Model Unspecific Search in CMS: Results with latest data

Paul Papacz
Erik Dietz-Laursonn, Thomas Hebbeker, Arnd Meyer,
Mark Olschewski, Holger Pieta, Stefan Schmitz
Nayeem Naimuddin, Shivali Malhotra

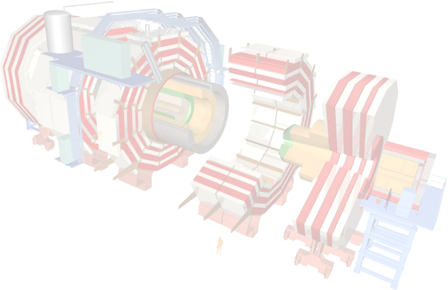
III. Physikalisches Institut A, RWTH Aachen
Department of Physics, University of Delhi

01.12.2010

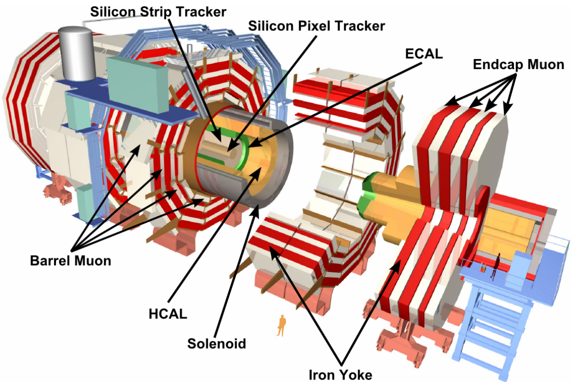
Terascale Alliance Workshop, Dresden 01.-03.12.2010

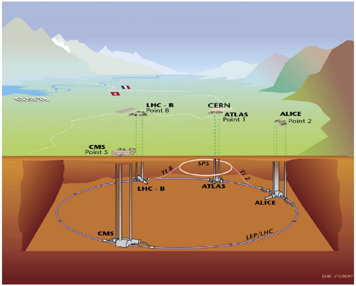
Outline

1. Introduction
2. Implementation
3. Looking at Data
4. Summary/Outlook



CMS @ LHC

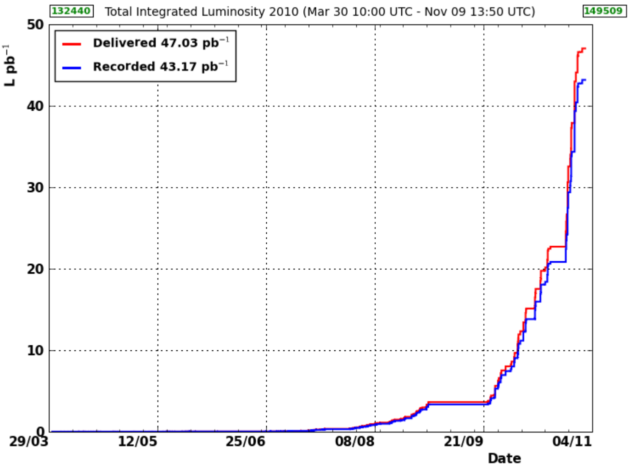




CMS

- length ≈ 21 m
- diameter ≈ 16 m
- mass ≈ 12500 t
- solenoid up to ≈ 4 T

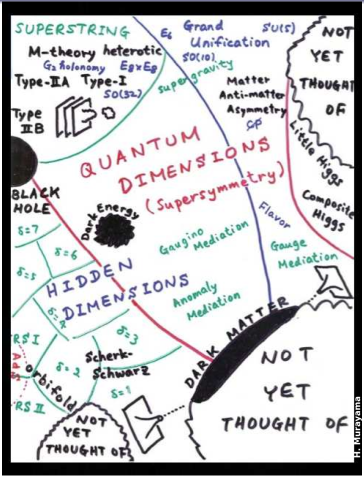
Integrated Luminosity



Data:

- Delivered: 47 / pb
- Recorded: 43 / pb
- Certified: 34 / pb

Motivation



Challenge

Many competing models!
Can we have (dedicated) analyses for all of these?
What are the models no one has thought of yet?

Example: SUSY

A whole “model framework” with many free parameters.

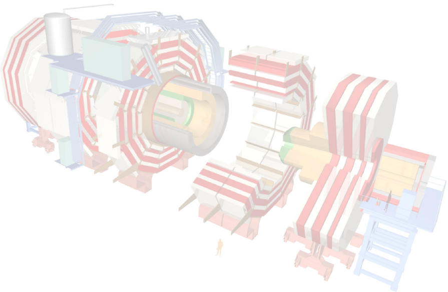
Idea:

Model independent search!

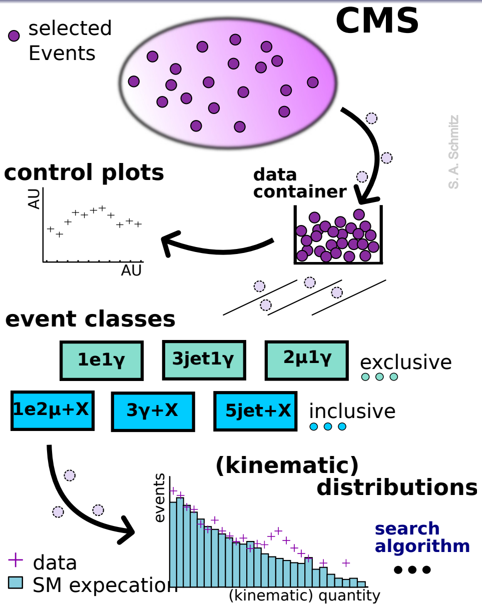
Minimise the theoretical bias:

- Assume only one model: The **Standard Model**
- Search for deviations from the SM expectation in many final states

Outline

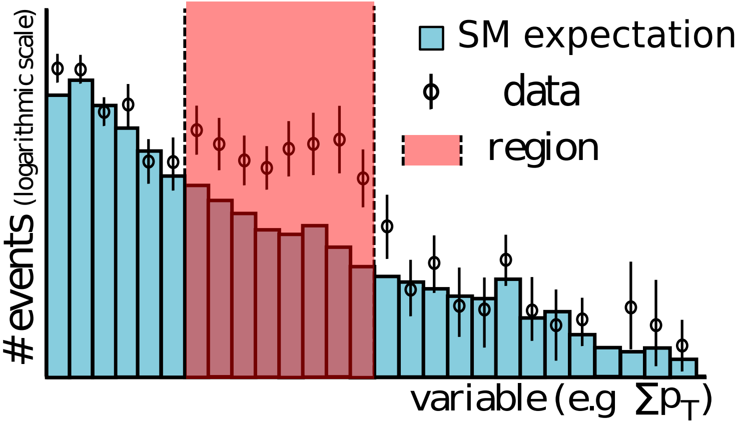


MUSiC Concept



- **Select** good events and objects
- Sort events by their physics object content (μ, e, γ, (b-)jets, $\not{E}_T$) into **event classes**
- **Kinematic distributions** of interest: $\sum p_T$, M_{inv} (, $\not{E}_T$)
- Run the **search algorithm** on these distributions
- Find the most significant **connected bin region** in every distribution

Search algorithm



Most significant region:

In **every distribution** of **every event class**: Find the region with the lowest probability of MC to deviate even more, i. e. the smallest p-**value**.

MUSiC's p-value

Standard treatment:

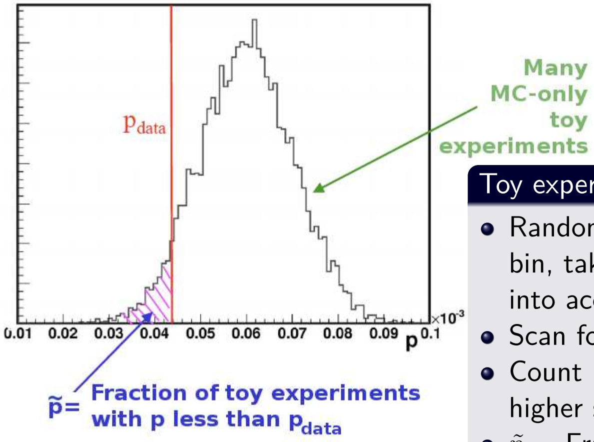
Convolute a Poisson (statics) with a **Gaussian** (systematics) to model the uncertainties on the mean.

Alternative approach:

Use a **Lognormal** prior.
$\rightarrow$ Better treatment of some uncertainties, worse for others.
$\rightarrow$ Good cross check.

"Look-elsewhere effect"

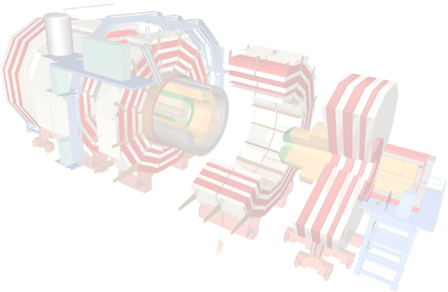
Considering **many regions** in **many distributions** it becomes more probable to see a deviation by chance due to statistical fluctuations.



Toy experiments:

- Randomise MC expectation bin by bin, taking all known uncertainties into account
- Scan for most significant region
- Count pseudo experiments with higher significance than data
- $\tilde{p}$ = Fraction of toy experiments with $p_{\text{toy}} < p_{\text{data}}$

Outline



MUSiC on 7 TeV data

Data:

Used SM backgrounds: Multijet, electroweak, top, γ + Jets, low mass resonances

Major systematic uncertainties:

MC statistics, luminosity (11 %), cross section (5 %), PDF, JES (5 %)

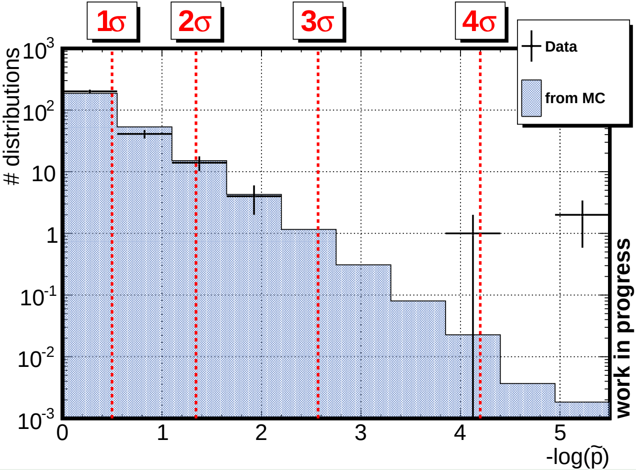
Cuts:

Require at least one e ($p_T > 60$ GeV) **or** μ ($p_T > 25$ GeV)
Allows **loose p_T**-Selection for other objects

Object	p_T^{min} / GeV
e	25
γ	25
μ	18
Jet	50
(MET	30)

$\tilde{p}$ distribution

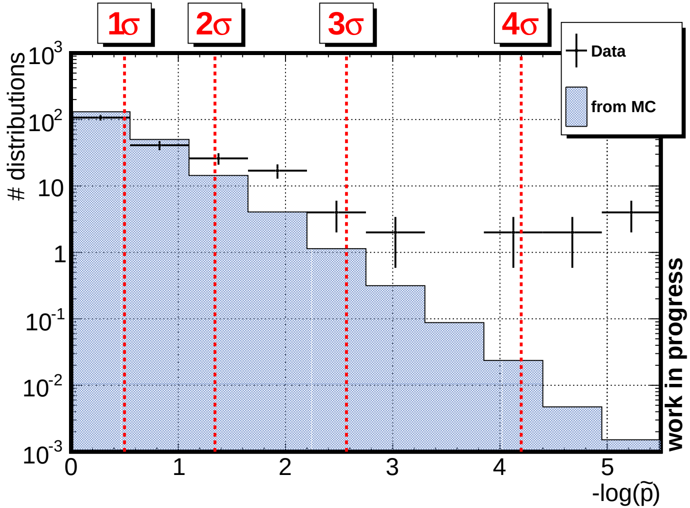
Without MET: Good agreement with MC prediction!



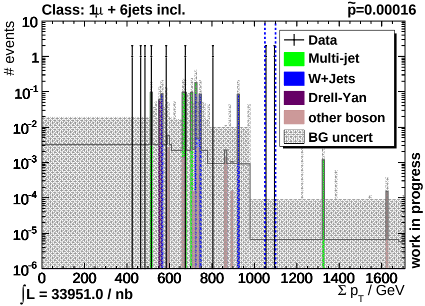
Few discrepant event classes can be explained by instrumental effects or simulation

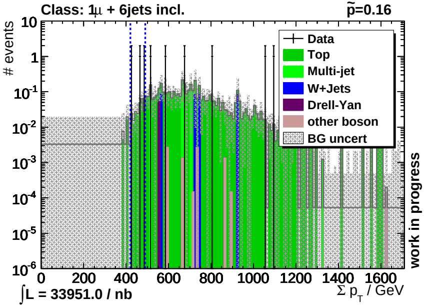
Cross check: Rediscovering the Standard Model

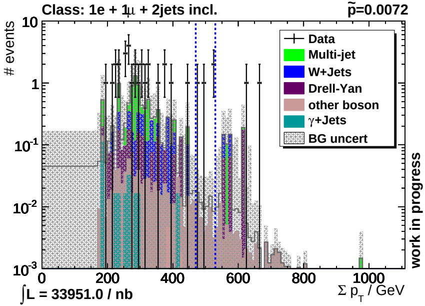
Removed the top quark MC samples:
excess in a wide range of event classes.

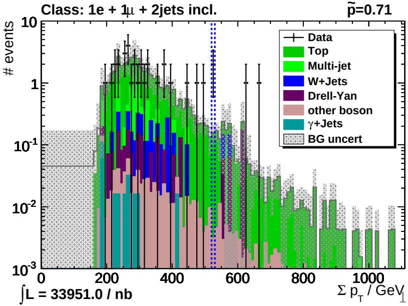


$t\bar{t}$ “most significant”



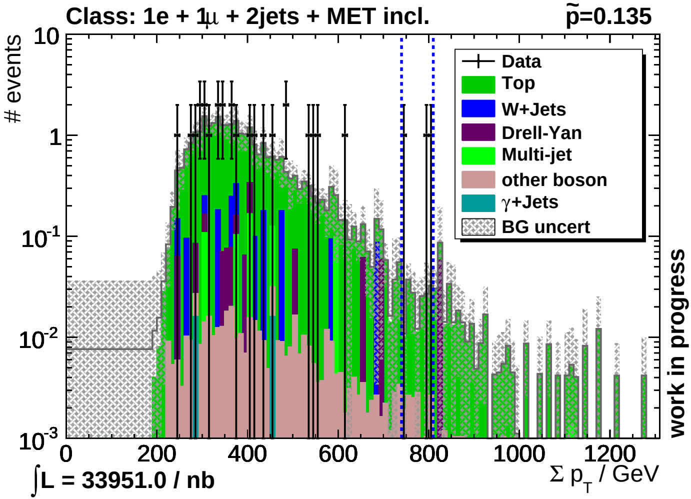




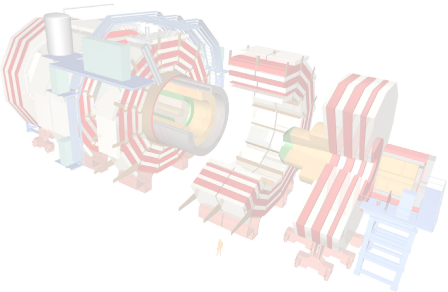


Preview: MET

MET not yet fully implemented in MUSiC but:
Promising first results



Outline



Summary and outlook

Today:

- Model independent search that looks for deviations from the SM
- Complementary to dedicated analyses
- Possible deviations need detailed investigation

Data results:

- Globally good agreement
- Few deviations left (under control)
- Able to rediscover SM physics (in non-standard final states)

Future:

- Analyse more final states

Thank you for your attention!

backup

p-value

Definition:

$$p \equiv \mathcal{P}(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \,|\, H_0)$$

With:

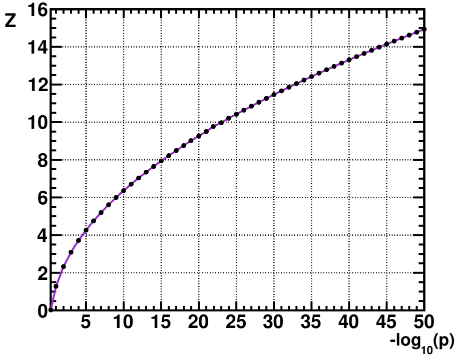
$\boldsymbol{X}$ = Possible data set

$\boldsymbol{x}$ = Observed data set

T = Test statistic

= (number of entries in our case)

H_0 = Null hypothesis = SM



Often p-values are expressed in terms of **standard deviations**.

MUSiC's p-value

Use a **Gaussian prior** to model the uncertainties on the mean of a *Poisson counting experiment* $\Rightarrow$ prior predictive p-value.

$$p^{\mathsf{N}} = \begin{cases} \underbrace{\sum_{i=N_{\text{data}}}^{\infty} C}_{\text{normalisation}} \cdot \int_0^\infty \mathrm{d}\lambda \, \underbrace{\exp\left(-\frac{(\lambda - N_{\text{SM}})^2}{2\,\sigma_{\text{SM}}^2}\right)}_{\text{systematics}} \cdot \underbrace{\frac{\mathrm{e}^{-\lambda}\,\lambda^i}{i!}}_{\text{statistics}} & \text{if } N_{\text{data}} \geq N_{\text{SM}} \\ \sum_{i=0}^{N_{\text{data}}} C \cdot \int_0^\infty \mathrm{d}\lambda \, \exp\left(-\frac{(\lambda - N_{\text{SM}})^2}{2\,\sigma_{\text{SM}}^2}\right) \cdot \frac{\mathrm{e}^{-\lambda}\,\lambda^i}{i!} & \text{if } N_{\text{data}} < N_{\text{SM}} \end{cases}$$

with:

N_{SM} = Pure SM (Monte Carlo) expectation in this region

$\sigma_{\text{SM}} = \sqrt{\sigma_{\text{stat}}^2 + \sum_i \sigma_{i,\text{syst}}^2}$ = Uncertainty on the MC prediction for the SM

Region with the smallest p-value: **Region of Interest**

Another approach: **Lognormal** prior

The "look-elsewhere effect"

Considering **many regions** in **many distributions** it becomes more probable to see a deviation by chance due to statistical fluctuations.

The single region p-value alone is not a good significance estimator.
$\Rightarrow$ Compute "new" estimator $\tilde{p}$:

$$\tilde{p} = \frac{\text{number of } H_0 \text{ experiments with a region featuring } p < p_{\text{data}}}{\text{total number of } H_0 \text{ experiments}}$$

Easy (toy) example: all regions **statistically independent**

$$\tilde{p} = 1 - (1 - p_{\text{data}})^n \;; \quad n = \text{number of regions}$$

In realistic cases regions are correlated due to shared bins and/or **correlated** systematic uncertainties!

Solution:

Dice a sufficiently large number of pseudo experiments in order to determine $\tilde{p}$.

Determination of $\tilde{P}$

① In every distribution the data gives **one most significant region**, i. e. smallest p-value (p_{data}).

② Dice the MC expectation according to uncertainties many times (pseudo experiments) to get a distribution of p-values.

③ Get the "relative number of pseudo experiments with $p < p_{\text{data}}$" $= \tilde{p}$.