Search for exotic resonances in $J/\psi\,\phi$ final state and GPU-based techniques for event reconstruction and statistical significance estimation at the CMS experiment

Dottorato in Fisica XXXI ciclo

Dottorando

Adriano Di Florio

Tutore

Dott. Alexis Pompili

Outline

- **I –** ***GPU-based techniques for event reconstruction and statistical significance estimation***
  - **GPU-based techniques for event reconstruction and statistical significance estimation**
  - **Convolutional Neural Networks for Track Seed Filtering at the CMS HLT**
- **II - Search of charmonium-like exotic states into the J/ψϕ mass spectrum**

Search of charmonium-like exotic states into the J/ψφ mass spectrum

Physics Motivation

The Y(4140) is one of the candidates for neutral exotic charmonium-like states, reconstructed in the $J/\psi\ \phi$ channel

CDF Collaboration:

2009: Y(4140) first evidence (confirmed in 2011 with higher statistics + evidence for Y(4274))

studying the $B^+ \to J/\psi\ \phi\ K^+$ decay (M=4143 MeV, Γ=11.7 MeV)

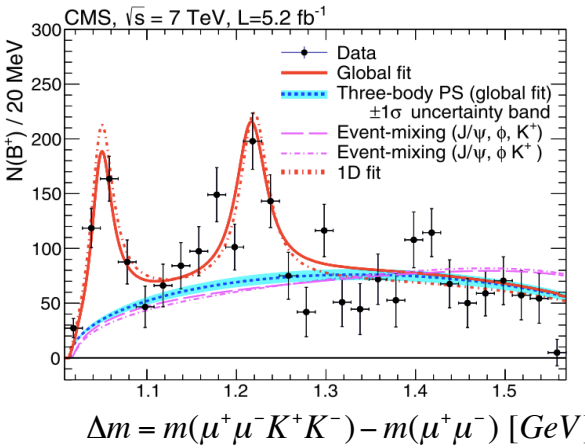
CMS Collaboration:

2014: confirmed peaking structure with a stat. signif. >5σ (hint for a 2nd)

studying the $B^+ \to J/\psi\ \phi\ K^+$ decay

$m = 4148.0 \pm 2.4(\text{stat}) \pm 6.3(\text{syst})MeV \qquad \Gamma = 28^{+15}_{-11}(\text{stat}) \pm 19(\text{syst})MeV$

(not enough statistics for an ***amplitude analysis ;*** issue of ϕK^+ resonances)

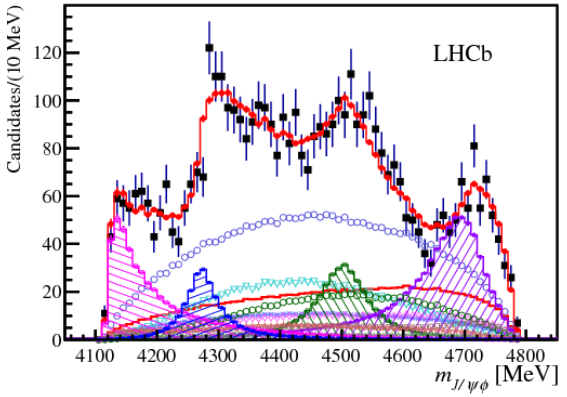


LHCb Collaboration:

2016: First amplitude analysis of the $B^+ \to J/\psi\ \phi\ K^+$ decay

Observing four resonant structures each with a significance >5σ

State	J^{PC}	signif.	Mass	Width	fit frac.
$X(4140)$	1^{++}	8.4σ	$4165 \pm 4.5^{+4.6}_{-2.8}$	$83 \pm 21^{+21}_{-14}$	$13.0 \pm 3.2^{+4.8}_{-2.0}$
$X(4274)$	1^{++}	6.0σ	$4273.3 \pm 8.3^{+17.2}_{-3.6}$	$56 \pm 11^{+8}_{-11}$	$7.1 \pm 2.5^{+3.5}_{-2.4}$
$X(4500)$	0^{++}	6.1σ	$4506 \pm 11^{+12}_{-15}$	$92 \pm 21^{+21}_{-20}$	$6.6 \pm 2.4^{+2.5}_{-2.3}$
$X(4700)$	0^{++}	6.1σ	$4704 \pm 10^{+14}_{-24}$	$120 \pm 31^{+42}_{-33}$	$12 \pm 5^{+9}_{-5}$



Physics Motivation

DØ experiment (2015) :

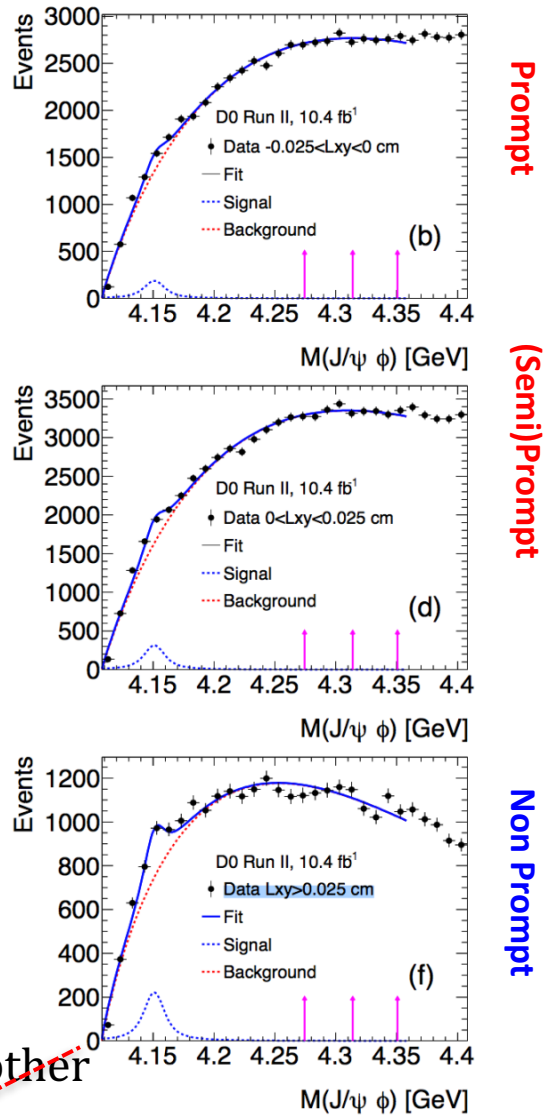
Y(4140) was inclusively searched in the mass spectrum $J/\psi\ \phi$

Evidence/Observation of **prompt** (4.7σ) and **non-prompt** (5.3σ)

in $p\bar{p}$ collisions

D0 experiment observed the inclusive production of *Y(4140)* : needs confirmation @ LHC !

If confirmed this would be the second exotic state [after X(3872)**] observed with two different production mechanisms (promptly produced & as resonances in 3-body B decays)**



We study (with CMS Run-II data) inclusively the spectrum separating *prompt* & non-prompt regions looking for not only the *Y(4140)* but also its 3 partners found by LHCb

In other words we study $J/\psi\ \phi$ spectrum

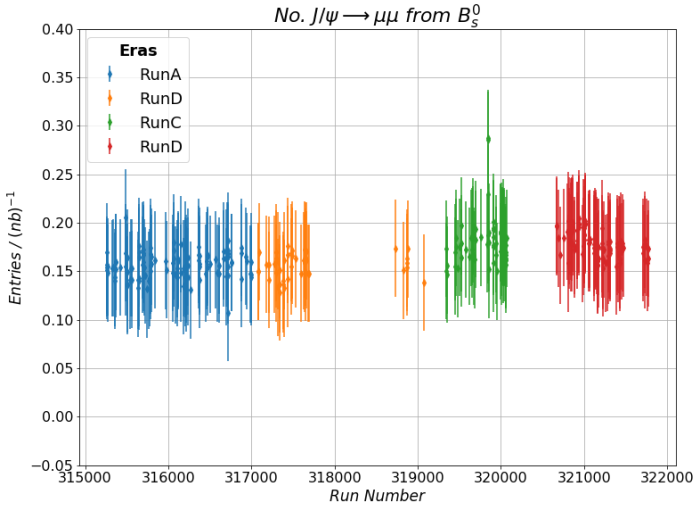
looking for $pp \to \boldsymbol{Y} + \cancel{other} \to J/\psi\ \phi + \cancel{\text{other}}$

& using as reference channel $pp \to \boldsymbol{B_s^0} + \cancel{other} \to J/\psi\ \phi + \cancel{\text{other}}$

$\sigma(L_{xy})/L_{xy} < 3.0$

2018 - $J/\psi\phi$ mass spectrum

- After having explored **Run I** data and both **2016 – 2017 Run II** data, for 2018 a novel dedicated HLT has been developed.

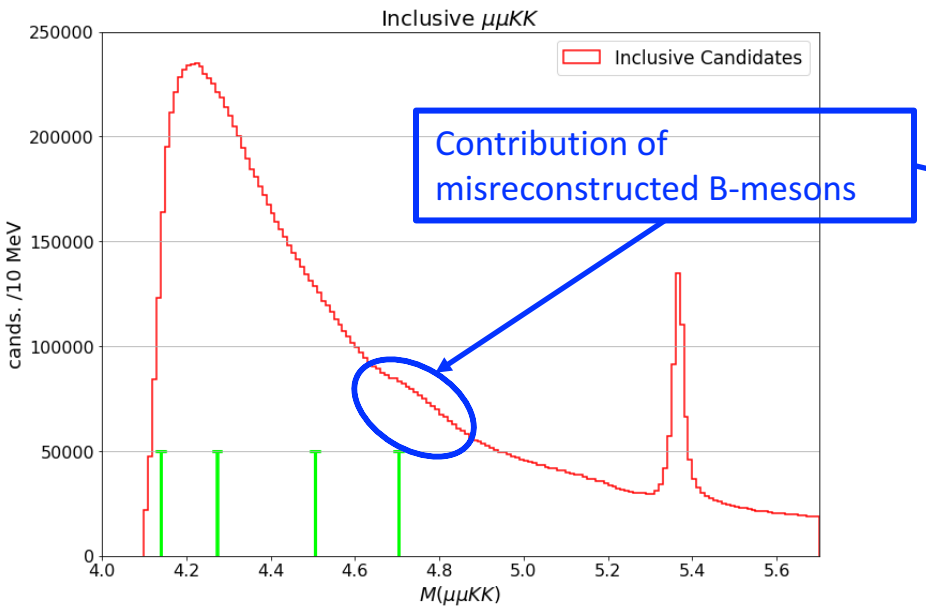


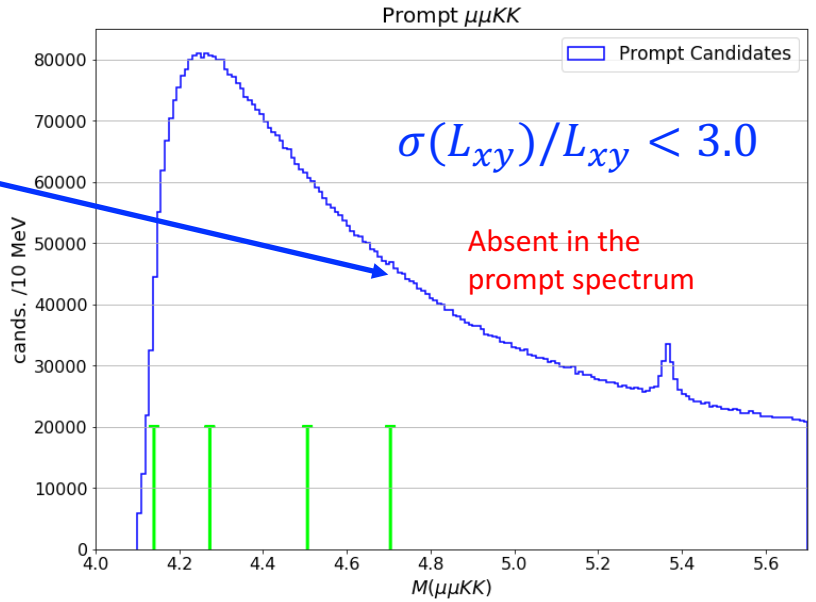
In order to keep as lowest as possible the p_T thresholds the HLT requests both a J/ψ and a ϕ in the event. The **stability of the yield of J/ψ , produced from a B_s^0 decay**, has been monitored along all the 2018 data taking with the aim to:

- to check eventual **biases in the evts selection**
- test the **behaviour of the new HLT**

- All 2018 data, available and certified until today and corresponding to $\mathcal{L}\sim 52\ fb^{-1}$ have been analysed.

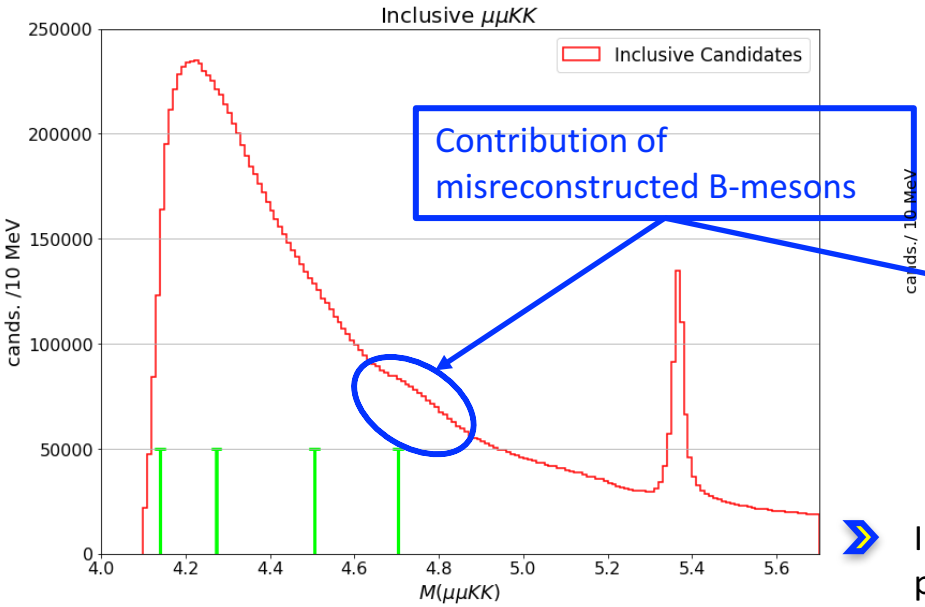
- After multiple studies on the event selection, the mass spectrum (**inclusive** and **prompt**) has been reconstructed.

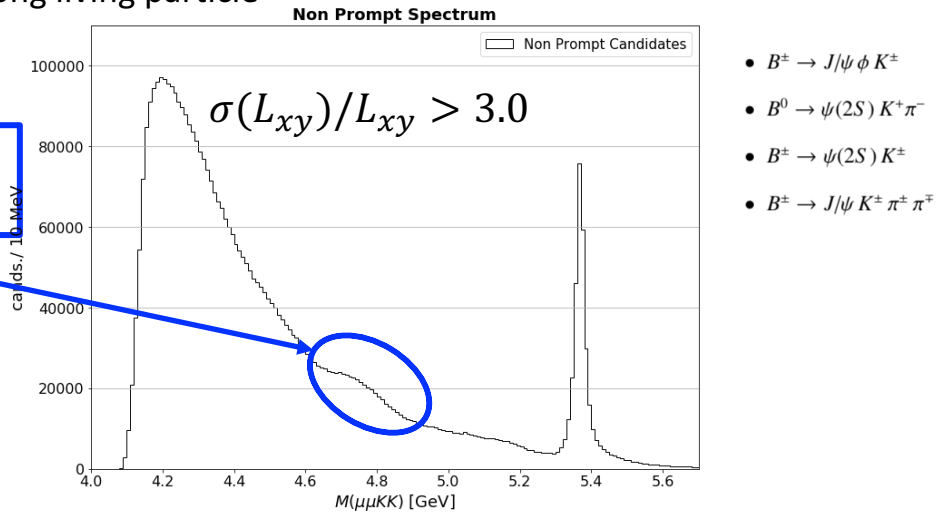




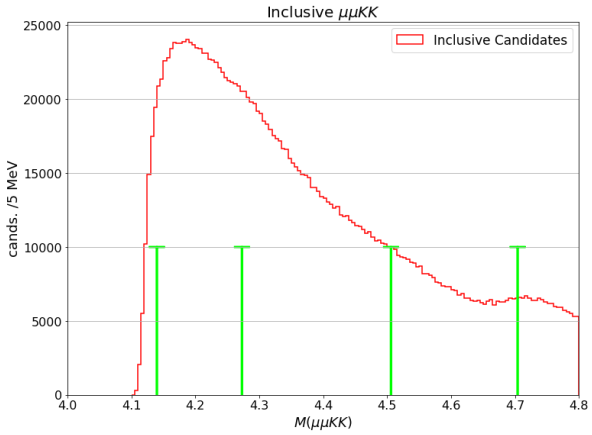
2018 - $J/\psi\phi$ mass spectrum - I

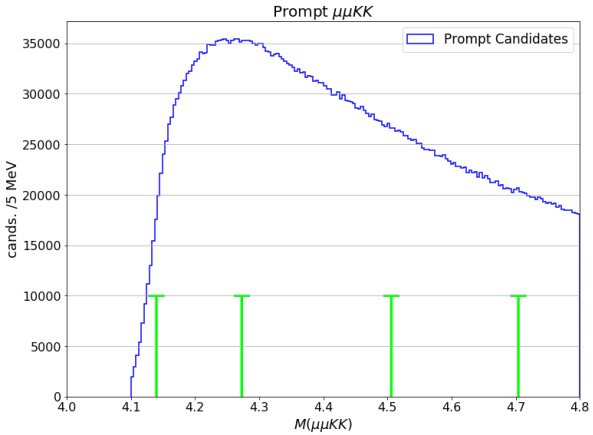
- A bump structure is clearly visible in the mass spectrum in the region [4.6-4.8] GeV. This stucture is expected to arise from **missreconstructed B-mesons**. As a matter of fact, while completely absent in the prompt spectrum it is still present in the **non-prompt** region indicating that it's a long living particle





- In order to confirm this hypothesis a generic $b\bar{b}$ **MC** has been produced and it **confirms the nature of such a bump**.

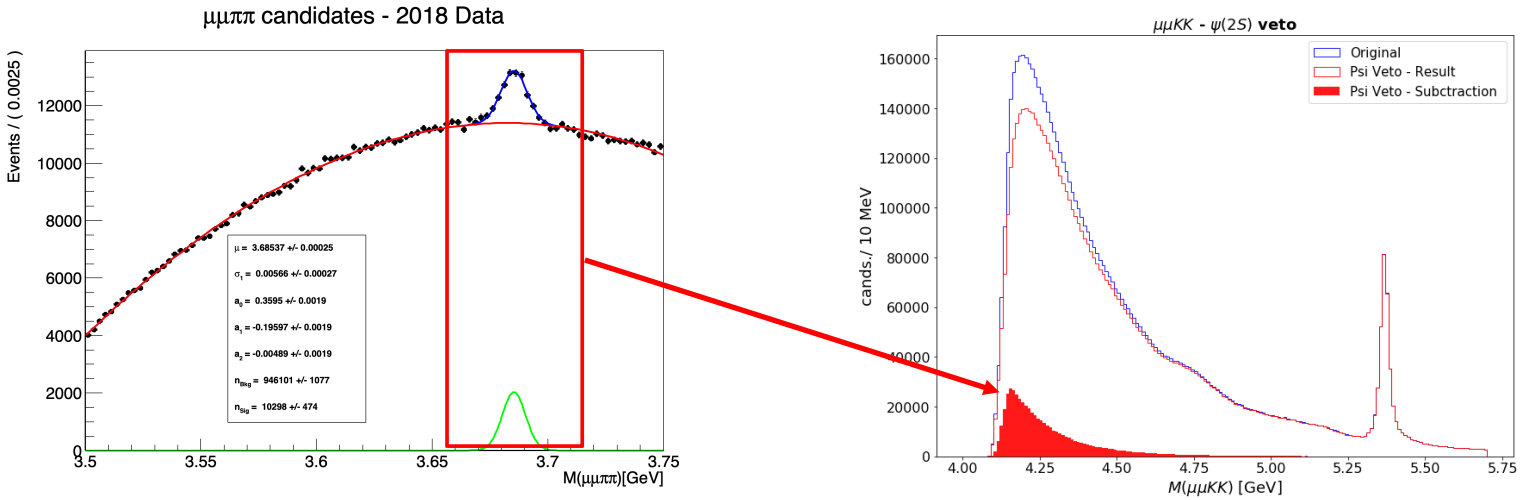




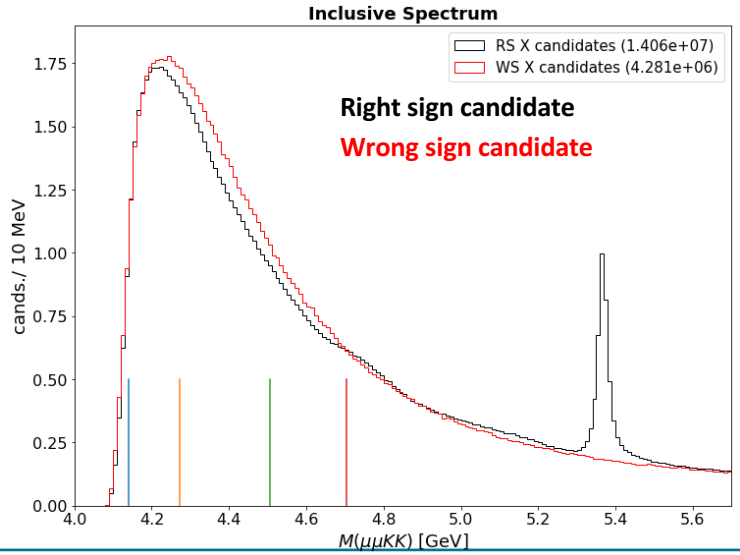
- By focussing on the region where the **Y signals** should appear no peaking behaviour is evident.

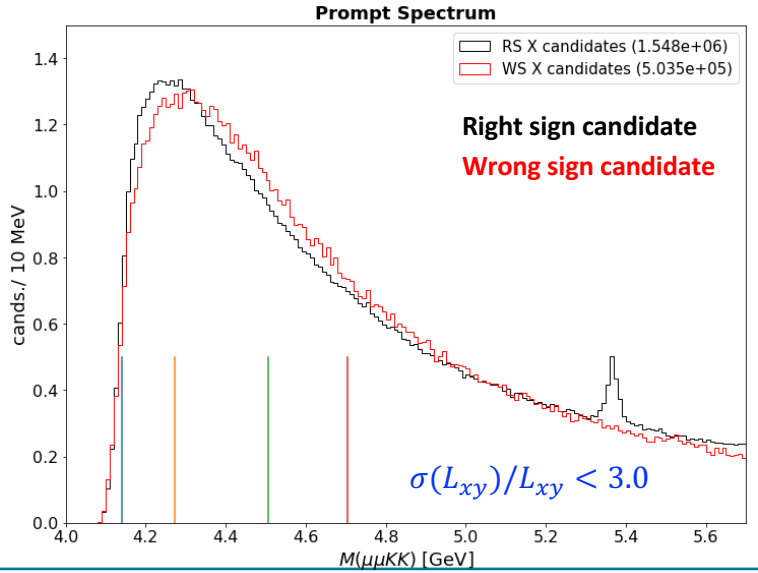
2018 – background studies

Multiple studies on the background sources has been conduced. Here we present two significant cases. Frist, since CMS does **not provide hadron identification**, each track is assigned a specific mass hypothesis. In our case the mass assigned to the two tracks is the ***kaon*** mass. With the **π** refit the $\psi(2S)$.



Selection of ϕ candidates with $q \neq 0$ **(WS)** and comparison with the signal ϕ (RS). Rather compatible, background estimation.



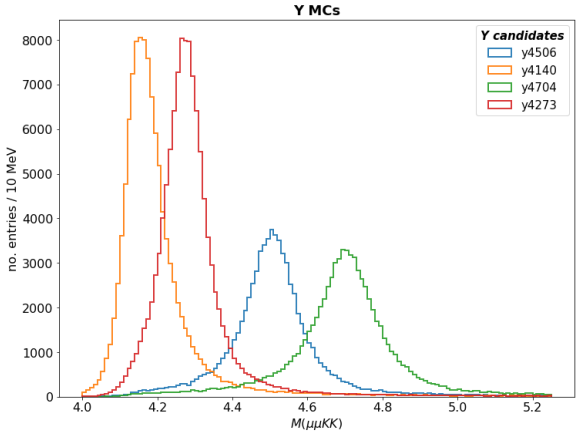


- **Since no evidence** of signal has been found in the selected data. The next step is then to calculate an upper limit on the production rate of **Υ** states with respect to $\mathbf{B^0_s}$ **,** our standard candle.

$$R = \frac{\sigma(pp \to Y + X) \times \mathcal{B}(Y \to J/\psi\phi)}{\sigma(pp \to B_s^0 + X) \times \mathcal{B}(B_s^0 \to J/\psi\phi)} = \frac{N_Y}{\epsilon_Y} \times \frac{\epsilon_{B_s^0}}{N_{B_s^0}}$$

Estimated from data

- A set of MC has been produced with the characteristics of the Υ states seen by *LHCb* in order to estimate the relative efficiency of each Υ signal with respect to B

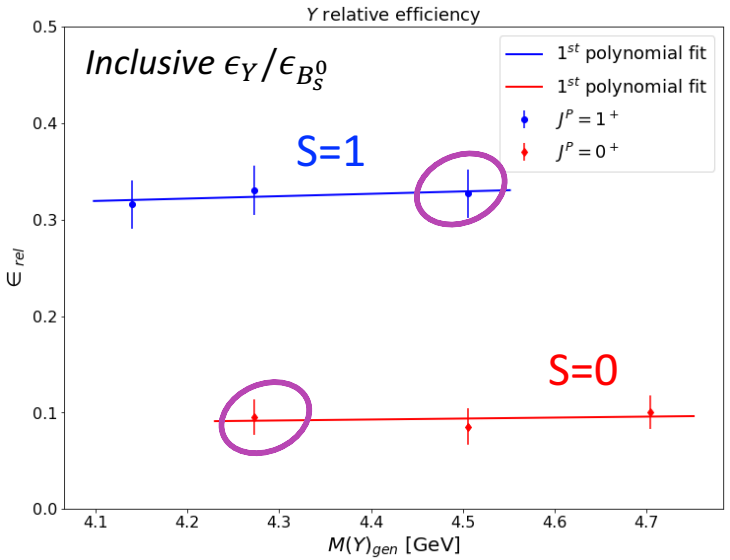


MCs			
Signal	J^P	M_{gen} (MeV)	N_{evt}
B_s^0	0^-	5366.79	346611
$\Upsilon(4140)$	1^+	4146.5	3395804
$\Upsilon(4273)$	1^+	4273.3	3742438
$\Upsilon(4506)$	1^+	4506.0	5131737
$\Upsilon(4273)$	0^+	4273.3	1346553
$\Upsilon(4506)$	0^+	4506.0	1979588
$\Upsilon(4704)$	0^+	4704.0	6236088

N.B. the $J^P=0^+$ states have been modelled in the generator with a χ_{c0} charmonium state, the $J^P=1^+$ with a χ_{c1} charmonium state.

$$\epsilon_{tot} = \epsilon_{gen} \cdot \epsilon_{reco} \cdot \epsilon_{trigger}$$

- From the MC we saw that the spin=1 states has an **harder p_T** spectrum and so the selection is more efficient at trigger level. Two extra MC has been produced with swapped spins to have a third point to determine **two trends** for the **two spins.**

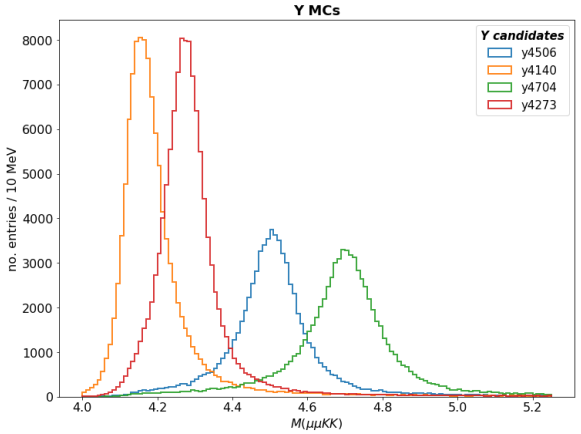


- **Since no evidence** of signal has been found in the selected data. The next step is then to calculate an upper limit on the production rate of **Υ** states with respect to $\mathbf{B^0_s}$ **,** our standard candle.

$$R = \frac{\sigma(pp \to Y + X) \times \mathcal{B}(Y \to J/\psi\phi)}{\sigma(pp \to B_s^0 + X) \times \mathcal{B}(B_s^0 \to J/\psi\phi)} = \frac{N_Y}{\epsilon_Y} \times \frac{\epsilon_{B_s^0}}{N_{B_s^0}}$$

Estimated from data

- A set of MC has been produced with the characteristics of the Υ states seen by *LHCb* in order to estimate the relative efficiency of each Υ signal with respect to B

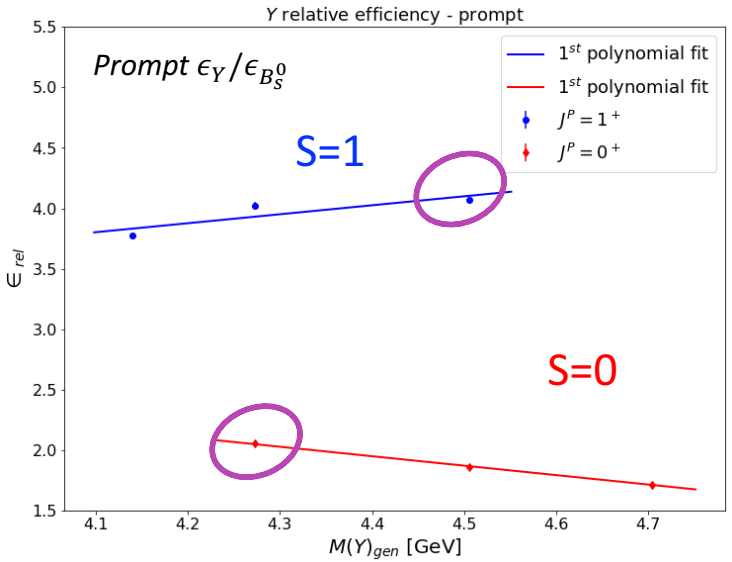


MCs			
Signal	J^P	M_{gen} (MeV)	N_{evt}
B_s^0	0^-	5366.79	346611
$\Upsilon(4140)$	1^+	4146.5	3395804
$\Upsilon(4273)$	1^+	4273.3	3742438
$\Upsilon(4506)$	1^+	4506.0	5131737
$\Upsilon(4273)$	0^+	4273.3	1346553
$\Upsilon(4506)$	0^+	4506.0	1979588
$\Upsilon(4704)$	0^+	4704.0	6236088

N.B. the $J^P=0^+$ states have been modelled in the generator with a χ_{c0} charmonium state, the $J^P=1^+$ with a χ_{c1} charmonium state.

$$\epsilon_{tot} = \epsilon_{gen} \cdot \epsilon_{reco} \cdot \epsilon_{trigger}$$

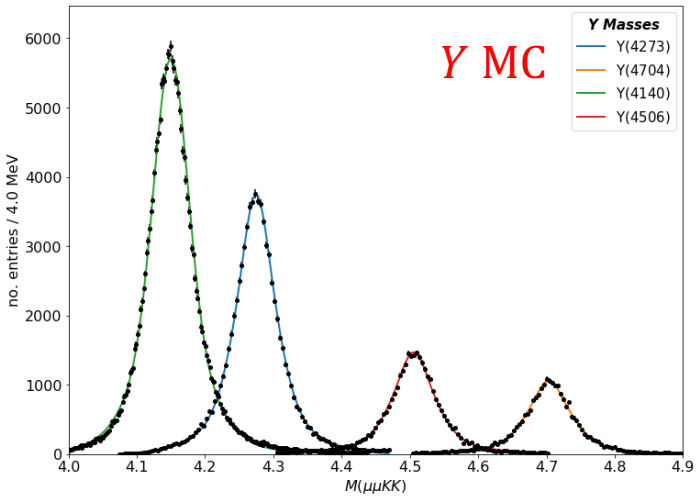
- From the MC we saw that the spin=1 states has an **harder p_T** spectrum and so the selection is more efficient at trigger level. Two extra MC has been produced with swapped spins to have a third point to determine **two trends** for the **two spins.**

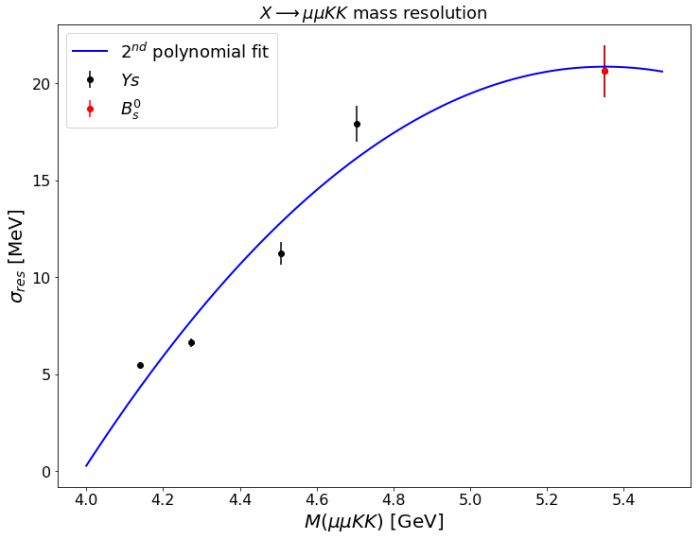


2018 – Υ & B^0_s MC – Signal modelling

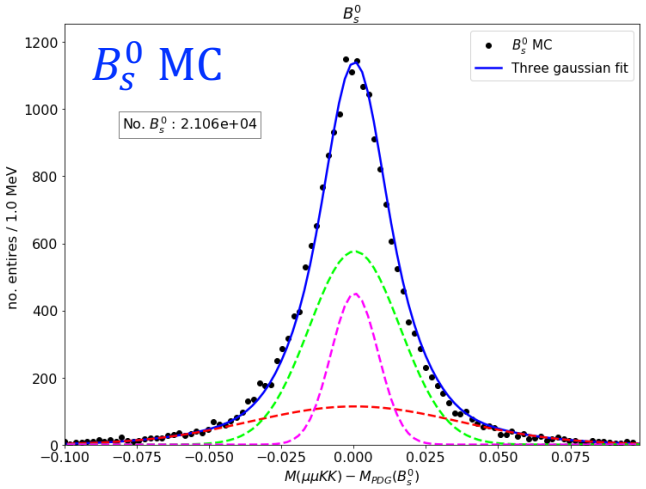
A second step for the calculation of the limit is the definition of the **signal model**. This is extracted from the MCs and is selected to be the convolution of a *B.W.* function *(with the widths fixed to the one estimated by LHCb)* and a Gaussian p.d.f. for the experimental resolution.

LHCb signals





Also the



N.B. the B^0_s natural width is negligible compared to experimental resolution. It has been modelled as a triple Gaussian.

$$\sigma_{eff} = [f_1\sigma_1^2 + f_2\sigma_2^2 + (1.0 - f_1 - f_2)\sigma_3^2]$$

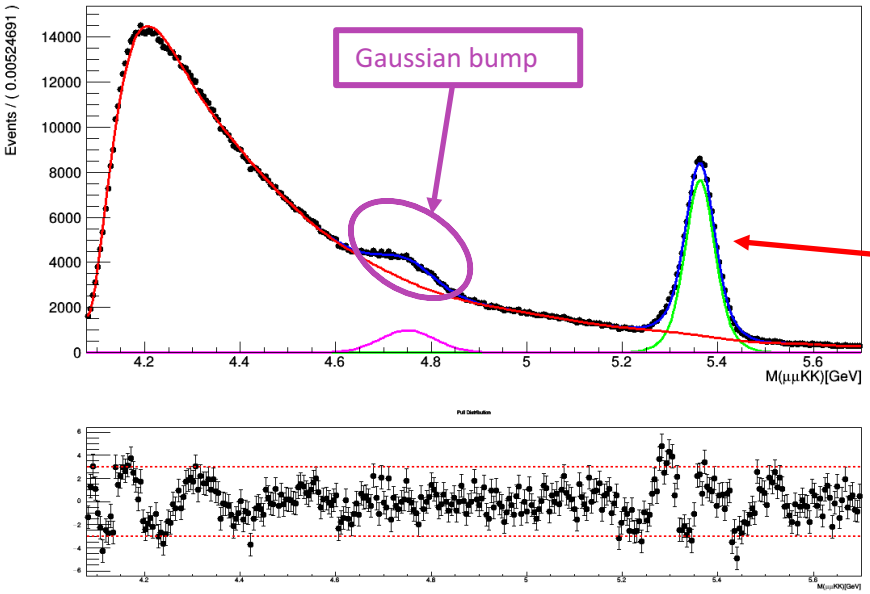
The resolution of the four **Υ** states together with the B^0_s signal are used to determine the trend of the mass resoltuon w.r.t. to the invariant mass of the μμkk candidate.

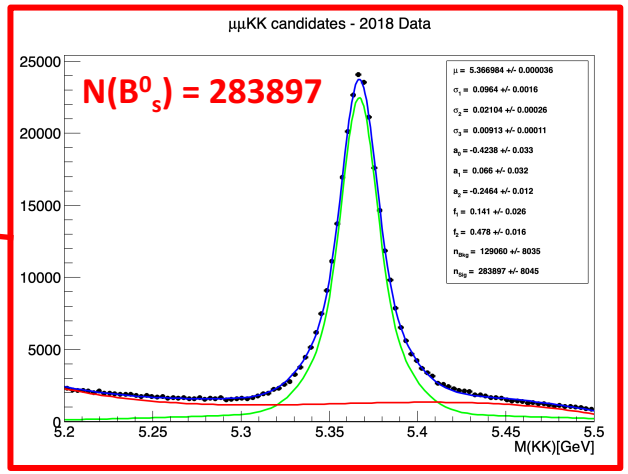
This mass-dependent resolution trend is used as **running experimental resolution** for the **Gaussian** component of the fit.

2018 – Spectrum Fit & B^0_s Yield

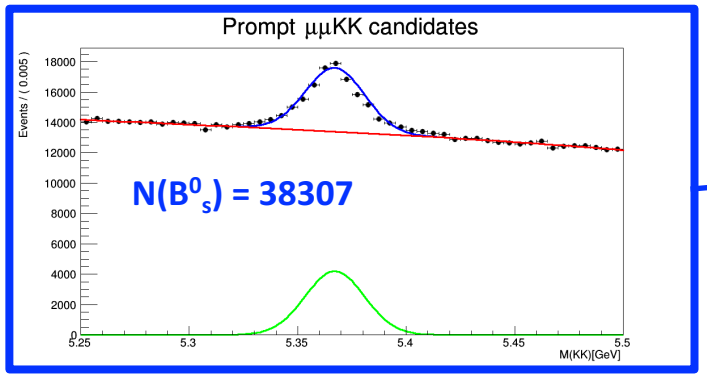
- The last ingredient for evaluating the ratio R is the estimation of the B_s^0 signal yield. To estimate that, the whole mass spectrum has been fitted in the two regions using a Chebychev polynomial p.d.f. to interpolate the background.

Inclusive

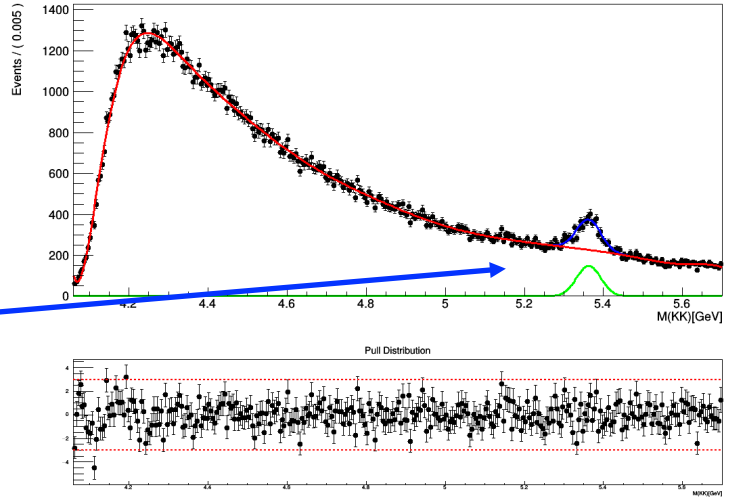




- For both **prompt** and **inclusive** region the signal has been modelled with a **double** gaussian fit.



Prompt



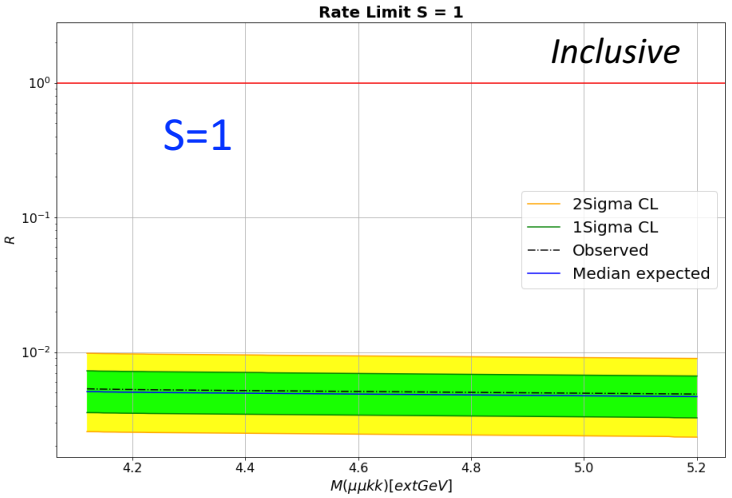
2018 – Limit calculation

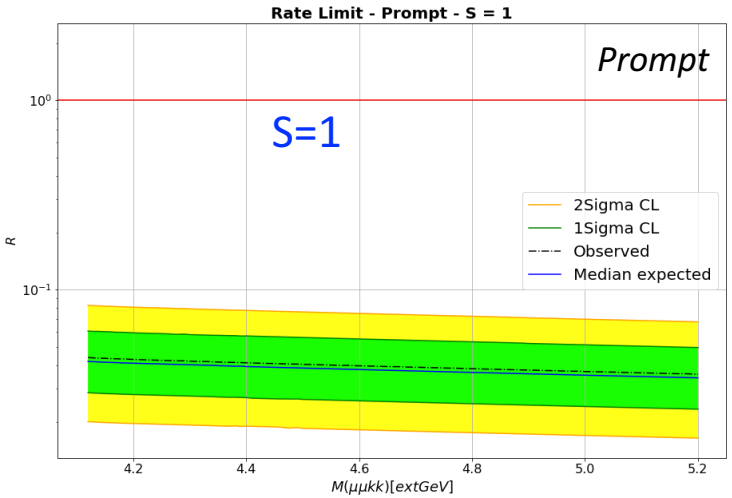
- Once the spectrum has been fitted and the background has been parametrized, it has been possible to estimate the upper limit on the ratio, corrected with the relative efficiencies estimated from the MC

$$R = \frac{\sigma(pp \to Y + X) \times \mathcal{B}(Y \to J/\psi\phi)}{\sigma(pp \to B_s^0 + X) \times \mathcal{B}(B_s^0 \to J/\psi\phi)} = \frac{N_Y}{\epsilon_Y} \times \frac{\epsilon_{B_s^0}}{N_{B_s^0}}$$

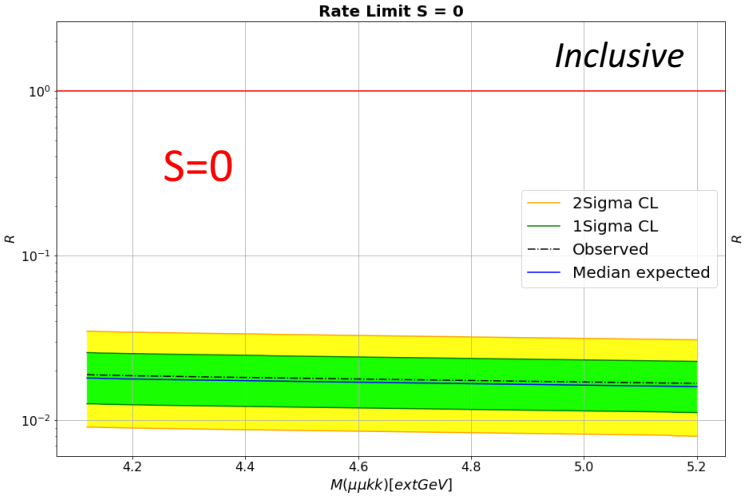
The calculation has been peformed with the HiggsCombine tool developed within CMS software framework

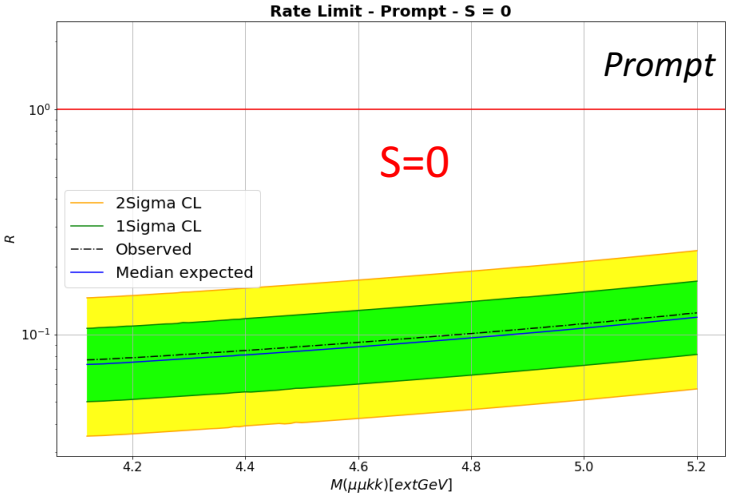
- The limit is split in four possible configurations. Two arising from the **prompt** and **inclusive** search, and two arising from the **two polarization hypothesis** that affects the relative efficiency,





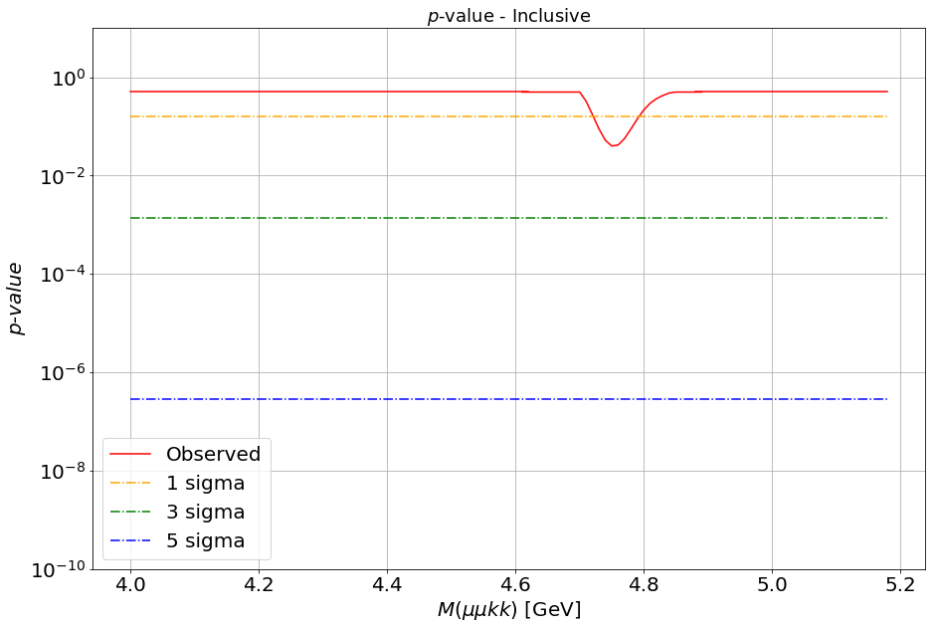
- In the case of **prompt** production, due to the lower B_s^0 yield, the $R \sim O(0.1)$ while for the **inclusive** search, it reaches $R \sim O(0.01)$

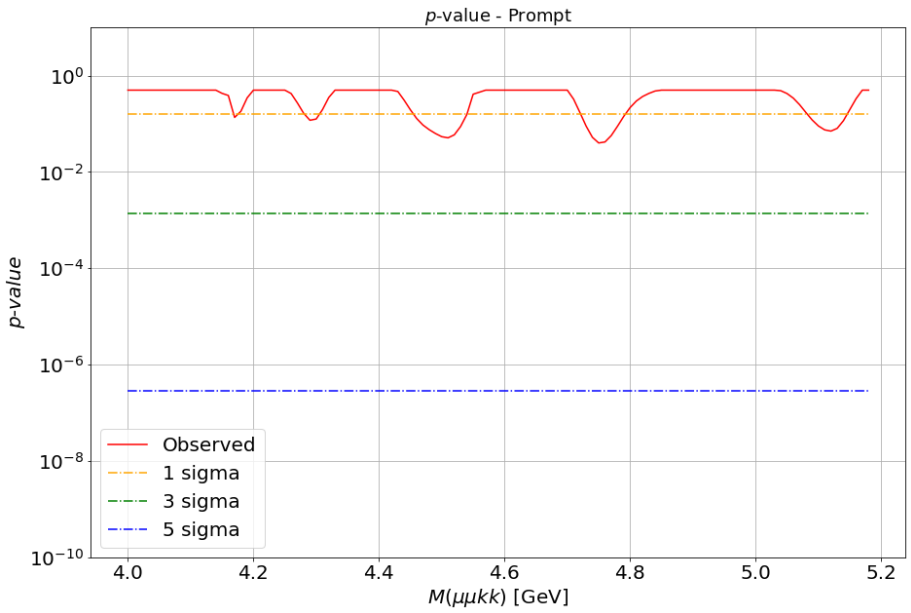




Local p-Value estimation

- Using the same tool used in the case of the rate upper limit calculation, it has been possible to estimate also the local p-value of a possible Y signal along the whole mass spectrum.





- For both searches (**prompt** and **inclusive**) non of the mass point shows a local significance greater than 1σ. No significant structure has been found.

GPU-based techniques for event reconstruction and statistical significance estimation

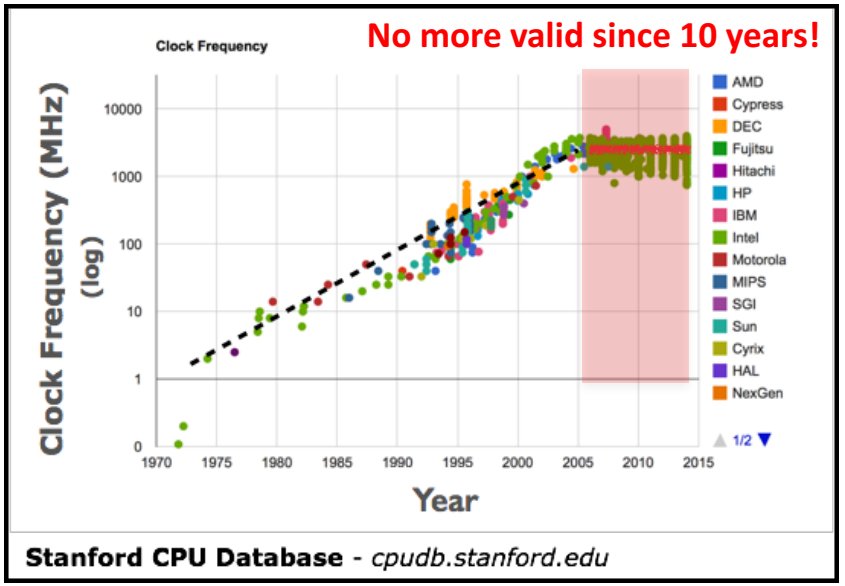
GPU Computing in HEP analysis: the GooFit framework

Moore's Law *:"the number of transistors per unit area would double approximately every two years"*

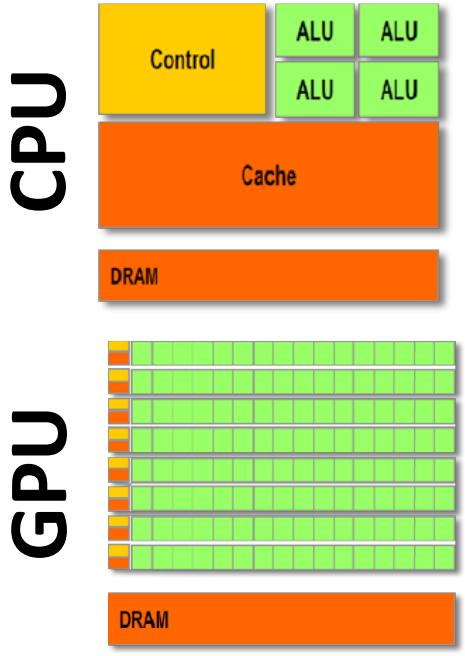
Physical limit: **heat dissipation**

$$P = \mathrm{C} \times V \times f^2$$

V – working tension
C – capacity
f – clock frequency



Future developments **cannot** rely anymore on an **exponential growth of frequency.** A new approach is needed: a possible solution is ***GPU-computing***.



What is a GPU?

Graphic Processing Unit

- **1970s:** first graphical user interface produced requiring dedicated microchips
- **Video games** and **3D graphics**: strong economic stimulus for GPU development

Consequences on GPU architecture:

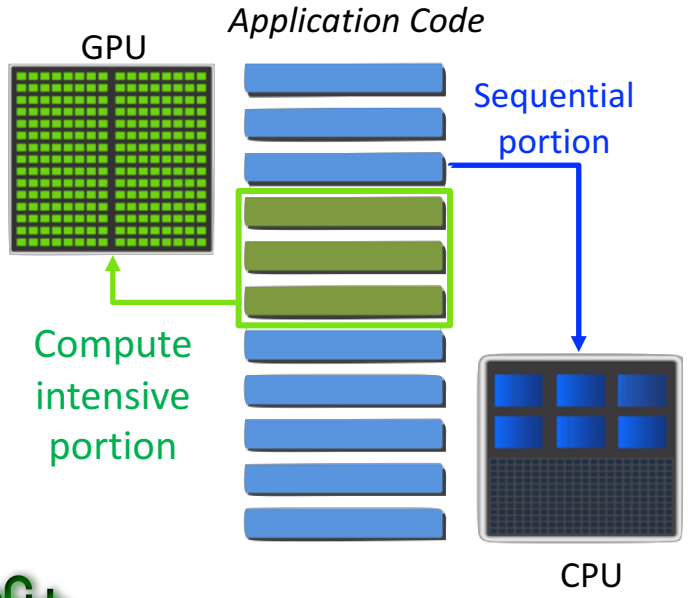
- **Thousands** of cores
- **Big loads of data**
- **Low frequency clock** (~1GHz)
- **Arithmetical operations in a single clock cycle** (*sin,cos,sqrt,1/x, …*)

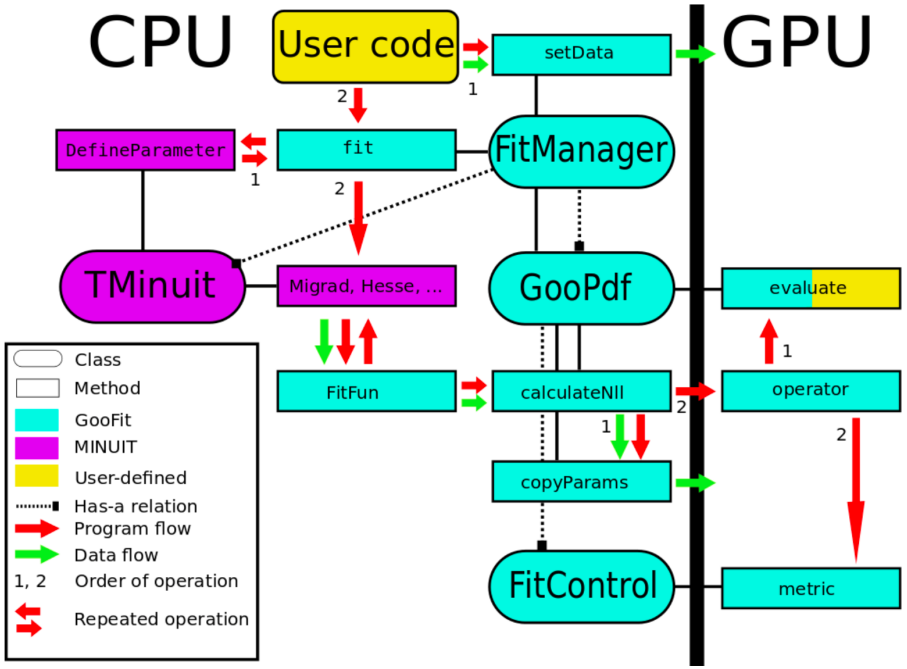
Estimation of statistical significance of a new signal within the GooFit framework on GPUs

GPU Computing in HEP analysis: the GooFit framework

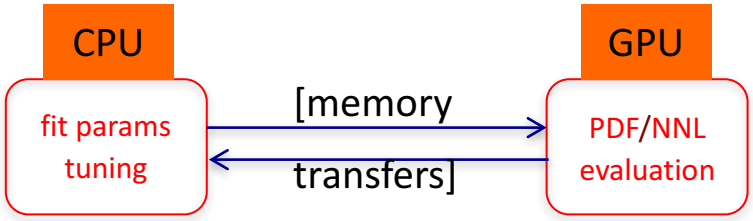
Hetherogeneous GPU-acccelerated computing is the use of a Graphics Processing Unit to accelerate scientific applications (among other apps).

We explored the capabilities of GPU compuiting in the context of the 'end-user HEP analyses' by using *GooFit*.





Goofit is a data analysis tool for HEP, that interfaces ROOT/RooFit to CUDA parallel computing platform on *nVidia* GPU. It also supports OpenMP.



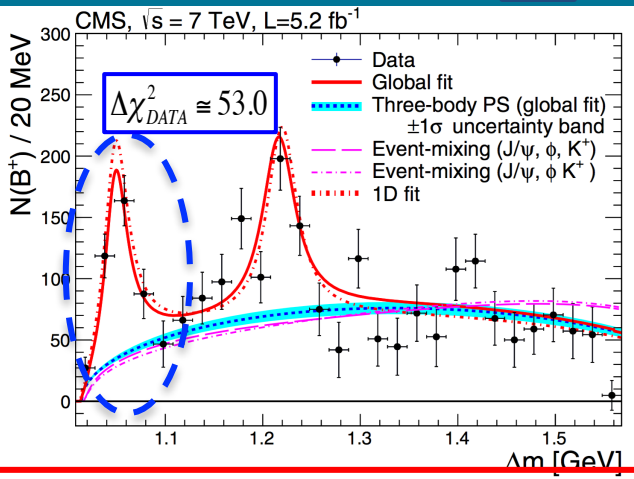
Since v2.0 **Goofit** is completely integrated in python through **PyBindings** and it can run within jupyter notebooks that makes its use even easier.

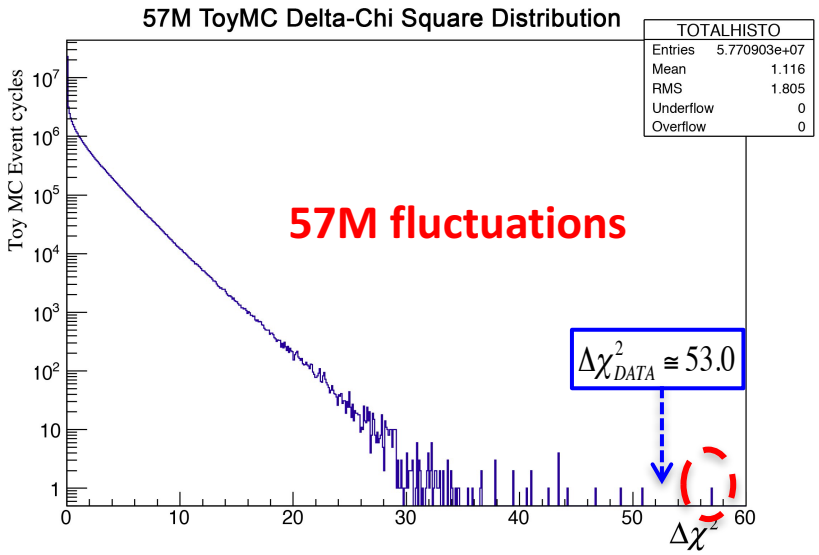
GooFit for local statistical significance

When a **known** signal is found the ***local significance*** must be estimated. MC pseudo-experiments are used to estimate the probability (*p-value*) that **background fluctuations** would - alone - **give rise to a signal as much significant as that seen in the data.**

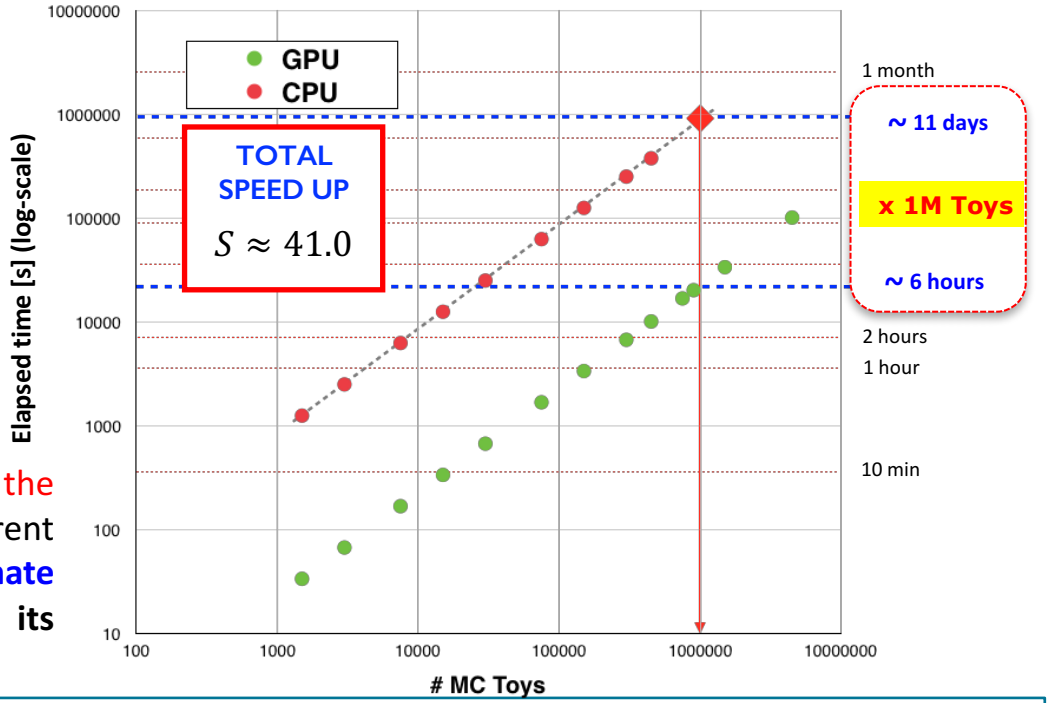
To get a signal significance>5σ, a *p-value* < *3x10*$^{-7}$ is needed, namely at least 3.3M toys are needed. To estimate a signal signif. much more toys are actually needed.

The **final obtained** $\Delta\chi^2$ **distribution** (MC toys production was stopped once **a fluctuation** with $\Delta\chi^2 > \Delta\chi^2_{DATA}$ was found) is built from **more than 57M fluctuations.**





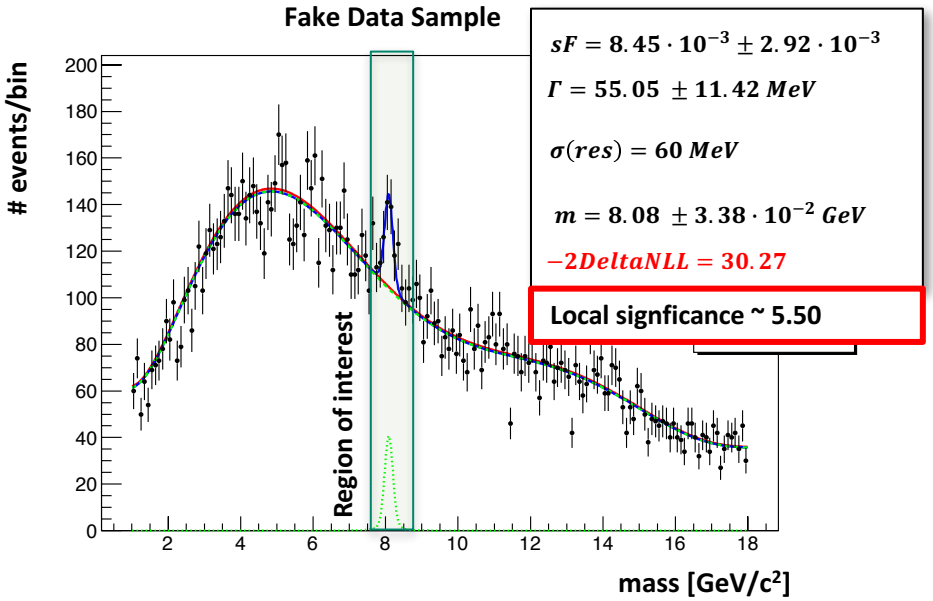
The *GooFit* application running on GPU provides a striking speed-up performance with respect to the *RooFit* application on CPU. It has allowed us to get the final result in **few weeks** instead of **months**!



By means of *GooFit* it has also been easier to explore the (asymptotic) behaviour of a $\Delta\chi^2$ test statistic in different situations in which **the Wilks Theorem (used to estimate the p-value) may apply or does not apply because its regularity conditions are not satisfied**.

GooFit for global statistical significance

When dealing with an unexpected new signal, a *global statistical significance* must be estimated and the Look-Elsewhere-Effect (LEE) must be taken into account. This implies to consider – within the same background-only fluctuation and everywhere in the relevant mass spectrum – any peaking behavior with respect to the expected background model and then a scanning technique must be implemented.



In order to test the effects of the LEE we generated a **pseudo-data inv. mass distribution** of 15K candidates in a generic region of interest (1-18GeV)

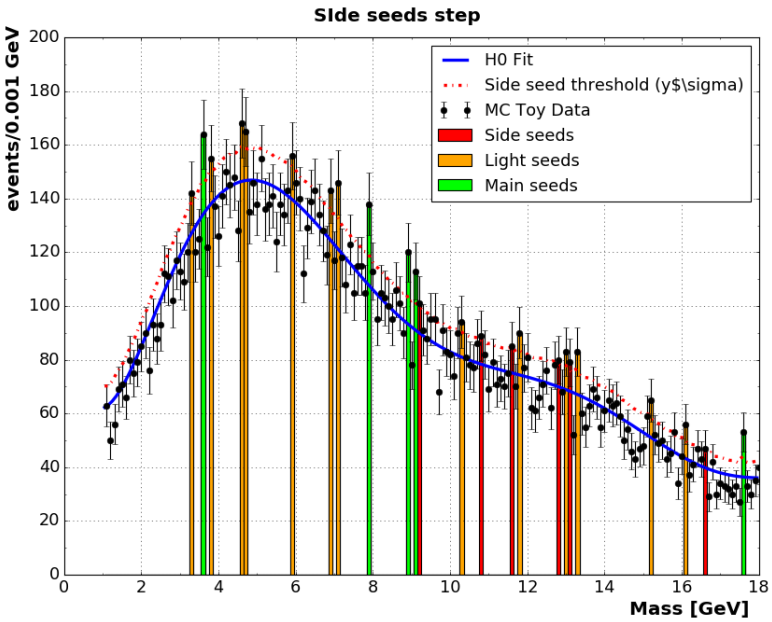
- ***Background* mode (H0) :** 7^{th} order polynomial on
- ***Signal model:*** convolution of a B.W. and a Gaussian (resolution) p.d.f.s, **artificially added @ ~8GeV**

The **scanning technique** has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do not miss any interesting fluctuation

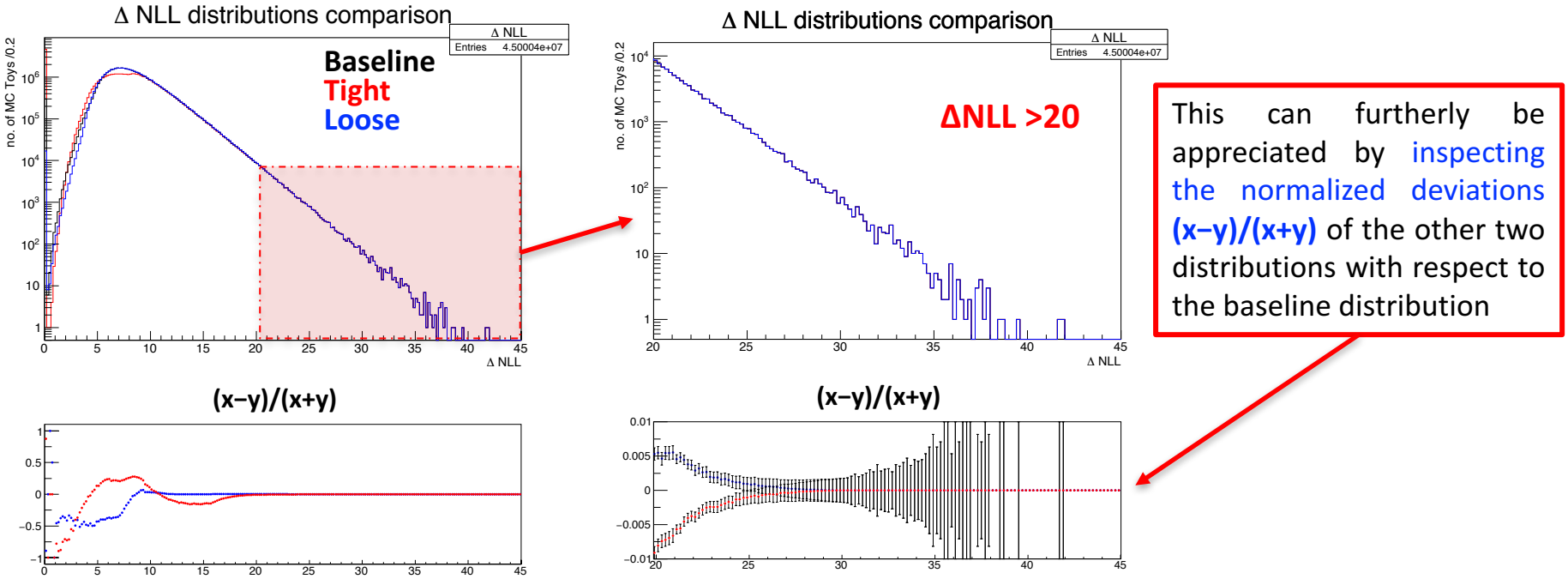
B) Do not select too many small fluctuations

The technique is based on selecting bins flcutuating above the **H0 background expectation value** more than a pre-tuned set of thresholds denoted as ***(x,y,z)***



Scanning technique: systematical bias

The use of a scanning technique **introduces the necessity to to study the possible bias** of this method to the estimation of a global significance. Three clustering configurations are selected: a ***baseline configuration***, a ***looser*** and a ***tighter*** one.



Also we can examine the estimated global significances for the **p-values** corresponding to **different values of local significances**

Clustering configs.	$< fit_{H1} >$	f_{nofit}	Local Significance	4.0σ	4.5σ	5.0σ	5.5σ	6.0σ
Tight (3.00, 1.75, 1.00)	2.2	∼10%	Tight (3.00, 1.75, 1.00)	2.21	2.91	3.58	4.23	5.19
Baseline (2.25,1.50, 1.00)	4.5	∼1%	Baseline (2.25,1.50, 1.00)	2.20	2.91	3.58	4.23	5.19
Loose (2.00, 1.25, 1.00)	6.6	0.1%	Loose (2.00, 1.25, 1.00)	2.19	2.92	3.58	4.23	5.19

It can be concluded that the systematic uncertainty on the p-values associated to the method is negligible.

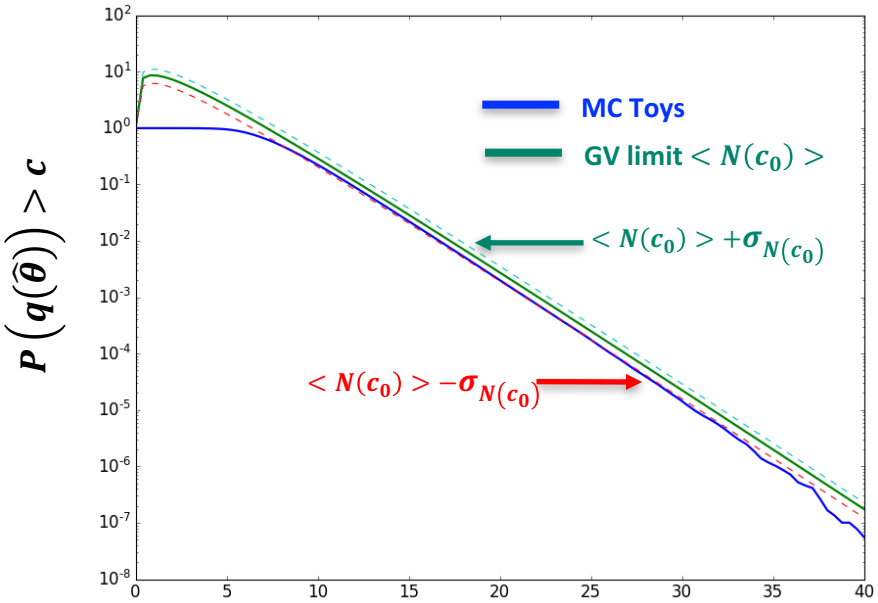
2018 – Limit calculation

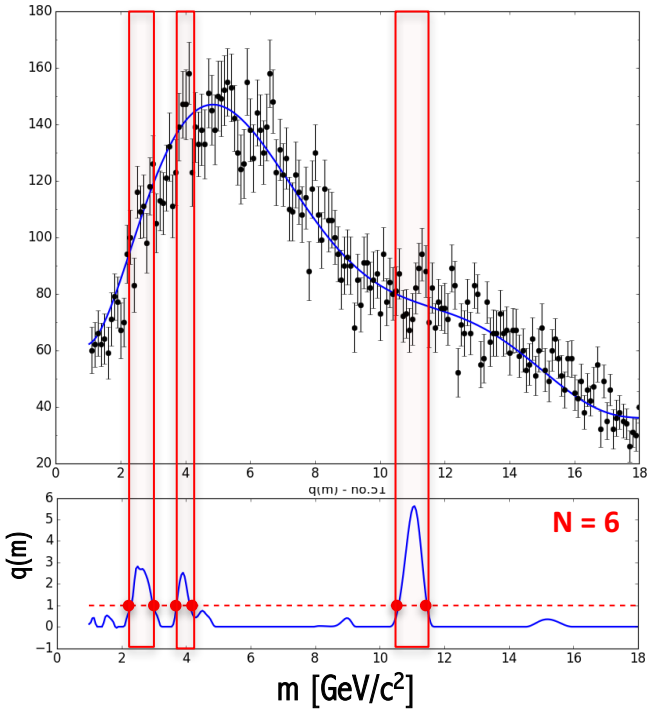
By means of ***GooFit*** , given the speed ups shown, it has also been feasible to compare our results with a method to estimate an **upper bound** for the **global p-value** proposed by E. Gross and O. Vittels [*]. The **G-V method** relies on the estimation of the **average number of *upcrossings*** $< N(c) >$ of $q(\vec{\theta})$, spanning along the $\vec{\theta}$ parameter space, w.r.t. to a desired threshold **c** for the test statistics (in our case the ΔNLL_{data}):

$$P\left(q(\hat{\theta}) > c\right) \leq P\left(\chi_s^2 > c\right) + \langle N(c_0)\rangle \left(\frac{c}{c_0}\right)^{(s-1)/2} e^{-(c-c_0)/2}$$

We set up a procedure [within **GooFit** framework] to estimate $< N(c_0) >$ for **our pseudo-data configuration.** **10k** toys are produced and for each toy a **complete scan** (in **1000** steps) of the mass spectrum is performed.

$< N(c_0) > = 7.3$ $\quad \sigma_{N(c_0)} = 2.4$ $\quad c_0$=s-1=1





Local Sig.	4.0σ	4.5σ	5.0σ	5.5σ	6.0σ
GV method	2.09	2.82	3.48	4.10	4.71
MC Toys	2.20	2.91	3.58	4.22	4.87

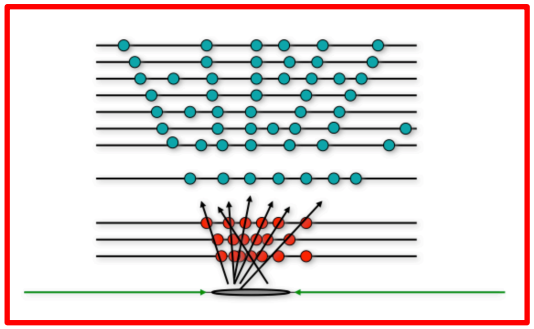
The limit is perfectly compatible with our results with the MC toys procedure

[*] Eur. Phys. J. C (2010) 70: 525–530

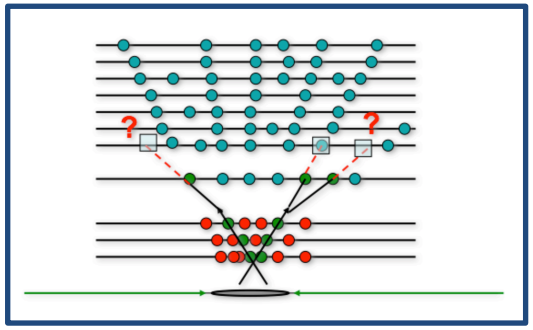
Convolutional Neural Networks for Track Seed Filtering at the CMS HLT

CMS Tracking at HLT

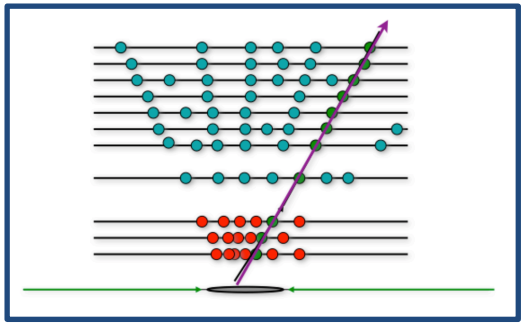
Event selection at CMS is divided in two steps. A first selection is performed by the **L1 trigger** that, based on hardware readout of the *calorimeters and the muon systems*, filter events in **less than 4μs**. The second stage is the **High Level Trigger**. The HLT selection is based on a simplified global track reconstruction.



Seeding



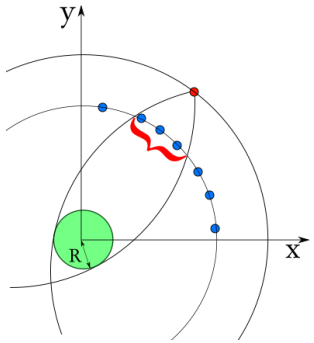
Tracks Building

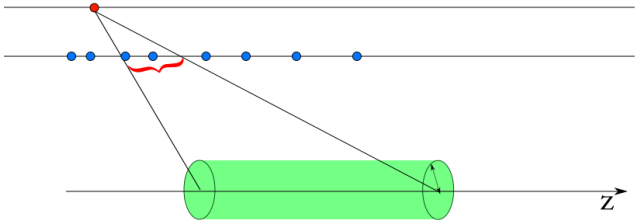


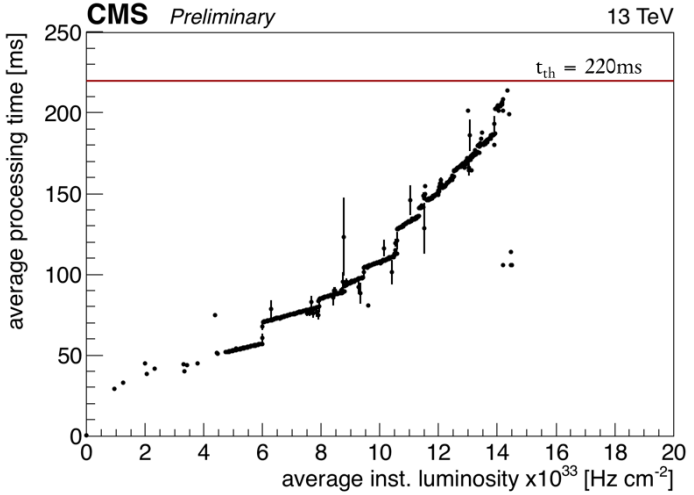
Track fitting

Online reconstruction (HLT) is practically the same reconstruction procedure as the one run offiline. It has to undergo stringent time limits : O(100) ms. It is based on **pixel-only reconstruction.**

Pixel doublet generation: bottleneck due to huge combinatorial background.



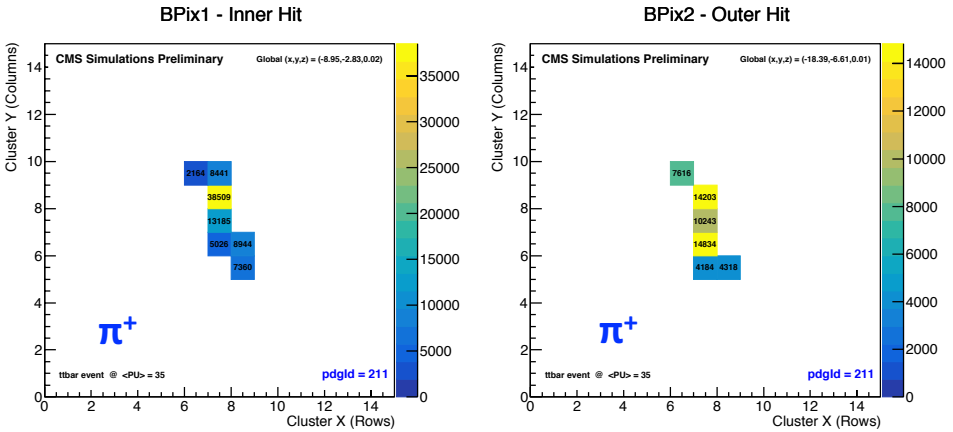




For a single $t\bar{t}$ at $\sqrt{s} = 13\text{TeV}$ with $< PU > = 35$ simulated event: $O(10^6)$**doublets** produced with ***fake ratio*** $\sim O(10^2)$ corresponding to $O(10^4)$ **true doublets**.

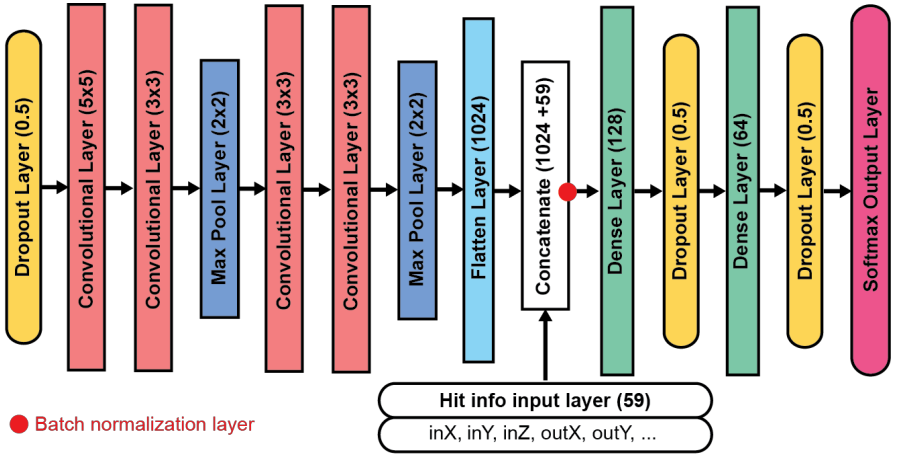
CNN Filtering: the Layer Map Model

- Each seed is build from a couple of hits on the **silicon pixel tracker detector**. Each hit is not simply a point on the detector but it is a collection o **pixlels** (2D) **on or off**. Each pixel is associated with an **ADC** (16 bit levels) level proportional to the charge deposited by a particle.



- We considered each hit as a **15x15** pixel **pad/image** centred in the cluster <u>center of charge</u>
- *A single doublet is considered as a couple of 16x16 matrices*
- Pattern recognition problem : suitable for a Convolutional Neural Network approach

- A single doublet is a **multi levels** image. *Concatenates:*
  - ***CNN architecture***: stack of convolutional layers (4) and max pooling (2)
  - ***"DENSE" architecture***: dense layers (2) fed with the 1-dim reduced images + **doublets infos** (inX,inY,inZ …)



- ***Dropouts & early stopping*** to prevent overfitting
- ***Train & val*** *datasets balanced (0.5)*

Train dataset 3M doublets: generation of $t\bar{t}$ at $\sqrt{s} = 13\text{TeV}$ with $< PU > = 35$ simulated events (via PYTHIA integrated in CMS reconstruction software, CMSSW): $O(10^5)$**doublets** produced with ***fake ratio*** $\sim O(100)$ equals to a $O(1000)$ **true doublets**.

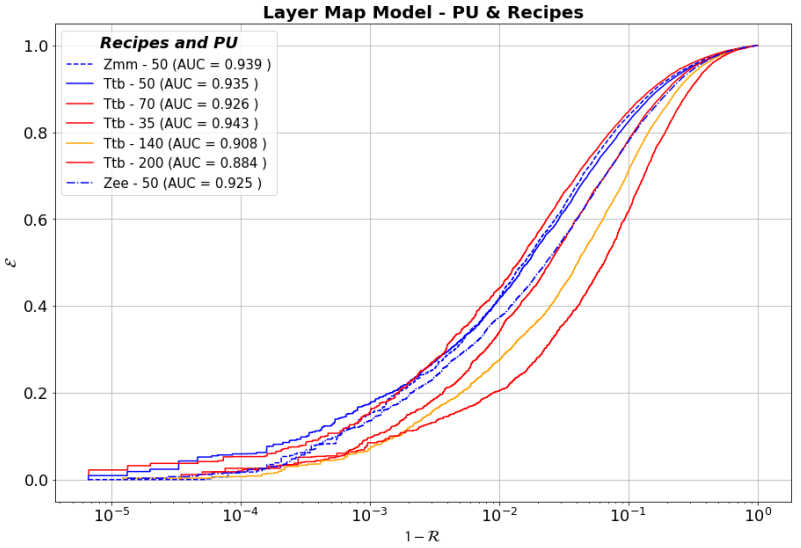
Model testing and validation

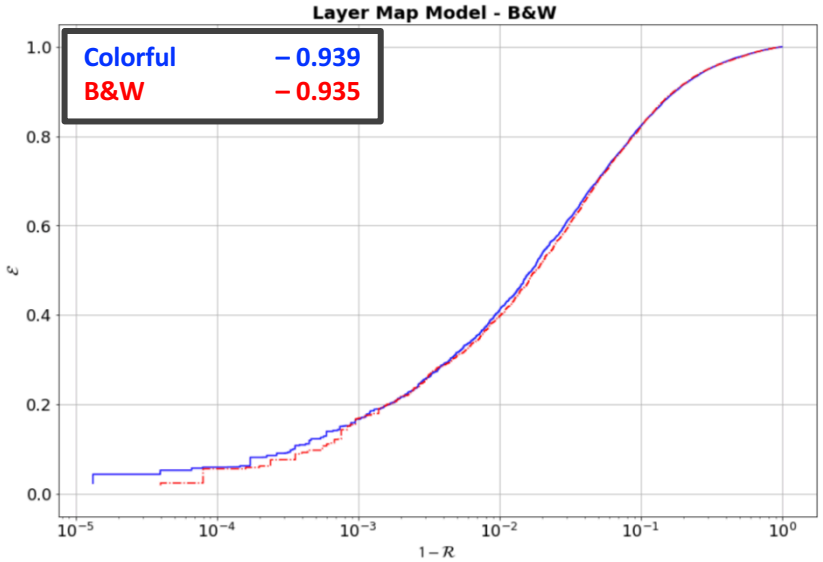
Once the model has been trained, a double cross-check has been carried out generating a set of events both with different **simulated processes conditions** and with increased number of **simultaneous collisions (PU)** with respect to the training sample.

Training configuration

	Data	$<PU>$	AUC	$\mathcal{R}_{90}$	$\mathcal{R}_{95}$	$\mathcal{R}_{98}$	$\mathcal{R}_{99}$	t_{99}	t_{90}
0	Zmm	50	0.940	0.839	0.727	0.556	0.432	0.951	0.499
1	Ttb	50	0.935	0.824	0.703	0.530	0.413	0.951	0.497
2	Ttb	70	0.927	0.783	0.641	0.457	0.338	0.969	0.711
3	**Ttb**	**35**	**0.944**	**0.847**	**0.738**	**0.566**	**0.439**	**0.949**	**0.510**
4	Ttb	140	0.909	0.708	0.548	0.369	0.273	0.974	0.762
5	Ttb	200	0.883	0.615	0.435	0.277	0.207	0.971	0.783
6	Zee	50	0.925	0.782	0.649	0.480	0.372	0.958	0.622

For the training configuration the classifier retain 0.98 efficiency (typical target) reducing of 50% the fake doublets.

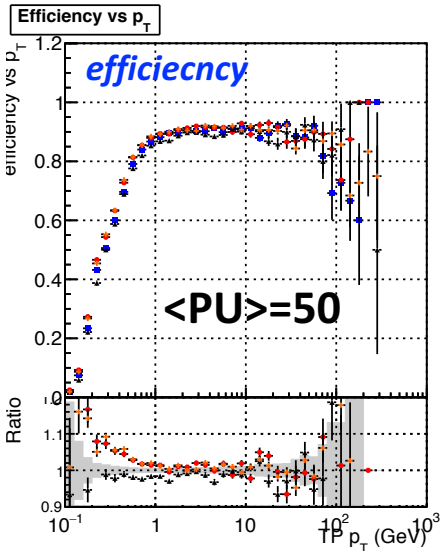


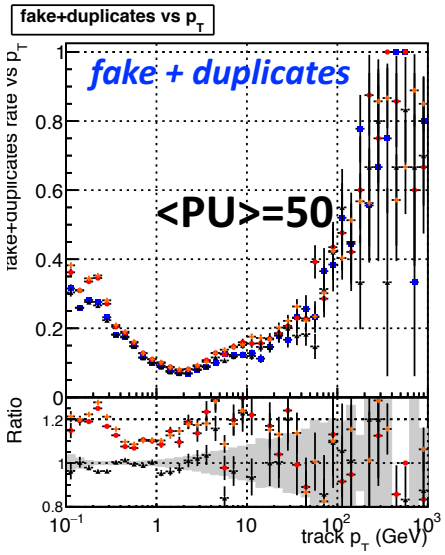


Another has been performed taking into account that, **starting from Run III**, the ADC readout for the **Silicon Pixel** detector is expected to become purely digital, i.e. each pixel will loose any information on the amount of charge collected reducing the output from 16 bits to 1 bit.

CMSSW Integration and Test

All the tests described above have been performed **offline** once the simulated event samples have been produced and processed by the CMSSW. Through **TensorFlow AOT Compilation** it has been possible to integrate the classifier into the **CMS software framework**



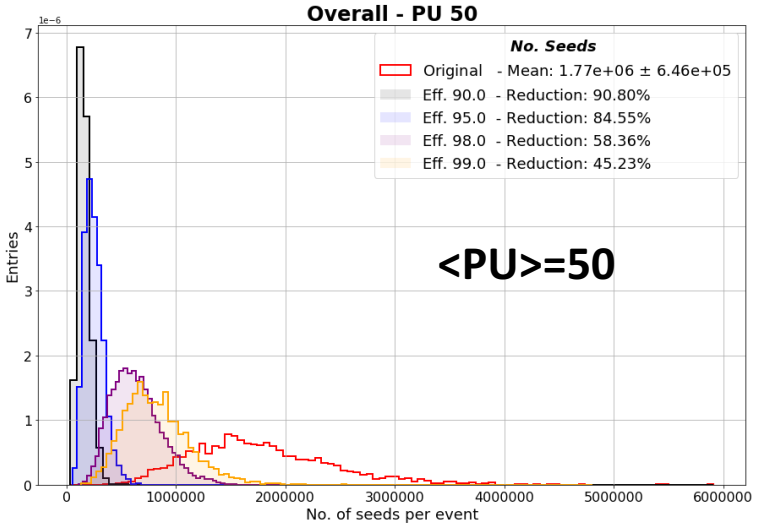


The physics performance of the track reconstruction in which the pixel doublet seeds are filtered by the Layer Map Model **(at different working points)** has been compared to the standard event reconstruction chain.

The overall reconstruction efficiency are almost identical for all the thresholds, with respect to the original configuration.

The comparison of the size of the doublet collection for the main reconstruction step, between the original path and the plugged-in Layer Map Model shows that the doublet set size is reduced at least by 40% of the original size and sometimes even much more depending on the particular step considered.

The outcome of the pixel doublet producer is much more stable and that this feature may be used to better tune the subsequent steps.



THANK YOU

"I am putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do"

HAL9000