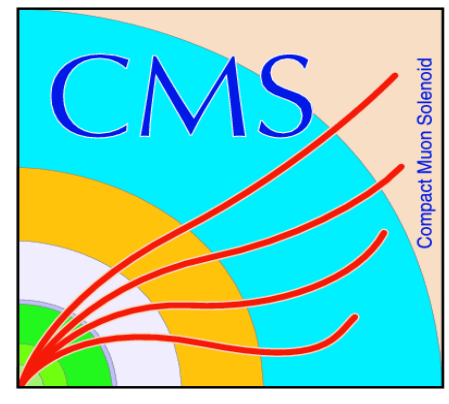


# PERFORMANCE STUDIES of GooFit on GPUs WHILE ESTIMATING the GLOBAL STATISTICAL SIGNIFICANCE of A NEW PHYSICAL SIGNAL



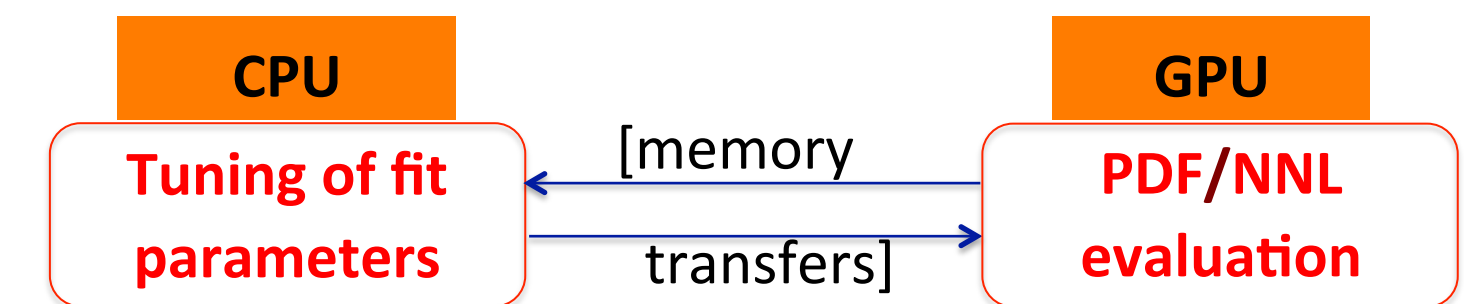
ALEXIS POMPILI<sup>1,2</sup>  
(for the CMS Collaboration)

(1): INFN SEZIONE DI BARI, ITALY  
(2): UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO, ITALY



In the context of High Energy Physics analysis applications, **GooFit** is an **open source data analysis tool** that interfaces **ROOT/RooFit** to the **CUDA** parallel computing platform on **nVidia GPU**. It is exploited in applications enabling the modeling of event data distributions and using (unbinned) maximum likelihood parameter estimation technique. Parameter estimation is a crucial part of many physics analyses. The **Probability Density Function (PDF)** represents a physical model and **its evaluation on large datasets is usually the bottleneck in the minimization task**.

GooFit acts as an interface between the MINUIT minimization algorithm (running on CPU) and a parallel processor (GPU) which allows a **PDF** to be evaluated in parallel. Fit parameters are estimated at each **Neg-Log-Like-lihood** minimization step on the **host side** (CPU) while the **PDF/NLL** is evaluated on the **device side** (GPU).



Description and details about GooFit : R.Andreassen et al., *GooFit: a library for massively parallelising maximum-likelihood fits*, J. Phys.: Conf. Ser. 513 (2014) 052003 [CHEP2013 Proceedings].

To test the computing capabilities of GPUs with respect to CPU cores, a high-statistics pseudo-experiments (toys) technique has been implemented in **RooFit & GooFit** frameworks in order to estimate a **p-value** and thus the (local or global) statistical significance of a signal reconstructed from data. The p-value is the probability that background fluctuations would - alone - give rise to a signal as much significant as that seen in the data.

**Hardware setup** consists in 2 servers (hosted @ ReCas-Bari Data Center): one equipped with 2 nVidia TeslaK20 and 32 cores (16+16 by HT), the other with 1 nVidia TeslaK40 and 40 cores (20+20)

## RooFit with PROOF-LITE

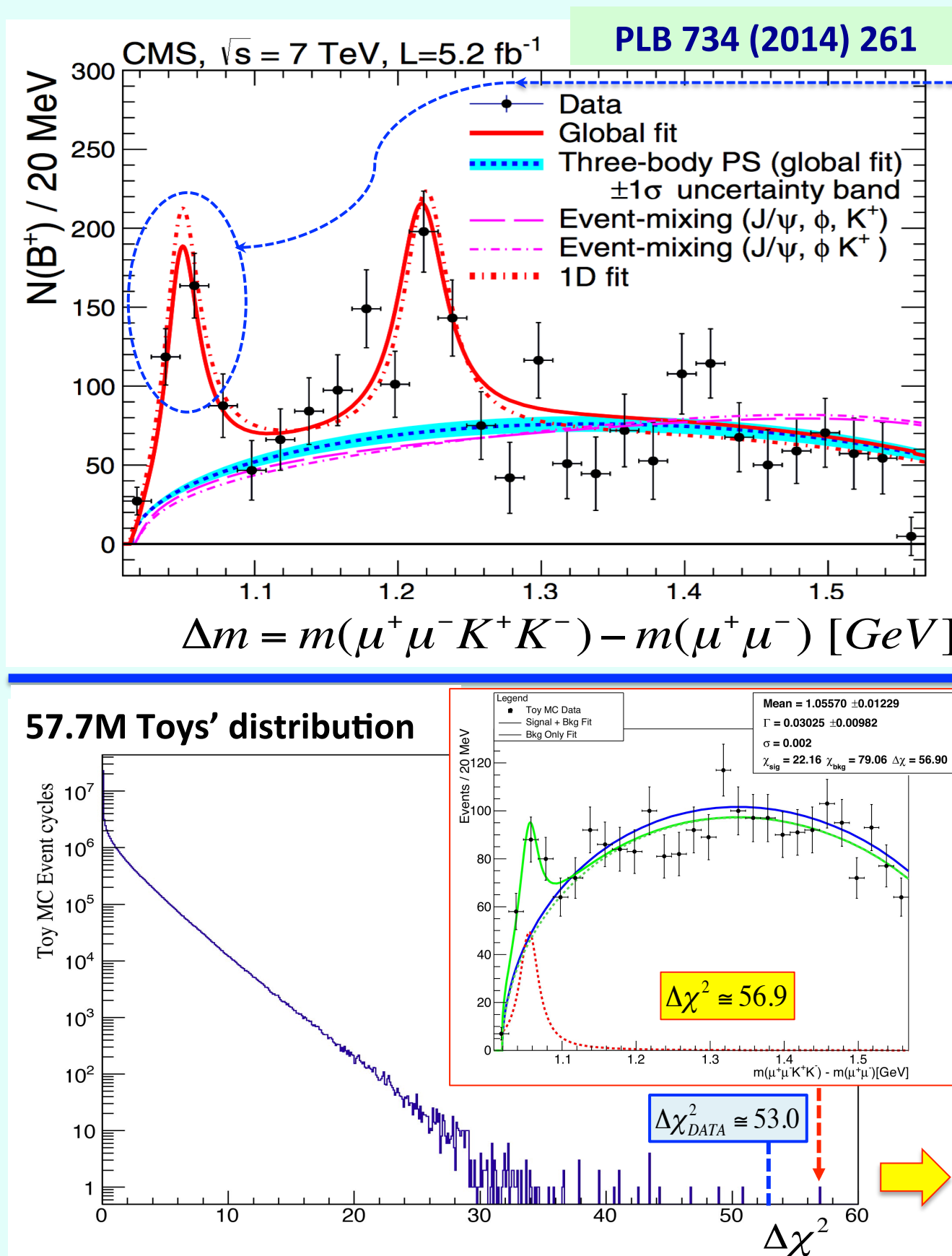
To efficiently run **RooFit MC toys** in parallel on the 72 CPUs available on the 2 servers hosting the GPUs, **PROOF-Lite** is used. It has a **pull architecture**.

## GooFit with MULTI PROCESS SERVER

The **nVidia Multi Process Server (MPS)** tool allows the execution of - up to 16 - simultaneous processes on the same GPU acting as a **scheduler** and allowing a **balanced full use** of the GPU.

MPS / PROOF-LITE show a **similar** behaviour of the **speed up** as a function of processes / workers; both their **Amdhal fits** indicate a serial overhead of ~3% for the MC toys' application execution.

## MC TOYS for LOCAL SIGNIFICANCE



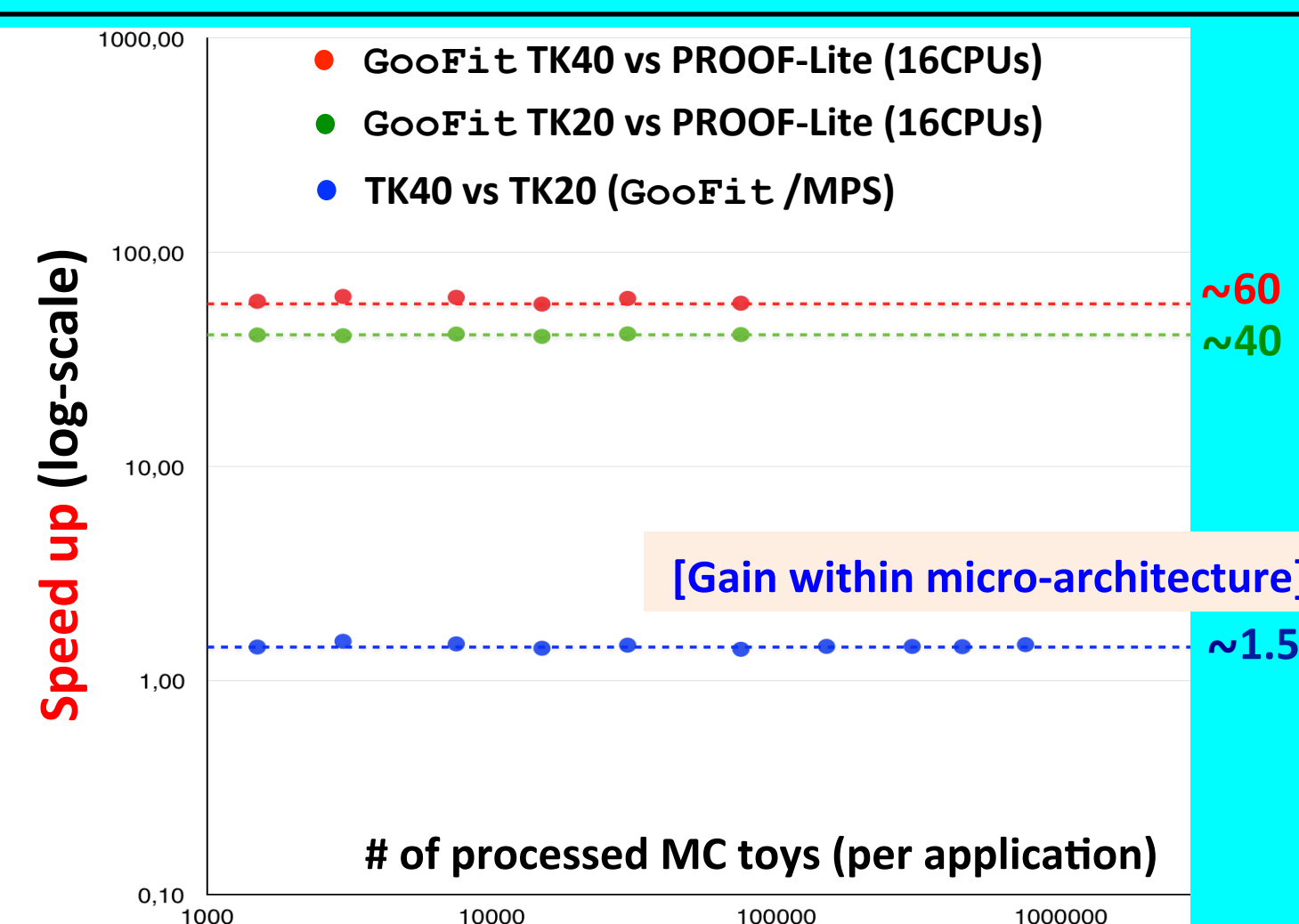
**Aim** : to estimate the **local statistical significance** of the structure observed by CMS close to the kinematical boundary of the  $J/\psi \phi$  inv. mass, in the 3-body decay  $B^+ \rightarrow J/\psi \phi K^+$  [PLB 734 (2014) 261], and compatible with an exotic charmonium-like signal already observed by CDF.

### ToyMC cycle :

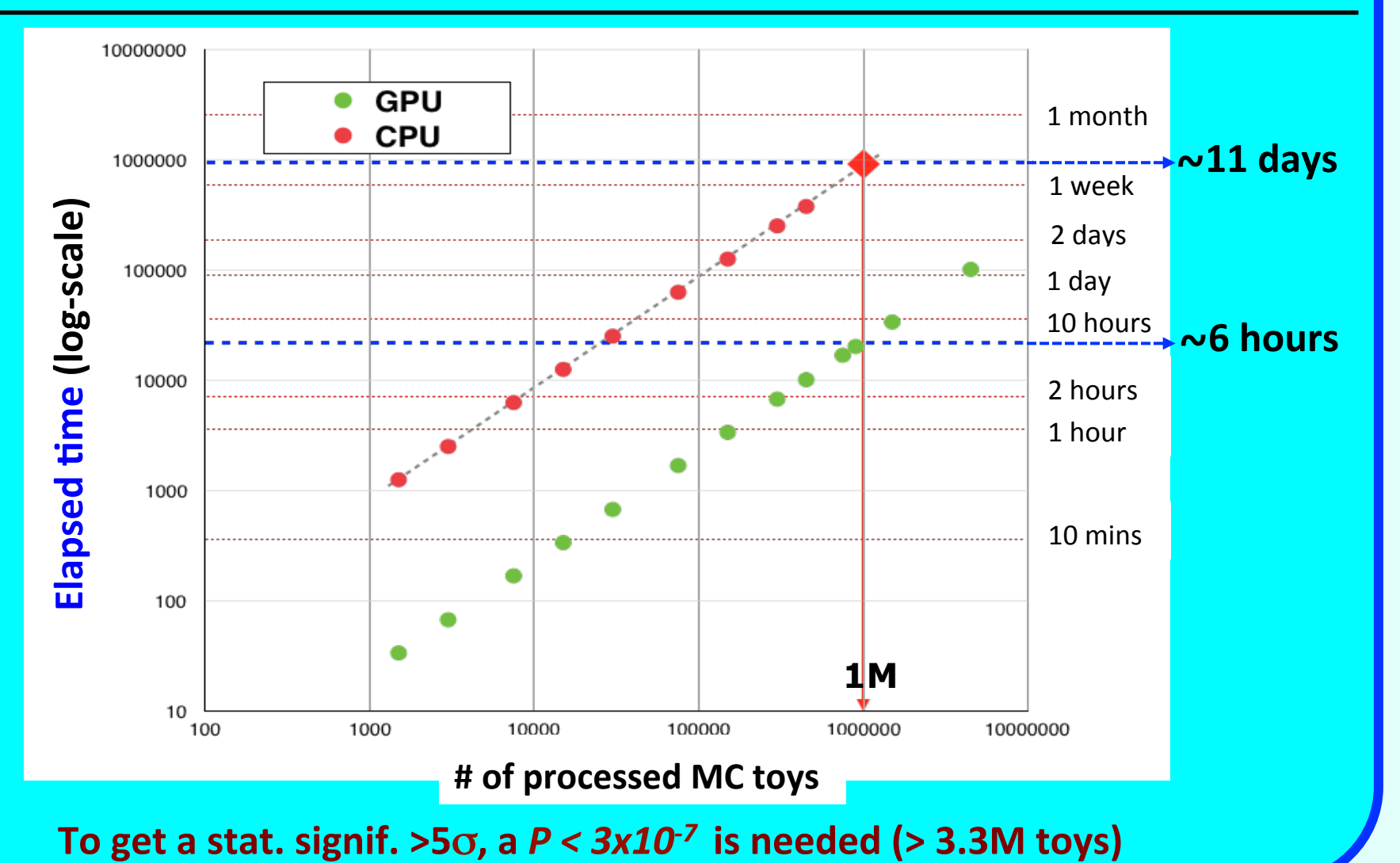
- 0) Generate a background distribution (3-body phase-space model)
- 1) Perform **H0 Binned ML fit** with same bkg model @ generation
- 2) Make **8 H1 BML fits** (signal model is a **truncated Voigtian** function; resolution fixed @2MeV) in the  $\Delta m$  region of interest by trying different starting values (2 masses & 4 widths). Signal yield constrained to be >0.
- 3) Choose the fit with the best  $\Delta\chi^2$  (test statistic). Fill its distribution over the sample of MC toys.

$$p\text{-value} : P = \int_{\Delta\chi^2_{DATA}}^{\infty} f(\Delta\chi^2) d(\Delta\chi^2) \approx (57.7 \cdot 10^6)^{-1} \approx 1.73 \cdot 10^{-8} \Rightarrow Z\sigma = \Phi^{-1}(1-P)\sigma \approx 5.52\sigma$$

Compare: 1 PROOF-Lite job using 16 workers (on 16 CPUs) with: 1 GooFit/MPS job running 16 processes on TK40/TK20



From the point-of-view of the end-user/analyst having at its own disposal 72 CPUs and 3 GPUs (1 TK40 & 2 TK20) on 2 servers :



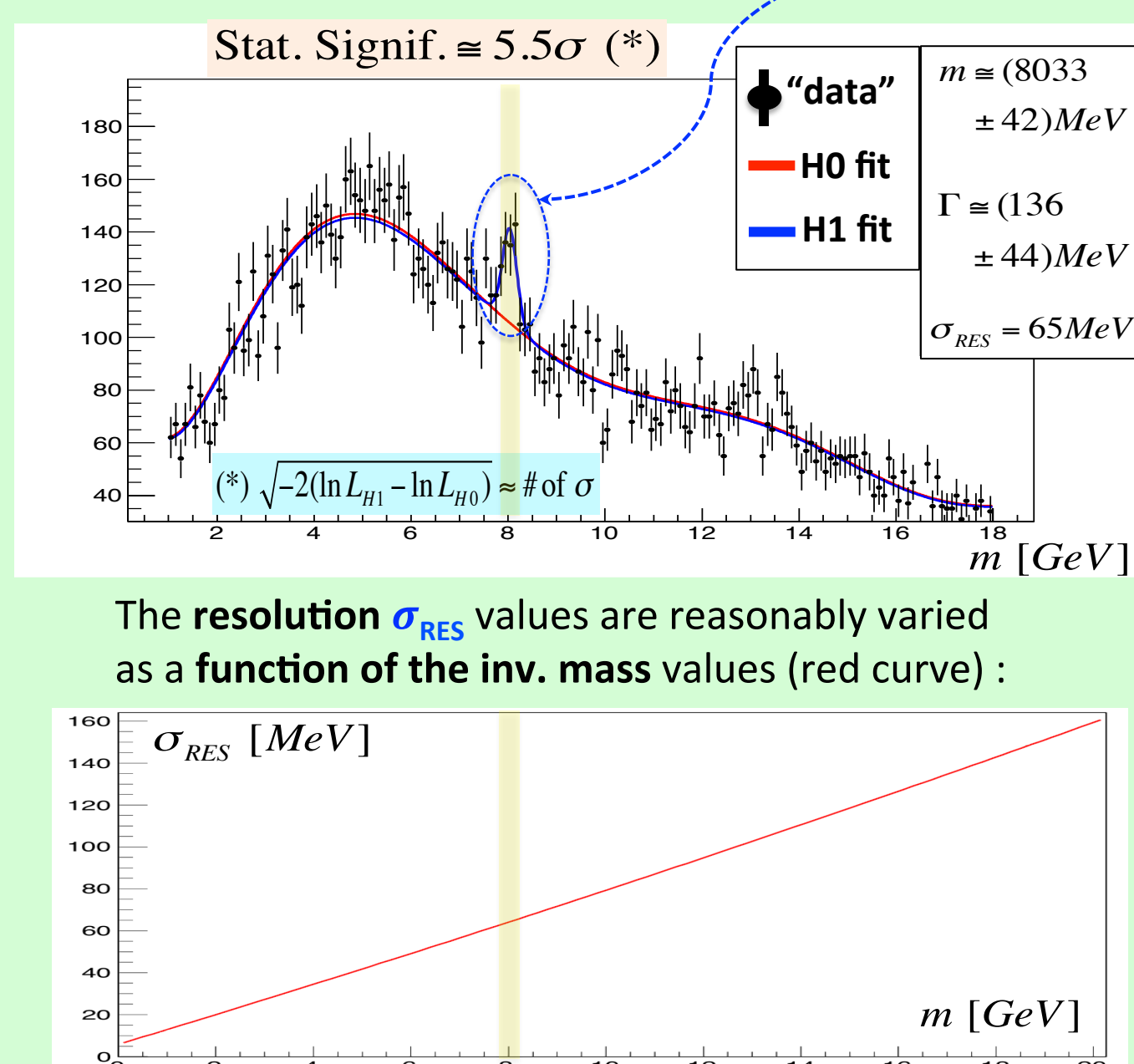
Compatible with the lower limit of  $5\sigma$  for the stat. signif. quoted in the CMS paper on the basis of 50.5 millions of MC toys (by **RooFit**)

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 behaviour with respect to the expected background model.

## MC TOYS for GLOBAL SIGNIFICANCE

A **pseudo-data** inv. mass distribution of 15K candidates in a generic region of interest (1-18GeV) is generated according to an invented **background model** (7th order polynomial) on the top of which any desired amount of **significant signal** can be **artificially added** @  $\sim 8\text{GeV}$



The **LEE inclusion** is addressed with a mass scan **coupled to a clustering technique** to **identify** the peaks from significant fluctuations

The **clustering approach** is designed to satisfy **two concurrent requirements** :

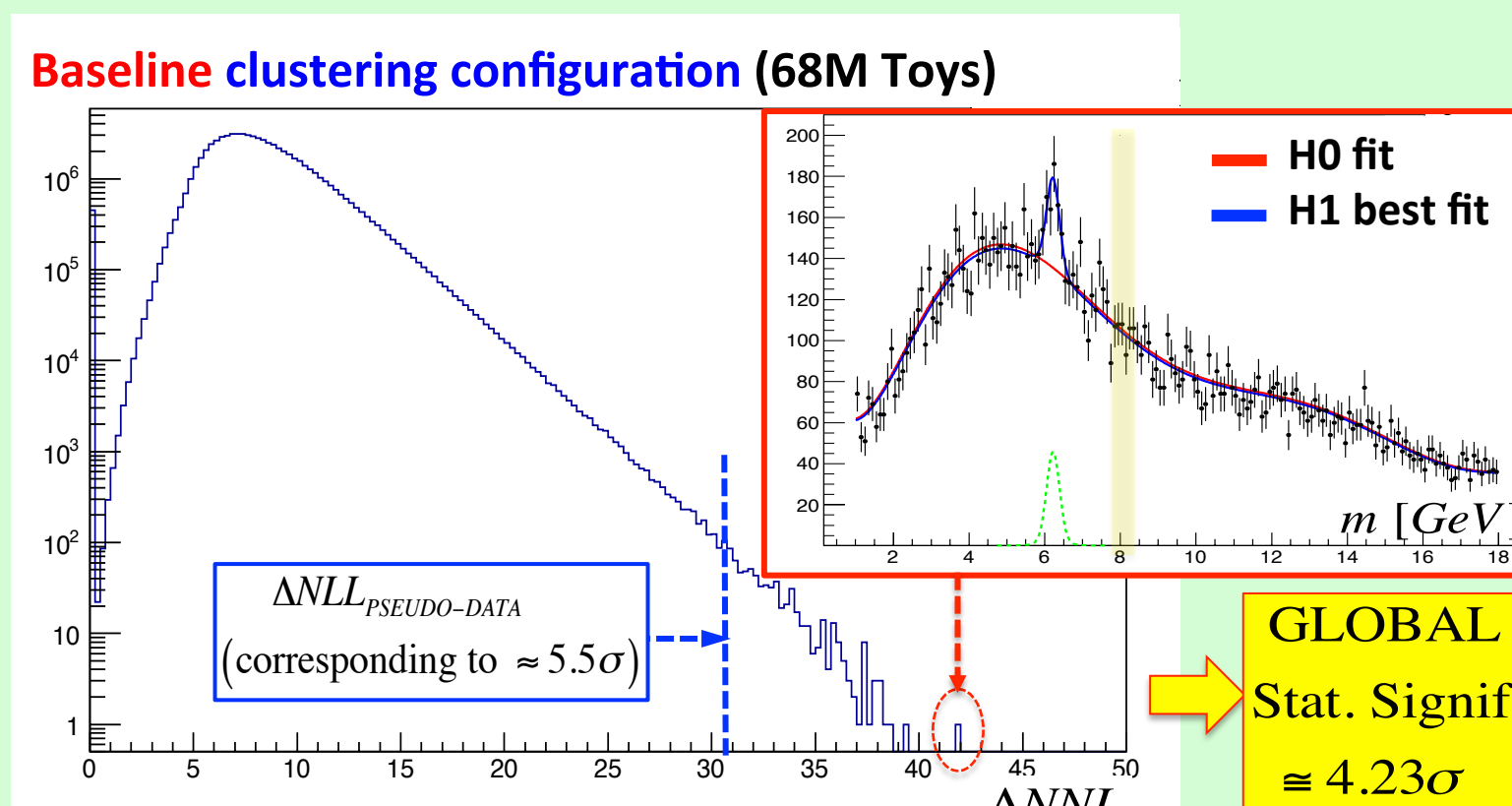
- do not miss any interesting fluctuation
- do not select too many small fluctuations

For each toy, the procedure starts from the **H0 Binned ML fit** (same bkg model @ generation).

While scanning the mass bins :

- 1) search for a **seed-bin** (its content fluctuates more than  $x\sigma$  strictly above the H0 fit);
- 2) add any **side-bin** to the seed-bin if fluctuating more than  $z\sigma$ , otherwise seed-bin forms a 1-bin cluster;
- 3) check for a **light seed-bin** fluctuating more than  $y\sigma$  (with  $z < y < x$ ) if it has at least one other side-bin with more than  $z\sigma$ .

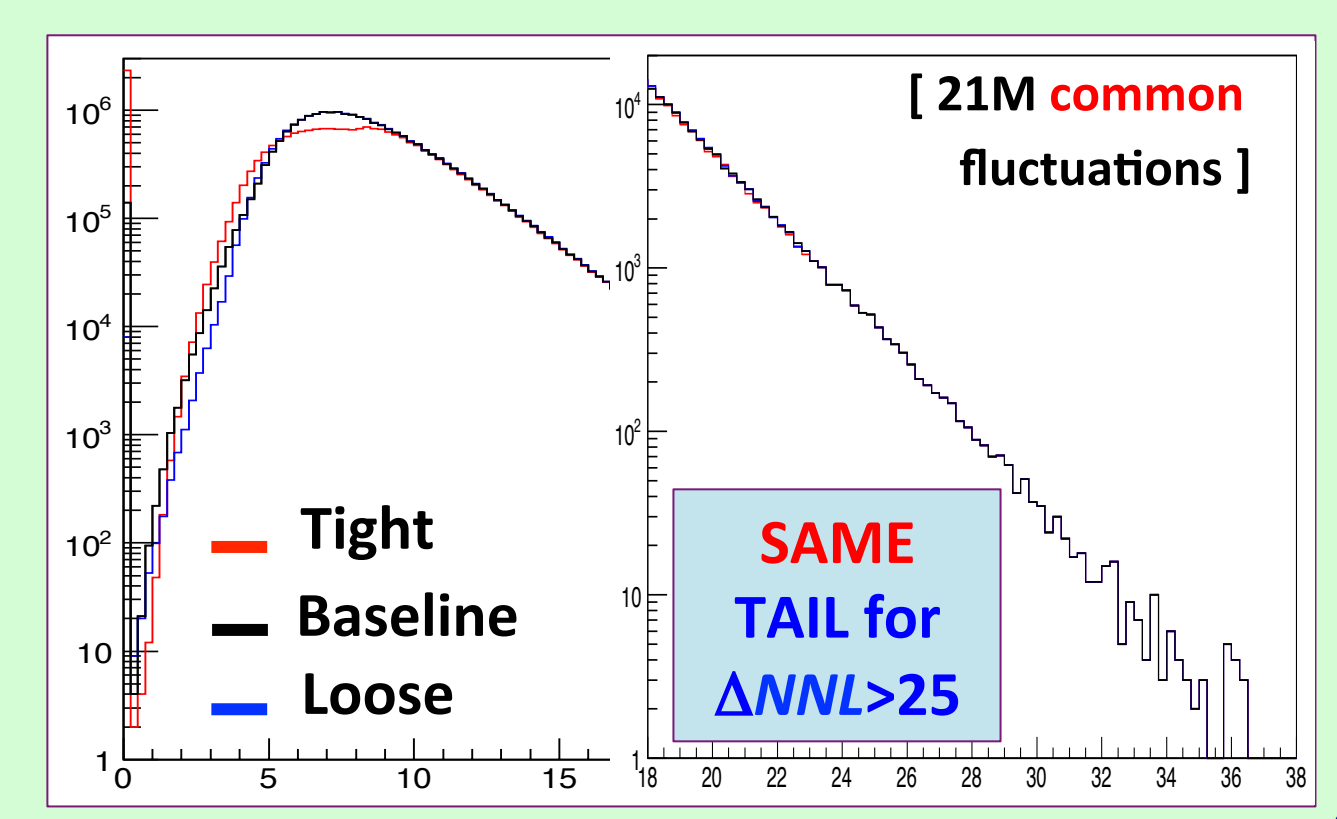
For each toy, make **all the H1 BML fits** (signal model is a Voigtian function; yield constrained to be >0) and choose the fit with the best  $\Delta\text{NNL}$  (test statistic). Get its distribution over all the MC toys.



### GLOBAL STATISTICAL SIGNIFICANCE (#sigma)

Approx. LOCAL Stat. Signif. (*)	4.0sigma	4.5sigma	5.0sigma	5.5sigma	6.0sigma
Tight	2.21	2.91	3.58	4.23	5.19
Baseline	2.20	2.91	3.58	4.23	5.19
Loose	2.19	2.92	3.58	4.23	5.19

The method behaves **suitably stable** and its associated **systematic uncertainty is negligible**



18<sup>th</sup> Advanced Computing and Analysis Techniques in Physics Research  
21-25 August 2017 – University of Washington – Seattle

