MACHINE LEARNING FOR B-JET TAGGING

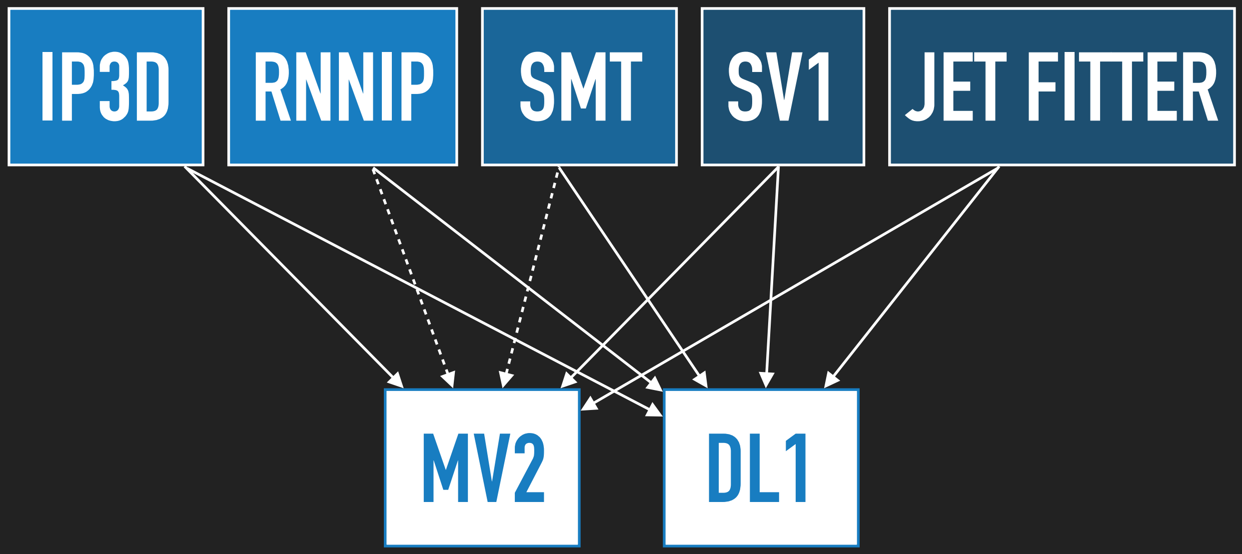
MICHELA PAGANINI
ON BEHALF OF THE ATLAS COLLABORATION

IN THIS TALK

ATLAS-PHYS-PUB-2017-003

ATLAS-PHYS-PUB-2017-013

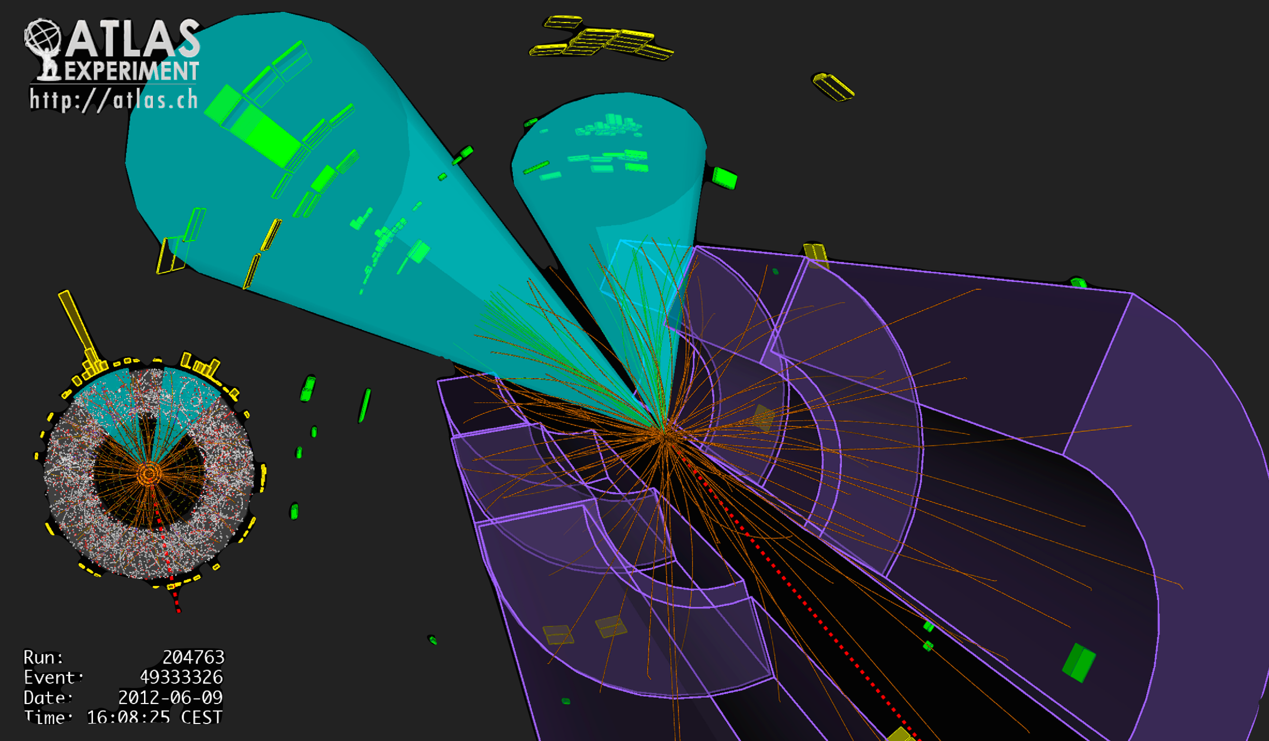
- Description of flavor tagging ecosystem for the 2017-18 run
- Focus on Deep Learning-powered improvements

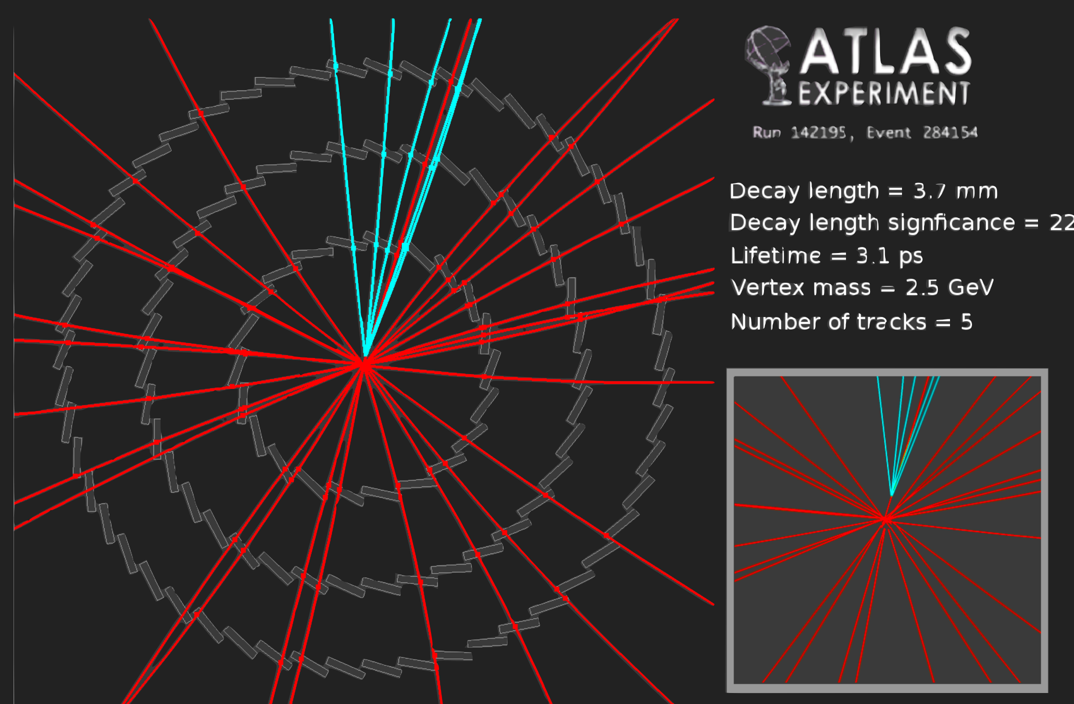


ATLAS
EXPERIMENT

GOAL OF FLAVOR TAGGING

Separate jets that contain ***b*-hadrons** from jets initiated by lighter quark flavors





- Average *b*-hadron lifetime → distance travelled before decaying (~mm) ideal for detection in ATLAS

B-HADRON DECAY

- *b*-hadron contains *b* quark,
  which decays through a cascade

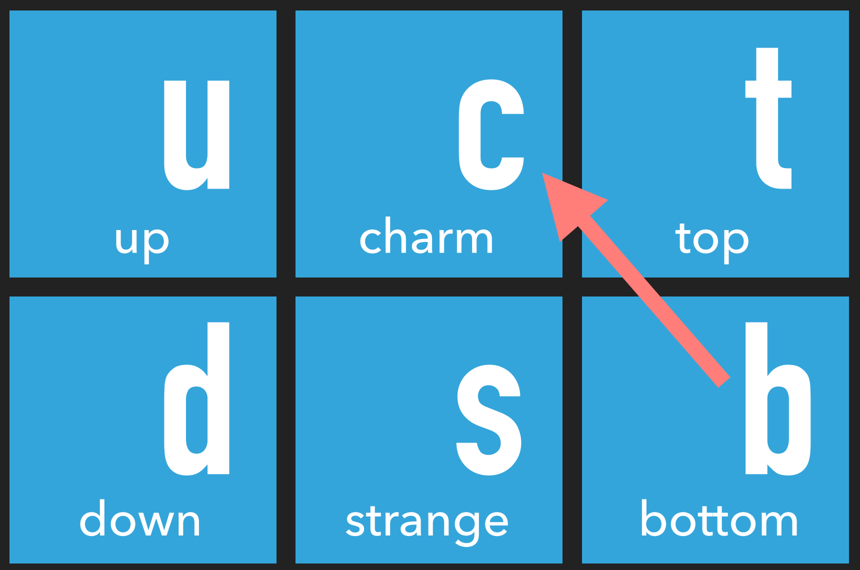
u up	c charm	t top
d down	s strange	b bottom

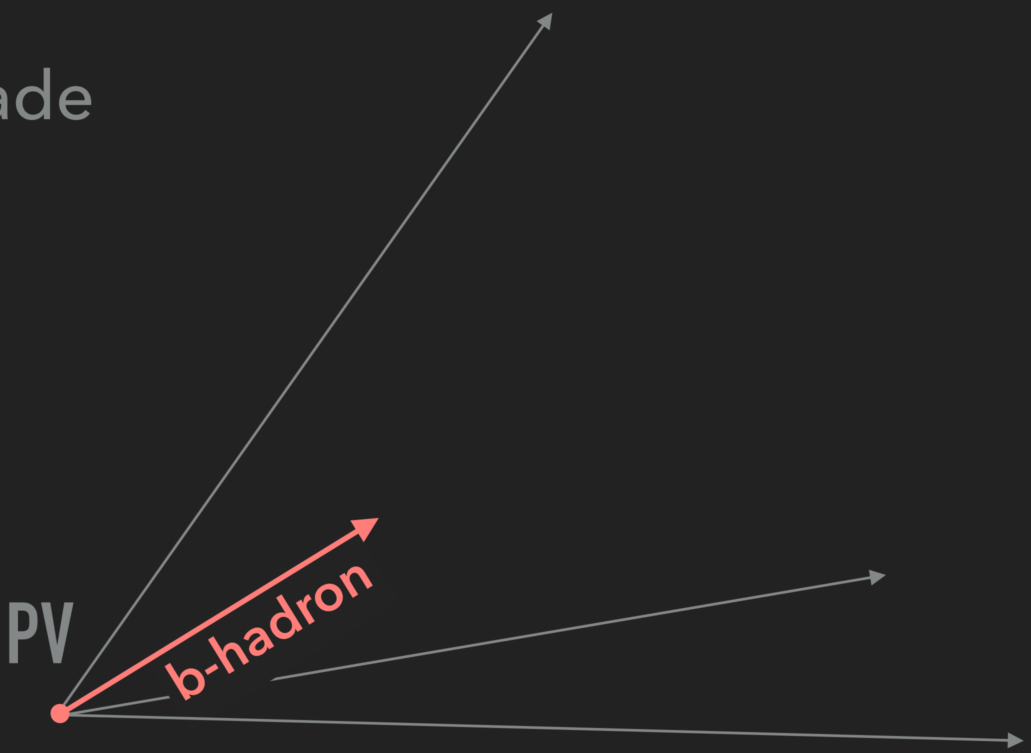
Limited by detector resolution, pileup, tracking inefficiency, material interactions, and long-lived decays for light jets

Slide inspired by D. Guest

B-HADRON DECAY

- *b*-hadron contains *b* quark, which decays through a cascade





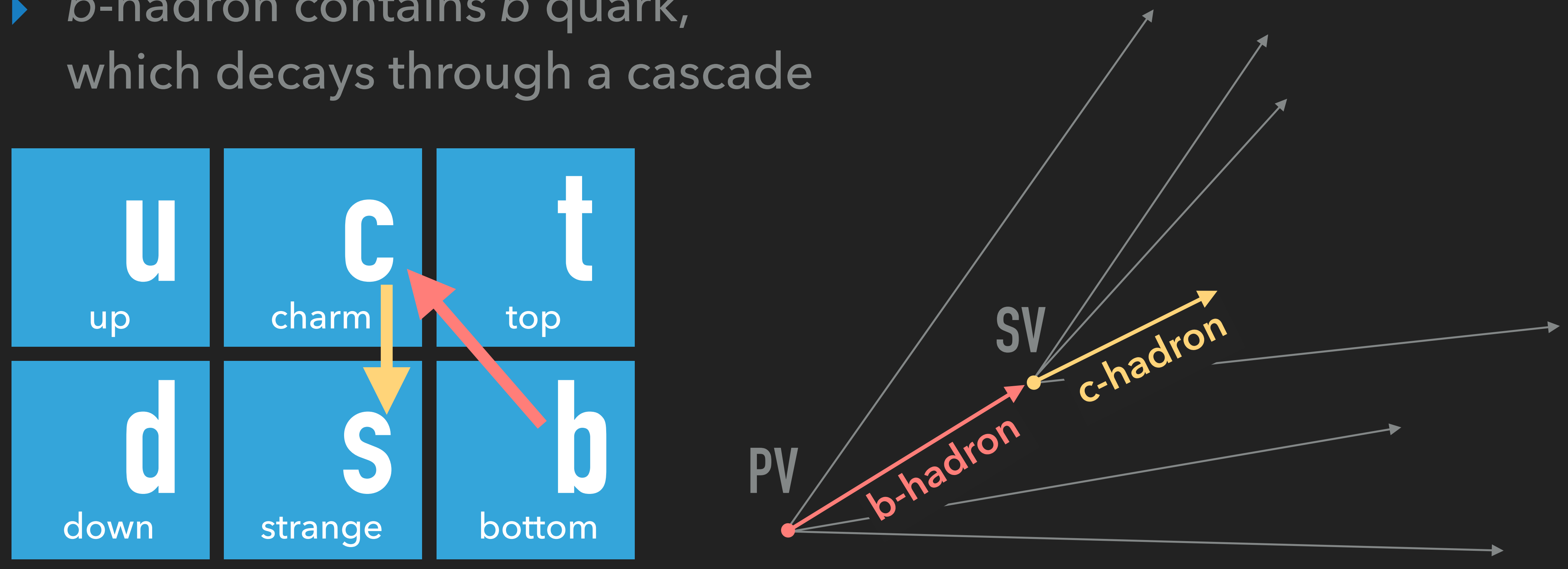
Limited by detector resolution, pileup, tracking inefficiency, material interactions, and long-lived decays for light jets

ATLAS EXPERIMENT

Slide inspired by D. Guest

B-HADRON DECAY

- *b*-hadron contains *b* quark, which decays through a cascade

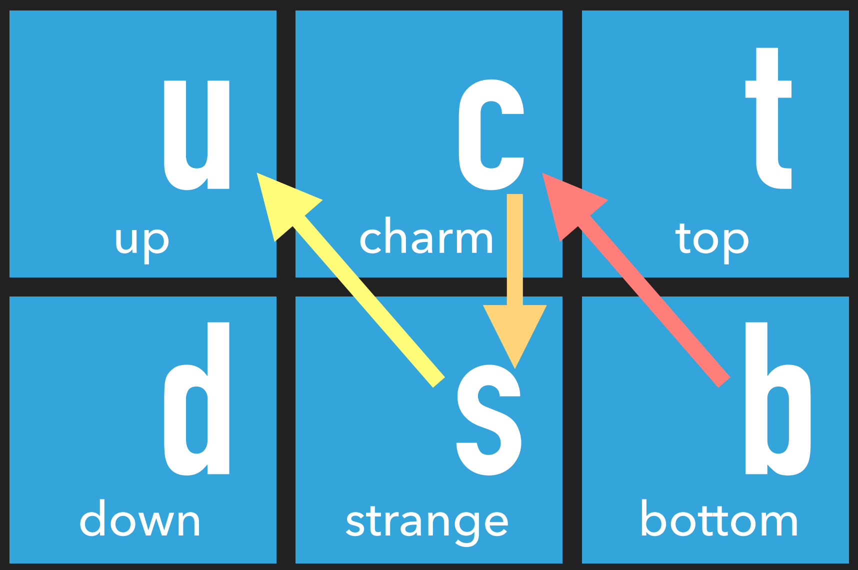


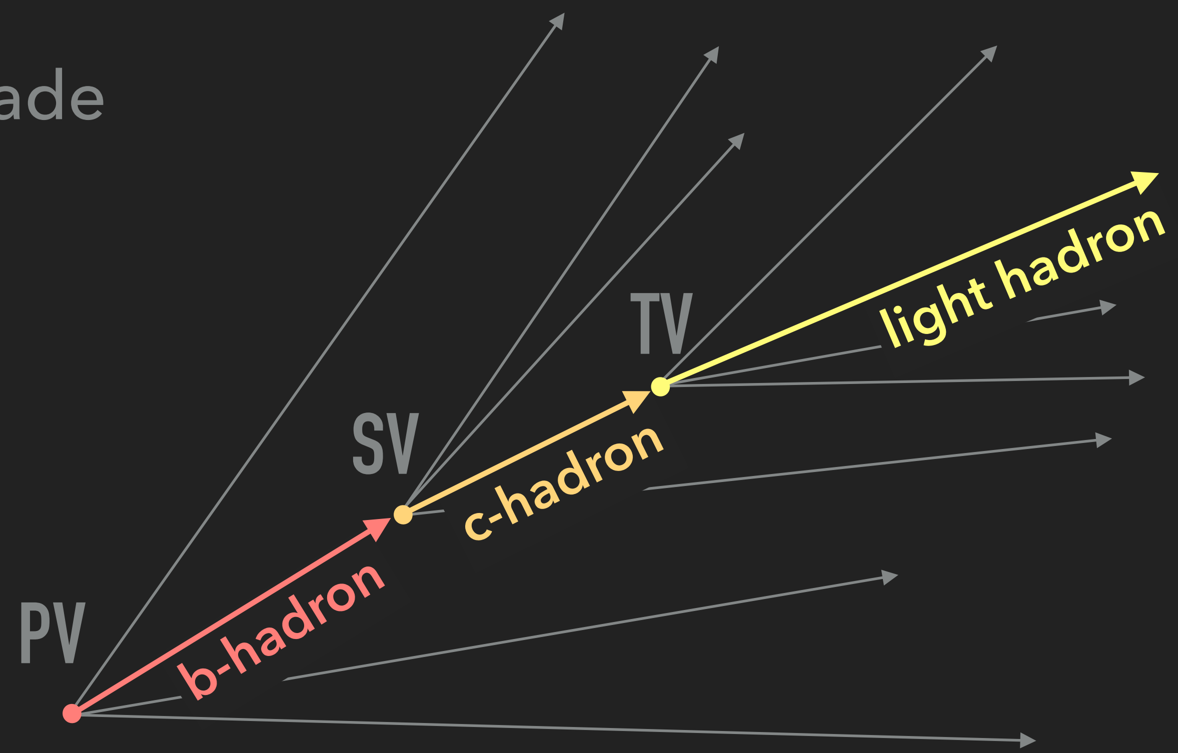
Limited by detector resolution, pileup, tracking inefficiency, material interactions, and long-lived decays for light jets

Slide inspired by D. Guest

B-HADRON DECAY

- *b*-hadron contains *b* quark, which decays through a cascade



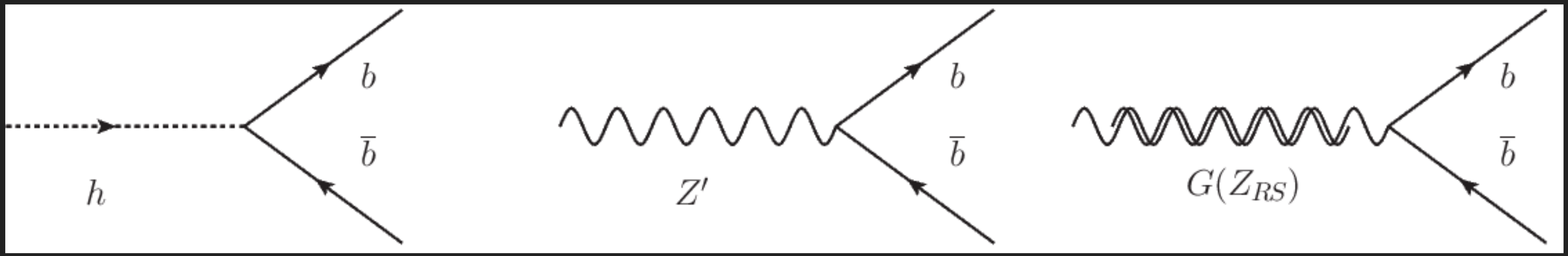


Limited by detector resolution, pileup, tracking inefficiency, material interactions, and long-lived decays for light jets

ATLAS EXPERIMENT

Slide inspired by D. Guest

- Truth labels:
  - *b*: if *b*-hadron with p_T > 5 GeV within ΔR=0.3 of jet axis
  - *c*: if not *b* & *c*-hadron with p_T > 5 GeV within ΔR=0.3 of jet axis
  - τ: if not *b* or *c* & τ-lepton with p_T > 5 GeV within ΔR=0.3 of jet axis
  - light: otherwise
- Important for ATLAS Physics program ($H \rightarrow bb$, SUSY, …)

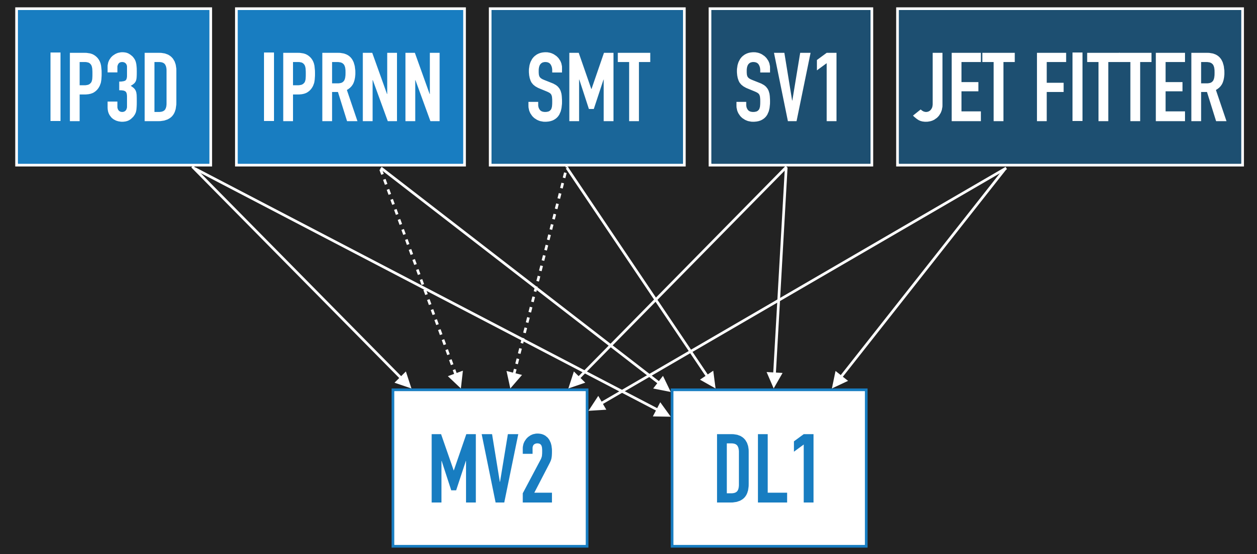


ATLAS EXPERIMENT

Diagrams from Quantum Diaries

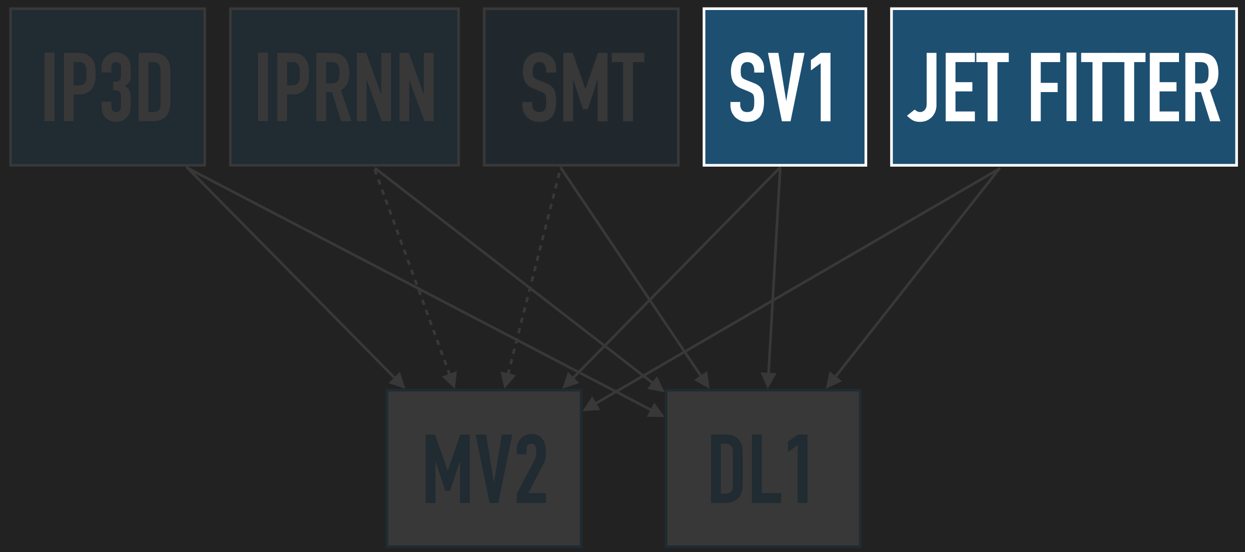
LOW LEVEL TAGGERS

VERTEX FINDING ALGORITHMS



LOW LEVEL TAGGERS

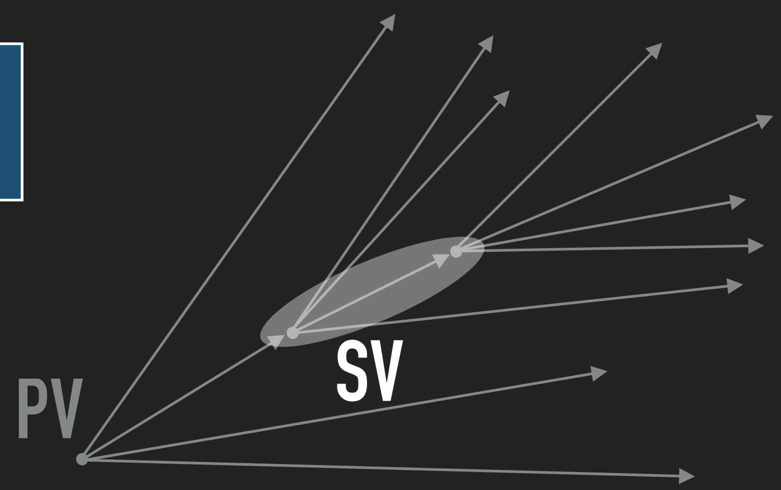
VERTEX FINDING ALGORITHMS



ATLAS
EXPERIMENT

VERTEX FINDING ALGORITHMS

SV1

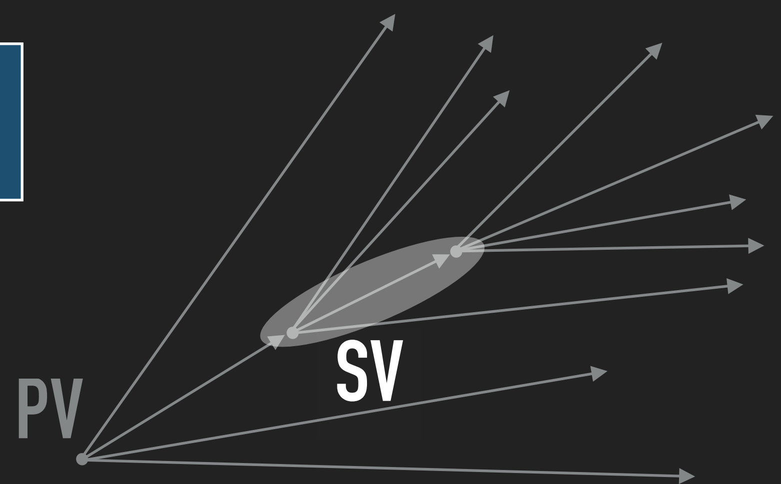


JET FITTER

- reconstructs a single displaced vertex

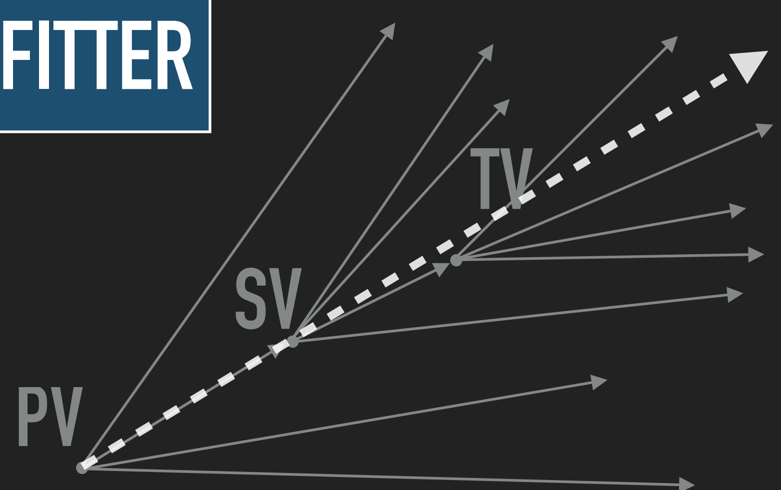
VERTEX FINDING ALGORITHMS

SV1



- reconstructs a single displaced vertex

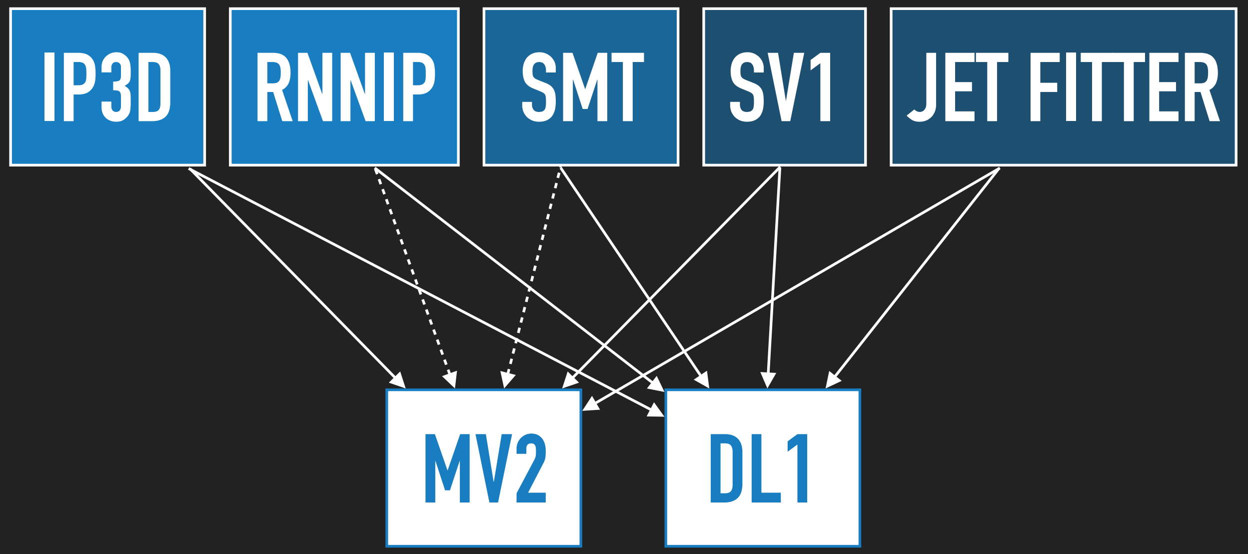
JET FITTER



- performs a topological decay reconstruction along the *b*-hadron line of flight

LOW LEVEL TAGGERS

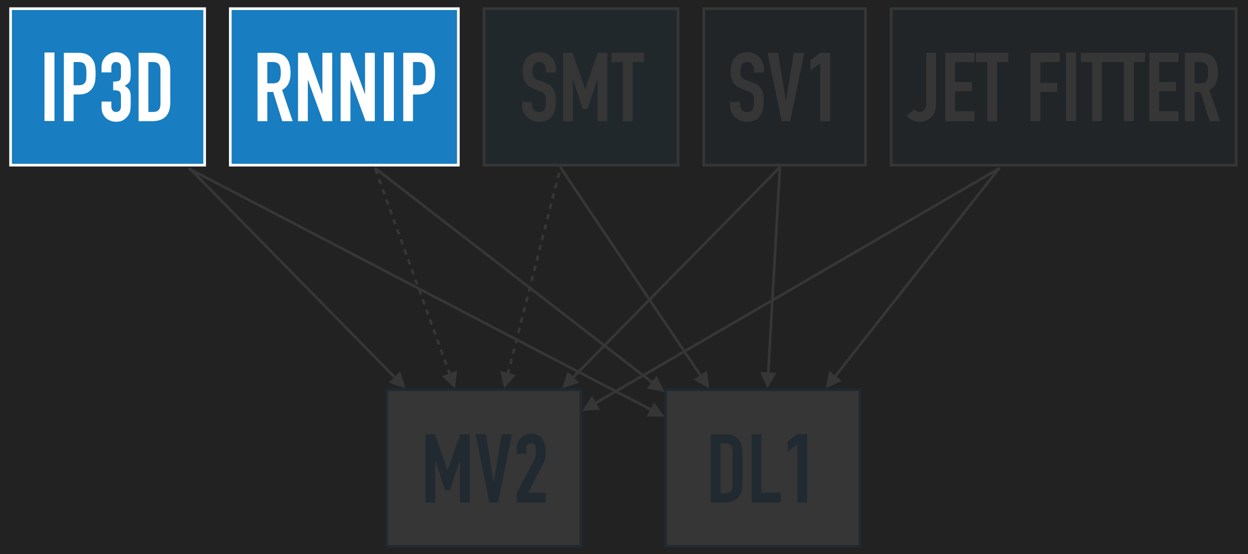
IMPACT PARAMETER TAGGERS



ATLAS
EXPERIMENT

LOW LEVEL TAGGERS

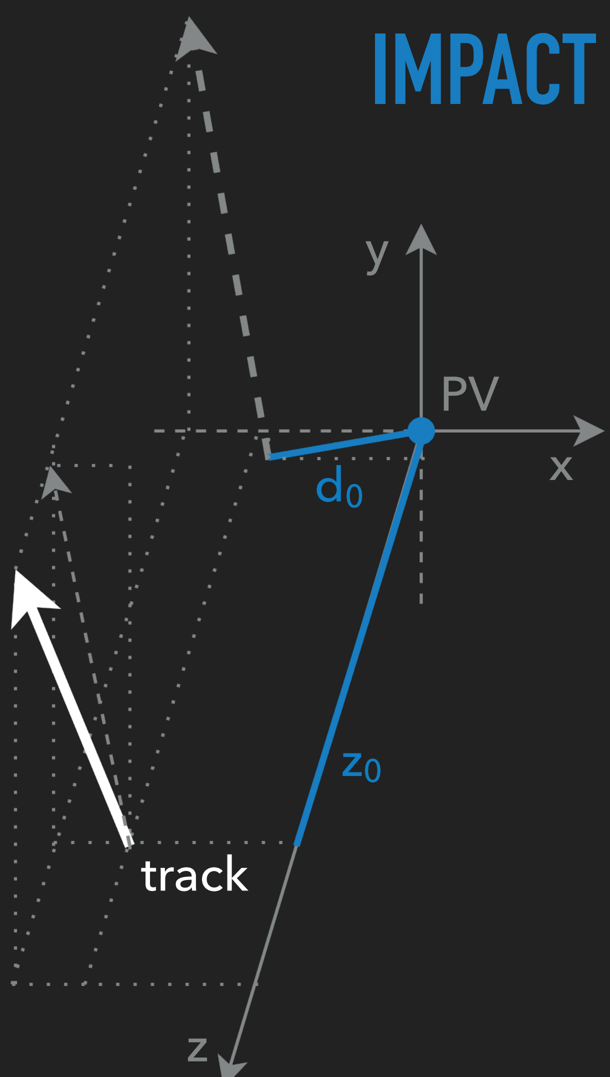
IMPACT PARAMETER TAGGERS



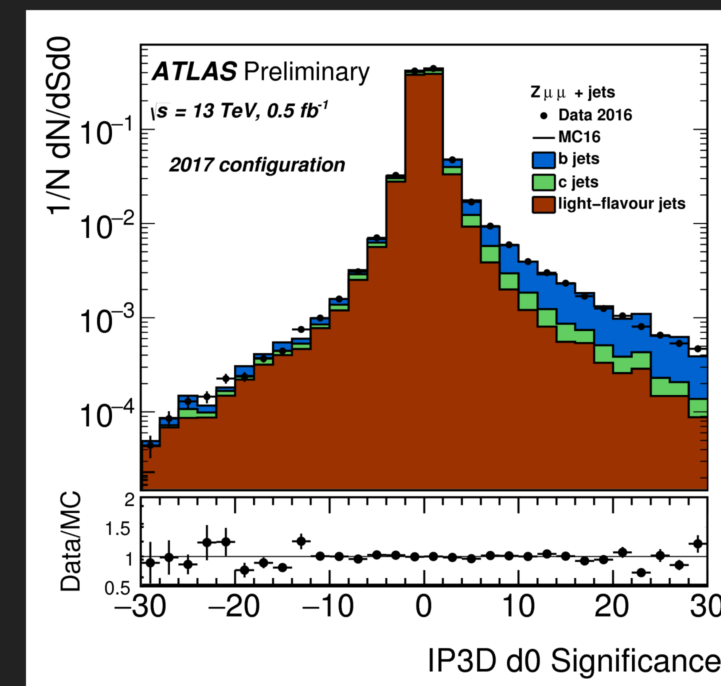
ATLAS EXPERIMENT

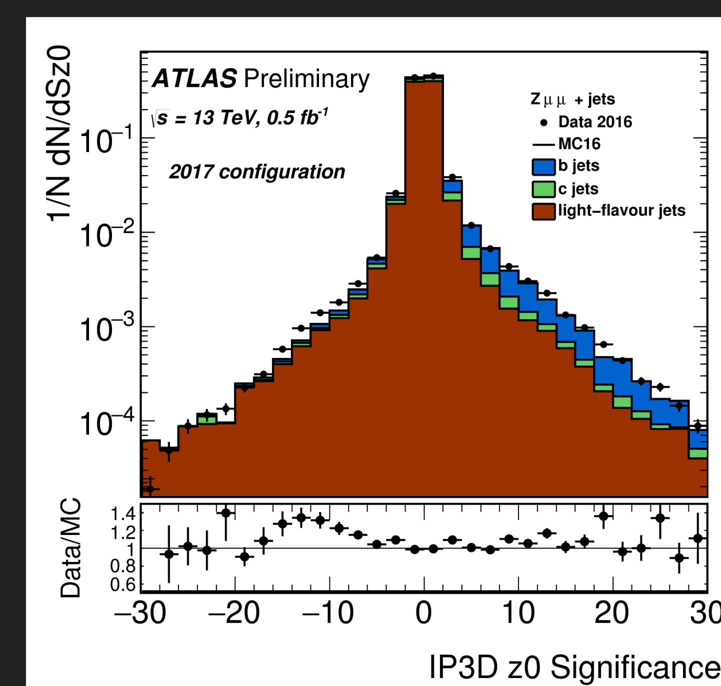
IMPACT PARAMETER TAGGERS

IP3D RNNIP



- measures compatibility of track with primary vertex hypothesis
- binned 2D likelihood per grade category using each track's transverse ($S_{d_0}=d_0/\sigma_{d_0}$) and longitudinal ($S_{z_0}=z_0/\sigma_{z_0}$) impact parameter significances
- light: significance consistent with 0

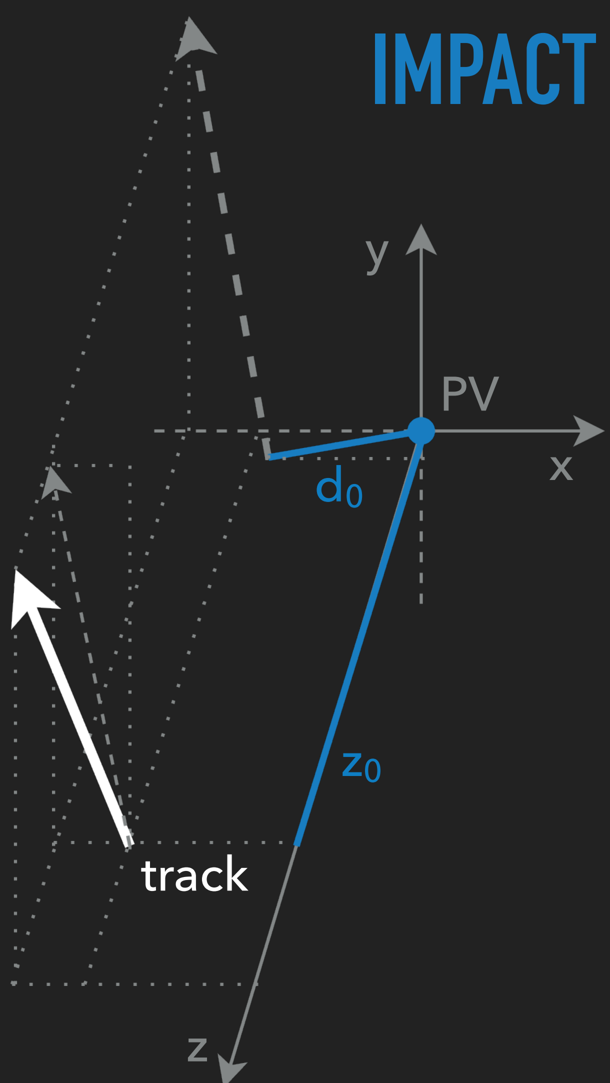




ATLAS EXPERIMENT

IMPACT PARAMETER TAGGERS

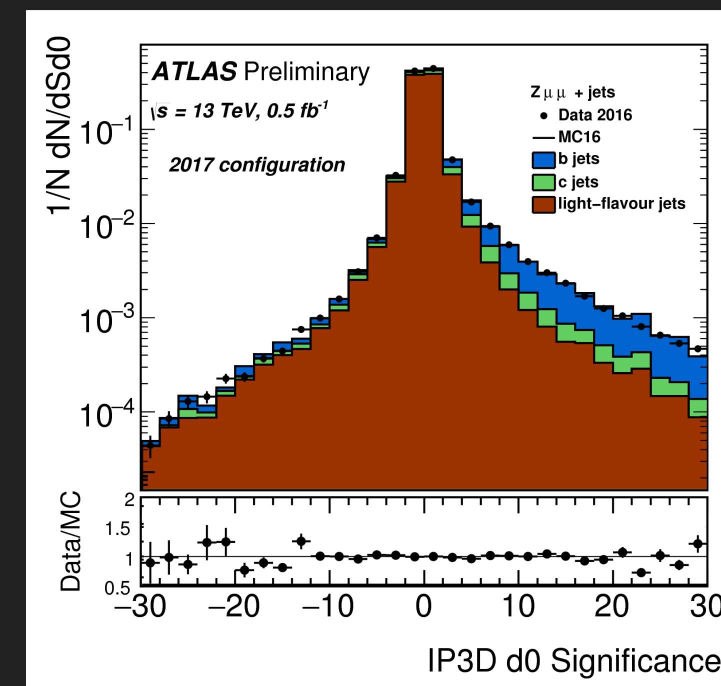
IP3D RNNIP

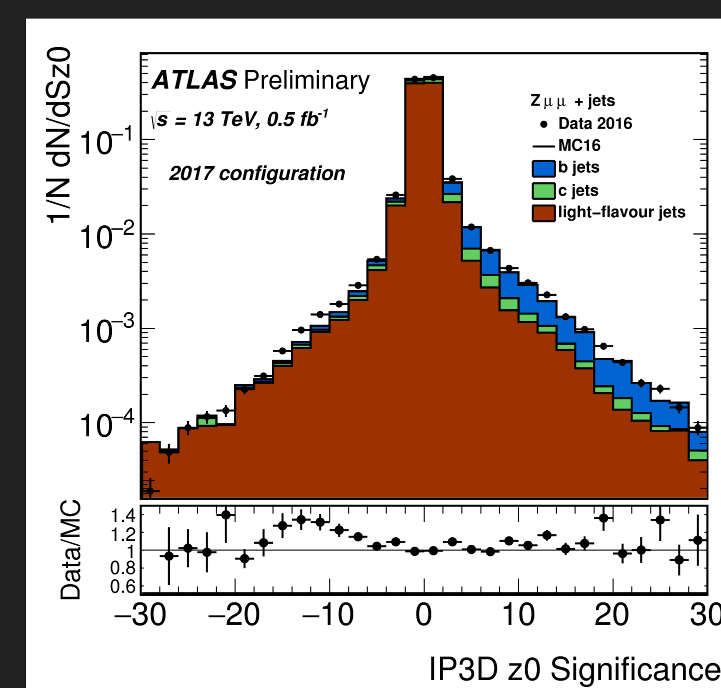


- measures compatibility of track with primary vertex hypothesis
- binned 2D likelihood per grade category using each track's transverse ($S_{d_0}=d_0/\sigma_{d_0}$) and longitudinal ($S_{z_0}=z_0/\sigma_{z_0}$) impact parameter significances
- light: significance consistent with 0

$$\text{IP3D LLR} = \sum_{i=1}^{N} \log \frac{p_{b_i}}{p_{u_i}}$$

sum over tracks in a jet

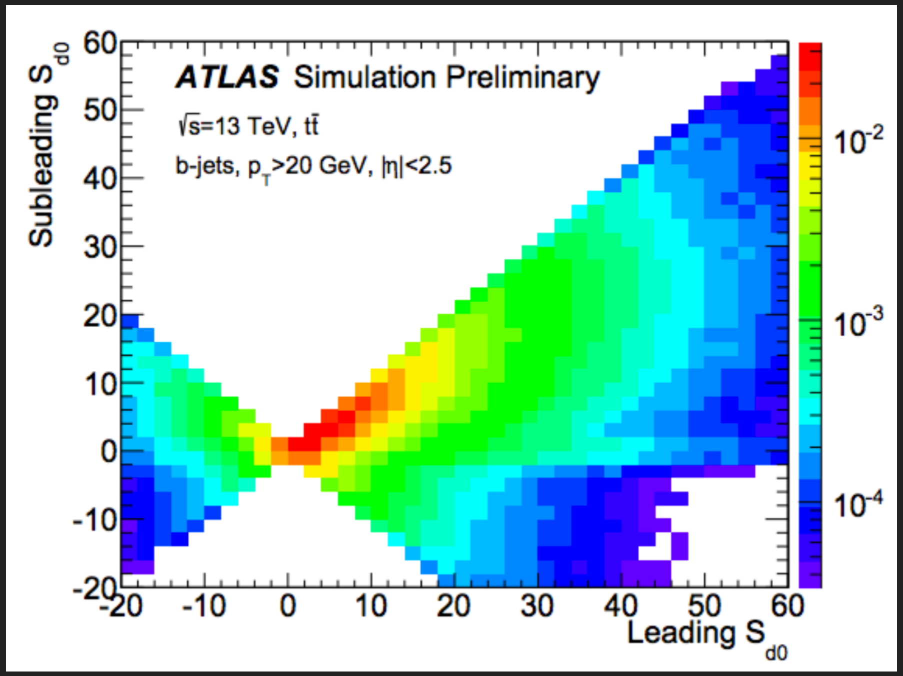




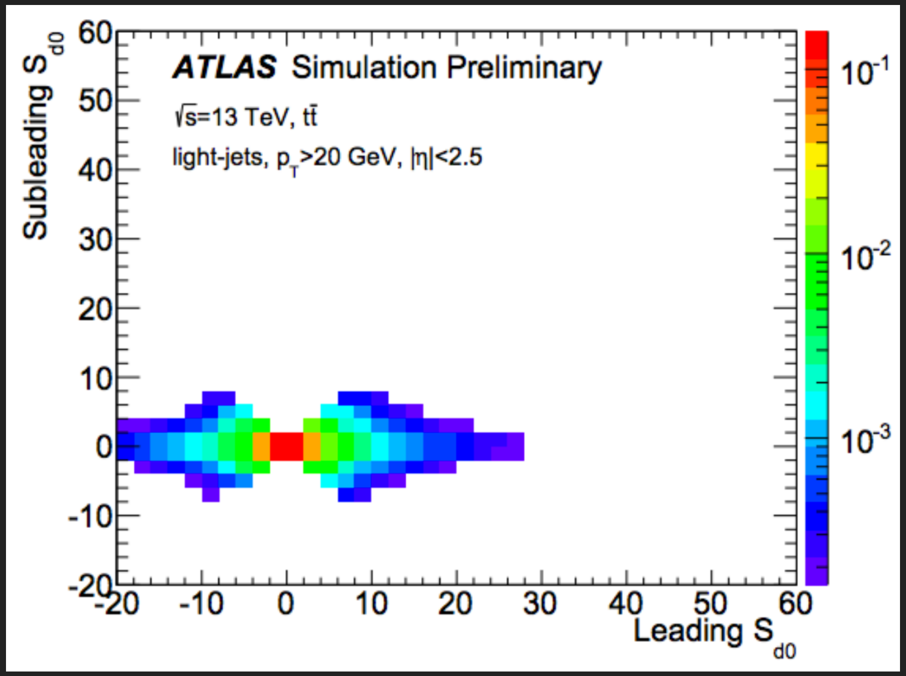
IMPACT PARAMETER TAGGERS

IP3D RNNIP

- Based on Recurrent Neural Networks
- Exploits correlation among tracks, neglected by IP3D



b-jets

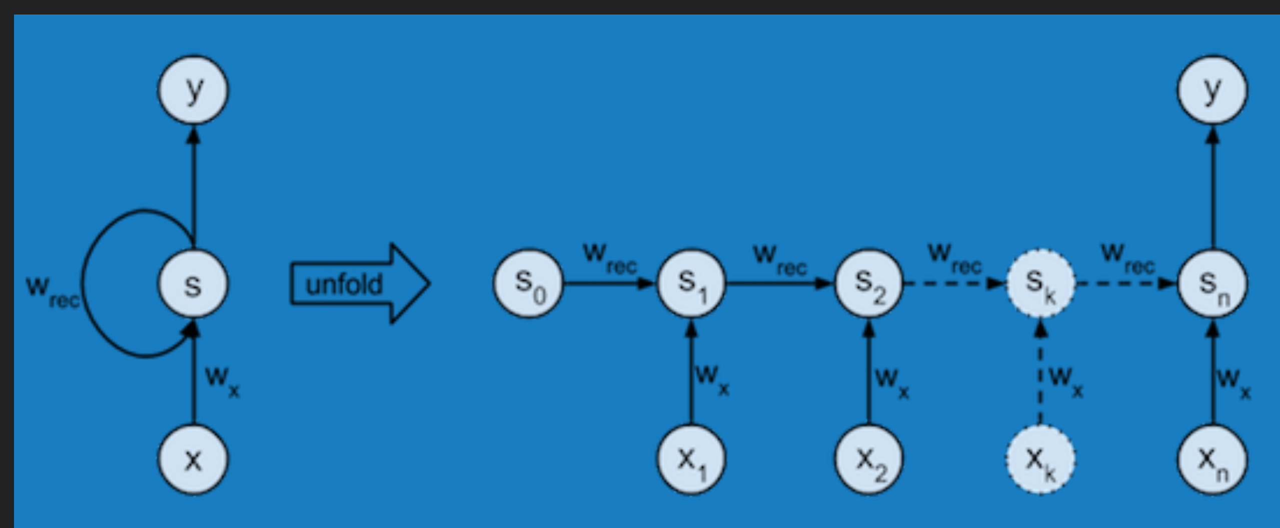


light jets

ATLAS EXPERIMENT

RECURRENT NEURAL NETWORKS

- Neural network unit to learn **sequence-based dependencies** for arbitrary-length input sequences

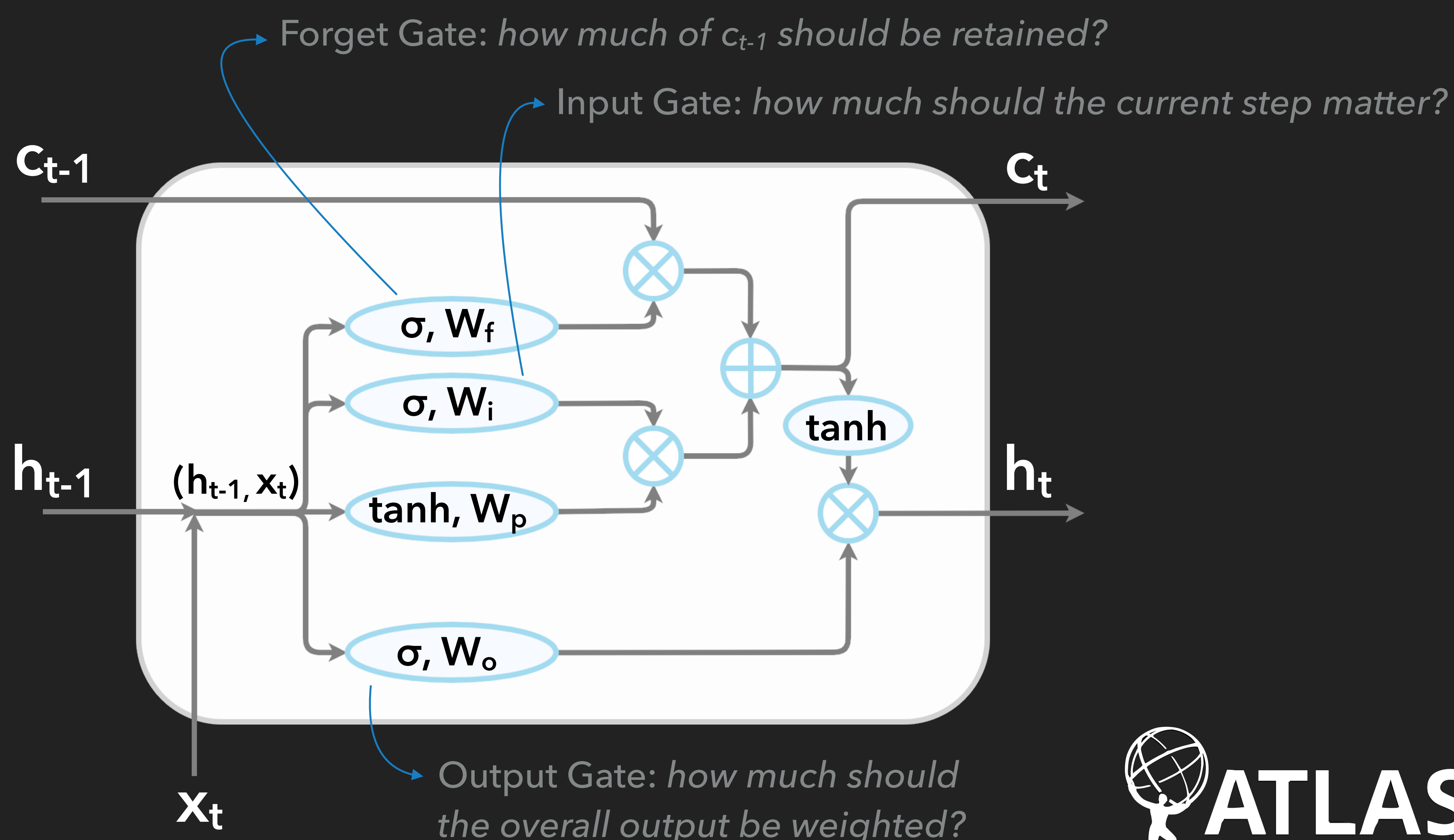


from Peter Roelants

- Cell holds internal state vector
- Identically applied to every entry in sequence
- Recurrent loop feeds back into cell

LSTM

- Long-Short Term Memory units

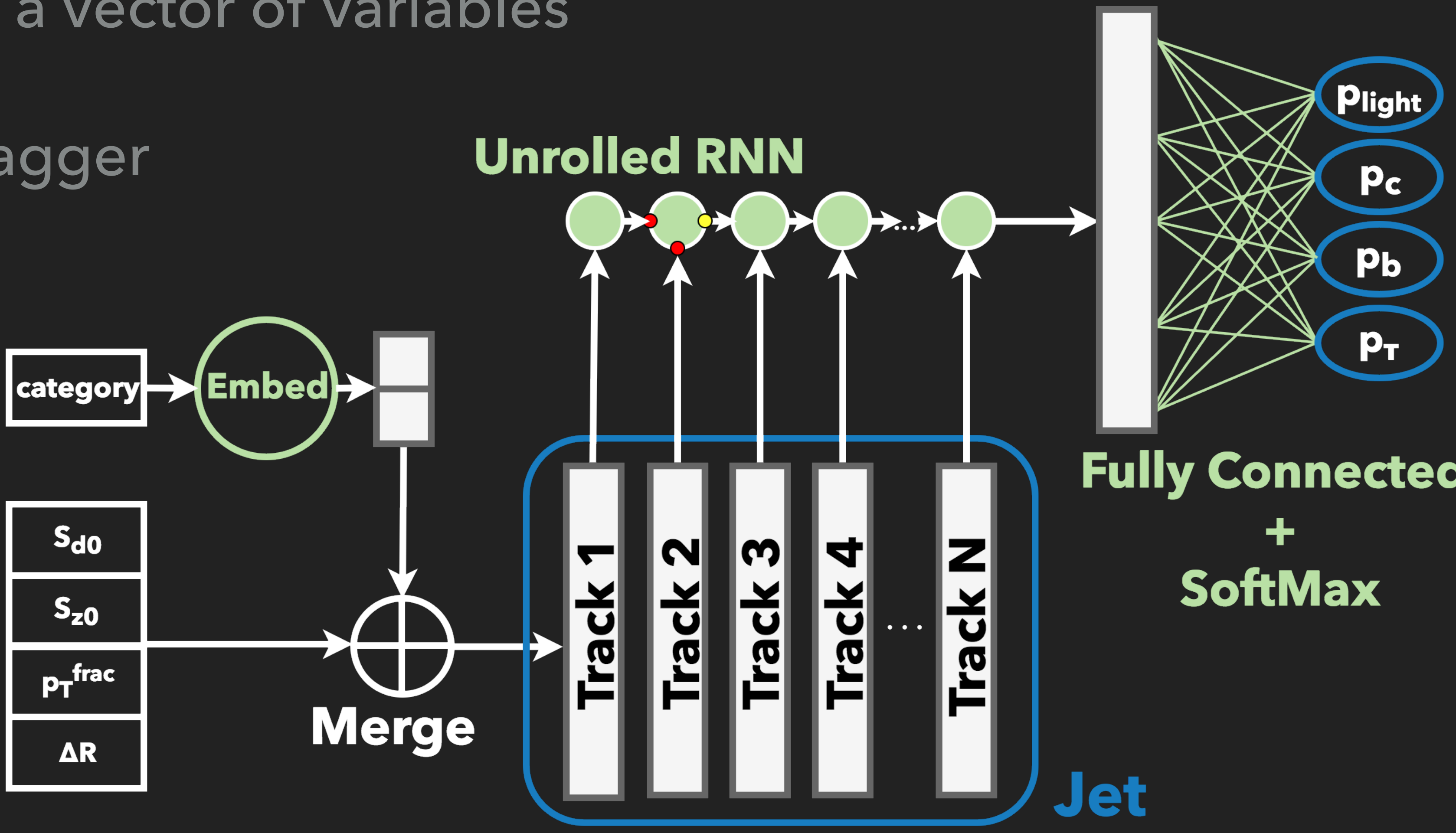


ATLAS EXPERIMENT

IMPACT PARAMETER TAGGERS

RNNIP

- Represent jets as a sequence of tracks ordered by $|S_{d_0}|$
- Each track is a vector of variables
- Multi-class tagger



ATLAS EXPERIMENT

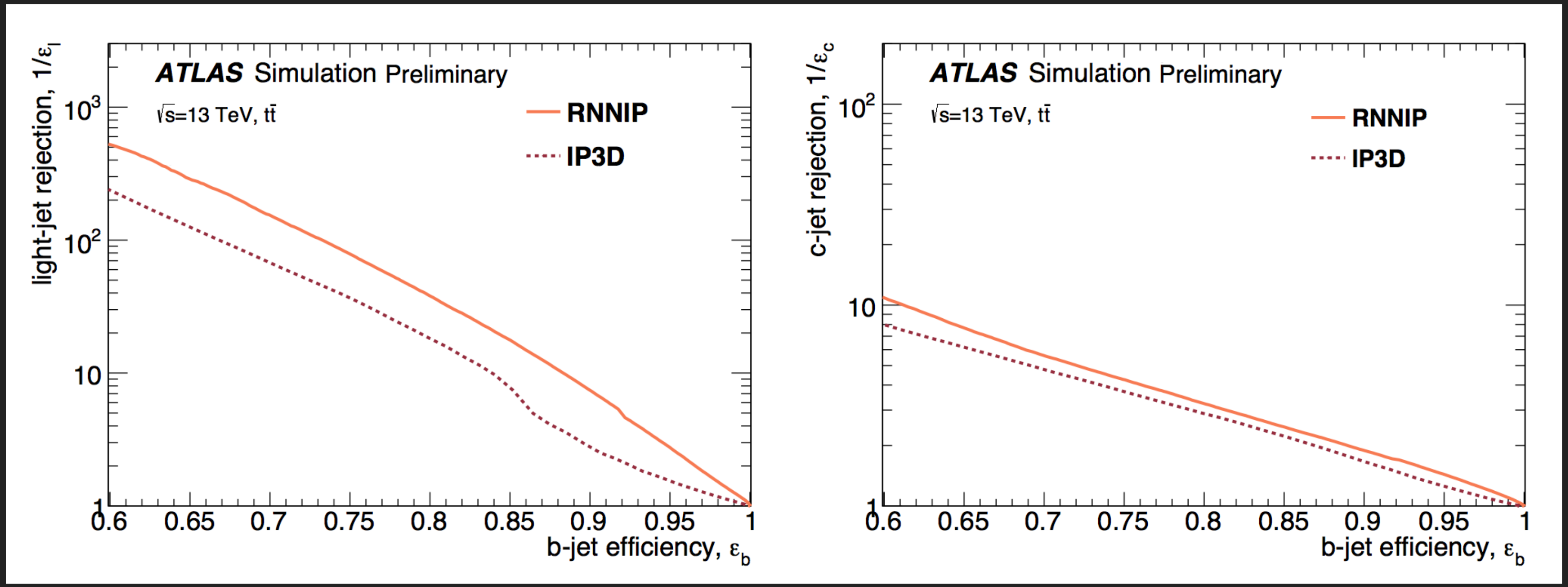
IMPACT PARAMETER TAGGERS

RNNIP

- Combine output in discriminant:

$$D_{\mathrm{RNN}}(b) = \ln \frac{p_b}{f_c p_c + f_\tau p_\tau + (1 - f_c - f_\tau) p_{\mathrm{light}}}$$

- Can be tuned after training



- IP3D and RNNIP tagged jets are partly complementary → increased performance when both are inputs to subsequent tagger

ATLAS EXPERIMENT

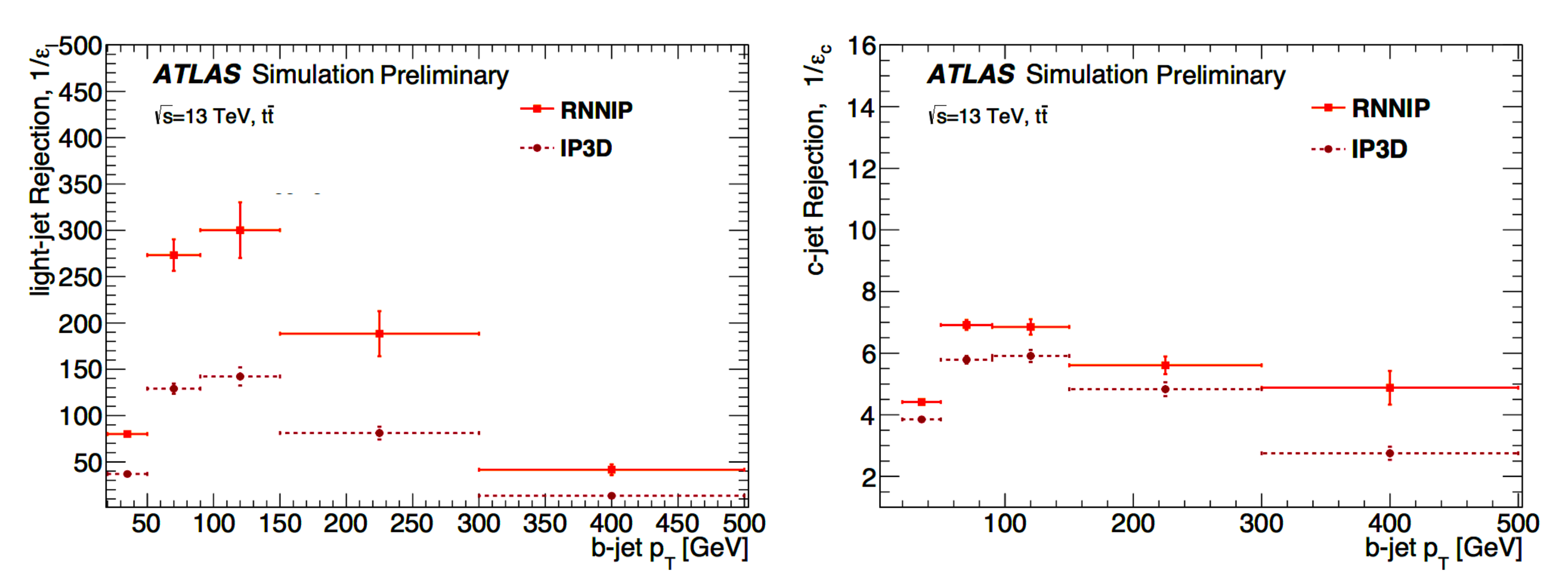
IMPACT PARAMETER TAGGERS

RNNIP

- Combine output in discriminant:

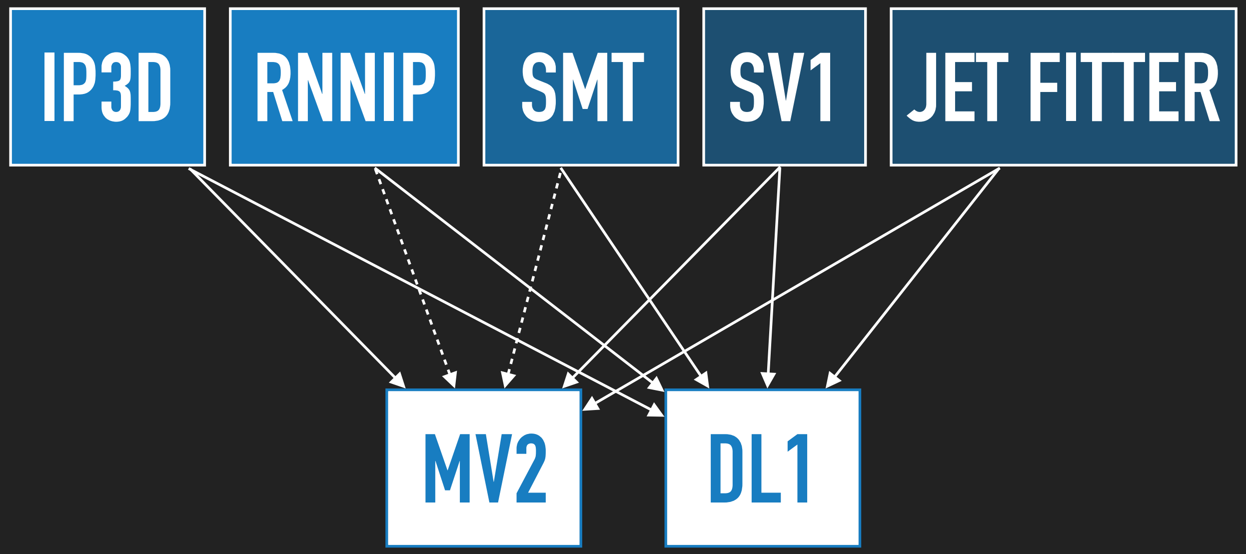
$$D_{\mathrm{RNN}}(b) = \ln \frac{p_b}{f_c p_c + f_\tau p_\tau + (1 - f_c - f_\tau) p_{\mathrm{light}}}$$

- Can be tuned after training



- IP3D and RNNIP tagged jets are partly complementary → increased performance when both are inputs to subsequent tagger

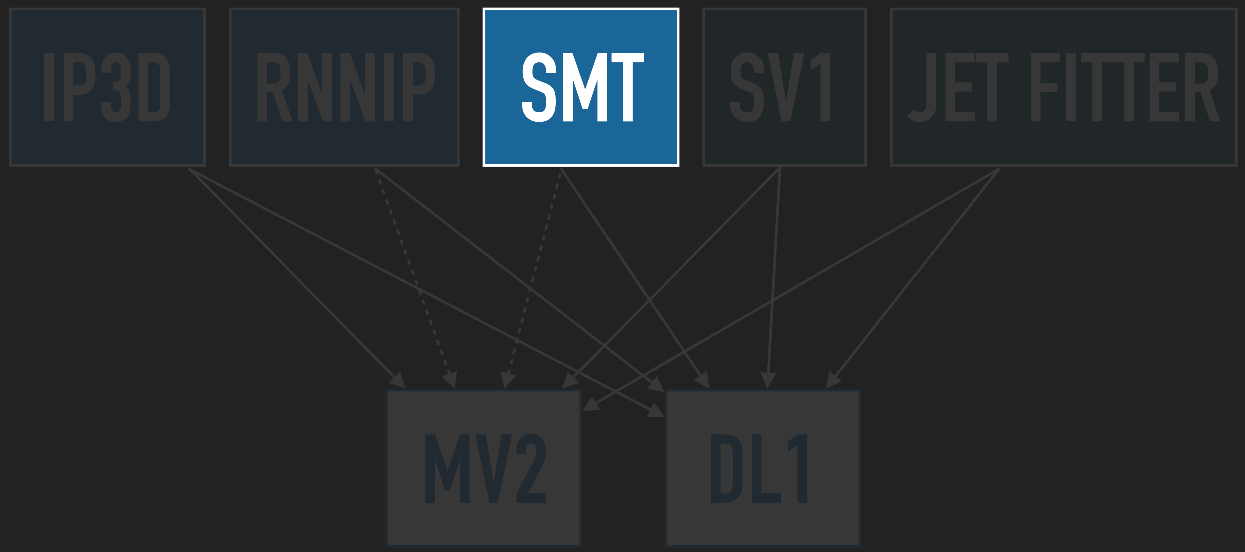
SOFT MUON TAGGER



ATLAS EXPERIMENT

LOW LEVEL TAGGERS

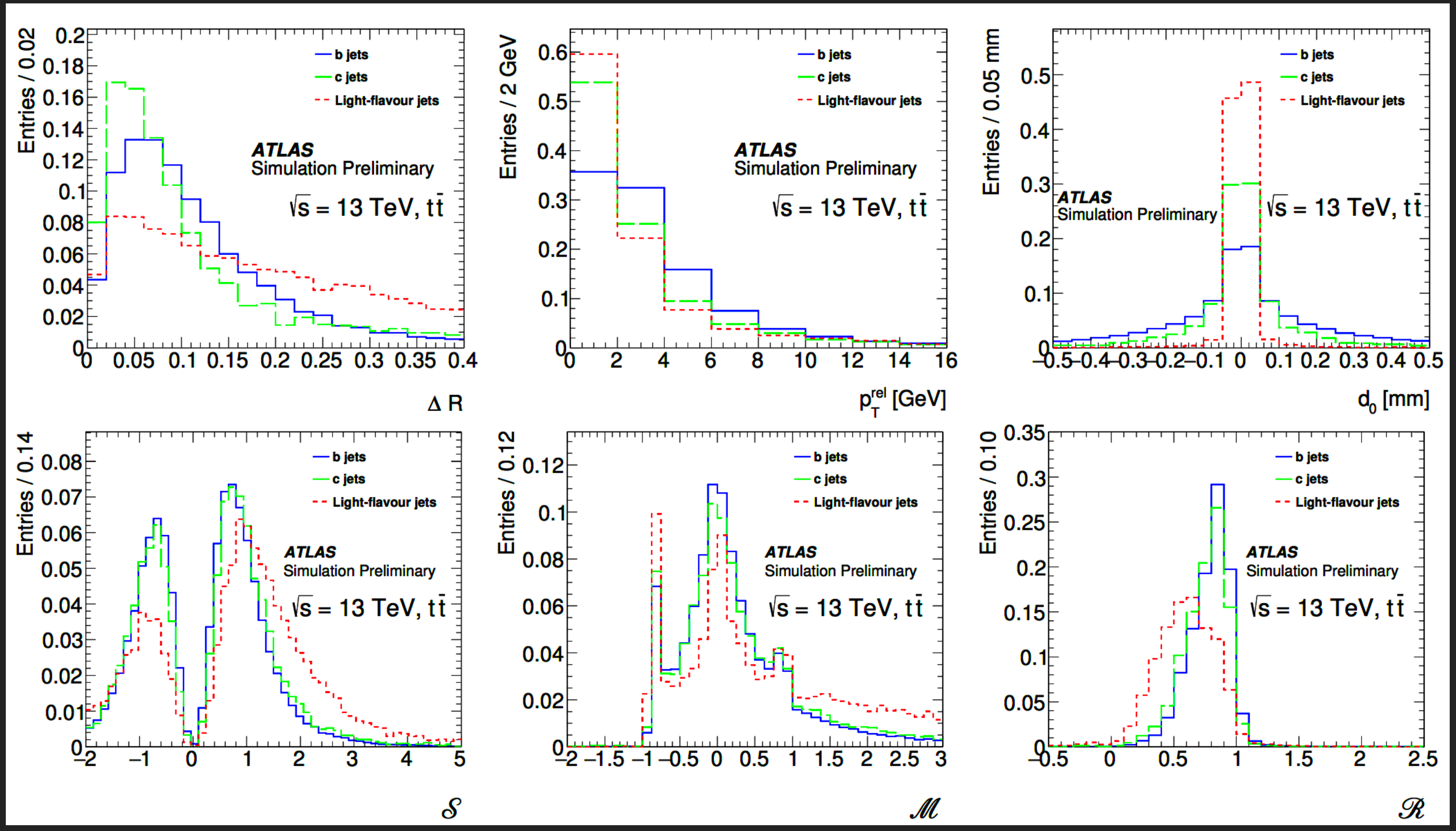
SOFT MUON TAGGER



SOFT MUON TAGGER

SMT

- Reconstructs muons from semi-leptonic decays
- Limited by the semi-leptonic branching ratio BR(b → μ ν X) + BR(b → c → μ ν X) ≈ 21%
- Complementary to other low level taggers that are based on lifetime information



Defined new variables to separate muons from *b*-decays, and bkg muons from decays in flight of pions and kaons:

$$\mathcal{S} = q \sum_i \frac{\Delta\phi^i_{\text{scat}}}{\sigma_{\Delta\phi^i_{\text{scat}}}}$$

$$\mathcal{M} = \frac{p_{\text{ID}} - p^{\text{extr}}_{\text{MS}}}{\sigma_{E_{\text{loss}}}}$$

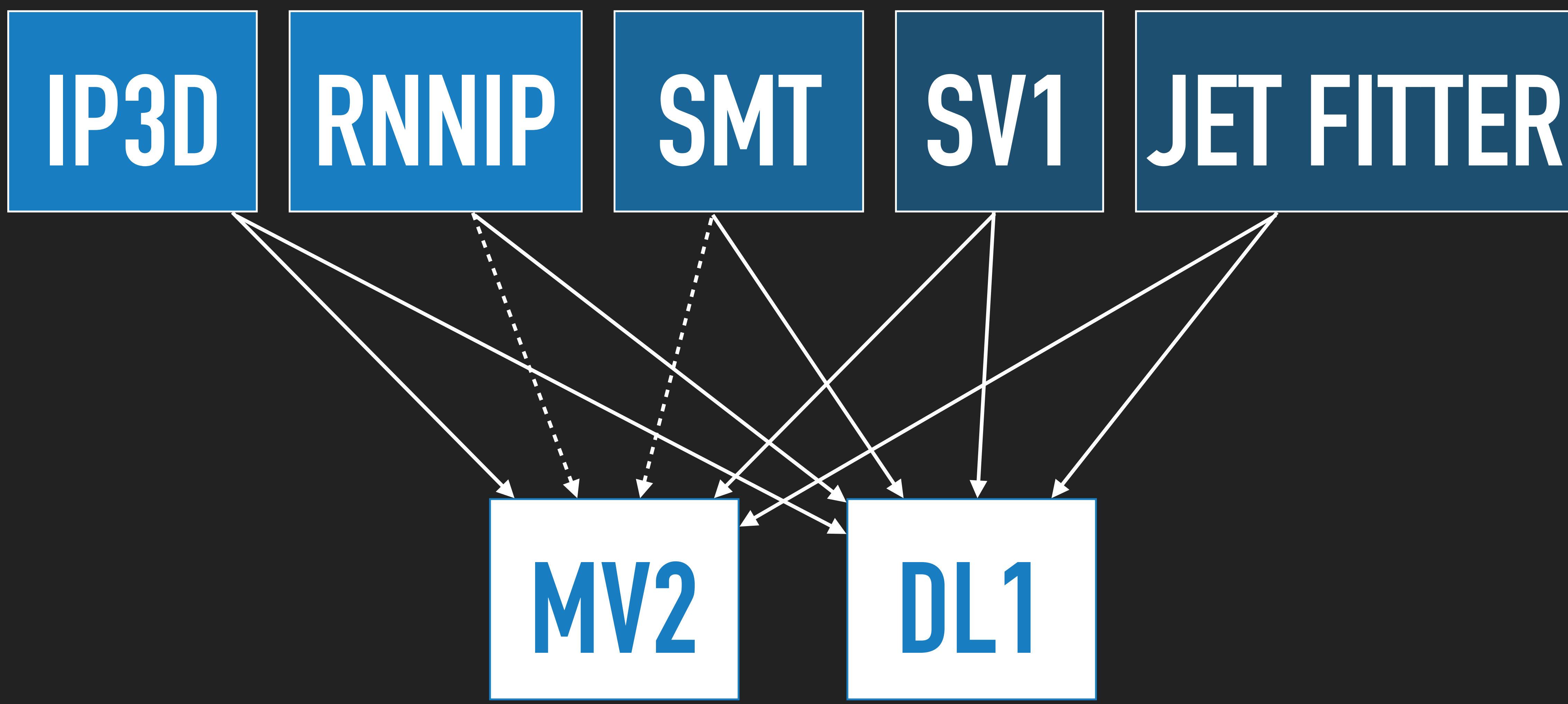
$$\mathcal{R} = \frac{(q/p)_{\text{ID}}}{(q/p)_{\text{MS}}}$$

ATLAS EXPERIMENT

ATLAS EXPERIMENT

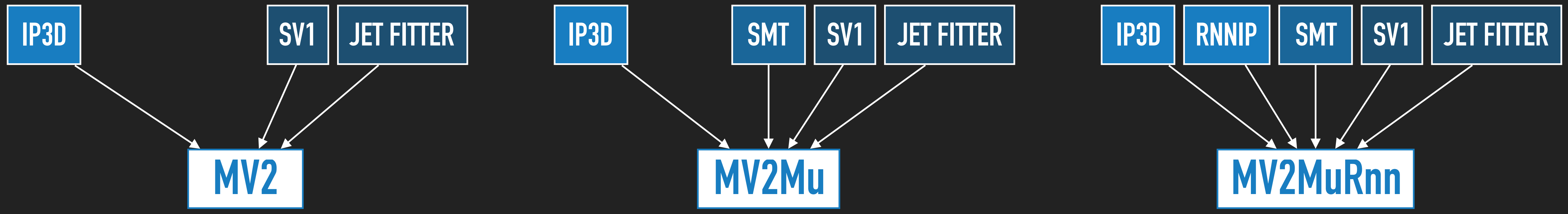
IP3D	RNNIP	SMT	SV1	JET FITTER

ATLAS EXPERIMENT



MV2 GRADIENT BOOSTED DECISION TREE

- Trained with ROOT TMVA
- *b* vs non-*b* — Default non-*b* background: 7% charm and 93% light
- Various versions:



- For *c*-tagging:
  - MV2c100 trained on 100% *c* background
  - MV2cl100 for *c* vs light, no *b*

ATLAS EXPERIMENT

DL1 DEEP NEURAL NETWORK

- Trained with `Keras` (`Theano` backend)
- In ATLAS codebase using `LWTNN`
- Multi-class (*b*, *c*, light)
- Architecture: fully connected + maxout + ReLU + batch norm layers

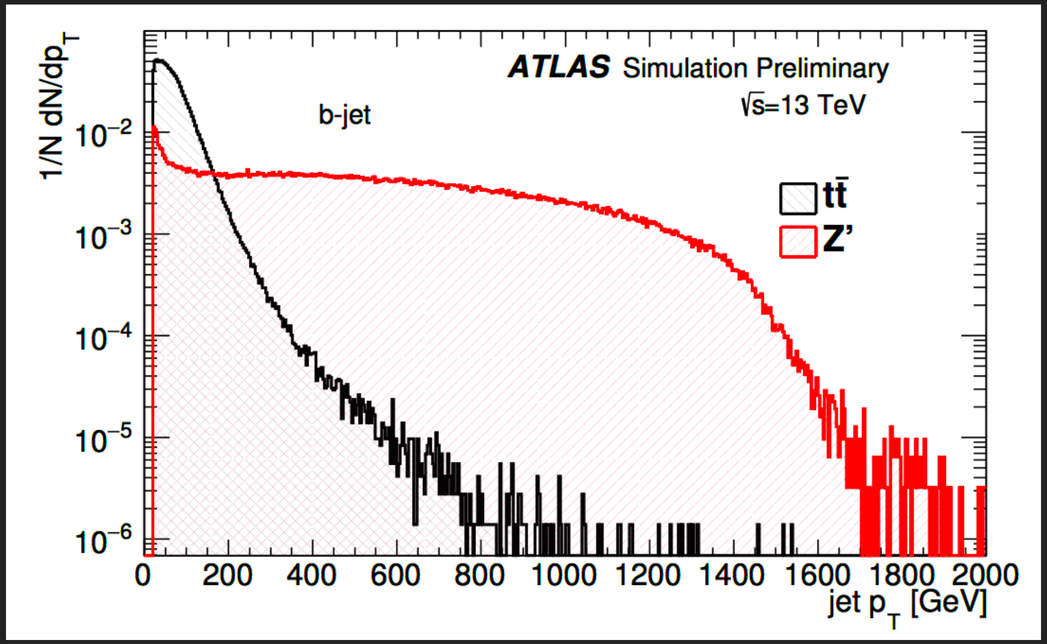
ADVANTAGES

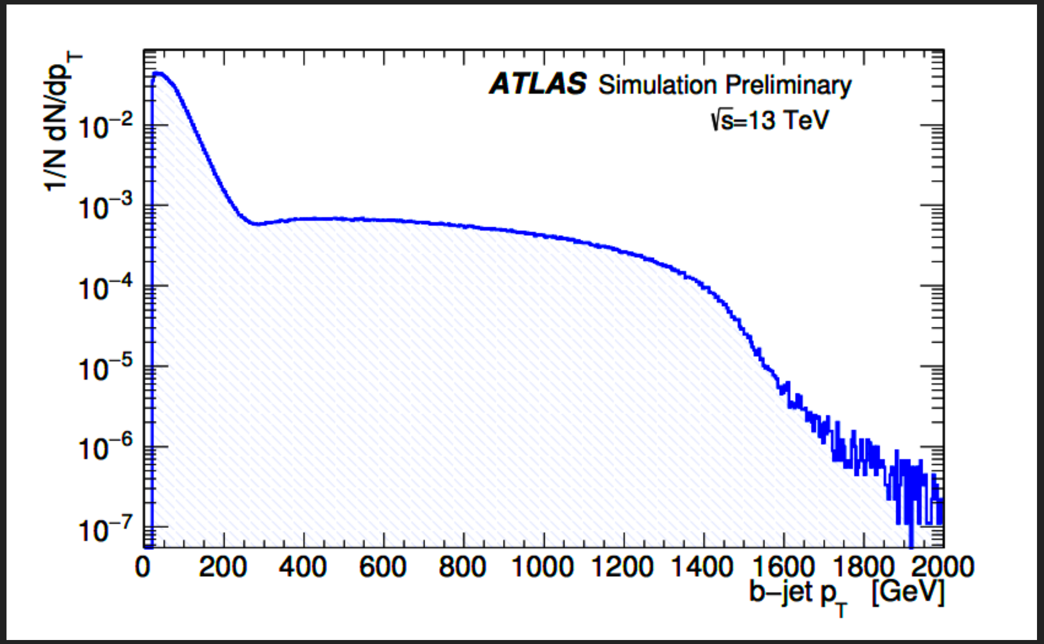
- flexibility in future R&D
- easy to train
- min standalone code
- GPU enabled
- modular
- easy to extend to new input variables
- can be trained adversarially
- can be trained end-to-end with RNNIP

ATLAS EXPERIMENT

HYBRID SAMPLE

- Join ttbar and Z′ samples ~250 GeV to extend kinematic range

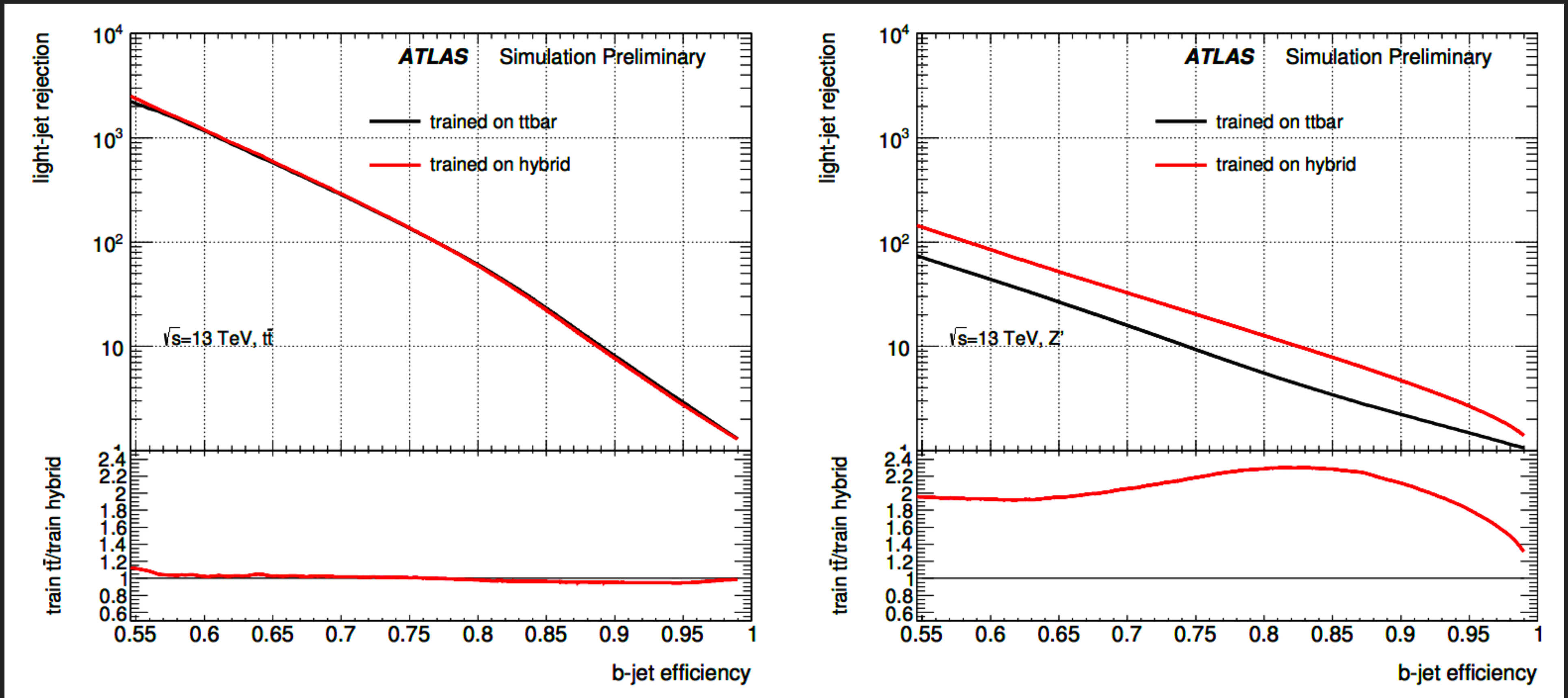




ATLAS EXPERIMENT

HYBRID SAMPLE

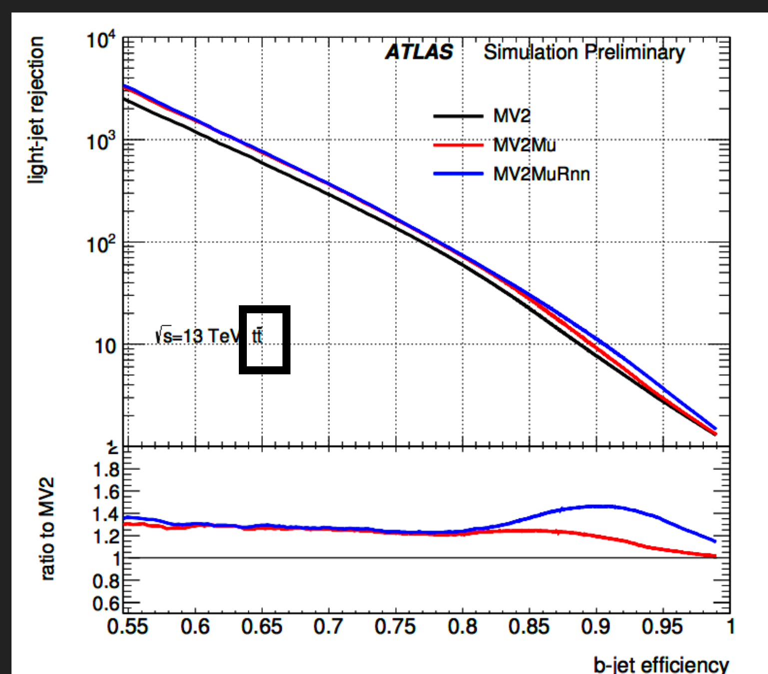
- Improves MV2 performance at high p_T with no performance degradation for ttbar

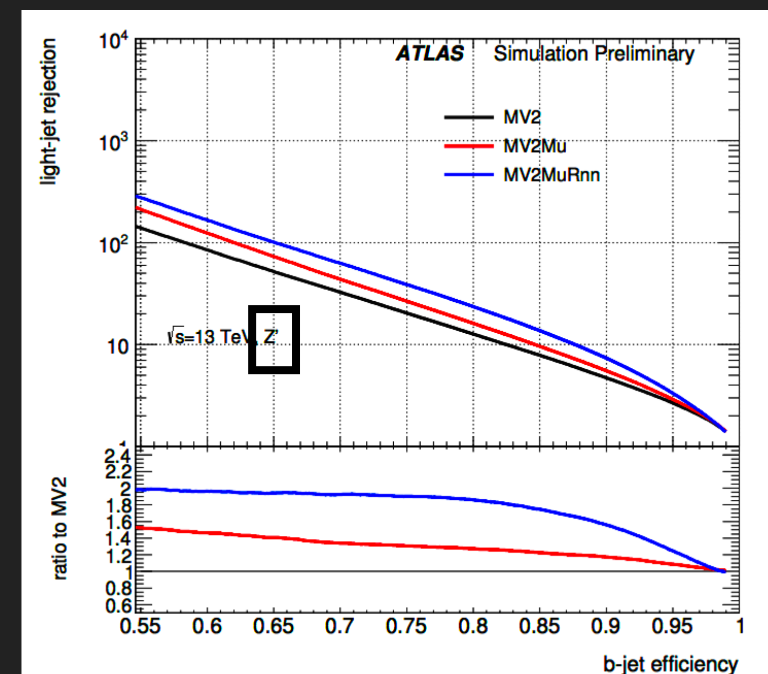


ATLAS EXPERIMENT

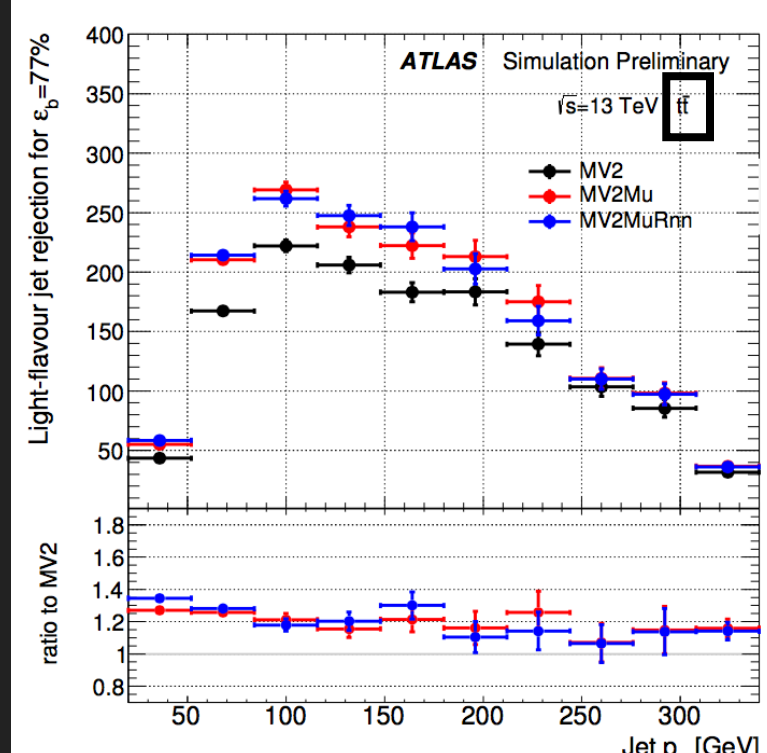
MV2 VARIANTS EVALUATION – B VS LIGHT

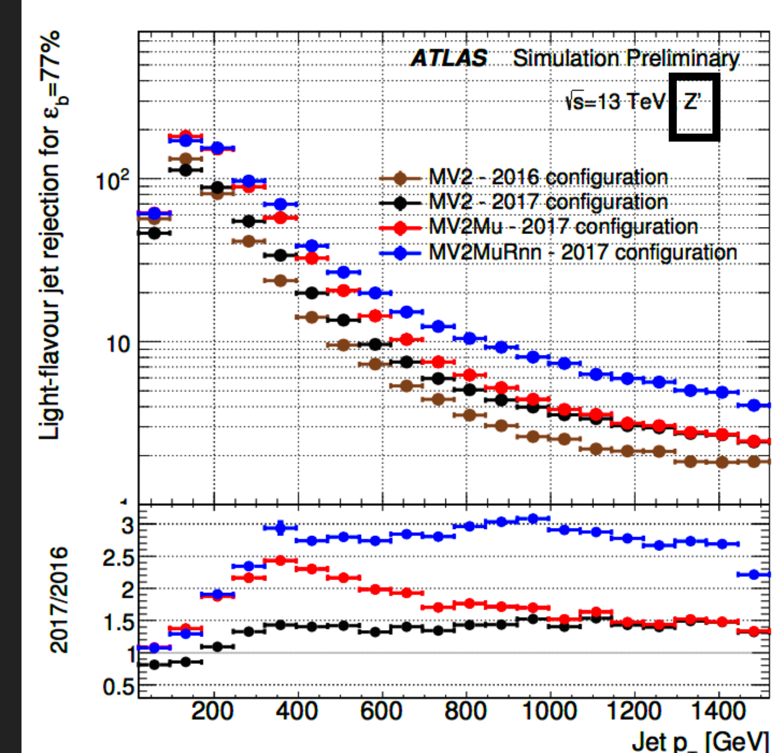
ROC Curves





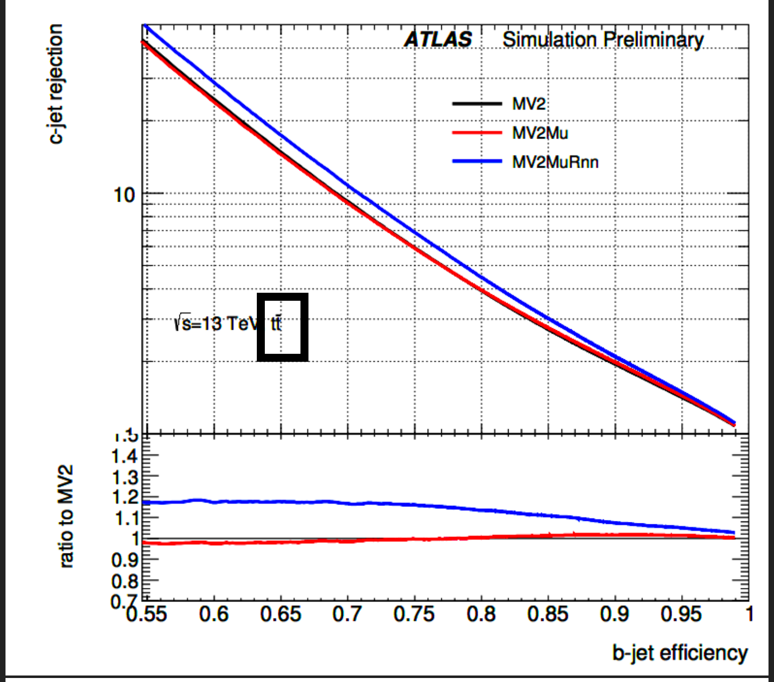
Performance in bins of p_T

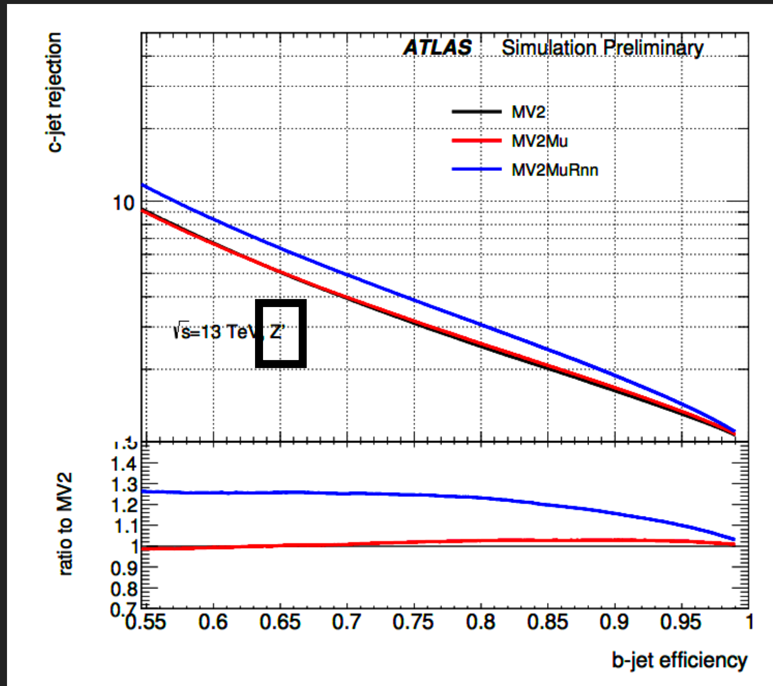




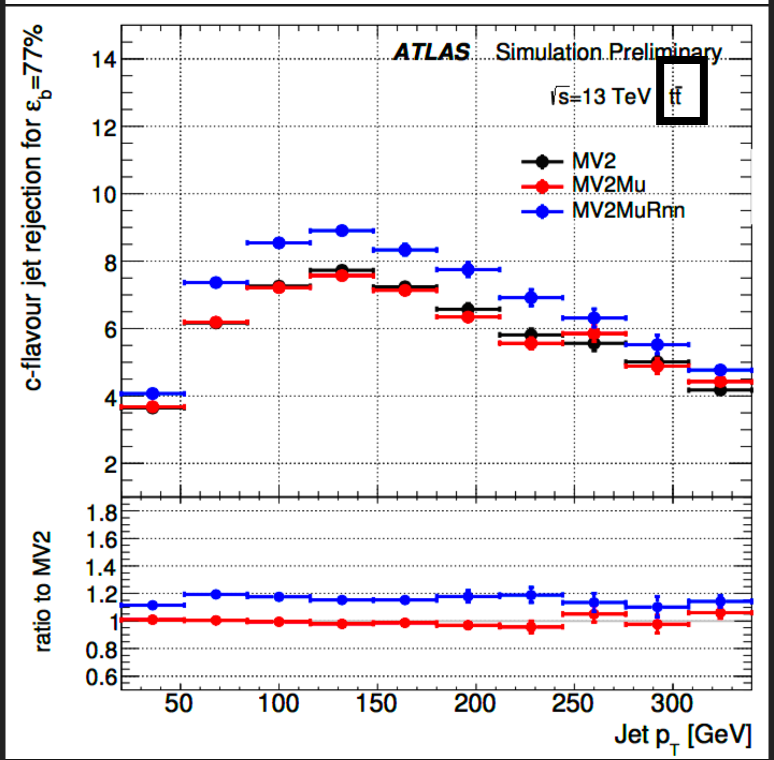
MV2 VARIANTS EVALUATION – B VS C

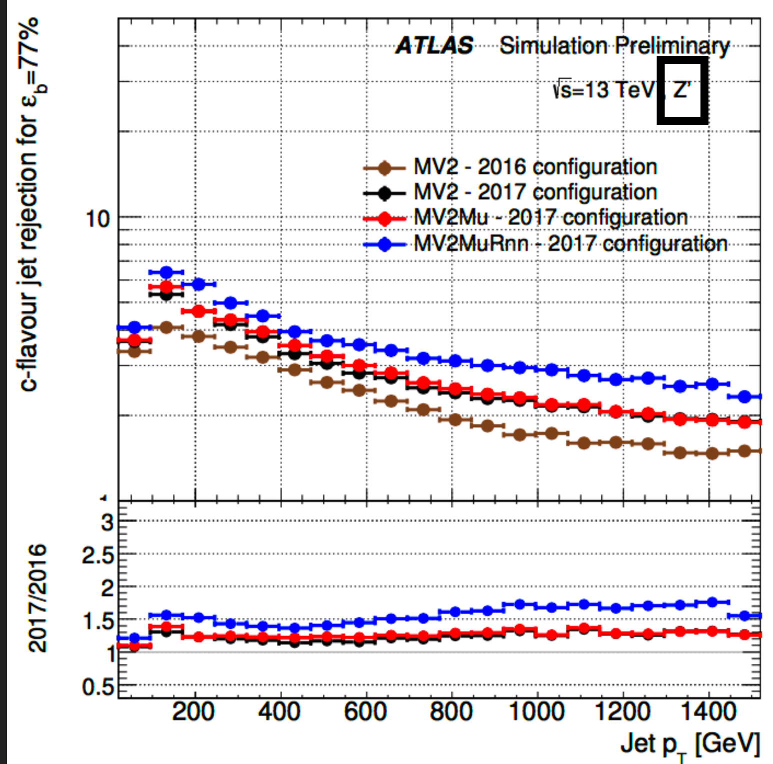
ROC Curves





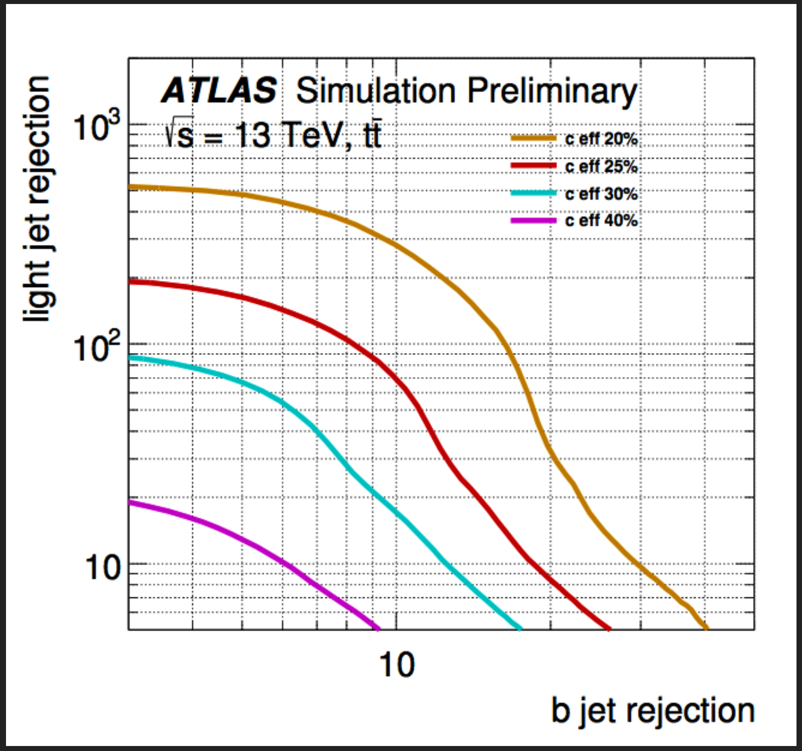
Performance in bins of p_T





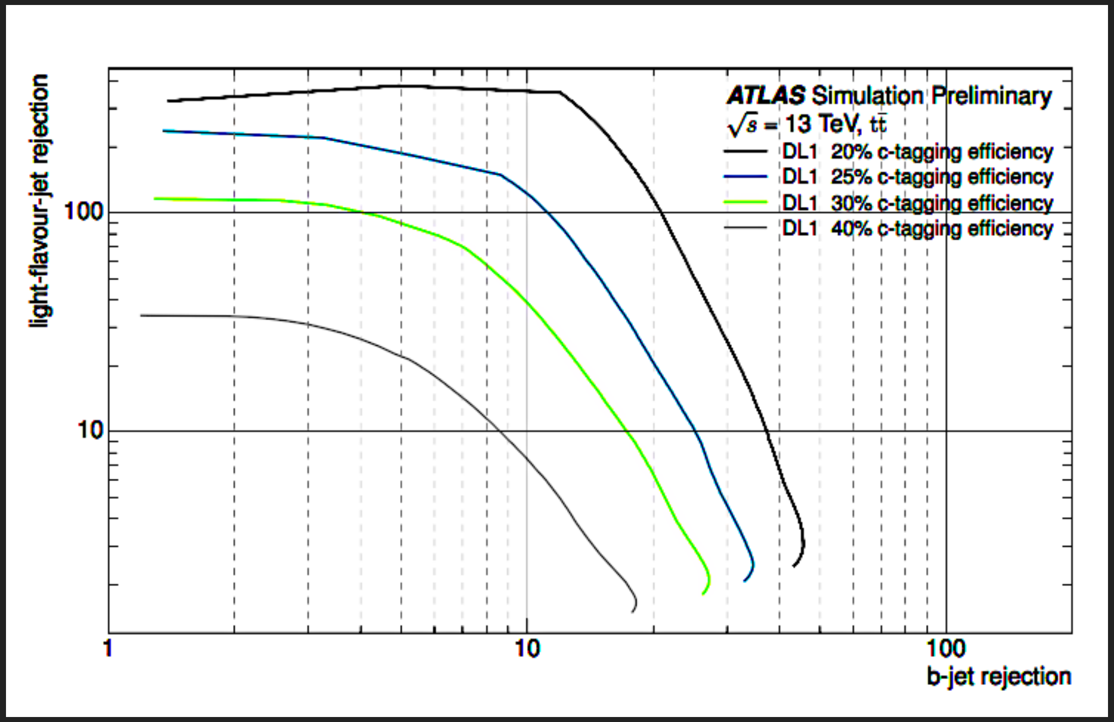
c-EFFICIENCY ISO-CURVES

- When multi-label tagging is enabled, can look at tagging rejections trade-off at constant c-efficiency

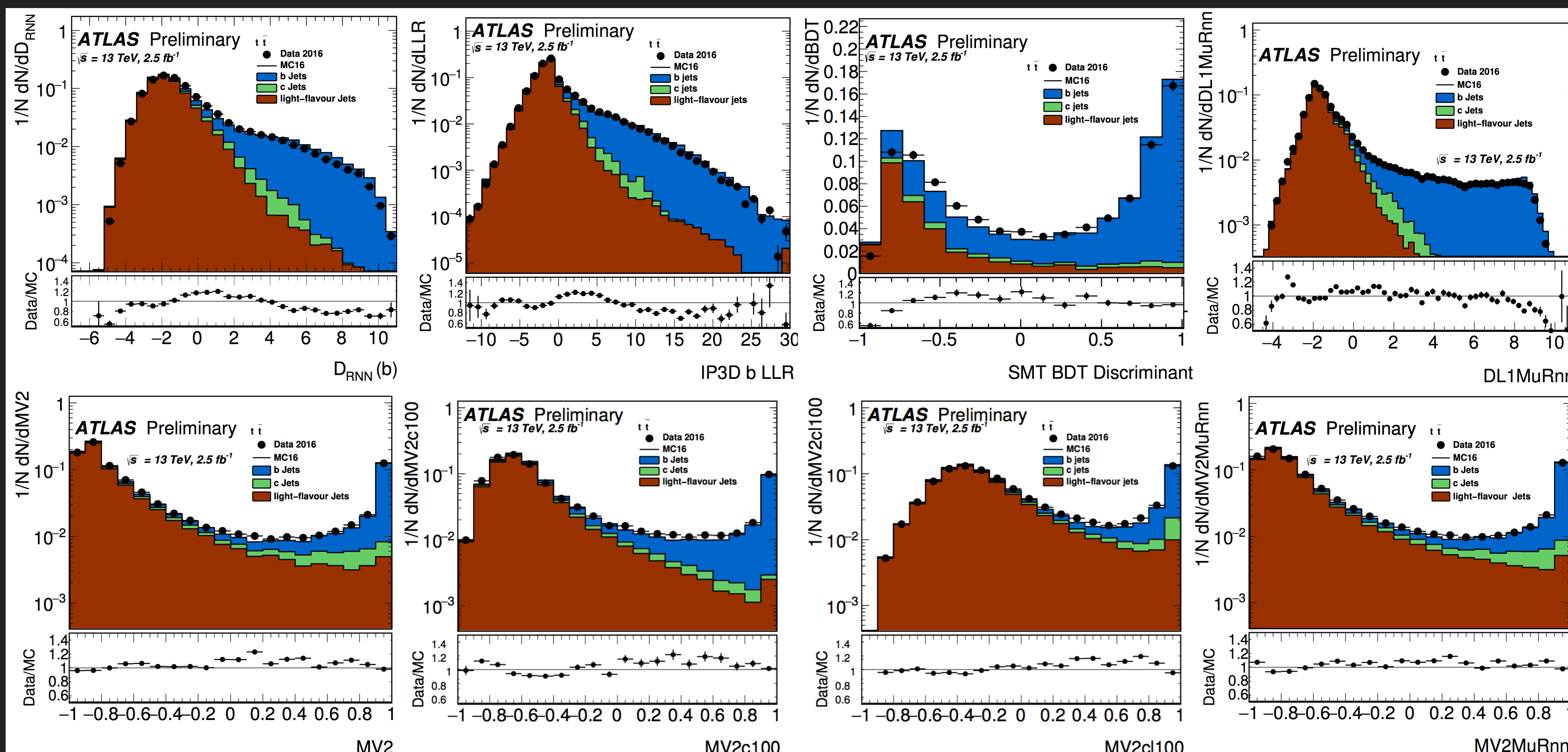


MV2c100 & MV2cl100

DL1



IMPROVED DATA-MC AGREEMENT



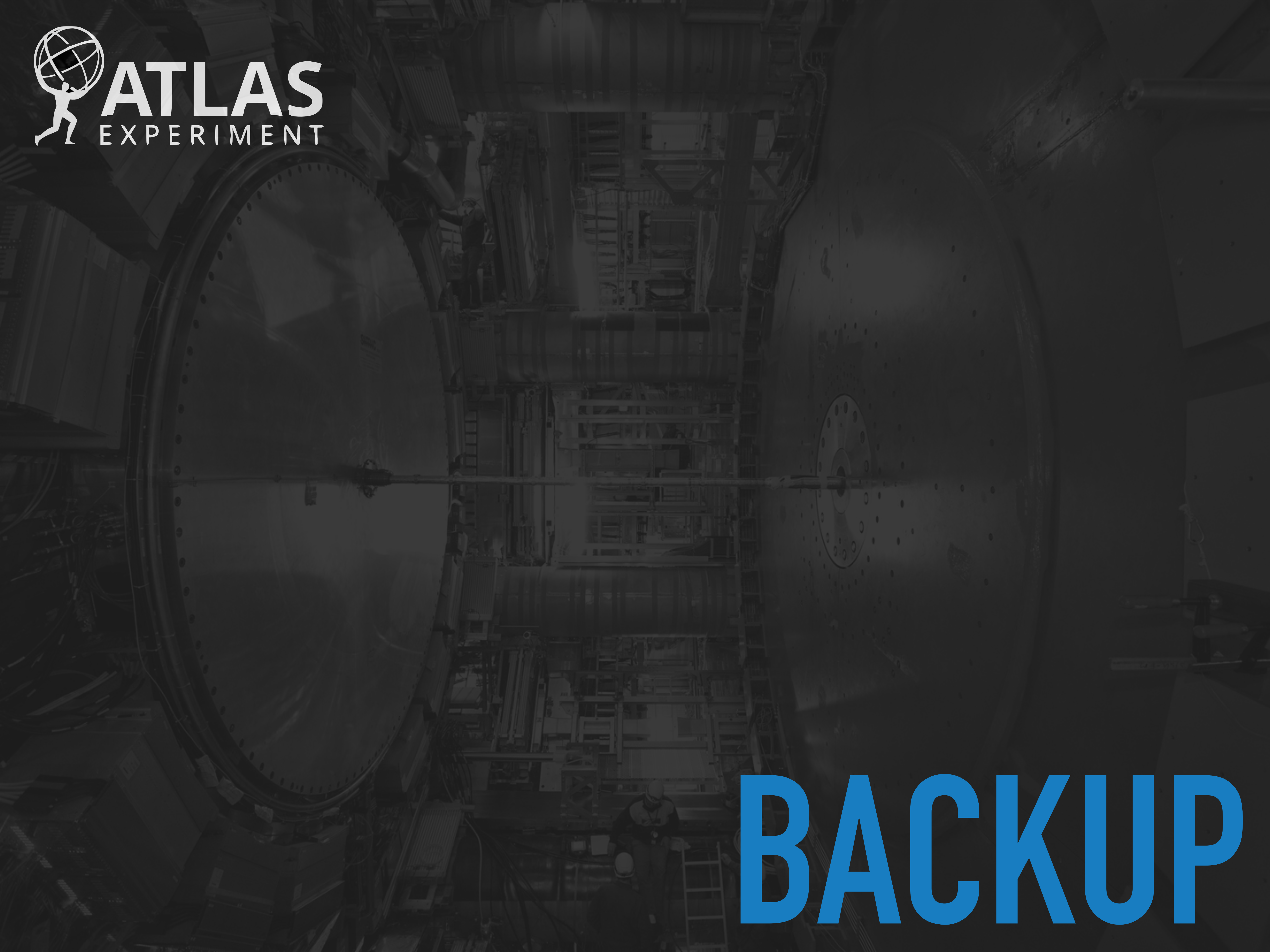
- Due to improvement in tracking simulation
- Minor local discrepancies

WHAT TO EXPECT

- Better data - Monte Carlo agreement
- More performant flavor tagging, due to:
  - availability of new hybrid training sample to extend p_T range
  - improvements and innovations in low level taggers, such as **RNNIP▪** and SMT
  - improvements and innovations in high level taggers, such as **DL1▪**

▪ = deep learning taggers

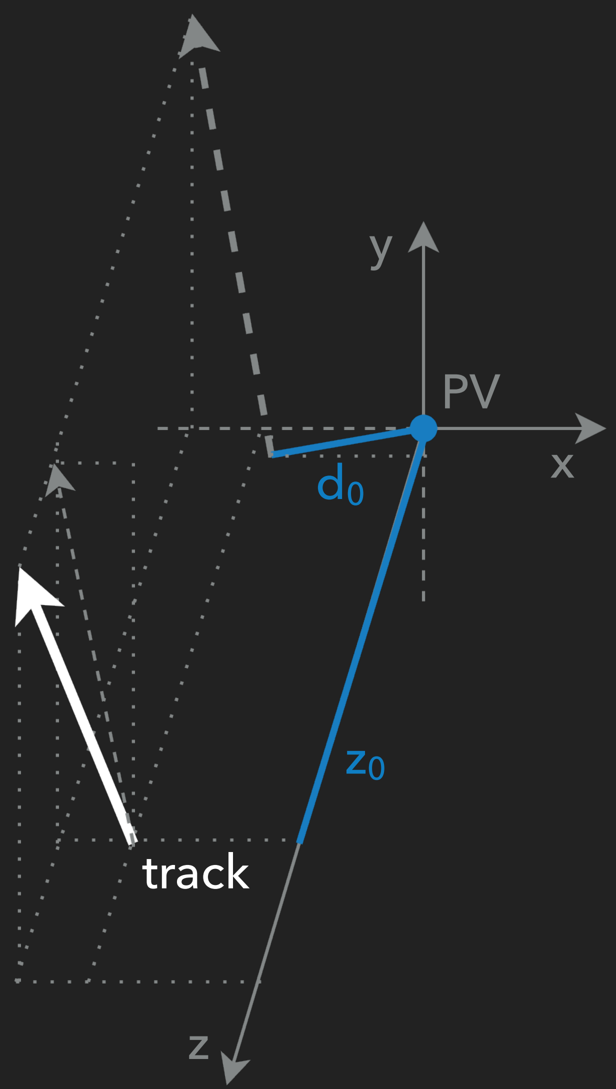
ATLAS EXPERIMENT



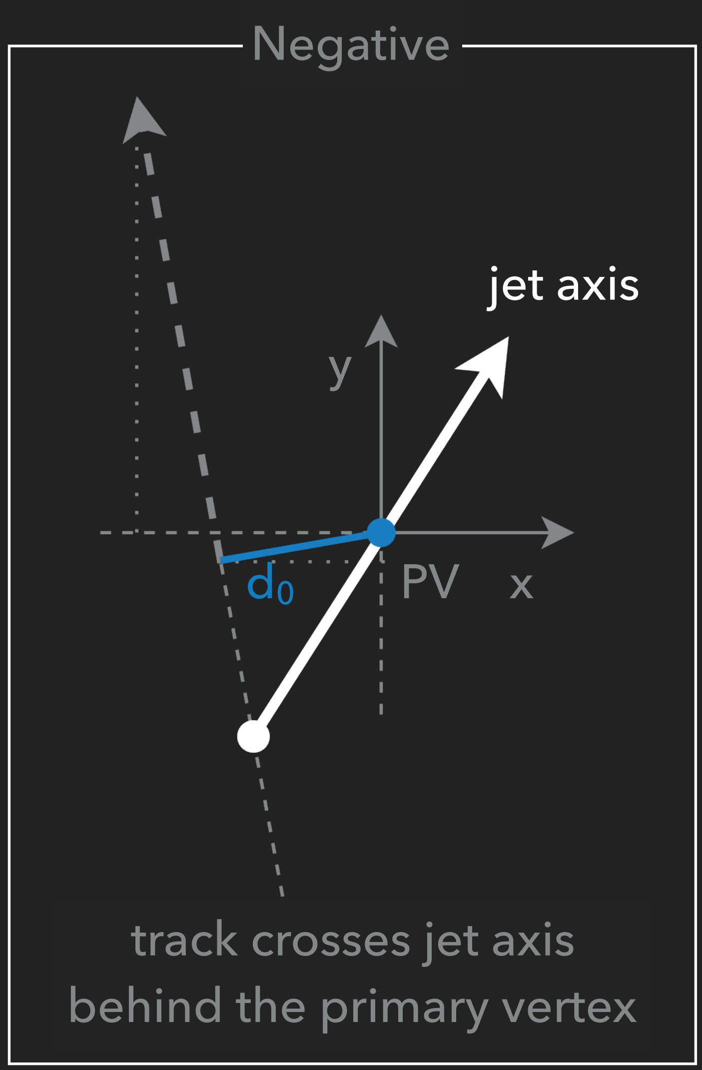
ATLAS EXPERIMENT

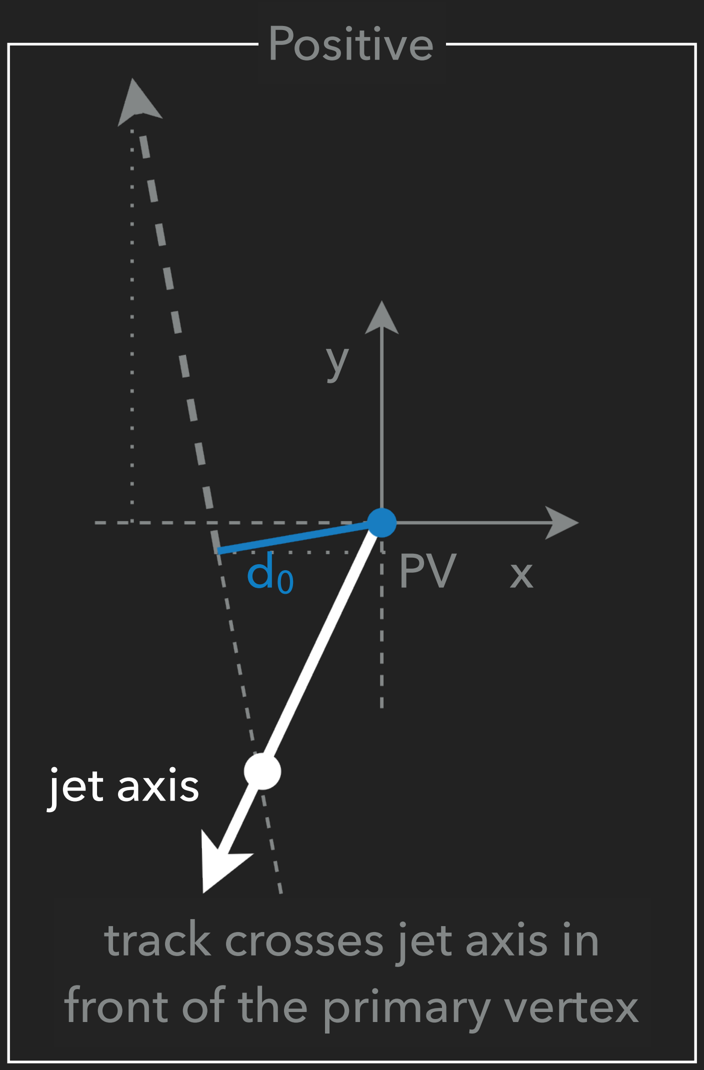
BACKUP

IMPACT PARAMETER DEFINITION



- Sign:





ATLAS EXPERIMENT

LSTM & GRU

- Mitigate issues with **exploding and vanishing gradients**
- Improve knowledge persistence of long-term dependencies
- Internal gating mechanisms to read, write, reset memory
- Classical RNN:

$$\begin{cases} \boldsymbol{s}_t = \boldsymbol{W}_{\mathrm{rec}}\phi(\boldsymbol{s}_{t-1}) + \boldsymbol{W}_x\boldsymbol{x}_t \\ \boldsymbol{y}_t = \boldsymbol{W}_y\boldsymbol{s}_t \end{cases}$$

- Train by optimizing objective function: $\mathcal{L}$

EXPLODING & VANISHING GRADIENTS

$$\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{W}_{\text{rec}}} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{y}_t} \frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{s}_t} \boxed{\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{s}_k}} \frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{W}_{\text{rec}}}$$

Slide inspired by N. de Freitas

EXPLODING & VANISHING GRADIENTS

$$\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{W}_{\text{rec}}} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{y}_t} \frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{s}_t} \boxed{\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{s}_k}} \frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{W}_{\text{rec}}}$$

product of Jacobians

$$\prod_{i=k+1}^{t} \frac{\partial \boldsymbol{s}_i}{\partial \boldsymbol{s}_{i-1}} = \prod_{i=k+1}^{t} \boldsymbol{W}_{\text{rec}}^{T} \text{diag}[\phi'(\boldsymbol{s}_{\boldsymbol{i-1}})]$$

Slide inspired by N. de Freitas

ATLAS EXPERIMENT

EXPLODING & VANISHING GRADIENTS

$$\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{W}_{\text{rec}}} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{y}_t} \frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{s}_t} \boxed{\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{s}_k}} \frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{W}_{\text{rec}}}$$

product of Jacobians
$$\prod_{i=k+1}^{t} \boxed{\frac{\partial \boldsymbol{s}_i}{\partial \boldsymbol{s}_{i-1}}} = \prod_{i=k+1}^{t} \boldsymbol{W}_{\text{rec}}^{T} \text{diag}[\phi'(\boldsymbol{s}_{\boldsymbol{i-1}})]$$

norm is bounded above
$$\left|\left|\frac{\partial \boldsymbol{s}_i}{\partial \boldsymbol{s}_{i-1}}\right|\right| \leq ||\boldsymbol{W}_{\text{rec}}^{T}|| \, ||\text{diag}[\phi'(\boldsymbol{s}_{i-1})]|| \leq \gamma_w \gamma_\phi$$

Slide inspired by N. de Freitas

ATLAS EXPERIMENT

EXPLODING & VANISHING GRADIENTS

$$\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{W}_{\text{rec}}} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{y}_t} \frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{s}_t} \boxed{\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{s}_k}} \frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{W}_{\text{rec}}}$$

product of Jacobians $$\prod_{i=k+1}^{t} \boxed{\frac{\partial \boldsymbol{s}_i}{\partial \boldsymbol{s}_{i-1}}} = \prod_{i=k+1}^{t} \boldsymbol{W}_{\text{rec}}^T \text{diag}[\phi'(\boldsymbol{s}_{\boldsymbol{i-1}})]$$

norm is bounded above $$\left|\left| \frac{\partial \boldsymbol{s}_i}{\partial \boldsymbol{s}_{i-1}} \right|\right| \leq ||\boldsymbol{W}_{\text{rec}}^T||\, ||\text{diag}[\phi'(\boldsymbol{s}_{i-1})]|| \leq \gamma_w \gamma_\phi$$

$$\boxed{\left|\left| \frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{s}_k} \right|\right| \leq (\gamma_w \gamma_\phi)^{t-k}}$$

for long sequences:
- goes to 0 if arg < 1
- diverges for arg >1

Slide inspired by N. de Freitas