Development of a Fast Search Algorithm for the MUSiC Framework

Bachelor Presentation

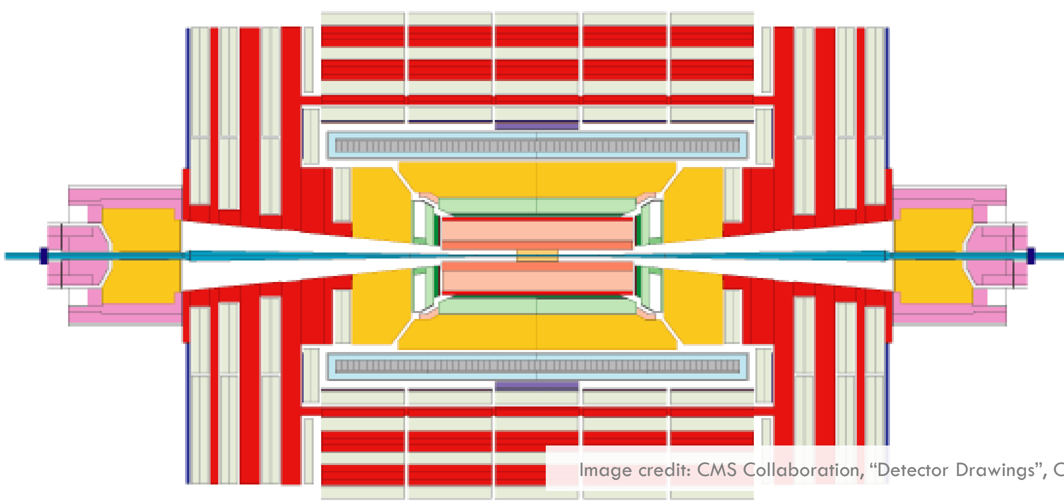
Jonas Lieb, 21.09.2015

Outline

- Introduction to CMS and MUSiC
- Motivation for a fast search algorithm
- Concept of the fast search algorithm
- Optimization and validation
- Summary and outlook

LHC and CMS

- LHC: Large Hadron Collider (CERN), Proton-Proton accelerator, center-of-mass energy of 8 TeV (2012), hosts 4 detector experiments
- CMS: Detector at the LHC, barrel around the beam pipe, featuring silicon trackers, calorimeters, a solenoid magnet and muon chambers



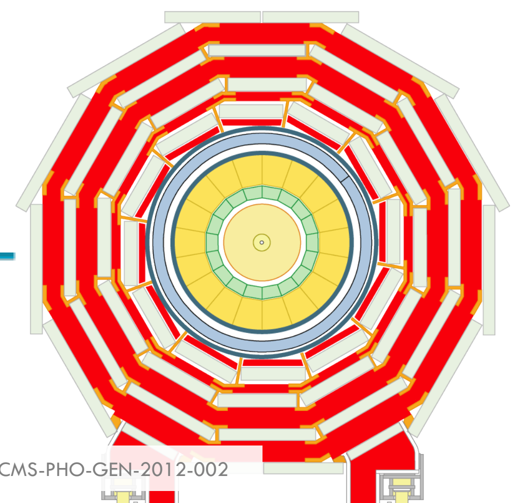
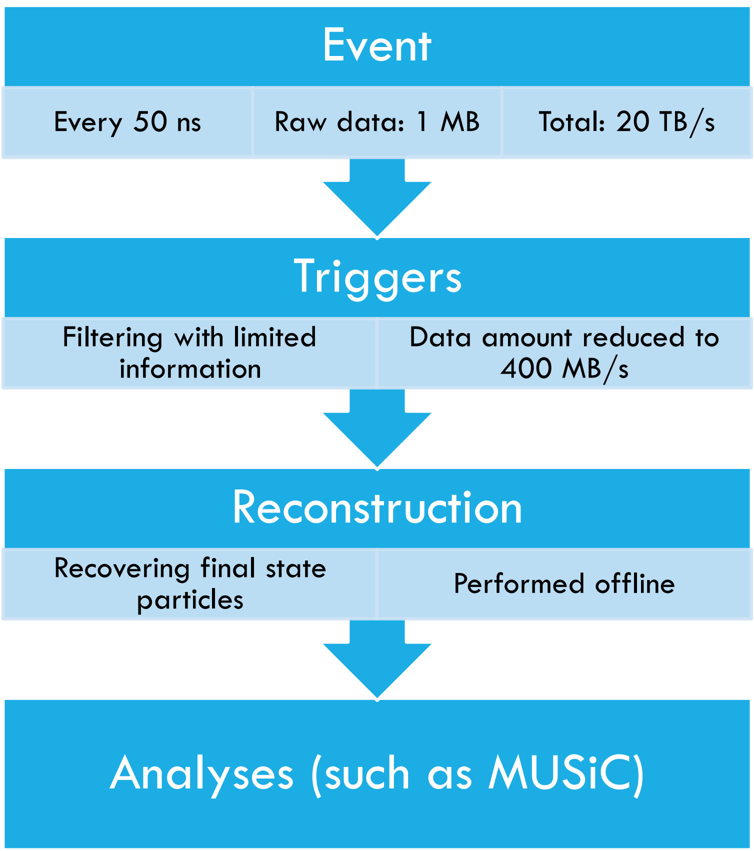


Image credit: CMS Collaboration, “Detector Drawings”, CMS-PHO-GEN-2012-002

Data Pipeline



- Huge amount of data
- 15 PB (10^{15} byte) per year
- Many different final states possible
- Cannot be processed solely by dedicated analyses

→ Complementary method necessary to be sensitive to signs of new physics

MUSiC – The Model Unspecific Search

- MUSiC: Model Unspecific Search in CMS
- Goal: Find new physics beyond the standard model
- Compares measurement from CMS with standard model expectations from Monte Carlo simulations
- Does not focus on one final state, regard many final states at the same time
- Three steps: „Skimming“, „Classification“ and „Scanning“

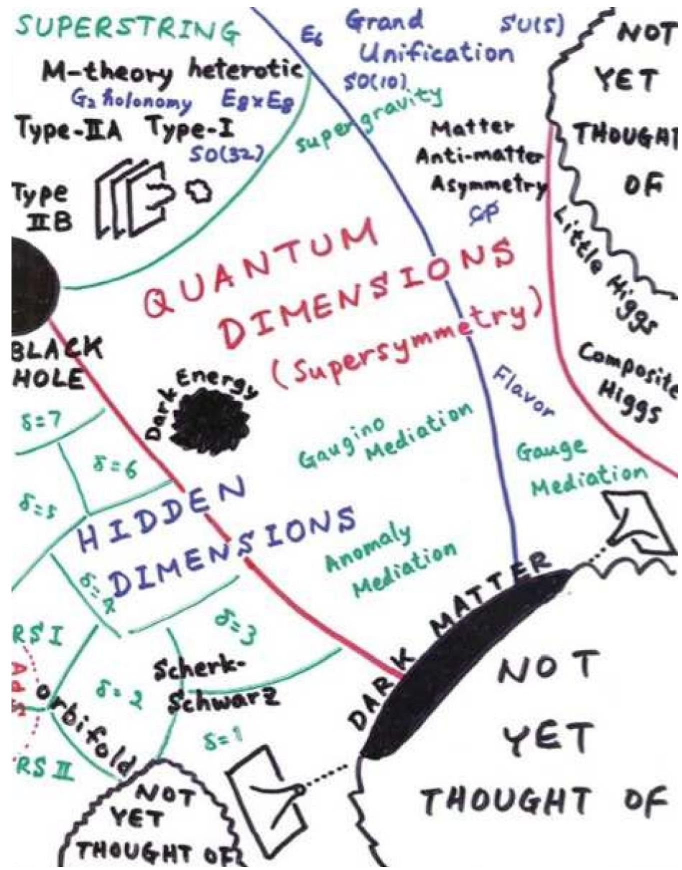
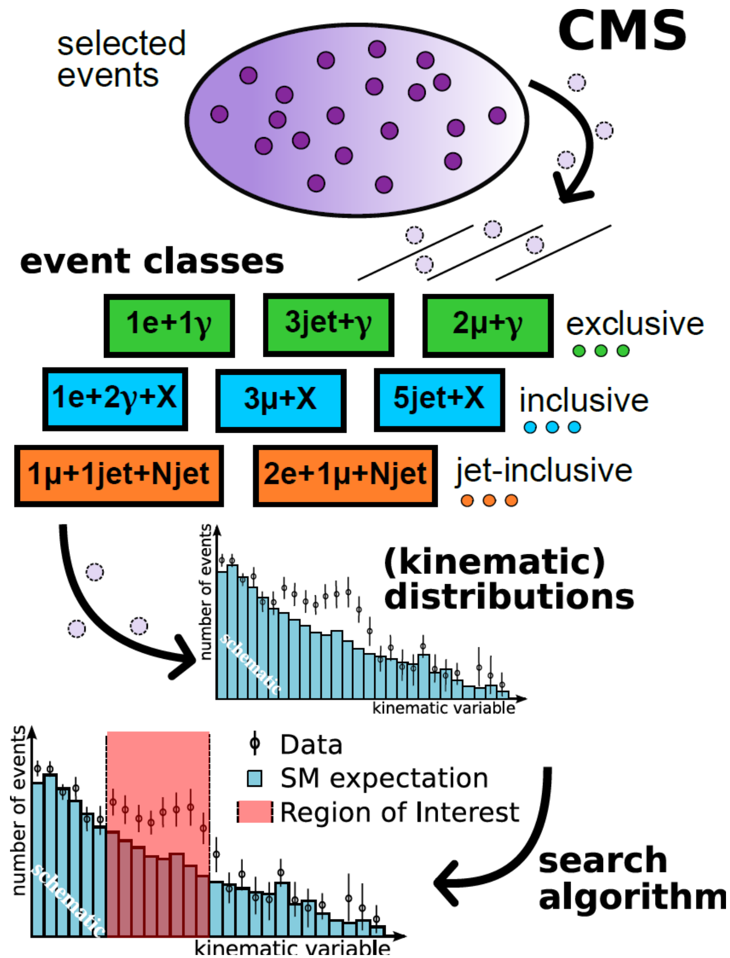


Image credit: Hitoshi Murayama, Berkeley National Laboratory

Skimming, Classification and Scanning

- Skimming: Import CMS data, apply cuts and significantly reduce data amount
- Classification: Group events by physics content into „Event Classes“
- Objects: Electrons, Muons, Photons, Jets, Missing transverse energy
- For each event class, build three kinematic distributions:
  - Sum of transverse momenta $\sum |\overrightarrow{p_T}|$
  - Invariant mass M_{inv}
  - Missing transverse energy MET
- Scanning: In each distribution, search all connected bin regions for the one with the most significant deviation (with the smallest p-value)
- → Region of Interest



The p-Value

- Measure for the probability to observe a deviation between measurement and null hypothesis as least as large as the observed one
- Calculated from the observed event count N_{data}, the expected event count N_{SM} and its systematic uncertainty σ_{SM}
- Poissonian probabilities is used to model statistical nature
- Smear Poisson mean with a Gaussian to model systematic effects

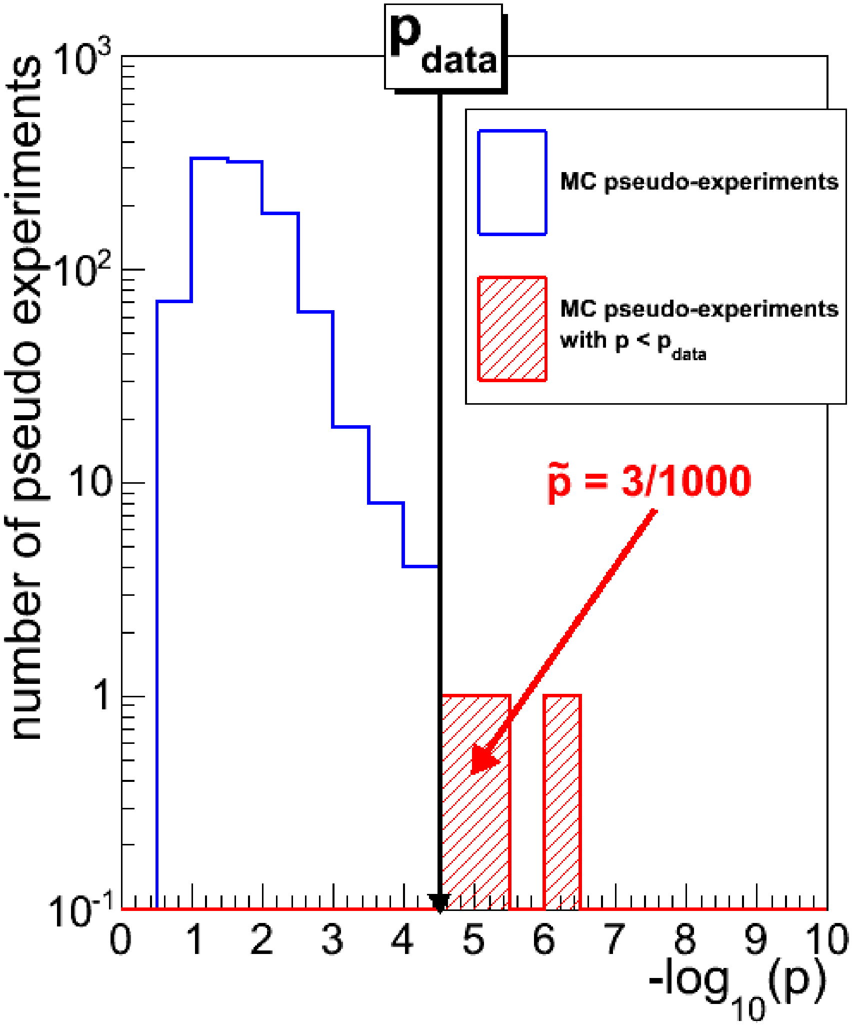
$$p_{\mathrm{data}} = \begin{cases} \displaystyle\sum_{N=N_{\mathrm{data}}}^{\infty} C \cdot \int_0^{\infty} \mathrm{d}\theta \, \exp\left(-\frac{(\theta - N_{SM})^2}{2\,\sigma_{SM}^2}\right) \frac{e^{-\theta}\theta^N}{N!}, & \text{if } N_{\mathrm{data}} \geq N_{\mathrm{SM}} \\ \displaystyle\sum_{N=0}^{N_{\mathrm{data}}} C \cdot \int_0^{\infty} \mathrm{d}\theta \, \exp\left(-\frac{(\theta - N_{SM})^2}{2\,\sigma_{SM}^2}\right) \frac{e^{-\theta}\theta^N}{N!}, & \text{if } N_{\mathrm{data}} < N_{\mathrm{SM}} \end{cases}$$

Correcting for the Look-Elsewhere-Effect

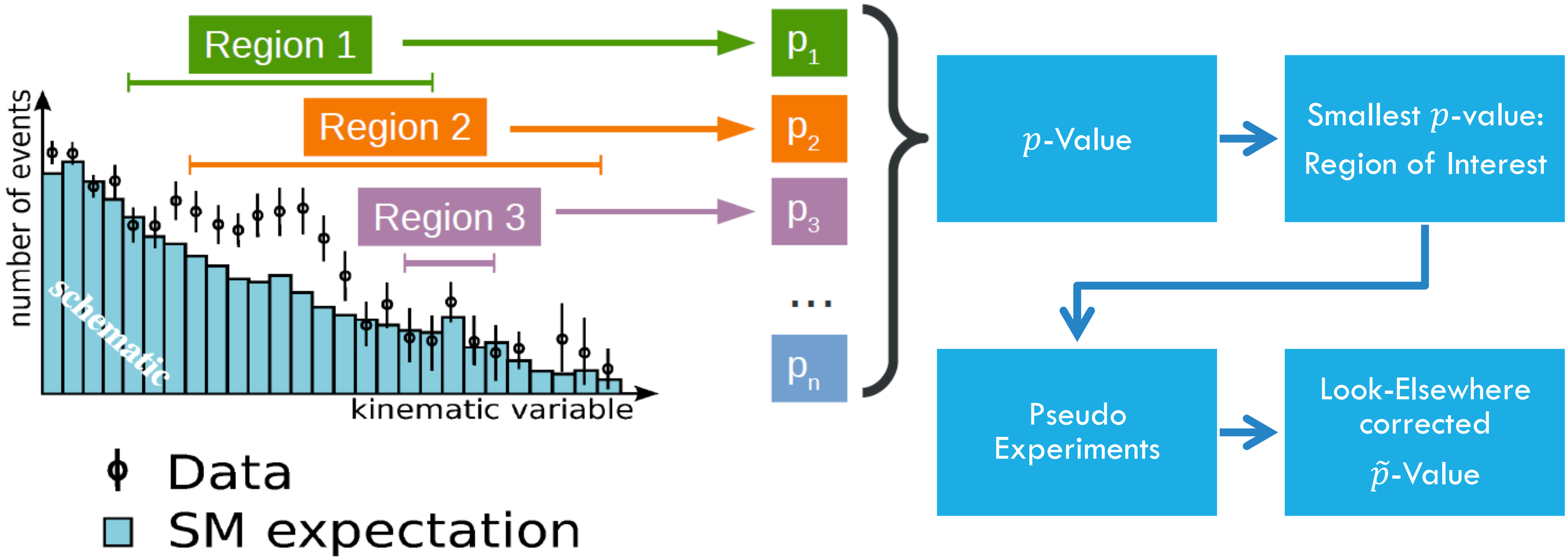
- Problem: The probability to find a significant deviation just by chance rises with the number of considered search regions (look-elsewhere-effect)
- Solution: Dice pseudo distribution according to MC estimate (and systematics), calculate correction using the number of pseudo-experiments with a significance larger than the observed one
- Count pseudo-experiments with a more significant outcome

$$\tilde{p} = \frac{\text{pseudo experiments with } p < p_{data}}{\text{number of pseudo experiments}}$$

- One value per distribution and event class



Scanning Summarized



Why a Fast Search Algorithm?

- Example: 100000 pseudo experiments for ~100 event classes with ~1000 connected bin regions per distribution
- 200 µs per p-value → 560 h computing time (for one kinematic distribution only!)
- Integral calculation slow, but already using the best available implementation

- Remedy: reduce the number regions considered for the p-value integral

Quickscan – The Fast Search Algorithm

- Use on pseudo-experiments only
- Still consider all regions
- Calculate a simpler, less computation intensive significance estimator χ for each region (instead of the full p-value)
- Keep a list of the N most significant candidate regions
- Choose final region of interest from this candidate list, using the p-value integral

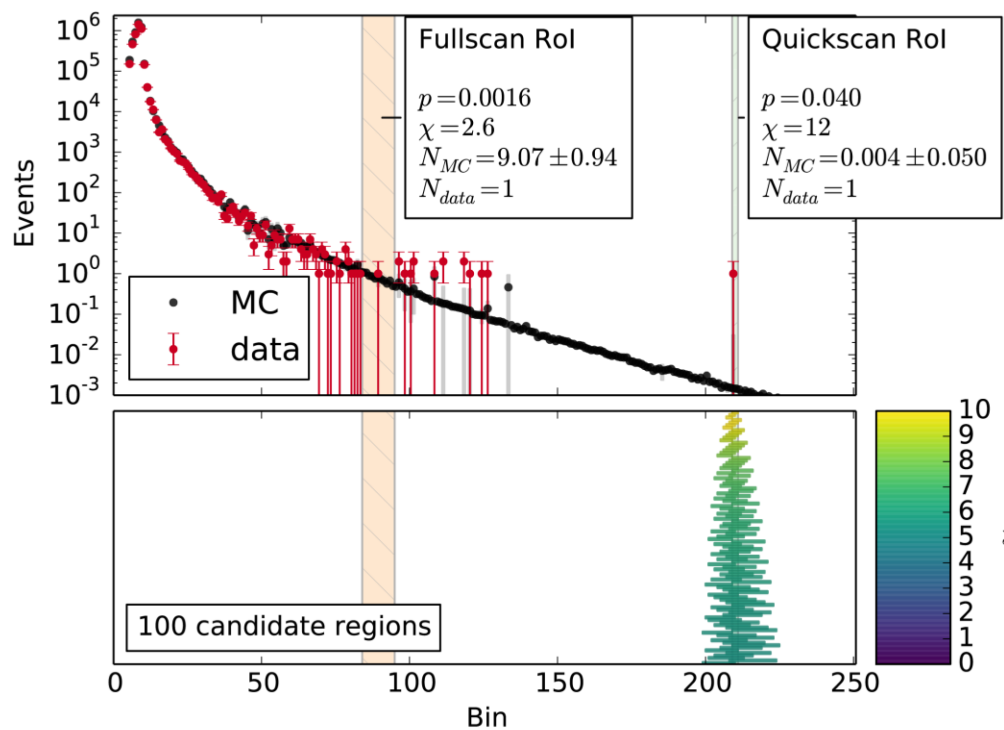
Quickscan – Estimator

$$\chi = \frac{|N_{obs} - N_{MC}|}{\sqrt{\sigma_{MC}^2 + N_{MC}}}$$

- Ratio of observed deviation to expected deviation
- Denominator includes expected statistical deviation $\sigma_{stat} = \sqrt{N_{MC}}$ and systematical deviation σ_{MC}, combined in quadrature
- Only expected to hold in the Poissonian regime of high N_{MC}

Quickscan – Nested Regions

- Problem in high-energy tail of distributions
- Additional significance criterion for nested regions viable
- Given regions A and B:
  - Region A is nested in region B
  - A and B are excesses
  - A and B have the same amount of data
- → A is more significant
- Successfully suppresses unnecessary regions



Situation without nested region handling, Pseudo-experiments

Optimization & Evaluation Criteria

Statistical Accuracy

- Compare $\tilde{p}$ for each event class between full scan and Quickscan
- For each event class: calculate relative $\tilde{p}$ difference

$$\sigma_{rel} = \frac{\tilde{p}_{full} - \tilde{p}_{Quickscan}}{\tilde{p}_{full}}$$

- Situation without Quickscan: → next slide

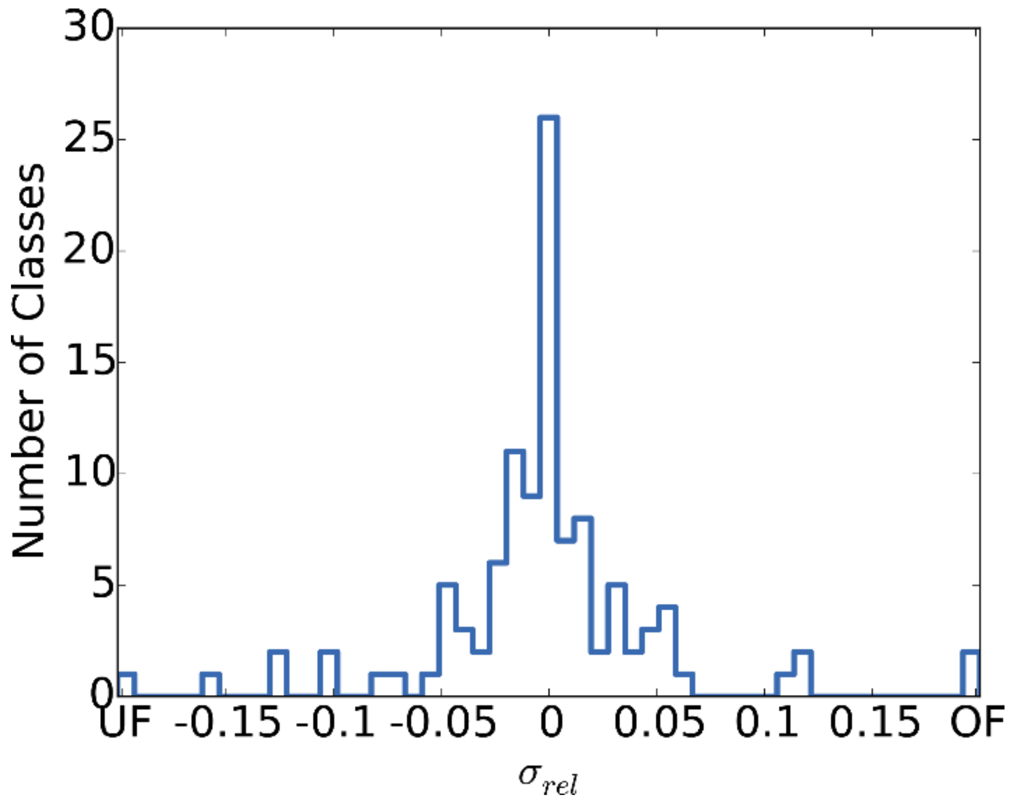
Runtime

- Evaluated as wall-clock time
- Includes input/output
- Compared to runtime of complete scan
- $\text{speedup} = \frac{T_{full}}{T_{Quickscan}}$
- Situation without Quickscan: ~ 1h 30min

Optimization subset: excl. event classes, $\sum |\overrightarrow{p_T}|$ distribution, 1000 pseudo experiments, summarize events with >2 jets in one event class

Statistical Accuracy without Quickscan

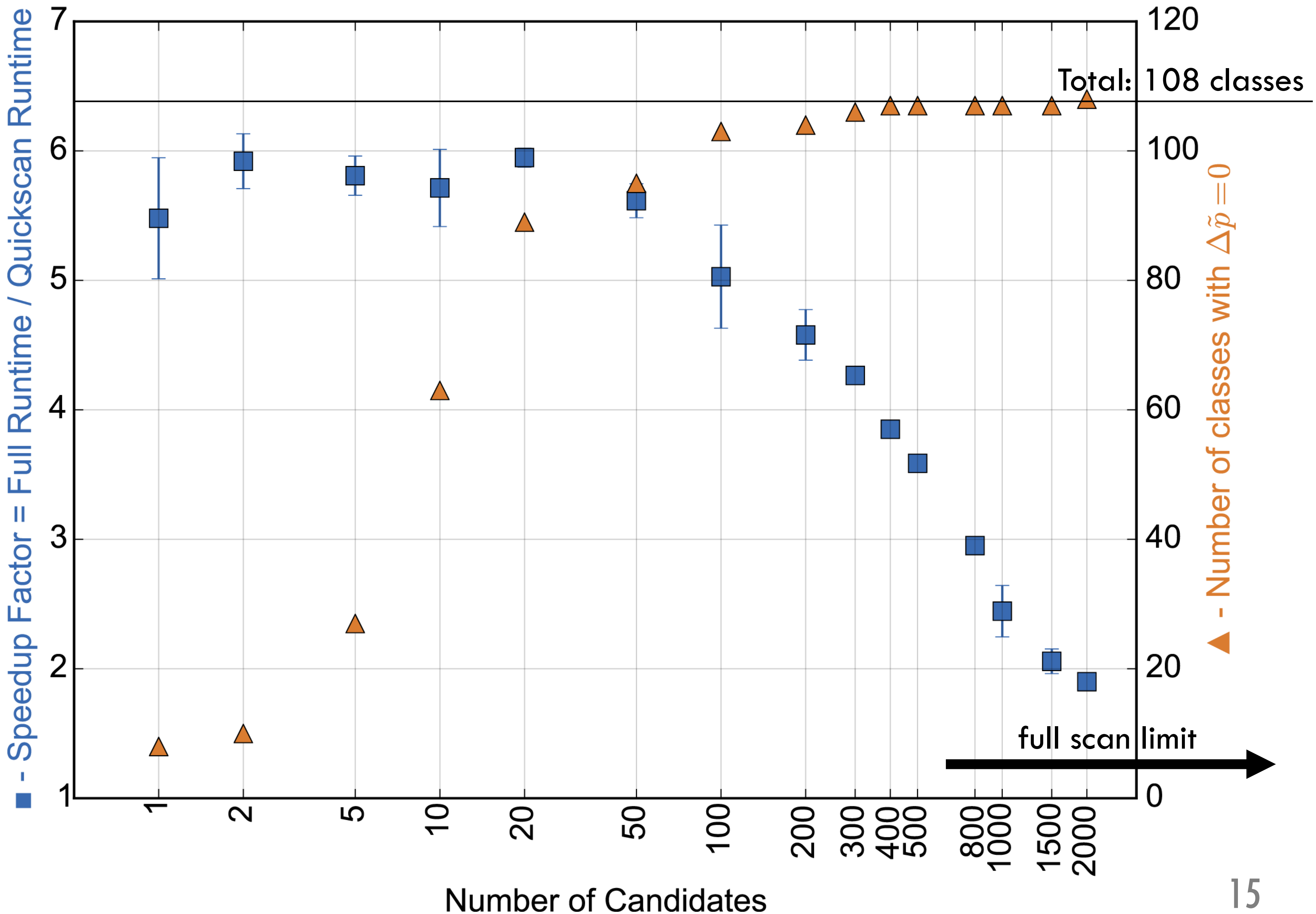
- Deviation of results due to random dicing of pseudo-experiments
- Width of the distribution about 5%
- Optimization of the number of candidate regions N to keep width of distributions stable



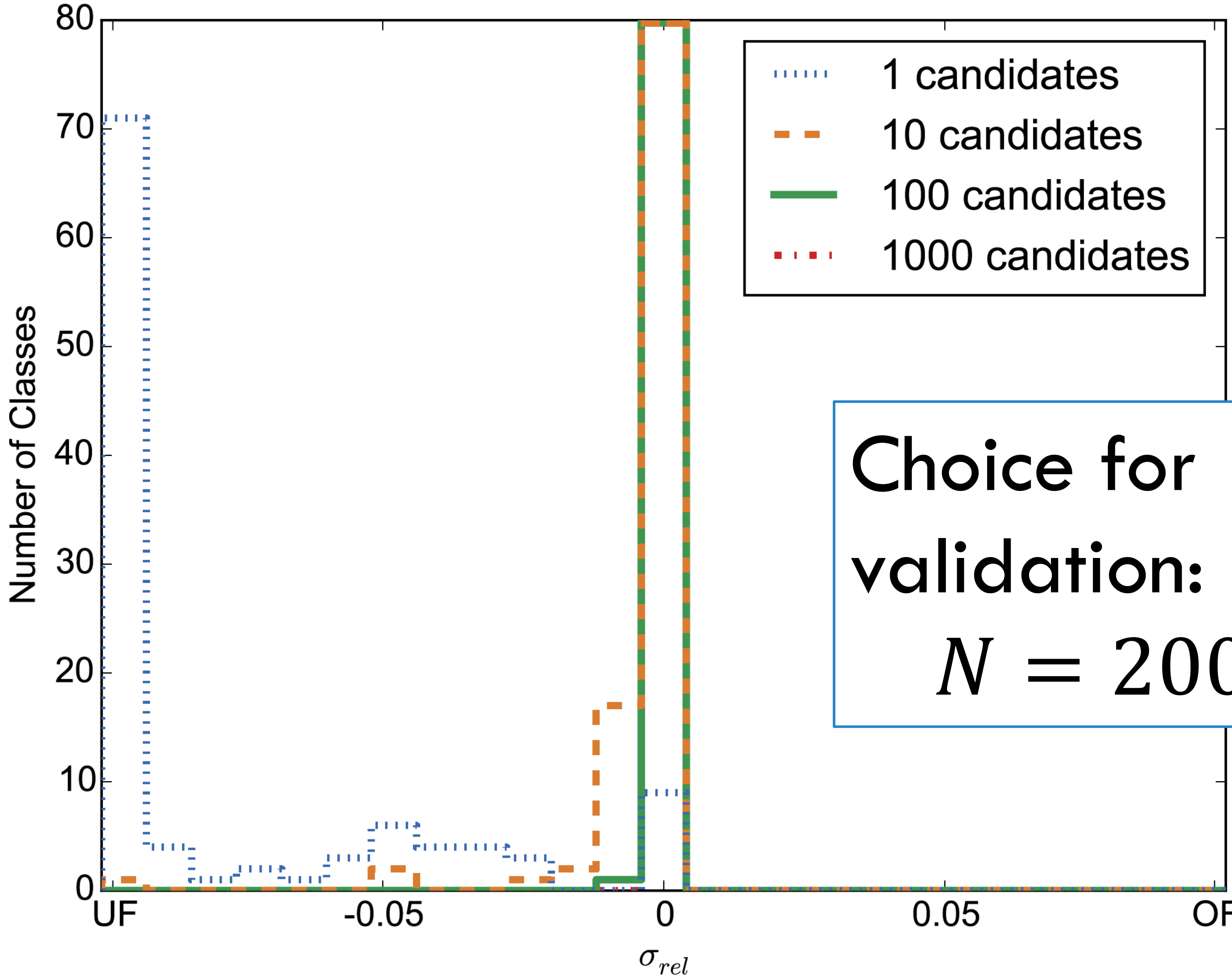
Two full scans compared to each other

$$\sigma_{rel} = \frac{\tilde{p}_{full_2} - \tilde{p}_{full_1}}{\tilde{p}_{full_1}}$$

Results for Different N



Selected Results, Choice of N



Choice for validation:

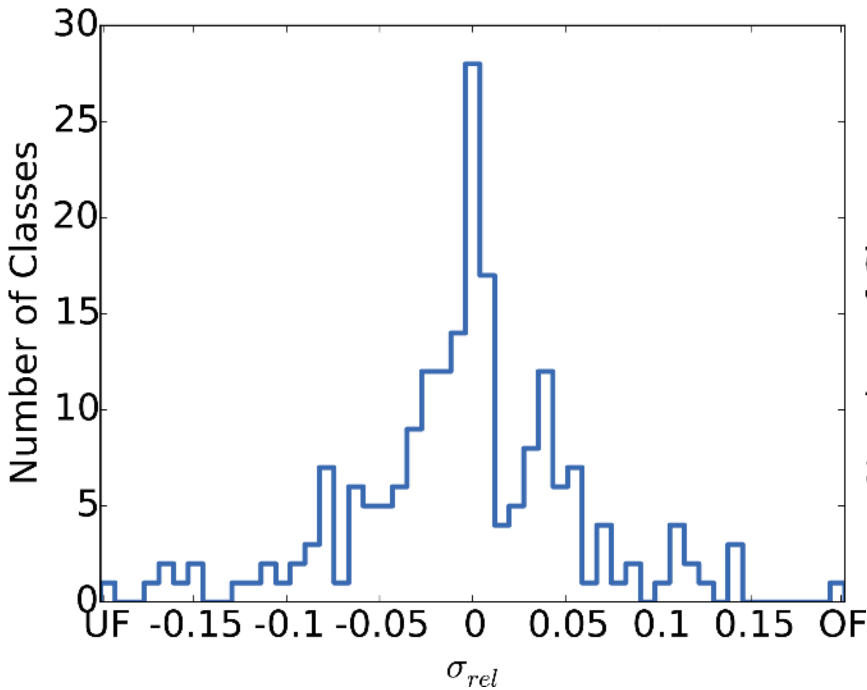
$$N = 200$$

Validation Data Set

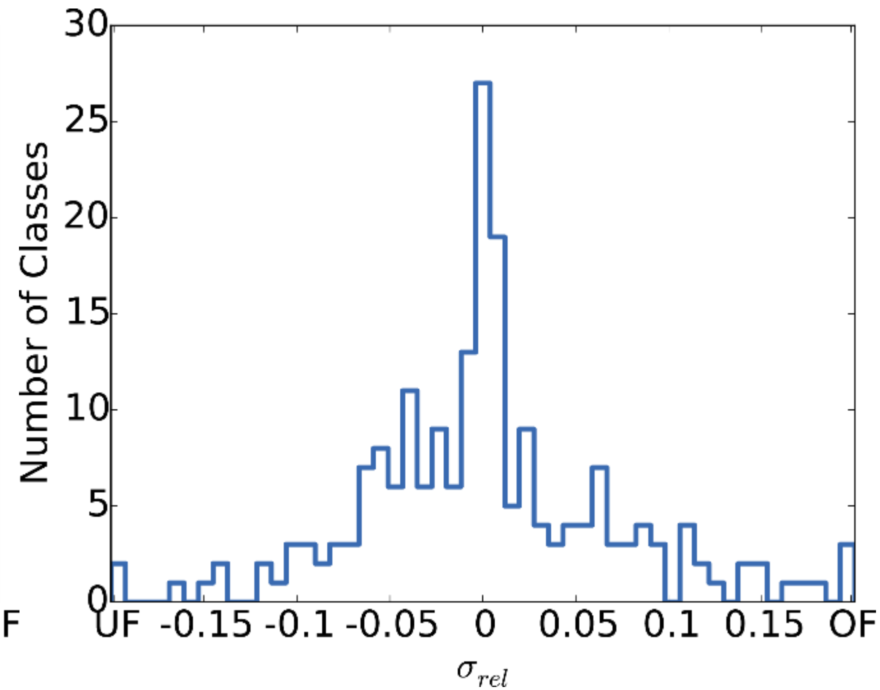
- Run on a different subset:
  - Use M_{inv} distribution (instead of $\sum |\overrightarrow{p_T}|$)
  - Turn off jet threshold, any number of jets generates a new event class (instead of summarizing >2 jets in one class)
  - Generate 100000 pseudo experiments (instead of 1000)

Validation Result

2 full scans compared with each other



A full scan compared with a Quickscan



→ Statistical sensitivity unchanged, ~9 times faster

Conclusion & Outlook

- MUSiC analysis is very complex and computation intensive
- This work has introduced and implemented an additional step called „Quickscan“
- Quickscan allows for a speedup of about 9 times, keeping the statistical sensitivity

- Outlook: Quickscan could benefit from a different estimator, maybe making the nested region handling superfluous
- The MUSiC scan could benefit from a better parallelization, allowing it to run on the CERN computing grid

Backup: Data Details

- Recorded at CMS in 2012
- Center of momentum energy $\sqrt{s} = 8\ \mathrm{TeV}$
- Integrated luminosity $L = 19.7\ \mathrm{fb}^{-1}$

Backup: CMS Barrel

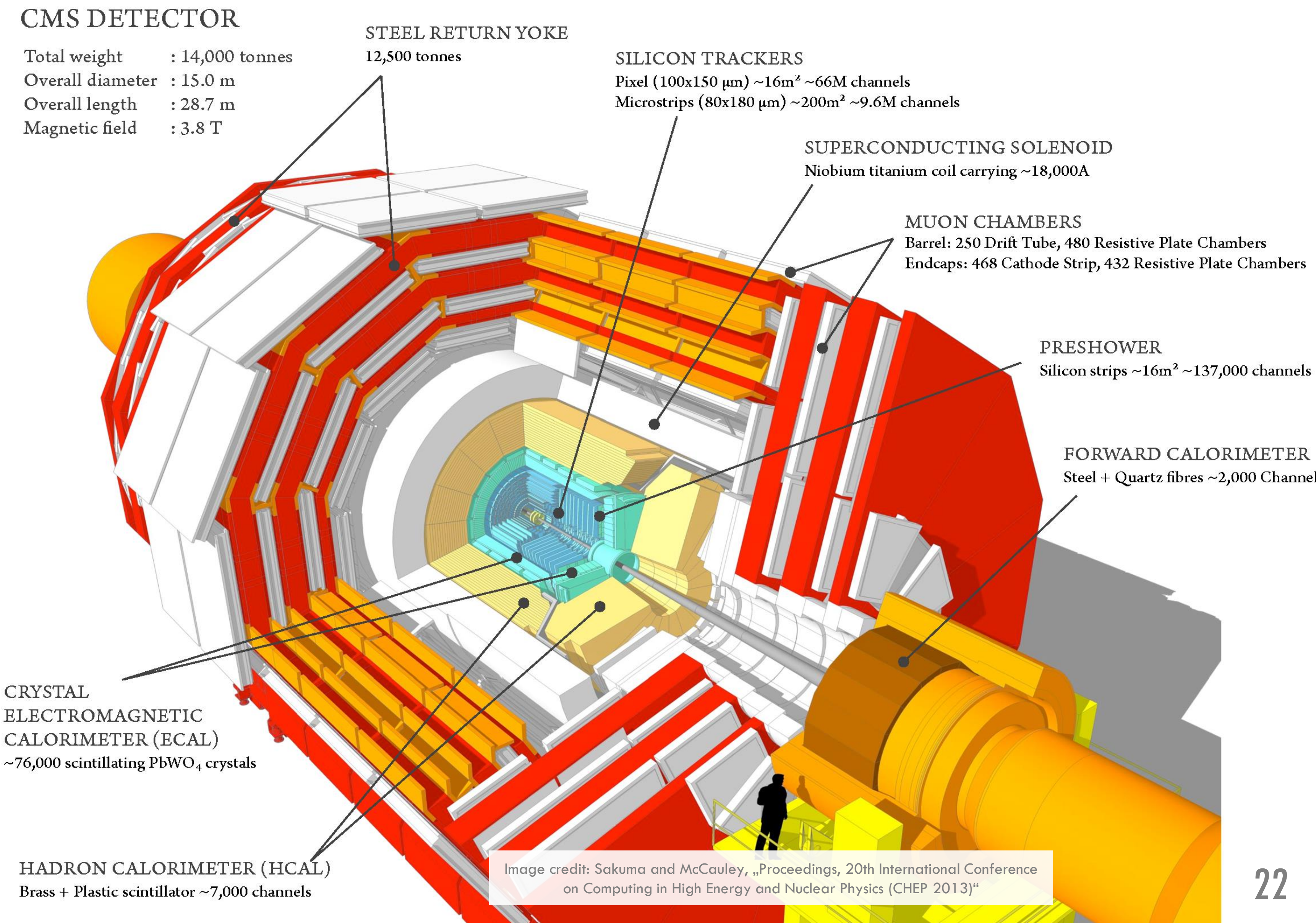
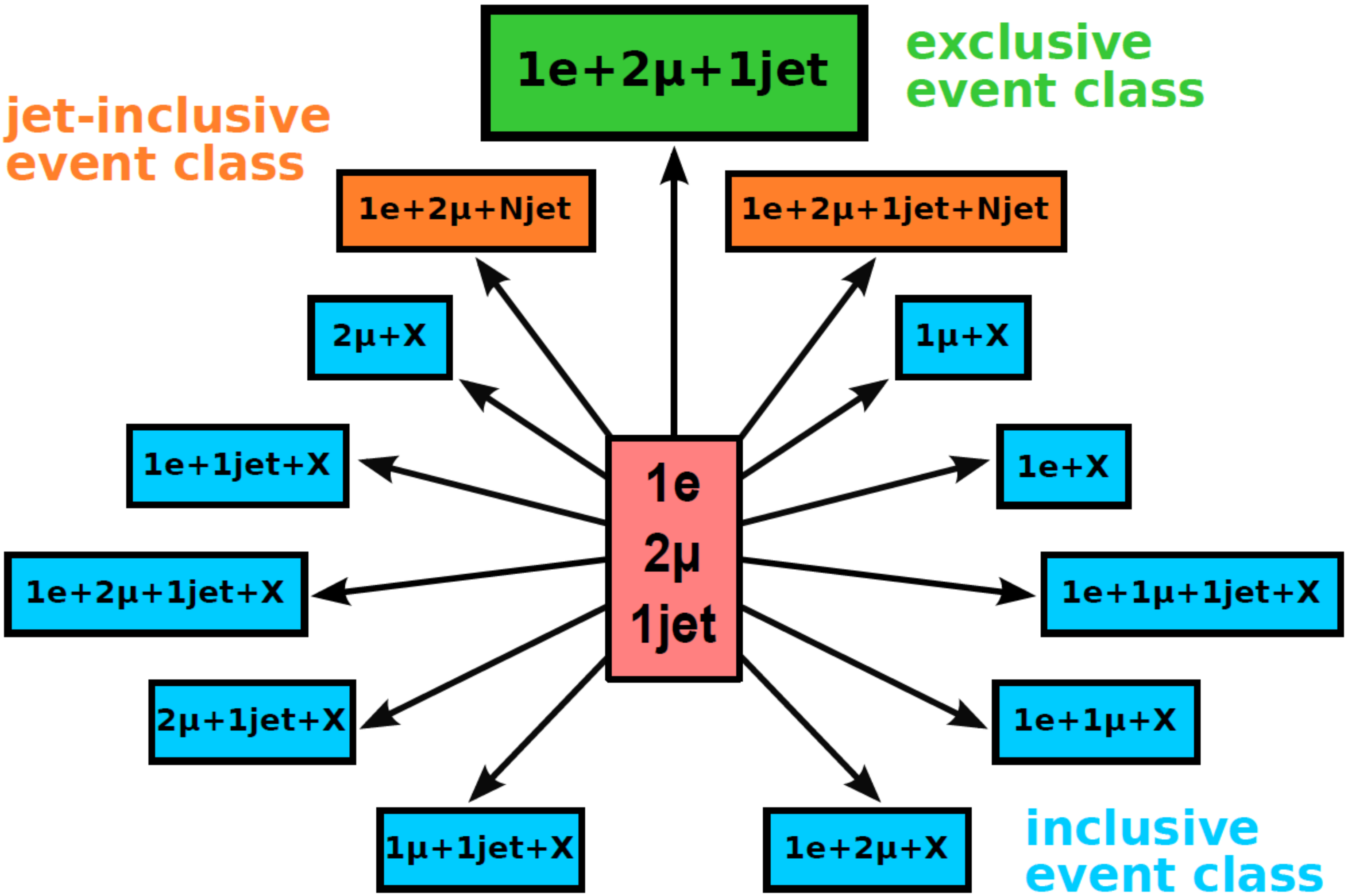


Image credit: Sakuma and McCauley, „Proceedings, 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2013)“

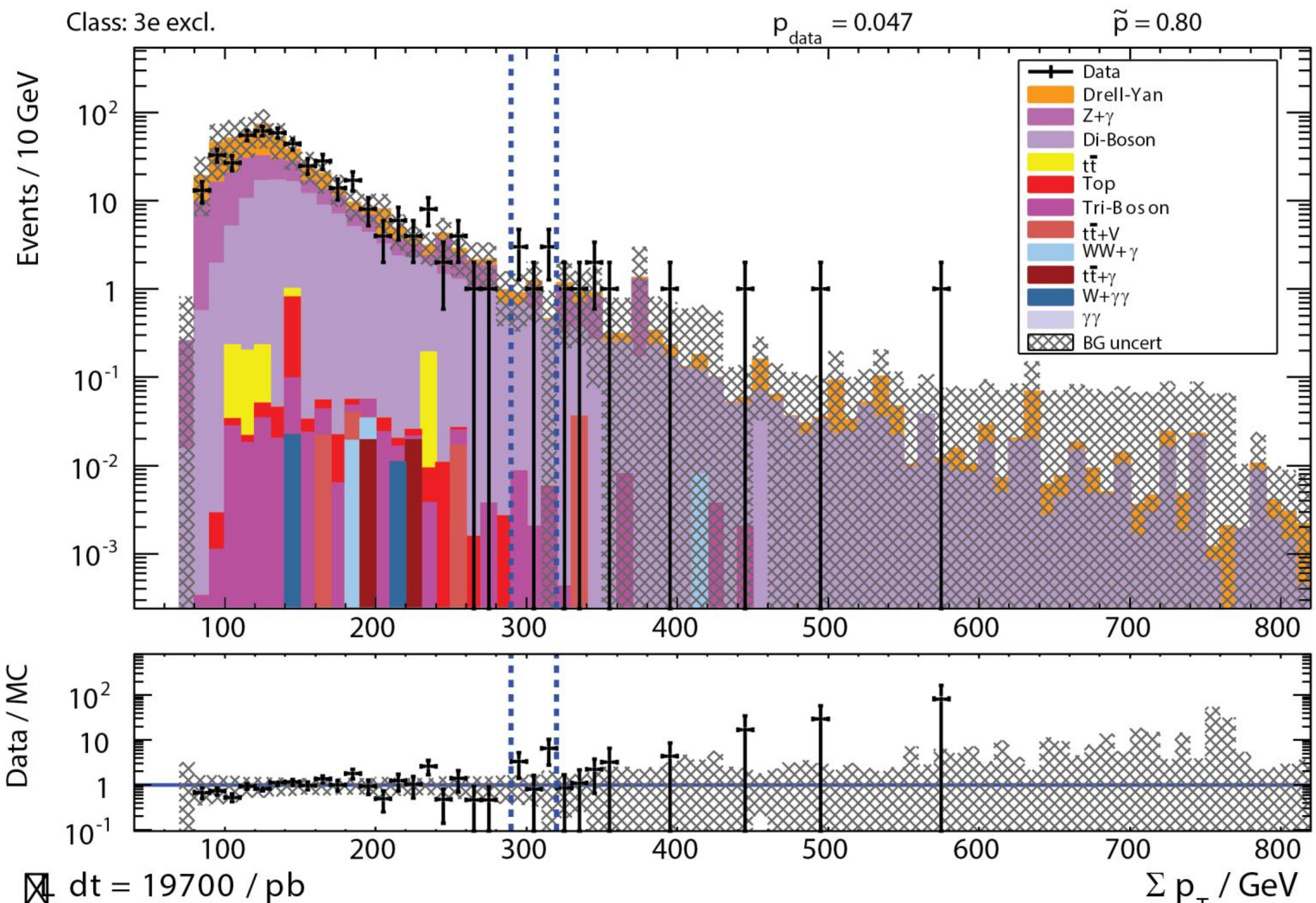
Backup: Object Identification

Object	$\vec{p}_{\mathrm{T}}$/ GeV	$\lvert\eta\rvert$	Identification Summary
μ	>25	<2.1	track quality, isolation, dedicated high-$\vec{p}_{\mathrm{T}}$
e	>25	<2.5	track quality, isolation, dedicated high-E_{T}
γ	>25	<1.442	isolation, veto against e from conversions
jet	>50	<2.4	anti-k_{t} algorithm ($R = 0.5$)
$\not\!\!E_{\mathrm{T}}$	>50		

Backup: Event Class Building

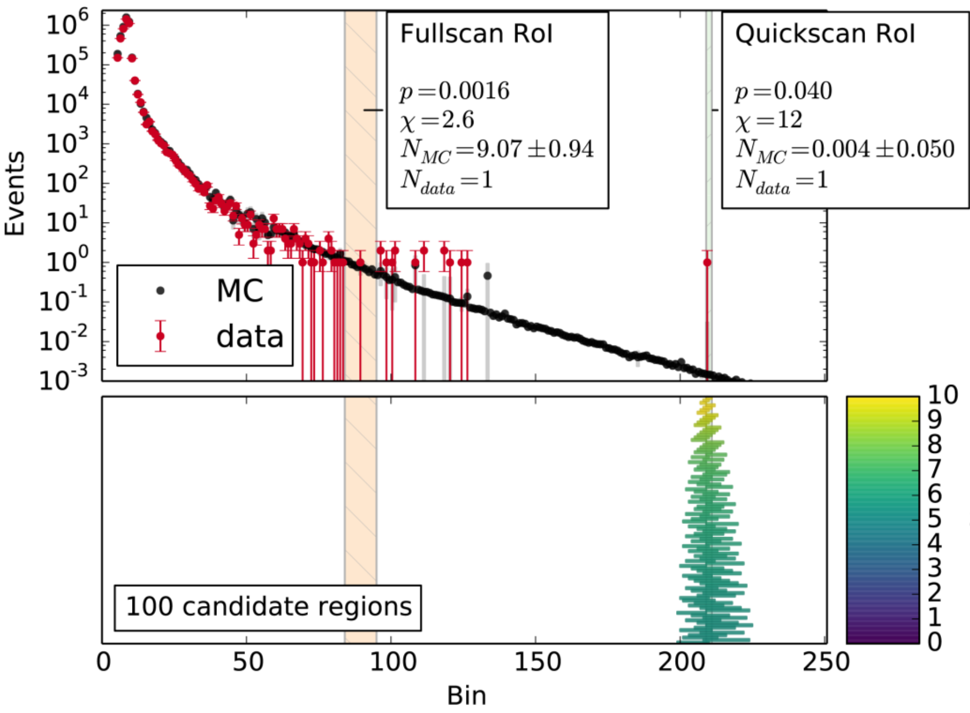


Backup: That One Class



Backup: Nested Region Handling

Without nested region handling



With nested region handling