The NNPDF4.0 global analysis of the proton structure

Roy Stegeman

University of Milan and INFN Milan

On behalf of the NNPDF Collaboration

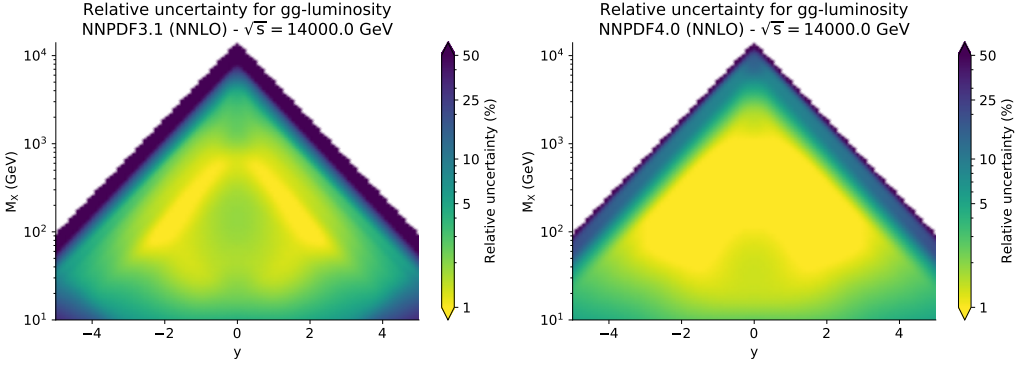
Based on: arXiv:2109.02653

DIS 2022, 3 May 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006.

High-precision: gluon

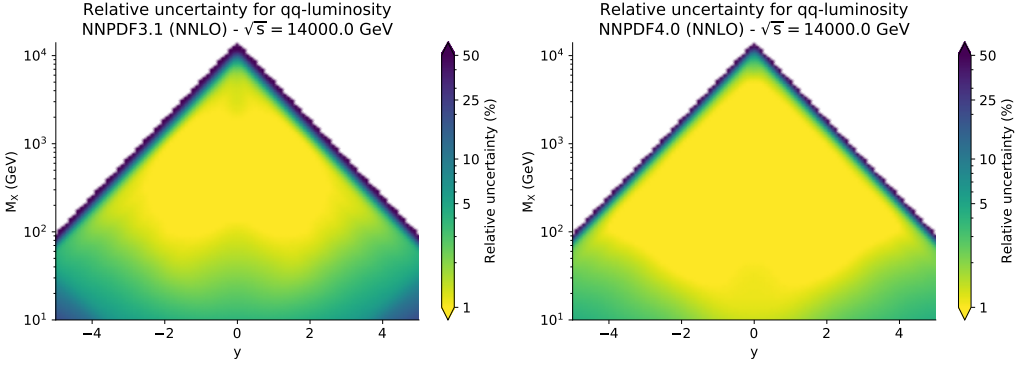
$$\mathcal{L}_{ij}\left(M_X, y, \sqrt{s}\right) = \frac{1}{s}\sum_{i,j} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$



How did we get here?

High-precision: singlet

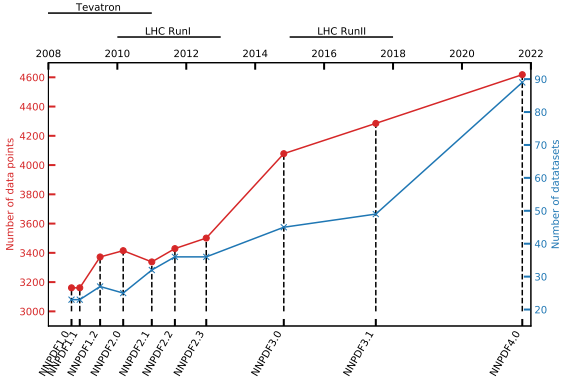
$$\mathcal{L}_{ij}\left(M_X, y, \sqrt{s}\right) = \frac{1}{s}\sum_{i,j} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$



How did we get here?

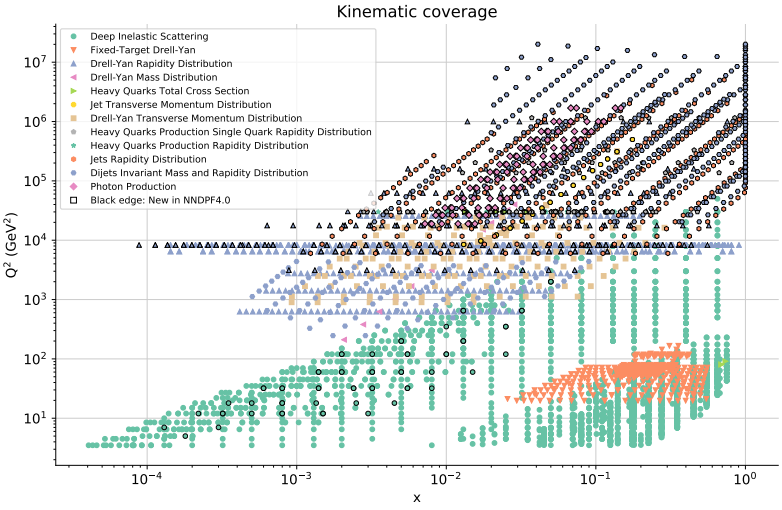
Data

Data from NNPDF1.0 to NNPDF4.0



The number of datasets – normally corresponding to different processes – is generally more relevant than the number of datapoints

Experimental data in NNPDF4.0



New processes:

- direct photon
- single top
- dijets
- W+jet
- DIS jet

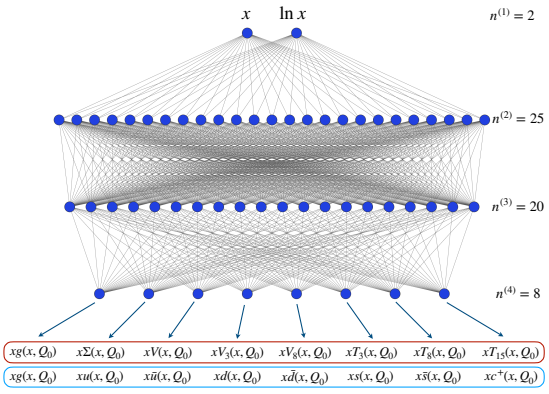
Methodology

Improved fitting methodology

- **Stochastic Gradient Descent** for NN training using TensorFlow
- Automated optimization of **model hyperparameters**
- Methodology is validated using **closure tests** (data region), **future tests** (extrapolation region), and **parametrization basis independence**

Physical constraints:

- PDF positivity
- Integrability of nonsinglet distributions (Gottfried sum rules)



$$f_i\left(x, Q_0\right) = x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}_i(x)$$

Automated model selection

NNPDF aims to minimize sources of bias in the PDF:

- Functional form $\rightarrow$ Neural Network
- Model parameters $\rightarrow$?

Automated model selection

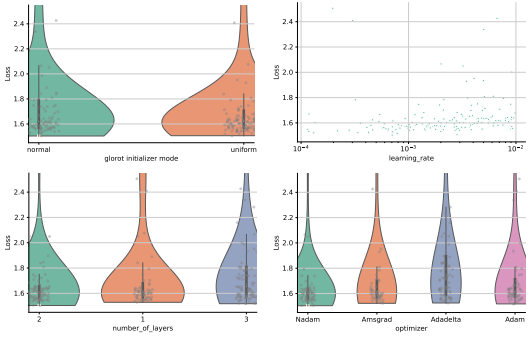
NNPDF aims to minimize sources of bias in the PDF:

- Functional form $\rightarrow$ Neural Network
- Model parameters $\rightarrow$ **Hyperoptimization**

Scan over thousands of hyperparameter combinations and select the best one

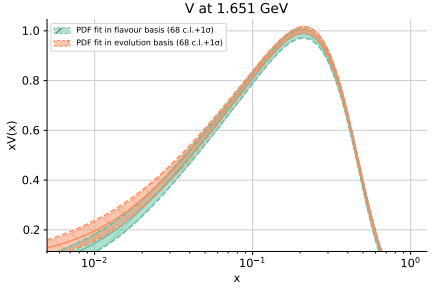
k-fold cross-validation: used to define the reward function based on a **test dataset**

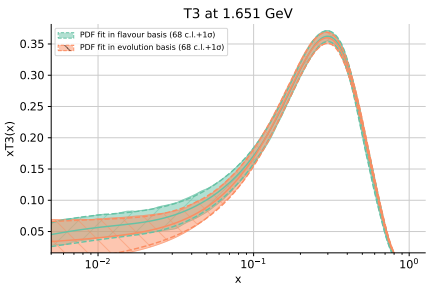
Objective function:
$L = \mathrm{mean}(\chi^2_1, \chi^2_3, \chi^2_2, \ldots, \chi^2_k)$



Stability

Parametrization basis independence





Evolution Basis:

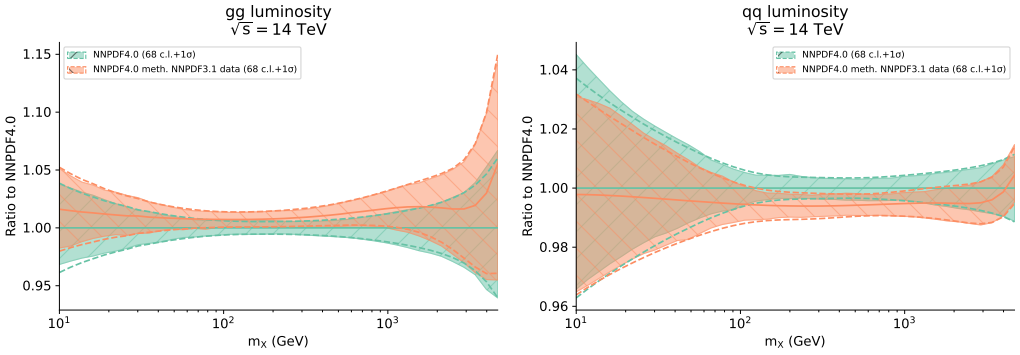
$$xV\left(x,Q_0\right) \propto \mathrm{NN}_V(x)$$
$$xT_3\left(x,Q_0\right) \propto \mathrm{NN}_{T_3}(x)$$

Flavour Basis:

$$xV\left(x,Q_0\right) \propto \left(\mathrm{NN}_u(x) - \mathrm{NN}_{\bar{u}}(x) + \mathrm{NN}_d(x) - \mathrm{NN}_{\bar{d}}(x) + \mathrm{NN}_s(x) - \mathrm{NN}_{\bar{s}}(x)\right)$$
$$xT_3\left(x,Q_0\right) \propto \left(\mathrm{NN}_u(x) + \mathrm{NN}_{\bar{u}}(x) - \mathrm{NN}_d(x) - \mathrm{NN}_{\bar{d}}(x)\right)$$

Different strategies to parametrize the quark PDF flavour combinations leave the uncertainties essentially unchanged

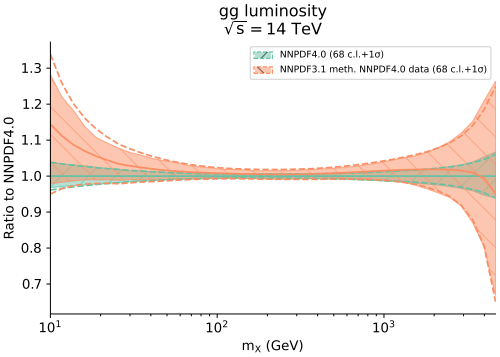
Impact of the new data

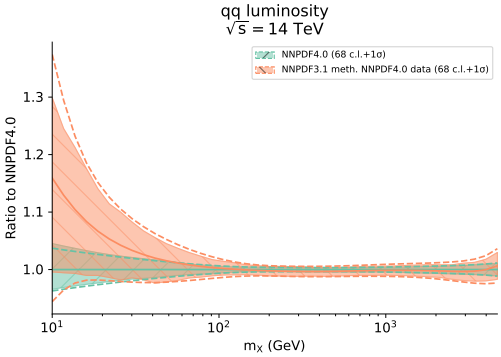


Individual datasets have a limited impact, but collectively they result in:

- Moderate reduction of PDF uncertainties
- Shifts in central value at the one-sigma level

Impact of the new fitting methodology





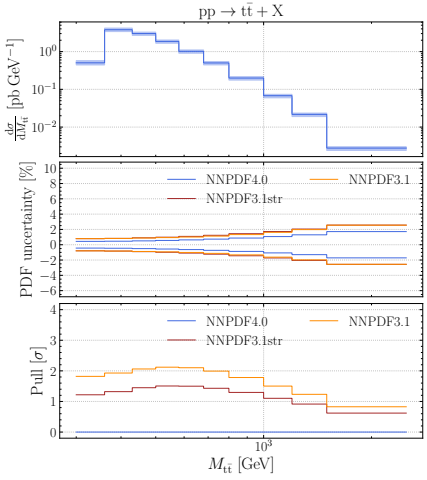
- Significant reduction of PDF uncertainties
- Good agreement between the central values

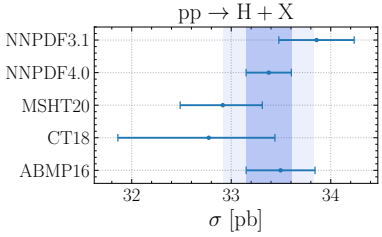
PDF uncertainties are validated using closure tests and future tests
Validation tests successful for both NNPDF4.0 and NNPDF3.1

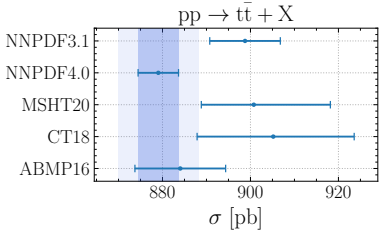
LHC phenomenology

Implications for LHC phenomenology

Reduced luminosity uncertainties → Reduced uncertainty at the level of observables







Open-source code

The open-source NNPDF code

The full NNPDF code has been made public along with user friendly documentation

This includes: fitting, hyperoptimization, theory, data processing, visualization

It is possible to reproduce all results of NNPDF4.0 and more!

Eur.Phys.J.C 81 (2021) 10, 958
`https://github.com/NNPDF/nnpdf`
`https://docs.nnpdf.science`

Summary and Outlook

Summary and Outlook

- NNPDF4.0 is the latest release in the NNPDF family of PDF sets
- 44 new datasets from many new processes are included
- Improved methodology with Stochastic Gradient Descent and hyperoptimization
- Validation of PDF uncertainties using closure test, future test and parametrization basis independence

$\Rightarrow$ NNPDF4.0 achieves a high precision over a broad kinematic range

- The current level of PDF uncertainties challenges the accuracy of theoretical predictions and demands an increased effort towards the systematic inclusion in the fit of theoretical uncertainties (nuclear, higher orders, SM parameters, ...) and higher-order QCD and EW corrections

Summary and Outlook

- NNPDF4.0 is the latest release in the NNPDF family of PDF sets
- 44 new datasets from many new processes are included
- Improved methodology with Stochastic Gradient Descent and hyperoptimization
- Validation of PDF uncertainties using closure test, future test and parametrization basis independence

$\Rightarrow$ NNPDF4.0 achieves a high precision over a broad kinematic range

- The current level of PDF uncertainties challenges the accuracy of theoretical predictions and demands an increased effort towards the systematic inclusion in the fit of theoretical uncertainties (nuclear, higher orders, SM parameters, ...) and higher-order QCD and EW corrections

Thank you!

Backup

Experimental data in NNPDF4.0

- 44 new datasets included
- 323 more data points in NNPDF4.0 than in NNPDF3.1
- New data is mostly from the LHC RUN II

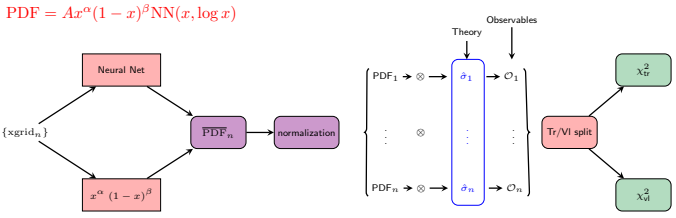
Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
ATLAS W, Z 7 TeV ($\mathcal{L} = 35$ pb^{-1})	[51]	✓	✓	✓	✓	✓
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ fb^{-1})	[52]	✓	✓	✗	(✓)	✓
ATLAS low-mass DY 7 TeV	[53]	✓	✓	✗	(✓)	✗
ATLAS high-mass DY 7 TeV	[54]	✓	✓	✗	(✓)	✓
ATLAS W 8 TeV	[79]	✗	(✓)	✗	✗	✓
ATLAS DY 2D 8 TeV	[78]	✗	✓	✗	✗	✓
ATLAS high-mass DY 2D 8 TeV	[77]	✗	✓	✗	(✓)	✓
ATLAS $\sigma_{W,Z}$ 13 TeV	[81]	✗	✓	✓	✗	✗
ATLAS W+jet 8 TeV	[93]	✗	✓	✗	✗	✓
ATLAS Z p_T 7 TeV	[260]	(✓)	✗	✗	(✓)	✗
ATLAS Z p_T 8 TeV	[63]	✓	✓	✗	✓	✓
ATLAS $W + c$ 7 TeV	[83]	✗	✓	✗	(✓)	✗
ATLAS σ_{tt}^{tot} 7, 8 TeV	[65]	✓	✓	✓	✗	✗
ATLAS σ_{tt}^{tot} 7, 8 TeV	[261–266]	✗	✗	✓	✗	✗
ATLAS σ_{tt}^{tot} 13 TeV ($\mathcal{L} = 3.2$ fb^{-1})	[66]	✓	✗	✓	✗	✗
ATLAS σ_{tt}^{tot} 13 TeV ($\mathcal{L} = 139$ fb^{-1})	[134]	✗	✓	✗	✗	✗
ATLAS σ_{tt}^{tot} and Z ratios	[267]	✗	✗	✗	✗	(✓)
ATLAS $t\bar{t}$ lepton+jets 8 TeV	[67]	✓	✓	✗	✓	✓
ATLAS $t\bar{t}$ dilepton 8 TeV	[89]	✗	✓	✗	✗	✓
ATLAS single-inclusive jets 7 TeV, R=0.6	[73]	✓	(✓)	✗	✓	✓
ATLAS single-inclusive jets 8 TeV, R=0.6	[86]	✗	✓	✗	✗	✗
ATLAS dijets 7 TeV, R=0.6	[148]	✗	✓	✗	✗	✗
ATLAS direct photon production 8 TeV	[100]	✗	(✓)	✗	✗	✗
ATLAS direct photon production 13 TeV	[101]	✗	✓	✗	✗	✗
ATLAS single top R_t 7, 8, 13 TeV	[94, 96, 98]	✗	✓	✓	✗	✗
ATLAS single top diff. 7 TeV	[94]	✗	✓	✗	✗	✗
ATLAS single top diff. 8 TeV	[96]	✗	✓	✗	✗	✗

Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
CMS W asym. 7 TeV ($\mathcal{L} = 36$ pb^{-1})	[268]	✗	✗	✗	✗	✓
CMS Z 7 TeV ($\mathcal{L} = 36$ pb^{-1})	[269]	✗	✗	✗	✗	✓
CMS W electron asymmetry 7 TeV	[55]	✓	✓	✗	✓	✓
CMS W muon asymmetry 7 TeV	[56]	✓	✓	✓	✓	✗
CMS Drell-Yan 2D 7 TeV	[57]	✓	✓	✗	(✓)	✓
CMS Drell-Yan 2D 8 TeV	[270]	(✓)	✗	✗	✗	✗
CMS W rapidity 8 TeV	[58]	✓	✓	✓	✓	✓
CMS W, Z p_T 8 TeV ($\mathcal{L} = 18.4$ fb^{-1})	[271]	✗	✗	✗	(✓)	✗
CMS Z p_T 8 TeV	[64]	✓	✓	✗	(✓)	✗
CMS $W + c$ 7 TeV	[76]	✓	✓	✗	(✓)	✓
CMS $W + c$ 13 TeV	[84]	✗	✓	✗	✗	(✓)
CMS single-inclusive jets 2.76 TeV	[75]	✓	✗	✗	✗	✓
CMS single-inclusive jets 7 TeV	[147]	✓	(✓)	✗	✓	✓
CMS dijets 7 TeV	[74]	✗	✓	✗	✗	✗
CMS single-inclusive jets 8 TeV	[87]	✗	✓	✗	✓	✓
CMS 3D dijets 8 TeV	[149]	✗	(✓)	✗	✗	✗
CMS σ_{tt}^{tot} 5 TeV	[88]	✗	✓	✗	✗	✗
CMS σ_{tt}^{tot} 7, 8 TeV	[146]	✓	✓	✗	✗	✗
CMS σ_{tt}^{tot} 8 TeV	[272]	✗	✗	✗	✗	✓
CMS σ_{tt}^{tot} 5, 7, 8, 13 TeV	[68, 273–281]	✗	✗	✓	✗	✗
CMS σ_{tt}^{tot} 13 TeV	[69]	✓	✓	✓	✗	✗
CMS $t\bar{t}$ lepton+jets 8 TeV	[70]	✓	✓	✗	✗	✓
CMS $t\bar{t}$ 2D dilepton 8 TeV	[90]	✗	✓	✗	✓	✓
CMS $t\bar{t}$ lepton+jet 13 TeV	[91]	✗	✓	✗	✗	✗
CMS $t\bar{t}$ dilepton 13 TeV	[92]	✗	✓	✗	✗	✗
CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV	[95]	✗	✓	✓	✗	✗
CMS single top R_t 8, 13 TeV	[97, 99]	✗	✓	✓	✗	✗
CMS single top 13 TeV	[282, 283]	✗	✗	✗	✗	(✓)

Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
LHCb Z 7 TeV ($\mathcal{L} = 940$ pb^{-1})	[59]	✓	✓	✗	✗	✓
LHCb $Z \to ee$ 8 TeV ($\mathcal{L} = 2$ fb^{-1})	[61]	✓	✓	✓	✓	✓
LHCb W 7 TeV ($\mathcal{L} = 37$ pb^{-1})	[284]	✗	✗	✗	✗	✓
LHCb $W, Z \to \mu$ 7 TeV	[60]	✓	✓	✓	✓	✓
LHCb $W, Z \to \mu$ 8 TeV	[62]	✓	✓	✓	✓	✓
LHCb $W \to e$ 8 TeV	[80]	✗	(✓)	✗	✗	✗
LHCb $Z \to \mu\mu, ee$ 13 TeV	[82]	✗	✓	✗	✗	✗

NNPDF4.0 model

For more information see EPJ C79 (2019) 676



Main changes:

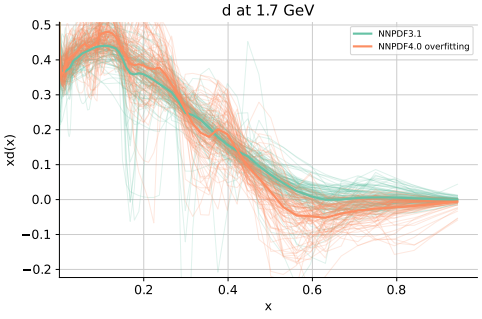
- `Python` codebase
  - Easier and faster development
- Freedom to use external libraries (default: `TensorFlow`)
- Modularity $\Rightarrow$ ability to vary all aspects of the methodology

Performance benefit - time per replica

	NNPDF3.1	NNPDF4.0 (CPU)	NNPDF4.0 (GPU)
Fit timing per replica	15.2 h	38 min	6.6 min
Speed up factor	1	24	140
RAM use	1.5 GB	6.1 GB	NA

Hyperoptimization: the reward function

Choosing as the hyperoptimization target the χ^2 of fitted data results in overfitting.



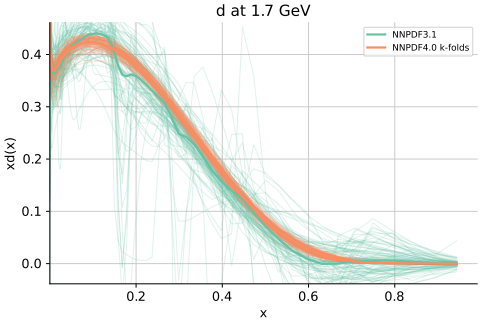
Hyperoptimization: the reward function

Choosing as the hyperoptimization target the χ^2 of fitted data results in overfitting.

We solve this using **k-fold cross-validation**:

1. Divide the data into k representative subsets
2. Fit $k-1$ sets and use k-th as test set
   $\Rightarrow$ k values of χ^2_{test}
3. Optimize the average χ^2_{test} of the k test sets

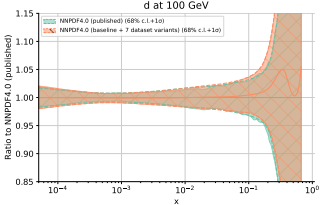
$\Rightarrow$ The hyperoptimization target is not based on data that entered the fit.

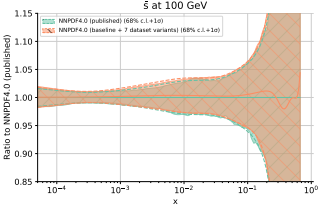


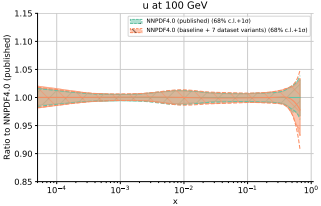
- No overfitting
- Compared to NNPDF3.1:
  - Increased stability
  - Reduced uncertainties

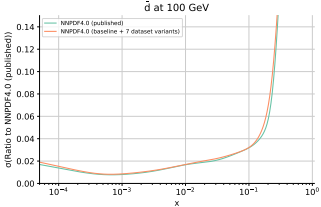
The (negligible) impact of datasets with tension

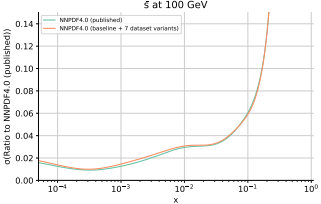
Excluding datasets with large $(\chi^2 - 1)/\sigma_{\chi^2}$ one at a time and combining the resulting PDFs following the conservative PDF4LHC15 prescription shows stability at the level of statistical fluctuations.

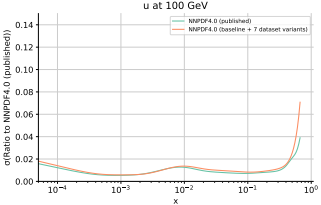






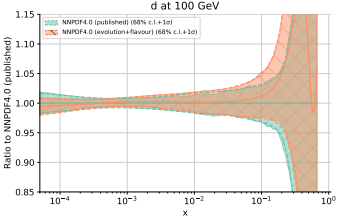


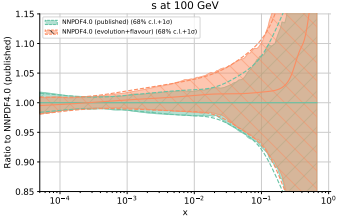


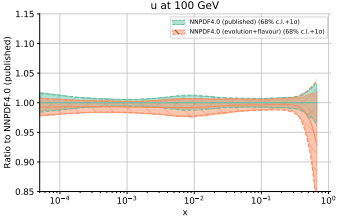


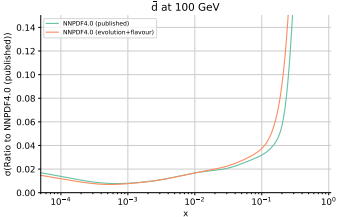
Envelope of fits with different arametrization bases

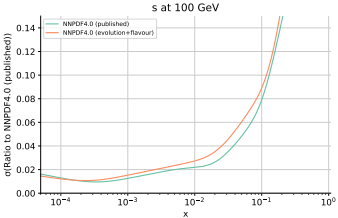
Different strategies to parametrize the PDF flavour combinations lead to the same result

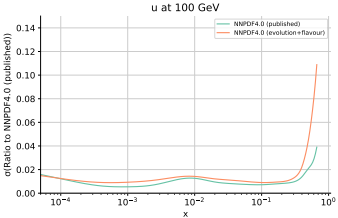




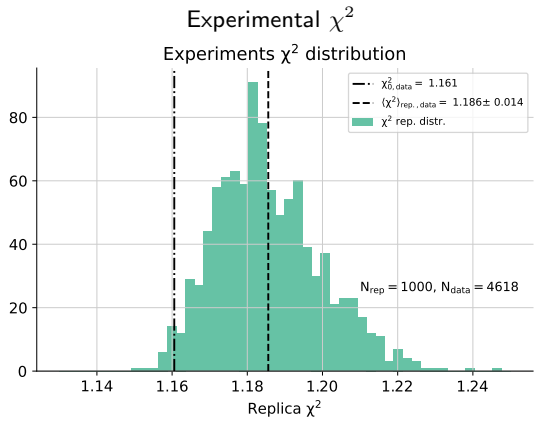


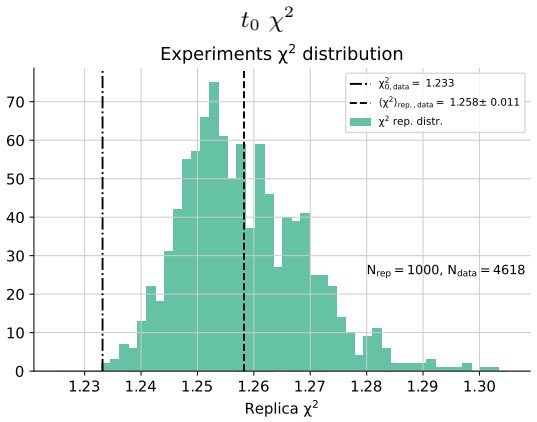




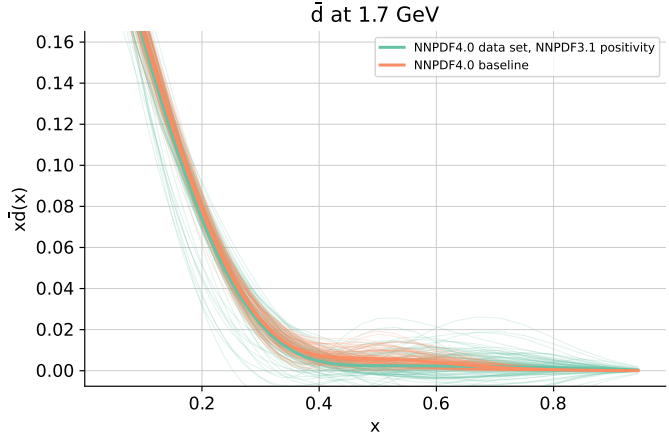
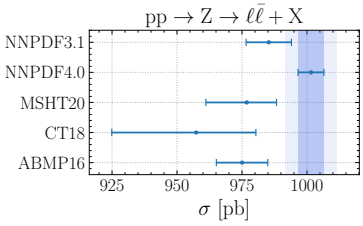
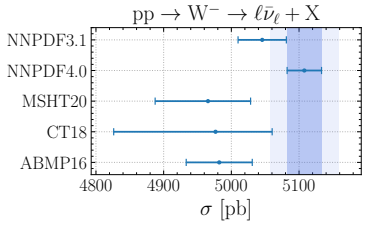
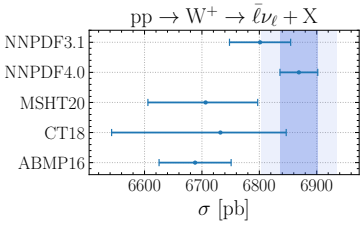


Understanding the χ^2 distribution





Impact of positivity on the PDFs



More implications for phenomenology