Available on the CERN CDS information server **CMS PAS EXO-08-005**

CMS Physics Analysis Summary

Contact: cms-pag-conveners-exotica@cern.ch 2008/10/08

MUSiC – An Automated Scan for Deviations between Data and Monte Carlo Simulation

The CMS Collaboration

Abstract

We present a model independent analysis approach, systematically scanning the data for deviations from the Monte Carlo expectation. Such an analysis can contribute to the understanding of the detector and the tuning of the event generators. Due to the minimal theoretical bias this approach is sensitive to a variety of models, including those not yet thought of. Events are classified into event classes according to their particle content (muons, electrons, photons, jets and missing transverse energy). A broad scan of various distributions is performed, identifying significant deviations from the Monte Carlo simulation. We outline the importance of systematic uncertainties, which are taken into account rigorously within the algorithm. Possible detector effects and generator issues, as well as models involving supersymmetry and new heavy gauge bosons have been used as an input to the search algorithm.

1 Introduction

The start-up of the Large Hadron Collider (LHC) promises to enter an unknown territory in particle physics. Still it is not clear which effects of new physics will appear in the first data, and which theory beyond the Standard Model (SM) will describe them. In addition, the initial understanding of the detector and the validation of the Monte Carlo (MC) generator predictions is an important task which precedes any search for new physics. In order not to miss any unexpected signal and in order to give a consistent picture of a large part of the phase space, we plan to systematically analyze the data with as little bias as possible. For this purpose a special algorithm called "MUSiC" (Model Unspecific Search in CMS) has been developed. Details of the strategy, the algorithm and some representative results will be presented, assuming a statistics of 1 fb^{-1} of simulated pp collisions at 14 TeV.
MUSiC does not aim to repeat, nor to compete with, the work of the many dedicated analyses in CMS. As a global physics monitor it is an alternative approach and a complementary addition to the CMS physics program. Significant deviations found by MUSiC have to be interpreted by physicists. The suspicious final state(s) will have to be investigated in detail in order to determine the reason for any deviation: it may be a detector effect, a lack of understanding of the event generation and simulation, or a new physics signal. In this way MUSiC can contribute to the process of understanding the detector and tuning the MC simulation. Similar strategies have already been applied successfully at other accelerator experiments, see e.g. [1–4].
Such an ambitious strategy has also its drawbacks: For signals where a targeted analysis already exists it is likely to be less sensitive. Since a variety of final states is investigated, one has to rely more on the background predictions made by Monte Carlo generators. Each single final state cannot be studied in such great detail as within a model-driven analysis. There is a possible trade-off between trying to cover a large amount of data and describing all of it properly. One key issue is to estimate and implement uncertainties according to one's best knowledge such that problematic areas of the phase space have a reasonable error assigned.
Before introducing the concept of the MUSiC approach we set the following "guidelines":

- Robustness: well-understood physics objects, i.e. high p_{T}, central $|\eta|$, solid object ID
- Model independence: no optimization of selection cuts
- Simplicity: the steps of the algorithm should be easy to follow, standard statistical estimators and methods should be preferred
- Completeness: include any possible systematic differences between data and MC
- Allow for deviations that contributes predominantly to a single channel (resonances) and physics that produces deviations in numerous final states (SUSY).

In order to have a well-defined trigger stream and in order to reduce the QCD multi-jet background we restrict the analysis to events containing **at least one lepton** (electron or muon). Events are classified into so called **event classes** which group events according to their final state topology. Each event class is defined by the number of physics objects in the event, e.g. 1μ 3jet. We consider *exclusive* and *inclusive* event classes: In the inclusive classes we require only a minimal number of particles (e.g. 1μ 3jet $+ X$, so at least one muon and 3 jets).
We consider the following **physics objects** measured by the CMS detector: *Muons (μ), electrons (e), photons (γ), hadronic jets (jet) and missing transverse energy ($E_{\text{T}}^{\text{miss}}$)*. This leads to approximately 300-400 event classes [1] to be considered. Including additional objects once they are well-understood such as τ's or b-jets is planned for the future.

[1] including photons, however dedicated physics results with photons are not presented; note that with data this number will be defined by the amount of classes populated by the data

The LHC is designed to probe the high-energy frontier, thus the analysis assumes that new physics will appear in events with high-p_T objects. Variables sensitive to new physics which are analyzed systematically by the search algorithm within each MUSiC event class are:

- the **total cross section**, i.e. number of events per class.
- **kinematic distributions**, e.g. scalar sum $\sum p_T$ of all its physics objects.

2 Monte Carlo Simulation

The CMS software framework [5] is used in order to process the simulated samples and to reconstruct the physics objects. All samples are generated with the full detector simulation and originate from the MC production during the Computing, Software and Analysis Challenge 2007 (CSA07). The physics objects are reconstructed assuming the 100 pb^{-1} scenario for detector alignment and calibration.
The Standard Model backgrounds W+jets, Z+jets and $t\bar{t}$ +jets backgrounds are simulated with the ALPGEN [6] generator. Diboson samples, lepton enriched QCD, bottomonia and charmonia events as well as QCD multi-jets, photons+jets backgrounds and minimum bias events are all simulated using PYTHIA [7]. Some SUSY samples are generated using the SoftSusy [8] program to calculate the mass spectrum, and PYTHIA for the event generation. For other processes beyond the SM, like Z′, PYTHIA is used as well.
For the dominant backgrounds of many new physics signals ($t\bar{t}$ +jets, W+jets and Z+jets) a constant k-factor has been applied consistently for all sub-samples in order to reweight the leading order cross section to the next-to-leading order prediction (obtained from MCFM [9]). For SUSY the Prospino 2 [10] next-to-leading order cross sections are used.

3 Event Selection

In this analysis, we use the single and di-lepton HLT triggers (electron/muon) [11], with and without isolation requirement. This specific trigger menu should be robust, even at the start-up of the machine. The efficiency of the trigger menu after all cuts is typically $80-90\%$ for SUSY events with respect to the selected ones.
The aim of the selection cuts is to analyze standard physics objects which are well-studied within the experiment, even if this implies some loss of statistics. Relying on standard physics objects MUSiC can benefit from dedicated studies which e.g. determine efficiencies from data or develop selection cuts well-suited for rejecting misreconstructed objects in real data.
Muons which are measured by the muon system and the inner tracker are selected with $p_T(\mu) > 30\,\text{GeV}$ (well above the trigger threshold) and $|\eta(\mu)| < 2.1$ (inside the trigger acceptance). Isolation is required, i.e. the sum of the transverse momenta of all tracks in a $\Delta R = 0.3$ cone around the muon (excluding the muon track itself) divided by the muon p_T has to be less than 10%, mainly to reject non-isolated muons from heavy flavor decays. Additional offline criteria, e.g. on the number of hits of the track or on the χ^2 of the track-fit, are applied to reject misreconstructed muon candidates.
Electrons are identified using tracker and calorimeter information. We select, as in the muon case, isolated well-identified electrons with $p_T\,(e) > 30\,\text{GeV}$ and $|\eta(e)| < 2.5$. The identification takes energy ratios, the shape of the electromagnetic cluster and the matching between the track and the calorimeter-cluster in η and ϕ into account.
Jets reconstructed with the "iterative cone" algorithm with a radius of $R = \sqrt{\Delta\phi^2 + \Delta\eta^2} = 0.5$ are selected. Jet energy scale corrections are applied and jets with p_T (jet) $> 60\,\text{GeV}$ and $|\eta(\text{jet})| < 2.5$ are selected. The hadronic fraction of the jets is required to be $E_{\text{had}}/E_{\text{tot}} > 0.05$.

Finally, the missing transverse energy is considered when exceeding $E_{\mathrm{T}}^{\mathrm{miss}} > 100\,\mathrm{GeV}$, after accounting for jet energy scale corrections and subtracting muon momenta.
Further cleaning cuts based on ΔR criteria are performed to reduce ambiguities, misidentifications and duplications of the various physics objects.

4 The Search Algorithm

The composition of the selected events, i.e. the number of muons, jets etc. determines to which event class it is assigned. At the present time three distributions are investigated 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 $\sum p_{\mathrm{T}}$ of all physics objects, e.g. for the class 1μ 2jet $E_{\mathrm{T}}^{\mathrm{miss}} + X$ one calculates $\sum p_{\mathrm{T}} = p_{\mathrm{T}}(\mu) + p_{\mathrm{T}}(\mathrm{jet}_1) + p_{\mathrm{T}}(\mathrm{jet}_2) + E_{\mathrm{T}}^{\mathrm{miss}}$.
- Invariant mass M_{inv} of all physics objects (transverse mass M_{T} for classes with $E_{\mathrm{T}}^{\mathrm{miss}}$).
- For classes with missing transverse energy: $E_{\mathrm{T}}^{\mathrm{miss}}$.

The $\sum p_{\mathrm{T}}$ distribution is the most general observable. The invariant mass has an obvious advantage for new particles produced as resonances. Models beyond the SM which aim to provide a dark matter candidate might be spotted with the $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution. The implementation of even further variables is straightforward.

All distributions are input to the MUSiC algorithm (similar to H1 analysis [2]) which scans them systematically for deviations, comparing the MC prediction with the measured data. Each **connected bin region** is considered within the distributions, i.e. individual bins (bin 10 or bin 200) as well as broad regions (e.g. bins $3 - 100$). For each connected region, a counting experiment is performed, adding up the various Monte Carlo contributions (N_{SM}) and comparing this sum to the measured data (N_{data}). In addition to these two numbers also the systematic uncertainty of the prediction δN_{SM} is used. A Poisson probability is computed, determining how likely the prediction fluctuates to the number seen in the data. The systematic uncertainties (section 5), taking correlations into account, are included using a convolution with a Gaussian function:

$$p = \sum_{i=N_{\mathrm{data}}(0)}^{\infty(N_{\mathrm{data}})} A \cdot \int_0^{\infty} \mathrm{d}b \exp\left(\frac{-(b - N_{\mathrm{SM}})^2}{2(\delta N_{\mathrm{SM}})^2}\right) \cdot \frac{e^{-b}\, b^i}{i!} \quad \text{if} \quad N_{\mathrm{data}} \geq (<) N_{\mathrm{SM}} \quad , \tag{1}$$

where A ensures the normalization. From all possible combinations of connected bins, the region with the smallest p-value ($p_{\mathrm{min}}^{\mathrm{data}}$) is chosen. This is called the **Region of Interest**. This approach is sensitive to an excess of data as well as a deficit – be it spread out or narrow. A bin width of 50 GeV is chosen for all three distributions to absorb detector resolution effects.
One should emphasize that this definition of p represents a Bayesian-frequentist hybrid since the true value of the background is one of the b's in the Gaussian integration. As pointed out in [16] correct frequentist coverage cannot be guaranteed for all the parameter space and comparisons with other classes of algorithms are desirable for the future.
The statistical estimator p alone is not sufficient to claim any signal. A statistical penalty factor has to be included to account for the number of regions investigated. To obtain the **event class significance** (per distribution) of the deviation found in the first step, toy Monte Carlo experiments (HDH = Hypothetical Data Histogram) are performed including all uncertainties, assuming the background-only hypothesis, and scanning again all possible regions. The event class significance of the deviation is defined as:

$$\widetilde{P} = \frac{\text{number of HDH with } p_{\text{min}}^{\text{SM}} \leq p_{\text{min}}^{\text{data}}}{\text{total number of HDH}} \quad . \tag{2}$$

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. The $\widetilde{P}$ can be translated into standard deviations and is comparable to the widely used CL_{b} [17].

At the present time, we have no N_{data} to compare with the MC prediction. However, we can pick typical models and test the sensitivity of MUSiC. Instead of only producing pseudo-data for the background-only hypothesis, we can also assume *signal + background*, i.e. add a signal to the SM distributions. In this way we can repeat several pseudo-CMS experiments and determine the expected event class significance of a possible signal.
Figure 1 illustrates this procedure: The green curve represents the pseudo-experiments where a toy signal plus background are assumed, including all errors. With data this would correspond to a single line. The red curve on the other hand displays the multiple repetition of the SM expectation, thus this represents the second step of the algorithm. The p and $\widetilde{P}$ values stated in the plot 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 a significant deviation can be identified.

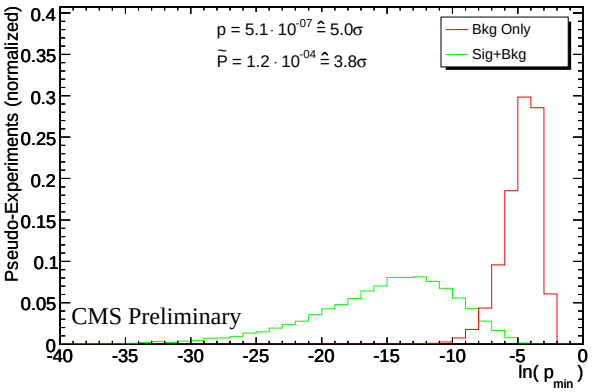

Figure 1: Many CMS experiments using signal+background and bkg-only (SM) hypotheses.

5 Systematic Uncertainties

It is crucial to rigorously implement systematic uncertainty estimates correctly 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 1 fb^{-1} of data, but the values can be changed easily:

- σ(integrated luminosity) = 5%
- σ(cross sections) = 10%, used for all Standard Model background processes
- σ(jet energy scale) = 5%, the uncertainty is also propagated into $E_{\text{T}}^{\text{miss}}$
- σ(efficiency correction factor) = 2% for e, μ and 1% for jets
- σ(fake probability) = 100% for e, μ
- statistical uncertainty of the MC prediction

The 10% cross section uncertainty reflects our current theoretical knowledge which might improve in the future. Also assuming different numbers for different processes is possible within MUSiC. The uncertainty on possible efficiency correction factors refers to differences in the reconstruction efficiencies between data and simulation; the uncertainty on fake probabilities accounts for the fact that complex processes producing fake objects are unlikely to be modeled perfectly by the simulation. It is important to stress that the algorithm also accounts for known correlations within one error when computing p-values for a certain region and when generating pseudo-data for the whole distribution. The contributions of other background sources can be regarded as another systematic uncertainty: cosmic ray muons, beam halo particles and pile-up are an irreducible background which can affect data-MC comparisons. In the future it is planned to mix events from these external sources of particles under the hard collision, using dedicated generators [14] or extracting them from the data. Loose cuts on the extrapolation of the tracks to the vertex (Δz) can help in further reducing these backgrounds.

6 QCD Background Estimation from Data

QCD multi-jet production with its enormous cross section is notoriously difficult to simulate both in terms of computational resources and the underlying physics processes. In the present analysis, we investigate 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. It is envisioned to estimate the QCD contribution from the data. Since in a generic search one is looking at a diversity of final states, it is not practical to define control regions for each specific event class. A more general estimate of the QCD background applicable to all classes is needed. The uncertainties of such cross-class extrapolations have to be absorbed by an appropriate global uncertainty of the QCD estimate, which can be 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 [15]. A single selection cut, which is prominent for distinguishing "fake" leptons from well-measured isolated ones, is inverted or relaxed. The sample with the relaxed 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. Two control regions are defined, $110 - 150\,\text{GeV}$ in the $\sum p_\text{T}$ distribution of the class 1μ 1jet $+ X$, and $130 - 180\,\text{GeV}$ in the $\sum p_\text{T}$ distribution of the class 1μ $E_\text{T}^\text{miss} + X$. These two inclusive classes represent quite different corners of the MUSiC analysis space. 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 the signal contamination is expected to be small, and QCD plus other SM processes dominate. From the control regions we obtain the scale factor with its uncertainty:

$$f_\text{QCD} = \frac{\text{"data"} - \text{SM MC without QCD}}{\text{relaxed "data"} - \text{relaxed SM MC without QCD}} = 0.2 \pm 0.1. \tag{3}$$

The comparison of the QCD estimate from "data" with respect to the QCD Monte Carlo samples is shown in Fig. 2, where the 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.

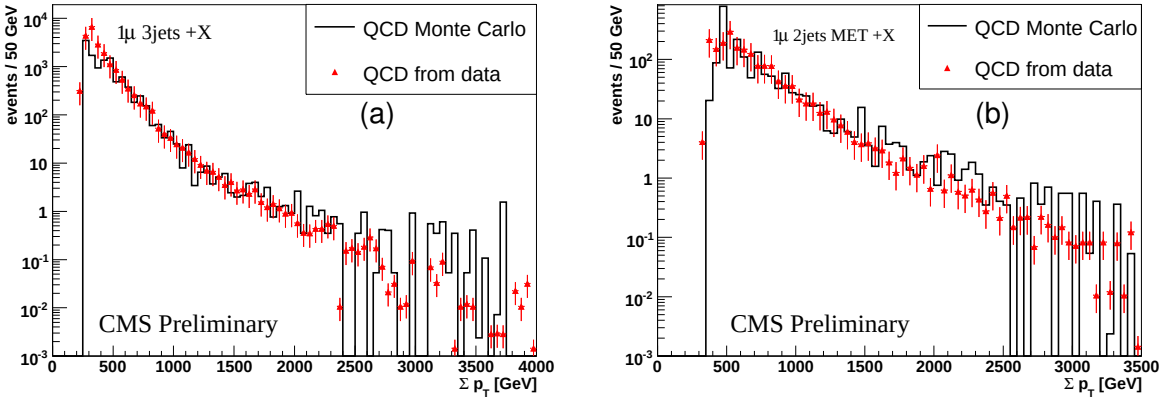

Figure 2: QCD Monte Carlo and estimate using cut inversion for two typical event classes.

7 Results

For a model independent analysis it is difficult to summarize its performance without data to work with. One can only pick archetypical use cases and toy signals which surely can represent only a small part of the phase space covered by such a generic Ansatz. Still we will try to discuss three scenarios which fit to the concept of a model independent search. In order to set a reasonable threshold for a significant deviation we label event classes with $\widetilde{P} < 1 \cdot 10^{-3}$ ($\widehat{=} > 3.3\,\sigma$) as "interesting". As a global physics monitor the main focus of MUSiC is to spot discrepancies in the data worth looking at in detail. In this context it should also be noted that the precise value of $\widetilde{P}$ in the statistical sense is not of major importance. It is rather just a flag defining which deviations are worth to be investigated further.

7.1 MUSiC timeline

One should emphasize that the succession of the following results also reflects the possible timeline of MUSiC in the context of first data analysis:
With the first pb^{-1} the focus will not be on discovering new physics but on re-establishing the Standard Model, understanding the detector and validating the Monte Carlo predictions. One would concentrate on the high statistics parts of the distributions where the SM candles dominate. In this way it is also possible to ensure that initially one is not overwhelmed by deviations found by the algorithm, thus reducing the amount of distributions to be studied in detail. In this phase of data analysis MUSiC can contribute to the understanding of the detector and the tuning of the generators, as shown in the examples below.
In a next phase the focus will shift also to the tails of the distributions where higher order effects like jet-multiplicities become important. The validation of the MC predictions will be crucial and comparisons of different event generators, e.g. PYTHIA vs ALPGEN, will be important. Here MUSiC can contribute, comparing data and MC prediction in a large part of the phase space. While one event generator might describe one part of the data properly, it might fail in another part. Each time new generator parameter tunes are available MUSiC can compare them to data in a general way and thus play an important role in the overall generator validation.
After all initial problems have been solved and confidence in the understanding of the detector and the MC prediction has been established, the full dataset available can be analyzed and one can start looking for deviations from the SM, as demonstrated in the following sections.

7.2 Detector effects and generator tuning

In the context of this demonstration of a detector/generator effect we do not include the multi-jet background since its large statistical fluctuations distort the results and since it is irrelevant for the message here. With the data multi-jets will be considered.
As a first example we introduce a detector effect which is totally unexpected: We disable the jet energy scale (JES) uncertainty of 5% and assume a JES of 10%-up in the pseudo-data. In this context we reduce the assumed cross section uncertainty from 10% to 3%. Here we focus on regions already probed at previous accelerator experiments where uncertainties of 3% or below are realistic. Figure 3 (a) shows the $\sum p_{\mathrm{T}}$ distribution of the inclusive class with 1 muon and 4 jets for a single CMS pseudo-experiment. One can see that the variation in the jet energy scale leads to a considerable excess of pseudo-data in the complete distribution. The region of interest selected by the algorithm is very wide and the p- and $\widetilde{P}$-values are very small. Given the limited amount of background-only repetitions (step 2 of algorithm) only a lower limit for $\widetilde{P}$ can be given, resulting in a significant deviation of at least 4.4 σ. In general many event classes with jets show significant deviations thus revealing the presence of a problem.
If we now re-introduce the assumed JES uncertainty of 5%, things change: Only an expected $\widetilde{P}$ of 12% $\widehat{=}$ 1.6 σ is left, thus no significant deviation is found. This is caused by the fact that the assumed 5% JES uncertainty absorbs the 10% variation in the pseudo-data. This example underlines the importance of including all major systematic uncertainties correctly.

In the second toy example we exercise the possibility of a MC prediction not describing the data properly. In order to evaluate this scenario we compare pseudo-data which are produced using ALPGEN W+jets samples to a SM expectation which relies on a PYTHIA inclusive $W \rightarrow e\nu$ sample. Thus one would expect pseudo-data with higher jet-multiplicities and harder jet momenta. Since these higher order effects become important especially in the tails of the distributions we re-introduce the 10% cross section uncertainty. Given the fact that the PYTHIA sample has only small statistics it is clear that in the very high-p_{T} tails and for high jet-multiplicities the PYTHIA sample does not contribute at all, leading to huge deviations with respect to the ALPGEN-pseudo-data. Thus the results presented here should not be interpreted as a detailed generator comparison but only as an illustration of the MUSiC performance.
Due to the lack of statistics we restrict ourselves to the inclusive $1e\ 1jet\ E_{\mathrm{T}}^{\mathrm{miss}} + X$ event class where we still get PYTHIA contributions. We set an upper bound of 1350 GeV in $\sum p_{\mathrm{T}}$ for the search regions since above PYTHIA suffers from a lack of statistics. Figure 3 (b) shows the $\sum p_{\mathrm{T}}$ distribution of this inclusive event class, comparing ALPGEN-pseudo-data to a MC prediction using the PYTHIA $W \rightarrow e\nu$ sample. One can see a clear overall excess of pseudo-data and the algorithm picks out the tail of the distribution where the observed excess well exceeds the systematic uncertainties. The overall excess can be explained by the fact that this event class is inclusive and that > 100 GeV $E_{\mathrm{T}}^{\mathrm{miss}}$ is demanded. Since ALPGEN predicts considerably more events with many and/or energetic jets these can all contribute to the event class. As the jets are more energetic and more numerous the boosts of the decaying W-boson are stronger, leading to more events with large $E_{\mathrm{T}}^{\mathrm{miss}}$. The deviation in the Region of Interest is significant ($\widetilde{P} > 4.4\sigma$) and the deviation would even increase when regions > 1350 GeV would be included.

7.3 Prominent single deviation

Another important aspect of a model independent search would be the detection of a signal in corners of the phase space possibly not covered by specific analyses so far.
As a toy example a Z′ signal with a mass of 1 TeV and a cross section of $\sigma = 365$ fb is used. As it is clear that a dedicated analysis (see e.g. [18]) or a search optimized for identifying mass resonances is superior to our approach, the results should be interpreted as a proof of principle

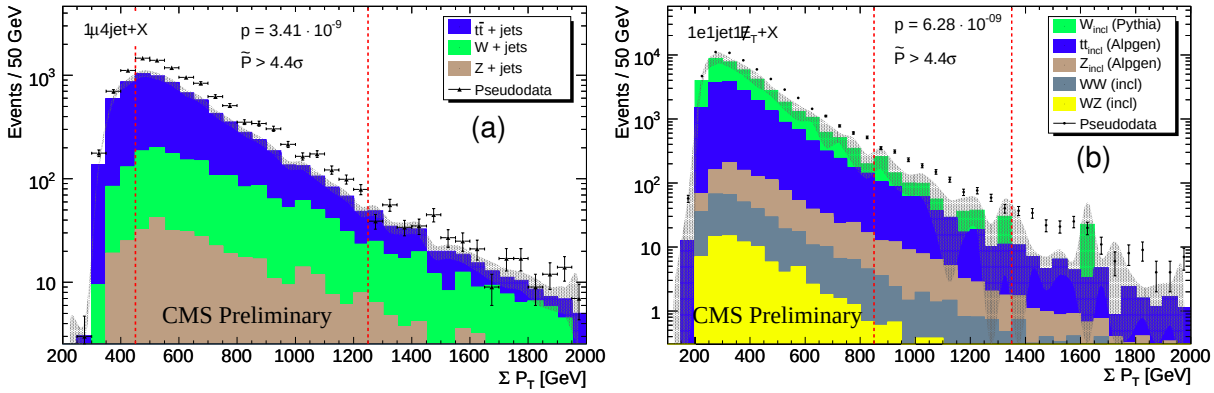

Figure 3: (a) Event class 1μ 4jet+X with JES 10% up in pseudo-data. Region of interest: $N_{\text{data}} = 8831$ and $N_{MC} = 6202 \pm 448$. (b) Event class $1e\,1jet\,\not{E}_{\text{T}}^{\text{miss}} + X$ in case of pseudo-data following an ALPGEN W+jets sample and MC expectation using PYTHIA $W \rightarrow e\nu$ instead. Region of interest: $N_{\text{data}} = 1588$ and $N_{\text{MC}} = 799 \pm 133$. Shaded area: syst. SM MC uncertainty.

that the algorithm is capable of finding a narrow excess. Scanning all event classes the biggest discrepancy between pseudo-data and SM expectation is found in the M_{inv} distribution of the class $2e + X$, as one would expect ($\sum p_{\text{T}}$ as well as the exclusive 2 electron class show also significant deviations). The region of interest (between $950 - 1050$ GeV with $N_{\text{data}} = 52$ and $N_{\text{MC}} = 4.0 \pm 0.7$) clearly selects the Z′-peak at 1 TeV and the p-value of 10^{-36} indicates that the deviation is very large. The $\widetilde{P}$ gives a lower limit of $> 4.4\,\sigma$.

7.4 Complex multiple deviations

There are several reasons why a generic analysis strategy might be a good supplement to SUSY searches. First of all the large number of unconstrained parameters in most SUSY extensions of the SM also leads to an almost unlimited parameter space from which nature can pick the scenario realized. Thus it might be inappropriate to rely solely on analyses optimized on specific sets of parameters. 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 $\not{E}_{\text{T}}^{\text{miss}}$ due to the LSP (lightest SUSY particle). 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 the SUSY signatures above the SM prediction.

We highlight the search results using a typical mSUGRA [13] point:

- **SUSY LM4:** $m_0 = 210\,\text{GeV}$; $m_{1/2} = 285\,\text{GeV}$; $\tan\beta = 10$; $sgn(\mu) = +$; $A_0 = 0$;
  σ (LO) $= 20.5$ pb ; σ (NLO)$= 27.7$ pb

For point LM4, the decay of the $\tilde{\chi}_2^0$ into on-shell Z's is characteristic. $\tilde{q}\tilde{g}$ production is dominant, contributing about half of the total cross section. The squark and gluino masses are below ~ 700 GeV, leading to relatively large total cross sections at 14 TeV center of mass energy. Still LM4 is well beyond the Tevatron exclusion limits.

As a first test we perform a global scan of all event classes, assuming 1 fb^{-1} of data. In total 375 inclusive classes and 315 exclusive classes are populated (either signal and/or background MC). Deviations are found in a large part of the pseudo-data:

- LM4 contributes to 160 (260) *exclusive* (*inclusive*) classes, 94 (170) classes with $\not{E}_{\text{T}}^{\text{miss}}$:
  15% (36%) show significant deviations with $\widetilde{P}$ (expected) $< 1 \cdot 10^{-3}$ in $\sum p_{\text{T}}$
  38% (59%) show significant deviations with $\widetilde{P}$ (expected) $< 1 \cdot 10^{-3}$ in $\not{E}_{\text{T}}^{\text{miss}}$

In the case of inclusive classes deviations found are partially "duplicated" 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 expected p-values and smaller event class significances than the exclusive ones. 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 this LM4 example. On the other hand one can observe a clear gain when using the E_T^{miss} distribution which is a prominent signature of the LSP.

In general there are several classes with a single lepton plus several jets and E_T^{miss} which show significant deviations. Figure 4 (a) shows the E_T^{miss} distributions for a single CMS pseudo-experiment[2]. There is a considerable $t\bar{t}$ +jets and W+jets background left. The systematic uncertainties of these backgrounds, especially the jet energy scale, lead to relative uncertainties of around $20\% - 30\%$, thus signal and background cannot be separated perfectly. Still the distribution shows a deviation well above $4\,\sigma$.
The other type of class with significant deviations contains multi-leptons plus jets ($+E_T^{miss}$). These channels look even more promising in the context of MUSiC since many combinations of 2 or 3 leptons show small p-values and $\widetilde{P}$. Without a dedicated cut optimization with respect to the SUSY signal the mere topology cut of the particle content suppresses the SM background considerably. One example would be the $1e$ 1μ 3jet $E_T^{miss} + X$ event class where a significant deviation $> 4.4\ \sigma$ can be found in $\sum p_T$ (Region of interest between $1000 - 2650$ GeV with $N_{data} = 188$ and $N_{MC} = 61 \pm 18$ leading to $p = 2.6 \cdot 10^{-9}$).

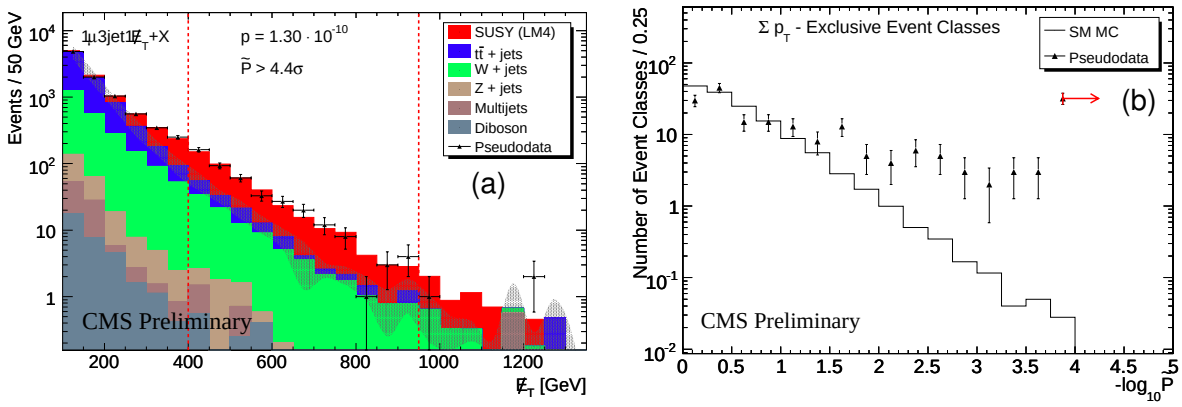

Figure 4: (a) E_T^{miss}-distribution using an event class with single lepton+jets+E_T^{miss}. Region of interest: $N_{data} = 424$ and $N_{MC} = 142 \pm 40$. (b) Frequency distribution of the $\widetilde{P}$ values using all exclusive event classes which have pseudo-data entries, using $\sum p_T$ distribution. Curve: averaged CMS experiment (SM-only), points: single experiment with LM4.

7.5 Statistical interpretation of all event classes

So far the individual event classes have been interpreted apart from the complete set of events. In order to quantify the global compatibility of data and Standard Model expectation one can plot the frequency distribution of the $\widetilde{P}$ values using all event classes analyzed. In a dataset where no signal beyond the SM is present, these $\widetilde{P}$ should peak at large values. If there is a signal leading to significant deviations in several event classes, one would expect the tails of this global distribution to differ from the SM-only scenario. More entries than expected with small $\widetilde{P}$ should be observed, thus a discrepancy in the tails of this distribution between a SM-only scenario and a dataset including some signal should be seen.

[2] note that the QCD MC simulation is used in the figure and not the estimate from data

Figure 4 (b) gives an example for such a distribution, using the $\sum p_T$ distribution in the exclusive case. Here the $\widetilde{P}$ values ($-log_{10}\widetilde{P}$, thus $3 \mathrel{\widehat{=}} 3.3\,\sigma$) of all event classes with pseudo-data entries are shown as a histogram. The black curve refers to the expectation of a SM-only dataset. Here the distributions of several single CMS experiments without any signal are averaged in order to give a reliable prediction. The points correspond to a single CMS experiment assuming SUSY LM4 as signal. One can clearly see that the SUSY contribution leads to significant deviations in numerous classes. Note that classes where only an upper limit can be set ($\widetilde{P} < X$, indicated by the red arrow) all contribute to the rightmost bin.
The issue of global trial factors implies that it might be desirable to reduce the number of distributions examined to a minimum. In the context of MUSiC it is clear that $\sum p_T$ of all event classes will be scanned for deviations in a generic way minimizing any bias towards a certain model beyond the SM. Including transverse mass, E_T^{miss} or possible additional distributions might look promising for certain models, e.g. M_{inv} for Z′ or E_T^{miss} for SUSY.

8 Summary and Conclusions

We have introduced a novel analysis strategy in CMS, MUSiC (Model Unspecific Search in CMS), which systematically scans various final state topologies for significant deviations from the Monte Carlo simulation. Only robust selection criteria are utilized, focusing on well- understood high-p_T objects, without any optimization towards a certain signal. A dedicated algorithm selects the region of biggest discrepancy within a certain distribution and estimates the significance, taking into account statistical as well as systematic uncertainties.
In the context of LHC start-up such a strategy might be an autonomous cross check and a good supplement to the CMS physics program since reliable predictions of how physics beyond the Standard Model could appear are not known. We have discussed several representative scenarios of deviations in the data and demonstrated that as a global physics monitor MUSiC is sensitive to detector effects and generator issues, gold plated signatures as well as complex models such as Supersymmetry.

References

[1] B. Abbott *et al.* (DØ Collaboration), Phys. Rev. D 62 (2000) 092004;
V.M. Abazov *et al.* (DØ Collaboration), Phys. Rev. D 64 (2001) 012004.

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

[3] T. Aaltonen *et al.* (CDF Collaboration), Phys. Rev. D 78 (2008) 012002.

[4] T. Hebbeker, L3 note 2305 (1998),
`http://web.physik.rwth-aachen.de/~hebbeker/l3note_2305.pdf`.

[5] C.D. Jones *et al.*, The new CMS data model and framework, CHEP 06 Proceedings, 2007.

[6] M.L. Mangano *et al.*, ALPGEN, JHEP 0307:001 (2003), hep-ph/0206293.

[7] T. Sjostrand, S. Mrenna and P. Skands, PYTHIA, JHEP 0605:026 (2006), hep-ph/0603175.

[8] B.C. Allanach, Comput. Phys. Commun. 143 (2002) 305-331, hep-ph/0104145.

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

[10] W. Beenakker, R. Hopker and M. Spira, arXiv:hep-ph/9611232.

[11] The CMS Collaboration, The TriDAS Project, Technical Design Report, Volume 1, CERN LHCC 2000-38, 2000; Volume 2, CERN LHCC 2002-36, 2002.

[12] The CMS Collaboration, CMS Physics TDR: Volume I, CERN/LHCC 2006-001(2006).

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

[14] P. Biallass *et al.*, CERN CMS NOTE 2007/024.

[15] See e.g. V.M. Abazov *et al.* [DØ Collaboration], arXiv:0803.0739 [hep-ex].

[16] R.D. Cousins *et al.*, Nucl. Instr. Meth. A 595 (2008) 480–501

[17] A.L. Read, The CL(s) technique, J. Phys. G28, 2693-2704, 2002.

[18] The CMS Collaboration, PAS EXO-08-001, 2008.