HF jets analysis

09.03.2020 ALICE@IFJ meeting

Sebastian Bysiak

Outline

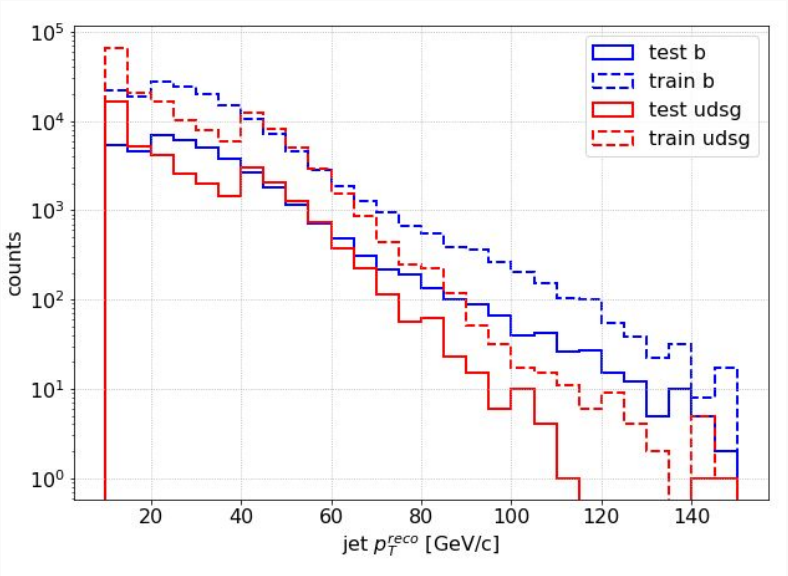
1. Progress in analysis
    - ML model
    - dataset
    - experiments
2. Questions & issues
3. Plans for next week

Model & data - basic info

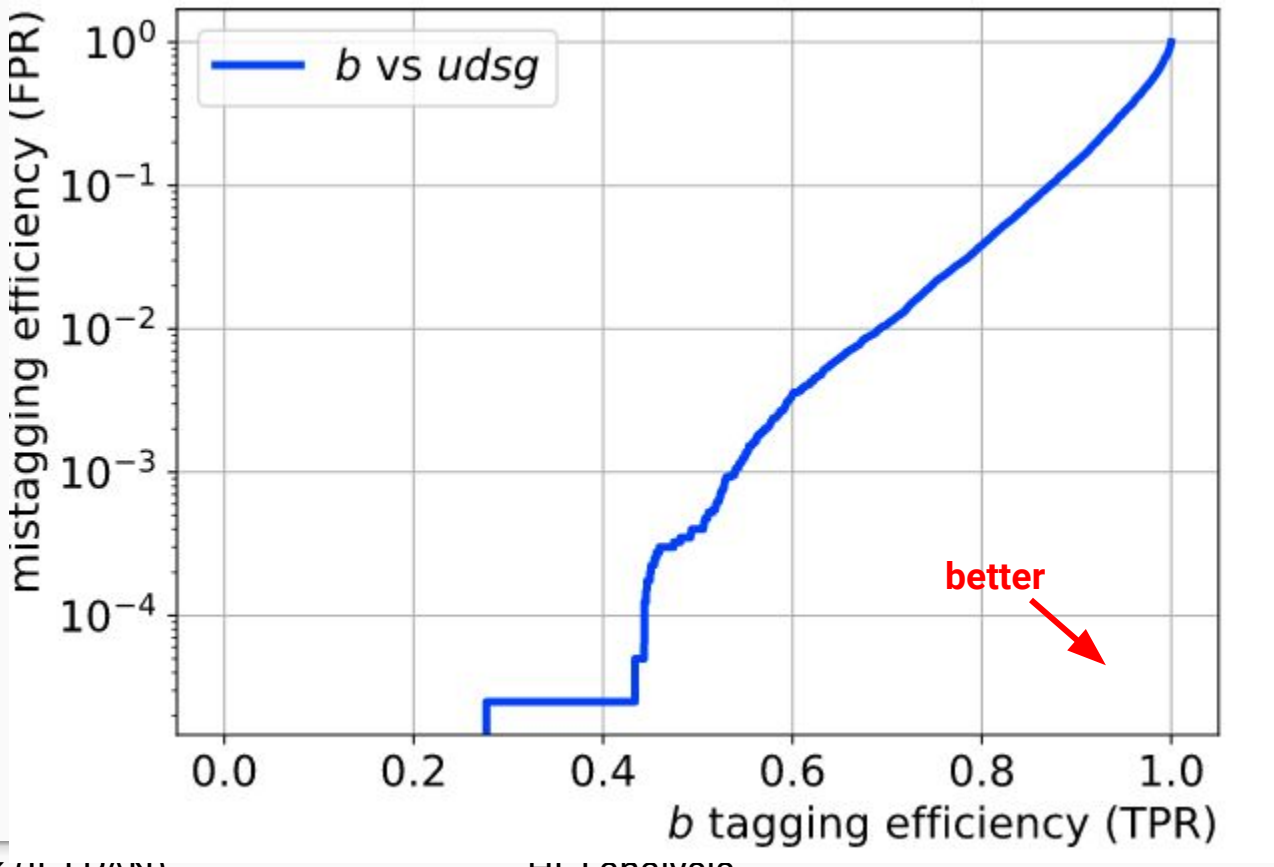
1. Default dataset:
   - sample: 2 x 500k (or 2 x 200k for experiments) jets
   - up to 15 tracks & 15 SV
   - **jet-level**: Pt, Area, Eta, Phi, NumTracks, NumSV
   - **SV**: Lxy, SigmaLxy, Chi2, Dispersion, Mass
   - **tracks**: IPd, IPz, CovIPd, CovIPz, Pt, Eta, Phi
   - ... if not stated otherwise
2. ML model (BDT) trained on MC data
   - binary classification: b-jets vs udsg-jets
   - test set: 20%, pT-statified

   b- & usdg-jets have different spectra

   but no difference between train and test set

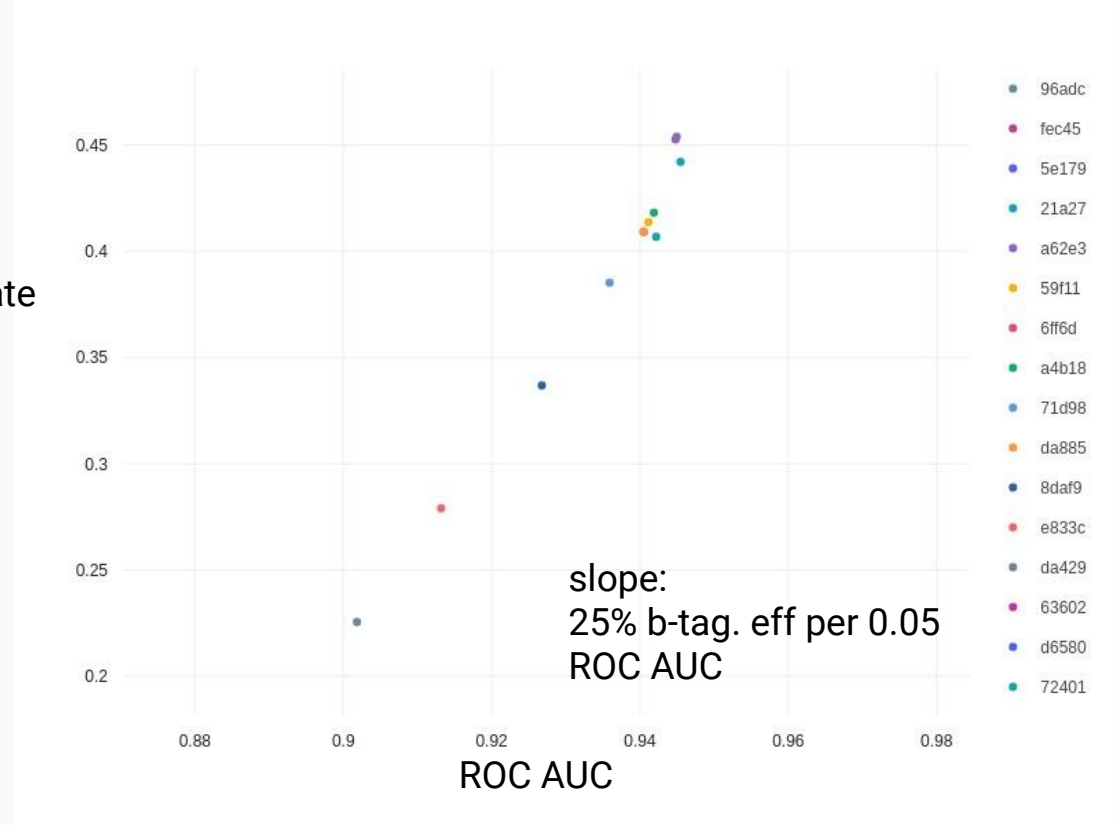


Metrics


b vs udsg
mistagging efficiency (FPR)
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0.0
0.2
0.4
0.6
0.8
1.0
b tagging efficiency (TPR)
better

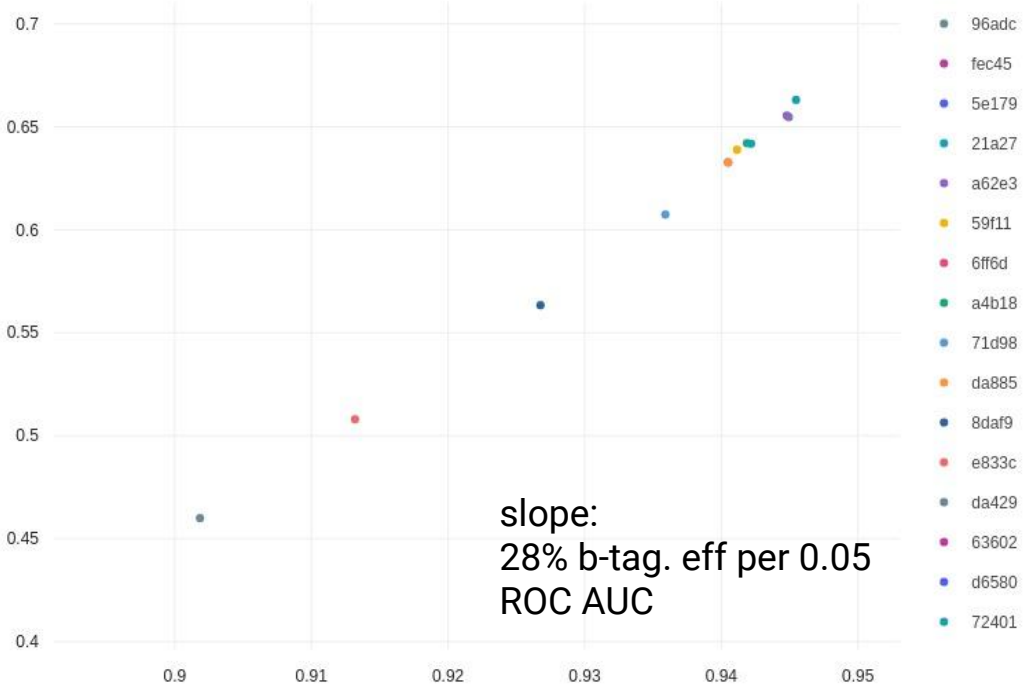
Metrics

b tag. eff.
for **0.001**
mistag. rate



Metrics

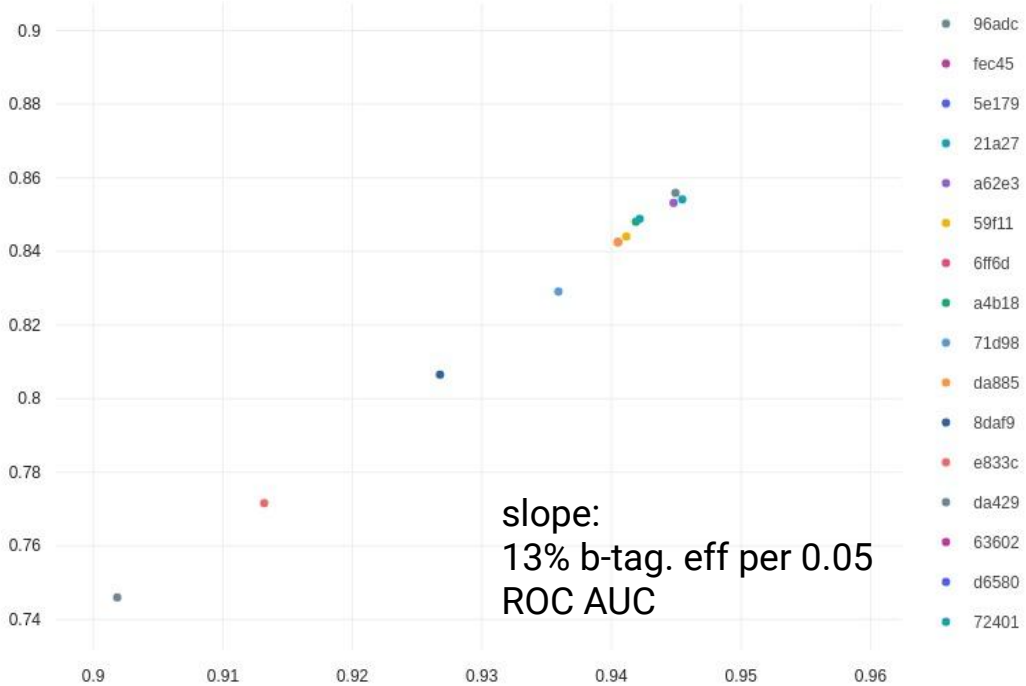
b tag. eff.
for **0.01**
mistag. rate



slope:
28% b-tag. eff per 0.05 ROC AUC

Metrics

b tag. eff.
for **0.1**
mistag. rate



slope:
13% b-tag. eff per 0.05 ROC AUC

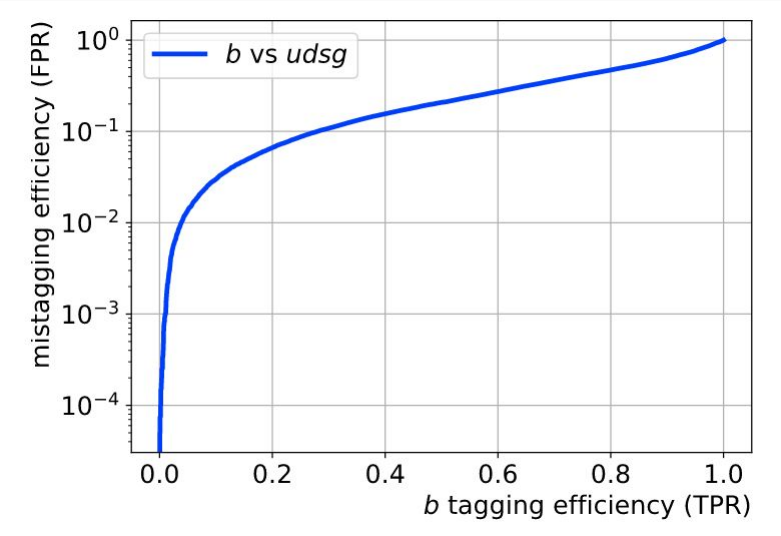
Metrics

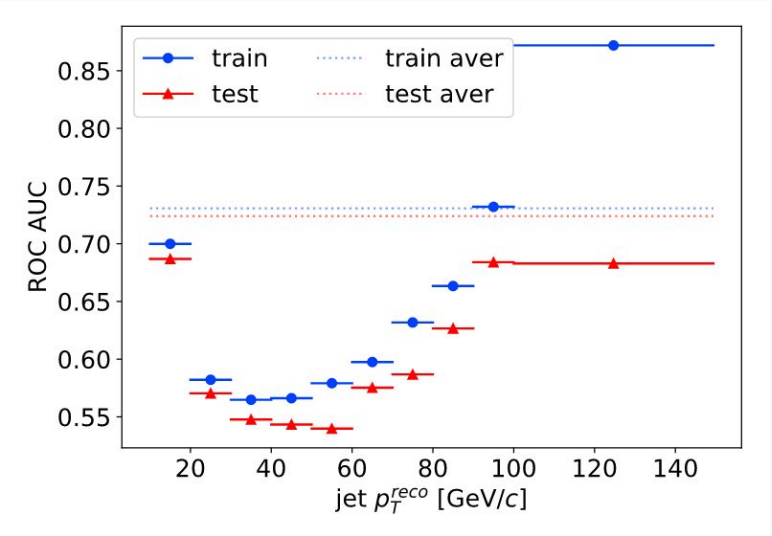
I will report ROC AUC,
0.01 difference in ROC AUC corresponds to

- 5% eff. @ mistagging rate = 0.001
- 5.6% eff. @ mistagging rate = 0.01
- 2.6% eff. @ mistagging rate = 0.1

Experiments

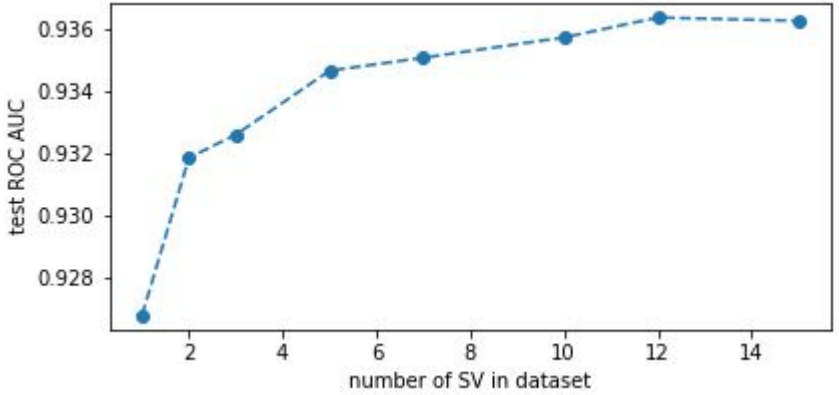
- Model trained only on jet-level observables (only pT, phi, eta, area)
  - terrible performance
  - better for very low and very high pT due to different spectra shape
  - just for comparison, not to be used

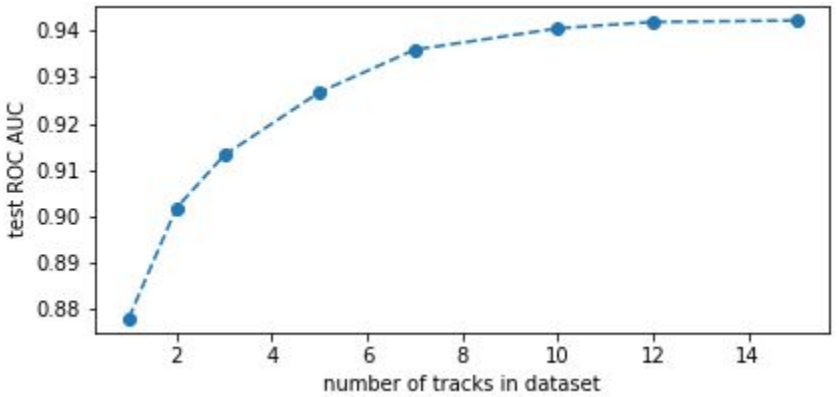




Experiments

- Vary (lower) number of SV and tracks models trained only on tracks/SV + jet level parameters
  - adding more SV does not improve much
  - adding more tracks does, but ~10 should be enough





Experiments

- feature engineering on track's observables
  almost no diff. when trained with IP (up to 3% b-tag. eff), but without them it's quite large
- more complex models may benefit more from these transf. (simpler models: IP dominates)

features	roc_auc_train	roc_auc_test	bEff@mistag_1e-03	bEff@mistag_1e-02	bEff@mistag_1e-01
default	0.9434980755837...	0.94049911871...	0.40926023150578766	0.6327158178954474	0.8425210630265757
Track_Pt -> Track_PtFrac	0.9442104357400...	0.94115147652...	0.4138103452586315	0.6388659716492913	0.8440461011525288
phi,eta -> deltaPhi, deltaEta, deltaR	0.9478629633964...	0.94479494434...	0.45276131903297584	0.6554663866596665	0.8531963299082477
pt -> ptFrac & phi,eta -> deltaPhi, deltaEta, delt...	0.9486315104862...	0.94546739121...	0.44221105527638194	0.6630665766644166	0.8541213530338259
no IP	0.7601006840335...	0.74974257296...	0.016450411260281506	0.0703017575439386	0.3407085177129428
no IP & pt -> ptFrac & phi,eta -> deltaPhi, delta...	0.84672232995495	0.83923138271...	0.05912647816195405	0.21035525888147202	0.5742643566089152
add Nsigma of IPd/IPz/IP3d	0.9478492635526...	0.94494417809...	0.45406135153378835	0.6546913672841821	0.8558963974099353

Plans for next week

1. Include c-jets
2. Compare with other's (Rudiger, Hadi Hassan, my MSc thesis)
3. Apply on data

4. Design selection criteria for SV (just sort by chi2 ?), at least vizualize sth

- add train curve on TPR-vs-FPR plot
- plot metrics as a func. of train iterations (train & test)

BACKUP

lower_edges=(5 7 9 12 16 21 28 36 45 57 | 70 85 99 115 132 150 169 190 212 235)
higher_edges=(7 9 12 16 21 28 36 45 57 70 | 85 99 115 132 150 169 190 212 235 -1)

momentum dispersion: $$p_T D = \frac{\sqrt{\sum_{i \in jet} p_{T,i}^2}}{\sum_{i \in jet} p_{T,i}}.$$

angularity: