Details of b-tagging Algorithm

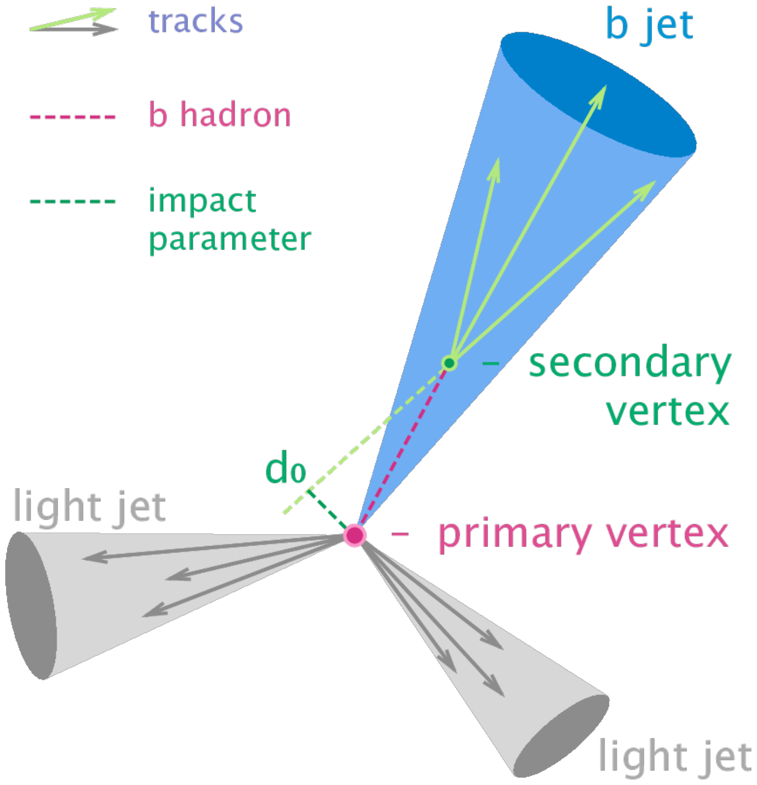
Dec 16, 2019

Description of Task Background ([AFT-455]())

- B-tagging: the identification of jets containing b-hadrons
  - Several dedicated algorithms exploiting specific proprieties like long lifetime, high mass and decay multiplicity of b-hadrons and the hard b-quark fragmentation.

  - Low Level Discriminants:
    - IP3D & IP2D: Impact Parameter (d0, z0) based
    - SV1: Secondary Vertex based
    - JetFitter: Secondary Vertex based

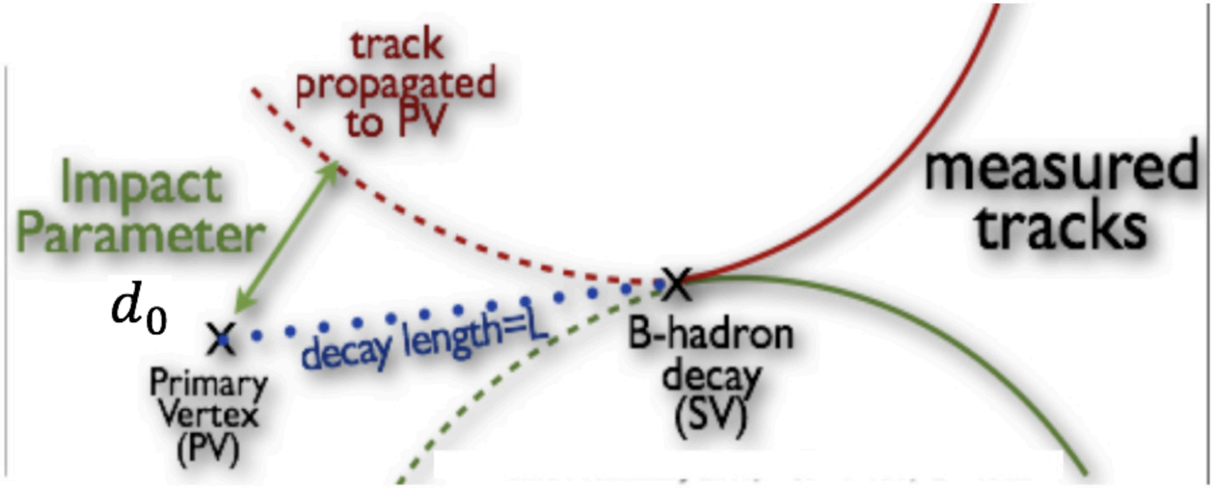
  - As input of High Level Discriminants:
    - MV2 and DL1



- Algorithms are trained on *pp* simulations.
- B-tagging algorithms can now run heavy ion data from AFT-233.

Track-based b-tagging Algorithms (IP2D & IP3D)

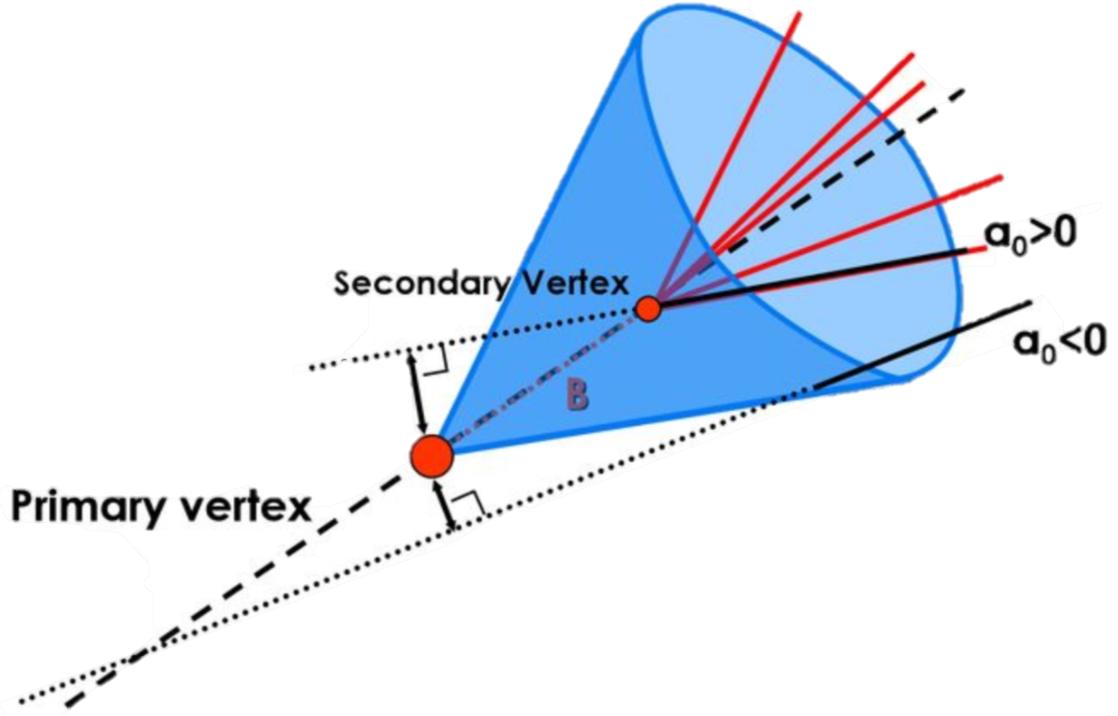
- Impact parameter-based algorithms.
  - Transverse impact parameter d_0, distance in r-ϕ plane between point of nearest approach to the primary vertex.
  - Longitudinal impact parameter z_0, separation in z-direction between the point that gives d_0 and primary vertex.
  - $z_0 sin\theta$, projection of z_0 onto the plane perpendicular to track direction.
  - IP2D only uses the d_0 information where IP3D uses both d_0 and $z_0 sin\theta$ information.



- Motivation: b-quark containing hadrons typically have a long lifetime, and thus the track vertex is displaced from primary vertex, giving a larger impact parameter.

- http://cdsweb.cern.ch/record/2273281/files/ATL-PHYS-PUB-2017-013.pdf

Using the Impact Parameters

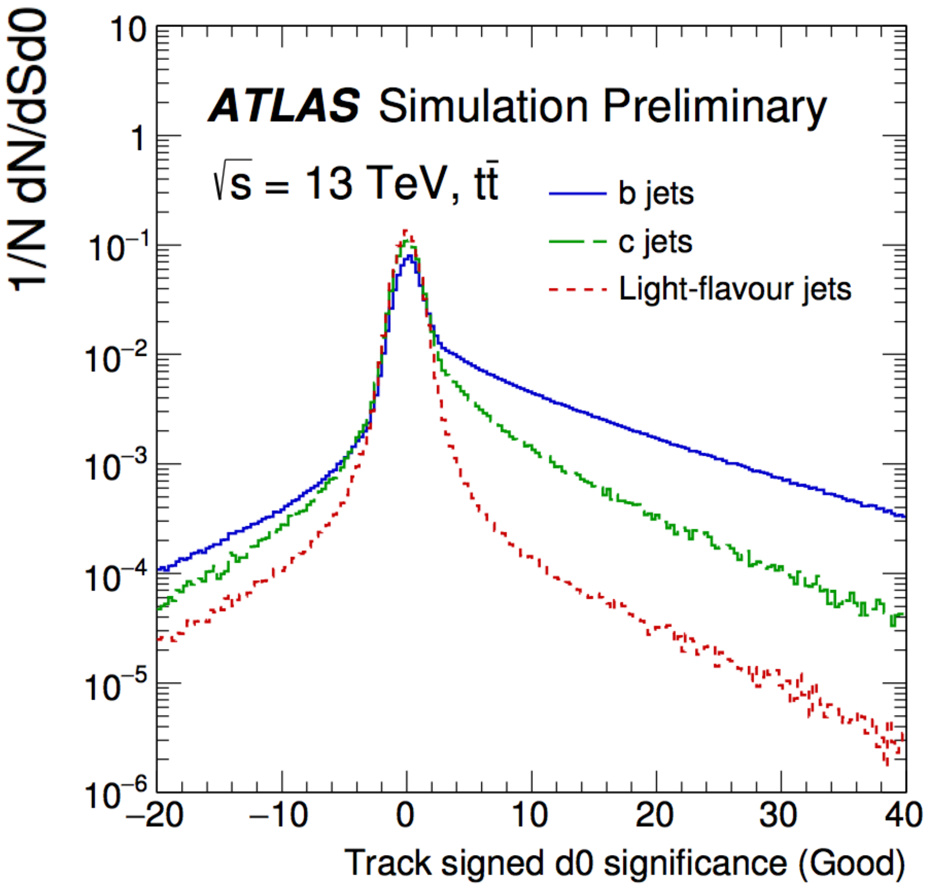


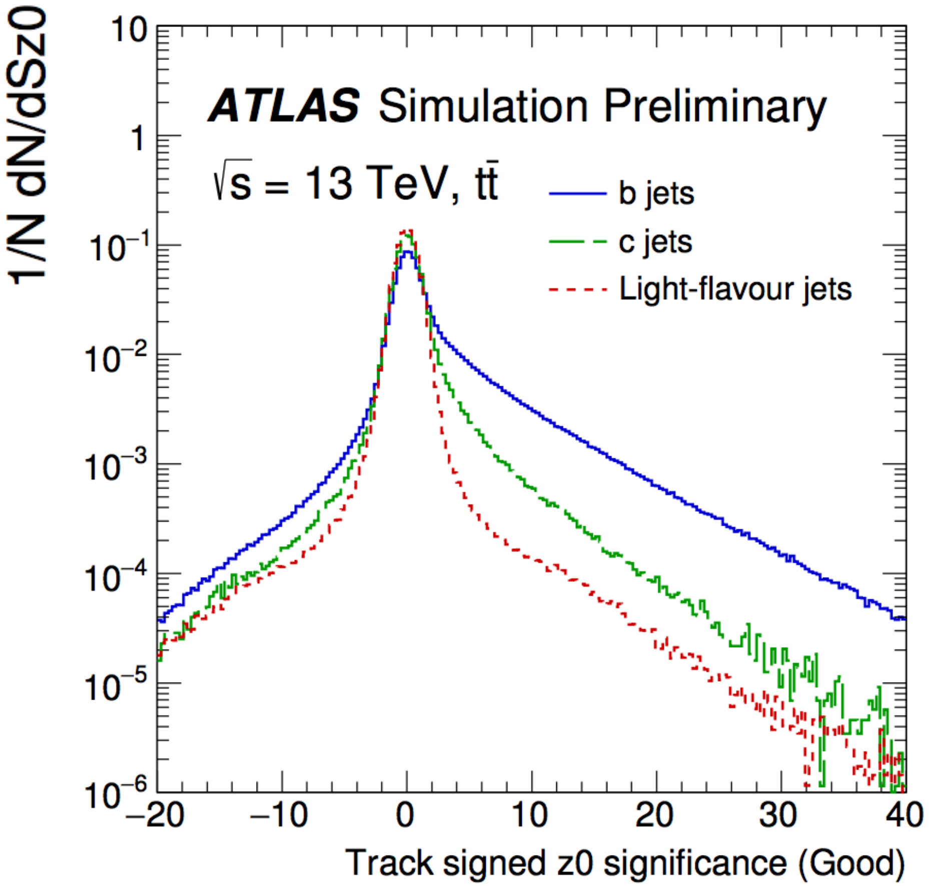
- The impact parameter is given a positive sign if the position of crossing of the track with the jet axis is located upstream to the PV along the jet axis which defines the direction, and negative otherwise.

- Transverse impact parameter significance: $S_{d_0} = d_0/\sigma_{d_0}$
- Longitudinal impact parameter significance: $S_{z_0} = z_0/\sigma_{z_0}$

- Tracks generated from b-hadron decays have impact parameter significances which differ significantly from zero, whereas those from tracks in light jets are typically consistent with zero

Impact Parameter Significance Distribution of Different Jets





- Shown above is the Monte Carlo of $Z' \to t\bar{t}$ events.
- The category tag (good) is explained in next slide.
- http://cds.cern.ch/record/2160731/files/ATL-PHYS-PUB-2016-012.pdf?version=2

Categories of Tracks (track grade)

- Exclusive categories (the track grade) that depend on the hit pattern of a given track to increase the discriminating power.

#	Category	Fractional contribution [%]		
		b-jets	*c*-jets	light-jets
0	No hits in first two layers; expected hit in IBL and b-layer	1.9	2.0	1.9
1	No hits in first two layers; expected hit in IBL and no expected hit in b-layer	0.1	0.1	0.1
2	No hits in first two layers; no expected hit in IBL and expected hit in b-layer	0.04	0.04	0.04
3	No hits in first two layers; no expected hit in IBL and b-layer	0.03	0.03	0.03
4	No hit in IBL; expected hit in IBL	2.4	2.3	2.1
5	No hit in IBL; no expected hit in IBL	1.0	1.0	0.9
6	No hit in b-layer; expected hit in b-layer	0.5	0.5	0.5
7	No hit in b-layer; no expected hit in b-layer	2.4	2.4	2.2
8	*Shared* hit in both IBL and b-layer	0.01	0.01	0.03
9	At least one *shared* pixel hits	2.0	1.7	1.5
10	Two or more *shared* SCT hits	3.2	3.0	2.7
11	*Split* hits in both IBL and b-layer	1.0	0.87	0.6
12	*Split* pixel hit	1.8	1.4	0.9
13	*Good*	83.6	84.8	86.4

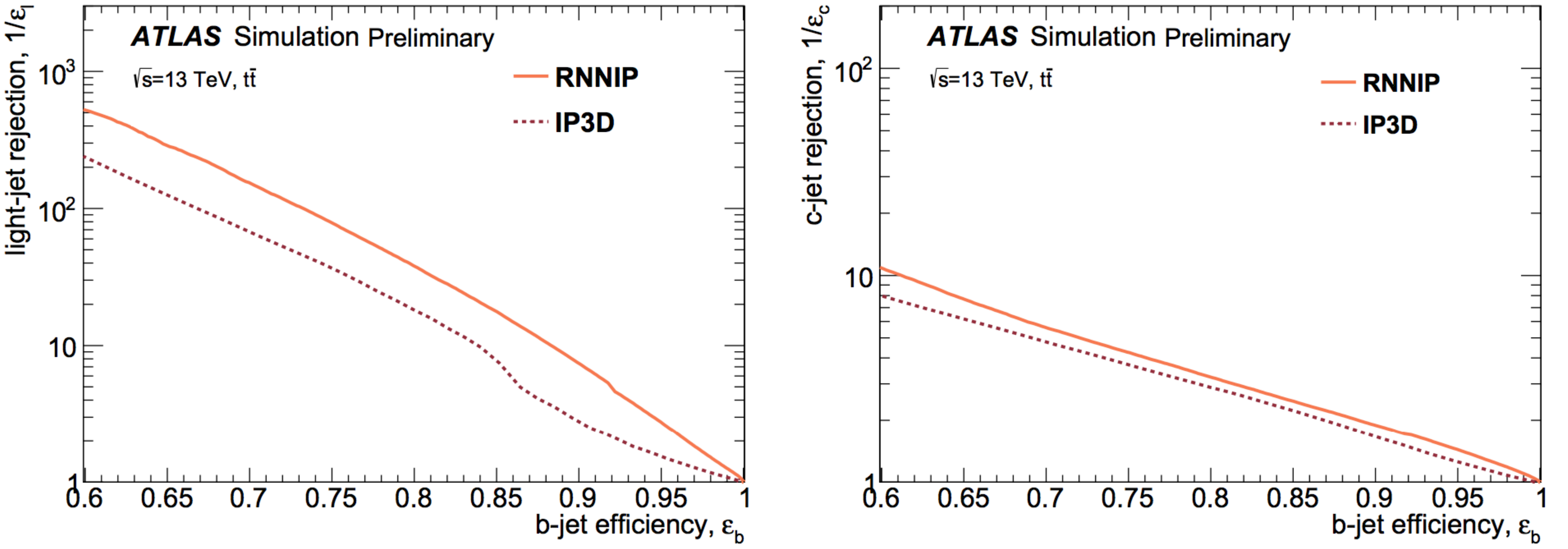
LLR (Log-Likelihood Ratio): how tagging is done

- Goal: determine whether the given jet is more likely to be one kind of jet or the other. For example, whether a jet is more likely to be b-jet or a light-jet.
- $\sum_{i=1}^{N} \log\left(\frac{p_b}{p_u}\right)$ the LLR is calculated for each jet, by summing over individual tracks within the jet, assuming no correlation between jets.

- There're 14 categories of tracks, values of S_{d_0} is binned into 35 bins, values of S_{z_0} is binned into 20 bins, and there're three flavors of jets, giving a bin count of 29400. Each bin is assigned a probability based on the MC samples.
- p_b is the probability for b-jet of a track with a given S_{d_0}, S_{z_0}, and category. Similarly, p_u is the probability for light-jet of a track with a given S_{d_0}, S_{z_0}, and category.

- IP3D assumes no correlation between tracks from a jet for tagging, to account for this effect, a neural network algorithm which utilizes the same information is also used.
- RNNIP: http://cds.cern.ch/record/2255226/files/ATL-PHYS-PUB-2017-003.pdf?version=1

Comparison of RNNIP and IP3D



(a)

(b)

Figure 3: Light-flavour jet (a) and *c*-jet (b) rejection as a function of the *b*-tagging efficiency comparing the IP3D and RNNIP *b*-tagging algorithms for selected jets in simulated $t\bar{t}$ events.

Secondary Vertex Based Algorithm

- Secondary Vertex Reconstruction in SVF
  - Basic procedure:
    - Apply cuts on tracks within a jet before reconstructing, excluding tracks too close to primary vertex and of bad quality
    - Find all possible two-track secondary vertices from these tracks
    - Clean the secondary vertices, excluding vertices from irrelevant processes
    - Merge into one secondary vertex iteratively
  - Output used by SVx (secondary vertex-based tagger), available in MC and data.
    - The mass of reconstructed Secondary vertex. (branch SVx_masssvx)
    - The ratio of energy of tracks in the SV to the energy of tracks in the jet. (branch SVx_efracsvx)
    - The number of good two-track vertices. (branch SVx_N2Tpair)
    - Del_R between the jet direction and the SV-PV line (Primary to Secondary vertex distance).
    - Significance (dist/error) for 3D distance between secondary and primary vertex. (branch sig3d)
  - SV2 location in xyz is also available. (
- B-tagging tool using Secondary vertex:

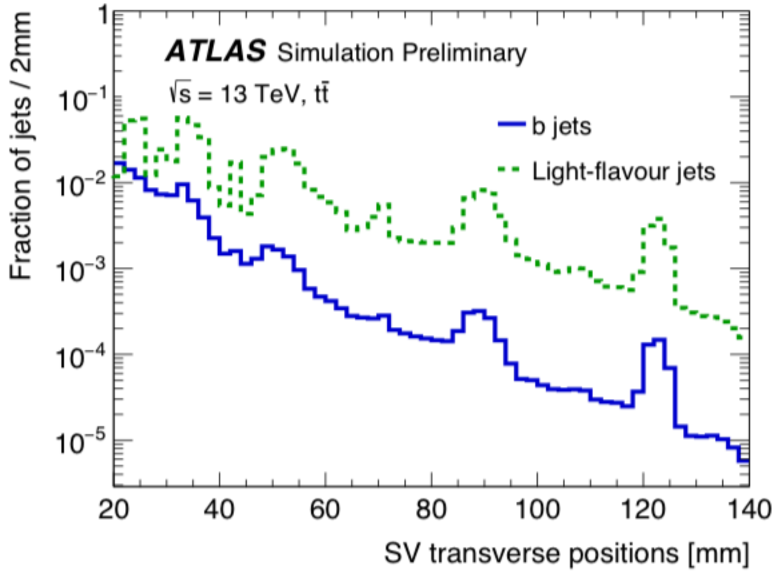
https://svnweb.cern.ch/trac/atlasoff/browser/InnerDetector/InDetRecTools/InDetVKalVxInJetTool/tags/InDetVKalVxInJetTool-00-09-03?order=name

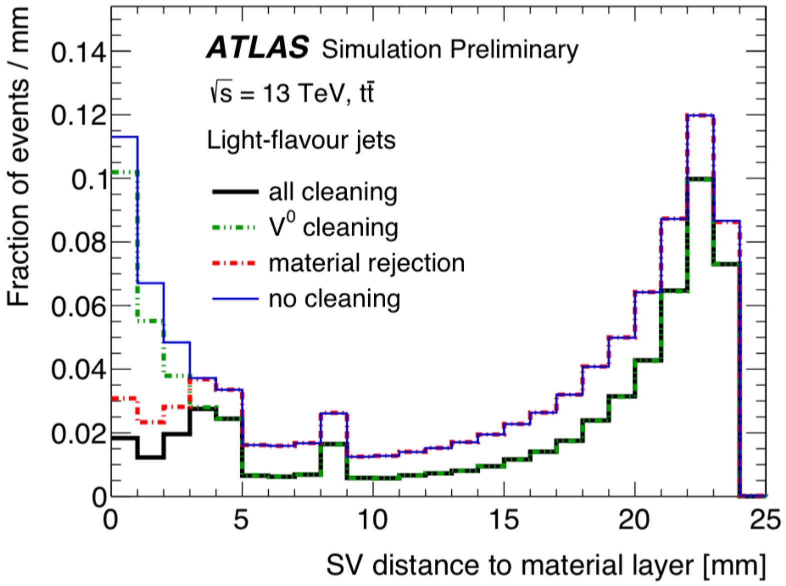
Vertexing from Tracks and Cleaning of 2-track Vertices Candidates

Selecting for 2-track Vertices Candidate Tracks

- Track quality
  - Number of silicon hits ≥ 7
  - Quality of track fit chi2 ≤ 3
  - At most one pixel shared hit with another track
  - At most two shared hits in the precision detectors (pixel and SCT)
- Pile-up, hadronic interaction with detector material, long lifetime particle vertices cleaning
  - tracks with low $d0/\sigma d0$ (<2) and high $z0/\sigma z$ (>6) are removed (remove pile-up)
  - secondary tracks cannot have associated detector hits with radii smaller than the production vertex radius
  - Vertices with radius within detector regions are excluded

http://cds.cern.ch/record/2270366/files/ATL-PHYS-PUB-2017-011.pdf?version=1



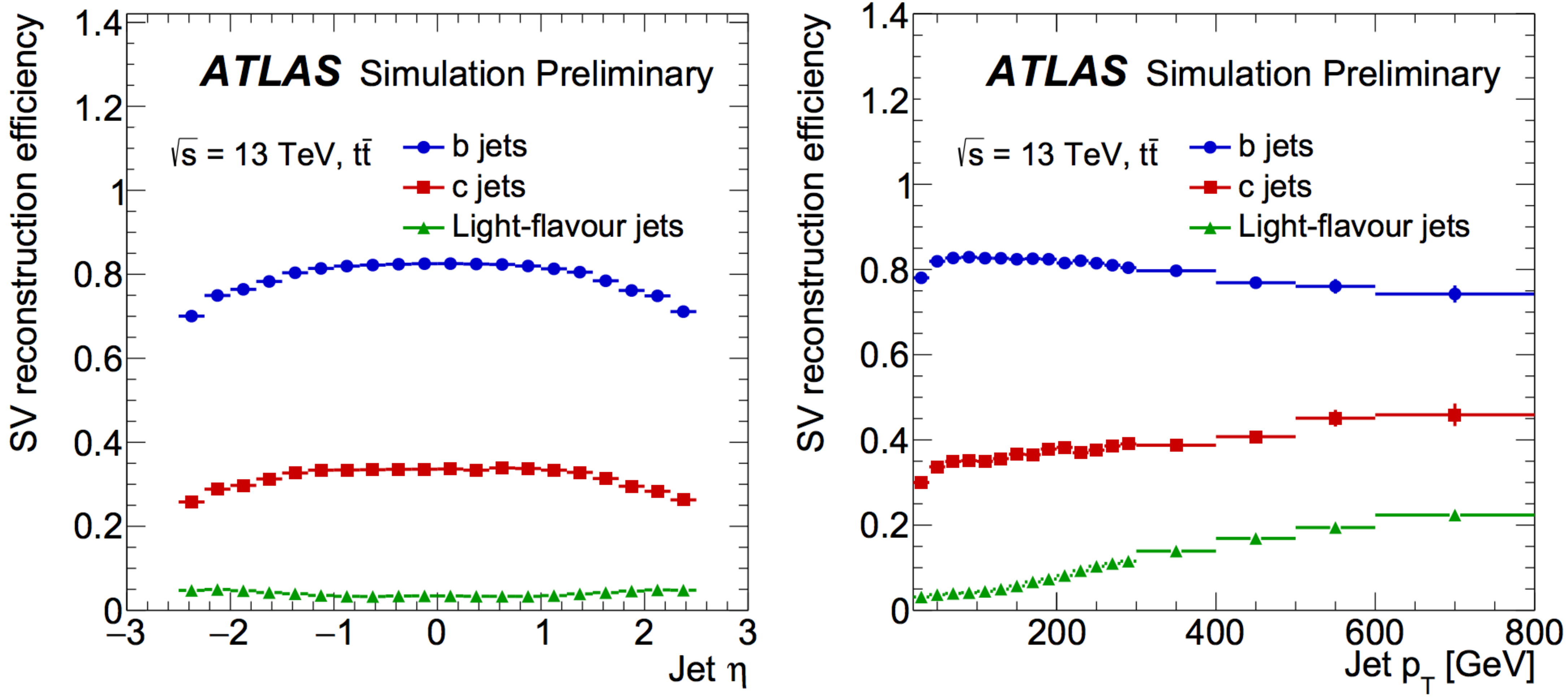


Vertexing from N tracks

- Utilizes the package VKalVrt, which iteratively removes track of highest chi^2 with initial vertex position and recompute until all tracks have acceptable chi^2.

https://cds.cern.ch/record/685551/files/phys-2003-031.pdf?version=1

Secondary Vertex Reconstruction Performance



- Efficiency: fractions of jets with a secondary vertex reconstructed.
- 13 TeV pp->ttbar MC events
- http://cds.cern.ch/record/2270366/files/ATL-PHYS-PUB-2017-011.pdf?version=1

Secondary Vertex Reconstruction Performance

- Reconstructed SV xyz directly available from b-tagging object and is already saved in branches in retagging
- Truth primary vertex: signalProcessVertex(), truth interaction position.
- Truth secondary vertex: decayVtx(), decay position.
- To do:
  - Plot 3D decayVtx()-signalProcessVertex().

To do: SV0/SV1/SV2 Tagger