NNPDF1.0: parton distribution functions with faithful errors

Luigi Del Debbio

luigi.del.debbio@ed.ac.uk

University of Edinburgh

arXiv:0808.1231 [hep-ph], Nucl. Phys. B (in press)

The NNPDF Collaboration

R.D.Ball[1], L.Del Debbio[1], S.Forte[2], A.Guffanti[3], J.I.Latorre[4], A. Piccione[2], J. Rojo[2], M.U.[1]

[1] PPT Group, School of Physics, University of Edinburgh

[2] Dipartimento di Fisica, Università di Milano

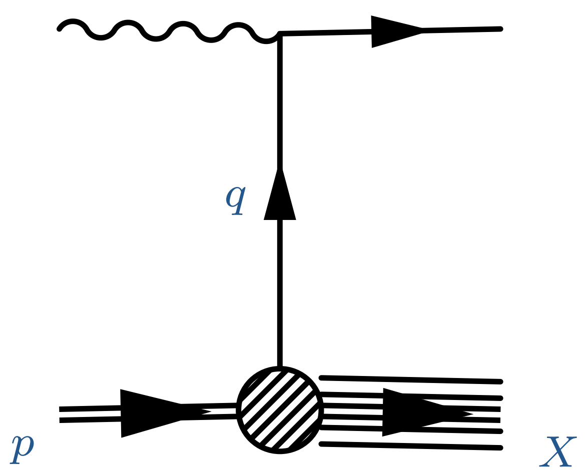
[3] Physikalisches Institut, Albert-Ludwigs-Universität Freiburg

[4] Departament d'Estructura i Constituents de la Matèria, Universitat de Barcelona

DIS Parton distribution functions

Deep inelastic observables:

$$F_I(x,Q^2) = \sum_j C_{Ij}(x,\alpha_s(Q^2)) \otimes f_j(x,Q^2)$$



Scale–dependence of PDFs:

$$Q^2 \frac{\partial}{\partial Q^2} f_i(x,Q^2) = \sum_j P_{ij}(x,\alpha_s(Q^2)) \otimes f_j(x,Q^2)$$

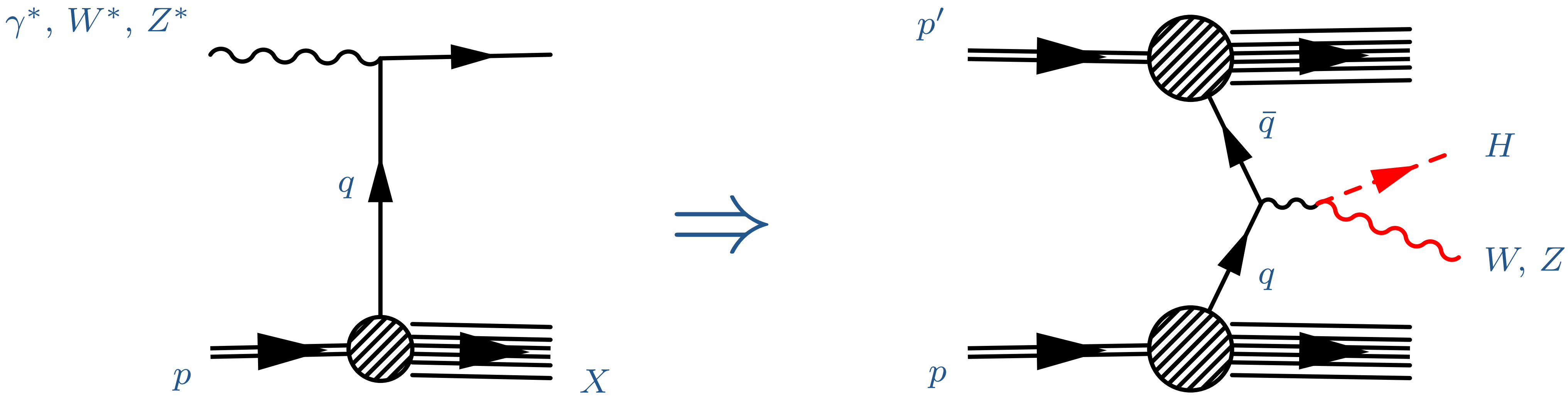
$$f_i(x,Q^2) = \sum_j \Gamma_{ij}(x,\alpha_s,\alpha_s^0) \otimes f_j(x,Q_0^2)$$

Back to the observables:

$$F_I(x,Q^2) = \sum_{jk} C_{Ij}(x,\alpha_s) \otimes \Gamma_{jk}(x,\alpha_s,\alpha_s^0) \otimes f_k(x,Q_0^2)$$

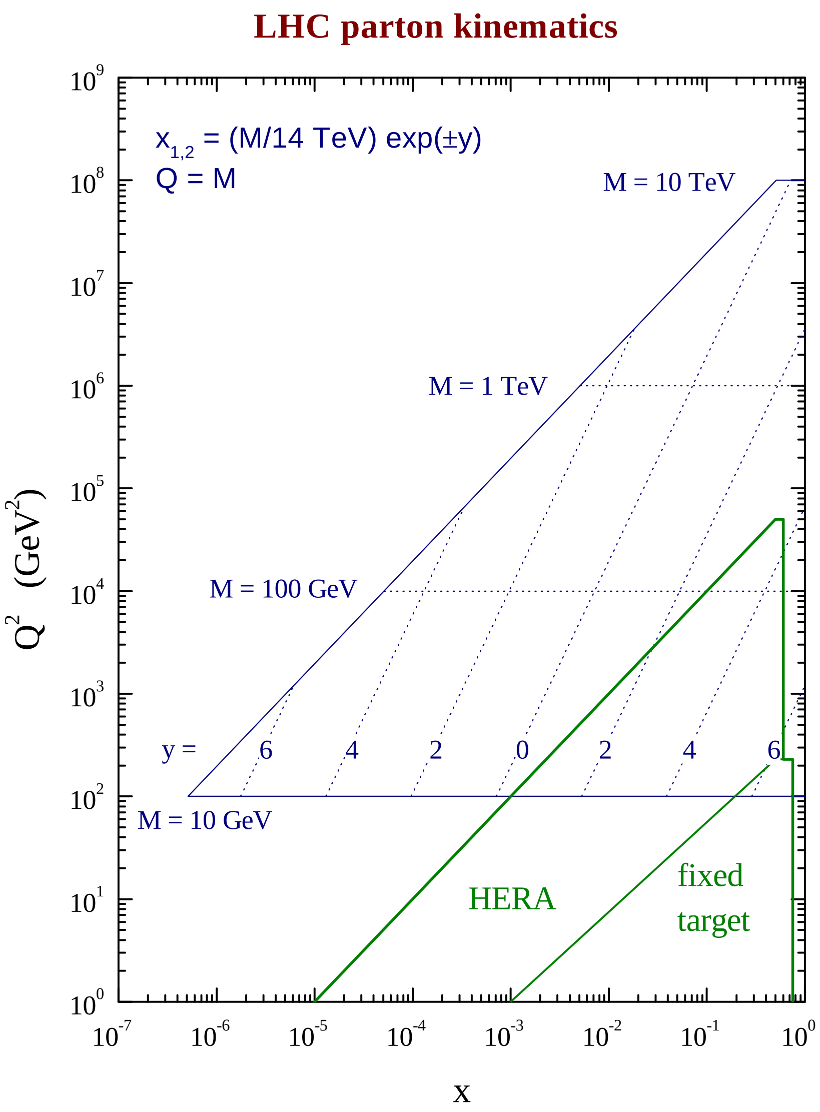
$$= \sum_j K_{Ij}(x,\alpha_s,\alpha_s^0) \otimes f_j(x,Q_0^2)$$

Parton distributions for LHC



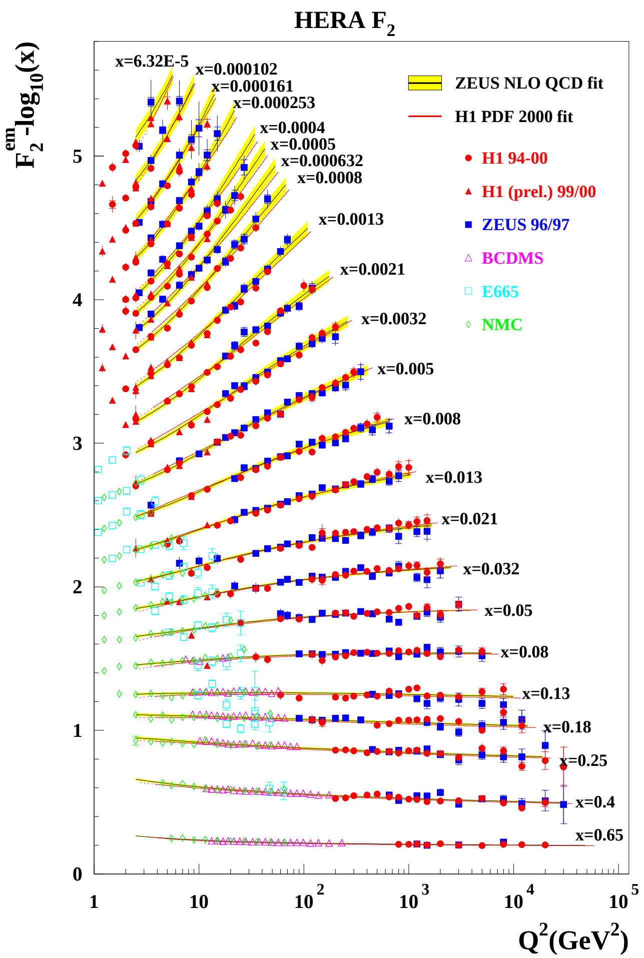
- different kinematics
- nonperturbative nucleon structure described by the *same* PDFs
- evolved to the relevant scales

LHC kinematics



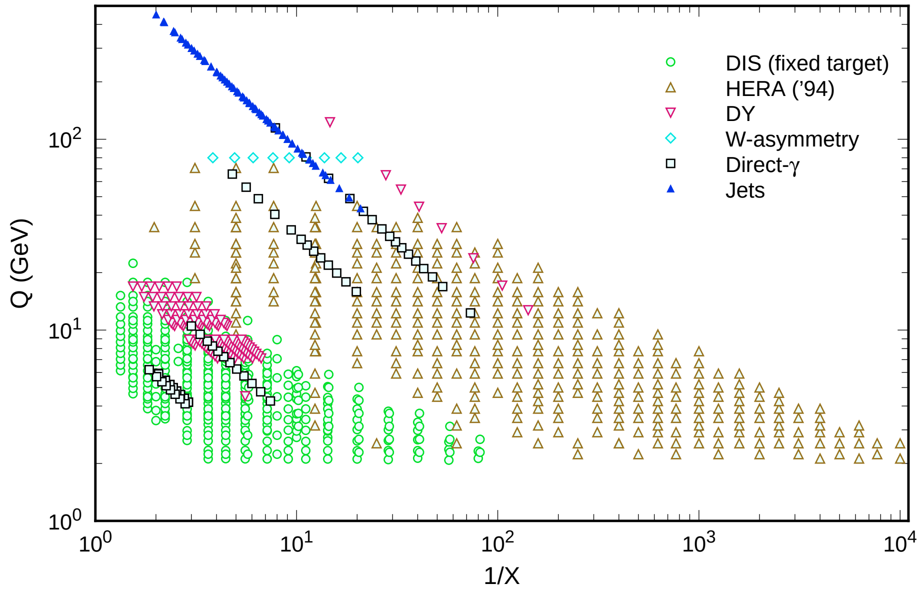
- PDFs need to be extrapolated
- uncertainty in the extrapolation region
- uncertainty propagates into LHC physical observables
- important for modelling the QCD background at LHC
- necessity to develop PDFs with faithful errors

Partons with errors



given a set of data points, determine a set of functions with errors

data included in CTEQ5 parton fit



What's the problem? [Kosower 99]

- for a single quantity, we quote 1 sigma errors: value $\pm$ error
- for a pair of numbers, we quote a 1 sigma ellipse
- for a function, we need an "error bar" in a space of functions

we must determine the probability density (measure) $\mathcal{P}[f_i(x)]$ in the space of parton distribution functions $f_i(x)$ (i=quark, antiquark, gluon)

EXPECTATION VALUE OF $\mathcal{F}\left[f_i(x)\right] \Rightarrow$ FUNCTIONAL INTEGRAL

$$\left\langle \mathcal{F}\left[f_i(x)\right] \right\rangle = \int \mathcal{D}f_i \, \mathcal{F}\left[f_i(x)\right] \, \mathcal{P}[f_i],$$

we must extract from the data a description of the probability distribution $\mathcal{P}$

The standard solution

- choose a parameterization at a reference scale
- evolve to desired scale & compute physical observables
- determine best-fit values of parameters
- determine error by propagation of error on parameters (’hessian method’) or by parameter scans (’lagrange multiplier method’)

> problem projected onto the finite–dimensional space of parameters

Comparing global fits

W production cross-section Tevatron

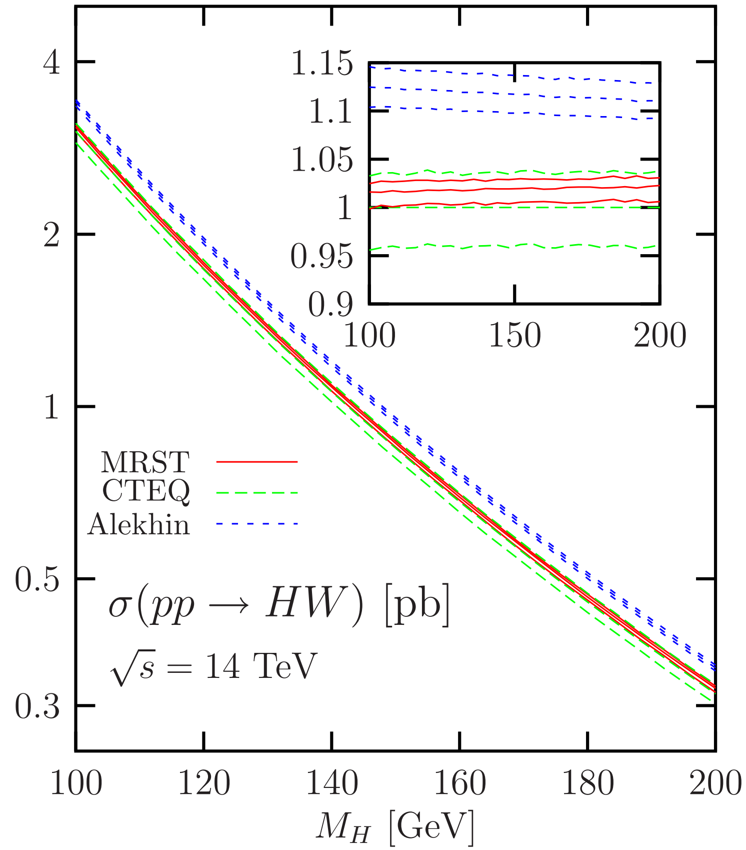
PDF set	xsec [nb]	PDF uncertainty
Alekhin	2.73	$\pm$ 0.05
MRST2002	2.59	$\pm$ 0.03
CTEQ6	2.54	$\pm$ 0.10

[Thorne 03]

Alekhin vs. MRST/CTEQ $\rightarrow$ W production xsect at tevatron do not agree within respective errors

Alekhin vs. MRST/CTEQ $\rightarrow$ predictions for associate Higgs W production LHC do not agree within respective errors

Higgs production at LHC

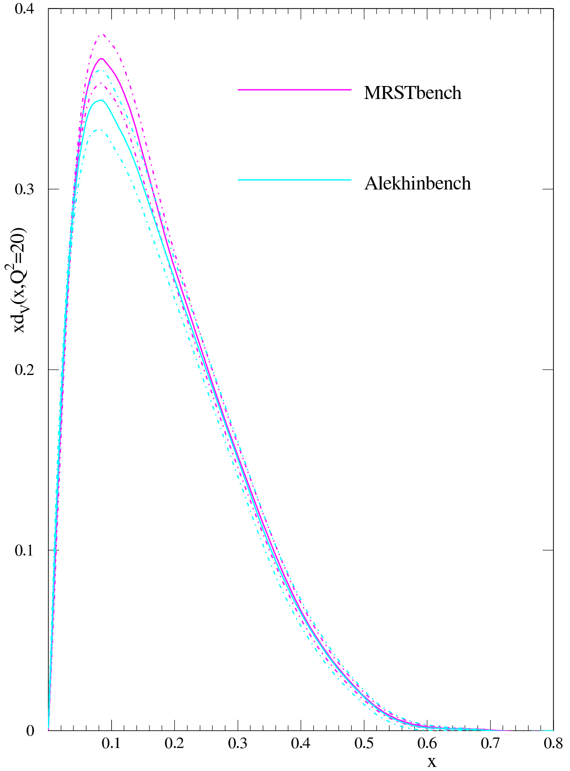


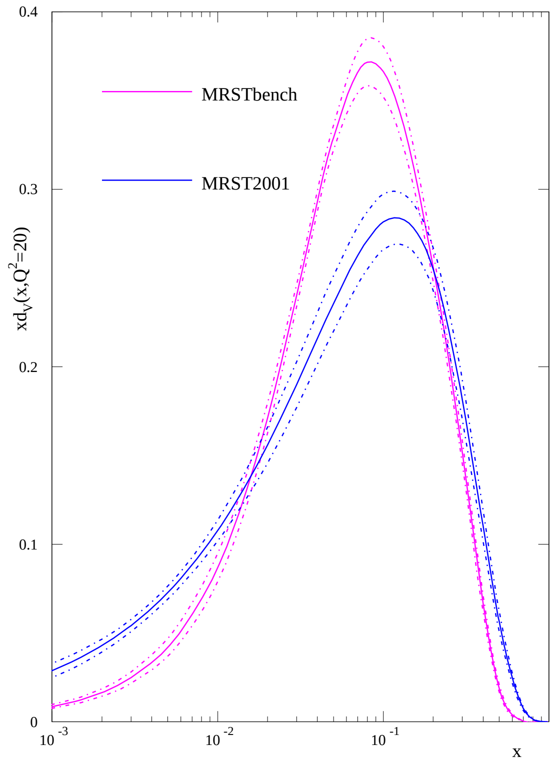
[Djouadi and Ferrag 04]

Troubles with error bars

PDF4LHC workshop at CERN 08

- benchmark fits on reduced sets do not agree with global fits **within errors**
- incompatible experiments?
- lack of generality in the parametrization?
- tolerance criterion $\Delta\chi^2 > 1$?



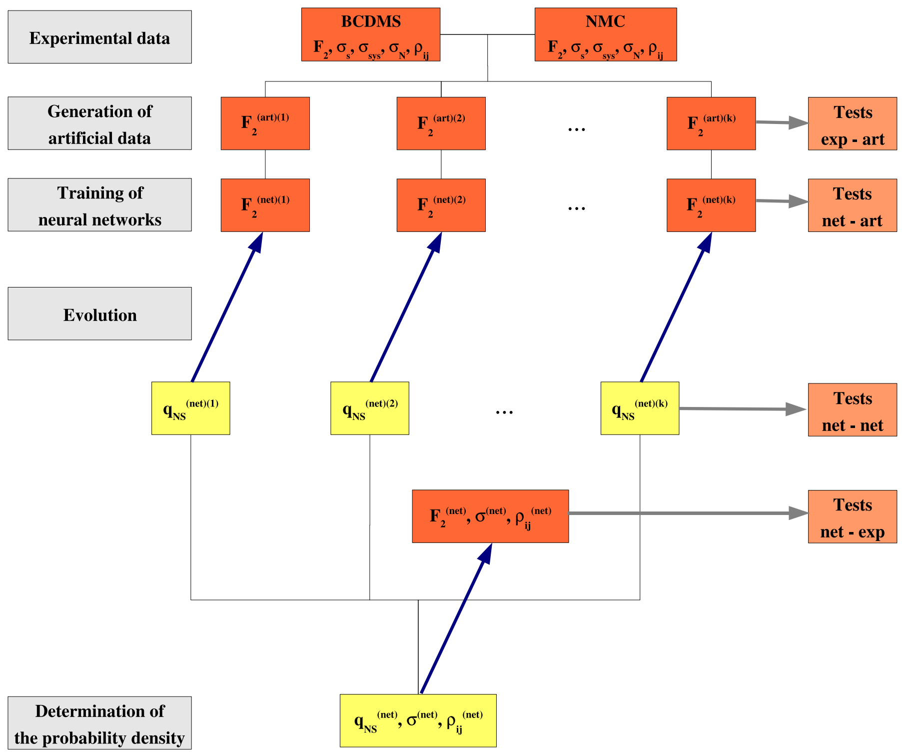


The Neural Monte Carlo

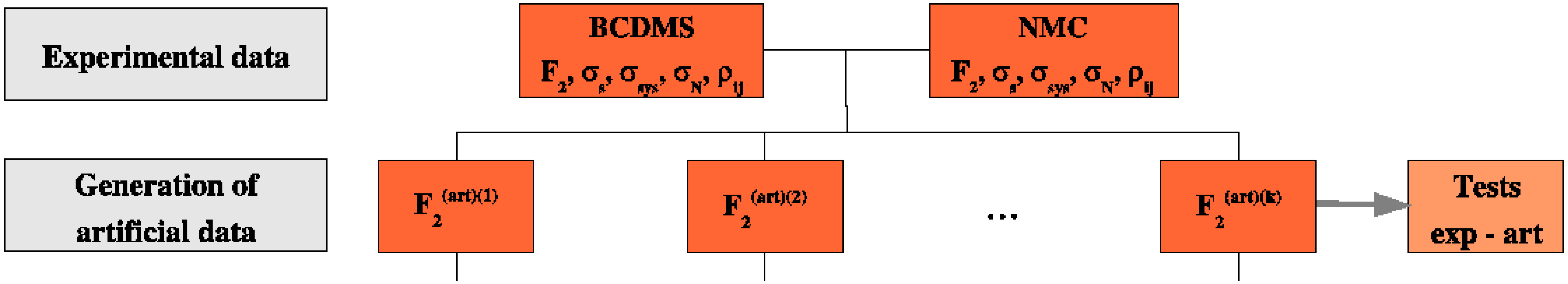
NNPDF collaboration (2004: Idd, Forte, Latorre, Piccione, Rojo; 2007: + Ball, Guffanti, Ubiali)

- Monte Carlo replicas $F_I^{(k)}(p_i)$ of the original dataset $F_I^{(\mathrm{data})}(p_i)$
  $\Rightarrow$ representation of $\mathcal{P}[F_I(p_i)]$ at discrete set of points p_i
- train a neural net for each pdf on each replica, $\longrightarrow$ neural representation of the pdfs $f_i^{(net),(k)}$
- The set of neural nets is a representation of the probability density:

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



MC sampling of exp data



$$F_{I,p}^{(\mathrm{art})(k)} = S_{p,N}^{(k)} F_{I,p}^{(\mathrm{exp})} \left(1 + \sum_{l=1}^{N_c} r_{p,l}^{(k)} \sigma_{p,l} + r_p^{(k)} \sigma_{p,s} \right) \ , \ k = 1, \ldots, N_{\mathrm{rep}} \ ,$$

where

$$S_{p,N}^{(k)} = \prod_{n=1}^{N_a} \left(1 + r_{p,n}^{(k)} \sigma_{p,n} \right) \prod_{n=1}^{N_r} \sqrt{1 + r_{p,n}^{(k)} \sigma_{p,n}} .$$

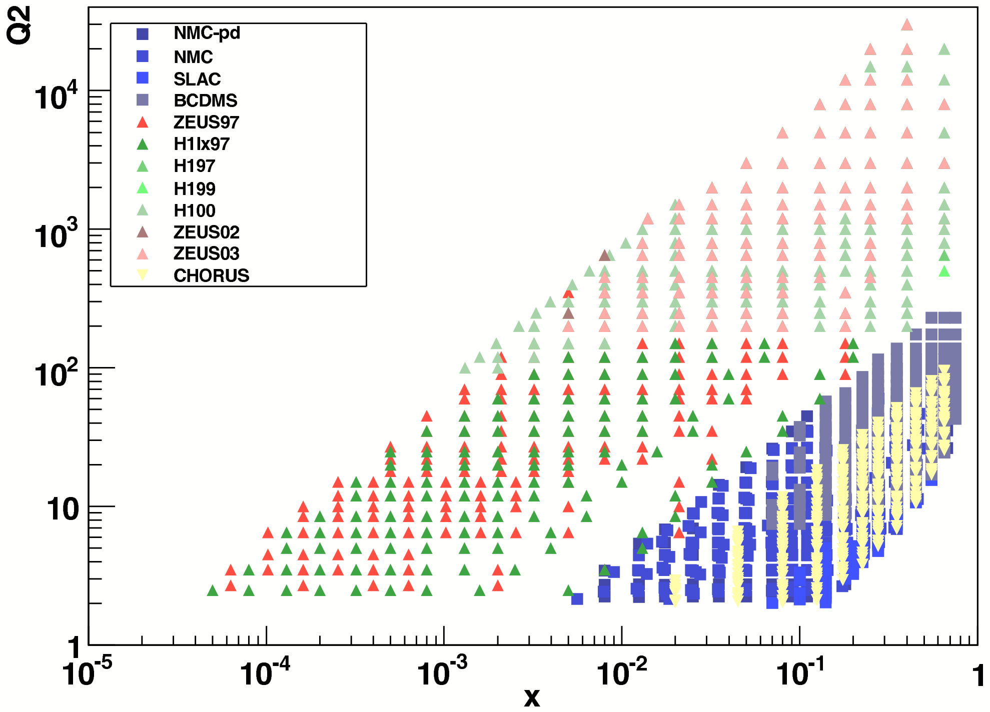
Monte Carlo errors

- for each replica (k) we fit one set of PDFs
- the ensemble of fitted replicas represents the probability distribution in the space of PDFs
- statistical properties of any function of the PDFs can be computed using standard methods:

$$
\begin{aligned}
\langle \mathcal{F}[f(x)] \rangle &= \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} \mathcal{F}[f^{(k)(\rm net)}(x)] \\
\sigma_{\mathcal{F}[f(x)]} &= \sqrt{\langle \mathcal{F}[f(x)]^2 \rangle - \langle \mathcal{F}[f(x)] \rangle^2} \\
\rho[f_a(x_1,Q_1^2), f_b(x_2,Q_2^2)] &= \frac{\langle f_a(x_1,Q_1^2) f_b(x_2,Q_2^2) \rangle - \langle f_a(x_1,Q_1^2) \rangle \langle f_b(x_2,Q_2^2) \rangle}{\sigma_a(x_1,Q_1^2)\sigma_b(x_2,Q_2^2)}
\end{aligned}
$$

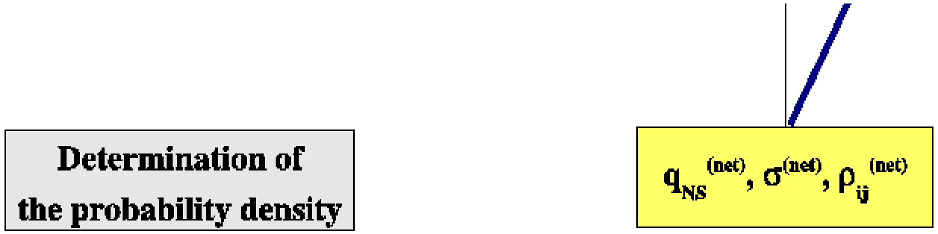
Experimental data

OBS	Data set	OBS	Data set
F_2^p	NMC	σ_{NC}^{-}	ZEUS
	SLAC		H1
	BCDMS	σ_{CC}^{+}	ZEUS
F_2^d	SLAC		H1
	BCDMS	σ_{CC}^{-}	ZEUS
σ_{NC}^{+}	ZEUS		H1
	H1	$\sigma_{\nu}, \sigma_{\bar{\nu}}$	CHORUS
F_2^d/F_2^p	NMC-pd	F_L	H1



- Kinematical cuts:
  $Q^2 >$2 GeV^2
  $W^2 = Q^2(1-x)/x >$12.5 GeV^2
- $\sim$ **3000** points.

Neural Networks?



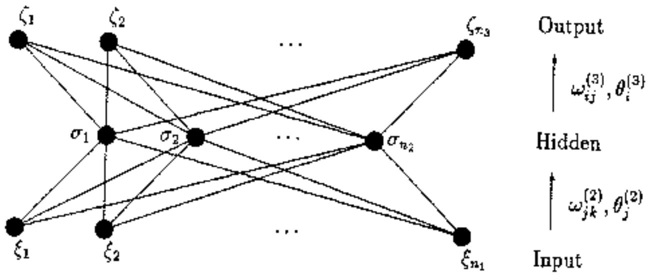
each PDF at the reference scale is parametrised by one NN:

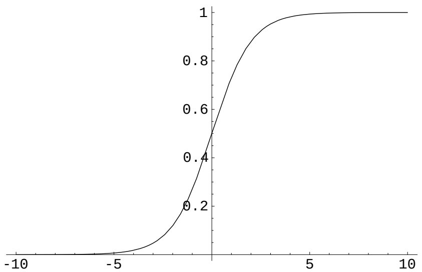
$$\boxed{f_i(x, Q_0^2) = N_i(x)}$$

NN is a non–linear mapping:

$$N_i : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

Some details about Neural Networks





Multilayer feed-forward networks

- Each neuron receives input from neurons in preceding layer and feeds output to neurons in subsequent layer
- Activation determined by weights and thresholds
  $\xi_i = g\left(\sum_j \omega_{ij}\xi_j - \theta_i\right)$
- Sigmoid activation function
  $g(x) = \frac{1}{1+e^{-\beta x}}$

... just another set of basis functions!

eg, a 1-2-1 NN: $\xi_1^{(3)}(\xi_1^{(1)}) = \frac{1}{1+\exp[\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1+e^{\theta_1^{(2)}-\xi_1^{(1)}\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1+e^{\theta_2^{(2)}-\xi_1^{(1)}\omega_{21}^{(1)}}}]}$

Thm: any function can be represented by a sufficiently big neural network

Basis set

- Each independent PDF at the initial scale $Q_0^2 = 2\mathrm{GeV}^2$ is parameterized by an individual NN.
- Little constraint on strange $\longrightarrow$ Flavor Assumptions:
  - Symmetric strange sea $s(x) = \bar{s}(x)$
  - Strange sea proportional to non-strange sea $\bar{s}(x) = \frac{C}{2}(\bar{u}(x) + \bar{d}(x))$ (C = 0.5)
  - Intrinsic heavy quarks contributions neglected.
- Parametrization of **(4+1)** combinations of PDFs at $Q_0^2 = 2$ GeV^2:

Singlet : $\Sigma(x)$	$\longmapsto \mathrm{NN}_\Sigma(x)$	2-5-3-1 37 pars
Gluon : $g(x)$	$\longmapsto \mathrm{NN}_g(x)$	2-5-3-1 37 pars
Total valence : $V(x) \equiv u_V(x) + d_V(x)$	$\longmapsto \mathrm{NN}_V(x)$	2-5-3-1 37 pars
Non-singlet triplet : $T_3(x)$	$\longmapsto \mathrm{NN}_{T3}(x)$	2-5-3-1 37 pars
Sea asymmetry : $\Delta_S(x) \equiv \bar{d}(x) - \bar{u}(x)$	$\longmapsto \mathrm{NN}_\Delta(x)$	2-5-3-1 37 pars

185 parameters

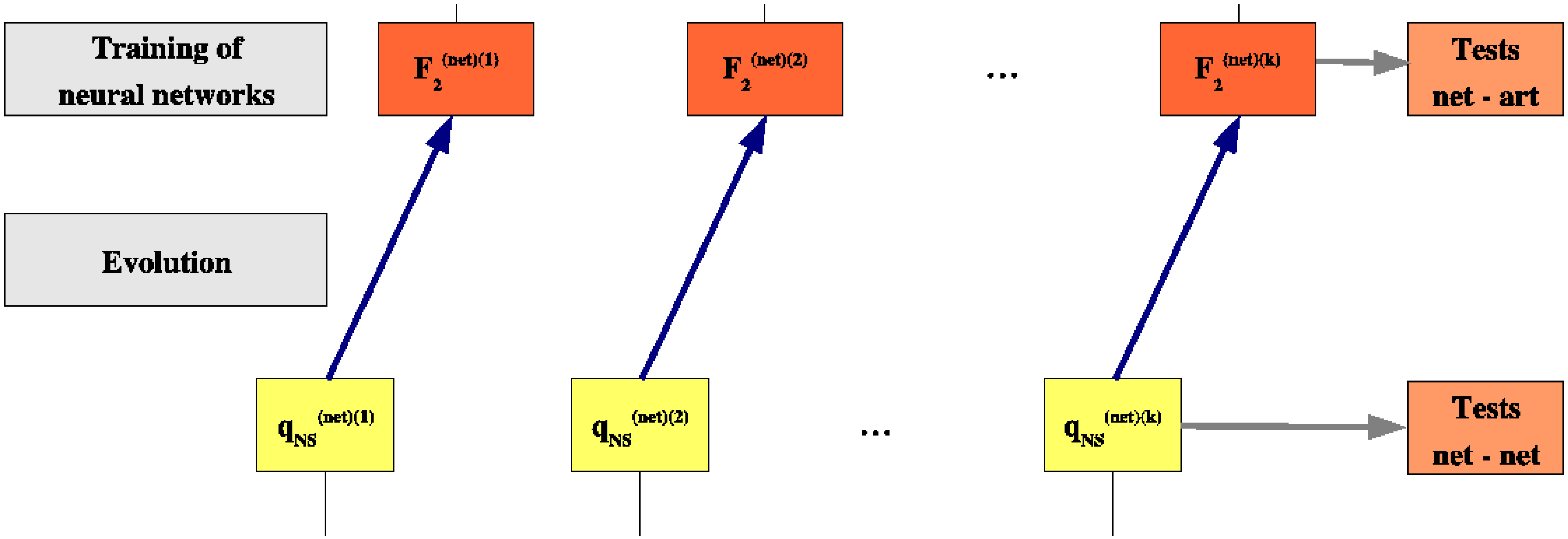
Normalization and sum rules

$$
\begin{aligned}
\Sigma(x,Q_0^2) &= (1-x)^{m_\Sigma} x^{-n_\Sigma} \mathrm{NN}_\Sigma(x) \,, \\
V(x,Q_0^2) &= A_V (1-x)^{m_V} x^{-n_V} \mathrm{NN}_V(x) \,, \\
T_3(x,Q_0^2) &= (1-x)^{m_{T_3}} x^{-n_{T_3}} \mathrm{NN}_{T_3}(x) \,, \\
\Delta_S(x,Q_0^2) &= A_{\Delta_S} (1-x)^{m_{\Delta_S}} x^{-n_{\Delta_S}} \mathrm{NN}_{\Delta_S}(x) \,, \\
g(x,Q_0^2) &= A_g (1-x)^{m_g} x^{-n_g} \mathrm{NN}_g(x) \,.
\end{aligned}
$$

- Polynomial Preprocessing $\longrightarrow$ Training Efficiency
- Normalization $\longrightarrow$ Fixed by valence and momentum sum rules

$$
\begin{aligned}
\int_0^1 dx\, x\left(\Sigma(x)+g(x)\right) &= 1 \\
\int_0^1 dx\, \left(u(x)-\bar{u}(x)\right) &= 2 \\
\int_0^1 dx\, \left(d(x)-\bar{d}(x)\right) &= 1 \,.
\end{aligned}
$$

Evolution



Observables are a convolution over x of PDFs and Coefficient Functions:

$$F_I(x,Q^2) = \sum_j C_{Ij}(x,\alpha_s) \otimes f_j(x,Q^2) = \sum_{j,k} C_{Ij}(x,\alpha_s) \otimes \Gamma_{jk}(x,\alpha_s,\alpha_s^0) \otimes f_k(x,Q_0^2)$$

Kernels for a physical observable

F_2 proton structure function

$$F_2^p = x\{\frac{5}{18}C^s_{2,q}\otimes\Sigma + \frac{1}{6}C_{2,q}\otimes(T_3 + \frac{1}{3}(T_8 - T_{15}) + \frac{1}{5}(T_{24} - T_{35})) + \langle e_q^2\rangle C_{2,g}\otimes g\}$$

$$F_2^p = x\{K_{\mathrm{F2},\Sigma}\otimes\Sigma_0 + K_{\mathrm{F2},g}\otimes g_0 + K_{\mathrm{F2},+}\otimes\left(T_{3,0} + \frac{1}{3}(T_{8,0} - T_{15,0})\right)\}$$

In Mellin space

$$K_{\mathrm{F2},\Sigma} = \frac{5}{18}C^s_{2,q}\Gamma_{\mathrm{S}}^{qq} + \frac{1}{30}C_{2,q}(\Gamma_{\mathrm{S}}^{24,q} - \Gamma_{\mathrm{S}}^{35,q}) + \langle e_q^2\rangle C_{2,g}\Gamma_{\mathrm{S}}^{gq}$$

$$K_{\mathrm{F2},g} = \frac{5}{18}C^s_{2,q}\Gamma_{\mathrm{S}}^{qg} + \frac{1}{30}C_{2,q}(\Gamma_{\mathrm{S}}^{24,g} - \Gamma_{\mathrm{S}}^{35,g}) + \langle e_q^2\rangle C_{2,g}\Gamma_{\mathrm{S}}^{gg}$$

$$K_{\mathrm{F2},+} = \frac{1}{6}C_{2,q}\Gamma_{\mathrm{NS}}^{+}$$

Theoretical errors

- Higher perturbative orders $\longrightarrow$ NLO fit
- Heavy quark treatment $\longrightarrow$ Zero Mass Variable Flavor Number scheme. quarks are radiatively generated at thresholds. [Thorne,Tung, arXiv:0809.0714]
- Target Mass Corrections included and factorized into the hard kernels.

$$\widetilde{F}_2(\xi,Q^2) = \frac{x^2}{\tau^{3/2}}\frac{F_2(\xi,Q^2)}{\xi^2} + 6\frac{M_N^2}{Q^2}\frac{x^3}{\tau^2}I_2(\xi,Q^2)$$

$$\tau = 1 + \frac{4M_N^2 x^2}{Q^2}$$

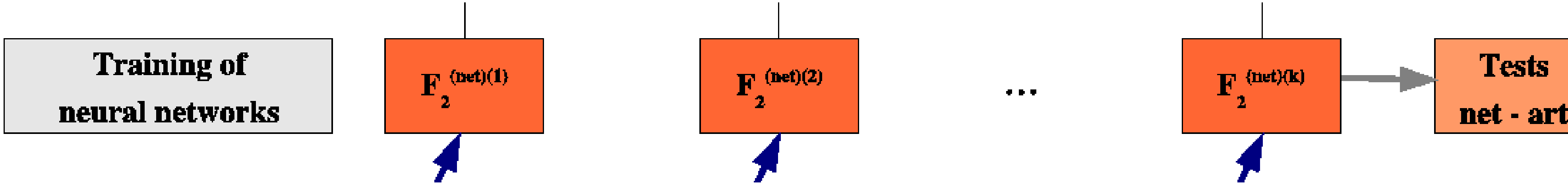
$$\xi = \frac{2x}{1+\sqrt{\tau}}$$

$$I_2(\xi,Q^2) = \int_\xi^1 \frac{dz}{z^2}F_2(z,Q^2).$$

Taking Mellin transforms with respect to ξ, defines a new target mass corrected coefficient function

$$\widetilde{C}_{2,j}(N,\alpha_s,\tau) = \frac{(1+\tau^{1/2})^2}{4\tau^{3/2}}\left(1 + \frac{3\left(1-\tau^{-1/2}\right)}{N+1}\right)C_{2,j}(N,\alpha_s)$$

Training strategy

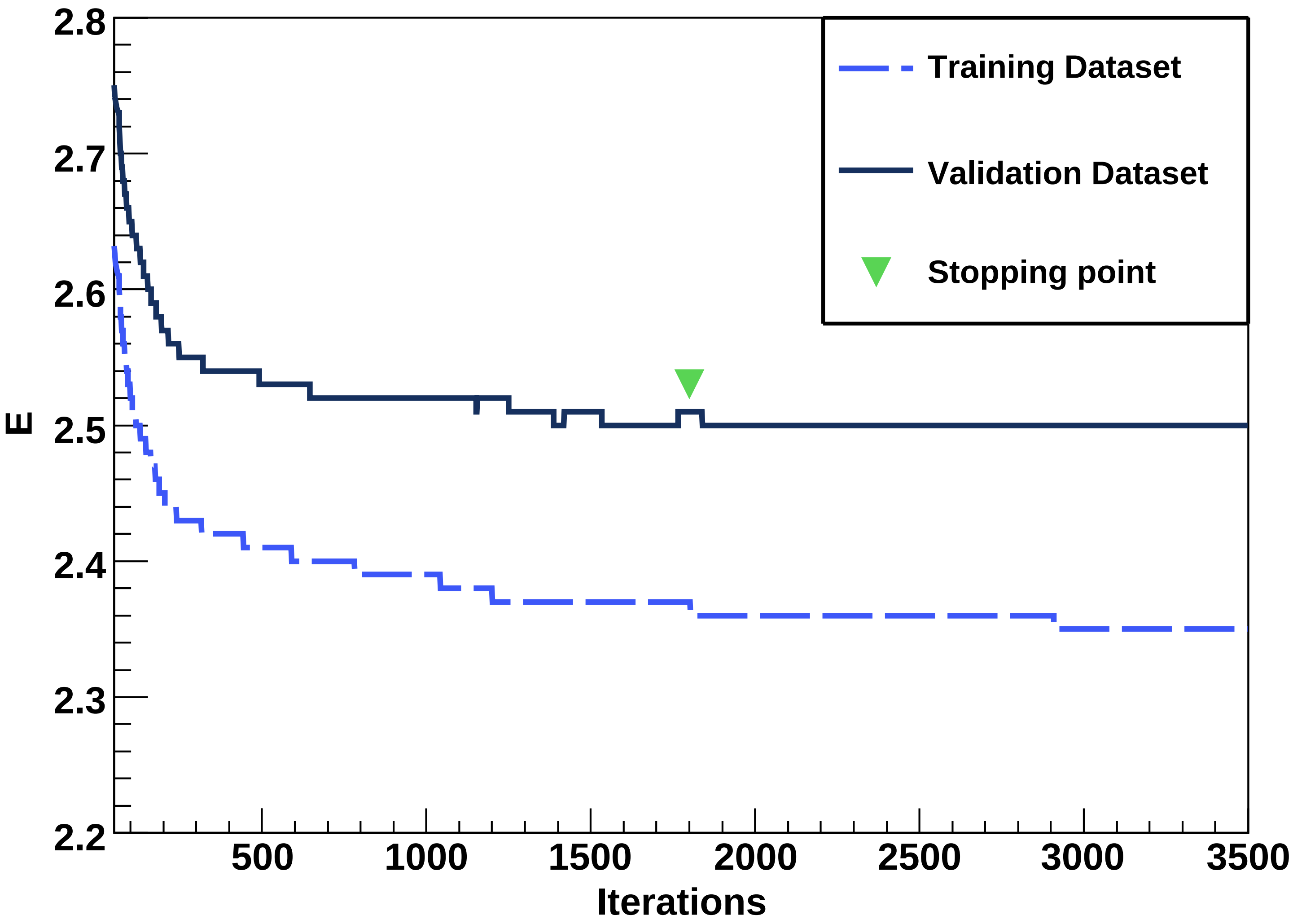


- unbiased basis of functions, parametrized by a **large** number of parameters
- genetic algorithms for minimization

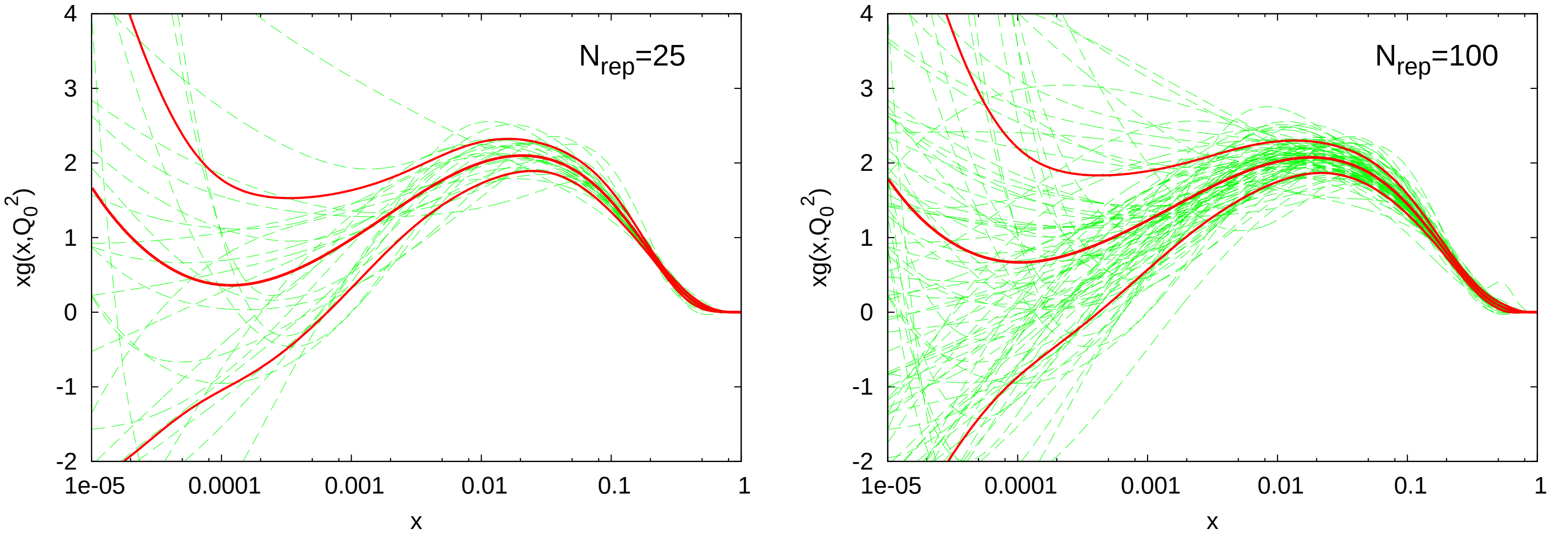
- might accomodate statistical fluctutations of the data
- optimal training, beyond which the fit is just adjusting to statistical fluctutations

- dynamical stopping by cross validation
- for each replica divide the data randomly into **training** and **validation**
- minimization performed on the training set **only**
- when the training χ^2 still decreases while the validation χ^2 stops decreasing $\longrightarrow$ STOP

Dynamical stopping



Ensemble of replicas



- individual replicas fluctuate significantly
- averages are smooth as the number of replicas is increased

PDFs uncertainties

- Monte Carlo prescription (NNPDF)

$$\sigma_{\mathcal{F}} = \left(\frac{N_{\rm set}}{N_{\rm set}-1} \left(\langle \mathcal{F}[\{f\}]^2 \rangle - \langle \mathcal{F}[\{f\}] \rangle^2 \right) \right)^{1/2}$$

- HEPDATA prescription (CTEQ and MRST/MSTW)

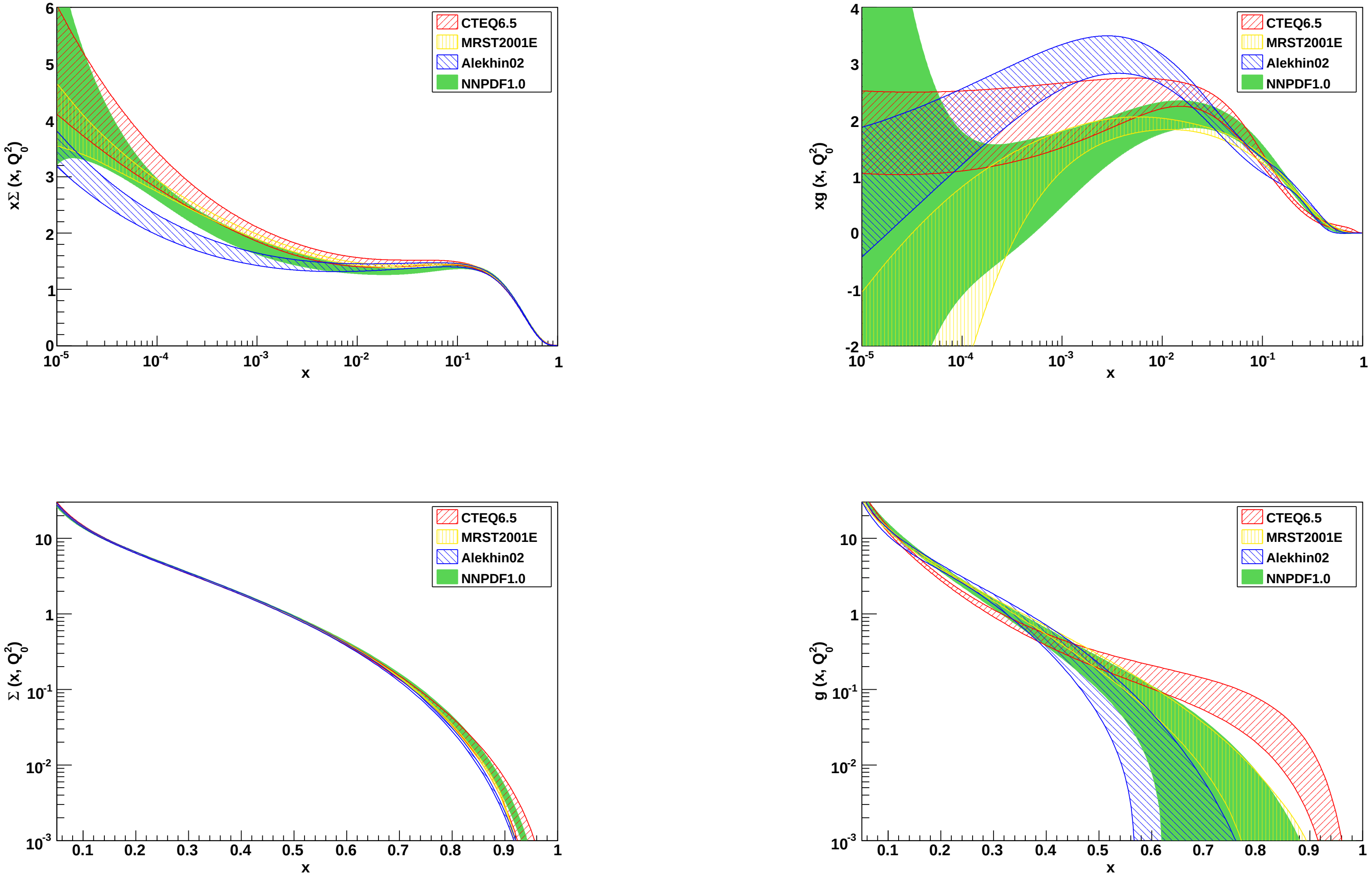
$$\sigma_{\mathcal{F}} = \frac{1}{2C_{90}} \left(\sum_{k=1}^{N_{\rm set}/2} \left(\mathcal{F}[\{f^{(2k-1)}\}] - \mathcal{F}[\{f^{(2k)}\}] \right)^2 \right)^{1/2} , \quad C_{90} = 1.64485$$

C_{90} accounts for the fact that the upper and lower parton sets correspond to 90% confidence levels rather than to one-σ uncertainties.

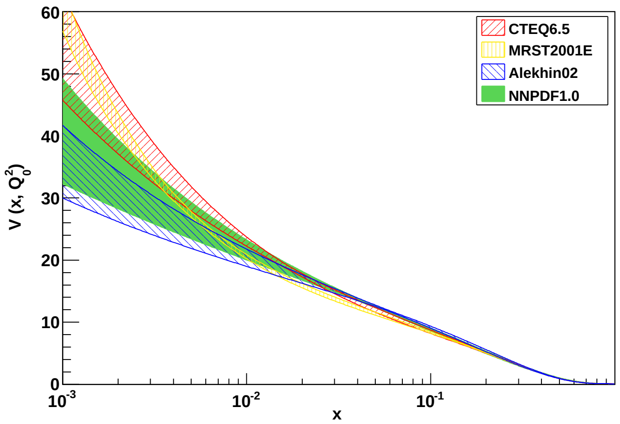
- HEPDATA* prescription (Alekhin)

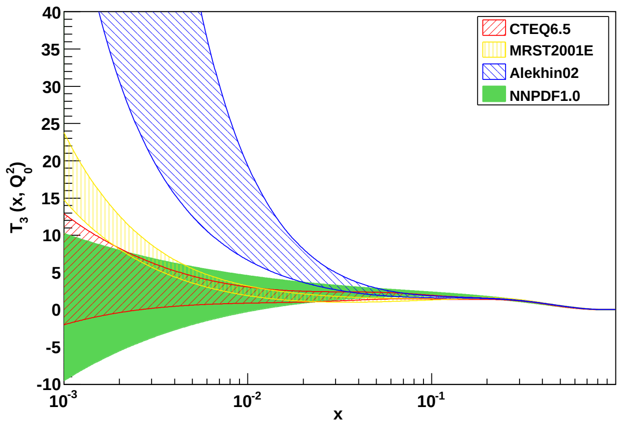
$$\sigma_{\mathcal{F}} = \left(\sum_{k=1}^{N_{\rm set}} \left(\mathcal{F}[\{f^{(k)}\}] - \mathcal{F}[\{f^{(0)}\}] \right)^2 \right)^{1/2} .$$

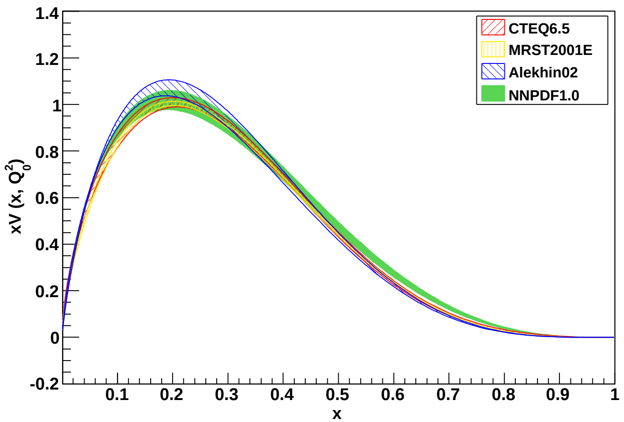
The NNPDF1.0 parton set

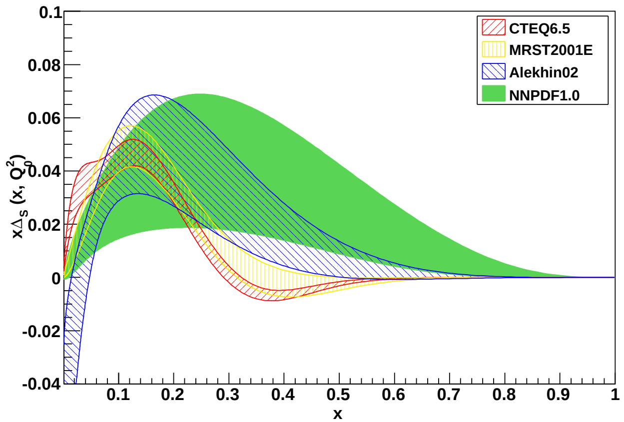


The NNPDF1.0 parton set

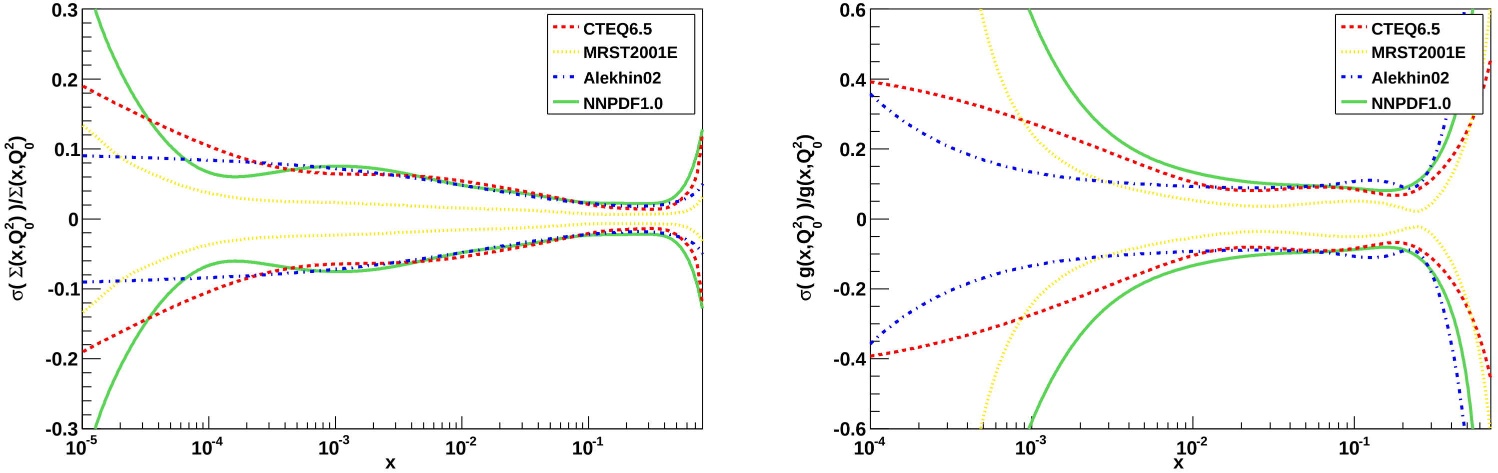








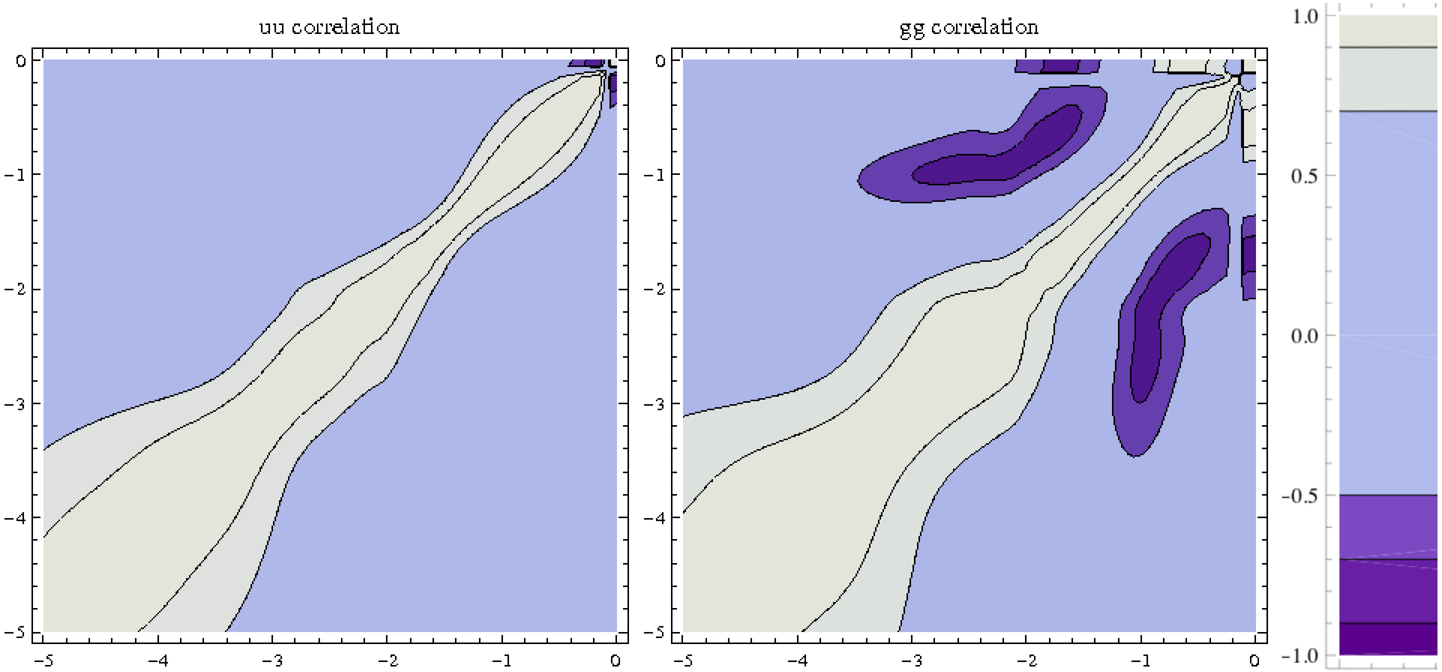
Relative uncertainties



- comparable uncertainty in the data region
- BUT larger uncertainties in the extrapolation

PDFs correlations

Correlations between $u-u$ and $g-g$ (Q=85GeV) [nadolsky 08]



$$\rho\left[f_a(x_1,Q_1^2)f_b(x_2,Q_2^2)\right]=\frac{\langle f_a(x_1,Q_1^2)f_b(x_2,Q_2^2)\rangle_{\mathrm{rep}}-\langle f_a(x_1,Q_1^2)\rangle_{\mathrm{rep}}\langle f_b(x_2,Q_2^2)\rangle_{\mathrm{rep}}}{\sigma_a(x_1,Q_1^2)\sigma_b(x_2,Q_2^2)}.$$

Distance between MC ensembles.

- Stability of the NNPDF parton set can be assessed by using standard statistical tools.
- Distances between two probability distributions: $\left\{ f_{ik}^{(1)} = f_k^{(1)}(x_i, Q_0^2) \right\}$

$$\langle d[f] \rangle = \sqrt{\left\langle \frac{\left(\langle f_i \rangle_{(1)} - \langle f_i \rangle_{(2)} \right)^2}{\sigma^2[f_i^{(1)}] + \sigma^2[f_i^{(2)}]} \right\rangle_{\rm pts}}$$

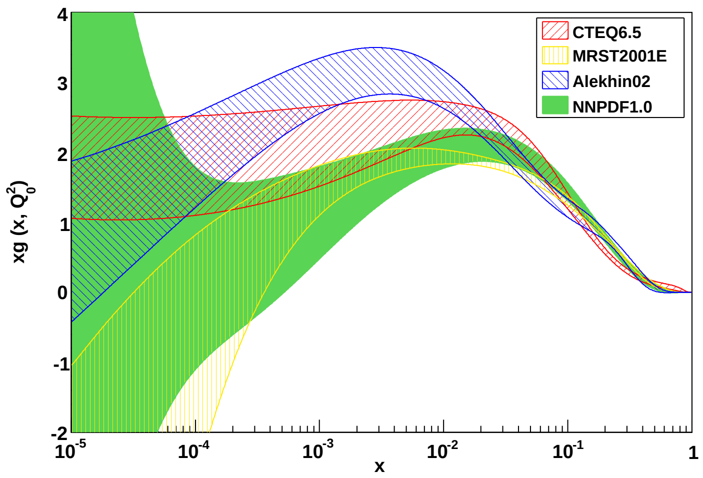
- where:

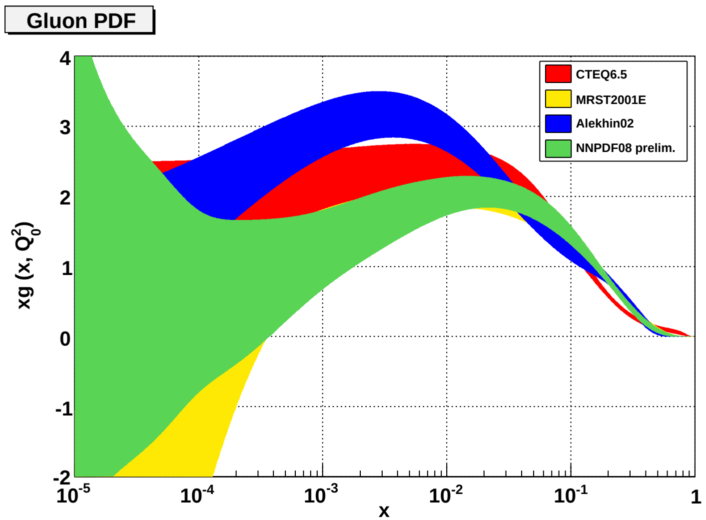
$$\langle f_i \rangle_{(1)} \equiv \frac{1}{N_{\rm rep}^{(1)}} \sum_{k=1}^{N_{\rm rep}^{(1)}} f_{ik}^{(1)} \; ,$$

$$\sigma^2[f_i^{(1)}] \equiv \frac{1}{N_{\rm rep}^{(1)}(N_{\rm rep}^{(1)} - 1)} \sum_{k=1}^{N_{\rm rep}^{(1)}} \left(f_{ik}^{(1)} - \langle f_i \rangle_{(1)} \right)^2$$

- For statistically equivalent PDF sets: $\langle d[f] \rangle \sim \langle d[\sigma_f] \rangle \sim 1$

Stability under variation of the parametrization





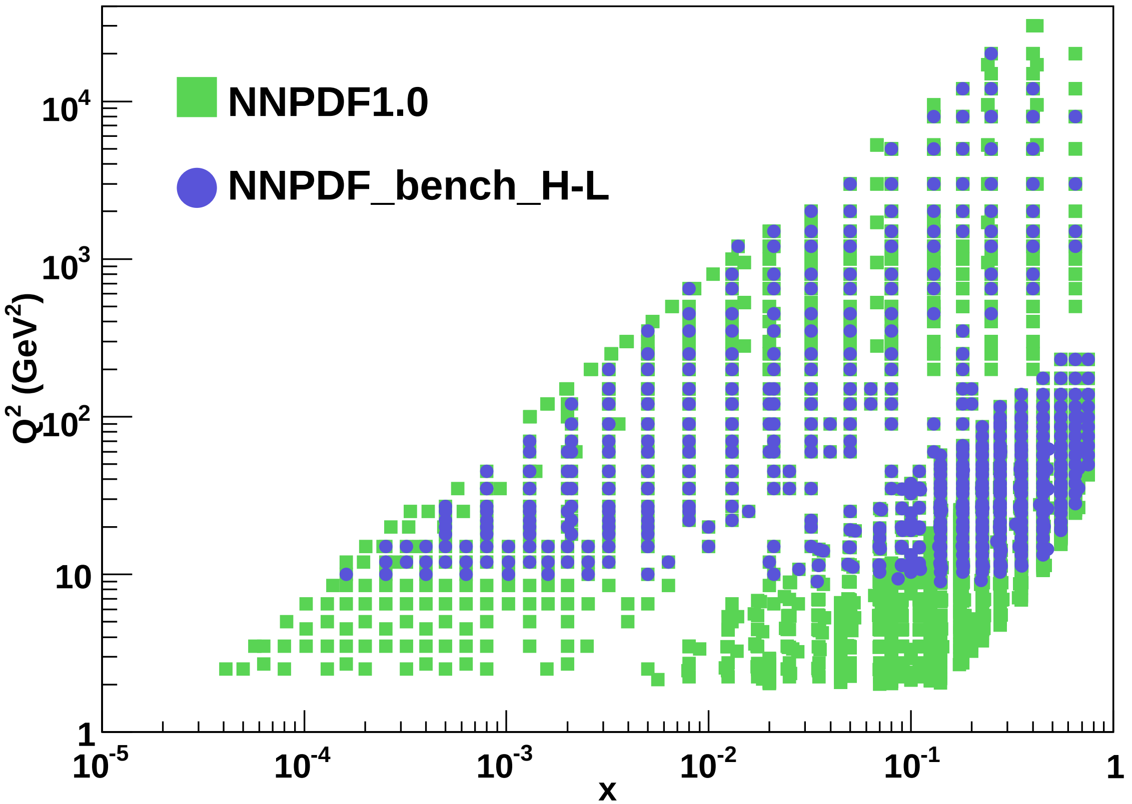
	Data	Extrapolation
$\Sigma(x, Q_0^2)$	$5\,10^{-4} \le x \le 0.1$	$10^{-5} \le x \le 10^{-4}$
$\langle d[f] \rangle$	0.98	1.25
$\langle d[\sigma] \rangle$	1.14	1.34
$g(x, Q_0^2)$	$5\,10^{-4} \le x \le 0.1$	$10^{-5} \le x \le 10^{-4}$
$\langle d[f] \rangle$	1.52	1.15
$\langle d[\sigma] \rangle$	1.16	1.07
$T_3(x, Q_0^2)$	$0.05 \le x \le 0.75$	$10^{-3} \le x \le 10^{-2}$
$\langle d[f] \rangle$	1.00	1.11
$\langle d[\sigma] \rangle$	1.76	2.27
$V(x, Q_0^2)$	$0.1 \le x \le 0.6$	$3\,10^{-3} \le x \le 3\,10^{-2}$
$\langle d[f] \rangle$	1.30	0.90
$\langle d[\sigma] \rangle$	1.10	0.98
$\Delta_S(x, Q_0^2)$	$0.1 \le x \le 0.6$	$3\,10^{-3} \le x \le 3\,10^{-2}$
$\langle d[f] \rangle$	1.04	1.91
$\langle d[\sigma] \rangle$	1.44	1.80

- Stability under change of architecture of the nets:
  37 pars $\longrightarrow$ **31 pars**
- Independence on the parametrization!

Dependence on data sets

HERA-LHC benchmark

Benchmark PDF fit to a reduced consistent set of DIS data [hep-ph/0511119]



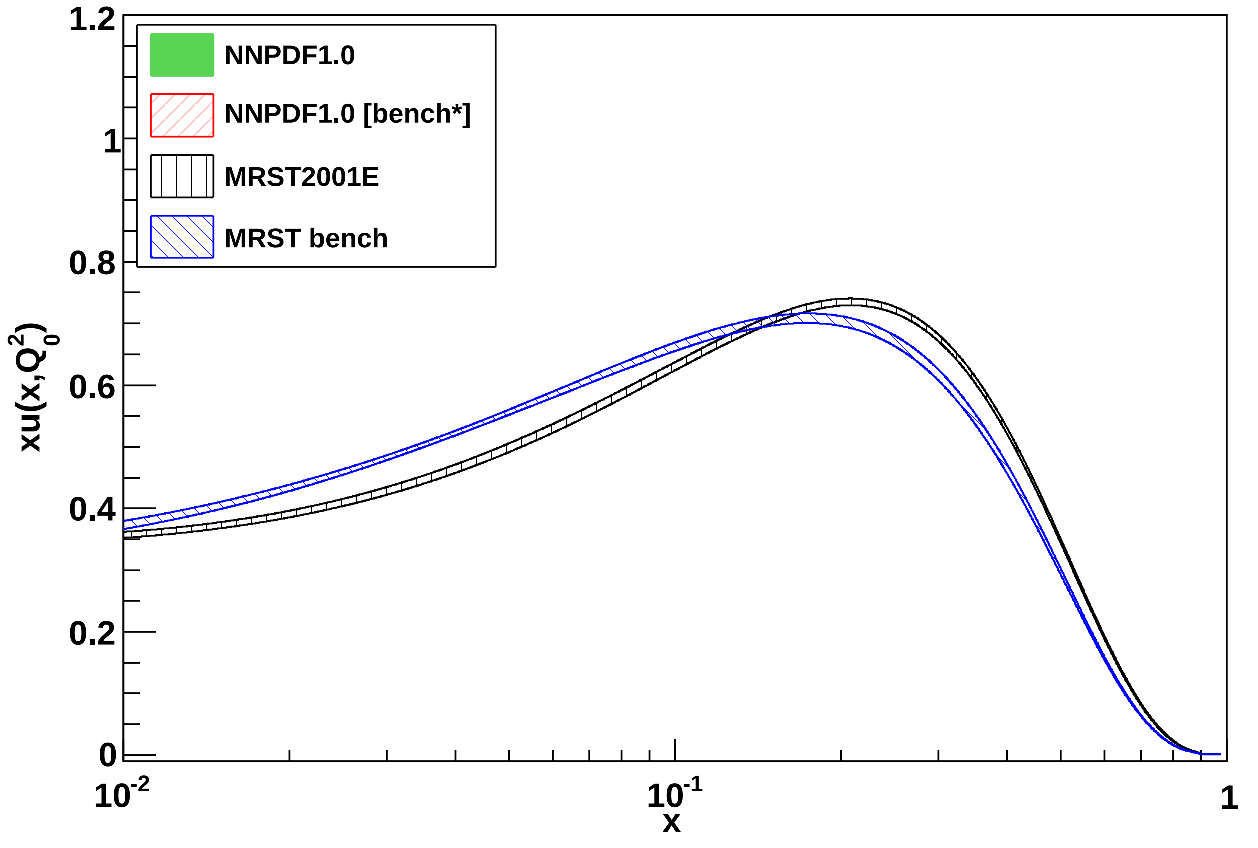
3163 data $\longrightarrow$ **773** data

Dependence on data sets

HERA-LHC benchmark

Comparison between collaborations and between benchmark/global partons.

$u(x, Q^2 = 2\mathrm{GeV}^2)$: MRST data region

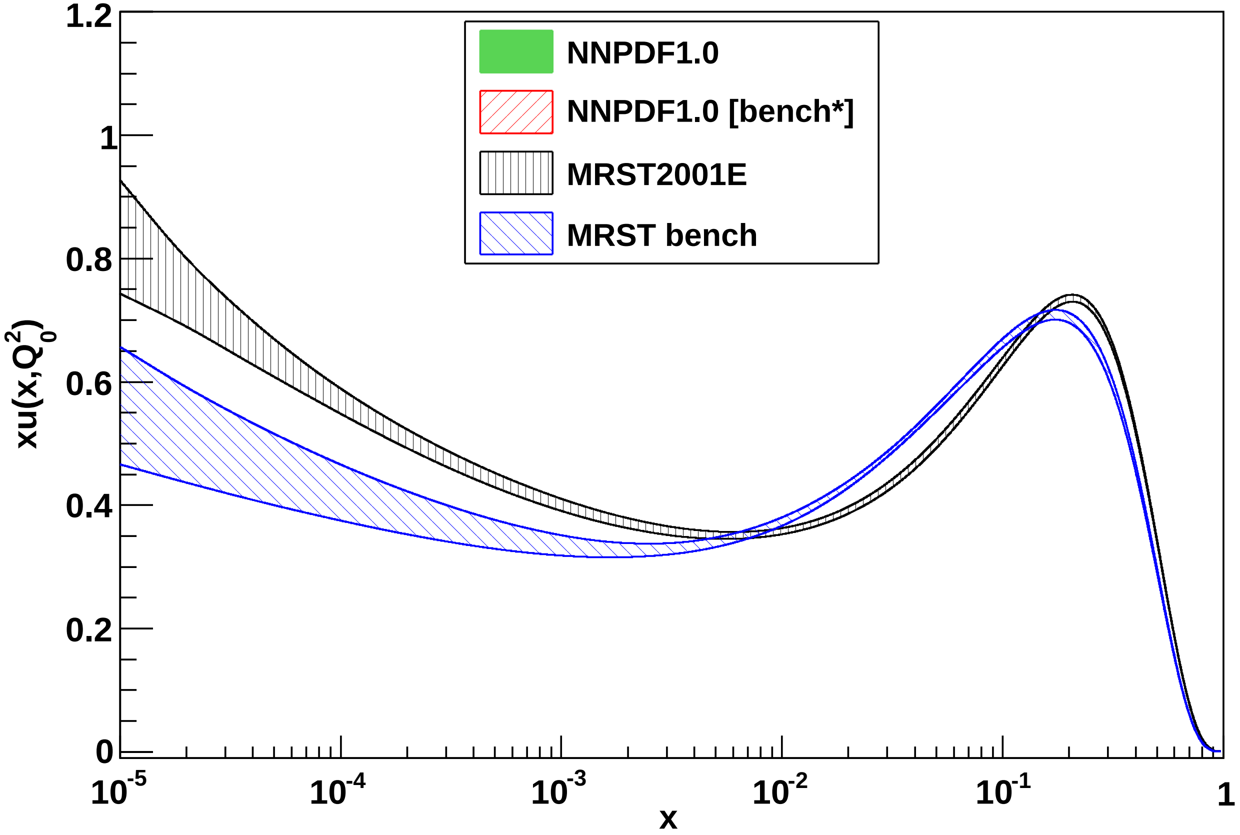


Dependence on data sets

HERA-LHC benchmark

Comparison between collaborations and between benchmark/global partons.

$u(x, Q^2 = 2\mathrm{GeV}^2)$: MRST extrapolation region



Dependence on data sets

HERA-LHC benchmark

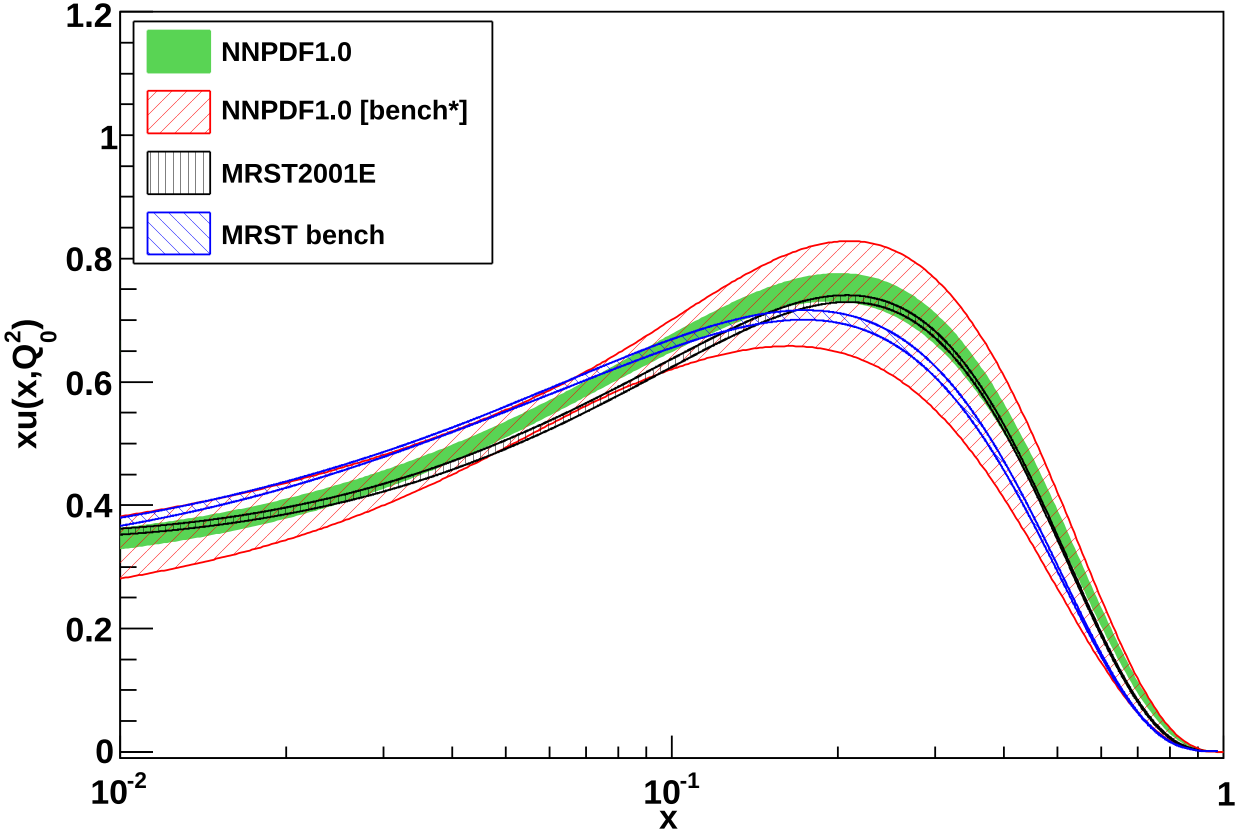
- benchmark partons and global partons do not agree within error!
- note that PDFs input parametrization, flavor assumptions and statistical treatment ($\Delta\chi^2_{\mathrm{global}} = 50$, $\Delta\chi^2_{\mathrm{bench}} = 1$) are tuned to data.
- not satisfactory especially to predict the behaviour of PDFs in the extrapolation region (LHC)

Dependence on data sets

HERA-LHC benchmark

Comparison between collaborations and between benchmark/global partons.

$u(x, Q^2 = 2\mathrm{GeV}^2)$: data region

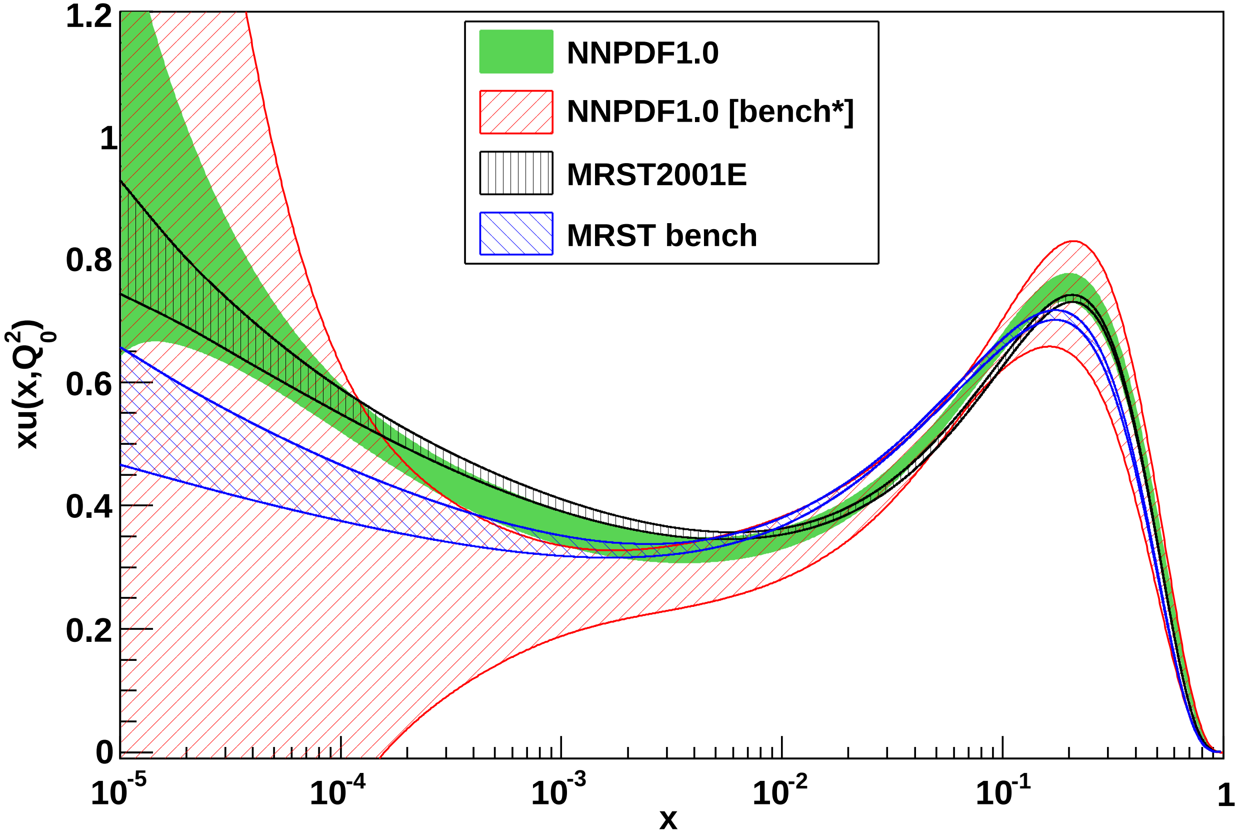


Dependence on data sets

HERA-LHC benchmark

Comparison between collaborations and between benchmark/global partons.

$u(x, Q^2 = 2\mathrm{GeV}^2)$: extrapolation region

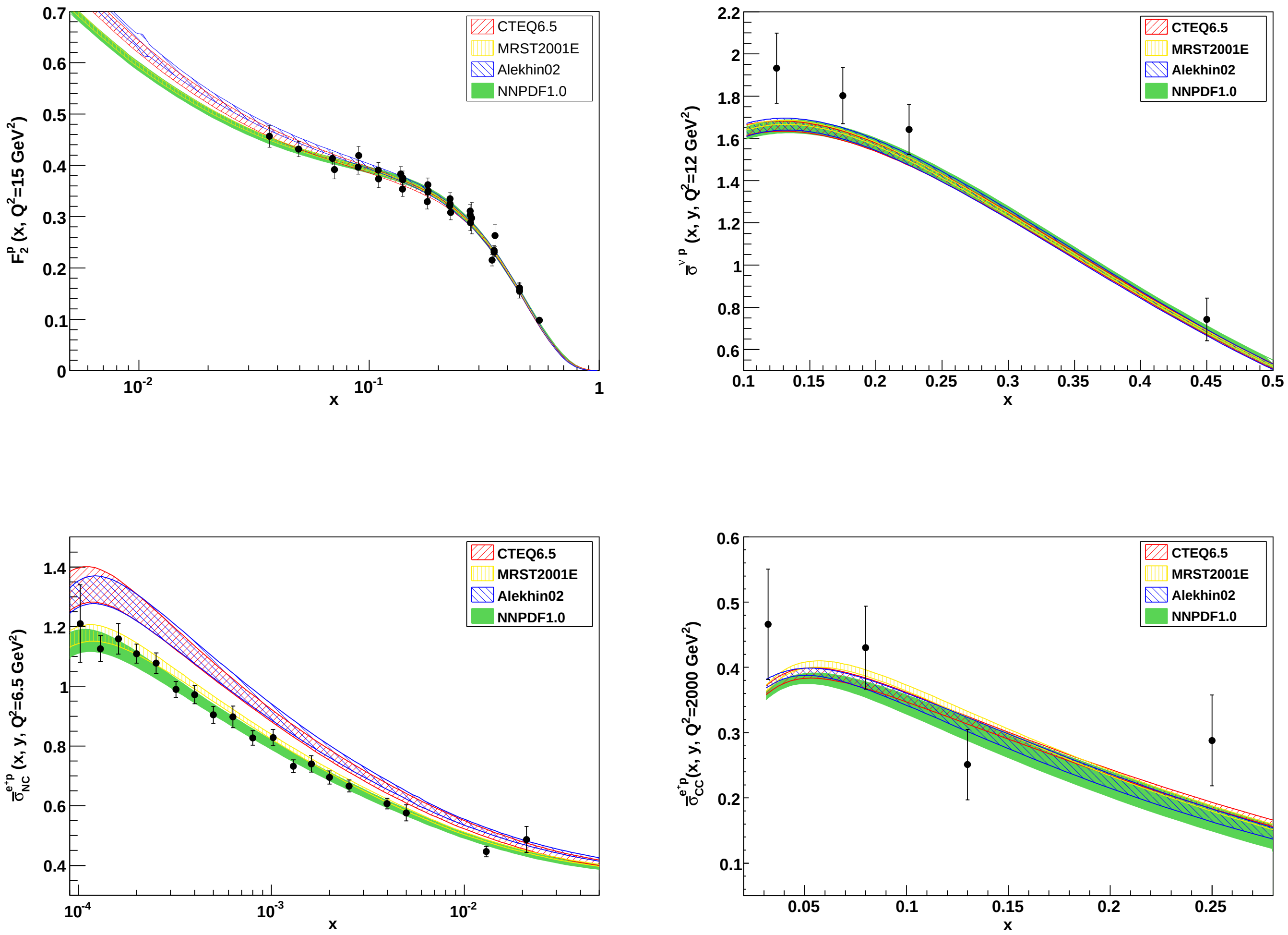


Dependence on data sets

HERA-LHC benchmark

- NNPDF1.0 is consistent with MRST global fit.
- NNPDFbench is consistent with NNPDF1.0 and MRST.
- Same parametrization and flavour assumption.
- Same statistical treatment.
- Underestimation of the error in the standard approach.

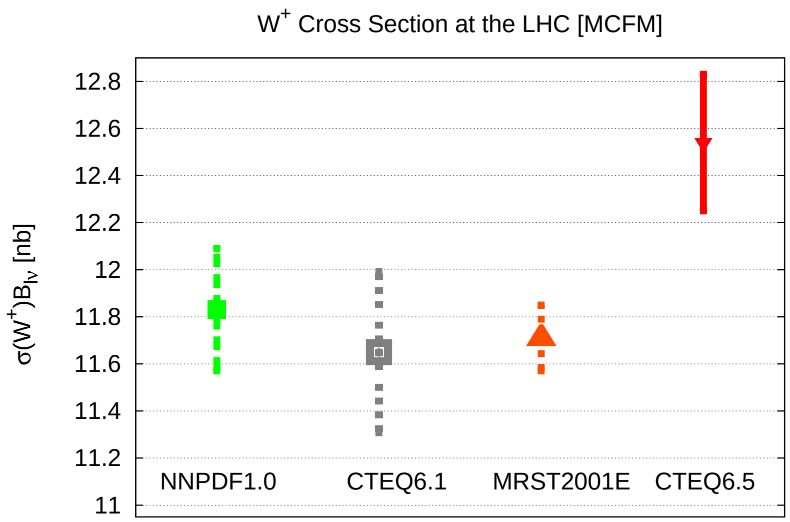
Comparison with present experimental data

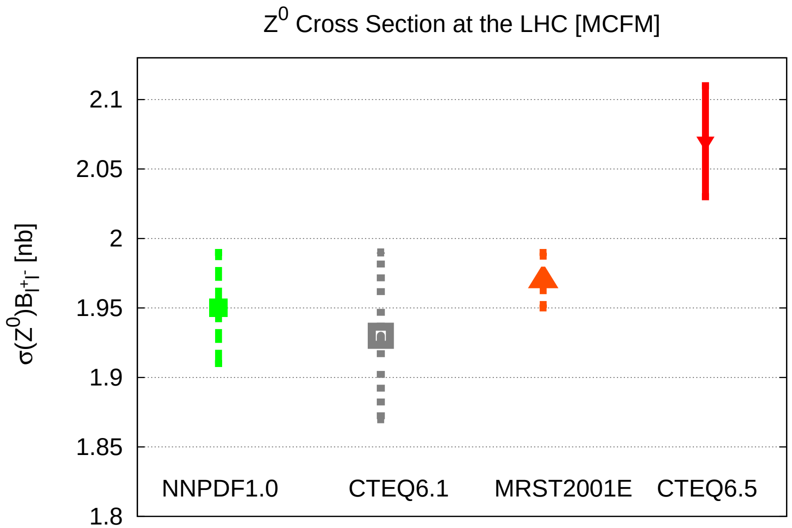


LHC standard candle processes

- All quantities have been computed at NLO with MCFM [http://mcfm.fnal.gov]
- Quoted uncertainties are the 1σ bands due to the PDF uncertainty only.

	$\sigma_{W^+}\mathcal{B}_{l^+\nu_l}$	$\Delta\sigma_{W^+}/\sigma_{W^+}$	$\sigma_Z\mathcal{B}_{l^+l^-}$	$\Delta\sigma_Z/\sigma_Z$
NNPDF1.0	11.83 ± 0.26	2.2%	1.95 ± 0.04	2.1%
CTEQ6.1	11.65 ± 0.34	2.9%	1.93 ± 0.06	3.1%
MRST01	11.71 ± 0.14	1.2%	1.97 ± 0.02	1.0%
CTEQ6.5	12.54 ± 0.29	2.3%	2.07 ± 0.04	1.9%





Conclusions

- Standard approaches with fixed parametrization tend to underestimate uncertainties unless experimental errors are inflated by essentially arbitrary amount.
- Monte Carlo ensemble
  - Any statistical property of PDFs can be calculated using standard statistical methods.
  - No need of any tolerance criterion.
- The Neural Network parametrization
  - Small uncertainties come from an underlying physical law, not from parametrization bias.
  - Inconsistent data or underestimated uncertainties do not require a separate treatment and are automatically signalled by a larger value of the χ^2.
- The first NNPDF parton set [arXiv:0808.1231] is available on the common LHAPDF interface [http://projects.hepforge.org/lhapdf].

Outlook

- Inclusion of hadronic data to
  - improve the accuracy of gluon at large x (jets)
  - determine the light antiquark sea asymmetry (Drell-Yan)
  - allow for a direct determination of the strange distribution (dimuon data)
- More accurate treatment of Heavy Quark thresholds.
- LO parton set in view of its use in Monte Carlo generators.
- More sophisticated theoretical treatment: NNLO parton distributions, large and small x resummation corrections should also be considered.
- Study of the impact of PDFs uncertainties on LHC phenomenology