MUSiC – A Generic Search for Deviations from Standard Model Predictions in CMS

For the CMS Collaboration (published in CMS PAS EXO-08-005)

- Introduction
- Selection Cuts
- Algorithm
- Results
- Summary

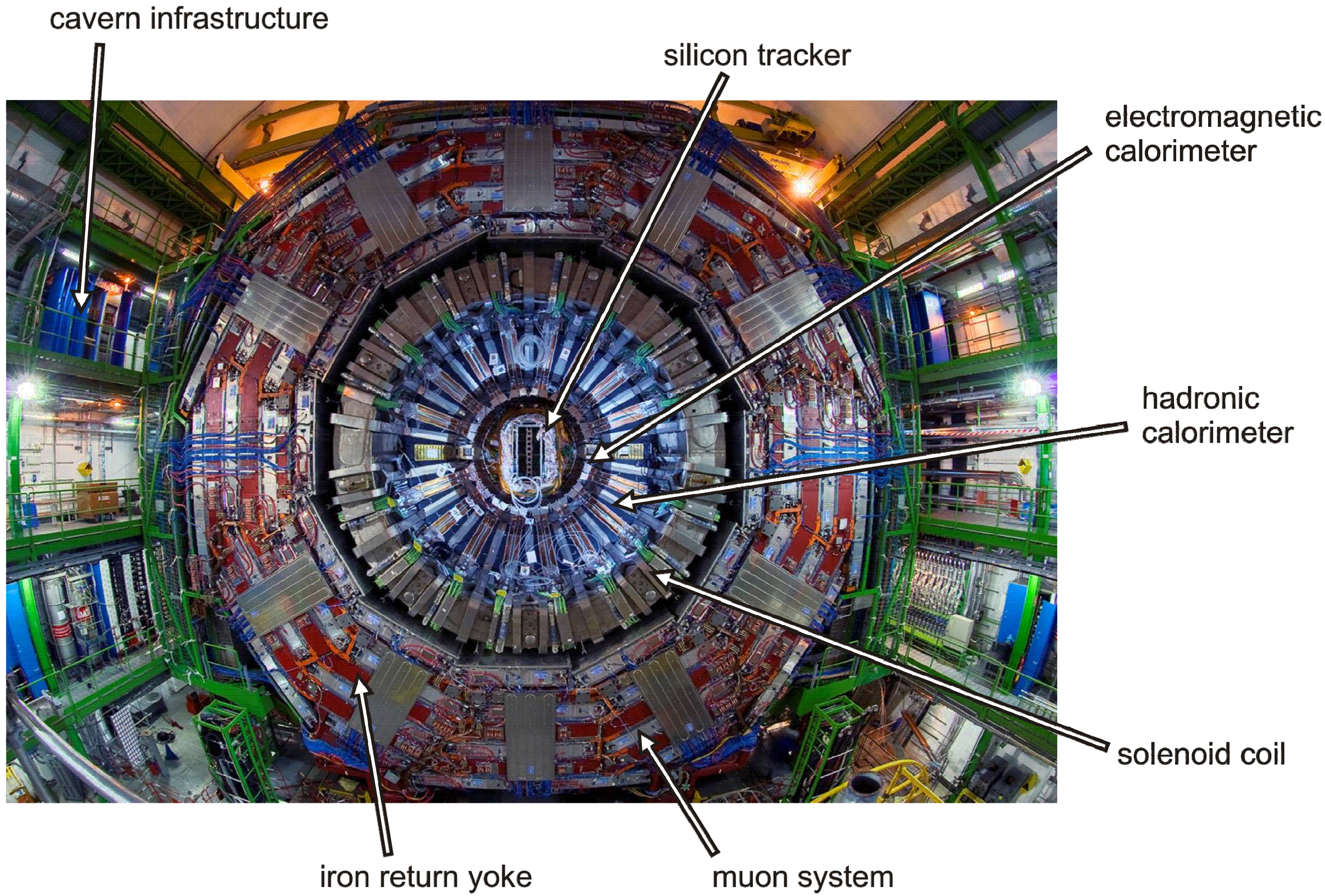
Philipp Biallass

RWTH AACHEN UNIVERSITY

DISCRETE'08, 13. December

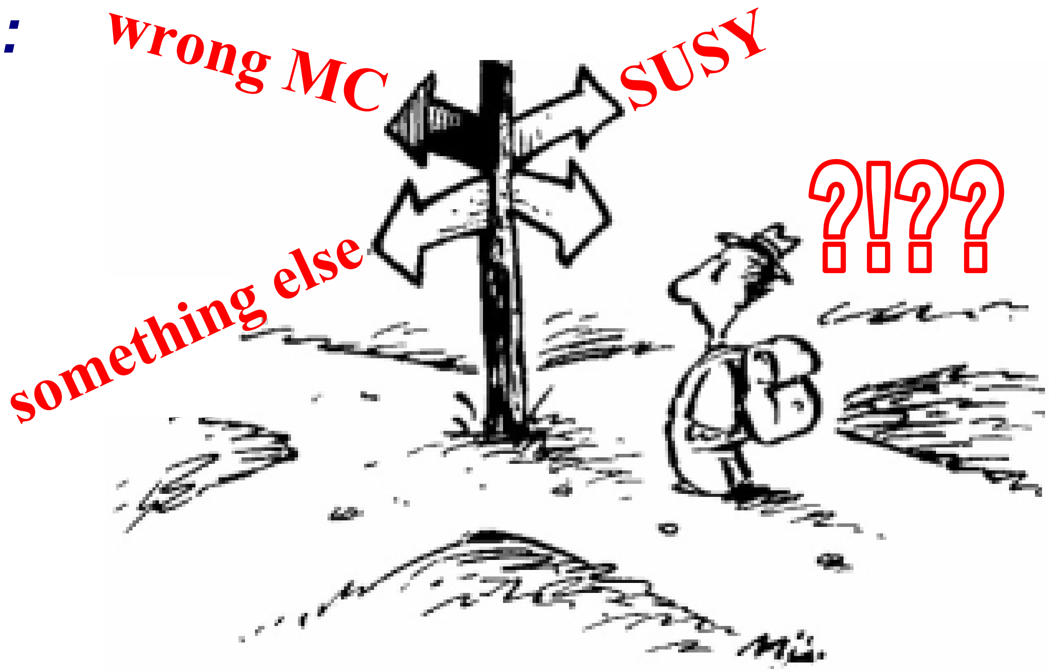
The CMS Experiment

- ***While our colleagues from theory have been working on innovative models beyond the SM in the past 10 years, this what experimentalists have been playing with lately:***

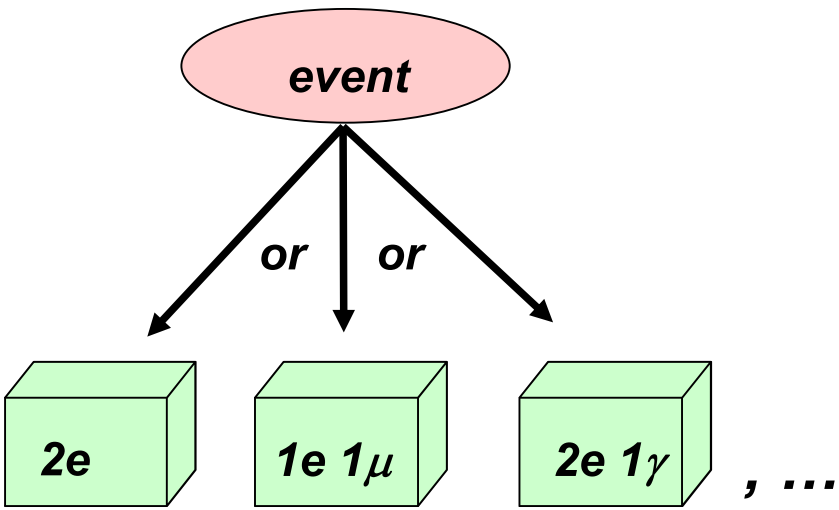


Why Generic?

Our situation at LHC start-up:



- *Good idea to analyze events without expecting certain signal*
- *Broad search, but with less detail*
- *Expect the unexpected!*



Maybe not as easy as it sounds…

- *For us difficult to look into each detail of each final state*

  → have to rely more on Monte Carlo estimate

- *Thus a deviation found is not directly a "discovery" but rather a deviation from the expectation (=Standard Model MC) → need to study deviation*

Common Misunderstanding:

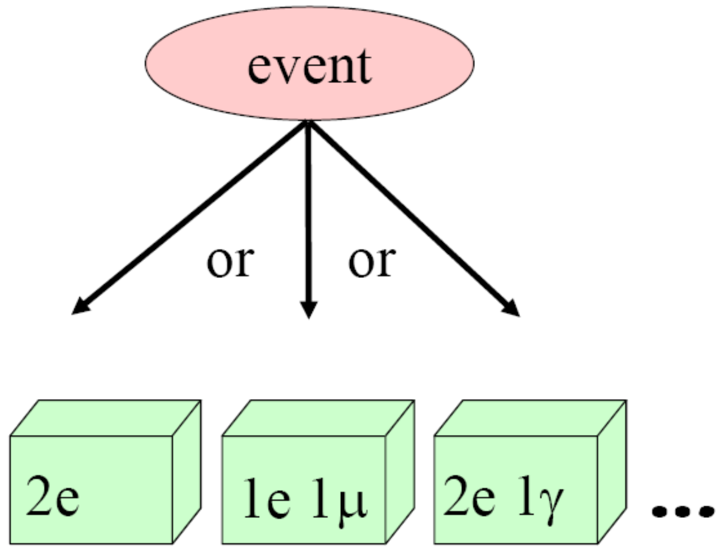
- *MUSiC is not an automated discovery tool, rather a global physics monitor*

History of Generic Searches:

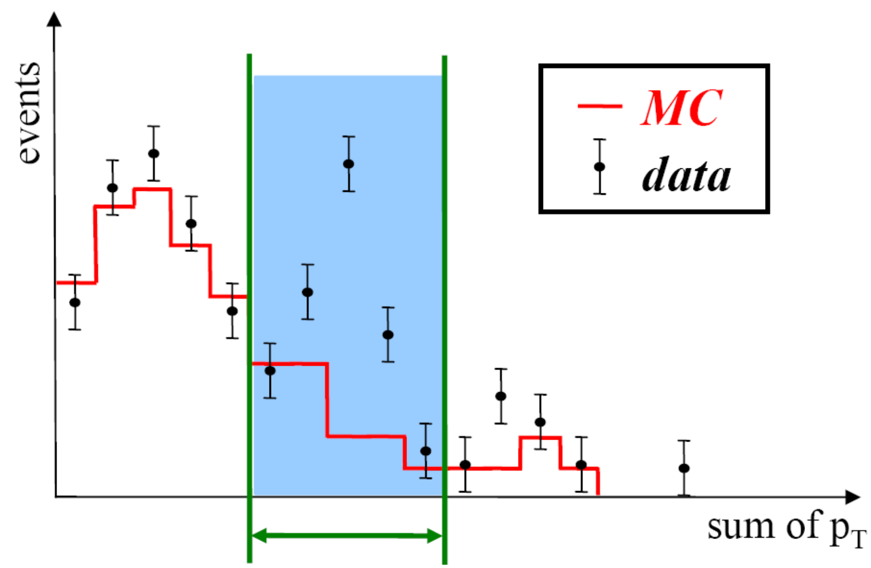
- *Similar strategies already successfully applied at various other accelerator experiments (L3, DØ, H1, CDF)*
  - *e.g. Phys. Rev. D 64 (2001), Phys. Lett. B 602 (2004), Phys. Rev. D 78 (2008)*
- *MUSiC (Model Unspecific Search in CMS) first effort at LHC conditions*

MUSiC: General Concept

- ***Classify* events by particle content**
  - ***Single isolated lepton always required (→easy trigger, less QCD)***
  - ***Exclusive and inclusive (+X) final states (≈300 classes)***
  - ***e, μ, γ, jet, MET***



- ***Scan distributions* for statistically significant deviations**
  - ***Presently Σp_T, invariant (transverse) mass, MET-distribution***
  - ***Dedicated algorithm [H1 publication] searching for biggest discrepancy (excess OR deficit)***
- ***Takes systematic uncertainties into account***



Setup and Selection

- *Assume 1 fb^{-1} of data and 14 TeV center-of-mass energy*
- *Consider realistic composition of SM backgrounds (ALPGEN or Pythia MC)*
- *Use data-driven estimate for QCD background*
- *General strategy:*
  - *Focus on standard objects with standard efficiencies, cuts etc.*
  - *Focus on well-understood objects, even if statistics lost*

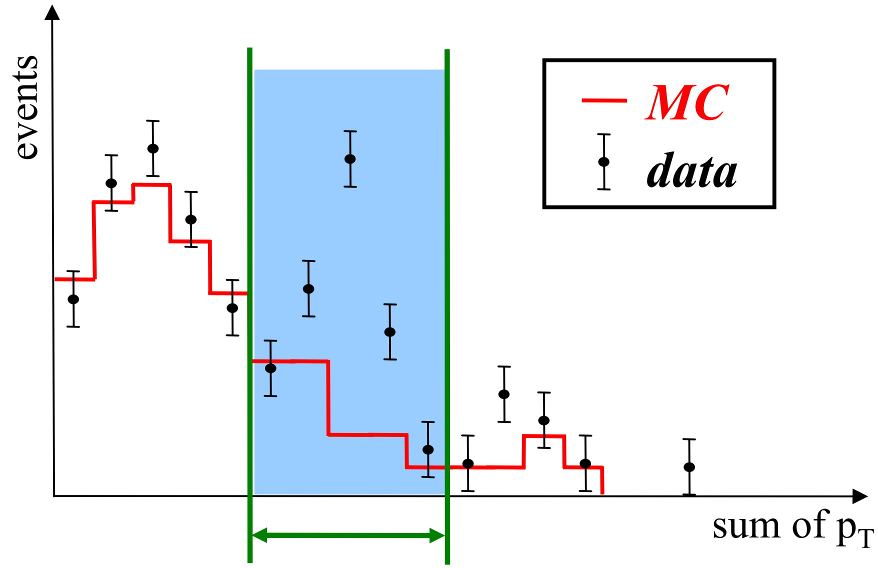
	e/ γ	μ	ItCone Jet (R = 0.5)	MET
p_T cut	30 GeV	30 GeV	60 GeV	100 GeV
$\vert\eta\vert$ cut	2.5	2.1	2.5	/
isolation	tracks	tracks	/	/

→ *high p_T, central η plus several quality criteria (N_{hits}, χ^2, …)*

- *High Level Trigger:*
  Single muon/electron "OR" di-muon/electron HLT (with/without isolation)
- *Trigger efficiency ε_{HLT} typically 80-90%*

MUSiC: Algorithm (following H1 analysis)

- *Define all possible connected regions in every distribution*



- *For each region count N_{data} and N_{MC}*
- *LHC has not started yet: Dice pseudo-data according to uncertainties*

First step:

identify region where "probability" for N_{MC} to fluctuate to N_{data} is smallest
→ *Region of Interest* → p_{data}

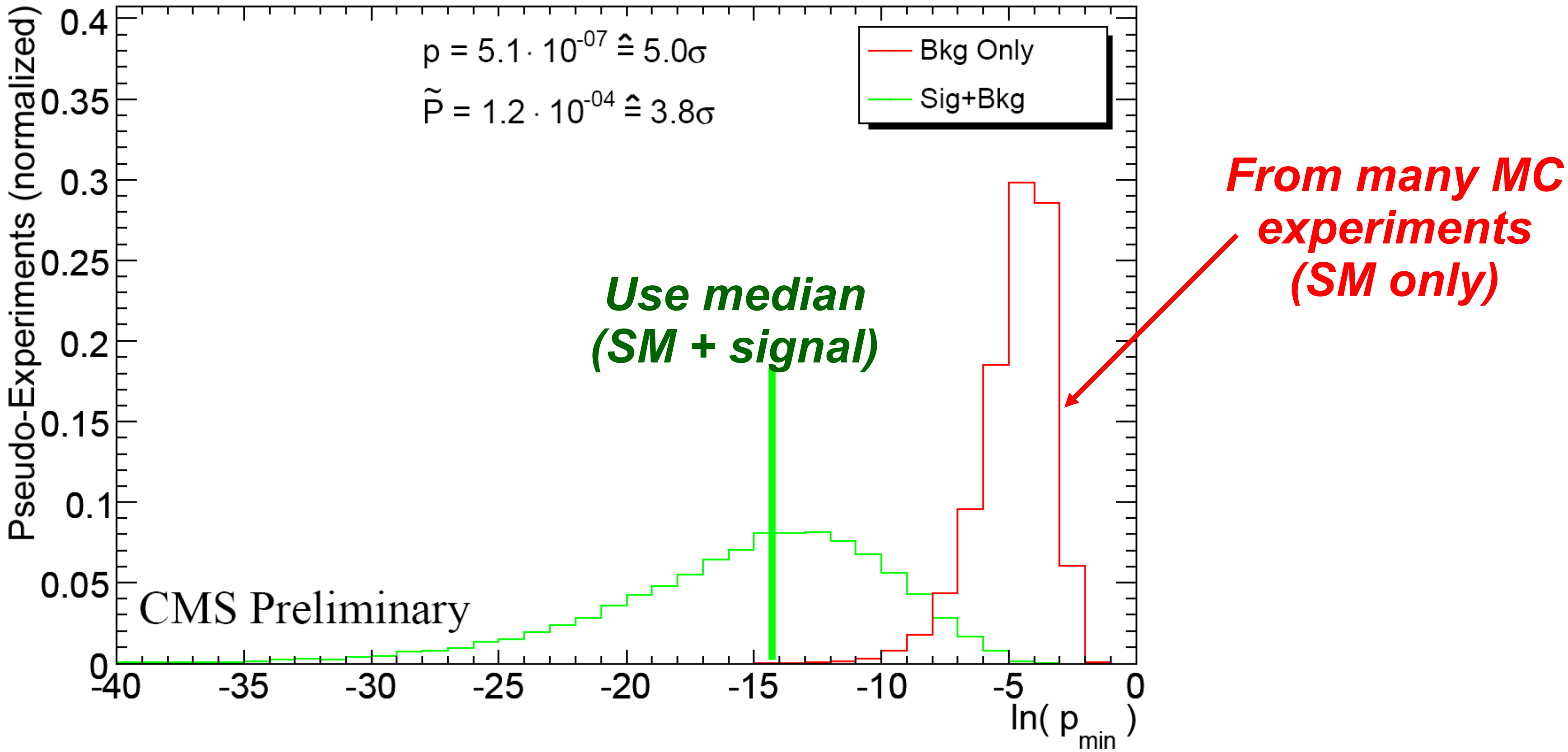
Second step:

Account for "look-elsewhere-effect"

- *repeat "experiment" with bkg-only hypothesis many times (scan all regions)*
- *determine probability $\tilde{P}$ for finding value $p \leq p_{data}$*

Example: $p_{data} = 10^{-6}$ could lead to $\tilde{P}$ = 10% (≈1.6 σ)

MUSiC: From p to $\tilde{P}$



$\tilde{P}$ = fraction of MC experiments (SM only) with p less than p_{median}

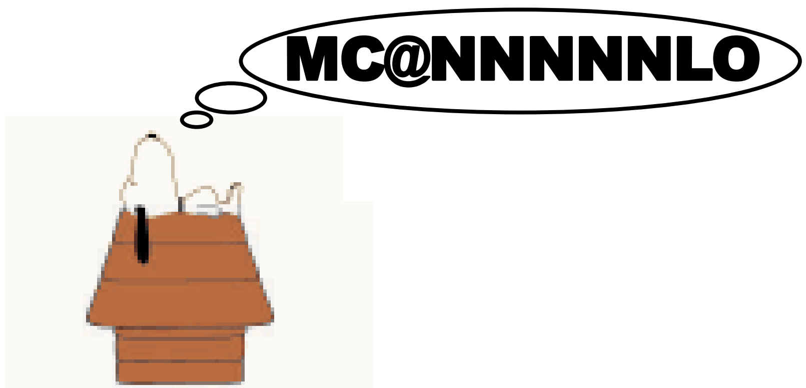
- ***Comparable to widely-known CL_S method, $\tilde{P}$ can be interpreted as CL_B***

Systematic Uncertainties

- *Our limited detector/MC-understanding should be absorbed by systematics*

- *Various systematic uncertainties included, respecting correlations*
  - *Several experimental uncertainties, e.g. 5% luminosity*
  - *Main theoretical uncertainty:*
    10% cross sections (e.g. detailed PDF variation studies yield 2% - 8%)

- *Used flat k-factors for W/Z/tt NLO estimate*



- *Of course this is not the final answer, values are nevertheless reasonable for this kind of analysis approach*

→ *These are only „starting values“, should be refined in the future*

MUSiC Timeline

Focus with first data:

- *Understand the detector, tune the MC generators, re-establish the SM*

After initial problems:

- *Higher order effects in tails, compare e.g. PYTHIA ↔ ALPGEN*

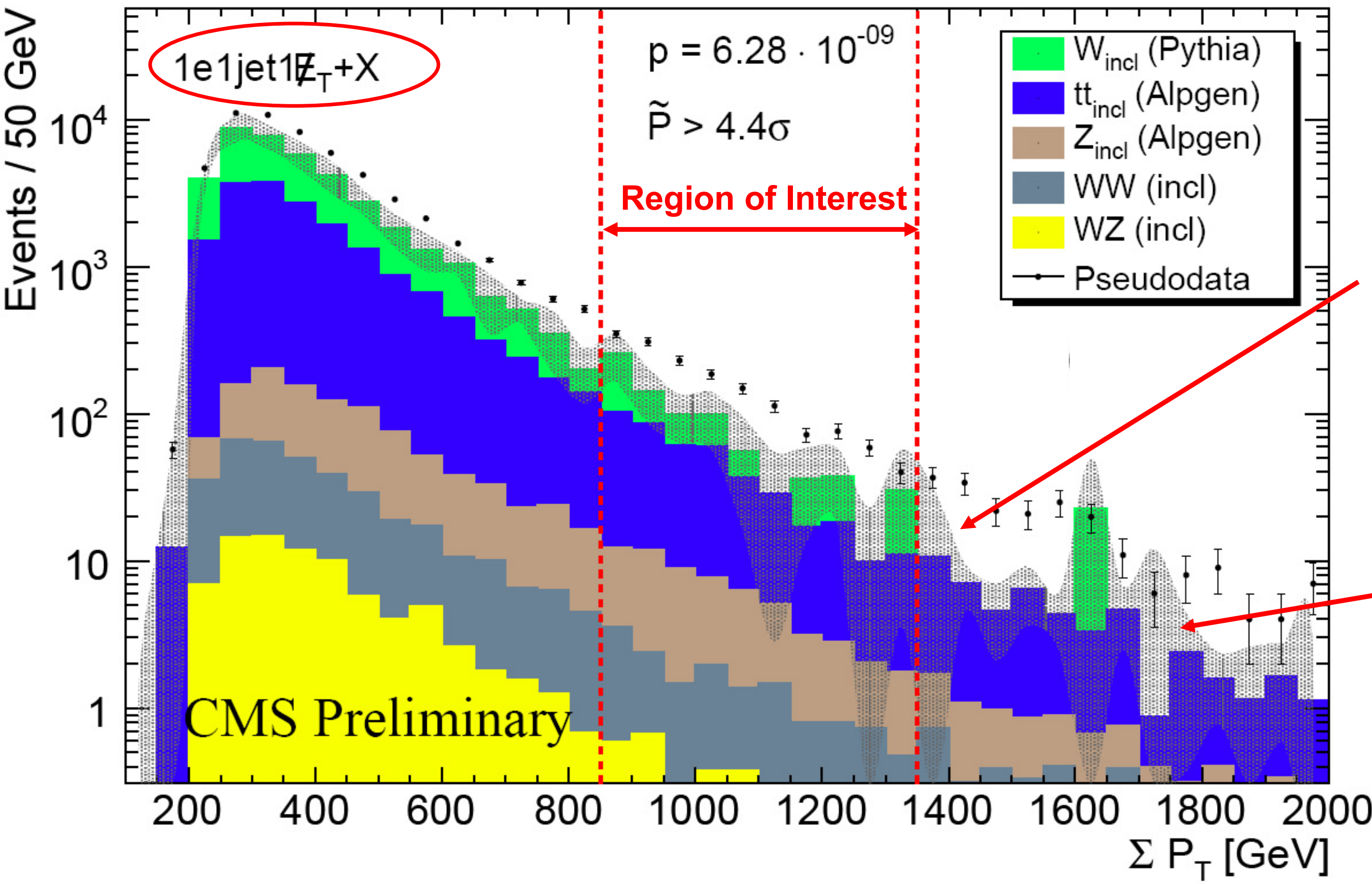
→ ***MUSiC can contribute** to all these points, gives **global picture** of data-MC comparison*

Confidence in detector and MC:

- *Start looking for deviations from the SM, possible signals not covered by specific analyses yet*

Generator Tuning Example

- ***No a dedicated generator comparison, just a toy example!***
- ***Assume W-sample (ALPGEN) for pseudo-data and W-sample (PYTHIA) for MC expectation***



lack of statistics in Pythia sample, constrain to regions **<** 1350GeV

shaded area = syst. uncertainty

- ***Significant deviation (>4.4σ) due to more events with many and/or hard jets predicted by ALPGEN***

New Physics Examples

Prominent single excess: ***1TeV Z' ($\sigma \approx 365$ fb) @ 1 fb^{-1}***

- ***As expected significant deviation ($p = 10^{-36}$, $\tilde{P} > 4.4\sigma$) in M_{inv} of 2e+X class***
- ***Z'-peak nicely selected as Region of Interest → proof of principle***

Complex deviations: ***SUSY***

- ***General search might be complementary strategy for SUSY:***
  - ***do not know which parameters of SUSY-space nature has chosen***
  - ***large SUSY crosssections, can be seen early and with simple cuts***
  - ***long decay chains → complex topologies with many different particles***

- ***LM4-mSUGRA benchmark point @ 1 fb^{-1}***

  ($\sigma_{NLO} \approx 28$pb, $m_0 = 210$ GeV, $m_{1/2} = 285$ GeV, $\tan\beta = 10$, sgn(μ) = +, $A_0 = 0$)

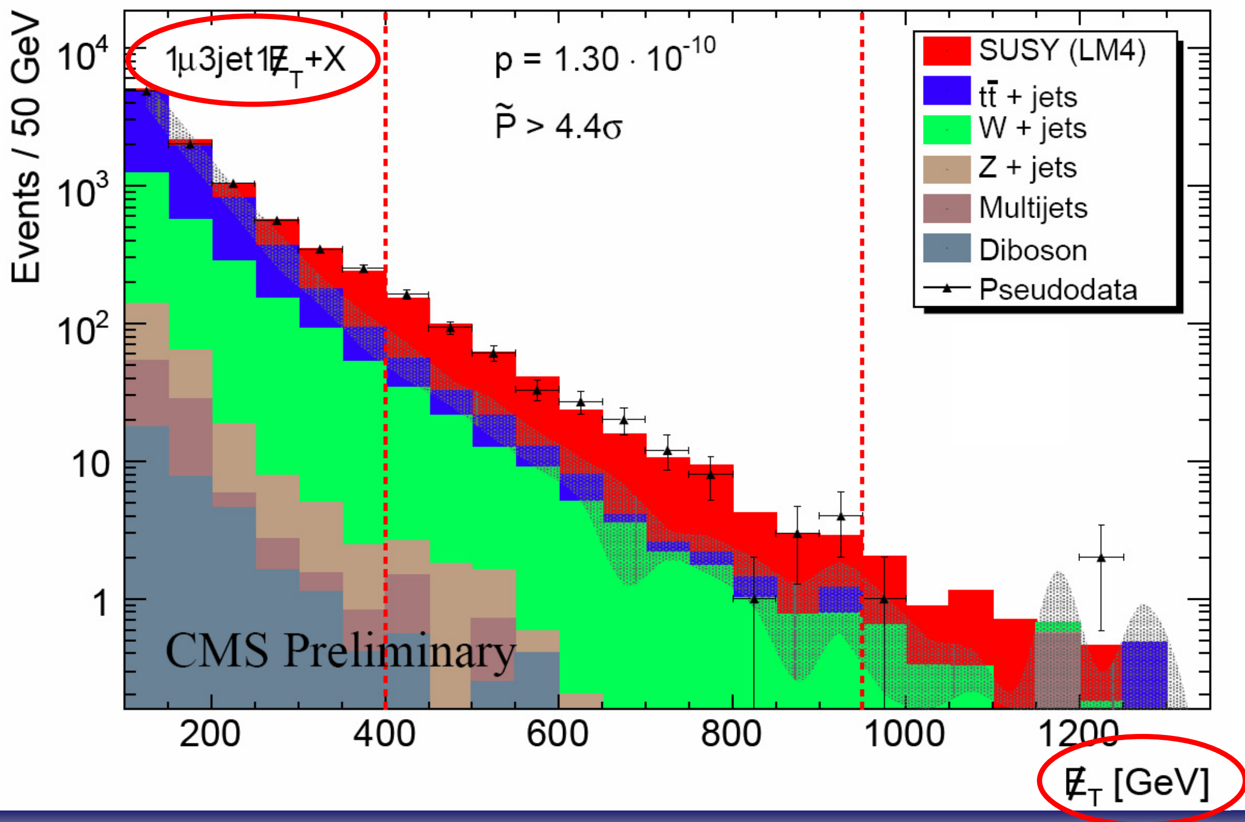
  - ***contributes to large number of event classes***
  - ***several classes with large „data-excess“***
  - ***some classes with only a few events over <<1 event SM background***

SUSY LM4 Results

- ***In total 375 inclusive and 315 exclusive classes are populated***
  - LM4 contributes to 160 (260) *exclusive (inclusive)* classes, 94 (170) classes with E_T^{miss}:
    15% (36%) show significant deviations with $\tilde{P}$ (expected) $< 1 \cdot 10^{-3}$ in $\sum p_T$
    38% (59%) show significant deviations with $\tilde{P}$ (expected) $< 1 \cdot 10^{-3}$ in E_T^{miss}

→ ***Deviations (>3σ) found in many classes, two examples:***

single lepton + jets + MET:



multi leptons + jets + MET:

1e 1μ 3jet MET +X

using Σp_T

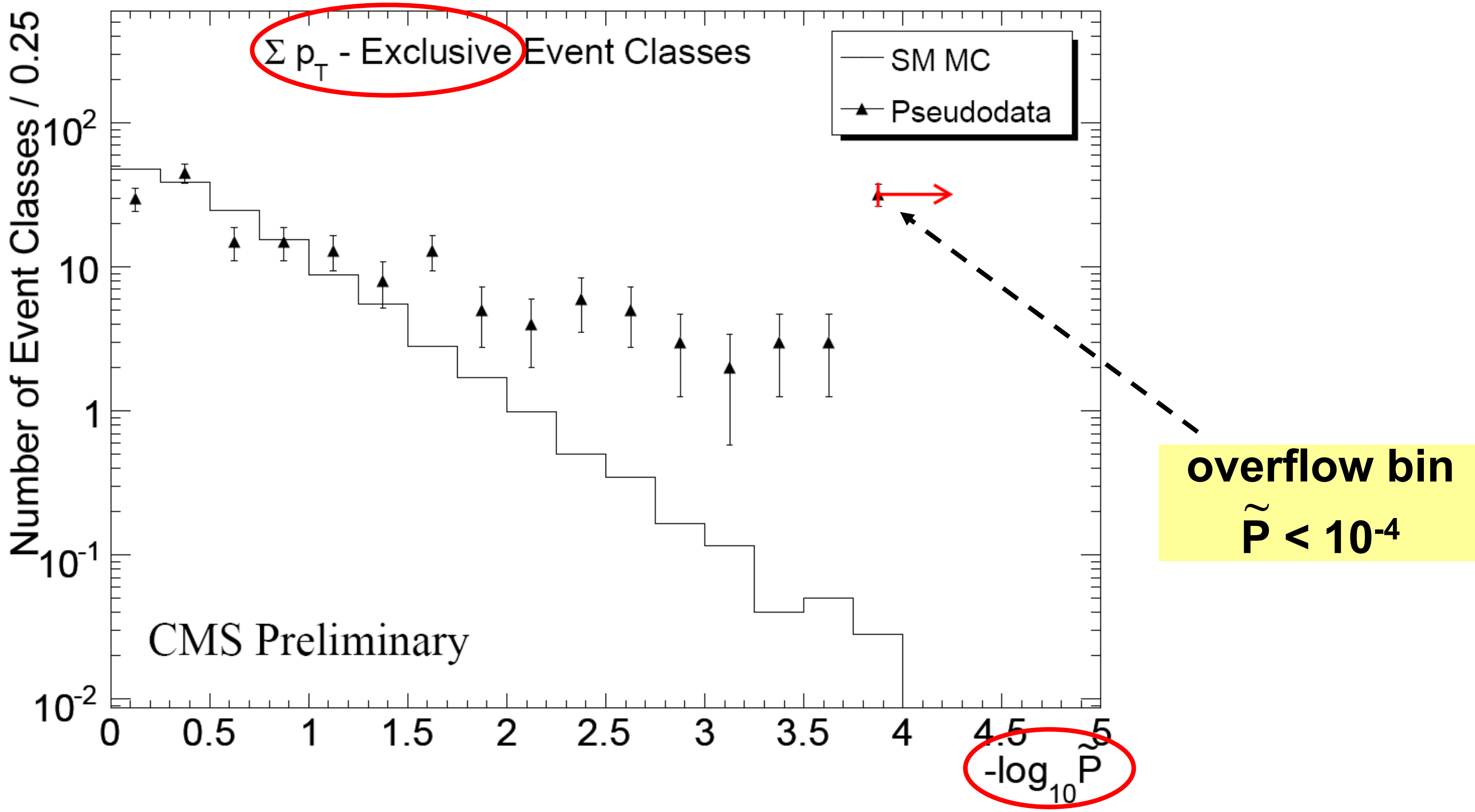
ROI between 1000-2650 GeV

N_{data} = 188, N_{MC} = 61 +/- 18

p = 2.6 ·10^{-9} , $\tilde{P}$ > 4.4σ

SUSY Global Picture @ 1fb^{-1}

- *Plot $\tilde{P}$ of all event classes, shows global data ↔ MC agreement*



- *Deviations in SM-only case compatible with expectation, but with SUSY LM4 tails "explode"*

- *Pseudo-data with SUSY globally disagree with SM-expectation*

MUSiC Summary

- *MUSiC is a complementary analysis strategy with rather global sensitivity*

- *Keep it simple: focus on well-understood objects and selection cuts*
- *Do not over-automatize: deviations need to be interpreted by physicist*

- *Be alert to all possibilities → model independent search*
  - *Helpful to understand detector and backgrounds initially, then physics potential*
- *Demonstrated sensitivity to various models of new physics (Z', SUSY)*

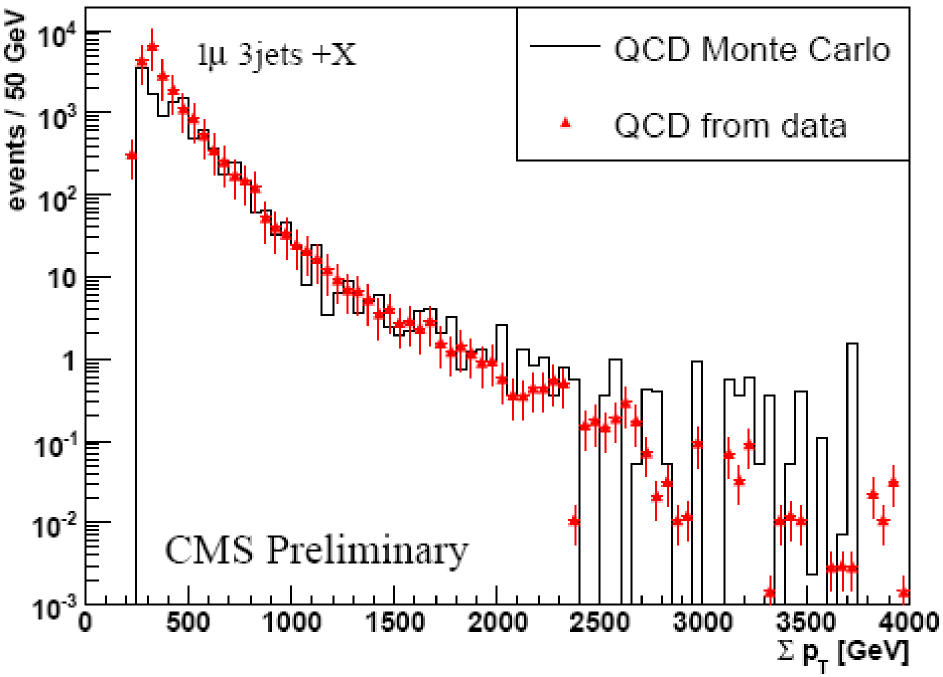
Future Plans:

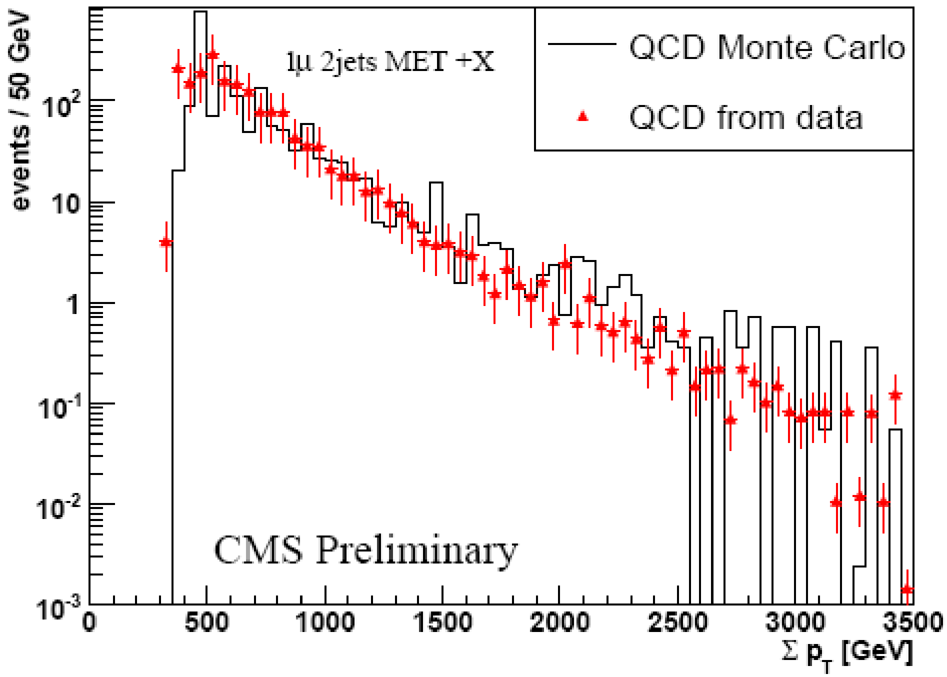
- *Refine treatment of theoretical uncertainties*
- *Include taus/b-jets, use lepton charges, include photon-trigger-stream*

Backup

Data-driven Estimates

- ***Will need QCD estimation from data due to lack of MC statistics***
  - ***"Cut inversion" technique*** $0.1 < R_{track-isolation} = \frac{\sum p_T \text{ of tracks in 0.3 cone}}{p_T(\mu)} < 0.5$
    - ***Relax track isolation to obtain shape of QCD sample***
    - ***Define two control regions to determine normalization and systematics (f = 0.2 +/- 0.1)***
    - ***Extrapolation to other final states seems to work, see plots***





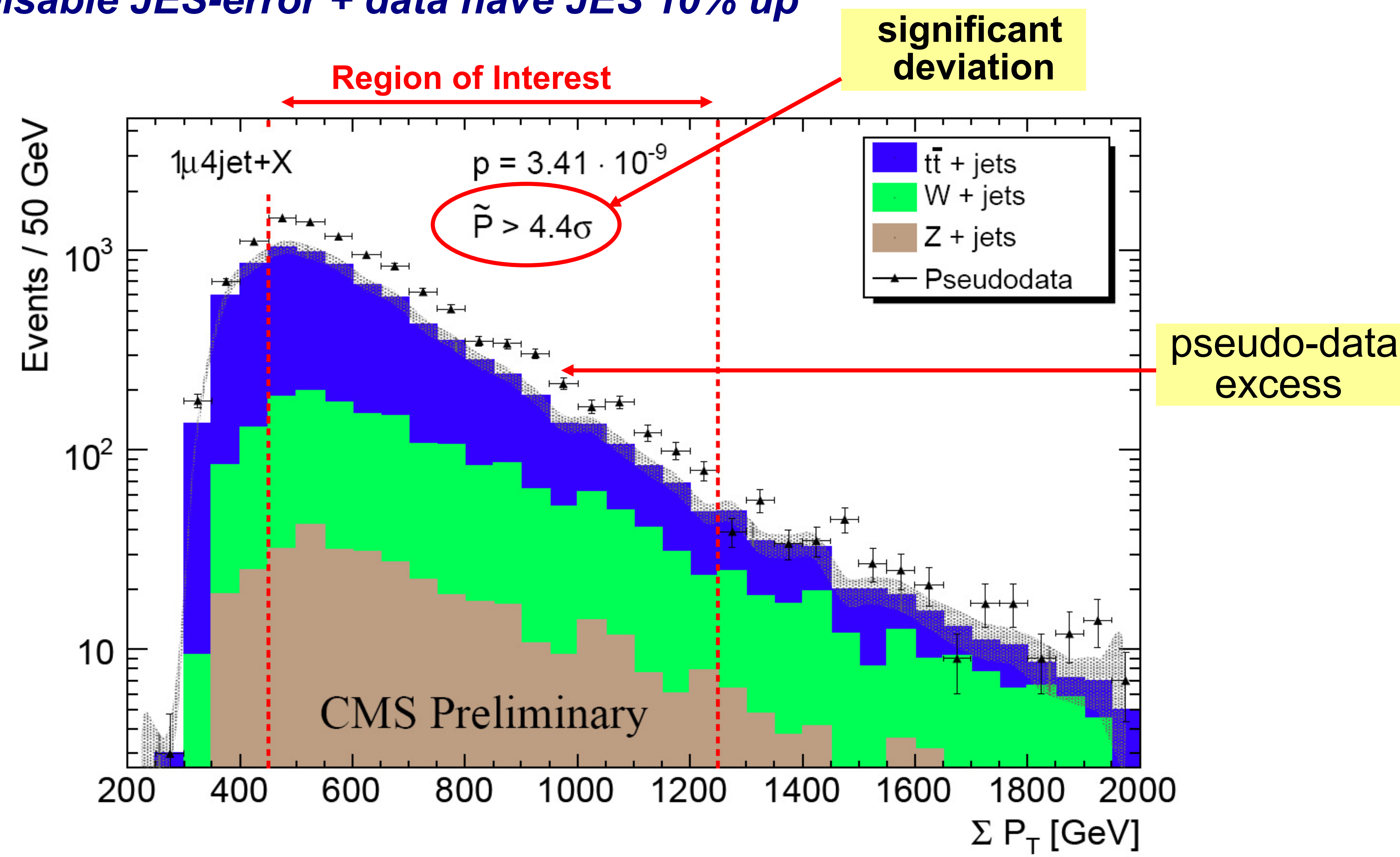
Definition of p-value

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

- ***Convolution of Gaussian (systematics) and Poisson (statistics)***
- ***This is a Baysian-frequentist hybrid method, has reasonable coverage***
- ***Since N_{data}, N_{SM} and δN_{SM} are always stated one can easliy check using alternative statistical methods***
- ***Including syst. errors in statistical estimator long discussed problem, see e.g.*** R.D. Cousins et al., arXiv:physics/0702156v3
- ***MUSiC is an alarm-system for interesting deviations, precise value of p not of major importance !***

Detector Effect Example

- ***Assume unknown detector effect:***
  Disable JES-error + data have JES 10% up



- ***Possible to spot problem in many classes with jets → consistent picture***
- ***Re-enable 5% JES-error: Only 1.6σ effect left***

MUSiC Formula and Algo

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

$$\tilde{P} = \frac{\text{number of HDH with } p_{min}^{SM} \leq p_{min}^{data}}{\text{total number of HDH}}$$

HDH = hypothetical data histogram