The inclusion of theory errors in NNPDF fits.

The NNPDF4.0MHOU and NNPDF4.0N3LO PDFs sets

A. Barontini

University of Milan and INFN Milan

IRN Terascale, Laboratori Nazionali di Frascati

16/04/2024

Based on arXiv:1906.10698, arXiv:2401.10319 and arXiv:2402.18635

Outline.

PHYSICS

- How can we evaluate PDFs?
- What are theory errors?
- Why is it relevant to include them in a PDF fit?

METHODOLOGY AND VALIDATION

- How can we include MHOU in a NNPDF fit?
- Can we validate our estimation?
- What about N3LO?

RESULTS

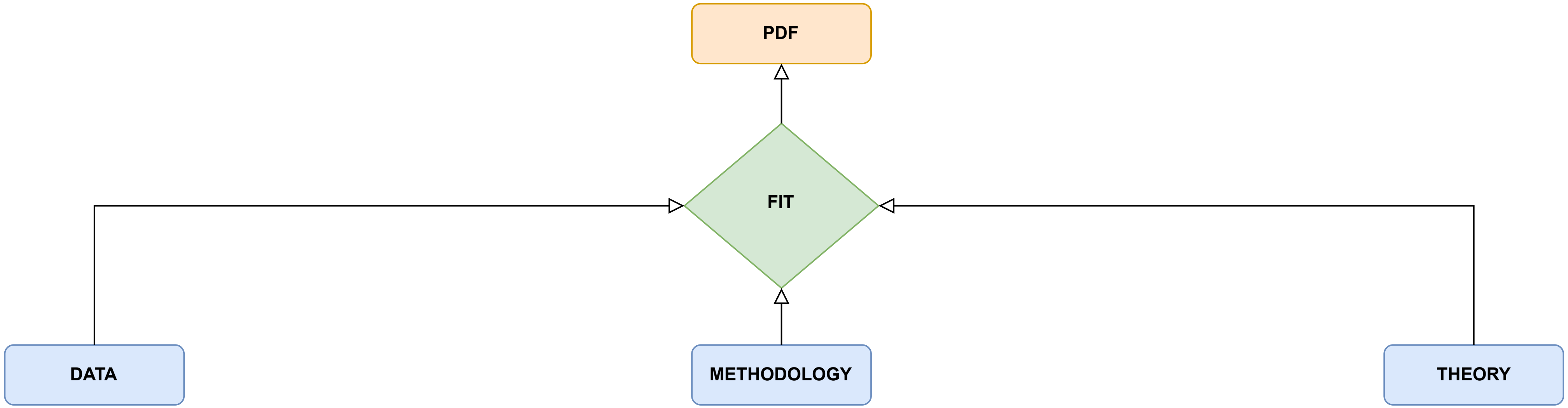
- Does the fit quality improve upon inclusion of theory errors?
- What is the impact on the PDFs?
- What about phenomenology?

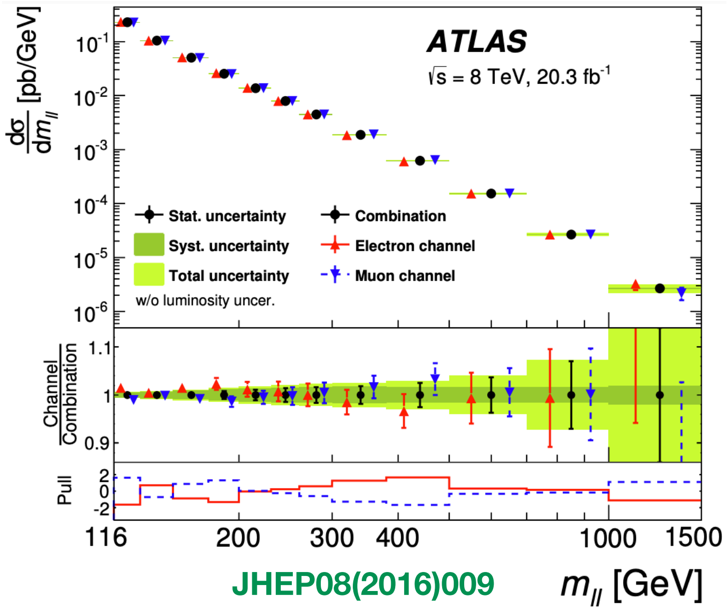
PHYSICS

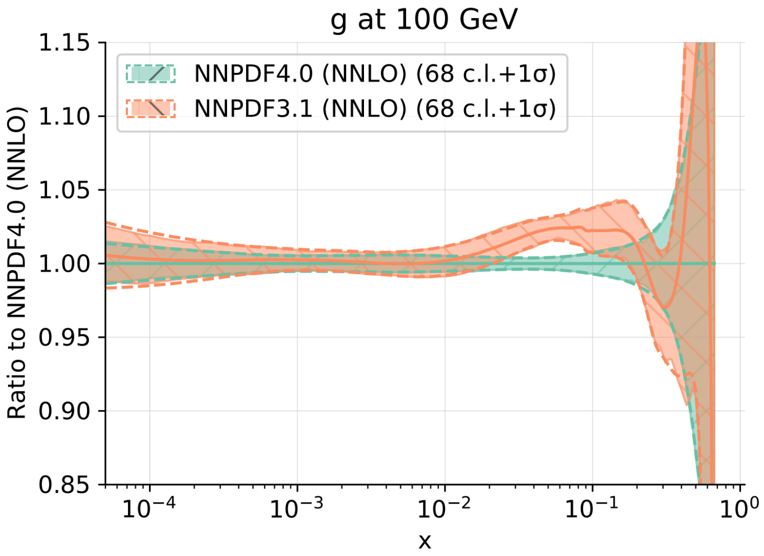
- How can we evaluate PDFs?
- What are theory errors?
- Why is it relevant to include them in a PDF fit?

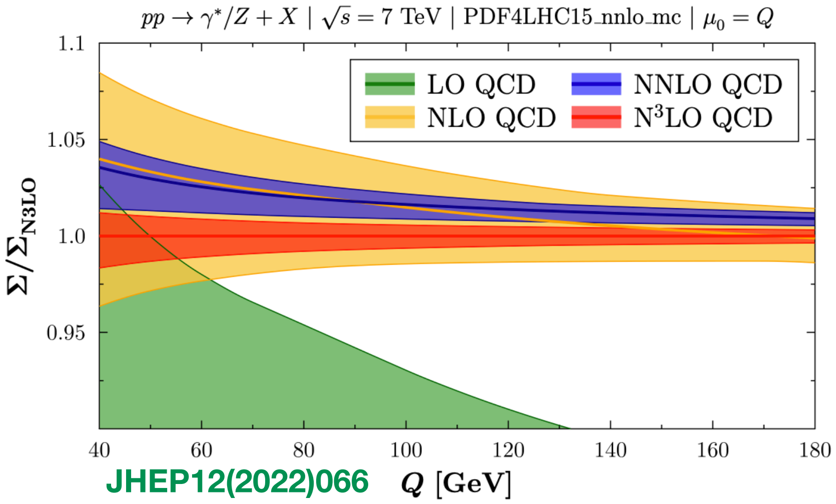
“We are not strangers, only the introduction is missing”
(Jesus Apolinaris)

Motivation.


PDF
FIT
DATA
METHODOLOGY
THEORY


ATLAS
√s = 8 TeV, 20.3 fb-1
dσ/dm_ll [pb/GeV]
Stat. uncertainty
Syst. uncertainty
Total uncertainty
w/o luminosity uncer.
Combination
Electron channel
Muon channel
Channel/Combination
Pull
116
200
300
400
1000
1500
JHEP08(2016)009
m_ll [GeV]


g at 100 GeV
NNPDF4.0 (NNLO) (68 c.l.+1σ)
NNPDF3.1 (NNLO) (68 c.l.+1σ)
Ratio to NNPDF4.0 (NNLO)
x


pp → γ*/Z + X | √s = 7 TeV | PDF4LHC15_nnlo_mc | μ0 = Q
LO QCD
NLO QCD
NNLO QCD
N3LO QCD
Σ/Σ_N3LO
JHEP12(2022)066
Q [GeV]

PDF extraction.

Factorization theorem

unknown

$$\sigma(x,Q^2) = \hat{\sigma}_{ij} \otimes f_i \otimes f_j = \int dz_1 dz_2 \hat{\sigma}(z_1, z_2, Q^2) f_i\left(\frac{x}{z_1}, Q^2\right) f_j\left(\frac{x}{z_2}, Q^2\right)$$

→ **Inverse problem**

Measured in experiments

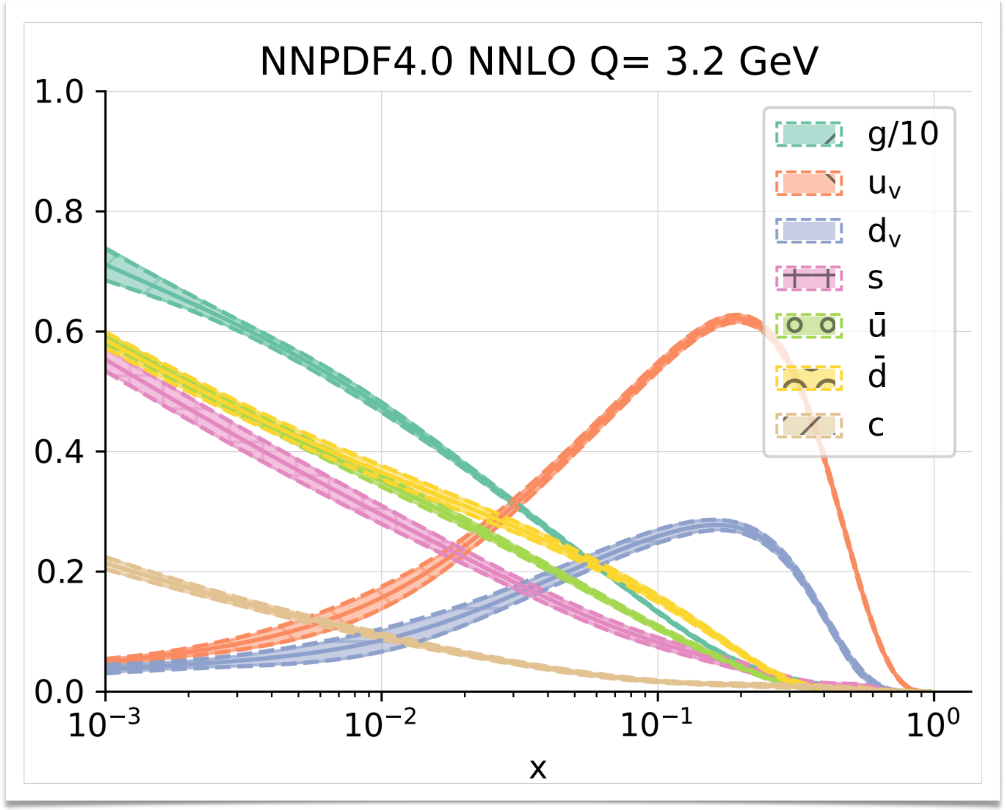
computed in perturbation theory

Also, **DGLAP equations** allow us to compute the PDFs at all scale Q^2, once known at a certain scale Q_0^2

$$f_i(Q^2) = E_{ij}(Q^2 \leftarrow Q_0^2) f_j(Q_0^2)$$

PDFs are then just a set of **unknown functions**

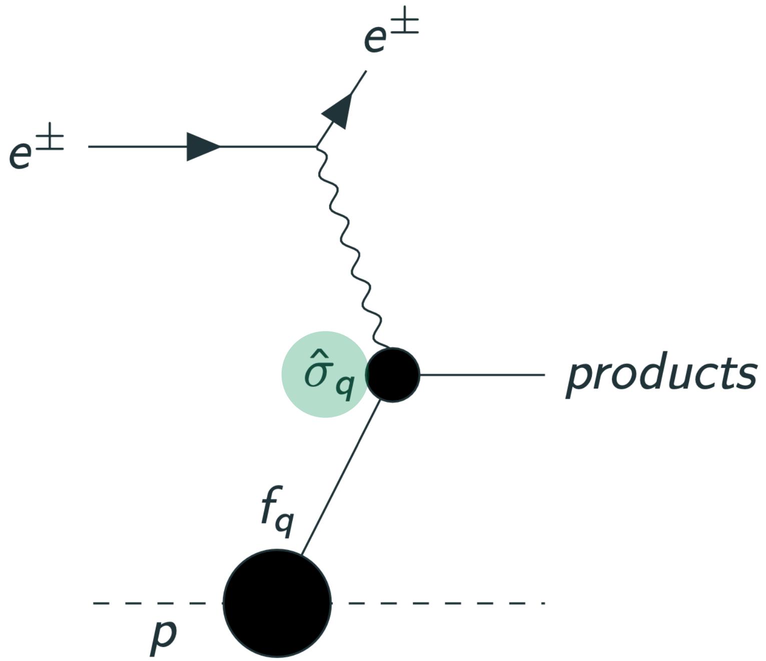
$$f_i : [0,1] -> \mathbb{R}$$



Theory errors.

$$F(Q) = \hat{\sigma}(Q^2) \otimes E_{ij}(Q^2 \leftarrow Q_0^2) \otimes f_j(Q_0^2)$$

Deep Inelastic Scattering (DIS)



→ ***Partonic cross sections*** are computed in perturbation theory

→ ***Anomalous dimensions*** inside DGLAP operator are computed in perturbation theory

$$\hat{\sigma}^{NLO} = \hat{\sigma}^{(0)} + \alpha_s \hat{\sigma}^{(1)} + \mathcal{O}(\alpha_s^2)$$
$$\gamma^{NLO} = \alpha_s \gamma^{(0)} + \alpha_s^2 \gamma^{(1)} + \mathcal{O}(\alpha_s^3)$$

MHOU

(Missing Higher Order Uncertainties)

How can we estimate them?

Theory errors: estimation.

Scale Variations

$$\overline{F}^{NLO}(\mu_f = \kappa_f Q, \mu_r = \kappa_r Q) - F^{NLO}(\mu_f = Q, \mu_r = Q) = \mathcal{O}(NNLO)$$

Factorization scale

Estimates MHOU of anomalous dimensions

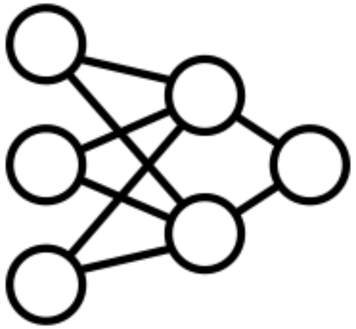
$$E^{NLO}(Q \leftarrow Q_0) \rightarrow \overline{E}^{NLO}(Q \leftarrow Q_0, \kappa_f)$$

Renormalization scale

Estimates MHOU of partonic cross sections

$$\hat{\sigma}^{NLO}(Q) \rightarrow \overline{\hat{\sigma}}(Q, \kappa_r)$$

$\kappa_f, \kappa_r \in (0.5, 2.0)$ is the most common choice



METHODOLOGY AND VALIDATION

- How can we include MHOU in a NNPDF fit?
- Can we validate our estimation?
- What about N3LO?

“Truth has nothing to do with the conclusion, and everything to do with the methodology”
(Stefan Molyneux)

MHOU in a PDF fit: the *theory covmat*.

How to use it

→ Experimental and theoretical uncertainties enter in a symmetric way in the figure of merit used for PDF determination.

→ The theory covariance matrix S describes theoretical uncertainties and correlations.

→ Include it both in figure of merit and in pseudodata generation.

FIT WITHOUT THEORY ERRORS		FIT WITH THEORY ERRORS
$\chi^2 \propto (D_i - T_i) C_{ij}^{-1} (D_j - T_j)$	→	$\chi^2 \propto (D_i - T_i)(C+S)_{ij}^{-1}(D_j - T_j)$
Pseudodata replica $\propto C$	→	*Pseudodata replica* $\propto C + S$

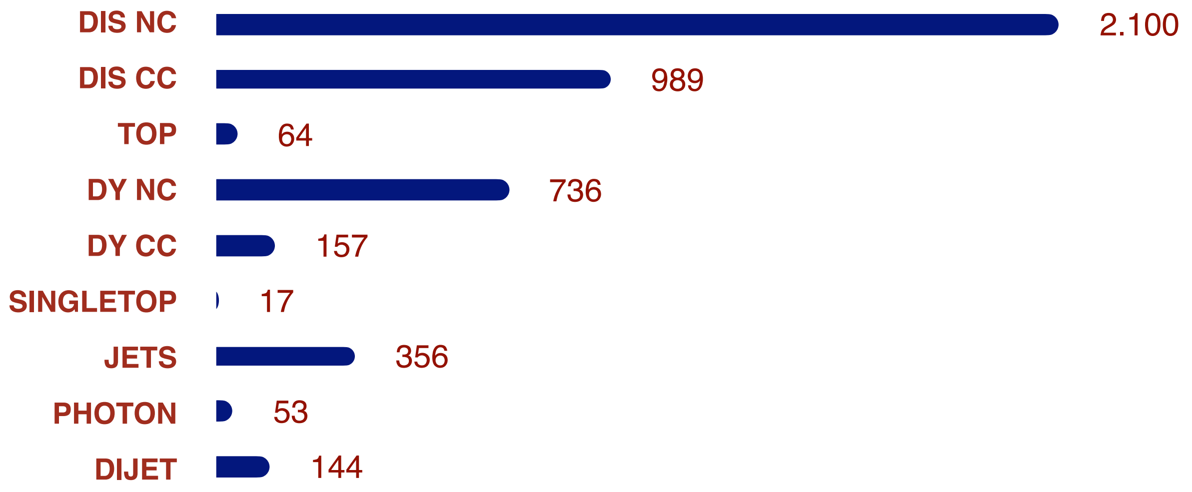
MHOU in a PDF fit: the *theory covmat*.

How to construct it

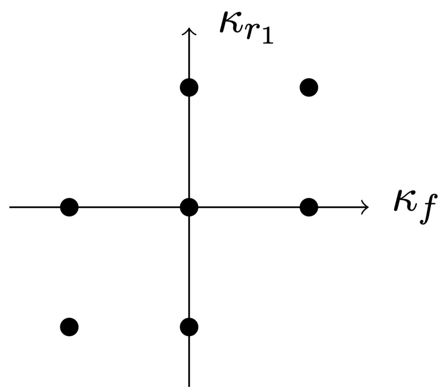
$$S_{ij} = n_m \sum_{V_m} \left(\overline{F}(\kappa_f, \kappa_{r_a}) - F \right)_{i_a} \left(\overline{F}(\kappa_f, \kappa_{r_b}) - F \right)_{j_b}$$

→ *Factorization scale* **correlates** all the points

→ *Renormalization scale* **correlates** points belonging to the same process



7 points

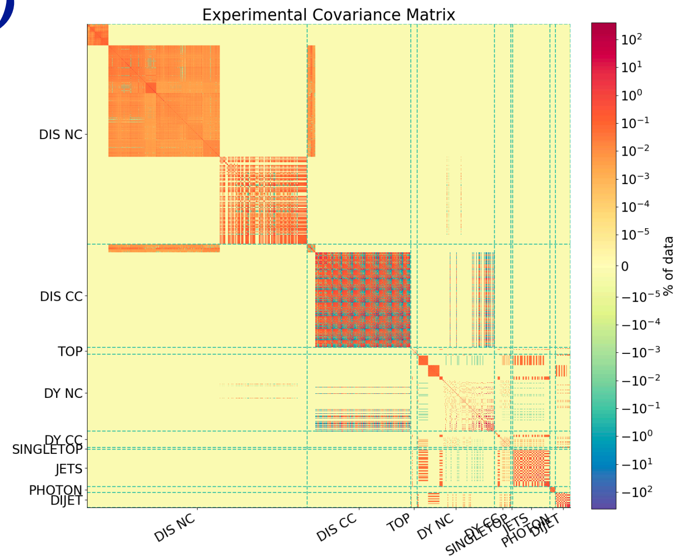


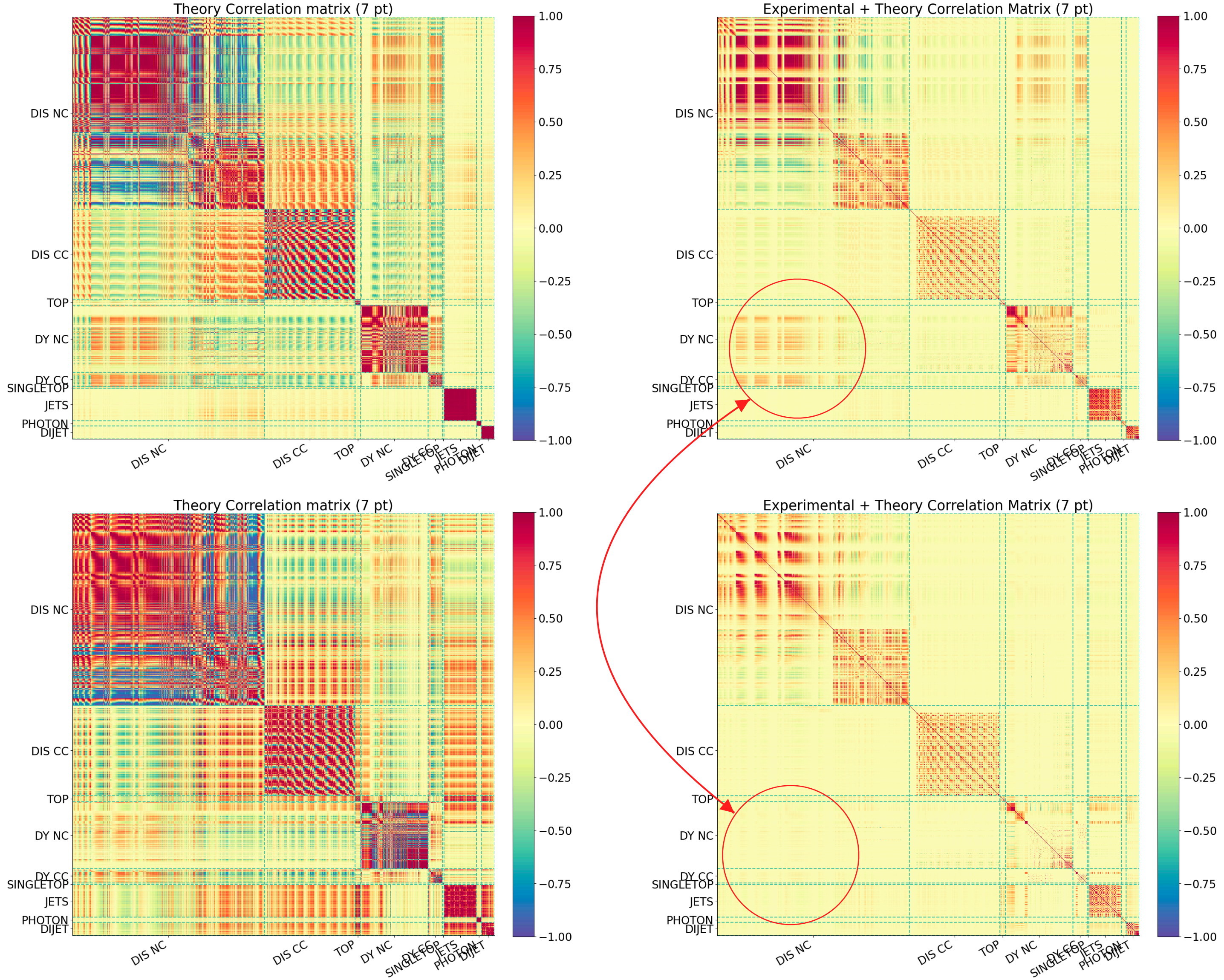
How do they look like?

C S $C + S$

→ NLO

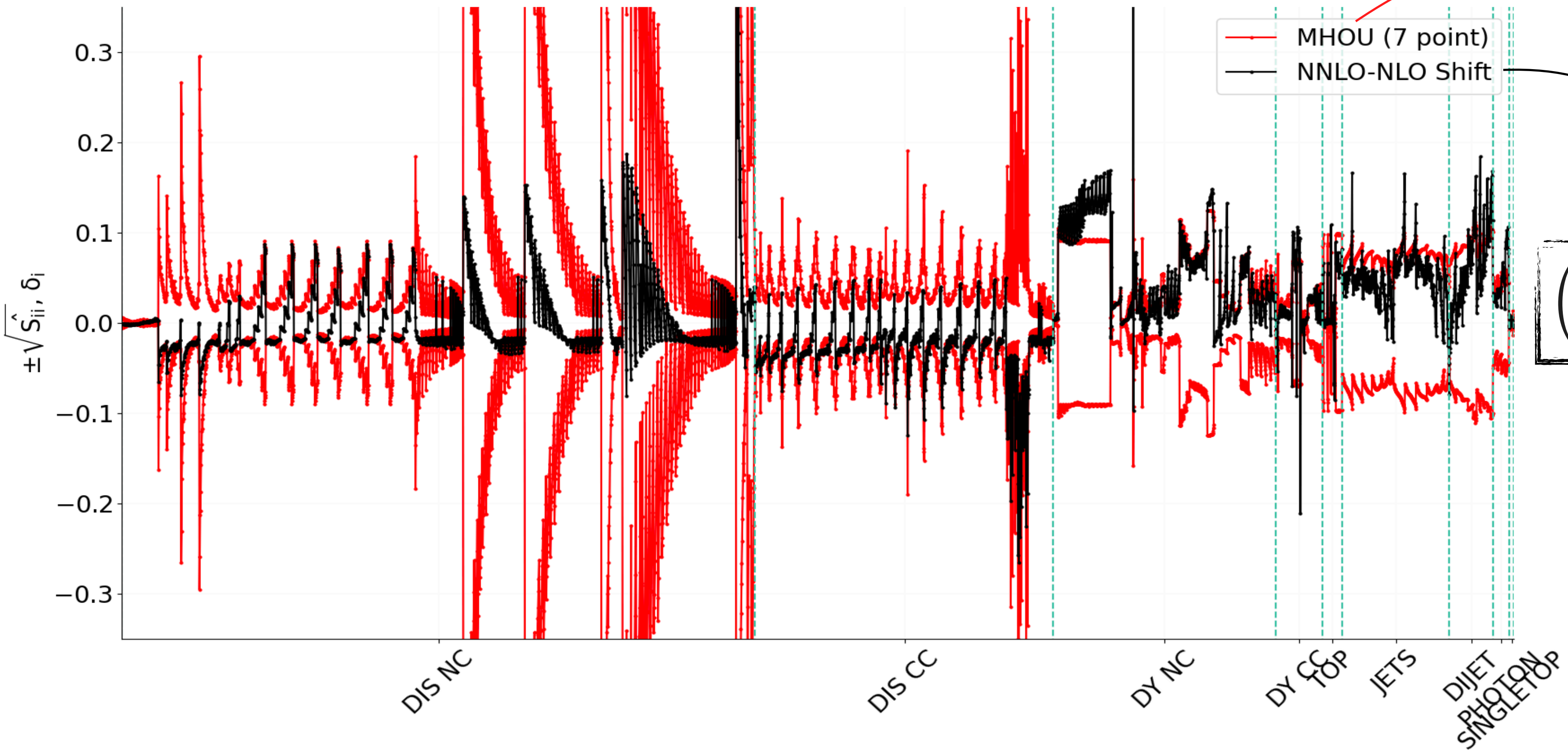
→ $NNLO$





Validation: is it reproducing the known NNLO?

Most of the predictions are currently known up to $\mathcal{O}(NNLO)$:
we can test the NLO MHOU !



$$\left(\frac{\sqrt{S_{ii}^{NLO}}}{F_i^{NLO}}\right) \times 100$$

$$\left(\frac{F_i^{NNLO} - F_i^{NLO}}{F_i^{NLO}}\right) \times 100$$

N3LO.

Thanks to G.Magni for the slides

➔ Not all the ingredients (*coefficient functions, anomalous dimensions,…*) are available at N3LO yet

↓ Incomplete Higher Order Uncertainties **(IHOU)**

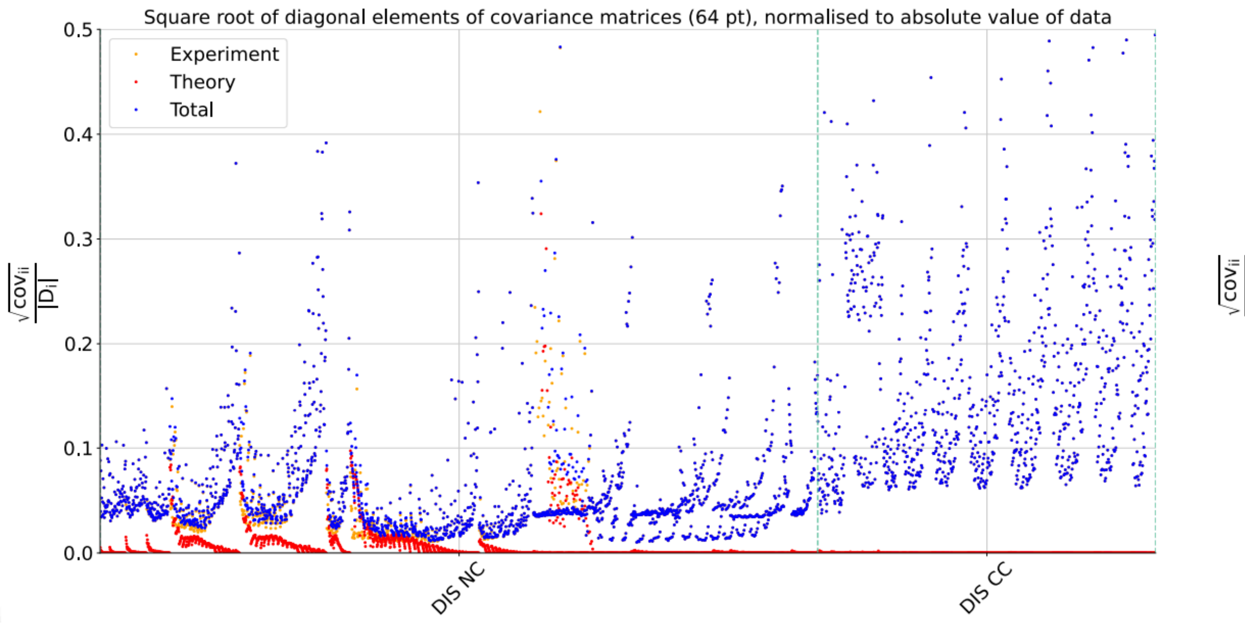
↓

- *Splitting functions @ N3LO*
- *Massive DIS coeff. functions @ N3LO*

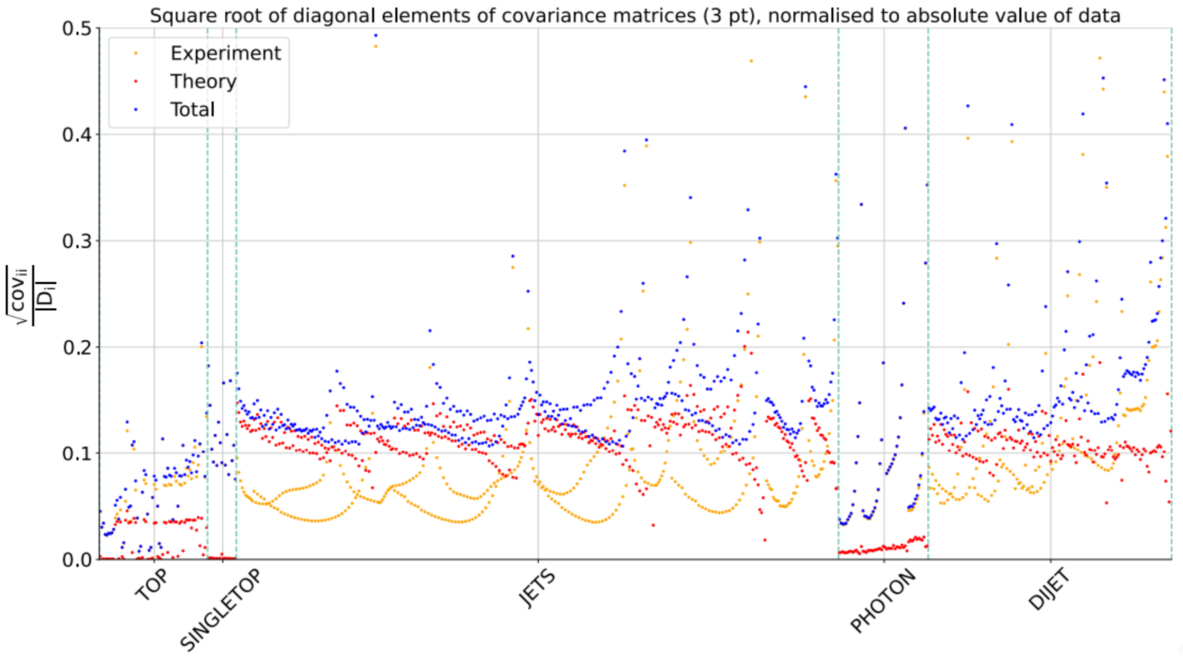
↓ Missing Higher Order Uncertainties **(MHOU)**

↓

- *NNLO MHOU for all data but DIS*



IHOU have a larger effect on the **small-x**, **low-Q** DIS data



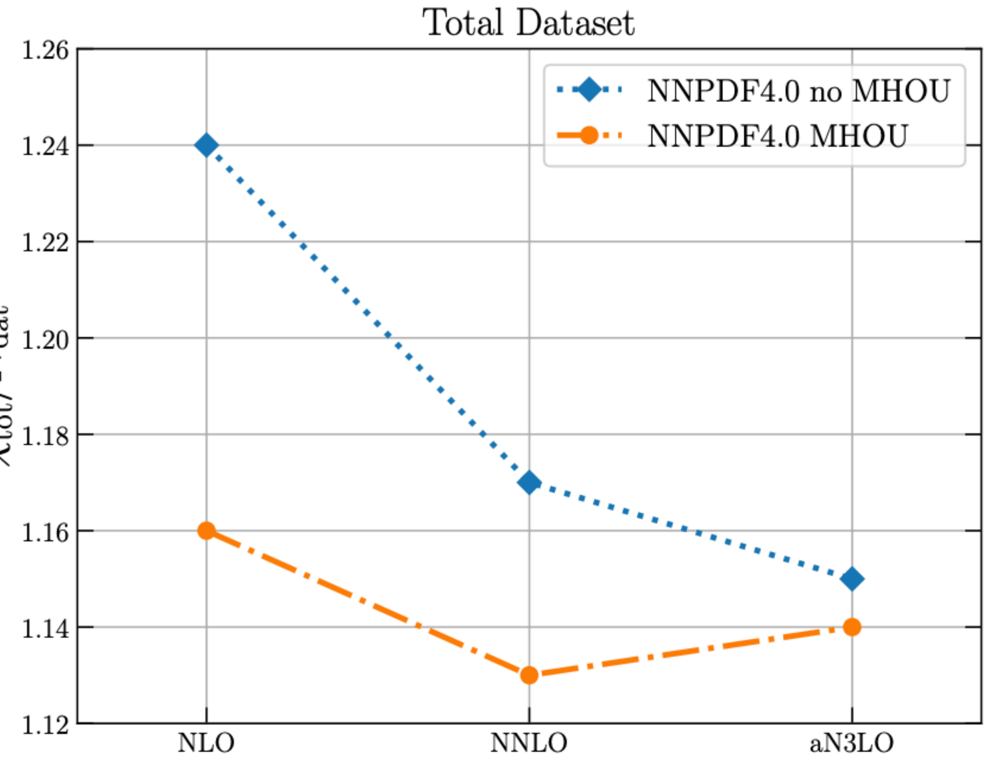
MHOU are included when ME @ N3LO are not available

RESULTS

- Does the fit quality improve upon inclusion of theory errors?
- What is the impact on the PDFs?
- What about phenomenology?

"We're always, by the way, in fundamental physics, always trying to investigate those things in which we don't understand the conclusions. After we've checked them enough, we're okay"
(Richard P. Feynman)

Fit quality.



➔ The χ^2 of the fit with MHOU is much more flat as a function of the perturbative order

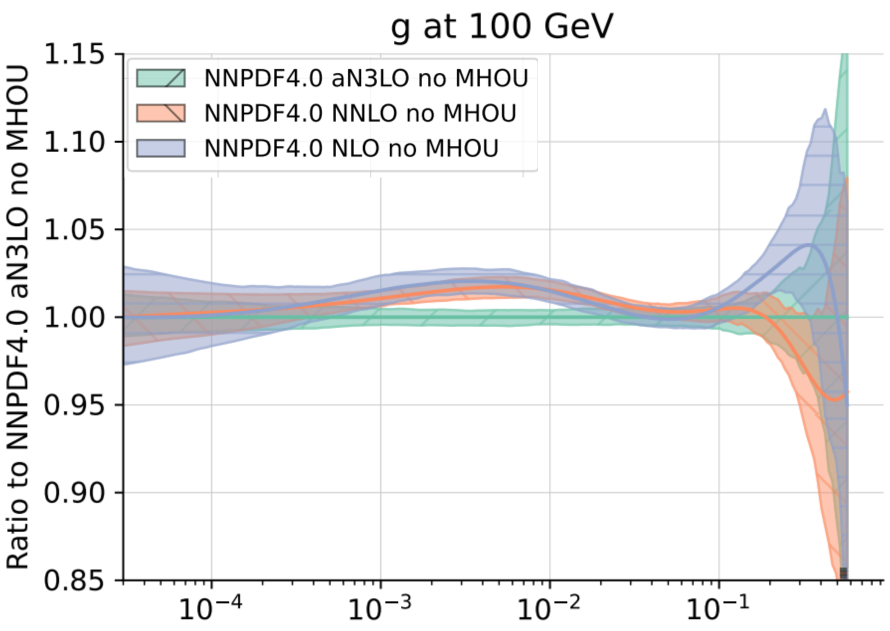
➔ The N3LO fit quality is the same irrespective of whether MHOU are included or not.

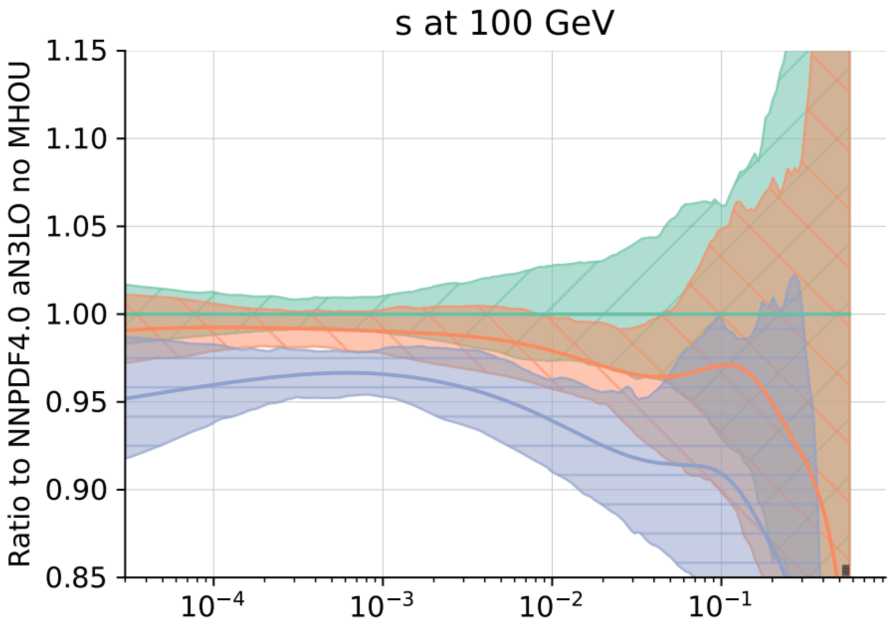
χ^2

Dataset	$N_{\rm dat}$	NLO no MHOU	NLO MHOU	$N_{\rm dat}$	NNLO no MHOU	NNLO MHOU	$N_{\rm dat}$	aN^{3}LO no MHOU	aN^{3}LO MHOU
DIS NC	1980	1.30	1.22	2100	1.22	1.20	2100	1.22	1.20
DIS CC	988	0.92	0.87	989	0.90	0.90	989	0.91	0.92
DY NC	667	1.49	1.32	736	1.20	1.15	736	1.17	1.16
DY CC	193	1.31	1.27	157	1.45	1.37	157	1.37	1.36
Top pairs	64	1.90	1.24	64	1.27	1.43	64	1.23	1.41
Single-inclusive jets	356	0.86	0.82	356	0.94	0.81	356	0.84	0.83
Dijets	144	1.55	1.81	144	2.01	1.71	144	1.78	1.67
Prompt photons	53	0.58	0.47	53	0.76	0.67	53	0.72	0.68
Single top	17	0.35	0.34	17	0.36	0.38	17	0.35	0.36
Total	4462	1.24	1.16	4616	1.17	1.13	4616	1.15	1.14

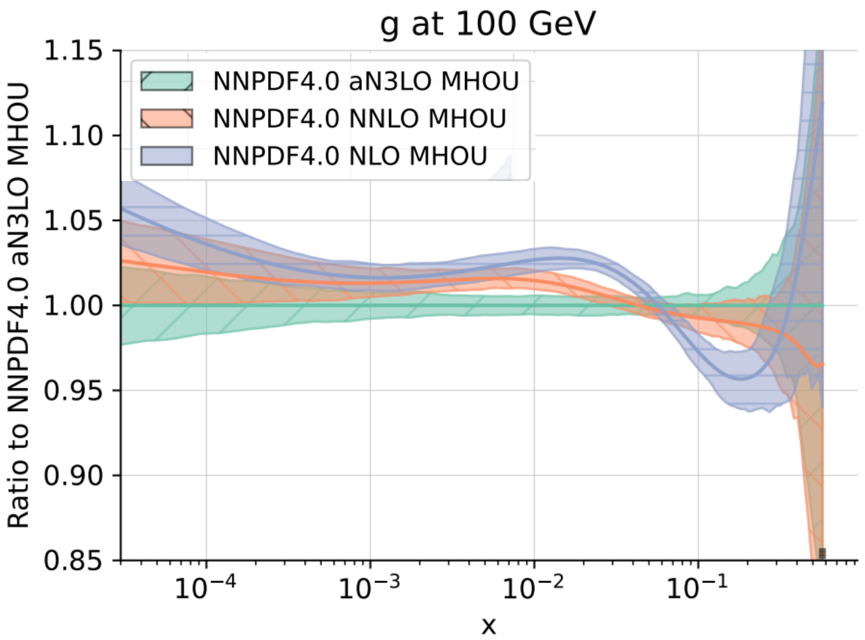
PDF comparison.

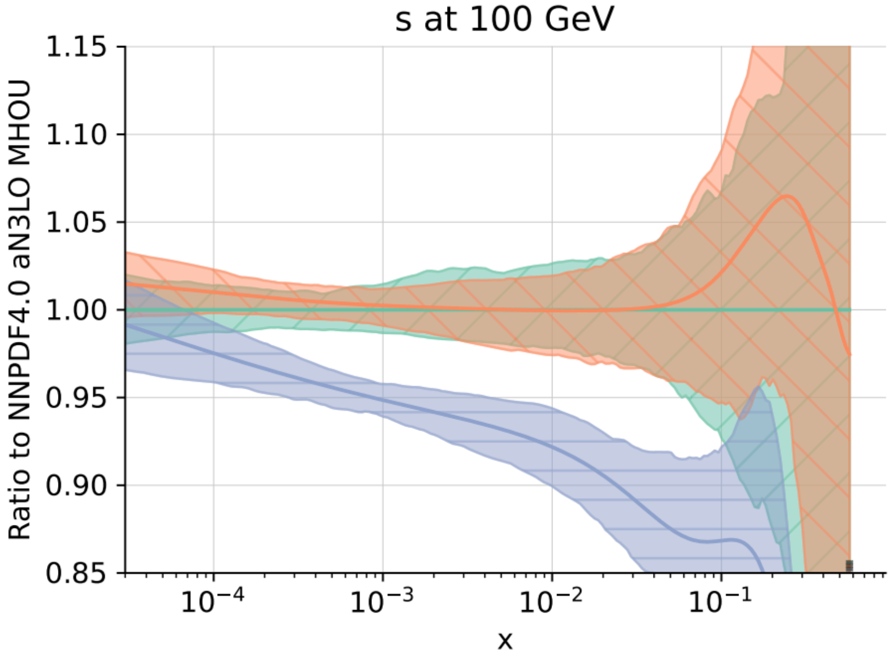
noMHOU



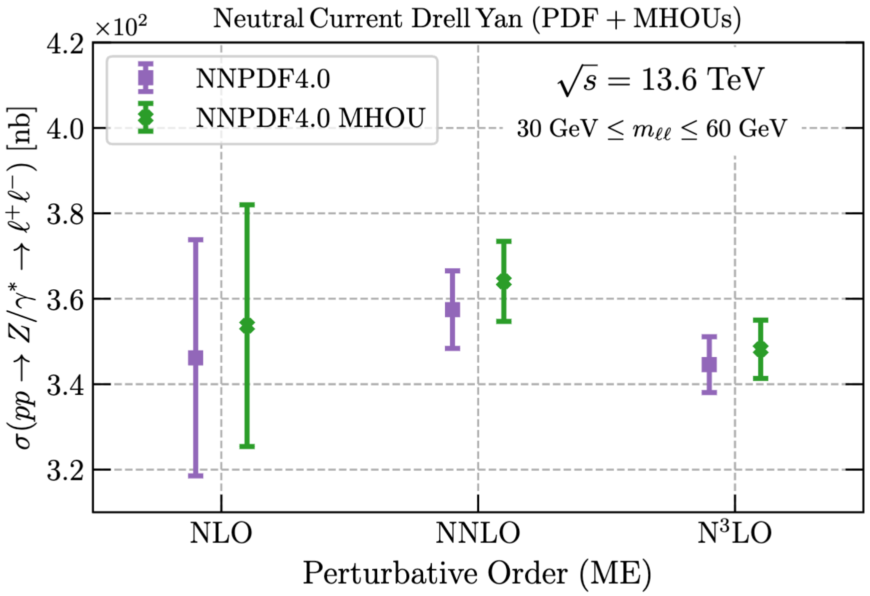


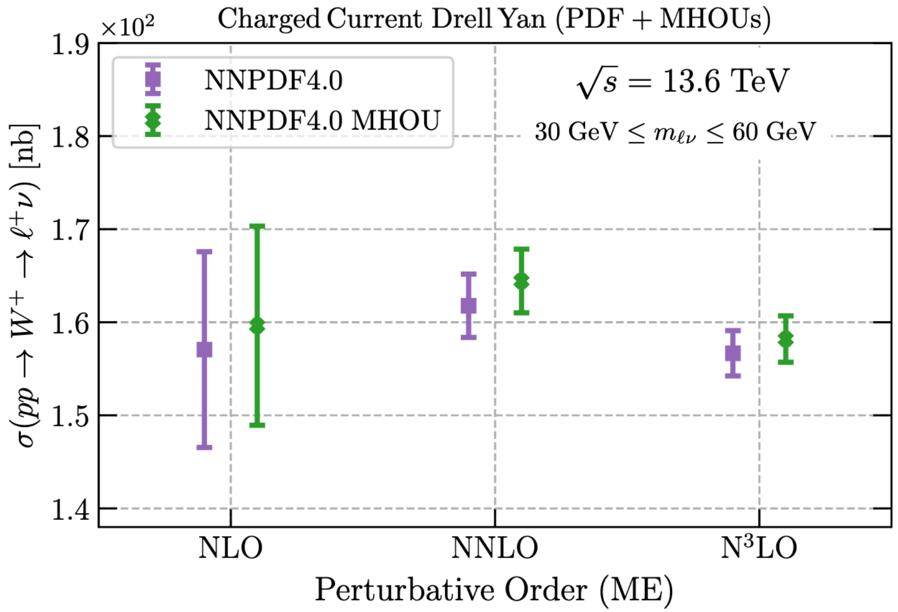
MHOU

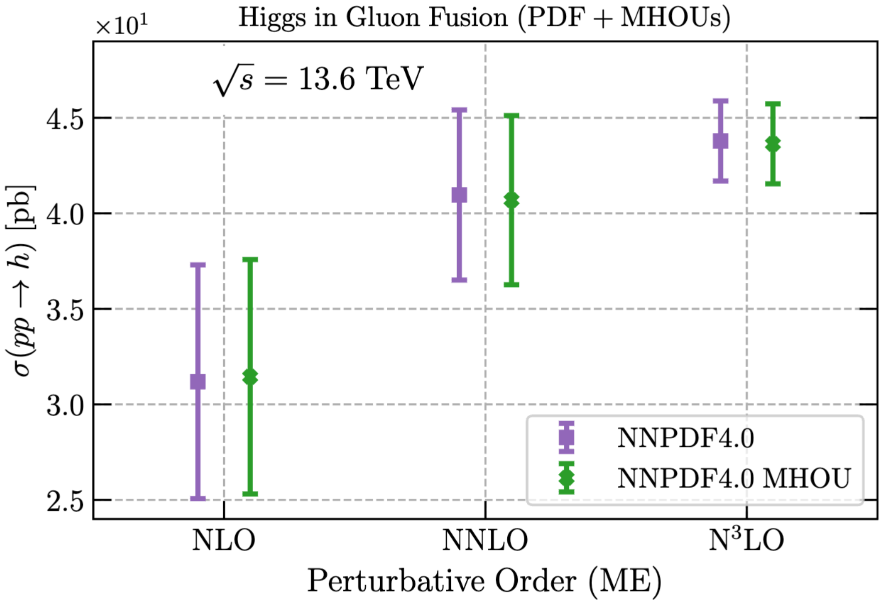


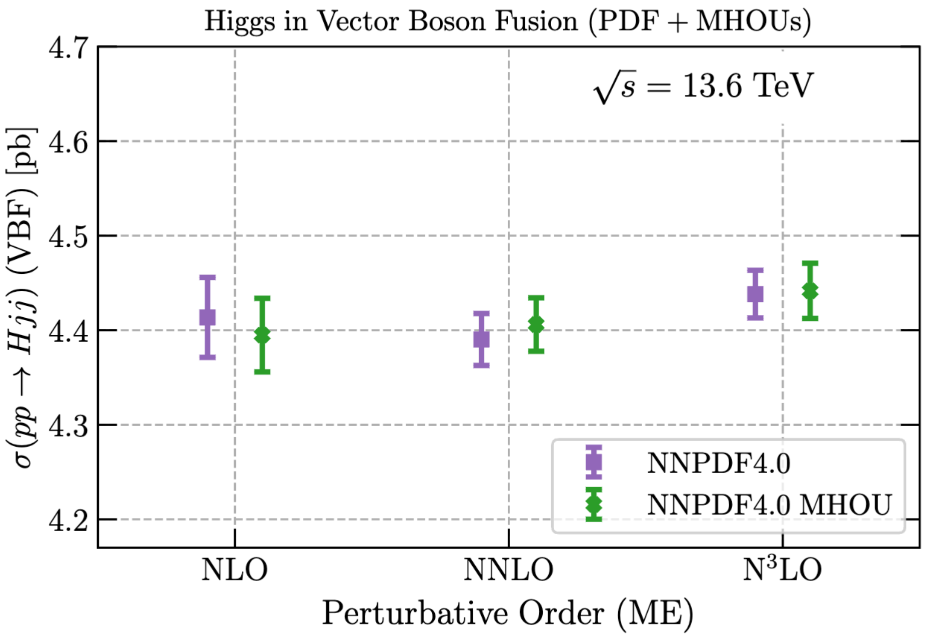


Perturbative convergence.









Conclusions and outlooks.

- Thanks to *scale variations* it is possible to estimate MHOU while, thanks to the theory covmat formalism, it is possible to include such estimation in a PDF fit.

- Including MHOU in a PDF fit is necessary to have faithful uncertainties and central values.

- The perturbative convergence from NLO to N3LO improves once theory errors are accounted for.

Thanks for your attention!

BACKUP

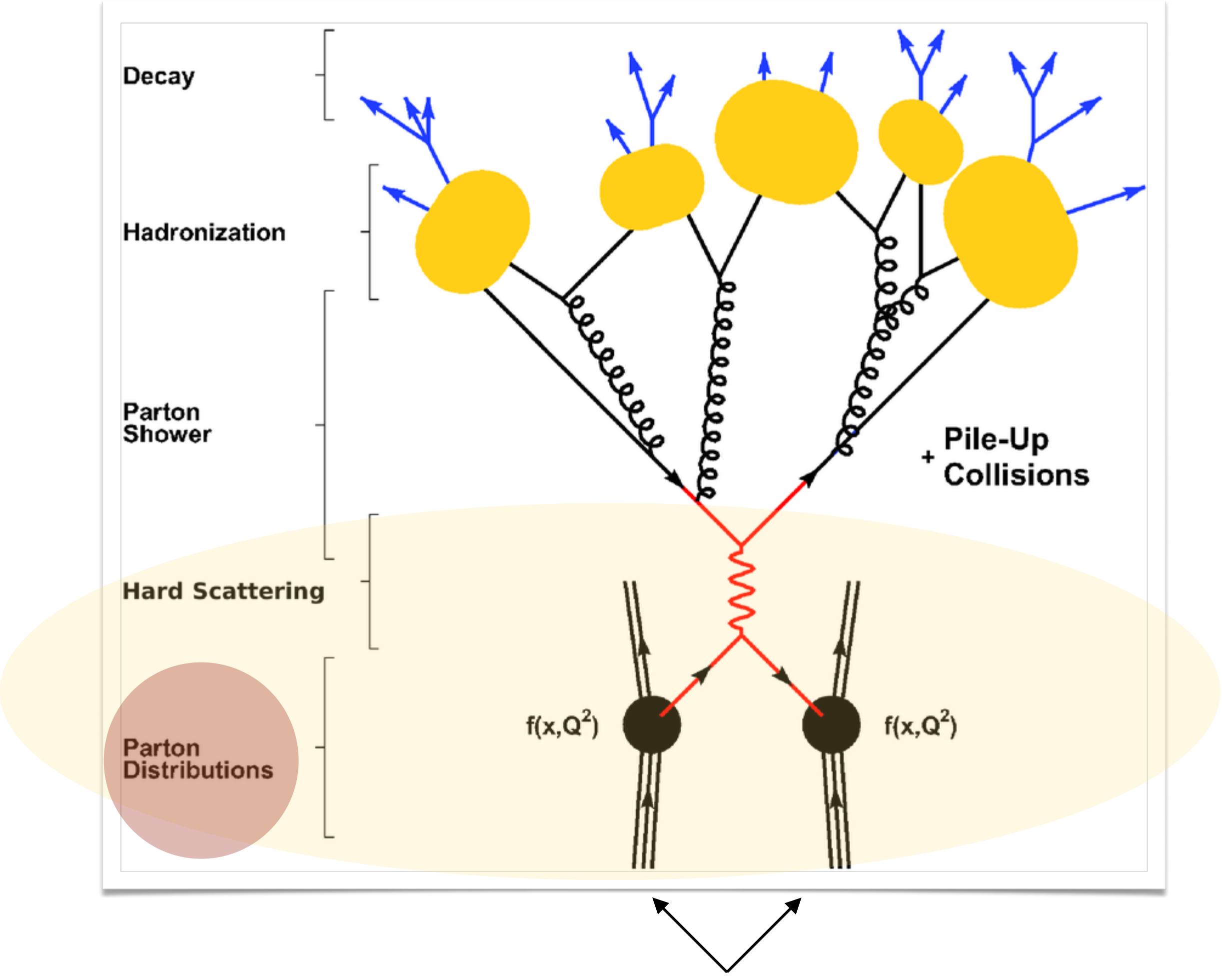
Describing a collision.

The theoretical description of a **collision** involves several **QCD** (Quantum ChromoDynamics) ingredients

We are going to focus on

And, in particular, on Parton Distribution Functions **(PDFs)**

Describe the hadronic initial state in terms of their *partonic components*



Initial state = hadrons (protons, neutrons ,...)

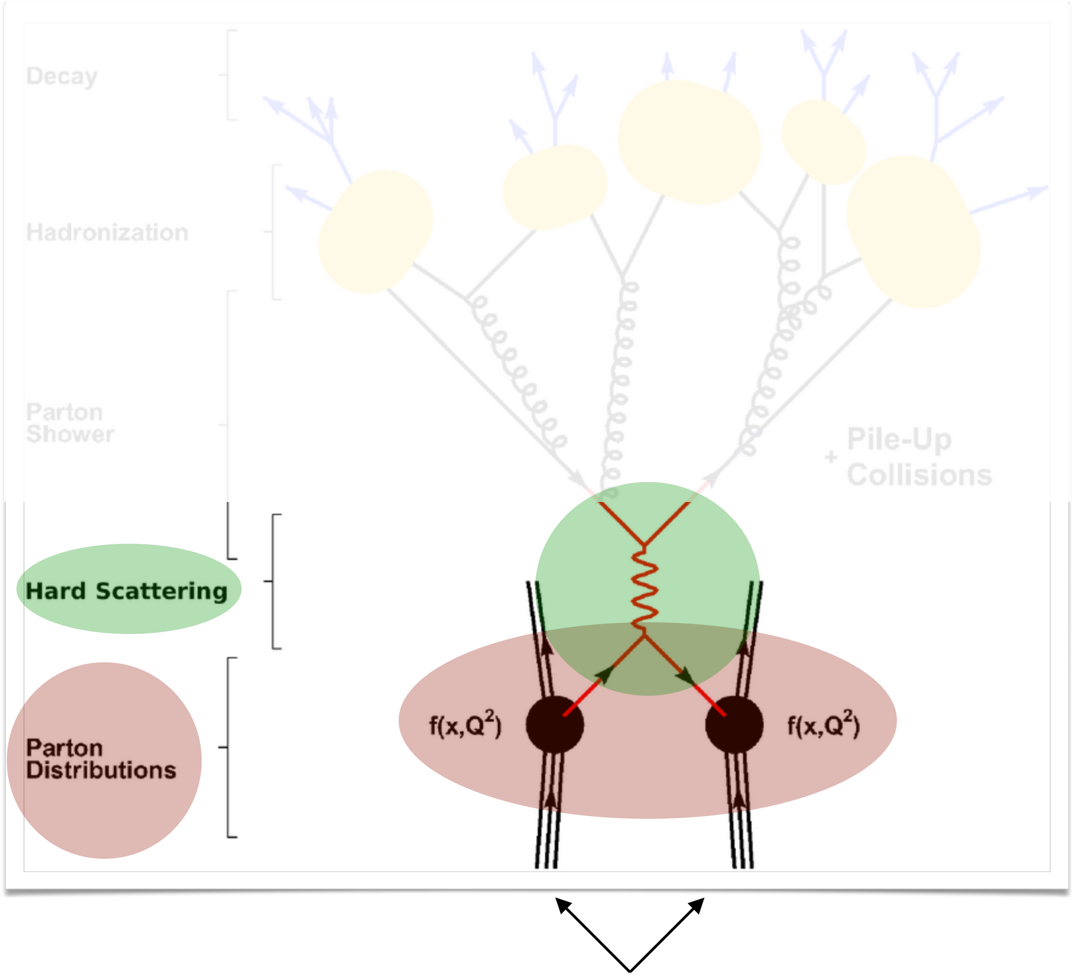
Describing a collision.

Thanks to **Factorization theorem**

PDFs

$$\sigma(x, Q^2) = \hat{\sigma}_{ij} \otimes f_i \otimes f_j = \int dz_1 dz_2 \hat{\sigma}(z_1, z_2, Q^2) f_i\left(\frac{x}{z_1}, Q^2\right) f_j\left(\frac{x}{z_2}, Q^2\right)$$

Partonic (hard) cross sections



Initial state = hadrons (protons, neutrons ,...)

- $\sigma(x, Q^2)$ is our **observable**
- Q^2 is the energy scale of the process
- $\hat{\sigma}(z_1, z_2, Q^2)$ can be computed in **perturbation theory**
  NLO, NNLO, ...
- $f_{i/j}(x, Q^2)$ **cannot** be computed in perturbation theory (and they are **universal**)

Non perturbative objects

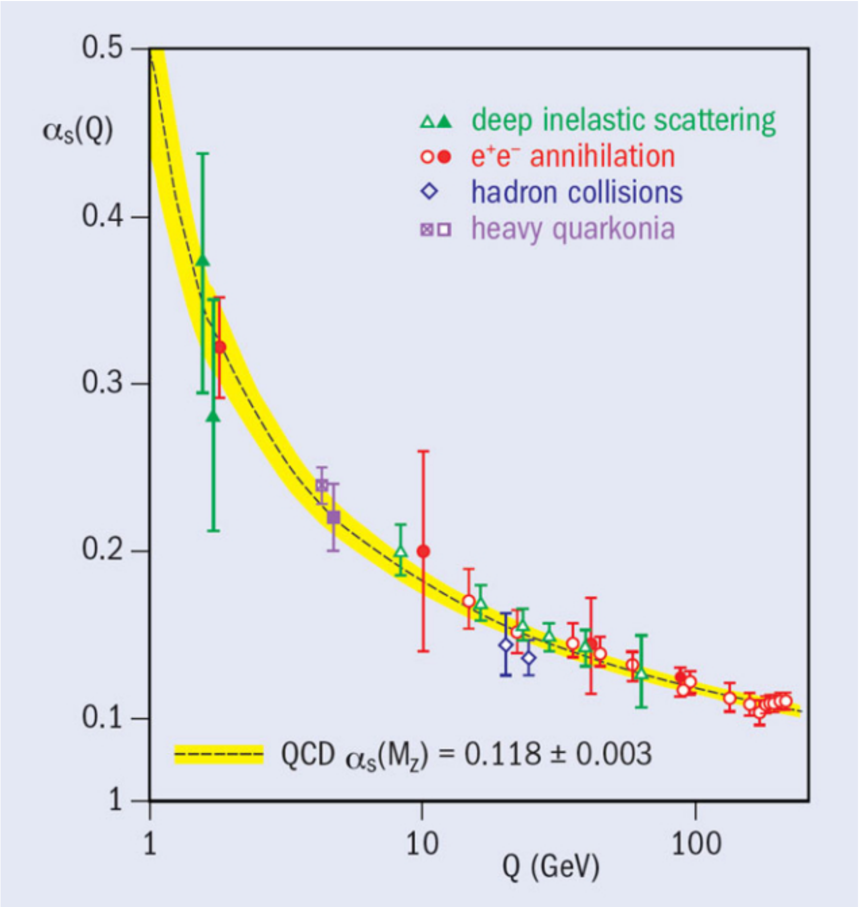
Asymptotic freedom.

In **QCD** we are usually expand quantities in terms of the **strong coupling** $\alpha_s(Q^2)$ → (Notable counterexample is lattice QCD)

$$\hat{\sigma}^{NLO}(z_1, z_2, Q^2) = \hat{\sigma}^{(0)}(z_1, z_2, Q^2) + \alpha_s(Q^2)\hat{\sigma}^{(1)}(z_1, z_2, Q^2) + \mathcal{O}(\alpha_s^2)$$

(NLO = Next-to-leading order)

But $\alpha_s(Q^2)$ is a decreasing function of the energy scale →



Physik Journal 3 12 31

perturbative QCD (pQCD) from ~ 1 GeV → **Partonic cross sections**

Non perturbative QCD below ~ 1 GeV → **PDFs** (Mass of the proton ~ 0.938 GeV)

How can we extract them?

Inverse problems.

Number of datapoints is **finite** while function space is **infinite-dimensional**

↓

Fitting PDFs is always an **under-determined** problem

↓

ASSUMPTIONS

Fixed parametrization

→ Reduce the number of parameters

→ Assumptions = choice of the parameters to be fitted

↓

→ Needs theoretical insight on PDFs shape

→ Can be biased by human prejudice

Neural Network

NNPDF choice

→ Applies a **regularization**

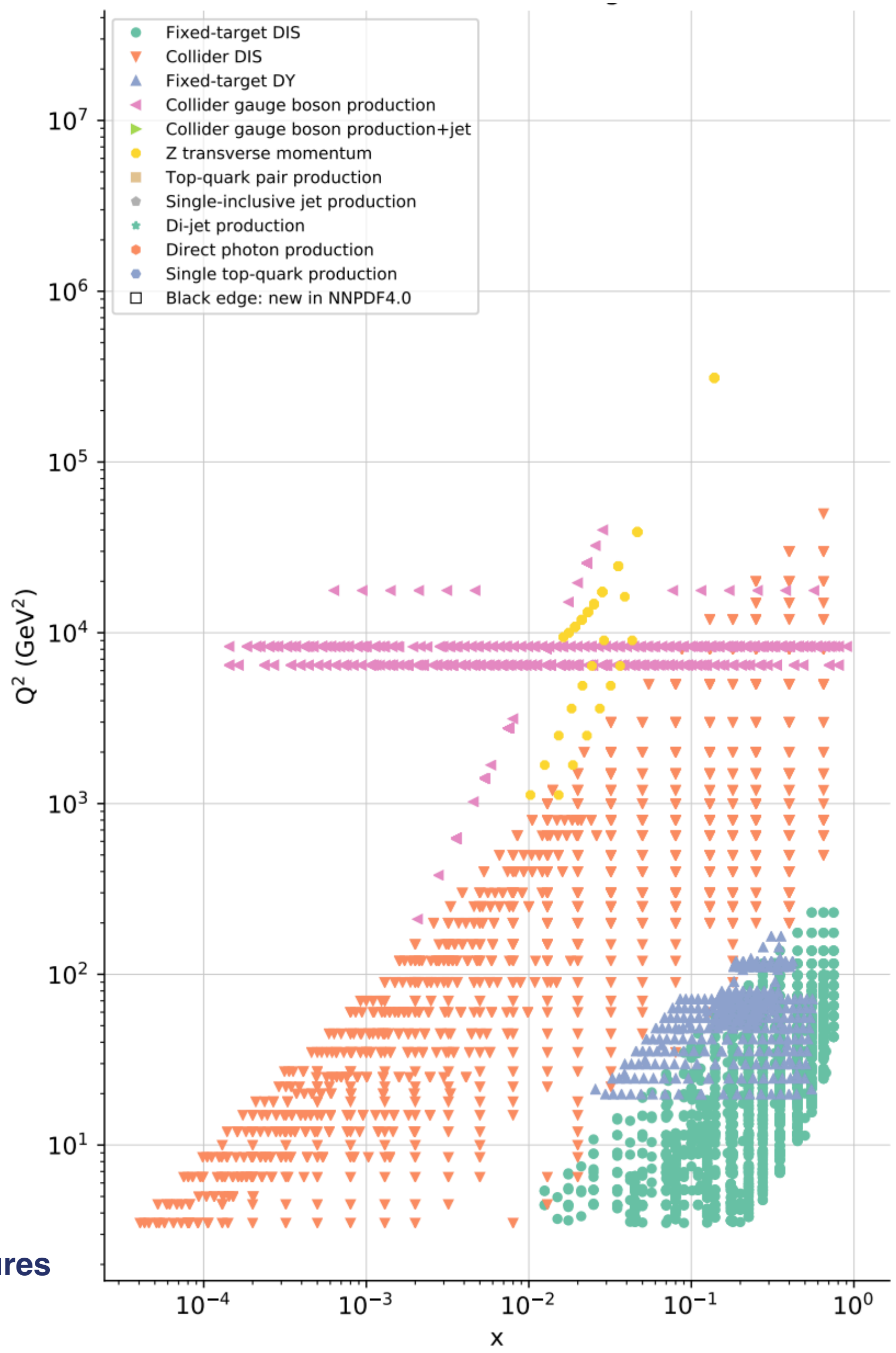
→ Assumptions = encoded in the network (and not only…)

↓

→ Needs theoretical insight on more **abstract features**

→ Human prejudice effect can be minimized

Which is better?

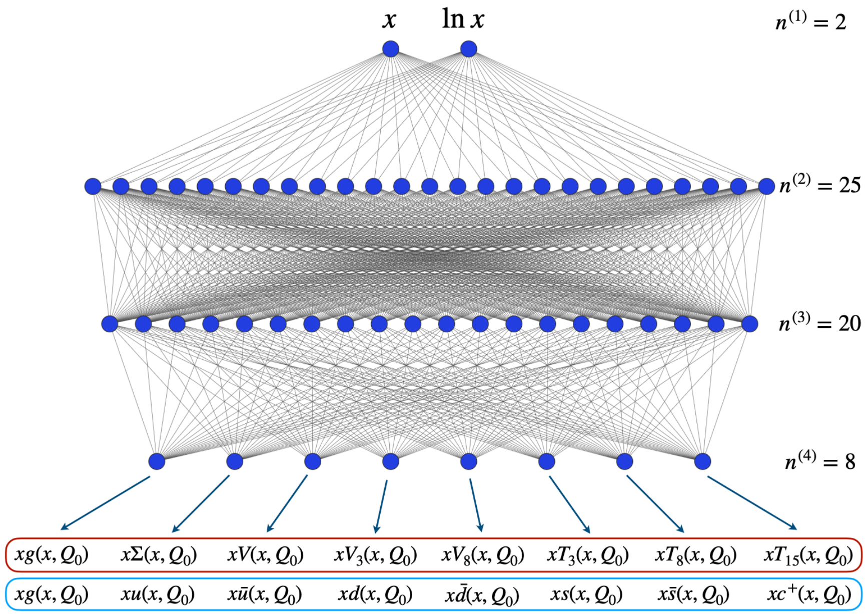


~4600 datapoints in NNPDF4.0

Parametrization: the Neural Network.

$$f(x) = A_k x^{-\alpha_k}(1-x)^{\beta_k}\mathbf{NN}(x)$$

Architecture: 2-25-20-8
Activation functions: hyperbolic; linear for the last layer



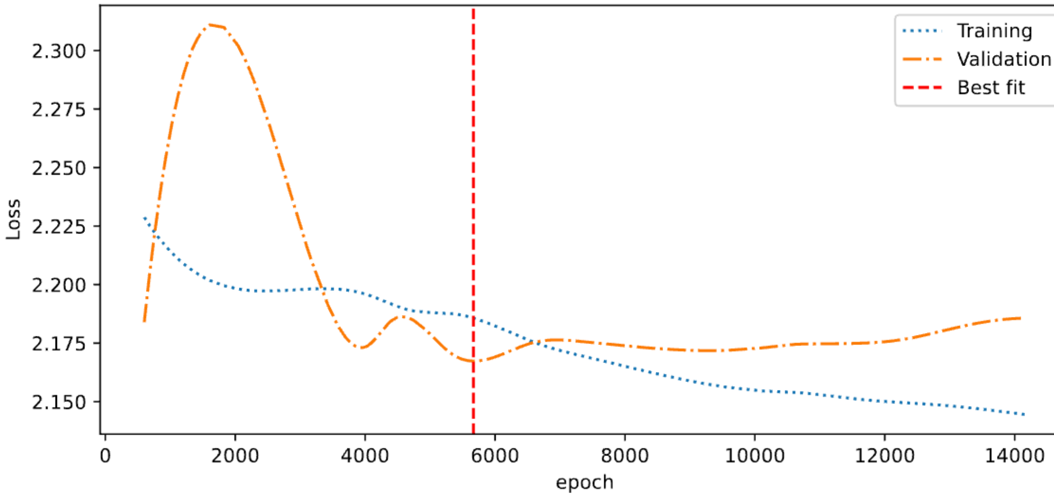
Neural Network: universal interpolator

Avoid fitting the noise (**overfitting)**

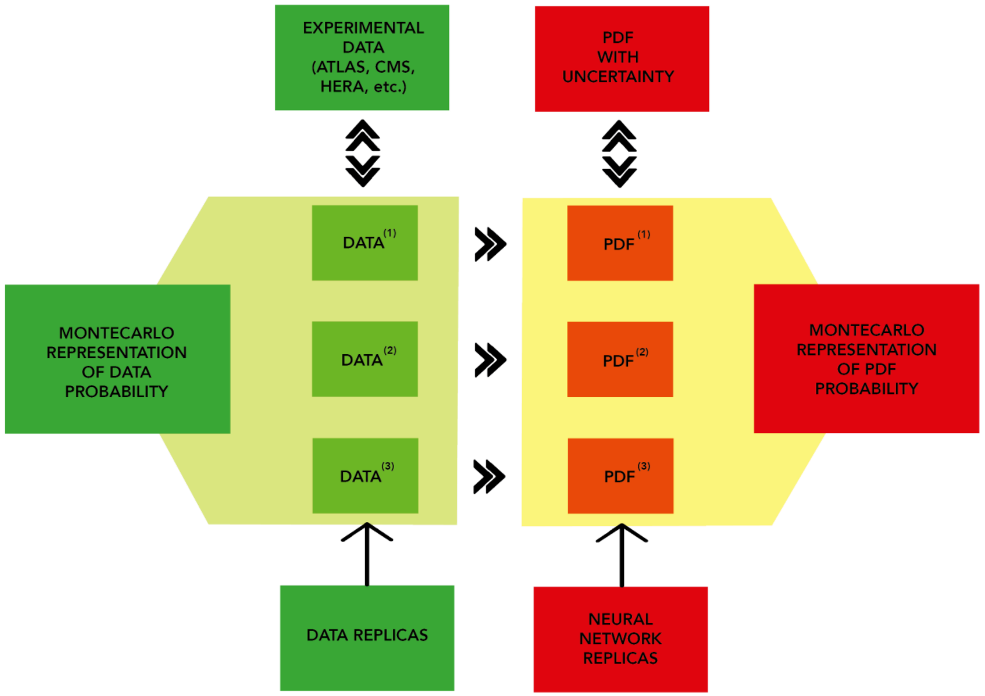
Cross-validation

Stopping

1. Divide data **D** into **training** set and **validation** set
2. Minimize training χ^2
3. Stop if validation χ^2 no longer improves
4. Take best validation χ^2



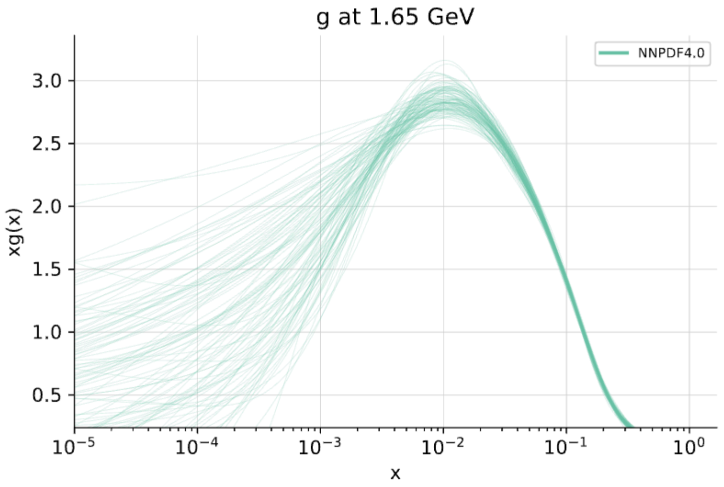
Propagating uncertainties: data to PDF.

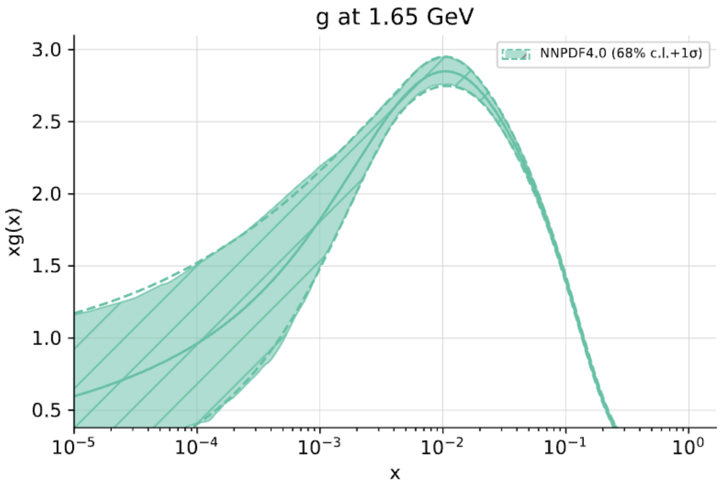


NNPDF adopts a **Monte Carlo** approach

1. Start with the original dataset **D** and its **covariance matrix C**
2. Generate N_{rep} **pseudodata** D_i according to C
3. Fit a **Neural Network** $\mathbf{NN}_i$ to each of the pseudodata replica
4. Deliver the full set of replicas

PDFs uncertainties are given by the distribution of the Monte Carlo set





NB: Another possibility is the Hessian approach. The two methods can be converted one in the other (hep-ph:1505.06736)

Automated model selection

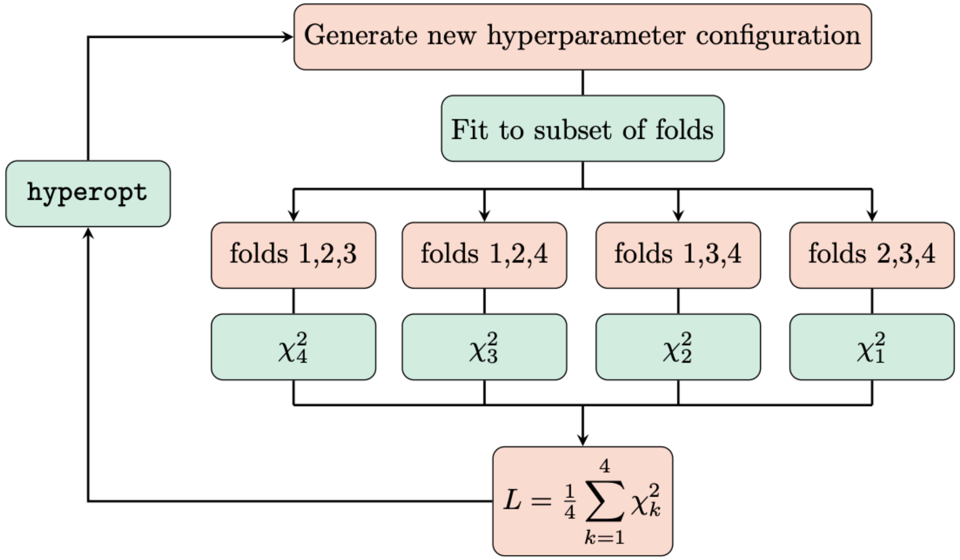
Minimize sources of **bias** in the PDFs:

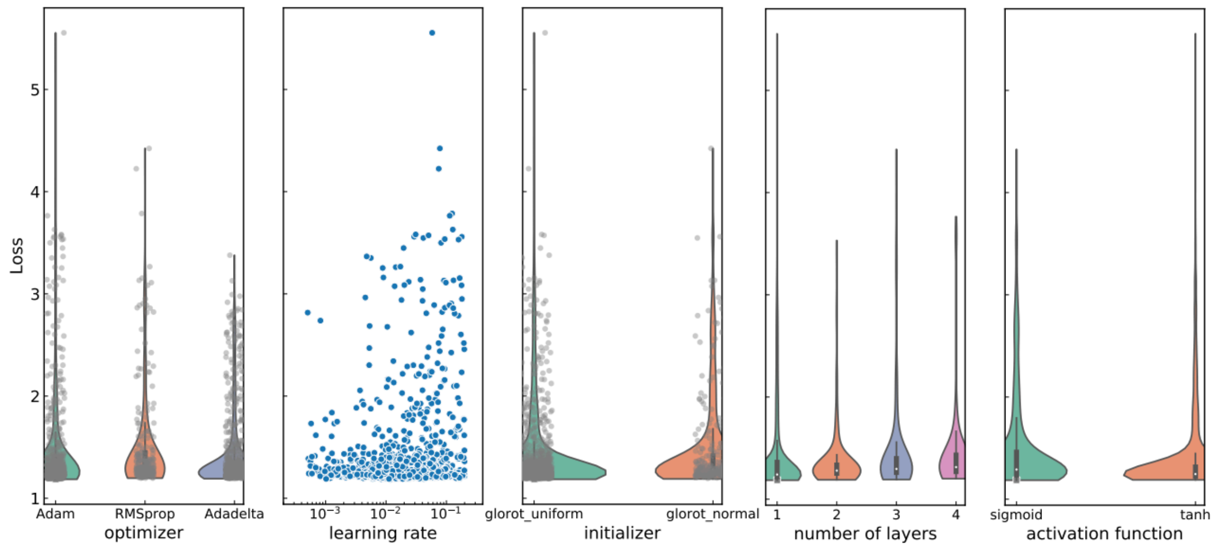
- Functional form ⟶ Neural Network
- Model parameters ⟶ **Hyperoptimization**

Idea is to scan over a large enough hyperparameter space and select the best set

Best ⟶ best χ^2 on a **test dataset** (never seen by the NN)

NB: Still requires some human input (more on this later)





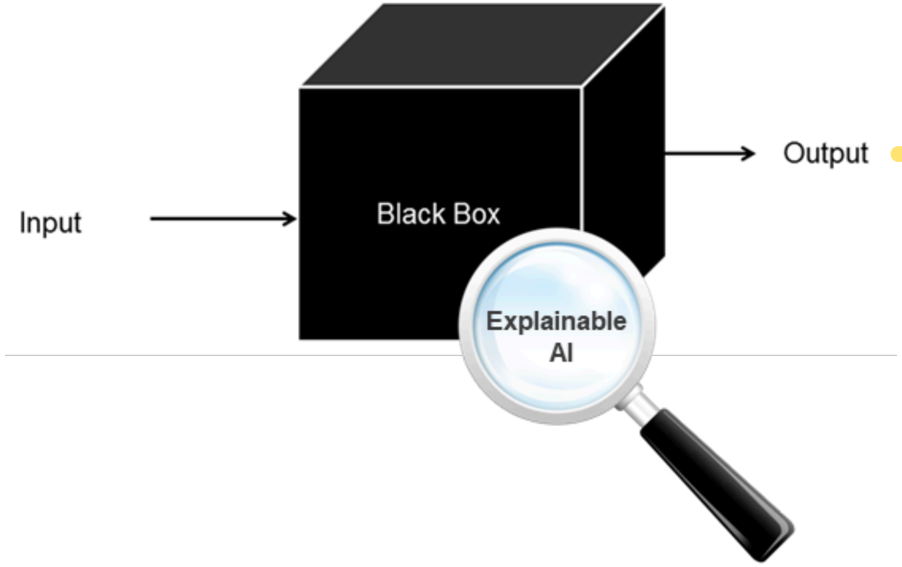
Can we trust our results?

Downside of Neural Networks:
we lack a **full analytical insight** on the process

NN is often considered to be a **black box**

Tests a priori (WIP)

- Test internal features of the NN
- “Analytical” approach



Tests a posteriori

- Test properties of the results
- Empirical approach

Focus on these!

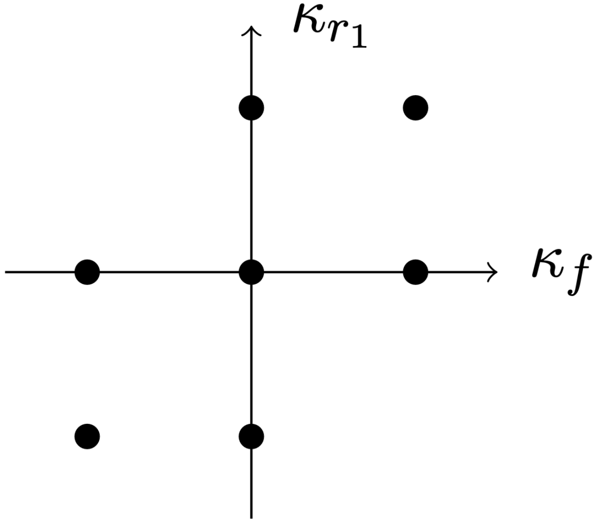
More on the construction: point prescriptions.

Depending on how many (κ_f, κ_r) points among the 9 possible points, one has a different **point prescription**

$$\Delta_{i_a}^{(\pm,0);(\pm,0)} = (\overline{F}(\kappa_f, \kappa_{r_a}) - F)_{i_a} \rightarrow \quad + \rightarrow \kappa_{f,r} = 2.0 \quad - \rightarrow \kappa_{f,r} = 0.5 \quad 0 \rightarrow \kappa_{f,r} = 1.0$$

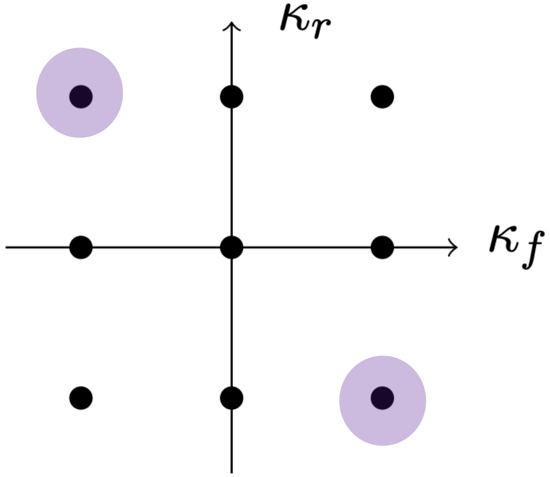
7 points

$$S_{i,j} = \frac{1}{3}\left[\Delta_i^{+0}\Delta_j^{+0} + \Delta_i^{-0}\Delta_j^{-0} + \Delta_i^{0+}\Delta_j^{0+} + \Delta_i^{0-}\Delta_j^{0-} + \Delta_i^{++}\Delta_j^{++} + \Delta_i^{--}\Delta_j^{--}\right]$$



9 points

$$S_{i_1,j_2} = \frac{1}{4}\left[\Delta_i^{+0}\Delta_j^{+0} + \Delta_i^{-0}\Delta_j^{-0} + \Delta_i^{0+}\Delta_j^{0+} + \Delta_i^{0-}\Delta_j^{0-} + \Delta_i^{++}\Delta_j^{++} + \Delta_i^{+-}\Delta_j^{+-} + \Delta_i^{-+}\Delta_j^{-+} + \Delta_i^{--}\Delta_j^{--}\right]$$

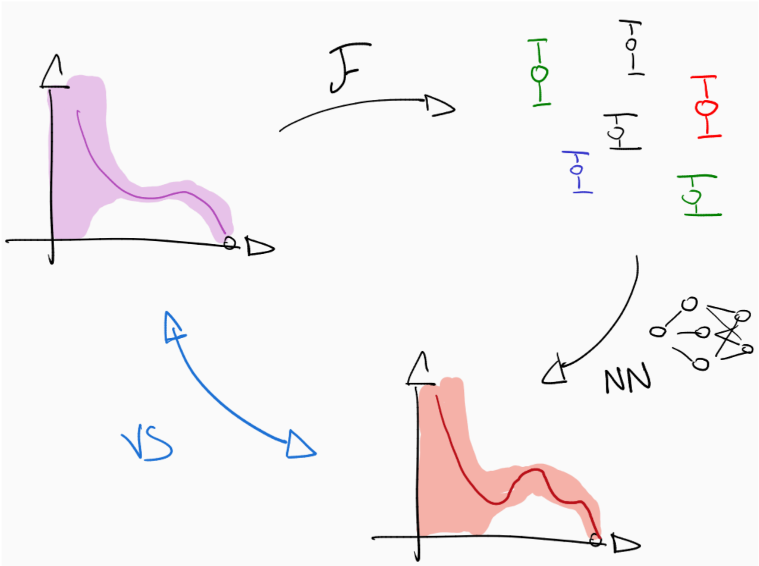


Closure and future tests

Closure test

Test the algorithm in a controlled environment where the **“truth” is known**

1. Choose a PDF as underlying truth
2. Generate central fake data (**LEVEL 0**)
3. Generate smeared fake data with the experimental covariance matrix (**LEVEL 1**)
4. Generate and fit pseudodata replica (**LEVEL 2**)
5. Compare the results with known distribution

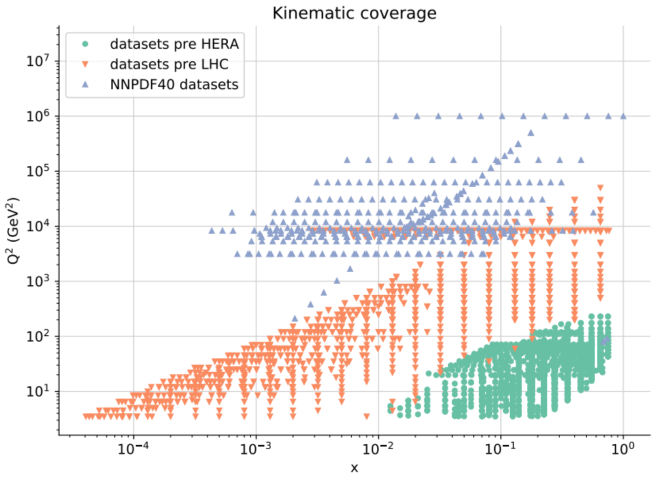


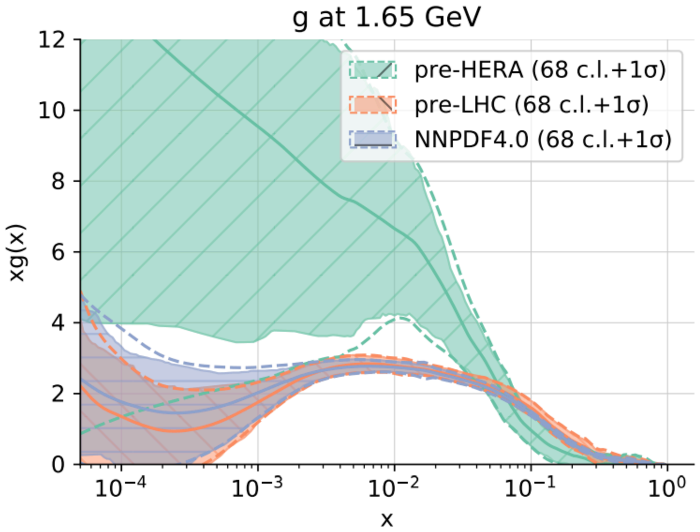
Future test

What about data you have not seen yet?

Traveling in time is not possible but I know **history!**

Divide the dataset **chronologically** and perform a fit for each set: **yesterday’s extrapolation region is today’s data region**



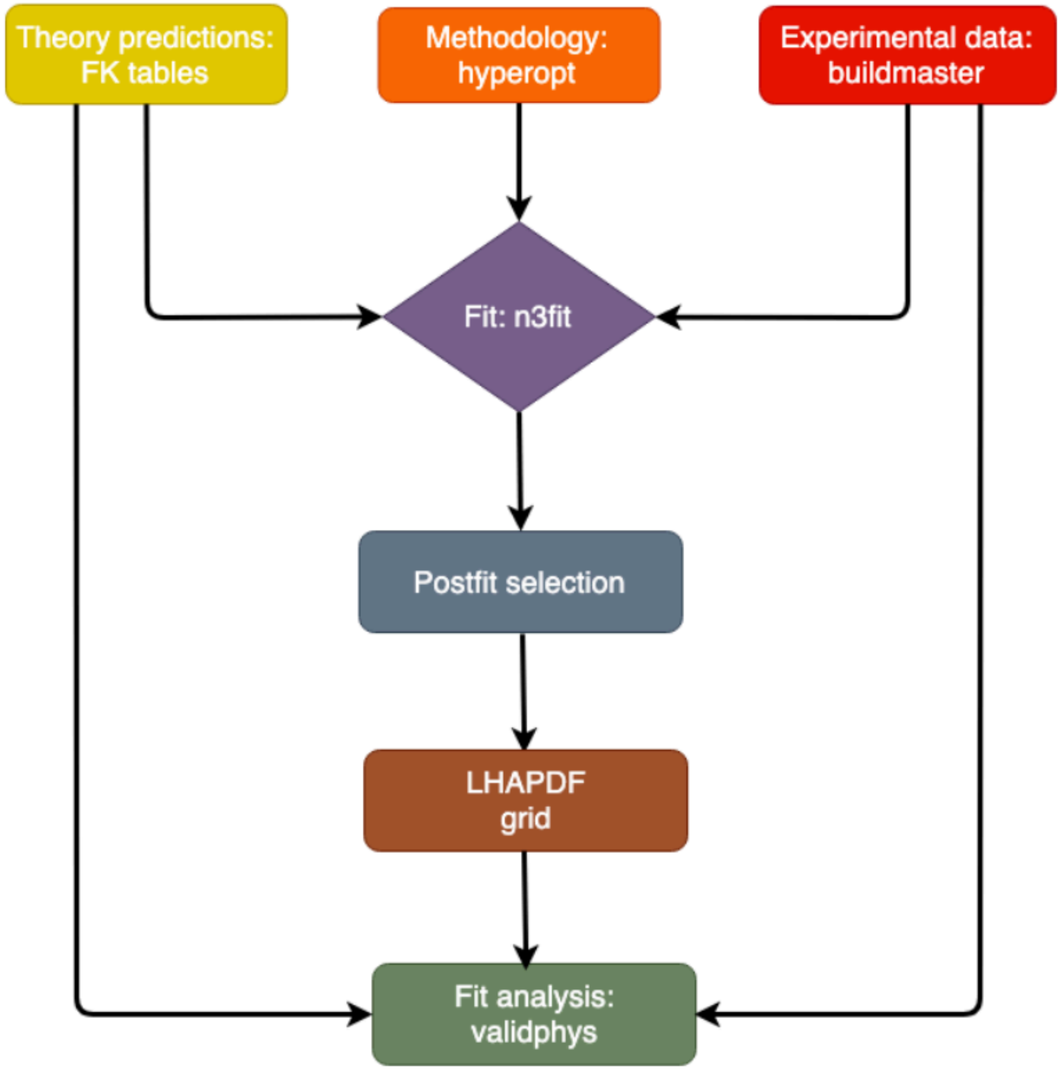


The NNPDF code is open-source

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

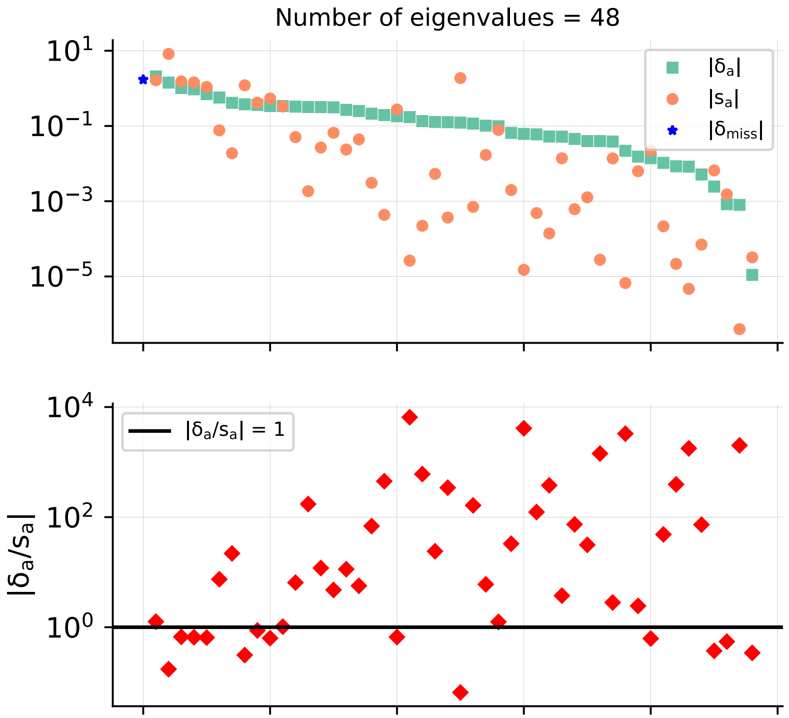
https://github.com/NNPDF/nnpdf

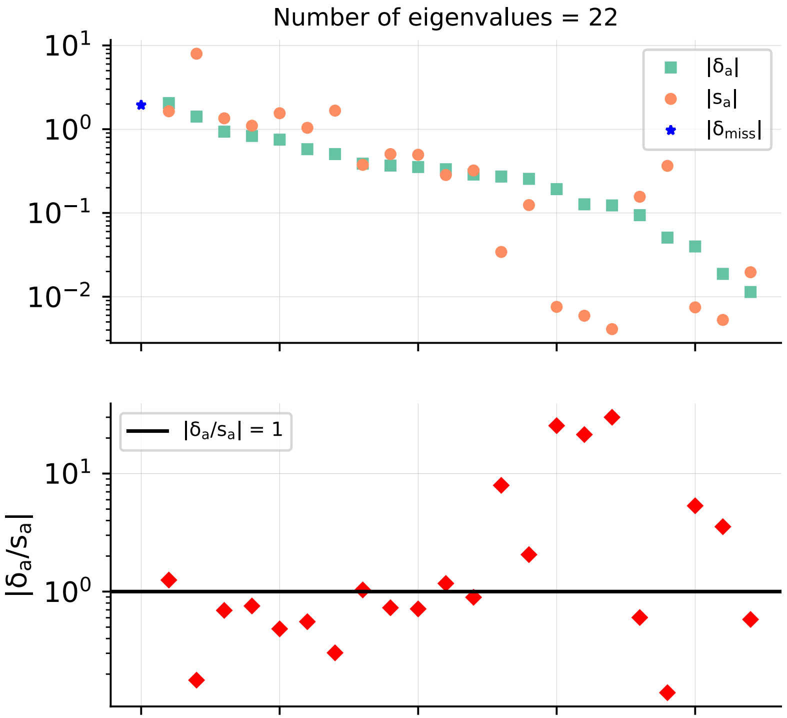
https://docs.nnpdf.science/



Validation: comparing point prescriptions.

$$\delta_i = \left(\frac{F_i^{NNLO} - F_i^{NLO}}{F_i^{NLO}}\right) \rightarrow \delta^\alpha = \sum_{i=1}^{N_D} \delta_i e_i^\alpha \rightarrow \delta_i^S = \sum_{\alpha=1}^{N_{sub}} \delta^\alpha e_i^\alpha \rightarrow \theta = \arccos\left(\frac{|\delta^S|}{|\delta|}\right) \rightarrow \delta_i^{miss} = \delta_i - \delta_i^S$$





Where e^α are the eigenvectors of the theory covariance matrix with eigenvalue $\lambda^\alpha = (s^\alpha)^2$ such that $s^\alpha > 0$

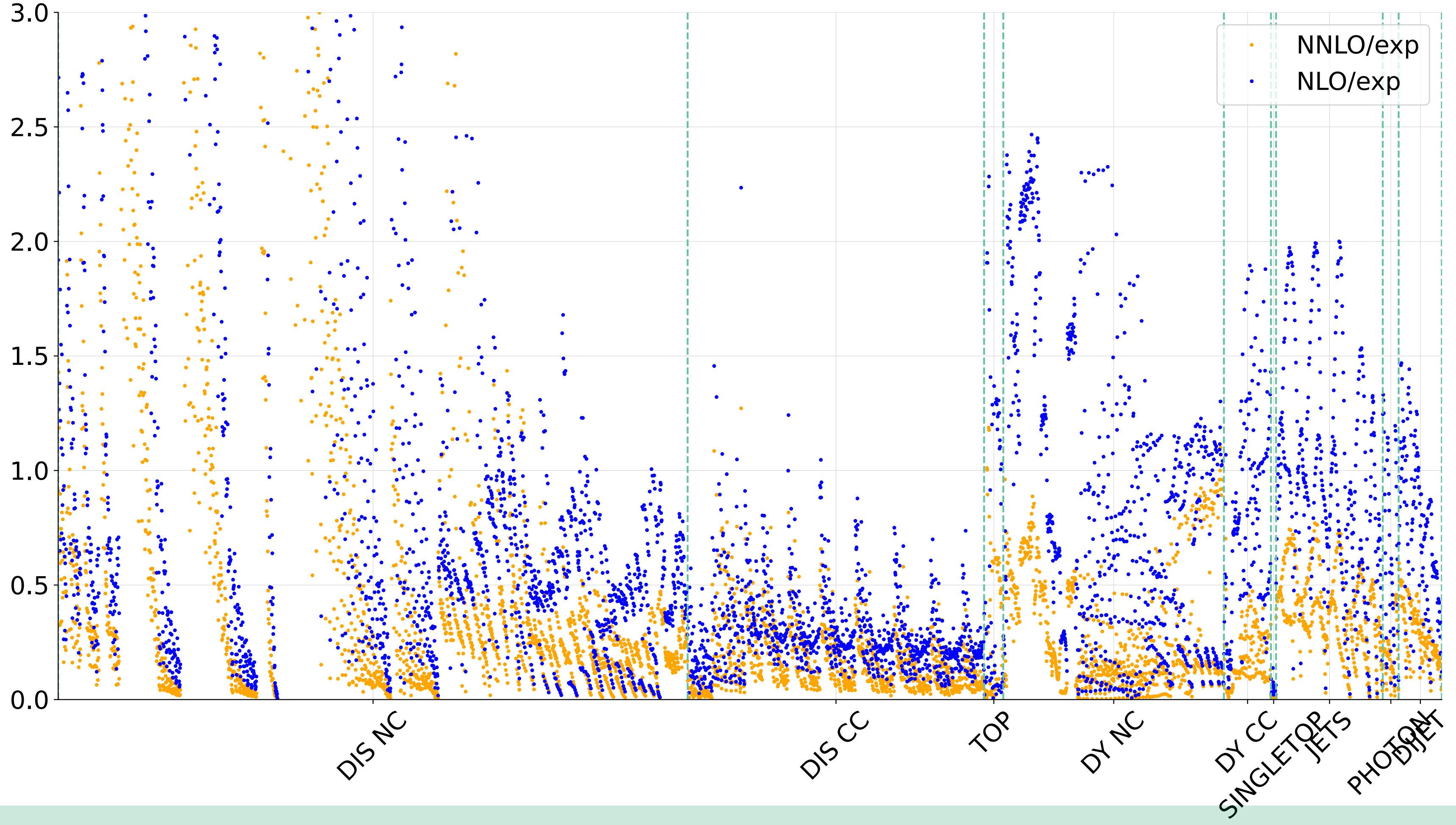
Good agreement for the **largest eigenvalues** with both prescriptions.

9 pts prescription **underestimates** the size of the shift for smaller eigenvalues

However 9 pts prescription performs better in terms of the **angles** θ

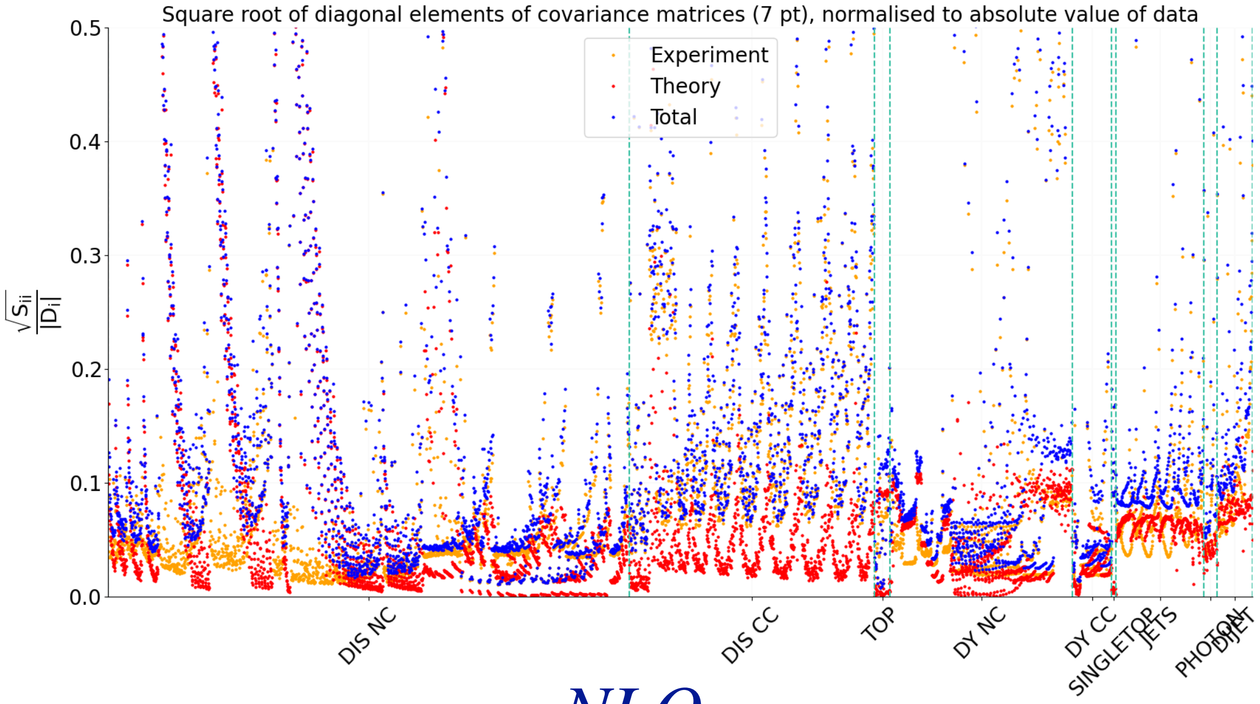
Prescription	$N_{\rm sub}$	DIS NC	DIS CC	TOP	DY NC	DY CC	SINGLETOP	JETS	PHOTON	DIJET	TOTAL
		θ [°]									
7-point	22	39	18	24	23	38	14	15	12	12	32
9-point	48	37	15	20	23	34	12	13	7	12	28

Diagonal elements.

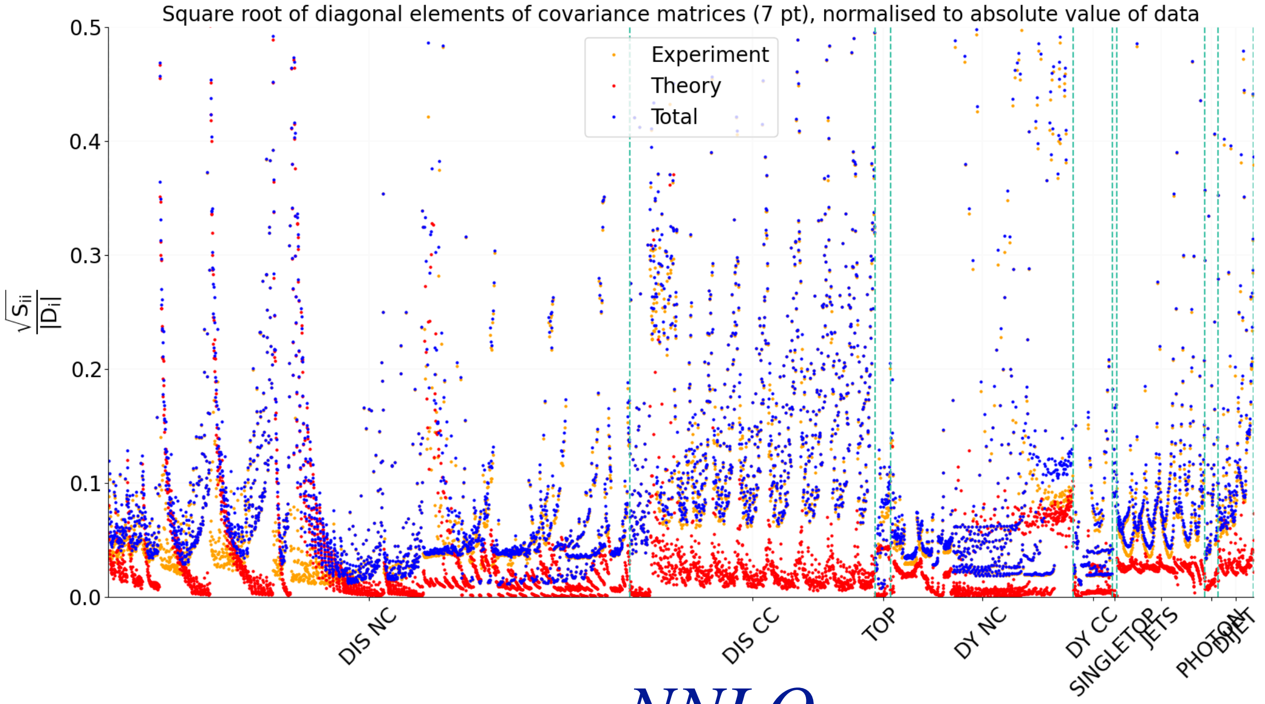


At **NNLO** theory errors are clearly subdominant, while at **NLO** they are of the same size of experimental errors

Diagonal elements.



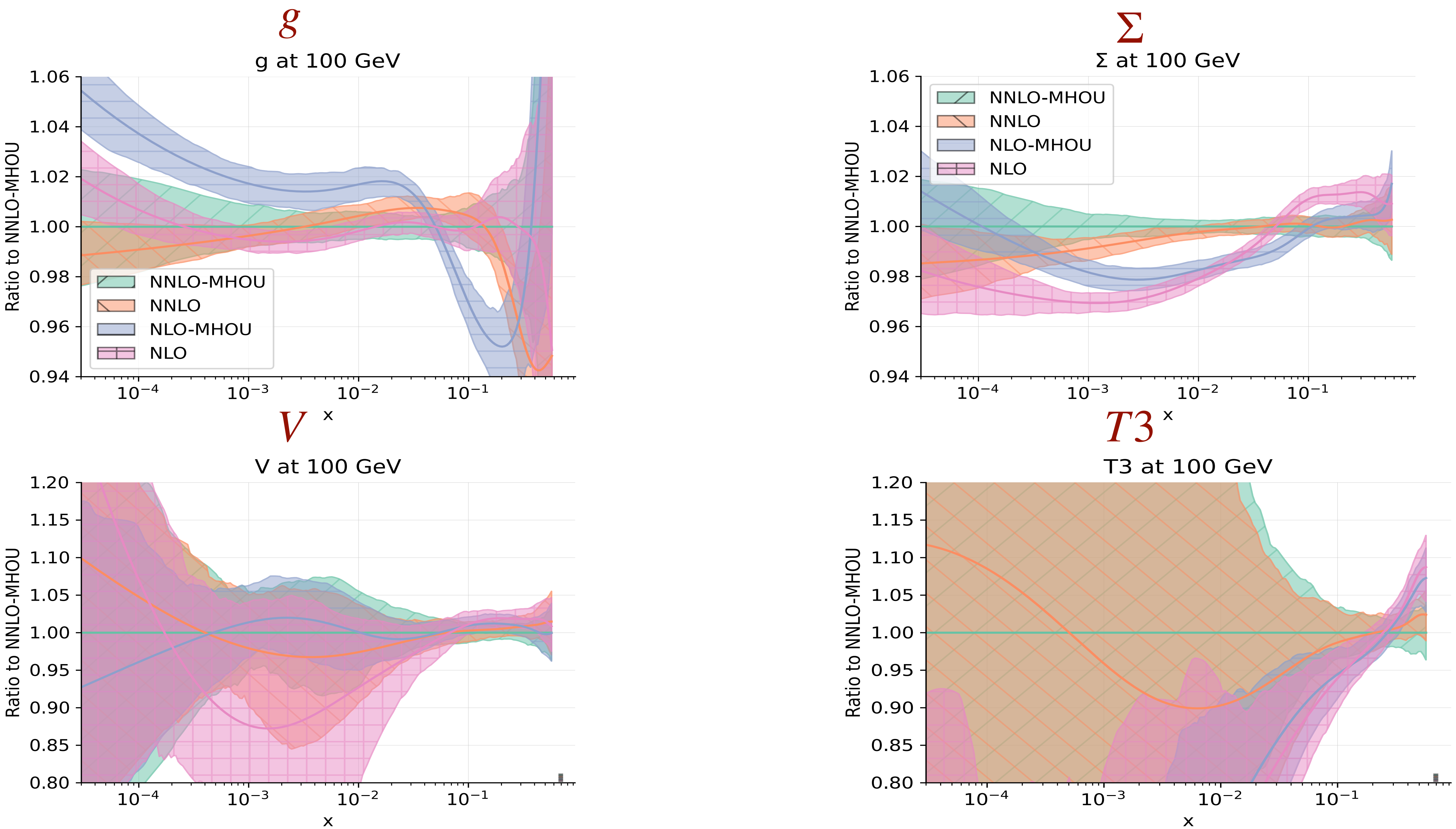
NLO



NNLO

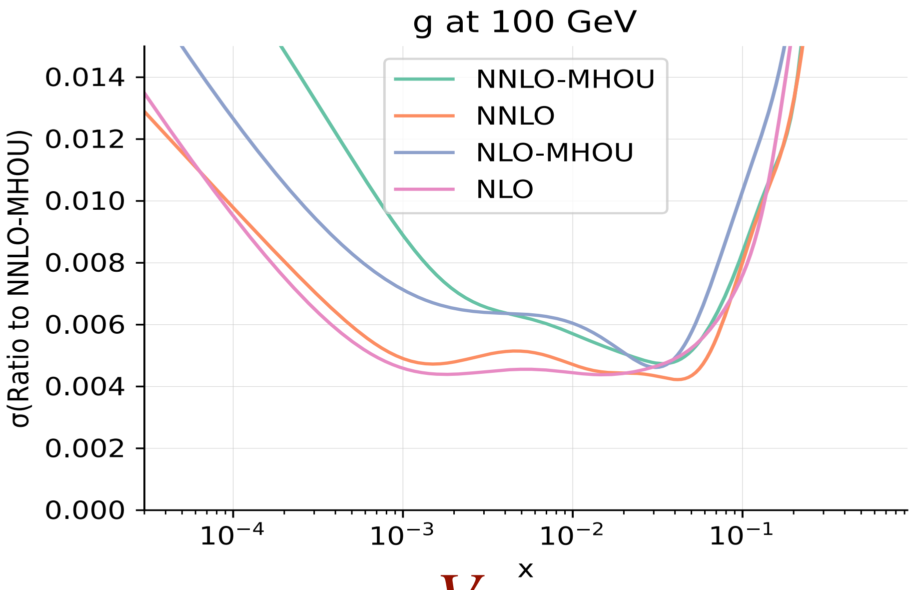
At **NNLO** theory errors are clearly subdominant, while at **NLO** they are of the same size of experimental errors

PDF comparison.

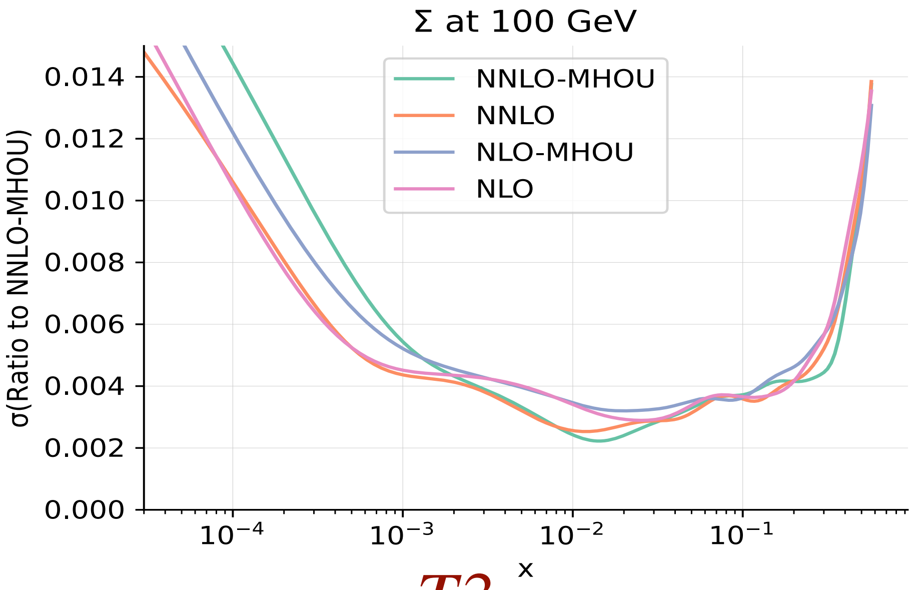


PDF uncertainties.

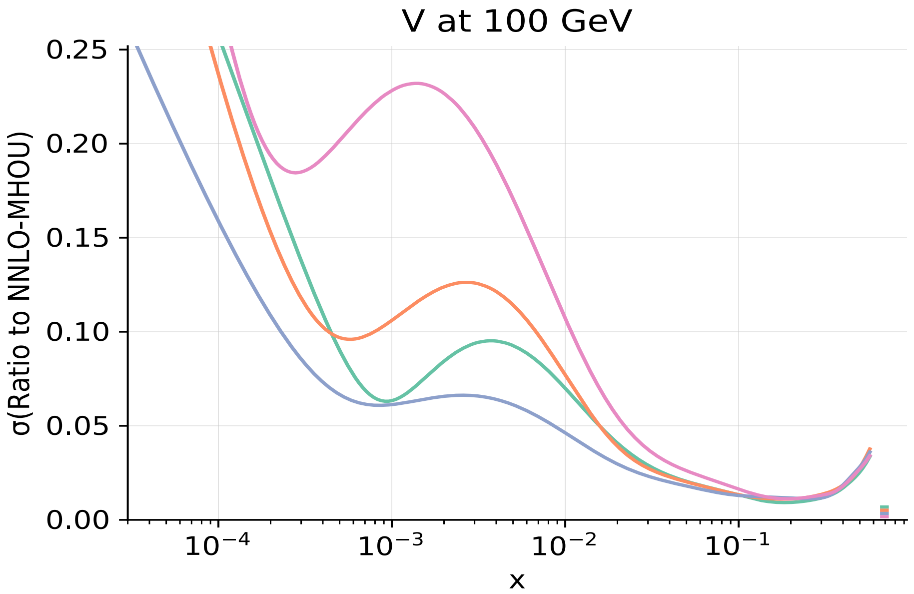
g



Σ



V



$T3$