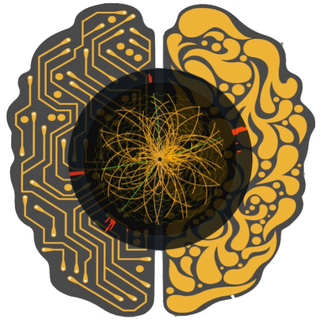
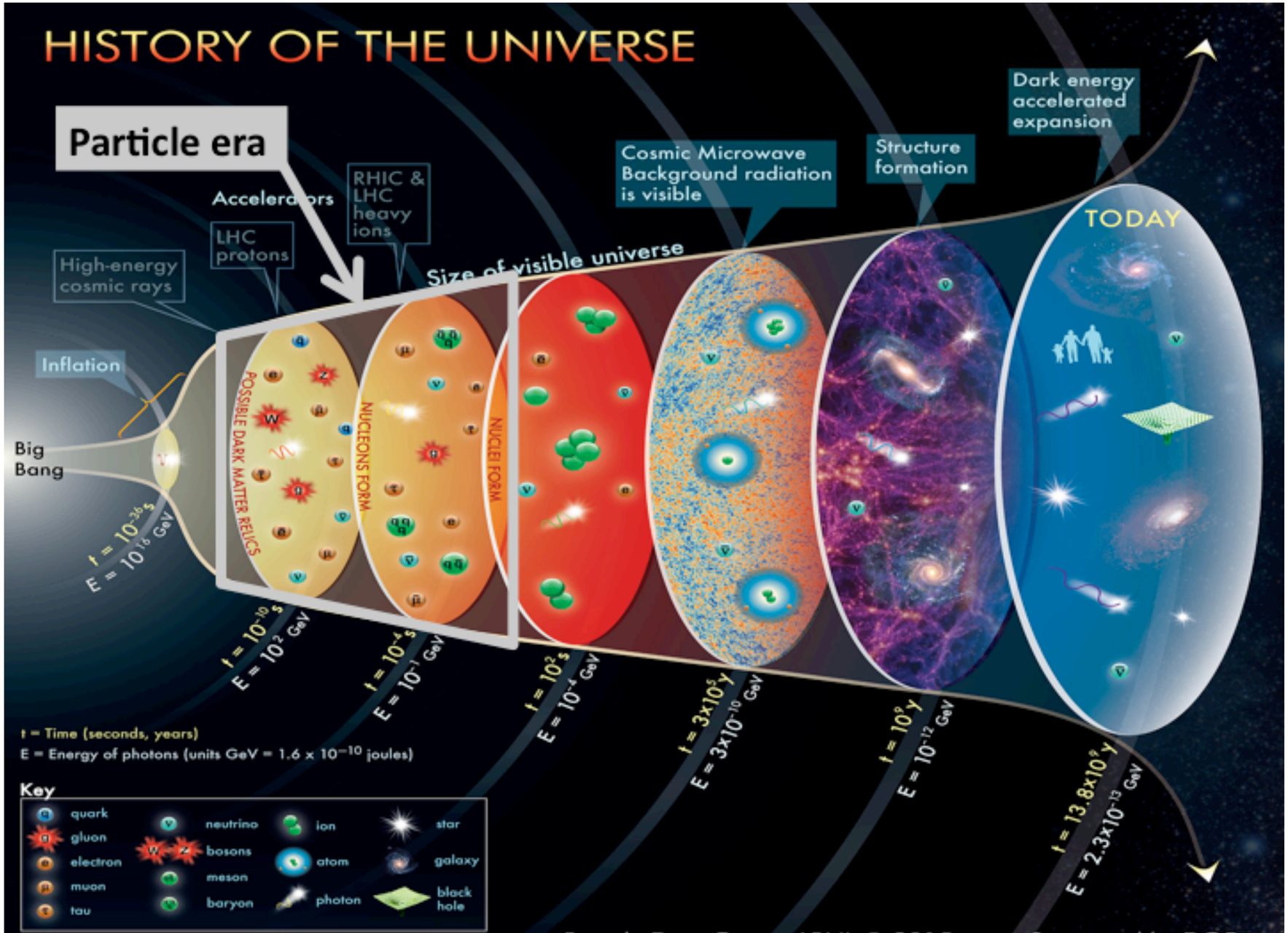
Artificial Intelligence Accelerated Discoveries At the Large Hadron Collider

Mia Liu
Purdue University

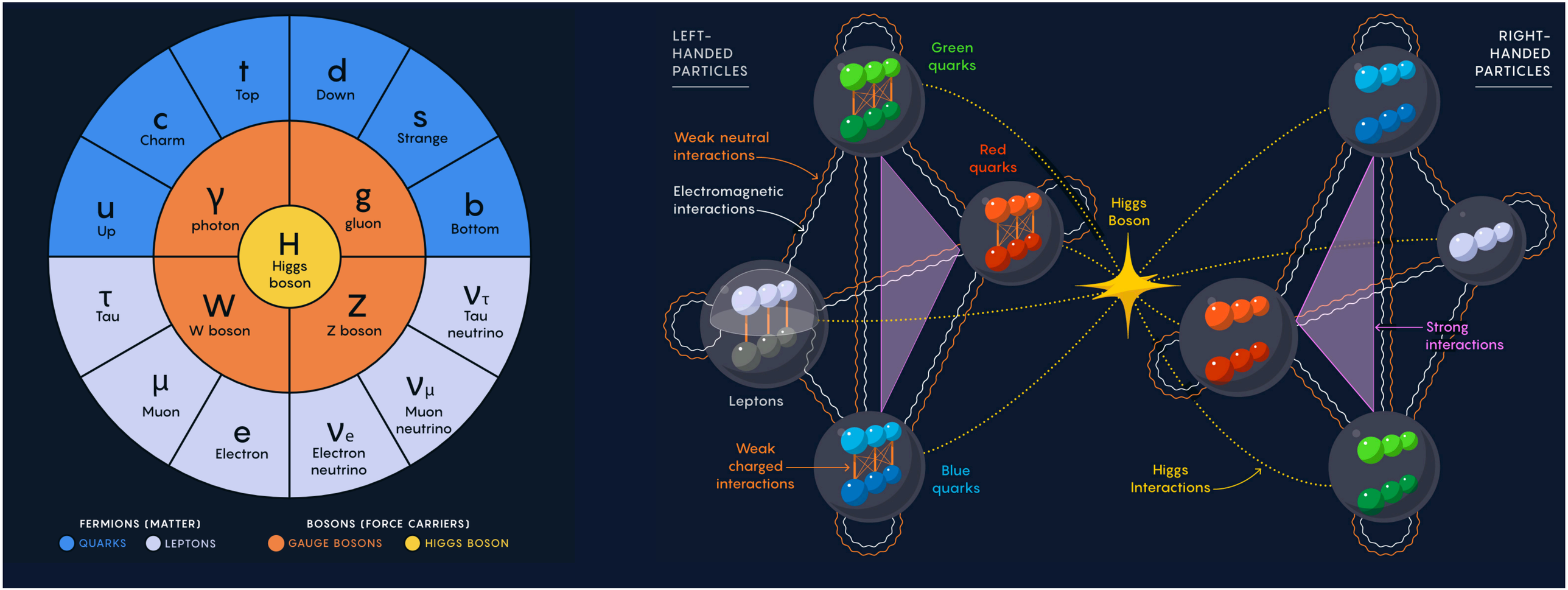
Oct 18, 2021
A3D3 pre kick-off seminar



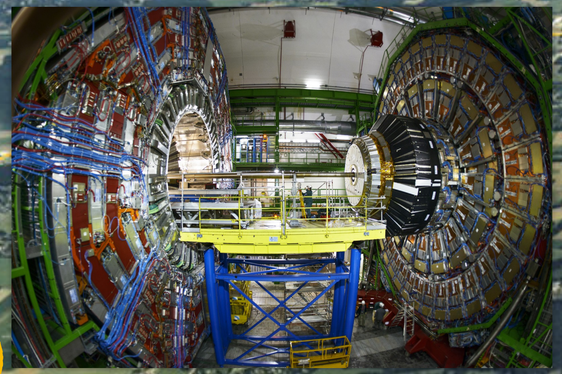
Recreating the universe after the big bang 2

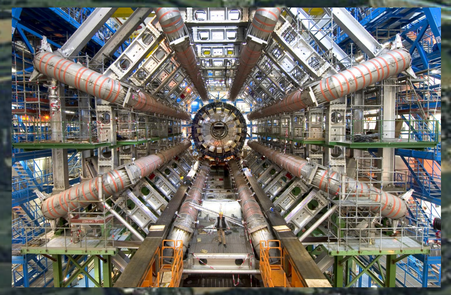


The Standard Model of Fundamental Particles

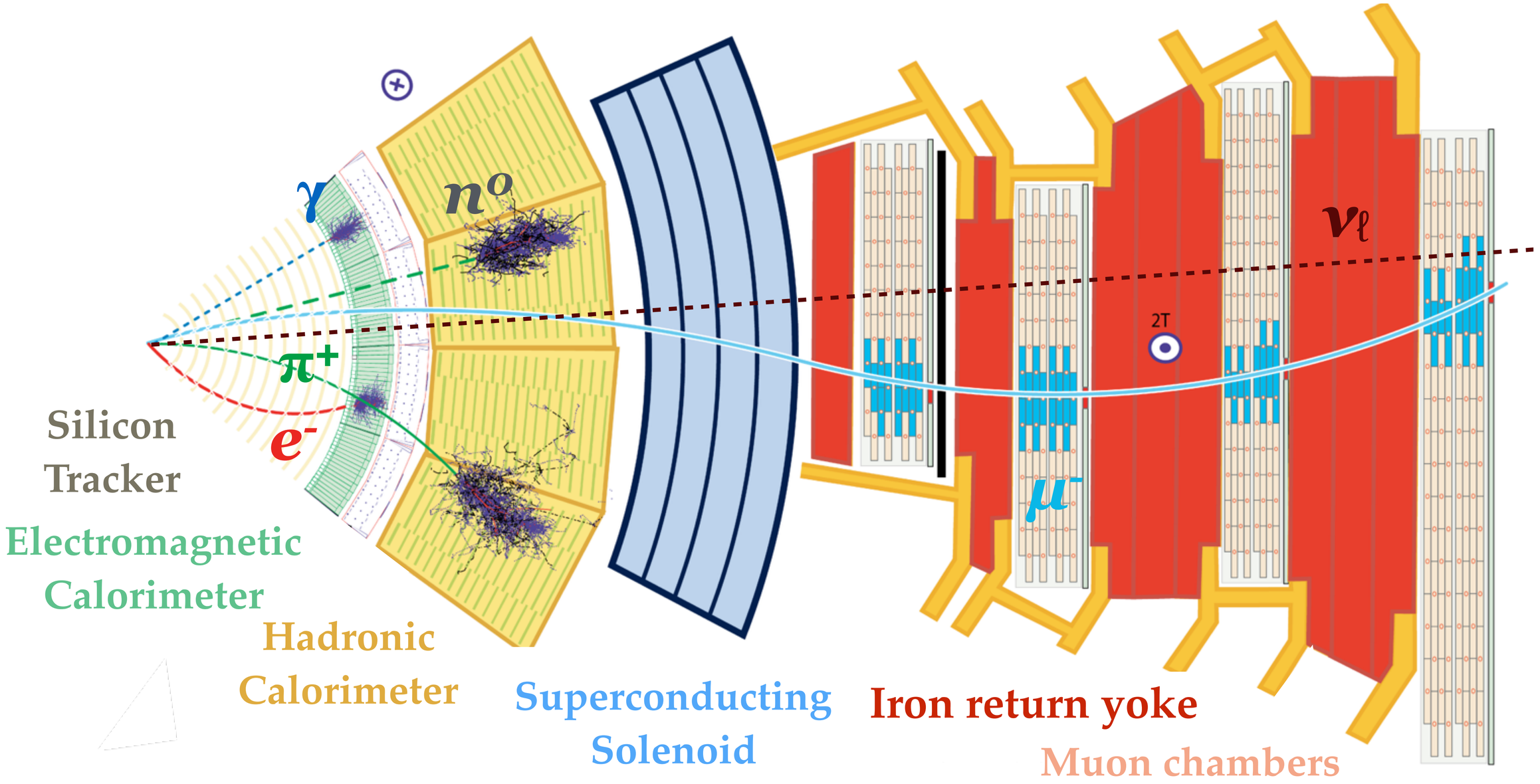


LAST MISSING PIECE
2012: HIGGS DISCOVERED AT THE LHC

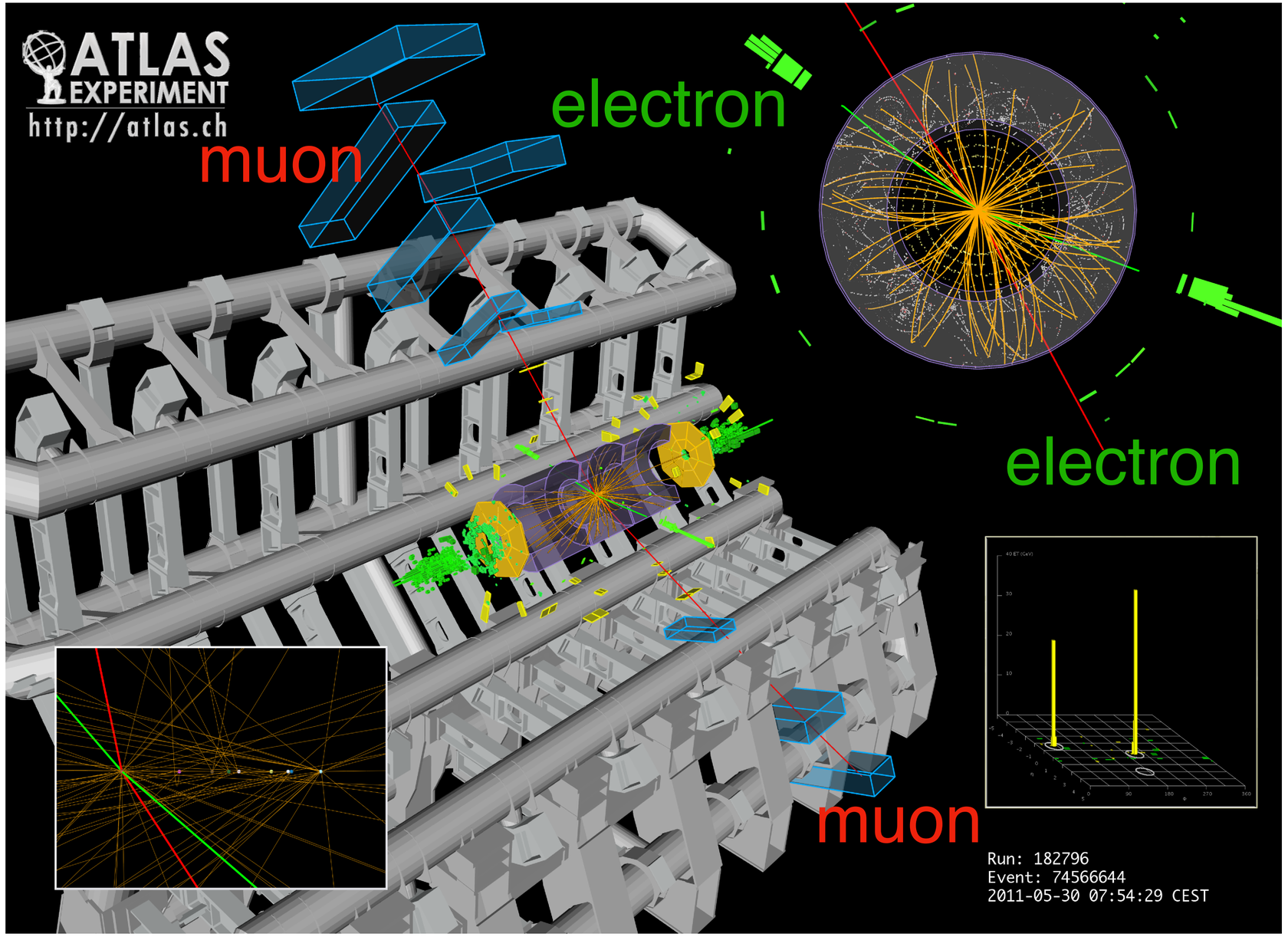




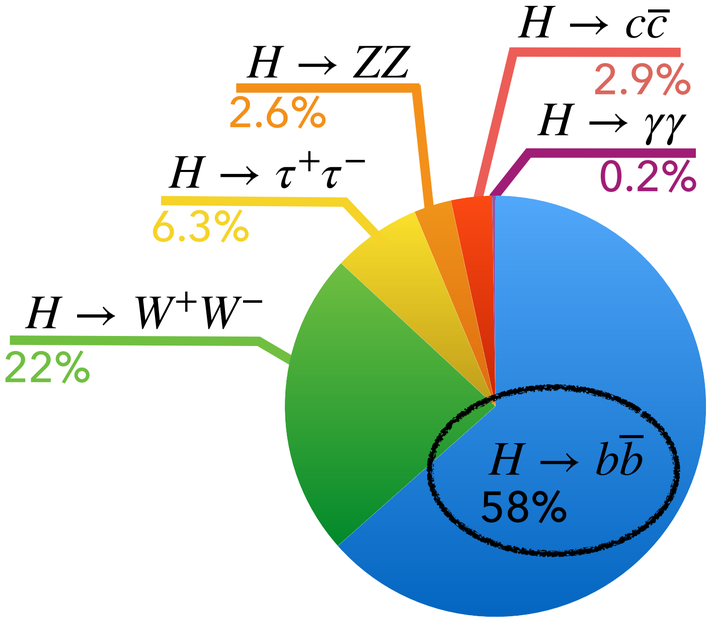
Particle Detection in CMS



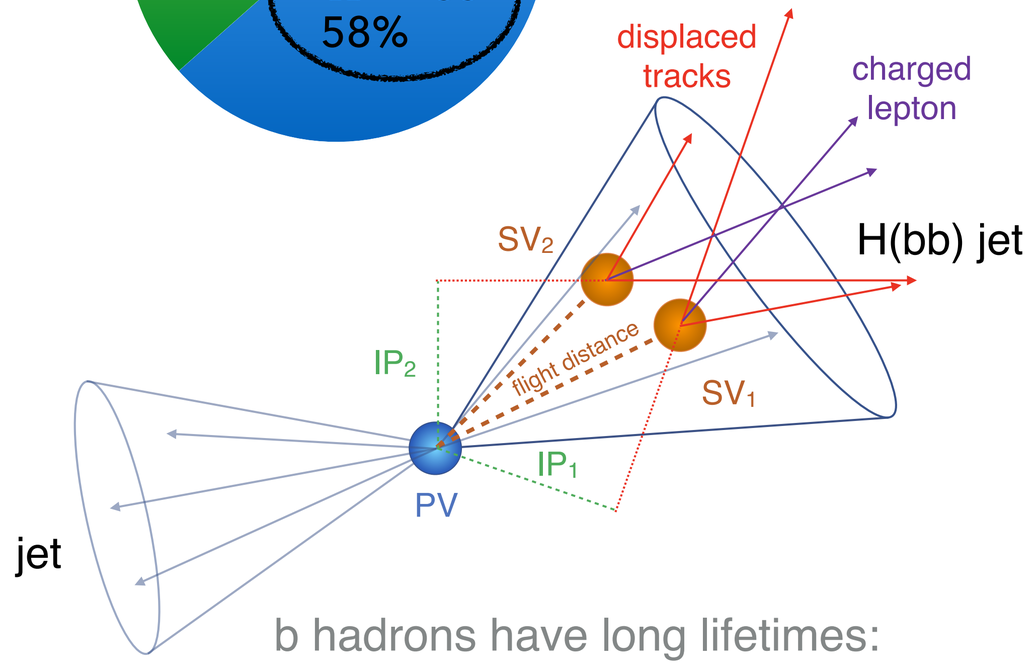
Discovering the Higgs boson (H->ZZ->eeμμ)



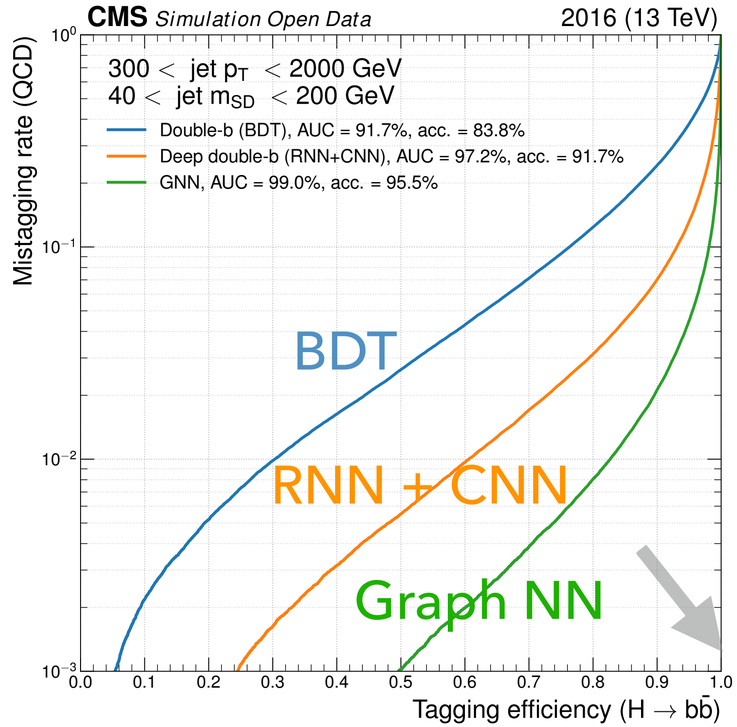
Tagging boosted Higgs with Machine Learning



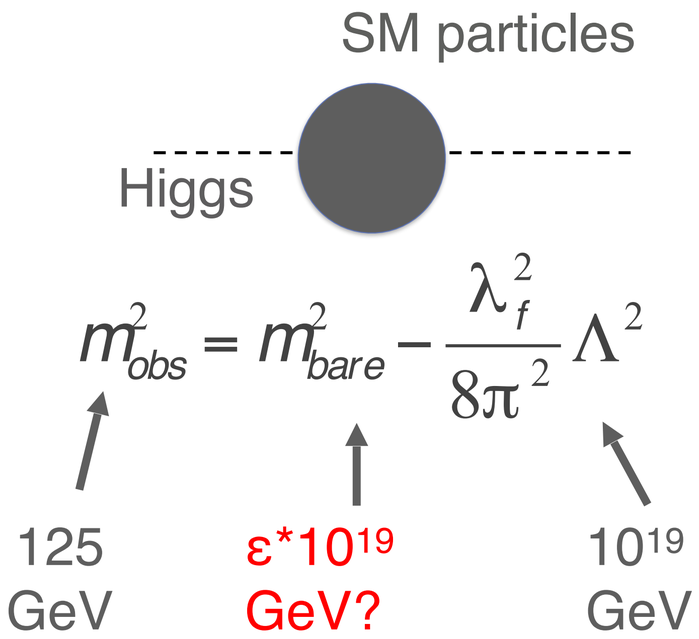
- 10^7 more frequent background
- Complex signature:
  - secondary vertices
  - displaced tracks
  - large impact parameters
  - soft leptons



b hadrons have long lifetimes:
travel O(mm) before decay!



Puzzles



Fine tuning

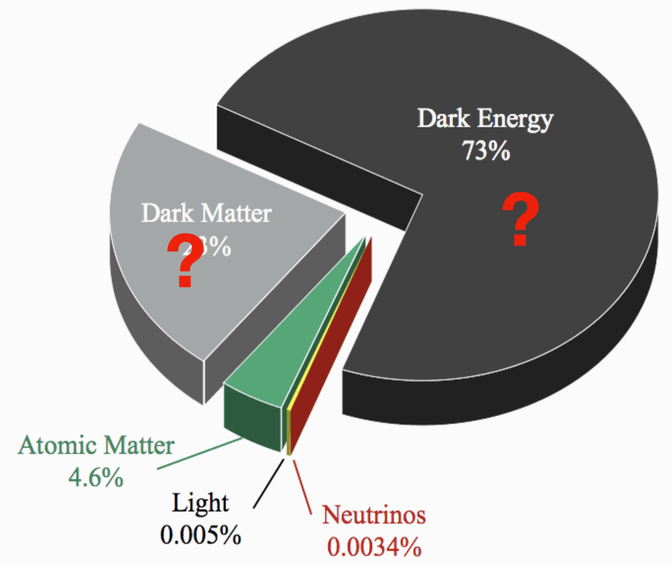
Dynamical origin

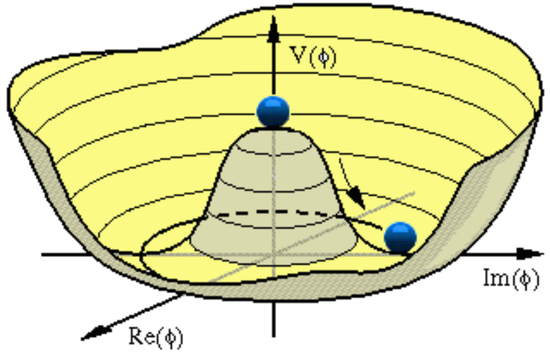
…

Experimental:

Dark matter/dark energy

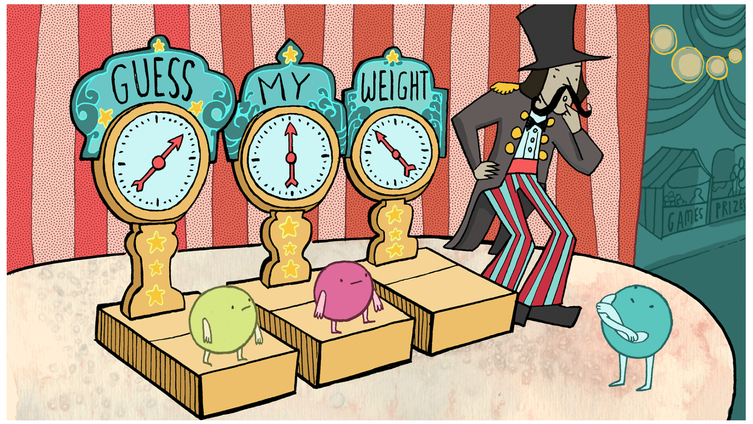
Not in SM



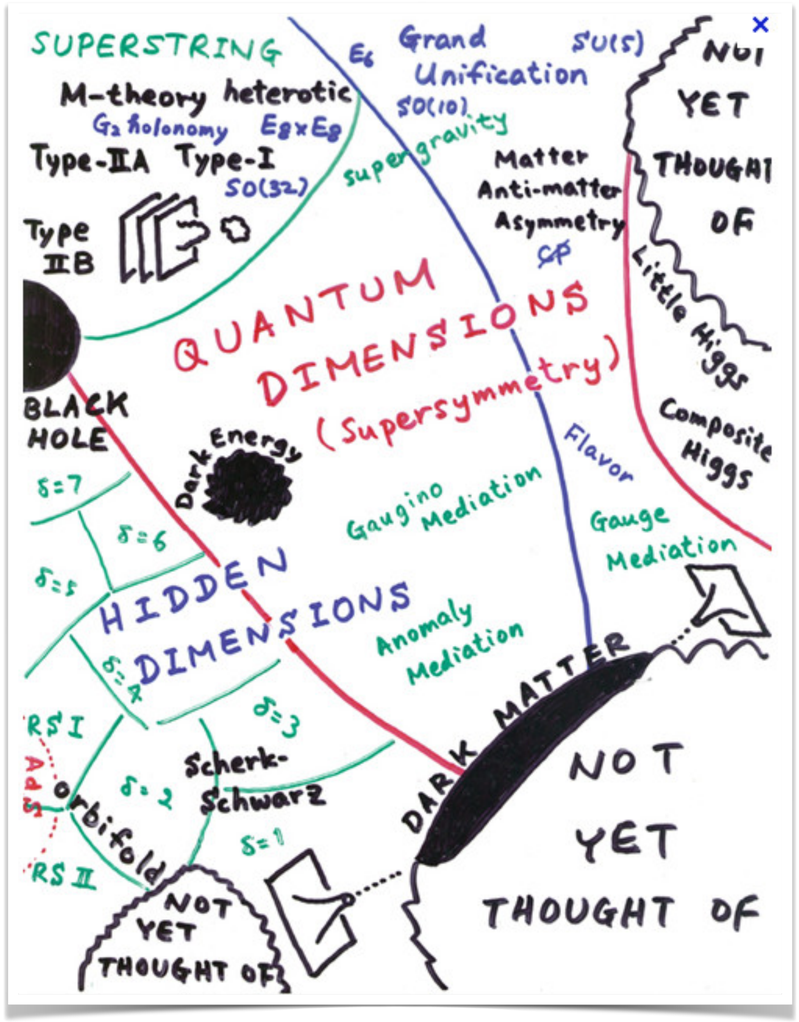


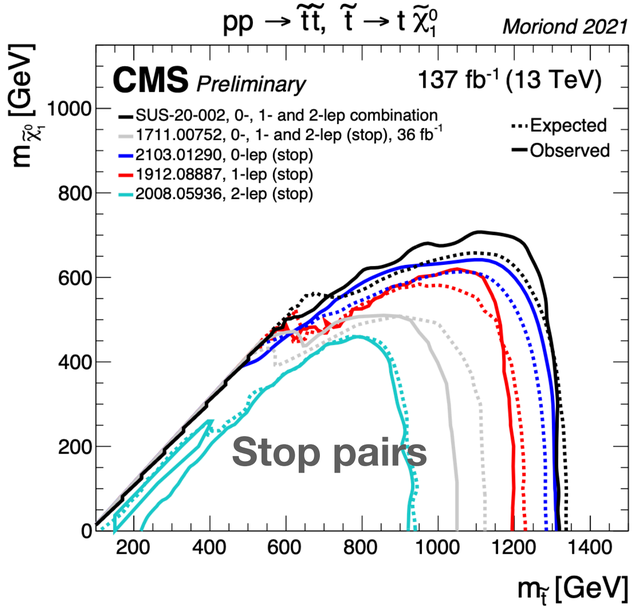
Neutrinos

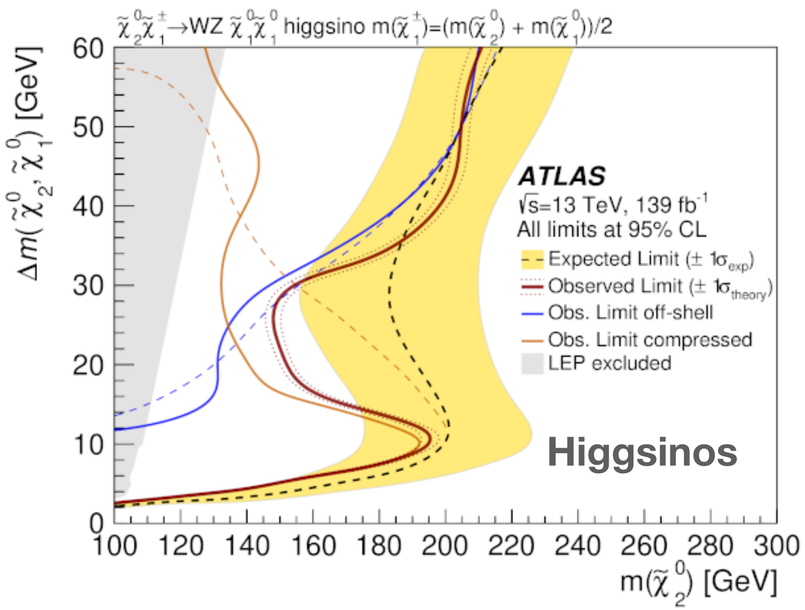
Anomalies: Muon g-2, LHCb lepton flavour universality

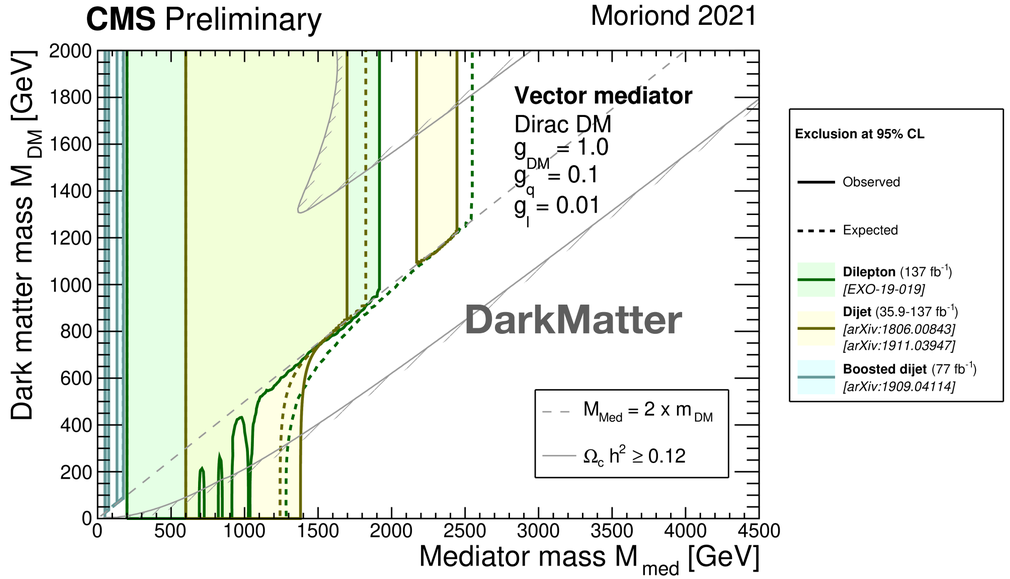


Extensive searches performed at the LHC

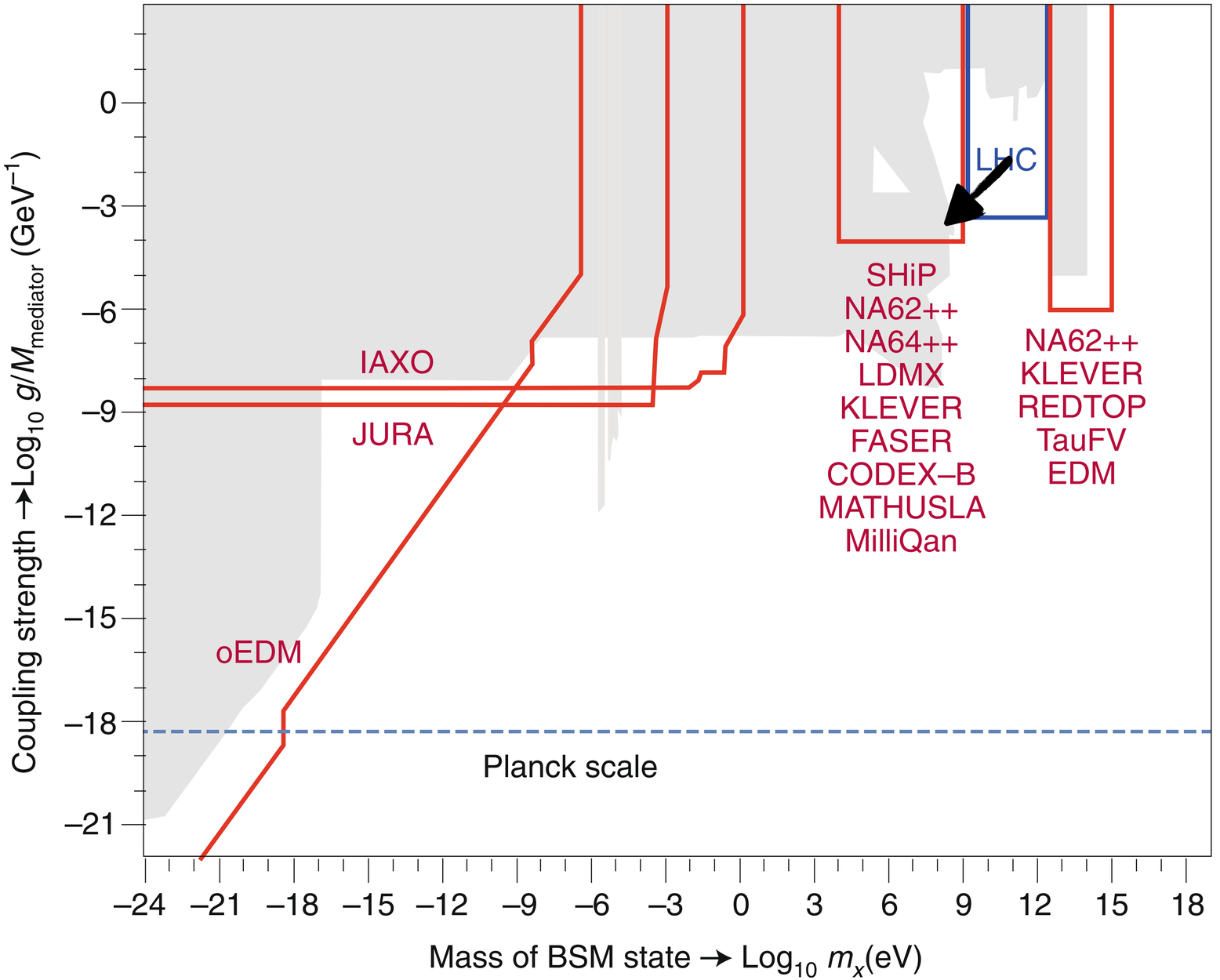








Improving search space at the LHC

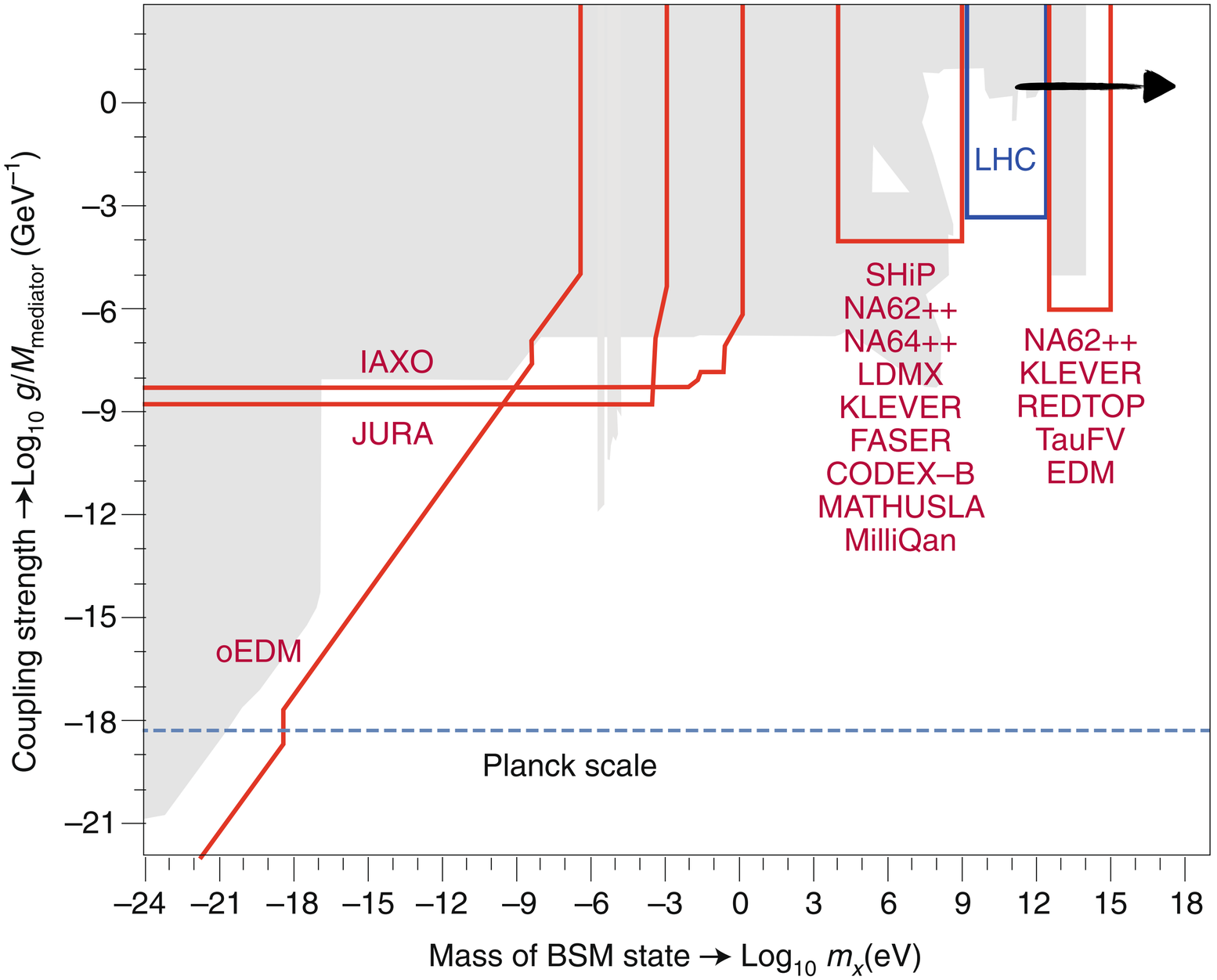


Low mass, weakly coupled BSM particles:

Unconventional signatures:
Long lived particles that decay outside beam pipe

Many are limited by online reconstruction/ filtering (trigger):
Currently relying on additional object present in the event

Improving search space at the LHC



Improve high energy probe:

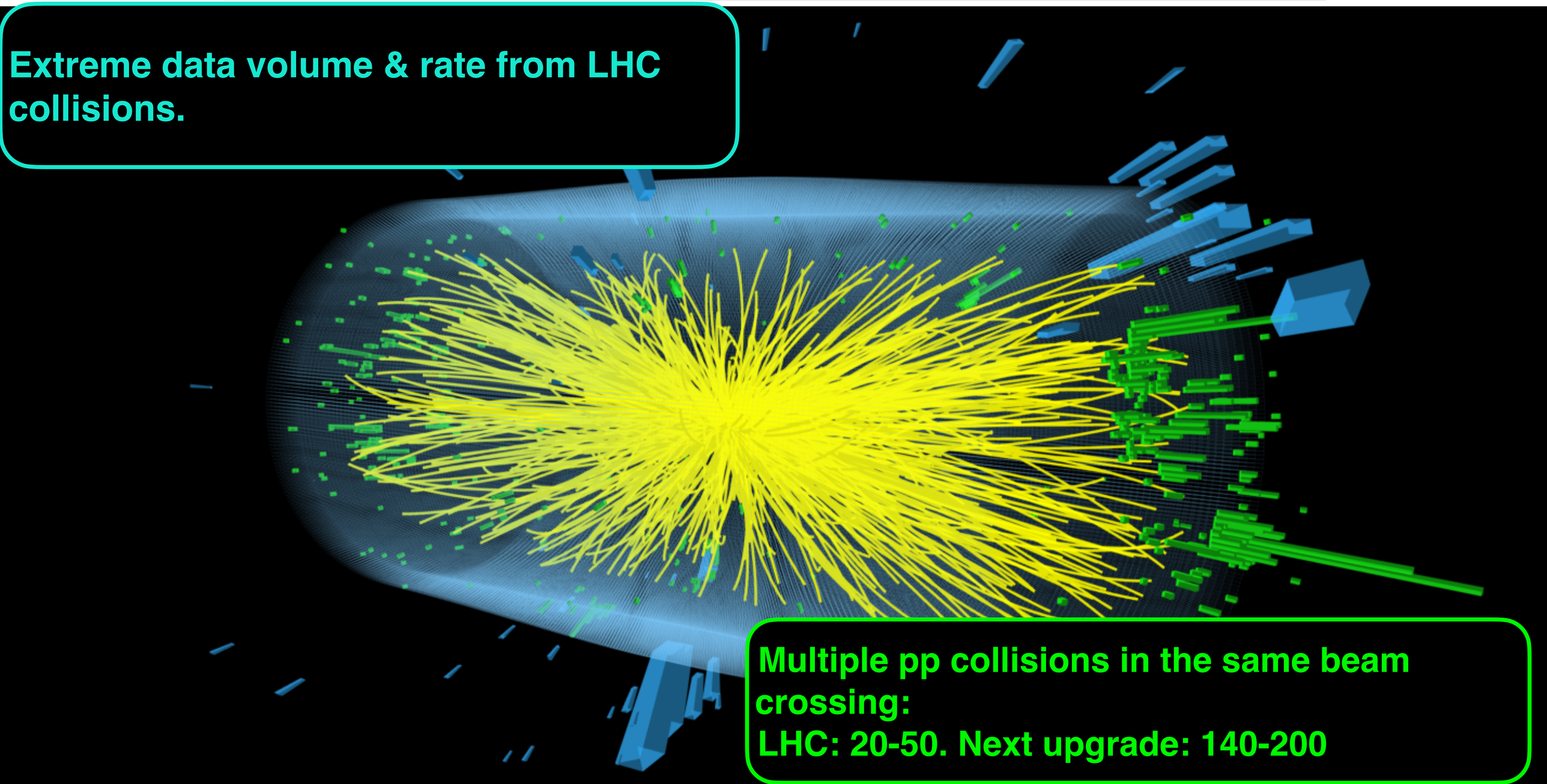
high pT object detection. e.g. boosted Higgs/W/Z/top etc.

Loop contributions to ultra-rare processes.e.g lepton flavor violation

- **soft signatures that are limited by online reconstruction/filtering (trigger) as well**

The Fast and Furious

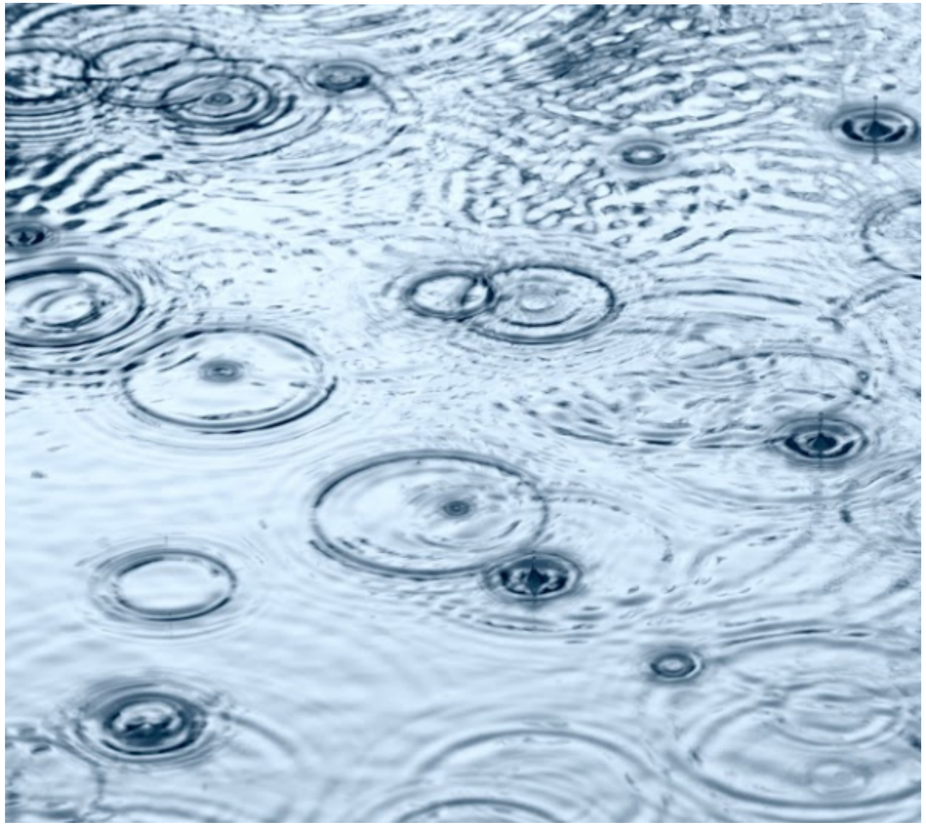
Extreme data volume & rate from LHC collisions.



Multiple pp collisions in the same beam crossing:
LHC: 20-50. Next upgrade: 140-200

Capturing raindrops

Now: The LHC



LHC Run-4

Faster and fine granularity detectors to collect higher pileup collisions

ML to maintain/improve detection of (un)conventional signatures

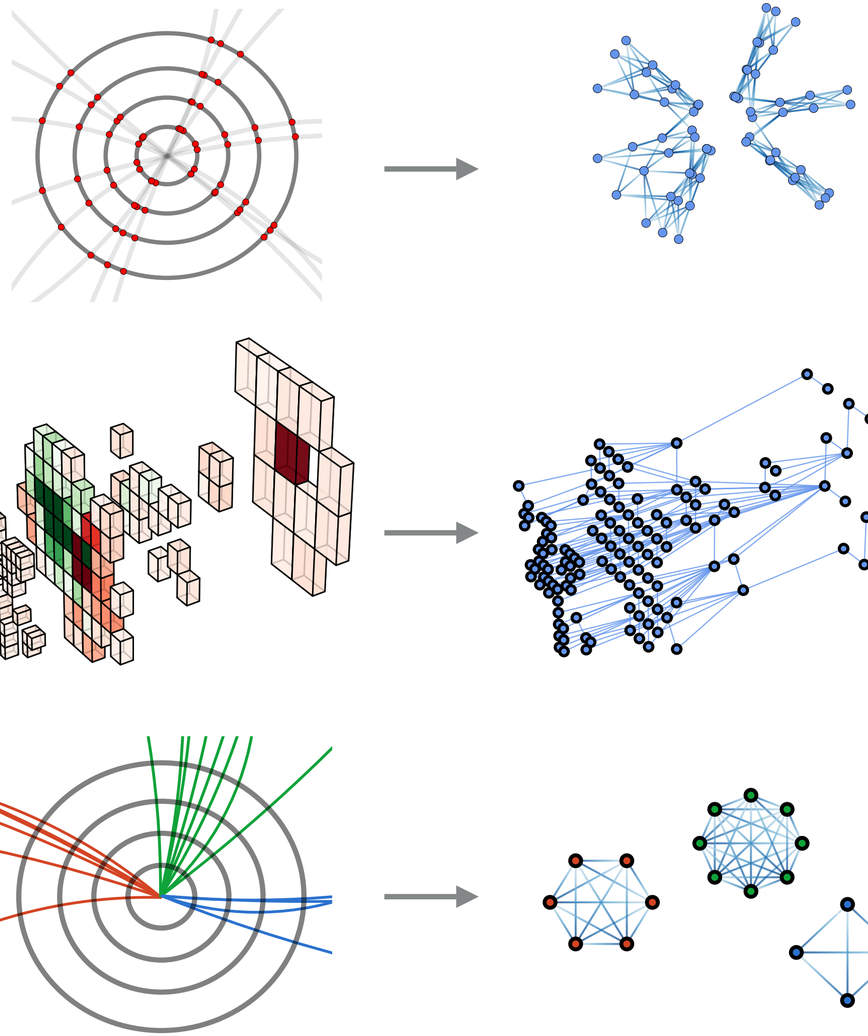
More and more… and more GNNs

Represent data as nodes and edges

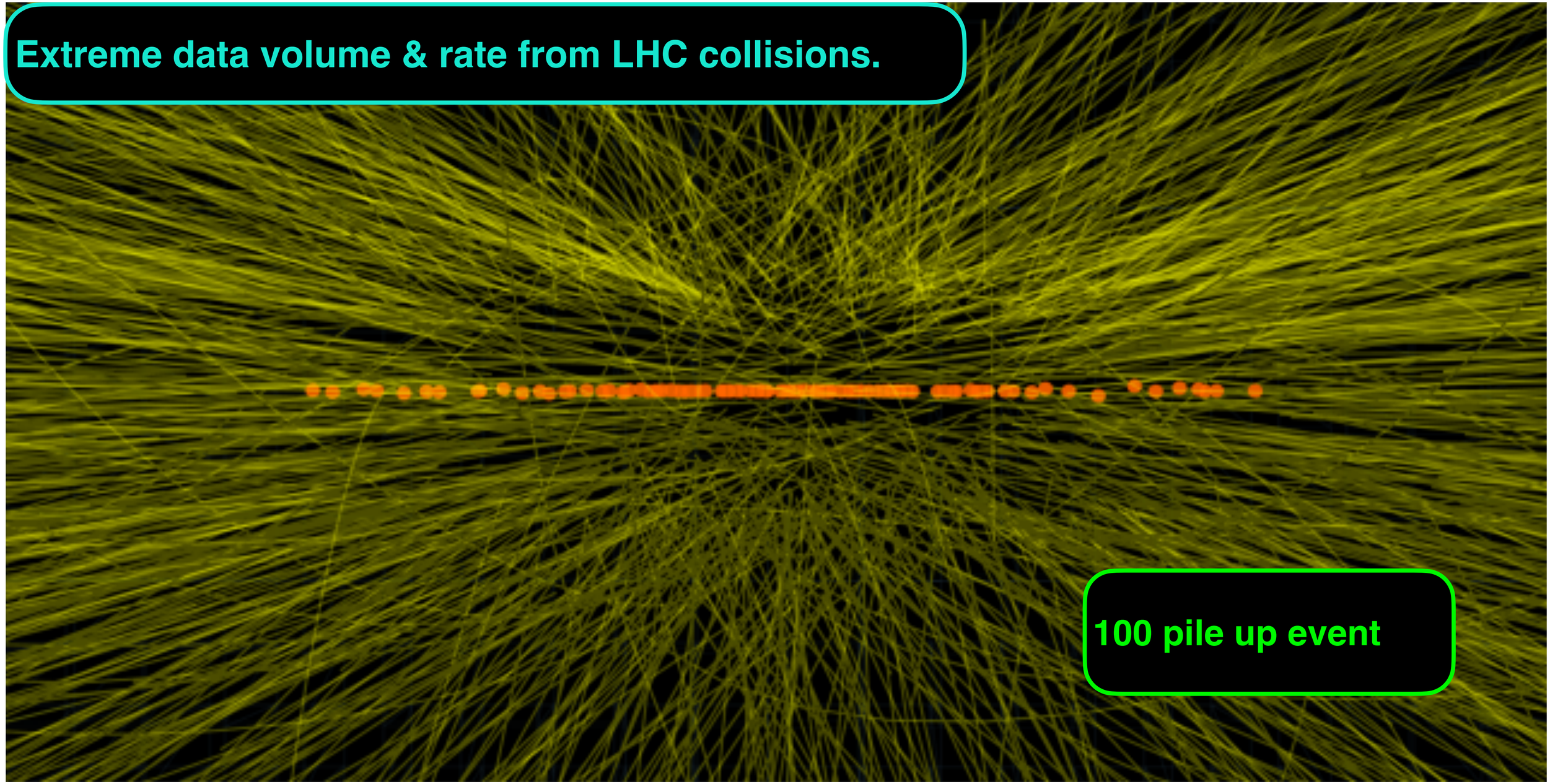
appropriate representation of particle physics data: irregular, structural, relational

GNN developments driven by social media, how to adapt/adopt to domain knowledge

- node classification: Pileup mitigation
- graph classification: τ->3μ
- edge classification: charged particle tracking
- identifying subgraphs/sets: particle flow



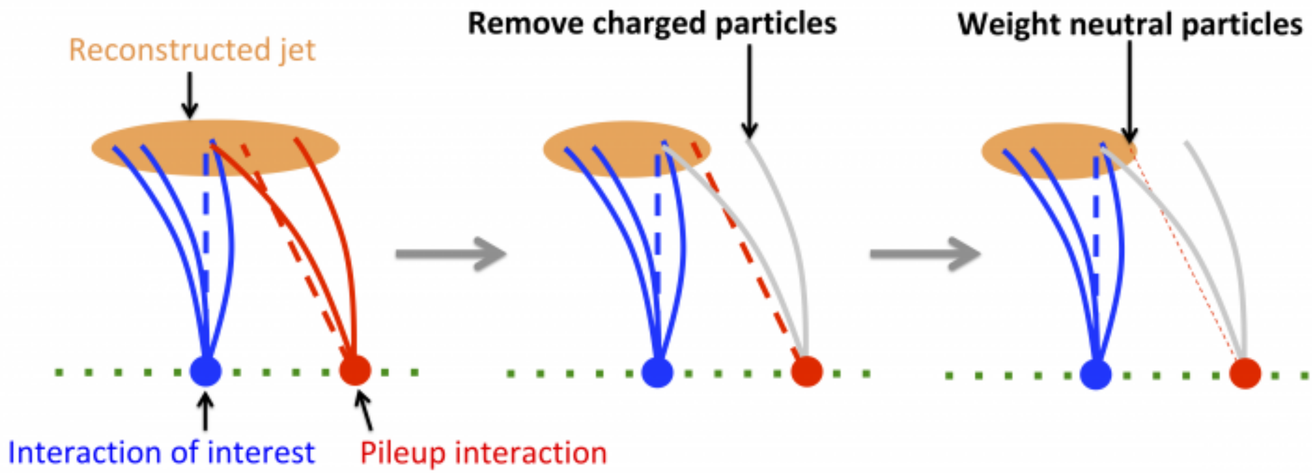
The Fast and Furious



Pileup mitigation

Essential to the LHC physics program.

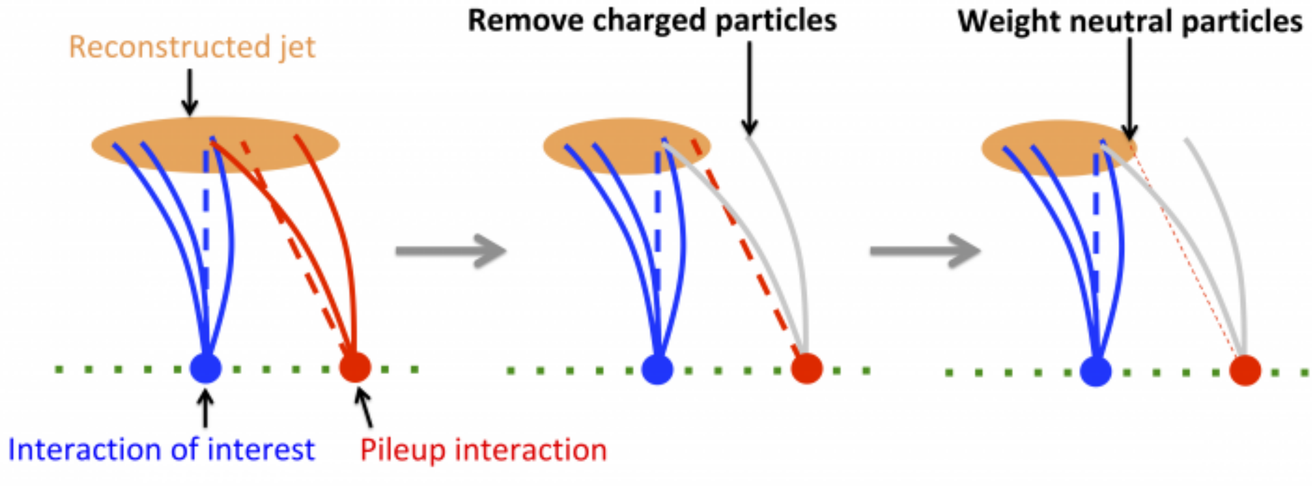
- “Noise” needs to be subtracted from interaction of interest

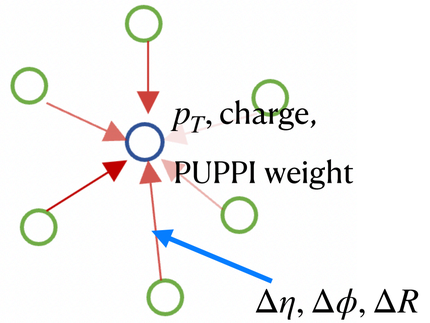


Semi-supervised Pileup mitigation

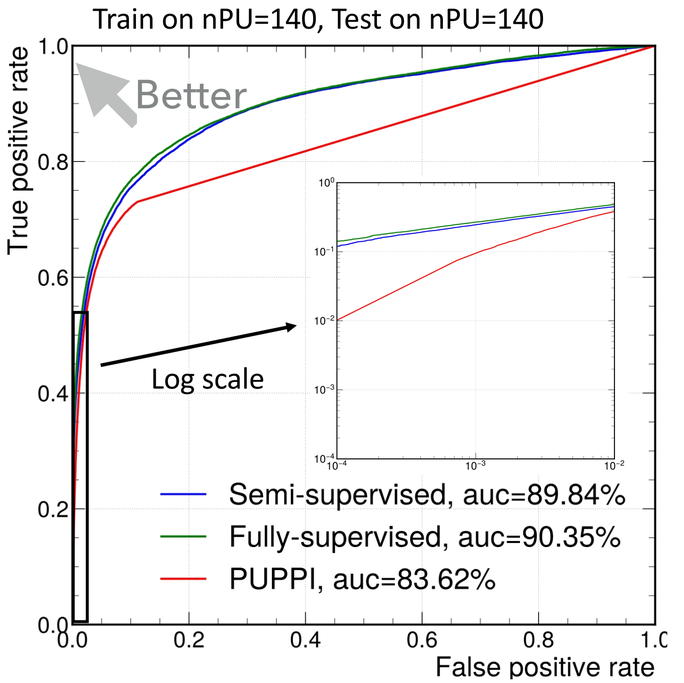
Trained on labeled charged particles and applied to neutral ones —> can learn from data.

- Outperforms expert feature method (PUPPI), comparable to fully supervised method.
- Presented at BOOST 2021, Papers coming soon.
- Next: Apply to CMS simulation & data. Neutral particle vertex association for the forward region. Physics with PU interactions?

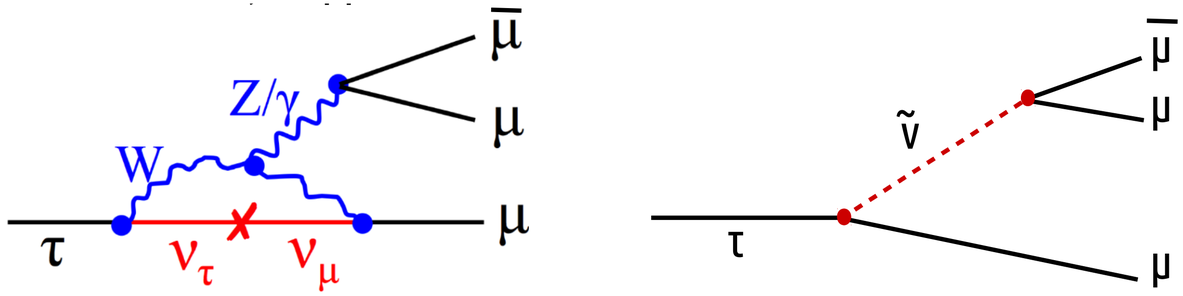




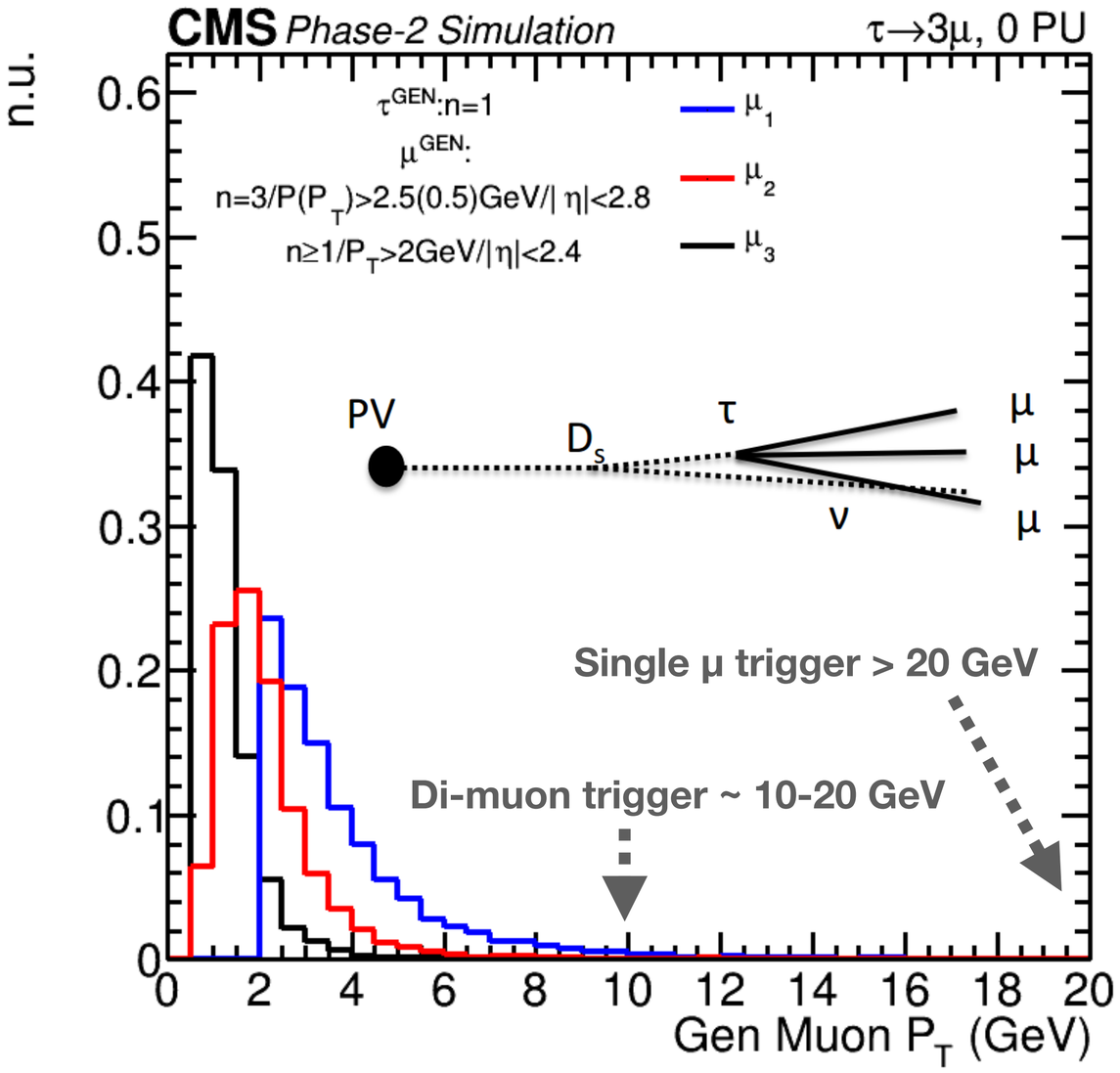
Neighboring structure



Graph Classification τ->3μ



- **Detecting lepton flavor violating τ->3μ decays**
  - Very (super ultra) rare in the SM, Neutrino oscillations: BR ~O(10^{-14})
  - BSM e.g. R-parity violating SUSY enhanced: O(10^{-8})
- **Collimated, low pT and very forward muons**
- Next upgrade uses tracking information, **only 25% efficient** (already 10 times better than current method)



Graph Classification τ->3μ

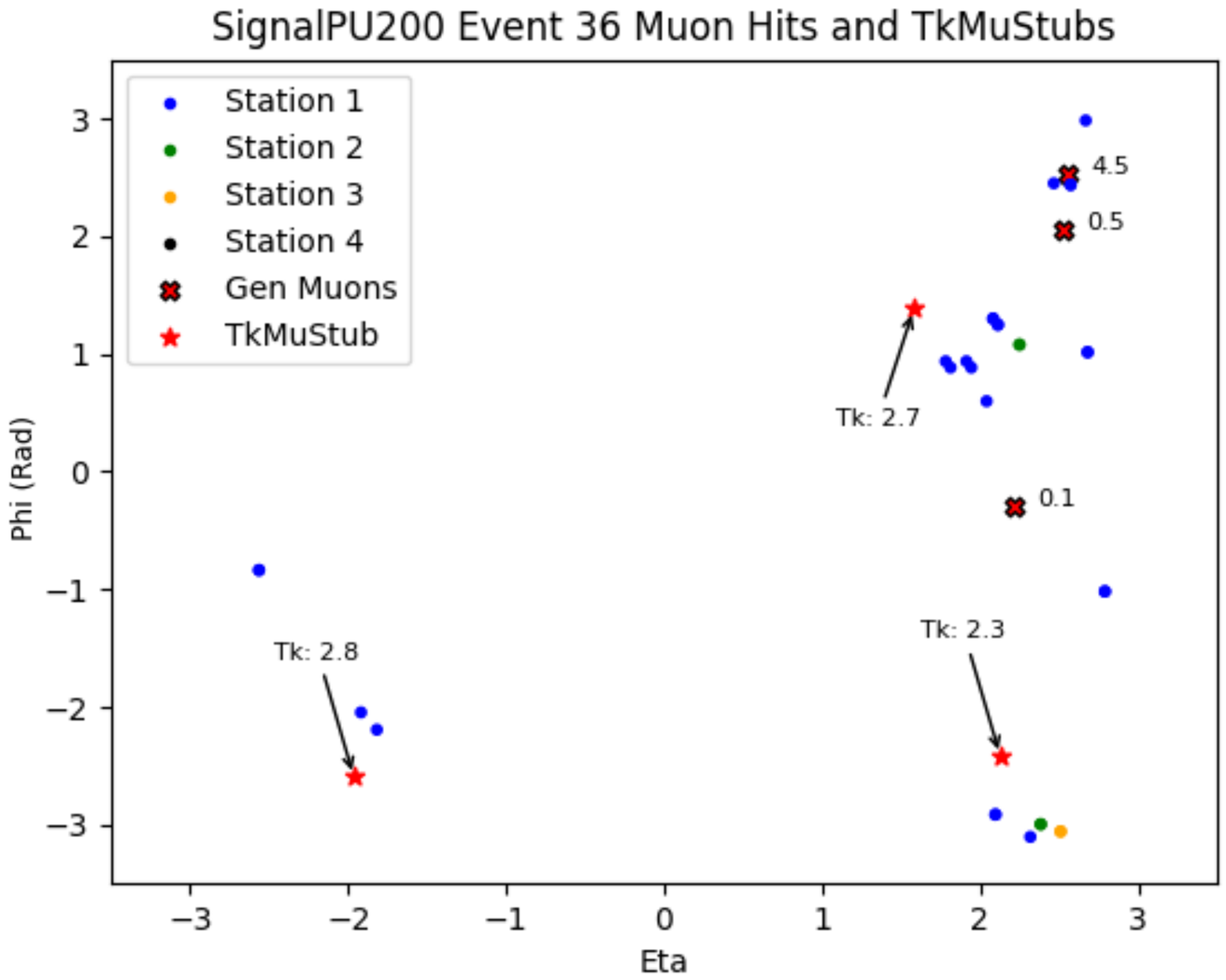
Graph Inputs: Muons hits (Coordinates and bending angle), represented as nodes in graph.

Attention based GNNs for information aggregation between local nodes and global node

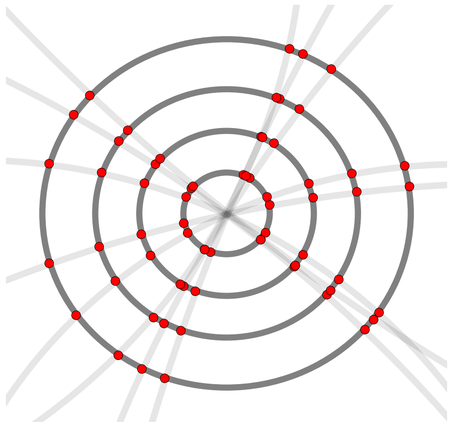
Training setting: τ->3μ signals mixed with PU backgrounds vs pure PU sample

> 90% efficient within trigger bandwidth.

End-to-end approach. Model interpretations, Model adaption (to other signals & other experiments),Data augmentation. Regression of τ/μ kinematics

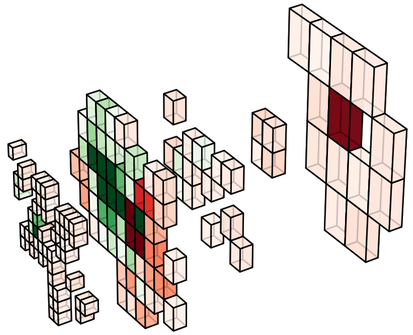


Reconstruction at the LHC



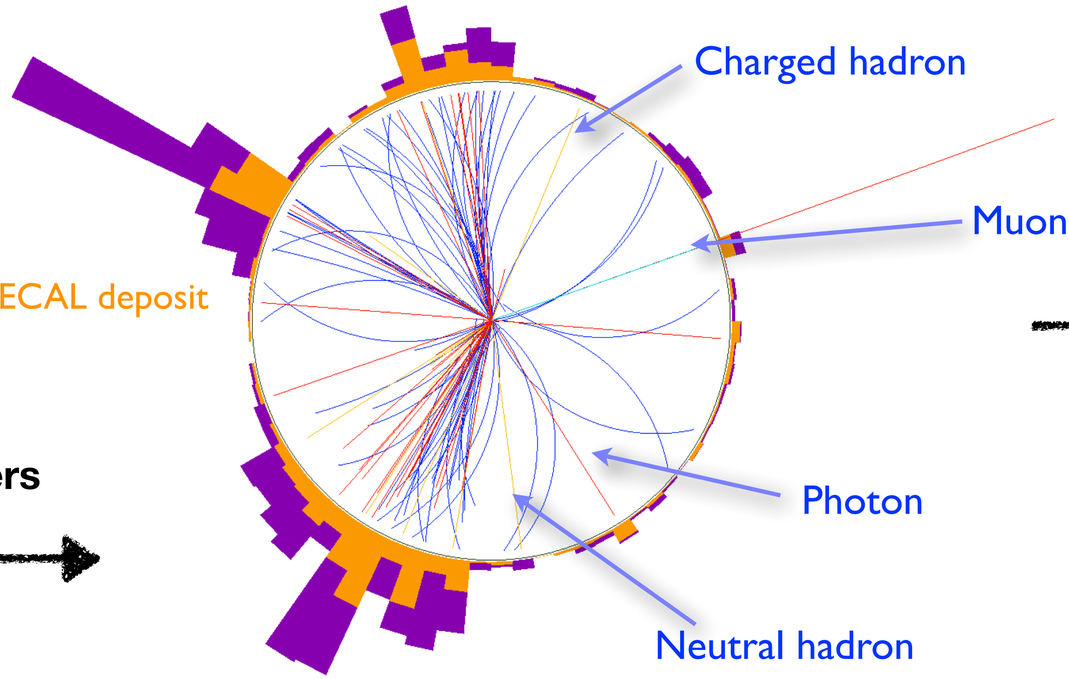
Charged particle tracks

Connecting the dots



Energy clusters

Find the clusters

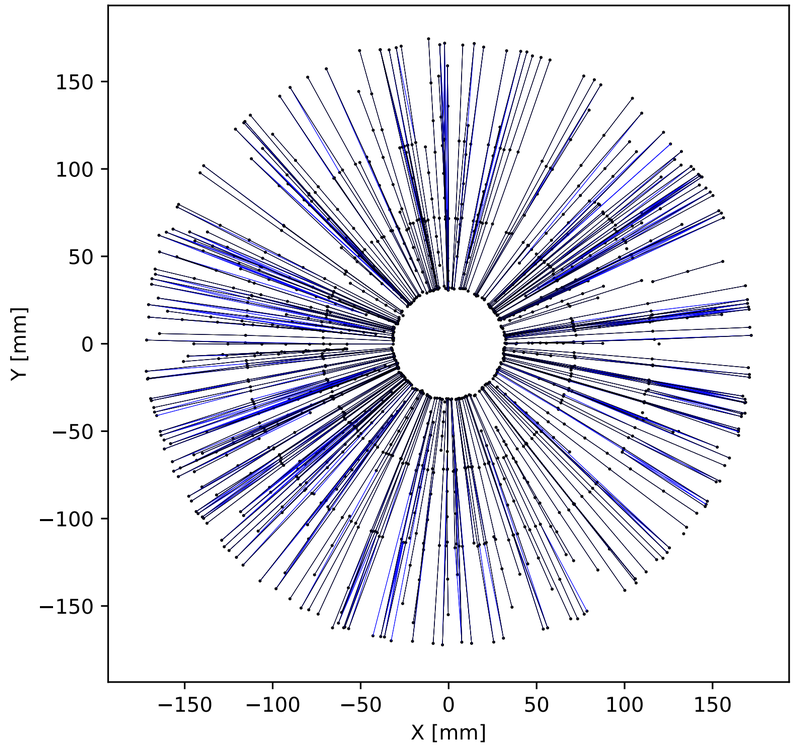


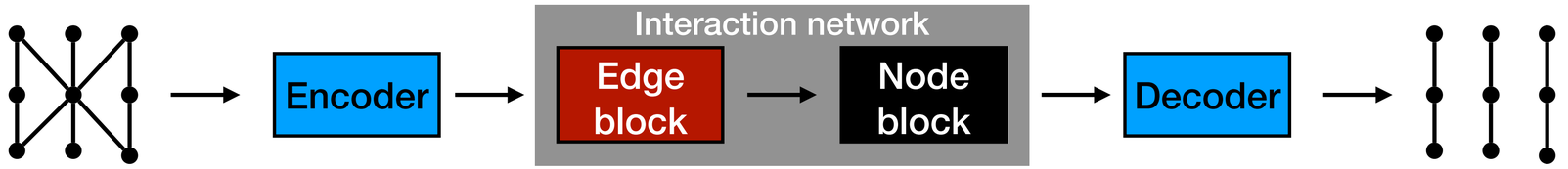
Particle Flow Reconstruction

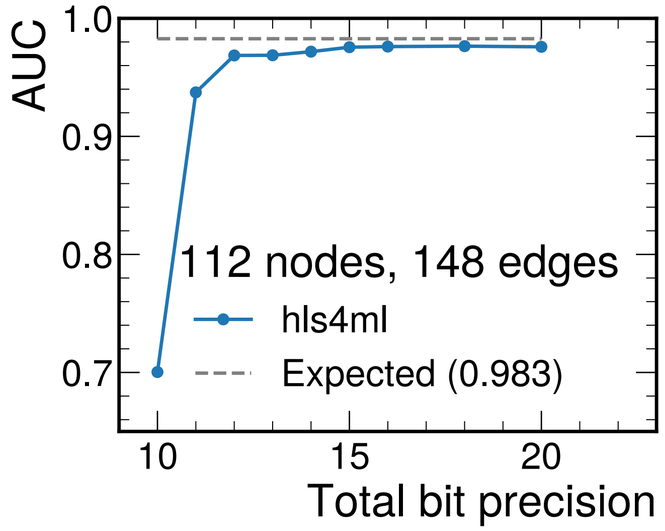
Higgs?
Top Quark Pair?
W boson?
Z boson?
Multiple boson?
.
.
.
.
New Physics?

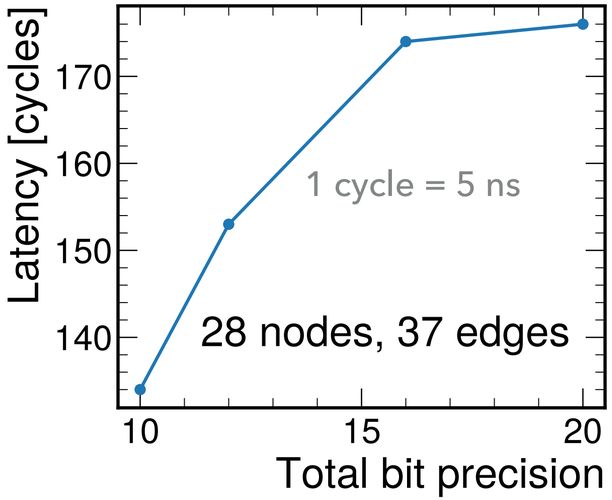
Analysis Reconstruction

Speed up tracking with GNN









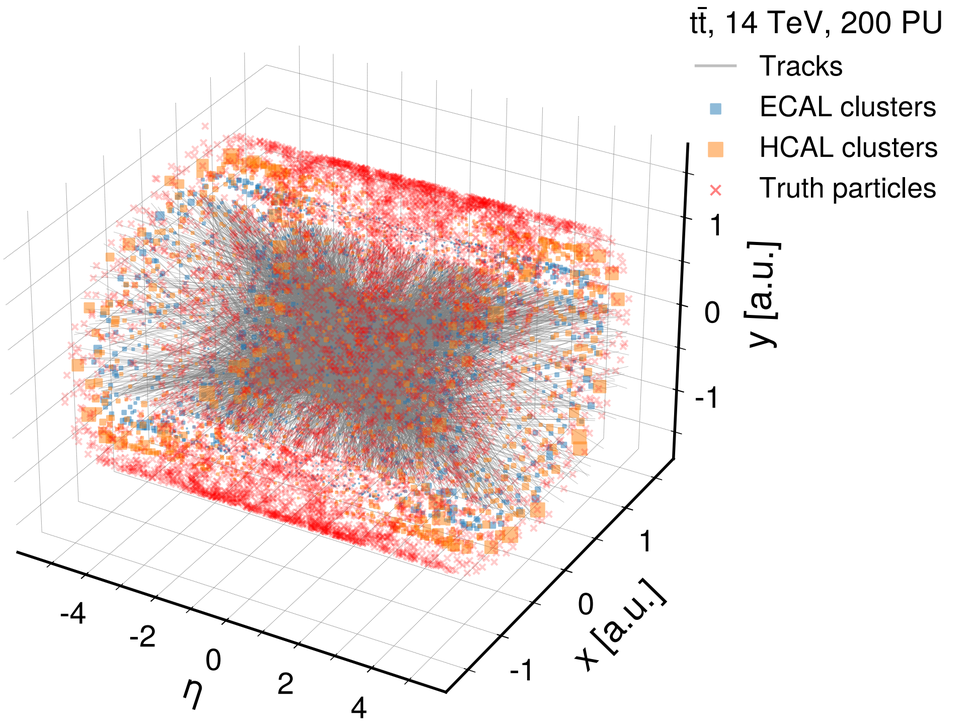
- Traditional Kalman Filter: computational expensive, serial operations
  - Efficient tracking with ML.
- GNN: Identifying edges compatible with real particle trajectories
- Implemented a GNN on an FPGA for particle tracking on small sectors
  - Found it can complete a tracking task in < 1 μs.

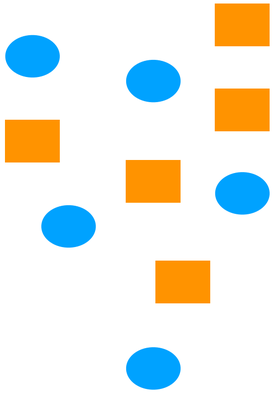
Particle Flow with GNN

- Input set (or point could or heterogeneous graph) of tracks,
- Target set of truth particles $Y = \{y_i\}$
- Goal: construct a mapping $\mathcal{U}(X) = Y' \sim Y$ that minimizes some distance

$$||Y - Y'|| \equiv \sum_{i \in \text{event}} L(y_i, y_i'),$$

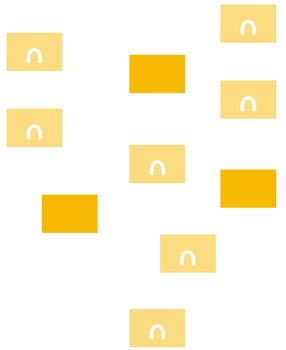
$$L(y_i, y_i') \equiv \text{CLS}(c_i, c_i') + \alpha \text{REG}(p_i, p_i'),$$





Input set $X = \{x_i\}$

$\{\text{track, cluster}\}$

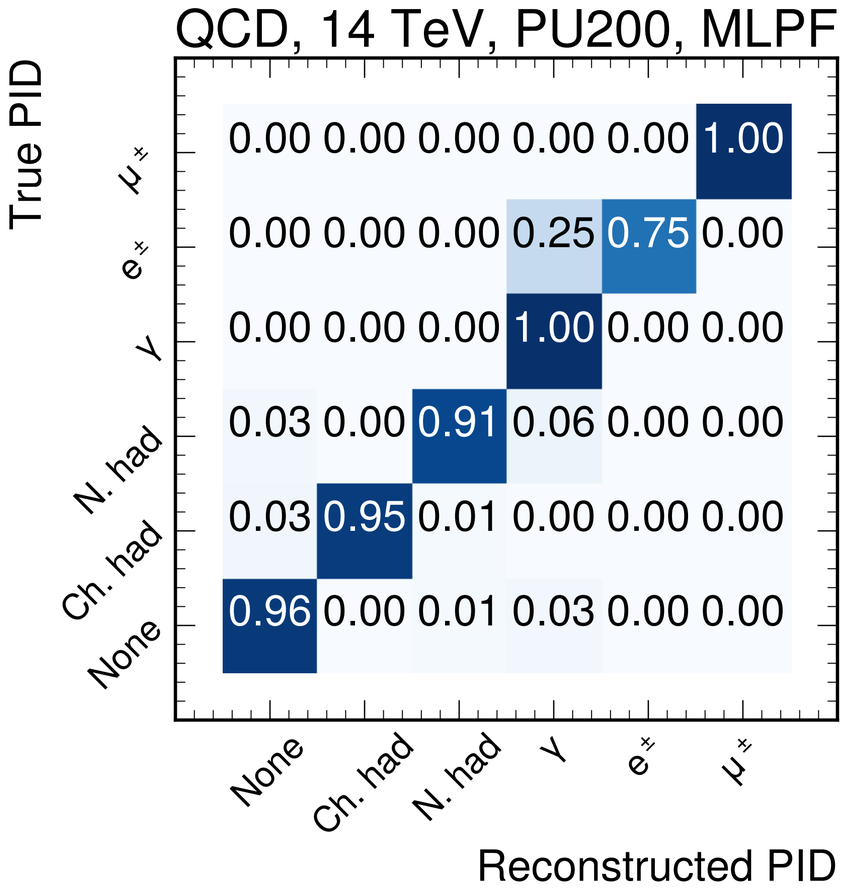


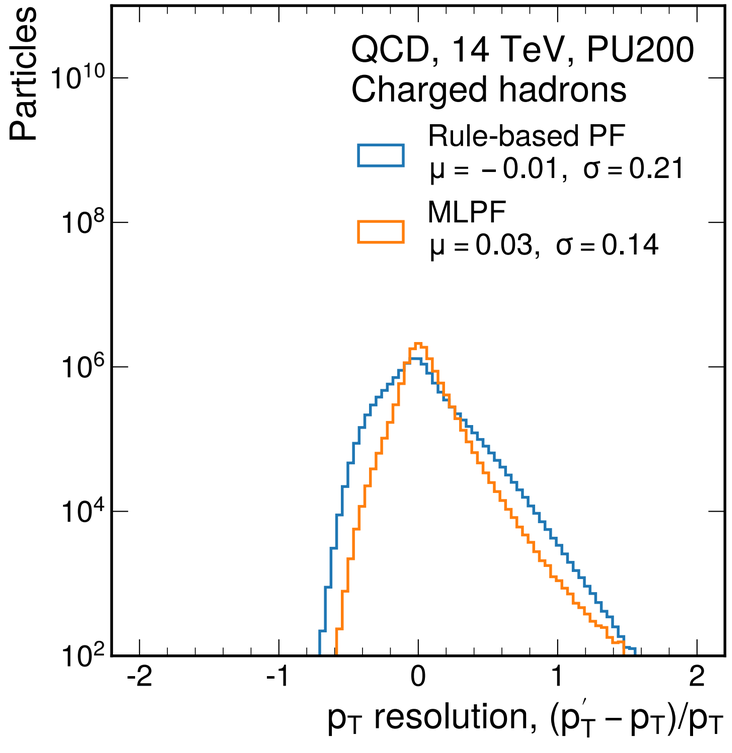
Target set $Y = \{y_i\}$

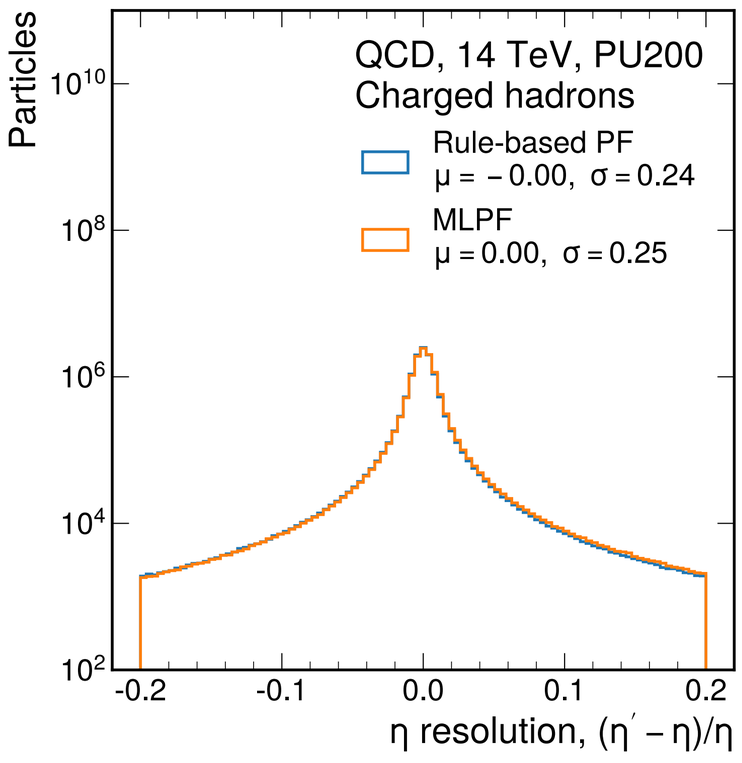
$\{\text{charged hadron, neutral hadron, } \gamma, e^{\pm}, \mu^{\pm}\}$

Classification/regression performance

arXiv:2101.08578

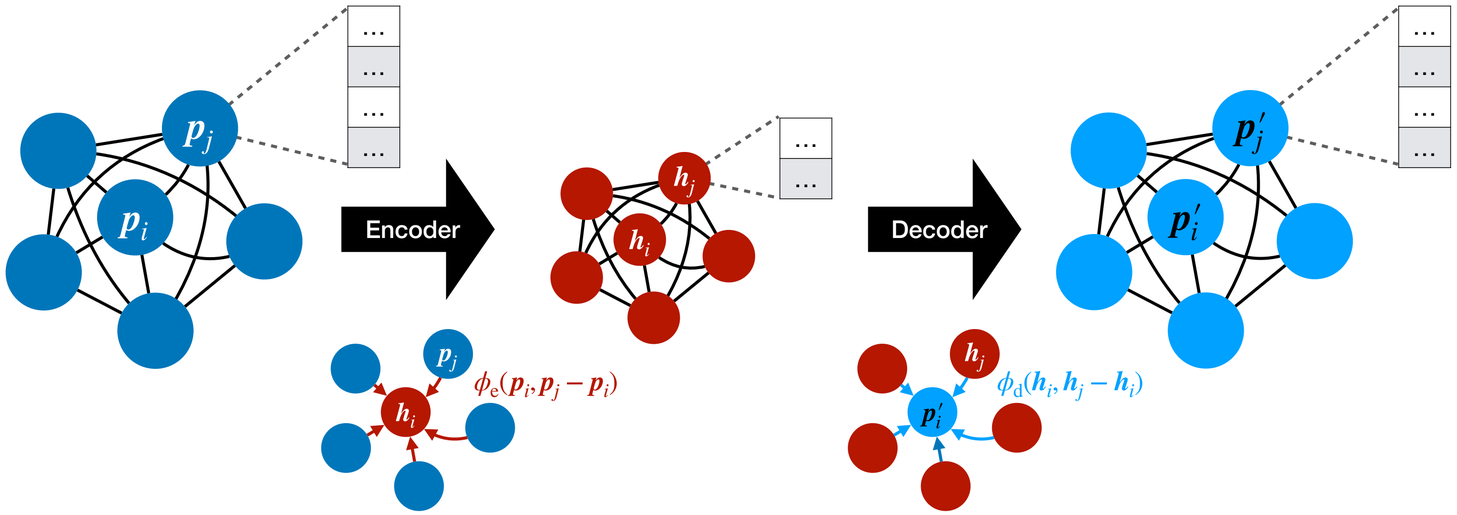




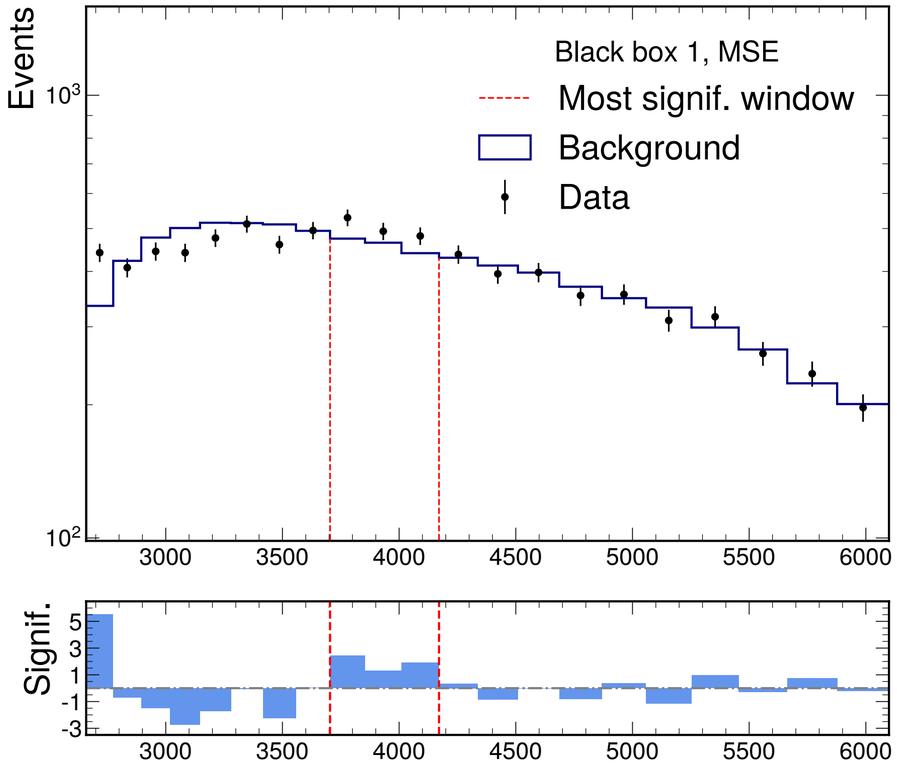


- MLPF same or better than rule-based PF algorithm

Model-agnostic searches



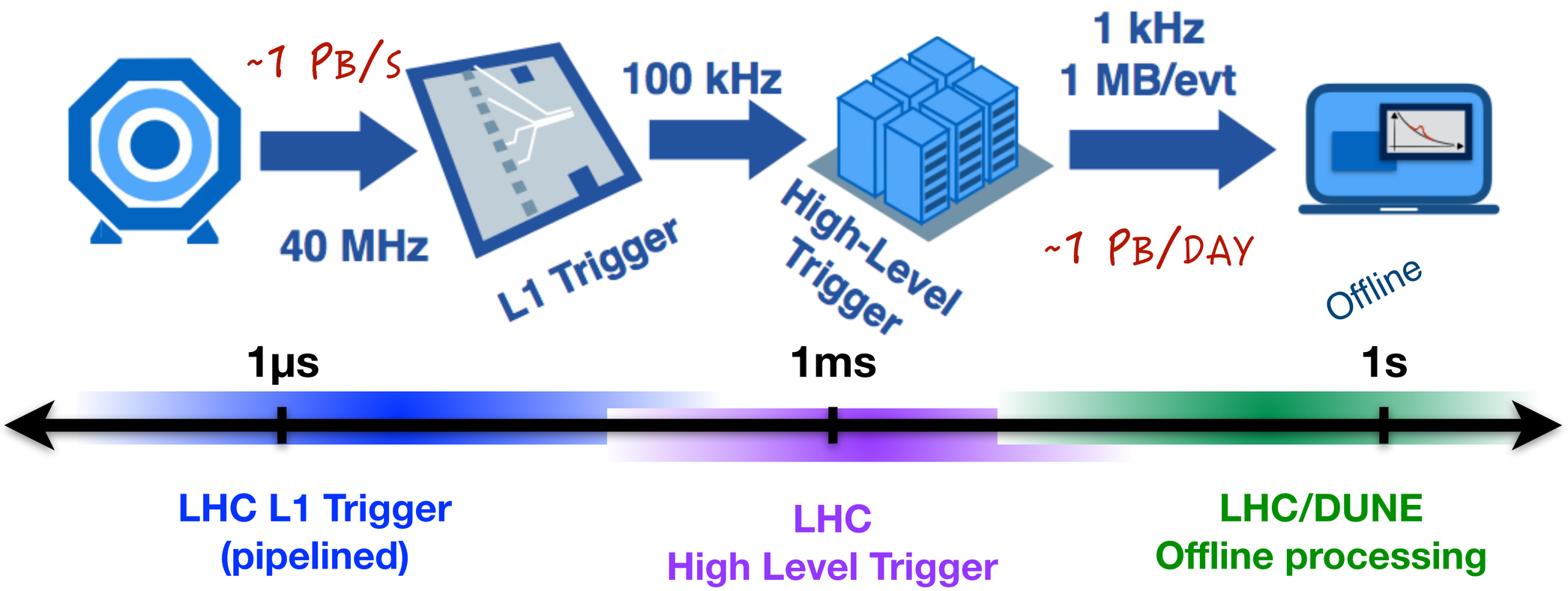
arXiv:2101.08320



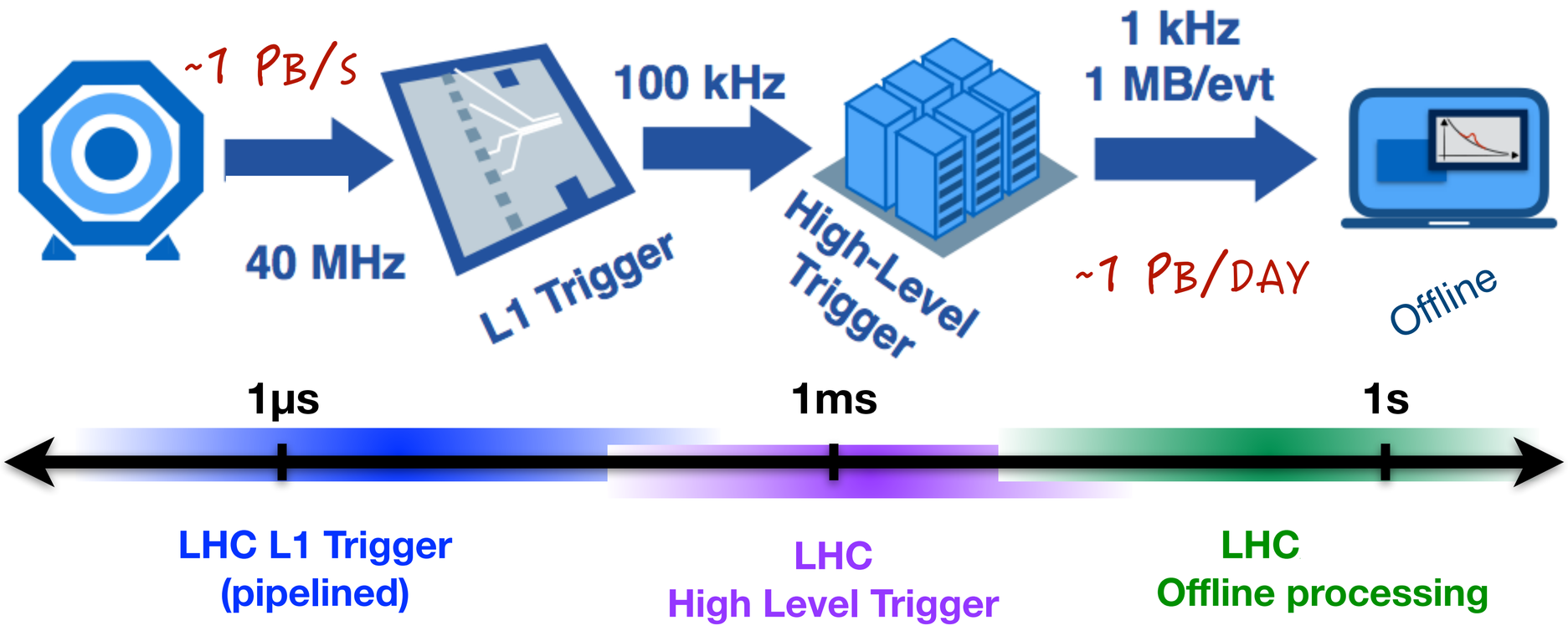
- **Autoencoders** compress data to a smaller representation and reconstruct it
  - Apply it to particles (E, ***p***): train autoencoder on known SM data
- Found to be effective on a mock dataset with a signal injected at 3.8 TeV

The reality:
Have to be fast enough

From Collisions to Discoveries



Accelerate ML landscape



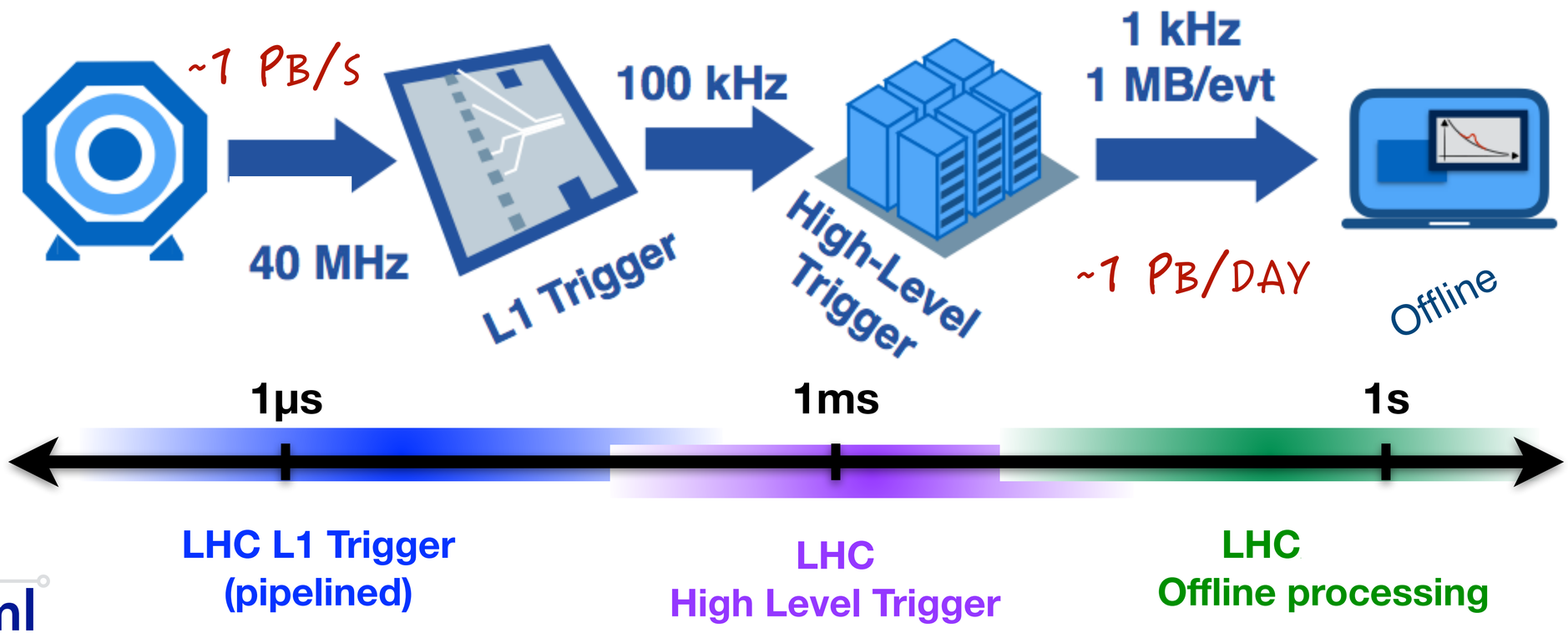
FPGAs/ASICs - high bandwidth low latency specialized compute hardware

"Targeted systems"

CPUs/increasing heterogeneous

"Heterogeneous systems"

Products developed in Fast ML



hls 4 ml

High Level Synthesis 4 Machine Learning

Applications in CMS: muon pT in endcap L1T, VBF tagger in L1T, BSM AE in L1T, ASIC data compression, AE in HGCAL, b-tagging/MET in L1T

ML-as-a-service (SONIC):

Cost-effective, flexible, scalable, reduced software stack maintenance

Applications in CMS data processing workflows (ML or CUDA)

Summary

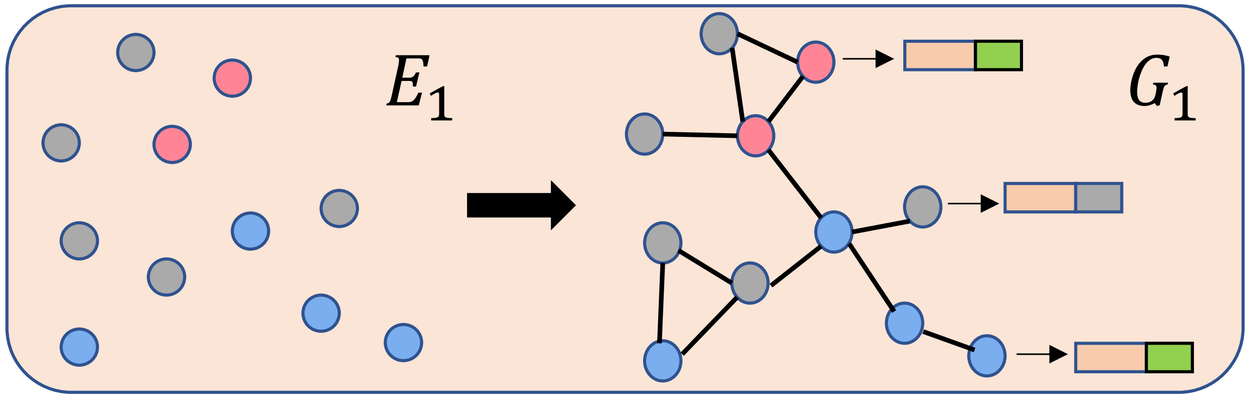
Machine learning offers opportunities to significantly boost the discovery potential at the LHC.

- Need accelerated machine learning inference in online & offline processing.
  - Multidisciplinary teams to realize optimal ML on targeted (e.g.L1 trigger) or heterogeneous systems (e.g. CMS HLT & offline).
  - User-friendly prototype tools for domain experts to streamline ideas to implementation.
  - Discussions focused not limited to reconstructions. e.g. fast simulation, experiment control/ real-time calibration..etc

Looking forward to continuing with tackling these challenges in A3D3 community

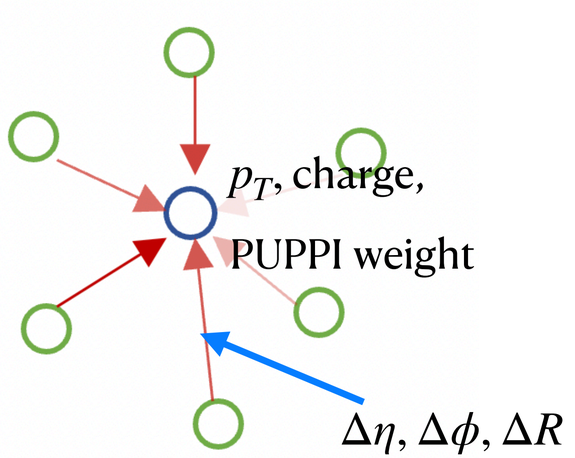
  - Cross-domain synergies in ML mathematical formulations
  - Advanced optimal ML methods in targeted&heterogeneous systems

Semi-supervised Graph Puppi

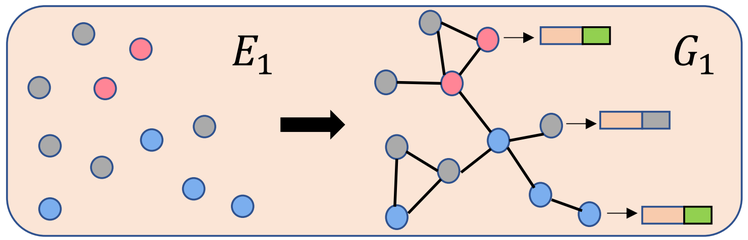


Charged LV particles
Charged PU particles
Neutral particles

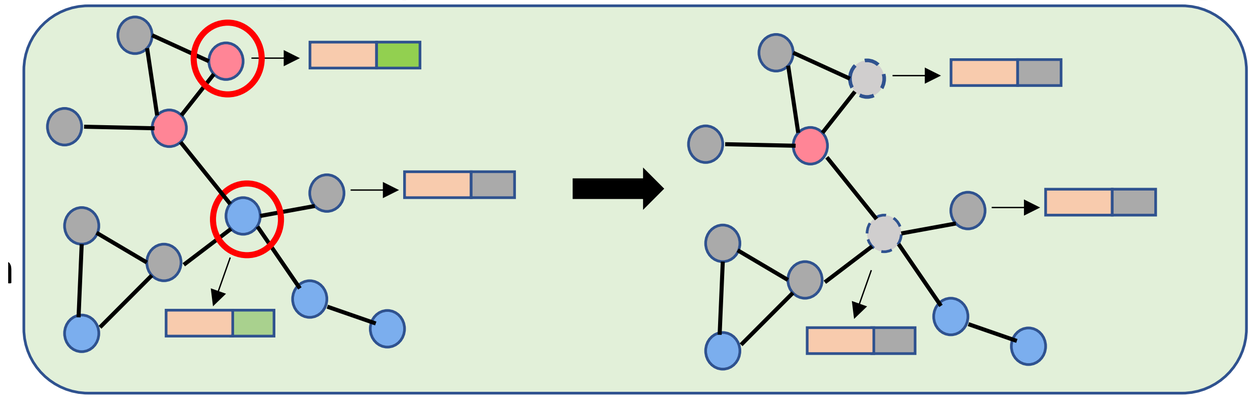
Common feature domain
Charged-specific feature domain
Neutral-specific feature domain

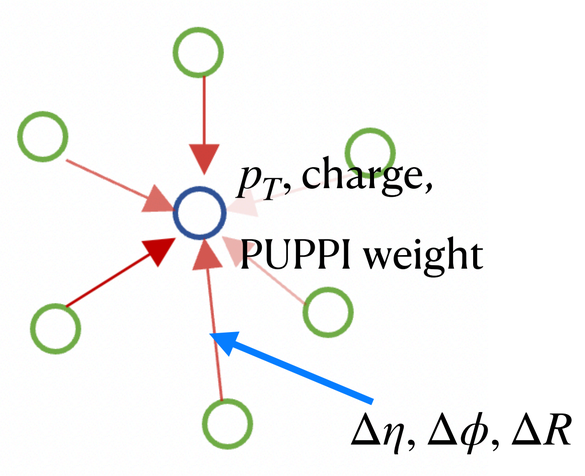


Semi-supervised Graph Puppi



(b). Randomly select charged LV/PU particles, and mask charged-specific feature domain for training



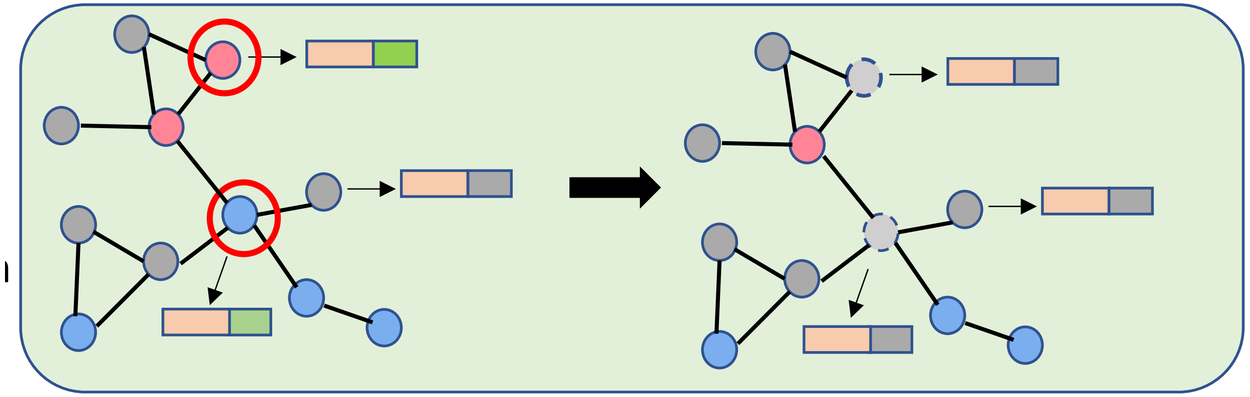


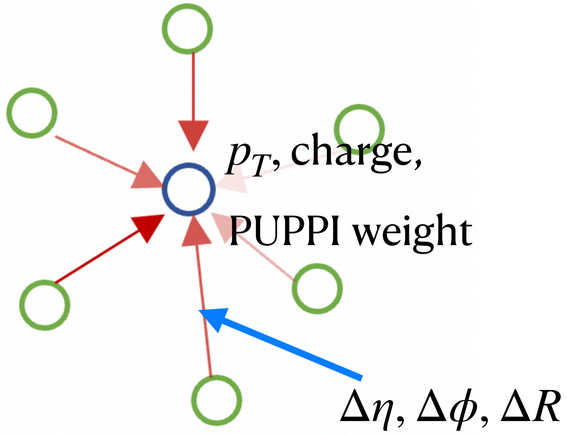
Charged LV particles
Charged PU particles
Neutral particles

Common feature domain
Charged-specific feature domain
Neutral-specific feature domain

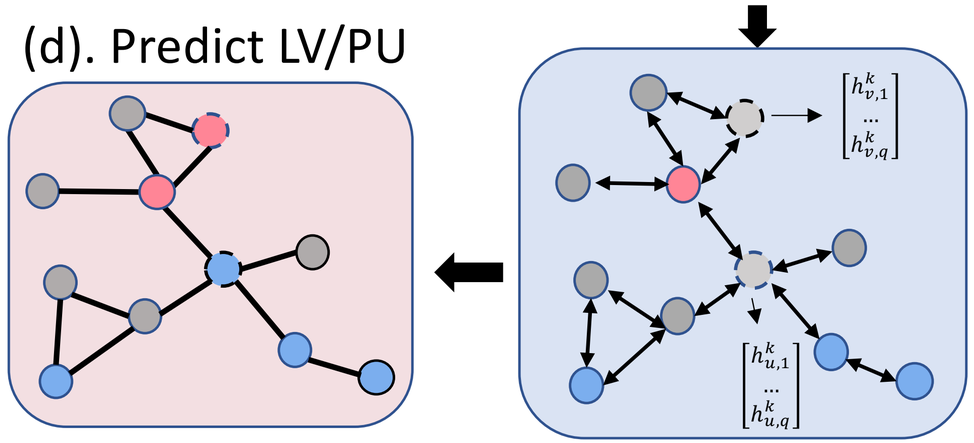
Semi-supervised Graph Puppi

(b). Randomly select charged LV/PU particles, and mask charged-specific feature domain for training



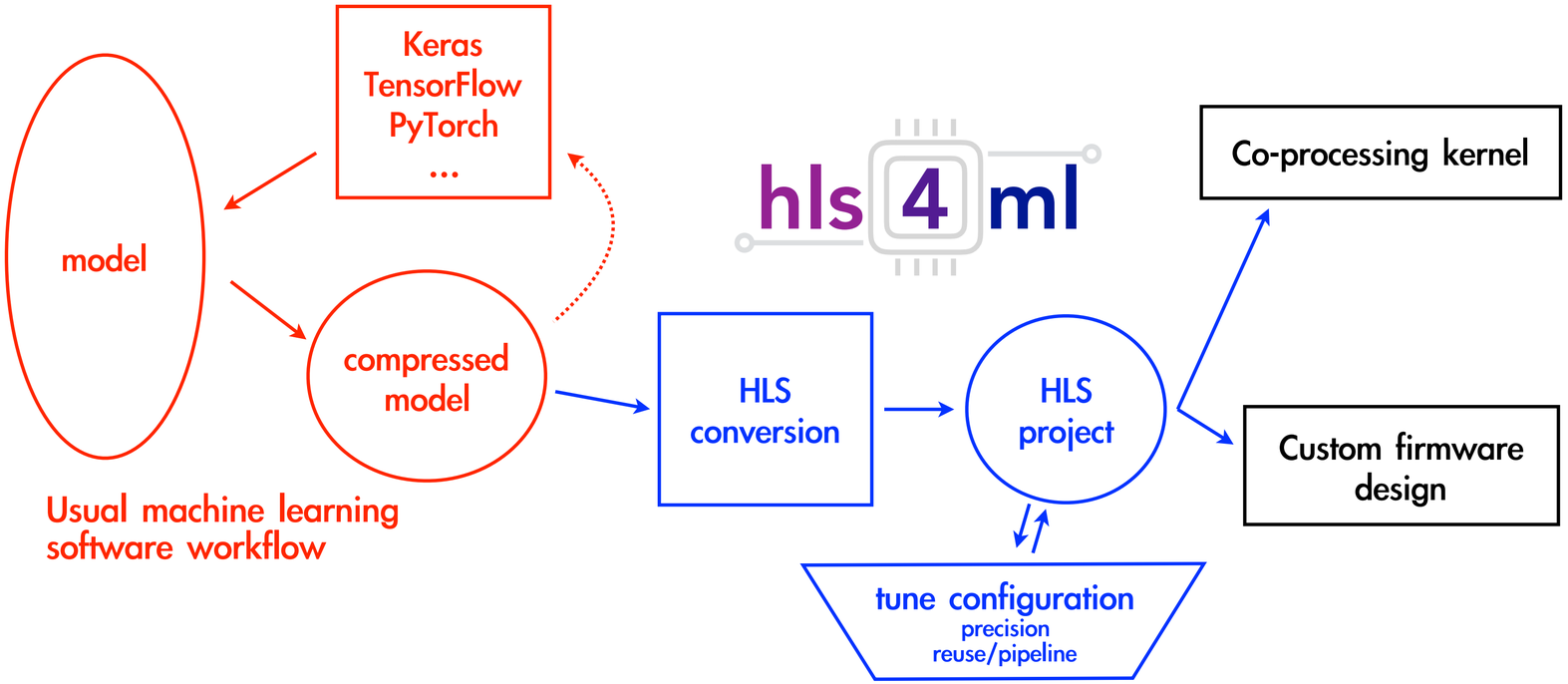


(d). Predict LV/PU



- Algorithm outperforms Puppi, comparable to fully supervised method. Can be adapted to different pile up conditions. No need for tuning as the particle itself is represented as a node.
- Presented at BOOST 2021, Short version of the paper submitted to NeurIPS 2021 AI for Science Workshop. Long version paper targeting PRD in preparation.
- Next: Apply to CMS simulation & data. Neutral particle vertex association in for the forward region.

High-Level Synthesis 4 Machine learning



HLS - High Level Synthesis - compiler for C, C++, SystemC into FPGA IP cores

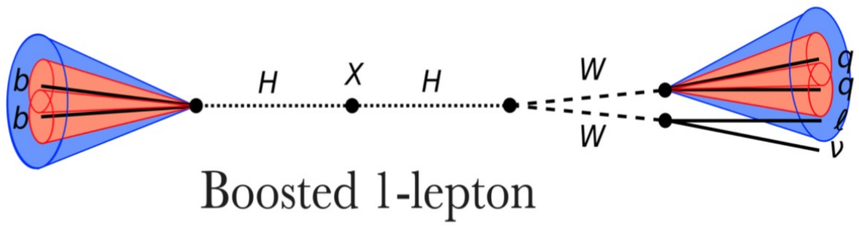
HLS 4 Machine learning :**Prototype** ML algorithms for FPGA WITHOUT Verilog/ VHDL: firmware in a few hours

Boosted Higgs Boson Tagging using ML

Boosted H→bb

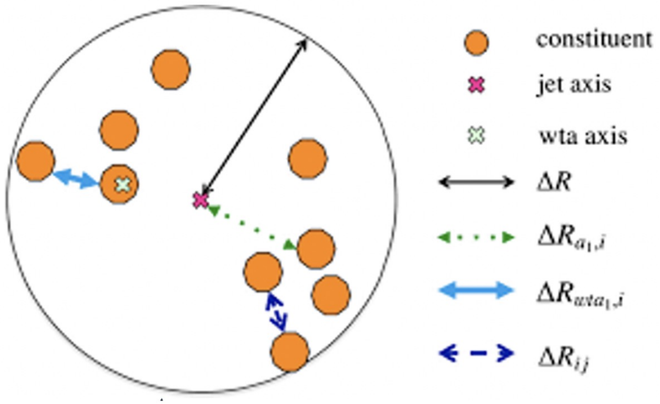
Boosted H→WW$^{(*)}$

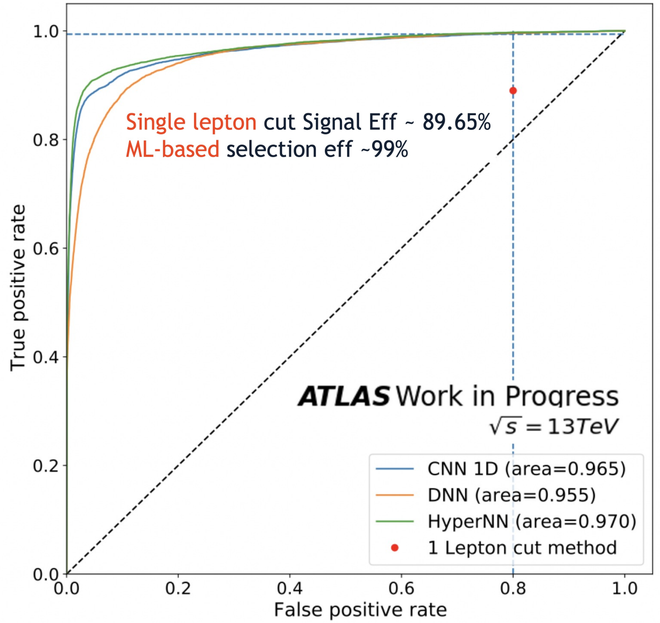
High p_T H→WW$^{(*)}$ semileptonic decays, e.g. from a heavy particle X→HH decay, are difficult to identify due to merging of lepton and jets



Boosted 1-lepton

Train Neural Network taggers on jet 4-vectors and jet constituents





As-a-service Computing Model

How many CPU can this GPU serve?

CPU-to-GPU ratio:

t_total_cpu - t_othercpu = t_ml

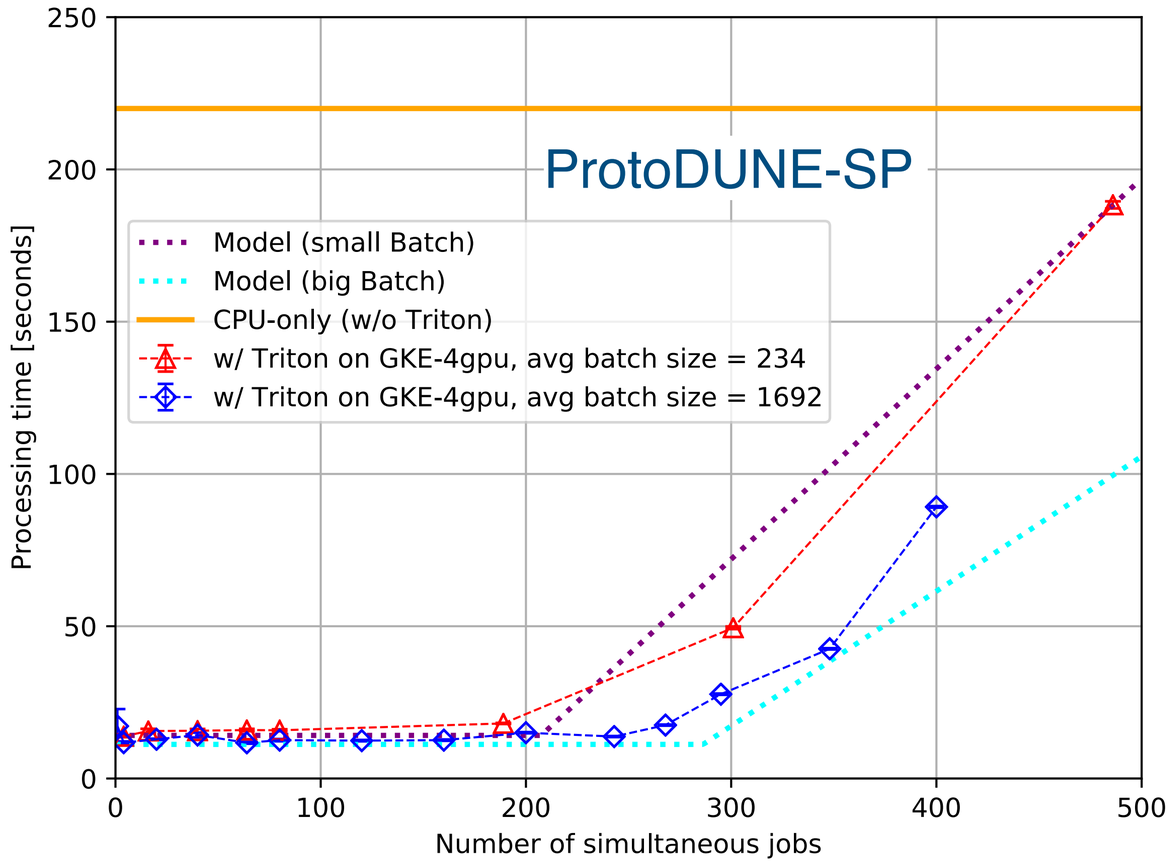
t_total_sonic - t_othercpu - t_sonic_cpu = t_transfer + t_scheduling + t_sonic_gpu

t_ml/ (t_sonic_cpu + t_transfer + t_scheduling + t_sonic_gpu) =

tsonic_cpu_part = t_sonic - t_transfer - t_scheduling

ratio = t_sonic_cpu_part/t_gpu

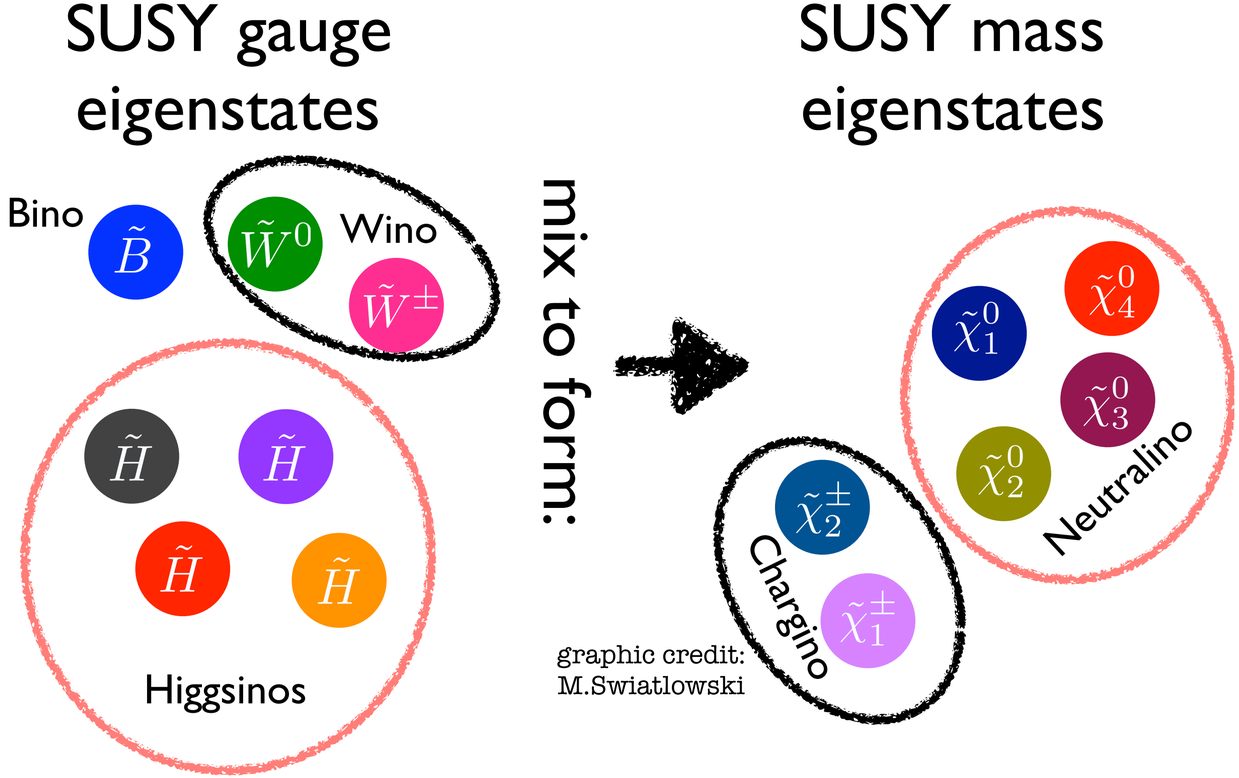
Passes this ratio, GPU saturates and average processing time increase.



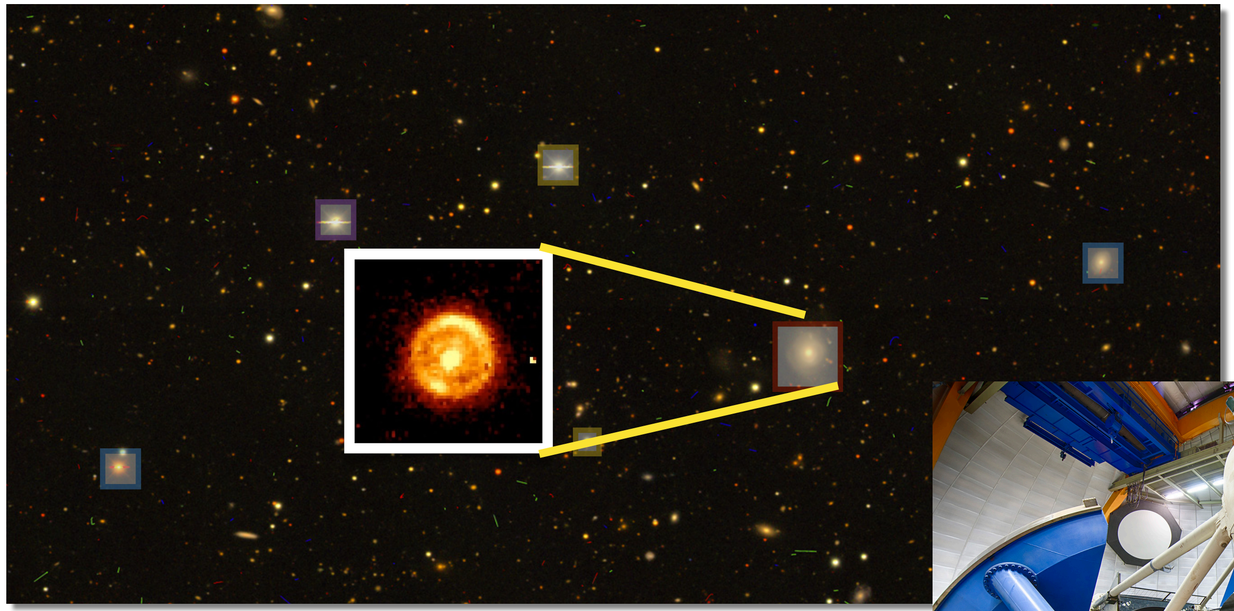
Electroweakinos 101

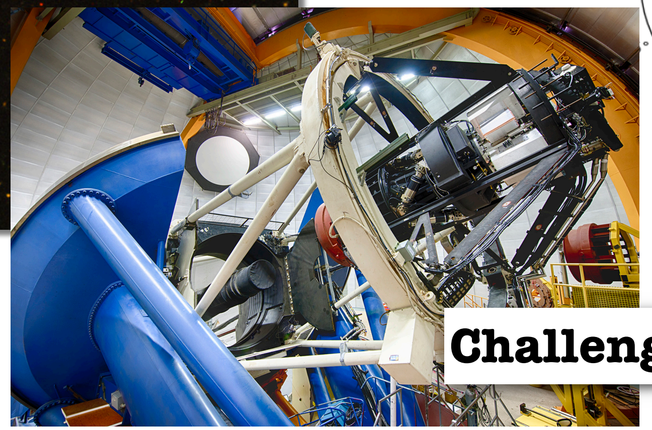
- SUSY partners of the SM electroweak sector:
  - U(1)-> Bino, SU(2)->Winos
  - Higgs-> Higgsinos
  - Leptons-> sleptons
- Could be light and accessible at the LHC

Important to search for them at the LHC!

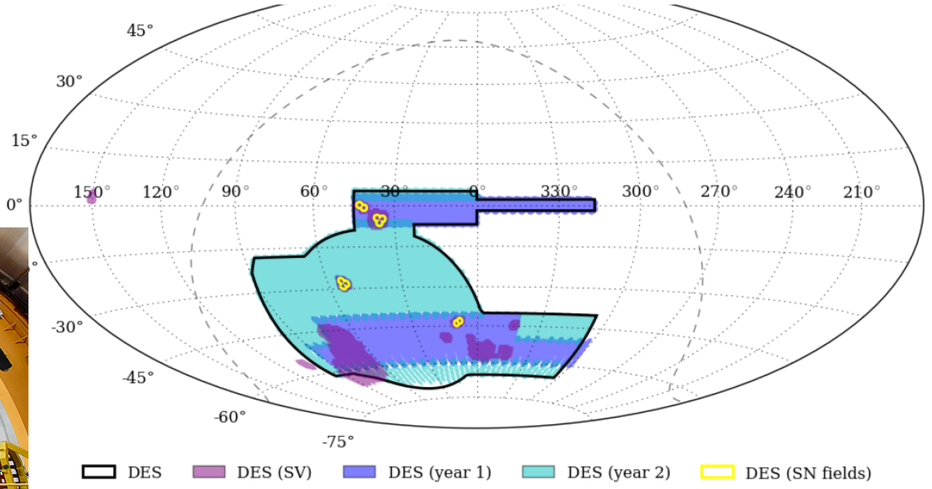


ML in the Sky: it's full of stars





- **Dark Energy Survey: Sky Footprint of Observations**



Challenge of scheduling on multiple time

Populations of objects show **dark matter, dark energy**

Region-based CNNs on heterogeneous compute devices

- LSST: 20 Tb / night
- 1 Billion transient alerts /night

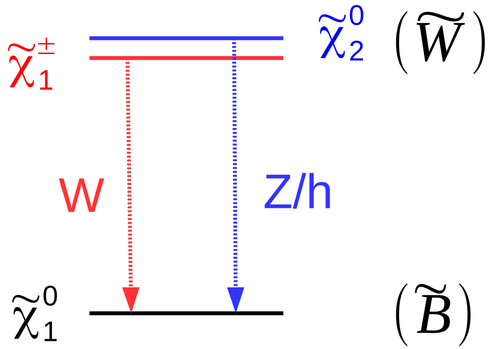
Long: competition between **faint galaxies**, **transient objects**

Short: Weather, annual modulation of sky positions

Smart telescopes: reinforcement learning for optimal scheduling and control

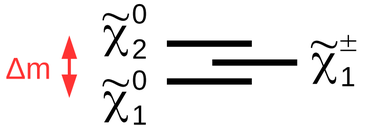
Some topics need HL-LHC dataset

- 3000 fb-1 data expected at the HL-LHC
- e.g. Higgsinos: Low cross section, challenging signatures
  - Δm ~ tens of GeV : Soft decay products
  - Δm ~ hundreds of MeV : Long-lived signatures



Wino-like
45 fb*

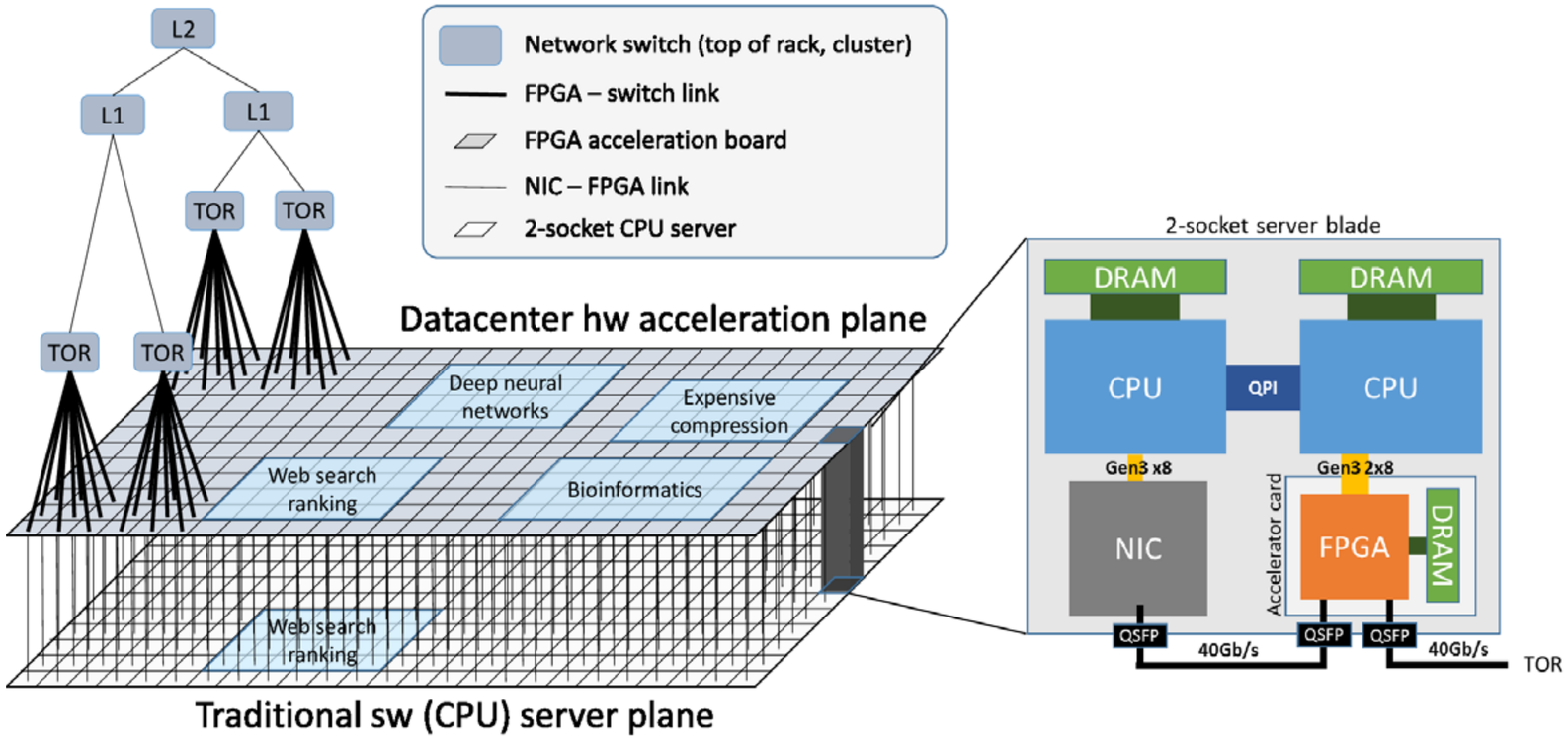
Δm~ hundreds of MeV to tens of GeV



Higgsino-like
11 fb*

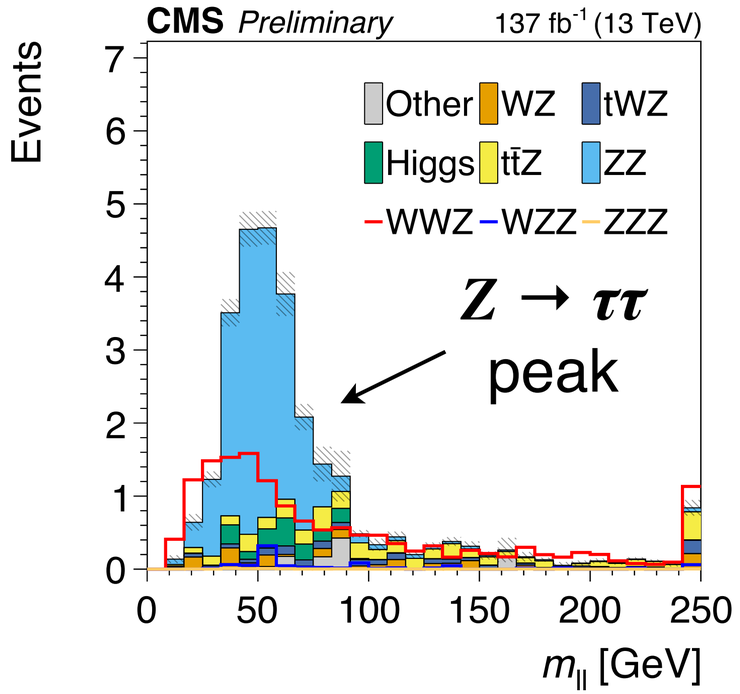
*Cross-sections for 500 GeV sparticles @ 13 TeV($\tilde{\chi}_2^0\,\tilde{\chi}_1^{\pm}$ only)

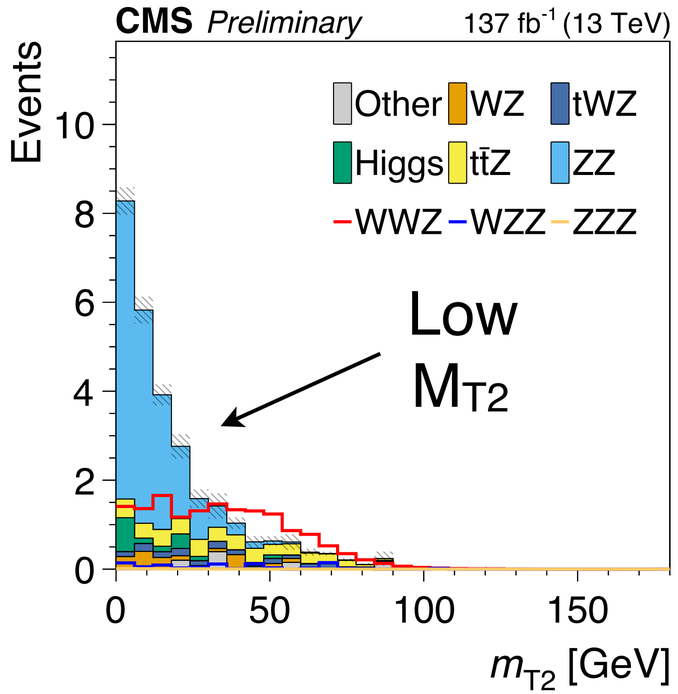
Microsoft Brainwave



- **Mature service** at scale (more than just a single co-processor)
- Multi-FPGA/CPU fabric accelerates *both* **computing** and **network**
- Models supported:
  - ResNet50, ResNet152, DenseNet121 ,VGGNet16…
  - Partially fixed neural network architecture. weights can be retuned.

WWZ: smaller rate but clean

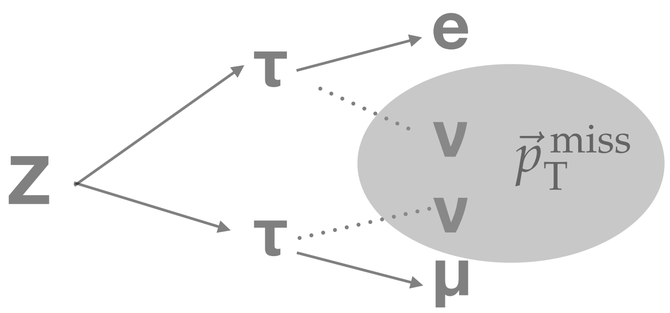




40% smaller than WWW , leptonic decays

4ℓ: Tag Z->$\ell\ell$, WW-> (eμ/ee/μμ)

- eμ shown as example: kinematic selections against ZZ



$$m_{\mathrm{T2}} = \min_{\vec{p}_{\mathrm{T}}^{\,\nu(1)} + \vec{p}_{\mathrm{T}}^{\,\nu(2)} = \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}} \left[\max \left(m_{\mathrm{T}}^{(1)}(\vec{p}_{\mathrm{T}}^{\,\nu(1)}, \vec{p}_{\mathrm{T}}^{\,\mathrm{e}}), m_{\mathrm{T}}^{(2)}(\vec{p}_{\mathrm{T}}^{\,\nu(2)}, \vec{p}_{\mathrm{T}}^{\,\mu}) \right) \right]$$

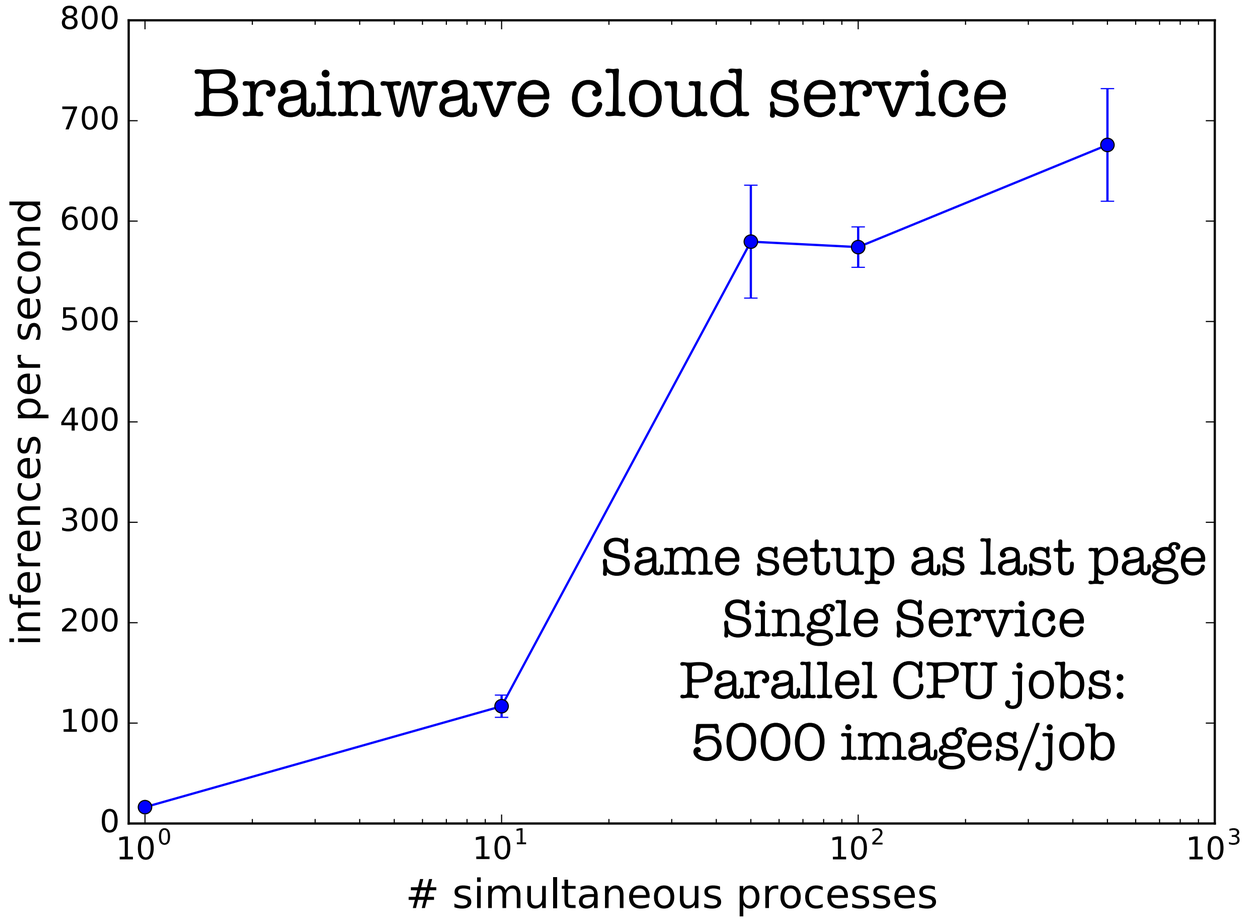
EDGE DATA BOX @ Fermilab

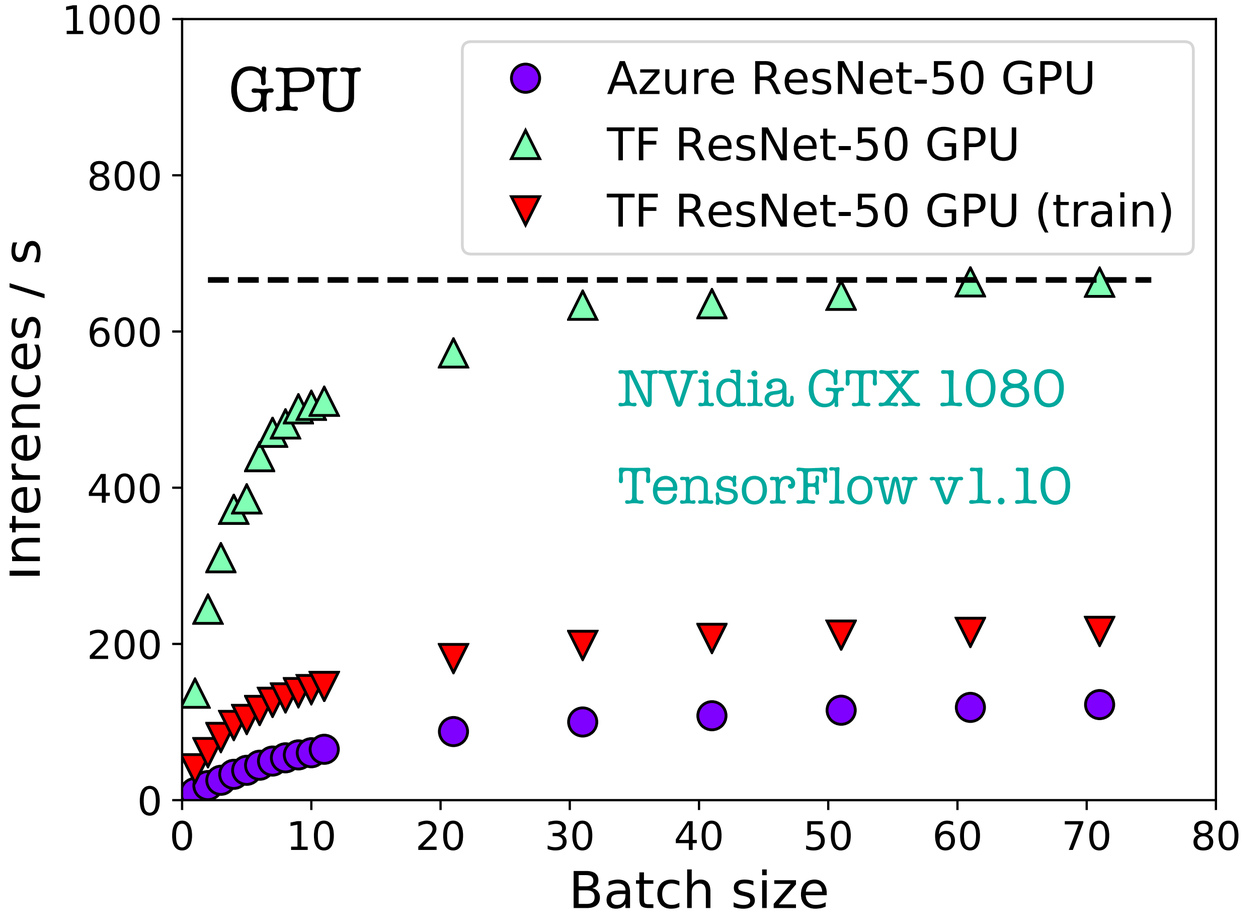
Edge data box at Feymann computing center

Docker container on server (PCIe connection):	14 ± 25 ms
Fermilab computing cluster:	20 ± 30 ms
Local laptop:	68 ± 27 ms
CERN (Geneva):	168 ± 62 ms

- Gain experience in deploying co-processors in local clusters with cloud native tools: docker image, kubernetes
- Benchmark latency and scaling performance, compare with previous studies
- Can be used for neutrino and cosmology experiments as of ~today—> next slide

data throughout compared to





Comparable max data throughout: 600-700 images/sec

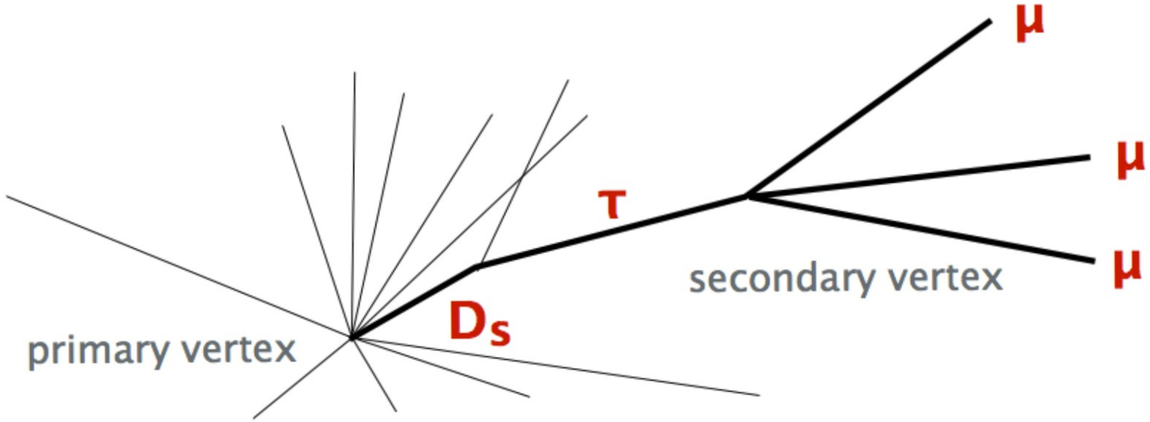
Beyond fully connected neural networks

Neutrinos oscillate:

Lepton number not conserved

At the LHC

What about in charged leptons? τ->3μ



SM: 10^{-25}

Currently:< 10^{-8}

Pixel detector of CMS at the LHC

Need to cope with more challenging LHC environment in Run 2 & Run 3 (300 fb^{-1}) until HL-LHC upgrade (2023).

Module designed to reduce dynamic inefficiency

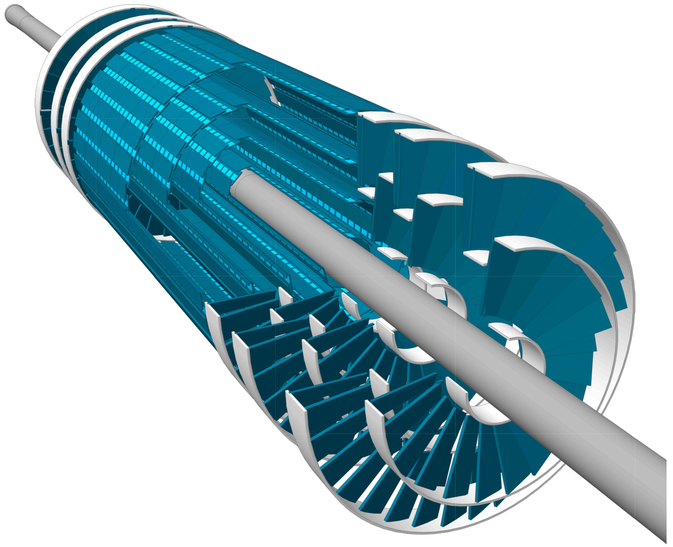
Digital readout chip (ROC). Faster readout.

Geometry design: ensure tracking and vertex quality

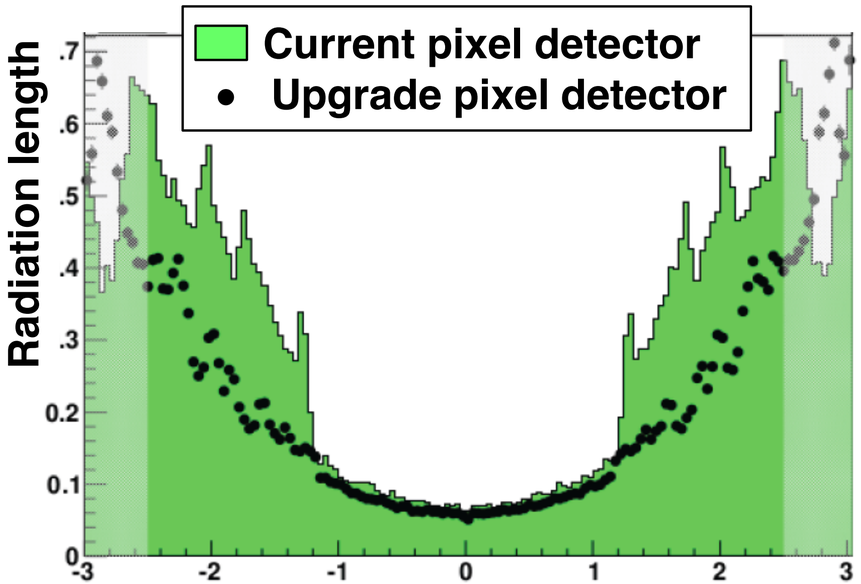
Added layers, channels doubled

Services: reduce material budget

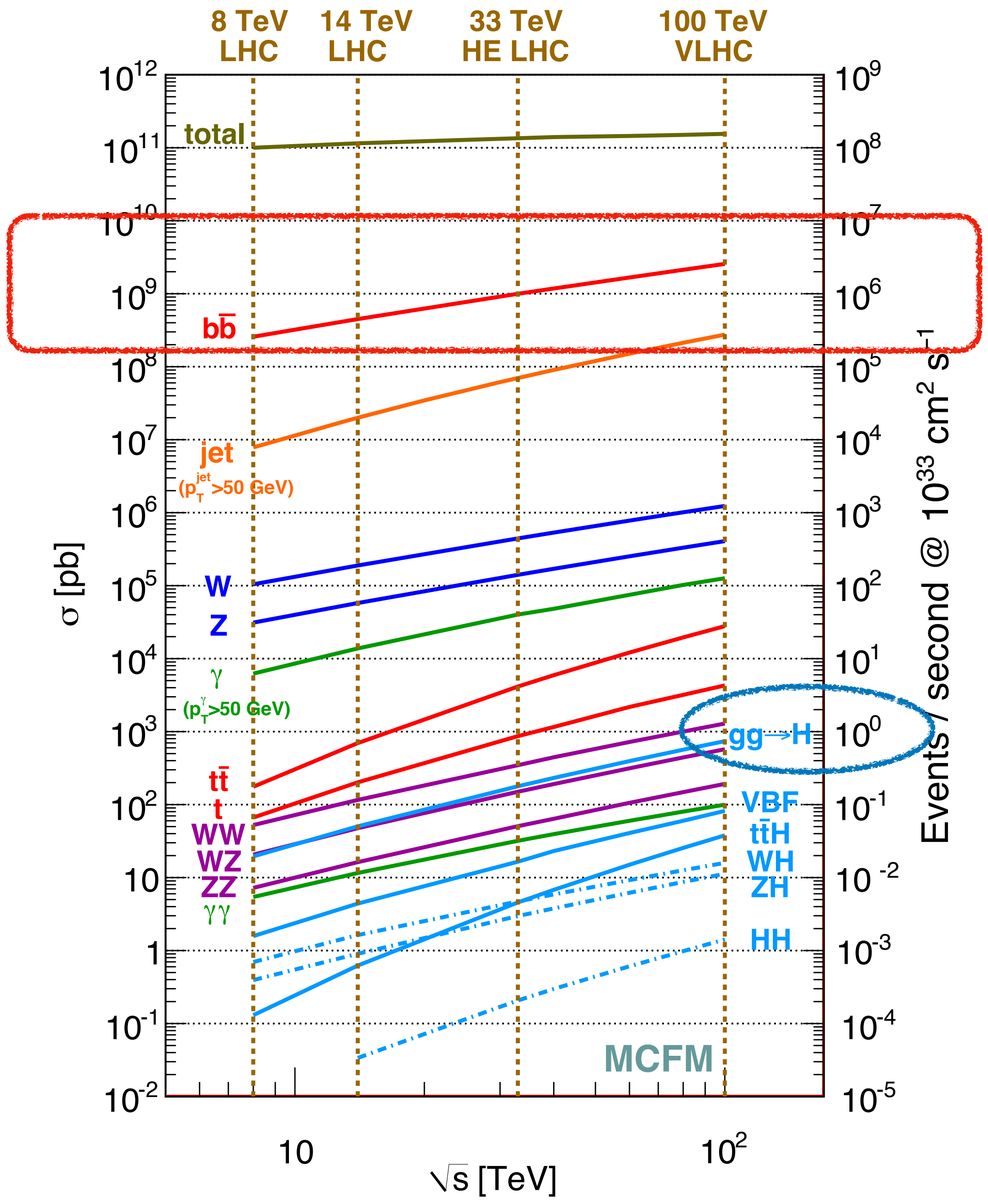
CO2 cooling, DCDC powering, Service electronics out of tracker volume.

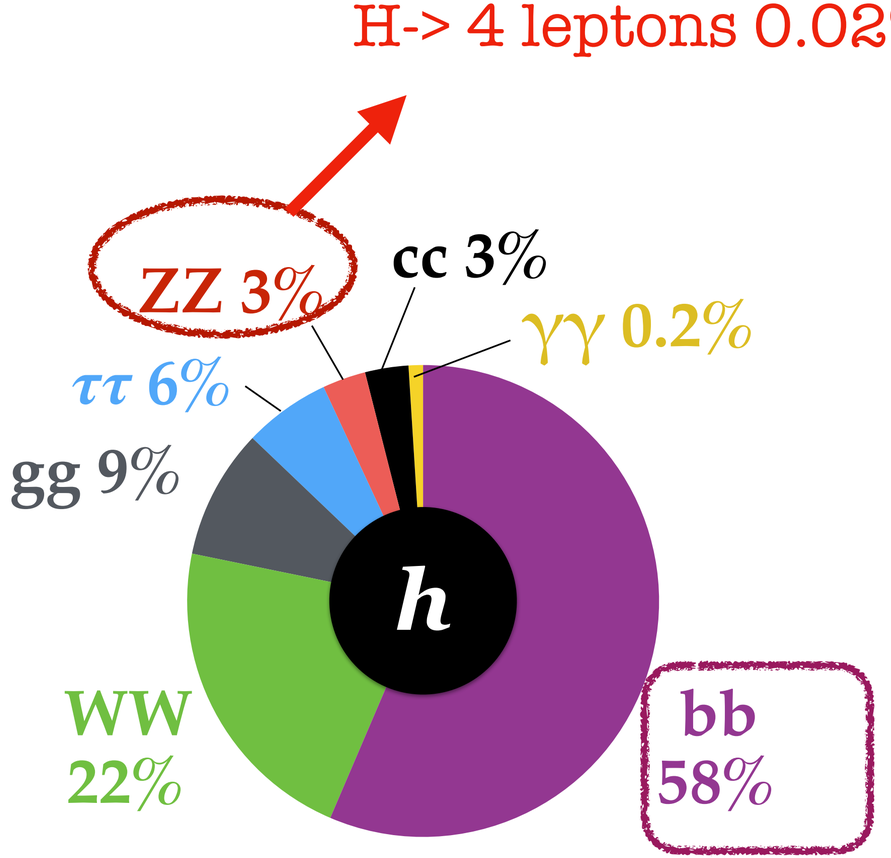


Material budget



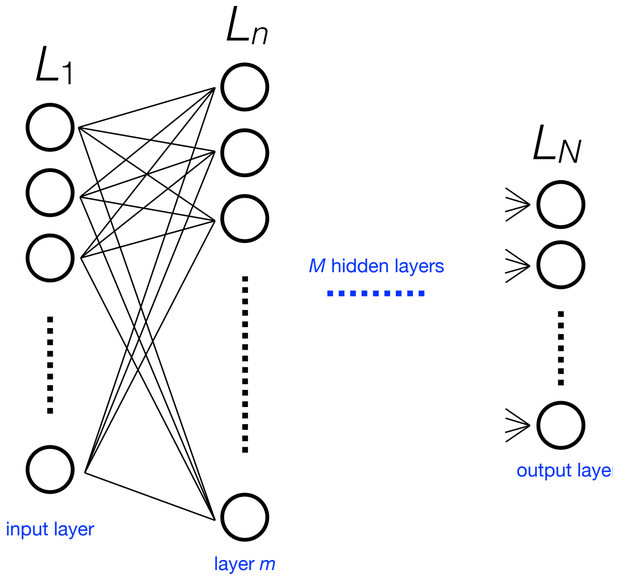
Signal vs background





Higgs boson discovery: decay modes of lower backgrounds (WW/ZZ/γγ).

NN on FPGAs



FPGA diagram

$$\mathbf{x}_n = g_n(\mathbf{W}_{n,n-1}\mathbf{x}_{n-1} + \mathbf{b}_n)$$

Activation functions

Precomputed, and stored in BRAMs

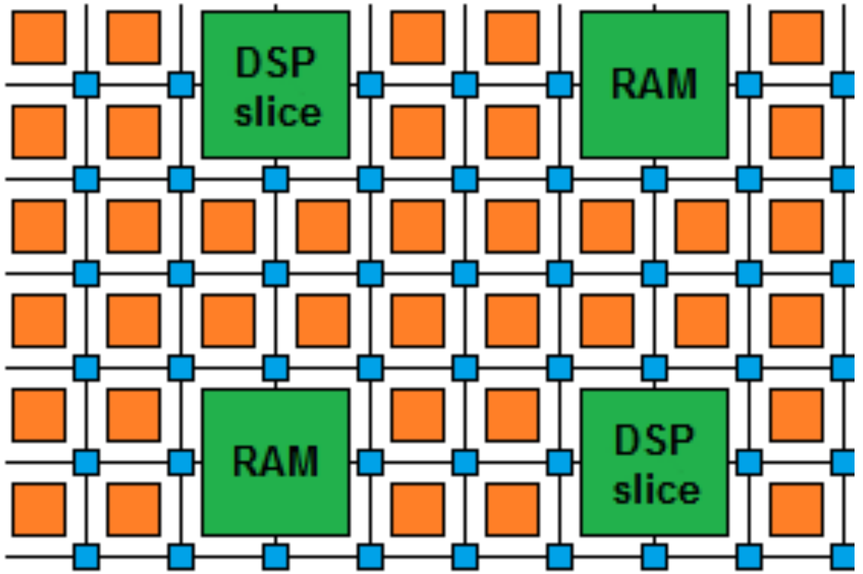
Multiplications

Digital Signal Processing

DSPs

Addition

Logic cells



Natural fit for FPGAs… limited resources

$$\mathbf{x}_n = g_n(\mathbf{W}_{n,n-1}\mathbf{x}_{n-1} + \mathbf{b}_n)$$

Activation functions

Precomputed, and stored in BRAMs

Multiplications

DSPs

Addition

Logic cells

$$N_{\text{multiplications}} = \sum_{n=2}^{N} L_{n-1} \times L_n$$

Small network: thousands of connections

Limitation: Number of DSPs

Virtex Ultrascale+ VU9P

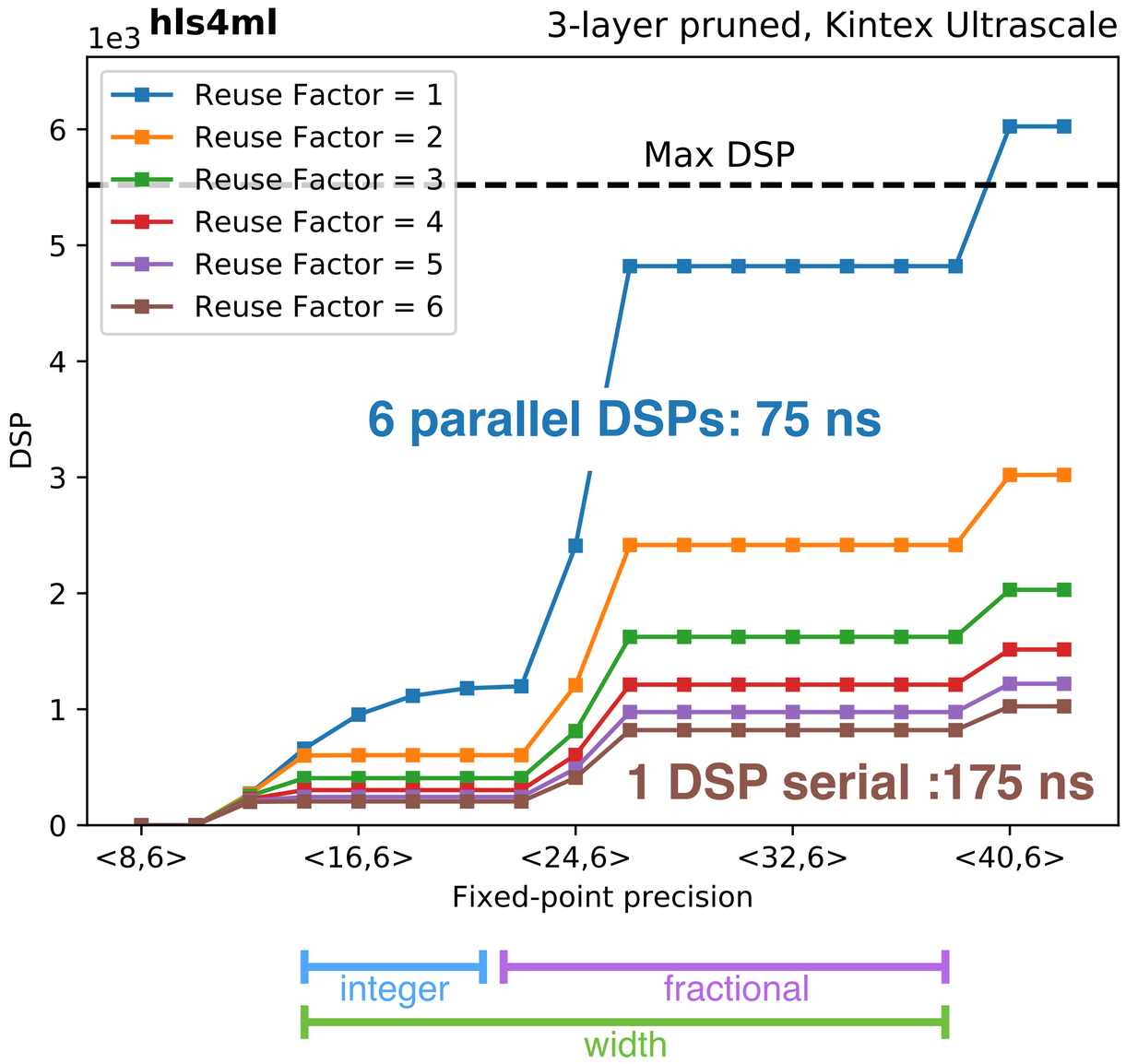
6800 DSPs

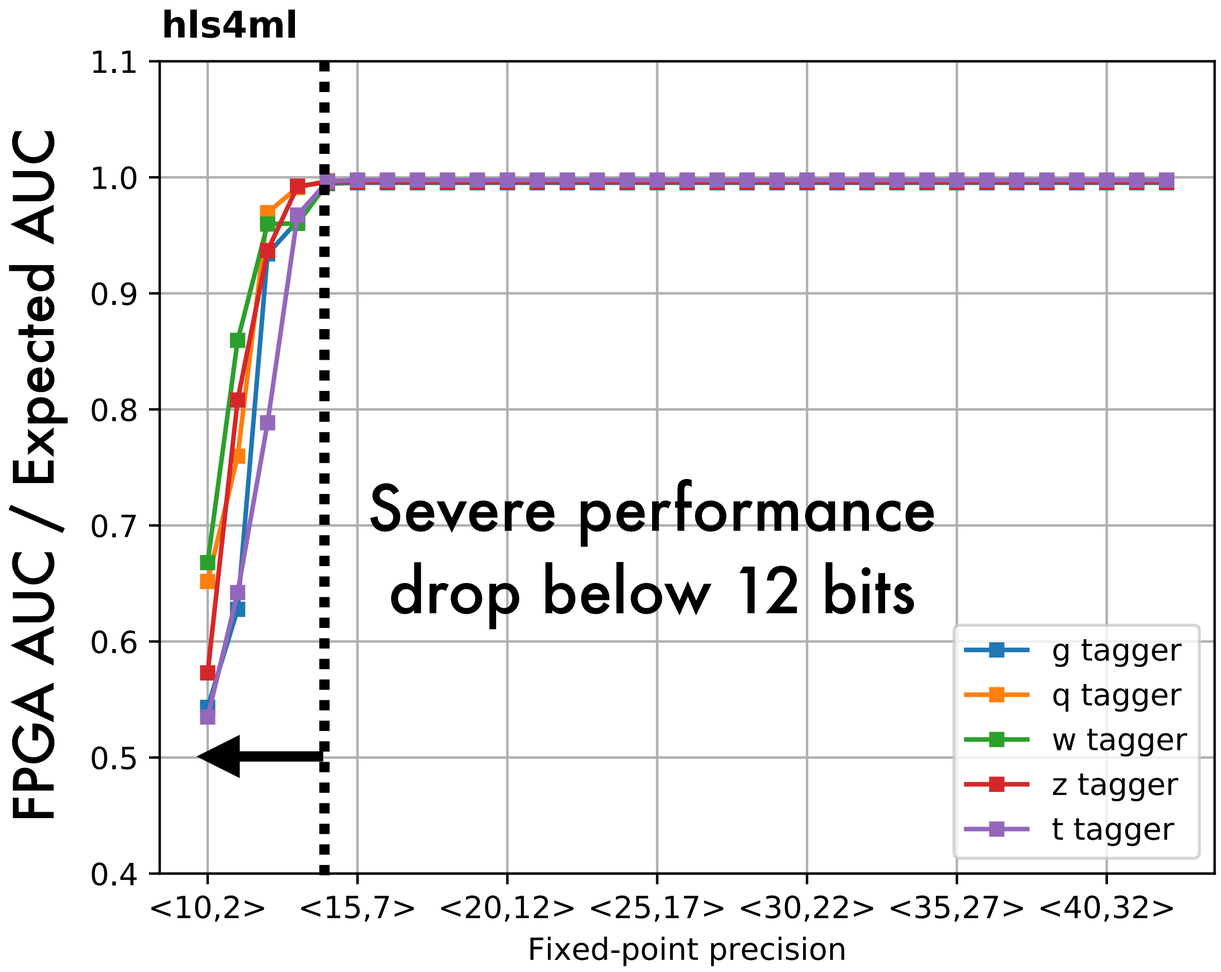
1M LUTs

2M FFs

75 Mb BRAM

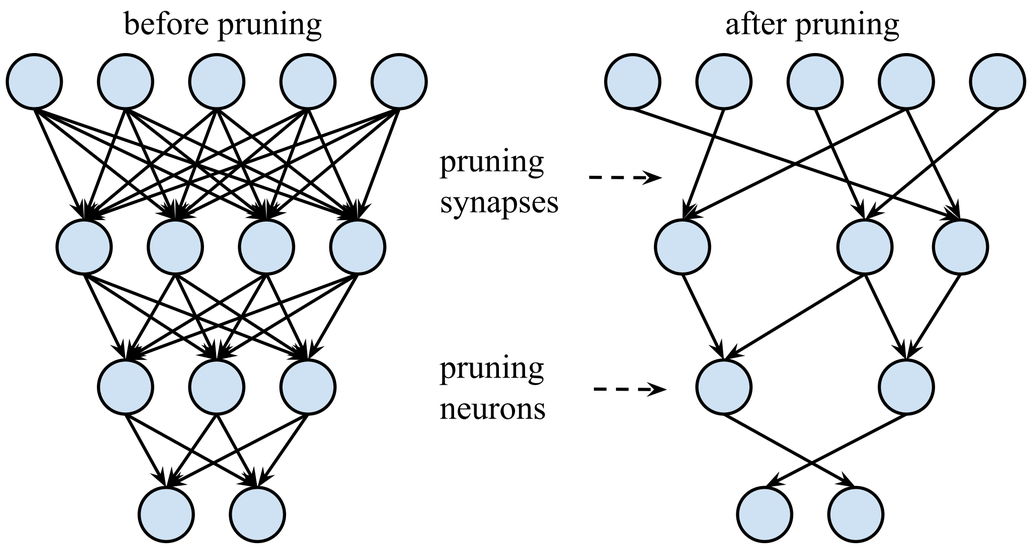
Fit NN on FPGA: Quantization & Reuse

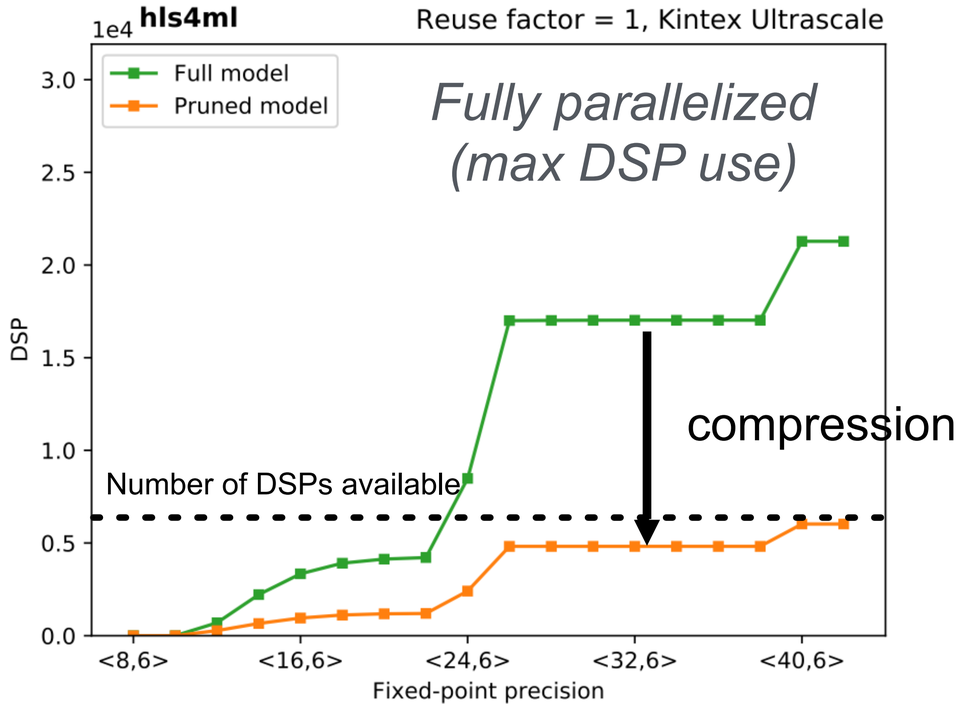




Efficient NN design: compression

[https://arxiv.org/abs/1804.06913]





Neural Network compression is a widespread technique to reduce the size, energy consumption, and overtraining of deep neural networks

Several approaches in literature

arxiv.1510.00149, arxiv.1712.01312, arxiv.1405.3866, arxiv.1602.07576, doi:10.1145/1150402.1150464

Recent developments

First paper demonstrated a fully connected NN in 100 ns.

HLS4ML in CMS

Run 3: muon momentum regression in CMS

More models demonstrated for Phase-2 trigger upgrade TDR

Advanced models:

binary/ternary, CNNs, RNNs, auto-encoders. Support for Graph Neural Network Models

Advanced Pruning/quantization:

Quantization-aware training with QKeras/Quantization-aware pruning

On ASICs and Low power devices.

For latest status: please check hls4ml website, CPAD 2021 talk,

Try it out: hls4ml tutorials

High Level Synthesis 4 Machine Learning