Accelerated AI Algorithms for Data-Driven Discovery

Real-time AI and Heterogeneous compute

Philip Harris

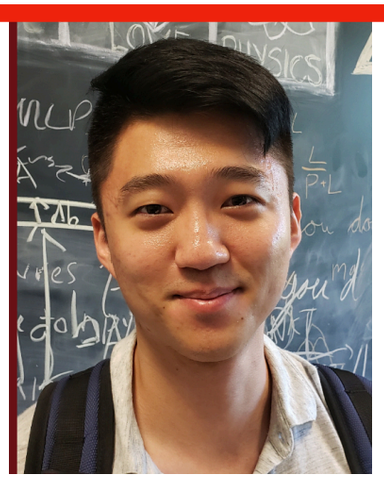
About Me

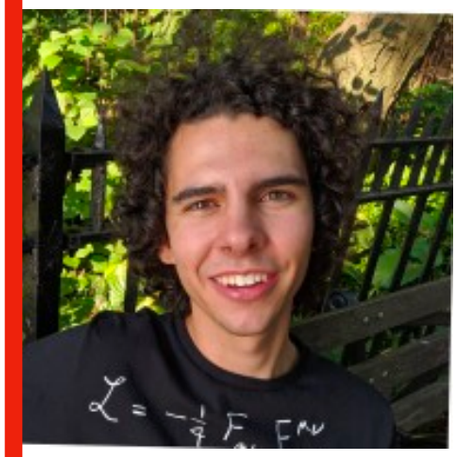
- PH: Associate Professor w/o tenure since 2023
  - CERN Fellow and CERN Staff before that
- On the CMS experiment @LHC :
  - L1 Correlator Trigger Upgrade Covener
  - GPU integration into offline computing w/SONIC project
  - CMS BSM Physics Representative for the LHC (2024)
- Outside the CMS experiment (Not DOE-HEP):
  - Founding member of the Fast Machine Learning Group
    - A3D3 Deputy director
  - Member of the Spinquest collaboration
  - IAIFI Experimental Physics coordinator
  - LIGO-Virgo-Kagra Analysis & MMA trigger with AI/ML

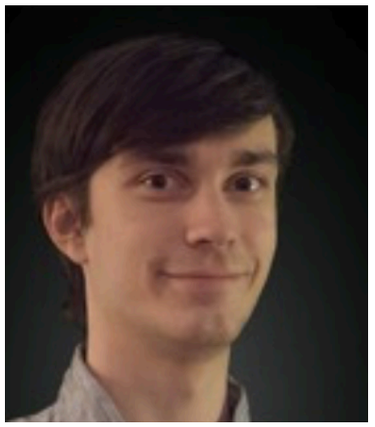
The Group

Professor

Postdocs







Ph.D/Postbacs

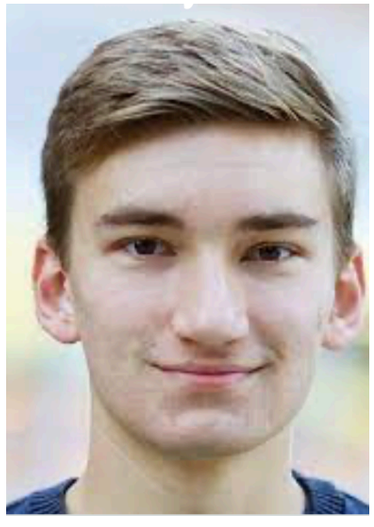
DOE-HEP

Current funding from DOE-HEP is just early career award
(Other NSF/ASCR/SNF)

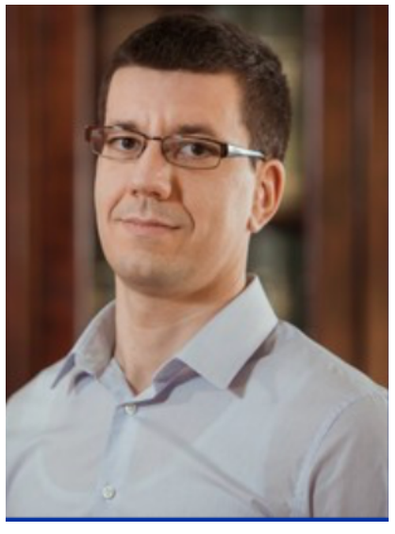
Group Migration(2 years)

Left Jan, 2023
Faculty at U Penn
(ECA 2024)

Left Jan 2024
Fidelity AI



Ph.D 2024
Postdoc at SLAC
(Started Yesterday)



CERN Staff
Scientist
(Started Last week)

M.S 2023
Left Sept, 2023
Ph.D Student JHU

Driving Questions

What can we say about new physics w/Higgs Boson?

What can we say about the nature of dark matter?

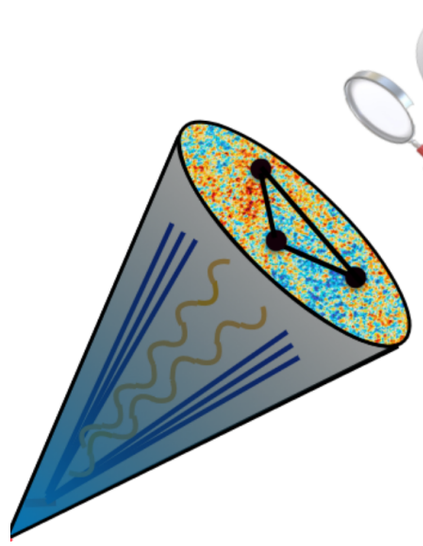
How do we harness the AI/ML revolution?

How do we automate/ensure full new physics coverage?

How do we bring these ideas into real-time?

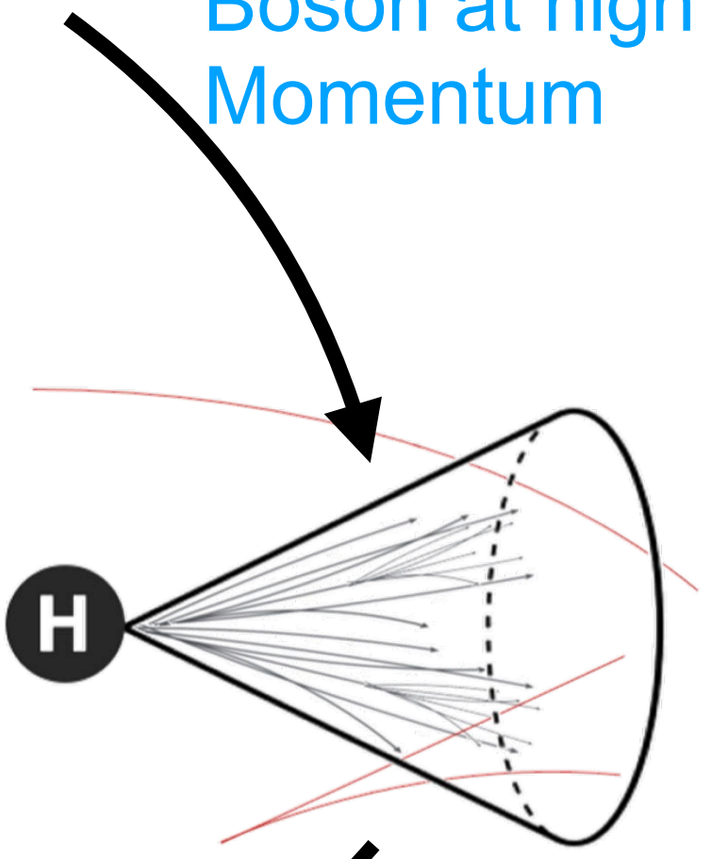
Cycle of LHC Research

We focus on jets of all kinds

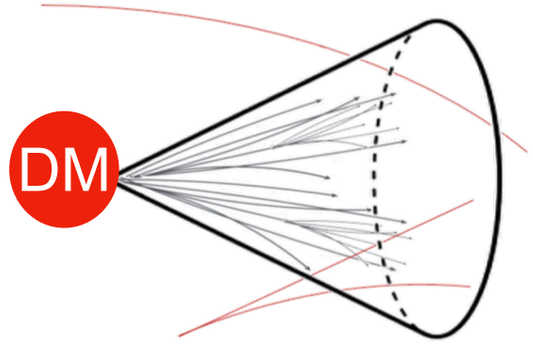


Core QCD Measurements

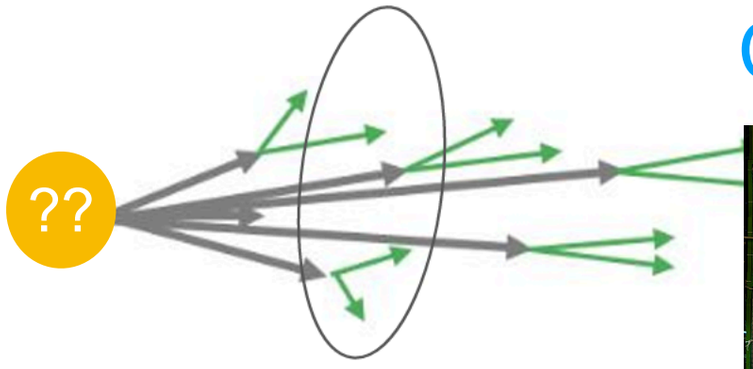
Probing Higgs Boson at high Momentum



Searching For (light) DM

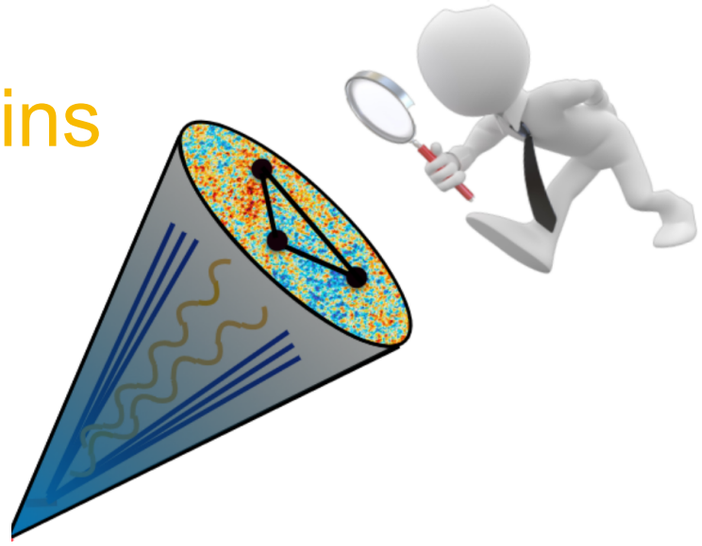


Searching For any anomaly Using New(AI) tech

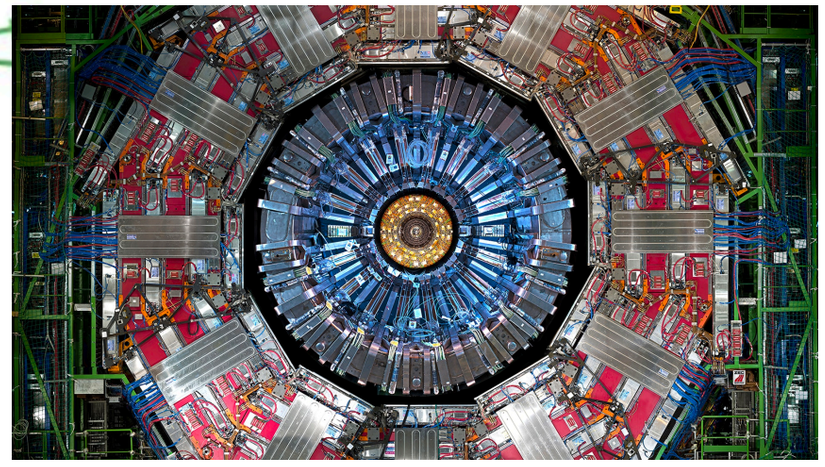


Cycle of LHC Research

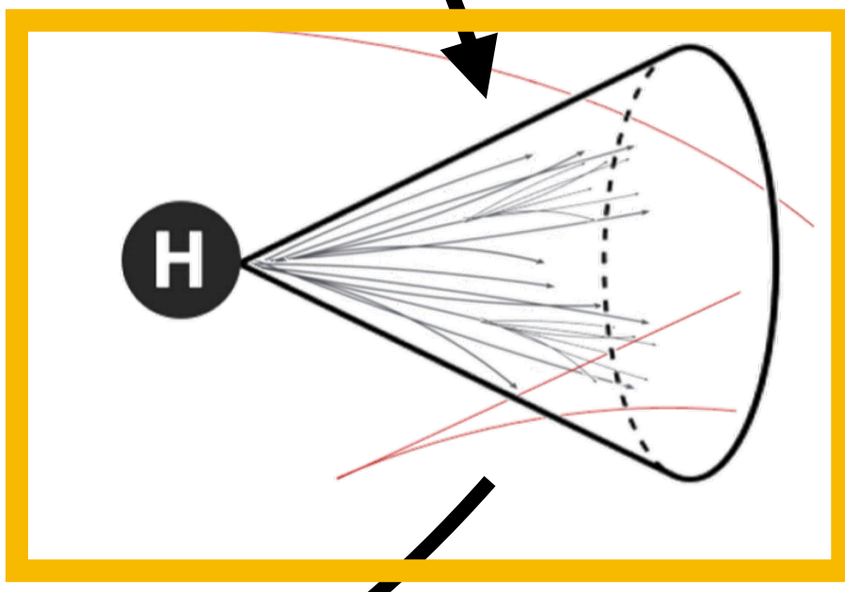
Bringing AI/ML to these domains
Has been a focus of
recent times



Core QCD Measurements

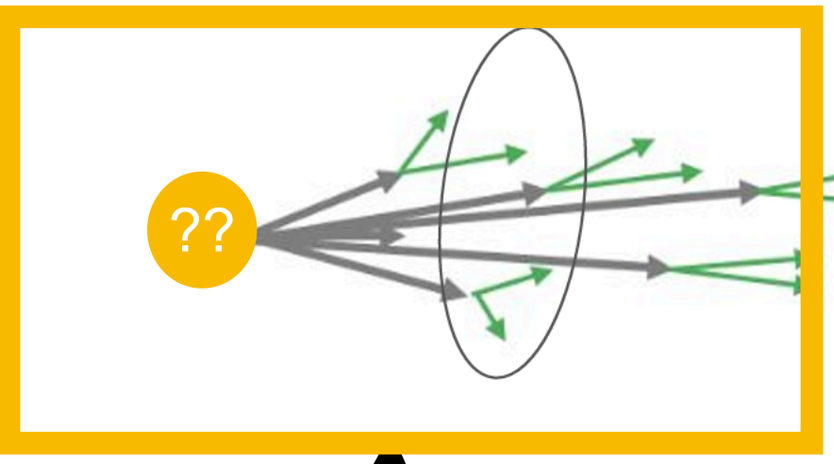


Probing Higgs
Boson at high
Momentum



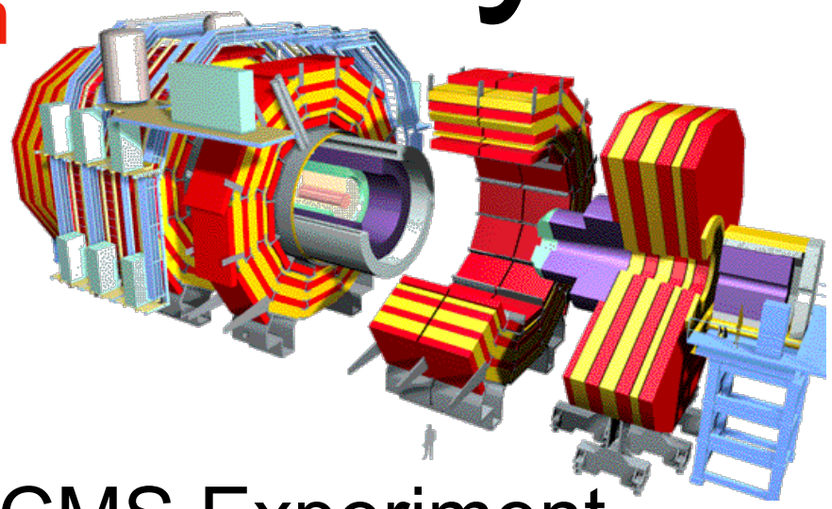
Searching
For (light) DM

Searching
For any anomaly
Using New(AI) tech



How do we bring
Ideas from the offline
Workflow into the data chain
Of the LHC

Cycle of Data

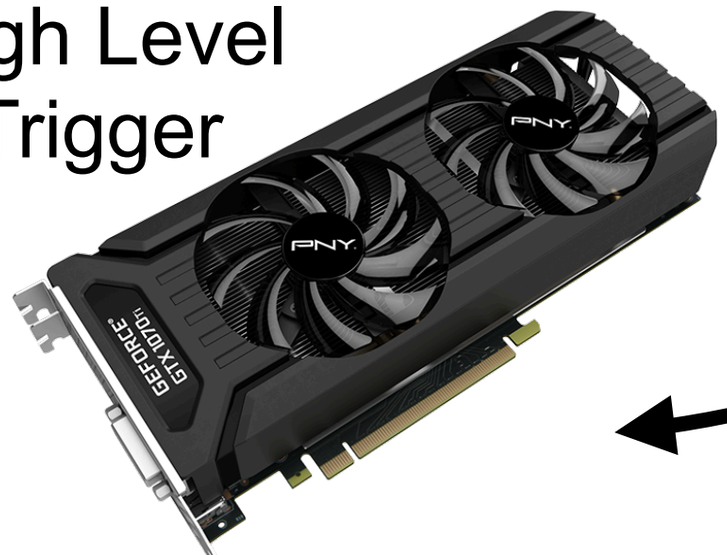


CMS Experiment

40 MHz

L1
Trigger

100 kHz



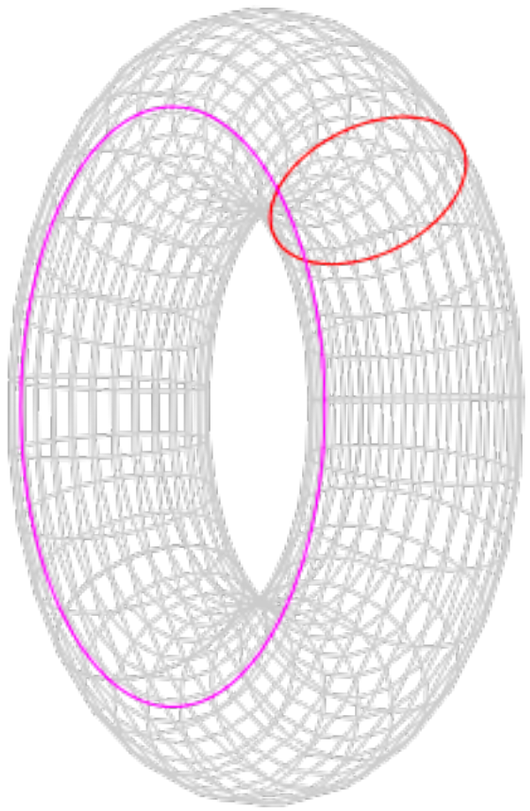
High Level
Trigger

1 kHz

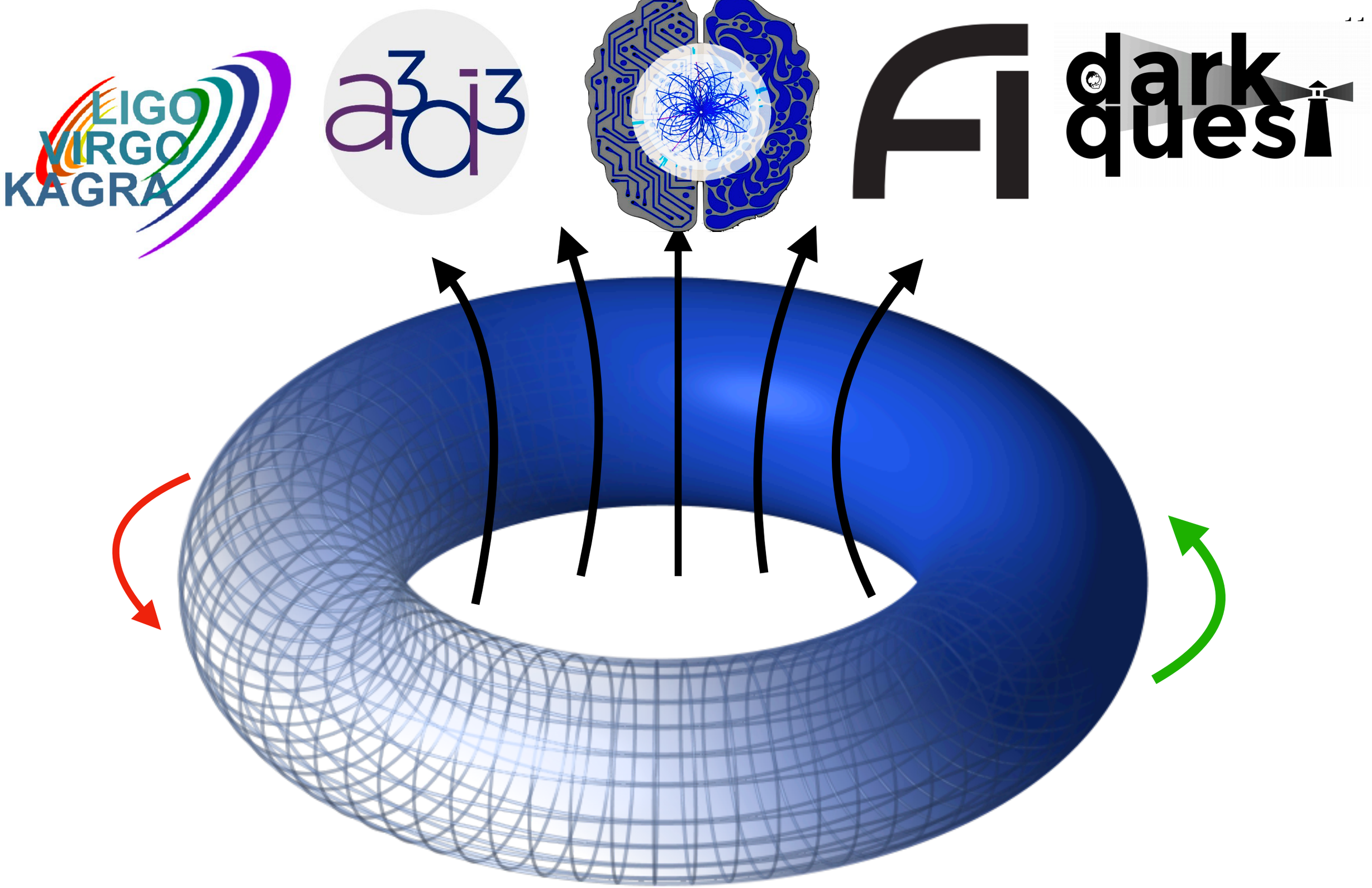
Offline
Data Reconstruction

High Level Focus

- Our Strategy
  - Focus on most critical elements in DAQ/Analysis
  - Continually work to bring AI/ML to LHC throughout
  - Focus less on the work internally in collaboration(ATLAS/CMS)
    - Be concerned about the overall impact of this work

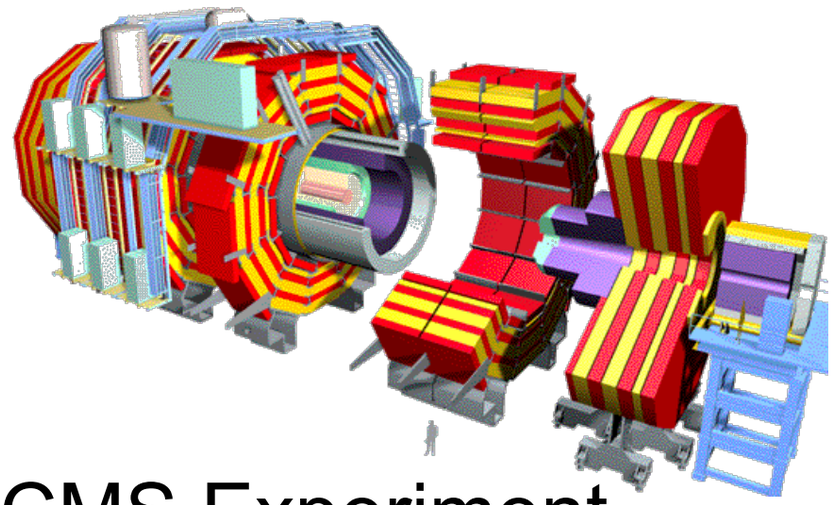


Embedding two circular flows yields a toroid


LIGO
VIRGO
KAGRA
a3d3
dark
quest
External Applications
Inductive Bias

Detector Work

Detector Research



CMS Experiment

40 MHz

L1
Trigger

100 kHz

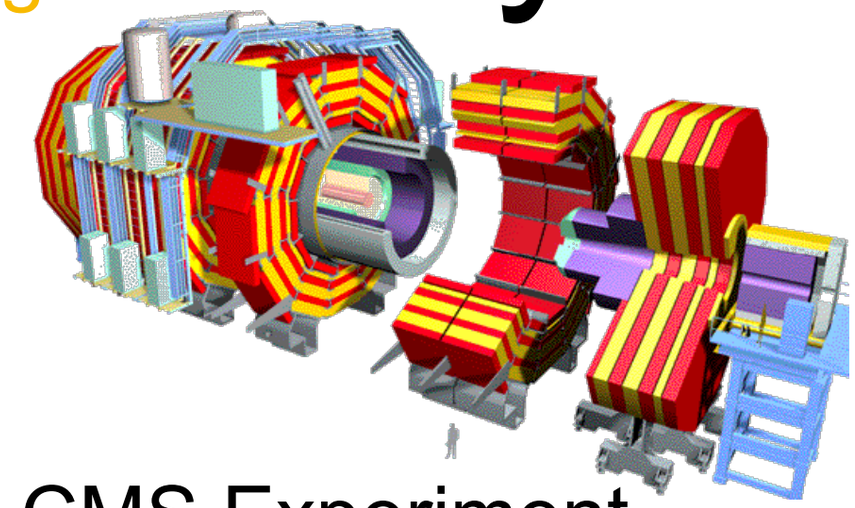
High Level
Trigger

1 kHz

Offline
Data Reconstruction

How do we integrate AI/ML
& Heterogeneous computing
Throughout the LHC

Cycle of Data



CMS Experiment

40 MHz

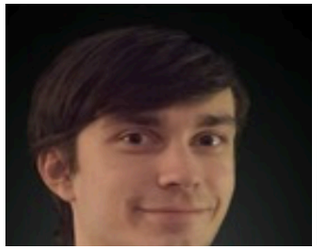
L1
Trigger

100 kHz

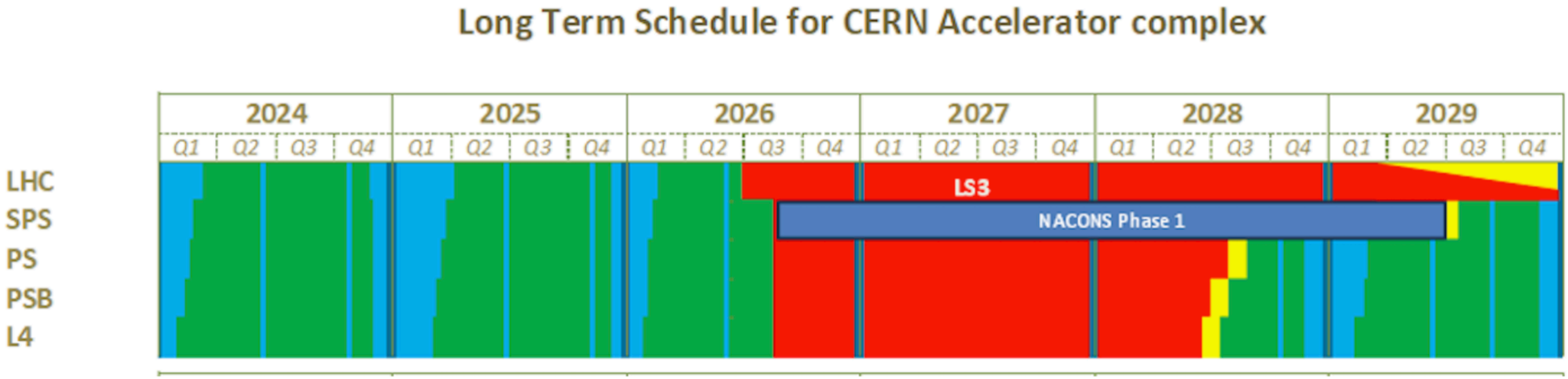
High Level
Trigger

1 kHz

Offline
Data Reconstruction




Goals



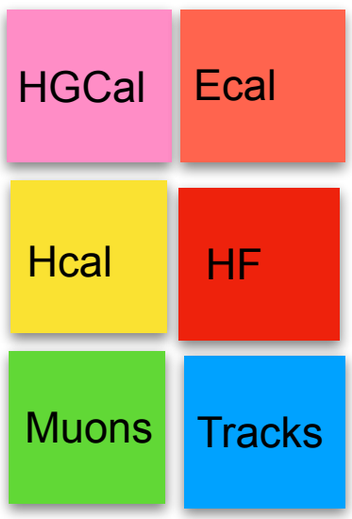
- Our focus is on resources for the HL-LHC upgrade
  - L1 Trigger: leading correlator effort
  - Computing: leading offline GPU integration(SONIC) effort
- Run 3 operations:
  - We participate in operations of L1 Trigger and Hcal

L1 Trigger Upgrade

- When started at MIT, wanted to build PUPPI for the L1
  - Appeared possible given the planned CMS upgrade

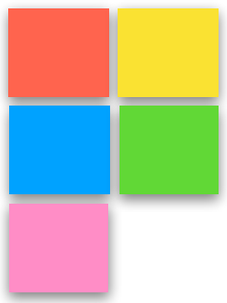
PF takes in everything

Assemble All detectors



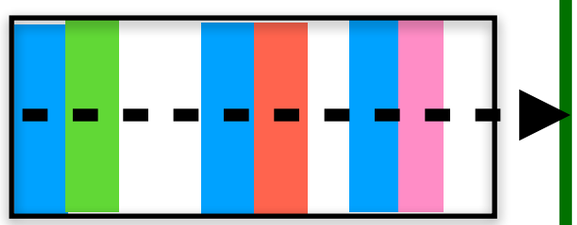
PF is local

Assemble Things in local regions η x φ



PF Links

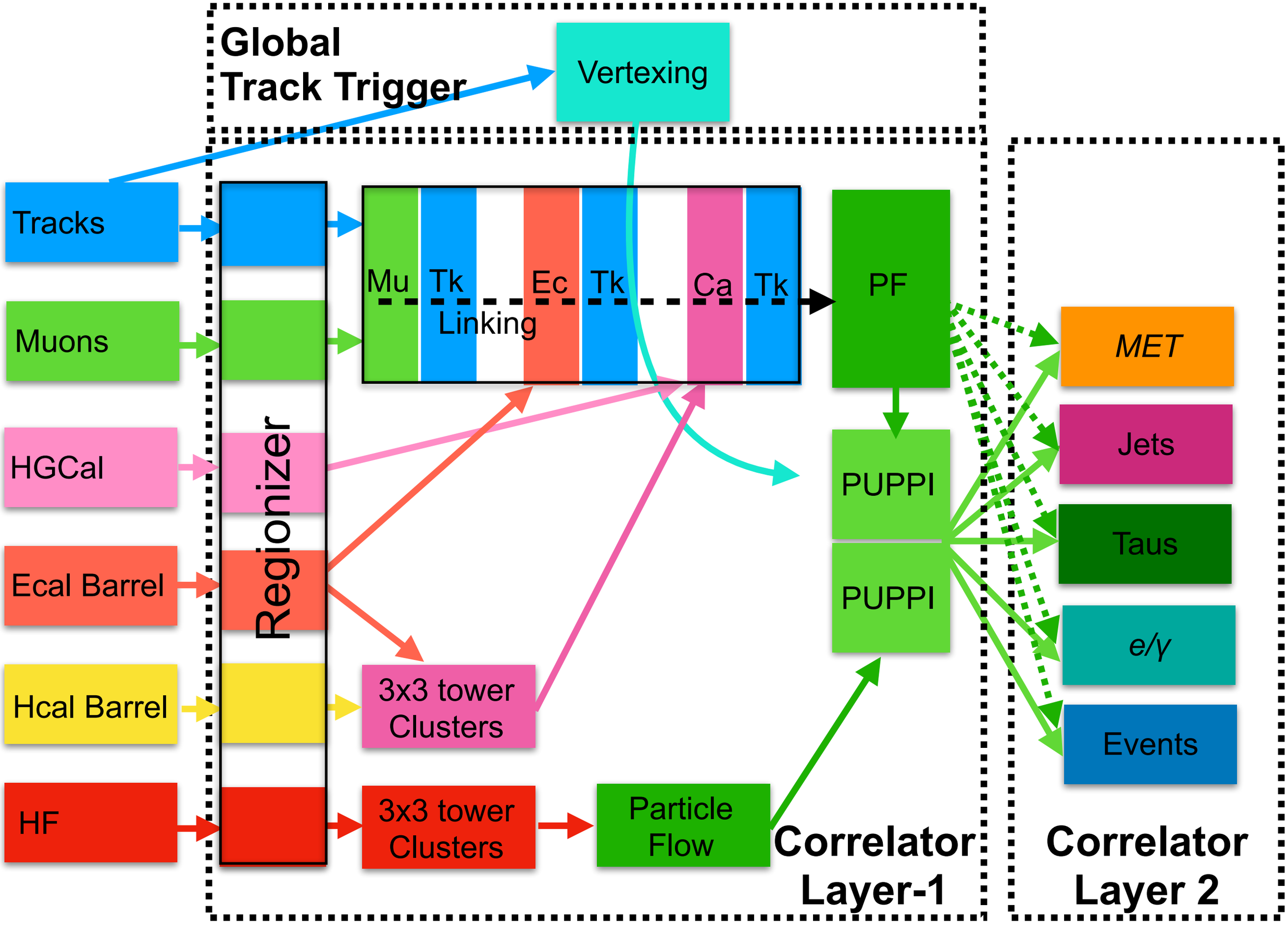
Particle flow combines everything together to Particls



PU is Local

PUPPI is local
Can parallelize it by region

Also its the CMS default PU algo


Global
Track Trigger
Vertexing
Tracks
Muons
HGCal
Ecal Barrel
Hcal Barrel
HF
Regionizer
Mu
Tk
Ec
Tk
Ca
Tk
Linking
3x3 tower
Clusters
3x3 tower
Clusters
Particle
Flow
PF
PUPPI
PUPPI
MET
Jets
Taus
e/γ
Events
Correlator
Layer-1
Correlator
Layer 2

Progress

Implemented

Working Implementation

To be done

Global Track Trigger

Vertexing

Tracks

Muons

HGCal

Ecal Barrel

Hcal Barrel

HF

Regionizer

Mu Tk Ec Tk Ca Tk

Linking

PF

Tuning & Improvement

Tuning & Improvement

3x3 tower Clusters

3x3 tower Clusters

Particle Flow

PUPPI

PUPPI

Correlator Layer-1

Our Focus Right Now

MET

Jets

Taus

e/γ

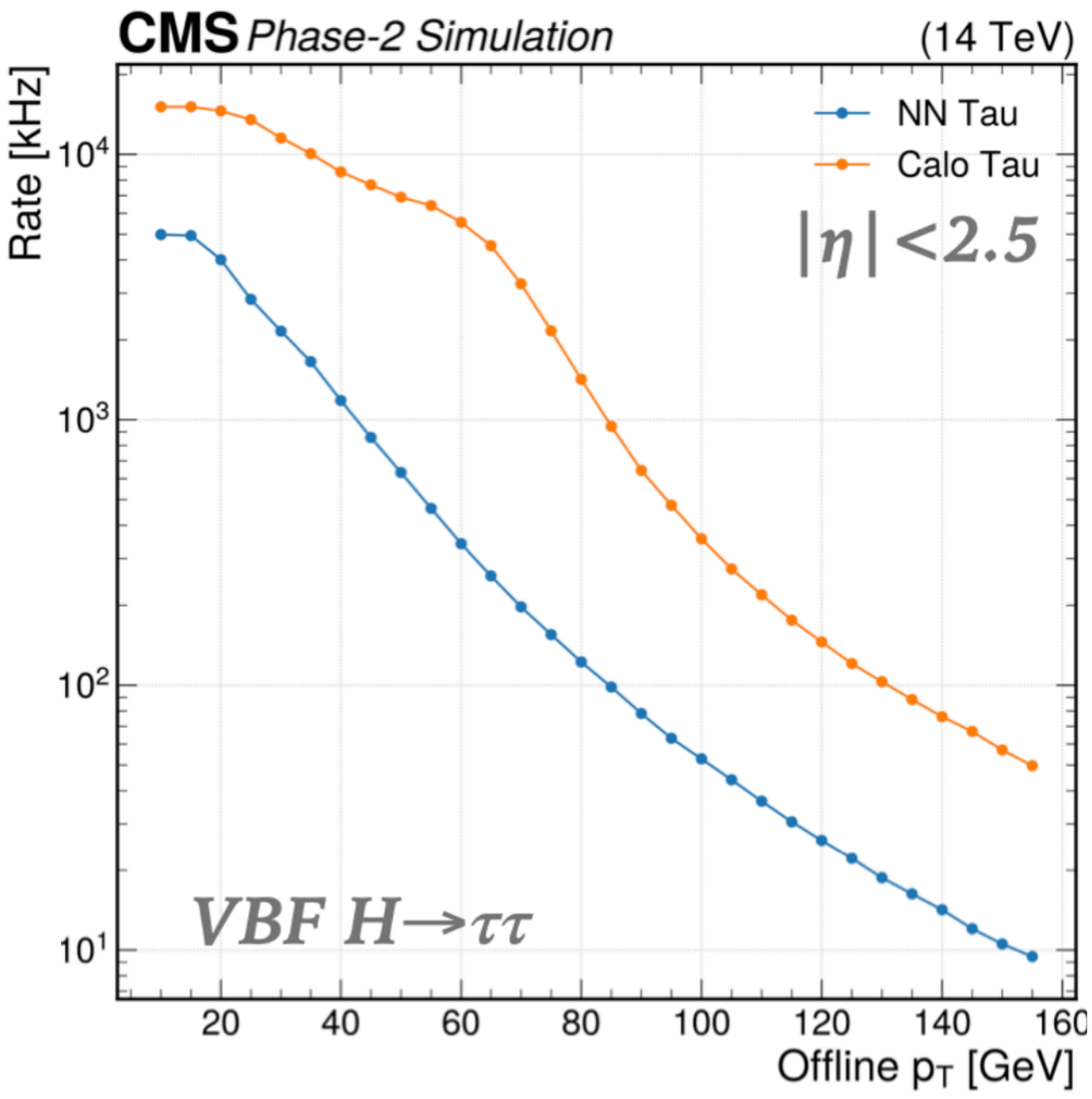
Events

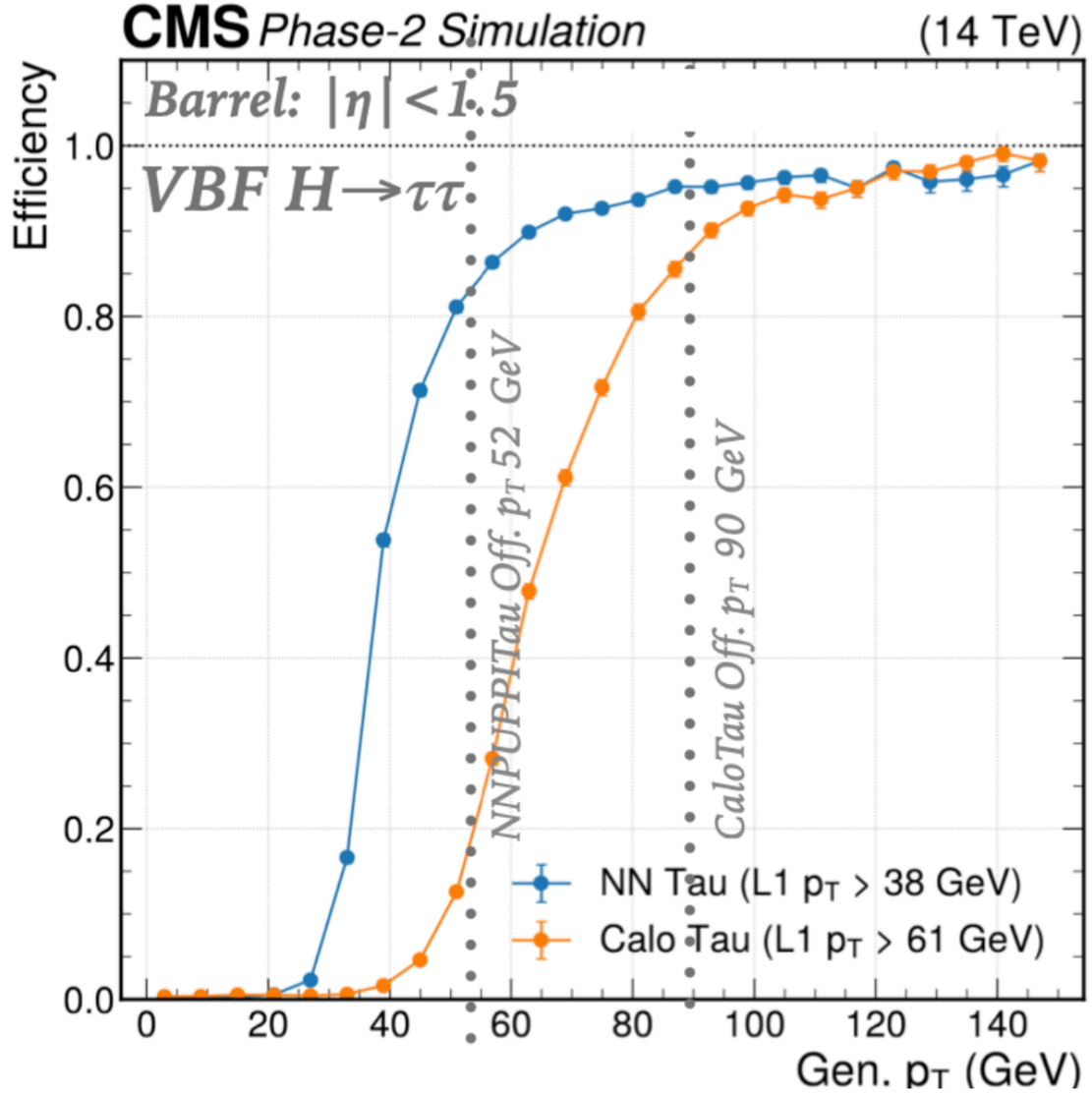
Correlator Layer 2



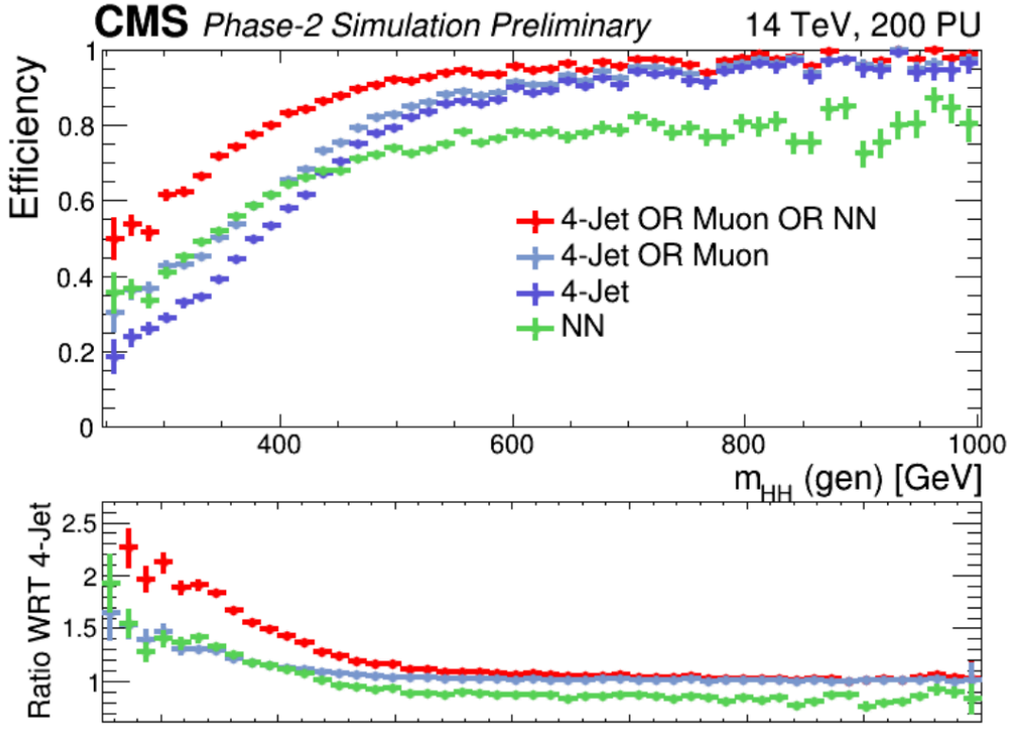
L1 Tau Tagging

- Developed/implemented a Deep Neural Network for Tau Tagging
  - Fully impelmented and operational in test systems





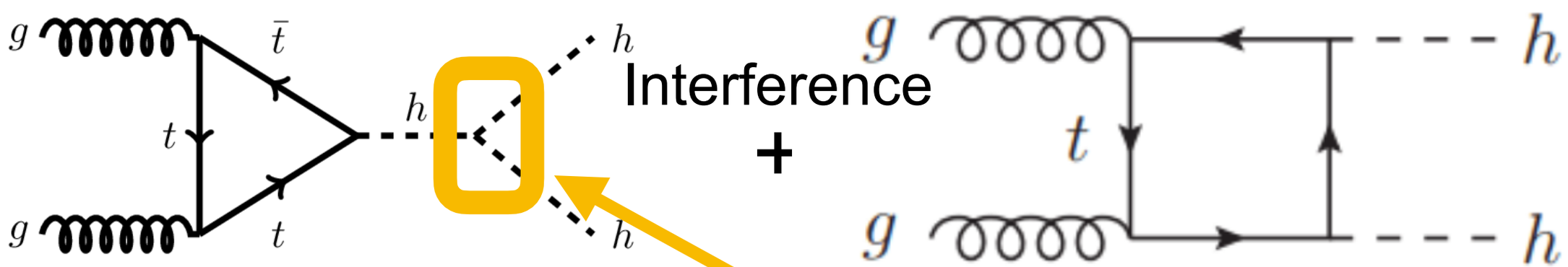
L1 B-Tagging



Decided to investigate to see if it helped

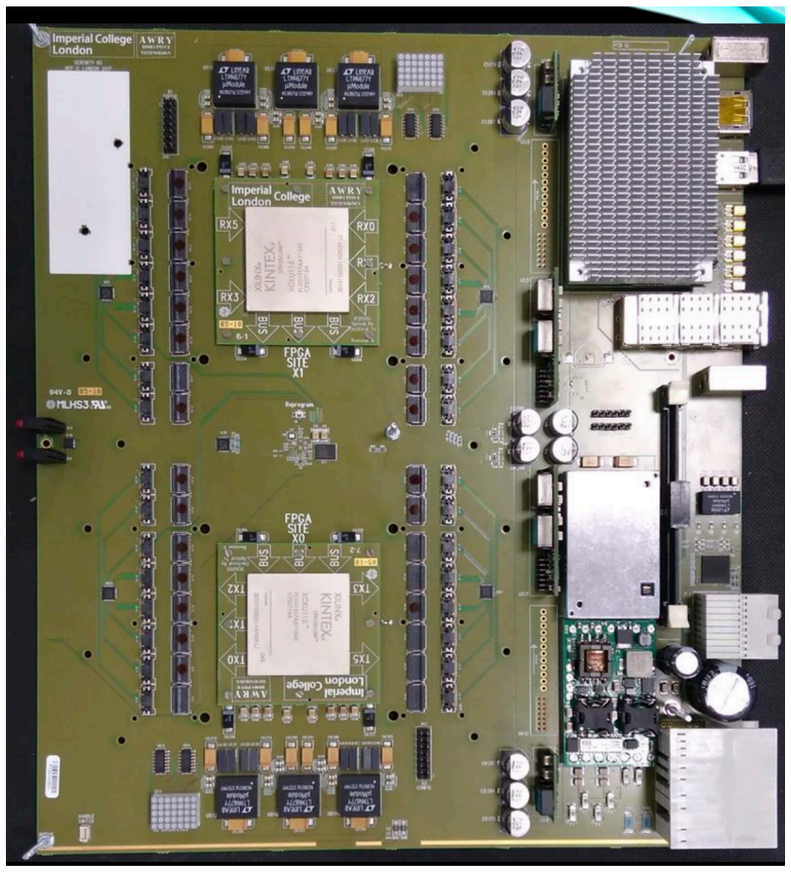
Turns out 30-40% improvement in Higgs self-coupling

Network runs in 80 nanosecond



=Strategy

Serenity Board (UK/EU/…)

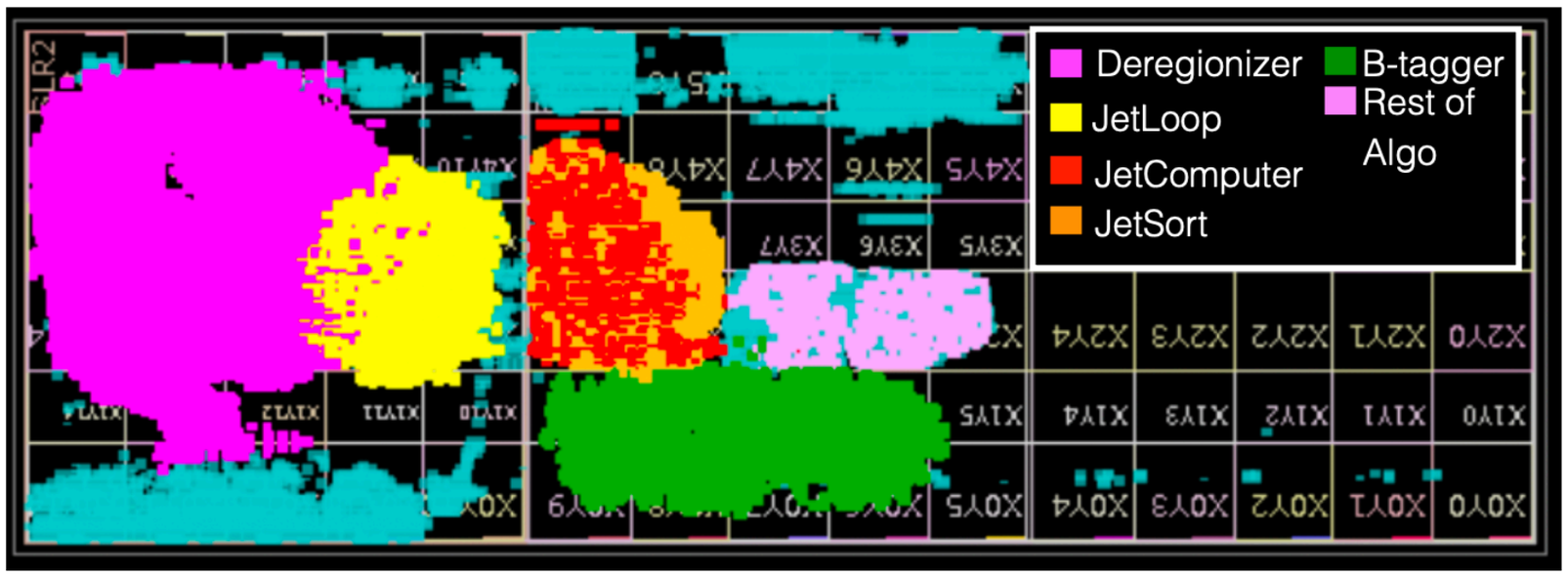


APx Board (Wisconsin/…)

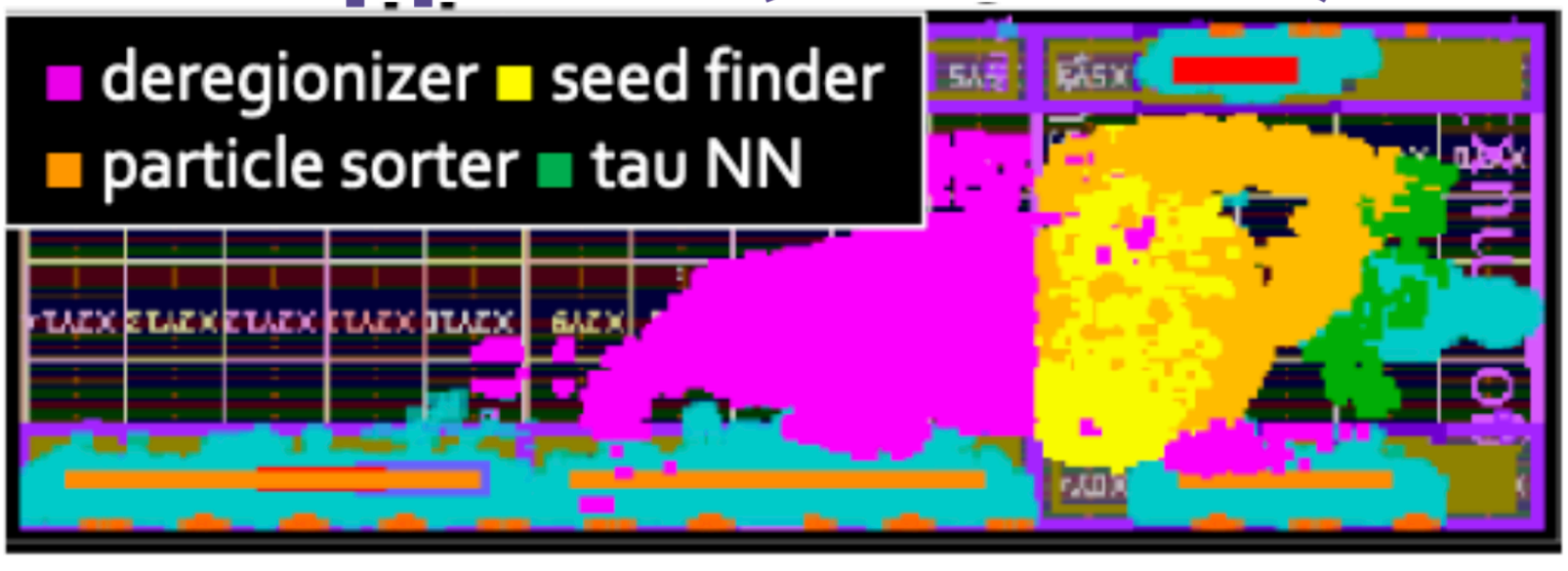
- Hardware is being developed by two groups
  - Serenity in the UK and Wisconsin in the US
  - Roughly 50% of the system is each board
- Our focus has been on centralizing/connecting the two efforts

Strategy

B-tagger *(Serenity VU9P-2)*

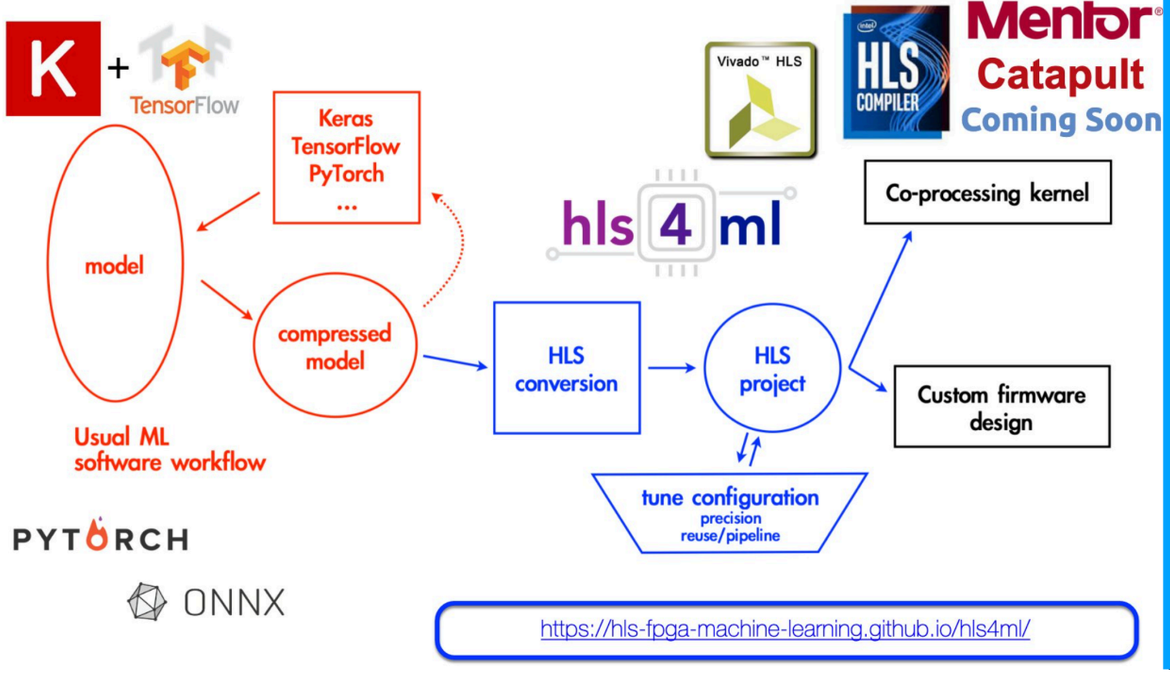


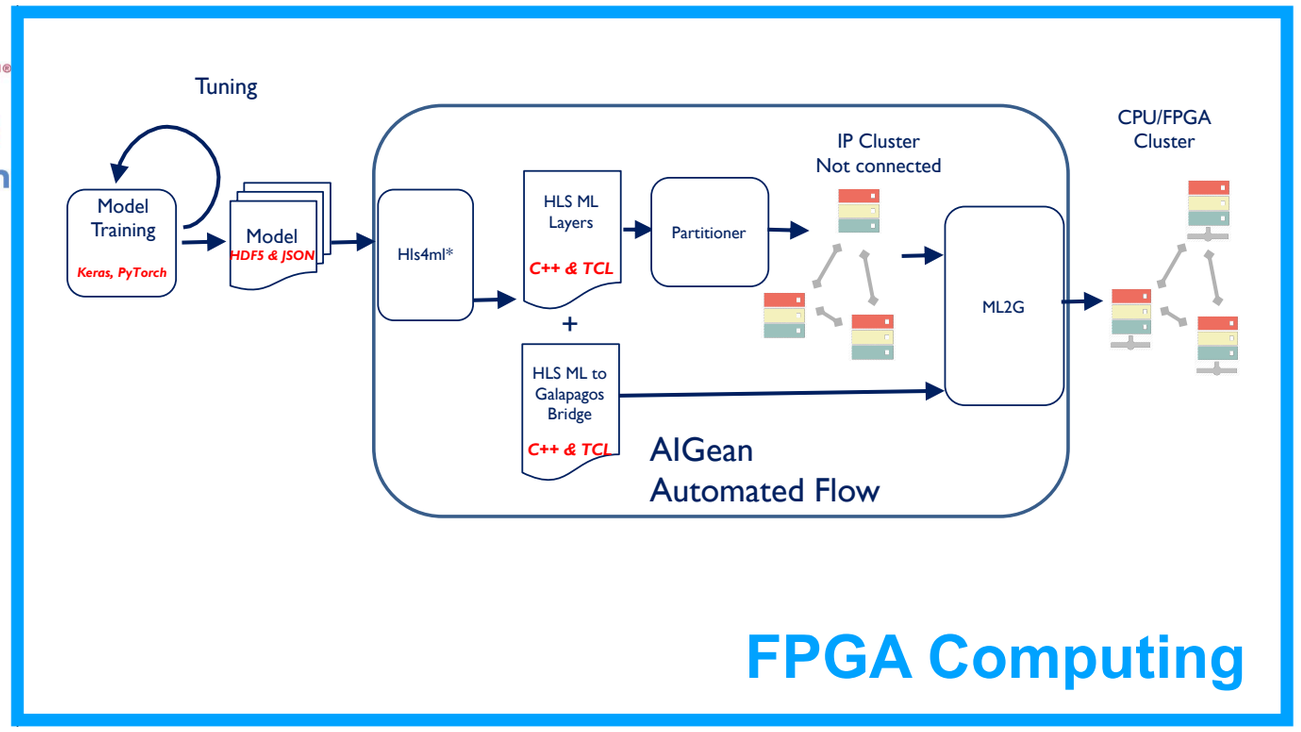
NN Puppi Tau *(APx VU9P-1)*



FPGA & Deep Learning

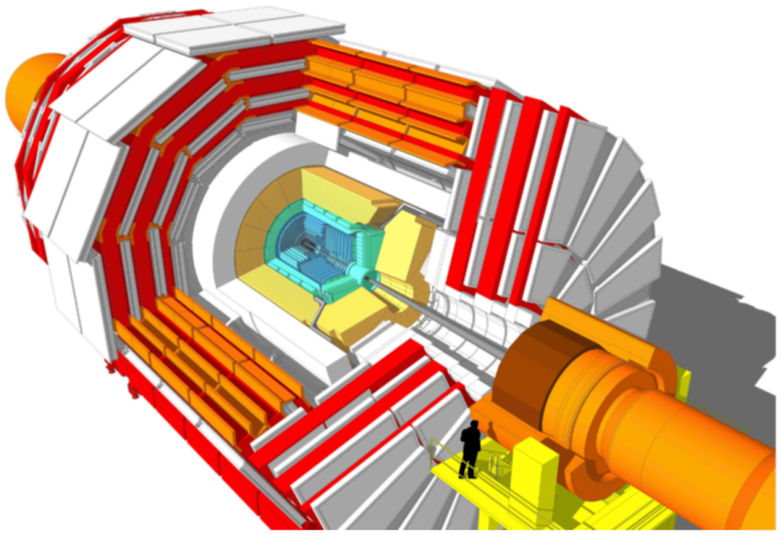
- Following our work on the L1 Trigger
  - Developed a toolkit to deploy optimized ML on FPGAs HLS4ML
  - This work is quickly becoming a major tool in industry
  - Built some fastest neural network inferences in the world





Future of our system

- We are investigating strategies for scouting system



Detector Readout

L1 Trigger System

New System (40 MHz Scouting)

750 kHz (Phase Out)

New Data Flow

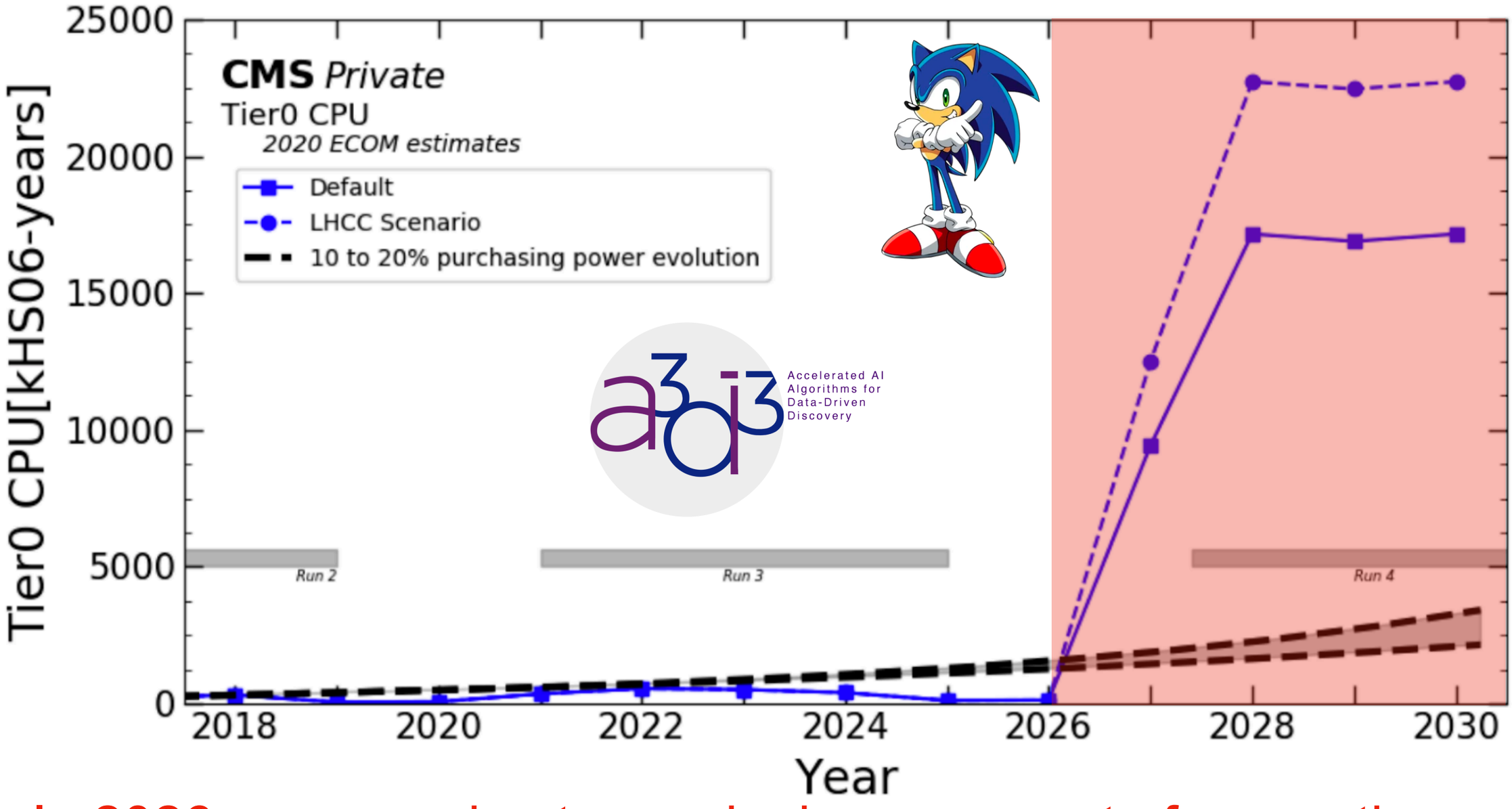
HLT System

5 kHz Data Output

Save To Disk Analyze Over&Over

Extensions of HLS4ML are critical as well as GPU/ML integration

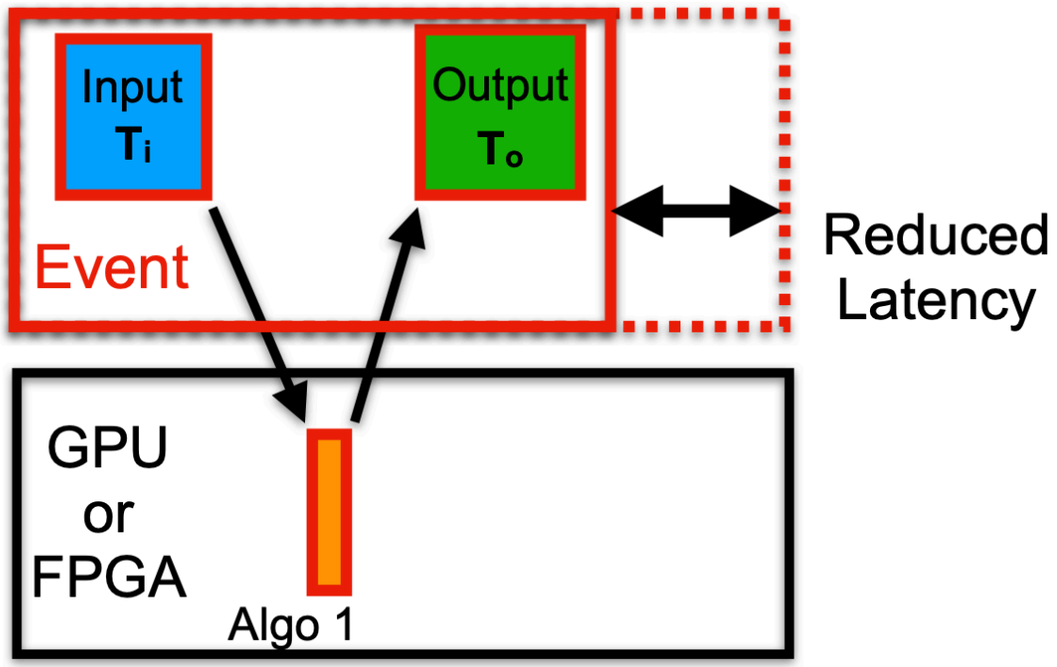
Computing Challenge

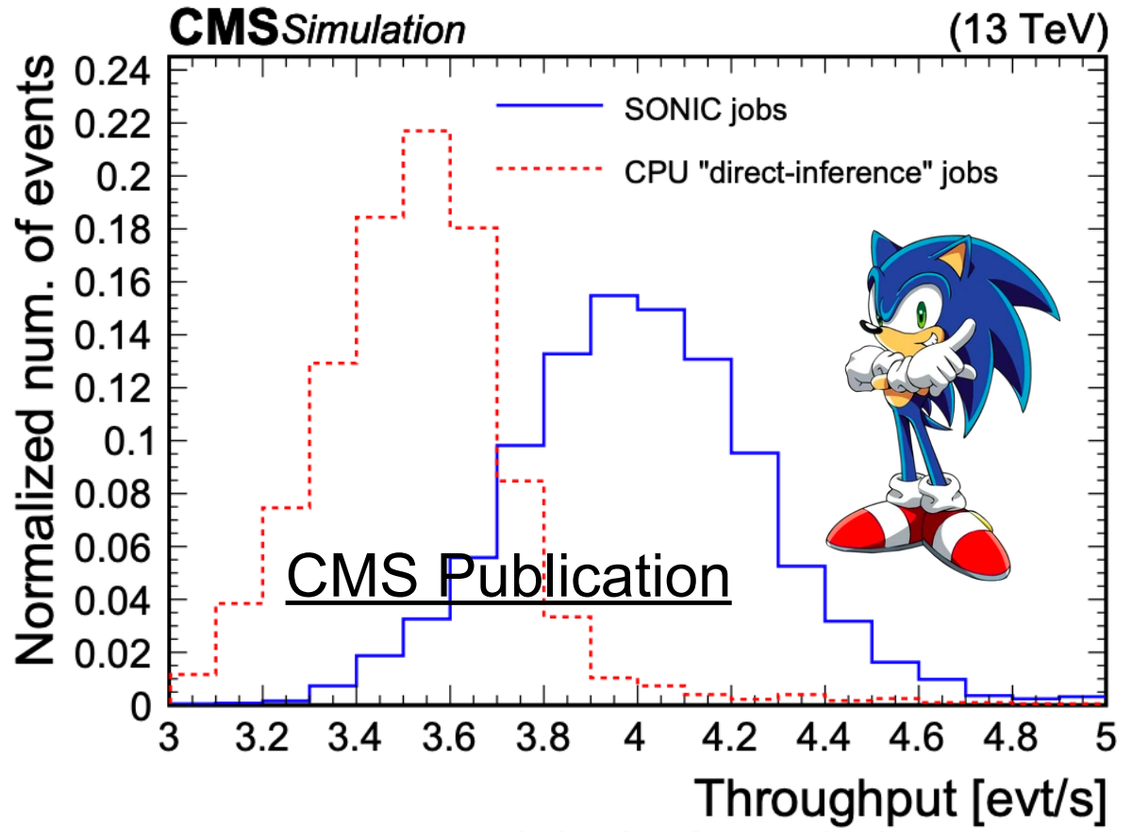


In 2026 we are going to need a huge amount of computing

Developing transformative ideas to solve this problem

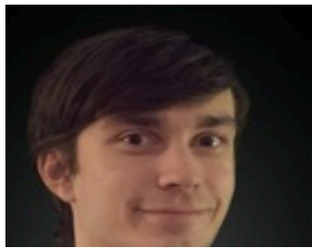
Overall Speed Ups



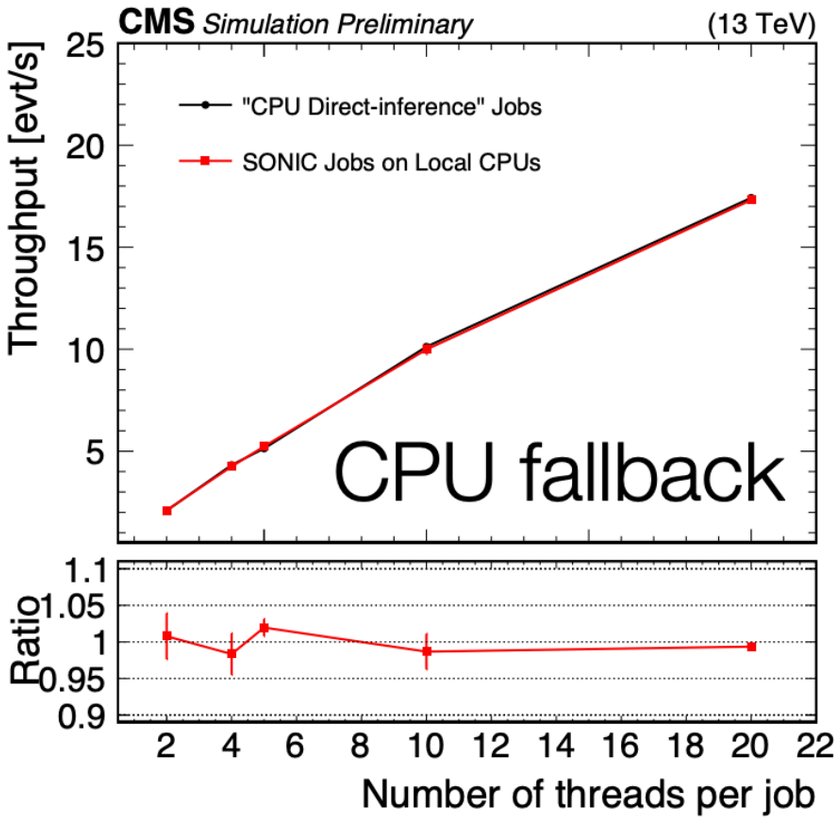


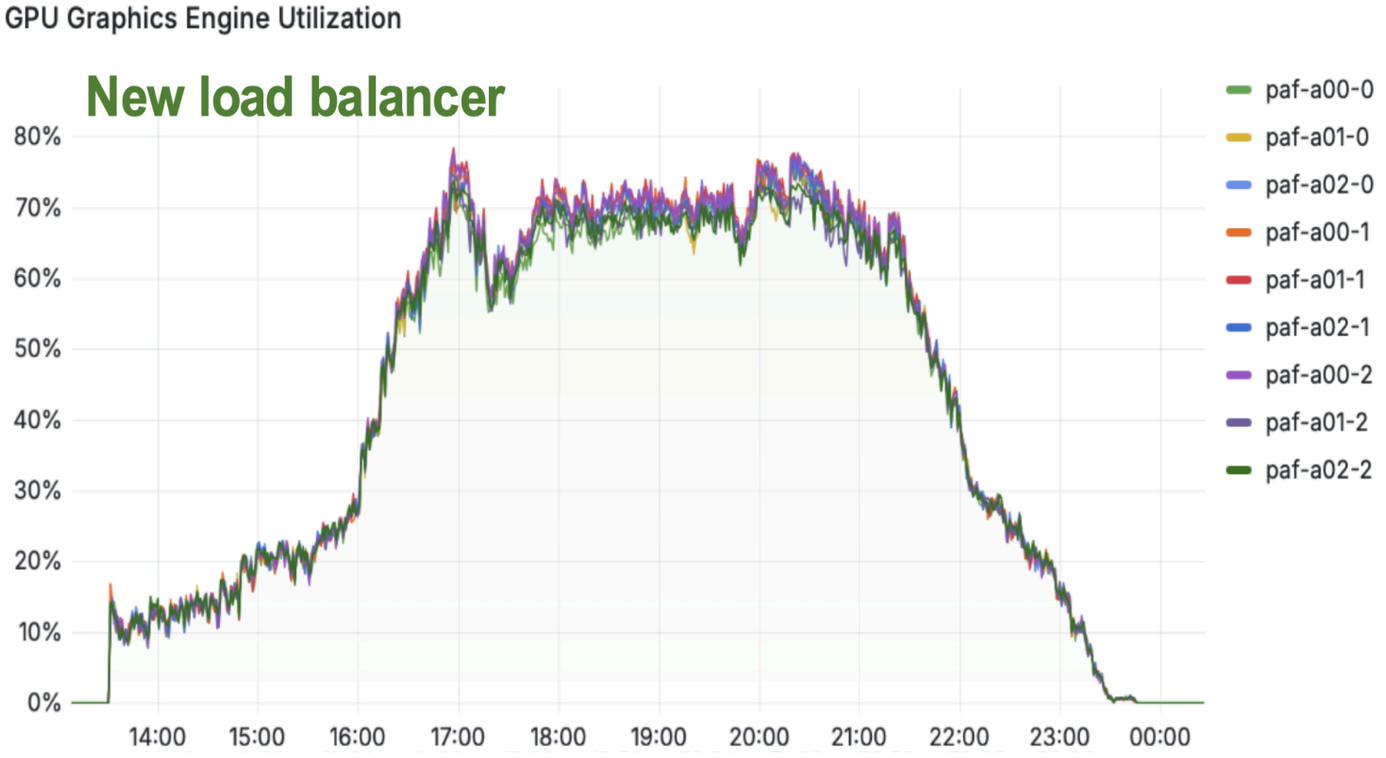
- **SONIC**: Integrated GPU/FPGAs in CMS SW and computing
  - Already demonstrated to scale the use case
- Next step is to start running official computing workflows
  - Aiming to have GPU based running at scale by end of year

a3i3 Accelerated AI Algorithms for Data-Driven Discovery



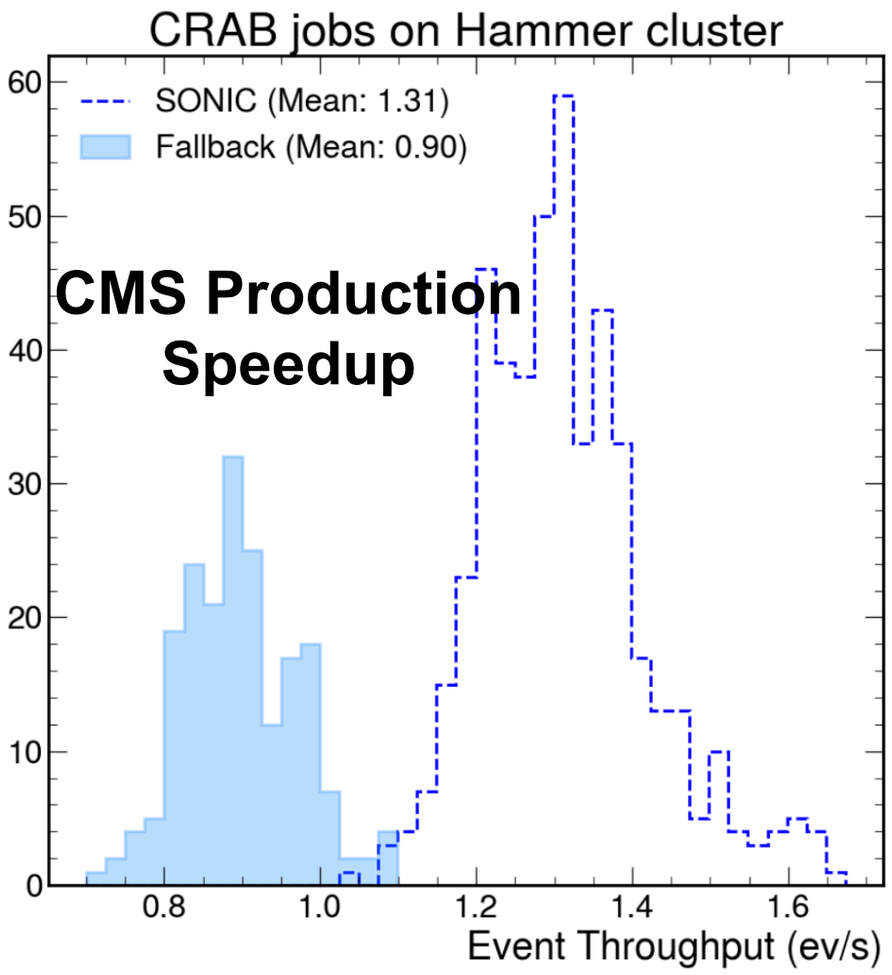
Integrating Into CMS

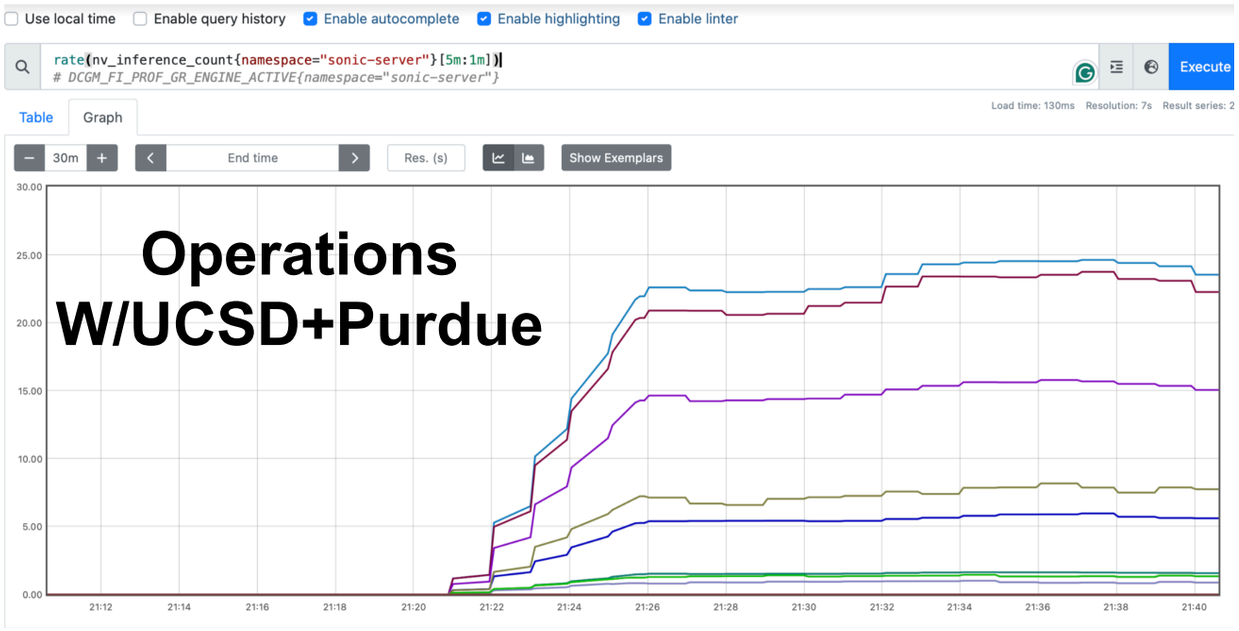




- Current focus production infrastructure for CMS
  - Optimized load balancer that ensures balanced GPU usage
  - Computational model built on Amdahls' law
- Active integration with CMS computing
  - First full-scale GPU production w/SONIC targetting this fall

Future of SONIC

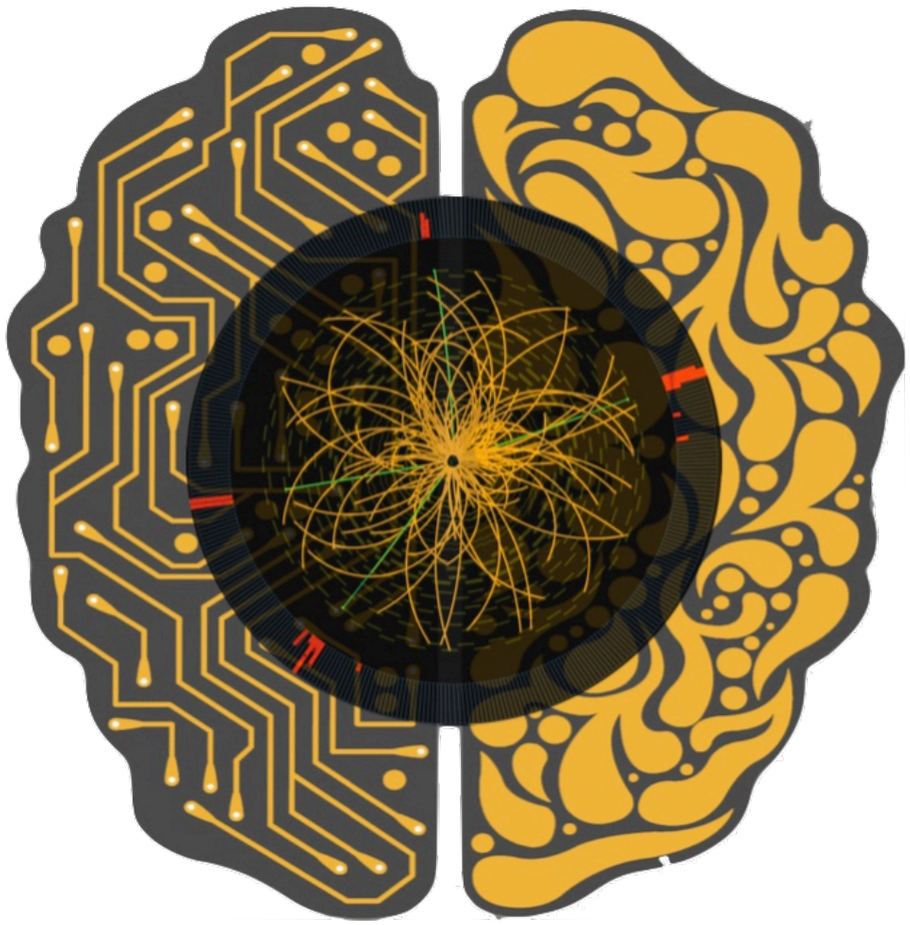




Parallel investigation with NERSC

- Aiming for integration into CMS production for HL-LHC
  - Most of the software and toolkit already there
- Synergy with DUNE/IceCube/LIGO/NERSC

FastMachineLearning.org



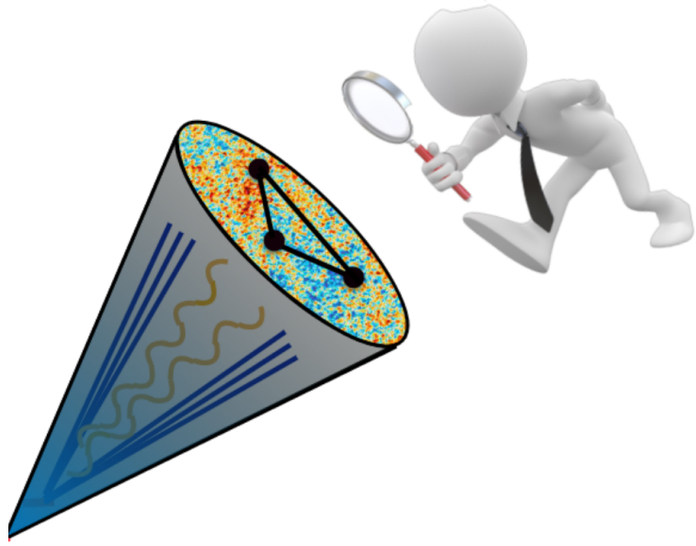
FML

Group Founded by P. Harris and N. Tran (FNAL)https://indico.cern.ch/event/822126/

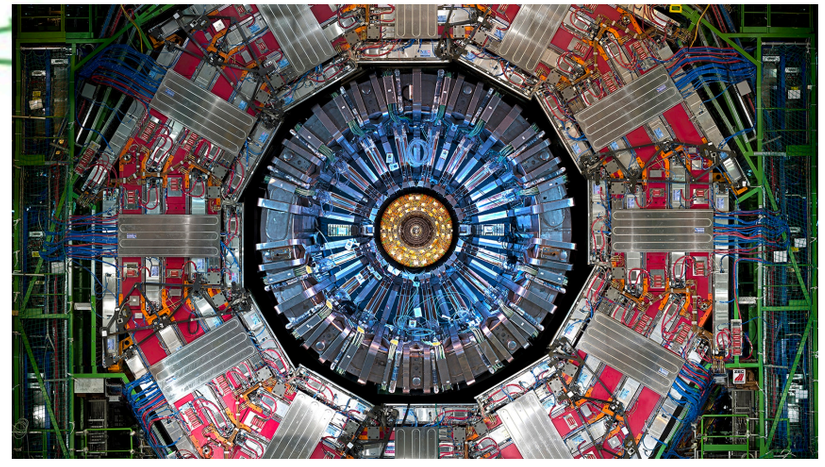
- Project now covers some LHC, DUNE, LIGO, Materials science….
- Slack(collaboration) is now > 1000 members across globe
- All 3 DOE-HEP Early Career Award this year uses FastML

Cycle of LHC Research

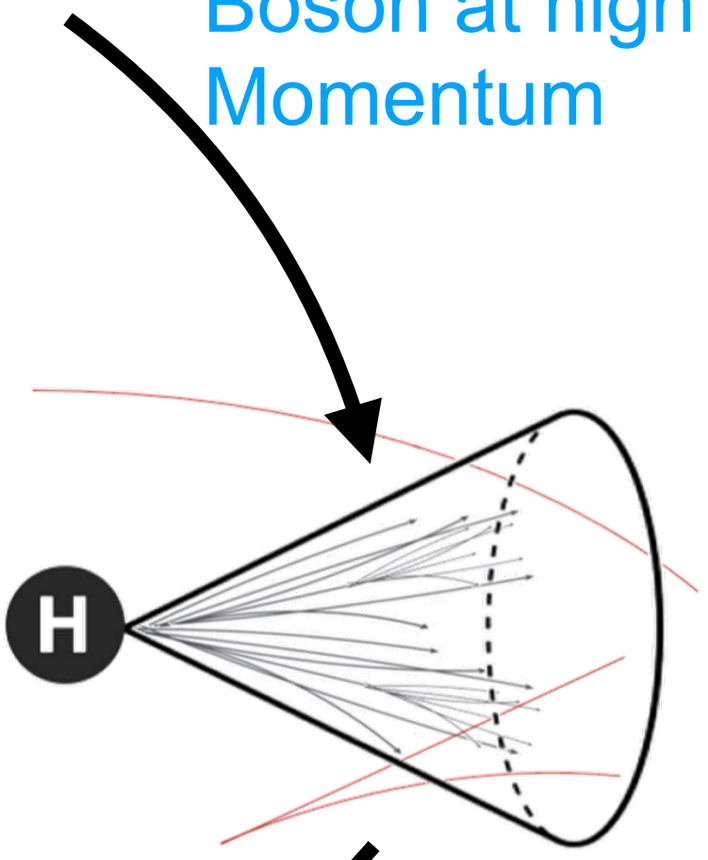
We focus on jets of all kinds



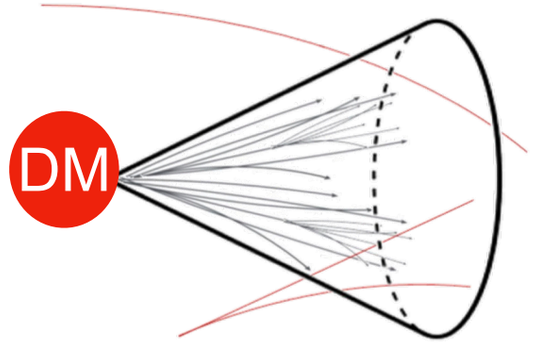
Core QCD Measurements



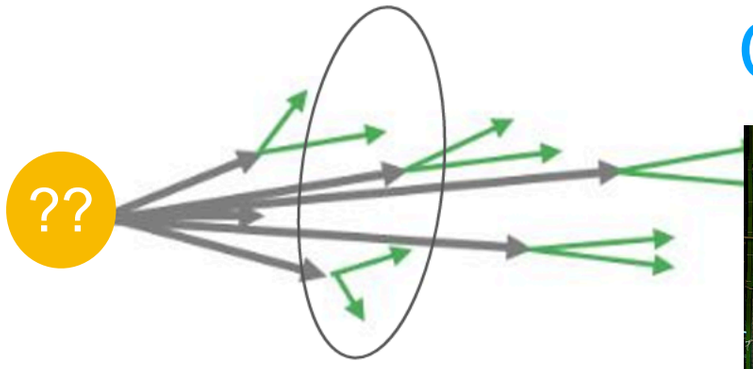
Probing Higgs Boson at high Momentum

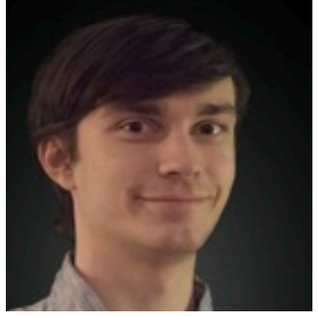


Searching For (light) DM



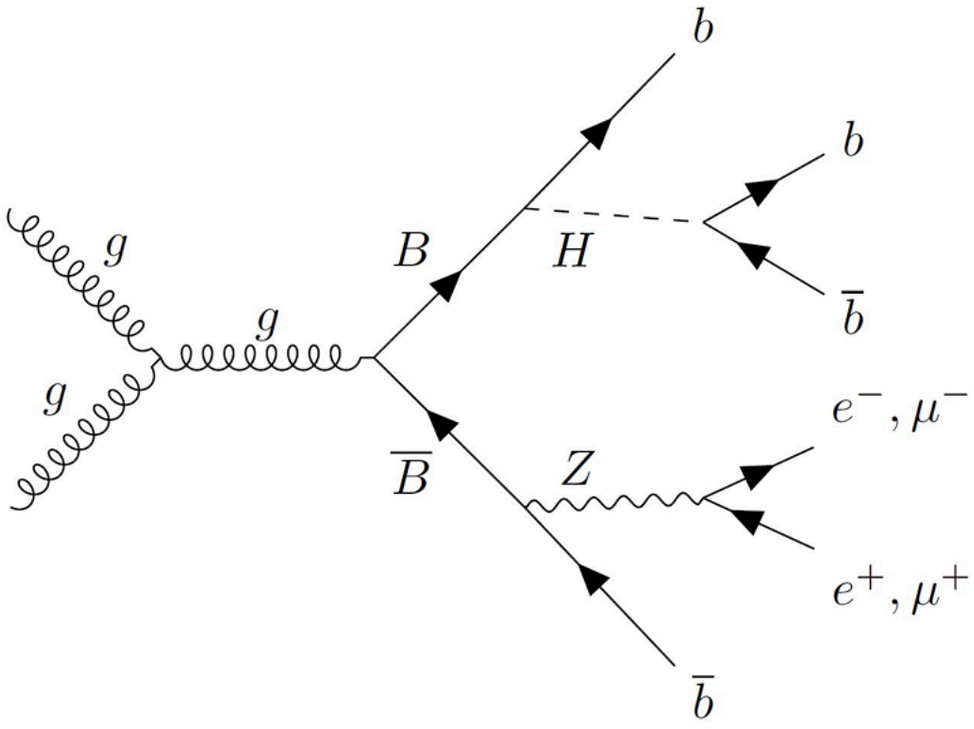
Searching For any anomaly Using New(AI) tech

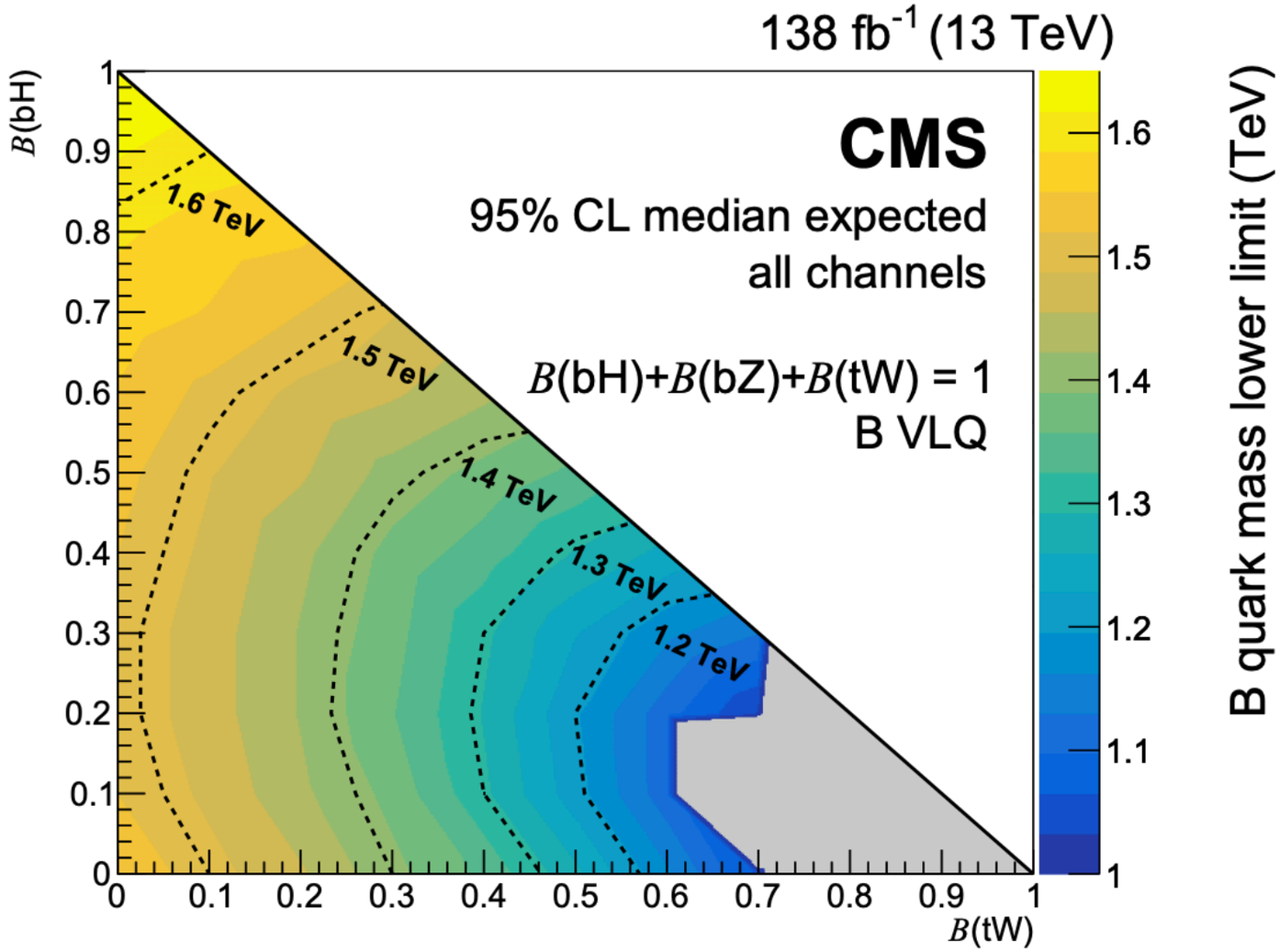




VLQ Analysis

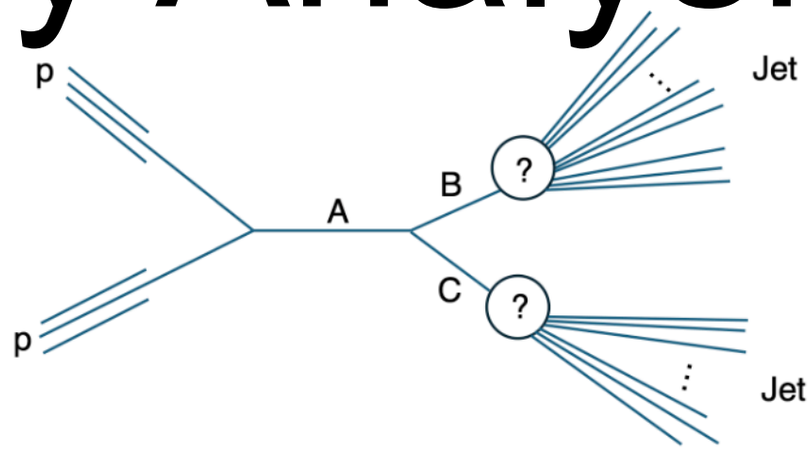
- World leading results on B-quark type Vector Like Quark
  - Complex All hadronic final state, many jets



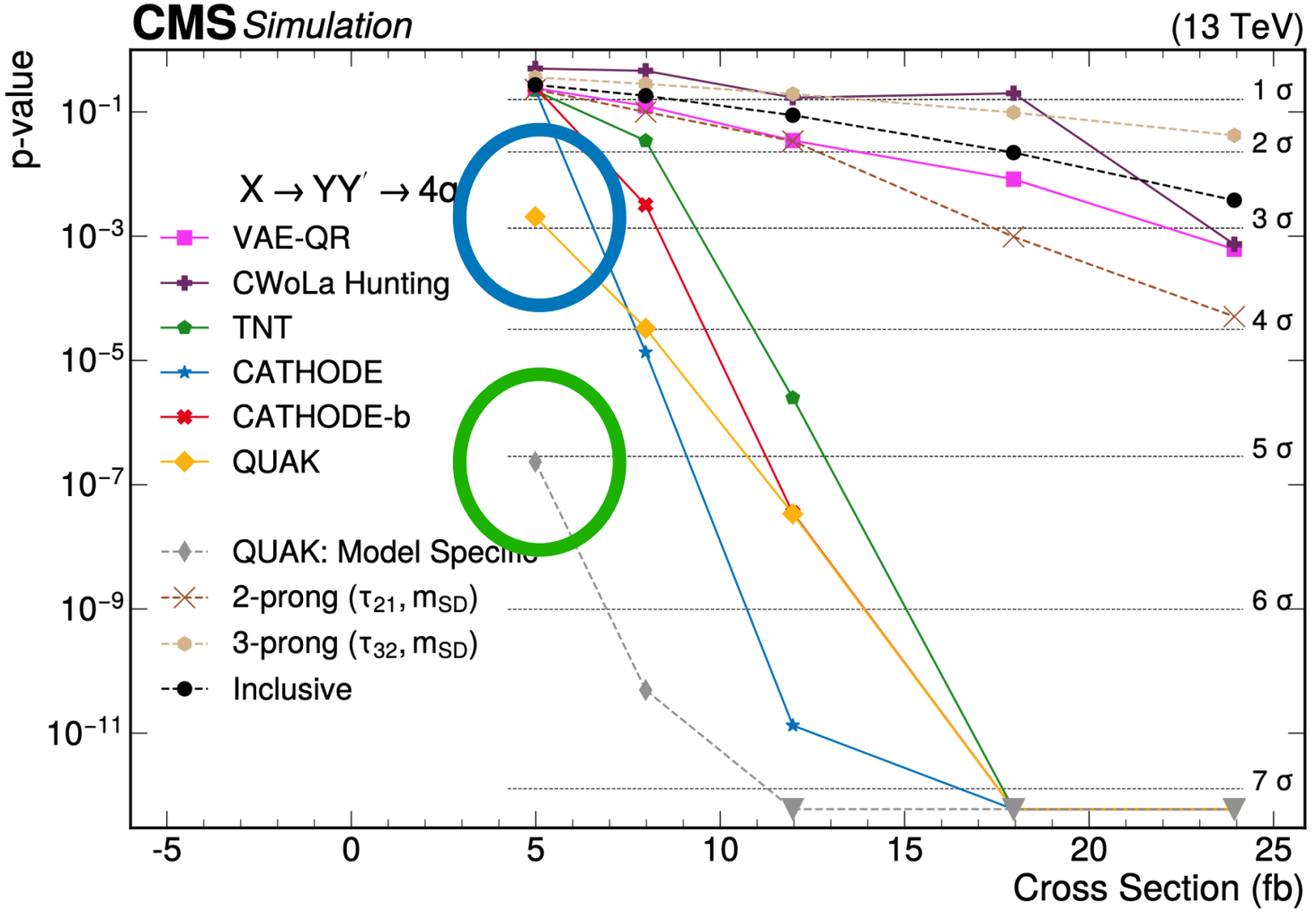


https://arxiv.org/abs/2402.13808

AI-Anomaly Analysis

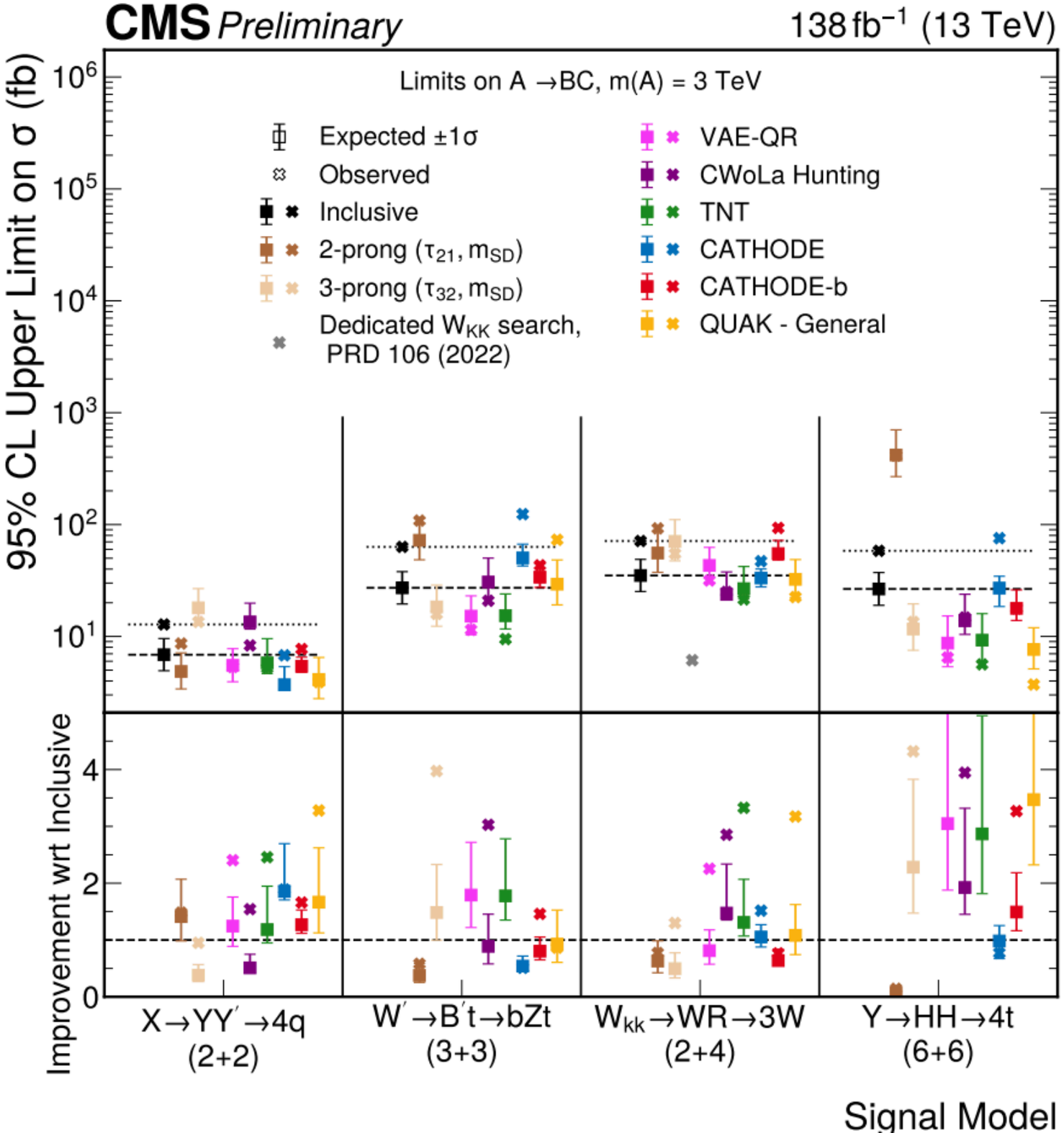


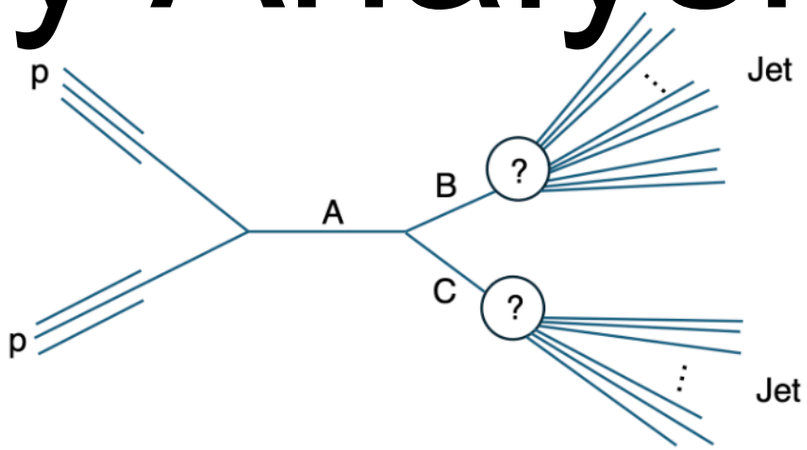
- Recently finished AI-anomaly analysis
  - Compared 6 different AI-anomaly strategies



https://cds.cern.ch/record/2

AI-Anomaly Analysis

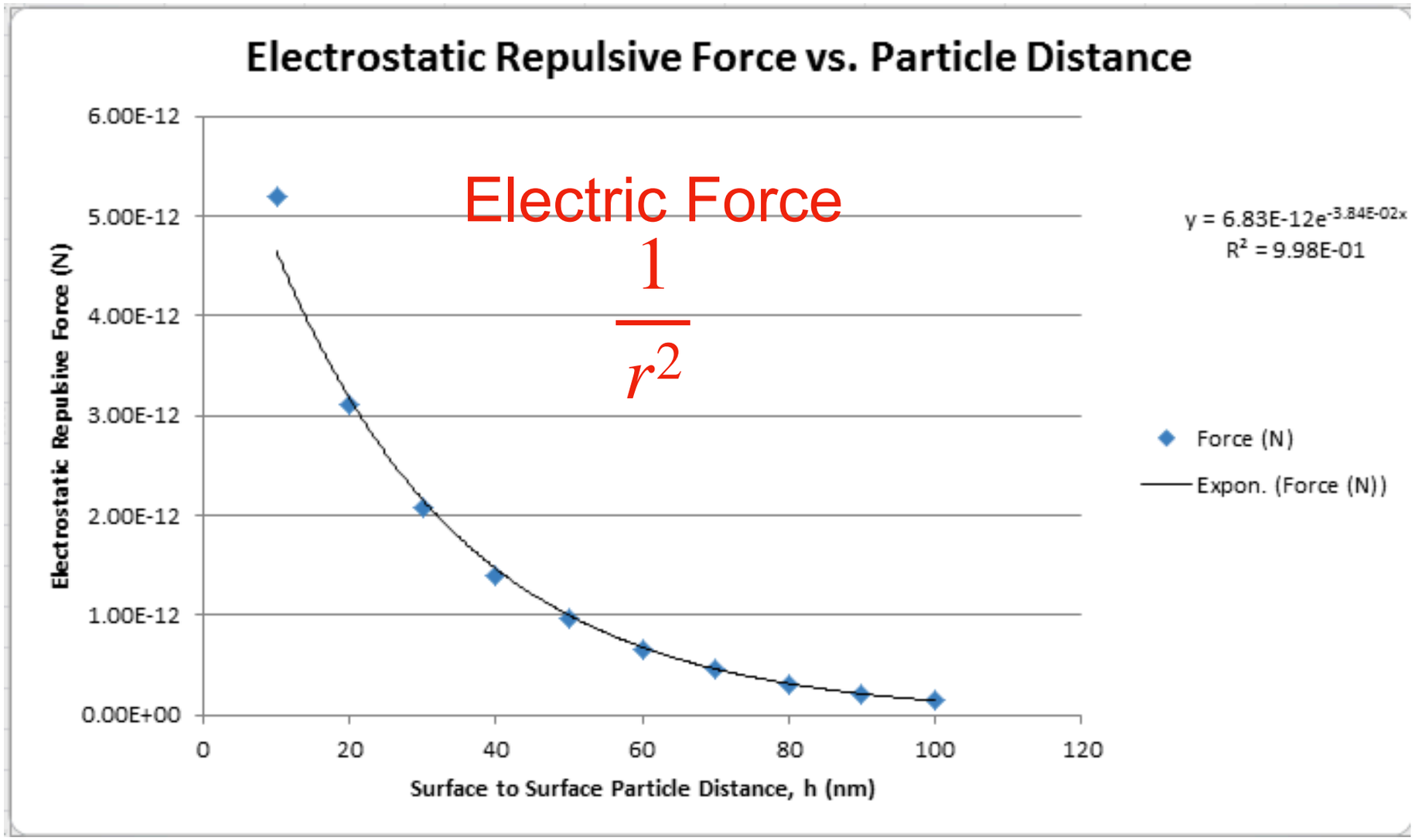




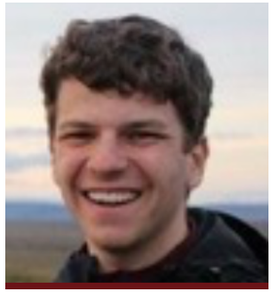
- Pathway towards automated physics searches
  - Here we put bounds on O(30) different models all at once
  - Future analyses will build on this idea

https://cds.cern.ch/record/2892677

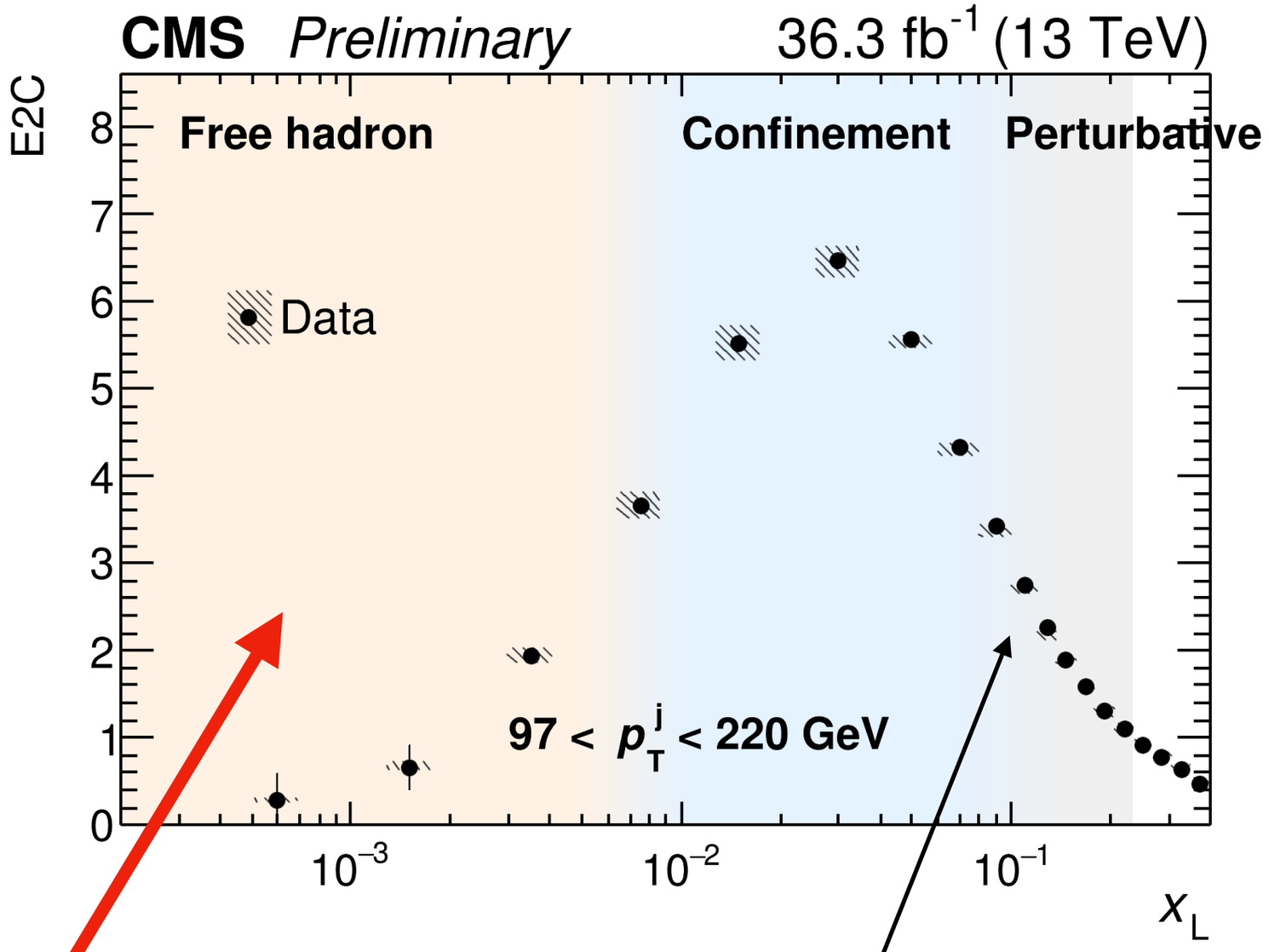
Studying Strong Force



Electric Force builds a $\frac{1}{r^2}$ distribution with radius

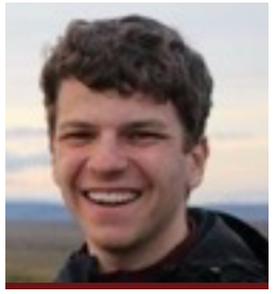


Studying Strong Force

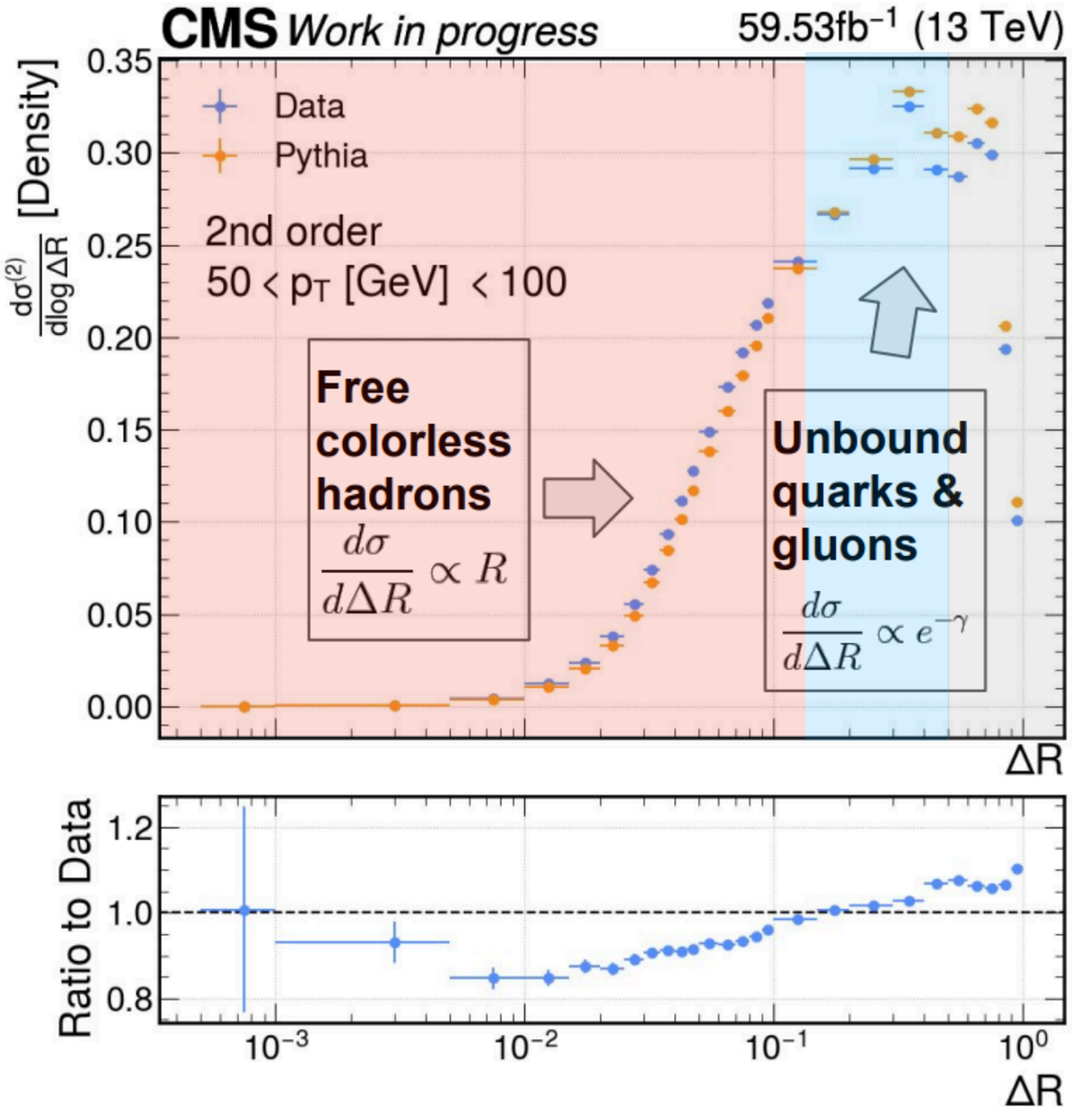


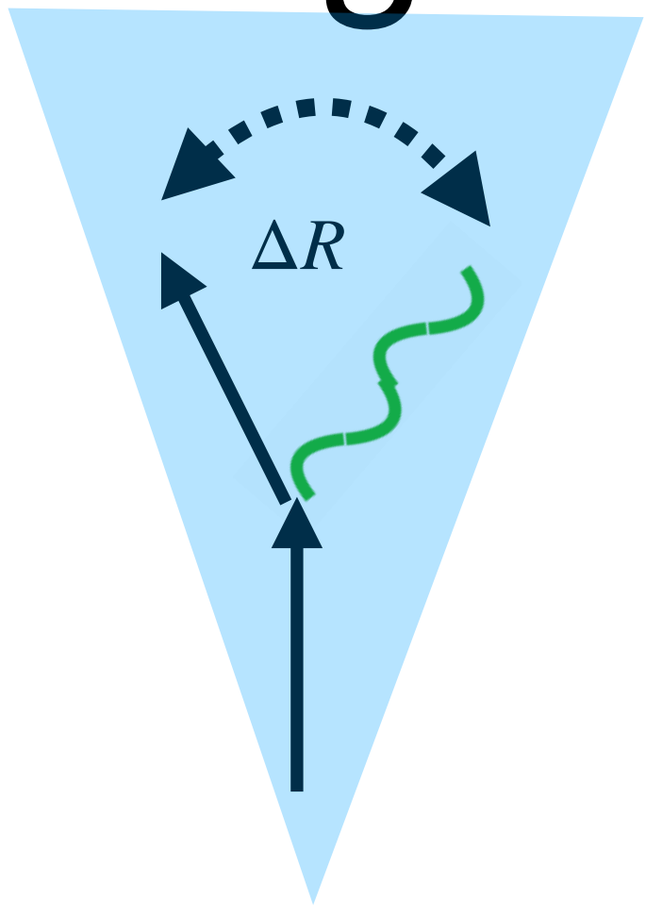
QCD is constant at small distance

QCD has $\frac{1}{r^2}$ distribution at large distance

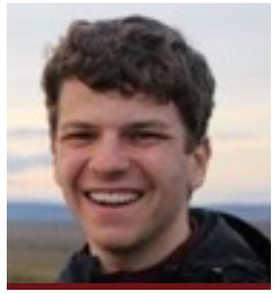


Understanding QCD

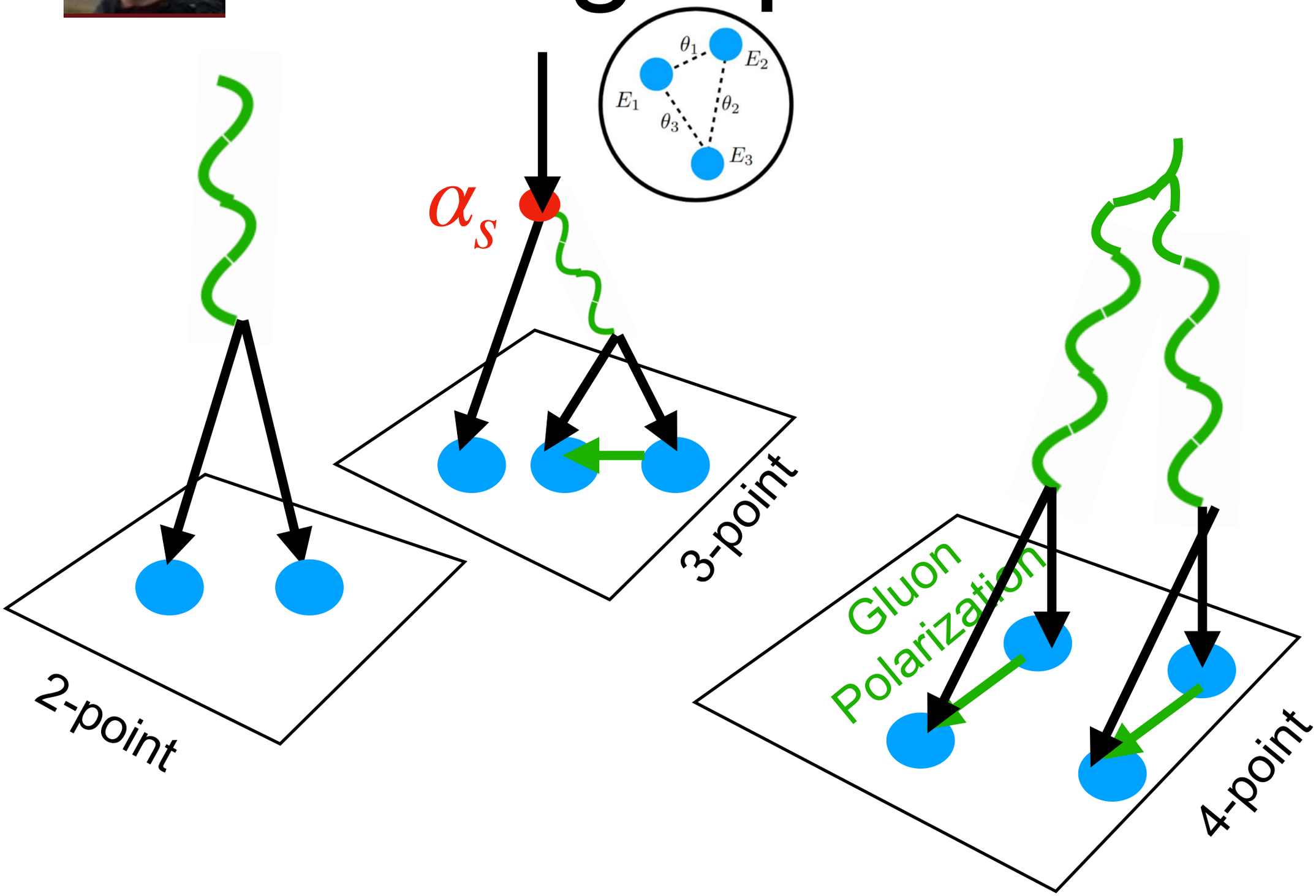




Pairwise Force of two particles vs distance

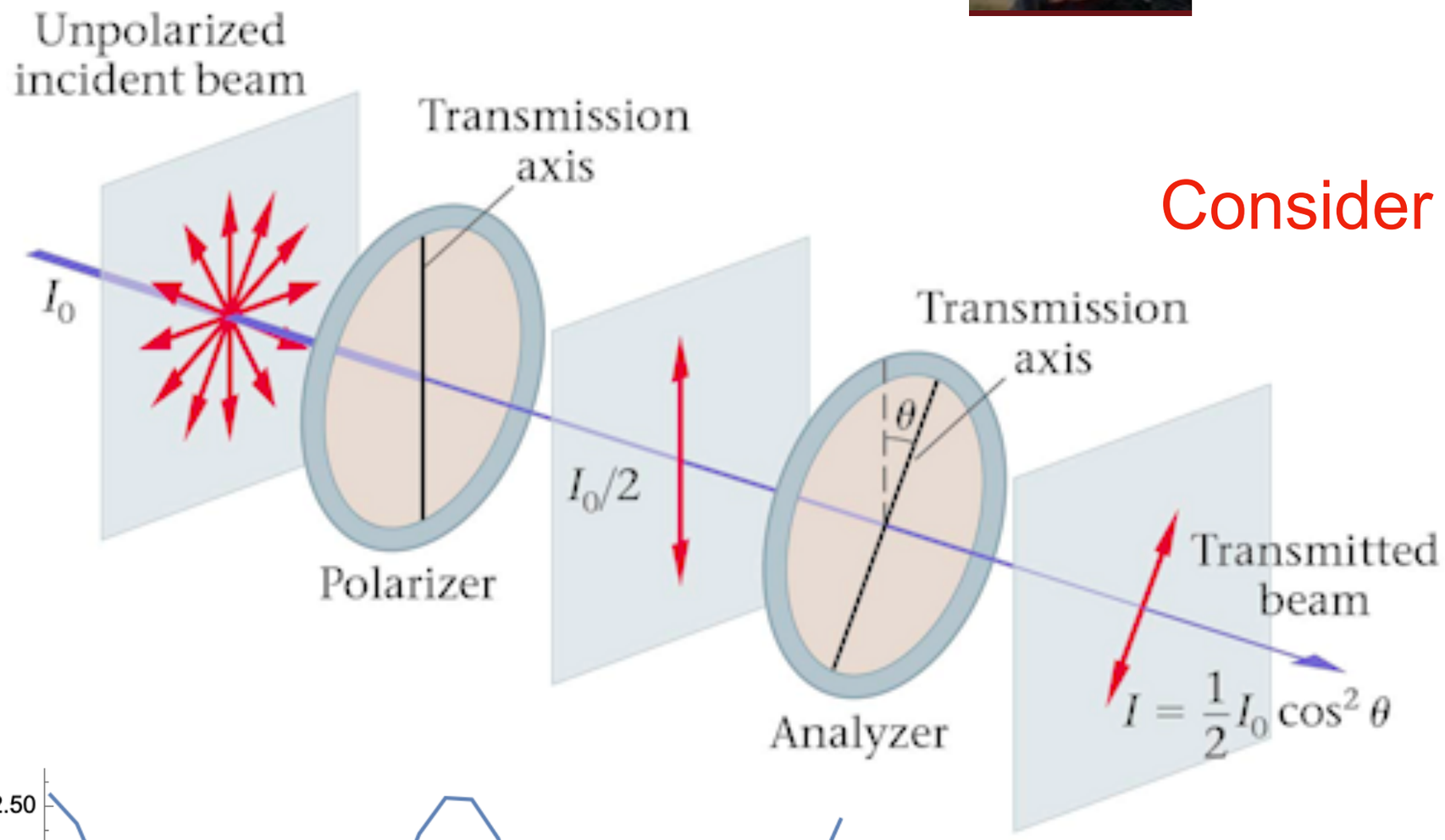


Adding Spin Information



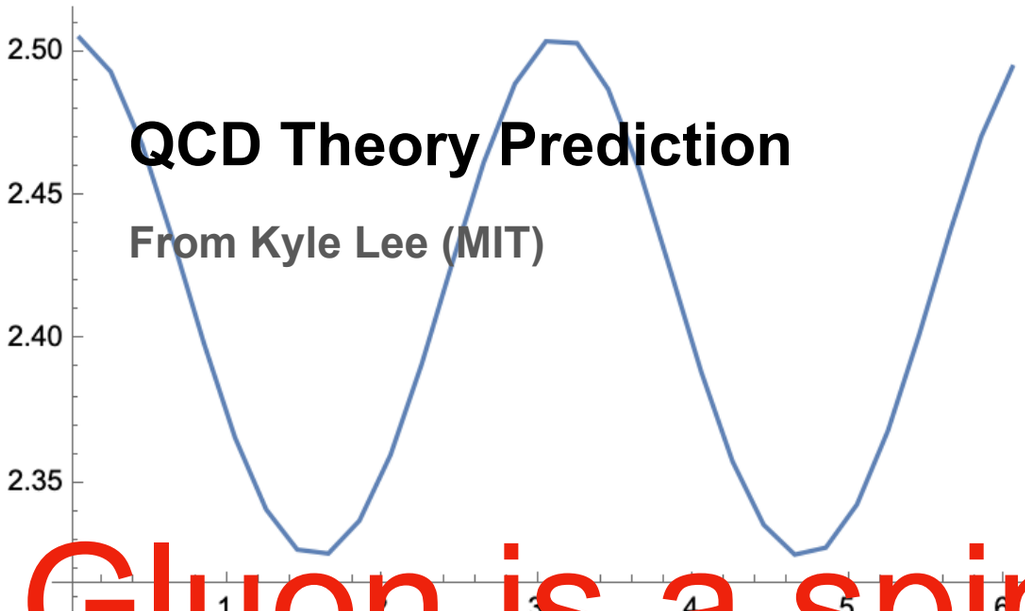
With the 4-point we start to probe spin effects

Visualizing



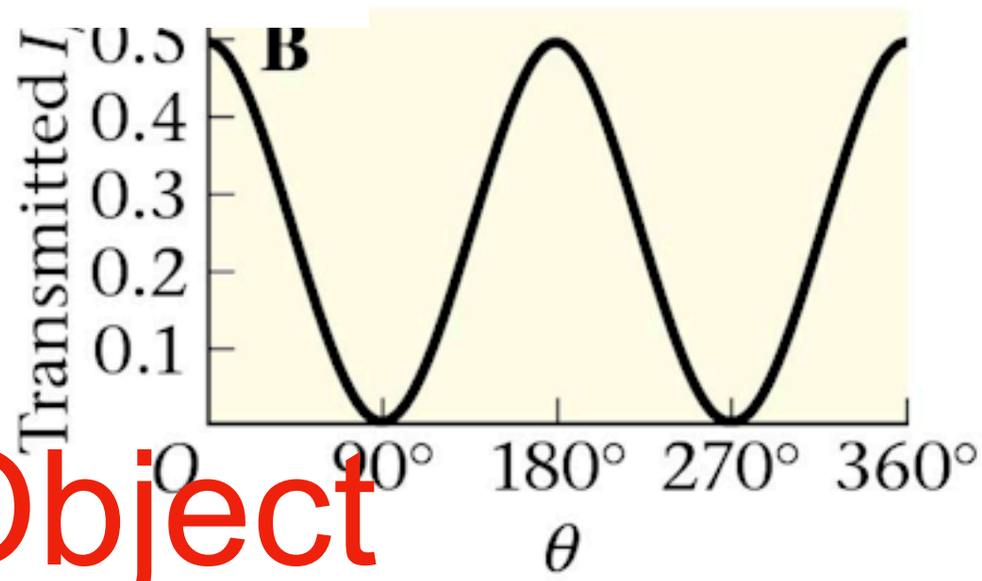
Consider a double polarizer

Yields Cos(2θ)



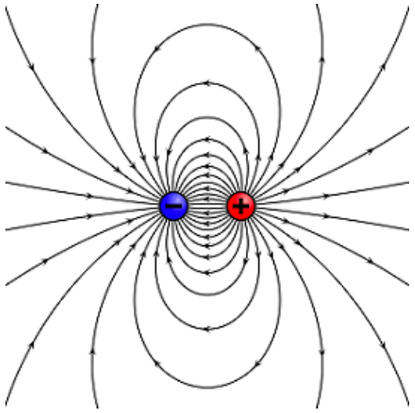
QCD Theory Prediction

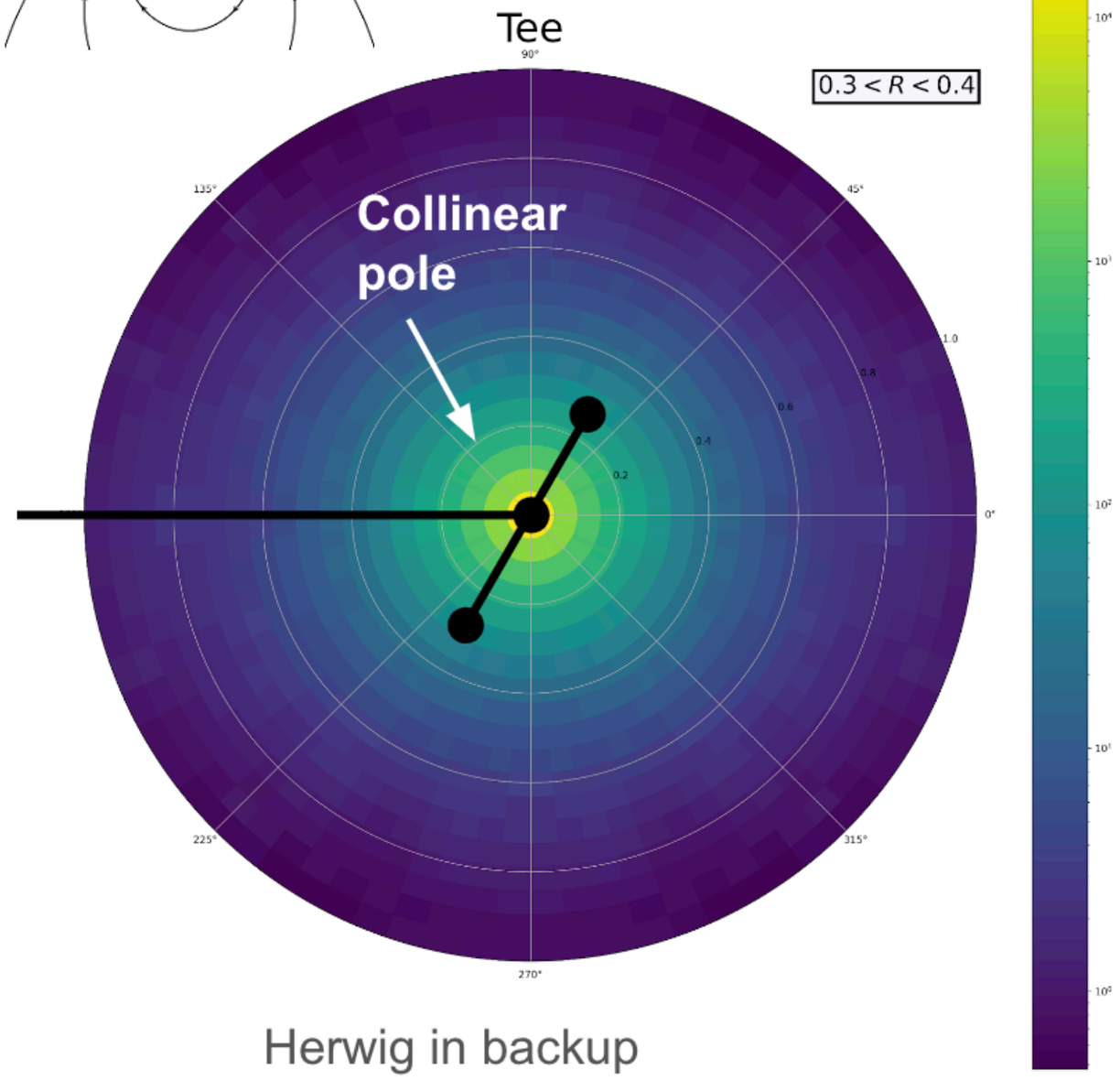
From Kyle Lee (MIT)



Gluon is a spin-1 Object

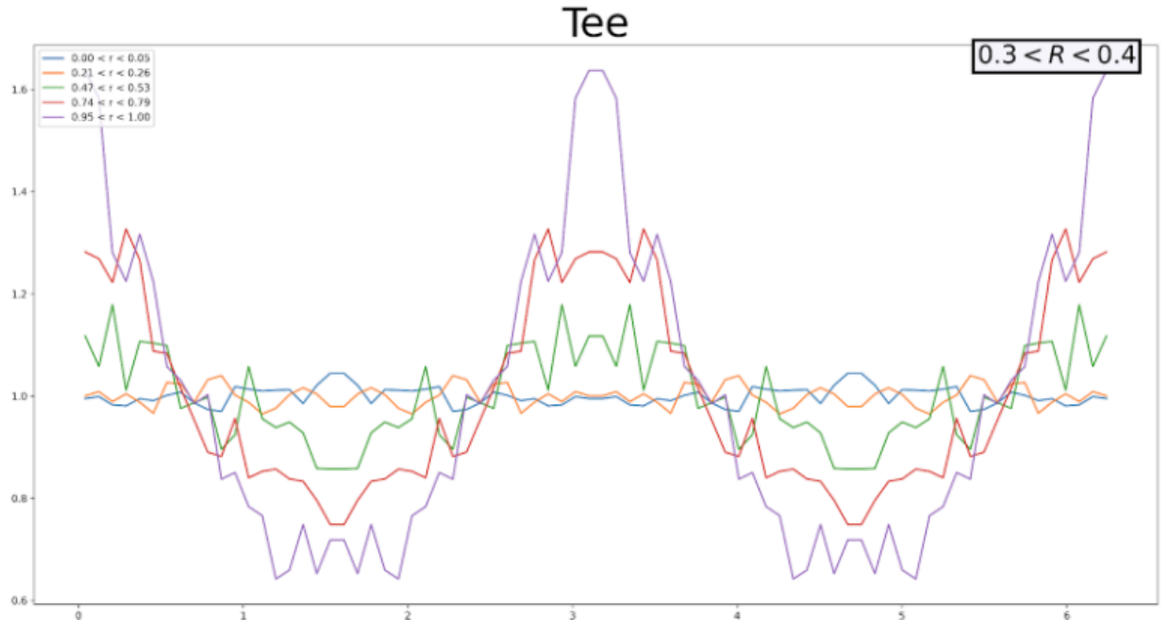
Emergence of Spin Interference

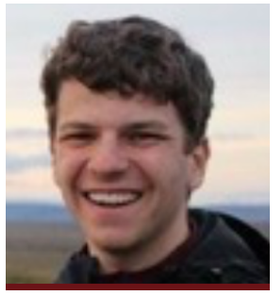




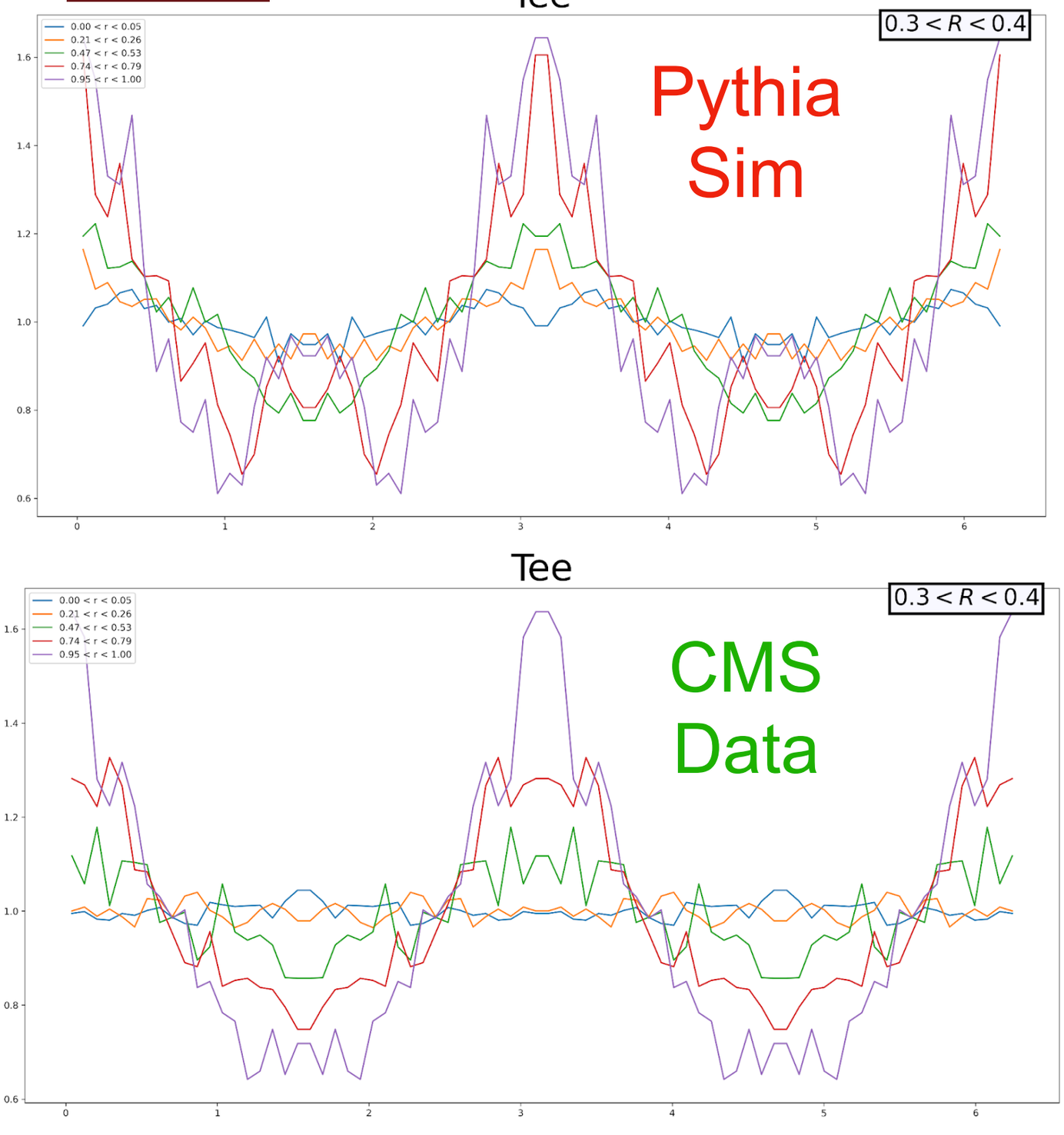
Herwig in backup

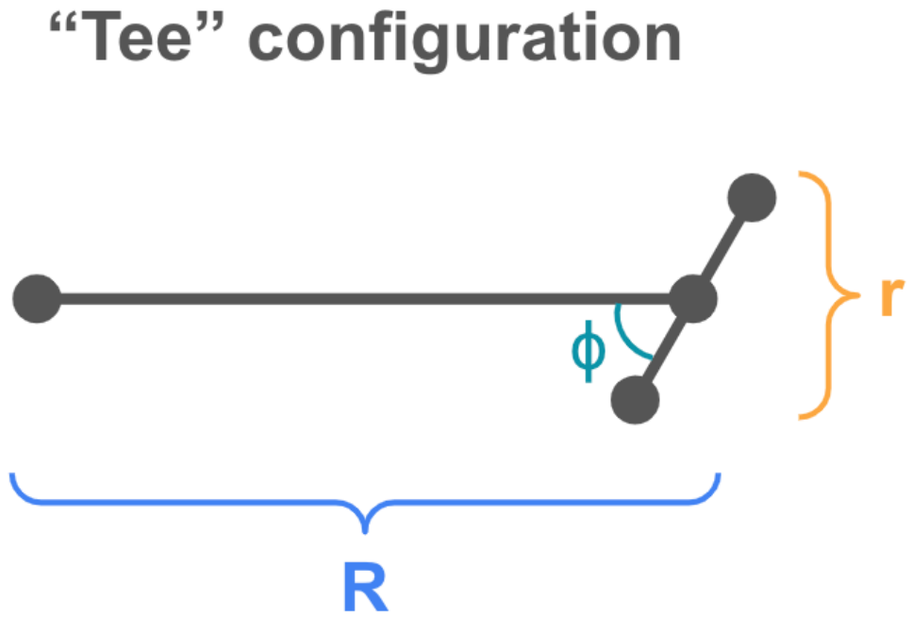
“Tee” configuration





Data vs MC(Not Public)



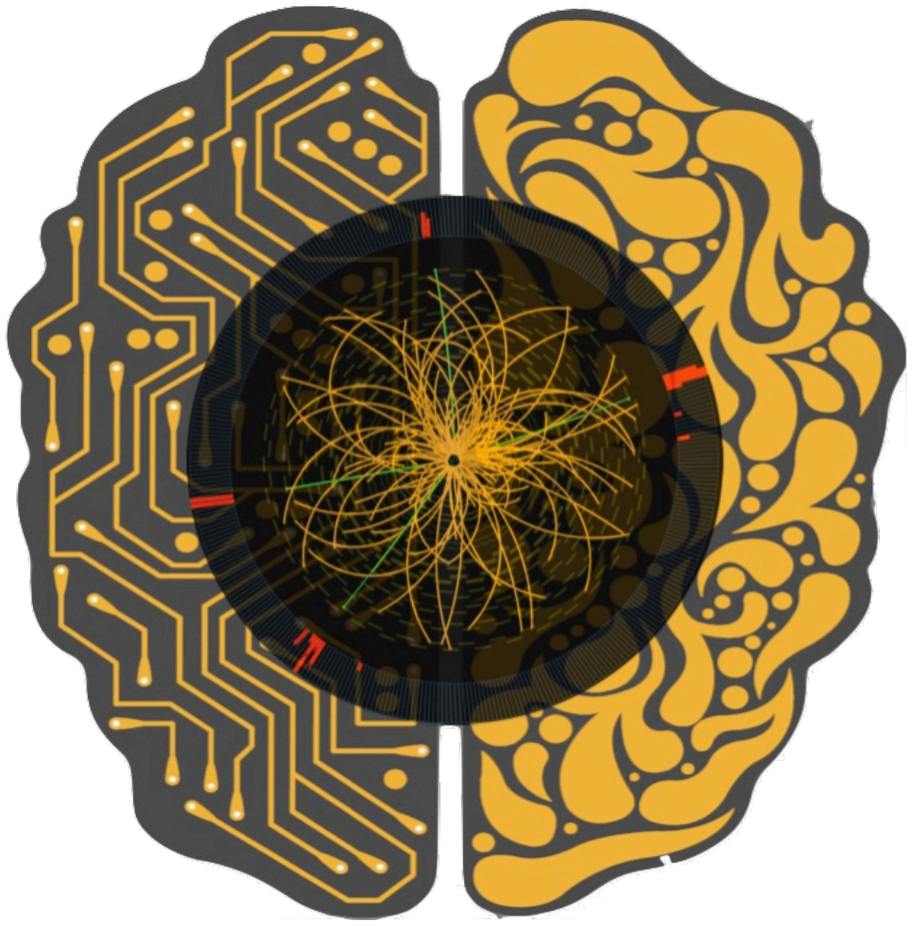


Start to observe the known lack of spin information in the Parton shower

CMS Result out targetting Winter conferences

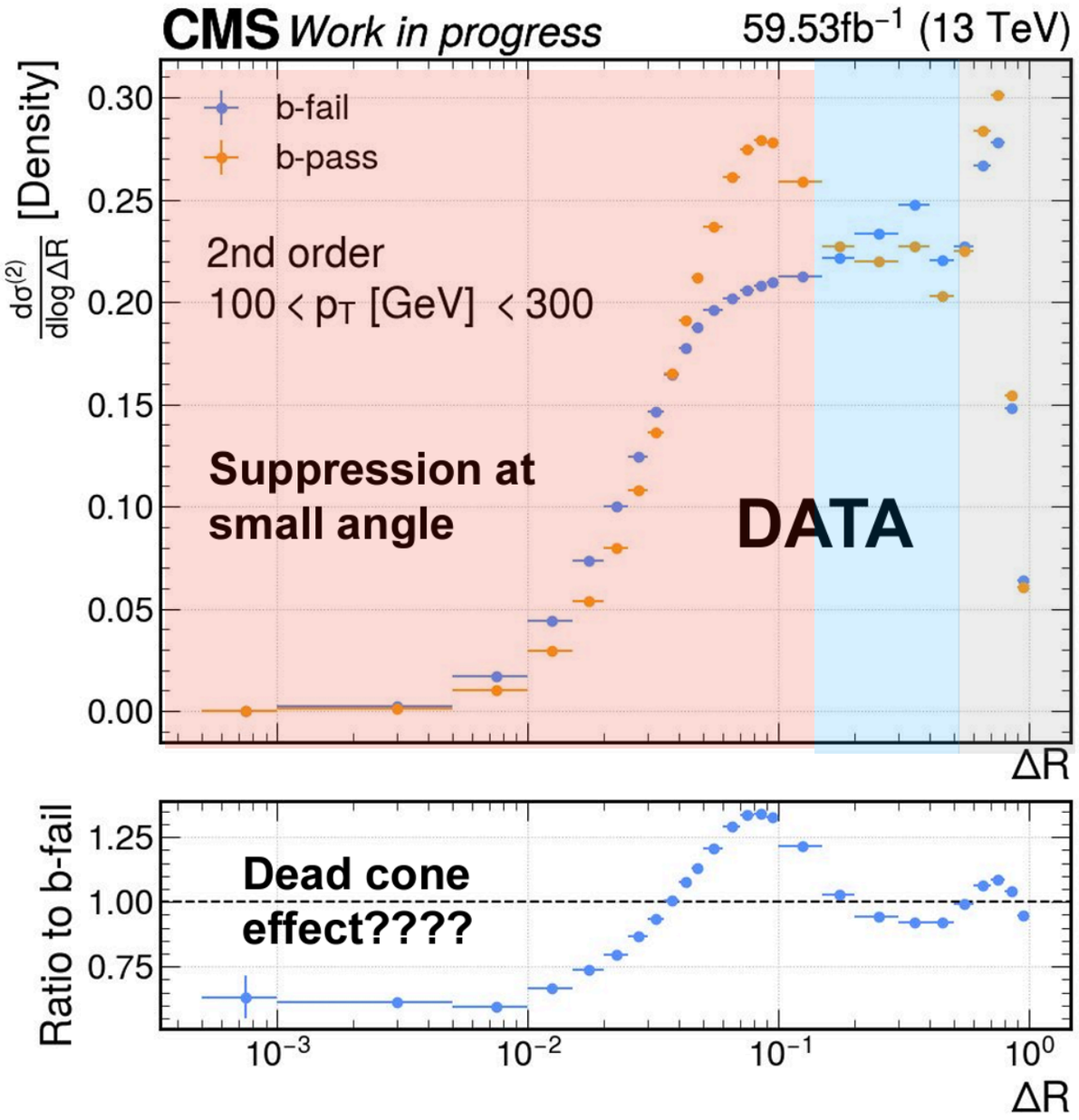
Looking Forward

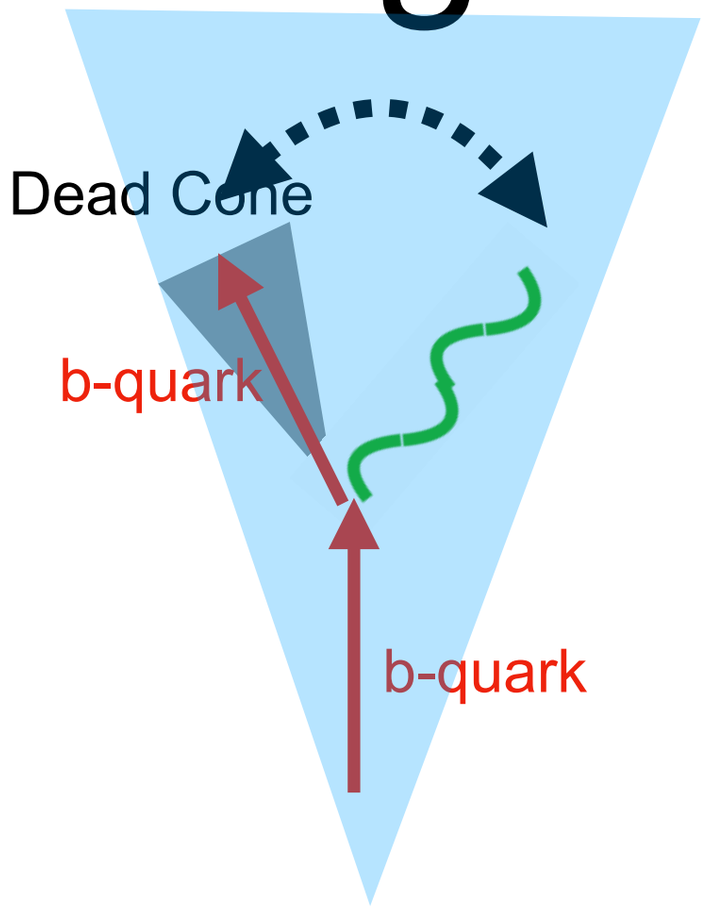
- We are preparing for future HL-LHC running
  - Upgrading core elements of CMS L1 Trigger
    - Bringing Deep Learning to the masses
  - Integrating GPUs into CMS computing model
    - Allow for rapid and flexible deployment of algos
- Pursuing novel QCD/Anomaly detection meaurements
  - Many new results and more in next few months



Thanks!

Understanding QCD



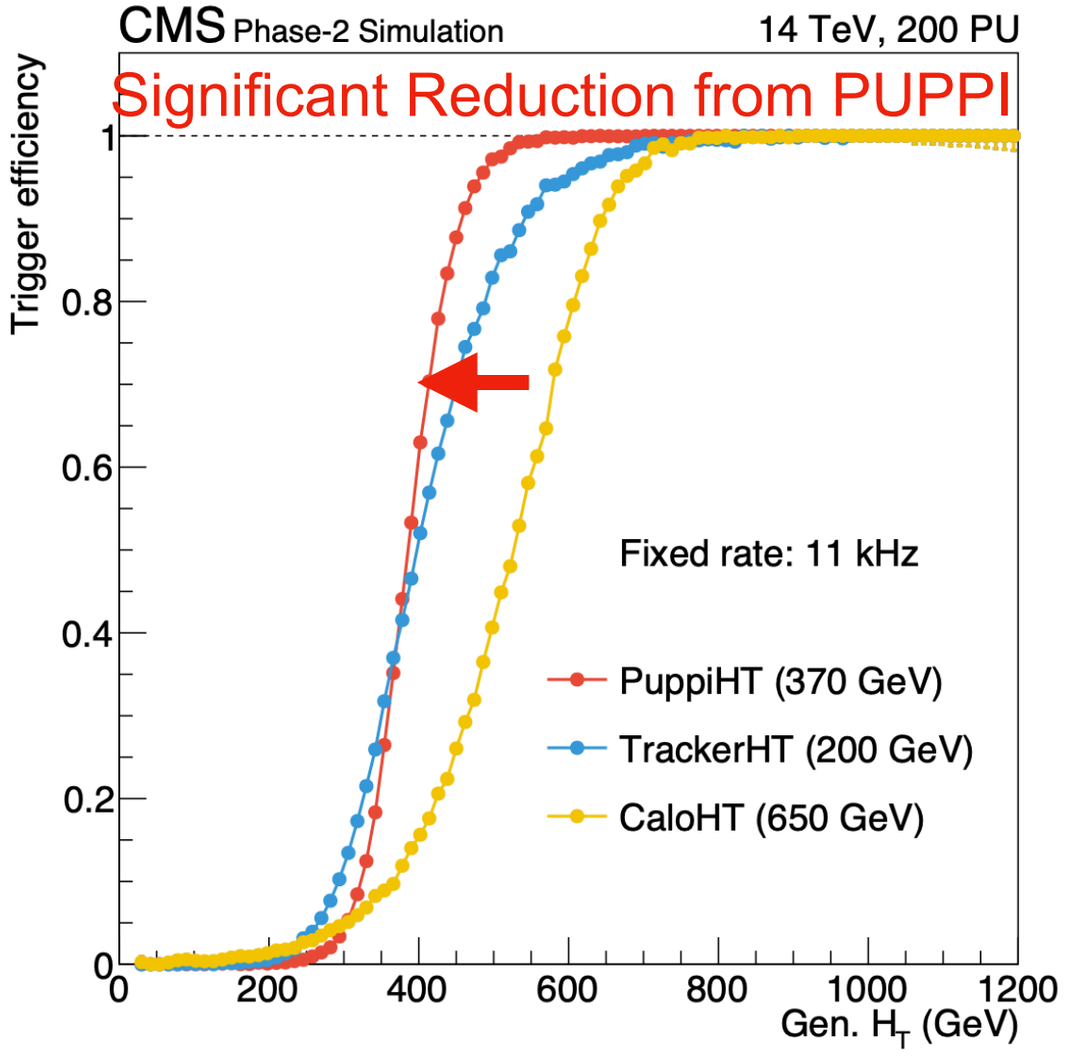


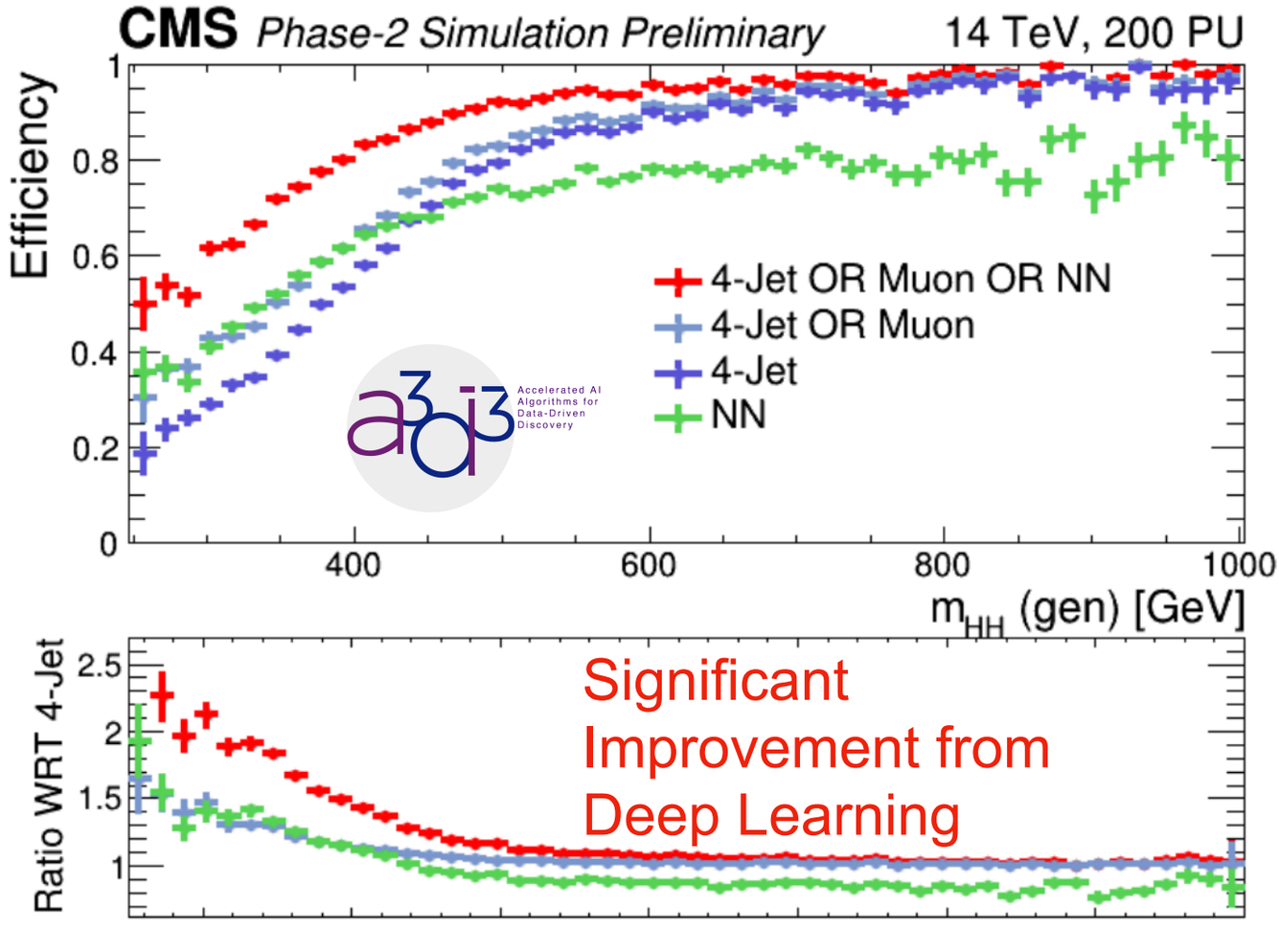
Pairwise Force of tow particles vs distance

Taking into account Flavor considerations

Enhancing Trigger

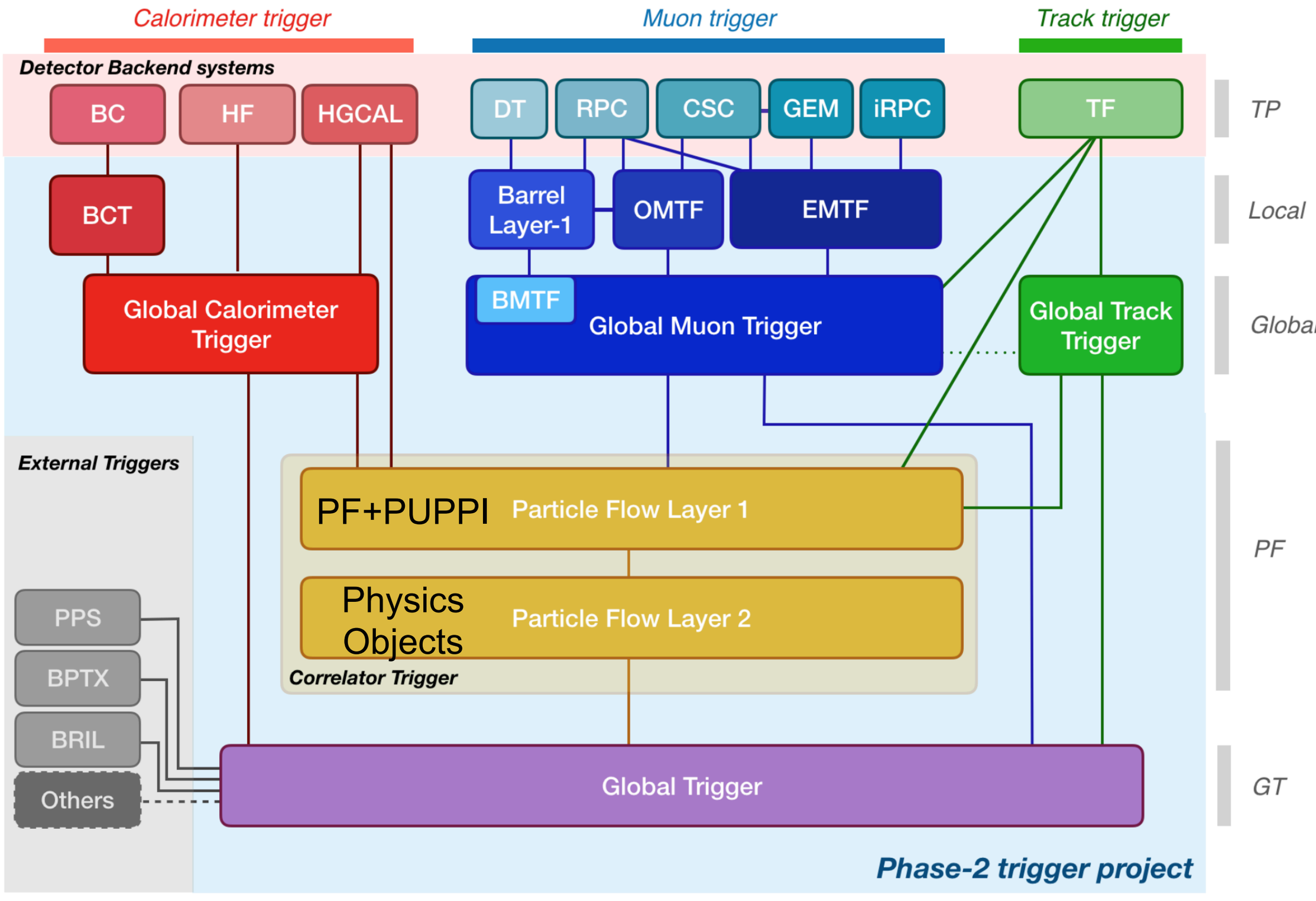
- On top of dramatic gains our reconstruction
  - Now pursuing deep learning algorithms to further enhance



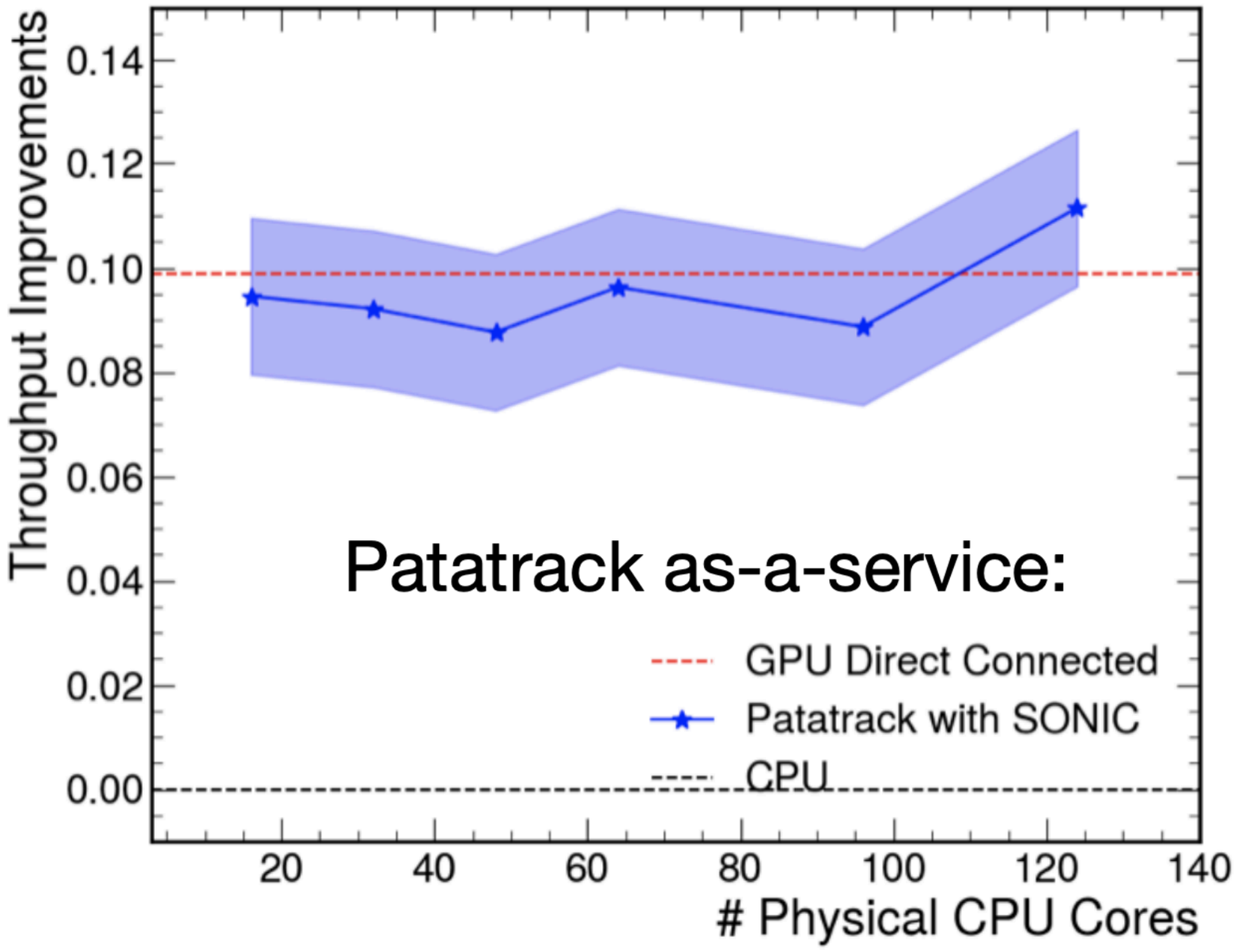


Deep learning based our toolkit will be in Run 3 CMS (possibly ATLAS)

Larger System



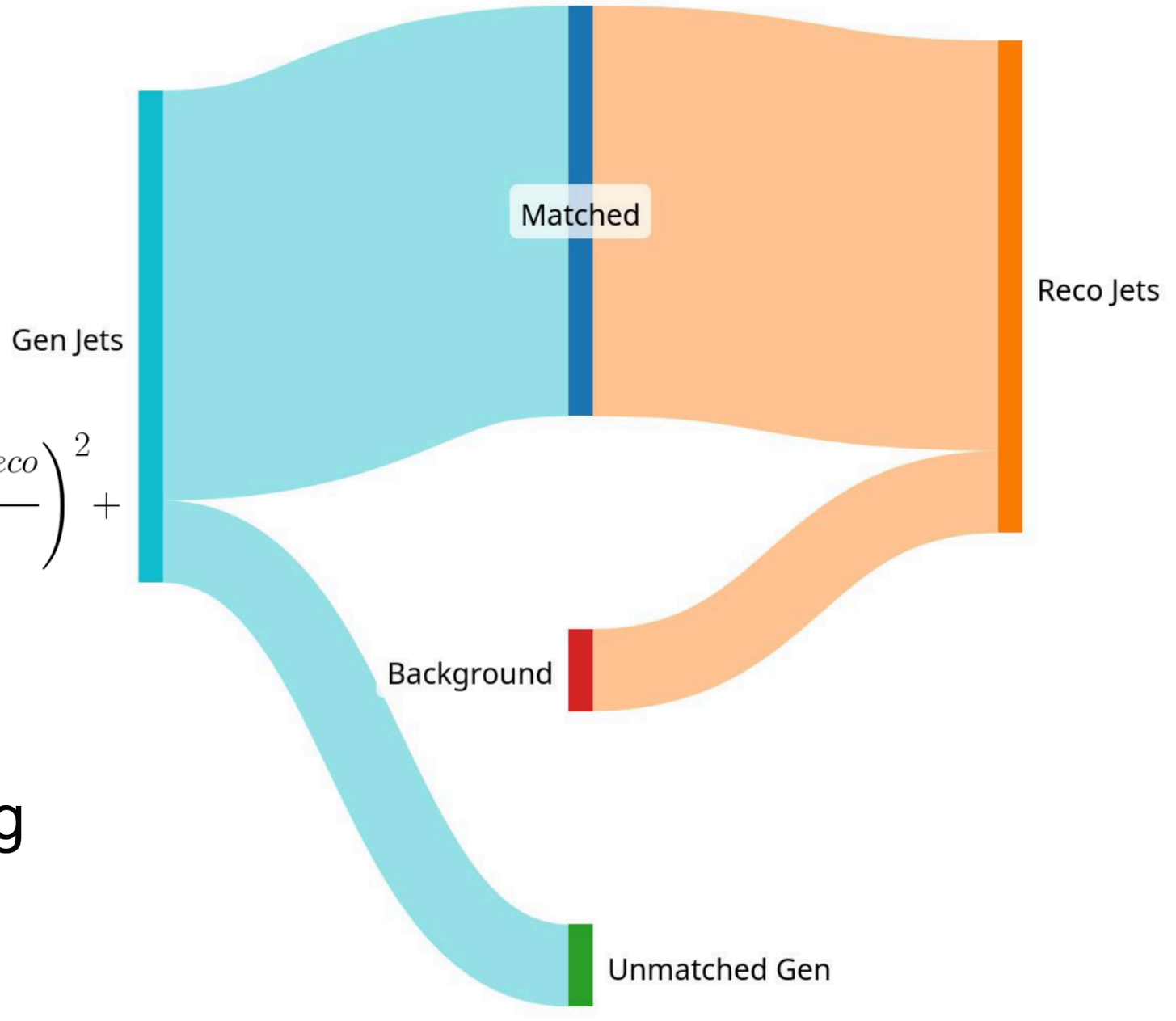
Rule Based Algorithms



Towards a Measurement

Event by Event kinematic fit
Matching Gen Particles
with Reco Particles

$$\chi^2 = \left(\frac{p_T^{Matched} - p_T^{Reco}}{\sigma_{p_T}}\right)^2 + \left(\frac{\eta^{Matched} - \eta^{Reco}}{\sigma_\eta}\right)^2 + \left(\frac{\phi^{Matched} - \phi^{Reco}}{\sigma_\phi}\right)^2$$



Obtain the exact matching
of each particle
Allows for factorized
unfolding of any distribution