RNN Tau Identification

IN THE

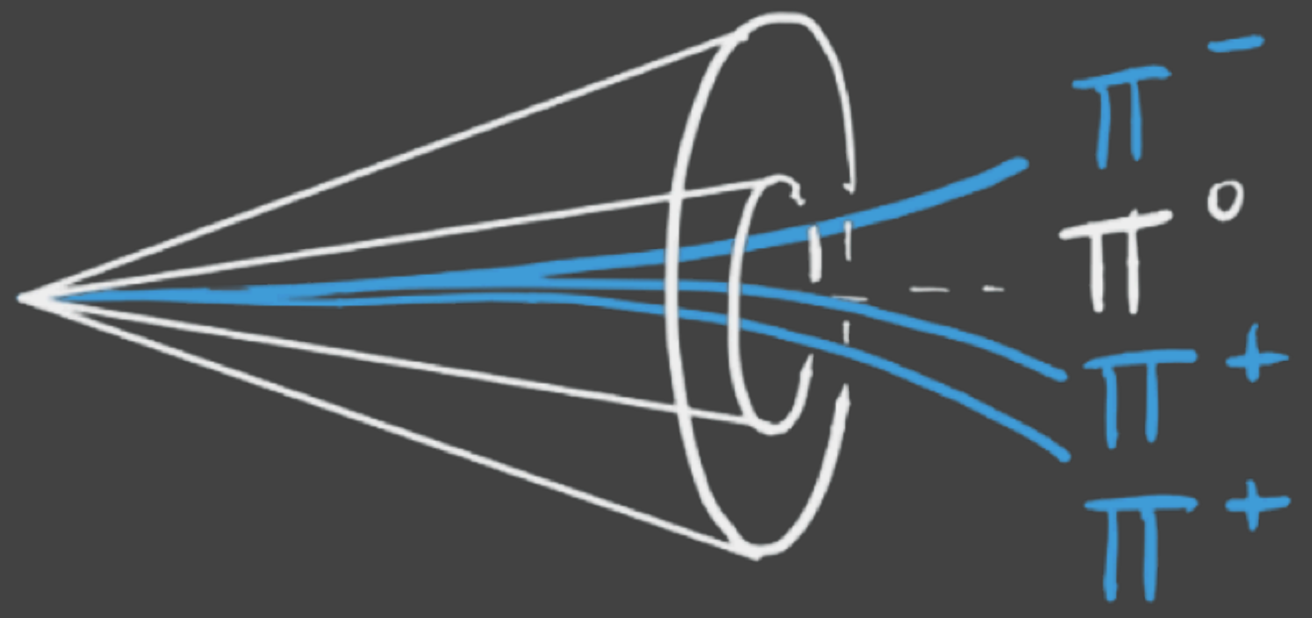
ATLAS High-Level Trigger

Mariel Pettee

October 26th, 2018

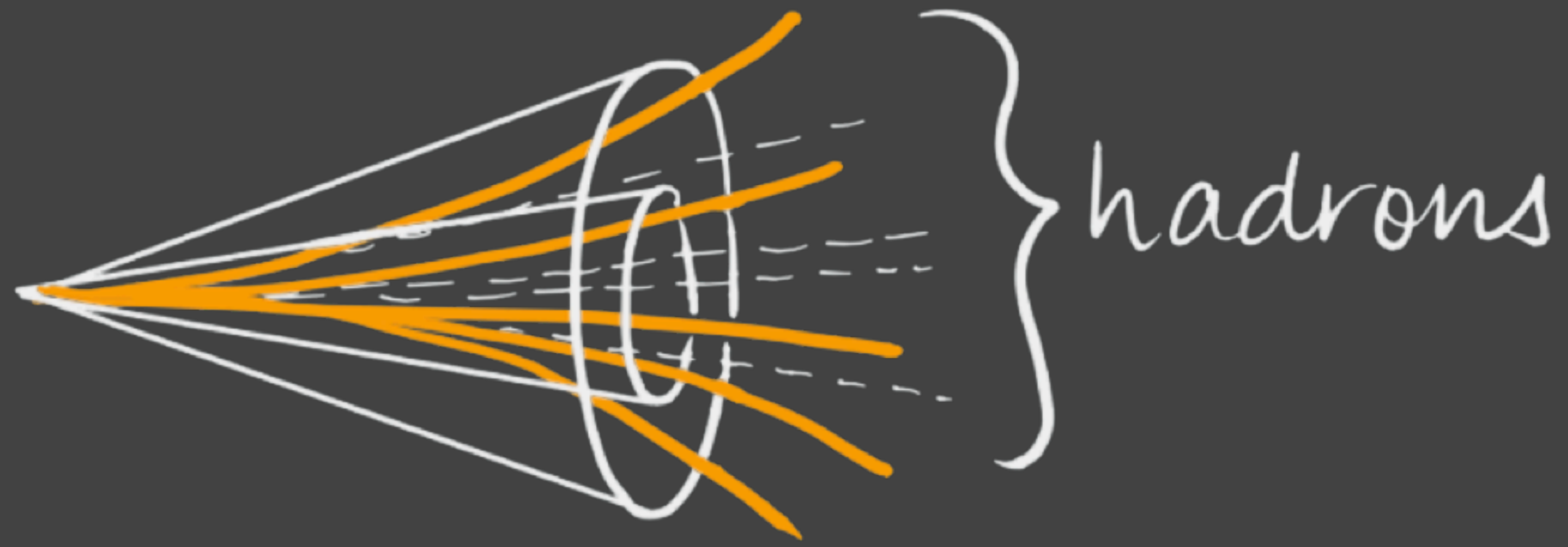
US LUA Annual Meeting

Hadronic Tau Decay:



- Typically 1 or 3 tracks
- Highly collimated
- No surrounding tracks

Jet Background:



- Many tracks
- Less collimated
- Tracks in isolation region

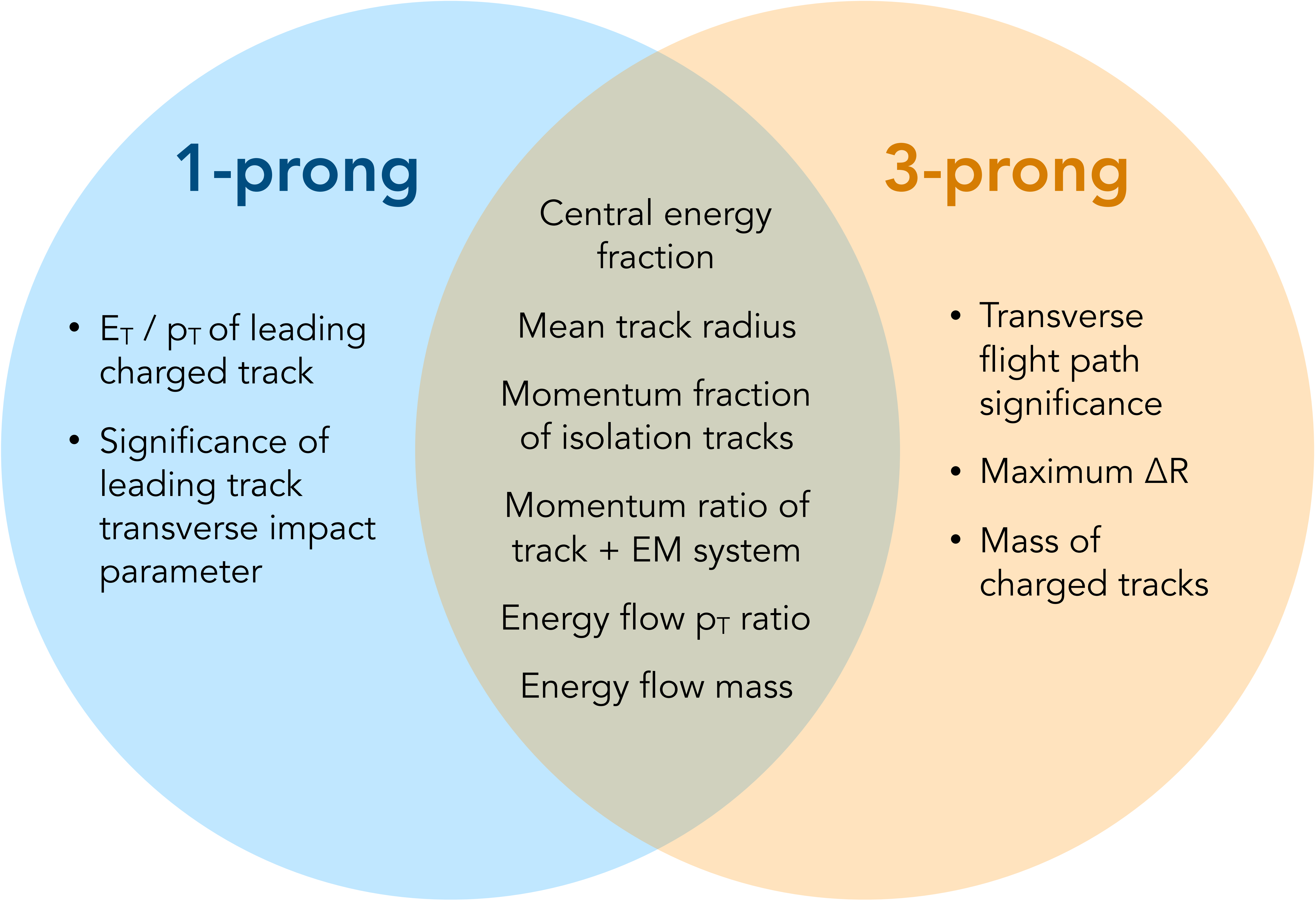
BDT τ Identification

BDT Tau Identification

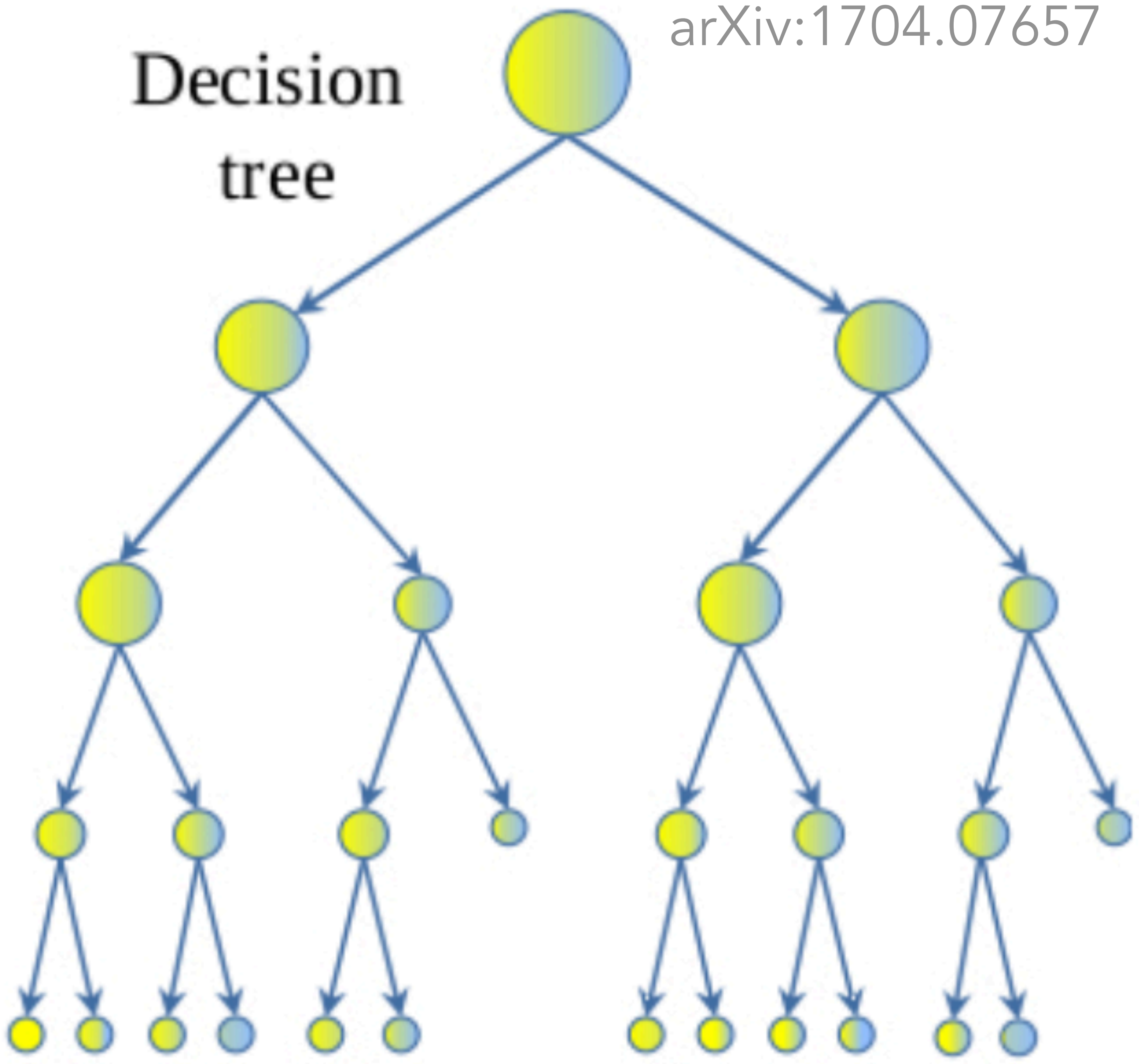
- BDT = Boosted Decision Tree
- **Inputs:** 12 "high-level" tau ID variables
- Separate models for **1-** and **3-prong taus**

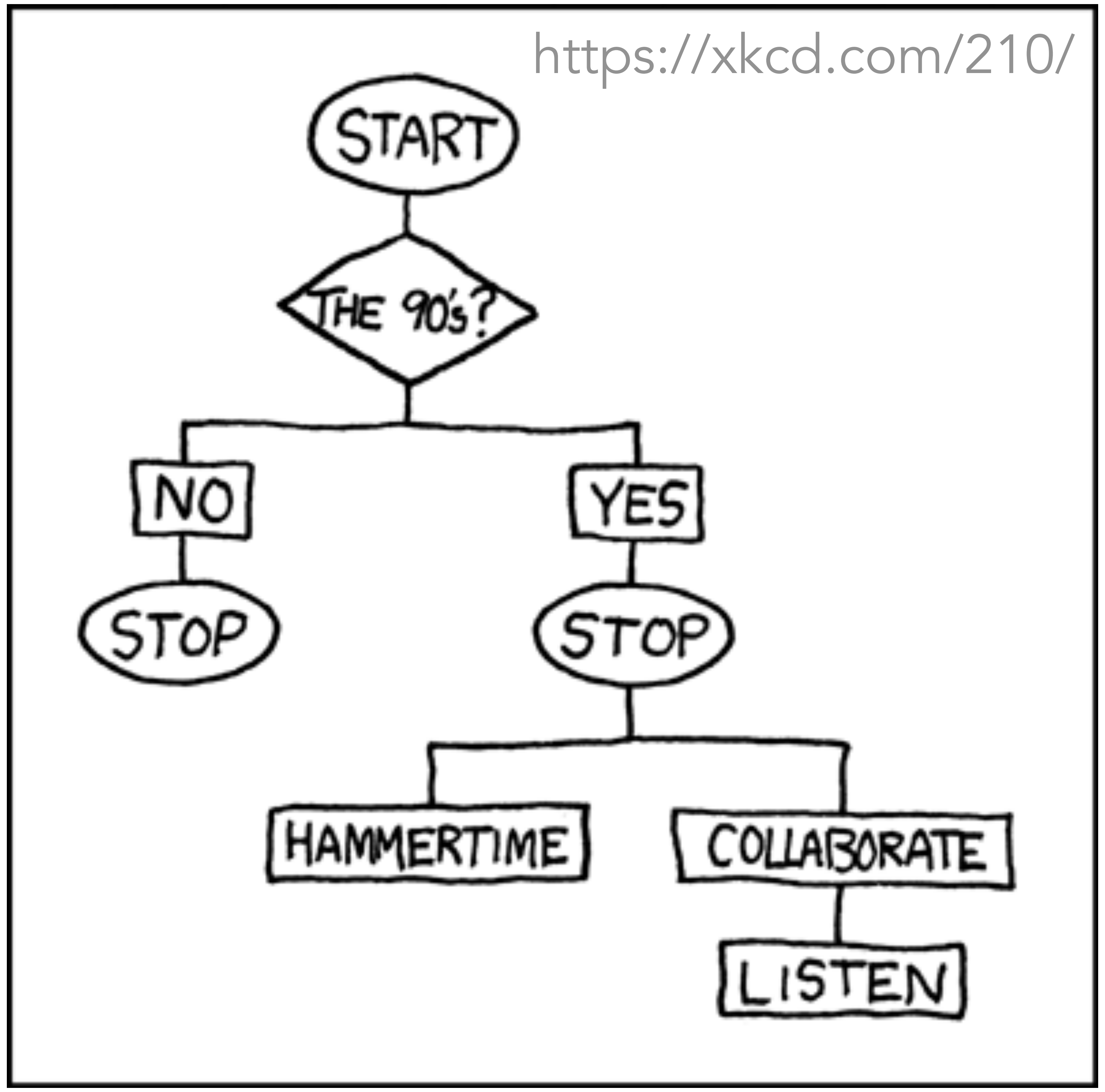
Variable	Description
f_{cent}	Central energy fraction
$f^{-1}_{\text{leadtrack}}$	Leading track momentum fraction
R_{track}	Track radius
$\lvert S_{\text{leadtrack}} \rvert$	Leading track impact parameter significance
$f^{\text{track}}_{\text{iso}}$	Fraction of p_{T} from tracks in the isolation region
ΔR_{Max}	Maximum ΔR
$S^{\text{flight}}_{\text{T}}$	Transverse flight path significance
m_{track}	Track mass
$f^{\text{track-HAD}}_{\text{EM}}$	Fraction of EM energy from charged pions
$f^{\text{EM}}_{\text{track}}$	Ratio of EM energy to track momentum
$m_{\text{EM+track}}$	Track-plus-EM-system mass
$p^{\text{EM+track}}_{\text{T}}/p_{\text{T}}$	Ratio of track-plus-EM-system to p_{T}

Table: ATLAS-CONF-2017-061



BDT Tau Identification





- After training, classify each tau candidate with a single score ∈ (0,1):

0 = background-like 1 = signal-like

BDT Tau Identification

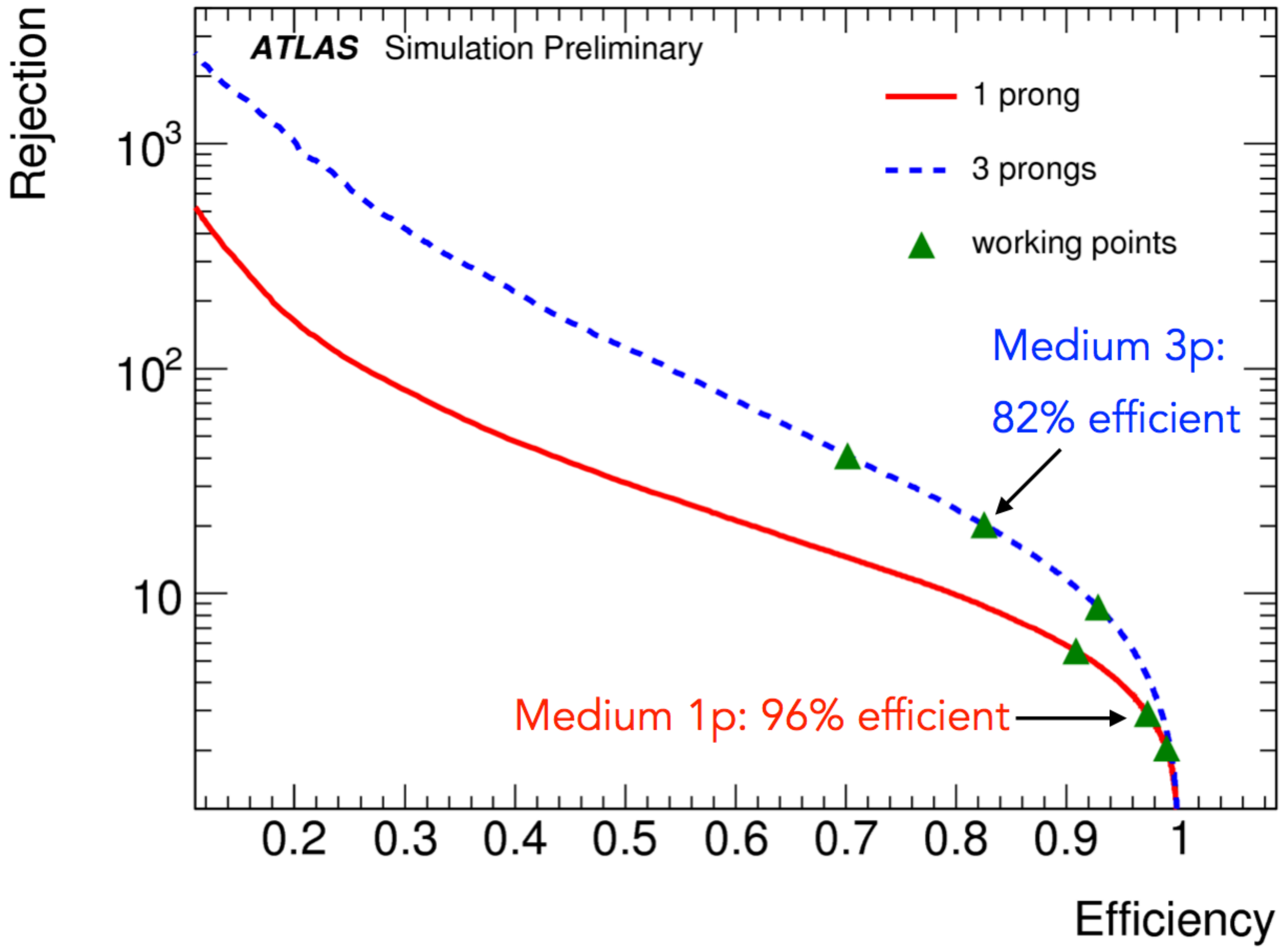
Benefits of the BDT architecture:

- Highly-optimized classification
- Physical intuition & insight

But… **are we throwing away information** by only considering these select high-level variables?

Could adding low-level input variables into the mix improve our signal/background discrimination?

BDT Tau ID Performance in the HLT (2017)



Plot: ATLAS-CONF-2017-061

RNN τ Identification

RNN Tau Identification

- RNN = Recurrent Neural Network
- Approaches τ identification as a **sequence classification** problem
- **Inputs:**
  - **Track-level variables** (for at most 10 tracks; full list in backup slides)
  - **Cluster-level variables** (for at most 6 clusters; full list in backup slides)
  - BDT input variables

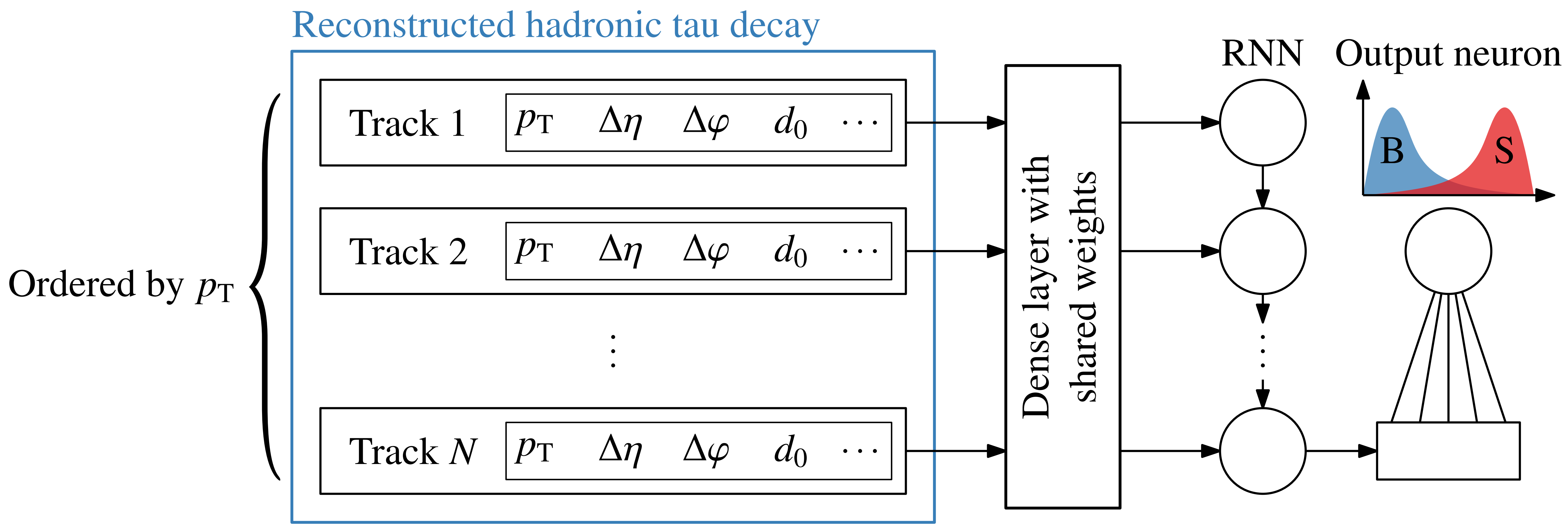
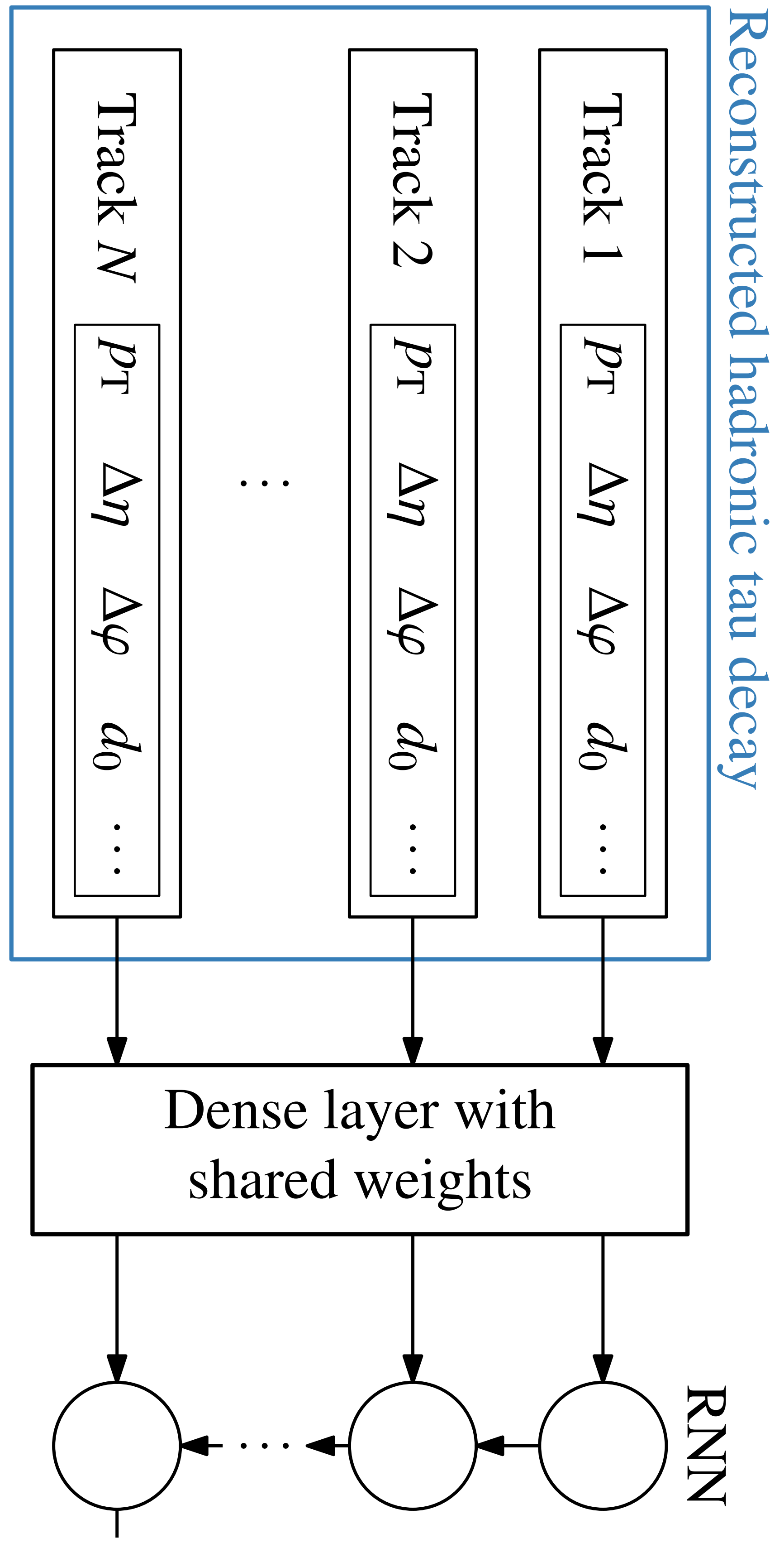


Image: C. Deutsch

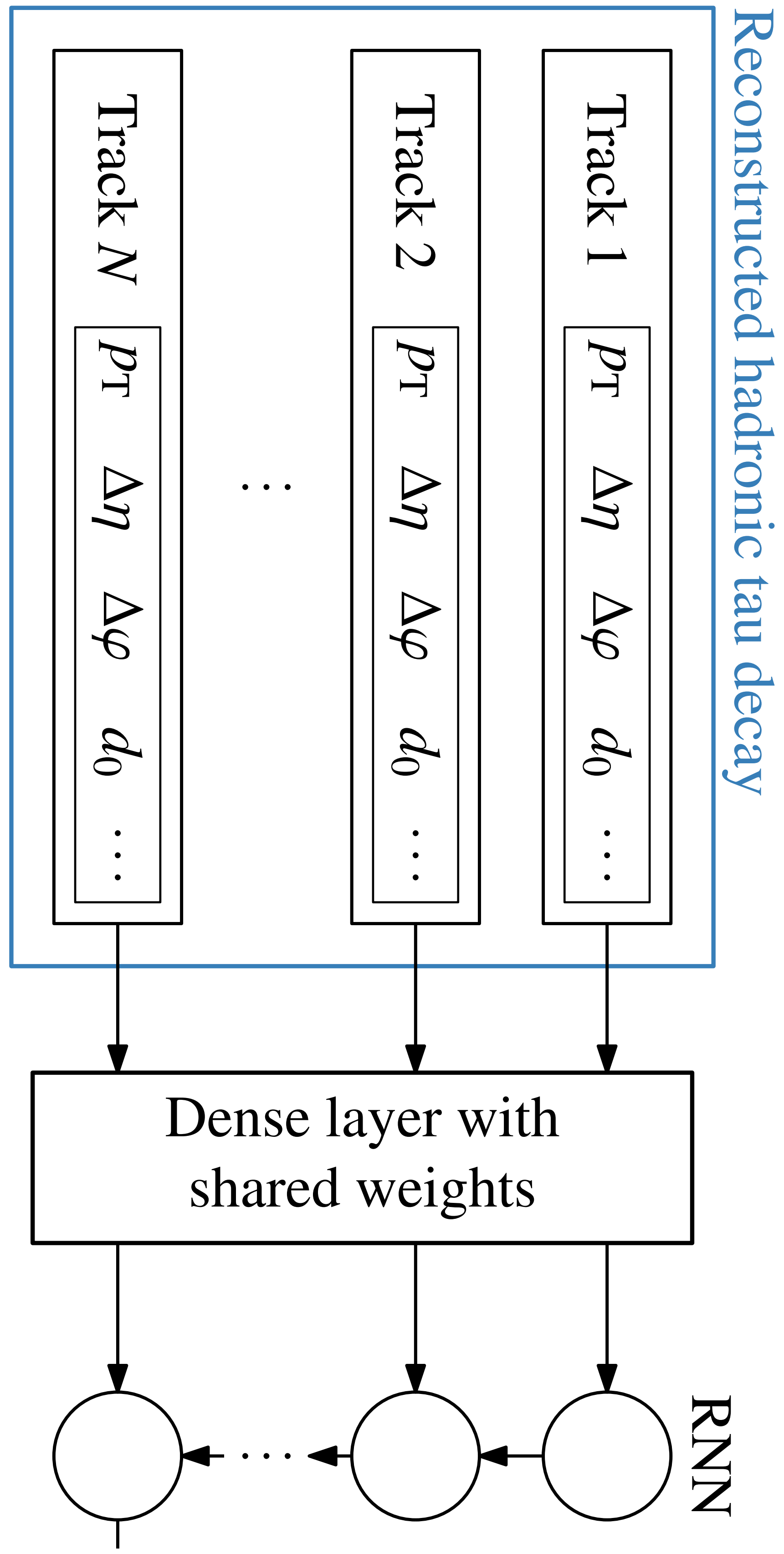
RNN Tau Identification

Tracks



+

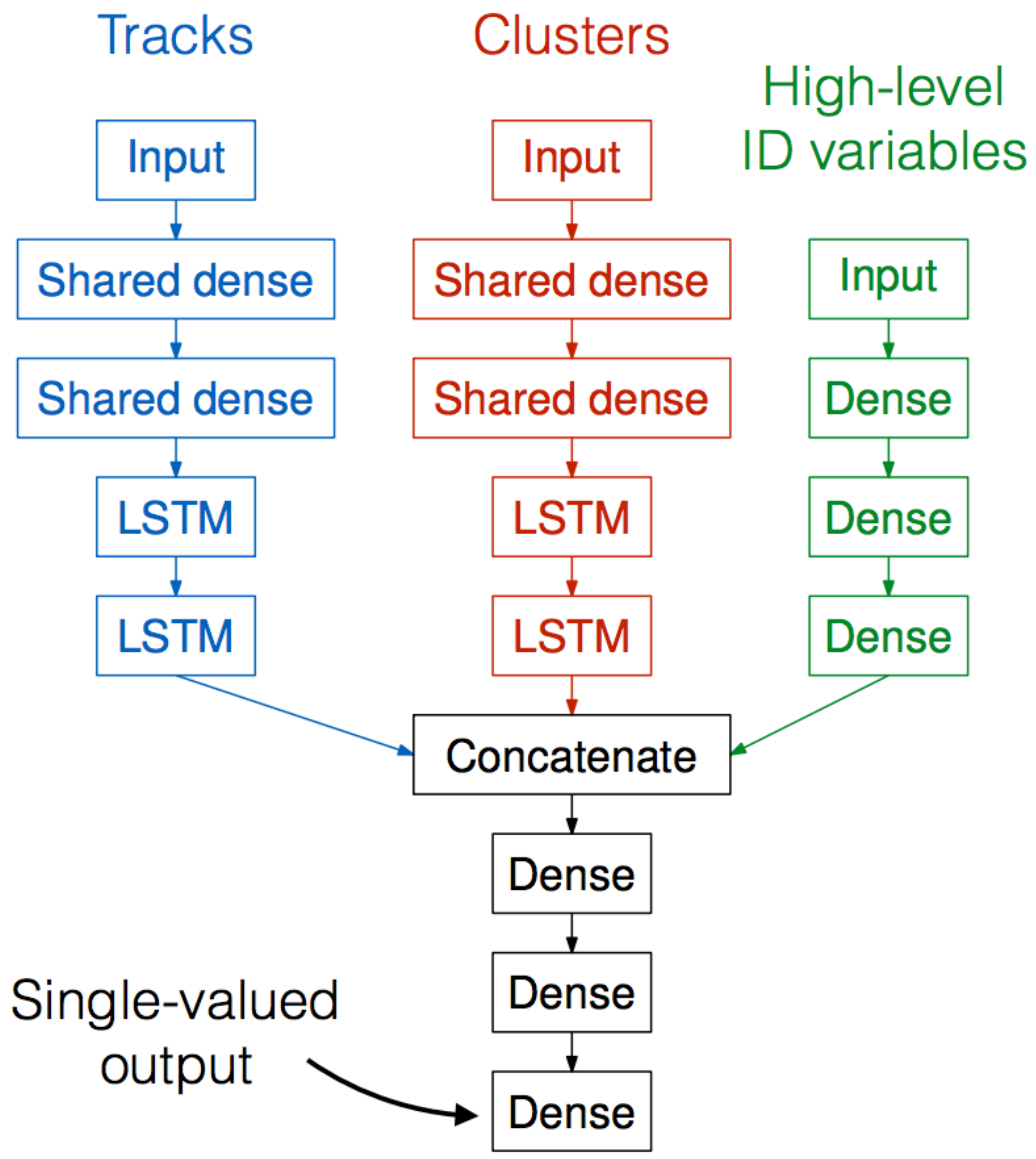
Clusters



+

BDT Input Variables

=



$N_{parameters} \approx 56{,}000$

Tau-ID BDT: 38k

Images: C. Deutsch

RNN Tau Identification

- Shared dense & **LSTM** (Long Short-Term Memory) layers preserve contextual information from multiple tracks/clusters to improve decision-making

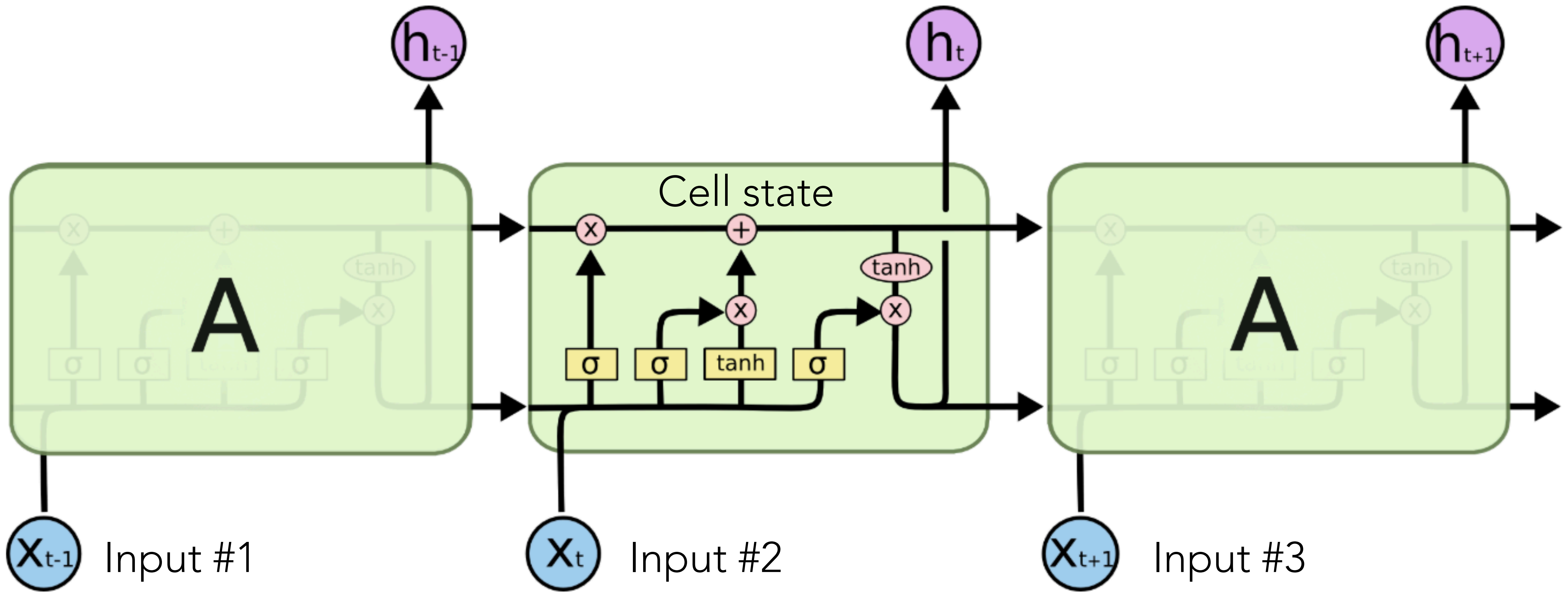
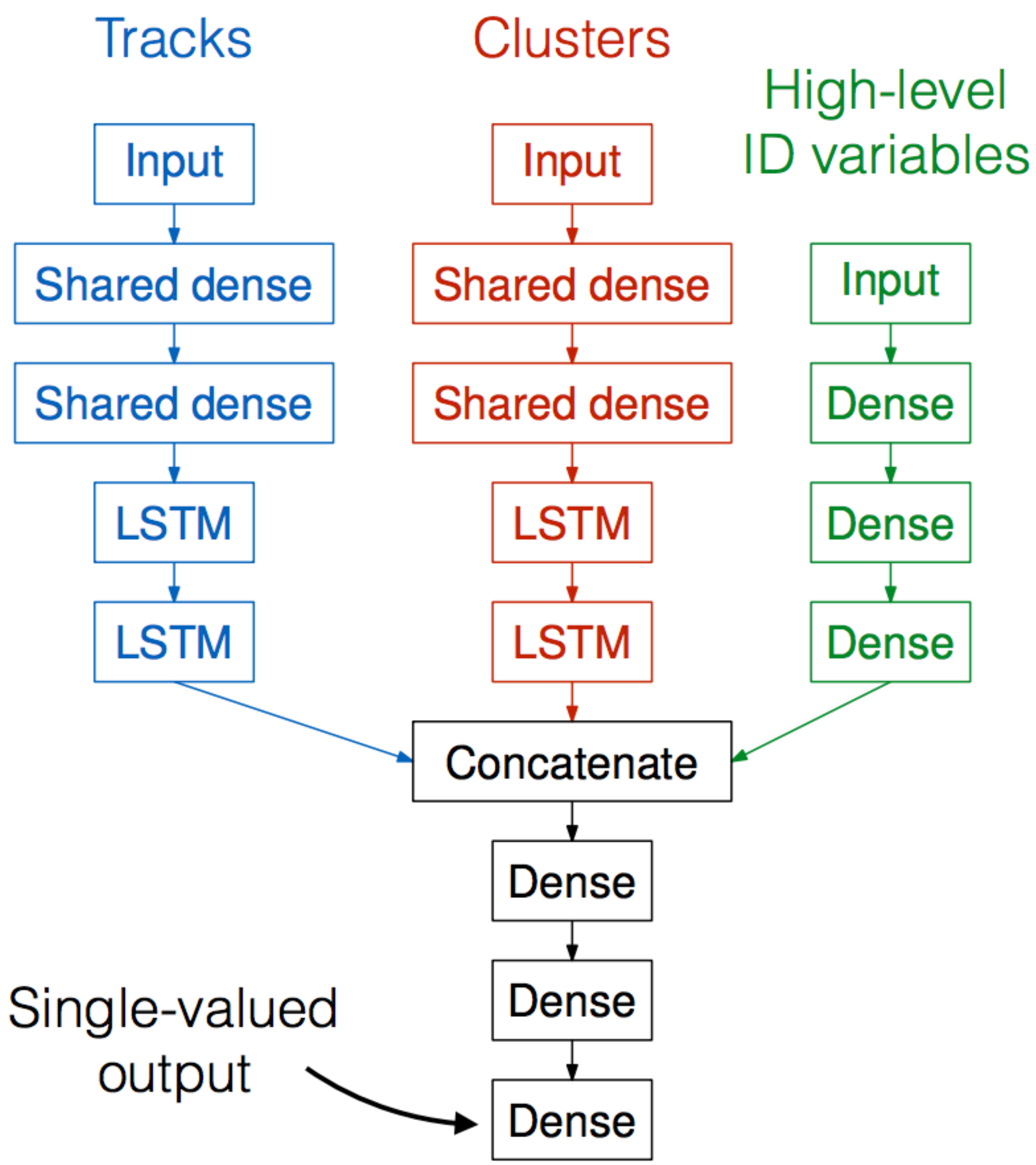


Image: http://colah.github.io/posts/2015-08-Understanding-LSTMs/



$N_{parameters} \approx 56{,}000$

Tau-ID BDT: 38k

Image: C. Deutsch

RNN Tau Identification

- BDT τ identification: **1-prong** and **3-prong taus**
- RNN τ identification: **0-prong**, **1-prong** and **≥2-prong taus**

Goal: recover true 1p taus for which the charged track has been poorly reconstructed, especially at low-p_T and high-μ.

Goal: recover true 3p taus for which at least one charged track has been mis-reconstructed.

Image: C. Deutsch

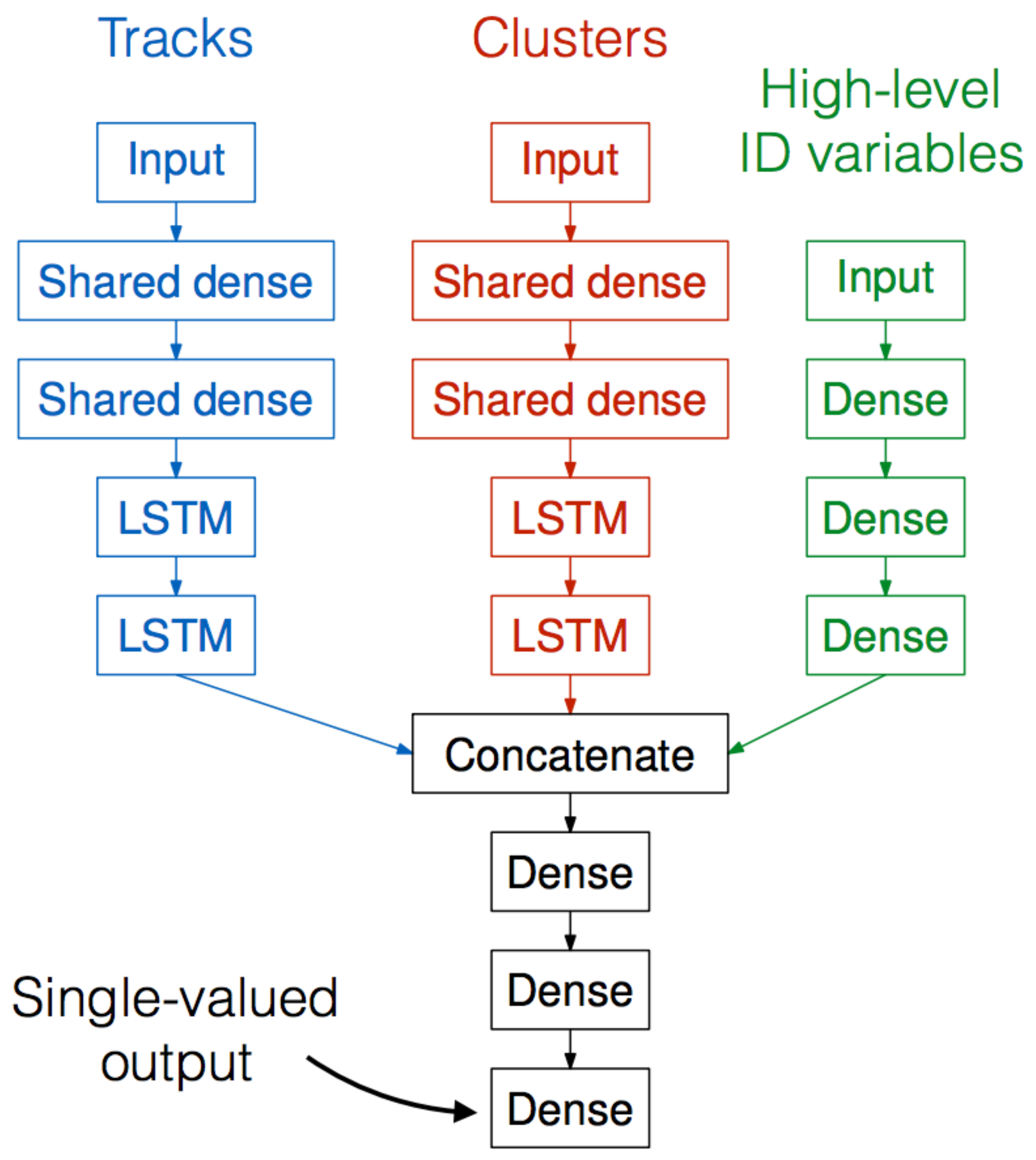
RNN Tau Identification: Online

- **Training samples:**
  - $\gamma^* \rightarrow \tau\tau$ and dijets
  - Fewer statistics than for offline, but same # of trainable parameters
  - **0-prong:** Signal ~100k events, Background ~50k events
  - **1-prong:** Signal ~2M events, Background ~175k events
  - **≥2-prong:** Signal ~700k events, Background ~3M events

- Reweight signal p_T spectrum to match background p_T spectrum for determining same-rejection working points as the BDT (since **trigger rates are dominated by low-p_T jets**).

RNN Tau Identification: Outlook

- RNN tau identification **successfully deployed** in the ATLAS high-level trigger in mid-2018! (Public results forthcoming)
- Will likely be the default ATLAS tau identification algorithm in the **Run 3** trigger
- Until then:
  - Continue optimizing models
  - Investigate RNN input variable modeling
  - Consider training on data rather than MC



$N_{parameters} \approx 56{,}000$

Tau-ID BDT: 38k

Backup Slides

Tau ID RNN Input Variables

of tracker (pixel, SCT) hits → # of tracker hits + # of dead sensors

- Designed to **protect against varying detector conditions**

# of IBL hits → If a hit is expected, use actual # of IBL hits. If not, set # IBL hits = 1.

- If # of IBL hits were set to 0 in the latter case, might be wrongly classified as a bad track

- **Track-level variables**
  - pt_log
  - pt_jetseed_log
  - d0_abs_log
  - z0sinThetaTJVA_abs_log
  - dEta
  - dPhi
  - nIBLHitsAndExp
  - nPixelHitsPlusDeadSensors
  - nSCTHitsPlusDeadSensors
- **Cluster-level variables**
  - et_log
  - pt_jetseed_log
  - dEta
  - dPhi
  - SECOND_R
  - SECOND_LAMBDA
  - CENTER_LAMBDA

Tau ID RNN Input Variables

Central energy fraction (f_{cent}): Transverse energy at EM scale deposited in calorimeter cells with a barycentre in a cone of radius $\Delta R < 0.1$ divided by the transverse energy of cells in a cone of radius $\Delta R < 0.2$ with respect to the reconstructed tau axis. For noise suppression the calorimeter cells must be part of a TopoCluster.

Inverse momentum fraction of the leading track ($f^{-1}_{\text{leadtrack}}$): Transverse energy at EM scale deposited in calorimeter cells (as part of TopoClusters) with a barycentre in a cone of radius $\Delta R < 0.2$ with respect to the tau axis divided by the transverse momentum of the highest transverse momentum track classified as *charged* according to the track classification.

Track radius (R_{track}): Mean ΔR-distance of tracks classified as *charged* and the tau axis weighted by the transverse momentum of each track.

Maximum track ΔR (ΔR_{max}): Maximum ΔR-distance of all tracks classified as *charged* with respect to the tau axis. Equivalent to R_{track} for 1-track $\tau_{\text{had-vis}}$.

Transverse impact parameter significance of the leading track ($|S_{\text{leadtrack}}|$): Absolute value of the transverse impact parameter of the leading *charged* track with respect to the reconstructed primary vertex divided by its uncertainty estimate from the track and vertex fit.

Transverse flight path significance ($S^{\text{flight}}_{\text{T}}$): Distance between the secondary vertex reconstructed using tracks classified as *charged* and primary vertex in the transverse plane divided by the estimated uncertainty from the secondary vertex fit. Defined only for multi-track $\tau_{\text{had-vis}}$.

Momentum fraction of isolation tracks ($f^{\text{track}}_{\text{iso}}$): Sum of transverse momenta of *modified isolation* tracks (cf. Section 3.2) divided by the sum of transverse momenta of *modified isolation* and *charged* tracks.

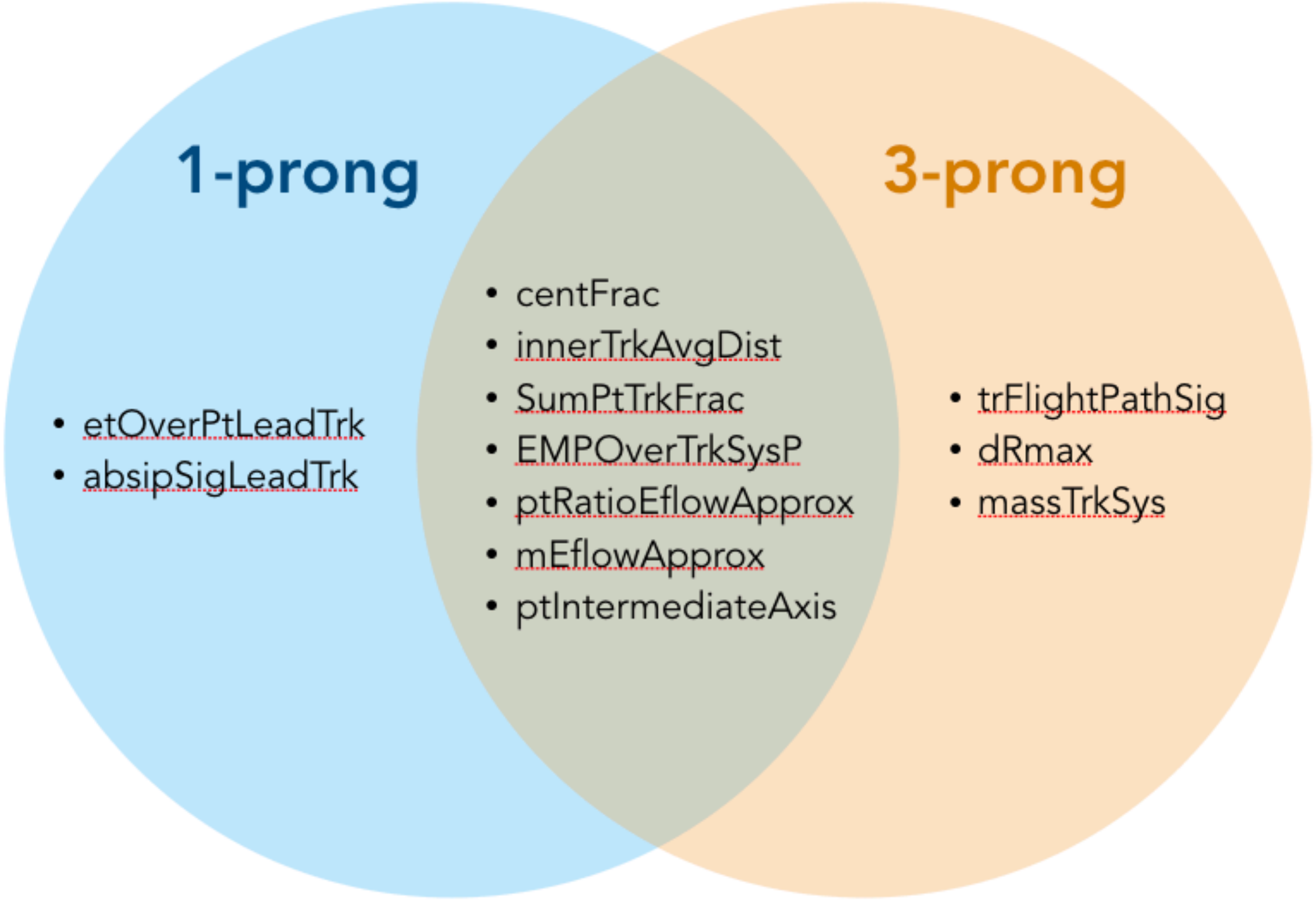
EM energy fraction of charged pions ($f^{\text{track-HAD}}_{\text{EM}}$): Energy deposited by charged pions in the electromagnetic part of the calorimeter estimated by subtracting the energy contained in the hadronic part of the calorimeter (in TopoClusters) from the energy of the track system, consisting of tracks classified as *charged*, estimated by the scalar sum of track momenta (assuming zero mass). This energy is divided by the energy contained in the electromagnetic part of clusters associated with the jet seeding the $\tau_{\text{had-vis}}$. All cluster energies are calibrated at LC scale.

Transverse impact parameter significance of the leading track ($|S_{\text{leadtrack}}|$): Absolute value of the transverse impact parameter of the leading *charged* track with respect to the reconstructed primary vertex divided by its uncertainty estimate from the track and vertex fit.

Transverse flight path significance ($S^{\text{flight}}_{\text{T}}$): Distance between the secondary vertex reconstructed using tracks classified as *charged* and primary vertex in the transverse plane divided by the estimated uncertainty from the secondary vertex fit. Defined only for multi-track $\tau_{\text{had-vis}}$.

Momentum fraction of isolation tracks ($f^{\text{track}}_{\text{iso}}$): Sum of transverse momenta of *modified isolation* tracks (cf. Section 3.2) divided by the sum of transverse momenta of *modified isolation* and *charged* tracks.

EM energy fraction of charged pions ($f^{\text{track-HAD}}_{\text{EM}}$): Energy deposited by charged pions in the electromagnetic part of the calorimeter estimated by subtracting the energy contained in the hadronic part of the calorimeter (in TopoClusters) from the energy of the track system, consisting of tracks classified as *charged*, estimated by the scalar sum of track momenta (assuming zero mass). This energy is divided by the energy contained in the electromagnetic part of clusters associated with the jet seeding the $\tau_{\text{had-vis}}$. All cluster energies are calibrated at LC scale.



Tau HLT in 2018

CALORIMETER CLUSTERING → **INITIAL TRACKING** → **PRECISION TRACKING** → **TAU IDENTIFICATION**

CALORIMETER CLUSTERING

- **Inclusive MVA TES:**

  MVA Tau Energy Scale calibration applied to all tau candidates

- **Minimum p_T cut**

INITIAL TRACKING

- **"Fast Track Finder" (FTF):**

  Look for a track in a narrow cone ($\Delta R < 0.1$) around the center of the cluster & along the full beam line

PRECISION TRACKING

- **Track refit:**

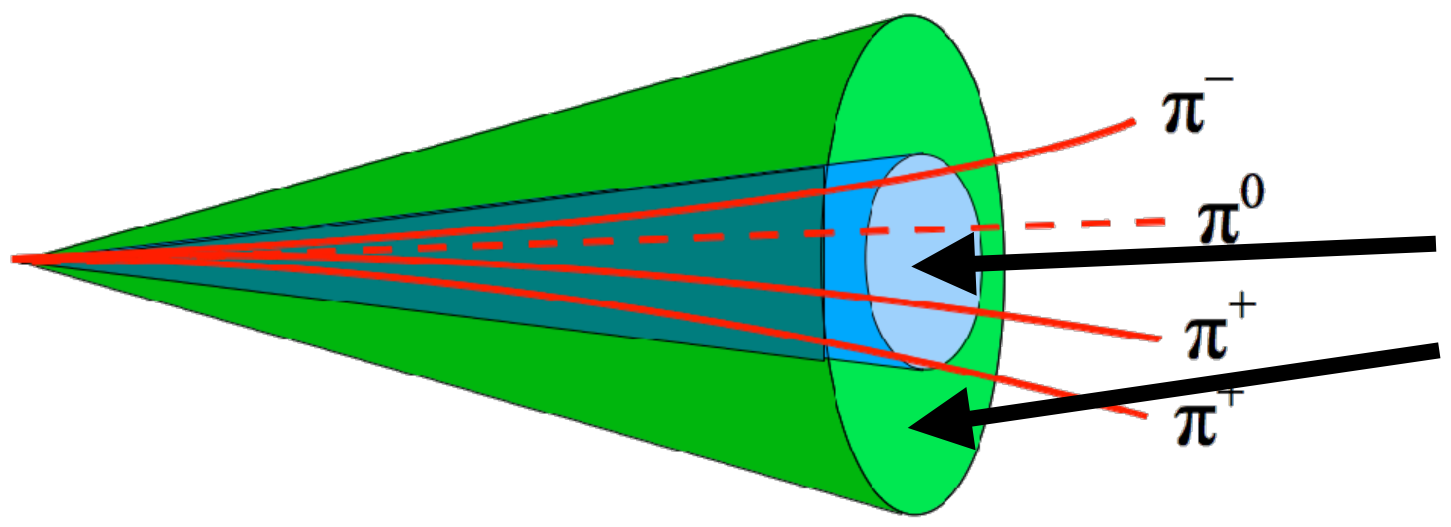
  Run precision tracking on FTF seeds

- **nTracks cut**

TAU IDENTIFICATION

- **RNN**

  0-/1-/multi-prong RNN with tau ID variables + track + cluster variables as inputs



nTracks Cut: only pass taus with…

✅ 1-3 tracks in core region

✅ ≤1 tracks in isolation region