HEPData entries for precision measurements

Louie Corpe (UCL)

LHC Electroweak Working Group General Meeting

October 2020

Introduction and Context

- As part of the Yellow Report process, a set of new recommendations have been drafted for what material is needed to be preserved in HEPData
- These recommendations have been discussed in relevant fora in CMS, ATLAS, LHCb and ALICE, with broad agreement
- But some technical challenges remain for precision analyses, due to the large amount of information to be stored
- This talk will review the recommendations for precision analysis, and will use examples of measurements which throw up some technical challenges for storing the HEPData material

HEPData recommendations in the Yellow Report

HEPData

Repository for publication-related High-Energy Physics data

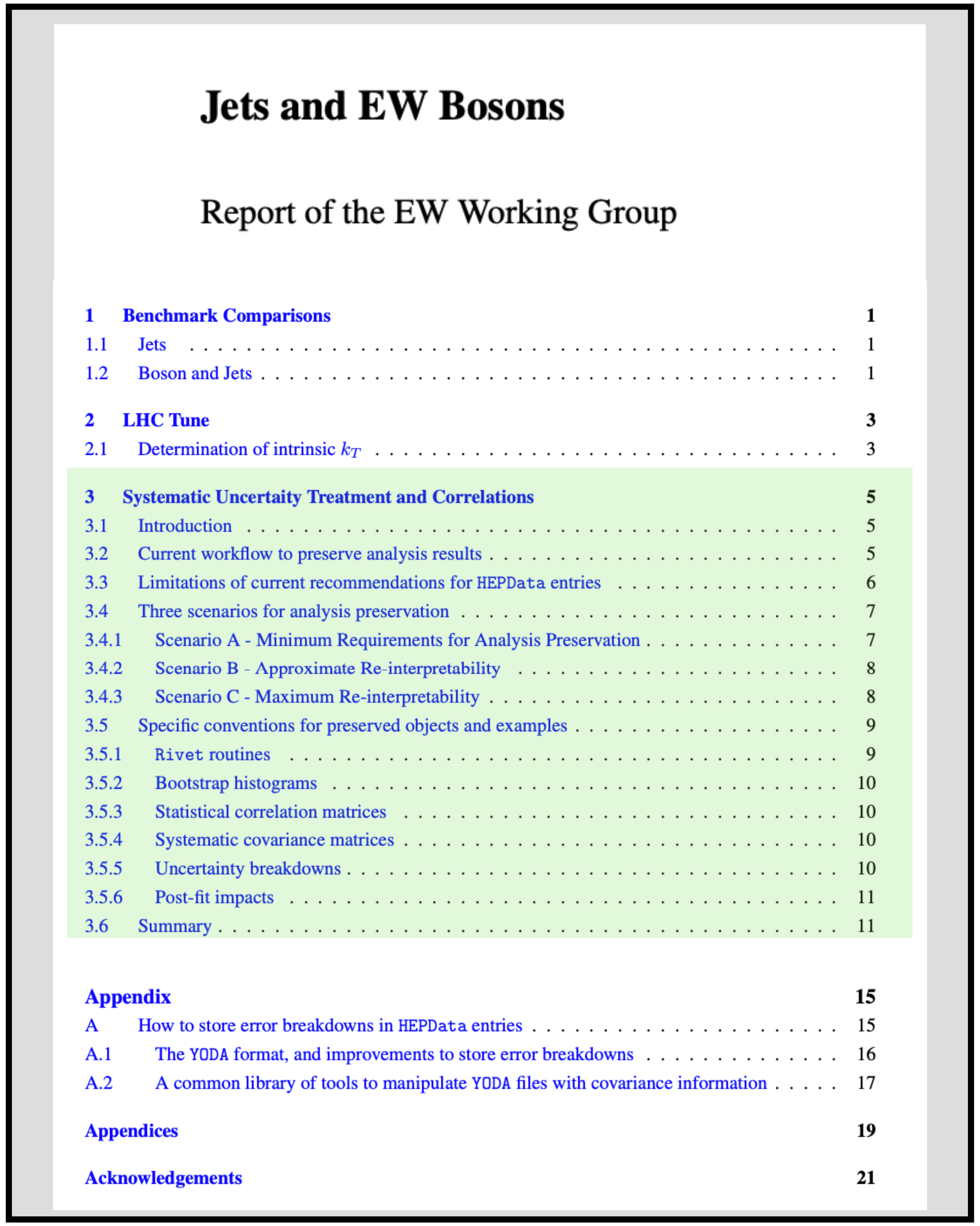
Update to HEPData recommendations

HEPData recommendations to be made in YR

Defines 3 scenarios for levels of information to provide on HEPData

Gives concrete recommendations for the format of objects which are to be stored

Has been discussed in various fora, in CMS, ATLAS, LHCb, ALICE… with lots of great input from the community



Jets and EW Bosons

Report of the EW Working Group

3 levels of re-interpretation

- Identify different levels of recommendations, depending on the analysis type and how re-interpretable it needs to be:

Scenario A - Minimum Requirements for Analysis Preservation
Scenario B - Approximate Re-interpretability
Scenario C - Maximum Re-interpretability

Minimum for a search to be re-useable

Not necessarily enough for strict combinations... but good enough for many analyses (especially searches)

Best case - aims to provide maximal information for reinterpretations. Should be gold standard for precision measurements

- The scenarios are **not intended to be “strict”**, but are more **designed to get groups thinking about what their intended level of re-interpretability** is, and **what they should preserve** as a result

A - Minimal Scenario

- Minimum amount of info for result to be re-used meaningfully.
  e.g if **only rough estimate of MC/data agreement** or **sensitivity to new models** needed

- **Phase Space Definition**: Ideally, **runnable code snippet (eg Rivet…)** if not...
  - **detailed description of the region of interest**
  - **per-object efficiency tables for non-standard objects**
  - **explicit definitions of each variable used in the selection**,
  - **cutflows of the effect of each selection** on well-defined signals
- **Statistical correlations:** omitted in this minimal scenario. **Stat error per bin** still needed (assumed uncorrelated between bins) separate from systematics.
- **Systematic correlations: uncertainty breakdown** of **major sources for multi-bin SRs**
- **Generator Prediction**: **SM prediction of MC generators**, with theory uncertainty if possible

B - Medium Scenario

- For standard measurements or searches to be **re-interpreted approximately**. E.g **tuning , and recasting of searches, repeat of statistical analysis**

- **Phase space definition**: **Runnable Code Snippet analysis** must be provided concurrently with arXiv submission
- **Statistical correlations**: **correlation matrices**. *Can't infer correlations between analyses, but OK if re-interpreting in isolation.* ***Not needed if likelihood given (eg pyhf)***
- **Systematic correlations**:
  - EITHER likelihood in eg `pyhf` format **(unc. breakdown/cov matrices not needed in this case)**
  - OR **uncertainty breakdown:** effect of each major uncertainty source/NP on each bin
  - OR, covariance matrix for each distribution: e.g. for simplified likelihoods
- **Generator Prediction**: include **SM prediction from latest MC generators** with breakdown of theory uncertainty if possible

C - Maximal Scenario

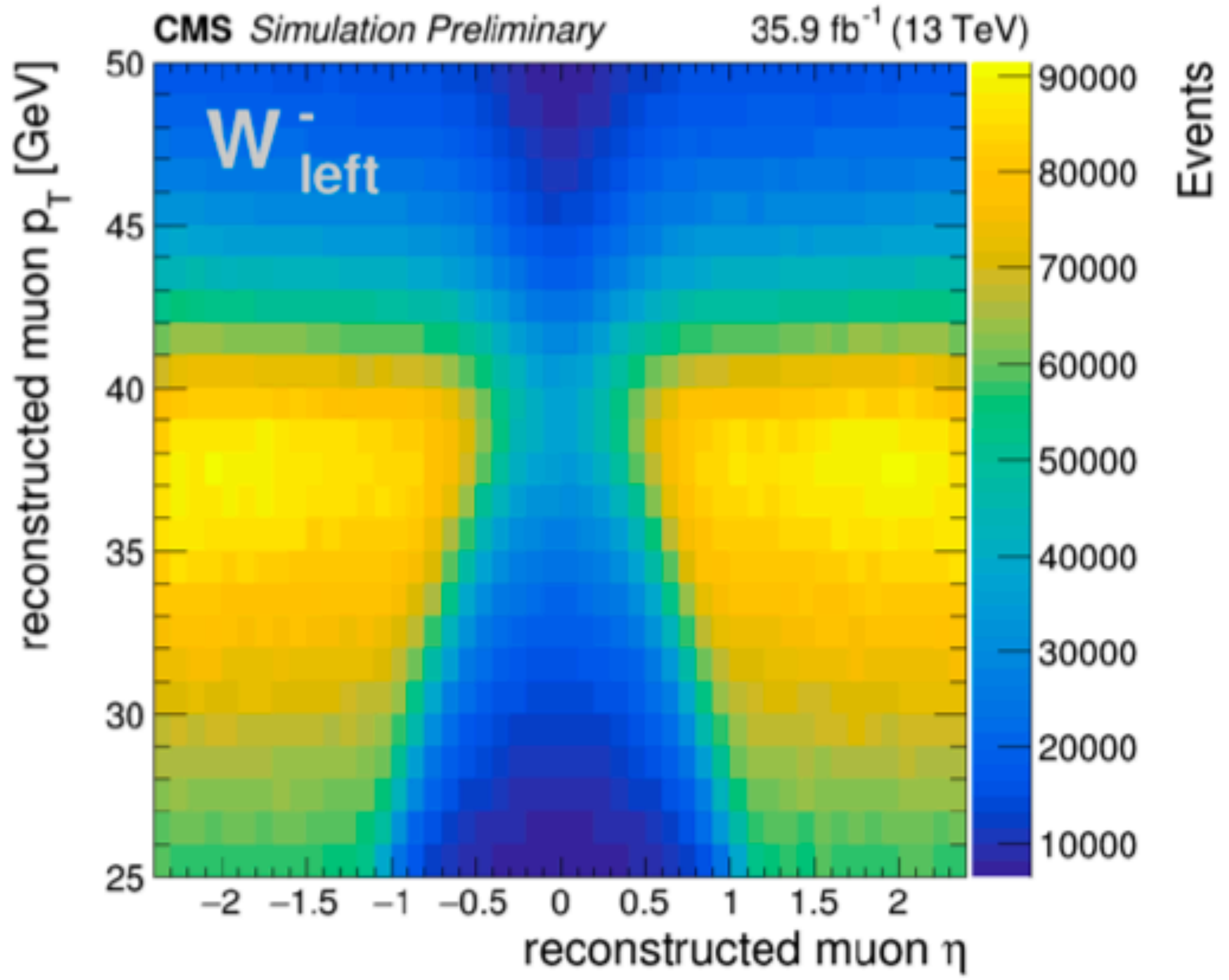
- **For precision analyses:** for **future combinations**, **measurements of SM parameters, PDF fitting...** Enough info for exact combination

- **Phase space definition:** **Particle-level Rivet analysis** must be provided concurrently with arXiv submission
- **Stat correlations: Bootstrap Replicas** attached to HEPData entry [plan for Bootstrap code to be made public]
- **Syst correlations** as detailed uncertainty breakdown, with **no grouping of NPs** (e.g. for JES, use full granularity of NPs) OR likelihood (eg `pyhf`) OR “enlarged” covariance matrix with columns for each bin and uncertainty source
- **Generator Prediction**: include **SM prediction from latest MC generators** w/ breakdown of theory uncertainty if possible

Some examples of technical hurdles

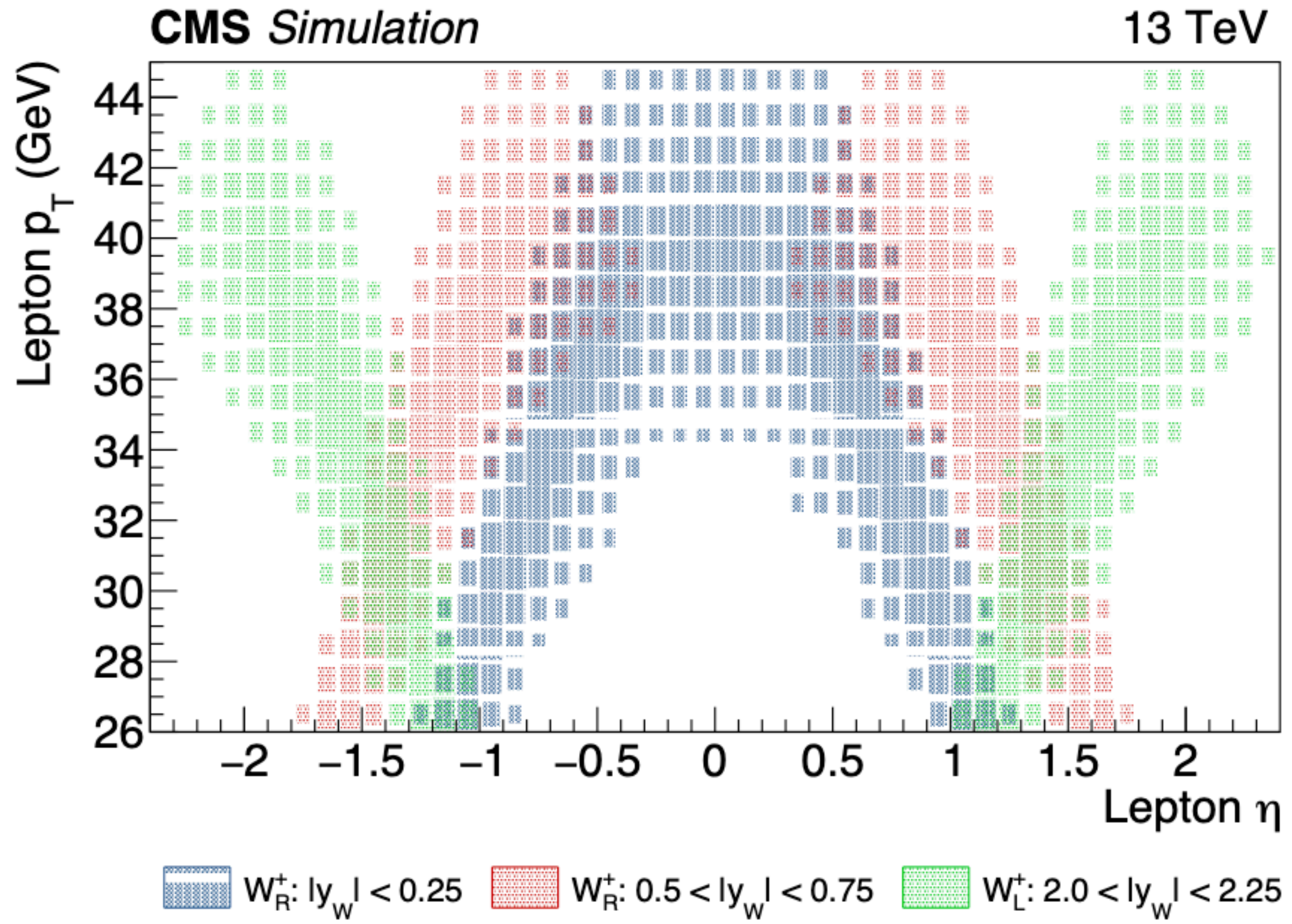
(when ambition and reality collide)

CMS W polarisation



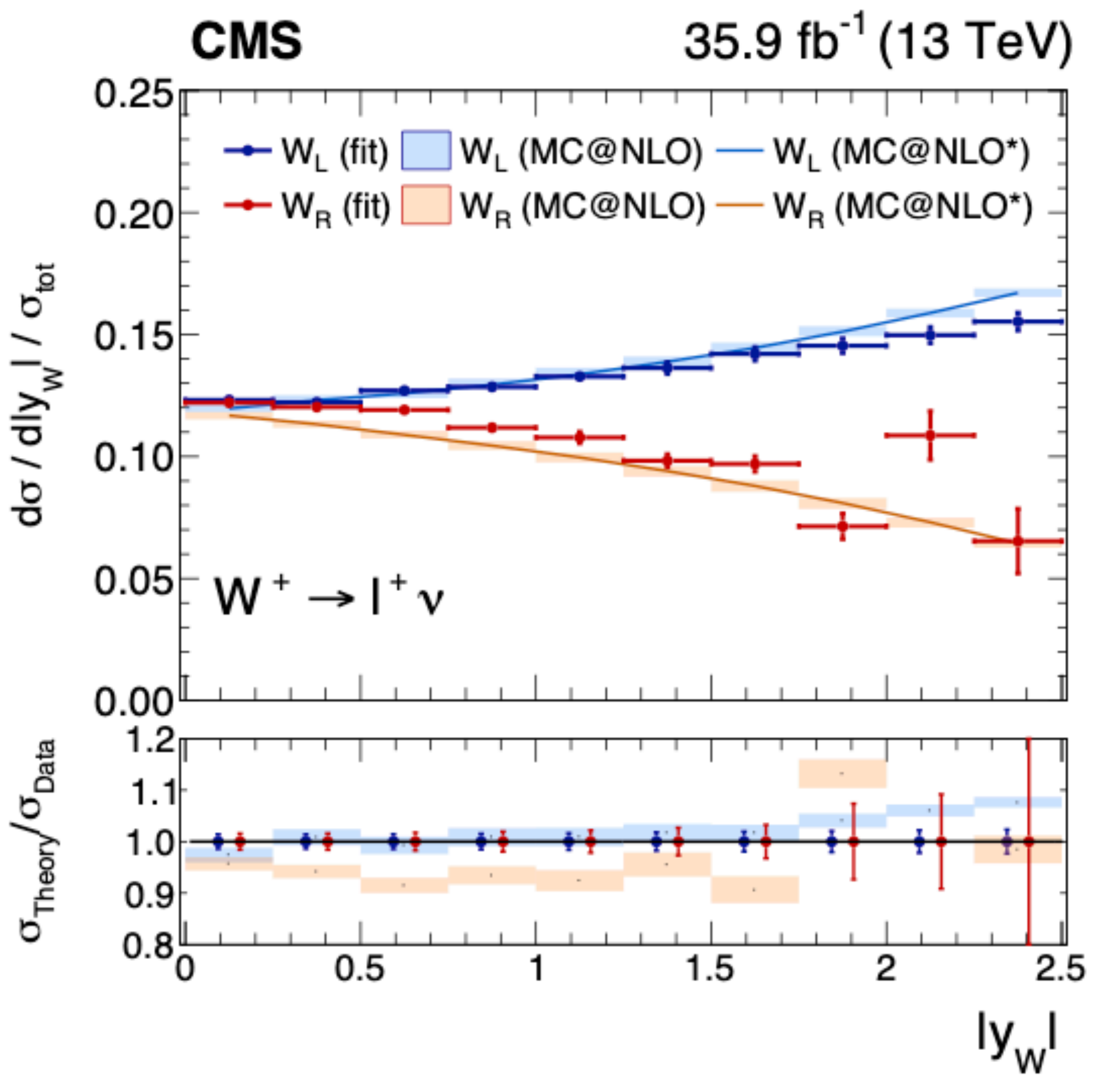
CMS measurement of W helicity/rapidity

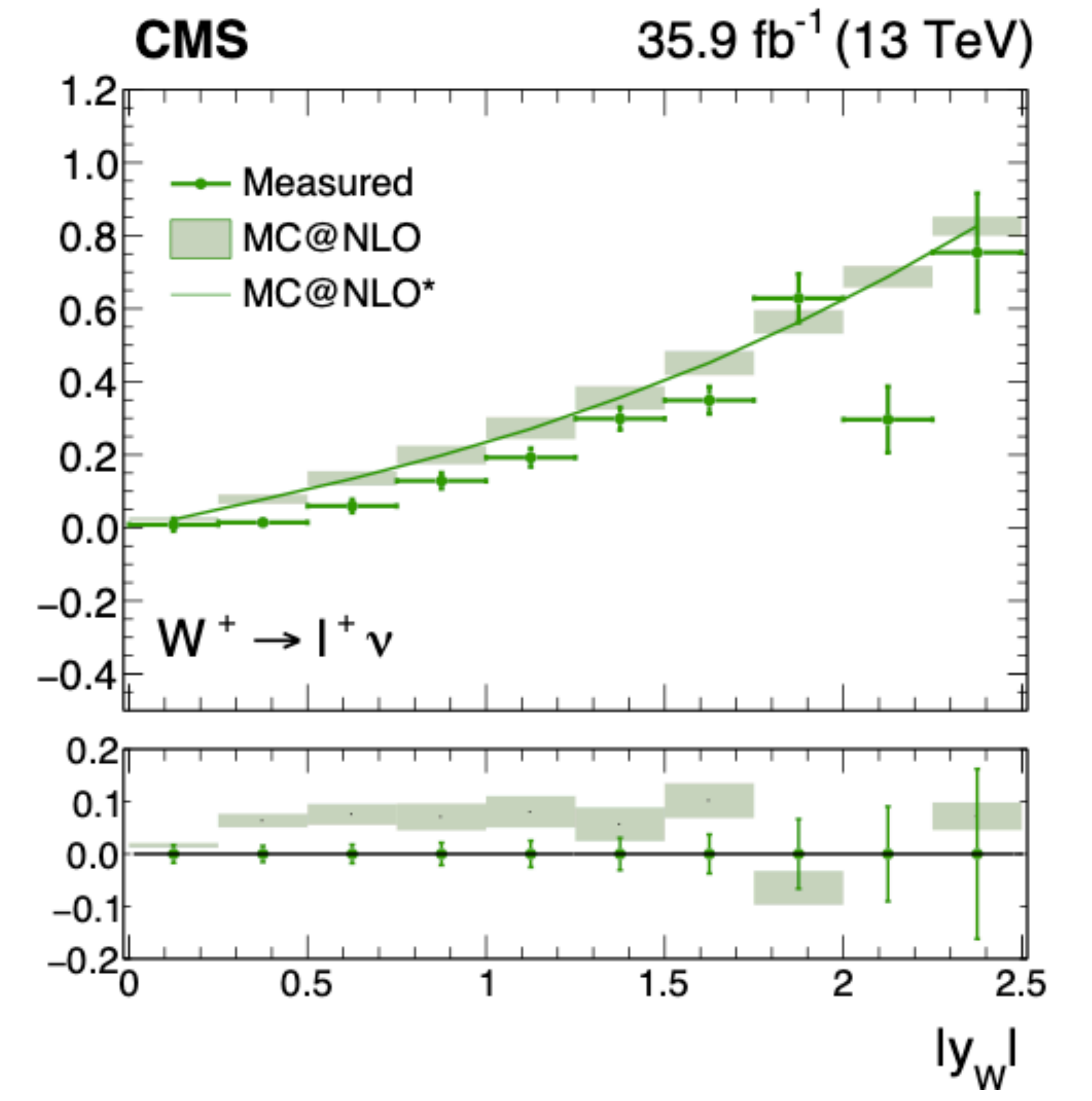
- Submitted to PRD (https://arxiv.org/abs/2008.04174). See also: https://indico.cern.ch/event/891674
- W/Z decays characterised by 5-dim diff cross-section, as a function of p_T^V, y^V, m^V, φ, θ (decay angles of leptons in Collins-Soper frame)
- Differential cross-section and charge asymmetries sensitive to proton PDFs, but can lead to circular dependence on PDF results, and loss of info if integrating over variables
- W production is qqbar-induced at LHC, so helicity determined by direction of W wrt q. Only 2 amplitude/helicity states!
  - Full information on valence quark PDFs is contained in differential cross-section as a function of rapidity, broken down into 2 helicity states.
  - Information can be extracted from template fit to charged-lepton p_T/η

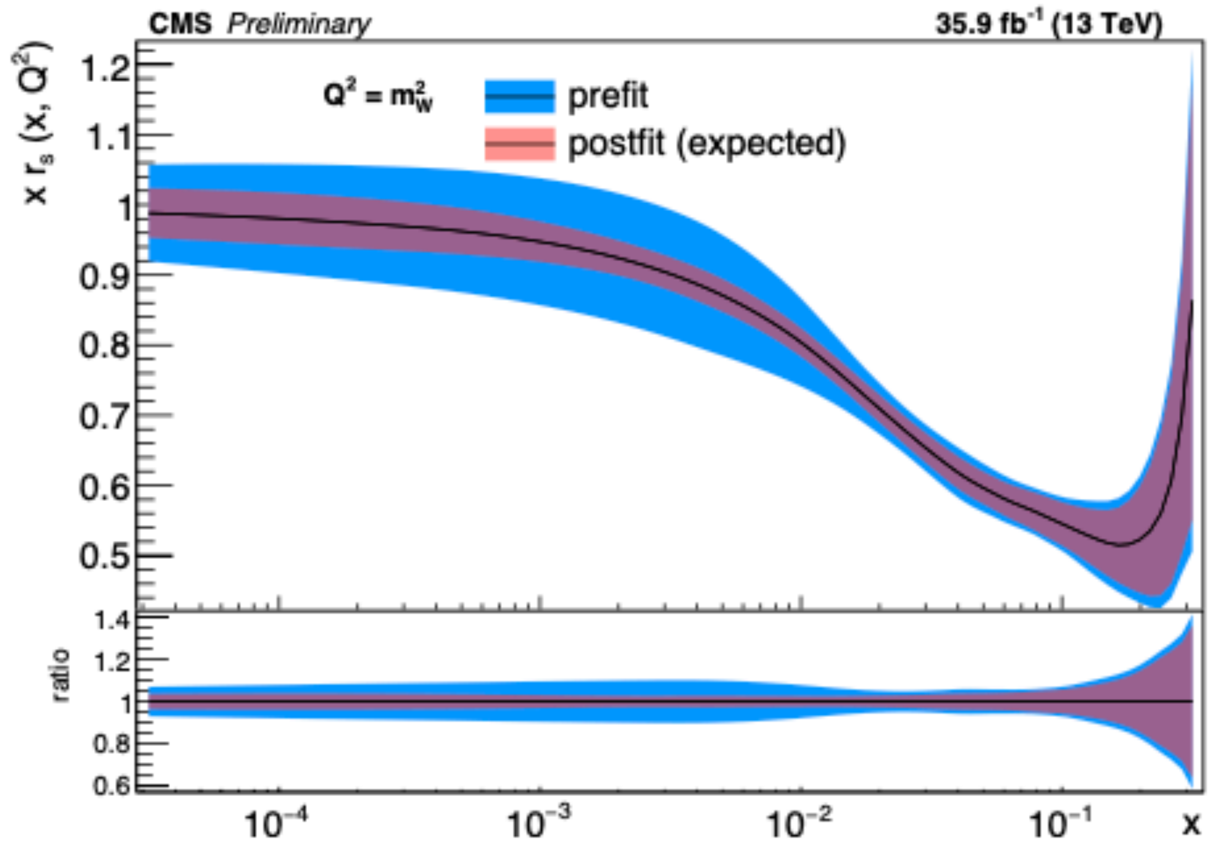


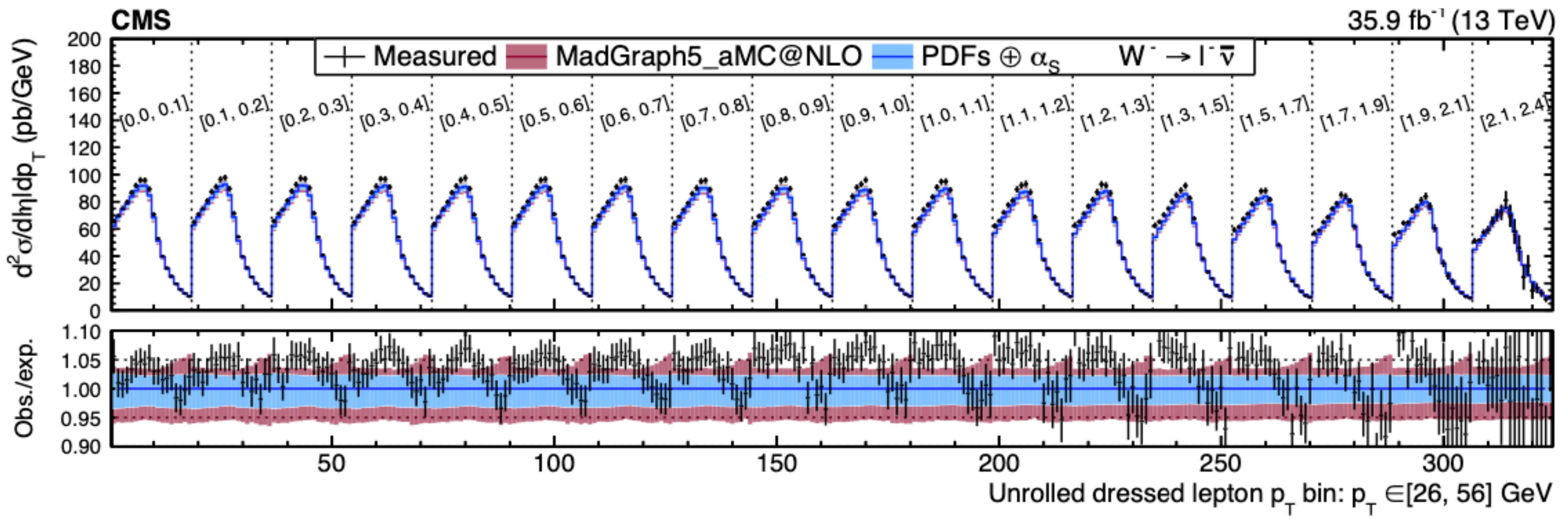
CMS measurement of W helicity/rapidity

- Measurement extracts polarised+unpolarized cross-sections, asymmetries and double-differential lepton cross-sections
- Clear potential to constraint PDFs !
  - Need to ensure all relevant information is made public to allow future PDF studies



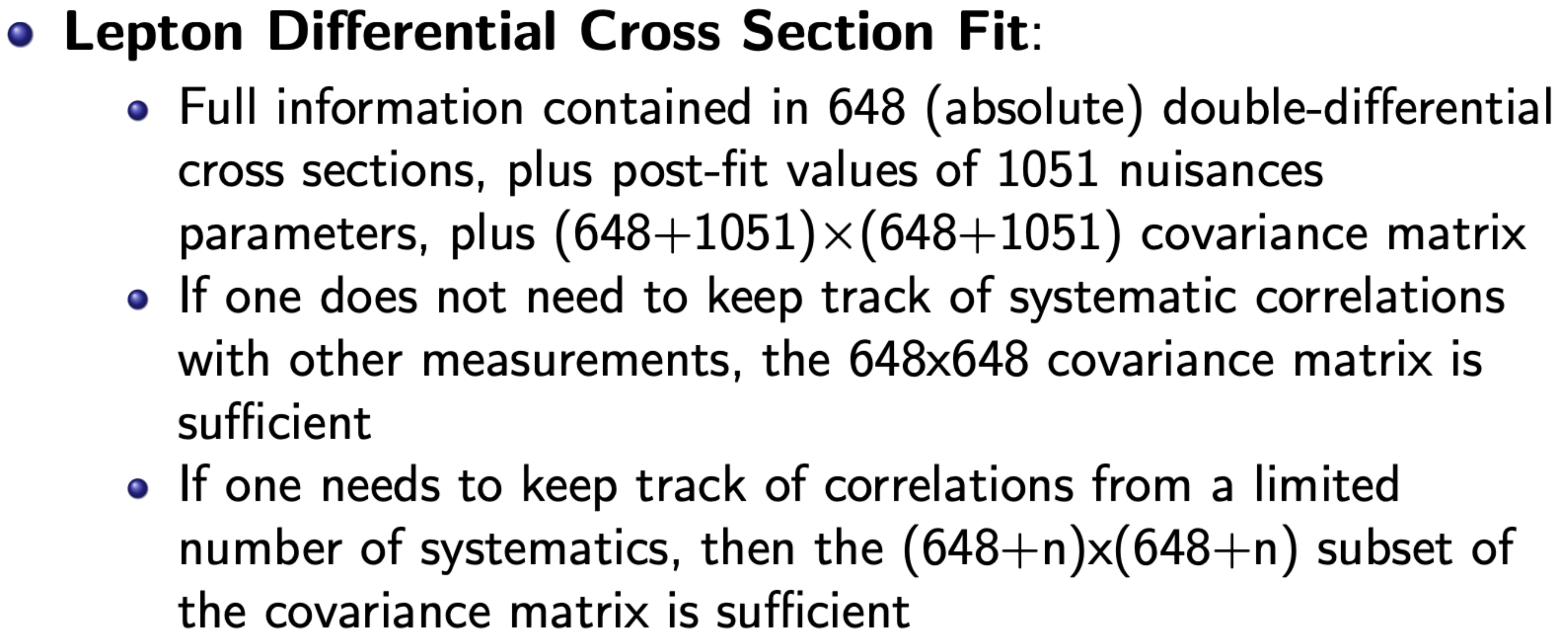




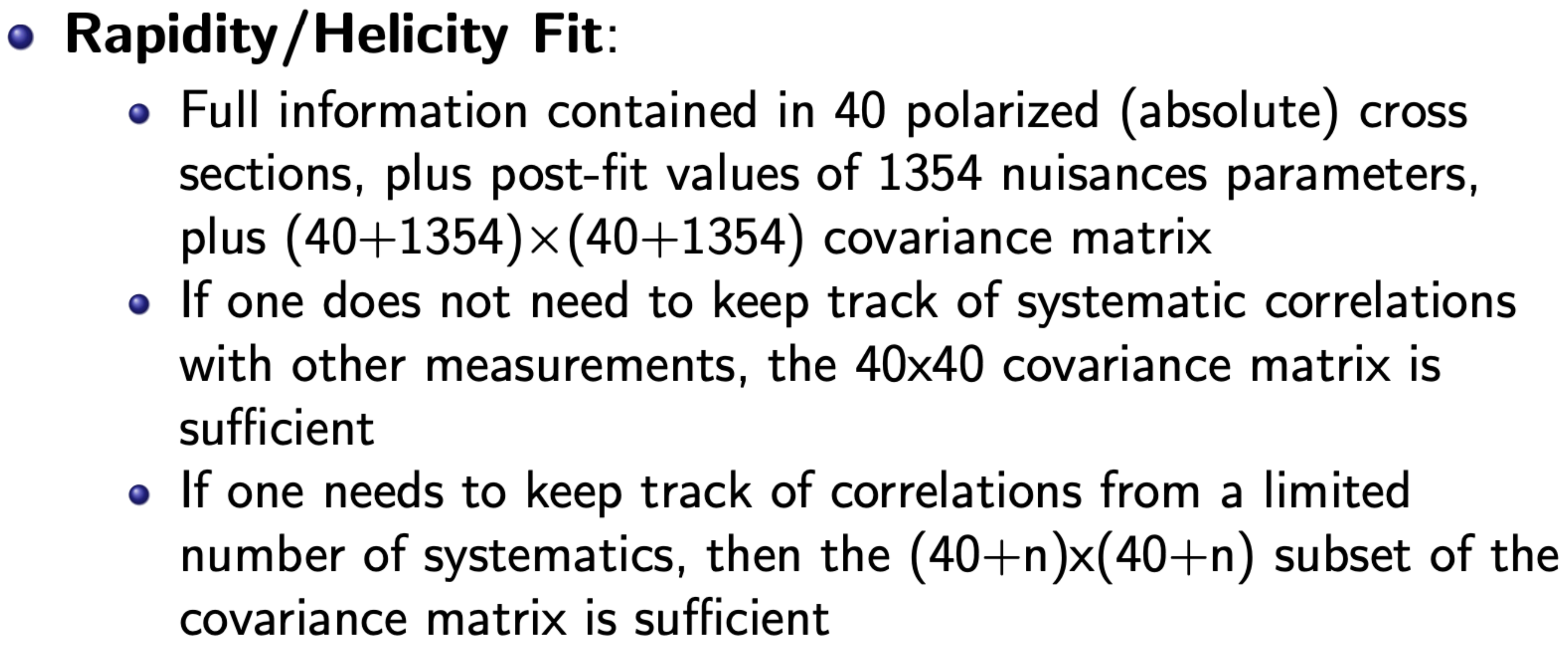


CMS measurement of W helicity/rapidity

- All results derived from “just” three likelihood fits: y/helicity fit, lepton double-differential XS fit, fixed POI fits for PDF constraints
  - Normally these results would be included in HEPData entry, but the O(1500x1500) matrices are technical challenge
    - Include simply a link to CMS public page, or perhaps store in Root-based format as additional material?
  - Other questions:
    - Is full granularity of results needed on HEPData? Or just reduced subset?
    - Will a Rivet plugin be needed?
    - Is it useful to include Bootstraps Replicas for the stat uncertainties ?
    - Also provide a Hessian in addition to covariance?

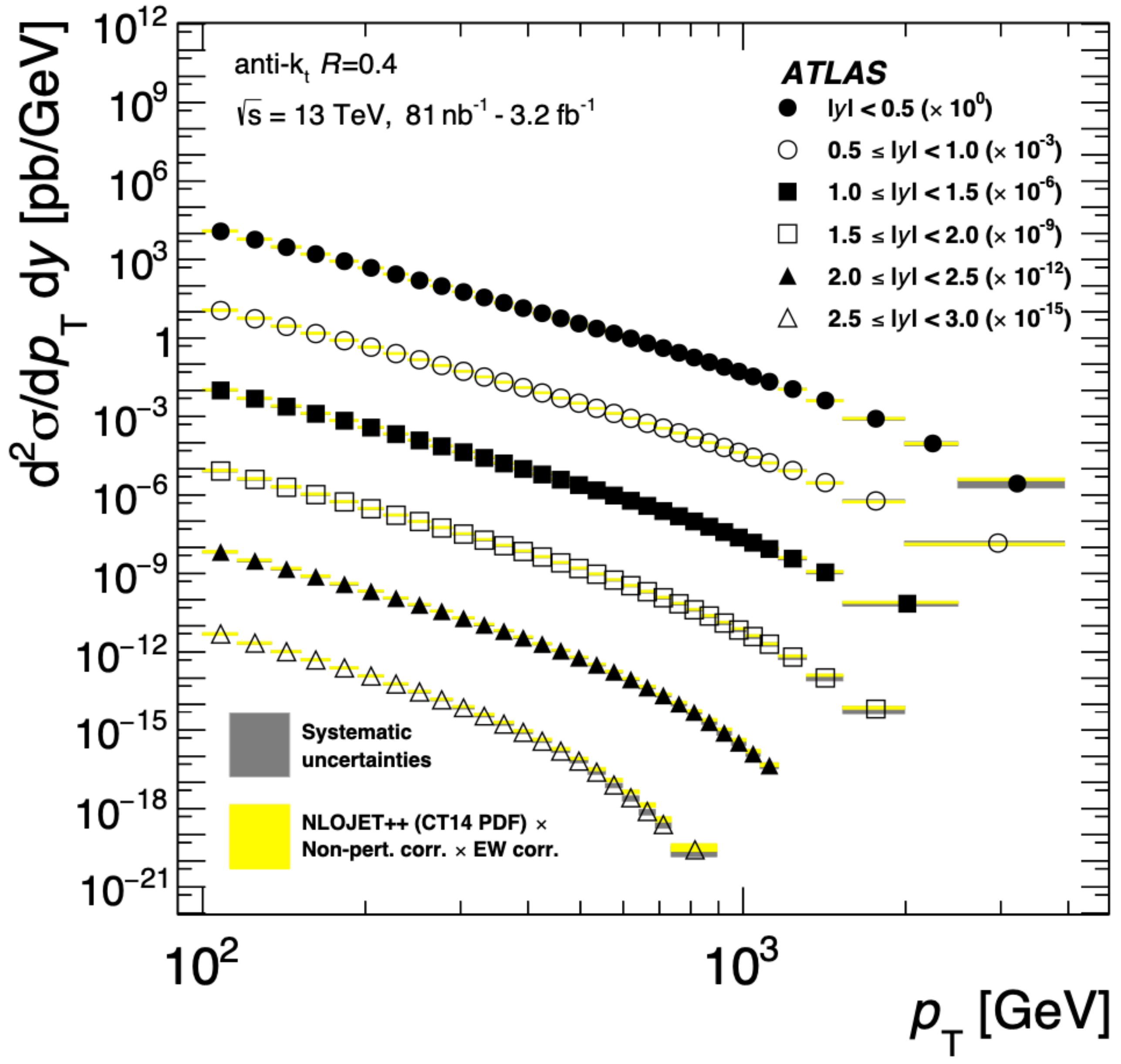


- **Lepton Differential Cross Section Fit**:
  - Full information contained in 648 (absolute) double-differential cross sections, plus post-fit values of 1051 nuisances parameters, plus (648+1051)×(648+1051) covariance matrix
  - If one does not need to keep track of systematic correlations with other measurements, the 648x648 covariance matrix is sufficient
  - If one needs to keep track of correlations from a limited number of systematics, then the (648+n)x(648+n) subset of the covariance matrix is sufficient



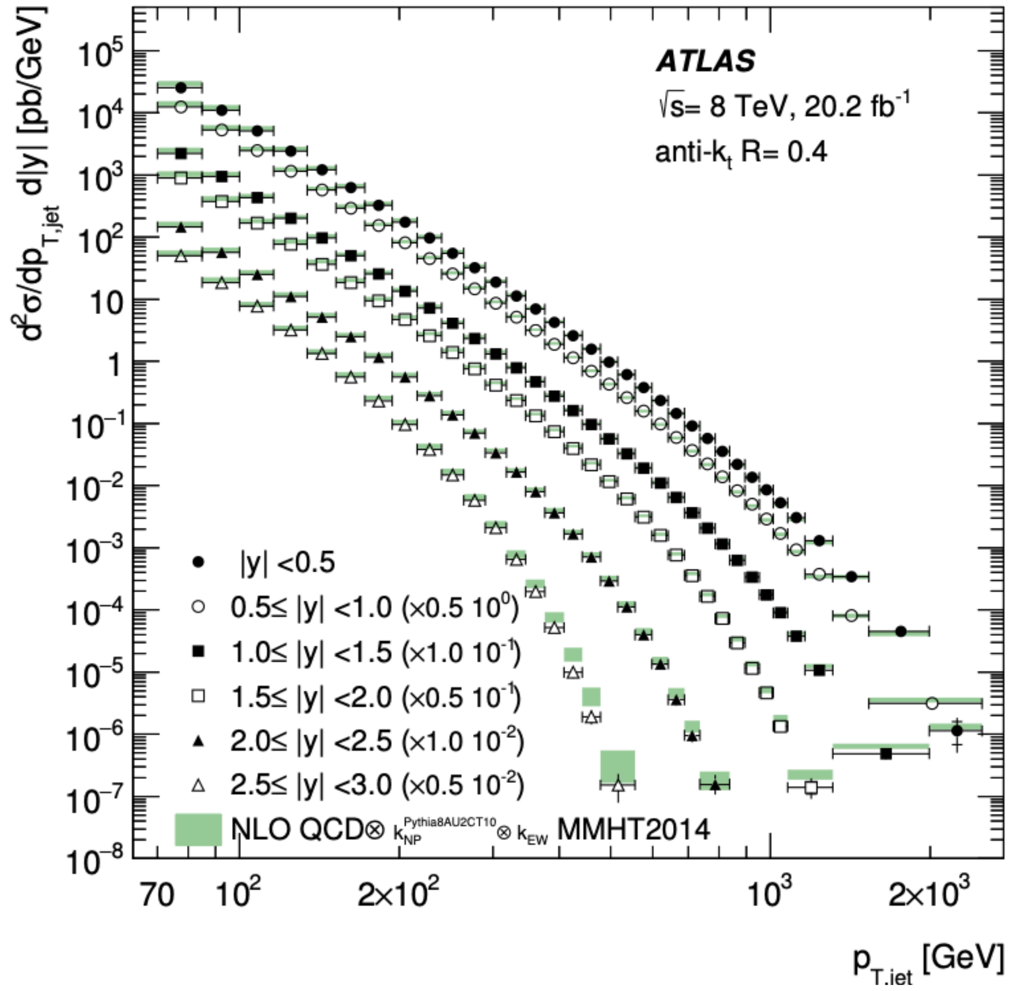
- **Rapidity/Helicity Fit**:
  - Full information contained in 40 polarized (absolute) cross sections, plus post-fit values of 1354 nuisances parameters, plus (40+1354)×(40+1354) covariance matrix
  - If one does not need to keep track of systematic correlations with other measurements, the 40x40 covariance matrix is sufficient
  - If one needs to keep track of correlations from a limited number of systematics, then the (40+n)x(40+n) subset of the covariance matrix is sufficient

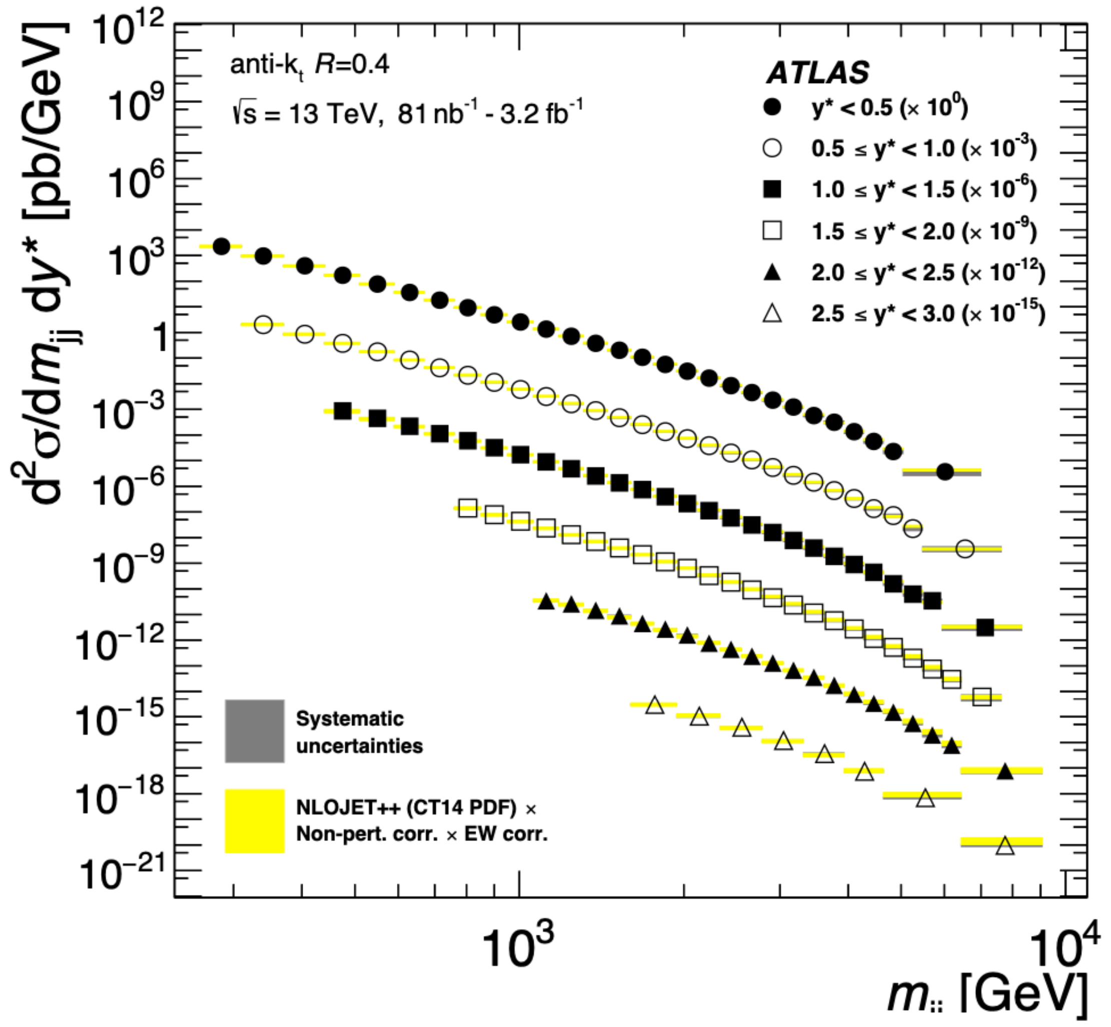
ATLAS inclusive jet and dijet cross-sections at 13 TeV

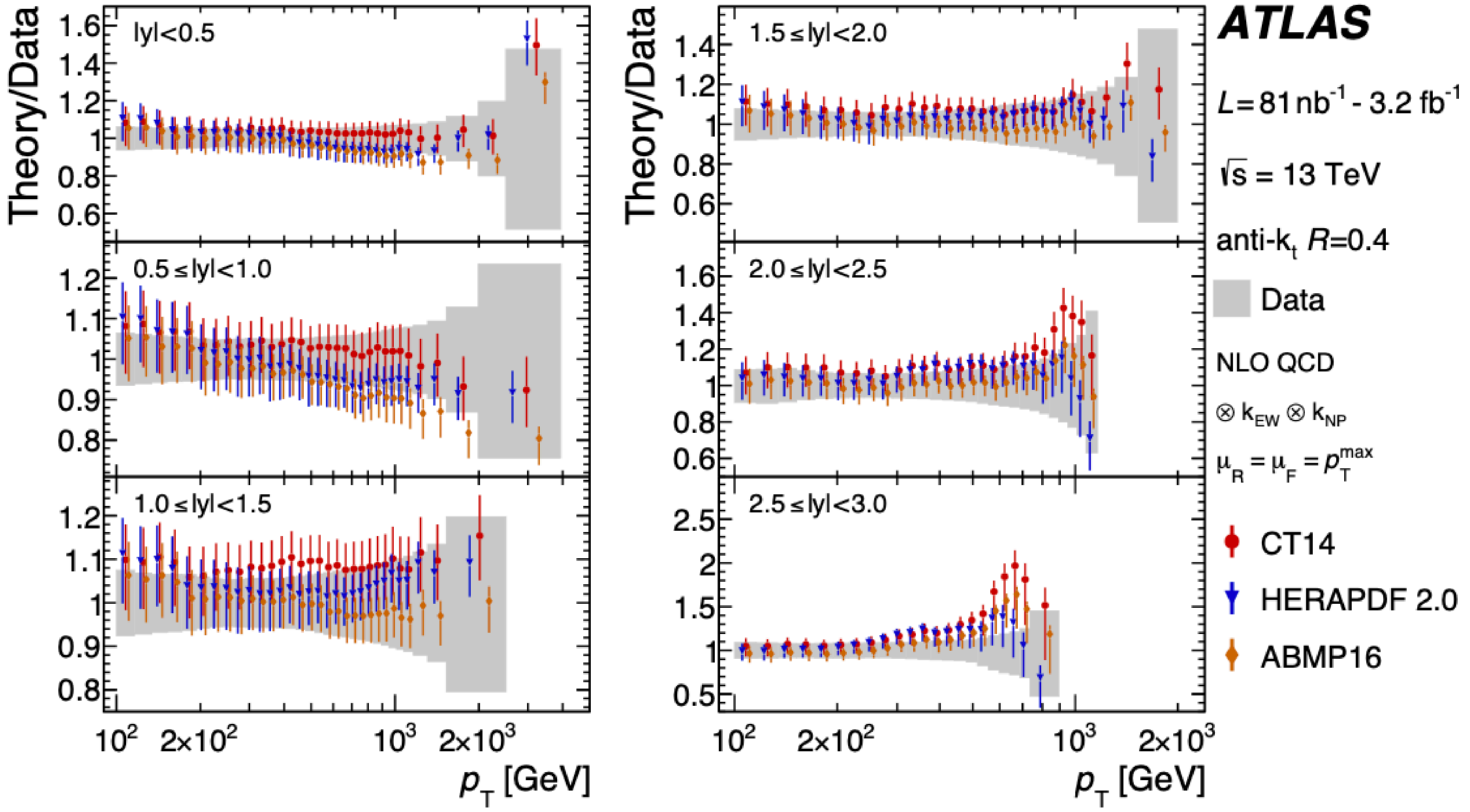


ATLAS jet and dijet cross-sections

- Results from 8 and 13 TeV, measure inclusive jet cross-section double-differentially in jet p_T and $|y|$.
- 13 TeV results also measure dijet cross-section
  - https://arxiv.org/pdf/1711.02692.pdf
  - https://arxiv.org/pdf/1706.03192.pdf
- Jets clustered with anti-kt (R=0.4)
- Results are corrected for detector effects (unfolded) using Iterative Bayesian technique (dynamically stabilised)
- Test of perturbative QCD, and various PDF sets



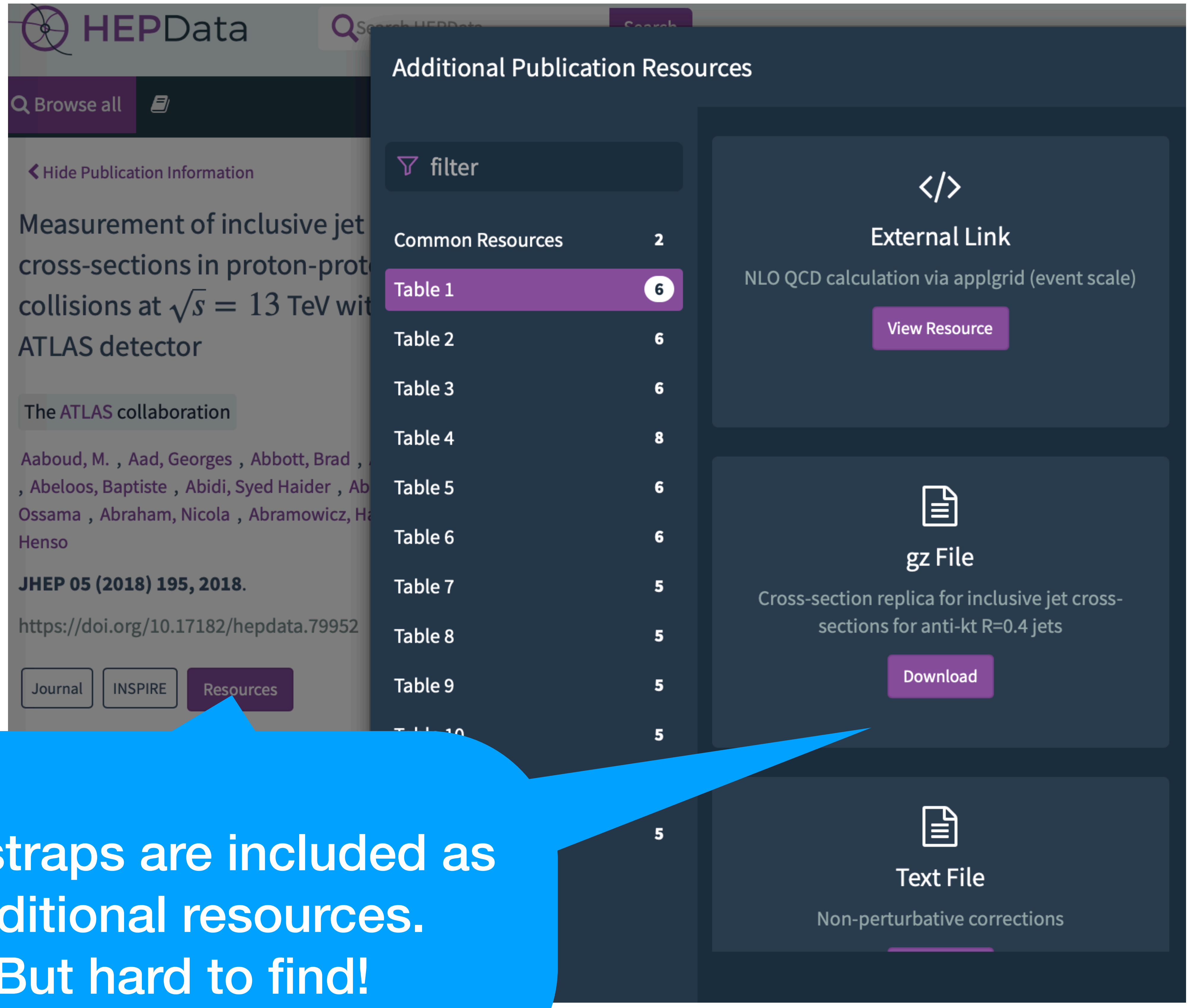




ATLAS jet and dijet cross-sections

- Propagating statistical uncertainties through unfolding + challenge of statistical fluctuations in propagation/evaluation of uncertainties
- Bootstrap method: assign unique seed to each event/run/sample number, and produce N (~10,000) Replicas. Each time histogram is filled, replicas are filled varied by a random Poisson(1). Whole analysis, including unfolding and uncertainty evaluation, automatically repeated for each replica, and the RMS of final replica results gives the final uncertainty
    - Note being prepared to accompany ATLAS's bootstrap ROOT classes
    - Since seeds are set uniquely by run/event/sample number, the bootstraps can be used for future evaluation of statistical correlation between operate measurements
- But how to store this information on HEPData ?
- Measurement also has technical challenge of large number of bins (~170) and uncertainties (~300)

https://www.hepdata.net/record/ins1634970?version=1



Measurement of inclusive jet cross-sections in proton-prot collisions at $\sqrt{s} = 13$ TeV wit ATLAS detector

The ATLAS collaboration

Aaboud, M. , Aad, Georges , Abbott, Brad , , Abeloos, Baptiste , Abidi, Syed Haider , Ab Ossama , Abraham, Nicola , Abramowicz, Ha Henso

JHEP 05 (2018) 195, 2018.

https://doi.org/10.17182/hepdata.79952

Additional Publication Resources

Common Resources	2
Table 1	6
Table 2	6
Table 3	6
Table 4	8
Table 5	6
Table 6	6
Table 7	5
Table 8	5
Table 9	5
Table 10	5
	5

External Link

NLO QCD calculation via applgrid (event scale)

gz File

Cross-section replica for inclusive jet cross-sections for anti-kt R=0.4 jets

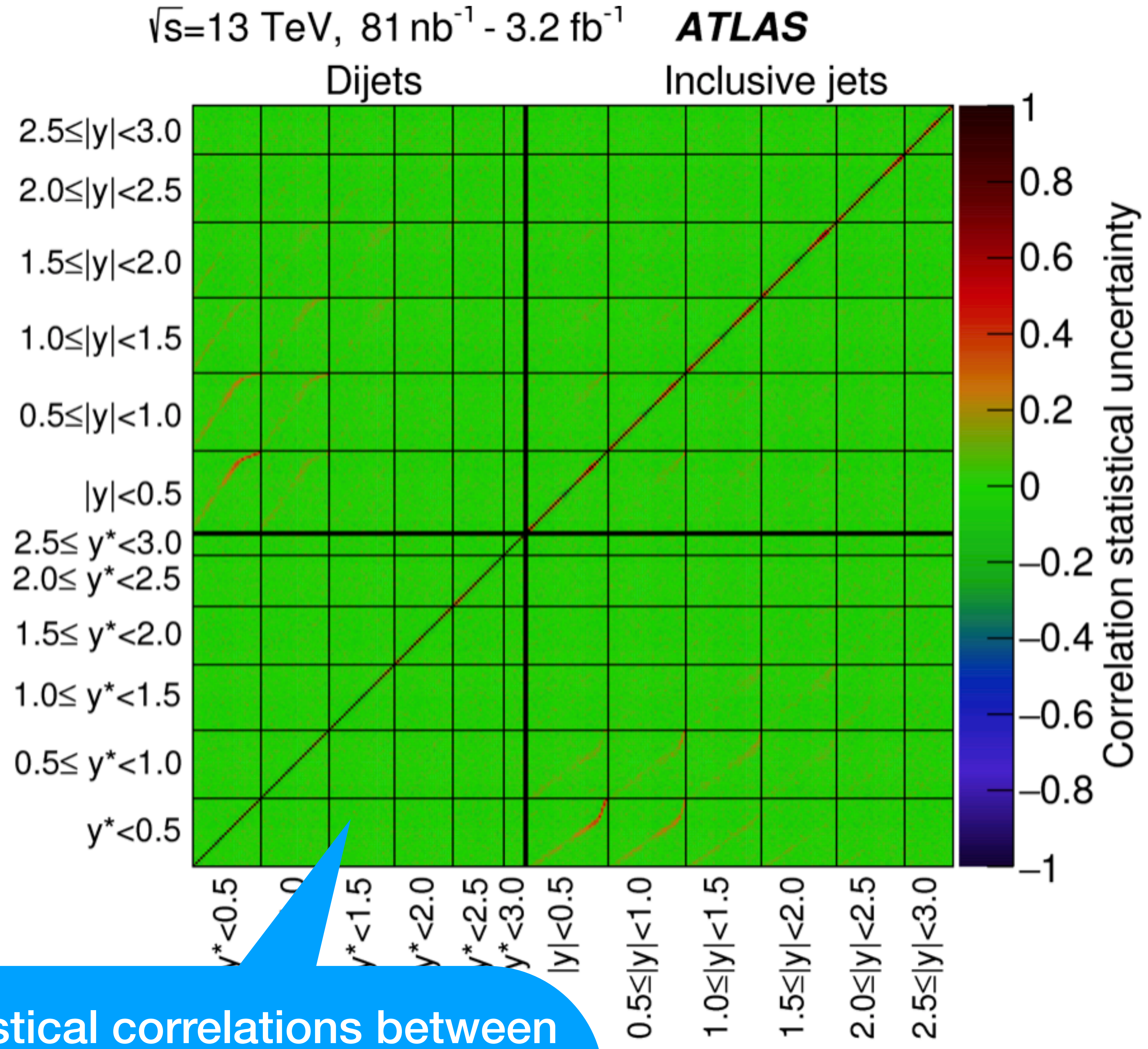
Text File

Non-perturbative corrections

Bootstraps are included as additional resources.
But hard to find!

ATLAS jet and dijet cross-sections

- Propagating statistical uncertainties through unfolding + challenge of statistical fluctuations in propagation/evaluation of uncertainties
- Bootstrap method: assign unique seed to each event/run/sample number, and produce N (~10,000) Replicas. Each time histogram is filled, replicas are filled varied by a random Poisson(1). Whole analysis, including unfolding and uncertainty evaluation, automatically repeated for each replica, and the RMS of final replica results gives the final uncertainty
  - Note being prepared to accompany ATLAS's bootstrap ROOT classes
  - Since seeds are set uniquely by run/event/sample number, the bootstraps can be used for future evaluation of statistical correlation between operate measurements
- But how to store this information on HEPData ?
- Measurement also has technical challenge of large number of bins (~170) and uncertainties (~300)

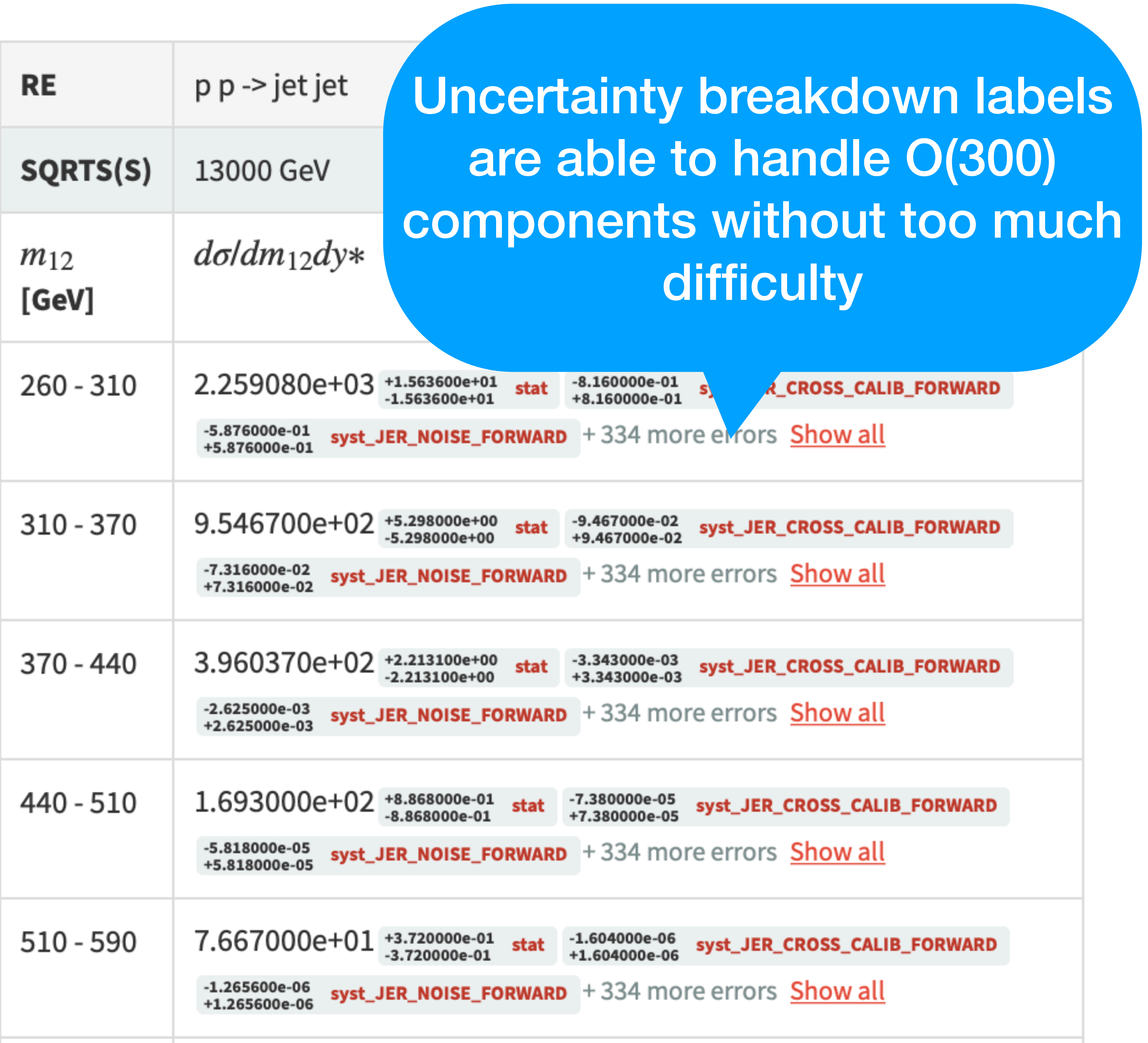


statistical correlations between inclusive jets and dijets obtained from Bootstrap replicas

ATLAS jet and dijet cross-sections

- Propagating statistical uncertainties through unfolding + challenge of statistical fluctuations in propagation/evaluation of uncertainties
- Bootstrap method: assign unique seed to each event/run/sample number, and produce N (~10,000) Replicas. Each time histogram is filled, replicas are filled varied by a random Poisson(1). Whole analysis, including unfolding and uncertainty evaluation, automatically repeated for each replica, and the RMS of final replica results gives the final uncertainty
  - Note being prepared to accompany ATLAS's bootstrap ROOT classes
  - Since seeds are set uniquely by run/event/sample number, the bootstraps can be used for future evaluation of statistical correlation between operate measurements
- But how to store this information on HEPData ?
- Measurement also has technical challenge of large number of bins (~170) and uncertainties (~300)

https://www.hepdata.net/record/ins1634970?version=1



RE	p p -> jet jet
SQRTS(S)	13000 GeV
m_{12} [GeV]	$d\sigma/dm_{12}dy*$
260 - 310	2.259080e+03 +1.563600e+01 -1.563600e+01 stat -8.160000e-01 +8.160000e-01 sy R_CROSS_CALIB_FORWARD -5.876000e-01 +5.876000e-01 syst_JER_NOISE_FORWARD + 334 more e rors Show all
310 - 370	9.546700e+02 +5.298000e+00 -5.298000e+00 stat -9.467000e-02 +9.467000e-02 syst_JER_CROSS_CALIB_FORWARD -7.316000e-02 +7.316000e-02 syst_JER_NOISE_FORWARD + 334 more errors Show all
370 - 440	3.960370e+02 +2.213100e+00 -2.213100e+00 stat -3.343000e-03 +3.343000e-03 syst_JER_CROSS_CALIB_FORWARD -2.625000e-03 +2.625000e-03 syst_JER_NOISE_FORWARD + 334 more errors Show all
440 - 510	1.693000e+02 +8.868000e-01 -8.868000e-01 stat -7.380000e-05 +7.380000e-05 syst_JER_CROSS_CALIB_FORWARD -5.818000e-05 +5.818000e-05 syst_JER_NOISE_FORWARD + 334 more errors Show all
510 - 590	7.667000e+01 +3.720000e-01 -3.720000e-01 stat -1.604000e-06 +1.604000e-06 syst_JER_CROSS_CALIB_FORWARD -1.265600e-06 +1.265600e-06 syst_JER_NOISE_FORWARD + 334 more errors Show all

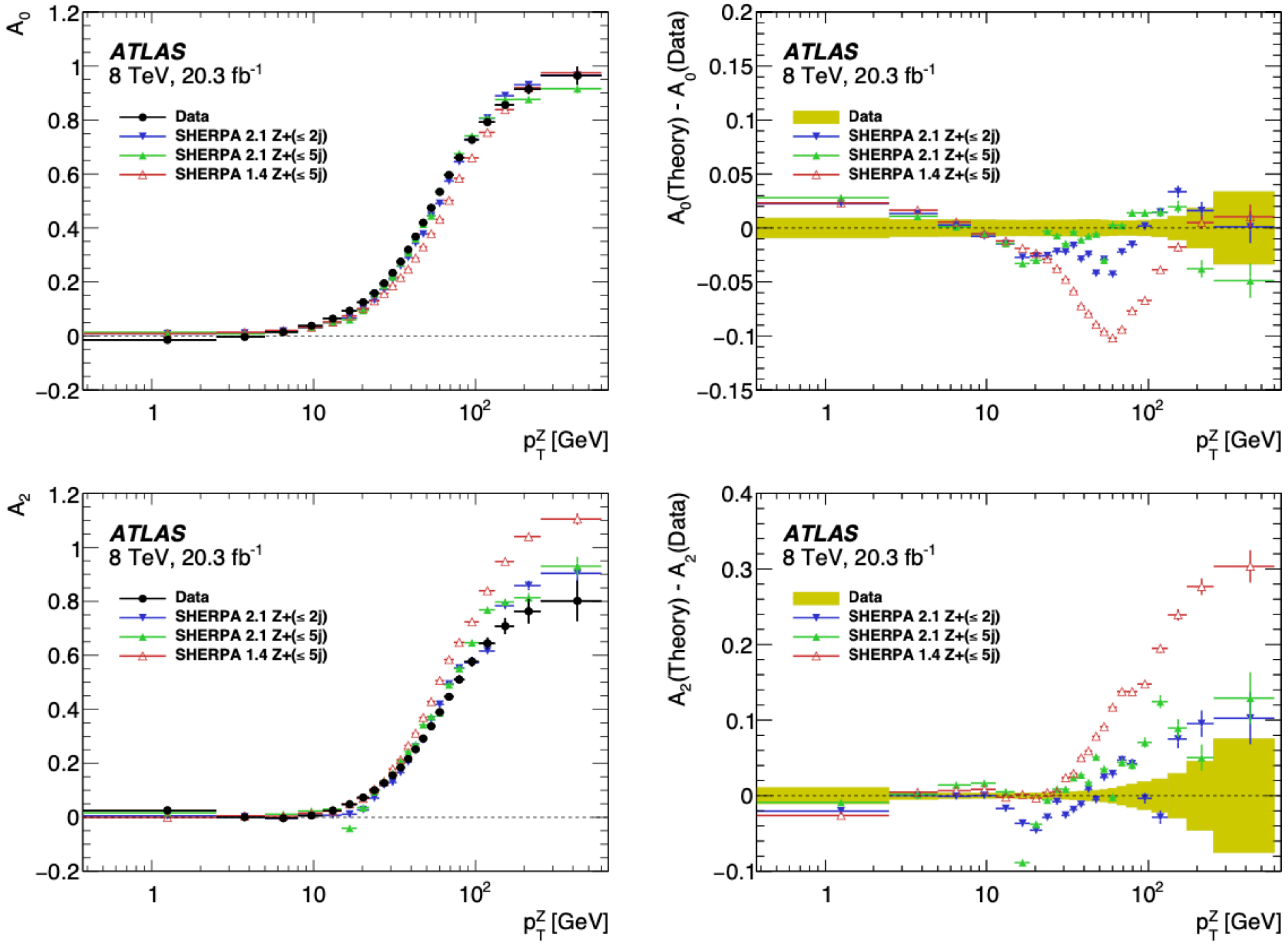
Uncertainty breakdown labels are able to handle O(300) components without too much difficulty

ATLAS 8 TeV angular coefficients in Z events

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{T}}^{Z}\,\mathrm{d}y^{Z}\,\mathrm{d}m^{Z}\,\mathrm{d}\cos\theta\,\mathrm{d}\phi} = \frac{3}{16\pi}\frac{\mathrm{d}\sigma^{U+L}}{\mathrm{d}p_{\mathrm{T}}^{Z}\,\mathrm{d}y^{Z}\,\mathrm{d}m^{Z}}$$
$$\Big\{(1+\cos^{2}\theta) + \frac{1}{2}\,A_{0}(1-3\cos^{2}\theta) + A_{1}\;\sin 2\theta\;\cos\phi$$
$$+\frac{1}{2}\,A_{2}\;\sin^{2}\theta\;\cos 2\phi + A_{3}\;\sin\theta\;\cos\phi + A_{4}\;\cos\theta$$
$$+A_{5}\;\sin^{2}\theta\;\sin 2\phi + A_{6}\;\sin 2\theta\;\sin\phi + A_{7}\;\sin\theta\;\sin\phi\Big\}.$$

ATLAS 8 TeV angular coefficients in Z events

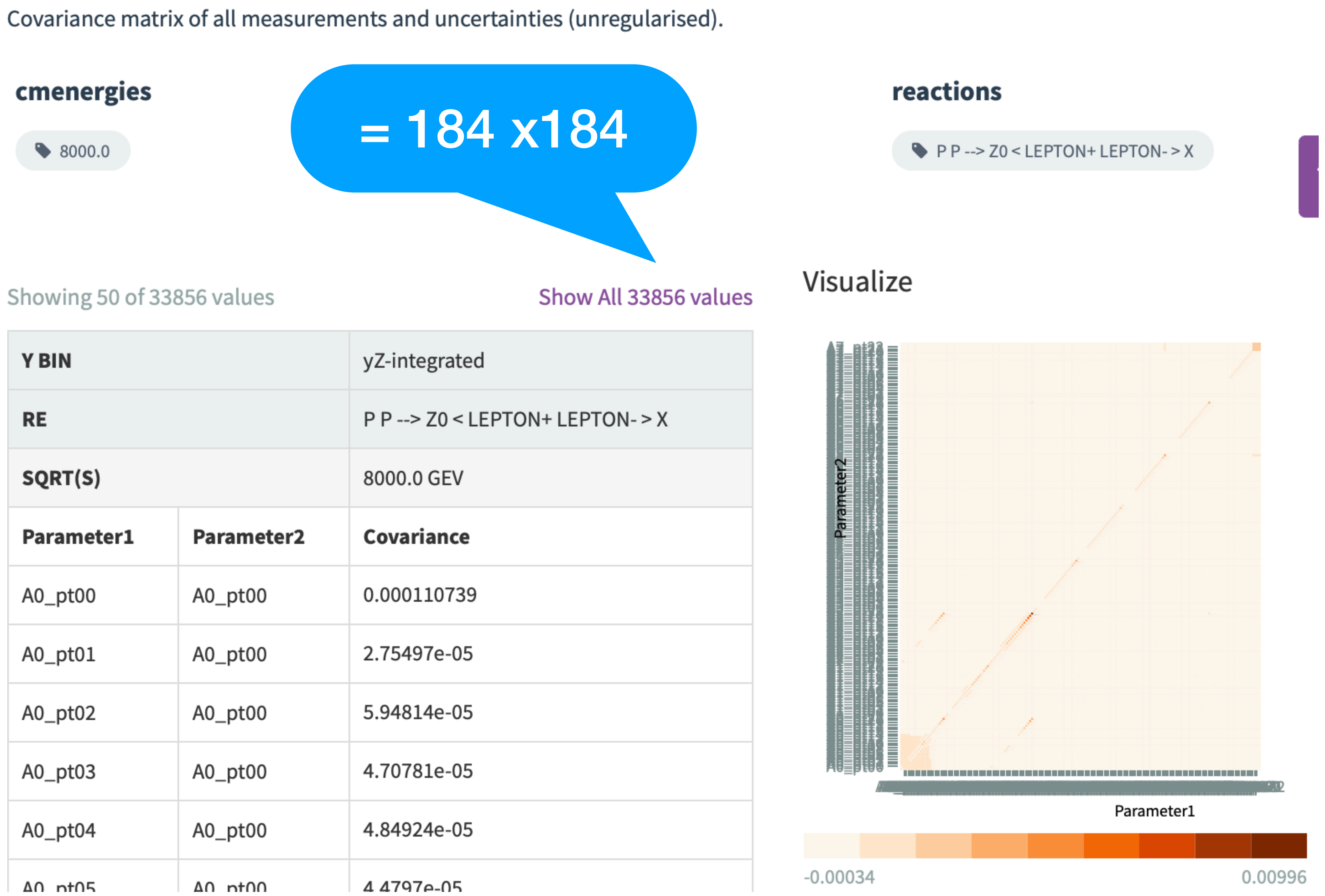
- Angular distributions of charged lepton pairs near the Z-peak probe the underlying QCD dynamics of Z boson production.
- The spin correlations of the leptons are described by helicity density matrix elements, which can be calculated in pQCD
- The cross-section can be factorised, and the polarised part can be expressed in terms of angular coefficients $A_{0\text{-}7}$, which encapsulate the $p_T{}^Z$, y^Z and m^Z dependence
- The coefficients are extracted from the data by fitting templates of the polynomial terms to the reconstructed angular distributions



ATLAS 8 TeV angular coefficients in Z events

- Full covariance matrix would be O(6000x6000)… !
- Instead, all covariance matrices are only 184 x 184, corresponding to the 23*8 = 184 Ais, so the stat. and syst. are folded into this reduced covariance matrix already.
- Would it be better to simply upload a binary file to HEPData, like for the Bootstraps?
- Usefulness of the full covariance matrix including all NPs in this case would be pretty limited, since the theory uncertainties are very small.
- One could always provide the covariance matrix of only the POIs and NPs that might be interested, and ignore the rows and columns of the uninteresting ones (like MC stats, for instance).

https://www.hepdata.net/record/76986



Covariance matrix of all measurements and uncertainties (unregularised).

cmenergies

8000.0

= 184 x184

reactions

P P --> Z0 < LEPTON+ LEPTON- > X

Showing 50 of 33856 values

Show All 33856 values

Y BIN		yZ-integrated
RE		P P --> Z0 < LEPTON+ LEPTON- > X
SQRT(S)		8000.0 GEV
Parameter1	**Parameter2**	**Covariance**
A0_pt00	A0_pt00	0.000110739
A0_pt01	A0_pt00	2.75497e-05
A0_pt02	A0_pt00	5.94814e-05
A0_pt03	A0_pt00	4.70781e-05
A0_pt04	A0_pt00	4.84924e-05
A0_pt05	A0_pt00	4.4797e-05

Visualize

Summary and Discussion

22

Louie Corpe, UCL (l.corpe@ucl.ac.uk)

Summary and Discussion

- In this talk I recalled the proposed HEPData recommendations which could be part of the YR.
- Formalises what is often standard practice anyway…
- But highlighted some precision analyses where the size of the HEPData material becomes challenging.
- Do we need to cater for these cases in the YR recommendations? What should we recommend?
  - Reduced covariance matrices?
  - Make more use of “additional material” section?
  - Direct links to collaboration pages?
  - Can HEPData be made more flexible wrt large uploads?