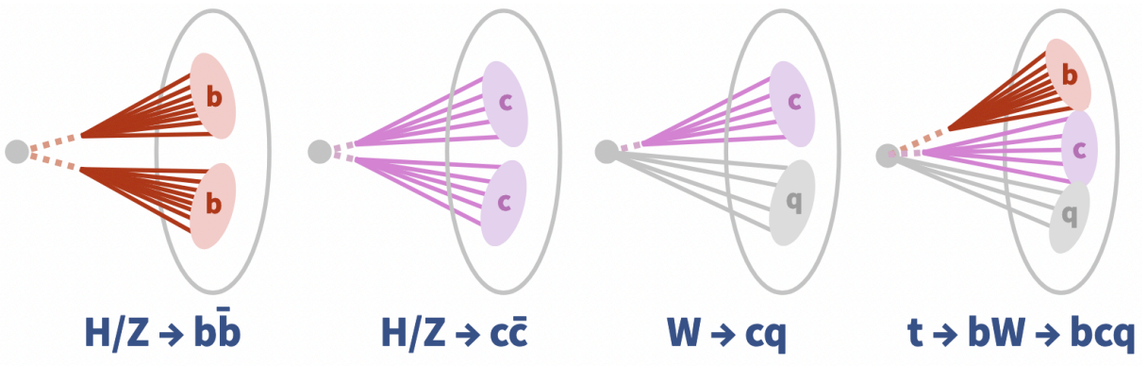
ML models re-usability: inputs from CMS

Jennifer Ngadiuba (Fermilab), Gregor Kasieczka (Universitat Hamburg)
on behalf of the CMS Collaboration

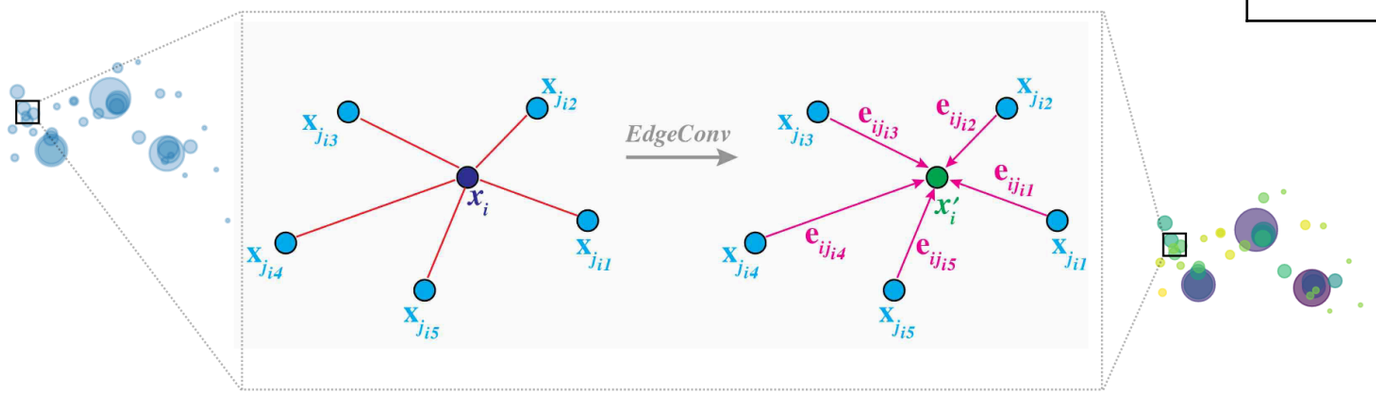
Reinterpretation of LHC results for new physics
December 12—15, 2022
CERN

ML usage in CMS: physics objects

- Widely used for **physics objects identification (or regression)** — **ex: boosted jet tagging**
  - applied to AK jets with R=0.8 — default in CMS
  - identification of jets arising from hadronic decay of boosted W/Z/H/t — common signature of large variety of BSM models
- Latest model (ParticleNet) uses point-cloud representation of the particles inside the jet and graph NN architecture for multi-classes classification
  - used by several CMS Run 2 analyses [ex: Search for non-res HH→4b, PRL 129 (2022) 081802 or VH→cc search, HIG-21-008]



	Higgs			Z			W	Top
	bb	cc	4q	bb	cc	qq	qq	qqb
ParticleNet	✓	✓	✓	✓	✓	✓	✓	✓
ParticleNet-MD	✓	✓			✓		✓	



For details on the architecture see:

Phys. Rev. D 101, 056019 (2020)

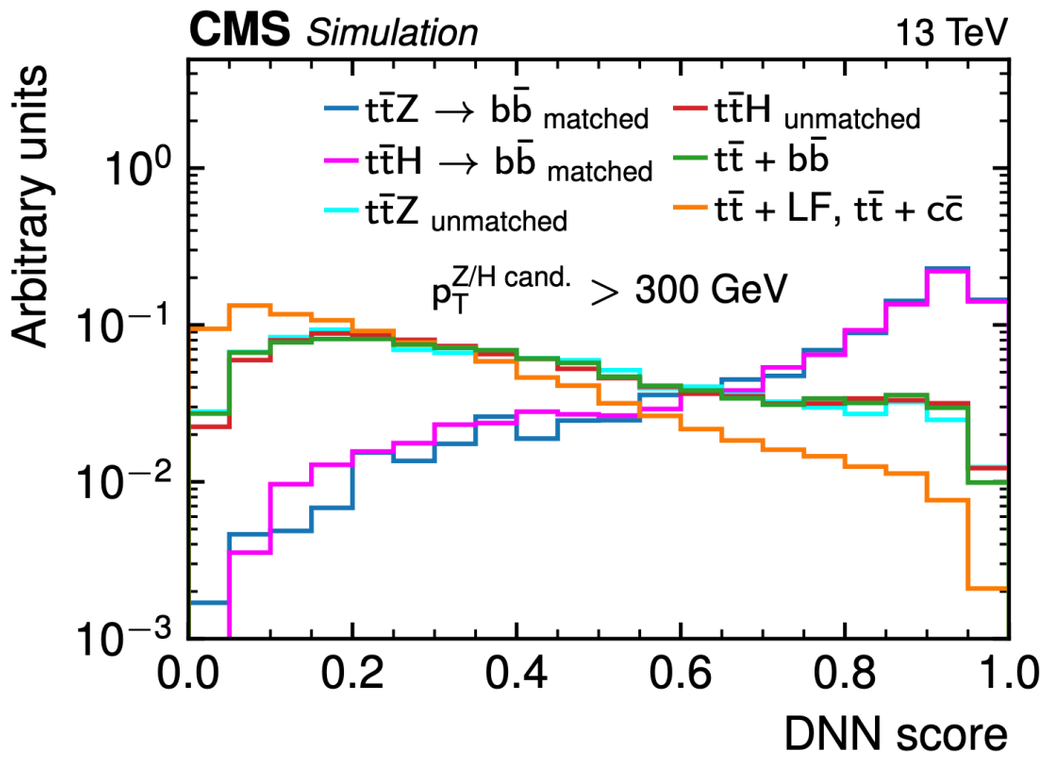
ML usage in CMS: analyses

- Standard usage in many analyses is to build a **signal versus background classifier** (could also target multiple backgrounds at once)

Ex: Search for dim-6 EFT in boosted ttH and ttZ final states

- DNN used to separate ttZ/ttH from tt+jets
- 50 input variables, 2 dense layers, 3 output nodes [one for signals and two for backgrounds]
- Signal extraction performed over multiple bins of three discriminant observables: p_T, jet softdrop mass and DNN score

CMS-TOP-21-003
Submitted to PRD



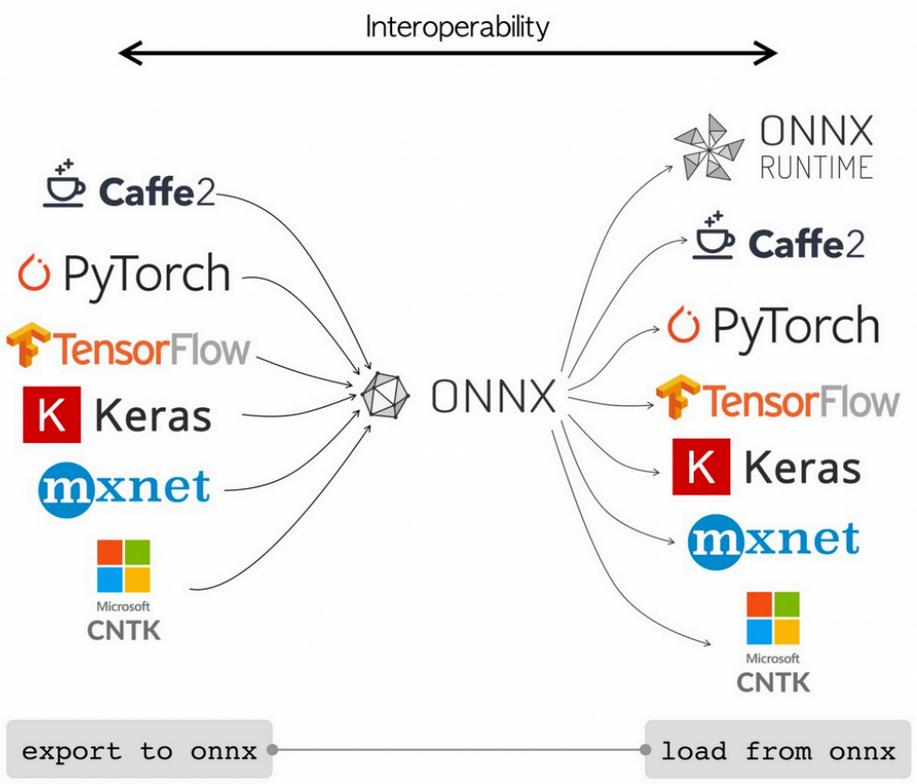
Name	Description
$t\bar{t}$ system	
b p_T	p_T of the leading (subleading) b jet
b score	DeepCSV score of the leading (subleading) b jet
q p_T	p_T of the leading (subleading) non-b jet
q score	DeepCSV score of the leading (subleading) non-b jet
$\Delta R(b,q)$	minimum ΔR between the leading (subleading) b jet and any non-b jet
$\Delta R(q,q)$	ΔR between the non-b jets closest and next-to-closest to the leading (subleading) b jet
$m(q+q)$	invariant mass of the non-b jets closest and next-to-closest to the leading (subleading) b jet
$\Delta R(b,q+q)$	ΔR between the leading (subleading) b jet and the sum of the nearest and next-to-nearest non-b jets
$m(b+q+q)$	invariant mass of the leading (subleading) b jet and the nearest and next-to-nearest non-b jets
$\Delta R(Z/H,b+q+q)$	ΔR between the Z or Higgs boson candidate and the sum of the leading (subleading) b jet and the non-b jets nearest and next-to-nearest to the leading (subleading) b jet
$\Delta R(Z/H,b+b+q+q+\ell)$	ΔR between the Z or Higgs boson candidate and the sum of the leading and subleading b jets, the non-b jets nearest and next-to-nearest to the leading (subleading) b jet, and the lepton
$m_T(b+\ell,\vec{p}_T^{miss})$	transverse mass of the subleading b jet and the lepton
$m(Z/H+b)$	invariant mass of the Z or Higgs boson candidate and the nearest b jet
$m(b+b)$	invariant mass of the leading and subleading b jets
$\Delta R(b,b)$	ΔR between the leading and subleading b jets
$\Delta R(Z/H,q)$	ΔR between the Z or Higgs boson candidate and the leading non-b jet
$\Delta R(Z/H,b)$	ΔR between the Z or Higgs boson candidate and the leading b jet
$\Delta R(Z/H,\ell)$	ΔR between the Z or Higgs boson candidate and the lepton
$m(Z/H+\ell)$	invariant mass of the Z or Higgs boson candidate and the lepton
$\Delta R(b,\ell)$	ΔR between the leading (subleading) b jet and the lepton
$m(b+\ell)$	invariant mass of the leading (subleading) b jet and the lepton
$N(b_{out})$	number of b jets outside the Z or Higgs boson candidate cone ($\Delta R > 0.8$)
$N(q_{out})$	number of non-b jets outside the Z or Higgs boson candidate cone ($\Delta R > 0.8$)
Event topology	
N(AK8 jets)	number of AK8 jets, including the Z or Higgs boson candidate
N(AK4 jets)	number of AK4 jets
N(Z/H)	number of AK8 jets with a minimum AK8 $b\bar{b}$ tagger score of 0.8
AK8 m_{SD}	maximum m_{SD} of AK8 jets, excluding the Z or Higgs boson candidate
$H_T(b_{out})$	H_T of the b jets outside the Z or Higgs boson candidate cone ($\Delta R > 0.8$)
$H_T(b_{out},q_{out},\ell)$	H_T of all AK4 jets outside the Z or Higgs boson candidate cone ($\Delta R > 0.8$) and the lepton
sphericity	sphericity calculated from the AK4 jets and the lepton [**?**]
aplanarity	aplanarity calculated from the AK4 jets and the lepton [**?**]
Z or Higgs boson candidate substructure	
b_{in} score	maximum (minimum) DeepCSV score of AK4 jets within the Z or Higgs boson candidate cone ($\Delta R < 0.8$)
$\Delta R(b_{in},b_{out})$	ΔR between one b jet within the Z or Higgs boson candidate cone ($\Delta R < 0.8$) and the leading b jet outside of the Z or Higgs boson candidate cone
$N(b_{in})$	number of b jets within the Z or Higgs boson candidate cone ($\Delta R < 0.8$)
$N(q_{in})$	number of non-b jets within the Z or Higgs boson candidate cone ($\Delta R < 0.8$)
Z/H $b\bar{b}$ score	AK8 $b\bar{b}$ tagger score of the Z or Higgs boson candidate

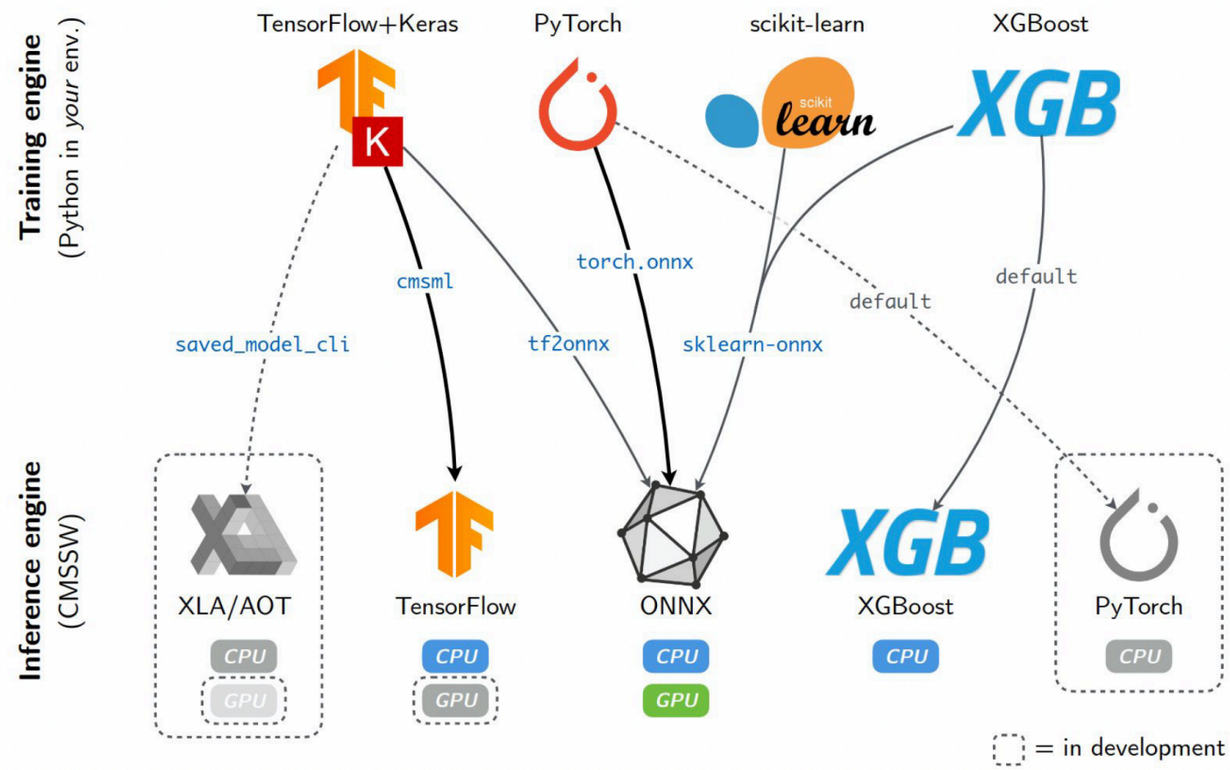
ML models preservation in CMS

- For **centrally-supported ML models** (i.e. for physics objects) the full model exists in the central CMS software release used to produce the datasets and MC samples used by the analyses

- CMSSW supports several model implementations — **ex: ParticleNet uses ONNX** [1,2]

ONNX is an open format built to represent machine learning models. It is designed to improve interoperability across a variety of frameworks and platforms in the AI tools community—most deep learning frameworks (e.g. XGBoost, TensorFlow, PyTorch which are frequently used in CMS) support converting their model into the ONNX format or loading a model from an ONNX format.





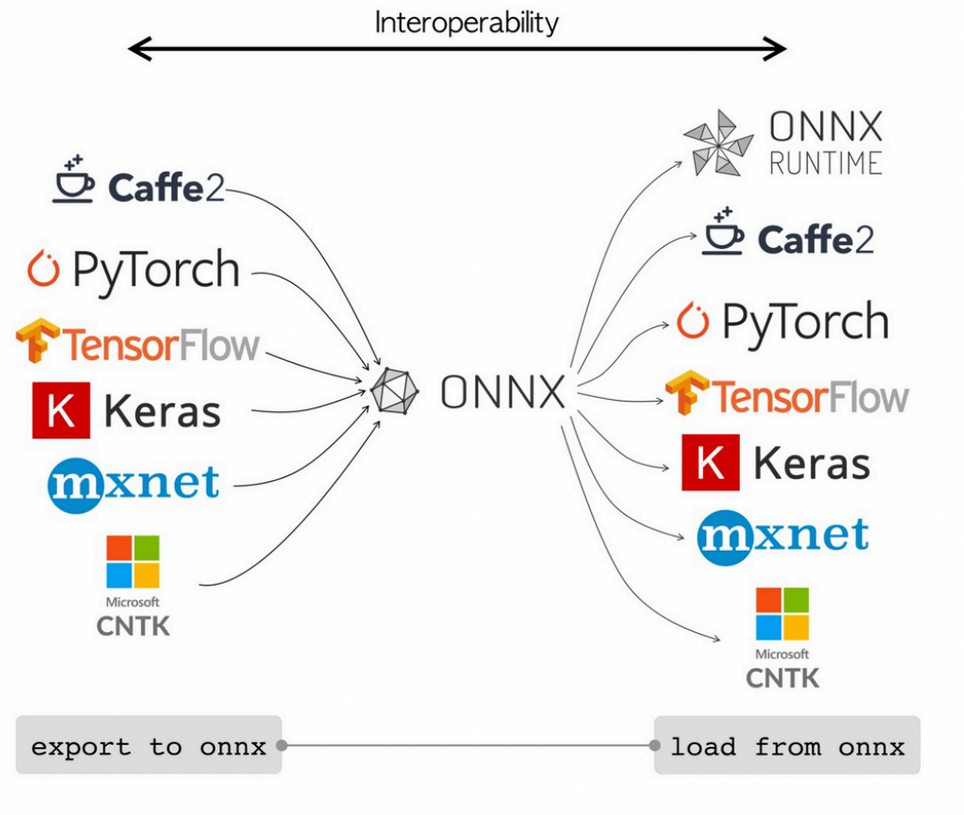
More details in our CMS ML living documentation at:
https://cms-ml.github.io/documentation/
[fully public]

ML models preservation in CMS

- For **centrally-supported ML models** (i.e. for physics objects) the full model exists in the central CMS software release used to produce the datasets and MC samples used by the analyses

- CMSSW supports several model implementations — **ex: ParticleNet uses ONNX** [1,2]

- While preservation is handled properly with **github repositories fully public**, the accessibility of the useful information is currently very limited



- **Ways to improve this could be:**

  1) *publish trained model in ONNX format on Zenodo with the required documentation*

  2) *publish ML model efficiencies*

- Option 2) is already in place but not structured and most of the time not well accessible

- Option 1) is something new that should be pursed but there are caveats

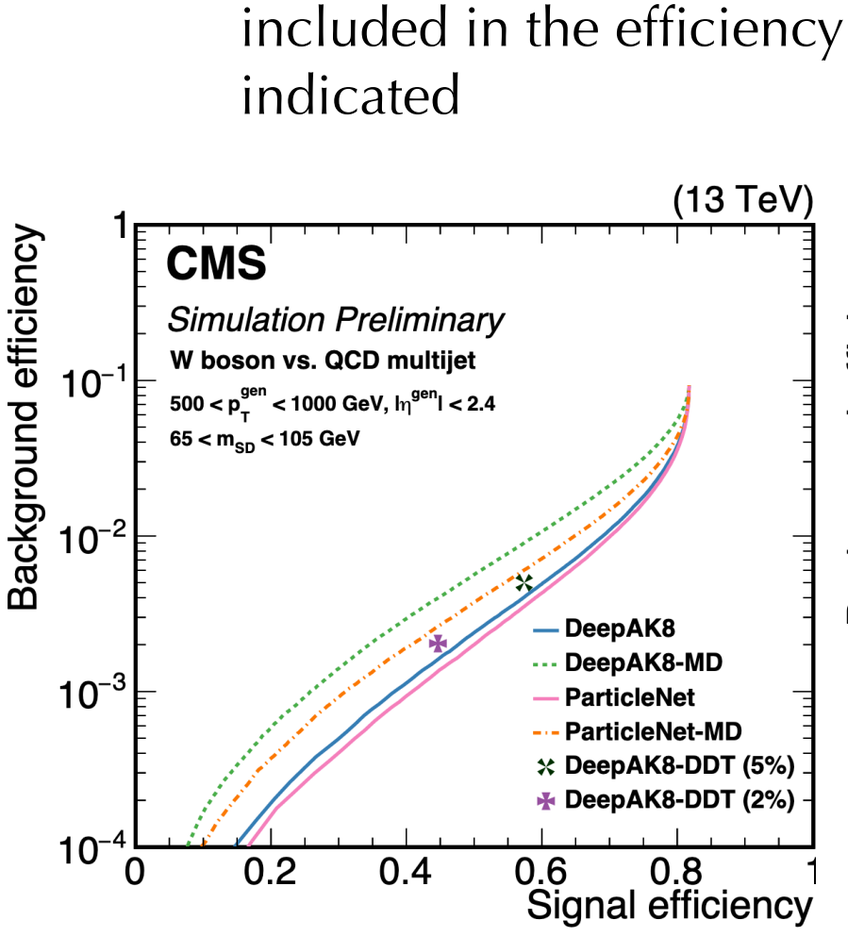
ML models reusability

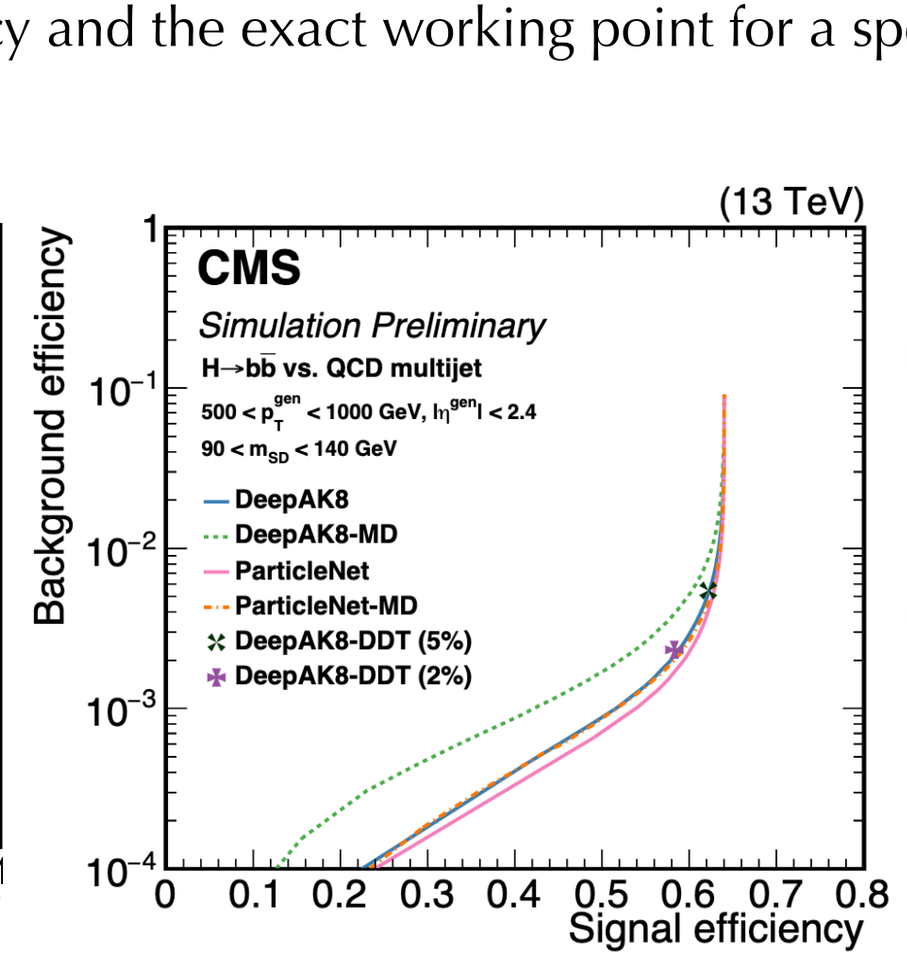
- **Publishing the trained ML model with dedicated documentation can be done** and would be certainly important for preservation motives
  - significant effort still needed that people have to be willing to take
- The **question remain on the reusability** of inference-only information, i.e. without the associated training code and data
  - training code typically lives in someone's (the developer) repository
  - data are only made available to the public as CMS Open Data after several years [together with the associated CMSSW ML inference code and weights]
- **Training code might not be needed** if documentation is clear and descriptive enough
  - thinking in particular about input data preprocessing — however ONNX enables to store more than the ML model itself
- While **applying a ML model pre-trained on CMS-reconstructed data on privately produced signal simulation** (ex, with Delphes) **might be less trivial**
  - the outside user would have to also fully simulate the backgrounds and retrain from scratch

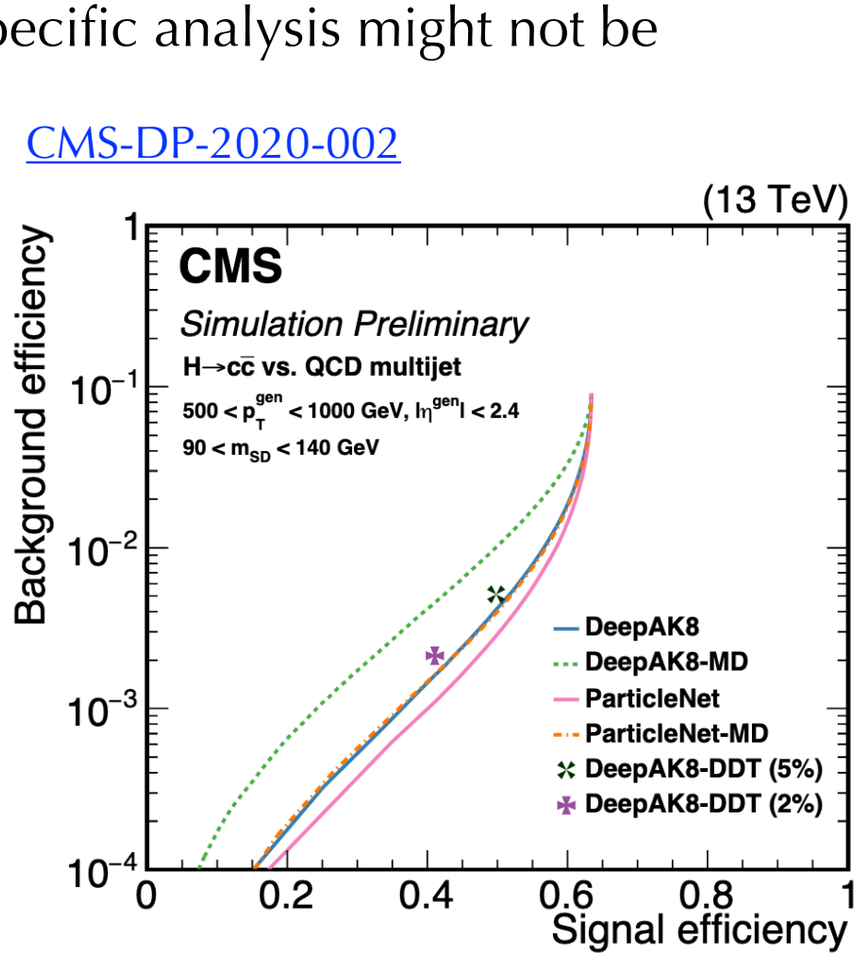
ML models reusability

- **Publishing the ML model efficiency for physics objects** might be more straightforward for reusability although not fully structured in CMS
- Taking again the **ParticleNet example**: *if you know where to look* you will find the CMS Detector Note in CDS with some efficiency plots
- And even if you know where to look it does not necessarily mean that what you need is contained in that document
  - ex, the ParticleNet efficiency for H→4q are not there, the efficiency for the jet mass cut is included in the efficiency and the exact working point for a specific analysis might not be indicated

CMS-DP-2020-002





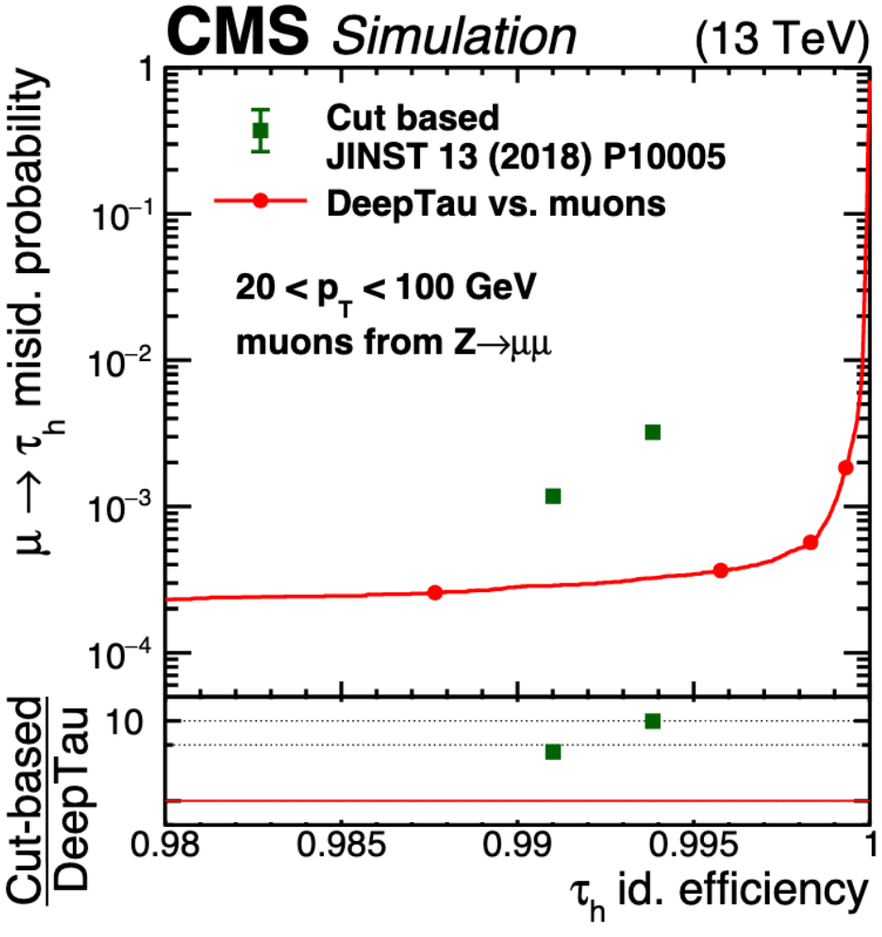


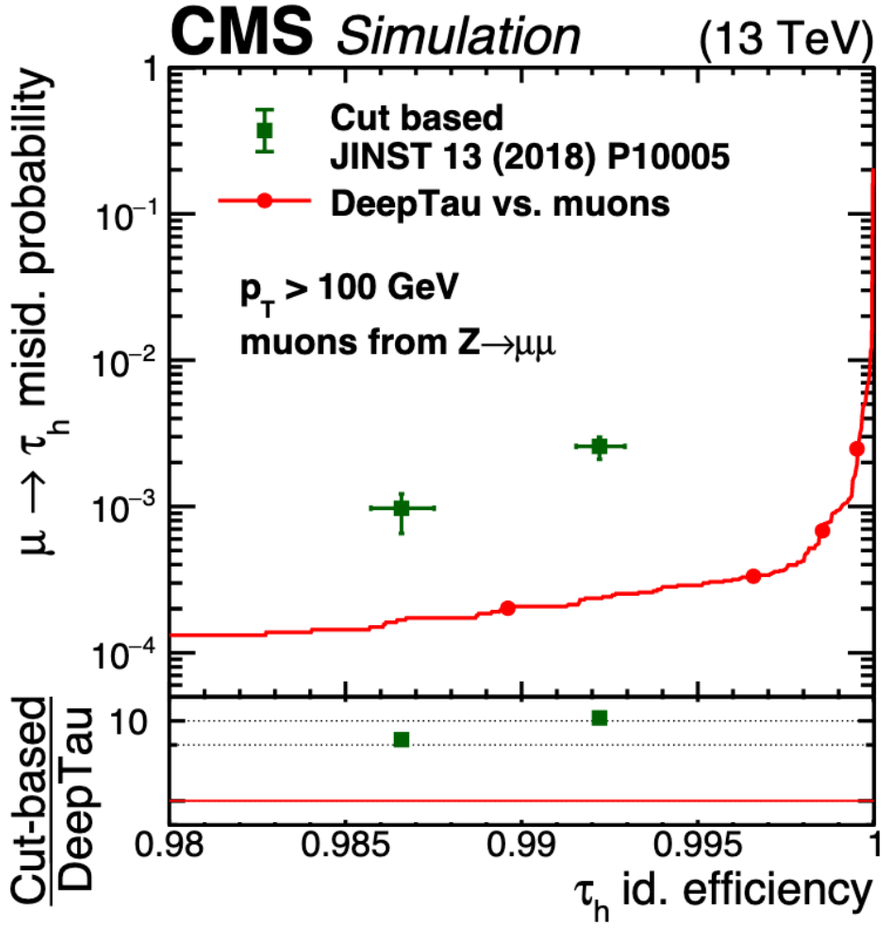
ML models reusability

- **Publishing the ML model efficiency for physics objects** might be more straightforward for reusability although not fully structured in CMS

- A better **example from the DeepTau case**: particle-based CNN to discriminate real τ_h decays from QCD jets

  - technical publication in JINST with associated HEP Data of ROC curves and maps of efficiency and mistag rates versus p_T and η





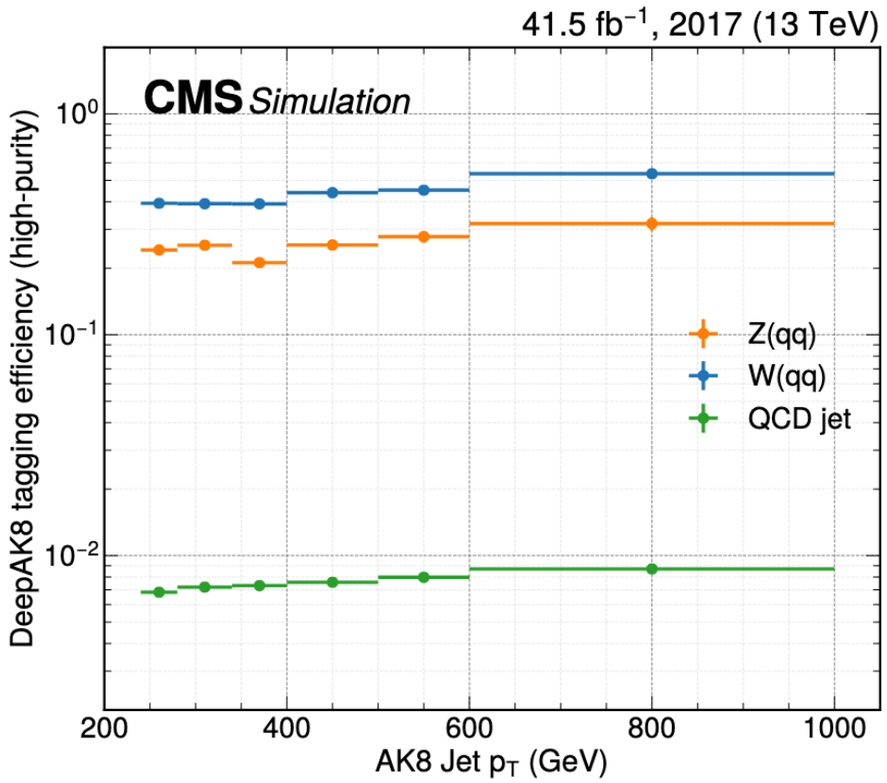
ML models reusability

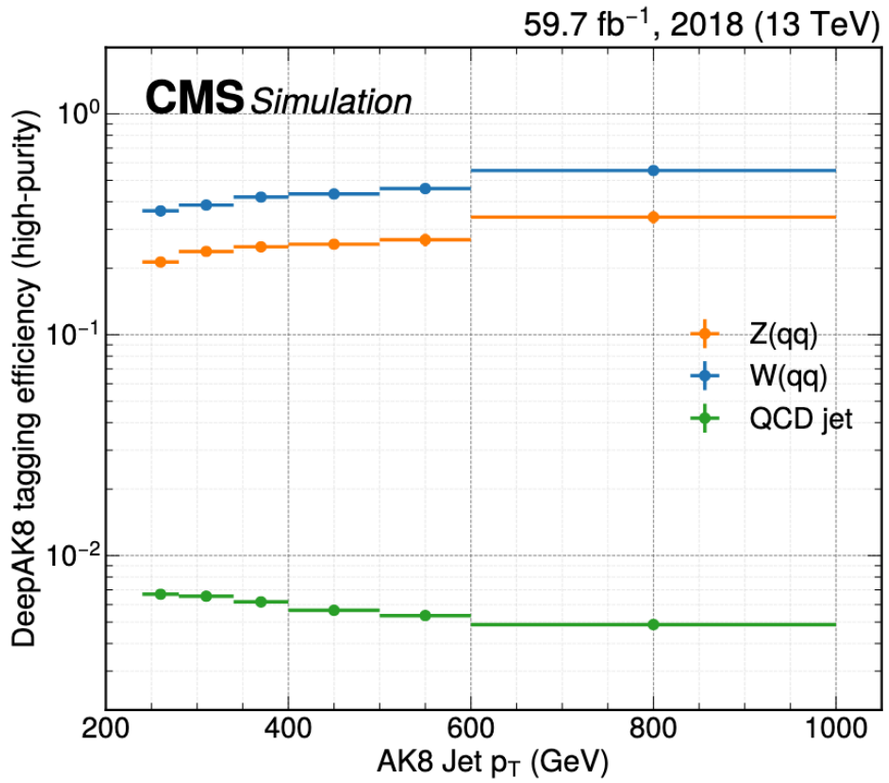
- **Publishing the ML model efficiency for physics objects** might be more straightforward for reusability although not fully structured in CMS

- See also an interesting example from a mono-jet dark matter search in CMS EXO

  - case in CMS for a general effort (i.e. not strictly ML related) to aid re-interpretation of analysis through HEP data and MADANALYSIS

  - includes also generator-level matched efficiencies in HEP data for the ML-based boosted jet tagging model



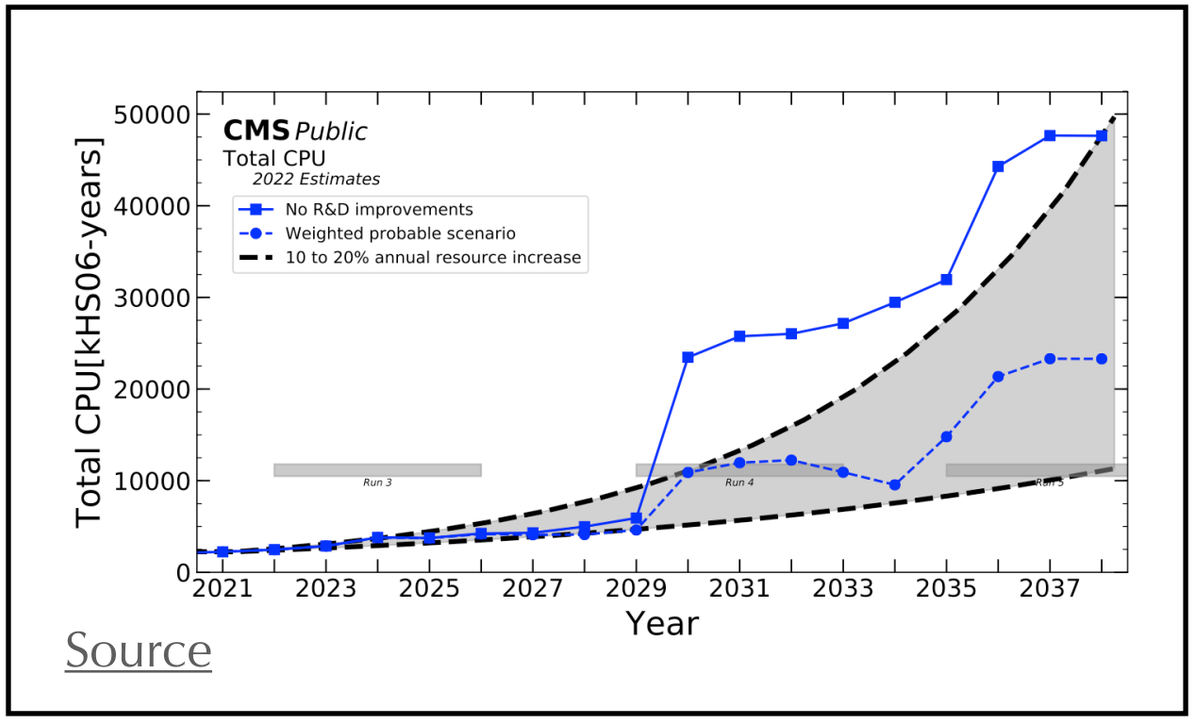


ML models reusability: analyses

- For the **standard usage in analyses most of what discussed applies in a similar way**
- However, this is even less trivial as custom (not CMS central) ML models are typically developed for these cases
- At the same time, **most cases use the DNN score in the likelihood** such that such application can be handled in a separate way
- There were proposals of publishing the likelihood as DNN weights → see separate talk from the CMS statistics committee
- **No standard usage include unsupervised learning** but given that such effort, although rapidly growing, just started in LHC collaborations it is hard to have a full picture in this exact moment
  - certainly we'll need to think about!

Future usage of ML models

- To cope with growing computing demands a large effort being put to **replace event generation and GEANT4 simulation with surrogate ML-based models**
- Different **generative models** being explored:
  - Variational Autoencoder
  - Generative Adversarial Network
  - Normalizing Flows
- And different **data representations:**
  - images and point cloud
- In CMS nothing in production yet (ATLAS already deploys GANs for Run 3 FastSim)



Source

- **Opportunities to increase data reusability by publishing full or partial detector simulation as compact ML models**

See recent mini-workshop by CERN IT on
"Foundation models and fast detector simulation"

Conclusions

- With the rapidly increasing usage of ML models for physics analyses it is the right time to bring the discussion about their preservation and reusability as a priority in LHC Collaborations

- We can take the problem step by step starting from the most standard usage of ML

- CMS has ways implemented although not fully optimal in terms of accessibility nor fully structured throughout the collaboration

- As substantial effort would be needed to improve it, feedbacks from the users would be beneficial to fully understand how to prioritise and channel the efforts!

- Can take inspiration from independent efforts on structuring the principles that a reusable AI model should follow

  - see for example the FAIR for AI project (Findable, Accessible, Interoperable, and Reusable)