INNOVATIVE TECHNIQUES FROM CMS: FROM SCOUTING TO MACHINE LEARNING

Plus: W mass!

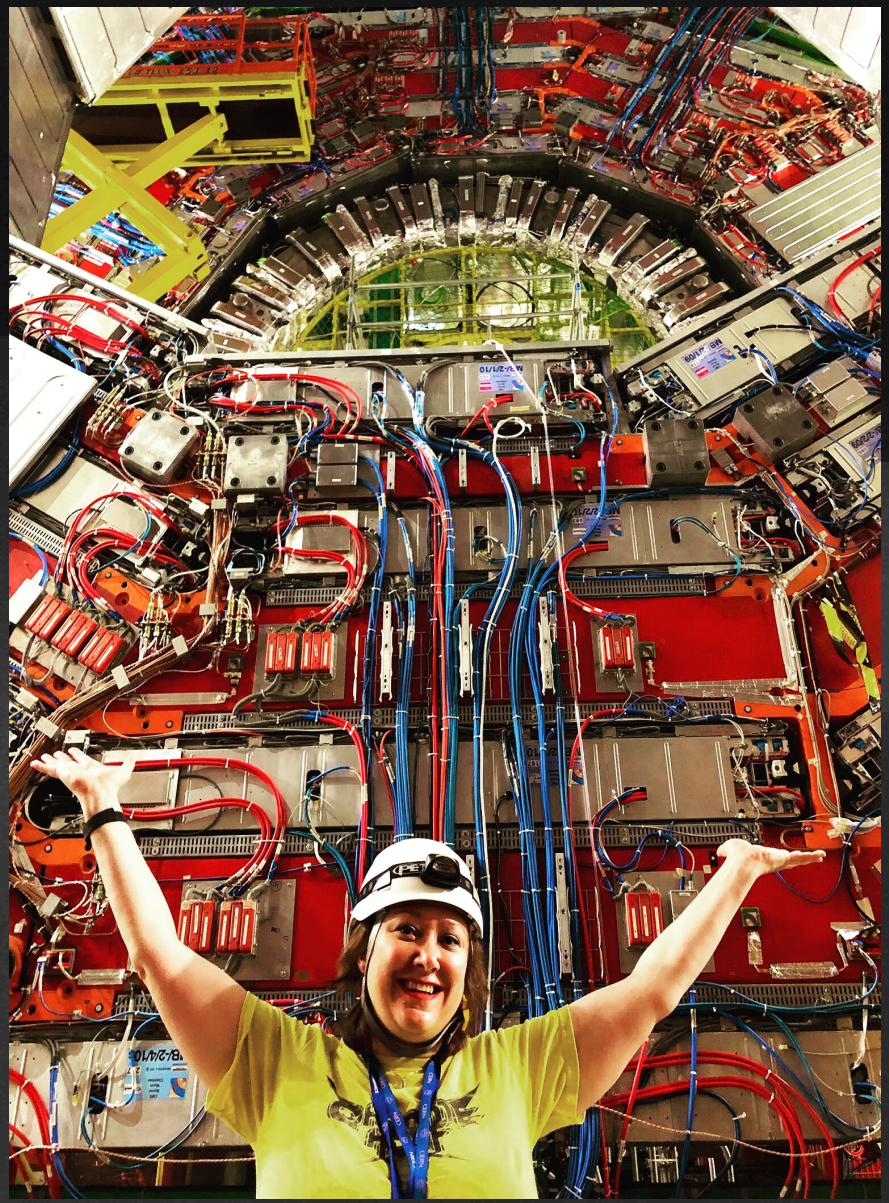
Freya Blekman

NNPDF Collaboration Meeting

23 September 2024

Disclaimer: heavily based on some recent CMS overview talks (e.g. from ICHEP, last week's seminar)

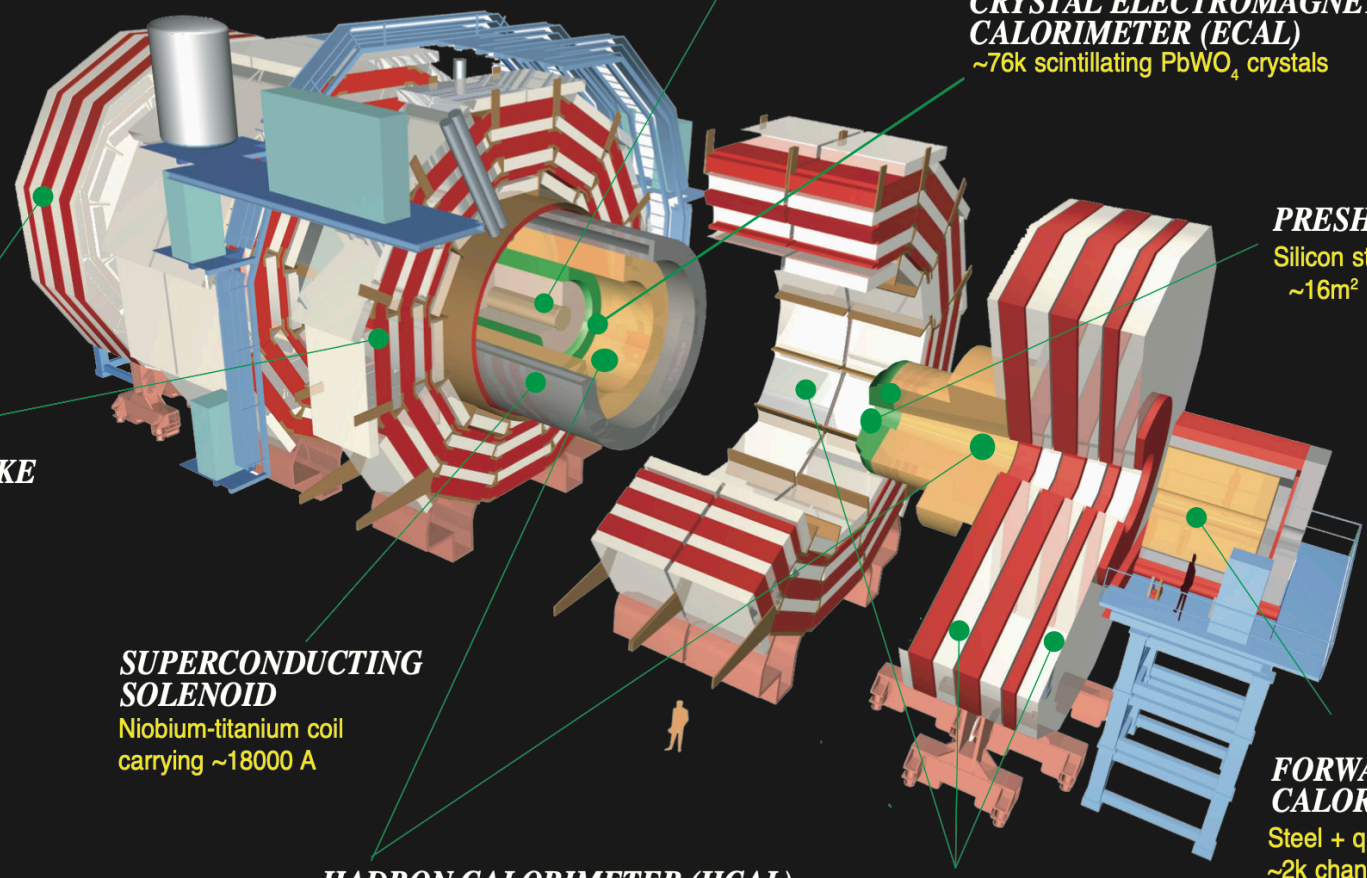
socials: freyablekman



socials: freyablekman

CMS Detector

Pixels
Tracker
ECAL
HCAL
Solenoid
Steel Yoke
Muons



SILICON TRACKER
Pixels (100 x 150 μm²)
~1m² ~66M channels
Microstrips (80-180μm)
~200m² ~9.6M channels

CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76k scintillating $PbWO_4$ crystals

PRESHOWER
Silicon strips
~16m² ~137k channels

STEEL RETURN YOKE
~13000 tonnes

SUPERCONDUCTING SOLENOID
Niobium-titanium coil
carrying ~18000 A

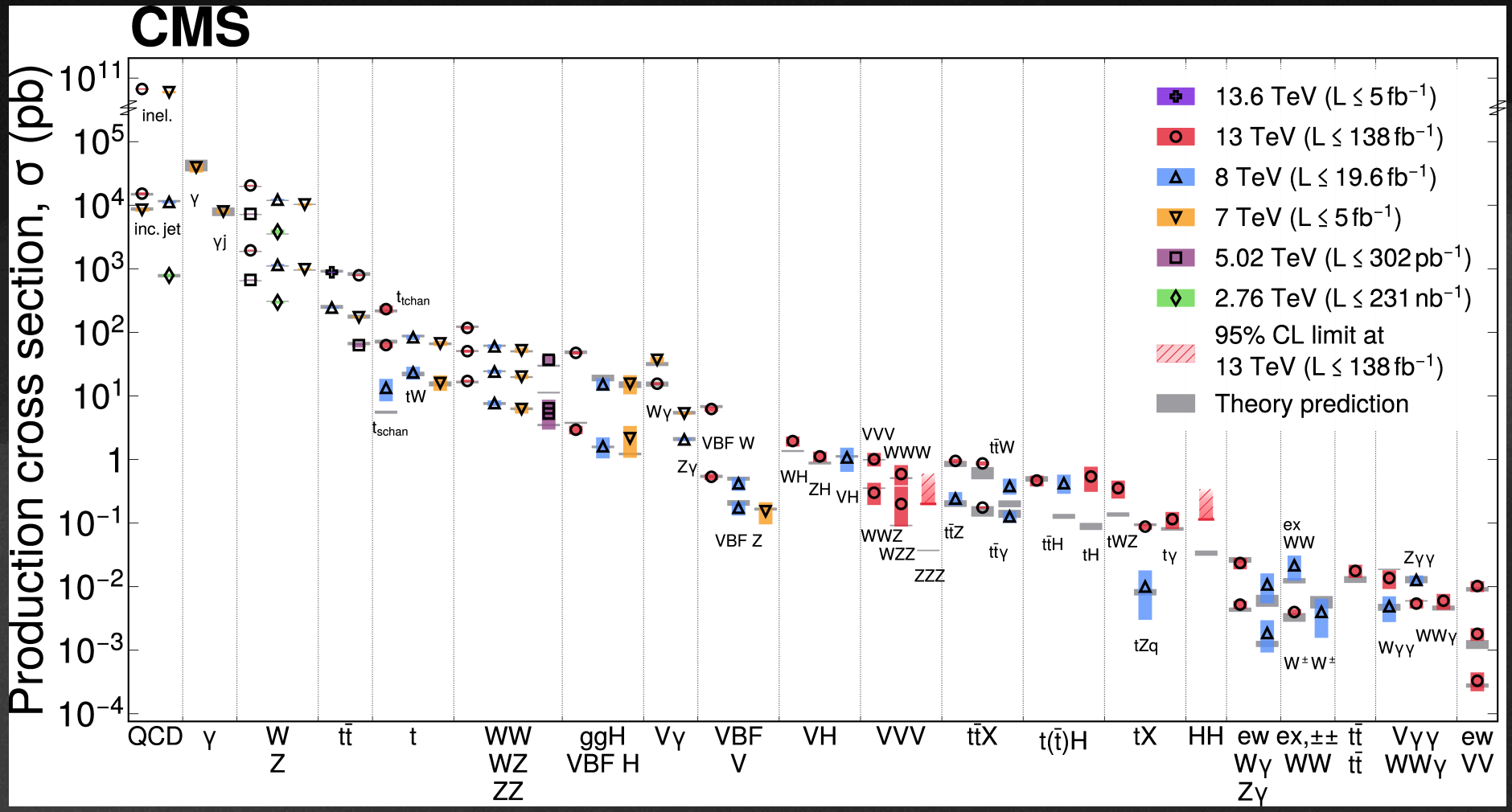
HADRON CALORIMETER (HCAL)
Brass + plastic scintillator
~7k channels

FORWARD CALORIMETER
Steel + quartz fibres
~2k channels

MUON CHAMBERS
Barrel: 250 Drift Tube & 480 Resistive Plate Chambers
Endcaps: 473 Cathode Strip & 432 Resistive Plate Chambers

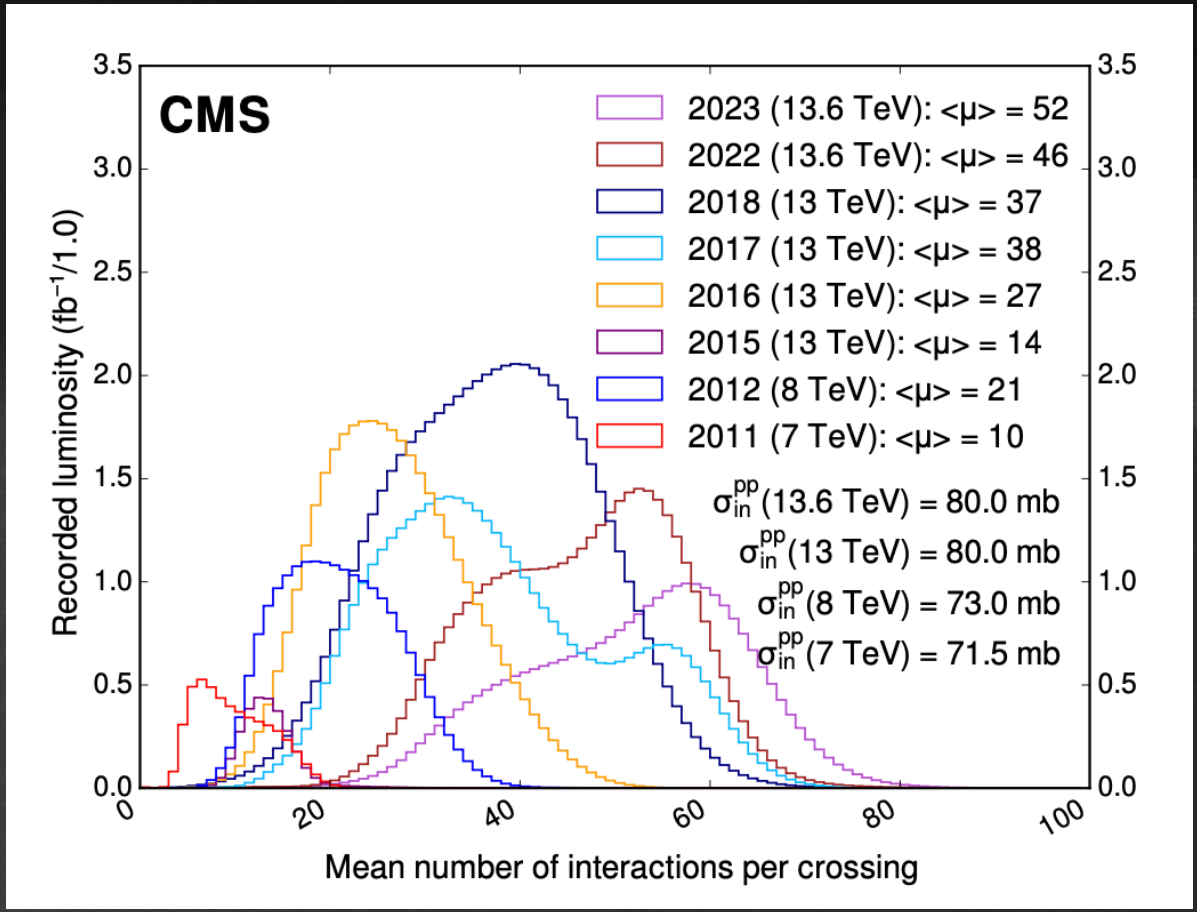
Total weight	**: 14000 tonnes**
Overall diameter	**: 15.0 m**
Overall length	**: 28.7 m**
Magnetic field	**: 3.8 T**

STANDARD MODEL AT THE LHC: ORDERS OF MAGNITUDE



Ref: CMS "stairway to discovery" SM crosss section paper SMP-23-004
arxiv:2045..18661 (Submitted to Physics Reports)

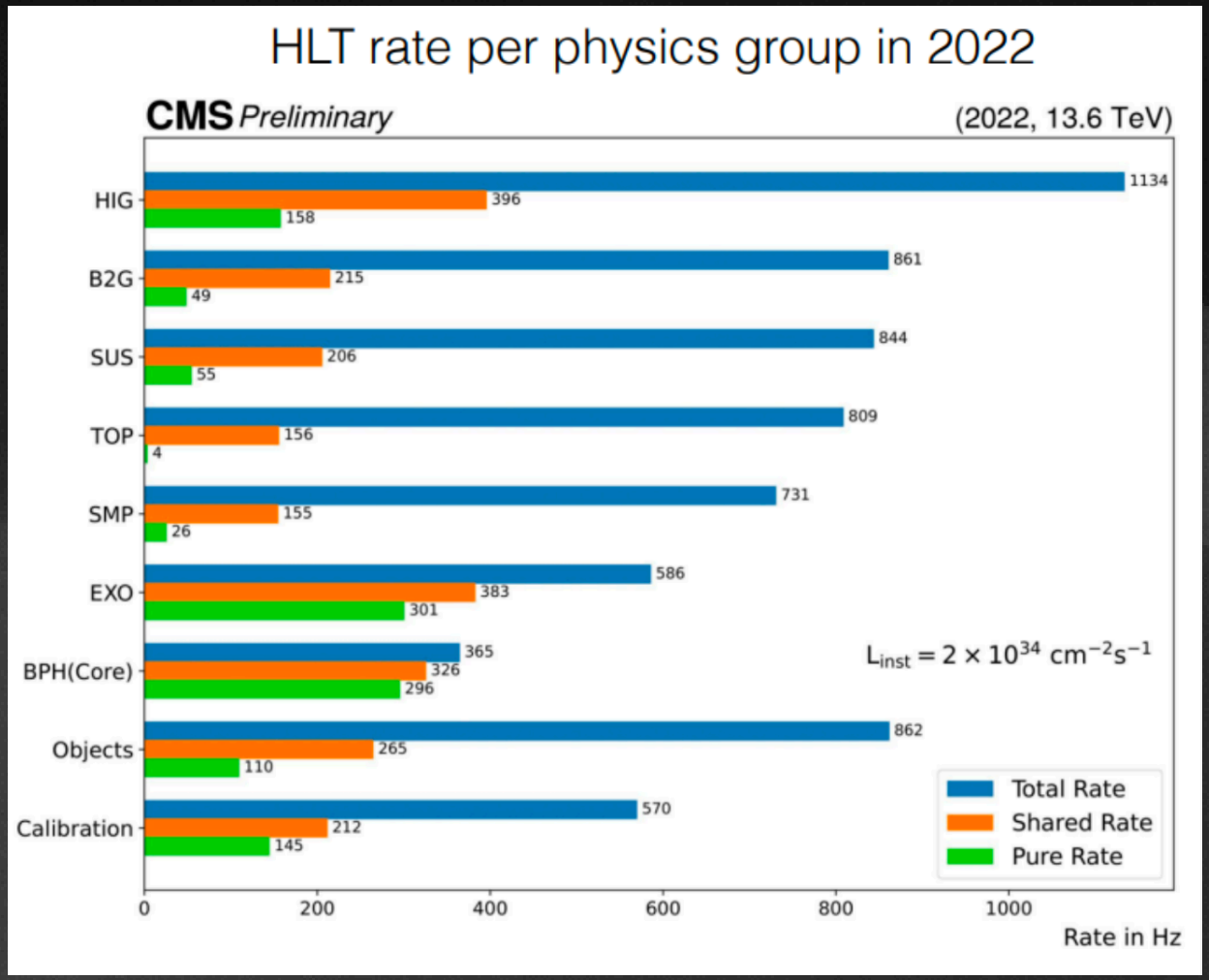
LHC PERFORMANCE OVER THE YEARS



But: we throw away most of this data!

So my apologies to you, but we are going to talk about triggers!

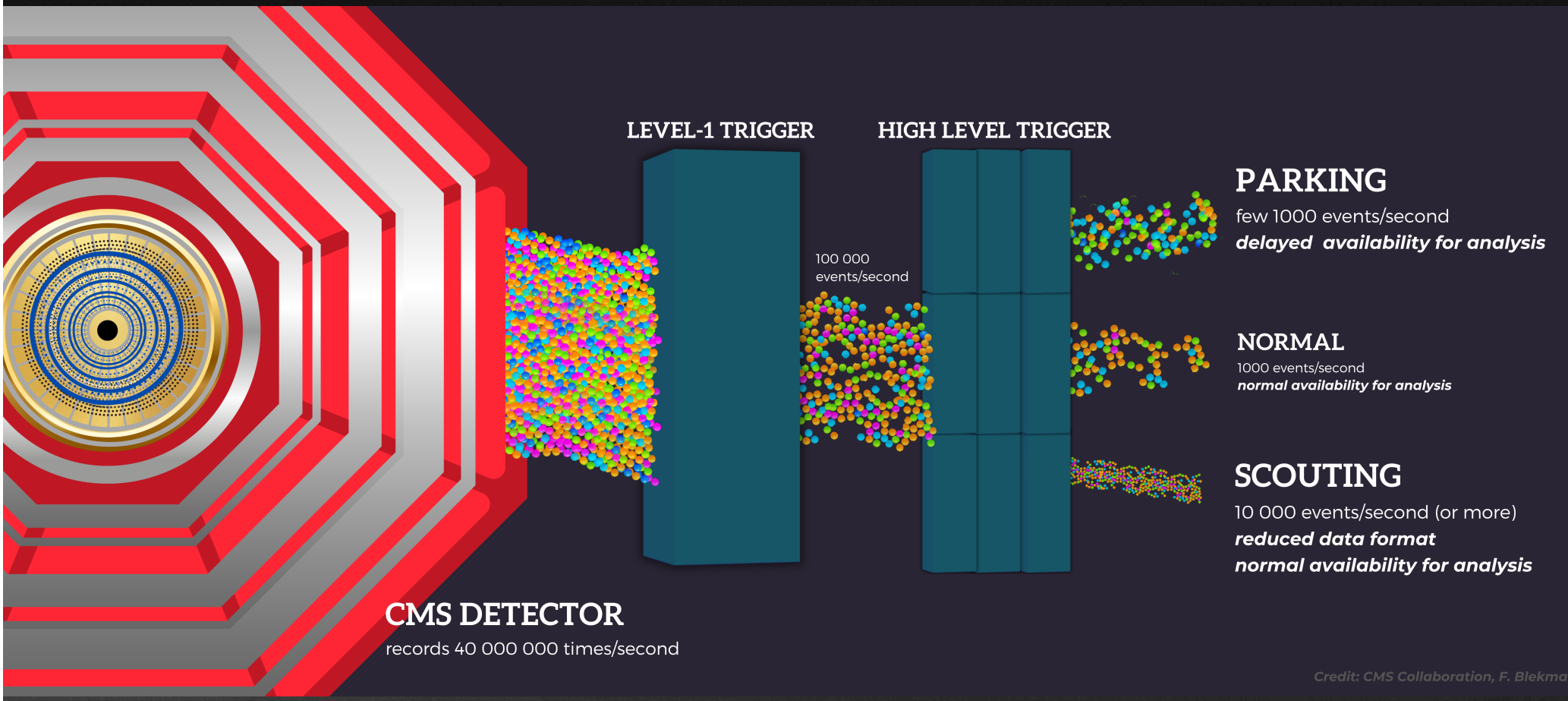
LHC PERFORMANCE OVER THE YEARS



But: we throw away most of this data!

So my apologies to you, but we are going to talk about triggers!

THE CMS TRIGGER PARKING AND SCOUTING



CMS OVERVIEW PAPER

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CMS

CERN-EP-2024-068
2024/03/26

CMS-EXO-23-007

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

The CMS Collaboration*

Abstract

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

To be submitted to Physics Reports

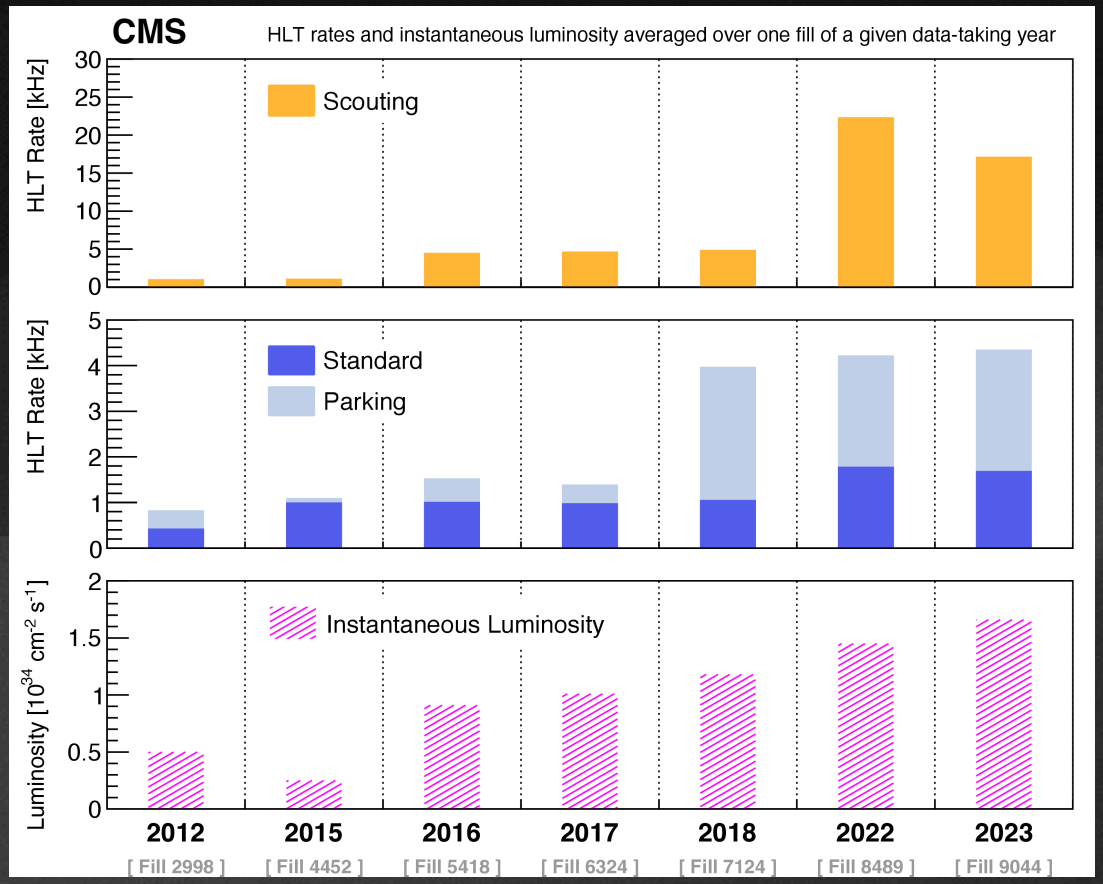
*See Appendix 8 for the list of collaboration members

https://arxiv.org/abs/2403.16134

EXO-23-007
(Accepted by Physics Reports)

it's long! I won't do it justice. Check it out if you're interested in lower pT than typical at LHC, large cross sections, displaced signatures, signatures with large backgrounds

THE DATA CONSTRAINT



LHC bunch crossing 30 MHz

'Standard" trigger cross section ~100 nb -> 1kHz

Main bottleneck rate: prompt offline reconstruction

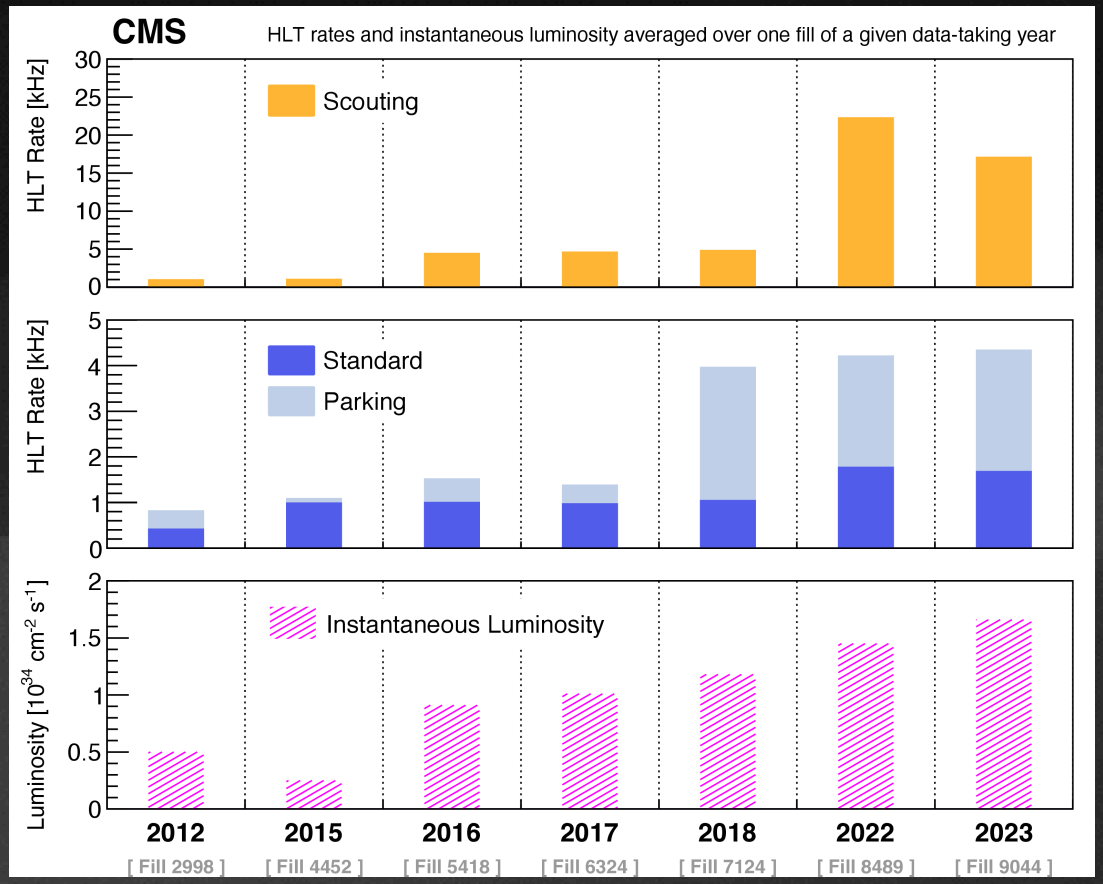
Delayed reconstruction can be used to bypass the rate limit

In 2018 CMS collected 10B events just with displaced single muon triggers that were 'parked' until later analysis

CMS (and ATLAS, "delayed reconstruction"!) has expanded strategy

socials: freyablekman

THE DATA CONSTRAINT



LHC bunch crossing 30 MHz

'Standard" trigger output ~100 nb -> 1kHz

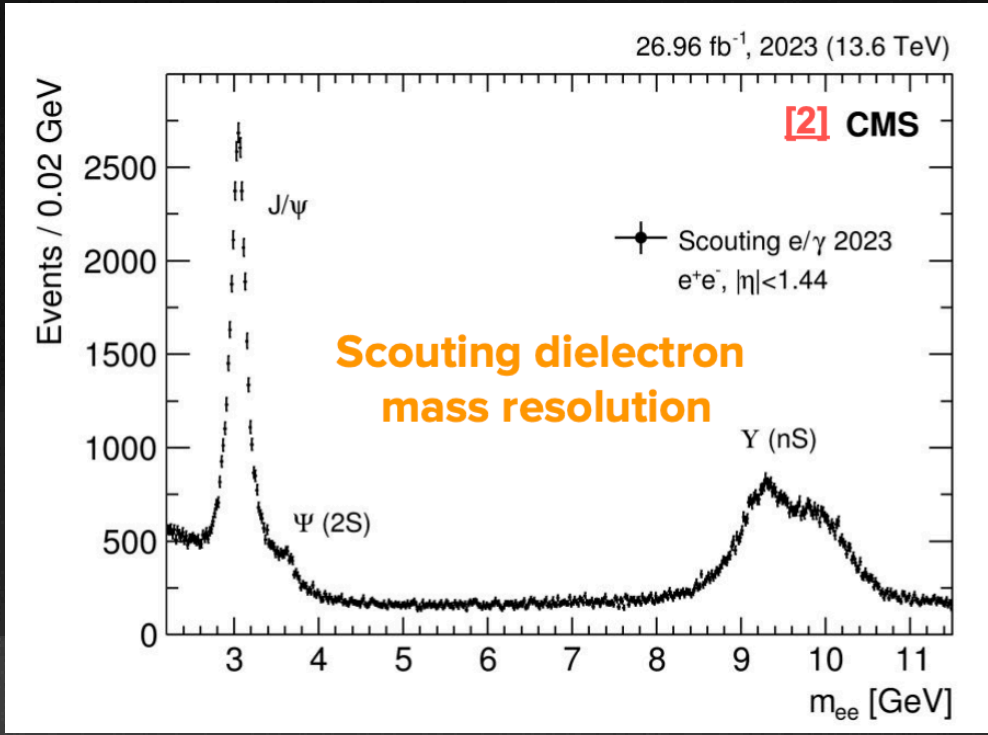
Main bottleneck rate: prompt offline reconstruction and storage

Scouting means directly saving high-quality trigger objects

Event size 10 kB/event instead of 1MB/event

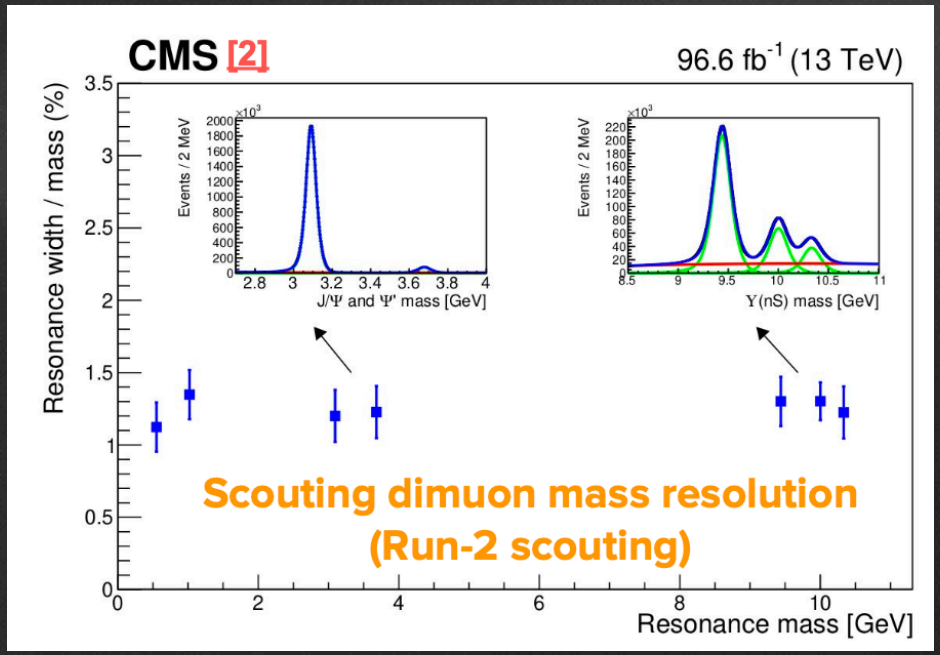
(ATLAS, "trigger level analysis". Also done by LHCb btw, 100% in HL-LHC)

SCOUTING OBJECTS

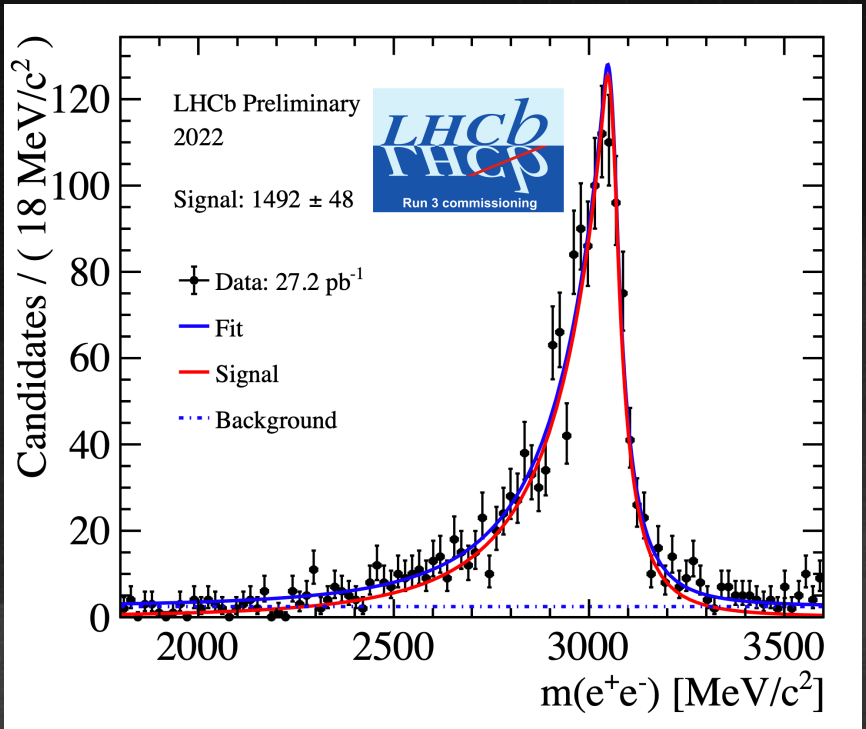


MUCH lower momentum threshold

Opens effectively a flavour-physics program a la LHCb



SCOUTING OBJECTS



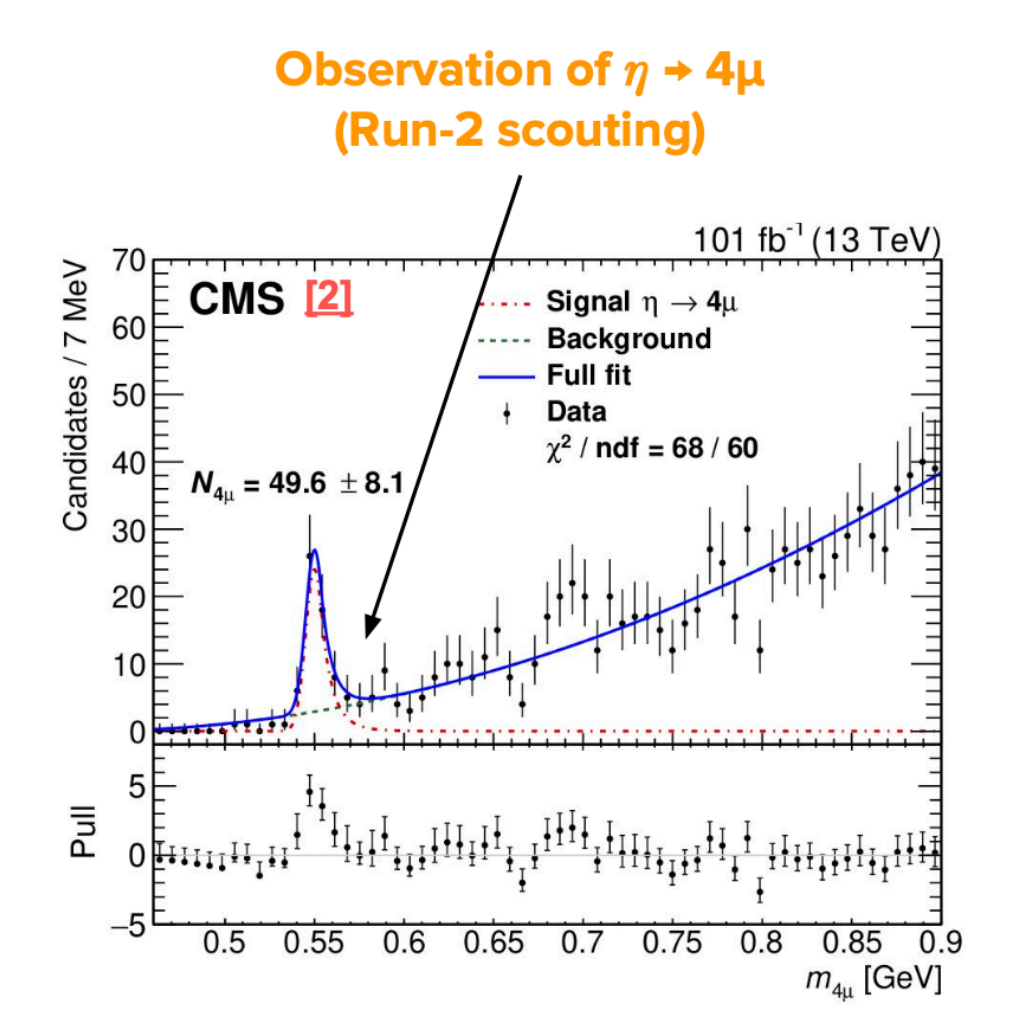
MUCH lower momentum threshold

Opens effectively a flavour-physics program a la LHCb

Including lepton ratio measurements!

Src: LHCb performance monitoring https://lbfence.cern.ch/alcm/public/figure/details/444

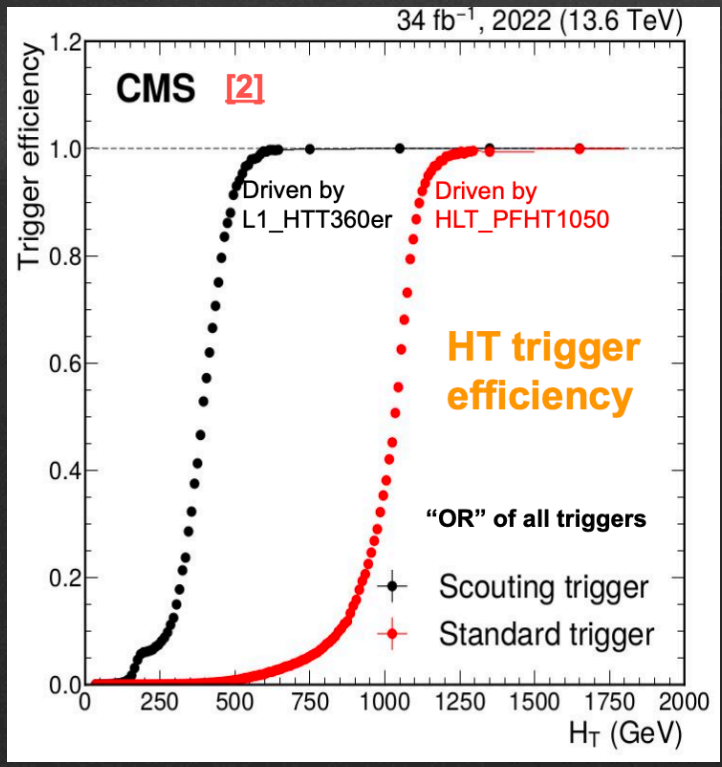
SCOUTING OBJECTS



MUCH lower momentum threshold

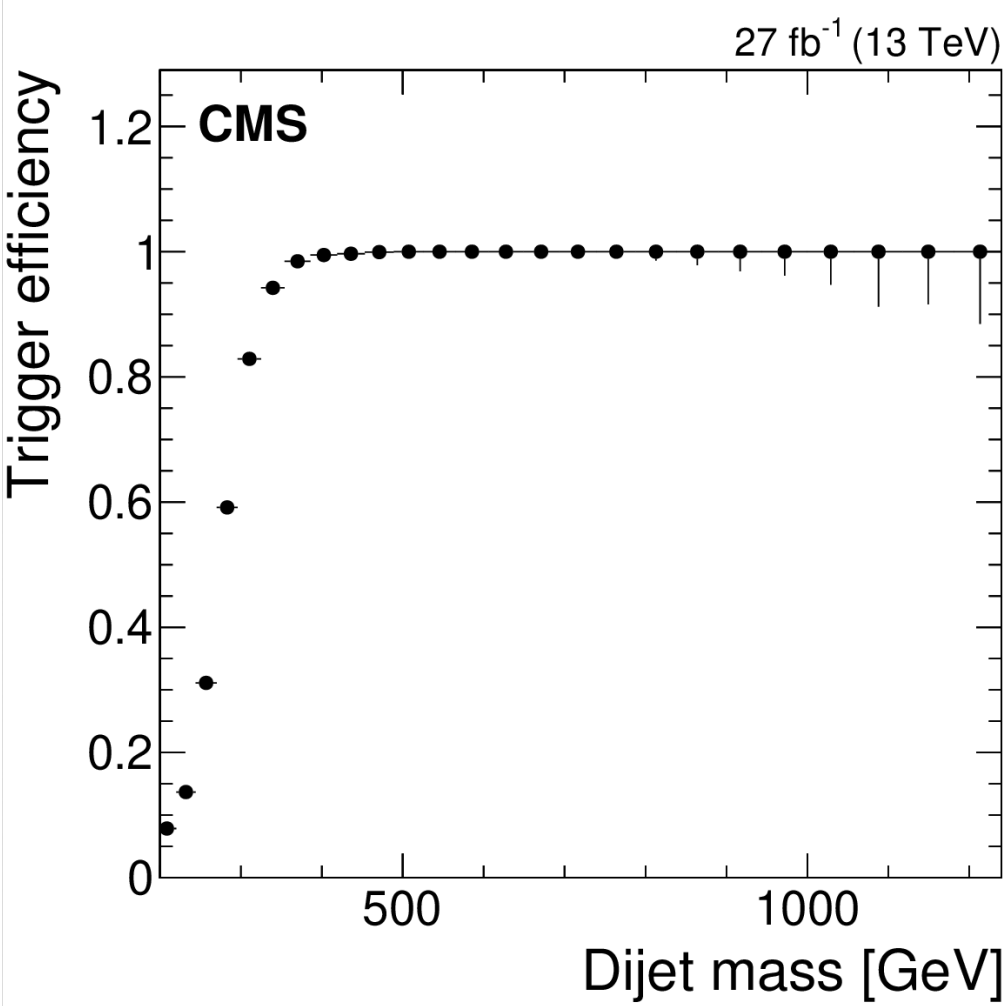
Opens effectively a flavour-physics program a la LHCb

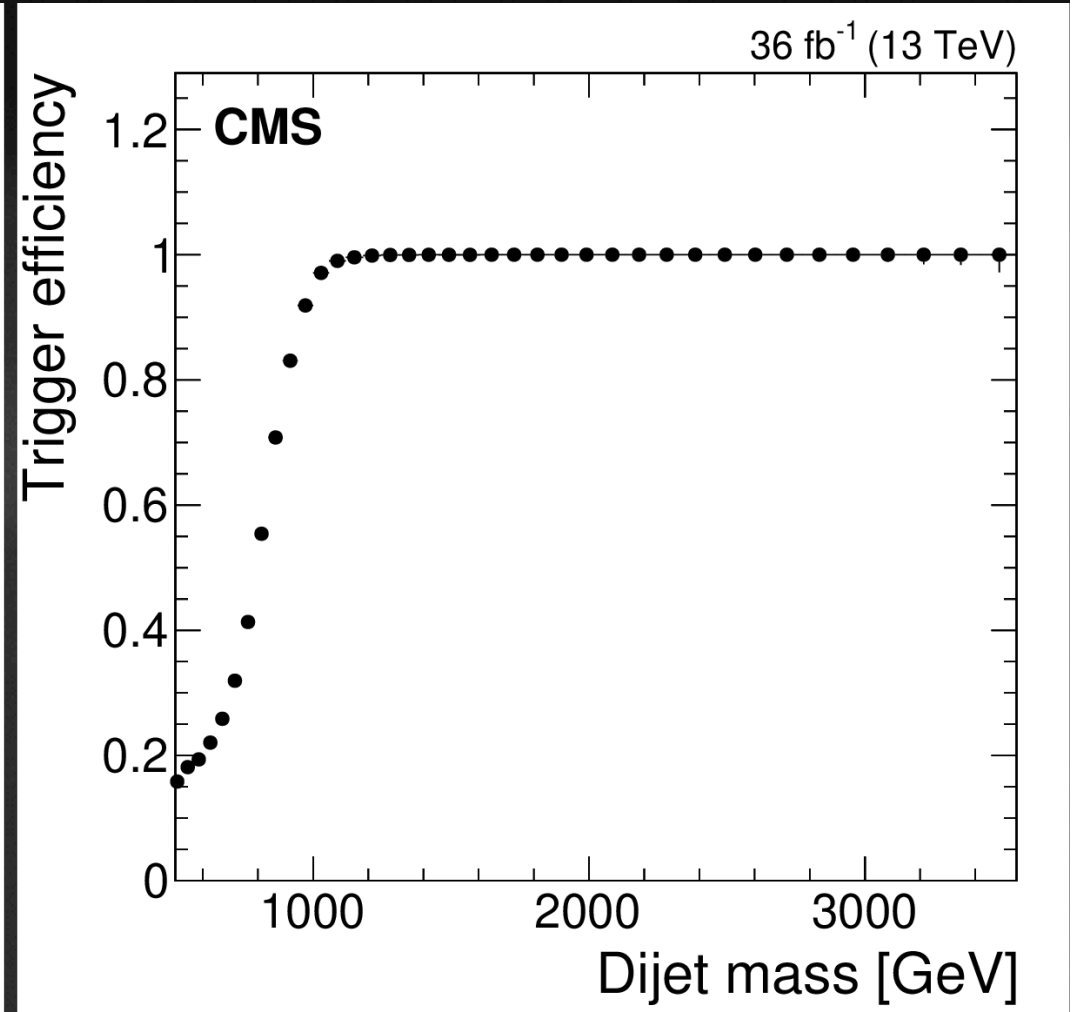
But also: lower HT thresholds



SCOUTING OBJECTS

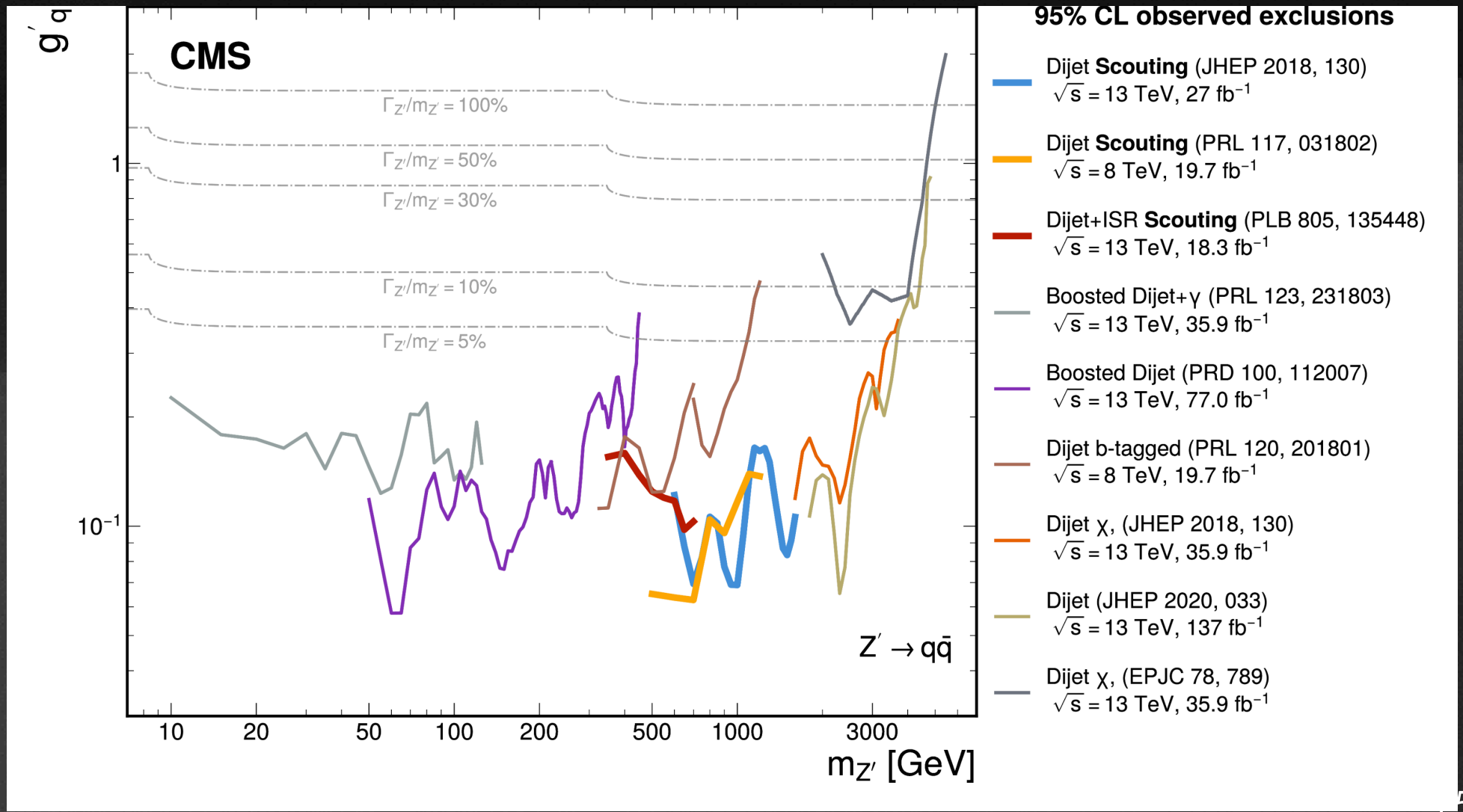
Two dijet trigger selection efficiencies (left is scouting)





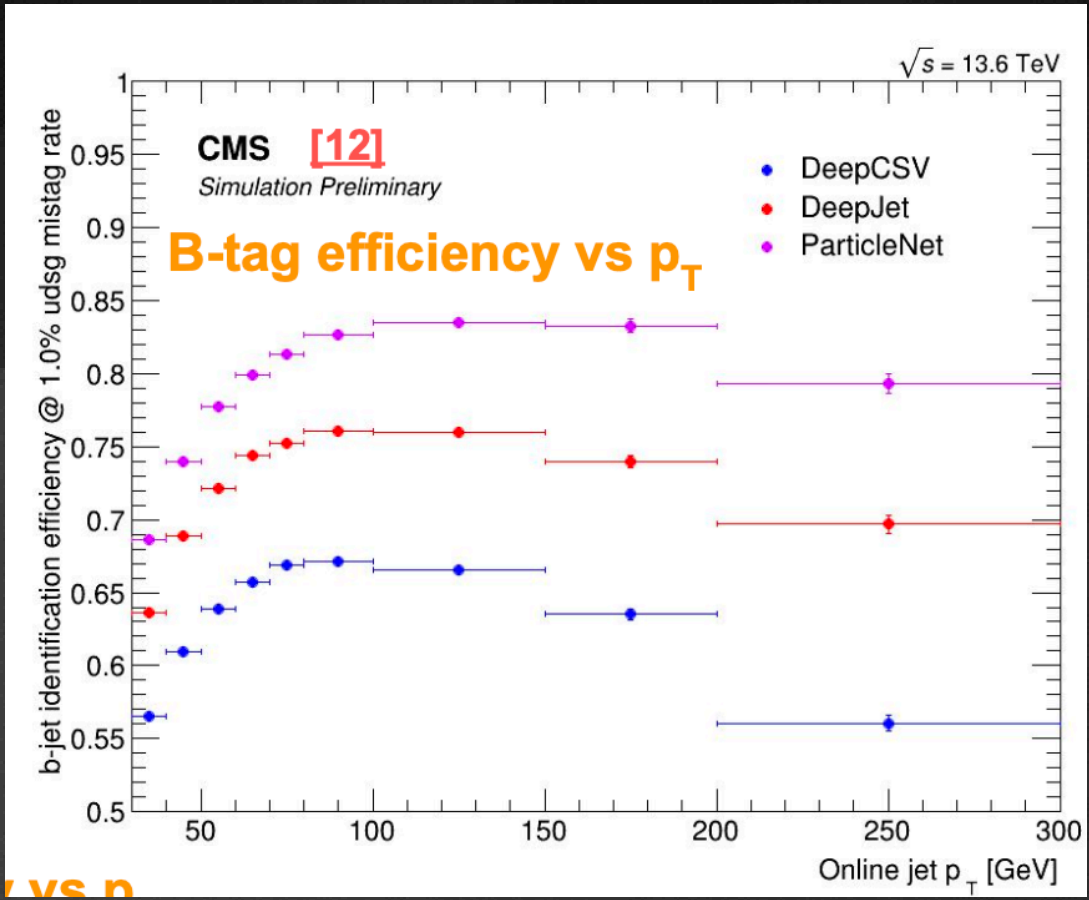
SCOUTING OBJECTS

Improvements on the physics (Z'->di-jet)



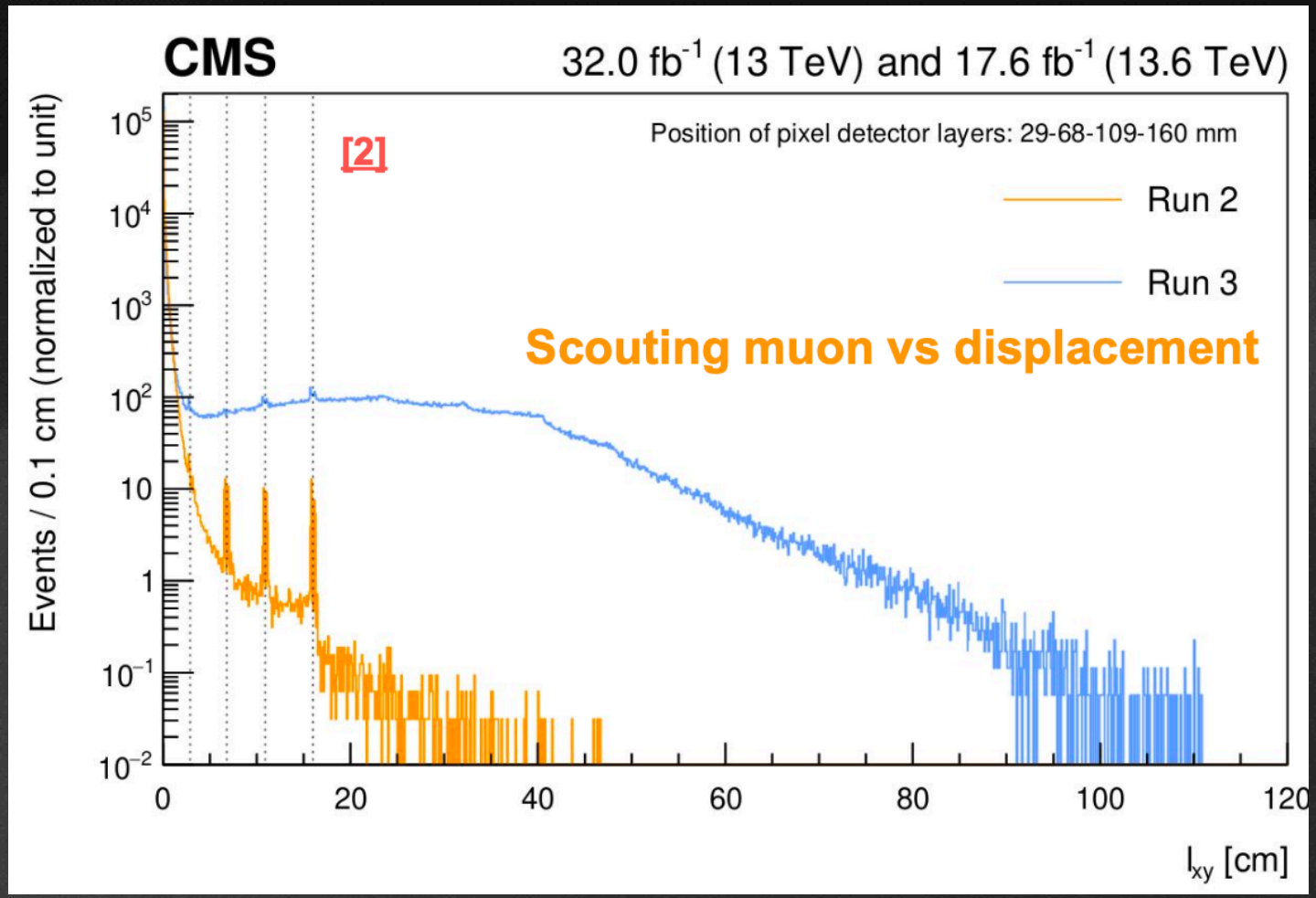
SCOUTING OBJECTS

By now objects like taus and b-jets
Also available!



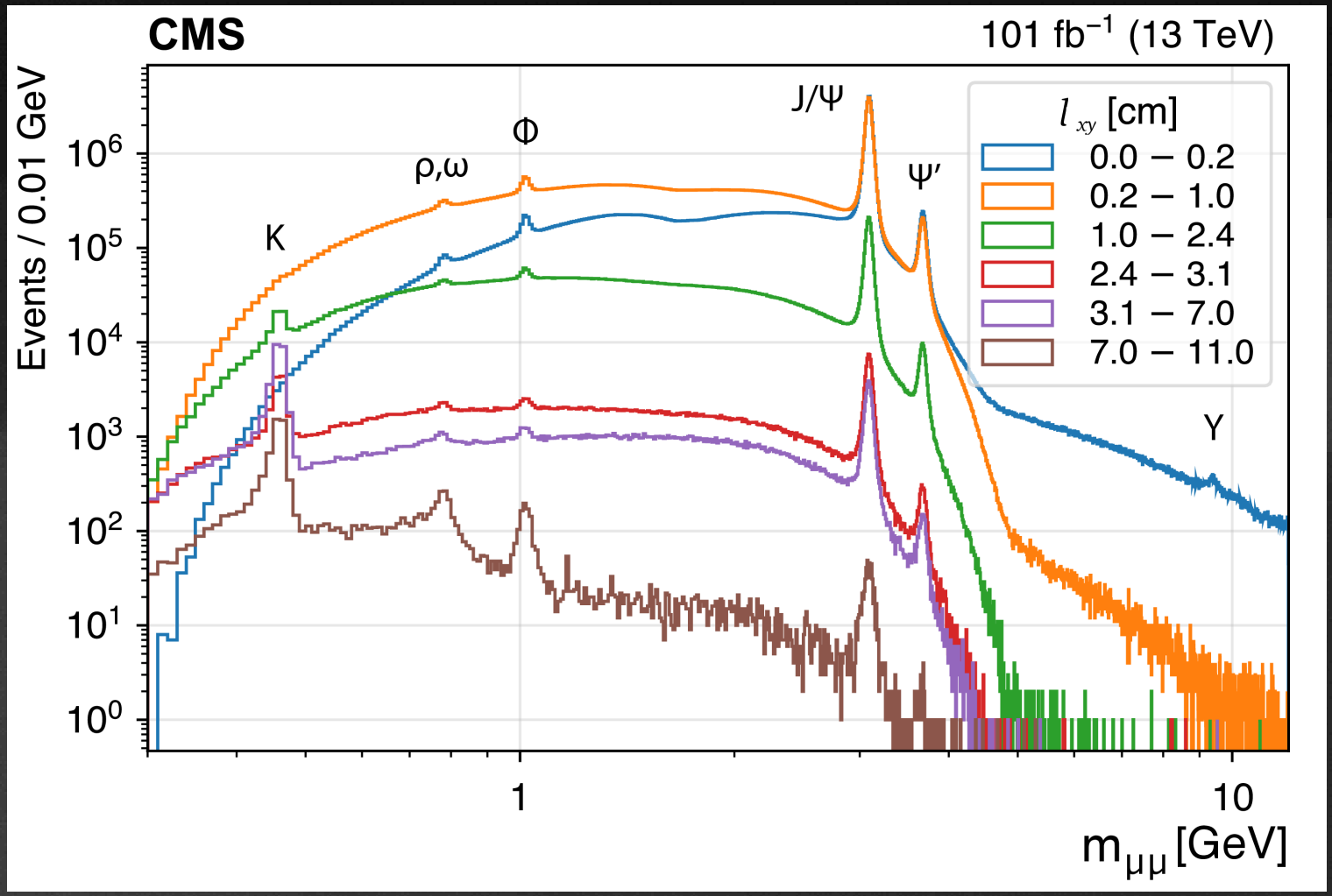
SCOUTING OBJECTS

And long-lived signatures



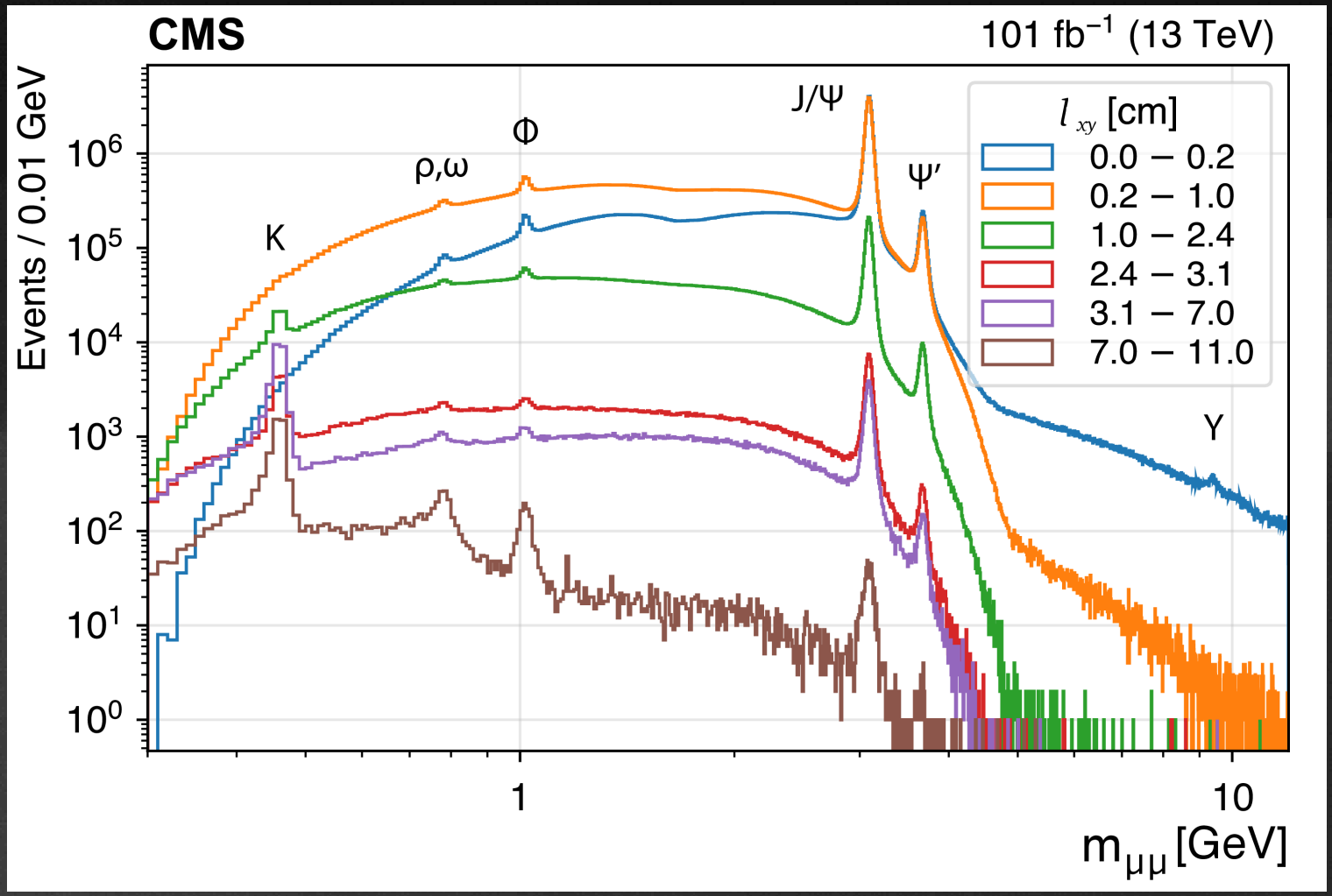
SCOUTING OBJECTS

And long-lived signatures

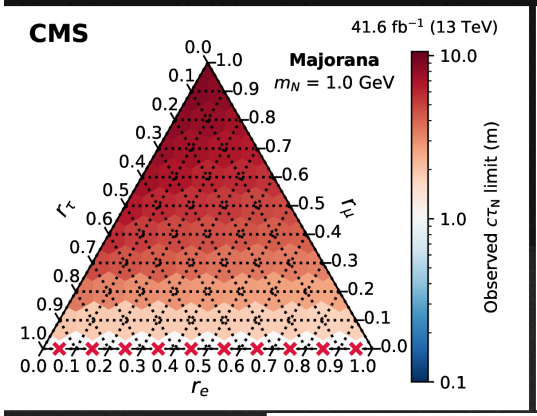


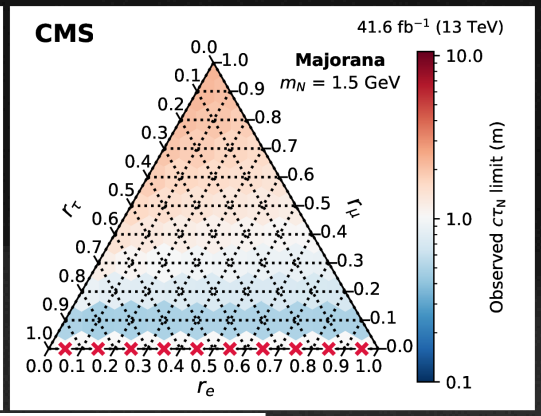
SCOUTING OBJECTS

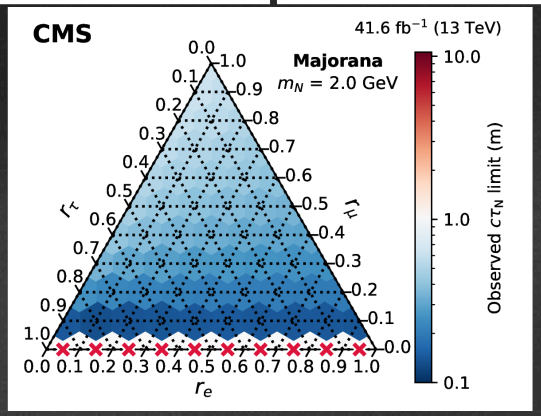
And long-lived signatures



SCOUTING - LET'S DO IT







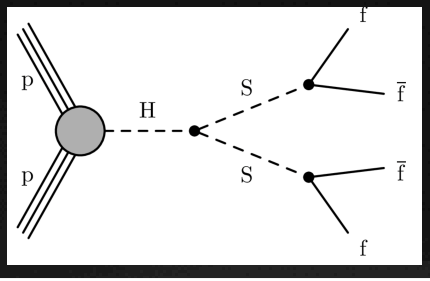
Example: low-mass HNL /Majorana neutrino limits

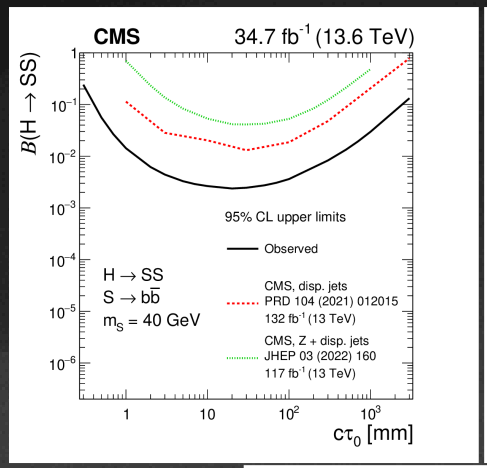
CMS by now collecting factors 10-30 more data through scouting than 'normal' trigger+reconstruction

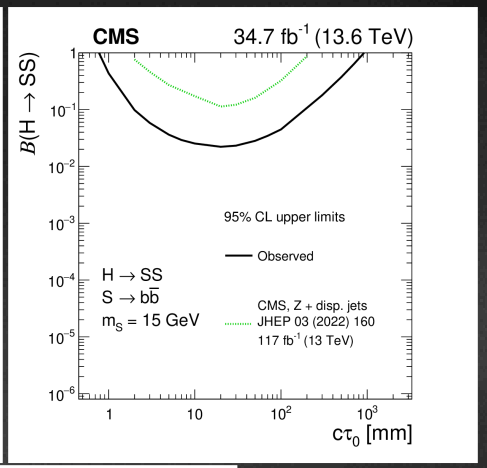
Particularly good for low-mass and low-pT signatures that are normally rejected by triggers

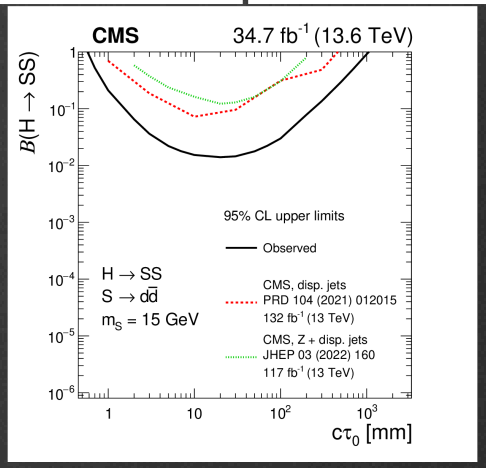
Let's look at it :)

SCOUTING - LET'S DO IT









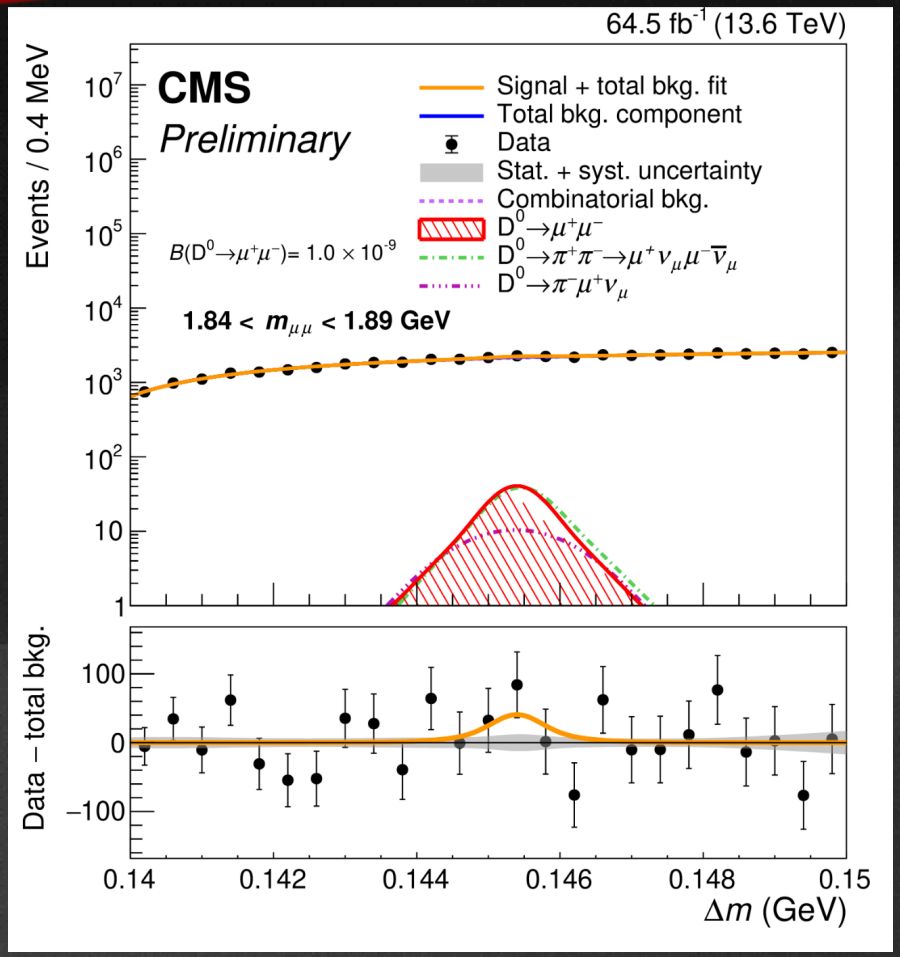
CMS by now collecting factors 10-30 more data through scouting than 'normal' trigger+reconstruction

Particularly good for low-mass and low-pT signatures that are normally rejected by triggers

Let's look at it :)

Light long-lived particles decaying to displaced jets
EXO-23-013, arXiv:2409.10806

SCOUTING - LET'S DO IT



CMS by now collecting factors 10-30 more data through scouting than 'normal' trigger+reconstruction

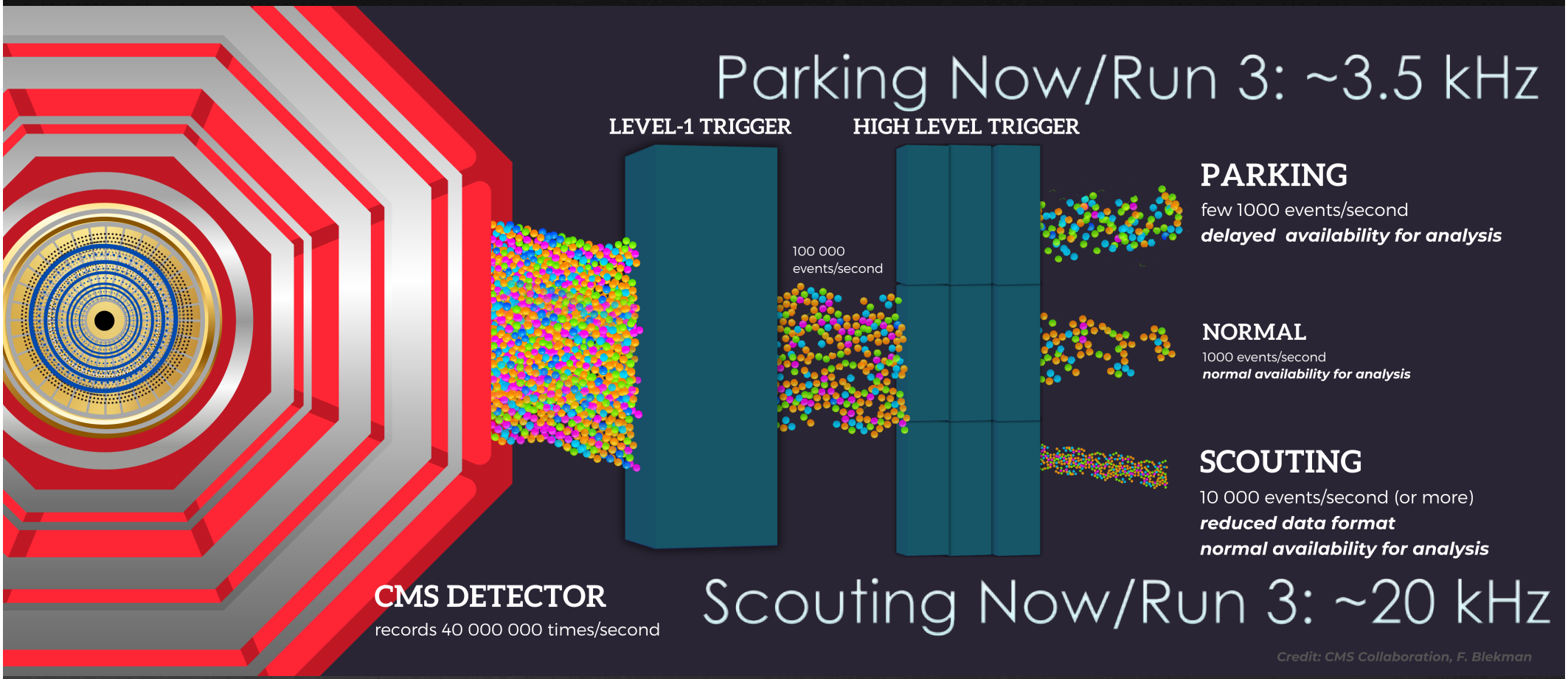
Particularly good for low-mass and low-pT signatures that are normally rejected by triggers

Let's look at it :)

World's strongest limits on D to μμ (BR < 2.6 10^-9)
BPH-23-008 (PAS)

socials: freyablekman

TIME FOR DISCUSSION: WHAT CAN THIS COMMUNTITY GAIN FROM SCOUTING AND PARKING?



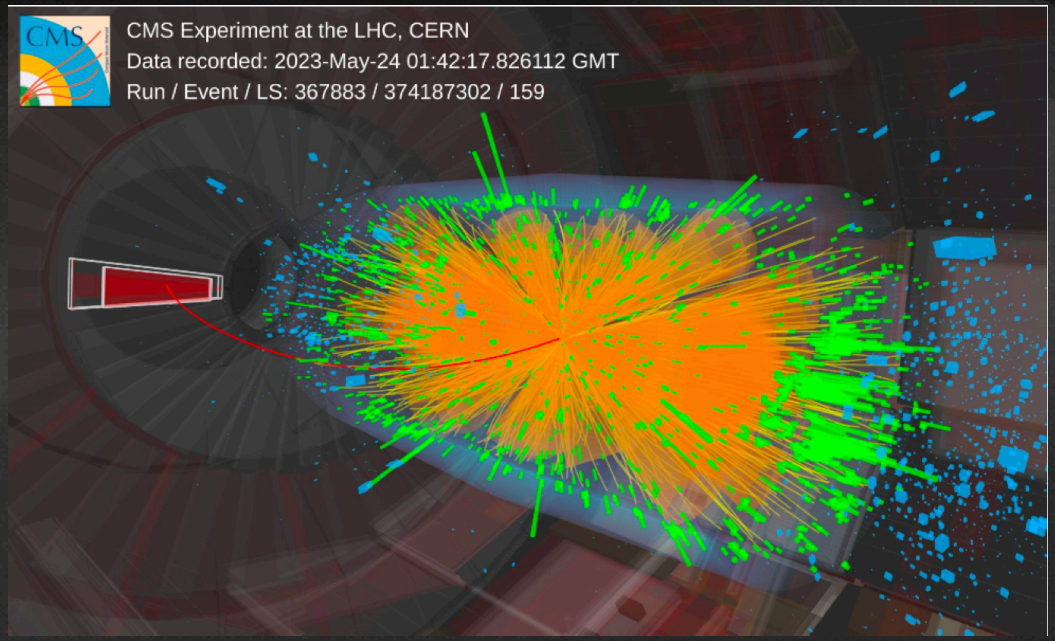
socials: freyablekman

MACHINE LEARNING: WHAT IS THE CUTTING EDGE FOR CMS?

- Loads happening, so only highlights
- CMS is frequently publishing ML method papers now: code: MLG
- ML in event selection
- ML in analysis:
  - Event reconstruction
  - Background estimation
  - Classic: signal vs background separation
- ML in interpretation:
  - Likelihood-free inference
  - Reweighting
  - Unfolding

MACHINE LEARNING: ARE WE MISSING SOME EVENTS?

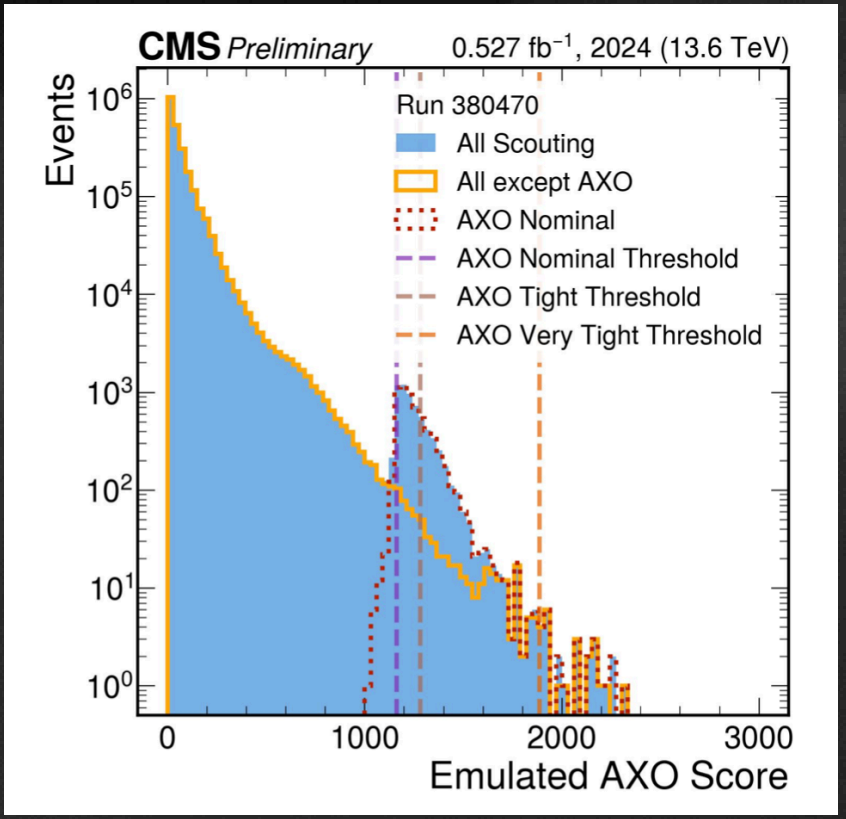
- Running various ML-based anomaly detection algorithms in the trigger - and selecting events we normally would reject
  - Trained on data (random/ low bias selection)
  - Shown: 12-jet event likely from double hard scatter (75 PU) that would not be selected by normal triggers (or offline, only 7 jets there)



AXOL1TL Anomaly detection in global trigger (in Run 3): CMS-DP-2023-079

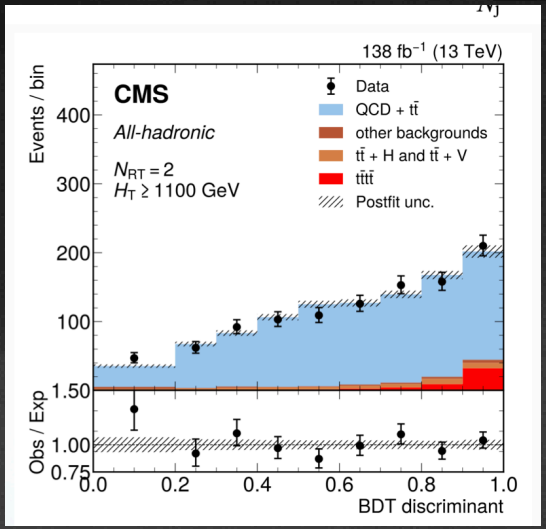
AXOL1TL IN SCOUTING

- Selects collisions with high HT but from many low pT objects (with a preference for high-PU events)
-



AXOL1TL in 2024 data: CMS-DP-2024-059

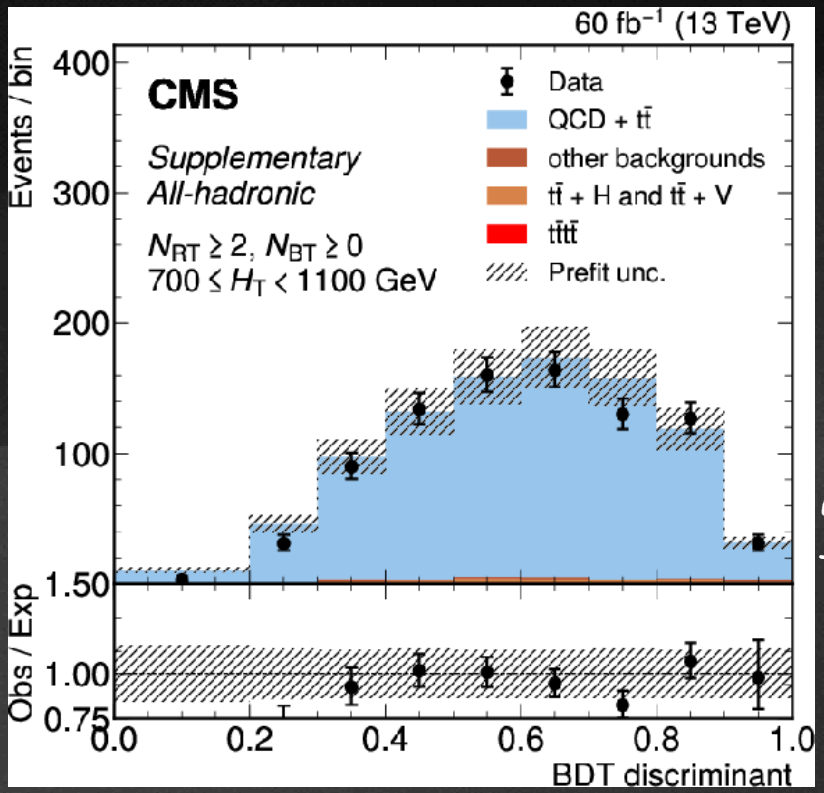
EXAMPLE : FOUR TOP QUARK PRODUCTION (TICKS MANY BOXES IN ML)



- Case study: All hadronic channel (9+jets)
- Of course:
  - Relies heavily on ML-based top and b tagging
  - uses ML (BDT in this case) for signal-background discrimination.
- But also: Using Neural autoregressive flow NN to predict BDT shapes from data control region to signal region

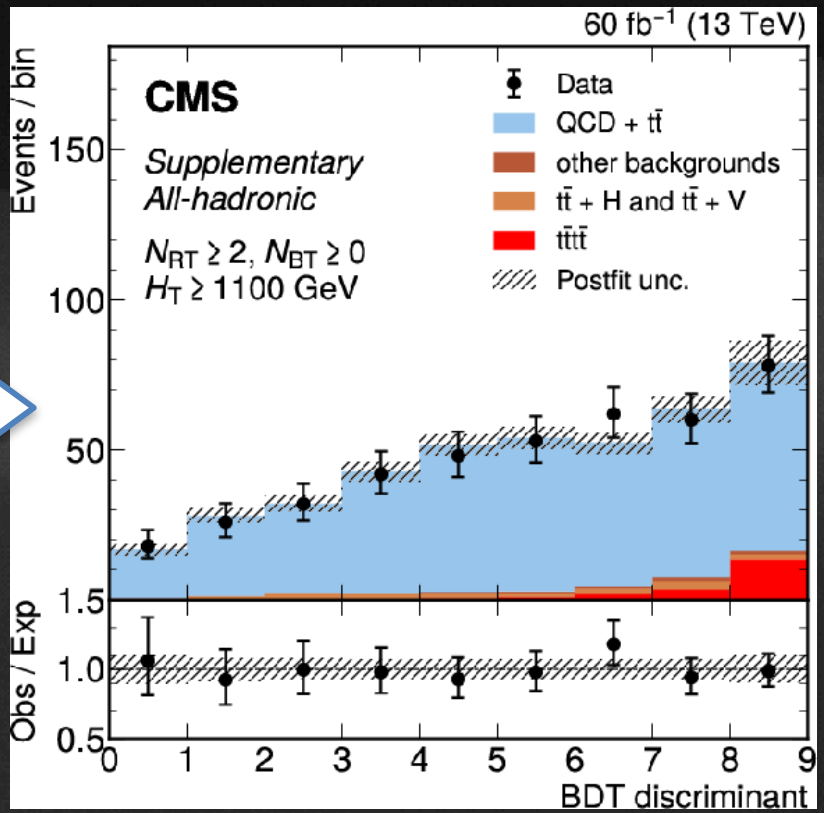
Evidence for tttt in final states with few leptons (all-hadronic, 1 slepton, 2 OS leptons) TOP-21-005 in PRD, arXiv:2303.03864

EXAMPLE : FOUR TOP QUARK PRODUCTION (TICKS MANY BOXES IN ML)



Neural net does transformation

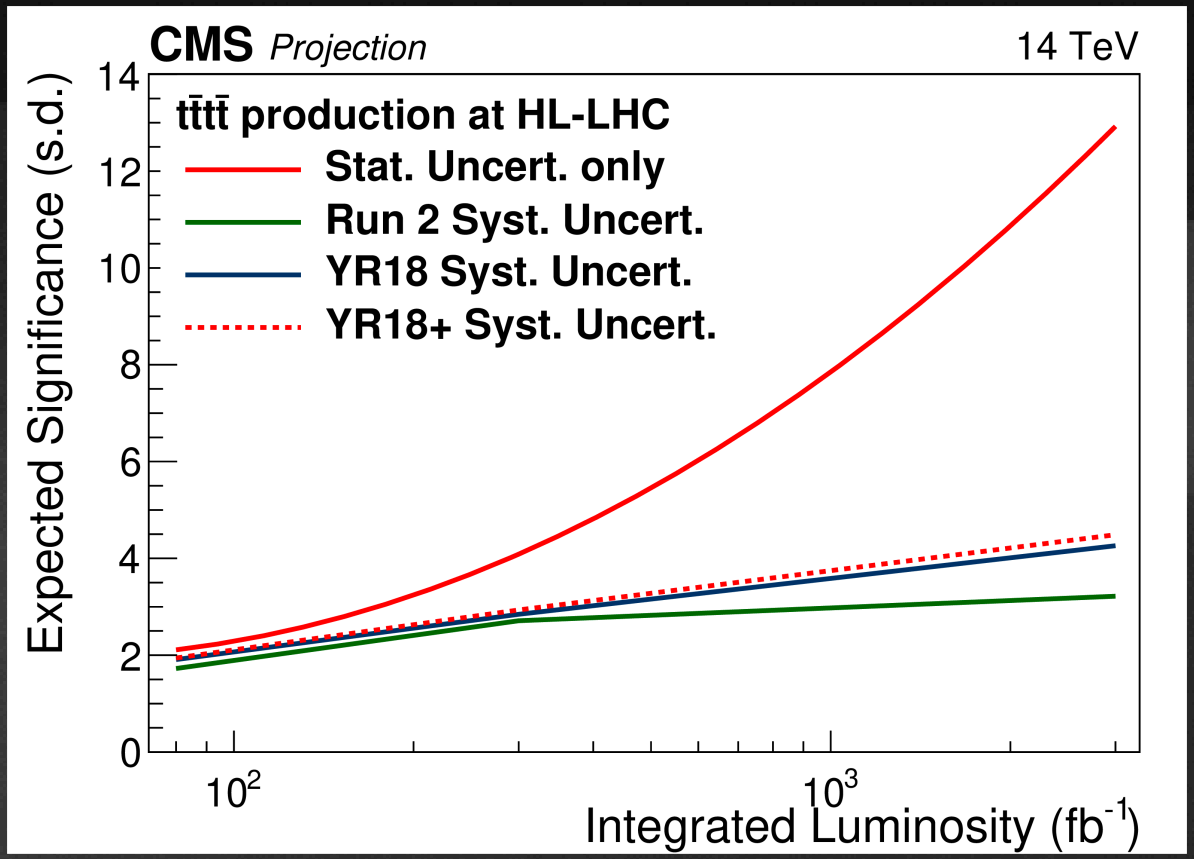
auto-regressive flow DNN



Evidence for tttt in final states with few leptons (all-hadronic, 1 slepton, 2 OS leptons)
TOP-21-005 in PRD, arXiv:2303.03864

socials: freyablekman

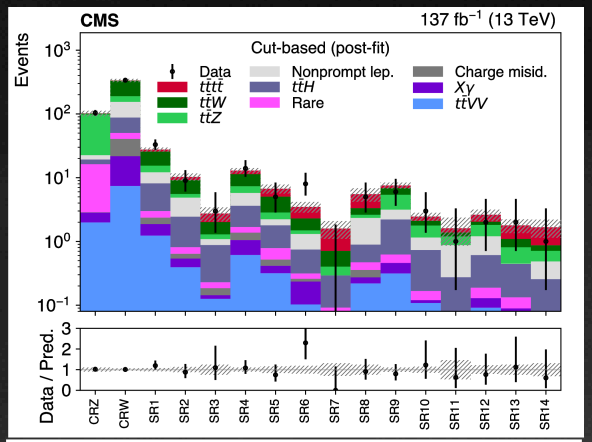
EXAMPLE : FOUR TOP QUARK PRODUCTION (TICKS MANY BOXES IN ML)

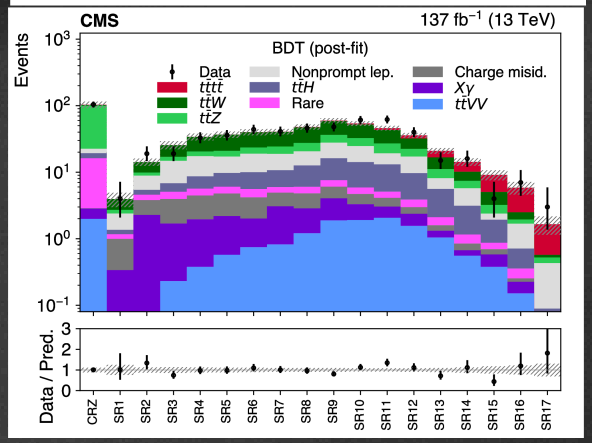


CMS-PAS-FTR-18-031

(recast of 2016SS dilepton tttt analysis with HL-LHC assumptions, also in HL-LHC YR)

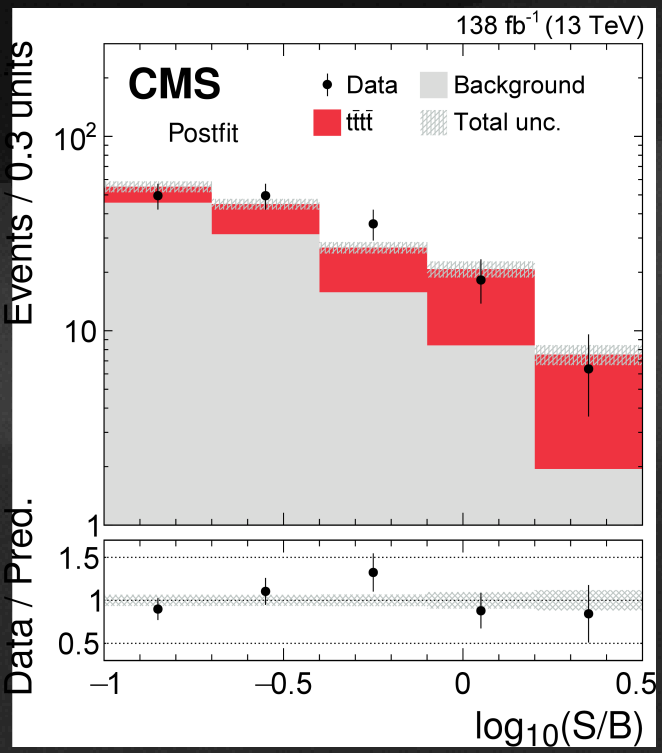
EXAMPLE : FOUR TOP QUARK PRODUCTION (TICKS MANY BOXES IN ML)





Same data
Better reconstruction

+much more ML
Gets you from
2.7 sigma
To
Over 5 sigma

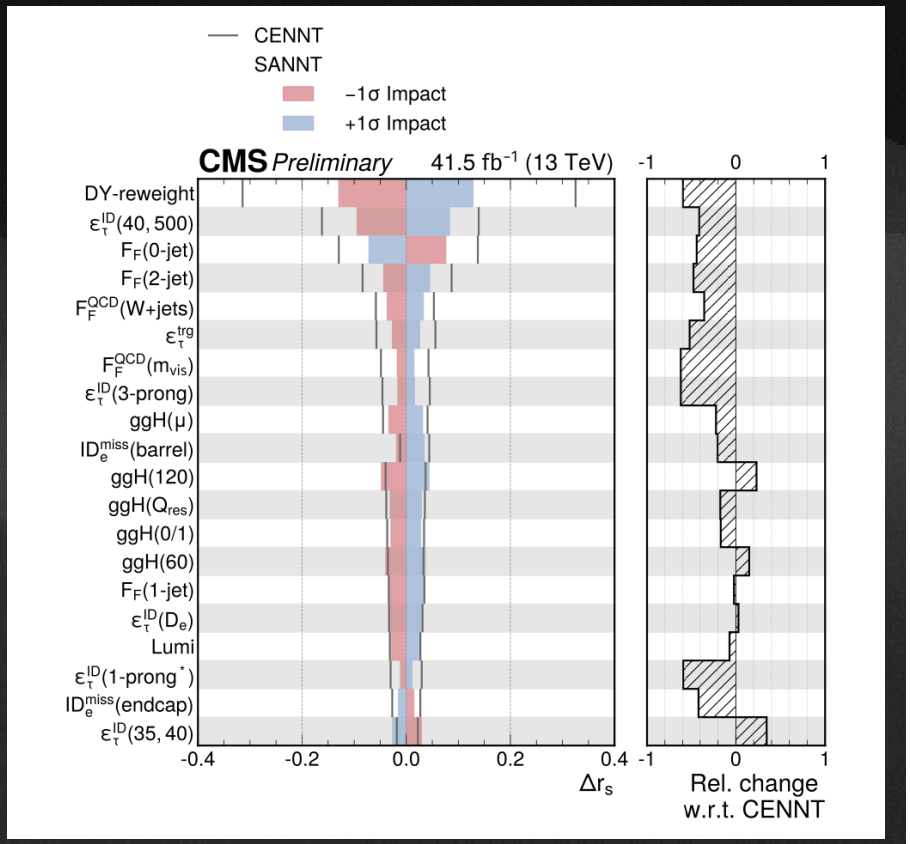


Observation for tttt in final states with many leptons (2 SS leptons, 3/4 leptons)
TOP-22-013 in PLB, arXiv:2305.13439
(old result TOP-19-017)

SIGNAL - BACKGROUND SEPARATION: PITFALLS!

- how reliable is the simulation?
  - simulation @ LO → might miss extra channels
  - QCD multijet simulation in analysis phase space challenging
  - optimal binning scheme of classifier output score? → might leave performance on the table
  - classification = not the best training target?
  - optimize for discovery significance : systematic (& profiling)-aware training
    → CMS-PAS-MLG-23-005, INFERNO

fits also in ML transparency/reproducibility discussion (relevant in wider ML community) - Q: are we teaching our NNs the "LO SM signature" or the "real SM signature"?



CMS-PAS-MLG-23-005 : systematic uncertainty aware NNs (examples with Higgs boson production in gluon fusion and vector-boson fusion signatures),

socials: freyablekman

W BOSON MASS

W BOSON MASS

- Long time in the making
- Completely new method that does not rely (as much) extrapolation from Z to W
- Loads of expertimental improvement (including on muon resolution, new tracking, corrections to magnetic field, etc etc)
- Loads of theory development as well!
- These slides won't do the result credit, LHC CERN seminar by Josh Bendavid available (incl. web cast) here: https://indico.cern.ch/event/1441575/
- PAS is out: SMP-23-002
- The paper is planned to be very similar to the PAS, so important to give feedback on missing information NOW so it is added (**really *now, do not wait for arXiv***!)

W BOSON MASS

CMS W mass measurement strategy **(on one slide! And for this audience :))**

- Extremely well-understood (relatively large) sample of single muon events (no electron channel) from 2016
- Directly measure the 2D binned distribution of yields vs muon pT, muon eta and separately per muon charge (so: not MT)
    - Also using helicity components (statistically less sensitive but valuable check)
- Key point: get precise calibration of muons from J/ψ
    - that way the Z can be used for testing/cross-checks
- Relies on theory uncertainties from N3LL accuracy and QCD@NNLO calculations. Uncertainties are constrained in situ in the very finely binned fit to pT, eta, charge
- PDFs: Uses methods from:
    - E. Manca, et al,"About the rapidity and helicity distributions of the W bosons produced at LHC", JHEP, arXiv:1707.09344
    - S. Farry, et.al, "Understanding and constraining the PDF uncertainties in a W boson mass measurement with forward muons at the LHC", EPJC, arxXiv:1902.04323

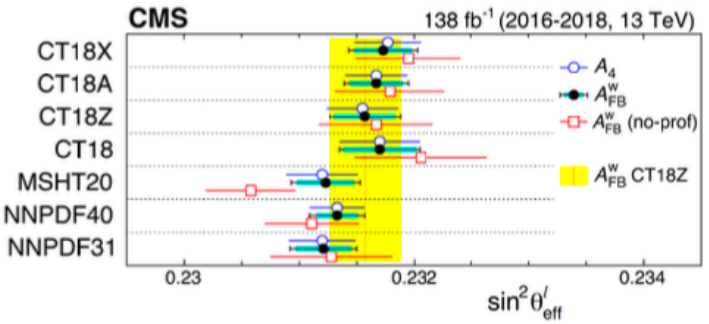
PDF SET CHOICE

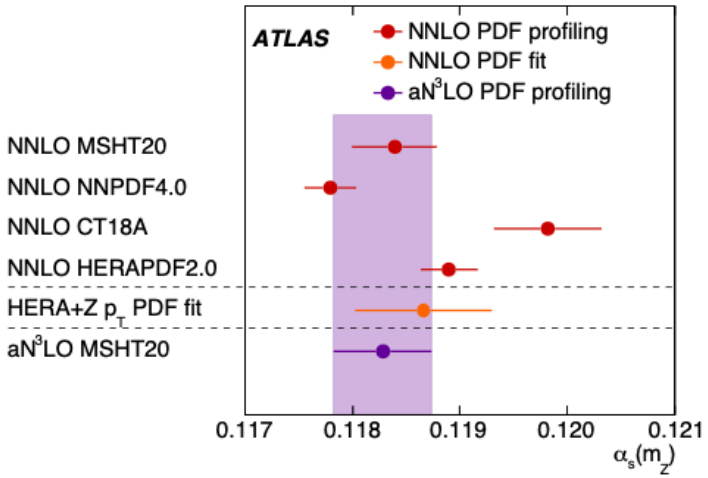
Theoretical Considerations

CMS /ATLAS DY AFB/sin2theta and alpha-S results (plus blinded uncertainty) used to pick 'main' PDF

arXiv:2408.07622, arXiv:2309.12986

- PDFs are a challenge! In recent precision measurements at hadron colliders often a significant spread in measured values depending on the choice of PDF set
- Angular dependence of W and Z production can be decomposed in terms of angular coefficients/helicity cross sections:
- This can be a useful way to factorize theoretical corrections and uncertainties

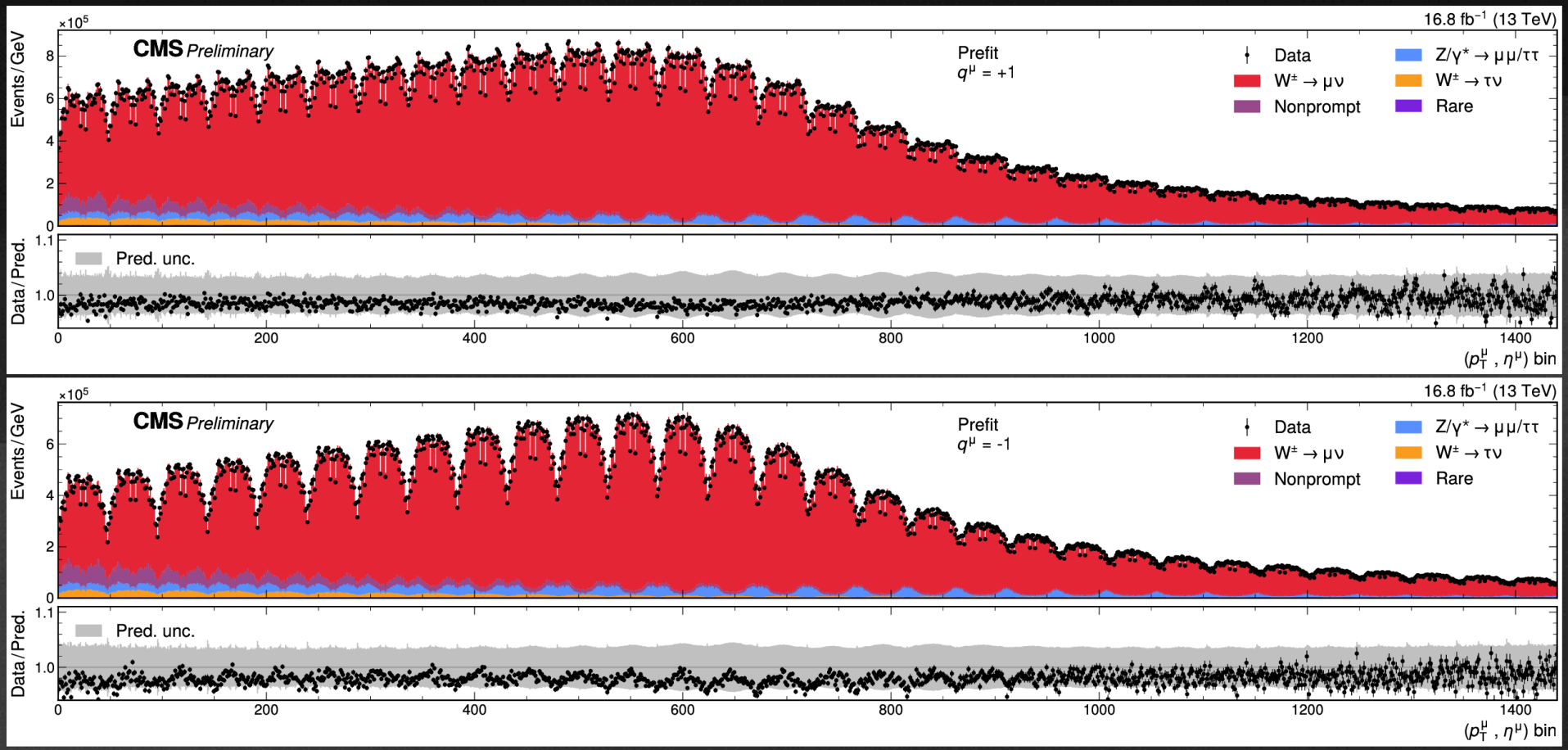




$$\frac{\mathrm{d}^5\sigma}{\mathrm{d}q_T^2\,\mathrm{d}y\,\mathrm{d}m\,\mathrm{d}\cos\theta\,\mathrm{d}\phi} = \frac{3}{16\pi}\frac{\mathrm{d}^3\sigma^{U+L}}{\mathrm{d}q_T^2\,\mathrm{d}y\,\mathrm{d}m}[(1+\cos^2\theta)+\frac{1}{2}A_0(1-3\cos^2\theta)+A_1\sin 2\theta\cos\phi$$
$$+\frac{1}{2}A_2\sin^2\theta\cos 2\phi+A_3\sin\theta\cos\phi+A_4\cos\theta+A_5\sin^2\theta\sin 2\phi+A_6\sin 2\theta\sin\phi+A_7\sin\theta\sin\phi]$$

THE FIT (PRE-FIT)

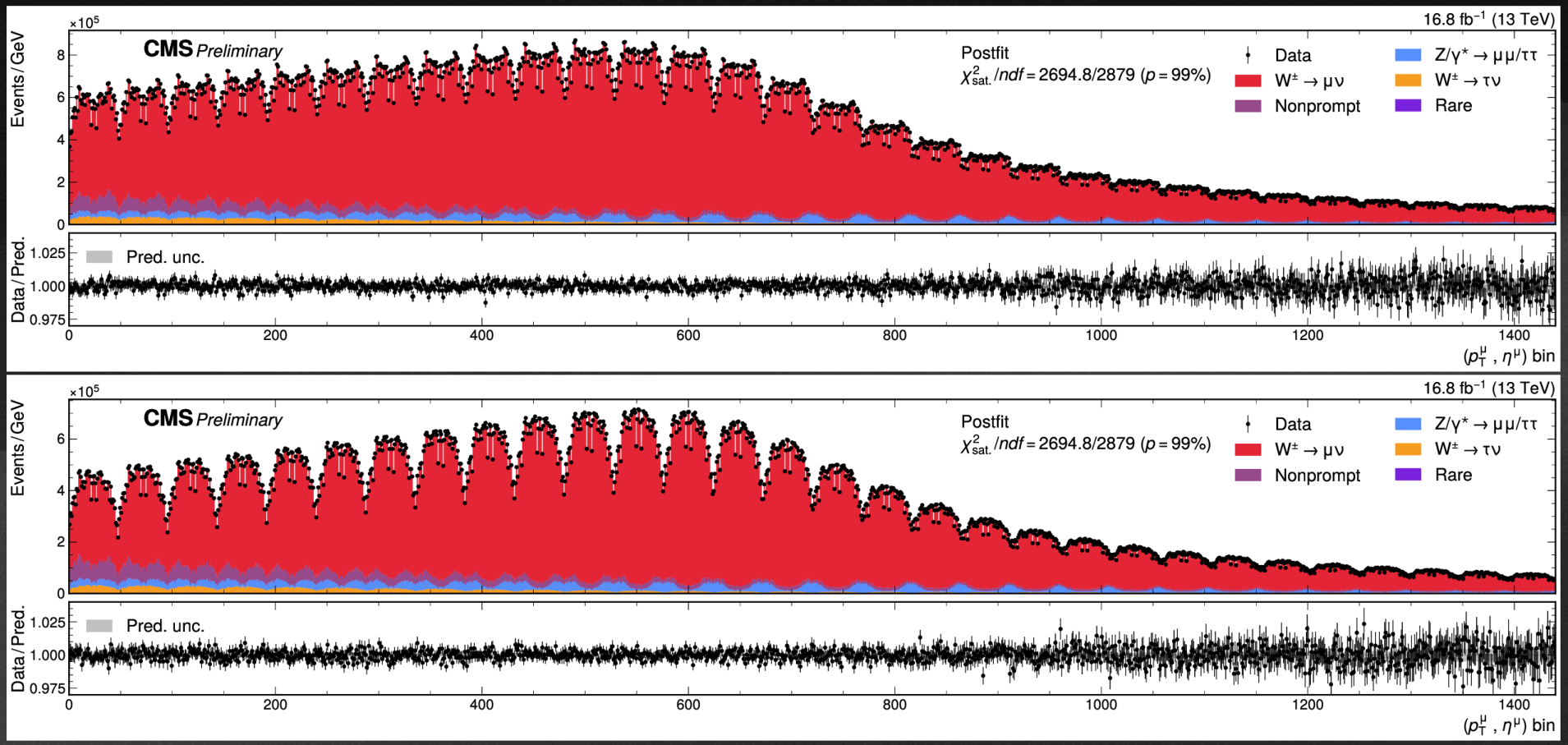
- After 10 years of work….



CMS-PAS-SMP-22-002 (give feedback now! particularly on extra appendices)
as these distributions are in the paper they will be in HEPDATA!

THE FIT (POST-FIT)

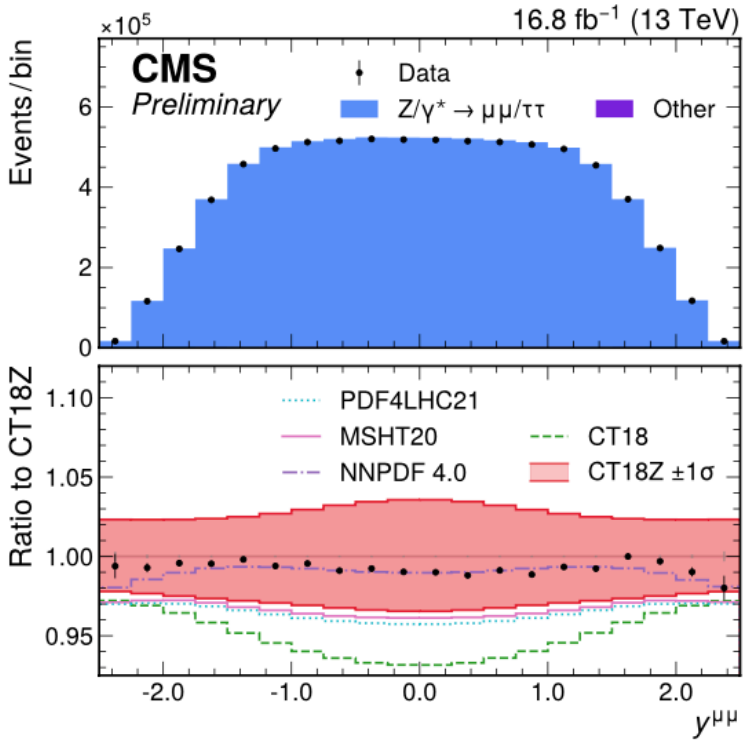
- After 10 years of work….



CMS-PAS-SMP-22-002 (give feedback now! Particularly on extra appendices) as these distributions are in the paper they will be in HEPDATA!

W MASS VS PDFS

Parton Distribution Functions



- **Good:** PDF sets are accompanied by uncertainty models with well defined correlations across phase space and between processes
- **Bad:** Different PDFs don't necessarily agree within their uncertainties
- Missing higher order uncertainties, resummation corrections in predictions usually not included
  - Partly mitigated by tolerance factors, etc

W MASS VS PDFS

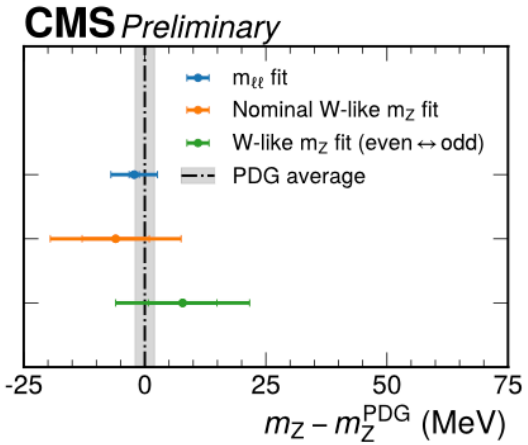
Parton Distribution Functions

PDF set	Scale factor	Impact in m_W (MeV) Original σ_{PDF}	Scaled σ_{PDF}
CT18Z	–	4.4	
CT18	–	4.6	
PDF4LHC21	–	4.1	
MSHT20	1.5	4.3	5.1
MSHT20aN3LO	1.5	4.2	4.9
NNPDF3.1	3.0	3.2	5.3
NNPDF4.0	5.0	2.4	6.0

- **Strategy:** Scale prefit PDF uncertainties to ensure consistency between sets for measured m_W value
- This procedure does **not** prove that e.g. NNPDF4.0 uncertainty is underestimated, only that it's too small to cover the central value of the other sets
- CT18Z is chosen as the nominal since it covers the others without scaling and with small uncertainty
  - But note that this set is amongst the largest in terms of nominal uncertainty

Z MASS, THE FINAL CROSS CHECK

W-like m_Z result: Uncertainty Breakdown

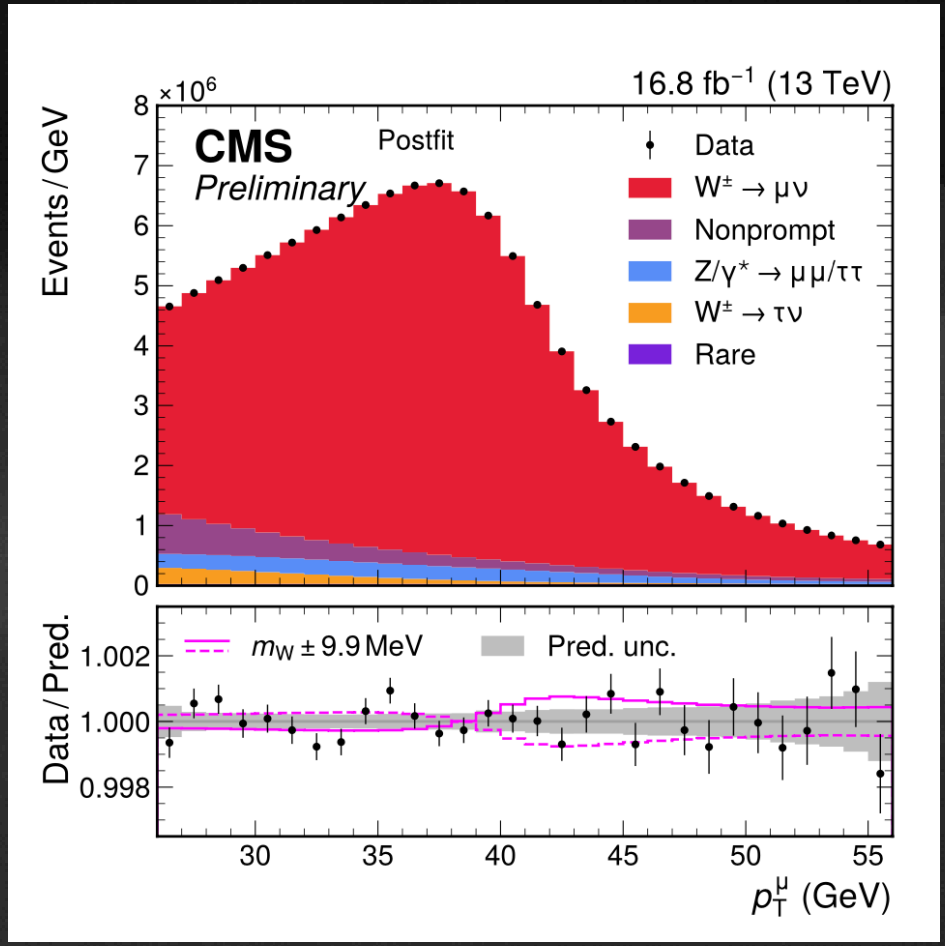


Source of uncertainty	Impact (MeV) Nominal	Global
Muon momentum scale	5.6	5.3
Muon reco. efficiency	3.8	3.0
W and Z angular coeffs.	4.9	4.5
Higher-order EW	2.2	2.2
p_T^V modeling	1.7	1.0
PDF	2.4	1.9
Integrated luminosity	0.3	0.2
MC sample size	2.5	3.6
Data sample size	6.9	10.1
Total uncertainty	13.5	13.5

- Largest uncertainties are statistical, muon calibration, angular coefficients
- Total uncertainty is well defined, but several different ways of decomposing statistical and systematics uncertainties
- When uncertainties are constrained in-situ, "global" impacts (used e.g. for ATLAS 2024 m_W measurement) tends to count them as part of the statistical uncertainties

W BOSON MASS

- 80350.2 ± 9.9 MeV
- Largest uncertainties:
  - experimental muon momentum, reconstruction, non-prompt background, MC sample size (!)
  - Theory: PDFs, higher order EW, pT-V modelling



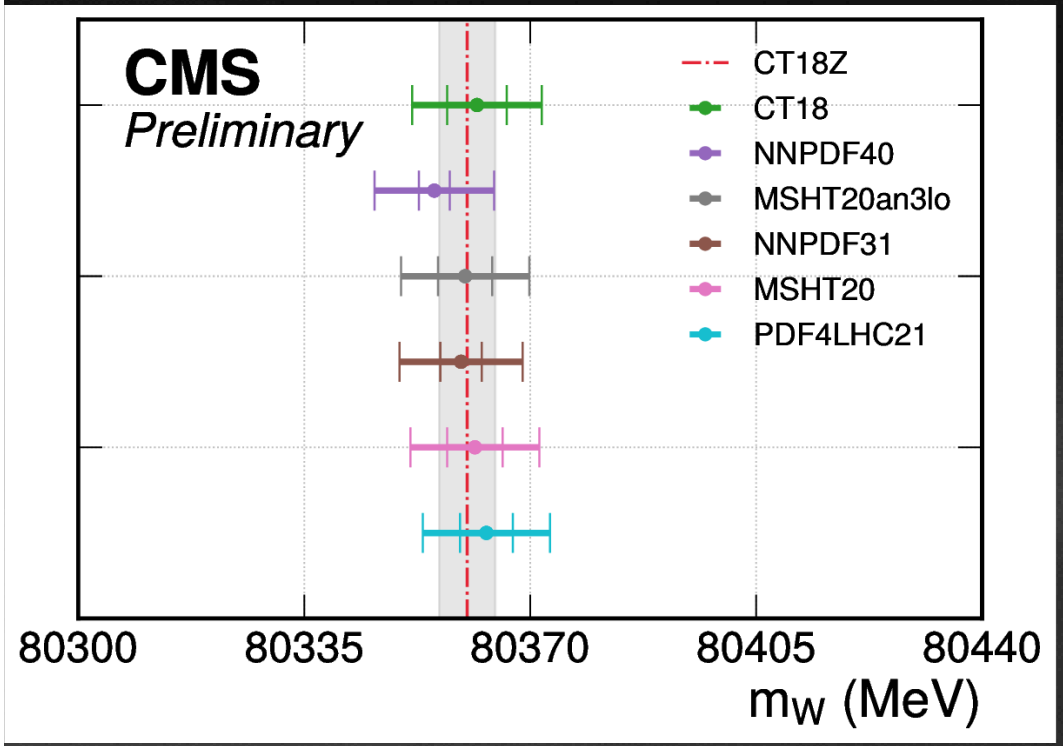
CMS-PAS-SMP-22-002 (give feedback now! Particularly on extra appendices) as these distributions are in the paper they will be in HEPDATA!

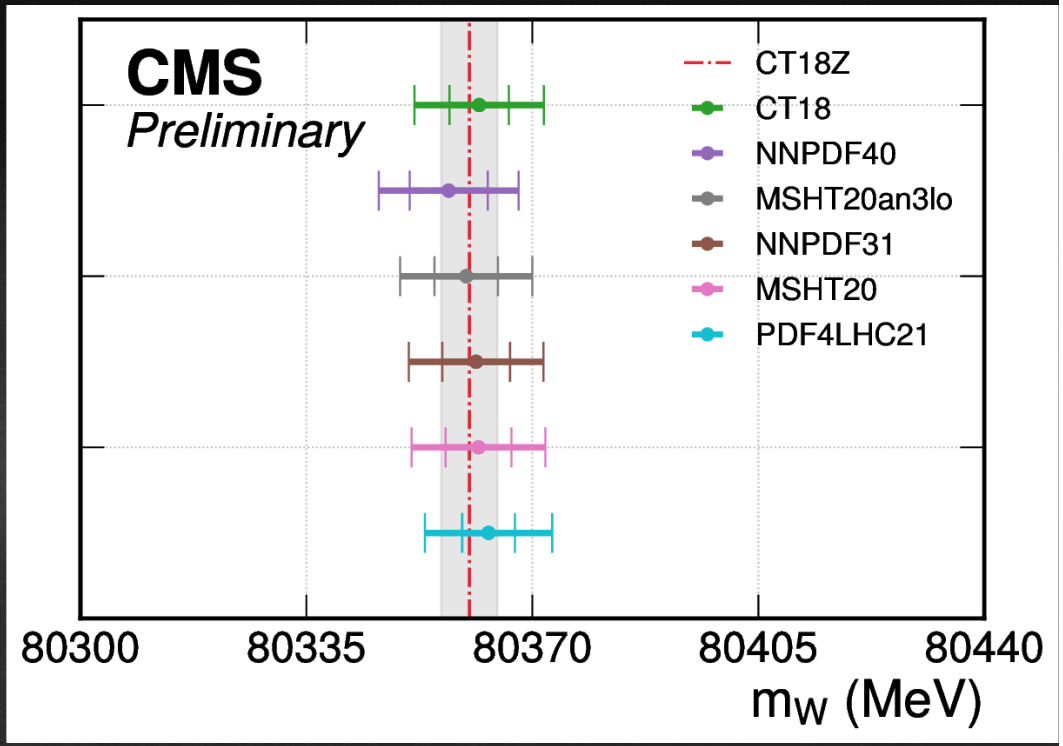
W BOSON MASS

Source of uncertainty	Impact (MeV) Nominal		Global	
	in m_Z	in m_W	in m_Z	in m_W
Muon momentum scale	5.6	4.8	5.3	4.4
Muon reco. efficiency	3.8	3.0	3.0	2.3
W and Z angular coeffs.	4.9	3.3	4.5	3.0
Higher-order EW	2.2	2.0	2.2	1.9
p_T^V modeling	1.7	2.0	1.0	0.8
PDF	2.4	4.4	1.9	2.8
Nonprompt background	–	3.2	–	1.7
Integrated luminosity	0.3	0.1	0.2	0.1
MC sample size	2.5	1.5	3.6	3.8
Data sample size	6.9	2.4	10.1	6.0
Total uncertainty	13.5	9.9	13.5	9.9

CMS-PAS-SMP-22-002 (give feedback now! Particularly on extra appendices)
as these distributions are in the paper they will be in HEPDATA!

W BOSON MASS PER PDF SET





- "Scaled" effectively means differences in the MT/pT spectrum peak of the nominal PDF were corrected before fitting

CMS-PAS-SMP-22-002 (give feedback now! Particularly on extra appendices)
as these distributions are in the paper they will be in HEPDATA!

socials: freyablekman

W BOSON MASS PER PDF SET

PDF set	Extracted m_W (MeV)	
	Original σ_{PDF}	Scaled σ_{PDF}
CT18Z	$80\,360.2 \pm 9.9$	
CT18	$80\,361.8 \pm 10.0$	
PDF4LHC21	$80\,363.2 \pm 9.9$	
MSHT20	$80\,361.4 \pm 10.0$	$80\,361.7 \pm 10.4$
MSHT20aN3LO	$80\,359.9 \pm 9.9$	$80\,359.8 \pm 10.3$
NNPDF3.1	$80\,359.3 \pm 9.5$	$80\,361.3 \pm 10.4$
NNPDF4.0	$80\,355.1 \pm 9.3$	$80\,357.0 \pm 10.8$

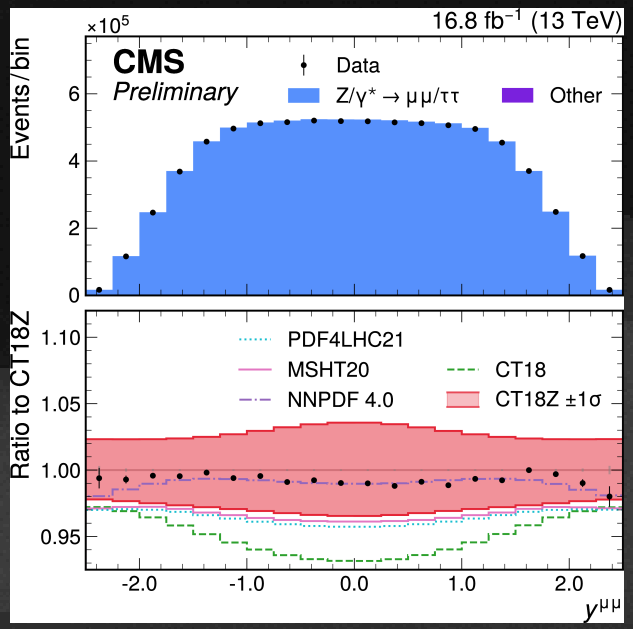
- "Scaled" effectively means differences in the MT/pT spectrum peak of the nominal PDF were corrected before fitting
- Information (pre-fit on more advanced pdfs (e.g. heavy quark dependence) are available internally
- Plan is to provide theory-propagated PDF uncertainties for all PDF sets, once the theory model is fully published etc.

CMS-PAS-SMP-22-002 (give feedback now! Particularly on extra appendices)
as these distributions are in the paper they will be in HEPDATA!

socials: freyablekman

W BOSON MASS PER PDF SET

PDF set	Extracted m_W (MeV)	
	Original σ_{PDF}	Scaled σ_{PDF}
CT18Z	$80\,360.2 \pm 9.9$	
CT18	$80\,361.8 \pm 10.0$	
PDF4LHC21	$80\,363.2 \pm 9.9$	
MSHT20	$80\,361.4 \pm 10.0$	$80\,361.7 \pm 10.4$
MSHT20aN3LO	$80\,359.9 \pm 9.9$	$80\,359.8 \pm 10.3$
NNPDF3.1	$80\,359.3 \pm 9.5$	$80\,361.3 \pm 10.4$
NNPDF4.0	$80\,355.1 \pm 9.3$	$80\,357.0 \pm 10.8$



CMS-PAS-SMP-22-002 (give feedback now! Particularly on extra appendices)
as these distributions are in the paper they will be in HEPDATA!

CONCLUSION

- CMS is probing lower pT through scouting and parking
  - Also means "on demand" triggers, and lower thresholds on "main" analyses
  - ideas welcome!
- Machine learning is pushing the edge
  - But standard "signal-background discrimination" is starting to reach its limits
  - Innovation in data-driven modelling, ML in low-level selection, simplification of complex prodedures
- Completely new mW method that does not rely (as much on) extrapolation from Z to W
  - But it does rely on theory inputs, of course, including PDFs!
  - Important to give input now ! (easiest: cms-pag-conveners-smp@cern.ch)

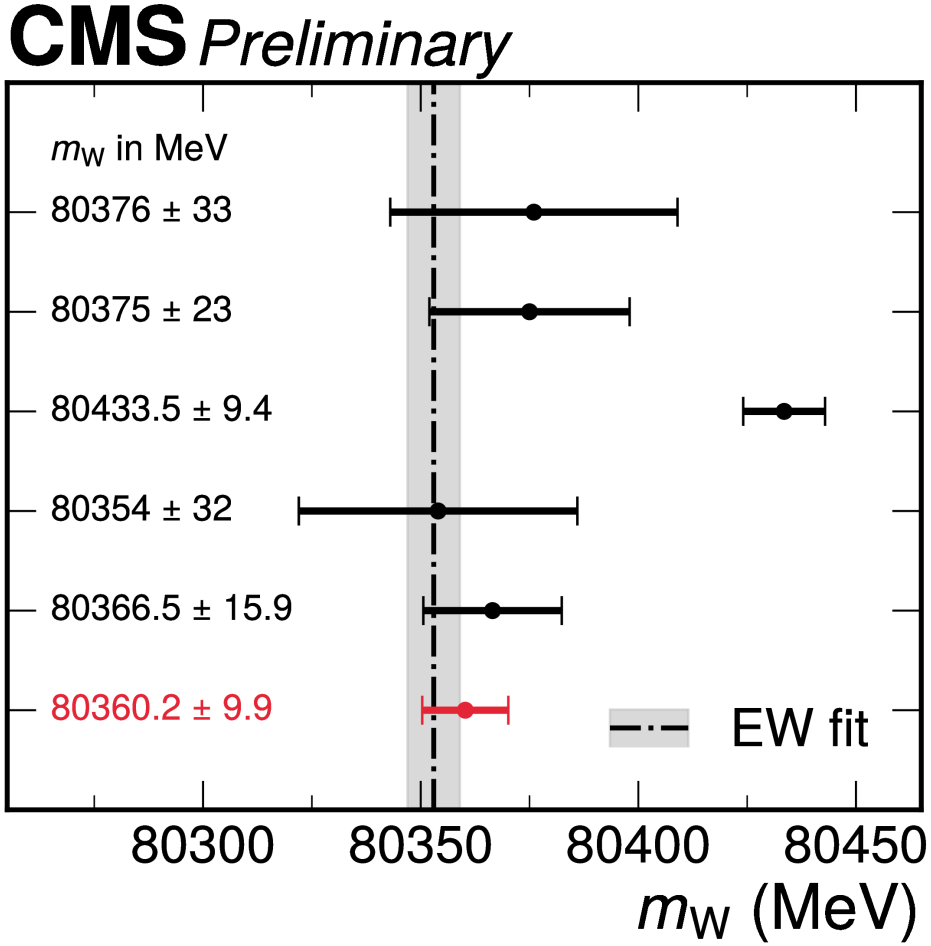
Thank you

for the invitation and your attention

socials: freyablekman

CONCLUSION

LEP combination
Phys. Rep. 532 (2013) 119
D0
PRL 108 (2012) 151804
CDF
Science 376 (2022) 6589
LHCb
JHEP 01 (2022) 036
ATLAS
arxiv:2403.15085, subm. to EPJC
CMS
This Work



Thank you

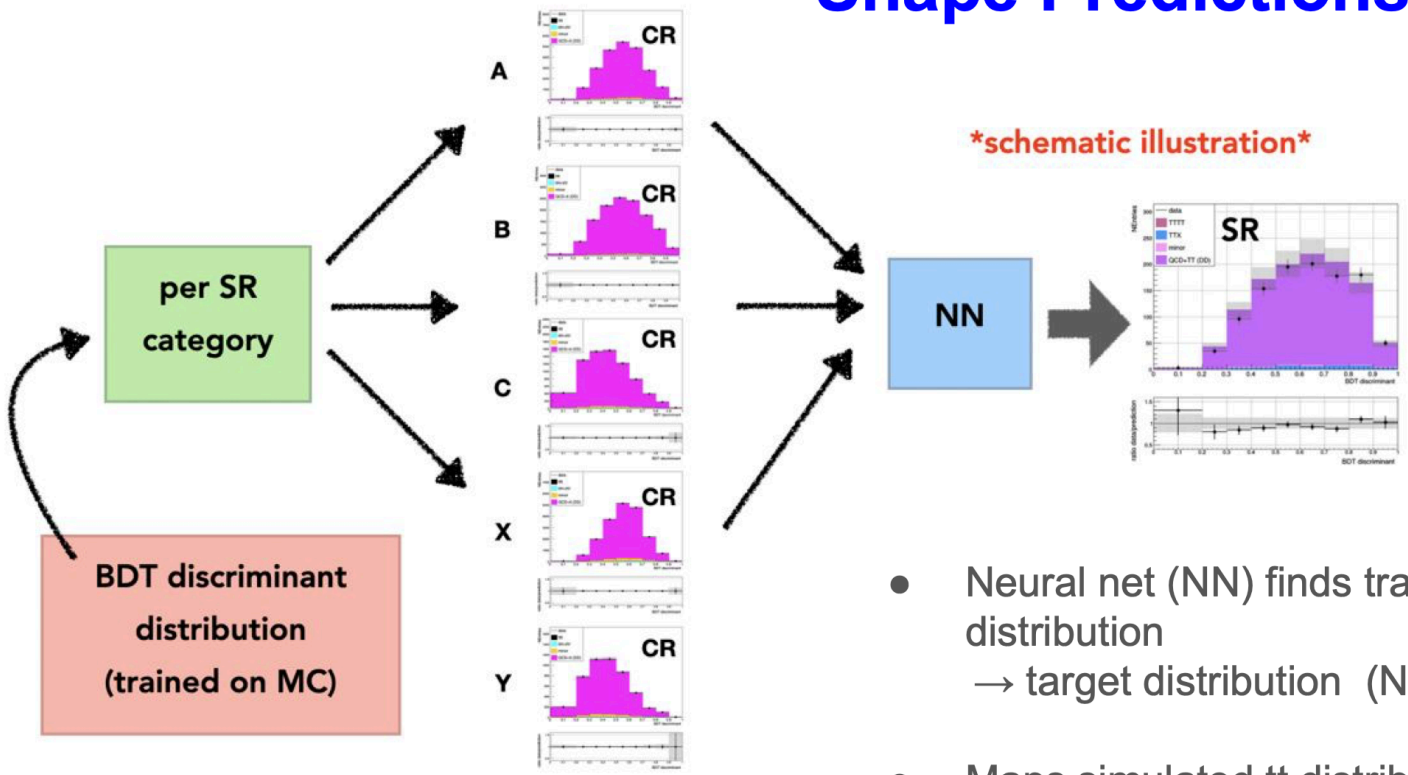
for the invitation and your attention

socials: freyablekman

BACKUP

AUTOREGRESSIVE FLOW (1)

Shape Predictions for BDT and H_T



- Neural net (NN) finds transformation from input distribution
  → target distribution (Neural autoregressive flow)
- Maps simulated tt distributions onto tt + QCD distributions in 5 CR distributions for BDT & H_T simultaneously

*Huang, Krueger, Lacoste, Courville. *Neural Autoregressive Flows.* arXiv:1804.00779
*S. Choi. arXiv:2008.0363

AUTOREGRESSIVE FLOW (2)

- Normalizing flow

  Phys. Lett. B 844 (2023) 138076

  - idea: use autoregressive normalizing flow to map
  - learn to map background distribution CR → SR
  - applied in full hadronic 4t analyses by CMS
  - trained on $\mathrm{t\bar{t}}$ MC
  - 2D transform of $(H_{\mathrm{T}}, \mathrm{BDT\ score})$