TStnSVF Mistag Matrix and B-tagging Scale Factors

Shawn Kwang
Mel Shochet

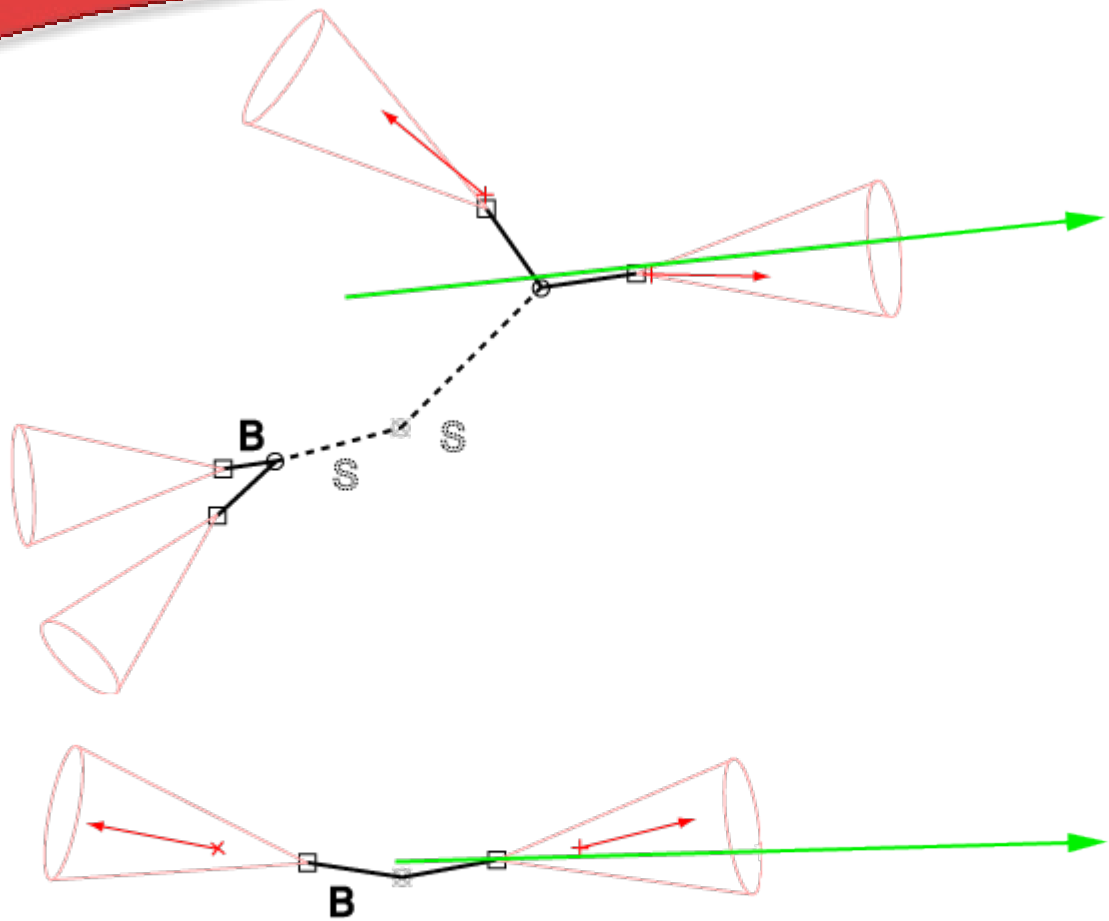
University of Chicago

Introduction

- Motivations – a reminder
- Description of TStnSVF and its difference with SecVtx
- Validation Analyses
  - Mistag Matrix
  - B-tagging Scale Factors – muon method
- Conclusions

Motivation – My Analysis

Top diagram - Signal
H_0->a_0,a_0->bbbar,bbbar
with A0 $c\tau \sim 1$ cm



Bottom diagram - Bkgd
QCD dijet bbbar

- Top diagram is the signal, bottom diagram is a typical dijet bbbar event.
- Primary vertex is the gray X inscribed in a circle.
  - The a_0 (S) is a heavy pseudo-scalar from the Hidden-Valley model.
- A d0 cut on a track (in green) would remove tracks from any signal.

TStnSVF

- TStnSVF is a Stntuple Secondary Vertex Finder.
- The algorithm reproduces the procedures (or strategies) of SecVtx but on an Ntuple level.
  - This is considerably faster than reprocessing Production-level data.
- The program allows the user to adjust the same parameters as SecVtx does.
  - By default it uses the tight level cuts from SecVtx
  - It supports both loose and ultra-tight level cuts as well.

Differences Between Algorithms

- Stntuple does not contain 100% of the information that Production has, and thus my algorithm and SecVtx will have differences.
  - Si databases are not available (or at least not readily available) and thus some track quality cuts cannot be reproduced.
  - However, these wind up being second-order effects.
  - The main source of differences are what I call “resolution” effects where a variable lies on either side of a parameter cut.
  - Example: For a d_0 significance cut of 3.5;
    - one data source has a value if 3.4,
    - the other has a value of 3.6.
- In addition, my algorithm does not recalculate the primary vertex location.
  - Instead it mimics SecVtx by asking the ZVertex algorithm for the best class 12 vertex, and finding the closest PrimeVtx Finder vertex from this “seed” vertex.
  - Functionally, this has the result of asking for the best vertex class 12 in the PrimeVtx Finder block since the two almost always find the same vertex.

Validation

- The purpose of this talk is to validate the TStnSVF b-tagger.
  - This is to prove that not only it works, but that is behaves as expected.
- To this end I have run and am running three analyses that compare TStnSVF with SecVtx.
  - Mistag Matrix Analysis – done
  - B-tagging Scale Factor Analysis – done
  - Mistag Asymmetry Analysis – ongoing
- All three analysis are being run...
  - with data and MC from periods 0-8.
  - with tight-level parameter cuts.
  - without calculating systematic uncertainties.
- It is faster and easier to deal with p0-8, tight level cuts.
- As for systematic uncertainties, the thinking is to see what the results are first, then consider the systematics.

Mistag Matrix Analysis

- Mistag matrix (MM) analysis is considered the easiest of the three and was tackled first. (CDF Note 9280 & 8519)
- It consists of constructing a six-dimensional histogram of the positive and negative tag rates.
  - The numerator is the number of positive (negative) tags.
  - The denominator is the number of taggable jets.
    - Taggable is defined here as a jet with two "good" tracks. Tracks that pass all selection cuts except the d_0 significance requirement.
  - The histogram is parameterized in terms of jet E_T, jet η, number of good tracks, ΣE_T, number of primary verticies, event primary vertex Z position.
- When we run over a dataset, we can plot both the observed tag rate, and then use the MM to determine a predicted tag rate.
- Three different analyses are run:
  - Even-Odd Analysis, TStnSVF only
  - Tagger Comparison, TStnSVF vs SecVtx
  - Independent Sample, TStnSVF only

MM Data

- The MM(s) were constructed using data from the four JET triggers: JET_20, JET_50, JET_70, and JET_100; periods 0-8
  - gjt10x, gjt20x, gjt30x, gjt40x
  - Where x is “d,” “h,” and part of “i.”
- Only unique events were considered. Events for JET_20 and JET_50 were weighted for some runs due to the dynamic prescale of those two triggers.
- The independent sample used was the Jet Sum ET trigger
  - gset0x
- Only a mistag matrix for tight-level cuts was created.
- For each jet, the positive and negative tag rates are recorded. This is the observed rate.
- In addition, the positive and negative tag rates that are predicted from the mistag matrix are calculated as well.
  - The comparison between the predicted and observed rates are presented in the tables on the next slides(s).
  - The corresponding table(s) for SecVtx can be found in CDF Note 8519.

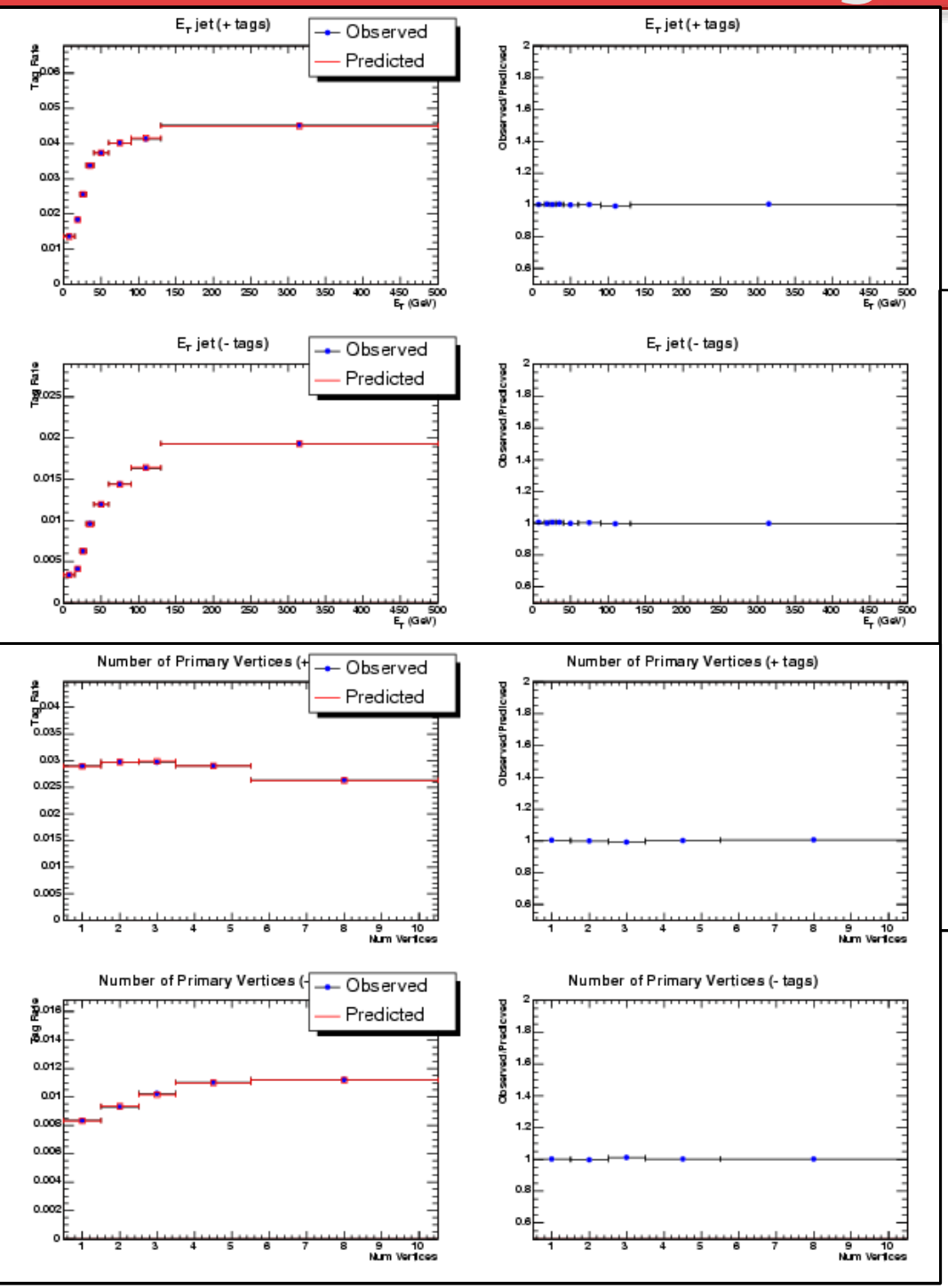
MM Results

- Even Odd Analysis- A mistag matrix was built using only even events from all jet samples using the TStnSVF b-tagger.
- The observed rate is from the odd events in the same jet samples.
- The predicted rate is from the mistag matrix.
  - The ratio is observed/predicted.

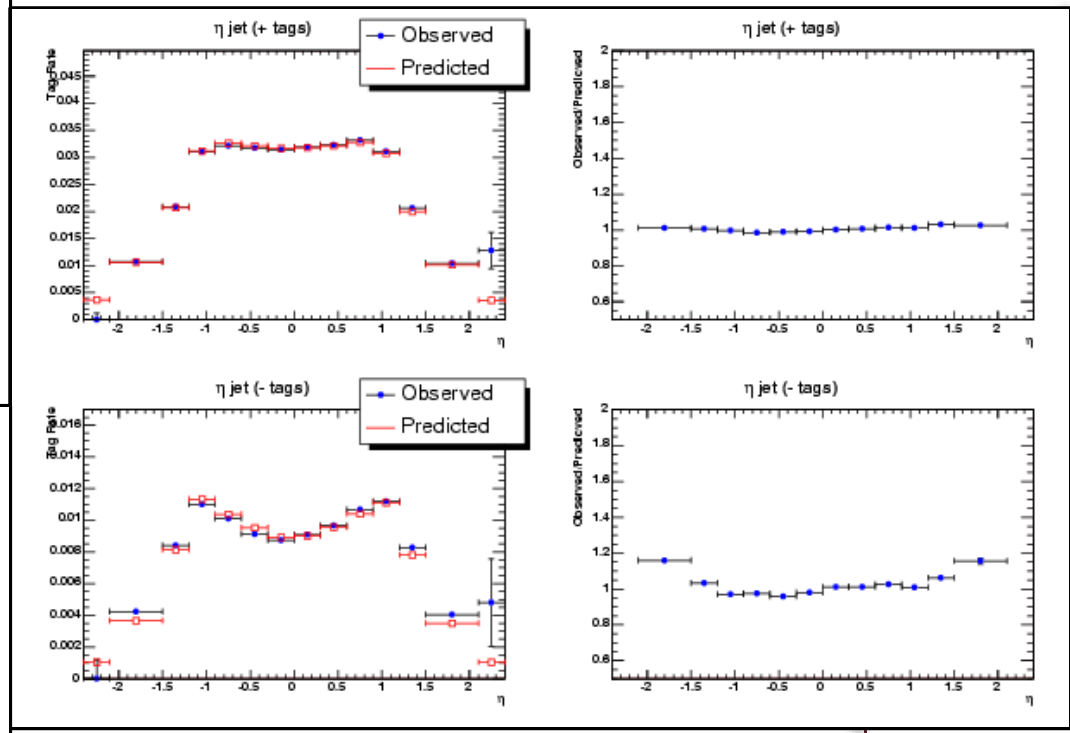
Tag Rate	**JET_20**	**JET_50**	**JET_70**	**JET_100**	**All JET**
Observed Positive	0.01572	0.02850	0.03329	0.03982	0.02929
Stat. Error (+/-)	3.6242E-05	6.2335E-05	6.2755E-05	5.7323E-05	1.2580E-05
Predicted Positive	0.01605	0.02884	0.03338	0.03921	0.02930
Stat. Error (+/-)	5.4716E-06	1.2919E-05	1.3222E-05	1.3558E-05	5.5888E-06
Ratio	0.97920	0.98815	0.99737	1.01553	0.99990
Stat. Error (+/-)	2.2824E-03	2.2063E-03	1.9211E-03	1.5036E-03	4.6986E-04
Observed Negative	0.00324	0.00816	0.01064	0.01471	0.00925
Stat. Error (+/-)	1.6549E-05	3.3701E-05	3.5883E-05	3.5292E-05	7.1432E-06
Predicted Negative	0.00336	0.00831	0.01071	0.01439	0.00924
Stat. Error (+/-)	2.0008E-06	5.7639E-06	5.8695E-06	6.9389E-06	2.6467E-06
Ratio	0.96306	0.98222	0.99257	1.02213	1.00121
Stat. Error (+/-)	4.9578E-03	4.1140E-03	3.3927E-03	2.5017E-03	8.2422E-04

- Overall the positive and negative tag rates are consistent with 1.0.

MM Histograms

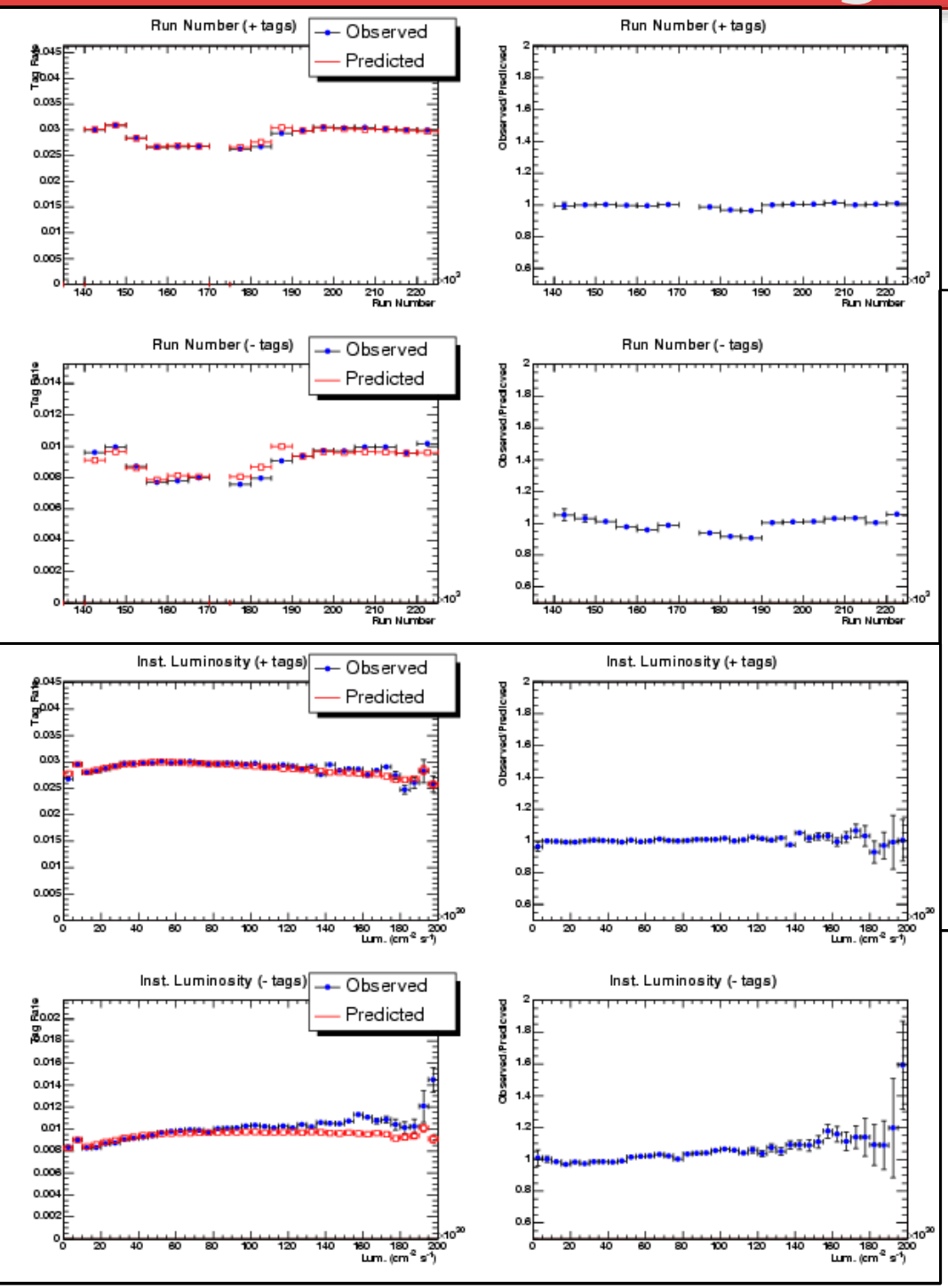


Each of these variables, E_T, η, Num. primary vertices, are parameterized in the mistag matrix.

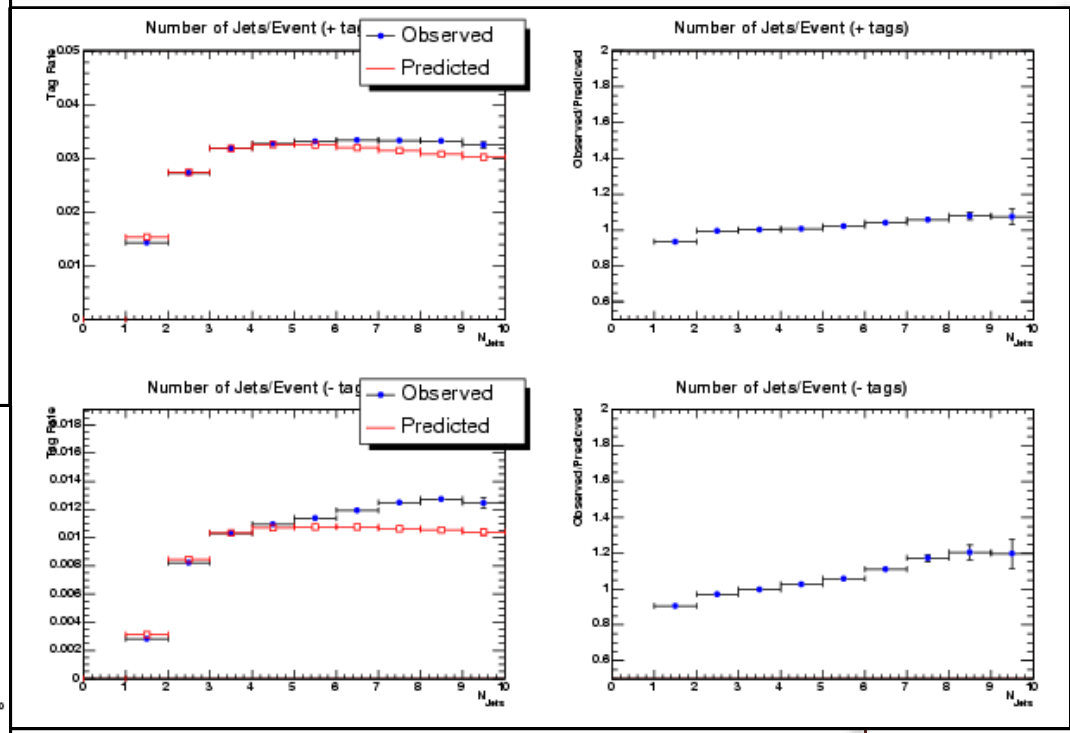


Even-Odd Analysis
All variables are consistent with 1.0

MM Histograms



These variables are *not* parameterized in the mistag matrix, Run number, Inst. Luminosity, and the number of jets/event.



Even-Odd Analysis

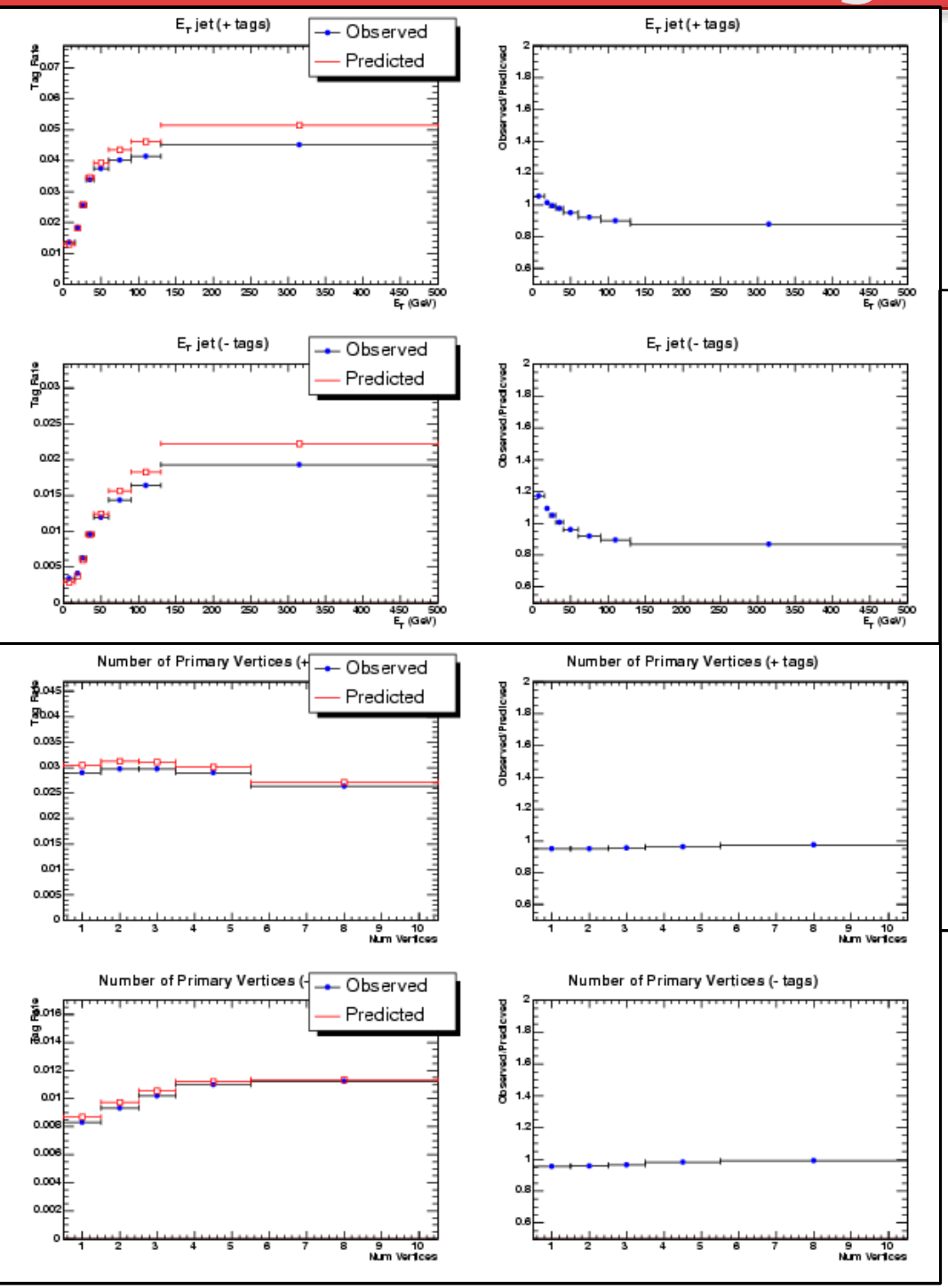
MM Results

- Tagger Comparison Analysis: A mistag matrix was built using SecVtx from all jet samples, all events.
- The observed rate is from the TStnSVF tagger from the jet samples as well.
- The predicted rate is from the mistag matrix.
  - The ratio is observed/predicted.

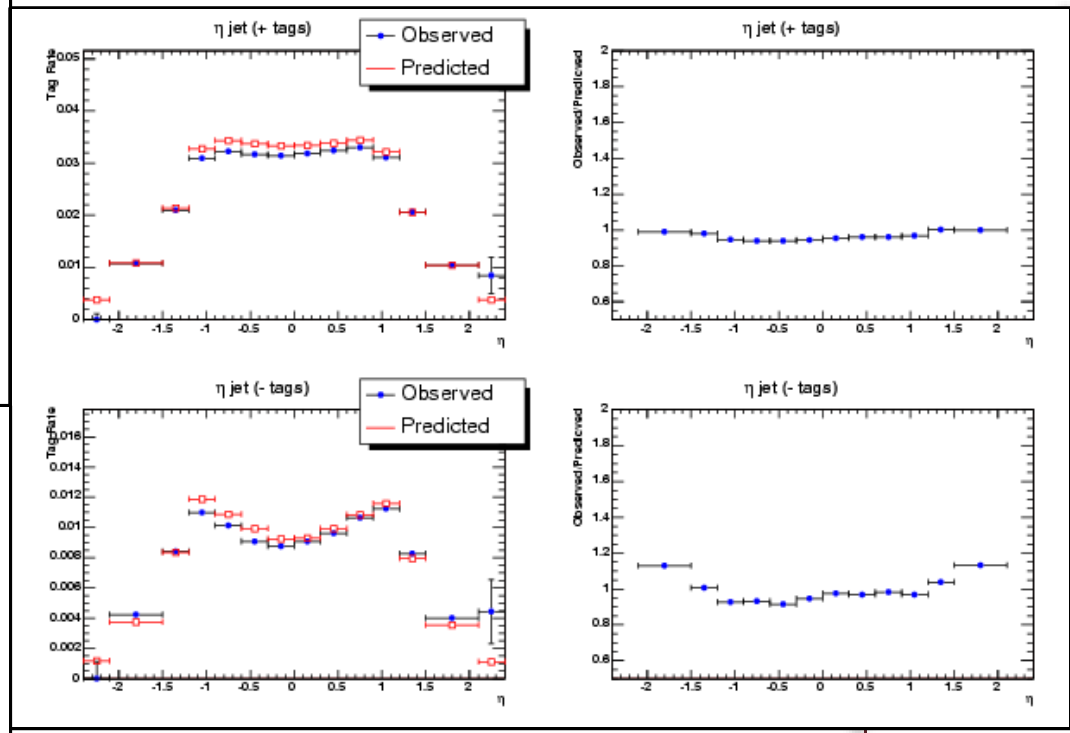
Tag Rate	**JET_20**	**JET_50**	**JET_70**	**JET_100**	**All JET**
Observed Positive	0.01574	0.02840	0.03325	0.03985	0.02928
Stat. Error (+/-)	2.5640E-05	4.3989E-05	4.4345E-05	4.0549E-05	1.2577E-05
Predicted Positive	0.01605	0.02981	0.03506	0.04204	0.03070
Stat. Error (+/-)	3.7826E-06	9.1439E-06	9.5346E-06	9.8640E-06	3.9931E-06
Ratio	0.98024	0.95261	0.94839	0.94780	0.95366
Stat. Error (+/-)	1.6139E-03	1.5042E-03	1.2910E-03	9.8973E-04	4.2798E-04
Observed Negative	0.00323	0.00814	0.01059	0.01472	0.00924
Stat. Error (+/-)	1.1691E-05	2.3802E-05	2.5318E-05	2.4961E-05	7.1382E-06
Predicted Negative	0.00315	0.00841	0.01114	0.01542	0.00960
Stat. Error (+/-)	1.1425E-06	3.5406E-06	3.7455E-06	4.5108E-06	1.6595E-06
Ratio	1.02390	0.96895	0.95079	0.95440	0.96223
Stat. Error (+/-)	3.7243E-03	2.8609E-03	2.2955E-03	1.6428E-03	7.6157E-04

- Overall my TStnSVF b-tagger is a lower in its tag rate than SecVtx.
  - I believe these can be accounted to the differences in the algorithms described earlier.

MM Histograms



Each of these variables, E_T, η, Num. primary vertices, are parameterized in the mistag matrix.



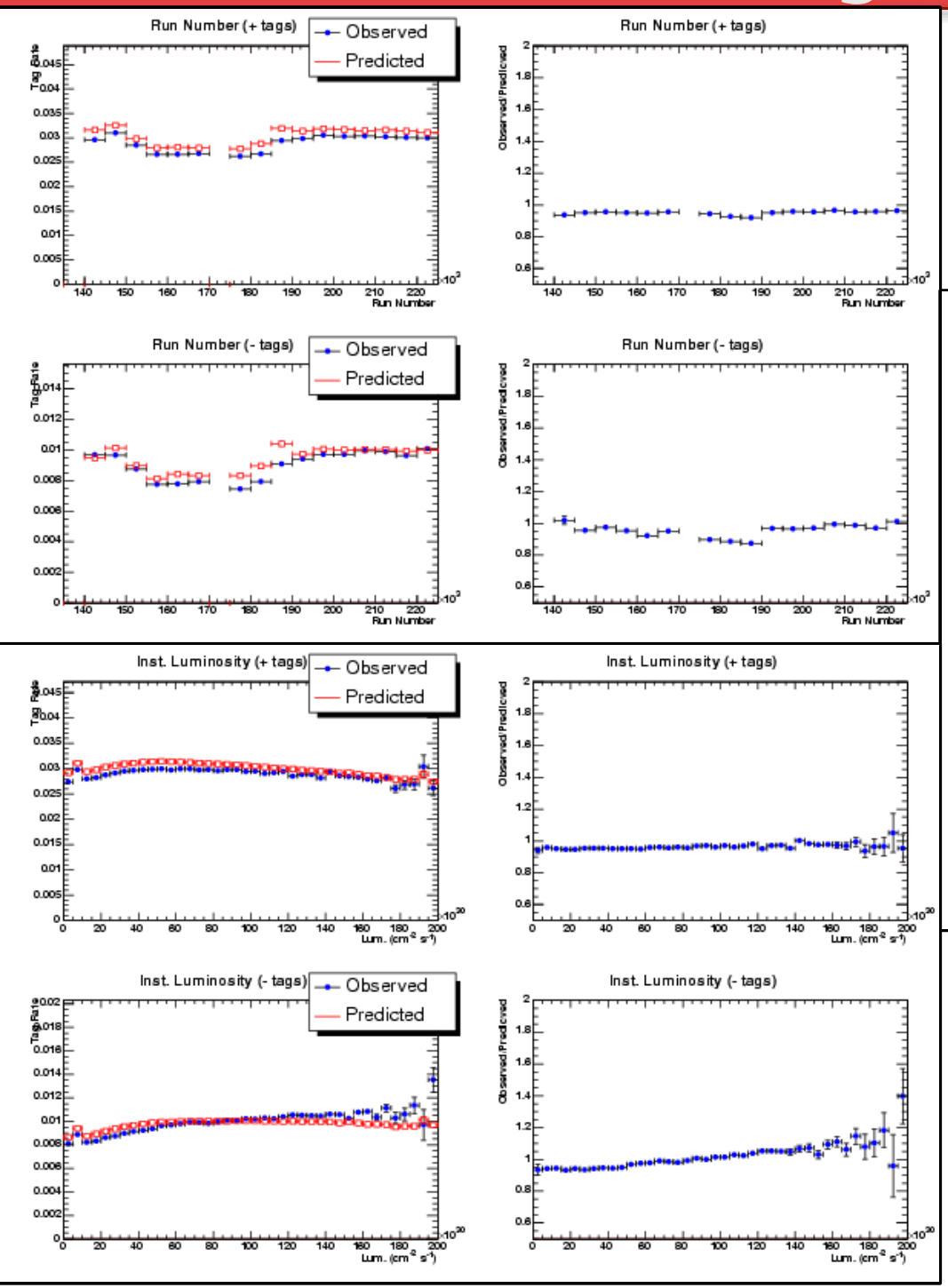
Tagger Comparison Analysis

TStnSvf (Observed) is a bit lower than SecVtx.

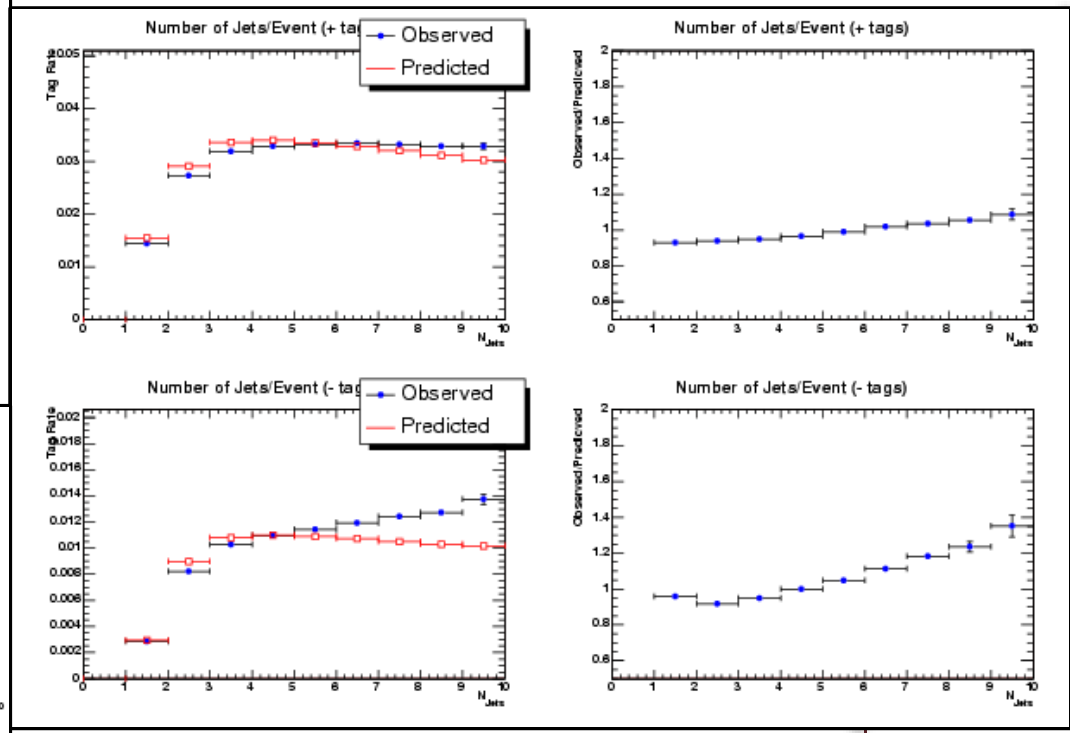
2009-01-26

Shawn Kwang

MM Histograms



These variables are *not* parameterized in the mistag matrix, Run number, Inst. Luminosity, and the number of jets/event.



Tagger Comparison Analysis
TStnSvf (Observed) is a bit lower than SecVtx.

MM Results

- Independent Sample: A mistag matrix was build using TStnSVF from all jet samples, all events.
- The observed rate is from the TStnSVF algorithm run over the Jet ΣE_T trigger.
- The predicted rate is from the mistag matrix.
  - The ratio is observed/predicted.

Tag Rate	All JET	Note8519
Observed Positive	0.03081	
Stat. Error (+/-)	4.0304E-05	
Predicted Positive	0.02951	
Stat. Error (+/-)	8.3832E-06	
Ratio	1.04422	1.03700
Stat. Error (+/-)	1.3977E-03	2.0000E-03
Observed Negative	0.00864	
Stat. Error (+/-)	2.1580E-05	
Predicted Negative	0.00901	
Stat. Error (+/-)	3.8978E-06	
Ratio	0.95890	0.93700
Stat. Error (+/-)	2.4318E-03	3.0000E-03

- The numbers on the right are from CDF Note 8519.

MM Conclusion

- Our conclusion is that my TStnSVF b-tagger performs as expected and computes comparable results to SecVtx.
  - My SecVtx calculated numbers are also comparable to the ones in CDF Notes 8519.

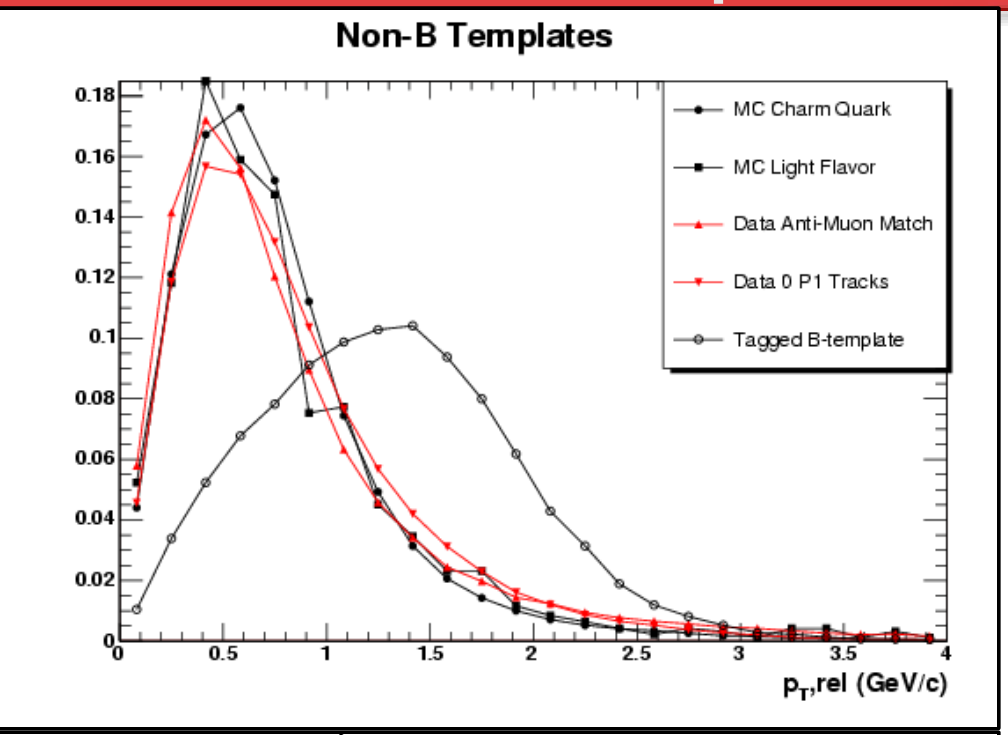
B–Tagging Scale Factor

- For the next validation analysis I calculated the B-Tagging Scale Factor (SF) using the muon method. (CDF Note 8640)
- What we are trying to calculate the is efficiency of the b-tagger in Data divided by the efficiency of the b-tagger in MC.
  - Scale Factor = ϵ^{Data}_{tag} / ϵ^{MC}_{tag}
- We can easily calculate ϵ^{MC}_{tag}
  - ϵ^{MC}_{tag} = Number of tagged b-jets in MC / Total number of b-jets in MC
- For the data (ϵ^{Data}_{tag}) this is more difficult.
  - We know the number of tagged jets not the number of tagged b-jets. We need to determine how many of the former in the data sample came from a b-quark.
  - We have MC which can tell us if a jet comes from a b-quark or not.
- The analysis consists of looking at dijet pairs, where one jet is identified with having a muon in it (μ-jet) and the other is tagged with the loose tagger (the away-jet). See CDF Note 8640 for details.
- The $p_{T,rel}$ between the muon and the jet is used as the primary variable in this analysis.
  - This variable is different for both b-quark jets and non-b quark jets.
  - B-quark and non b-quark templates were constructed using this variable.
  - The data was fitted using ROOT's TFractionFitter to determine the fraction of jets (in the data), both tagged and untagged, from a b-quark.

SF Data

- For this analysis all the data comes from the 8 GeV muon calibration trigger: MUON_CMUP8.
  - bmcl0x
  - Where x is “d,” “h,” and part of “i.”
- This trigger uses dynamic prescaling so the data was weighted by the dynamic prescale value for a given run section.
- For the MC, two sets were used.
  - btopla – Pythia QCD bb with e(μ) p_T = 18 GeV/c
  - btopqb, btoprb – Pythia QCD dijet
- $p_{T,rel}$ histograms were constructed for data, b-quark templates, and four non b-quark templates; two MC driven, two data driven.
  - c-quark jets, from btopla
  - light-flavor jets, from btop[qr]b
  - anti-muon jet match, from bmcl0x
    - These are jets where the muon has the opposite CMU Δx and CMP Δx requirements.
  - zero pass1 tracks in the jet, from bmcl0x
- Because this analysis requires loose-level tags, the SFs were calculated for both the tight and the loose level parameter cuts.

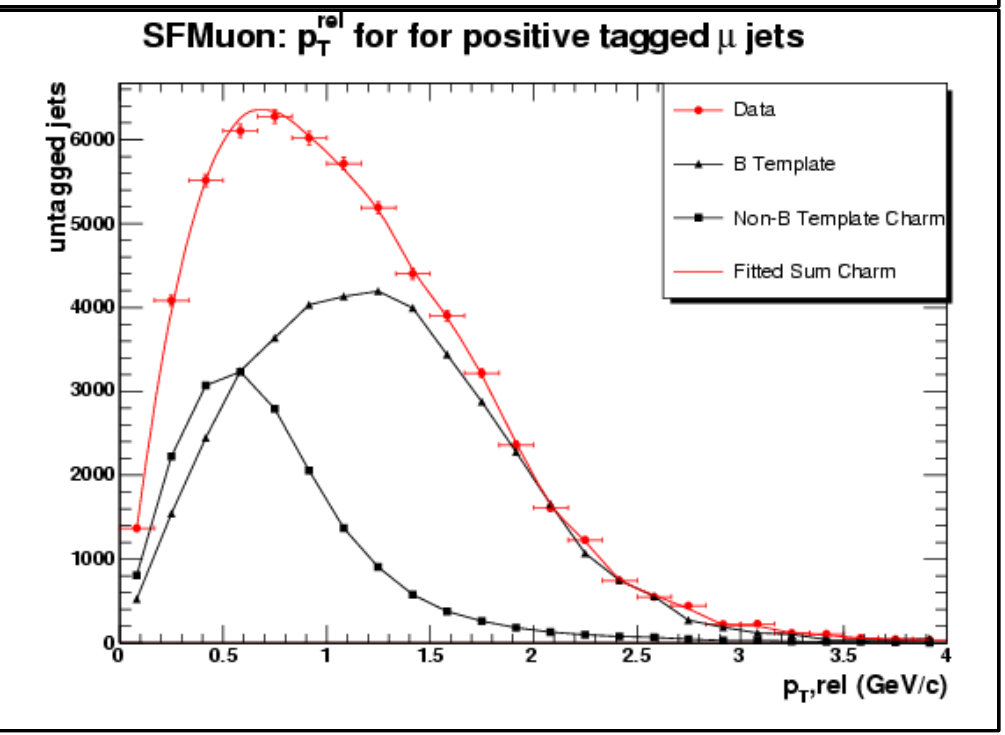
SF Template Example

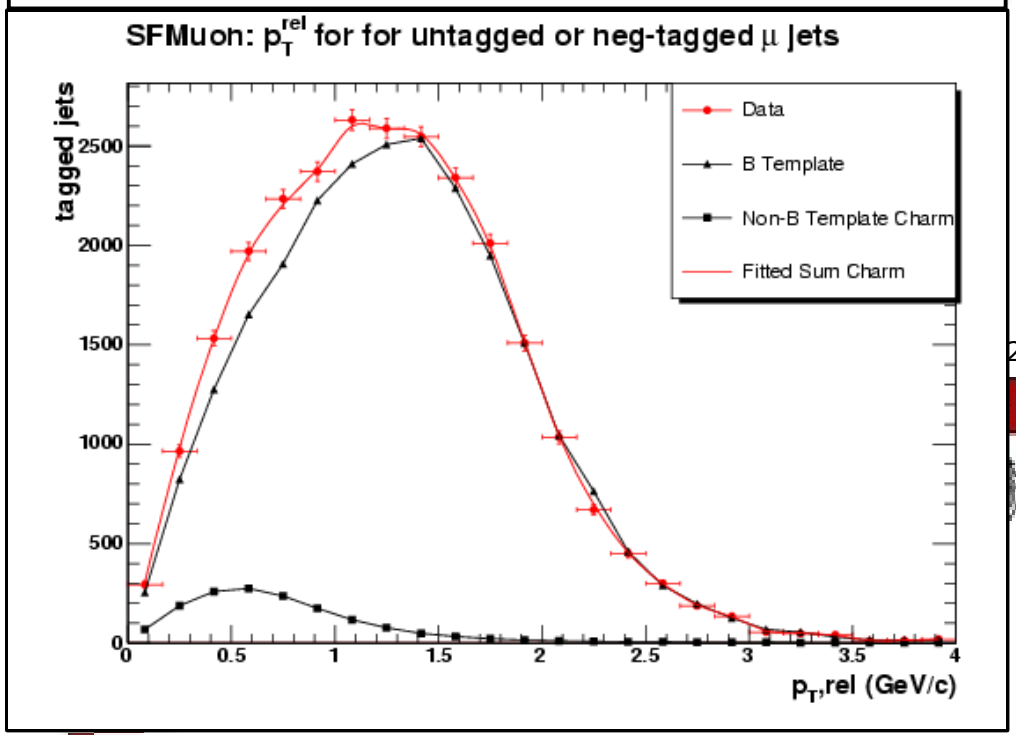


Upper-Left:
The four Non-B Templates for all events and the B Template for positively tagged μ-jets.

Lower-Left:
The Scale Factor fraction fit for the Charm Non-B template for μ-jets without a positive tag.

Lower-Right:
The Scale Factor fraction fit for the Charm Non-B template for μ-jets with a positive tag.





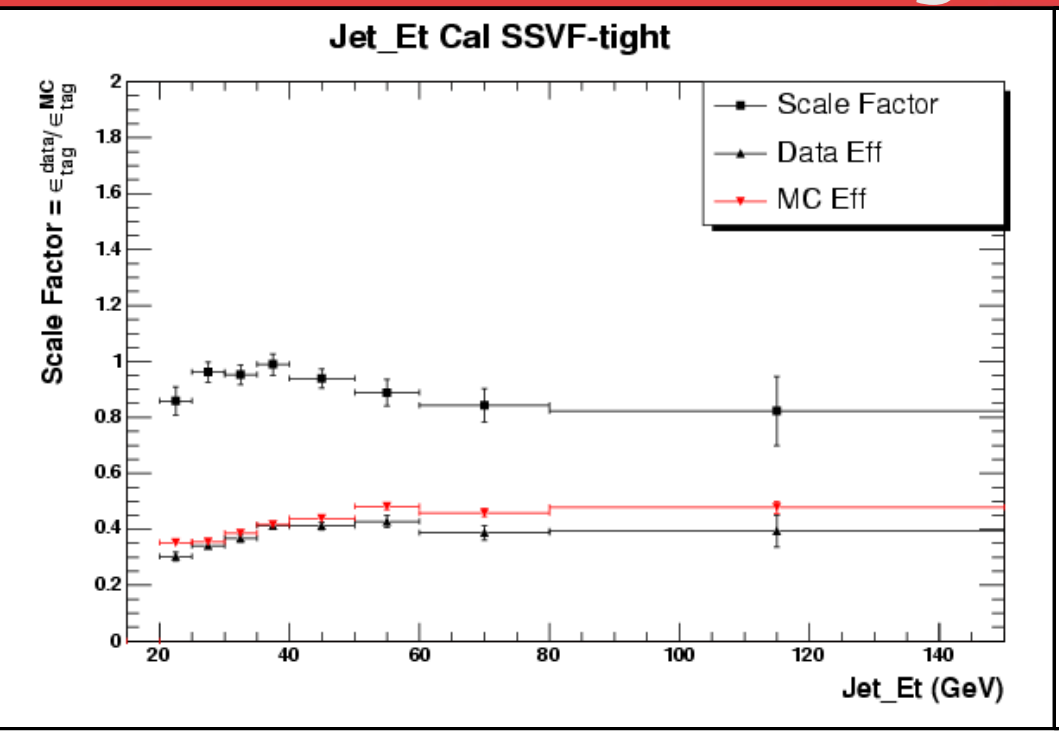
SF Results

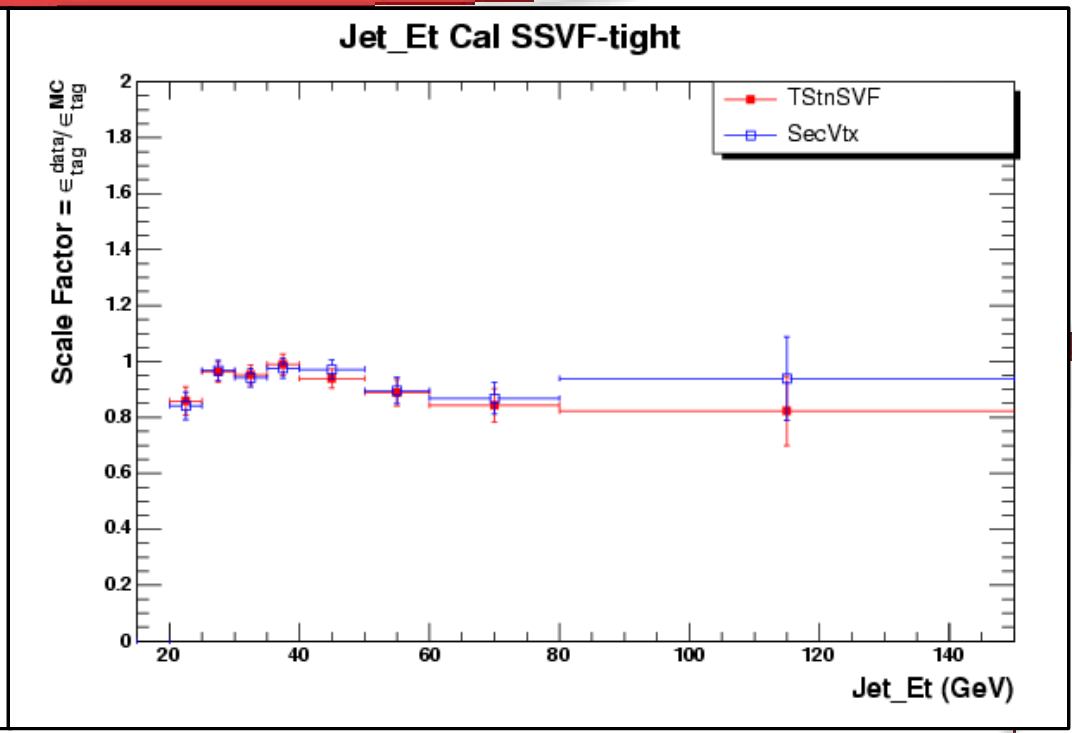
- B-Tagging Scale Factor results for calorimeter based jets, and tight-level parameter cuts

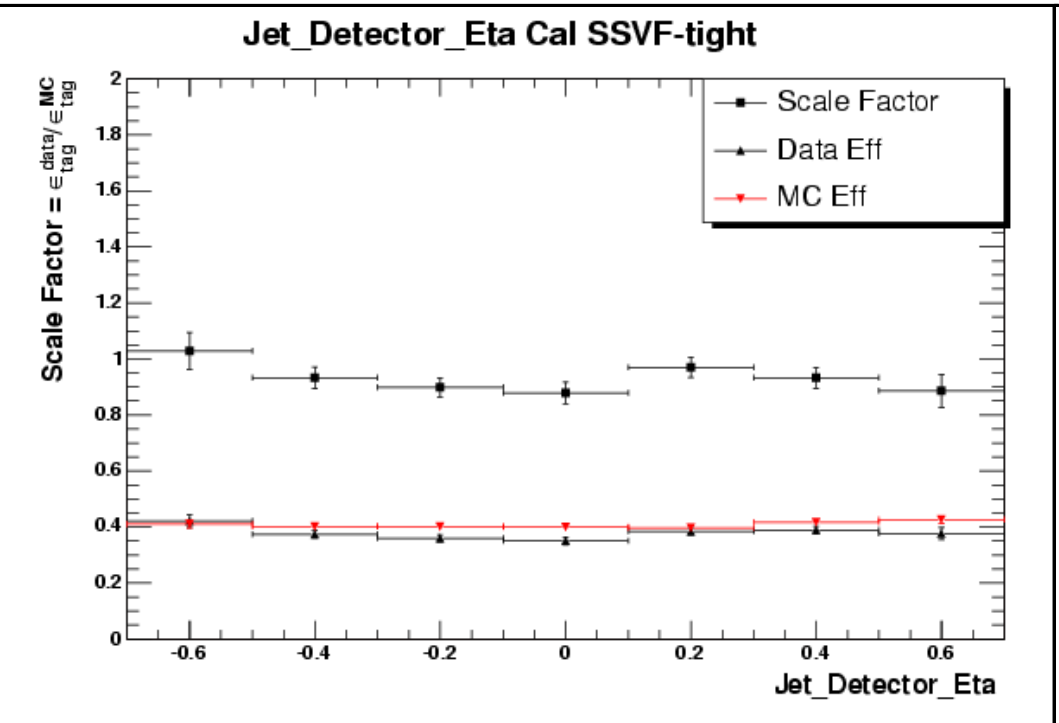
B-Tagging SF for Cal-jets	TStnSVF – Tight			SecVtx – Tight		
	Value	Stat Err	Sys Err	Value	Stat Err	Sys Err
Total number of muon jets considered	85478			84200		
Number of not positive tagged jets	59528			57596		
Number of positive tagged jets	25950			26604		
Data Efficiency, charm template	0.37218	0.00536		0.39189	0.00545	
Data Efficiency, light flavor template	0.37979	0.00673		0.39884	0.00685	
Data Efficiency, Anti-Match template	0.37063	0.00525		0.39095	0.00534	
Data Efficiency, 0 P1 Tracks template	0.38723	0.00577		0.40844	0.00586	
Avg Data B-Tagging Efficiency	0.37746	0.00578	0.00764	0.39753	0.00588	0.00808
MC B-Tagging Efficiency	0.40412	0.00315		0.42205	0.00315	
B-Tagging Scale Factor	**0.93402**	**0.01606**	**0.01892**	**0.94190**	**0.01561**	**0.01915**

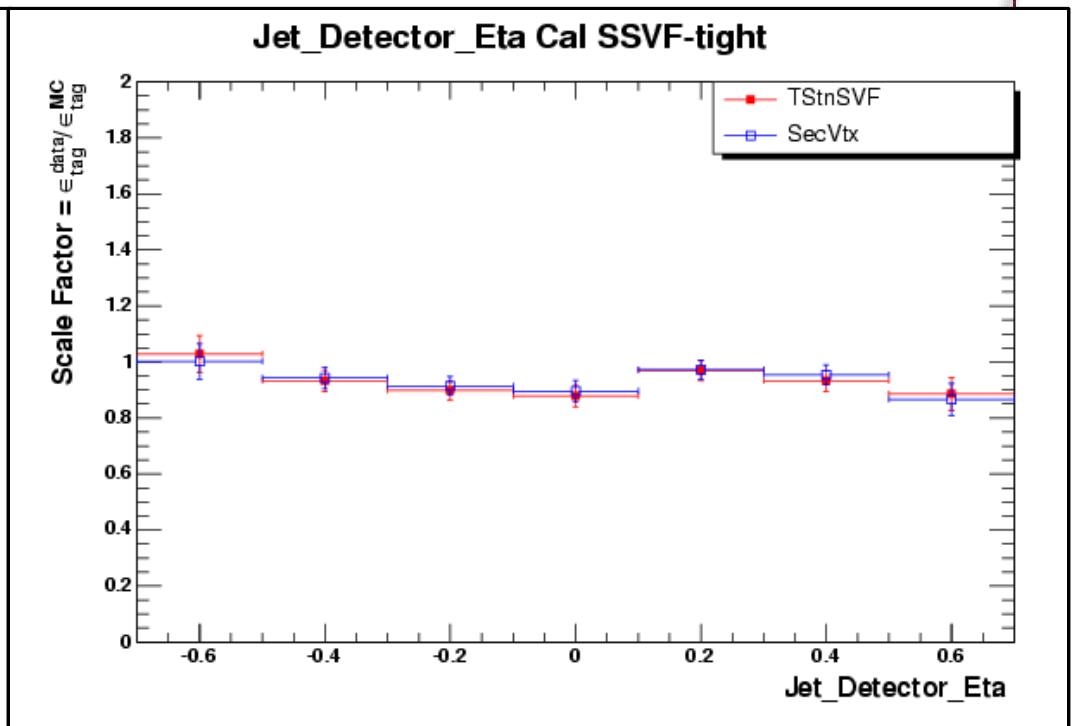
- The TStnSVF and SecVtx results are within statistical errors.
  - The statistical error is the linear combination of the uncertainties of each non-b template. This assumes 100% correlation between the each measurement.
  - The systematic error is the RMS of the individual non-b templates w.r.t the average value calculated.

SF Histograms

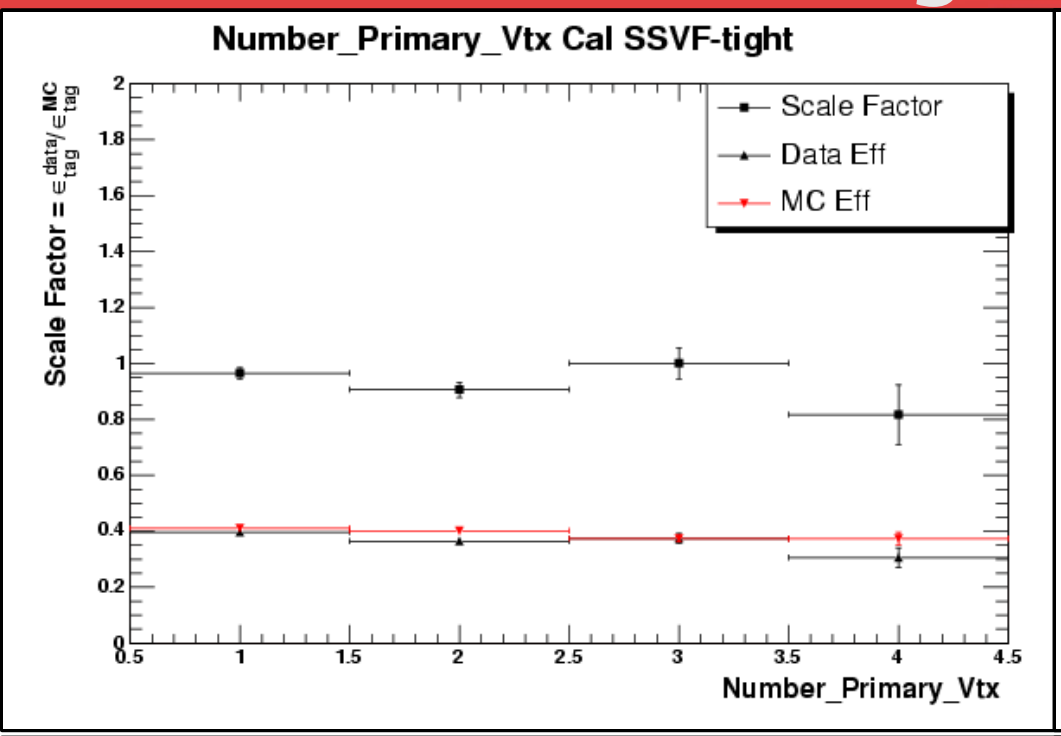


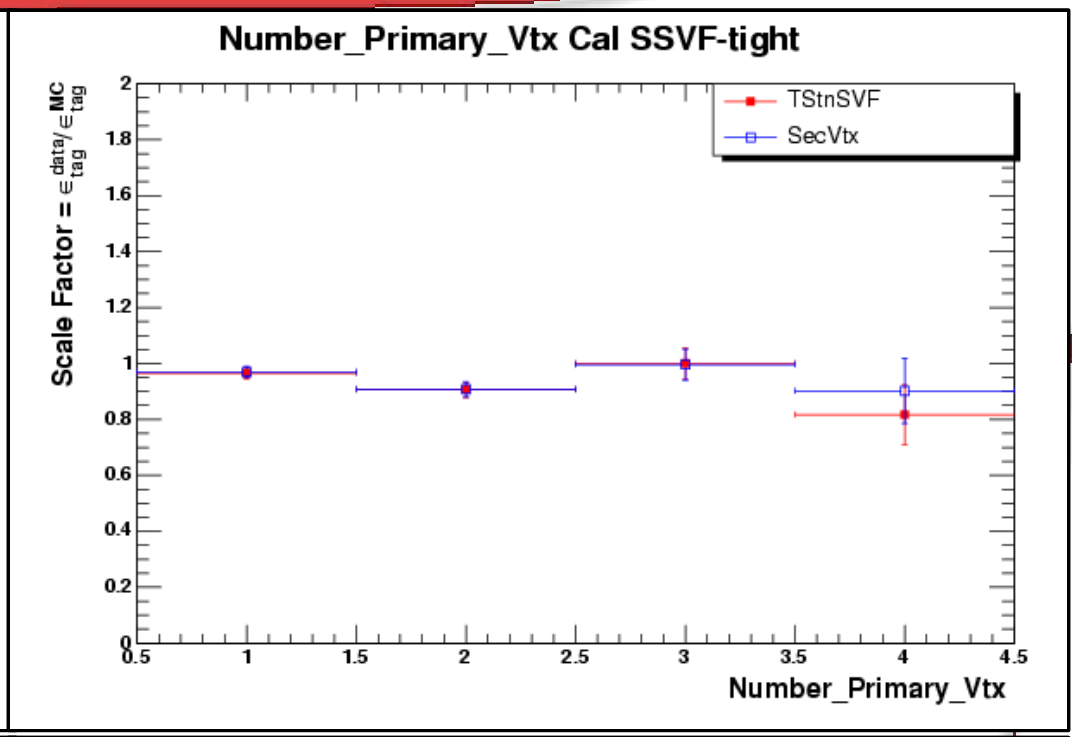


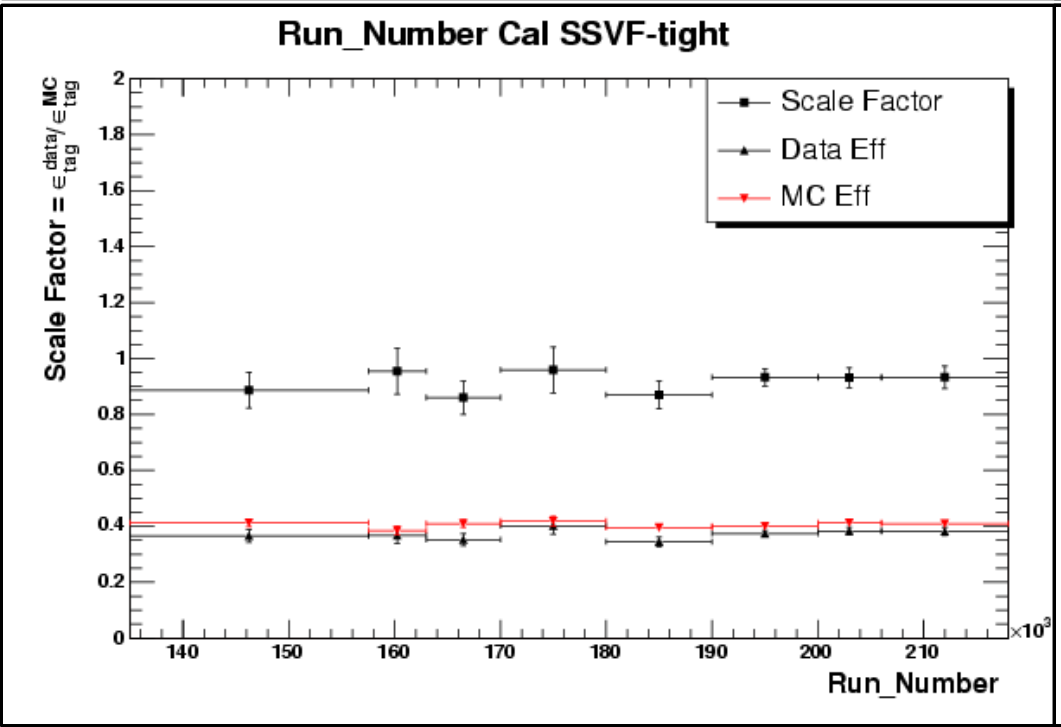


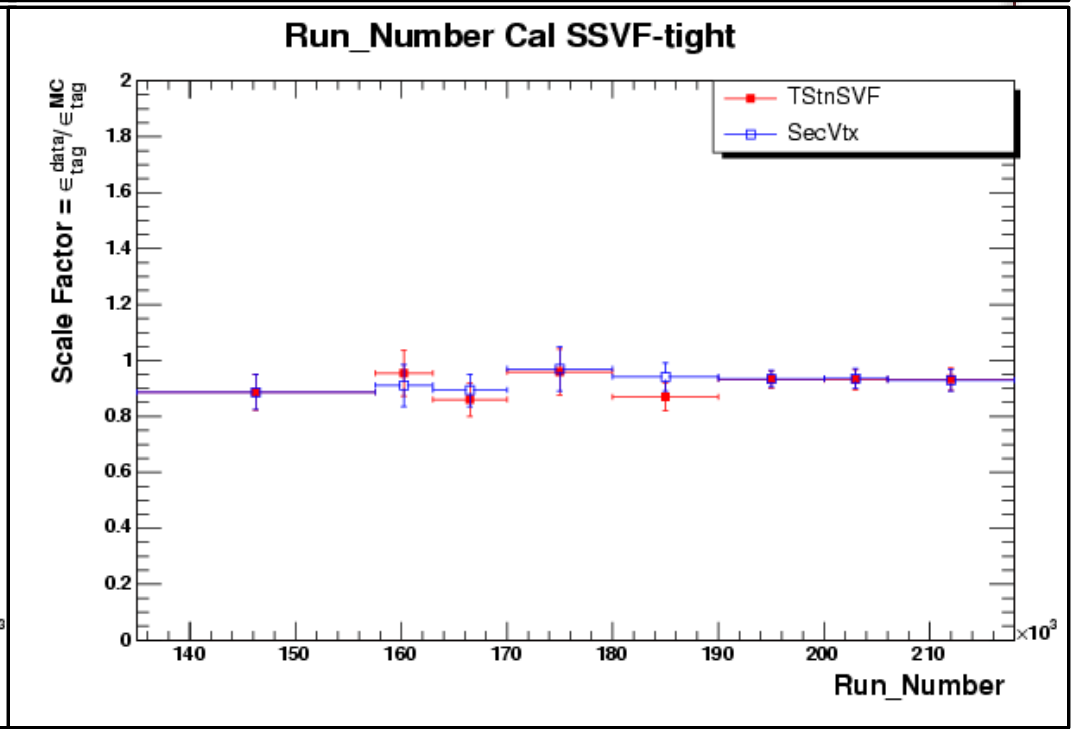


SF Histograms









BTSF Conclusion

- The SF I calculate for SecVtx and TStnSVF are comparable.
  - Additionally, the SFs are I calculate for SecVtx are within one sigma from the SFs from CDF Note 8640.
- The trends show that my tagger behaves as SecVtx does over the different variables.

Overall Conclusions

- So far my TStnSVF b-tagger is performing well and behaving normally.
- All my numbers so far are comparable to SecVtx.
- Upcoming... is the Mistag Asymmetry analysis.
- In addition, I will have to consider the systematic uncertainties as well.

- Thanks to Tom J., Enrique P., and Tom W. for all their help.
- A CDF Note is in the works regarding TStnSVF and what you've seen in this presentation.
- Web page at URL:
  - http://www-cdf.fnal.gov/htbin/twiki/bin/view/ZtoBBbar/TStnSVF
  - Additional Histograms

Backup

- Raw b-tagging rate for positive tags, calorimeter based jets.

B-Tagging Rate	Tight	Loose
TStnSVF	0.30359	0.36448
SecVtx	0.31596	0.37201