End-to-End Jet Classification of Boosted Top Quarks with CMS Open Data

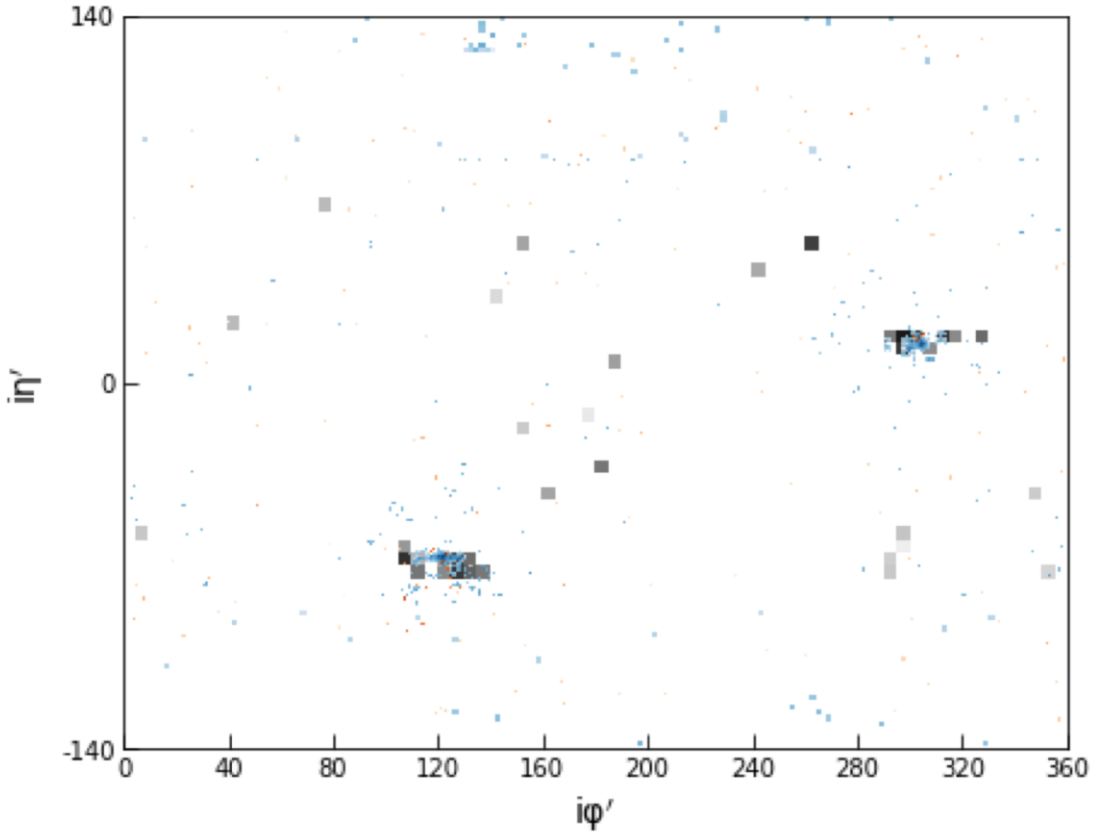
Michael Andrews, **Bjorn Burkle**, Shravan Chaudhari, Davide Di Croce,
Ulrich Heintz, Meenakshi Narain, Manfred Paulini, Emanuele Usai

BROWN

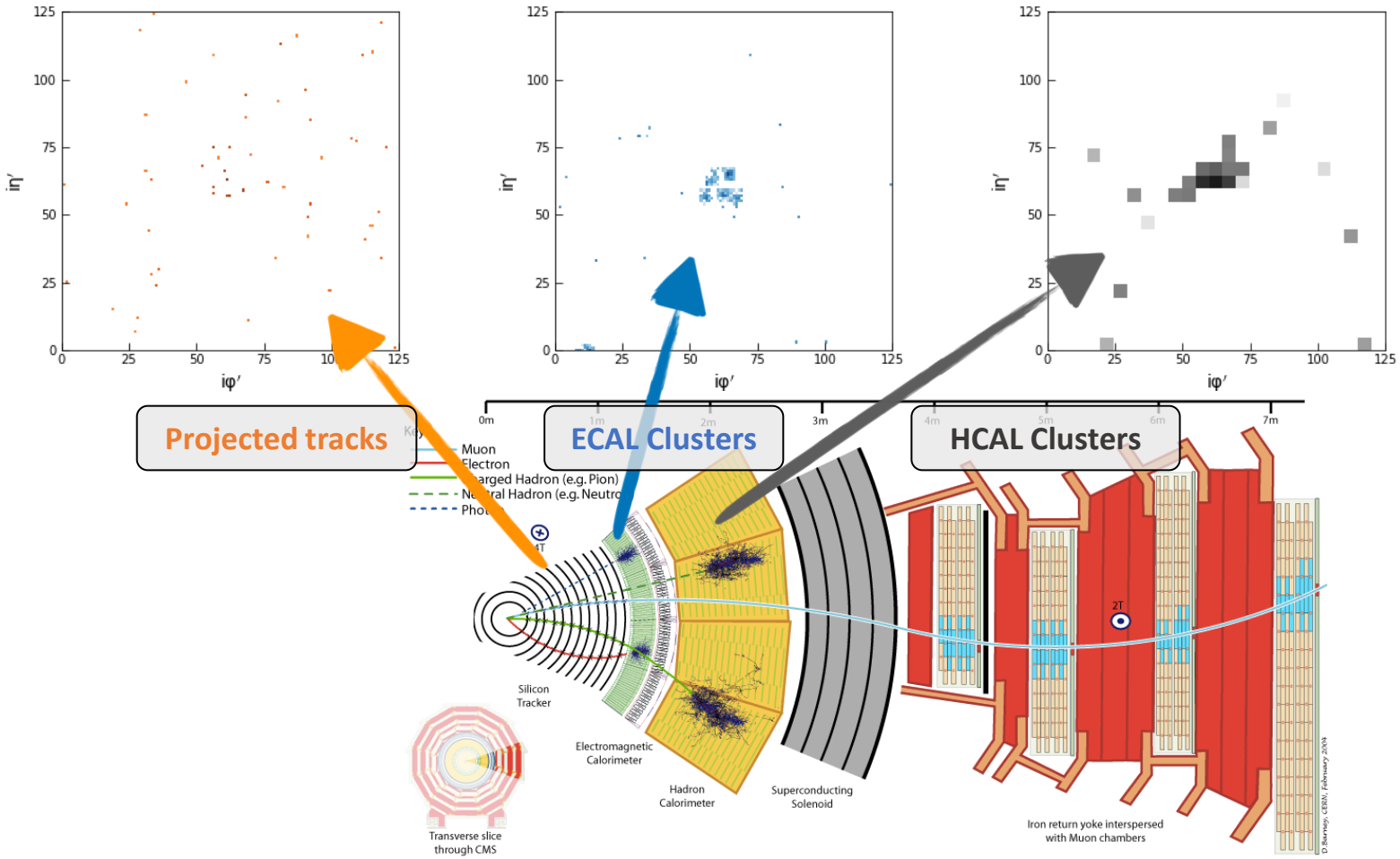
Carnegie Mellon University

The End-To-End Philosophy

- Traditional tagging algorithms use reconstructed physics objects as inputs to classifiers
- Circumvent reconstruction, and train networks directly on low-level information
- Construct images or graphs made from low detector readouts, simulate using Geant-4 full simulation
- Multiple papers on the use of image-based jet identification [1, 2]

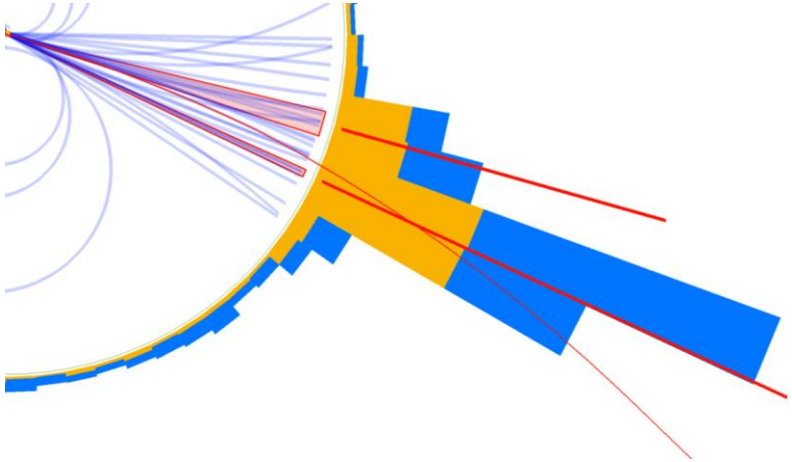


The Use of Detector Images

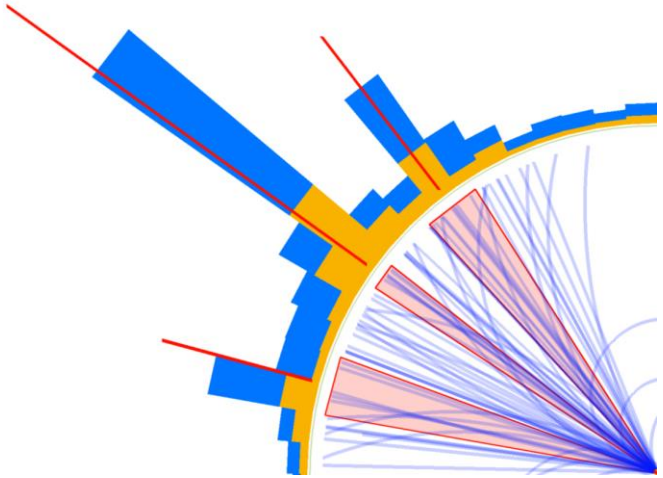


Boosted Top vs QCD Discrimination

- Previous end-to-end studies extended to discriminate between jets originating from boosted hadronic top quark decays and boosted QCD jets
- Study uses same ResNet architecture as our previous work



QCD Jet



Hadronic Top

Jet Selections

- CMS Open Data samples used to obtain top and QCD jets
  - Public datasets produced with the purpose of machine learning experimentation
  - Events simulated at 8 TeV center of mass energy and using 2012 detector conditions
  - Current work is first study to make use dataset availability

Sample	Description	Dataset Location
TTJets_HadronicMGDecays	$t\bar{t}$, $p_T(t)$ > 400 GeV, all-jet top decays	10.7483/OPENDATA.CMS.OPKY.OJMJ
QCD_Pt-300to600_TuneZ2star_Flat	Non-top multijet, flat 300 < $\hat{p}_T$ < 600 GeV	10.7483/OPENDATA.CMS.HUED.7R3E
QCD_400to600_TuneZ2star_Flat	Non-top multijet, flat 400 < $\hat{p}_T$ < 600 GeV	10.7483/OPENDATA.CMS.YWDZ.KSLK
QCD_Pt-600to3000_TuneZ2star_Flat	Non-top multijet, flat 600 < $\hat{p}_T$ < 3000 GeV	10.7483/OPENDATA.CMS.CWTT.8Q3E

Jet Selections

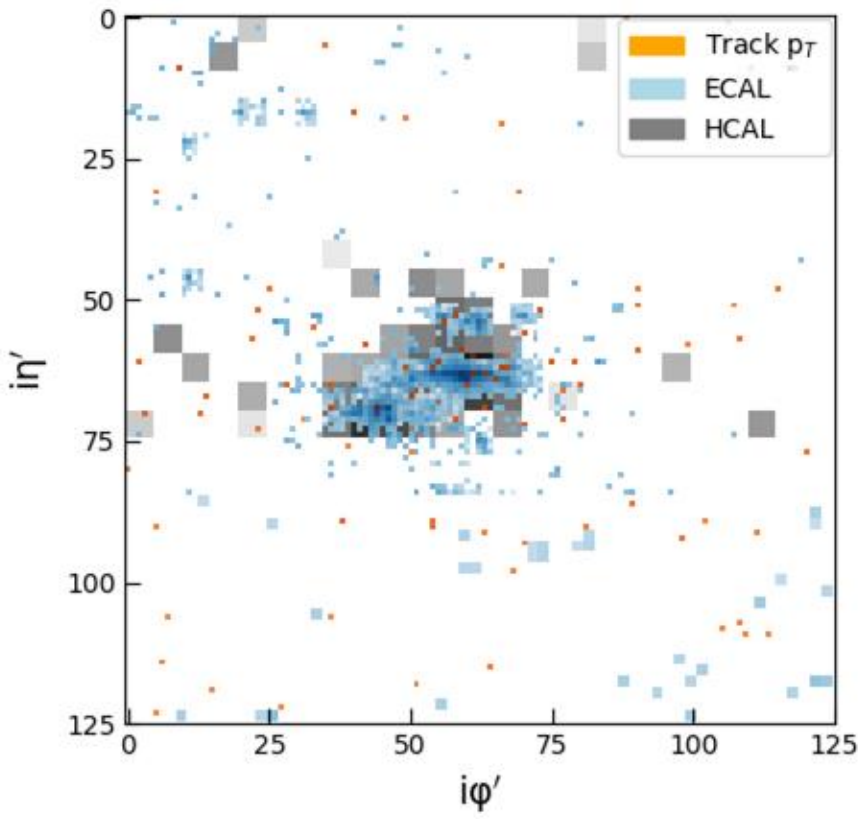
- Selected AK8 jets with $p_T > 400$ GeV and $|\eta| < 1.57$
  - η cut ensures that jet image is within the tracker η acceptance
  - Gen level information used to tag top jets and assure that t, W, and b trajectory within jet cone
- QCD jets pseudo-randomly resampled to match p_T distribution of top jets

Category	Top quark jets	Non-top quark jets	Total Jets
Train	1280830	1279179	2560000
Validation	47859	48141	96000
Test	319819	320181	640000

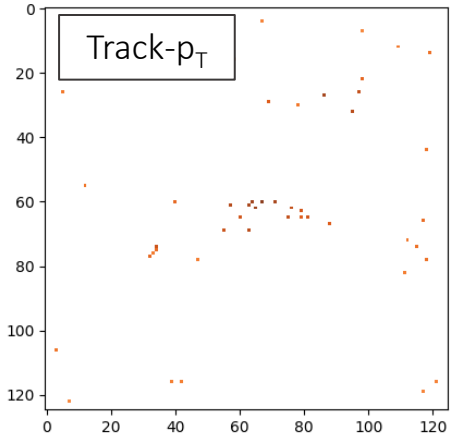
The Need For More Features

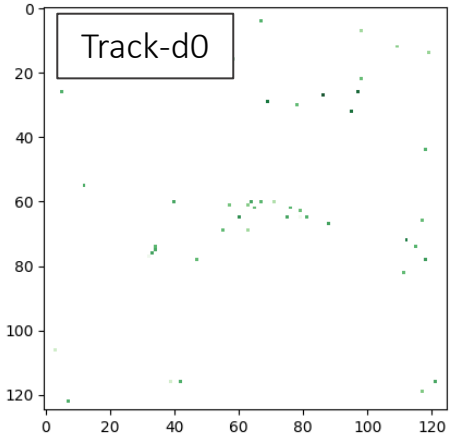
Layer Combinations	ROC-AUC
Track + ECAL + HCAL	0.967±0.002

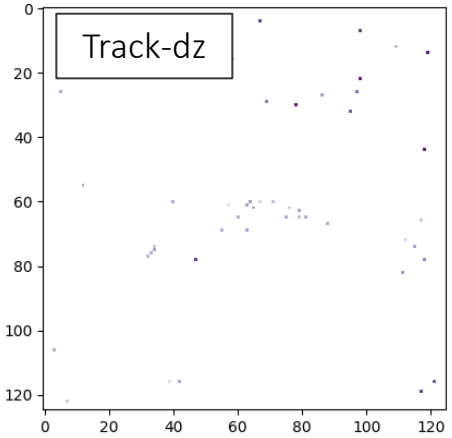
- Good performance, but could be better
  - Previous end-to-end studies were agnostic to track vertexing information
  - Identification of SV imperative for top-tagging
- SV information added in two forms
  1. Additional track channels weighted by impact parameter variables
  2. Addition of pixel hits used for track reconstruction



IP-Weighted Tracks



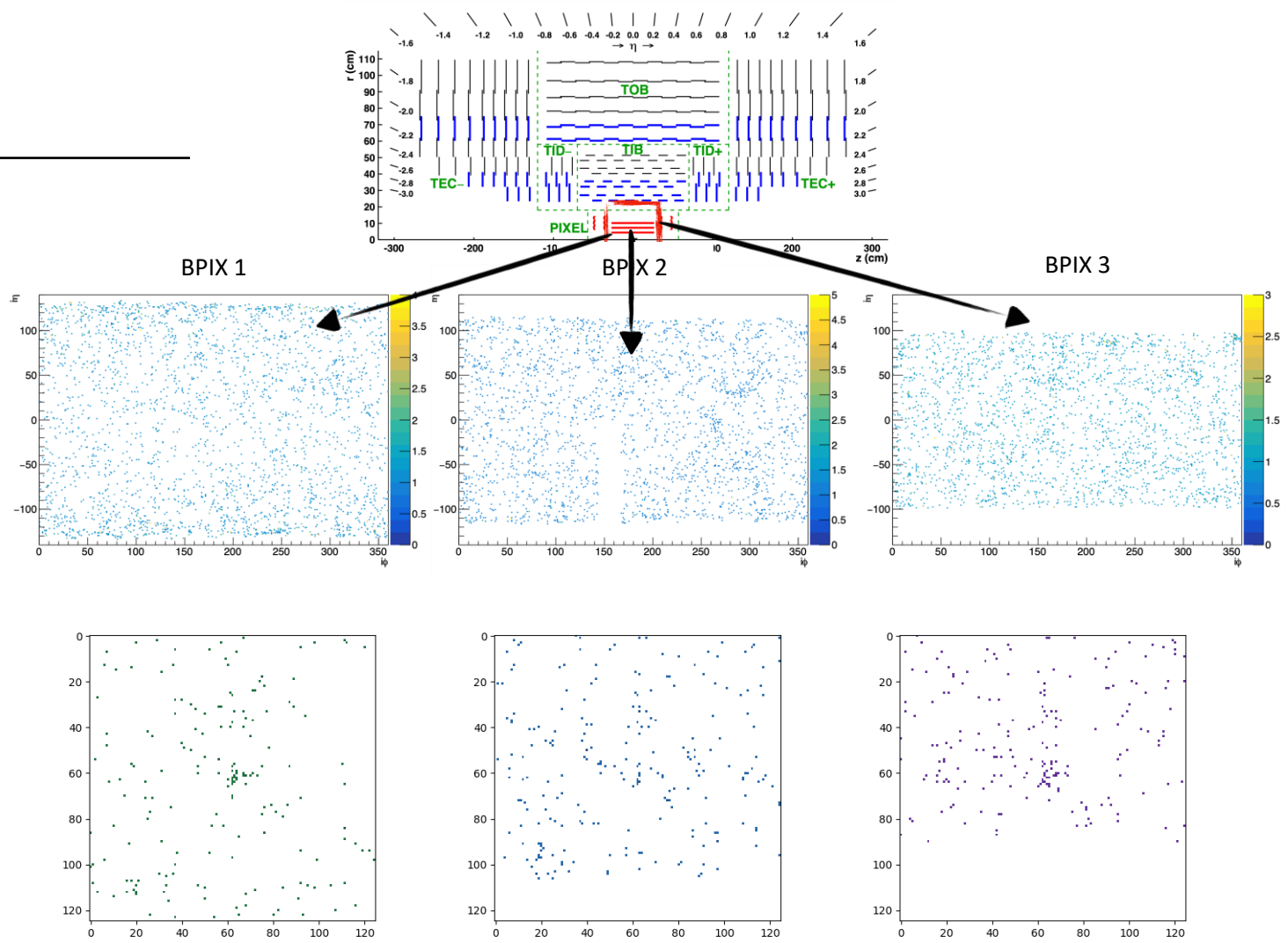


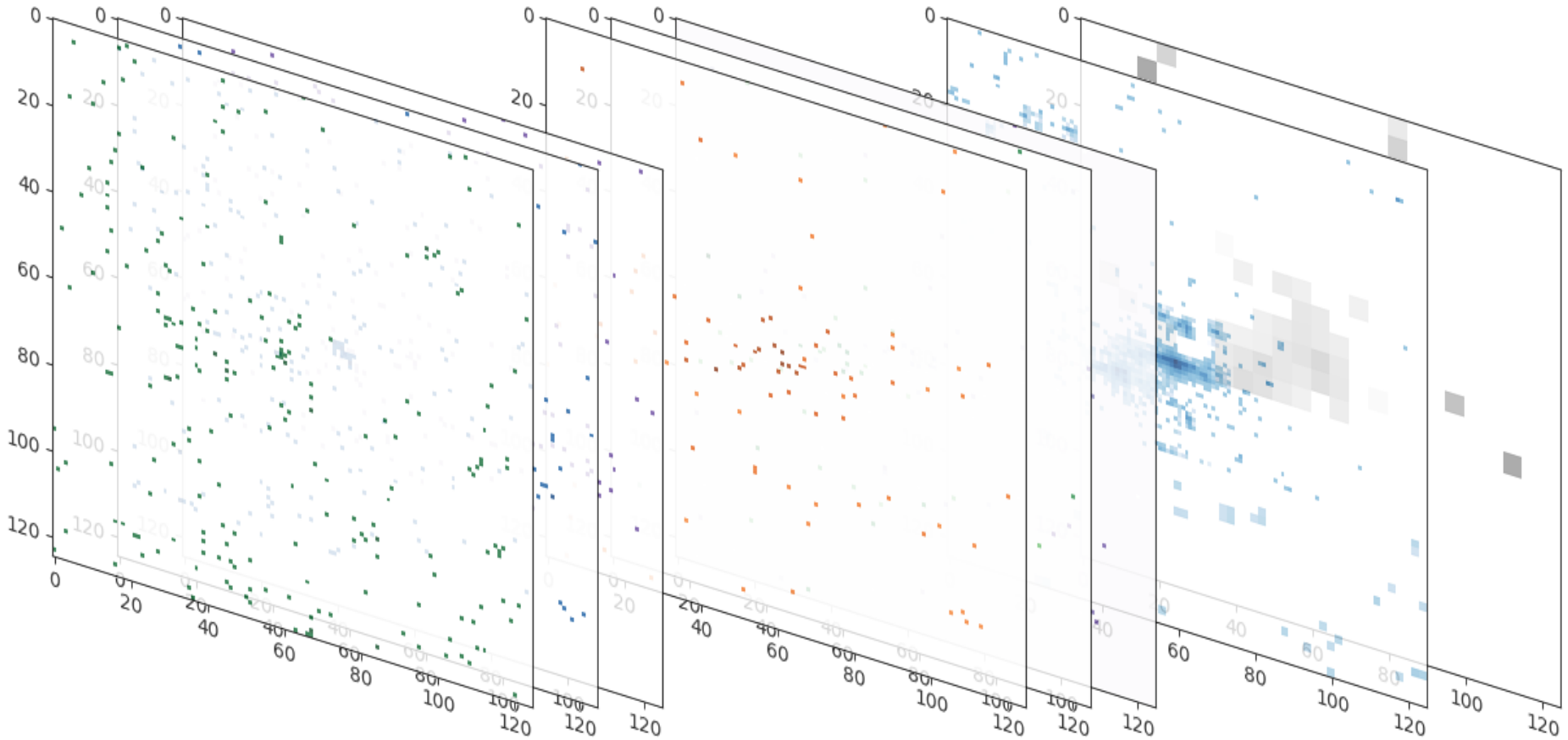


- Two additional track layers are added corresponding to longitudinal (dz) and transverse ($d0$) impact parameter significance
- Spatial location of activated pixels are identical among track layers
- Extra cut made on tracks where impact parameter is too large

Pixel RecHits

- 3 new image channels added, corresponding to RecHits in three layers of barrel region of pixel detector
- RecHits binned in $i\eta$ and $i\phi$ to match resolution of other channels
  - RecHit η is recalculated w.r.t. PV
- Constructed as binary layers – 1 if there is a hit, otherwise 0





Pixel RecHits

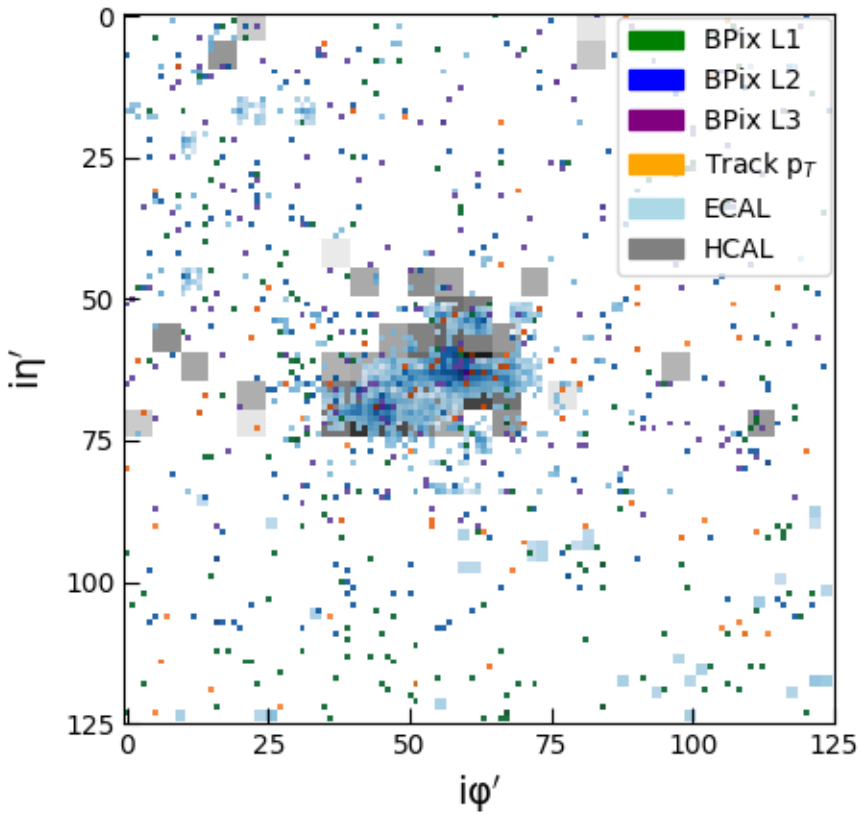
Track Information

Energy Clusters

Network Performance

Layer Combinations	ROC-AUC
Track p_T **(baseline)**	0.955±0.002
Track p_T + ECAL + HCAL **(nominal)**	0.967±0.002
Track p_T + $d0$ + dZ	0.972±0.002
Track p_T + $d0$ + dZ + ECAL + HCAL	0.981±0.002
BPIX1–3	0.947±0.002
BPIX1–3 + Track p_T	0.965±0.002
BPIX1–3 + ECAL + HCAL **(no reconstructed variables)**	0.975±0.002
BPIX1–3 + Track p_T + $d0$ + dZ	0.977±0.002
BPIX1–3 + Track p_T + $d0$ + dZ + ECAL + HCAL **(full image)**	0.9824±0.0013

- Observe large increase in discriminating performance after inclusion of new features
- Reconstructed track variables lead to stronger discriminating power than RecHits
  - RecHits still give powerful discriminating capabilities when mixed with calorimeter information



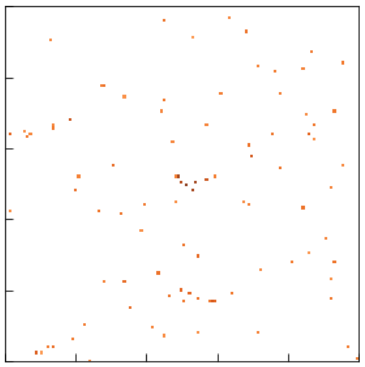
Conclusion

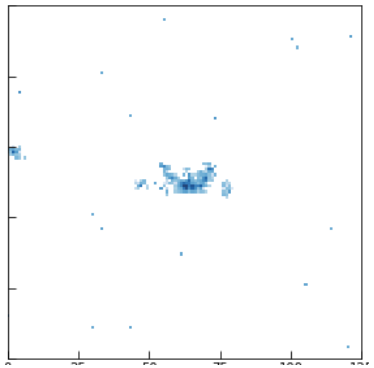
- End-to-end framework has been expanded to discriminate between boosted hadronic top decays and QCD jets
  - ROC-AUC of 0.9824±0.0013 set as first benchmark for publicly available open datasets
- First ever use of low-level tracker RecHit information as input feature for jet discrimination
  - Able to obtain a ROC-AUC score of 0.975±0.002
- Full paper now available at arXiv:2104.14659

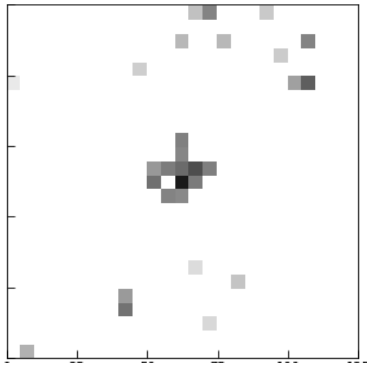
Backup

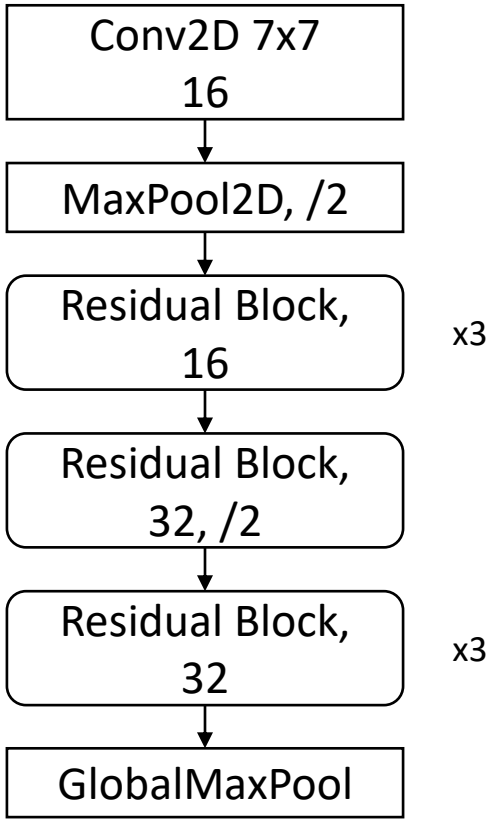
Quark vs Gluon Discrimination

- ResNet architecture used to discriminate between quark and gluon jets
- Trained on single subdetectors, and combinations of subdetectors









Quarks vs Gluon Discrimination

- Network gains most discrimating power from tracks
- E2E jet image approach out-performed all other algorithms

E2E jet image	**ROC AUC**
Generated EM+Had	0.854
Tracks+ECAL+HCAL	0.808
Tracks+ECAL	0.804
ECAL+HCAL	0.781
Tracks	0.782
ECAL	0.760
HCAL	0.682

Jet ID Algorithm	**ROC AUC**
E2E jet image, Tracks+ECAL+HCAL	0.8077 ± 0.0003
RecNN, ascending-p_T	0.8017 ± 0.0003
RecNN, descending-p_T	0.802
RecNN, anti-k_T	0.801
RecNN, Cambridge/Aachen	0.801
RecNN, no rotation/re-clustering	0.800
RecNN, k_T	0.800
RecNN, k_T-colinear10-max	0.799
RecNN, random	0.797