmark points, give similar numbers for background and signal expectation, and consequently for the expected statistical signficance. Similar to the cut based analysis, the dominant backgrounds remaining are $t\bar{t}$ and W+jets.

4.6 Interpretation

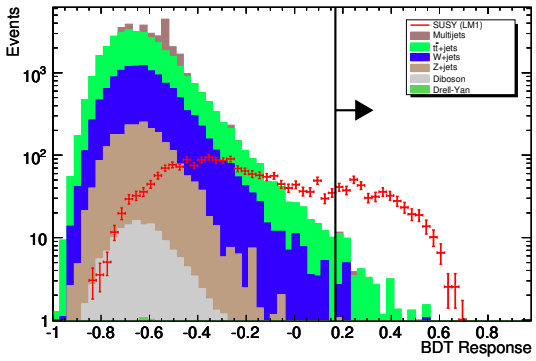
Using the same approach as in Section 3.5, we obtain the expected significances shown in Tables 10 to 15. There is a modest improvement compared to the cut based analysis in terms of the expected significances, and correspondingly in the amount of required luminosity for a discovery. As there is more improvement for smaller luminosities, the selection performance of the BDTs is better than of the conventional analysis, but they are more affected by systematic uncertainties. This is expected, as the systematic effects are not taken into account during the training, only during the optimization of the minimal response. With the current set of variables and training parameters, there are only small differences between the BDTs trained specifically for a certain LM point.



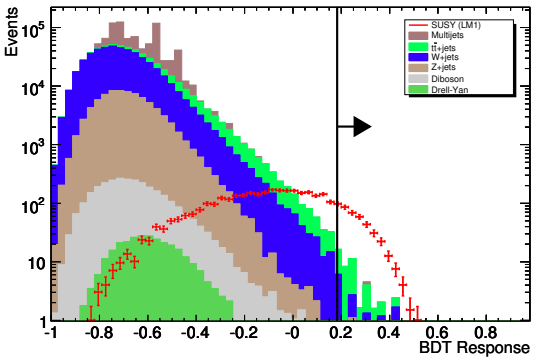
(a) Approach 1; Trained on LM1



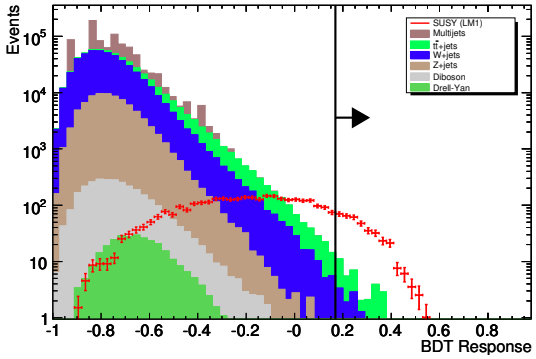
(b) Approach 1; Trained on LM2



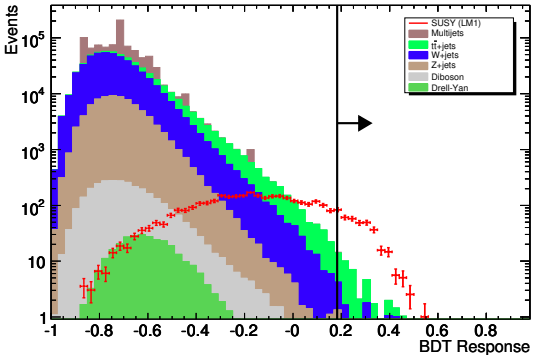
(c) Approach 1; Trained on LM4



(d) Approach 2; Trained on LM1



(e) Approach 2; Trained on LM2



(f) Approach 2; Trained on LM4

Figure 10: *BDT response for signal and background for* 1 fb^{-1}.

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	4.6	5.0	2.3	0.97	–
Z+jets	0.03	1.7	1.5	0.39	–
$t\bar{t}$	29	29	4.4	8.4	–
Diboson	0.25	0.4	0.22	0.16	–
Drell-Yan	0.004	0.072	0.064	0.016	–
QCD	0.0	15	8.9	3.3	–
Background sum	34	51	10	13	–
LM1	465	–	19	70	38
LM2	135	–	4.5	17	15
LM4	263	–	10	41	25

Table 4: *Number of events in approach 1, trained on LM1, scaled to* 1 fb^{-1}.

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	13	13	4.2	4.6	–
Z+jets	0.14	1.8	1.5	0.42	–
$t\bar{t}$	29	29	4.3	11	–
Diboson	0.12	0.42	0.23	0.2	–
Drell-Yan	0.0	0.073	0.064	0.016	–
QCD	0.83	15	8.9	3.4	–
Background sum	43	59	11	19	–
LM1	547	–	20	93	38
LM2	172	–	5.0	25	17
LM4	362	–	12	61	31

Table 5: *Number of events in approach 1, trained on LM2, scaled to* $1\ fb^{-1}$.

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	7.4	7.8	3.2	3.1	–
Z+jets	0.061	1.8	1.5	0.45	–
$t\bar{t}$	37	37	4.9	13	–
Diboson	0.25	0.54	0.26	0.2	–
Drell-Yan	0.001	0.073	0.064	0.017	–
QCD	0.61	15	8.9	3.6	–
Background sum	45	62	11	20	–
LM1	572	–	21	95	38
LM2	169	–	5.0	23	17
LM4	396	–	13	64	33

Table 6: *Number of events in approach 1, trained on LM4, scaled to* $1\ fb^{-1}$.

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	12	12	3.6	5.9	–
Z+jets	1.8	1.8	1.6	0.89	–
$t\bar{t}$	41	41	5.1	12	–
Diboson	0.5	0.65	0.28	0.22	–
Drell-Yan	0.062	0.063	0.059	0.013	–
QCD	0.83	15	8.9	3.3	–
Background sum	56	70	11	21	–
LM1	572	–	21	89	38
LM2	141	–	4.6	17	14
LM4	305	–	11	44	26

Table 7: *Number of events in approach 2, trained on LM1, scaled to* $1\ fb^{-1}$.

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	14	14	3.5	2.6	–
Z+jets	0.06	1.8	1.5	0.42	–
$t\bar{t}$	37	37	4.9	13	–
Diboson	0.37	0.52	0.26	0.14	–
Drell-Yan	0.012	0.08	0.064	0.02	–
QCD	0.0	15	8.9	3.3	–
Background sum	52	69	11	18	–
LM1	566	–	21	101	38
LM2	181	–	5.2	25	17
LM4	391	–	13	66	31

Table 8: *Number of events in approach 2, trained on LM2, scaled to* 1 fb^{-1}.

Sample	# Events @ 1 fb^{-1}		σ_{stat}	σ_{sys}	$\sigma_{Poisson}$ @ 1 fb^{-1}
	NL	UL			
W+jets	12	12	3.9	3.8	–
Z+jets	0.24	1.9	1.5	0.5	–
$t\bar{t}$	43	43	5.3	15	–
Diboson	0.37	0.52	0.26	0.2	–
Drell-Yan	0.007	0.075	0.064	0.021	–
QCD	0.83	15	8.9	3.5	–
Background sum	57	73	11	22	–
LM1	509	–	20	92	38
LM2	153	–	4.8	20	14
LM4	390	–	13	61	31

Table 9: *Number of events in approach 2, trained on LM4, scaled to* 1 fb^{-1}.

Sample	10 pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	3.7	6.5	8.5	11
LM2	2.0	2.8	3.3	4.0
LM4	2.7	4.1	5.4	6.6

Table 10: *Expected significance of approach 1, trained on LM1*

Sample	10 pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	3.7	6.0	7.5	8.5
LM2	2.1	2.8	3.3	3.7
LM4	3.1	4.7	5.4	6.1

Table 11: *Expected significance of approach 1, trained on LM2*

Sample	10 pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	4	6.3	7.7	9.1
LM2	2.1	2.8	3.2	3.5
LM4	3.2	4.9	5.8	6.6

Table 12: *Expected significance of approach 1, trained on LM4*

Sample	10‘pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	3.9	6.6	8.2	10
LM2	1.9	2.5	2.9	3.4
LM4	2.7	3.8	5.0	5.8

Table 13: *Expected significance of approach 2, trained on LM1*

Sample	10‘pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	3.8	6.6	8.0	9.8
LM2	2.1	3.0	3.4	4.1
LM4	3.1	5.3	6.3	7.6

Table 14: *Expected significance of approach 2, trained on LM2*

Sample	10‘pb^{-1}	40 pb^{-1}	100 pb^{-1}	1 fb^{-1}
LM1	3.5	5.5	6.4	7.4
LM2	2.0	2.5	2.9	3.2
LM4	3.1	4.4	5.4	6.1

Table 15: *Expected significance of approach 2, trained on LM4*

5 MUSiC – A Model Unspecific Search for New Physics in CMS

In the context of a generic search for New Physics in early LHC data a special algorithm called "MUSiC" has been developed. A dedicated note documenting all the details of this analysis Ansatz is currently in preparation. In the scope of this note we will only highlight the basic features and the algorithm. We focus on the comparison with the previous analysis strategies and discuss the sensitivity to selected SUSY signals. More models beyond the Standard Model will be discussed in a forthcoming MUSiC note.

5.1 The concept of model independent searches

Model independent searches aim to cover a large fraction of data, systematically scanning them for deviations from the Standard Model. Therefore the selection cuts are not optimized for any expected new physics signal; however, the quality of the measurement is ensured by selecting well measured and well understood physics objects such as isolated high-p_T leptons. Similar strategies have already been applied successfully at accelerator experiments, see e.g. [20–23].
The motivation for such an analysis strategy is evident: The LHC will enter unkown territory and there are multiple reasons why new physics is expected to appear. Unlike in experiments of the past there is an almost infinite number of predictions from theory of how exactly these new physics models will look like. Following the saying "expect the unexpected", a model independent search tries to cover a wide range of the phase space and is not limited to a specific topology. In this way it should be sensitive to surprises with spectacular final states such as mini black holes, give a consistent picture of the various channels where a SUSY signal could contribute, or quickly discover discrepancies caused by detector effects or effects not properly accounted for in the Monte Carlo simulation.
However, such an ambitious strategy also has its drawbacks: For some signals it could be less sensitive than a targeted analysis. In addition, since a variety of final states is looked at, one has to rely more on the background predictions made by Monte Carlo generators. While some SM backgrounds may be reasonably well predicted in particular thanks to the availability of Monte Carlo programs based on matrix element calculations like ALPGEN or MC@NLO, there are others where this is likely not to be the case. As one obvious example, the QCD multijet background will have to be estimated from the data in such a generic search.

In order to have a well defined trigger stream and in order to reduce the QCD backgound we restrict the analyzed phase space by requiring:

Topology cut: Events with $>= 1$ lepton (electron or muon)

All events which fulfill this requirement are then classified into so called event classes. Each event class is defined by the amount of physics objects in the event, e.g. 1μ 3jet. We consider *exclusive* and *inclusive* event classes, where in the exclusive case the exact number of particles is required, e.g. $1\mu\ 0e\ 0\gamma$ 2jet $\not\!\!E_T$. The inclusive classes which require only a minimal number of particles, e.g. 1μ 3jet $+ X$, so at least one muon and 3 jets, are denoted with a suffix *+X*. Given the complex decay chains of e.g. SUSY events such inclusive classes might be useful.
We consider the following physics objects measured by the CMS detector:

Physics objects: Muons (μ), electrons (e), photons (γ), hadronic jets (jet) and missing transverse energy ($\not\!\!E_T$)

Given the complexity of the analysis, the strategy will be to focus on well measured and well understood objects (high p_T, central η), even if this implies some loss in efficiency. Selection cuts are desired to remain as simple as possible. We note that similar stategies are useful for any start-up physics study.

5.2 Technical setup

Standard CMS software is used in order to process the simulated samples and to reconstruct the physics objects, using version CMSSW_1_6_8 [13]. Various official reconstruction tools are used and will be discussed below in the context of selection cuts. All samples are generated with the full detector simulation and originate from the MC production during the Computing, Software and Analysis Challenge 2008. The complete soups Chowder, Stew and Gumbo have been processed, using the susyElectron skim (logical "OR" of various electron HLT triggers) and the susyMuon skim ("OR" of various muon HLT triggers).

In addition, various CSA07 "signal" samples have been processed in order to supplement the Soup-SM-cocktail:

- ALPGEN $W + N$jets with $N = 0, .., 5$ ($300 < p_T(W) < 1600$ GeV)

- ALPGEN $Z + N$jets with $N = 0, .., 5$ ($300 < p_T(Z) < 1600$ GeV)
- PYTHIA $Z \to 2l$ ($l = e, \mu$) with $M_{ll} > 200$ GeV (ALPGEN samples include a cut on $M_{ll} < 200$ GeV)
- PYTHIA WW, WZ and ZZ samples
- SOFTSUSY LM1, LM2 and LM4 signal samples

For the dominant SUSY backgrounds, $t\bar{t}$, W+jets and Z+jets, a constant k-factor has been applied consistently for all jet-bins and p_T-bins in order to correct the LO cross section to the NLO prediction (obtained from MCFM [24]). For SUSY LM1, LM2 and LM4 the Prospino 2 NLO predictions are used.

5.3 Selection criteria

Muons

Global muons [25] are selected, i.e. a central track matched to the measurements in the muon system. In addition to this the standard muon isolation tool is used. The following selection cuts are applied to ensure high reconstruction efficiency and low fake probability:

- $p_T(\mu) > 30$ GeV
- $|\eta(\mu)| < 2.1$
- $N_{\text{tracker hits}} > 8$
- $R_{\text{track isolation}} = \frac{\sum p_T \text{ of tracks in 0.3 cone}}{p_T(\mu)} < 0.1$
- $\frac{\chi^2/dof}{N_{\text{valid hits}}} < 1$

The chosen η-acceptance is given by the instrumentation of the L1 single muon trigger. The cuts on number of hits and χ^2 of the muon track are designed to suppress mis-measured muon candidates which tend to have unphysically high p_T.

Electrons

So called Gaussian-Sum-Filter-Pixel-Matched electrons [26] are selected, i.e. an inner track matched to a super-cluster measured by the electromagnetic calorimeter (ECAL). In addition to this the standard EGamma Isolation tool and the official "cut based" electron identification criteria are used. The following selection cuts are applied in order to ensure high reconstruction efficiencies and low misidentification rates:

- $p_T(e) > 30$ GeV
- $|\eta(e)| < 2.5$
- Electron ID classification *tight*
- $R_{\text{track isolation}} = \frac{\sum p_T \text{ of tracks in 0.3 cone}}{p_T(e)} < 0.1$

The cut based ID tool uses the variables E/p, E_{had}/E_{tot}, cluster shape $\sigma_{\eta\eta}$ and the distances between track and supercluster in η and ϕ. Different cuts are used for different electron classifications, also distinguishing between endcap and barrel detectors.

Photons

Corrected photons [27] are selected, based on a supercluster measured in the ECAL. In addition to this the standard isolation tool of the e/γ-group and an official tool to identify photon conversions are used. Since no standardized photon ID tool existed at the time the analysis was designed, a private likelihood algorithm was used in order to identify photons. Separately for endcap and barrel regions, this likelihood discriminator uses energy ratios (r_9, r_{19}, e_{3x3}/e_{5x5} and E_{had}/E_{tot}) and shape variables ($\sigma_{\eta\eta}$, $\sigma_{\phi\phi}$ and $\sigma_{\eta\phi}$). The decision if a photon is "good" or not (needed as input to LLH) is taken matching generated photons to reconstructed ones. The following selection cuts are applied in order to ensure high reconstruction efficiency and low misidentification rates:

- $p_T(\gamma) > 30$ GeV
- $|\eta(\gamma)| < 2.5$
- veto on a matched pixel seed
- select only unconverted photons
- $E_{had}/E_{tot} < 0.2$
- $R_{\text{track isolation}} = \frac{\sum p_T \text{ of tracks in 0.3 cone}}{p_T(\gamma)} < 0.1$
- $likelihood > 0.1$

These cuts are chosen to ensure that only electromagnetic objects in the calorimeter without any tracks matched to them are considered, therefore photon which convert in the tracker are excluded, and thus the selection is orthogonal to the electrons.

Jets

The "iterative cone" jet algorithm[28] with a radius of $R = \sqrt{\Delta\phi^2 + \Delta\eta^2} = 0.5$ is used. L2-L3 jet energy scale corrections are applied in order to have a proper estimate of the jet at particle level. The following selection cuts are applied in order to ensure high reconstruction efficiency and a reasonable understanding of the jet energy scale:

- $p_T(jet) > 60$ GeV
- $\eta(jet) < 2.5$
- $E_{had}/E_{tot} > 0.05$

A certain amount of hadronic energy is required in order to seperate jets from electromagnetic objects in the calorimeter such as electrons or photons.

Missing tranverse energy

Since any error of the measurement of the physics objects discussed above propagates into the determination of $\not\!\!E_T$, a relatively high threshold of

- $\not\!\!E_T > 100$ GeV

is used. Standard corrections [29] are applied to the $\not\!\!E_T$ object, accounting for jet energy scale corrections of the jets in the event, and subtracting muon momenta from the calorimeter based $\not\!\!E_T$ estimate.

Within the scope of this note we do not present detailed control plots and efficiency curves for all the physics objects selected. As an example we only want to present the muon reconstruction efficiency after all selection criteria have been applied (Figure 11). The turn-on curve has been created using a SUSY LM4 sample, matching generated muons to reconstructed global muons with a $\Delta R < 0.2$ criterion. An overall reconstruction efficiency of more than 95% is found, almost flat in p_T.

5.4 Suppression of instrumental backgrounds

While it is clear that with the arrival of first data various "cleaning" steps will be needed to select "good" runs without detector problems, also on the level of physics object reconstruction additional criteria are needed in order to minimize instrumental and background from "fakes". This cleaning mainly refers to the removal of duplicate objects and the ambiguous interpretation of objects in the detector, e.g. a supercluster can be an electron as well as a photon candidate.
The cleaning steps are carried out in the following sequence:

- Muon candidates which are closer than $\Delta R < 0.2$ to each other are cleaned, keeping only the one measured best (smaller χ^2/dof). This is designed to remove ghost muons and other sources of duplicate muons.

- Electron candidates which are closer than $\Delta R < 0.2$ to each other, and which share either the inner track or the supercluster seed are cleaned, keeping only the more energetic one.
- Photon candidates which are closer than $\Delta R < 0.2$ to each other and which share the supercluster seed are cleaned, keeping only the more energetic one. Also photon candidates closer than $\Delta R < 0.2$ to an already selected electron are removed if the photon has the same supercluster seed as the electron. This should remove the ambiguity imposed by the fact that all superclusters can be interpreted as electrons as well as photons. Thus well measured electrons receive a higher priority than photons.
- Jet candidates closer than $\Delta R < 0.2$ to an already selected electron or photon are removed to avoid an overlap of those collections.

5.5 High Level Trigger

The choice of the trigger menu used in this analysis is driven mainly by the requirement to combine triggers with prescale factors 1 (or at least identical prescales) and HLT triggers which are expected to be "standard" at the LHC start-up and therefore commonly used and well understood.
An "OR" of various single lepton and di-lepton HLT-triggers [14] is used; they can be combined into a muon stream on the one hand and an electron stream on the other:

- single muon (with isolation) OR di-muon HLT (with and without isolation)

  [HLT1MuonIso || CandHLT2MuonIso || HLT2MuonNonIso]
- single electron OR di-electron HLT (both with and without isolation)

  [HLT1Electron || HLT2Electron || HLT1ElectronRelaxed || HLT2ElectronRelaxed || HLT1EMVeryHighEt || HLT1EMHighEt]

After all selection criteria have been applied, the muon and electron stream are merged into a single dataset, avoiding double counting in events where both electron and muon triggers have fired.

Again we present one exemplary efficiency plot. Figure 11 shows the HLT trigger efficiency using an "OR" of all used muon trigger bits. Events from the SUSY LM4 sample have been used, and the trigger efficiency has been determined after all selection cuts have been applied, including the topology requirement of at least a single reconstructed muon. The efficiency curve as a function of η shows an overall level of well above 80%; the dips clearly reflect the wheel structure of CMS. Note that unfortunately the single muon trigger without isolation has not been included in the SUSY-HLT-skim used for the CSA07 samples. Including this trigger in future skims would increase the overall efficiency to about 90%.
In addition to specific turn-on curves also a single number for the global HLT trigger efficiency using an "OR" of all the triggers mentioned above is of interest. We obtain, again using the SUSY LM4 sample and with respect to selected events (statistical errors only):

- Muon Stream: $\varepsilon_{HLT} = 82 \pm 0.4\%$
- Electron Stream: $\varepsilon_{HLT} = 88 \pm 0.3\%$

5.6 The search algorithm

At this point of the analysis, the events have been processed and physics objects satisfying the criteria mentioned above are identified. The composition of the event, i.e. the number of muons, jets etc. then determines to which event class it is assigned. At the present time two distributions are considered for each event class, thus limiting the number of distributions looked at and focusing on distributions which seem to be promising for spotting new physics:

- Scalar sum of the transverse momentum $\sum p_T$ of all physics objects.
  For example for the class 1μ 2jet $\not\!\!E_T + X$ one calculates $\sum p_T = p_T(\mu) + p_T(jet_1) + p_T(jet_2) + \not\!\!E_T$.
- Invariant mass M_{inv} of all physics objects. For classes with $\not\!\!E_T$ the transverse mass M_T is investigated.

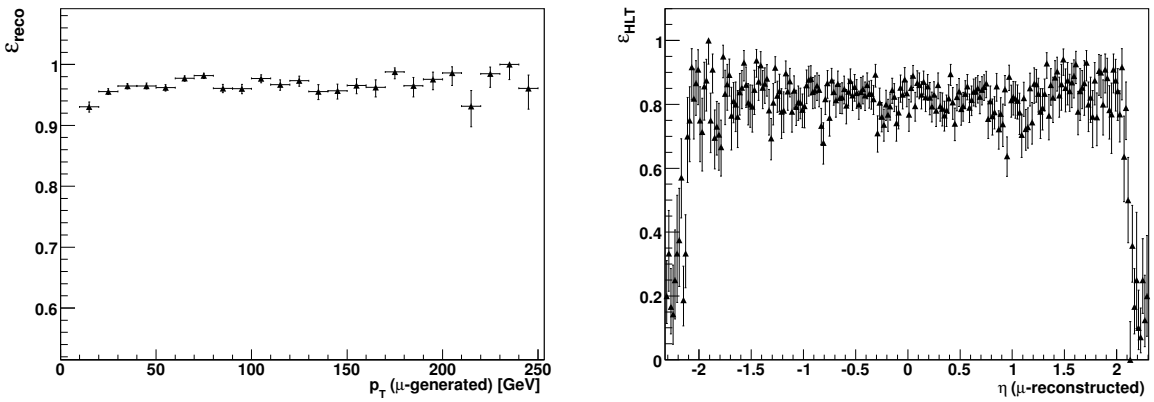


Figure 11: *Left: Muon-reconstruction efficiency as a function of p_T(gen) using the SUSY LM4 sample. Right: HLT trigger efficiency using "OR" of all used muon triggers as a function of η(rec) using the LM4 sample, with respect to selected events.*

These two distributions are input to the MUSiC algorithm (similar to the H1 analysis [21]) which scans them systematically for deviations, comparing the Standard Model expectation (Monte Carlo prediction) with the measured data.

First part

In order to do so, *each connected* bin region is considered within the distributions, i.e. single bins (bin 10 or bin 200) as well as broad regions (bin $3-100$ or bin $300-305$). We do not consider it meaningful to combine unconnected bins, e.g. combine bin 20 and bin 100 and bin 314 into one region.

For each connected region, a counting experiment is performed, adding up the various Monte Carlo contributions (N_{SM}) and comparing this sum to the amount of measured data (N_{data}). In addition to these two numbers also the uncertainty of the prediction ($\delta(N_{SM})$) is used, i.e. the systematic and statistical uncertainties of simulated events contributing to this specific region. Then a Poisson probability is computed, determining how likely the prediction fluctuates to the number seen in the data. The systematic errors, taking correlations into account, are included using a convolution with a Gaussian function:

$$p = \begin{cases} \displaystyle\sum_{i=N_{data}}^{\infty} A \cdot \int_0^{\infty} \mathrm{d}b \exp\left(\frac{-(b-N_{SM})^2}{2(\delta N_{SM})^2}\right) \cdot \frac{e^{-b}\, b^i}{i!} & \text{if} \quad N_{data} \geq N_{SM} \\ \displaystyle\sum_{i=0}^{N_{data}} A \cdot \int_0^{\infty} \mathrm{d}b \exp\left(\frac{-(b-N_{SM})^2}{2(\delta N_{SM})^2}\right) \cdot \frac{e^{-b}\, b^i}{i!} & \text{if} \quad N_{data} < N_{SM} \end{cases} , \tag{6}$$

where A ensures the normalization. From all the possible combinations of connected bins, the region with the smallest p-value (p_{min}^{data}) is chosen. This is the place in the distribution where the biggest discrepancy between data and Monte Carlo prediction is found. It is called the **Region of Interest**.
This simple but effective approach is sensitive to an excess of data as well as a deficit; it can detect large single bin fluctuations as well as possible signals spread over a large part of the distribution. A bin width of 50 GeV for both distributions ($\sum p_T$ and M_{inv}) is chosen in order to absorb detector resolution effects.

Second part

It is important to understand that the statistical estimator p alone is not sufficient to claim any signal. A statistical penalty factor has to be included to account for the large number of regions investigated. This is done in a second step of the algorithm, determining the **event class significance** (per distribution) of the deviation found in the first step:
Toy Monte Carlo experiments are performed, assuming the background-only hypothesis. Therefore hypothetical data histograms (HDH) are created numerous times by varying the Monte Carlo prediction for each bin according to its statistical and systematic uncertainty. These hypothetical data are then fed again into the first step of the

algorithm and compared to the Monte Carlo mean (results in p_{min}^{SM}). Again all possible connected regions are examined, not only the Region of Interest from step 1. The event class significance of the deviation is defined as:

$$\widetilde{P} = \frac{\text{number of HDH with } p_{min}^{SM} \leq p_{min}^{data}}{\text{total number of HDH}} \; . \tag{7}$$

The value of $\widetilde{P}$ is the fraction of toy experiments where a deviation even bigger than the one observed in the data is found. Thus performing these pseudo-experiments one jitters the Standard Model expectations and tests for signal-like fluctuations of the Standard Model. The $\widetilde{P}$ can directly be translated into standard deviations (see Figure 17) and is comparable to the widely used CL_b.

Sensitivity study with simulated events

Since the LHC has not started yet it is clear that we have no N_{data} to compare with the Monte Carlo prediction. Still we can pick exemplary models beyond SM and test the sensitivity of MUSiC with them. Instead of only producing pseudo-data for the backgound-only hypothesis in step 2 we can also create toy data as input to step 1 assuming *signal + background*, i.e. add the signal distributions on top of the SM ones. In this way we can repeat several pseudo-CMS experiments and determine the expected event class significance of a possible signal present in the data. Figure 12 illustrates this procedure, using the event class $1e\ 5\text{jet} + X$ as an example: The left curve represents the pseudo-experiments where signal plus background are assumed. With data this would correspond to a single line. The right curve on the other hand displays the multiple repetition of the SM expectation including its errors, thus this represents step 2 of the algorithm. The p and $\widetilde{P}$ values stated in the plots refer to the median of the left curve and then integrating the red curve beyond this median p_{min}. The interpretation of the two curves is clear: In the case that they are well separated, $\widetilde{P}$ will be quite low and discovery is easy, as shown in the left plot where no systematic errors are assumed. But if we include systematic uncertainties in the algorithm, the two hypotheses move close to each other and less than a 3σ deviation ($\approx 10^{-3}$) remains. This also underlines the importance of implementing systematic uncertainties into MUSiC.

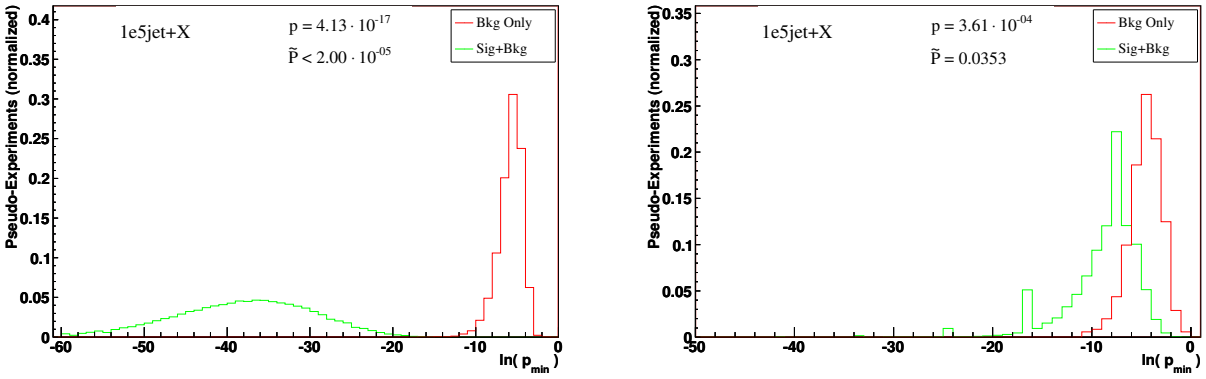


Figure 12: *Signal plus background and background-only hypotheses for an LM4 event class, on the left without systematic uncertainties and on the right with all uncertainties included. The striking difference between both plots shows the importance of systematic uncertainties.*

5.7 Systematic uncertainties

As mentioned in the previous section, it is crucial to implement correct systematic uncertainty estimates in the algorithm in order to distinguish a true signal from a "fake" deviation caused by an unanticipated detector effect or an incorrect theoretical estimation of the Standard Model expectation. The following systematic uncertainties are assumed and included in MUSiC; their magnitude is estimated in the context of $1fb^{-1}$ of data, but the values can be adapted very easily:

- σ(integrated luminosity) $= 5\%$
- σ(cross sections) $= 10\%$
- σ(jet energy scale) $= 5\%$, change in jets propagated also into $\not\!\!E_T$ estimate

- statistical uncertainty of the Monte Carlo prediction, based on the amount of originally produced events per sample

It is important to stress that the algorithm also accounts for correlations in the context of systematic errors. For global factors like cross sections all bins in a distribution are correlated, for the integrated luminosity even all physics processes are correlated. In addition to this variations are not always just "up or down", the JES uncertainty actually redistributes the bins and is again correlated for all generated samples. These correlations have to be taken into account when computing p-values and when generating pseudo-data.

The assumed cross section uncertainty of 10% is used for all Standard Model background processes, not distinguishing between different jet multiplicity bins or p_T bins of the generated samples. Studying the expected PDF uncertainties using the PDF reweighting method (see [30]) one can see that these typically range at the order of 2% to 8%, see [31] for details. The discussion of this and other theoretical uncertainties is still in flux, and our understanding of these numbers is likely to change in the future; therefore, we decided to use a single "conservative" number.

5.8 QCD background estimation from data

While for Standard Model processes like W+jets or $t\bar{t}$+jets we have Monte Carlo generator tools which can produce fairly reliable predictions of shapes for the various distributions with high statistics, it is clear that for QCD multijet production the enormous cross sections exceed the computational resources available. In addition the theoretical uncertainties for multijet events are orders of magnitude larger than in the case of electroweak processes. This analysis investigates events with at least a single isolated lepton; these are produced in QCD events only via non-prompt mechanisms or via misidentification, e.g. muons from b-jets or electrons from misidentified jets with a large pion fraction. Compared to the inclusive di-jet cross section these "fake" leptons are very rare, and thus difficult to model using inclusive QCD Monte Carlo samples.

For our MUSiC approach we aim to estimate the QCD contribution from the data, thus do not want to rely on the simulated prediction. Since in a generic search one is looking at many different distributions and a diversity of final states, it seems not practical to define control regions for each specific event class. One has to use a more general estimate of the QCD background applicable to all classes. The uncertainties of such cross-class extrapolations have to be absorbed by an appropriate global uncertainty of the QCD estimate, which can be easily incorporated into the search algorithm.

The strategy used to estimate the QCD from the data is similar to the methods commonly applied at the Tevatron [32]. A single selection cut, which is prominent for distinguishing "fake" leptons from well measured isolated ones, is inverted or relaxed. The sample with the inverted cut is then used to model the shape of the QCD background, and a control region is defined where the sample is scaled to fill up the gap between the remaining SM Monte Carlo samples and the data.
We exercise this method using final states with muons. Here the isolation cut is relaxed

$$0.1 < R_{track-isolation} = \frac{\sum p_T \text{ of tracks in 0.3 cone}}{p_T(\mu)} < 0.5 \tag{8}$$

In this way we get a sample with similar kinematics compared to the QCD events entering the final selection. Relaxing the cut even further one would risk to introduce larger differences in the distributions due to muons well within hadronic jets.
Two control regions are defined,

- $110 - 150$ GeV in the $\sum p_T$ distribution of the class 1μ 1jet $+ X$
- $130 - 180$ GeV in the $\sum p_T$ distribution of the class $1\mu \not{E}_T + X$

These two inclusive classes represent quite different corners of the phase space analyzed with MUSiC, once requesting a lepton and a jet and once the combination of a lepton and missing transverse energy. In this way we get two independent estimates of the scale factor to be used. Furthermore the two regions are both located at the very low p_T edge of the distributions, where signal contamination is expected to be small and QCD plus other SM processes dominate. From these control regions we obtain the following scale factor with its uncertainty:

$$f_{QCD} = \frac{\text{“data” − SM MC without QCD}}{\text{relaxed “data” − relaxed SM MC without QCD}} = 0.2 \pm 0.1 \tag{9}$$

The relative error of 50% indicates that the estimation of QCD from data for all event classes is not very precise. Nevertheless, since QCD contamination in the signal region is not very large, in most cases even such a large error should have a minor impact on the search sensitivity. It is more vital to get a proper shape of QCD in all classes without the enourmous single bin fluctuations of a QCD Monte Carlo sample caused by the lack of MC statistics. Note that since we exercise this method only using a QCD Monte Carlo sample (as “data”) the subtraction of the other SM samples is not needed. In any case the contribution in the denominator from relaxed SM MC without QCD should be negligable since these mostly fulfill $R_{track-isolation} < 0.1$.

Figure 13 shows the comparison of the QCD estimate from data with respect to the QCD Monte Carlo samples used to perform the cut inversion, errors correspond to the uncertainty of the scaling factor. The sample with relaxed cuts and the one fulfilling all final selection cuts agree well in terms of the shape. Note that the event classes shown here do not contain the control regions, thus the agreement within the assumed errors serves as a good indication that the extrapolation from one final state topology to another works reasonably well.

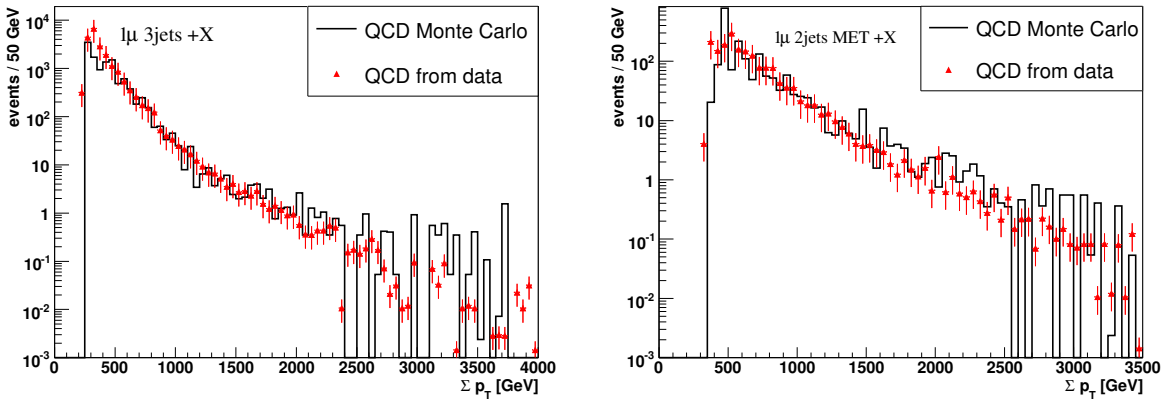


Figure 13: *QCD Monte Carlo and estimate using cut inversion in comparison, for two exemplary event classes.*

5.9 SUSY sensitivity

There are several evident reasons why such a generic search strategy as presented here might be a good supplement to more conventional analyses hunting for SUSY. First of all the large number of unconstrained parameters in most supersymmetric extensions of the SM also leads to an almost unlimited parameter space where nature can pick the SUSY scenario realized. Simplified models like MSSM try to shrink the number of free parameters using well or not so well founded physics assumptions. Still the parameter space remains huge, with a multitude of possible “benchmark” points leading to completely different decay channels and signal topologies. Thus it might be dangerous to rely solely on analyses optimized on specific SUSY points.

Through their decay chains, SUSY particles often lead to spectacular cascades in the final state with high multiplicities of leptons and jets and a large amount of $\not\!\!E_T$ due to the LSP (lightest SUSY particle). So unlike single resonance production as for example $Z \rightarrow \mu\mu$, SUSY does not predominantly favor a single topology, but does contribute to a multitude of event classes within MUSiC. Thus this generic search aims to give a consistent picture of SUSY particles appearing on top of the Standard Model prediction. The information of significant deviations found in several classes could be combined to provide additional evidence for SUSY.

MUSiC performance in the specific channel $1\mu, \geq 3$jet, $\not\!\!E_T + X$ for LM1, LM2 and LM4

In order to be comparable more easily to alternative analysis strategies (cut-based and BDT), we will now focus on the specific channel these analyses are optimized for. Of couse in a way this contradicts the basic philosophy of MUSiC which does not aim to look only at a single topology. Still it is interesting to investigate how well the algorithm performs compared to the specific analyses in this channel, assuming three exemplary SUSY points. In a next step we will report about the global MUSiC results on LM4.

As introduced ealier we can repeat pseudo experiments assuming once the signal plus background hypothesis and compare these to multiple background-only estimates. This gives us the median p-value and the expected event class significance $\widetilde{P}$. The results for exemplary inclusive channels for all three SUSY points are shown in Table 16, assuming 1 fb^{-1} of luminosity.

SUSY Point	Event Class	Luminosity	p-value (median)	$\widetilde{P}$ (expected)	$\widetilde{P}$ Standard Dev.
LM1	1μ 3jet $\not\!\!E_T + X$	1 fb^{-1}	$2.5 \cdot 10^{-9}$	$1.1 \cdot 10^{-04}$	3.9 σ
LM2	1μ 4jet $\not\!\!E_T + X$	1 fb^{-1}	$8.1 \cdot 10^{-4}$	0.09	1.7 σ
LM4	1μ 5jet $\not\!\!E_T + X$	1 fb^{-1}	$3.5 \cdot 10^{-7}$	$7 \cdot 10^{-05}$	4 σ
LM1	$2e$ 3jet $\not\!\!E_T + X$	1 fb^{-1}	$2.7 \cdot 10^{-21}$	$< 2 \cdot 10^{-05}$	$>$4.3 σ

Table 16: *Results for repeating S+B and B hypothesis multiple times, all for Σp_T distribution and assuming 1 fb^{-1}.*

SUSY Point	Event Class	Luminosity	p-value (median)	$\widetilde{P}$ (expected)	$\widetilde{P}$ Standard Dev.
LM1	1μ 3jet $\not\!\!E_T + X$	100 pb^{-1}	$3.7 \cdot 10^{-5}$	$4.7 \cdot 10^{-3}$	2.8 σ
LM2	1μ 4jet $\not\!\!E_T + X$	100 pb^{-1}	$3.9 \cdot 10^{-3}$	0.25	1.2 σ
LM4	1μ 5jet $\not\!\!E_T + X$	100 pb^{-1}	$1.9 \cdot 10^{-4}$	$1.6 \cdot 10^{-2}$	2.4 σ
LM1	$2e$ 3jet $\not\!\!E_T + X$	100 pb^{-1}	$1.3 \cdot 10^{-6}$	$8 \cdot 10^{-5}$	3.9 σ

Table 17: *Results for repeating S+B and B hypothesis multiple times, all for Σp_T distribution and assuming 100 pb^{-1}.*

These numbers clearly show that if LM1 or LM4 is realized in nature, within 1 fb^{-1} of luminosity MUSiC can detect a significant deviation from the Standard Model expectation. With significances well above 3σ ($\approx 10^{-3}$, see Figure 17), a deviation inconsistent with the SM only assumption could be found in these parts of the SUSY parameter space. Using the single lepton channel a 5σ discovery is not within reach due to the large SM backgrounds and the jet energy scale uncertainties assumed. In the lepton + jets + $\not\!\!E_T$ channel these lead to a considerable variation of the SM background, on top of which the SUSY events are accumulated. More specific analysis strategies might be able to decrease these uncertainties by optimizing selection cuts. Thus a synergy between MUSiC spotting some significant deviation and a specific analysis then maximizing the sensitivity for a discovery in the same channel looks promising.

Taking a closer look at the numbers one can see that in LM1 with its large cross sections a significant deviation can even be found in the 3 jet channel. For LM4 events with at least 4 or 5 jets look more promising.

Since for cut-based and BDT analysis also the reach with only 100 pb^{-1} has been investigated, Table 17 summarizes these numbers. One can see that in the single lepton channels the expected event class significance decreases a lot. Even for LM1 the obeserved deviation from the SM remains just under 3σ. On the other hand the last line presents the results for a class with more than one lepton ($2e$ 3jet $\not\!\!E_T + X$). Here also within 100 pb^{-1} a significant deviation can be found. Assuming more luminosity (Table 16) only upper limits can be set for the general significance. The tiny p-value of 10^{-21} indicates that in this channel even a 5σ discovery seems possible.

Figure 14 shows the Σp_T distribution for the event classes with significant deviations discussed above, assuming 1 fb^{-1} of luminosity. Here only a single pseudo experiment for the signal+backgound hypothesis is shown, thus the data points correspond to a single CMS experiment. All uncertainties are included when creating the toy data, and the shaded area corresponds to the total error of the SM backgound estimate for each bin. The event class significance $\widetilde{P}$ is of course computed repeating the background-only hypothesis numerous times.

The dominant background sources in both distributions are $t\bar{t}$ and W+jets. In both cases the region of interest with the biggest discrepancy found in the "data" lies in the high p_T tail where the SUSY contribution becomes bigger than or equal to the Standard Model expectation. The deviations in both cases well exceed the 3σ level.

So far only the results for LM1 and LM4 have been discussed. The $\widetilde{P}$ of 9% for LM2 indicates that the algorithm has no chance to discover anything here. The small cross section of LM2 leads to only few SUSY events in the lepton + jet + $\not\!\!E_T$ final state, thus data and SM-only hypothesis cannot be distinguished. Nevertheless, this is the point where the proper meaning of MUSiC sets in: MUSiC looks at a variety of final state topologies, thus there is hope that at least in other channels LM2 can be revealed. Unfortunately the small cross section of LM2 leads to significant deviations only in a small number of event classes, all belonging to the type tri-leptons + jets (+$\not\!\!E_T$). Here, the SM backgrounds are tiny, so events are found in topologies where no Monte Carlo events have been generated.

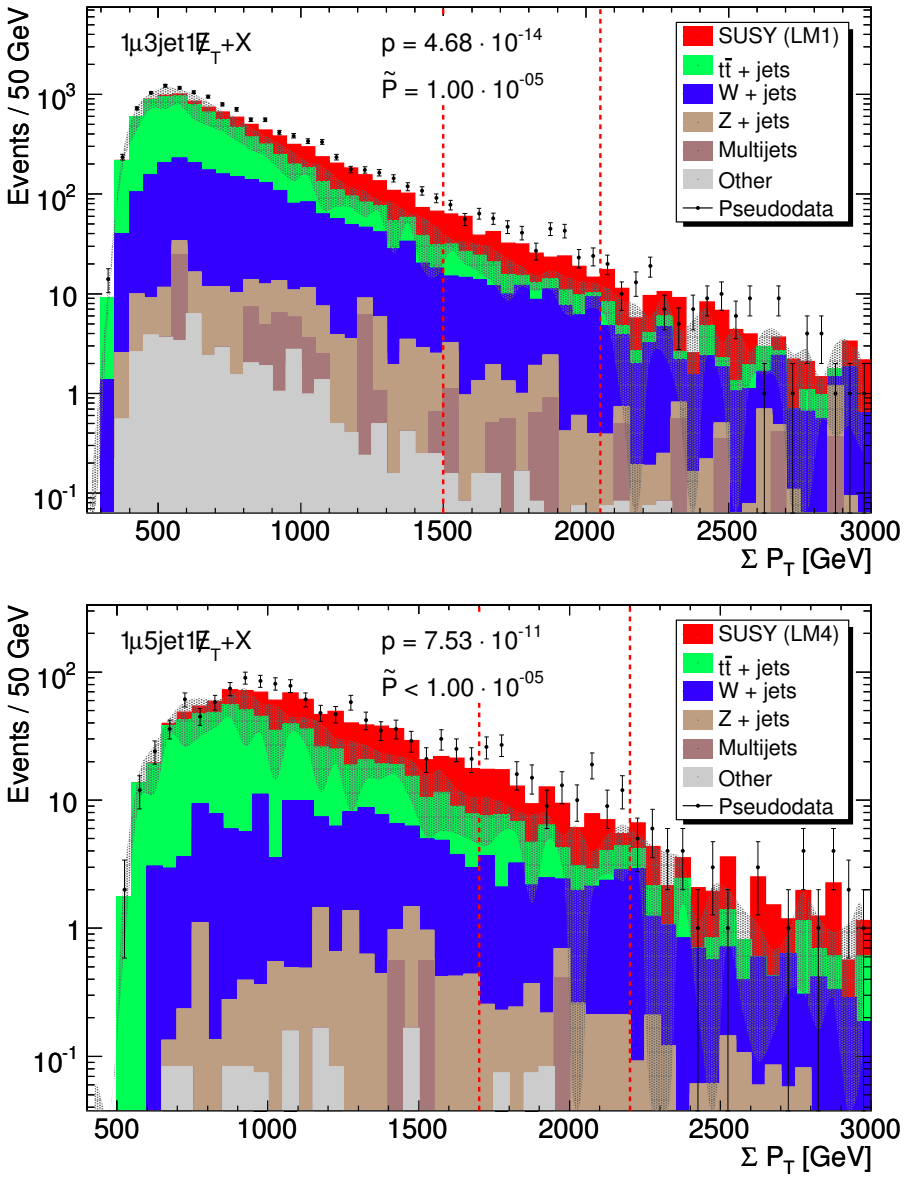


Figure 14: *Results of an exemplary single pseudo-experiment assuming* $1\ fb^{-1}$*. The dotted lines indicate the region of interest where the smallest* p*-value has been found,* $\widetilde{P}$ *corresponds to the significance assuming this single set of toy data. The shaded area corresponds to the total uncertainty of the SM expectation. Numbers for the regions of interest (dotted lines): upper plot* $N_{data} = 505$ *and* $N_{MC} = 191 \pm 42$*; lower plot* $N_{data} = 156$ *and* $N_{MC} = 52 \pm 12$

Figure 15 shows the Σp_T distribution for the event classes $2e\ 1\mu$ 3jet $\not{E}_T + X$. Only a single Z+jets event at 1800 GeV is contributing. Since a p-value for zero background is not defined, we conservatively assume a 68% upper confidence level for the bins with "data" entries, thus using 1.15 generated Z-events instead of zero. Still an event class significance above 3σ is found, thus giving evidence for LM2. It is clear that as a consequence one would look at this channel in more detail, i.e. by using a $Z + jets$ Monte Carlo with a higher statistic and carefully estimating possible other SM background constributions, or investigating the possibility of an additional contamination from "fake" leptons. Nevertheless, MUSiC is able to serve as an early warning system and points to final state topologies where interesting physics beyond the Standard Model could be present. Some more general features of MUSiC will be discussed in the following, taking LM4 as an example.

General MUSiC performance assuming LM4

As mentioned earlier, the advantage of a generic search is that it is sensitive to deviations from the SM expectation in a variety of final state topologies, thus possibly giving a more consistent picture of SUSY appearing within the

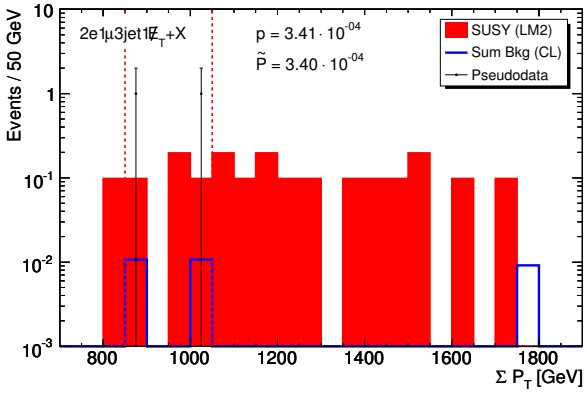


Figure 15: *Results of an exemplary single pseudo experiment for LM2 assuming* 1 *fb*$^{-1}$*. Since no SM background contributes in the bins with data points, a* 68% *upper confidence level is used.*

first LHC data. In order to test the more global results of MUSiC, a complete scan of all possible final states using the distributions $\sum p_T$ and M_{inv} (M_T) has been performed, assuming SUSY point LM4 and 1 fb^{-1}.

In a first step, signal plus backgound hypothesis and background-only hypothesis have been repeated numerous times in order to get an estimate for the expected magnitude of any discrepancy (p_{median} and $\widetilde{P}_{expected}$). We can summarize the results in the following way:

- LM4 contributes to 160 *exclusive* event classes, 15% show significant deviations with $\widetilde{P}_{expected} \leq 10^{-4}$
- LM4 contributes to 260 *inclusive* event classes, 30% show significant deviations with $\widetilde{P}_{expected} \leq 10^{-4}$

Of course in the case of inclusive classes deviations found are "duplicated" in some way since 1μ 5jet events contribute to 1μ 2jet $+ X$, 1μ 3jet $+ X$ and so on. Nevertheless, when comparing similar final state topologies, the inclusive classes tend to have smaller p_{median} values and larger event class sigificances than the exclusive ones. In the exclusive channels only multi-lepton final states with 2 or 3 leptons look promising. For the inclusive classes also the single lepton plus jets plus $\not\!\!E_T$ show indications of SUSY. Still in general di-leptons or events with even 3 or 4 leptons show bigger SUSY excesses over the SM background.

The two kinematic distributions examined, $\sum p_T$ and M_{inv} (M_T), lead to similar results. A systematic advantage of one of them cannot be observed in the presented LM analyses.

The result that the inclusive channels look more promising is consistent with the expectation one would have for SUSY events: The complex decay chains produce a certain number of jets only with a certain probability, there is no single (exclusive) topology enhanced over others. Thus by summing up events of a certain topology inclusively more SUSY contributions can be integrated, and the excess observed in the data is more striking.

Also one should note that in the case of inclusive classes a number of final states *without* $\not\!\!E_T$ (e.g. 3μ1jet+X, $\sum p_T$: p-value(median) = $4.4 \cdot 10^{-7}$, $\widetilde{P}$(expected) $\leq 2 \cdot 10^{-5}$) show significant deviations. This could be useful if, in the case of early CMS running, the understanding and description of $\not\!\!E_T$ proves to be more difficult than anticipated.

Two exemplary distributions corresponding to a single pseudo experiment are shown in Figure 16. These represent final states with same or different flavor di-leptons plus several jets and a large amount of $\not\!\!E_T$. Here the amount of SM background is relatively small (mostly $t\bar{t}$), while SUSY still contributes with a relatively large number of events. The MUSiC algorithm chooses in both event classes regions where the sum of SUSY and SM-background well exceeds the SM-only expectations and its corresponding errors.

5.10 Statistical interpretation of all event classes

So far the individual event classes have been interpreted and compared to the sensitivity of the specific analyses apart from the complete set of events. For each event class a significance $\widetilde{P}$ has been computed which easily can be translated into standard deviations, see left plot of Figure 17.
In addition a final trial factor can be estimated and applied to account for the multiple number of final state topologies looked at. A similar punishment factor could also be used when considering the large number of

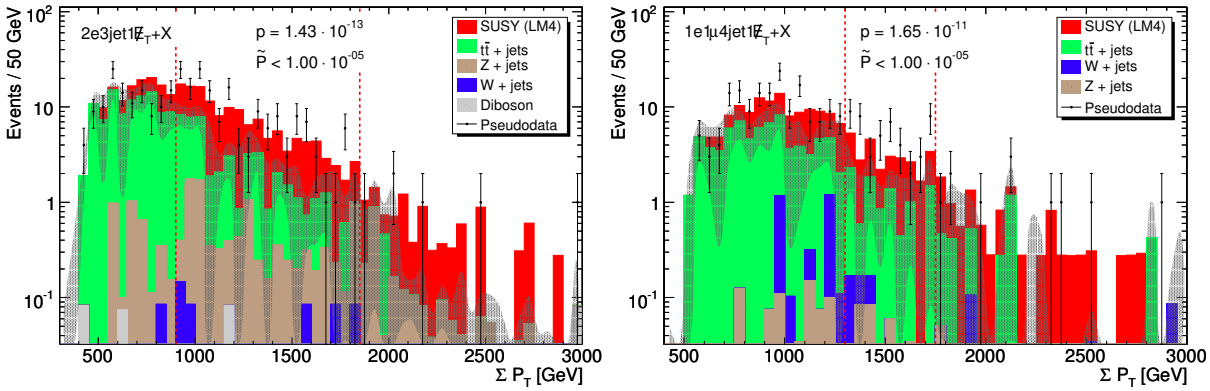


Figure 16: *Results of an exemplary single pseudo experiment assuming* $1\ fb^{-1}$. *The dotted lines indicate the region of interest where the smallest* p*-value has been found,* $\widetilde{P}$ *corresponds to the significance assuming this single set of toy data. The shaded area corresponds to the total uncertainty of the SM expectation. Numbers for the regions of interest: left plot* $N_{data} = 164$ *and* $N_{MC} = 49 \pm 12$, *right plot* $N_{data} = 49$ *and* $N_{MC} = 10 \pm 3$.

independent analyses conducted by the whole CMS collaboration.
Conservatively neglecting correlations between the event classes (which is not true for the inclusive ones for sure), the final statistical estimator for the overall degree of agreement with the Standard Model can be quantified using the formula

$$P_{\text{CMS}} = 1 - (1 - \widetilde{P})^n, \tag{10}$$

where $\widetilde{P}$ is the significance of a certain event class and n refers to the total number of distributions analysed. The right plot of Figure 17 displays this translation for various number of event classes considered. As an example, if 1000 classes are used a *local* 5-sigma deviation in a certain topology leads to roughly 3.5-sigma for *global CMS*. An interesting approach to lower the penalty of the trial factor is to use a so called hypothesis ranking as described here [33]. Its feasibility within the MUSiC algorithm is currently investigated and will be a topic of a future note.

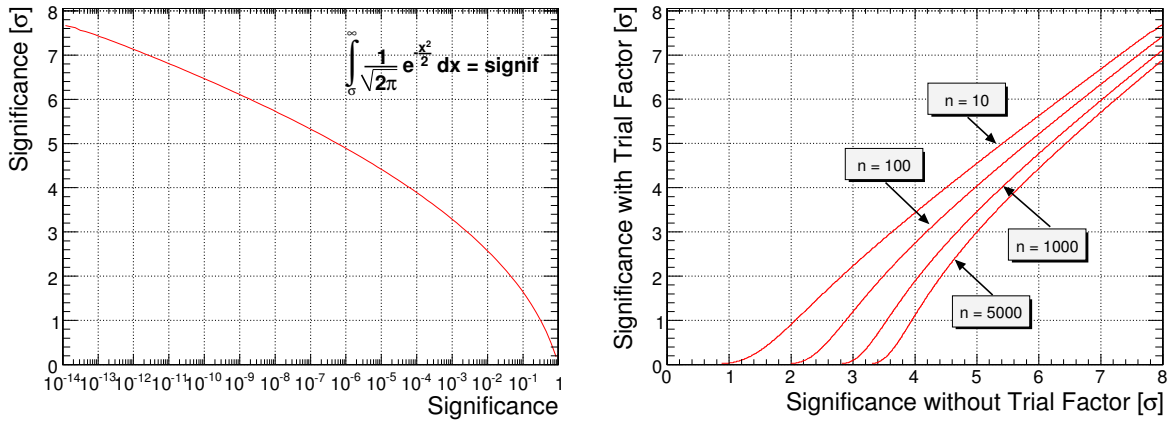


Figure 17: *Left: Translation of significance* ($\widetilde{P}$) *into number of standard deviations* σ. *Right: Translation of significance per event class* ($\widetilde{P}$) *into global-experiment significance after accounting for trial factors* (P_{CMS}).

6 Summary and conclusions

The sensitivities of the three analyses follow the expectations. The model independent search MUSiC performs best when several channels are combined into a comprehensive review in accordance with this strategy. For individual channels, this search has, for an integrated luminosity of $1\ \text{fb}^{-1}$, discovery sensitivity for LM1 and LM4. The same is true for the cut based and the BDT analysis, but their expected significances are higher – or correspondingly, the required discovery luminosities are lower. The best sensitivity is achieved with BDTs.

The LM2 point is difficult for all analyses with $1\ \mathrm{fb}^{-1}$ of integrated luminosity. Both the cut based analysis would require more luminosity, or a reduction of systematic uncertainties. MUSiC would find LM2 when including additional signatures.

All three searches confirm the feasibility of an early discovery of low mass SUSY. For early data taking, assuming a signal with a very large cross section (e.g. LM1), a cut based analysis which can be implemented also with only limited detector understanding would be preferable. Once an excellent detector understanding has been achieved, BDTs could lead to discovery, in particular if the signal cross sections are small and the backgrounds high. MUSiC will help to stay alert to all possibilities, be it standard SUSY or an unexpected signal, or the understanding of the detector and backgrounds.

All three analyses outline the importance of systematic uncertainties, but also the importance of sufficient MC statistics or data-driven techniques to reduce the reliance on MC.

A Cross Sections

Dataset	NLO Cross Section [fb]	Number of Events	Generator	Comments
LM1	61122	120k	Pythia	NLO (Prospino 2)
LM2	10520	110k	Pythia	NLO (Prospino 2)
LM4	27691	100k	Pythia	NLO (Prospino 2)

Table 18: *Signal Datasets used in the analyses. The total NLO cross section from Prospino 2 (with CTEQ6M PDF) are quoted as well as the number of analysed events. SoftSusy has been used to calculate the mass spectrum, SUSY-Hit to decay the particles and Pythia for the simulation of the parton shower.*

Dataset	LO Cross Section [fb]	# of Events	Generator	Comments
W_0jets	50736000	8.8M	Alpgen	Chowder, NLO (k = 1.12)
W_1jets_0_100	10348800	9.1M	Alpgen	Chowder, NLO (k = 1.12)
W_1jets_100_300	287280	247k	Alpgen	Chowder, NLO (k = 1.12)
W_1jets_300_800	3259	57k	Alpgen	NLO (k = 1.12)
W_1jets_800_1600	17.8	60k	Alpgen	NLO (k = 1.12)
W_2jets_0_100	2842560	2.4M	Alpgen	Chowder, NLO (k = 1.12)
W_2jets_100_300	252000	287k	Alpgen	Chowder, NLO (k = 1.12)
W_2jets_300_800	4525	25k	Alpgen	NLO (k = 1.12)
W_2jets_800_1600	35.3	54k	Alpgen	NLO (k = 1.12)
W_3jets_0_100	658560	353k	Alpgen	Chowder, NLO (k = 1.12)
W_3jets_100_300	120288	118k	Alpgen	Chowder, NLO (k = 1.12)
W_3jets_300_800	3438	107k	Alpgen	NLO (k = 1.12)
W_3jets_800_1600	33.7	53k	Alpgen	NLO (k = 1.12)
W_4jets_0_100	138432	126k	Alpgen	Chowder, NLO (k = 1.12)
W_4jets_100_300	42336	40k	Alpgen	Chowder, NLO (k = 1.12)
W_4jets_300_800	1747	29k	Alpgen	NLO (k = 1.12)
W_4jets_800_1600	21.1	55k	Alpgen	NLO (k = 1.12)
W_5jets_0_100	84300	62k	Alpgen	Chowder, NLO (k = 1.12)
W_5jets_100_300	44352	44k	Alpgen	Chowder, NLO (k = 1.12)
W_5jets_300_800	3394	40k	Alpgen	NLO (k = 1.12)
W_5jets_800_1600	66.3	17k	Alpgen	NLO (k = 1.12)
Z_0jets	5040000	3.3M	Alpgen	Chowder, NLO (k = 1.12)
Z_1jets_0_100	1041600	945k	Alpgen	Chowder, NLO (k = 1.12)
Z_1jets_100_300	33600	36k	Alpgen	Chowder, NLO (k = 1.12)
Z_1jets_300_800	403	30k	Alpgen	NLO (k = 1.12)
Z_1jets_800_1600	2.24	13k	Alpgen	NLO (k = 1.12)
Z_2jets_0_100	302400	289k	Alpgen	Chowder, NLO (k = 1.12)
Z_2jets_100_300	31584	35k	Alpgen	Chowder, NLO (k = 1.12)
Z_2jets_300_800	616	29k	Alpgen	NLO (k = 1.12)
Z_2jets_800_1600	4.48	22k	Alpgen	NLO (k = 1.12)
Z_3jets_0_100	77280	73k	Alpgen	Chowder, NLO (k = 1.12)
Z_3jets_100_300	14448	24k	Alpgen	Chowder, NLO (k = 1.12)
Z_3jets_300_800	448	28k	Alpgen	NLO (k = 1.12)
Z_3jets_800_1600	4.26	16k	Alpgen	NLO (k = 1.12)
Z_4jets_0_100	15456	33k	Alpgen	Chowder, NLO (k = 1.12)
Z_4jets_100_300	4704	7k	Alpgen	Chowder, NLO (k = 1.12)
Z_4jets_300_800	224	25k	Alpgen	NLO (k = 1.12)
Z_4jets_800_1600	2.8	12k	Alpgen	NLO (k = 1.12)

Table 19: *Standard Model background datasets used in the analyses. The leading order crossection of the quoted generator and the approximate number of events used in the analyses are stated. The comment column contains the k-factor and the "soup" from which the dataset has been extracted.*

Dataset	LO Cross Section [fb]	# of Events	Generator	Comments
Z_5jets_0_100	9744	12k	Alpgen	Chowder, NLO (k = 1.12)
Z_5jets_100_300	5712	6k	Alpgen	Chowder, NLO (k = 1.12)
Z_5jets_300_800	504	25k	Alpgen	NLO (k = 1.12)
Z_5jets_800_1600	8.3	37k	Alpgen	NLO (k = 1.12)
$t\bar{t}$_0jets	618825	1.5M	Alpgen	Chowder, NLO (k = 1.85)
$t\bar{t}$_1jets	176490	362k	Alpgen	Chowder, NLO (k = 1.85)
$t\bar{t}$_2jets	33670	81k	Alpgen	Chowder, NLO (k = 1.85)
$t\bar{t}$_3jets	5920	14k	Alpgen	Chowder, NLO (k = 1.85)
$t\bar{t}$_4jets	1480	5.3k	Alpgen	Chowder, NLO (k = 1.85)
WW	70000	850k	Pythia	LO
WZ	26870	360k	Pythia	LO
ZZ	11200	140k	Pythia	LO
QCD_0_15	5.3e+13	14M	Pythia	Gumbo, LO
QCD_15_20	1.46e+12	1.7M	Pythia	Gumbo, LO
QCD_20_30	6.3e+11	2.7M	Pythia	Gumbo, LO
QCD_30_50	1.63e+11	2.5M	Pythia	Gumbo, LO
QCD_50_80	2.16e+10	2.5M	Pythia	Gumbo, LO
QCD_80_120	3.08e+09	1.2M	Pythia	Gumbo, LO
QCD_120_170	4.94e+08	1.3M	Pythia	Gumbo, LO
QCD_170_230	1.01e+08	1.2M	Pythia	Gumbo, LO
QCD_230_300	2.45e+07	1.2M	Pythia	Gumbo, LO
QCD_300_380	6.24e+06	1.2M	Pythia	Gumbo, LO
QCD_380_470	1.78e+06	1.2M	Pythia	Gumbo, LO
QCD_470_600	683000	1.2M	Pythia	Gumbo, LO
QCD_600_800	204000	500k	Pythia	Gumbo, LO
QCD_800_1000	35100	100k	Pythia	Gumbo, LO
QCD_1000_1400	10900	30k	Pythia	Gumbo, LO
QCD_1400_1800	1600	30k	Pythia	Gumbo, LO
QCD_1800_2200	145	20k	Pythia	Gumbo, LO
QCD_2200_2600	23.8	10k	Pythia	Gumbo, LO
QCD_2600_3000	4.29	10k	Pythia	Gumbo, LO
QCD_3000_3500	0.844	10k	Pythia	Gumbo, LO
QCD_3500_inf	0.108	10k	Pythia	Gumbo, LO
PhotonJets_0_15	1.7e+11	300k	Pythia	Gumbo, LO
PhotonJets_15_20	2.57e+08	520k	Pythia	Gumbo, LO
PhotonJets_20_30	1.32e+08	600k	Pythia	Gumbo, LO
PhotonJets_30_50	4.1e+07	510k	Pythia	Gumbo, LO
PhotonJets_50_80	7.2e+06	520k	Pythia	Gumbo, LO
PhotonJets_80_120	1.3e+06	530k	Pythia	Gumbo, LO
PhotonJets_120_170	275000	560k	Pythia	Gumbo, LO
PhotonJets_170_300	87000	200k	Pythia	Gumbo, LO
PhotonJets_300_500	8200	30k	Pythia	Gumbo, LO
PhotonJets_500_7000	870	30k	Pythia	Gumbo, LO
emQCD	5.335e+11	8.7M	Pythia	Stew, LO
emQCD_bbbar_5_50	1.7005e+10	3.0M	Pythia	Stew, LO
emQCD_bbbar_50_170	1.6524e+08	3.0M	Pythia	Stew, LO
emQCD_bbbar_170_inf	2.535e+06	2.6M	Pythia	Stew, LO
muQCD	4.4e+10	20M	Pythia	Stew, LO
B_JPsi	4.26738e+07	500k	Pythia	Stew, LO
BottomOnia_0_20	3.20423e+07	1M	Pythia	Stew, LO
BottomOnia_20_inf	425427	1M	Pythia	Stew, LO

Table 20: *More Standard Model background datasets used in the analyses. The leading order crossection of the quoted generator and the approximate number of events used in the analyses are stated. The comment column contains the k-factor and the "soup" from which the dataset has been extracted.*

Dataset	LO Cross Section [fb]	# of Events	Generator	Comments
CharmOnia_0_20	3.274e+08	1M	Pythia	Stew, LO
CharmOnia_20_inf	2.473e+06	1M	Pythia	Stew, LO
DrellYan_ee_200	1656	42k	Pythia	LO
DrellYan_ee_500	86.15	65k	Pythia	LO
DrellYan_ee_1000	7.45	12k	Pythia	LO
DrellYan_ee_1500	1.22	4k	Pythia	LO
DrellYan_ee_2000	0.264	10k	Pythia	LO
DrellYan_mumu_200	1656	42k	Pythia	LO
DrellYan_mumu_500	86.15	42k	Pythia	LO
DrellYan_mumu_1000	7.45	13k	Pythia	LO
DrellYan_mumu_1500	1.22	14k	Pythia	LO
DrellYan_mumu_2000	0.264	13k	Pythia	LO

Table 21: *More Standard Model background datasets used in the analyses. The leading order crossection of the quoted generator and the approximate number of events used in the analyses are stated. The comment column contains the k-factor and the "soup" from which the dataset has been extracted.*

References

[1] Y.A. Golfand and E.P. Likhtman, JETP Lett. 13 (1971) 323;
D.V. Volkov and V.P. Akulov, JETP Lett. 16 (1972) 438;
D.V. Volkov and V.P. Akulov, Phys. Lett. 46 B (1973) 109;
J. Wess and B. Zumino, Nucl. Phys. B 70 (1974) 39.

[2] P. Fayet, Phys. Lett. 69 B (1977) 489;
P. Fayet, Phys. Lett. 70 B (1977) 461.

[3] V.M. Abazov *et al.* (D0 Collaboration), Phys. Lett. B 660 (2008) 449.

[4] *CMS Physics Technical Design Report, Volume II: Physics Performance*, J. Phys. G: Nucl. Part. Phys. 34 (2007) 995.

[5] W. Beenakker, R. Hopker and M. Spira, `http://www.ph.ed.ac.uk/~tplehn/prospino/` arXiv:hep-ph/9611232.

[6] J. Pumplin *et al.*, J. High Energy Phys. **0207** (2002) 012;
D. Stump *et al.*, J. High Energy Phys. **0310** (2003) 046.

[7] `https://twiki.cern.ch/twiki/bin/view/CMS/CSA07Physics`

[8] B.C. Allanach, Comput. Phys. Commun. 143 (2002) 305, hep-ph/0104145.

[9] A. Djouadi *et al.*, Decays of Supersymmetric Particles: the program SUSY-HIT (SUspect-SdecaY-Hdecay-InTerface), hep-ph/0609292.

[10] T. Sjostrand, S. Mrenna and P. Skands, PYTHIA 6.4 Physics and Manual, hep-ph/0603175.

[11] M.L. Mangano *et al.*, ALPGEN, a generator for hard multiparton processes in hadronic collisions, JHEP 0307:001 (2003), hep-ph/0206293.

[12] M.U. Mozer *et al.*, Description of CSA07 SUSYBSM Skims, CMS IN-2008/009.

[13] C. D. Jones et al., The new CMS data model and framework, CHEP06 Conference Proceedings, 2007.
CMS Software (CMSSW), `https://twiki.cern.ch/twiki/bin/view/CMS/WorkBook`.

[14] The CMS Collaboration, The TriDAS Project, Technical Design Report, Volume 1: The Trigger Systems, CERN LHCC 2000-38, 2000;
The CMS Collaboration, The TriDAS Project, Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger, CERN LHCC 2002-36, 2002.

[15] P. Biallass *et al.*, Simulation of Cosmic Muons and Comparison with Data from the Cosmic Challenge using Drift Tube Chambers, CMS NOTE 2007/024;
M. Aldaya *et al.*, Results of the First Integration Test of the CMS Drift Tubes Muon Trigger, CMS NOTE 2006/072;
M. Aldaya *et al.*, Fine synchronization of the muon drift tubes local trigger, CMS NOTE 2006/002.

[16] T. Junk, Nucl. Instrum. Meth. A **434** (1999) 435 [arXiv:hep-ex/9902006].

[17] A.L. Read 2002 J. Phys. G: Nucl. Part. Phys. 28 2693-2704,
`http://www.iop.org/EJ/abstract/0954-3899/28/10/313/`

[18] R. Brun and F. Rademakers, Nucl. Instrum. Meth. A **389** (1997) 81, `http://root.cern.ch/`.

[19] A. Hocker *et al.*, arXiv:physics/0703039, `http://tmva.sourceforge.net/`.

[20] B. Abbott *et al.* (D0 Collaboration), Phys. Rev. D 62 (2000) 092004;
V.M. Abazov *et al.* (D0 Collaboration), Phys. Rev. D 64 (2001) 012004;
B. Abbott *et al.* (D0 Collaboration), Phys. Rev. Lett. 86 (2001) 3712.

[21] A. Aktas *et al.* (H1 Collaboration), Phys. Lett. B 602 (2004) 14.

[22] T. Aaltonen *et al.* (CDF Collaboration), arXiv:0712.1311 (submitted to Phys. Rev. D).

[23] T. Hebbeker, http://web.physik.rwth-aachen.de/~hebbeker/l3note_2305.pdf

[24] J.M. Campbell and R.K. Ellis, Phys. Rev. D **60**, 113006 (1999), http://mcfm.fnal.gov/.

[25] The CMS Collaboration, The Muon Project, Technical Design Report, CERN/LHCC 97-31, CMS TDR 2 (1997).

[26] S. Baffioni *et al.*, Electron Reconstruction in CMS, CMS Note 2006/40;
The CMS Collaboration, The Electromagnetic Calorimeter, Technical Design Report, CERN/LHCC 97-33, CMS TDR 4 (1997).

[27] N. Marinelli, Track finding and identification of converted photons, CMS NOTE 2006/005;
The CMS Collaboration, The CMS Physics TDR, Volume I, CERN/LHCC 2006-001 (2006).

[28] P. Schieferdecker *et al.*, Performance of Jet Algorithms in CMS, CMS AN 2008/001;
S. Esen *et al.*, Plans for Jet Energy Corrections at CMS, CMS AN 2007/055;
R. Harris *et al.*, MC Truth L2 & L3 Factorized Jet Corrections at CMS, CMS AN 2008/003.

[29] S. Esen *et al.*, $\not\!\!E_T$ Performance in CMS, CMS AN 2007/041;
The CMS Collaboration, The CMS Physics TDR, Volume I, CERN/LHCC 2006-001 (2006).

[30] D. Bourilkov, R. C. Group and M. R. Whalley, arXiv:hep-ph/0605240.

[31] https://twiki.cern.ch/twiki/bin/view/CMS/AachenPdfUncertainties

[32] The methods are discussed in numerous D0 and CDF publications. A recent example is:
V. M. Abazov *et al.* [D0 Collaboration], arXiv:0803.0739 [hep-ex].

[33] S. D. Biller, Astropart. Phys. **4** (1996) 285.