Statistical tools for global QCD analyses

GFI 1st miniworkshop

Emanuele R. Nocera

Università degli Studi di Torino and INFN — Torino

Oasi di Cavoretto – 29th June 2023

UNIVERSITÀ DI TORINO

PDF fitting in Bayesian language

Inverse Problem

Given a data set F, determine $p(f|F)$ in the sapce of functions $f : [0,1] \to \mathbb{R}$

$$E[\mathcal{O}] = \int \mathcal{D}f\, p(f|F)\mathcal{O}(f) \qquad V[\mathcal{O}] = \int \mathcal{D}f\, p(f|F)[\mathcal{O}(f) - E[\mathcal{O}(f)]]^2$$

Choose a basis in the space of functions

Choose a parametrisation for the functions and a fitting model where
τ: kinematic variables; θ: parameters; h: hyperparameters

$$f = f(\tau, \theta, h) \qquad p(f|F) \to p(\theta|F, h)$$

Sample the statistical distribution of the data with Monte Carlo replicas $F^{(k)}$

The posterior distribution is the product of the likelihood and the prior (Bayes theorem)

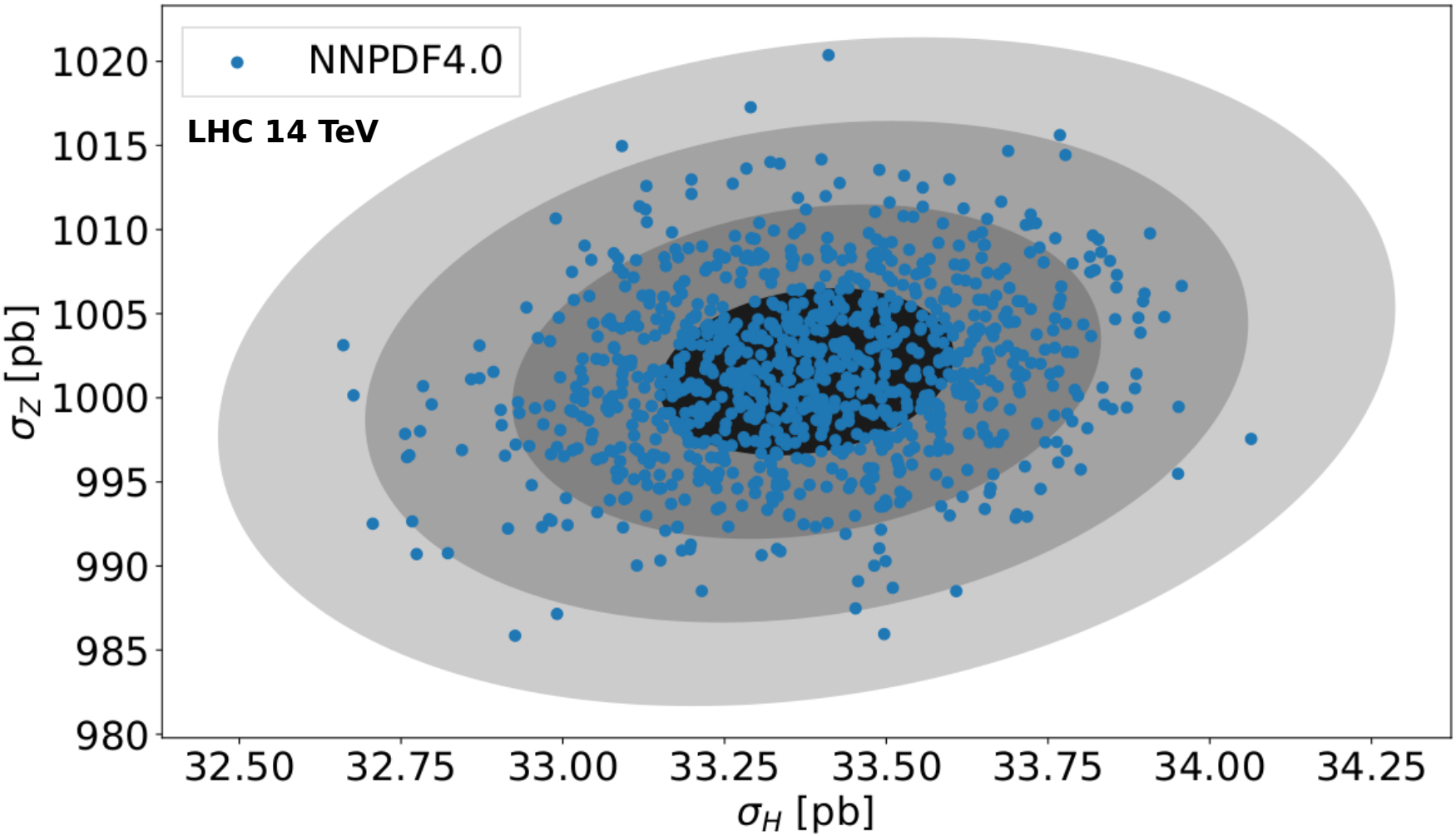
$$p(\theta|F^{(k)}, h) \propto p(F^{(k)}|\theta, h)p(\theta|h)$$

Each replica is the mode of the posterior distribution given one instance of the data

$$\theta^{*(k)} = \arg\max_{\theta} p(\theta|F^{(k)}, h)$$

Hyperoptimise h and determine $\theta^{*(k)}$, for each replica, by $\chi^{2(k)}$ minimisation

Importance sampling

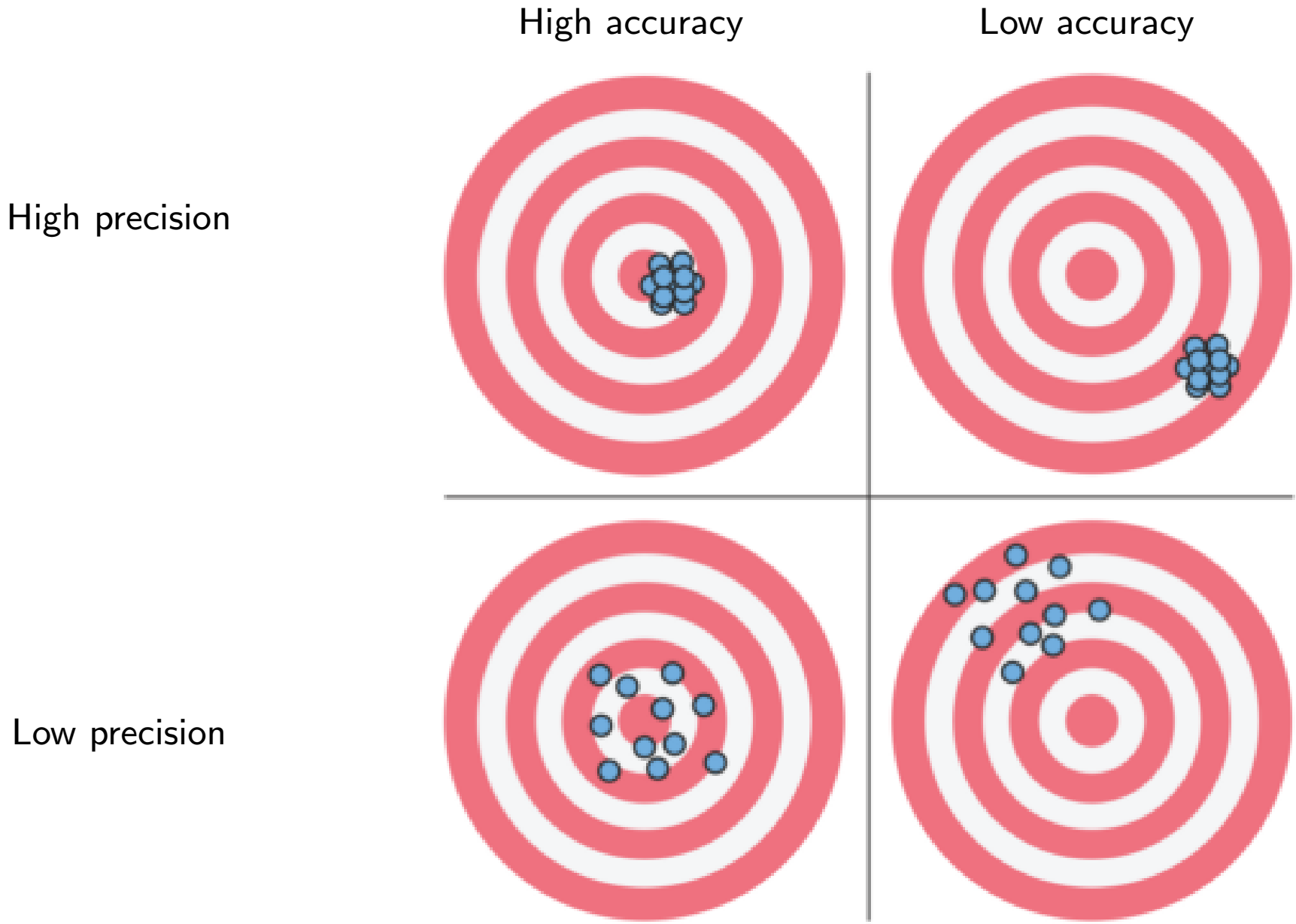


$N_{\rm rep} = 100$ replicas are sufficient to reproduce mean values and uncertainties to 1%

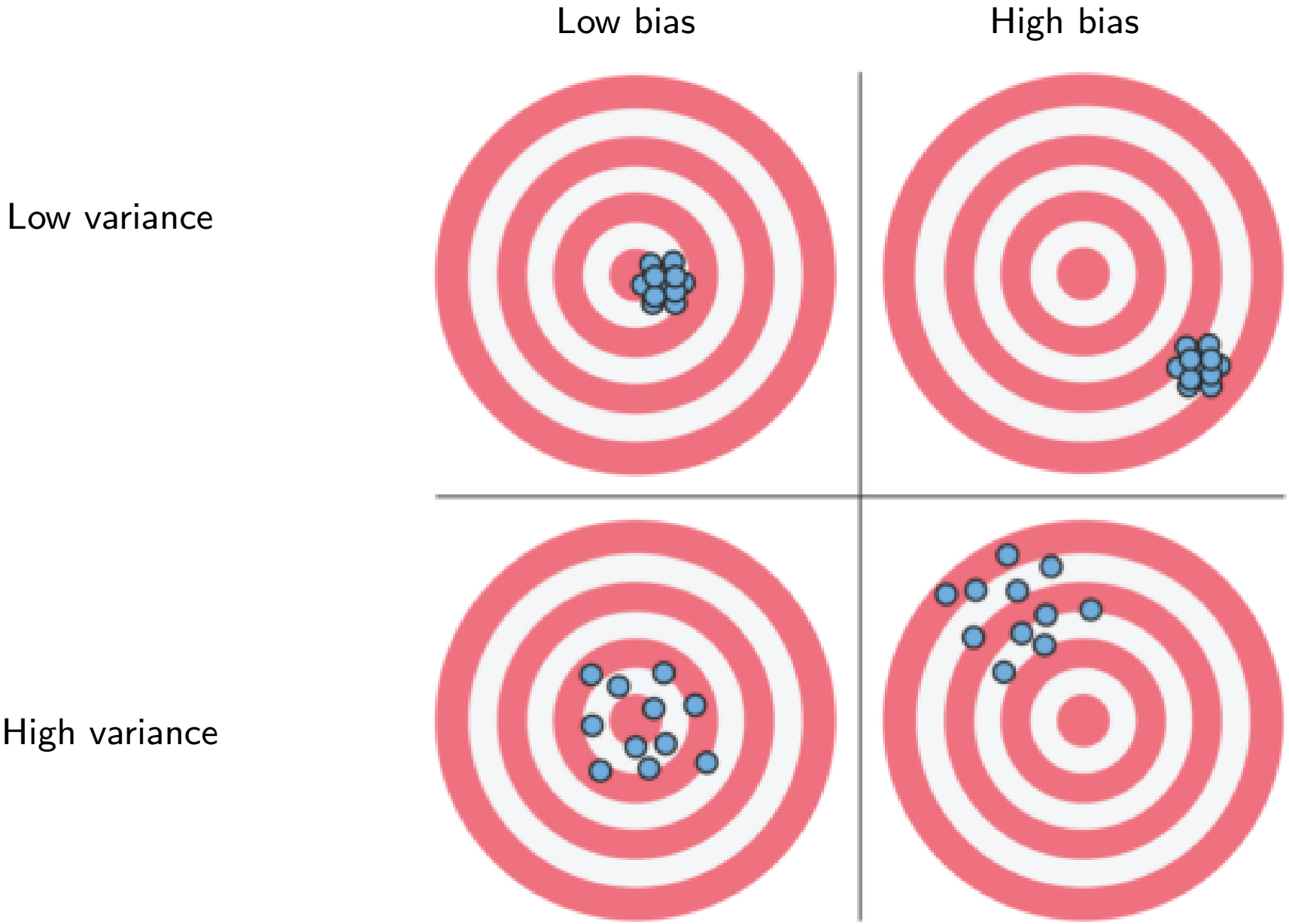
Importance sampling is reproduced correctly (14 replicas out of 1000 vs 98.9% C.I.)

Outliers in the distribution of fitted replicas are good fits to unlikely fluctuations of the data

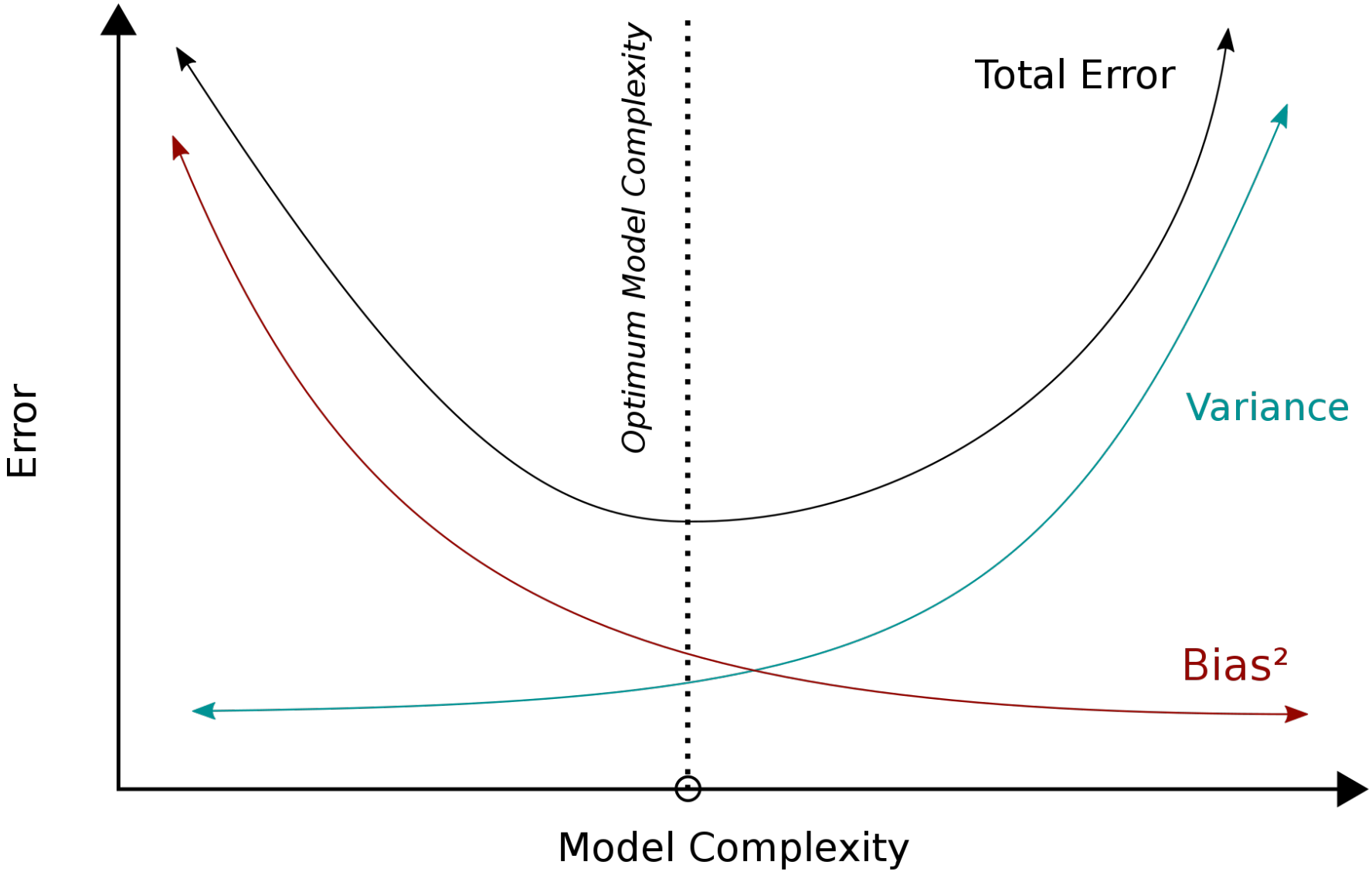
A tale of accuracy and precision



A tale of bias and variance



The bias-variance trade-off



$$\text{squared mean error} = \text{bias}^2 + \text{variance} + \text{irreducible error}$$

1. Goodness-of-fit

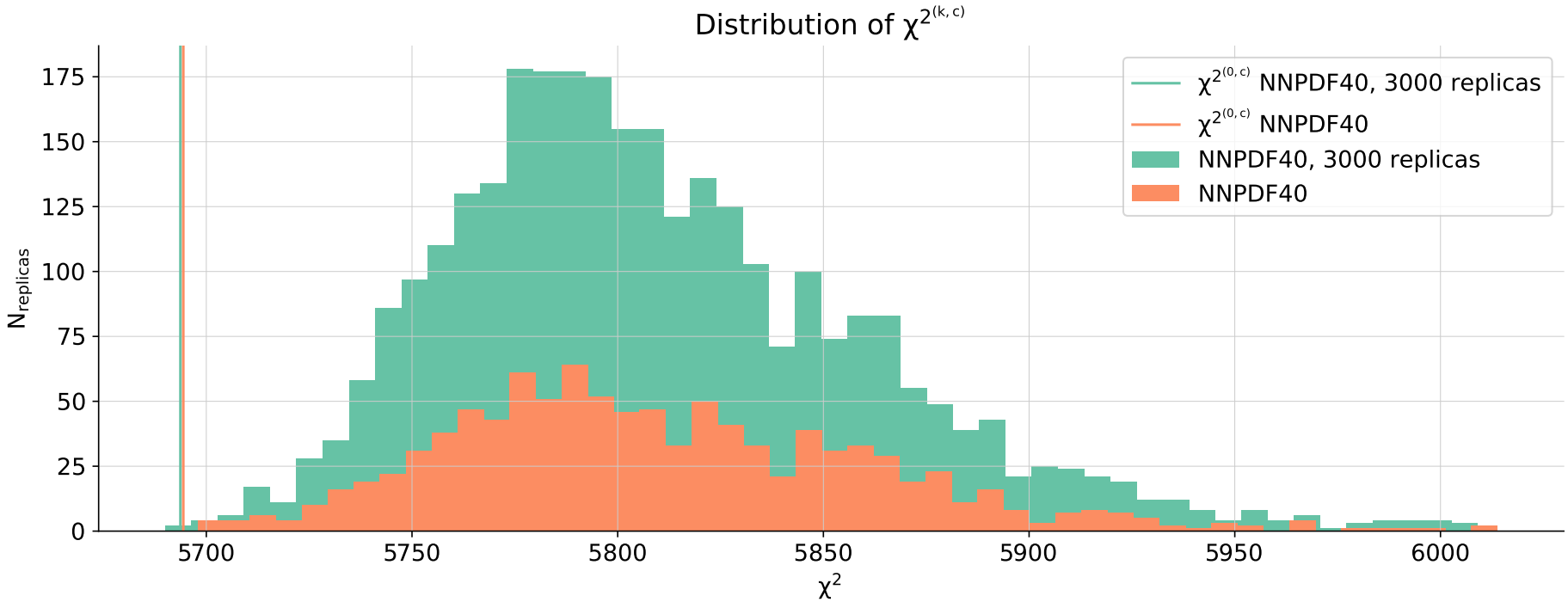
Defining the χ^2

$$\chi^2 = \frac{1}{N_{\rm dat}} \sum_{i,j}^{N_{\rm dat}} \Delta_i \left({\rm cov}_{t_0}^{-1} \right)_{ij} \Delta_j$$

$$\begin{cases} \chi^{2(k)} & \Delta_i = T_i^{(k)} - D_i^{(k)} \\ \chi^{2(k,c)} & \Delta_i = T_i^{(k)} - D_i^{(c)} \\ \chi^{2(0,c)} & \Delta_i = T_i^{0} - D_i^{(c)} \end{cases}$$

t_0 used to avoid D'Agostini bias
[JHEP 05 (2010) 075]
[JHEP 04 (2013) 125]

$$({\rm cov}_{t_0})_{ij} = \delta_{ij} s_i^{\rm (uncorr)} \, s_j^{\rm (uncorr)} + \sum_{m=1}^{N_{\rm norm}} \sigma_{i,m}^{\rm (norm)} \sigma_{j,m}^{\rm (norm)} T_i^{(0)} T_j^{(0)} + \sum_{l=1}^{N_{\rm corr}} \sigma_{i,l}^{\rm (corr)} \sigma_{j,l}^{\rm (corr)} D_i D_j ,$$



Normalisation uncertainties: D'Agostini bias [JHEP 1005 (2010) 075]

1. Consider one experiment with $N_{\rm dat}$ data d_i of one theoretical quantity t

$$\chi^2(t) = \sum_{i,j}^{N_{\rm dat}} (t - d_i) \left({\rm cov}^{-1}\right)_{ij} (t - d_j)$$

2. The best-fit theoretical quantity t_0 and its variance v_t are given by

$$\left.\frac{d\chi^2}{dt}\right|_{t=t_0} = 0 \Longleftrightarrow t_0 = \frac{\sum_{i,j}^{N_{\rm dat}} \left({\rm cov}^{-1}\right)_{ij} d_j}{\sum_{i,j}^{N_{\rm dat}} \left({\rm cov}^{-1}\right)_{ij}} \qquad v_t = \left(\frac{1}{2}\frac{d^2\chi^2}{dt^2}\right)^{-1} = \frac{1}{\sum_{i,j}^{N_{\rm dat}} \left({\rm cov}^{-1}\right)_{ij}}$$

3. Consider completely uncorrelated additive errors: $({\rm cov})_{ij} = s_i^2 \delta_{ij}$

$$t_0 = w = \Sigma^2 \sum_i^{N_{\rm dat}} \frac{d_i}{s_i^2} \qquad v_t = \Sigma^2 \qquad \text{with } \frac{1}{\Sigma^2} = \sum_i^{N_{\rm dat}} \frac{1}{s_i^2}$$

4. Consider an additional common normalisation error: $({\rm cov})_{ij} = (s_i^2 + \sigma^2 d_i^2)\delta_{ij}$

$$t_0 = \frac{w}{1 + r^2\sigma^2 w^2/\Sigma^2} \qquad v_t = \frac{\Sigma^2 + \sigma^2 w^2 (1 + r^2)}{1 + r^2\sigma^2 w^2/\Sigma^2} \qquad \text{with } r^2 = \frac{\Sigma^2}{w^2} \sum_i^{N_{\rm dat}} \frac{(d_i - w)^2}{s_i^2}$$

5. Both t_0 and v_t are affected by a downward bias
   smaller values of d_i have a smaller normalization uncertainties σd_i and are thus preferred

Normalisation uncertainties: D'Agostini bias [JHEP 1005 (2010) 075]

1. The penalty trick: redefine the fit quality

$$\chi^2(t) \to \chi^2(t,\mathcal{N}) = \sum_i^{N_{\rm dat}} \frac{(t/\mathcal{N}-d_i)^2}{s_i^2} + \frac{(\mathcal{N}-1)^2}{\sigma^2}$$

$$\left.\frac{\partial\chi^2}{\partial t}\right|_{t=t_0} = \frac{\partial\chi^2}{\partial\mathcal{N}} = 0 \Longleftrightarrow t_0 = w \qquad v_t = \left(\frac{1}{2}\frac{d^2\chi^2}{dt^2}\right)^{-1} = \Sigma^2 + \sigma^2 w^2$$

$\longrightarrow$ recover the unbiased estimators for t_0 and v_t

2. The t_0 method: redefine the covariance matrix

$$(\mathrm{cov})_{ij} \to (\mathrm{cov}_{t_0})_{ij} \Longleftrightarrow (s_i^2 + \sigma^2 d_i^2)\delta_{ij} \to s_i^2\delta_{ij} + \sigma^2 t_0^2$$

$$\left(\mathrm{cov}_{t_0}^{-1}\right)_{ij} = \frac{\delta_{ij}}{s_i^2} - \frac{\sigma^2 t_0^2}{s_i^2 s_j^2}\frac{\Sigma^2}{\Sigma^2 + \sigma^2 t_0^2} \Longleftrightarrow t_0 = w \qquad v_t = \Sigma^2 + \sigma^2 w^2$$

$\longrightarrow$ recover the unbiased estimators for t_0 and v_t, provided that w is tuned to t_0
$\longrightarrow$ w can be tuned to t_0 via an iterative procedure

The d'Agostini bias and its solution can be generalised to more than one experiment with different normalisation errors (per experiment/per data point)

2. Consistency of data

Data inconsistency: tensions between data sets

Give more weight to a data set p

$$\chi^2 \to \chi^2 + w\chi^2_p$$

Refit: the total χ^2 will increase

Which data sets get worse? How much?

Refit: the data set χ^2_p will decrese

Self-consistency? Inconsistency?

Examples: ATLAS W, Z and $t\bar{t}$

Inconsistency clearly spotted

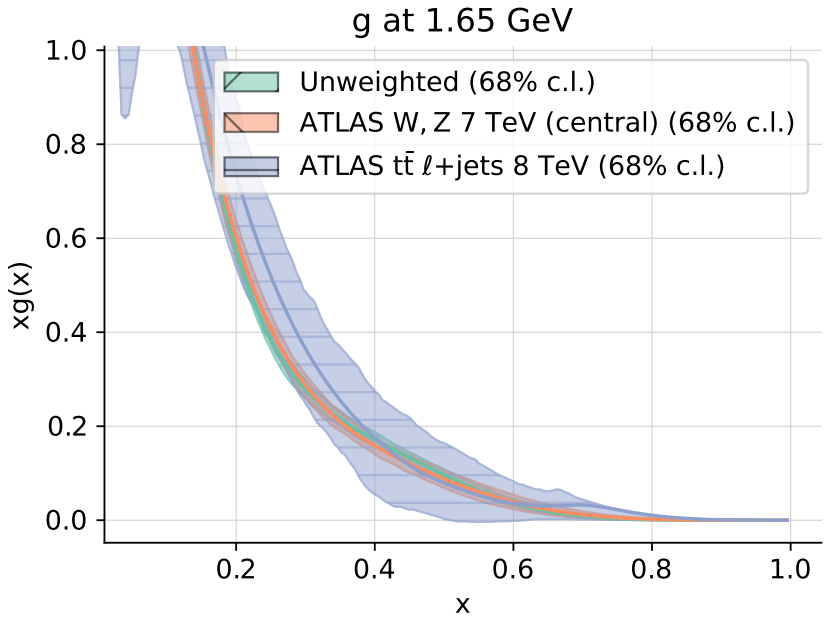
unnatural PDF shapes appear

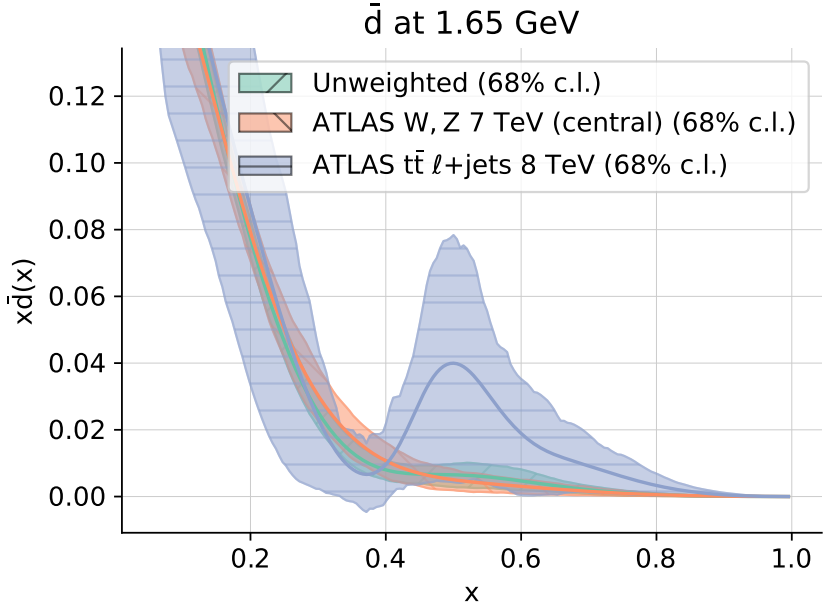
error in other data sets increases

Otherwise global fit quality

and PDFs remain unaltered

Data set	baseline	rw W, Z	rw $t\bar{t}$
ATLAS W, Z 7 TeV	1.86	1.23	—
ATLAS $t\bar{t}$ 8 TeV	4.11	—	1.21
Total	1.20	1.21	1.73





Data inconsistency: experimental correlations

Single inclusive jet data from ATLAS 7 TeV

default correlations: terrible χ^2

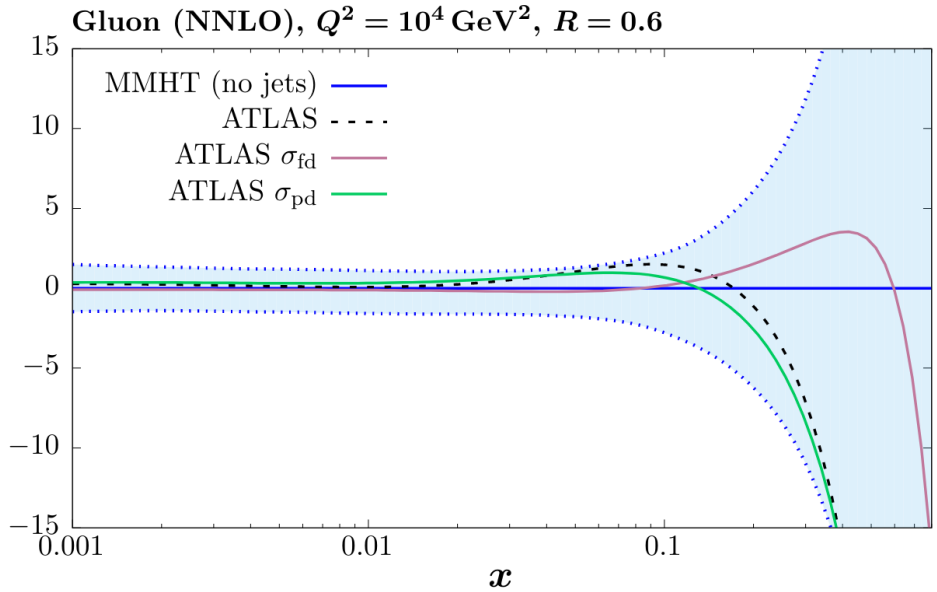
(correlations across rapidity bins)

decorrelation models: improve the fit a lot

$n_{\rm dat}$	default	part. decorr.	full decorr.
140	1.89	1.28	0.83

no significant effect on the extracted gluon

similar gluon irrespective of the rapidity bin



[EPJ C78 (2018) 248; EPJ C80 (2020) 797]

Top pair production from ATLAS 8 TeV

default correlations: terrible χ^2

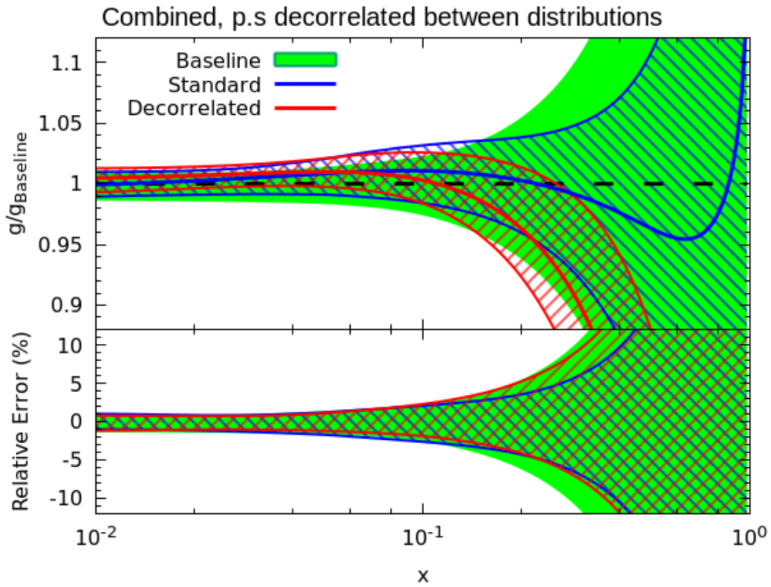
(correlations across different spectra)

decorrelation models: improve the fit a lot

$n_{\rm dat}$	default	stat. uncorr.	p.s. uncorr
25	7.00	3.28	1.80

appreciable effect on the extracted gluon

different gluon depending on the top spectrum



[EPJ C80 (2020) 1; Les Houches proceedings, 2019]

Good knowledge of experimental correlations is important

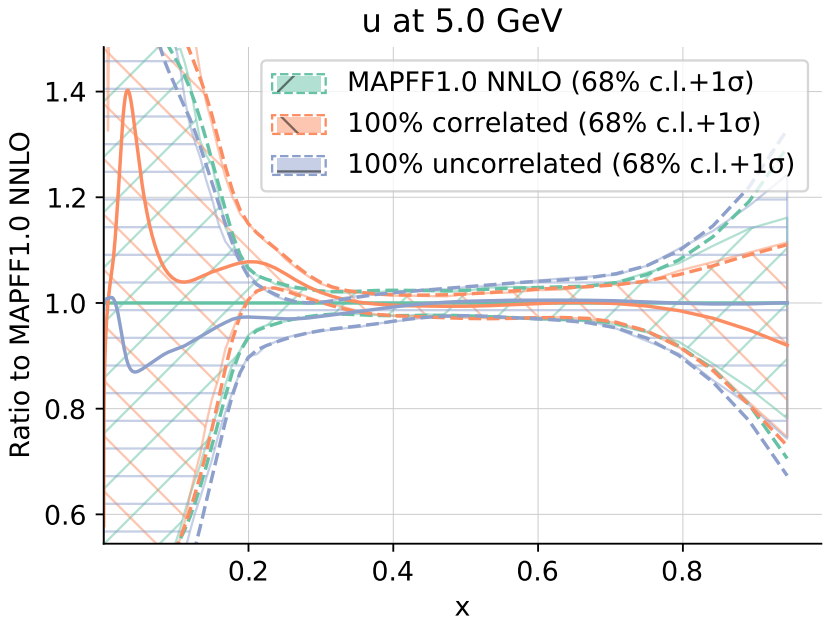
Consider the COMPASS $\pi^{\pm}$ multiplicities

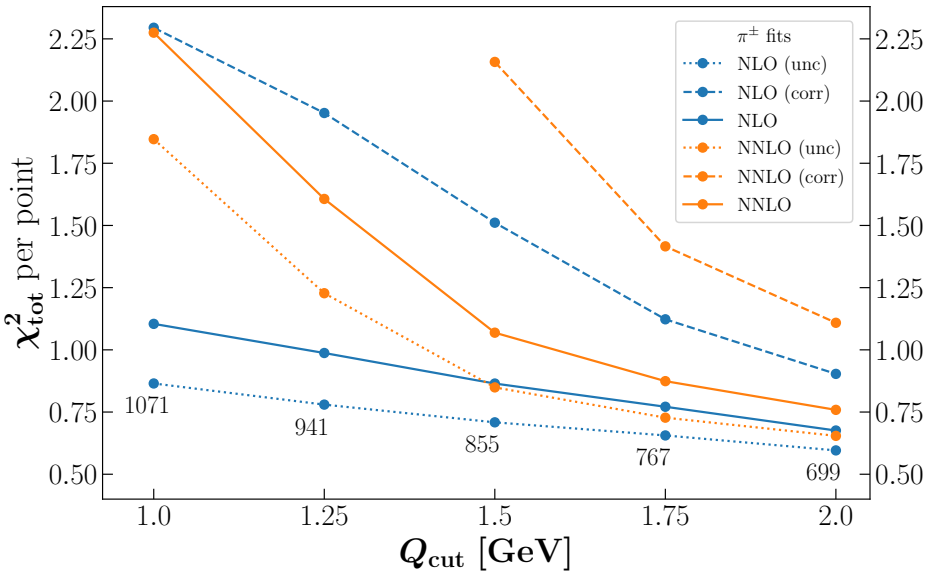
[PLB 764 (2017) 1]

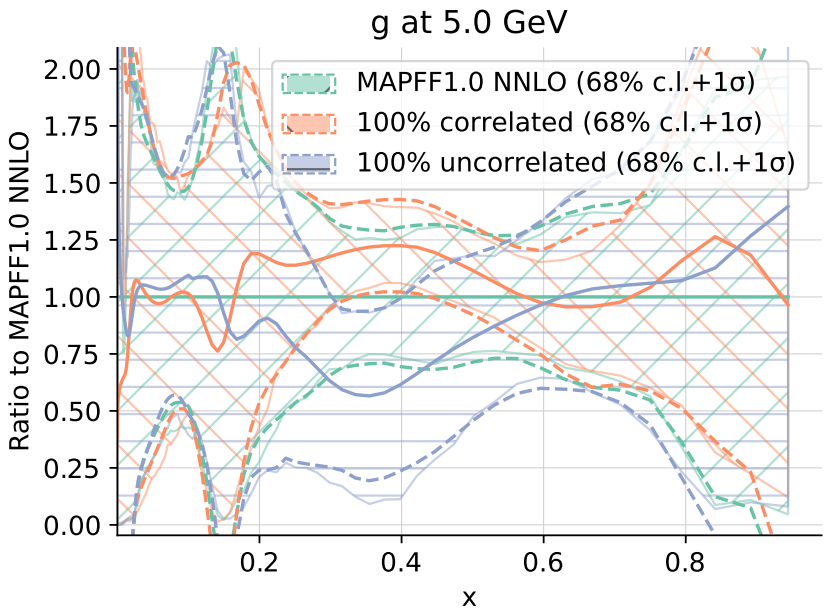
Only 80% of the systematic uncertainty is bin-by-bin correlated

What if you incorporate a different piece of information in a FF fit?

Consider two cases: [arXiv:2204.10331]
full correlation; full decorrelation







3. Consistency of methodology

Closure tests [EPJ C77 (2017) 663; EPJ C82 (2022) 330]

A test to validate PDF uncertainties in the data region

Fit PDFs to pseudodata generated
assuming a known underlying law

Define bias and variance

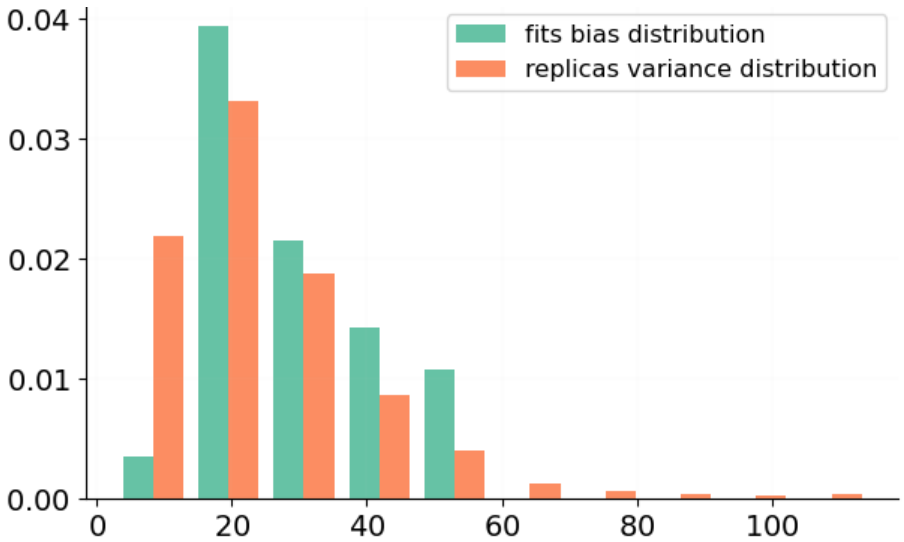
bias difference of central prediction and truth

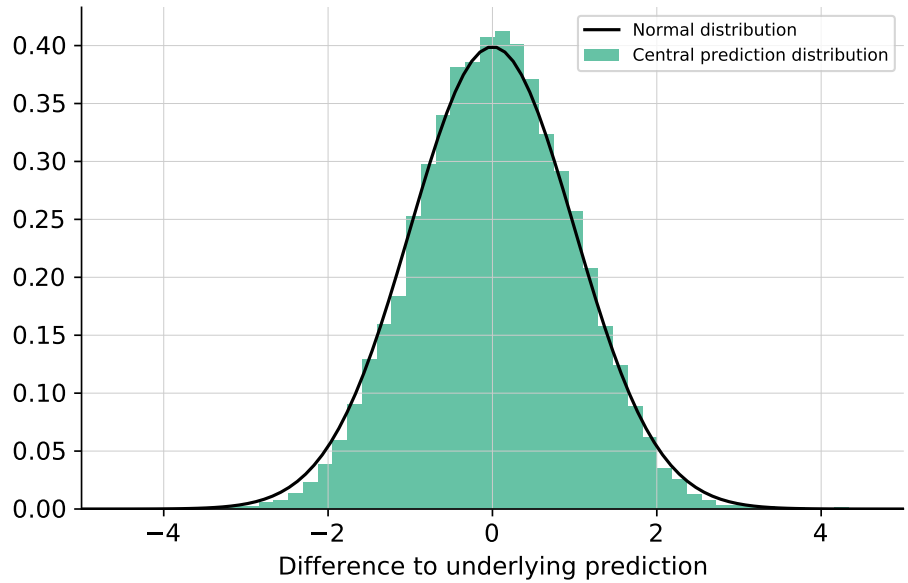
variance uncertainty of replica predictions

If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each





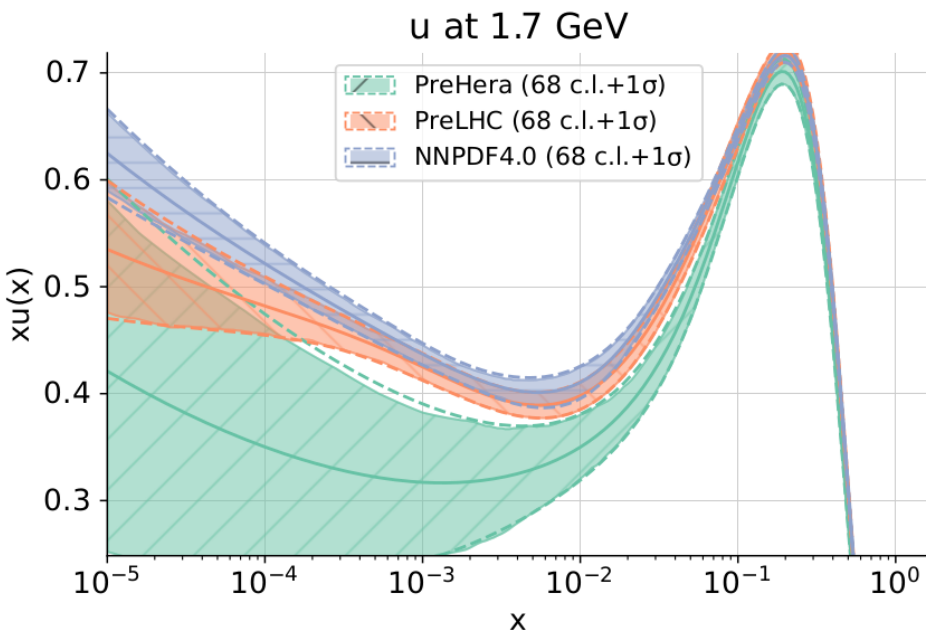
Future tests [Acta Phys.Polon. B52 (2021) 243]

A test to validate PDF uncertainties in the extrapolation regions

Test PDF uncertainties on data sets
not included in a given PDF fit
that cover unseen kinematic regions

Data set	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

Only exp. cov. matrix



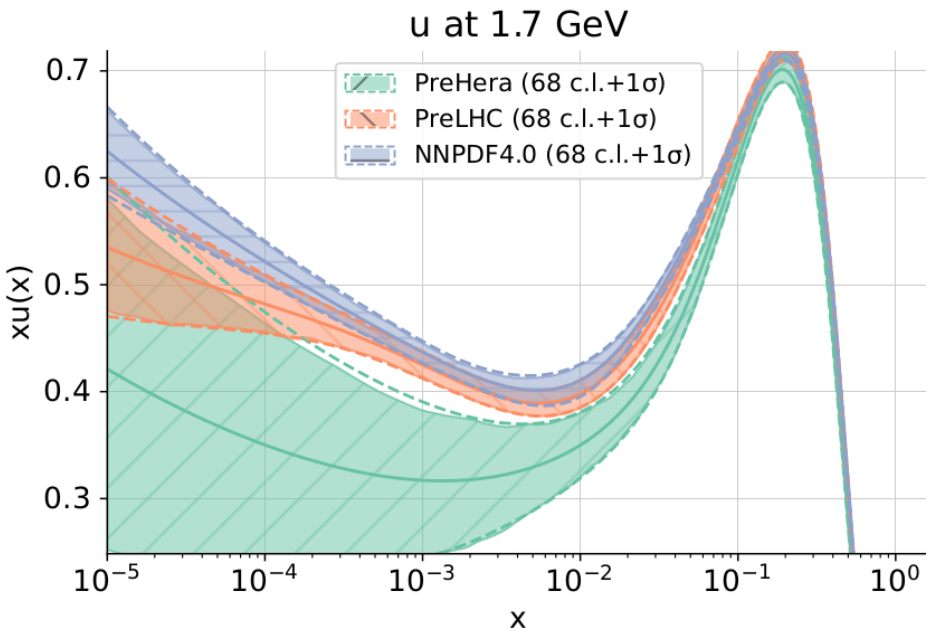
Future tests [Acta Phys.Polon. B52 (2021) 243]

A test to validate PDF uncertainties in the extrapolation regions

Test PDF uncertainties on data sets
not included in a given PDF fit
that cover unseen kinematic regions

Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA			0.86
pre-LHC		1.17	1.22
NNPDF4.0	1.12	1.30	1.38

Exp+PDF cov. matrix

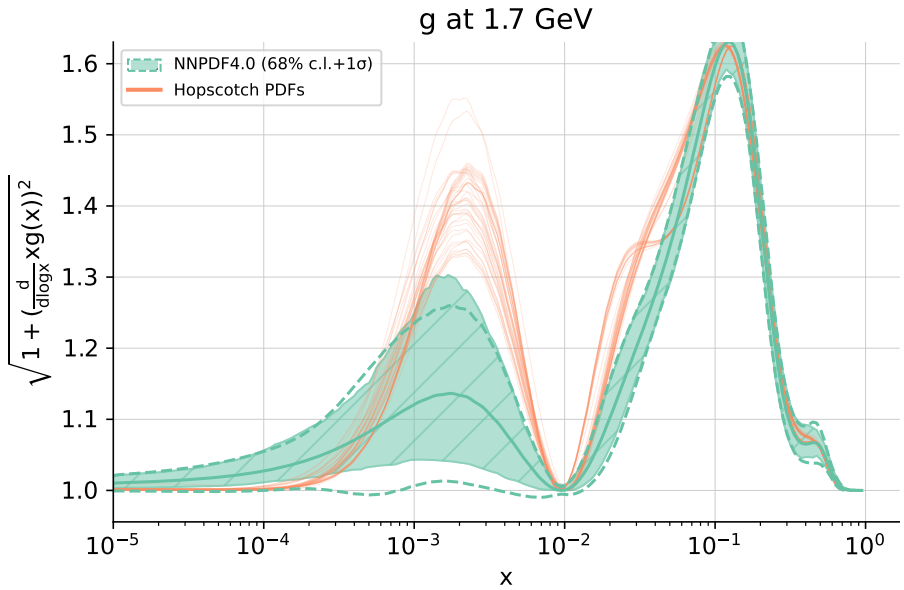


How to test for overfitting?

Metric 1: Kinetic Energy

$$\mathrm{KE} = \sqrt{1 + \left(\frac{d}{d\ln x} x f(x, Q^2)\right)^2}$$

A measure of PDF wiggliness:
the higher the KE, the longer the curve that joins two fixed points



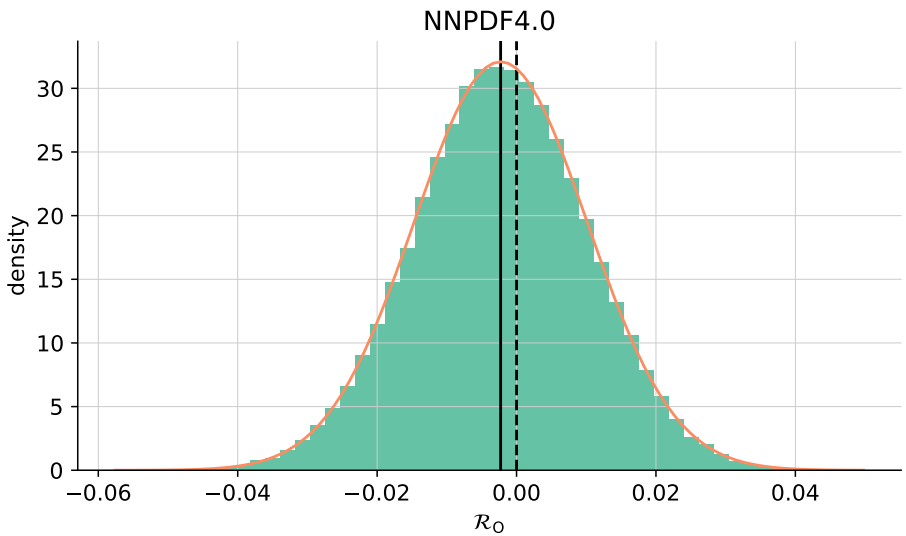
How to test for overfitting?

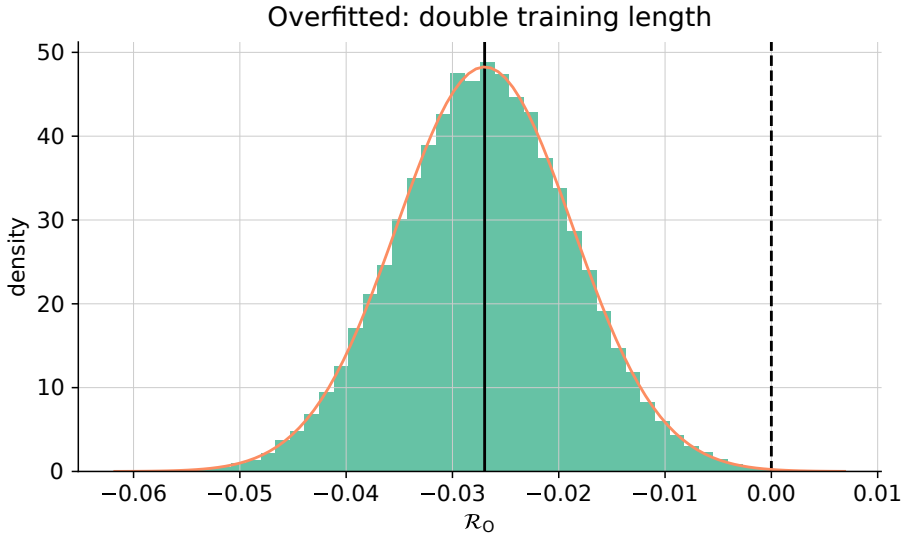
Metric 2: Overfit metric

$$\mathcal{R}_O = \chi^2_{\rm val}\left[T^{(k)}, D^{(k)}\right] - \overline{\chi^2_{\rm val}}\left[T^{(k)}, D^{(k)}\right]$$

$$\overline{\chi^2_{\rm val}}\left[T^{(k)}, D^{(k)}\right] \equiv \frac{1}{N}\sum_{k'=1}^{N} \chi^2_{\rm val}\left[T^{(k)}, D^{(k')}\right]\Big|_{\rm fixed\ mask}$$

A measure of PDF overfitting:
for each replica, there is a different fluctuation and training/validation split (mask)
if $\mathcal{R}_O < 0$ the replica is overfitted, *i.e.* it contains information specific to D^k





$\mathcal{R}_O = -0.001 \pm 0.013$ (NNPDF4.0)

$\mathcal{R}_O = -0.027 \pm 0.001$ (overfitted NNPDF4.0)

4. Consistency of theory

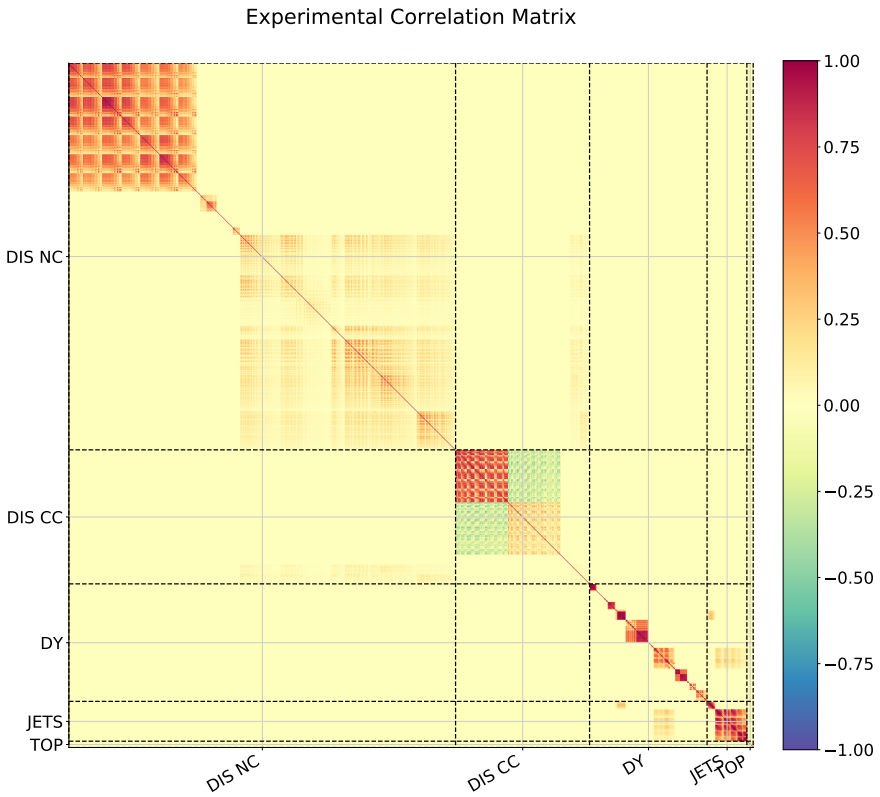
Theory uncertainties in PDF determination

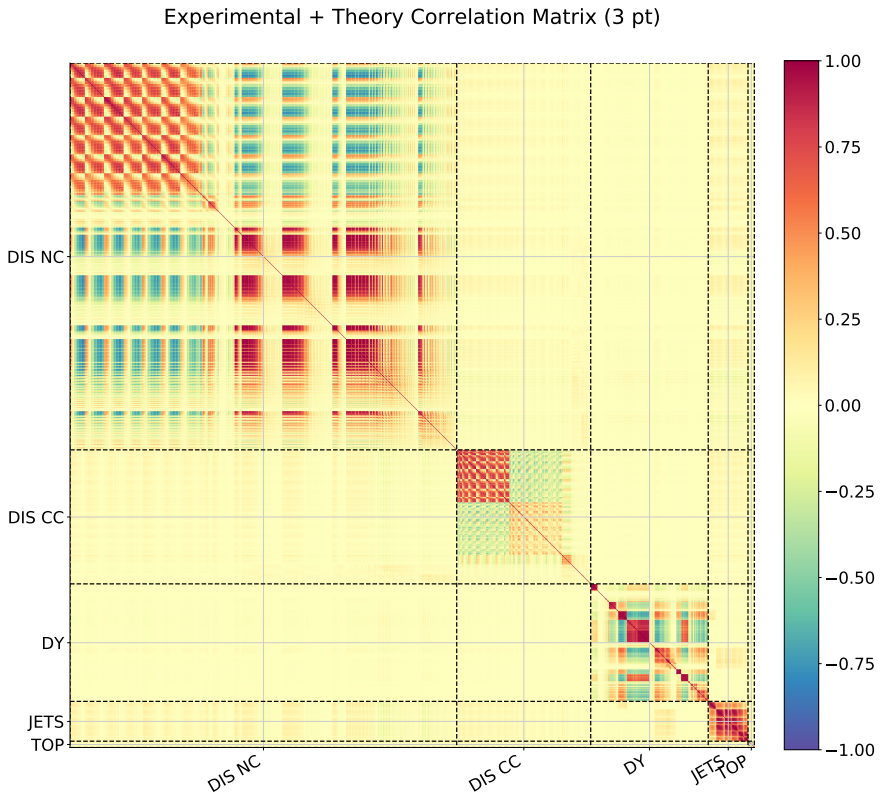
Assuming that theory uncertainties are (a) Gaussian and (b) independent from experimental uncertainties, modify the figure of merit to account for theory errors

$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} (D_i - T_i)(\mathrm{cov}_{\rm exp} + \mathrm{cov}_{\rm th})^{-1}_{ij}(D_j - T_j);\ (\mathrm{cov}_{\rm th})_{ij} = \frac{1}{N}\sum_k^N \Delta_i^{(k)}\Delta_j^{(k)};\ \Delta_i^{(k)} \equiv T_i^{(k)} - T_i$$

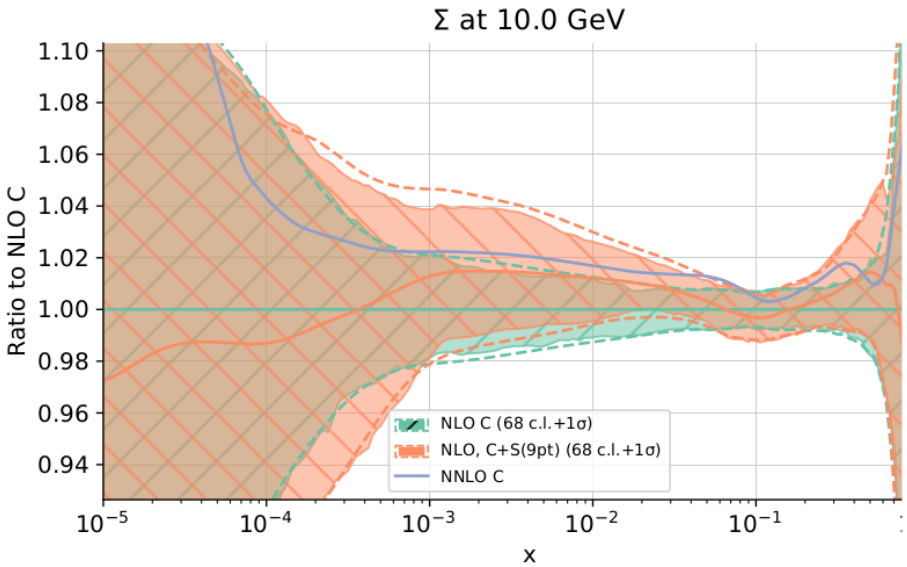
Problem reduced to estimate the th. cov. matrix, *e.g.* in terms of nuisance parameters

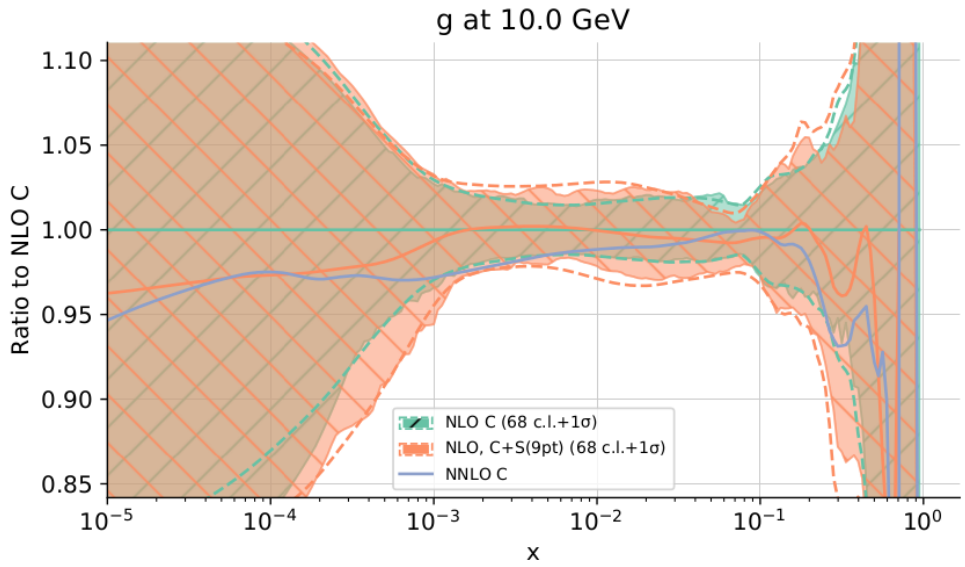
$$\Delta_i^{(k)} = T_i(\mu_R, \mu_F) - T_i(\mu_{R,0}, \mu_{F,0});\ \text{vary scales in } \frac{1}{2} \leq \frac{\mu_F}{\mu_{F,0}}, \frac{\mu_R}{\mu_{R,0}} \leq 2$$





Theory uncertainties in PDF determination





PDF uncertainty increase encapsulates NLO-NNLO shift

Overall (rather small) increase in uncertainties

Increase in PDF uncertainties due to replica generation
is counteracted by extra correlations in fitting minimisation

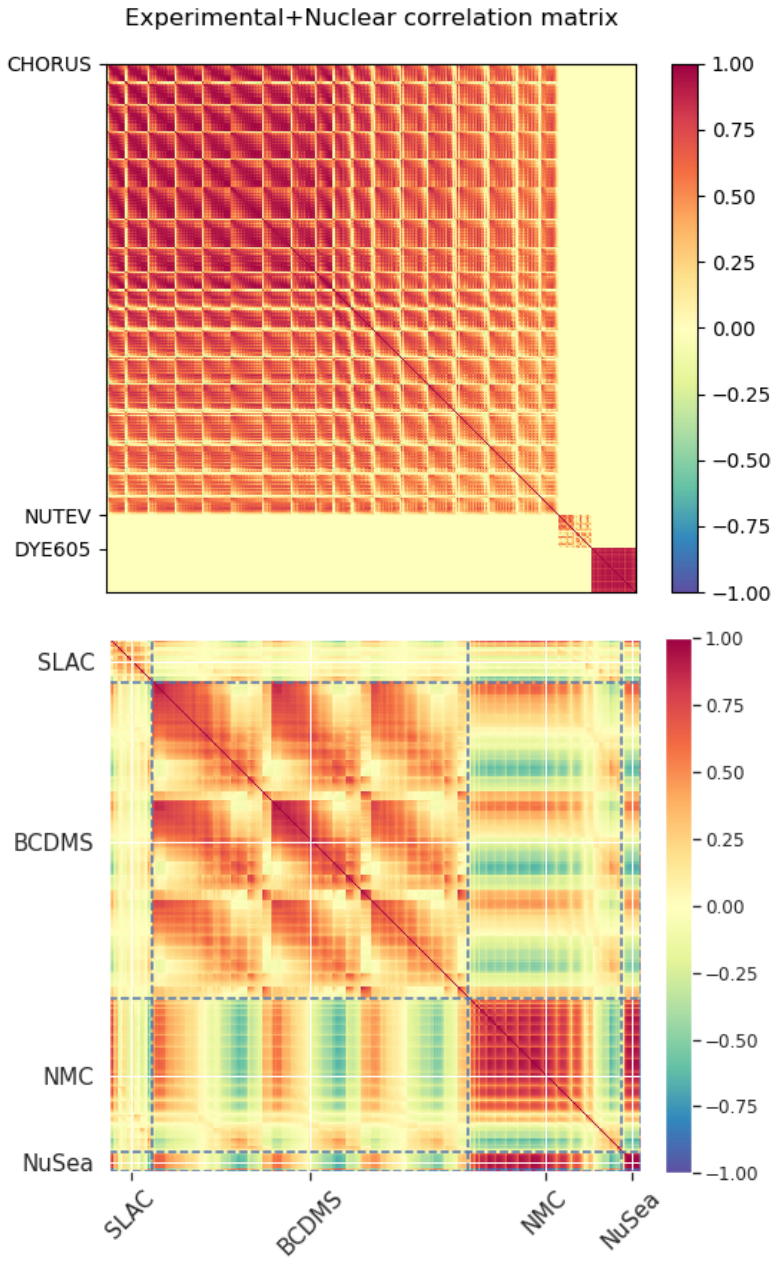
Tensions relieved: improvement in χ^2

exp only: $\chi^2/N_{\rm dat} = 1.139$ exp+th: $\chi^2/N_{\rm dat} = 1.110$

Data whose theoretical descrition is affected by large scale uncertainties
are deweighted in favour of more perturbatively stable data

[EPJ C79 (2019) 838; ibid. 931]

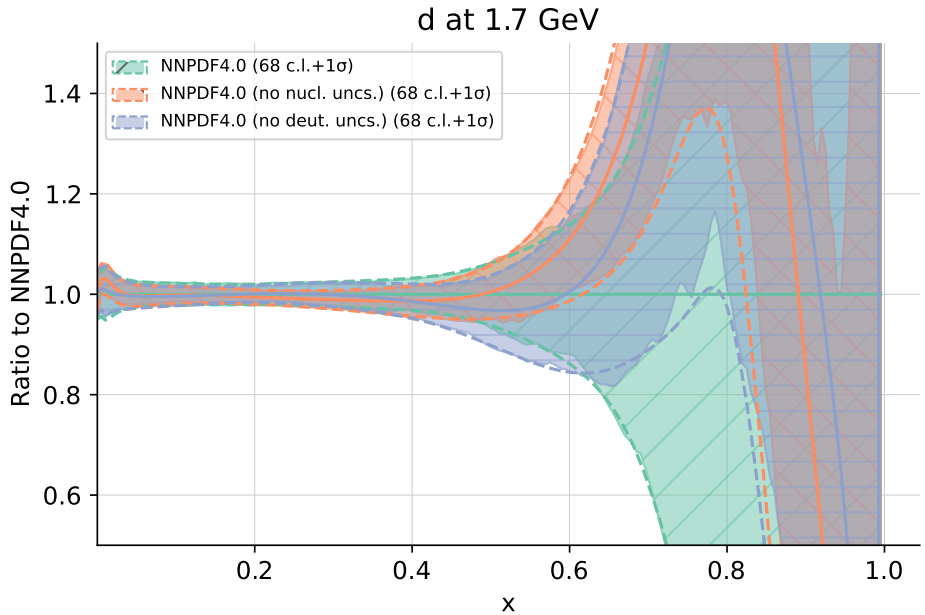
Theory uncertainty in PDF determination



Effect of nuclear uncertainties relevant at large x
to reconcile FT DIS with LHC DY data

$$\chi^2_{\rm tot} = 1.17 \to \chi^2_{\rm tot} = 1.26 \text{ (no nucl. uncs.)}$$
$$\chi^2_{\rm LHCb} = 1.54 \to \chi^2_{\rm tot} = 1.76 \text{ (no nucl. uncs.)}$$

The bulk of the effect is due to nuclear uncertainties for heavy nuclei
deuteron uncertainties have a comparatively smaller effect at inermediate values of x



[EPJ C79 (2019) 282; EPJ C81 (2021) 37]

5. To conclude

Summary and outlook

Performing a global QCD analysis is a daunting task (in psychological terms)

Performing a global QCD analysis is an inverse problem (in Bayesian terms)

There is an inherent bias-variance trade-off

In order to maximise precision and accuracy one has to minimise bias and variance

One has to take into account the experimental, methodological and theoretical inputs

Summary and outlook

Performing a global QCD analysis is a daunting task (in psychological terms)

Performing a global QCD analysis is an inverse problem (in Bayesian terms)

There is an inherent bias-variance trade-off

In order to maximise precision and accuracy one has to minimise bias and variance

One has to take into account the experimental, methodological and theoretical inputs

Thank you