End-to-end particle and event identification at the LHC with CMS Open Data

John Alison, Sitong An, Michael Andrews, Patrick Bryant, Bjorn Burkle, Sergei Gleyzer, Ulrich Heintz, Meenakshi Narain, Manfred Paulini, Barnabas Poczos, **Emanuele Usai**

APS DPF 2019 — 31 July 2019

Outline

Introduction: the end-to-end approach, building images

previous work: photon vs. electron, discriminator bias

jet and event ID: quark vs. gluon

top tagging: new layers configurations

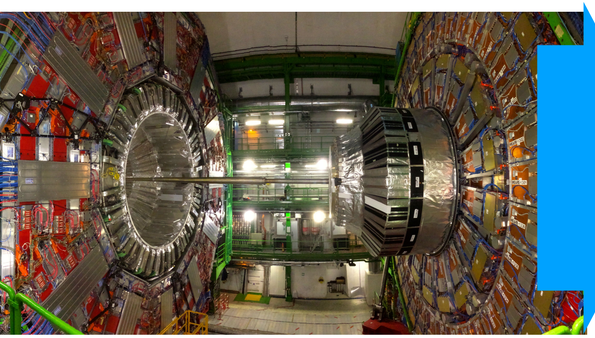
conclusions & outlook

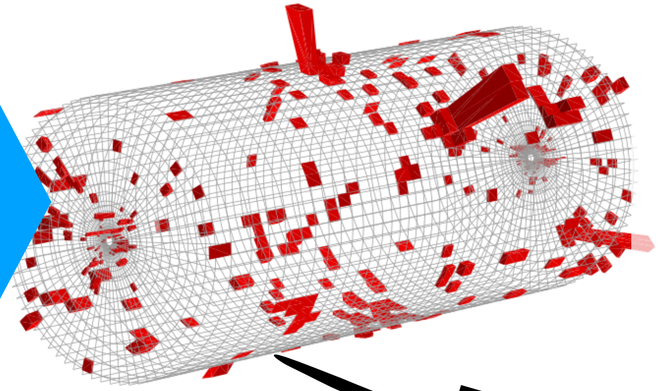
What is E2E

Train Particle/Jet/event IDs starting from low-level detector hits

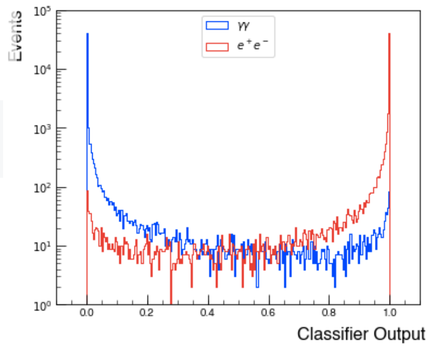
End-to-end

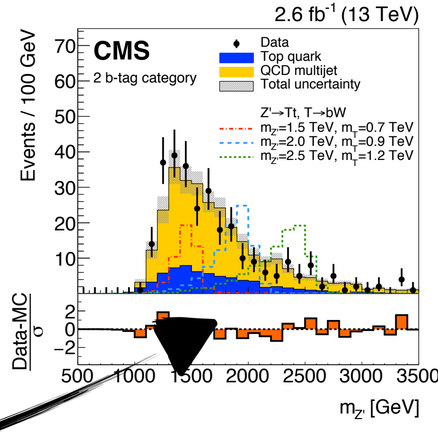
Sensors → Hits/Energy deposits → Reconstructed objects → Particle/Jet Discriminators → Event interpretation





	Energy	p_T	η	φ
1	74.0	70.6	0.44	1.92
2	47.2	45.3	0.69	0.39



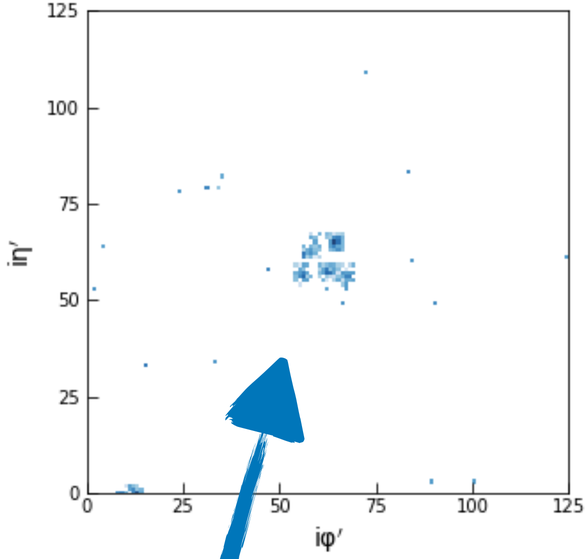


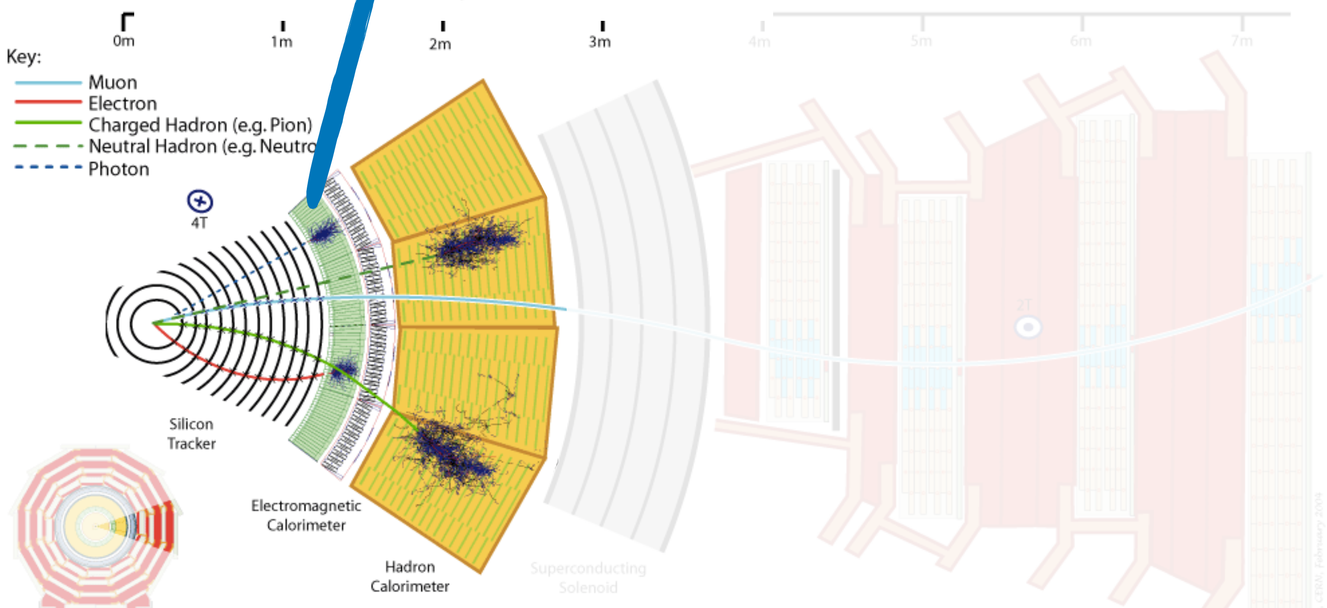
End-to-end

Proof of concept, not a readily usable classifier.

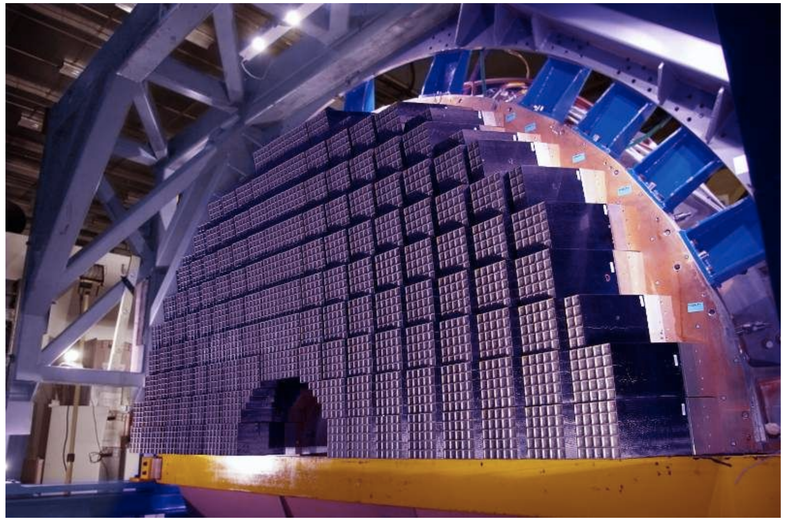
Detector images

ECAL, energy deposits, 1 pixel per crystal

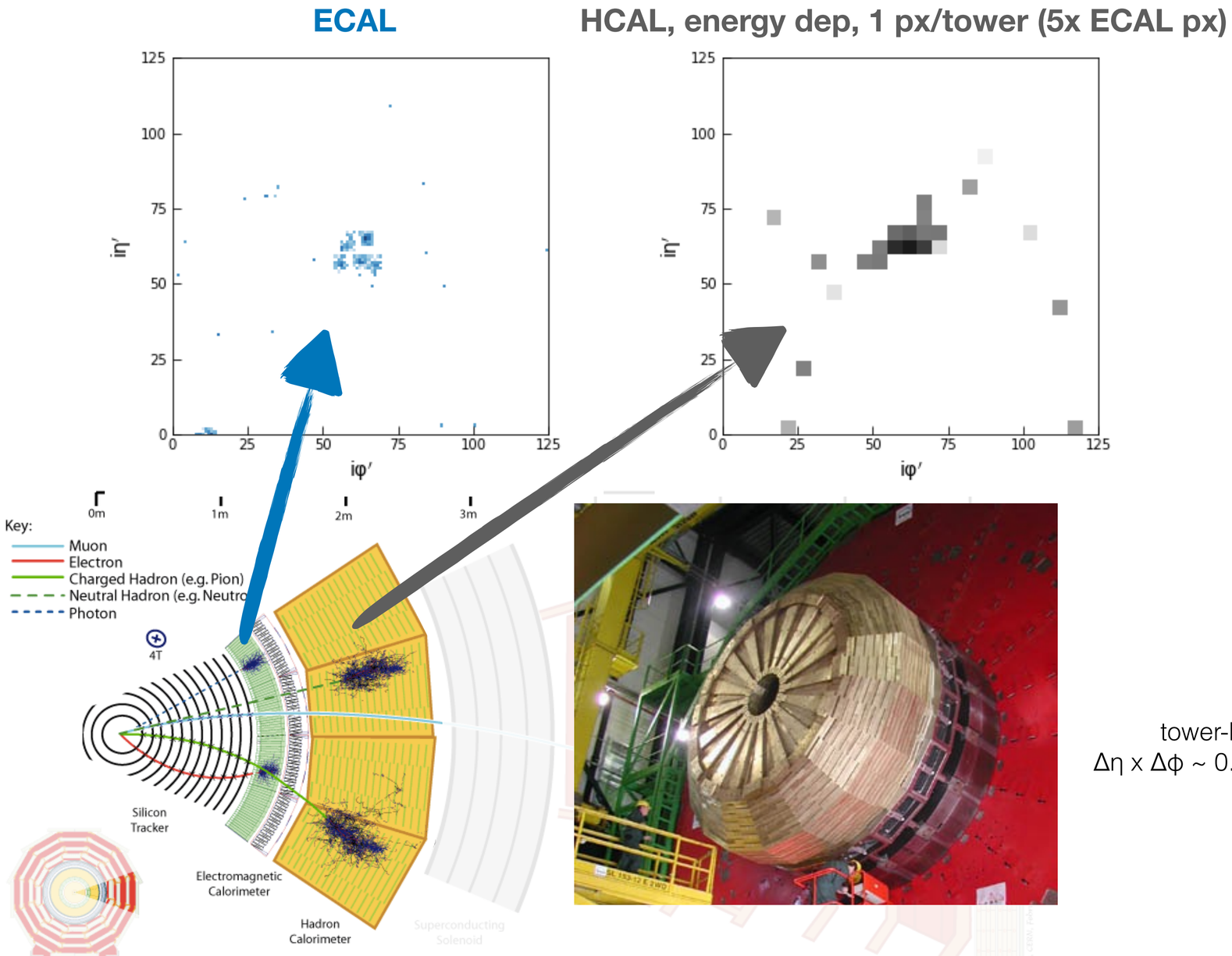




crystal-based
$\Delta\eta \times \Delta\phi \sim 0.0174 \times 0.0174$

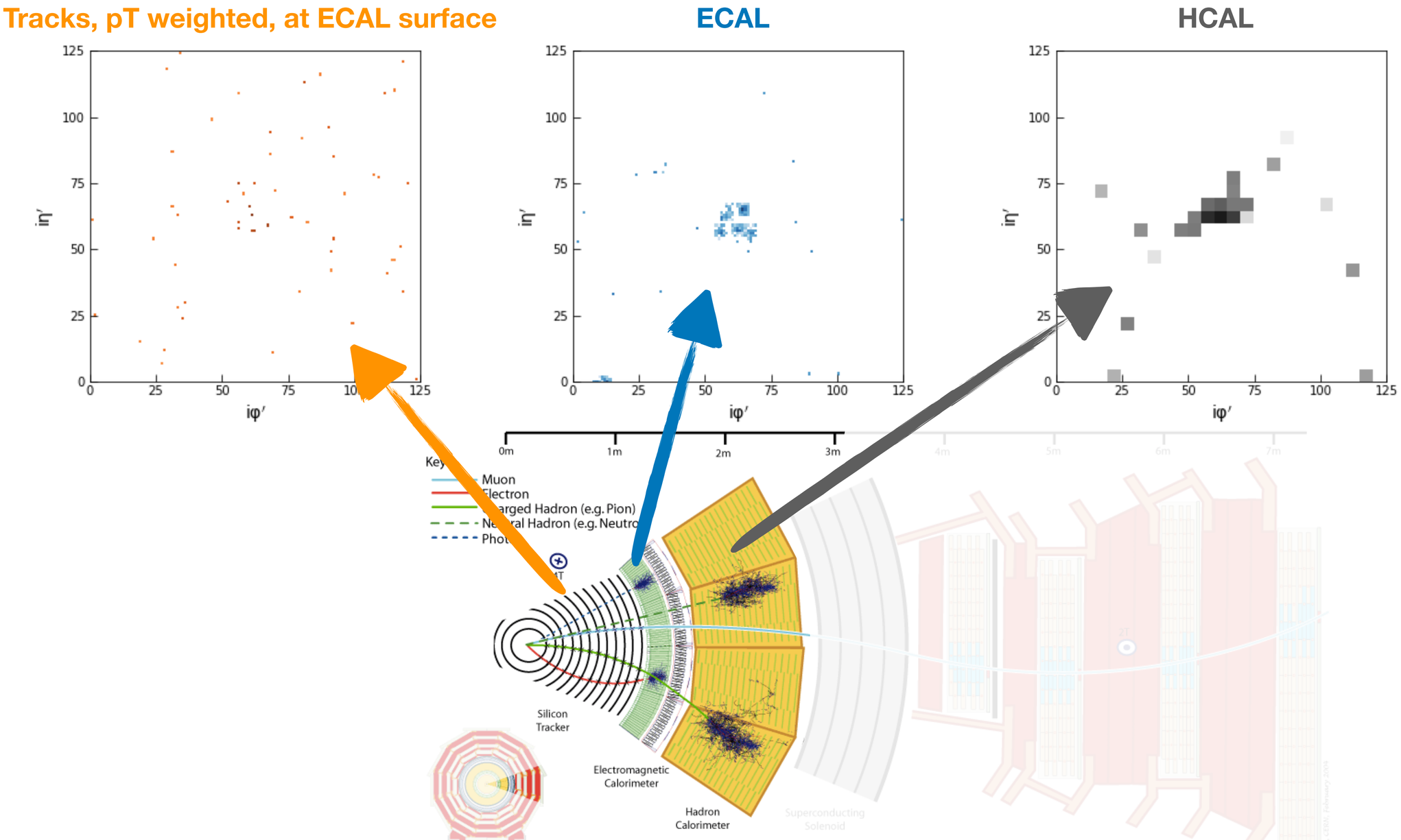


Detector images

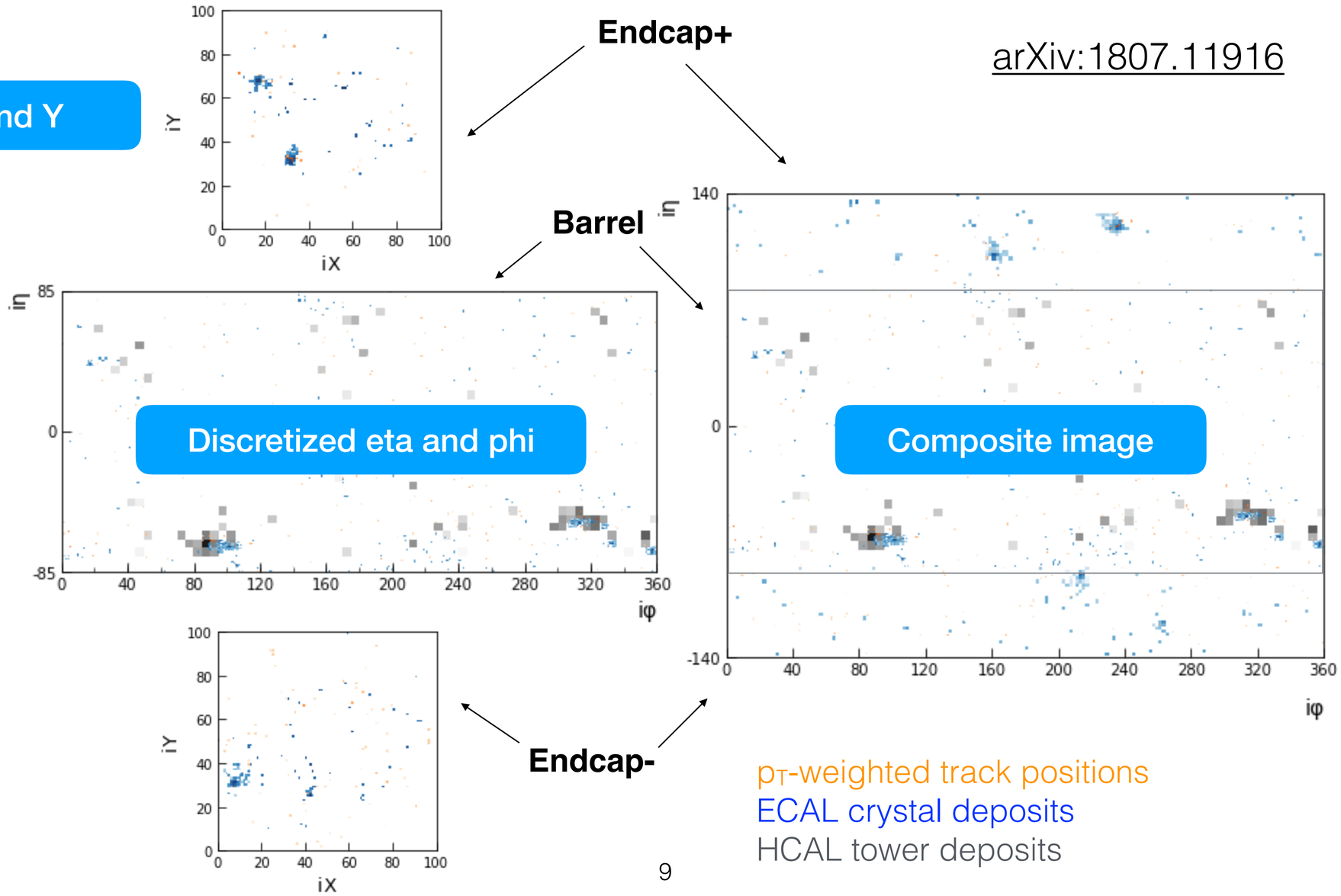


tower-based
$\Delta\eta \times \Delta\phi \sim 0.087 \times 0.087$

Detector images



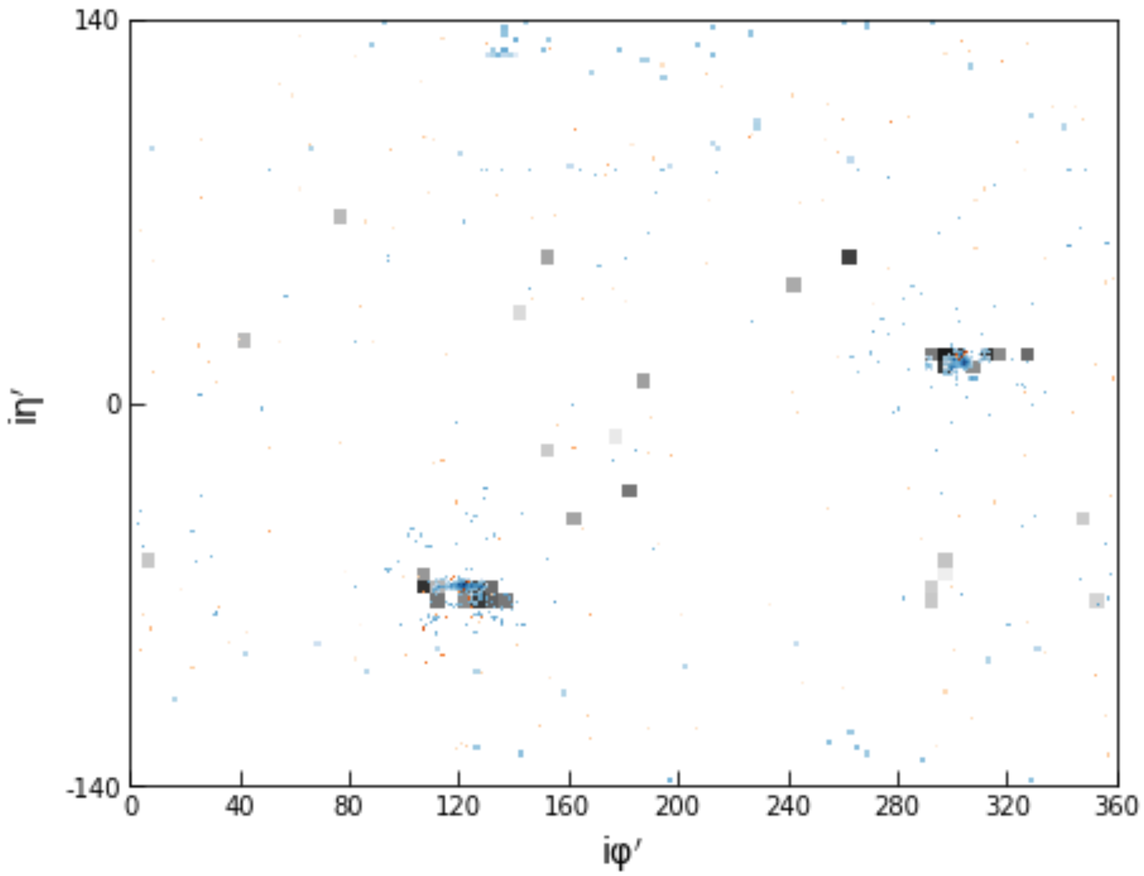
Discretized X and Y



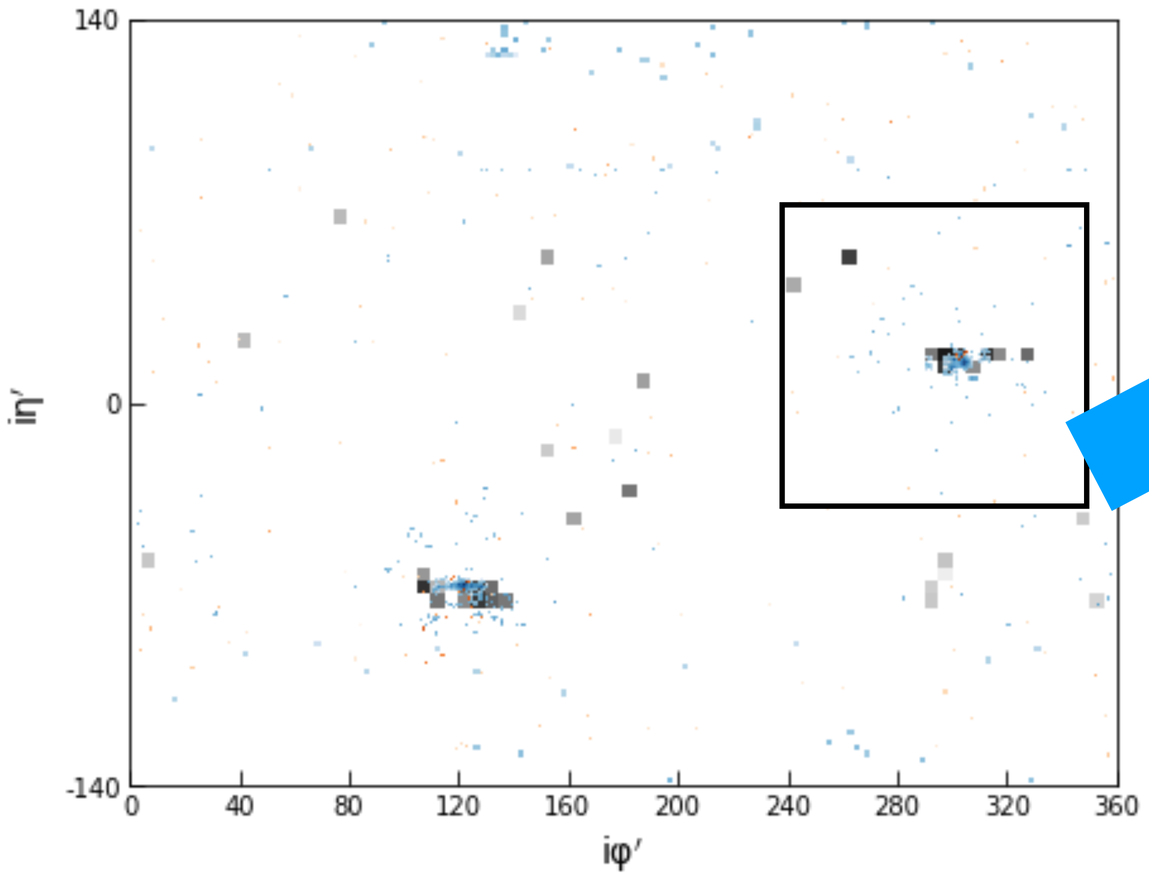
arXiv:1807.11916

p_T-weighted track positions
ECAL crystal deposits
HCAL tower deposits

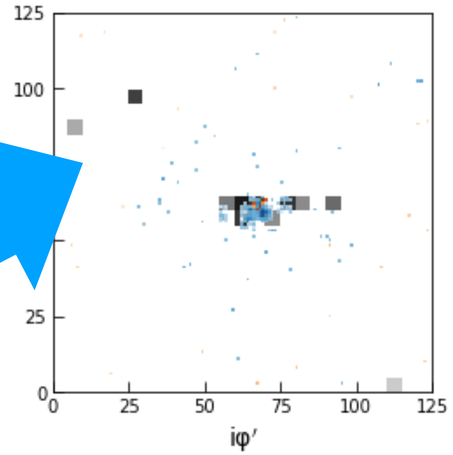
Full-detector image



Full-detector image

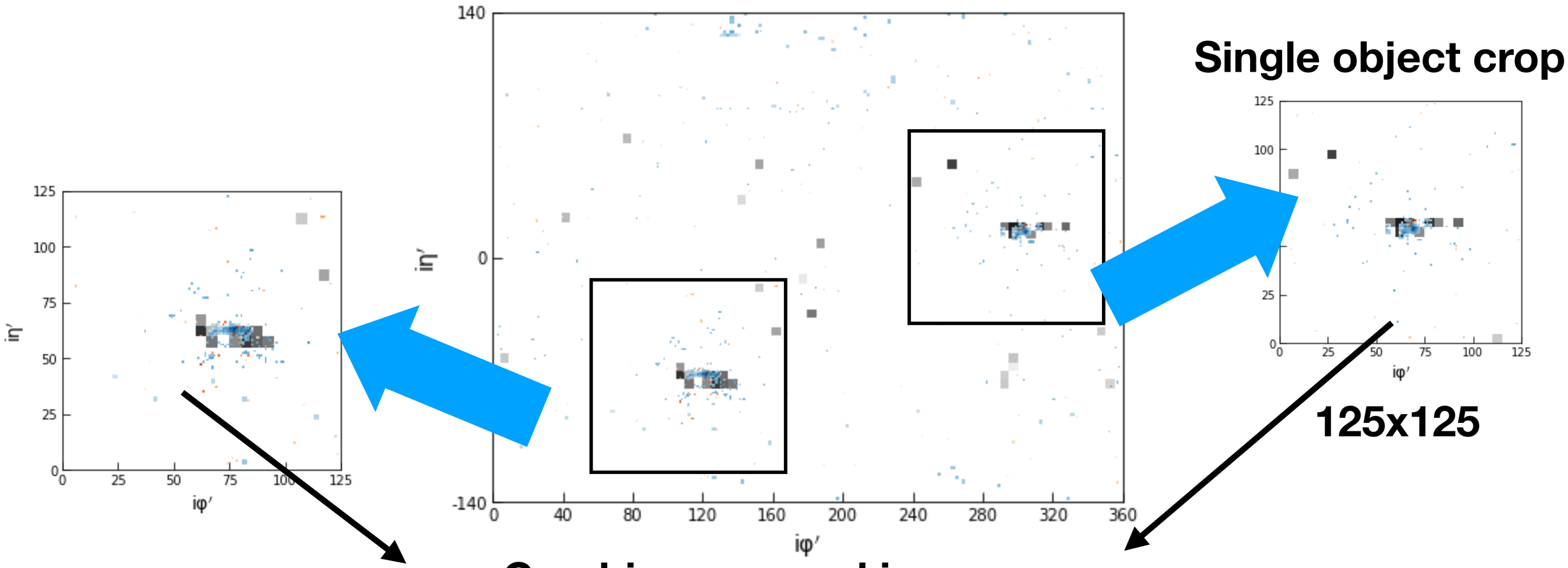


Single object crop



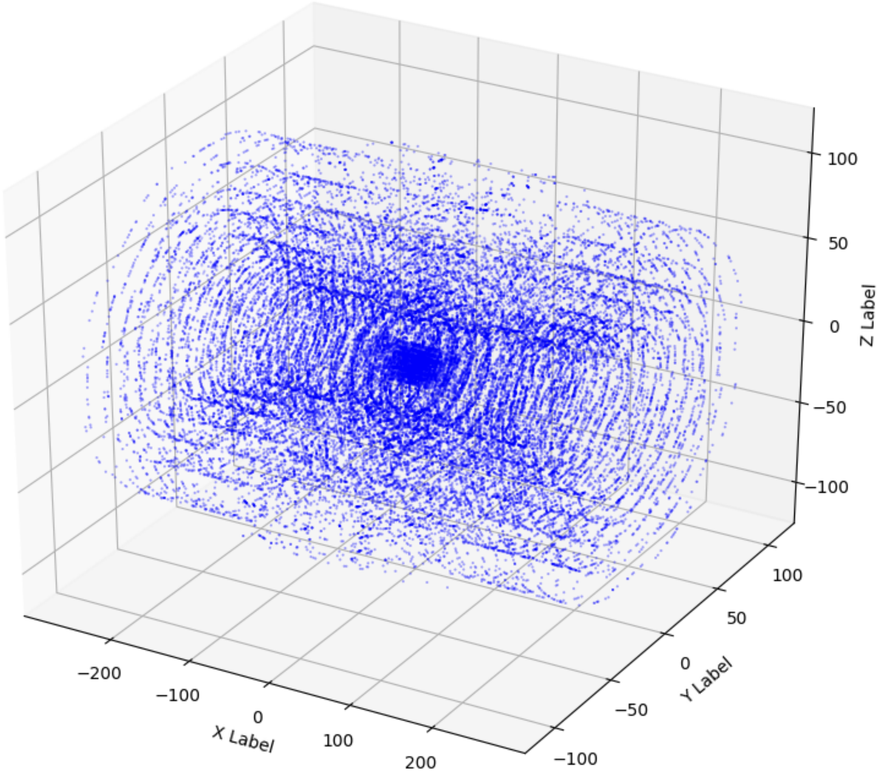
125x125

Full-detector image

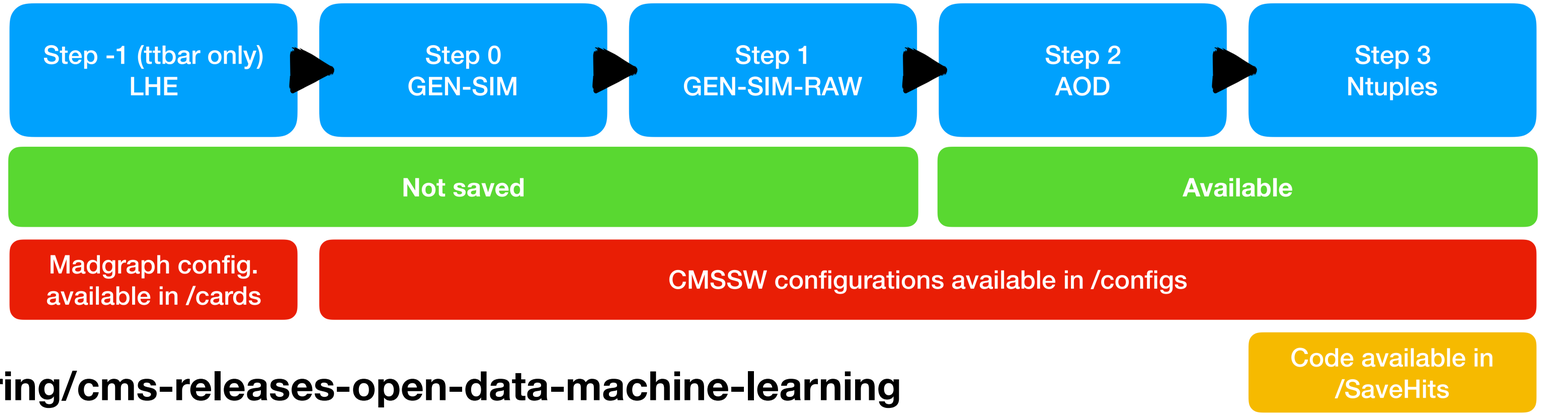


This work uses 8 TeV CMS Open Data

- Essential to access full-simulation, low level detector information
- New release just published focusing on ML application
- Tracker clusters saved in high-level AOD format
- Can reconstruct tracker hits (“RecHits”) on the fly



Full simulation stack reproducible in OD (with code & instructions)



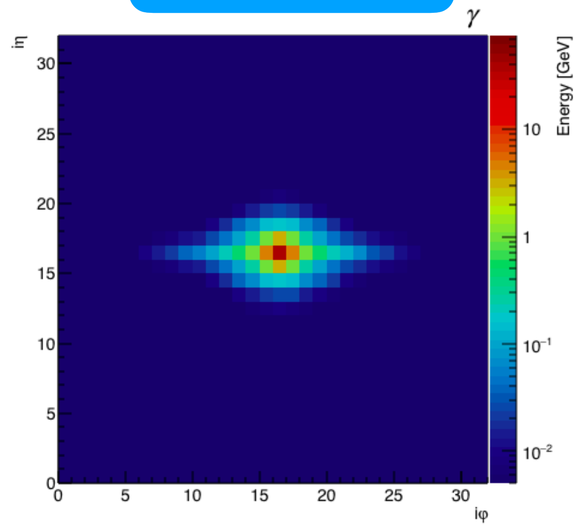
https://home.cern/news/news/knowledge-sharing/cms-releases-open-data-machine-learning

Previous work

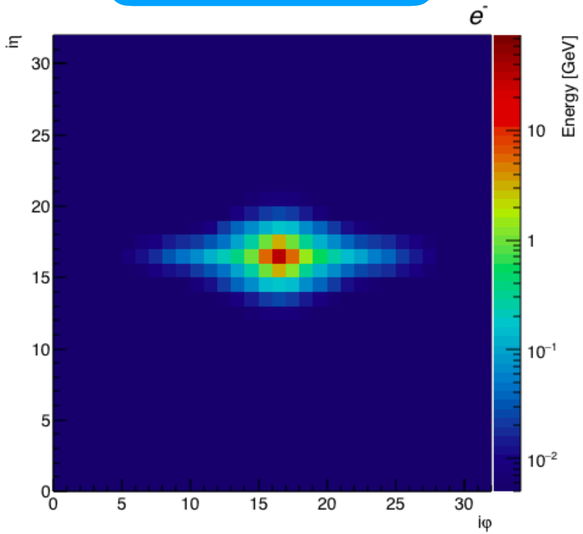
Full CMS detector simulation

Photon vs Electron

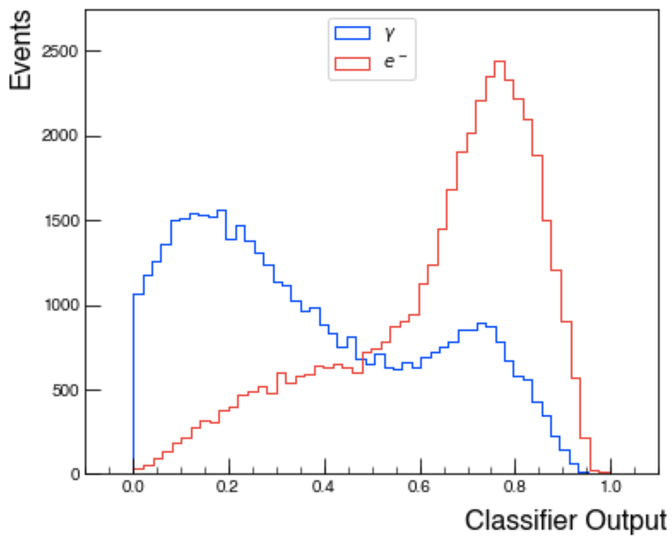
Photon



Electron

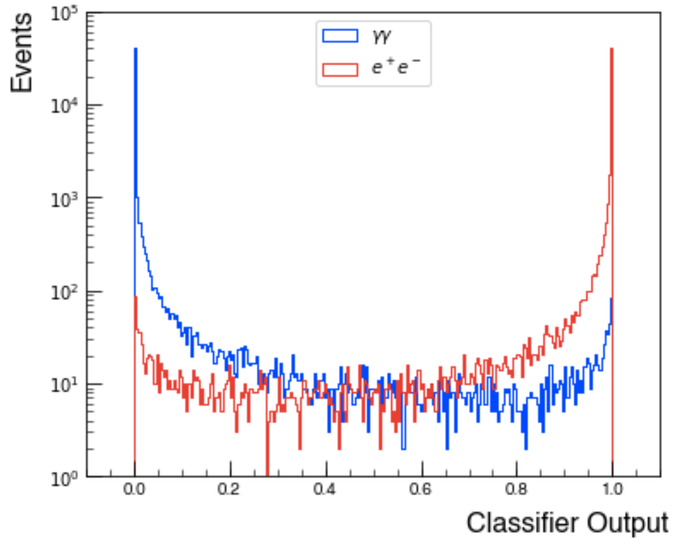


e^- vs γ



ResNet-15	
ROC AUC	0.788

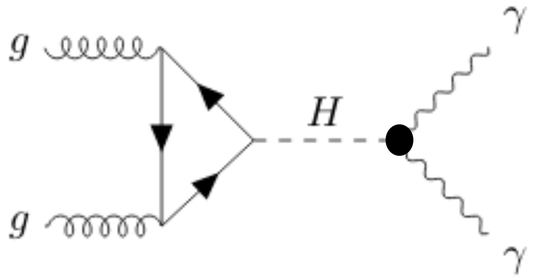
e^+e^- vs $\gamma\,\gamma$

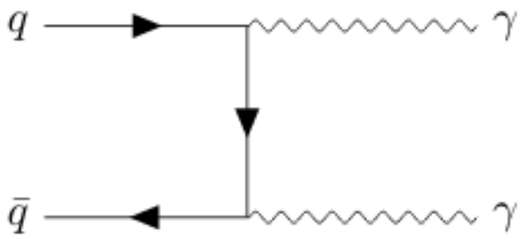


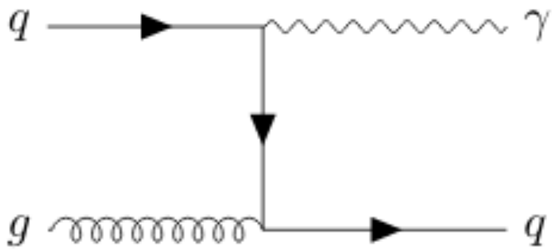
ResNet-23	
ROC AUC	0.997

https://arxiv.org/abs/1807.11916

Event ID: discriminator bias







Resonant diphoton, signal

Nonresonant diphoton

Photon + misidentified jet

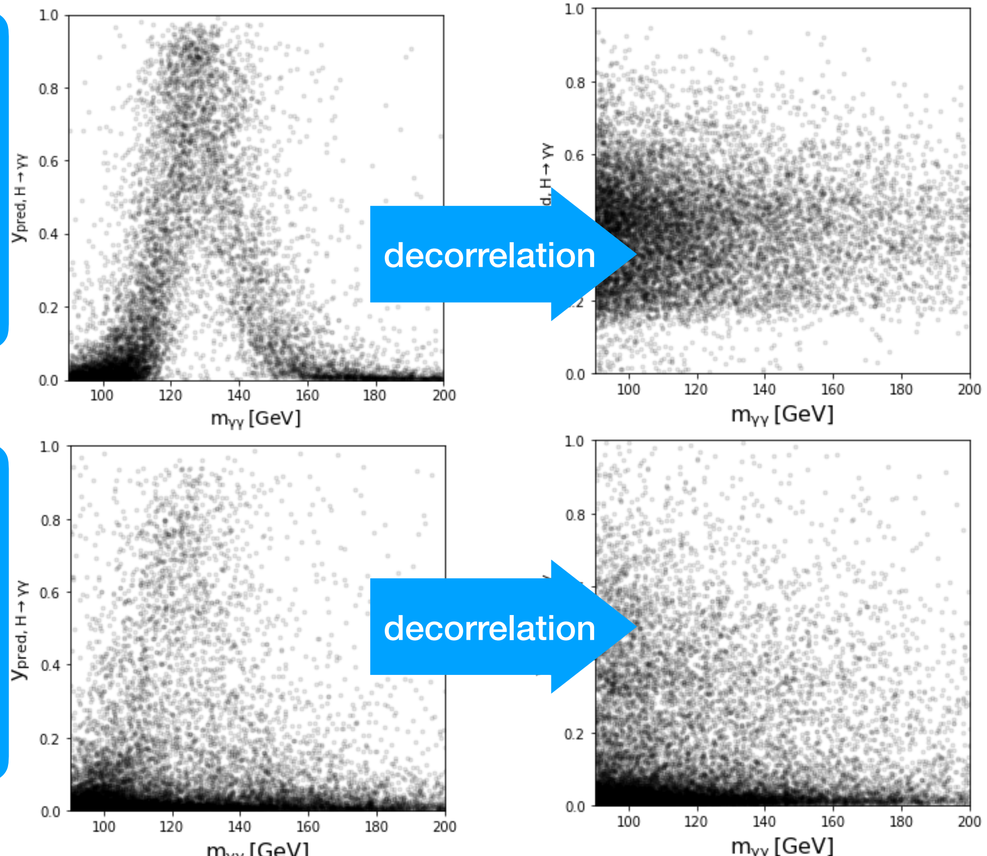
Mass-biased discriminators

Decorrleated discriminators

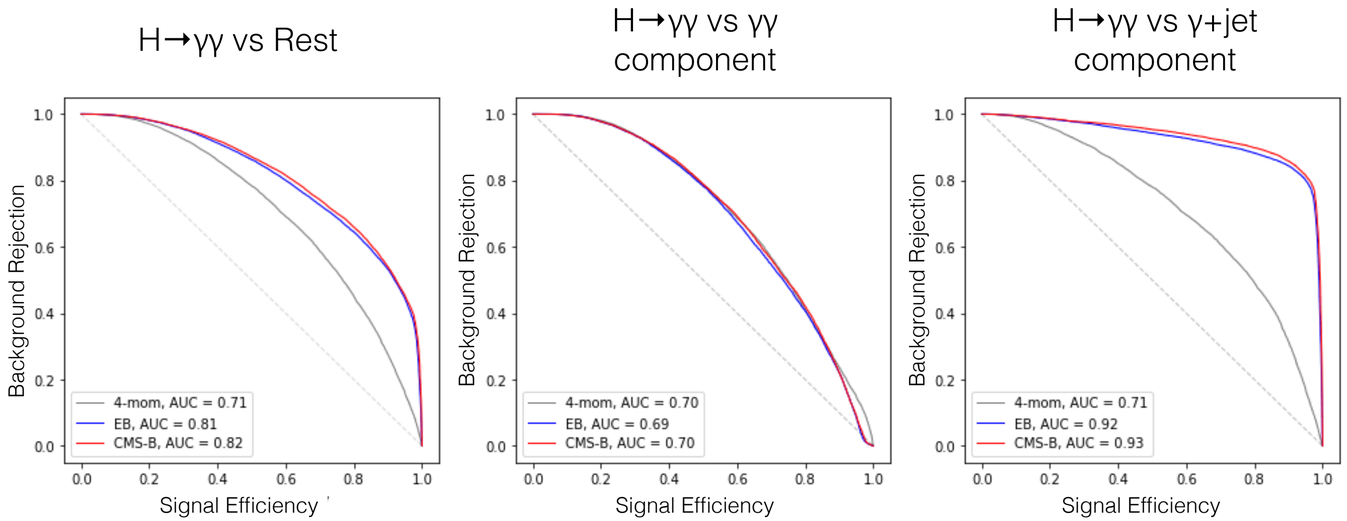
Normalize images to gg mass or sum of pT

signal

background



Final discriminator ROC



AUC **0.82** **0.70** **0.93**

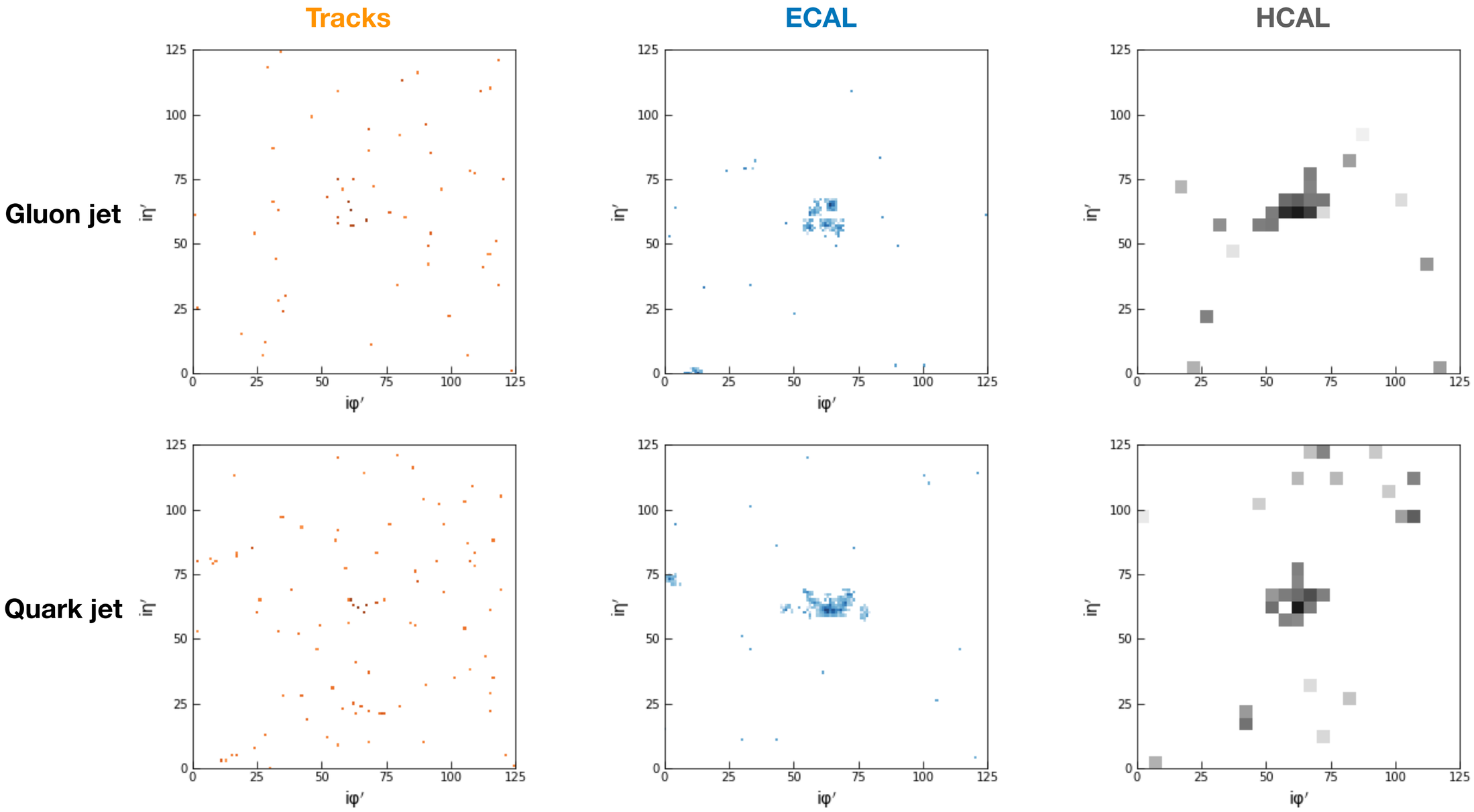
https://arxiv.org/abs/1807.11916

Quark vs gluon jet

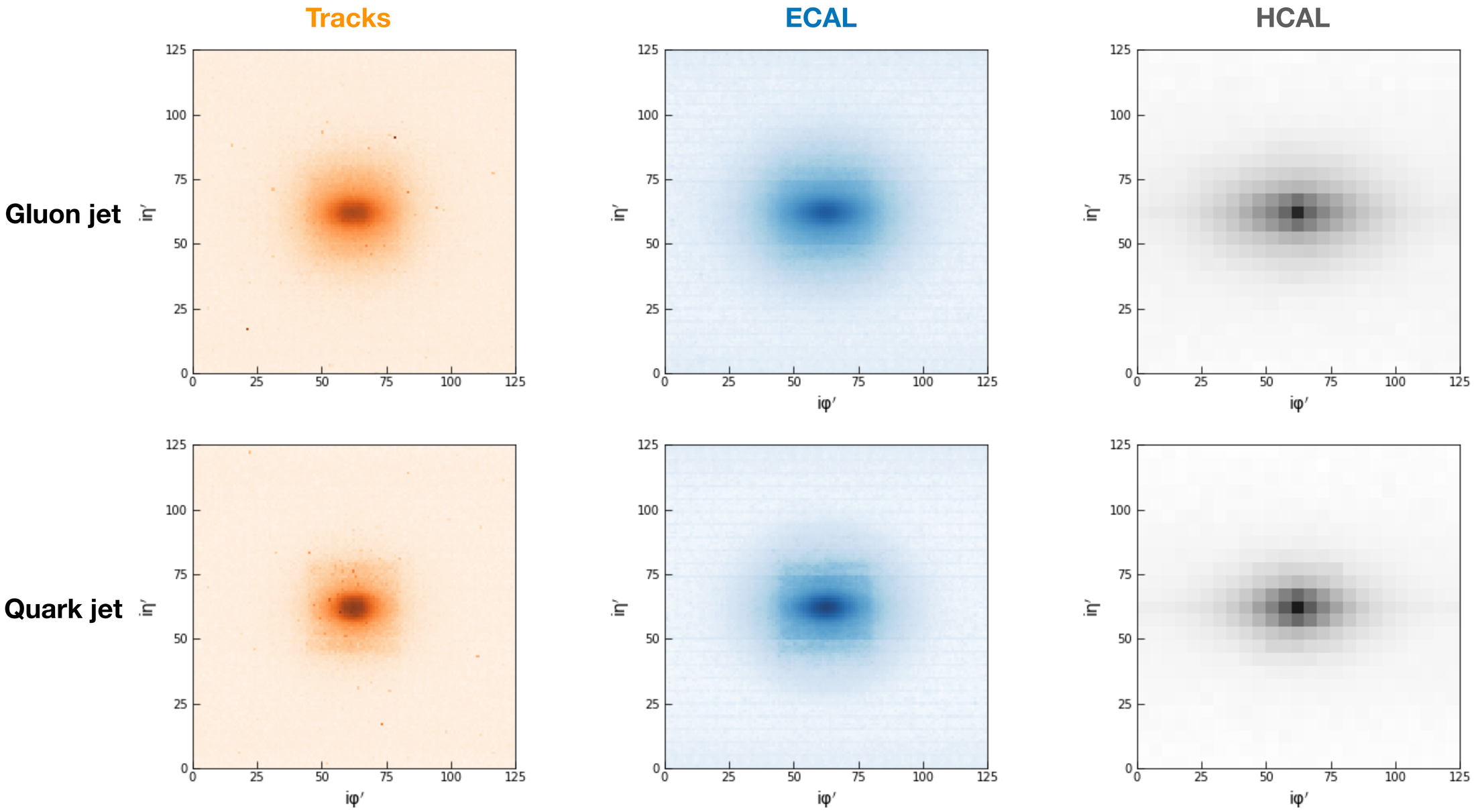
Full CMS detector simulation — open data

arxiv:1902.08276

Quark vs gluon jet

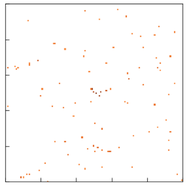


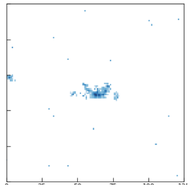
Quark vs gluon jet

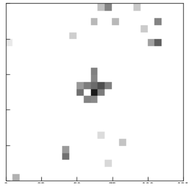


Quark vs gluon jet

Single subdetector



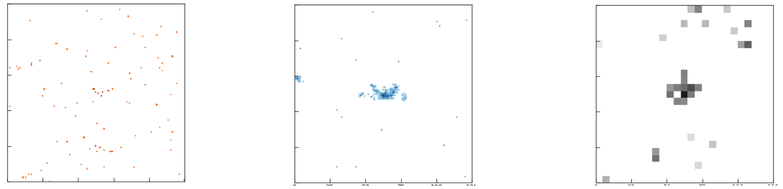




	ROC AUC
E2E jet image, Tracks	0.782
E2E jet image, ECAL	0.760
E2E jet image, HCAL	0.682

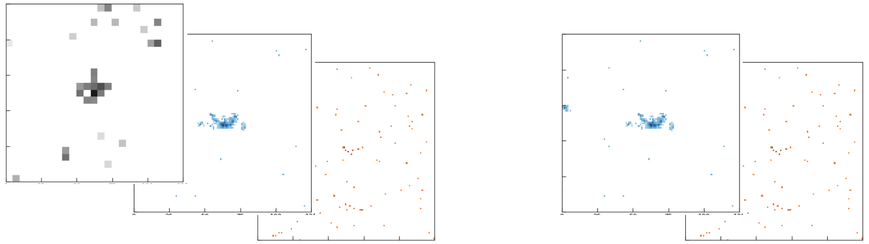
Quark vs gluon jet

Single subdetector



	ROC AUC
E2E jet image, Tracks	0.782
E2E jet image, ECAL	0.760
E2E jet image, HCAL	0.682

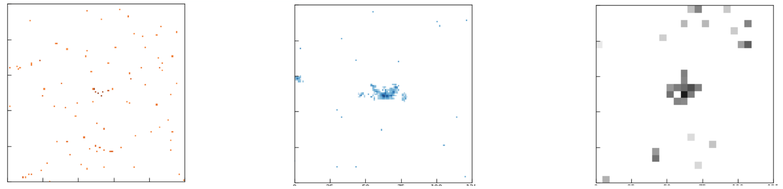
Subsystems combined



	ROC AUC
E2E jet image, ECAL+Tracks	0.804
E2E jet image, Tracks	0.782
E2E jet image, ECAL+HCAL	0.781
E2E jet image, ECAL+HCAL+Tracks	0.808

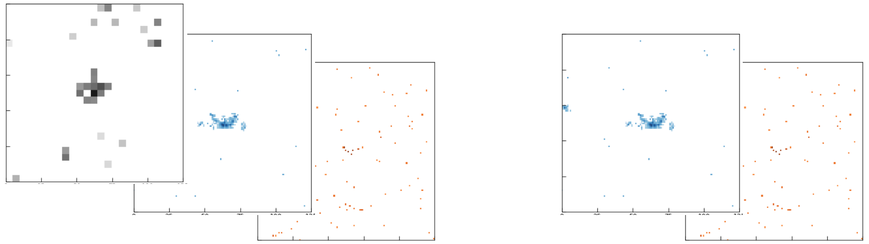
Quark vs gluon jet

Single subdetector



	ROC AUC
E2E jet image, Tracks	0.782
E2E jet image, ECAL	0.760
E2E jet image, HCAL	0.682

Subsystems combined



	ROC AUC
E2E jet image, ECAL+Tracks	0.804
E2E jet image, Tracks	0.782
E2E jet image, ECAL+HCAL	0.781
E2E jet image, ECAL+HCAL+Tracks	0.808

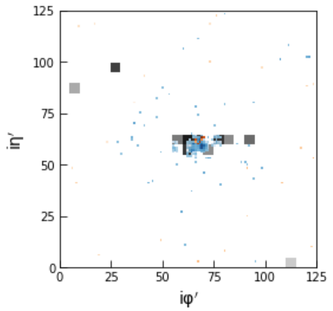
	ROC AUC
E2E image, ECAL+HCAL+Tracks	**0.8077 ± 0.0003***
RecNN, ascending-p_T	0.8017 ± 0.0003*
RecNN, descending-p_T	0.802
RecNN, anti-k_T	0.801
RecNN, Cambridge/Aachen	0.801
RecNN, no rotation/reclustering	0.800
RecNN, k_T	0.800
RecNN, k_T-colinear10-max	0.799
RecNN, random	0.797
Traditional jet images	0.720

Comparison with RecNN

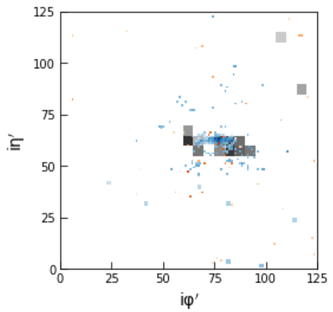
- **RecNN, Jet ID for QCD vs boosted *W* jet**
  - *K. Cranmer et al.:* https://arxiv.org/pdf/1702.00748.pdf
  - DELPHES detector simulation
  - Applied to quark vs gluon by *T. Cheng:* https://arxiv.org/pdf/1711.02633.pdf
  - **Traditional jet images perform less well than 4-momenta**

Qq vs gg

Scenario A: 2 separate images



ResNet-15,
convolutional output

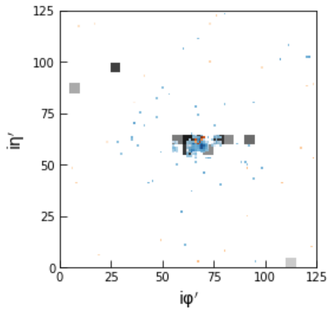


ResNet-15,
convolutional output

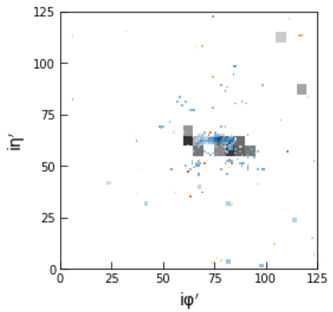
Fully-connected, 128 x 2

Qq vs gg

Scenario A: 2 separate images



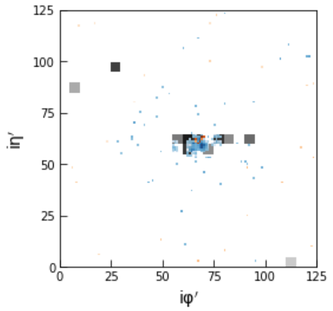
ResNet-15,
convolutional output



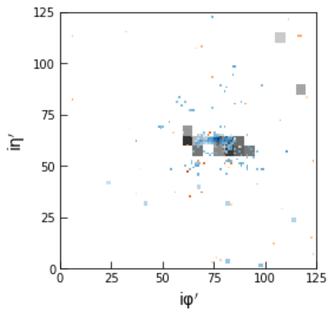
ResNet-15,
convolutional output

Fully-connected, 128 x 2

Scenario B: A + dijet 4-momenta



ResNet-15,
convolutional output



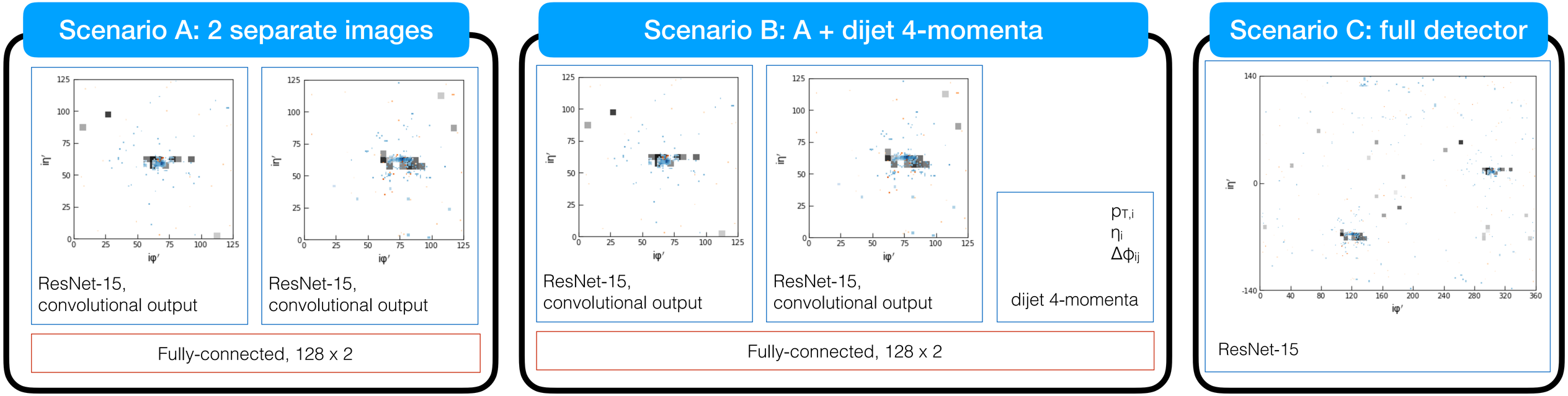
ResNet-15,
convolutional output

$p_{T,i}$
η_i
$\Delta\phi_{ij}$

dijet 4-momenta

Fully-connected, 128 x 2

Qq vs gg



- **Local or global physics?**
  - Performance dominated by jet-level differences (Scenario A vs. B or C)
  - Both dijets are non-resonant decays, so jet 4-momenta doesn't hold much discrimination power (Scenario B vs. A)
  - **Fully E2E approach (Scenario C) picking up on subtle, event-level effects not captured by either B or A.**

	ROC AUC
Scenario A	0.876
Scenario B	0.878
Scenario C	0.889

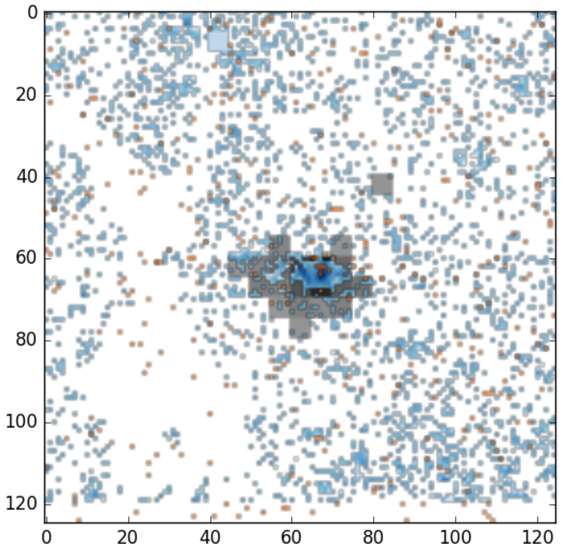
Top quark ID

Full CMS detector simulation — open data

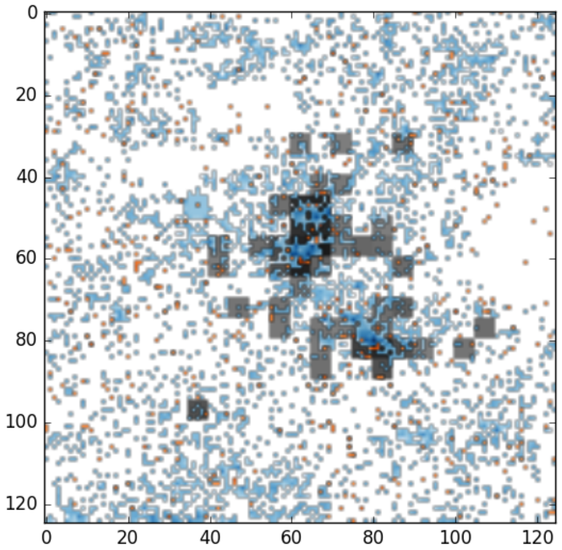
Top tagging

600<pT<800 GeV

Non-top jet



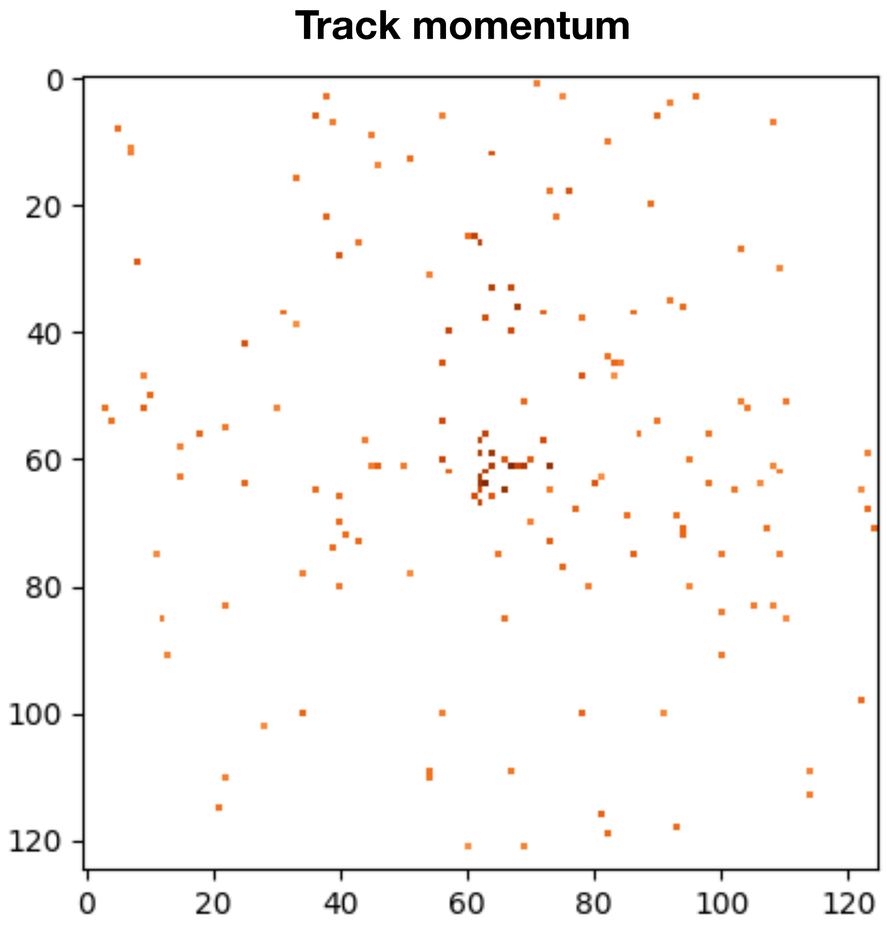
Top jet



- ~ 5M top-antitop pair events
- Transverse momentum > 400GeV, |eta|<2.4
- Natural pT distribution from SM top-antitop
- Non-top jets sampled in from same momentum distribution as top quark

More track information

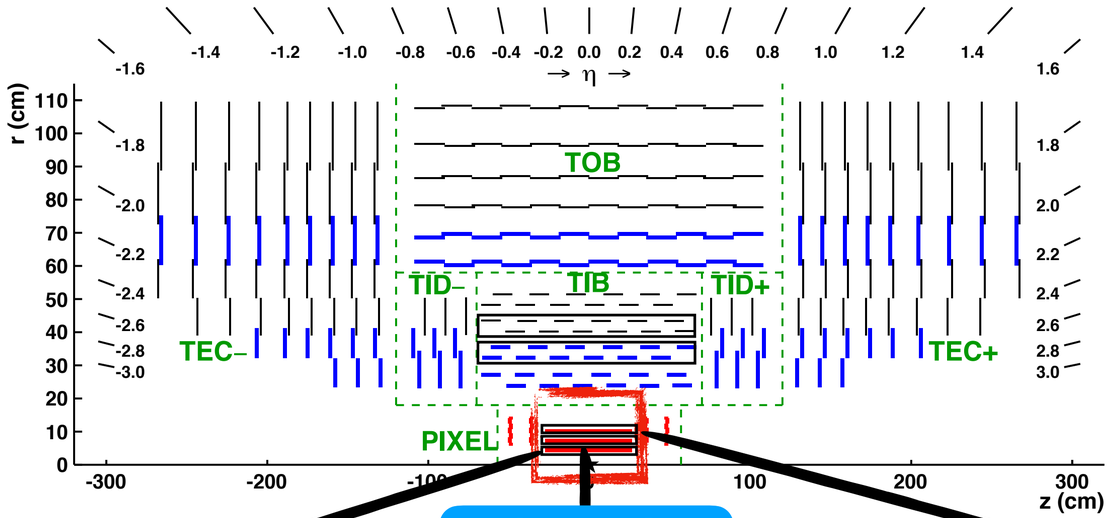
- track position weighted by a different variable:
  1) Transverse Momentum
  2) d0 impact parameter
  3) dz impact parameter



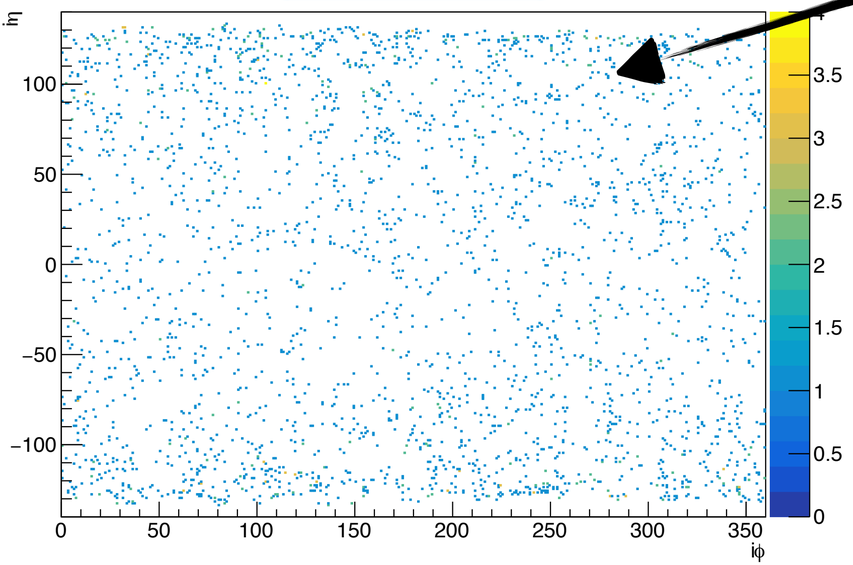
Tracking rechits

Add information tracker/muon information beyond reconstructed tracks

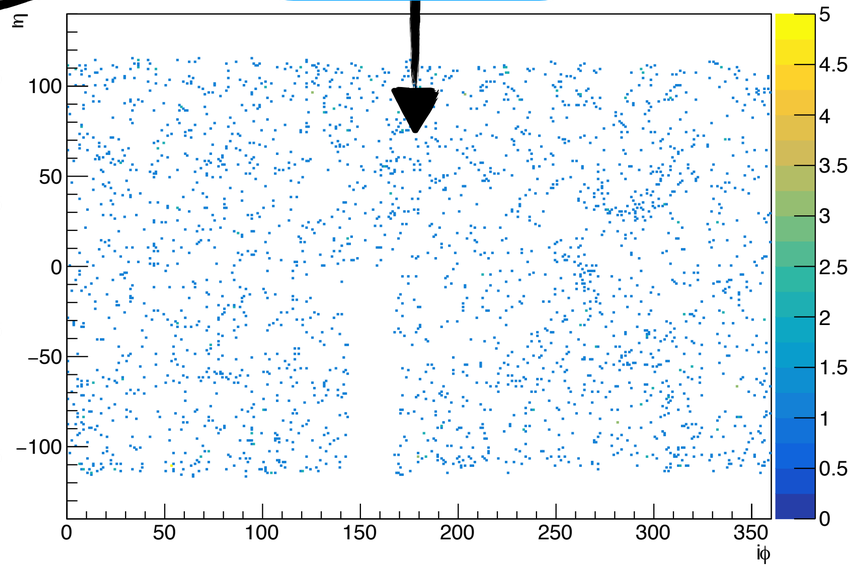
Pixel Barrel



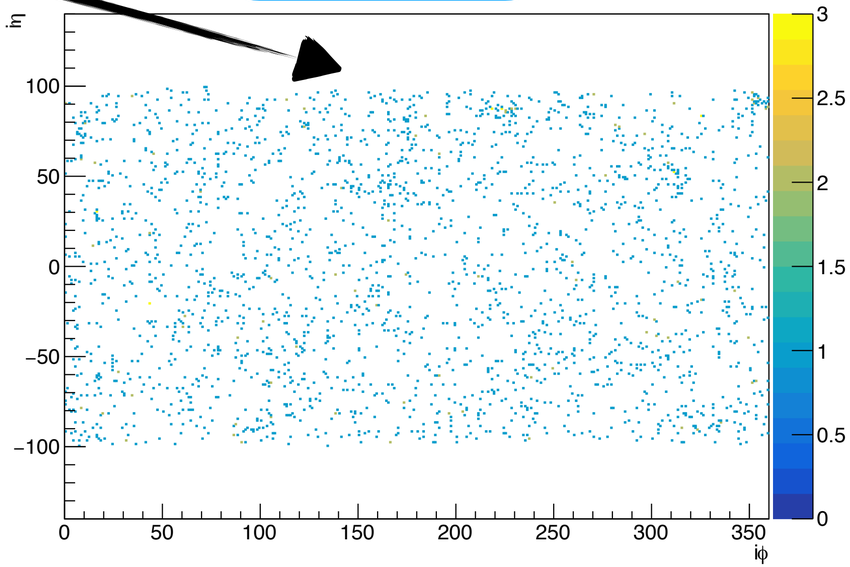
Layer 1



Layer 2



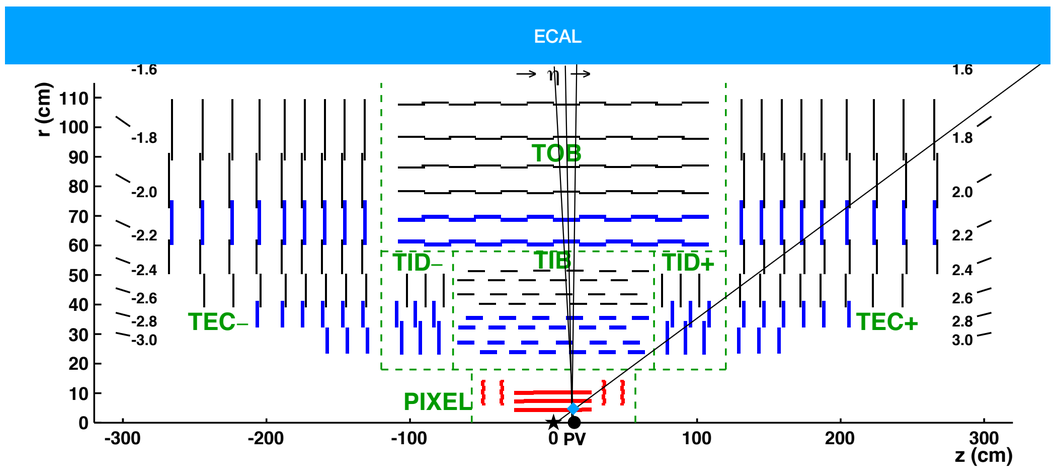
Layer 3

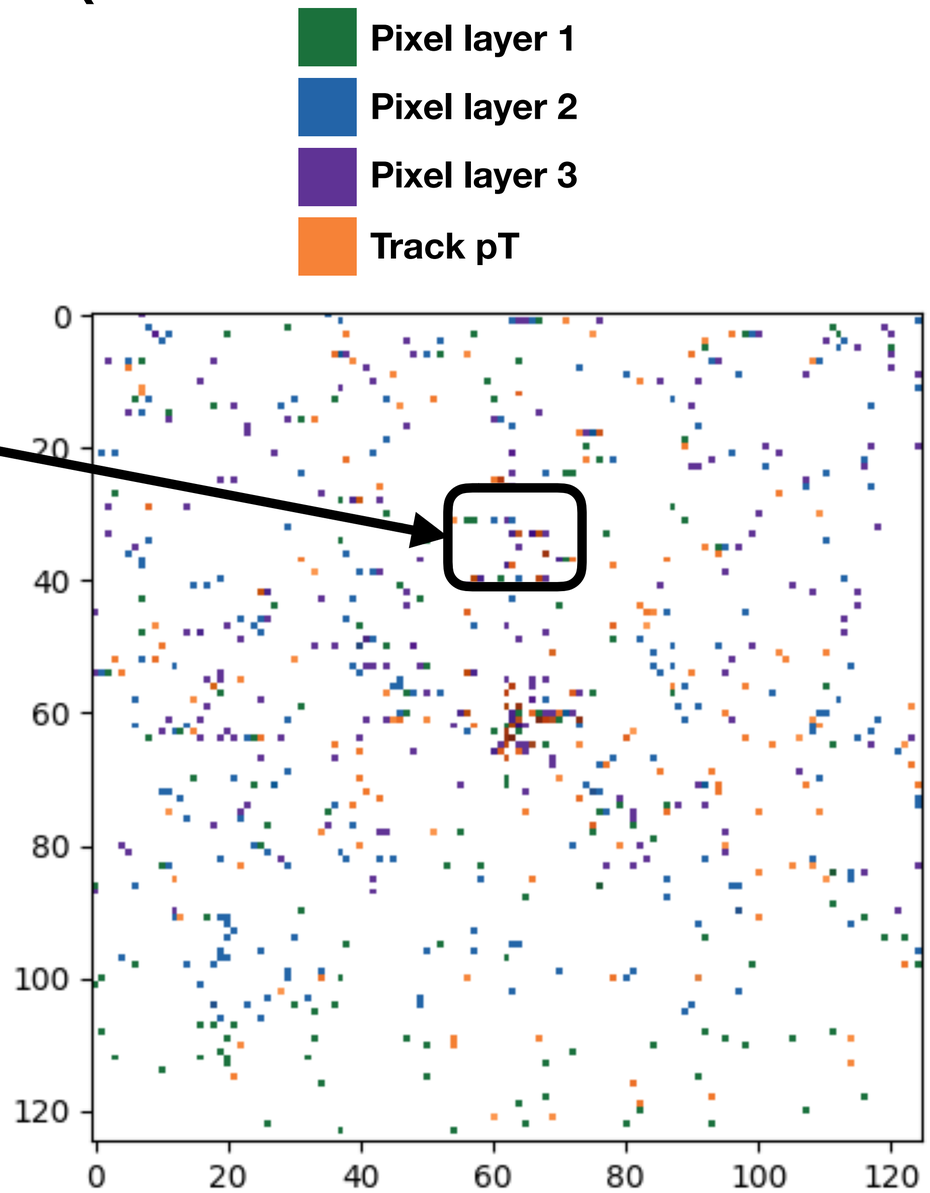


Jet Images - track pT (at ECAL surface)

low-pT tracks bending

Pixel detector very close to beam line
Must correct for PV position



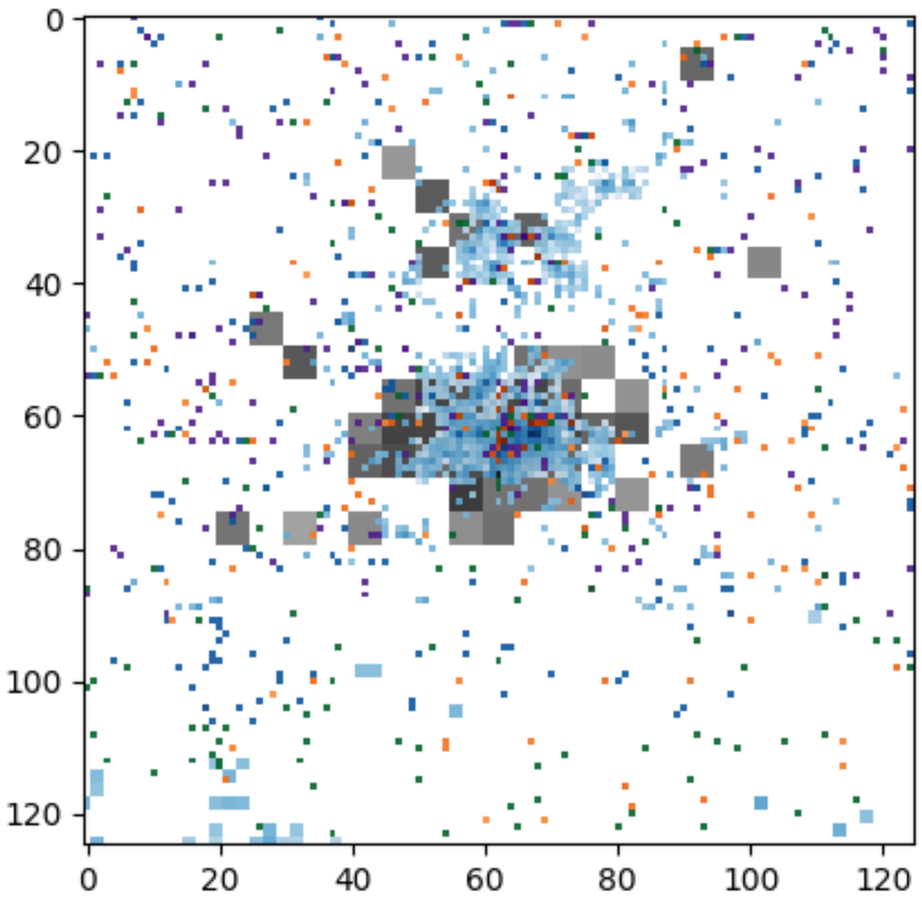


Jet Images - All Image Channels

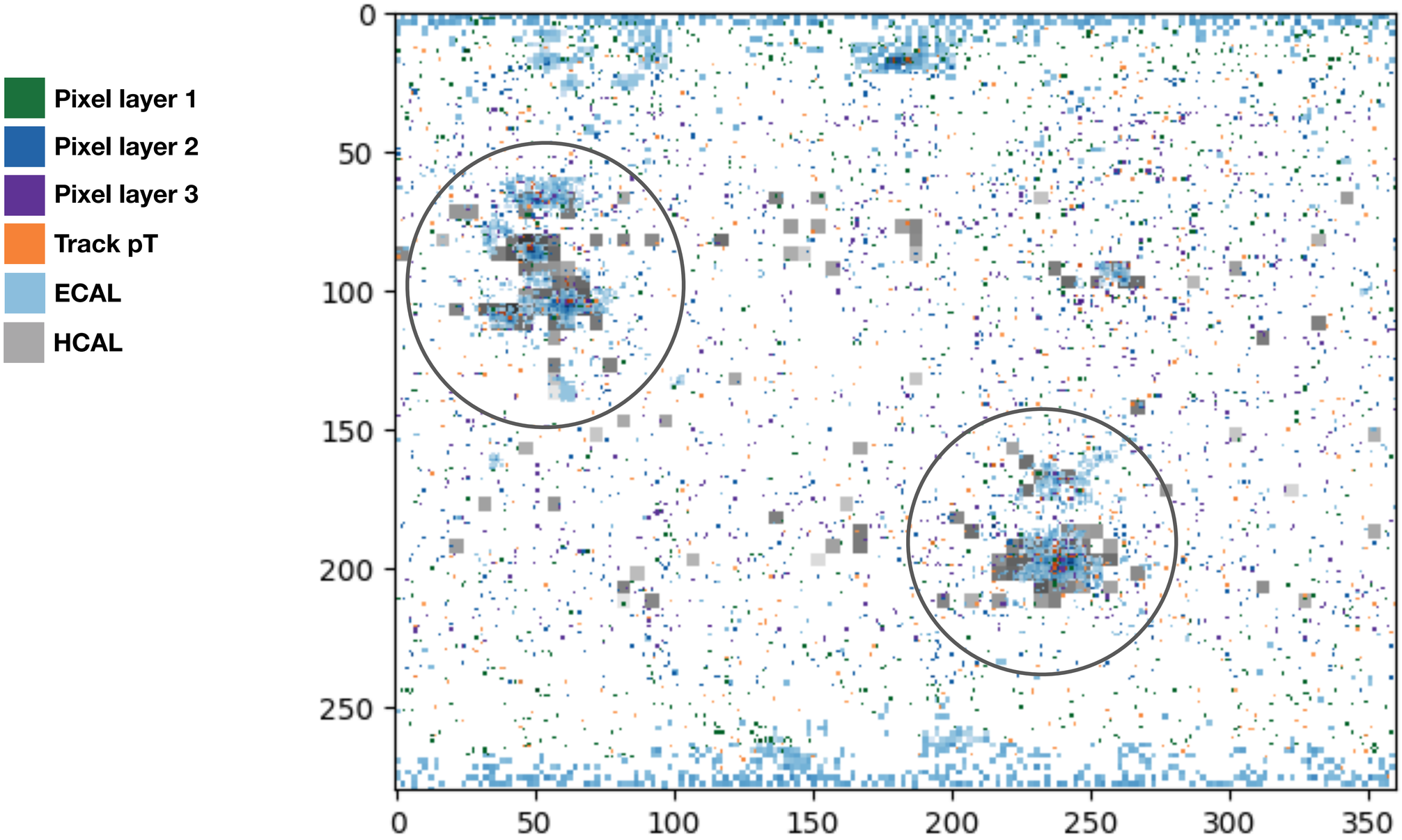
final image for a 628 GeV top jet.

list of the image channels:

1. Pixel layer 1 rechits
2. Pixel layer 2 rechits
3. Pixel layer 3 rechits
4. pT weighted tracks
5. d0 weighted tracks
6. dz weighted tracks
7. ECAL rechits
8. HCAL rechits



Full Detector Image



Summary

E2E Particle ID:

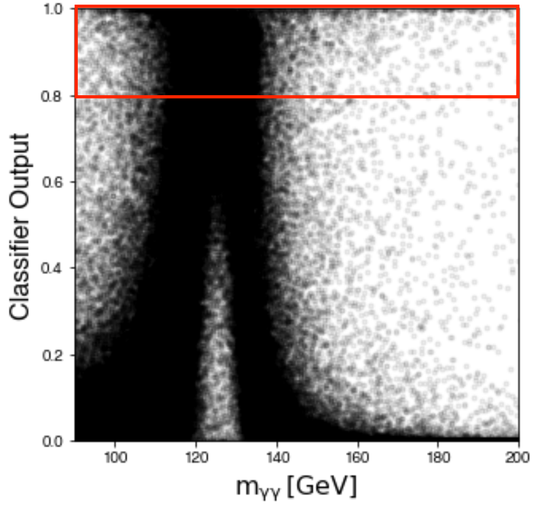
- Able to learn particle kinematics and shower shapes

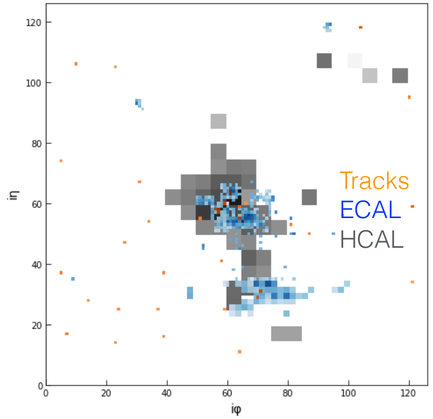
E2E q vs. g Jet ID:

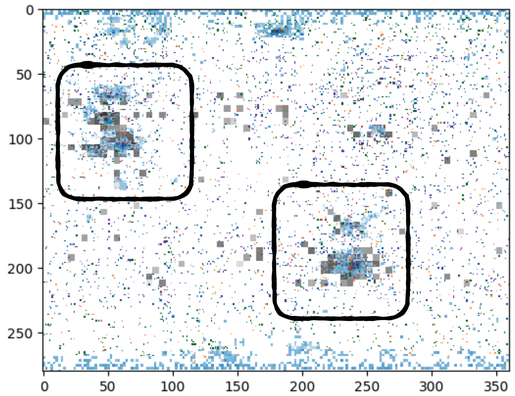
- Competitive with existing state-of-the-art jet ID classifiers
- E2E approach exploits full detector performance
- Event ID Captures event-level correlations lost at jet-level.

E2E Top ID:

- Work in progress
- Adding more tracking information and tracker rechits
- Increase resolution
- Expect results soon







 https://arxiv.org/abs/1902.08276 https://arxiv.org/abs/1807.11916

Backup

Related Work

- CNNs in various neutrino experiments
  (A. Aurisano et al., see DS@HEP 2017, IML Workshop 2017, 2018)
- Particle ID CNNs on 4-momenta of jet constituents *(Luke de Oliveira et al., see DS@HEP 2017, IML Workshop 2017, 2018)*
- RNNs on 4-momenta of jet constituents a la Nat. Lang. Proc.
  (Kyle Cranmer et al., Jean-Roch Vlimant et al.)
- Particle ID CNNs on photon cluster detector data
  (Andre Holzner et al.)
- Event ID CNNs on whole detector images
  (*Wahid Bhimji et al.*)
- **Our approach emphasizes high detector fidelity:**
  True detector-level data, Geant4 detector sim, most accurate CMS model. *Results representative of real physics analysis!*

Credit: Michael Andrews

Particle ID I: Network

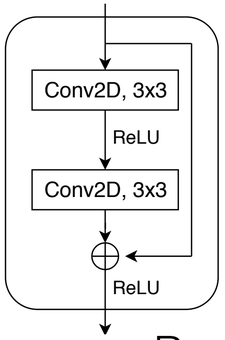
- **Choose best-in-category for each of:**
  - Convolutional NN (**CNN**): VGG, Inception, ResNet
  - Conv-LSTM (**LSTM**): TimeDist(CNN)→LSTM, LSTM(CNN)
  - Fully-Connected NN (**FCN**): 2-, 3-, 6-hidden layers, 256 nodes
- **Try a variety of inputs:**
  - energy, (energy, time), (DIGI), (energy, time, DIGI)*
- **Try different concatenation schemes:**
  - @input, @convolutional output, @FC output

Credit: Michael Andrews

Particle ID II: Network

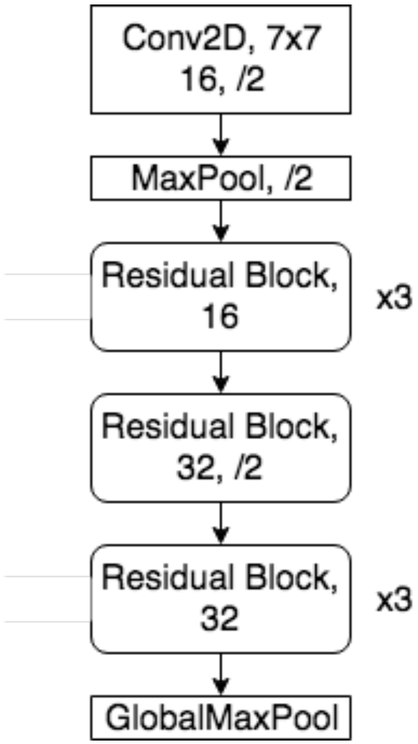
- **VGG does not scale with image size**
  - Scales with CNN output *volume* x FCN nodes causing weights explosion!
  - VGGs also subject to *degradation* with increasing depth

- **Residual Nets scale much better**
  - Scales with CNN output *layers,* no need for FCN
  - Skip connections mitigate *degradation* with depth



Residual Block

ResNet-15



Credit: Michael Andrews

Jet ID | quark vs gluon

- **CMS OpenData QCD Samples**
  - Leading jet from QCD dijet qq' (*uds*) or gg, EMenriched @ 8 TeV
  - CMS GEANT4 full detector simulation, PTYHIA 6
  - $\hat{p}_T$: 80-170 GeV, reco p_T > 70 GeV, $|\eta| < 1.8$
  - Run-dependent ⟨PU⟩ : 18-21
  - Produced and ntuplized with CMSSW 5_3_32
  - **Sample split:**
    - Training set: 576k jets (of which, 26k jets for validation)
    - Test set: 139k jets
    - Balanced samples per class
    - Balanced PU representation per class
  - **Architecture:** ResNet-15 trained from scratch on an NVIDIA Titan X/p using Pytorch 0.4

Credit: Michael Andrews