QCD @ LHC 2012

20th-24th August 2012 at Michigan State University

Bundesministerium
für Bildung
und Forschung

Studies on PDF Uncertainties

Sasha Glazov, Voica Radescu

Outline:

- Motivation
- Methods
- Comparisons
- Summary

HELMHOLTZ
GEMEINSCHAFT

DESY

PDF Parametrisation methods

- Extraction of the PDFs from fits relies on ansatz such as parametrisation which still needs to be understood.
- Currently there are various approaches to take this ansatz into account:
  1. HERAPDF estimates uncertainty on the PDF parametrisation by scanning the parameter space starting from a basic functional form:

$$x f(x, Q_0^2) = A x^B (1-x)^C$$

and adds a parameter at a time till no improvement in chisquare is observed hence the number of parameters is chosen by saturation of the χ^2 method.

→ there is some arbitrariness in the method that has been criticized

  2. NNPDF uses redundant parametrisation (>200p) and introduces a stopping criteria based on data (Data Driven Regularisation) - see Alberto's talk [NNPDF]
  3. Chebyshev Polynomial with the regularisation criteria using chisquare penalty term
     [J. Pumplin DIS 2011, S. Glazov, S. Moch, VR PLB27193]

- There are different approaches to deal with experimental uncertainties:
  1. Hessian Method [J. Pumplin method – see CTEQ talk of Dan Stump]
  2. MC Method [a la NNPDF, [arXiv:1101.0536 [hep-ph]]]

The study here combines these ideas and it is based on the HERAFitter Platform

[see Ringaile's talk] http://herafitter.hepforge.org/

HERA PDF Parametrisation 13p vs 22p

- This is study is based on HERA I published data, using initial HERAPDF settings (NNLO)
- Following PDFs are parametrised at the starting scale (below the charm threshold)

$$xg, xu_v, xd_v, x\bar{U} = x\bar{u}, x\bar{D} = x\bar{d} + x\bar{s}$$

- The PDF parametrisation form is able to describe data with few input paramaters

$$xf(x) = Ax^B(1-x)^C P(x);$$
$$P(x) = 1 + Dx + Ex^2$$

- A - normalisation
- B - low x behaviour
- C - high x behaviour
- D,E - medium x tuning

Additional Contraints:

- Quark Number Sum Rules
- Momentum Sum Rule
- $B_{\bar{U}} = B_{\bar{D}}$ & $A_{\bar{U}} = A_{\bar{D}}(1\text{-}f_s)$

$\bar{u} \to \bar{d}$ as $x \to 0$

$$\begin{aligned}
xu_v(x) &= A_{u_v}x^{B_{u_v}}(1-x)^{C_{u_v}}(1+E_{u_v}x^2), \\
xd_v(x) &= A_{d_v}x^{B_{d_v}}(1-x)^{C_{d_v}}, \\
x\bar{U}(x) &= A_{\bar{U}}x^{B_{\bar{U}}}(1-x)^{C_{\bar{U}}}, \\
x\bar{D}(x) &= A_{\bar{D}}x^{B_{\bar{D}}}(1-x)^{C_{\bar{D}}}, \\
xg(x) &= A_g x^{B_g}(1-x)^{C_g} - A'_g x^{B'_g}(1-x)^{25}.
\end{aligned}$$

HERA PDF Parametrisation 13p vs 22p

- This is study is based on HERA I published data, using initial HERAPDF settings (NNLO)
- Following PDFs are parametrised at the starting scale (below the charm threshold)

$$xg, xu_v, xd_v, x\bar{U} = x\bar{u}, x\bar{D} = x\bar{d} + x\bar{s}$$

- The PDF parametrisation form is able to describe data with few input paramaters

$$xf(x) = Ax^B(1-x)^C P(x);$$
$$P(x) = 1 + Dx + Ex^2$$

- A - normalisation
- B - low x behaviour
- C - high x behaviour
- D,E - medium x tuning

- In HERAPDF 13 free parameters are considered to describe HERA I data with the 14 free parameters used to evaluate the parametrisation uncertainty

- Here we allow for extra 2 free parameters for every PDF distribution, hence we end up with 22p

Experimental uncertainties are estimated using MC method

MC method to estimate exp uncertainties

- Method consists in preparing replicas of data sets allowing the central values of the cross sections to fluctuate within their systematic and statistical uncertainties taking into account all point to point correlations.
  - Various assumptions can be considered for the error distributions: Gauss, Log-Normal, etc. ...
- Shift central values randomly within their uncorrelated errors assuming Gauss distributions of the errors:

$$\sigma_i = \sigma_i(1 + \delta_i^{uncorr} RAND_i)$$

- Shift central values with the same probability of the corresponding correlated systematic shift assuming Gauss distribution of the errors:

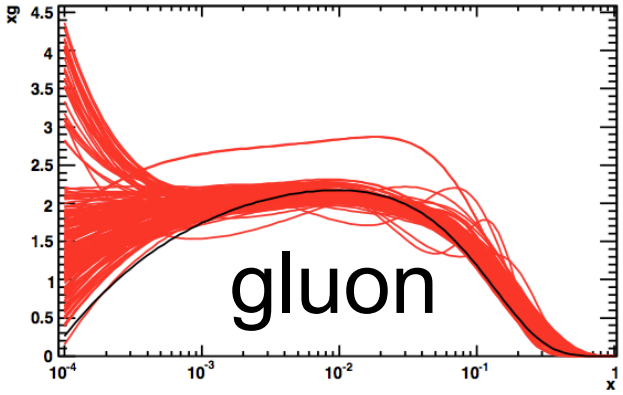
$$\sigma_i = \sigma_i(1 + \delta_i^{uncorr} RAND_i + \sum_j^{N_{sys}} \delta_{ij}^{corr} RAND_j)$$

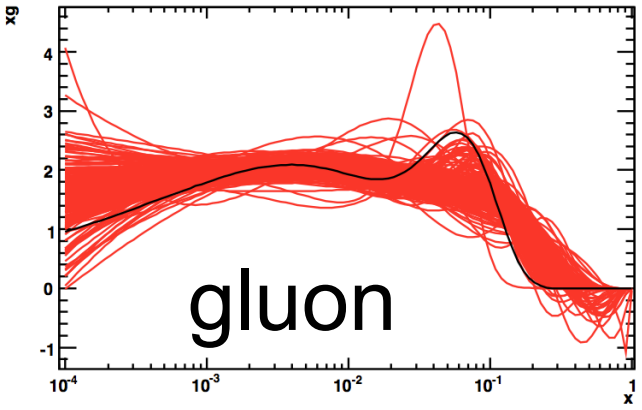
- Preparation of the data is repeated for N times (N>100):
  - For each replicas NLO QCD fit is performed to extract the PDF set
- Errors on the PDFs are estimated from the RMS of the spread of the N curves corresponding to the N individual extracted PDFs.

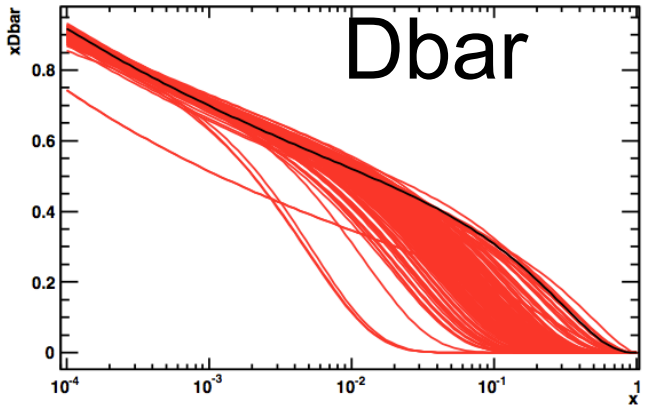
MC 13p vs 22p

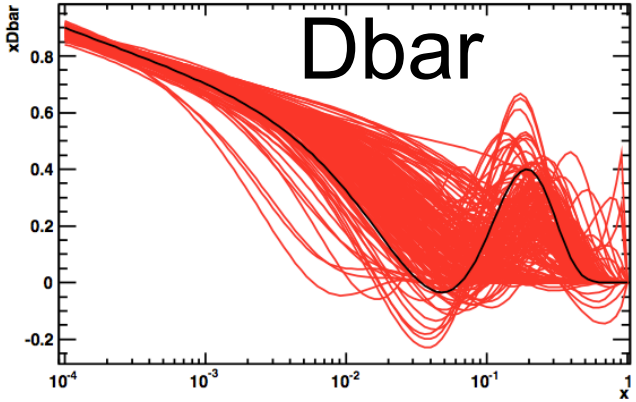
- An illustration using MC method to estimate experimental uncertainties with 200 replicas based on 13p fit (left) vs 22p fit (right)
  - More freedom observed for the 22p fit ➔ sufficient freedom for HERA I only data.

13p Q^2=1.9 22p



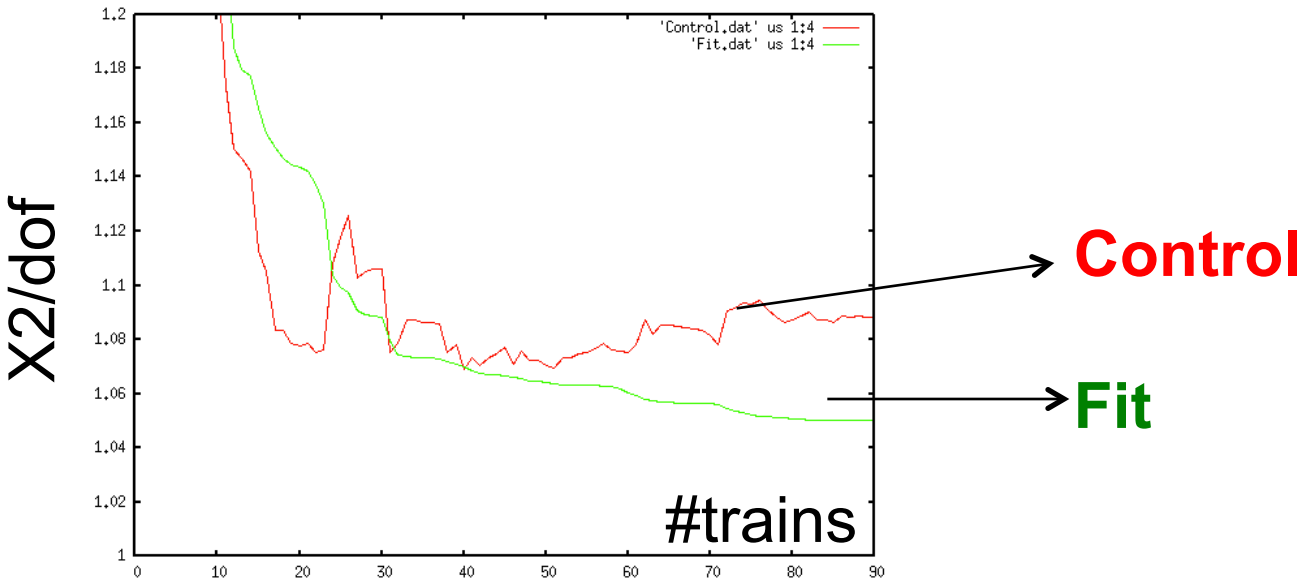






1. Data Driven Regularisation method

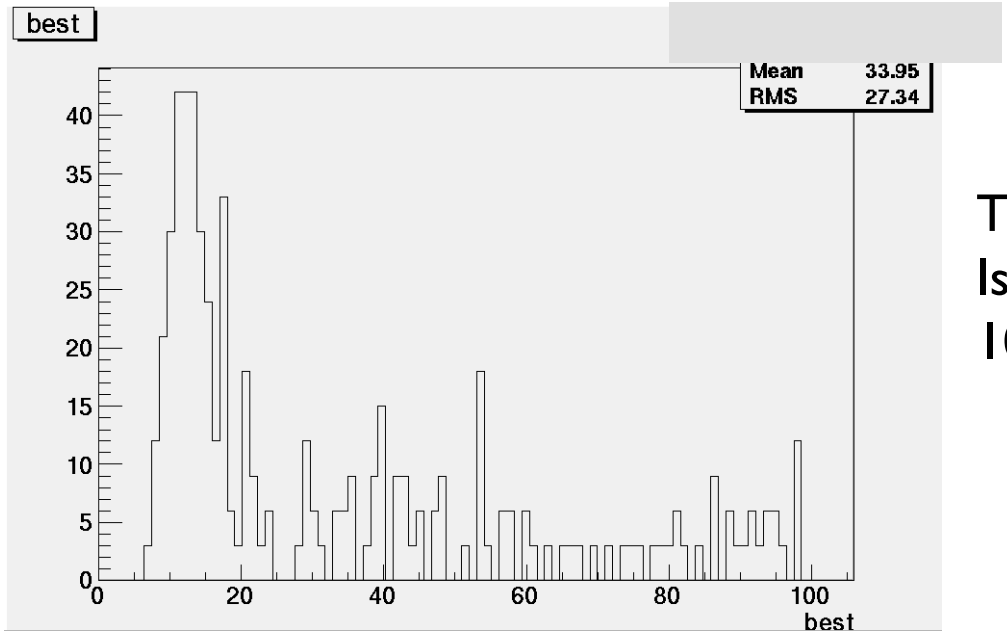
1. Data driven regularisation by splitting data randomly into "fit" and "control" samples
   - **"Fit"** sample used for determining PDF parameters, semi-monotonically decreases in chisquare
   - **"Control"** sample used to protect against over-fitting: it starts by decrease in chisquare but later on gets to increase due to fluctuation of the sample.
   - Technically:
     - Request MIGRAD minimisation with 100 calls and repeat this procedure 100 times (trains)
     - Save the output of minuit minimisation for each train



https://znwiki3.ifh.de/HERAFitter/HERAFitter/HERAFitterMeetings/Meeting2012-Aug-01?action=AttachFile&do=view&target=status.pdf

1. Data Driven Regularisation method

1. Data driven regularisation by splitting data randomly into "fit" and "control" samples
   - **"Fit"** sample used for determining PDF parameters, it semi-monotonically decreases in chisquare
   - **"Control"** sample used to protect against over-fitting: it starts by decrease in chisquare but later on gets to increase due to fluctuation of the sample.
   - Technically:
     - Request MIGRAD minimisation with 100 calls and repeat this procedure 100 times (trains)
     - Save the output of minuit minimisation for each train

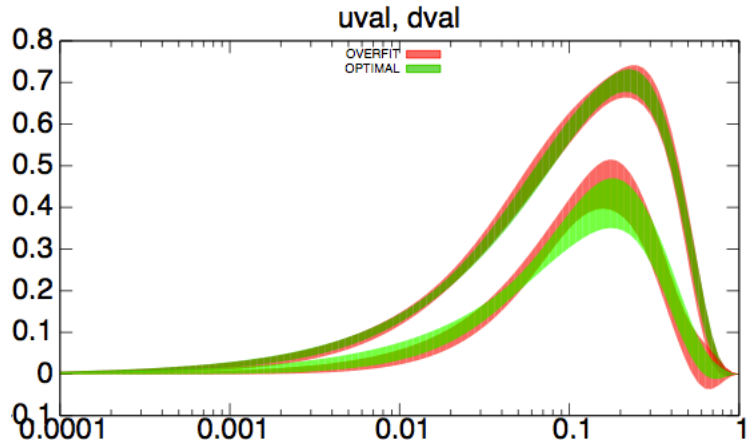


The best (optimal) fit
Is found well within the
100 trains

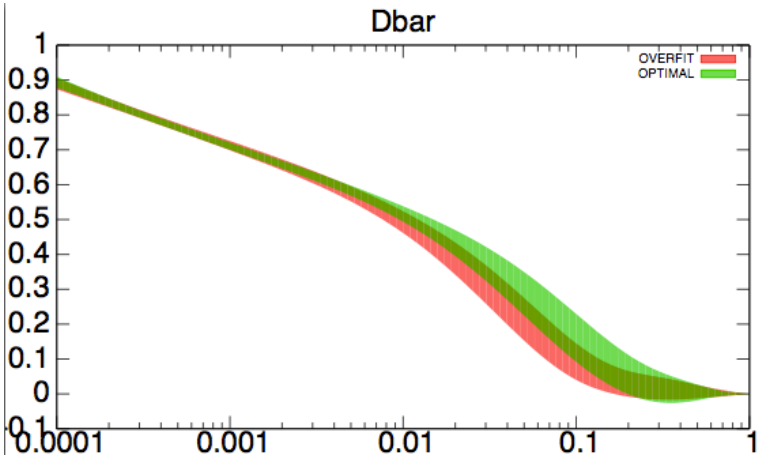
https://znwiki3.ifh.de/HERAFitter/HERAFitter/HERAFitterMeetings/Meeting2012-Aug-01?action=AttachFile&do=view&target=status.pdf

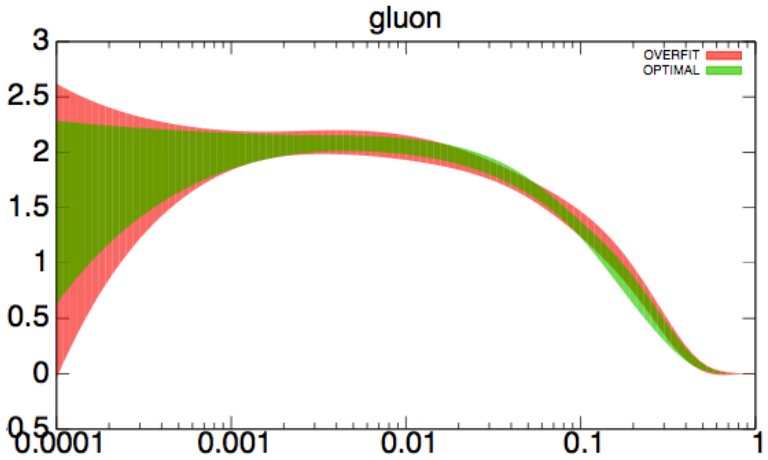
optimal vs overfit (using 200 random splits)

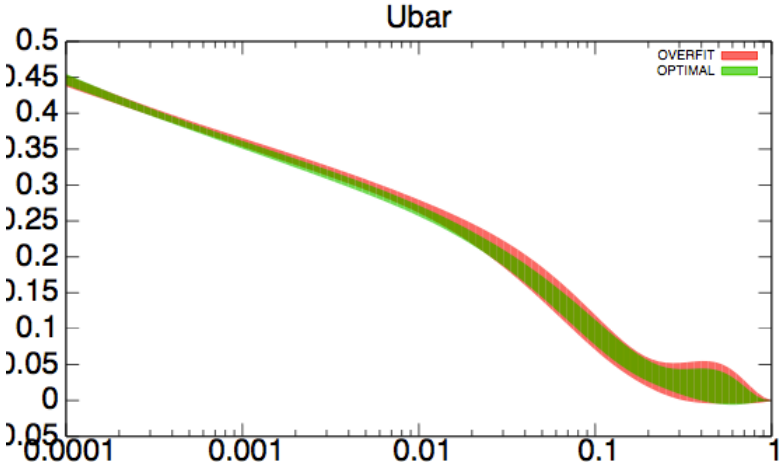
- Last train from the fit sample is called here Overfit,
- Optimal fit corresponds to the train with minimal chisquare of the control sample
- Optimal fit provides smaller uncertainties and smoother shapes for PDFs



$Q^2=1.9$



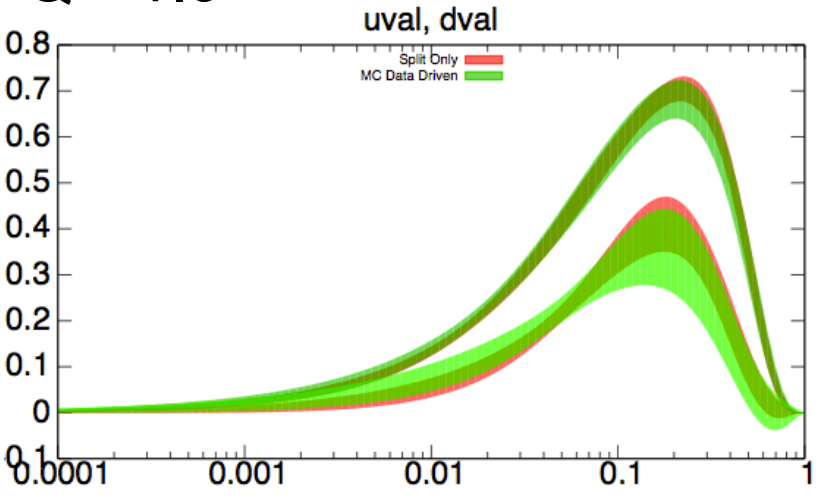


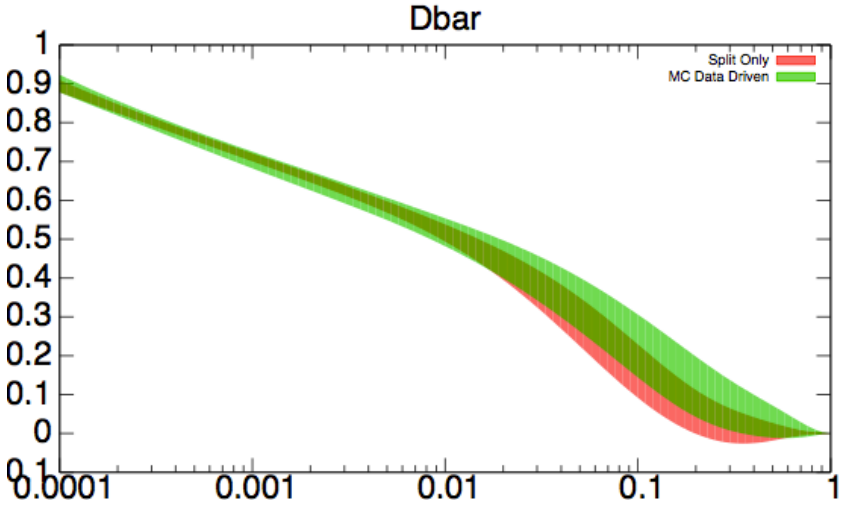


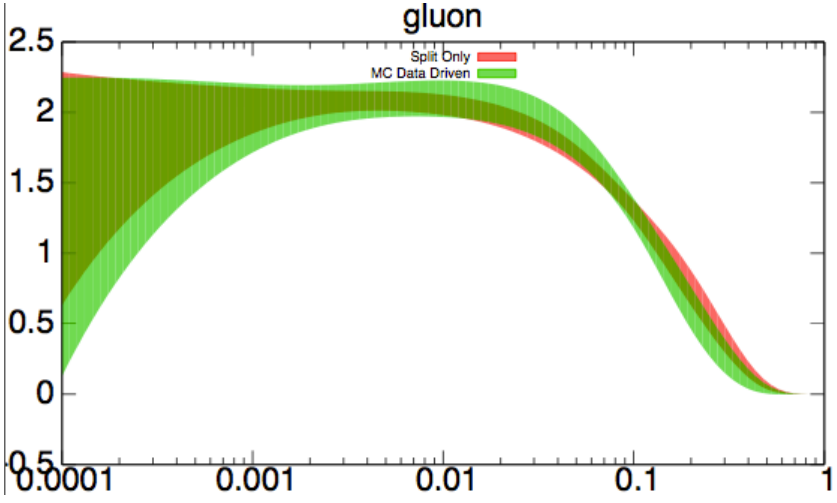
Optimal random split only vs random split+MC

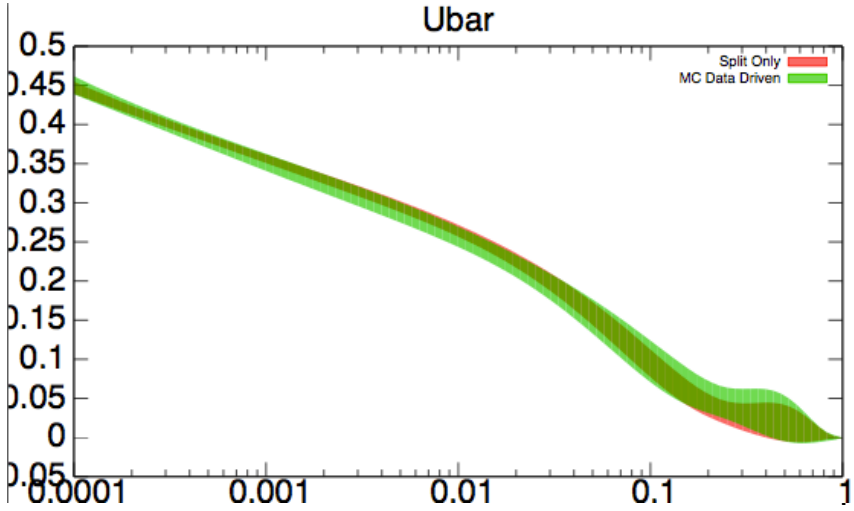
- Random split only provides sensitivity to the parametrisation uncertainty only, while Random split of the data combined with MC method provides both experimental + parametrisation uncertainties

$Q^2=1.9$





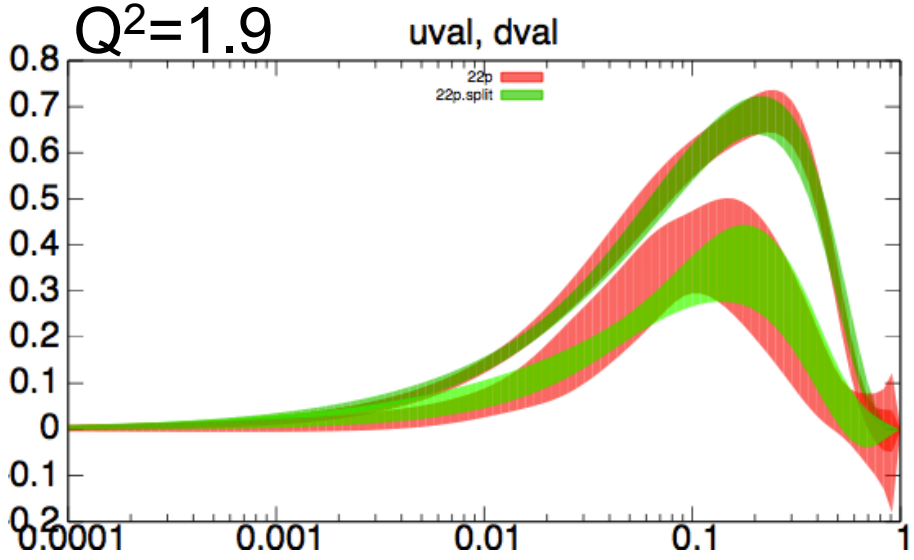


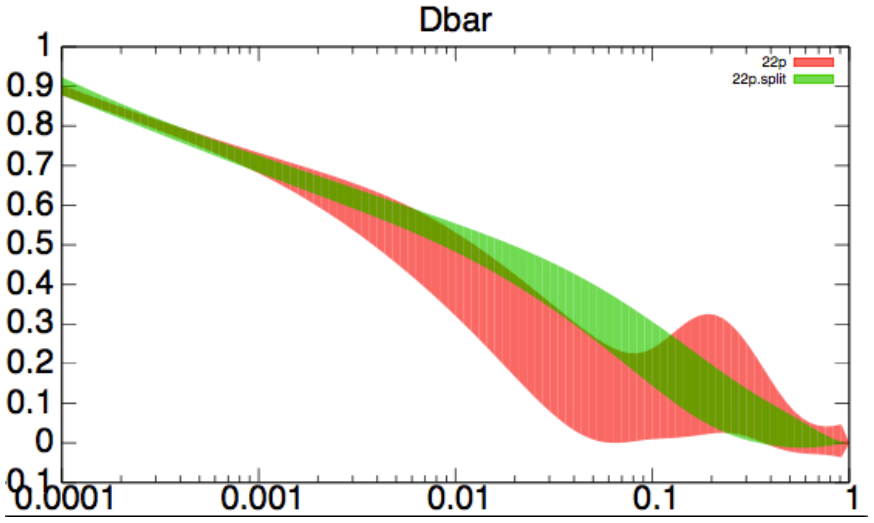


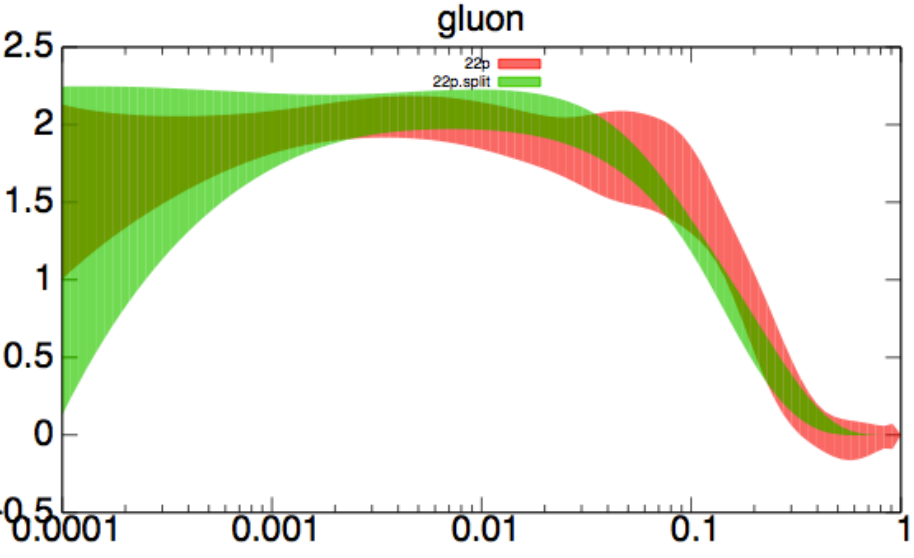
Optimal vs standard (using MC errors)

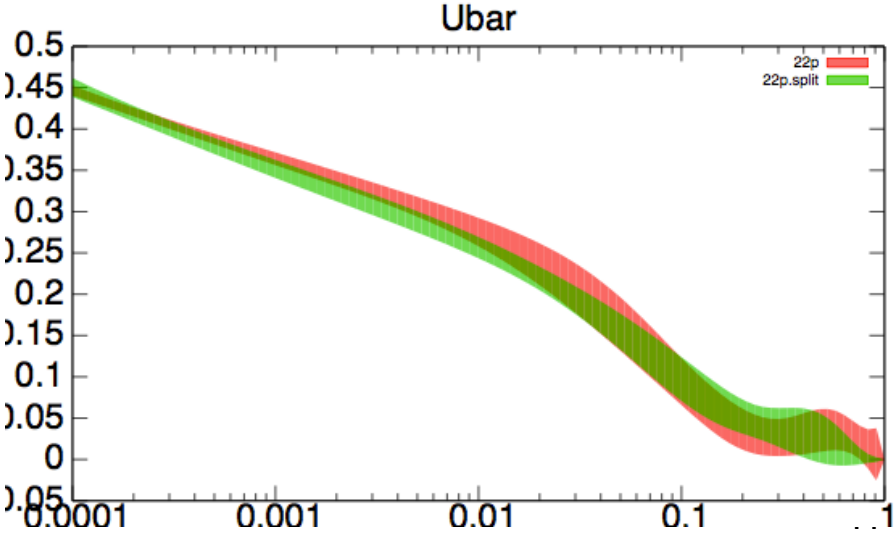
- Optimal fit provides smaller uncertainties and smoother shapes for PDFs

  ➔ the Data Driven Regularisation method works.









External Regularisation based on penalty term in χ^2

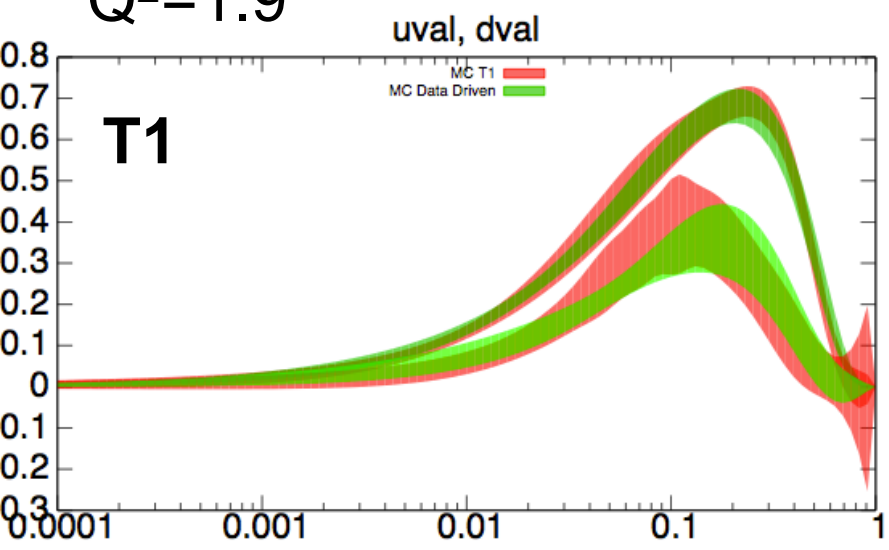
2. A simple χ^2 penalty term for a deviation from a simple PDF parametrisation form:

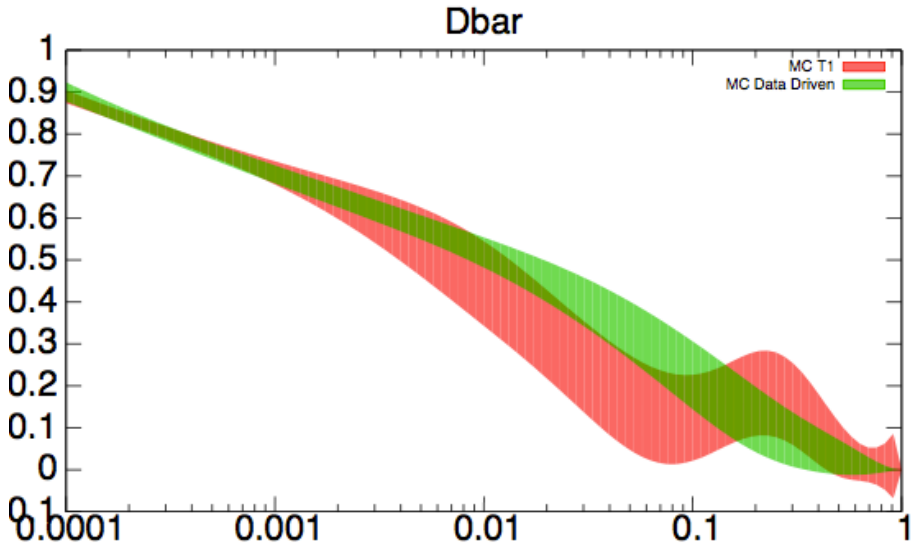
$$\chi^2_{\text{reg}} = T \sum_f \left(\left(\frac{D_f}{\Delta_D} \right)^2 + \left(\frac{E_f}{\Delta_E} \right)^2 \right)$$

   - with $\Delta D = \Delta E = 100$, such that for large D, E the ratio will approach 1,
   - T the regularisation parameter: T=0, no penalty, T>>0 strong penalty

- The idea is to compare the Data Driven Regularisation method with the external regularisation procedure to tune the T parameter.

$Q^2=1.9$





External Regularisation based on penalty term in χ^2

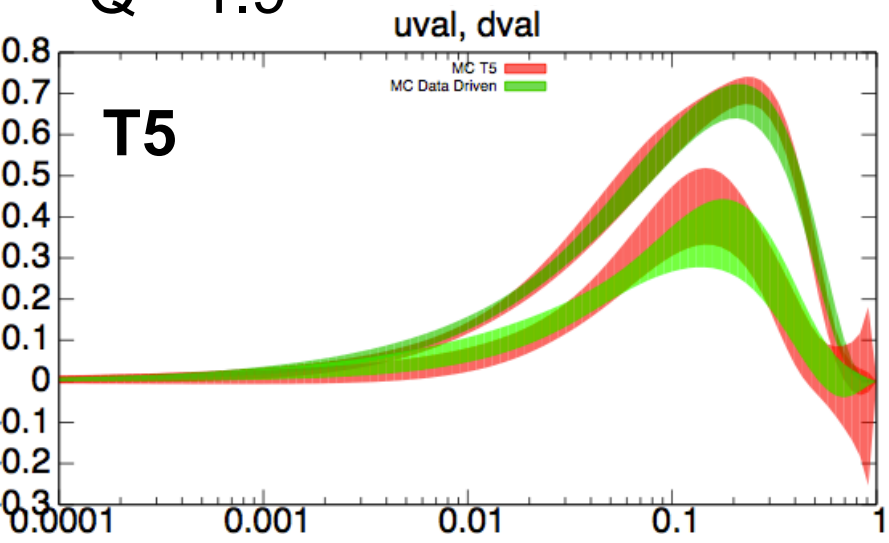
2. A simple χ^2 penalty term for a deviation from a simple PDF parametrisation form:

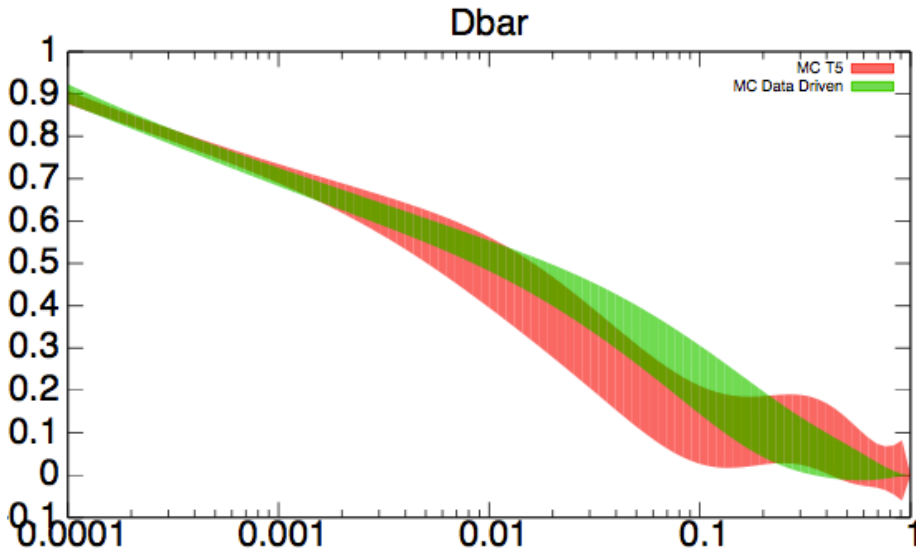
$$\chi^2_{\text{reg}} = T \sum_f \left(\left(\frac{D_f}{\Delta_D} \right)^2 + \left(\frac{E_f}{\Delta_E} \right)^2 \right)$$

   - with $\Delta D = \Delta E = 100$, such that D, E large approach 1,
   - T the regularisation parameter: T=0, no penalty, T>>0 strong penalty

- The idea is to compare the Data Driven Regularisation method with the external regularisation procedure to tune the T parameter.

$Q^2=1.9$





External Regularisation based on penalty term in χ^2

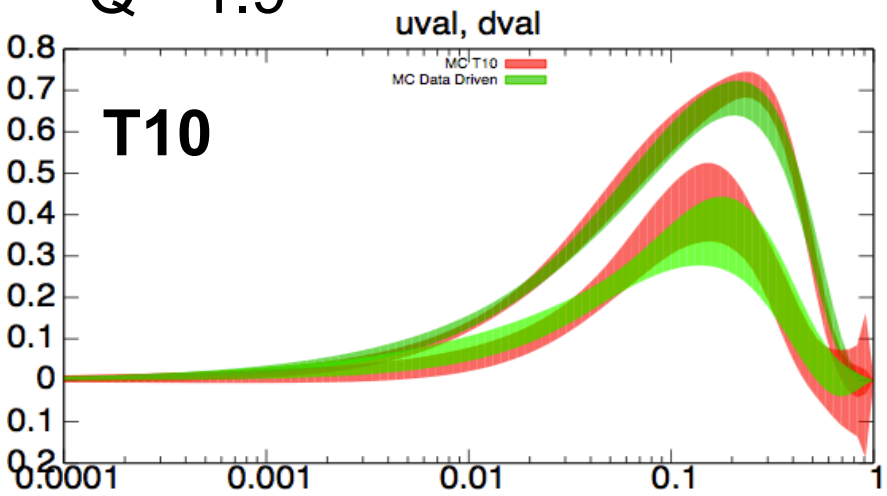
2. A simple χ^2 penalty term for a deviation from a simple PDF parametrisation form:

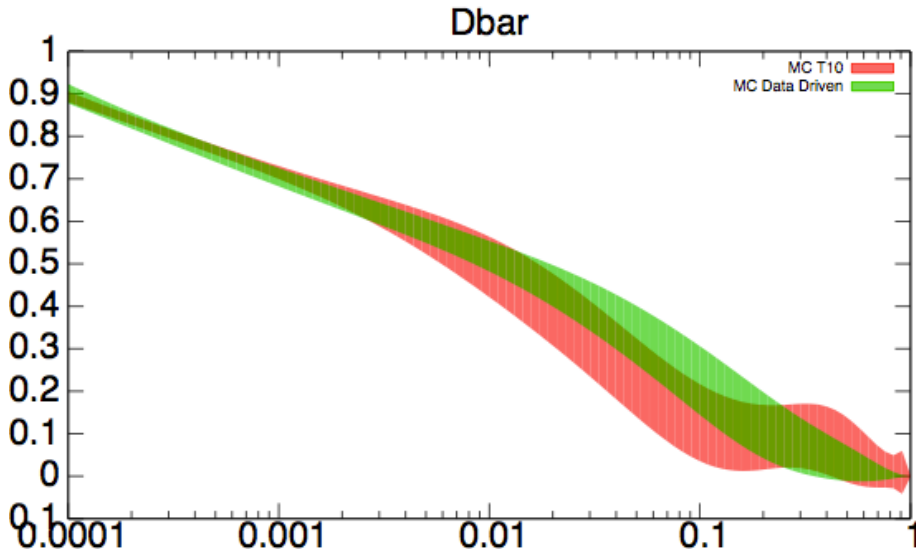
$$\chi^2_{\text{reg}} = T \sum_f \left(\left(\frac{D_f}{\Delta_D} \right)^2 + \left(\frac{E_f}{\Delta_E} \right)^2 \right)$$

   - with $\Delta D = \Delta E = 100$, such that D, E large approach 1,
   - T the regularisation parameter: T=0, no penalty, T>>0 strong penalty

- The idea is to compare the Data Driven Regularisation method with the external regularisation procedure to tune the T parameter.

$Q^2=1.9$





External Regularisation based on penalty term in χ^2

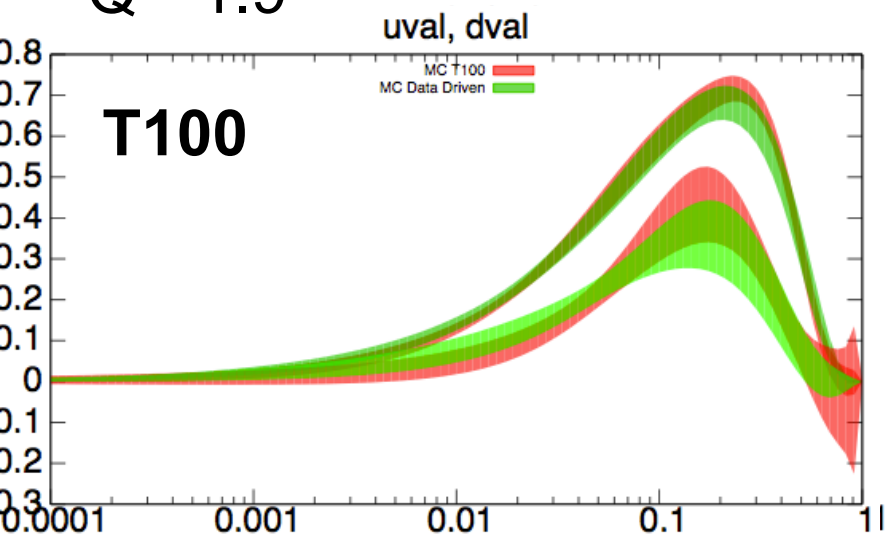
2. A simple χ^2 penalty term for a deviation from a simple PDF parametrisation form:

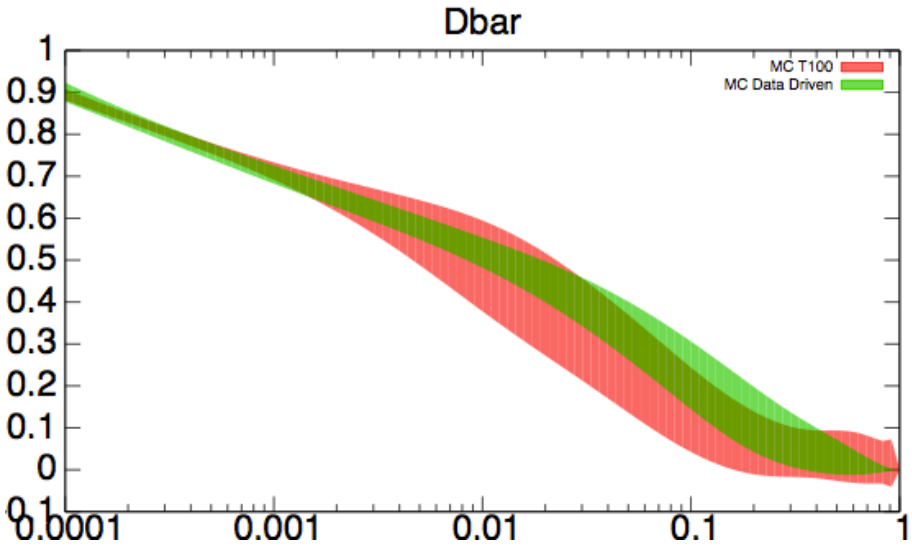
$$\chi^2_{\text{reg}} = T \sum_f \left(\left(\frac{D_f}{\Delta_D} \right)^2 + \left(\frac{E_f}{\Delta_E} \right)^2 \right)$$

   - with $\Delta D = \Delta E = 100$, such that D, E large approach 1,
   - T the regularisation parameter: T=0, no penalty, T>>0 strong penalty

- The idea is to compare the Data Driven Regularisation method with the external regularisation procedure to tune the T parameter.

$Q^2=1.9$





External Regularisation based on penalty term in $\chi 2$

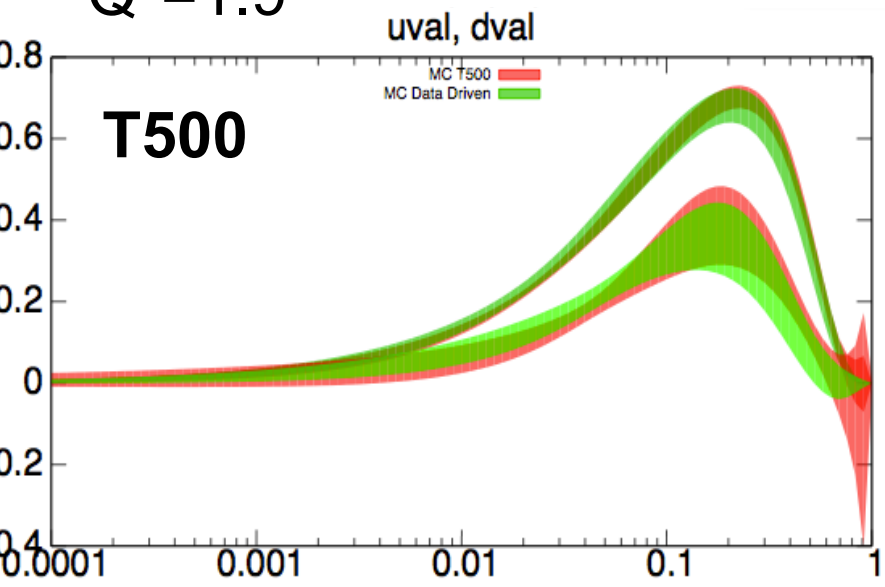
2. A simple χ^2 penalty term for a deviation from a simple PDF parametrisation form:

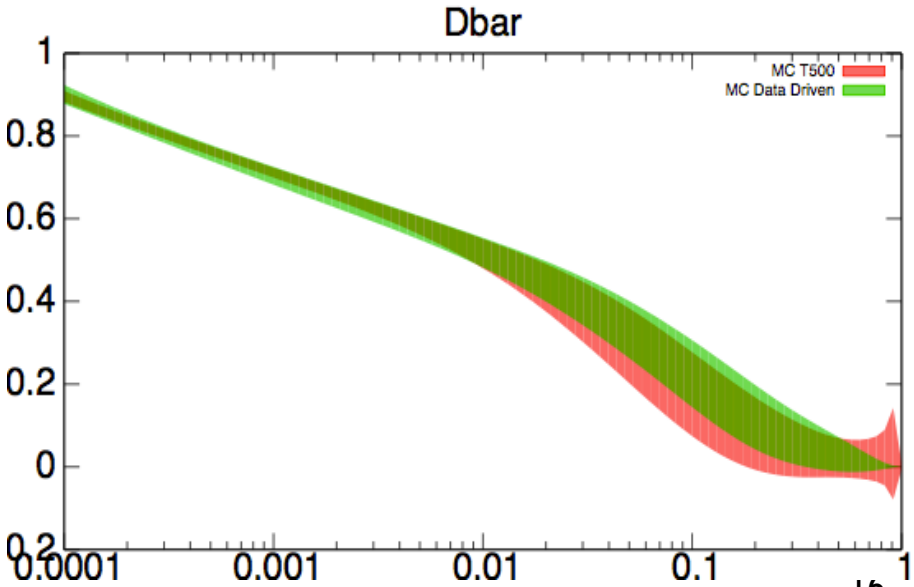
$$\chi^2_{\text{reg}} = T \sum_f \left(\left(\frac{D_f}{\Delta_D} \right)^2 + \left(\frac{E_f}{\Delta_E} \right)^2 \right)$$

   - with $\Delta D = \Delta E = 100$, such that D, E large approach 1,
   - T the regularisation parameter: T=0, no penalty, T>>0 strong penalty

- The idea is to compare the Data Driven Regularisation method with the external regularisation procedure to tune the T parameter.

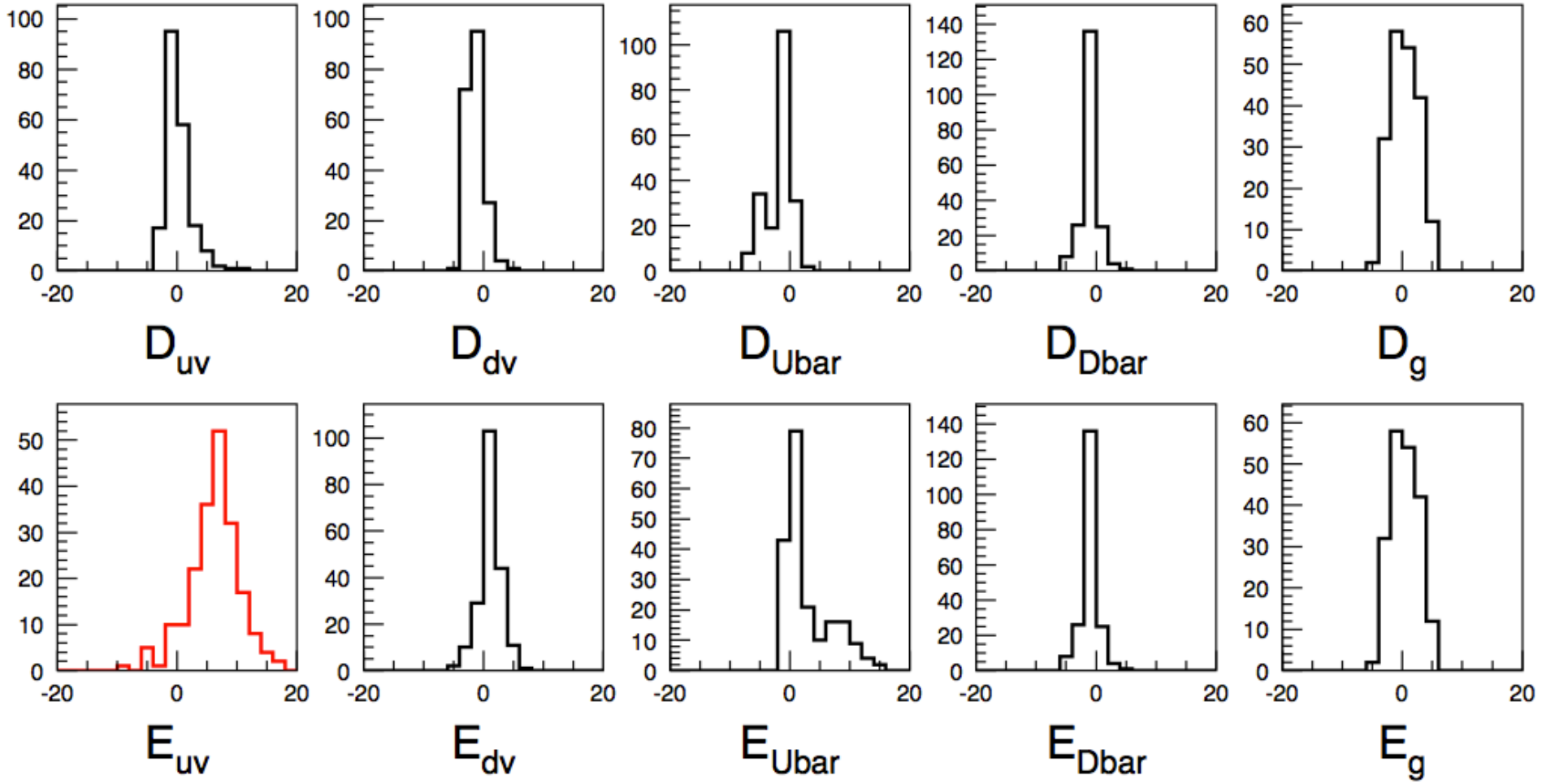
$Q^2=1.9$





Distributions of extra parameters for T=500 case

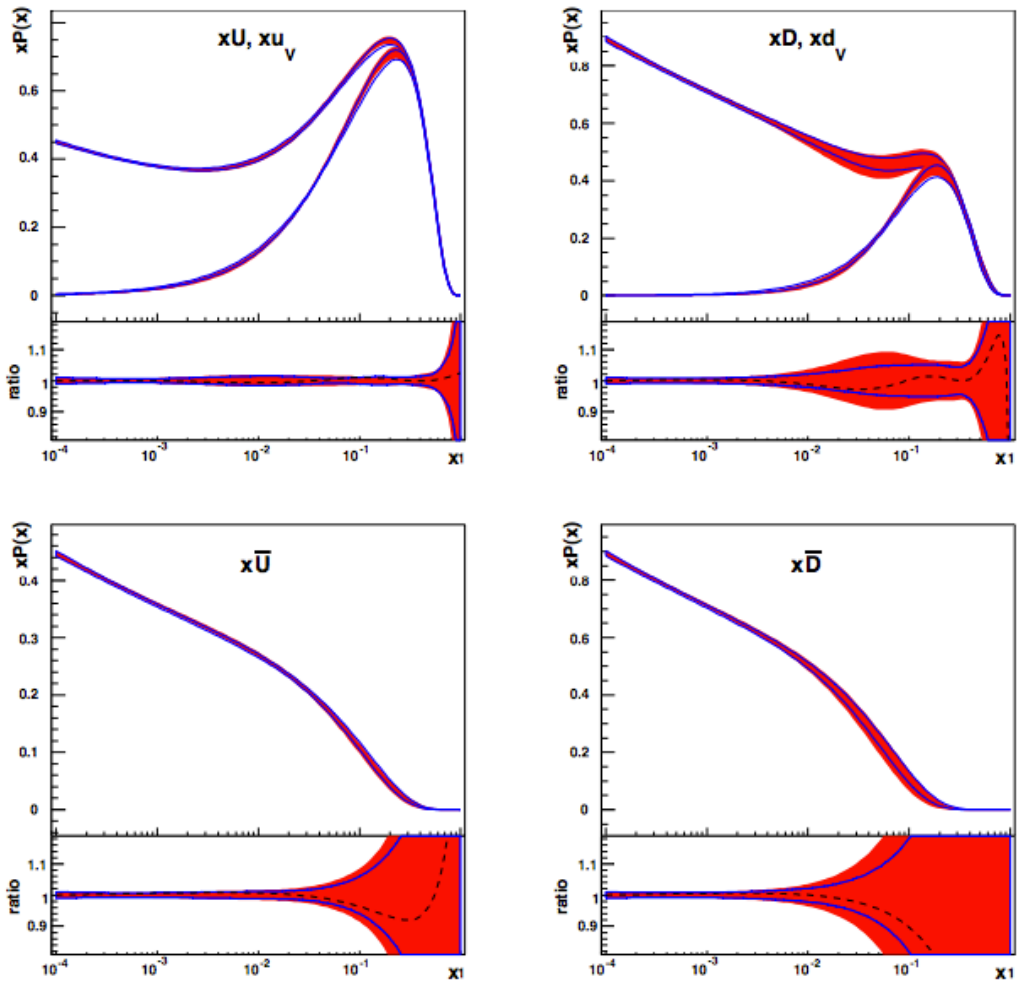
- Interesting to observe that for the case with external regularisation (T=500) which is the closest to data driven method the extra parameter that's non zero is Euv → similar to 13p fit



Hessian (13p+T=0) vs (22p+T=500)

- We compare here the unregularised method with 13p vs regularised with more flexibility T=500 and observe that the method of using external regularisation does work as well.

$Q^2=1.9$



Summary

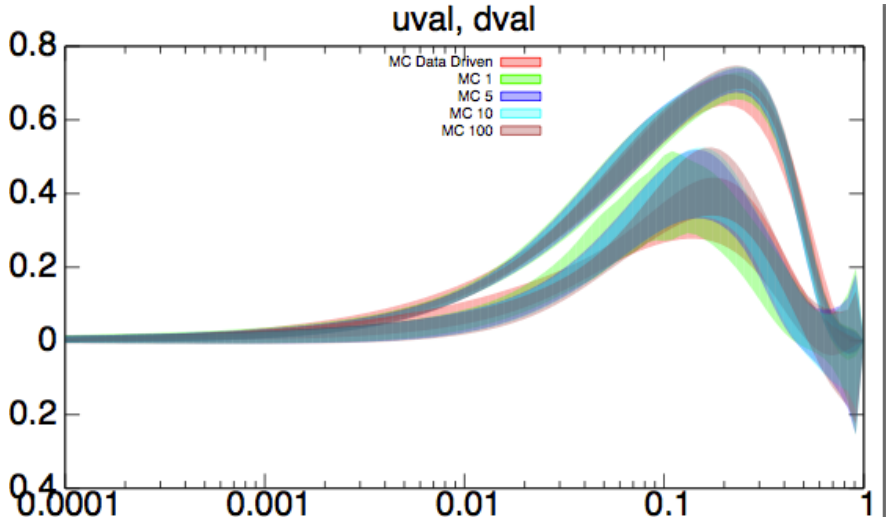
- There is need to study the parametrisation assumptions on PDFs:
- There are various methods to estimate the ansatz uncertainties.
- HERAFitter framework provides the means to study and compare various methods:
  - Data Driven Regularisation
  - External Regularisation through chisquare penalty terms
    - Both methods have been shown to work when using large data set from HERA
- Next step is to study these methods when including small data sets.

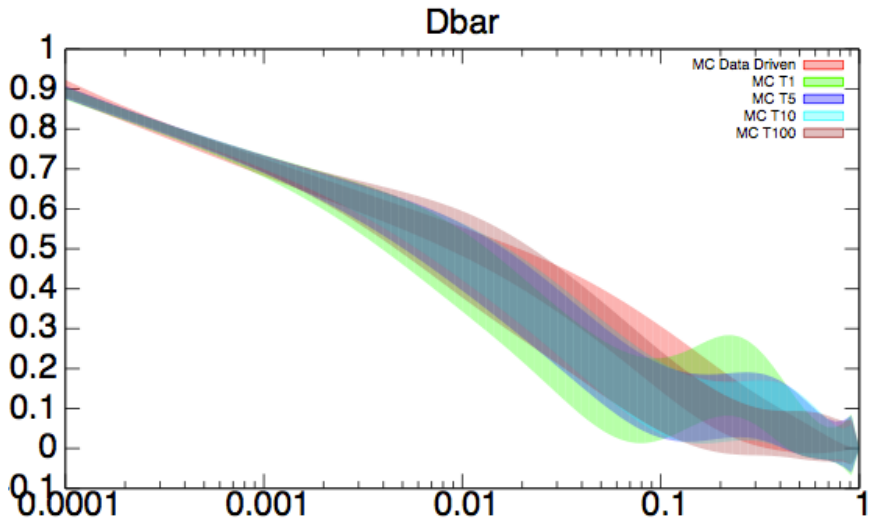
External Regularisation based on penalty term in χ^2

2. A simple χ^2 penalty term for a deviation from a simple PDF parametrisation form:

$$\chi^2_{\text{reg}} = T \sum_f \left(\left(\frac{D_f}{\Delta_D} \right)^2 + \left(\frac{E_f}{\Delta_E} \right)^2 \right)$$

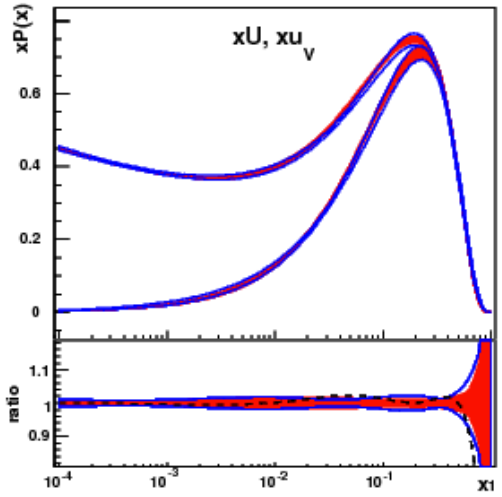
   - with $\Delta D = \Delta E = 100$, such that D, E large approach 1,
   - T the regularisation parameter: T=0, no penalty, T>>0 strong penalty

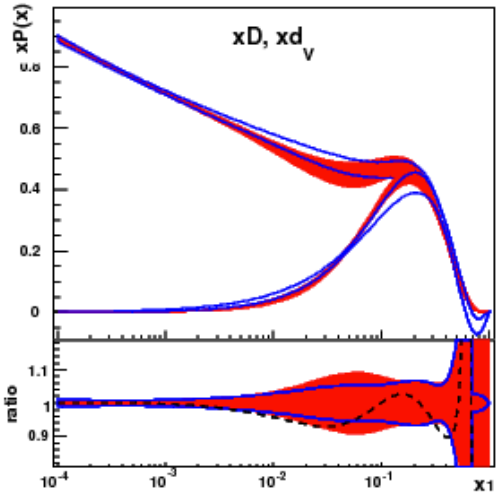


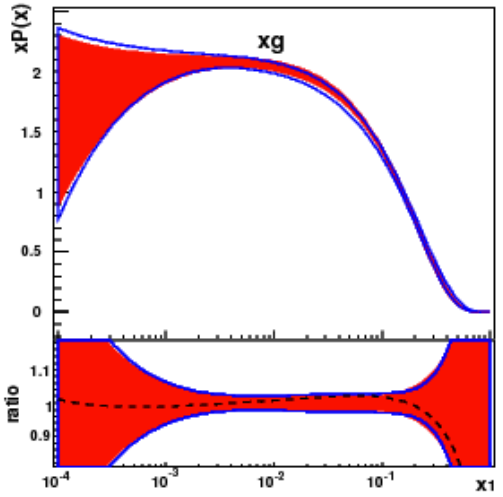


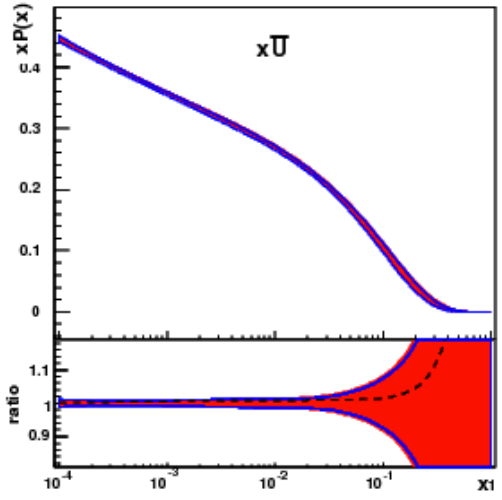
Hessian method: effect of regularisation

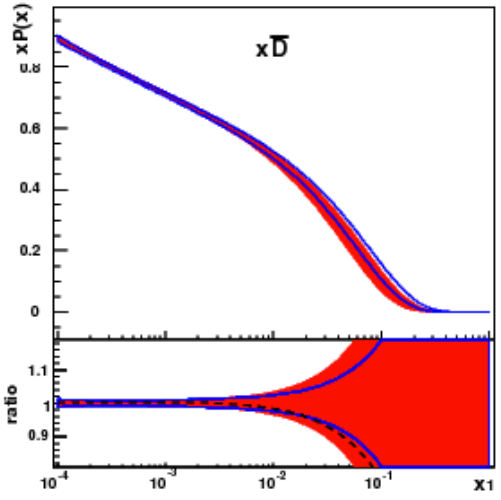
- Here is the effect of the regularisation T=500 vs T=1000











QCDatLHC.hessian.heraonly.T1000.nnlo.fs05/output

QCDatLHC.hessian.heraonly.T500.nnlo.fs05/output

$Q^2 =$ 1.90 GeV^2