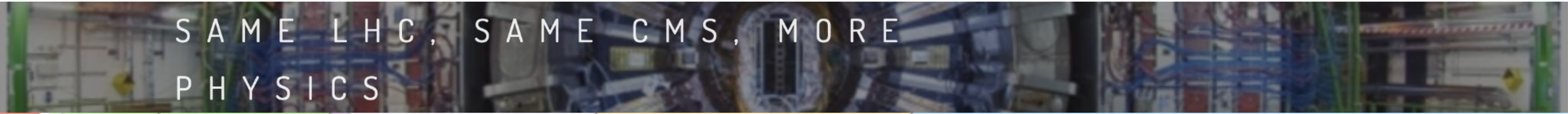
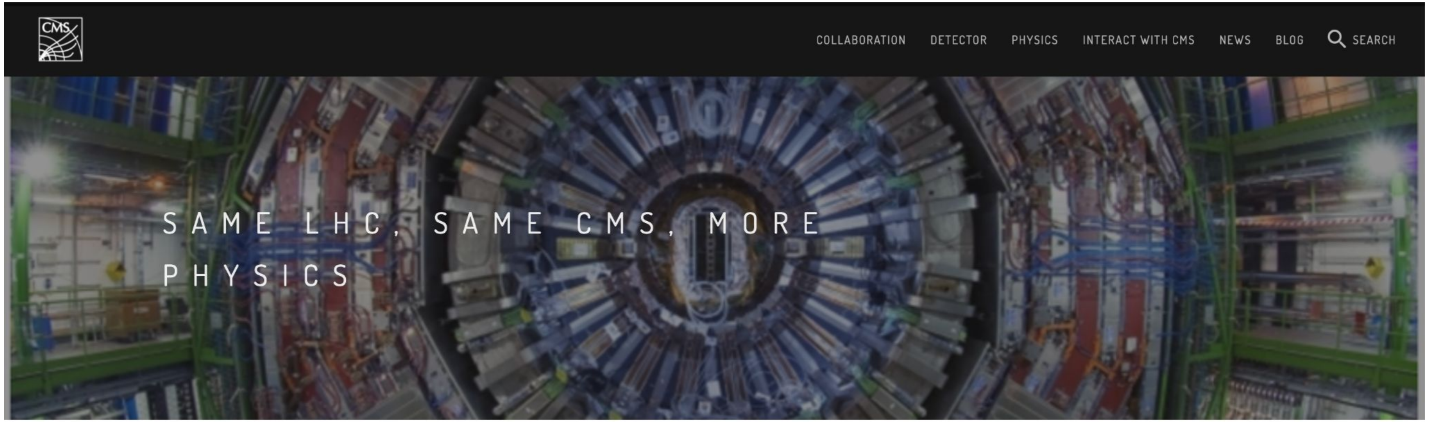
H → bb measurement exploiting data scouting during run 3 at CMS

Patin Inkaew, Henning Kirschenmann, Mikko Voutilainen
Helsinki Institute of Physics

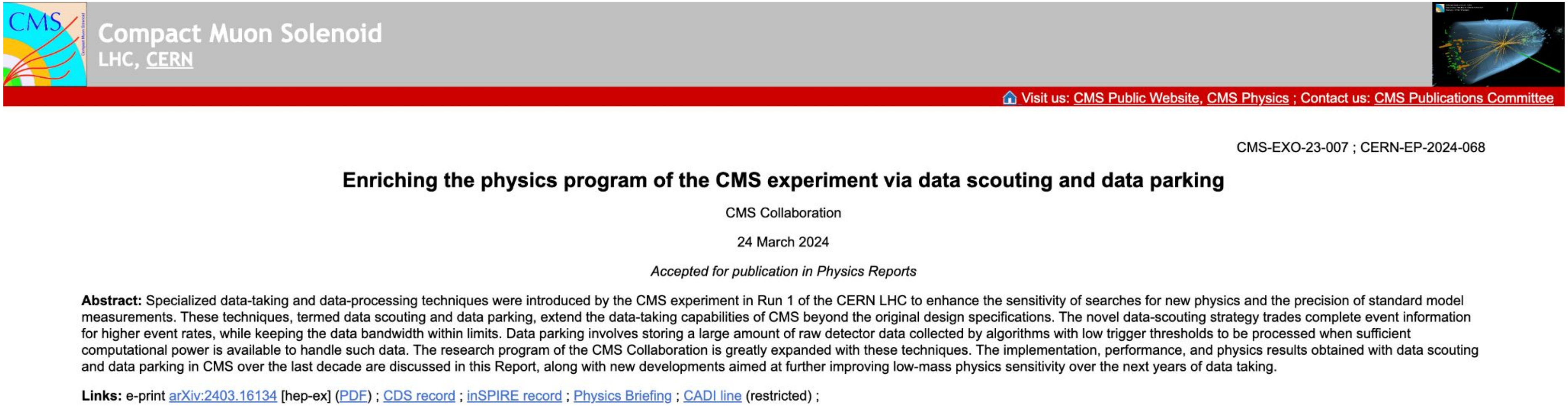
SMARTHEP Yearly Meeting (01.10.2024)
Milano-Bicocca University, Italy





PAPER IS OUT!

[CMS physics briefing](#)
[CMS public results](#)


CMS
Compact Muon Solenoid
LHC, CERN
Visit us: CMS Public Website, CMS Physics ; Contact us: CMS Publications Committee

CMS-EXO-23-007 ; CERN-EP-2024-068

Enriching the physics program of the CMS experiment via data scouting and data parking

CMS Collaboration

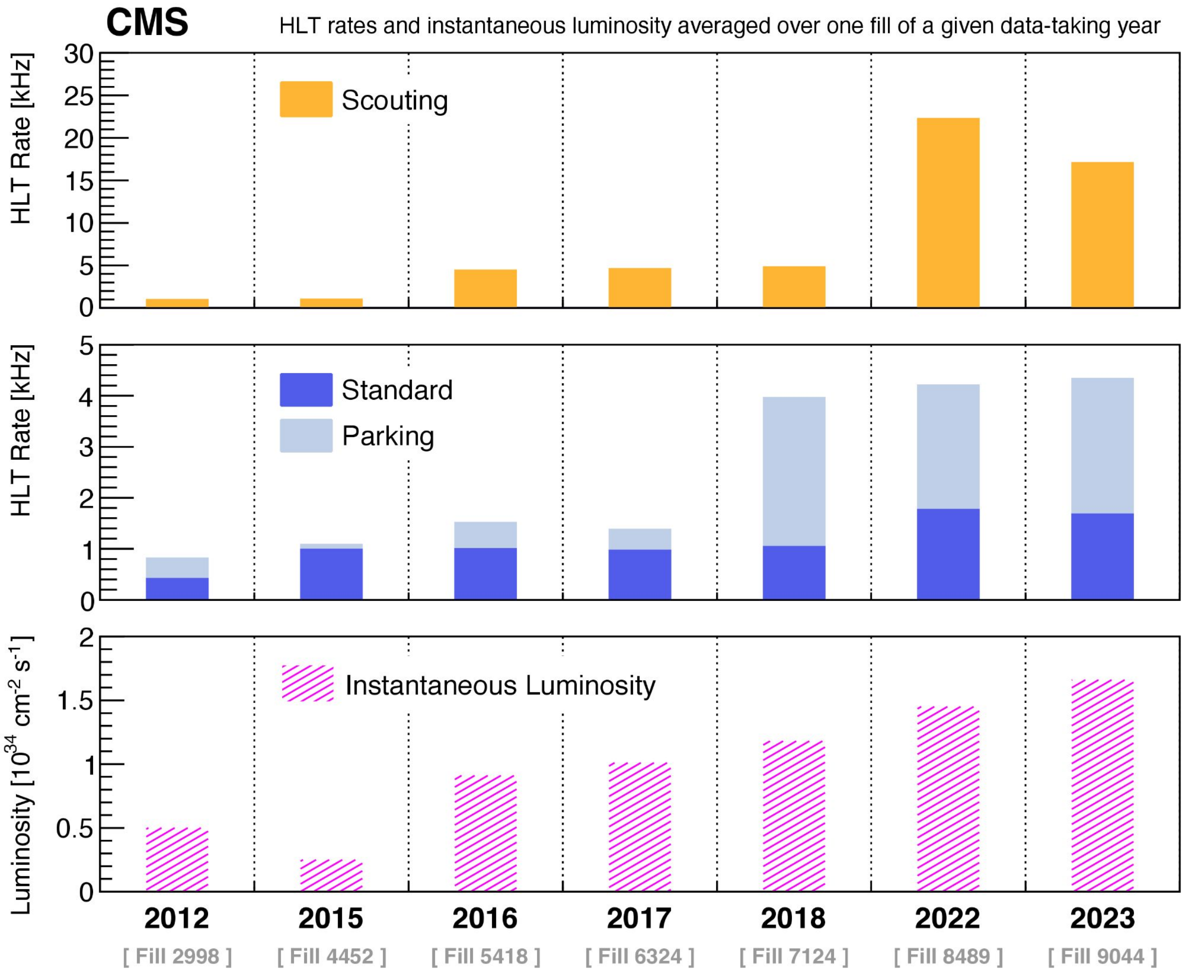
24 March 2024

Accepted for publication in Physics Reports

Abstract: Specialized data-taking and data-processing techniques were introduced by the CMS experiment in Run 1 of the CERN LHC to enhance the sensitivity of searches for new physics and the precision of standard model measurements. These techniques, termed data scouting and data parking, extend the data-taking capabilities of CMS beyond the original design specifications. The novel data-scouting strategy trades complete event information for higher event rates, while keeping the data bandwidth within limits. Data parking involves storing a large amount of raw detector data collected by algorithms with low trigger thresholds to be processed when sufficient computational power is available to handle such data. The research program of the CMS Collaboration is greatly expanded with these techniques. The implementation, performance, and physics results obtained with data scouting and data parking in CMS over the last decade are discussed in this Report, along with new developments aimed at further improving low-mass physics sensitivity over the next years of data taking.

Links: e-print arXiv:2403.16134 [hep-ex] (PDF) ; CDS record ; inSPIRE record ; Physics Briefing ; CADI line (restricted) ;

HLT rates



Evolution from 2012-2023 (CMS-EXO-23-007)

Numbers for 2024

Scouting	: 25 kHz
Parking	: 4.9 kHz
Standard	: 2.5 kHz

~an order of magnitude higher than standard

>200 Billion scouting events collected in Run3

>100 Billion scouting events collected in 2024 alone

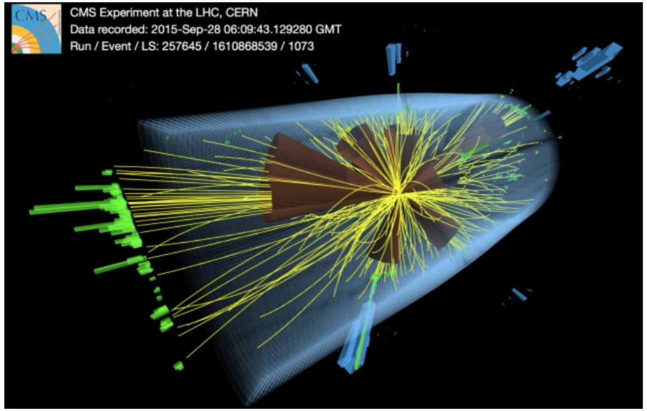
>1.5 PB of scouting data stored in Run3

Can we even analyse these data?

Outline

- Introduction
- ScoutingNano
- H → bb measurement using data scouting during run 3 at CMS
- JEC for scouting jets
- Other activities
- Conclusion

Welcome to CMS and CMSSW

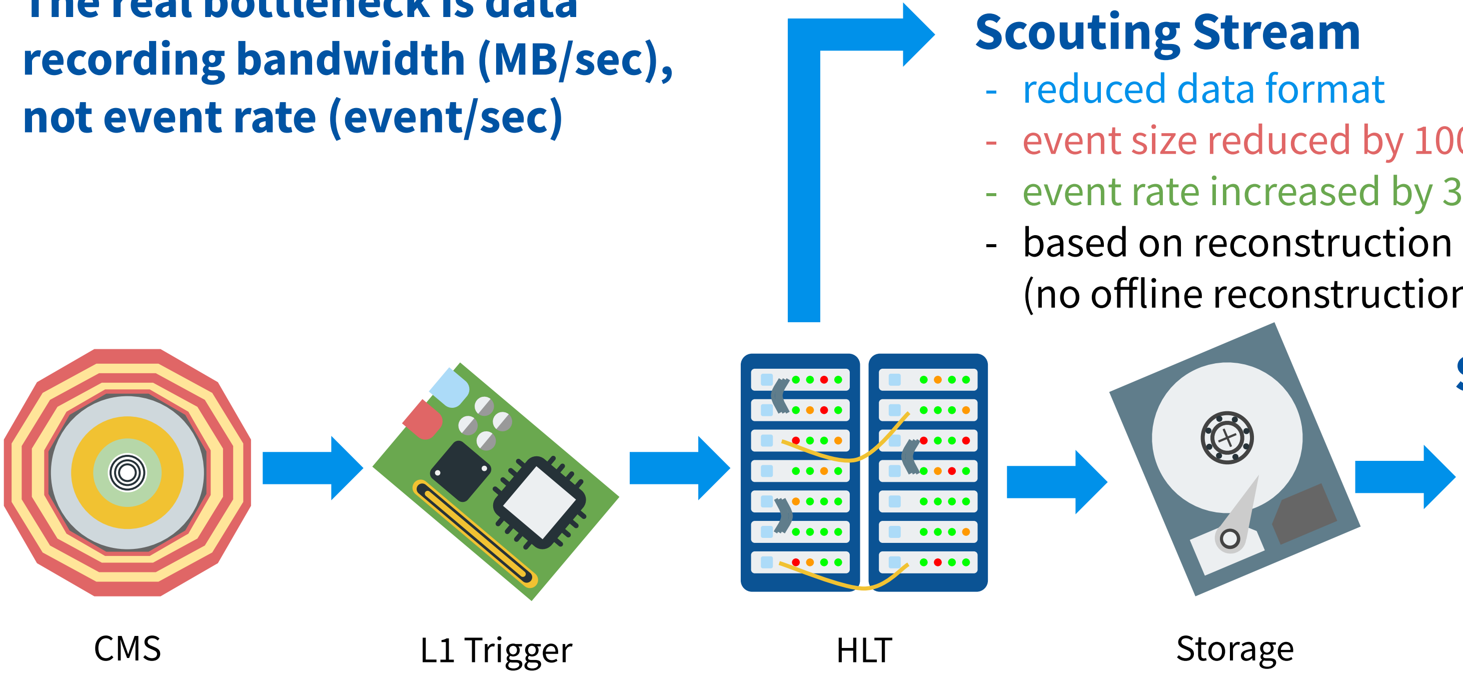

CMS
CMS Experiment at the LHC, CERN
Data recorded: 2015-Sep-28 06:09:43.129280 GMT
Run / Event / LS: 257645 / 1610868539 / 1073

cms-sw.github.io

Introduction

HLT Scouting in a nutshell

The real bottleneck is data recording bandwidth (MB/sec), not event rate (event/sec)



Scouting Stream

- reduced data format
- event size reduced by 100x
- event rate increased by 30x
- based on reconstruction at HLT (no offline reconstruction)

Standard Stream

- ~ 2.5 kHz
- ~ 1.1 MB/event
- full event
- reconstructed at Tier0 after 48 hours

CMS datatier

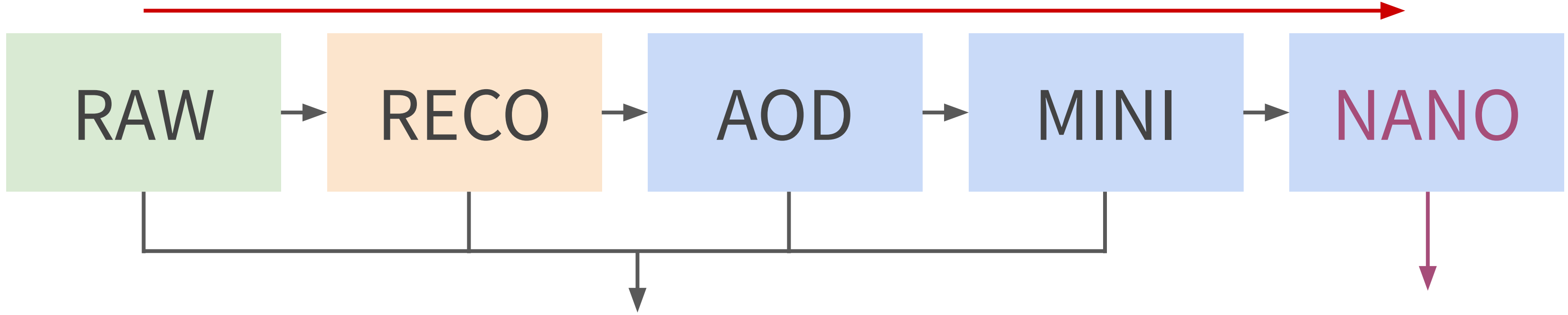
Modified from CERN-CMS-NOTE-2023-003

Raw detector readout
data coming out of the detector

Reconstruction
to physics objects

Analysis Object Data
suitable for physics analysis

Prompt processing up to NanoAOD



RAW → RECO → AOD → MINI → NANO

required CMSSW

flat ntuple-like does not require CMSSW

CMSSW = CMS Software

NanoAOD as exchange format

Bamboo

mkShapesRDF

columnflow

ROOT::RDataFrame

Awkward Array

mplhep

ROOT
Data Analysis Framework

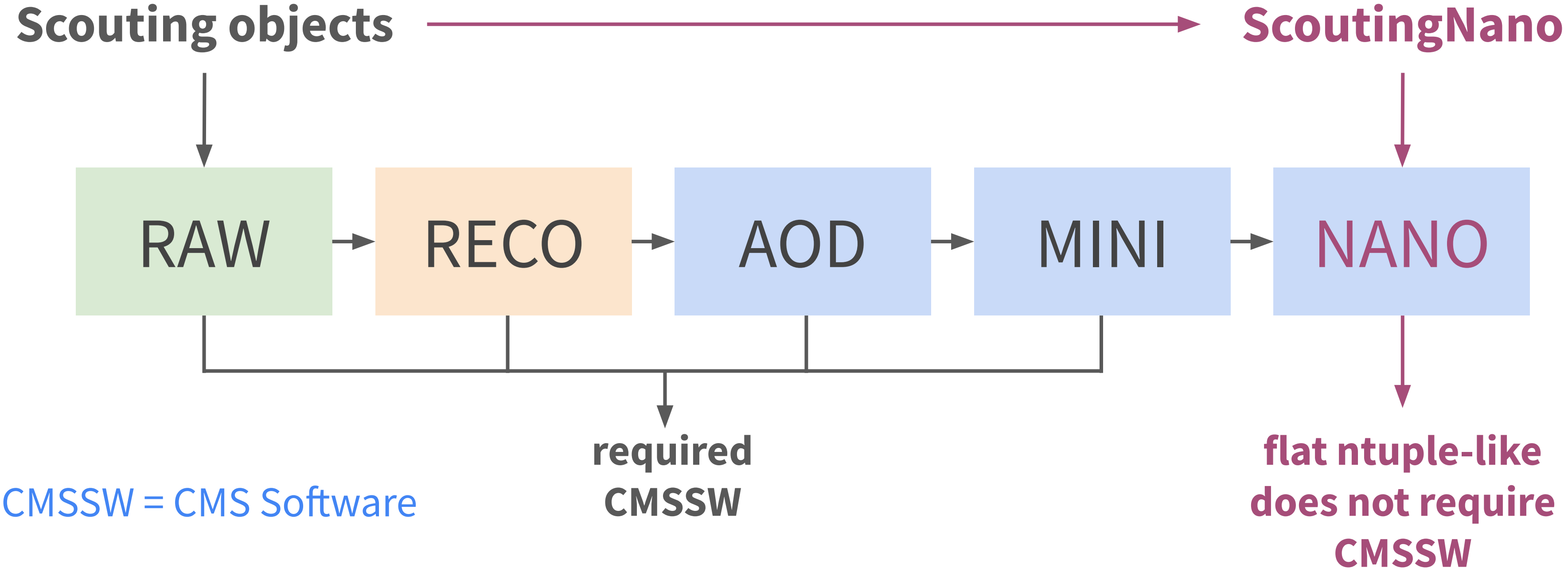
uproot

NANO

NanoAOD give analysers flexibilities to analyse data outside CMSSW

HLT Scouting in CMS datatier

Modified from CERN-CMS-NOTE-2023-003



Scouting objects → ScoutingNano

RAW → RECO → AOD → MINI → NANO

RAW, RECO, AOD, MINI → required CMSSW

NANO → flat ntuple-like does not require CMSSW

CMSSW = CMS Software

ScoutingNano

NanoAOD and Scouting

- **NanoAOD** is a flat ntuple-like format, suitable for most physics analyses
  - creating ntuple (ntuplising) is common across CMS analysers
  - central production of flat ntuple-like format helps reduce memory requirement across the collaboration
- to keep the size small, most used objects and their attributes are selected and the rest are drop during production
- However, some analyses require more → **custom NanoAOD**
- Since Scouting is a special stream, scouting objects are **not** included in standard NanoAOD → custom NanoAOD is needed

ScoutingNano is a custom NanoAOD with scouting objects

ScoutingNano

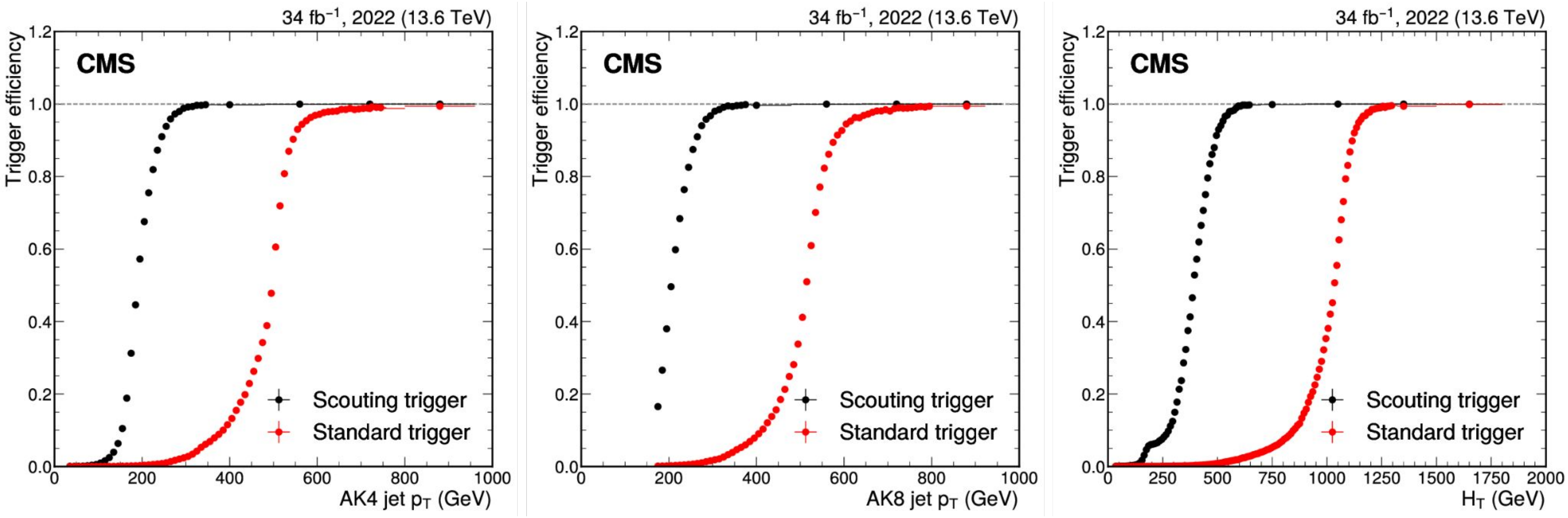
- ScoutingNano was initiated by previous PhD student for H→bb with Scouting analysis
- I continue her work, including ScoutingNano
  - Currently, developer and maintainer of ScoutingNano from this year
- ScoutingNano adds scouting objects, but also post-processes
  - Current: Jet tagging, Development: Jet Energy Correction, Secondary Vertexing
- NanoAOD can also ease quality monitoring → **good to have them in fast (prompt)**
- All ingredients for prompt processing are ready and tests passed last week

Prompt processing of ScoutingNano SOON

NanoScouting in Prompt #4991

ScoutingNano: usages so far

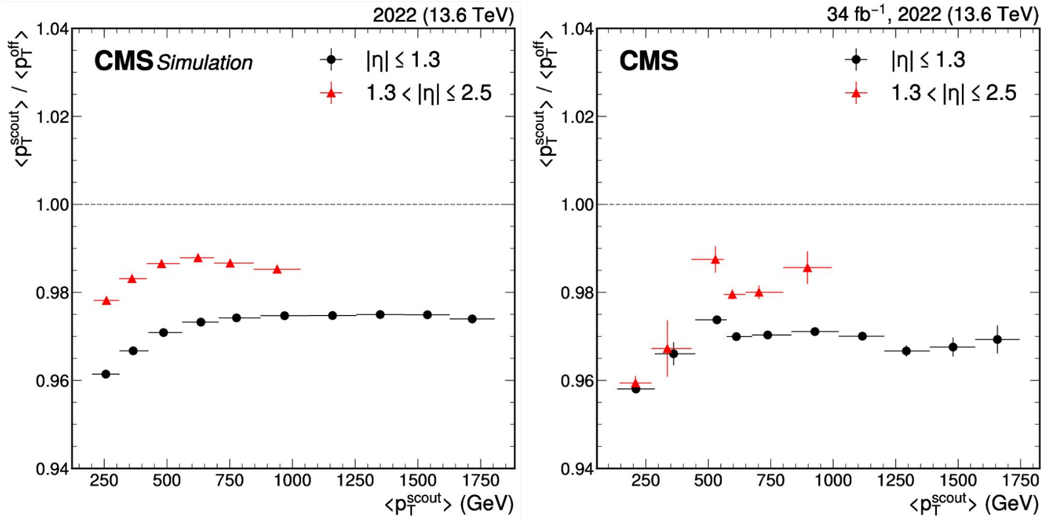
- Trigger efficiency studies (CMS-EXO-23-007)



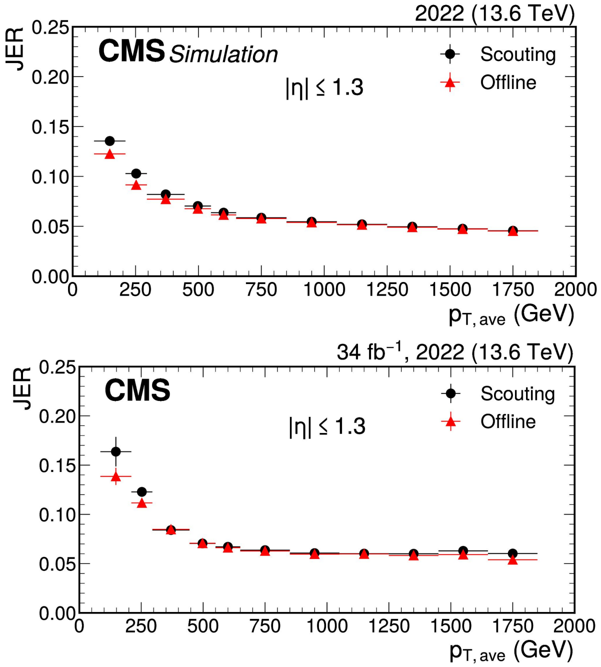
Jet and HT Trigger Efficiency

ScoutingNano: usages so far

- Object performance studies, e.g. Jet Energy Scale and Resolution (CMS-EXO-23-007)



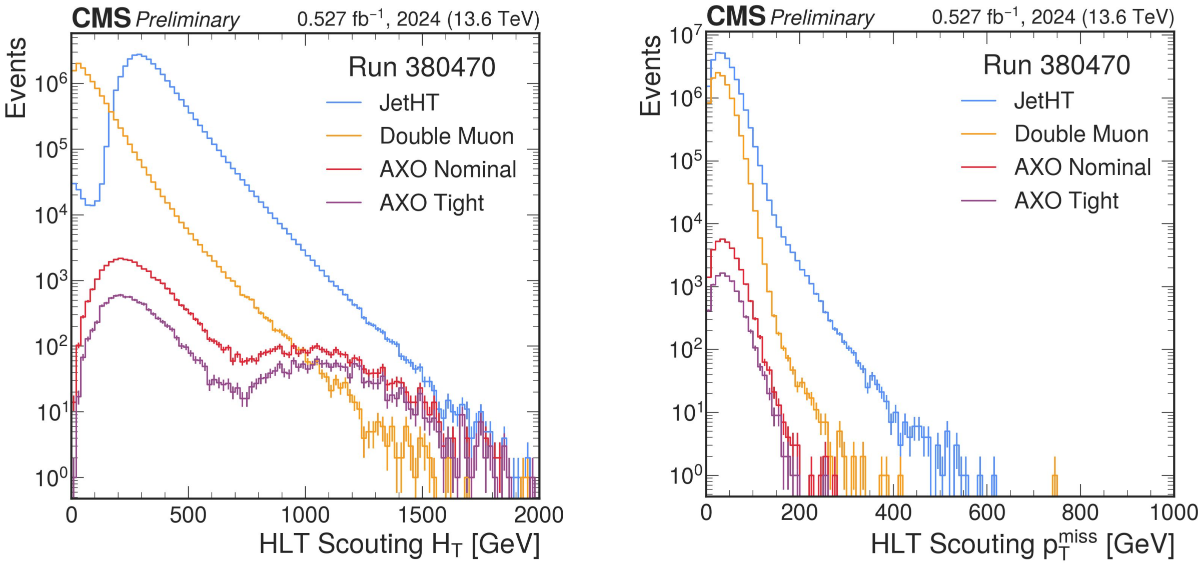
Jet Energy Scale



Jet Energy Resolution

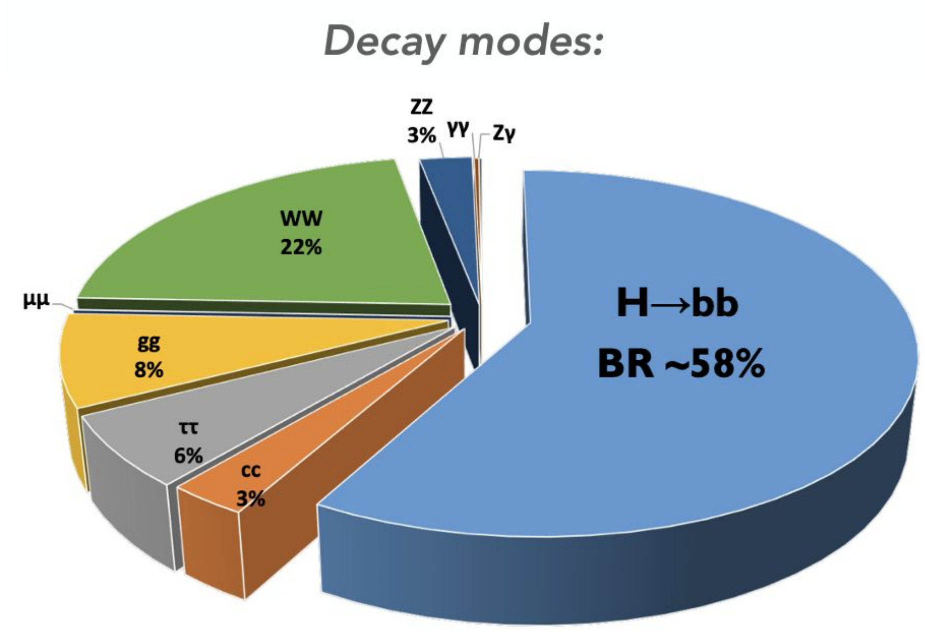
ScoutingNano: usages so far

- AXOL1TL paths in Scouting stream
- *2024 Data Collected with AXOL1TL Anomaly Detection at the CMS Level-1 Trigger* CMS-DP-2024-059

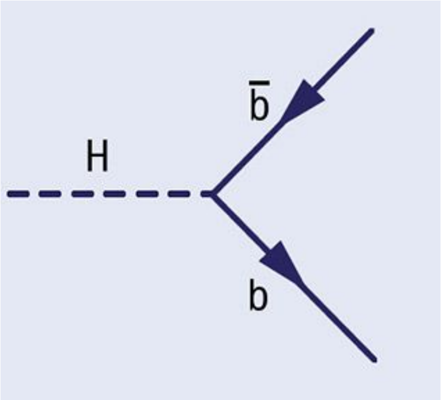


H → bb measurement using data scouting during run 3 at CMS

Motivation: Higgs decay modes



ATL-PHYS-SLIDE-2022-013



- H → bb is the most probable decay mode
- However, suffer from enormous QCD background

Motivation: Higgs production modes

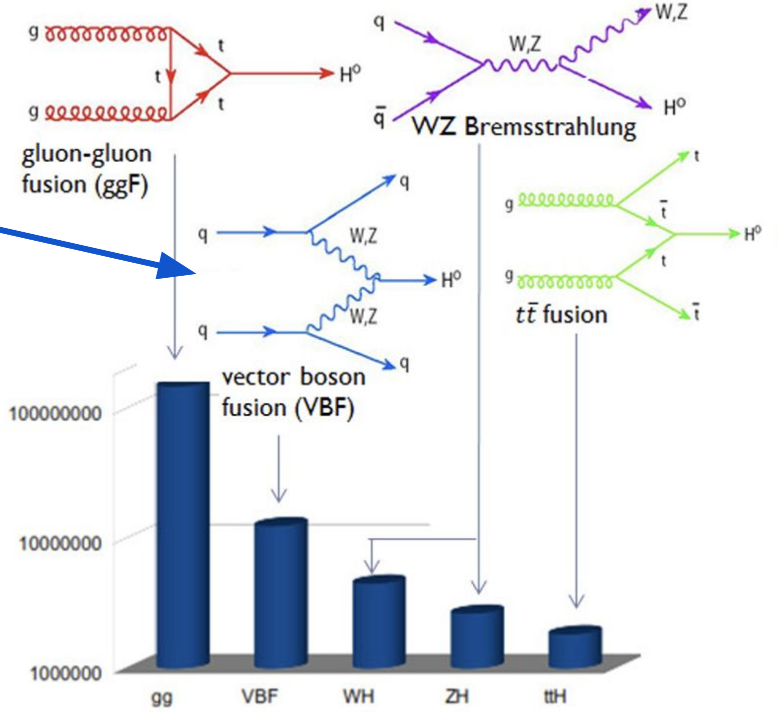
Gluon Fusion (ggF) and

Vector Boson Fusion (VBF)

→ most probable at LHC

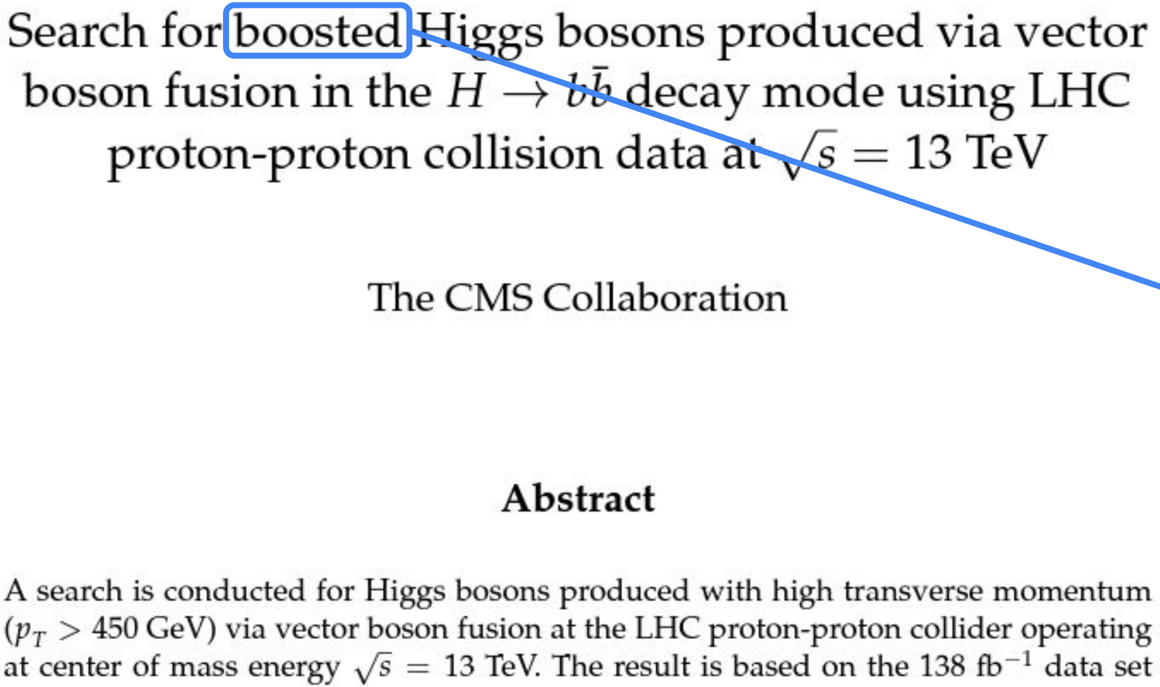
ggF/VBF H → bb is full hadronic search

→ challenging to observe due to large QCD background



ATL-PHYS-PROC-2014-205

Motivation: latest result



Search for boosted Higgs bosons produced via vector boson fusion in the $H \to b\bar{b}$ decay mode using LHC proton-proton collision data at $\sqrt{s} = 13$ TeV

The CMS Collaboration

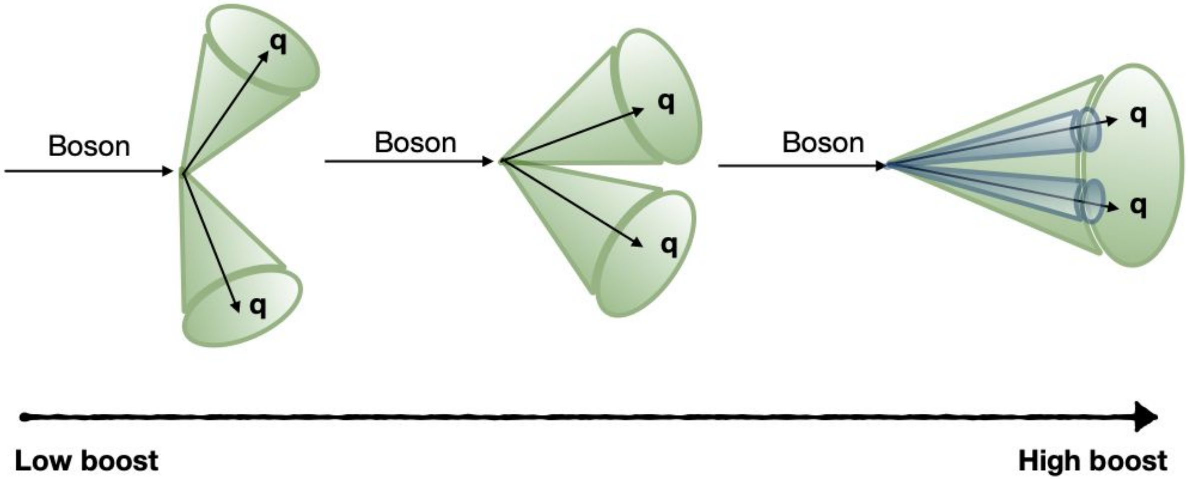
Abstract

A search is conducted for Higgs bosons produced with high transverse momentum ($p_T > 450$ GeV) via vector boson fusion at the LHC proton-proton collider operating at center of mass energy $\sqrt{s} = 13$ TeV. The result is based on the 138 fb^{-1} data set

Boosted?

CMS-PAS-HIG-21-020

Motivation: boosted jets

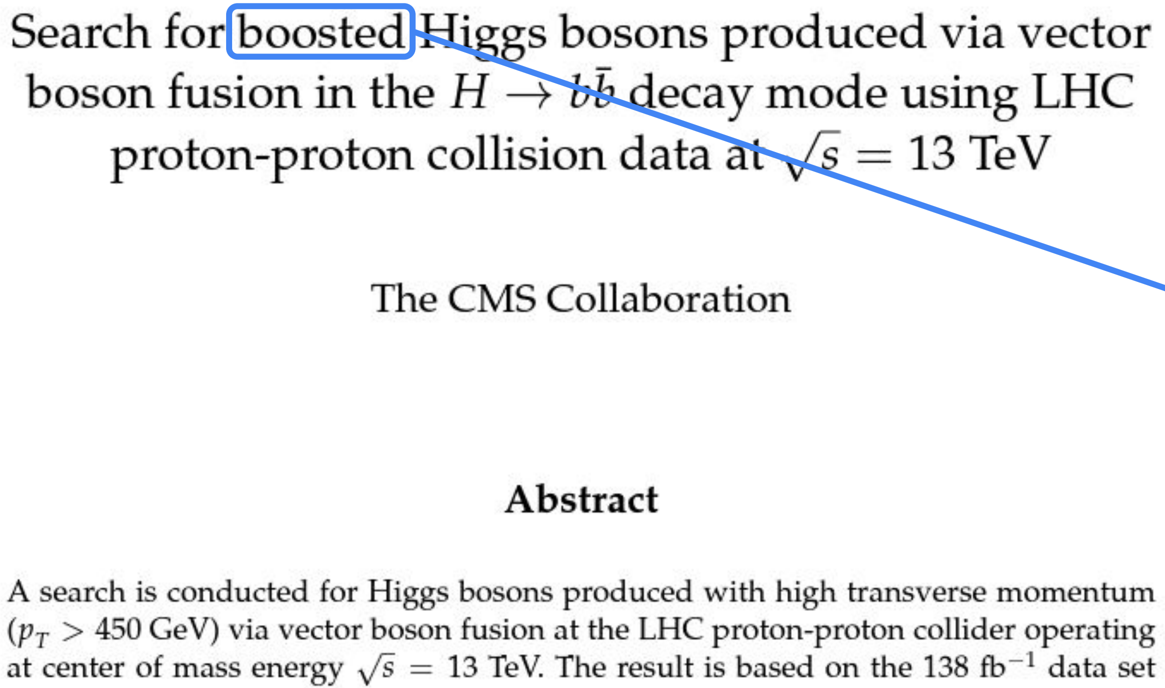


With high boost, two jets originating from single boson merge into single large jet.

Probing jet substructure can improve signal sensitivity from QCD background.

CMS-PHO-EVENTS-2022-018

Motivation: latest result



Search for boosted Higgs bosons produced via vector boson fusion in the $H \to b\bar{b}$ decay mode using LHC proton-proton collision data at $\sqrt{s} = 13$ TeV

The CMS Collaboration

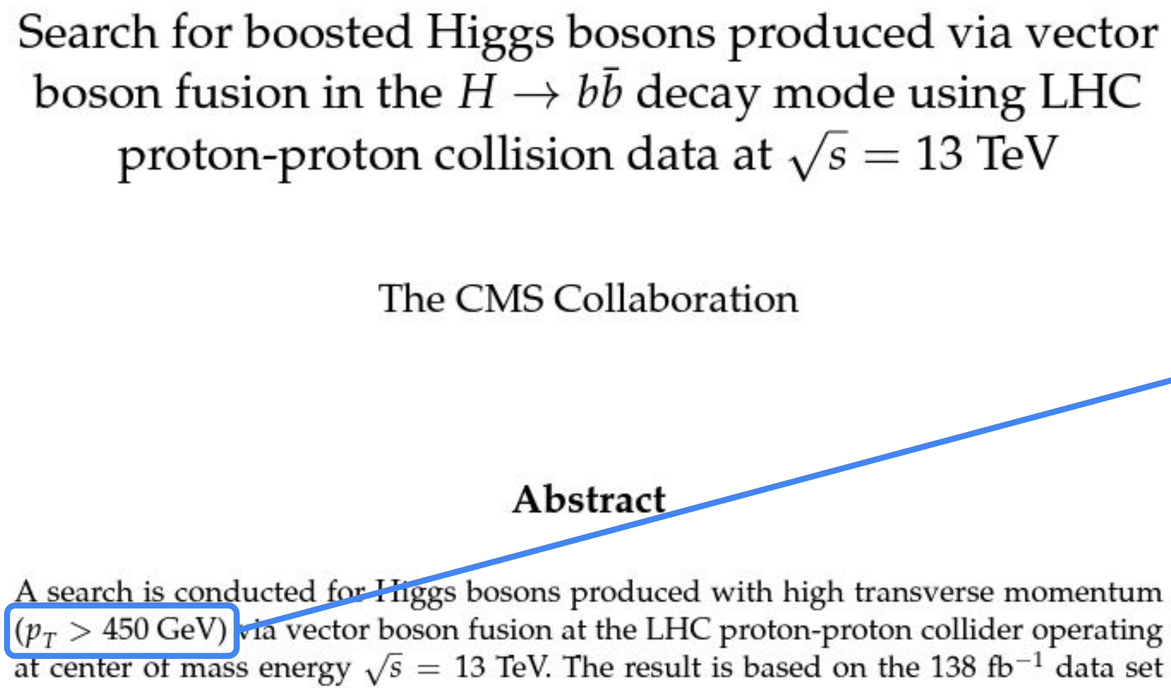
Abstract

A search is conducted for Higgs bosons produced with high transverse momentum ($p_T > 450$ GeV) via vector boson fusion at the LHC proton-proton collider operating at center of mass energy $\sqrt{s} = 13$ TeV. The result is based on the 138 fb^{-1} data set

For AK8, H→bb merged ≳ 300 GeV

CMS-PAS-HIG-21-020

Motivation: latest result



Search for boosted Higgs bosons produced via vector boson fusion in the $H \to b\bar{b}$ decay mode using LHC proton-proton collision data at $\sqrt{s} = 13$ TeV

The CMS Collaboration

Abstract

A search is conducted for Higgs bosons produced with high transverse momentum ($p_T > 450$ GeV) via vector boson fusion at the LHC proton-proton collider operating at center of mass energy $\sqrt{s} = 13$ TeV. The result is based on the 138 fb^{-1} data set

pT > 450 GeV? required from trigger efficiency

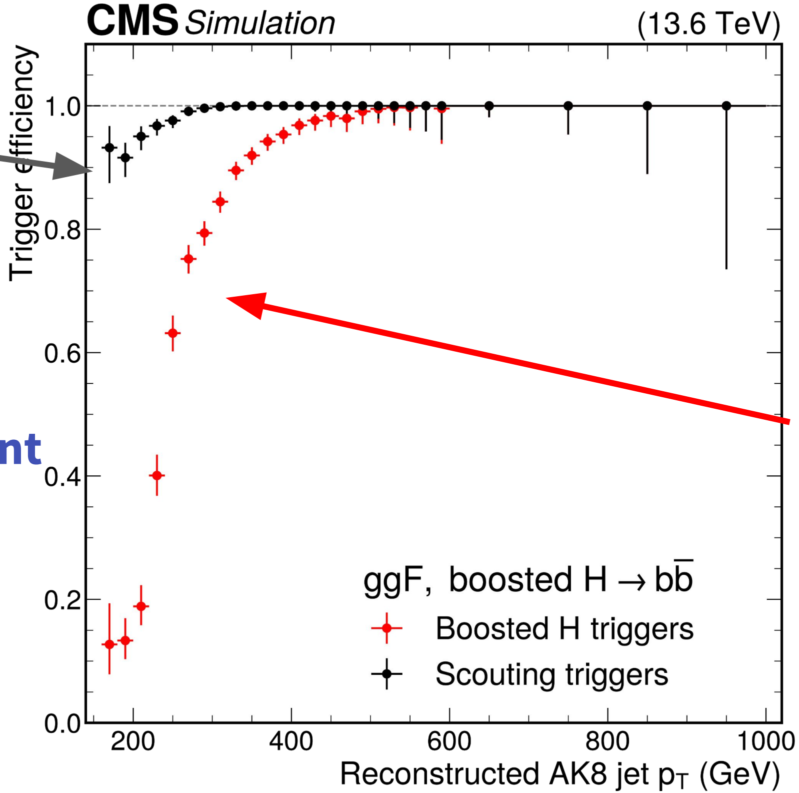
Can we lower this?

CMS-PAS-HIG-21-020

Trigger efficiency studies of the CMS Run-3 Data Scouting

Scouting Triggers (Run 3) mostly pass through for L1 decisions

Faster turn-on
→ can lower pT requirement



Preliminary study on simulation (ggF, VBF)

HLT triggers targeting boosted H→bb
based on ParticleNet

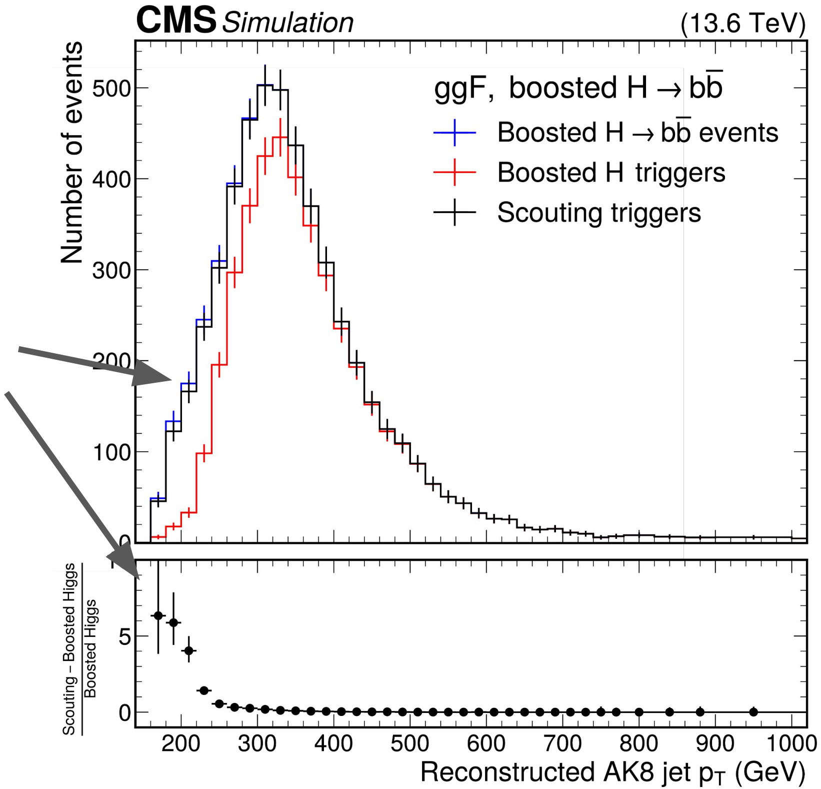
arxiv.org/abs/1902.08570

CMS-EXO-23-007

Trigger efficiency studies of the CMS Run-3 Data Scouting

particularly gains in low pT region

recover phasespace inefficient for standard triggers

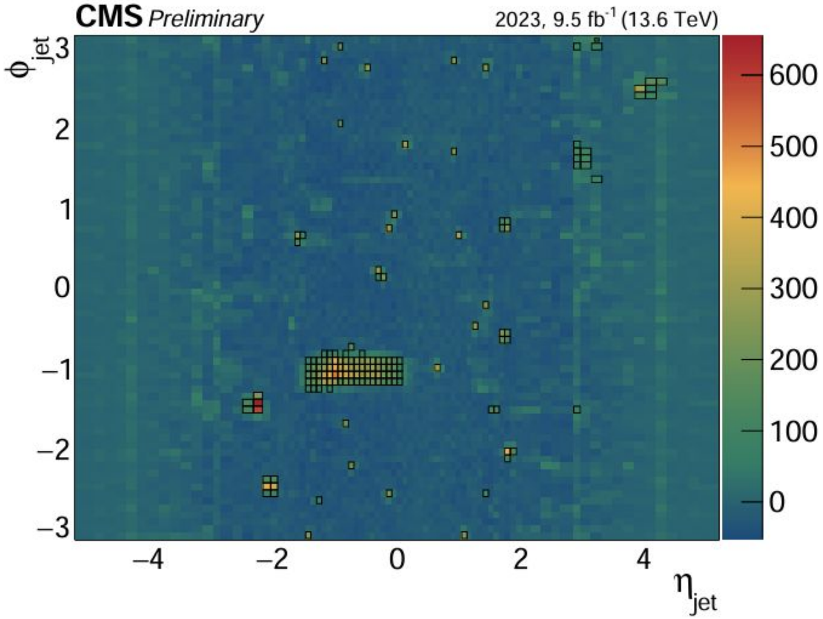


Overall number of events gain ~20%

CMS-DP-2023-076

Hbb: status and plan

- Development and performance studies of AK8 jet tagger
    - ParticleNet tagger was trained, but the performance can still be improved
    - Retraining to adapt for changing detector condition
    - ParticleNet → Particle Transformer (ParT)
- Analysis code
    - Update coffea to newer version
    - Learn and try Combine



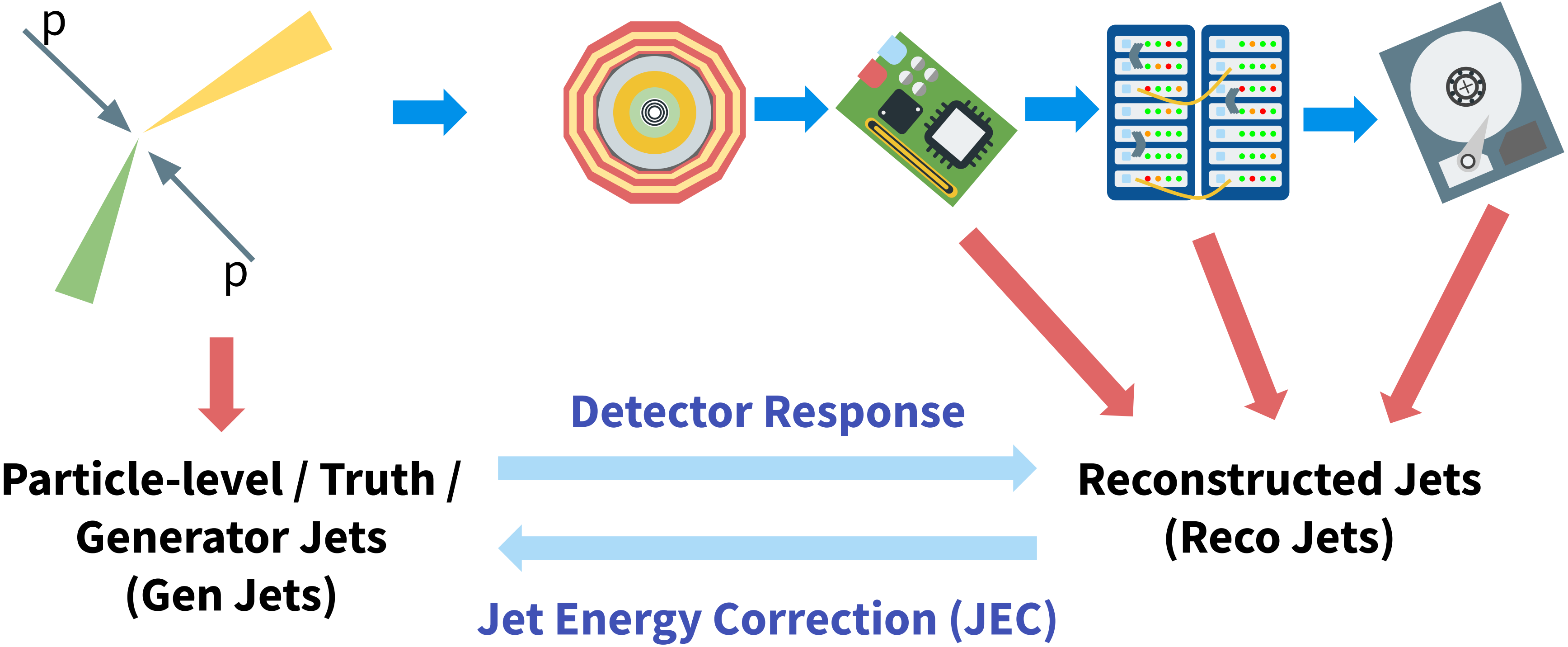
The CMS statistical analysis and combination tool: COMBINE

CMS-CAT-23-001

CMS-DP-2024-039

Jet Energy Correction (JEC) studies on scouting jets

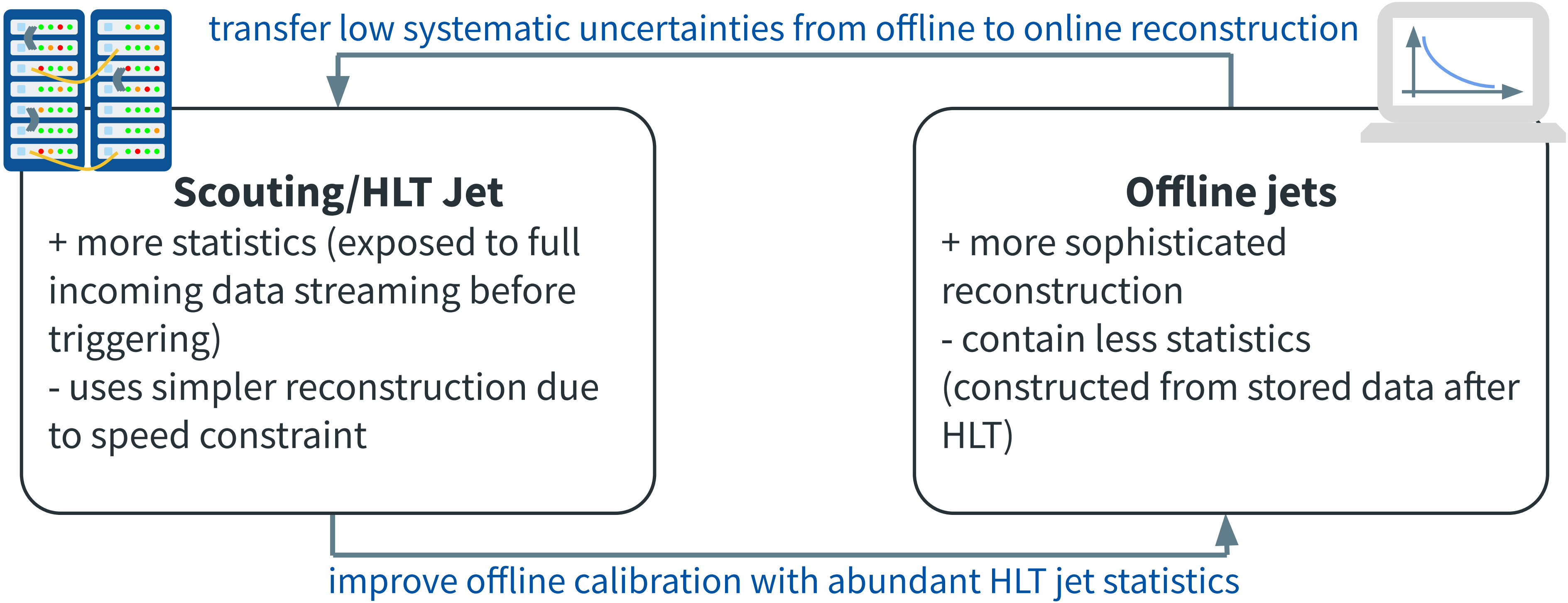
Introduction: JEC



Particle-level / Truth / Generator Jets (Gen Jets)

Reconstructed Jets (Reco Jets)

JEC for scouting jets



transfer low systematic uncertainties from offline to online reconstruction

Scouting/HLT Jet

+ more statistics (exposed to full incoming data streaming before triggering)

- uses simpler reconstruction due to speed constraint

Offline jets

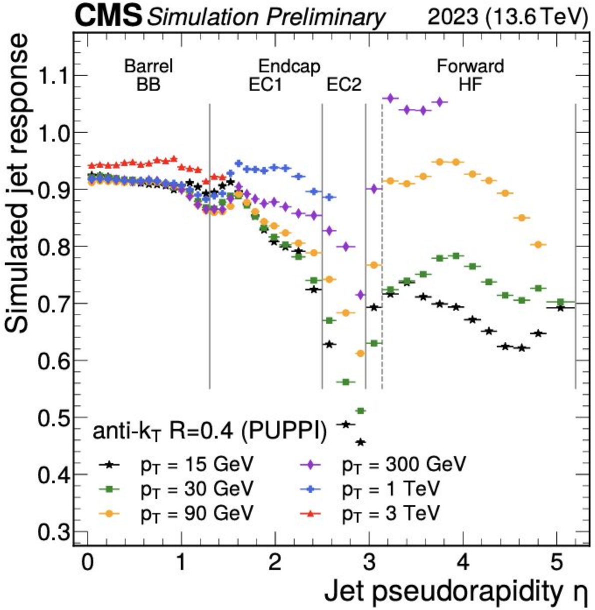
+ more sophisticated reconstruction

- contain less statistics (constructed from stored data after HLT)

improve offline calibration with abundant HLT jet statistics

JEC: status and plan

- was a starting project and some results were made → resume activities
- redo event yields and trigger efficiency to ensure enough statistics to prepare for next year data-taking
- update analysis code
- try ML, e.g. symbolic regression

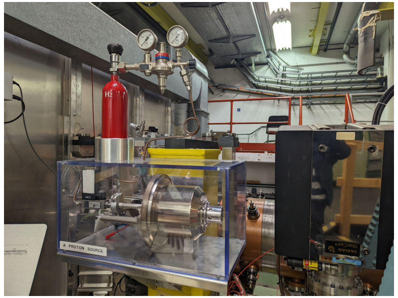


CMS-DP-2024-039

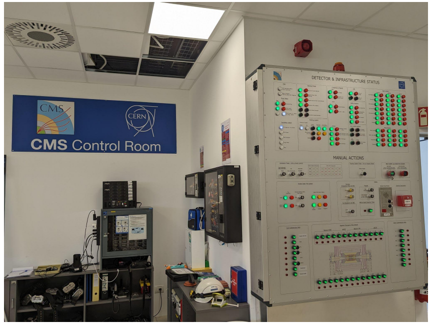
Other activities

Other activities

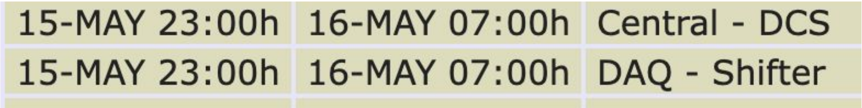
- New CMS role: JME-BTV-PF AlCa contact
- CERN summer stay May-June
  - DAQ, TRG, DCS shifters
  - guide training:
    - CMS underground guide
    - CERN visit guide: ATLAS visitor center, ALICE exhibition, Data Center, Antimatter factory (AD), LIER/LINAC2, SM18, CERN Control Center (CCC)



Proton source Model



New CMS control room



15-MAY 23:00h	16-MAY 07:00h	Central - DCS
15-MAY 23:00h	16-MAY 07:00h	DAQ - Shifter

Double DCS-DAQ night shift

Murder in the control room: who killed the DAQ duck !?

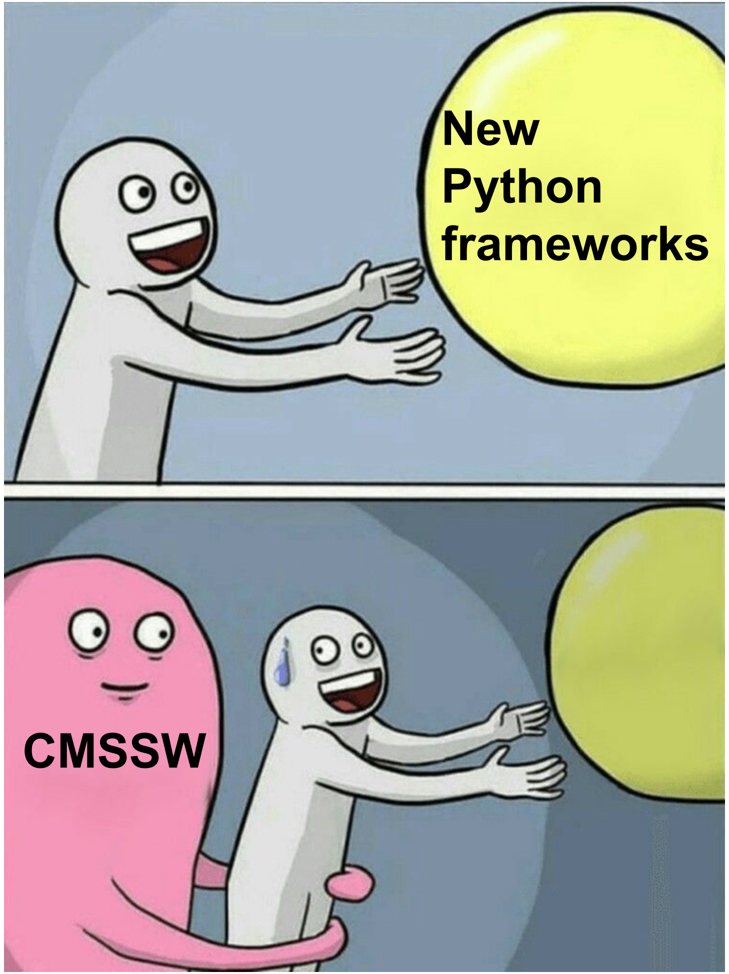
Other activities

- PAPU Fall Seminar (22 November 2022): **lightning talk!**
- CMS Week December 2022 (5 - 9 December 2022)
- Spåtind 2023: Nordic Conference on Particle Physics (3-8 January 2023): **talk!**
- JetMET Workshop (15 - 17 May 2023)
- Stay at CERN (1 June - 20 August 2023): **shifts + summer project supervision!**
- CMS Data Analysis School (5 - 10 June 2023)
- CMS Week June 2023 (12 - 16 June 2023)
- 13th Patatrack Hackathon (26 - 30 June 2023)
- Advanced Artificial Intelligence for Precision High Energy Physics (16 - 28 July 2023)
- CERN School of Computing (20 August - 2 September 2023): **lightning talk!**
- Researcher Night (29 September 2023): **outreach!**
- Particle Physics Day (12 October 2023)
- ML4Jets (4 - 6 November 2023)
- ML@L1 Workshop (11 - 15 December 2023)
- Physics Day (4-6 March 2024): **organisation!**
- Group retreat (3-8 March 2024)
- Midsummer school in QCD (24 June - 6 July 2024)
- Edge ML school (23-27 September 2024)

Conclusion

Can we even analyse these data?

I don’t know, but hopefully we can


New Python frameworks
CMSSW

Omg this CMSSW sucks, I hate it
Then just stop playing it

Enjoying Finland

Kotka

Kilpisjärvi

Saariselkä

Porvoo

Conclusion

- ScoutingNano is custom NanoAOD with scouting objects,
  allowing more accessibilities to scouting data
  and potentially utilising analysis frameworks currently in development
- ScoutingNano in prompt processing at T0 soon^{TM}
- Exploiting scouting stream in H→bb can increase
  overall statistics by ~20% with particularly gain in low p_T region,
  inefficient by standard trigger
- Lots of works to do:
  - Retrain tagger and update to ParT
  - Resume JEC activities
- VERIZON Secondment: plan this week → start next week

Espoo

Backup

About Me

Name: Patin Inkaew (PI ~ 3.14)

Nickname: Earth

Birthday: 22 July 1998 (22/7 ~ 3.14)

Hometown: Bangkok, Thailand

Institution: University of Helsinki (UH), Helsinki Institute of Physics (HIP)

Contract start: 01/10/2022

Education

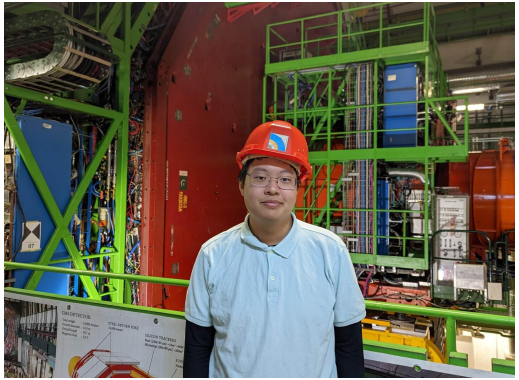
Stanford University, CA, USA (Thai Government Scholarship)

Coterminal program (Joint BS+MS) in 4 years

BS: Physics, Minor: Mathematics, East Asian Studies (Japan subplan)

MS: Computer science (AI track)

Research: Many things: laser, detector design, ML, CV, CG, ComBio,
particle physics analysis



ESR1:

Machine learning and Real-Time Analysis for Higgs boson measurements and fleet safety

PhD:
University of Helsinki (UH) &
Helsinki Institute of Physics (HIP),
Finland

Supervisors:
Mikko Voutilainen,
Henning Kirschenmann

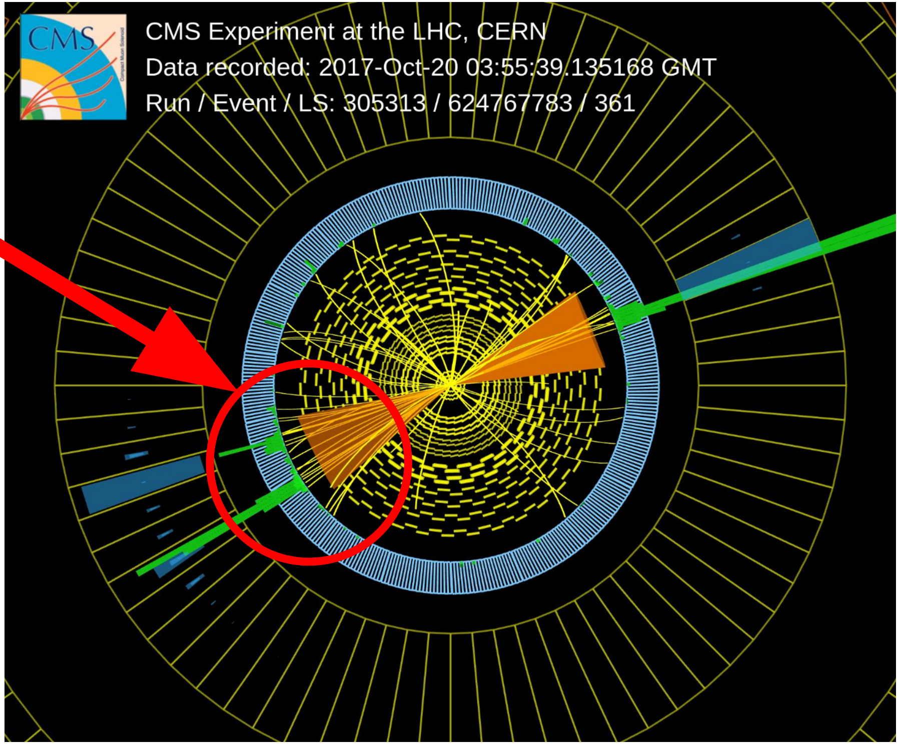
Secondment:
CERN, Switzerland

Collaborator:
Maurizio Pierini

Secondment:
Verizon Connect, Italy

Collaborators:
Leonardo Taccari,
Francesco Sambo

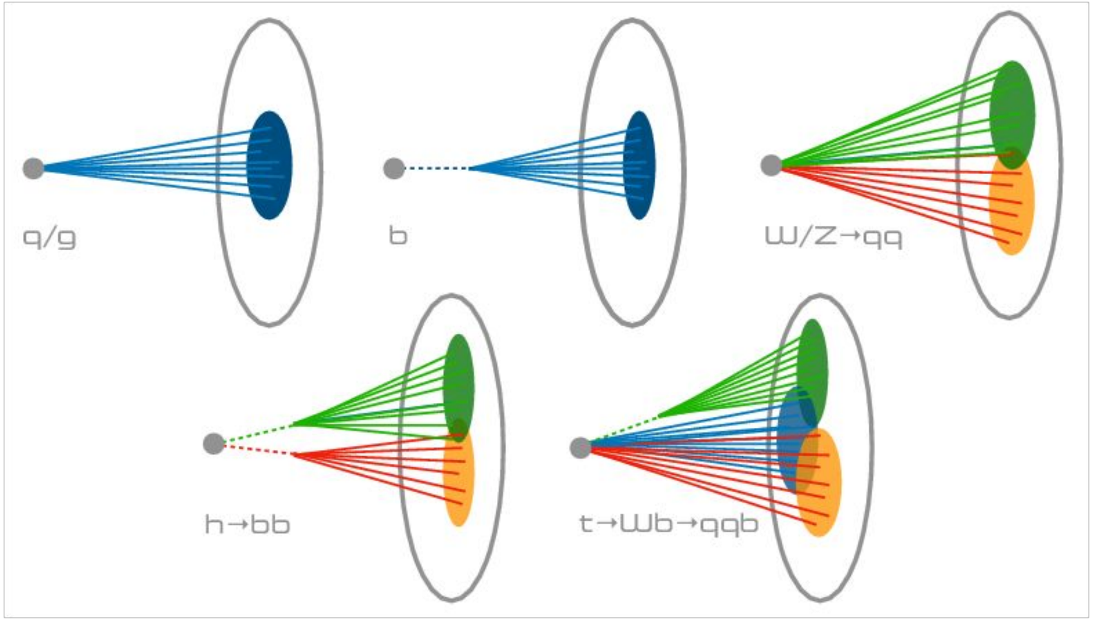
Motivation: boosted jets


CMS Experiment at the LHC, CERN
Data recorded: 2017-Oct-20 03:55:39.135168 GMT
Run / Event / LS: 305313 / 624767783 / 361

2 subjets

CMS-PAS-HIG-19-003

Motivation: jet substructure

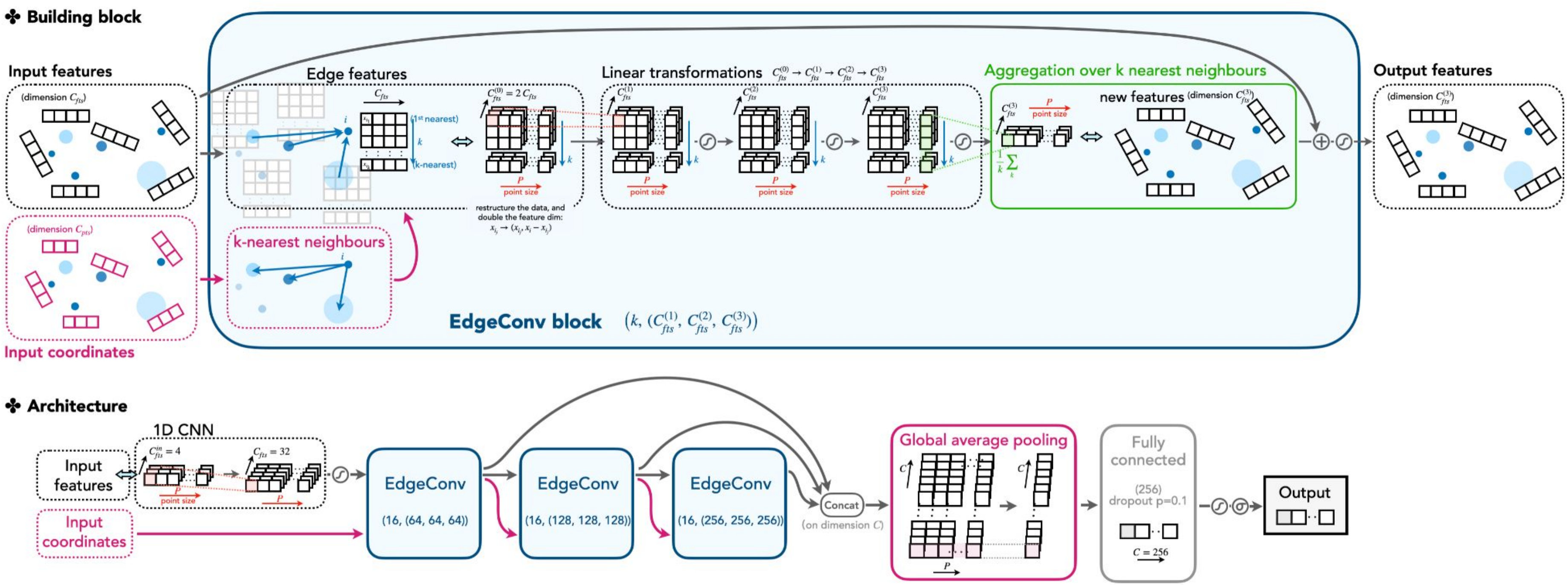


arxiv.org/abs/1909.12285

Jet structure indicates type of original particles

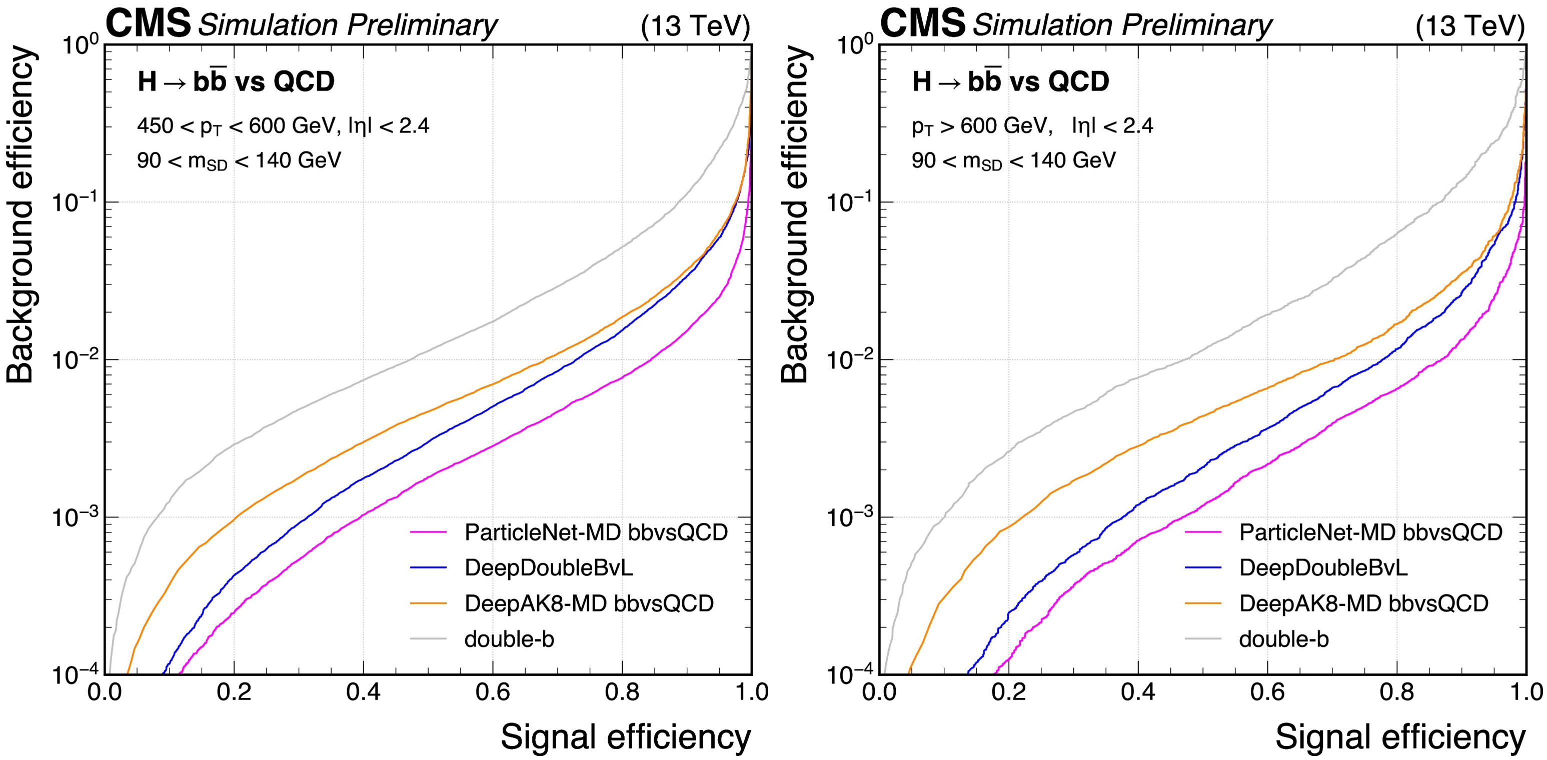
→ **jet tagging,** e.g. with neural network (ParticleNet, ParT, etc.)

ParticleNet



[CMS Machine Learning Documentation - ParticleNet]

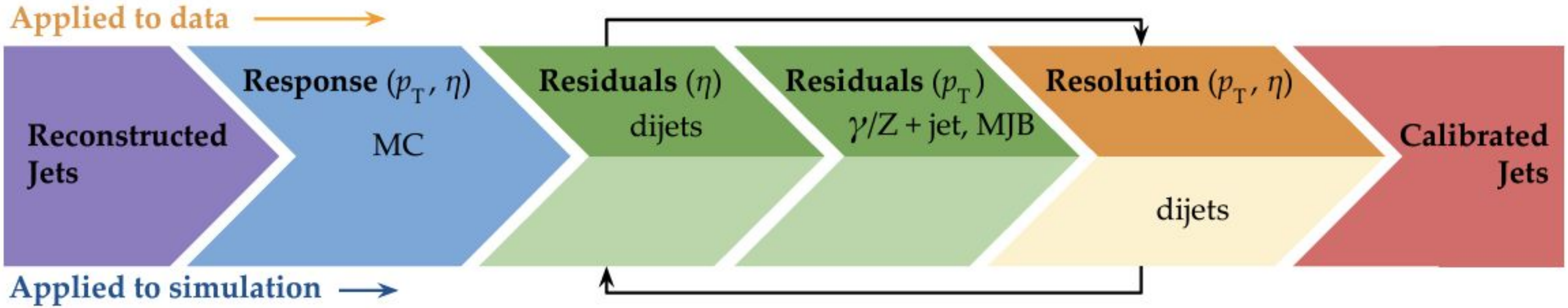
B-tagging performance (Run 2)



CMS-PAS-BTV-22-001

JEC in CMS Run 3

- Jet is clustered from PF candidates by **anti-kt algorithm** with R=0.4 or R=0.8
- **PUPPI (PileUp Per Particle Identification)** is applied to mitigate effects from pileup
- JEC is then applied: factorized approach - each step aims to correct specific effect



CMS-DP-2022-054