Machine Learning for PDF determination

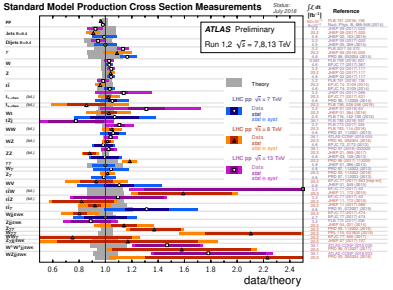
Roy Stegeman
Supervisor: Stefano Forte

University of Milan and INFN Milan

University of Milan, 19 December 2022

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

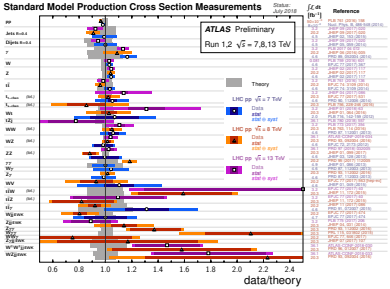
Precision physics at the LHC



SM cross-sections [ATLAS, PRD 87, 112003 (2013)]

- Currently **good agreement** between data and theory
- Observing deviations requires **sub-percent** accuracy

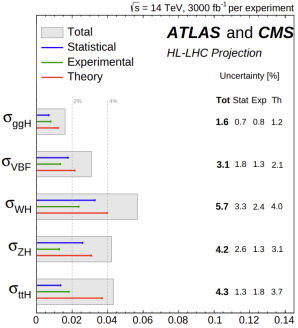
Precision physics at the LHC

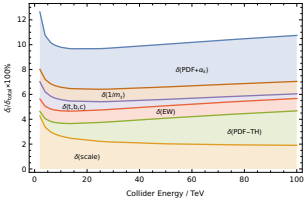


SM cross-sections [ATLAS, PRD 87, 112003 (2013)]

- Currently **good agreement** between data and theory
- Observing deviations requires **sub-percent** accuracy

PDFs uncertainties is often the **dominant** source of uncertainty for LHC cross-sections

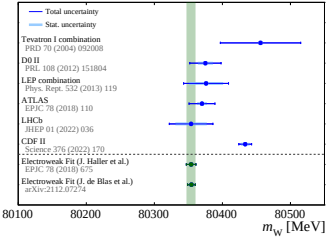




Projection for HL-LHC [HL/HE-LHC WG2, 2019]

Importance of faithful uncertainties

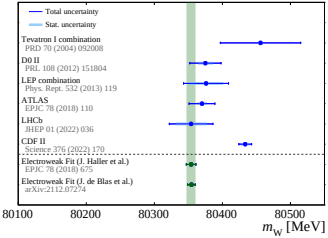
- This years' m_W determination highlights the importance of **understanding all sources of uncertainty**
- Understand apparent **discrepancies** between measurements
- Important that predictions are **accurate**



SM parameters: m_W [LHCb-FIGURE-2022-003]

Importance of faithful uncertainties

- This years' m_W determination highlights the importance of **understanding all sources of uncertainty**
- Understand apparent **discrepancies** between measurements
- Important that predictions are **accurate**
- Important role for **PDF uncertainties**

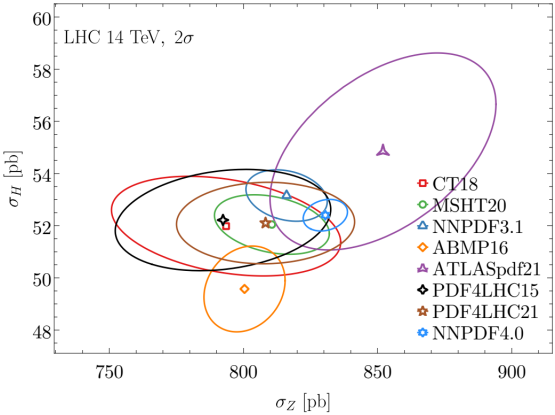


SM parameters: m_W [LHCb-FIGURE-2022-003]

Source	Size [MeV]
Parton distribution functions	9
Theory (excl. PDFs) total	17
Transverse momentum model	11
Angular coefficients	10
QED FSR model	7
Additional electroweak corrections	5
Experimental total	10
Momentum scale and resolution modelling	7
Muon ID, trigger and tracking efficiency	6
Isolation efficiency	4
QCD background	2
Statistical	23
Total	32

SM parameters: m_W [JHEP01(2022)036]

Status of modern PDF sets

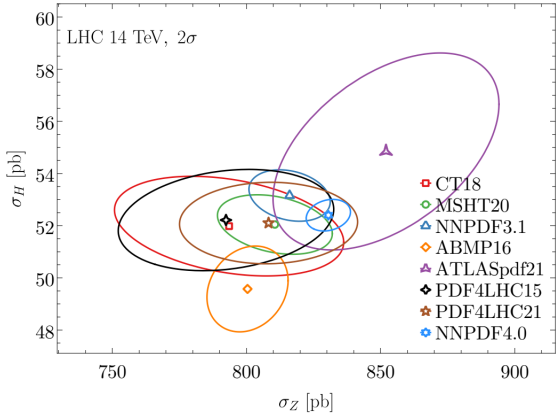


[Snowmass (2022), 2203.13923]

- PDF predictions are **consistent** but with **different uncertainties**
- Ingredients of a PDF fit:
  - Data
  - Theory
  - Methodology

Main difference between fitting groups is the methodology

Status of modern PDF sets



[Snowmass (2022), 2203.13923]

- PDF predictions are **consistent** but with **different uncertainties**
- Ingredients of a PDF fit:
  - Data
  - Theory
  - Methodology

Main difference between fitting groups is the methodology

"Does the NNPDF methodology produce faithful uncertainties?"

Particularly relevant today!

Parton distributions need representative sampling

Aurore Courtoy,[1,*] Joey Huston,[2,†] Pavel Nadolsky,[3,‡] Keping Xie,[4,§] Menghsi Yan,[5,¶] and C.-P. Yuan[2,**]
[1] *Instituto de Física, Universidad Nacional Autónoma de México, Apartado Postal 20-364, 01000 Ciudad de México, Mexico*
[2] *Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA*
[3] *Department of Physics, Southern Methodist University, Dallas, TX 75275-0181, USA*
[4] *Pittsburgh Particle Physics, Astrophysics, and Cosmology Center, Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA*
[5] *School of Physics and State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing 100871, China*
(Dated: September 27, 2022)

In global QCD fits of parton distribution functions (PDFs), a large part of the estimated uncertainty on the PDFs originates from the choices of parametric functional forms and fitting methodology. We argue that these types of uncertainties can be underestimated with common PDF ensembles in high-stake measurements at the Large Hadron Collider and Tevatron. A fruitful approach to quantify these uncertainties is to view them as arising from sampling of allowed PDF solutions in a multidimensional parametric space. This approach applies powerful insights gained in recent statistical studies of large-scale population surveys and quasi-Monte Carlo integration methods. In particular, PDF fits may be affected by the big data paradox, which stipulates that more experimental data do not automatically raise the accuracy of PDFs – close attention to the data quality and sampling of possible PDF solutions is as essential. To test if the sampling of the PDF uncertainty of an experimental observable is truly representative of all acceptable solutions, we introduce a technique ("a hopscotch scan") based on a combination of parameter scans and stochastic sampling. With this technique, we show that the PDF uncertainty on key LHC cross sections at 13 TeV obtained with the public NNPDF4.0 fitting code is larger than the nominal uncertainty obtained with the published NNPDF4.0 Monte-Carlo replica sets. For example, the uncertainties on the charm distribution at a large momentum fraction x and gluon PDF at small x are enlarged. In PDF ensembles obtained in the analytic minimization (Hessian) formalism, the tolerance on the PDF uncertainty must be based on sufficiently complete sampling of PDF functional forms and choices of the experiments.

May 2022 [2205.10444]

Edinburgh 2022/19
TIF-UNIMI-2022-21

Response to "Parton distributions need representative sampling"

The NNPDF Collaboration:
Richard D. Ball,[1] Juan Cruz-Martinez,[2] Luigi Del Debbio,[1] Stefano Forte,[3]
Zahari Kassabov,[4] Emanuele R. Nocera,[5] Juan Rojo,[6,7] Roy Stegeman,[1] Maria Ubiali[4]

[1] *The Higgs Centre for Theoretical Physics, University of Edinburgh, JCMB, KB, Mayfield Rd, Edinburgh EH9 3JZ, Scotland*
[2] *CERN, Theoretical Physics Department, CH-1211 Geneva 23, Switzerland*
[3] *Tif Lab, Dipartimento di Fisica, Università di Milano and INFN, Sezione di Milano, Via Celoria 16, I-20133 Milano, Italy*
[4] *DAMTP, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, United Kingdom*
[5] *Dipartimento di Fisica, Università degli Studi di Torino and INFN, Sezione di Torino, Via Pietro Giuria 1, I-10125 Torino, Italy*
[6] *Department of Physics and Astronomy, Vrije Universiteit, NL-1081 HV Amsterdam*
[7] *Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands*

November 2022 [2211.12961]

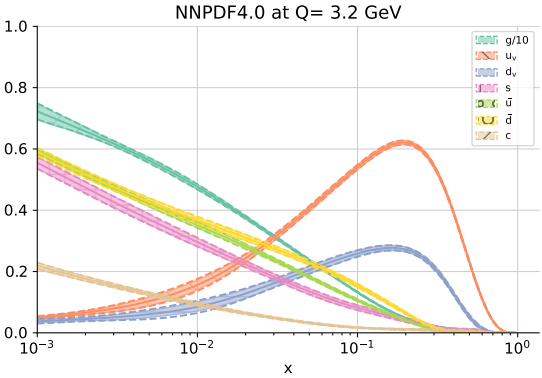
Outline

NNPDF4.0

2109.02653 and 2109.02671

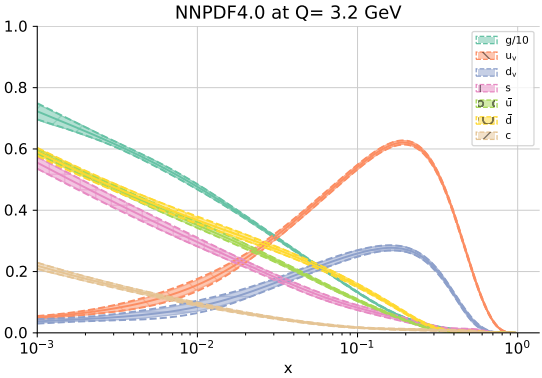
PDFs from data

- Evaluating LHC cross-sections:
  - $\sigma_{ab} = \sum_{ab} f_a \otimes f_b \otimes \hat{\sigma}_{ab}$
  - PDF f_a of flavor a (non-perturbative, **from data**)
  - hard-scattering matrix element $\hat{\sigma}_{ab}$ (perturbative QCD)
  - $\otimes$ denotes a convolution over momentum fraction x
- PDFs f_a depends only on x and Q^2
- Other kinematic variables in $\hat{\sigma}$



PDFs from data

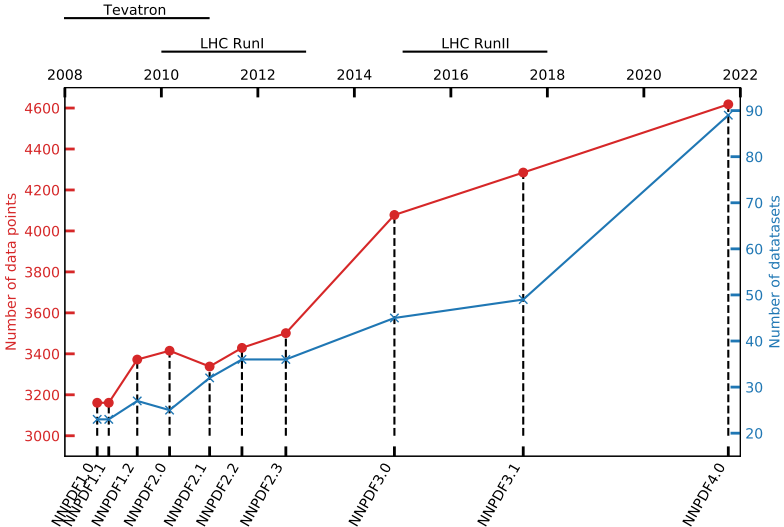
- Evaluating LHC cross-sections:
  - $\sigma_{ab} = \sum_{ab} f_a \otimes f_b \otimes \hat{\sigma}_{ab}$
  - PDF f_a of flavor a (non-perturbative, **from data**)
  - hard-scattering matrix element $\hat{\sigma}_{ab}$ (perturbative QCD)
  - $\otimes$ denotes a convolution over momentum fraction x
- PDFs f_a depends only on x and Q^2
- Other kinematic variables in $\hat{\sigma}$



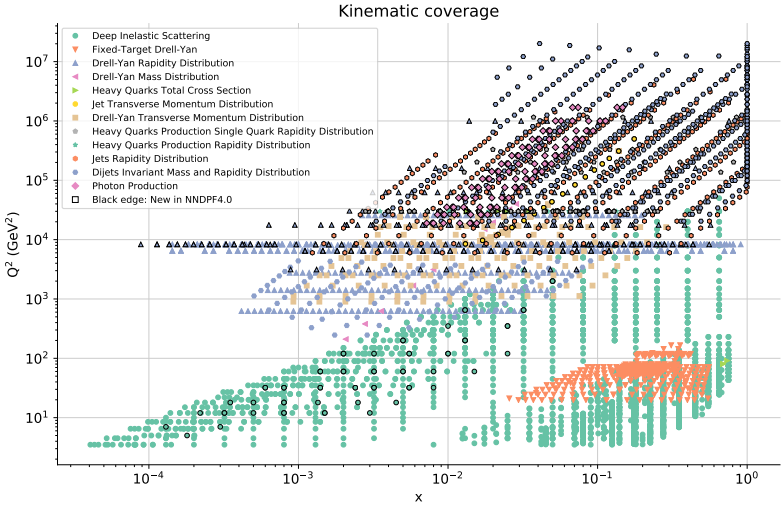
The problem

- Given a dataset D, determine $p(f|D)$ in the space of PDFs $f : [0, 1] \to \mathbb{R}$
- PDFs are multivariate probability distributions in **infinite dimensional** space
- However data is **discrete**

Data in NNPDF4.0



Data in NNPDF4.0



More than 4000 datapoints and 13 processes!

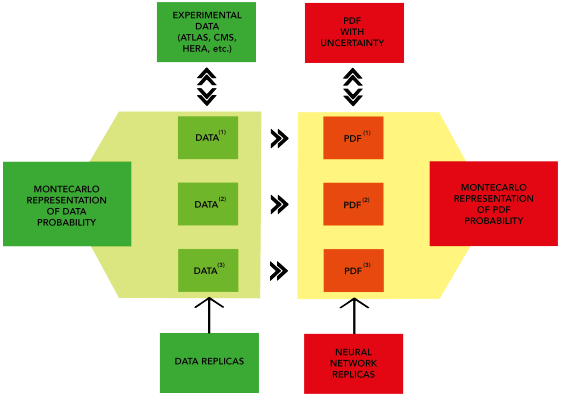
New processes:

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

The NNPDF approach: probabilities in a space of function

Data is fully defined by central values μ_i and covariance matrix cov_{ij}

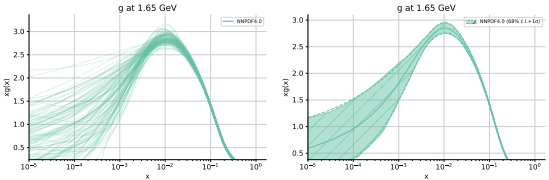
1. Generate $N_{\rm rep}$ Monte Carlo data "replicas" $\hat{\mu}_i$ such that as $N_{\rm rep} \to \infty$
   $\mu_i = \frac{1}{N_{\rm rep}} \sum_{i=1}^{N_{\rm rep}} \hat{\mu}_i$
   $\mathrm{cov}_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$
2. Perform a PDF fit to each replica
3. Compute observables X and their uncertainties
   $\langle X\left[f\right]\rangle = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N} X\left[f^{(r)}\right]$
   $\mathrm{Var}\left[X\left[f\right]\right] = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N_{\rm rep}} \left(X\left[f^{(r)}\right] - \langle X\left[f\right]\rangle\right)^2$

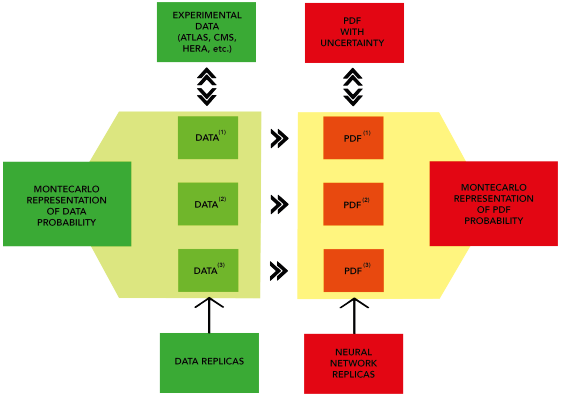


The NNPDF approach: probabilities in a space of function

Data is fully defined by central values μ_i and covariance matrix cov_{ij}

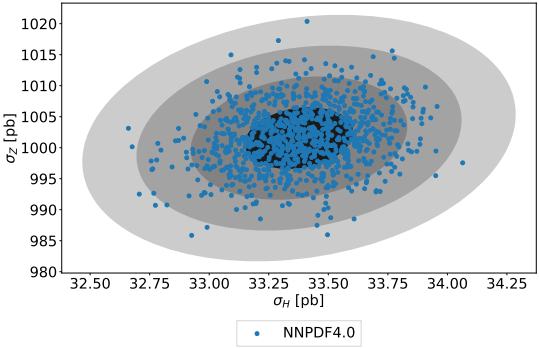
1. Generate $N_{\rm rep}$ Monte Carlo data "replicas" $\hat{\mu}_i$ such that as $N_{\rm rep} \to \infty$
   $\mu_i = \frac{1}{N_{\rm rep}} \sum_{i=1}^{N_{\rm rep}} \hat{\mu}_i$
   $\mathrm{cov}_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$
2. Perform a PDF fit to each replica
3. Compute observables X and their uncertainties
   $\langle X\left[f\right]\rangle = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N} X\left[f^{(r)}\right]$
   $\mathrm{Var}\left[X\left[f\right]\right] = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N_{\rm rep}} \left(X\left[f^{(r)}\right] - \langle X\left[f\right]\rangle\right)^2$





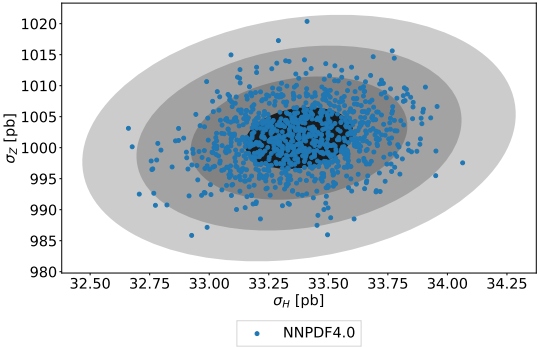
The NNPDF approach: an importance sampling

- Importance sampling produces Gaussian posterior distribution
- All replicas equally likely



The NNPDF approach: an importance sampling

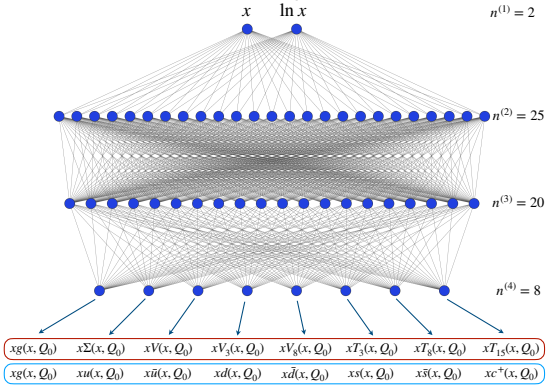
- Importance sampling produces Gaussian posterior distribution
- All replicas equally likely



How did we obtain this distribution?

Parametrization

Neural network: universal interpolator



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

Physical constraints:

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

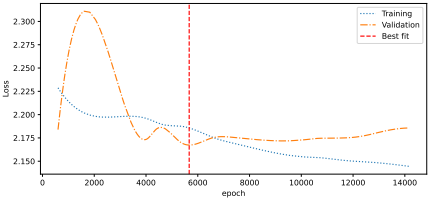
- Train by minimizing χ^2 loss function comparing data to prediction

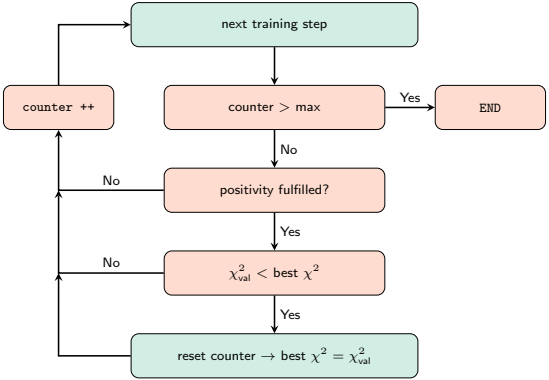
Training the neural network

- Optimize using a gradient descent algorithm
- NN should generalize the underlying law, but if trained to long noise is fitted

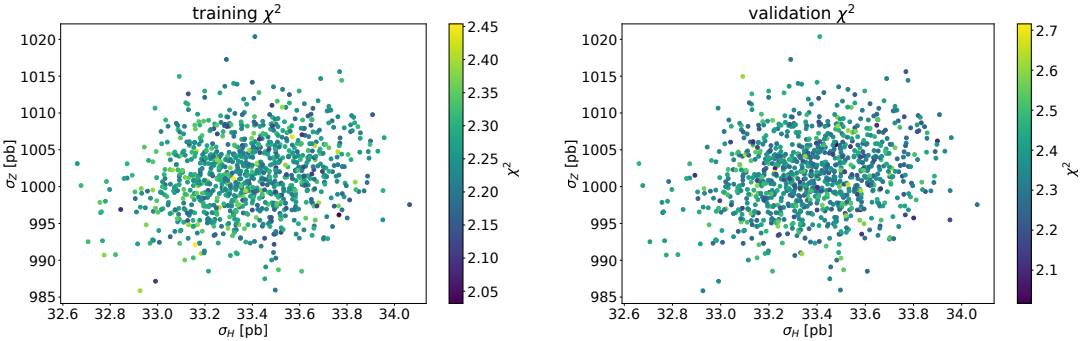
Cross-validation

- Divide data into training and validation
- Minimize training χ^2
- Stop if validation χ^2 no longer improves





Verify the importance sampling assumption



- All PDF replicas are fitted equally well to their data replica
- Thus outliers correspond to unlikely data replicas

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:

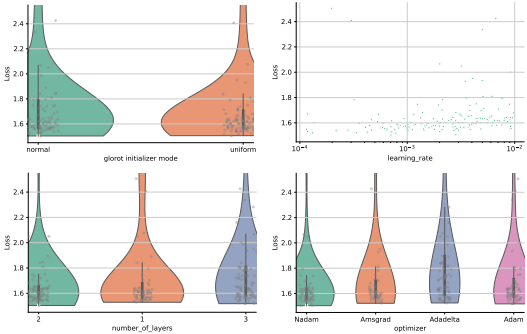
- Functional form → Neural Network
- Model parameters → **Hyperoptimization**

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**

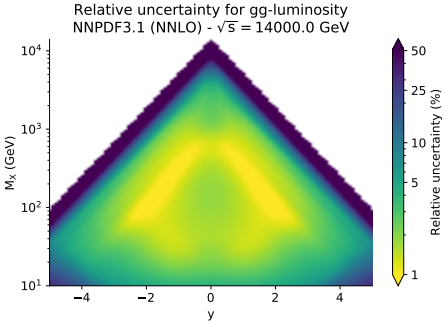
Objective function:
$L = \mathrm{mean}(\chi_1^2, \chi_3^2, \chi_2^2, \ldots, \chi_k^2)$

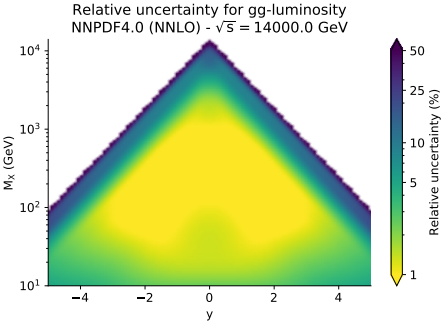
Final step requires human input



High-precision: gluon

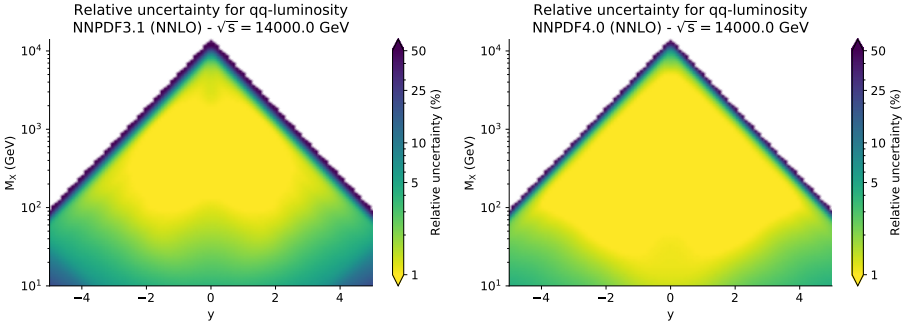
$$\mathcal{L}_{ij}\left(M_X, y, \sqrt{s}\right) = \frac{1}{s}\sum_{i,j} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$





High-precision: singlet

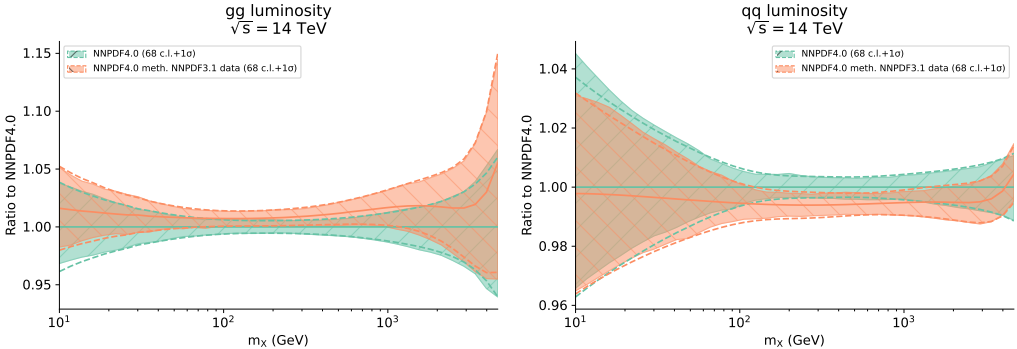
$$\mathcal{L}_{ij}\left(M_X, y, \sqrt{s}\right) = \frac{1}{s}\sum_{i,j} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$



Typical uncertainties in the data region:

- singlet: from $\sim 3\%$ to $\sim 1\%$
- nonsinglet: from $\sim 5\%$ to $\sim 2\%$

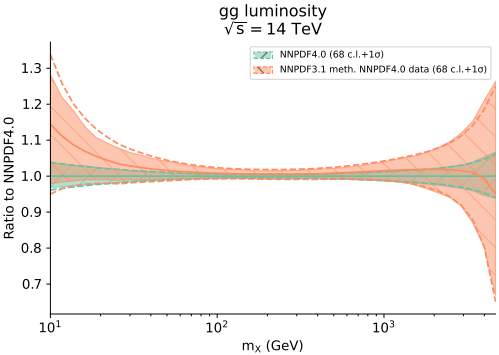
Impact of the new data

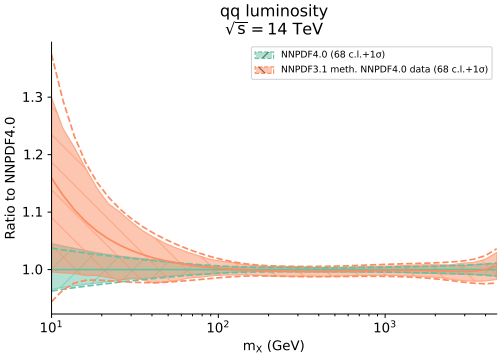


Individual datasets have a limited impact, but collectively they result in:

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

Impact of the new fitting methodology



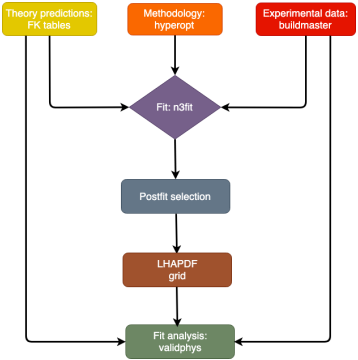


- Significant reduction of PDF uncertainties
- Good agreement between the central values

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

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`

PDF correlations

2110.08274

PDF correlations: definition and application

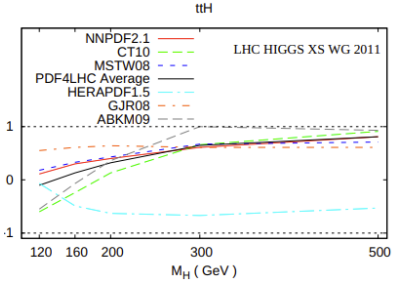
Definition:

Covariance:
$$\mathrm{Cov}\left[f_a, f_b\right](x,x') = E\left[f_a\left(x,Q_0^2\right) f_b\left(x',Q_0^2\right)\right] - E\left[f_a\left(x,Q_0^2\right)\right] E\left[f_b\left(x',Q_0^2\right)\right]$$

Correlation:
$$\rho\left[f_a, f_b\right](x,x') = \frac{\mathrm{Cov}[f_a,f_b](x,x')}{\sqrt{\mathrm{Var}[f_a](x)\,\mathrm{Var}[f_a^q](x')}}$$

In the MC approach this can be calculated using
$$E\left[f_a\left(x,Q_0^2\right) f_b\left(x',Q_0^2\right)\right] = \frac{1}{N}\sum_{r=1}^{N} f_a^{(r)}\left(x,Q_0^2\right) f_b^{(r)}\left(x',Q_0^2\right)$$

Correlation **induced by** data, theory (e.g. sumrules), and methodology (e.g. preprocessing).



Application: determination of PDF induced signal-background correlation between e.g. ggH and ttH

Correlations between different PDF sets

General **cross-covariance** between PDFs in **different sets**, e.g. NNPDF4.0 and MSHT20:

$$\mathrm{Cov}\left[f_a^{\mathrm{NNPDF}}, f_b^{\mathrm{MSHT}}\right](x,x') = E\left[f_a^{\mathrm{NNPDF}}\left(x,Q_0^2\right) f_b^{\mathrm{MSHT}}\left(x',Q_0^2\right)\right] - E\left[f_a^{\mathrm{NNPDF}}\left(x,Q_0^2\right)\right] E\left[f_b^{\mathrm{MSHT}}\left(x',Q_0^2\right)\right]$$

Special cases of cross-correlation:

- F-correlation - same PDF, different flavor $\rho\left[f_a^{\mathrm{NNPDF}}, f_b^{\mathrm{NNPDF}}\right]$
- S-correlation - different PDF, same flavor $\rho\left[f_a^{\mathrm{NNPDF}}, f_a^{\mathrm{MSHT}}\right]$

The same replica must be used when calculating covariance

If $f^{(r)}$ and $f^{(r')}$ are uncorrelated, covariance vanishes:
$E\left[f_a f_b\right] = \frac{1}{N}\sum_{r=1}^{N} f_a^{(r)} f_b^{(r')} = \left[f_a\right]\left[f_b\right]$

Problem: What is the meaning of "same replica" across PDF sets?

Correlations between different PDF sets

General **cross-covariance** between PDFs in **different sets**, e.g. NNPDF4.0 and MSHT20:

$$\mathrm{Cov}\left[f_a^{\mathrm{NNPDF}}, f_b^{\mathrm{MSHT}}\right](x,x') = E\left[f_a^{\mathrm{NNPDF}}\left(x,Q_0^2\right) f_b^{\mathrm{MSHT}}\left(x',Q_0^2\right)\right] - E\left[f_a^{\mathrm{NNPDF}}\left(x,Q_0^2\right)\right] E\left[f_b^{\mathrm{MSHT}}\left(x',Q_0^2\right)\right]$$

Special cases of cross-correlation:

- F-correlation - same PDF, different flavor $\rho\left[f_a^{\mathrm{NNPDF}}, f_b^{\mathrm{NNPDF}}\right]$
- S-correlation - different PDF, same flavor $\rho\left[f_a^{\mathrm{NNPDF}}, f_a^{\mathrm{MSHT}}\right]$

The same replica must be used when calculating covariance

If $f^{(r)}$ and $f^{(r')}$ are uncorrelated, covariance vanishes:
$E\left[f_a f_b\right] = \frac{1}{N}\sum_{r=1}^{N} f_a^{(r)} f_b^{(r')} = \left[f_a\right]\left[f_b\right]$

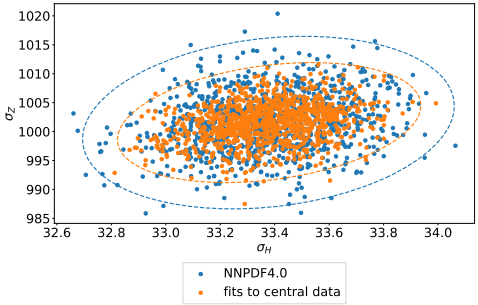
Problem: What is the meaning of "same replica" across PDF sets?

Possible solution: PDF replicas fitted to the same data replica

- Fit $f^{\mathrm{MSHT}(r)}$ and $f^{\mathrm{NNPDF}(r)}$ to the same data replica r
- Calculate covariance as $E\left[f_a f_b\right] = \frac{1}{N}\sum_{r=1}^{N} f_a^{\mathrm{NNPDF}(r)} f_b^{\mathrm{MSHT}(r)}$

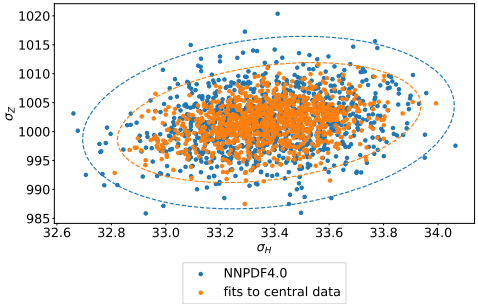
Uncorrelated methodological aspects

The data replica **does not** uniquely determine the PDF replica



Uncorrelated methodological aspects

The data replica **does not** uniquely determine the PDF replica



PDF depends on uncorrelated methodological aspects:

- initialization of the neural network
- preprocessing exponents
- training/validation mask
- ...

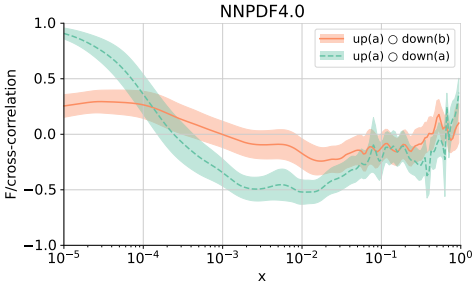
Data-induced correlation

Let us distinguish

- data replicas r
- methodological replicas r'

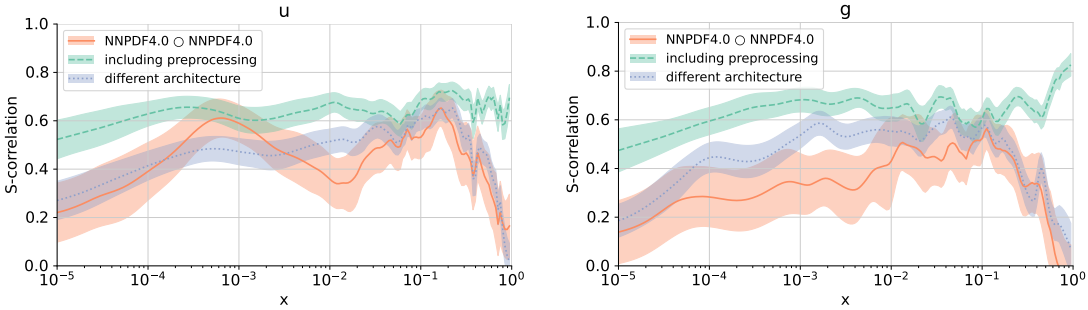
$$\left| \frac{1}{N} \sum_{r=1}^{N} f_a^{(r,r')} f_b^{(r,r'')} - E\left[f_a\right] E\left[f_b\right] \right| \leq \left| \frac{1}{NM} \sum_{r=1}^{N} \sum_{r'=1}^{M} f_a^{(r,r')} f_b^{(r,r')} - E\left[f_a\right] E\left[f_b\right] \right|$$

Only **data-induced** contributions are calculated



Correlated methodological replicas

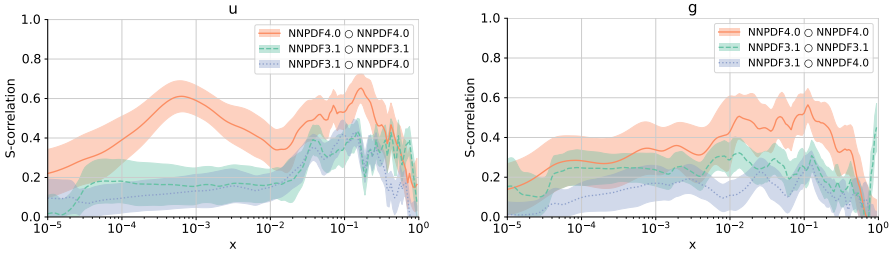
- Can easily be done for **parametric** components such as preprocessing or architecture
- Noticeable impact of **preprocessing**, negligible for **architecture**



Non-trivial for **non-parametric** aspects of the methodology

Data-induced self-correlation

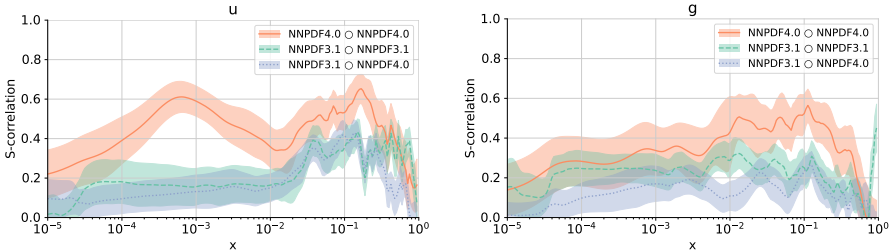
- Correlation **between two determinations** of the same PDF set (e.g. NNPDF4.0)
- The **deviation from 100%** correlation is due to uncorrelated aspects of the methodology



Data-induced self-correlation

- Correlation **between two determinations** of the same PDF set (e.g. NNPDF4.0)
- The **deviation from 100%** correlation is due to uncorrelated aspects of the methodology
- Calculate **correlation between different methodologies**
- "weakest link" NNPDF3.1∘NNPDF4.0 ≈ NNPDF3.1∘NNPDF3.1

Higher correlation indicates a more efficient methodology



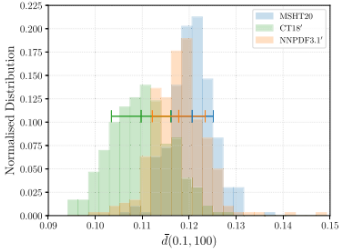
PDF4LHC combination

OPEN ACCESS
IOP Publishing
Journal of Physics G: Nuclear and Particle Physics
J. Phys. G: Nucl. Part. Phys. **49** (2022) 080501 (94pp)
https://doi.org/10.1088/1361-6471/ac7216

Major Report

The PDF4LHC21 combination of global PDF fits for the LHC Run III*

- All PDFs assumed to have **equal probability**
- **Monte Carlo combination** of 300 replicas from each of the underlying PDF sets



Uncertainty of the Monte Carlo combination:
$\mathrm{Var}[f_a^{\mathrm{comb}}] = \frac{1}{2}\left(\mathrm{Var}[f_a^{\mathrm{MSHT}}] + \mathrm{Var}[f_a^{\mathrm{NNPDF}}]\right) + \frac{1}{4}(E\left[f_a^{\mathrm{MSHT}}\right] - E\left[f_a^{\mathrm{NNPDF}}\right])^2$
combined uncertainty bigger than average uncertainty if central values disagree

Correlated PDF combination?

Idea:

- Combine PDF determinations as independent observations
- Correlated combination produces a **weighted average**

Assuming two sets of same variance, $\mathrm{Var}[f_a^{\mathrm{NNPDF}}] = \mathrm{Var}[f_a^{\mathrm{MSHT}}]$:

$$\mathrm{Var}[f_a^{\mathrm{comb}}] = \frac{1}{2}\left(1 + \rho[f_a^{\mathrm{MSHT}}, f_a^{\mathrm{NNPDF}}]\right)\mathrm{Var}[f_a^{\mathrm{NNPDF}}]$$

Correlated PDF combination?

Idea:

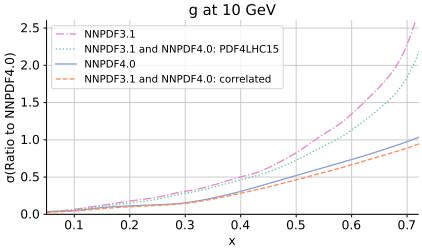
- Combine PDF determinations as independent observations
- Correlated combination produces a **weighted average**

Assuming two sets of same variance, $\mathrm{Var}[f_a^{\mathrm{NNPDF}}] = \mathrm{Var}[f_a^{\mathrm{MSHT}}]$:

$$\mathrm{Var}[f_a^{\mathrm{comb}}] = \frac{1}{2}\left(1 + \rho[f_a^{\mathrm{MSHT}}, f_a^{\mathrm{NNPDF}}]\right)\mathrm{Var}[f_a^{\mathrm{NNPDF}}]$$

Problems:

- Does not account for difference in central values
- How to compute correlations reliably? I.e. methodological components



Underestimated data-induced correlations leads to **underestimated uncertainty of combination**

Correlated PDF combination?

Idea:

- Combine PDF determinations as independent observations
- Correlated combination produces a **weighted average**

Assuming two sets of same variance, $\mathrm{Var}[f_a^{\mathrm{NNPDF}}] = \mathrm{Var}[f_a^{\mathrm{MSHT}}]$:

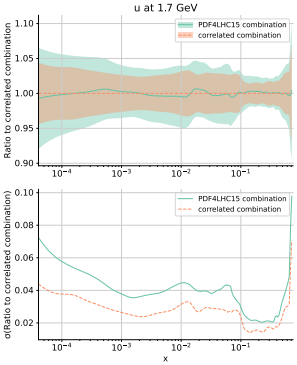
$$\mathrm{Var}[f_a^{\mathrm{comb}}] = \frac{1}{2}\left(1 + \rho[f_a^{\mathrm{MSHT}}, f_a^{\mathrm{NNPDF}}]\right)\mathrm{Var}[f_a^{\mathrm{NNPDF}}]$$

Problems:

- Does not account for difference in central values
- How to compute correlations reliably? I.e. methodological components

Underestimated data-induced correlations leads to **underestimated uncertainty of combination**

What if we combine many repeated determinations of the same PDF set?

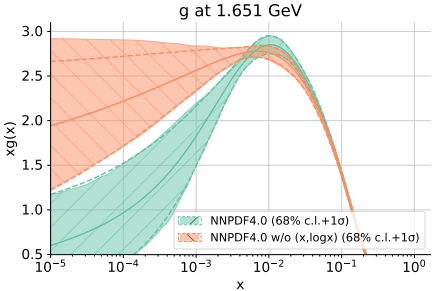


Future ML developments

2111.02954 and 2211.12961

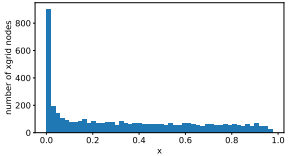
Preprocessing

- PDF model: $f_a = A_a x^{\alpha_a}(1-x)^{\beta_a}\mathrm{NN}_a(x, \log x)$
- Exponents α_a and β_a sampled **randomly** per replica
- Range determined through **self-consistent iterative procedure**
- x and $\log x$ inputs

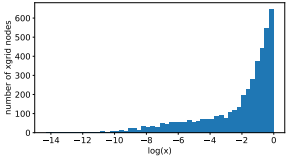


- Need to **iterate**
- Hierarchy in input scale. A source of **bias?**
- Sampled exponents add **noise** during hyperopt

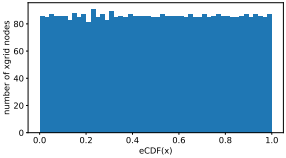
Feature scaling



x



$\log x$



eCDF(x)

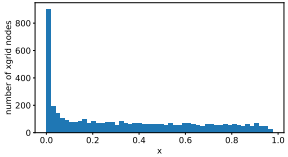
Avoid potential bias:

- generate flat distribution: effective cumulative distribution function (eCDF)
- Add interpolation
- Rerun hyperopt

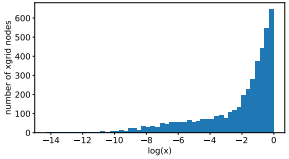
$\Rightarrow$ Only single input

$\Rightarrow$ No preprocessing needed

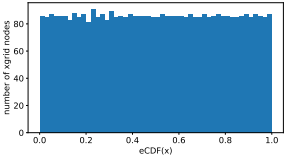
Feature scaling



x



$\log x$



eCDF(x)

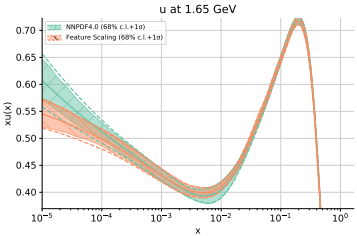
Avoid potential bias:

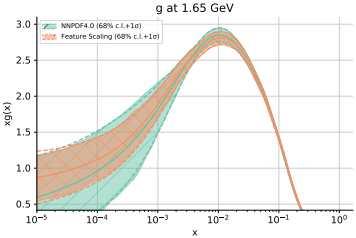
- generate flat distribution: effective cumulative distribution function (eCDF)
- Add interpolation
- Rerun hyperopt

⇒ Only single input

⇒ No preprocessing needed

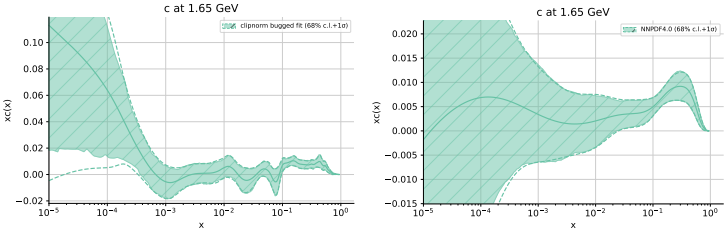
Good agreement between NNPDF4.0 and feature scaling!





Overfitting in hyperopt

- Statistical fluctuations affecting hyperopt results
- Hyperopt solutions can be overfitted or underfitted



- What does it mean for a PDF to be overfitted?
- More wiggles → overfitting?
- More wiggles → better agreement to data

overfitting metric

Validation data only used for early stopping

1. Take set of PDF replicas
2. Calculated expected validation loss
3. Compare to real validation loss

$$\mathcal{R}_O = \chi^2_{\mathrm{val,r}}\left[\mathcal{T}\left[f^{(r)}\right], \mathcal{D}^{(r)}\right] - \frac{1}{N}\sum_{r'=1}^{N} \chi^2_{\mathrm{val,r}}\left[\mathcal{T}\left[f^{(r)}\right], \mathcal{D}^{(r')}\right].$$

Negative $\mathcal{R}_O \to$ overfitting

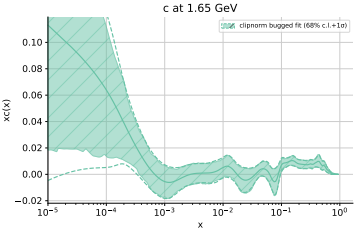
overfitting metric

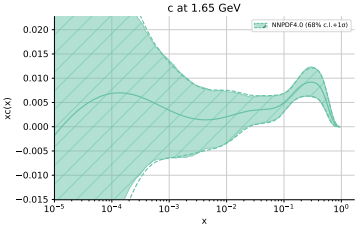
Validation data only used for early stopping

1. Take set of PDF replicas
2. Calculated expected validation loss
3. Compare to real validation loss

$$\mathcal{R}_O = \chi^2_{\mathrm{val,r}}\left[\mathcal{T}\left[f^{(r)}\right], \mathcal{D}^{(r)}\right] - \frac{1}{N}\sum_{r'=1}^{N} \chi^2_{\mathrm{val,r}}\left[\mathcal{T}\left[f^{(r)}\right], \mathcal{D}^{(r')}\right].$$

Negative $\mathcal{R}_O \to$ overfitting





old candidate: $\mathcal{R}_O = -0.024 \pm 0.012$ NNPDF4.0: $\mathcal{R}_O = -0.001 \pm 0.013$

I did not discuss other sources of uncertainty

- Theory and missing higher order uncertainties (Andrea, next days)
- The negligible impact of data inconsistencies (see 2212.07703)
- Future tests and closure tests

Possible future directions

- Optimize hyperopt folds
- Overfitting metric in hyperopt
- Parallelization on GPU
- Understanding PDF fitting in a Bayesian framework

More work still to be done!

I did not discuss other sources of uncertainty

- Theory and missing higher order uncertainties (Andrea, next days)
- The negligible impact of data inconsistencies (see 2212.07703)
- Future tests and closure tests

Possible future directions

- Optimize hyperopt folds
- Overfitting metric in hyperopt
- Parallelization on GPU
- Understanding PDF fitting in a Bayesian framework

More work still to be done!

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]	✗	✓	✗	✗	✗

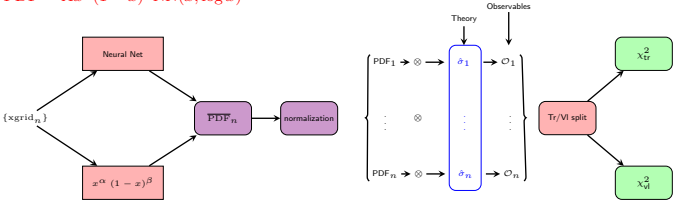
Data set	Ref.	NNPDF3.1	NNPDF4.0	ABMP16	CT18	MSHT20
CMS W asym. 7 TeV ($\mathcal{L} = 36$ pb^{-1})	[268]	✗	✗	✗	✗	✓
CMS Z 7 TeV ($\mathcal{L} = 36$ pb^{-1})	[269]	✗	✗	✗	✗	✓
CMS W electron asymmetry 7 TeV	[55]	✓	✓	✗	✓	✓
CMS W muon asymmetry 7 TeV	[56]	✓	✓	✓	✓	✗
CMS Drell-Yan 2D 7 TeV	[57]	✓	✓	✗	(✓)	✓
CMS 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

$$\mathrm{PDF} = Ax^{\alpha}(1-x)^{\beta}\mathrm{NN}(x, \log x)$$



Main changes:

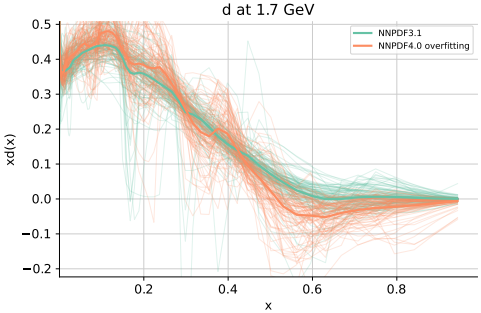
- `Python` codebase
  - Easier and faster development
- Freedom to use external libraries (default: `TensorFlow`)
- Modularity ⇒ ability to vary all aspects of the methodology

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.



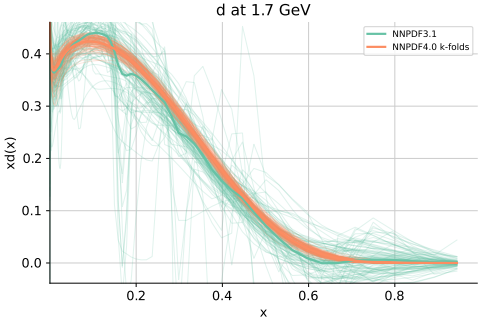
Hyperoptimization: the reward function

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

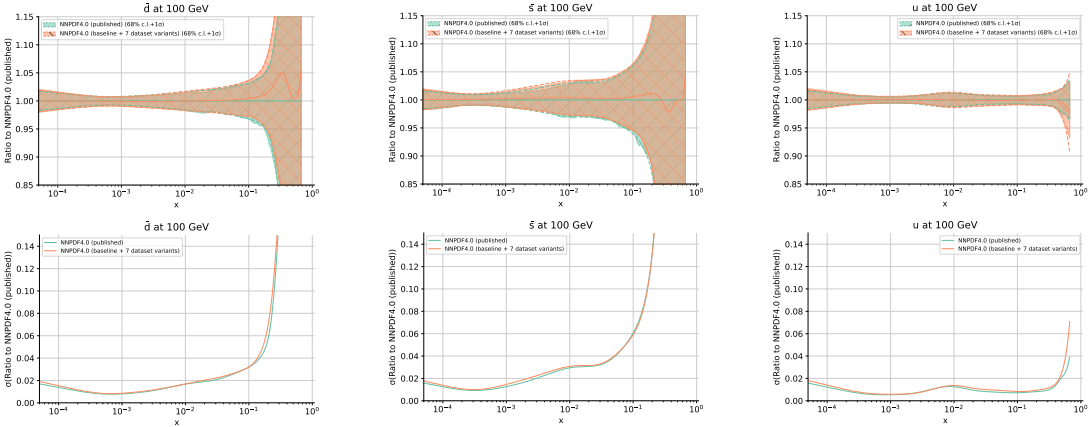
$\Rightarrow$ The hyperoptimization target is not based on data that entered the fit.



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

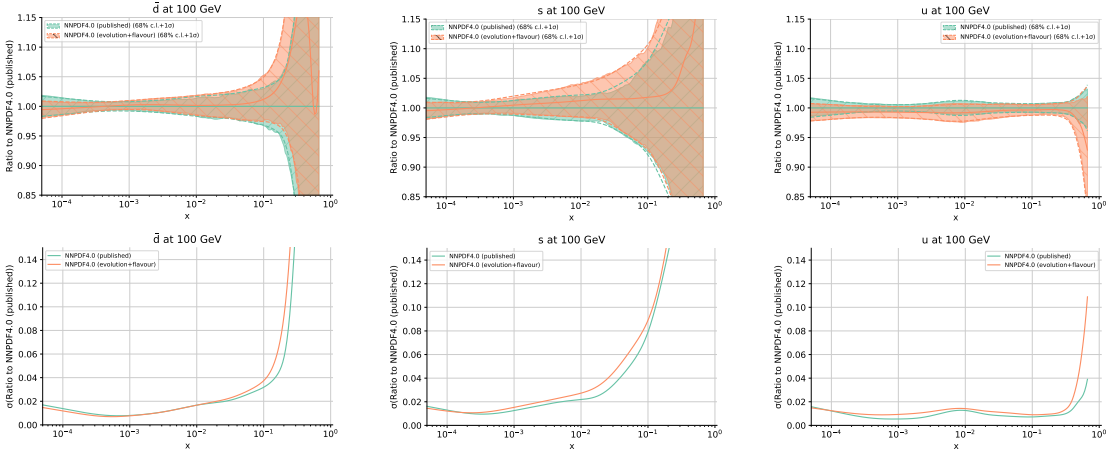
The (negligible) impact of datasets with tension

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.



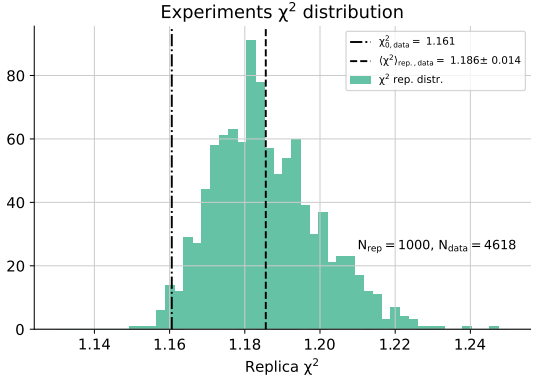
Envelope of fits with different arametrization bases

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

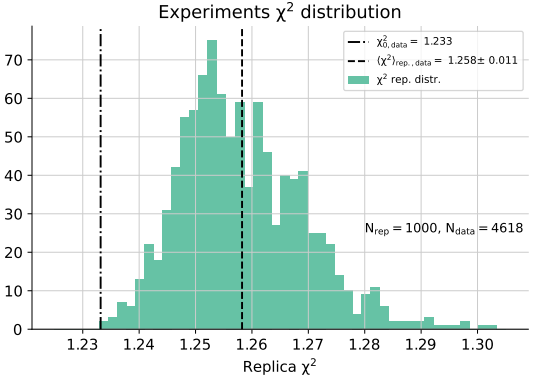


Understanding the χ^2 distribution

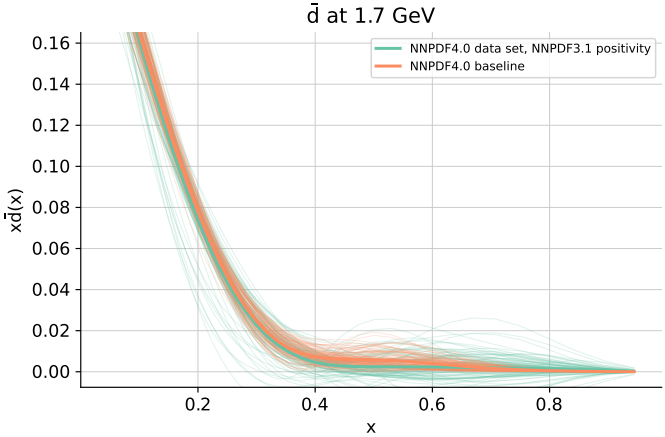
Experimental χ^2



t_0 χ^2



Impact of positivity on the PDFs

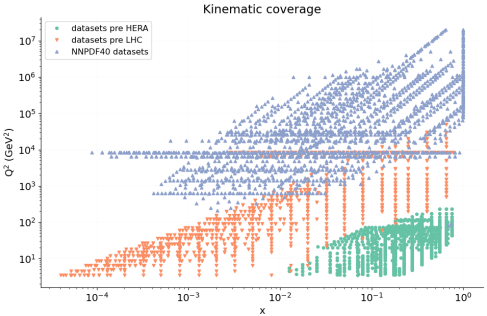


Trusting uncertainties outside the data region

- The improved methodology and extended dataset result in a reduction of the PDF uncertainties
- 'Closure test' to validate uncertainty in the data region: arxiv:1410.8849
- Can we trust the uncertainties in the extrapolation region?

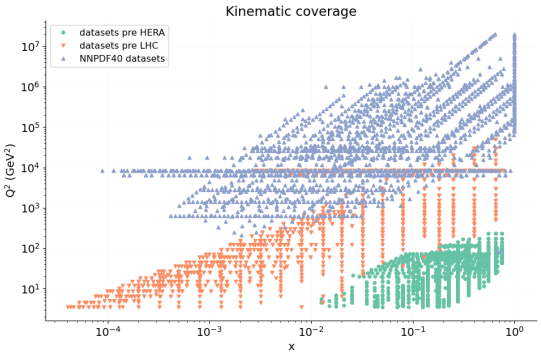
Idea:

1. Take a historic dataset
   e.g. pre-HERA or pre-LHC
2. Perform fit
3. Compare predictions to "future" data



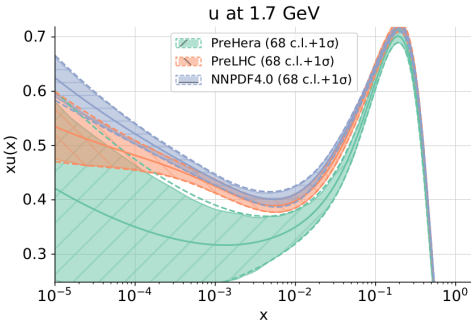
Future tests

For more information see arxiv:2103.08606



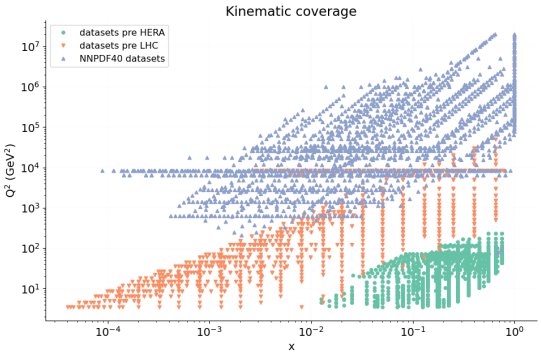
χ^2/N (only exp. covmat)

(dataset)	NNPDF4.0	pre-LHC	pre-Hera
pre-HERA	1.09	1.01	0.90
pre-LHC	1.21	1.20	**23.1**
NNPDF4.0	1.29	**3.30**	**23.1**



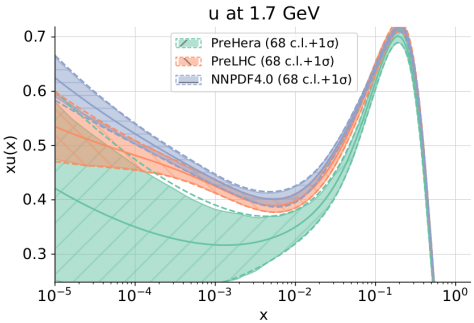
Future tests

For more information see arxiv:2103.08606



χ^2/N (exp. and PDF covmat)

(dataset)	NNPDF4.0	pre-LHC	pre-Hera
pre-HERA			0.86
pre-LHC		1.17	**1.22**
NNPDF4.0	1.12	**1.30**	**1.38**



The total uncertainty increases, and accommodates for difference between predictions and new data.

Closure test

See Eur.Phys.J.C 82 (2022); arxiv:2111.05787

Closure test of a known input assumption

1. Assume a "true" underlying PDF (e.g. a single PDF replica)
2. Produce data distributed according to the experimental covariance matrices
3. Perform a fit to this data

Examples of statistical estimators:

- **Bias**: squared difference between central value and true observable
  Variance: variance of the model predictions
  Faithful uncertainties require $E[\text{bias}] = \text{variance}$
- Is truth within one sigma in 68% of cases?

$\sqrt{\text{bias}/\text{variance}}$	$\xi_{1\sigma}^{(\text{data})}$
1.03 ± 0.05	0.68 ± 0.02