CMS Draft Analysis Note

2013/11/20
Head Id: 217244
Archive Id: 217401M
Archive Date: 2013/11/20
Archive Tag: trunk

Calibration of the Combined Secondary Vertex b-Tagging discriminant using dileptonic ttbar and Drell-Yan events

Nazar Bartosik[1], Sarah Boutle[2], Andrew Brinkerhoff[3], Alexis Descroix[4], Tyler Dorland[1], Jasone Garay Garcia[1], Johannes Hauk[1], Michael Hildreth[3], Richard Hughes[5], Ulrich Husemann[4], Colin Jessop[3], Jan Kieseler[1], Jeff Kolb[3], Kevin Lannon[3], Wuming Luo[3], Rainer Mankel[1], Andreas Meyer[1], Hannes Mildner[4], Chris Neu[2], Carmen Diez Pardos[1], Patricia Lobelle Pardo[4], Darren Puigh[5], Alexei Raspereza[1], Jason Slaunwhite[3], Geoff Smith[5], Nil Valls[3], Jeannine Wagner-Kuhr[4], Shawn Williamson[4], Brian Winer[5], Matthias Wolf[3], John Wood[2], and Anna Woodard[3]

[1] DESY
[2] University of Virginia
[3] University of Notre Dame
[4] Karlsruhe Institute of Technology
[5] The Ohio State University

Abstract

This note describes a tag-and-probe technique for obtaining scale factors for correcting the per-jet combined secondary vertex (CSV) distribution for both heavy flavor and light flavor jets. The scale factor is calculated in exclusive bins of jet CSV output, jet p_{T}, and (in the case of the light flavor scale factor) jet η. Systematic uncertainties are assessed for tag-and-probe sample purities, jet energy scale uncertainties, and sample statistics. The scale factors are validated in control regions and found to provide a good description of the data.

This box is only visible in draft mode. Please make sure the values below make sense.

PDFAuthor:	Sarah Boutle, Kevin Lannon, Wuming Luo, Darren Puigh, Jason Slaunwhite, Geoff Smith, John Wood, Anna Woodard
PDFTitle:	Calibration of the Combined Secondary Vertex b-Tagging discriminant using dileptonic ttbar and Drell-Yan events
PDFSubject:	CMS
PDFKeywords:	CMS, physics, software

Please also verify that the abstract does not use any user defined symbols

1 Introduction

This note describes a new calibration technique for the combined secondary vertex (CSV) algorithm [1–3]. As described in more detail below, this technique calibrates the CSV distribution for both heavy flavor and light flavor quarks by means of a scale factor parameterized in terms of jet CSV value, jet p_T and jet η, applied on a per-jet basis. This calibration technique was designed to meet the specific needs of the search for $t\bar{t}H$ production, described in Ref. [4]. Although the studies of this calibration technique have focused on the use of the scale factors within the $t\bar{t}H$ analysis, the scale factors derived here could be employed in other analyses.

The CSV tagger plays a role in the $t\bar{t}H$ analysis in two ways. First, we identify jets as being b-tagged if they pass the medium CSV threshold (CSVM, CSV > 0.679). The number of CSVM-tagged jets are then used to separate the events into different categories with different signal purities and background contributions. Furthermore, we use the CSV output for jets as inputs to our MVA techniques for separating the $t\bar{t}H$ signal from the backgrounds. Therefore, we need scale factors for the CSV that correct, not only the rates of observing jets in MC with a CSV value above a given threshold, but also the overall shape of the CSV distribution. Note that, if the shape of the full CSV distribution is corrected so that data and MC agree, then this should also provide a correct description of the rates of events passing any CSV cut threshold.

At the time this note was written, the BTV POG provides two different sets of CSV scale factors, but neither of these sets meets all the requirements of the $t\bar{t}H$ analysis:

- **μ-Jet Scale Factors** [5, 6]: These scale factors are determined from QCD dijet events where one of the jets contains a muon. They are extracted for both heavy-flavor (HF) and light flavor (LF) jets, and they are parameterized in terms of jet p_T for HF and jet p_T and η for LF. However, these scale factors are only available for the rates of jets passing one of the three official CSV working points (loose, medium, and tight), and therefore, applied as provided, do not correct the shape of the CSV distribution. A technique has been developed to try to interpolate a differential scale factor from integrated scale factors for the three working points. This technique has been used in the VH, $H \to b\bar{b}$ search [7]. Unfortunately, attempts to use this technique in the $t\bar{t}H$ analysis have not yielded a sufficiently good description of the data.
- **$t\bar{t}$ Scale Factors** [8–10]: These scale factors are extracted by looking at the number of jets passing a given CSV threshold in a sample selected such that there is an enhanced contribution from $t\bar{t}$ production. As these scale factors come from $t\bar{t}$ production, only the HF scale factor is extracted. The scale factor is parameterized as a function of the CSV threshold and includes a larger range of values than just the three official CSV working points. There is no p_T or η dependence included in the scale factor parameterization. Although this scale factor does provide more CSV shape information than the previous method, there are still a number of issues with applying it to the $t\bar{t}H$ analysis. This approach does not supply a scale factor for the LF events. Furthermore, although it is parameterized for any value of the CSV threshold, it is still provided as an integrated scale factor and would have to be converted into a differential one using numerical differentiation techniques. Finally, this SF method matches the MC to the number of tagged jets distribution in data. This distribution also plays an important role in the $t\bar{t}H$ analysis, where the $t\bar{t}H$ signal is expected to have more tagged jets than the $t\bar{t}$ background. As the $t\bar{t}H$ fit is implemented, the b-tagging SF is also constrained *in-situ* during the $t\bar{t}H$ limit extraction, using a dataset that overlaps with the ones used in [8–10]. Therefore, to avoid double-fitting the data, we seek a method that extracts the SF from a sample

not used in the $t\bar{t}H$ analysis.

The technique described in detail below falls into the category of a "Tag-and-Probe" analysis. This approach relies on a sample of events selected to have two high-p_T, charged leptons and exactly two jets. Requirements placed on the lepton pair, the missing transverse energy E_T^{miss}, and the tag jet select a sample that is either enhanced in $t\bar{t}$ production (for the HF scale factor) or Z+jets production (for the LF scale factor). No CSV requirements are applied to the probe jet, to leave the CSV distribution of this jet unbiased. After correcting the data for unwanted contamination (for example, subtracting b-jet and c-jet contributions when calculating the LF scale factor), the SF is extracted by comparing the CSV distribution observed in data to that predicted by MC. This scale factor is determined separately for *exclusive* bins of CSV output, p_T, and (in the case of LF) η of jet. Uncertainties are assessed arising from contamination of the tag and probe sample by unwanted jet flavors, effects of JEC uncertainties, and the statistics of the samples used to obtain the scale factors. The scale factors are validated by checking their performance in samples independent from the ones used to extract them, and good agreement is found.

2 Data and MC Samples

The full list of data and MC samples used for this note is the same as is used for the $t\bar{t}H$ analysis [4]. In the following section, we list the primary samples used. See Ref. [4] for the full list (including the less signficant backgrounds).

2.1 Data Samples

The results presented here are based on the full 19.5 fb^{-1} of the 2012 CMS dataset. The double-lepton datasets shown in Table 1 were used to perform the tag-and-probe measurement.

Dataset	Run Range	Integrated Luminosity
/DoubleMu/Run2012A-13Jul2012-v1/AOD	190456–193621	0.81 fb^{-1}
/DoubleMu/Run2012A-recover-06Aug2012-v1/AOD	190782–190949	0.08 fb^{-1}
/DoubleMu/Run2012B-13Jul2012-v4/AOD	193834–196531	4.40 fb^{-1}
/DoubleMu/Run2012C-24Aug2012-v1/AOD	198022–198523	0.50 fb^{-1}
/DoubleMu/Run2012C-PromptReco-v2/AOD	198941–203746	6.40 fb^{-1}
/DoubleMu/Run2012D-PromptReco-v1/AOD	203768–208686	7.27 fb^{-1}
Total DoubleMu	**190645–208686**	**19.5 pb^{-1}**
/DoubleElectron/Run2012A-13Jul2012-v1/AOD	190456–193621	0.81 fb^{-1}
/DoubleElectron/Run2012A-recover-06Aug2012-v1/AOD	190782–190949	0.08 fb^{-1}
/DoubleElectron/Run2012B-13Jul2012-v1/AOD	193834–196531	4.40 fb^{-1}
/DoubleElectron/Run2012C-24Aug2012-v1/AOD	198022–198523	0.50 fb^{-1}
/DoubleElectron/Run2012C-PromptReco-v2/AOD	198941–203746	6.40 fb^{-1}
/DoubleElectron/Run2012D-PromptReco-v1/AOD	203768–208686	7.27 fb^{-1}
Total DoubleElectron	**190645–208686**	**19.5 fb^{-1}**
/MuEG/Run2012A-13Jul2012-v1/AOD	190456–193621	0.81 fb^{-1}
/MuEG/Run2012A-recover-06Aug2012-v1/AOD	190782–190949	0.08 fb^{-1}
/MuEG/Run2012B-13Jul2012-v1/AOD	193834–196531	4.40 fb^{-1}
/MuEG/Run2012C-24Aug2012-v1/AOD	198022–198523	0.50 fb^{-1}
/MuEG/Run2012C-PromptReco-v2/AOD	198941–203746	6.40 fb^{-1}
/MuEG/Run2012D-PromptReco-v1/AOD	203768–208686	7.27 fb^{-1}
Total MuEG	**190645–208686**	**19.5 fb^{-1}**

Table 1: The datasets analyzed for this analysis.

2.2 Monte Carlo Samples

The full list of MC samples from [4] is used to model the data samples listed above when extracting the CSV scale factors. However, most of the contribution comes from the $t\bar{t}$+jets and Z+jets samples listed in Table 2. For information on the smaller contributions from single top, $t\bar{t}V$, and diboson samples, see Ref. [4].

Sample	Dataset	Cross Sect.
$t\bar{t}$+ jets		
$t\bar{t} \to$ jets	/TTJets_HadronicMGDecays_8TeVmadgraph/Summer12_DR53X-PU_S10_START53_V7A_extv1/AODSIM	106.94 pb
$t\bar{t} \to \ell\nu + 4$ jets	/TTJets_SemiLeptMGDecays_8TeVmadgraph/Summer12_DR53X-PU_S10_START53_V7A_extv1/AODSIM	102.49 pb
$t\bar{t} \to \ell\nu\ell\nu + 2$ jets	/TTJets_FullLeptMGDecays_8TeVmadgraph/Summer12_DR53X-PU_S10_START53_V7Av2/AODSIM	24.57 pb
Z/γ^*+ jets		
10 GeV/c^2 $< M_{\ell\ell} <$50 GeV/c^2	/DYJetsToLL_M-10To50_TuneZ2Star_8TeV-madgraph/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM	14702 pb
$M_{\ell\ell} > 50$ GeV/c^2	/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM	3505.7 pb
$Z/\gamma^* + 1$ jet	/DY1JetsToLL_M-50_TuneZ2Star_8TeV-madgraph/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM	666.7 pb
$Z/\gamma^* + 2$ jets	/DY2JetsToLL_M-50_TuneZ2Star_8TeV-madgraph/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM	215.1 pb
$Z/\gamma^* + 3$ jets	/DY3JetsToLL_M-50_TuneZ2Star_8TeV-madgraph/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM	66.07 pb
$Z/\gamma^* + 4$ jets	/DY4JetsToLL_M-50_TuneZ2Star_8TeV-madgraph/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM	27.38 pb

Table 2: List of background MC datasets and cross sections used for normalization.

2.3 MC pileup reweighting

During the 2012 data taking period, the LHC provided increasingly large instantaneous luminosities to the experiments. As a result, the average number of overlapping events per time interval also increased. These overlapping pileup events that occur along with the physics events of interest can have an effect on everything from lepton isolation to jet reconstruction. Therefore, it is important that our simulated events have the same distribution of pilup events as found in the data.

When the simulation events were generated, the average amount of pileup that we would see in 2012 was unknown. Therefore, the pileup distribution in the simulation needs to be reweighted to match the data. For the simulation, it is known how many additional interactions were added to every generated event. For the data, the number of pileup interactions for each unit of time depends on the instantaneous luminosity for each bunch pair and the total inelastic cross section, $\sigma_{\text{inelastic}}$. For the inelatistic cross section, we use the standard CMS value $\sigma_{\text{inelastic}} = 69.4$ mb.

2.4 Top p_{T} reweighting

It has been generally observed that the spectra of leptons and jets produced from top quark decays have softer p_{T} distribution than are predicted by the Monte Carlo. Investigations in the top group have traced this mismodeling to the top quark p_{T} distribution. It was found in Refs [11, 12] that the fully corrected differential cross section for top pair production as a function of the top quark p_{T} is softer than the predictions of any $t\bar{t}$+jets Monte Carlo and more consistent with calculations done at approximate NNLO accuracy. Whenever we use $t\bar{t}$ MC in

deriving these CSV scale factors, we apply the same reweighting used in [4] to ensure the best possible description of the data by the MC.

3 Tag-and-Probe

We use the full 8 TeV DoubleMu, DoubleElectron and MuEG datasets taken in 2012. First, we find control regions with high purity for heavy flavor or light flavor jets. Dilepton $t\bar{t}$ events, $t\bar{t} \to \ell\nu\ell\nu b\bar{b}$, are well suited for the former since that sample is dominated by events which have two b flavor jets from the top pair decay. We also use a sample dominated by Z+jets where there are two light flavor jets. For events with one jet passing the tag requirements, we plot the CSV distribution for the probe jet in given p_T and η bins. We normalize the total MC yields to the data yields. In order to account for heavy or light flavor contamination, we divide our MC samples into heavy flavor and light flavor components and then subtract the non-relevant part from data. The scale factor is then just the ratio of subtracted data CSV distribution and the relevant MC CSV distribution, as shown below:

$$SF(\text{CSV}, p_T, \eta) = \frac{\text{Data} - \text{MC}_A}{\text{MC}_B} \tag{1}$$

where A, B = heavy flavor component or light flavor component.

3.1 Heavy Flavor Scale Factor

To calculate the heavy flavor scale factor, we need a control region which has high b-jet purity. The following cuts are designed to select a sample enriched in dileptonic ttbar events. The jets in these events will be enriched in b content.

- The event has two leptons ($ee/\mu\mu/e\mu$) and exactly 2 jets.
- For $ee/\mu\mu$ events, they have to pass a Z boson veto cut. The Z boson veto use $m_{\ell\ell}$ and E_T^{miss} to reject events in which the dilepton pair likely comes from a Z boson decay. To pass this cut, the event has to satisfy at least one of the following conditions:
  - $m_{\ell\ell} < (65.5 + 3E_T^{miss}/8)$
  - $m_{\ell\ell} > (108 - E_T^{miss}/4)$
  - $m_{\ell\ell} < (79 - 3E_T^{miss}/4)$
  - $m_{\ell\ell} > (99 + E_T^{miss}/2)$

  Figure 1 shows a diagram of the Z boson veto region in the $m_{\ell\ell}$ vs. E_T^{miss} plane. Note: This veto was optimized in the context of the $t\bar{t}H$ search in the dilepton channel. Most, but not all, of the region rejected by this veto is also rejected by the E_T^{miss} cut listed below. Nonetheless, we apply both cuts to ensure maximal removal of Z+jets events.
- $E_T^{miss} > 50\,\text{GeV}$.
- The tag jet must pass the CSVM working point, CSV > 0.679.

Based on MC expectation, the selected sample is composed of 91% $t\bar{t}$, 5% single top, 3% Z+jets and 1% other backgrounds. To account for contamination from light flavor jets, we divide the MC sample into a b-jet part and a non-b-jet part based on the flavor of the probe jet and then subtract the non-b-jet part from data. For this scale factor calculation, jets containing a c quark are included with the non-b-jet contribution. For the probe jet, the fraction of b-jets in MC is

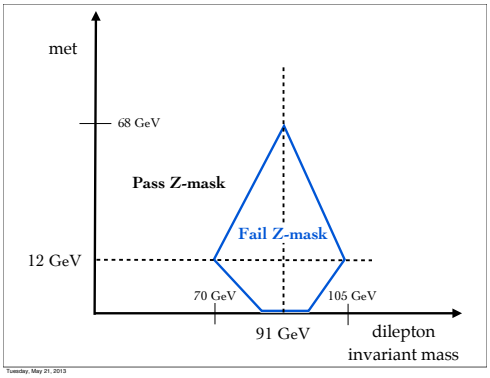

Figure 1: Diagram of the Z boson veto cuts in the $m_{\ell\ell}$ vs $E_{\mathrm{T}}^{\mathrm{miss}}$ plane.

about 80%. The heavy flavor scale factor is calculated as below:

$$\text{HF } SF(\text{CSV}, p_{\mathrm{T}}) = \frac{\text{Data} - \text{MC}_{\text{non-b-jets}}}{\text{MC}_{\text{b-jets}}} \tag{2}$$

To account for the p_{T} depedence of the scale factor, the control region is divided into five subsamples based on the probe jet p_{T}, and the scale factor is extracted separately for each subsample. The dependence of the HF scale factor was also checked as a function of probe jet η, but no dependence was observed.

The p_{T} bins used for the HF scale factor are given below:

- $30\,\text{GeV}/c \leq p_{\mathrm{T}} < 40\,\text{GeV}/c$
- $40\,\text{GeV}/c \leq p_{\mathrm{T}} < 60\,\text{GeV}/c$
- $60\,\text{GeV}/c \leq p_{\mathrm{T}} < 100\,\text{GeV}/c$
- $100\,\text{GeV}/c \leq p_{\mathrm{T}} < 160\,\text{GeV}/c$
- $p_{\mathrm{T}} \geq 160\,\text{GeV}/c$.

Although the HF scale factor is not binned in η, scale factors for jets with $|\eta| > 2.4$ are not evaluated in this note.

3.2 Light Flavor Scale Factor

To calculate the light flavor scale factor, we need a control region which has a high purity of light flavor jets. We select events for the control region with the following requirements:

- The event has two leptons ($ee/\mu\mu$) and exactly 2 jets
- Fail the Z boson veto cut, described above
- $E_{\mathrm{T}}^{\mathrm{miss}} < 30\,\text{GeV}$
- $|m_{\ell\ell} - 91| < 10\,\text{GeV}/c^2$
- The tag jet must fail the CSVL working point: CSV < 0.244.

Based on MC expectation, the selected control region is composed of 99% Z+jets and 1% others. In order to account for heavy flavor contamination, we divide the MC sample into a light flavor part and a non-light flavor part based on the flavor of the probe jet and then subtract the non-light flavor part from the data. For the purpose of this calculation, charm jets are included with the heavy flavor. For the probe jet, the fraction of light flavor jets in MC is about 90%. The light

flavor scale factor is calculated as below:

$$\text{LF } SF(\text{CSV}, p_{\text{T}}, \eta) = \frac{\text{Data} - \text{MC}_{\text{non-light flavor}}}{\text{MC}_{\text{light flavor}}} \tag{3}$$

To account for the p_{T} and η depedence of the scale factor, the control region is divided into nine subsamples based on the probe jet p_{T} and η, and the scale factor is extracted separately for each subsample. Each subsample is nonoverlapping in both p_{T} and in η. Unlike the HF case, it was necessary to paremeterize the scale factor in terms of both p_{T} and η for the LF jets.

The p_{T} and η bins used for the LF scale factor have the following ranges in p_{T} and η. To isolate a specific bin, cuts must be applied both on p_{T} and η:

- p_{T} bins:
  - $30\,\text{GeV}/c \leq p_{\text{T}} < 40\,\text{GeV}/c$
  - $40\,\text{GeV}/c \leq p_{\text{T}} < 60\,\text{GeV}/c$
  - $p_{\text{T}} \geq 60\,\text{GeV}/c$
- η bins:
  - $|\eta| < 0.8$
  - $0.8 \leq |\eta| < 1.6$
  - $1.6 \leq |\eta| < 2.4$.

Scale factors for jets with $|\eta| > 2.4$ are not evaluated in this note.

3.3 Iterative Technique and Smoothing

There is a bit of a logical problem with the approach outlined above. In order to calculate the b-jet scale factor, the LF scale factor must be known, so that the correct shape for the LF contamination can be subtracted. Likewise, to calculate the LF scale factor, the b-jet scale factor must be known. To overcome this conundrum, we take an iterative approach. For the first iteration, the contamination is subtracted with no CSV scale factor applied. After this, the scale factors calculated from the previous iteration are used to subtract the contamination for the current one. For example, when we calculate the HF scale factors for a given iteration, we apply the LF scale factors from the previous iteraction to improve when subtracting the estimated LF contamination. We stop iterating once the new SFs do not change much with respect to the old ones; this was achieved after three iterations. Figure 2 shows examples of HF and LF scale factors converging.

Instead of using the scale factors binned in CSV, we parameterize the scale factor curves. This is done to minimize the impact of statistical fluctuations and to ensure a smooth and continuous CSV scale factor. We fit the light flavor scale factor to a 6th-order polynomial function. Since the heavy flavor scale factors have a complicated shape, we interpolate between the points in each CSV bin to obtain a smooth function. Jets with negative CSV values, indicating insufficient information to run the complete CSV algorithm, are included in a single bin below 0 which is not smoothly connected to values above 0. Figure 3 shows an example of the smoothing process.

3.4 Scale Factor Results

The results for one p_{T} and η bin for both HF and LF scale factors is shown in Figs 4 and 5. Each of these figures contains three plots. The first plot shows the data selected in the given control region, compared to the sum of MC expectations, broken down in the the component

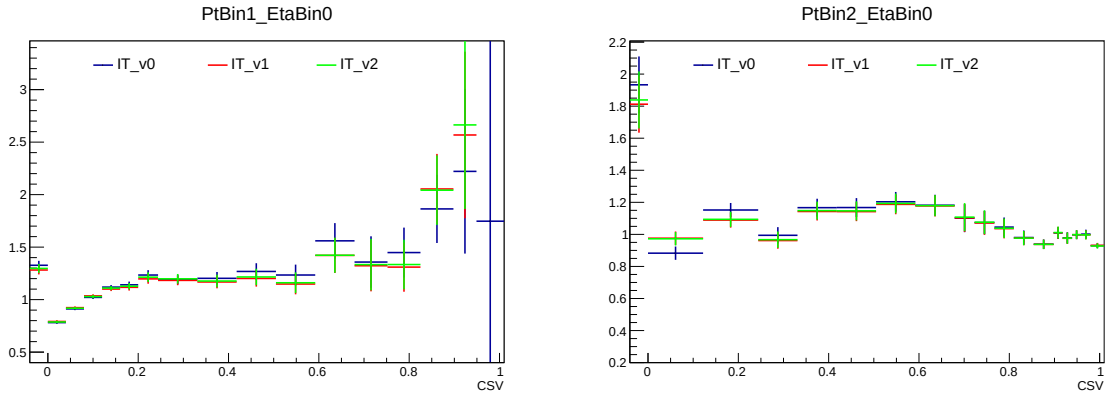

Figure 2: Iterative technique for the scale factor. Points in blue, red, and green are iterations 0, 1, and 2, respectively. Left is an example of the light flavor scale factor and right is an example of the heavy flavor scale factor.

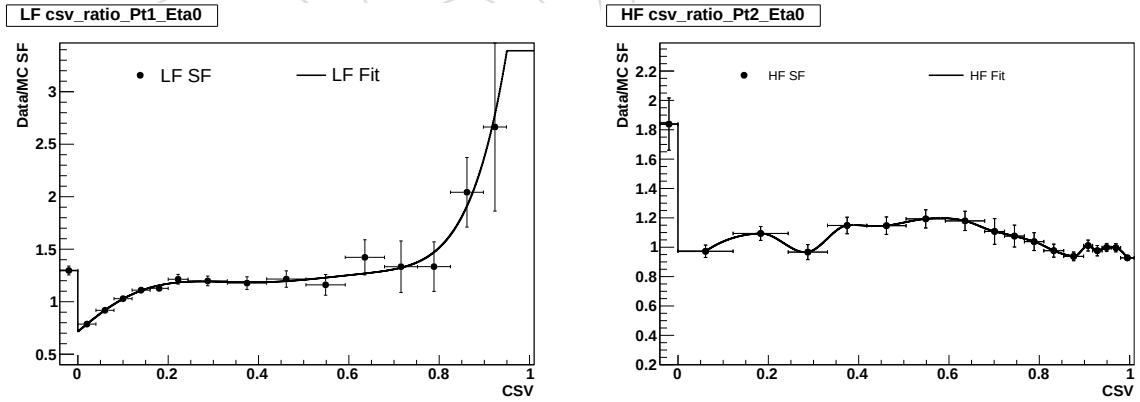

Figure 3: Left: fitting the light flavor scale factor. Right: interpolating for the heavy flavor scale factor.

of interest (either b quarks or LF) and the contamination from other sources. The sum of the MC expectations is normalized to the total data yield. The second plot shows a comparison of the data shape with the contamination subtracted out (subtracting c and LF from the b-jet scale factor, and subtracting b and c from the LF scale factor), compared to the relevant (b-jet or LF) component from the MC. The last plot shows the resulting scale factor that comes from taking the ratio of the subtracted data. The plots for all p_T and η bins are shown in Appendices A and B.

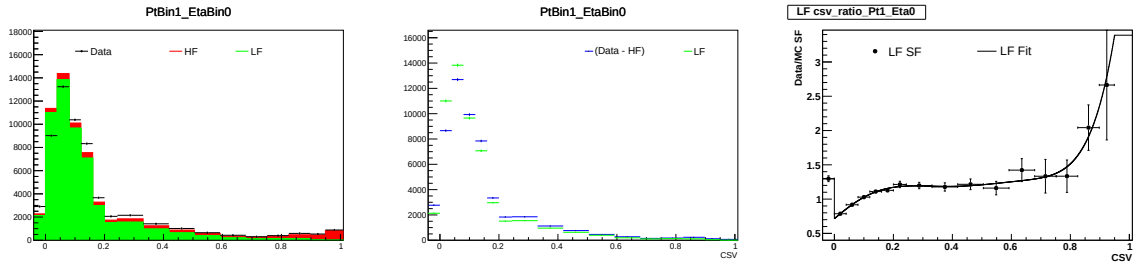

Figure 4: Left: Data/MC CSV distribution comparison, with MC normalized to Data yields. Middle: CSV distributions for (Data - MC_{HF}) compared with MC_{LF}. Right: scale factor as a function of p_T along with the fitted function. The plots shown here are from the bin with $40\,\text{GeV}/c \leq p_T < 60\,\text{GeV}/c$ and $|\eta| < 0.8$.

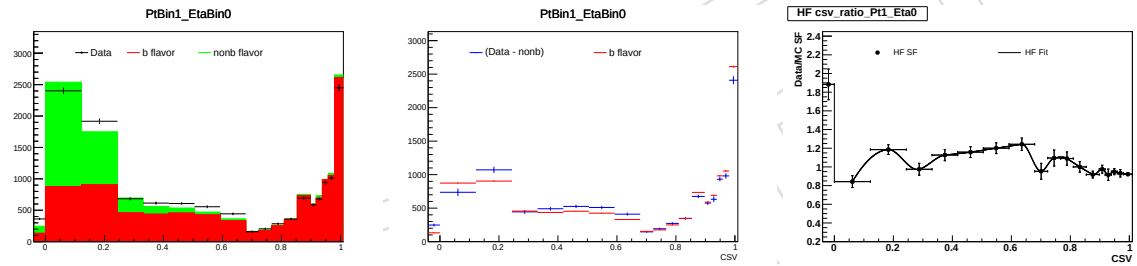

Figure 5: Left: Data/MC CSV distribution comparison, MC stack are normalized to Data yields. Right: CSV distributions for (Data - MC_{non-b}) and MC_b. The plots shown here are from the bin with $40\,\text{GeV}/c \leq p_T < 60\,\text{GeV}/c$. (Recall that the HF scale factor is not binned in η.)

3.5 Scale Factor Application

To apply the scale factors, for each MC event, we loop through each jet passing our analysis selection. If the jet is a b flavor jet, we assign a heavy flavor scale factor to it. Otherwise, if it is a light flavor jet, we assign a light flavor scale factor to it. The BTV POG convention is to use the heavy flavor scale factors for c flavor jets, just as for b flavor jets, but with twice the uncertainty. However, we found this convention not quite appropriate for our special needs where we not only have to correct the b-tagging rates but also the CSV shapes. The main problem is that the CSV output shape for b-jets in the data is quite different from what the MC simulation predicts for charm jets. Figure 6 shows two examples of the jet CSV shape comparison: the black line is the CSV distribution for b-jets in the control sample used for the heavy flavor scale factor derivation, the red and green lines are the charm jets CSV distribution in $t\bar{t}H$ (m_H= 125 GeV/c^2) and $t\bar{t}$ samples, respectively.

If we were to apply the heavy flavor scale factors we derived to charm jets, for example, multiplying the CSV shapes of charm jets as those in Figure 6 by the corresponding heavy flavor scale factors, the normalization of the charm jets CSV distribution would change by 3% to 20%,

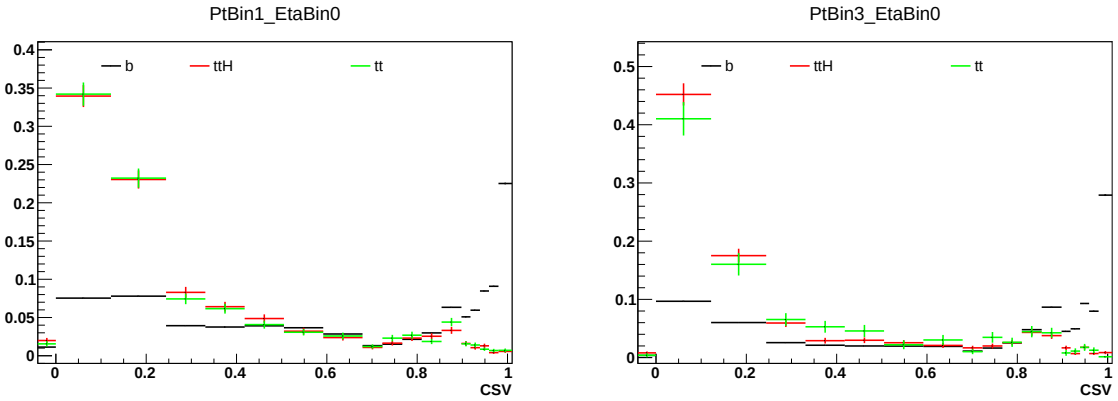

Figure 6: CSV distribution comparison between charm jets in $t\bar{t}H$ (m_H= 125 GeV/c^2) and $t\bar{t}$ samples and b-jets in the control sample for the heavy flavor scale factor derivation. The events in the $t\bar{t}H$ and $t\bar{t}$ samples must have two oppositely charged leptons, at least 3 jets, and pass the Z-veto cut, but without any b-tagged jets requirements.

depending on the jet p_T bin. Moreover, this would also have a non-negligible impact on the jet CSV shapes of samples that have a significant amount of charm jets, such as $t\bar{t}c\bar{c}$, which is an important background for $t\bar{t}H$ analysis.

In the absence of a data-driven calibration sample for charm jets, we set the scale factors for charm jets to 1.0 and retain the relative uncertainty from the calibration for b-jets. Details of how we treat the uncertainties for charm jets are covered in the following systematic uncertainty chapter.

To summarize, the scale factors are applied as follows:

- For b flavor jets, assign heavy flavor scale factors
- For light flavor jets, assign light flavor scale factors
- For c flavor jets, assign a flat scale factor of 1.

The total scale factor for the event is the product of all the scale factors of the jets:

$$SF_{\text{total}} = \prod_{i}^{N_{\text{jets}}} SF_{\text{jet}_i} = SF_{\text{jet}_1} \cdot SF_{\text{jet}_2} \cdot ... \tag{4}$$

In the end, the CSV value for each jet remains unchanged, but each event gets a weight of SF_{total}.

4 Systematic Uncertainties

Not only do we have to get the correct b-tag scale factors, but we also need to consider any systematic uncertainty related to the scale factor calculation procedure to make sure the systematic uncertainties are sufficient to cover any possible CSV shape discrepancy between data and MC. There are three sources of b-tag uncertainty for both light flavor and b flavor jets considered below: jet energy scale, purity and statistics of the control sample. For c flavor jets, the scale factor uncertainties are treated separately.

4.1 Jet Energy Scale Uncertainty

Instead of the nominal MC samples, we use the JES shifted samples. Otherwise, the basic method remains unchanged. Shifting the JES will only change the p_T of each jet, which could lead to events migrating in or out of the selected sample and jets migrating to different p_T bins for the scale factor. Figure 7 shows an example of the effect of shifting JES up and down on the derived SF.

Depending on the details of the analysis using this scale factor, the JES uncertainty may also impact other aspects of the analysis. If the JES uncertainty is important, it may be advisable to correlate this part of the b-tag uncertainty with the JES. This is the approach taken for the $t\bar{t}H$ analysis. When assessing the JES uncertainty for that analysis, we use the CSV scale factors corresponding to the appropriate JES shift.

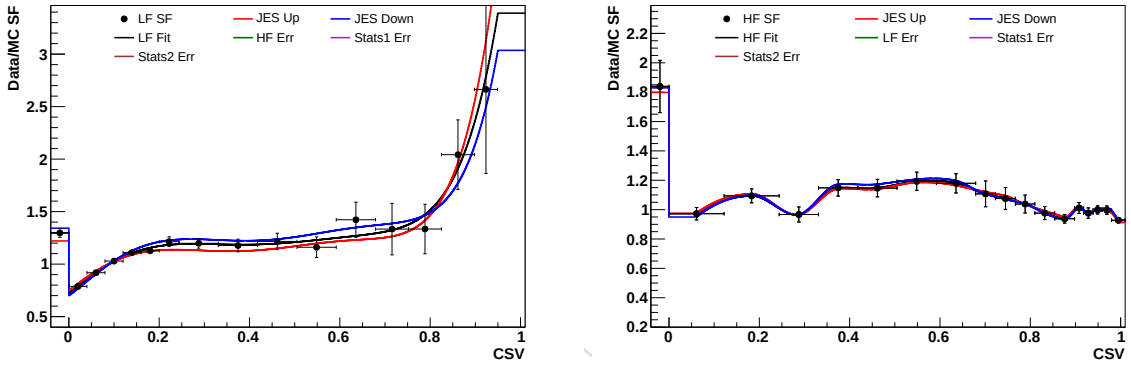

Figure 7: Scale factor comparison for the JES uncertainty. Left: the light flavor SF. Right: the heavy flavor SF.

4.2 Purity Uncertainty

In both our heavy flavor and light flavor scale factor calculations, we use MC to model the ratio of heavy flavor and light flavor components in the data for the purpose of subtracting the non-relevant part. Therefore, we need to consider the uncertainty on the MC prediction for this ratio and apply that as an uncertainty on our calculated scale factors. The uncertainty assessed for the HF and LF scale factors is described below.

4.2.1 LF Scale Factor

When calculating the LF scale factor, we must assess the uncertainty on the contamination from b- and c-jets. The expected contribution to the LF scale factor control sample from processes other than Z+jets is negligible, so the uncertainty on the HF contamination is dominated by the uncertainty on the fraction of b and c production in Z+jets events. Recent measurements from CMS find the rate of Z+b-jet production to be well described by our MC, with the agreement in inclusive rates being within 5% for both the single b and double b topologies [13]. However, a study of the differential cross section of $Z+b\bar{b}$ production suggests that depending on the kinematics of the event, the predictions from MADGRAPH may differ from data by as much as a factor of two [14]. Therefore, we must use control regions to assess the level of agreement between our MC and the data in terms of contributions from Z+HF.

We begin by assessing the level of agreement between the observed and predicted yields in a sample selected to be enhanced in Z+b-jet content, using the selection criteria listed below. For

this check, the scale factors after iteration are applied for both HF and LF.

- The event has two leptons ($ee/\mu\mu$) and exactly 2 jets.
- We require the same Z selection as we use for the sample from which the LF scale factor is selected, namely that the events fail the Z boson veto cut, have $E_{\mathrm{T}}^{\mathrm{miss}} < 30\,\mathrm{GeV}$ and have $|m_{\ell\ell} - 91| < 10\,\mathrm{GeV}/c^2$. (See Sec. 3.2 for more details.)
- For both jets, CSV > 0.898 (CSVT).

We observed 974 data events and 1095 MC events in this control region. The MC prediction breaks down as follows:

- 964 Z+jets events.
  - 876 of the above events are either Z+$b\bar{b}$ or Z+$c\bar{c}$ events.
  - The remaining 88 events come from mistagged Z+dijet events with fewer than two heavy flavor jets.
- 131 events from sources other Z+jets, such as $t\bar{t}$ or single top.

The plot in Figure 8 shows the distributions of the invariant mass of the two leptons for the data events after subtracting the non Z+jets MC events as well as for Z+jets events. We see good agreement between data and MC for Z boson production with two HF jets within 12%. This check relies on the measurement of the LF scale factor, the uncertainty on this quantity being the thing we are attempting to assess. However, even if we increased the LF scale factor by a factor of 2 from our measured value, the agreement in this check would remain in the 20% range. Therefore, we conservatively estimate this HF contamination uncertainty at the 20% level. In assessing this uncertainty, we vary the contribution from Z+b and Z+c simultaneously upward or downward by 20% and recalculate the LF scale factor. It should be noted that a 20% uncertainty on the HF contamination in this sample is comparable to the uncertainty assessed by the BTV POG in the determination of the mistag rate using the QCD dijet dataset [6].

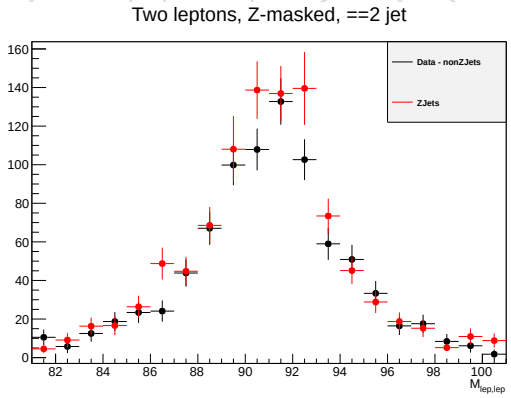

Figure 8: The distributions of the invariant mass of the two leptons for the data events after subtracting the non Z+jets MC events as well as for Z+jets events.

The possible impact of mismodeling of the rate of Z+b compared to Z+$b\bar{b}$ topologies is investigated in Sec. 5 and found to be consistent within the uncertainties assessed here.

4.2.2 HF Scale Factor

As for the uncertainty on the LF fraction while calculating the HF scale factors, we know the largest contribution to the control sample is from $t\bar{t}$ (with small amounts from other processes).

So, the relevant quantity is the uncertainty on the rate of LF jets in $t\bar{t}$ (and, to a lesser extent, single top) production. To first approximation, all jets coming from $t\bar{t}$ production will be b-jets. However, when $t\bar{t}$ is produced along with additional partons, there is the possibility that one of the b-jets will fail to be reconstructed and a light flavor jet will be reconstructed instead. From studies in the $t\bar{t}H$ search [4] we know that the rate of $t\bar{t}$+extra jets varies by up to 20% with the variation on Q^2 scale in MADGRAPH. The amount of variation depends on the number of extra jets. The maximum variation of 20% is seen for cases where there are ≥ 2 additional partons produced with the $t\bar{t}$ pair. For $t\bar{t}+1$ parton final states, the variation is approximately 10%. Based on these studies, we assess an uncertainty of 20% on the LF contamination for the probe jet in the HF tag-and-probe sample.

4.2.3 Impact of the Purity Uncertainties

Examples of the comparison of the new scale factors after changing the LF/HF fraction to the nominal ones is shown in Figure 9. The assessed uncertainties produce relatively large variations of the scale factors in regions where the contamination is large. For example, in the LF scale factor, the HF purity uncertainty is largest for high CSV values. However, this uncertainty turns out not to have a very large impact on analyses like the $t\bar{t}H$ analysis because there are few LF MC events with CSV values in the region where the uncertainty is largest.

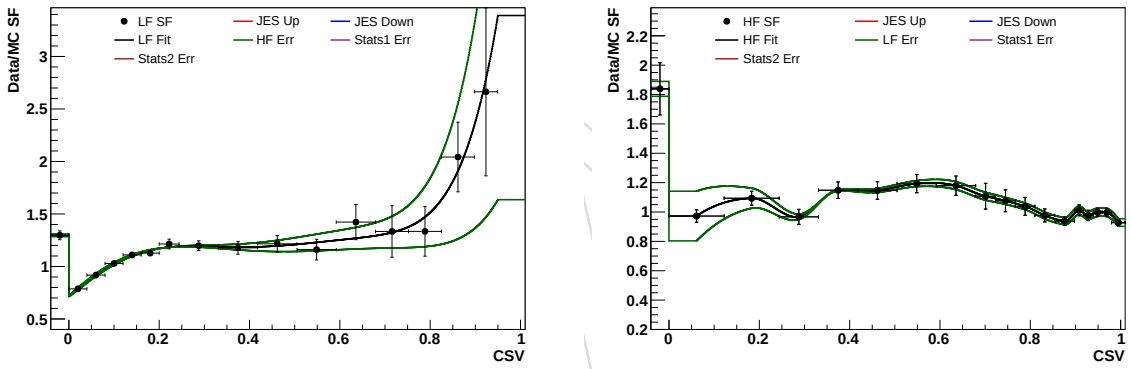

Figure 9: Scale factor comparison for the purity uncertainty. Left: the light flavor SF. Right: the heavy flavor SF.

4.3 Statistical Uncertainty

The statistics of the samples used to extract the scale factors must also be taken into account. Although there are generally sufficient statistics in these samples to make the statistical uncertainty negligible, limited statistics in certain regions, such as at the high CSV end for the LF SF and the low CSV end for the HF SF, require us to account for this uncertainty. In these regions, since we are calculating the ratio of two small numbers, any statistical fluctuation would change the SF.

Since the scale factors calculated here merely change the shape of the inclusive CSV distribution, but not the rate of events, random fluctuations up and down across the full CSV distribution have little impact. On the other hand, fluctations that affect primarily one part of the distribution will cause a noticeable shape change. It is these flucations that are most important. To address these fluctuations, we introduce first order and second order polynomial functions $f_1(x)$ and $f_2(x)$ such that the scale factor value for each CSV bin is varied according to the following equations where x is the central CSV value of that bin, and σ represents the statistical

uncertainty on the scale factor in that bin:

$$
\begin{aligned}
Up &= +\sigma \cdot f(x) \\
Down &= -\sigma \cdot f(x) \\
f_1(x) &= 1 - 2 \cdot x \\
f_2(x) &= 1 - 6 \cdot x \cdot (1 - x).
\end{aligned}
$$

A zeroth-order polymonial would simply adjust the overall rates without changing the CSV distribution, so the contribution from that class of variations is ignored. The first order polynomial catches the effect of statistical fluctations that would tend to tilt the SF distribution, while the second-order polynomial describes fluctations that increase or decrease the scale factor in the center of the CSV distribution compared to the ends, as shown in Figure 10. To good approximation, an arbitrary distrortion in the scale factor shape can be described using the sum of these two variations.

The varied SF distributions are then fit using the same procedure as the nominal ones, and these fitted functions define two independent sets of systematically varied SF functions to account for the statistical uncertainty in the SF determination.

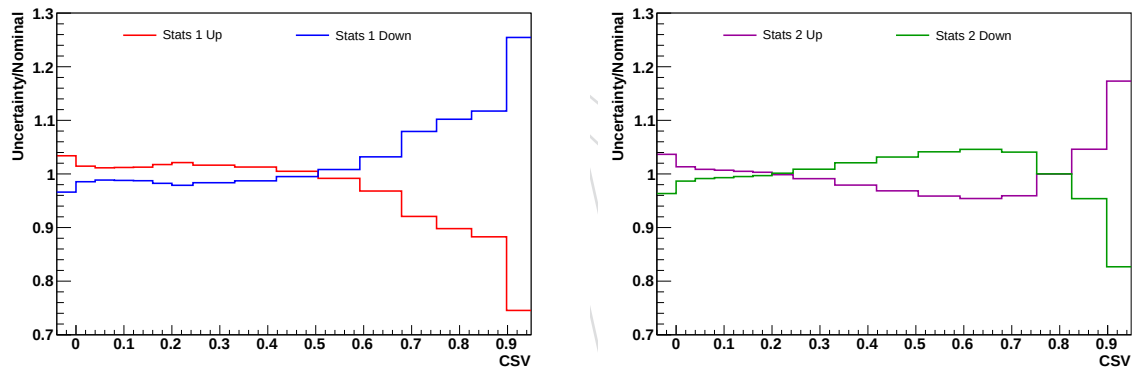

Figure 10: First order (left) and second order (right) variations for the statistical uncertainty.

Since these variations should change the shape but not the normalization of the CSV distribution, f_1 and f_2 are chosen to preserve the overall normalization. The functions are also chosen to be orthogonal so the two uncertainties can be applied independently. An example of how the SF changes is shown in Figure 11. Each of the statistical fluctuations will be considered as a new systematic uncertainty, namely first/second order HF/LF statistics.

The above treatment of statistical uncertainty is, in some sense, the worst case (most conservative) scenario for the effect of statistics on the scale factor: that a set of neighboring bins fluctuate up/down, causing a neighboring set of bins somewhere else to fluctuate down/up (to preserve overall normalization). Another approach would be to have a separate systematic scale factor for each CSV bin corresponding to the statistical uncertainty in that bin. Here, the up/down fluctuation in a CSV bin is given by the statistical uncertainty on the scale factor in that bin, and all other bins must fluctuate down/up (by the same fraction) to preserve overal normalization. Figure 12 shows how the scale factor changes due to the statistical uncertainty in each of the first few bins. The trade-off from using this second approach over the first is that it requires 90 new systematics for the HF scale factor (18 CSV bins, 5 p_{T} bins) and 144

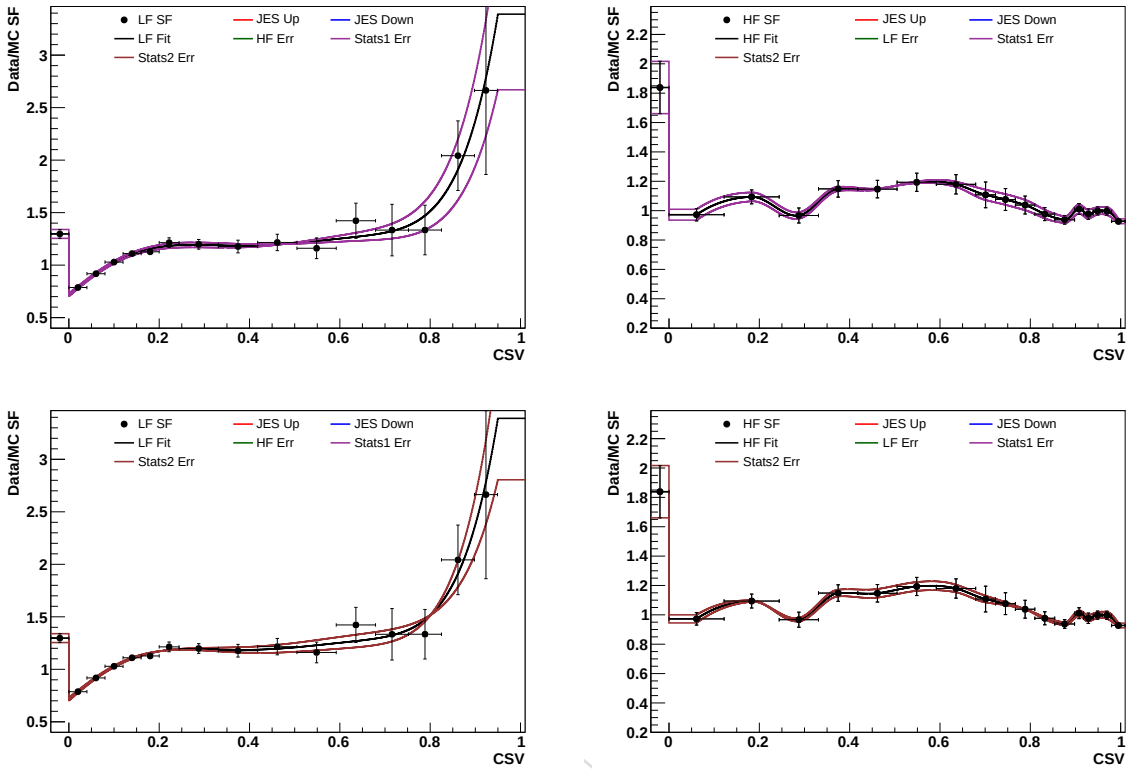

Figure 11: Scale factor comparison for the statistical uncertainty. Left: light flavor SF. Right: heavy flavor SF. Top: linear component. Bottom: quadratic component.

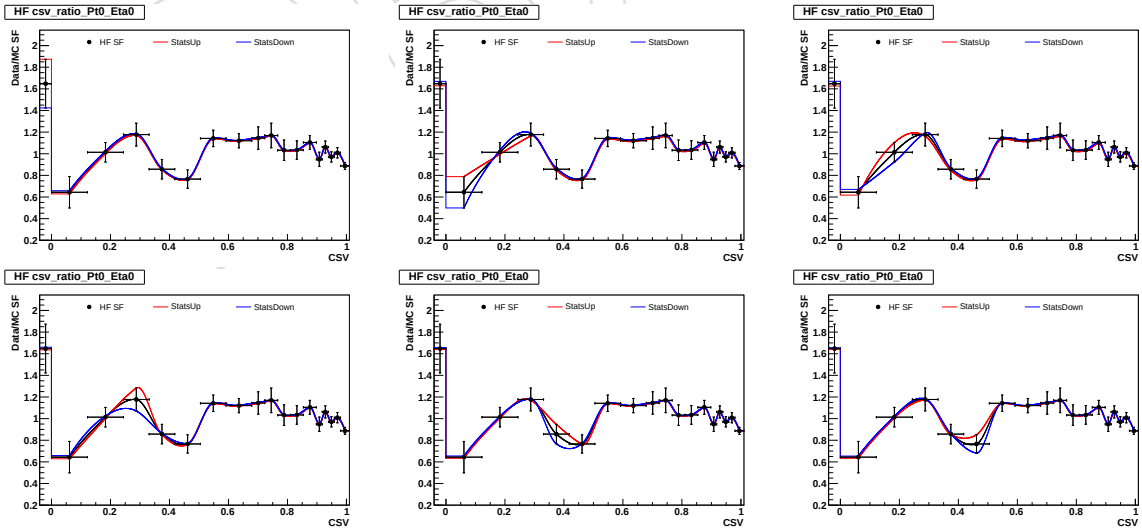

Figure 12: Comparison of the nominal heavy flavor scale factor to the systematic up/down scale factors obtained from the statistical uncertainty in the first six bins of CSV.

new systematics for the LF scale factor (16 CSV bins, 3 p_T bins, 3 η bins), as opposed to 4 total systematics in the previous approach for both HF and LF.

4.4 Scale Factor Uncertainties for Charm Jets

For c-jets, we set the scale factor to 1 and use twice the relative uncertainty as we use for b-jets. For example, if the SF (in a certain p_T, η, CSV bin) for b-jets is 0.80 ± 0.20, then the SF for c-jets would be 1.00 ± 0.50. As discussed above, there are four sources of uncertainty for the b-jets SF: JES, purity, first order and second order statsitics. We take all the uncertainties from b-jets, make them relative uncertainties, and then add them up in quadrature to get an overall uncertainty, as shown in the left plot of Figure 13. As the BTV group convention recommended, we double the overall relative scale factor uncertainty for b-jets in size and then use it as the relative scale factor uncertainty for c-jets. Following the same approach as in Section 4.3, we construct two separate uncertainties for c-jets: a linear piece and a quadratic piece. In this case, σ represents the relative uncertainty from above. The right plot in Figure 13 shows an example of the final CSV SFs for c-jets along with the two separate uncertainties (red curve for the linear piece and blue curve for the quadratic piece), which will be treated separately from all the uncertainties for light flavor and b-jets.

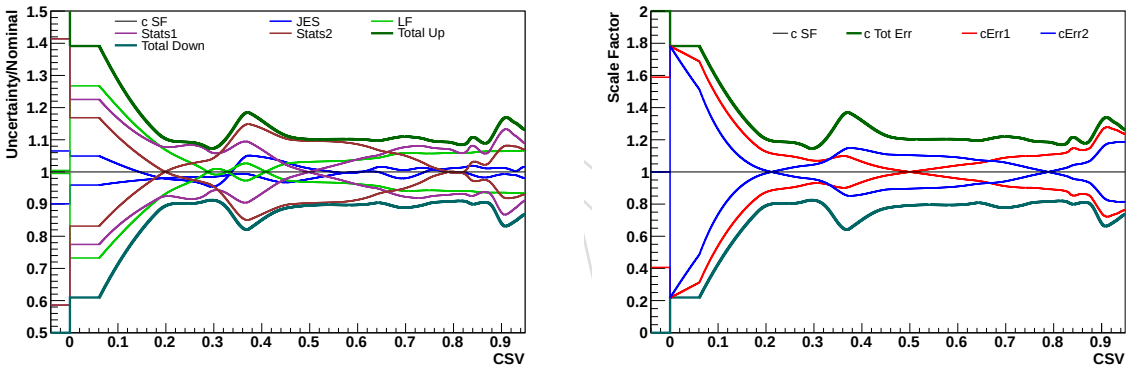

Figure 13: Relative uncertainties from b-jets SFs on the left and final SFs with uncertainties for c-jets on the right (red curve is for the linear uncertainty and blue curve is for the quadratic uncertainty).

5 Validation

To some extent, there is an unavoidable degeneracy in this technique between incorrectly estimating sample composition and the extracted scale factor. For example, underestimating the contribution of Z+b events to the LF tag-and-probe sample would result in a higher CSV scale factor for light flavor jets than the true value. The uncertainites evaluated in this note should account for such effects; however, it is useful to confirm these uncertainties with additional checks.

A powerful check of our method is to apply the obtained scale factors in samples where the sample composition is significantly different than the ones used to extract the scale factors in the first place. If a scale factor is mismeasured, for example, because of an underestimation of a particular HF contamination source, when the scale factors are applied in a different sample with a different mixture of HF and LF jets, the agreement will be poor. Below we discuss two such checks.

5.1 Z+1 b-jet check

The level of possible disagreement between data and MADGRAPH shown in differential distributions measured in Ref. [14], as well as the difference in scale factors assessed on the production of Z+b and Z+$b\bar{b}$ topologies in Ref. [7], may raise concerns about the uncertainty assessed on the Z+HF contamination in the LF scale factor extraction. As a cross check, we apply our extracted scale factors in a sample that has a different mixture of Z+HF and Z+LF than the sample used to extract the scale factors, as described below.

- The event has two leptons ($ee/\mu\mu$) and exactly 2 jets.
- We require the same Z selection as we use for the sample from which the LF scale factor is selected, namely that the events fail the Z boson veto cut, have $E_{\mathrm{T}}^{\mathrm{miss}} < 30\,\mathrm{GeV}$ and have $|m_{\ell\ell} - 91| < 10\,\mathrm{GeV}/c^2$. (See Sec. 3.2 for more details.)
- At least one jet must pass CSV > 0.898 (CSVT).

Unlike the sample used to calculate the LF scale factors, in which at most one jet can be b-tagged with the CSVL operating point, this sample requires that at least one jet be b-tagged with the CSVT operating point. This control sample is dominated by Z+$\geq$1 b-jet events. The ratio of events in this sample having only one b-tagged jet to those events with two b-tagged jets will be sensitive to the agreement between data and MADGRAPH for Z+b versus Z+$b\bar{b}$ rates. To count b-tagged jets for this check, we use the CSVM operating point. Every event has at least one b-tagged jet, and, for some events, both jets are b-tagged.

Figure 14 shows the ratio of our observation in data to the MADGRAPH prediction for the rate of events with 1 b-tagged jet and events with 2 b-tagged jets. The agreement in the 1-b-tagged jet bin is within 1%, well within the statistical uncertainty on that bin. The level of agreement in the 2-b-tagged jets bin is consistent with the check performed to assess the purity uncertainty on the LF scale factor in Sec. 4.2.1. Based on this level of agreement, there appears to be no motivation to increase the size of the uncertainty on the HF contamination in the LF scale factor control sample.

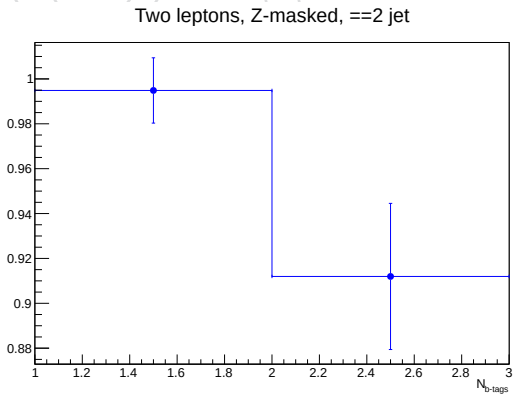

Figure 14: Data/MC ratio for the number of b-tagged jets distribution.

5.2 Testing in Lepton+Jets

The sample of events selected to be consistent with the semileptonic decay of a top pair–the so called lepton + jets $t\bar{t}$ sample–provides yet another opportunity to test the performance of these scale factors. Unlike the dilepton $t\bar{t}$ sample, each event in the lepton plus jets sample should have at least two non-b-jets from the hadronic decay of a W boson, as well as two b-jets

from the top quark decay. The jet flavor composition of this sample is different from both the dilepton $t\bar{t}$ and Z+jets samples used to derive the scale factor. Correctly describing the CSV shape in this sample is a powerful test of these scale factors.

Figure 15 shows the data/MC agreement before and after applying the CSV scale factors derived in this note. Two different selections are consider: The first involves lepton + jets events with exactly four jets, two of which are b-tagged. The second selects events with ≥ 6 jets, three of which are b-taggged. The jet flavor distribution of jets in these two samples is quite different from the samples where the scale factor is derived. Applying the scale factors significantly improves the agreement between the data and MC, and after applying the scale factors, the agreement is quite good. As this level of agreement would be difficult to achieve without having an accurate set of scale factors, we take this as a good sign that the scale factors are correct and that the assessed systematic uncertainties are well justified.

This check can be made even stronger by trying to use kinematic information to separate the jets from selected $t\bar{t}$ events into jets from W decays and b-jets from top quark decays. For this check, we use the lepton + jets sample selected to have exactly four jets with exactly two of them b-tagged, as this would in principle represent a potentially fully reconstructable final state with no extra jets. For events in this selected sample, we consider all possible assignments of jets to partons from the $t\bar{t}$ decay, and for each assignment calculate a χ^2 value as follows:

$$\chi^2 = \frac{\left(M(\ell\nu b) - m_{\text{top}}\right)^2}{\sigma^2_{\text{MC}}(\ell\nu b)} + \frac{\left(M(jjb) - m_{\text{top}}\right)^2}{\sigma^2_{\text{MC}}(jjb)} + \frac{\left(M(jj) - m_{\text{W}}\right)^2}{\sigma^2_{\text{MC}}(jj)},$$

where the σ_{MC} values are taken from the distribution of MC events where the correct parton assignment has been made. See Table 3 for values used. In evaluating different jet-parton assignments, no distinction is made between b-tagged and untagged jets; only the kinematic information is used. The combination with the lowest χ^2 value is selected and the CSV distribution is plotted separately for jets matched to the partons from the hadronic W and jets matched to the b-quarks from the top decay. The comparison between data and MC expectation for these plots is shown in Fig. 16. Again, the agreement is very good both for the LF enhanced jets matched to the W boson and the b-quark enhanced jets matched to the b from the top decay.

Parameter	Value
$\sigma_{\text{MC}}(\ell\nu b)$	$10.02\,\text{GeV}/c^2$
$\sigma_{\text{MC}}(jjb)$	$17.97\,\text{GeV}/c^2$
$\sigma_{\text{MC}}(jj)$	$10.51\,\text{GeV}/c^2$

Table 3: The σ_{MC} values extracted from correctly matched MC events and used in the kinematic fit.

5.3 Comparing to BTV POG Scale Factors

It can be instructive to compare the scale factors derived in this notes to the ones obtained by other methods and recommended by the BTV POG. When doing this comparison, care must be taken to use equivalent numbers. All scale factors provided by the BTV are integral; that is, they provide the scale factor for events passing a cut at a given CSV threshold. The scale factors derived here are differential, in that they are parameterized in exclusive bins in CSV. We convert our differential measurement to an integral one weighted by the CSV distribution from MC for the jet flavor of interest. Figure 17 shows the comparison between the b-tagging SF derived with our approach and the scale factors provided by the BTV POG using $t\bar{t}$ data [8–10].

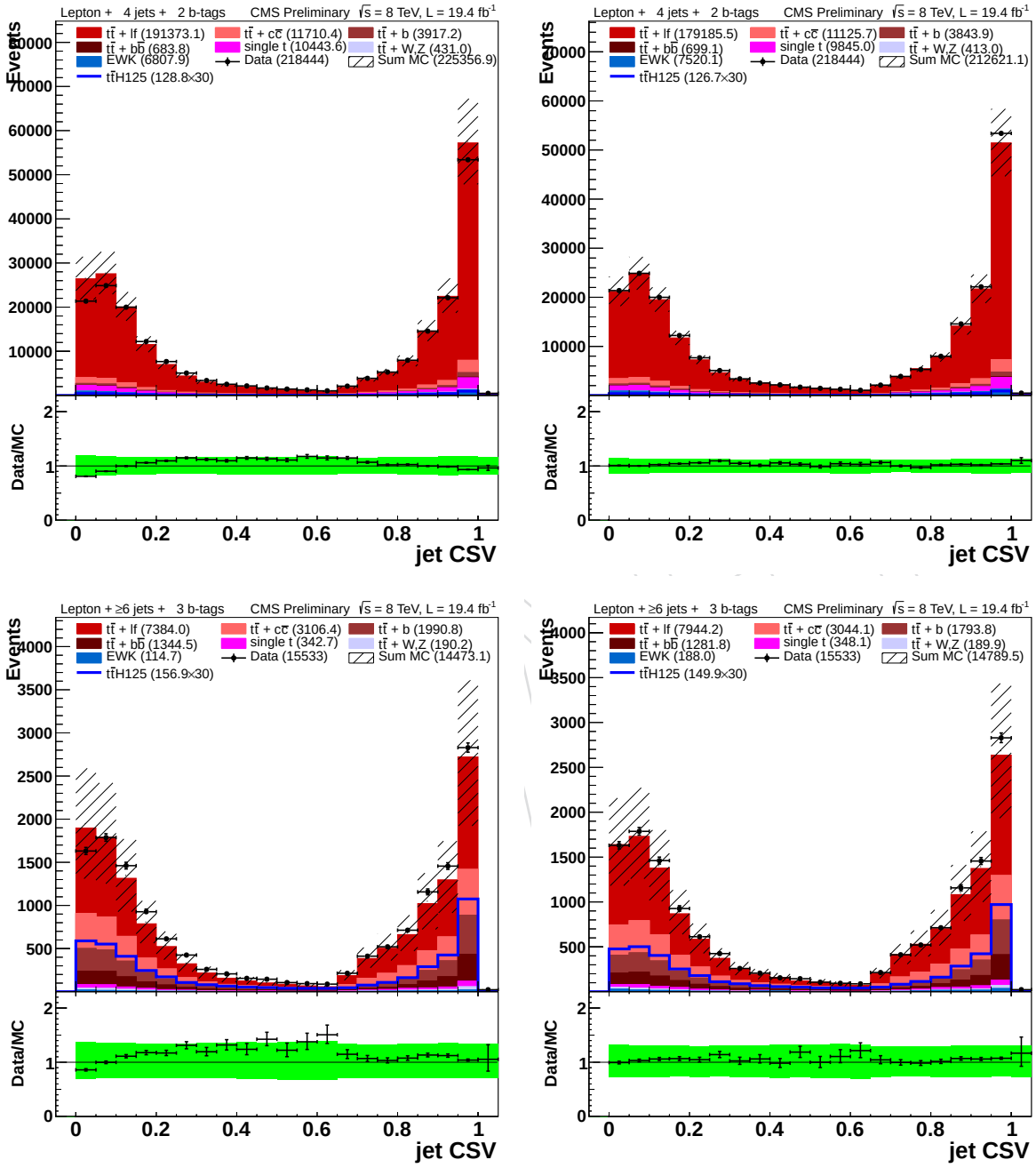

Figure 15: Distribution of jet CSV before (left) and after (right) applying this correction in a control region requiring exactly one tight lepton (electron or muon) with 4 jets and 2 b-tagged jets (top row) or $\geq$6 jets and 3 b-tagged jets (bottom row).

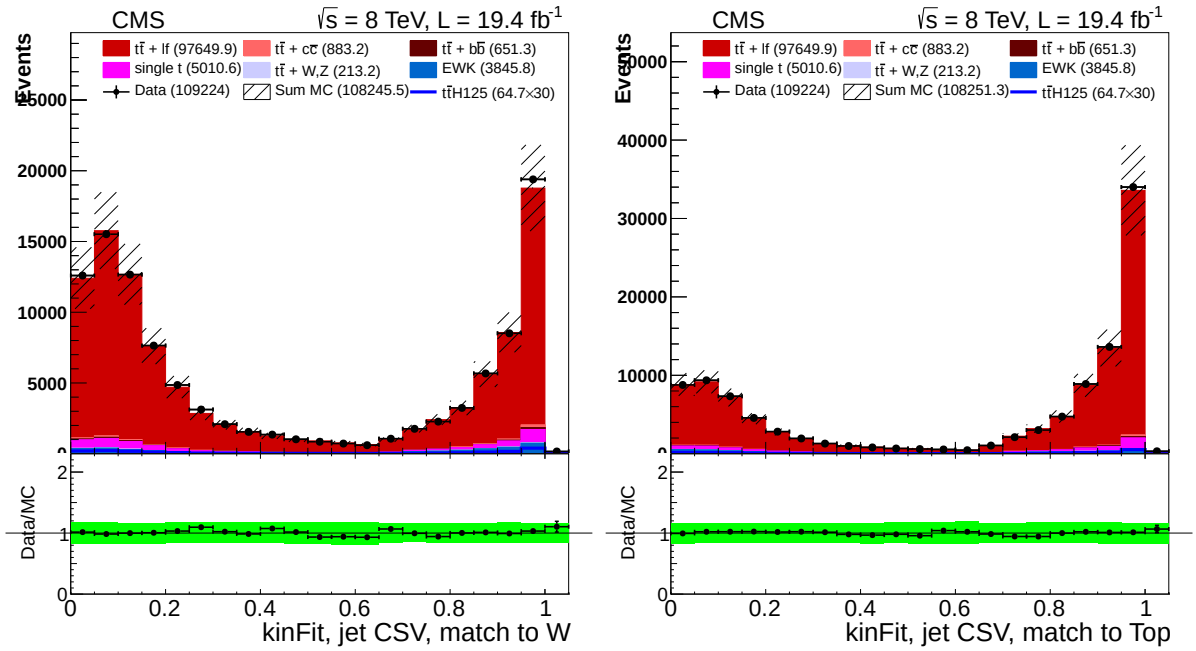

Figure 16: Comparison of the CSV distribution to originating from the W decay (left) or the v-quark from the top decay (right). This assignment is based on the permutation with the best χ^2 from the kinematic fit described above.

The scale factors from the two techniques agree well, although the scale factors derived here have larger uncertainties. Table 4 shows the comparison for different working points. Again the agreement is very good.

Working Point	Tag-and-Probe	BTV POG
CSVL	0.988 ± 0.032	1.008 ± 0.023
CSVM	0.958 ± 0.039	0.963 ± 0.020
CSVT	0.944 ± 0.039	0.947 ± 0.024

Table 4: Comparison of HF scale factors derived in this note to the ones provided by the BTV POG. The CSV threshold range over which the comparisons are performed is determined by the range over with the BTV scale factor is provided.

Figure 18 shows the comparison of the scale factor derived in this note to the one provided by the BTV POG for LF jets for three different η ranges, using the CSVM operating point. For the light flavor, there is more of a difference between the two methods in the scale factor values obtained. Like the HF scale factor, the techniques from this note result in a larger uncertainty, driven by the purity systematic. Nonetheless, we prefer to use the scale factor values derived here. This is because, as shown above, the agreement in overall CSV distribution shape using the scale factors from this note is quite good, and the scale factors provided by this approach are differential, which is critical for the ttH analysis.

We also compared our integrated scale factors for charm jets to the ones from the BTV group which are just the scale factors for b-jets but with twice the size of uncertainty. The scale factors we derived are binned in different jet p_T ranges, but here we show the results integrating over all p_T bins to have a direct comparison to the BTV scale factors. As you can see in Figure 19, the scale factors for charm jets we use are compatible with those from the BTV group within uncertainties.

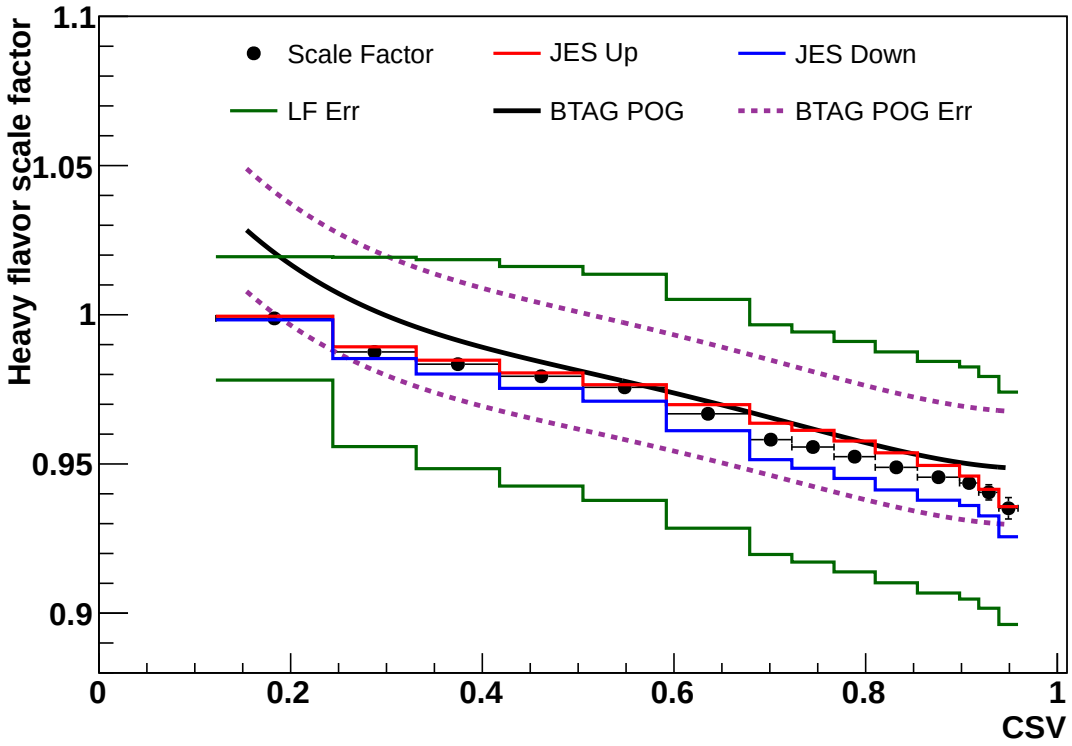

Figure 17: Comparison of our HF scale factors to the BTV POG recommendation.

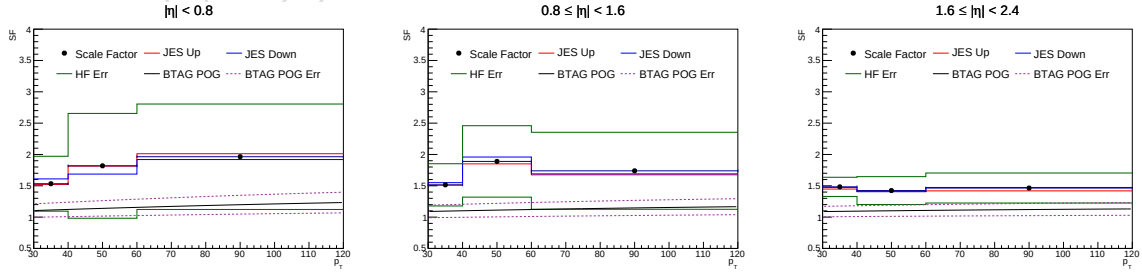

Figure 18: Comparison of the scale factors derived in this note to the ones provided by the BTV POG in three different η bins.

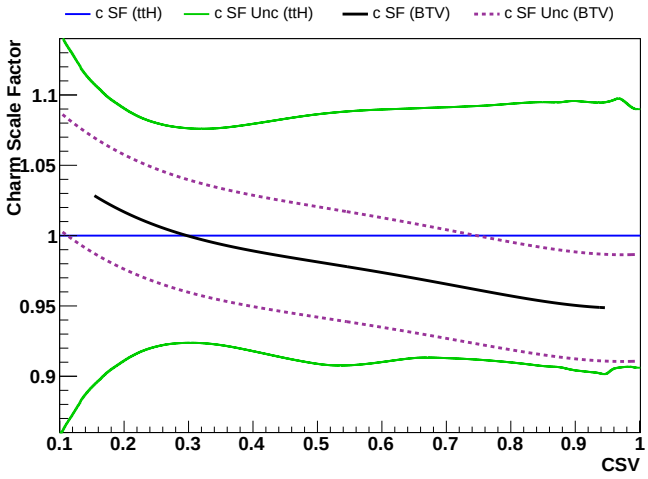

Figure 19: Comparison of the scale factors of charm jets derived in this note to the ones provided by the BTV POG.

Table 5 shows the comparison for different working points. The scale factors from two approaches are consistent within uncertainties.

Working Point	Tag-and-Probe	BTV POG
CSVL	1.000 ± 0.080	1.008 ± 0.046
CSVM	1.000 ± 0.100	0.963 ± 0.040
CSVT	1.000 ± 0.100	0.947 ± 0.050

Table 5: Comparison of charm jets scale factors derived in this note to the ones provided by the BTV POG. The CSV threshold range over which the comparisons are performed is determined by the range over with the BTV scale factor is provided.

5.4 Heavy Flavor Scale Factors Using Other b-tagging Discriminators

The CSV b-tagging discriminator is known to have a complicated structure. The non-smooth features in the HF scale factors calculated above might be related to this. One way to test this hypothesis is to calculate the HF scale factors using a different b-tagging discriminator and then compare to those using CSV. Recall that in the tag and probe method described in Sect. 3, we use the CSV distribution of the probe jet to calculate the b-tagging scale factors. Instead of CSV, we can use a different b-tagging discriminator, such as jet probability (JP), track counting high purity (TCHP), simple secondary vertex high efficiency (SSVHE), or simple secondary vertex high purity (SSVHP) of the probe jet. Everything else during the scale factor calculation is unchanged. The resultant scale factors will be functions of the new b-tagging discriminator used.

Figures 20 and 21 show examples of the distributions of different b-tagging discriminators for the probe jet as well as the corresponding HF scale factors. By comparing the different b-tagging discriminators, we can see that CSV has the most complicated structure; the shape changes dramatically at each working point and gets more complicated as we go to high CSV values. JP has some small structure with a few peaks in the distribution. TCHP, SSVHE or SSVHP is relatively smooth, especially in the region of positive b-tagging discriminator values. We also see features for the HF scale factors using CSV or JP. These features usually show up at bins or adjacent bins to where the discriminator itself has a structure, while for TCHP, SSVHE, or SSVHP, the HF scale factors are relatively smooth.

Figure 22 shows the overlay of the HF scale factors in different p_T ranges for each of the b-tagging discriminators. For a complicated discriminator such as CSV, the differences on the HF scale factors among different p_T bins are relatively large. While for simple discriminators like SSVHE or SSVHP, the differences on the HF scale factors among different p_T bins are relatively small. We also plotted the scale factor dependence on the jet p_T for each CSV bin. All the plots for the 18 CSV bins can be found in Appendix D. The scale factor dependence on p_T does not display "random" behavior. This suggests that the scale factor difference among different p_T bins are not purely statistical.

The CSV HF scale factors do have some features. We found through the above studies that these features are not purely statistical, and they can be explained by the non-trivial structure of the CSV.

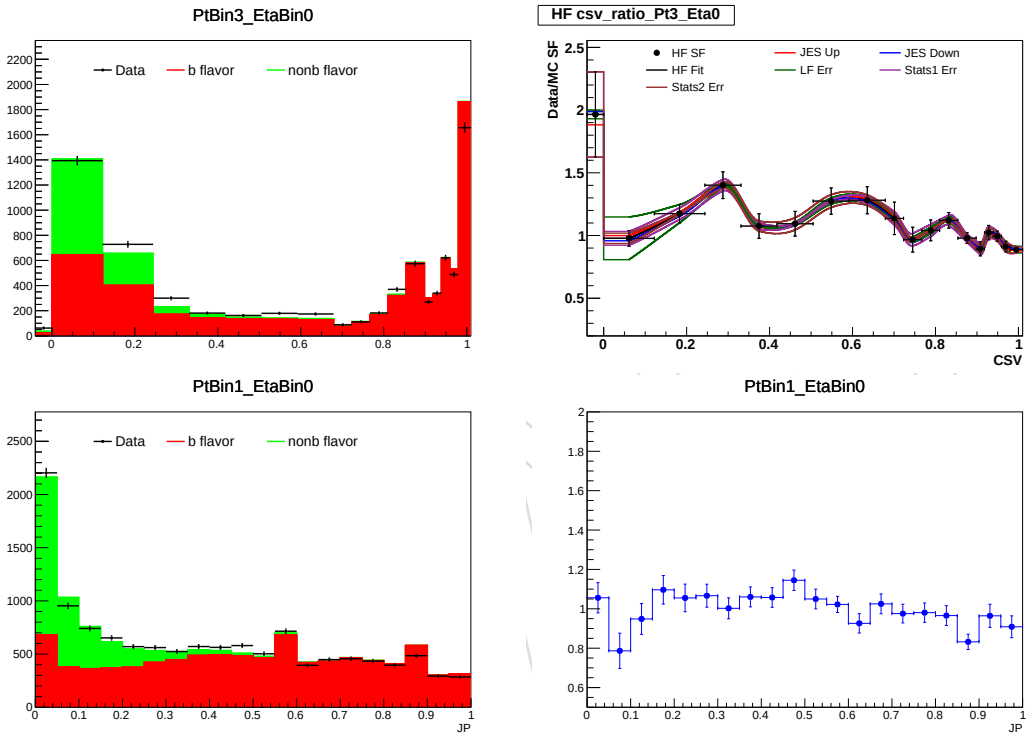

Figure 20: Example HF scale factors using CSV or JP b-tagging discriminator. Left plots show the b-tagging discriminator distribution of the probe jet for data and MC (b flavor and non-b flavor overlayed). Right plots show the corresponding HF scale factors.

5.5 Comparing CSV distribution for b-jets from different sources

Our tag-and-probe method assumes that all the scale factor differences in b-jets can be captured by their p_T and η dependence. This is not special to our method. All the other official BTV scale factors do not have separate scale factors for different sources of b-jets. We investigated the CSV distribution for b-jets from different sources, such as top quark decays, Higgs decays or parton shower. The $t\bar{t}H$ ($m_H = 125$ GeV/c^2) sample was used for this study. Generator level particles are matched to reconstructed jets to identify b-jets from different sources.

Figure 23 shows the comparison of the CSV shape for b-jets from top decays, Higgs decays, or parton shower in different p_T bins. Across all the p_T bins, the CSV shapes for b-jets with different parentage are very similar. There is some small difference for b-jets from parton shower at low CSV values. Figure 24 shows the comparison of the b-jet p_T, corresponding parton p_T and

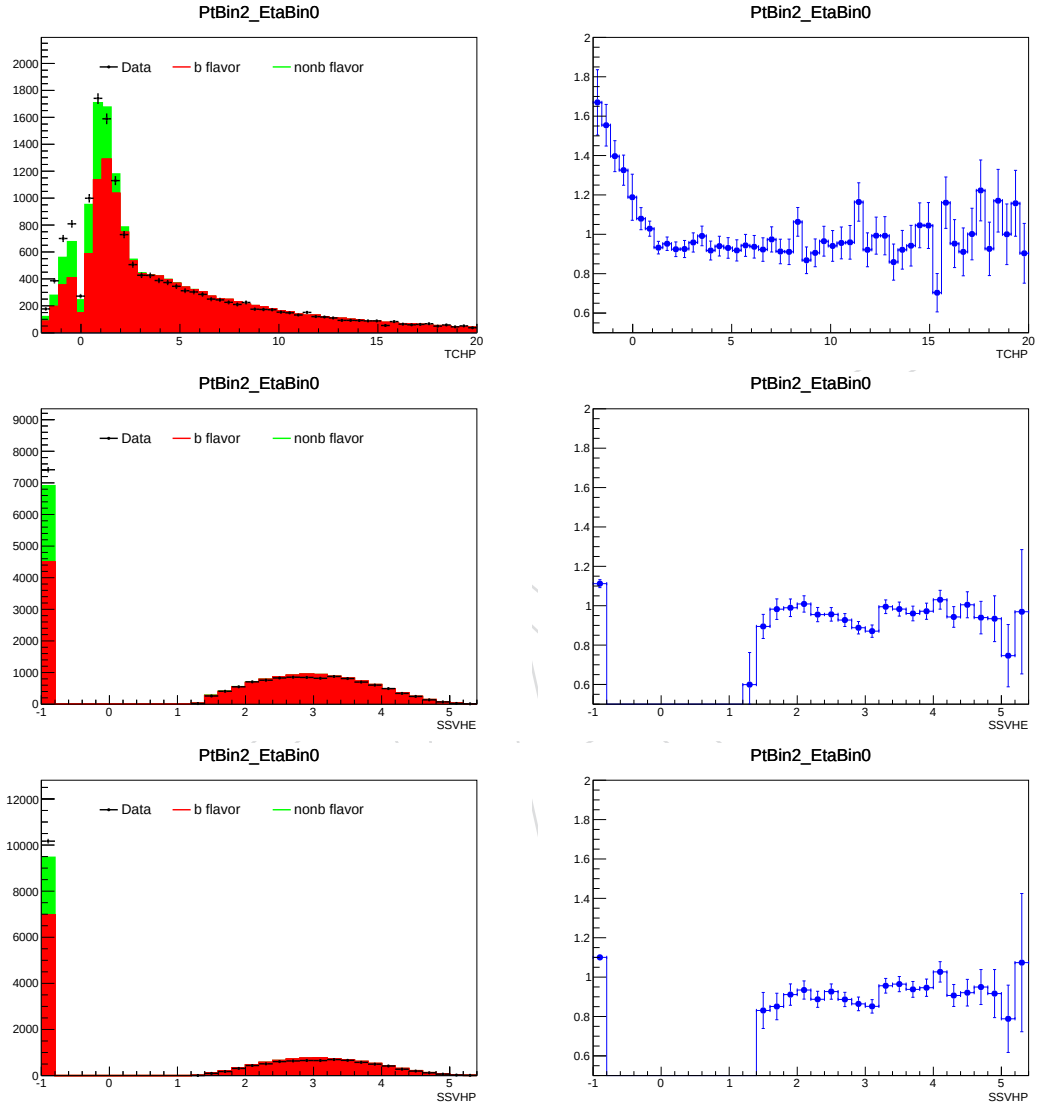

Figure 21: Example HF scale factors using TCHP, SSVHE or SSVHP b-tagging discriminators. Left plots show the b-tagging discriminator distribution of the probe jet for data and MC (b flavor and non-b flavor overlayed). Right plots show the corresponding HF scale factors.

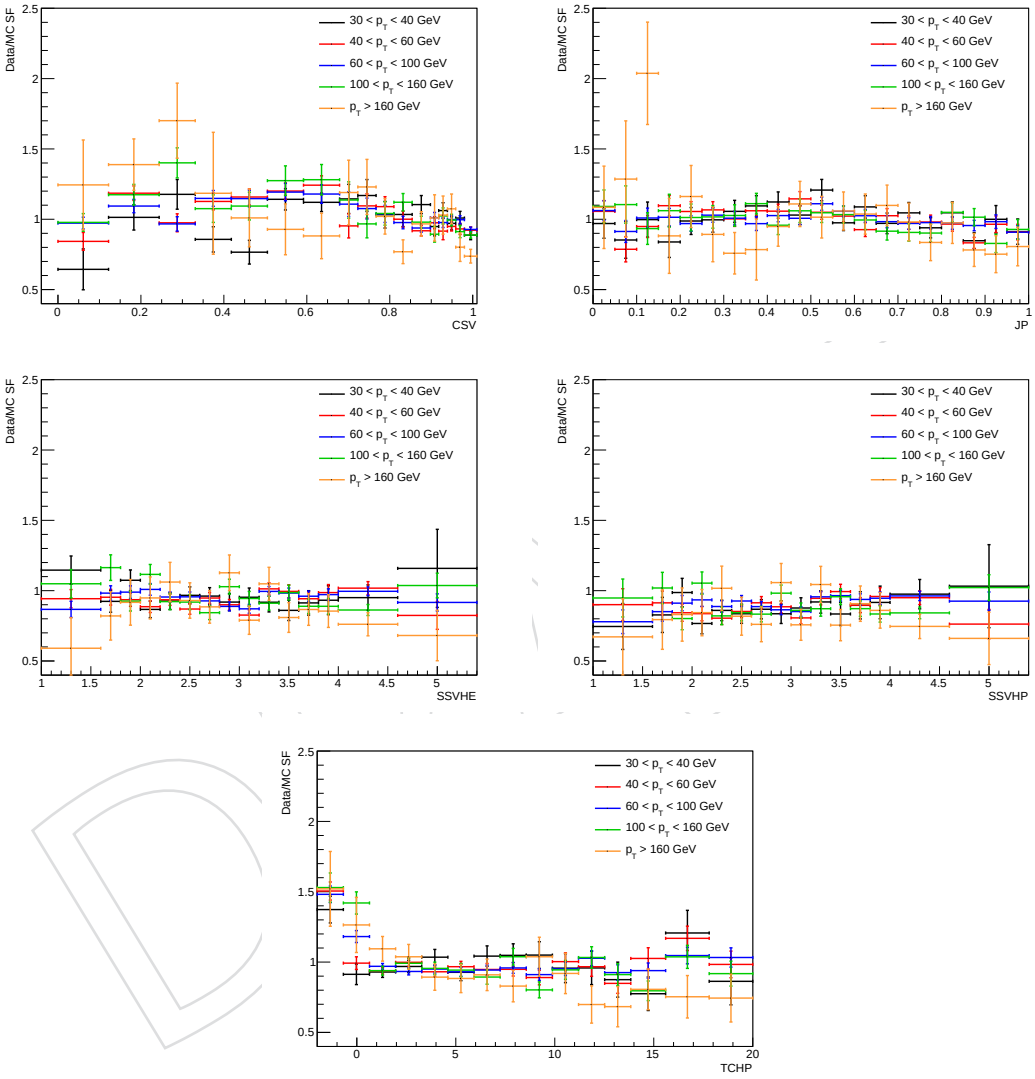

Figure 22: Overlay of the HF scale factors in different p_{T} bins for different b-tagging discriminators.

the ratio of parton p_T to jet p_T distributions for b-jets from different sources. As shown in the ratio plot, for b-jets from parton shower, it often happens that the corresponding b-parton only carries a small fraction of the b-jet p_T. Figure 25 shows the 2D correlation plots where Y-axis is the ratio of parton p_T to b-jet p_T and x-axis is b-jet CSV for b-jets from different sources. For b-jets from parton shower, we can clearly see the correlation between low CSV values and low parton p_T to b-jet p_T ratio. Other than that, all these 2D plots look quite similar.

We have checked the CSV distributions for b-jets from different sources. We found that they are quite similar except for the slight differences for b-jets from parton shower at low CSV values. We further traced the small difference down to the fact that some b-jets originating from parton shower include other partons besides the b-parton. This is a global issue for all b-tagging scale factors parameterized using jet kinematics, since the jet CSV should in principle depend on the parton p_T not the jet p_T. In most cases, the jet p_T is a good approximation of the parton p_T. The difference we observed here is small and it is not expected to cause a significant impact.

6 Conclusions

In this note, we have described a new approach for calibrating the CSV tagger. This approach not only corrects the rate of b-tagged jets predicted by MC, but it also corrects the shape of the overall CSV distribution. Three separate sources of uncertainty are evaluated. The performance of the scale factors is assessed in independent control regions, and they are shown to provide an excellent description of the CSV distributions observed in data.

A LF Scale Factor Results

Figures 26 through 28 show the scale factors for all CSV p_T and η bins, as well as the distributions from which the scale factors are derived.

B HF Scale Factor Results

Figure 29 shows the scale factors for all CSV p_T, as well as the distributions from which the scale factors are derived.

C Comparing charm SF and bottom SF

Figure 30 compares the CSV scale factors for charm and bottom jets in the five different p_T ranges used for the calculation of HF scale factors. The relative uncertainty on the charm scale factor is 2x the relative uncertainty from the bottom scale factor. The largest b-jet SF variations occur at low CSV values, where we expect charm to behave more like light-flavor jets.

D HF scale factor as a function of p_T

Figure 31 shows the HF scale factor as a function of jet p_T for each of the different bins of the CSV distribution.

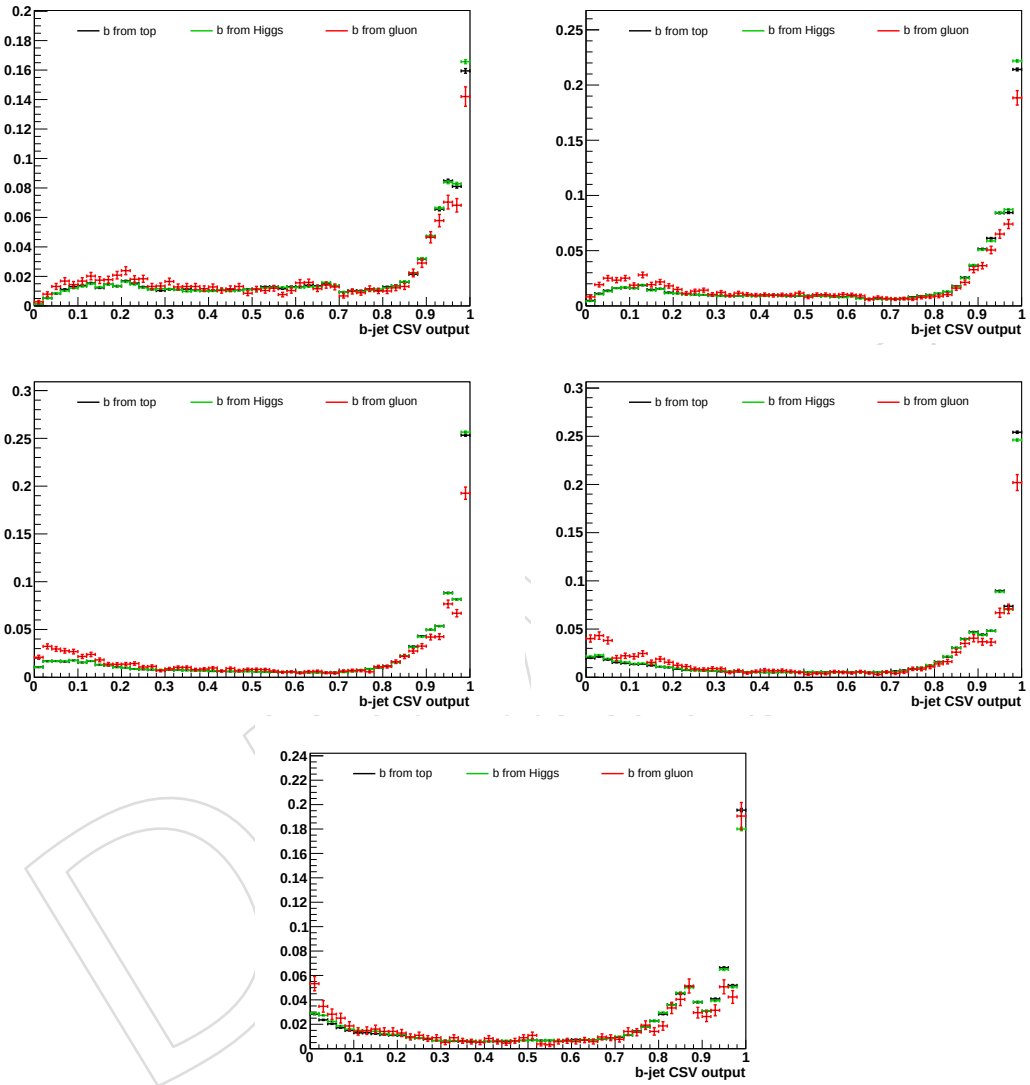

Figure 23: Comparison of the CSV distributions for b-jets from different sources in the 5 different p_T ranges used for HF scale factors calculation. The black, green, and red lines are for b-jets from top decays, Higgs decay, and the parton shower, respectively. All plots are normalized to unit area.

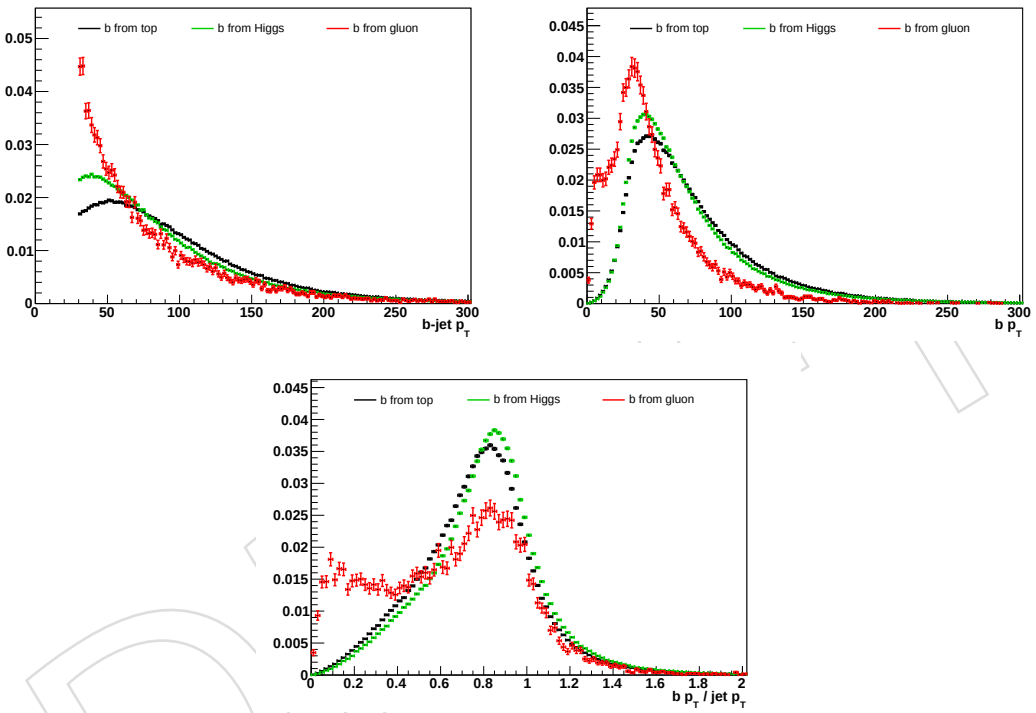

Figure 24: Comparison of the b-jet p_T(top left), corresponding parton p_T(top right) and the ratio of parton p_T to jet p_T(bottom) distributions for b-jets from different sources. The black, green, and red lines are for b-jets from top decays, Higgs decay, and the parton shower, respectively. All plots are normalized to unit area.

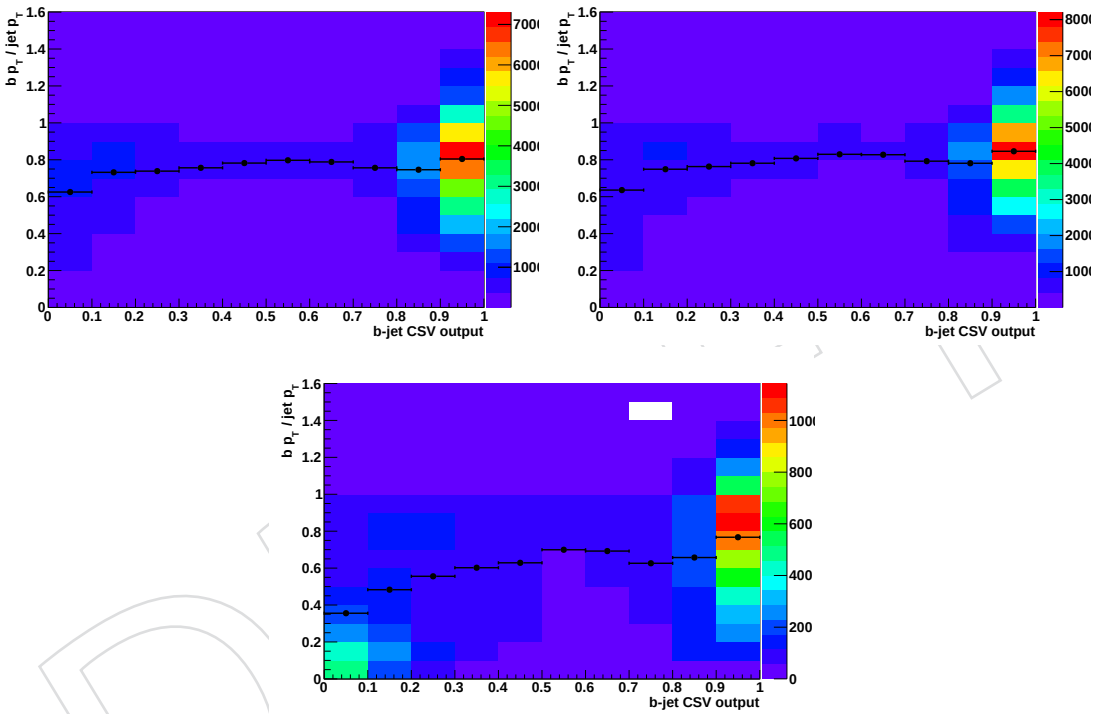

Figure 25: Comparison of the ratio of parton p_T to jet p_T versus CSV 2D plots for b-jets from different sources. Top left is for b-jets from top decays, top right is for b-jets from Higgs decays, and bottom is for b-jets from parton shower.

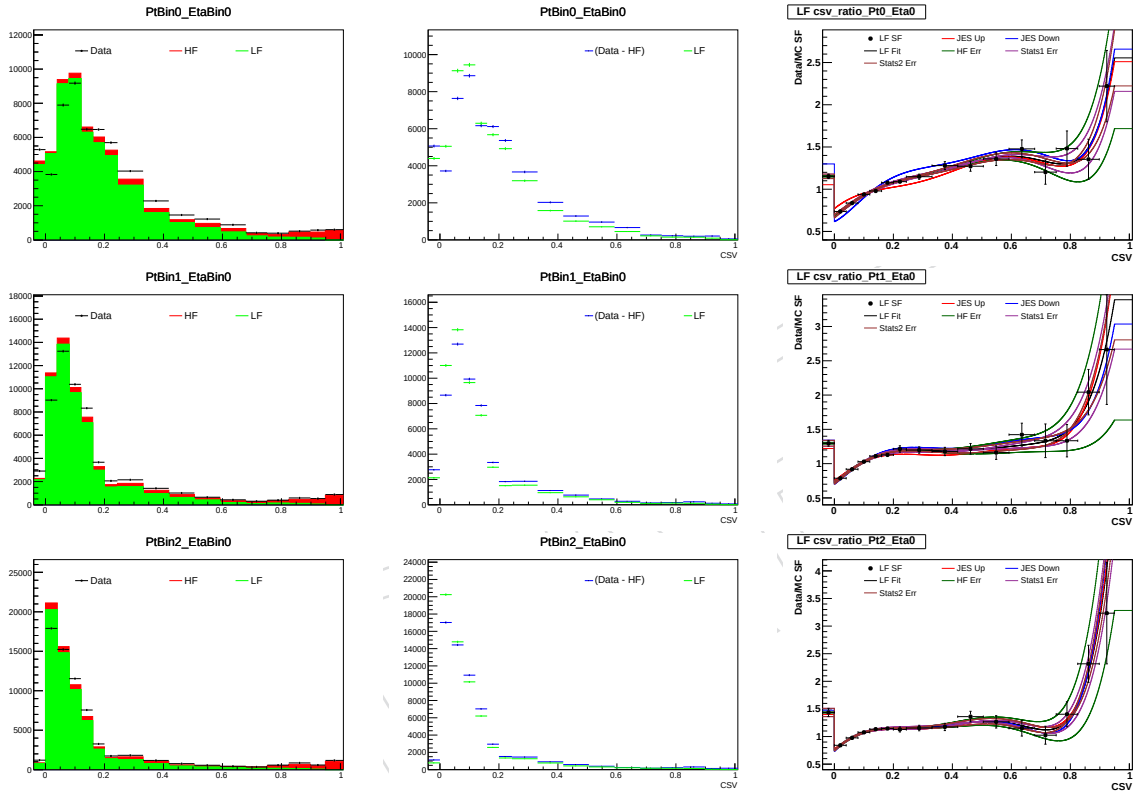

Figure 26: These plots are for probe jets with $|\eta| < 0.8$. Left: Comparison of CSV distribution between data and MC. MC are normalized to data yields. Center: CSV distributions for (Data - MC_{HF}) and MC_{LF}. Right: The CSV SF, including the fitted function and all systematically varied curves. The top row is for $30\,\text{GeV}/c \leq p_T < 40\,\text{GeV}/c$. The middle row is for $40\,\text{GeV}/c \leq p_T < 60\,\text{GeV}/c$. The top row is for $p_T > 60\,\text{GeV}/c$.

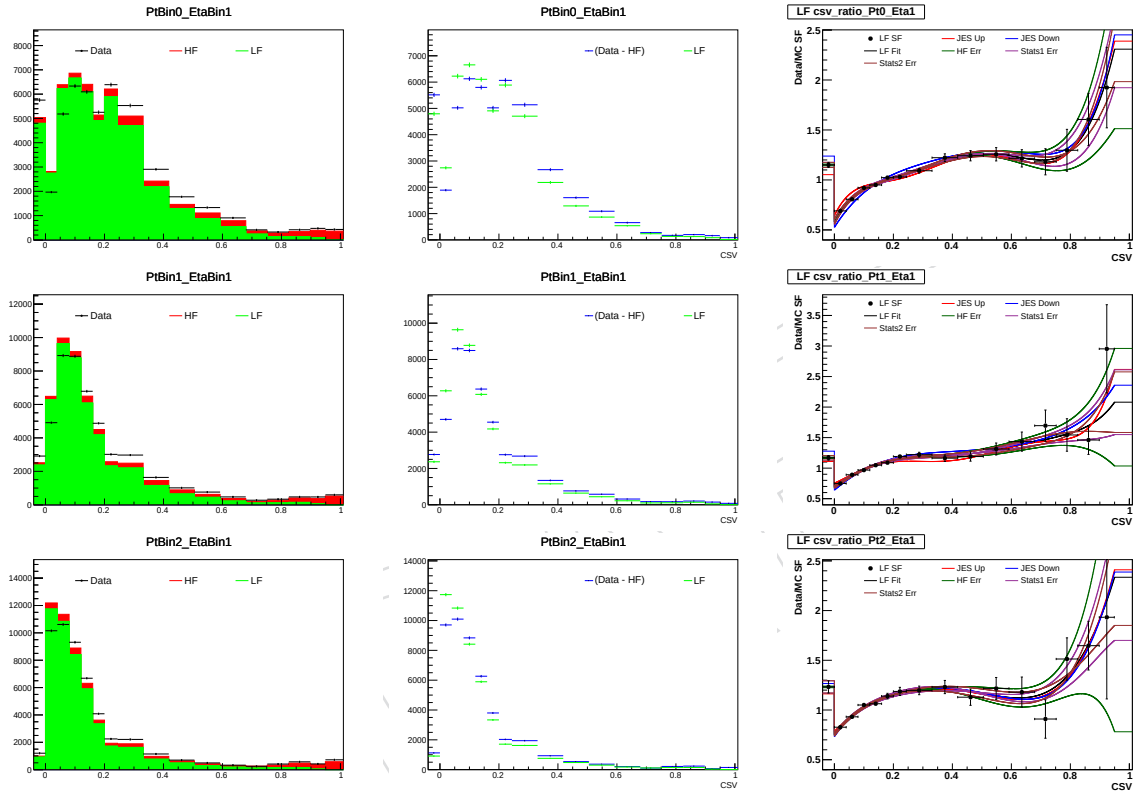

Figure 27: These plots are for probe jets with $0.8 \leq |\eta| < 1.6$. Left: Comparison of CSV distribution between data and MC. MC are normalized to data yields. Center: CSV distributions for (Data - MC_{HF}) and MC_{LF}. Right: The CSV SF, including the fitted function and all systematically varied curves. The top row is for $30\,\text{GeV}/c \leq p_T < 40\,\text{GeV}/c$. The middle row is for $40\,\text{GeV}/c \leq p_T < 60\,\text{GeV}/c$. The top row is for $p_T > 60\,\text{GeV}/c$.

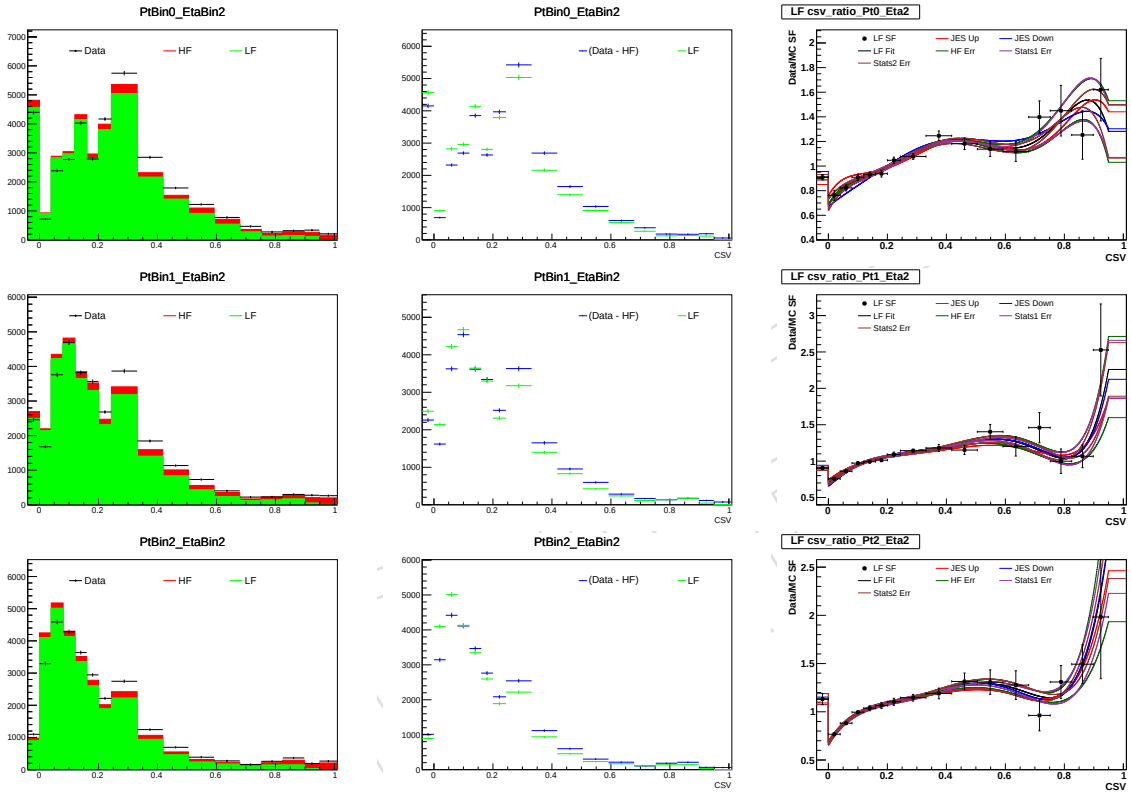

Figure 28: These plots are for probe jets with $1.6 \leq |\eta| < 2.4$. Left: Comparison of CSV distribution between data and MC. MC are normalized to data yields. Center: CSV distributions for (Data - MC_{HF}) and MC_{LF}. Right: The CSV SF, including the fitted function and all systematically varied curves. The top row is for $30\,\text{GeV}/c \leq p_T < 40\,\text{GeV}/c$. The middle row is for $40\,\text{GeV}/c \leq p_T < 60\,\text{GeV}/c$. The top row is for $p_T > 60\,\text{GeV}/c$.

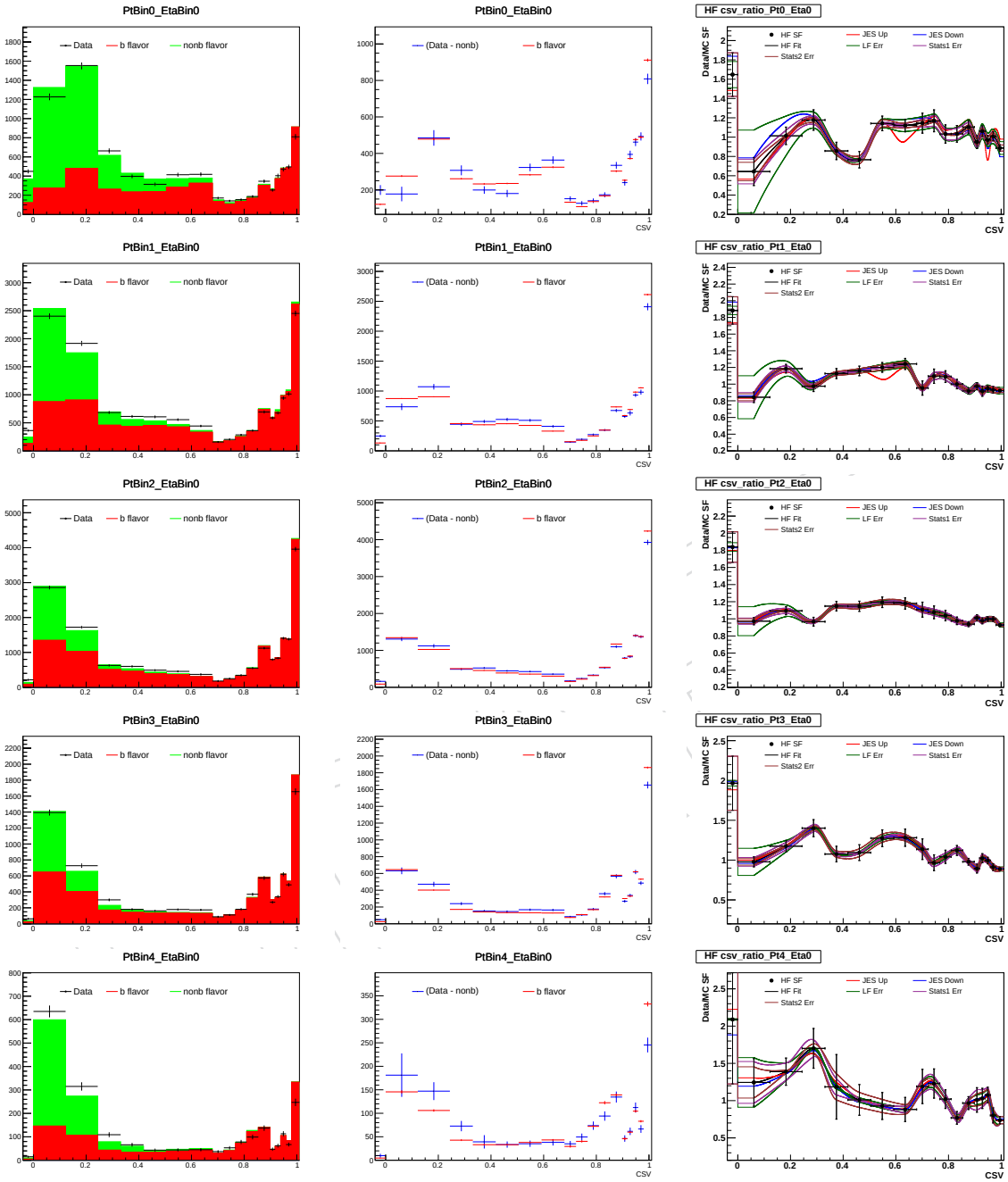

Figure 29: Left: Comparison of CSV distribution between data and MC. MC are normalized to data yields. Center: CSV distributions for (Data - MC_{non-b}) and MC_b. Right: The CSV SF, including the fitted function and all systematically varied curves. The top row is for 30 GeV/c $\leq$ p_T $<$ 40 GeV/c. The next row is for 40 GeV/c $\leq$ p_T $<$ 60 GeV/c. The row after that is for 60 GeV/c $\leq$ p_T $<$ 100 GeV/c. The second to last row is 100 GeV/c $\leq$ p_T $<$ 160 GeV/c. Finally, the bottom row is for p_T $>$ 160 GeV/c.

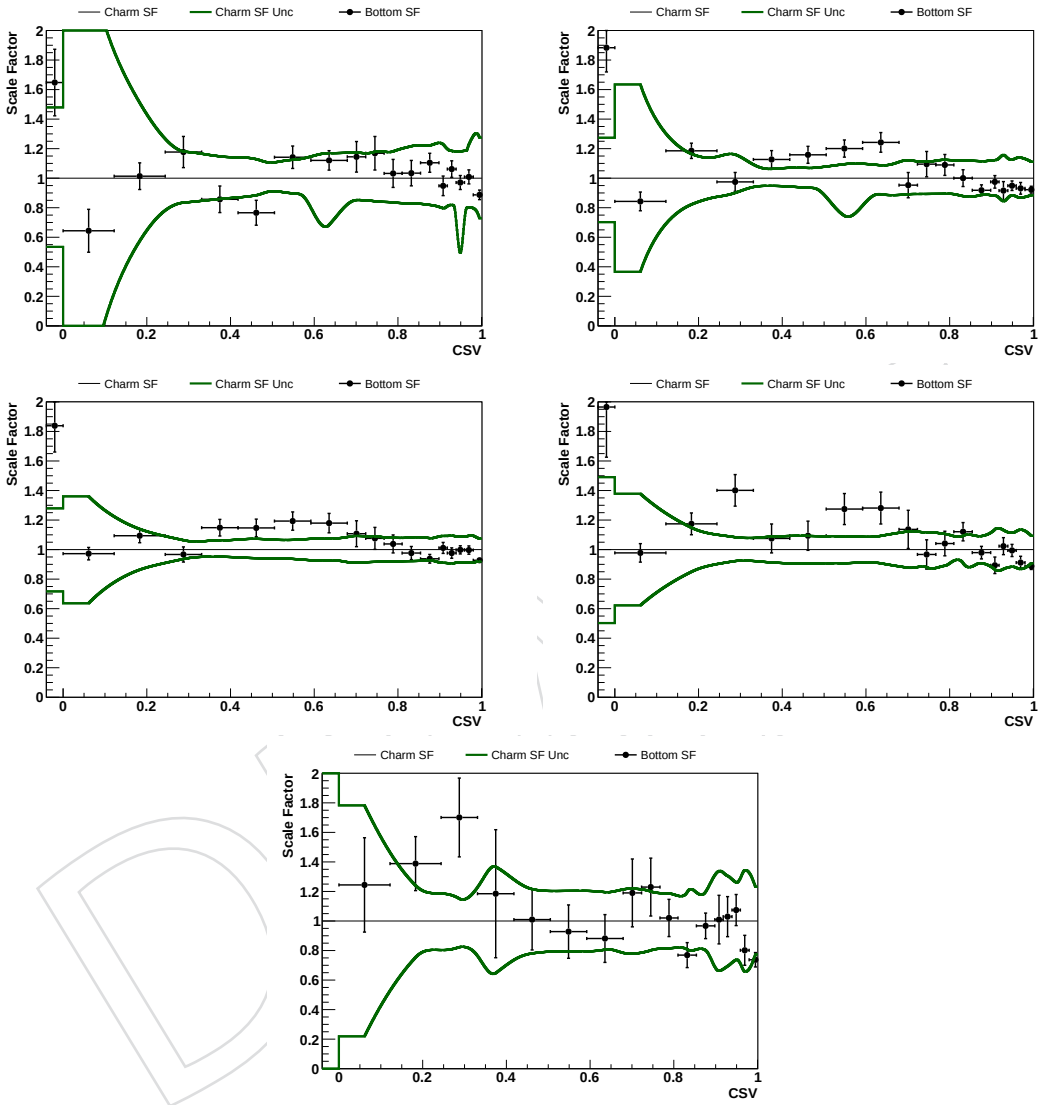

Figure 30: Comparison of the CSV scale factor for charm jets (solid black line, equal to 1.0) and its uncertainty (solid green line) to the scale factor for b-jets (black markers with error bars showing statistical uncertainty) for in the five different p_{T} ranges used for HF scale factors calculation.

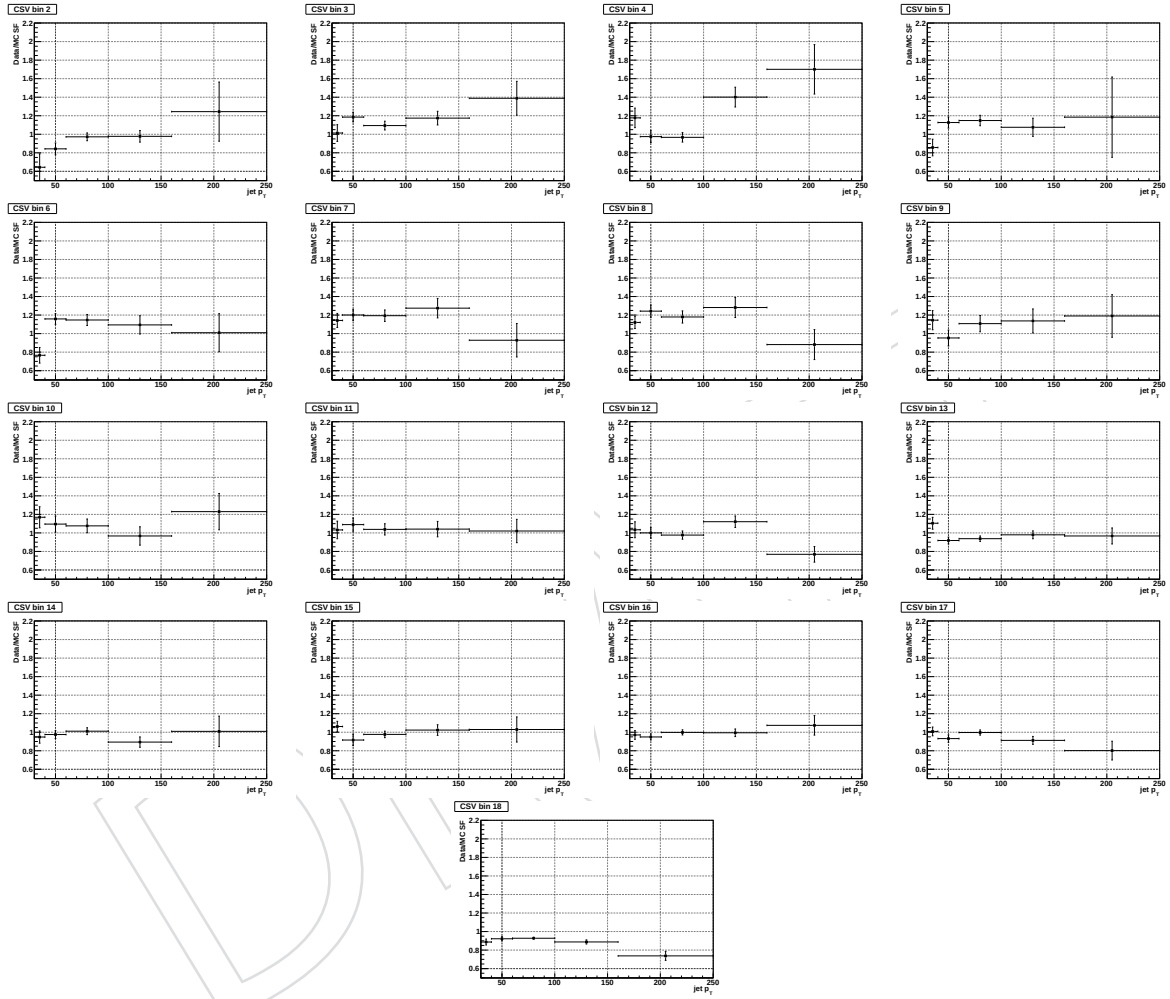

Figure 31: Comparison of the p_{T} dependence of the heavy flavor scale factor for each of the different CSV bins.

References

[1] CMS Collaboration, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS", CMS Analysis Note AN-05-042, (2005).

[2] CMS Collaboration, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS", CMS Analysis Note NOTE-06-014, (2006).

[3] CMS Collaboration Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462.

[4] CMS Collaboration, "Search for Higgs Produced in Association with a Top Quark Pair and Decaying to Bottom Quarks or Tau Leptons", CMS Analysis Note AN-13-145, (2013).

[5] CMS Collaboration, "Combination of b-tagging efficiency measurements in 2012 data at 8 TeV pp collision", CMS Analysis Note AN-12-470, (2012).

[6] CMS Collaboration, "Mistag rate on b-tagging in 2012 data", CMS Analysis Note AN-12-195, (2012).

[7] CMS Collaboration, "Search for the Standard Model Higgs Boson Produced in Association with W and Z and Decaying to Bottom Quarks (Full 2012 dataset)", CMS Analysis Note AN-13-069, (2013).

[8] CMS Collaboration, "Measurement of b-tagging efficiency in semi-leptonic decays of $\mathrm{t\bar{t}}$ events using the Flavor-tag Consistency Method at $\sqrt{s} = 8$ TeV", CMS Analysis Note AN-12-187, (2012).

[9] CMS Collaboration, "Measurement of $R = B(t \to Wb)/B(t \to Wq)$ using a $\mathrm{t\bar{t}}$ dilepton sample selected in proton-proton collisions", CMS Analysis Note AN-12-302, (2012).

[10] CMS Collaboration, "Measurement of the b-tagging efficiency using μ+jets events at 8 TeV", CMS Analysis Note AN-12-432, (2012).

[11] CMS Collaboration, "Measurement of the differential top-quark pair production cross section in the dilepton channel in pp collisions at $\sqrt{s}$ = 8 TeV.", CMS PAS TOP-12-028, (2012).

[12] CMS Collaboration, "Measurement of the differential top-quark pair production cross section in the lepton+jets channel in pp collisions at $\sqrt{s}$ = 8 TeV.", CMS PAS TOP-12-027, (2012).

[13] "Measurement of the Z/gamma* + bb-jets cross section in pp collisions at sqrt(s) = 7 TeV", Technical Report CMS-PAS-SMP-13-004, CERN, Geneva, (2013).

[14] "Measurement of B hadron angular correlations in association to a Z boson", Technical Report CMS-PAS-EWK-11-015, CERN, Geneva, (2012).