Karlsruhe Institute of Technology

New applications of the multi variate analysis framework NeuroBayes for an inclusive b-jet cross section measurement at CMS

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
von der Fakultät für Physik des
Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl. Phys. Simon Honc

aus Stuttgart

Tag der mündlichen Prüfung: 13.05.2011
Referent: Prof. Dr. M. Feindt, Institut für Experimentelle Kernphysik
Korreferent: Prof. Dr. Th. Müller, Institut für Experimentelle Kernphysik

”Der Vorwurf, meine Doktorarbeit sei ein Plagiat, ist abstrus.”
(K. T. zu Guttenberg am 16. Februar 2011)

für Lisa

Contents

Chapter 1

Introduction

This thesis is about the measurement of the inclusive b-jet cross section and new applications of the multivariate analysis framework NeuroBayes. As for most PhD theses the title is a accumulation of technical terms and for most people more or less incomprehensible. But nevertheless the title is deliberate.

New applications of the multi variate analysis framework NeuroBayes for an inclusive b-jet cross section measurement at CMS

The title covers the main topics of this thesis. These are the new applications of NeuroBayes and the measurement of a physical quantity: the inclusive b-jet cross section. The reader will be brought to these topics after a extensive introduction of the required background.

In chapter 2 I will highlight the theoretical basis of the physical contents. Therefor I will give a short essay about the historical process up to the formulation of the quantum chromo dynamics (QCD). QCD is the underlying theory behind the measurement presented in this thesis and describes the strong interactions between the particles. Its intellectual father Murray Gell-Mann won the nobel prize "for his contributions and discoveries concerning the classification of elementary particles and their interactions" as early as in 1969 [Nob].
Many years of research enhanced the list of elementary particles. Until now we found twelve elementary fermions, six of them, the quarks, are able to do the strong interaction. The quarks vary in electromagnetic charge and mass. One of these particles is the so called bottom quark (b-quark). Its mass is about 4.2 GeV. The name bottom is chosen in analogy to the down-quark, which is part of the proton.
The b-quark was discovered in 1977 at Fermilab [HHL$^+$77]. This event was the starting point of a huge field of research: b-physics. There are three main topics in b-physics: b production, B spectroscopy and b flavor physics. The first two cover physical effects caused by the strong interaction, while the last treats the description of the weak decay of the b-quark, which is very important for discovery of physics beyond the standard model. The LHCb experiment at CERN was built especially for analyses in this interesting sector.
But also studies on the strong interacting sector of the b-quark play an important role. On the one hand b-quarks contribute to the background distributions for many analyses. A good understanding of the b quark production mechanism may lead to significant improvements. Especially for new particles, which decay into b-quarks, this is non-negligible. Furthermore for these processes a good identification of the b-quarks is important.
On the other hand the production of the b-quarks itself is very interesting. In the last decades analyses on this topic lead to curious results. At the beginning of the new millenium they already

claimed new production mechanisms beyond QCD predictions [Ber01], [Jun03]. It was not until ten years later thet the results could be brought in line after a recalculation of the old theory. Above all the difficulty to solve the perturbation calculations for heavy particles and the insufficient modeling of the hadronization process led to those discrepancies between theory end experiment.
Today it is possible to reach a new regions in energy with the experiments at the Large Hadron Collider (LHC). Thus it becomes again very interesting to check whether the theory is able to describe the new measurements. In this thesis I will present the first analysis which covers the production of b-quarks at such energies. The studies are done at the CMS experiment.

The third and the forth part of the thesis cover the description of the experiment. I will picture the splendid history of CERN and its experiments which culminate in the construction of the LHC and its experiments. I will introduce the layout of the CMS detector in chapter 3. With this apparatus we are able to measure the physical processes which happen after the collision of protons at a center of mass energy of 7 TeV. Further it is planned to increase the energy up to 14 TeV in the near future. The obtained data must be transformed into physical objects. In chapter 4 I will present how these objects are reconstructed.
The main topics are presented in chapter 5 and 6.

But let us have a more detailed look into the main parts. The syntax of the title is chosen to emphasize these topics: For the measurement new applications where developed. An important part of this thesis deals with the aggregation of known and new methods based on NeuroBayes. NeuroBayes is unknown to most people so it is described in more detail.
NeuroBayes is a tool to do multi variate analysis. In the simplest case, this means that many input variables are used to do a classification of two targets: background and signal or in more general: target 0 and target 1. The inputs are combined and transformed to a single output variable which carries all information to do the classification. The multi variate phase space is reduced to a one dimensional. In section 5.1 a detailed summary of NeuroBayes is presented.
In the thesis itself I developed different applications for this. They range from general derivations how to combine known methods with NeuroBayes to new approaches based on NeuroBayes and specific applications for the physical analysis.
In section 5.2 I will focus on the interpretation of the NeuroBayes output. NeuroBayes is constructed in a way that it is possible to transform the results into a probability. This depends among other things on the specification of the input samples. I will discuss two different setups. The first I will call Monte Carlo (MC) based, the other data based. In fact the main difference is the type of one of the targets for the classification. Either we take simulations of the background distribution in the MC based approach or a data sample for the data based approach. The derivations of this I will show in section 5.2.1.
The knowledge on the probability can be used to execute a so called boost training. In section 5.2.2 I will introduce the basic idea, the implementation and the resulting possibilities.
Another application of NeuroBayes that was developed is the transformation of its output into sPlot weights. sPlot is a method to determine the inclusive distribution of a specific variable using the inclusive informations of a source variable which is uncorrelated to the former. In section 5.2.3 I will derive the connection between NeuroBayes and sPlot. With this it is possible to take the NeuroBayes output directly as source for the sPlot method. The inspected variable has to be uncorrelated to the NeuroBayes output.

Having many tools based on NeuroBayes in place it is obvious to apply them on physical topics. Thus, it is used for a classification of jets. A jet is an observable pattern in the detector. Quarks and gluon create many particles with a similar direction. On the one hand this happens because

of the hadronization of these particles. On the other hand most of the created hadrons decay into lighter particles in turn. All these particles were combined to so called jets. If such a jet is created by b-quarks it is called b-jet. The aim of a b-jet tagger is to classify jets into b-jets and non-b-jets. In section 5.3.1 I will present a NeuroBayes b-jet tagger based on Monte Carlo simulations (MC) and in a second case based on data. For the data based tagger differences between data and MC have to be studied. This was also done with NeuroBayes. Further I applied another method, called boost, to show how much improvements are possible when doing this procedure.

In the last chapter I will focus on the inclusive b-jet cross section measurement done by CMS. I will present the recent analysis done on early CMS data in section 6.1. This measurement was done on data with an integrated luminosity at 60/nb. During the first year of data taking CMS already collected 36/pb of data. Therefor an update of the recent measurement is planned. In section 6.2 I will present our contributions to this analysis and prospects for future results.
At the end I will start a discussion on alternative approaches to do this inclusive b-jet cross section measurement. I will present a method based on the jet classification performed with the NeuroBayes framework. In section 6.3 I will show how the results change if the newly developed NeuroBayes b-jet tagger is used. The new tagger is used in the same manner as for the former measurement.

Finally the results of this thesis will be summarized and discussed in chapter 7.

Chapter 2

The Standard Model of particle physics

In this chapter I will introduce the Standard Model of particle physics. The first part is an overview of the parameters of the model. Precise mesurements of these quantities are needed to make further estimates on the behaviour of the nature at the elemantary level.
The second part is about a special part of the model, perturbative quantum chromo dynamics (pQCD). For this thesis we want to compare the experimental data with the predictions made by the Standard Model. Therfore we need the calculations done by theorists. I will focus on the problematics that came up over the years doing such calculations and present the current status. Finally, I will work out the prospects of a further measurement of the inclusive b-jet cross section at CMS.

2.1 Parameter of the Standard Model

Particle physics had its genesis with the discovery of the electron by Thomson in 1897 [Tho97]. Not that they were aware of the further particles we may discover, but the door was opened for a new field of physics. In the beginning of the 20th century the picture of the atom was completed and also the first anti-particle, the positron, was discovered in 1932 by C.D. Anderson [And33].
Nobody expected what would happen in the following years. It started with the discovery of a new particle in cosmic rays: the muon in 1937 [SS37]. Its discovery led to the formulation of the quantum electro dynamics (QED) in the 1940s by J. Schwinger, R. Feynman and S. Tomonaga [Dys49]. Henceforward many further particles were found, many were detected in cosmic rays, but also first accelerators were built. The first discovery of a new particle produced in collisions was the neutral pion in 1949 [SPS50]. Until 1961 the number of particles rapidly increased. They found many of the ground states of the today so called mesons and baryons, and also discovered the first neutrinos in 1956 [RC56]. Up to then they found more then 20 different particles.
1961 a particle called η was discovered [PRS61]. This was the needed ingredient for a new phenomenological classification of the particles: the eightfold way. Dependent on the measured quantum numbers a systematic ordering was possible (see figure 2.1).
1964 the model was confirmed by the discovery of the Ω^-, which was the last missing particle to complete the structure of the eightfold way. Knowing this, Gell-Mann and Zweig independently saw the possibility of a underlying theory using group theory with an SU(3) symmetry. This was the birth of today's quark model and finally, together with the quantum electro dynamics (QED),

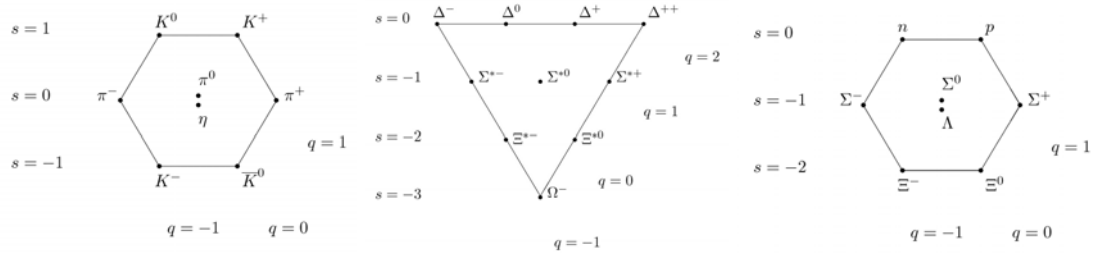


Figure 2.1: The eightfold way is an ordering of the ground state particles as proposed in 1961. The particles are ordered dependent on their quantum numbers S and q. This composition leads to the prediction of the later discovered particle Ω^{--} and is the base for the later formulated quark model.

		charge				charge		
		e.m.	weak	color	mass	e.m.	weak	color
leptons						right handed		
I	neutrino ν_e	-	+1/2	-	< 2 eV	-	-	-
	electron e	-e	-1/2	-	511 keV	-e	-	-
II	neutrino ν_μ	-	+1/2	-	< 0.19 MeV	-	-	-
	muon μ	-e	-1/2	-	106 MeV	-e	-	-
III	neutrino ν_τ	-	+1/2	-	< 18.2 MeV	-	-	-
	tau τ	-e	-1/2	-	1.8 GeV	-e	-	-
quarks						right handed		
I	up u	+2/3e	+1/2	rgb	2.49 MeV	+2/3e	-	rgb
	down d	-1/3e	-1/2	rgb	5.05 MeV	-1/3e	-	rgb
II	charm c	+2/3e	+1/2	rgb	1.27 GeV	+2/3e	-	rgb
	strange s	-1/3e	-1/2	rgb	101 MeV	-1/3e	-	rgb
III	top t	+2/3e	+1/2	rgb	172 GeV	+2/3e	-	rgb
	bottom b	-1/3e	-1/2	rgb	4.2 GeV	-1/3e	-	rgb

Table 2.1: List of elementary particles of the Standard Model. The parameters are taken from [N+10].

the Standard Model of particle physics.

In the following I will present the elementary particles and the interactions of the Standard Model. This section is partially extracted from [Ind04].

2.1.1 The elementary particles of the Standard Model

The one part of the Standard Model are the so called fermions. All particles with a half-integral spin quantum number are assigned to this class. The elementary particles of this kind are divided in leptons and quarks. For leptons, as well as for quarks, three generations of doublets exist. In its generations the particles differ only by their mass. The quantum numbers are the same. Table 2.1 shows an overview of the different elementary particles.
Particles only participate in interaction where they have charge. It is to remark that the leptons have no color charge. Further the neutrinos have even no electromagnetic charge. The weak interaction only couples to left handed fermions. Therefore right handed neutrinos do not interact with other particles except by gravitation. Because of the weakness of the gravitation these kind

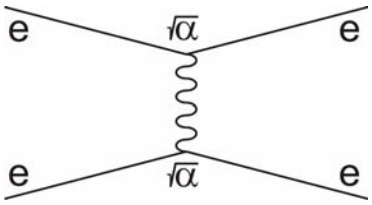


Figure 2.2: Feynman diagram of the electron scattering. A virtual photon is exchanged for the interaction.

of neutrinos are not detectable. It is not known if such particles exist.
For each particle further exists an anti particle with opposite quantum numbers but same mass.
Quarks appear only in neutral colored combinations. Combination consisting of two quarks are called mesons, combinations with three quarks baryons. Combinations with more than three components are also possible but not observed in nature, yet.

2.1.2 The interactions of the Standard Model

There are four types of interactions, which we are able to experience: gravitation, electromagnetic (e.m.), weak and strong interaction.

Gravitation

The gravitation is not included in the Standard Model of particle physics. I will briefly discuss its main properties.
The pull of the gravitation is very small for energies below the Planck scale of 10^{19} GeV compared to other interactions. We can neglect it in the model of particle physics. The gravitation has no repulsion. Looking at cosmic phenomena it gets a dominant contribution e.g. for the motion of the planets, stars and galaxies.
A boson called graviton, which carries the interaction, has not been discovered until now.

Electromagnetic interaction

The electromagnetic interaction is described by the quantum electro dynamics (QED). QED is a group theory with a abelian symmetry group U(1). The U(1) implies only one generator which represents the electromagnetic charge. The mediators of this charge are virtual photons. The photons are bosons with spin 1 and have no mass. Further they have no charge themselves. Thus they are not able to couple to each other.
Figure 2.2 shows a leading order Feynman diagram of QED process. There are two electrons which interact to each other via a virtual photon. The coupling at the vertices is proportional to the coupling constant $\sqrt{\alpha}$. $\alpha \approx \frac{1}{137}$ for small energy transfers. Therefore it is possible to calculate the QED in terms of perturbation theory. Quite good approximations are already reached at $\mathcal{O}(\alpha^2)$.

Weak interaction

The weak interactions are described by the non abelian symmetry group SU(2). This group implies three generators. One possible exposure of these generators is the use of the Pauli matrices J_i, $i = 1, 2, 3$ multiplied by a factor of $\frac{1}{2}$. Three generators leads to three gauge bosons. All three gauge bosons were discovered. There is the neutral Z boson with a mass of 91.2 GeV and the two charged $W^{\pm}$ bosons with a mass of 80.4 GeV. Z and $W^{\pm}$ bosons are spin 1 particles.

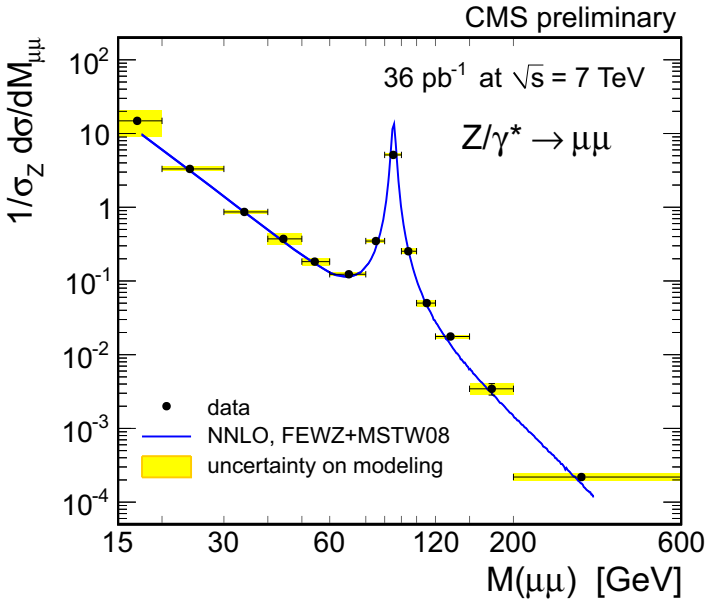


Figure 2.3: The normalized Drell-Yan mass spectrum, $(1/\sigma Z)d\sigma/dM$, obtained in the di-muon channel and compared to theoretical predictions. The uncertainty on the modeling accounts for differences in the acceptance corrections obtained with POWHEG and FEWZ. [CMS11]

Further the high masses lead to a small range of the weak force, because of the Heisenberg uncertainty principle. The product of the energy taken from the vacuum ΔE for a specific time scale Δt is limited. Therefore the weak interaction is strongly suppressed. The raw estimate of the maximum range l for the weak interaction is defined by $l = c \cdot \Delta t$, where c is the speed of light.
The gauge bosons couple to the third component of the weak isospin I_3. Right handed particles and left handed antiparticles have $I_3 = 0$. Therefore the weak interaction couples only to half of the particles. The Lorentz invariant property of the handedness is the chirality. It is a generalization of helicity h, which is defined as the normalized product of the spin $\vec{S}$ of the particle and its direction of the momentum $\vec{p}$: $h = \vec{S}\hat{p}$. For massive particles the helicity is not Lorentz invariant.
Figure 2.3 shows the Drell-Yan mass spectrum in the range of the Z mass obtained in the di-muon channel at CMS. The Z bosons are easy to produce if the center of mass energy is next to the boson mass. We see a resonant structure in the spectrum. Away from the resonance we see the expected dependency $\frac{1}{s}$ (remark the double logarithmic scale), where s is the square of the center of mass energy $\sqrt{s}$.
Another difference between weak and electromagnetic interaction is the possibility of the gauge bosons to couple to each other. This due to the non abelian structure of the gauge theory. It is possible to have vertices with three or four Z and $W^{\pm}$ bosons.

Strong interaction

The strong interaction couples only with quarks and not with leptons. The quark model is formulated by a non abelian gauge theory with the symmetry group SU(3). The charge is called color charge in analogy to the color mixture of light. There is a red, green and blue charge and the corresponding anti colors. All physical objects are uncolored. This can be achieved by a mixture of a color with the same anti-color or a mixture of all three colors. This leads to mesons, composed by a quark and an anti-quark, and baryons, composed by three quarks.
The SU(3) implies eight generators which lead to eight various gluons. The generator are the

lambda matrices. The gluons carry a color and an anti-color. Because of the non abelian structure of the SU(3) again couplings between the gluons to each other are possible. There exists three as well as four gluon vertices. The gluons are spin 1 particles.
In contrast to the electromagnetic theory the coupling constant of the strong interaction α_s is not small. This has the effect that only for large momentum transfer it is possible to calculate physical processes of strong interactions by perturbation theory. α_s becomes smaller for large momentum transfer and reaches asymptotically zero. This is called asymptotic freedom. If the momentum transfer is small the higher order contributions $\mathcal{O}(\alpha_s^n)$ become large and approximations are very difficult.
Another effect of the strong interactions is the increase of the potential with increasing distance. This leads to the confinement. If one tries to displace two quarks the requiered energy rises with the distance of the two particles until enough energy is available to generate a new quark/anti-quark pair out of the vacuum.
Looking at hadron collider this leads to the so called hadronization. Instead of single quarks, many particles are produced which form so called jets. The jets consist of hadrons moving in similar direction. The jets itself are reconstructed as objects with properties corresponding to the properties of the leading quarks.

2.2 Heavy quark production

I found an meaningful introduction of heavy quark production in the proceedings of the heavy flavor working group at the HERA-LHC Workshop in 2006 [BBB$^+$06]. The main parts are extracted here:

Perturbation QCD is expected to provide reliable predictions for the production of bottom and (to a lesser extent) charm quarks since their masses are large enough to assure the applicability of perturbation calculations. Anyway a direct comparison of perturbation QCD predictions to heavy flavor production data is not straightforward. Difficulties arise

- from the presence of scales, which are very different from the quark masses that reduce the predictability of fixed-order theory,
- from the non-perturbation ingredients, which are needed to parametrize the fragmentation of the heavy quarks into the observed heavy hadrons and
- from the limited phase space accessible to present detectors.

Moreover a breakdown of the standard collinear factorization approach can be expected at low momenta of the partons. The study of heavy quark production in hadronic interactions together with the nice results ot the electron-proton collisions at HERA has been therefore an active field in the effort to overcome these difficulties and to get a deeper understanding of hard interactions. Besides its intrinsic interest, a precise understanding of heavy quark production is important at LHC because charm and beauty from QCD processes are relevant backgrounds to other interesting processes from the Standard Model (e.g. Higgs to $b\bar{b}$ or beyond). Moreover, theoretical and experimental techniques developed at HERA in the heavy quark field, such as heavy-quark parton densities or b-tagging, are also of great value for future measurements at the LHC.

After exciting years with 'rise and fall of the bottom quark production excess' [Cac04] oil was put on troubled waters and they came up with a rational route for further investigations in this interesting topic on physics.

In my description of the theory behind the analysis I will refer mainly to these proceedings and

summarize the ideas and prospects they made for the measurements at LHC. In the following I will introduce the basics of the analysis and draw the picture of the main physics behind this thesis. I will point to the complex approaches needed for predictions in QCD, and therefore point to the problems in the perturbation as well as in the non-perturbation parts of the calculations. To complete the picture of how confusing measurements of heavy flavor productions were, I will reflect the historical curiosities and finalize this chapter with the paradigm claimed by Matteo Cacciari [Cac04].

2.2.1 b-jet cross section

From the point of view of standard perturbation QCD calculations, the situation has not changed since the beginning of the 90s: fully massive next-to-leading order (NLO) calculations were made available for hadron-hadron, photon-hadron (i.e. photo production) and electron-hadron (i.e. Deep Inelastic Scattering, DIS) collisions. These calculations still constitute the state of the art as far as fixed order results are concerned, and they form the basis for all modern phenomenological predictions.
This statement was given by [BBB$^+$06] on the theory of heavy flavor production in 2006. Therefore the perturbation QCD calculations are the base for a inclusive b-jet cross section measurement. From the experimental point of view the cross section σ is defined by the number of events N produced at a certain integrated luminosity $\int \mathcal{L}$:

$$\sigma = \frac{N}{\int \mathcal{L}}.$$

In this thesis we are interested in the differential cross section of b-jets:

$$\frac{d^2\sigma_{\text{b-jet}}}{dp_T dy} = \frac{\partial^2 N_{\text{b-jet}}}{\partial p_T \partial y \int \mathcal{L}}$$

For the measurement of this quantity the number of b-jets have to be counted in different ranges of the transverse momentum p_T of the jet and its rapidity y. In addition the integrated luminosity has to be measured. The latter was already done by the CMS collaboration [CMS10g]. The remaining part, the analysis of the b-jets was first done during the summer 2010 on the very early data of the CMS experiment [CMS10e]. In this thesis the update and the improvement of the former measurement will be discussed.
Let us start with the already mentioned theoretical base, the perturbation QCD calculations:

2.2.2 QCD predictions

A nice overview of the heavy quark production is given in [FNW03]. The following explanations are extracted from them.
For the heavy flavor production we distinguish three production mechanisms: the flavor creation (FCR), the flavor excitation (FEX) and the gluon splitting (GSP). FCR processes occur already at $\mathcal{O}(\alpha_s^2)$ while FEX and GSP appear primary at $\mathcal{O}(\alpha_s^3)$.

At leading order (LO) we have the following FCR processes:

$$gg \to Q\bar{Q} \qquad q\bar{q} \to Q\bar{Q}$$

where g specifies the gluons, q light quarks and Q the heavy quarks. At next to leading (NLO) order we have the following:

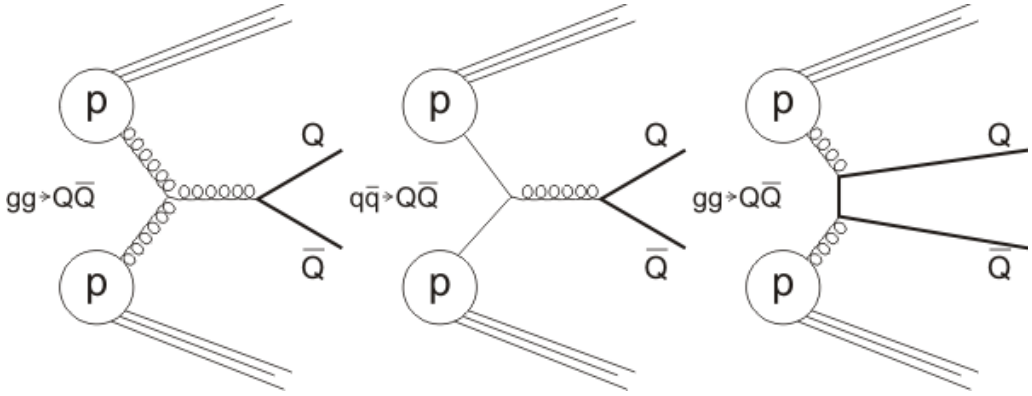


Figure 2.4: Heavy quark Q production mechanisms at leading order (LO). These processes are called flavor creation (FCR).

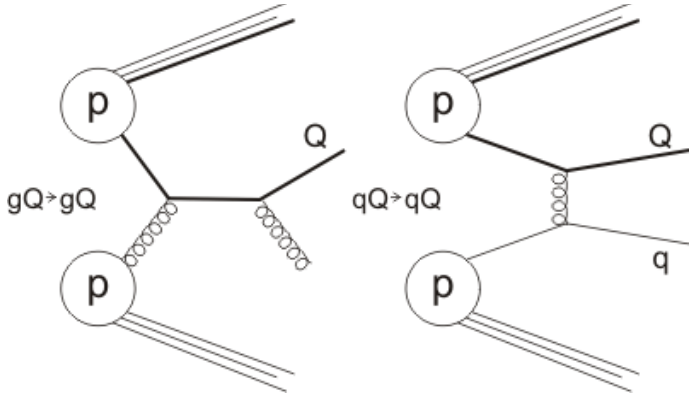


Figure 2.5: At next to leading order the production mechanism are classified in three kinds. Here an example for the flavor excitation (FEX) is illustrated. The processes are assigned to it, if the scattering is accomplished by a heavy quark out of the proton.

$$gg \to Q\bar{Q}g \qquad q\bar{q} \to Q\bar{Q}g \qquad gq \to Q\bar{Q}q \qquad g\bar{q} \to Q\bar{Q}\bar{q}.$$

There we can specify two further kinds of production mechanism. For the FEX we define:

$$qQ \to q\bar{Q} \qquad q\bar{Q} \to q\bar{Q} \qquad g\bar{Q} \to g\bar{Q}$$

and the GSP is a hard $gg \to gg$ process followed by:

$$g \to Q\bar{Q}.$$

Figure 2.4 to 2.6 illustrate the different production mechanisms.

FEX and GSP processes are well defined only in the case of large transverse momenta of the heavy quark. Their extrapolation to the low transverse momentum region can at best be considered a very rough model of higher-order heavy flavor production processes. Figure 2.7 shows the transverse momentum p_T spectrum of the different production mechanism, modeled by a Pythia 6 TuneZ2 event generator. It is nice to see how the guon splitting process becomes more dominant for the high momenta jets.
In [BBB$^+$06] the problems for such a calculation were discussed:
Perturbation calculations of heavy quark production contain badly converging logarithmic terms of quasi collinear origin in higher orders when a second energy scale is present and it is much

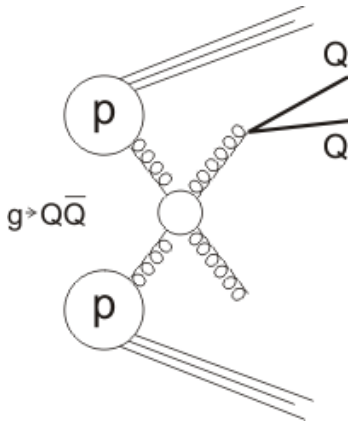


Figure 2.6: The gluons splitting category (GSP) refers to a hard gluon production process which is followed by the intrinsic gluon splitting.

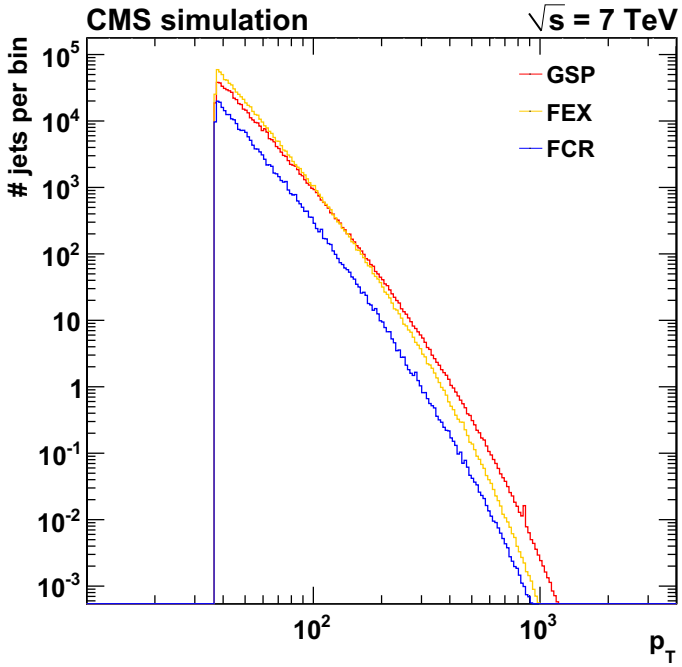


Figure 2.7: Transverse jet momentum spectrum of the different b production mechanisms. For high p_T jets the production is dominated by gluon splitting processes. The plot is done on a Pythia 6 TuneZ2 sample were $p_T > 37$ GeV.

larger than the heavy quark mass m. Examples are the (square root of the) photon virtuality Q^2 in DIS and the transverse momentum p_T in either hadroproduction or photoproduction. Naming generically E the large scale, we can write schematically the cross section for the production of the heavy quark Q as

$$\sigma_Q(E,m) = \sigma_0 \left(1 + \sum_{n=1} \alpha_s^n \sum_{k=0}^{n} c_{nk} \ln^k \left[\frac{E^2}{m^2} + \mathcal{O}\left(\frac{E}{m}\right)\right]\right),$$

where σ_0 is the Born cross section, and the coefficients c_{nk} can contain constants as well as functions of m and E, vanishing as powers of m/E when $E \gg m$. Solving this equation for next to leading order processes needs advanced resummation approaches. Various are developed with the goal to resum the leading logarithms ($\alpha_s^n \ln^n(E^2/m^2)$, LL) and next-to-leading logarithms ($\alpha_s^n \ln^{n-1}(E^2/m^2)$, NOLL).

Over the years, and with increasing experimental accuracies, it however became evident that perturbation QCD alone did not suffice. In fact, real particles - hadrons and leptons - are observed in the detectors, not the quarks and gluons of perturbation QCD. A proper comparison between theory and experiment requires that this gap is bridged by a description of the transition. Of course, the accuracy of such a description will reflect on the overall accuracy of the comparison. When the precision requirements were not too tight, one usually employed a Monte Carlo description to correct the data, deconvoluting hadronization effects and extrapolating to the full phase space. The final experimental result could then easily be compared to the perturbation calculation. This procedure has the inherent drawback of including the bias of our theoretical understanding (as implemented in the Monte Carlo) into an experimental measurement. This bias is of course likely to be more important when the correction to be performed is very large. It can sometimes become almost unacceptable, for instance when exclusive measurements are extrapolated by a factor of ten or so in order to produce an experimental result for a total photoproduction cross section or a heavy quark structure function.
The alternative approach is to present (multi)differential experimental measurements, with cuts as close as possible to the real ones, which is to say with as little theoretical correction and extrapolation as possible. The theoretical prediction must then be refined in order to compare with the real data that it must describe. This has two consequences. First, one has to deal with differential distributions which, in certain regions of phase space, display a bad convergence in perturbation theory. All-order resummations must then be performed in order to produce reliable predictions. Second, differential distributions of real hadrons depend unavoidably on some non-perturbation phenomenological inputs, fragmentation functions. Such inputs must be extracted from data and matched to the perturbation theory in a proper way, pretty much like parton distribution functions of light quarks and gluons are.
To satisfy these claims Stefano Frixione and Bryan R. Webber propose in [FW02] the MC@NLO method for matching the next-to-leading order calculation of a given QCD process with a parton shower Monte Carlo simulation. For almost all analysis on QCD this method is used to compare the measurements with the NLO prediction.

2.2.3 Historical context

A nice overview of the strange historical progress of b quark production measurements can be found in [Cac04]. The main points are listed here:

Measurements of the bottom transverse momentum spectrum at collider began in the late 80s, when the UA1 Collaboration, taking data at the CERN S$p\bar{p}$S with $\sqrt{s} = 546$ and 630 GeV, published

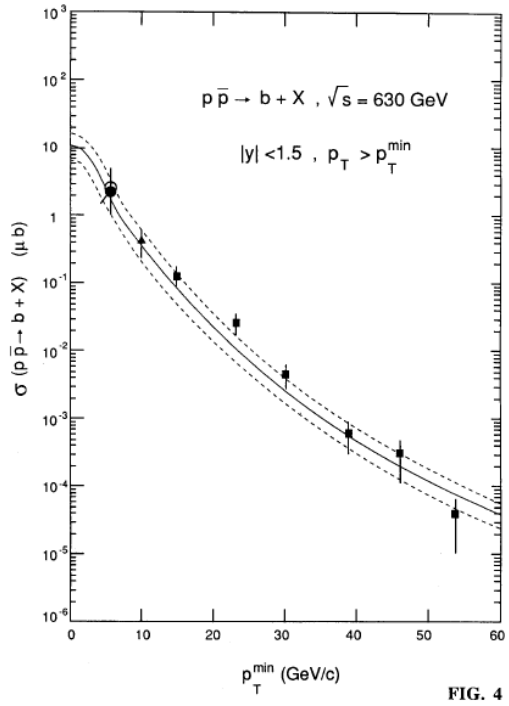


Figure 2.8: UA1 b-quark cross section measurement [A$^+$91]. The experimental points origin from independent measurements: $b \to J/\Psi X$ (solid circle), high mass dimuons (open circle), low mass dimuons (triangles) and muon jets (squares). These results were compared to the then recently completed next-to-leading order calculations (NLO).

results for the $p_T > m_b$ (the bottom quark mass) region.
The UA1 collaboration published two papers about beauty production: [A$^+$87] and [A$^+$91].
Figure 2.8 shows the result from [A$^+$91]. The inclusive b cross section is plotted for rapidity $|y| < 1.5$. The experimental points come from independent measurements: $b \to J/\Psi X$ (solid circle), high mass dimuons (open circle), low mass dimuons (triangles) and muon jets (squares). These results were compared to the then recently completed next-to-leading order (NLO), i.e. order α_s^3, calculation ([NDE88] and [NDE89]), and were found to be in good agreement.

During the 90s the CDF and D0 Collaborations also measured the bottom quark p_T distribution in $p\bar{p}$ collisions at the Fermilab Tevatron at $\sqrt{s} = 1800$ GeV. The main difference to the UA1 measurements is that they measured mainly the b cross sections out of the production rates of specific B hadrons. This includes more difficult parts of non perturbation QCD which are needed to describe the hadronization process and were not well modeled at that time.

There were seven papers published by the CDF collaboration:

- Measurement of the B-meson and b-quark cross sections at $\sqrt{s}$ =1.8 TeV using the exclusive decay $B^\pm \to J/\Psi K^\pm$ [A$^+$92]
- Measurement of the bottom quark production cross section using semileptonic decay electrons in $p\bar{p}$ collisions at $\sqrt{s}$ =1.8 TeV [A$^+$93b]
- Measurement of bottom quark production in 1.8 TeV $p\bar{p}$ collisions using muons from b-quark decays [A$^+$93a]
- Measurement of the B meson and b quark cross sections at $\sqrt{s}$ =1.8 TeV using the exclusive decay $B^0 \to J/\Psi K^*(892)^0$ [A$^+$94]

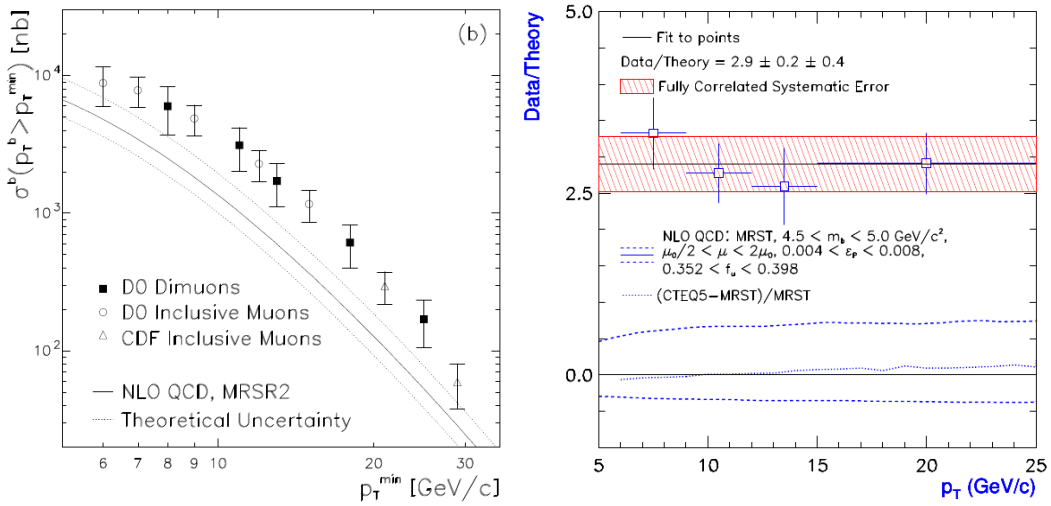


Figure 2.9: b quark production cross section for $|y_b| < 1.0$ compared with the inclusive single muon results and the NLO QCD prediction. On the right is the result of the B^+ meson differential cross section measurements from CDF normalized to the NLO predictions.

- Measurement of the B Meson Differential Cross Section $d\sigma/dp_T$ in $p\bar{p}$ Collisions at $\sqrt{s}$=1.8 TeV [A$^+$95b]
- Measurement of the B^+ total cross section and B^+ differential cross section $d\sigma/dp_T$ in $p\bar{p}$ collisions at $\sqrt{s}$=1.8TeV [A$^+$02a]
- Measurement of the ratio of b quark production cross sections in $p\bar{p}$ collisions at $\sqrt{s}$=630GeV and $\sqrt{s}$=1800GeV [A$^+$02b]

Three papers were published by the D0 collaboration:

- Inclusive μ and b-Quark Production Cross Sections in $p\bar{p}$ Collisions at $\sqrt{s} = 1.8$ TeV [A$^+$95a]
- Small-Angle Muon and Bottom-Quark Production in $p\bar{p}$ Collisions at $\sqrt{s} = 1.8$ TeV [A$^+$00a]
- The $b\bar{b}$ production cross section and angular correlations in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV [A$^+$00b]

In figure 2.9 results of the b cross section measurements done at Tevatron Run1 are shown. The left is taken from [A$^+$00b] and shows the b quark production cross section for $|y_b| < 1.0$ compared with the revised inclusive single muon results and the NLO QCD prediction. The error bars on the data represent the total error. The theoretical uncertainty shows the uncertainty associated with the factorization and renormalization scales and the b quark mass. Also shown are the inclusive single muon data from CDF [A$^+$93a]. On the right hand side of the figure the result of the B^+ meson differential cross section measurements from CDF normalized to the NLO predictions is shown [A$^+$02a]. Both plots show a large discrepancy between data and prediction.
Apparently at odds with the UA1 results, the Tevatron data seemed to display an excess with respect to NLO QCD predictions.

At the same time, rates for bottom production that appeared higher than QCD predictions were also observed in the so called $\gamma\gamma$ collisions by three LEP experiments: L3 ([A$^+$01],[A$^+$05]), OPAL [C$^+$00] and DELPHI [Sil04]. A $\gamma\gamma$ collison at electron positron collider means that both initial particles remain after the interaction. The collison is happens by interchanging photons γ.

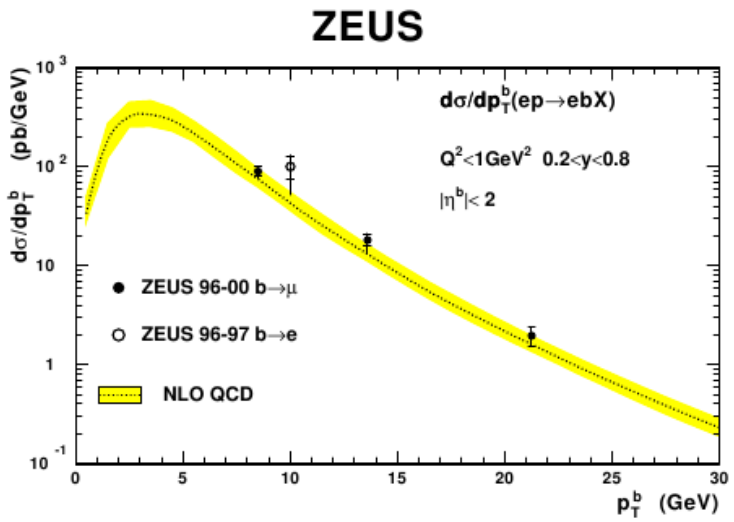


Figure 2.10: Photoproduction of beauty quarks in events with two jets and a muon. The filled points show the ZEUS results from this analysis and the open point is the previous ZEUS measurement in the electron channel [B$^+$01]. The full error bars are the quadratic sum of the statistical (inner part) and systematic uncertainties. The dashed line shows the NLO QCD prediction with the theoretical uncertainty shown as the shaded band.

Also the differences were found by the H1 [A$^+$99] and ZEUS [B$^+$01] Collaborations in ep collisions at HERA.

All these analyses measured the open beauty production in $\gamma\gamma$ collisions. This high order process is needed to be sensitive on the pQCD NLO calculation at electron collider.

But despite this seemingly overwhelming evidence of a excess of b quarks, theorists argue that QCD is instead rather successful in predicting bottom production rates. Improved theoretical analyses and more recent experimental measurements by the CDF and ZEUS Collaborations support this claim, which is also borne out by a critical reconsideration of previous results.

At ZEUS they measured the photoproduction of beauty quarks in events with two jets and a muon [C$^+$04]. The resulting b cross section is shown in figure 2.10. The improved theoretical predictions are comparable with the data.

In Tevatron Run 2 two results were presented by the CDF collaboration. One covers the inclusive b cross section [CDF05] and the other measured the $b\bar{b}$ di-jet production [CDF07]. Both show a good agreement with the NLO predictions (Figure 2.11 and 2.12).

Finally we have results from the LHC. Until now the data seems to be in the predicted regions although we reached already the regions not covered by the Tevatron experiments, where $p_t > 400$ GeV or the rapidity y is in very forward direction. The forward region was explored by the LHCb experiment [A$^+$10].

Further results are published by the CMS collaboration. They have performed measurements of the inclusive b-hadron production cross section with muons [K$^+$11b] and the $B^{\pm}$ production cross section [K$^+$11a] were published. Also studies on the angular correlations of two b quarks were analyzed [K$^+$11c].

The inclusive b-jet cross section measurement is presented in this thesis in chapter 6. For this until now only a preliminary result exists [CMS10e].

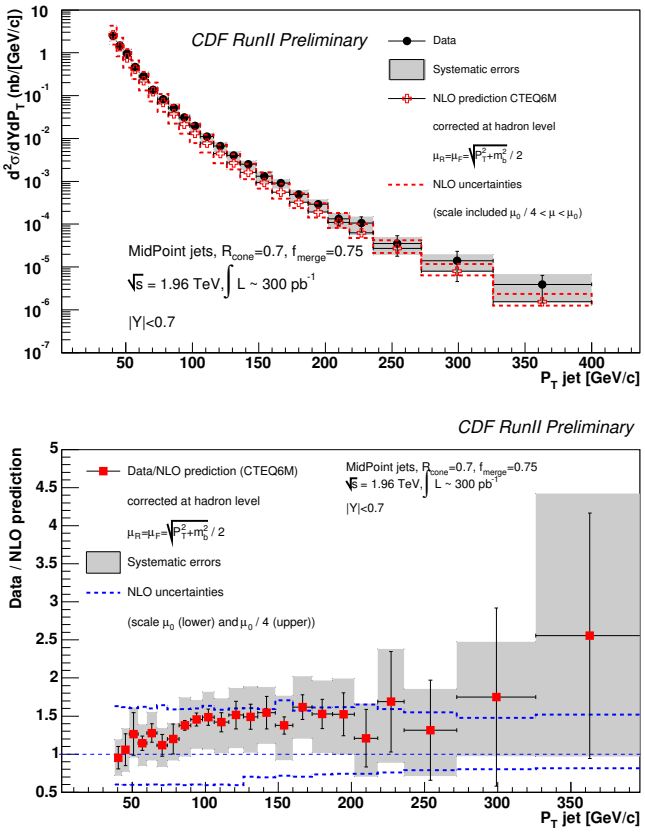


Figure 2.11: The upper plot shows the inclusive b-jet cross section over a Pt range between 38 and 400 GeV measured at CDF Run2. In the lower the same is plotted relative to the theory predictions.

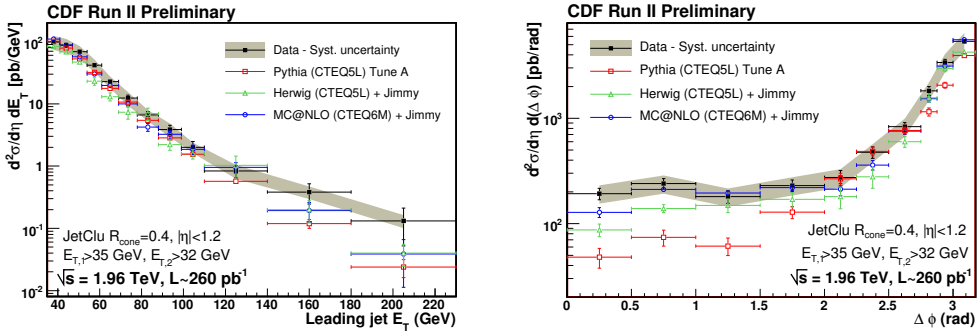


Figure 2.12: The differential $b\bar{b}$ cross section as a function of the leading jet E_t is plotted on the left. On the right the angle between the two b quarks is shown. This distribution is sensitive to the fractions of the three production mechanisms FCR, FEX and GSP.

2.2.4 Conclusion

Looking back the measurement of b quark productions rates causes a few ambiguities and misinterpretations. Therefore Matteo Cacciari formulated in [Cac04] a paradigm which points the problems in comparisons between measurements and prediction and proposes an procedure for analysis on the b quark production:

> 'We shall take NLO QCD calculations as a benchmark for comparisons. We shall require the experimental measurements to be genuine observable quantities. By this we mean that as a matter of principle we do not wish to compare data for, e.g., b-quark p_T distributions, since such a quantity is clearly an unphysical one: the quark not being directly observed, its cross sections have to be inferred rather than directly measured.
>
> A meaningful comparison will therefore be one between a physical cross section and a QCD calculation with at least NLO accuracy. Non-perturbation information, where needed, will have to be introduced in a minimal and self-consistent way. This means that we refrain from using unjustified models, and we shall only include non-perturbation information that has been extracted from one experiment and then employed in predicting another observable, using the same underlying perturbation framework in both cases. Such a precaution allows for a good matching between the perturbation and the non-perturbation phases, a necessity in that only the combination of the two steps leads to an unambiguous measurable quantity.
>
> In practice, the non-perturbation information relative to the hadronization of the b- quarks into B-hadrons is extracted from LEP data with a calculation which has NLO + NLL accuracy. ... the LEP (or SLD) data are translated to Mellin moments space, and only the moments around $N = 5$ are fitted. This ensures that it is the relevant part of the non-perturbation information which is properly determined. These non-perturbation moments are then used together with a calculation having the same perturbation features, FONLL (Fixed Order plus Next-to-Leading Log - in this case $\log(p_T^2/m_b^2)$), to evaluate the cross sections in $p\bar{p}$ collisions.
>
> The expectation is then that total cross sections be reproduced by the NLO calculations for b quarks, and that differential distributions for B hadrons be correctly described by a proper convolution of the FONLL perturbation spectrum for b quarks and the non-perturbation information extracted from LEP data. Notice that a minimalist use of non-perturbation information is made: there is no attempt to fully describe the hadronization process. Only the relevant phenomenological information is determined from data and used in the predictions.
>
> A successful comparison will see data and theory in agreement within their combined uncertainties. The theoretical ones will be assessed by varying as extensively as reasonable the parameters and the unphysical scales entering the predictions. As for the experimental errors, it is perhaps worth reminding that only 1-sigma errors are usually shown on the plots, so that non-overlapping bands do not necessarily point to a solid disagreement.'

Chapter 3

The CMS experiment

In this section I will outline the multifarious publication of the facilities, which stick together with the successful realization of this analysis. Starting with one of the most important center of scientific research, where the accelerator and experiment are built, I will go step by step into more details until I reach the single detector components, which are relevant for my studies.
The following chapter is a summary with extractions of the public CERN web page: `cern.ch` and in addition parts of the most informative papers I found to give a complete impression of the large effort, which was made to realize such a project.

3.1 CERN - Conseil Europeen pour la Recherche Nucleaire

'CERN, the European Organization for Nuclear Research, is one of the worlds largest and most respected centers for scientific research. Its business is fundamental physics, finding out what the Universe is made of and how it works. At CERN, the worlds largest and most complex scientific instruments are used to study the basic constituents of matter - the fundamental particles. By studying what happens when these particles collide, physicists learn about the laws of Nature.
The instruments used at CERN are particle accelerators and detectors. Accelerators boost beams of particles to high energies before they are made to collide with each other or with stationary targets. Detectors observe and record the results of these collisions.
Founded in 1954, the CERN Laboratory sits astride the Franco-Swiss border near Geneva. It was one of Europes first joint ventures and now has 20 Member States.' [CER08a]
This is the short introduction on their web page. It is addressed to the interested to enter the fascinating world of CERN. Over 50 year CERN is a leader in scientific and technical inventions, which results in various highlights of research.

- 1954 Foundations for European science. CERN was ratified by the 12 founding Member States: Belgium, Denmark, France, the Federal Republic of Germany, Greece, Italy, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom, and Yugoslavia. On 29 September 1954 the European Organization for Nuclear Research officially came into being.
- 1957 The first accelerator began operation. The 600 MeV Synchrocyclotron (SC) was CERNs first accelerator and it provided beams for CERNs first particle and nuclear physics experiments.
- 1959 The PS started up. The Proton Synchrotron (PS) accelerated protons for the first time. With a beam energy of 28 GeV, the PS became host to CERNs particle physics program, and provides beams for experiments to this day.

- 1968 Georges Charpak revolutionized detection. Georges Charpak developed the multiwire proportional chamber, a gas-filled box with a large number of parallel detector wires, each connected to individual amplifiers. Linked to a computer, it could achieve a counting rate a thousand times better than existing detectors. The invention revolutionized particle detection, which passed from the manual to the electronic era. Charpak was awarded by the 1992 Nobel Prize in Physics for his work on particle detectors.

- 1971 The worlds first proton-proton collider. The Intersecting Storage Rings (ISR) produced the worlds first proton-proton collisions, providing CERN with valuable knowledge and expertise for its subsequent colliding-beam projects.

- 1973 Neutral currents are revealed. In an experiment conducted by Andr Lagarrigue and colleagues, an invisible neutrino passed through the Gargamelle bubble chamber at CERN jolting an electron in its wake.

- 1976 The SPS is commissioned. Measuring 7 km in circumference, the Super Proton Synchrotron (SPS) was the first of CERNs giant rings. Built in a tunnel, it was also the first accelerator to cross the Franco-Swiss border. Initially conceived as a proton accelerator with a beam energy of 300 GeV, the SPS operates today at up to 450 GeV, and has handled many different kinds of particles.

- 1983 Discovery of the W and Z particles. In 1983, CERN announced the discovery of the W and Z particles. The discovery was so important that Carlo Rubbia and Simon van der Meer, the two key scientists behind the discovery, received the Nobel Prize in physics only a year after.

- 1986 Heavy-ion collisions begin. CERN began to accelerate heavy ions - nuclei containing many neutrons and protons - in the Super Proton Synchrotron (SPS). The aim was to deconfine the quarks by smashing the heavy ions into appropriate targets.

- 1989 Giant LEP started up. LEP was commissioned in July 1989. During 11 years of research, LEP and its experiments provided a detailed study of the electroweak interaction based on solid experimental foundations. Measurements performed at LEP also proved that there are three - and only three - generations of particles of matter. LEP was closed down on 2 November 2000 to make way for the construction of the LHC in the same tunnel.

- 1990 Tim Berners-Lee invented the Web. Berners-Lee had defined the Webs basic concepts, the URL, http and html, and he had written the first browser and server software.

- 1993 Precise results on matter-antimatter asymmetry. The NA31 experiment at CERN published the first precise results on what is known as direct CP symmetry breaking, which indicates more clearly the physics underlying the phenomenon.

- 1995 First observation of antihydrogen. A team led by Walter Oelert created atoms of antihydrogen for the first time at CERNs Low Energy Antiproton Ring (LEAR) facility. Nine of these atoms were produced in collisions between antiprotons and xenon atoms over a period of three weeks.

- 2002 Capturing antihydrogen atoms. Two CERN experiments, ATHENA and ATRAP, took a major step towards understanding antimatter in 2002 by creating thousands of atoms of antimatter in a cold state.

- 2004 CERN celebrates its 50th anniversary. The inauguration of the Globe in 2004 coincided with the official celebration of CERNs anniversary, attended by representatives of the Organizations 20 Member States including the heads of state of France, Spain and Switzerland.
- 2009/10 The LHC started up.

More details on the highlights can be seen on [CER08a].

3.2 LHC - Large hadron collider

The Large Hadron Collider (LHC) is a gigantic scientific instrument near Geneva, where it spans the border between Switzerland and France about 100 m underground. It is a particle accelerator used by physicists to study the smallest known particles - the fundamental building blocks of all things. It will revolutionise our understanding, from the minuscule world deep within atoms to the vastness of the Universe.
Two beams of subatomic particles called hadrons - either protons or lead ions - will travel in opposite directions inside the circular accelerator, gaining energy with every lap. Physicists will use the LHC to recreate the conditions just after the Big Bang, by colliding the two beams head-on at very high energy. Teams of physicists from around the world will analyse the particles created in the collisions using special detectors in a number of experiments dedicated to the LHC.
There are many theories as to what will result from these collisions, but whats for sure is that a brave new world of physics will emerge from the new accelerator, as knowledge in particle physics goes on to describe the workings of the Universe. For decades, the Standard Model of particle physics has served physicists well as a means of understanding the fundamental laws of Nature, but it does not tell the whole story. Only experimental data using the higher energies reached by the LHC can push knowledge forward, challenging those who seek confirmation of established knowledge, and those who dare to dream beyond the paradigm. [CER08b]

The LHC, the worlds largest and most powerful particle accelerator, is the latest addition to CERNs accelerator complex. It mainly consists of a 27 km ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way.
Inside the accelerator, two beams of particles travel at close to the speed of light with very high energies before colliding with one another. The beams travel in opposite directions in separate beam pipes - two tubes kept at ultrahigh vacuum. They are guided around the accelerator ring by a strong magnetic field, achieved using superconducting electromagnets. These are built from coils of special electric cable that operates in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires chilling the magnets to about 271°C - a temperature colder than outer space! For this reason, much of the accelerator is connected to a distribution system of liquid helium, which cools the magnets, as well as to other supply services.
Thousands of magnets of different varieties and sizes are used to direct the beams around the accelerator. These include 1232 dipole magnets of 15 m length which are used to bend the beams, and 392 quadrupole magnets, each 5-7 m long, to focus the beams. Just prior to collision, another type of magnet is used to squeeze the particles closer together to increase the chances of collisions. The particles are so tiny that the task of making them collide is akin to firing needles from two positions 10 km apart with such precision that they meet halfway!
All the controls for the accelerator, its services and technical infrastructure are housed under one roof at the CERN Control Centre. From here, the beams inside the LHC will be made to collide at four locations around the accelerator ring, corresponding to the positions of the particle detectors.[CER08b]

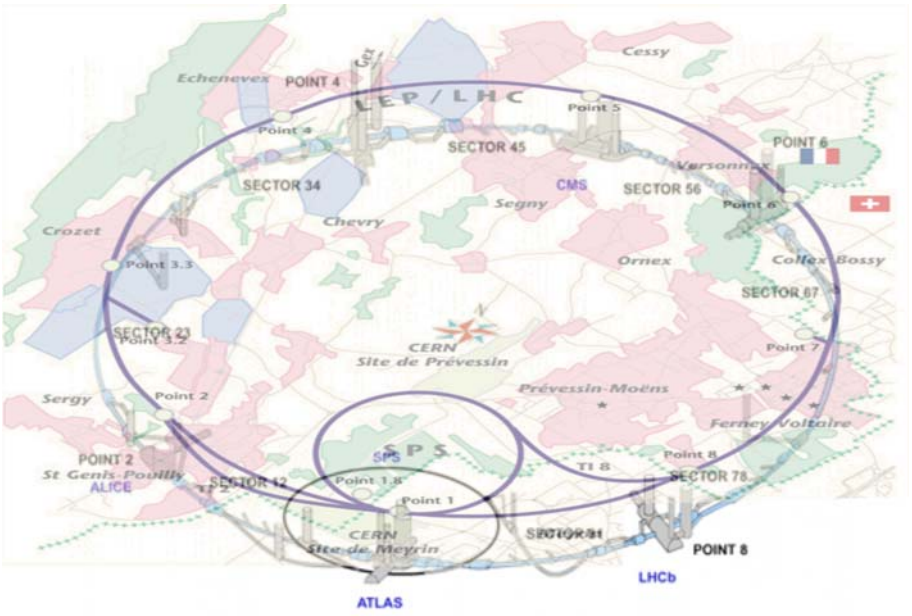


Figure 3.1: Map of the LHC and its hinterland. The red regions correspond to the villages next to CERN. The for experiments are drawn on their location in the accelerator ring.

Figure 3.1 shows the detectors and their location. The six experiments at the LHC are all run by international collaborations, bringing together scientists from institutes all over the world. Each experiment is distinct, characterized by its unique particle detector.

The two large experiments, ATLAS and CMS, are based on general-purpose detectors to analyse the myriad of particles produced by the collisions in the accelerator. They are designed to investigate the largest range of physics possible. Having two independently designed detectors is vital for cross-confirmation of any new discoveries made.

Two medium-size experiments, ALICE and LHCb, have specialized detectors for analyzing the LHC collisions in relation to specific phenomena.

Two experiments, TOTEM and LHCf, are much smaller in size. They are designed to focus on forward particles (protons or heavy ions). These are particles that just brush past each other as the beams collide, rather than meeting head-on

The ATLAS, CMS, ALICE and LHCb detectors are installed in four huge underground caverns located around the ring of the LHC. The detectors used by the TOTEM experiment are positioned near the CMS detector, whereas those used by LHCf are near the ATLAS detector.[CER08b]

3.3 CMS detector - Compact muon solenoid detector

CMS stands for Compact Muon Solenoid: compact because it is small for its enormous weight, muon for one of the particles it detects, and solenoid for the coil inside its huge superconducting magnet. It is a high-energy physics experiment in Cessy, France, part of the Large Hadron Collider (LHC) at CERN.

CMS is designed to see a wide range of particles and phenomena produced in high-energy collisions in the LHC. Like a cylindrical onion, different layers of detector stop and measure the different particles, and use this key data to build up a picture of events at the heart of the collision.

Scientists then use this data to search for new phenomena that will help to answer questions such as: What is the Universe really made of and what forces act within it? And what gives

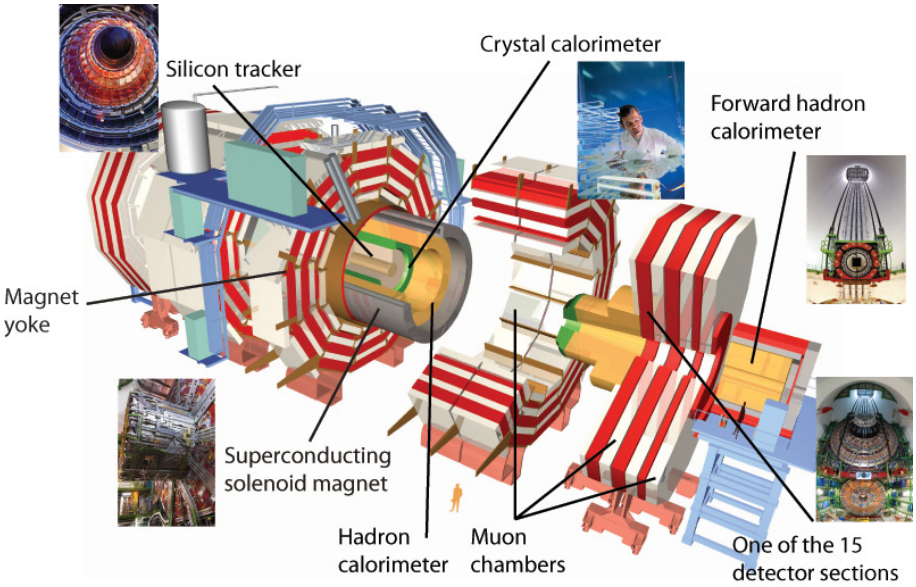


Figure 3.2: model of the CMS detector decorated with pictures of the different detector components [2].

everything substance? CMS will also measure the properties of previously discovered particles with unprecedented precision, and be on the lookout for completely new, unpredicted phenomena. [CMS08b]

Detectors consist of layers of material that exploit the different properties of particles to catch and measure the energy and momentum of each one. CMS was designed around getting the best possible scientific results, and therefore to look for the most efficient ways of finding evidence for new physical theories. This put certain requirements on the design. CMS needed:

- a high performance system to detect and measure muons,
- a high resolution method to detect and measure electrons and photons (an electromagnetic calorimeter),
- a high quality central tracking system to give accurate momentum measurements, and
- a hermetic hadron calorimeter, designed to entirely surround the collision and prevent particles from escaping.

With these priorities in mind, the first essential item was a very strong magnet. The higher a charged particles momentum, the less its path is curved in the magnetic field, so when we know its path we can measure its momentum. A strong magnet was therefore needed to allow us to accurately measure even the very high momentum particles, such as muons. A large magnet also allowed for a number of layers of muon detectors within the magnetic field, so momentum could be measured both inside the coil (by the tracking devices) and outside of the coil (by the muon chambers).
The magnet is the Solenoid in Compact Muon Solenoid (CMS). The solenoid is a coil of superconducting wire that creates a magnetic field when electricity flows through it; in CMS the solenoid has an overall length of 13m and a diameter of 7m, and a magnetic field about 100,000 times stronger than that of the Earth. It is the largest magnet of its type ever constructed and allows

[2] taken from http://bigscience.web.cern.ch/bigscience/en/cms/cms2.html

the tracker and calorimeter detectors to be placed inside the coil, resulting in a detector that is, overall, compact, compared to detectors of similar weight.
The design of the whole detector was also inspired by lessons learnt from previous CERN experiments at LEP (the Large Electron Positron Collider). Engineers found that building sections above ground, rather than constructing them in the cavern with all its access and safety issues, saved valuable time. Another important conclusion was that sub-detectors should be made more easily accessible to allow for easier and faster maintenance.
Thus CMS was designed in fifteen separate sections or slices that were built on the surface and lowered down ready-made into the cavern. Being able to work in parallel on excavating the cavern and building the detector saved valuable time. This slicing, along with the careful design of cabling and piping, also ensures that the sections can be fully opened and closed with minimum disruption, and each piece remains accessible within the cavern.
These considerations, along with the unique conditions of the LHC, affected the design of each layer of the detector. [CMS08b]

3.3.1 Tracking system

The tracking system consists of two main components the pixel detector next to the beam pipe and the silicon strip detectors next to it.

Pixels

Momentum of particles is crucial in helping us to build up a picture of events at the heart of the collision. One method to calculate the momentum of a particle is to track its path through a magnetic field; the more curved the path, the less momentum the particle had. The CMS tracker records the paths taken by charged particles by finding their positions at a number of key points. The tracker can reconstruct the paths of high-energy muons, electrons and hadrons (particles made up of quarks) as well as see tracks coming from the decay of very short-lived particles such as beauty or b quarks that will be used to study the differences between matter and antimatter.
The tracker needs to record particle paths accurately yet be lightweight so as to disturb the particle as little as possible. It does this by taking position measurements so accurate that tracks can be reliably reconstructed using just a few measurement points. Each measurement is accurate to 10 μm, a fraction of the width of a human hair. It is also the inner most layer of the detector and so receives the highest volume of particles: the construction materials were therefore carefully chosen to resist radiation.
The final design consists of a tracker made entirely of silicon: the pixels, at the very core of the detector and dealing with the highest intensity of particles, and the silicon microstrip detectors that surround it. As particles travel through the tracker the pixels and microstrips produce tiny electric signals that are amplified and detected. The tracker employs sensors covering an area the size of a tennis court, with 75 million separate electronic read-out channels: in the pixel detector there are some 6000 connections per square centimeter.[CMS08b]

Silicon Strip Detectors

After the pixels and on their way out of the tracker, particles pass through ten layers of silicon strip detectors, reaching out to a radius of 130 centimeters.
The tracker silicon strip detector consists of four inner barrel (TIB) layers assembled in shells with two inner endcaps (TID), each composed of three small discs. The outer barrel (TOB) consists of

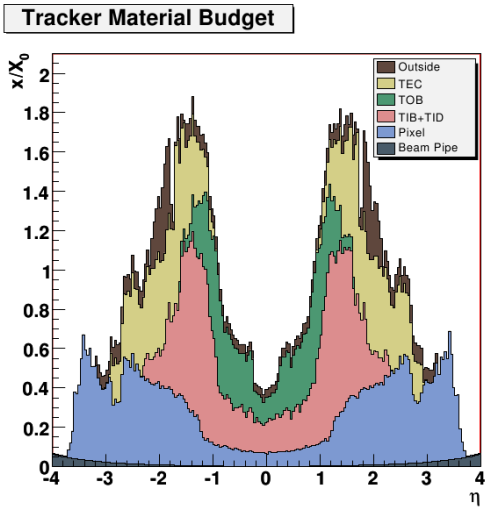


Figure 3.3: model of the CMS detector decorated with pictures of the different detector components.

six concentric layers. Finally two endcaps (TEC) close off the tracker. Each has silicon modules designed differently for its place within the detector.

This part of the tracker contains 15,200 highly sensitive modules with a total of 10 million detector strips read by 80,000 microelectronic chips. Each module consists of three elements: a set of sensors, its mechanical support structure and readout electronics.

Silicon sensors are highly suited to receive many particles in a small space due to their fast response and good spatial resolution. The silicon detectors work in much the same way as the pixels: as a charged particle crosses the material it knocks electron from atoms and within the applied electric field these move giving a very small pulse of current lasting a few nanoseconds. This small amount of charge is then amplified by APV25 chips, giving us hits when a particle passes, allowing us to reconstruct its path.[CMS08b]

Looking at the two tracking detector components it is easy to see that they include a particular amount of material in the innerst regions of the detector. Therefore electrons, photons and pions are stimulated to react before reaching the calorimeters for their energy measurements. This implies a more difficult reconstruction of the physical objects, but opens a new field of particle identification dependent on bremsstrahlung. Figure 3.3 shows the material budget of the CMS tracker in units of radiation length. [CMS09b]

3.3.2 Calorimeter

Outside the tracker are calorimeters that measure the energy of particles. In measuring the momentum, the tracker should interfere with the particles as little as possible, whereas the calorimeters are specifically designed to stop the particles in their tracks.

The Electromagnetic Calorimeter (ECAL) - made of lead tungstate, a very dense material that produces light when hit - measures the energy of photons and electrons whereas the Hadron Calorimeter (HCAL) is designed principally to detect any particle made up of quarks (the basic building blocks of protons and neutrons). The size of the magnet allows the tracker and calorimeters to be placed inside its coil, resulting in an overall compact detector.[CMS08b]

Electromagnetic calorimeter (ECAL)

In order to build up a picture of events occurring in the LHC, CMS must find the energies of emerging particles. Of particular interest are electrons and photons, because of their use in finding the Higgs boson and other new physics.

These particles are measured using an electromagnetic calorimeter (ECAL). But to find them with the necessary precision in the very strict conditions of the LHC - a high magnetic field, high levels of radiation and only 25 nanoseconds between collisions - required very particular detector materials. Lead tungstate crystal is made primarily of metal and is heavier than stainless steel, but with a touch of oxygen in this crystalline form it is highly transparent and scintillates when electrons and photons pass through it. This means it produces light in proportion to the particles energy. These high-density crystals produce light in fast, short, well-defined photon bursts that allow for a precise, fast and fairly compact detector.

Photodetectors that have been especially designed to work within the high magnetic field, are also glued onto the back of each of the crystals to detect the scintillation light and convert it to an electrical signal that is amplified and sent for analysis.

The ECAL, made up of a barrel section and two endcaps, forms a layer between the tracker and the HCAL. The cylindrical barrel consists of 61,200 crystals formed into 36 supermodules, each weighing around three tonnes and containing 1700 crystals. The flat ECAL endcaps seal off the barrel at either end and are made up of almost 15,000 further crystals.

For extra spatial precision, the ECAL also contains Preshower detectors that sit in front of the endcaps. These allow CMS to distinguish between single high-energy photons (often signs of exciting physics) and the less interesting close pairs of low-energy photons.[CMS08b]

Hadron Calorimeter (HCAL)

The Hadron Calorimeter (HCAL) measures the energy of hadrons, particles made of quarks and gluons (for example protons, neutrons, pions and kaons). Additionally it provides indirect measurement of the presence of non-interacting, uncharged particles such as neutrinos.

Measuring these particles is important as they can tell us if new particles such as the Higgs boson or supersymmetric particles (much heavier versions of the standard particles we know) have been formed.

As these particles decay they may produce new particles that do not leave record of their presence in any part of the CMS detector. To spot these the HCAL must be hermetic, that is make sure it captures, to the extent possible, every particle emerging from the collisions. This way if we see particles shoot out one side of the detector, but not the other, with an imbalance in the momentum and energy (measured in the sideways transverse direction relative to the beam line), we can deduce that we are producing invisible particles.

To ensure that we are seeing something new, rather than just letting familiar particles escape undetected, layers of the HCAL were built in a staggered fashion so that there are no gaps in direct lines that a familiar particle might escape through.

The HCAL is a sampling calorimeter, meaning it finds a particles position, energy and arrival time using alternating layers of absorber and fluorescent scintillator materials that produce a rapid light pulse when the particle passes through. Special optic fibers collect up this light and feed it into readout boxes where photodetectors amplify the signal. When the amount of light in a given region is summed up over many layers of tiles in depth, called a tower, this total amount of light is a measure of a particles energy.

As the HCAL is massive and thick, fitting it into compact CMS was a challenge, as the cascades of

particles produced when a hadron hits the dense absorber material (known as showers) are large, and the minimum amount of material needed to contain and measure them is about one meter.
To accomplish this feat, the HCAL is organized into barrel (HB and HO), endcap (HE) and forward (HF) sections. There are 36 barrel wedges, each weighing 26 tonnes. These form the last layer of detector inside the magnet coil whilst a few additional layers, the outer barrel (HO), sit outside the coil, ensuring no energy leaks out the back of the HB undetected. Similarly, 36 endcap wedges measure particle energies as they emerge through the ends of the solenoid magnet.
Lastly, the two hadronic forward calorimeters (HF) are positioned at either end of CMS, to pick up the myriad particles coming out of the collision region at shallow angles relative to the beam line. These receive the bulk of the particle energy contained in the collision so must be very resistant to radiation and use different materials to the other parts of the HCAL.[CMS08b]

3.3.3 Muon detector

As the name Compact Muon Solenoid suggests, detecting muons is one of CMSs most important tasks. Muons are charged particles that are just like electrons and positrons, but are 200 times heavier. We expect them to be produced in the decay of a number of potential new particles; for instance, one of the clearest "signatures" of the Higgs Boson is its decay into four muons.
Because muons can penetrate several meters of iron without interacting, unlike most particles they are not stopped by any of CMSs calorimeters. Therefore, chambers to detect muons are placed at the very edge of the experiment where they are the only particles likely to register a signal.
A particle is measured by fitting a curve to hits among the four muon stations, which sit outside the magnet coil and are interleaved with iron "return yoke" plates. By tracking its position through the multiple layers of each station, combined with tracker measurements the detectors precisely trace a particles path. This gives a measurement of its momentum because we know that particles traveling with more momentum bend less in a magnetic field. As a consequence, the CMS magnet is very powerful so we can bend even the paths of very high-energy muons and calculate their momenta.
In total there are 1400 muon chambers: 250 drift tubes (DTs) and 540 cathode strip chambers (CSCs) track the particles positions and provide a trigger, while 610 resistive plate chambers (RPCs) form a redundant trigger system, which quickly decides to keep the acquired muon data or not. Because of the many layers of detector and different specialities of each type, the system is naturally robust and able to filter out background noise.
DTs and RPCs are arranged in concentric cylinders around the beam line (the barrel region) whilst CSCs and RPCs, make up the endcaps disks that cover the ends of the barrel.[CMS08b]

Chapter 4

Event reconstruction

Recording the collisions with the CMS detector (see 3.3) is the first part of a physics analysis. But it is almost impossible to study physics behaviour on the raw data. Most of the manpower is needed to transform this data into a usable structure. The main goal of this transformation is to reconstruct objects with well defined physics properties. Such objects are tracks, which are mainly reconstructed with the tracking system (see 3.3.1) in the middle of the detector. Tracks are produced by charged particles. Most of them are pions, but also protons, kaons, muons and electrons make a visible signal which is recorded. With the additional information of the calorimeter (see 3.3.2) it is possible to build objects called jets, which correspond to elementary particles produced in the QCD process. To study QCD processes a well understood reconstruction of jet objects is very important. The muon detector (see 3.3.3) helps us to find muon candidates with a large purity. Further it is possible to reconstruct electron candidates, tau candidates or b-jet candidates, which need more advanced algorithms of pattern recognition to identify such objects. The following section will introduce how the recorded data is filtered by the CMS trigger system. Further I present the samples, which are used for this analysis, explain the different physics objects stored in their files and how they are reconstructed within the CMS software framework (CMSSW).

4.1 Trigger system

The trigger system is the important infrastructure which selects the samples for further analysis. It does a rough classification of each event. In doing this the main job is the reduction of the huge amount of data and the dispersion into so called trigger streams. A trigger stream provides an enriched sample of interesting events which are needed for the analysis. A nice description of the trigger system can be found in [A$^+$09]. The main parts in this section are extracted from there. It is summarized as much as possible in spite of getting an introducing idea, how the large amount of data measured by the CMS experiment is filtered and recorded for analysis.
The trigger system consists of two modules: The CMS trigger [B$^+$00] and data acquisition system [CRS02]. They are designed to cope with unprecedented luminosities and interaction rates. At the LHC design luminosity of 10^{34} cm^{-2} s^{-1}, and bunch-crossing rates of 40 MHz, an average of about 20 interactions will take place at each bunch crossing. The trigger system must reduce the bunch-crossing rate to a final output rate of O(100) Hz, consistent with an archival storage capability of O(100) MB/s.
Only two trigger levels are employed in CMS. The first one, the Level-1 Trigger (L1T) [B$^+$00], is implemented using custom electronics and is designed to reduce the event rate to 100 kHz. The second trigger level, the High Level Trigger (HLT), provides further rate reduction by analyzing

full-granularity detector data, using software reconstruction and filtering algorithms running on a large computing cluster consisting of ordinary CPUs, the Event Filter Farm.

4.1.1 Level 1 trigger

In [B+00] the Level 1 trigger is explained in detail: The CMS L1 trigger is based on the identification of muons, electrons, photons, jets, and missing transverse energy. The trigger must have a sufficiently high and understood efficiency at a sufficiently low threshold to ensure a high yield of events in the final CMS physics plots to provide enough statistics and enough efficiency for these events so that the correction for this efficiency does not add appreciably to the systematic error of the measurement.
Given the high event rate at the nominal LHC luminosity, only a limited portion of the detector information from the calorimeters and the muon chambers is used by the L1T system to perform the first event selection, while the full granularity data are stored in the detector front-end electronics modules, waiting for the L1T decision. The overall latency to deliver the trigger signal (L1A) is set by the depth of the front-end pipelines and corresponds to 128 bunch crossings. The L1T processing elements compute the physics candidates (muons, jets, e/γ, etc.) based on which the final decision is taken.
Relevant for this analysis are only the jet trigger. The definition of the level-1 jet trigger can be found in [VPB07]. Level-1 jets are defined using the transverse energy sums in 12x12 calorimeter trigger tower windows. A calorimeter trigger tower is defined as an array of 5x5 crystals in the ECAL of dimensions 0.087x0.087 ($\Delta\eta$x$\Delta\phi$), which corresponds 1:1 to the physics tower size of the HCAL. The algorithm uses a sliding-window technique that steps in units of 4x4 trigger towers, called trigger regions, to give complete (η, ϕ) coverage of the calorimeter. The four highest jets in the central and forward calorimeters, as well as four central τ jets are selected. Also selected are single, double, triple and quad-jet triggers with varying thresholds and prescale factors.

4.1.2 High level trigger

As described in the CMS Technical Design Reports on the DAQ/HLT and on the Physics Performance of the experiment [CMS06], the HLT selection is implemented as a sequence of reconstruction and selection steps of increasing complexity, reconstruction refinement and physics sophistication. The fully programmable nature of the processors in the Event Filter Farm enables the implementation of very complex algorithms utilizing any and all information in the event.
At HLT, jets are reconstructed using an iterative cone algorithm with cone size $R = 0.5$. The algorithm is identical to the one used in the offline analysis. The inputs to the jet algorithm are calorimeter towers, which are constructed from one or more projected HCAL cells and corresponding projected ECAL crystals, and satisfy certain threshold requirements. For inclusion in the jet finding algorithm, the calorimeter towers must have $p_T > 0.5$ GeV and at least one tower must satisfy the jet seed requirement of $p_T > 1$ GeV. After jet finding, a correction for the calorimeter response is applied to the reconstructed jets. This correction was obtained using QCD di-jet events generated by PYTHIA and run through the full CMS detector simulation in CMSSW.
More details on the high level trigger, as well as this short summaray can be found in [VPB07].

4.2 Luminosity measurement

For any cross section analysis the measurement of the luminosity is important. The online and offline methods on measuring the CMS luminosity are summarized in [CMS10g]. The following is

extracted from there:

Online methods The CMS online luminosity measurement employs signals from the forward hadronic calorimeter (HF), which covers the pseudorapidity range $3 < |\eta| < 5$. Two methods for extracting a real-time relative instantaneous luminosity with the HF have been implemented in firmware. The first is based on zero counting, in which the average fraction of empty towers is used to infer the mean number of interactions per bunch crossing. The second method exploits the linear relationship between the average transverse energy per tower and the luminosity. Although all HF towers are outfitted with luminosity firmware, the best linearity is obtained by limiting the coverage to four azimuthal (2π) rings in the range $3.5 < |\eta| < 4.2$. The principal reason for restricting the η range is to avoid non-linearities introduced by averaging the tower occupancy over a range of η rings with very different probabilities for having an occupied tower in a single interaction event. In this case, the average fraction of empty towers becomes a sum over exponentials and is no longer linear with the number of interactions per bunch crossing. The digital outputs of the circuits used to read the signals from the HF PMTs are monitored in a non-invasive way and used to collect channel-occupancy and E_T-sum data in histograms that have one bin for each of the 3564 possible bunch crossings.

Both methods can operate up to the full luminosity of the LHC (10^{34} cm^{-2} s^{-1}). At very low luminosities (10^{25} cm^{-2} s^{-1} and below) the algorithms just described are subject to small noise backgrounds, but were demonstrated to function well for the luminosities delivered by the LHC during the initial stages of the 2010 run. Since the tower occupancy method offers somewhat better performance at the relatively low luminosities delivered by the LHC thus far, it has been adopted as the default method. Results referred to as HF online are based on tower occupancy unless stated otherwise.

Offline methods As a cross check on the HF-based online luminosity monitor, two offline algorithms were developed for luminosity monitoring. One of these methods is based on energy depositions in the HF, while the other makes use of tracking and vertex finding. The offline methods have the drawback of long latency (typically 24 hours elapse before the offline information from a given run is available), but allow for better background rejection than the online methods. Most importantly, the offline techniques employ a largely independent data-handling path, and in the case of the vertex-counting method, involve a completely separate set of systematic uncertainties. They thus complement the online method nicely.

The offline HF method is based on the coincidence of $\sum E_T$ depositions of at least 1 GeV in the forward and backward HF arrays (the sum in each HF runs over all towers). Timing cuts, where $|t_{HF}| < 8$ ns for both HF+ and HF-, are imposed to eliminate non-collision backgrounds.

A second offline method requires that at least one vertex with at least two tracks be found in the event. The z-position of the vertex is required to lie within 150 mm of the center of the interaction region. This method provides good efficiency for minimum bias (MB) events, while suppressing non-collision backgrounds to the few per mil level.

An overview of the increasing luminosity can be seen in figure 4.1. For our analysis we decide to ignore the very first data recorded in the commissioning era. This is arguable because of the small luminosities during these runs (run 132440-135802). The end of the commissioning era is around the end of May. Further the integrated luminosity is plotted on the right.

[2]taken from https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults

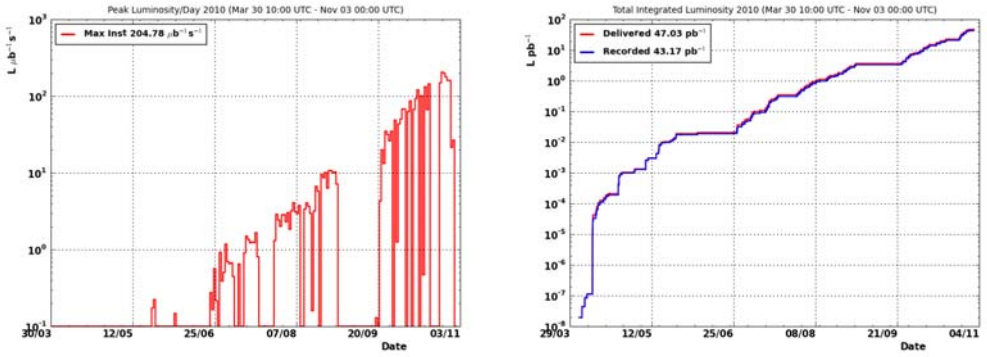


Figure 4.1: Peak luminosity per day for the first year of data taking with the CMS experiment. The period from the first collisions until June is called the commissioning era, the period after Run2010 era. On the right the integrated luminosity is shown. [2]

4.3 Event reconstruction and object identification

This section addresses the reconstruction of physics objects appearing during the proton collisions provided by the LHC. As already explained the measurements of all detector components are filtered by the trigger system (see 4.1) and stored on tape. The data is saved in a so called RAW data format. The information of all detector components is available.
The RAW dataset is the base for almost all physics analysis at CMS. But most do not use it before the reconstruction of physics objects. The creation of the RECO data format is done promptly after the data taking in the grid infrastructure. The grid is a global network of high performance computing centers, placed all over the world, to execute the data streams from the CMS experiment as well as for the other LHC experiments. The processing is done in multiple steps. Finally the reconstructed data is stored distributed all over the world, but still available for all members of the collaboration via the grid.
Using the CMS software, every collaboration member is able to select the physics objects of interest for his analysis. In the majority of the cases the analyst produces relative small files with a flat structure for his studies.
In my case this structure is related to jet objects. A jet is a particular structure in the detector, which was generated by high energetic quarks or gluons created in the QCD process. Each jet is further linked to other physics objects like tracks, electron, muons and vertices. The whole constellation of the different objects is used to analyse the inclusive b cross section.
This section is a summary of all the important publications which studied and explained the reconstruction of the physics objects needed for my analysis. The information is mostly extracted from the papers related to the commissioning of the CMS experiments and its components.

4.3.1 Track reconstruction

The default track reconstruction at CMS is performed by the combinatorial track finder (CTF). Starting from the reconstructed hits, the track reconstruction is decomposed in four logical parts [AMST06]:

- Seed generation
- Pattern recognition, or trajectory building
- Ambiguity resolution

- Final track fit

Triplets of hits in the tracker or pairs of hits with an additional constraint from the beamspot or a vertex are used as initial estimates, or seeds, of tracks [CKKT06]. The seeds are then propagated outward in a search for compatible hits. As hits are found, they are added to the seed trajectory and the track parameters and uncertainties are updated. This search continues until either the limit of the tracker is reached or no more compatible hits can be found, yielding the collection of hits that belong to the track. In the final step, this collection of hits is fit to obtain the best estimate of the track parameters.

The CTF performs multiple iterations. Between each iteration, hits that can be unambiguously assigned to tracks in the previous iteration are removed from the collection of tracker hits to create a smaller collection that can be used in the subsequent iteration. At the end of each iteration, the reconstructed tracks are filtered to remove tracks that are likely fake and to flag the expected purity of the tracks. More details can be found in [CMS10m].

4.3.2 Primary vertex reconstruction

To get a complete overview of the important components of the CMSSW, which are needed for this thesis, the main parts of primary vertex reconstruction are extracted from [CMS10l]:

In the primary vertex reconstruction, the measurements of the location and uncertainty of an interaction vertex are computed from a given set of reconstructed tracks. The prompt tracks originating from the primary interaction region are selected based on the transverse impact parameter significance with respect to the beam line, number of strip and pixel hits, and the normalized track χ^2.

The beam line represents the three-dimensional profile of the luminous region where the LHC beams collide at CMS. The beam line is determined in an average over many events, in contrast to the event-by-event primary vertex which gives the precise position of a single collision. A good measurement of the position and slope of the beam line is an important component of the event reconstruction.

The selected tracks are then clustered based on their z coordinates at the point of closest approach to the beam line. Vertex candidates are formed by grouping tracks that are separated in z by less than a distance $z_{sep} = 1$ cm from their nearest neighbor. Candidates containing at least two tracks are then fit with an adaptive vertex fit to compute the best estimate of vertex parameters such as position and covariance matrix, as well as the indicators of the success of the fit, such as the number of degrees of freedom of the vertex and track weights of the tracks in the vertex. The adaptive vertex fitter does not reject an outlying track; rather it down-weights the outliers with a weight w_i. The weight w_i depends on the compatibility of track i with the vertex, as measured by χ^2 [FWV07]. For a track consistent with the common vertex, its weight is close to 1. The number of degrees of freedom is defined as $ndof = 2\sum_{inTracks} w_i - 3$. It is thus strongly correlated to the ith number of tracks compatible with the primary interaction region. For this reason, the number of degrees of freedom of the vertex can be used to select real proton-proton interactions. The primary vertex resolution depends strongly on the number of tracks used in fitting the vertex and the p_T of those tracks.

Figure 4.2 shows exemplary the distribution of the reconstructed primary vertices from a single run. The plots show the result in one and two dimensions.

[3] taken from [CMS10m]

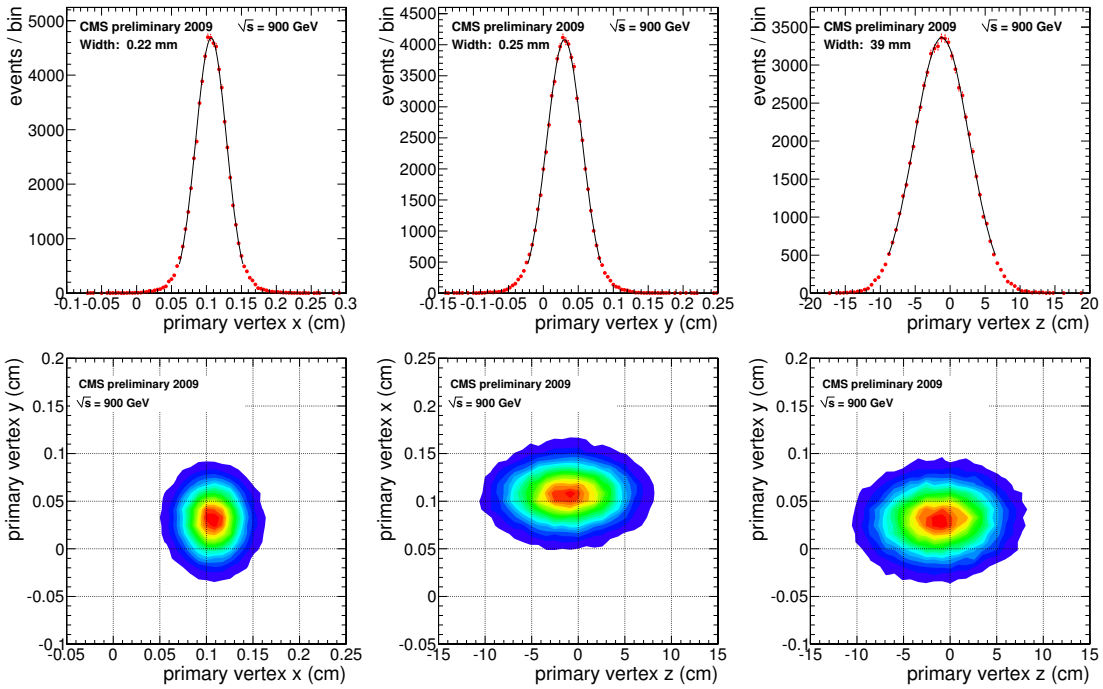


Figure 4.2: Plots of the primary vertex distributions from a single run.[3]

4.3.3 Secondary vertex reconstruction

The secondary vertex reconstruction is extracted from [MPQW06]:
Decay vertices, which result from long living particles are called secondary vertices. Most vertex finders are sensitive to primary (PV) and secondary vertices (SV), so a vertex filter is needed to select only the secondary vertex candidates. The discrimination is based on the distance of a vertex to the beam line or to an already reconstructed primary vertex.
The trimmed Kalman vertex finder [SPF+06] searches for vertex candidates among the input set of tracks, in an iterative way. During the first iteration, a Trimmed Kalman vertex fitter is applied to the complete input set of tracks, yielding as outputs a vertex candidate and a set of tracks which are incompatible with that vertex candidate. During the subsequent iterations, the same procedure is applied to the set of incompatible tracks identified at previous iterations.
The trimmed Kalman vertex finder is sensitive to primary and secondary vertices, so a vertex filter is used to select secondary vertex candidates. The vertex filter uses the following cuts on the vertices:

- The distance from the vertex to the beam line has to exceed 100 μm but must not exceed 2 cm. The lower limit should reject primary vertices, the upper limit photon conversions and nuclear interactions in the beampipe.

- The distance from the vertex to the beam line in the transverse (r-) plane divided by its uncertainty has to be greater than three: $\frac{L_t}{\sigma_{L_t}} > 3$.

- The total invariant mass of the vertex must be smaller than 6.5 GeV/c^2 to discard primary vertices.

- Vertices with two tracks with opposite charge and an invariant mass of the K_0 mass ($\pm$ 50 MeV) are rejected. The 100 μm cut and the 3σ cut on the transverse flight distance are

most important, because they reject most of the primary vertices. The effect of the cut of 6.5 GeV/c^2 on the total invariant mass of the vertex is smaller.

In most cases b-hadrons produce a tertiary vertex because the decay chain proceeds via charm production (the b-c-decay chain). The lifetime and the number of tracks from the decay vertex are smaller for weakly decaying c- than for weakly decaying b-hadrons. For this reason the secondary and the tertiary vertices are merged into one vertex in most cases. If tracks coming from a tertiary vertex are also used to fit the secondary vertex, the measured flight distance is shifted to a higher value. Another effect that corrupts the secondary vertex resolution are misassociated tracks from the primary vertex or from underlying events.

4.3.4 Electron reconstruction

The electron reconstruction of the CMSSW is described in [CMS10d]. The following summary is extracted from there:

Electron reconstruction uses two complementary algorithms at the track seeding stage: tracker driven seeding, more suitable for low p_T electrons as well as performing better for electrons inside jets and ECAL driven seeding.
The ECAL driven algorithm starts by the reconstruction of ECAL superclusters of transverse energy $E_T > 4$ GeV and is optimized for isolated electrons in the p_T range relevant for Z or W decays and down to $p_T > 5$ GeV/c. Supercluster is a group of one or more associated clusters of energy deposits in the ECAL constructed using an algorithm which takes account their characteristic narrow width in the η coordinate and their characteristic spread in ϕ due to the bending in the magnetic field of electrons radiating in the tracker material. As a first filtering step, superclusters are matched to track seeds (pairs or triplets of hits) in the inner tracker layers, and electron tracks are built from these track seeds. Trajectories are reconstructed using a dedicated modeling of the electron energy loss and fitted with a Gaussian Sum Filter (GSF).
The filtering performed at the seeding step is complemented by a preselection. For candidates found only by the tracker driven seeding algorithm, the preselection is based on a multivariate analysis as described in [CMS10b]. For candidates found by the ECAL driven seeding algorithm, the preselection is based on the matching between the GSF track and the supercluster in η and ϕ [BCF$^+$07]. The few ECAL driven electron candidates (1% for isolated electrons) not accepted by these matching cuts but passing the multivariate preselection are also kept.

4.3.5 Muon reconstruction

In the standard CMS reconstruction for pp collisions, tracks are first reconstructed independently in the silicon tracker (tracker track) and in the muon spectrometer (standalone-muon track). Based on these, two reconstruction approaches are used:

- Global Muon reconstruction (outside-in): starting from a standalone muon in the muon system, a matching tracker track is found and a global-muon track is fitted combining hits from the tracker track and standalone-muon track. At large transverse momenta (p_T 200 GeV/c), the global-muon fit can improve the momentum resolution compared to the tracker-only fit.

- Tracker Muon reconstruction (inside-out): in this approach, all tracker tracks with $p_T > 0.5$ GeV/c and $p > 2.5$ GeV/c are considered as possible muon candidates and are extrapolated to the muon system, taking into account the expected energy loss and the uncertainty due to

multiple scattering. If at least one muon segment (i.e. a short track stub made of DT or CSC hits) matches the extrapolated track in position, the corresponding tracker track qualifies as a tracker-muon track.

At low momentum (roughly $p < 5$ GeV/c) this approach is more efficient than the global muon reconstruction, since it requires only a single muon segment in the muon system, while global muon reconstruction typically becomes efficient with two or more segments. The majority of muons from collisions (with sufficient momentum) are reconstructed either as a Global Muon or a Tracker Muon, or very often as both. However, if both approaches fail and only a standalone-muon track is found, this leads to a third category of muon candidates:

- Standalone-muon track only: this occurs only for about 1% of muons from collisions, thanks to the high tracker-track efficiency. On the other hand, the acceptance of this type of muon track for cosmic-ray muons is a factor 10^2 to 10^3 larger, thus leading to a collision muon to cosmic-ray muon ratio that is a factor 10^4 to 10^5 less favorable than for the previous two muon categories.

The results of these three algorithms are merged into a single collection of muon candidates, each one containing information from the standalone, tracker, and global fit, when available. Candidates found both by the Tracker Muon and the Global Muon approach that share the same tracker track are merged into a single candidate. Similarly, standalone-muon tracks not included in a Global Muon are merged with a Tracker Muon if they share a muon segment. Additional muon identification information is stored for each candidate. The combination of different algorithms provides a robust and efficient muon reconstruction. A given physics analysis can achieve the desired balance between identification efficiency and purity by applying a selection based on the muon identification variables. Several standard selections are provided.
The basic selection important for this analysis is the Soft Muon Selection: This selection requires the candidate to be a Tracker Muon, with the additional requirement that a matching segment be found in the outermost station where a segment is expected (based on muon position and momentum), matching both in position and direction with the prediction of the track extrapolation. Segments that form a better match in position with a different tracker track are not considered. These additional requirements are optimized for low p_T (< 10 GeV/c) muons. This selection is presently used in B-physics analyses in CMS, in addition to Global Muons.[CMS10k]

4.3.6 Jets

In [CMS10f] I found a good explanation of the jet reconstruction:
Jets are experimental signatures of quarks and gluons, which are produced in high energy processes such as the hard scattering of partons in pp collisions. Four types of jets are reconstructed at CMS, which differently combine individual contributions from subdetectors to form the inputs to the jet clustering algorithm: calorimeter jets, Jet-Plus-Track (JPT) jets, Particle-Flow (PFlow or PF) jets, and track jets.
In this analysis only PFlow-jets are used. They are reconstructed using the Anti-kT [CSS08] clustering algorithm with the size parameter R = 0.5. In [CMS10c] they claim:
The Particle Flow algorithm combines the information from all CMS sub-detectors to identify and reconstruct all particles in the event, namely muons, electrons, photons, charged hadrons and neutral hadrons. Electrons and muons aside, the particle-flow algorithm can be roughly summarized in the following way. Tracks reconstructed in the central silicon tracker are extrapolated to the electromagnetic (ECAL) and hadron (HCAL) calorimeter. The charged hadron candidates,

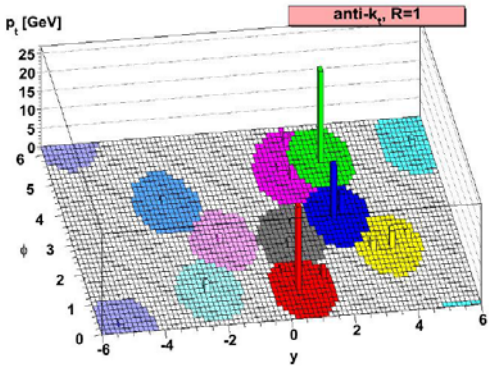


Figure 4.3: Exemplary clustering of the anti-kt jet algorithm. Larger transverse momenta are responsible for more conical clusters [5]

in particular their energies and directions, are reconstructed from these tracks. A track is linked to a calorimetric energy cluster in the ECAL and/or in the HCAL if the track extrapolation falls within the boundaries of one of the energy deposits of the cluster. Photons and neutral hadrons are reconstructed from calorimetric energy clusters: clusters separated from the extrapolated position of tracks in the calorimeters constitute a clear signature of these neutral particles; neutral particles overlapping with charged particles in the calorimeters can be detected as calorimeter energy excesses with respect to the sum of the associated track momenta.

Having the particle flow object they are fed to the anti-kt jet clustering algorithm. In [CSS08] the anti-kt algorithm is described as follows: The functionality of the anti-kt algorithm can be understood by considering an event with a few well separated hard particles with transverse momenta k_{t1}, k_{t2}, ... and many soft particles. Soft particles will tend to cluster with hard ones long before they cluster among themselves. If a hard particle has no hard neighbors within a distance $2R$, then it will simply accumulate all the soft particles within a circle of radius R, resulting in a perfectly conical jet. If another hard particle is present such that $R < \delta_{12} < 2R$ then there will be two hard jets. It is not possible for both to be perfectly conical. If $k_{t1} \gg k_{t2}$ then jet 1 will be conical and jet 2 will be partly conical, since it will miss the part overlapping with jet 1. Instead if $k_{t1} = k_{t2}$ neither jet will be conical and the overlapping part will simply be divided by a straight line equally between the two. Similarly one can work out what happens with $\delta_{12} < R$. Here particles 1 and 2 will cluster to form a single jet. If $k_{t1} \gg k_{t2}$ then it will be a conical jet centered on k1 . For $k_{t1} \sim k_{t2}$ the shape will instead be more complex, being the union of cones (radius $< R$) around each hard particle plus a cone (of radius R) centered on the final jet. Figure 4.3 shows the ϕ/η plane with an exemplary clustering of jets by the anti-kt algorithm.

CMS has developed jet quality criteria (Jet ID) for calorimeter jets and PFlow jets which are found to retain the vast majority of real jets in the simulation while rejecting most fake jets arising from calorimeter and/or readout electronics noise. These are studied in pure noise non-collision data samples such as cosmic trigger data or data from triggers on empty bunches during LHC operation. The PFlow jets are required to have a charged hadron fraction CHF > 0.0 if within the tracking fiducial region of $|\eta| < 2.4$, a neutral hadron fraction NHF < 1.0, a charged electromagnetic (electron) fraction CEF < 1.0, and a neutral electromagnetic (photon) fraction NEF < 1.0. These requirements remove fake jets arising from spurious energy depositions in a single sub-detector. In the studies presented jets are required to pass Jet ID criteria.

[5] taken from [CSS08]

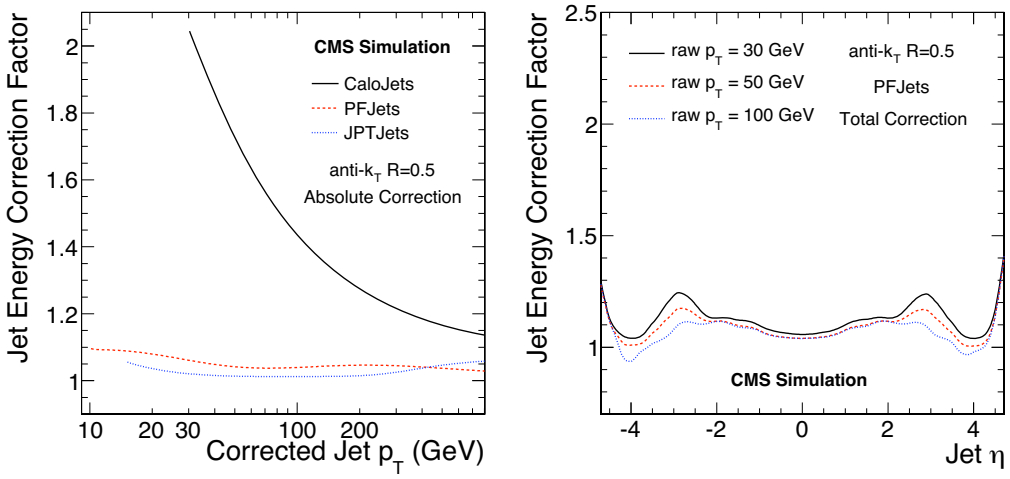


Figure 4.4: Jet energy corrections applied on the PFjets. The multi-step procedure for MC-truth jet energy corrections applies absolute (left), relative corrections (right). [7]

4.3.7 Jet energy corrections

Jet energy measured in the detector is typically different from the corresponding particle jet energy. The latter is obtained in the simulation by clustering, with the same jet algorithm, the stable particles produced during the hadronization process that follows the hard interaction. The main cause for this energy mismatch is the non-uniform and non-linear response of the CMS calorimeters. Furthermore, electronics noise and additional pp interactions in the same bunch crossing (event pile-up) can lead to extra unwanted energy. The purpose of the jet energy correction is to relate, on average, the energy measured in the detector to the energy of the corresponding particle jet. The information on the jet energy correction is extracted from [CMS08a] and [CMS10f]:

CMS has developed a factorized multi-step procedure for the jet energy calibration (JEC). The following three subsequent (sub-)corrections are devised to correct calorimeter, PFlow and JPT jets to the corresponding particle jet level: offset, relative and absolute corrections. The offset correction aims to correct the jet energy for the excess unwanted energy due to electronics noise and pile-up. The relative correction removes variations in jet response versus jet η relative to a central control region chosen as a reference because of the uniformity of the detector. The absolute correction removes variations in jet response versus jet p_T. CMS pursues two complementary approaches to determine the jet energy correction factors: utilizing MC truth information (MC truth JEC), and using physics processes from pp collisions for in-situ jet calibration. At the current initial stage of LHC running, MC truth JEC is used to correct jets in both data and MC simulation. In figure 4.4 the two correction steps for the MC-truth jet energy corrections are shown. The offset corrections are not factorized out.

Current physics analyses in CMS use 5% JEC uncertainties for PFlow jets, with an additional 2% uncertainty per unit rapidity.

4.3.8 Jet Flavor definition

There is no unambiguous answer to the correct underlying flavor of a reconstructed jet. Three definitions are used, reflecting three different points of view:

[7] taken from [CMS10f]

Physics definition Reconstructed jets are matched to initial partons from the primary physics process. They must be within the reconstructed jet cone with $\Delta R < 0.3$. For example, for $t\bar{t}$ events, the initial partons would be two b-jets from the decays of the top quarks, two non-b-jets per hadronic W decay, and no initial gluon jets. There is no matching if hard (FS) radiation occurred and the parton direction changes significantly. No flavor is assigned, if no unambiguous answer is possible when more than one initial parton is matched. Gluon jets splitting to c- or b-quarks are labeled as gluon-jets.

Algorithmic definition The parton that most likely determines the properties of the jet defines the true flavor of the jet. The final state partons, after showering and radiation, are analyzed. The partons must be within $\Delta R < 0.3$ of the reconstructed jet cone. Jets from radiation are matched with full efficiency. If there is a b-quark or a c-quark within the jet cone, it is labeled accordingly, otherwise the jet is assigned with the flavor of the hardest parton.

Energetic definition This definition applies to generated jets (GenJets), where the constituents of a jet are a set of generator objects (GenParticleCandidate). A variable is built for each jet computing the fraction of the energy of the jet which comes from b or c hadrons. These quantities can be used to attribute a flavor of the GenJet. A matched reconstructed jet can get the same flavor as the matched GenJet.

The main differences between the definitions effect mainly jets from gluon splitting. Only physics and energetic definitions see gluon splitting. The algorithmic definition is blind to it. Further the algorithmic definition causes some contamination from gluon splitting to b and c jets. All the three definitions can be applied to GenJets. Only the first 2 (Physics and Algorithmic definitions) can be applied to particle flow jets. [8]

For b-jet tagging the algorithmic definition is used.

4.3.9 b jet tagging

The CMS software contains already various b-jet tagging algorithms for different purposes. I took the summarized description of the different taggers from [CMS09a]. Each tagger produces an output value for each jet. The output of any algorithm is the so called discriminator, defined as a single number which the user can cut on to select different regions in the efficiency versus purity phase space. The discriminator can be a simple physics quantity like the IP significance for some taggers, or a complex variable like the output of likelihood ratio or neural network.

Track counting (TCHE,TCHP) The simplest way of producing a discriminator based on track impact parameters is an extension of the so called track counting algorithm. The track counting approach identifies a jet as a b-jet if there are at least N tracks each with a significance of the impact parameter exceeding S. This algorithm has two major parameters (N and S). The way of producing a continuous discriminator for this algorithm is to fix the value of N, and consider as discriminating variable the impact parameter significance of the Nth track (ordered in decreasing significance). If one is interested in a high efficiency for b-jets, the second track can be used; for higher purity selections the third track is a better choice. The discriminators obtained in this way are plotted for QCD events in Figure 4.5, and are simply the IP significance shapes for the chosen track.

[8] https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools
[9] taken from [CMS10a]

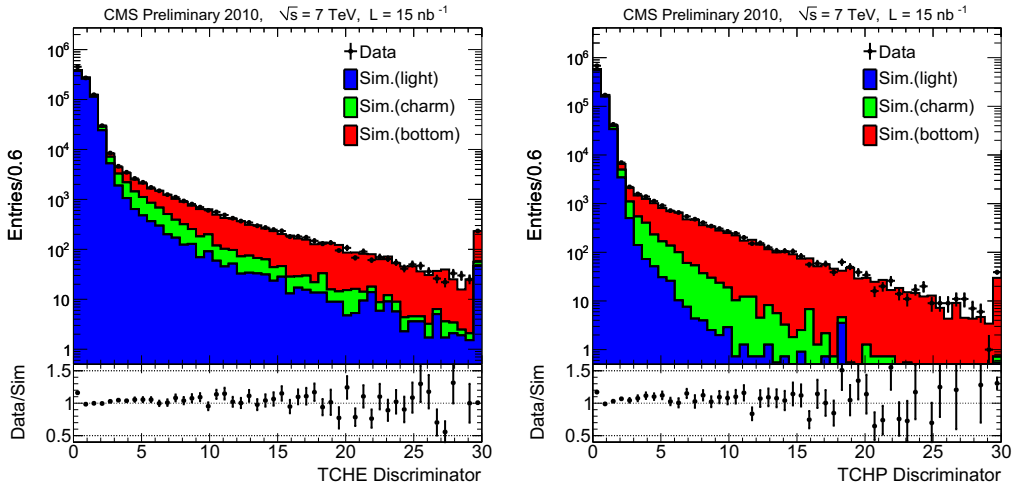


Figure 4.5: Discriminator of the track counting b-jet tagger[9]

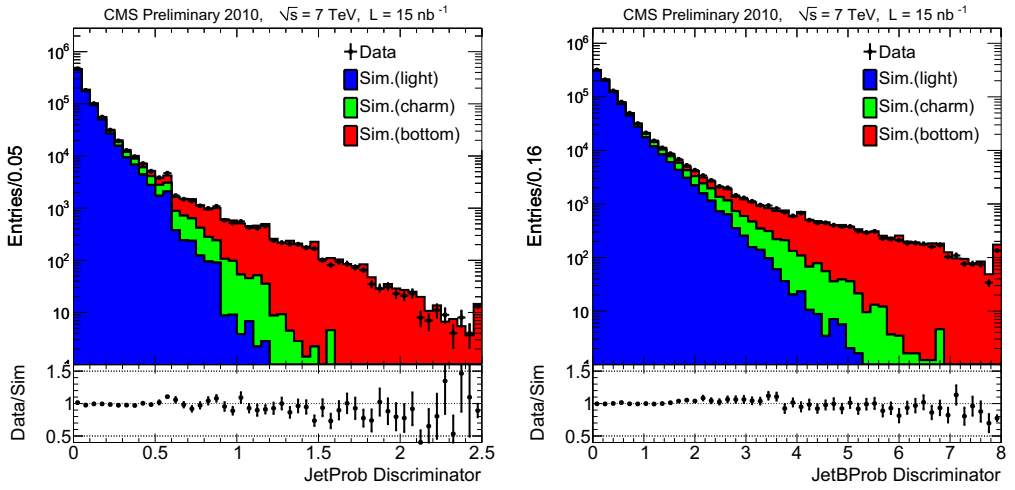


Figure 4.6: Discriminator of the jet probability b-jet tagger[10]

Jet probability (JPT,JBPT) The jet probability algorithms are a natural extension of the track counting algorithms [CMS09a]. The idea is to combine the information coming from all selected tracks.

For each track, the probability to come from the primary vertex is computed and these probabilities are combined to provide the jet probability. The track probability distribution is calibrated by means of the distribution of track impact parameters with negative signs. The negative part of the impact parameter distribution is used for this purpose because it is mainly made up of primary vertex tracks. The advantage of this method with respect to track counting is the fact that a single discriminator is used (i.e. there is no need to choose) and that information from all tracks is used at the same time. [RPS06] Two discriminators are provided; the first labeled `jet probability` is strictly related to the combined probability that all the tracks in the jet come from the primary vertex. The second, labeled `jet B probability` estimates how likely it is that the four most displaced tracks are compatible with the primary vertex; the selection comes from the fact that the average charged track multiplicity in weak b hadron decay is 5, and from the average track reconstruction efficiency, around 80% for tracks in jets. The shapes of the discriminant variable are presented in Figure 4.6.

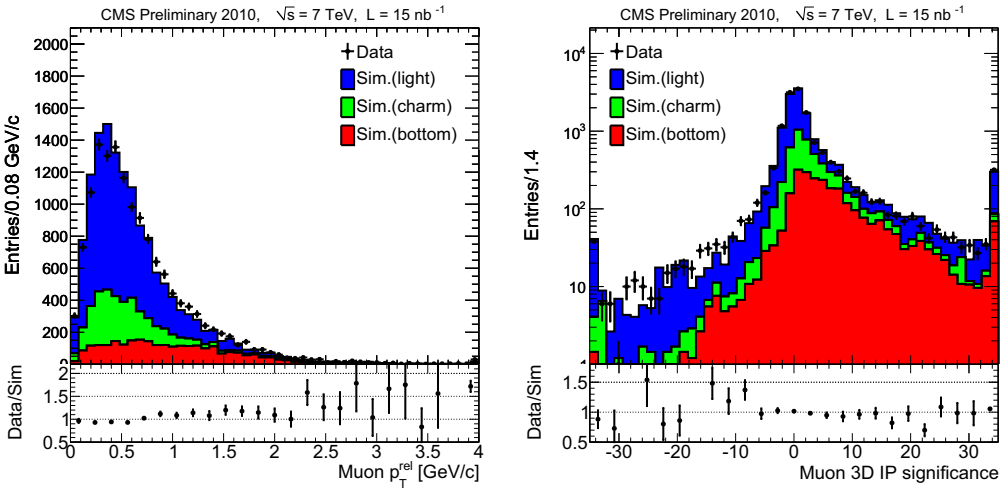


Figure 4.7: Input variables for the soft muon b-jet tagger[11]

Soft muon (SMT) The presence of a muon close to the jet is already a hint of a weak decay of a B hadron. This can be complemented with some additional quantity, in order to build a discriminator. In the `soft muon by` p_T `rel` algorithm the p_T of the muon with respect to the jet axis is used [CMS09a]; harder cuts yield higher purities. In the `soft muon by IP significance` the IP significance of the muon is used instead, but only when found to be positive. In all the cases, when more than one muon is reconstructed, the one with the highest discriminator value is used. Figure 4.7 shows shapes of the p_T and the IP significance of the soft muons, which is used to generate those taggers.

Soft electron (SET) It is also possible to create a soft electron b-jet tagger. Because of the large number of pions appearing in each event it is not possible to get a b-jet tagger based on pure soft electrons similar to the soft muon case. At the moment there is no official soft electron b-jet tagger at CMS. Anyhow a NeuroBayes soft electron tagger was developed in [Mar09].

Simple secondary vertex (SSV) Secondary vertices can be used to select jets from B hadrons with high purity. A simple version, called `simple secondary vertex` tagging algorithm is based upon the reconstruction of at least one secondary vertex. If no such vertex is found, the algorithm returns no discriminator, limiting its maximum b-jet efficiency to the probability of finding a vertex in the presence of weak B hadron decay (around 60-70%). The significance of the 3D flight distance is used as a discriminating variable for this tagger [CMS09a]. Two variants based on the minimum number of tracks attached to the vertex are considered: $N_{trk} \geq 2$ yields the `high efficiency` version (SSVHE). Further there is a `high purity` version (SSVHP), where $N_{trk} \geq 3$. [CMS10a] The distribution of this discriminator is shown Figure 4.8

Combined secondary vertex (CSV) A more complex approach involves the use of secondary vertices, together with other lifetime information, like the IP significance or decay lengths. By using these additional variables, the `combined secondary vertex` algorithm provides discrimination even when no secondary vertices are found, so the maximum possible b-tagging efficiency is not limited by the secondary vertex reconstruction efficiency [CMS09a]. In many cases, tracks with an

[10] taken from [CMS10a]
[11] taken from [CMS10a]
[12] taken from [CMS10a]

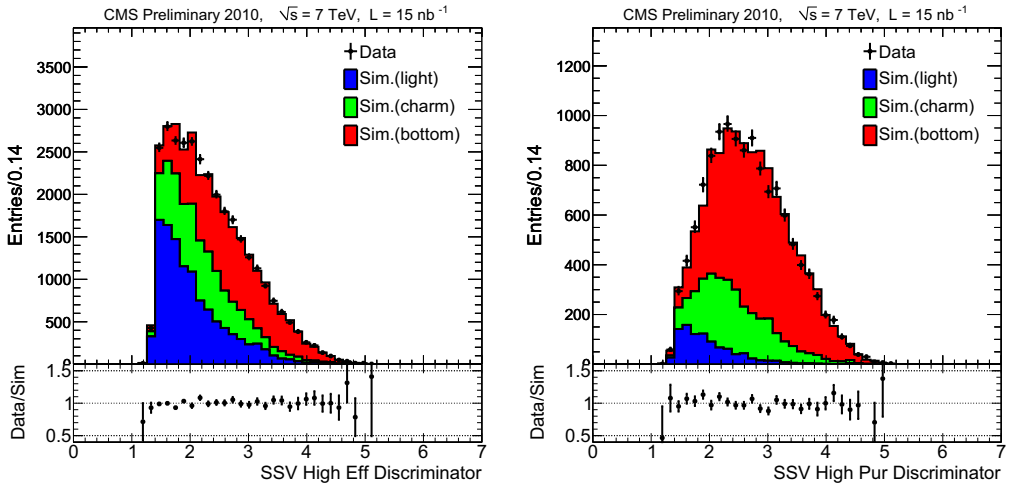


Figure 4.8: Discriminator for the simple secondary vertex b-jet tagger[12]

IP significance > 2 can be combined in a so-called pseudo vertex, allowing for the computation of a subset of secondary vertex based quantities even without an actual vertex fit. When even this is not possible, a no vertex category reverts simply to track based variables similarly to the jet probability algorithm. These variables are used as input to a Likelihood Ratio, used twice to discriminate between b- and c-jets and between b- and light jets, and then combined additively with a factor of 0.75 and 0.25 respectively. For the commissioning of the b-jet tagger, the combined secondary vertex algorithm was not incorporated. Because of its complex structure it needs a larger amount of data for its initiation.

Figure 4.9 shows the performance of the CMS b-jet taggers. The performance is calculated on the same Monte Carlo samples which are used for further comparisons in this thesis. On the y-axis the mistag rate is plotted. The x-axis shows the b-jet efficiency. The best point to perform is the lower right corner. As expected the more complex taggers, jet b probability and the combined secondary vertex, are more performant.

Tagging efficiency

Plans on how to measure the b-tagging efficiencies are presented in [CMS07b]. The following section is extracted from there. All the b-jet tagging algorithms rely upon the reconstruction of lower level objects like tracks, vertices, and jets, which might make it difficult for the Monte Carlo simulation to exactly reproduce the performance in data. The Tevatron collider experiments have developed methods to measure the performance of the lifetime tagging algorithms in collider data. The CMS collaboration adapted these methods to measure the b-tagging efficiency using data, where jets with muon appear. The `pTrel Method` relies directly on a fit to the $p_{T,rel}$ distribution of the muon before and after tagging the muon-jet; the `Counting Method` also relies on $p_{T,rel}$ fits but uses additional information derived from the data. The third method, `System8 Method`, consists of solving a system of eight equations constructed from the total number of events in two samples with different b-jet content, before and after tagging with two b-tagging algorithms.

pTrel Method The basic idea of the pTrel method is to measure the b-quark content of a muon+jet sample by fitting the $p_{T,rel}$ distribution of the muons to a linear combination of the b-quark and c/light-quark jet templates. The process is repeated after tagging the muon-jet. The b-tagging efficiency is calculated as the ratio between the number of b jets after and before tagging,

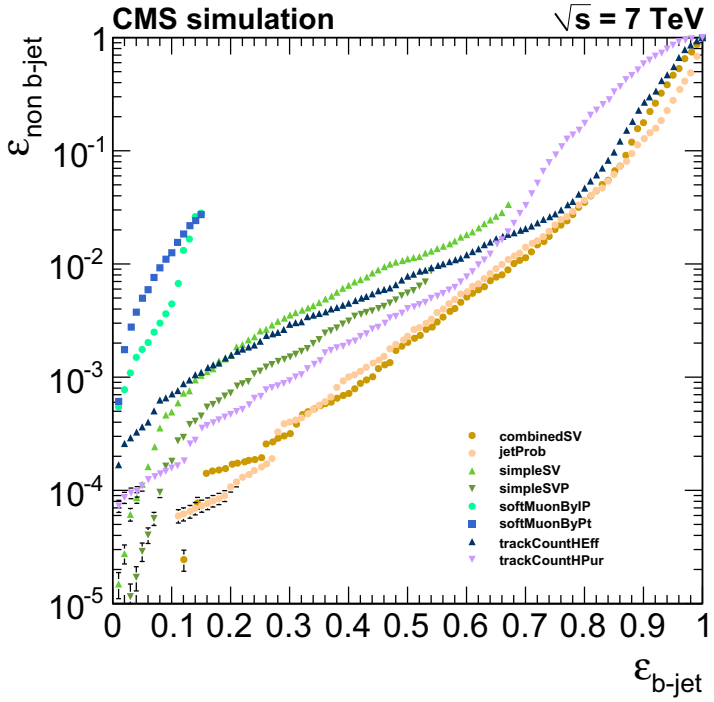


Figure 4.9: Performance of the CMS b-jet taggers. To compare the different taggers, they are plotted in the mistag rate/efficiency phase space. The lower right corner represents a tagger, where all non-b-jet could be suppressed without loosing any b-jet. The colors are chosen to distinguish the different kind of taggers. Blue: the muon tagger (SMT), green: the simple secondary vertex tagger (SSV), violet: are track counting tagger (TC), orange: are the two more complex: jet b probability and combined secondary vertex.

as determined by the $p_{T,rel}$ fits. The $p_{T,rel}$ fits can be applied to the muon-jet+away-jet sample (pTrel(n) method) or to the muon-jet+tagged-away-jet sample (pTrel(p) method) [CMS07b].

Counting Method The Counting method uses a different approach to estimate the b content of the sample before tagging; it assumes that the away-jets in the n sample are dominated by light jets, and that the average probability of tagging them can be estimated from light jets data sample with negative impact parameter with respect to the interaction point [CMS07b].
The major source of systematic uncertainty for methods that rely on the $p_{T,rel}$ fit is given by the modeling of the templates. To estimate this uncertainty, the templates were rederived using a different sample. This alternative set of templates is then used to remeasure the b-tagging efficiency, and the difference in the central value obtained is assigned as systematic uncertainty. Typical variations in the range between 10 and 20% are observed, with the larger values corresponding to bins with lower statistics on the samples used to derive the templates. The Counting method has an additional systematic uncertainty arising from the measurement of the mistag rate, which is evaluated by varying the number of cl jets before tagging by ±5% [CMS07b].

System8 Method The System8 method has been developed by the D0 collaboration [CDD+03]. It does not rely on $p_{T,rel}$ fits to extract the b-jet content of the samples; the Monte-Carlo simulation is only used to evaluate correlation factors between different tagging algorithms. For the current implementation of the System8 method, two data samples are used: the muon-jet+away-jet sample, and the muon-jet+tagged-away-jet sample.
The following system of eight equations is then obtained:

$$
\begin{aligned}
n &= n_b + n_{cl} \\
p &= p_b + p_{cl} \\
n^{tag} &= \varepsilon_b\, n_b + \varepsilon_{cl}\, n_{cl} \\
p^{tag} &= \beta\, \varepsilon_b\, p_b + \alpha\, \varepsilon_{cl}\, p_{cl} \\
n^{\mu} &= \varepsilon^{\mu}\, n_b + \varepsilon^{\mu}\, n_{cl} \\
p^{\mu} &= \varepsilon^{\mu}\, p_b + \varepsilon^{\mu}\, p_{cl} \\
n^{tag,\mu} &= \kappa_b\, \varepsilon_b^{tag}\, \varepsilon_b^{\mu}\, n_b + \kappa_{cl}\, \varepsilon_{cl}^{tag}\, \varepsilon_{cl}^{\mu}\, n_{cl} \\
p^{tag,\mu} &= \beta\, \kappa_b\, \varepsilon_b^{tag}\, \varepsilon_b^{\mu}\, p_b + \alpha\, \kappa_{cl}\, \varepsilon_{cl}^{tag}\, \varepsilon_{cl}^{\mu}\, p_{cl}
\end{aligned}
$$

The terms on the left hand side represent the total number of muon-jets in each sample before tagging (n, p) and after tagging with a lifetime tagger (n^{tag} , p^{tag}), the muon $p_{T,rel}$ cut (n^{μ}, p^{μ}), and both ($n^{tag,\mu}$, $p^{tag,\mu}$). The eight unknowns on the right hand side of the equations consist of the number of b and c+light jets in the two samples (n_b , n_{cl} , p_b , p_{cl}), and the tagging efficiencies for b and c+light jets for the lifetime tag and the muon $p_{T,rel}$ cut (ε_b^{tag}, ε_b^{μ}, ε_{cl}^{tag}, ε_{cl}^{μ}). The method assumes that the efficiency for tagging a jet with both the lifetime tag and the muon $p_{T,rel}$ cut can approximately be calculated as the product of the individual efficiencies.
Four additional parameters are needed to solve the system of equations: κ_b, κ_{cl}, α, and β. The first two parameters represent the correlation between the lifetime tag and the muon requirement for b jets (κ_b) and c+light jets (κ_{cl}), respectively. They are defined as

$$\kappa_b = \frac{\varepsilon_b^{tag,\mu}}{\varepsilon_b^{tag}\, \varepsilon_b^{\mu}} \quad \kappa_b = \frac{\varepsilon_{cl}^{tag,\mu}}{\varepsilon_{cl}^{tag}\, \varepsilon_{cl}^{\mu}}$$

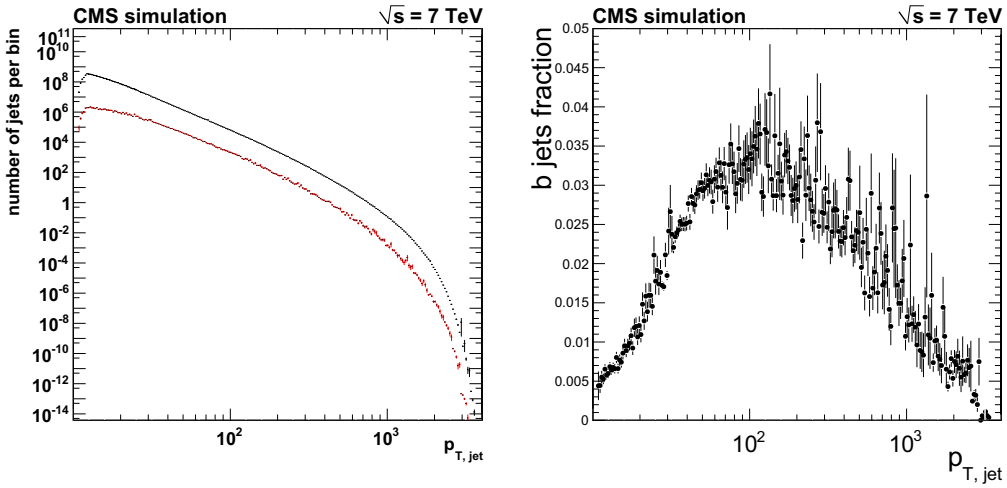


Figure 4.10: p_T spectrum of the all jets in black and b-jets in red (left) and the fraction of b-jets dependent on the jet p_T (right) for the Monte Carlo samples reconstructed in CMSSW version 3.6

The parameters α and β represent the ratio between the lifetime tagging efficiencies of the two data samples, used to solve System8, for b and c/light jets [CMS07b].

Further there is another method to determine the tagging efficiency of light quark and gluon jets. The method uses tracks with negative values of the signed impact parameter [CMS07a].

4.4 Monte Carlo samples

In this thesis different MC samples are used. Both were generated by PYTHIA6 with different tunes. They are also different in their GlobalTags and CMSSW release.
The samples are defined on specific $\hat{p}_T$ bins. The $\hat{p}_T$ ranges of each sample can be obtained from the sample names. To get meaningful, unbiased Monte Carlo statistics the samples must be combined. Before doing this it is needed to normalize all samples to the same integrated luminosity $\int \mathcal{L}$.

$$w = \frac{\text{cross section}}{\#\text{ events}}$$

Pythia 6 QCD DiJet

Table 4.4 contains the dataset names along with the corresponding number of events and the cross section. All values are extracted from the official MC generation page [13]. For the Summer 2010 reprocessing the CMS software version 3.6 is used. The GlobalTag of the reconstruction is START36_V10::All. The table shows all 20 samples which cover the $\hat{p}_T$ region from 0 GeV to 3500 GeV. Each sample has a specific number of events, created with the quoted cross section. In the last column the weights w are listed which adapt the distribution to an integrated luminosity $\int \mathcal{L} = 1/pb$.
After applying this weights the p_T spectrum of all jets (black) and also for b-jets can be seen in figure 4.10. On the right the expected fraction of b-jets is shown.
The smoothness of the curve is a proof of the right application of the weights. Further to check the amount of statistics of the Monte Carlo the number of events of each object is plotted in figure 4.11.

[13] https://twiki.cern.ch/twiki/bin/viewauth/CMS/ProductionReProcessingSummer10

name	# events	cross section [pb]	weight factor
QCDDiJet_Pt0to15	2197029	$4.844 \cdot 10^{10}$	$2.205 \cdot 10^{4}$
QCDDiJet_Pt15to20	2256430	$5.794 \cdot 10^{8}$	$2.568 \cdot 10^{2}$
QCDDiJet_Pt20to30	1032250	$2.361 \cdot 10^{8}$	$2.287 \cdot 10^{2}$
QCDDiJet_Pt30to50	1161768	$5.311 \cdot 10^{7}$	$4.571 \cdot 10^{1}$
QCDDiJet_Pt50to80	111289	$6.358 \cdot 10^{6}$	$5.713 \cdot 10^{1}$
QCDDiJet_Pt80to120	606771	$7.849 \cdot 10^{5}$	$1.294 \cdot 10^{0}$
QCDDiJet_Pt120to170	58888	$1.151 \cdot 10^{5}$	$1.955 \cdot 10^{0}$
QCDDiJet_Pt170to230	51680	$2.014 \cdot 10^{4}$	$3.897 \cdot 10^{-1}$
QCDDiJet_Pt230to300	52894	$4.094 \cdot 10^{3}$	$7.740 \cdot 10^{-2}$
QCDDiJet_Pt300to380	64265	$9.346 \cdot 10^{2}$	$1.454 \cdot 10^{-2}$
QCDDiJet_Pt380to470	52207	$2.338 \cdot 10^{2}$	$4.478 \cdot 10^{-3}$
QCDDiJet_Pt470to600	20380	$7.021 \cdot 10^{1}$	$3.445 \cdot 10^{-3}$
QCDDiJet_Pt600to800	22448	$1.557 \cdot 10^{1}$	$6.936 \cdot 10^{-4}$
QCDDiJet_Pt800to1000	26000	$1.843 \cdot 10^{0}$	$7.088 \cdot 10^{-5}$
QCDDiJet_Pt1000to1400	23956	$3.318 \cdot 10^{-1}$	$1.385 \cdot 10^{-5}$
QCDDiJet_Pt1400to1800	20575	$1.086 \cdot 10^{-2}$	$5.278 \cdot 10^{-7}$
QCDDiJet_Pt1800to2200	33070	$3.499 \cdot 10^{-4}$	$1.058 \cdot 10^{-8}$
QCDDiJet_Pt2200to2600	22580	$7.549 \cdot 10^{-6}$	$3.343 \cdot 10^{-10}$
QCDDiJet_Pt2600to3000	20644	$6.465 \cdot 10^{-8}$	$3.132 \cdot 10^{-12}$
QCDDiJet_Pt3000to3500	23460	$6.295 \cdot 10^{-11}$	$2.683 \cdot 10^{-15}$

Table 4.1: Monte Carlo samples: /*name*/Summer10-START36_V9_S09-v1/GEN-SIM-RECO, CMSSW 36X (GlobalTag: START36_V10::All)

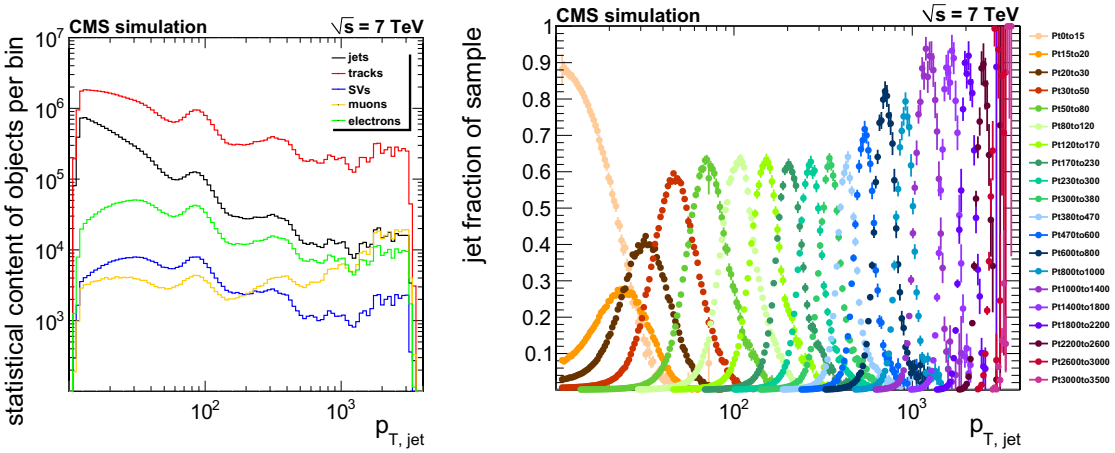


Figure 4.11: Left: Amount of statistics for the different objects, which are used in the analyses. MC samples for different $\hat{p}\_T$ bins are available. The different samples must be weighed to get the true p_T spectrum. Right: Composition of the weighed MC created from different samples with different $\hat{p}\_T$ bin ranges.

name	# events	cross section [pb]	weight factor
QCD_Pt_0to5	549809	$4.844 \cdot 10^{10}$	$8.810 \cdot 10^{4}$
QCD_Pt_5to15	1648096	$3.675 \cdot 10^{10}$	$2.230 \cdot 10^{4}$
QCD_Pt_15to30	5454640	$8.159 \cdot 10^{8}$	$1.496 \cdot 10^{2}$
QCD_Pt_30to50	3264660	$5.312 \cdot 10^{7}$	$1.627 \cdot 10^{1}$
QCD_Pt_50to80	3191546	$6.359 \cdot 10^{6}$	$1.992 \cdot 10^{0}$
QCD_Pt_80to120	3208299	$7.843 \cdot 10^{5}$	$2.445 \cdot 10^{-1}$
QCD_Pt_120to170	3045200	$1.151 \cdot 10^{5}$	$3.780 \cdot 10^{-2}$
QCD_Pt_170to300	3220080	$2.426 \cdot 10^{4}$	$7.534 \cdot 10^{-3}$
QCD_Pt_300to470	3171240	$1.168 \cdot 10^{3}$	$3.683 \cdot 10^{-4}$
QCD_Pt_470to600	2019732	$7.022 \cdot 10^{1}$	$3.477 \cdot 10^{-5}$
QCD_Pt_600to800	1979055	$1.555 \cdot 10^{1}$	$7.857 \cdot 10^{-6}$
QCD_Pt_800to1000	2084404	$1.844 \cdot 10^{0}$	$8.847 \cdot 10^{-7}$
QCD_Pt_1000to1400	1086966	$3.321 \cdot 10^{-1}$	$3.055 \cdot 10^{-7}$
QCD_Pt_1400to1800	1021510	$1.087 \cdot 10^{-2}$	$1.064 \cdot 10^{-8}$
QCD_Pt_1800	529360	$3.575 \cdot 10^{-4}$	$6.753 \cdot 10^{-10}$

Table 4.2: Monte Carlo samples: /*name*_TuneZ2_7TeV_pythia6/Fall10-START38_V12-v1/GEN-SIM-RECO, CMSSW 38X (GlobalTag: START38_V14::All)

To avoid running over samples, which have very small influence on the overall distribution, the fraction of the different samples was studied (figure 4.11, right). Analysis in specific p_T bins requires only samples with a sufficient contingent of weighted events.

These samples are used in the recent inclusive b-jet cross section measurement, which was performed on early CMS data. Further in this thesis these samples are used as an independent test dataset.

Pythia 6 QCD Tune2Z

The CMS collaboration also provides samples of Pythia 6 QCD Tune2Z. Tune2Z is a renewed fit of the parameters of the Monte Carlo generator.

Table 4.4 contains the dataset names along with the corresponding number of events and the cross section. All values are extracted from official MC generation page [14]. For the Full 2010 production CMSSW version 3.8 was used. The GlobalTag of the reconstruction is START38_V14::All.

The 15 samples provide more statistics than the former one. They are separated in $\hat{p}_T$ bins up to 1800 GeV. Further an additional inclusive sample which covers the high p_T regions above 1800 GeV is added. The application of the weight results in a fully inclusive spectrum up to large p_T values.

For this Monte Carlo sample the same distributions as above are plotted. In figure 4.12 the p_T spectrum and the b-jet fraction can be seen. Figure 4.13 show the influence of the different samples again.

Again the smoothness of the curve is a proof of the right application of the weights. Further to check the amount of statistics of the sample the number of events of each object are plotted in figure 4.13. The fraction of the different samples was studied as well.

These samples are used for all studies belonging to b-jet tagging as well as the b cross section measurements. Monte Carlo expectations for comparison with data are extracted from it.

[14] https://twiki.cern.ch/twiki/bin/view/CMS/ProductionFall2010

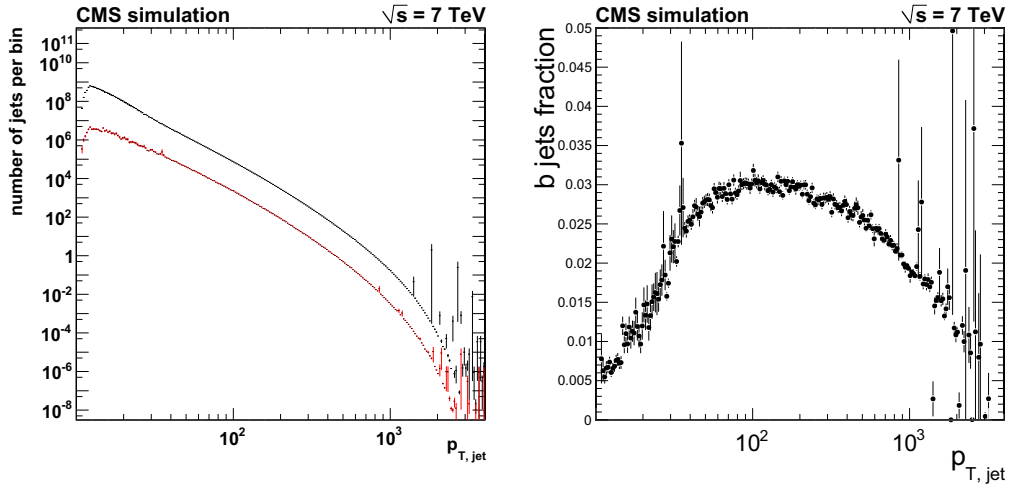


Figure 4.12: p_T spectrum of the all jets in black and b-jets in red (left) and the fraction of b-jets dependent on the jet p_T (right) for the Monte Carlo samples reconstructed in CMSSW version 3.8

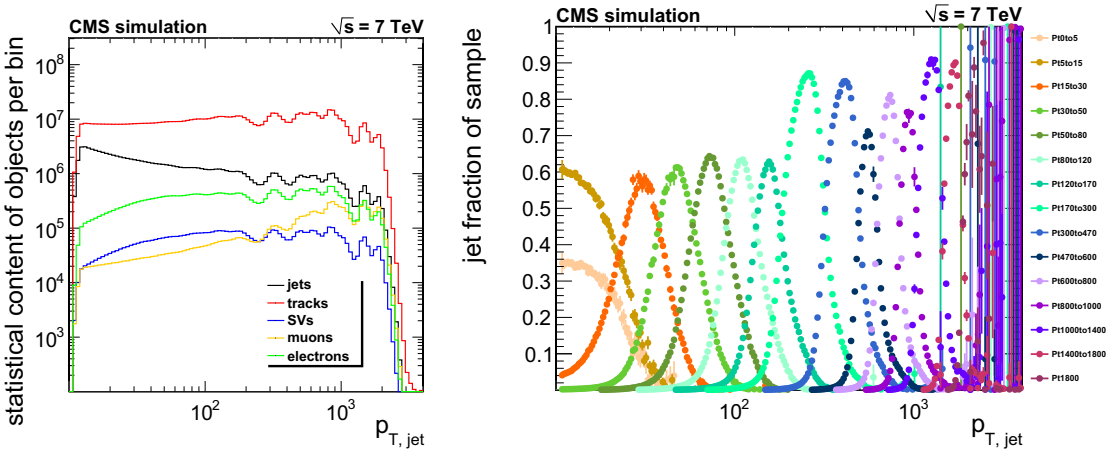


Figure 4.13: Left: Amount of statistics for the different objects, which are used in the analyses. MC samples for different $\hat{p}\_T$ bins are available. The different samples must be weighed to get the true p_T spectrum. Right: Composition of the TuneZ2 MC sample created from different samples with different $\hat{p}\_T$ bin ranges.

name	# events	Run range
JetMETTau/Run2010A	15 042 368	135821-141887
JetMET/Run2010A	24 064 576	141950-144114
Jet/Run2010B	20 270 640	146240-149711

Table 4.3: Trigger streams which arranges the dataset for the analysis. The jet stream, which is relevant for this analysis was combined with the missing E_T (MET) and the tau stream in earlier periods.

4.5 Data samples

For this analysis the complete amount of data of the Run2010 era is used, making use of the reprocessing in November 2010. It is reconstructed using the updated CMS software version 3.8. The recorded data is structured dependent on the different so called eras of data taking and the different trigger streams. The trigger stream changed two times, because of the increasing luminosity provided by the LHC.
While for the early data taking it was possible to pool events for jet, missing energy or tau studies, later the trigger streams only provided data for one analysis direction. The three datasets used are listed in table 4.3.
A former reconstruction of the first data sample was also used for an inclusive b cross section study on early CMS data, which was performed for the summer conferences 2010. Part of my thesis is to take my studies for the early analysis as base for further investigations and provide an update to the whole Run2010 dataset.

Having the right trigger streams it is also needed to check for an acceptable operation of the detector. For each run the lumi sections are centrally certified in so called good runs. Thus all detector components work well and we can believe in the reconstruction of the event. These certified lumi sections are provided by the CMS collaboration and listed in a published JSON file. We use the following official JSON files for that:

Cert_136033-149442_7TeV_Nov4ReReco_Collisions10_JSON.txt[15]

The attention of the CMS collaboration is focussed to high energy physics not yet covered by the Tevatron experiments. With the gain in luminosity it became necessary to prescale single, low energetic, jet triggers by some factor N. This must be done because of the technical limitations to record all collisions (4.1). This means instead of each event only every Nth event accepted by a specific trigger is recorded. Therefore the listing of the amount of data used for the analyses needs a separated look at the different triggers.
The integrated luminosities $\int \mathcal{L}$ of the data we analyzed are shown in table 4.4. They were measured by the CMS luminosity system [CMS10g]. For different trigger ranges the integrated luminosities are listed separately.
The prescaling of the low energetic triggers demands an efficient functionality of the higher energetic trigger. It is possible to test this and find a so called turn on point for each trigger. This can be achieved by comparing the trigger rate of the one we are interested in with another fully efficient one which is also unprescaled in this run range. The determined efficiency from such a comparison can be seen in figure 4.14. The estimated turn on points, where the trigger is more than 99% efficient is put in addition into the table 4.4.

[15] https://cms-service-dqm.web.cern.ch/cms-service-dqm/CAF/certification/Collisions10/7TeV/Reprocessing/

sample	first run	turn on	JetMETTau	JetMET	Jet	all
HLT_Jet15U	136035	37	0.0140	$9.60 \cdot 10^{-3}$	$1.86 \cdot 10^{-3}$	0.0256
HLT_Jet30U	136035	84	0.120	0.192	0.0374	0.352
HLT_Jet50U	136035	114	0.285	2.87	0.317	3.50
HLT_Jet70U	141956	153	-	2.87	5.99	9.17
HLT_Jet100U	141956	196	-	2.87	16.6	19.8
HLT_Jet140U	147196	245	-	-	27.8	36.0
HLT_Jet180U	148822	300	-	-	18.3	36.0

Table 4.4: integrated luminosity $\int \mathcal{L}$ in pb^{-1} for different samples and triggers. The different triggers are shown with the run number of its activation and the turn on position in $p_{T,jet}$, where the trigger becomes 99% efficient. The last column shows the integrated luminosity of this $p_{T,jet}$ range. The luminosities for the major trigger and the lower energy triggers are summed.

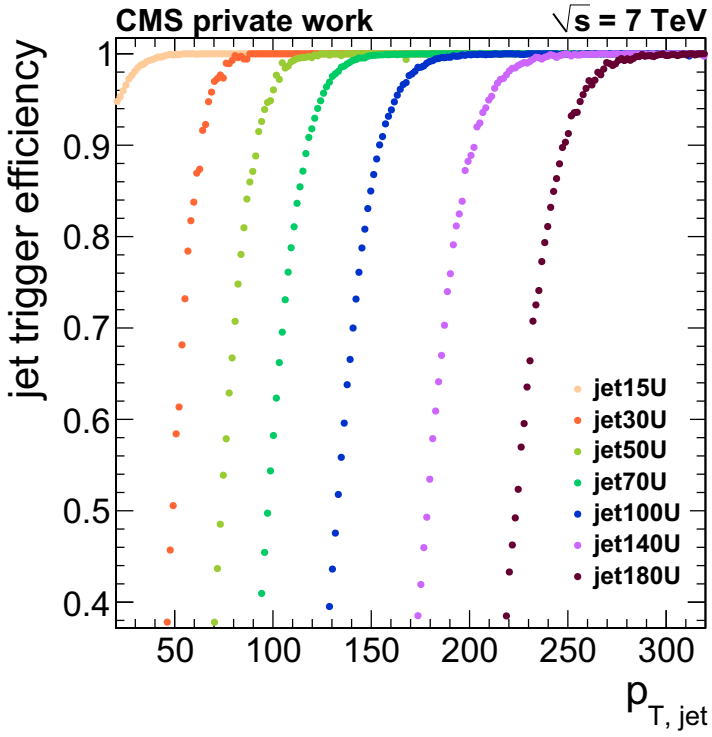


Figure 4.14: Efficiency of the different triggers. As turn on point the position where the trigger becomes 99% efficient is extracted.

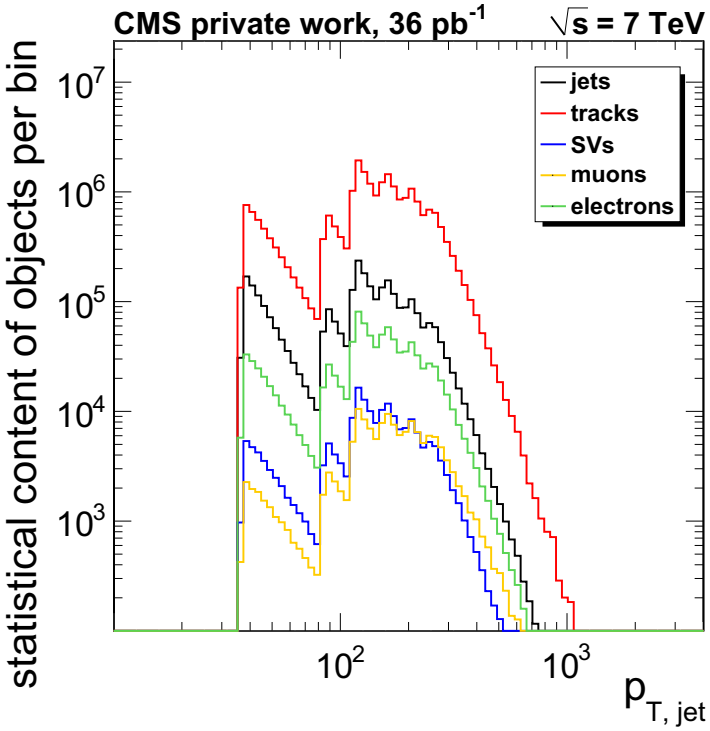


Figure 4.15: Available statistics for different bins in transverse momenta space of the jet. The amount is relevant for the precision of the measurement. Due to trigger prescaling in the low p_T regions, there the statistics are more or less limited perennially to the given number.

In the end we get the spectrum of the objects we want to analyse like it is plotted in figure 4.15. The structure in shape is caused by the prescaling of the trigger. The spectrum starts by a transverse jet momentum of 37 GeV. There the low energetic jet trigger (HLT_Jet15U) is barely efficient. It is also possible to see, that we have already jets with a transverse momentum of 1000 GeV collected. Accounting for the prescaling factor and the integrated luminosity we are able to plot the inclusive jet cross section of the reconstructed jets (see figure 4.16).

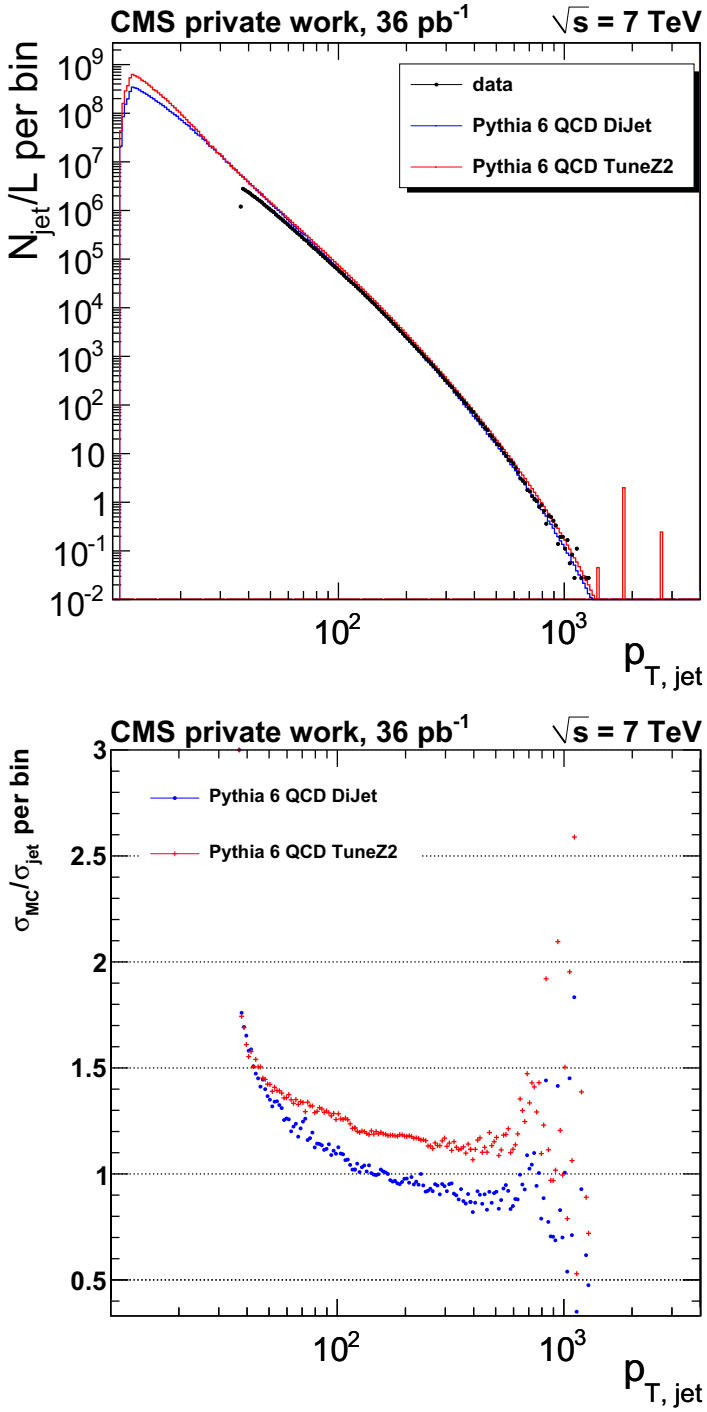


Figure 4.16: Top: the jet spectrum normalized by the integrated luminosity $\int \mathcal{L}$ is shown. It is compared with the two Monte Carlo expectations of the samples used in this thesis. Bottom: relative $p_{T,jet}$ spectrum. The MC spectra from the top plots are plotted relative to the data distribution.

Chapter 5

New applications of NeuroBayes

Many analyses for very different purposes are performed with NeuroBayes [Fei04]. It was not only applied to physics but also economics. With NeuroBayes interesting and important knowledge was obtained [1]. A summary of the different topics is available at the public web page `neurobayes.de`
In 2008 this knowledge was applied for the first time to the CMS experiment. A b-jet tagger for specific b-quarks decaying to electrons was developed [Mar09]. Based on this experience further applications of NeuroBayes for the CMS experiment were developed.
In this chapter I introduce the NeuroBayes framework and account for the new tasks I developed for CMS.

5.1 NeuroBayes

In this section I will give a complete overview of the multivariate analysis framework NeuroBayes. I will explain the architecture and the statistical methods included in this framework.

5.1.1 Introduction

NeuroBayes is a multivariate analysis framework, which was originally designed by Michael Feindt [PT10]. Like most frameworks for physics analysis it was developed to tackle one of the most important challenges in physics: the prediction of physics properties. This ranges from the binary case, where we are interested in whether or not an event belongs to a specific class, for example signal or background, to the continuous case, where the property of an object, for example the decay time, is estimated [Mor06]. In this case NeuroBayes delivers the probability density of the target for each event [Fei04].
These predictions are done by an intelligent combination of well known statistical methods [Fei04]. In this section I will introduce these methods and their contribution to NeuroBayes. The section is separated into two parts. The first describes the preprocessing of the input variables. The second part deals with correlation of the input variables to the target and the calculation of the prediction.

For a NeuroBayes analysis the various methods must be set up and calibrated on so called training samples. I will limit myself to the explanation to the classification mode of NeuroBayes. The classification is performed using two different samples as from now labeled as target 0 ($T0$) and target 1 ($T1$). The aim is to calibrate a so called NeuroBayes expertise/expert which is able to distinguish those two samples. In this section I will introduce the different methods which are

[1] http://neurobayes.phi-t.de/index.php/theses/jresearch-theses

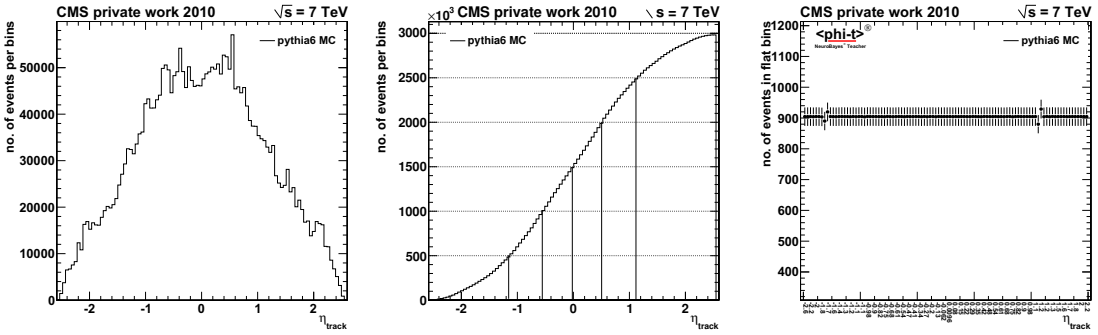


Figure 5.1: Probability integral transform for an exemplary variable. The left histogram shows the original distribution. In middle plot is the cumulative distribution created from the left. Dividing the y-axis in equally sized slices delivers the boundaries of the bins of the right histogram. The content of each bin of the last plot has the same amount of statistics.

implemented in NeuroBayes. An overview of the possible setup parameters can be found in [PT10]. This expertise file is used to get the predictions for each event of a data sample.

5.1.2 Preprocessing

Preprocessing is the umbrella term for all methods applied on the input variables x_i before the study of correlation to the target. This includes transformations of the single variables with and without knowledge of the target information as well as rotations of the complete input vector $\vec{x}$.

Probability integral transform

The probability integral transform is the transformation of random variables X distributed with density $f(x)$ to a uniform distribution. For a known cumulative distribution function $F(x) = \int_{-\infty}^{x} f(\tilde{x})\, d\tilde{x}$ the variable $Y = F(X)$ is distributed uniformly. If the distribution of X is unknown it is possible to estimate this transformation. For this the cumulative histogram is created which represents the cumulative distribution function. Dividing the y-axis in equal sized slices delivers us the boundaries of the bins in x for a new histogram (figure 5.1). The bins of this histogram all have the same amount of statistics.

Parametrization of the input variable distributions

Typically the distributions of the input variables for a given target are not known. For transformations of the input variable in a most convenient way it is recommended to parametrize these distributions. One possible procedure is the concept of orthogonal polynomials [BL98].
Each function $y_i(x_i)$ can be constructed by the linear combination of orthogonal polynomials $p_k(x) = \sum_i^k b_{ki}\, x^i$ in the following way

$$\tilde{y}_i(x_i) = \sum_j^m a_j p_j(x_i).$$

It is possible to get an estimate of the parameters a_i by a fit to the events. For a_j with expectation zero the estimate $\hat{a}_i$ follows a normal distribution with mean $\mathrm{E}[\hat{a}_i] = 0$ and variance $\mathrm{Var}(\hat{a}_i) = 1$. For the construction of the polynomial for the parametrization only the parameters a_i, which are significantly different from zero, are used.

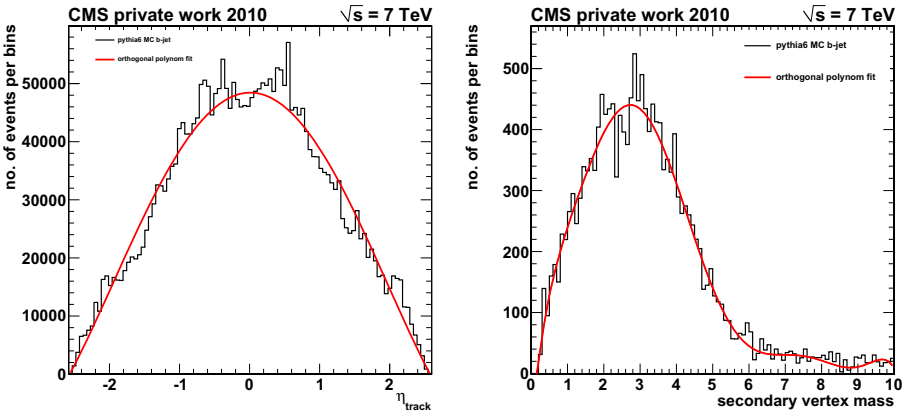


Figure 5.2: The target 1 distribution of two exemplary input variables (black) is fitted with the orthogonal polynomial method. The resulted function is plotted in red. Dependent on the shape of the variable may be difficult to find a good parametrization.

In figure 5.2 an orthogonal polynomial fit is performed. The fit looks good for the left plot. On the other hand it is possible to get a worse description of the events as shown on the right. This happens because of the low statistics in the corresponding bins of the histogram and the lack of flexibility of the fit function. To reduce such an effect a probability integral transformation can be applied before fitting.
Another important function to handle the parametrization is a fit of a spline function $S_n(x_i)$. This is a function defined piecewise by polynomials of degree n (see also [BL98]). With degree $n = 0$ it is a step function, indentical to the histogram. The natural cubic spline function has degree $n = 3$. It is twice continuously differentiable and the curvature of the endpoints a, b is defined by $S_3''(a) = S_3''(b) = 0$. This requirement leads to the smallest possible curvature. The knots for the spline function are constrained to the bin values.

Probability transformation

NeuroBayes makes use of both methods described above. First the distributions of the input variables x_i are transformed by the probability integral transform. Thus each bin of the input variable histogram has the same statistical power. In the next step we are only interested in the fraction of the events of one class. Figure 5.3 shows a histogram with the distribution of the two targets of a NeuroBayes classification in red and black. The plot is extracted from the output file of the official monitoring macro `analysis.C`. The plot can be identified by the label on the right. 'Flat' stands here for the result of the probability integral transform, splitted for the target 0 distribution in black and the target 1 distribution in red. Note the varying bin width, labeled on the x-axis.
Based on this histogram it is already possible to estimate a conditioned probability $P(T1|x_i)$ for each event,

$$P(T1|x_i) = \frac{N_{bin}(T1)}{N_{bin}(T0) + N_{bin}(T1)}.$$

N_{bin} is the number of $T0$ or $T1$ events per bin. The fraction of each bin is shown in figure 5.4. The plot is labeled with 'spline fit'. The 100 bins are simply labeled with the bin number.
To reduce binning effects and effects of the statistical uncertainties of each bin we can perform a fit by the method of orthogonal polynomials to the fraction.

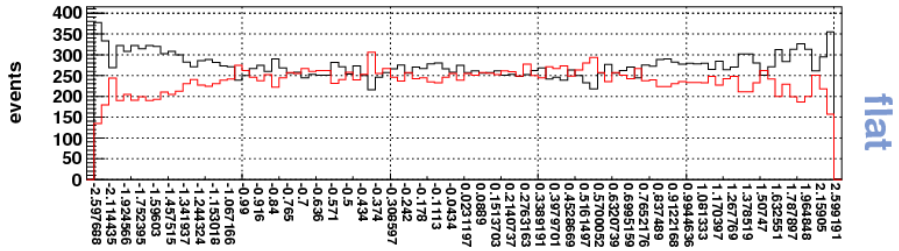


Figure 5.3: Target 1 (red) and target 0 (black) distribution after the probability integral transform.

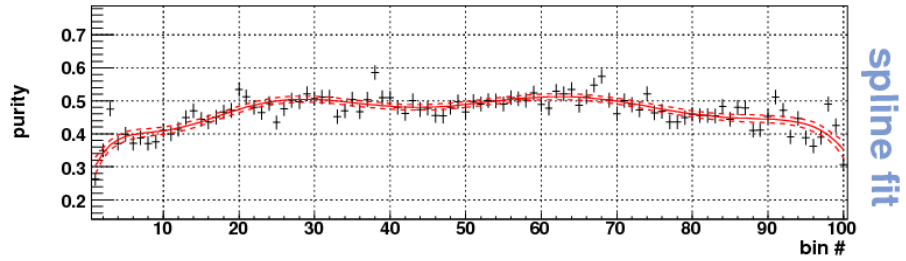


Figure 5.4: The fraction of the signal distribution of the flattened histogram (black) is fitted with the orthogonal polynomial method. The resulting function is plotted in red.

Now we are able to transform the input variables directly to an estimate of their probability $P(T1|x_i)$, where N is the normalization factor, which is given by the overall number of events,

$$N = \sum_{bins} \left(N_{bin}(T0) + N_{bin}(T1) \right).$$

This normalization factor cancels with the transformed a priori distribution $F(x_i)$, because this is an uniform distribution. We have constructed the simplest case of Bayes theorem with flat prior:

$$P(T1|x_i) = \frac{1}{N}\, \tilde{y}_i(x_i|T1)\, F(x_i) = \tilde{y}_i(x_i|T1).$$

NeuroBayes does this for all input variables. For the following calculations each x_i is replaced by its $P(T1|x_i)$.

Standardization and correlation coefficients

In preparation of the calculation of the correlation coefficients the distributions of different input variables are transformed once more. This time the variable distributions are standardized. The transformation is chosen in a way that the mean of the sole distributions is zero and the variance of them is one (figure 5.5)

$$y_i = \frac{\tilde{y}_i - \mathrm{E}[\tilde{y}_i]}{\sqrt{\mathrm{Var}(\tilde{y}_i)}}.$$

To calculate the correlation coefficients ϱ_{ij} of two input variables i and j we can sum the product of the transformed values for all events.

$$\varrho_{ij} = \sum_N y_i\, y_j$$

Figure 5.6 shows the matrix of the correlation coefficients of a exemplary NeuroBayes training.

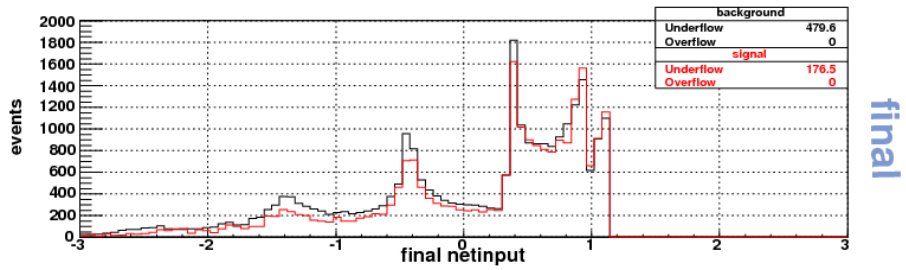


Figure 5.5: Standardization of the estimated signal probability of the variable from figure 5.4. The mean of this distribution is zero with a width of one. This is the final distribution of the transformed input variable after the preprocessing.

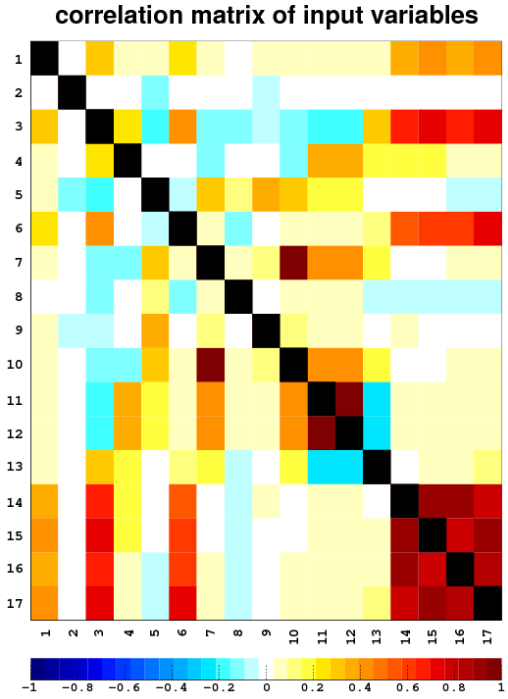


Figure 5.6: Matrix of correlation coefficients ϱ_{ij}. The first column is the target distribution, where target 1 is set to 1 and target 0 is set to 0. In this example there are many highly correlated variables $\varrho_{ij} \to 1$, which are painted in deep red and deep blue colors.

First order decorrelation

The main problem of multivariate analysis is the unknown correlation of the input variables. If the correlations were known we can construct the likelihood function and test any hypothesis by a likelihood ratio test [NP33]. Otherwise one has to find a procedure how to handle the correlated variables. The easiest way is to remove all variables which are highly correlated. This results in a more robust procedure but causes a loss of information which might be relevant for the discrimination.
Therefore a more advanced procedure is to attempt a decorrelation of the input variables. One possibility is to diagonalize the matrix of correlation coefficients ϱ. Such a diagonalization is a rotation in the n-dimensional phase space. The decorrelation is done in first order only. Second order correlations still remain.
Formally the diagonal matrix D can be described as

$$D = A\varrho A^{-1}$$

Technically the calculation of the rotation matrix A is almost impossible, because it needs the inverse of the n-dimensional matrix ϱ. But there are methods, which converge to the diagonalized form. The method used in NeuroBayes is the Jacobi rotation [BL98]. The idea is to do several 2-dimensional rotation steps until a close to diagonal matrix is formed. The merging of all this sub-rotations gives us the matrix A.
We are able to construct a set of uncorrelated variables $\tilde{z}_i = \sum A_{ij}\, y_j$.
The remaining correlation to the target of each row represents the information of the input variable finally added to the classification. With this information it is possible to prune variables with less relevance. This regards to variables with less information for the classification as well as to variables with more information but large correlation to others.

5.1.3 Target correlation and prediction

The preprocessing gives us a set of n almost uncorrelated variables $\tilde{z}_i$. There are many methods for hypothesis testing with these conditions in the literature. A lot of them have needlessly large running time, above all if some input variables have a very small correlation to the target. Depending on the correlation to the target, a selection of the relevant input variables is recommended.
NeuroBayes has an automatic sorting algorithm of the variables. Variables are sorted by relevance and, furthermore, it is possible to neglect variables with low significance.
To get the predictions NeuroBayes provides three different modes which lead to a calibrated expertise.

Zero iteration Training

The fastest and therefore most widely used procedure is an analytic method called zero iteration training. As described above we have a matrix of uncorrelated standardized variables. This represents a sphere in the n-dimensional phase space. So we have the freedom to decide a direction without influencing the variables itself. We can choose the direction with the most discriminating power with respect to the target. This direction is now called z_0. It is a linear combination of the $\tilde{z}_i$:

$$z_0 = \sum r_{ij}\, \tilde{z}_i$$

where r_{ij} are the coefficients of the rotation matrix, which provides the chosen direction.

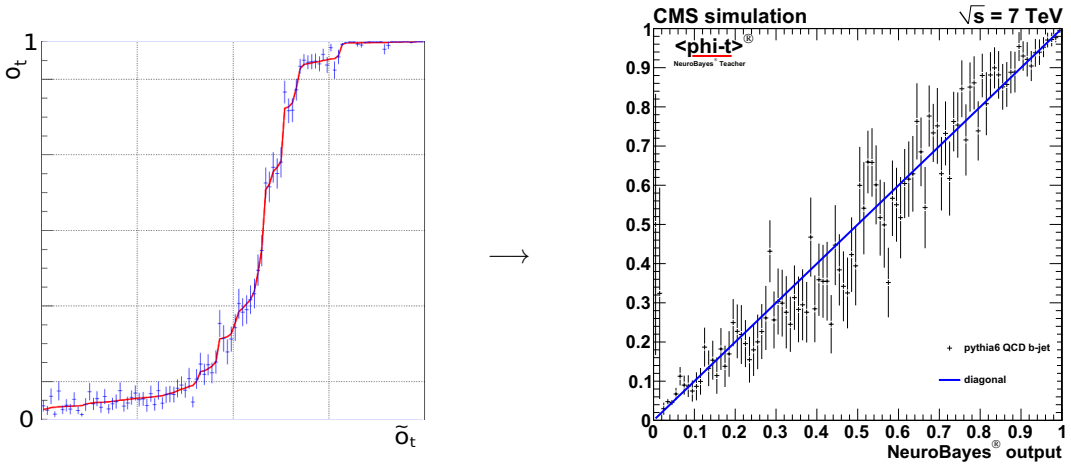


Figure 5.7: Diagonalization of the Zero iteration training. This is needed to get an probability interpretation of the NeuroBayes output. On the left the transformation functions is determined doing a fit of a monotonously rising spline function. On the right the resulting distribution of the target 1 purity $P(T1|o_t)$ to the final output value o_t is plotted. The expected behaviour, shown by the diagonal line is fullfilled.

This method works very well, because of the transformation we did for the input variables. The final input variables y_i have a monotonous rising dependency to the target 1 purity of each variable $P(T1|y_i)$. This dependency is conserved during the decorrelation. The projection of the variables z_i to the target is a very good discriminator z_0. But this z_0 does not fulfill the probability interpretation. Plotting the $T1$ fraction of z_0 and fitting with a monotonously spline function transforms this to the probability $P(T1|z_0)$. The probability transformation can be seen in figure 5.7 on the left. If the purity of each bin is plotted (as done on the right) it correspond directly to the mean value of each bin. That means the output can be interpreted as probability.
The output value is now called o_t. The index - finally combined with an identification number - is used to specify different NeuroBayes trainings t. Hence for the probability I take the following notation:

$$o_t = P(T1|o_t).$$

Neural Network Training

Another more time consuming method is an artificial neural network [Ros58]. With this we can handle higher order correlations too. In NeuroBayes just a simple feed-forward network with one hidden layer is implemented. The default value for the number of hidden nodes is the number of input nodes minus one. The number of nodes N of the output layer depends on the NeuroBayes mode. For a binary target it has one node, for the continuous mode $N = 20$. The output values $o_{t,i}$ are calculated by:

$$o_{t,i} = S\left(\sum_j w_{ji}^{2\to 3} S\left(\sum_k w_{kj}^{1\to 2} y_k\right)\right)$$

where the weights $w_{ji}^{a\to b}$ are the connections between the different layers, y_k is the transformed input value of variable k and $S(x)$ is the sigmoid function

$$S(x) = \frac{2}{1 - e^{-x}} - 1$$

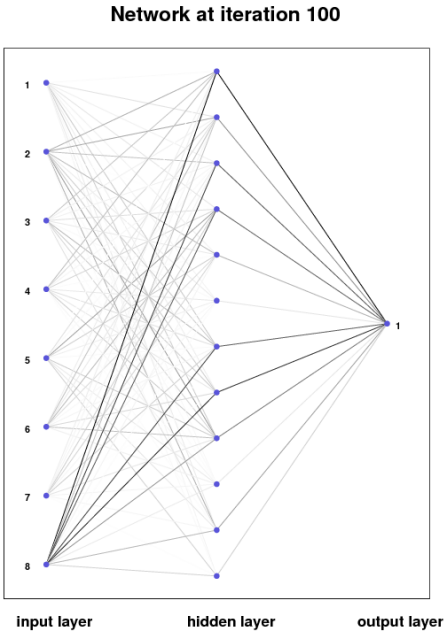


Figure 5.8: Example of an architecture of an artificial neural network calibrated in NeuroBayes. The thickness of the lines corresponds to the absolute values of the weight $w^{a\to b}$

used as the transfer function for each knot. A possible architecture of the artificial neural network implemented in NeuroBayes is shown in figure 5.8. There are three layers, the input layer, one hidden layer and the output layer with only one knot. The number of input layers is reduced, because of the minor correlation of the input variables y_i to the target. The thickness of the lines correspond to the absolute values of the weight $w^{a\to b}$. These are calculated by the back propagation mechanism [RHW87]. There is the possibility to use the so called BFGS mechanism [BRO70] to minimize the error function in a more efficient way. As shown in [Fei04] the output of such an artificial neural network can be interpreted as probability $o_t = P_t(T1|o_t)$.

5.2 NeuroBayes probability

In this section I will present applications of the probability interpretation of the NeuroBayes output. I will show how we must transform it to get the right probabilities of a given data sample and how it can be used to estimate the fraction of signal events. Further I will introduce the sPlot method and how it can be implemented if we have the NeuroBayes output distribution at our hands.

5.2.1 NeuroBayes probability transformation

For a given NeuroBayes classification with two samples $T0$ (target 0) and $T1$ (target 1) the result of the training t gives us for each of the events the probability $P_t(T1|o_t)$ with the NeuroBayes output value o_t. The overall number of events is given by $N = N_{T0} + N_{T1}$. For the output we have the following equations:

$$P_t(T1|o_t) = o_t,$$

$$P_t(T0|o_t) = 1 - o_t.$$

Being a probability is one of the main properties of the NeuroBayes output. Checking for this is therefore a crosscheck of a reliable calibration of the NeuroBayes expert. In figure 5.9 on the right

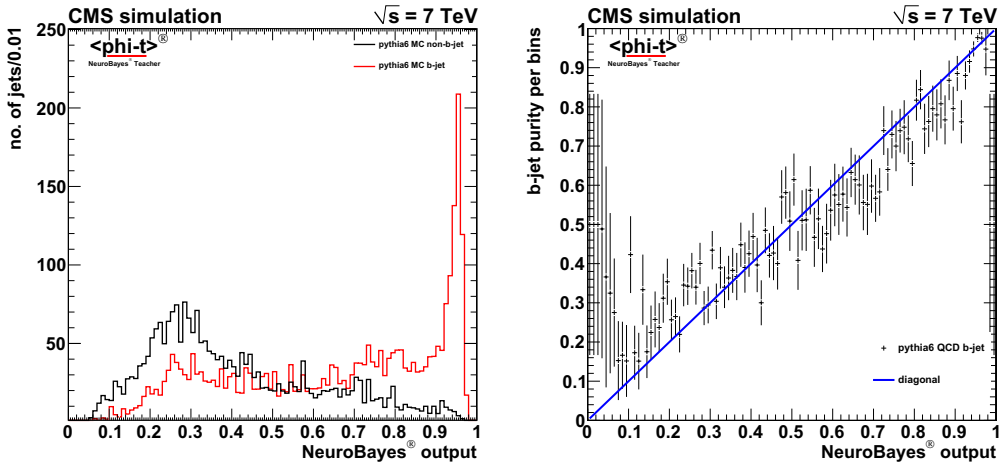


Figure 5.9: On the left the output distribution of an exemplary NeuroBayes calibration is shown. In black is the background and in red the signal distribution. On the right the purity of the signal distribution for each bin is plotted.

the purity $p_i = \frac{n_i(T1)}{n_i(T1 \vee T0)}$ of each bin for the exemplary NeuroBayes output variable o_t on the left is shown. As expected, all the calculated purity values lie on the diagonal axis, which corresponds to the quoted property.

If the target fraction differs form the analysis sample, a monotone transformation is needed to maintain the probability interpretation. Here I will discuss two cases, where such a transformation is needed.

Let us assume a general case, where we want to analyse a given data sample consisting of two classes: signal S and background B. The size of the sample is given by $N_d = N(S) + N(B)$. We are interested in the probability $P(S|o_{nb})$ of some event out of this sample to be a signal event dependent on the NeuroBayes output o_{nb}.

In the first case we have some simulations to study the differences between the two classes. Therefore we have one sample S_{MC}, where $pdf(\vec{x}|S_{MC}) \approx pdf(\vec{x}|S)$ and one sample B_{MC}, where $pdf(\vec{x}|B_{MC}) \approx pdf(\vec{x}|B)$, with given number of events $N(S_{MC})$ and $N(B_{MC})$. These two samples are called training samples.

The way to go is quite easy. A successful NeuroBayes classification for case one ($t1 : \vec{x} \to o_{t1}$) on the simulated samples gives us the probability $P_{t1}(S_{MC}|o_{t1})$.

$$o_{t1} = P_{t1}(S_{MC}|o_{t1})$$

and

$$1 - o_{t1} = P_{t1}(B_{MC}|o_{t1})$$

Bayes theorem says:

$$P_{t1}(S_{MC}|o_{t1})pdf(o_{t1}) = pdf(o_{t1}|S_{MC})P_{t1}(S_{MC})$$
$$P_{t1}(B_{MC}|o_{t1})pdf(o_{t1}) = pdf(o_{t1}|B_{MC})P_{t1}(B_{MC})$$

and gives us the ratio for the training sample as:

$$\frac{o_{t1}}{1 - o_{t1}} = \frac{P_{t1}(S_{MC}|o_{t1})}{P_{t1}(B_{MC}|o_{t1})} = \frac{pdf(o_{t1}|S_{MC})P_{t1}(S_{MC})}{pdf(o_{t1}|B_{MC})P_{t1}(B_{MC})} \approx \frac{pdf(o_{t1}|S)}{pdf(o_{t1}|B)} \frac{P_{t1}(S_{MC})}{P_{t1}(B_{MC})}$$

To simplify the formula in the further steps we can introduce the likelihood ratio:

$$\Lambda_{t1}(o_{t1}) = \frac{pdf(o_{t1}|S)}{pdf(o_{t1}|B)} = \frac{o_{t1}}{1-o_{t1}} \frac{P_{t1}(B_{MC})}{P_{t1}(S_{MC})}.$$

Bayes theorem is also true for the data sample so we get for the ratio on data:

$$\frac{P(S|o_{t1})}{P(B|o_{t1})} = \frac{pdf(o_{t1}|S)}{pdf(o_{t1}|B)} \frac{P(S)}{P(B)}.$$

With $P(B|o_{t1}) = 1 - P(S|o_{t1})$ and the ratio for the training sample we get the probability of an event to be signal for a given NeuroBayes value:

$$P(S|o_{t1}) = \frac{\Lambda_{t1}(o_{t1})}{\frac{P(B)}{P(S)} + \Lambda_{t1}(o_{t1})} = \frac{\Lambda_{t1}(o_{t1})P(S)}{1 + P(S)\left(\Lambda_{t1}(o_{t1}) - 1\right)}.$$

The likelihood ratio $\Lambda_{t1}(o_{t1})$ is easy to calculate from the known properties of the NeuroBayes training. So only the fraction of signal events $P(S)$ is needed to calculate the posterior probability $P(S|o_{t1})$. $P(S)$ can be estimated by a template fit using $P(o_t|S)$ and $P(o_t|B)$ as templates.
Using $P(S|o_{t1})$ as a weight on a data sample we can unfold the signal distribution of any variable, which is totally correlated to o_{t1}. In [PL05] this simple behavior is named inPlot.

In a second case we want to calculate this probability only with a given sample of simulated signal ($pdf(\vec{x}|S_{MC}) \approx pdf(\vec{x}|S)$). All information about the background must be taken from the data sample $D = S + B$. For the NeuroBayes classification ($t2 : \vec{x} \to o_{t2}$) we use the signal simulation as target 1 sample and the data sample as target 0 sample: $P_{t2}(D) + P_{t2}(S_{MC}) = 1$. Bayes theorem for the NeuroBayes training gives us:

$$o_{t2} = P_{t2}(S_{MC}|o_{t2}) = \frac{pdf(o_{t2}|S_{MC})P_{t2}(S_{MC})}{pdf(o_{t2}|S)P_{t2}(S) + pdf(o_{t2}|B)P_{t2}(B) + pdf(o_{t2}|S_{MC})P_{t2}(S_{MC})}$$

For the interesting probability $P(S|o_{t2})$ we know:

$$P(S|o_{t2}) = \frac{pdf(o_{t2}|S)P(S)}{pdf(o_{t2}|S)P(S) + pdf(o_{t2}|B)P(B)}$$

With $P_{t2}(S)/P_{t2}(B) = P(S)/P(B)$, $N_{t2}(S)/N_{t2}(D) = N(S)/N_d$ and Bayes theorem for the NeuroBayes training we get:

$$P(S|o_{t2}) = \frac{P_{t2}(S)}{P_{t2}(S_{MC})} \left(\frac{o_{t2}}{1-o_{t2}}\right) = P(S)\Lambda_{t2}(o_{t2})$$

In this case we have similar dependencies as in the first case. The likelihood ratio $\Lambda_{t2}(o_{t2})$ is well known, but we need an estimate of the unknown signal fraction $P(S)$. Another property of this equation is the limitation of $\Lambda_{t2}(o_{t2})$. Because of $\max(P(S|o_{t2})) = 1$ we get $\Lambda_{t2}(o_{t2}) < 1/P(S)$. This upper limit is smeared by the resolution of the NeuroBayes training.

5.2.2 Boost Training - NeuroBayes and weights

At last I want to focus on the procedure of boosting an already calibrated NeuroBayes expertise. Boosting is an umbrella term for calibrating more than one expertise to get a final result. The goal of each iteration step is to correct possible imprecisions of the former steps. For example if you have a very complex problem it is clever to learn the obvious thing in a first step and do the complicated things in a second. The second step will become easier because we are on a better initial position.
To implement such a boost for the new calibration all events must be weighted.

The easiest and most intuitive approach for a boost training is to weight the events e of the one target with the probability to be of the other target:

$$w_{T0} = P(T1|e),$$

$$w_{T1} = P(T0|e).$$

A given region r with a given number of events N out of two classes $T0$ and $T1$ has the probabilities $P(T0|r) = \frac{N(T0)}{N(T1)+N(T0)}$ and vice versa. Applying the weight from above results in the same effective numbers of events N_b:

$$N_b(T0) = w_{T0}N(T0) = \frac{N(T1)N(T0)}{N(T1)+N(T0)} = w_{T1}N(T1) = N_b(T1).$$

By knowing the true $P(T|r)$ of the given region, no further classification is possible. For any inclusive distribution is:

$$P(x|T1)w_{T0} = P(x|T1)P(T0|x) = P(x|T1)\int drP(T0|x,r)P(r|x)$$

A usual NeuroBayes classification results in an estimate of $\hat{P}(T|r) = P(T|o_t)$. The region r is defined by the input variables. If we construct a weight in the same way with this estimate, only the information gained by the training vanish. An additional so called boost training can find further quantities to separate the two classes. The combination of both improves the overall classification.

By construction the output of the two experts should be less correlated. The combination of the two results can be approached by the multiplication of their likelihood ratios, given by

$$\Lambda = \frac{pdf(o_t|T1)}{pdf(o_t|T0)} = \frac{o_t}{1-o_t}\frac{N(T0)}{N(T1)}.$$

Still correlations of the two experts can appear. It happens when the probability interpretation of the output variable o_t of the unboosted calibration is not entirely correct. Thus the weights for the boost training involve a bias from the ill estimated events. To control effects of this source it is advised to take o_t as an input for the boosted training. If everything is correct, the variable has no correlations to the target and does not influence the boost training. Any dependency between o_t and the target is a hint towards problems which must by investigated.
In most of the cases the boost training does only small corrections to the first. Therefore possible correlations are also small and the likelihood ratio combination can be used.

Such a boost training can be applied many times. In fact most of the improvements are already achieved by the first boost training. Maybe for very special cases a gain with more iterations is possible.

For the boost it is allowed to change the calibration settings in any imaginable way, while the effective number of events in any region is the same for both classes. This enables various implementations which I will explain in the following:

- It is possible to focus the calibration on specific regions of the samples. This can be arranged, if the weights are varied by any focusing function F_f.

$$w'_{T0} = P(T1|o_t)F_f$$

$$w'_{T1} = P(T0|o_t)F_f$$

The effective number of events is still the same for the two classes, but transformed by the function F_f.

$$N_b(T) = w'_T N(T) = \frac{N(T0)N(T1)}{N(T1) + N(T0)} F_f$$

If the multivariate analysis technique with a correct error propagation works without a preprocessing, which can cause binning effects, nothing should change. In most of the cases, e.g. using probability integral transformations or a binned fitting of a regularization function, the application of the focusing function uncloses the binning. Therefore it is possible to see the structure of the phase space on a subbin level. This way information lost during preprocessing is recovered and the classification is improved.

As an example for b-jet tagging it is interesting to enable very pure b-jet samples. Choosing $F_f = \frac{1}{P(T0|o_t)}$ focuses the boost training on this so called purity region.

- It is also possible to enlarge the number of events to be more precise for the next calibration. For this it is important to take the new statistics into account, when calculating the probabilities from the NeuroBayes output.

- It is allowed to add new variables to the boost training or leave some out.

- Another application is to study properties of the events independent of some other variable x. With the weighting we can remove the dependencies of this variable x and do a new classification. The calibration of this kind of boost training gives us an estimate independent of the variable x. This feature is interesting for b-jet tagging efficiency measurements. With a boost training it is possible for each existing b-jet tagger to create an uncorrelated partner. This can be used for efficiency measurements on the data sample (see 4.3.9).

NeuroBayes has the ability to handle each event with a specific weight. That is why it is easy to implement such a boost. NeuroBayes has also an internal boost mode where in a first step a zero iteration training is performed and in a second step it attempts to find second order correlations with an artificial neural network.

Weighting issue Indeed NeuroBayes is constructed for the use of weights. The idea is to have a correct error propagation included in the framework for the application of advanced algorithms as described in this section of the thesis. Unfortunately there is bug in version 20101026. Phi-T claimed to fix this issues for the updated versions. The bug occurs in the wrong uncertainty calculation, when large weights are applied (Figure 5.10). The effect of very large weights can be seen in the bins on the right. The purity distribution is effected by the large weight of single events. For the regularization a wrong estimate of the expected bin content is used.
This issue influences the result dramatically if very large weights are used. Therefore it is necessary for this thesis to avoid large weights.

5.2.3 sPlot

sPlot is an unfolding technique introduced 2005 by Muriel Pivk [PL05]. The sPlot technique is a method where the so called sPlot weights are applied on each event of a given sample. The weight is calculated corresponding to a target T. For any variable x not correlated to this weight, its distribution is transformed to the conditioned distribution of T:

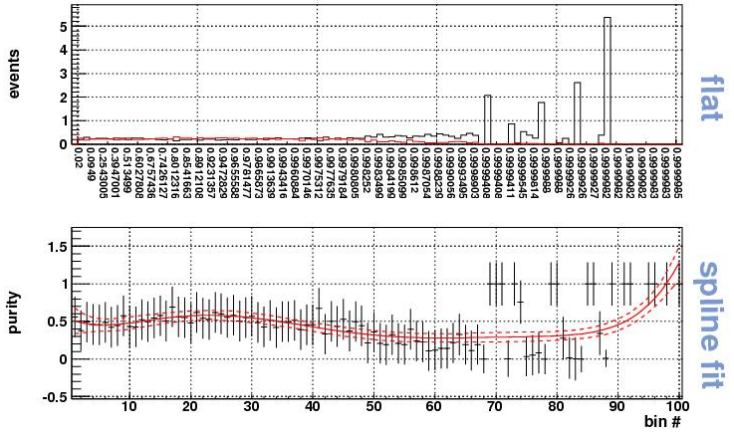


Figure 5.10: Effect of large weights. The bins on the right contain only one event with a large weight. The purity is effected by the target type of this single event. A more reasonable estimate of the purity should be around 0.5 with adequate uncertainty.

$$\int dw_{sPlot}\, pdf(x|w_{sPlot})\, w_{sPlot} = pdf(x|T)$$

The sPlot paper includes among other things the derivation of the sPlot weights. Because of the differing notations used in this thesis and analogies to a later method used for the b cross section, I will introduce the sPlot method in my own words.
I will explain the sPlot method for a special case of only two classes. Similar to the descriptions above we have one class called target 0 ($T0$) and the other called target 1 ($T1$). The sPlot weights w_{sPlot} are determined using the output values o_t of a given NeuroBayes expert.
Looking at the inclusive distribution of o_t for a given x we find:

$$\int do_t\, pdf(o_t|x) = \int do_t \sum_T P(T|x)\, pdf(o_t|T,x)$$

This equation is the usual base for any inclusive study. We have two variables. It is possible to define an inclusive region of one variable x and take the other variable for studies of the target. Therefore we need external knowledge about $pdf(o_t|T,x)$, e.g. templates from a Monte Carlo sample. If we do this in further inclusive regions we can get a picture of how the first variable is related to the target:

$$pdf(x|T) = \frac{P(T|x)\, pdf(x)}{\int dx\, P(T|x)\, pdf(x)}.$$

If o_t is uncorrelated to the variable x for both targets T, we are able to determine the x distribution in a more advanced way. This is called sPlot method. Therefore the following requirement

$$pdf(o_t|T,x) = pdf(o_t|T)$$

must be fulfilled. There is no correlation of the two variables for the different targets T. This brings us the following simplification:

$$\int do_t\, pdf(o_t|x) = \int do_t\, \left[P(T0|x)\, pdf(o_t|T0) + P(T1|x)\, pdf(o_t|T1)\right].$$

A simple trick allows us the calculation of the distributions $P(T|x)$. If we weight each event in o_t with the probabilities $P(T0)/P(T0|o_t)$ or $P(T1)/P(T1|o_t)$, we get two similar equations:

$$\int do_t\, pdf(o_t|x)\frac{P(T')}{P(T'|o_t)} = \sum_T pdf(T|x)\int do_t\, pdf(o_t|T)\frac{P(T')}{P(T'|o_t)}.$$

After applying Bayes' Theorem we can write this in matrix notation:

$$\begin{pmatrix} \int do_t\, pdf(o_t|x)\frac{P(T1)}{P(T1|o_t)} \\ \\ \int do_t\, pdf(o_t|x)\frac{P(T0)}{P(T0|o_t)} \end{pmatrix} = \begin{pmatrix} \frac{pdf(T0|x)}{P(T0)} \\ \\ \frac{pdf(T1|x)}{P(T1)} \end{pmatrix} V^{-1}$$

where V^{-1} is the matrix:

$$V^{-1} = \begin{pmatrix} \frac{1}{N}\sum \Lambda_{t1}^{-1} & 1 \\ \\ 1 & \frac{1}{N}\sum \Lambda_{t1} \end{pmatrix}$$

The integration about o_t is replaced by the sum over the finite number of events from the sample. Λ_{t1} is the likelihood ratio as defined in section 5.2.1 for the first case: $\Lambda_{t1} = \frac{pdf(o_t|T1)}{pdf(o_t|T0)}$. The integration of the normalized distribution $pdf(o_t)$ in the matrix elements next to the diagonal is one. After the determination of Λ_{t1} with a NeuroBayes training we are able to calculate the matrix V, which is the inverse of V^{-1}.

For $pdf(x|T0)$ and $pdf(x|T1)$ finally we get:

$$pdf(x|T1) = pdf(x)\int do_t\, pdf(o_t|x)\underbrace{\frac{P(T0)}{P(T0|o_t)}\left(V_{T1,T1} + \Lambda_{t1}\, V_{T1,T0}\right)}_{w_{sPlot}(T1)}$$

$$pdf(x|T0) = pdf(x)\int do_t\, pdf(o_t|x)\underbrace{\frac{P(T0)}{P(T0|o_t)}\left(V_{T0,T1} + \Lambda_{t1}\, V_{T0,T0}\right)}_{w_{sPlot}(T0)}$$

Here we can define the sPlot weights as requested in the beginning.

The sPlot weights w_{sPlot} have further properties shown in [PL05]. So the sum of the signal and the background weight is given by $w_{sPlot}(T0) + w_{sPlot}(T1) = 1$. The weights are not limited to an interval between $(0, 1)$. It is also possible to get weights smaller than zero and larger than one. The statistical uncertainties can be calculated by the sum of the squared weights ($\sigma_{sPlot} = \sqrt{\sum w_{sPlot}^2}$). The sPlot method is a nice feature to extract the signal and background shapes of variables, where the particular distributions are unknown. It is possible to get this distributions by running over the whole data sample. The uncertainty depends on amount of statistics, which is available for signal and background.

5.3 NeuroBayes b-jet tagger

In this section I will present two new methods for discriminating b-jets from non-b-jets. Both methods make use of the multi variante analysis framework NeuroBayes. For the first tagger the NeuroBayes expert is calibrated using a sample of simulated b-jets (signal target, T1) and non-b-jets (background target, T0). The other is calibrated using the real data sample as background target.
First I will explain the needed input variables and their quality for b-jet tagging. In the second part I explain the differences of the two b-jet taggers and show how they perform.

category	input variables of track objects
four vector	momentum p
	pseudo rapidity η
primary vertex	significance of the two dimensional signed impact parameter
	significance of the three dimensional signed impact parameter
	two dimensional signed impact parameter
	three dimensional signed impact parameter
	track decay length
jet position	track transverse momentum, relative to the jet axis
	track parallel momentum, along the jet axis
	ΔR of the track to the jet axis
	minimum track approach distance to jet axis
jet energy	transverse momentum, relative to the jet axis, normalized to its energy
	parallel momentum, along the jet axis, normalized to its energy
quality	χ^2 value of the track fit [SAF$^+$06]
	number of hits in the pixel detector
	number of hits in all tracking detectors
b hadron	distance to reconstructed b hadron axis
	significance of distance to reconstructed b hadron axis
	track weight for b hadron reconstruction

Table 5.1: Input variables of the track objects

5.3.1 b-jet tagging variables

For the two NeuroBayes b-jet tagger I decided to develop a framework of conditional NeuroBayes experts. To be most performant I use all available information concerning b-jets. This includes lifetime information as well as lepton information. For this it is necessary to match different objects to the jets. This objects are the tracks from the jet, secondary vertices, which are reconstructed from tracks, electron and muon tracks (see 4.3). A table of the available input variables is shown in the tables 5.1-5.5. Properties which correspond to the physical quantities as well as the quality of the objects are stored in the input variables.
All of these variables have to be compared to data. Only input variables, which compare to data are used for the NeuroBayes b-jet tagger. All of these objects will be used by NeuroBayes to decide, how b-like a jet is.

Tracks Details of the track reconstruction can be found in section 4.3.1. From the four momentum vector the transverse momentum and the pseudorapidity are extracted. Further geometrical properties of the track relative to the primary vertex and the jet are used, e.g. the impact parameter and its significance to be inconsistent with the primary vertex. The ratio of the sum of the track momenta to the jet energy measured in the calorimeter is calculated as well as properties, which describe the track kinematics relative to the jet. Finally some quality variables, the fit parameters and the number of hits in tracking detector components are used (table 5.1).

Muons Details of the muon reconstruction can be found in section 4.3.5. The muon input parameter list partially is the same as for the general track objects. There are two new variables, which depend on the jet energy. The momentum of the muon track is boosted into the jet rest frame. This and normalized by the jet energy is taken as additional input variable. The quality of

category	input variables of muon candidates
four vector	momentum p
	pseudo rapidity η
	angle ϕ
primary vertex	significance of the two dimensional signed impact parameter
	significance of the three dimensional signed impact parameter
jet position	track transverse momentum, relative to the jet axis
	track pseudorapidity, relative to the jet axis
	ΔR of the track to the jet axis
jet energy	track momentum along the jet axis, in the jet rest frame
	same, normalized to jet energy
quality	χ^2 value of the track fit [SAF$^+$06]

Table 5.2: Input variables of the muon objects

the muons track is described by the χ^2 value of the track fit. No information about the detector components and muon identification variables are used. The full list can be seen in table 5.2. The muon objects are very pure. The fraction of misidentified pions, kaons and protons is 0.26%, 0.3% and 0.05% [CMS10k].

Electron candidates Details of the electron reconstruction can be found in section 4.3.4. The same variables as for the muons are used to describe the electron candidates. Electrons are difficult to identify. Therefore some variables which deliver information about the electron likeliness of the candidates are added. The electrons have a small mass. Variables to get information on possible bremsstrahlung are created. Also the output of a classifier which separates electrons from pions is used (tabular 5.3).

Secondary vertices Details of the secondary vertex reconstruction can be found in section 4.3.3. Additional to the real reconstructed secondary vertices the secondary vertex objects in this analyses contain so called pseudo vertices. The pseudo vertices are the sum of the four vectors of tracks displaced from the primary vertex, which do not require the secondary vertex requirements. The variables formed from the properties of these objects are listed in tabule 5.4. For the real secondary vertices in addition the distance of the vertex position to the primary vertex is calculated.

Jets Details of the jet reconstruction can be found in section 4.3.6. For the jet classification the mean values of the NeuroBayes output from the classifications of the sole objects corresponding to the jet are taken as additional input variables. Further we have the corrected four vector of the jet and the discriminating variables of all existing b-jet taggers (table 5.5).

5.3.2 NeuroBayes MC tagger (NBMC)

MC training is the common case how NeuroBayes (see also 5.1) is used. For calibrating the NeuroBayes expert two simulated samples are needed: one sample for the signal target S and one sample for the background target B. The calibration procedure is often called: training. A fully calibrated NeuroBayes expert is able to discriminate events with signal target from events with background target. This expertise is applied on the data sample. Thus each jet is related to the transformed NeuroBayes output variable o_t of the interval $(0, 1)$. Small value of o_t represent

category	input variables of electron candidate
four vector	momentum p
	pseudo rapidity η
	angle ϕ
primary vertex	significance of the two dimensional signed impact parameter
	significance of the three dimensional signed impact parameter
jet position	track transverse momentum, relative to the jet axis
	track pseudorapidity, relative to the jet axis
	ΔR of the track to the jet axis
jet energy	track momentum along the jet axis, in the jet rest frame
	same, normalized to jet energy
quality	χ^2 value of the track fit [SAF$^+$06]
	output of a mva electron/pion classifier [CMS10j]
	position of first hit in z direction
	position of first hit in radial direction
	inversed ΔR of first and last hit of the track
	ΔR of electron candidate and Gaussian sum filter track
	inverted energy of bremsstrahlung
	energy loss before calorimeter

Table 5.3: Input variables of the electron candidate objects

category	input variables of secondary vertex (SV)
four vector	mass of track sum at secondary vertex
primary vertex	2D distance of the SV to the primary vertex
	significance of 2D distance of the SV to the primary vertex
	3D distance of the SV to the primary vertex
	significance of 3D distance of the SV to the primary vertex
jet	ΔR of the SV to the jet axis
	ratio of energy at secondary vertex over total energy
track	number of track connected to the vertex
	ΔR of the SV to the track sum
	ratio of energy at secondary vertex over track sum
quality	category of secondary vertex (Reco, Pseudo, No)

Table 5.4: Input variables of the secondary vertex objects

category	input variables of jets
four vector	corrected transverse momentum p_T
	bare jet energy
	pseudo rapidity η
	angle ϕ
b tag	combined SV tagger
	jet B probability tagger
	simple SV tagger
	simpleSV high purity tagger
	soft muon impact parameter tagger
	soft muon transverse momentum tagger
	track counting high efficiency tagger
	track counting high purity tagger
objects	number of track objects
	number of secondary vertex objects
	number of electron candidates
	number of muon candidates

Table 5.5: Input variables of the jet objects

background-like events, larger values stand for more signal-like events.

For the NeuroBayes b-jet tagger a multi-level architecture was designed. The architecture of the NeuroBayes b-jet tagger is shown in figure 5.11. Calibrations of the NeuroBayes experts on each of the five physical objects, the tracks, secondary vertices, electron candidates, muons and jets are needed.

We get an estimate for each object, how likely it is to be part of b-jet. This information were collected for each jet. For the final NeuroBayes calibrations on jet-level new input variables were defined. These were constructed out of the output values of object-level experts.

The jet-level consists of two steps. It is easy to achieve a good separation between b-jets and non-b-jets because of the lifetime of the b-hadron. This is done by a first NeuroBayes calibration. To be more effective for the jets, which are difficult to separate, in addition a boost training was performed. For this the number of target 0 events was increased and weighted by the procedure introduced in 5.2.2.

For all calibrations NeuroBayes is setup with default parameters. The number of hidden layers is the number of input nodes minus one. Each input node is fed by one of the input variables. For all calibrations of the experts NeuroBayes is used in classification mode with the global preprocessing flag 422, which represents preprocessing and zero iteration training. The result is boosted by an internal artificial neural network training. The maximal number of iterations for the neural network is 100. Further each variable needs at least 2σ significance to be used for the classification.

The decision to use the internal boost mode is basically a aesthetic and less a technical. In most of the cases the weights in the neural network converge to zero and the training is stopped before 100 iterations. This means that it is not possible to enhance the result found by the zero iteration training. But the runtime is enlarged without a qualitative improvement. Nevertheless the output distribution is slightly different, not in separation power, but in the shape. Figure 5.12 shows two output distributions on the same target for the different setup modes. The runtime in zero iteration mode was only 35.43s in contrast to the internal boost mode where 290.47s passed by.

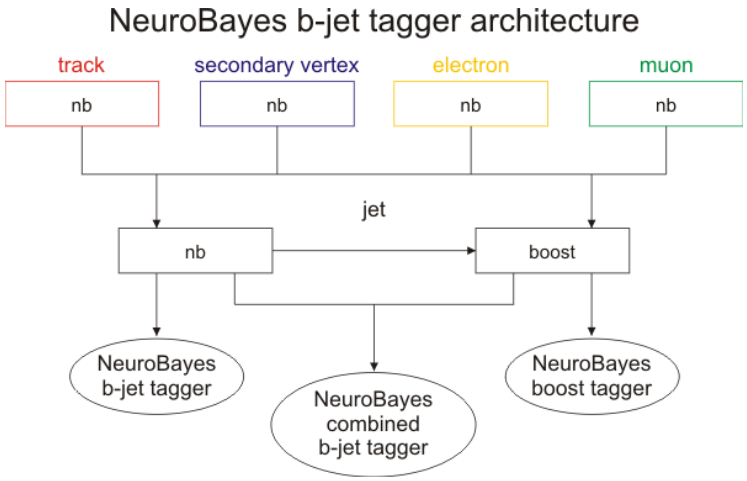


Figure 5.11: Architecture of the NeuroBayes b-jet tagger. Each box stands for a single NeuroBayes calibration. The arrows point up where the result of the expert is used. The colors clarify the different objects.

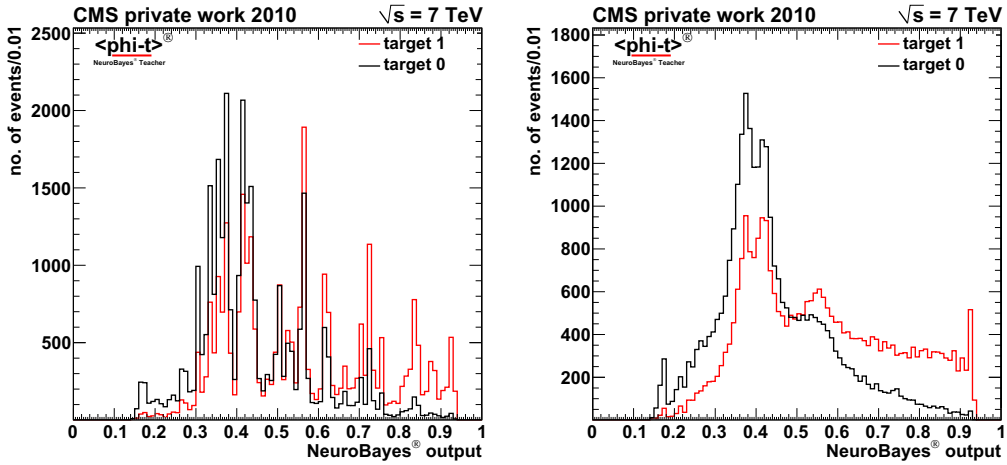


Figure 5.12: Left: the output distribution of the zero iteration mode. The training results in a Gini index of 19.9 calculated in 35.43s. A classification with the same setup parameters plus an additional internal boost results in a output distribution as shown on the right. The Gini index is also 19.9, but the shape is much smoother with a much longer runtime of 290.47s. The explanation of this effect is a different for the monotonous spline fit: NeuroBayes setup parameter DIAG instead of DIA2 5.1.3.

The explanation of this effect is a different value of the parameter which controls the curvature in the monotonous spline fit for the diagonalization (see 5.1.3). Instead of the common NeuroBayes shape parameter DIAG, the alternative DIA2 is used. The results of both setup modes are equivalent discriminators.
In this thesis the output values are used for another NeuroBayes training. Having a structure similar to many delta function can effect the preprocessing of the further NeuroBayes expert calibration. Therefore I decided use the smooth diagonalization mode (DIA2). Further I accepted the enlarged, but still small, runtime of the internal boost to get the best possible result.

For the calibrations of the experts the Pythia 6 QCD Tune2Z samples are used. As mentioned above the samples must be weighted by $w(\text{sample})$ to get a smooth realistic $p_{T,jet}$ spectrum (see 4.4). The available statistics are more or less flat in $log_{10}(p_{T,jet})$ (see also figure 4.13). This brings large weights which are problematic in the recent NeuroBayes version. To avoid weighting effects, the spectrum is transformed once more to a flat distribution. The $p_{T,jet}$ spectrum plotted in double logarithmic scale can be fitted by a polynomial function of the third order with a sufficient accuracy in the range 37 GeV $< p_{T,jet} <$ 1000 GeV:

$$f(p_{T,jet}) = \exp\left(a_0 + a_1 \log_{10}(p_{T,jet}) + a_2 \log_{10}(p_{T,jet})^2 + a_3 \log_{10}(p_{T,jet})^3\right)$$

The parameters are determined as follows:

$$a_0 = 39.86 \pm 0.89, \qquad a_1 = -24.79 \pm 0.50, \qquad a_2 = 8.39 \pm 0.17, \qquad a_3 = -1.582 \pm 0.048$$

Figure 5.13 shows the fitted spectrum. The final weights are calculated out of the Monte Carlo weights w multiplied with the extracted weight from the fit:

$$w_{final} = \alpha \cdot \frac{w(\text{sample})}{f(p_{T,jet})}$$

The constant factor α is chosen in a way, that the sum of all weights corresponds to the amount of statistics.

After this preprocessing the input variables are used for the various calibrations of NeuroBayes experts. The amount of statistics for the training is minimized as much as possible to avoid time consuming disk access operations. A main feature of NeuroBayes is, that good results for a discriminating output variable can be achieved, alreadz with a relative small number of training events. Further the calibration itself is very fast compared to other advanced multi variate analysis methods like boosted decision trees or artificial neural networks. A detailed study on this can be found in [Mar10]. Table 5.6 shows the amount of events used. Also the run time of the NeuroBayes training is listed. Large run time occure when the internal boost is able to improve the zero iteration result.
A list of the relevant input variables and the NeuroBayes output distributions are shown on the next pages. A complete list is attached in the appendix A.

Track training In table 5.7 the input variables of the track calibration are listed. They are sorted by their relevance for the classification.
The most important variable is the three dimensional significance of the impact parameter of the track (figure 5.14). The plots show the variable first in the usual exposition in logarithmic scale with equidistant bins and second with a binning calculated by the probability integral transformation 5.1.2. In both the distribution for tracks from b-jets and non-b-jets are plotted. It is easy to see, that with this input a good separation between signal and background can be established.

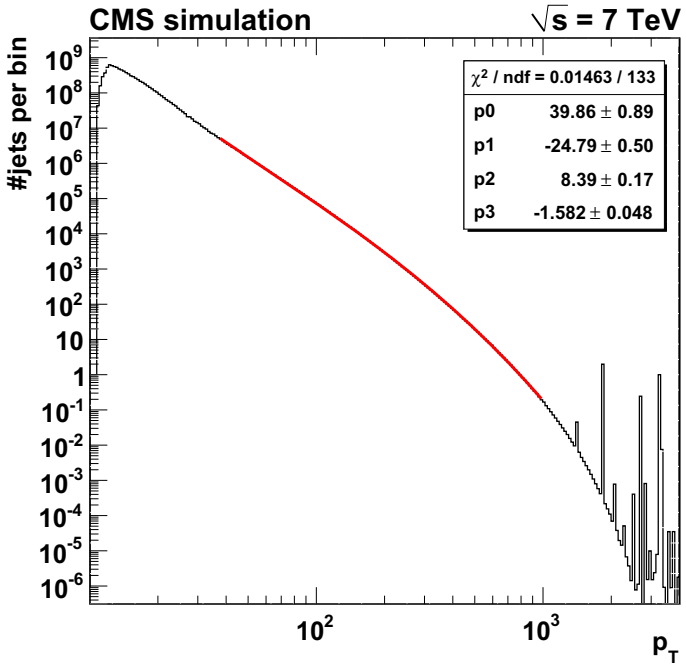


Figure 5.13: To transform the $p_{T,jet}$ spectrum into a flat distribution, it was fitted by a polynomial function at double logarithmic scale. The shown spectrum is calculated out of the avaliable MC events. Therefore different samples must be weighted by a specific value (see 4.4). This causes the binning effects on the right.

calibration	# t0 events	# t1 events	target 1 fraction	number of input vars	run time
track	184541	225460	50.0%	19	336.35s
vertex	362611	448418	51.1%	11	1969.51s
electron	145278	204932	53.3%	16	1396.76s
muon	253361	345744	48.3%	11	1335.45s
jet	192631	217312	48.2%	10	974.31s
boost	138772	433023	49.0%	10	167.72s

Table 5.6: Runtime of the different NeuroBayes calibrations. For the different trainings the number of events and the target 1 fraction is listed. Because of the weighing of the events this number does not correspond to the expected from the number of events. The last column shows how long it takes to get the calibration of the NeuroBayes expert.

name	added significance	only this	loss, when removed	correlation to others
trackSip3dSig	102.89	102.89	13.67	97.8%
trackEta	13.88	13.54	12.88	22.5%
trackBdistSig	8.81	53.16	6.01	98.4%
trackSip3d	7.71	98.48	7.23	97.4%
trackJetDist	9.85	67.93	7.24	82.9%
trackMom	7.70	10.32	6.86	24.9%
trackJetDeltaR	2.49	8.05	7.21	83.8%
trackPtRelFrac	5.71	8.73	7.28	85.2%
trackLxy	5.67	81.42	5.41	84.0%
trackChi2	4.10	2.71	4.27	5.8%
trackBDist	3.75	43.07	3.76	94.6%
trackHits	3.66	3.24	3.41	18.4%
trackPxHits	2.03	5.82	1.99	20.5%
trackBweight	1.82	53.53	1.83	97.8%
trackSip2dSig	1.64	92.41	0.56	97.0%
trackSip2d	0.39	86.67	0.39	96.3%
trackPparFrac	0.00	8.73	0.00	100.0%

Table 5.7: Input variables of the track object NeuroBayes classification. Only variables, which are more than 2σ in significance are used by NeuroBayes. The table also shows information about the classification power of the variable itself and how important it is, that this variable is used. The last column shows the correlation to the other variables.

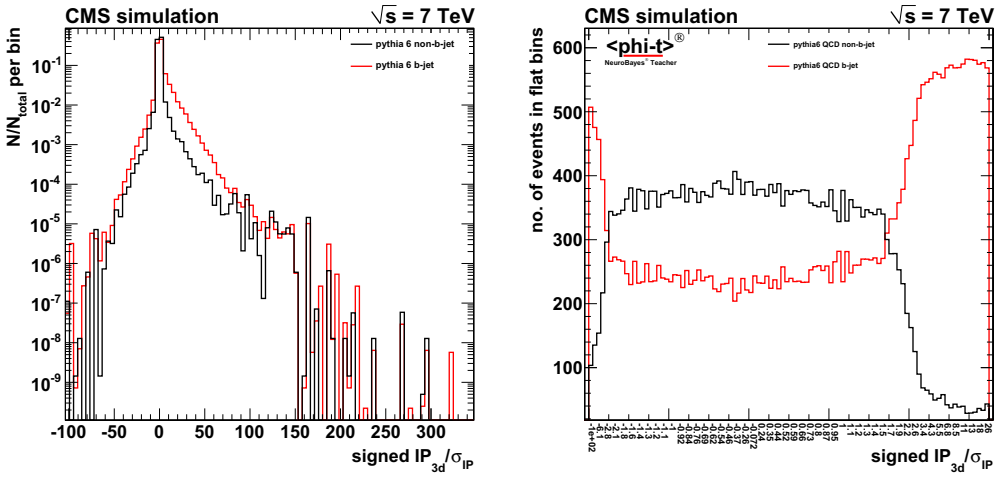


Figure 5.14: Track object training: The most important input variable is the significance of the signed impact parameter. The distribution of this variable is shown in the classical histogram with equidistant bins on logarithmic scale (left) and in probability integral transform (right). The b-jet tracks are plotted in red. The differences to the non-b-jet tracks (black) are clearly visible.

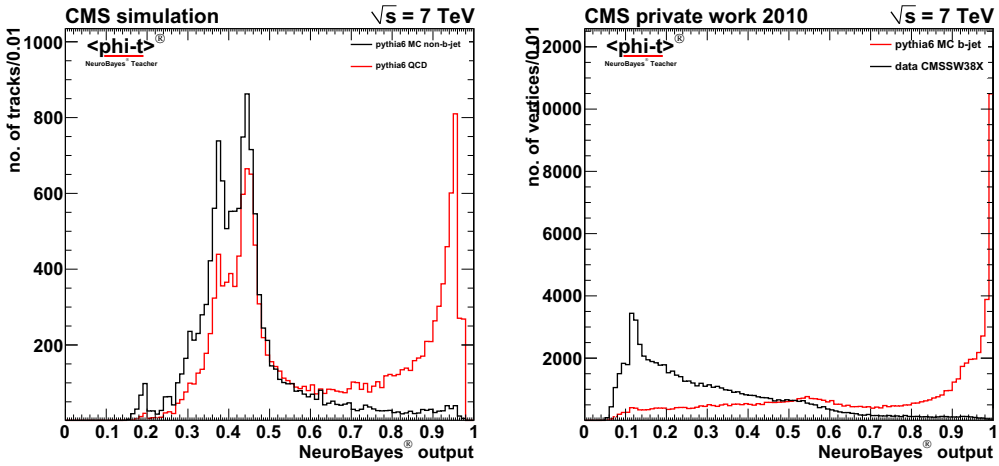


Figure 5.15: NeuroBayes output of the track and the vertex classification. In red are the objects coming from a b-jet in black are objects coming from other jets.

Having this variable in, the other variables are more or less corrections to this powerful one. Variables, which are highly correlated to it are ranked down. For example the two dimensional significance of the impact parameter is sorted out by NeuroBayes, because of the small additional information, which is left after the decorrelation. The parallel momentum of the track to the jet axis normalized by the jet energy is yet 100% correlated. This means NeuroBayes is able to reconstruct this variable from the other variables.
Figure 5.15 shows on the left the output distributions of track classification. Each event in this plot corresponds to one track. The two classes, tracks from a b-jet (red) and track from other jets (black), show the expected behavior. There is a good separation between signal and background. We see an interesting double peaking structure of the red curve. This results because of the fragmentation of the b-jet. The tracks of the right peak primarily correspond to tracks from the b hadron. The tracks from the left are mainly pions from the hadronization process. There are also large output values. This teaches the existence of tracks, which are very specific for b decays. On the other hand for low values there are no tracks in the first 20% of the output interval. This tells us that all kind of various tracks appear in b-jets. No single track can be excluded to stem from a b-jet.
This behavior is used to construct an additional input variable for the final b-jet tagger. Similar to the number of tracks corresponding to a secondary vertex, here the number of tracks which correspond to the b hadron candidate H_b, which should appear in the right peak, are counted. This is implemented by integrating the NeuroBayes output o_t of the track expert starting from a specific threshold $o_t > 0.5$ for each jet.

$$N_{track}(H_b) = N_{track}(jet) \int_{o_t=0.5}^{1} do_t(track)\, pdf(o_t)$$

Vertex training In tabular 5.8 the input variables of the vertex calibration are listed. The output of NeuroBayes is plotted in figure 5.15 on the right.
The most important variable is the number of tracks which are connected to the secondary vertex. This is caused by the large mass of the b hadron. The correlation of 77% to other variables, especially the secondary vertex mass confirms this statement. Another information contained in this variable is the fact, that the b-hadron needs at least one additional weak decay compared to the lighter quarks until it results in stable particles. This leads to an increased number of tracks

name	added significance	only this	loss, when removed	correlation to others
vertexNtracks	125.22	125.22	32.57	77.0%
vertexJetEFrac	63.54	122.79	35.01	72.2%
vertexPVSig2d	40.39	90.93	32.84	83.0%
vertexMass	28.87	112.81	25.16	78.7%
vertexTrackEFrac	18.65	11.99	17.19	35.6%
vertexPVDist3d	10.09	33.78	5.46	95.5%
vertexJetDeltaR	6.80	44.66	8.17	59.3%
vertexTrackDeltaR	5.57	8.39	5.57	40.9%
vertexCategory	0.41	66.40	0.39	69.1%
vertexPVDist2d	0.12	37.39	0.12	96.1%
vertexPVSig3d	0.00	91.15	0.00	100.0%

Table 5.8: Input variables of the vertex object classification. The columns show the relevance of each variable.

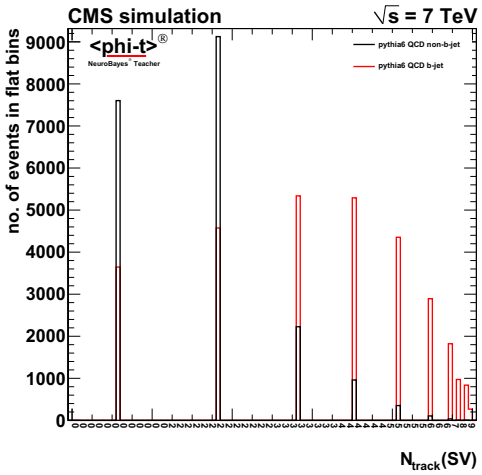


Figure 5.16: Secondary vertex object training: The most important input variable is the number of tracks connected to a reconstructed secondary vertex. The distribution of it is shown for vertices standing in b-jets (red) and non-b-jets (black). The bin with value zero shows only pseudo vertices.

connected to the secondary vertex. Figure 5.16 shows the distribution of this variable. The bin with no tracks correspond to pseudo vertices. If no secondary vertex is reconstructed, at least two tracks with large impact parameter are summed to this kind of object.
Other important variables are the vertex mass itself and the vertex energy compared to the jet energy. Both quantify the mass of the b hadron. Further the information on the lifetime of the b hadron are covered in the significance, how likely it is to have a secondary vertex away from the primary vertex (vertexPVSig2d).
All this information lead to a very good classification, if the reconstructed secondary vertex is part of a b-jet. Many of the secondary vertices are classified by almost 100% and appear in the last bin of the output distribution. The existence of a reconstructed secondary vertex is therefore already a good b-jet tagger.

Electron training In table 5.9 the input variables of the NeuroBayes electron calibration are listed.

name	added significance	only this	loss, when removed	correlation to others
eleSip3dSig	92.62	92.62	60.82	83.7%
elePtRel	19.91	18.71	12.08	66.9%
eleSip2dSig	13.09	69.81	12.82	83.9%
eleZpos	9.57	9.35	6.39	39.1%
eleInvDeltaR	8.75	6.12	6.58	28.7%
eleId	8.42	20.74	5.70	45.3%
eleMom	8.07	7.23	6.57	26.9%
eleChi2	6.78	5.10	7.15	17.0%
eleEta	6.76	17.79	5.83	68.8%
eleJetDeltaR	4.55	13.17	1.33	95.1%
eleBrem	4.40	10.43	4.63	44.9%
eleGSFDif	4.66	4.06	4.51	44.3%
elePhi	4.07	2.45	4.08	5.0%
eleJetPparFrac	0.98	5.13	1.14	69.7%
eleEtaRel	0.76	16.45	0.76	95.8%
eleJetPpar	0.00	18.71	0.00	100.0%

Table 5.9: Input variables of the electron candidate object classification. The columns show the relevance of each variable.

The electron candidates are more or less a subgroup of the track objects. So it is not surprising that also the three dimensional significance of the signed impact parameter contains the most important information for the classification. The same arguments due to the lifetime of the b hadron apply here. Again the other variables are more or less corrections to this powerful one.
But there is another interesting variable, which is more relevant to distinguish electron candidates from b-jets. This is the transverse momentum of the electron candidate relative to the jet axis $p_{T,rel}$. Because of the lepton decay of the b hadron into a electron it is possible, that the electron carries much of the momentum from its mother particle. This leads to larger relative momenta $p_{T,rel}$. The distribution of this variable is shown in figure 5.17.
Figure 5.18 shows the output distributions of the electron classification on the left. Each event in this plot corresponds to one electron candidate. The two classes, electron candidates from a b-jet (red) and from another jet (black), show the expected behavior. There is a good separation between signal and background.
We expects also two peaks in the signal distribution as seen for the tracks. There are two effects which cause the shape of the output distribution.
At first the electron reconstruction is difficult because of the multiplicity of tracks. Many of these tracks correspond to pions. To reduce the contribution of misidentified particles a selection dependent on a good electron identification is needed. For this additional electron quality variables are used. Nevertheless non electron particles remain and form the peaking structure on the left.
The other effect depends on real electrons not coming from b hadrons. Because of the material in the tracking detector, photons can create electron/positron pairs. These particles are part of the jets and arise for b-jets and non-b-jets in rather large impact parameter values. Compared to the track object NeuroBayes output distribution the expected peak on the right is reduced.
The events in the left peak correspond again to tracks not coming from the b hadron.
Finally an additional input variable is constructed for the final b-jet tagger. This is implemented by integrating the NeuroBayes output of the electron expert starting from a specific threshold

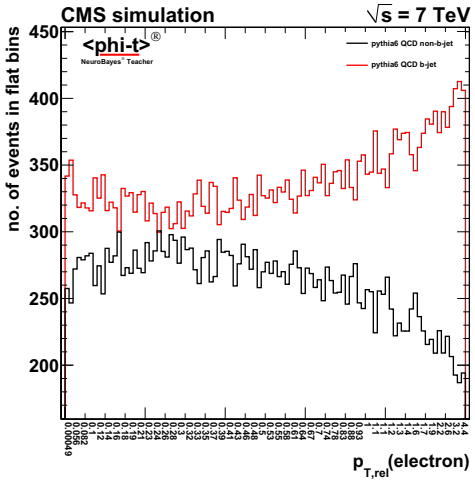


Figure 5.17: Electron candidate object training: Distribution of the transverse momentum of the electron relative to the jet axis.

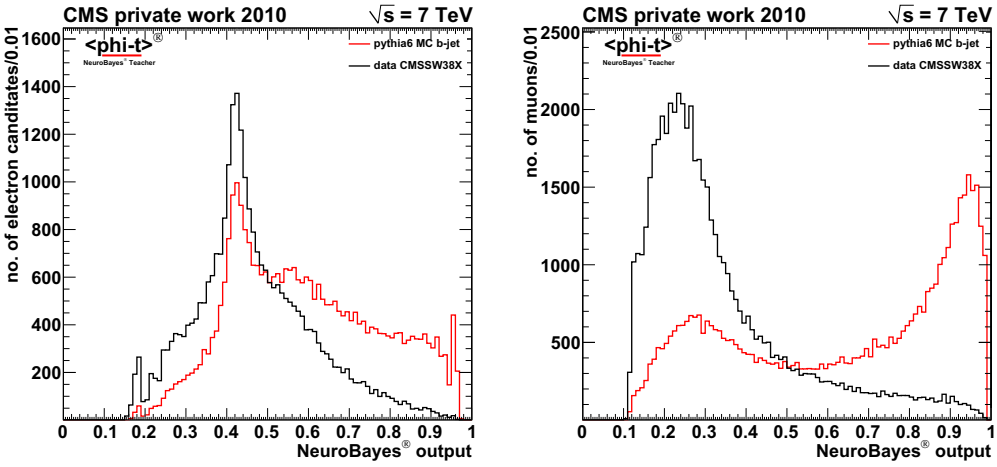


Figure 5.18: NeuroBayes output of the lepton classification. Left: electron candidates. Right: muons. The b-jet objects are plotted in red.

name	added significance	only this	loss, when removed	correlation to others
muonSip3dSig	92.21	92.21	42.00	86.3%
muonChi2	53.05	68.84	49.89	25.2%
muonPtRel	24.56	42.32	8.88	85.3%
muonJetPparFrac	16.41	31.97	11.95	85.8%
muonEta	10.70	14.69	9.05	19.4%
muonEtaRel	5.13	20.04	6.24	95.2%
muonJetDeltaR	5.82	18.98	5.67	94.7%
muonPhi	3.02	3.11	3.01	1.4%
muonMom	2.83	22.36	2.84	88.2%
muonSip2dSig	2.51	76.31	2.51	85.9%
muonJetPpar	0.00	42.32	0.00	100.0%

Table 5.10: Input variables of the muon candidate object classification. The columns show the relevance of each variable.

$o_t > 0.5$ for each jet.

$$N_{electron}(H_b) = N_{electron}(jet) \int_{o_t=0.5}^{1} do_t(electron)\, pdf(o_t)$$

Muon training In table 5.10 the input variables of the NeuroBayes muon calibration are listed. The most important variable is again to have a significant impact of the muon. The NeuroBayes output distribution is shown in figure 5.18 on the right. Opposite to the electron case, the muons are easy to detect. Because of the large muon system of CMS a detailed muon identification is not needed. More problematic is the extrapolation of the muons into the tracking detector and the mapping to a jet. Here the reconstruction quality becomes an important variable. The transverse momentum of the muon relative the jet axis is important again. The argument for this is the same as for the electrons, because the muons appear mostly from the weak decay of the b hadron.

Jet training Finally the input variables of table 5.11 are use for the NBMC b-jet tagger. The input variables are constructed from the NeuroBayes outputs for the different objects. For each jet the number of the specific objects found in the jet was created. For the tracks and the electron in addition the good candidates are counted.
For each jet a the NeuroBayes outputs o_i are combined. Under the presumption that each output is an independent estimate of the probability to be part of a b-jet, it is possible to combine the values by multiplying the likelihood ratios $\Lambda(o_i) = k \cdot o_i/(1 - o_i)$, where k is the ratio of the two targets used for the calibration.

$$\Lambda(\text{b-jet}|o) = \prod_{i}^{N} \Lambda(o_i)$$

Finally a jet probability $P(\text{b-jet}|o)$ is defined:

$$P(\text{b-jet}|o) = \frac{\Lambda(\text{b-jet}|o)}{1 + \Lambda(\text{b-jet}|o)}$$

The independence assumption is not entirely correct for objects like ours, because of correlations between them. To get the probability right, corrections have to be applied. Nevertheless without

name	added significance	only this	loss, when removed	correlation to others
jetTrackProb	196.07	196.07	58.38	90.6%
jetNSV	34.82	168.53	36.38	77.0%
jetNMuon	18.74	45.35	19.11	16.5%
jetVertexProb	15.74	50.93	15.05	35.1%
jetElectronProb	8.23	61.38	5.46	82.6%
jetNTrack	4.31	32.04	5.64	72.4%
jetNGoodTrack	3.97	163.61	4.09	91.3%
jetMuonProb	3.63	39.69	3.64	18.9%
jetNEle	2.26	26.08	2.11	30.8%
jetNGoodEle	0.68	52.67	0.68	82.5%

Table 5.11: Input variables of the jet classification. The columns show the relevance of each variable.

any correction the variables have good discrimination power and can be used for the construction of a b-jet tagger.
In table 5.11 the input variables of the NeuroBayes jet calibration are listed.
The most important input variable is the probability estimate calculated from the track objects. This is obvious, because it contains almost the whole lifetime information of the b hadron and is calculable for each jet. The other variables are more or less corrections to this main input variable. Further the existence of a secondary vertex or a muon are important informations for the identification of a b-jet. At last let us have a look at the variable of good tracks. Compared to the total number of tracks in the jet, the correlation to the target is strongly increased. The significance is in the order of the significance from the number of secondary vertices. The statement that we count tracks from b hadrons seem to be true.
The NeuroBayes output is shown in figure 5.19. The output is plotted in logarithmic scale for the y-axis. Because of the powerful separation between the two classes, most of the jets are in a few bins at low and high values. This was achieved using around 400000 events for the calibration of the NeuroBayes expert. The powerful separation indicates how simple a construction of a b-tagger is. More difficult is to improve this.

Boost training To gain the performance we need a boost training. Using the simplest boost weight will cause some problems. Because of the already powerful separation, most of the events would get a very small weight (see 5.2.2). This makes a further separation very difficult. The effective statistics after the weighting are small. To get into the whole advantages of a boost training, the number of events for the calibration must increase.
At first the probability interpretation of the NeuroBayes output must be tested. The right plot in figure 5.19 shows the existence of this property. The calculated purity for each bin compares to the diagonal within the statistical uncertainties. The probability interpretation is correct for the given binning.
For the external boost a NeuroBayes training with the same setup as the previous one was executed. As mentioned above there are different approaches to implement such a boost training. In this section I will focus on two different approaches. The first tries to improve the b-jet classification looking in more detail into the b-jet distributions. The weights are only applied on the background sample. The other ansatz does a weighting also on signal sample but balances the effective statistics to be in the same order as the real statistics.

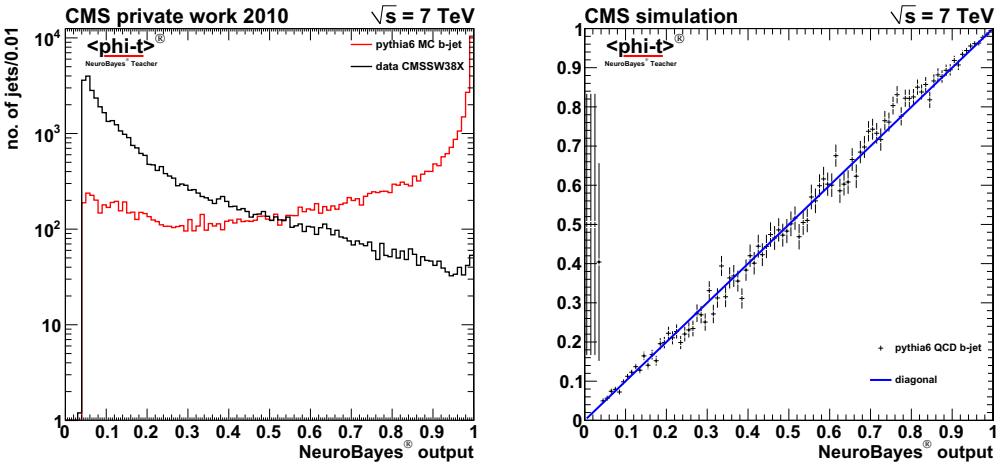


Figure 5.19: NeuroBayes output of the jet classification (left). red: the b-jets, black: other jets. There are no jets in the first 5% of the output distribution. No cut was applied here. B-jets without lepton and small lifetime look like a light jet. In the right plot the purity of the b-jet distribution is shown. If the purity of each bin matches with the value of the NeuroBayes output variable the probability interpretation is fulfilled. This property is required for the boost training.

Purity tagger. For the first we use the focusing function $F_f^{pur} = \frac{1}{1-P(S|o_t)}$. Thereby the weights for the target 1 events are always $w_S = 1$. Taking the changed statistics for the boost training into account brings us the following weights, which has to be applied on the target 0 events.

$$w_B = \frac{P(S|o_t)}{1-P(S|o_t)} = \Lambda_t \frac{N_b(S)}{N_b(B)} = \frac{o_t}{1-o_t}\frac{N_t(B)}{N_t(S)}\frac{N_b(S)}{N_b(B)}$$

$N_b(S)$ and $N_b(B)$ describe the number of events used for the boost training, while the number of events used in the first training are quoted as $N_t(S)$ and $N_t(B)$.

As control variable the output distribution of the unboosted training is added. This variable should not effect the boost training and does not have correlations to the target.

An overview of the input variables can be seen in table 5.12. More detailed information about the vertex is now the most important. Also the b-jet probability estimated from track informations is still an important variable. In the first training not the complete information it contains could be extracted and used for the classification. In figure 5.20 the distribution of this variable, as used for the first training, is shown. The background (black, simulated non-b-jets), and signal (red, simulated b-jets) are plotted separately and the good separation can be seen. To avoid overtraining, NeuroBayes reduces the dependency on statistical fluctuations. Instead of the red curve, the regularized blue curve is taken for the calibration. This procedure can affect some information loss, especially if the statistics are small in some regions of the variable.

Applying the weights calculated for the boost training, the black curve is transformed into one similar to the blue. For the boost training we want to be focused on the purity region, the statistics were increased and we get a distribution as shown on the right for this variable. With the enlarged statistics NeuroBayes is able to see the differences of the two shapes, which were not apparent in the first training, due to the low statistics of the background in this region.

One can observe that the $o_t(binned)$ gives a relevant contribution for the boost classification. This should not happen if the calibration of the first expert is perfect. Looking at the distribution shows us the cause of the remaining correlation to the target after the weighing (figure 5.21).

The target dependency occurs for large values of the output distribution at almost one. All these

name	added significance	only this	loss, when removed	correlation to others
jetVertexProb	20.21	20.21	15.65	48.1%
jetTrackProb	12.15	17.79	11.24	50.8%
$o_t(binned)$	11.85	4.78	6.75	64.1%
jetNTracks	7.49	7.83	6.67	25.4%
jetElectronProb	4.24	5.41	4.96	22.3%
jetNGoodEle	4.68	3.42	3.94	39.0%
jetNMuon	3.08	11.93	3.51	57.2%
jetNGoodTrack	2.57	8.47	2.71	41.7%
jetMuonProb	2.54	4.54	2.60	17.3%
jetNSV	2.02	2.05	2.12	25.3%
jetNEle	2.00	3.77	2.00	37.3%

Table 5.12: Input variables of the boost training for the jet classification. The columns show the relevance of each variable.

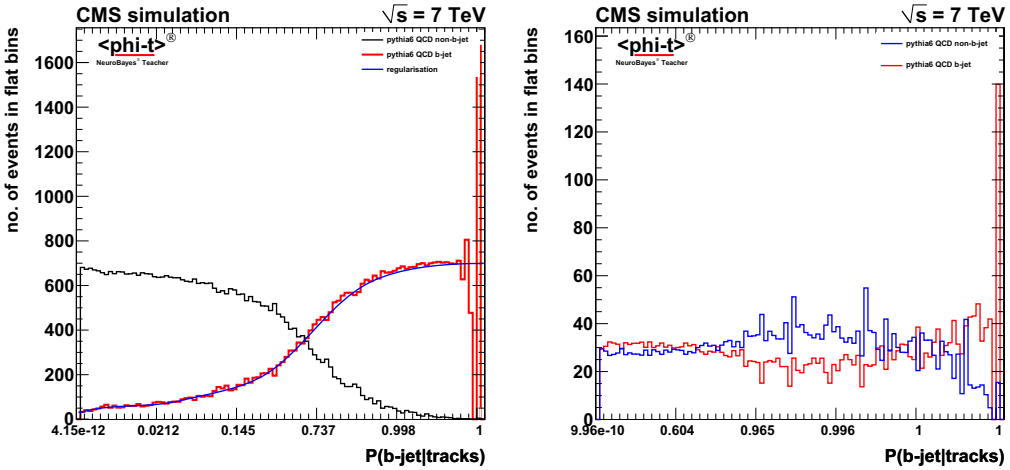


Figure 5.20: The plots show the distribution of the track probability, which is used as input variable, for the unboosted (left) and the boosted NeuroBayes training. The increase of statistics and the reweighing cause a gain of information for an improvement of the b-jet tagger.

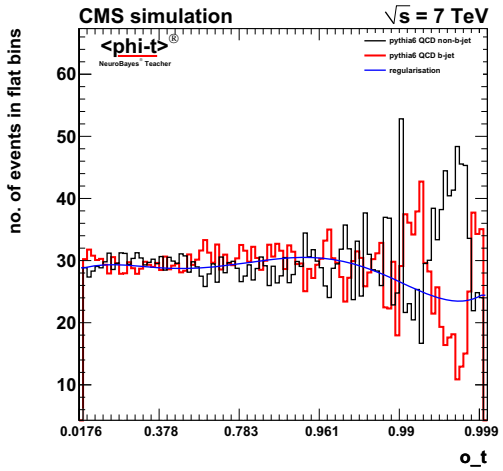


Figure 5.21: The plots show the distribution of the NeuroBayes output of the main expert, which is used as input variable, the boosted NeuroBayes training. The probability interpretation was tested on a well defined binning. The weighting allows us to see the structure within the bins.

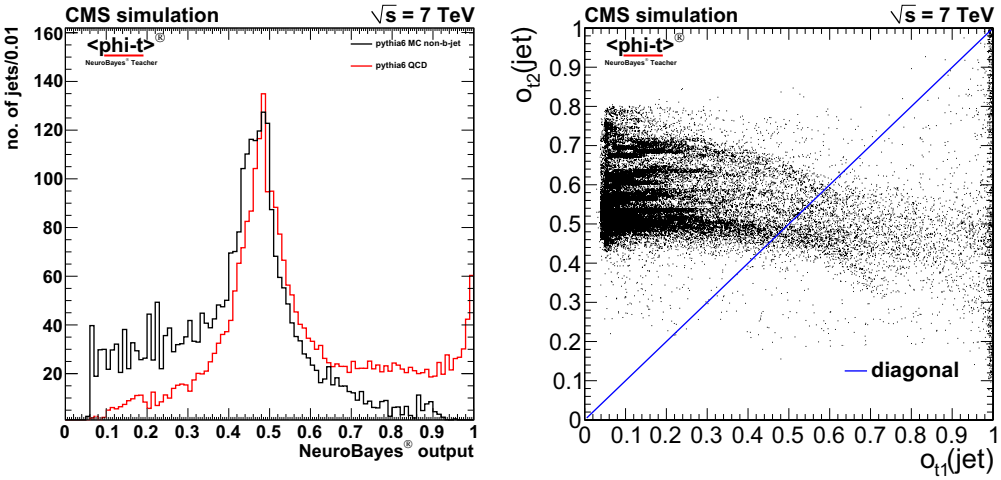


Figure 5.22: NeuroBayes output of boost training for the jet classification. In red are the b-jets, in black are other jets.

events belong to one bin of the diagonalization histogram. NeuroBayes is not able to resolve events in such a detail.

More insecure is another fact. The weights are large in this region. As shown before is there a problem with the correct error propagation for events with large weights. This can effect the regularization of this special input variable. For the final b-jet tagger I left this variable out.

Figure 5.22 shows the output distribution of the boosted NeuroBayes MC training on the right. Compared to the unboosted NBMC we see a smaller separation between the two classes. This is expected because all informations used in the first training are not included in the second one. The main advantage is that the both classification less correlated to each other. In figure 5.22 on the right a scatter plot of the two variables is shown.

To get a combined NeuroBayes b-jet tagger, which uses the results of the two calibrations, the likelihood ratios of the single training can be multiplied (see section 5.2.2). This results in a final powerful discriminator to identify b-jets. Starting from now I will call this b-jet tagger: NeuroBayes combined purity tagger (NB comb Pur). The performance of this tagger will be shown after the

name	added significance	only this	loss, when removed	correlation to others
jetNGoodEle	15.72	15.72	16.21	23.4%
jetTrackProb	13.11	11.78	11.82	16.8%
jetVertexProb	10.10	11.12	9.67	12.9%
jetNTracks	7.07	9.67	6.28	24.3%
jetMuonProb	5.65	6.07	5.83	10.0%
$o_t(binned)$	4.16	3.83	4.15	29.2%
jetElectronProb	3.88	4.52	4.06	8.0%
jetNSV	3.07	0.16	2.80	26.1%
jetNMuon	2.05	2.36	2.04	9.2%
jetNGoodTrack	2.02	1.41	1.86	18.0%
jetNEle	1.29	1.96	1.29	14.9%

Table 5.13: Input variables of boost training for the jet classification. The columns show the relevance of each variable.

introduction of another tagger optimized on the efficiency region.

Efficiency tagger. In addition I want to construct another b-jet tagger, which is more performant in the efficiency region. The focusing function is chosen in a way that the effective number of events is conserved over the spectrum of the NeuroBayes output distribution o_t. Therefore this variable was studied to create an additional weight factor dependent on their shape.
Normally the signal and background events are weighted with a specific weight: $w_S = P(B|o_t)$ and $w_B = P(S|o_t)$. The effective number of events in o_t is therefore reduced by $\alpha = P(B|o_t)\, pdf(o_t|S) + P(S|o_t)\, pdf(o_t|B)$. To balance this we need the following focusing function:

$$F_f^{eff} = \frac{pdf(o_t)}{P(B|o_t)\, pdf(o_t|S) + P(S|o_t)\, pdf(o_t|B)} = \frac{P(S)\, P(B)}{P(S|o_t)\, P(B|o_t)}$$

The application of the weights $w_T = (1 - P(T|o_t))\, F_f$ with the targets $t = S, B$ does not change the effective number of events.

A NeuroBayes expert calibration was arranged. The overview of the input variables is shown in table 5.13.
The ordering by relevance of the input variables changed when applying the other focusing function. Especially the electron informations are more important.
The variables are less correlated to each other which point to that the first calibration found some of the dependencies between the variables. Also the output variable of the former training is less important compared to the boost in the purity region, where we had the issue with large weights.
The NeuroBayes calibration results in the output distribution plotted in figure 5.23. The separation seems to be less developed than for the former boost training. This is deceiving because of the different weighing. The two output distributions are not comparable.
For the final b-jet tagger also a combination with o_t must be done. Again the likelihood ratios of the two training are multiplied as described in section 5.2.2. Starting from now I will call this b-jet tagger: NeuroBayes combined efficiency tagger (NB comb Eff).

Having two kinds of boosted NeuroBayes b-jet tagger available it would be interesting to compare them. This comparison is done on a independent sample. We must produce so called performance plots.

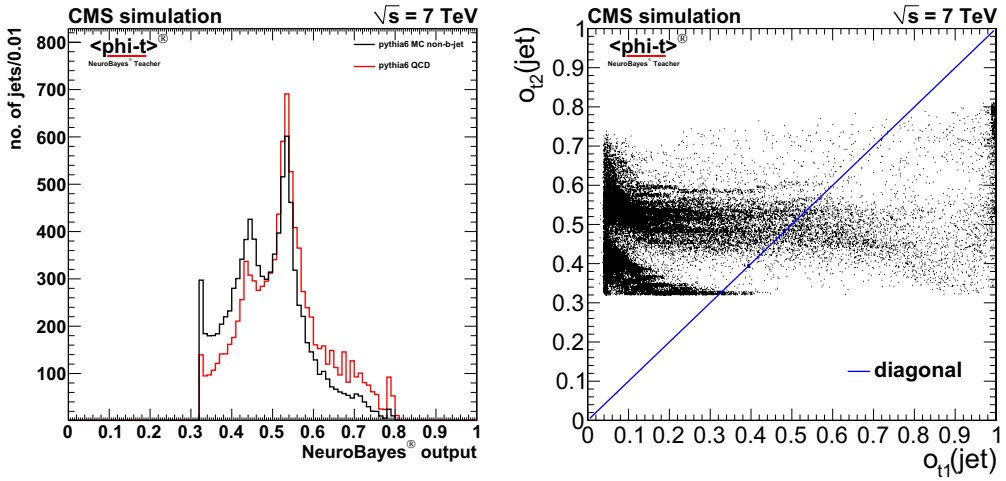


Figure 5.23: NeuroBayes output of the alternative boost training for the jet classification. In red are the b-jets, in black are other jets.

At CMS performance plots are used where the mistag rate is plotted against the b-jet efficiency. Further comparisons using the purity are also possible. The already existing b-jet taggers from CMS are presented in 4.3.9. To compare the NeuroBayes b-jet tagger only the most separating of the existing b-jet taggers are taken - the combined secondary vertex b-jet tagger. Figure 5.24 shows the performance plot for all NeuroBayes b-jet taggers: the unboosted, the boosted and the combined.

In the plot we see the already shown performance of the combined SV b-jet tagger and the NeuroBayes b-jet tagger we wanted to improve. The results of the two different boost calibrations are also plotted, called boost and alternative boost. As expected the boost on purity region is less performant over a wide range. Most of the events were weighted to a very small number, that improvement in this regions are impossible. The alternative boost is more sensitive to jets in this region. After the combination with the former NeuroBayes b-jet tagger is improved in the different regions dependent on their assignment. In the efficiency region there is not enough space for improvements. Only a tiny shift is visible. This differs in the purity region. Both combined tagger can gain in this region. The purity tagger, which is optimized for this region gives the strongest enhancements. On this sample we found a almost exponential separation over the whole range.

This brings us to the conclusion for the MC based b-jet tagger. The main advantage of a MC training is, that signal and background is well defined. Regardless which fraction of signal to background events is arranged, for the NeuroBayes classification we get an output with the best discrimination power for these two classes. But for applying the expertise on data, the simulation of our two classes must be quite well. For the b-jet we can be quite confident, that the simulations describe data well. Indeed the production mechanism, production rate and fragmentation is not well understood, but we have good understanding of the lifetime, the mass and the lepton branching ratio and these are the informations we use for b-jet tagging.

For the background sample it is more difficult. We have already seen that noise tracks are not good simulated. Further we have underlying events and pile up, which is hard to simulate. To get all background sources in good shape much work must be done, if it is possible at all.

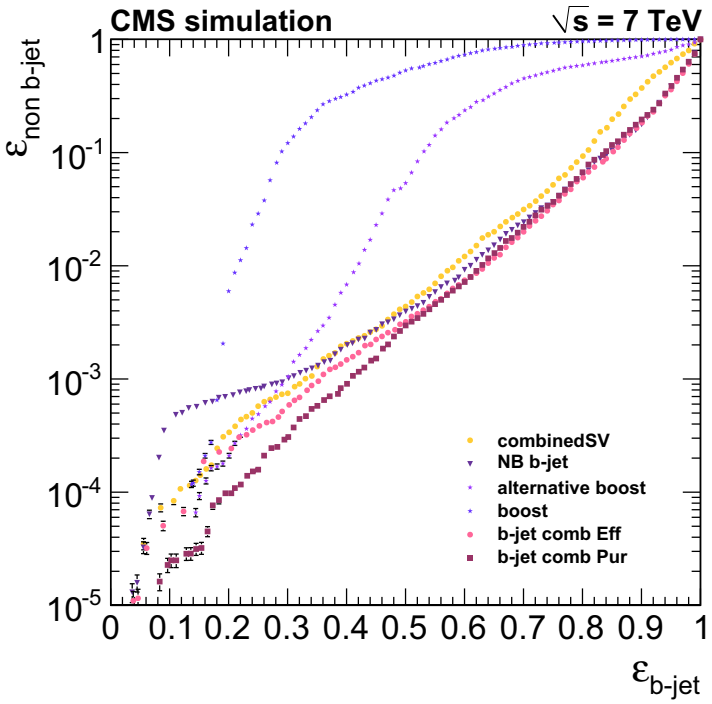


Figure 5.24: Performance of the NeuroBayes b-jet taggers. All new taggers are plotted together with the best existing tagger from CMS.

5.3.3 MC to data comparison

Having a good b-jet tagger on MC, a good performance on data is not guaranteed. All input variables have to be checked how they compare with data. If there is good agreement it is more likely that the b-jet tagger works on data.
Therefore this is one of the main tasks for the commissioning of the CMS experiment and b-jet tagging. If there is a good agreement between simulations and data, we know that the detector is well understood and usable for physics studies. Otherwise we have to restrict the field for analysis to the known regions and study the misunderstood areas for their issues.
It is possible to use NeuroBayes for this comparison of data and Monte Carlo simulations.

NeuroBayes is setup with default parameters. The number of hidden layers is chosen to the number of input layers minus one. Each input node is fed by one of the input variables. NeuroBayes is used in classification mode with the global preprocessing flag 422, which represents preprocessing with the internal boost training. The maximum number of iterations is set to 100. No BFGS algorithm is applied. Each variable needs at least 2σ significance to be used for the classification.
For such a study a NeuroBayes expert is calibrated on the sample of the simulated events as the one target and the events of the data sample as the other target. So the output distribution discriminates between simulation and reality. If data and simulation agree well, the NeuroBayes output distribution should be compatible with statistical fluctuations around the a priori fraction of the data sample and the simulated sample.
The output distributions for all NeuroBayes comparisons can be found in appendix B.

Besides testing the quality of the simulation, NeuroBayes provides the tools to identify the variables which cause differences between the two samples. It comes with an automatic calculation of the relevance of each variable to separate the two classes (see 5.1). In our case we are pointed directly to the variable with the largest discrepancy between data and simulation.

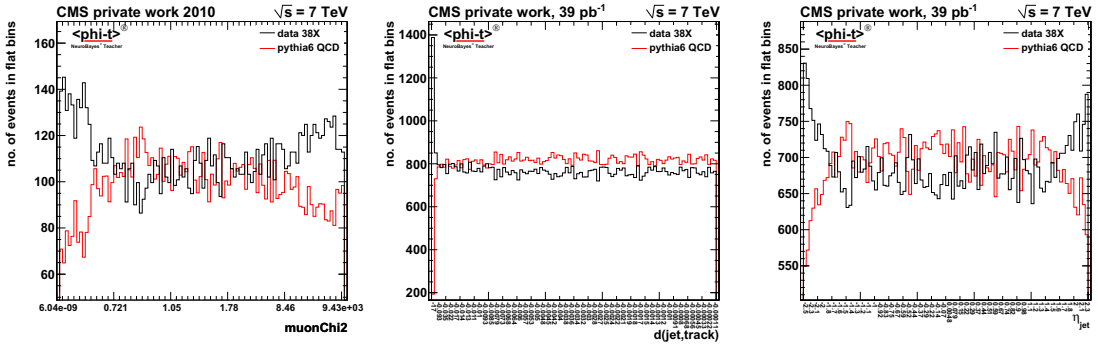


Figure 5.25: Left: Flat distribution given by the NeuroBayes comparion for the muon χ^2. Middle: Flat distribution of the distance of a track to the jet axis. For large distances effects of noise tracks can be seen. Right: Comparison of the η distributions of data and Pythia 6 QCD MC. In the barrel region $\eta < 1.5$ is a good agreement is observed. In the forward region large differences can be seen.

For our analysis we can use this feature to test how well simulation describes data. This study was done for all input variables and different triggers, because they correspond to different momentum ranges. The result is shown in tables 5.14 and 5.15. The order of the variables is the same as in the tables above, but for the description an abbreviation is taken. In the following I will use this abbreviations. The entries represent the correlation coefficients to the target.

When looking at the results of the comparison, one can notice the following things: Larger values represent more differences between data and simulations. This happens mostly for the muon candidates, which implies that muons in jets are not well simulated. The main difference appear in the χ^2 values of the muon track fit (figure 5.25). The exact cause could not be identified, because the muon objects contain different types of muons (see section 4.3.5) and it is possible that the difference is effected by one single type. For a final clarification we have to wait for an update from the CMS muon physics object group (POG).

Further we see an effect of the jet energy correction. Before the correction the discrepency is larger than after (jetEnergyUCorr vs. jetPt). This effects also the variables calculated relative to the jet energy, like the momenta of the tracks (PtRel and Ppar). If this quantities are normalized to the jet energy, the fraction looks quite fine (trackPtRelFrac and trackPparFrac). So I decide only to keep the normalized ones.

Another issue can be seen on the signed impact parameter variables. For the tracks the two dimensional $r\phi$- calculation is more reliable than the three dimensional one. Because the large correlation between these, I decided to take only the two dimensional ones.

The most important difference can be seen in the distance between track and jet axis (figure 5.25). Unfortunatly the variable is multiplied with a factor -1. But nevertheless we can see that in data tracks with a very large distance to the jet axis were found. This could happen by noise tracks which are connected to the jet object and not simulated in Monte Carlo. To reduce this noise I applied a cut at $d(jet, track) > -0.1$.

Also the ΔR distribution of the vertex to the jet axis as well as the sum of the jet tracks is not well simulated.

As already seen in the corrected jet spectrum of the transversal momentum is there a discrepancy for the low energy jets. In addition also a difference on jet and track level in η is found. This is in the forward region $|\eta| > 1.5$ (figure 5.25). For the construction of a b-jet tagger I will restrict to the barrel region and $p_T > 84$ GeV.

name	Jet15U	Jet30U	Jet50U	Jet70U	Jet100U	Jet140U	
trackMom	3.0	2.3	2.3	2.1	2.2	1.6	
trackEta	2.7	2.5	2.9	2.7	2.8	2.6	
trackSip2dSig	2.1	1.2	1.0	1.4	1.2	1.1	
trackSip3dSig	5.9	4.4	3.6	4.6	5.0	4.9	
trackSip2d	1.3	1.2	1.4	1.8	1.6	2.1	
trackSip3d	5.8	3.9	3.5	4.5	4.9	5.1	
trackLxy	6.3	4.3	3.4	4.3	3.8	3.9	
trackPtRel	5.3	5.2	4.9	5.1	5.1	4.7	
trackPpar	3.0	2.3	2.3	2.1	2.2	1.6	
trackJetDeltaR	1.4	1.9	1.8	2.6	2.9	3.0	
trackJetDist	8.0	6.3	5.0	6.4	6.6	6.8	
trackPtRelFrac	0.8	0.8	0.8	1.6	2.0	2.3	↗
trackPparFrac	0.8	0.8	0.8	1.6	2.0	2.3	↗
trackChi2	4.3	3.0	2.8	3.3	4.4	5.2	
trackPxHits	1.5	2.6	2.6	2.9	3.1	3.9	↗
trackHits	1.5	3.1	2.6	3.4	3.9	4.8	↗
trackBDist	6.9	5.5	4.6	5.8	5.8	6.2	
trackBdistSig	5.9	4.7	3.7	4.6	4.8	4.9	
trackBweight	6.6	5.1	3.9	5.1	5.1	5.2	
muonMom	1.5	5.0	4.7	3.8	2.6	2.1	
muonEta	3.1	1.1	7.2	6.3	4.1	4.0	
muonPhi	1.3	1.3	1.3	1.3	2.8	2.6	
muonSip2dSig	2.4	2.3	1.4	1.0	1.0	1.5	
muonSip3dSig	0.9	1.8	1.6	2.0	1.4	3.3	
muonPtRel	6.9	7.6	9.3	6.9	6.3	6.3	
muonEtaRel	5.8	5.8	5.9	4.5	5.7	4.8	
muonJetDeltaR	5.8	4.9	4.1	3.2	3.5	3.8	
muonJetPpar	6.9	7.6	9.3	6.9	6.3	6.3	
muonJetPparFrac	1.4	4.7	5.4	4.3	3.0	2.8	
muonChi2	7.4	8.1	10.5	9.5	9.8	11.3	
eleMom	0.4	1.1	1.3	2.2	1.5	2.2	
eleEta	2.5	0.9	0.4	1.6	1.4	1.4	
elePhi	0.7	0.2	1.2	1.1	0.7	2.1	
eleSip2dSig	1.1	0.8	1.1	1.6	1.8	2.7	
eleSip3dSig	2.3	0.8	0.5	2.3	0.5	3.2	
elePtRel	5.7	5.1	4.5	4.2	4.4	4.0	
eleEtaRel	3.6	3.3	3.7	3.8	3.9	4.5	
eleJetDeltaR	4.2	3.8	4.4	4.4	4.3	4.5	
eleJetPpar	5.7	5.1	4.5	4.2	4.4	4.0	
eleJetPparFrac	1.2	1.1	1.1	1.8	1.0	1.4	
eleChi2	3.0	3.3	2.8	3.7	4.4	4.1	

Table 5.14: Result of the data versus MC comparison. The single values represent the correlation coefficients of the listed variables to the target (data/MC). Larger values mean larger differences between data and simulation. In the last column some variables are marked with a small arrow. This illustrates that the differences between data and MC increase for the different jet momenta.

name	Jet15U	Jet30U	Jet50U	Jet70U	Jet100U	Jet140U	
eleId	1.2	1.5	1.9	2.2	2.2	1.8	
eleZpos	3.7	3.0	3.9	3.8	2.5	1.6	
eleInvDeltaR	5.3	3.2	4.1	3.0	3.3	3.1	
eleGSFDif	1.1	1.5	0.3	1.1	0.2	0.5	
eleBrem	1.1	0.6	0.9	1.4	1.7	2.2	
vertexMass	1.7	2.9	3.2	3.1	3.7	3.4	
vertexPVDist2d	2.9	1.5	3.8	2.5	2.7	4.9	
vertexPVSig2d	1.8	1.4	2.0	1.3	0.6	1.4	
vertexPVDist3d	1.7	1.0	2.5	2.1	2.4	4.5	
vertexPVSig3d	1.8	1.8	2.0	1.3	0.4	1.4	
vertexJetDeltaR	4.9	3.9	4.1	5.2	5.0	6.6	
vertexJetEFrac	1.9	4.0	3.3	2.7	2.9	1.2	
vertexNtracks	3.1	2.9	1.6	1.9	2.0	3.4	
vertexTrackDeltaR	7.0	7.0	7.7	8.5	9.2	8.0	
vertexTrackEFrac	2.4	3.5	4.1	3.5	3.2	3.7	
vertexCategory	1.6	1.6	0.2	0.2	0.3	1.5	
jetPt	**2.8**	0.6	0.2	0.2	0.2	0.4	
jetEnergyUCorr	3.8	3.8	4.2	3.6	3.6	2.6	
jetEta	3.8	3.9	4.3	3.9	3.8	3.1	
jetPhi	1.3	1.7	2.0	2.0	2.2	2.3	
jetNTrack	5.6	5.4	5.3	5.0	4.5	3.8	
jetNSV	0.2	0.2	0.3	0.1	0.4	0.3	
jetNEle	0.7	0.9	1.7	2.8	2.8	3.7	↗
jetNMuon	0.5	0.5	0.7	0.8	0.8	0.8	
jetCSV	4.7	4.8	4.2	4.7	4.7	5.7	
jetBProb	2.6	1.9	1.6	1.8	2.7	**4.2**	
jetSSV	0.2	0.5	0.9	0.4	0.7	0.6	
jetSSVP	0.6	0.6	0.8	0.6	0.9	0.8	
jetMuonIP	0.9	1.5	2.6	3.6	3.7	4.4	↗
jetMuonPt	1.4	2.2	3.3	4.0	4.1	4.8	↗
jetTCHE	4.7	2.9	3.0	3.0	3.0	3.4	
jetTCHP	5.3	3.5	3.3	3.8	3.7	4.1	

Table 5.15: Result of the data versus MC comparison. The single values represent the correlation coefficients of the listed variables to the target (data/MC). Larger values mean larger differences between data and simulation. In the last column some variables are marked with a small arrow. This illustrates that the differences between data and MC increase for the different jet momenta. The bold values points to unexpected large differences between data and simulation.

calibration	#t0 events	#t1 events	target 1 fraction	number of input vars	run time	expected signal fraction in data
track	415465	225460	47.8%	19	615.4s	0.036
vertex	164335	104582	52.6%	11	68.97s	0.26
electron	175666	102353	51.4%	16	144.69s	0.041
muon	110388	73216	49.8%	11	43.89s	0.11
jet	117432	86813	51.0%	10	180.3s	0.033
boost	150077	173899	51.3%	10	146.15s	

Table 5.16: Runtime of the different NeuroBayes calibrations. For the different trainings the number of events and the target 1 fraction is listed. Because of the weighing of the events this number does not correspond to the expected from the number of events. The last column shows how long it takes to get the calibration of the NeuroBayes expert.

At last we should discuss the existing b-jet taggers. As constructed, the simple secondary vertex b-jet taggers are well understood and very robust. The combined secondary vertex tagger is not yet calibrated and therefore shows discrepancies in the comparison. The muon b-jet tagger is dependent on the jet momentum. For larger jet momenta the differences between data and simulation rise. To indicate to this I added an arrow symbol into the last column of the table. The electron b-jet tagger only uses one variable of the electron properties. This variable looks also good. For the jet probability tagger a larger discrepancy in the high p_T region is found. The track counting taggers show differences but in an acceptable size.
The complete overview of the input variables and how they compare between data and MC can be seen in appendix A.

5.3.4 NeuroBayes data tagger (NBD)

In the last section we found a good agreement between data and simulations. This allows us to create another kind of b-jet tagger: a data based b-jet tagger. The idea is to be less dependent on the simulation of the background. The training is arranged in the same way as before, but for target 0 data samples are used. Target 1 is again a simulated sample of b-jets. The data includes correct background distributions, but also signal distributions. As shown in 5.2.1 for such a setup it is also possible to transform the NeuroBayes output distribution to the signal probability if the data is well described by MC.
We will see that a special preprocessing is needed to make the distributions comparable. I will start explaining the alternative b-jet tagger based on data samples (NBD) which has the same architecture as the classical MC based tagger (NBMC) described above. The NeuroBayes trainings are set up with the same settings as before. There are only two differences: at target 0 the data sample is used and for technical reasons the amount of statistics had changed. Table 5.16 shows the statistics used for the calibrations.
In the following I start with the studies on the track level calibration. There I will point out the special issues which appear for data based trainings. I will suggest a solution to reduce this problem and go on with an updated setup for the rest of the calibrations.

Track The calibration of the NeuroBayes expert results in the output distribution shown in figure 5.26 on the left.
The result looks quite promising. The shape looks as we would expect it with a prominent separation, which is caused by the large dependency on the track impact parameter. On the right side

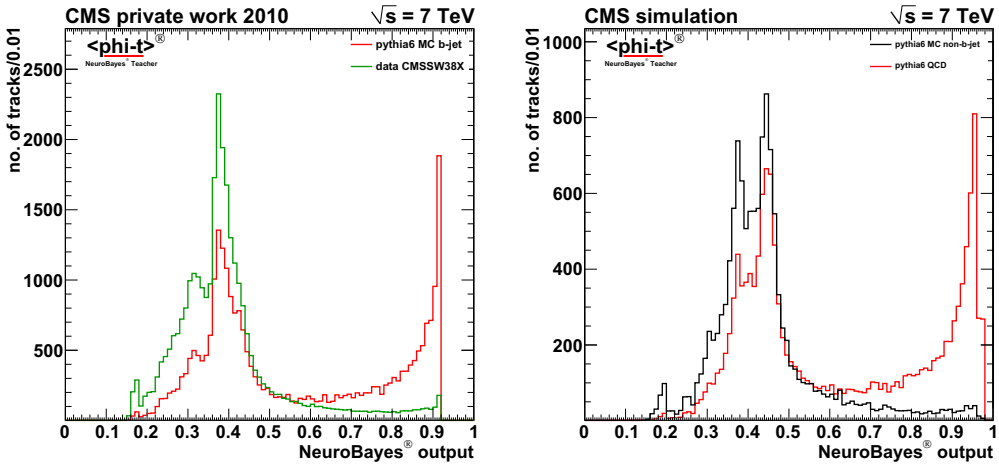


Figure 5.26: The output distributions of the data based training on track level on the left and for a classical MC based training on the right.

of the distribution we have tracks, which mainly come from B hadrons, the signal events on the left correspond to charged particles generated in the hadronization process. The main variation in shape appear because of the differing amount of statistics used for the calibration of the expert and the signal events present at target 0 for the NBD training. Let us have a more detailed discussion of the two distributions.

The first obvious difference between MC training $t1$ and data training $t2$ is the enlarged gap for large output values. The existence of signal events in target 0 leads to a shift of all output values. Using the probability transformation, calculated in section 5.2.1, we find for the two training scenarios the following dependency between the two output distributions:

$$o_d = \frac{o_{mc}\, f_{mc}/f_d}{1 - P(S) + o_{mc}(P(S)(f_{mc}+1) + f_{mc}/f_d - 1)}$$

o_d and o_{mc} are the output values of the two NeuroBayes experts. f_{mc} and f_d are the corresponding fractions of the training sample $f = \frac{N(T0)}{N(T1)}$. $P(S)$ is the unknown signal fraction of the data sample. The enlarged gap on the right can be directly extracted from the equation. Assuming a maximum value of $o_{mc} = 1$ for the MC based calibration the maximum of the data based training is limited to

$$\max(o_d) = \frac{1}{P(S)\, f_d + 1}.$$

But we can also go a step further and transform the complete distribution of the MC based training to an expected distribution for the data based training. For a fair comparison we applied the two experts on another MC sample (Pythia 6, QCD DiJet, CMSSW 36X), which is statistically independent from the samples used for their calibrations. The result is is shown in figure 5.27.
We see now a dominating structure produced by many tracks around 0.4. This is similar for the NBD distribution and the NBMC distribution. In shape there is a small difference between the two. The distribution consists of two peaking substructures.
The cause for this variation can be identified looking at the distributions of the input variables. The movement of the substructure is an effect of the changed distribution of the transverse track momentum relative to the jet axis $p_{T,rel}$. The result of the former comparison is shown in figure 5.28 on the left. We found the differences between data and MC already by the data/MC comparison introduced in section 5.3.3. In data we have more tracks at larger values than in MC. The jets in data are broader than simulated in MC.

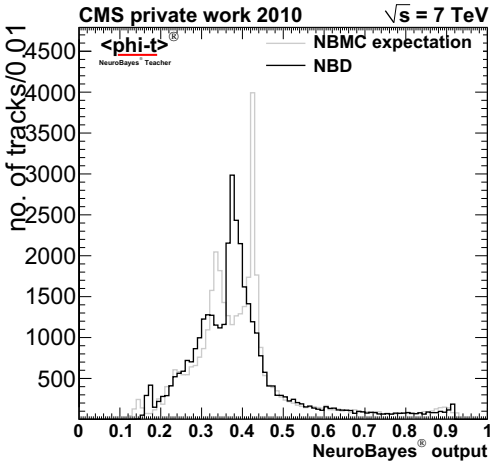


Figure 5.27: The output distribution of the data based training is plotted. Further a comparison of the distribution with the expectations from the NBMC training is shown. The NBMC output distribution (gray) was transformed by the function introduced in the text.

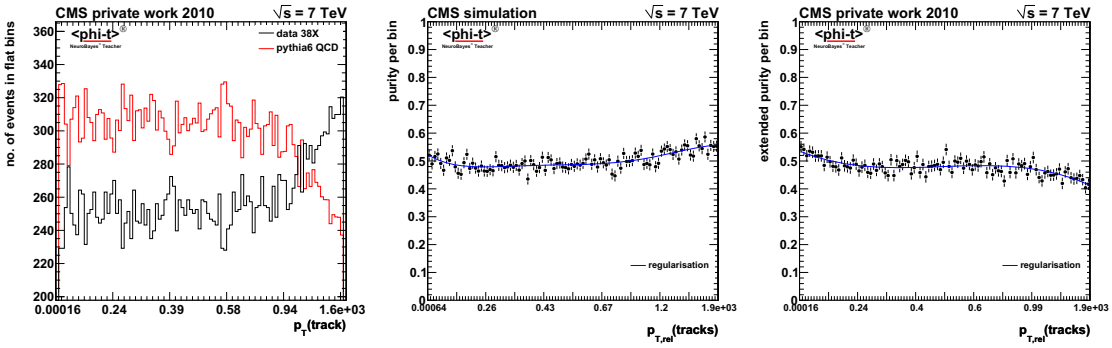


Figure 5.28: The left plot shows the differences between data and MC. In data more tracks with large $p_{T,rel}$ were found than simulated in MC. This effects also the classification of b-jets. The plot in the middle shows the purity of tracks coming from b-jets as calculated by MC, on the right the extended purity of the NBD training is shown.

This leads to a different purity estimate by the tagger calibrations at large $p_{T,rel}$ values. The plot in the middle shows the purity as extracted for the MC based training. The relative number of tracks from b-jets increases for large values. For the NBD training we can make a similar plot which does not correspond to the purity. In the following I will call it extended purity α. Extended means, that some signal events are present in the denominator, because of the construction of the tagger, where signal is trained against data.

$$\alpha = \frac{N_{MC}(S)}{N_d + N_{MC}(S)}$$

$N_{MC}(S)$ is the number of signal events from MC simulation and $N_d = N(B) + xN(S)$ is the number of events from the data sample. Compared to the real purity the shape is more uniform but should have the same variations at the same positions.
The right plot in figure 5.28 shows the α of the $p_{T,rel}$ variable. The shape is not more uniform than in the middle plot, but is different in the large $p_{T,rel}$ value region. The fraction of signal events is smaller. Knowing the differences between data and MC this is not surprising. We have less events in this region in MC.

The question is now: How do such effects affect a b-jet tagger which is based on data? We still achieve a good discrimination power. Also the probability interpretation is correct if we ask for objects like they are simulated in MC. It is true that the probability to find these tracks from a b-jet with large $p_{T,rel}$ is in reality smaller than expected from MC. But this does not mean that there are less tracks from b-jets in reality - only less tracks which are simulated.
Having a just commissioned detector it is not expected to have everything in perfect state. Looking at physics which is well known from other experiments we do not expect any new observation. Almost all discrepancies between data and MC can point to problems in the simulation, wrong assumptions for the resolution of detector components or efficiency calculations. Therefore we are able to do a simple correction. Under the assumption that the fraction of our signal is correctly simulated, and only the absolute numbers are wrongly simulated, we can do a reweighing of the MC samples.
The weight factor can be extracted looking at the purity $P(MC|x_i)$ of the data/MC comparison for the variables where we expect a correct simulated signal fraction. If we want to correct for more variables we have to account for the correlations between them. Doing a NeuroBayes classification gives us an estimate of the overall purity $P(MC|o_t)$. The application of the weights is similar to a boost training where the weights for the data events are $w_D = 1$. This is the focusing function $F_f = \frac{1}{P(MC|o_t)}$. The weights for the MC events are:

$$w_{MC} = \frac{1 - P(MC|o_t)}{P(MC|o_t)}.$$

A correct handling of this weights needs a detailed study of the input variables and the data. This is a ambitious goal but not really achievable. A big step could be made with the application of weights extracted from a overall data/MC comparison. The weights can be calculated with a NeuroBayes expertise. The sample will be corrected on the level of the comparison. For all inclusive subsamples still differences can occur.

Finally we apply the weights calculated from an overall data/MC comparison for the construction of the NBD b-jet tagger. Figure 5.29 shows how the distributions compare after this correction.
The two distributions become more similar. This means that it is worth to correct the MC samples. On the other hand we still found some differences. There is a clear structure on the left, where more tracks appear than expected from MC. It seems that we found another discrepancy between

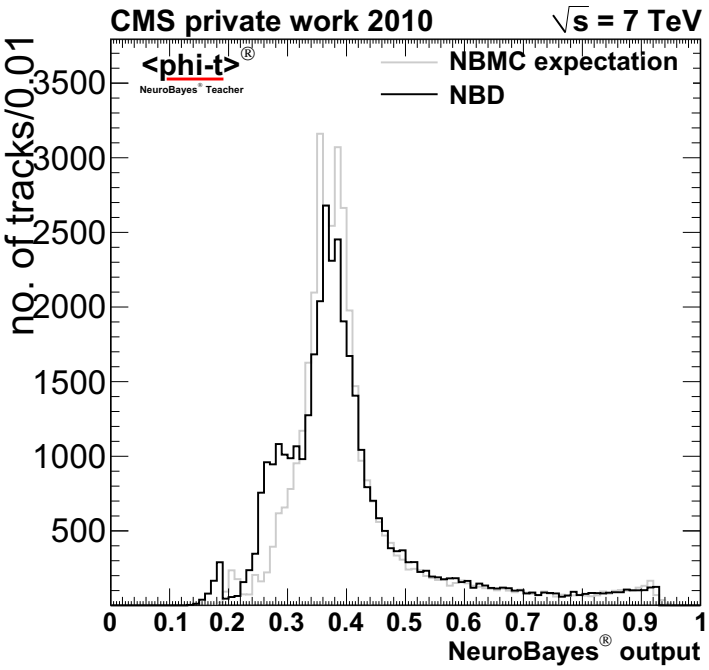


Figure 5.29: The output distribution of the data based training is plotted. For the calibration a correction of the MC distributions was applied. Further a comparison of the distribution with the expectations from the NBMC training is shown. The NBMC output distribution (gray) was transformed by the function introduced in the text.

data and MC. This points to differences between data and MC of the inclusive distributions of b-jets or non-b-jets.
This implies two things for the data based b-jet tagger. If these discrepancies are caused by insufficient simulation of the background processes we have a strong argument to do this correction. We will profit from the situation that for the NBD we are independent from the background simulations. This leads to an improvement for the b-jet tagger.
On the other hand it is also possible that the b-jet simulations are inadequate. Then the data based training leads to a misinterpretation of the sample. The probability interpretation is only valid for MC b-jets. Further real but not simulated b-jets are treated as background. To solve this problem a more general model of the signal distributions has to be developed.

In the following I decided to add another preprocessing step, where I calculate the weights to apply them for the correction. The very good knowledge on b hadrons and the excellent studies of b-jets at other experiments makes me believe in good simulations of this inclusive class.
The additional preprocessing must be included for all data based b-jet tagging classifications.

Vertex Figure 5.30 shows the output distributions of the secondary vertex classification for the NBD training and the expectation from the NBMC. The distributions looks similar to what we expect. We see a large gap on the right, caused by the signal events present in the target 0 (black) distribution. The shift is larger than for the tracks. This is expected because a reconstructed secondary vertices is already a good indication of a b-jet. Therefore in the target 0 sample a large fraction of secondary vertices from b-jets is present. From MC simulations we expect 26% of b-jets.

Further I will point to the number of tracks associated to the secondary vertex. The distributions

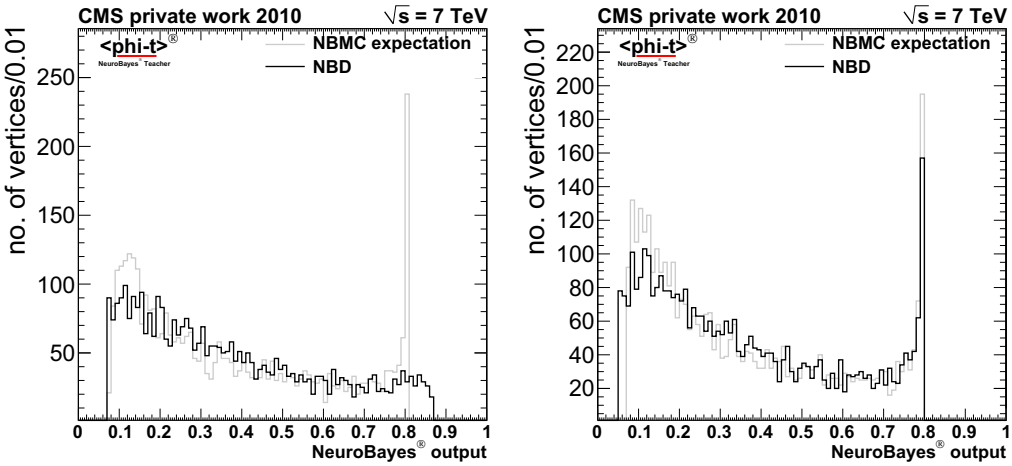


Figure 5.30: A comparison of the NeuroBayes output distributions of the NBD vertex training with the expectations from the NBMC training is shown. On the right a correction is applied on the MC sample.

of this illustrates the behavior in the NBD training very well. Figure 5.31 shows the purity of the variable as used for the NBMC training on the left and the same for the extended purity of the NBD training. For all bins we see a shift to the center. Especially for the 100% purity bins on the right the shift caused by the purity extension is nicely visible.
The values of these bins can be taken to make a rough estimate of the signal fraction. The extended purity α is around 0.8. We assume that the bins contain only signal: $\alpha = 1/(P(S)+1)$. This leads to a signal fraction around 25%, which agrees with the expectations from MC.

Leptons Figure 5.32 shows the improvements we get after applying the corrections for the calibrations on lepton level. It is interesting, that for the electron candidates the weighing is not needed to get a result similar to the expectations.

Jet We have all NeuroBayes calibrations of the jet objects in a good state and it is possible to take them for the final jet classification. Again we do the corrections. The weights were determined from the data/MC comparison. The result of the calibration can be seen in figure 5.33. The left plot shows the output distribution on the trained sample, while the right shows it on a independent sample. The method seems to work very well.
To finalize the data based b-jet tagger in a last step the boost training was performed. The improvements are similar to the the MC based training. The performance is plotted in figure 5.34.

In summary we created two different b-jet taggers. The first one is based on MC samples while the second uses data instead of background simulations. Both b-jet taggers are competitive to the existing ones or even better. More important are the comparisons between data and MC. With these studies we have a complete understanding of the tagger, which leads to a good belief in a accurate functionality for further analysis.
In the following I will present a first use case of the new b-jet tagger in an inclusive b-jet cross section measurement. But also further applications are imaginable, especially for the data based b-jet tagger. Above all in most of the analyses where signal has to be separated from a large amount of QCD background this ansatz is an interesting alternative to the usual approaches. In many cases the background distribution consists of many not perfectly known subprocesses. With

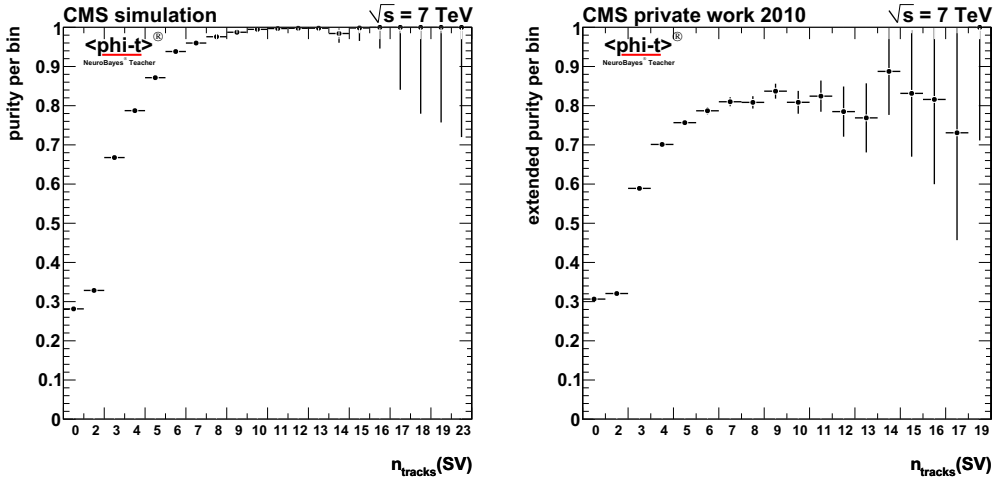


Figure 5.31: The left plot shows the purity of the number of tracks associated to the secondary vertex for the NBMC training. On the right the same is shown for the extended purity of the NBD training. Especially for the 100% purity bins on the right the shift caused by the purity extension is clearly visible.

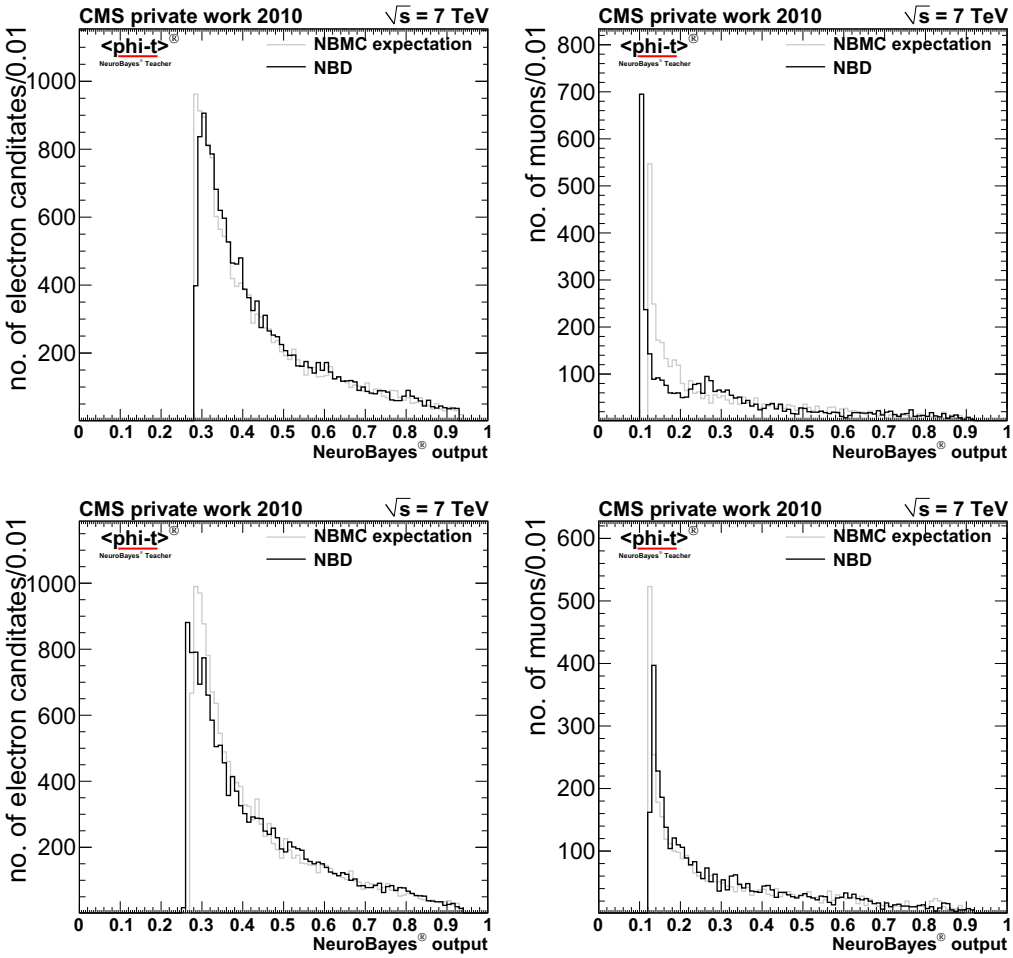


Figure 5.32: The output distributions of the data based training are plotted. The upper plots show the distributions when no correction is applied on the left for electron candidates and on the right for muons. The lower plots show the same with corrections.

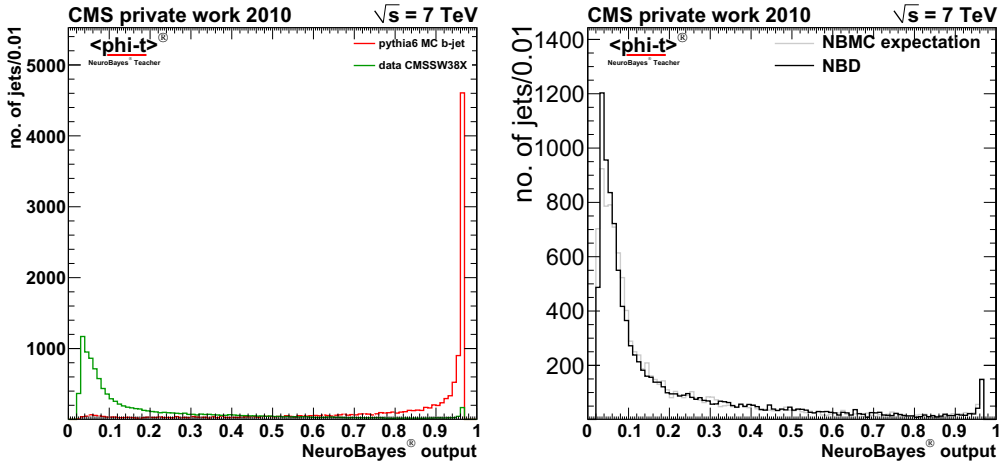


Figure 5.33: Left: the output distribution of the NBD training. This expertise applied on an independent sample results in the distribution plotted on the right. It is similar to the one expected from MC.

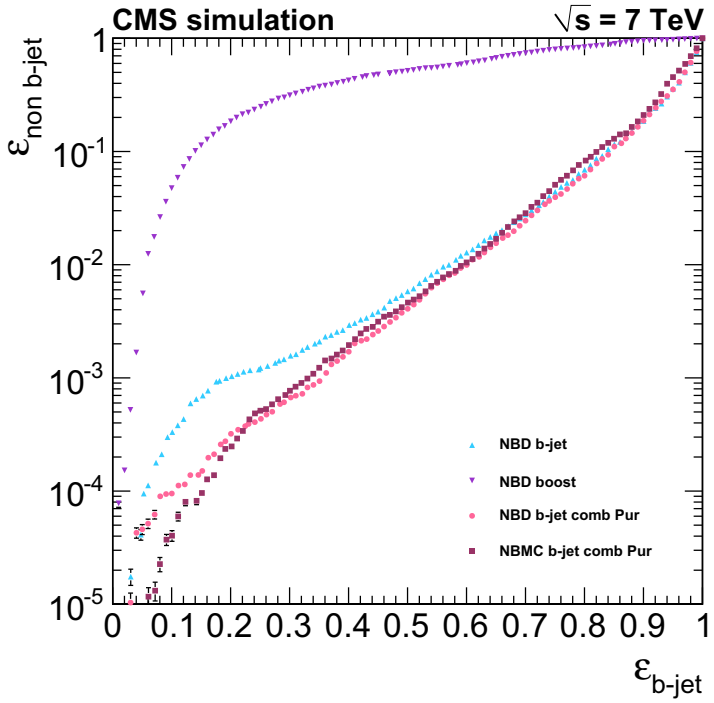


Figure 5.34: Performance of the NeuroBayes data based b-jet tagger. Instead of simulated background events data is used. The performance of this new kind of b-jet tagger is comparable to a MC based.

the data based approach one is able to bypass a detailed study on background and focus on the signal study.

Chapter 6

b jet cross section measurement

In this chapter I will present the analysis of the inclusive b-jet cross section measurement. As introduced in section 2.2 a measurement of this quantity is of large interest. It is important for searches of particles which decay into b-quarks. In the Standard Model we know three particles with this property. The t-quark, which decays to a b-quark with almost 100% probability, the Z boson, which is able to decay to a $b\bar{b}$ pair, and the W boson, where such a decay is strongly suppressed. In addition, as a fourth particle of the Standard Model, the expected Higgs boson is able to decay into $b\bar{b}$ pairs. For these and many of the new particles from models beyond the Standard Model the b-quark is an important indicator. Thus b-quark processes are also a large background source. With the results of an b-jet cross section measurement it is possible to scale the background for such analysis in a more reasonable way.
But also the analysis of the b-jet quantity itself is very interesting. Around 20 years ago, the same measurement was on the way to cause a sensation. The detectors at the hadron collider SPPS and Tevatron found a difference between theory and experiment (see section 2.2.3). New physics models were discussed, but in the end a recalculation of the next to leading order predictions solved this disparity and were approved by the Tevatron Run 2 experiments.
Among other things the old miscalculations were caused by an inadequate knowledge of the fragmentation functions and the parton distribution functions. These are still not understood completely today. A measurement of the b-jet cross section will give us a verification of the established model. Above all it is possible to test the QCD calculations in transverse momentum space, which is achieved by the collisions at LHC. Another discrepancy between experiment and theory will bring us closer to the discovery of physics beyond the Standard Model.
This chapter includes different approaches for such a measurement. In the first part I will describe the method as published in [CMS10e]. The last two sections will deal with updates done on the extended data and a alternative approach for a b-jet cross section measurement with the use of NeuroBayes.

6.1 Recent b cross section measurement at CMS

In the first part of this section I review the recent status of the CMS b-jet cross section measurement at an integrated luminosity of 60 nb^{-1} [CMS10e]. I performed this analysis together with colleagues from the CMS collaboration. Here the main parts are taken as published. Some information was added to bring it into the context of this thesis.
The review starts with the specification of the collected data at that time. The procedure of b-jet tagging is presented for a very small amount of data. It follows the measurement of the b-jet purity

and the b-jet efficiency. To get the final result, detector effects are unfolded. The uncertainties will be discussed.

6.1.1 Event and jet selections

The inclusive jet data was collected using a combination of Minimum Bias and single jet triggers (see section 4.1), which are consecutively used in the lowest p_T range where the triggers are fully efficient. Dependent on the samll amount of data events the quality selection was applied similar to that presented in section 4.5.

The p_T spectra from individual triggers are normalized using luminosity estimates [CMS10h] and then combined into a continuous jet p_T spectrum. The integrated luminosity corresponds to 60 nb^{-1}. Only one trigger is used per each p_T bin, to simplify the analysis. The raw p_T spectra are unfolded using the ansatz method [BBK71; FFF78], with the jet p_T resolution obtained from MC. The uncertainty of the jet p_T resolution is estimated using a comparison of dijet p_T balance between data and MC [CMS10f].

6.1.2 b-tagging

The b-jets are tagged using a secondary vertex high-purity tagger (SSVHP [CMS10a]). The secondary vertex is fitted with at least three charged particle tracks. A selection on the reconstructed 3D decay length significance is applied, corresponding to about 0.1% efficiency to tag light flavor jets and 60% efficiency to tag b-jets at $p_T = 100$ GeV.

The b-tagging efficiency and the mistag rates from c-jet and light jet flavors are taken from the MC simulation and constrained by a data/MC scale factor determined from data. This b-tag efficiency measurement relies on semileptonic decays of b-hadrons, the kinematics of which allow for discrimination between b and non-b-jets. Fits to the distribution of the relative transverse momentum of the muon with respect to the jet direction enable the extraction of the flavor composition of the data, and ultimately the efficiency for tagging b-jets. The mistag rate from light flavor jets is constrained separately by a study using a negative-tag discriminator [CMS10a].

The production cross section for b-jets is calculated as a double differential,

$$\frac{d^2\sigma_{\text{b-jets}}}{dp_T dy} = \frac{N_{\text{tagged}} f_b C_{\text{smear}}}{\epsilon_{\text{jet}} \epsilon_b \Delta p_T \Delta y \mathcal{L}} \quad ,$$

where N_{tagged} is the measured number of tagged jets per bin, Δp_T and Δy are the bin widths in p_T and y, f_b is the fraction of tagged jets containing a b-hadron, ϵ_b is the efficiency of tagging b-jets, ϵ_{jet} is the jet reconstruction efficiency and C_{smear} is the unfolding correction. ϵ_{jet}, ϵ_b and f_b are all calculated from MC in bins of reconstructed p_T and y, for consistency with the data-based methods. The correction factor C_{smear} unfolds the measured p_T back to particle level using the ansatz method, used also for the inclusive jet cross section measurement and described in [CMS10h].

b-tagging efficiency

The b-tagging efficiency with the selections used in this analysis is between 6% and 60% at $p_T >$ 18 GeV and $|y| < 2.0$. The efficiency rises at higher p_T as the b-hadron proper-time increases. The efficiencies estimated from MC are shown in Fig. 6.1. To smoothen out statistical fluctuations, the b-tagging efficiency in each rapidity bin is fitted versus p_T, and the fit result is used in the analysis.

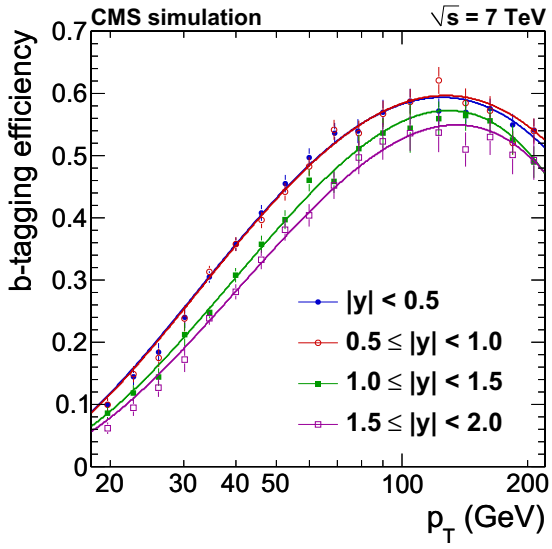


Figure 6.1: b-tagging efficiency in different rapidity bins.

b-tagged sample purity

The b-tagged sample purity is estimated using two complementary approaches. In the first method, the invariant mass of the tracks associated to the secondary vertex, denoted secondary vertex mass, is computed after the SSVHP selection. A fit to the secondary vertex mass distribution is performed, taking the shapes for light, c and b-jets from simulation and letting free the relative normalisations for c and b-jets, while fixing the small contribution from light jets to the MC expectation ("template fit"). This fit allows for a robust estimate of the b-tagged sample purity and constrains the mistag rate uncertainty from c jets. An example of the template fits is shown in Fig. 6.2.

In the second method the b-tagging efficiency ϵ_b as well as the mistag rates for light flavor ϵ_l and charm ϵ_c are estimated from MC. These are shown in Fig. 6.3. Multiplied by the expected relative fractions of b-jets F_b, c jets F_c and light flavor jets F_l, also shown in Fig. 6.3 in the inclusive jet sample (without b-tagging), the tag rates can be used to calculate the expected purity as

$$f_b = \frac{F_b\epsilon_b}{F_b\epsilon_b + F_c\epsilon_c + F_l\epsilon_l} \quad .$$

The b-tagging efficiencies of c and light jets in Fig. 6.3(left) are multiplied by their relative frequency to b-jets to illustrate the rough relative contributions of $F_b\epsilon_b$, $F_c\epsilon_c$ and $F_l\epsilon_l$ to the b-tagged sample at $p_T \approx 100$ GeV. The resulting estimates of b-tagged sample purity from data and from MC are shown in Fig. 6.4. The data and MC are found the be in good agreement, with an overall relative data/MC scale factor measured to be 0.976 ± 0.022 (0.996 ± 0.030) for b-jets in the p_T range 18–220 GeV (18–84 GeV) and rapidity $|y| < 2.0$.

Given the good agreement between data and MC, the central values for purity are taken from MC to properly take into account the p_T and y dependence.

b-tagging uncertainty estimates

The leading uncertainties for the inclusive b-jet production are those coming from jet energy scale, luminosity, b-tag efficiency, and mistag rates. The 11% luminosity uncertainty [CMS10g] cancels completely in the ratio to the inclusive jet p_T spectrum, and the JEC uncertainty produces only a

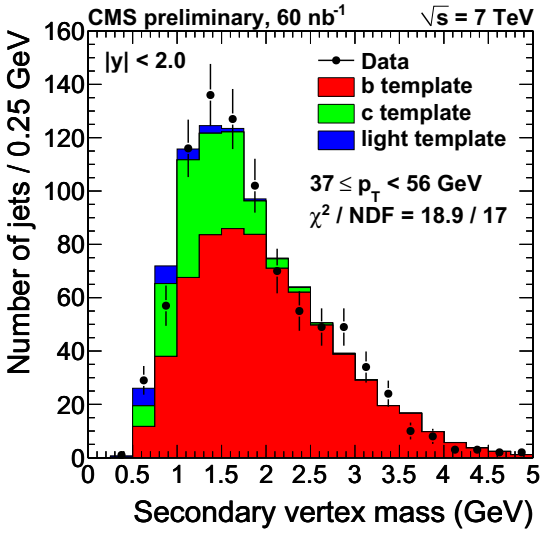


Figure 6.2: Example of secondary vertex mass fits.

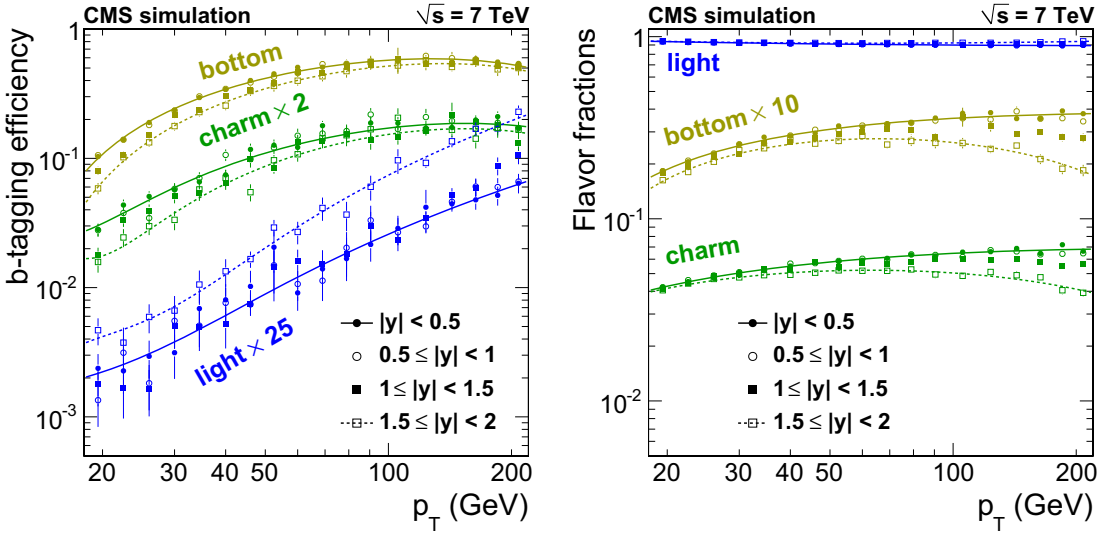


Figure 6.3: The b-tagging efficiency and light, charm mistag rates from MC truth (left). Bottom, charm and light fractions of inclusive jets from MC truth (right).

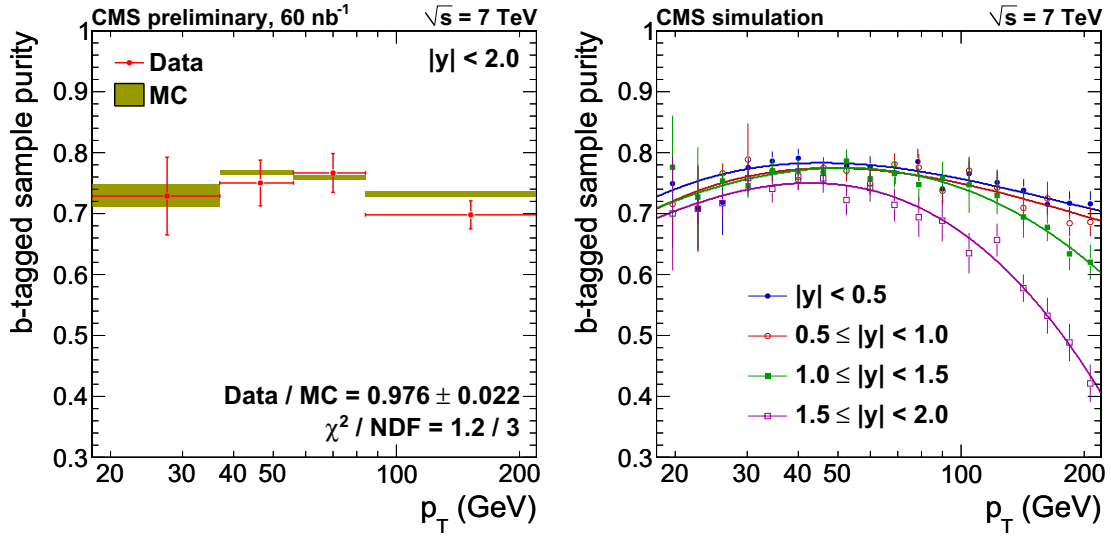


Figure 6.4: The b-tagged sample purity obtained using fits to secondary vertex mass (left). The b-tagged sample purity estimated using b-tagging efficiency and mistag rates from MC (right).

small residual uncertainty due to differences in p_T spectra and jet fragmentation between inclusive jets and b-jets.

The leading remaining uncertainties for the ratio between b-jet and inclusive jet production are the b-tagging efficiency and the charm mistag rate, both of which are currently in essence statistical uncertainties from the data-based methods to constrain the b-tagging efficiency and the b-tagged sample purity, and the b-jet specific JEC. The light quark mistag rate has a significant contribution to the total uncertainty at high p_T and forward rapidities, but is otherwise negligible due to the low mistag rate. The inclusive jet energy scale, on the other hand, only contributes at $p_T < 30$ GeV, where the b-jet spectrum flattens while the inclusive jet spectrum is still exponentially falling.

The b-tagging efficiency measurement relies on semimuonic decays of b-hadrons. The limiting factors for this measurement are the limited number of SSVHP tagged jets containing a muon, the uncertainty in the c- and light template shapes and the systematic uncertainty in generalizing the efficiency measured on semileptonically decaying b-jets to all b-jets. The obtained scale factor is $0.98 \pm 0.08(\text{stat}) \pm 0.18(\text{syst})$ for jets with $p_T > 20$ GeV and $|y| < 2.4$ [CMS10a].

The uncertainty on b-tagging efficiency arising from poorly known relative contributions of flavor creation (FCR), flavor excitation (FEX) and gluon splitting (GS) has also been studied in detail. The relative angle ΔR between the b-hadrons is strongly dependent on the production mechanism. The b-hadrons produced by GS, in particular, tend to be close to each other in ΔR, which leads to a reduced efficiency of the SSVHP tagger. This uncertainty is estimated by varying the relative contributions in MC within ±50%, constrained by studies of the ratio between secondary vertex energy and b-jet energy, which is sensitive to the contributions of FCR+FEX (large ratio) compared to GS (small ratio). The b-tagging efficiency as a function of the ΔR distance between the b-jets is shown in Fig. 6.5(left). The variation versus ΔR is observed to be up to 25%, but combined with the maximal variations of the GS and FCR+FEX by ±50% shown in Fig. 6.5(right) this uncertainty is found to be less than 2%.

The b-tagging efficiency uncertainty is dominated by the statistical uncertainty in the data-driven method. The uncertainty is conservatively taken as the statistical uncertainty of 8% in quadrature with the 18% systematic uncertainty and the 2% from the data/MC scale factor of 0.98 that is not applied in this analysis, giving 20% as the total systematic uncertainty for the b-tagging efficiency.

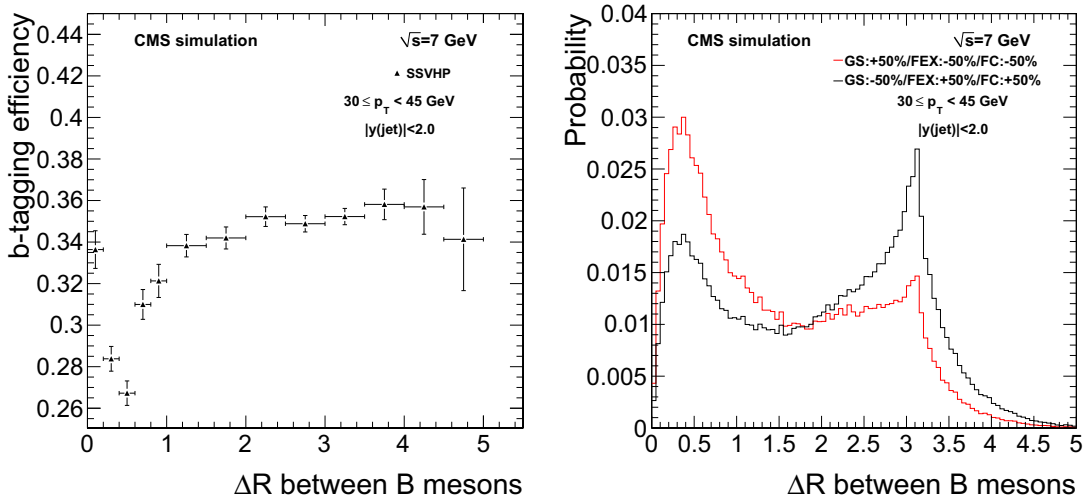


Figure 6.5: The b-tagging efficiency variation versus ΔR between b-hadrons (left). Distribution of ΔR between b-hadrons for $\pm 50\%$ variations of GS and FC+FEX (right).

It should be noted, however, that the robustness of the decay length observable can degrade at $p_T > 200$ GeV, which should be taken into account in future updates of the analysis that start to probe this kinematic region.

An additional 10% uncertainty at $p_T > 200$ GeV is taken into account for this, with the extra uncertainty log-linearly reduced to 0% at $p_T = 100$ GeV.

The light quark mistag rate calculated by MC simulation has been validated on data by studies using a negative-tag discriminator to within a systematic uncertainty of about 50% [CMS10a]. This uncertainty has been directly propagated to the light quark mistag rate used in the present analysis. This uncertainty is only a few percent across most of the kinematic range, but grows up to 15% at high p_T in the most forward rapidity bins.

The charm mistag rate is constrained by the secondary vertex mass template fits, whose results are shown in Fig. 6.4(left), with a data/MC scale factor of 0.976 ± 0.022. The template fit uncertainty is conservatively taken as the statistical uncertainty of 2.2% added in quadrature with the 2.4% from the data/MC scale factor of 0.976 that is not applied in this analysis, giving 3.3% as the total systematic uncertainty for b-tagged sample purity. The systematic uncertainty for the template fits due to fixing the light quark mistag rate to the MC prediction has been tested by varying the light quark mistag rate by $\pm 50\%$ and was found to be negligible compared to the statistical uncertainty. These studies constrain the charm mistag rate uncertainty to 20% or better, which is then propagated into an uncertainty in the analysis. The resulting uncertainty is around 3–4% and flat in p_T and y.

The difference of inclusive jet and b-JEC was studied by using the MC truth after applying the standard inclusive JEC. The residual difference in MC is less than 1% at $p_T > 30$ GeV where the b-JEC uncertainty contributes most, and the difference in data could be expected to be of the same magnitude. Due to the steeply falling p_T spectrum, a 1% b-JEC uncertainty leads to about 5% uncertainty on the ratio of b-jet and inclusive jet cross section. Here it is interesting to note that direct measurements done at CDF using $Z \to b\bar{b}$ observed a relative b-jet scale of 0.971 ± 0.011 [D$^+$08]. The significantly smaller relative b-jet correction expected at CMS can be attributed to the Particle Flow reconstruction, which natively includes muons from semileptonic decays and is more robust against differences in jet fragmentation than the calorimetric jets used in the CDF

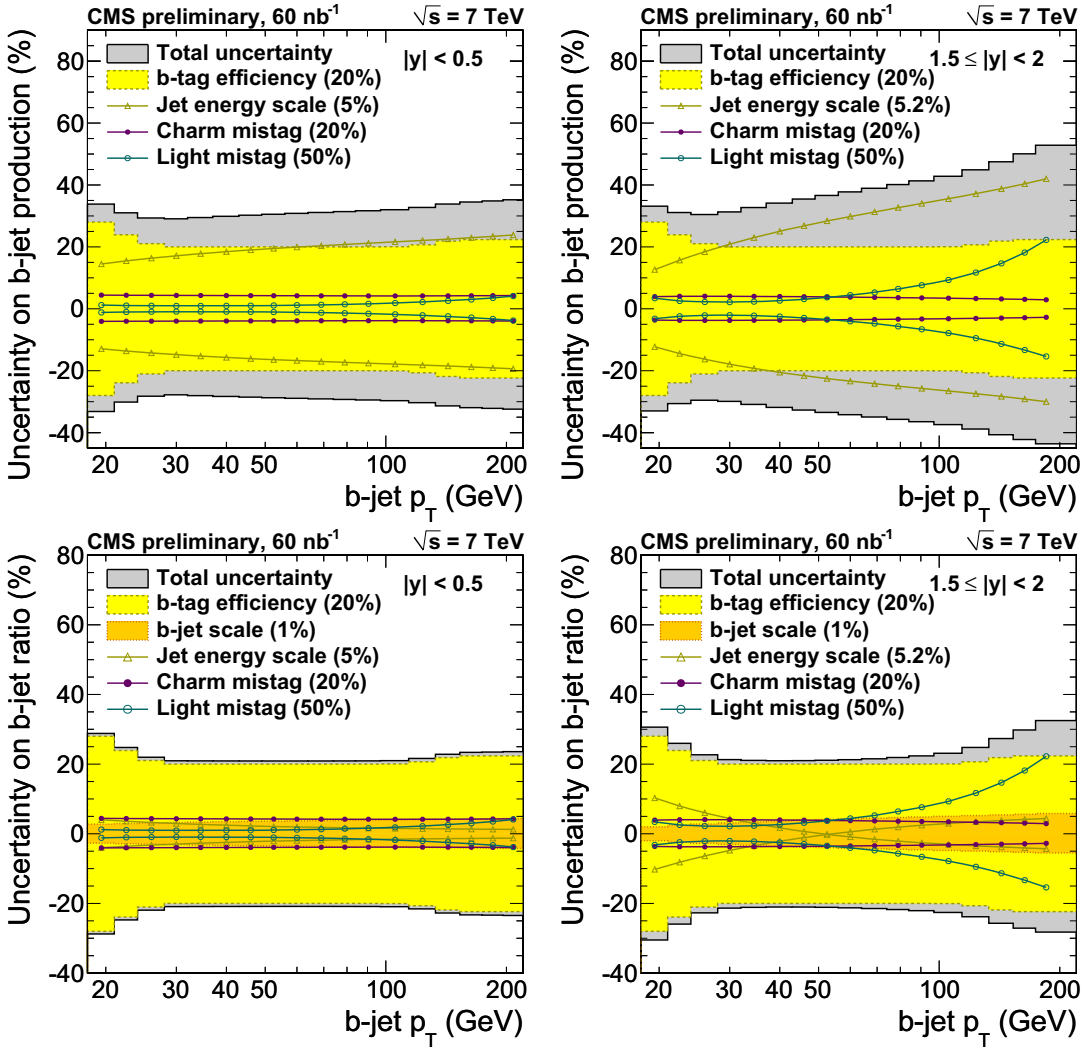


Figure 6.6: Leading sources of systematics uncertainty for the b-jet cross section measurement at $|y| < 0.5$ (top left) and at $1.5 \leq |y| < 2.0$ (top right), and for the ratio of b-jet and inclusive jet cross section measurements at $|y| < 0.5$ (bottom left), and $1.5 \leq |y| < 2.0$ (bottom right). The 11% luminosity uncertainty is not shown.

measurement.

Figure 6.6 shows a summary of the leading sources of uncertainty for the b-jet cross section and for the ratio of b-jet and inclusive jet cross sections. The contribution from luminosity uncertainty is completely canceled out in the ratio, and the contributions from JEC and JER [CMS10f] are largely reduced at $p_T > 20$ GeV. The remaining leading systematics for the ratio are b-tagging efficiency, relative b-jet scale and charm mistag rate, all contributing with similar weight and leading to a flat total uncertainty of about 20% at $p_T > 20$ GeV.

The reconstructed MC has been processed through the same analysis chain as the data, and the results have been compared to the MC truth results. This closure test found overall agreement to better than 1% (10%) at $p_T > 30$ GeV ($p_T > 15$ GeV) and $|y| < 2.0$. The worse closure test at low p_T can be explained by the large size (more than a factor of ten at $p_T < 20$ GeV) of the b-tagging correction at low p_T, combined with relatively poor MC statistics (10% uncertainty at 10 GeV).

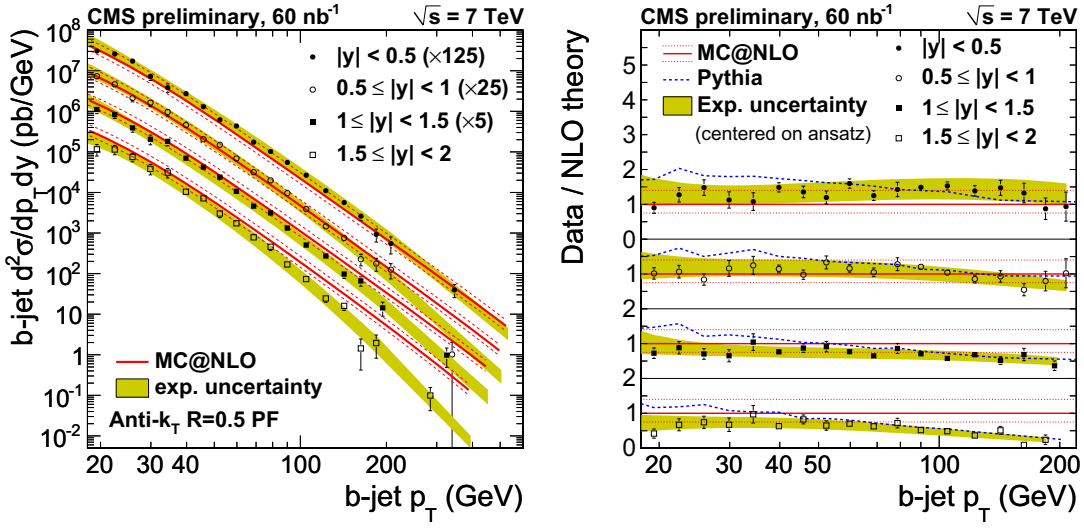


Figure 6.7: Measured b-jet cross section compared to the MC@NLO calculation, overlaid (left) and as a ratio (right). The Pythia prediction is also shown, for comparison.

6.1.3 Measurement

The measured b-jet cross section is shown as a stand-alone measurement in Fig. 6.7 and as a ratio to the inclusive jet p_T spectrum in Fig. 6.8. The inclusive jet NLO theory prediction is calculated with NLOJet++ [Nag02] using CTEQ6.6M PDF sets [P+02] and fastNLO [KRW06] implementation. The factorization and renormalization scales were set to $\mu_F = \mu_R = p_T$. The inclusive b-jet prediction is calculated with MC@NLO [FW02; FNW03] using the CTEQ6M PDF set and the nominal b-quark mass of 4.75 GeV, giving a total b cross section of 238 μb. The parton shower is modeled using Herwig 6.510 [M+92]. The results are compared to a NLO theory prediction (MC@NLO) and to the Pythia MC (tune D6T [Fan07]), and are found to be in good agreement with Pythia and in reasonable agreement with MC@NLO. The NLO calculation is found to describe the overall fraction of b-jets at $p_T > 18$ GeV and $|y| < 2.0$ well, but with significant shape differences in p_T and y.

Fitting the measured ratio of data to Pythia in the phase space window $30 < p_T < 150$ GeV and $|y| < 2.0$ to a constant, we obtain a global scale factor of $0.99 \pm 0.02(\text{stat}) \pm 0.21(\text{syst})$, where the systematic uncertainty is a weighted average over all the bins contributing to the fit. The fit has $\chi^2/NDF = 43.4/47$. Repeating the same fit for the ratio between reconstructed MC and generator-level MC results in a scale factor of 1.009 ± 0.005 with $\chi^2/NDF = 246/46$, confirming good closure of the analysis chain. Finally, the NLO/MC global scale factor is 1.04 ± 0.05.

The total b cross section of 238 μb from the MC@NLO calculation has a sizable uncertainty from the choice of renormalization scale between $\mu_R = 0.5$ and $\mu_R = 2$ (+40%, −25%), from CTEQ PDF variations (+10%, −6%), and from the choice of b-quark mass between 4.5 GeV and 5.0 GeV (+17%, −14%). The dominant scale uncertainty is overlaid as an uncertainty band around the MC@NLO prediction in Figs. 6.7(b) and 6.8.

The application of the former unfolding method is difficult to reuse. This is because of the strongly falling p_T spectrum, which is covered in the update measurement. The p_T distribution ranges many orders of magnitude, where the unfolding method leads to a bias to higher values. There are ongoing studies to solve these problems. The b-tagging measurements can be updated to the 36/pb.

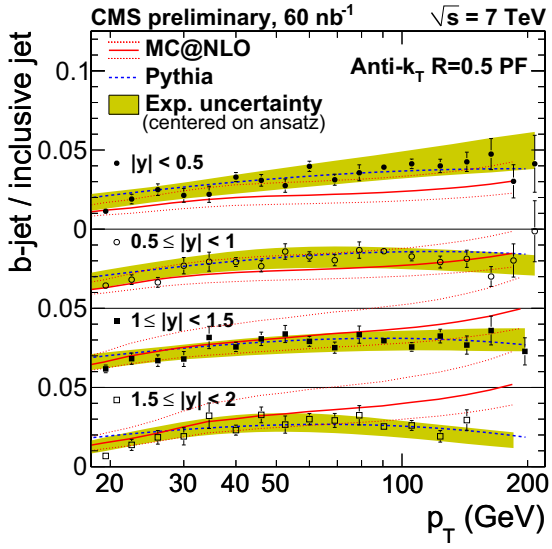


Figure 6.8: Measured b-jet cross section as a ratio to inclusive jet cross section. The NLO theory and Pythia MC predictions are shown for comparison.

6.2 Update of the flavor content fitter

In this section I will present the results of the b-jet cross section measurement using a flavor content fit to extract the fraction of b-jets in a so called tagged sample. Tagged means, only jets with a large probability to be a b-jet are selected. This is obtained by a cut on the significance, how likely it is, that the secondary vertex has a lifetime. This is an important property of b-jets, because they contain a decay vertex of long-lived B hadrons. To get the cross section right, further an estimate of the b-jet efficiency for the tagged sample is needed.
In the following I will introduce in the method of template fitting, which is used for the flavor content fit. Thereafter I will specify the area, where these fits are applied and present the results dependent on different intervals in transversal momentum and rapidity of the jet.

6.2.1 Template fit

Just for completeness I describe here the method of a binned log likelihood fit. Given a histogram with a known number of bins n_{bins} filled with the values from the variable of interest, the statistics d_i in each bin i follow a Poisson distribution. The goal of the binned log likelihood fit is to vary the parameters p of a given model F_i to the most probable values. This is done by a minimization of the extended log-likelihood function with TMinuit. The statistical uncertainties are calculated by Minos. The binned log likelihood function is shown in the following equation:

$$-2\log\mathcal{L} = \sum_{i=0}^{n_{bins}} d_i \cdot \log(F_{p,i}) - F_{p,i}$$

There are different ways for the parametrization p. The parametrization must be chosen dependent on the information one is interested in. The first parametrization ($p = 0$) estimates the number of events for each template. For n_t templates we have the same number of free parameters $N_0 ... N_{n_t}$.

$$F_{0,i} = F_i(N_0 ... N_{n_t}) = \sum_{k=0}^{n_t} N_k t_{k,i}$$

Each parameter is an estimate of how many events of a template class k are in a given data sample. The fit extracts the number of events and its statistical uncertainties.
The second parametrization ($p = 1$) tries to extract the fractions f_k of events to the total number of events in the given data sample N_{tot}. Therefore N_{tot} is one of the free parameters. A parametrization with all fractions f_k is not possible. One fraction can be replaced by the others: $f_k = 1 - \sum_{j \neq k} f_j$, because the sum of all fractions has to be 1.
Another difficulty is that the fractions have natural limits. They have to be between 0 and 1. The goal is to have a parametrization which can take this into account. So it is only possible to estimate one of the fractions $r_0 \in f_k$ at one time.

$$F_{1,i} = F_i(N_{tot}, r_0 ... r_{n_t-1}) = N_{tot} \left(\sum_{k=0}^{n_t-1} r_k t_{k,i} \prod_{j=0}^{k-1} (1 - r_j) + t_{n_t,i} \prod_{j=0}^{n_t-1} (1 - r_j) \right)$$

In principle it is possible to estimate the values of one parametrization out of the parameters from the other. Due to the asymmetric uncertainties of the values the error propagation is difficult and an additional fit is more reasonable.

6.2.2 $p_T/|y|$ binning

Choosing the p_T and y binning for a differential jet-cross-section measurement has to take two conflictive aspects into account. On the one hand, one wants to make a very fine binning in order to have enough well defined points to fit the assumed function to. On the other hand one has to have enough statistics in each bin for doing reasonable template fits which yield the fraction of b jets with a good enough precision.
Also it is not necessary to run over all MC samples for creating the templates. We only used the ones which influence the statistics by more then 0.1%. Also we try to avoid isolated events with very large weight. The different bins in p_T and the MC sample selection for each bin is listed in table 6.1. Further there is an additional binning into the barrel region $|y| < 1.5$ and the forward region $1.5 \leq |y| < 2.5$ of the detector. Studies of a finer binning in rapidity need a merging of the p_T bins. For the update of the CMS physical analysis summary [CMS10e] such a study was done, but is not presented in this thesis.

6.2.3 Fit results

The flavor content fit is performed to measure the fraction of b-jets in a tagged sample. An enriched b-jet sample was used. Therefore the jet has to pass the tight woking point of the simple secondary vertex purity b-jet tagger $SSVP > 2$. This represents a cut on the significance of the secondary vertex flight distance [Sch08]. It is required that the vertex is reconstructed with at least three tracks. The selection results in a pure sample of b-jets with a small fraction of light jets and c-jets. This selection is motivated to reduce the systematic uncertainties of the flavor content fit using the secondary vertex mass m_{SV}. Therefore we have to require a well understood secondary vertex.
Figure 6.9 shows the distribution of the simple secondary vertex purity b-jet tagger and the vertex mass as published for the CMS commissioning in [CMS10a]. There is a good agreement between data and MC
The flavor content fits are performed in the different bins of the transversal momentum p_T. Three templates for b-jets, c-jets and light-jets were created for each p_T/y region for the final result. As a cross check the b-jet fraction was also estimated with two templates (b-jet and non-b-jet). The results were the same in the statistical context. Table 6.2 show the results of the flavor content

p_T range [GeV]	MC samples													
	QCD0to5	QCD5to15	QCD15to30	QCD30to50	QCD50to80	QCD80to120	QCD120to170	QCD170to300	QCD300to470	QCD470to600	QCD600to800	QCD800to1000	QCD1000to1400	QCD1400to1800	QCD1800
$37 \leq p_T < 43$	x	x	x	x	x	x									HLT_Jet15U
$43 \leq p_T < 49$	x	x	x	x	x	x									
$49 \leq p_T < 56$	x	x	x	x	x	x									
$56 \leq p_T < 64$			x	x	x	x	x	x							
$64 \leq p_T < 74$			x	x	x	x	x	x							
$74 \leq p_T < 84$			x	x	x	x	x	x							
$84 \leq p_T < 97$			x	x	x	x	x	x							HLT_Jet30U
$97 \leq p_T < 114$			x	x	x	x	x	x							
$114 \leq p_T < 133$				x	x	x	x	x	x						HLT_Jet50U
$133 \leq p_T < 153$				x	x	x	x	x	x						
$153 \leq p_T < 174$					x	x	x	x	x						HLT_Jet70U
$174 \leq p_T < 196$					x	x	x	x	x						
$196 \leq p_T < 220$					x	x	x	x	x						HLT_Jet100U
$220 \leq p_T < 245$					x	x	x	x	x						
$245 \leq p_T < 272$						x	x	x	x	x					HLT_Jet140U
$272 \leq p_T < 300$						x	x	x	x	x					
$300 \leq p_T < 330$							x	x	x	x	x	x	x		
$330 \leq p_T < 362$							x	x	x	x	x	x	x		
$362 \leq p_T < 1000$	HLT_Jet180U						x	x	x	x	x	x	x		

Table 6.1: selected bins for the analysis

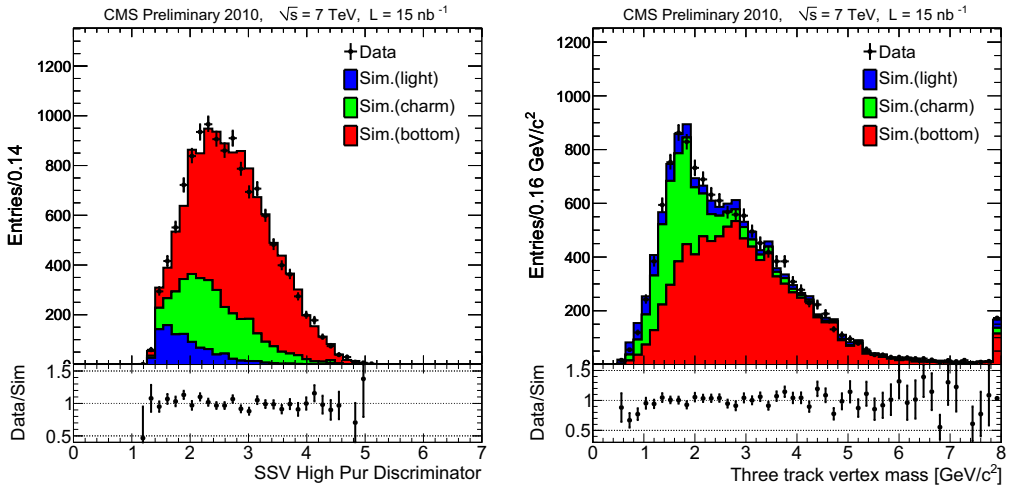


Figure 6.9: Distribution of the simple secondary vertex high purity b-jet tagger and the corresponding vertex mass reconstructed with at least three tracks.

	3 templates		2 templates	
p_T range [GeV]	f_b	err(stat)	f_b	err(stat)
$37 \leq p_T < 43$	0.718	0.014	0.719	0.014
$43 \leq p_T < 49$	0.722	0.017	0.721	0.017
$49 \leq p_T < 56$	0.747	0.019	0.750	0.019
$56 \leq p_T < 64$	0.761	0.022	0.761	0.022
$64 \leq p_T < 74$	0.745	0.027	0.742	0.027
$74 \leq p_T < 84$	0.815	0.037	0.815	0.038
$84 \leq p_T < 97$	0.741	0.013	0.741	0.013
$97 \leq p_T < 114$	0.719	0.016	0.720	0.016
$114 \leq p_T < 133$	0.693	0.008	0.691	0.008
$133 \leq p_T < 153$	0.713	0.012	0.713	0.012
$153 \leq p_T < 174$	0.720	0.011	0.717	0.011
$174 \leq p_T < 196$	0.702	0.015	0.702	0.015
$196 \leq p_T < 220$	0.729	0.016	0.727	0.016
$220 \leq p_T < 245$	0.689	0.024	0.690	0.023
$245 \leq p_T < 272$	0.678	0.026	0.685	0.025
$272 \leq p_T < 300$	0.671	0.035	0.671	0.034
$300 \leq p_T < 330$	0.710	0.044	0.712	0.045
$330 \leq p_T < 362$	0.735	0.061	0.755	0.065
$362 \leq p_T < 1000$	0.584	0.072	0.581	0.070

Table 6.2: Fractions of b-jets in a tagged jet sample extracted by fitting on the secondary vertex mass distribution

fits. The statistical uncertainties are also quoted. Further the result for the three template fit is plotted in figure 6.10. The fit result of each bin can be seen in appendix D.
Within its statistical limitation the fit results agree with the expected distribution. The expectation values are calculated from the samples which were used to determine the templates for the fit.
For the high p_T jets we see an overestimation of the b-jet fraction. To explain this the spectrum was studied dependent on the number of primary vertices existing in the event. The result of this is shown in figure 6.11 for the bins with enough statistics.
In line with the statistics it is possible to argue that the more flat distribution is caused by the existence of more than one primary vertex. These could be effects of an underlying event [CMS10i] or additional proton-proton interactions (pile up). To clarify this issue finally, more statistics is needed.
At last we tried to study the purity dependent on their rapidity. Therefore we had to reduce the number of bins in p_T to get sufficient statistics. The result is plotted in figure 6.12.
Again we find a good agreement between data and simulations.

6.2.4 Systematic uncertainties

In this section studies on the systematic uncertainties of the fitting procedure are presented.

Template statistics

The basic idea is that each template has a random status of the truth distribution for a given number of events. Each bin content fluctuates around an unknown truth mean μ. The fluctuations follow a Poisson distribution. The disagreement between these values and the truth contributes as

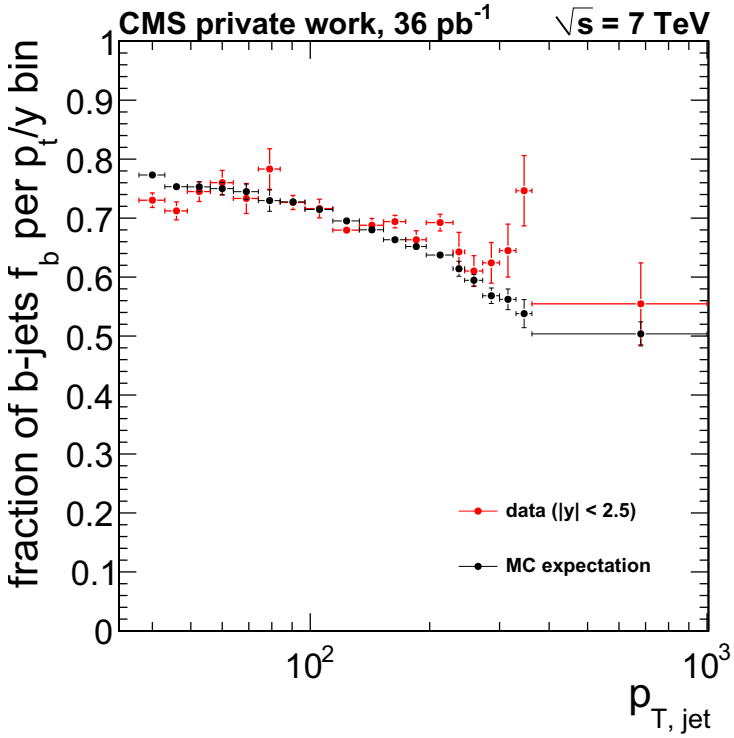


Figure 6.10: Result for the p_T spectrum determined by the flavor content fitter for $|y| < 2.5$. The parametrization is chosen to measure directly the b-jet fraction for tagged jets ($p = 1$). Only statistical errors are shown in this figure.

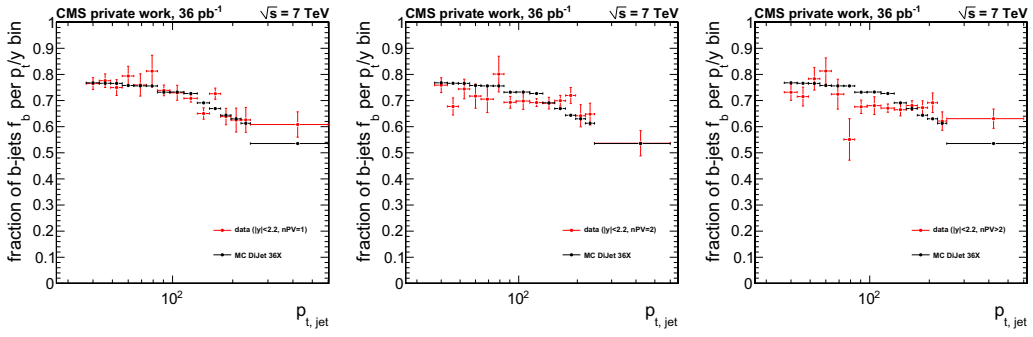


Figure 6.11: Study of the dependency on pile up effects. The histograms show the b-jet fraction of a tagged sample for different numbers of primary vertices. The statistics is too low to claim a final conclusion.

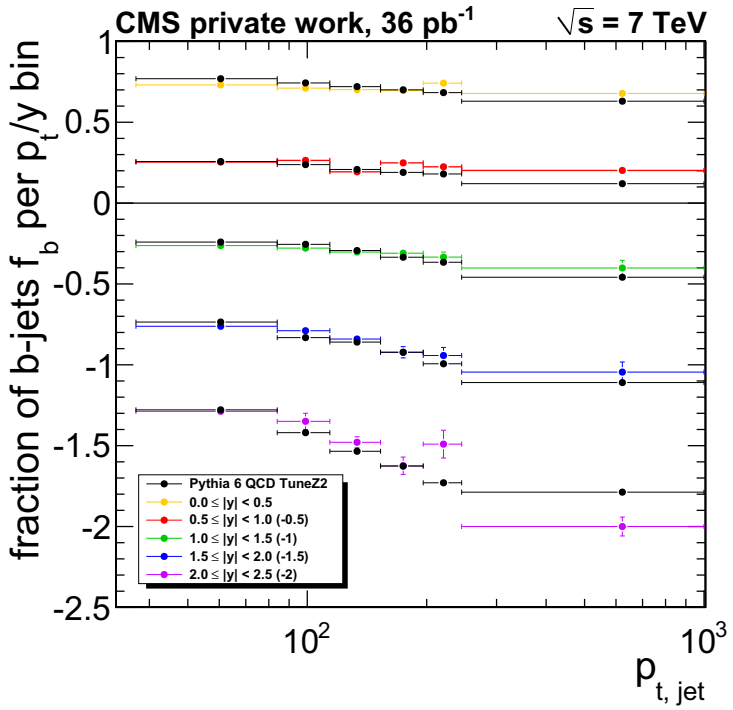


Figure 6.12: The plot show the measured dependencies of the b-jets in p_T and $|y|$ determined by the flavor content fitter. There is a agreement between data and simulations.

a systematic effect to our fitting procedure.
To estimate this systematics we vary the templates by changing the content of each bin with a random number following a Poisson distribution with the mean of the original bin value. These new templates are fed to the flavor content fitter. This is done many times. The results of such a variation can be seen in figure 6.13. The fit results vary around the original fit value. The width of this distribution can be taken as an estimate for the systematic uncertainty $\sigma_{tempStat}$.
Figure 6.14 shows the systematic uncertainty $\sigma_{tempStat}$ we presumed for the different bins. The systematics in the forward region are higher due to the smaller statistics in this region. Further we see a rise over p_T. This rising drops, when a bin is reached where a different Monte Carlo sample is used. This confirms the decision to use only the most relevant samples for a specific p_T regions. Overall the effects of large weights contribute to this kind of systematic uncertainty.

c and *light* fractions

The shapes of the templates look very similar for c-jets and light jets. The flavor content fit was performed with two and also with three templates. As shown in table 6.2 the result of both is the same for all fitting regions. The result for the b-jet fraction is independent fo the usage of the two or three template method.
Besides the measurement of the b-jet fraction it is also interesting to analyze the c-jet fraction. With the result of the three template method we also get an estimate on this quantity. Figure 6.14 shows on the right the fraction of the c-jets f_c. To calculate the c-jet fraction in the sample the fitted values must be multiplied by the non-b-jet fraction.

$$f_c = r_c(1 - f_b)$$

Dependent on the selections to get a pure b-jet sample the statistics for measuring this is poor and we get large statistical errors. A detailed study on f_c is not part of this thesis. Nevertheless

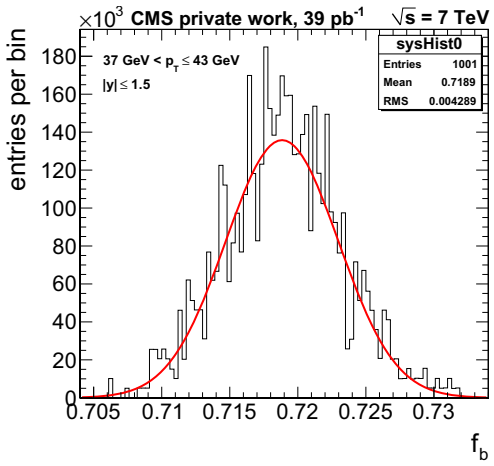


Figure 6.13: The fit results of different f_b estimations with varied templates describe nearly a Gaussian distribution with the mean of the original value. The width of this can be taken as systematic uncertainty $\sigma_{tempStat}$

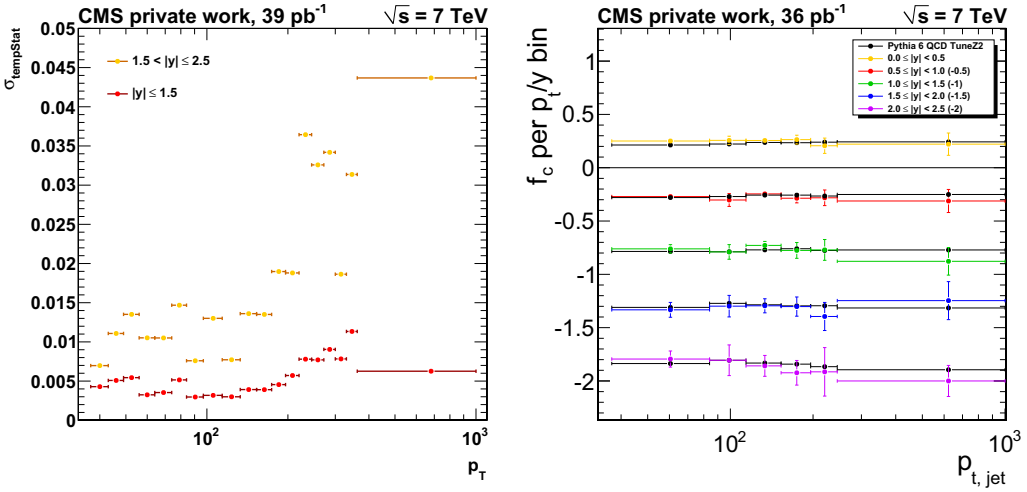


Figure 6.14: The plot on the left shows the variance of the flavor content fit, if the templates were varied within their statistical uncertainties. These contribute to the systematics. On the right the fraction of c-jets f_c is computed and compared to the MC expectations. The good agreement allows a more strict estimate on the uncertainty caused by badly simulated c-jets.

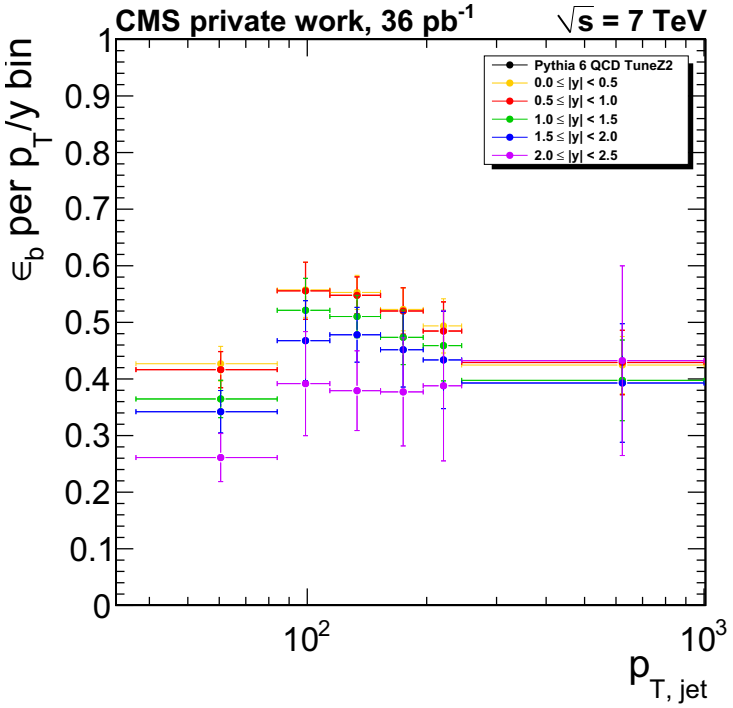


Figure 6.15: Efficiencies for b-jets passing the SSVHP>2.0 cut by p_T and y. The values are extracted from a Pythia 6 QCD Monte Carlo generator.

the agreement between data and simulations in the statistical context leads to a small systematic uncertainty involved by a possibly defective c-jet contribution. Using the three template method, systematic uncertainties caused by this are covered by a study of the variations of the template bin statistics.

6.2.5 Tagging efficiencies from Monte Carlo simulation

As described above, the jets are required to have a SSVHP discriminator greater than 2.0. This is the "tight" working point suggested by the b-tagging group. This cut rejects most of the light-jets which might not be well simulated in MC and could introduce huge differences between the light content of the data sample and the MC light-template. But on the other hand one has to measure the efficiency for a b-jet passing this discriminator requirement for each $p_T/|y|$ bin. Since the statistics on data is to low for an efficiency measurement, this is done on MC.

The b-tagging efficiency ε_b is defined as:

$$\varepsilon_b = \frac{N_{b_{tagged}}}{N_{b_{tagged}} + N_{b_{dumped}}}$$

where $N_{b_{tagged}}$ is the number of b-jets passing the $SSVHP > 2.0$ cut and $N_{b_{dumped}}$ is the number that fail this cut. So one only has to count the number of b-jets passing the cut and the number of those which are cut away.

Figure 6.15 shows the efficiency extracted from the Pythia 6 QCD Monte Carlo samples for the different bins.

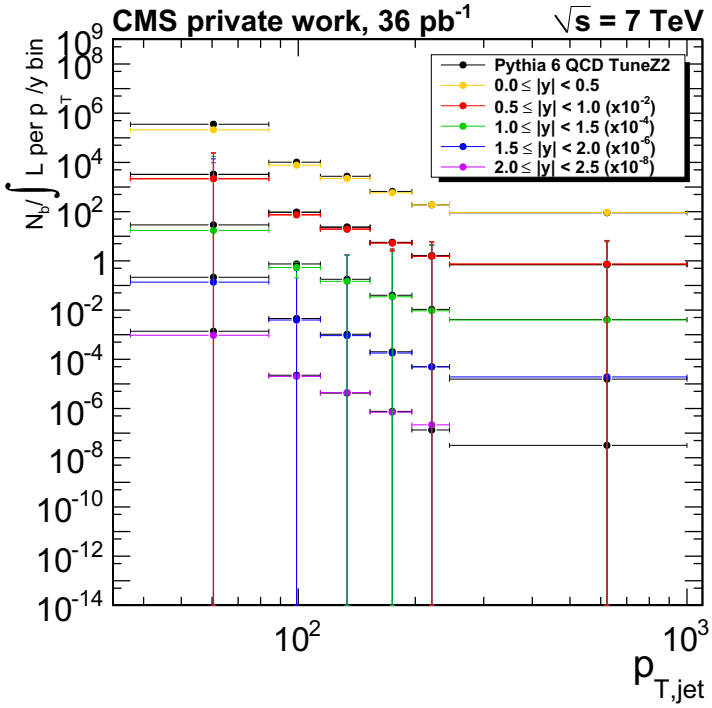


Figure 6.16: Normalized distribution for the number of b-jets extracted by the flavor content fit. The values are compared with a Pythia 6 QCD TuneZ2 expectations. There is good agreement.

6.2.6 Updated result

In the previous sections I showed the updates I did for the single parts of the analysis. Now we have to combine the results of all colleagues working at other parts of this analysis to get the final b-jet cross section measurement. Due to these constraints, for this thesis it was not possible to get this result.
Unfortunately the unfolding procedure does not work as expected for the increased statistics. The steeply falling p_T spectrum leads to a systematic effect while applying the ansatz fit. More discussion in the collaboration and maybe an alternative approach are needed to produce a final result.

Nevertheless I will present a picture of the measurement, where the all inputs are combined, but without performing the unfolding. Let us have a look at the differential b-jet cross section:

$$\frac{d^2\sigma_{\text{b-jet}}}{dp_T dy} = \frac{N_{\text{jet}}\, f_b}{\varepsilon_b \int \mathcal{L}} \cdot \frac{C_{\text{unfold}}}{\Delta p_T \Delta y} = \frac{N_{\text{b-jet}}}{\int \mathcal{L}} \cdot \frac{C_{\text{unfold}}}{\Delta p_T \Delta y}.$$

We have already measured the b-jet purity f_b and the number of jets N_{jet}. From our CMS colleagues we get the results of the measurements of the integrated luminosity $\int \mathcal{L}$. The b-jet efficiency is still extracted from MC. Apart from the unfolding correction factor we are able to create a normalized histogram for the number of b-jets. Figure 6.16 shows the determined values. In addition the expected distribution calculated by Pythia 6 QCD TuneZ2 is drawn. No disagreement between data and MC is found.
At last I will give an outlook at the expected b-jet cross section distribution. As claimed before the unfolding procedure is not yet accomplished. The plot in figure 6.17 shows the result with updated statistics.
The result is compared with MC@NLO simulations. Due to the unfolding challenge values are only calculated up to a certain threshold.

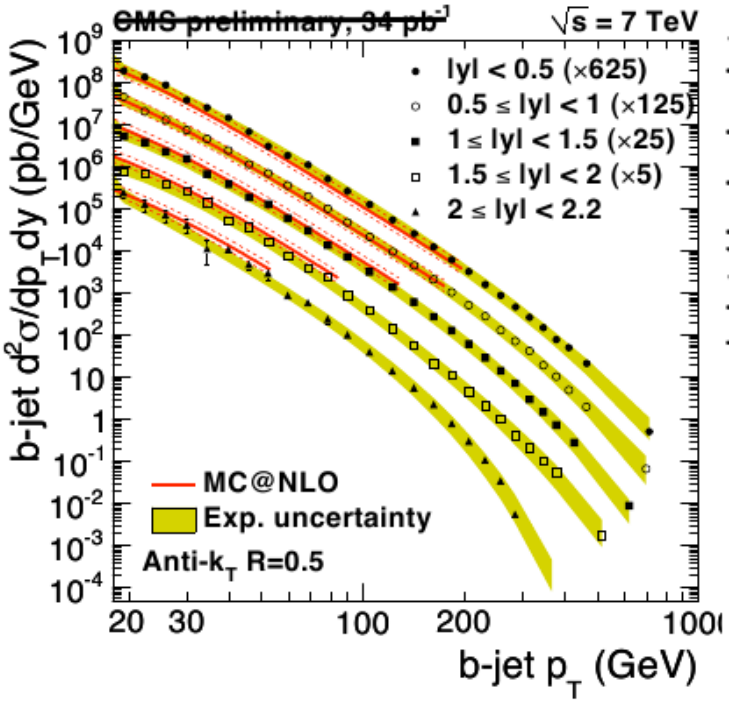


Figure 6.17: The complete analysis of the b-jet cross section measurement was updated. Most subparts are updated. The update of the purity measurement presented here is included. Having all analyss together a similar plot will go for publication.

6.3 NeuroBayes application

The methods presented for the b-jet cross section measurements do not yet use NeuroBayes at all. There are different possibilities to do this analysis with the use of multi variate analysis techniques. The apparent is the use of the NeuroBayes b-jet tagger, which was presented in this thesis. In the following I will describe, how the flavor content fit method can be changed for the new b-jet tagger.

6.3.1 NeuroBayes template fit

The obvious application of the NeuroBayes b-jet tagger is to use the discriminator variable within the flavour content fitter. There are different possibilities to include this tagger.
It is possible to change the target of the template fit. Recently the distribution of the secondary vertex mass was used. The secondary vertex mass is a reliable variable for analysis on early data. It has an adequate separation between b-jet and non-b-jets. Further it has a understandable shape, which makes it easier to notice possible problems. For first data the secondary vertx mass was a good choice.
Now we have studied all variables which are useful for b-jet tagging. The reconstruction software has been calibrated and we found a good agreement between data and simulations (see section 5.3.3). This allows us to use another variable with more discrimination power instead of the secondary vertex mass.

Further we can use another variable for the selection of the subsample which is used for the template fit. Therefore the simple secondary vertex b-tagger must be replaced by a new one. The cut down of the sample was made to reduce a possible dependency on badly simulated background distributions. But the application of this cut had its prize. The sample was reduced not only

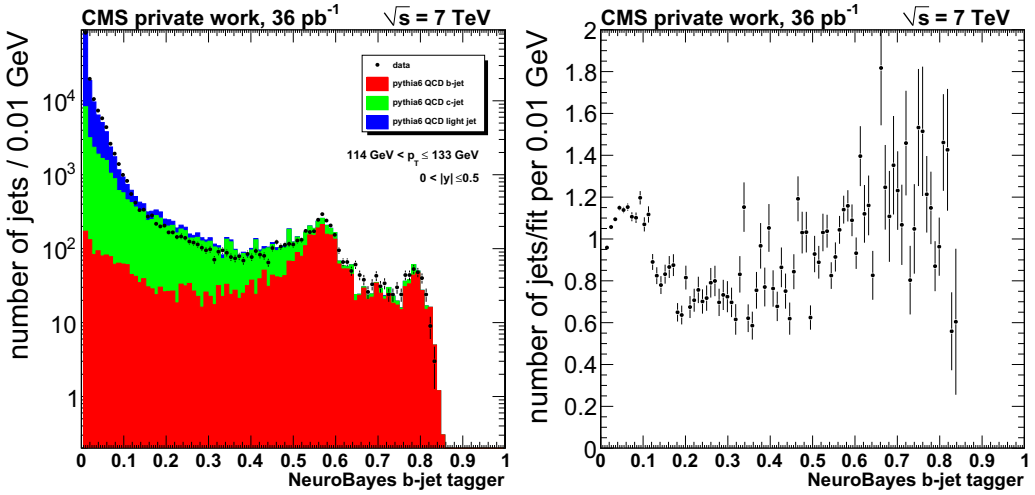


Figure 6.18: Exemplary result of the flavor content fitter. Target distribution is the NeuroBayes b-jet tagger output. On the left the result of the fit to data distribtuion of the three different classes is shown. The right plot shows the ratio between fitted templates and data distribution.

in light-jets and c-jets, but also in b-jets. For the b-jet cross section analysis an estimate on the b-tagging efficiencyiy ε_b is needed. For the recent measurement we took the expectation value from MC. This was necessary, because the accuracy we got by an efficiency estimate on data was not good enough. We did not gain by such a measurment.
Now we have enough data for a reasonable analysis of the b-tagging efficiency. Such measuments will be done at CMS coordinated by the b-tagging group (BTV POG) for all official CMS b-jet taggers. Unfortunately the new NeuroBayes b-jet tagger is not yet official. A measurement of its efficiency is planned at our institute in the future.

In the following I will present what happens if the flavour content fit (FCF) is applied to the whole data sample without a preselection. The FCF is used in the usual settings with three templates for light-jets, c-jets and b-jets. For the templates the inclusive distributions of the NeuroBayes combined b-jet tagger, which was introduced in section 5.3.2, are used. Figure 6.18 shows an exemplary result of a specific p_T/y bin, representative for all bins (see appendix E).
On the left the template distributions are plotted with a logarithmic scale on the y-axis. The amount is scaled to the numbers extracted from the fit. In black the data points are plotted for an easy comparison. It is already visible that in the central region we miss some jets. The plot on the right confirms this. Here the ratio between data and the final fit distribution is plotted.
It seems that the MC distribtuions are not as well simulated as expected. For small values of the NeuroBayes output distribution $o_t \approx 0.05 - 0.1$ we found an excess of jets in data. This is in a region dominated by light jets. It seems that they appear at larger o_t values than expected. Because of the large fraction of light jets this effects the fit of the other two templates. Jets from the other two classes are needed to compensate the missing light-jets. In our case this leads to an overestimation of c-jets. In the central region we can see this. The fit expects more jets than available in data. At last the b-jet template fits the rest of the distribution not yet covered by the c-template. The fitted fraction tends to be smaller than the expected values form MC.
Figure 6.19 shows the result of the flavour content fitter applied in all p_T/y bins for b-jets. All fit results lie below the expectations.
The simulations are not good enough for the application of the flavor content fitter at this level. A more detailed study on inclusive light jet distributions is needed to identify the objects which

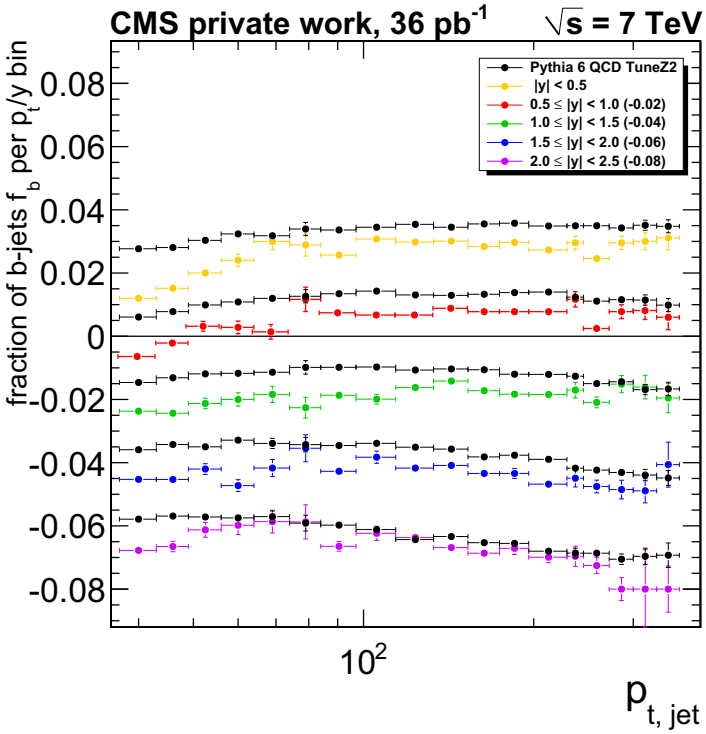


Figure 6.19: The flavour content fitter was used with the NeuroBayes b-jet tagger. the result of the fit in different p_T/y bins is shown. No preselection on the sample was applied. The NeuroBayes variable is sensitive on differences between b-jets and non-b-jets. Due to the differences in the inclusive shapes no convincing fit result is found. The bad application of the fittting procedure points to an insufficient simulation of the jet distributions.

cause the shape differences between data and MC.

Chapter 7

Conclusion

The modeling of b-quark production is one of the most challenging topics in elementary particle physics. Although the theory behind, QCD, was developed in the 1960s it was not possible to predict the correct heavy quark cross sections for a long time. Especially for the b-quarks this leads to curious discrepancies between measured results from experiments compared with insufficient predictions. In the 1990s the experiments at Tevatron (Run 1) as well as the experiments at LEP claimed an excess in b-quark appearance. Not until the revision of the next to leading order calculations, where the expansion of large logarithmic terms were incorporated, and the enhancement in the parametrization of the non-perturbative parts acceptable calculations of QCD were found. Following this measurements at the CDF detector during Tevatron Run 2 approved these calculations.

Today we redid the old CDF measurements at the CMS experiment. In contrast to the results at that time, further improvement on the theory were included. At HERA the proton structure function was measured in detail. This enables a further, more accurate verification of the theory.

The analysis published so far was done with very early CMS data with an integrated luminosity of 60/nb. The results were also presented in this thesis. We found an overall good agreement between data and Pythia in the jet transverse momentum range $30 < p_T < 150$ GeV and rapidity $|y| < 2.0$, within about 2% statistical uncertainty and 21% systematic uncertainty. In comparison with NLO@MC predictions we found significant differences in shape.

Furthermore this thesis presented the measurements on the b-jet purity for an update of the recent CMS analysis. The update includes the experiences of one year data taking which results in an integrated luminosity of 36/pb. The result will be published in the near future.

I showed, that the results from the recent and the updated measurement base on the simple secondary vertex b-jet tagger. This is a very robust tagger developed for the use on early data. For further improvements of the b-jet cross section analysis it is recommended to move to a more powerful tagger. For this a new b-jet tagger was constructed. This b tagger uses the multi variate analysis framework NeuroBayes.

The layout and the features of the NeuroBayes framework were described in detail in this thesis. I presented new multi variate tools based on NeuroBayes, which allows us to compare data and Monte Carlo simulations. These tools facilitate a quick and easy search for unexpected aspects of the data. This comparison was done for jets in the b-jet specific phase space. A good overall agreement was found.

This knowledge enabled the construction of a data based b-jet tagger. With this specific approach it is possible to ignore the possibly badly simulated background events from Monte Carlo. Instead,

this information is taken from the data sample.
I showed that it is further possible to correct for the small difference of the data/MC comparison. The correction factor was calculated from output values calculated by the NeuroBayes expertise. Further improvements were made by the so called boost method, which optimizes the classification for a pure b-jet selection.
The final data based b-jet tagger was compared to existing b taggers from the CMS collaboration. This is usually done on the different working points called loose, medium and tight, which correspond to the values in b tag efficiency calculated at mistag rates of 10%, 1% and 0.1%. Compared to the official b-jet probability tagger (JBP) I found efficiency improvement of 3%, 7% and 29%.

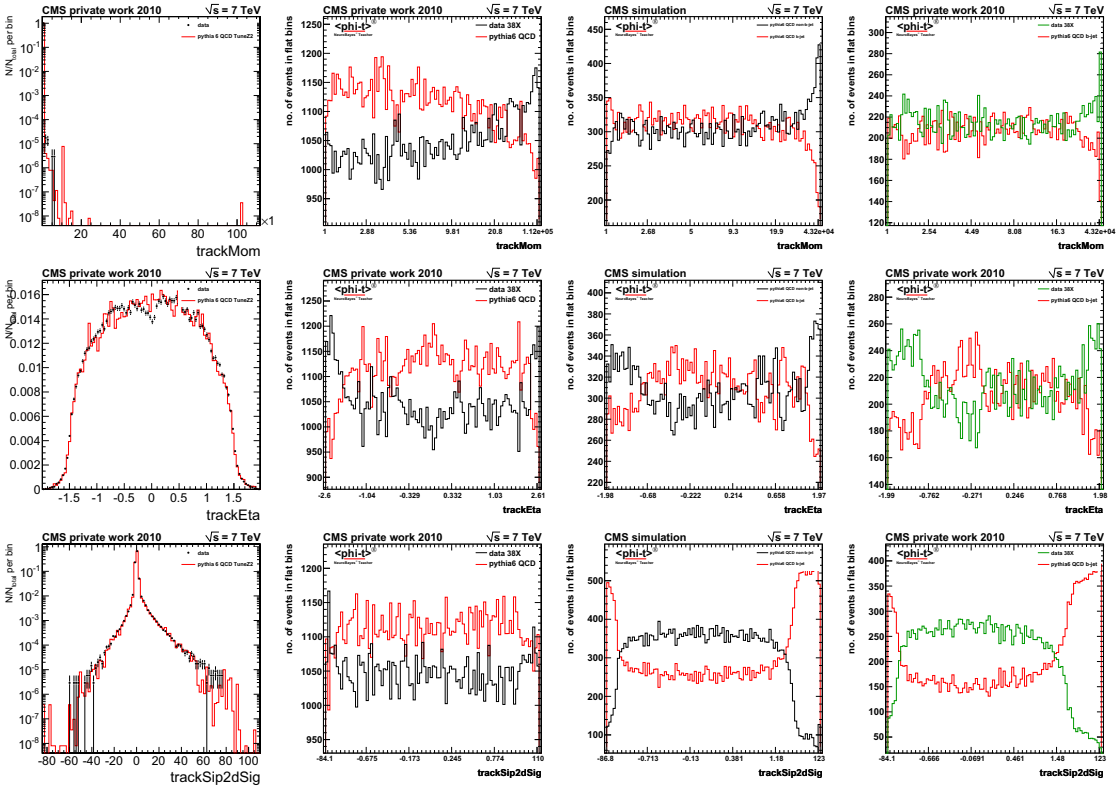
The new b-jet tagger is an additional tool for many analyses planned at the CMS experiment. Especially the possibility to identify b-jets with a small rate of misidentification qualifies for studies of heavy particles which decay in b-quarks. For SUSY searches, exotic particles and top physics the use of this tagger may play a decisive role.

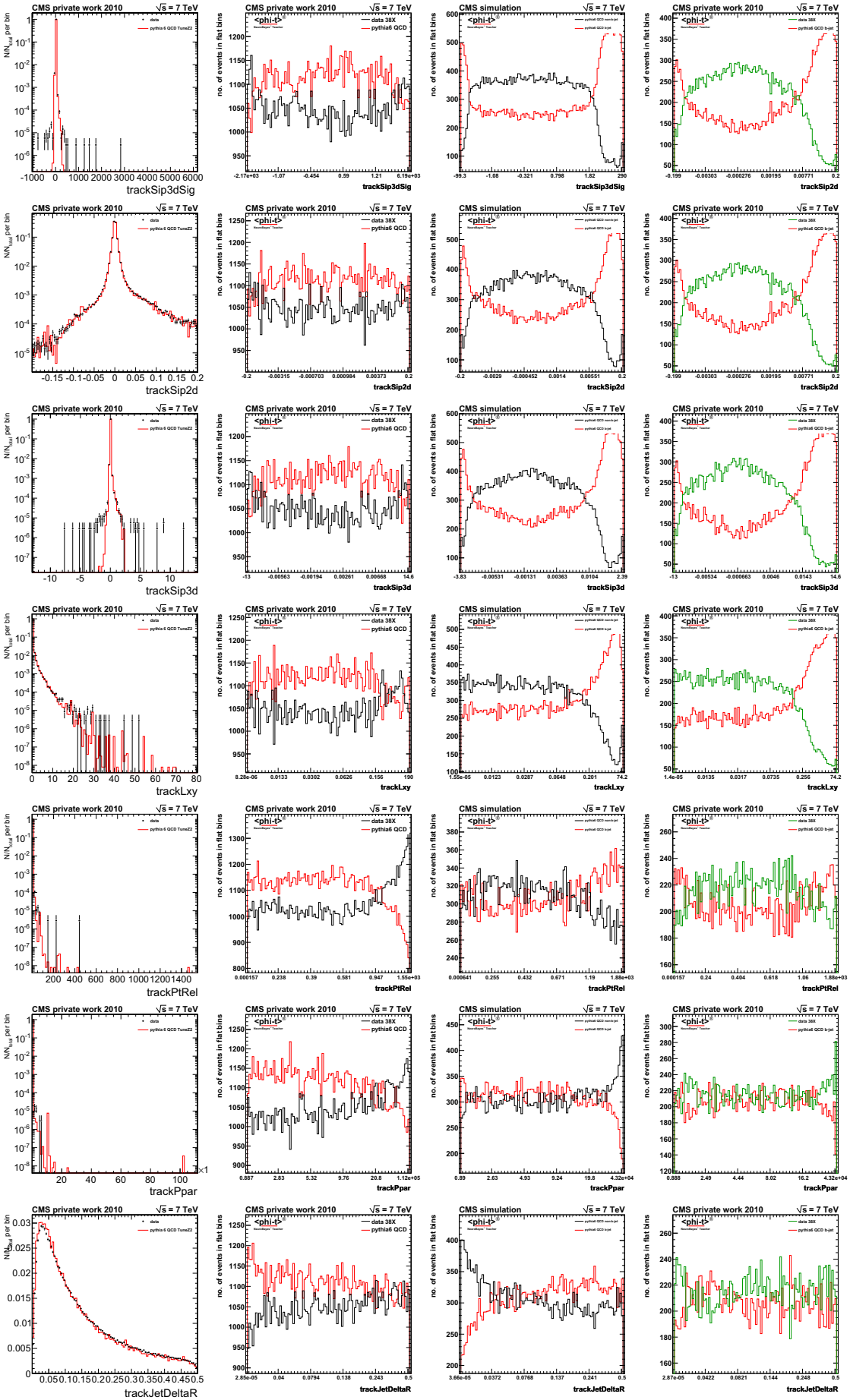
The full output of the new b-jet tagger is a strongly discriminating variable for inclusive jet distributions. Thus a measurement of the b-jet appearance was arranged. A large dependency on the light jet contribution was found. For a final result more detailed studies of the background distributions are needed. With this the measurement of the differential b-jet cross section can be improved.

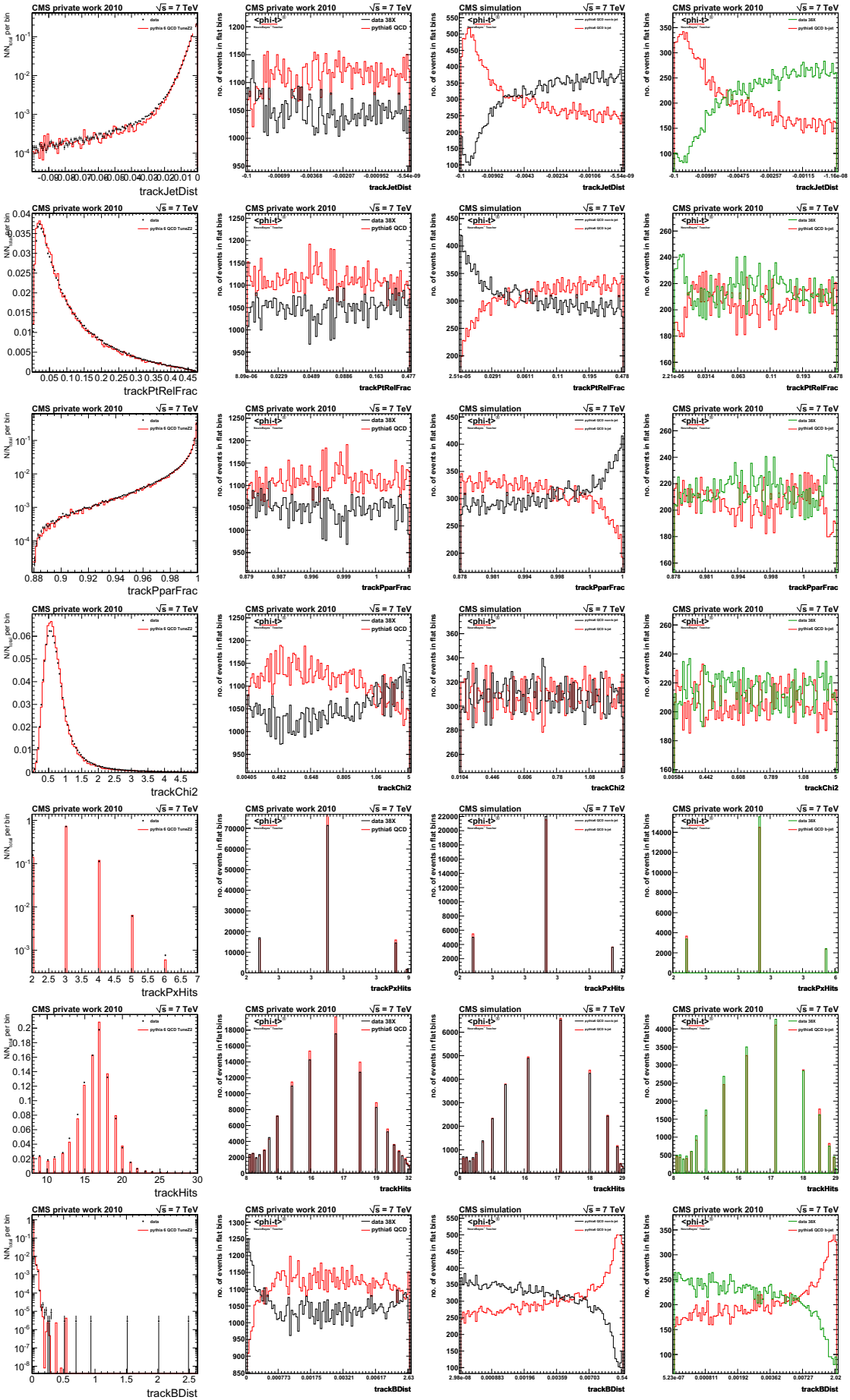
Appendix A

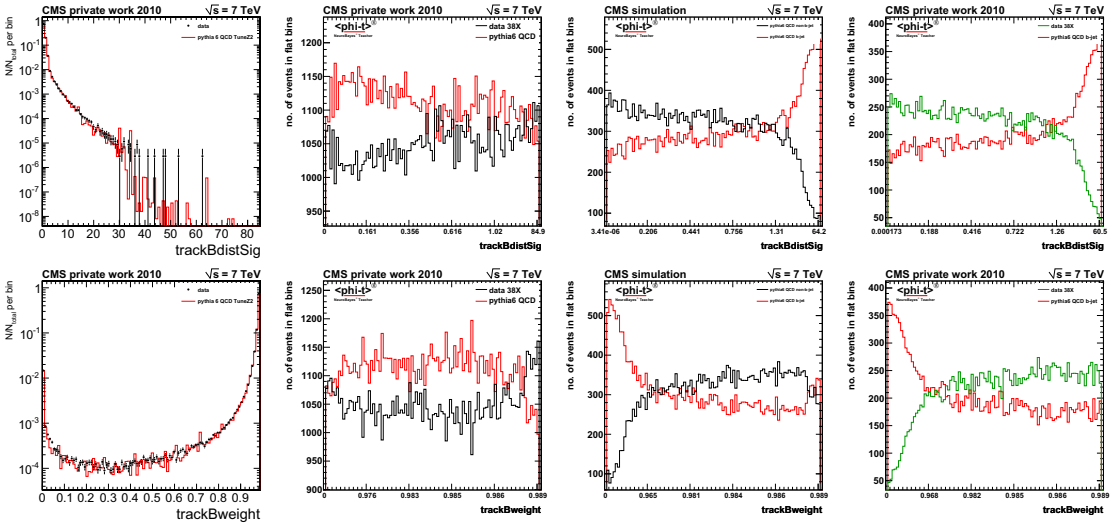
Distributions of b-jet tagging variables

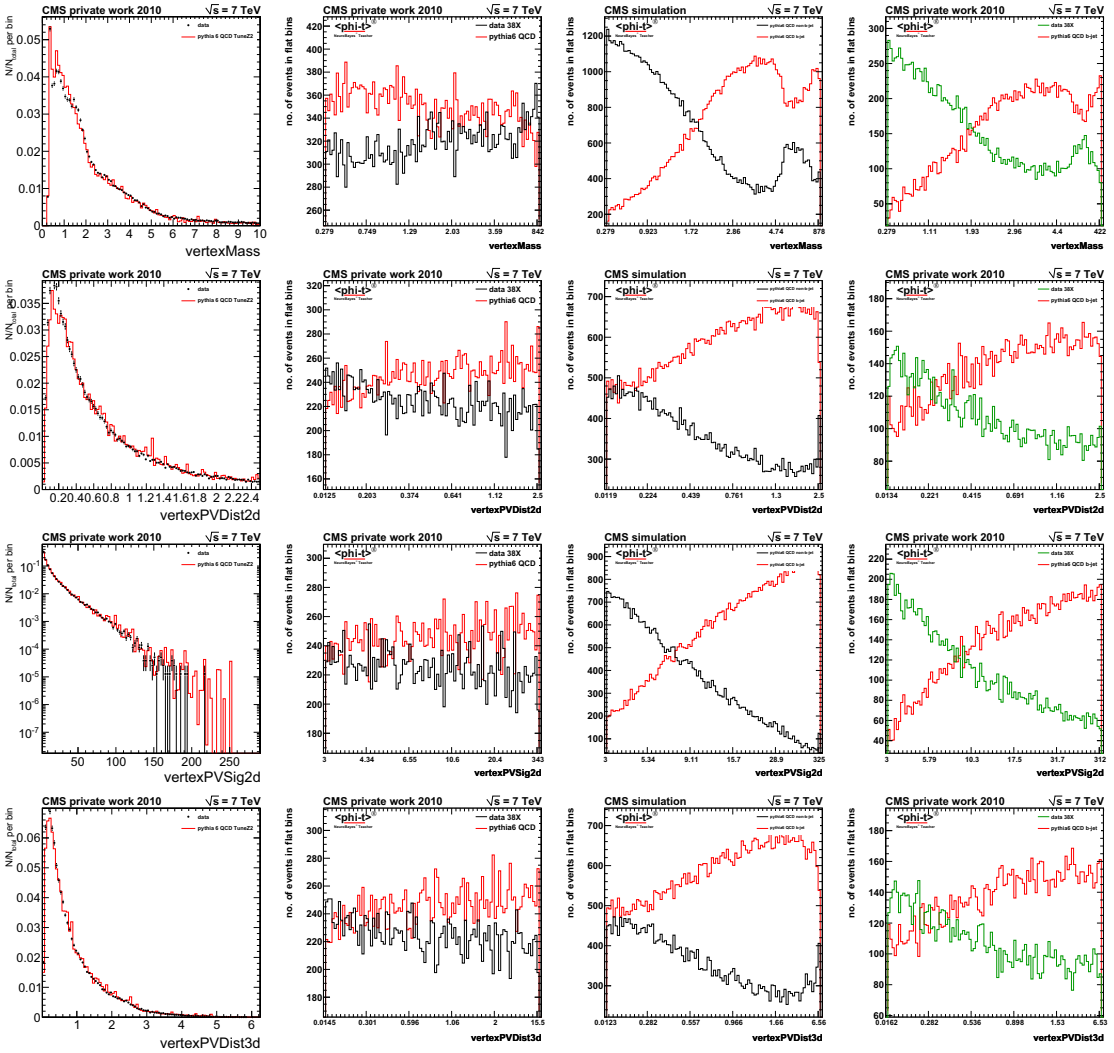
The following plots contain the distributions of all variables used in this thesis. The variable name is labeled on the x-axis as defined in section 5.3. The plots are sorted in the same order as introduced. The first plots show the data (black) and MC (red) distributions of each variable. For an quick and easy comparison the variables are plotted many times: in the first column they are the classical histograms (partly with logarithmic y-axis). The last three columns show the result of the probability integral transform: for data/MC, nonb-b-jet MC/b-jet MC and data/b-jet MC.

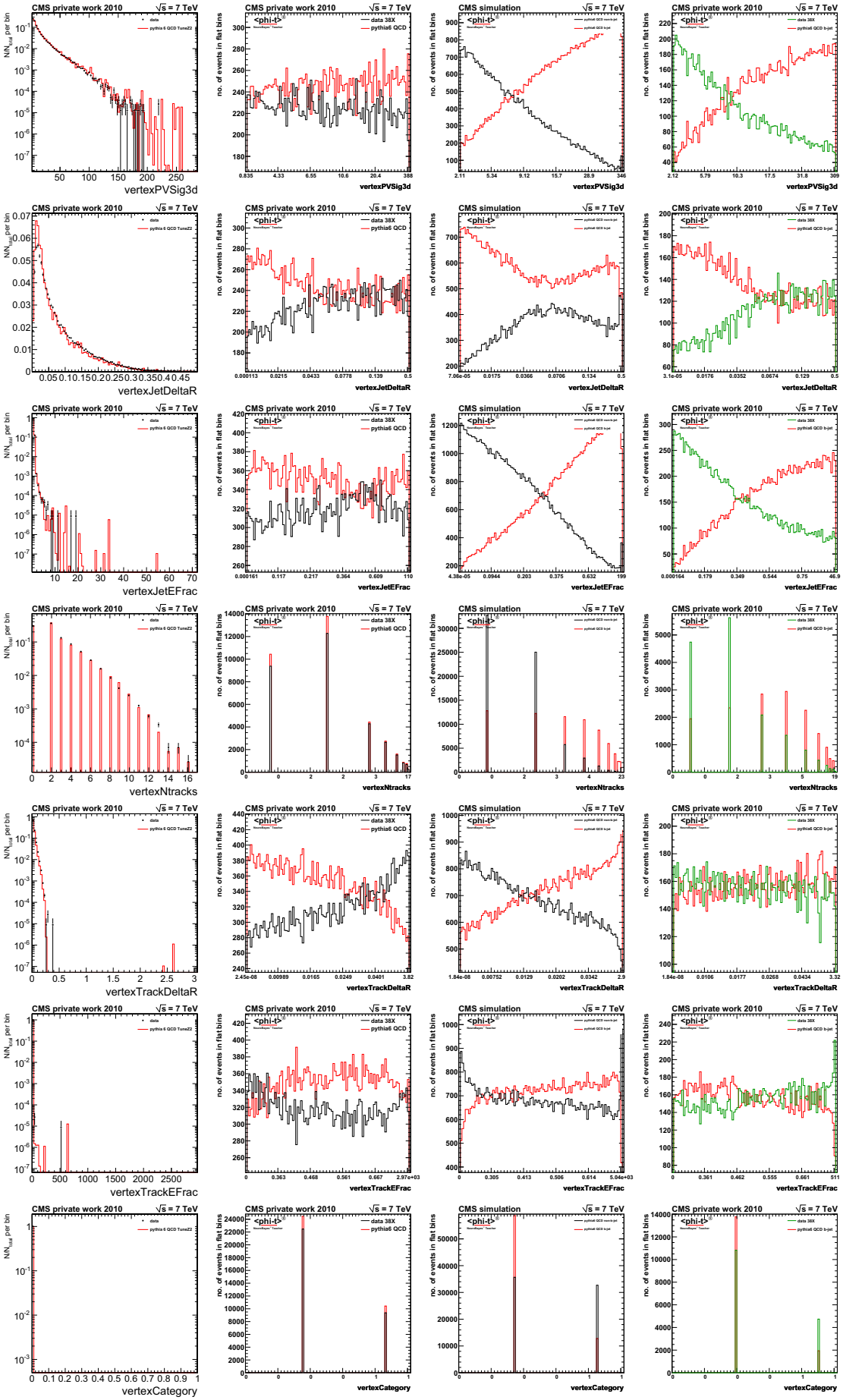


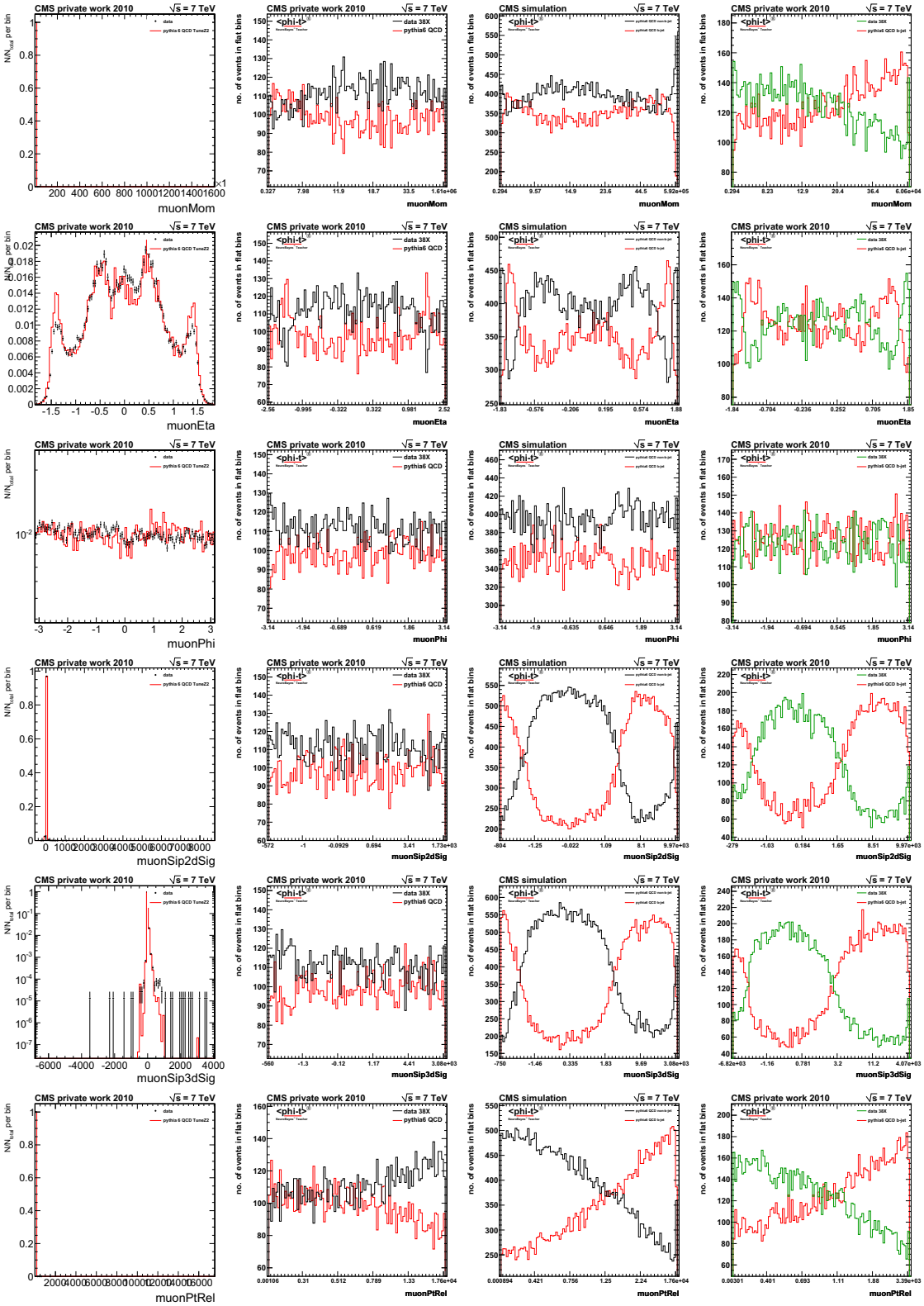

CMS private work 2010
CMS simulation
√s = 7 TeV
trackSip3dSig
trackSip2d
trackSip3d
trackLxy
trackPtRel
trackPpar
trackJetDeltaR

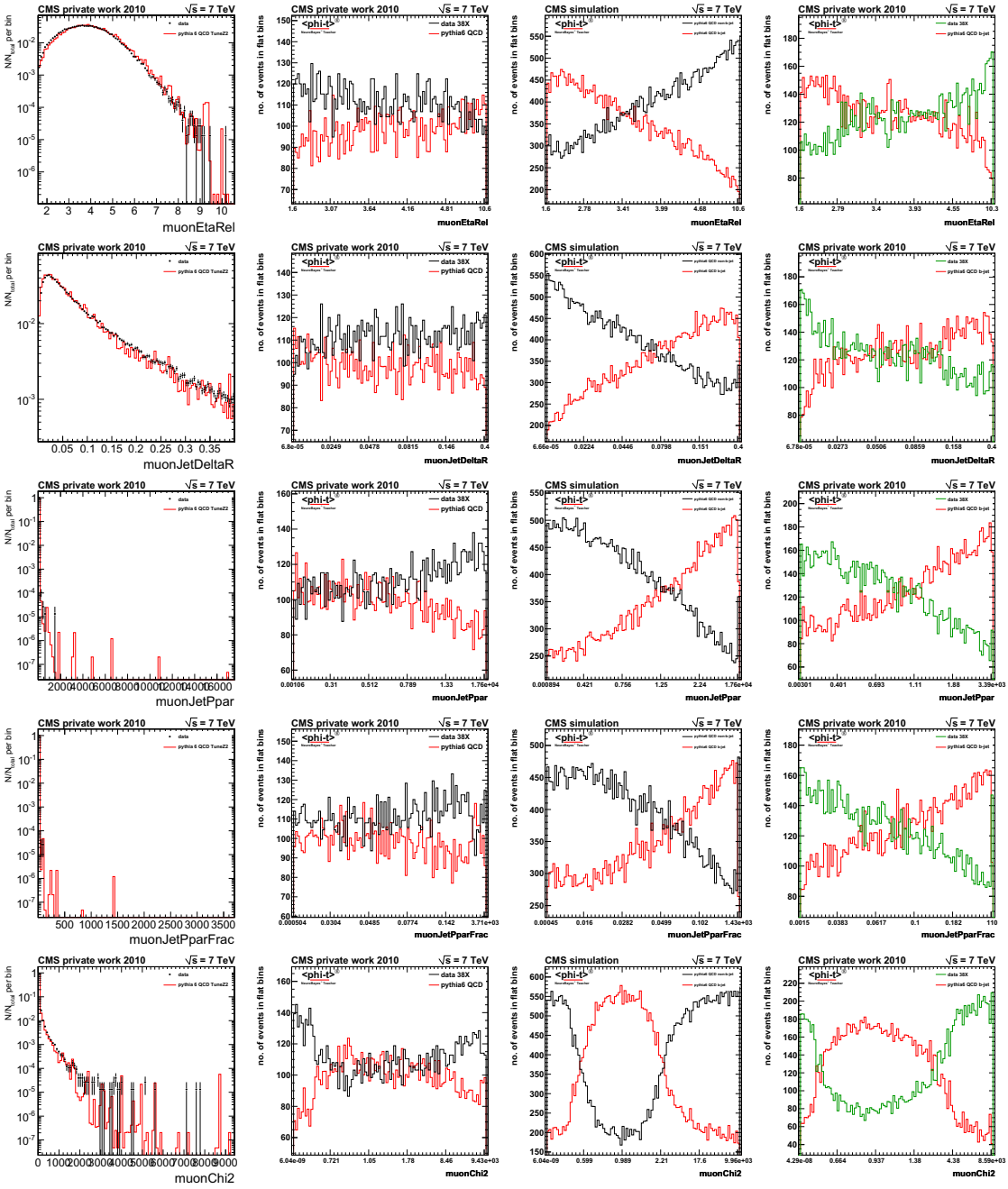

CMS private work 2010
CMS simulation
√s = 7 TeV
data
pythia 6 QCD TuneZ2
data 38X
pythia6 QCD
pythia6 QCD b-jet
<phi-t>
N/N_total per bin
no. of events in flat bins
trackJetDist
trackPtRelFrac
trackPparFrac
trackChi2
trackPxHits
trackHits
trackBDist

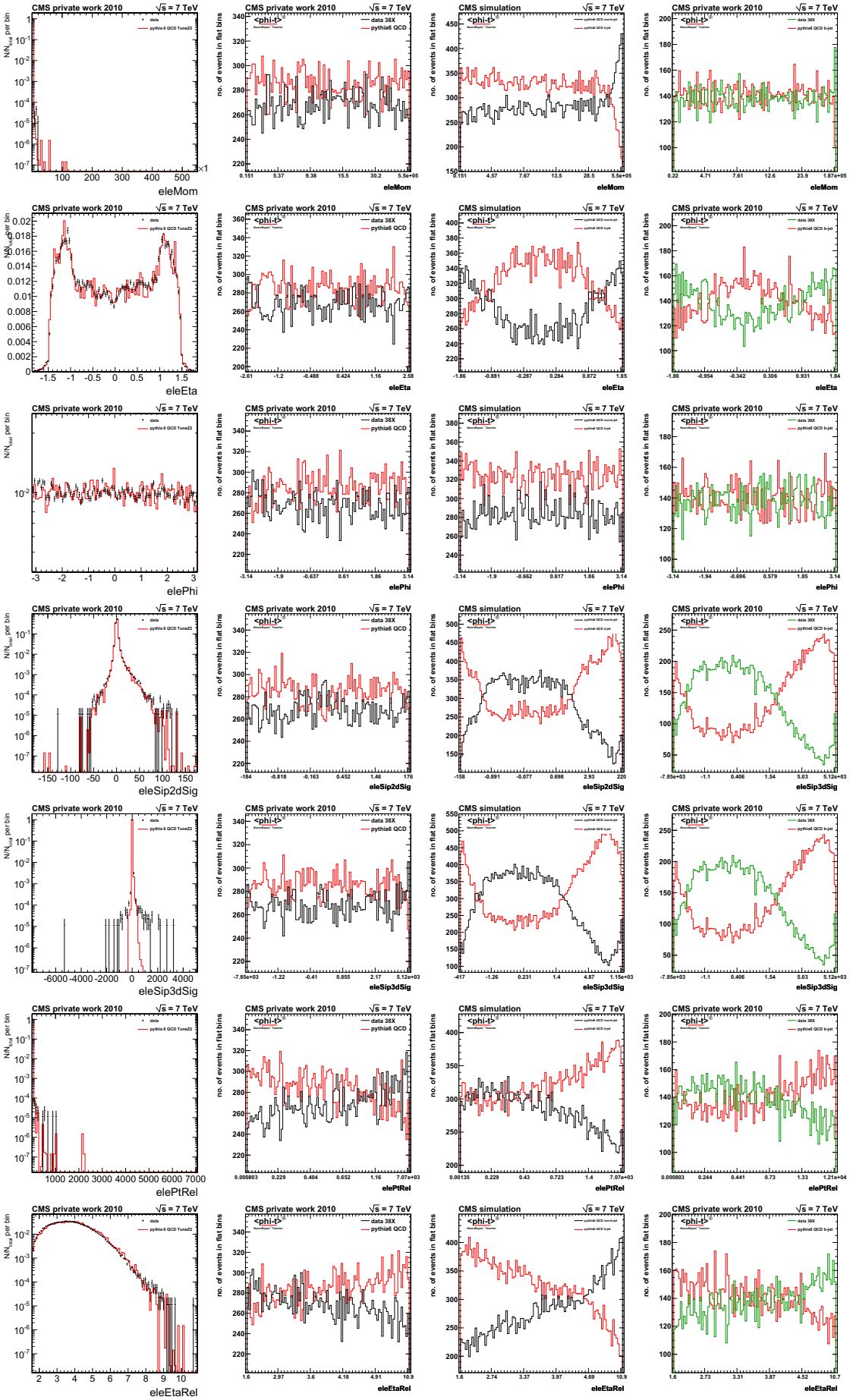

CMS private work 2010
√s = 7 TeV
CMS simulation
trackBdistSig
trackBweight

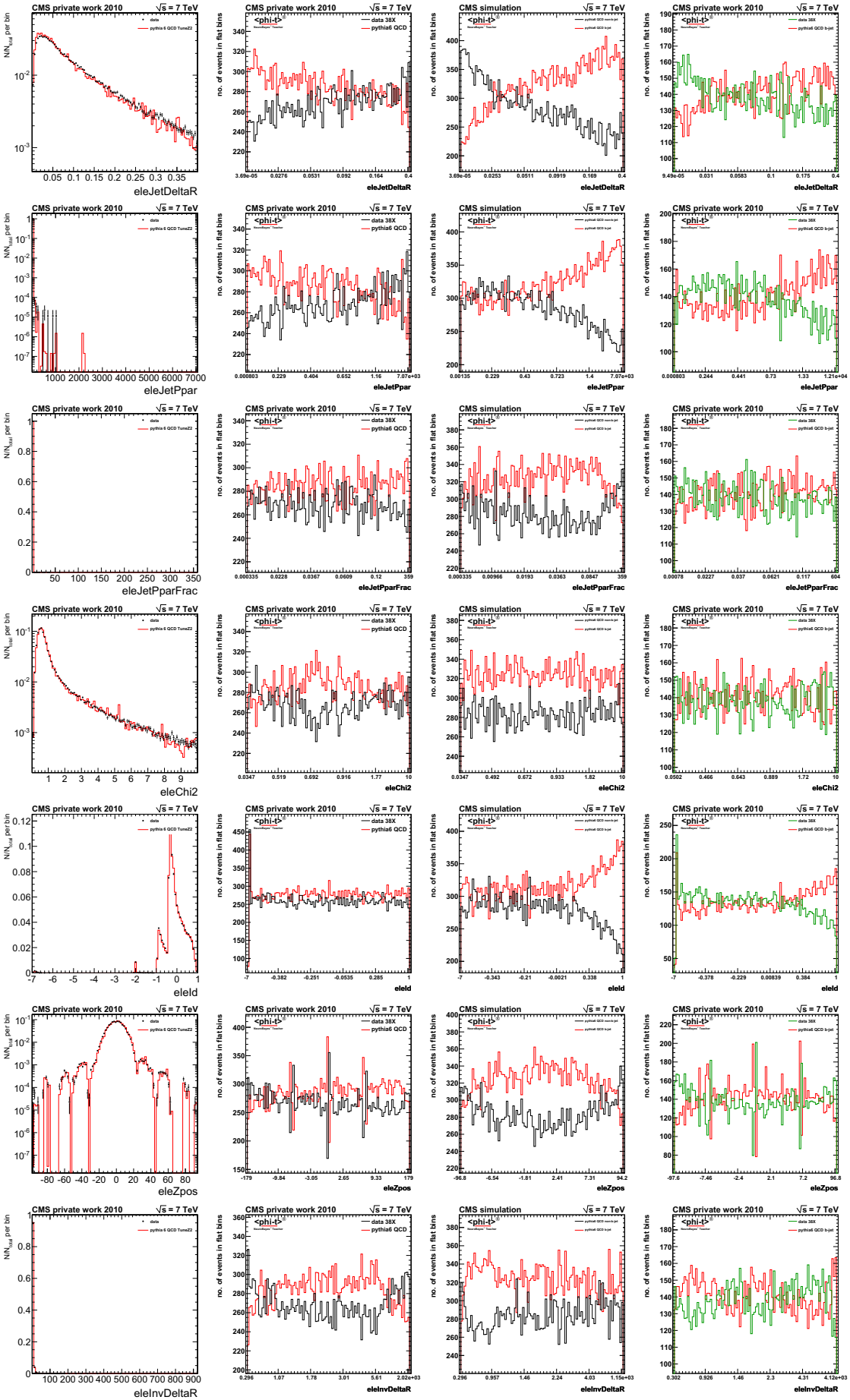

CMS private work 2010
√s = 7 TeV
CMS simulation
vertexMass
vertexPVDist2d
vertexPVSig2d
vertexPVDist3d

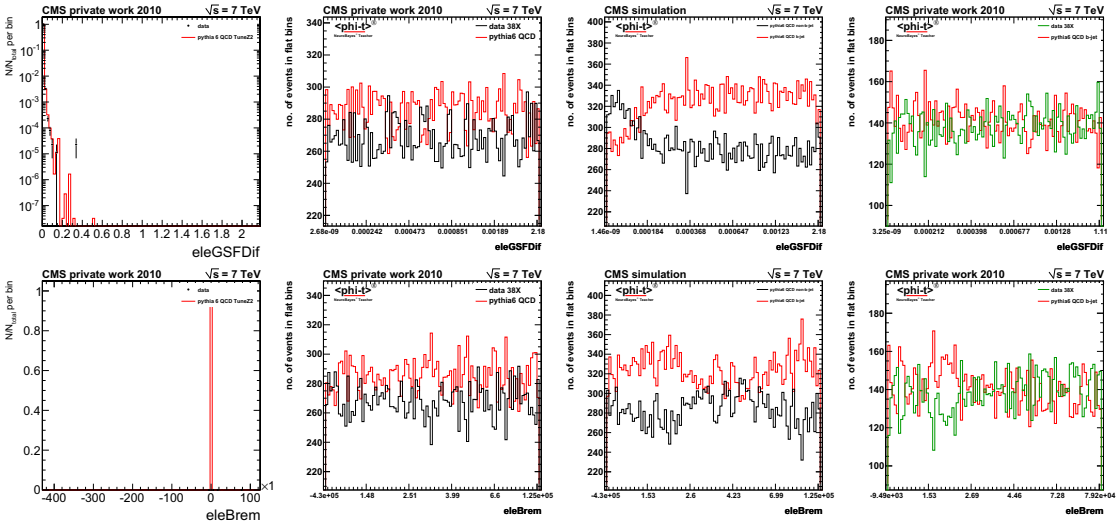

CMS private work 2010
CMS simulation
√s = 7 TeV
data
pythia 6 QCD Tune Z2
data 38X
pythia6 QCD
pythia6 QCD b-jet
<phi-t>
N/N_total per bin
no. of events in flat bins
vertexPVSig3d
vertexJetDeltaR
vertexJetEFrac
vertexNtracks
vertexTrackDeltaR
vertexTrackEFrac
vertexCategory

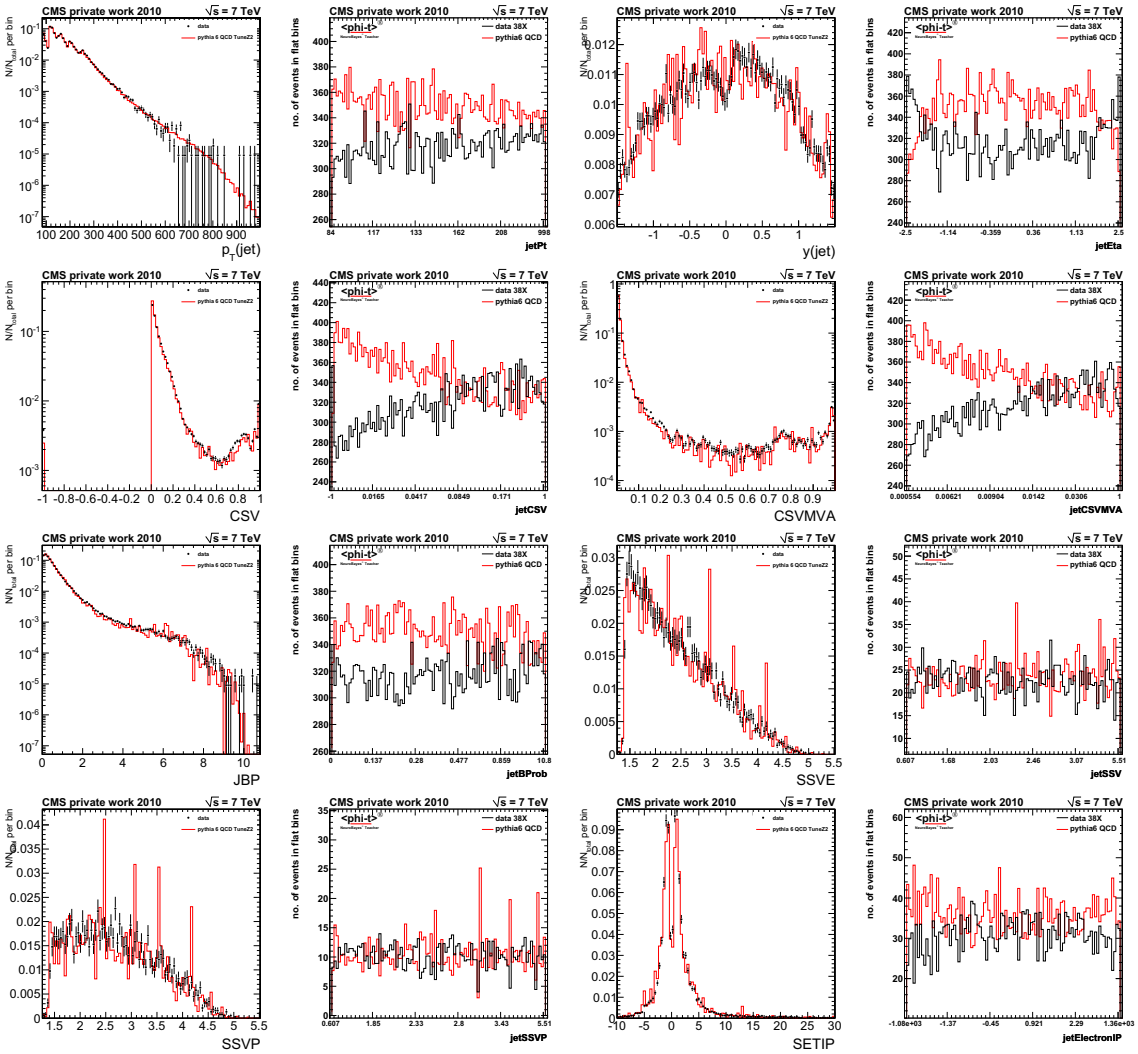

CMS private work 2010
CMS simulation
√s = 7 TeV
data
pythia 6 QCD TuneZ2
data 38X
pythia6 QCD
pythia6 QCD b-jet
N/N_total per bin
no. of events in flat bins
muonMom
muonEta
muonPhi
muonSip2dSig
muonSip3dSig
muonPtRel

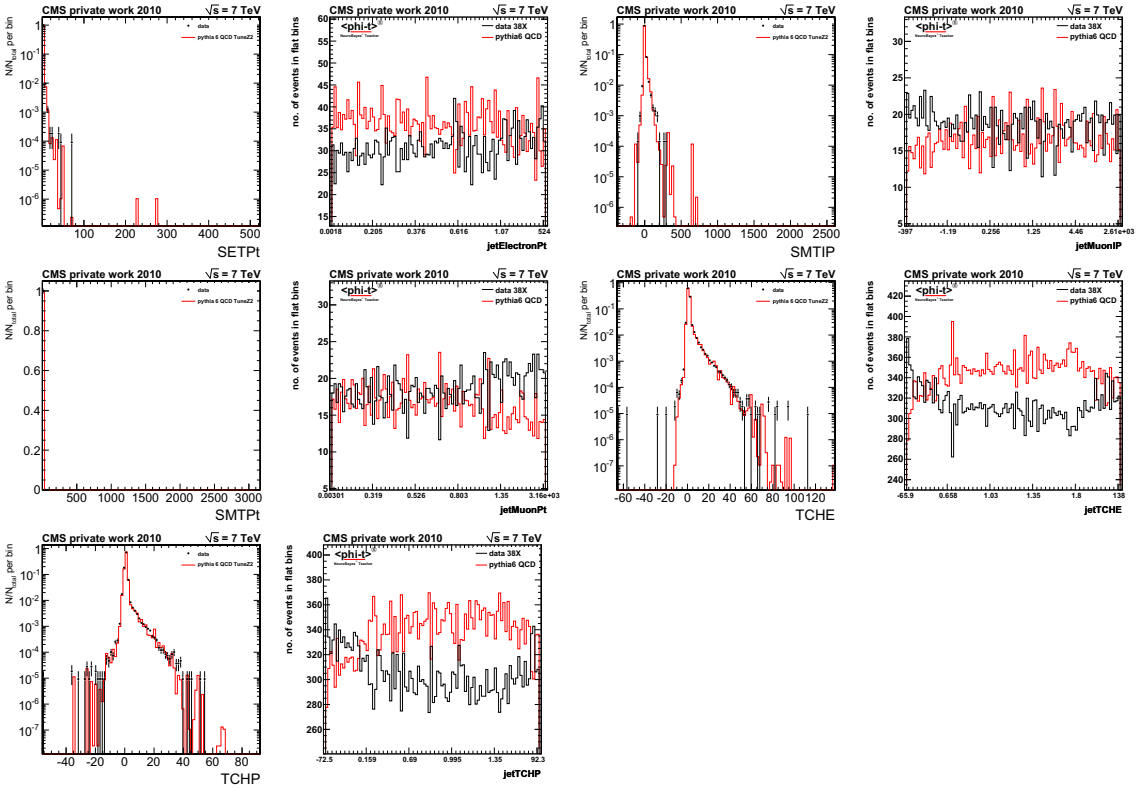

CMS private work 2010
CMS simulation
√s = 7 TeV
data
pythia 6 QCD Tune Z2
data 38X
pythia6 QCD
pythia6 QCD b-jet
N/N_total per bin
no. of events in flat bins
muonEtaRel
muonJetDeltaR
muonJetPpar
muonJetPparFrac
muonChi2

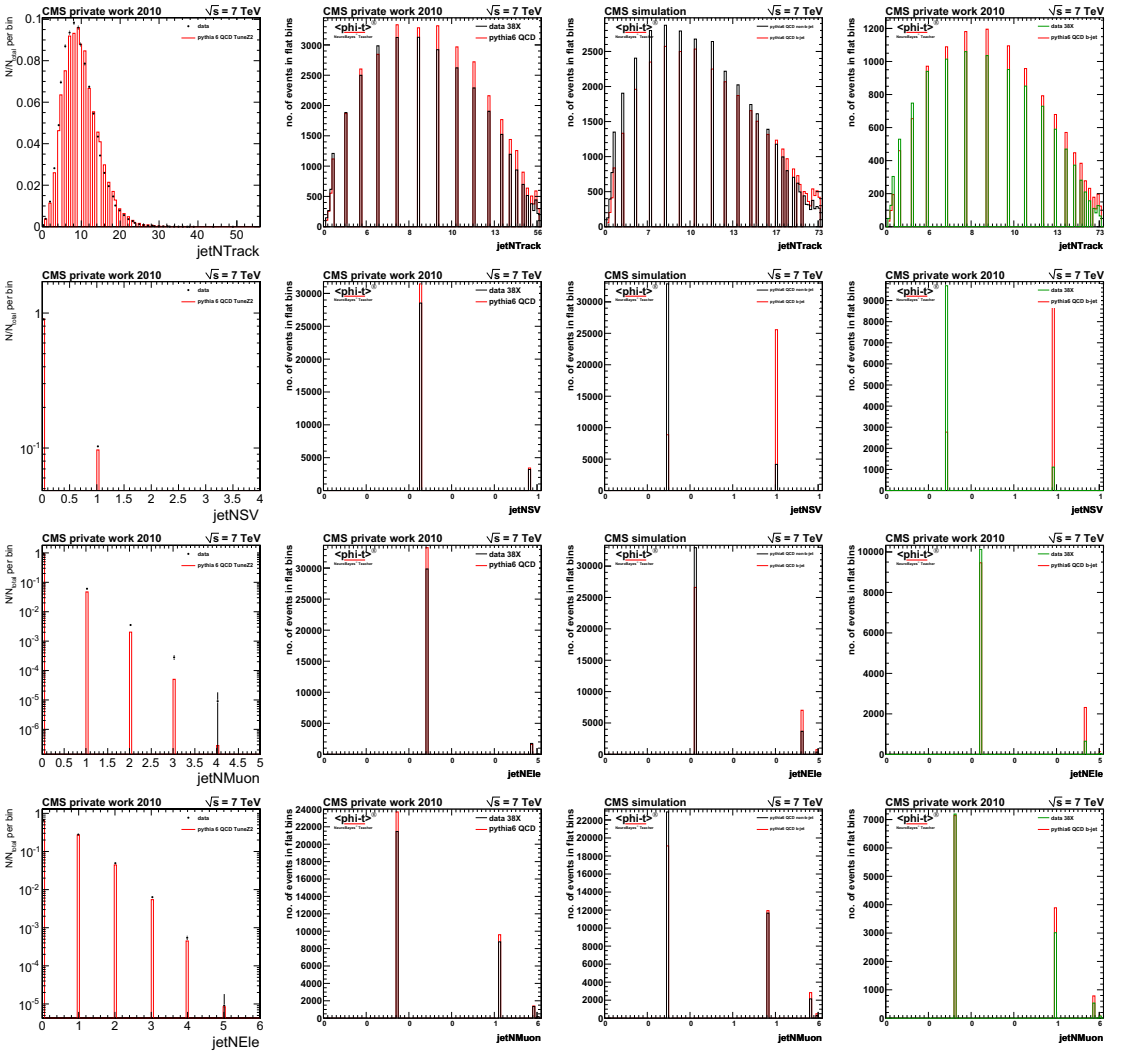

CMS private work 2010
CMS simulation
√s = 7 TeV
data
pythia 6 QCD TuneZ2
data 38X
pythia6 QCD
pythia6 QCD b-jet
<phi-t>
no. of events in flat bins
N/N_tot per bin
eleMom
eleEta
elePhi
eleSip2dSig
eleSip3dSig
elePtRel
eleEtaRel

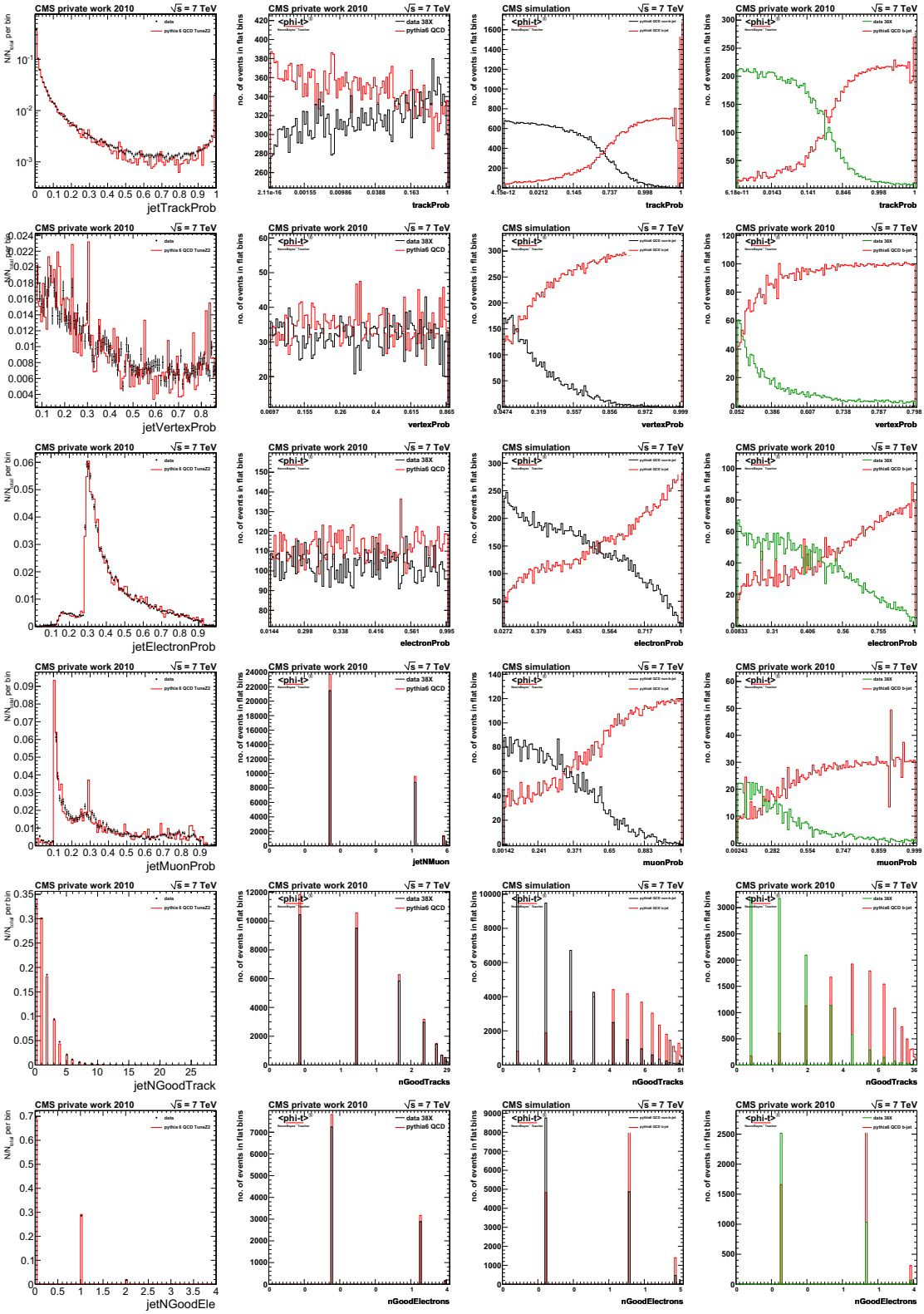

CMS private work 2010
√s = 7 TeV
CMS simulation
data
pythia 6 QCD TuneZ2
data 38X
pythia6 QCD
pythia6 QCD b-jet
<phi-t>
N/N_total per bin
no. of events in flat bins
eleJetDeltaR
eleJetPpar
eleJetPparFrac
eleChi2
eleId
eleZpos
eleInvDeltaR


CMS private work 2010
√s = 7 TeV
CMS simulation
eleGSFDif
eleBrem


CMS private work 2010
√s = 7 TeV
p_T(jet)
y(jet)
CSV
CSVMVA
JBP
SSVE
SSVP
SETIP
jetPt
jetEta
jetCSV
jetCSVMVA
jetBProb
jetSSV
jetSSVP
jetElectronIP


CMS private work 2010
√s = 7 TeV
SETPt
jetElectronPt
SMTIP
jetMuonIP
SMTPt
jetMuonPt
TCHE
jetTCHE
TCHP
jetTCHP


CMS private work 2010
CMS simulation
√s = 7 TeV
jetNTrack
jetNSV
jetNMuon
jetNEle


CMS private work 2010
CMS simulation
√s = 7 TeV
data
pythia 6 QCD TuneZ2
data 38X
pythia6 QCD
pythia6 QCD b-jet
N/N_tot per bin
no. of events in flat bins
jetTrackProb
trackProb
jetVertexProb
vertexProb
jetElectronProb
electronProb
jetMuonProb
jetNMuon
muonProb
jetNGoodTrack
nGoodTracks
jetNGoodEle
nGoodElectrons

Appendix B

Results of data to Monte Carlo comparison

The figures show the output distributions of the NeuroBayes experts, which were calibrated to compare physics objects from the detector (black) and simulated objects, created with Pythia 6 QCD Tune2Z event generation [SMS06] (red). The number of QCD events and Monte Carlo events are in the same order, so the a priori fraction is around 0.5. It is a good indication that the separation of the two classes is small. Nevertheless events with small values of the NeuroBayes output variable represent a kind of event which are underestimated in simulation. Events in the region around 0.5 are well simulated and events with larger values of the NeuroBayes output variable are overestimated in the simulation. The width of the output distribution is a measurement of the MC quality. If it is too broad, it is necessary to look into more detail of the related input variables. For each physics objects: tracks, secondary vertices, electron candidates, muons and jets the comparison was done in the six different trigger regions. The following plots show their output distributions of the NeuroBayes experts.

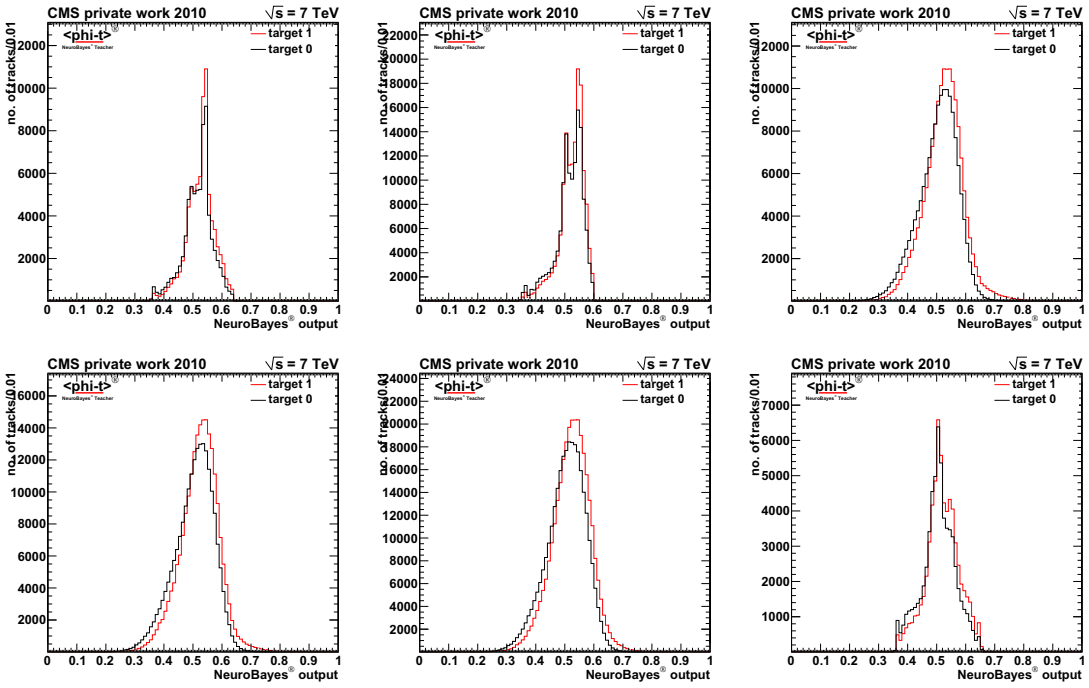


Figure B.1: Result of the comparison of data and MC for track objects

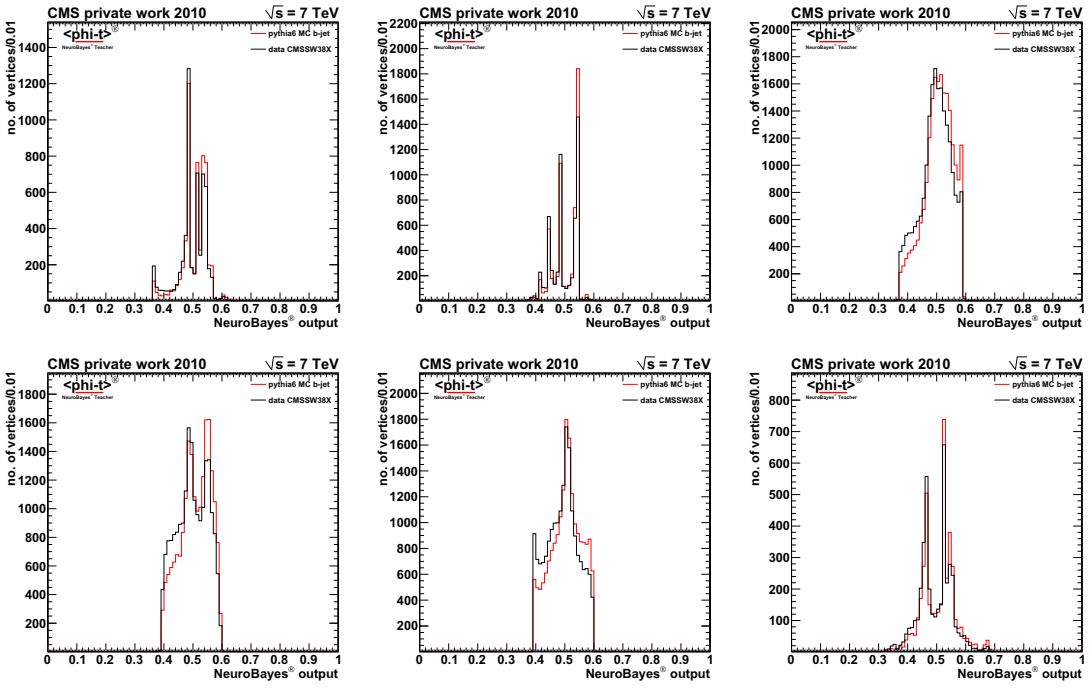


Figure B.2: Result of the comparison of data and MC for vertex objects

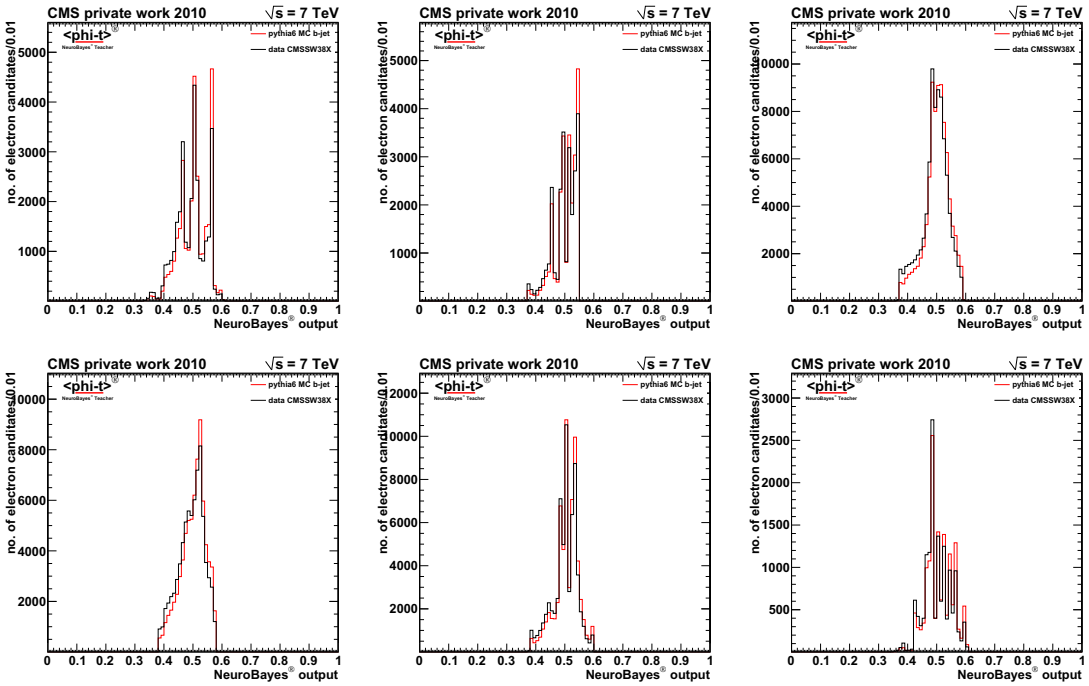


Figure B.3: Result of the comparison of data and MC for electron candidate objects

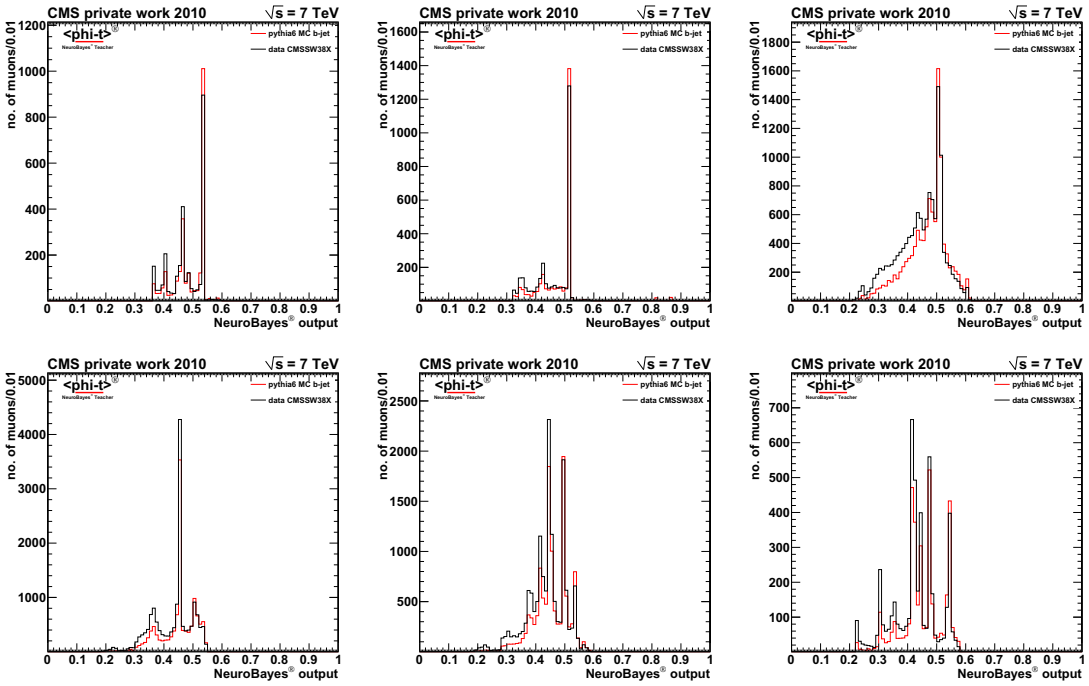


Figure B.4: Result of the comparison of data and MC for muon objects

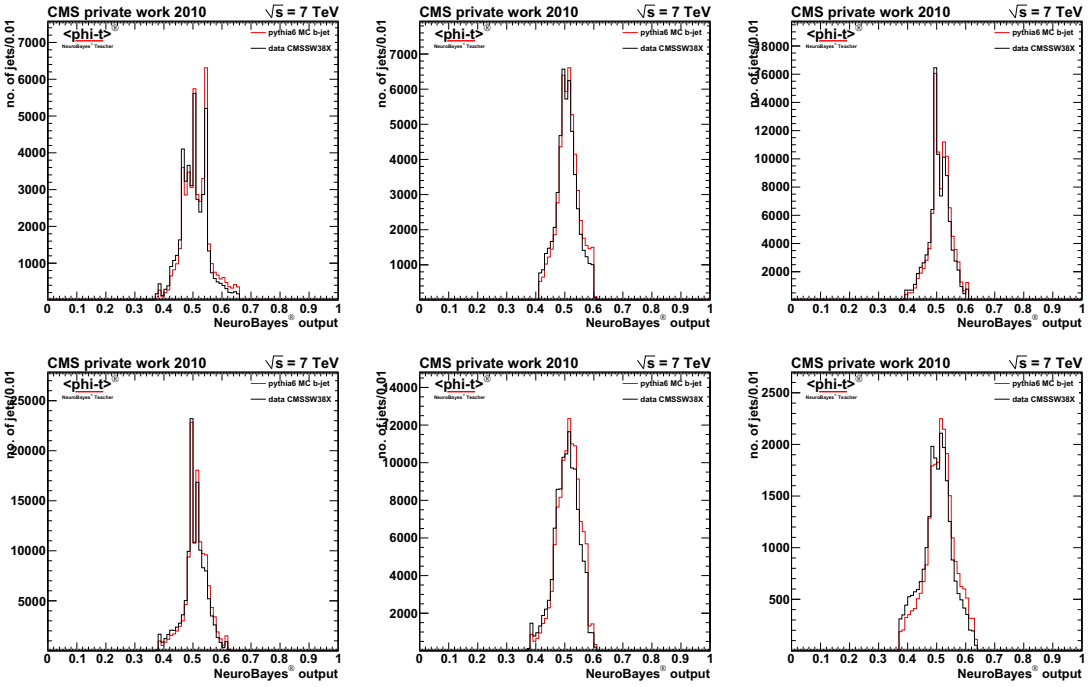


Figure B.5: Result of the comparison of data and MC for jet objects

Appendix C

Dependency check

In section 5.3.4 the differences between the data and the MC tagger were discussed. To visualize the dependencies of the two approaches the following equation was deduced:

$$o_{t2} = \frac{o_{t1}\, f_1/f_2}{1 - P(S) + o_{t1}(P(S)(f_1 + 1) + f_1/f_2 - 1)}$$

To check the correct implementation the framework was tested. First on MC two NeuroBayes calibrations were trained with different target 0 samples. Once with background simulations and once with complete data simulations. The expected $P(S) = 0.036$. The result is plotted in a scatter plot (figure C.1).
For the second test two NeuroBayes calibrations were trained on the same target but with different sample fraction. The result is also plotted to each other in a scatter plot. Here the expected $P(S) = 0$. Figure C.1 confirms that the determined dependency is correct.

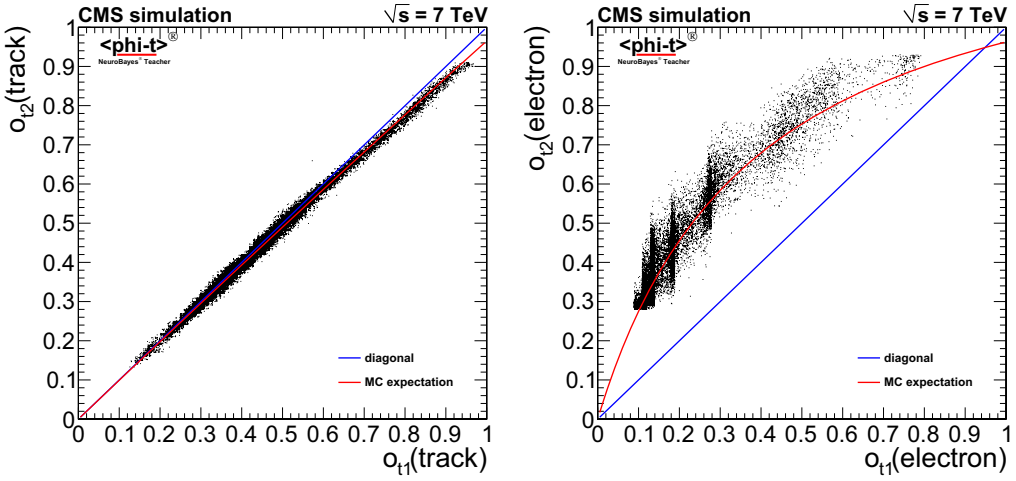
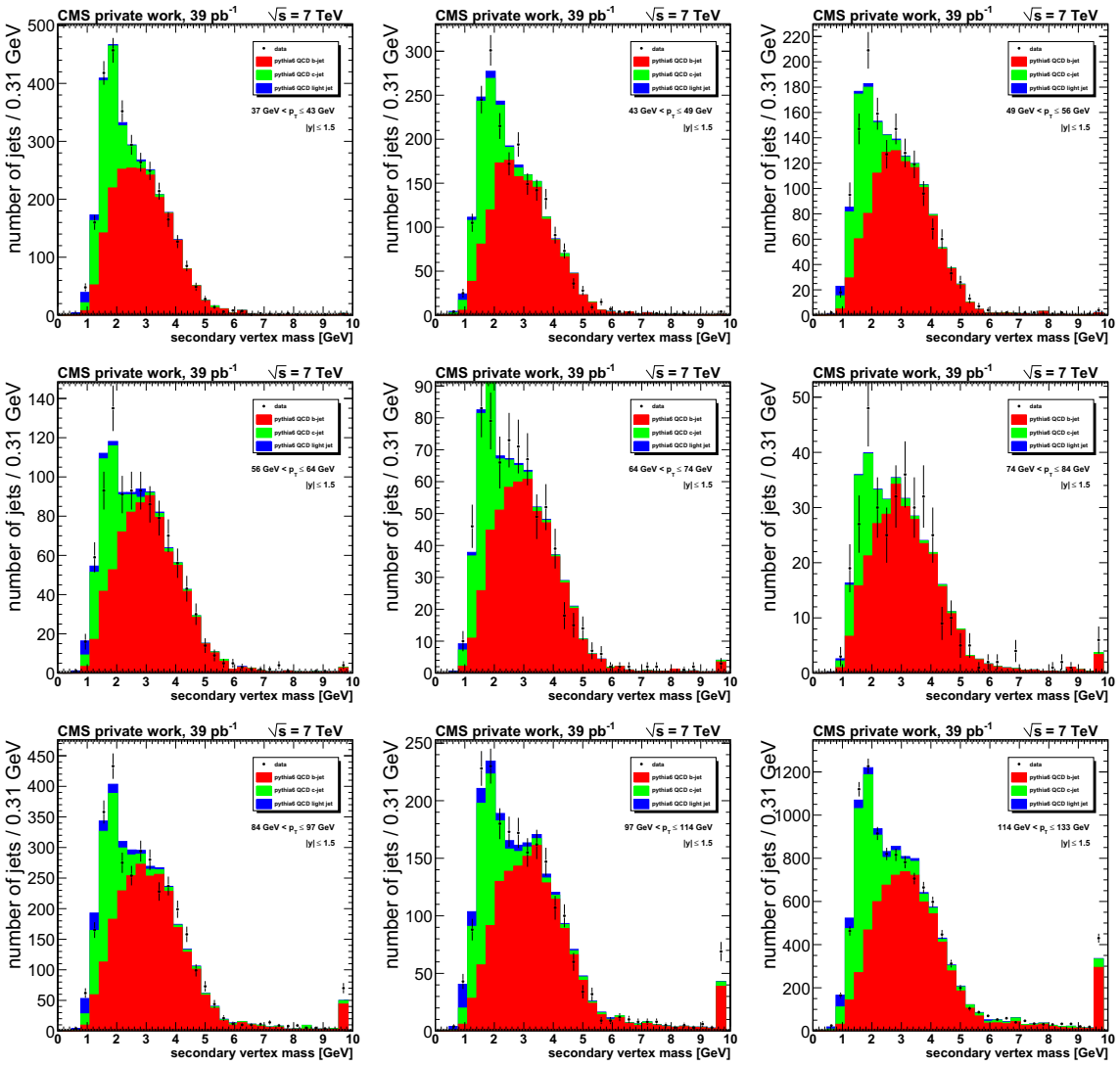


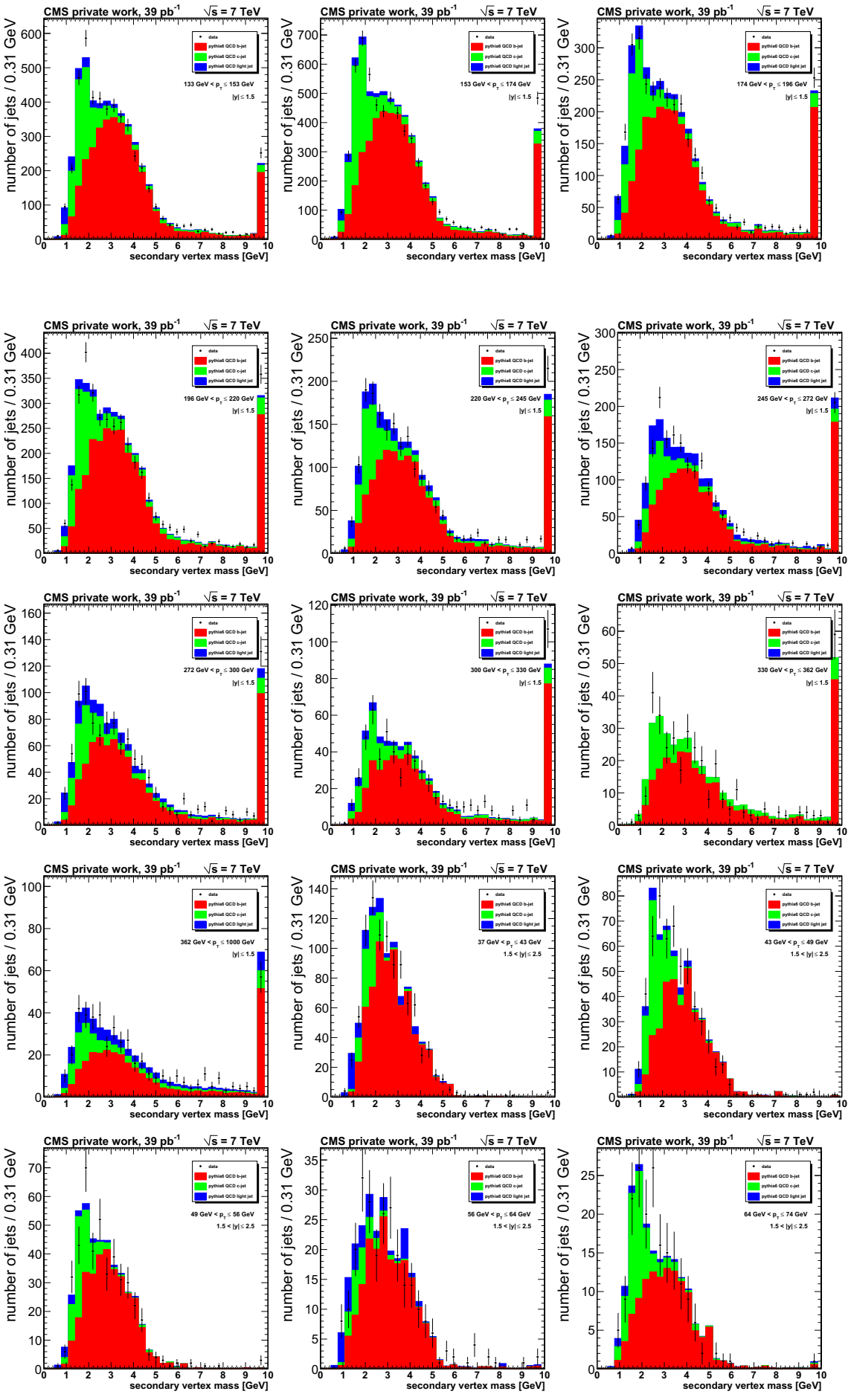
Figure C.1: Dependency check for two NeuroBayes calibrations trained on different scenarios. Left: signal events where added to the target 0 sample. On the right only the fraction of two samples changes. The points follow the red line which confirms the determined dependency equation.

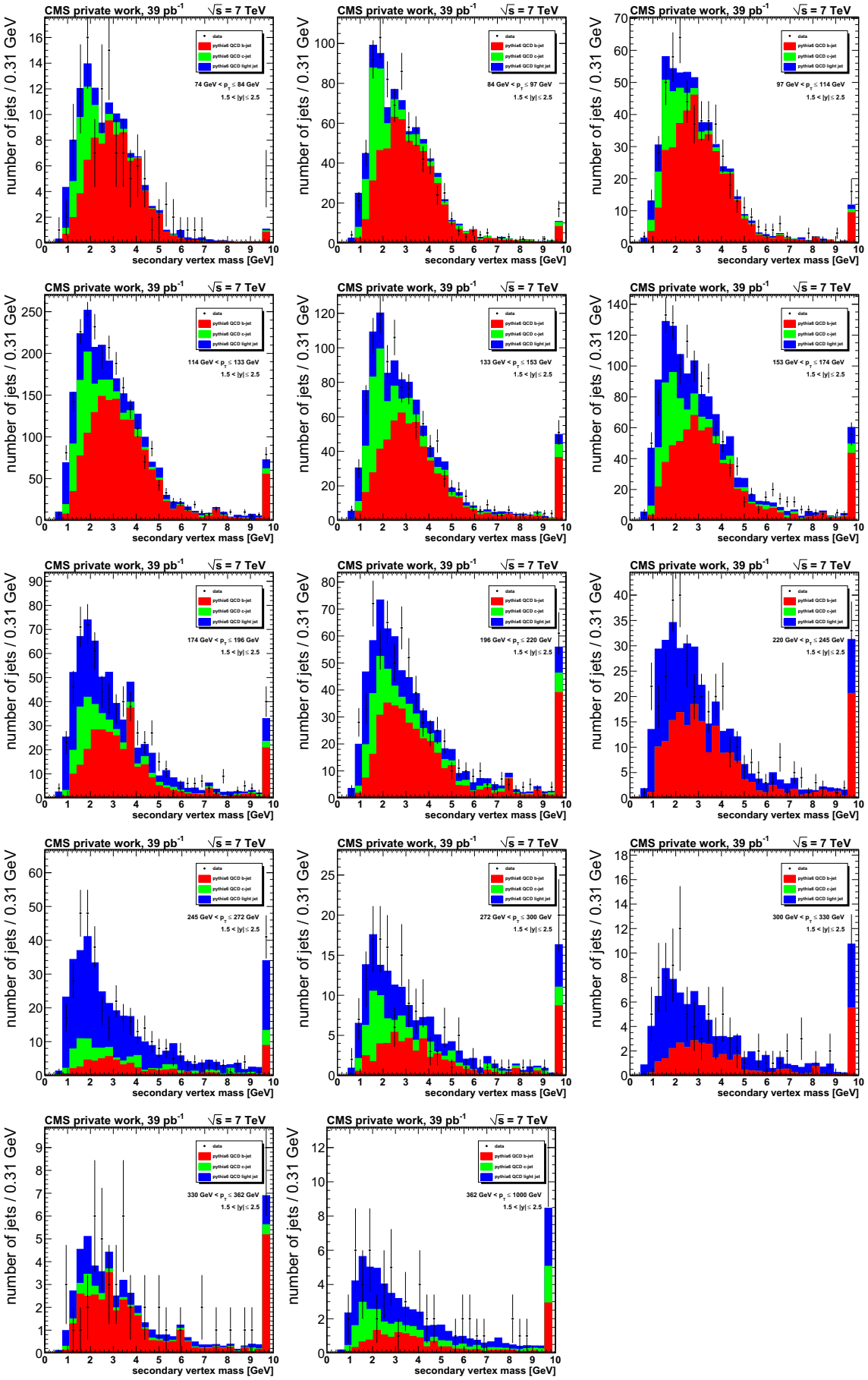
Appendix D

Fit histograms of flavour content fitter

The following histograms show the fits of the flavour content fitter. Each plot corresponds to a region in which the procedure was performed. It starts with the low p_T in the barrel region of the detector. After that the results for the forward regions are shown.



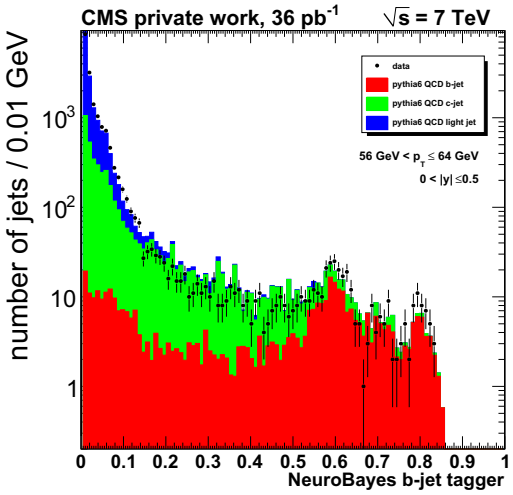

CMS private work, 39 pb-1 √s = 7 TeV
number of jets / 0.31 GeV
secondary vertex mass [GeV]
data
pythia6 QCD b-jet
pythia6 QCD c-jet
pythia6 QCD light jet
133 GeV < pT ≤ 153 GeV
153 GeV < pT ≤ 174 GeV
174 GeV < pT ≤ 196 GeV
196 GeV < pT ≤ 220 GeV
220 GeV < pT ≤ 245 GeV
245 GeV < pT ≤ 272 GeV
272 GeV < pT ≤ 300 GeV
300 GeV < pT ≤ 330 GeV
330 GeV < pT ≤ 362 GeV
362 GeV < pT ≤ 1000 GeV
37 GeV < pT ≤ 43 GeV
43 GeV < pT ≤ 49 GeV
49 GeV < pT ≤ 56 GeV
56 GeV < pT ≤ 64 GeV
64 GeV < pT ≤ 74 GeV
|y| ≤ 1.5
1.5 < |y| ≤ 2.5

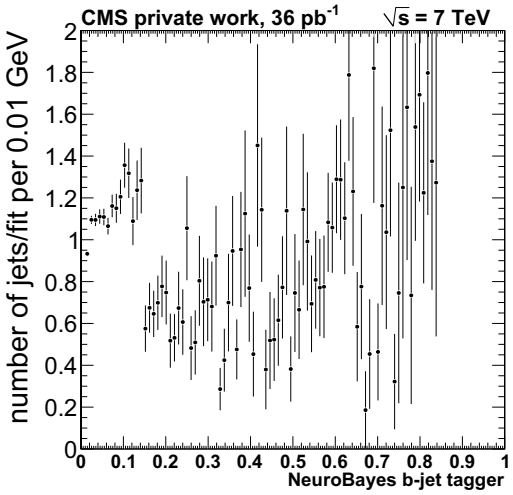

CMS private work, 39 pb-1 √s = 7 TeV
number of jets / 0.31 GeV
secondary vertex mass [GeV]
data
pythia6 QCD b-jet
pythia6 QCD c-jet
pythia6 QCD light jet
74 GeV < pT ≤ 84 GeV
84 GeV < pT ≤ 97 GeV
97 GeV < pT ≤ 114 GeV
114 GeV < pT ≤ 133 GeV
133 GeV < pT ≤ 153 GeV
153 GeV < pT ≤ 174 GeV
174 GeV < pT ≤ 196 GeV
196 GeV < pT ≤ 220 GeV
220 GeV < pT ≤ 245 GeV
245 GeV < pT ≤ 272 GeV
272 GeV < pT ≤ 300 GeV
300 GeV < pT ≤ 330 GeV
330 GeV < pT ≤ 362 GeV
362 GeV < pT ≤ 1000 GeV
1.5 < |y| ≤ 2.5

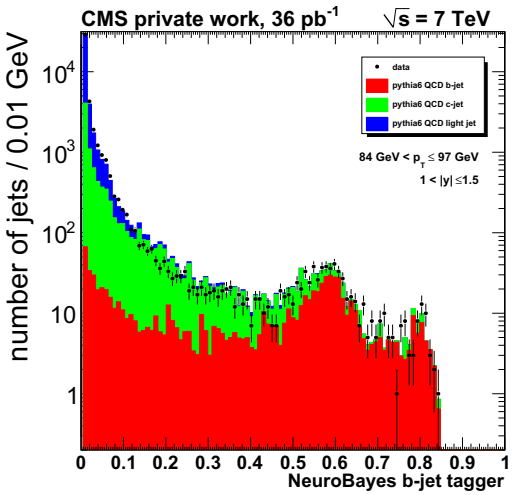
Appendix E

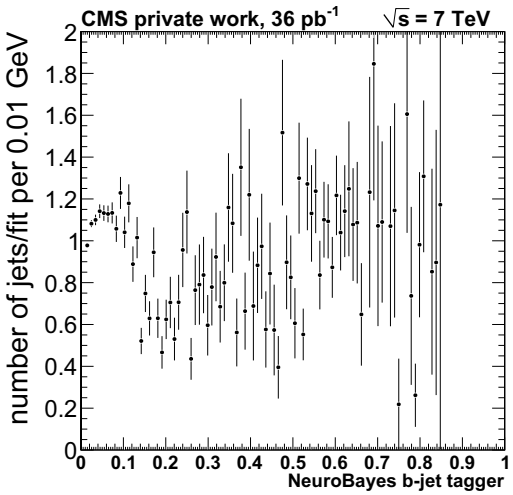
Fit histograms of NB flavour content fitter

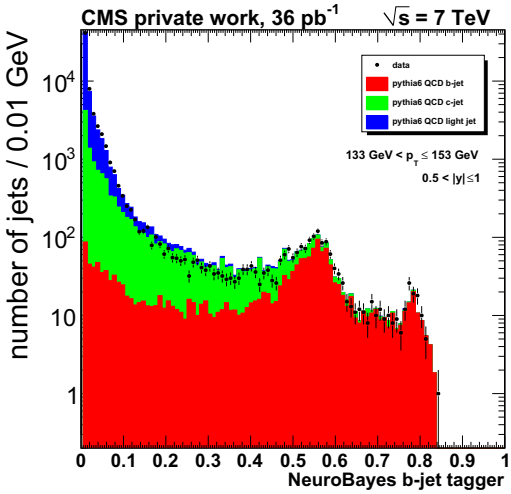
The following histograms show some of the fits of the flavour content fitter. Each plot corresponds to a region in which the procedure was performed. The bins are chosen in a way that it is possible to see the effect of the insufficent distributions in the whole p_T/y phase space.

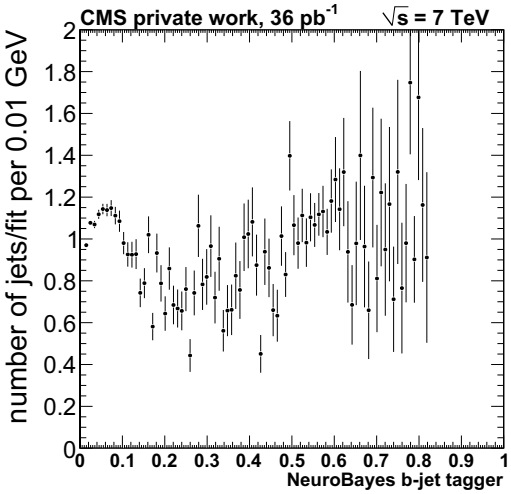


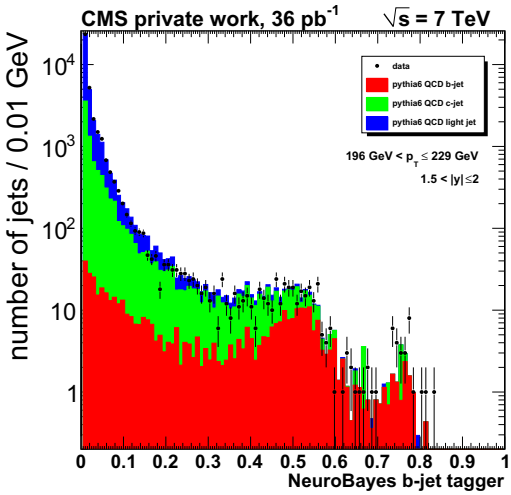


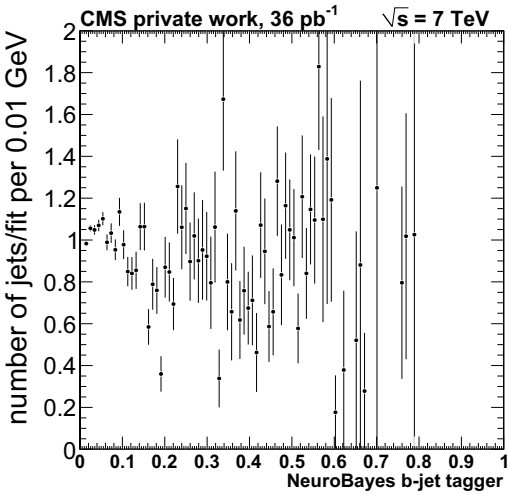


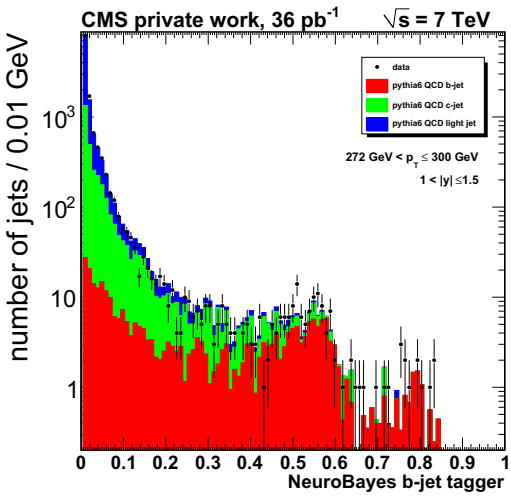


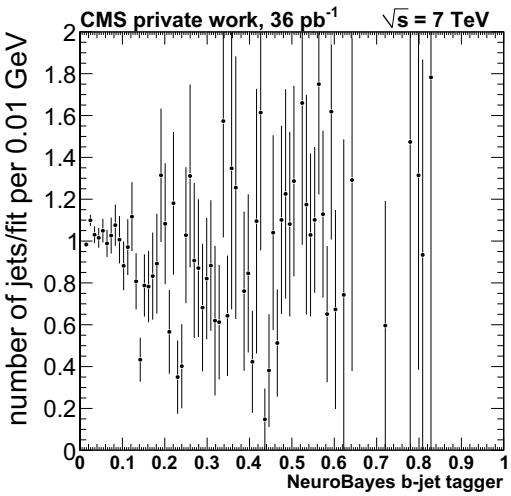

CMS private work, 36 pb-1 √s = 7 TeV
data
pythia6 QCD b-jet
pythia6 QCD c-jet
pythia6 QCD light jet
133 GeV < pT ≤ 153 GeV
0.5 < |y| ≤1
number of jets / 0.01 GeV
NeuroBayes b-jet tagger

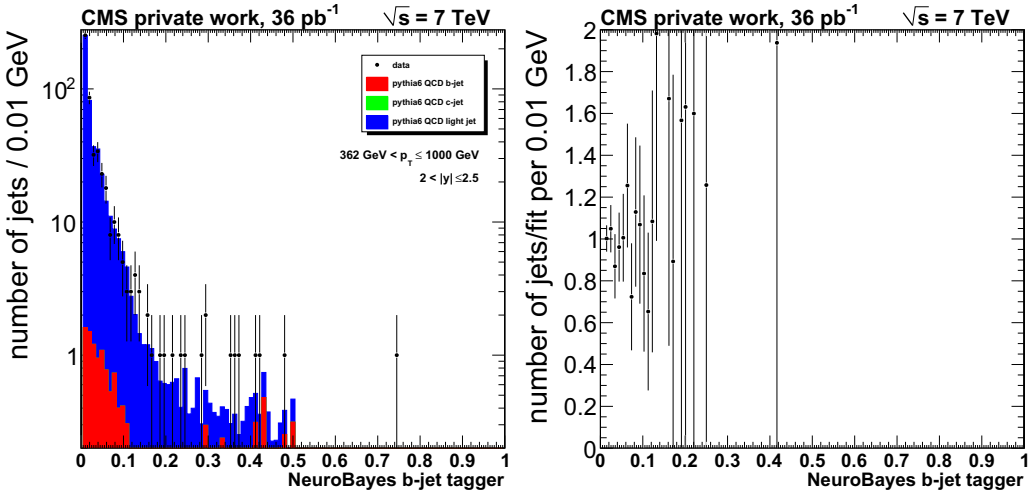

CMS private work, 36 pb-1 √s = 7 TeV
number of jets/fit per 0.01 GeV
NeuroBayes b-jet tagger


CMS private work, 36 pb-1 √s = 7 TeV
data
pythia6 QCD b-jet
pythia6 QCD c-jet
pythia6 QCD light jet
196 GeV < pT ≤ 229 GeV
1.5 < |y| ≤2
number of jets / 0.01 GeV
NeuroBayes b-jet tagger


CMS private work, 36 pb-1 √s = 7 TeV
number of jets/fit per 0.01 GeV
NeuroBayes b-jet tagger


CMS private work, 36 pb-1 √s = 7 TeV
data
pythia6 QCD b-jet
pythia6 QCD c-jet
pythia6 QCD light jet
272 GeV < pT ≤ 300 GeV
1 < |y| ≤1.5
number of jets / 0.01 GeV
NeuroBayes b-jet tagger


CMS private work, 36 pb-1 √s = 7 TeV
number of jets/fit per 0.01 GeV
NeuroBayes b-jet tagger


CMS private work, 36 pb-1
√s = 7 TeV
data
pythia6 QCD b-jet
pythia6 QCD c-jet
pythia6 QCD light jet
362 GeV < pT ≤ 1000 GeV
2 < |y| ≤2.5
number of jets / 0.01 GeV
NeuroBayes b-jet tagger
CMS private work, 36 pb-1
√s = 7 TeV
number of jets/fit per 0.01 GeV
NeuroBayes b-jet tagger

List of Figures

Bibliography

[A+87] C. Albajar et al. Beauty production at the CERN proton-antiproton collider. *Physics Letters B*, 186(2):237–246, 1987.

[A+91] C. Albajar et al. Beauty production at the CERN pp collider. *Physics Letters B*, 256(1):121–128, 1991.

[A+92] F. Abe et al. Measurement of the B-meson and b-quark cross sections at $\sqrt{s}$ =1.8 TeV using the exclusive decay $B^{\pm} \to J/\psi K^{\pm}$. *Phys. Rev. Lett.*, 68(23):3403–3407, Jun 1992.

[A+93a] F. Abe et al. Measurement of bottom quark production in 1.8 TeV $p\bar{p}$ collisions using muons from b-quark decays. *Phys. Rev. Lett.*, 71(15):2396–2400, Oct 1993.

[A+93b] F. Abe et al. Measurement of the bottom quark production cross section using semileptonic decay electrons in $p\bar{p}$ collisions at $\sqrt{s}$ =1.8 TeV. *Phys. Rev. Lett.*, 71(4):500–504, Jul 1993.

[A+94] F. Abe et al. Measurement of the B meson and b quark cross sections at $\sqrt{s}$ =1.8 TeV using the exclusive decay $B^0 \to J/\psi K^*(892)^0$. *Phys. Rev. D*, 50(7):4252–4257, Oct 1994.

[A+95a] S. Abachi et al. Inclusive μ and b-Quark Production Cross Sections in $p\bar{p}$ Collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 74(18):3548–3552, May 1995.

[A+95b] F. Abe et al. Measurement of the B Meson Differential Cross Section $d\sigma/dp\_T$ in $p\bar{p}$ Collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 75(8):1451–1455, Aug 1995.

[A+99] C. Adloff et al. Measurement of open beauty production at hera. *Physics Letters B*, 467(1-2):156–164, 1999.

[A+00a] B. Abbott et al. Small-Angle Muon and Bottom-Quark Production in $p\bar{p}$ Collisions at $\sqrt{s} = 1.8TeV$. *Phys. Rev. Lett.*, 84(24):5478–5483, Jun 2000.

[A+00b] B. Abbott et al. The b-bbar Production Cross Section and Angular Correlations in p-pbar Collisions at sqrt(s) = 1.8 TeV. *Physics Letters B*, 487(3-4):264–272, 2000.

[A+01] M. Acciarri et al. Measurements of the cross-sections for open charm and beauty production in $\gamma\gamma$ collisions at $\sqrt{s}$ = 189-GeV to 202-GeV. *Phys. Lett.*, B503:10–20, 2001.

[A+02a] D. Acosta et al. Measurement of the B^+ total cross section and B^+ differential cross section $d\sigma/dp\_T$ in $p\bar{p}$ collisions at $\sqrt{s} = 1.8\,TeV$. *Phys. Rev. D*, 65(5):052005, Feb 2002.

[A+02b] D. Acosta et al. Measurement of the ratio of b quark production cross sections in $p\bar{p}$ collisions at $\sqrt{s} = 630\ GeV$ and $\sqrt{s} = 1800\ GeV$. *Phys. Rev. D*, 66(3):032002, Aug 2002.

[A+05] P. Achard et al. Measurement of the cross section for open-beauty production in photon-photon collisions at LEP. *Phys. Lett.*, B619:71–81, 2005.

[A+09] L. Agostino et al. Commissioning of the CMS High Level Trigger. *J. Instrum.*, 4(CMS-NOTE-2009-012):P10005. 14 p, Jun 2009.

[A+10] R. Aaij et al. Measurement of sigma($pp \to b\bar{b}X$) at $\sqrt{(}s)$=7 TeV in the forward region. *Phys. Lett.*, B694:209–216, 2010.

[AMST06] Wolfgang Adam, Boris Mangano, Thomas Speer, and Teddy Todorov. Track Reconstruction in the CMS tracker. Technical Report CMS-NOTE-2006-041. CERN-CMS-NOTE-2006-041, CERN, Geneva, Dec 2006.

[And33] C. D. Anderson. THE POSITIVE ELECTRON. *Phys. Rev.*, 43:491–494, 1933.

[B+00] G L Bayatyan et al. *CMS TriDAS project: Technical Design Report; 1, the trigger systems*. Number CERN-LHCC-2000-038 in Technical Design Report CMS. 2000.

[B+01] J. Breitweg et al. Measurement of open beauty production in photoproduction at HERA. *Eur. Phys. J.*, C18:625–637, 2001.

[BBB+06] J. Baines, S.P. Baranov, O. Behnke, J. Bracinik, M. Cacciari, et al. Heavy quarks (Working Group 3): Summary Report for the HERA-LHC Workshop Proceedings. 2006.

[BBK71] S. M. Berman, J. D. Bjorken, and John B. Kogut. Inclusive Processes at High Transverse Momentum. *Phys. Rev.*, D4:3388, 1971.

[BCF+07] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois. Electron reconstruction in CMS. *The European Physical Journal C-Particles and Fields*, 49(4):1099–1116, 2007.

[Ber01] E.L. Berger. Supersymmetry explanation for the puzzling bottom quark production cross section. *Arxiv preprint hep-ph/0112062*, 2001.

[BL98] V. Blobel and E. Lohrmann. *Statistische und numerische Methoden der Datenanalyse*. Teubner Verlag, 1 edition, 1998.

[BRO70] C. G. BROYDEN. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[C+00] Akos Csilling et al. Charm and bottom production in two-photon collisions with OPAL. 2000.

[C+04] S. Chekanov et al. Bottom photoproduction measured using decays into muons in dijet events in e p collisions at s**(1/2) = 318-GeV. *Phys. Rev.*, D70:012008, 2004.

[Cac04] Matteo Cacciari. Rise and fall of the bottom quark production excess. 2004.

[CDD+03] B. Clement, Bloch D., Gele D., Greder S., and Ripp-Baudot I. SystemD or how to get signal, backgrounds and their efficiencies with real data only. *D0 Note*, 4159, June 2003.

[CDF05] CDF Collaboration. Inclusive b-jet production. *CDF NOTE*, (8418), September 2005.

[CDF07] CDF Collaboration. b-bbar dijet production using svt. *CDF NOTE*, (8939), April 2007.

[CER08a] CERN Public Web Pages. CERN in a nutshell, ff., 2008.

[CER08b] CERN Public Web Pages. CERN - The Large Hadron Collider, ff., 2008.

[CKKT06] Susanna Cucciarelli, Marcin Konecki, Danek Kotlinski, and Teddy Todorov. Track reconstruction, primary vertex finding and seed generation with the Pixel Detector. Technical Report CMS-NOTE-2006-026. CERN-CMS-NOTE-2006-026, CERN, Geneva, Jan 2006.

[CMS06] CMS Collaboration. CMS physics: Technical design report. *Volume II: Physics Performance, CERN/LHCC*, 21(CERN-LHCC-2006-001 ; CMS-TDR-008-1):2006, 2006.

[CMS07a] CMS Collaboration. Evaluation of udsg Mistags for b-tagging using Negative Tags. *CMS PAS*, BTV-07-002, 2007.

[CMS07b] CMS Collaboration. Performance Measurement of b tagging Algorithms Using Data containing Muons within Jets0. *CMS PAS*, BTV-07-001, 2007.

[CMS08a] CMS Collaboration. Plans for Jet Energy Corrections at CMS. *CMS PAS*, JME-07-002, Jul 2008.

[CMS08b] CMS Public Web Pages. CMS - Detector, ff., 2008.

[CMS09a] CMS Collaboration. Algorithms for b Jet identification in CMS. *CMS PAS*, BTV-09-001, Jul 2009.

[CMS09b] CMS Collaboration. Track reconstruction in the CMS Tracker. *CMS PAS*, TRK-09-001, 2009.

[CMS10a] CMS Collaboration. Commissioning of b-jet identification with pp collisions at $\sqrt{s} =$ 7 tev. *CMS PAS*, BTV-10-001, 2010.

[CMS10b] CMS Collaboration. Commissioning of the Particle-Flow Event Reconstruction with the First LHC collisions recorded in the CMS detector. *CMS PAS*, PFT-10-001, 2010.

[CMS10c] CMS Collaboration. Commissioning of the Particle-Flow Reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV. *CMS PAS*, PFT-10-002, 2010.

[CMS10d] CMS Collaboration. Electron reconstruction and identification at sqrt(s) = 7 TeV. *CMS PAS*, EGM-10-004, 2010.

[CMS10e] CMS Collaboration. Inclusive b-jet production in pp collisions at sqrt(s)=7 TeV. *CMS PAS*, BPH-10-009, 2010. CMS PAS BPH-10-009.

[CMS10f] CMS Collaboration. Jet Performance in pp Collisions at 7 TeV. *CMS PAS*, JME-10-003, 2010.

[CMS10g] CMS Collaboration. Measurement of cms luminosity. *CMS PAS*, EWK-10-004, 2010.

[CMS10h] CMS Collaboration. Measurement of the Inclusive Jet Cross Section in pp Collisions at 7 TeV using the CMS Detector. *CMS PAS*, QCD-10-011, 2010.

[CMS10i] CMS Collaboration. Measurement of the Underlying Event Activity at the LHC with $\sqrt{s}$=7TeV. *CMS-PAS*, QCD-10-010, 2010.

[CMS10j] CMS Collaboration. Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV. *CMS PAS*, PFT-10-003, 2010.

[CMS10k] CMS Collaboration. Performance of muon identification in pp collisions at s**0.5 = 7 TeV. *CMS PAS*, MUO-10-002, 2010.

[CMS10l] CMS Collaboration. Tracking and Primary Vertex Results in First 7 TeV Collisions. *CMS PAS*, TRK-10-005, 2010.

[CMS10m] CMS Collaboration. Tracking and Vertexing Results from First Collisions. *CMS PAS*, TRK-10-001, 2010.

[CMS11] CMS Collaboration. Measurement of Drell-Yan Cross Section (dsigma/dM). CMS-PAS-EWK-10-007, 2011.

[CRS02] Sergio Cittolin, Attila Rácz, and Paris Sphicas. *CMS trigger and data-acquisition project: Technical Design Report*. Number CERN-LHCC-2002-026 in Technical Design Report CMS. CERN, Geneva, 2002.

[CSS08] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *JHEP*, 04:063, 2008.

[D+08] Julien Donini et al. Energy Calibration of b Quark Jets with $Z \to b\bar{b}$ Decays at the Tevatron Collider. *Nucl. Instrum. Meth.*, A596:354–367, 2008.

[Dys49] F. J. Dyson. The Radiation theories of Tomonaga, Schwinger, and Feynman. *Phys. Rev.*, 75:486–502, 1949.

[Fan07] Livio Fano. Multiple parton interactions, underlying event and forward physics at lhc. Technical Report CMS-CR-2007-064. CERN-CMS-CR-2007-064, CERN, Geneva, Sep 2007.

[Fei04] Michael Feindt. A Neural Bayesian Estimator for Conditional Probability Densities. *ArXiv Physics e-prints*, physics/0402093, February 2004.

[FFF78] R. P. Feynman, R. D. Field, and G. C. Fox. Quantum-chromodynamic approach for the large-transverse- momentum production of particles and jets. *Phys. Rev.*, D18:3320, 1978.

[FNW03] Stefano Frixione, Paolo Nason, and Bryan R. Webber. Matching NLO QCD and parton showers in heavy flavour production. *JHEP*, 08:007, 2003.

[FW02] Stefano Frixione and Bryan R. Webber. Matching NLO QCD computations and parton shower simulations. *JHEP*, 06:029, 2002.

[FWV07] R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. Adaptive Vertex Fitting. Technical Report CMS-NOTE-2007-008. CERN-CMS-NOTE-2007-008, CERN, Geneva, Mar 2007.

[HHL+77] SW Herb, DC Hom, LM Lederman, JC Sens, HD Snyder, JK Yoh, JA Appel, BC Brown, CN Brown, WR Innes, et al. Observation of a dimuon resonance at 9.5 GeV in 400-GeV proton-nucleus collisions. *Physical Review Letters*, 39(5):252–255, 1977.

[Ind04] Andre S. Indenhuck. Das Standardmodell der Teilchenphysik. *web.physik.rwth-aachen.de*, February 2004.

[Jun03] H. Jung. kt-factorization and CCFM-the solution for describing the hadronic final states-everywhere? *Arxiv preprint hep-ph/0311249*, 2003.

[K+11a] Vardan Khachatryan et al. Measurement of the B+ Production Cross Section in pp Collisions at sqrt(s) = 7 TeV. 2011.

[K+11b] Vardan Khachatryan et al. Inclusive b-hadron production cross section with muons in pp collisions at sqrt(s) = 7 TeV. 2011.

[K+11c] Vardan Khachatryan et al. Measurement of B anti-B Angular Correlations based on Secondary Vertex Reconstruction at sqrt(s)=7 TeV. 2011.

[KRW06] T. Kluge, K. Rabbertz, and M. Wobisch. Fast pQCD calculations for PDF fits. In *14th International Workshop on Deep Inelastic Scattering (DIS 2006), 20-24 Apr 2006*, page 483, Tsukuba, Japan, April 2006.

[M+92] G. Marchesini et al. HERWIG: A Monte Carlo event generator for simulating hadron emission reactions with interfering gluons. Version 5.1 - April 1991. *Comput. Phys. Commun.*, 67:465–508, 1992.

[Mar09] D. Martschei. Developement of a soft electron based b-jet tagger for the CMS experiment - Entwicklung eines auf Elektronen basierenden B-Jet-Taggers für das CMS-Experiment. Master's thesis, Universität Karlsruhe (TH), 2009. IEKP-KA/2009-6.

[Mar10] Daniel Martschei. Different benchmarks for MVA methods - Comparison of methods from TMVA with NeuroBayes, December 2010.

[Mor06] J. Morlock. Optimization of the decay time resolution of semileptonic B_S decays using artificial neural networks. Master's thesis, Universität Karlsruhe (TH), 2006. IEKP-KA/2007-6.

[MPQW06] Thomas Müller, Christian Piasecki, Gunter Quast, and Christian Weiser. Inclusive Secondary Vertex Reconstruction in Jets. Technical Report CMS-NOTE-2006-027. CERN-CMS-NOTE-2006-027, CERN, Geneva, Jan 2006.

[N+10] K Nakamura et al. Review of particle physics. *J. Phys.*, G37:075021, 2010.

[Nag02] Zoltan Nagy. Three-jet cross sections in hadron hadron collisions at next-to-leading order. *Phys. Rev. Lett.*, 88:122003, 2002.

[NDE88] P. Nason, S. Dawson, and R.Keith Ellis. The Total Cross-Section for the Production of Heavy Quarks in Hadronic Collisions. *Nucl.Phys.*, B303:607, 1988.

[NDE89] P. Nason, S. Dawson, and R.Keith Ellis. The One Particle Inclusive Differential Cross-Section for Heavy Quark Production in Hadronic Collisions. *Nucl.Phys.*, B327:49–92, 1989.

[Nob] Nobelprize.org. "the nobel prize in physics 1969".

[NP33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:pp. 289–337, 1933.

[P+02] J. Pumplin et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 07:012, 2002.

[PL05] M. Pivk and F. R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nuclear Instruments and Methods in Physics Research A*, 555:356–369, December 2005.

[PRS61] E. Pickup, D. K. Robinson, and E. O. Salant. pi-pi Resonance in pi–p Interactions at 1.25 Bev. *Phys. Rev. Lett.*, 7:192–195, 1961.

[PT10] Phi-T. The NeuroBayes User's Guide. April 2010.

[RC56] Frederick Reines and Clyde L. Cowan. The neutrino. *Nature*, 178:446–449, 1956.

[RHW87] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation. In D. E. Rumelhart, J. L. McClelland, et al., editors, *Parallel Distributed Processing: Volume 1: Foundations*, pages 318–362. MIT Press, Cambridge, 1987.

[Ros58] F Rosenblatt. The perceptron: Aprobabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[RPS06] Andrea Rizzi, Fabrizio Palla, and Gabriele Segneri. Track impact parameter based b-tagging with CMS. Technical Report CMS-NOTE-2006-019. CERN-CMS-NOTE-2006-019, CERN, Geneva, Jan 2006.

[SAF+06] T. Speer, W. Adam, R. Frühwirth, A. Strandlie, T. Todorov, and M. Winkler. Track reconstruction in the CMS tracker. *Nuclear Instruments and Methods in Physics Research A*, 559:143–147, April 2006.

[Sch08] A. Scheurer. *Algorithms for the Identification of b-Quark Jets with First Data at CMS.* PhD thesis, Universität Karlsruhe (TH), 2008. IEKP-KA/2008-19.

[Sil04] W. Da Silva. Measurement of the open beauty and charm production cross sections in two photon collisions with delphi. *Nuclear Physics B - Proceedings Supplements*, 126:185–190, 2004. Proceedings of the International Conference on the Structure and Interactions of the Photon, Including the 15th International Workshop on Photon-Photon Collisions.

[SMS06] T. Sjostrand, S. Mrenna, and P. Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006.

[SPF+06] Thomas Speer, Kirill Prokofiev, R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. Vertex fitting in the cms tracker. Technical Report CMS-NOTE-2006-032. CERN-CMS-NOTE-2006-032, CERN, Geneva, Feb 2006.

[SPS50] J. Steinberger, W. K. H. Panofsky, and J. Steller. EVIDENCE FOR THE PRODUCTION OF NEUTRAL MESONS BY PHOTONS. *Phys. Rev.*, 78:802–805, 1950.

[SS37] J. C. Street and E. C. Stevenson. NEW EVIDENCE FOR THE EXISTENCE OF A PARTICLE OF MASS INTERMEDIATE BETWEEN THE PROTON AND ELECTRON. *Phys. Rev.*, 52:1003–1004, 1937.

[Tho97] J. J. Thomson. Cathode rays. *Phil. Mag.*, 44:293–316, 1897.

[VPB07] Tejinder Virdee, Achille Petrilli, and Austin Ball. Cms high level trigger. Technical Report LHCC-G-134. CERN-LHCC-2007-021, CERN, Geneva, Jun 2007. revised version submitted on 2007-10-19 16:57:09.

Danksagung

Zuletzt möchte ich noch all denen danken, ohne die die Anfertigung dieser Arbeit nicht möglich gewesen wäre.
Dazu muss man wissen, dass vor drei Jahren die Arbeitsgruppe von Michael Feindt ausschließlich Daten analysierte, die am CDF-Experiment des Tevatrons gesammelt wurden. Schwerpunkt dieser Studien waren und sind die Spektroskopie-Messungen von b-Hadronen und sogenannte b-flavour-Physik. Ich selbst habe dort auch schon im Rahmen meiner Diplomarbeit orbital angeregte Zustände des B_d Mesons entdeckt. Der Beginn meiner Doktorarbeit war genau in der Zeit, als die letzten Vorbereitungen für die Inbetriebnahme des LHC durchgeführt wurden. Noch innerhalb eines Jahres sollte dieses für mich sehr faszinierende Projekt starten.
Michael Feindt und Günter Quast ermöglichten es mir, in der CMS Kollaboration mitzuwirken. Es wurde eine neue Arbeitsgruppe am Institut für experimentelle Kernphysik (ekp) unter der Regie von Michael Feindt gegründet, deren Aufgabe es sein sollte, den hier vorgestellten NeuroBayes b-jet tagger zu entwickeln und wenn möglich zu etablieren. Die Gruppe umfasste neben mir noch Daniel Martschei.
Für die Anfangszeit sind vor allem Armin Scheurer und Christophe Saout zu erwähnen, die uns immer hilfsbereit zur Seite standen und es ermöglichten, uns in der komplexen Welt der CMS Software zurechtzufinden.
Finanziert wurde ich zu dieser Zeit durch ein Stipendium des Graduiertenkollegs (GK) für Teilchen- und Astroteilchenphysik der Fakultät für Physik der Universität Karlsruhe. Das GK unterstützte seine Mitglieder ebenso durch die Übernahme der Kosten von Dienstreisen und stellt eine Vielzahl von Möglichkeiten zur Weiterbildung bereit. Dadurch ist es möglich, effizient wissenschaftlich zu arbeiten. Diesem, den Professoren und Studentenvertretern, deren Einsatz das GK überhaupt möglich machte, danke ich hiermit. In der Hauptzeit wurde ich dann vom Land Baden-Württemberg finanziert. Diesem und dem Bundesministerium für Bildung und Forschung danke ich ebenso.
Nach der Phase der Einarbeitung haben wir es relativ schnell geschafft, einen neuen b-jet Tagger für CMS zu präsentieren. Vor allem die Hilfe von Christophe Saout sei hier nochmals erwähnt, dessen Sachverstand uns über viele Hürden geholfen hat. Dies ist umso bemerkenswerter, da er, als Frank-Peter Schilling seine b-tagging Tätigkeiten eingestellt hatte, zusätzlich auch noch die politischen Interessen des ekp in der CMS b-tagging Gruppe vertreten musste.
Die Suche zur Besetzung der nun freien PostDoc-Stelle für die ekp b-tagging Gruppe stellte sich dann als schwieriger heraus als erwartet. Dadurch ergab sich von organisatorischer Seite her eine schwierige Situation. Darum bedanke ich mich bei Thomas Kuhr und Jeannine Wagner-Kuhr, die uns in dieser Zeit als Ansprechpartner zur Verfügung standen und uns mit Rat und Tat zur Seite standen.
Anfang 2010 wurde die Stelle dann mit Jyothsna Komaragiri besetzt. Dadurch hatten wir wieder einen wichtigen Vertreter in der CMS b-tagging Gruppe. Nicht zuletzt durch ihren Einsatz wurde mir die Mitwirkung an der differentiellen b-Jet Wirkungsquerschnittsmessung ermöglicht. Dafür

bedanke ich mich.
Für die b-Jet Messung wurde eine Arbeitsgruppe, bestehend aus Fachleuten aus der Jet-Physik und b-Physik, gegründet. Die Analyse wurde von den b-tagging Konvenern Wolfgang Adam und Andrea Rizzi initiiert. Des weiteren danke ich Mikko Voutilainen für seine Leitung und Organisation dieser Gruppe. Ebenso bedanke ich mich bei den anderen Mitgliedern, vor allem bei Philipp Schieferdecker, Daniel Martschei, Hauke Held, Jyothsna Komaragiri und Niki Saoulidou.
Meine Doktorarbeit bestand dann aus den zwei großen Themen b-tagging und b-jet Wirkungsquerschnitt. Für die Analyse habe ich jeweils NeuroBayes vewendet, das von der Firma Phi-T vertrieben wird. Ich danke hiermit den Entwicklern der Software, vor allem Martin Hahn, aber auch Daniel Martschei, die mir immer wieder bei Fragen und Problemen zu NeuroBayes weitergeholfen haben.
Die Arbeit konnte auch nicht ohne die zahlreiche Hilfe meiner Kollegen am ekp fertiggestellt werden. Viele Gespräche und Diskussionen waren nötig. Dafür möchte ich allen am ekp herzlich danken. Im speziellen möchte ich mich bei den Korrektoren der Doktorarbeit Jyothsna Komaragiri, Iris Gebauer, Daniel Martschei, Sebastian Neubauer, Michael Feindt, Thomas Kuhr und Anže Zupanc bedanken.
Prof. Dr. Thomas Müller danke ich für die Übernahme des Korreferats und Prof. Dr. Ulrich Nierste für die Übernahme des Mentoriats.
Bei meinem Doktorvater Prof. Dr. Michael Feindt bedanke ich mich, dass ich bei ihm diese Arbeit anfertigen durfte. Ich danke für das ständige Vertrauen, dass er in mich und meine Arbeit hatte und dass mir die Freiheit gegeben wurde, diese so zu gestalten wie sie schlussendlich geworden ist.
Ich bedanke mich bei Daniel Martschei, dass er zusammen mit mir das CMS Experiment gewagt hat. Ich danke ihm für die letzten Jahre, dass wir trotz vieler Hoch und Tiefs nicht den Spaß an der Arbeit verloren haben und uns immer gegenseitig unterstützt haben.
Um mich ich im Bereich Teilchenphysik auf die Prüfung vorzubereiten, habe ich mich mit Susanne Mertens, Sebastian Neubauer und Felix Wick zusammen gesetzt und die großen Themen und Resultate der Teilchenphysik diskutiert. Allen dreien möchte ich hiermit danken. Des weiteren bedanke ich mich bei den Herrn Professoren, die mir die Prüfung abgenommen haben. Das sind neben den Referenten: Prof. Dr. Wim de Boer, Prof. Dr. Johann Kühn und Prof. Dr. Gerd Schön.
Ebenso danke ich meiner Familie für ihre Unterstützung: meinen Eltern Marlies und Norbert Honc und meinen Schwiegereltern Dina und Sepp Weintraut.
Am meisten danke ich meiner Frau Lisa, die ich unendlich liebe, und mit deren Ratschlägen, Ermutigungen, Humor und kluger Kritik ich die Zeiten des immerwiederkehrenden Frusts nicht überstanden hätte.
Euch allen danke ich nochmals herzlich für das Gelingen dieser Doktorarbeit.