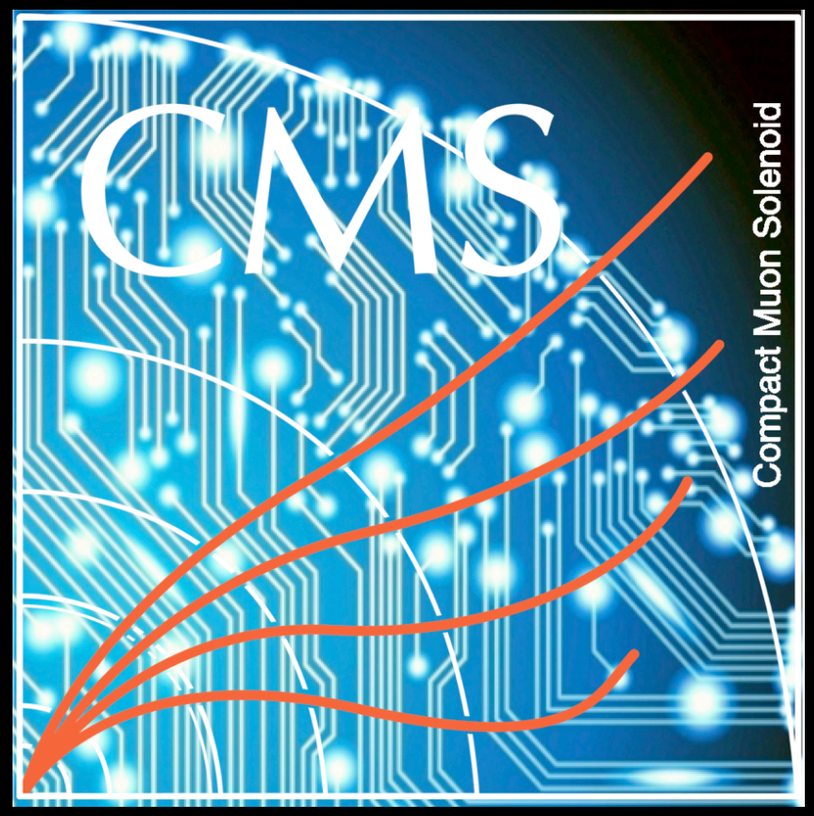
CMS at the frontier of Deep Learning applications in Particle Physics

Jennifer Ngadiuba (Fermilab)
on behalf of the CMS Collaboration

Aspen Winter Conference on Particle Physics
26-31 March, 2023
Aspen Center for Physics, Aspen (USA)



Fermilab

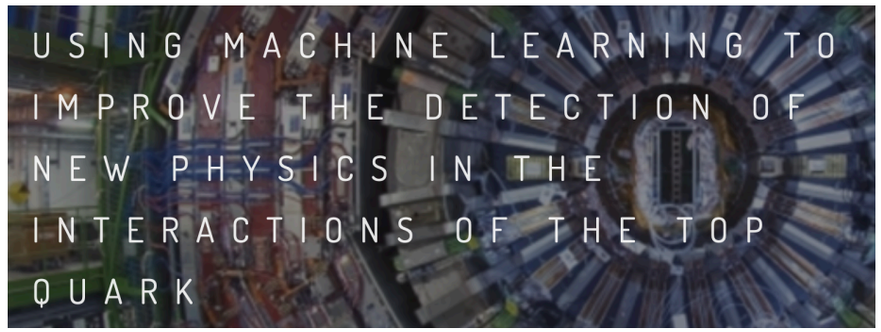
Intro: Machine Learning @ CMS

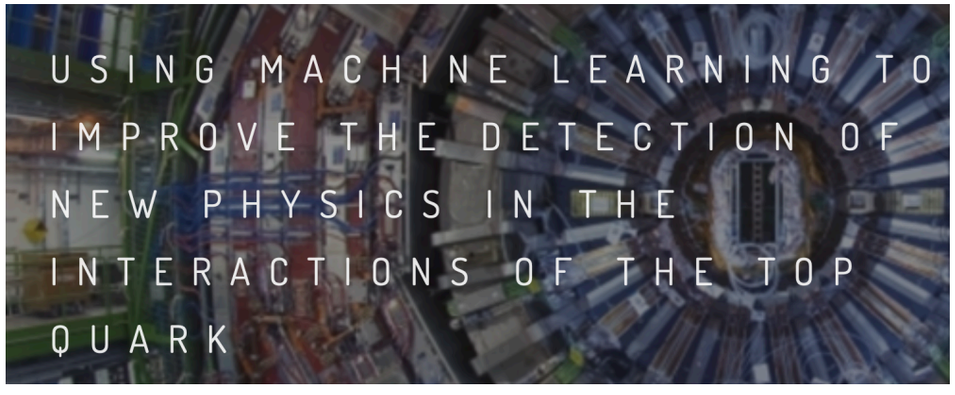
- **Growing number of applications of machine learning techniques in CMS**

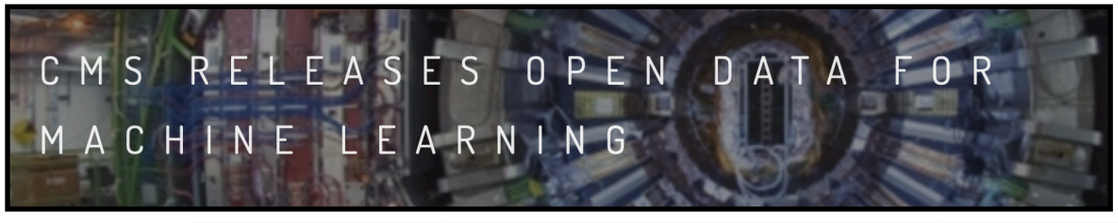
  - increasingly **more sophisticated** exploiting advances from non-scientific domains

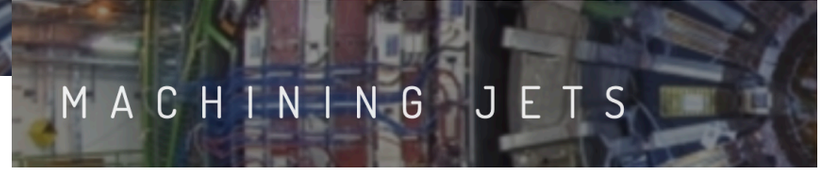
  - include also co-existing effort of **expanding central software and hardware infrastructures** to enable inference and training at scale

  - **the CMS ML group** at the border between computing and physics — coordinating and overseeing technical and innovation aspects across the experiment









Check CMS news page!

Intro: Machine Learning @ CMS

ML for classification:

heavy jet tagging
heavy flavour jet tagging
exotic jets
tau leptons
event-level (also unsupervised)

ML beyond classification:

mass and energy regression
background estimation
detector faults detection
reconstruction & simulation
triggering

Computing software & hardware for ML:

optimized inference in central software for CPU/GPU
GPU hardware on-site for software trigger system & grid sites
more powerful chips in hardware trigger system & development of portable tools*
ML-friendly central data format (NanoAOD) and scalable processing tools (Coffea)

DISCLAIMERS:

a lot ongoing that cannot be covered today
for more info contact CMS ML group conveners!

* see Javier's talk

CMS

Compact Muon Solenoid

ML for classification

Heavy jet tagging

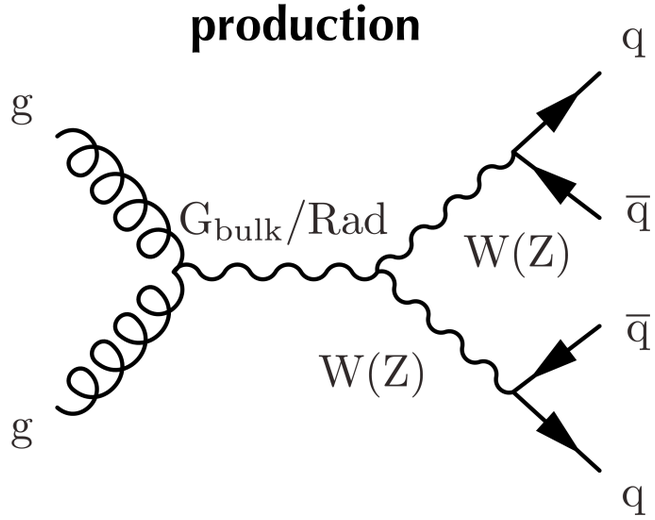
- **Identification of jets arising from hadronization of boosted W/Z/H/top is a key task in LHC physics:**

  - new physics searches, standard model measurements, higgs sector

  - **unique signature** from hadrons merging in single jet with substructure

  - exploit to suppress **overwhelming background from multijet** processes in most sensitive all-hadronic and semi-leptonic channels

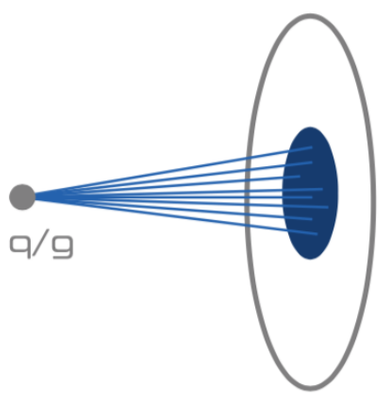
ex: Graviton or Radion production



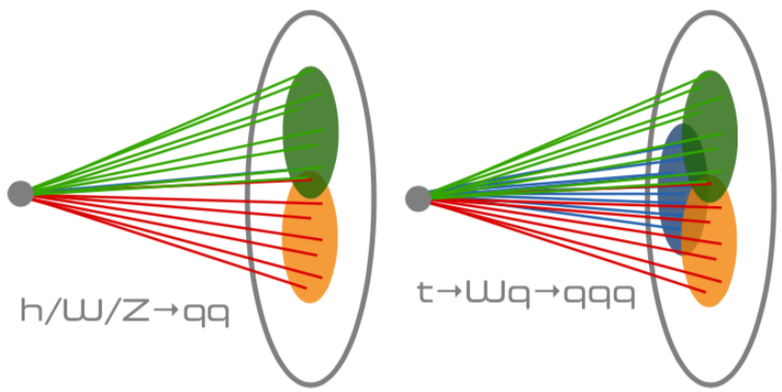
- A topic of interest in both theory and experiment communities since ~ 30 years

- Recent years advancement in ML enabled more powerful algorithms

BACKGROUND JET (single q/g)

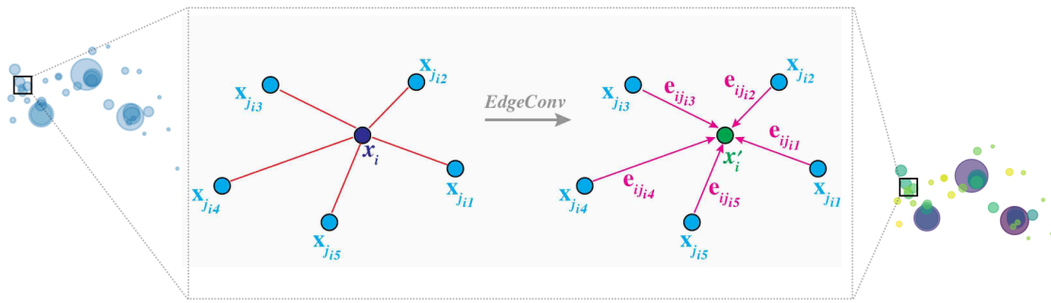


SIGNAL JETS

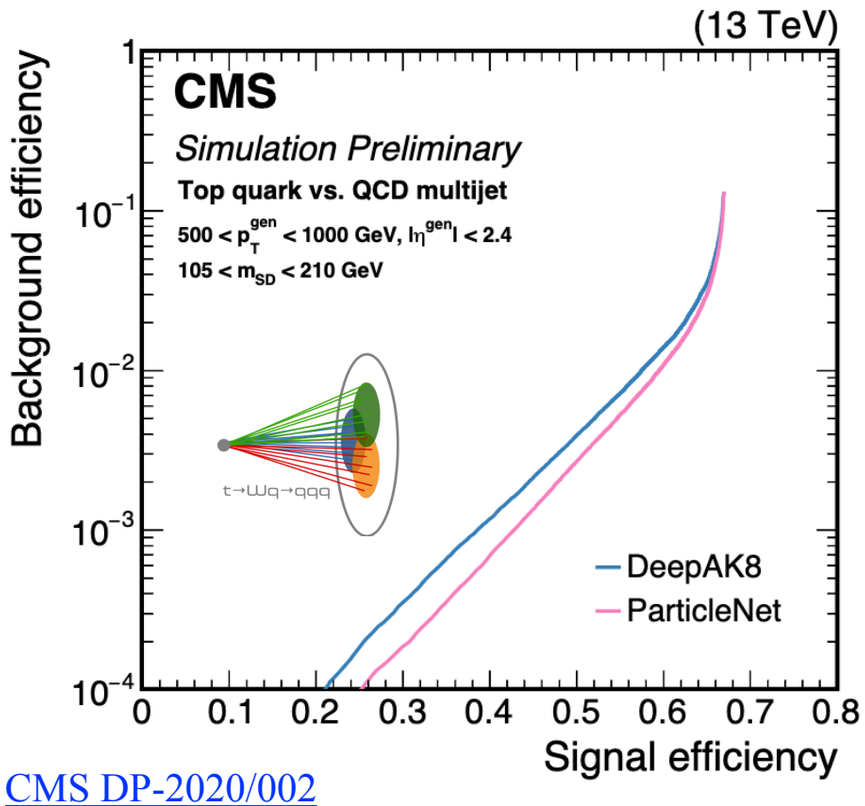


Heavy jet tagging @ CMS

- **A variety of ML algorithms developed and deployed in CMS since beginning of Run 2**
- **SOA: ParticleNet — first graph-based tagger at LHC !**
  - jet raw representation based on **point cloud formed by its constituents**
  - a **dynamic graph CNN** architecture used to fully exploit inherent permutation invariance of the representation *("EdgeConv"* block*)*



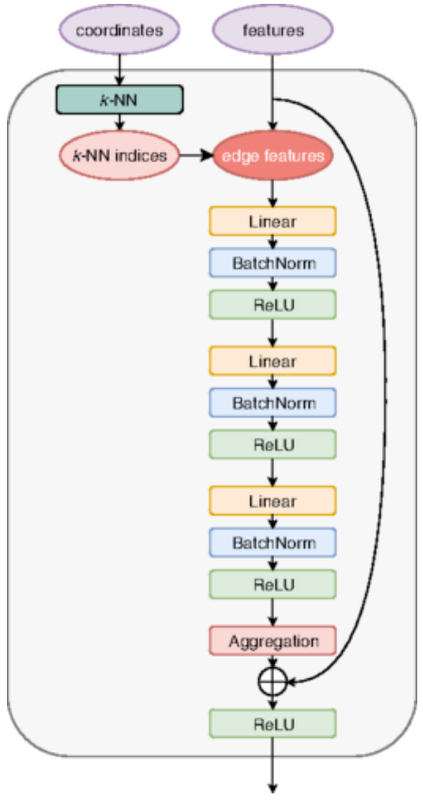
H. Qu et al.: Phys. Rev. D 101, 056019 (2020)



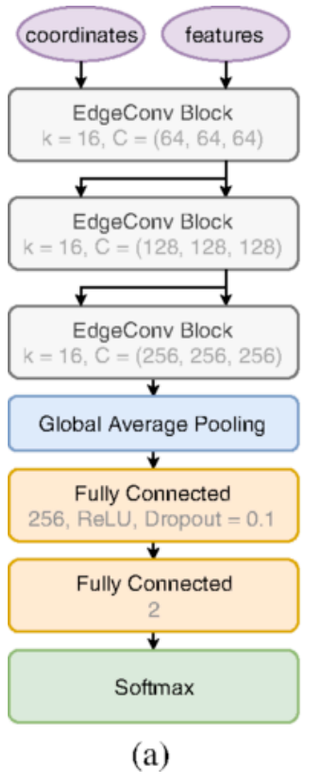
CMS DP-2020/002

DeepAK8:
Conv1D architecture on jet constituents
[CMS-JME-18-002]

EdgeConv block



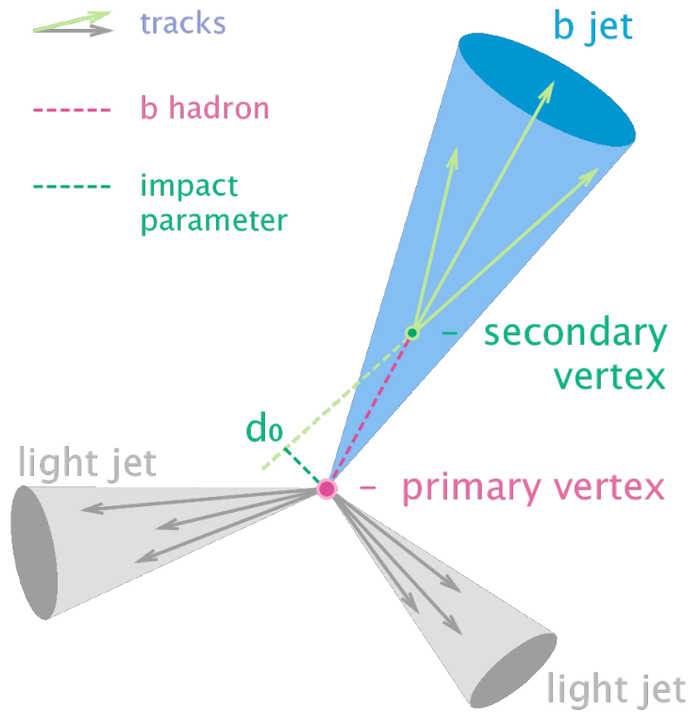
Stacked blocks

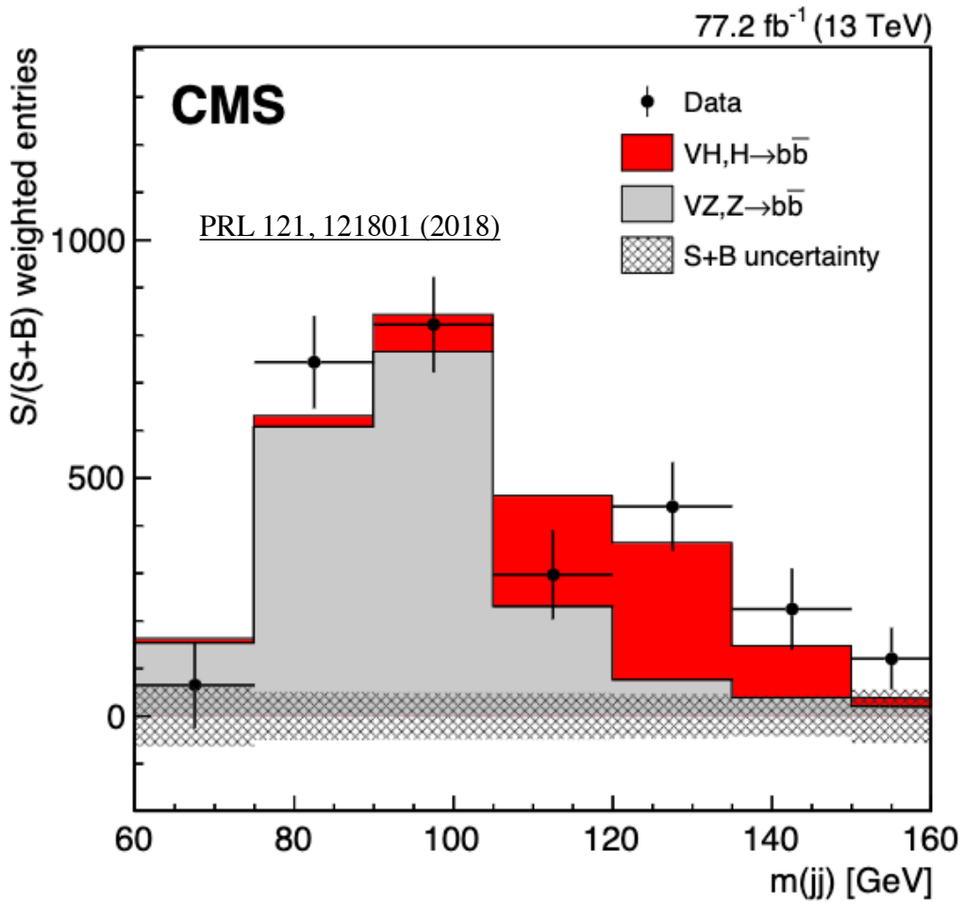


(a)

Heavy flavour jet tagging @ CMS

- Also heavy flavour (b/c) jet tagging allows to **access a variety of rare physics processes**
- Unique signature due to sizeable lifetime of B/C hadrons [$c\tau$ = O(mm)] resulting in displaced tracks and secondary vertices
- **We have built classifiers based on ML for HF jets at the LHC since Run 1**
  - successful in reducing time to discovery!

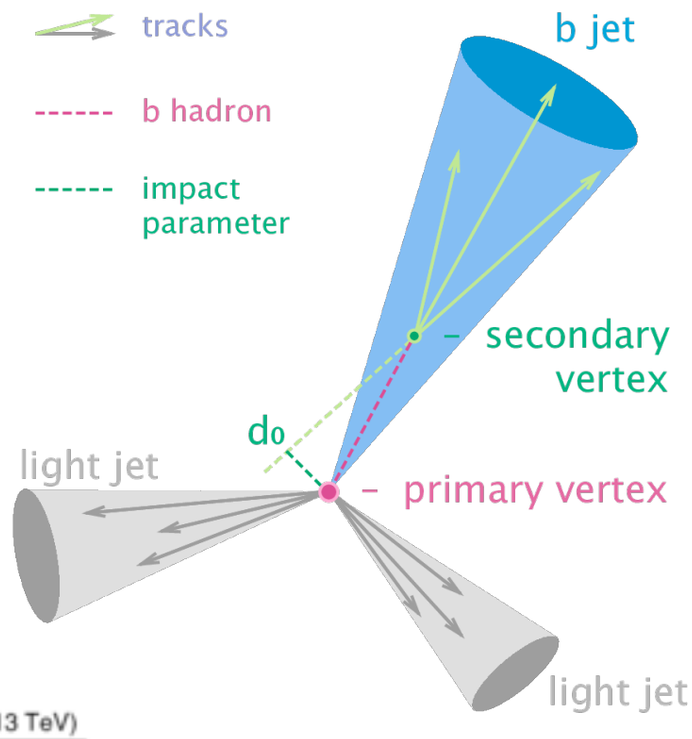


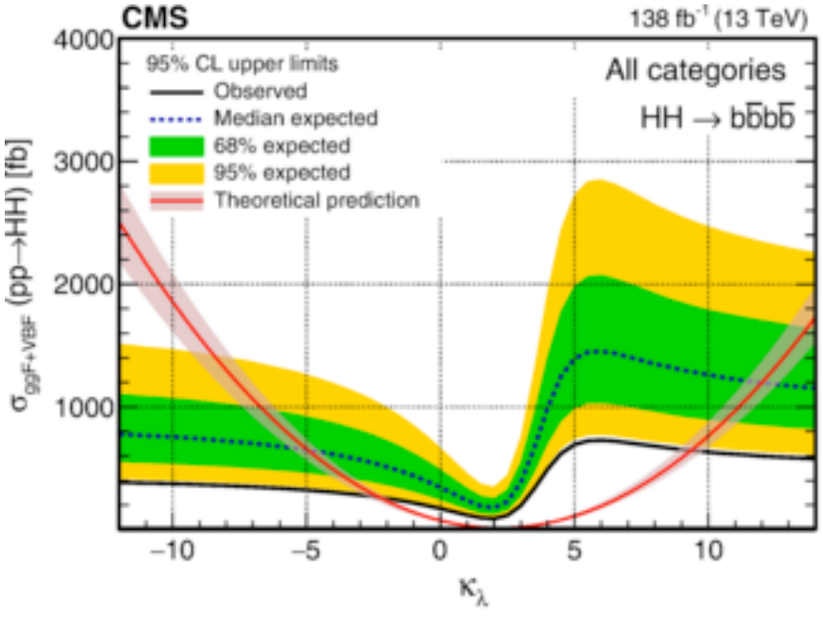


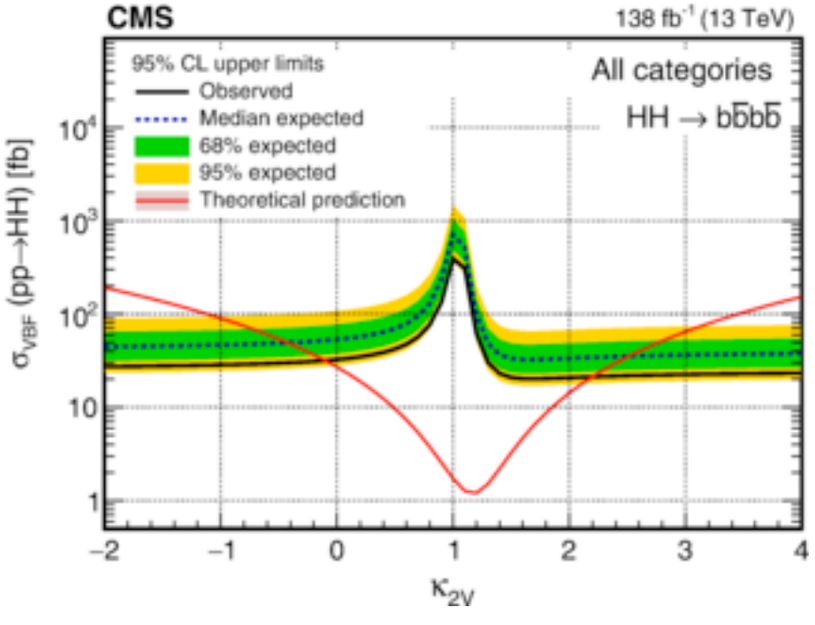
H→bb discovery in VH production

Heavy flavour jet tagging @ CMS

- Also heavy flavour (b/c) jet tagging allows to **access a variety of rare physics processes**
- Unique signature due to sizeable lifetime of B/C hadrons [$c\tau$ = O(mm)] resulting in displaced tracks and secondary vertices
- **We have built classifiers based on ML for HF jets at the LHC since Run 1**
  - successful in reducing time to discovery!
- **Application of more powerful ML architectures allowed recently setting the most stringent constraints on HH production!**



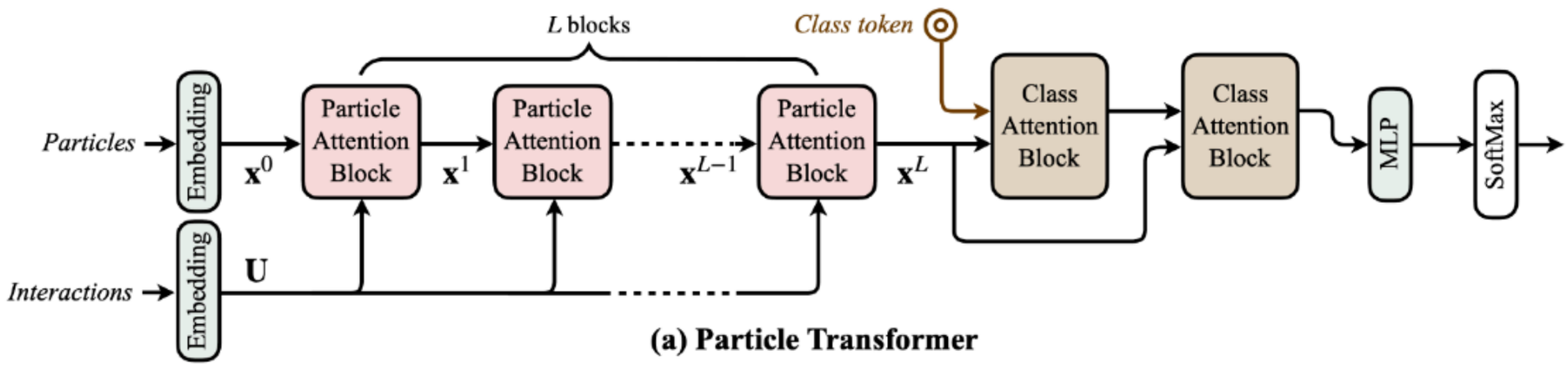




PRL 129, 081802 (2022)

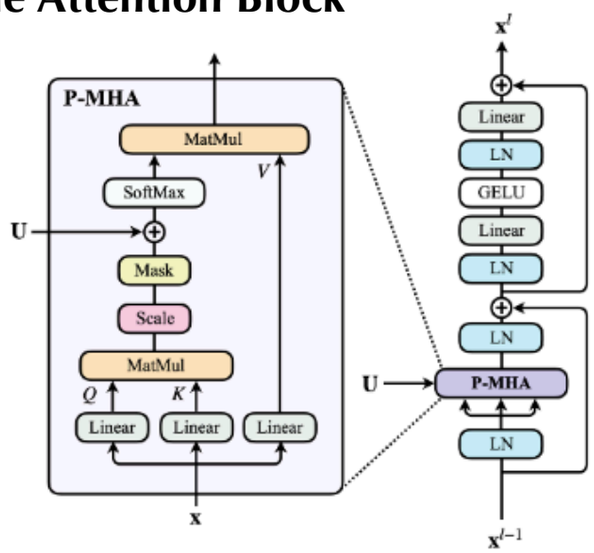
Heavy flavour jet tagging @ CMS

- Brand new development in CMS uses particle-based **transformers** — ***learn which neighbour particles are relevant through attention mechanism***
  - input embedding of both single particle and pair-wise features information
  - the pair-wise features encode physics principles → modifiers of standard dot-product attention weights in **Particle Attention Block**



(a) Particle Transformer

Particle Attention Block



(b) Particle Attention Block

$$\text{P-MHA}(Q,K,V) = \text{SoftMax}(QK^T/\sqrt{d_k} + \mathbf{U})V$$

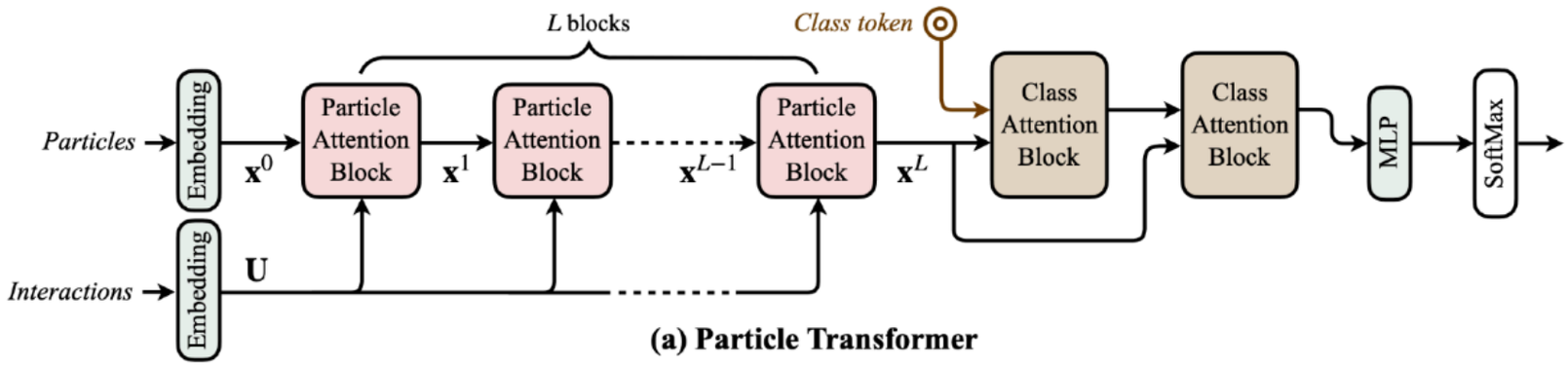
d_k: dimension of K

$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2}$$
$$k_T = \min(p_{T,a}, p_{T,b}) \cdot \Delta$$
$$z = \min(p_{T,a}, p_{T,b})/(p_{T,a} + p_{T,b})$$
$$m^2 = (E_a + E_b)^2 - ||\mathbf{p}_a + \mathbf{p}_b||^2$$

H. Qu et al.: arXiv.2202.03772 CMS-DP-2022/050

Heavy flavour jet tagging @ CMS

- Recent development in CMS uses particle-based **transformers** — ***learn which neighbour particles are relevant through attention mechanism***
  - input embedding of both single particle and pair-wise features information
  - the pair-wise features encode physics principles → modifiers of standard dot-product attention weights in **Particle Attention Block**
  - **Class Attention Block:** compute the attention between a global class and all the particles via standard Multi-Head Attention (MHA)



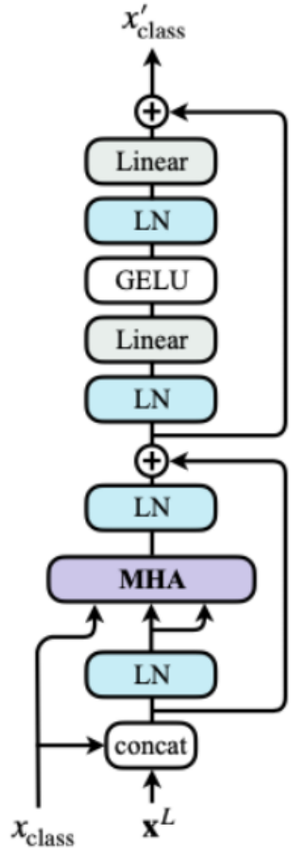
(a) Particle Transformer

$$\mathrm{MHA_C}(Q_\mathrm{C}, K_\mathrm{C}, V_\mathrm{C}) = \mathrm{SoftMax}(Q_\mathrm{C}K_\mathrm{C}^T/\sqrt{d_{k\mathrm{C}}})V_\mathrm{C}$$

$$Q_\mathrm{C} = W_{q\mathrm{C}}x_\mathrm{class} + b_{q\mathrm{C}} \qquad K_\mathrm{C} = W_{k\mathrm{C}}\mathbf{z} + b_{k\mathrm{C}} \qquad V_\mathrm{C} = W_{v\mathrm{C}}\mathbf{z} + b_{v\mathrm{C}} \qquad d_{k\mathrm{C}}\text{: dimension of } K_\mathrm{C}$$

$$\mathbf{z} = [x_\mathrm{class}, \mathbf{x}^L]$$

Class Attention Block

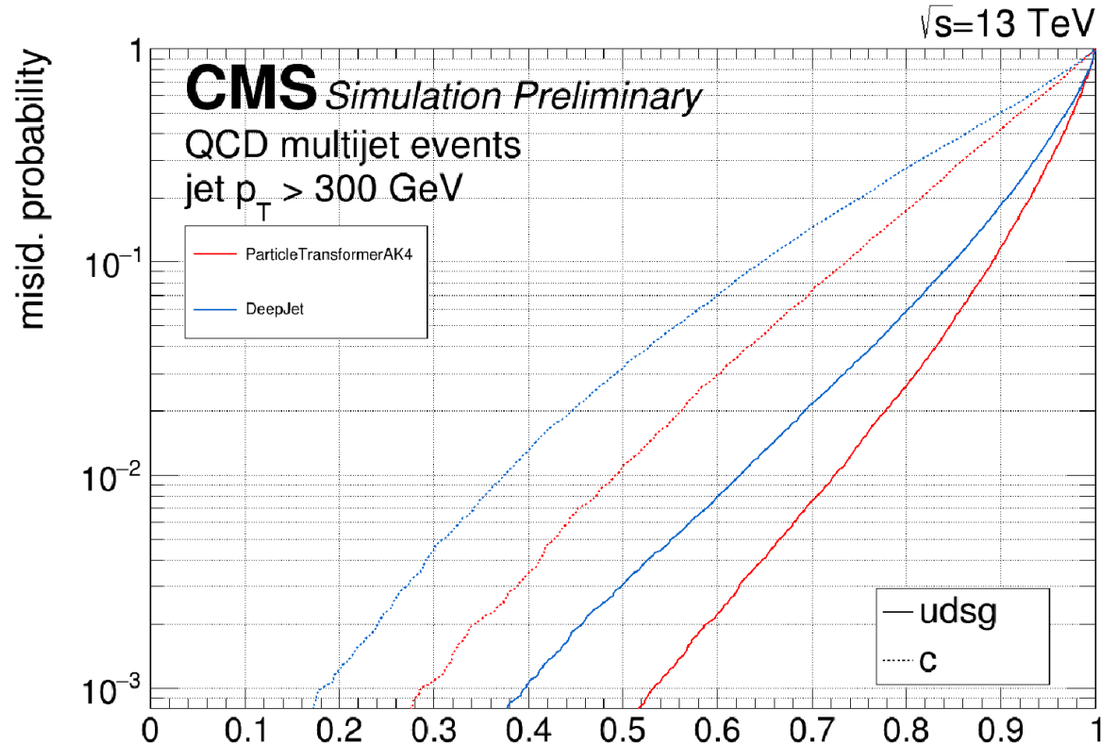


(c) Class Attention Block

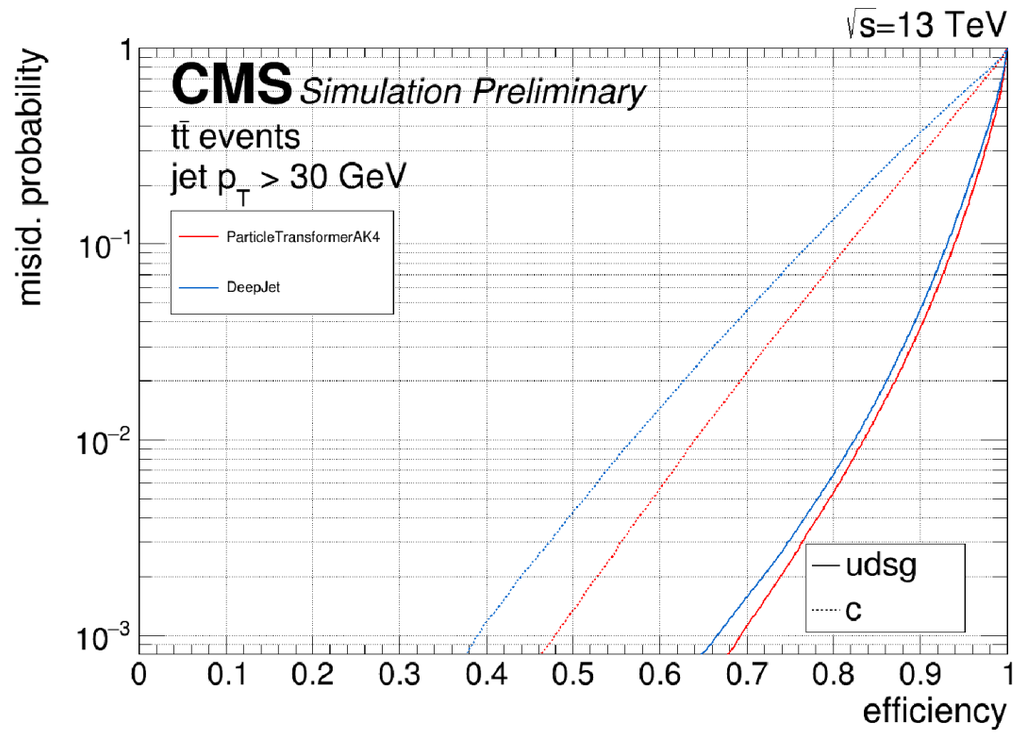
[CMS-DP-2022-050]

Heavy flavour jet tagging @ CMS

- Comparison with current default CMS algo DeepJet based on 1D CNN architecture on per-particle plus b-tagging features
- Known challenges in tagging high p_T jets — worse reconstruction of key features like impact parameter and secondary vertices
- **Introduction of pair-wise features mitigate performance loss thanks to increased expressivity**



Finalising integration in CMS software…
Looking forward for deployment in Run 3 analyses — first attention-based tagger at LHC!



CMS-DP-2022-050

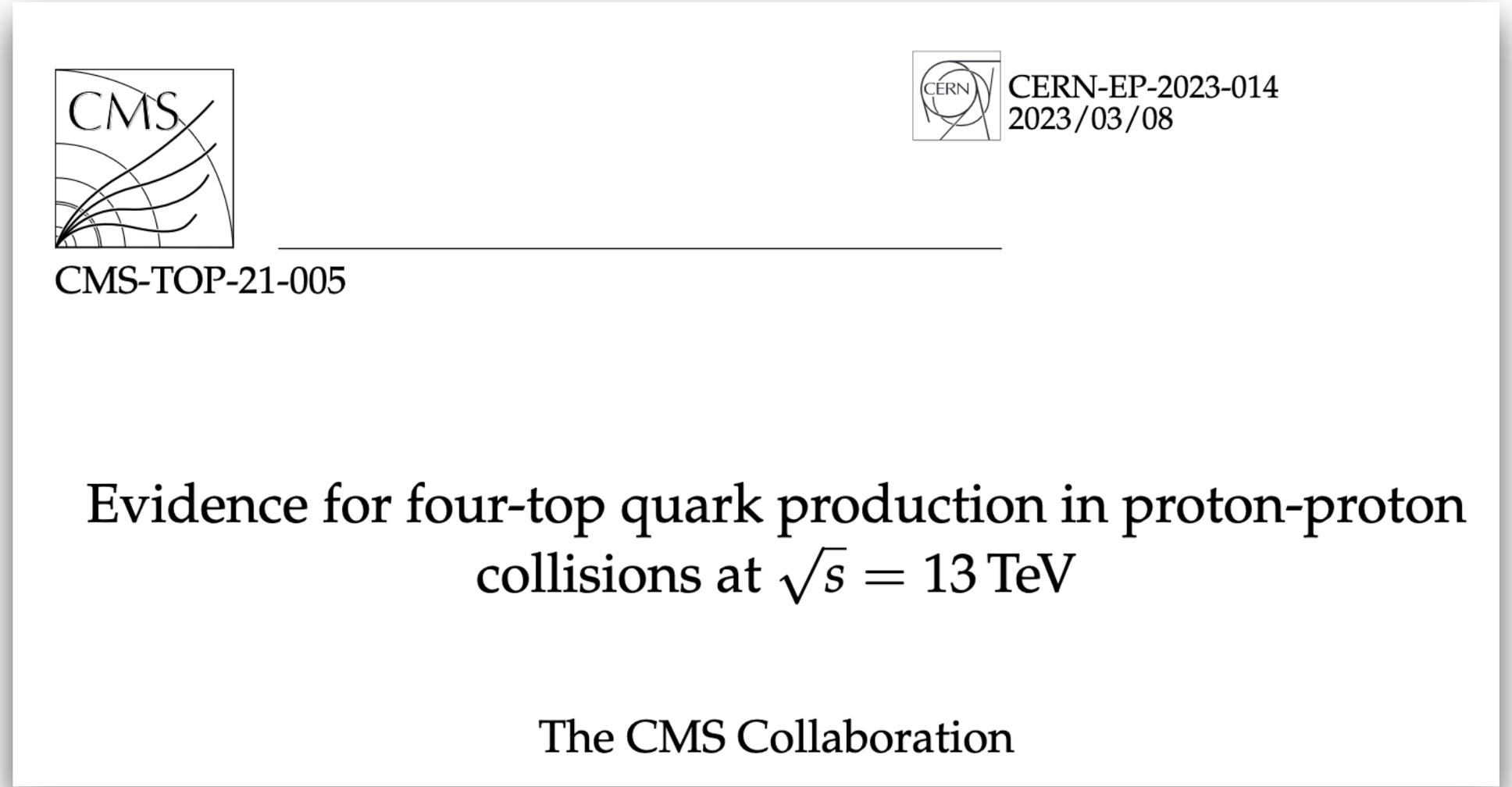
CMS

Compact Muon Solenoid

ML beyond classification

ML for background estimation

- Recent exciting CMS results reporting **observation of the rare 4 top quarks production process at 4σ** for combination of multiple final states [CMS-TOP-21-005 — Submitted to PLB]



CMS

CERN

CERN-EP-2023-014
2023/03/08

CMS-TOP-21-005

Evidence for four-top quark production in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$

The CMS Collaboration

OBSERVATION JUST REPORTED LAST WEEK AT MORIOND!

ML for background estimation

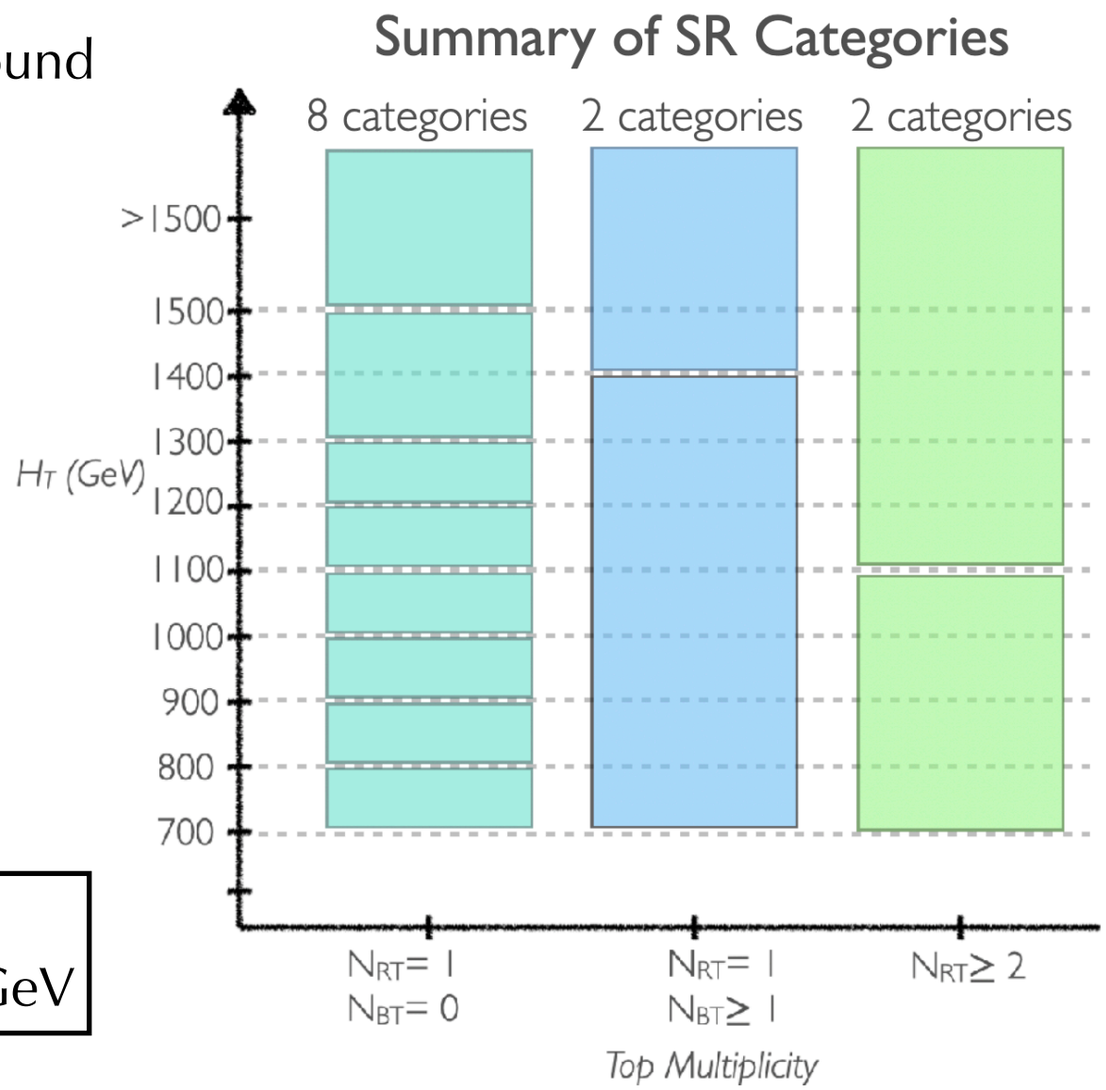
- Recent exciting CMS results reporting **evidence for the rare 4 top quarks production process at 4σ** for combination of multiple final states [CMS-TOP-21-005 — Submitted to PLB]

- Analysis of data in the challenging **all-hadronic final state performed for the first time**

  - after multijet preselection a BDT is trained to separate signal vs $t\bar{t}$+QCD background → final discriminating observable

  - multiple signal categories using tagged resolved tops (N_{RT}), tagged boosted tops (N_{BT}) and H_T

  - **background estimation with Neural Autoregressive Flows** novel in CMS!



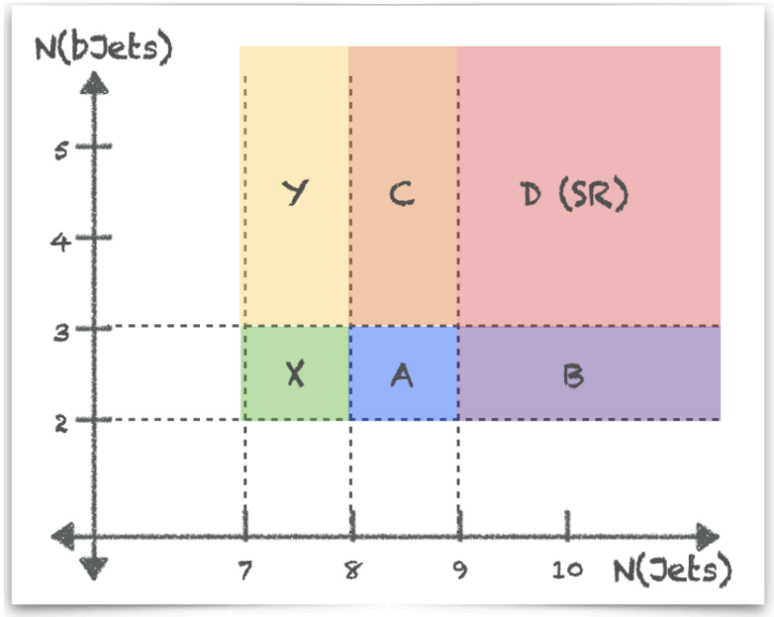
Baseline selections:
N(leptons) = 0, N(jets) ≥ 9, N(b-jets) ≥ 3, HT ≥ 700 GeV

ML for background estimation

- Start with defining 5 CRs as in “extended” ABCD method (i.e. number of CRs > 3)

- Main idea: for each CR learn transformation $\tau : \mathscr{P} \to \mathscr{P}'$ using NAF as universal approximator of bijective transformation with sigmoidal function σ

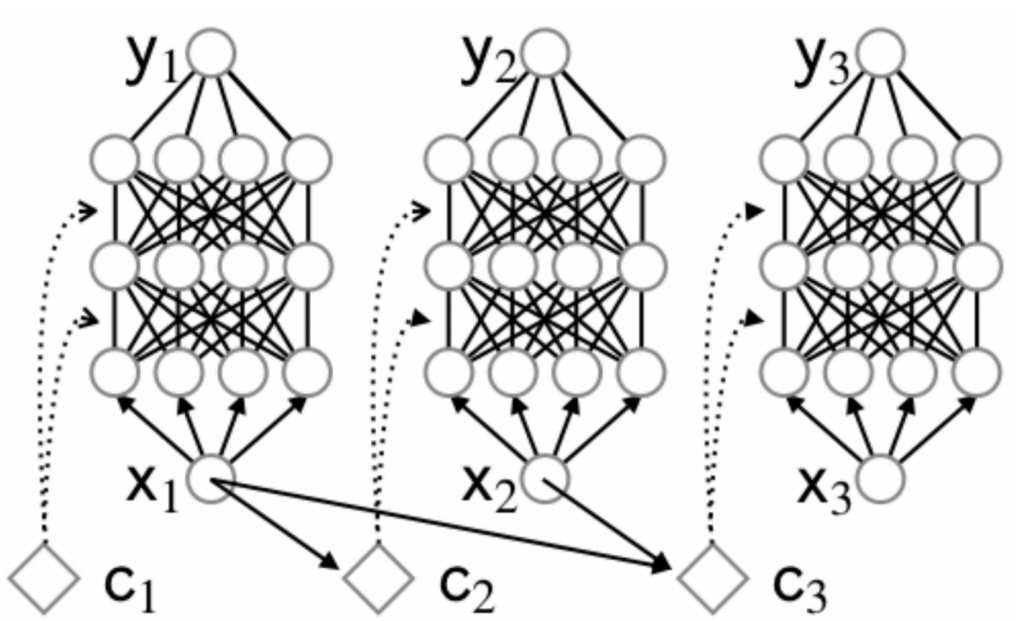
$$\mathscr{P}'(\vec{x}'\,|\,\vec{c}\,') = \tau(\vec{x}'; \vec{x}\,|\,\vec{c}\,'; \vec{c}\,) \otimes \mathscr{P}(\vec{x}\,|\,\vec{c}\,)$$



- $\mathscr{P}'$: $t\bar{t}$+QCD background = data - ($t\bar{t}t\bar{t}$ + minor MC backgrounds)
- $\mathscr{P}$: $t\bar{t}$ simulation
- $\vec{x}$: input variables
- $\vec{y}$: output variables

} **BDT and HT shapes**

- $\vec{c}$: conditions, including control variables N(Jets) and N(bJets) plus preceding inputs
- loss function: maximum-mean-discrepancy
- activation function: sigmoid

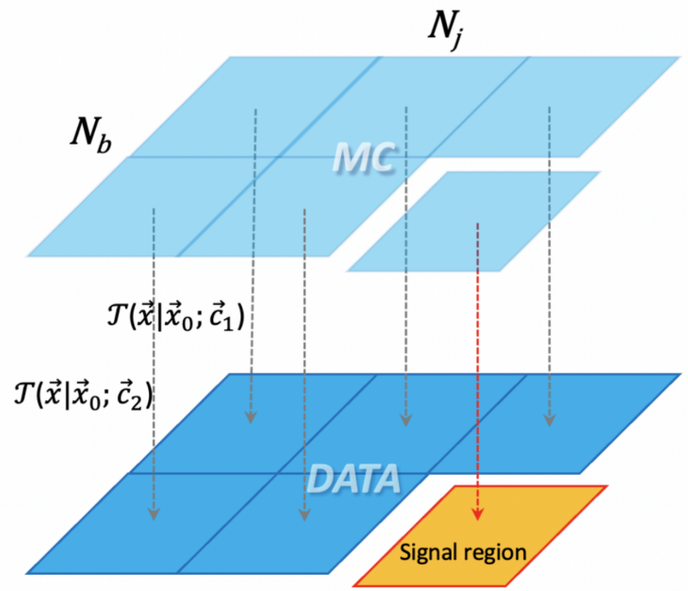


ML for background estimation

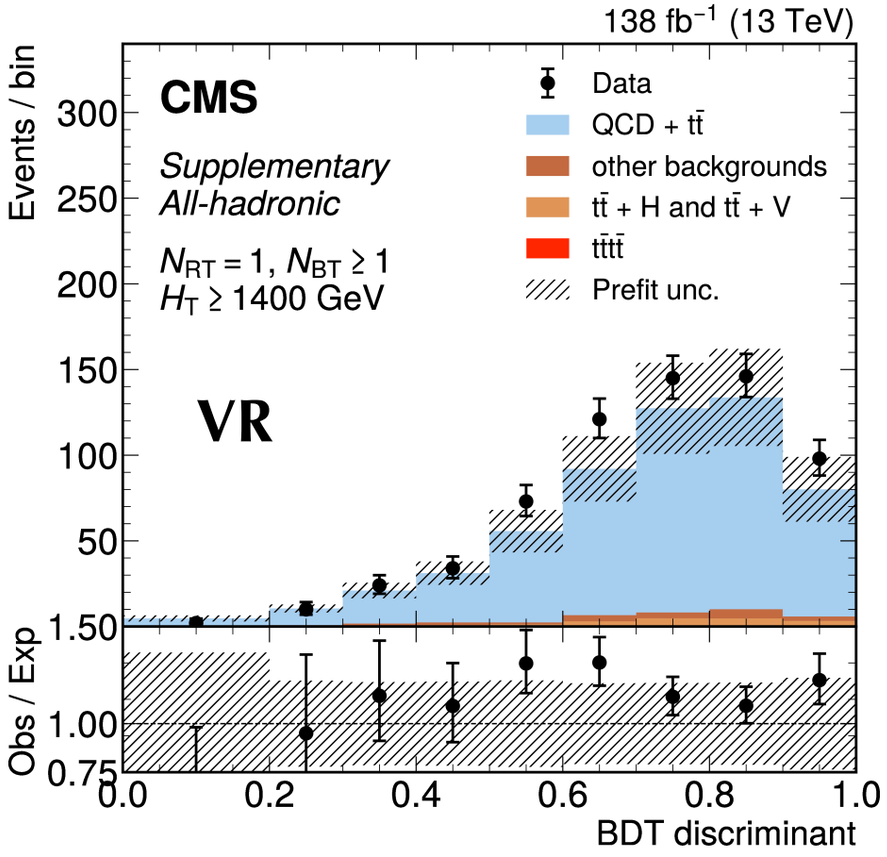
- After training in the 5 CRs, the transformation between MC and data is applied to $t\bar{t}$ simulation in the SR
  → morphed to predict the shape
  of the $t\bar{t}$+QCD multijet background in the SR

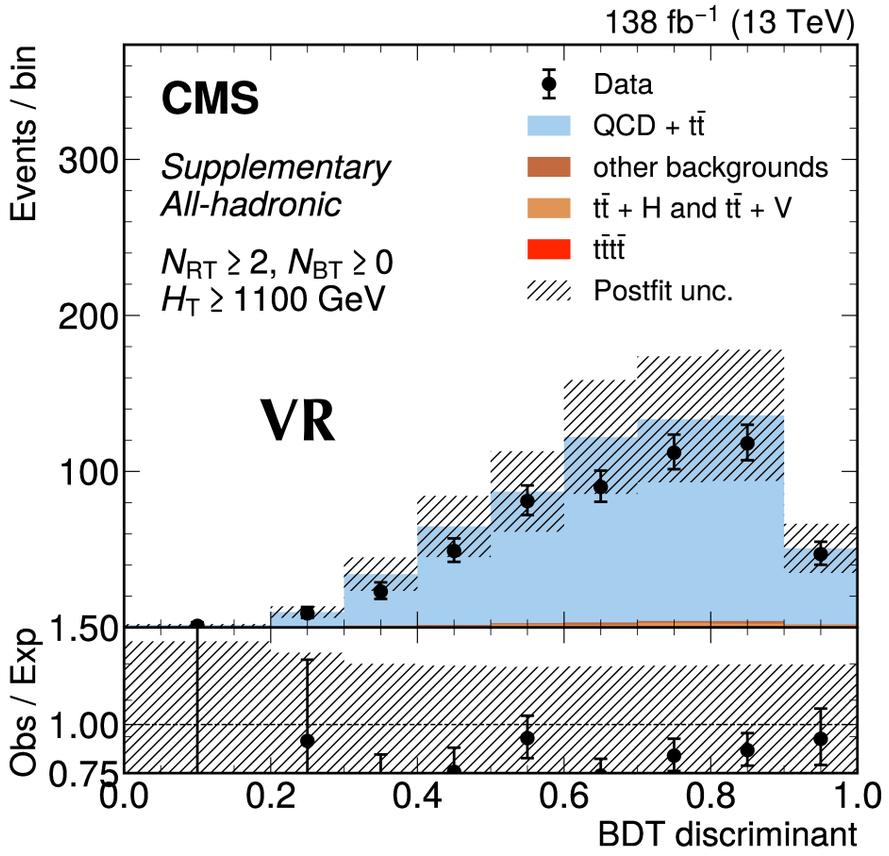
- **Extensive closure checks in validation region**:
  identical to SR but with N(jets) = 8

- Uncertainties derived from discrepancies in the VR and applied to corresponding SR



[CMS-TOP-21-005 Submitted to PLB]





ML for detector health monitoring

- The online data quality monitoring (DQM) of CMS aims at **detecting faults almost in real-time — O(s)**

  - initiate reaction from detector experts to promptly identify and fix problems

  - **minimize downtime, maximize time for physics!**

- **Standard workflow:** set of histograms that are populated with a set of events seen by the detector

- **Challenge:** anomalies come in all shapes and sizes
  → ***impossible to anticipate all possible failure modes***

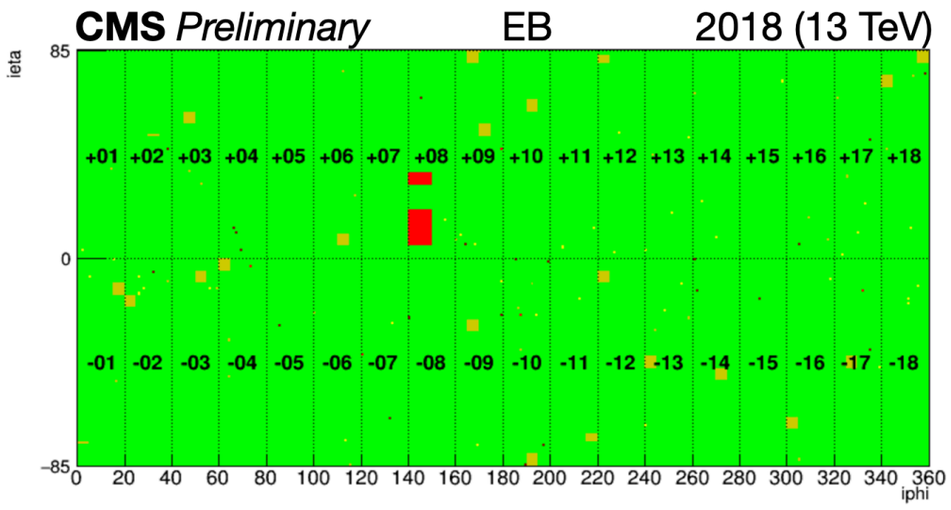
- **Why ML can help?**

  - provides robust anomaly detection and localization (via semi/unsupervised learning)

  - eliminate the need for hand coded rules for every possible fault

  - automated adaptivity to changing running conditions and experimental setup

Ex: CMS ECAL barrel



GREEN = good

RED = bad

BROWN = known problem

YELLOW = no data (which may or may not be problematic – depends on context).

ML for detector health monitoring

- **First development using an autoencoder in production since 2017 for Drift Tubes subsystem in muon detectors → first example of ML for online DQM at LHC!**

- More parallel efforts now ongoing mostly based on autoencoders **under or planned commissioning in Run 3 data taking**:
  - Resistive Plate Chambers subsystem of muon detectors [ACAT '22]
  - Electromagnetic and Hadronic Calorimeters subsystems [CMS-DP-2022-043, IML workshop]
  - Pixel Silicon Tracker subsystem [CMS-DP-2022-013]

Electromagnetic Calorimeter example:

- Training set of certified GOOD occupancy map images
- Test set of both synthetic and real Run 2 & 3 anomalies
- Autoencoder model: ResNet for both encoding and decoding
- Preprocessing to take into account variable spatial response and time-dependent nature of faults

Metric: False Discovery Rate (FDR)
at 99% anomaly detection
(i.e. what fractions of shifter calls will be false alarm?)

FDR at 99% anomaly detection							
Scenario	**Missing Supermodule**	**Zero Occupancy Tower**			**Hot Tower** (10% hot for EB, 20% hot for EE)		
	Barrel	**Barrel**	**EE+**	**EE-**	**Barrel**	**EE+**	**EE-**
No correction	3.6 %	51 %	86 %	87 %	2.8 %	0.01 %	< 1/30k
After *spatial* correction	3.1 %	49 %	13 %	14 %	2.9 %	0.06 %	0.05 %
After *spatial and time* correction	0.13 %	4.1 %	5.6 %	6.3 %	< 1/10k	< 1/30k	< 1/30k

Mostly due to **actual anomalies** contaminating certified GOOD data. (See Backup).

~10^4 is the size of the validation set

ML for detector health monitoring

- **First development using an autoencoder in production since 2017 for Drift Tubes subsystem in muon detectors → first example of ML for online DQM at LHC!**

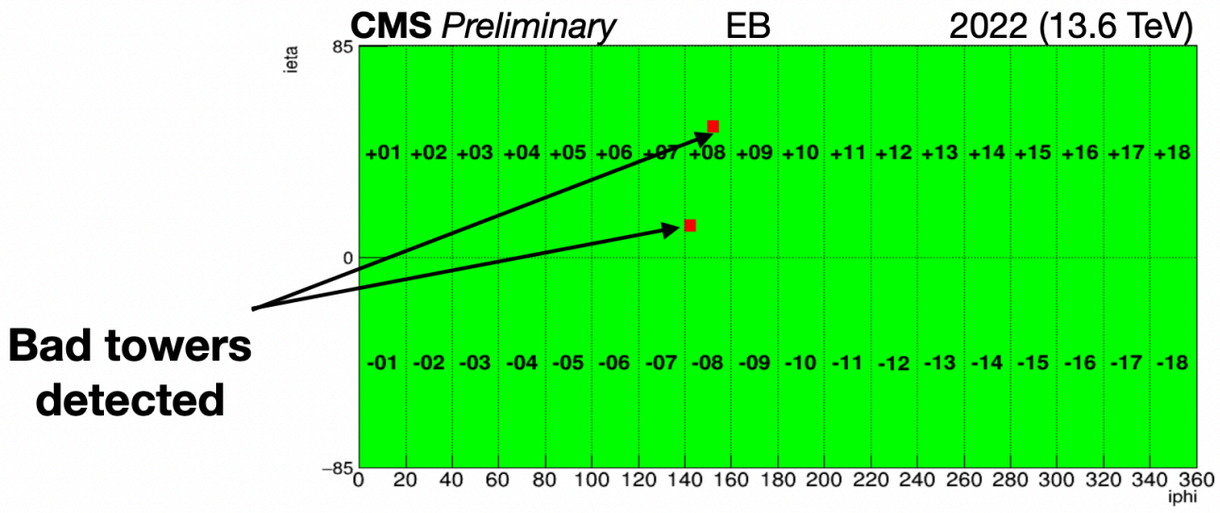
- More parallel efforts now ongoing mostly based on autoencoders **under or planned commissioning in Run 3 data taking**:
  - Resistive Plate Chambers subsystem of muon detectors [ACAT '22]
  - Electromagnetic and Hadronic Calorimeters subsystems [CMS-DP-2022-043, IML workshop]
  - Pixel Silicon Tracker subsystem [CMS-DP-2022-013]

Electromagnetic Calorimeter example:

- Training set of certified GOOD occupancy map images
- Test set of both synthetic and real Run 2 & 3 anomalies
- Autoencoder model: ResNet for both encoding and decoding
- Preprocessing to take into account variable spatial response and time-dependent nature of faults

Successful commissioning tests from last year run!

ML Quality plot from ECAL Online DQM during a Run3 run



CMS

Compact Muon Solenoid

Computing aspects

ML for particle flow reconstruction

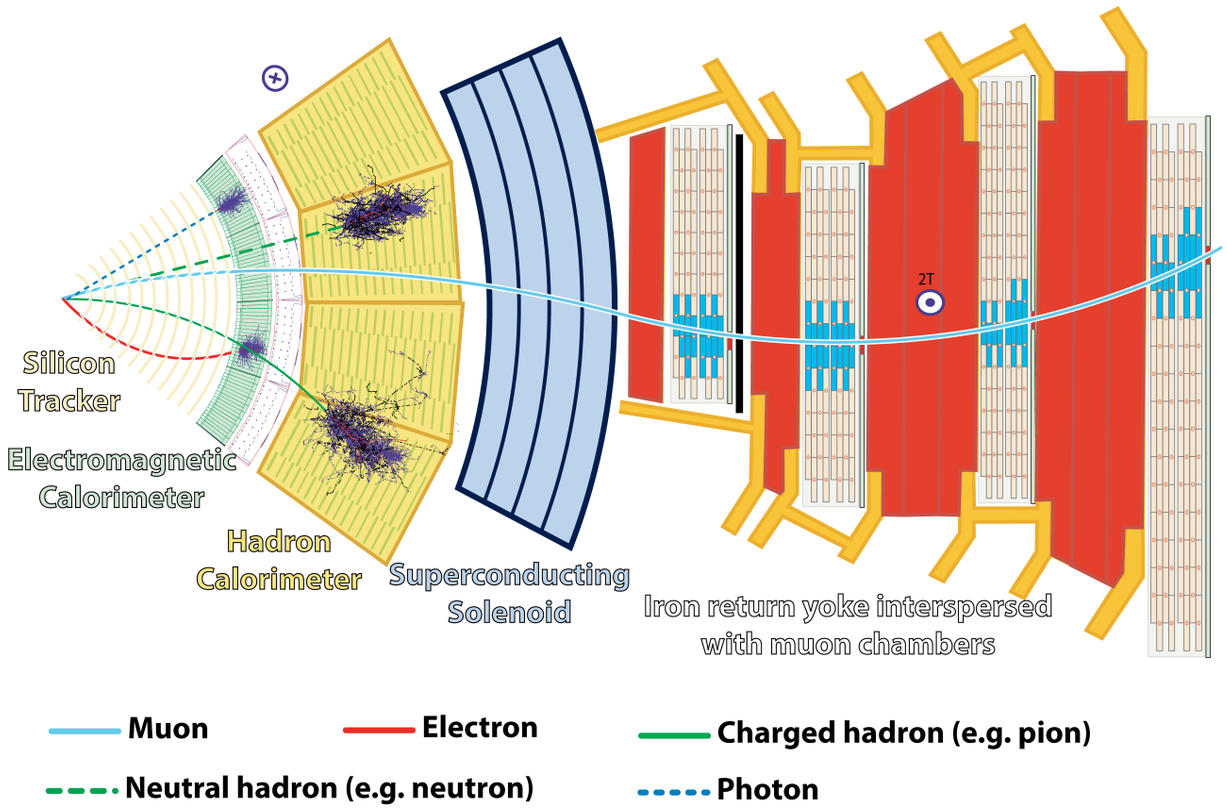
- General effort in the experimental community to **replace standard rule-based reconstruction algorithms with more computationally efficient and scalable ML models** in view of future increased challenges

- One example in CMS is the **Particle flow (PF) algorithm**:
  it combines information from all subdetector
  to reconstruct particles and thus improve the resolution

  - ex. track + hadronic energy
    = charged hadron

  - ex. no track + electromagnetic energy
    = photon

- It starts from calorimeter clusters & tracks and outputs particle candidates
  → **replace with ML model**



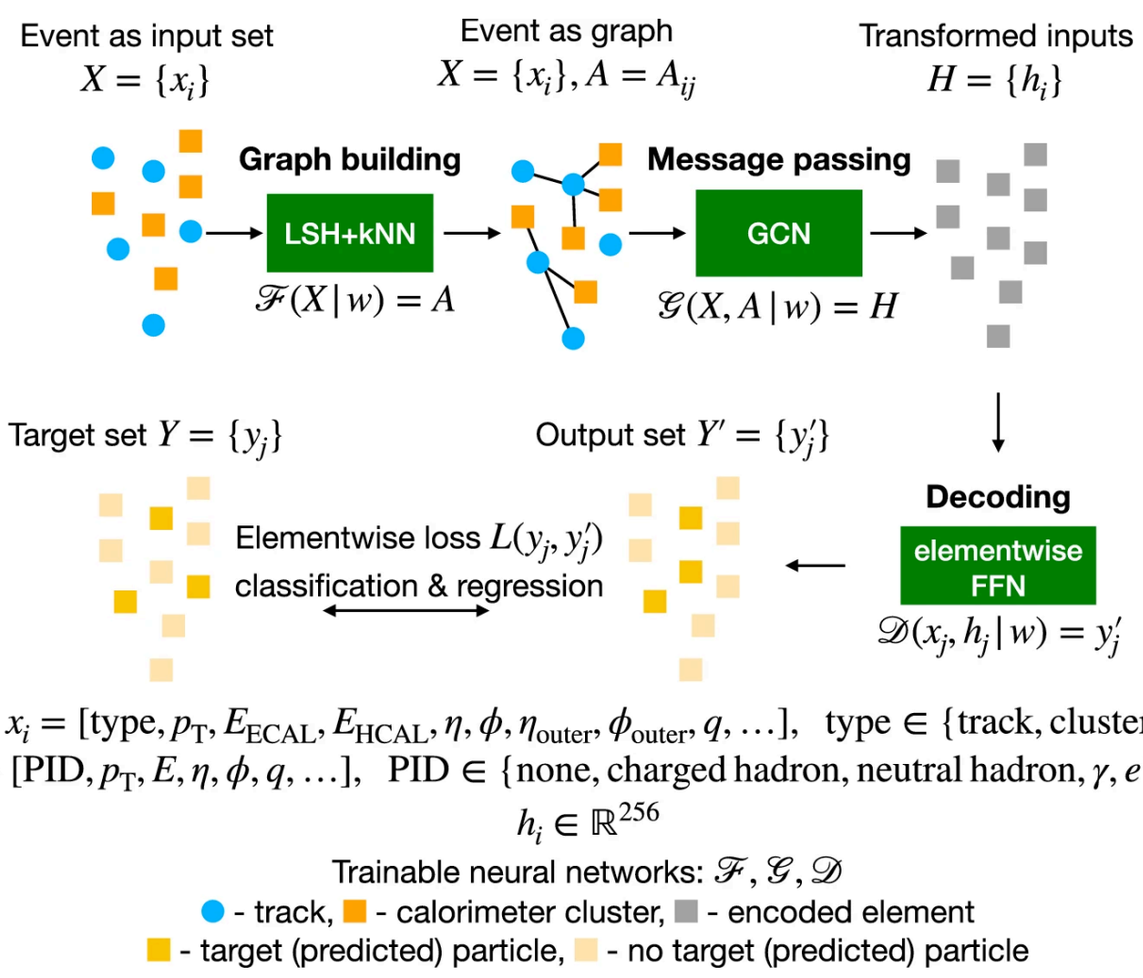
ML for particle flow reconstruction

- Model based on **dynamic graph CNN** generating on the fly multiple internal kNN graphs based on embedded features

- **Per-particle loss function** to simultaneously perform multi-classification and regression tasks

$$||Y - Y'|| \equiv \sum_{j \in \text{event}} L(y_j, y'_j)\,,$$

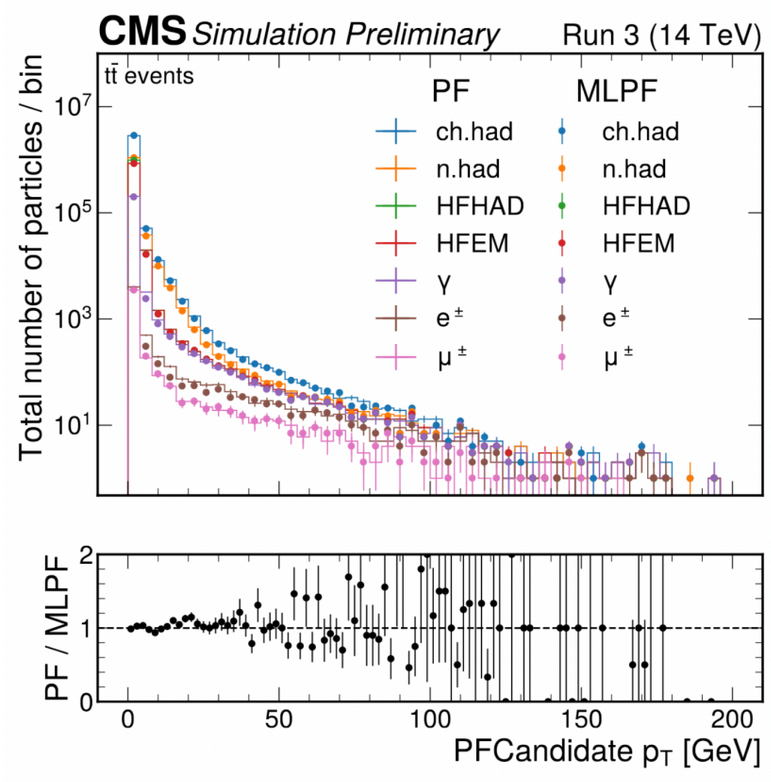
$$L(y_j, y'_j) \equiv \text{CLS}(c_j, c'_j) + \alpha \text{REG}(p_j, p'_j)\,,$$

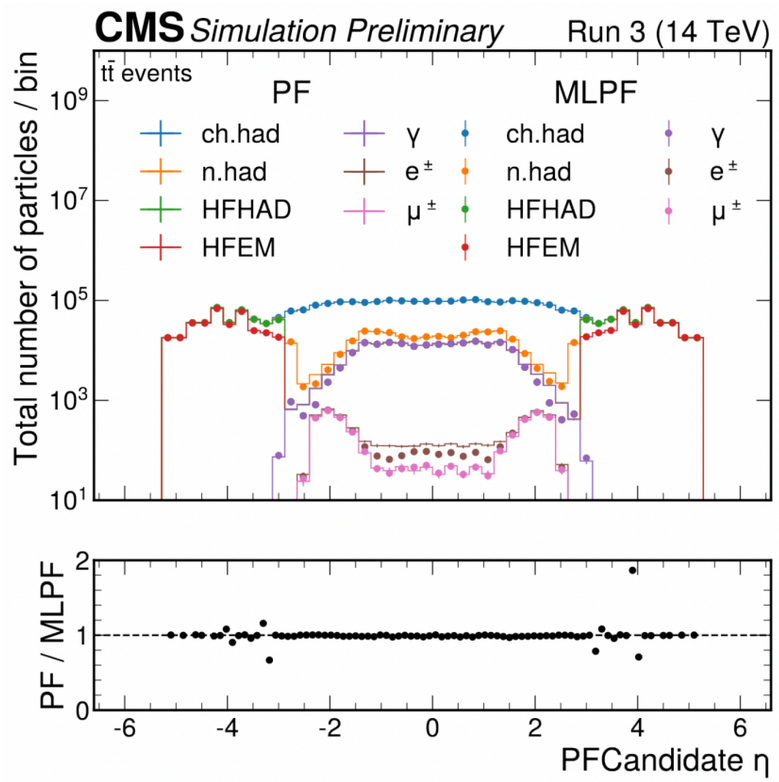


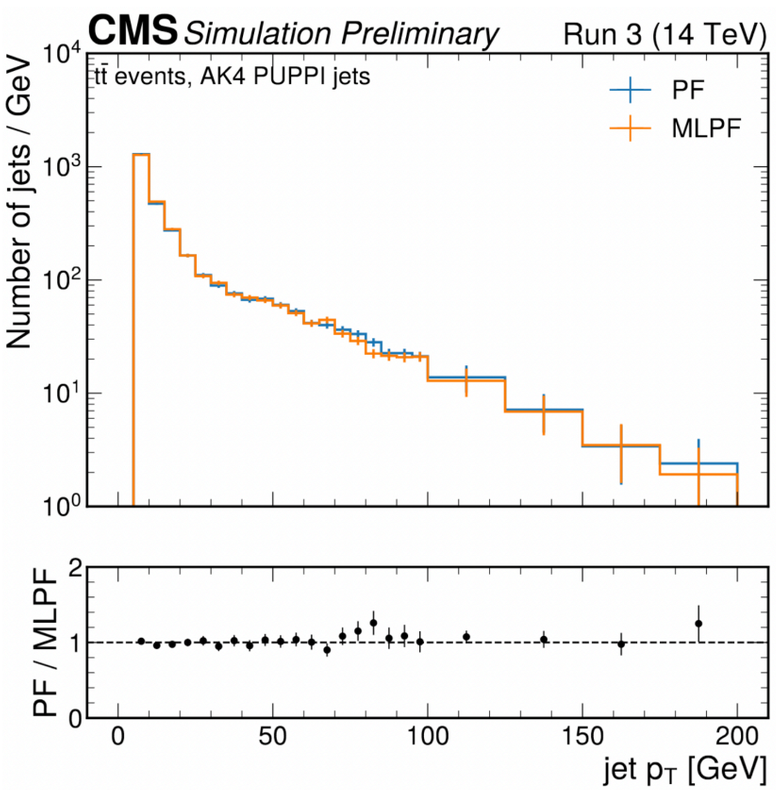
- **First version target baseline particle flow reconstruction** → does not allow to do better than that

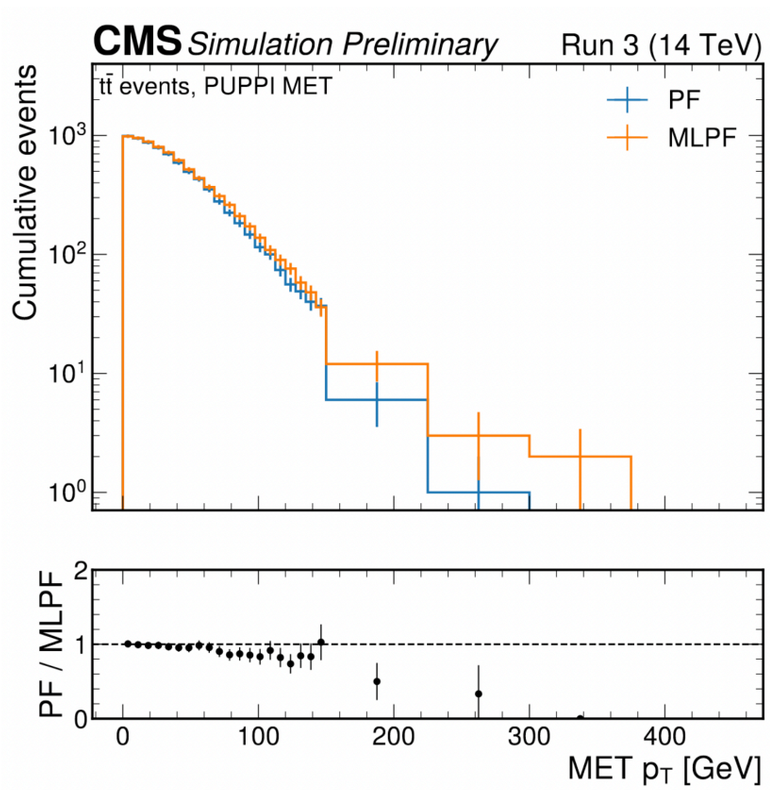
- **Second version target generator-level information** → can one improve response versus baseline?

ML for particle flow reconstruction









Eur. Phys. J. C (2021) 81: 381

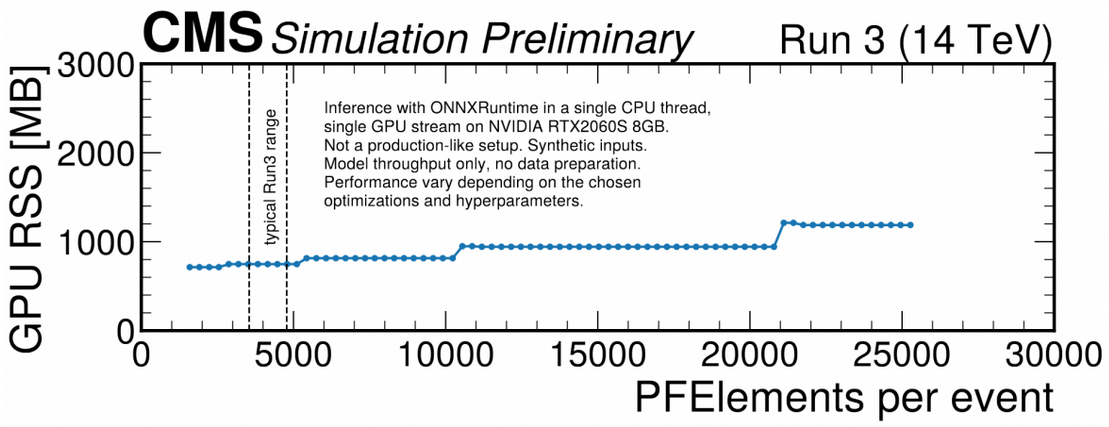
ACAT '21

ACAT '22

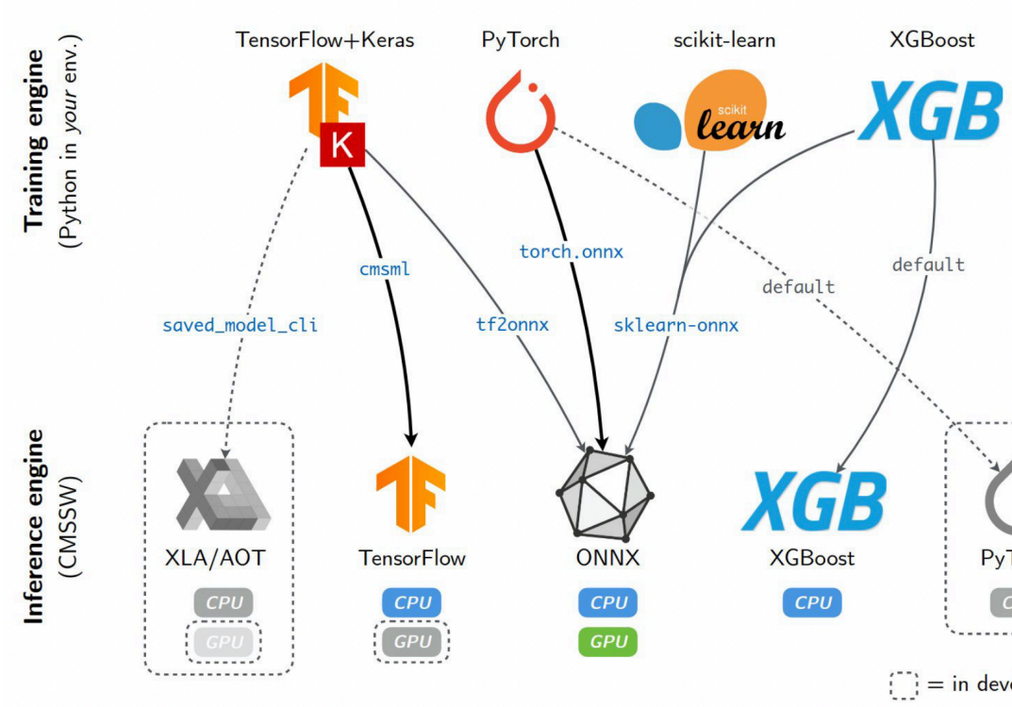
- Overall, a good particle-level agreement is observed between PF and MLPF algos
- Missing transverse energy mismodelling under investigation
- Using gen-level information gives similar performance

Computing infrastructure

- A crucial metric when developing ML models is the computational efficiency
- Cannot exploit full ML power without **extensive work to support and optimize ML inference in the CMS software**
  - continuous improvements in supporting direct inference on local GPU/CPU
  - promising effort in supporting inference as a service on cloud resources



Current CMS software support for ML

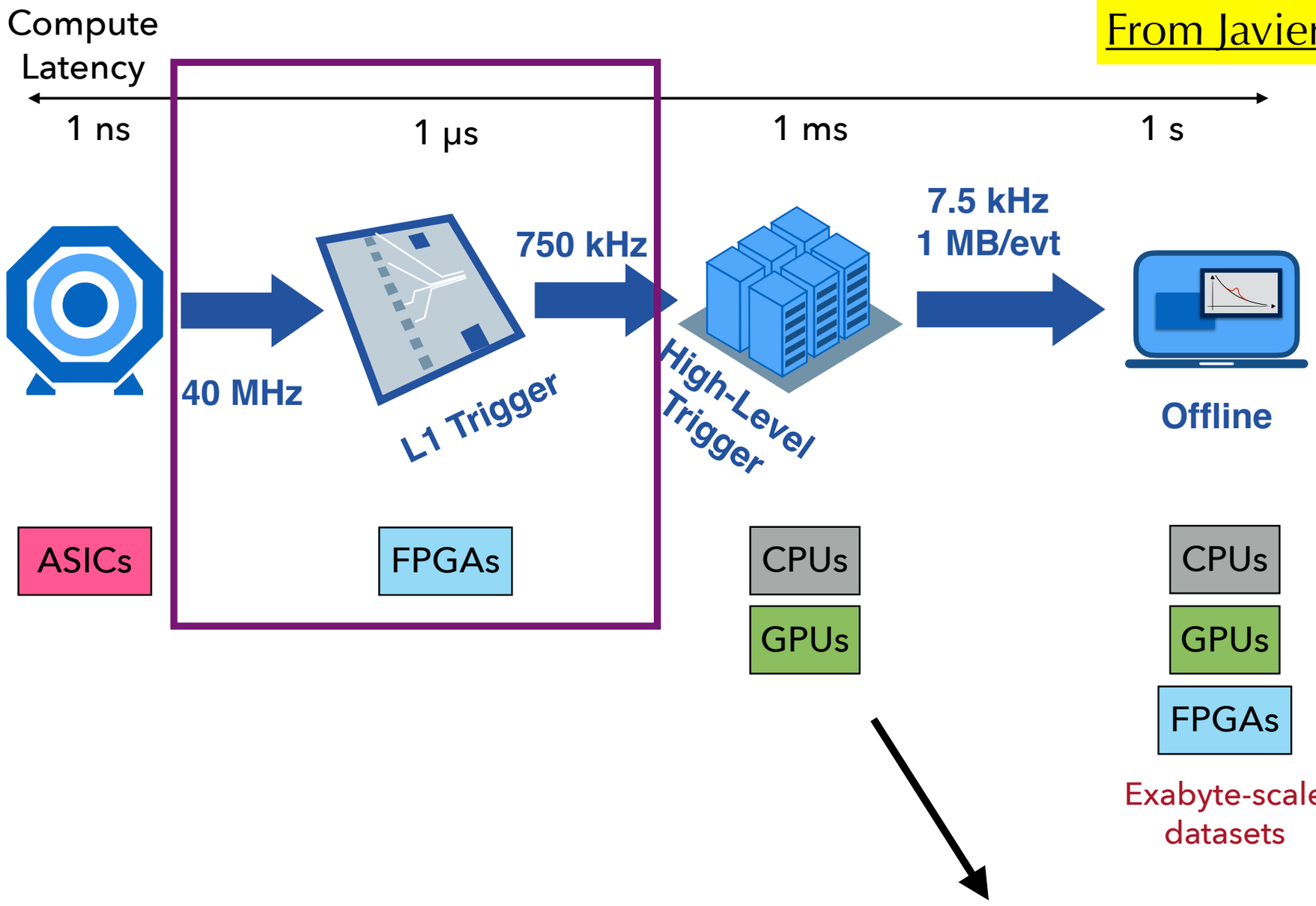


First results from MLPF in standalone setup promising: TENSORFLOW implementation and exported to ONNX

See CHEP 2023 this May for full results in CMS software stack!

ML for triggering

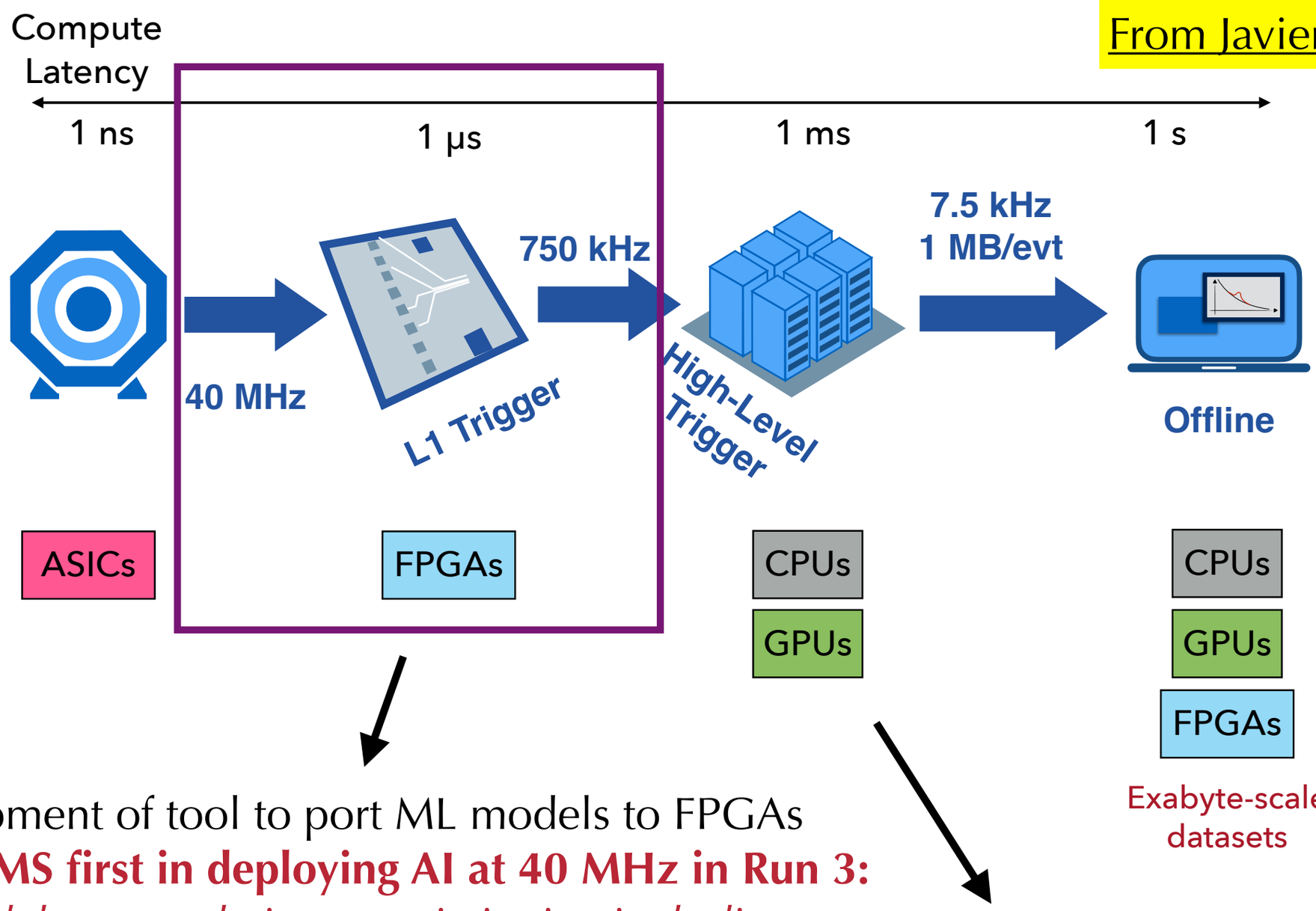
From Javier's talk on Tuesday



Thanks to advances in computing tools/infrastructures **DL possible in CMS HLT since Run 2 & SOA models like graph-based tagger possible in Run 3 data taking!**

ML for triggering

From Javier's talk on Tuesday



Exabyte-scale datasets

Development of tool to port ML models to FPGAs made **CMS first in deploying AI at 40 MHz in Run 3:** *few models currently in commissioning including anomaly detection!*

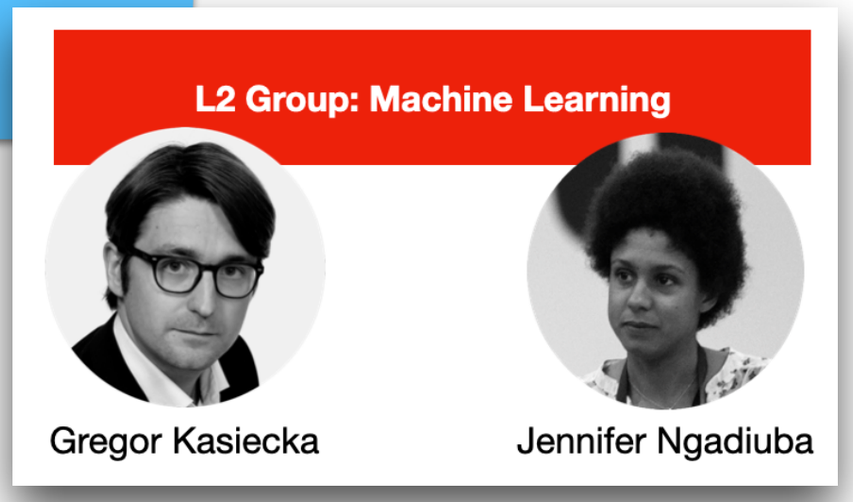
Effort also ongoing for Phase 2 developments where increased L1T system capabilities will allow to do even more!

Thanks to advances in computing tools/infrastructures **DL possible in CMS HLT since Run 2 & SOA models like graph-based tagger possible in Run 3 data taking!**

Conclusions

Continuous innovation in CMS on both algorithms and system side exploiting modern Deep Learning techniques

Push developments today and use acquired expertise in the future at HL-LHC



**Goal: enable, support,
guide and foster ML developments in CMS
computing, physics objects and physics analyses groups**

**Three subgroups coordinating multiple efforts
to achieve the goal!**

cms-conveners-ml@cern.ch

CMS

Compact Muon Solenoid

Thank you!