Università degli Studi di Bari

Dipartimento Interateneo di Fisica "M. Merlin"

PhD School in Physics XXXI Cycle

Activity report on the of the second year of the Doctoral School.

Phd Student: Adriano Di Florio

Supervisor: Dr. Alexis Pompili

6th November 2017

In this report, concerning the second year of Doctoral School in Physics, the research work is presented in Section 1, the participation to conferences, workshops and schools in Section 2 and the list of publications in Section 3.

1 Research work

During the second year the research work has mainly focused on three topics:

1. the study and development of applications of the GPU-based analysis tool known as `GooFit` in the context of charmonium-like exotic states' searches in the CMS experiment; specifically for amplitude analysis fitting and statistical significance estimation.
2. The search for the charmonium-like Y(4140) state and its partners in the $J/\psi\phi$ mass spectrum within CMS.
3. The study and development of machine learning techniques for filtering the track seeds (pixel hit doublets) at CMS High Level Trigger.

1.1 GooFit applications in the context of exotic QCD studies

1.1.1 Introduction to `GooFit`

The word heterogeneous computing refers to an enhancement of application performances that can be obtained by offloading compute-intensive portions to the GPU, while the remaining code still runs on the CPUs. In the context of High Energy Physics (HEP) analysis application, `GooFit` is an under development open source data analysis tool used in applications for parameters' estimation, that interfaces ROOT/RooFit to the CUDA parallel computing platform on nVidia's GPUs. `GooFit` acts as an interface between the MINUIT minimisation algorithm and a parallel processor which allows a Probability Density Function (PDF) to be evaluated in parallel. Fit parameters are estimated at each negative-log-likelihood (NLL)

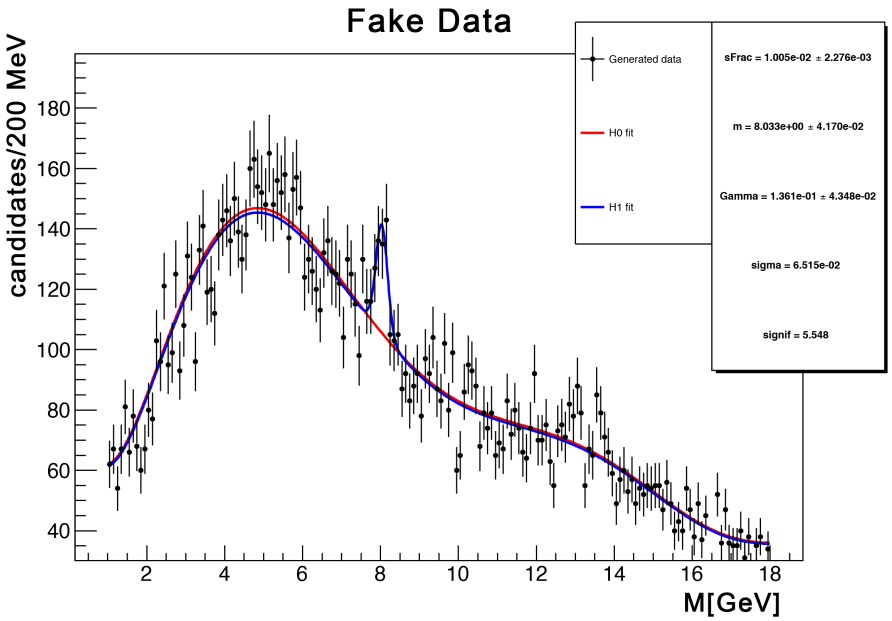

Figure 1: Pseudo-data simulated invariant mass distribution. In red H0 fit, in blue H1 fit.

minimisation step on the *host side* (CPU) while the PDF/NLL is evaluated on the *device side* (GPU).

1.1.2 Global statistical significance estimation by MC toys with GooFit

A high-statistics toy Monte Carlo technique has been implemented both in ROOT/RooFit and `GooFit` frameworks with the purpose to estimate the local statistical significance of an already known signal. The optimised `GooFit` application running on GPUs has provided striking speed-up performances with respect to the RooFit application parallelised on multiple CPUs by means of PROOF-Lite tool.

The ongoing work concerns the extension of this `GooFit` MC toys significance estimation method when a new unexpected signal is reconstructed and a global significance must be considered. In this case the Look Elsewhere Effect must be taken into account and a scanning technique needs to be implemented in order to consider any relevant random peaking activity with respect to the background model over the whole mass spectrum within the same fluctuation. Thus a scanning technique has been developed on the basis of a clustering approach defined as follows:

1. For each MC Toy iteration a distribution based on the background p.d.f. model is generated in the range of the whole mass spectrum and the *Null Hypothesis* fit is performed.

2. Search for a *seed* bin, namely for a bin whose content fluctuates more than $x\sigma$ strictly above the value of the background function in the center of that bin (where σ is the statistical error of the considered bin).

3. Add any side bin to the seed bin if it shows a content that fluctuates more than $z\sigma$ strictly above the value of the background function in the center of that bin, otherwise the seed bin forms a 1-bin cluster.

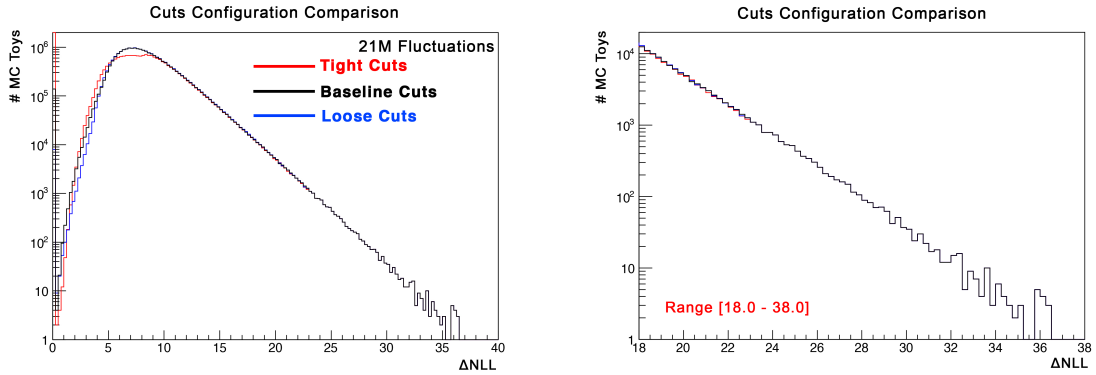

Figure 2: Left ΔNLL distributions for 21 millions of common fluctuations for the three configurations: baseline cut in black, tight in red and loose in blue. Right: zoom on the range 18.0-38.0.

Clustering configs.	$< fit_{H1} >$	f_{nofit}
Tight (3.00, 1.75, 1.00)	2.2	$\sim$10%
Baseline (2.25,1.50, 1.00)	4.5	$\sim$1%
Loose (2.00, 1.25, 1.00)	6.6	0.1%

Table 1: Man number of alternative hypothesis fits per toy ($< fit_{H1} >$), fraction of toys with no fit (f_{nofit}) for the three different clustering configurations.

4. Check also for "light" seeds: bins that fluctuates more than $y\sigma$ with $z < y < x$ and with at least a side bin fluctuating more than $z\sigma$. In case of positive result a cluster is formed.

The three configuration parameters x (single seed threshold), y (side bin threshold) and z (additional sided seed threshold) need to be tuned in order to avoid either missing any possible interesting fluctuation and selecting too many irrelevant fluctuations.

In order to test this procedure a pseudo-data invariant mass distribution of 15K candidates in a generic region of interest [1-18GeV] has been generated according to an invented background model (7th order polynomial) on the top of which a fake signal, mimicked with a Voigtian fucntion, is added close to 8 GeV [Figure 1]. The local statistical significance of this peak is 5.5σ.

Three clustering configurations, i.e. three values for the (x, y, z) parameters, have been tuned in order test the behavior of the clustering method and to estimate the systematic uncertainty associated to the clustering technique. After some tests with different cuts three configurations are chosen for (x, y, z) : a set of baseline cuts $(2.25, 1.50, 1.00)$, a set of tight values $(3.00, 1.75, 1.00)$ and a set of loose values $(2.00, 1.25, 1.00)$ (see Table **??** for specifics).

Local Significance	**4.0σ**	**4.5σ**	**5.0σ**	**5.5σ**	**6.0σ**
Tight (3.00, 1.75, 1.00)	2.21	2.91	3.58	4.23	5.19
Baseline (2.25,1.50, 1.00)	2.20	2.91	3.58	4.23	5.19
Loose (2.00, 1.25, 1.00)	2.19	2.92	3.58	4.23	5.19

Table 2: MCToys stimated global significance for the three configurations with respect to different local significance values associated to different pseudo-data distributions.

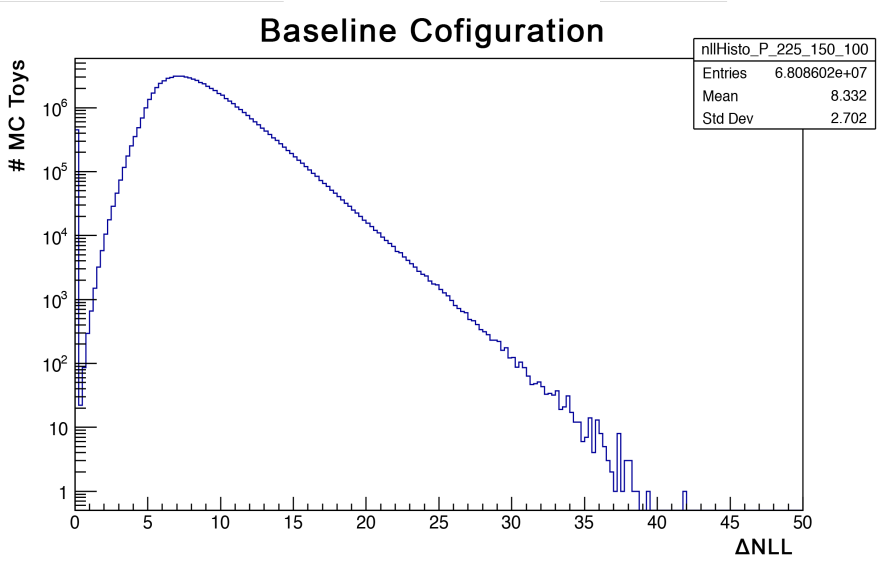

Figure 3: ΔNLL distribution for 68 millions of toys for the baseline configuration of clustering technique.

The three configurations are run on the same common set of 21 millions fluctuations and the three resulting ΔNLL distributions are shown superimposed in Figure 2. Focussing on the region of interest for the estimation of the statistical significance, i.e. the trail of the ΔNLL distribution ($\Delta NLL > 20$), it is evident that there is no sensitive difference by inspecting either the three distribution tails and the estimates of the global significance for the p-values (as reported in Table 1). Thus the systematic uncertainties associated with the method are negligible. Specifically the global significance for the pseudo-data signal presented in Figure 1 results to be 4.23σ for all the three configurations.

In conclusion, the whole ΔNLL distribution for the baseline clustering configuration with about 68M entries is shown in Figure 3. The ongoing work concerns the estimation of the statistical uncertainty affecting the estimation of the p-value.

1.1.3 Amplitude analysis fit of $B^0 \to J/\psi K^+\pi^-$

Traditional *Dalitz Plot* analyses deal with 3-body decays without any vector state as a daughter particle. For example many analyses of B o D meson decays at the *B-factories* dealt with pseudoscalars (pions and/or kaons, charged or neutral) in the final states. In these cases the decay amplitudes are calculated in a 2-dimensional parameter space, namely the *Dalitz Plot* space itself.

In the 3-body decays with vectors in the final state the decay amplitude has to be calculated on an *n-dimensional* parameter space within the helicity formalism. Among the searches for exotic QCD states in CMS, there are some decay modes that need a full *amplitude analysis* fit to explore the nature of possible exotic charmonium-like states appearing as intermediate resonances :

- $B^0 \to J/\psi K^+\pi^-$, to search for $Z(4430)$, $Z_c(4240)$ and $Z_c(3900)$ states;
- $B^+ \to J/\psi\phi K^+$, to search for $Y(4140)$ and other structures in the $J/\psi\phi$ system.

The former decay mode is being studied in the CMS Bari group.

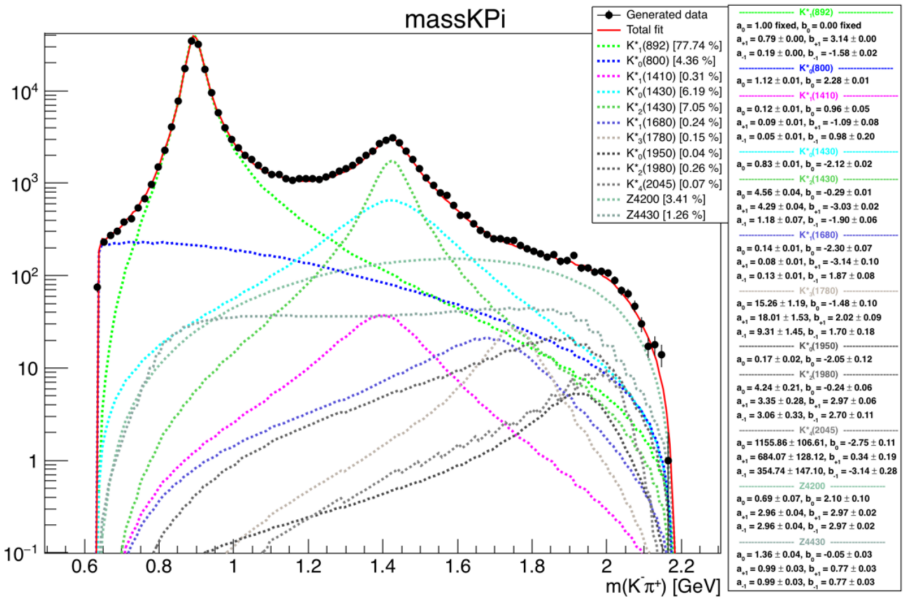

Figure 4: Projection on $m^2_{K\pi}$ spectrum of the four dimensional generated data with superimposed the projection of the total fit function with the K^*s and Zs components (displayed separately).

The *amplitude analysis (AA)* fit strategy requires to check if fitting the data with a model based only on the K^*s system is able to reproduce the real distributions without the need of extra Z states; if the fit will be not satisfactory a contribution for the $Z(4430)$, $Z_c(4240)$ and $Z_c(3900)$ have to be taken into account.

The most relevant intermediate K*s resonances are considered [$K_0^*(800)$, $K_1^*(892)$, $K_1^*(1410), K_0^*(1430), K_2^*(1430), K_3^*(1780)$] and they contribute to the p.d.f. with 28 fit parameters (one *absolute value* and one *phase* for each helicity amplitude; one amplitude for each spin-0 K^*, three for each K^* with spin greater than zero). Such a complex 4-dimensional fit requires high computational capabilities and very long fitting times (days) if carried out by means of `RooFit`. Relevant effort has been spent into the porting and partial re-development of the fitting code in the GooFit framework in order to run it on the ReCas servers equipped with GPUs. Specifically, in the last year the main tasks have been the following:

- the development of the necessary tools to perform the AA fit including also the efficiency correction and the combinatorial background interpolation.
- The implementation of a model that could distinguish B^0 and $\bar{B}^0$ events.
- The development of an additional model representing the $Z(4430)$ and the $Z_($4200) (see Figure 4).

Very encouraging results have been achieved since the `GooFit` fit takes only 25 minutes performing over 1000 `MIGRAD` calls with all the 28 parameters and the efficiency correction and the combinatorial background description included

Additional effort is being devoted to develop a *Goodness of fit* estimation procedure in order to conclude the validation fit procedure before fitting the real data.

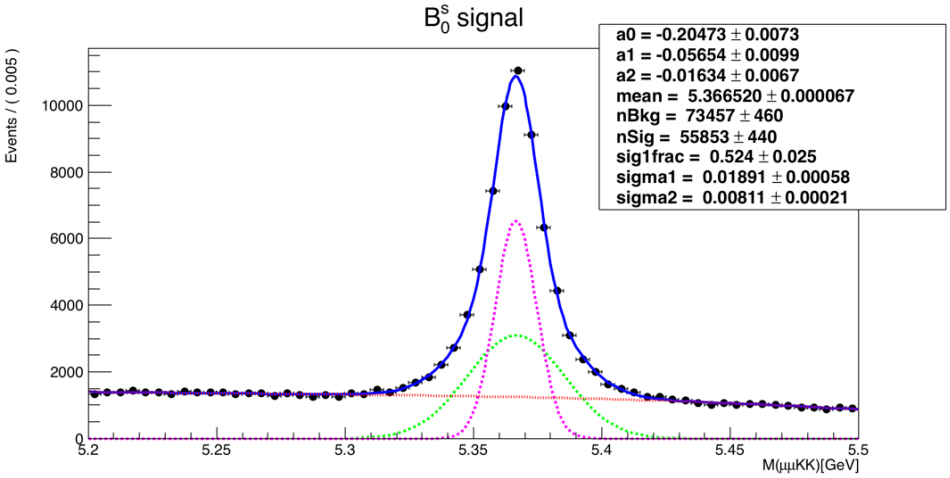

Figure 5: The B_s^0 signal fit in the control window (5.2 - 5.5) MeV. The signal is modelled as the sum of two gaussian (green and pink dotted lines) with the same mean and different widths σ_1 and σ_2. The combinatorial (red dotted line) background is parametrized as a 2^{nd} Chebychev polynomial p.d.f..

1.2 Inclusive search in the $J\psi/\phi$ spectrum of the prompt and non-prompt production of charmonium-like states in pp collisions

Hadron Spectroscopy has experienced a renaissance in the last decade thanks to the experimental findings at *B-factories*, *Tevatron* and recently at the *LHC*. Quarkonium has became again a tool for discoveries of new phenomena in the complex realm of low-energy QCD. A new wide zoology of quarkonium-like states, many of them with manifestly exotic characteristics (the so-called XYZ states), still needs to be understood within a possibly consistent framework. The development of theoretical models, such as tetraquark or hadron molecule models, is not yet able to provide an unified explanation of these states. The analyses of LHC Run-I data provided (and are still providing) new experimental observations and measurements for exotic hadrons.

Even if LHCb is the most suitable experiment to contribute to this new spectroscopy, CMS has proven, see e.g. [1] [2] [3] [4], to be able to provide a few important results despite the absence of hadronic identification.

One of the neutral charmonium-like XYZ meson states is the Y(4140) state. The CDF collaboration at the Tevatron reported the first evidence for the $Y(4140)$ state in 2009 and confirmed it later in 2011 with higher statistics. The Y(4140) state was appearing as an intermediate state in the decay $B^+ \to Y(4140)K^+ \to J/\psi\phi K^+$, close the $J/\psi\phi$ kinematic threshold. Since then several interpretations have been proposed for the Y (4140) decaying into $J/\psi\phi$, none of them entirely convincing: $D_s^*\bar{D}_s^*$ molecule, compact tetraquark $cs\bar{c}\bar{s}$, threshold kinematic effect, hybrid charmonium, weak transition with $D_s\bar{D}_s$ rescattering.

In 2014, the CMS Collaboration observed (with a statistical significance greater than 5σ) the $Y(4140)$ peaking structure in the $J/\psi\phi$ invariant mass spectrum. Later, in 2015, the D$\emptyset$ collaboration found evidence of prompt and non prompt direct production of this state, in $p\bar{p}$ collisions ($p\bar{p} \to Y(4140) + X$ with $Y \to J/\psi\phi$), by studying inclusively the $J/\psi\phi$ mass spectrum [5]. On the other hand LHCb collaboration observed [7], by means of the first amplitude analysis of the decay $B^+ \to J/\psi\phi K^+$, even four $J/\psi\phi$ structures, each with a statistical significance greater than 5σ. The lightest of them is the Y(4140) and it appears

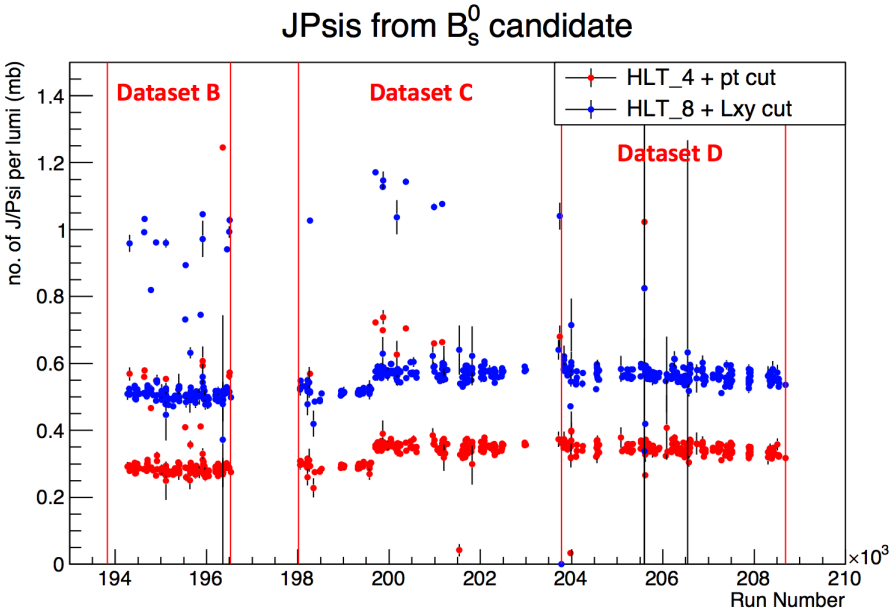

Figure 6: J/ψ yield normalized to integrated luminosity per run with respect to run number. For boht, red and blu events, it is requested that $p_t(J\psi) > 8.0$ GeV and $L_{xy}/\sigma(L_{xy} > 3.0$. In red events passing HLT_DoubleMu4_Jpsi_Displaced trigger selection. In blue events passing HLT_Dimuon8_Jpsi trigger selection.

to be possibly described as a cusp, even if a resonant interpretation is also possible with a mass consistent with, but a width a much larger than, previous measurements of the claimed Y(4140) state.

In [8] the authors propose these structures to be interpreted as S-wave tetraquarks, with $[cs][\overline{c}\,\overline{s}]$ diquark-antidiquark composition.

Given the latest experimental results it becomes crucial to confirm or refute the D$\emptyset$ result for the inclusive search. In case of a positive confirmation this would rule out a cusp interpretation of the Y(4140) because a state that appears both as an intermediate state in a 3-body B meson decay and as a state promptly produced in pp collisions could definitely be considered as a genuine resonance.

By exploiting Run-I and Run-II data, CMS has the capability to look for the presence of the Y(4140) and its partners inclusively in the $J/\psi\phi$ mass spectrum either promptly produced or coming by decays of beauty mesons. In order to validate the analysis procedure, two regions of $J/\psi\phi$ invariant mass spectrum need to be explored; a control window centered on the B_s^0 meson mass region and a low mass region (search window) in which these states may appear. In both regions one looks for known and unknown states decaying to $J/\psi\phi$ (where $J/\psi \to \mu\mu$ and $\phi \to KK$).

The full 2012 $\sqrt{s} = 8$ TeV Run-I pp collisions data sample have been analyzed so far and two main results extracted in the control window region study have been achieved:

- Reconstruction of the $B_s^0 \to J/\psi\phi$ signal as a control channel (see the fit to the B_s^0 signal in Figure 5). The extracted control signal shows good agreement with already published B_s^0 signals for Run I data.

- Study of the stability of the yield of J/Ψ, produced from a $B_s^0 \to J/\psi\phi$ decay, along all the Run-I data taking. This study has been done both to check eventual biases in the

events selection and to test the effect of two different possible HLT triggers selection. The results show any biases as can be deduced from the steady trend for the yield normalized to integrated luminosity (Figure 6).

The search strategy of a Y resonance in the $\mu\mu KK$ final state is suitable for both the full 2012 $\sqrt{s} = 8$ TeV Run-I pp collisions data sample and the $\sqrt{s} = 13$ TeV Run-II data being collected in 2016-2018. In addition, for the Run II data taken in 2017/2018, an additional strategy can be explored, thanks to the development of thanks to the development of two new specific high level trigger paths (*HLT_DoubleMu2_ Jpsi_ DoubleTrk1_Phi* and *HLT_DoubleMu2_ Jpsi_DoubleTkMu0 _Phi*). Since the $J/\psi\phi$ background is expected to arise mainly from combinatorial background of kaon track pairs, it would be interesting to investigate the $J/\psi\phi$ invariant mass spectrum also through the rare $\phi \to \mu\mu$ decay channel, namely in a much cleaner (and rare) topology [6]. Therefore $J\psi\phi$ system could be explored relying only on muons reconstruction, bypassing the issue related to the lack of particle identification for hadrons.

Therefore the efforts for 2018 will aim to:

- reproduce the analysis done (in the $\mu\mu KK$ mass spectrum) for the 2012 data on the new data coming from 2015-2018 Run-II data.
- For the 2017/2018 data analysis, explore the $J/\psi\phi$ mass spectrum not only for the $\mu\mu KK$ final state but also for the $\mu\mu\mu\mu$ one.
- For the same data explore also the $J/\psi\omega(\omega \to \mu\mu)$ spectrum in the final state looking for Y(3940) signal.

These will be also the ingredients of the main data analysis foreseen in the Ph.D. thesis

1.3 Convolutional Neural Network for Track Seed Filtering at the CMS High-Level Trigger

Future development projects for the Large Hadron Collider will allow to integrate higher luminosity than that already collected. The ultimate goal will be to reach a peak luminosity of $5 \cdot 10^{34} cm^{-2} s^{-1}$ for ATLAS and CMS experiments, as planned for the High Luminosity LHC (HL-LHC) phase. This will directly result in an increased number of simultaneous proton collisions (pileup), up to 200, that will pose new challenges for the CMS detector and, specifically, for track reconstruction in the Silicon Pixel Tracker.

One of the first steps of the track finding workflow is the creation of track seeds, i.e. compatible pairs of hits from different detector layers, that are subsequently fed to the higher level pattern recognition steps. However the set of compatible hit pairs is highly affected by combinatorial background and thus would have a strong impact onto the next steps of the tracking algorithm. To avoid the processing of a significant fraction of fake doublets a new method is being developed. It relies on taking into account the shape of the hit pixel cluster to check the compatibility between two hits. To each doublet a collection of two images built with the ADC levels of the pixels forming the hit cluster is associated , see Figure **??** for an example. Thus the task of fakes' rejection can be considered as an image classification problem for which Convolutional Neural Networks (CNNs) have been widely proven to provide reliable results.

To test the feasibility of this kind of approach, a Monte Carlo $t\bar{t}$ within the CMS software framework at the center of mass energy of $\sqrt{s} = 13$ TeV, with average pileup $< PU >= 35$ and bunch time spacing of 25 ns. About 10^6 doublets are produced per each event and the

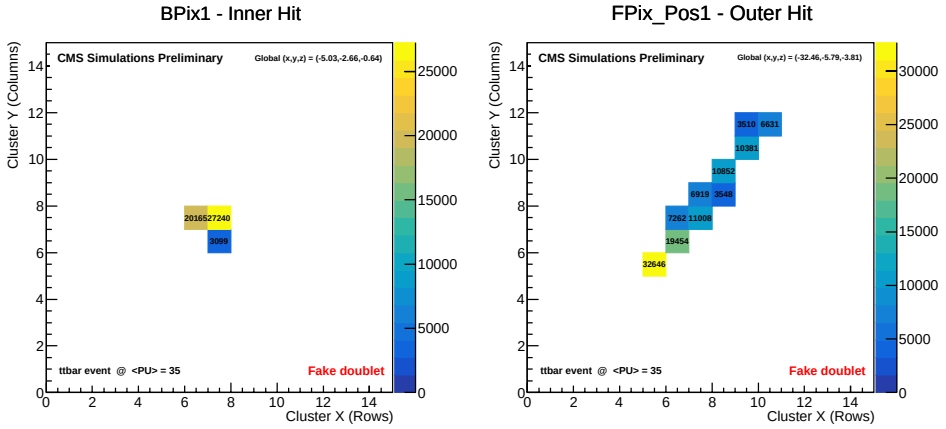

Figure 7: Example of a BPix1-BPix2 true doublet generated from *ttbar* simulated events corresponding to a π^+ track.

ratio between *true* and *fake* doublets is between 300-400 (i.e. only about 3000 true doublets are produced per event by the seed generator). For each doublet 537 parameters are stored (450 pixels, 63 doublet info and 24 track parameters). The core structure for the doublet filtering classifier (*layer map model*) is shown in Figure 8 and it concatenates:

1. *CNN block*: conventional stack of convolutional layers (4) with 5×5 or 3×3 filters and max pooling layers (2) whose output is reduced to a one dimensional structure through a flatten layer that returns a 1024 element vector.

2. *Dense block*: stack of two fully connected layers that are fed with the one dimension reduced images from the previous block and 59 further doublets info (such as hits' detectors and coordinates).

On the whole, 1000 events are simulated with 3 millions of doublets, 800 for the training dataset, 150 for testing and 50 for validation. The ROC curve for the CNN classifier is shown in Figure 9. The area under the curve (AUC) is greater than 0.96: while rejecting half of the fake doublets the network reaches an efficiency greater than 99%. These first promising results show that a CNN approach for this kind of applications can be very effective and reliable.

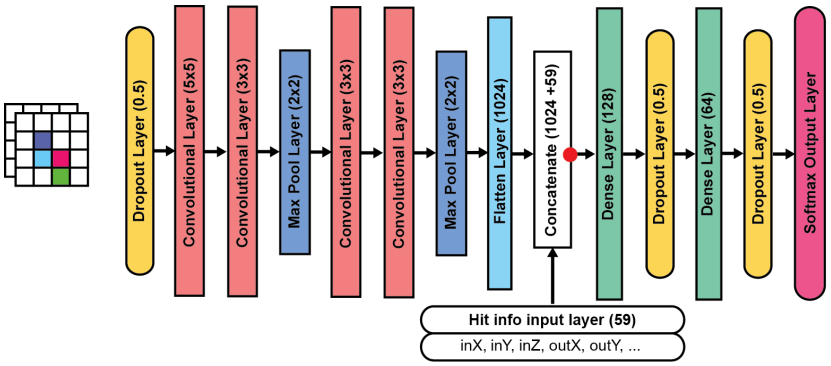

Figure 8: Layer map model architecure

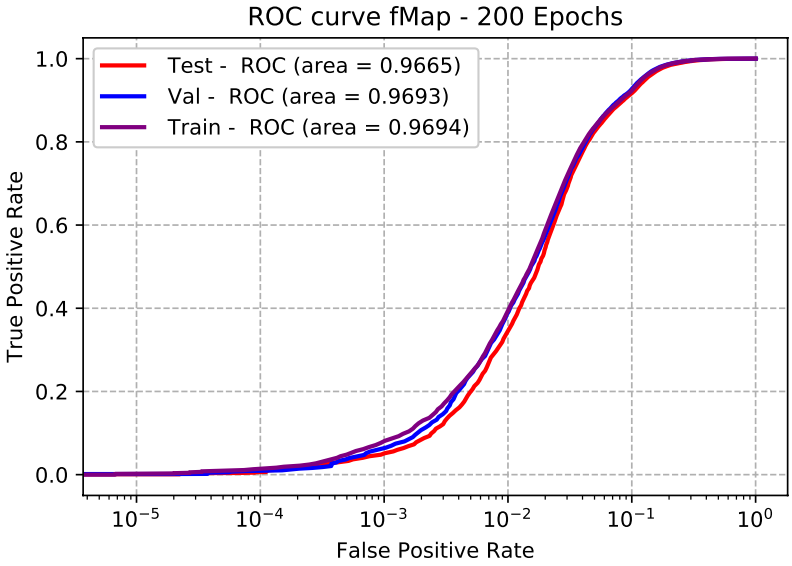

Figure 9: ROC curve for validation, testing and training dataset.

2 Schools, Conferences and Workshops

Here the learning activity concerning schools and workshops attended during the first year is reported.

International School Attended

- Third Machine Learning in High Energy Physics Summer School, Reading (UK), 16-23 Jul 2017;
- CMS Physics Object School (CMSPOS), Bari (Italy), 4-8 Sep 2017;

Participation to Workshops and Conferences

- 103° Congresso Nazionale della Società Italiana Di Fisica, Trento(Italy), 11-15 Sep 2017 - Oral presentation at SIF parallel session with title: "Recent CMS results in the search of new signals in quarkonium physics";

3 List of publications

- A.Pompili and A. Di Florio (on behalf of CMS collaboration), *"GPUs for statistical data analysis in HEP: a performance study of GooFit on GPUs vs. RooFit on CPUs"*, J.Phys.Conf.Ser. 762 (2016), 012044, Proceedings of "17th International workshop on Advanced Computing and Analysis Techniques (ACAT-2016)".

- A.Di Florio et al. *Statistical significance estimation of a signal within the GooFit framework on GPUs*", J.Phys.Conf.Ser. 762 (2016), 012044 Proceedings of "12th Conference on Quark Confinement and the Hadron Spectrum (Confinement XII)".

- A. Di Florio, *"Performance studies of GooFit on GPUs versus RooFit on CPUs while estimating the statistical significance of a new physical signal"*, Proceedings of "22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP-2016)". Already reviewed, going to print.

- A. Di Florio et al., *Convolutional Neural Network for Track Seed Filtering at the CMS High-Level Trigger*, Proceedings of "18th International workshop on Advanced Computing and Analysis Techniques in physics research (ACAT-2017)", in preparation.

- A. Di Florio et al., *Performance studies of GooFit on GPUs versus RooFit on CPUs while estimating the global statistical significance of a new physical signal*, Proceedings of "18th International workshop on Advanced Computing and Analysis Techniques in physics research (ACAT-2017)",in preparation.

- CMS Collaboration Author since 5 October 2017.

Adriano Di Florio

References

[1] CMS Collaboration., *Measurement of the $X(3872)$ production cross section via decays to $J/\psi\ \pi\pi$ in pp collisions at $\sqrt{s}$ = 7 TeV*, JHEP **1304** (2013) 154.

[2] CMS Collaboration, *Observation of a peaking structure in the J/psi phi mass spectrum from B(+/-) to J/psi phi K(+/-) decays*, Phys. Lett. B **734** (2014) 261.

[3] CMS Collaboration., *Search for a new bottomonium state decaying to $\Upsilon(1S)\pi^+\pi^-$ in pp collisions at $\sqrt{s}$ =8 TeV*, Phys. Lett. B **727** (2013) 57.

[4] CMS Collaboration, CMS PAS BPH-16-002. Paper draft in final reading stage.

[5] DØ Collaboration, *Inclusive production of the $X(4140)$ state in $p\overline{p}$ collisions at DØ*, Phys.Rev.Lett. **115** (2015) 23.

[6] The Particle Data Group, 2017 Review, http://pdg.lbl.gov .

[7] LHCb Collaboration, *Observation of $J/\psi\phi$ structures consistent with exotic states from amplitude analysis of $B^+ \to J/\psi\phi K^+$ decays*, Phys.Rev.Lett. **118** (2017), 022003.

[8] L. Maiani, A.D. Polosa, V. Riquer, *Interpretation of Axial Resonances in J/psi-phi at LHCb*, Phys. Rev. D **94**, 054026 (2016).