NNPDF4.0: Towards a high-precision Determination of the Proton Structure

Roy Stegeman

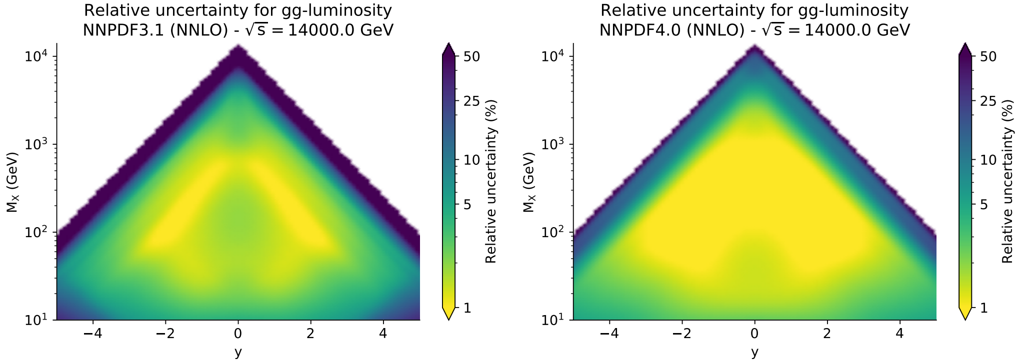
University of Milan and INFN Milan

EPS-HEP 2021, 26 July 2021

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

High-precision: gluon

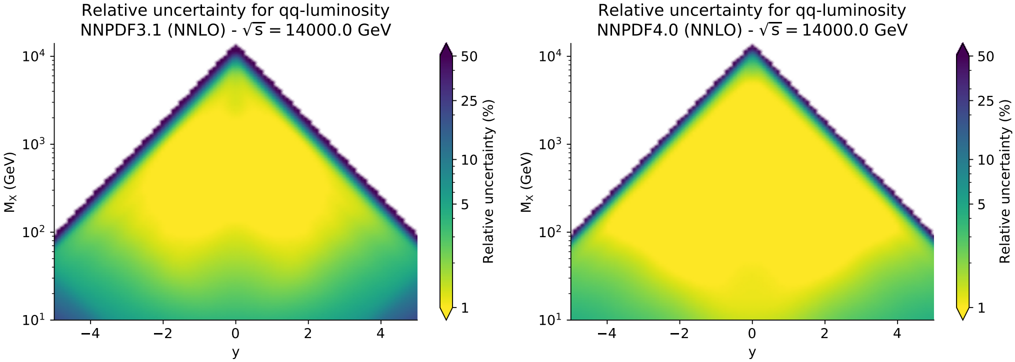
$$\mathcal{L}_{ij}\left(M_X, y, \sqrt{s}\right) = \frac{1}{s}\sum_{i,j} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$



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?

The path to NNPDF4.0

Progress towards extending data, theory and methodology

06/2017	**NNPDF3.1**	[EPJ C77 (2017) 663]
10/2017	NNPDF3.1sx: PDFs with small-x resummation	[EPJ C78 (2018) 321]
12/2017	NNPDF3.1luxQED: consistent photon PDF à la luxQED	[SciPost Phys. 5 (2018) 008]
02/2018	NNPDF3.1+ATLASphoton: inclusion of direct photon data	[EPJ C78 (2018) 470]
12/2018	NNPDF3.1alphas: α_s from a correlated-replica method	[EPJ C78 (2018) 408]
12/2018	NNPDF3.1nuc: heavy ion nuclear uncertainties in a fit	[EPJ C79 (2019) 282]
05/2019	NNPDF3.1th: missing higher-order uncertainties in a fit	[EPJ C79 (2019) 838; ibid. 931]
07/2019	Gradient descent and hyperoptimisation in PDF fits	[EPJ C79 (2019) 676]
12/2019	NNPDF3.1singletop: inclusion of single top t-channel data	[JHEP 05 (2020) 067]
05/2020	NNPDF3.1dijets: comparative study of single- and di-jets	[EPJ C80 (2020) 797]
06/2020	Positivity of $\overline{\mathrm{MS}}$ PDFs	[JHEP 11 (2020) 129]
08/2020	PineAPPL: fast evaluation of EW$\times$QCD corrections	[JHEP 12 (2020) 108]
08/2020	NNPDF3.1strangeness: assessment of strange-sensitive data	[EPJ C80 (2020) 1168]
11/2020	NNPDF3.1deu: deuteron uncertainties in a fit	[EPJ C81 (2021) 37]
03/2021	Future tests	[arXiv:2103.08606]
2021	**NNPDF4.0**	[to appear]

Experimental data in NNPDF4.0

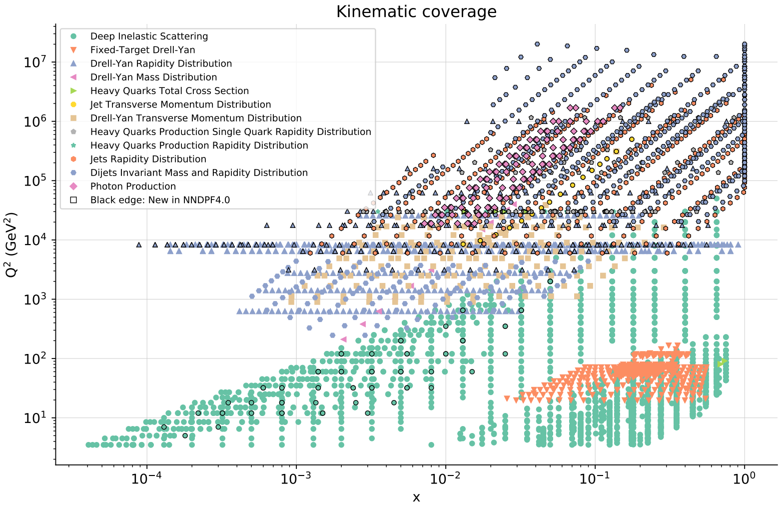
- $\mathcal{O}(35)$ datasets investigated
- $\mathcal{O}(400)$ more data points in NNPDF4.0 than in NNPDF3.1
- New data is mostly from the LHC RUN II

Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
ATLAS W, Z 7 TeV ($\mathcal{L} = 35$ pb^{-1})	[50]	✓	✓	✓	✓	✓
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ fb^{-1})	[51]	✓	✓	✗	✗	✓
ATLAS low-mass DY 7 TeV	[52]	✓	✓	✗	✗	✗
ATLAS high-mass DY 7 TeV	[53]	✓	✓	✗	✗	✓
ATLAS W 8 TeV	[84]	✗	✓	✗	✗	✓
ATLAS DY 2D 8 TeV	[83]	✗	✓	✗	✗	✓
ATLAS high-mass DY 2D 8 TeV	[82]	✗	✓	✗	✗	✓
ATLAS $\sigma_{W,Z}$ 13 TeV	[86]	✗	✓	✓	✗	✗
ATLAS W^++jet 8 TeV	[98]	✗	✓	✗	✗	✓
ATLAS Z p_T 8 TeV	[62]	✓	✓	✗	✓	✓
ATLAS $\sigma_{tt}^{\rm tot}$ 7, 8 TeV	[64]	✓	✓	✓	✗	✗
ATLAS $\sigma_{tt}^{\rm tot}$ 7, 8 TeV	[213–218]	✗	✗	✓	✗	✗
ATLAS $\sigma_{tt}^{\rm tot}$ 13 TeV ($\mathcal{L} = 3.2\,{\rm fb}^{-1}$)	[65]	✓	✗	✓	✗	✗
ATLAS $\sigma_{tt}^{\rm tot}$ 13 TeV ($\mathcal{L} = 139\,{\rm fb}^{-1}$)	[70]	✗	✓	✗	✗	✗
ATLAS $t\bar{t}$ lepton+jets 8 TeV	[66]	✓	✓	✗	✓	✓
ATLAS $t\bar{t}$ dilepton 8 TeV	[94]	✗	✓	✗	✗	✓
ATLAS single-inclusive jets 7 TeV, R=0.6	[73]	✓	✗	✗	✓	✓
ATLAS single-inclusive jets 8 TeV, R=0.6	[91]	✗	✓	✗	✗	✗
ATLAS dijets 7 TeV, R=0.6	[146]	✗	✓	✗	✗	✗
ATLAS direct photon production 13 TeV	[106]	✗	✓	✗	✗	✗
ATLAS single top R_t 7, 8, 13 TeV	[99, 101, 103]	✗	✓	✓	✗	✗
ATLAS single top diff. 7 TeV	[99]	✗	✓	✗	✗	✗
ATLAS single top diff. 8 TeV	[103]	✗	✓	✗	✗	✗

Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
CMS W asym. 7 TeV ($\mathcal{L} = 36\,{\rm pb}^{-1}$)	[219]	✗	✗	✗	✗	✓
CMS Z 7 TeV ($\mathcal{L} = 36\,{\rm pb}^{-1}$)	[220]	✗	✗	✗	✗	✓
CMS W electron asymmetry 7 TeV	[54]	✓	✓	✗	✓	✓
CMS W muon asymmetry 7 TeV	[55]	✓	✓	✓	✓	✗
CMS Drell-Yan 2D 7 TeV	[56]	✓	✓	✗	✗	✓
CMS W rapidity 8 TeV	[57]	✓	✓	✓	✓	✓
CMS Z p_T 8 TeV	[63]	✓	✓	✗	✗	✗
CMS $W + c$ 7 TeV	[80]	✓	✓	✗	✗	✓
CMS $W + c$ 13 TeV	[143]	✗	✓	✗	✗	✗
CMS single-inclusive jets 2.76 TeV	[75]	✓	✗	✗	✗	✓
CMS single-inclusive jets 7 TeV	[145]	✓	✗	✗	✓	✓
CMS dijets 7 TeV	[74]	✗	✗	✗	✗	✗
CMS single-inclusive jets 8 TeV	[92]	✗	✓	✗	✓	✓
CMS 3D dijets 8 TeV	[147]	✗	✓	✗	✗	✗
CMS $\sigma_{tt}^{\rm tot}$ 5 TeV	[93]	✗	✓	✗	✗	✗
CMS $\sigma_{tt}^{\rm tot}$ 7, 8 TeV	[144]	✓	✓	✗	✗	✗
CMS $\sigma_{tt}^{\rm tot}$ 8 TeV	[221]	✗	✗	✗	✗	✓
CMS $\sigma_{tt}^{\rm tot}$ 5, 7, 8, 13 TeV	[67, 222–230]	✗	✗	✓	✗	✗
CMS $\sigma_{tt}^{\rm tot}$ 13 TeV	[68]	✓	✓	✓	✗	✗
CMS $t\bar{t}$ lepton+jets 8 TeV	[69]	✓	✓	✗	✗	✓
CMS $t\bar{t}$ 2D dilepton 8 TeV	[95]	✗	✓	✗	✓	✓
CMS $t\bar{t}$ lepton+jet 13 TeV	[96]	✗	✓	✗	✗	✗
CMS $t\bar{t}$ dilepton 13 TeV	[97]	✗	✓	✗	✗	✗
CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV	[100]	✗	✓	✓	✗	✗
CMS single top R_t 8, 13 TeV	[102, 104]	✗	✓	✓	✗	✗

Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
LHCb Z 7 TeV ($\mathcal{L} = 940\,{\rm pb}^{-1}$)	[58]	✓	✓	✗	✗	✓
LHCb $Z \to ee$ 8 TeV ($\mathcal{L} = 2\,{\rm fb}^{-1}$)	[60]	✓	✓	✓	✓	✓
LHCb W 7 TeV ($\mathcal{L} = 37\,{\rm pb}^{-1}$)	[231]	✗	✗	✗	✗	✓
LHCb $W, Z \to \mu$ 7 TeV	[81]	✓	✓	✓	✓	✓
LHCb $W, Z \to \mu$ 8 TeV	[61]	✓	✓	✓	✓	✓
LHCb $Z \to \mu\mu, ee$ 13 TeV	[87]	✗	✓	✗	✗	✗

Experimental data in NNPDF4.0



New processes:

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

Theoretical improvement

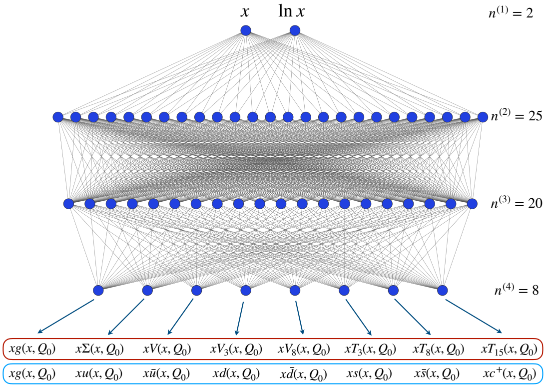
Nuclear uncertainties are included

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 [JHEP 11 (2020) 129]
- Integrability of nonsinglet distributions (Gottfried sum rules)



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

Different strategies to parametrize the quark PDF flavour combinations lead to **identical results**

Automated model selection

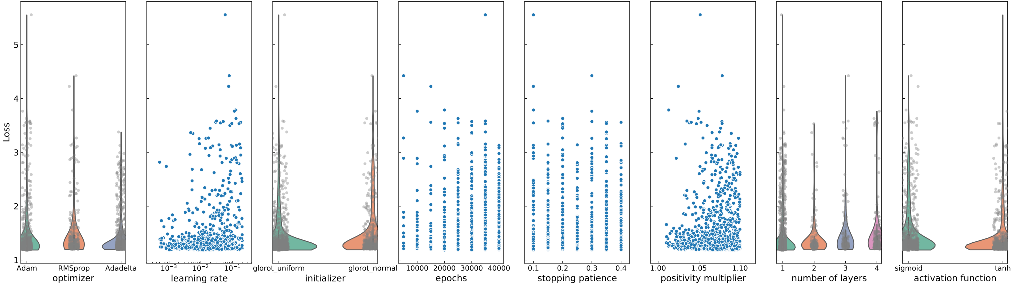
NNPDF aims to minimize sources of bias in the PDF:

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

Automated model selection

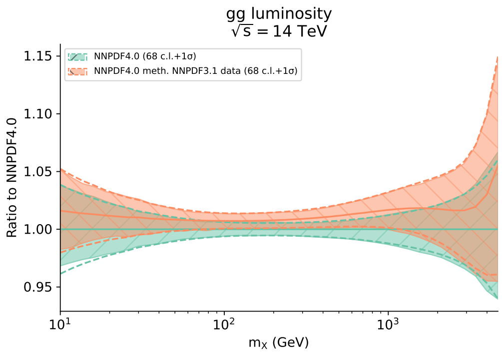
NNPDF aims to minimize sources of bias in the PDF:

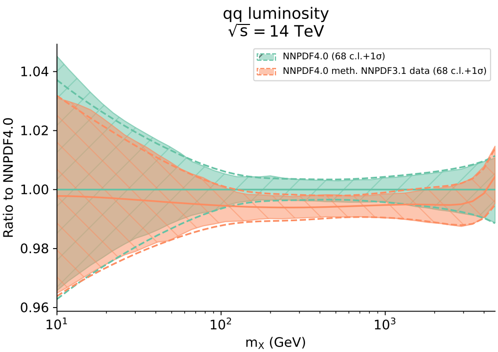
- Functional form → Neural Network
- Model parameters → **Hyperoptimization**



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**

Impact of the new data

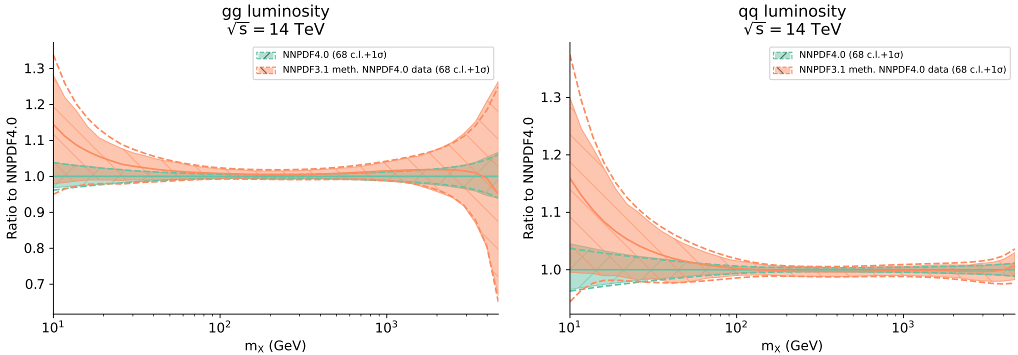




Individual datasets have a limited impact, but collectively they result in:

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

Impact of the new fitting methodology

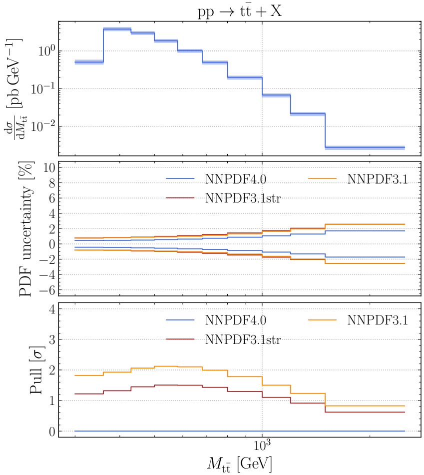


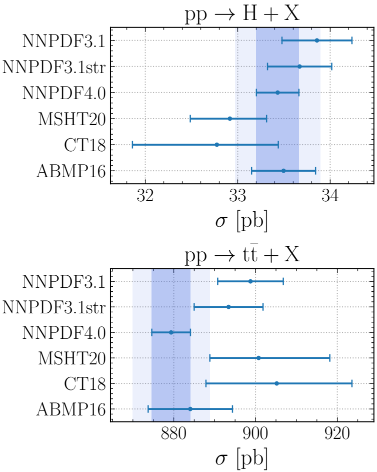
- Significant reduction of PDF uncertainties
- Good agreement between the central values

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

Implications for phenomenology

Reduced luminosity uncertainties → Reduced uncertainty at the level of observables





Summary

- Added $\mathcal{O}(400)$ new data points from many new processes
- 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 **NNPDF code** will be made **publicly available** along with user-friendly documentation

Summary

- Added $\mathcal{O}(400)$ new data points from many new processes
- Improved methodology with Stochastic Gradient Descent and hyperoptimization
- Validation of PDF uncertainties using closure test, future test and parametrization basis independence

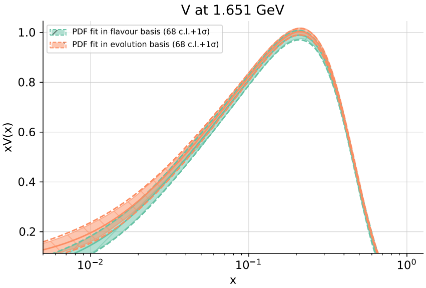
$\Rightarrow$ NNPDF4.0 achieves a high precision over a broad kinematic range

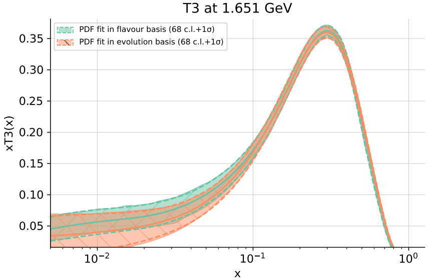
The **NNPDF code** will be made **publicly available** along with user-friendly documentation

Thank you!

Backup

Parametrization basis independence





Evolution Basis:

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

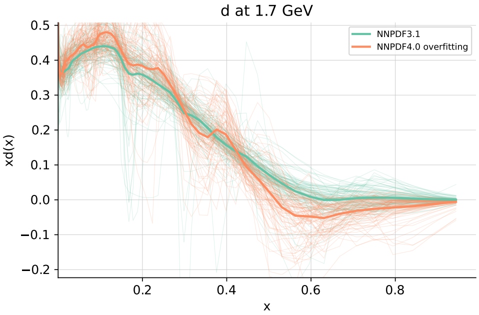
Different strategies to parametrize the quark PDF flavour combinations lead to **identical results**

Flavour Basis:

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

Hyperoptimization: the reward function

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



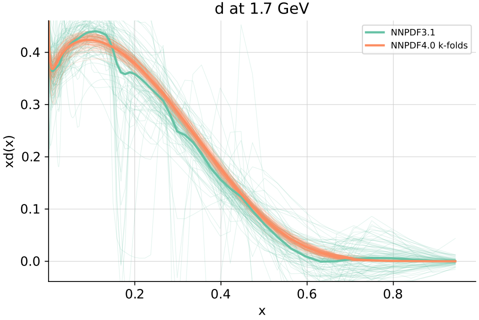
Hyperoptimization: the reward function

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

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

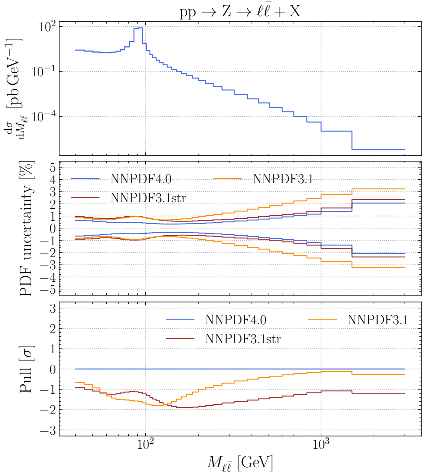
1. Divide the data into k representative subsets
2. Fit $k-1$ sets and use k-th as test set
   $\Rightarrow$ k values of χ^2_{test}
3. Optimize the average χ^2_{test} of the k test sets

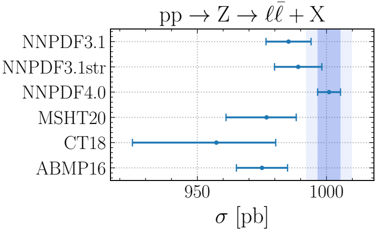
$\Rightarrow$ The hyperoptimization target is not based on data that entered the fit.



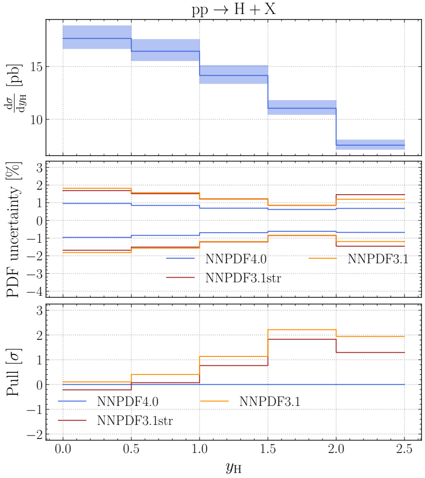
- No overfitting
- Compared to NNPDF3.1:
  - Increased stability
  - Reduced uncertainties

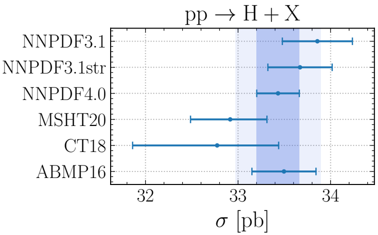
More implications for phenomenology



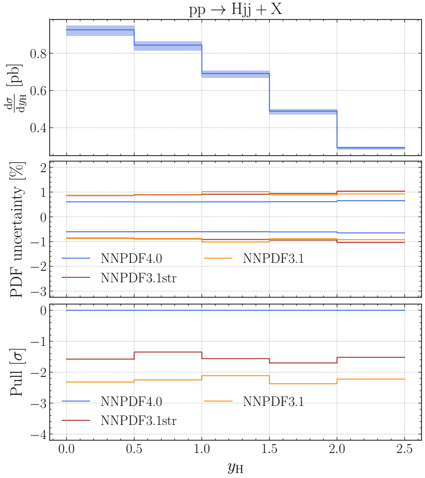


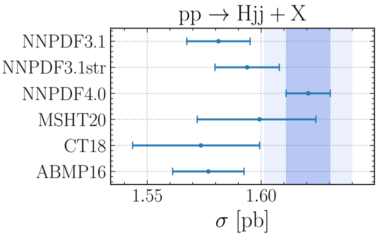
More implications for phenomenology



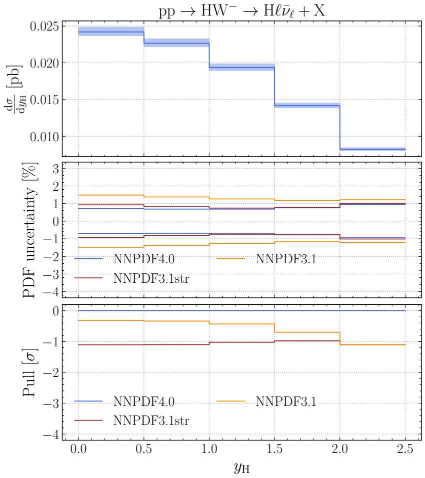


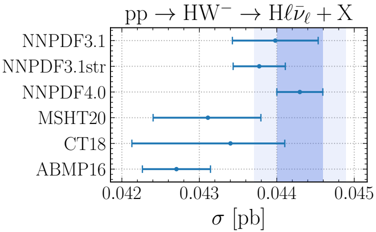
More implications for phenomenology



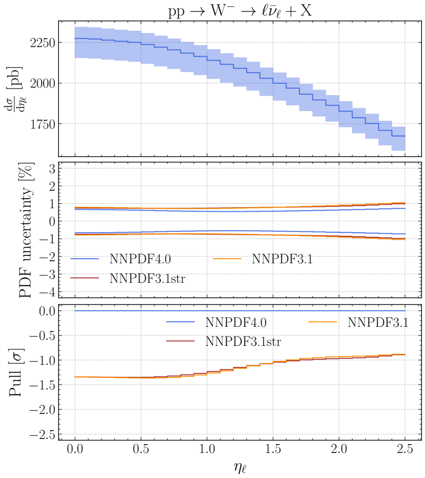


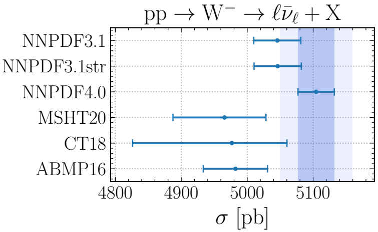
More implications for phenomenology



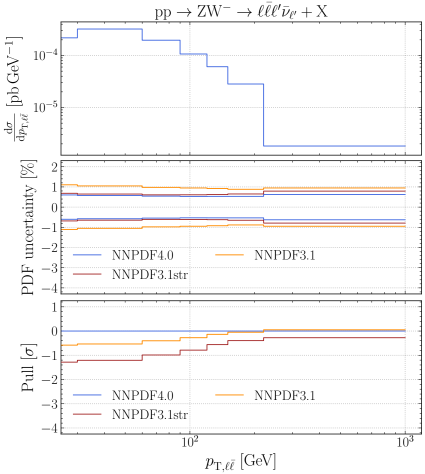


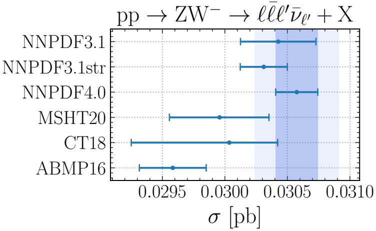
More implications for phenomenology



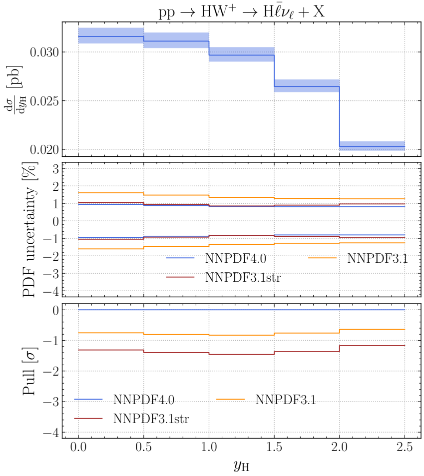


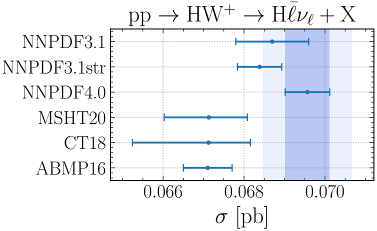
More implications for phenomenology



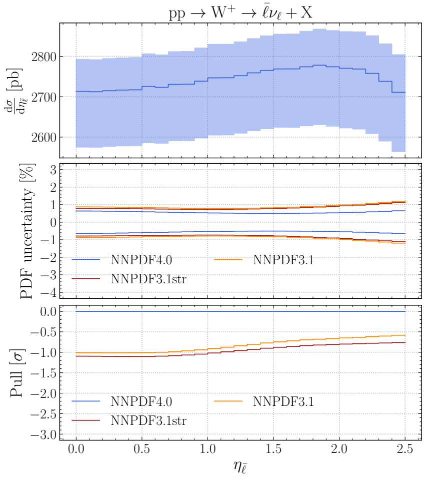


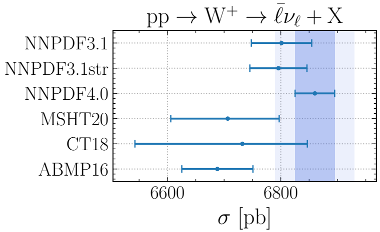
More implications for phenomenology



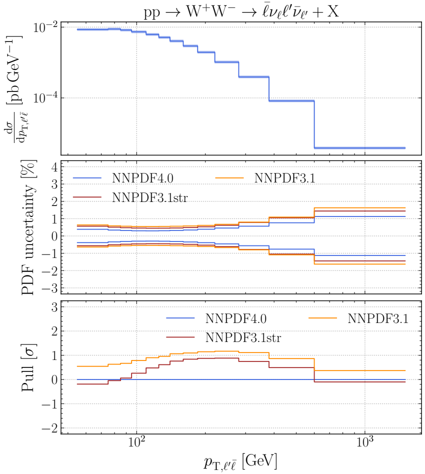


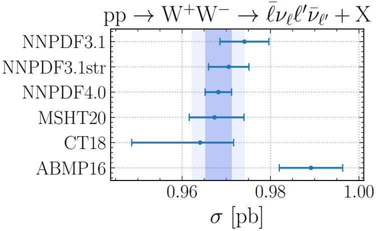
More implications for phenomenology





More implications for phenomenology





More implications for phenomenology