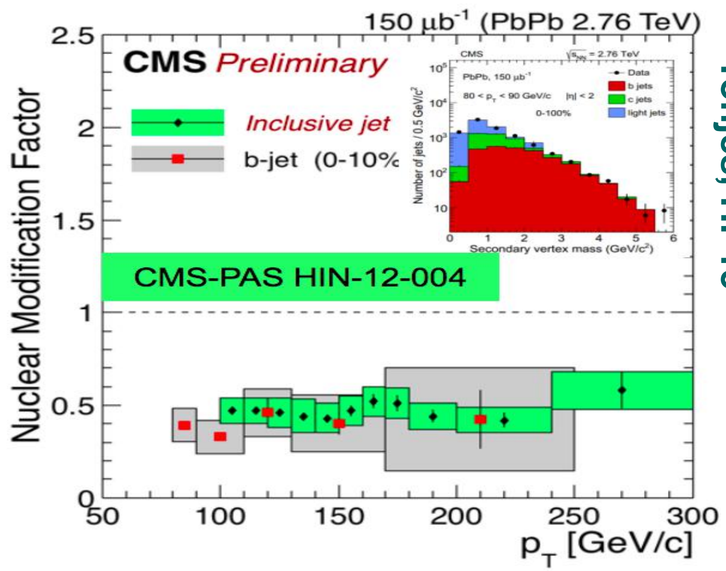
ALICE b-jet tagging
in Run2 5 TeV pPb collisions with the SV method

Róbert Vértesi (Wigner RCP Budapest)
vertesi.robert@wigner.mta.hu
Filip Krizek, Artem Isakov (NPI Prague)
Ashik Ikbal Sheikh (VECC Kolkata)

This work has been supported by the Hungarian NKFIH/OTKA K 120660 grant and the János Bolyai scholarship of the Hungarian Academy of Sciences

goals and datasets

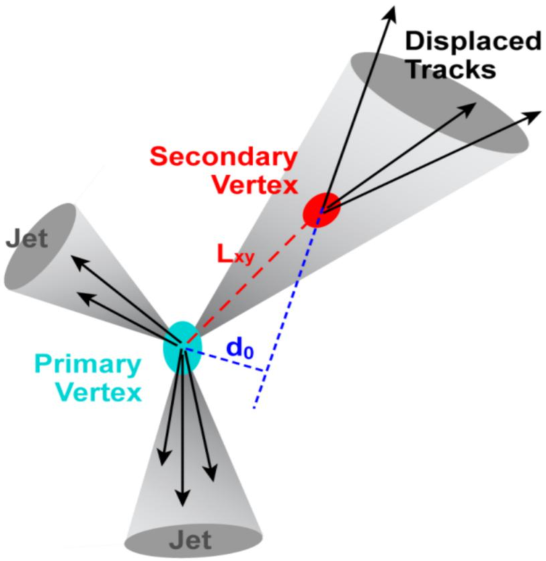
- **pp**:
  - pQCD benchmark and baseline for nuclear modification
  - Color charge vs. mass/flavor effects?
- **p-A**:
  - CNM effects?
    Baseline for modification in hot medium
- A-A:
  - mass ordering?
  - Low/intermediate pT - unique
  - contribution of gluon-splitting to direct b quark production?
  - Radiative or collisional energy loss?



CMS-PAS-HIN-12-014
Tonjes, HP15

- Experimental data:
  - pp 2017 data at 5 TeV (IP method) ~600M evts
  - p-Pb 2016 data at 5 TeV (IP and SV methods) ~900M Minimum Bias evts
  - ITS+TPC tracks, p_T>0.15
  - Anti-kT jets, R=0.4, $|\eta|<0.5$

extracting the b-jet cross section



A. Jet Reconstruction

B. b-jet selection

b-jet: presence of a b-hadron inside a cone with given R centered on the jet axis

1. ***Impact parameter significance method***
   based on the closest approach to the primary vertex of tracks inside a jet
   Hadi Hassan (Linus Feldkamp, Min Jung Kweon, Minjung Kim)
2. ***Displaced secondary vertex method***
   secondary vertex reconstruction and evaluation of its distance from the primary vertex - *Ashik Ikbal Sheikh, Filip Křížek, Artem Isakov, R.V. (Elena Bruna, Lukás Kramárik, Gyulnara Eyyubova)*

C. Statistically remove non-b jets from tagged sample

D. Unfolding

E. Efficiency correction

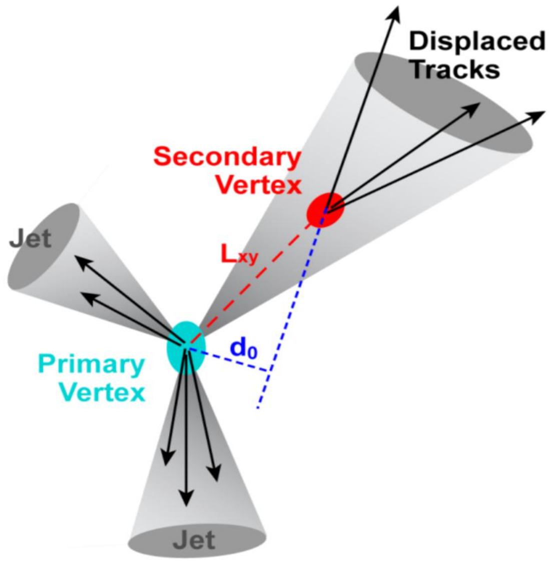
extracting the b-jet cross section

A. Jet Reconstruction

B. b-jet selection

b-jet: presence of a b-hadron inside a cone with given R centered on the jet axis

1. ***Impact parameter significance method***
   based on the closest approach to the primary vertex of tracks inside a jet
   Hadi Hassan (Linus Feldkamp, Min Jung Kweon, Minjung Kim)



2. ***Displaced secondary vertex method***
   secondary vertex reconstruction and evaluation of its distance from the primary vertex - *Ashik Ikbal Sheikh, Filip Křížek, Artem Isakov, R.V. (Elena Bruna, Lukás Kramárik, Gyulnara Eyyubova)*

C. Statistically remove non-b jets from tagged sample

D. Unfolding

E. Efficiency correction

spectrum extraction

- Extract SV tagging **efficiency** from Monte Carlo

$$\epsilon_{b,c,udsg}(p_{\mathrm{T,ch.\ jet}}^{det.}) = \frac{N_{b,c,udsg}^{tagged}(p_{\mathrm{T,ch.\ jet}}^{det.})}{N_{b,c,udsg}^{gen.}(p_{\mathrm{T,ch.\ jet}}^{det.})}$$

==> Reliable MC is essential!

- Extract **purity** from template fit and MC

$$P(p_{\mathrm{T,ch.\ jet}}^{det.}) = \frac{N_{\mathrm{b-jets}}(p_{\mathrm{T,ch.\ jet}}^{det.})}{N_{\text{all jets}}(p_{\mathrm{T,ch.\ jet}}^{det.})}$$

- Correct tagged inclusive raw spectrum:

$$\frac{1}{N}\frac{dN_{\mathrm{measured,b}}}{dp_{\mathrm{T}}} = \frac{1}{N}\frac{1}{\epsilon_b}\cdot P \cdot \frac{dN_{\mathrm{tagged}}}{dp_{\mathrm{T}}}$$

- Unfolding and final correction step $p_{\mathrm{T,ch.\ jet}}^{det.} \rightarrow p_{\mathrm{T,ch.\ jet}}$

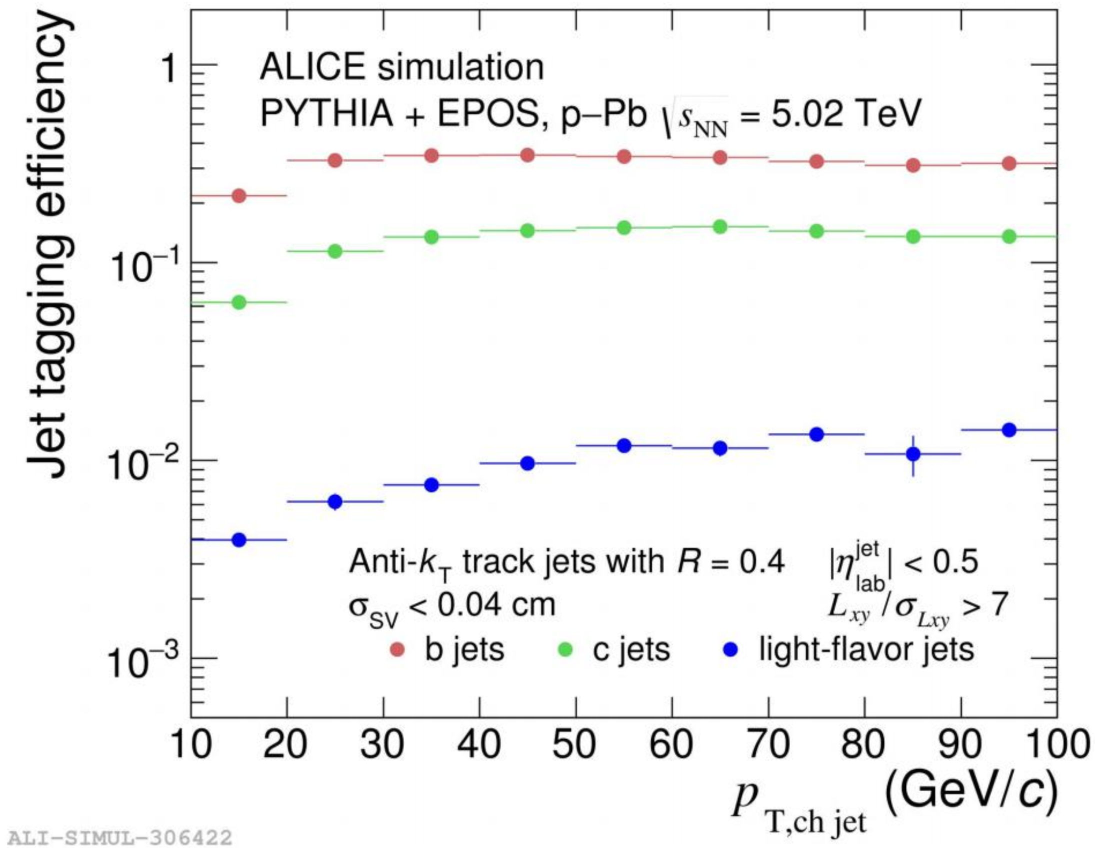
$$\frac{d\sigma_{\mathrm{b-jet}}}{dp_T} = \frac{1}{L_{\mathrm{int}}}\cdot \mathrm{Unfolded}\left(\frac{dN_{\mathrm{measured,b}}}{dp_T}\right)$$

analysis

- Data extraction (Run2 p-Pb, LHC16{q,t}):
- Efficiency and purity corrections
  - Efficiency: LHC17h6{a,b,c,d,e,f}2
  - Purity #1 "POWbc": real inclusive jets and POWHEG c,b spectra
  - Purity #2: "data-driven": template fits from LHC17h6_2 simulations

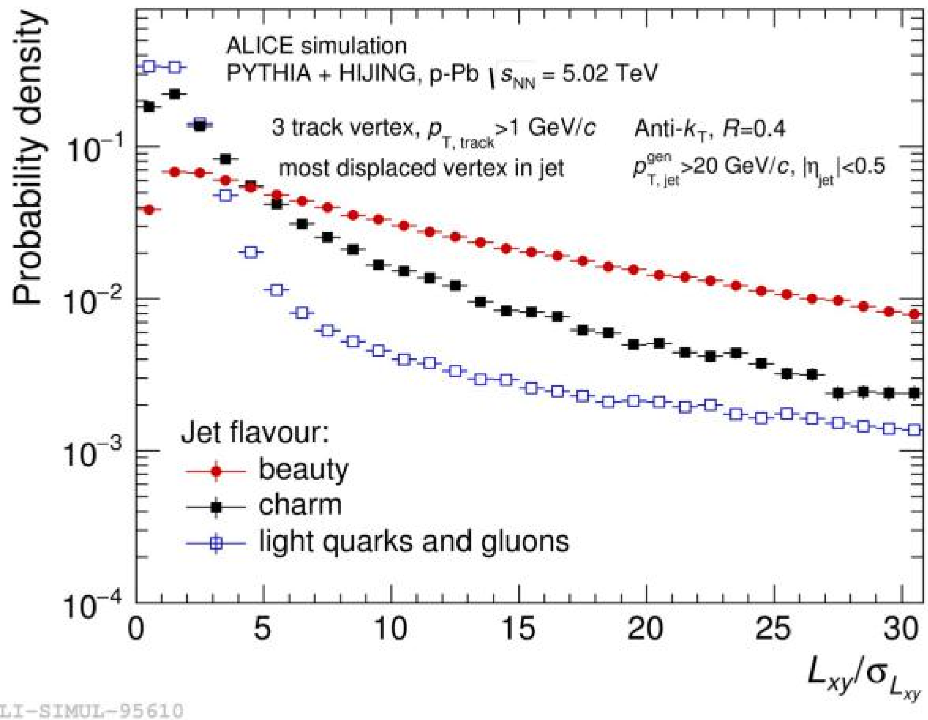
    ==> a combined “hybrid” method
- Unfolding (SVD & Bayesian, binned)
  - Matrix based on LHC17h6_2, outliers removed ($p_T < p_T$hard x4)
  - PYTHIA hard processes + EPOS underlying event
- Systematics
  - Tracking & jet reconstruction related
  - b-tagging related

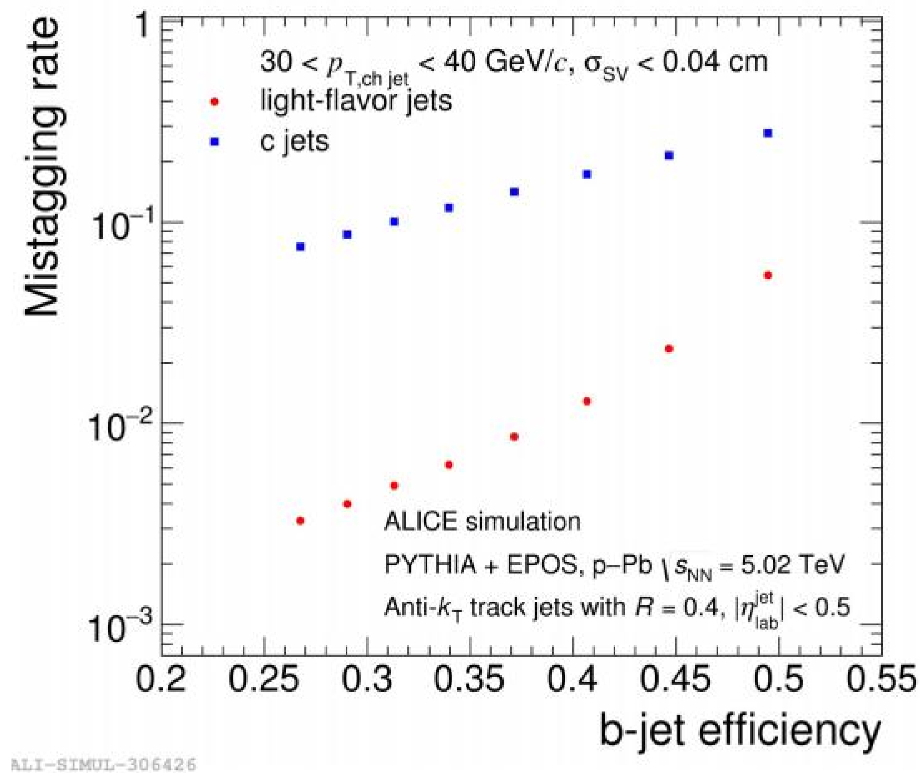
tagging efficiencies vs. p_T



- Tagging cuts
  - Dispersion of reconstructed secondary vertex $\boldsymbol{\sigma_{vtx}}$
  - Significance of primary-secondary vertex distance $SL_{xy} = L_{xy}/\boldsymbol{\sigma_{Lxy}}$

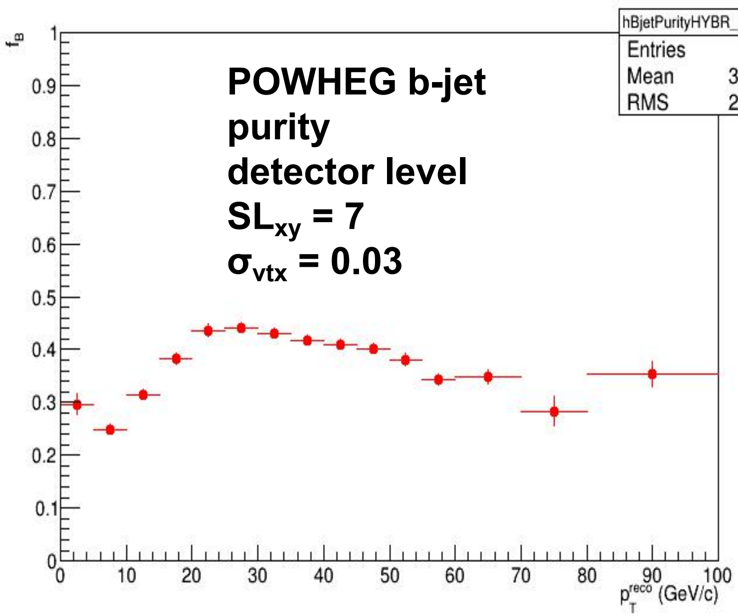
tagging performance

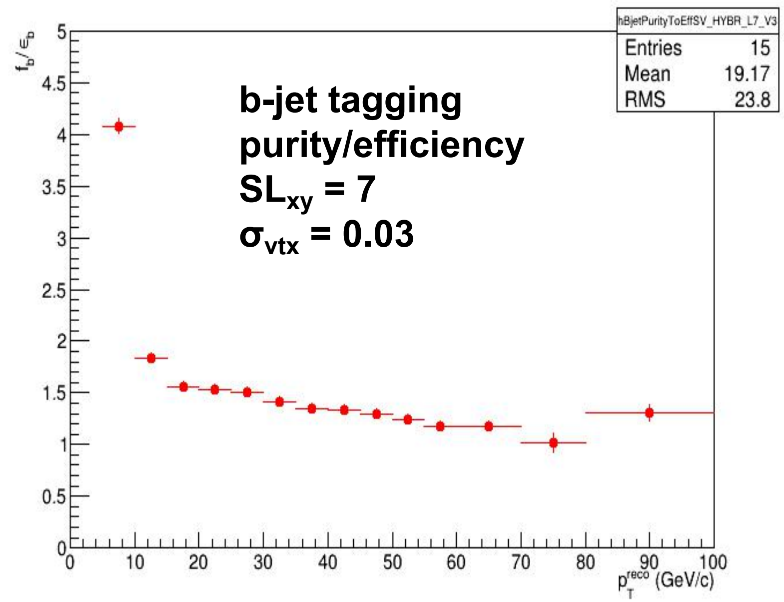




- Evolutions of efficiencies and mistagging rates with SL_{xy}
  - Left: efficiency vs. SLxy, no sigvtx cut applied
  - Right: efficiency vs. mistagging rates for different SLxy values

purity & tagging correction, POWHEG



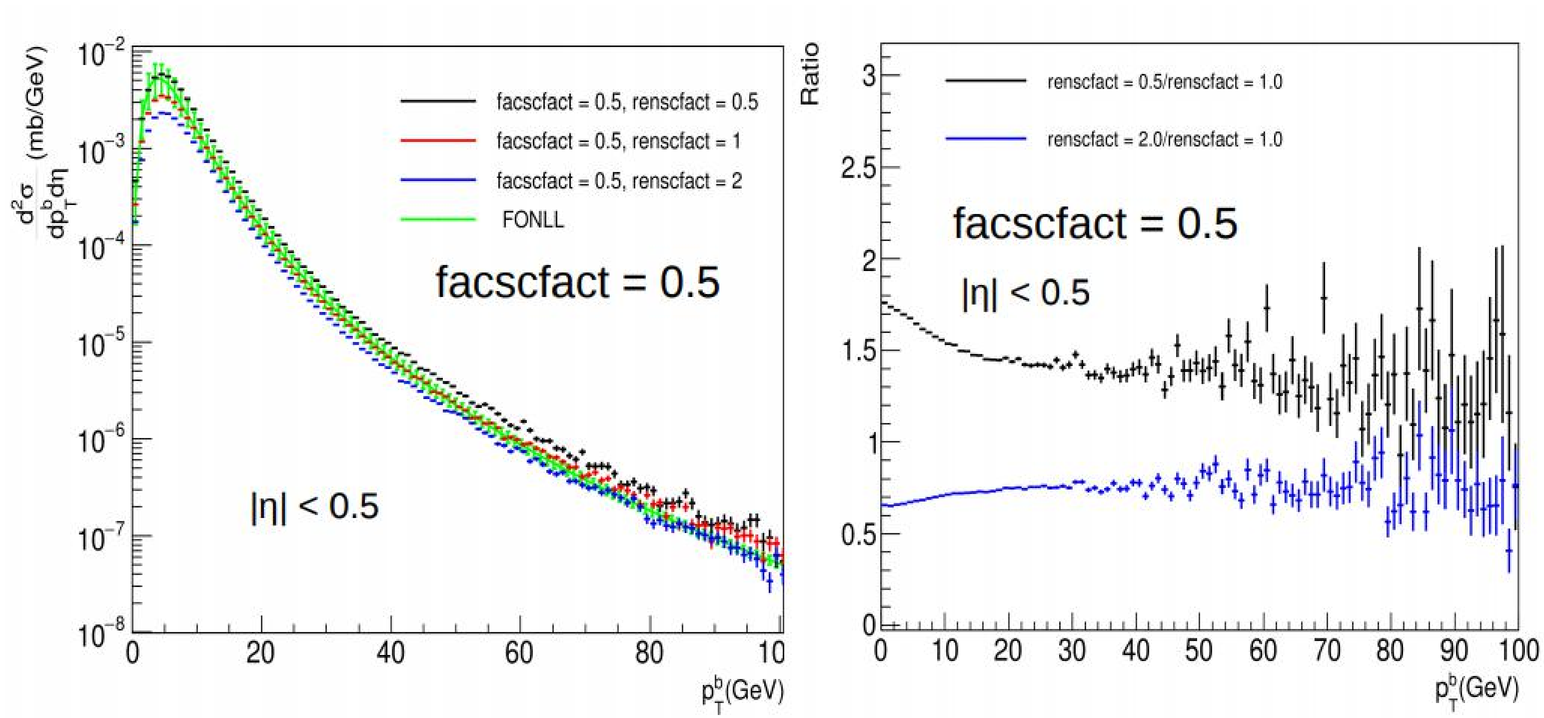


- Purity obtained as:
  - Going to detector level: POWHEG spectrum * detector matrix
  - Using the good old formulae

$$f_b(p_T^{\det}) = \frac{N_b^{tagged}(p_T^{\det})}{N_{inclusive}^{tagged}(p_T^{\det})}$$

$$N_b^{tagged}(p_T^{\det}) = N_{inclusive}^{tagged}(p_T^{\det}) - N_b^{Powheg}(p_T^{\det})\cdot\varepsilon_b(p_T^{\det}) - N_c^{Powheg}(p_T^{\det})\cdot\varepsilon_c(p_T^{\det})$$
$$-\left(N_{inclusive}(p_T^{\det}) - N_c^{Powheg}(p_T^{\det}) - N_b^{Powheg}(p_T^{\det})\right)\cdot\varepsilon_{lf}(p_T^{\det})$$

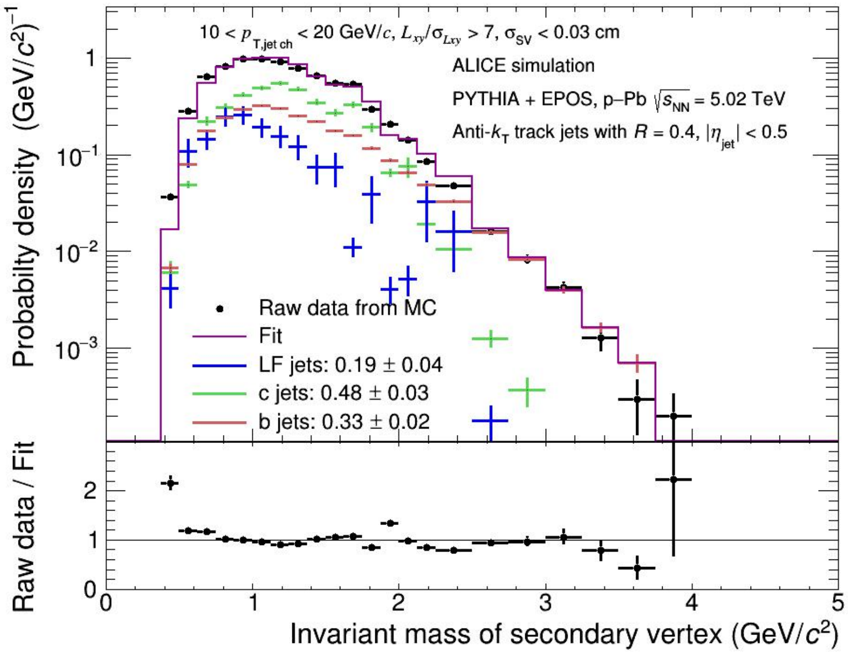
uncertainty in POWHEG



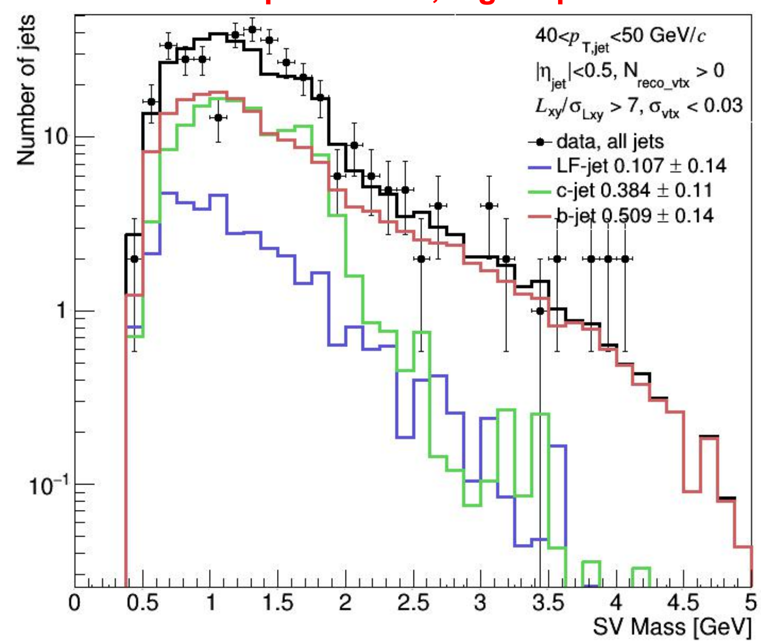
- Several variations (defaults in bold) and cross-variations:
  - m_b=4.5, **4.75**, 5.0 GeV ; m_c=1.3, **1.5**, 1.7 GeV
  - factorization scale = 0.5, **1.0**, 2.0 ; renormalization scale = 0.5, **1.0**, 2.0
  - Translates to a factor ~2 uncertainty on the corrected spectrum (later)
  - See more: backup slides and https://twiki.cern.ch/twiki/pub/ALICE/BtagSecVtx/PowhegSystematicsBeauty_ashik.pdf

purity, data-driven - template fits

Example: Minuit, lowest pT bin



Example: RooFit, higher pT bin

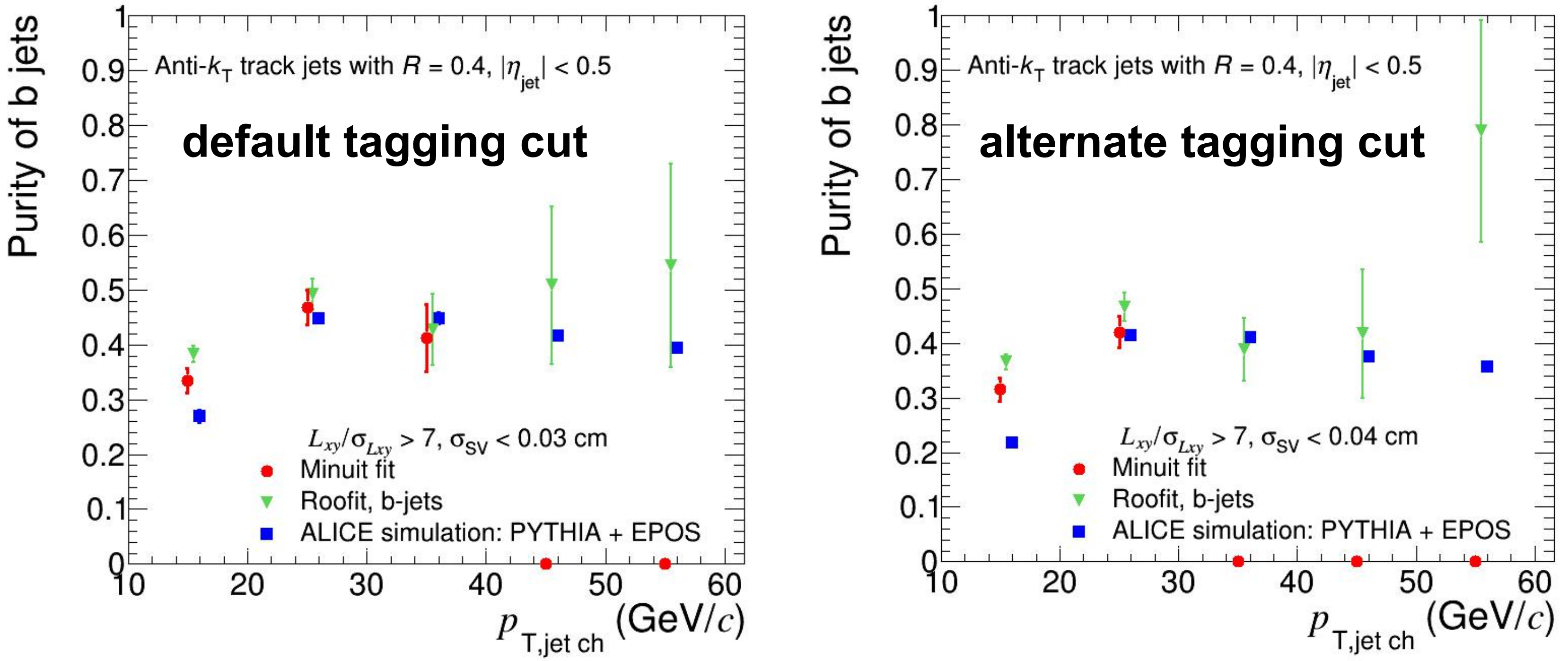


- Minuit vs. RooFit on **measured data**
  - RooFit: template errors ignored - can be a problem
  - Minuit: correct treatment of errors;

$$F(n) = \sum_{i=1}^{nbins} \frac{(Data_i - B_i * p_B - C_i * p_C - LF_i * p_{LF})^2}{\sigma^2_{Data_i} + (\sigma_{B_i} * p_B)^2 + (\sigma_{C_i} * p_C)^2 + (\sigma_{LF_i} * p_{LF})^2}$$

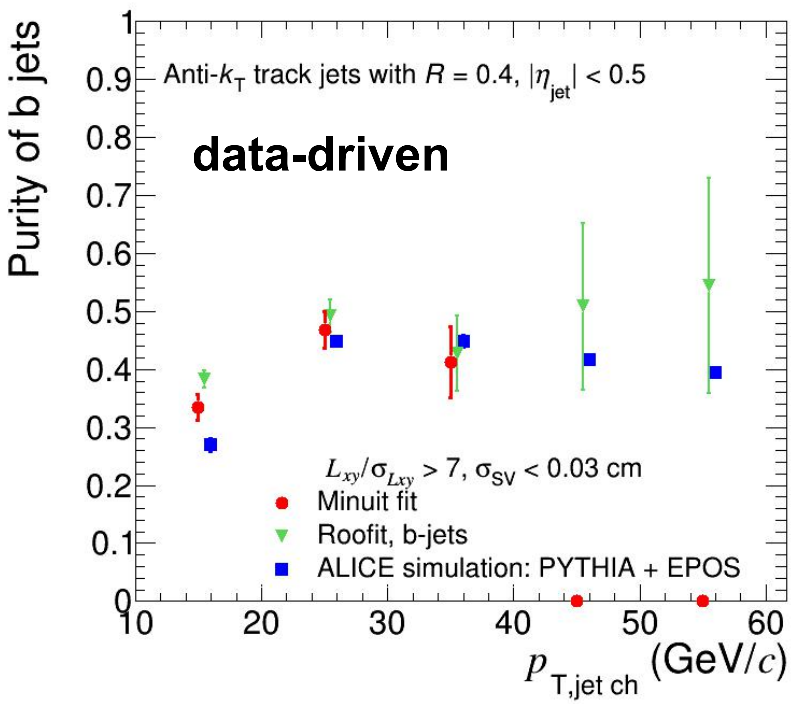
but: convergence problems at higher p_T

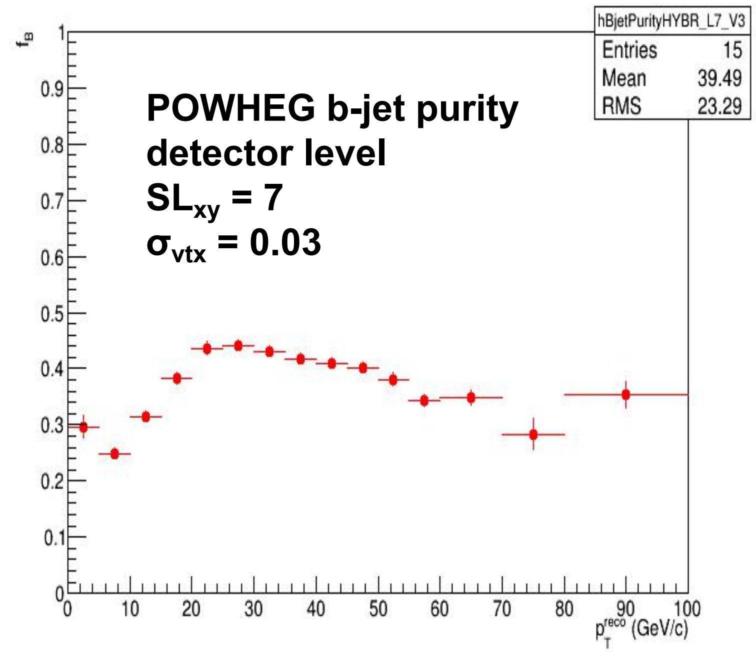
purity, data-driven - p_T-dependence



- Minuit convergence problems already from 30 GeV/c in some cases, above 40 in most cases
  - Note: merging the bins did not help
- RooFit different at low-p_T (and we trust it less than Minuit)
- *But:* At higher p_T-bins, RooFit and Minuit always match
  - Perhaps less effect of template errors because of wider distributions

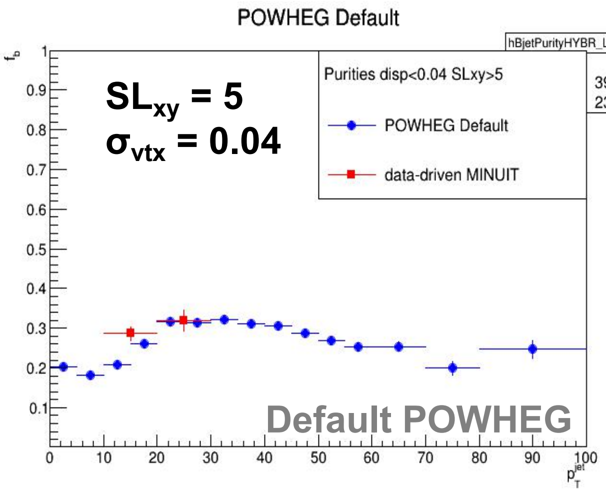
purity - comparison of methods

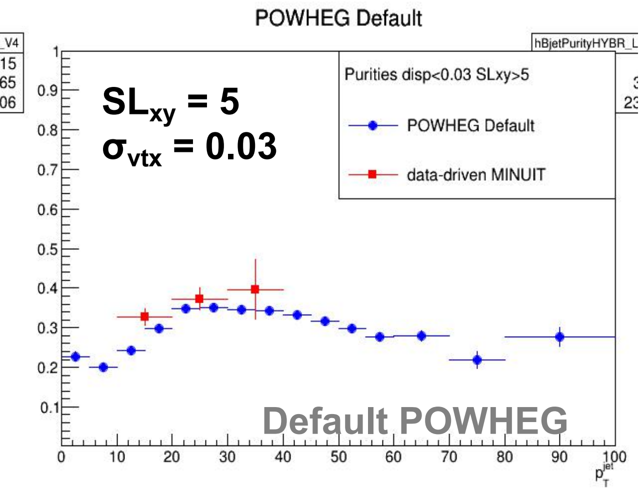


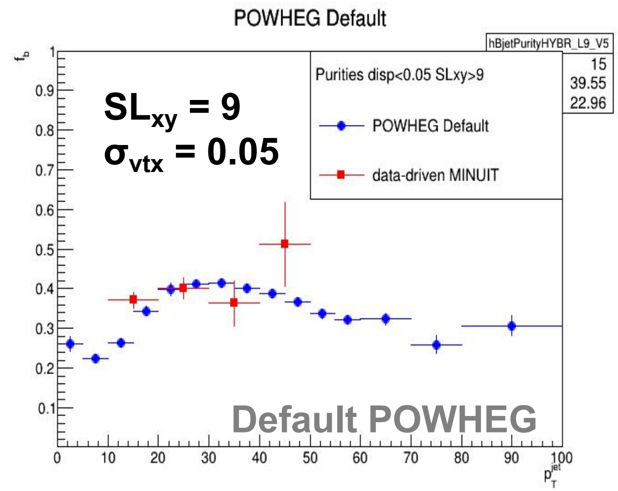


- Good news: very good consistency with the POWbc method
  - POWbc and data-driven MC template closure: good match
  - POWbc and data-driven with real data consistent within errors
- Strategy: data-driven constraints to be used to constrain purities from POWHEG

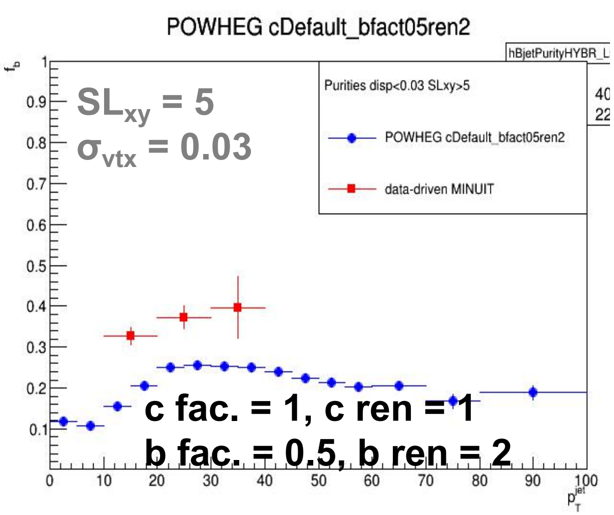
purity comparison examples

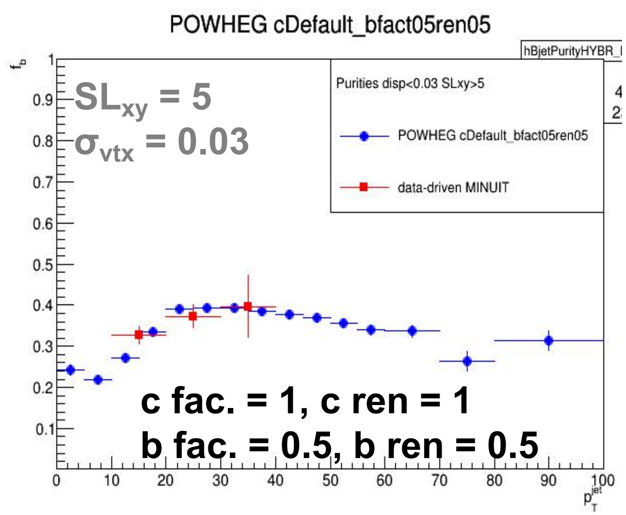


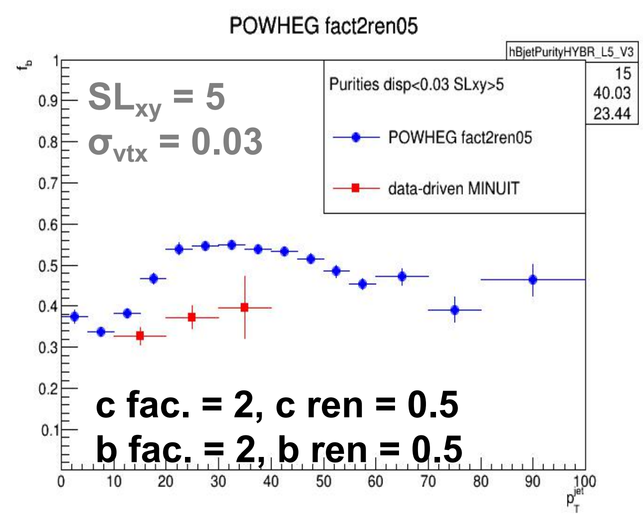




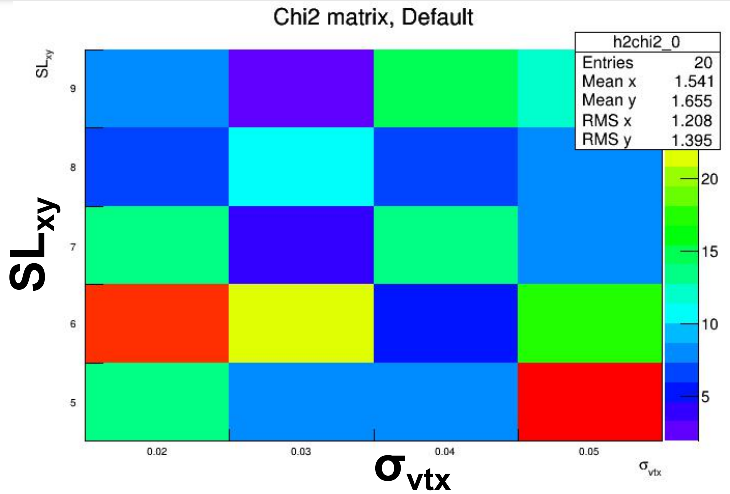
- Default POWHEG describes data regardless of tagging cut
- Scale variations cause big differences

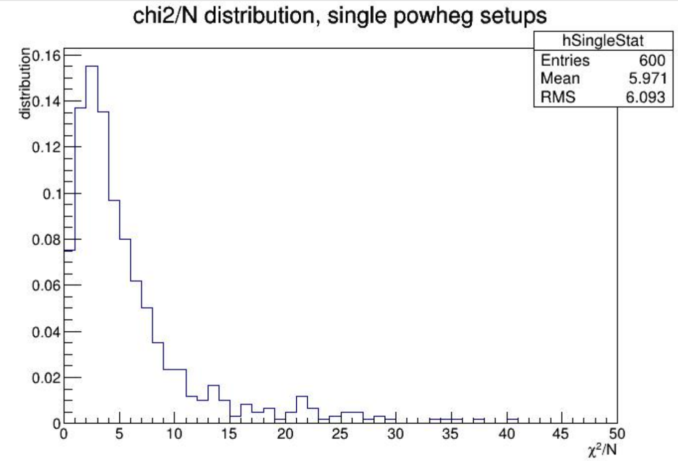




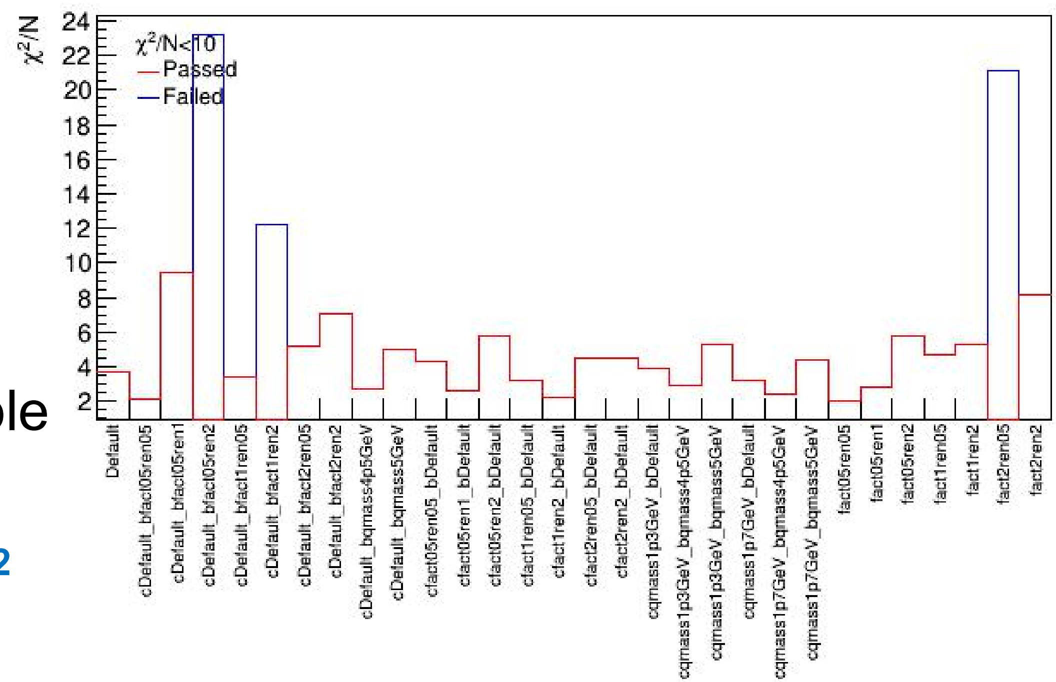


hybrid method: statistical exclusion





chi2/N values of the simultaneous test

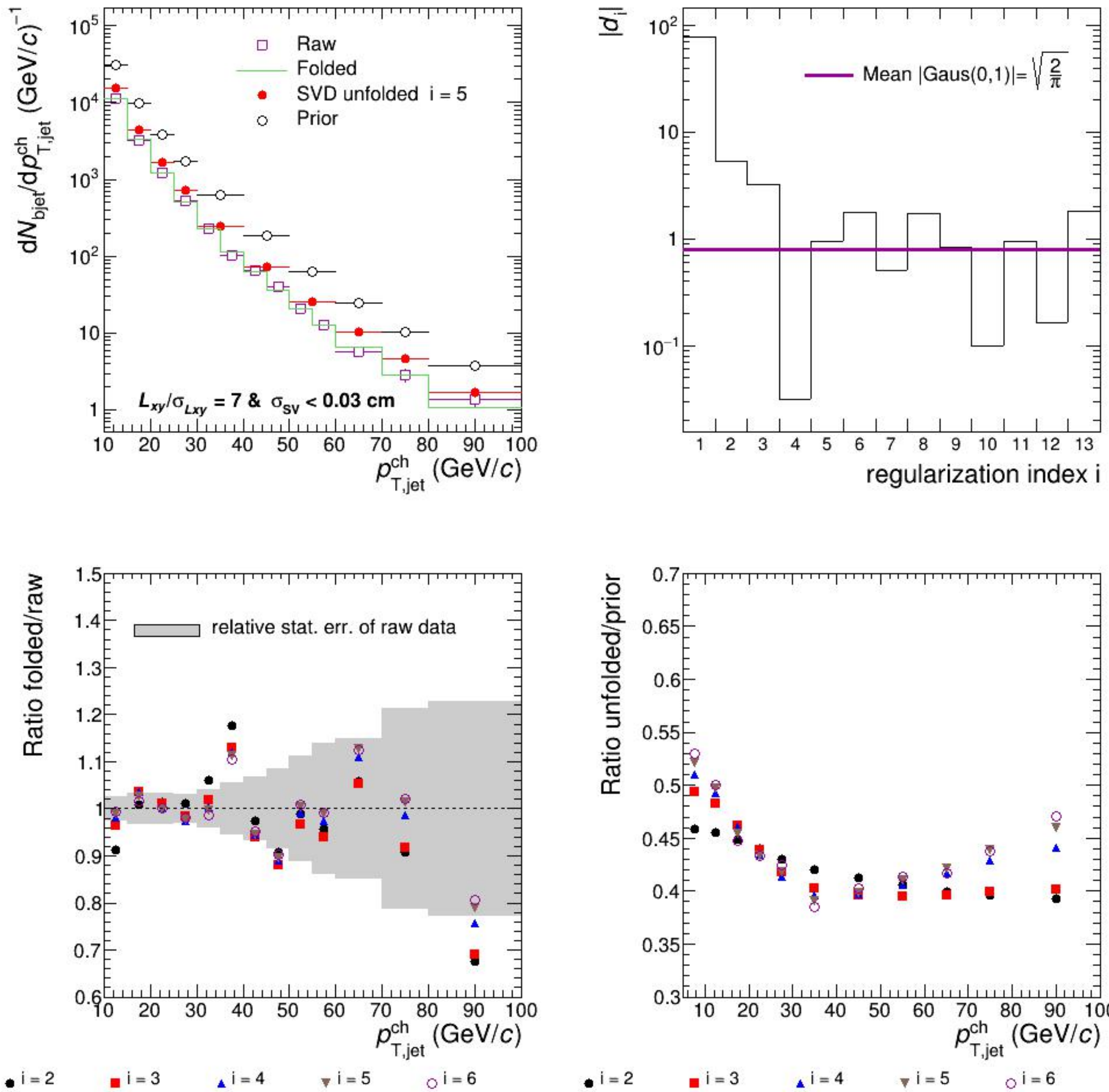


For each POWHEG setting:

- Compute χ^2 for each tagging cut
- sum them up
- Divide by sum of N_{points}
- Keep statistically acceptable settings only ($\chi^2/N<10$)
  - **c fac=1 c ren=1 b fac=0.5 b ren=2**
  - **c fac=1 c ren=1 b fac=1 b ren=2**
  - **c fac=2 c ren=2 b fac=2 b ren=2**

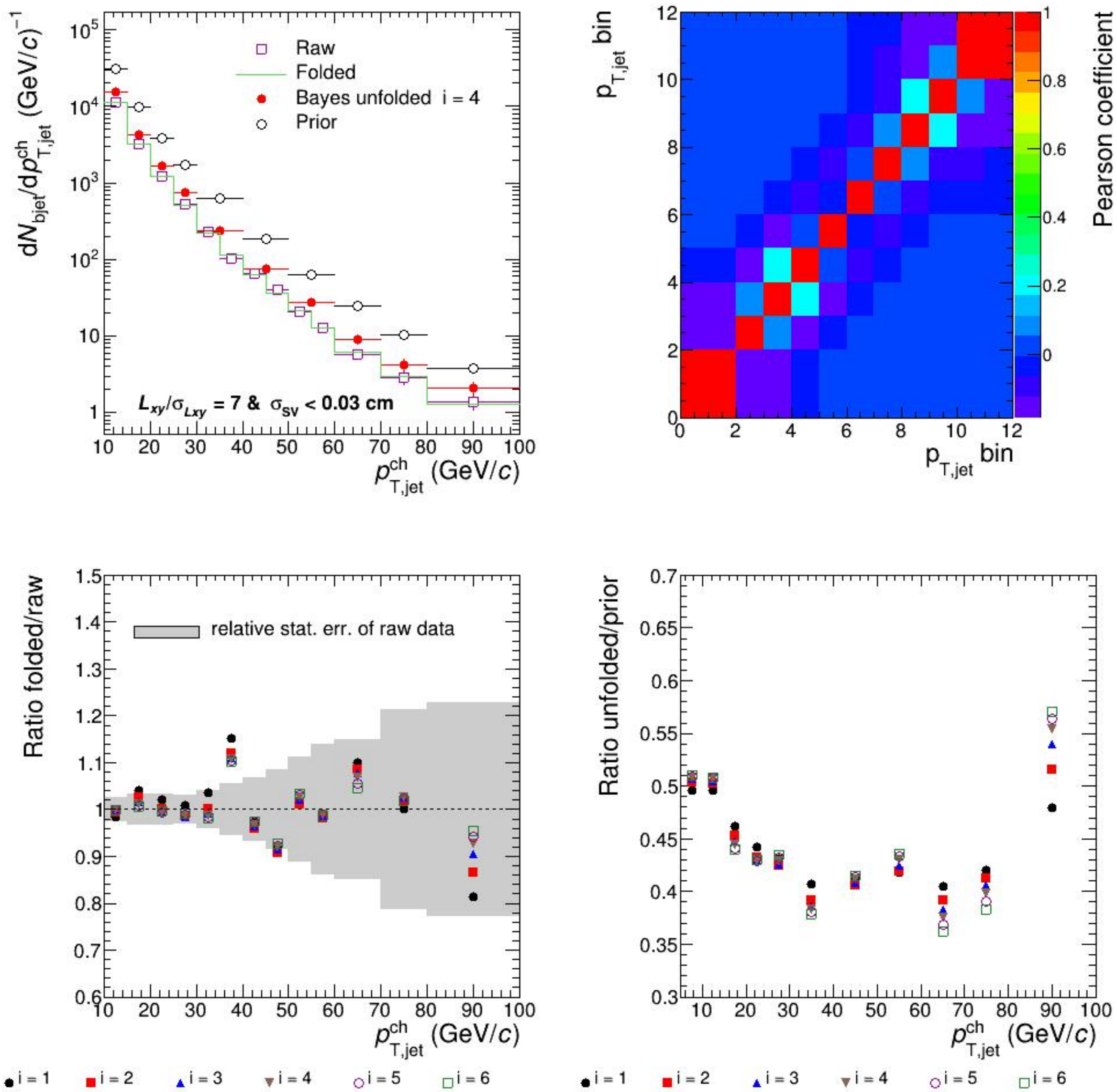
unfolding - SVD

- Generally good
  - Folded/raw ~1
  - Uncertainties below fluctuations
  - Convergent iteration (unfolded/prior)
  - kSVD>=4
- Some oscillation
  - $\sim 2\sigma$; plan to take care by rebinning or cropping the response matrix

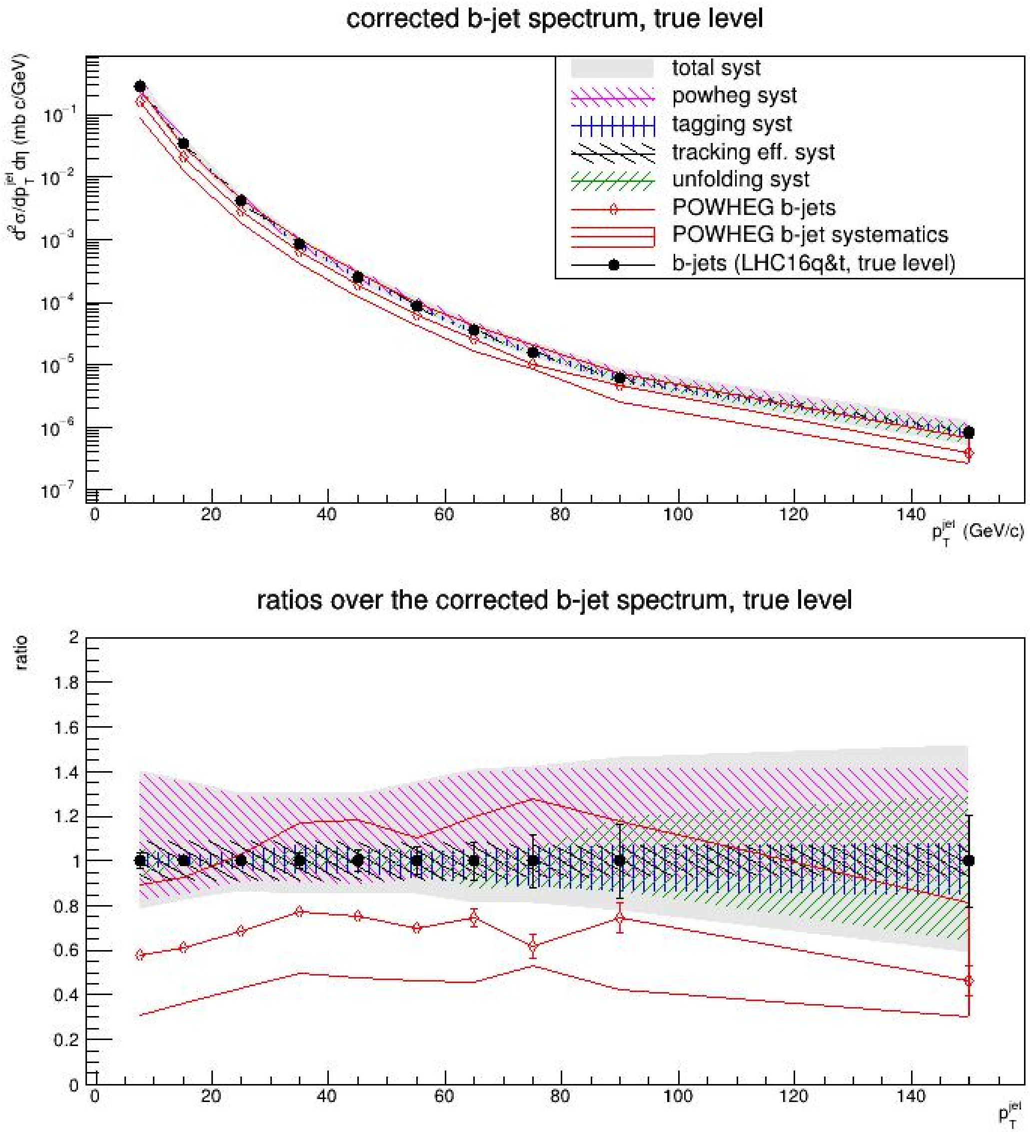


unfolding - bayesian

- Generally good
  - Folded/raw ~1
  - Uncertainties below fluctuations
  - Convergent iteration
  - kBayes>=4
- Some oscillation
  - $\sim 2\sigma$; plan to take care by rebinning or cropping the response matrix

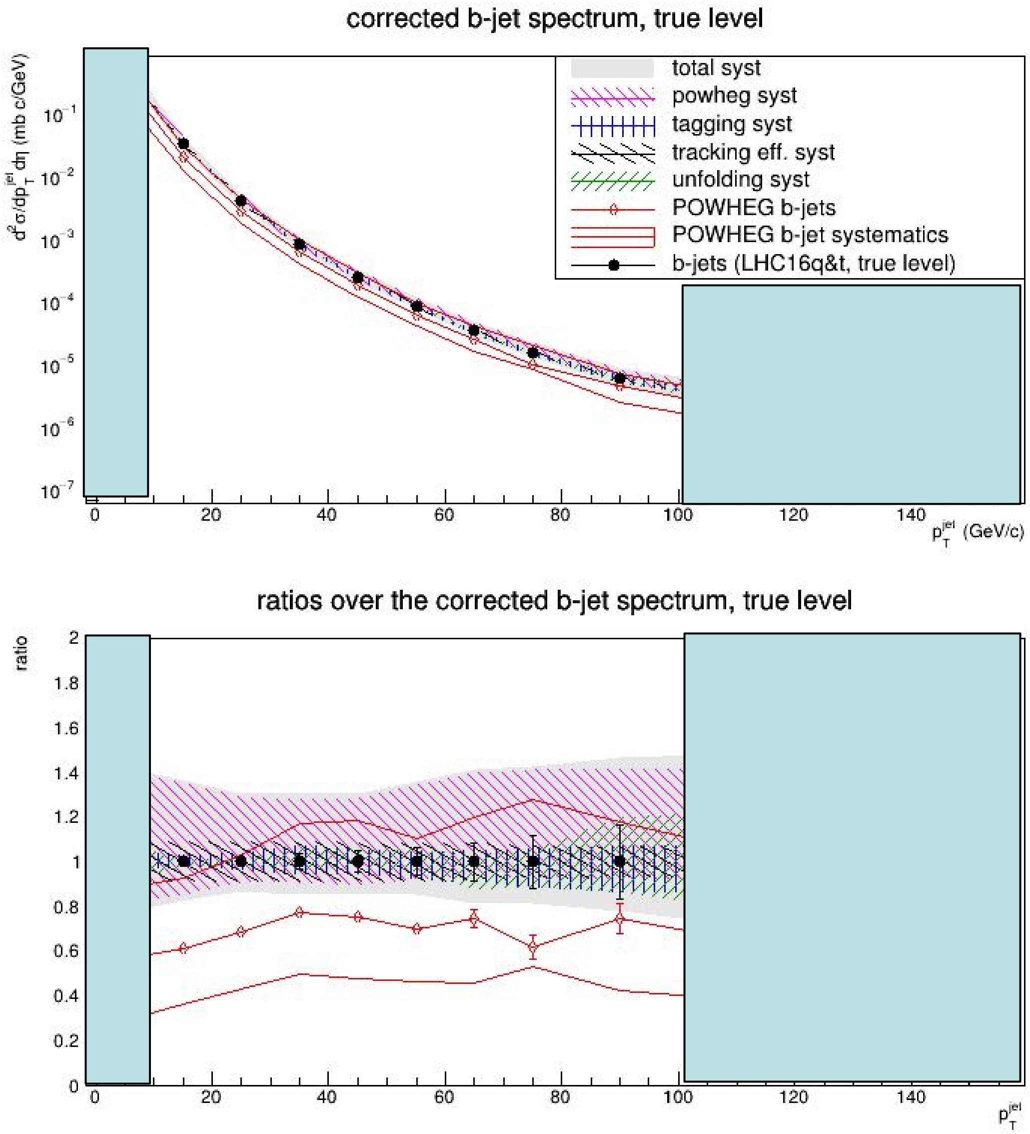


spectrum with systematics



- Corrections with the Hybrid method
- Principal analysis: $SL_{xy}>7$, $\sigma_{vtx}<0.03$
- Dominant uncertainties:
  - hybrid purity
  - unfolding (including method, regularization and prior)
  - tracking
  - tagging
- Consistent with POWHEG within errors

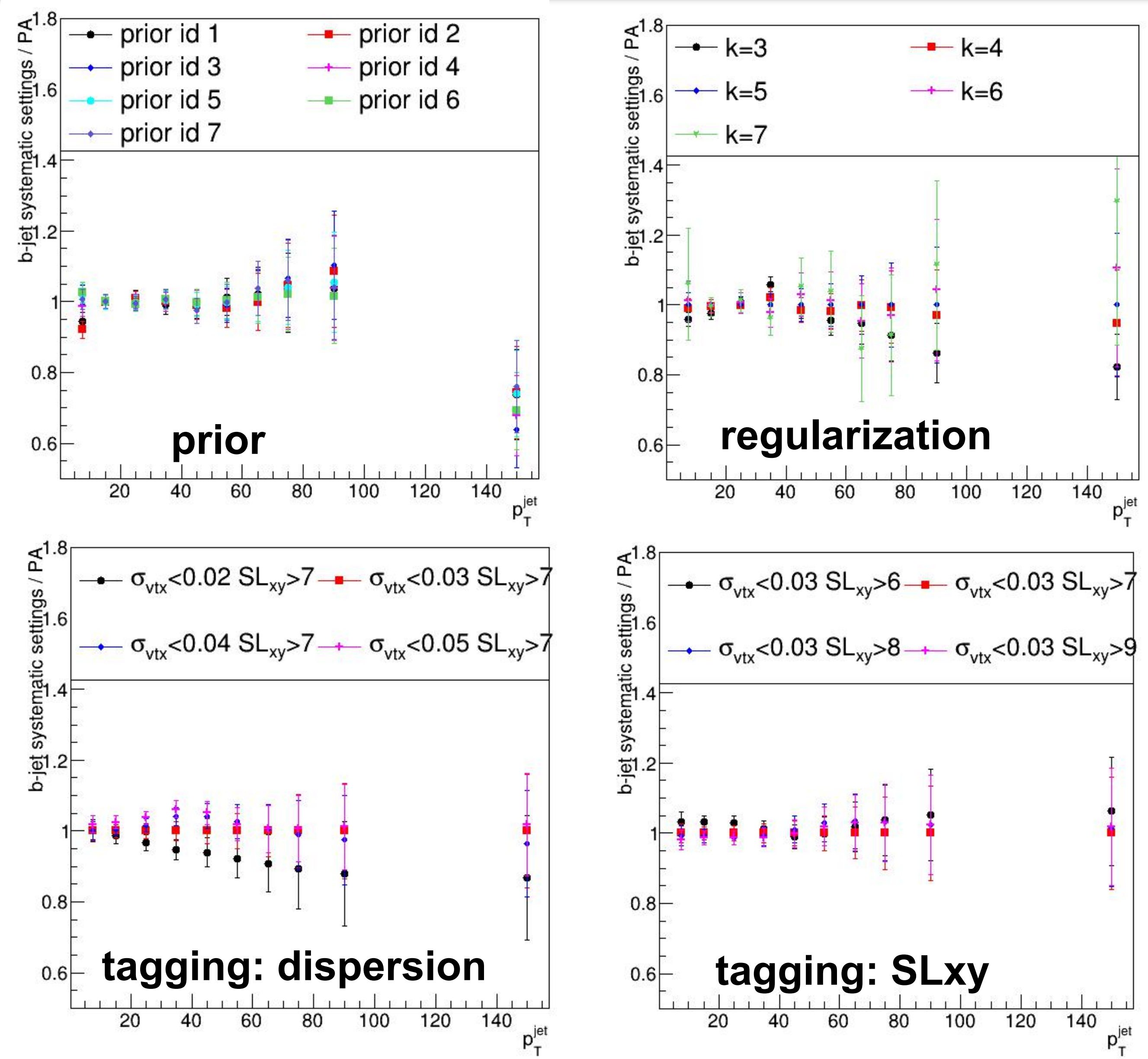
spectrum with systematics



- Corrections with the Hybrid method
- Principal analysis: $SL_{xy}>7$, $\sigma_{vtx}<0.03$
- Dominant uncertainties:
  - hybrid purity
  - unfolding (including method, regularization and prior)
  - tracking
  - tagging
- Consistent with POWHEG within errors
- **Range: 10-100 GeV/c**

some systematics (visualization)

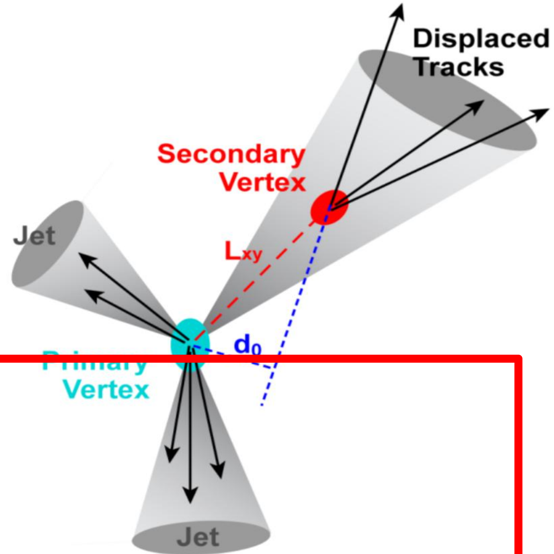
Ratios of systematic variations compared to principal analysis



Status Summary

- We computed the 5 TeV pPb b-jet spectrum
  - New hybrid purities and efficiencies
  - Corrections are consistent
  - Most of the systematics are at hand
  - Detector matrix from EPOS+PYTHIA - the extent of background effect low p_T needs to be addressed
  - New unfolding method, slight oscillations - crop matrix?
  - Some minor (?) systematics needed:
    - unfolding: test with different binning
    - contamination of primary tracks by secondary tracks
    - track p_T smearing
- Next step: Preliminary for SQM

extracting the b-jet cross section



A. Jet Reconstruction

B. b-jet selection

b-jet: presence of a b-hadron inside a cone with given R centered on the jet axis

1. ***Impact parameter significance method***
   based on the closest approach to the primary vertex of tracks inside a jet
   Hadi Hassan (Linus Feldkamp, Min Jung Kweon, Minjung Kim)

2. ***Displaced secondary vertex method***
   secondary vertex reconstruction and evaluation of its distance from the primary vertex - *Ashik Ikbal Sheikh, Filip Křížek, Artem Isakov, R.V. (Elena Bruna, Lukás Kramárik, Gyulnara Eyyubova)*

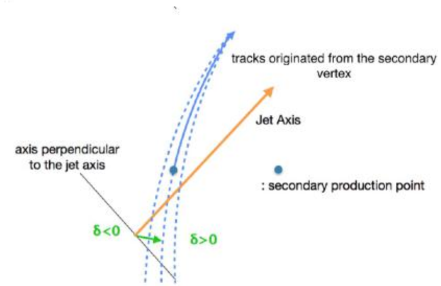
C. Statistically remove non-b jets from tagged sample

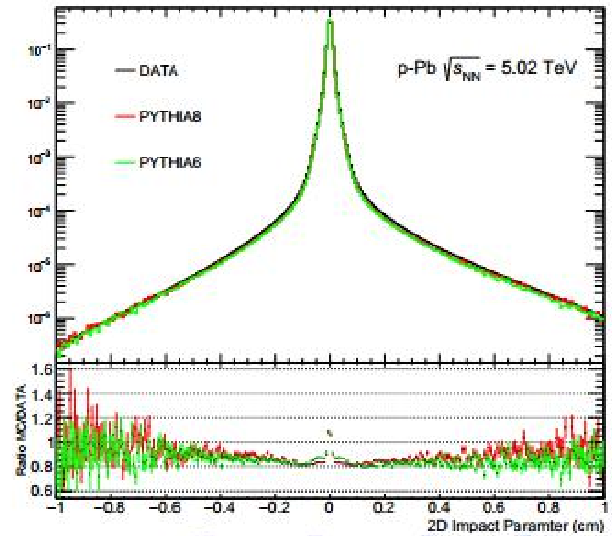
D. Unfolding

E. Efficiency correction

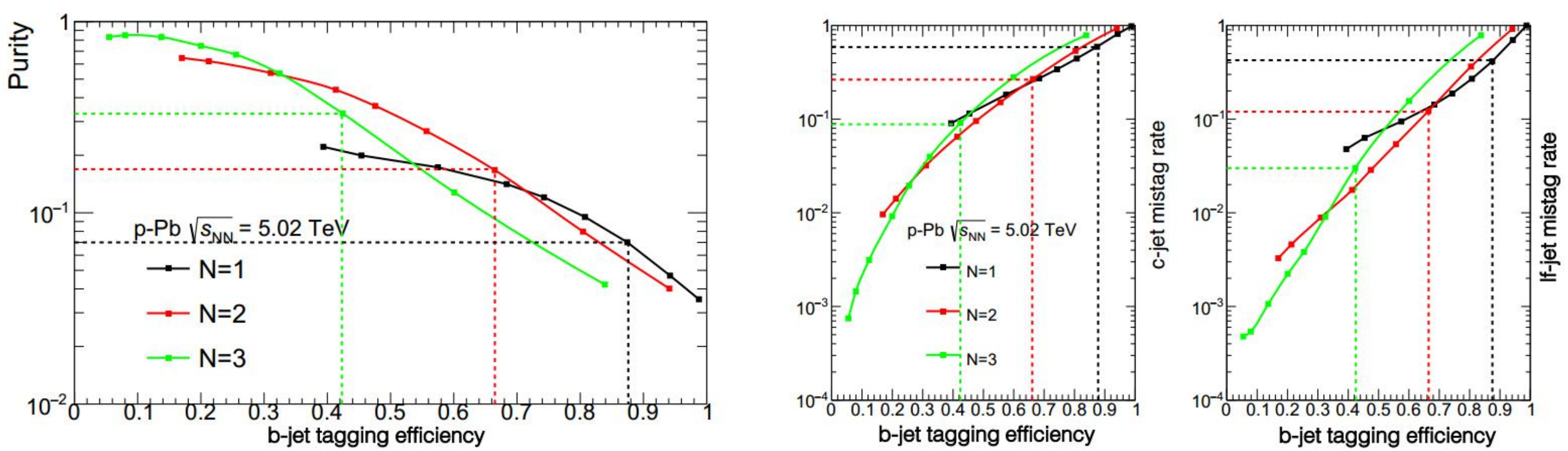
Impact parameter significance method

- Discriminator: $sd_{xy} = \delta . d_{xy}$, where δ is the impact parameter sign: $sign(\vec{d_{xy}} . \vec{p_{jet}})$.



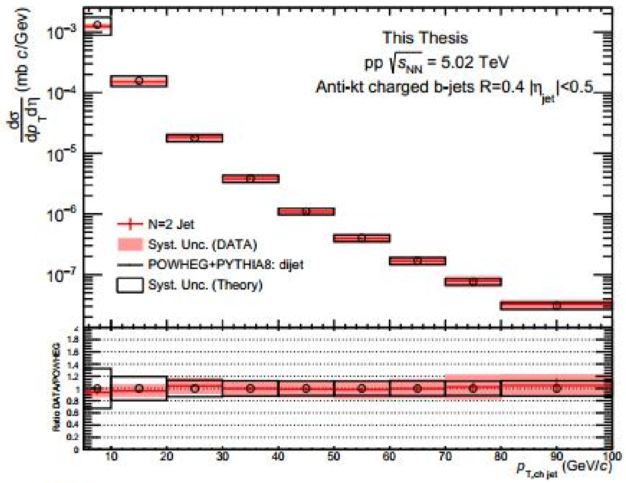


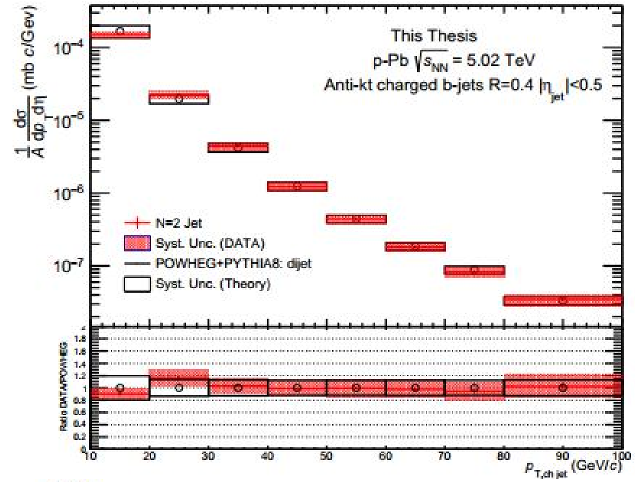
- **Track counting**
  Based on #tracks fulfilling threshold
  - N=1 : high-efficiency ; N=2 ; N=3 : high purity
- Efficiency/purity curves:



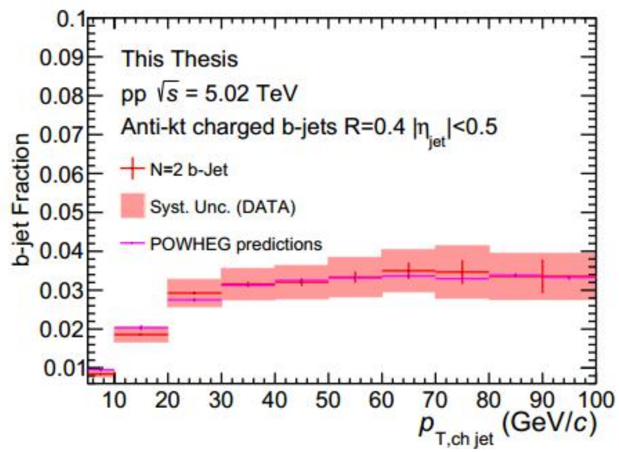
  - Working point: $\mathbf{d_{xy} > 0.008}$ threshold

Production of b-jets (pp and pPb)

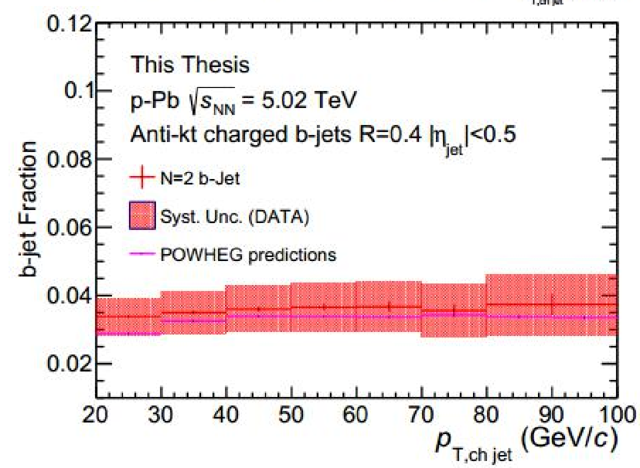




pp **x-section** pPb

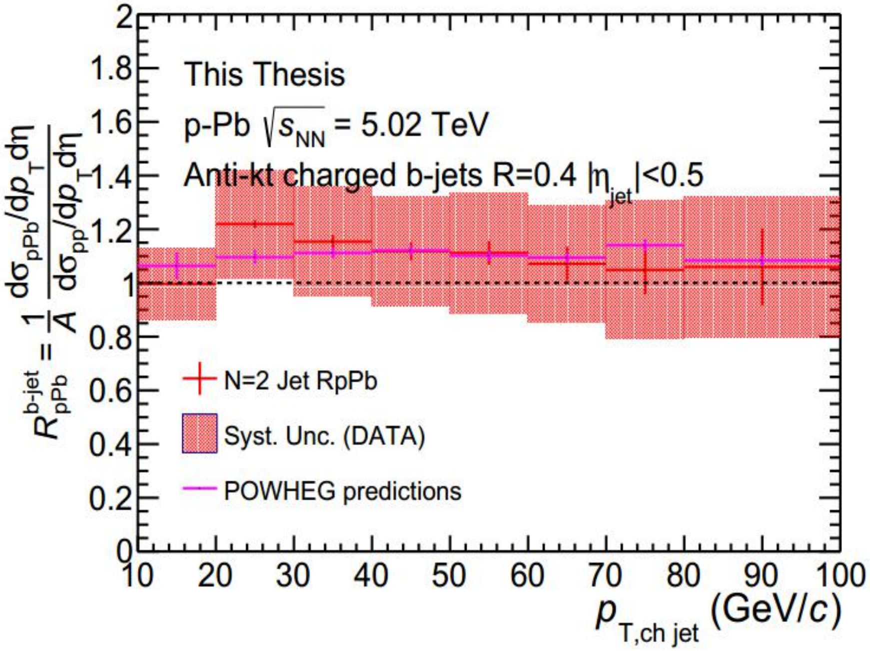


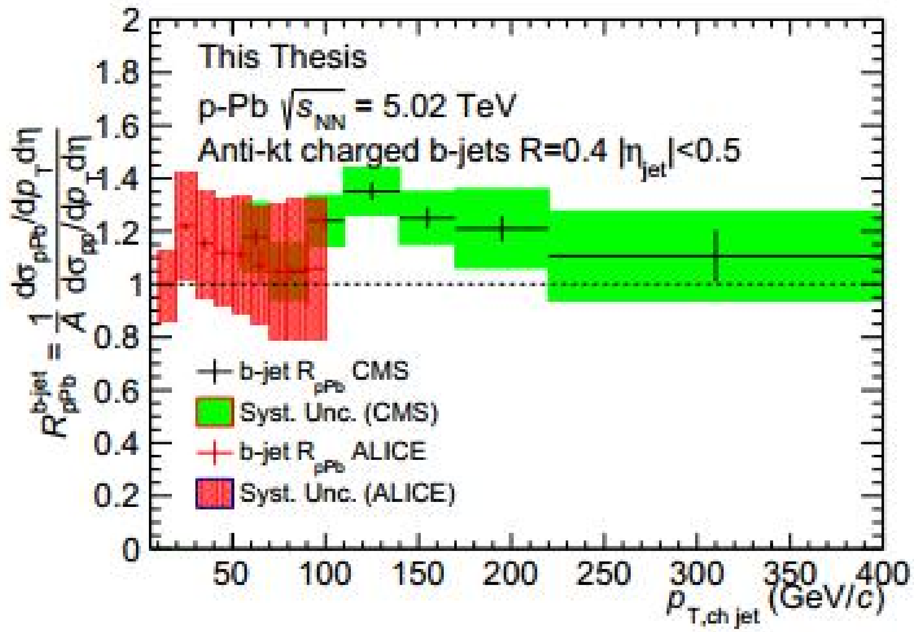
b-fraction



- Production is consistent within N=1, **2** (shown), 3
- Constistent with POWHEG w/ scale variation
- b-jet fraction drops at low-p_T in pPb

R_{pPb} of b-jets





- The R_{pPb} of b-jets is consistent with unity
- ...consistent with CMS measurements
- ...and with theory predictions within uncertainties
- The interaction of the b-jet with the cold nuclear matter has no effect on the b-jet within uncertainties.

Systematics and ToDo

Uncertainty source		p_T bins	
		10-20	**20-30**
p–Pb collisions	Unfolding algorithm	3.14%	
	Regularization parameter	2.03 %	
	Prior	2.09%	2.91%
	Unfolding range	1.40%	1.43%
	δp_T	0.12%	0.29%
	Tracking Efficiency	7.67%	10.60%
	Tagger working point	0.31%	0.24%
	V^0 rejection	0.20%	0.05%
	Normalization uncertainty	3.24%	
	Total	9.47%	11.7%
pp collisions	Unfolding algorithm	3.23%	
	Regularization parameter	3.3%	
	Prior	1.19%	0.19%
	Unfolding range	0%	
	Tracking Efficiency	9.3%	10.6%
	Tagger working point	0.13%	0.36%
	Normalization uncertainty	2.29%	
	Total	10.7%	11.7%

ToDo (Hadi)

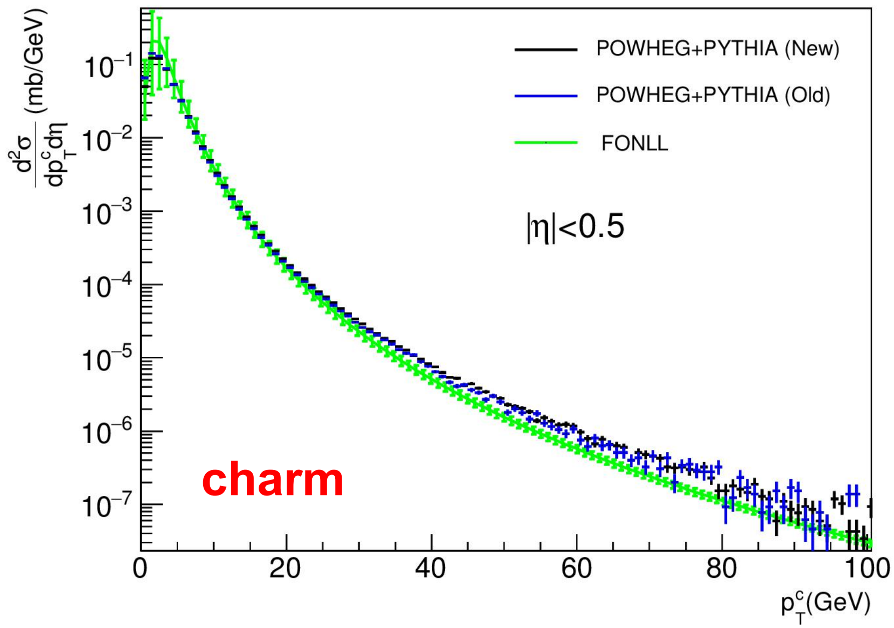
- Change the jet probability distribution, and use another discriminator for the templates used in the tagging efficiency determination.
- Use another distribution to fit the purity.
- Cancel the correlated uncertainties on the b-jet fraction and the R_{pPb}

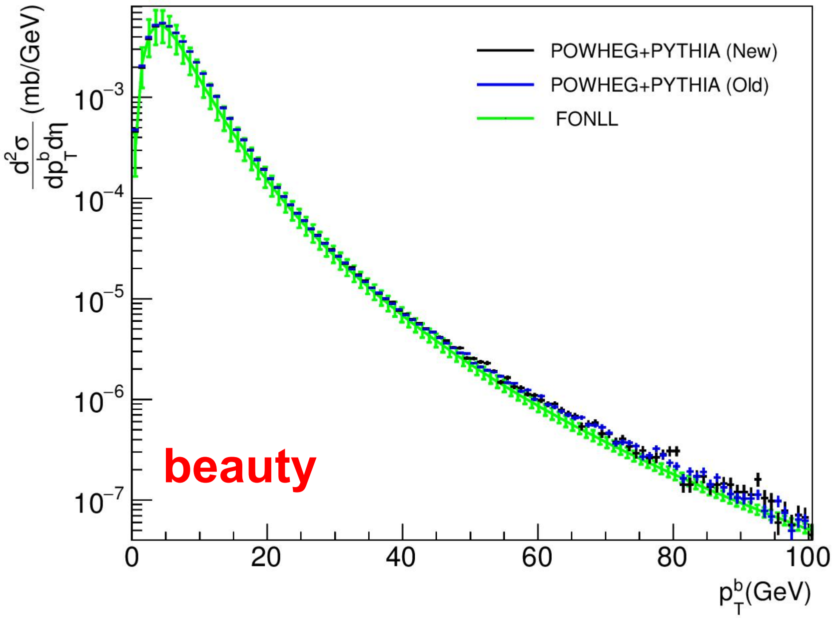
15 March 1848
Hungarian Revolution
against the Habsburg rule

Hungarian Academy of Sciences
Founded 1830
count István Széchenyi

This work has been supported by the Hungarian NKFIH/OTKA K 120660 grant and the János Bolyai scholarship of the Hungarian Academy of Sciences

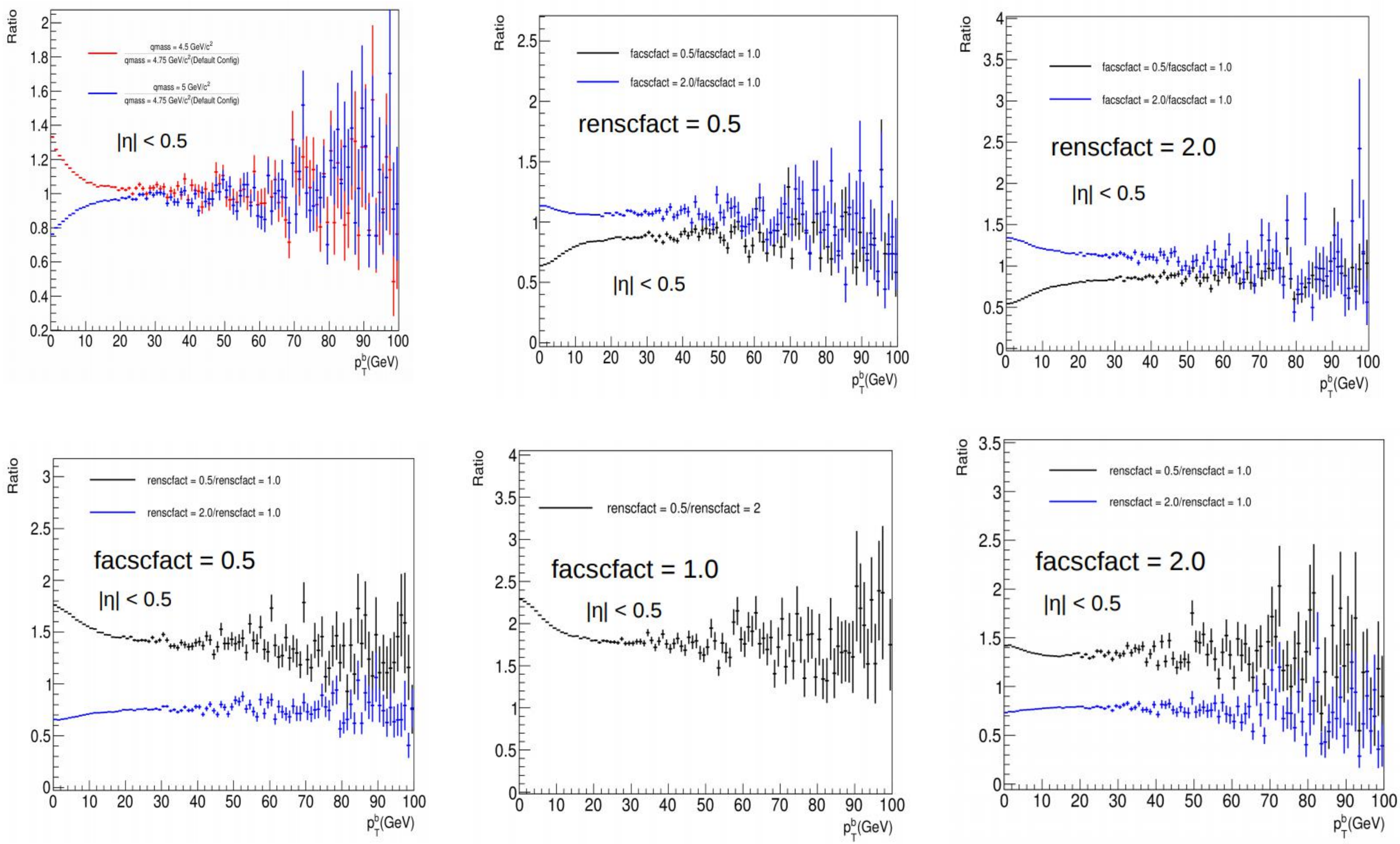
POWHEG simulations



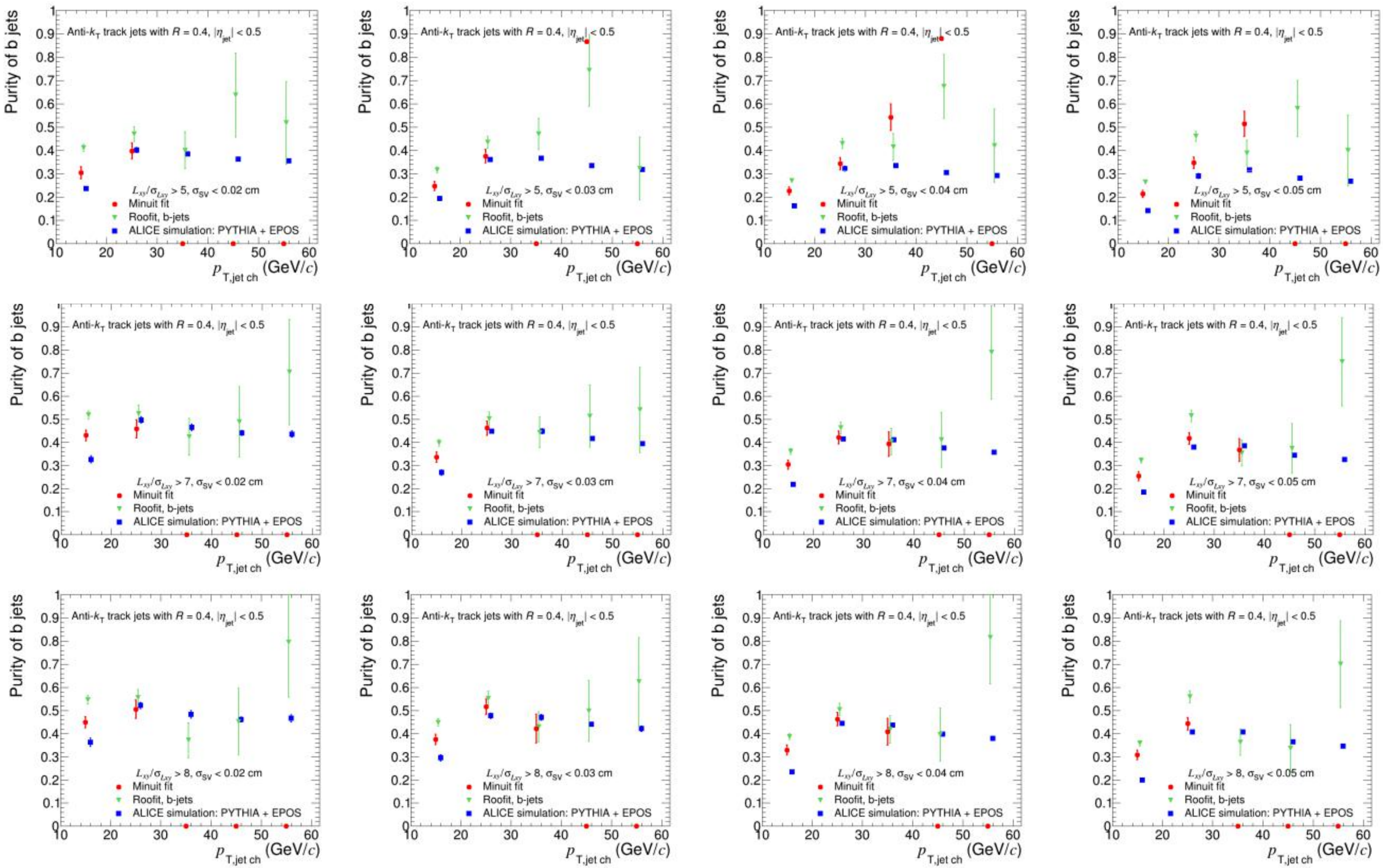


- Changes in "new" since the "old" one:
  - acceptance |eta|<0.5 instead of |eta|<0.6 scaled by 1/1.2
  - more suitable 1-GeV/c binning
  - Lorentz-boost applied
  - p-Pb nPDF applied
- No significant difference between "old" and "new"
- Marginal match to FONLL

POWHEG systematics



Data-driven fits to real data



Unfolding closure test - inclusive

Unfolding of inclusive spectrum of jets filtered PYTHIA

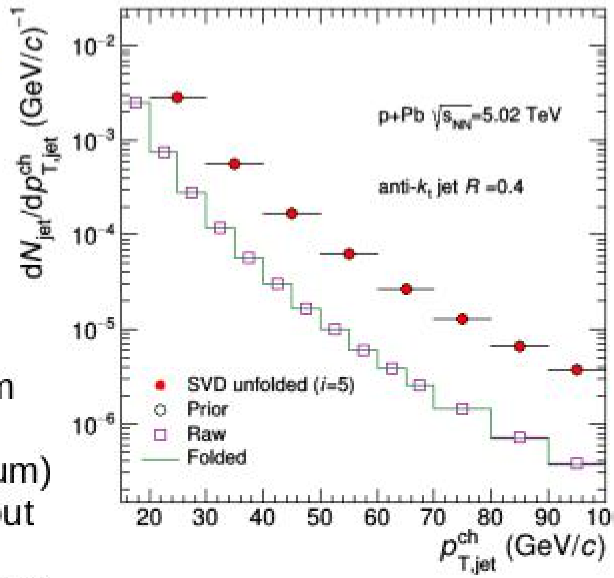
response matrix was filled with inclusive jet

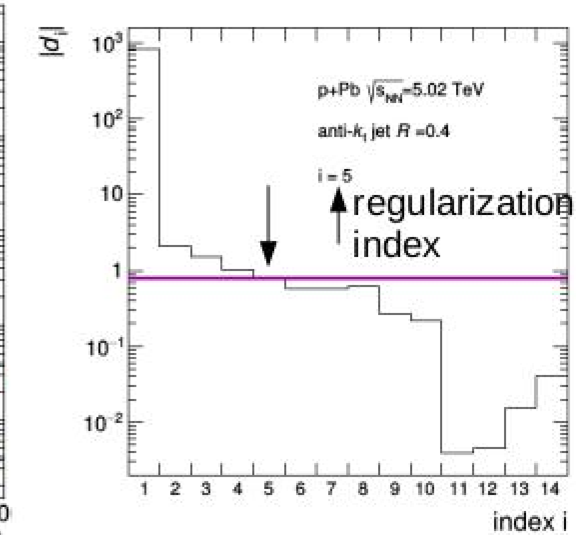
normalization spectrum for response matrix (=here the true spectrum) are inclusive jets without the requirement that the jet was reconstructed

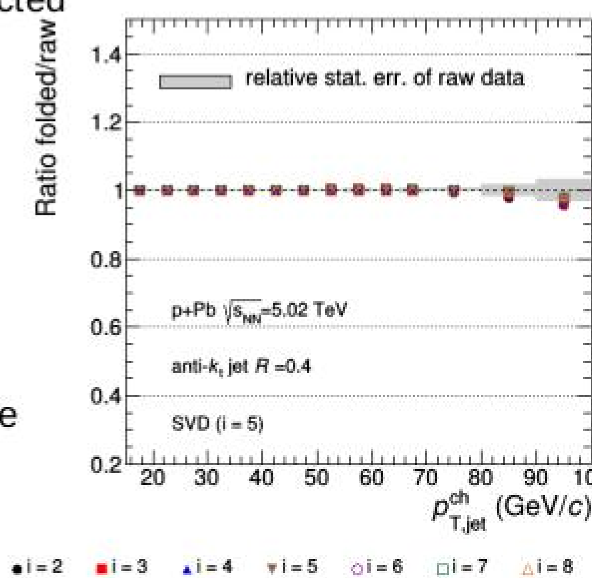
measured spectrum is projection of the response matrix on X axis

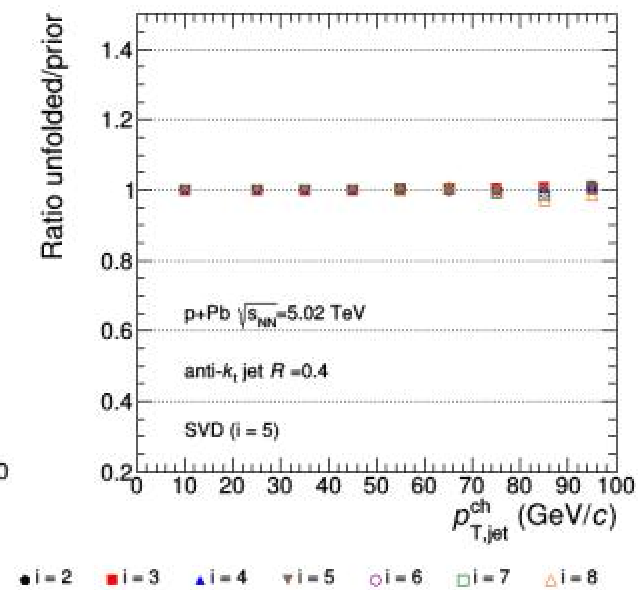
prior spectrum = true spectrum

in this most ideal case everything works

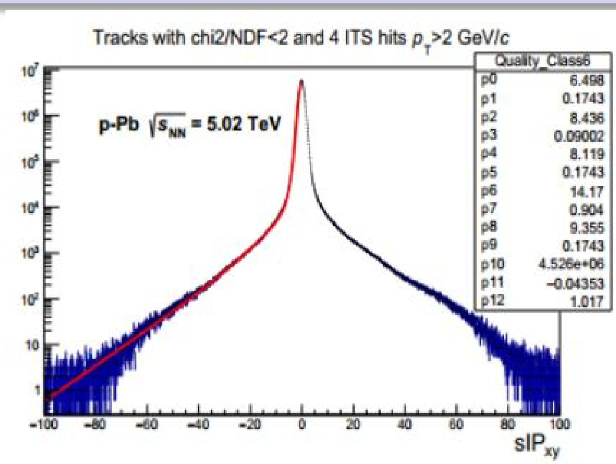


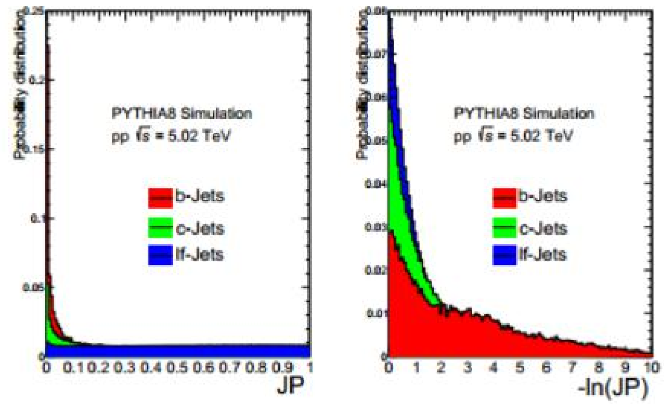






Jet probability algorithm

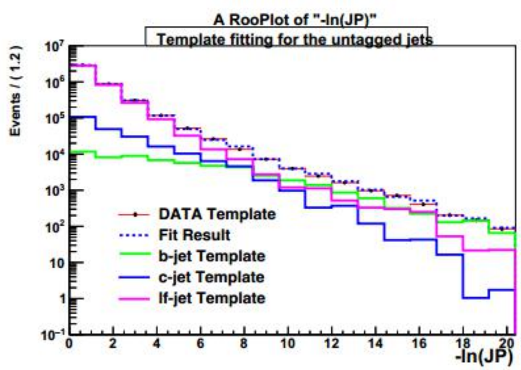


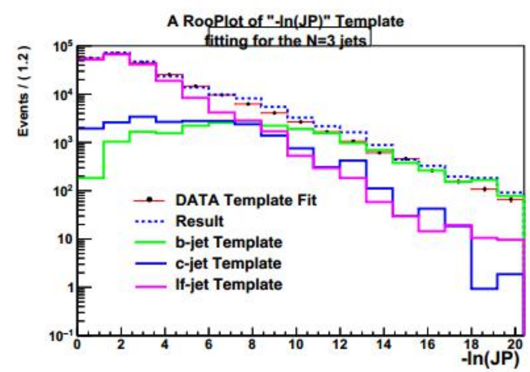


- Calculate the track probability (P_{tr})

$$P_{tr}(S_{d_0}) = \frac{\int_{-\infty}^{-|S_{d_0}|} R(S)dS}{\int_{-\infty}^{0} R(S)dS}$$

- Calculate the jet probability:
  $JP = \prod \times \sum_{k=0}^{N_{trk}-1} \frac{(-\log \prod)^k}{k!}$ where
  $\prod = \prod_{i=1}^{N_{trk}} P_{tr}$



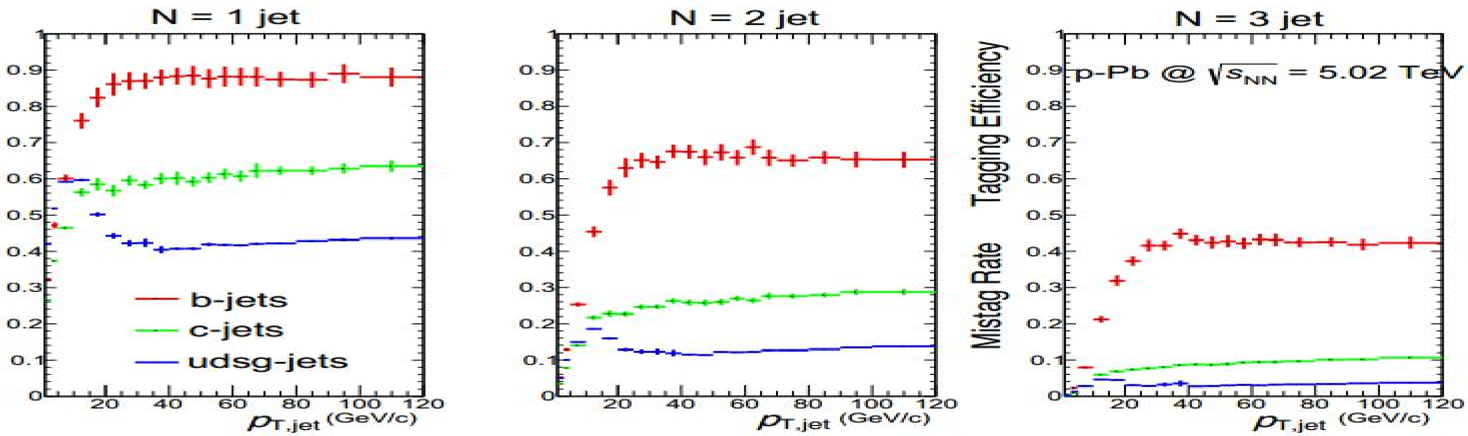


- Construct the $-\ln(JP)$ templates for the b,c, lf-jet from MC
- Apply the track counting tagger.
- Fit to data for tagged and untagged $-\ln(JP)$ distributions.
- Determine efficiency by: $\epsilon_{\mathbf{b}} = \frac{\mathbf{f_b^{tag}.N_{data}^{tag}}}{\mathbf{f_b^{untag}.N_{data}^{untag}}}\mathbf{C_b}$.

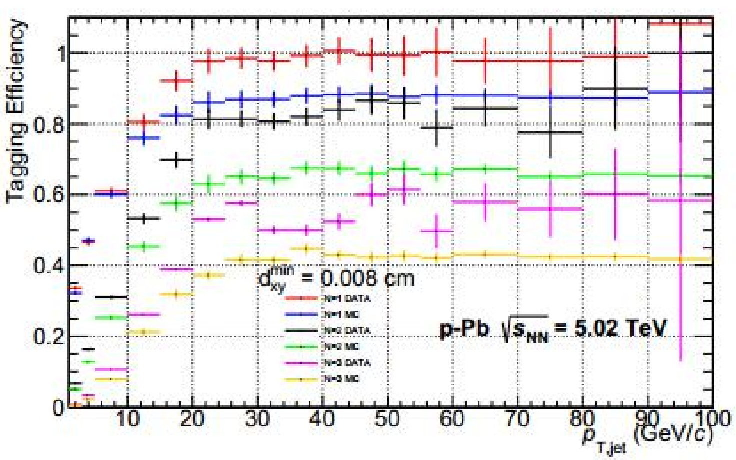
Efficiencies

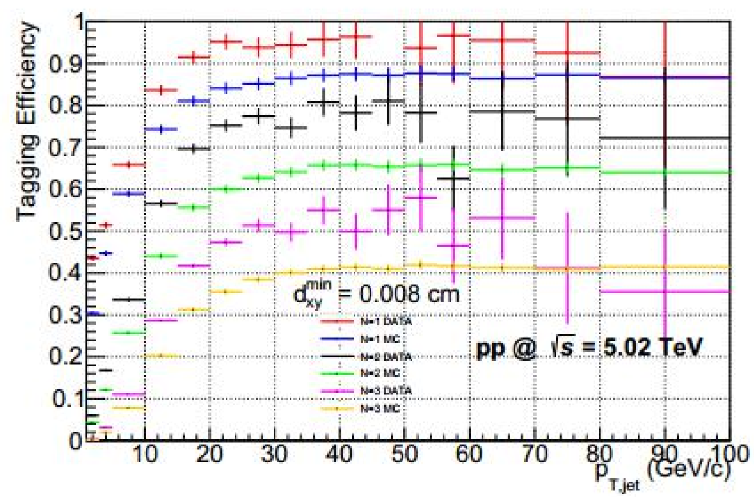
- MC efficiency

$$\epsilon_b(p_{T,ch.jet}^{det}) = \frac{N_b^{tagged}(p_{T,ch.jet}^{det})}{N_b^{Total}(p_{T,ch.jet}^{det})}$$



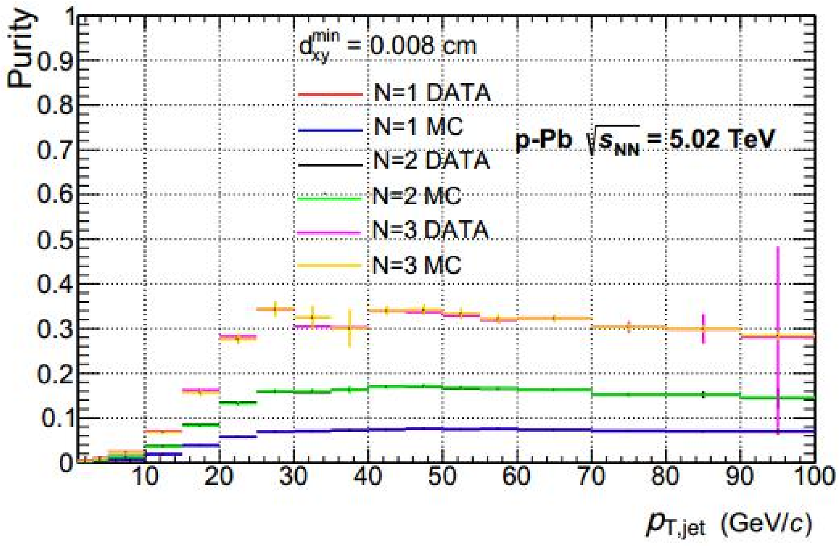
- Data-driven efficiency underestimated my MC!

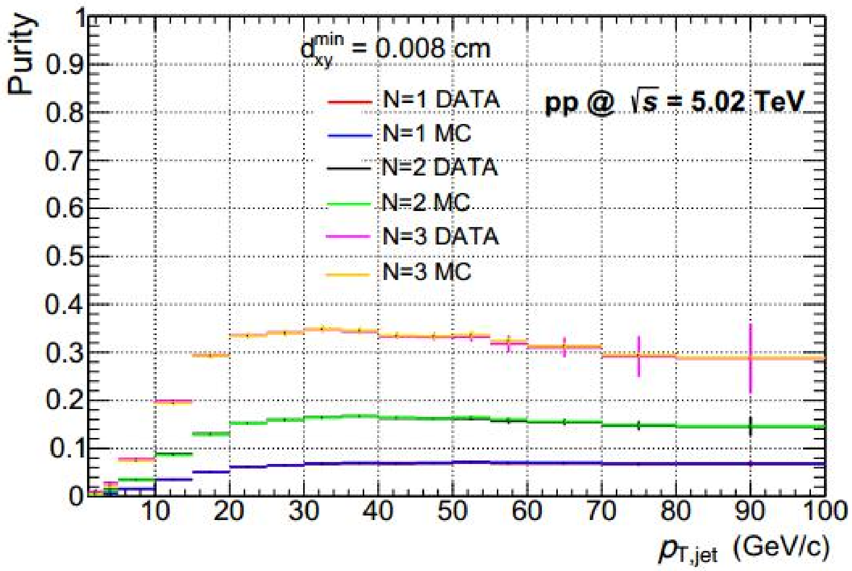




b-jet purity

- b-jet purity: $\mathrm{P} = \frac{N^{tagged}_{b-jet}}{N^{tagged}_{total}}$.
- N^{tagged}_{b-jet} can be estimated using data-driven methods:
  - by fitting templates for the JP.

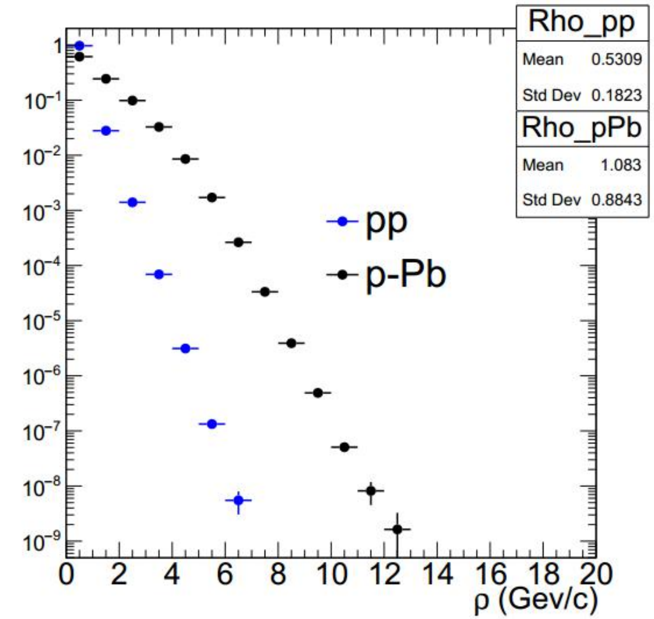


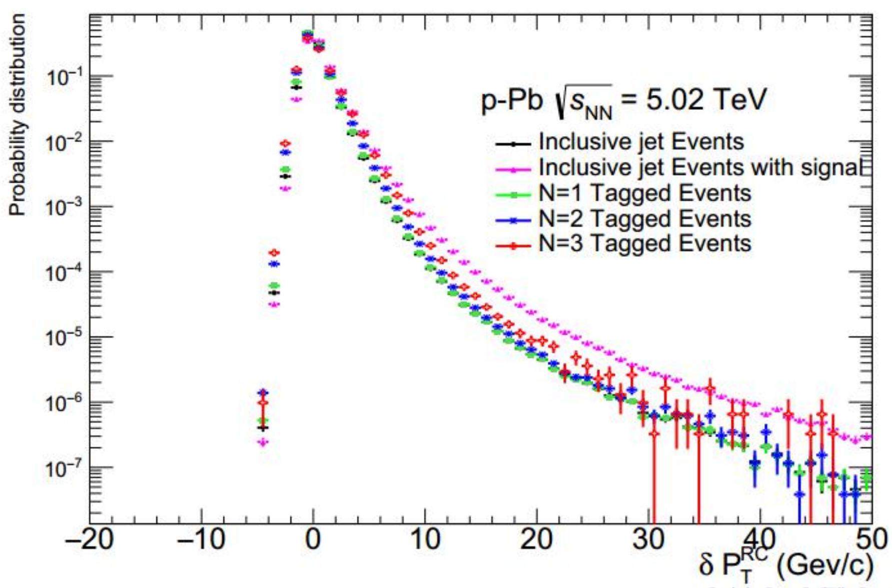


- b-jet tagging purity is consistent between data and MC.

Underlying event

- UE density:
  $\rho = median(\frac{p_{T,i}}{A_i}).C$, where
  $C = \frac{CoveredArea}{TotalArea}$.
- Correct the jet p_T:
  ${p_{T,j}}^{Sub} = p_{T,j} - \rho A_j$.





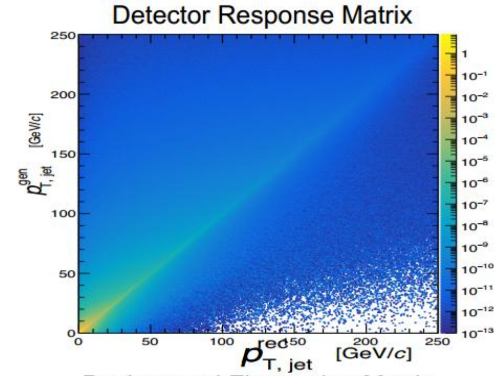
- UE fluctuation for unfolding
- Random cone method

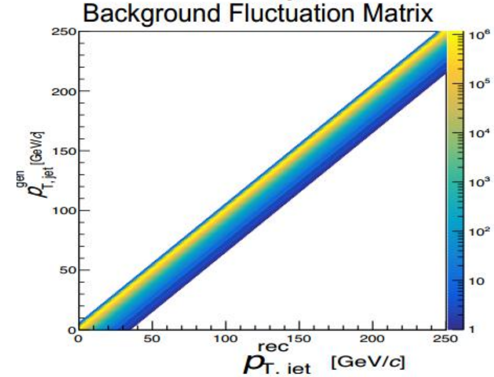
$$\delta p_T = p_T^{RC} - \rho\pi R^2$$

  - If overlap with signal jet, throw again

Unfolding

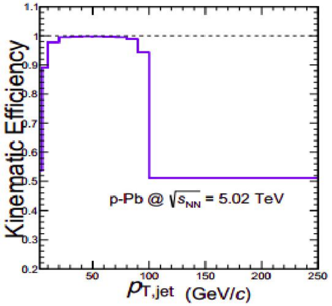
- Correction for detector effects.
  ⇒ Detector response (DR) matrix is needed.
- The DR is built by matching jets at the detector level to jet in the generated level: $\Delta R_{jet1,jet2} = \sqrt{(\eta_{jet1} - \eta_{jet2})^2 + (\phi_{jet1} + \phi_{jet2})^2} < 0.25$
- Correction for UE fluctuations (for p–Pb collisions).
  ⇒ background fluctuation (F) matrix need,
- The F matrix built from the δp_T distribution.
- The SVD unfolding was used (A. Hoecker et al).
- Prior: PYTHIA b-jet spectrum (jet-jet MC).





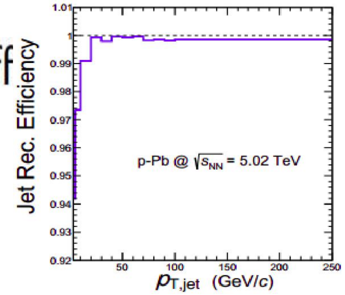
- Combine both matrices for p–Pb : $R = F \times DR$.
- Closure test shows that the measured spectrum is correctly unfolded

- Correct for kinematic efficiency: fraction of remaining jets after rebinning.

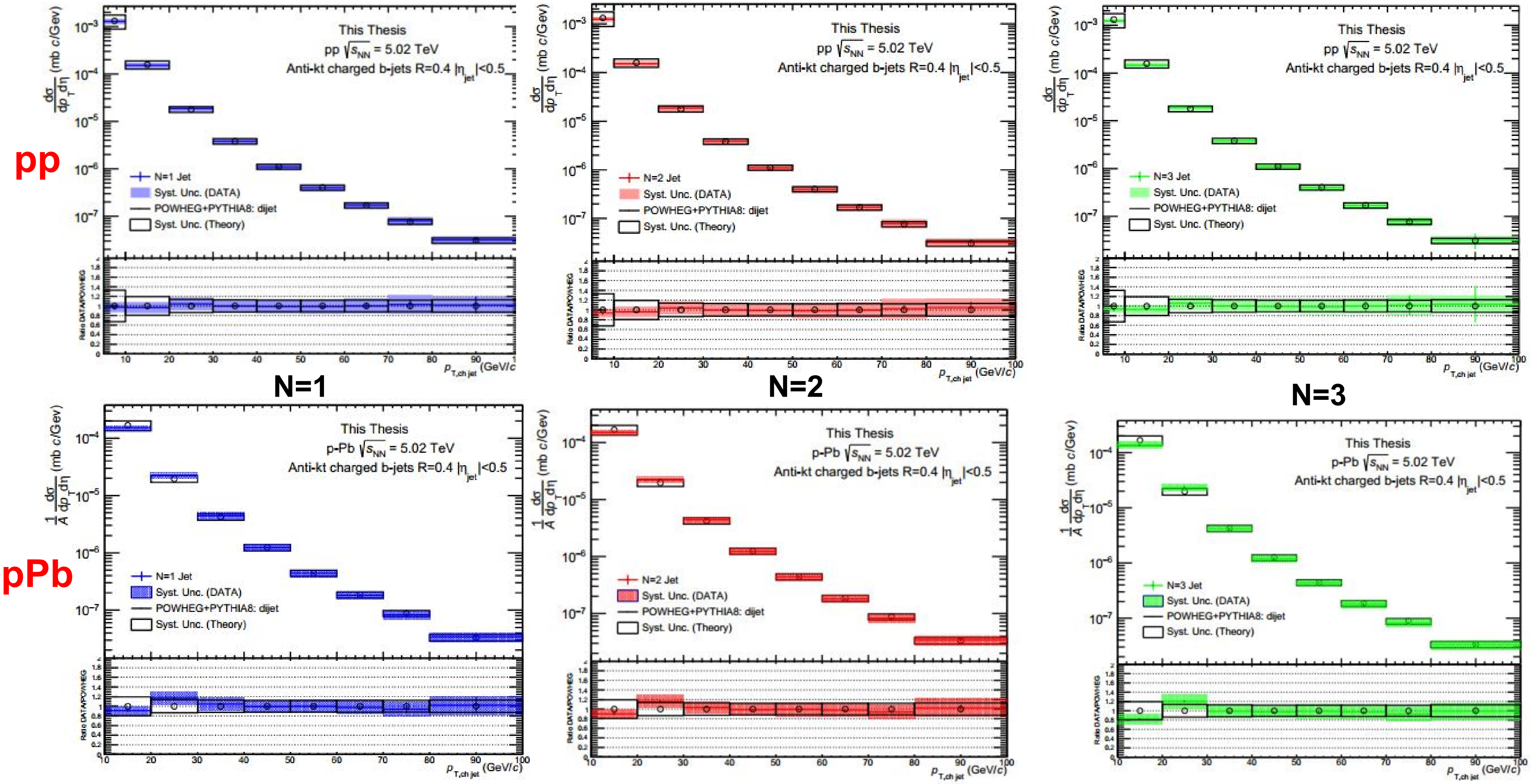


- b-jet reconstruction eff

$$\epsilon_{jet}^{rec} = \frac{N_{jet}^{matched}}{N_{jet}^{total}}$$

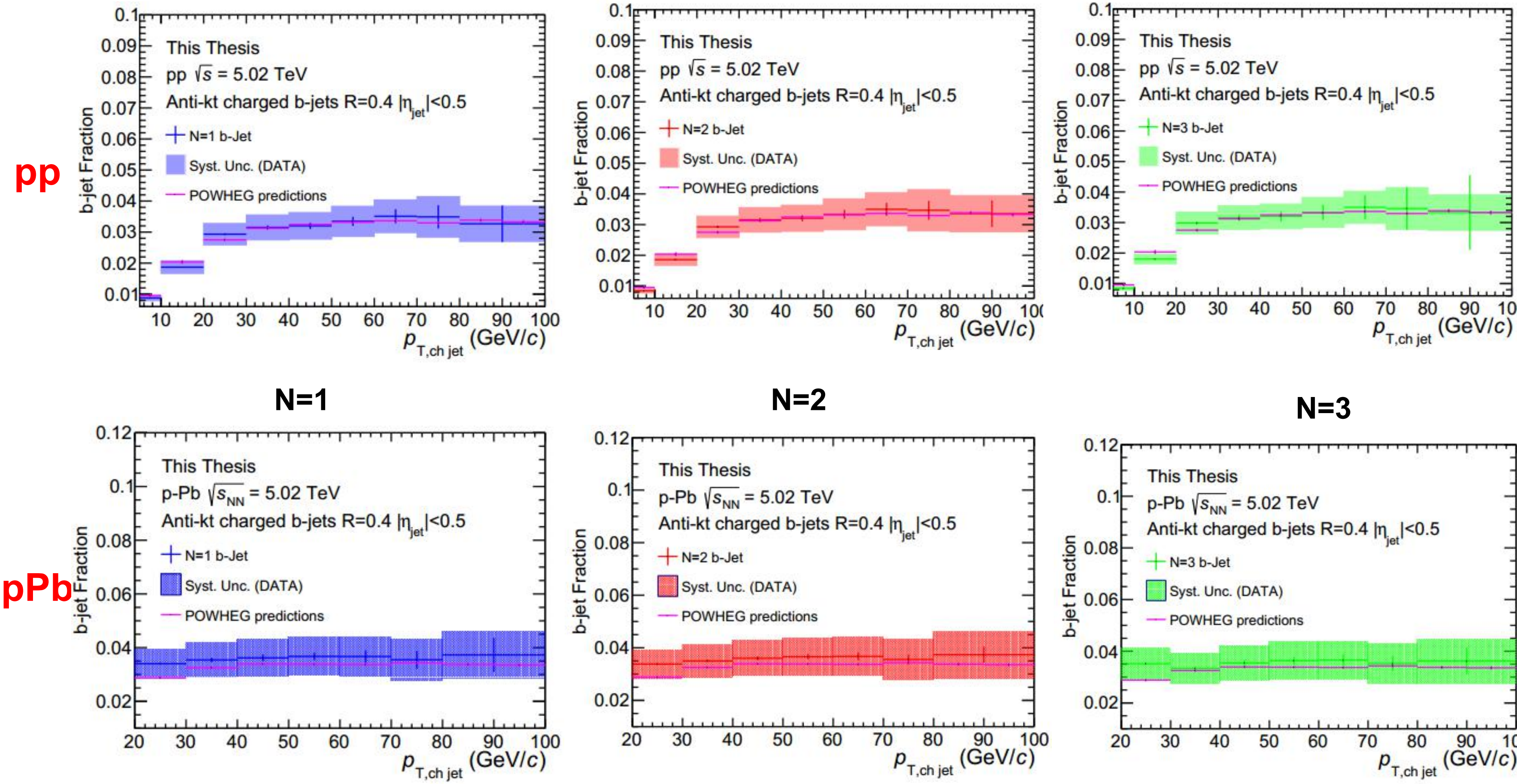


Production cross-section (pp and pPb)



- Production is consistent within N=1,2,3
- Constistent with POWHEG w/ scale variation

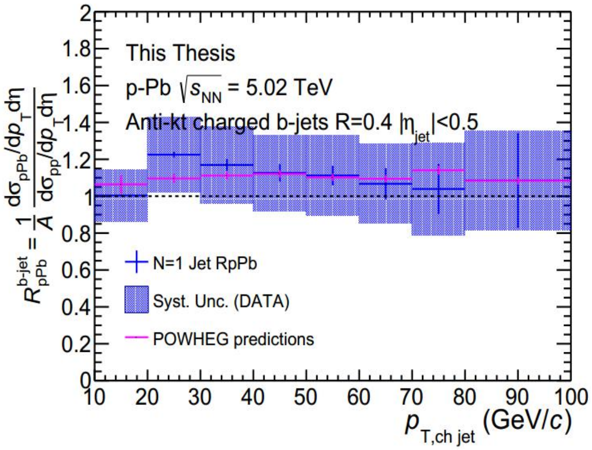
b-jet fraction (pp and pPb)



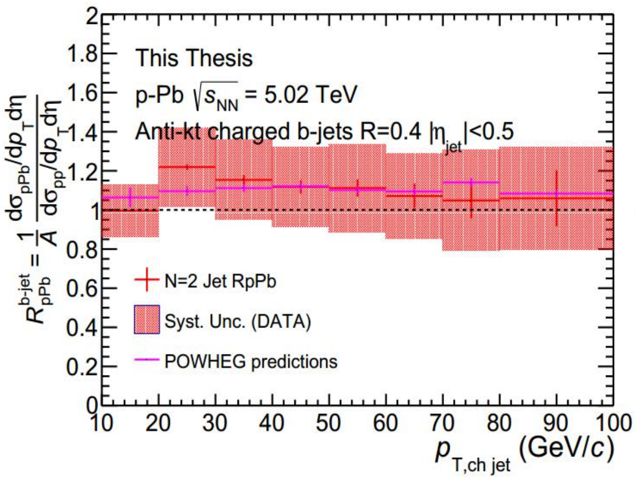
- Production is consistent within N=1,2,3 and POWHEG

R_{pPb} of b-jets

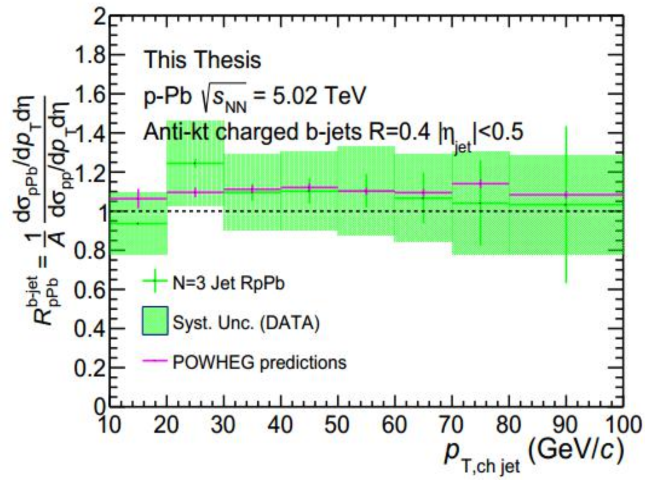
N=1



N=2

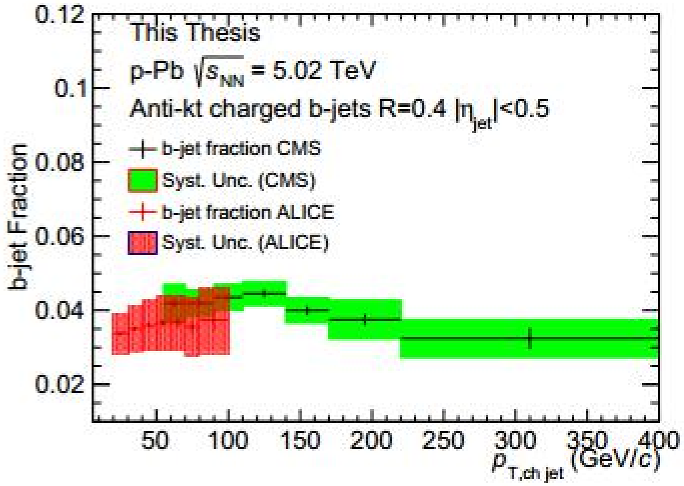


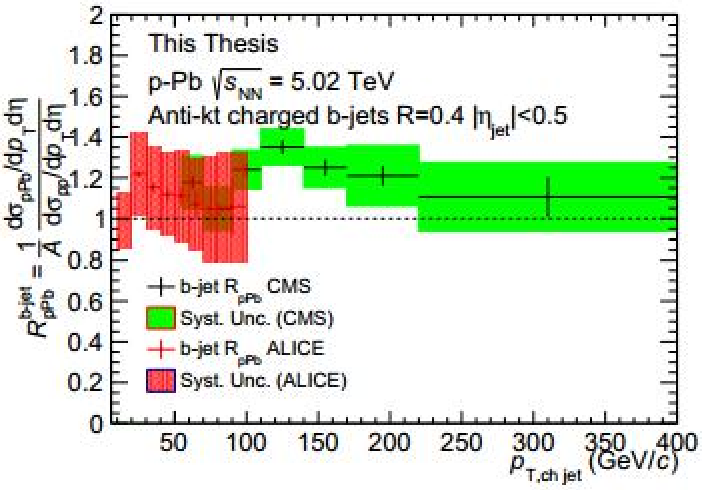
N=3



- The R_{pPb} of b-jets for N=1,2,3 is...
  - consistent with unity
  - and with theory predictions within uncertainties
- The interaction of the b-jet with the cold nuclear matter has no effect on the b-jet within uncertainties.

Comparison to CMS and ToDo





ToDo (Hadi)

- Change the jet probability distribution, and use another discriminator for the templates used in the tagging efficiency determination.
- Use another distribution to fit the purity.
- Cancel the correlated uncertainties on the b-jet fraction and the R_{pPb}