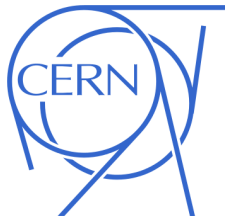
Experimental Constraints on PDFs for Precision Electroweak Measurements

Josh Bendavid



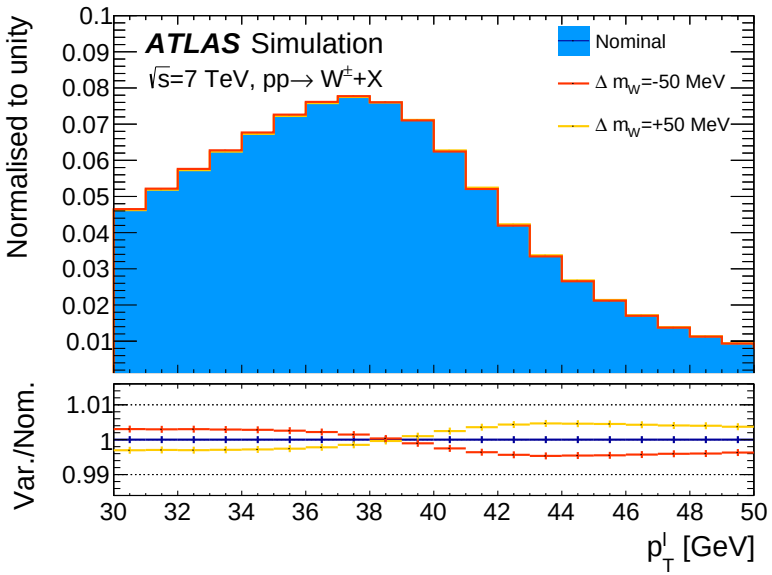
July 19, 2019

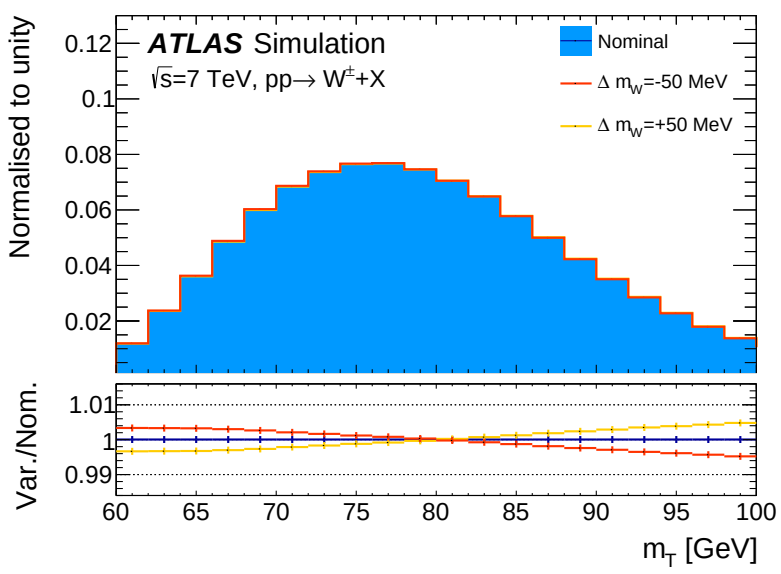
Introduction

- PDF's are an important or dominant uncertainty to precision measurements at the LHC of e.g. $\sin^2\theta_W$ and M_W
- There are many measurements, in particular of W and Z production at the LHC which can constrain PDFs in the relevant phase space
- In-situ constraints can also be used to constrain PDFs in the context of the measurements themselves
- This talk:
  - Overview of particularly interesting and relevant measurements
  - Overview of related phenomenological studies
  - Some personal thoughts
  - Some "unrelated" technical work in progress

W mass at LHC

- Current ATLAS measurement of m_W performed using 1D p_T^ℓ and M_T distributions (in bins of η^ℓ)
- Highest possible precision required on lepton momentum and hadronic recoil scale/resolution
- p_T^ℓ (and p_T^ν) distributions depend not only on m_W but also critically on p_T^W as well as polarization $\rightarrow$ strong dependence on QCD calculation and PDFs
- M_T distribution still sensitive to p_T^W and polarization due to finite detector acceptance





Eur. Phys. J. C 78 (2018) 110 (ATLAS)

W mass: PDF Uncertainties

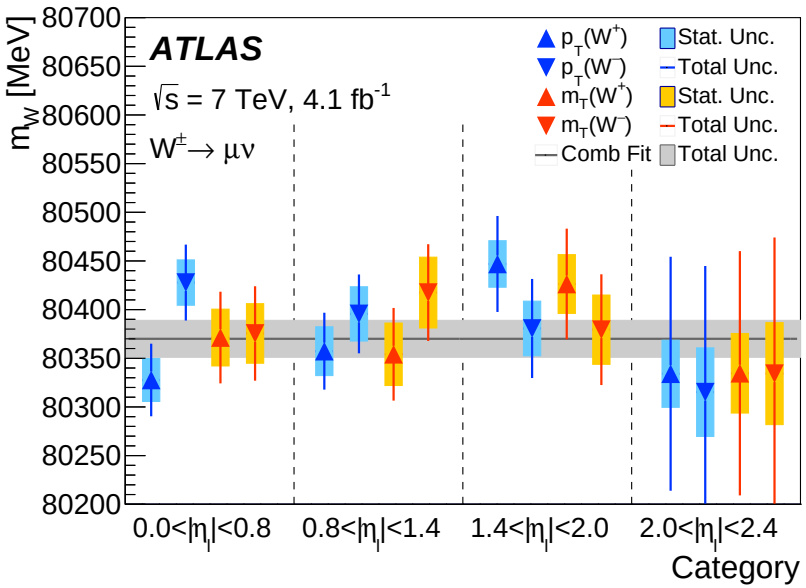
Eur. Phys. J. C 78 (2018) 110 (ATLAS)

$m_W = 80370 \pm 7(\text{stat.}) \pm 11(\text{exp. syst}) \pm 14(\text{mod. syst.})$ MeV

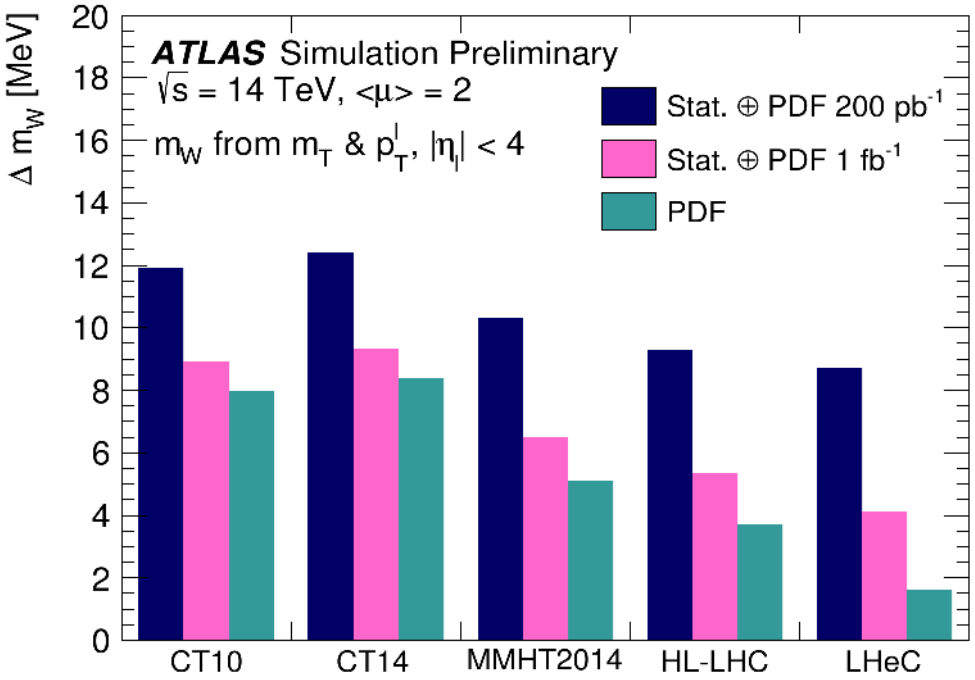
$m_W = 80370 \pm 7(\text{stat.}) \pm 11(\text{exp.}) \pm 8.3(\text{QCD}) \pm 5.5(\text{EWK}) \pm 9.2(\text{PDF})$ MeV

	PDF Uncertainty (MeV)
per $\lvert\eta\rvert$-charge cat.	20-34
per-charge	14-15
full combination	9.2

- PDFs determine the W rapidity spectrum and lepton decay angles through W polarization
- Well-defined correlations between phase space regions and processes which are already partly exploited in present measurement to reduce uncertainty
- Can be further exploited in the future



W mass: PDF Uncertainties

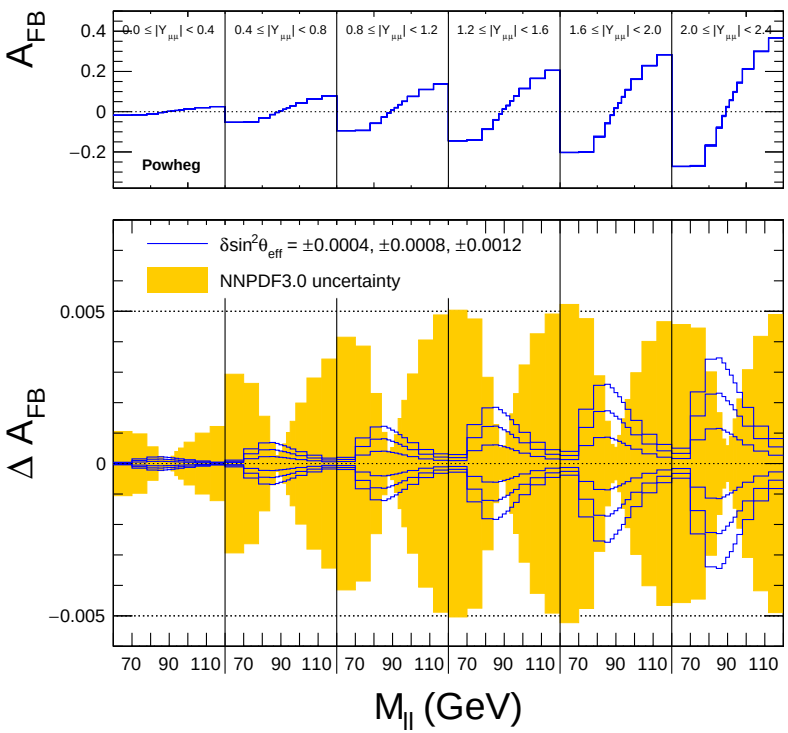


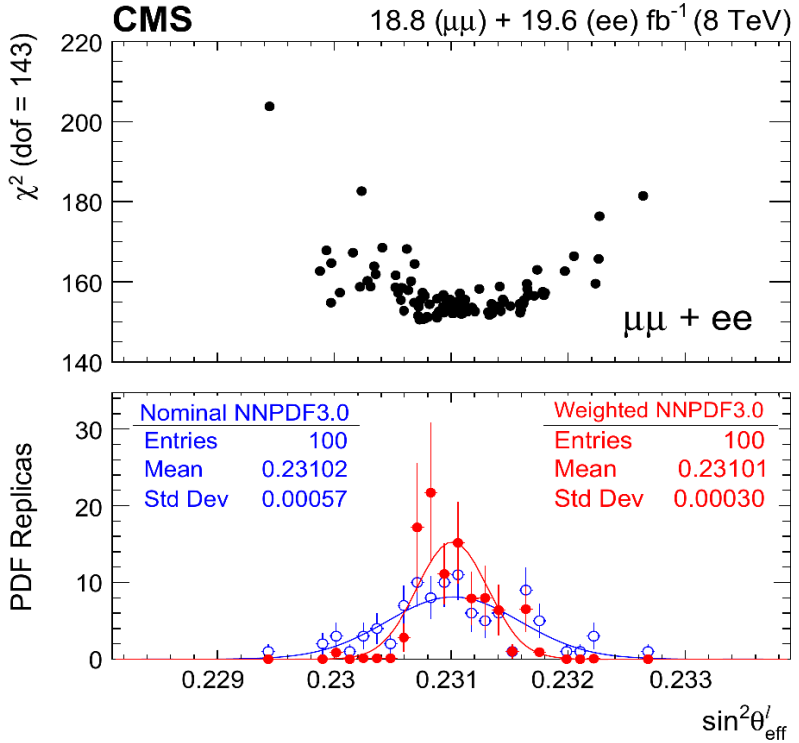
arXiv:1902.10229 (HL-LHC Yellow Report)

- Projected additional reduction in PDF uncertainties from additional measurements HL-LHC (or LHeC) could significantly reduce PDF uncertainty on m_W

In-situ PDF constraints: Weak Mixing Angle Case

- CMS weak mixing angle measurement exploits in-situ constraints to reduce PDF uncertainties with Bayesian reweighting of Monte Carlo replicas (equivalent to profiling of nuisance parameters associated with Hessian representation)



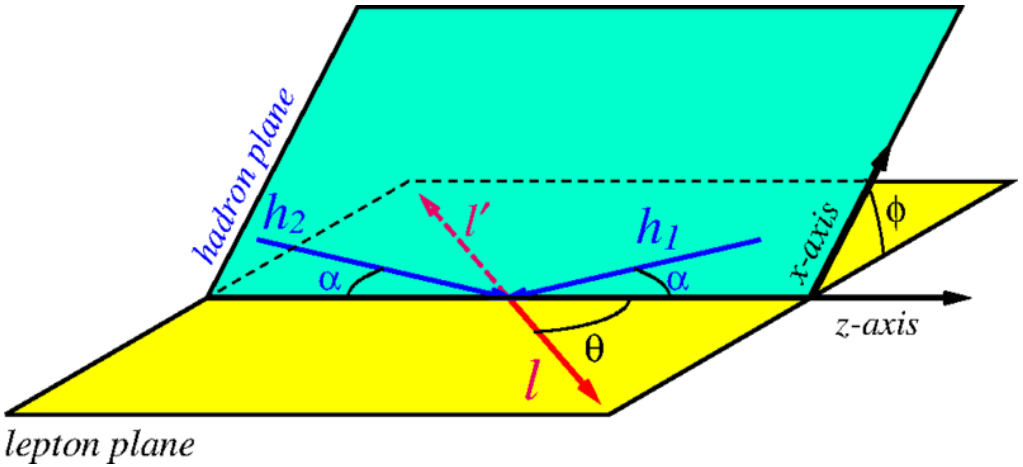


Eur. Phys. J. C (2018) 78: 701 (CMS)

Drell Yan Production at the LHC

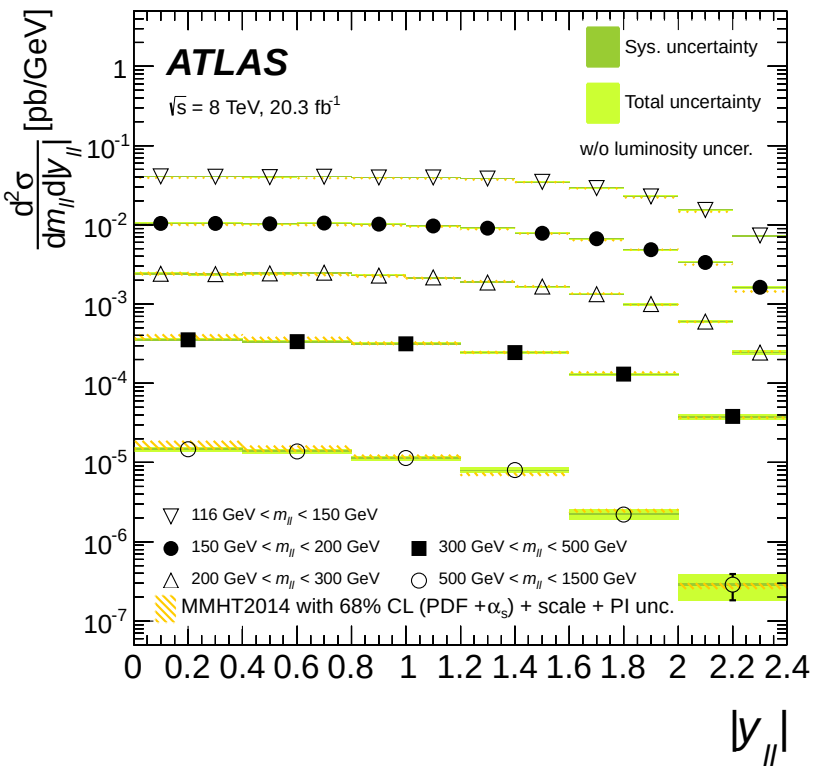
- Production and decay of $Z/\gamma^* \to \ell^+\ell^-$ or $W \to \ell\nu$ at the LHC, inclusive in additional hadronic activity, can be characterized by a 5-dimensional differential cross section

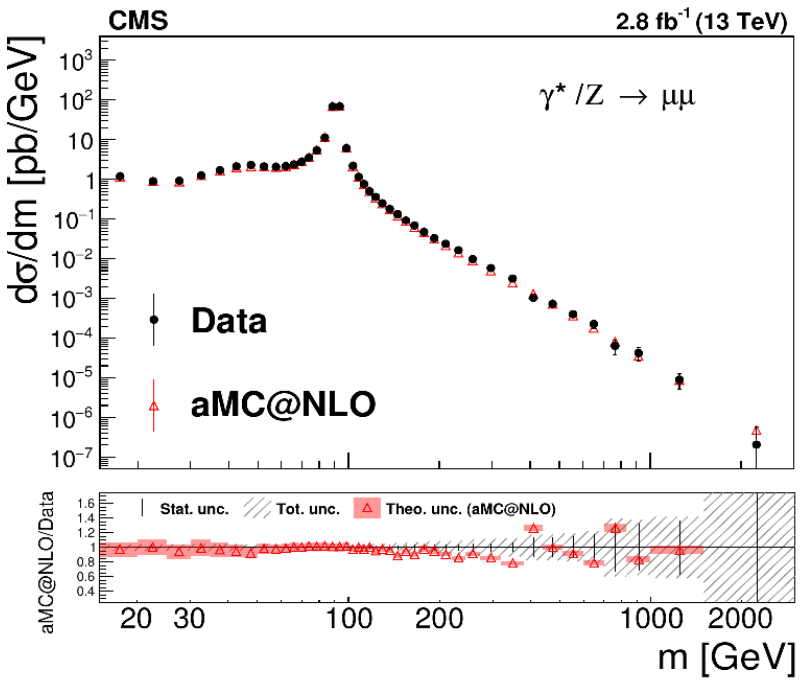
$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{T}}^{Z}\,\mathrm{d}y^{Z}\,\mathrm{d}m^{Z}\,\mathrm{d}\cos\theta\,\mathrm{d}\phi} = \frac{3}{16\pi}\frac{\mathrm{d}\sigma^{U+L}}{\mathrm{d}p_{\mathrm{T}}^{Z}\,\mathrm{d}y^{Z}\,\mathrm{d}m^{Z}} \times \Big\{(1+\cos^2\theta) + \frac{1}{2}\,A_0(1-3\cos^2\theta) + A_1\,\sin 2\theta\,\cos\phi + \frac{1}{2}\,A_2\,\sin^2\theta\,\cos 2\phi + A_3\,\sin\theta\,\cos\phi + A_4\,\cos\theta + A_5\,\sin^2\theta\,\sin 2\phi + A_6\,\sin 2\theta\,\sin\phi + A_7\,\sin\theta\,\sin\phi\Big\}. \tag{1.1}$$



- θ and ϕ are the decay angles of the lepton/neutrino in the rest-frame of the Z/γ^* or W, defined e.g. in the Collins-Soper frame

Drell Yan Measurements

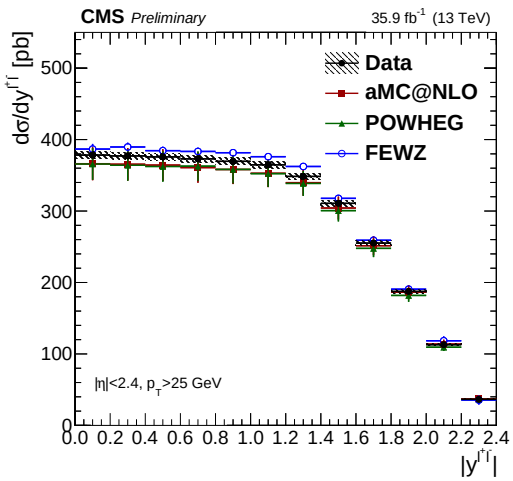




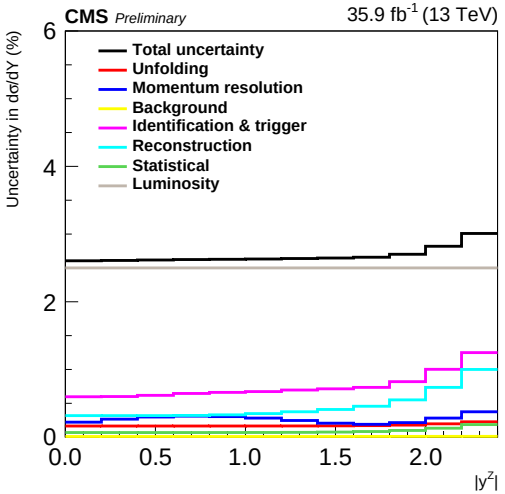
JHEP 08 (2016) 009 (ATLAS), arxiv:1812.10529 (CMS)

- Drell-Yan differential cross sections can be measured very precisely
- Sensitivity to PDFs
- Measurements integrated in $p_T^{\ell\ell}$ reduce the impact of higher-order QCD corrections (but acceptance cuts on lepton p_T and η are necessarily present for all fiducial cross sections...)

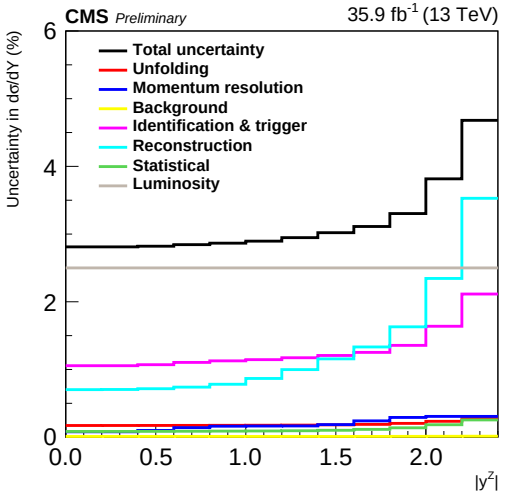
Drell Yan Measurements



(a) combined



(b) muons

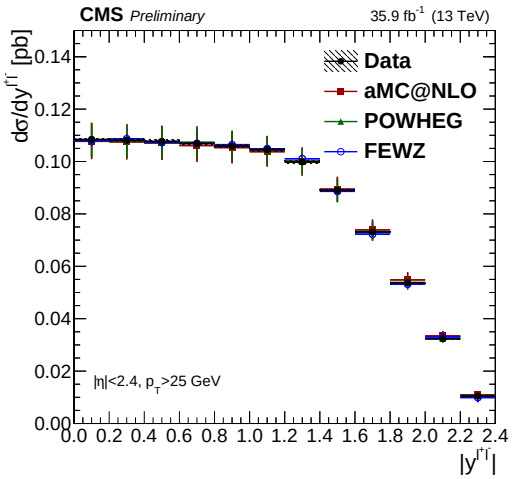


(c) electrons

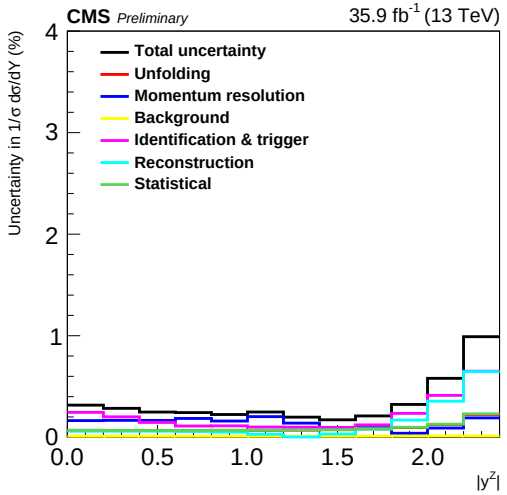
CMS-SMP-17-010

- Absolute cross sections near Z peak limited by luminosity uncertainty

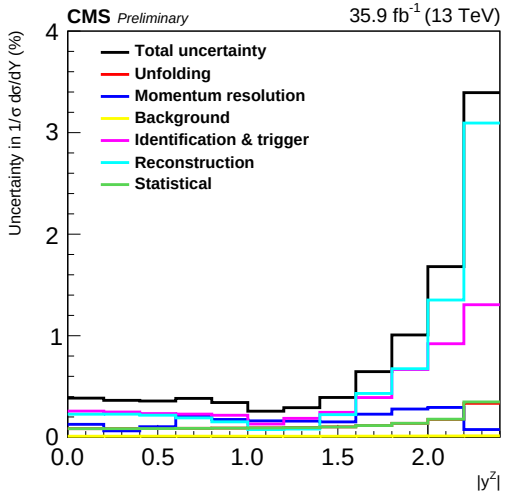
Drell Yan Measurements



(a) combined



(b) muons

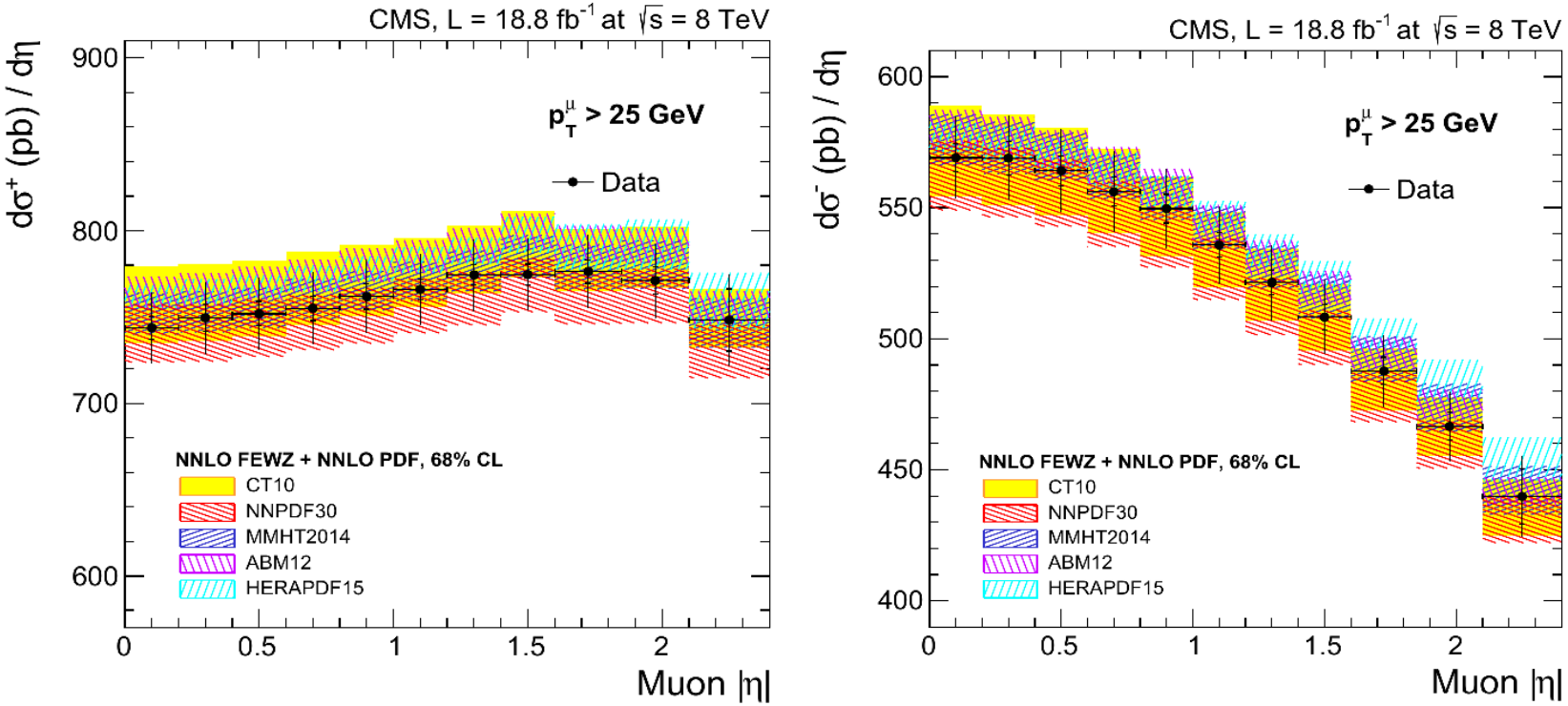


(c) electrons

CMS-SMP-17-010

- Precise normalized cross sections/shape also relevant, limited by lepton efficiencies $\rightarrow$ correlations across phase-space crucial

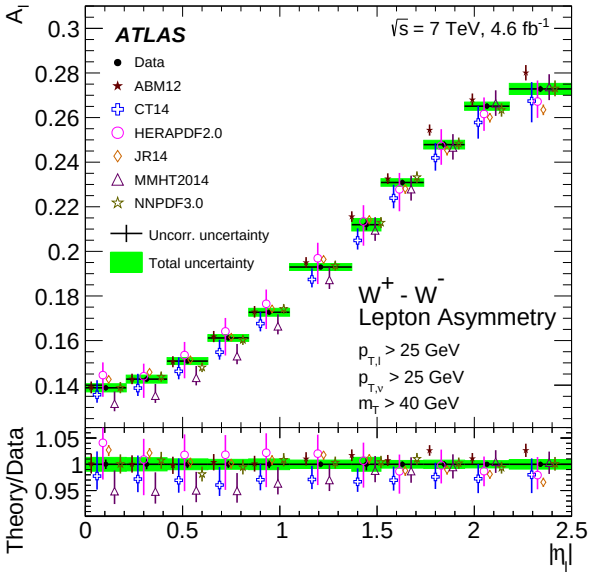
W Differential Cross Sections/charge asymmetry

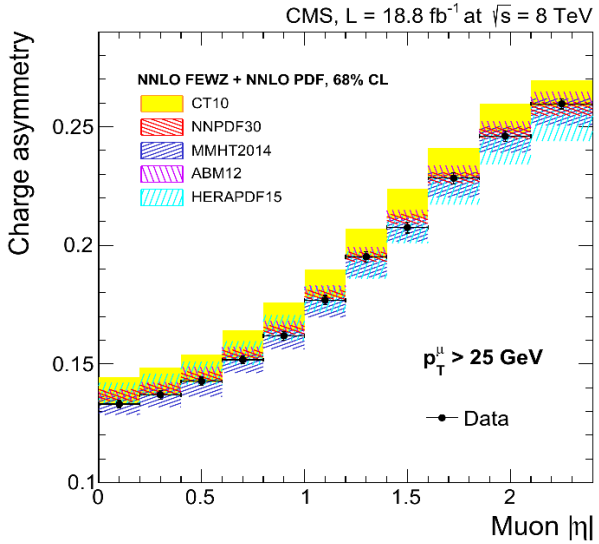


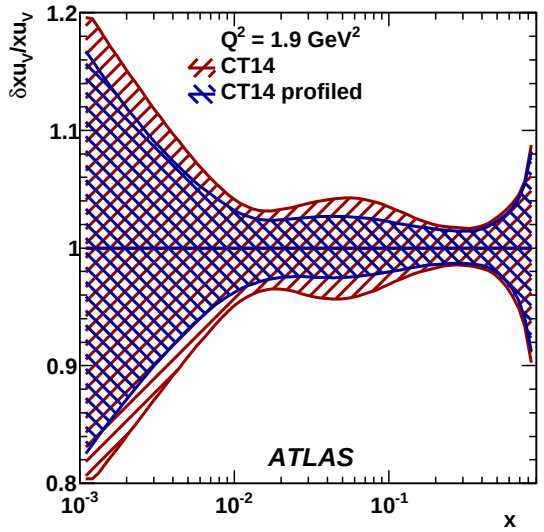
Eur. Phys. J. C 76 (2016) 469 (CMS)

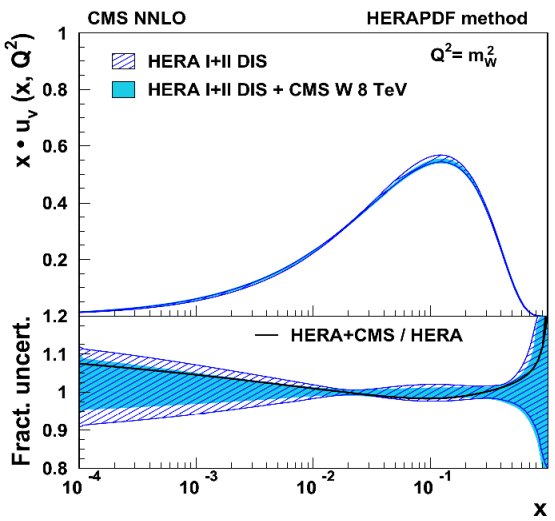
- W differential cross sections and charge asymmetries provide constraints on the valence quark pdfs
- Main systematic uncertainties: Lepton efficiencies, multi-jet background estimate

W Differential Cross Sections/charge asymmetry





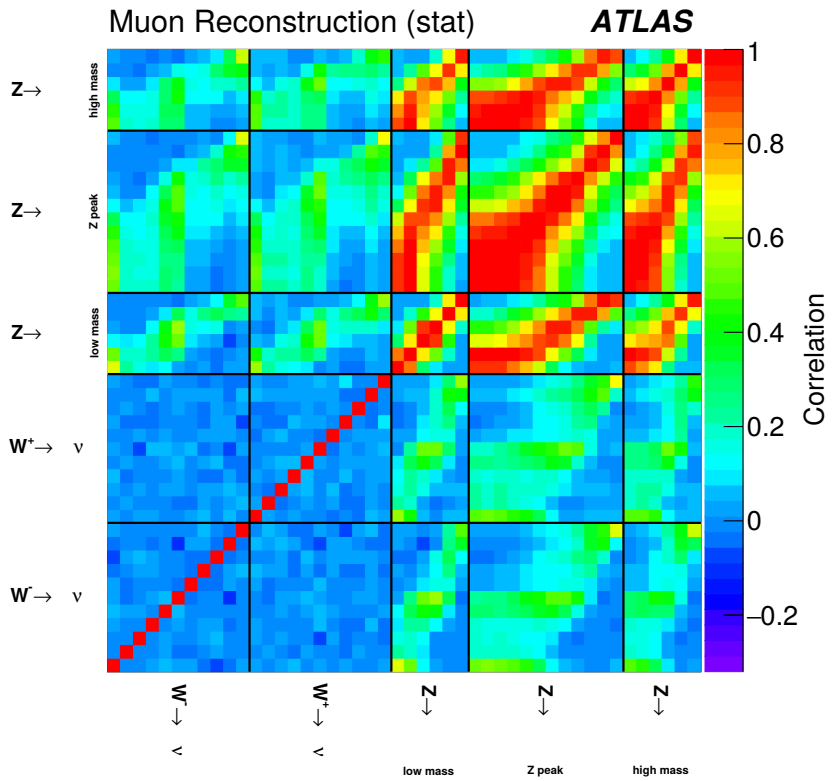




- W differential cross sections and charge asymmetries provide constraints on the valence quark pdfs
- Main systematic uncertainties: Lepton efficiencies, multi-jet background estimate

Eur. Phys. J. C 77 (2017) 367 (ATLAS), Eur. Phys. J. C 76 (2016) 469 (CMS)

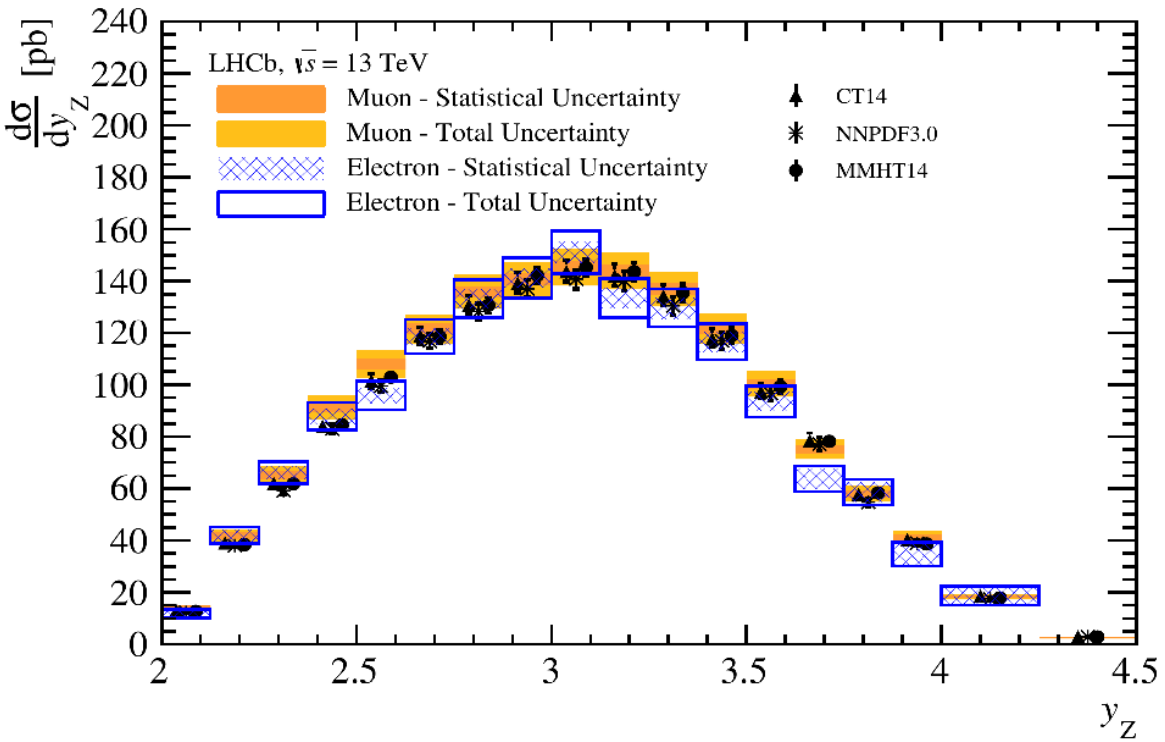
Correlations of Lepton Efficiency Uncertainties



Eur. Phys. J. C 77 (2017) 367 (ATLAS)

- Example shown here for **statistical** component of uncertainty on muon reconstruction efficiency for ATLAS W/Z measurement
- Underlying uncertainty is uncorrelated in bins of single muon p_T and η in which efficiencies were measured with tag and probe, leading to non-trivial correlations in particular for $Z/\gamma^* \to \mu\mu$ measurements

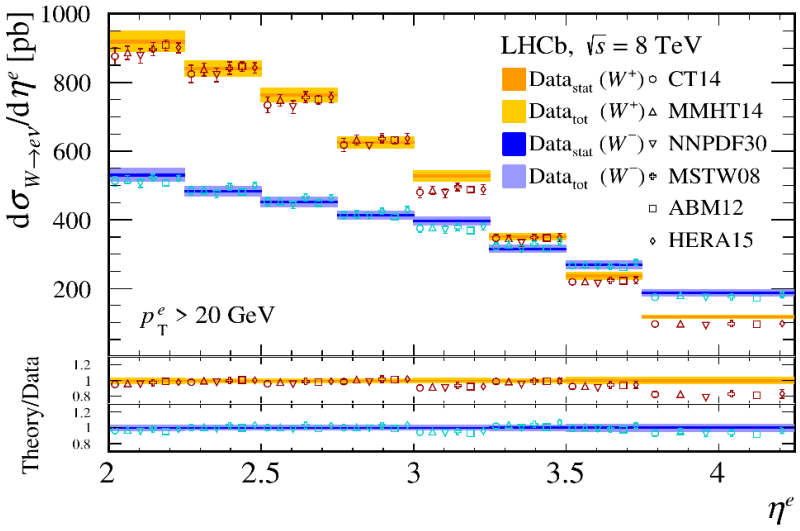
Forward W/Z Production at LHC (LHCb)

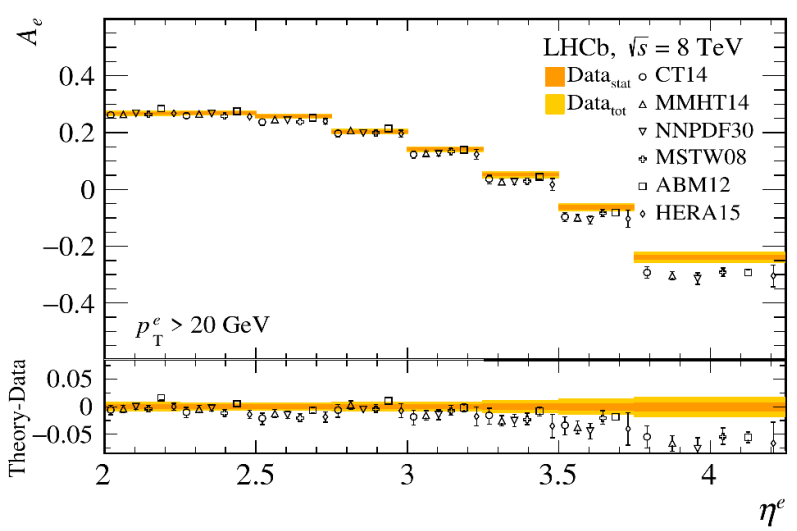


JHEP 09 (2016) 136

- LHCb has complementary coverage for charged leptons, starting from $\eta > 2$ up to $\eta < 4.25(4.5)$
- Provides complementary information on PDFs

Forward W/Z Production at LHC (LHCb)

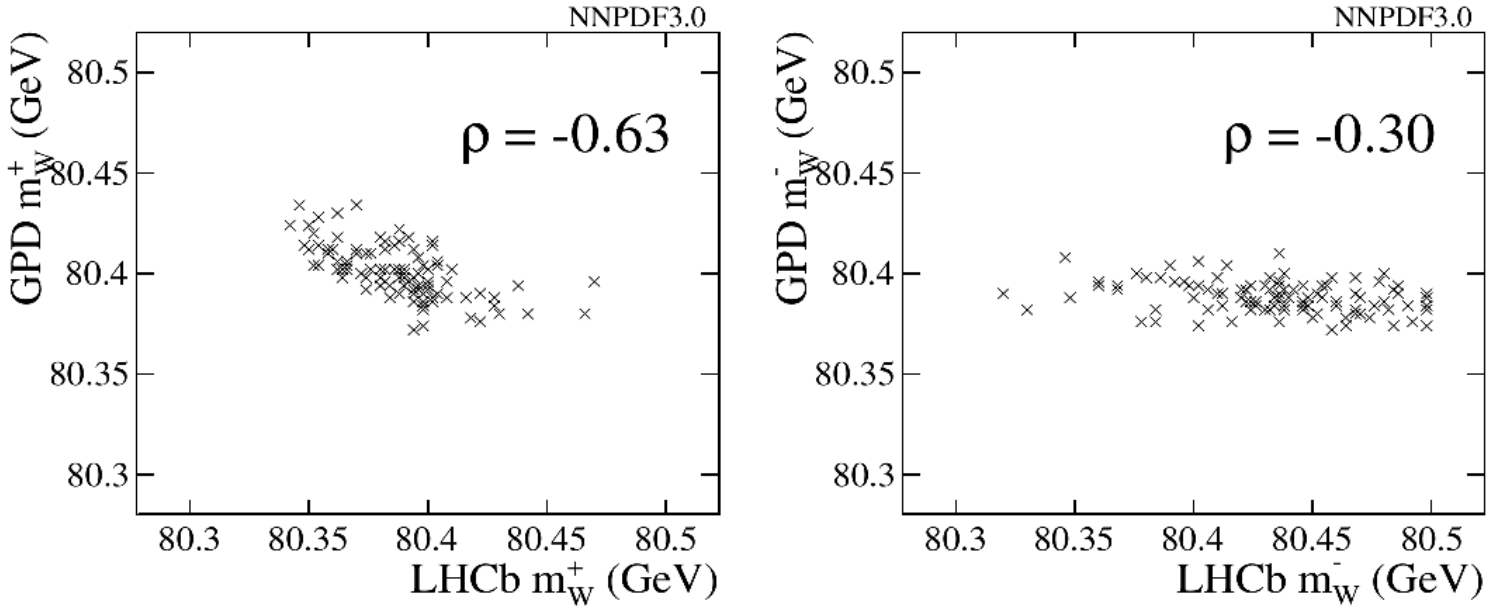




JHEP 10 (2016) 030 (LHCb)

- LHCb has complementary coverage for charged leptons, starting from $\eta > 2$ up to $\eta < 4.25(4.5)$
- Provides complementary information on PDFs

Forward W/Z Production at LHC (LHCb)

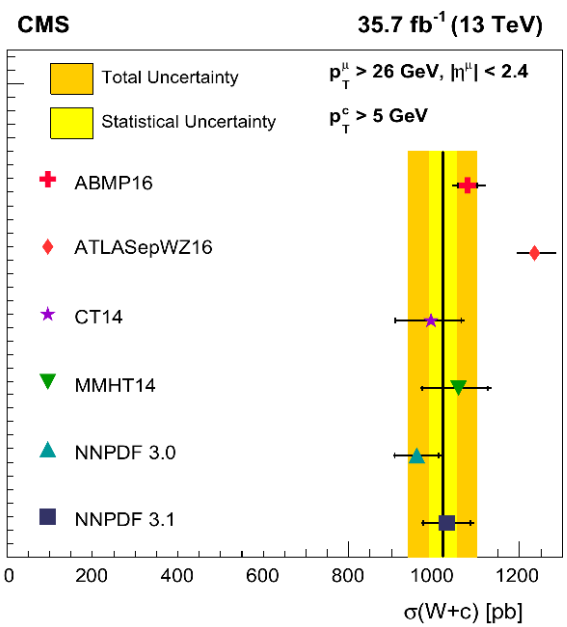


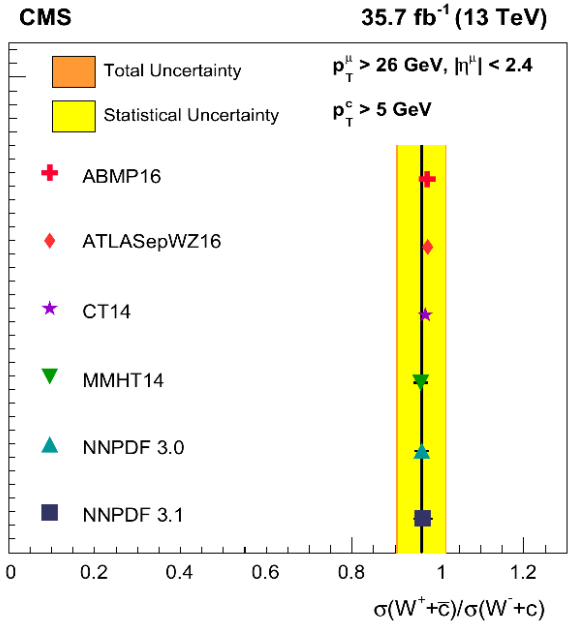
Scenario	Experiments	δm_W (MeV) Tot	Exp	PDF
Default	2×GPD + LHCb	9.0	4.7	7.7
Default	1×GPD + LHCb	10.1	6.5	7.7
Default	2×GPD	12.0	5.8	10.5

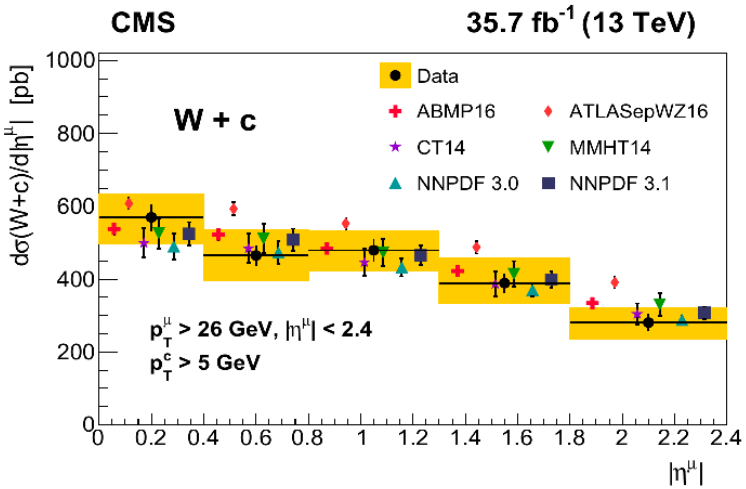
arXiv:1508.06954 G. Bozzi, L. Citelli, M. Vesterinen, A. Vicini

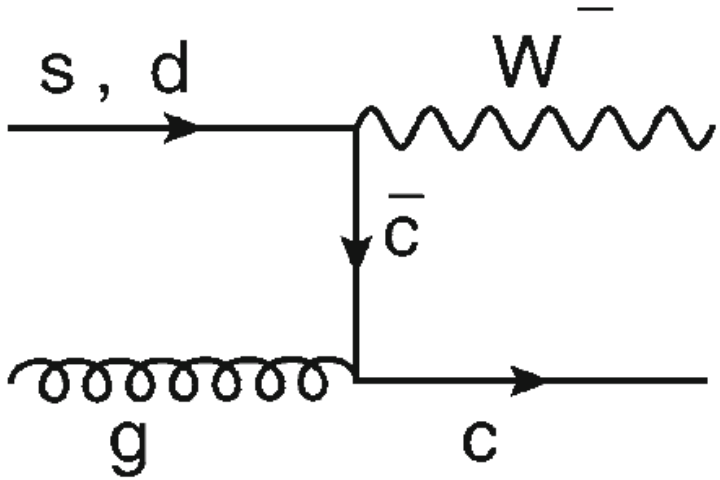
- Complementarity of forward phase space for PDFs in W mass context demonstrated in simplified phenomenological studies

W+charm





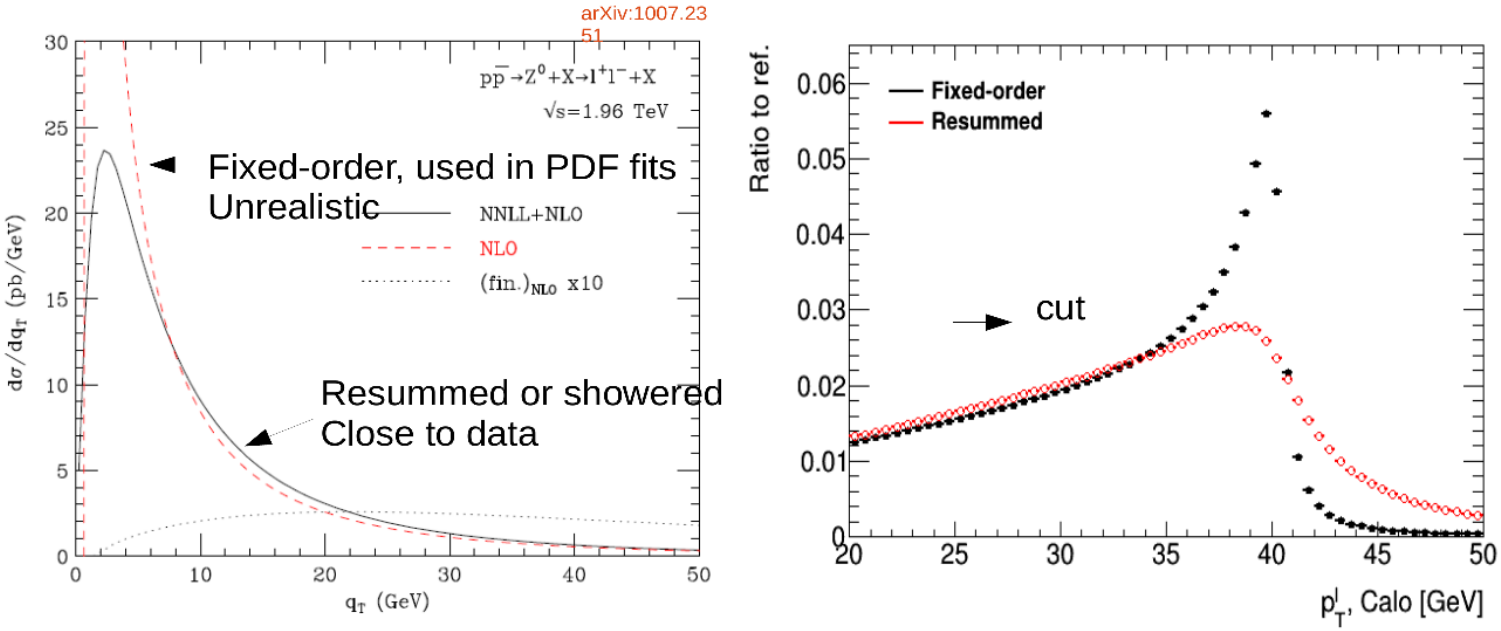




Eur. Phys. J. C (2019) 79:269 (CMS)

Impact of Resummation in predictions for PDF Fits

Motivation



Resummation effects affect the pT distributions, hence the acceptance of fiducial cuts

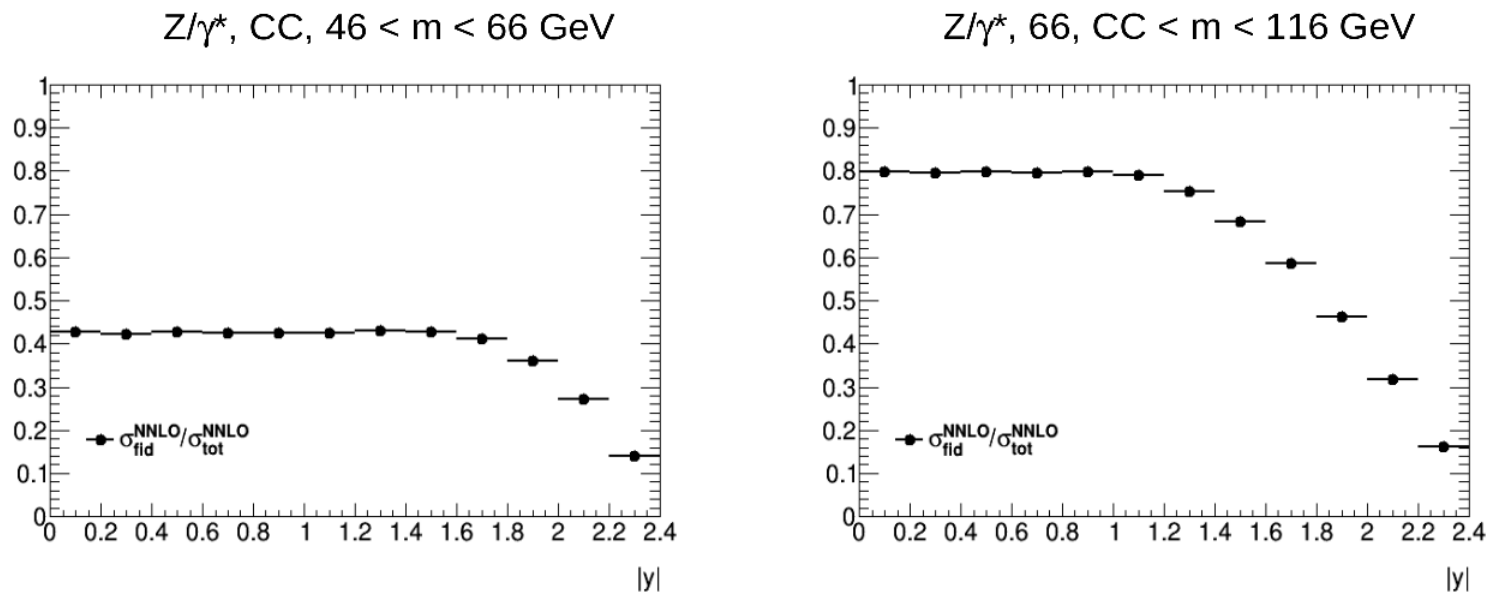
For same total cross section, fixed-order and resummed ***fiducial*** cross sections differ.

This leads to a small inconsistency when interpreting fiducial cross section measurements in terms of PDFs, which typically use fixed-order predictions

M. Boonekamp `https://indico.cern.ch/event/801961/contributions/3368455/attachments/1824716/2985820/psKfactors_050419.pdf`

Impact of Resummation in predictions for PDF Fits

Fixed-order acceptance



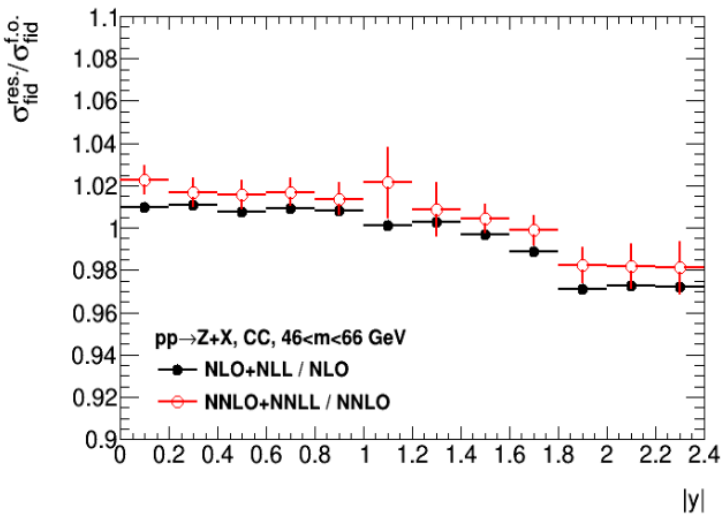
$$p_T^\ell > 20 \text{ GeV}, \ |\eta^\ell| < 2.5$$

M. Boonekamp https://indico.cern.ch/event/801961/contributions/3368455/attachments/1824716/2985820/psKfactors_050419.pdf

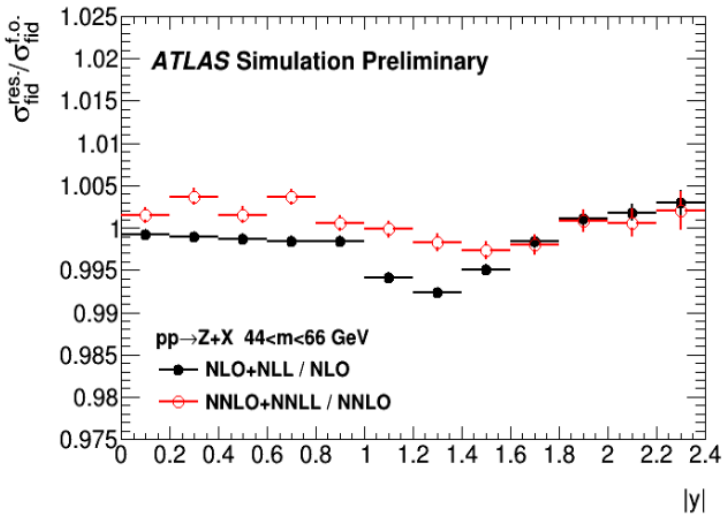
Impact of Resummation in predictions for PDF Fits

Effect of pT resummation – Z

CC, 46 < M < 66 GeV



CC, 66 < M < 116 GeV



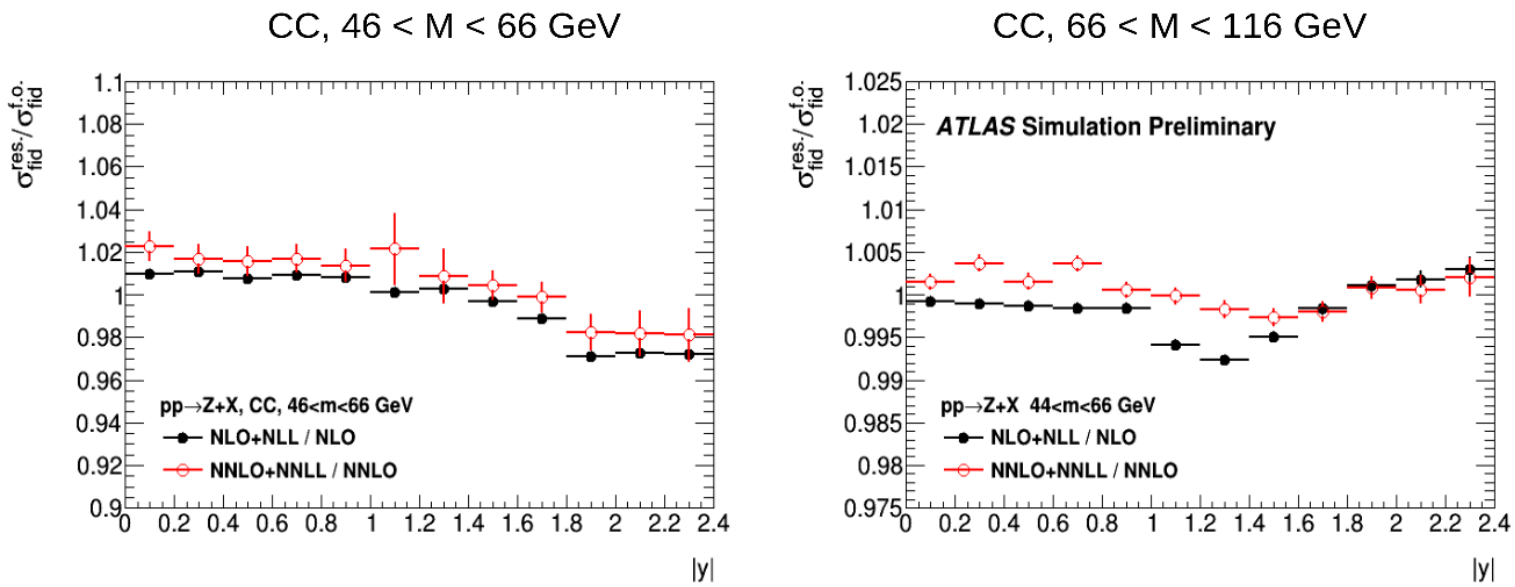
3-4% drop towards high eta

NLL : ~1% dip near |eta|=1.4
Reduced to .5% at NNLL

$$p_T^\ell > 20 \text{ GeV}, \ |\eta^\ell| < 2.5$$

M. Boonekamp https://indico.cern.ch/event/801961/contributions/3368455/attachments/1824716/2985820/psKfactors_050419.pdf

Impact of Resummation in predictions for PDF Fits

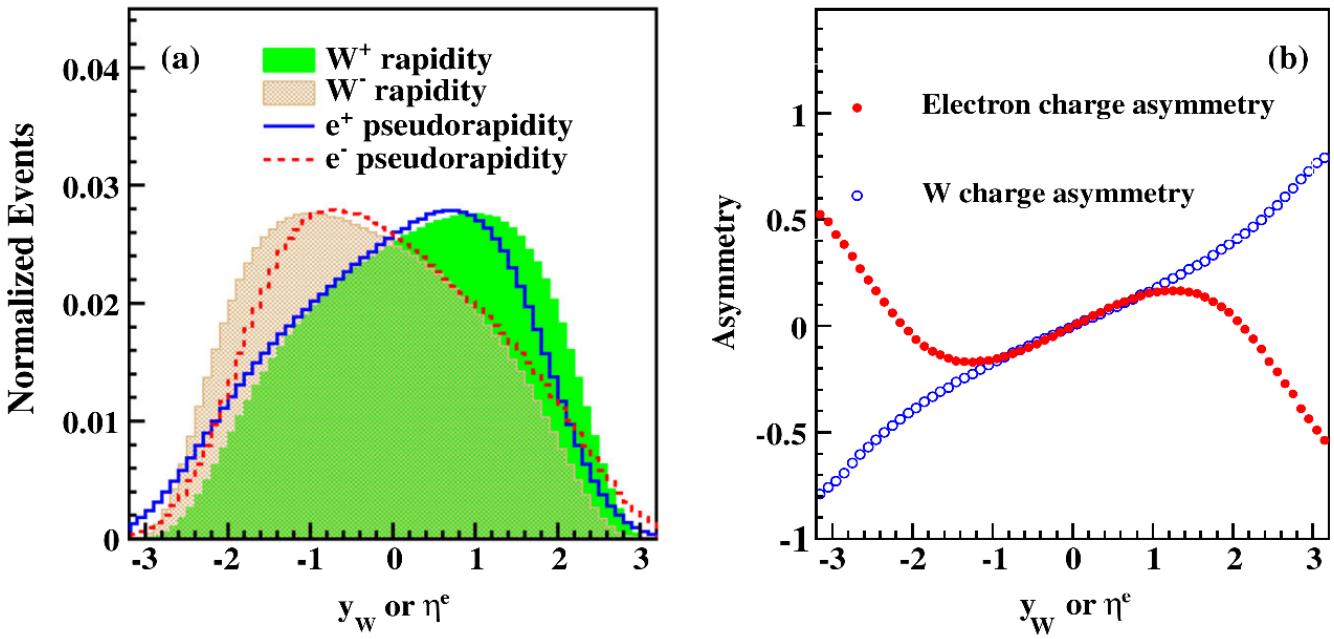


$$p_T^\ell > 20 \text{ GeV}, \ |\eta^\ell| < 2.5$$

M. Boonekamp https://indico.cern.ch/event/801961/contributions/3368455/attachments/1824716/2985820/psKfactors_050419.pdf

- Resummation corrections are relevant for predictions of W and Z differential fiducial cross sections (mainly due to lepton pT cuts in fiducial phase space definition)
- Effect may be small in absolute terms, but is relevant compared to the precision of the experimental measurements, in particular for normalized cross sections

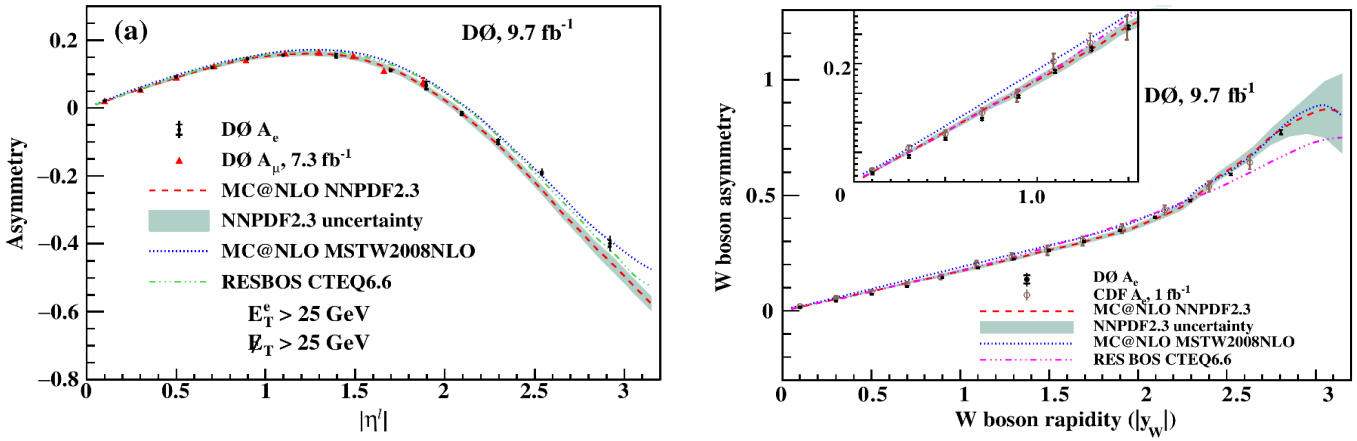
W vs lepton charge asymmetry at the Tevatron



Phys. Rev. D 91, 032007 (2015) (D0)

- Lepton charge asymmetry vs η is a convolution of PDF effect with V-A structure of W decay
- W charge asymmetry as a function of W rapidity more directly probes the PDFs (but less directly accessible experimentally)
- Tevatron experiments historically provided both measurements
- n.b. at Tevatron, asymmetries are sensitive to sign of η or y due to $p\bar{p}$ collisions $\rightarrow$ final results are "CP" folded $A(-\eta/y) \rightarrow -A(\eta/y)$

W vs lepton charge asymmetry at the Tevatron

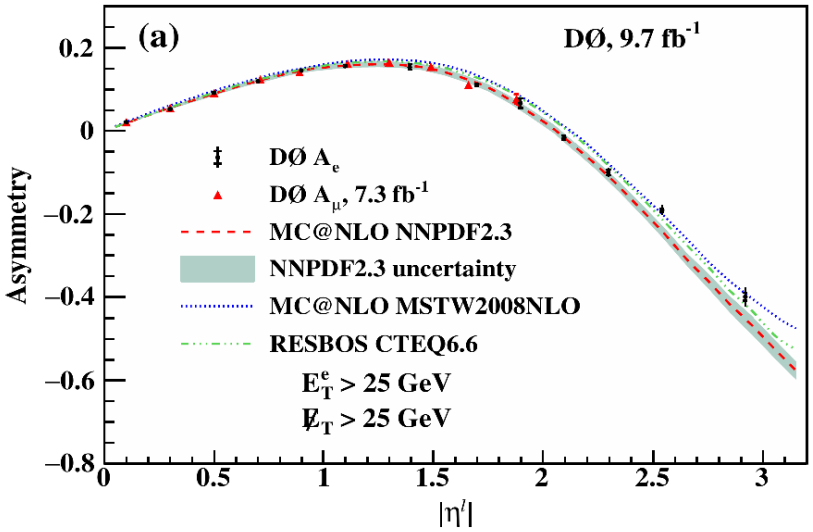


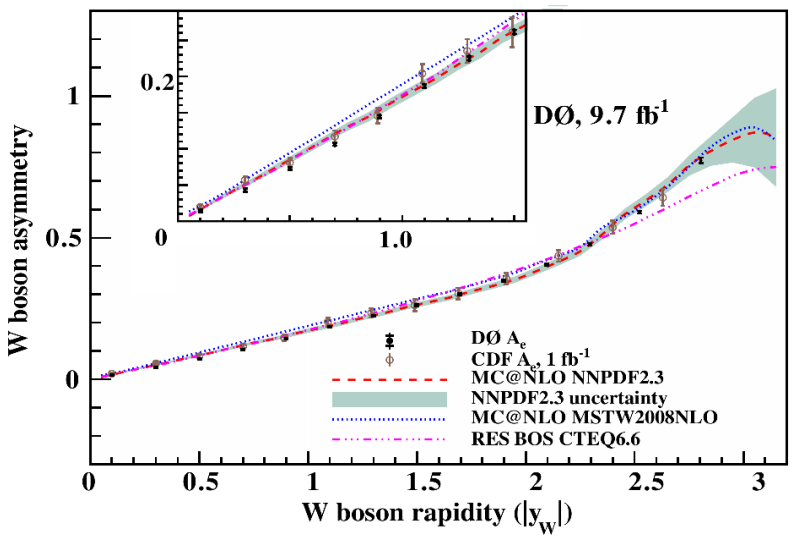
(a) Lepton Charge Asymmetry (b) W Charge Asymmetry

Phys. Rev. D 91, 032007 (2015) (D0), Phys. Rev. Lett. 112, 151803 (2014) (D0)

- Unfolding to W rapidity using missing transverse momentum and M_W constraint
- Resolving resulting twofold ambiguity requires assumption about relative fractions of incoming quark vs antiquark in proton beam (plus smaller effect from gluon-initiated production) $\rightarrow$ 10% effect in total, with non-negligible uncertainty from PDF's $\rightarrow$ some circularity in using data in this form for PDF determination

W vs lepton charge asymmetry at the Tevatron



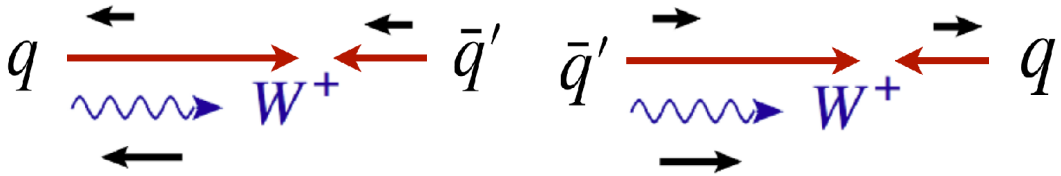


(a) Lepton Charge Asymmetry (b) W Charge Asymmetry

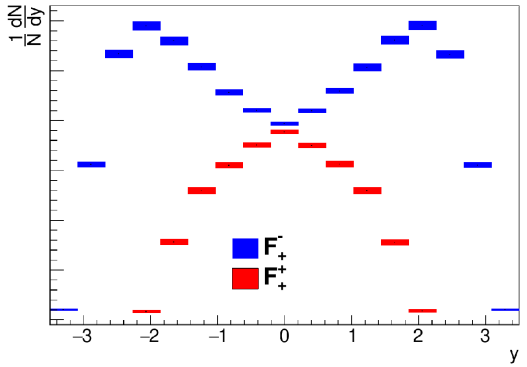
Phys. Rev. D 91, 032007 (2015) (D0), Phys. Rev. Lett. 112, 151803 (2014) (D0)

- On the other hand, lepton charge asymmetry vs η^ℓ does not contain all available information, since information on p_T^ℓ, p_T^ν and $\Delta\phi_{\ell,\nu}$ are lost

W Helicity/Rapidity at LHC



(a) left-handed W^+ (b) right-handed W^+

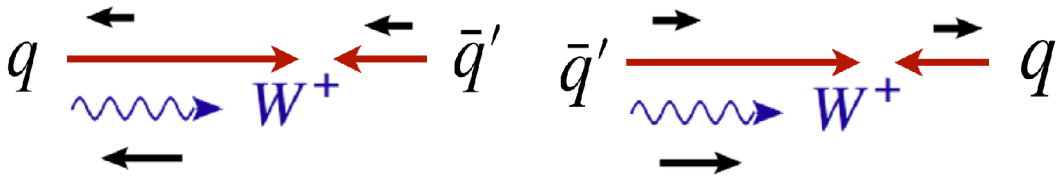


(c) W^+ Rapidity

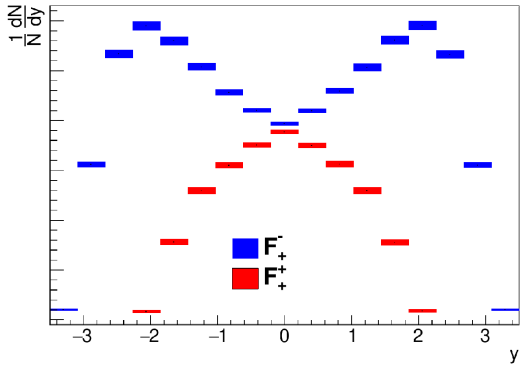
- At tree level:
  - All W production at LHC is $q\bar{q}$ induced
  - Direction of the W relative to the incoming quark determines the helicity
  - Only two helicity amplitudes/polarization states
  - W has zero transverse momentum
  - **Full information on valence quark PDF's in the relevant x range contained in $d\sigma/dy$ broken down into the two helicity states**

JHEP12(2017)130 E. Manca, O. Cerri, N. Foppiani, G. Rolandi

W Helicity/Rapidity at LHC



(a) left-handed W^+ (b) right-handed W^+

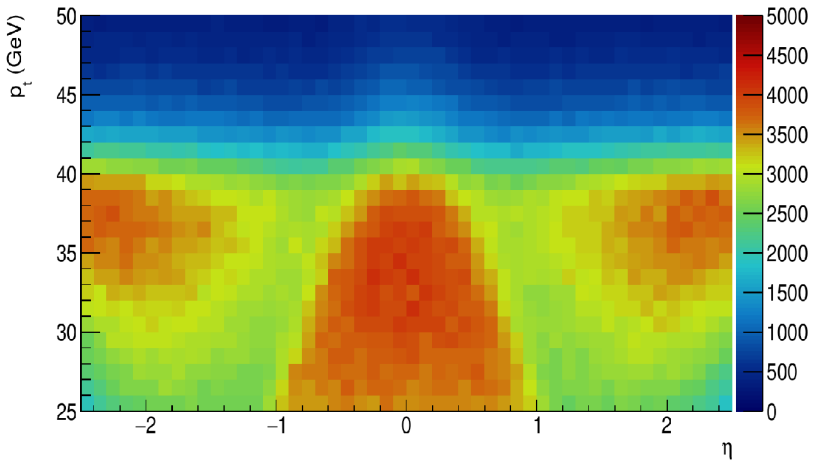


(c) W^+ Rapidity

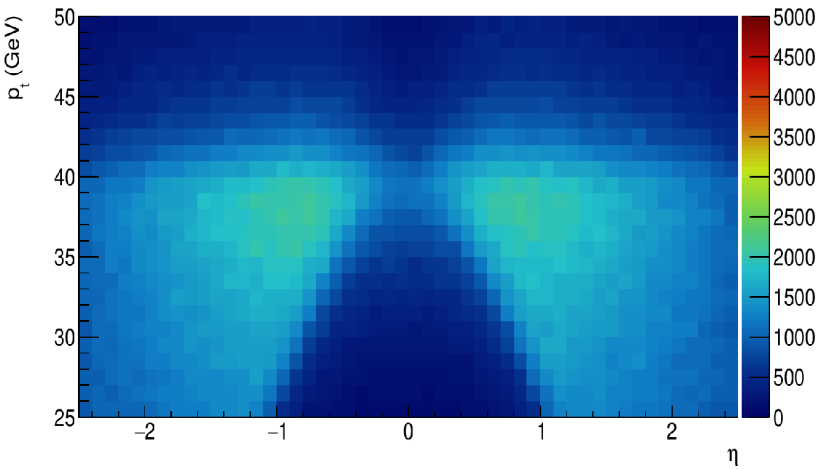
- Direction of incoming quark depends even more on PDF's in pp vs $p\bar{p}$ collisions
- gluon-induced contribution from higher order effects larger and more uncertain (also due to higher E_{cm} compared to Tevatron)

JHEP12(2017)130 E. Manca, O. Cerri, N. Foppiani, G. Rolandi

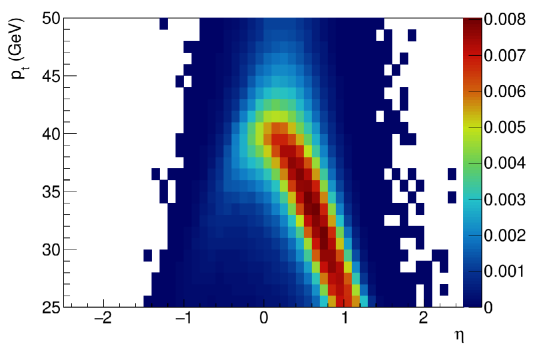
W Helicity/Rapidity at LHC

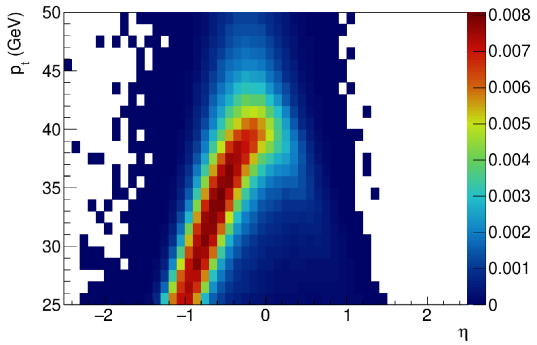


(a) left-handed W^+



(b) right-handed W^+





- 2D distribution of charged lepton p_T and η can discriminate between helicity states as well as rapidity of the W

JHEP12(2017)130 E. Manca, O. Cerri, N. Foppiani, G. Rolandi

W Helicity/Rapidity at LHC

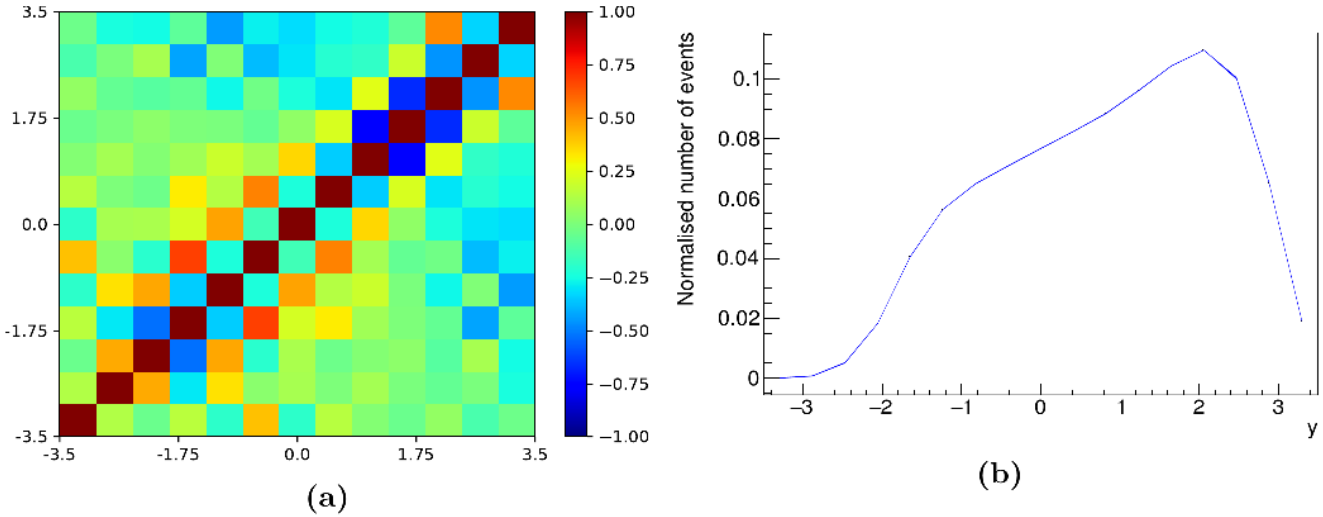


Figure 4. a) Correlation matrix of the fit b) Rapidity distribution for W^+ with spin pointing to the negative z axis as measured in the fit.

(a) Gen-Level Fit

- Left and right polarization components can be extracted simultaneously as a function of W rapidity, using only charged lepton kinematics
- Avoids dependence on less precisely measured missing transverse momentum (at the cost of some statistical dilution)
- Avoids circular dependence on PDFs since quark vs anti-quark fraction for each rapidity is **measured**

JHEP12(2017)130 E. Manca, O. Cerri, N. Foppiani, G. Rolandi

W Helicity/Rapidity at LHC

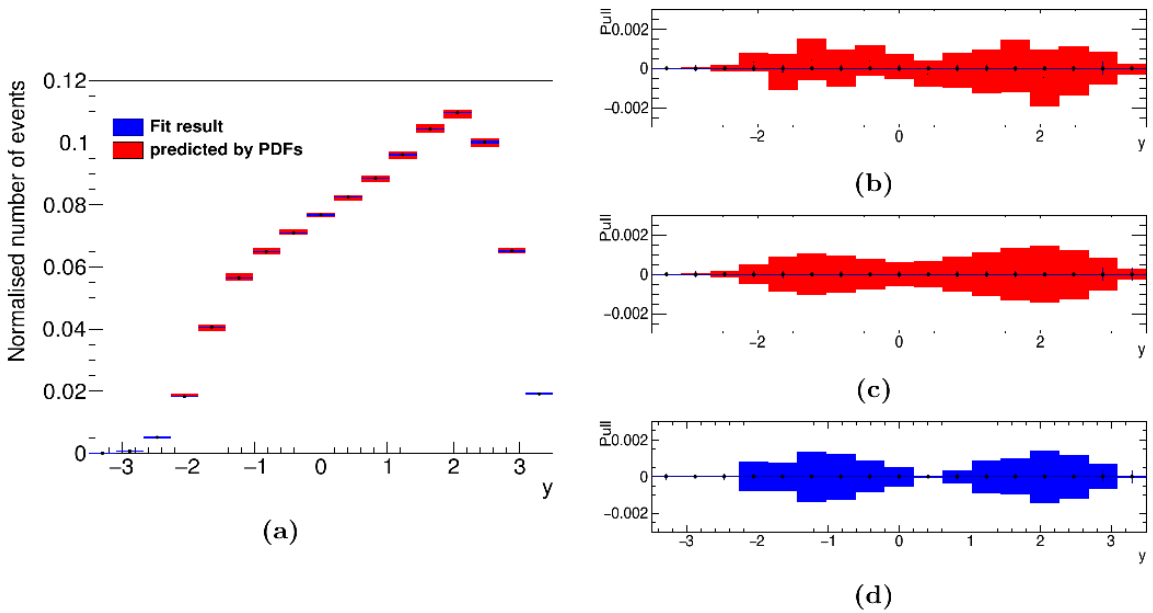


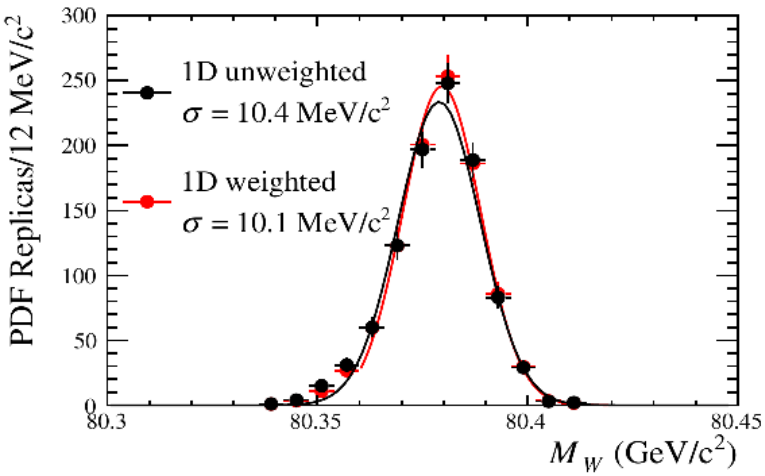
Figure 5. a) The result of the fit is compared to the PDFs prediction. b) Difference between fit and PDFs prediction. c) Difference between fit and PDFs prediction shifted at the same central value of the fit. d) Shape variation when modifying the p_t^W spectrum.

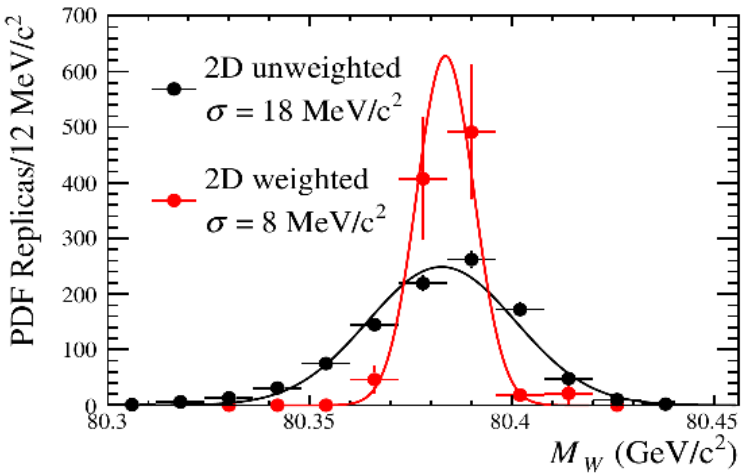
- Left and right polarization components can be extracted simultaneously as a function of W rapidity, using only charged lepton kinematics
- Resulting measurement would have sensitivity to PDFs
- Important systematic uncertainty from W p_T distribution (which is implicitly integrated over)

JHEP12(2017)130 E. Manca, O. Cerri, N. Foppiani, G. Rolandi

W mass from charged lepton kinematics

- Phenomenological study in LHCb-like acceptance ($30 < p_T^\ell < 50$GeV,$2 < \eta^\ell < 4.5$) comparing 1D fit of p_T^ℓ distribution to 2D fit of (p_T^ℓ, η^ℓ) distribution, with and without posterior weighting ($\sim$ equivalent to profiling) of PDF's
- 2D fit is closely related to previous study, differs from ATLAS fit to p_T^ℓ in categories of η^ℓ because the latter effectively leaves the normalization in each η^ℓ bin freely floating (and uses coarser η^ℓ binning)





Eur. Phys. J. C (2019) 79: 497, S. Farry, O. Lupton, M. Pili, M. Vesterinen

Towards experimental measurement

- Phenomenological studies motivate a $2D$ fit to charged lepton p_T,$|\eta|$ distribution to extract W rapidity distribution decomposed into left, right (longitudinal) polarization states
- Equivalent to measuring unpolarized cross section, and A_0, A_4 angular coefficients integrated over mass and p_T^W
- (Can and should also measure unfolded 2D p_T^ℓ, $|\eta^\ell|$ distribution)
- Such a fit can also be used to extract M_W with strong in-situ PDF constraints
- Experimentally challenging, must control all the usual ingredients to maximum precision:
  - Lepton energy/momentum scale
  - Lepton efficiencies
  - QCD background estimate
  - Theoretical modeling of boson production and decay (+QED effects)
- Must correctly model normalization as well as shape for predictions and experimental/theoretical uncertainties, with proper correlations
- At least a minimal cut on missing transverse energy and/or transverse mass might be necessary/desirable to suppress QCD background

In-situ PDF Constraints

- PDF Bayesian reweighting and/or profiling strongly desired to reduce uncertainties
- This implies **detector level** extraction of PDF constraints, using e.g. Monte Carlo with full shower/hadronization/MPI + detector simulation for prediction
- Could still be done fully consistently at NNLO in QCD if using Powheg-MINLO-NNLOPS or GENEVA for Monte Carlo
- In both cases predictions would include resummation

In-situ PDF Constraints

- Ideally detailed information on in-situ constraints should be provided in a useful form:
  - Reweighting of MC replicas: Post-fit weight for each replica (n_{replica} floats)
  - Profiling of Hessian uncertainties: Post-fit value of nuisance parameters associated with eigenvectors, plus postfit covariance matrix ($n_{\text{eigenvector}} + (n^2_{\text{eigenvector}} + n_{\text{eigenvector}})/2$ floats)
- In principle the two are equivalent, but converting from a small number of replicas to a covariance matrix of this nature likely suffers from large numerical inaccuracies, so the latter may be preferred where gaussian uncertainties is an acceptable approximation...
- Interesting possibilities for comparison between detector level constraints, constraints from several different variants of unfolded cross sections

Avoiding Double Counting of PDF Constraints

- Avoiding double counting of PDF constraints is critical for correct statistical interpretation $\rightarrow$ W (and/or Z) inclusive/differential cross sections and asymmetries from the same dataset must be excluded from input PDFs (or else careful factorization of observables)
- In particular when exploiting normalization information as well as shape, any fit to lepton p_T, η distribution in W production is $\sim$fully overlapping with W differential cross section and/or charge asymmetry data
- To the extent that theoretical modeling of W (and/or Z) production is crucial, may also be desirable to exclude closely related datasets (e.g. W and Z data from other LHC experiments and/or CM energies)
- Thanks to NNPDF collaboration for providing at request of CMS additional NNPDF3.1 sets with some/all W/Z data removed

Towards experimental measurement

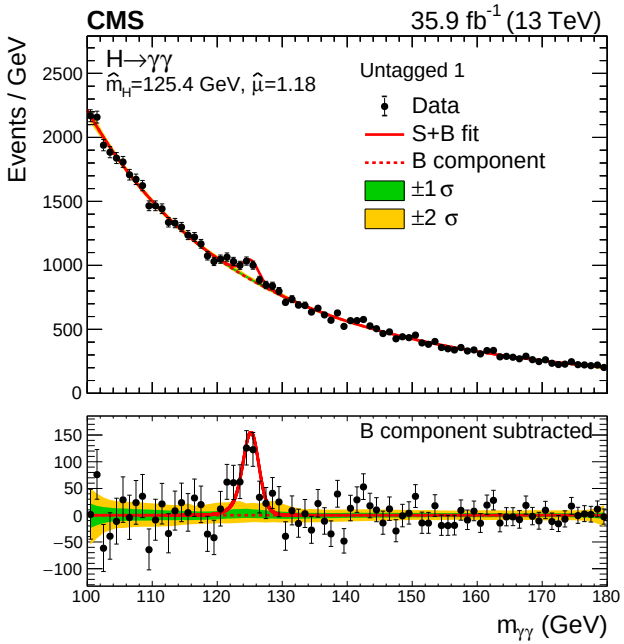
- PDF Bayesian reweighting and/or profiling strongly desired to reduce uncertainties
- One possibility which allows consistent treatment of PDF uncertainties with other experimental and theoretical systematic uncertainties: (binned) profile likelihood fit
- M_W fit has one parameter of interest, but helicity-rapidity extraction, or finely binned differential cross section measurement could have tens or hundreds
- Detailed model for all uncertainties could imply a large number of nuisance parameters $\rightarrow$, likelihood fit could become technically complicated/challenging...

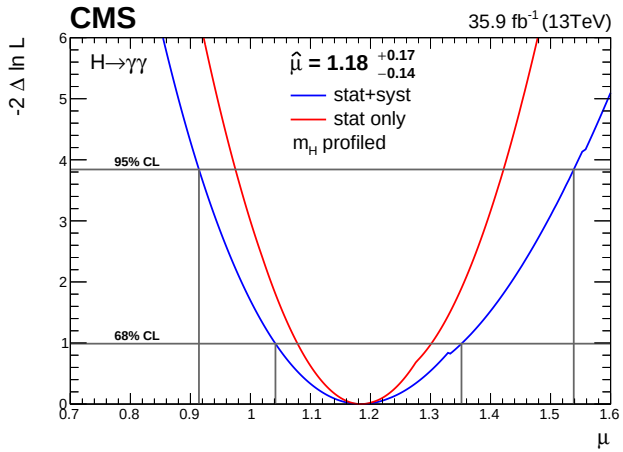
Introduction: Maximum Likelihood Fits

- Common framework for statistical interpretation of HEP data: **Maximum Likelihood Fits**
  - Maximize the joint probability of the data $\vec{x}$ given some parameters of the model $\vec{\theta}$ which may include both **parameters of interest** (POIs) such as production cross sections, particle masses, etc, as well was **nuisance parameters**, e.g. reconstruction efficiency or energy scale allowed to vary within some prior constraint
- Two variants:
  - **Unbinned Maximum Likelihood Fit:** Typically a small number of observables (often 1, rarely more than 3) with a large number of events, evaluate the continuous probability density for each data event: $-\ln L = -\sum_{events} \ln p(\vec{x}_{ievent}|\vec{\theta})$
  - **Binned Maximum Likelihood Fit:** Likelihood is evaluated using bin counts in a histogram: $-\ln L = -\sum_{bins} \ln p(N_{ibin}|\vec{\theta})$

Introduction: Maximum Likelihood Fits

- **Unbinned Maximum Likelihood Fit:** Typically a small number of observables (often 1, rarely more than 3) with a large number of events, evaluate the continuous probability density for each data event: $-\ln L = -\sum_{events} \ln p(\vec{x}_{ievent}|\vec{\theta}) \rightarrow$ **small feature space, many examples**
- **Binned Maximum Likelihood Fit:** Likelihood is evaluated using bin counts in a histogram: $-\ln L = -\sum_{bins} \ln p(N_{ibin}|\vec{\theta}) \rightarrow$ **Moderately sized feature space, 1 example**



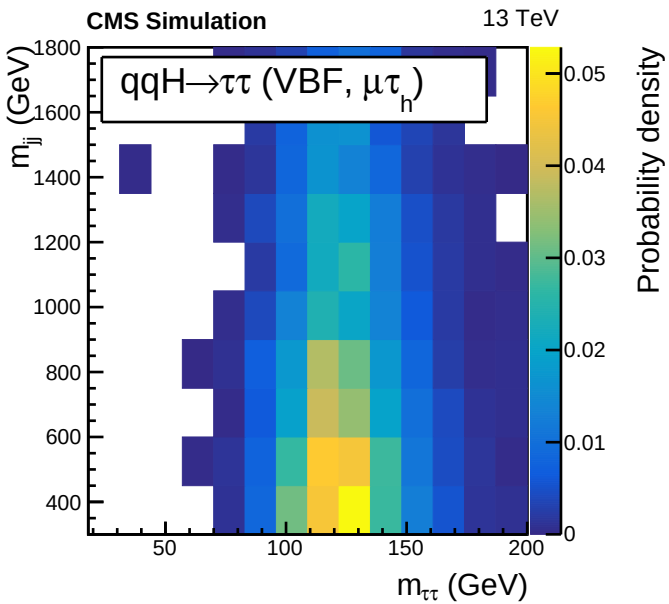


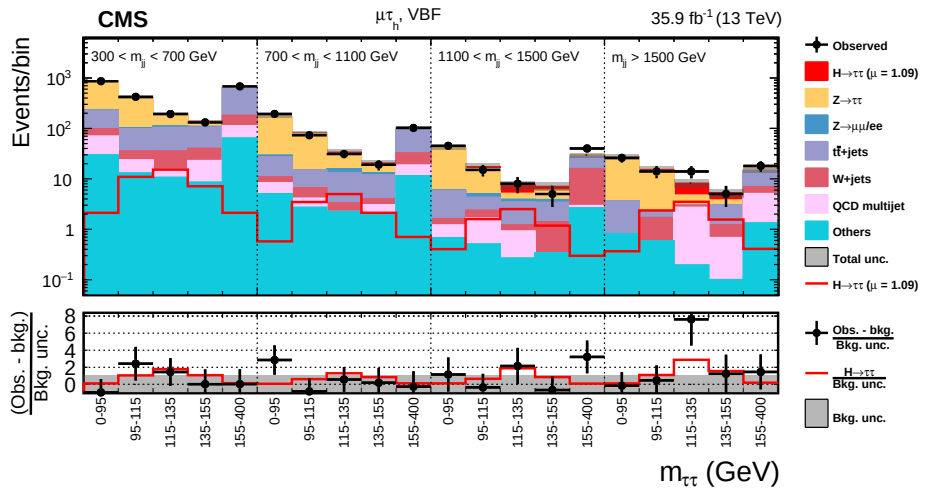
JHEP 1811 (2018)

(CMS) 185

Introduction: Maximum Likelihood Fits

- Special case: **Binned template fits**: Probability for observing a given number of counts in a given histogram bin is itself encoded in a set of histogram "templates" which are scaled and/or interpolated as a function of the model parameters
- Multi-dimensional histograms can always be "unrolled"





Phys. Lett. B 779 (2018) 283 (CMS)

Binned Maximum Likelihood Fits at the LHC

- Large scale binned maximum likelihood fits used e.g. for some Higgs measurements typically performed with combination of RooFit+Minuit2
- e.g. in CMS implemented in Higgs combination tool:
  - Likelihood constructed and computed in RooFit
  - Minimization with Minuit2
  - Gradient for minimization evaluated numerically with some variation of finite difference method (implemented internally in Minuit)
- Numerical precision of gradients strongly limited by finite difference method (though maybe smarter adapative algorithms could be used here)
- Large number of POI's and/or nuisance parameters can make convergence slow or unstable
- Large number of events can further exacerbate numerical precision and stability issues (looking for very small relative change in likelihood value)

What is TensorFlow?

- TensorFlow is a library for high performance numerical computation
- Typical workflow:
  - Construct a **computational graph** using TensorFlow library in python
  - Execute graph (transparent-to-user compilation and execution on threaded/vectorized CPU's, GPU's, etc)
- Originally developed at Google for deep learning applications
- **Efficient and numerically stable computation of gradients by backpropagation**, needed for Stochastic Gradient Descent in training of deep neural networks

Binned Maximum Likelihood Fits in Tensorflow

- Most important in this context: **Efficient and numerically stable computation of gradients** (using standard backprop)
- Parallelization, use of GPU's etc also interesting
- Goal is to be free of any technical constraints for what concerns the level of detail and sophistication needed for modeling experimental and theoretical systematic uncertainties

JB, work in progress

Likelihood Construction in TensorFlow

- Any template shape fit can be expressed as a many-channel counting experiment
- Negative log-likelihood can be written as

$$L = \sum_{ibin} \left(-n^{obs}_{ibin} \ln n^{exp}_{ibin} + n^{exp}_{ibin} \right) + \frac{1}{2} \sum_{ksyst} \left(\theta_{ksyst} - \theta^{0}_{ksyst} \right)^2 \quad (1)$$

$$n^{exp}_{ibin} = \sum_{jproc} r_{jproc} n^{exp}_{ibin,jproc} \prod_{ksyst} \kappa^{\theta_{ksyst}}_{ibin,jproc,ksyst} \quad (2)$$

- $n^{exp}_{ibin,jproc}$ is the expected yield per-bin per-process
- r_{jproc} is the signal strength multiplier per-process
- θ_{ksyst} are the nuisance parameters associated with each systematic uncertainty
- $\kappa_{ibin,jproc,ksyst}$ is the size of the systematic effect per-bin, per-process, per-nuisance
- (The above assumes all shape uncertainties are implemented as log-normal variations on individual bin yields, which is appropriate for e.g. PDF/QCD scale variations, but not for things like momentum scale/resolution variations)

Likelihood Construction in TensorFlow

- Full contents of datacards can be represented by a few numpy arrays:
  - nbin $\times$ nproc 2D tensor for expected yield per-bin per-process
  - nbin $\times$ nproc $\times$ nsyst 3D tensor for κ (actually $ln\kappa$) values parameterizing size of systematic effect from each nuisance parameter on each bin and process (actually two tensors, one each for $\ln \kappa_{up}$ and $\ln \kappa_{down}$ to allow for asymmetric uncertainties)
- POI's and nuisance parameters implemented as TensorFlow Variables
- Full likelihood constructed as TensorFlow computation graph with observed data counts as input
- Some details:
  - Precompute as much as possible with numpy arrays which are loaded into graph via tf data api from h5py arrays on disk
  - Double precision everywhere
  - Offsetting of likelihood in optimal placement within the graph to minimize precision loss

Likelihood Construction in TensorFlow

- Any template shape fit can be expressed as a many-channel counting experiment
- Negative log-likelihood can be written as

$$L = \sum_{ibin} \left(-n^{\text{obs}}_{ibin} \ln n^{\text{exp}}_{ibin} + n^{\text{exp}}_{ibin} \right) + \frac{1}{2} \sum_{ksyst} \left(\theta_{ksyst} - \theta^{0}_{ksyst} \right)^2 \tag{3}$$

$$n^{\text{exp}}_{ibin} = \sum_{jproc} r_{jproc} n^{\text{exp}}_{ibin,jproc} \prod_{ksyst} \kappa^{\theta_{ksyst}}_{ibin,jproc,ksyst} \tag{4}$$

- Likelihood evaluation reduced to essentially two large tensor contractions (matrix multiplications)
- Both dense and sparse implementations are used as appropriate

Minimization

- Minimization in TensorFlow normally done with variations on Stochastic Gradient Descent, appropriate for very large number of parameters in deep learning (10's of thousands to millions)
- For O(100's-1000's) of parameters, more appropriate to use second-order minimization techniques
- **Particularity:** Loss function needs to be minimized **exhaustively**. There is a global minimum, and further statistical analysis (determining confidence interevals etc) requires finding it to high accuracy
- Hessian can be computed analytically but still slow and not very optimal $\rightarrow$ use quasi-newton methods which approximate hessian from change in gradient between iterations (the MIGRAD algorithm in Minuit/Minuit2 belongs to this class of algorithms, as does BFGS)

Minimization

- **While the likelihood has a global minimum and is well behaved in the vicinity, it is (apparently) NOT convex everywhere in the parameter space**
  - BFGS-type quasi-Newton methods are not appropriate since the Hessian approximation can never capture non-convex features
  - Line search is not a good strategy even with a well-approximated (or exact) Hessian, since this will tend to get stuck or have slow convergence near saddle points/in non-convex regions
  - Major source of non-convexity is the polynomial interpolation of $\ln \kappa$ for asymmetric log normal uncertainties
- Started with trust-region based minimizer with SR1 approximation for hessian, as implemented in SciPy (minimal adaptation required for existing TensorFlow-SciPy interface)
  - Bonus: this also supports arbitrary non-linear constraints
  - **Caveat:** Only likelihood and gradient evaluation done in Tensorflow, rest of minimizer is in python/numpy

Some Performance Tests

	Likelihood	Likelihood+Gradient	Hessian
Combine, TR1950X 1 Thread	10ms	830ms	-
TF, TR1950X 1 Thread	70ms	430ms	165s
TF, TR1950X 32 Thread	20ms	71ms	32s
TF, 2x Xeon Silver 4110 32 Thread	17ms	54ms	24s
TF, GTX1080	7ms	13ms	10s
TF, V100	4ms	7ms	8s

- (1444 bins, 96 POI's, 70 nuisance parameters, 180M expected events)
- n.b. these numbers are with an older implementation, all have improved
- Single-threaded CPU calculation of likelihood is 7x **slower** in Tensorflow than in Roofit (to be understood and further optimized)
- Gradient calculation in combine/Minuit is with 2n likelihood evaluations for finite differences (optimized with caching)
- Xeons are lower clocked than Threadripper, but have more memory channels and AVX-512
- Back-propagation calculation of gradients in Tensorflow is much more efficient (in addition to being more accurate and stable)
- Best-case speedup is already a factor of 100

Some Performance Tests: Minimization

	Minimization		
	L+Gradient	scipy trust-constr	scipy cpu usage
TF, TR1950X 32 Thread	71ms/call	200ms/iteration	2107%
2x Xeon Silver 4110 32 Thread	54ms/call	237ms/iteration	2587%
TF, GTX1080 (+TR1950X)	13ms/call	84 ms/iteration	1081%
TF, V100 (+2x Xeon 4110)	7ms/call	78ms/iteration	1558%

- Each iteration of the SR1 trust-region algorithm requires exactly 1 likelihood+gradient evaluation
- Significant amount of processing power (and CPU bottleneck) in scipy+numpy parts of the minimizer (non-trivial linear algebra)

Further Optimizing Minimization

- Current SR1 trust-region implementation in scipy based on conjugate gradient method for solving the quadratic subproblem $\rightarrow$ large number of inexpensive sub-iterations which don't parallelize well
- Have implemented several variants of quasi-newton trust region minimizers natively in TensorFlow
- Most advanced based on L-SR1 Orthonormal basis minimization (arXiv:1506.07222), including a new non-limited-memory variant with direct update to eigen-decomposition of Hessian
- Hessian-free methods (e.g "trust-krylov" in SciPy) are also interesting since they can be used with exact Hessian-vector products computed efficiently with backprop, but in practice these require many Hessian-vector product evaluations per-iteration

Some Performance Tests: Minimization

	Minimization		
	L+Gradient	scipy trust-constr	TF TrustSR1Exact
TF, TR1950X 32 T	71ms/call	200ms/iteration	89ms/iteration
2x Xeon Silver 4110 32 T	54ms/call	237ms/iteration	63ms/iteration
TF, GTX1080 (+TR1950X)	13ms/call	84ms/iteration	55ms/iteration
TF, V100 (+2x Xeon 4110)	7ms/call	78ms/iteration	51ms/iteration

- Example here with iterative Cholesky decomposition to solve TR subproblem (a la Nocedal and Wright algo 4.3)
- Substantial reduction of overhead relative to bare likelihood+gradient call
- Relative remaining overhead much larger on GPU
- n.b, this fit converges in about 500 iterations with the TrustSR1Exact algorithm, about 25s/fit with GPU
- Using gradient descent methods available in Tensorflow requires O(10k) iterations

Updated Performance Tests

(Newer TensorFlow, further optimized, but larger model)

	Likelihood	L+Grad	Hessian	MaxRSS
TF, TR1950X 1 Thread (pfor)	26ms	73ms	7.9s	3000MB
TF, TR1950X 32 Thread (pfor)	39ms	83ms	1.1s	3900MB
TF, GTX1080 (+TR1950X) (loop)	64ms	69ms	3.0s	2900MB
TF, GTX1080 (+TR1950X) (pfor)	64ms	69ms	0.8s	2900MB

- (1824 bins, 101 processes, 96 POI's, 257 nuisance parameters)
- Size of raw arrays is 760MB
- non-pfor hessian calculation failed with "Already exists: Resource" errors without " on CPU

Updated Performance Tests: Large/Sparse Model

	Likelihood	L+Grad	Hessian	MaxRSS
Sparse TF, TR1950X 1 Thread	24ms	40ms	52s	980MB
Sparse TF, TR1950X 32 Thread	40ms	70ms	3.7s	1200MB
Dense TF, TR1950X 1 Thread	245ms	540ms	-	6800MB
Dense TF, TR1950X 32 Thread	237ms	534ms	-	7000MB

- (1296 bins, 655 processes, 648 POI's, 444 nuisance parameters)
- GPU not available with standard build (SparseTensorDenseMatMul)
- Size of raw arrays in dense mode is $6GB$
- pfor for Hessian not available in Sparse case (SparseTensorDenseMatMul not supported)
- Hessian computation in dense mode caused OOM with pfor, and "Already exists: Resource" errors without
- Dense model too big for my GPU

Optimizing Memory Consumption

- This type of model has a peculiar feature of very large constants (3-tensor representing systematic variations on templates can be several GB especially in dense mode with larger numbers of processes and systematic variations)
- To optimize memory consumption for graphs with large constants:
  - **Don't** include large constants in the graph definition (there is also a hardcoded 2GB limit in doing so)
  - **Don't** read large numpy arrays from disk (unless using memmapping, but then can't use compression)
  - **Don't** store large constants in tf Variables (because it's apparently impossible to initialize them without having at least a second copy of the contents in memory)

Optimizing Memory

- Adopted solution
  - HDF5 arrays with chunked storage and compression
  - Numpy arrays are stored as flattened HDF5 arrays to allow reading chunk by chunk while preserving the order of the array and maintaining flexibility in choice of chunk size
  - Read chunk by chunk using tf data API with tf py_func to interface with h5py
  - Use batching to reassemble full array into a single tensor, then use the in-memory cache so the read only happens once (reshaping and possible truncation of the overflow from the last batch have near-zero cpu or memory footprint)
  - Text+root histogram conversion has been adapted to write hdf5 arrays instead of a tf graph with in-built constants
- (Avoiding a second copy in memory took some patience and was not obvious how to achieve)

Covariance Matrices

- Irrespective of minimization algorithm, often want to compute covariance matrix at the end for interpreting uncertainties $\rightarrow$ compute Hessian and invert it
- New vectorized pfor construction gives large speedups for this (so much that full second-order minimization methods are even feasible in some cases)

Other Optimization Opportunities

- Detailed study of scaling of minimization overhead/performance with number of free parameters is needed
- Most likely there is further room for improvement with better algorithms/ones more suited for GPU's
- Efficiency of specific matrix factorization steps to be carefully checked/profiled
- Batch evaluation of likelihood feasible/useful? (parallel minimization algorithm? Multiple toys in parallel?)
- Implement simpler χ^2/Gaussian approximation to likelihood for high statistics cases

Implementation

- Code lives here: `https://github.com/bendavid/HiggsAnalysis-CombinedLimit/tree/tensorflowfit` (not very streamlined for the moment, since the priority has been on a particular set of physics analyses in progress with it, and currently somewhat intertwined with existing CMS fitting tools)
- Two scripts:
  - **scripts/text2hdf5.py**: Create tensorflow graph from datacards/ROOT histograms (outputs hdf5 file containing flattened arrays for large constant tensors)
  - **scripts/combinetf.py**: Construct graph, load constant arrays into tensors, run fits/toys/scans with graph
- Some interesting bits related to reading hdf5 arrays, some sparse tensor operations, and minimization in python area
- Second order minimizers will be interesting to contribute upstream (and some work already on L-SR1 algorithms for more conventional deep learning applications, e.g arXiv:1807.00251)

Tensorflow Maximum Likelihood Fitting: Other Related Work

- Some related efforts by others in parallel (some more focus on general frameworks, less on large-scale performance optimization so far):
  - pyhf: `https://github.com/diana-hep/pyhf`
  - ZFit: `https://github.com/zfit/zfit`
- Small working group formed amongst interested people

Conclusions

- Many existing measurements targeting constraints on PDFs in or near relevant phase space for precision electroweak measurements at LHC
- Existing measurements of $\sin^2\theta_W$ and M_W already exploit in-situ constraints on PDFs to varying degrees
- Phenomenological studies and projections from experiments indicate further potential in this direction
- Ultimate precision in this direction may also require technical improvements in statistical interpretation
- Interesting new measurements are expected in the coming months and years