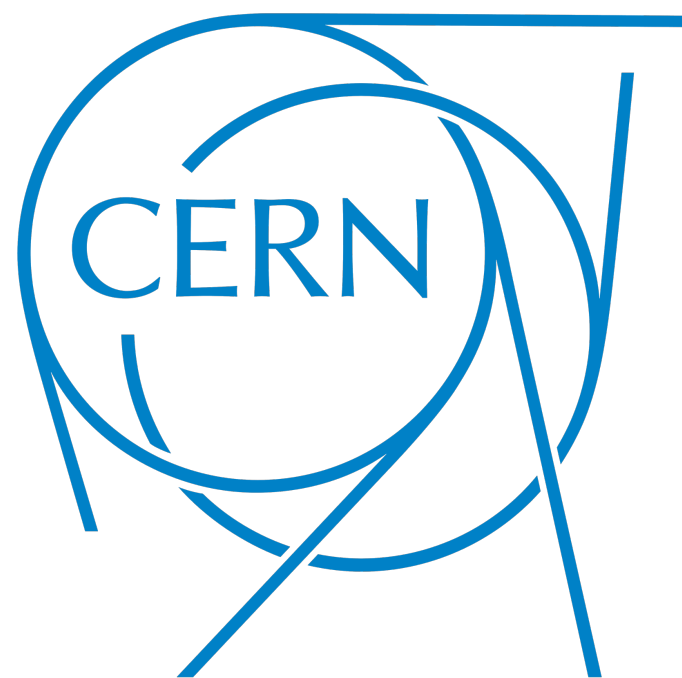


ETH zürich

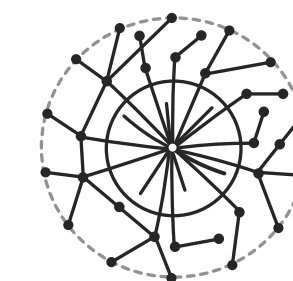
 **Fermilab**

IMPERIAL



R

FP



NextGen
Next Generation Triggers

COLLIDE-2V

750 Million Dual-View LHC Event Dataset for Low-Latency ML



Eric A. Moreno, Philip Ploner, Ranit Das, Jennifer Ngadiuba, Abhijith Gandrakota, Benedikt Maier, Thea Aarrestad, Maciej Glowacki, Shiqi Kuang, Philip Harris, Javier Duarte, David Shih, Alex Tapper, Mia Liu



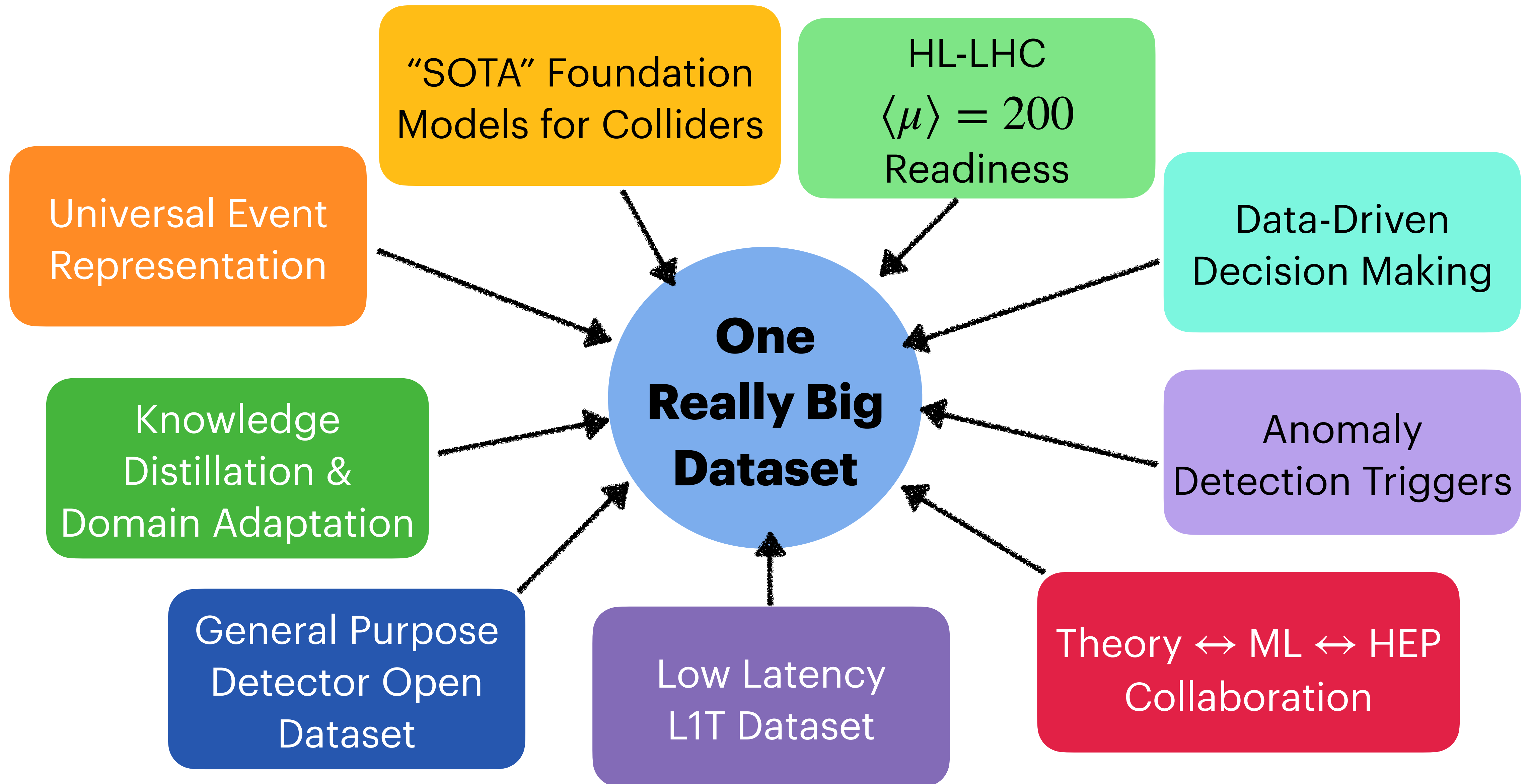
**Phil
Harris
Lab@MIT**

TL;DR

Public CernBox

HuggingFace

Motivations

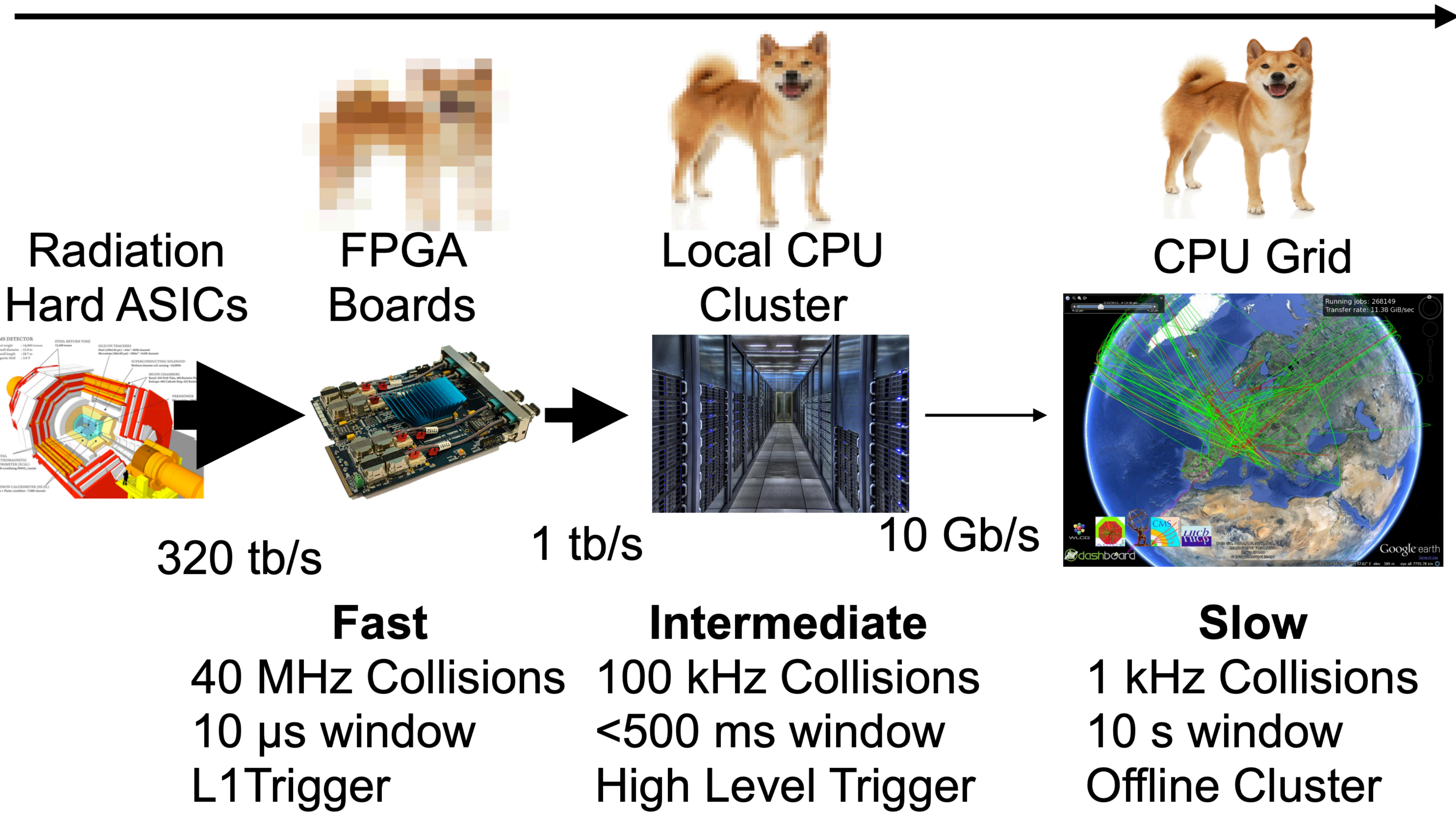




Quick Background CMS L1T

40 MHz

1 kHz



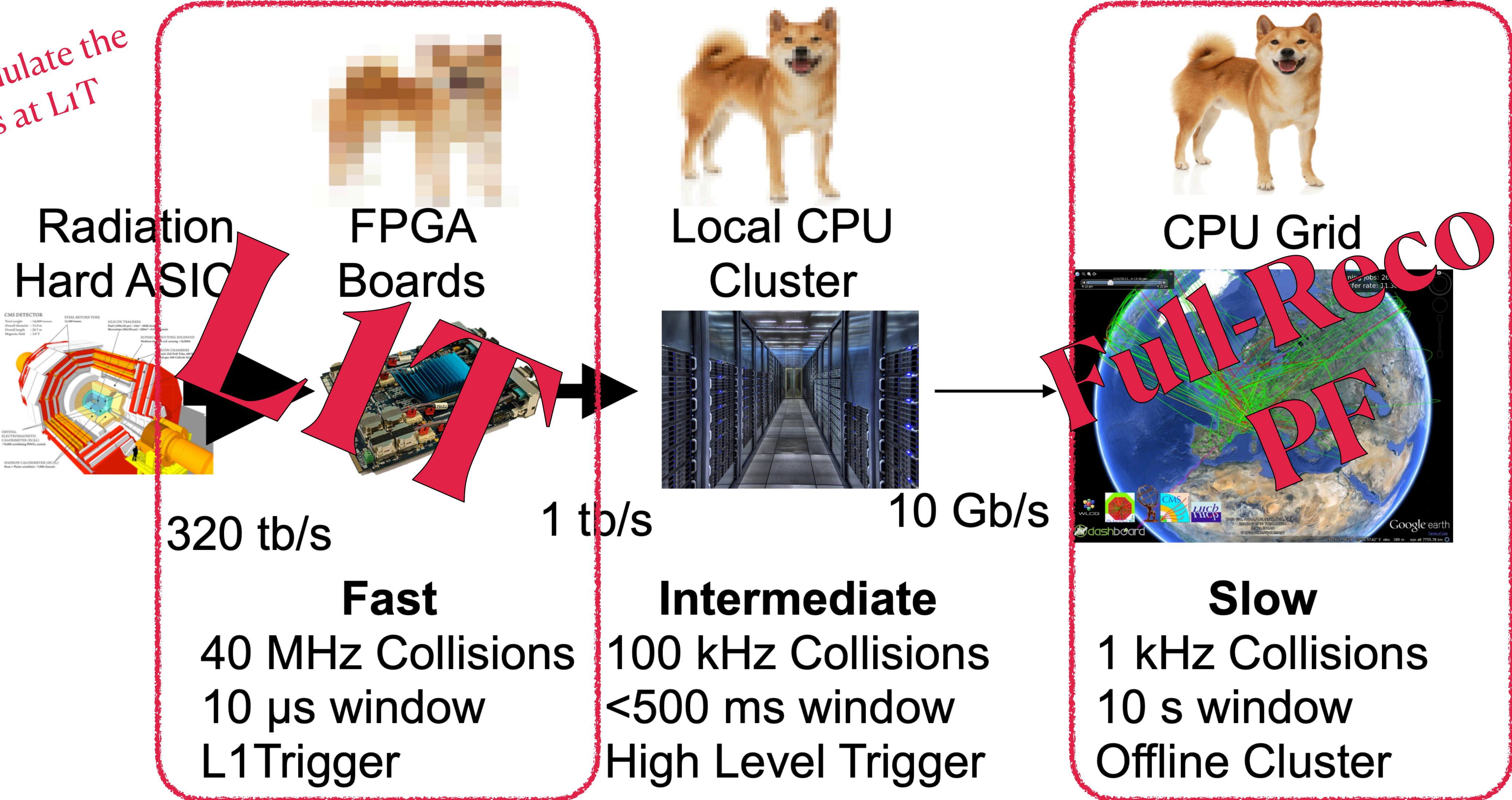


Quick Background CMS L1T

40 MHz

1 kHz

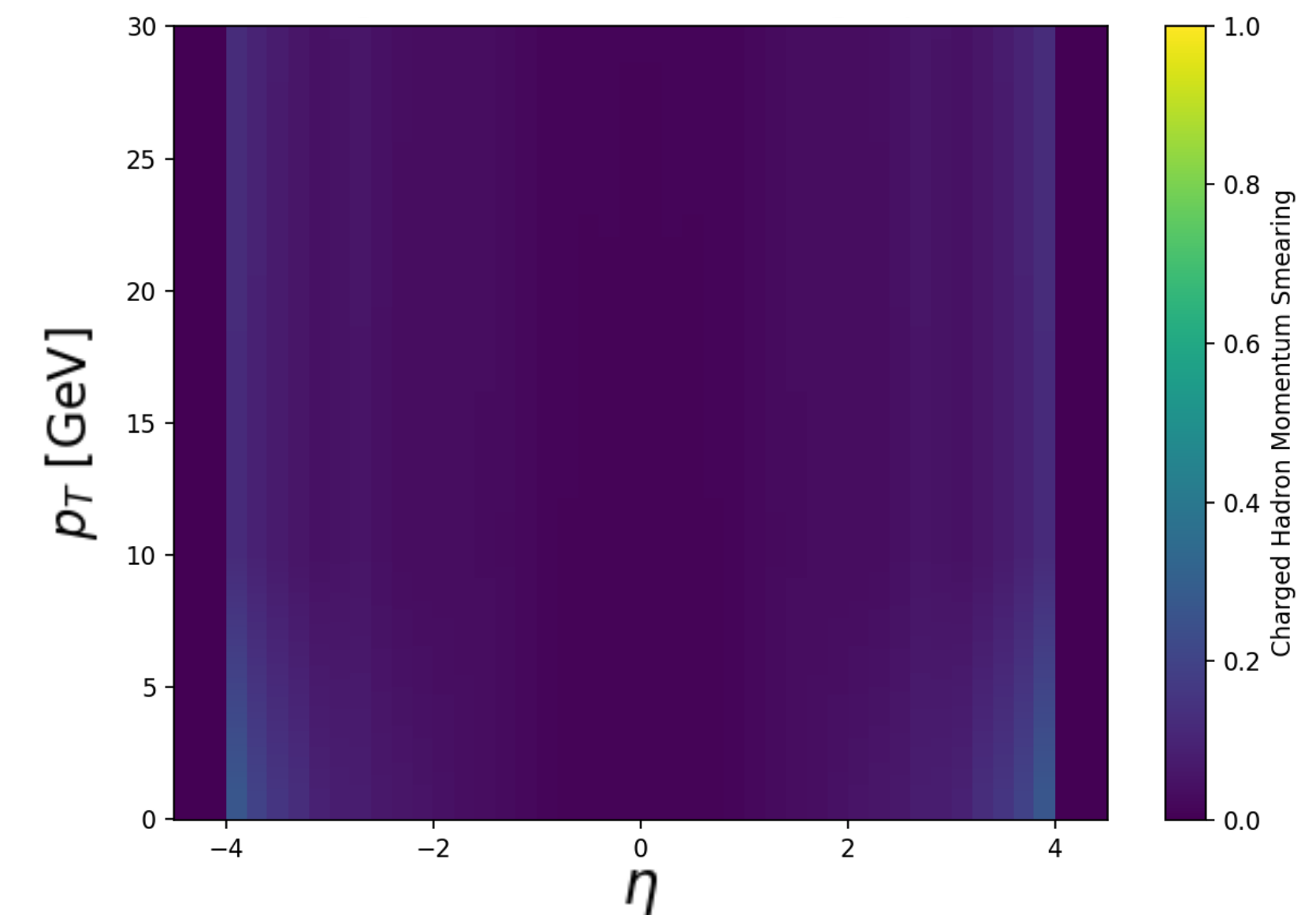
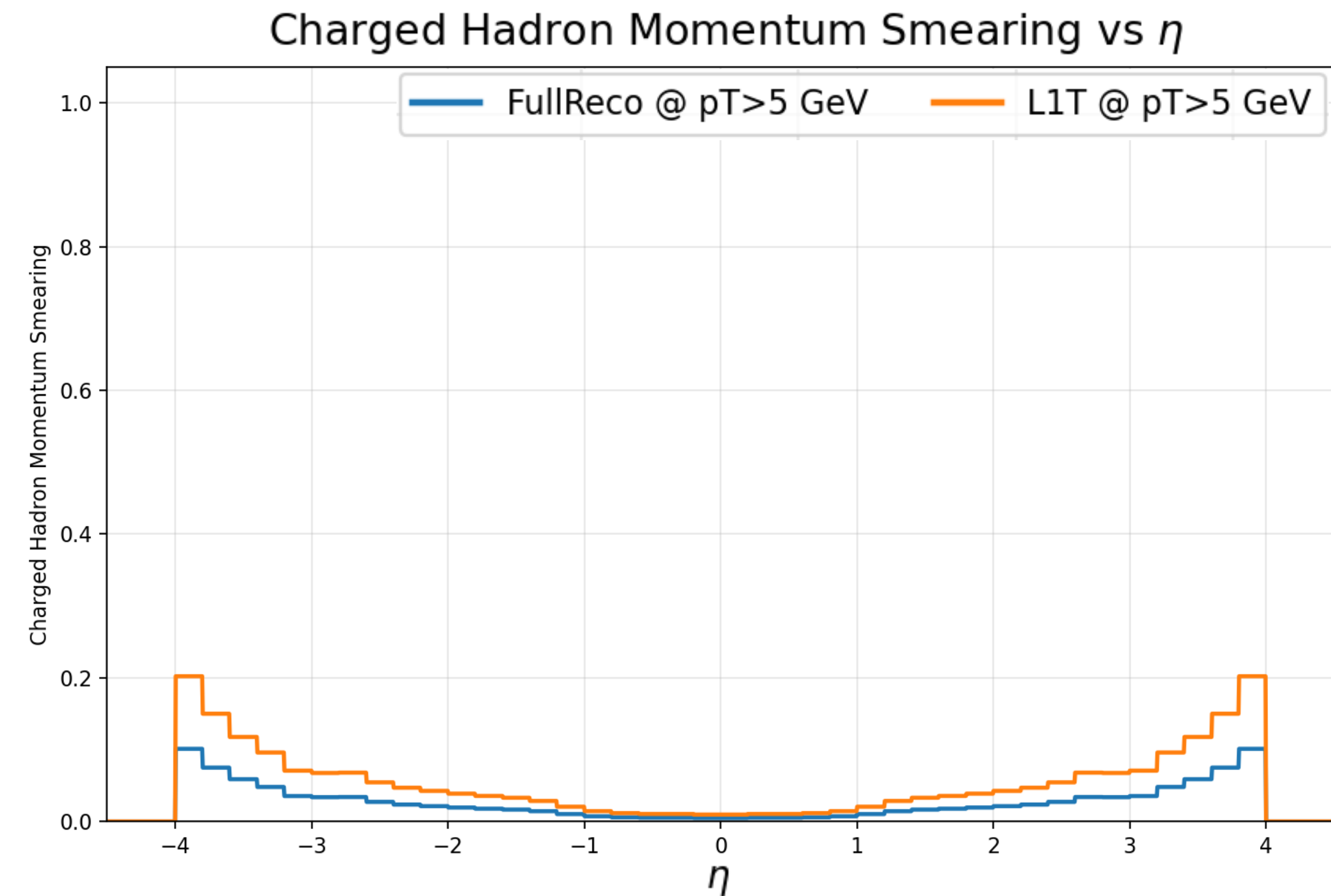
Want to emulate the conditions at L1T



Custom Delphes L1T

From the CMS L1T TDR & Phase-2 TDR

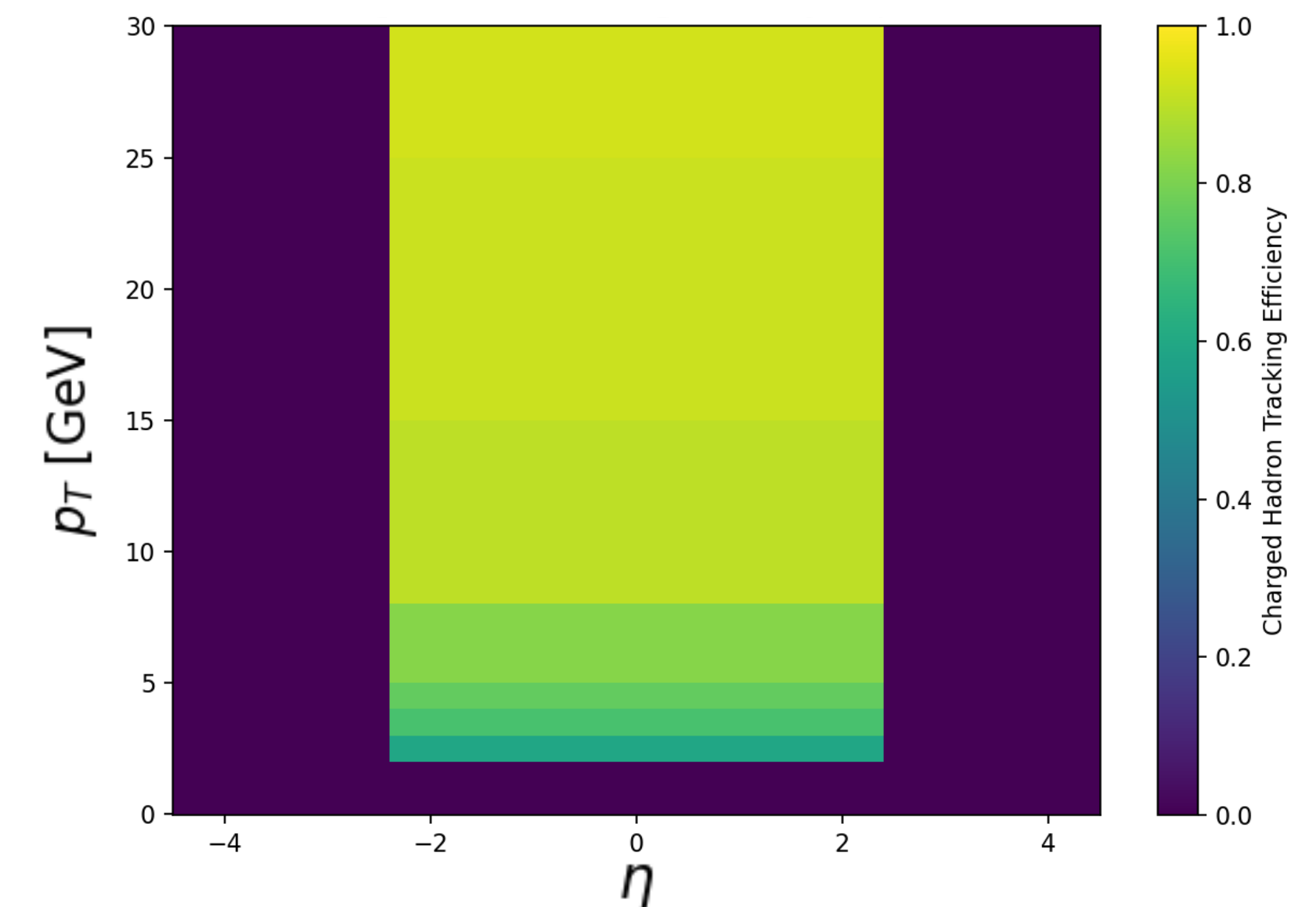
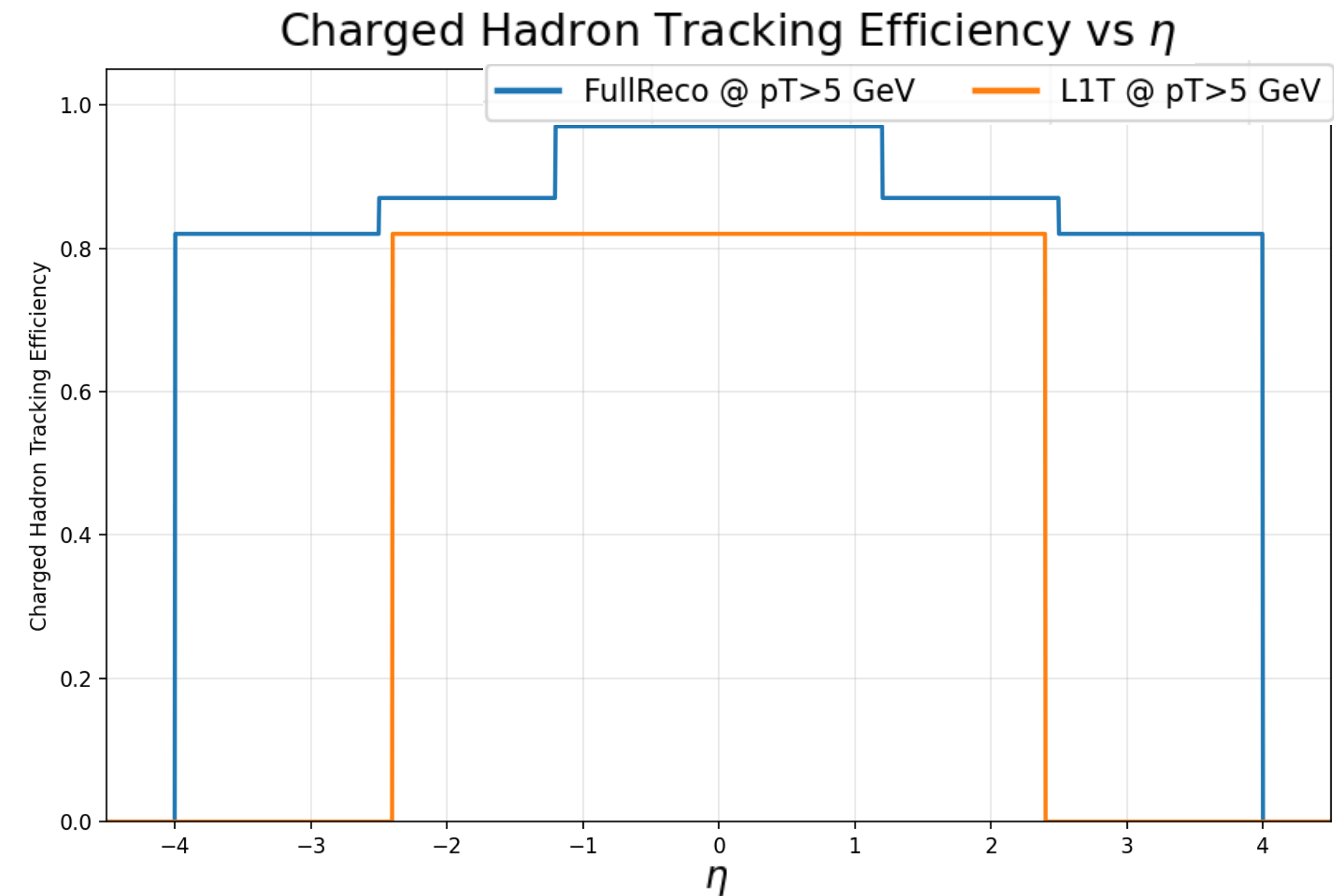
- Main differences between Full-Reco and CMS L1T
 - **Worse vertex resolution**
 - **Worse track resolution**
- Tracking can go up to eta 2.5
- Track efficiency goes to zero below 2 GeV
- PUPPI algorithm (PUPPI-tune) different
- Muons, barrel, calorimeters perform similarly



Custom Delphes L1T

From the CMS L1T TDR & Phase-2 TDR

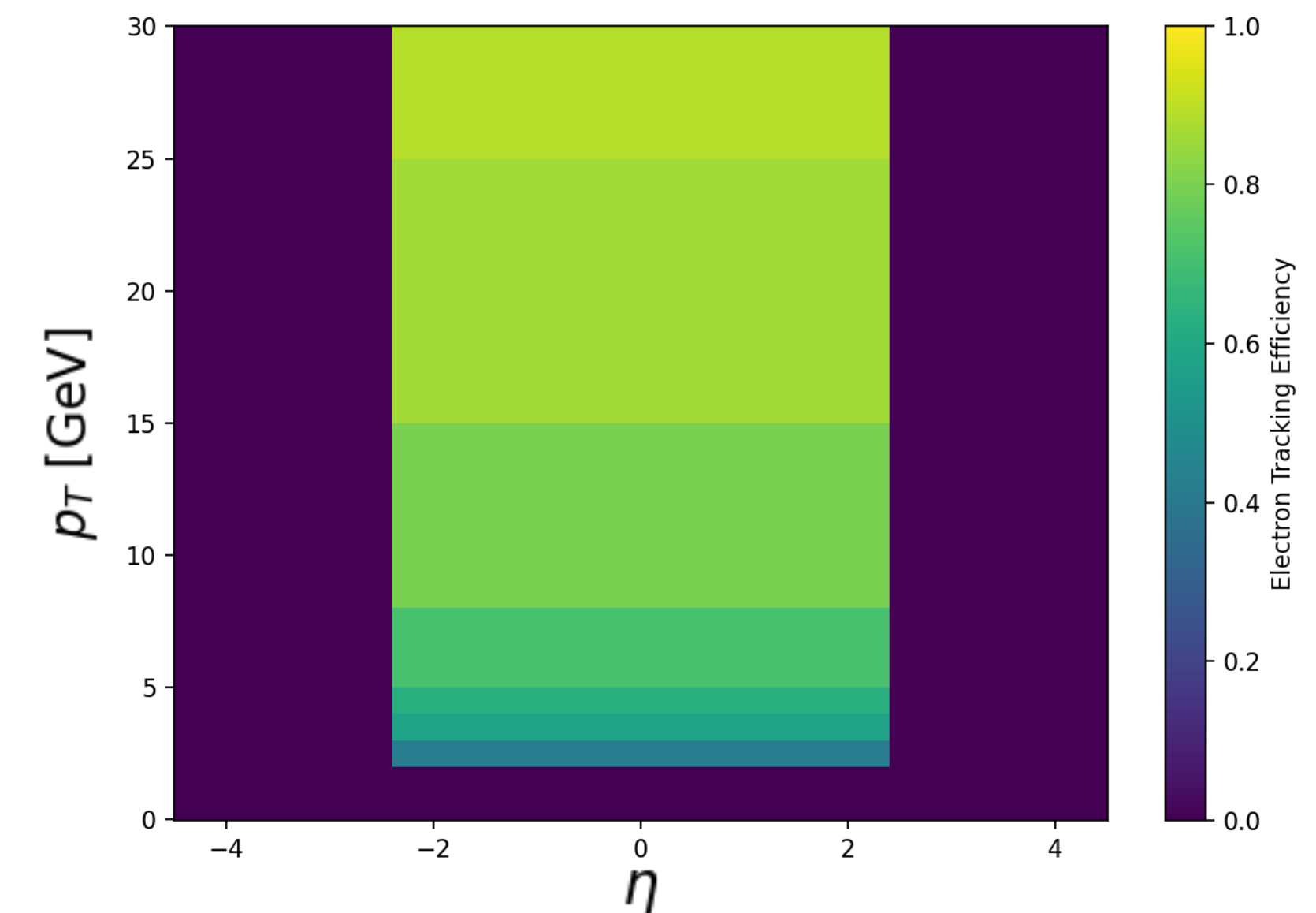
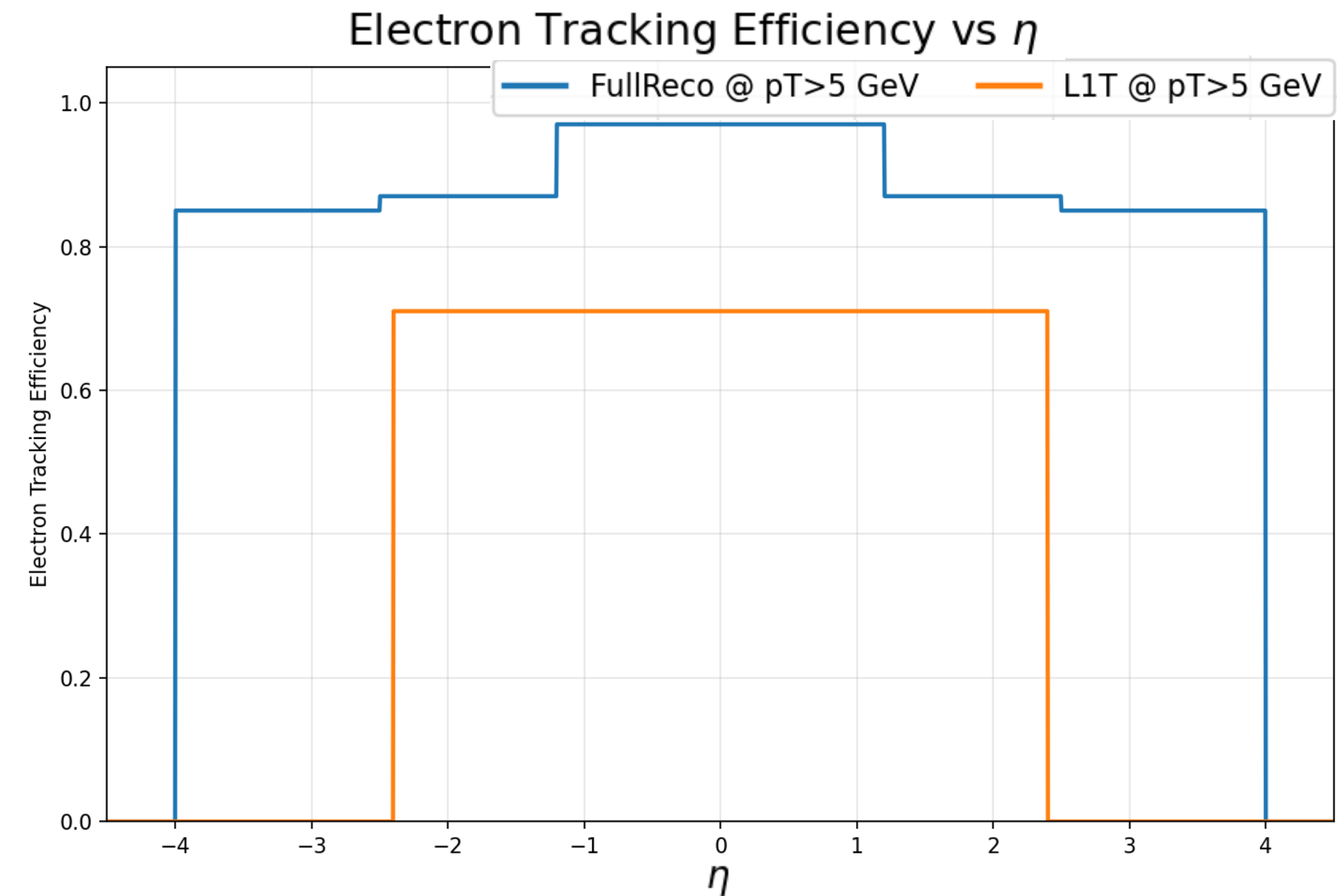
- Main differences between Full-Reco and CMS L1T
- Worse vertex resolution
- Worse track resolution
- **Tracking can go up to eta 2.5**
- **Track efficiency goes to zero below 2 GeV**
- PUPPI algorithm (PUPPI-tune) different
- Muons, barrel, calorimeters perform similarly



Custom Delphes L1T

From the CMS L1T TDR & Phase-2 TDR

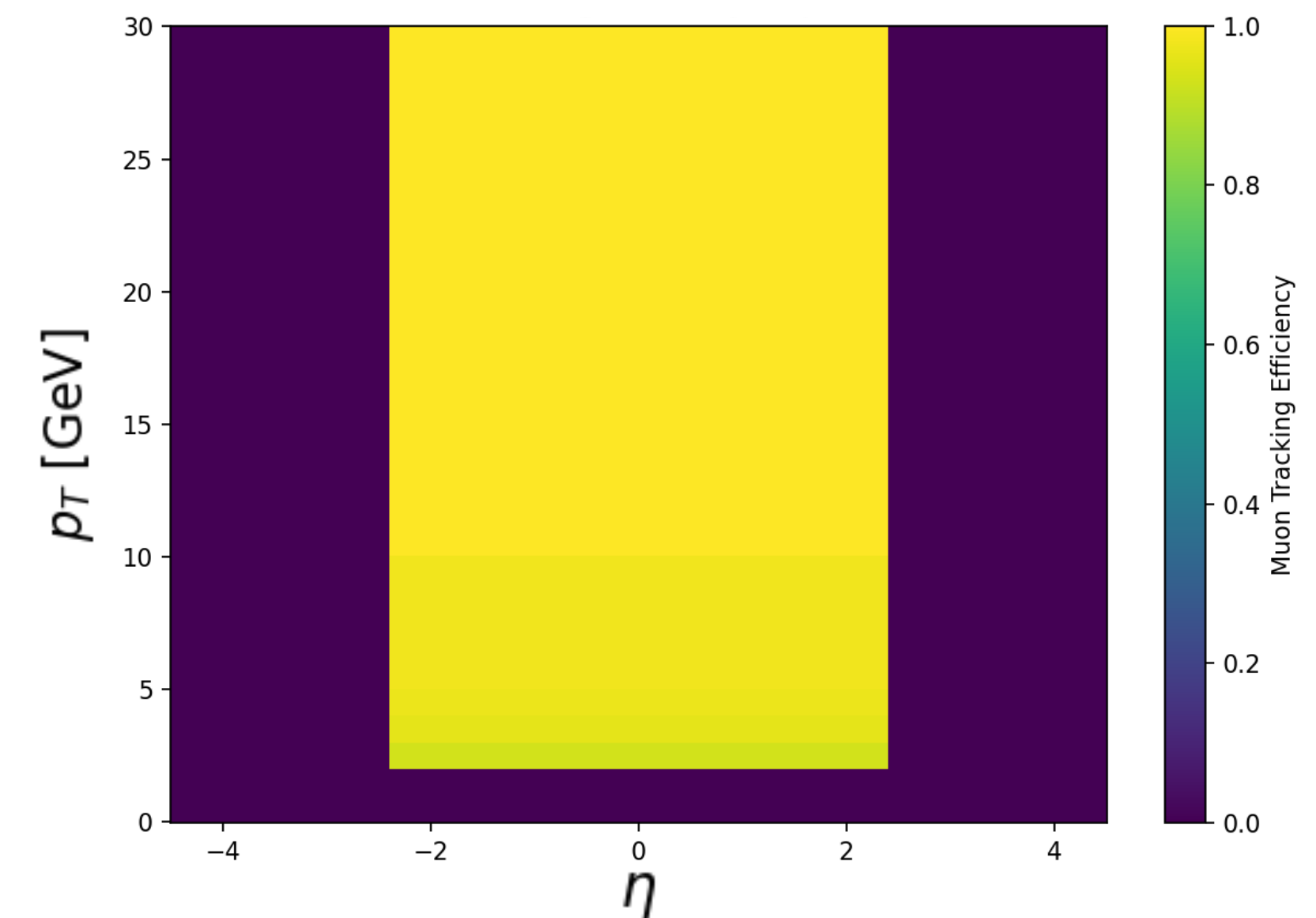
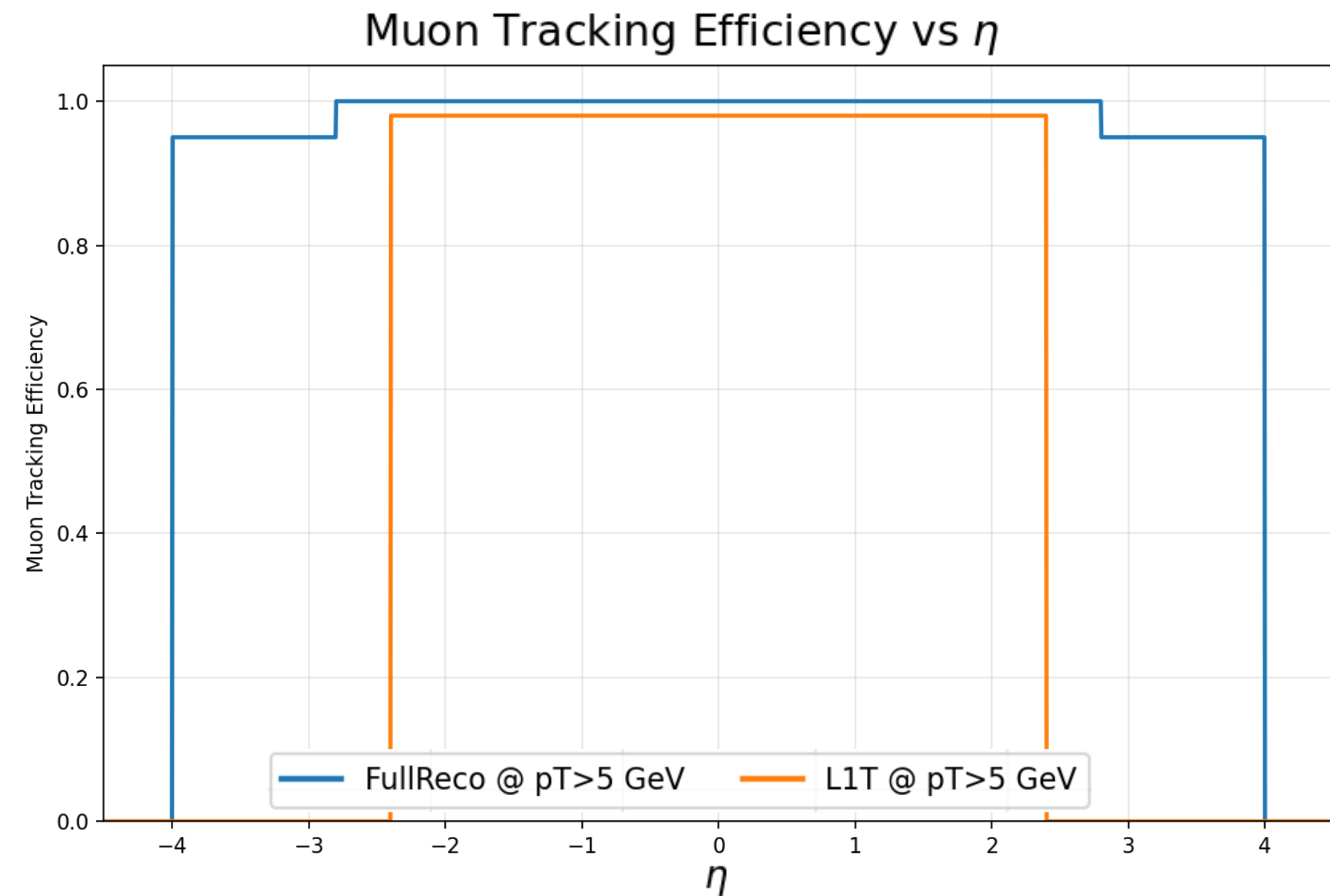
- Main differences between Full-Reco and CMS L1T
- Worse vertex resolution
- Worse track resolution
- **Tracking can go up to eta 2.5**
- **Track efficiency goes to zero below 2 GeV**
- PUPPI algorithm (PUPPI-tune) different
- Muons, barrel, calorimeters perform similarly



Custom Delphes L1T

From the CMS L1T TDR & Phase-2 TDR

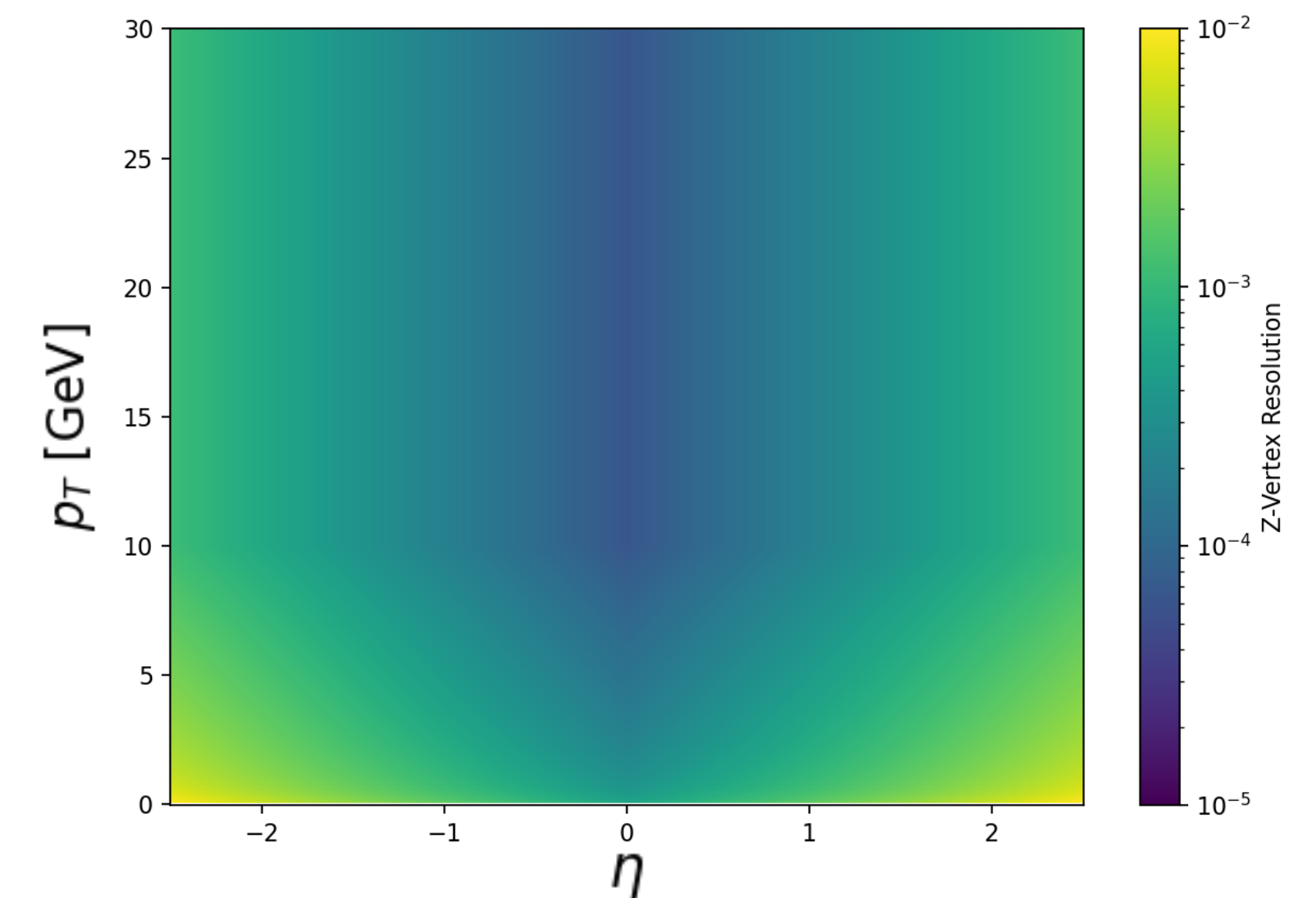
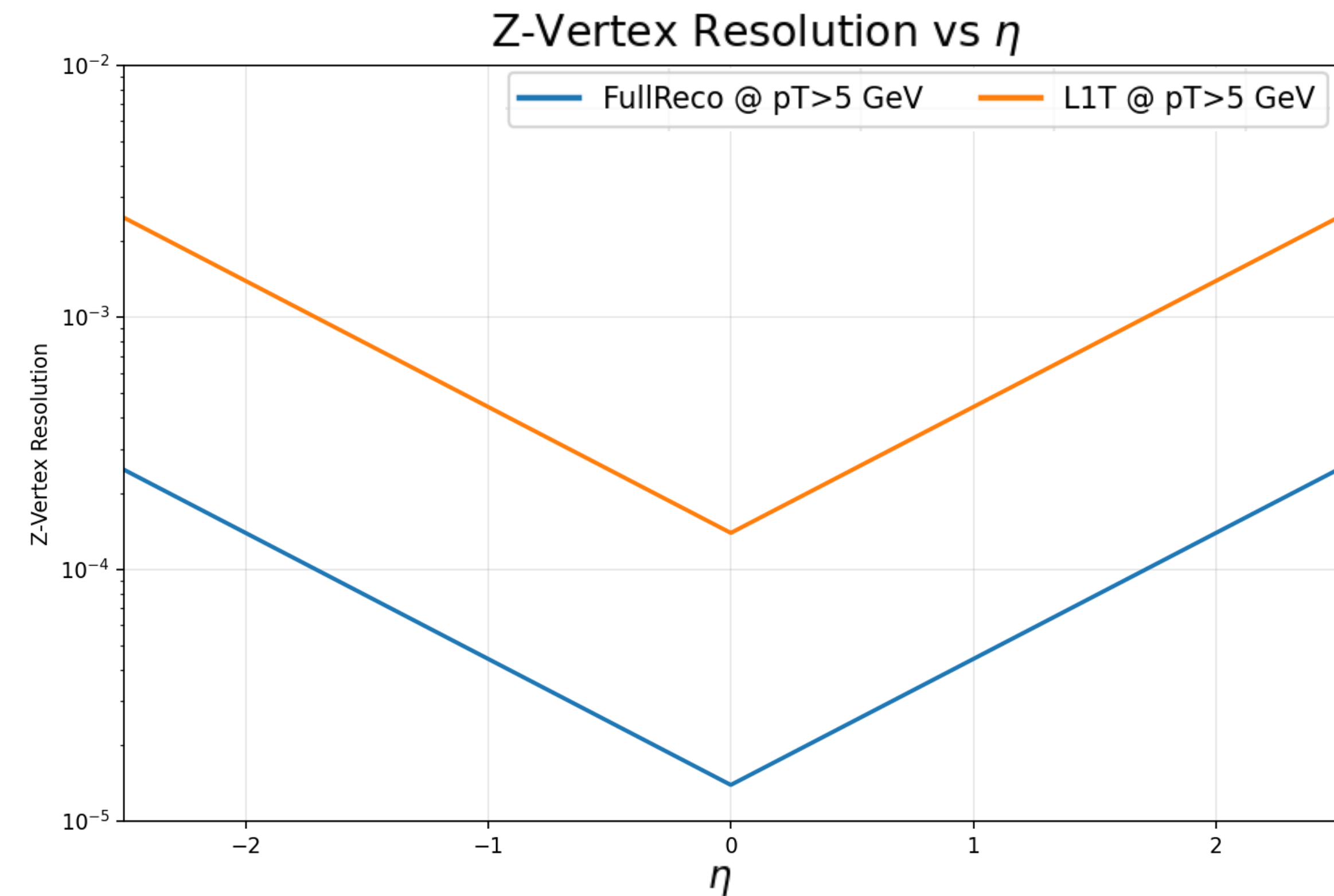
- Main differences between Full-Reco and CMS L1T
- Worse vertex resolution
- Worse track resolution
- **Tracking can go up to eta 2.5**
- **Track efficiency goes to zero below 2 GeV**
- PUPPI algorithm (PUPPI-tune) different
- Muons, barrel, calorimeters perform similarly



Custom Delphes L1T

From the CMS L1T TDR & Phase-2 TDR

- Main differences between Full-Reco and CMS L1T
 - **Worse vertex resolution**
 - Worse track resolution
 - Tracking can go up to eta 2.5
 - Track efficiency goes to zero below 2 GeV
 - **PUPPI algorithm (PUPPI-tune) different**
 - Muons, barrel, calorimeters perform similarly





Custom Delphes L1T

From the CMS L1T TDR & Phase-2 TDR

- Main differences between Full-Reco and CMS L1T
 - Worse vertex resolution
 - Worse track resolution
 - Tracking can go up to eta 2.5
 - Track efficiency goes to zero below 2 GeV
 - **PUPPI algorithm (PUPPI-tune) different**
 - Muons, barrel, calorimeters perform similarly

PUPPI FullReco

	bin1	bin2	bin3
EtaMinBin	0.0	1.5	4.0
EtaMaxBin	1.5	4.0	10.0
PtMinBin	0.0	0.0	0.0
ConeSizeBin	0.2	0.2	0.2
RMSPtMinBin	0.1	0.5	0.5
RMSScaleFactorBin	1.0	1.0	1.0
NeutralMinEBin	0.2	0.2	0.5
NeutralPtSlope	0.006	0.013	0.067
ApplyCHS	TRUE	TRUE	TRUE
UseCharged	TRUE	TRUE	FALSE

PUPPI L1T

	bin1	bin2	bin3	bin4
EtaMinBin	0.0	1.5	2.5	3.0
EtaMaxBin	1.5	2.5	3.0	10.0
PtMinBin	0.0	0.0	0.0	0.0
ConeSizeBin	0.3	0.3	0.3	0.3
RMSPtMinBin	0.1	0.5	0.5	0.5
RMSScaleFactorBin	1.0	1.0	1.0	1.0
NeutralMinEBin	1.0	1.0	4.0	10.0
NeutralPtSlope	0.006	0.013	0.03	0.03
ApplyCHS	TRUE	TRUE	TRUE	TRUE
UseCharged	TRUE	TRUE	FALSE	FALSE

Processes

Soft & QCD Multijet

Minbias / Soft QCD | 100
QCD Inclusive | 100
QCD bb (heavy-flavor enriched) | 25

Quarkonium Control Sample

$\Upsilon \rightarrow \text{leptons}$ | 25

Single-Higgs

Gluon Fusion ($ggH \rightarrow X$) | 10
 $ggHbb, ggHcc, ggH\tau\tau, ggH\gamma\gamma, ggHgg, ggHZZ, ggHWW$
Vector-Boson Fusion ($VBFH \rightarrow X$) | 10
 $VBFHbb, VBFHcc, VBFH\tau\tau, VBFH\gamma\gamma, VBFHgg, VBFHZZ, VBFHWW$

Single Bosons, V+jets & DY

$W \rightarrow \ell\nu, W \rightarrow qq$ | 25
 $DY \rightarrow \ell\ell$ | 25
 $Z \rightarrow \nu\nu + jet$ | 25
 $Z \rightarrow qq (uds), Z \rightarrow bb, Z \rightarrow cc$ | 25

Top-quark family

$t\bar{t}$ (all-had, semi-lep, all-lep) | 25
 $t\bar{t}t\bar{t}$ (4-top) | 2
 $t\bar{t}W$ (incl), $t\bar{t}Z$ (incl) | 5
 $t\bar{t}H$ (incl) | 10

Photon Processes

γ (prompt photon + jets) | 25
 $\gamma + V$ ($W\gamma/Z\gamma$ -like) | 10
Tri- $\gamma(3\gamma)$ | 2

Associated Higgs

VH (incl) (W/Z + H; decay-agnostic) | 10

Diboson & Triboson

WW (all-had, semi-lep, all-lep) | 3
 WZ (all-had, semi-lep, all-lep) | 3
 ZZ (all-had, semi-lep, all-lep) | 3
 VVV (triboson) | 2

Di-Higgs ($HH \rightarrow \text{final states}$)

$HH \rightarrow b\bar{b}b\bar{b}$ | 2
 $HH \rightarrow b\bar{b}\tau\tau$ | 2
 $HH \rightarrow b\bar{b}WW, HH \rightarrow b\bar{b}ZZ$ | 2
 $HH \rightarrow b\bar{b}\gamma\gamma$ | 2

BSM Processes (tbd)

Legend: Process | Size (Millions)

Dataset Features / Collections



Low-Level Constituents (PF-like)

PFCands / EFlowCands

PT, Eta, Phi, PID, Charge, Mass, D0, DZ, ErrorD0, ErrorDZ, fUniqueID, PuppiW

PUPPI Particles

PT, Eta, Phi, PID, Charge, Mass, D0, DZ, ErrorD0, ErrorDZ, fUniqueID, PuppiW

Jets (small-R / large-R; CHS vs PUPPI)

Jet-AK4

PT, Eta, Phi, Mass, BTag, BTagPhys, Charge, Constituents

Jet-AK8

PT, Eta, Phi, Mass, BTag, BTagPhys, Charge, Constituents

JetPUPPI-AK4

PT, Eta, Phi, Mass, BTag, BTagPhys, Charge, Constituents

JetPUPPI-AK8

PT, Eta, Phi, Mass, BTag, BTagPhys, Charge, Constituents

Generator-level supervision (truth)

GenPart (Particle)

PT, Eta, Phi, PID, M1, M2, D1, D2, Status, IsPU

GenJets(AK4/AK8)

PT, Eta, Phi, Mass

GenMET

MET, Eta, Phi

High-Level Constituents

Electrons

PT, Eta, Phi, IsolationVarRhoCorr
EhadOverEem

MuonsTight

PT, Eta, Phi, IsolationVarRhoCorr

PhotonsTight

PT, Eta, Phi

Event-Level MET

MET (MissingET)

MET, Phi, Eta

PUPPI-MET

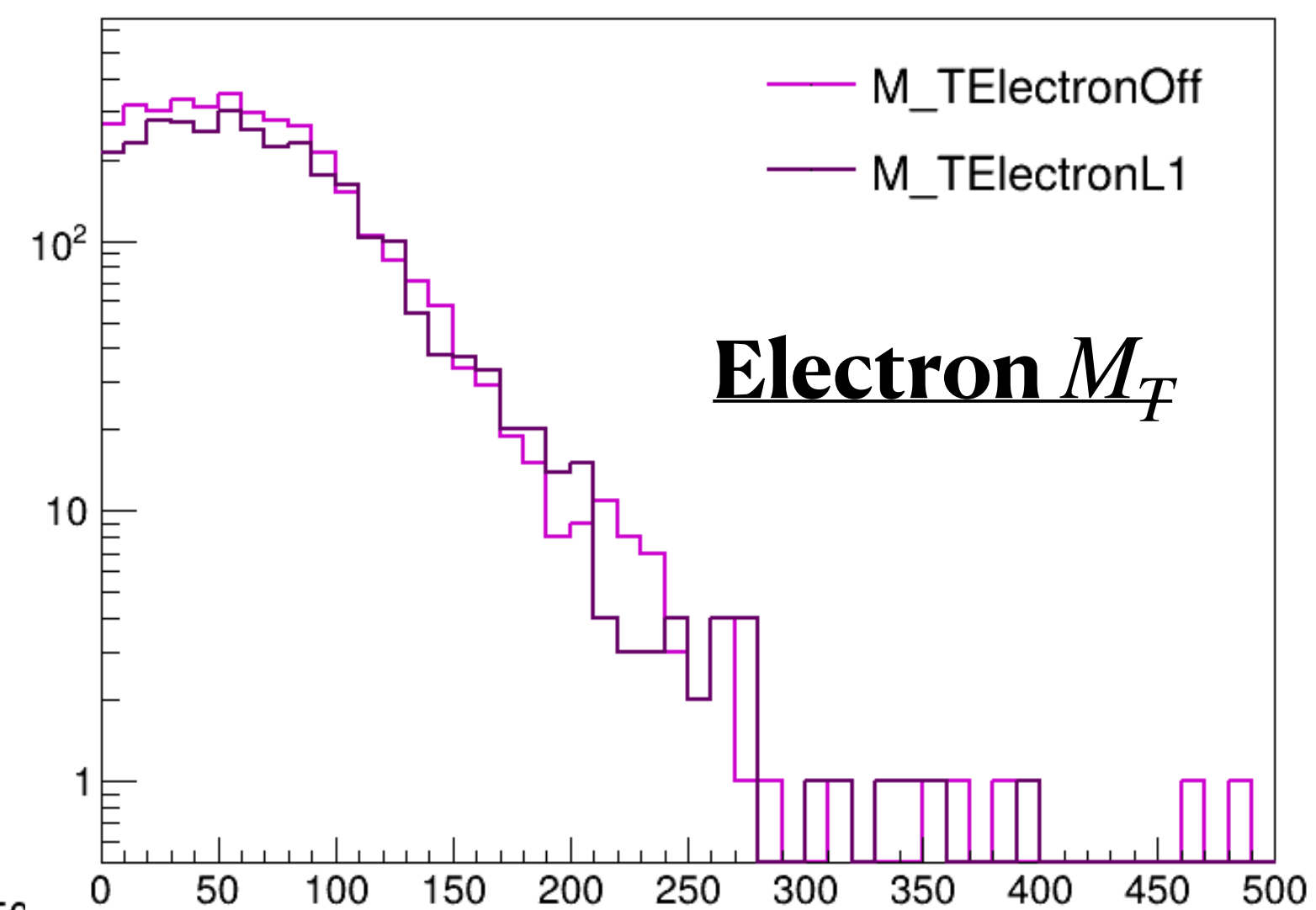
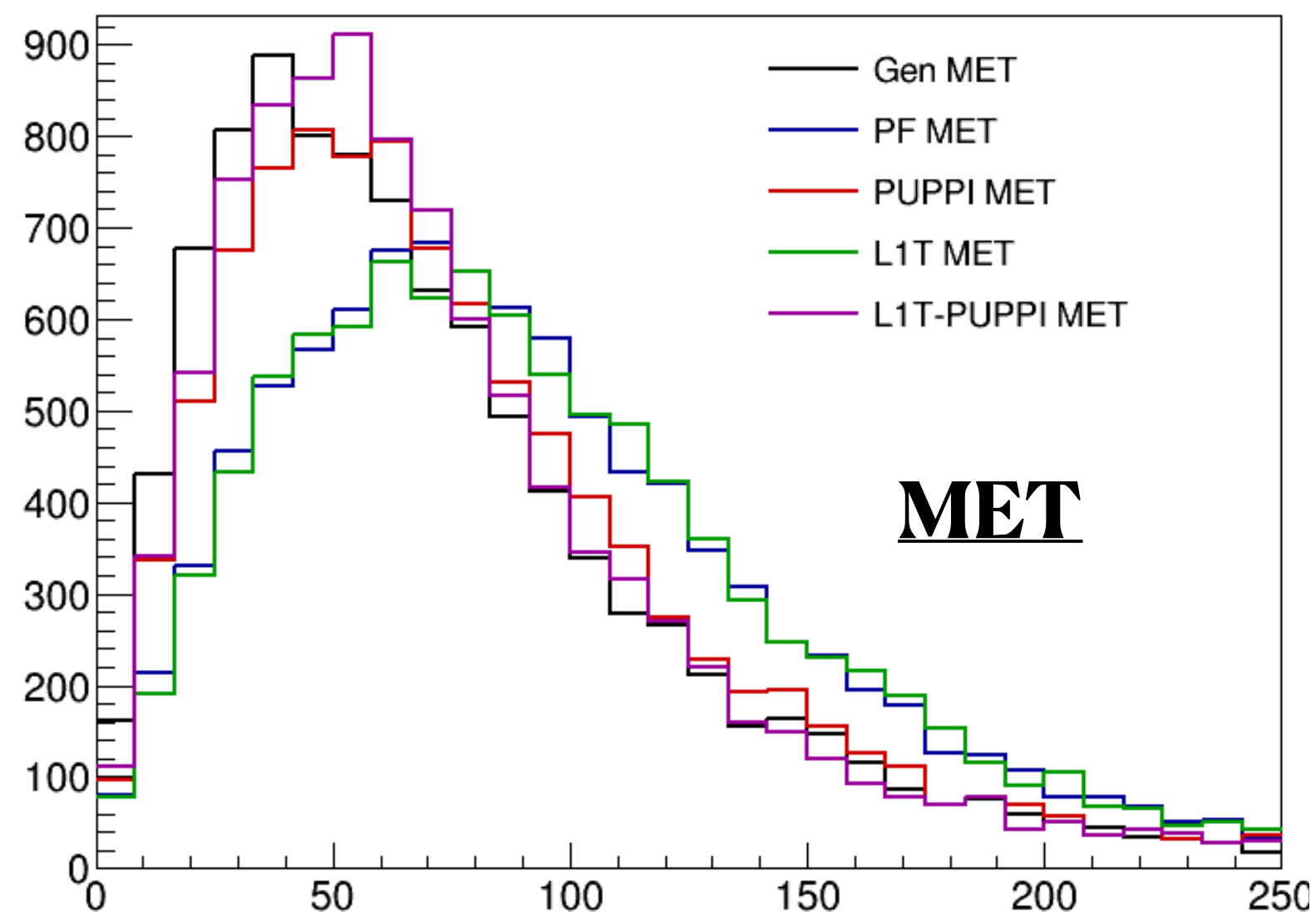
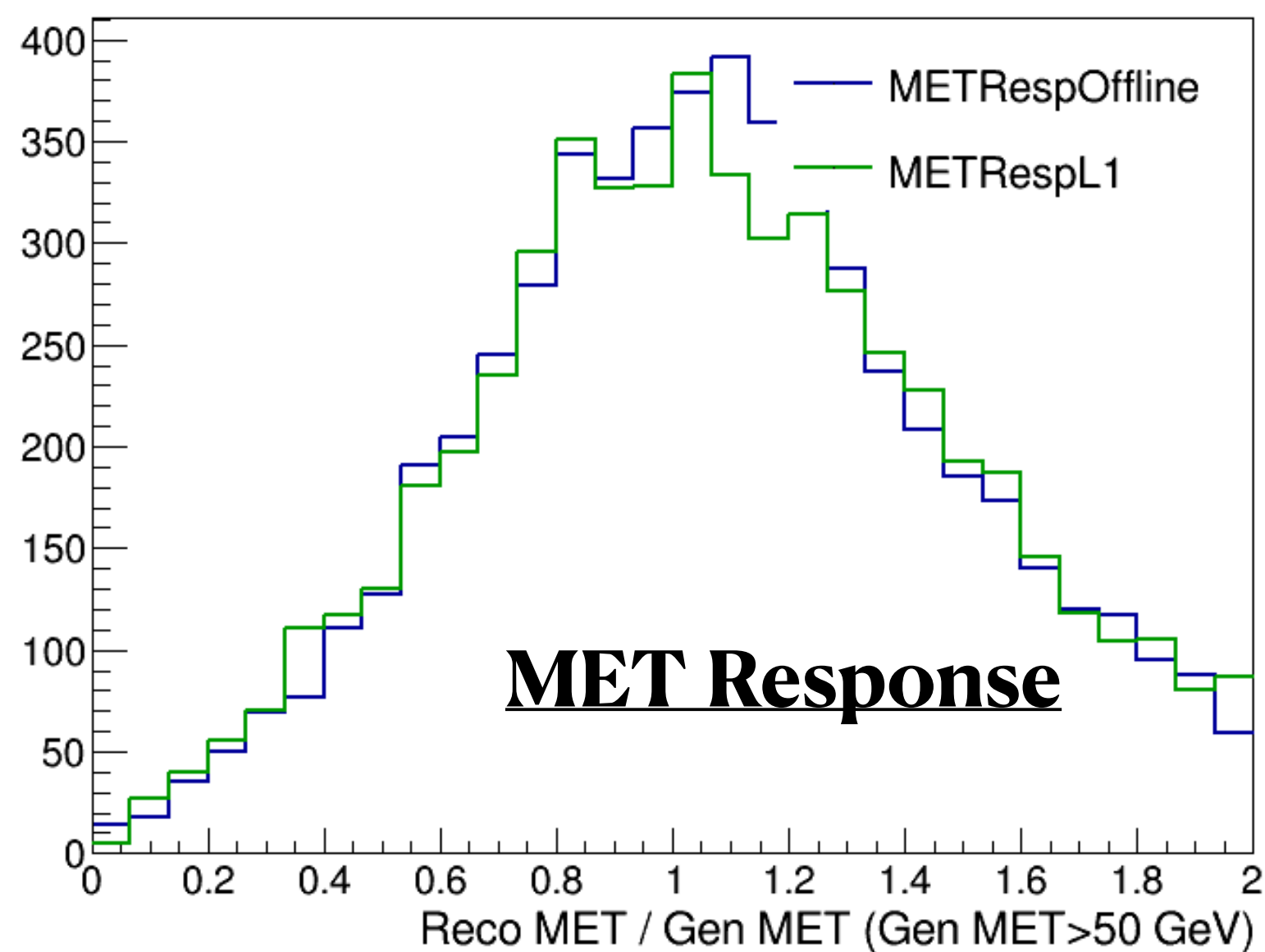
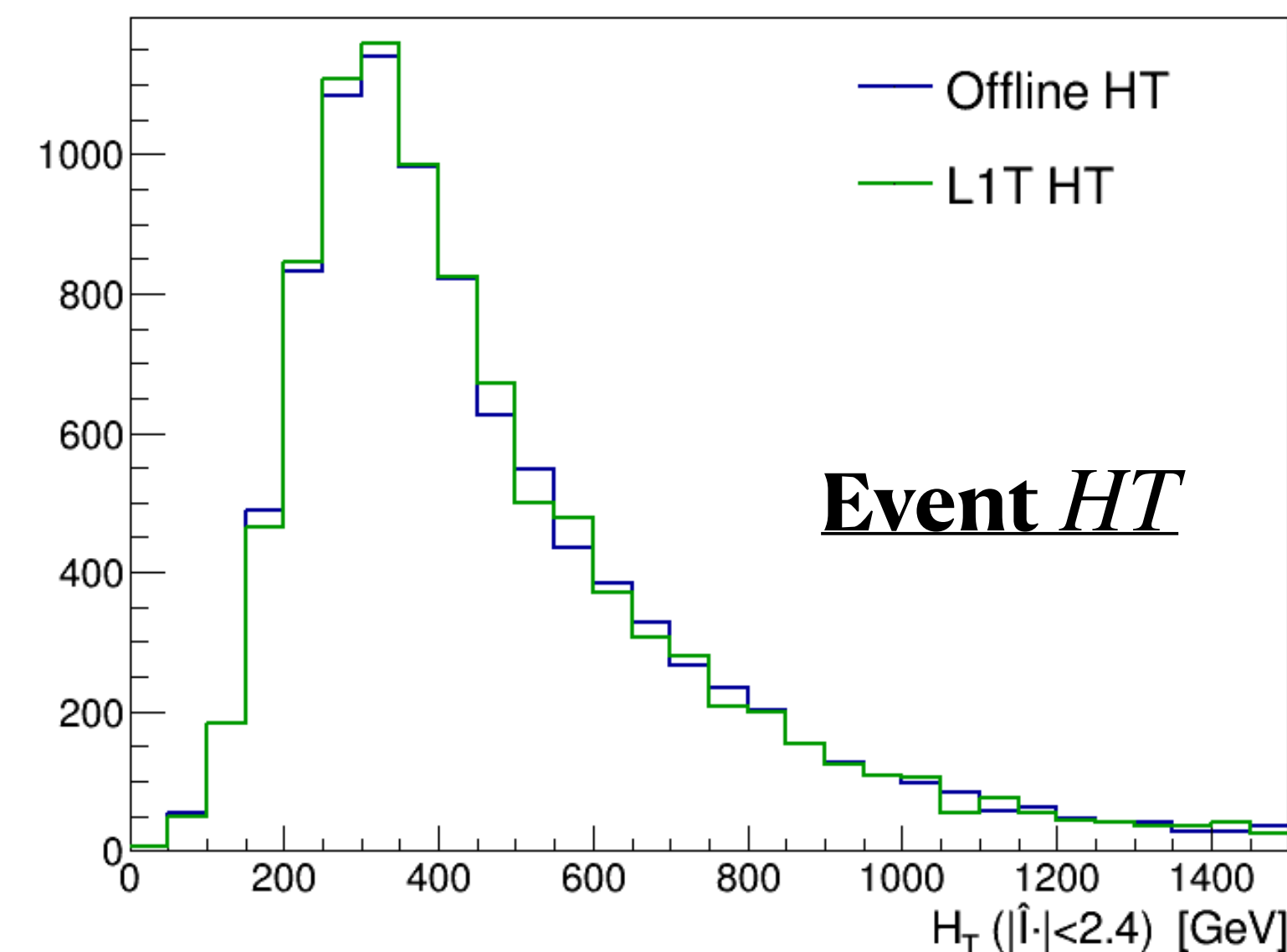
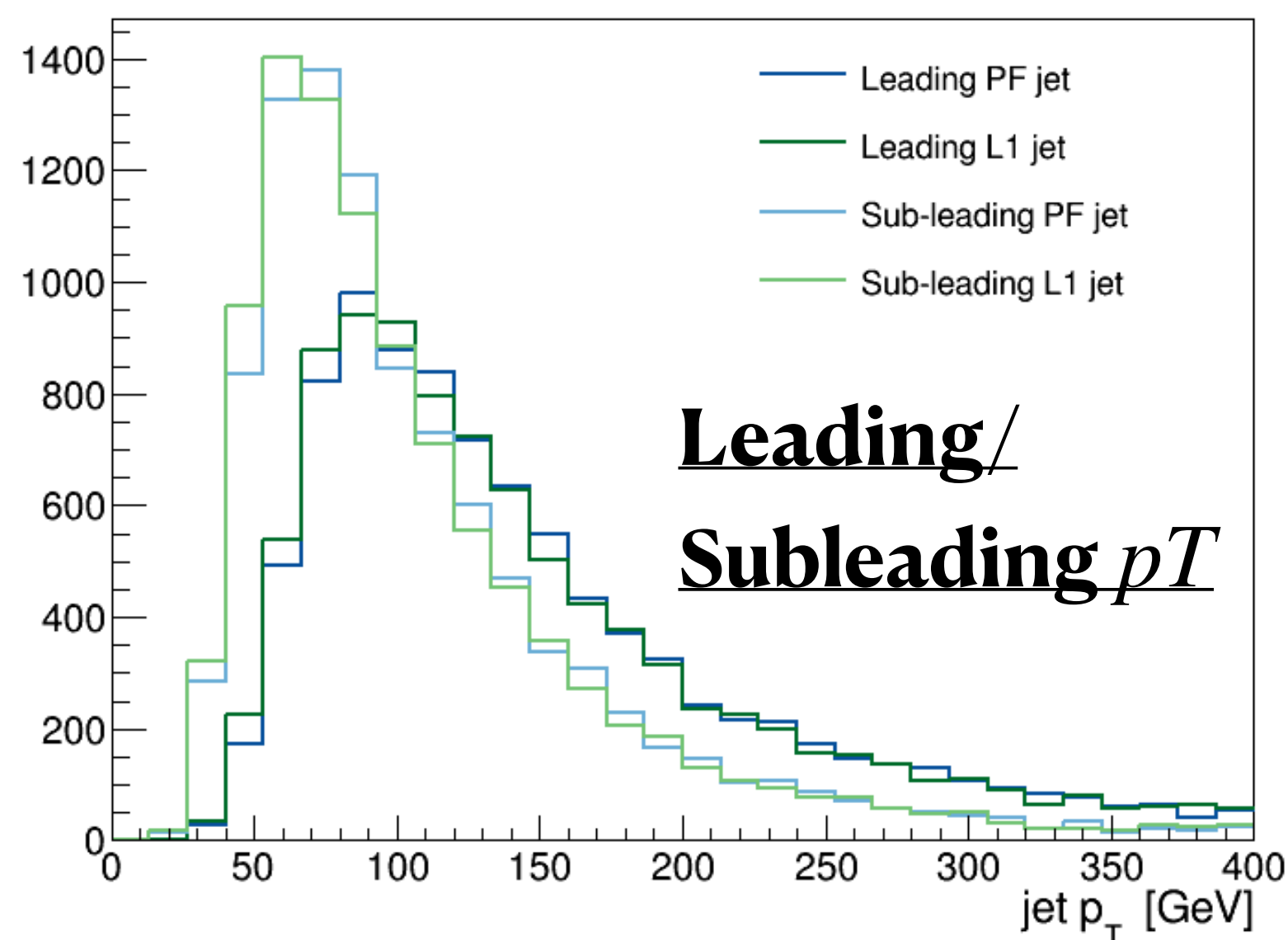
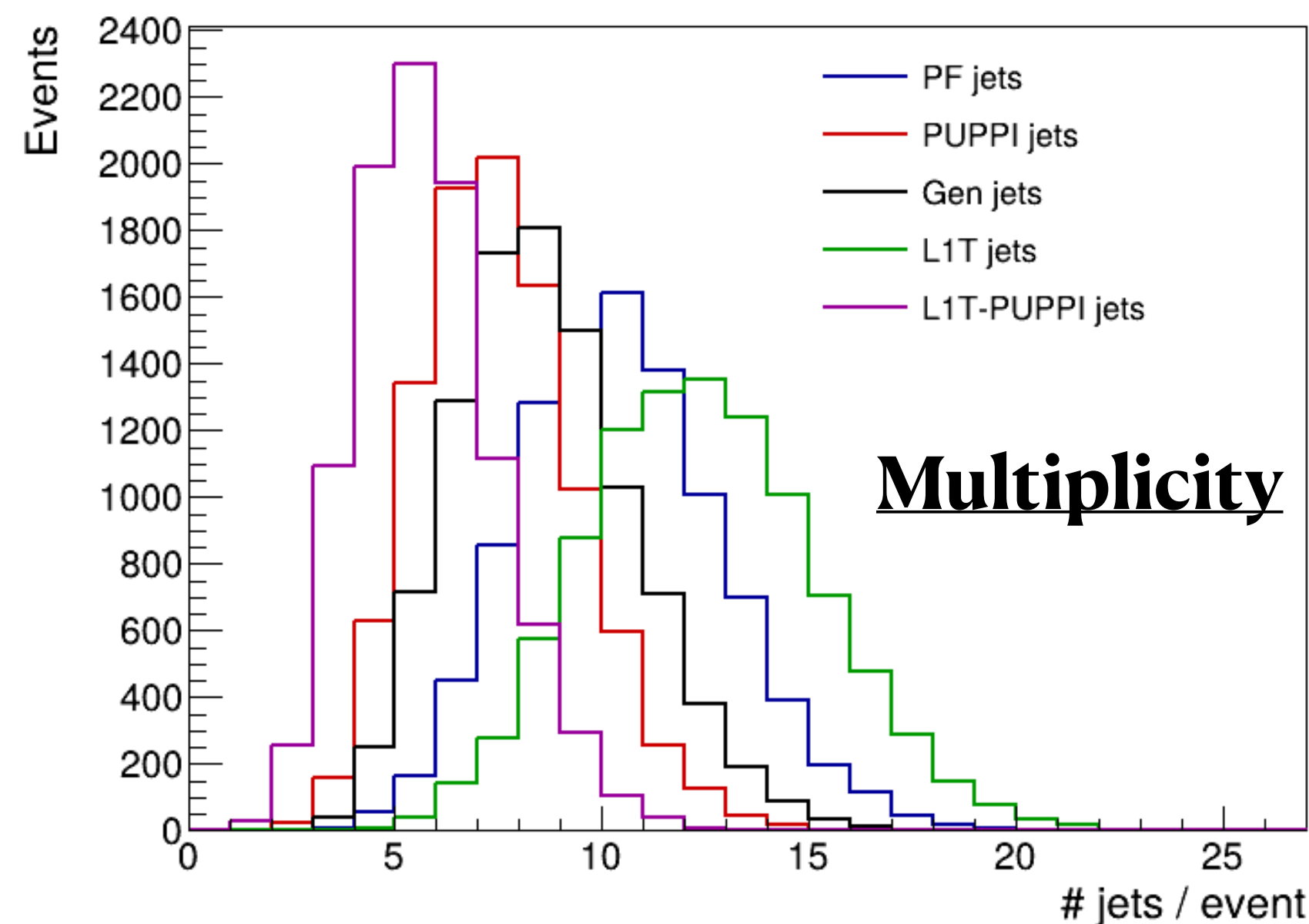
MET, Phi, Eta

PrimaryVertex (Vertex)

X, Y, Z, T, SumPT2



Example Process: $t\bar{t}$ semi-leptonic



Data Structure

Event Num	L1T_PFCand_pT (GeV)	L1T_JetAK4_pT (GeV)	L1T_JetAK4_Constituents	...
0	[75.1, 4.2, 43.8, ...]	[12.1, 2.4, ...]	[[..., ...] [..., ...] [..., ...]]	
1	[94.0, 3.1, 4.7, 9.8, ...]	[3.4, 4.1, 5.5, ...]	[[..., ..., ...], [..., ..., ...], ...]	
2	[77.4, 41.2, ...]	[11.6, 252.1, 80.4...]	[[..., ..., ...], ..., [..., ...]]	
⋮	⋮	⋮	⋮	

All branches saves as columns with jagged and nested arrays

Nested structure of one column: [event index][particle/jet index]([jet constituent index])

Constituent column maps to respective PFCand features

Datasets:

fastmachinelearning/collide-1m

like

1

Follow

Fast Machine Learning ...

4

Modalities:

Time-series

Formats:

parquet

Size:

1M - 10M

Tags:

physics

Libraries:

Datasets

Dask

Dataset card

Data Studio

Files and versions

xet

Community

2

main

collide-1m

pploner

Update COLLIDE2V_example_notebook.ipynb

114dd2a

VERIFIED

DYJetsToLL_13TeV-madgraphMLM-pythia8

HH_4b

HH_bbWW

HH_bbZZ

HH_bbgammagamma

HH_bbtautau

QCD_HT50toInf

QCD_HT50tobb

VBFHWW

VBFHZZ

VBFHbb

VBFHcc

Example Streaming Call

```
# load dataset of one process folder
# if one wants to load the entire dataset, skip the data_dir or data_files argument
dataset = load_dataset("fastmachinelearning/collide-1m",
                        data_dir="WJetsToLNu_13TeV-madgraphMLM-pythia8",
                        streaming=True)

dataset = dataset["train"]
```

Features

```
cols = list(dataset.features.keys())
for c in cols:
    print(c)

FullReco_PFCand_PT
FullReco_PFCand_Eta
FullReco_PFCand_Phi
FullReco_PFCand_PID
FullReco_PFCand_Charge
FullReco_PFCand_Mass
FullReco_PFCand_D0
```

Values

```
example_arr = ak.Array(row1['FullReco_PFCand_PT'])
example_arr

[0.291,
 0.226,
 0.454,
 0.468,
 0.412,
 0.22,
 0.755,
 0.963,
```

Training

HF Training Notebook

```
class TinyMLP(nn.Module):
    def __init__(self, d=184, h=256, num_classes=6):
        super().__init__()
        self.net = nn.Sequential(
            nn.Linear(d, h), nn.ReLU(),
            nn.Linear(h, h//2), nn.ReLU(),
            nn.Linear(h//2, num_classes),
        )
    def forward(self, x):
        return self.net(x)
```

```
for batch in train_loader:
    x = batch['x'].to(DEVICE, non_blocking=True)
    y = batch['y'].to(DEVICE, non_blocking=True)

    logits = model(x)
    loss = loss_fn(logits, y)
    opt.zero_grad(set_to_none=True)
    loss.backward()
    opt.step()
```

```
[epoch 1] val acc: 49.24% | classes: ['DY', 'QCD', 'SingleHiggs', 'top', 'diboson', 'diHiggs']
epoch 2 step 20 | loss 0.4491
[epoch 2] val acc: 52.77% | classes: ['DY', 'QCD', 'SingleHiggs', 'top', 'diboson', 'diHiggs']
[epoch 3] val acc: 55.13% | classes: ['DY', 'QCD', 'SingleHiggs', 'top', 'diboson', 'diHiggs']
```

Adding BSM Models in Future



**Reproducible
Statistical Analysis**

reana

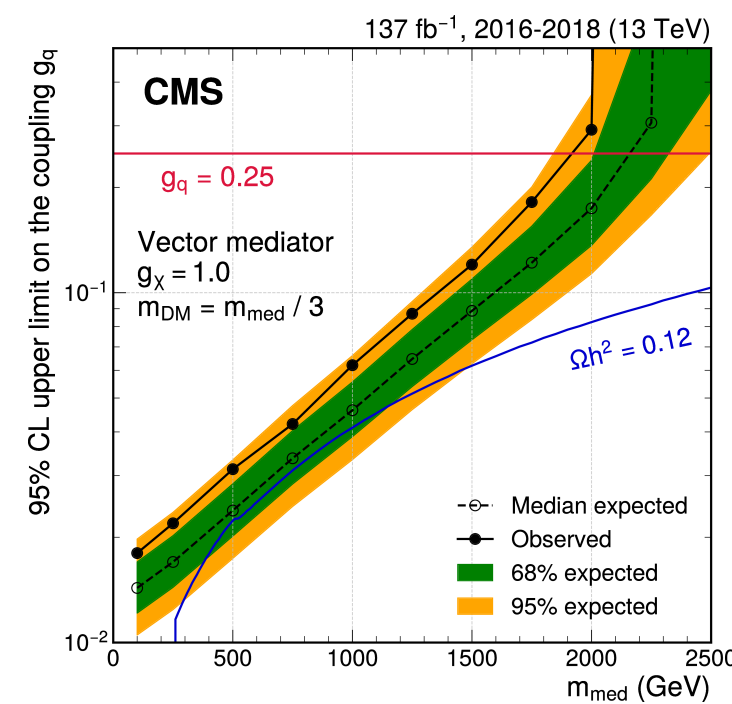
Reproducible research data analysis platform

Open Data

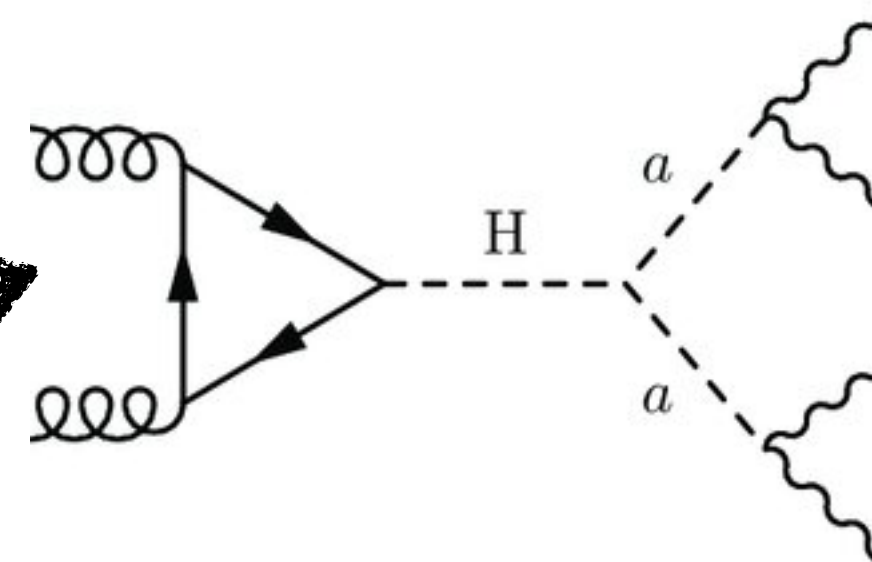
COLLIDE-2V

Reconstruction

**Physics
Results**



MadGraph Card

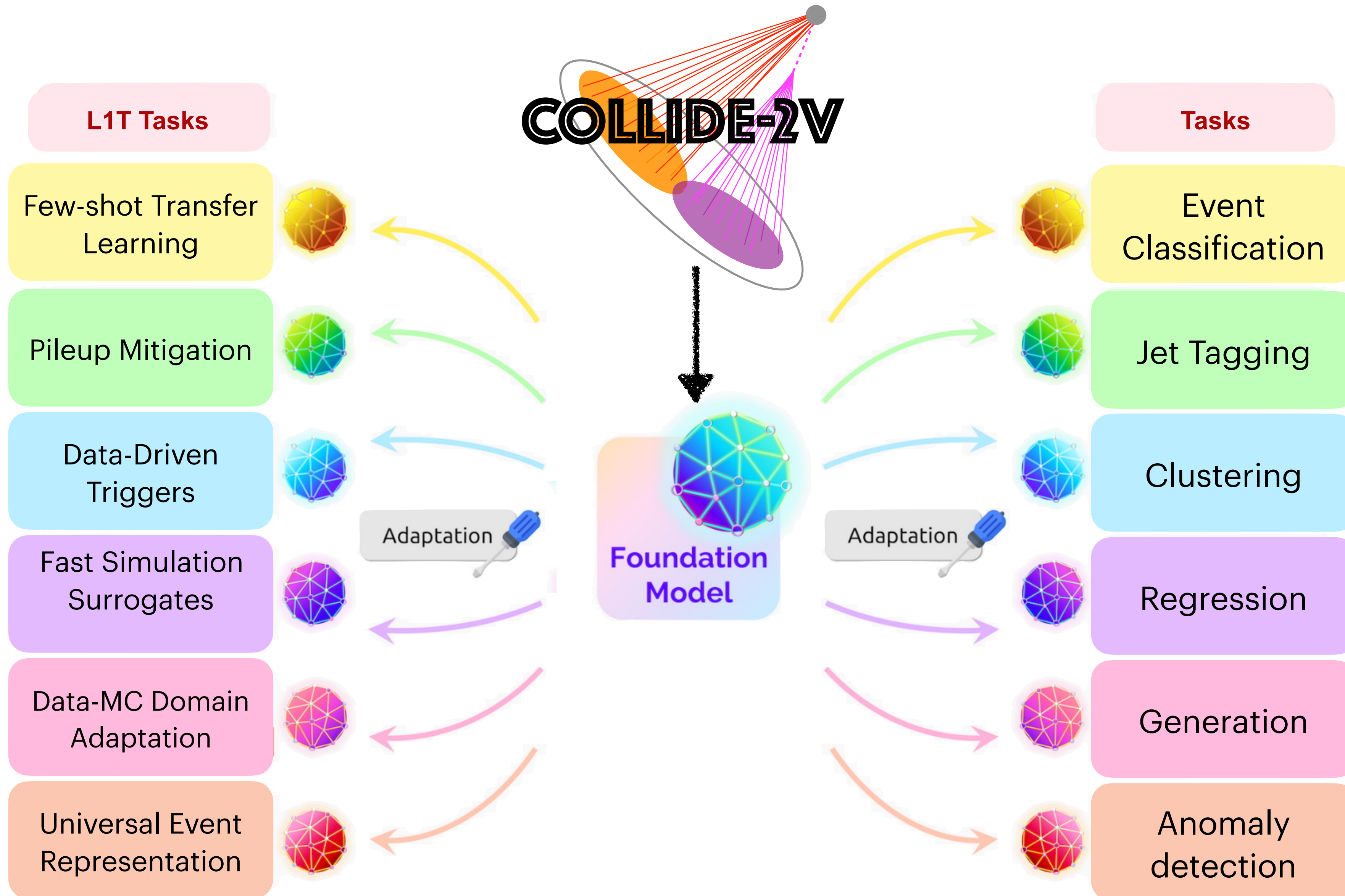


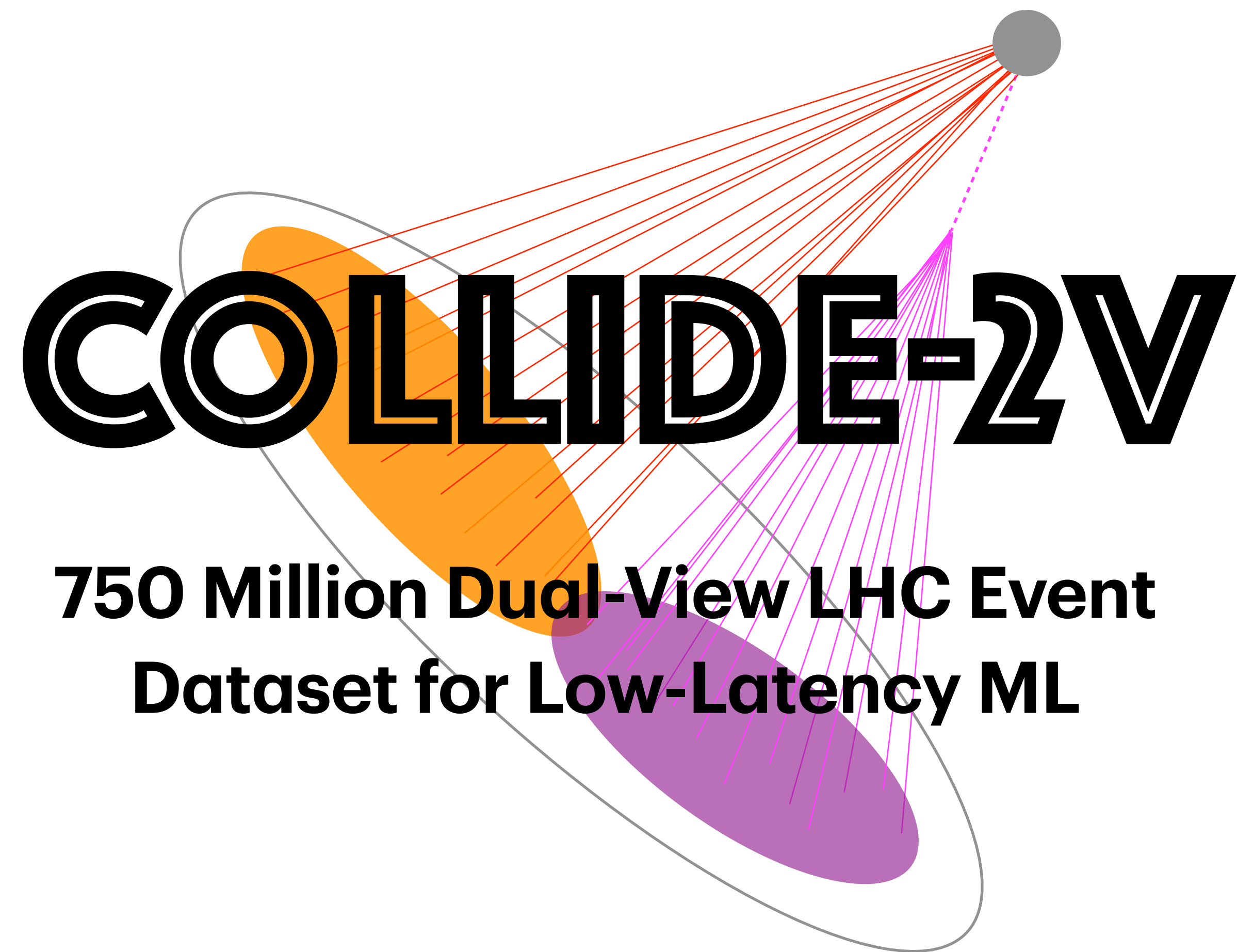
**Detector
Simulation
(Dual-View)**

BSM Processes (planned)

Dark shower prompt (including SUEPs)
 Dark shower semi-visible
 Dark shower long lived
 RPV multijets
 $Z' \rightarrow qq/\tau\tau$
 $H \rightarrow 2a \rightarrow 4b/4\tau/2b2\tau/4\gamma$
 Light dark photon to di-lepton/di-pion
 Heavy neutral leptons
 Light pseudoscalar $\rightarrow 2\gamma$
 $H \rightarrow 4b$ displaced

Towards a Foundation Model





Thanks for listening!

[Public CernBox](#)

[HuggingFace](#)

If you have any questions or suggestions please let us know at emoreno@mit.edu or plonerp@ethz.ch

Example Process: $DY \rightarrow \ell\ell$

