Searching for Hqq–>bbqq

P. Azzurri, J. Bernardini, D. Caiulo, R. Dell'Orso,
F. Fiori, F. Palla, P. Spagnolo
INFN Pisa

S. Alderweireldt, S. Bansal, X. Janssen, S. Luyckx, N. van Remortel
Universiteit Antwerpen

Konstantinos Kousouris
CERN

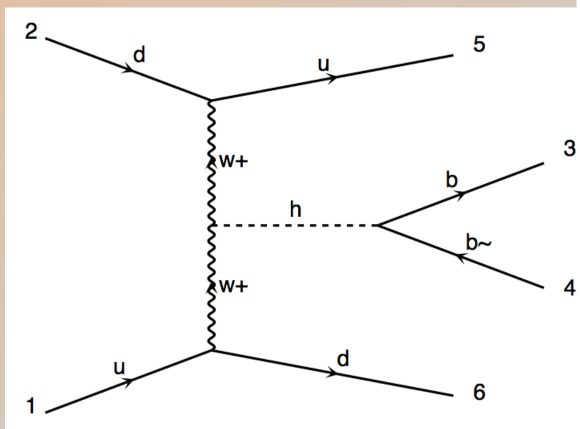
CMS meeting Pisa, 16/12/2011

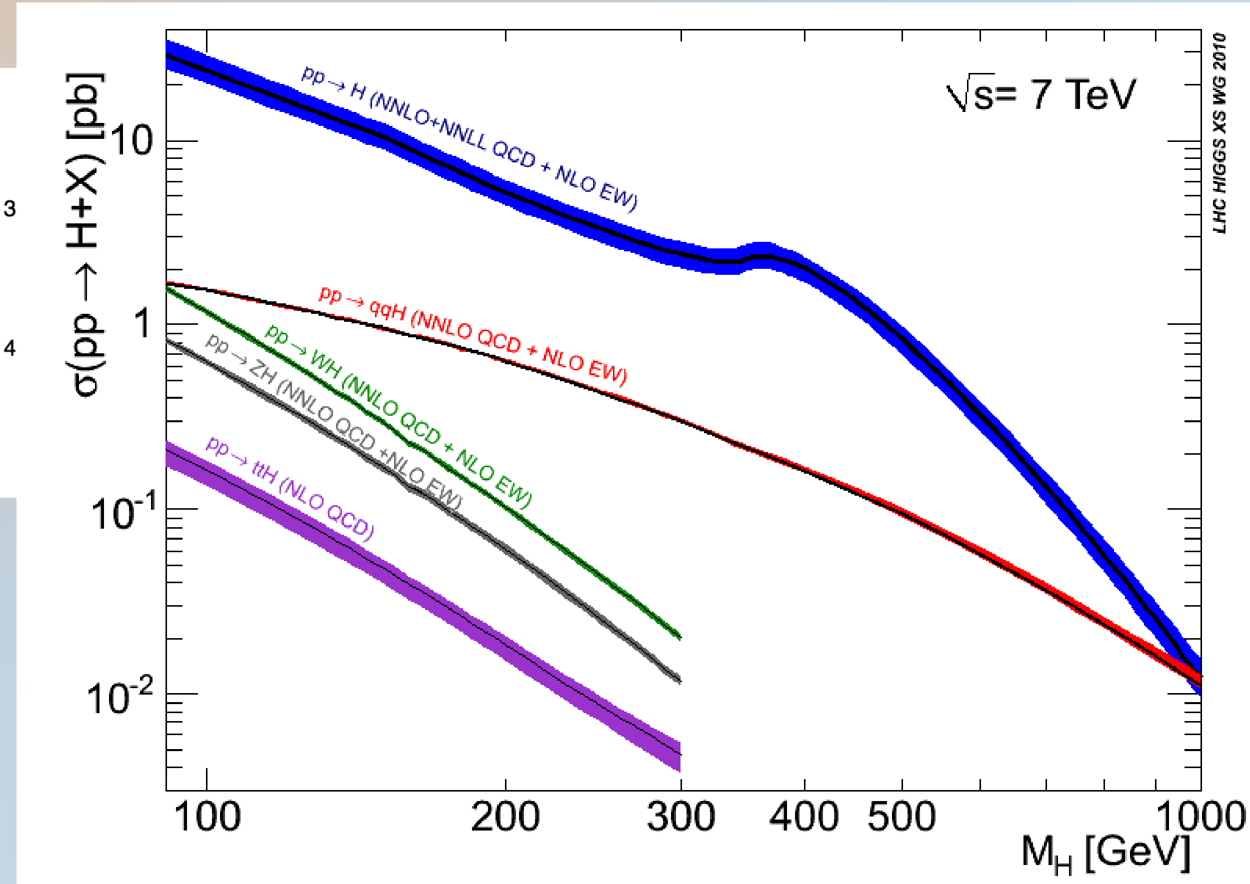
Outline

- Signal and Backgrounds
- Trigger strategy
- Discriminating variables
- MVA results
- A look at the 2011 data

The qqH->qqbb signal

VBF is the second most important Higgs production mechanism at the LHC





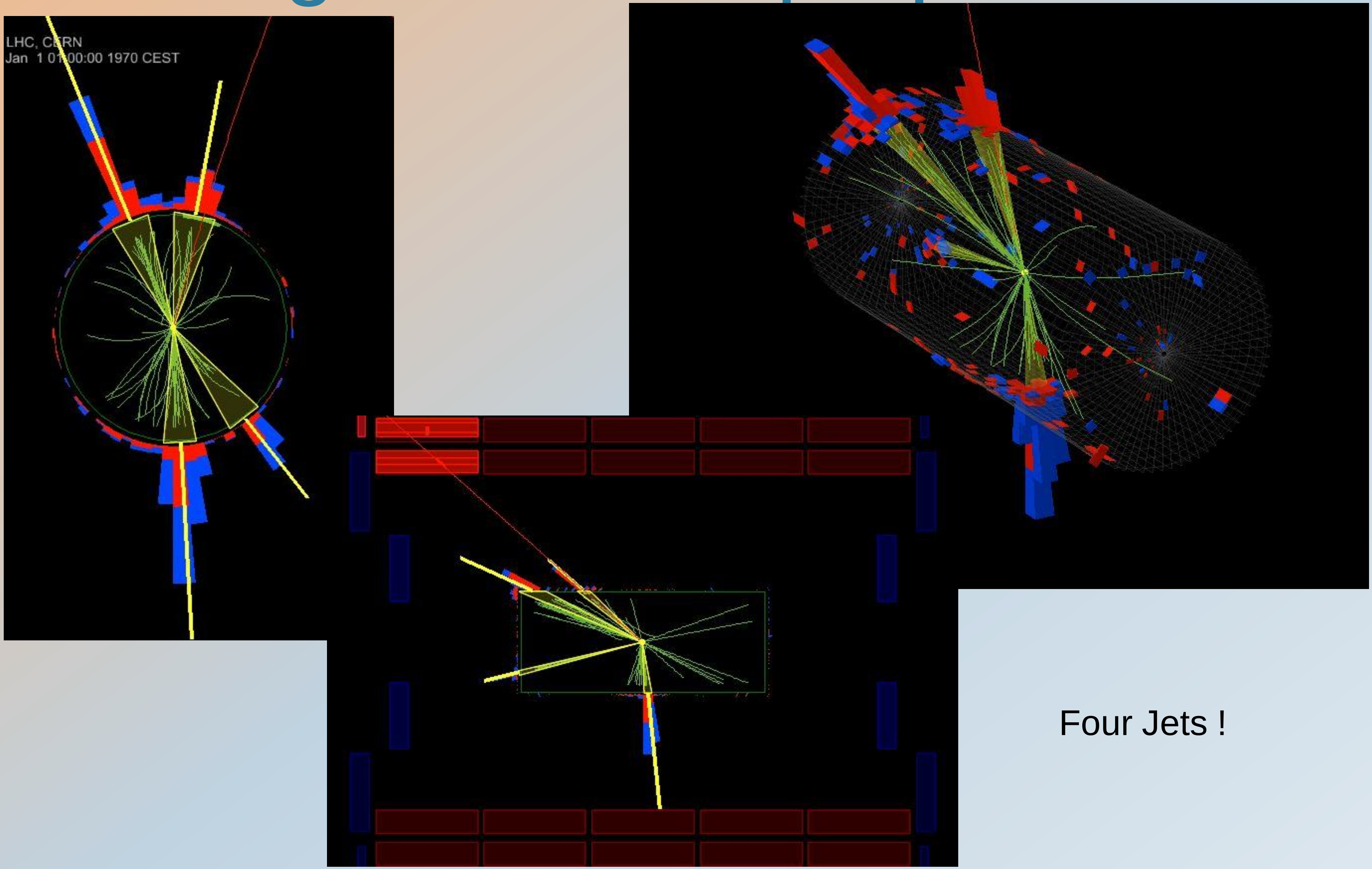
Generated with PYTHIA

m_H=120 GeV

BR(H−>bb)≈70%

at √s=7 TeV (123K)

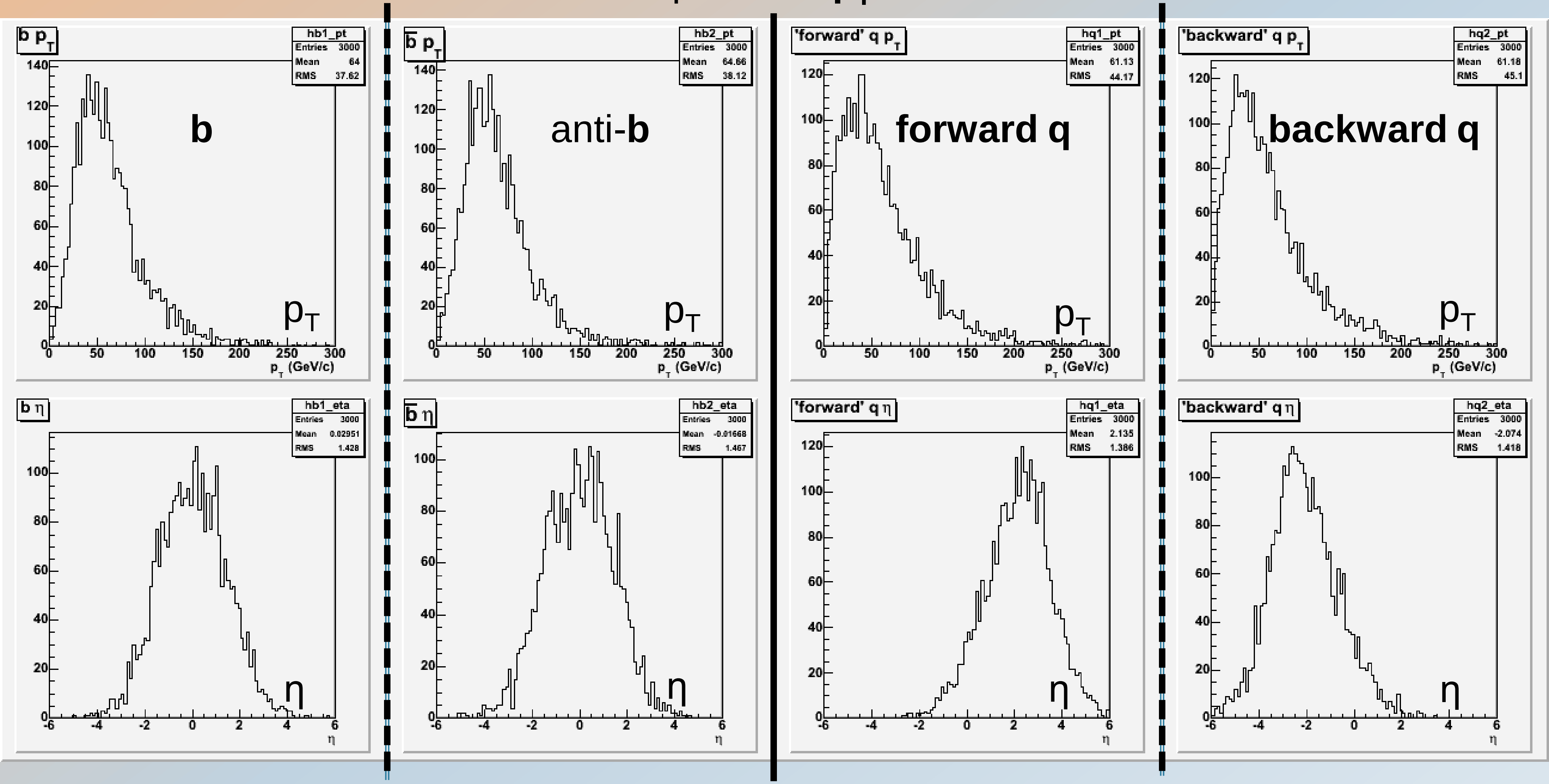
Signal events properties



Four Jets !

Signal four quarks: bbqq

For each parton $\langle p_T \rangle$ **≈62 GeV**



$\langle \Delta\eta_b \rangle$ ≈1.4 : b-jets quite well contained in the tracker acceptance

$\langle \eta_q \rangle$ ≈±2.1

Signal four quarks: bbqq

p_T-ordered parton spectra : (1=hardest … 4=softest)

$\langle p_{T1} \rangle$ =98 GeV **$\langle p_{T2} \rangle$ =77 GeV** **$\langle p_{T3} \rangle$ =48 GeV** **$\langle p_{T4} \rangle$ =28 GeV**

the fourth (softest) parton is soft !

η-ordered parton distributions : (1=most central … 4=most forward)

$\langle |\eta_1| \rangle$ =0.5 **$\langle |\eta_2| \rangle$ =1.2** **$\langle |\eta_3| \rangle$ =2.0** **$\langle |\eta_4| \rangle$ =3.1**

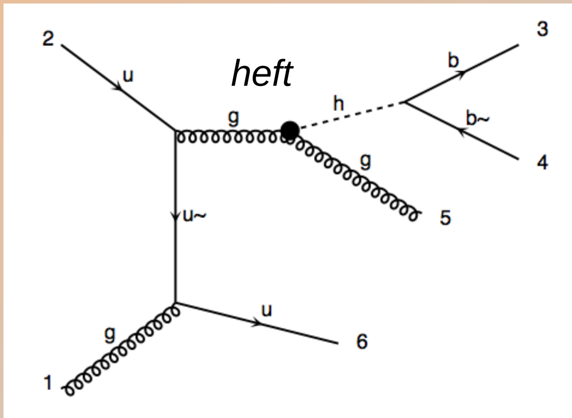
In pairs, the two most central partons can be:

- the two b-quarks (45%)
- one b-quark and one associated-quark (54%)
- the two associated-quarks (1%)

total four-quark system : Total p_T ~ 7 GeV. Total p_L ~ 600 GeV. Total invariant mass ~ 700 GeV. Total H_T ~ 250 GeV

Collateral Signal

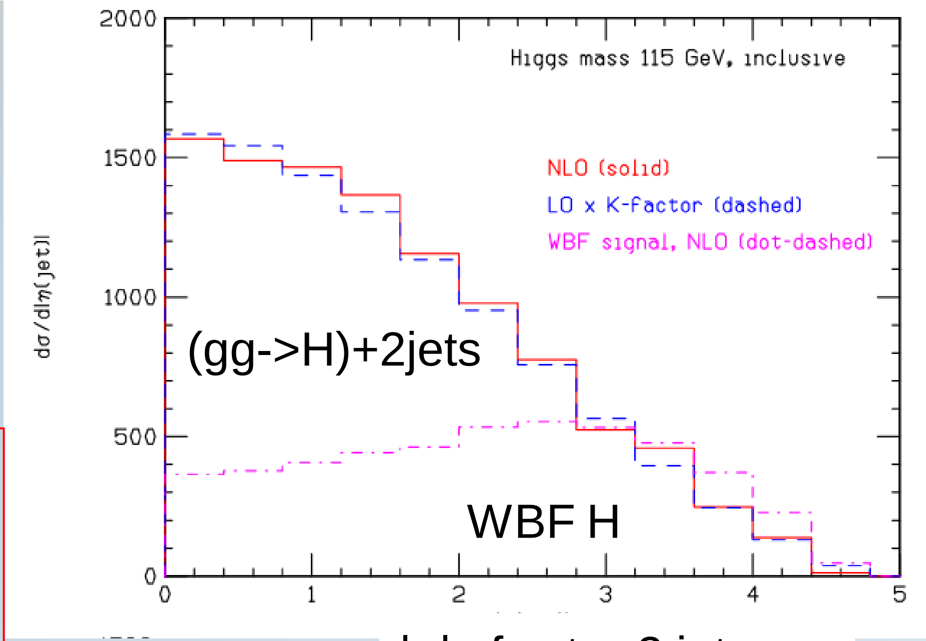
Gluon Fusion Higgs (bb) + jets



Campbell, Ellis, Zanderighi *Next-to-leading order Higgs + 2 jet production via gluon fusion* http://arxiv.org/pdf/hep-ph/0608194v2

Higgs mass	115 GeV	115 GeV
σ_{LO} [pb]	3.50	271 fb
σ_{NLO} [pb]	4.03	346 ± 5
σ_{WBF} [pb]	1.77	911

$$p_t(\text{jet}) > 40\ \text{GeV}, \quad |\eta_{\text{jet}}| < 4.5, \quad R_{\text{jet,jet}} > 0.8$$



|η| of extra 2 jets

Prepared a MadGraph gridpack for a LO sample production of H+jets (up to 3)

… waiting in the GEN pipeline …

Backgrounds

Summer11

	#events	sigma(pb)	lumi(fb-1)
/QCD_TuneZ2_HT-100To250_7TeV-madgraph	14437469	4194000.0	**0.00344**
/QCD_TuneZ2_HT-250To500_7TeV-madgraph	20674219	198500.0	0.1042
/QCD_TuneZ2_HT-500To1000_7TeV-madgraph	14437469	5856.0	2.465
/QCD_TuneZ2_HT-1000_7TeV-madgraph	6294851	122.6	51.597
/WW_TuneZ2_7TeV_pythia6_tauola	4225916	27.83	151.8
/WZ_TuneZ2_7TeV_pythia6_tauola	4265243	10.47	407.4
/ZZ_TuneZ2_7TeV_pythia6_tauola	4187885	4.287	976.9
/TTJets_TuneZ2_7TeV-madgraph-tauola	3701947	94.76	39.066
/ZJetsToQQ_HT-100_7TeV-madgraph	7647683	3488	2.192

(dedicated production)

requiring 4 Jets p_T>10GeV σ(QCD)~10μb ~10^7 σ(VBF H)

Trigger strategy : L1

L1TripleJet_X_Y_Z_VBF = (3 central >= X, Y, Z) OR (2 central >= Y, Z and 1 Fwd >= X) OR (2 central >= X, Z and 1 Fwd >= Y)*

- L1TripleJet_64_44_24_VBF as main
- L1TripleJet_64_48_28_VBF as backup

*No option for Z to be forward : Stay save from increased PU in the forward region for a low Z cut value

	5 / 7 *10^{33} (KHz)	Efficinecy	Pure (KHz)
L1_TripleJet_64_44_24_VBF	10.94 / 0	62 %	1.1
L1_TripleJet_64_48_28_VBF	7.44 / 0	56 %	
L1_TripleJet_64_48_32_VBF	5.17 / 8.34	50 %	1.8

https://twiki.cern.ch/twiki/pub/CMS/L1Menu2012WorkingPage/Draft3_7e33_NewNaming.pdf

Trigger strategy : HLT

- 1 track pT   > 80
- 2 tracks pT > 58
- 3 tracks pT > 45
- 4 tracks pT > 20

- Sort the 4 jets with the highest pT on η. Take the jet with the lowest and highest η and abel them as « qq »
  - Mqq > 200
  - |Δηqq | > 2.5

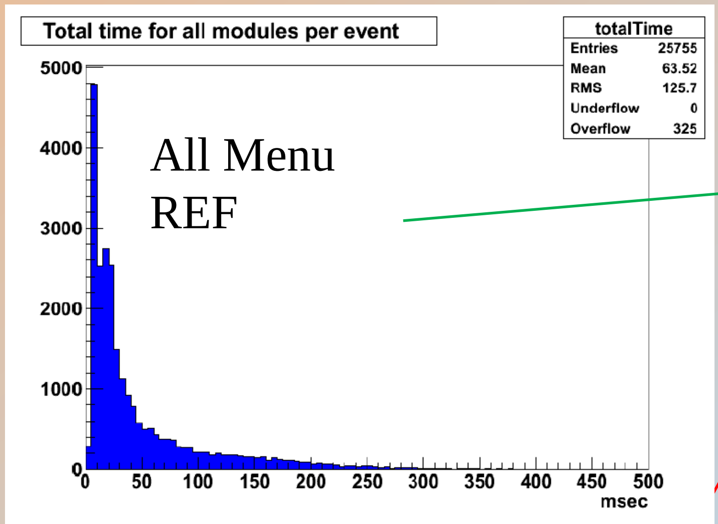
- 1 jet with BtagL25 > 2.5

- 1 jet with BtagL3   > 7.5

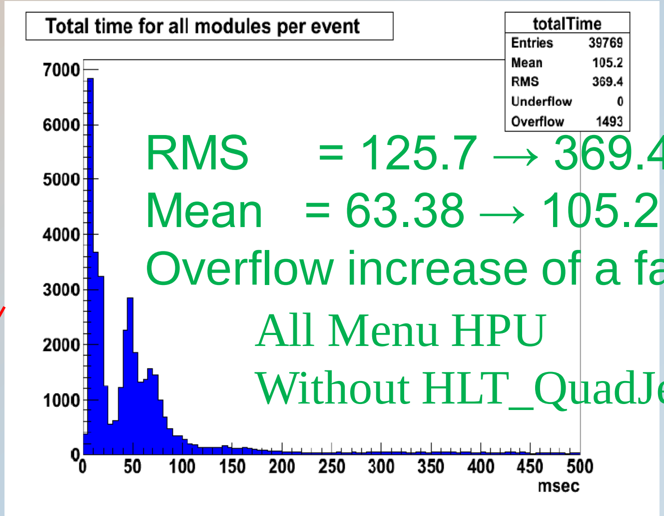
- Sort the 4 jets on Btag. Label the 2 lowest btagged jets « qq »
  - Mqq > 200
  - | Δηqq | > 2.5

pT cuts (GeV)	**[80 58 45 20]**	**[80 68 45 20]**	**[80 72 48 24]**	
Rate (Hz)	**3.34 +- 0.44**	**2.24 +- 0.36**	**1.69 +- 0.32**	**@ $5*10^{33}$**
Eff (%)	**8.05 +- 0.3**	**7.26 +- 0.26**	**6.1 +- 0.24**	

Trigger strategy : CPU Time

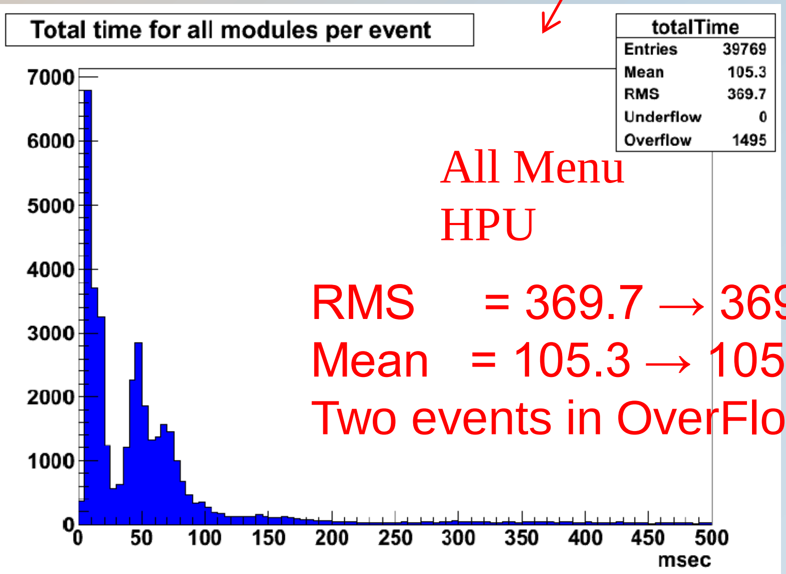


All Menu
REF



RMS = 125.7 → 369.4 ms
Mean = 63.38 → 105.2 ms
Overflow increase of a factor 3

All Menu HPU
Without HLT_QuadJet_BTag



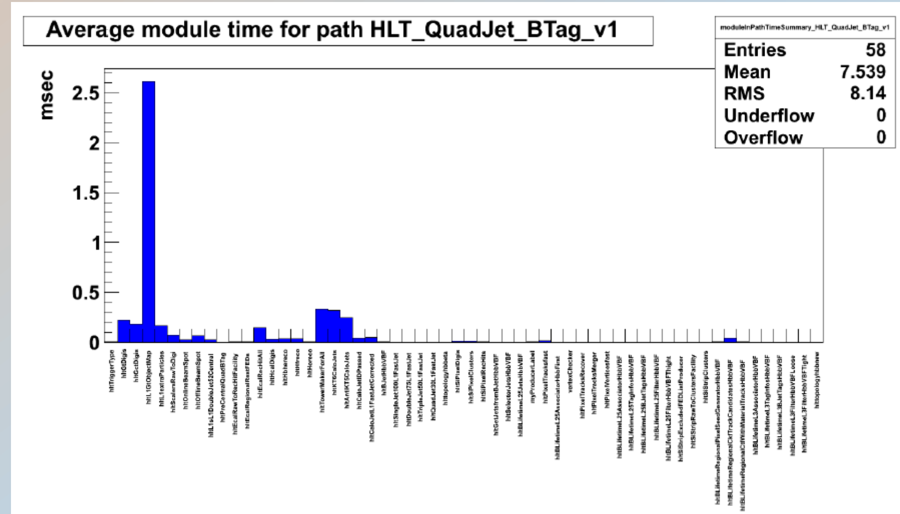
All Menu
HPU

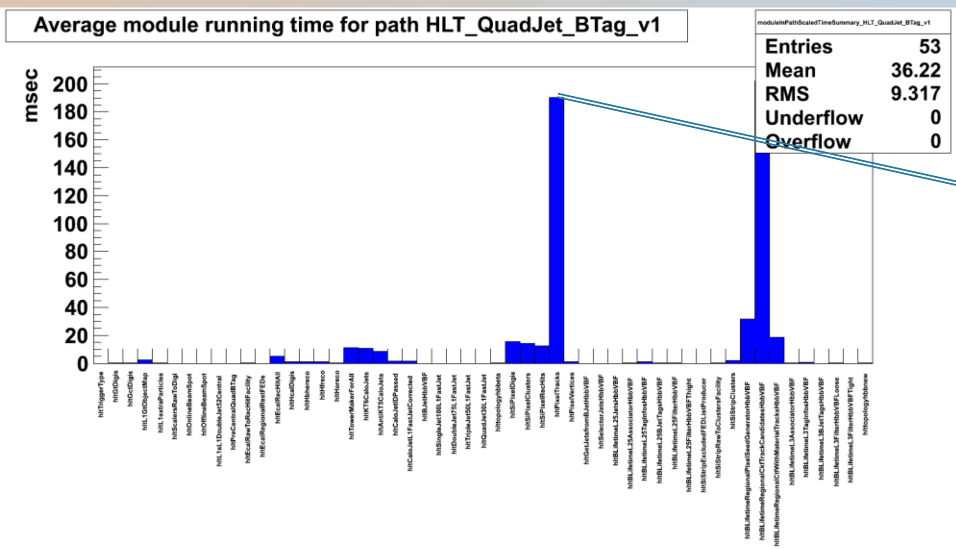
RMS = 369.7 → 369.4 ms
Mean = 105.3 → 105.2 ms
Two events in OverFlow

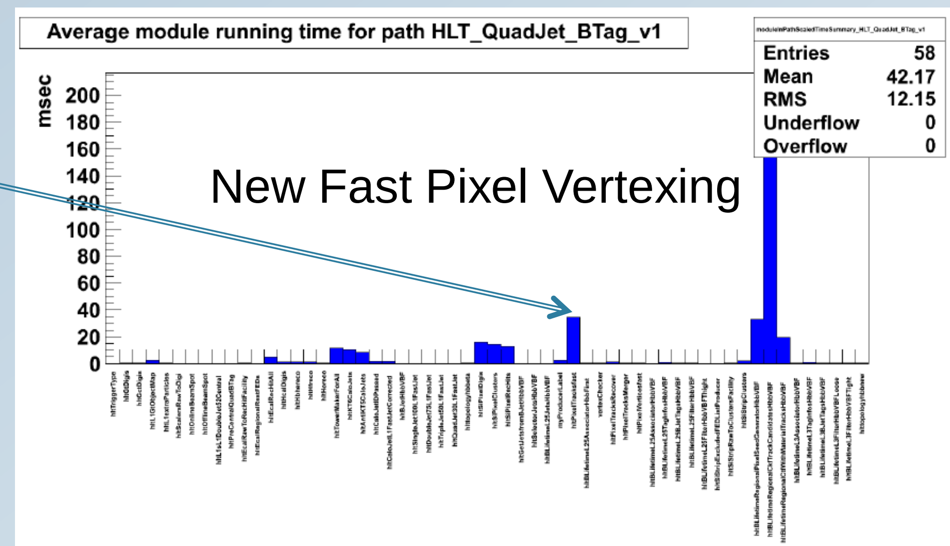
@ 6.3 10 [33]

Trigger strategy : CPU Time

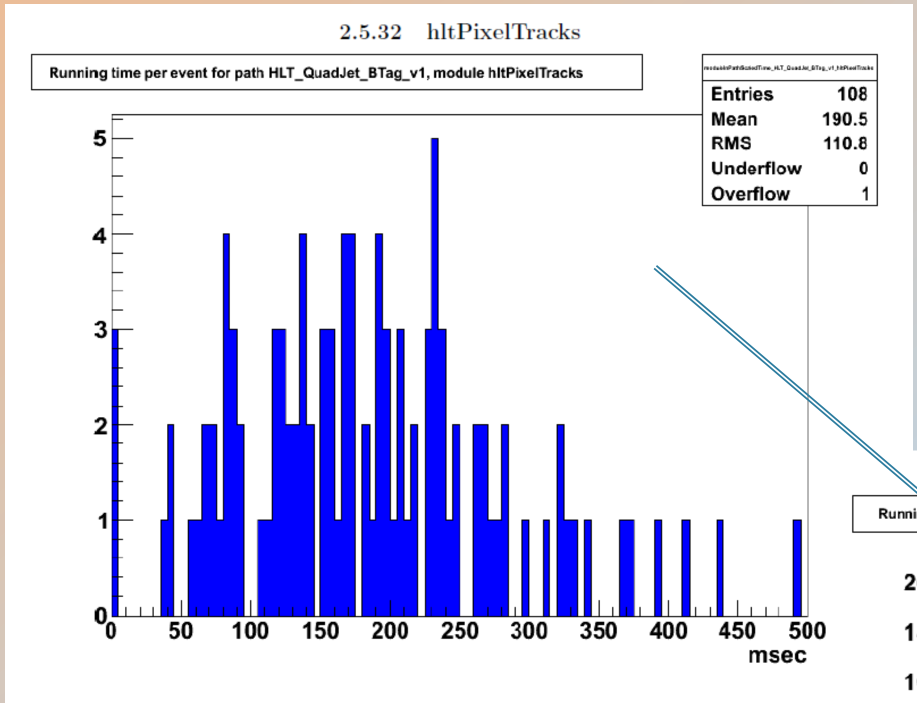
6.1517 (path HLT_DiCentralJet20_BTagIP_MET65_v11)
3.3988 (path HLT_DiCentralJet36_BTagIP3DLoose_v5)
4.7298 (path HLT_QuadJet_BTag_v1)
23.8505 (path HLT_CentralJet46_CentralJet38_DiBTagIP3D_v7)
18.0986 (path HLT_CentralJet60_CentralJet53_DiBTagIP3D_v6)
3.3998 (path HLT_QuadJet40_v11)
3.6878 (path HLT_QuadJet45_DiJet40_v3)
9.0117 (path HLT_QuadJet50_DiJet40_v5)
11.2076 (path HLT_QuadJet50_DiJet40_L1FastJet_v2)
4.1720 (path HLT_QuadJet40_IsoPFTau40_v18)
13.617 (path HLT_QuadJet45_IsoPFTau45_v13)



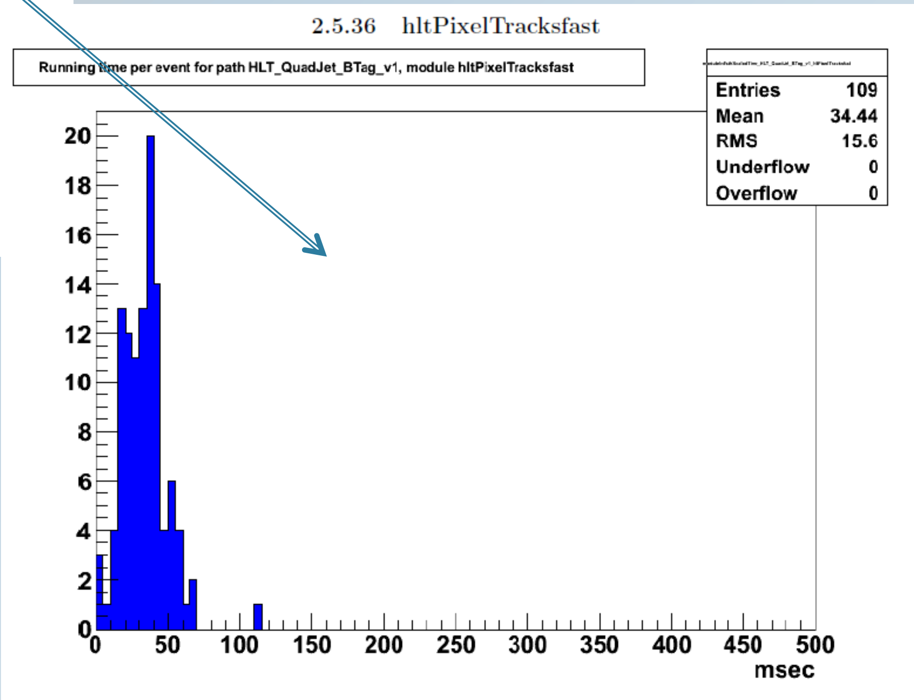




Trigger strategy : CPU Time



New Fast Pixel Vertexing



Offline analysis

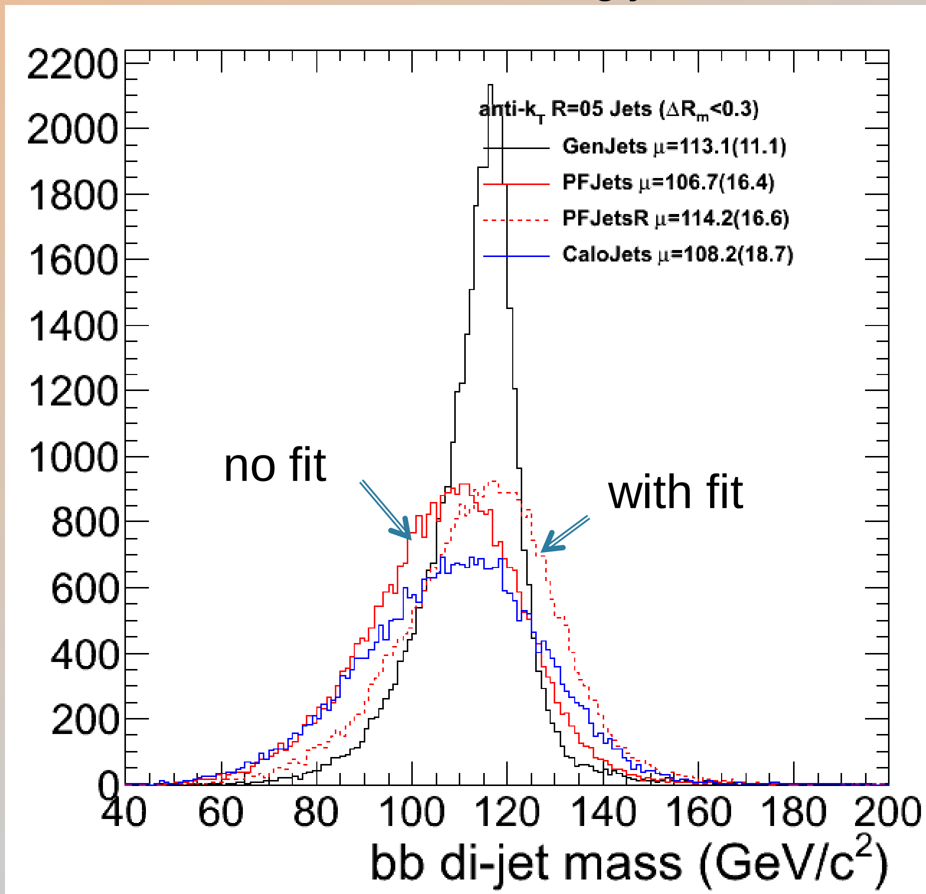
After the trigger selections we expect (events in 10/fb)

QCD-100_250	54 pb	(540 000)
QCD-250_500	782 pb	(7 820 000)
QCD-500_1000	91 pb	(910 000)
QCD-1000	2.3pb	(23 000)
tt	2.8pb	(28 000)
Zqq	2.4 pb	(24 000)
WW	0.012pb	(120)
WZ	0.010pb	(100)
ZZ	0.006pb	(60)
signal	0.13 pb	(1 300)

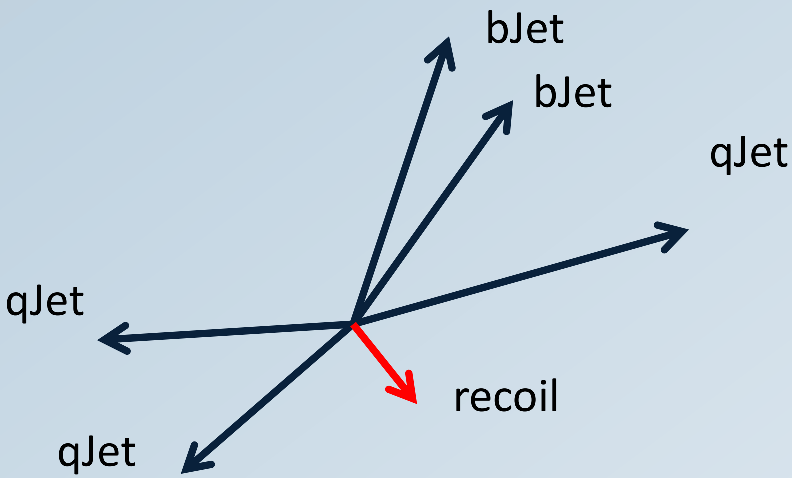
the total background is 934pb (9M events)
we still have S/B around 10^{-4}.

Kinematic 2C fit

Rescale the 5 leading jets in the event to impose transverse balance with the recoil system (pT=0)



Use lagrange multipliers to minimize a chi-square based on the jet response and resolution functions measured on signal MC



Very small improvement in M(bb): **+6% in μ/σ**
(reach +10% cutting on the fit χ^2 with small loss of signal events)

Jet axes in the (η,φ) plane

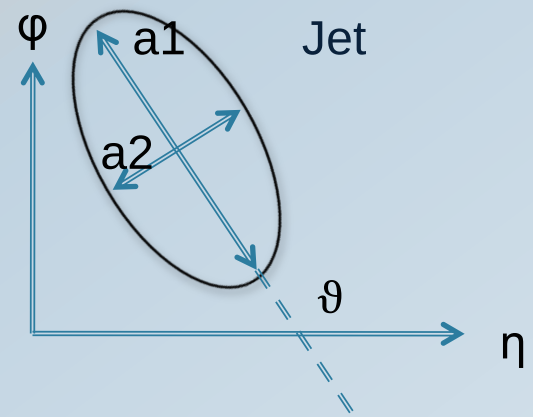
Measure the jet width to discriminate QCD gluon jets from signal light quarks

$$M_{11} = \sum w_i \Delta\eta_i^2 = a$$

$$M_{22} = \sum w_i \Delta\phi_i^2 = b$$

$$M_{12} = M_{21} = -\sum w_i \Delta\phi_i \Delta\eta_i = c$$

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$



Eigenvectors: $$\lambda_{1,2} = \frac{a + b \pm \sqrt{(a - b)^2 + 4c^2}}{2}$$

Major axis $$a_1 = \left(\lambda_1 / \sum w_i\right)^{\frac{1}{2}}$$

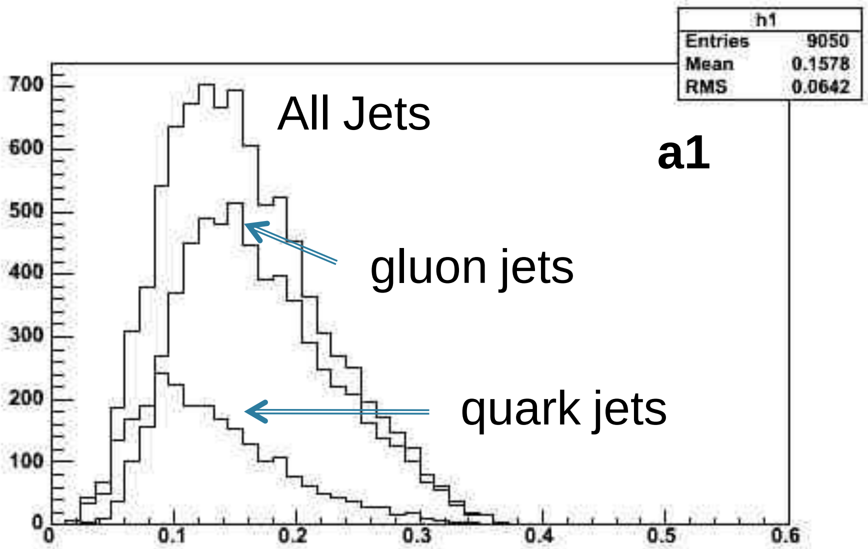
Minor axis $$a_2 = \left(\lambda_2 / \sum w_i\right)^{\frac{1}{2}}$$

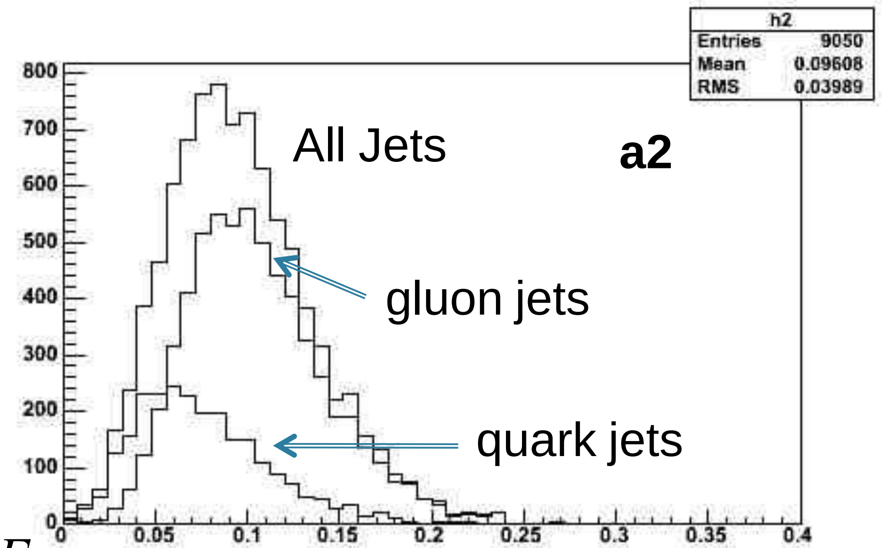
Minor/Major axis rotation angle in the (η,φ) plane:

$$\tan\theta = \frac{\lambda - a}{c} = \frac{c}{\lambda - b}$$

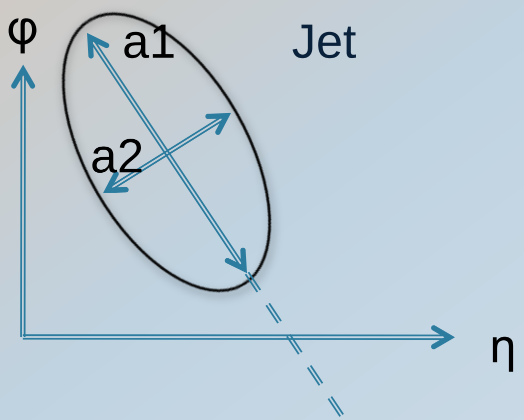
Jet axes in the (η,φ) plane

MC di-jets pT=80-120 GeV





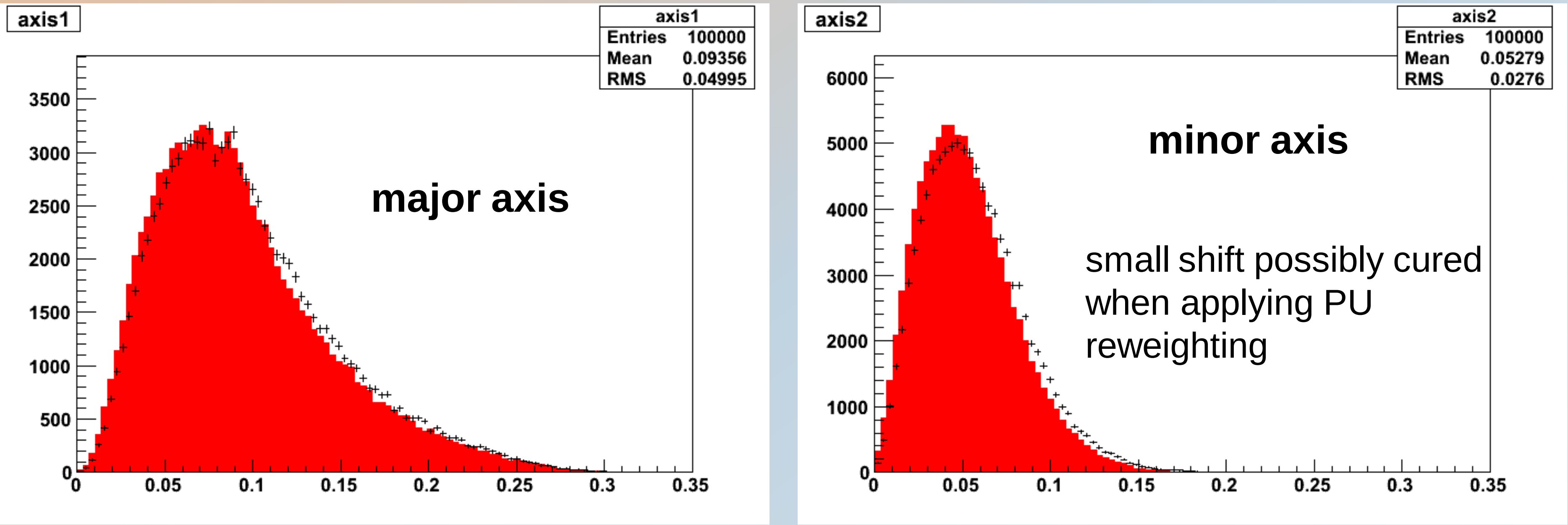
$w_i = E_{Ti}$



optimal weight to assign to jet components is found to be $\boldsymbol{w_i=E_{Ti}^2}$

Jet axes validation

Selected **di-jet** events with pT=50-80 GeV



MC (di-jets) / Data (2011A Data DiJetAve30)

good agreement also in η bins

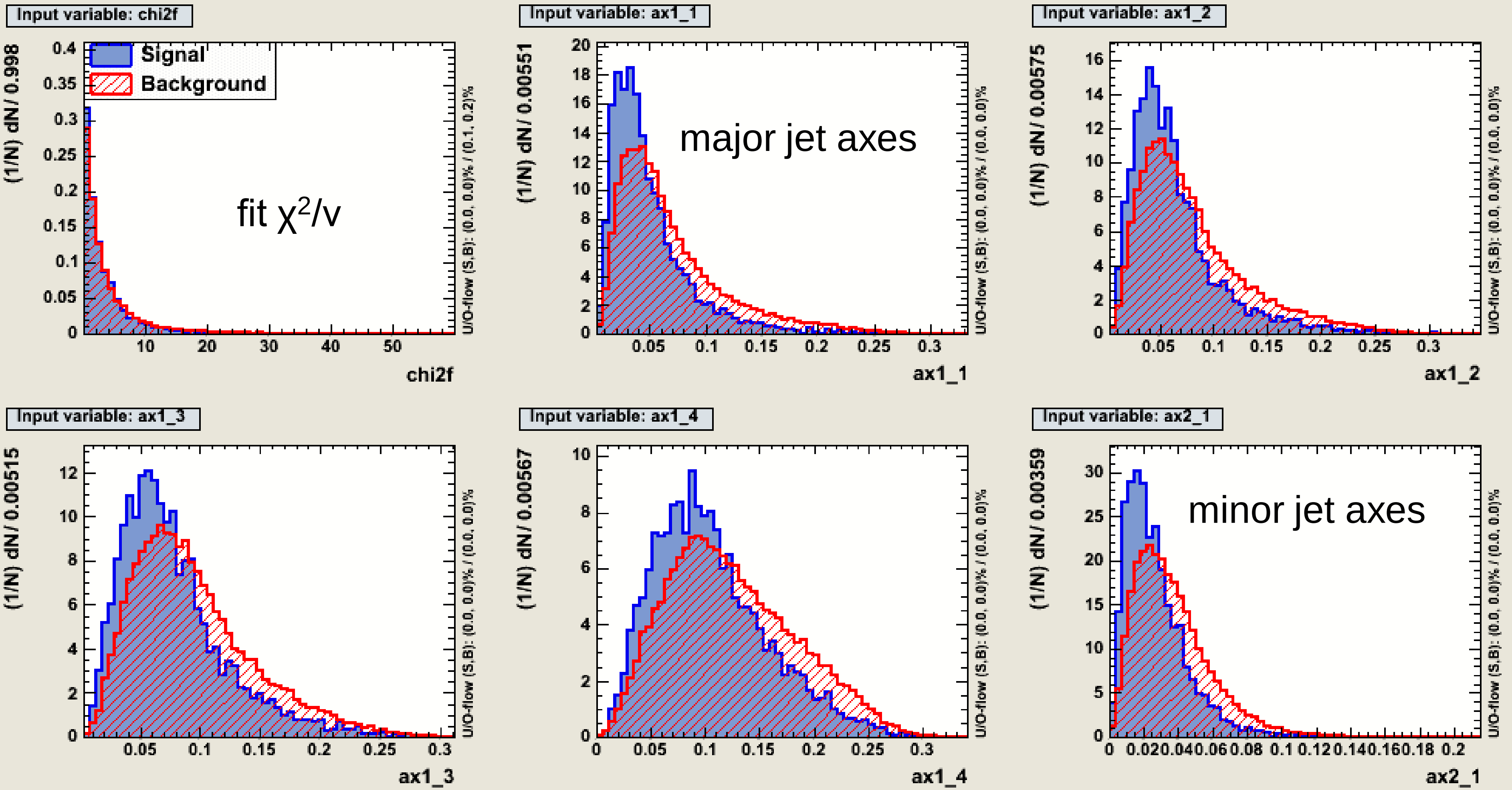
Discriminating variables

- p_{T1} p_{T2} p_{T3} p_{T4} of the four leading ak5 PF jets (fully corrected)
- p_{T5} of the fifth ak5 PF jet
- minor/major axis of the 4 leading jets in the η-ϕ plane
- largest two btag values (JetBProbability on ak5PFBJets)
- max $\Delta\eta$ between all pairs among the leading 4-jets
- $\Delta\eta$ between two most b-tagged (bb) and least b-tagged (qq) 4-jets
- invariant mass of the least b-tagged (qq) 4-jets
- χ^2 of event kinematical fit
- Total p_T of charged tracks (>300MeV/c) associated to the hardest PV and not to the 4 leading PF Jets.

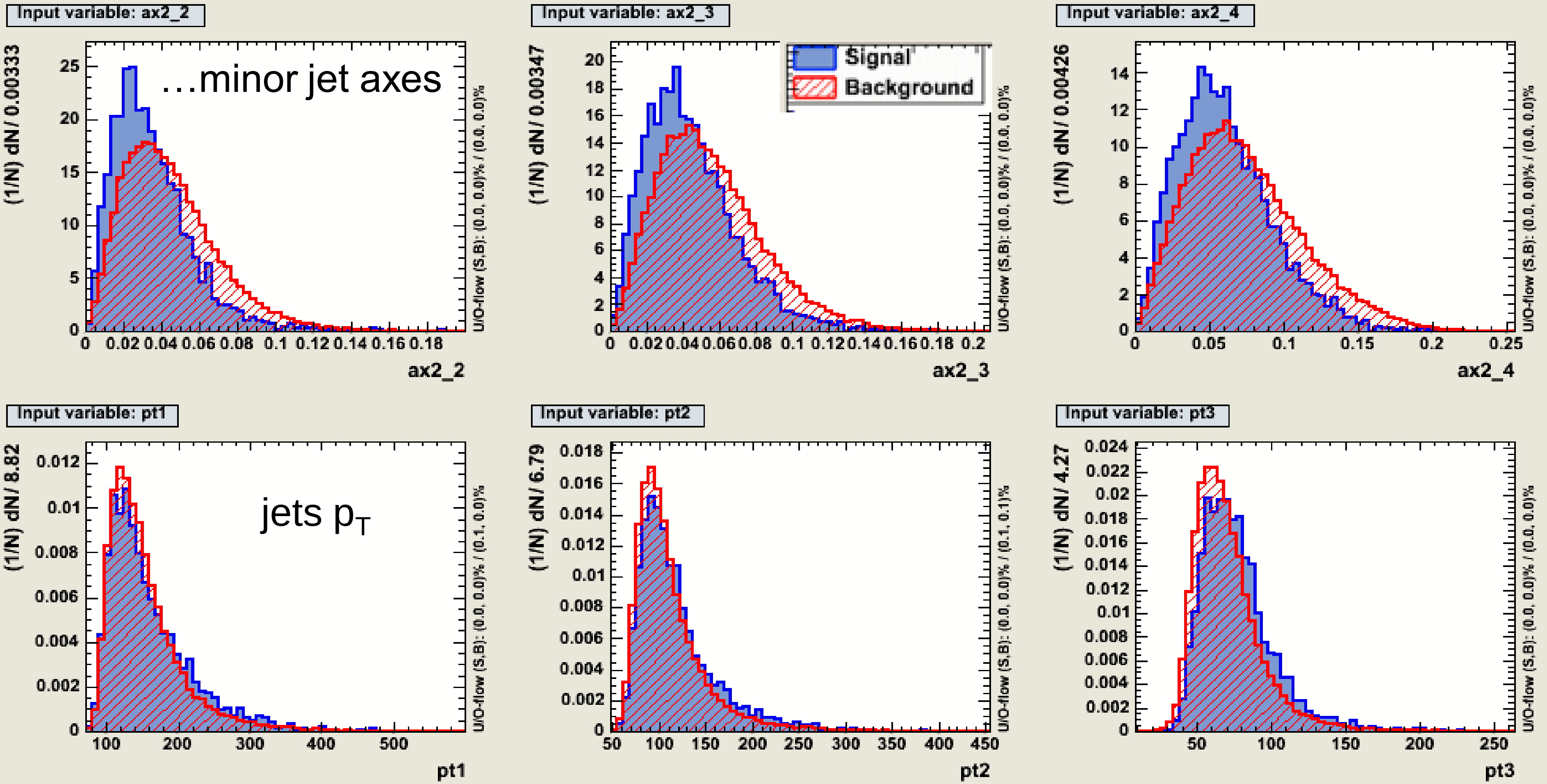
Fed to a Multi Variate Analysis with the TMVA Root package

- trained only against QCD background
- tried both Neural Networks (MLP) and Boosted Decision Trees
- results shown are rescaled to expected ones with 10/fb

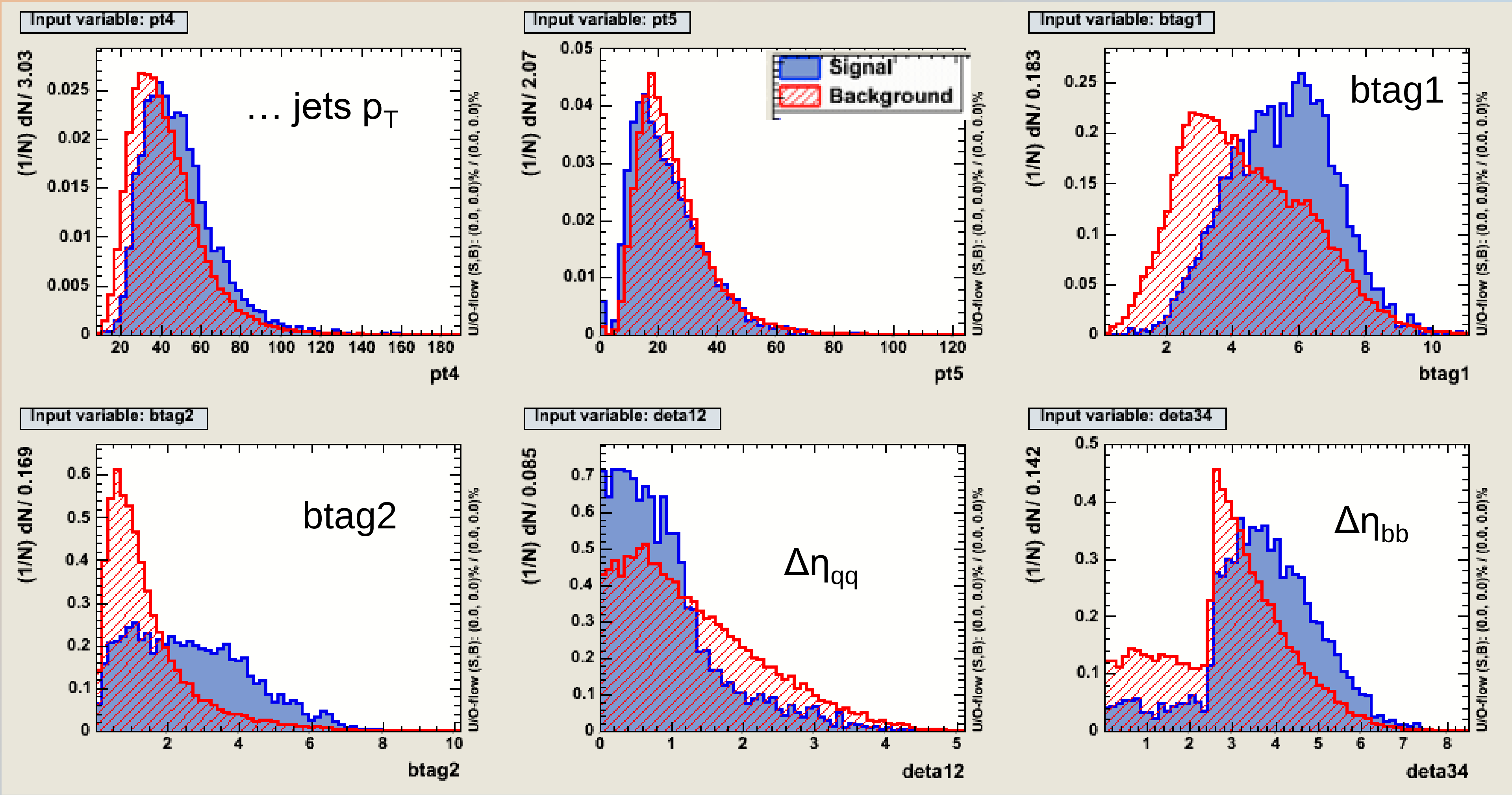
Variables in TMVA



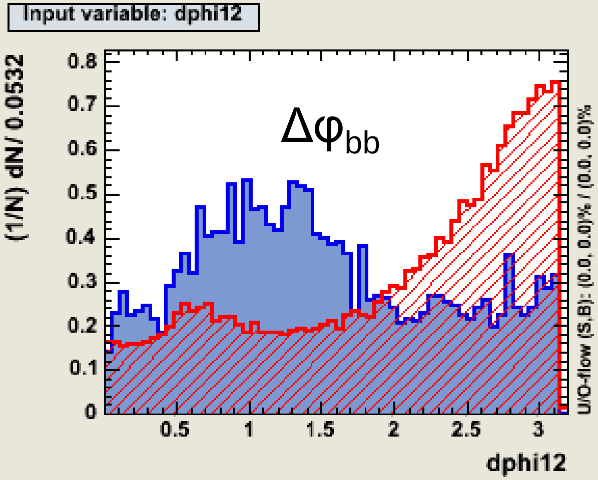
Variables in TMVA

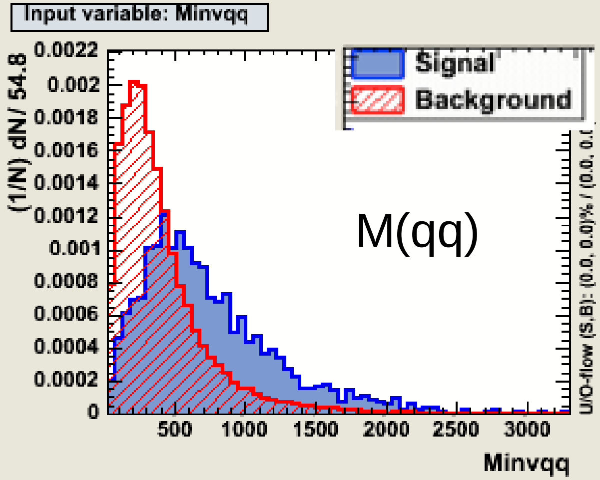


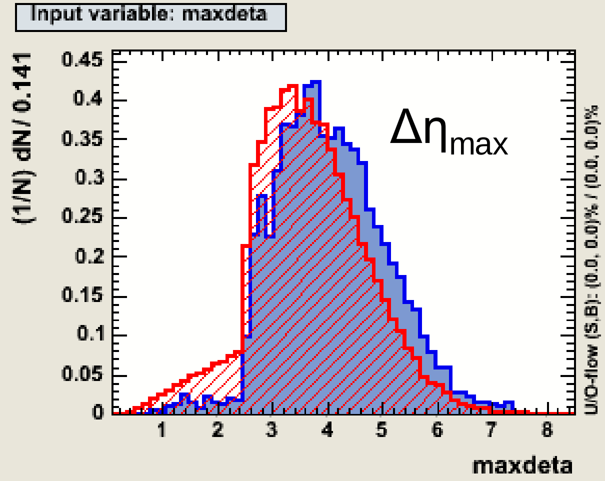
Variables in TMVA

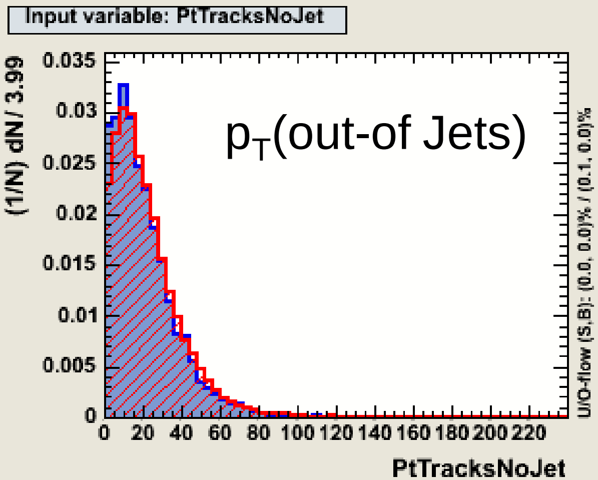


Variables in TMVA

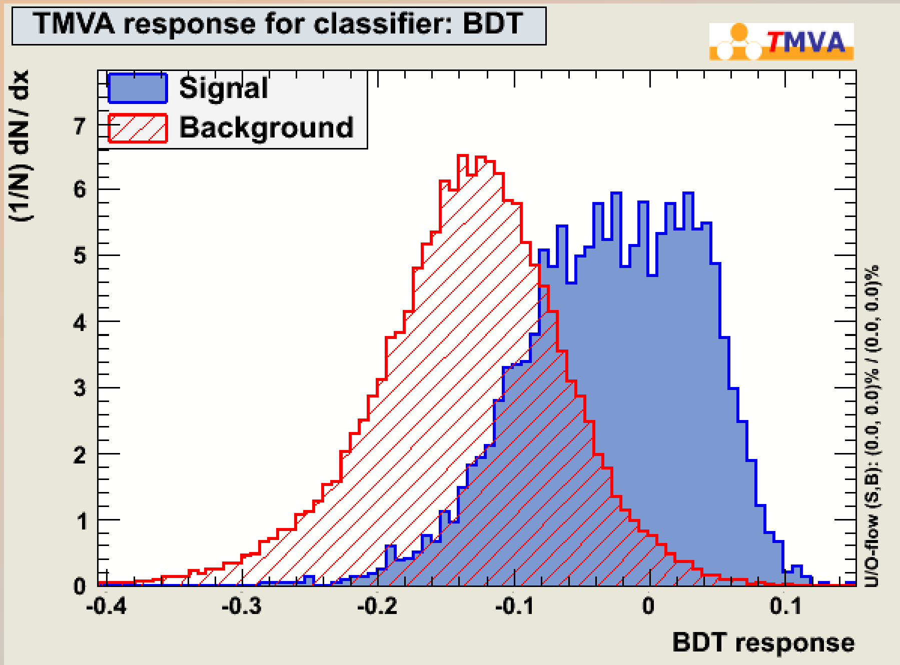


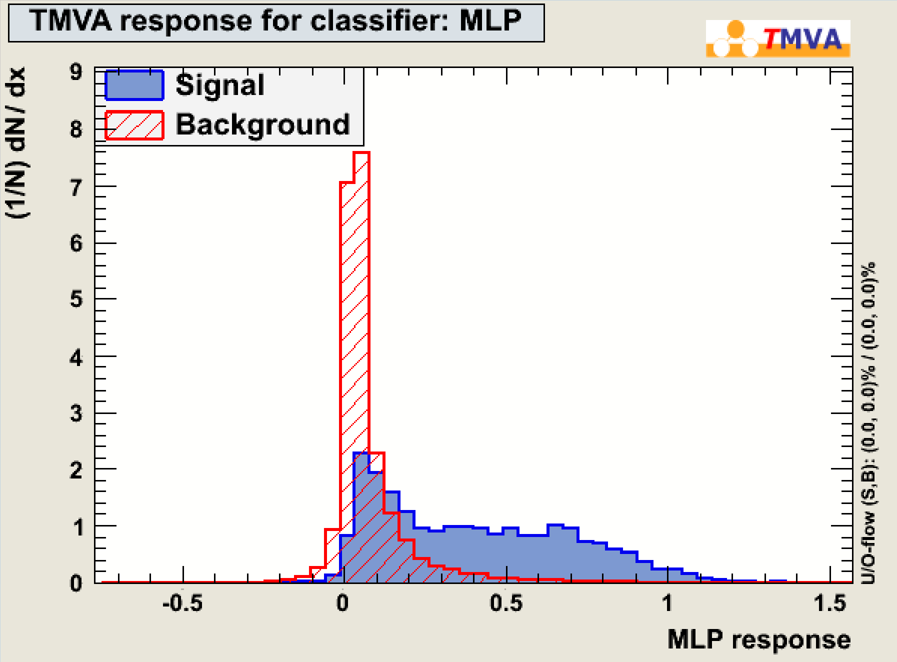






MVA outputs

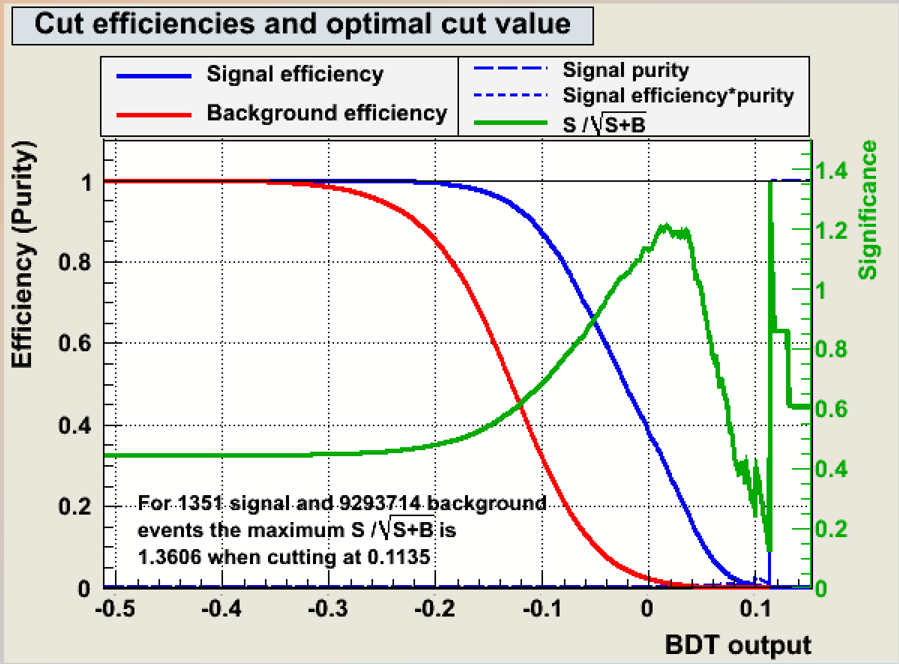


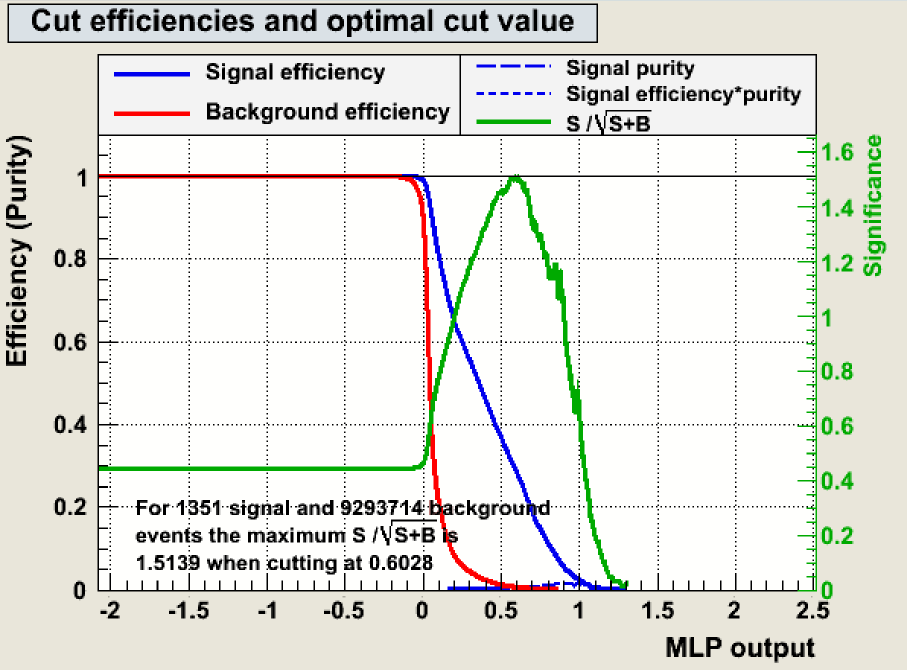


BDT has a overall worse separation power wrt MLP
(for configurations tried up to now)

MVA performance

- N BKG in 10 fb^{-1} ~ 9.3M   N Signal in 10 fb^{-1}~ 1.3 K
  - 9.232.714 from QCD
  - 70.680 from ttbar, Zqq, ZZ, WW, WZ





- MLP cut at 0.6 yields a 1.5σ Significance
- BDT cut at 0.011 yields a 1.2σ Significance

Selected events

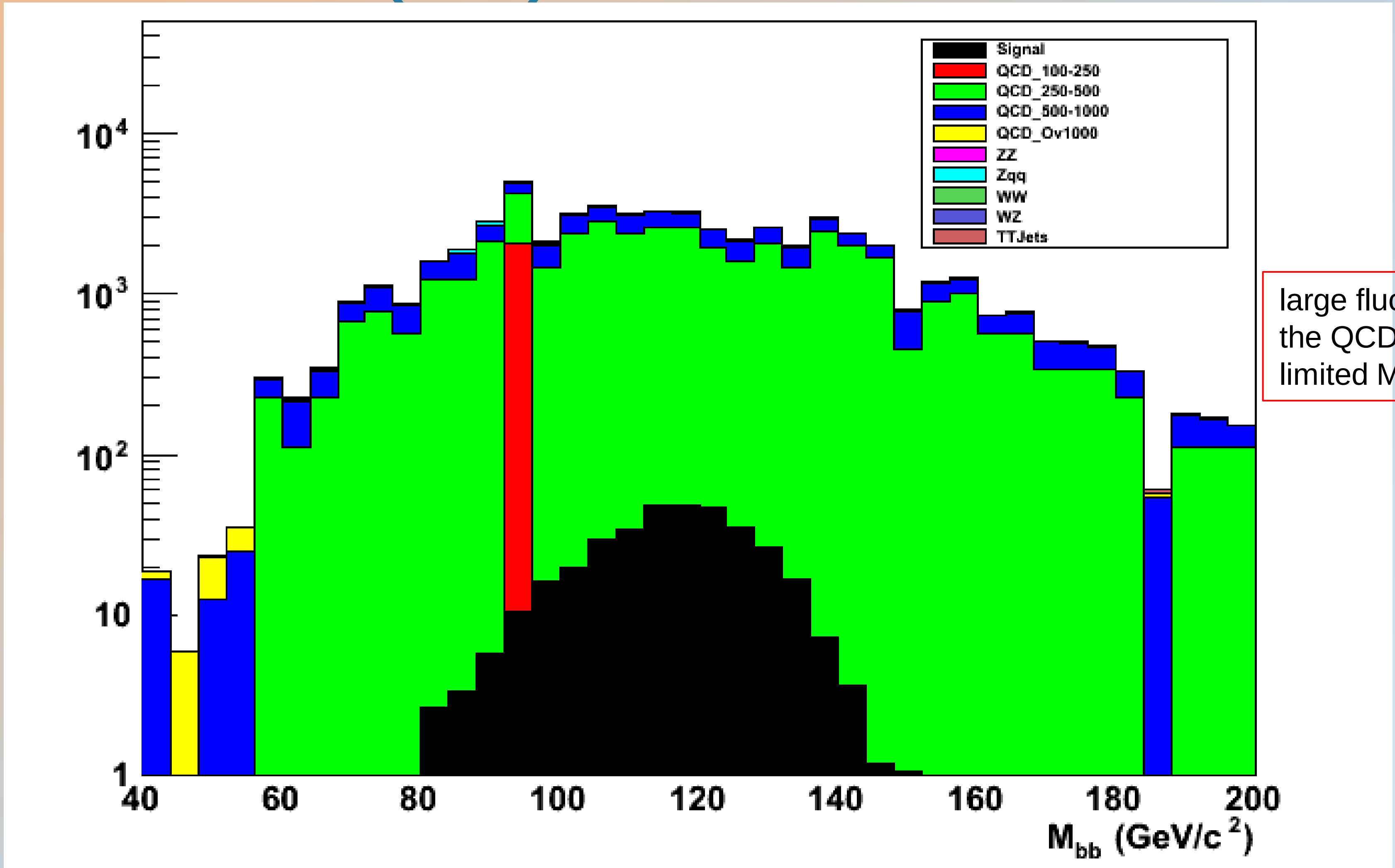
After the MVA cut (MLP>0.6) we expect (events in 10/fb)

QCD-100_250	0.21 pb	(2 100)
QCD-250_500	3.91 pb	(39 100)
QCD-500_1000	1.17 pb	(11 700)
QCD-1000	0.01pb	(100)
ttbar	0.031 pb	(310)
Zqq	0.057 pb	(570)
WW	0.00005pb	(0.5)
WZ	0.00010pb	(1)
ZZ	0.00008pb	(1)
signal	0.038 pb	(380)

signal (380), Zqq (570) and ttbar (310) floating on 50k QCD events
now S/B is around 10^{-2}.

…but didn't use the bb mass yet

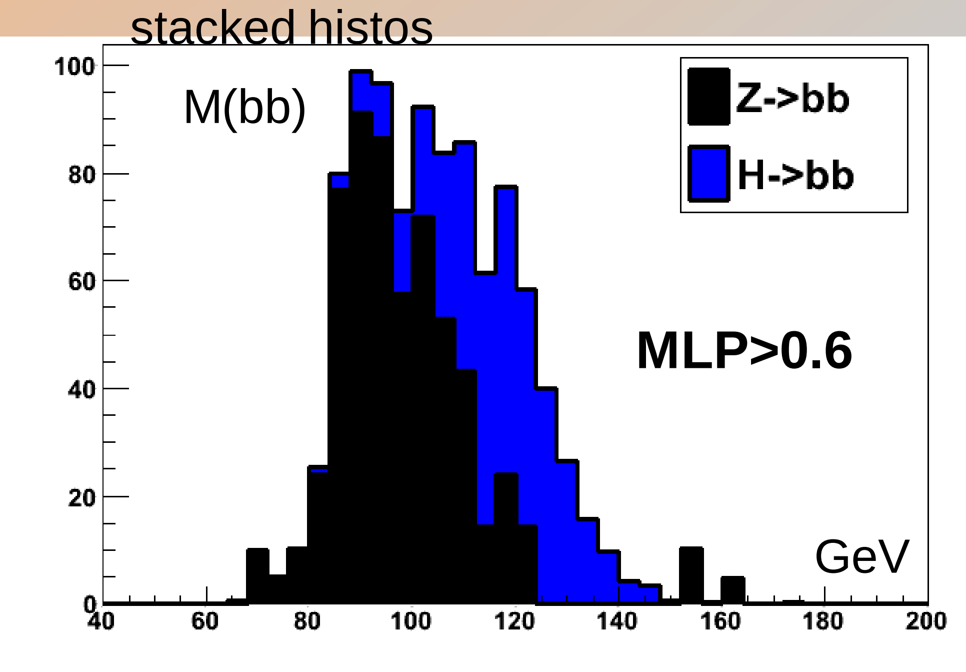
M(bb) after selection

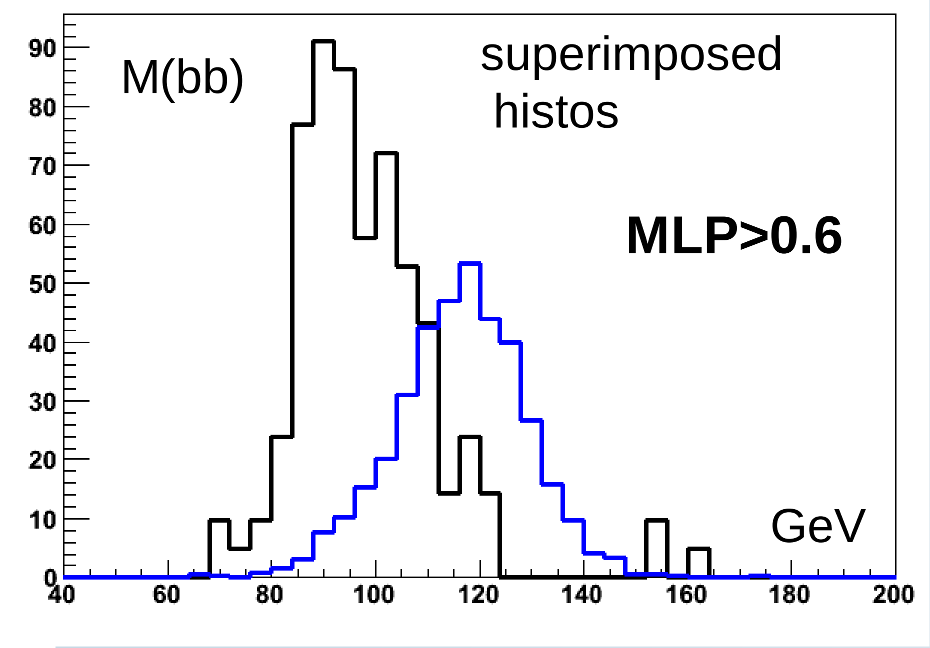


large fluctuations in the QCD due to limited MC stats

Signal significance MLP>0.6: 1.51σ
1.85σ in the [100-140] interval

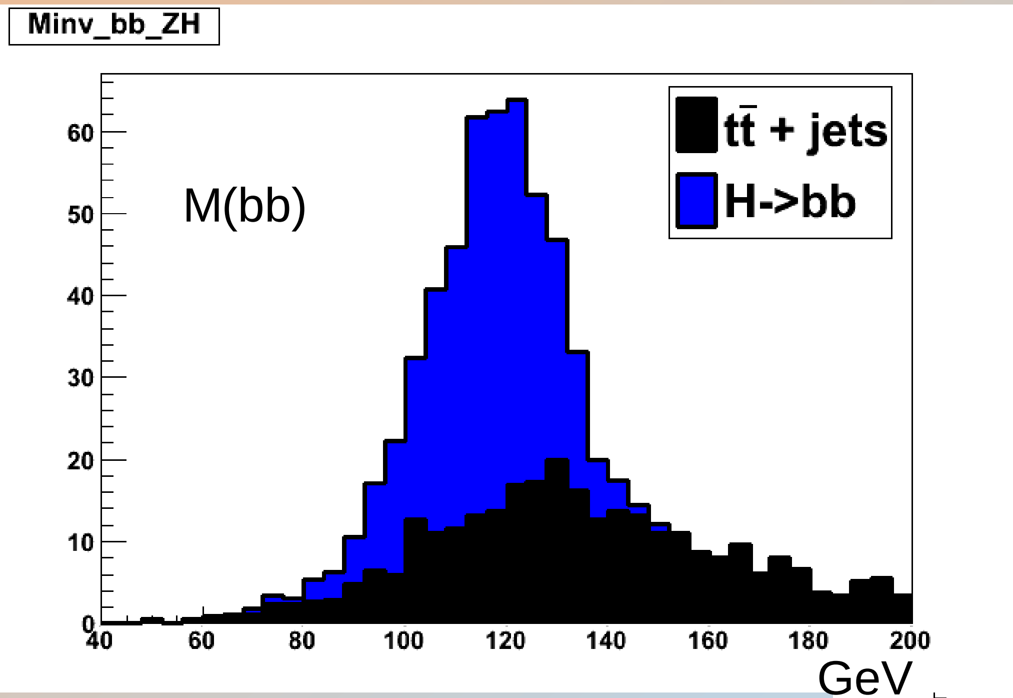
(Z->qq)+jets



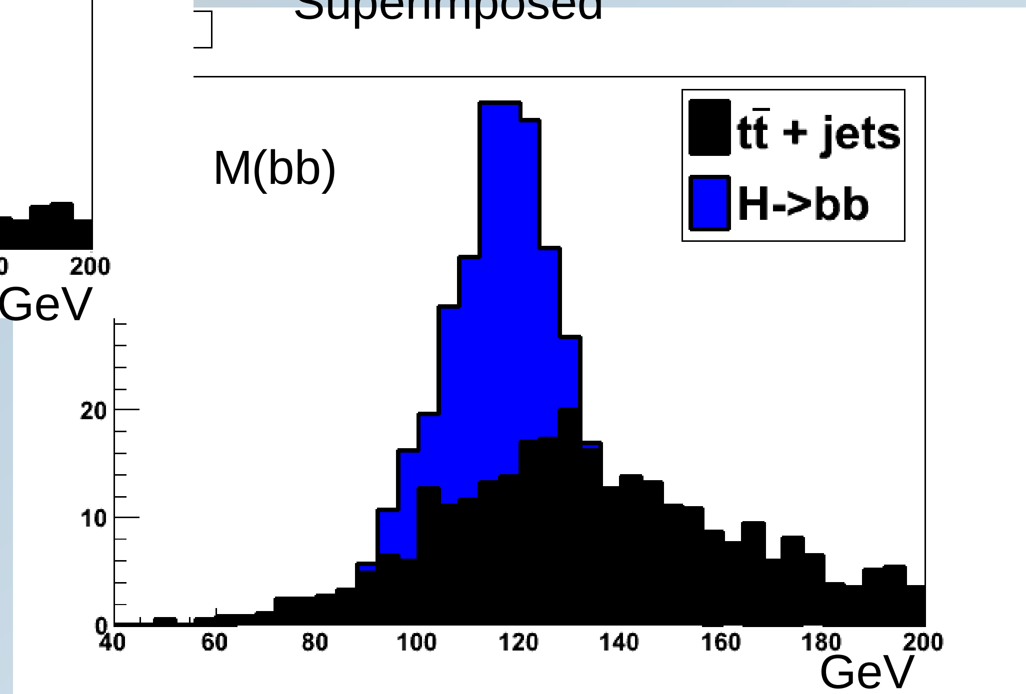


Use the Z peak presence as a reference candle

ttbar contribution



Superimposed



a look at the 2011 data

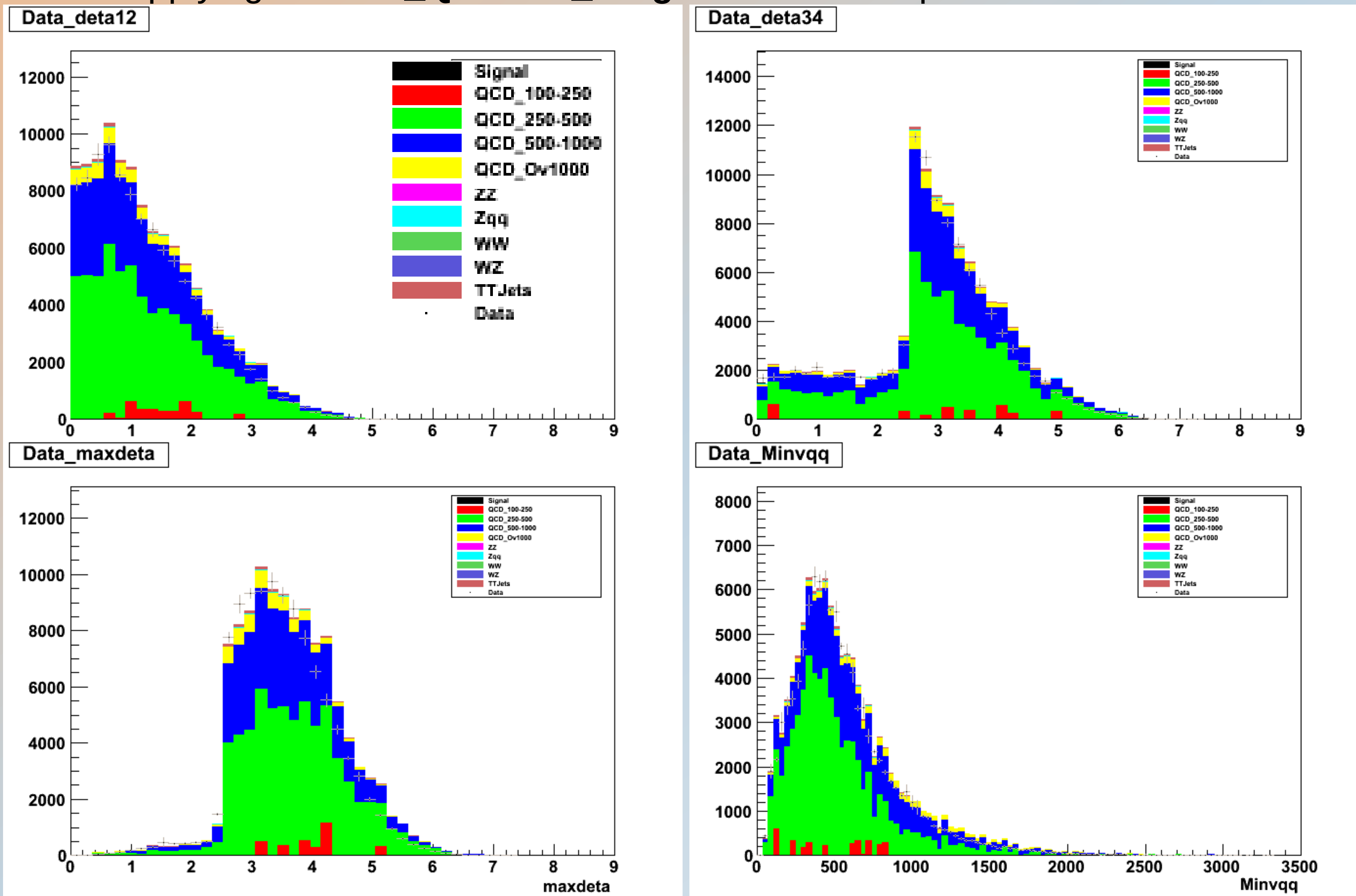
run the analysis on the **HLT QuadJet70** data sample (mostly unprescaled in 2011A) : luminosity **1.3 fb^{-1}**
HLT signal efficiency is 0.8% (0.4% after MLP>0.6 cut)

Monte Carlo reweighted to data PU distribution.
Data 21M events:

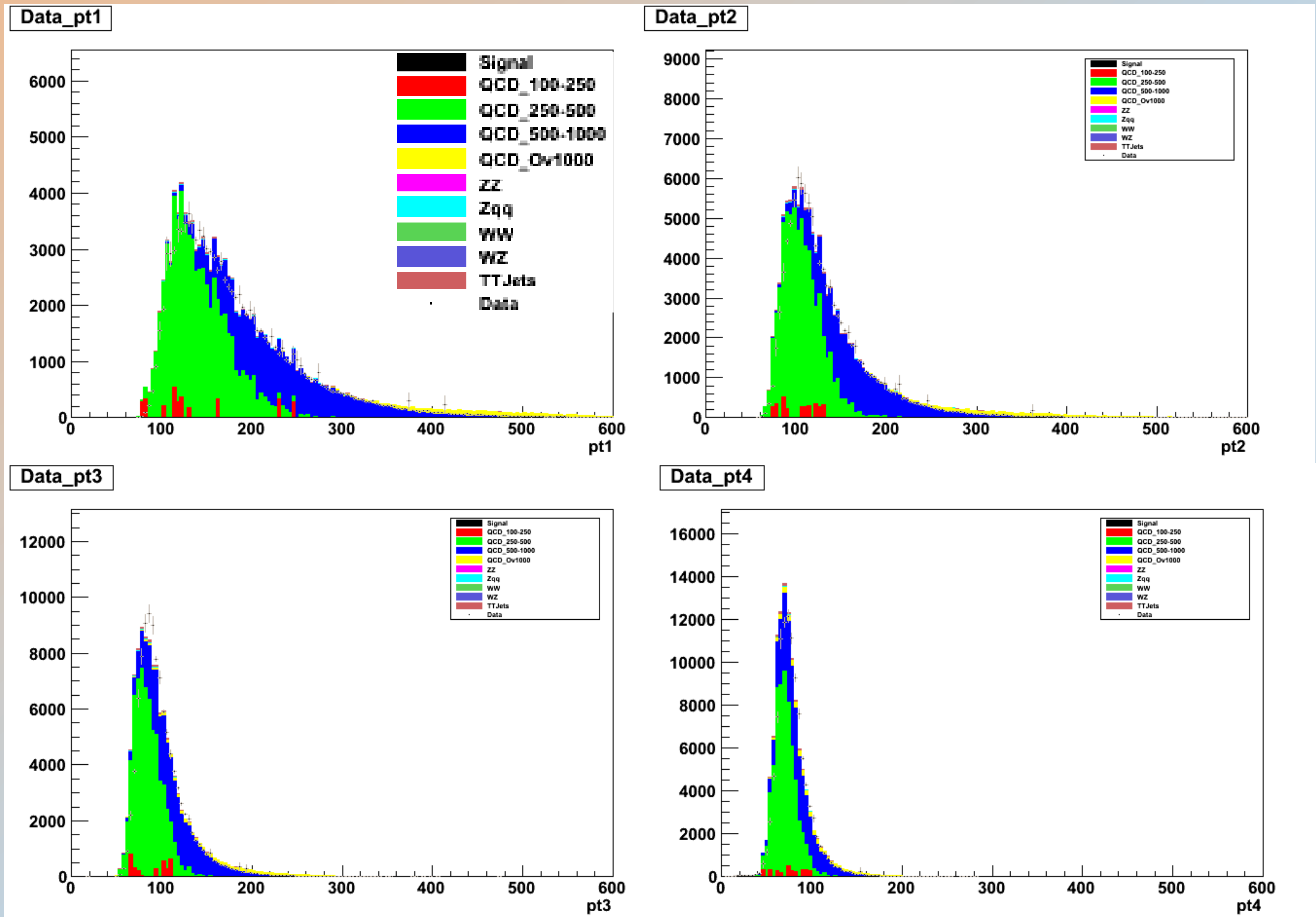
events	after HLT cuts	with MLP>0.6
QCD-100_250	872	0
QCD-250_500	26 493	794
QCD-500_1000	8 485	688
QCD-1000	454	12
ttbar	422	11
Zqq	152	19
WW	1	0.001
WZ	1	0.02
ZZ	0.4	0.03
Signal	**21**	**11.5 (0.28σ)**
All MC	**36 881 ± 820**	**1536 ± 92**
Data	**34 703**	**1676**

QuadJet70 data

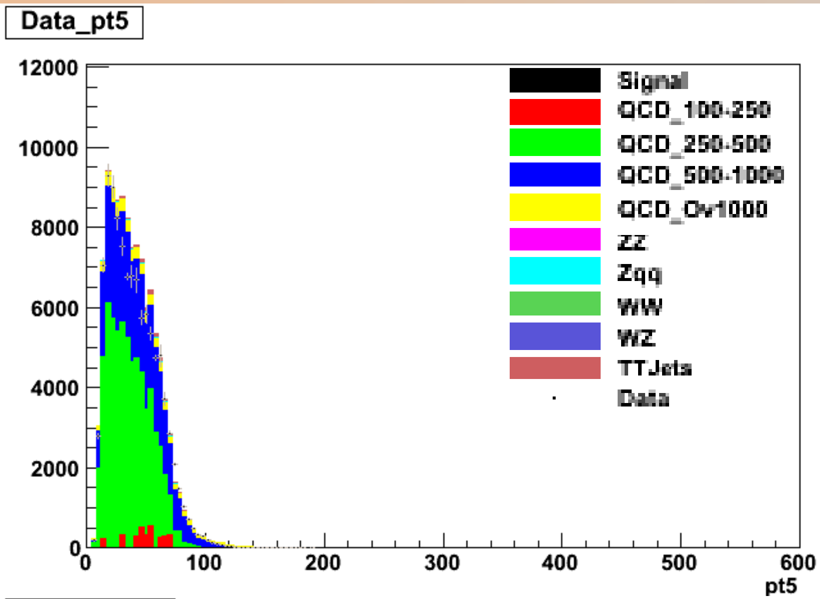
After applying our **HLT_QuadJet_Btag** selection on top

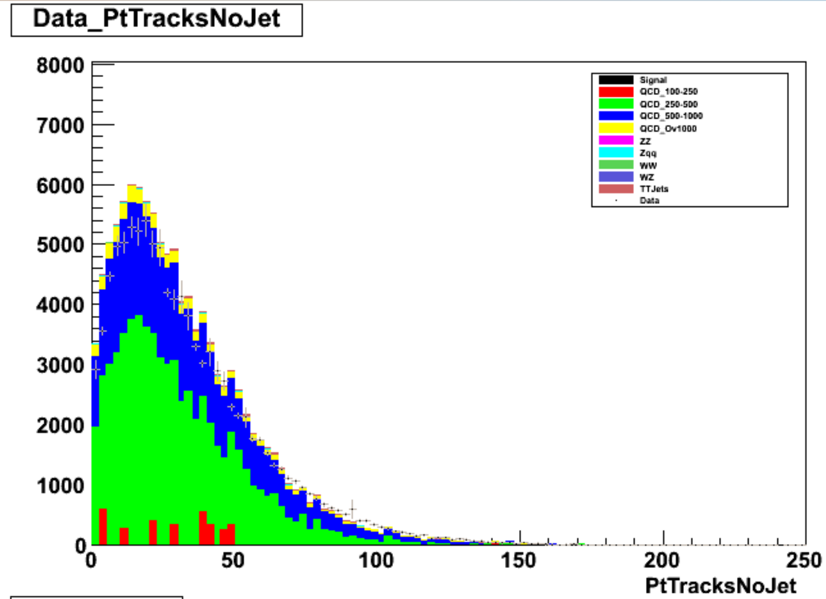


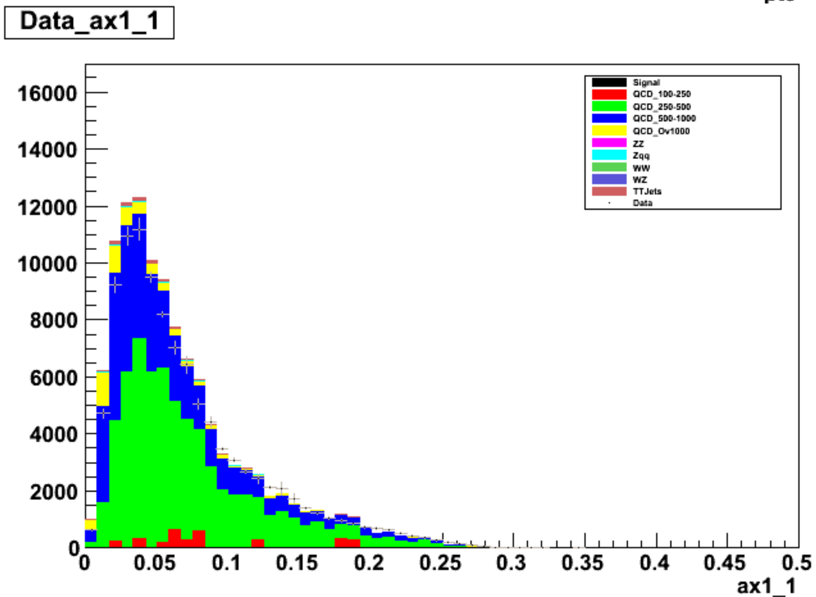
QuadJet70 data

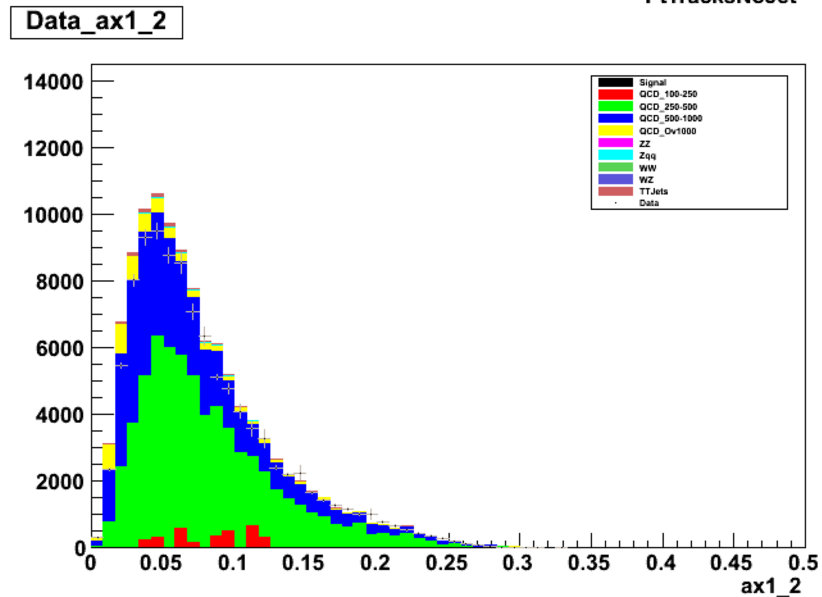


QuadJet70 data

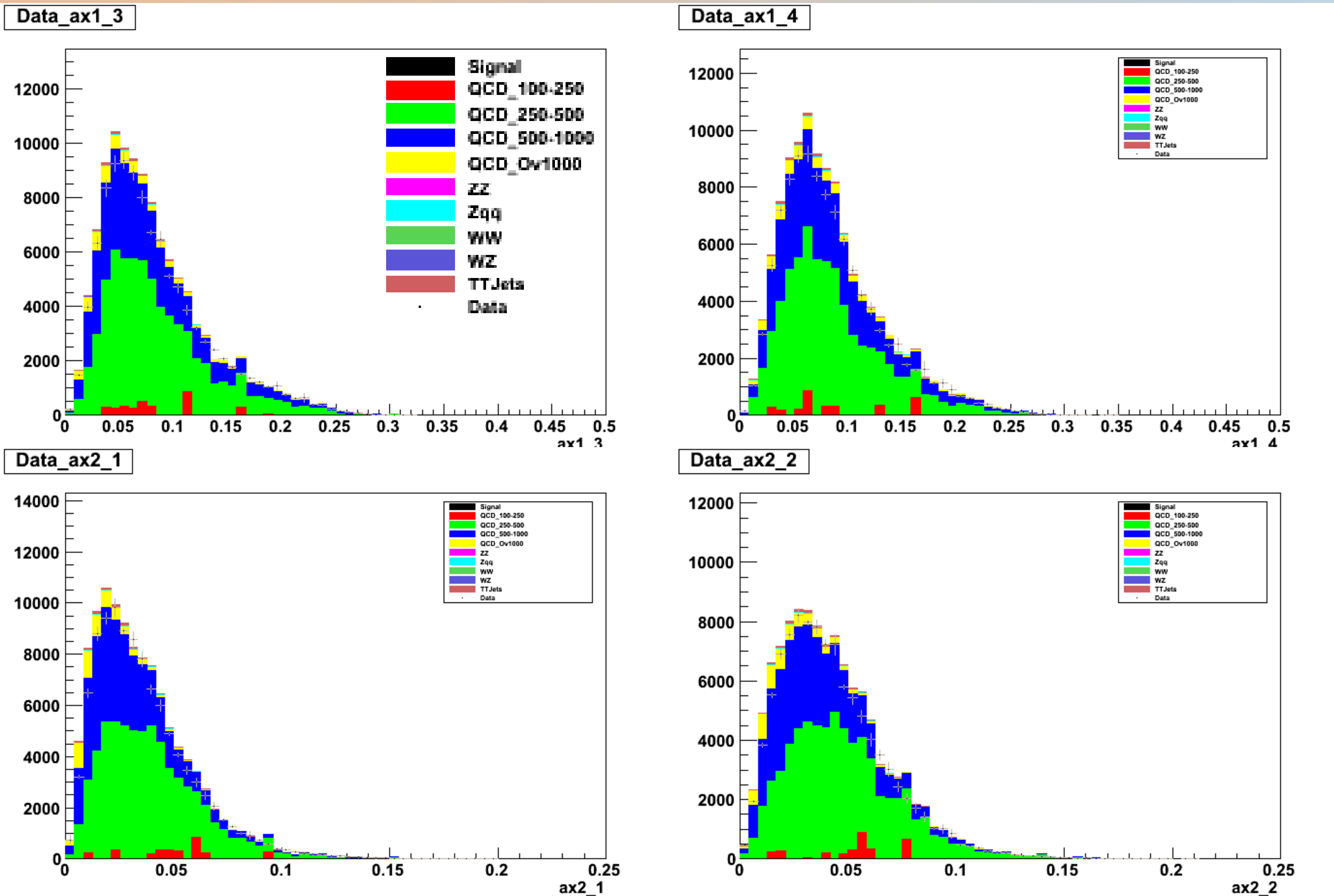




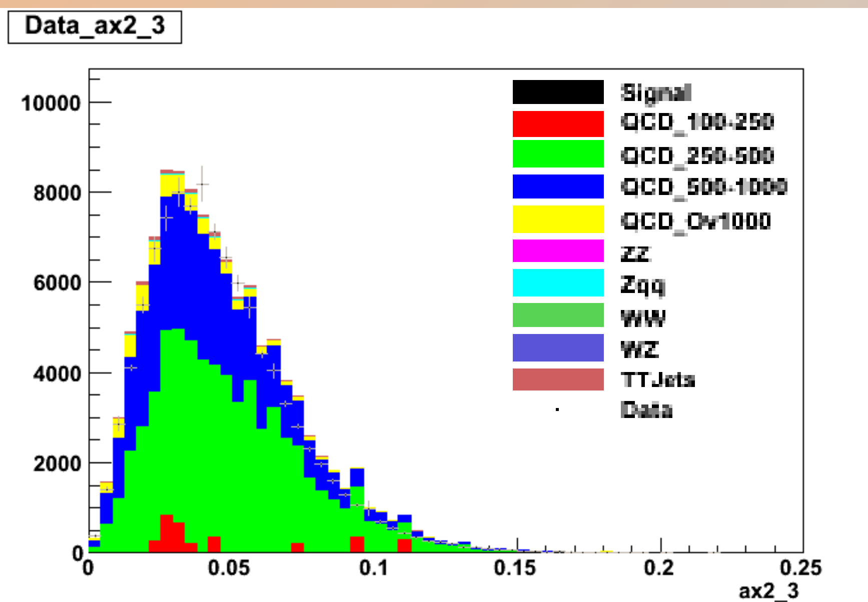


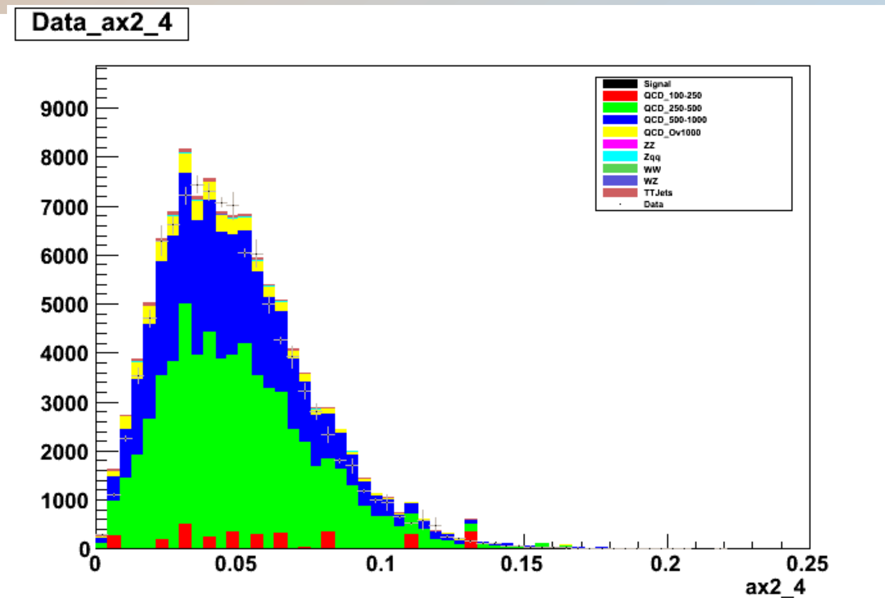


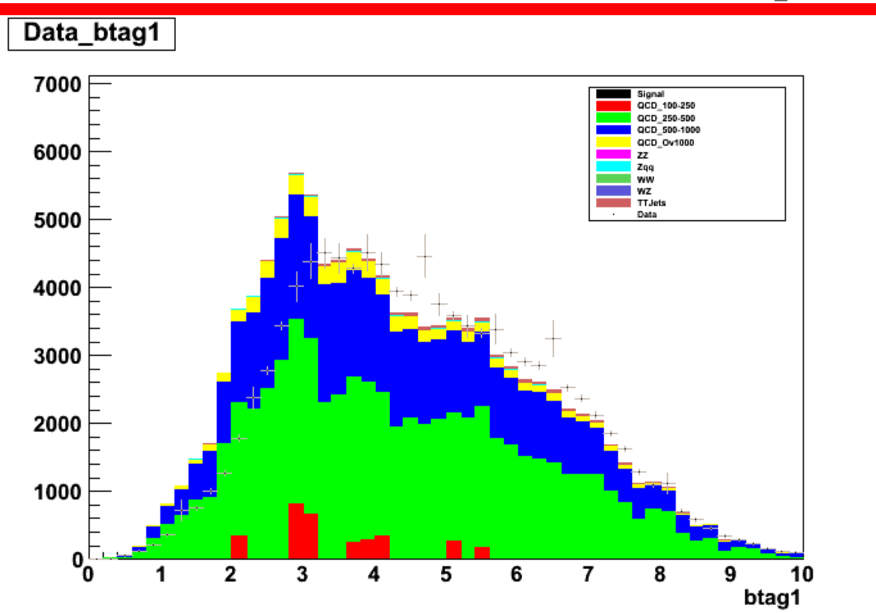
QuadJet70 data

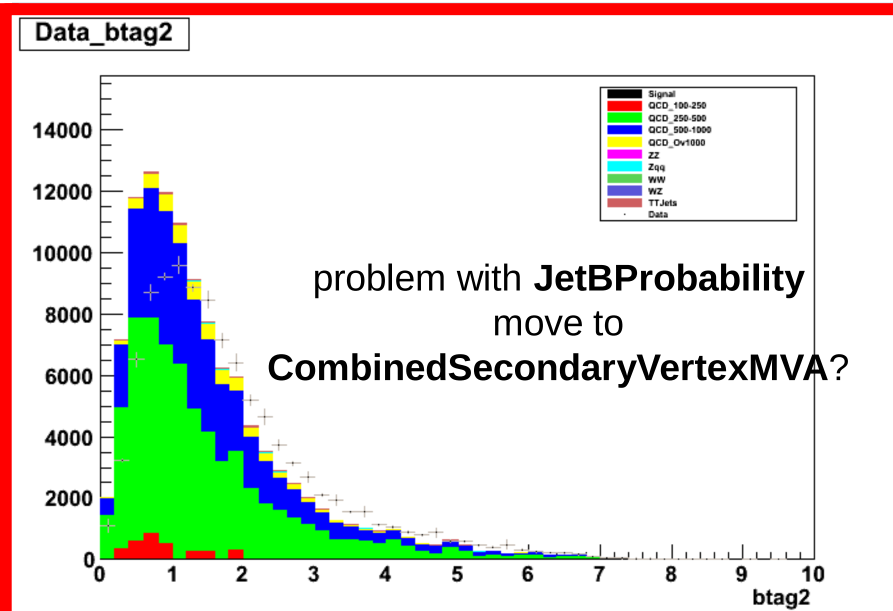


QuadJet70 data



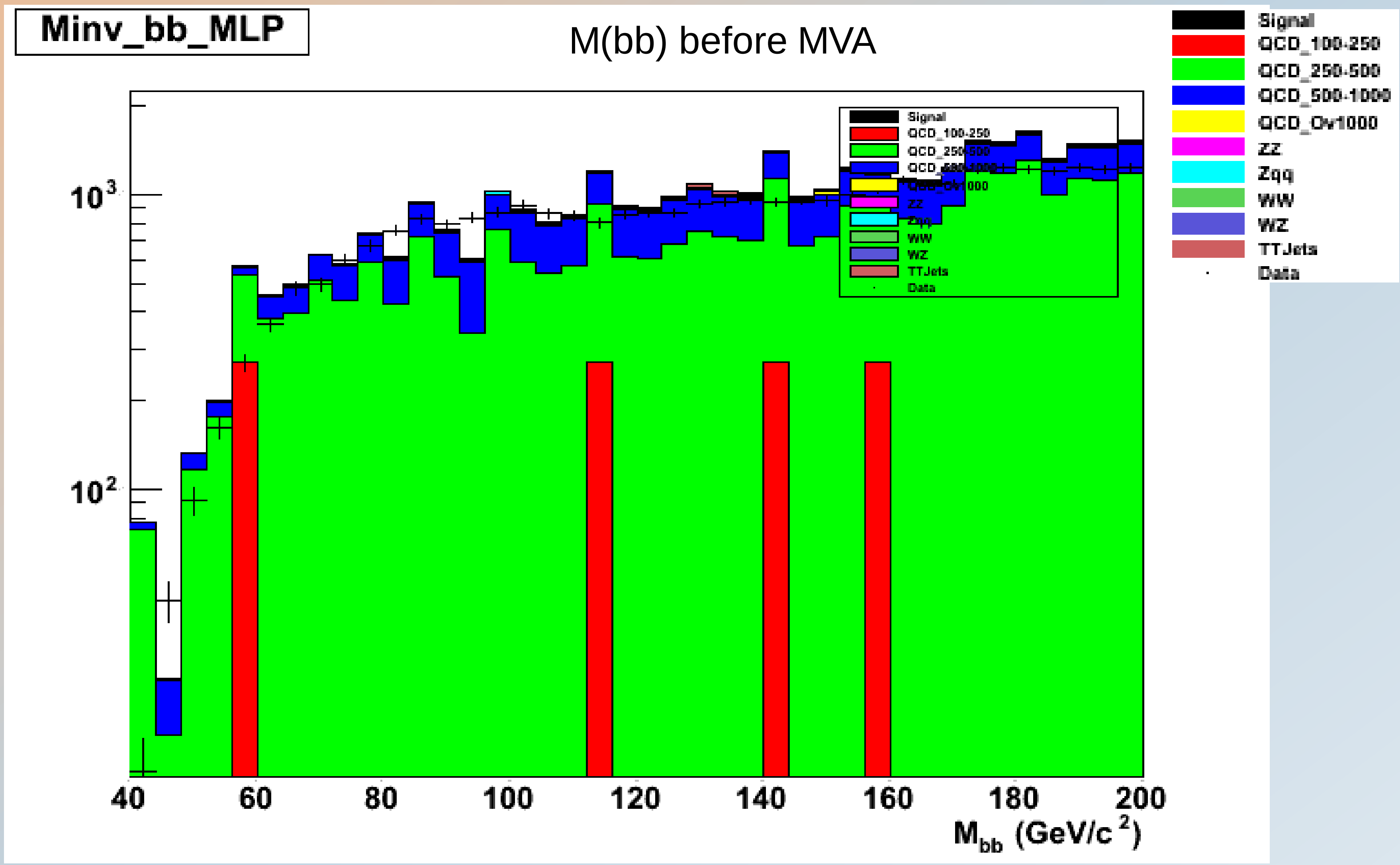




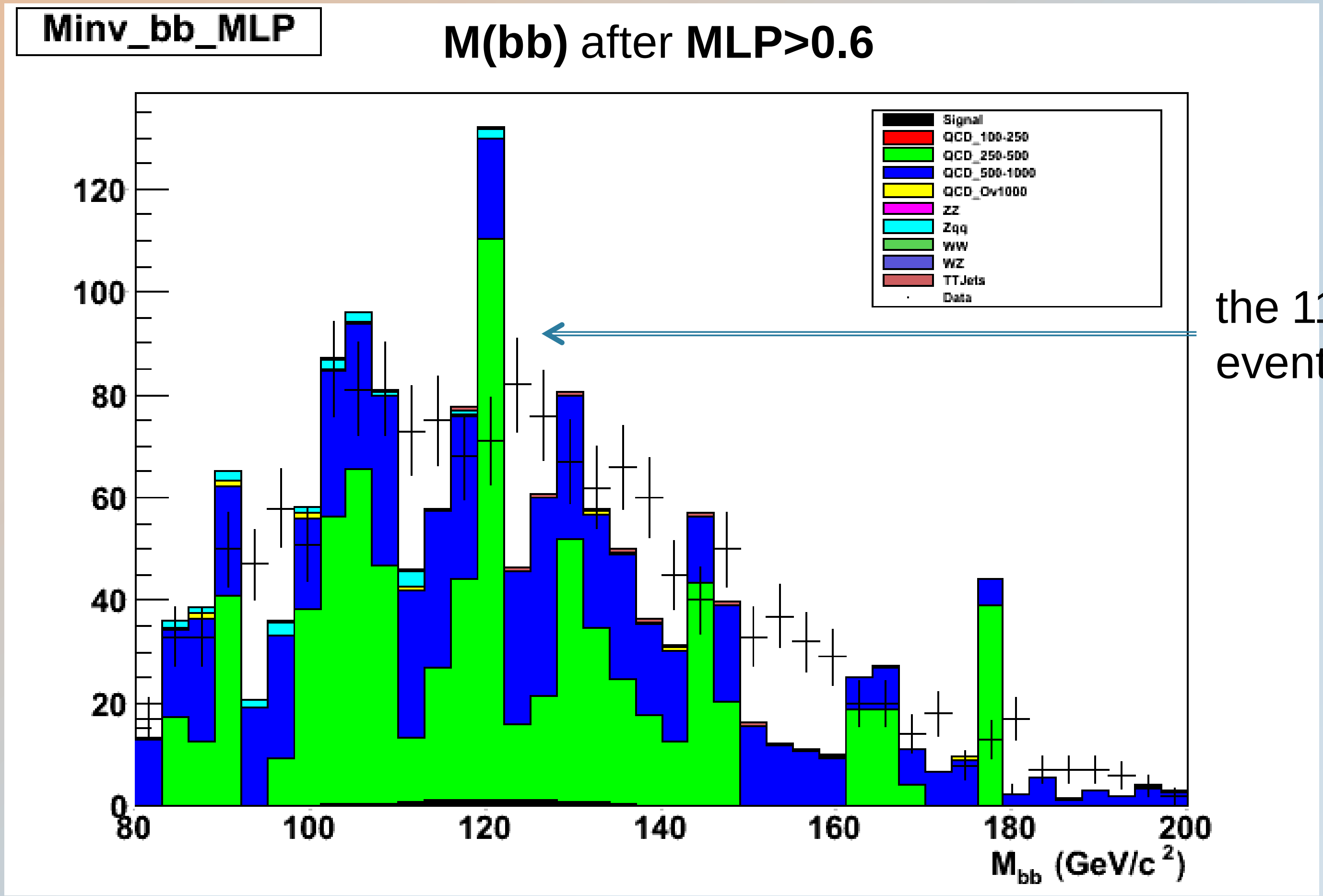


problem with **JetBProbability**
move to
CombinedSecondaryVertexMVA?

QuadJet70 data



QuadJet70 data



the 11.5 signal events are here ;-)

A forward shift in the data M(bb) ? (could be a JEC issue)

Conclusions & Outlook

Trigger approved and HLT Path submitted yesterday

The VBF H->bb channel could potentially bring extra significance to the CMS discovery of a light Higgs boson in the 115-125 range

This channel is extremely arduous but the final sensitivity could be not so far from other complementary channels (VH->bb or tau channels)

Need to start a laborious work on the signal extraction with a multi component fit of the final M(bb) distribution, and on accessing and estimating systematic uncertainties.

Help is welcome

(actually there are other interested people)

Backup

Kinetical Fit to “close” the event in the trasverse plane

$$\chi^2 = \Sigma_{i=1,n} (\alpha_i - m_i)^2/\sigma_i^2 + \lambda_1(\Sigma_{i=1,n} \alpha_i P_{xi} + P_x^{recoil}) + \lambda_2(\Sigma_{i=1,n} \alpha_i P_{yi} + P_y^{recoil})$$

i = index running on the number of the jet

n = total number of jets to fit

α_i = correction factor applied to the Lorentz vector (P_x, P_y, P_z, E)

m_i = response function of P_t^{reco}/P_t^{parton} as a function of P_t^{parton}

σ_i = rms of P_t^{reco}/P_t^{parton} as a function of P_t^{parton}

$P_{x(y)}^{recoil} = \Sigma_{i=n+1,N} P_{x(y)i}$

$\lambda_{1(2)}$ = Lagrange multiplier

Minimizing χ^2 respect to $\alpha_i, \lambda_{1(2)}$

$2(\alpha_i - m_i)/\sigma_i^2 + \lambda_1 P_{xi} + \lambda_2 P_{yi} = 0$

$\alpha_i P_{xi} + P_x^{recoil} = 0$

$\alpha_i P_{yi} + P_y^{recoil} = 0$

The linear system can be written with a symmetric matrix (n+2,n+2)

c_1 0....0	P_{x1}	P_{y1}	α_1		d_1
0 c_2 ..0	P_{x2}	P_{y2}			
.........c_n	P_{xn}	P_{yn}	α_n	=	d_n
....	0	0	λ_1		$-P_x^{recoil}$
....	0	0	λ_2		$-P_y^{recoil}$

$ci = 2/\sigma_i^2$

$di = 2m_i/\sigma_i^2$

Parametrization of m_i and σ_i as a function of P_t^{parton}

Matching jet-parton as the nearest jet respect to the parton direction and with a matchcng cut $\Delta R < 0.1$

Parametrization for q-jet and b-jet separately for central ($|\eta|<2.4$) and forward ($|\eta|>2.4$) region

offset=mean of $(P_t^{reco}/P_t^{parton} - 1)$ as a function of P_t^{parton}

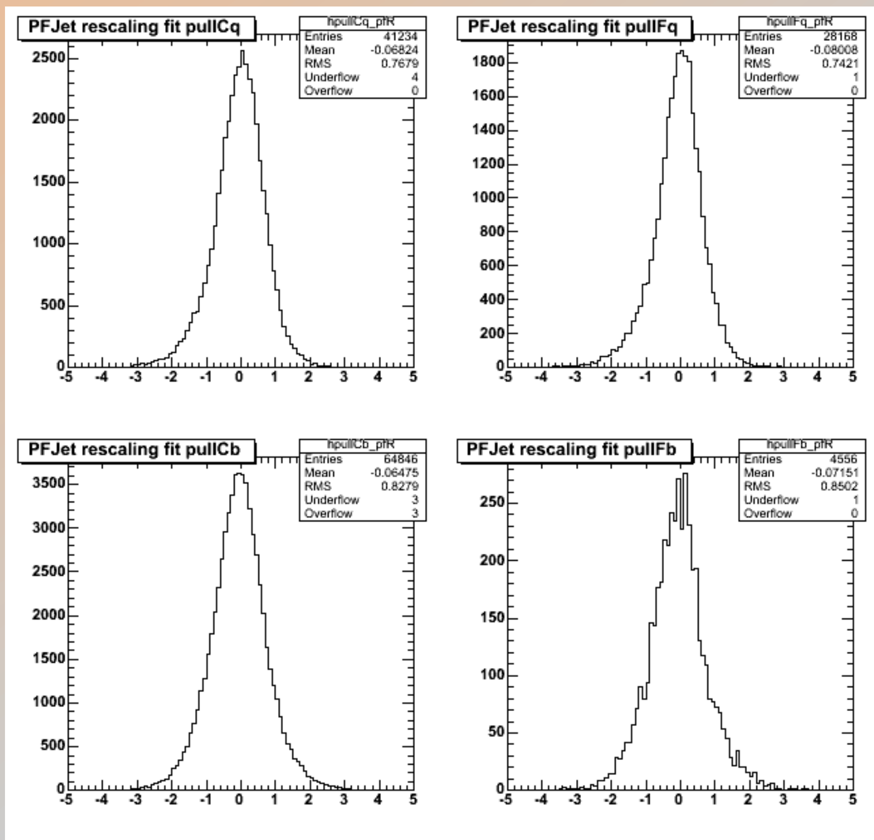
offset =a+b*pt

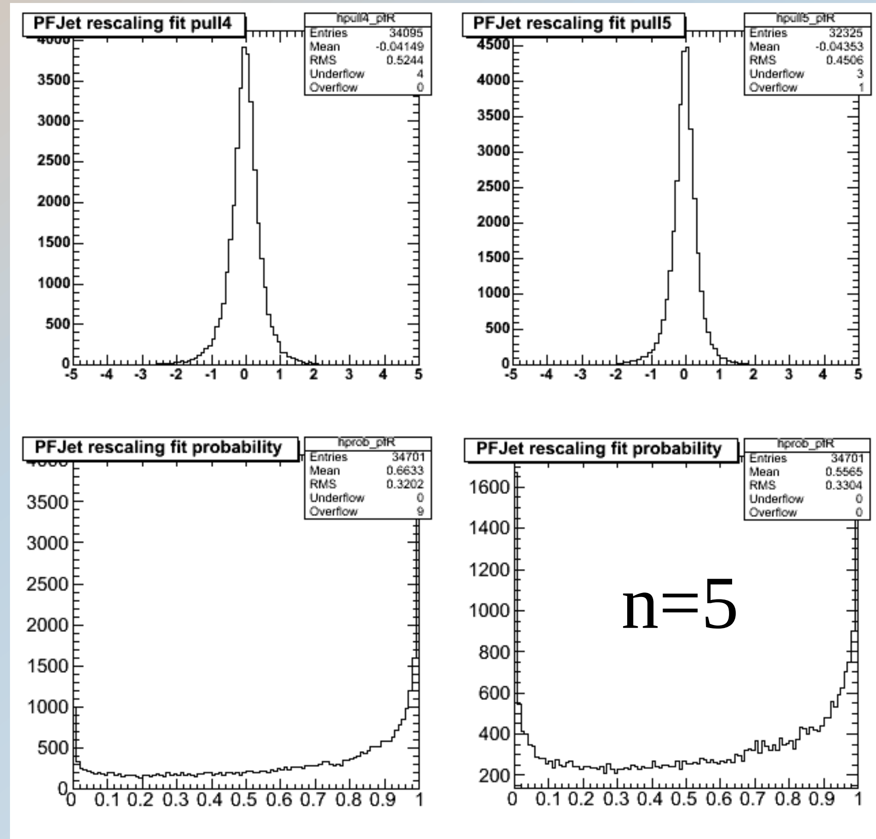
σ_i = rms of $(P_t^{reco}/P_t^{parton} - 1)$ as a function of P_t^{parton}

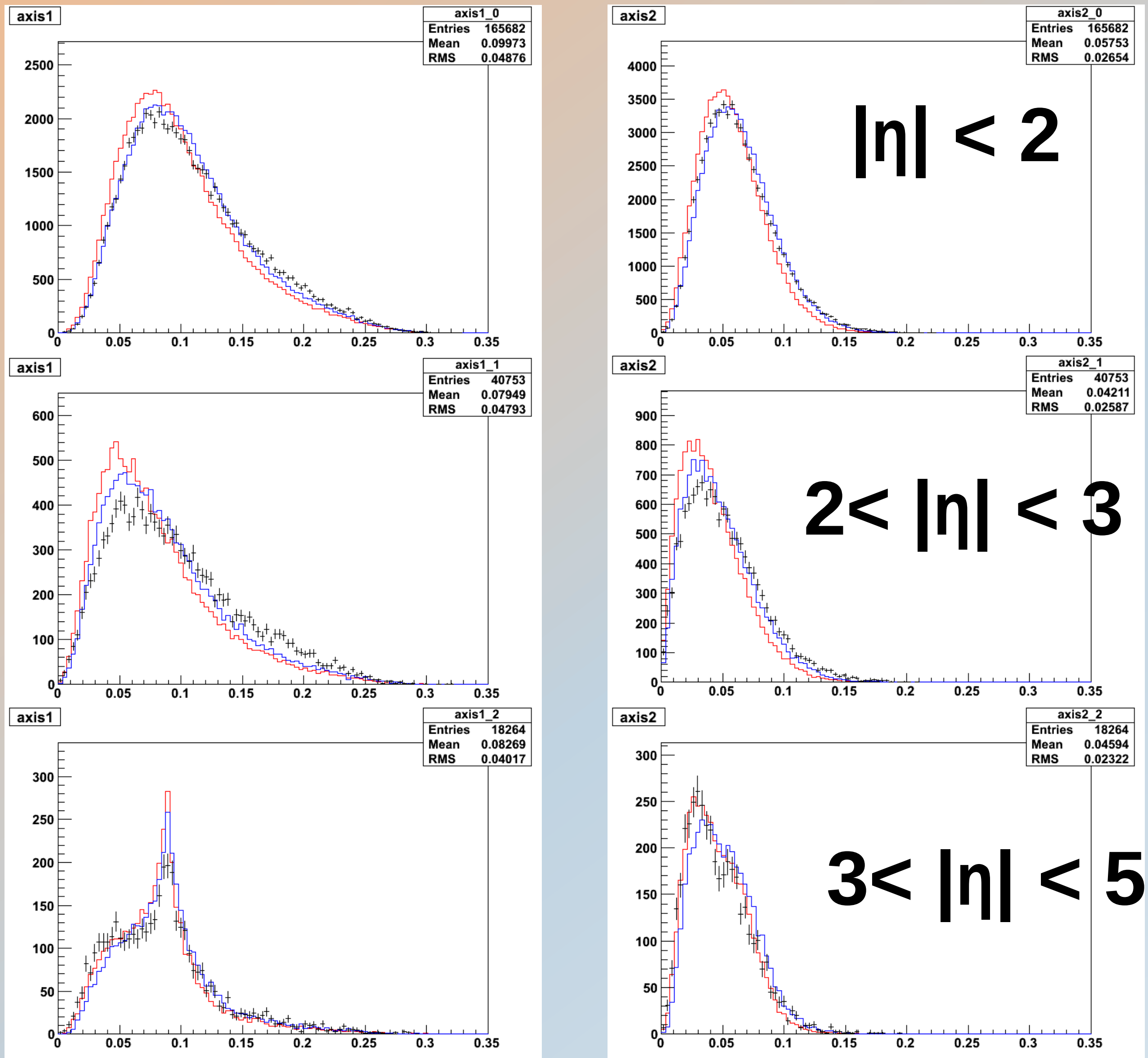
$\sigma_i = \sqrt{(a^2+b^2/pt)}$

$m_i = 1/(1+\text{offset})$

Fit pulls and Probability (n=6), anti-kT 0.5



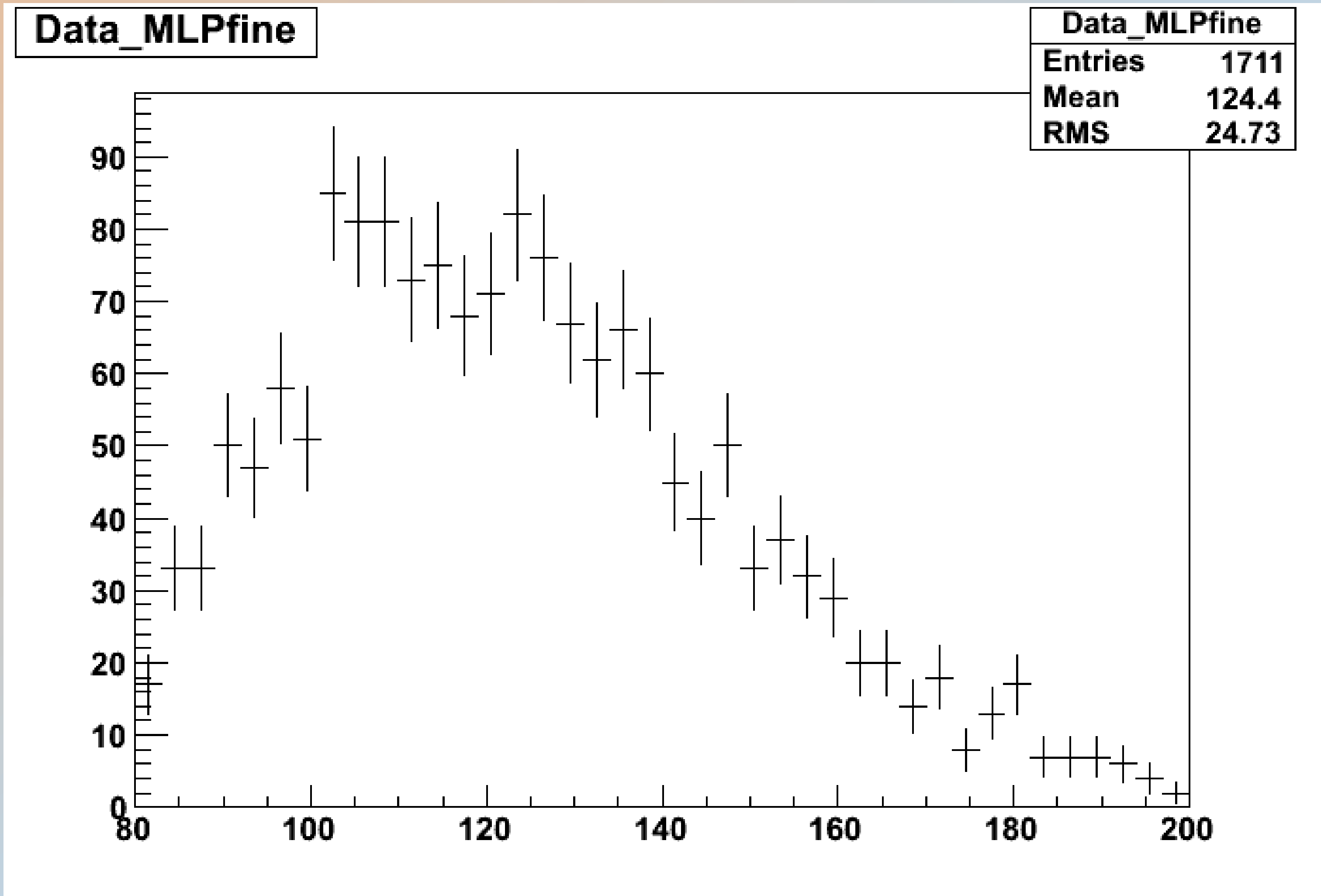


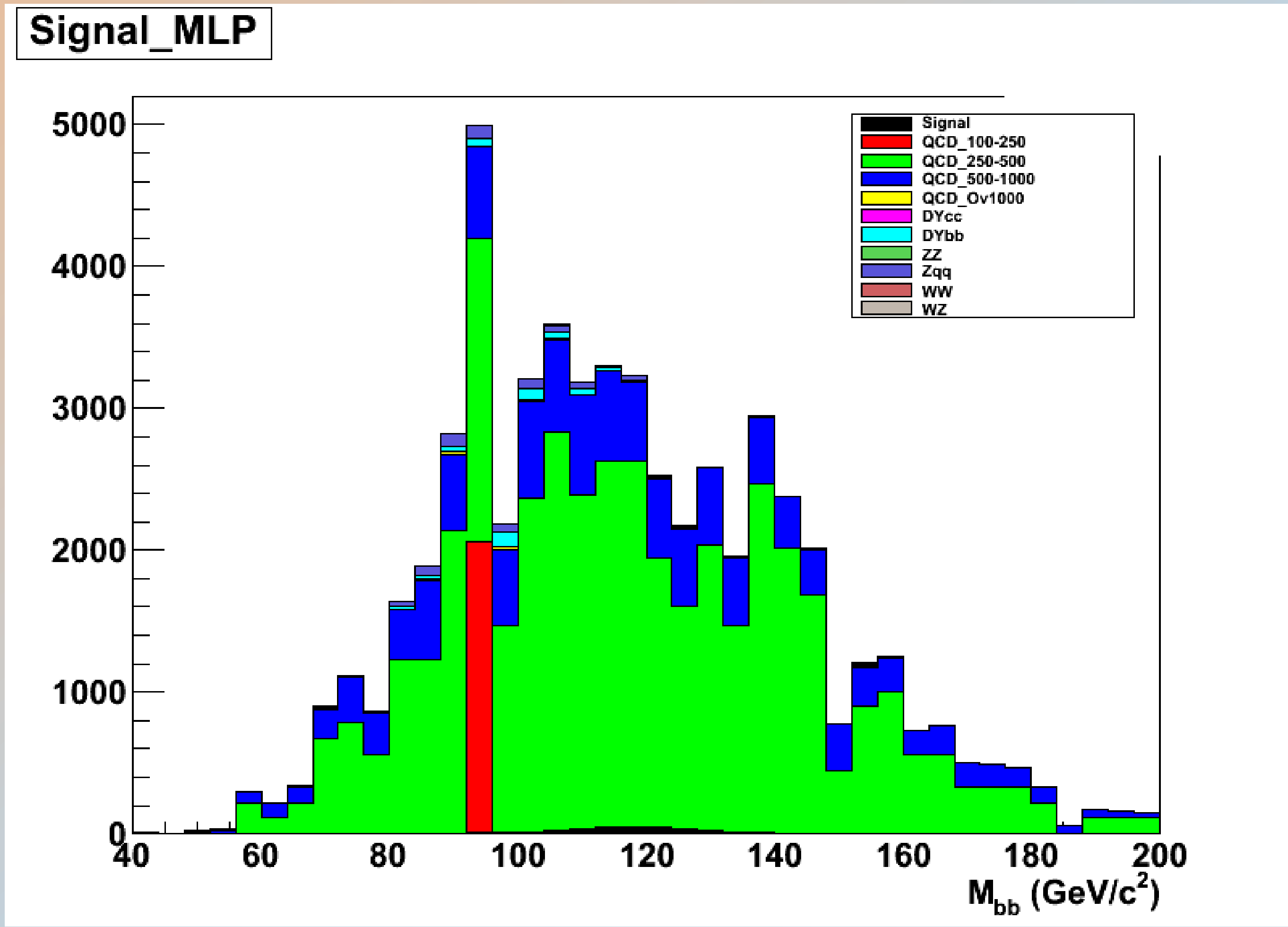


MC (No PU)

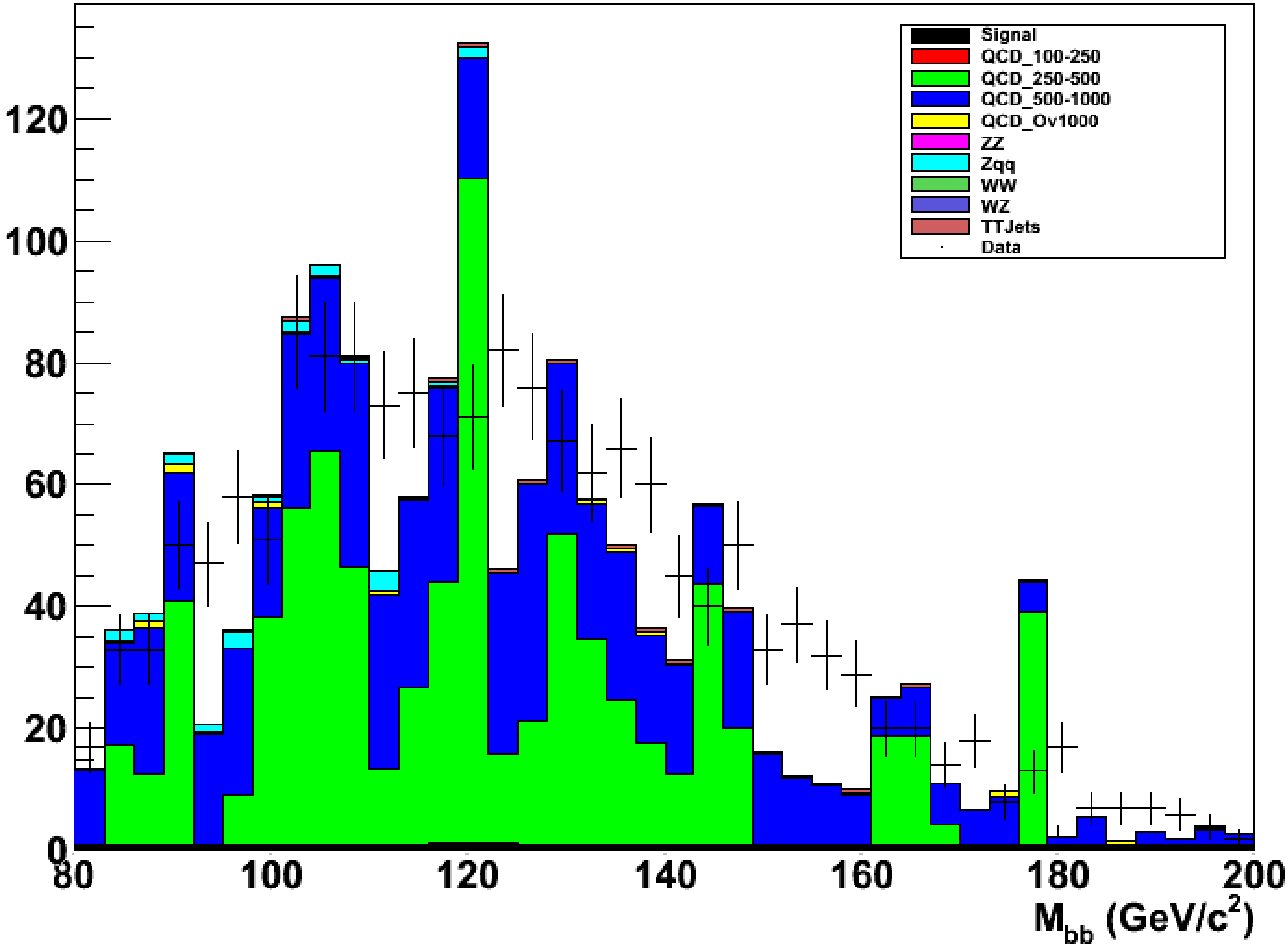
MC (2011 PU)

Data (2011A Data DiJetAve30)


Data_MLPfine
Data_MLPfine
Entries 1711
Mean 124.4
RMS 24.73
90
80
70
60
50
40
30
20
10
0
80
100
120
140
160
180
200


Signal_MLP
5000
4000
3000
2000
1000
0
40
60
80
100
120
140
160
180
200
M_{bb} (GeV/c^2)
Signal
QCD_100-250
QCD_250-500
QCD_500-1000
QCD_Ov1000
DYcc
DYbb
ZZ
Zqq
WW
WZ

Minv_bb_MLP



Signal
QCD_100-250
QCD_250-500
QCD_500-1000
QCD_Ov1000
ZZ
Zqq
WW
WZ
TTJets
Data
0
20
40
60
80
100
120
80
100
120
140
160
180
200
M_{bb} (GeV/c^2)