Heavy-flavour jet tagging in ATLAS

... and how we use Machine Learning

Philipp Windischhofer
University of Oxford
philipp.windischhofer@cern.ch

on behalf of the ATLAS Collaboration

ML4Jets, January 15-17, 2020, NYU
Why flavour tagging?

quarks:
“elementary”

Quantum numbers:
symmetries

momentum \( p^\mu \)

charge, flavour, … \( \sigma \)
Why flavour tagging?

quarks: “elementary”

jets: observables

TeV \rightarrow QCD \rightarrow GeV

Quantum numbers: symmetries

momentum $p^\mu$

charge, flavour, … $\sigma$
Why flavour tagging?

Quarks: “elementary”

\[ \approx f(\text{observables}) \]

Jets: observables

TeV \quad QCD \quad GeV

Reconstruction

Quantum numbers:
- Symmetries

Momentum: \( p^\mu \)

Charge, flavour, …: \( \sigma \)
Why flavour tagging?

quarks: “elementary”

jets: observables

TeV QCD GeV

reconstruction

Quantum numbers: symmetries

momentum $p^\mu$ $f = \text{jet clustering algorithm}$

charge, flavour, … $\sigma$ $f = \text{jet tagger}$
Why flavour tagging?

quarks: “elementary”

jets: observables

TeV - QCD - GeV

\[ f = \text{jet clustering algorithm} \]

\[ f = \text{jet flavour tagger} \]

Quantum numbers: symmetries

momentum \( p^\mu \)

charge, flavour, … \( \sigma \)
By the end of this talk, you will know ...

- ... how flavour tagging helps ATLAS
- ... how the ATLAS flavour tagging tools work
- ... how machine learning helped to improve them
- ... what their performance is on data and simulation
The market for b-tagging

ATLAS publications that...

- mention any b-tagger
- mention MV1
- mention MV2
- mention DL1
- mention DL1r

MV1: Run 1 workhorse
MV2: Run 2 workhorse
DL1, DL1r: the future!

https://gitlab.cern.ch/phwindis/arxivscraper
The market for b-tagging

MV2

Stop search (arXiv:1606.03903v1)

ATLAS

- Observed limit (\( \pm \sigma_{\text{exp}} \))
- Expected limit (\( \pm \sigma_{\text{th}} \))
- Limit at 95% CL

\( m_{t^+} \) production, \( g \rightarrow t^+t, m_{t^+} \leq m_t \), \( \Delta m(t^+, t^-) = 5 \text{ GeV} \)

Limit at 95% CL

ATLAS

\( s = 13 \text{ TeV}, 3.2 \text{ fb}^{-1} \)

- ATLAS incl. 1L 8 TeV, 20.3 fb^{-1}
- ATLAS mono-jet 13 TeV, 3.2 fb^{-1}

\( m_g \) (GeV)

\( m_{t^+} \) (GeV)


DL1, DL1r:

the future!

MV1:
Run 1 workhorse

MV2:
Run 2 workhorse

https://gitlab.cern.ch/phwindis/arxivscraper
The market for b-tagging

**MV2**

$H \rightarrow b\bar{b}$ observation
(arXiv:1808.08238)

*ATLAS*

- Data
- $VH, H \rightarrow b\bar{b}$ ($\mu = 1.06$)
- Diboson
- Uncertainty

**MV1:**
Run 1 workhorse

**MV2:**
Run 2 workhorse

**DL1, DL1r:**
the future!

https://gitlab.cern.ch/phwindis/arxivscraper
The market for b-tagging

ATLAS publications that ...
- mention any b-tagger
- mention MV1
- mention MV2
- mention MV2
- mention DL1
- mention DL1r

MV1:
Run 1 workhorse

MV2:
Run 2 workhorse

DL1, DL1r:
the future!

https://gitlab.cern.ch/phwindis/arxivscraper
The market for b-tagging

**DL1r** di-b-jet resonance search
(arXiv:1910.08447)

**ATLAS**
\( \sqrt{s} = 13 \text{ TeV}, 139 \text{ fb}^{-1} \)
2 b-tag

- Data
- Background fit
- BumpHunter interval
- DM \( Z' \), \( m_{Z'} = 2 \text{ TeV} \)
- DM \( Z' \), \( m_{Z'} = 3 \text{ TeV} \)
- DM \( Z' \) \( g_q = 0.25, \sigma \times 10 \)
- \( p \)-value = 0.83

**MV1**: Run 1 workhorse
**MV2**: Run 2 workhorse
**DL1, DL1r**: the future!

**MV2, DL1, DL1r**

https://gitlab.cern.ch/phwindis/arxivscraper
The ATLAS strategy for b-tagging

\[ \approx f \left( \begin{array} \text{TeV} \\ \text{QCD} \\ \text{GeV} \\ \begin{array} \text{Quark} \\ \text{Low-dimensional} \\ \text{properties} \\ \begin{array} \text{Jet} \\ \text{High-dimensional} \\ \text{features} \\ \end{array} \\ \end{array} \right) \]

momentum \( p^\mu \)

charge, flavour, … \( \sigma \)
The ATLAS strategy for b-tagging

Low-level taggers
- Physics-based feature extraction
- Detector-specific

High-level taggers
- Likelihood ratio estimation
- Detector agnostic
- Fast turnaround

$D \sim \log \frac{p(d_i | b \text{- jet})}{p(d_i | \neg b \text{- jet})}$

$D_{MV2, DL1, DL1r}$
Low-level taggers

$b$-jet formation:

primary $pp$ vertex

$\tau_{\text{had.}} \sim \frac{1}{1 \text{ GeV}} \sim 10^{-24} s$
Low-level taggers

\( b \)-jet formation:

primary \( pp \) vertex

\[ \tau_{\text{had.}} \approx \frac{1}{1 \text{ GeV}} \approx 10^{-24} \text{s} \]

\( B \)-hadron: large quark momentum fraction

\[ \langle \frac{p_{T,B\text{-hadron}}}{p_{T,b\text{-quark}}} \rangle \approx 1 - \frac{1 \text{ GeV}}{m_b} \approx \frac{3}{4} \]

ATLAS Simulation Preliminary

POWHEG \( pp \rightarrow t \bar{t} \)

Anti-\( k_t \) \( R=0.4 \) \( b \)-jets

\( \mathbf{B} \)-hadron: large quark momentum fraction
Low-level taggers

*b-jet formation:*

\[ \langle p_T, B-\text{hadron} \rangle \approx 1 - \frac{1 \text{ GeV}}{m_b} \approx \frac{3}{4} \]

\[ z = \frac{p_{\text{jet}} \cdot p_{c'}}{p_{\text{jet}}^2} \]
Low-level taggers

*b*-jet formation:

**B-hadron: high decay multiplicity**

**B-hadron: large quark momentum fraction**

\[ \langle p_T, b \text{-hadron} \rangle \approx 1 - \frac{1 \text{ GeV}}{m_b} \approx \frac{3}{4} \]

\[ z = \frac{p_{\text{jet}} \cdot p_c}{|p_{\text{jet}}|^2} \]
$b$-jet formation:

**Primary vertex**

**$B$-hadron: high decay multiplicity**

**$B$-hadron: large quark momentum fraction**

**$\langle p_T, B^{-\text{hadron}} \rangle \approx 1 - \frac{1 \text{ GeV}}{m_b} \approx \frac{3}{4}$**

**$z = p_{\text{jet}} \cdot p_{c} / l^2_{\text{jet}}$**
Low-level taggers

$b$-jet formation:

primary $pp$ vertex
Low-level taggers

\[ b\text{-Hadron decay: long lifetime for } b \to c \]

\[ c \tau_B \sim \frac{1}{|V_{bc}|^2} \]

**b-jet formation:**

- Primary vertex
- \( pp \) vertex

**Compute properties of Secondary Vertex:**

- SV mass: \( m(SV) \sim m_B \)
- SV decay length significance: \( S_{xyz} \sim c \tau_B \)
- ...
**b-jet formation:**

**b-Hadron decay:** long lifetime for $b \rightarrow c$

$$c \tau_B \sim \frac{1}{|V_{bc}|^2}$$

*primary pp vertex*

**Compute properties of Secondary Vertex:**

- SV mass $m(SV) \sim m_B$
- SV decay length significance $S_{xyz} \sim c \tau_B$
- ...
Low-level taggers

$b$-jet formation:

b-Hadron decay: long lifetime for

\[ b \rightarrow c \]

primary vertex

impact parameter
Low-level taggers

**b-jet formation:**

![Diagram showing b-jet formation with primary pp vertex and impact parameter]

**b-Hadron decay:** long lifetime for $b \rightarrow c$

**b-jets: tracks w/ large impact parameter**

![Graph showing impact parameter with b-jets, c-jets, and light-flavour jets]

- **ATLAS Simulation Preliminary**
  - $\sqrt{s}=13$ TeV, tt
  - b jets
  - c jets
  - Light-flavour jets

**IP3D**

**RNNIP**
Low-level taggers

$b$-jet formation:

primary vertex

$pp$ vertex
Low-level taggers

*b*-jet formation:

Exploit structure of full $b \rightarrow c$ decay chain

ATLAS Simulation Preliminary

$\sqrt{s}=13$ TeV, $t\bar{t}$

JetFitter
The ATLAS strategy for b-tagging

- **MV2** vs. **DL1**: different architecture, same inputs
- **DL1r**: also add RNNIP
The ATLAS strategy for b-tagging

Focus on taggers with strong ML component:

- **MV2** vs. **DL1**: different architecture, same inputs
- **DL1r**: also add RNNIP
RNNIP

Unrolled LSTM

Tracks ordered by transverse IP sig.

Jet

ΔR association

IP3D

Naive Bayes model:

$$D_{IP3D} = \sum_i \log \frac{p_b^i}{p_u^i}$$

Neglect correlations between tracks

• LSTM vs. Naive Bayes: **correlations are important**!
  • ... but hard to model and exploit
Tracks contain a lot of information:

- Jet flavour discrimination:
  - Impact parameter, track momentum, ...

- Track quality:
  - Number of (shared) pixel hits

(more information in backup)

**Provides higher light (and charm-) rejection compared to other low-level taggers!**

RNNIP now part of officially recommended taggers:

- Ensure that data/simulation scale factors can be reliably measured
  - Transform tagger to have same light-rejection, worse b-efficiency
  - Calibrate transformed tagger, then extrapolate to original tagger
• Fully connected (deep) neural network
• Estimate likelihood ratio of low-level tagger outputs

\[ D_{\text{DL1}}^{b-\text{tag}} \sim \log \frac{p_b}{p_c + p_u} \]

• Supports \( b \)- and \( c \)-tagging

Separate trainings for both supported jet collections:
  • Particle-flow jets: new ATLAS baseline
  • Variable-radius track jets: invaluable for boosted decays
Training dataset: “hybrid” sample

- simulated $t\bar{t}$ for $p_T < 250$ GeV
- $Z' \rightarrow q\bar{q}$ for $p_T > 250$ GeV

Taggers well-behaved even for multi-TeV jets!

\[
\frac{dN_{\text{jet}}}{dp_T} \sim \frac{1}{p_T^5} \quad \text{(for } t\bar{t})
\]
\[
\frac{dN_{\text{jet}}}{dp_T} \sim \text{const.} \quad \text{(for } Z')
\]

- Models implemented in Keras + Tensorflow
- More efficient training and optimisation pipeline, heavily containerised
Tagger performance in simulation

- **MV2 → DL1**: very similar performance
- **DL1 → DL1r**: adding RNNIP (+ optimising network architecture) significantly improves light-and charm rejection
Tagger performance on data

- Good modelling is **essential** for training
- Scale factor determination very complex
  - Measure efficiency in data for b, c, light jets
  - b-SF: precision top measurement!!

\[
\text{SF}_b = \frac{\varepsilon_{b\text{-jets}}^{\text{data}}}{\varepsilon_{b\text{-jets}}^{\text{sim.}}}
\]

(more plots in backup)
**Summary**

**Two-stage approach:**
- Robust, physics-driven low-level taggers
- Detector-agnostic, ML-based high-level taggers

- New LSTM-based low-level tagger
- Improved training pipeline for high-level taggers

**Significant performance gain compared to Run-2 baseline tagger**
Backup
References

(1) Comparison of Monte Carlo generator predictions for bottom and charm hadrons in the decays of top quarks and the fragmentation of high pT jets, ATL-PHYS-PUB-2014-008

(2) Expected performance of the ATLAS b-tagging algorithms in Run-2, ATL-PHYS-PUB-2015-022


(5) ATLAS flavour-tagging calibration results with 139 ifb, ATL-FTAG-2019-004

(6) Hyper-parameter scan with the Deep Learning heavy-flavour tagger (DL1), ATL-FTAG-2019-001

(7) Machine learning algorithms for b-jet tagging at the ATLAS experiment, ATL-PHYS-PROC-2017-211
Improvement from high-level taggers

Combination of low-level taggers leads to huge performance gain!
Tagger performance in simulation

- **MV2 → DL1**: very similar performance
- **DL1 → DL1r**: adding RNNIP (+ optimising network architecture) significantly improves light-and charm rejection

![Graph showing Tagger performance in simulation](image)

**ATLAS** Simulation

\( \sqrt{s} = 13 \text{ TeV}, \ell \bar{\ell} \)

- Fraction of jets / 0.50
- \( D_{DL1} \)
- b-jets
- c-jets
- Light-flavour jets
IP sign convention

\[ \alpha < \frac{\pi}{2} \implies d_0 = |d_0| \]
\[ z_0 = |z_0| \]

\[ \alpha > \frac{\pi}{2} \implies d_0 = -|d_0| \]
\[ z_0 = -|z_0| \]
Tagger performance in data

- b-SF: top-precision measurement!!
- Uncertainties O(1%)
- Simulation-based extrapolation of uncertainties to high-pT

\[
SF_b = \frac{\varepsilon_{b-jets}^{data}}{\varepsilon_{b-jets}^{sim.}}
\]

**ATLAS** Preliminary
\( \sqrt{s} = 13 \) TeV, 139 fb\(^{-1}\)

- MV2 \( \tau_b = 70 \% \) Single Cut OP
- anti-k, R=0.4 EMPFlow Jets

**b-SF:**
- Measured Scale Factor (total unc.)
- Smoothed Scale Factor (total unc.)

**ATLAS** Preliminary
\( \sqrt{s} = 13 \) TeV, 139 fb\(^{-1}\)

- DL1 \( \tau_b = 70 \% \) Single Cut OP
- anti-k, R=0.4 EMPFlow Jets

**b-SF:**
- Measured Scale Factor (total unc.)
- Smoothed Scale Factor (total unc.)
Tagger performance in data

- light-SF measured in Z + jets
- Calibrate “flipped” tagger, then extrapolate to nominal tagger

\[
SF_b = \frac{\epsilon_{b\text{-jets}}^{\text{data}}}{\epsilon_{b\text{-jets}}^{\text{sim.}}}
\]
Tagger performance in data

- charm-SF measured in semileptonic ttbar

\[ SF_b = \frac{\epsilon_{\text{data}}}{\epsilon_{\text{sim.}}} \]

\(\epsilon_{\text{data}}\) = \(\epsilon_b - \text{jets}\)

\(\epsilon_{\text{sim.}}\) = \(\epsilon_b - \text{jets}\)

**ATLAS** Preliminary

\(\sqrt{s} = 13\) TeV, 139 fb\(^{-1}\)

c-jet Calibration with \(t\bar{t}\) Events

MV2 \(\epsilon_b = 70\%\) Single Cut OP

anti-\(k_t\), R=0.4 EMPFlow Jets

- Measured Scale Factor (total unc.)
- Smoothed Scale Factor (total unc.)

**ATLAS** Preliminary

\(\sqrt{s} = 13\) TeV, 139 fb\(^{-1}\)

c-jet Calibration with \(t\bar{t}\) Events

DL1 \(\epsilon_b = 70\%\) Single Cut OP

anti-\(k_t\), R=0.4 EMPFlow Jets

- Measured Scale Factor (total unc.)
- Smoothed Scale Factor (total unc.)

Tagger performance in data
**ATLAS Simulation Preliminary**

- **Subleading S**
  - $\sqrt{s}=13$ TeV, $t\bar{t}$
  - b-jets, $p_T>20$ GeV, $|\eta|<2.5$

- **Leading S**
  - $\sqrt{s}=13$ TeV, $t\bar{t}$
  - light-jets, $p_T>20$ GeV, $|\eta|<2.5$

**ATLAS Simulation Preliminary**

- **light-jet rejection, $1/\epsilon_l$**
  - $\sqrt{s}=13$ TeV, $t\bar{t}$
  - $p_T>20$ GeV, $|\eta|<2.5$

- **b-jet efficiency, $\epsilon_b$**
  - $\sqrt{s}=13$ TeV, $t\bar{t}$

- **c-jet rejection, $1/\epsilon_c$**
  - $\sqrt{s}=13$ TeV, $t\bar{t}$
  - $p_T>20$ GeV, $|\eta|<2.5$
RNNIPFlip

- RNNIP leads to significant enhancement of light-jet rejection

**How to measure light-jet rejection in data?**
- Flavour composition after tagging dominated by heavy-flavour jets
- Cannot establish pure enough sample of light jets to perform calibration

**Instead:**
- Define another tagger “RNNIPFlip”, with the same light rejection, but much worse b-efficiency
- Easy to calibrate in data, then extrapolate to actual RNNIP
**RNNIPFlip**: flip sign of transverse and longitudinal impact parameters, then evaluate RNNIP network.

I.e. sequence order is reversed as well.

ATLAS Simulation Preliminary

- Symmetric around zero for tracks in light-jets
- Highly asymmetric for tracks in heavy-flavour jets
RNNIP network architecture

- LSTM with 50 hidden units
- Dense layer with 10 hidden units before softmax

Tracks ordered by transverse IP sig.

Jet $\Delta R$ association

Fully connected

$pb$ $pc$ $pu$ softmax output

- LSTM with 50 hidden units
- Dense layer with 10 hidden units before softmax
Old RNNIP architecture

Track quality embedding

2D unit vector

Unrolled RNN

Ordered by magnitude of transverse IP significance

Fully Connected + SoftMax

Jet

ordered by $|S_{do}|$

Track 1, Track 2, Track 3, Track 4, ..., Track N

Merge

Embed

category

S_{do}, S_{z0}, p_T^{frac}, $\Delta R$
RNNIP training details

Training technicalities:

- Train on same “hybrid” sample as DL1, DL1r
- Sort delta-R associated tracks by transverse IP sig.
- Use first 15 tracks for training, zero-pad if shorter

Track features:

- 4 continuous features per track:
  - Transverse & longitudinal IP sig.
  - $\Delta R(\text{track}, \text{jet})$
  - $p_T, \text{track}/p_T, \text{jet}$
- Additional track quality features:
  - Number of (shared) hits in (innermost pixel layer | pixel | Si-strip tracker)
What does the network “learn”?

- The track multiplicity of the B-hadron decay, large impact parameters for these tracks
- Tracks with large IPs tend to be *harder* and *wider* for b-jets