Reweighting PDFs

K. Lohwasser (DESY) & A. Guffanti (NBI)

“Proton Structure in the LHC Era - School and Workhop”
DESY, 29/9 - 2/10/2014

Reweighting PDFs

- Including **new data in global PDF fits** is usually a job for a restricted set of people who have access to **complicated, massive and (mostly) private codes**

- This is especially true when considering **hadronic processes**, for which the computations of **higher order corrections** is **computationally very intensive**

- Wouldn’t it be nice to have a method to **quickly assess the impact of new data** on a PDF determination, which is **based only on public tools** and is as fast as producing a theory-experiment comparison plot for the dataset under study?

- **Bayesian reweighting** provides such a method ... so let’s have a closer look!

Bayesian Reweighting

- **First implementation** suggested by Giele and Keller, who thought of it as method for producing new PDF fits

  [W. Giele & S. Keller, hep-ph/9803393]

- **Reformulated in the context of the NNPDF fits** (based on Monte Carlo methodology for uncertainties estimation) and applied for the first time to include data in a global PDF fit (NNPDF2.2)

  [R.D. Ball et al., arXiv:1012.0836]
  [R.D. Ball et al., arXiv:1108.1758]

- Recently extended by G. Watt and R. Thorne to PDF based on the **Hessian method** for estimation of uncertainties

  [G. Watt & R.S. Thorne, arXiv:1205.4024]
  [LHCb, De Lorenzi et. al, arXiv:1011.4260]

Bayesian Reweighting

- The replicas of a Monte Carlo PDF set provide a sampling of the probability density in the space of Parton Distribution Functions

- Expectation values for observables which depend on PDFs are obtained by taking the average for a given observable over the replica set

$$\langle \mathcal{F}[f_i(x,Q^2)] \rangle = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \mathcal{F}\left(f_i^{(net)(k)}(x,Q^2)\right)$$

  ... with corresponding expressions for variances, correlations, etc.

- The central idea of Bayesian reweighting is to assess the impact of including new data in a PDF determination by updating the probability density of PDFs without performing a complete refit

Bayesian Reweighting

- We can apply Bayes Theorem to determine the conditional probability of the PDF upon inclusion of the new data

$$\mathcal{P}_{\rm new}(\{f\}) = \mathcal{N}_\chi \mathcal{P}(\chi^2|\{f\})\mathcal{P}_{\rm init}(\{f\}), \quad \mathcal{P}(\chi^2|\{f\}) = [\chi^2(y,\{f\})]^{\frac{n_{dat}-1}{2}} e^{-\frac{\chi^2(y,\{f\})}{2}}$$

- Averages over the sample are no weighted sums

$$\langle \mathcal{F}[f_i(x,Q^2)]\rangle = \sum_{k=1}^{N_{rep}} w_k \mathcal{F}\left(f_i^{(net)(k)}(x,Q^2)\right)$$

… and the weights are given by

$$w_k = \frac{[\chi^2(y,f_k)]^{\frac{n_{dat}-1}{2}} e^{-\frac{\chi^2(y,f_k)}{2}}}{\sum_{i=1}^{N_{\rm rep}} [\chi^2(y,f_i)]^{\frac{n_{dat}-1}{2}} e^{-\frac{\chi^2(y,f_i)}{2}}}$$

Bayesian Reweighting

- The original sample of replica was constructed through importance sampling of the old probability distribution and it is thus maximally efficient (i.e. all replicas are equiprobabile and it gives the best representation of the probability density for a given number replicas)

- **After reweighting** the new replicas set will **not** give anymore a **maximally efficient** representation of the new probability distribution

- This **loss of efficiency** can be quantified using the **Shannon Entropy**

$$N_{\rm eff} \equiv \exp\left\{\frac{1}{N}\sum_{k=1}^{N} w_k \ln(N/w_k)\right\}$$

to estimate the ***effective number of replicas*** **after reweighting**

- The **smaller** the **Shannon entropy** the **more constraining** the new **data** are

Bayesian Reweighting

- If the value of the Shannon entropy obtained after reweighting a prior set with a given dataset becomes too small the reweighting procedure becomes unreliable

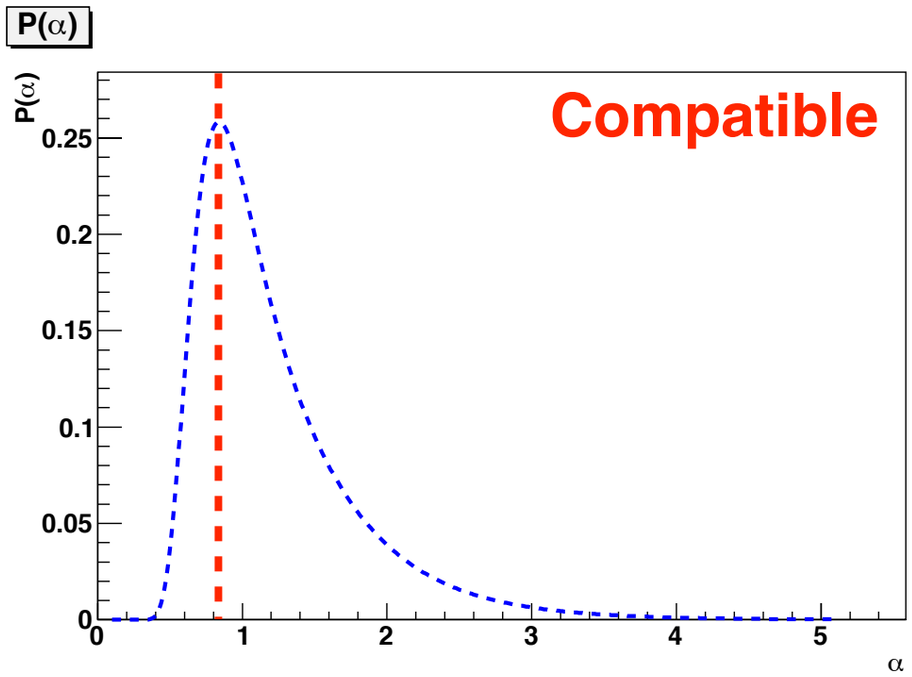
- There are two reasons why that can happen
  - the new data contain a lot of information on PDFs not present in the prior fit (⇒ refit)
  - the new data are incompatible with data included in the prior PDF set

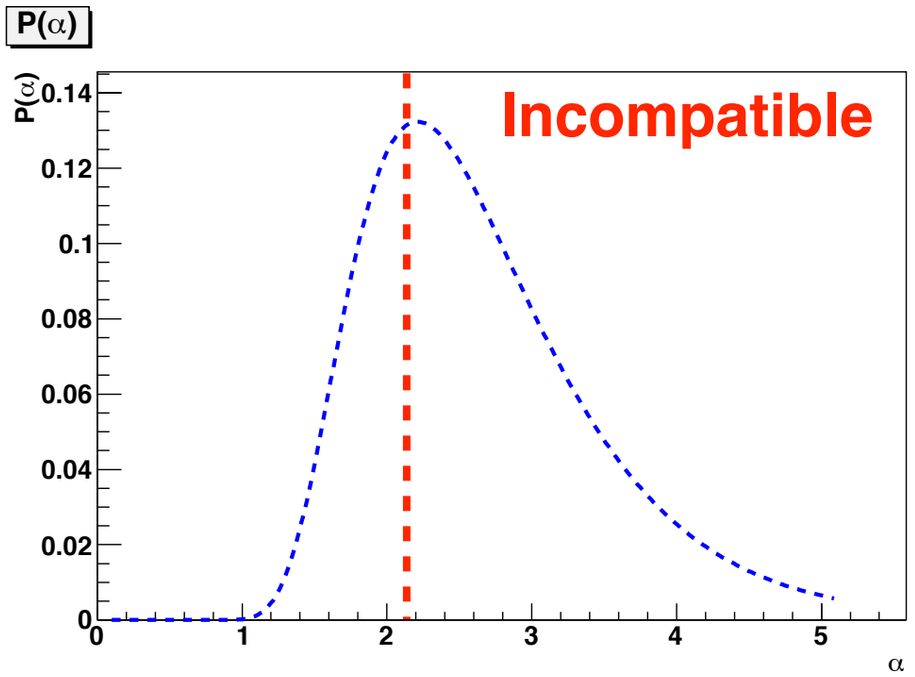
- We can distinguish the two cases by looking at the probability density of the nuisance parameter (α), defined as a rescaling factor for the uncertainties on the new data

$$\mathcal{P}(\alpha) \propto \frac{1}{\alpha} \sum_{k=1}^{N} w_k(\alpha)$$

Bayesian Reweighting

- If the **probability density peaks** close to **one** (or below one) the **new data** are **compatible** with the data included in the prior fit
- If the **probability density peaks far above one** the **uncertainties** on the new data are **probably underestimated** and these **data** are thus **incompatible** with the data included in the prior fit





Unweighting PDFs

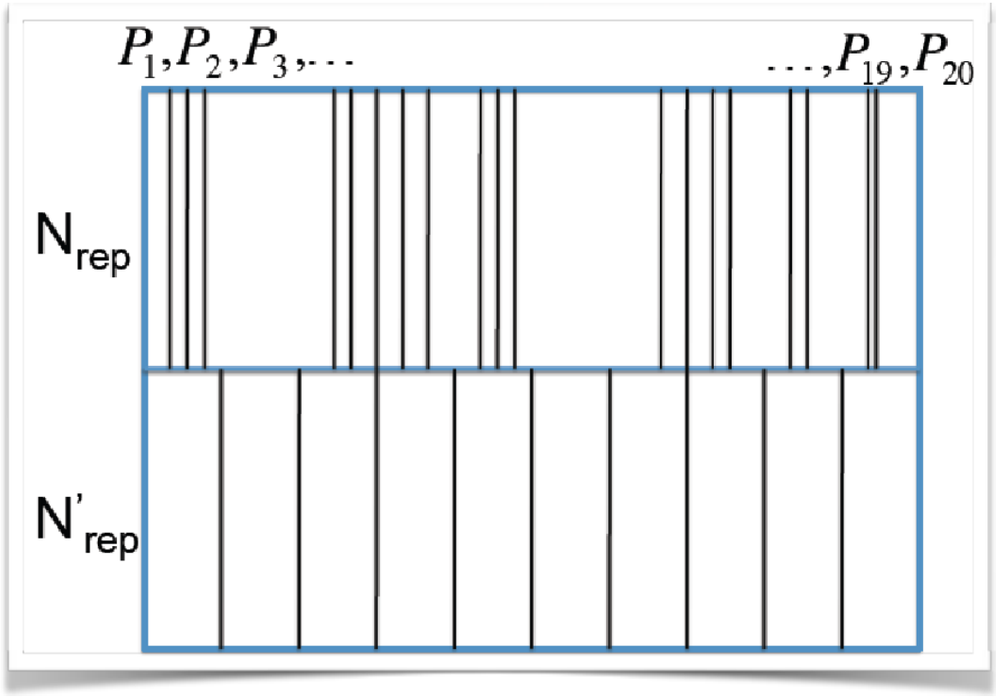
- We want to **define a procedure to generate an unweighted PDF set** (*i.e.* a set in which all replicas are equiprobable) starting from one we constructed via reweighting

- To do that we define the probability cumulants of the original fit as

$$P_k \equiv P_{k-1} + p_k = \sum_{j=0}^{k} p_j$$

- The weights in the new, unweighted, set are the defined as

$$w'_k = \sum_{j=1}^{N'_{\rm rep}} \theta\Big(\frac{j}{N'_{\rm rep}} - P_{k-1}\Big)\theta\Big(P_k - \frac{j}{N'_{\rm rep}}\Big)$$



This is equivalent to the graphical procedure of picking a number of replicas from the original reweighted set which is equal to the number of lower segments whose right edge is contained in the upper segment corresponding to the specific replica

Validating Bayesian reweighting

- We now **validate the reweighting** procedure checking that the methodology yields results which satisfy a number of consistency tests

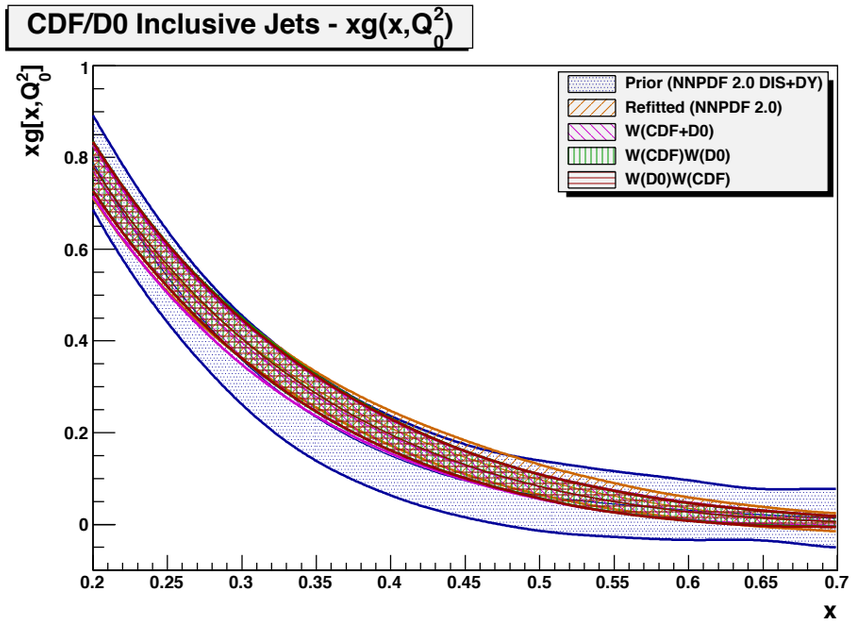
- Including a given dataset in a prior fit by **reweighting** or **refitting** should **yield statistically equivalent results**

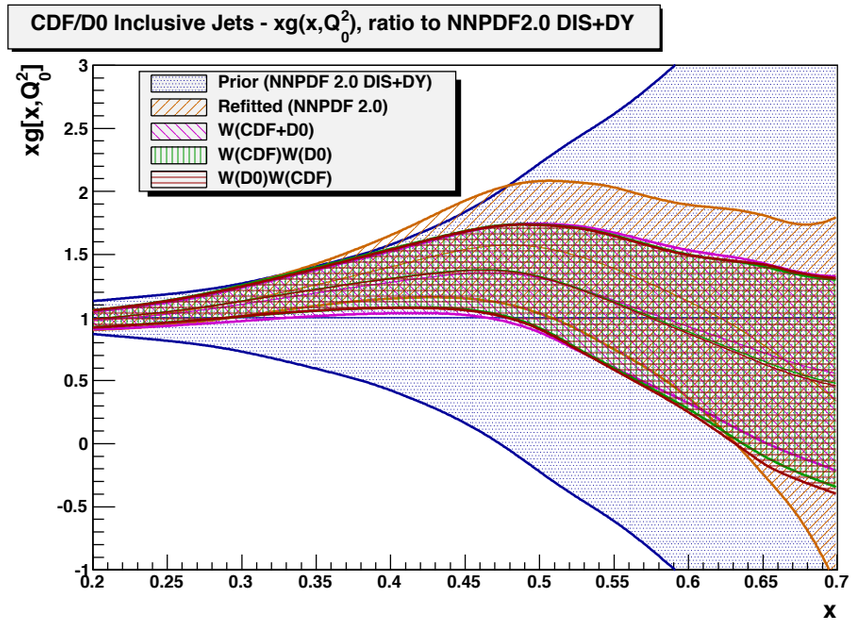
- If we **include two or more datasets** we can choose to include them in a **single step** (as a single dataset) or in **successive steos**: the two procedures should yield **statistically equivalent results**

- When i**ncluding sets in successive steps** results should **not depend on the order** in which the reweighting is performed

Validating Bayesian reweighting

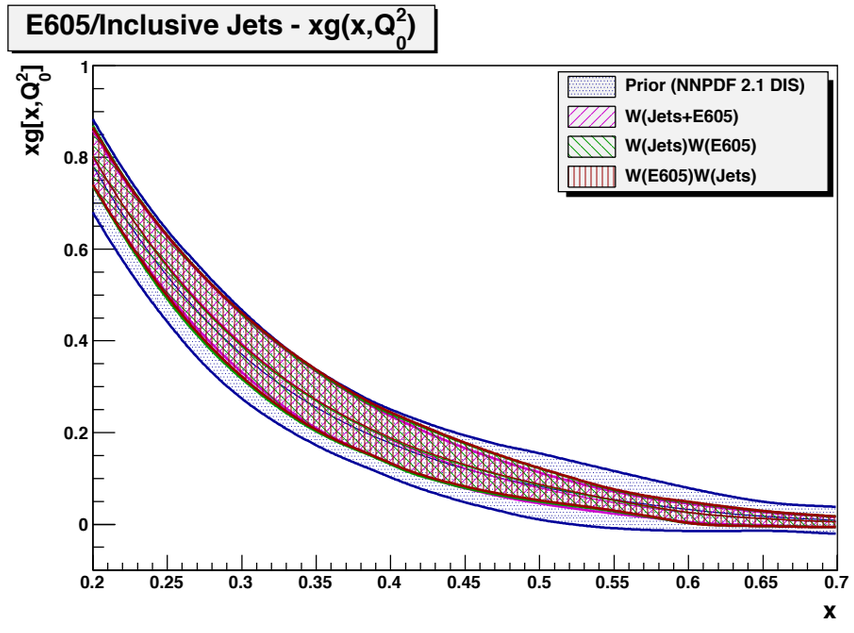
- Start from NNPDF2.0 DIS+DY fit as a prior fit
- Add CDF and D0 inclusive jet data through refitting (NNPDF2.0)
- Add CDF & D0 jet data via reweighting as a single dataset
- Add CDF data then unweight and then add D0 data
- Add D0 data then unweight and then add CDF data

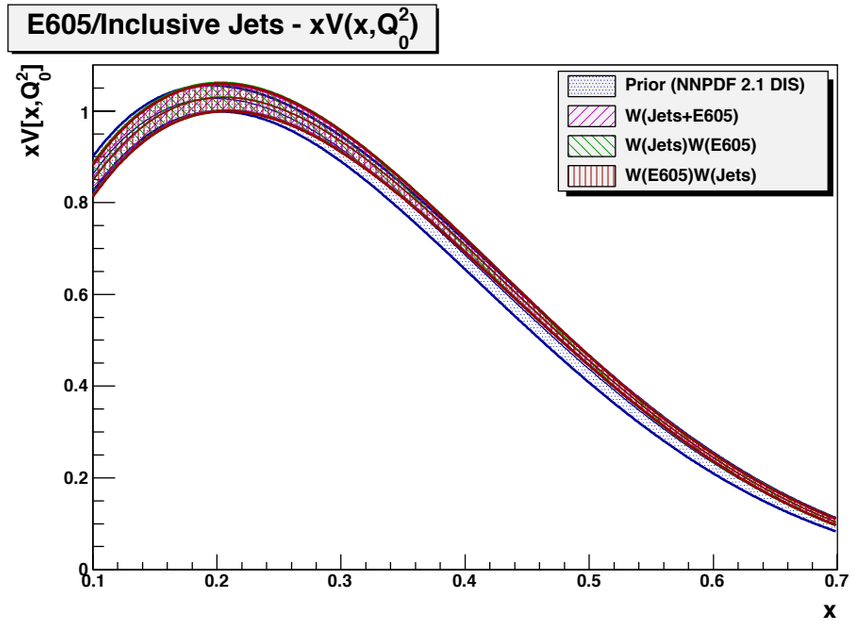




Validating Bayesian reweighting

- Start from NNPDF2.1 DIS fit as a prior fit
- Add E605 Drell-Yan & Tevatron jet data via reweighting as a single dataset
- Add Tevatron jet data then unweight and then add E605 Drell-Yan data
- Add E605 Drell-Yan data, unweight and then add Tevatron jet data
- All procedures yield statistically equivalent results





Reweighting for Hessian sets

[G. Watt & R.S. Thorne, arXiv:1205.4024]

- For **Hessian sets** (assuming symmetric uncertainties) the uncertainties on a given observables can be computed as

$$\Delta F = \frac{1}{2}\sqrt{\sum_{k=1}^{n}\left[F(S_k^+) - F(S_k^-)\right]^2}$$

- A set of **Monte Carlo replicas** can then be **generated according** to

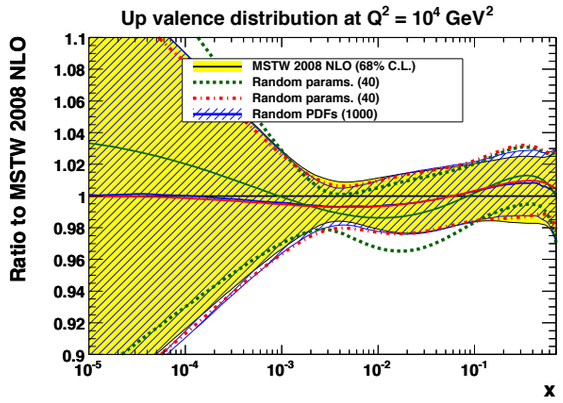
$$F(\mathcal{S}_k) = F(S_0) + \frac{1}{2}\sum_{j=1}^{n}\left|F(S_j^+) - F(S_j^-)\right| R_{jk} \qquad (k = 1, \ldots, N_{\text{pdf}})$$

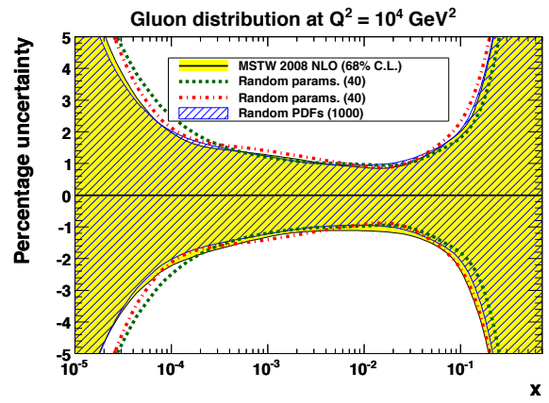
where S_0 is the "central set" and R_{jk} are normally distributed random numbers

Reweighting for Hessian sets

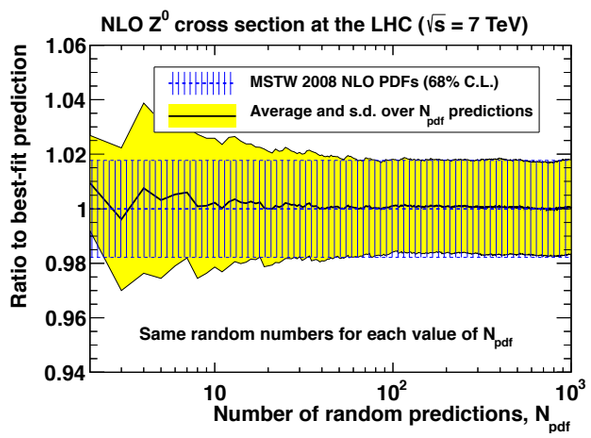
[G. Watt & R.S. Thorne, arXiv:1205.4024]

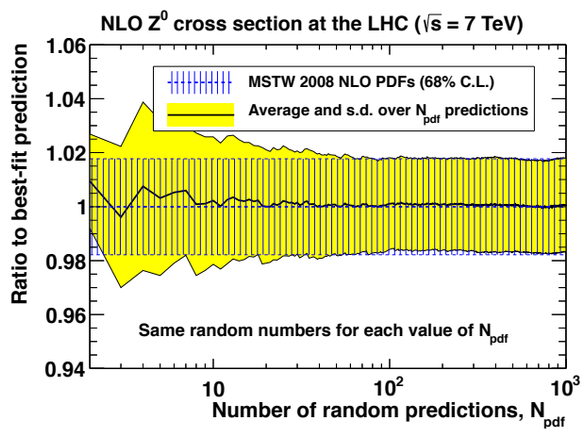
- How well does the Monte Carlo PDF ensemble we generated reproduce the original probability distribution of PDFs given by the Hessian eigenvectors?





... and what about observables?



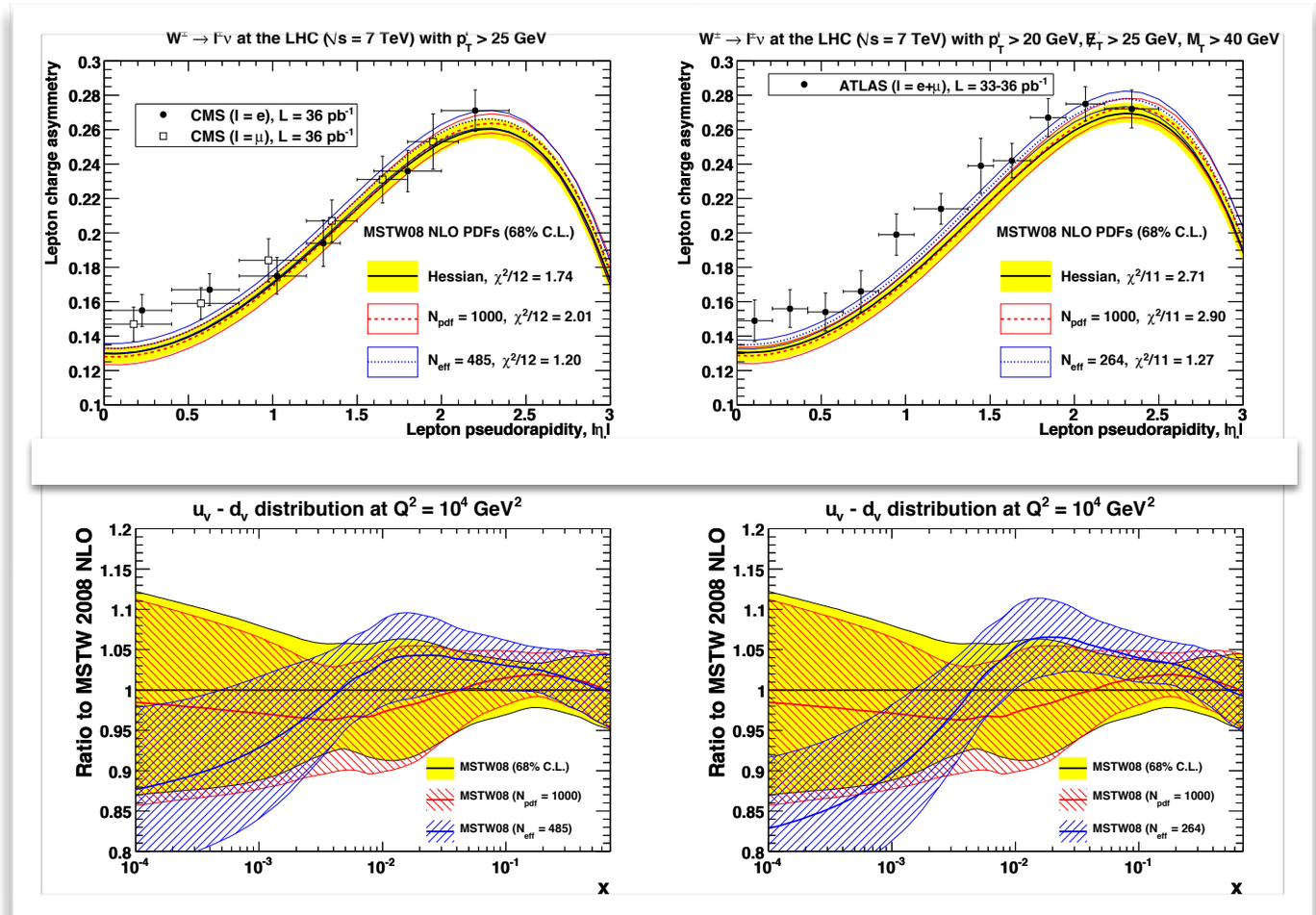


Reweighting for Hessian sets

[G. Watt & R.S. Thorne, arXiv:1205.4024]

- Once replicas are generated according to the recipe given, the very same Bayesian reweighting methodology described for NNPDF can be used to compute weights and generate a reweighted PDF set including new data

- Used to assess the impact of LHC W lepton asymm. on MSTW08 PDFs
- Valence quark distribution mostly affected in the medium-/small-x region



Conclusions & Outlook

- **Inclusion of new data** in global PDF fits is usually performed by PDF fitting collaborations using **complicated**, **massive** and (mostly) **private codes**

- **Bayesian reweighting** provides **a method to quickly assess the impact of new data** on a PDF determination and is based on the use of public tools

- Bayesian reweighting was **initially developed for Monte Carlo sets** but the same techniques have recently been **extended to** use with **Hessian sets**

- Available as a module of the HERAfitter package … so let us play around with it a bit!