HF jets analysis

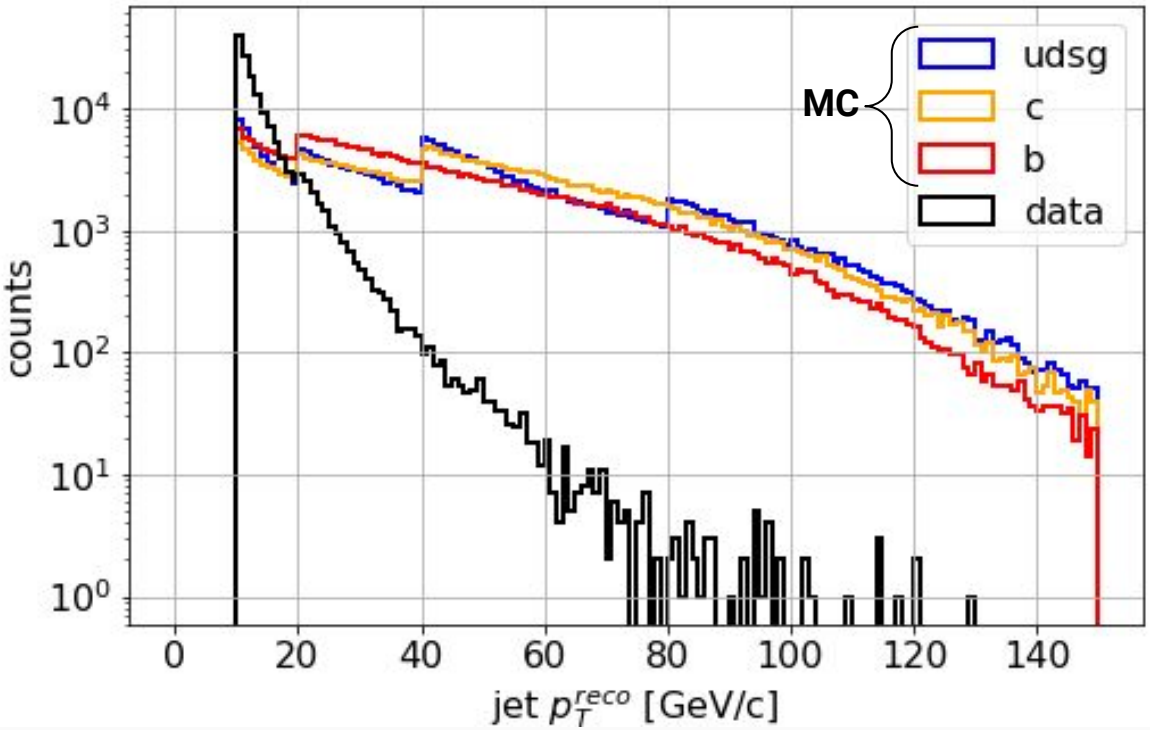
23.03.2020 ALICE@IFJ meeting

Sebastian Bysiak

Outline

1. Progress in analysis
    - model applied on data
        - 1D & 2D score distribution
        - c- and b-fraction estimation
    - data-vs-MC classifiaction
2. Questions & issues
3. Plans for next week

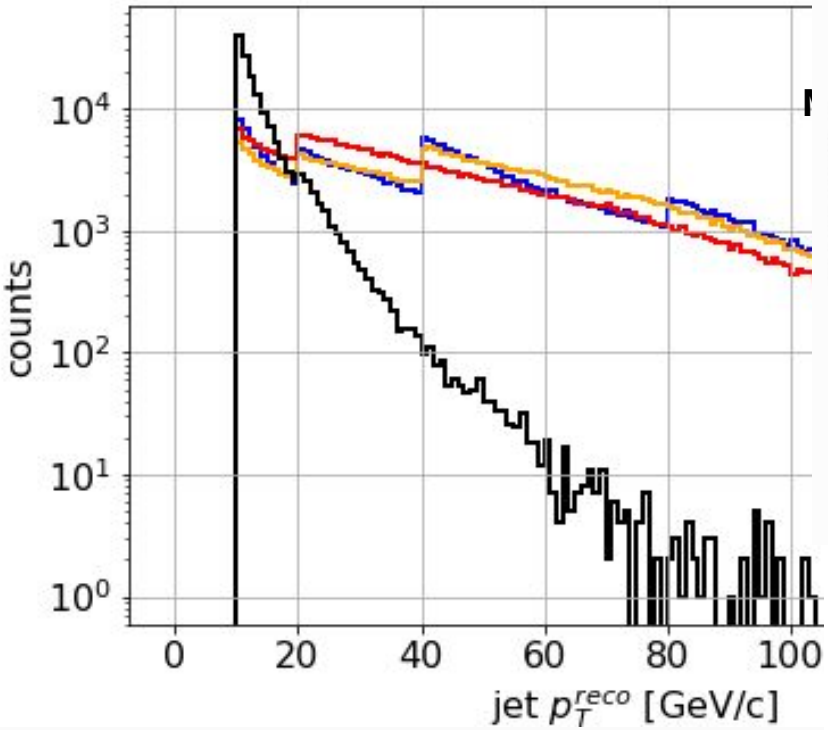
pT spectrum & statistics in data

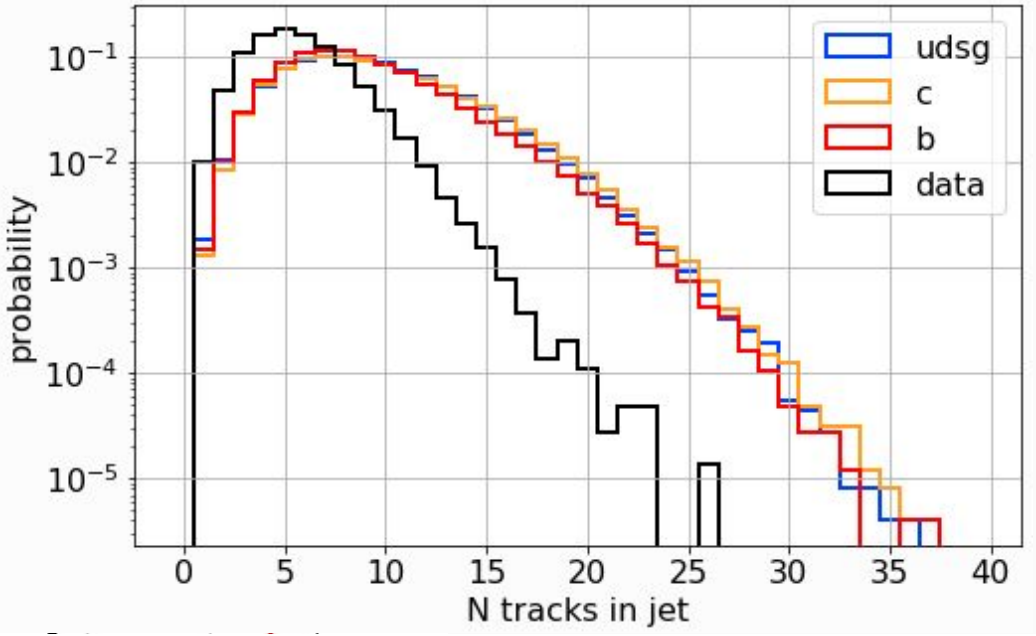


- pT spectrum in data is much more steep than in MC (hard-pt-bin production)
- steps on MC due to downsampling (e.g. b-jets were not downsampled)
- statistics in data (LHC15n):

10-20 GeV/c --	129000
20-30 GeV/c --	15000
30-40 GeV/c --	2500
40-50 GeV/c --	700
> 50 GeV/c --	450

pT spectrum & statistics in data

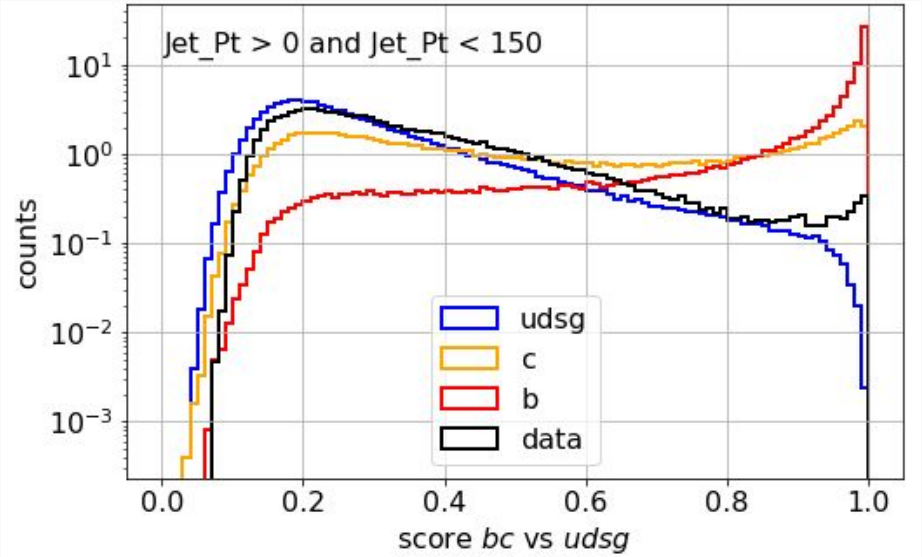


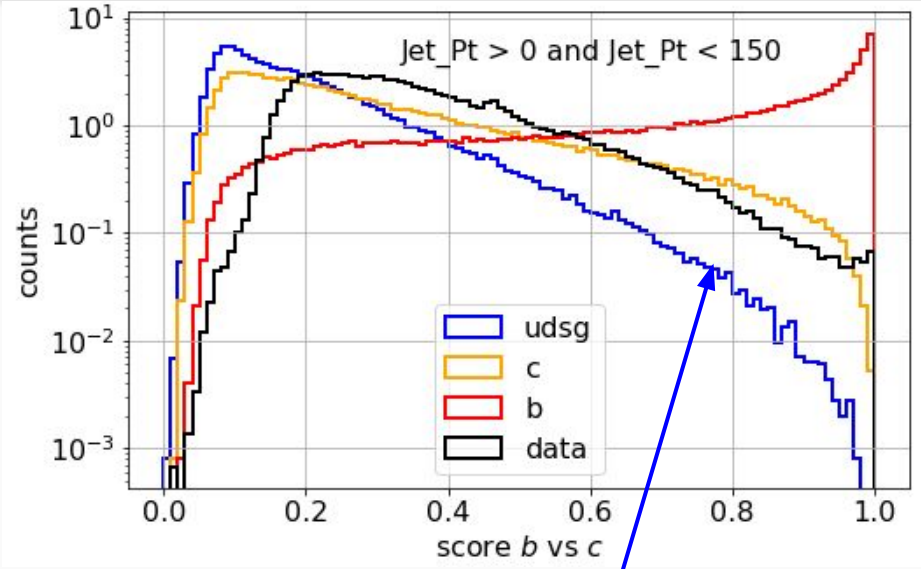


it impacts many other observables so has to be taken into account

example above: NumTracks - difference between data & MC but no difference when only jets with similar pT selected

score distributions: any pT

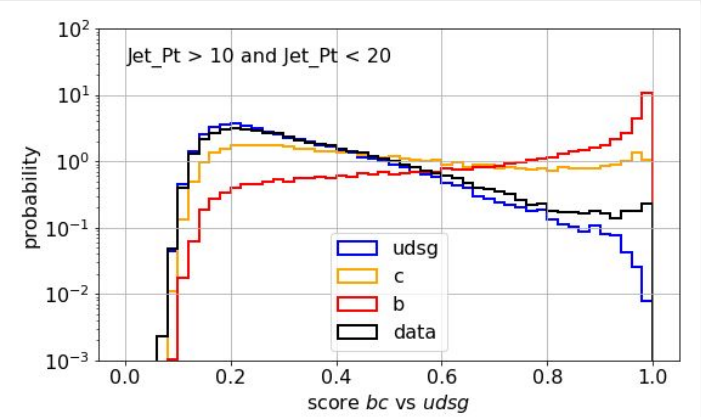


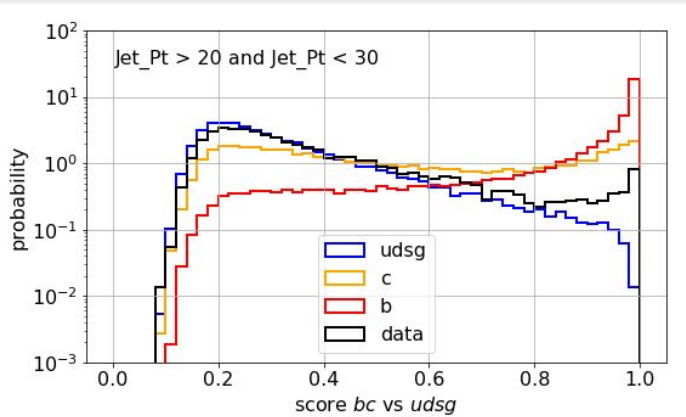


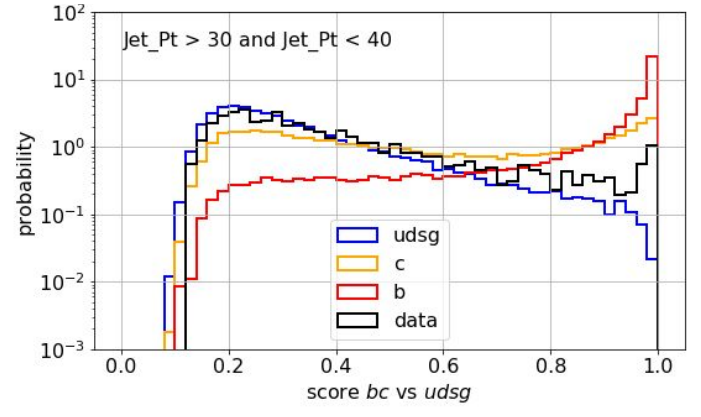
large separation of udsg-jets is just a side effect as this model was trained only on b- and c-jets

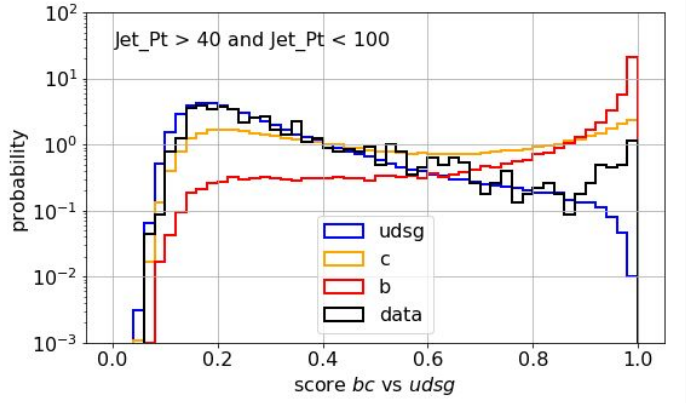
In score distr. of model “bc-vs-udsg” (left) **data** looks more less like *udsg* with small admixture of *b* and *c*
In score distr. of model “b-vs-c” (right) it does not => requires pT differential view

score distributions: “bc-vs-udsg” (logscale)





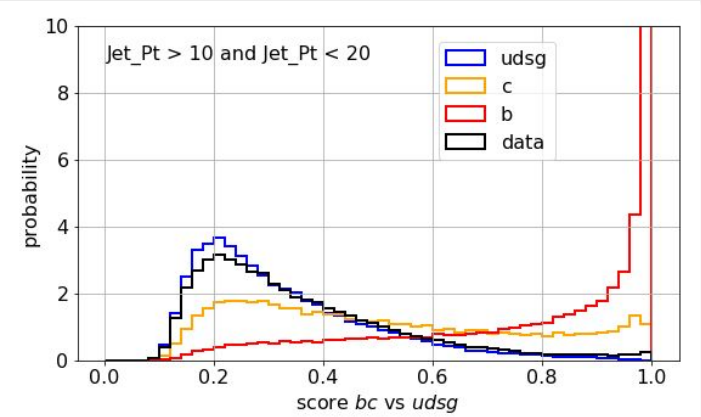


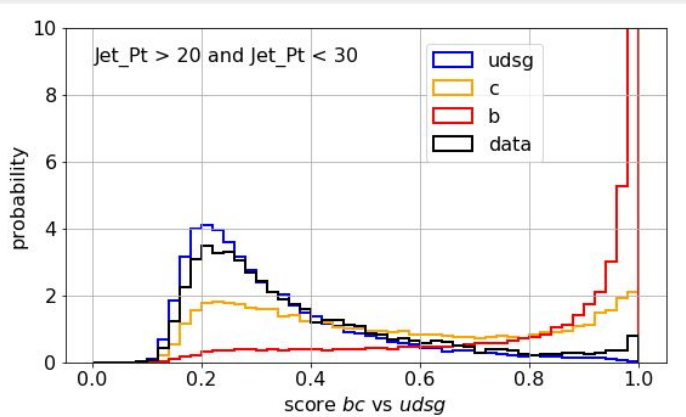


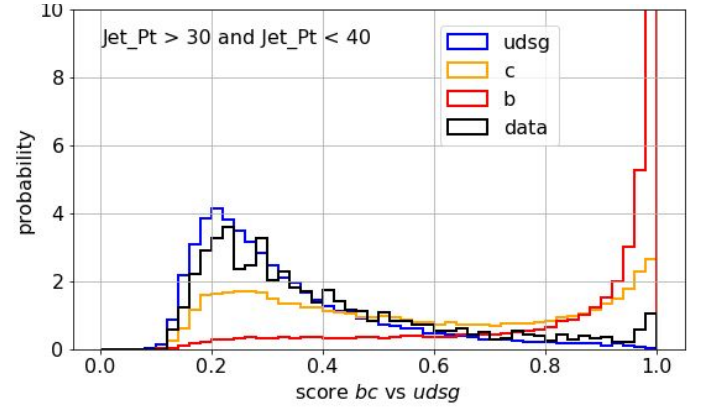
data looks more less like *udsg* with small admixture of *b* and *c*

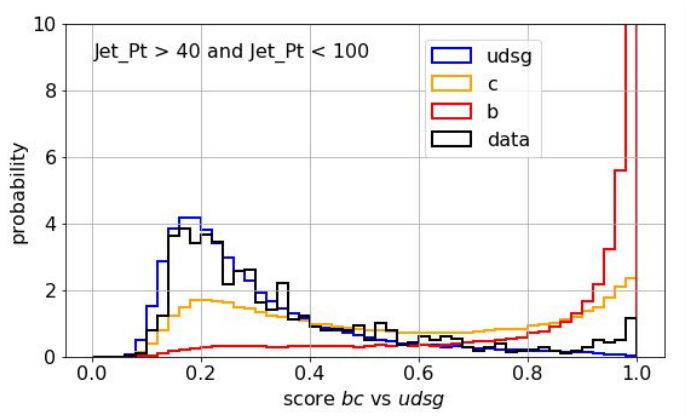
maybe even this admixtures grow with pT (this we expect)

score distributions: “bc-vs-udsg” (linscale)





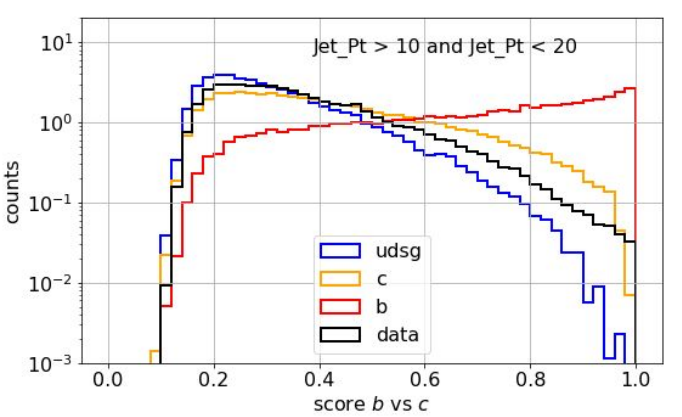


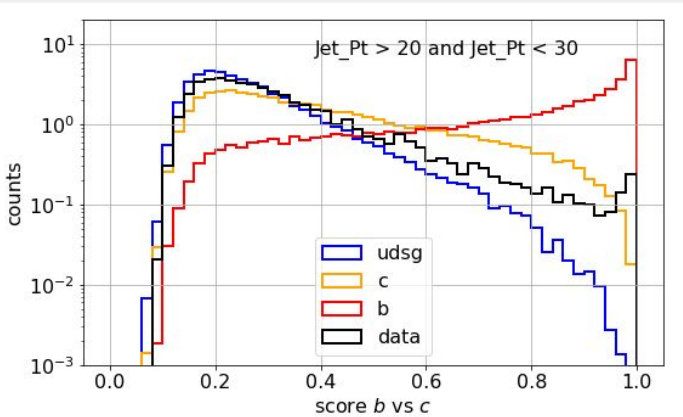


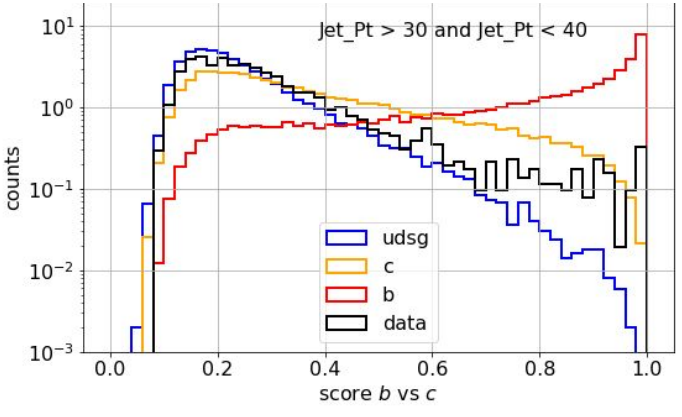
data looks more less like *udsg* with small admixture of *b* and *c*

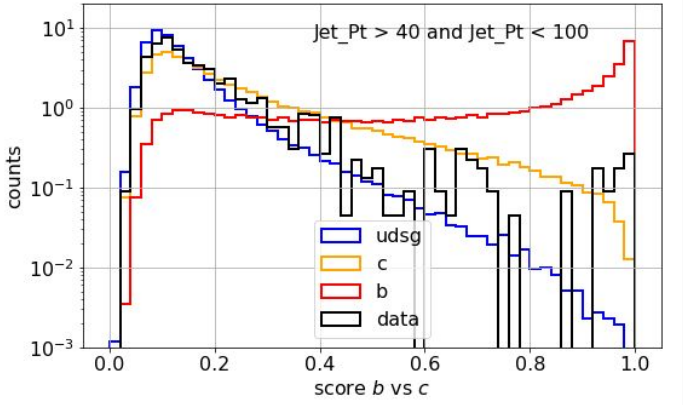
maybe even this admixtures grow with pT (this we expect)

score distributions: “b-vs-c” (logscale)





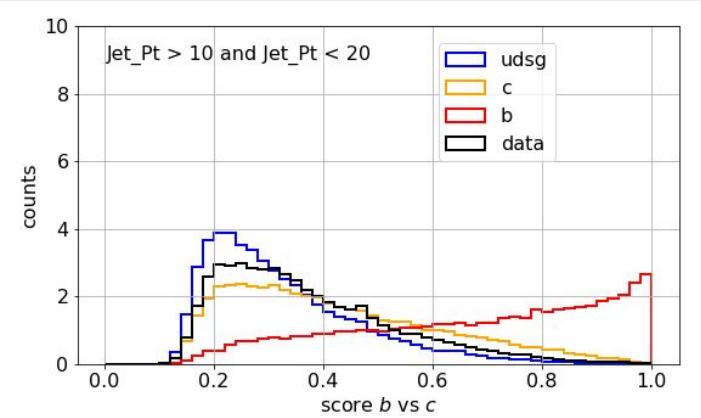


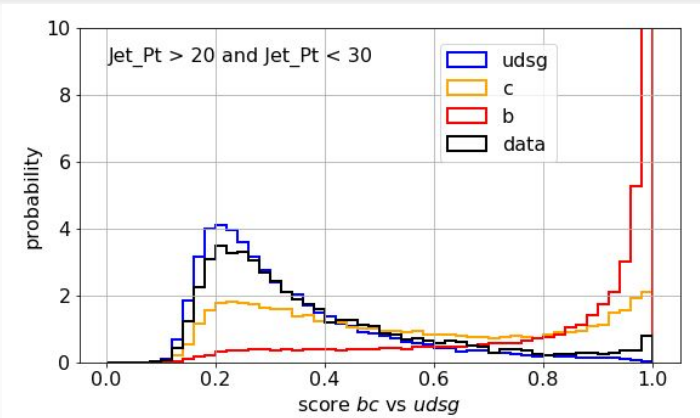


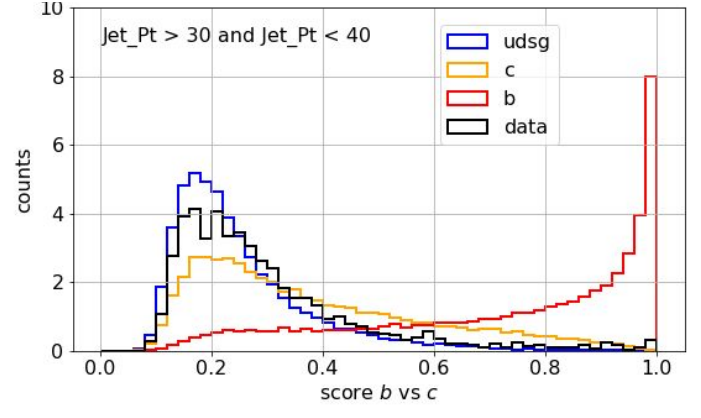
data looks more less like *udsg* with small admixture of *b* and *c*

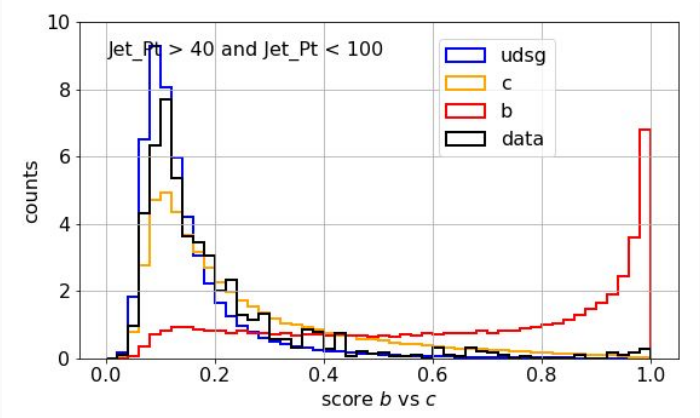
maybe even this admixtures grow with pT (this we expect)

score distributions: “b-vs-c” (linscale)







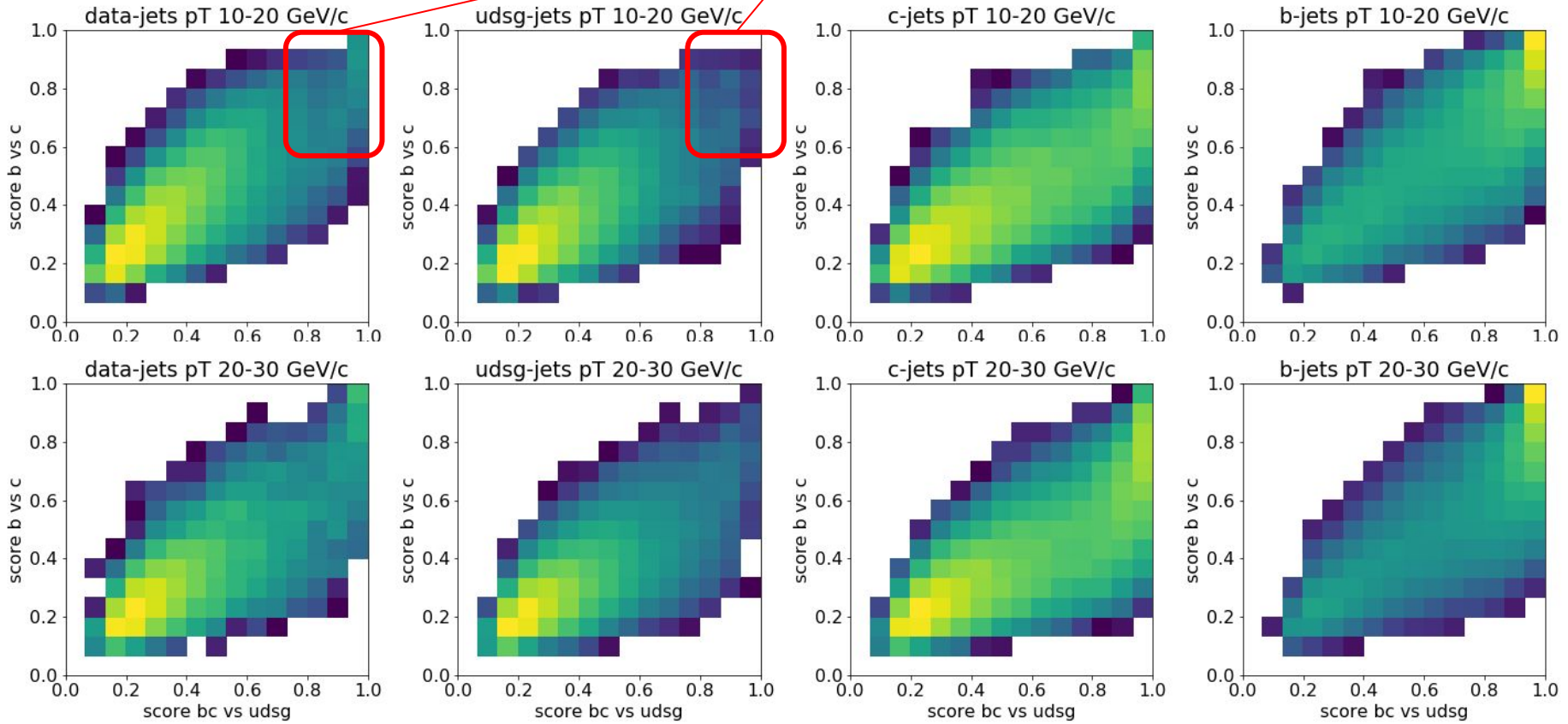


data looks more less like *udsg* with small admixture of *b* and *c*

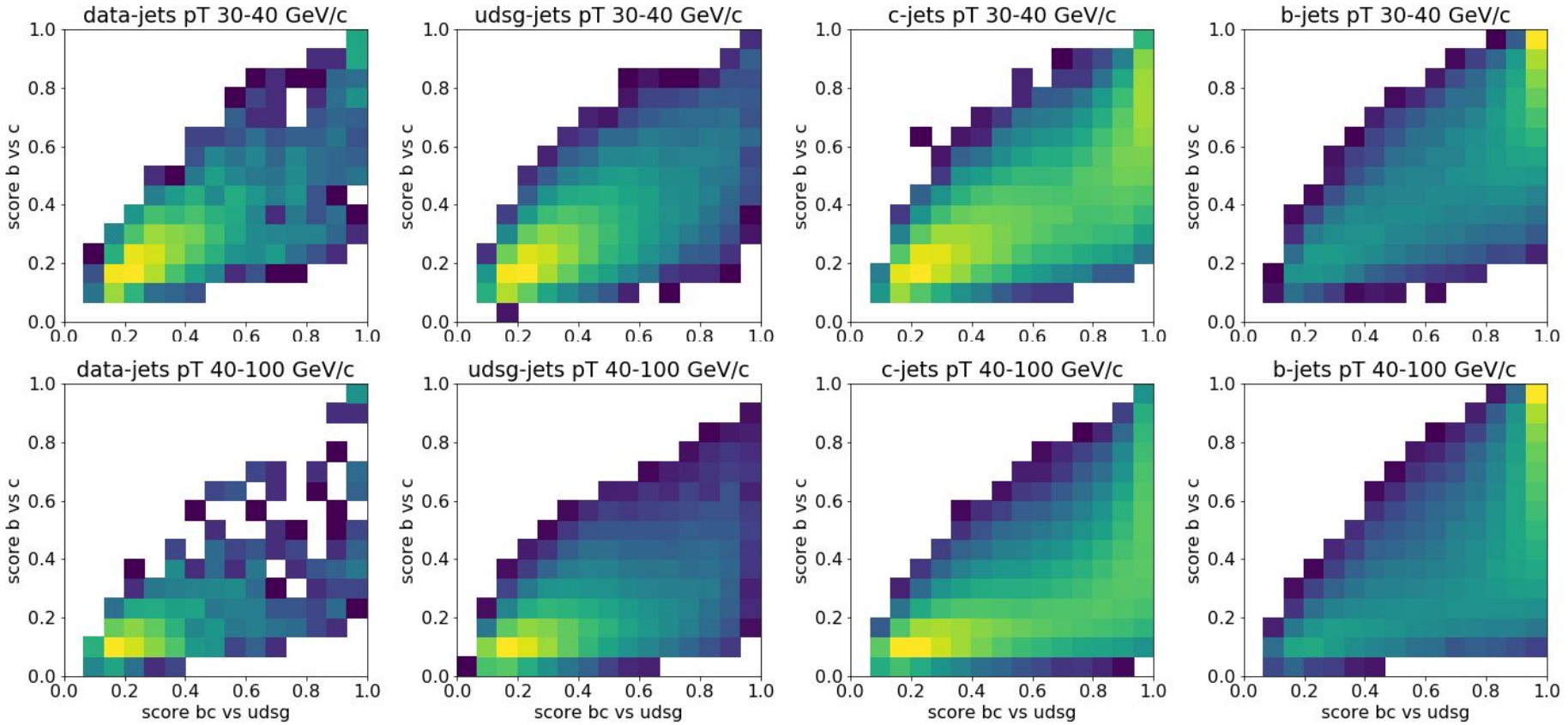
maybe even this admixtures grow with pT (this we expect)

score distributions: 2D

data looks like *udsg* + small admixture of heavy flavour



score distributions: 2D



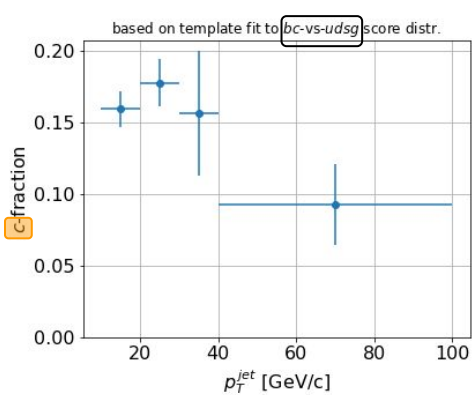
b- and c-fraction estimation

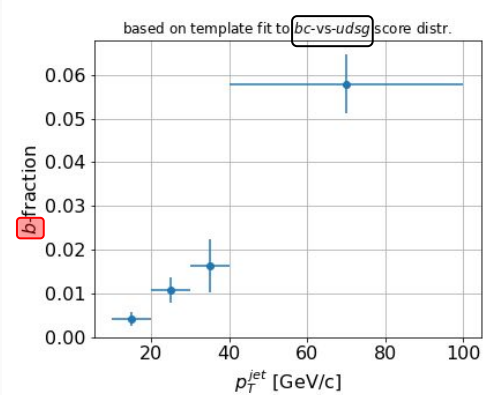
Approach to estimate frac. of c- and b-jets in the sample:

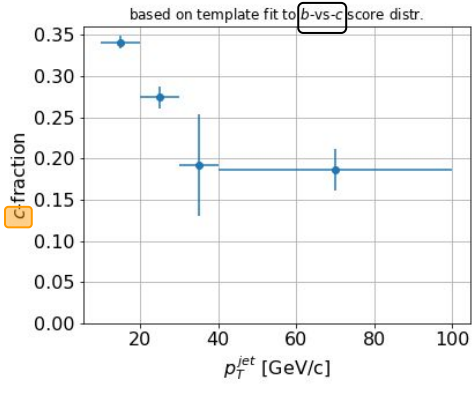
use 1D scores distributions for *udsg*, *c* and *b* as templates to fit data distribution

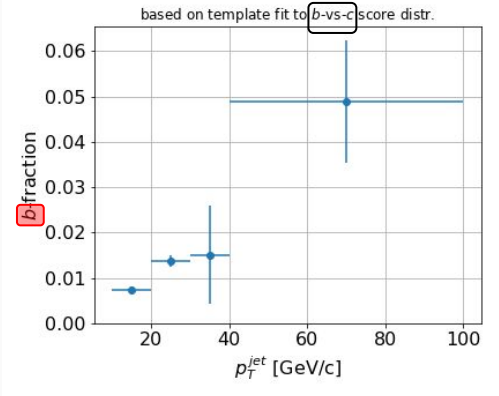
LINK TO PLOTS

b- and c-fraction estimation









- very rough estimation of *c*- and *b*-fraction
- errorbars = stability of the fit
  stddev when some fit parameters varied:
  #bins = 30,50,100
  bin edges +/- 10%
- *c*-fraction:
  - templates quite similar to *udsg*
  - results from two fits not consistent
  - drops with pT - not expected behaviour
  - --> probably wrong
- *b*-fraction:
  - completely different template shape
  - assigned weight strictly related to last couple bins where other templates are much smaller
  - reasonable behaviour & numerical values compared to Fig. 59-64 in:
    https://alice-notes.web.cern.ch/system/files/notes/analysis/982/2019-10-06-ALICE_analysis_note.pdf

b- and c-fraction estimation

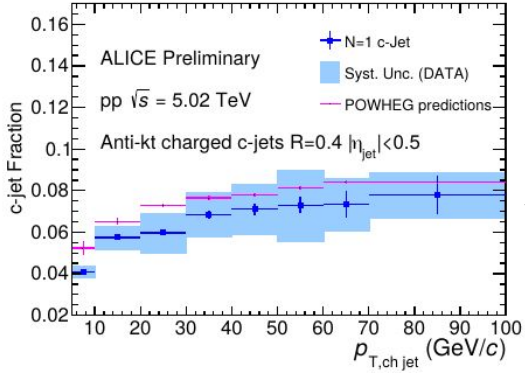


Fig. 62: The N=1 c-jet fraction in pp collisions at $\sqrt{s_{NN}} = 5.02$ TeV.

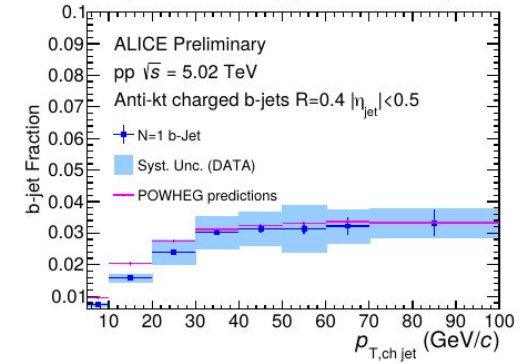


Fig. 59: The N=1 b-jet fraction in pp collisions at $\sqrt{s_{NN}} = 5.02$ TeV.

- very rough estimation of *c*- and *b*-fraction
- errorbars = stability of the fit
  stddev when some fit parameters varied:
  #bins = 30,50,100
  bin edges +/- 10%
- *c*-fraction:
  - templates quite similar to *udsg*
  - results from two fits not consistent
  - drops with pT - not expected behaviour
  - --> probably wrong
- *b*-fraction:
  - completely different template shape
  - assigned weight strictly related to last couple bins where other templates are much smaller
  - reasonable behaviour & numerical values compared to Fig. 59-64 in:
    https://alice-notes.web.cern.ch/system/files/notes/analysis/982/2019-10-06-ALICE_analysis_note.pdf

Data-vs-MC classification [WIP]

model similar to those used for bc-vs-udsg and b-vs-c classification

performance reached: ROC AUC = 0.75, accuracy = 68% - quite high

but:

- feature_importance analysis shows that contribution of all columns is between 1% and 3%, which means there are no features which differ significantly in data and MC.
  Such behaviour is typical when we try to fit some random data or with random labels
- with limited depth of the tree, e.g. to 3-5, the performance was much worse

Plans for the next week

1. Step back:
   - plot distributions of observables like Lxy, IP
     plot it also whenever I classify anything in the data
2. HF mesons reconstruction:
   - use PDG (if stored) to check if there was D meson or J/Psi in jet
   - invariant mass (all tracks OR juts with proper PID: D -> pi+K)
3. Data-MC classifier
   - if needed -> reweighting
   - visualize features first (see point 1)
4. Extend template fits (?):
   - use fit to some physical observable, like Lxy instead
5. SV - selection criteria, needed before plotting SV features other than columns, like `Jet_SecVtx_1_Lxy__sortby__LxyNsigma__desc`

BACKUP

Data-vs-MC classification

model similar to those used for bc-vs-udsg and b-vs-c classification

performance reached: ROC AUC = 0.75, accuracy = 68% - quite high

but: feature_importance analysis shows that contribution of all columns is between 1% and 3%, which means there are no features which differ significantly in data and MC

- pT aligned by sampling - show plot
- maybe dealing with None is required for good perf. that's why MLP and GB cannot make it (GB really cannot? - check XGB and GB with similar key params)
- data leakage: same examples in train and test set - how to deal with that?

Plans for the next week [OLD]

1. Apply on data:
   - 1D & 2D score distribution
   - check b- and c-jets fractions (vs pT)
   - train classifier: MC vs data
2. Design selection criteria for SV (just sort by chi2 ?), at least vizualize sth

- check available MC prod. for LHC17p, LHC17q -- hard pt-bins, hf-enhanced
- model improving (ML-side): PCA before BDT, feat. eng., incl. jet shapes, vary sorting, N_tracks and N_Sv
- model improving (physics-side):
  technically easy: PID (e.g. e- and its energy)
  technically hard: Lund diagram, D and B meson reconstruction