Towards Strange Tagging with ILD
A summary of our progress to-date

ILD Software Meeting – January 13, 2021

Matt Basso (University of Toronto) & Valentina Cairo (SLAC),
On behalf of everyone on the Snowmass 2021 LoI

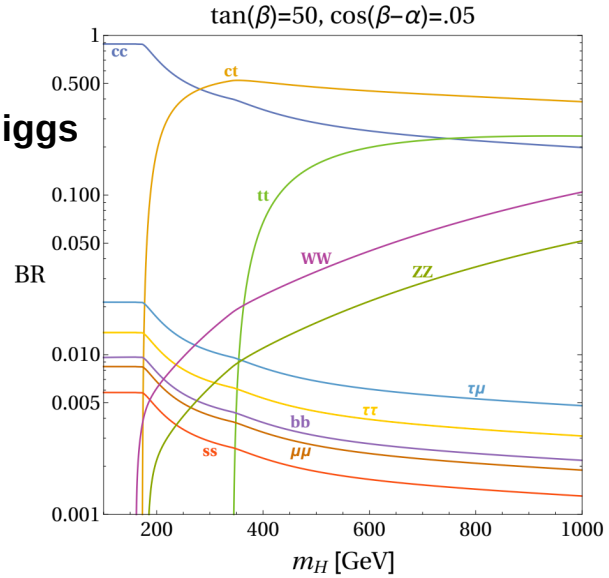
Overview

- LoI: “Strange Quark as a probe for new physics in the Higgs Sector” – in line with ILC Snowmass 2021 study questions (2007.03650)
  - *Basic goal*: develop a strange tagger using ILD and apply the tagger to a simple SM H->ss or BSM H->cs analysis
    - Interplay with the instrumentation: strange tagging capabilities strong depend on the detector (e.g., PID)
  - Collaboration between SLAC, Brown, Oregon, KEK, and Toronto
  - Two working meetings since August:
    - September 24th, 2020
    - November 24th, 2020

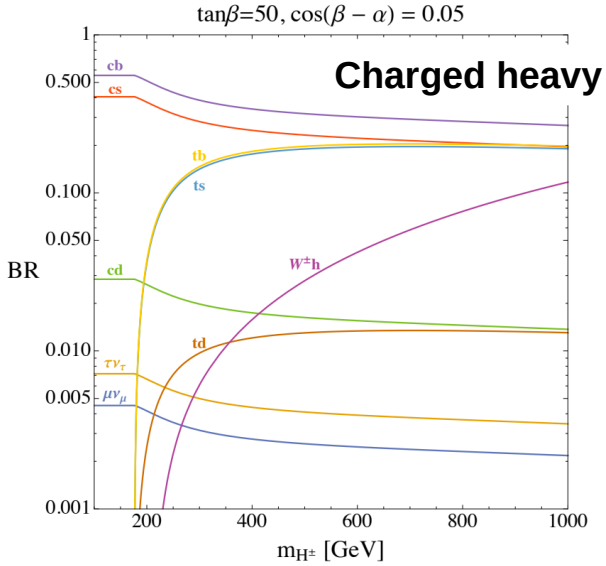
H->ss and H->cs

- H->ss: likely to remain out of experimental reach unless enhanced relative to SM expectations
- H->cs: some BSM models allow for the 1st and 2nd generation fermion masses to be an additional source of EW symmetry break, resulting in a “SM” Higgs doublet (125 GeV) and a “heavy” Higgs doublet (see 1610.02398 for instance, figures on the right taken from Figs. 3 and 6 of that paper)
  - Predicts an enhancement to Higgs cross section
  - Charged heavy Higgs can undergo flavour violating decays (e.g., cs) – s/c-tagging can help with identifying these

Neutral heavy Higgs



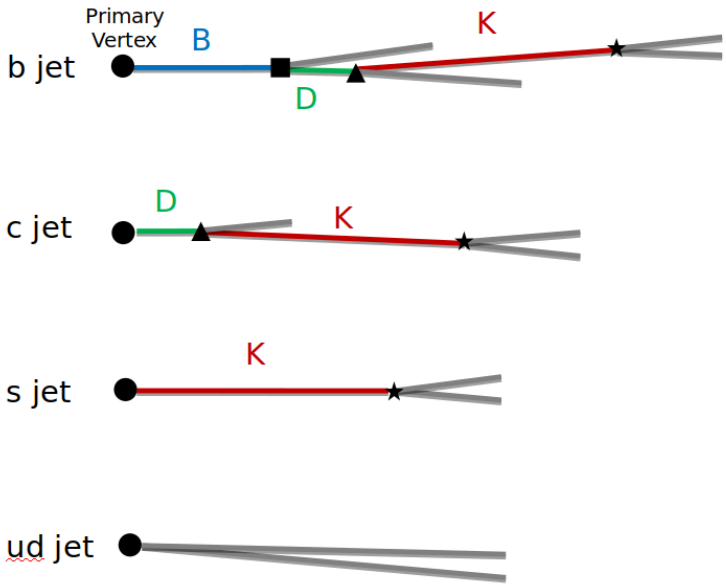
Charged heavy Higgs



Different jet types, pictorially

Discriminants

Taken from Slide 5 of Tomohiko Tanabe's presentation



Charged Kaon track

- Zero track impact parameter w.r.t. primary vertex
- Momentum fraction relative to the jet momentum carried by the leading Kaon
  - (Longitudinal vs transverse components?)

V^0 (K_S^0,Λ^0)

- Vertex momentum & displacement must point in the same direction
- Mean vertex distance smaller compared to b/c

+ the usual b/c discriminants (vertex mass, impact parameter for all tracks, etc.)

Remember to normalize the discriminants to make them boost invariant (as much as possible)

5

Analysis workflow

- Build iLCSoft @ v02-15-02, run macros which closely follow the macro used in Daniel Jean's tutorial
  - **Is this tag still recommended?** e.g., should we update to v02-16?
- Workflow (done in C++ & Python, a similar workflow works equally in Julia):
  - (√) Run ROOT macros on input miniDSTs, dump variables of interest to ntuples
  - (~) Load the ntuples into Python (uproot), train an MVA with TensorFlow+Keras
  - (X) Apply trained MVA to analysis macro running on input miniDSTs
    - **Recommended way** to deploy MVAs in ROOT macros/LCIO? lwtnn?
    - **Background samples??**

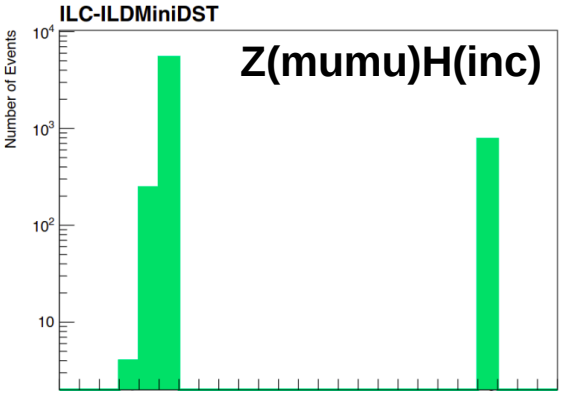
Testing things out

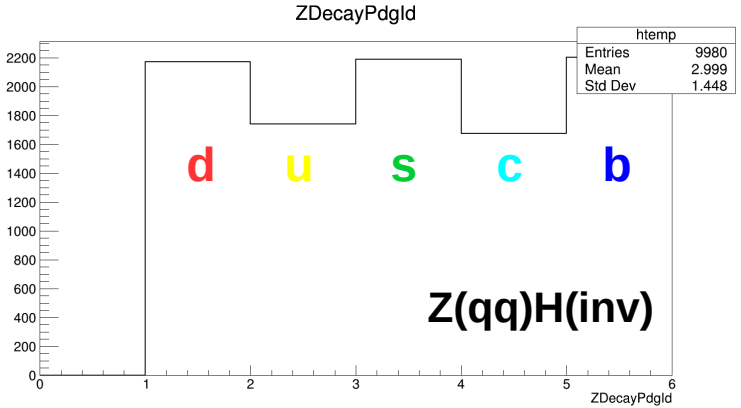
- Wrote dedicated macros for Z(mumu)H(inc) [1] and Z(qq)H(inv) [2] samples
  - BR(H->ss) ≈ 0.1%, so expect very few events for H inclusive sample
  - BR(Z->dd+ss+bb)/3 = 15.6% and BR(Z->uu+cc)/2 = 11.6%, so expect to see strange jet kinematics better with Z(qq)H(inv)

[1] rv01-16-p10_250.sv01-14-01-p00.mILD_o1_v05.E250-TDR_ws.I106479.Pe2e2h.eL.pR-00001-ILDminiDST.slcio
[2] rv01-16-p10_250.sv01-14-01-p00.mILD_o1_v05.E250-TDR_ws.I108079.Pqqh_zz_4n.eL.pR-00001-ILDminiDST.slcio

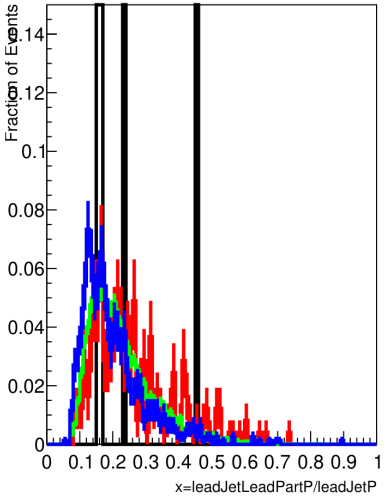
A few sanity checks

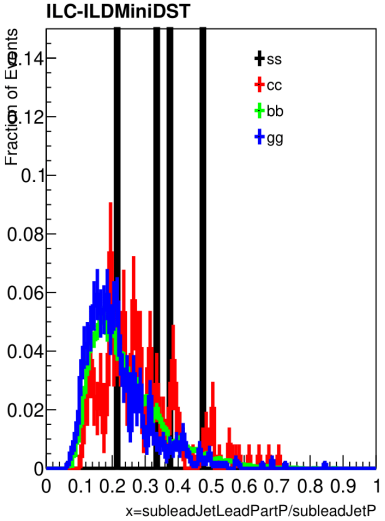
Looks as expected!

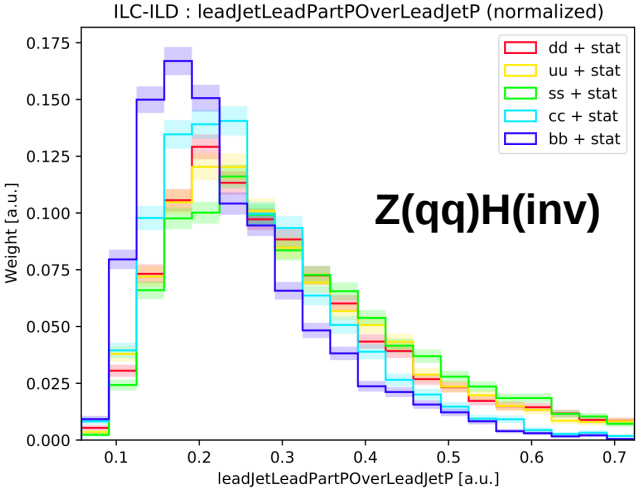




Z(mumu)H(inc)





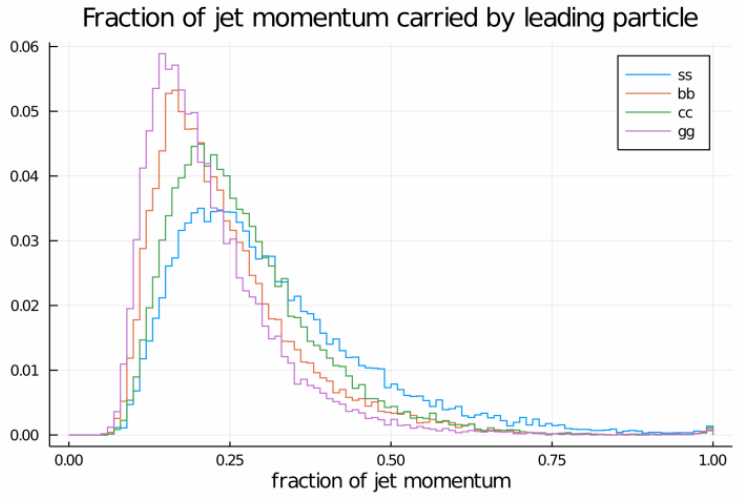


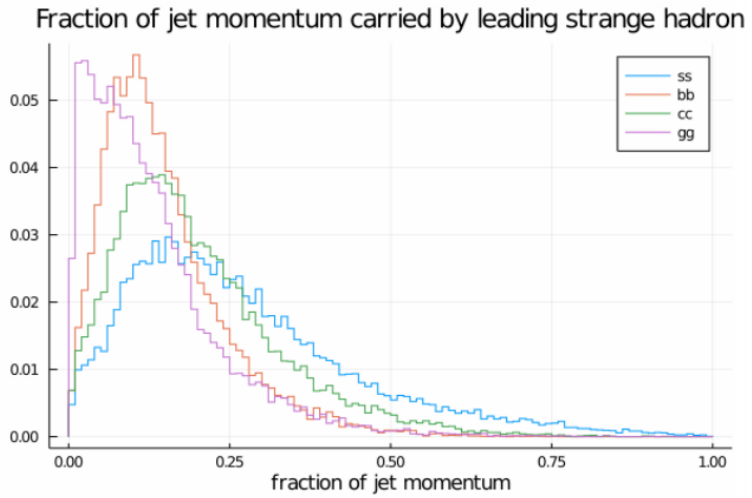
...and consistent with colleagues

SiD →

Fragmentation properties

Using perfect PID (MCParticle)





Taken from Slide 7 of Jan Strube's presentation

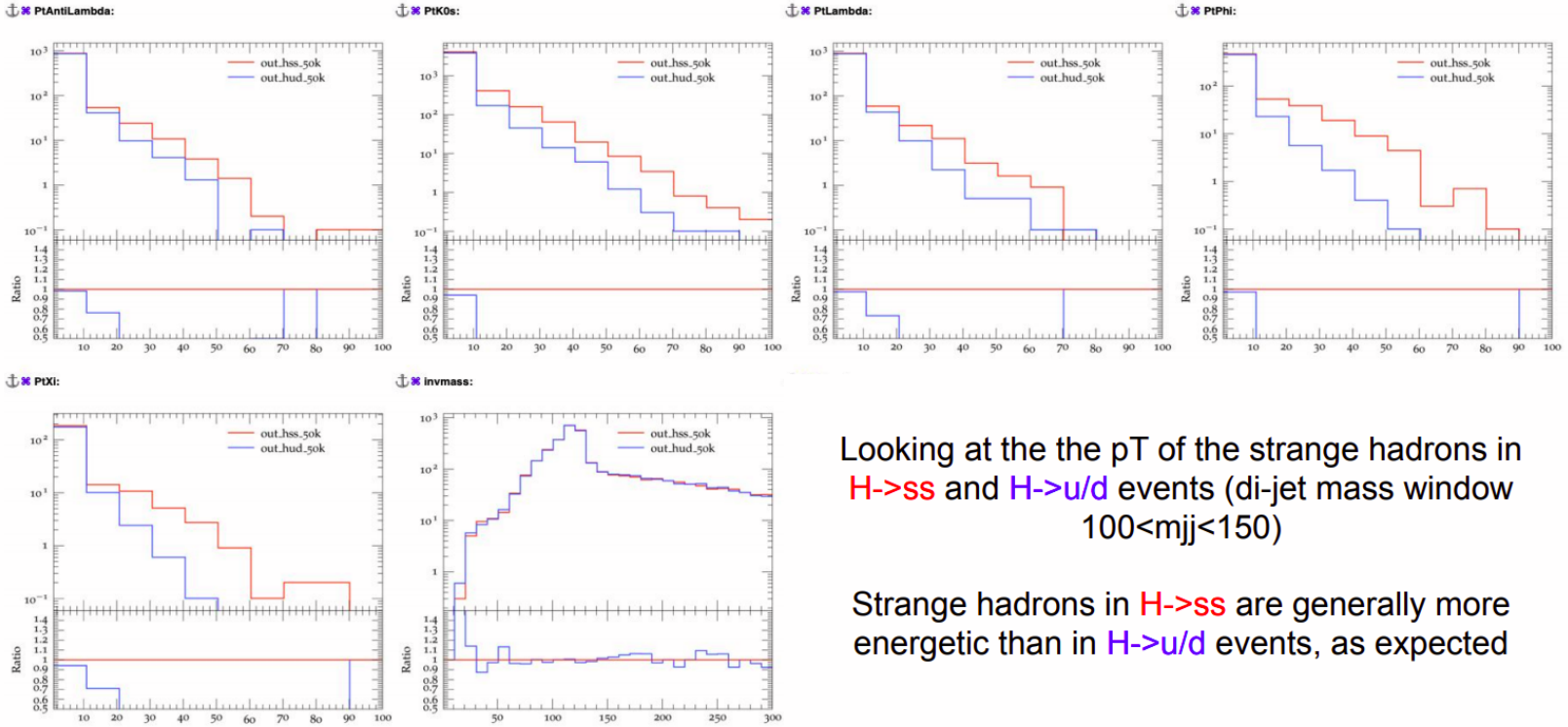
7

...and true at truth level

(in the meanwhile...)

Standalone Py8 generation

Taken from here – thanks to Deepak Kar for his help with producing these!



Looking at the the pT of the strange hadrons in H->ss and H->u/d events (di-jet mass window 100<mjj<150)

Strange hadrons in H->ss are generally more energetic than in H->u/d events, as expected

14

H->qq/gg miniDSTs

- To improve statistics for training, we've switched to dedicated Z(inv)H(qq/gg) samples (thanks to Jenny List and Shin-ichi Kawada!)
  - 50,000/events per flavour
  - Available: /nfs/dust/ilc/group/ild/miniDST/E250-SetA/ILD/flavortag/ (accessible on DESY-NAF)
- **No issues** with running on the samples, but some confusion as how to access the dE/dx, TOF, PID, etc. – more on this

(Towards) training an s-tagger

- **Haven't** gotten this far, *some* considerations:
  - Training events will likely see the MVA deployed on them too – **need** to kfold inputs:
    - evt->getEventNumber() % N == {0, 1, ..., N-1}, N := # of kfolds
  - Inputs will likely consist of jet variables + per-track variables within each jet
    - In H->qq, there are two jets in each event: do we want to use only **1** of the jets in training? If so, **leading** or **subleading** or **random**?
    - Track momentum redefined wrt to the jet momentum axis, 4-vector normalized to jet momentum
    - Sensible ordering of tracks? In order of highest **track+calo weight** or **momentum**?

Inputs and outputs

- Outputs: could imagine the network provides bottom, charm, strange, and light output scores
  - **Multiclassifier** provides more freedom for output class
- Jets: p4, ILD tagger scores (b-, c-, o-, and category?), ...
  - **Anything else which is sensible/useful to include?**
- Tracks (jet constituent particles): p4, momentum / jet momentum, dE/dx (+ uncertainty?), different PID likelihoods, ...
  - **Anything else?**

Tagger architecture(s)

- Possible architectures from the literature include:
  - “Maximum performance of strange-jet tagging at **hadron** colliders” (2011.10736 – published in November 2020)
    - {Recurrent neural network for track inputs} + {jet inputs} -> Concatenate -> multilayer perceptron (MLP) -> output
    - Could also use MLP on the jet inputs prior to concatenation
  - “ParticleNet: Jet Tagging via Particle Clouds” (1902.08570)
    - Proposed for flavour tagging at FCC-ee (see talk here)
    - *Complex*: represent particles in jet as a graph and apply EdgeConv (1801.07829) units to relationships between a given particle and its nearest neighbours

dE/dx+TOF for kaon separation

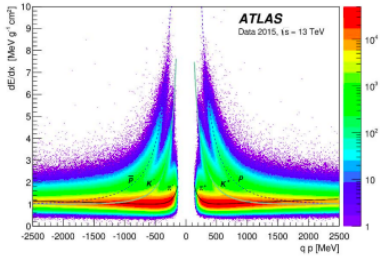
Taken from here

Strange Tagging

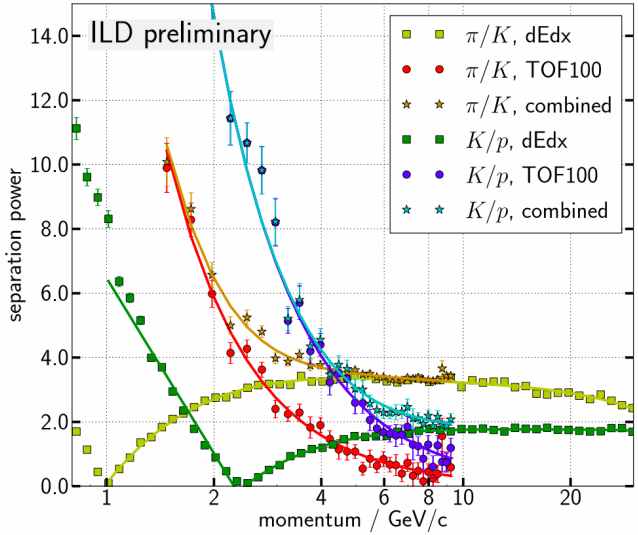
SLAC

- **Existing strange tagging** studies suffer from **low efficiency and very large mis-tag probability** from u and d quarks, even when using sophisticated machine learning algorithms
- To complement existing studies (more on this in the next slides), we thought we would put **more emphasis on exploiting Particle Identification** to get a better handle on pions/kaons identification, and consequently on s/d quark discrimination
  - This implies looking at new detector concepts
  - Current general purpose detectors use the well known **d*E*/d*x* dependence on βγ**, but this only allows to get to good PID up to ~1 GeV
  - Alternatively, as foreseen for the HL-LHC detectors, timing information can be used to deduce a **velocity** that, in combination with the standard measurement of **momentum from track curvature in the magnetic field**, yields a measure of the **charged particle mass.**
  - Another very effective way to achieve particle identification is through **Cherenkov detectors**, as done in the ALICE and LHCb experiments at the LHC

Phys. Rev. D 93, 112015 (2016)



In either scenario, **kaons beyond the 10 GeV range** will have to be identified in order for this to be relevant for strange tagging.



Plot taken from Slide 14 of Uli Einhaus' presentation

dE/dx seems to reach a 2 sigma pi/K separation power throughout the desired momentum range: is this good enough?

Technical questions

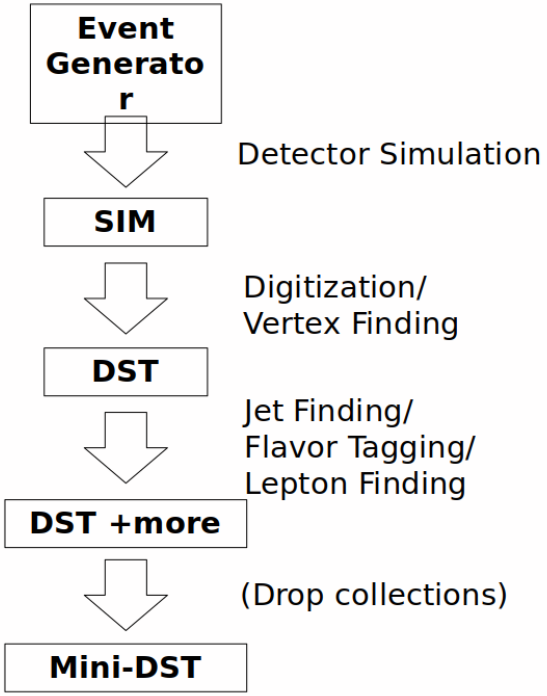
- Constituents of jet accessed with ReconstructedParticle::getParticles()?
- How to access track(s) associated to constituent particle? ReconstructedParticle::getTracks()?
  - Returns vector of nullptrs – understood?
  - Is it possible to access impact parameters, dE/dx otherwise?
- Likelihood seems access for algorithm “LikelihoodPID” here and for algorithm “dEdxPID” here – what is the difference between the two?
  - e.g., see Backup for first attempt at accessing this info
- Is there a way to access TOF?

DST vs. miniDST

Do miniDSTs have the **links** from the particles in a jet to their respective tracks? Is there a good reason why we should have **all** tracks?

ILD Samples

Taken from Slide 7 of Tomohiko Tanabe's presentation



	DST	Mini-DST
Tracks	✔	
MC Particles	✔	✔
Reconstructed Particles	✔	✔
Primary Vertex	✔	✔
Isolated e/mu/tau		✔
Jets (2, 3, 4, 5, 6)		✔
b/c/o tagging		✔
Particle ID	✔	✔
Secondary Vertices	✔	
V0 Vertices [pandora]	✔	
V0 Vertices [lcfiplus]	✔	

Do we want to request mini-DSTs with the full collections (for a select few channels)?

7

Conclusion

- Making **steady** progress, a long way to go though!
  - Workflow for running on flavour tag samples is straightforward, still need to add PID info to ntuples
  - Framework exists for training in Keras, still need to figure out how to define RNN+MLP network (starting with the simpler of the architectures in the literature)
- In terms of achieving nice results, we will profit from the delayed Snowmass timeline
  - There are also parallel efforts in 4D tracking technology (see LoI)

Questions?

Backup

ALICE PID Performance

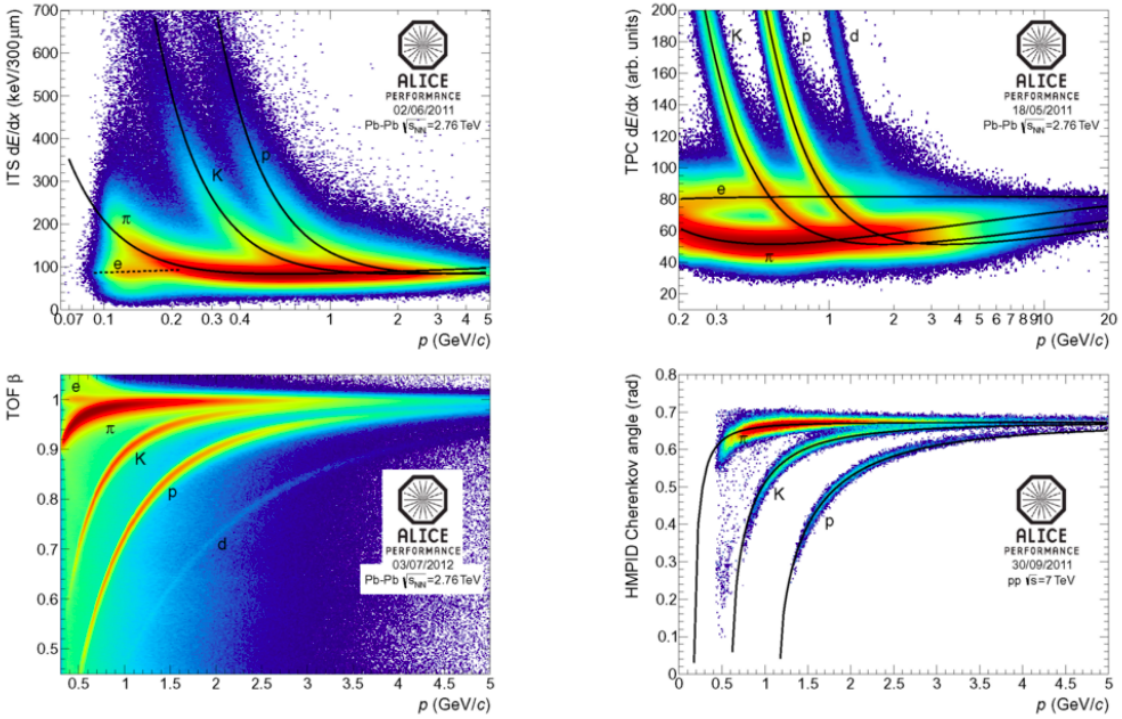


Figure 2: The PID performance of the ALICE detector. The figure shows the ITS dE/dx vs p, the TPC dE/dx vs p, the TOF β vs p, and the HMPID Cherenkov angle vs p.

Taken from Fig. 2 in: M. Ivanov, Nuclear Physics A 904–905 (2013) 162c–169c

Kaon Likelihood for H->ss