Uncertainties in Global Fits: Insights from the NNPDF Framework

15th COMPASS Analysis Phase international mini-workshop (COMAP-XV)

Emanuele R. Nocera

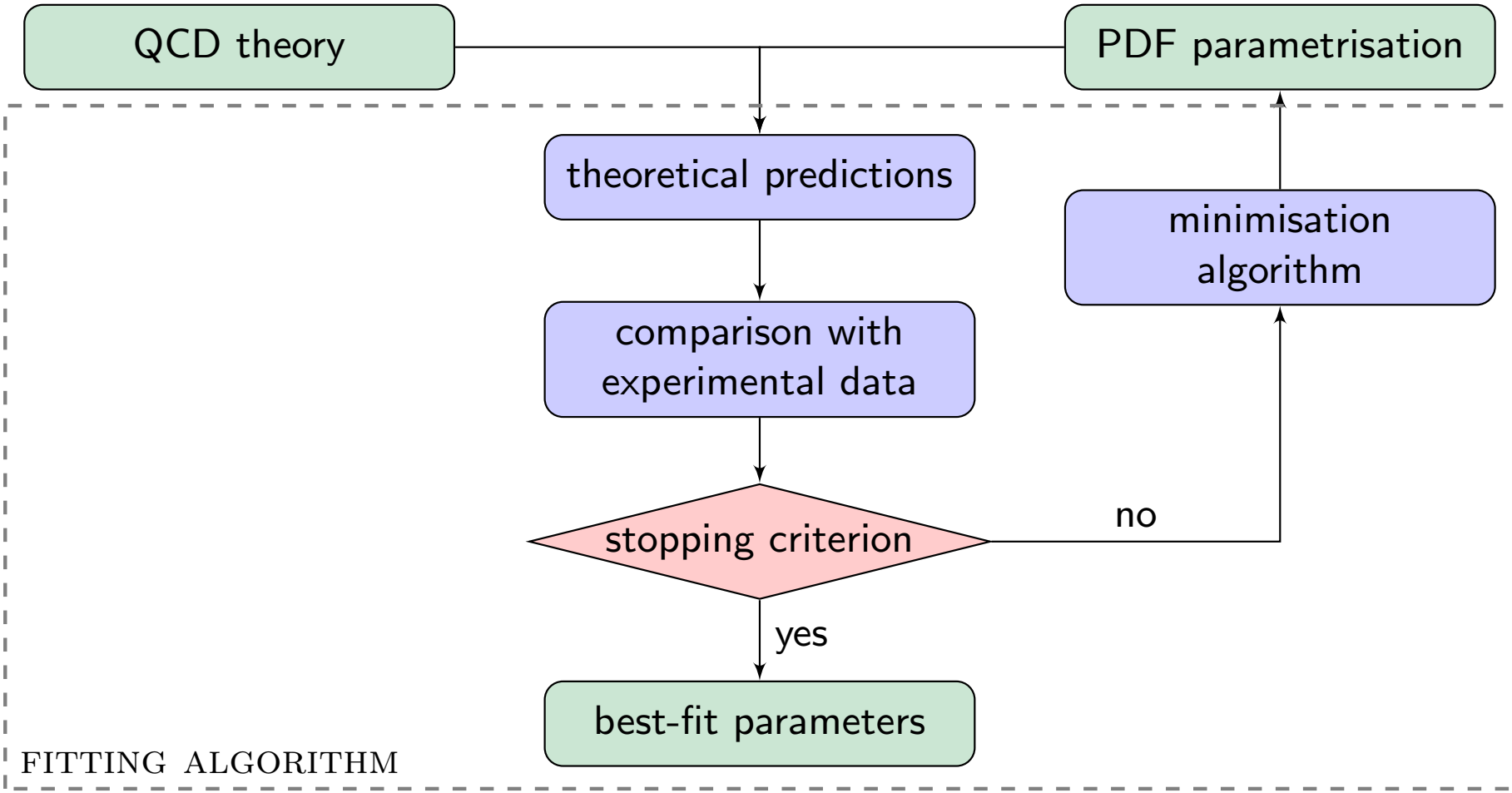
Università degli Studi di Torino and INFN, Torino

22 April 2026

UNIVERSITÀ DI TORINO

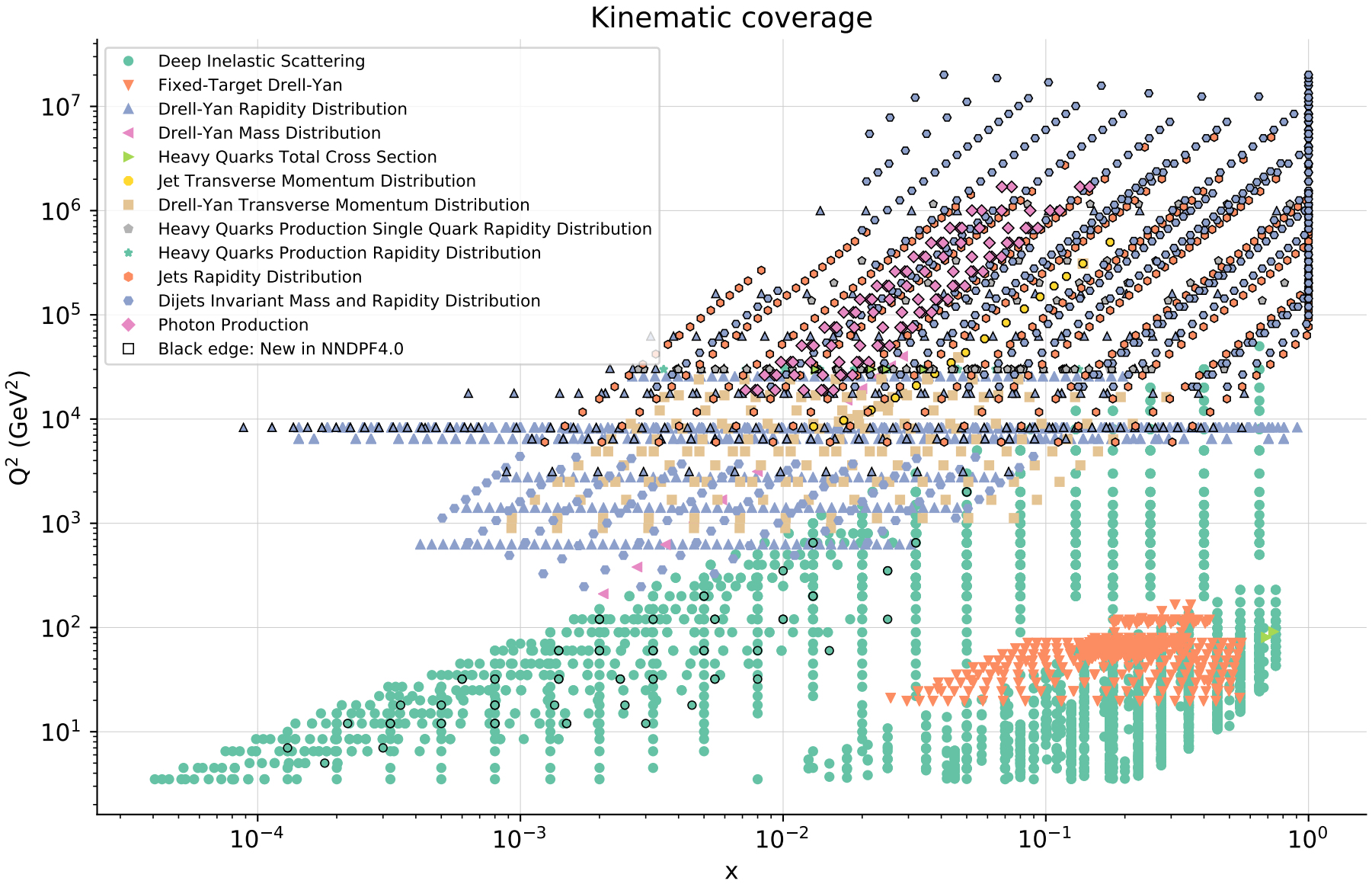
Determining PDFs from (LHC) experimental data

$$\sigma(Q^2,\tau,\mathbf{k}) = \sum_{ij}\int_\tau^1 \frac{dz}{z}\mathcal{L}_{ij}(z,Q^2)\hat{\sigma}_{ij}\left(\frac{\tau}{z},\alpha_s(Q^2),\mathbf{k}\right) \qquad \mathcal{L}_{ij}(z,Q^2) = (f_i^{h_1}\otimes f_j^{h_2})(z,Q^2)$$



$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} [T_i[\{\vec{a}\}] - D_i](\mathrm{cov}^{-1})_{ij}[T_j[\{\vec{a}\}] - D_j] \qquad \text{with } \{\vec{a}\} \text{ the set of parameters}$$

What is the typical quality of a global fit of PDFs?



$N_{\rm dat} = 4618$ $\chi^2/N_{\rm dat} \sim 1.19$ (NNLO) $1\sigma = \sqrt{2/N_{\rm dat}} \sim 0.02$

Underlying assumption: uncertainties are Gaussian

Assumption: Uncertainties are well-behaved Gaussian errors

Sometimes they are NOT

$Y = \text{'}best\ value\text{'}^{+\Delta_+}_{-\Delta_-}$ $\quad$ Δ_+ and Δ_- can be positive or negative

Possible origins of asymmetric uncertainties in LHC data:

non-parabolic χ^2 or log-likelihood curves

non-linear error propagation

systematic uncertainties (example: two-point systematic uncertainties)

Let us indicate with $\mathbf{X}$ the set of quantities that concur to construct Y, *i.e.* $Y = Y(\mathbf{X})$

with $\mathbf{X} = \{x_1, x_2, \ldots x_i, \ldots, x_n\}$

In a Bayesian framework, it can be shown that [physics/0403086]

$$E(Y) \approx Y(E[X]) + \sum_i \delta_i$$

$$\sigma^2(Y) \approx \sum_i \bar{\Delta}_i^2 + 2\sum_i \delta_i^2$$

with $\delta_i = \frac{\Delta_+ - \Delta_-}{2}$ and $\bar{\Delta} = \frac{\Delta_+ + \Delta_-}{2}$

What goes into the computation of the χ^2?

$$\chi^2 = \sum_{i,j}^{\mathrm{N_{dat}}} \left(T_i[\{\mathbf{a}\}] - D_i\right) \left(\mathrm{cov}^{-1}\right)_{ij} \left(T_j[\{\mathbf{a}\}] - D_j\right) \qquad \mathrm{cov} = \mathrm{cov_{exp}} + \mathrm{cov_{th}}$$

- experimental covariance matrix

$$\left(\mathrm{cov_{exp}}\right)_{ij} = \delta_{ij} s_i^2 + \left(\sum_{\alpha}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} + \sum_{\alpha}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \right) D_i D_j$$

  s_i are N_{dat} uncorrelated uncertainties (statistical + uncorrealted systematic uncertainties)
  $\sigma_{i,\alpha}^{(c)}$ are $N_{\mathrm{dat}} \times N_c$ additive correlated uncertainties
  $\sigma_{i,\alpha}^{(\mathcal{L})}$ are $N_{\mathrm{dat}} \times N_{\mathcal{L}}$ multiplicative uncertainties

- theory covariance matrix

$$\left(\mathrm{cov_{th}}\right)_{ij} = \frac{1}{N} \sum_{k}^{N} \Delta_i^{(k)} \Delta_j^{(k)}$$

$\Delta_{i,j}^{(k)}$ are nuisance parameters *e.g.* $\Delta_i^{(k)} = T_i^{7pt}(\mu_R^{(k)}, \mu_F^{(k)} - T_i(\mu_R^0, \mu_F^0)$

Treat conveniently uncorrelated/correlated and additive/multiplicative uncertainties

1 Correlated uncertainties

Experimental correlations: what are they?

1. What is correlated with what?
   Correlations between data points in a data set
   Easy (clear). Identify the various sources (~ 300) of uncertainty.
   Between data sets in the same experiment
   Medium (usually clear). Put in correspondence uncertainties with the same name.
   Between different experiments
   Difficult (typically obscure). Usually not clear how to match uncertainties.
2. How much are uncertainties correlated? Assumption: 100%.
   Sometimes this is NOT realistic. There exist decorrelation models.
3. Do experimentalists release complete information to properly treat correlations?
   Information on correlation/decorrelation provided years after publication.

Systematic uncertainty	8 TeV W + jets	8 TeV Z + jets	8 TeV $t\bar{t}$ lepton + jets	13 TeV $t\bar{t}$ lepton + jets	8 TeV inclusive jets
Jet flavour response	JetScaleFlav2	Flavor Response	flavres-jes	JET29NP JET Flavour Response	syst JES Flavour Response*
Jet flavour composition	JetScaleFlav1Known	Flavor Comp	flavcomp-jes	JET29NP JET Flavour Composition	syst JES Flavour Comp
Jet punchthrough	JetScalepunchT	Punch Through	punch-jes	-	syst JES PunchThrough MC15
Jet scale	JetScalePileup2	PU OffsetMu	pileoffmu-jes	-	syst JES Pileup MuOffset
	-	PU Rho	pileoffrho-jes	JET29NP JET Pileup RhoTopology	syst JES Pileup Rho topology*
	JetScalePileup1	PU OffsetNPV	pileoffnpv-jes	JET29NP JET Pileup OffsetNPV	syst JES Pileup NPVOffset
	-	PU PtTerm	pileoffpt-jes	JET29NP JET Pileup PtTerm	syst JES Pileup Pt term
Jet JVF selection	JetJVFcut	JVF	jetvxfrac	-	syst JES Zjets JVF
B-tagged jet scale	-	btag-jes	JET29NP JET BJES Response	-	-
Jet resolution	-	jeten-res	JET JER SINGLE NP	-	-
Muon scale	-	-	mup-scale	MUON SCALE	-
Muon resolution	-	-	muonms-res	MUON MS	-
Muon identification	-	-	muid-res	MUON ID	-
Diboson cross section	-	-	dibos-xsec	Diboson xsec	-
Z + jets cross section	-	-	zjet-xsec	Zjets xsec	-
Single-t cross section	-	-	singletop-xsec	st xsec	-

[EPJ C82 (2022) 438]

Experimental correlations and data inconsistency

Single inclusive jet data from ATLAS 7 TeV

default correlations: terrible χ^2

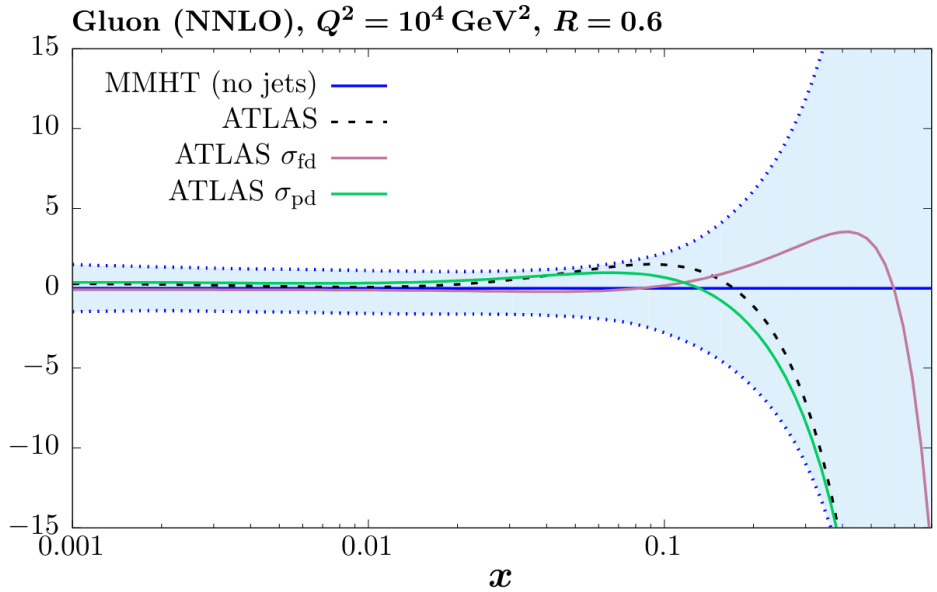
(correlations across rapidity bins)

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

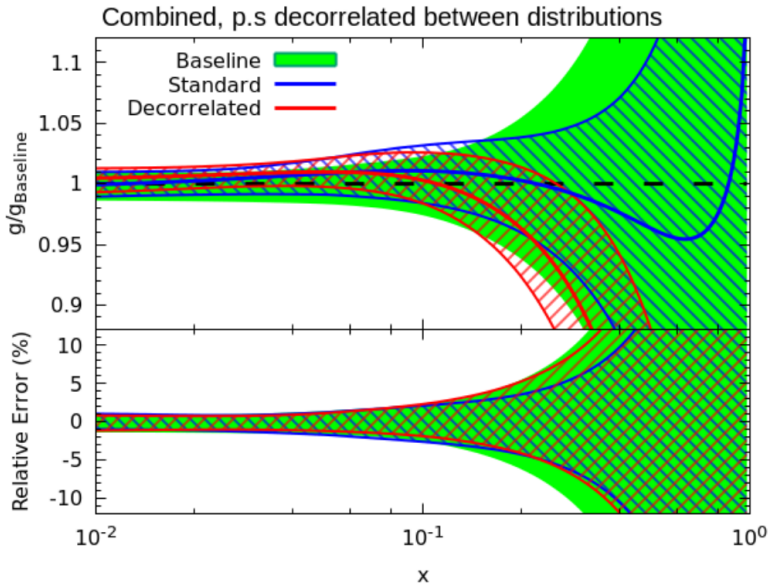
(correlations across different spectra)

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]

Experimental correlations and data inconsistency

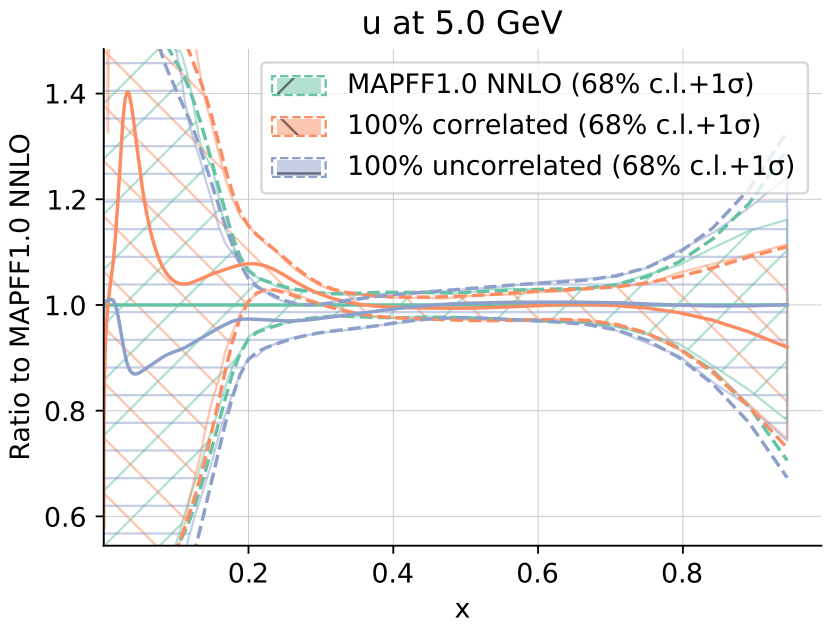
Consider the COMPASS $\pi^{\pm}$ multiplicities

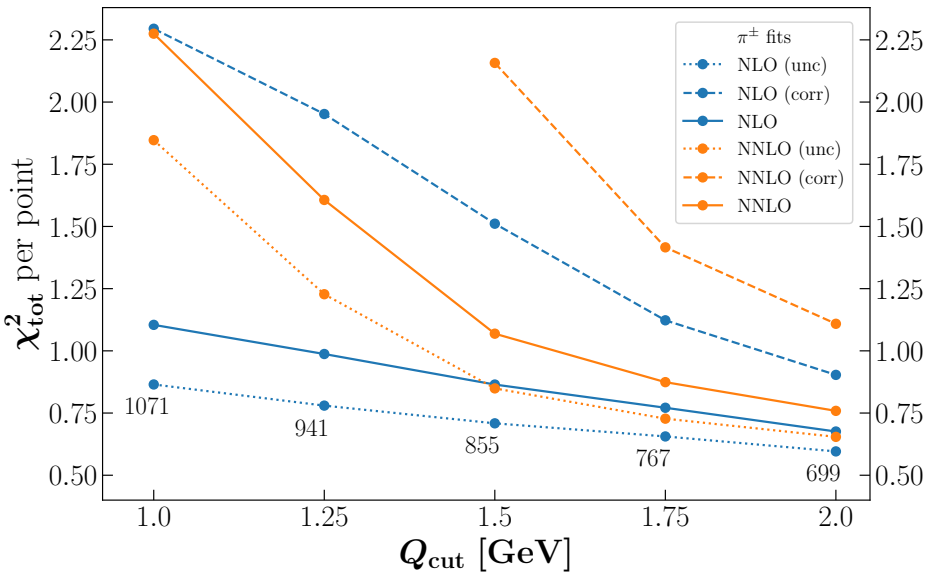
[PLB 764 (2017) 1]

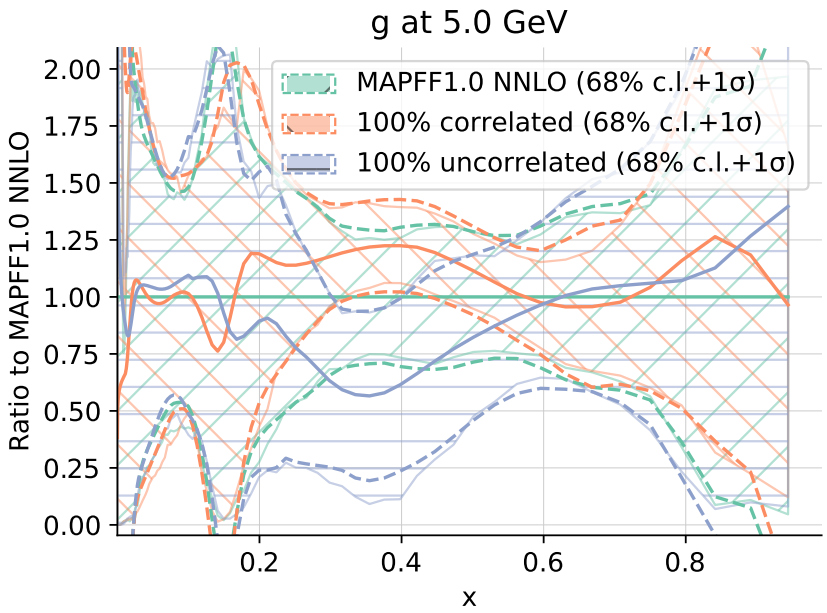
Only 80% of the systematic uncertainty is bin-by-bin correlated

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

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







Good knowledge of experimental correlations is important

Let us call A the $N_{\rm dat} \times N_{\rm err}$ matrix of uncertainties, such that cov$=AA^t$

If the theory is known, fixed and correct:

$$\langle \chi^2_{\rm true} \rangle = \|A^+ A\|_F = N_{\rm dat}$$

If we know $\bar{A}$ instead of A:

$$\langle \bar{\chi}^2 \rangle = \|\bar{A}^+ A\|_F$$

The χ^2 is *stable* if:

$$\langle \bar{\chi}^2 \rangle - \langle \chi^2 \rangle = \|\bar{A}^+ A\|_F - N_{\rm dat} < \sqrt{2N_{\rm dat}}$$

If not, define $A_{\rm reg}$ by clipping the singular values of the correlated part of $\bar{A}$ to δ, whenever these are smaller than δ; the rest of the singular vectors are left unchanged

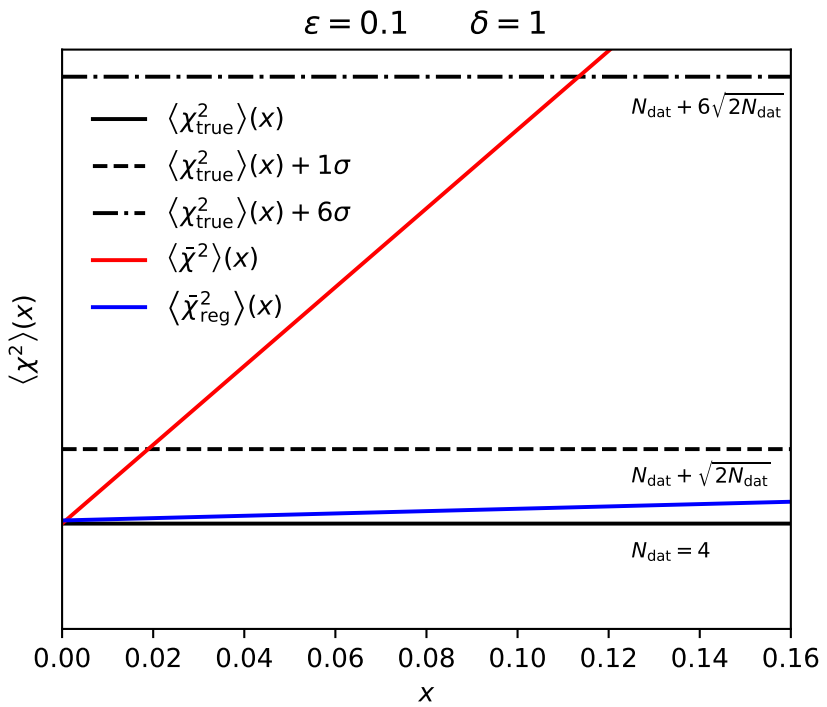
$$\langle \chi^2_{\rm reg} \rangle = \|A^+_{\rm reg} A\|_F$$

Assumptions:

correlations are determined less precisely than variances and inaccuracy is limited to a small number of uncertainties

$$A(x) = \begin{pmatrix} \epsilon & 0 & 0 & 0 & 1 & 0 \\ 0 & \epsilon & 0 & 0 & 1 & 0 \\ 0 & 0 & \epsilon & 0 & 1 & 0 \\ 0 & 0 & 0 & \epsilon & 1-x & \sqrt{1-(1-x)^2} \end{pmatrix}$$

$$\bar{A} = \begin{pmatrix} \epsilon & 0 & 0 & 0 & 1 & 0 \\ 0 & \epsilon & 0 & 0 & 1 & 0 \\ 0 & 0 & \epsilon & 0 & 1 & 0 \\ 0 & 0 & 0 & \epsilon & 1 & 0 \end{pmatrix}$$

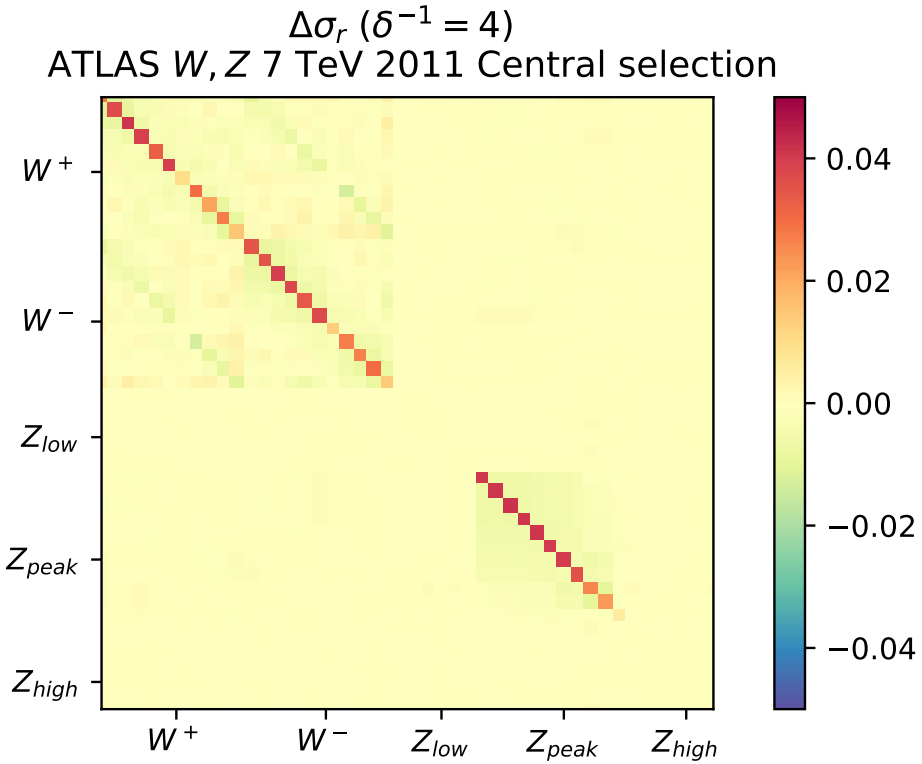


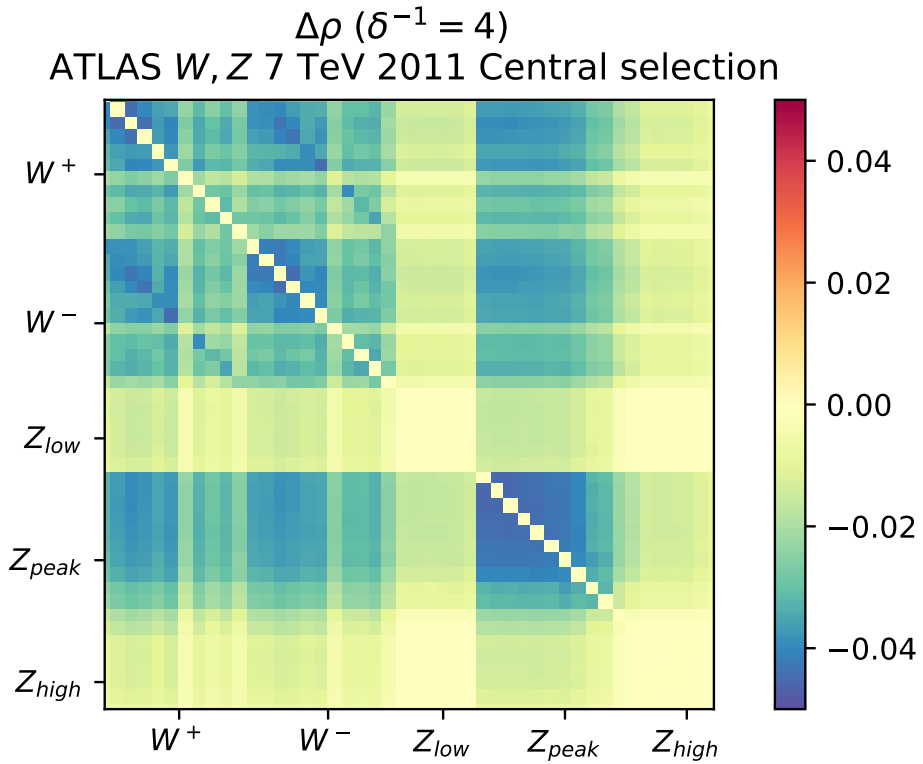
[EPJ C82 (2022) 956]

Regularising the NNPDF4.0 data set [EPJ C82 (2022) 956]

Let us test the regularisation procedure on the NNPDF4.0 data set

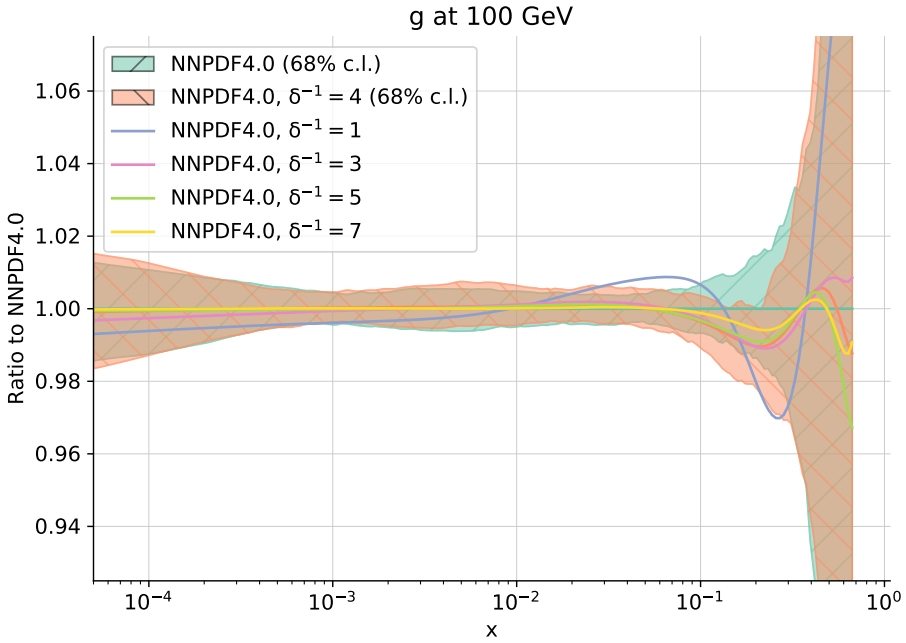
As an example, let us focus on a specific data set:
ATLAS W, Z 7 TeV 2011 central selection [EPJ C77 (2017) 367]

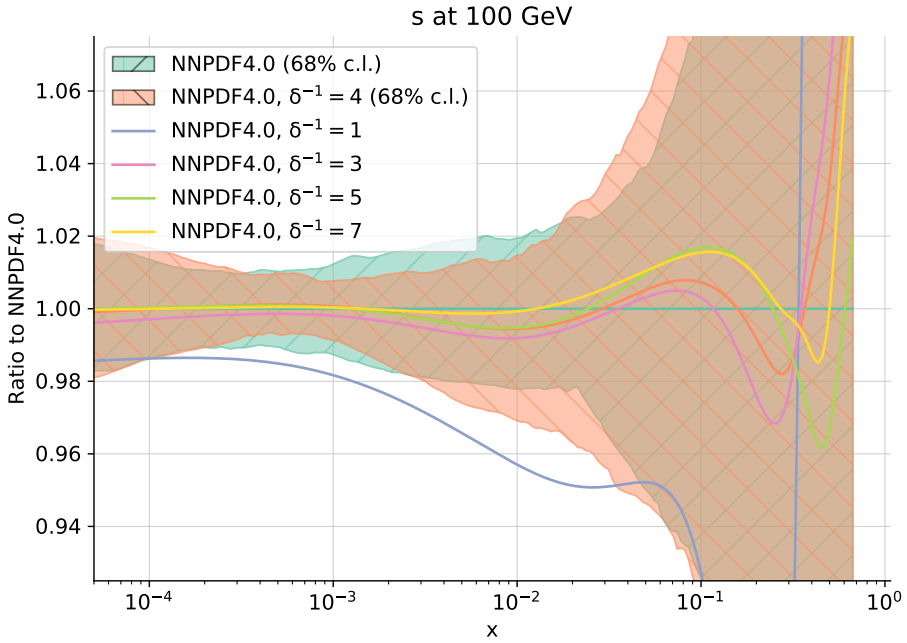




Data set	Z	$\delta^{-1}=1$ $\Delta\sigma_r$	$\lvert\Delta\rho\rvert$	$\delta^{-1}=2$ $\Delta\sigma_r$	$\lvert\Delta\rho\rvert$	$\delta^{-1}=3$ $\Delta\sigma_r$	$\lvert\Delta\rho\rvert$	$\delta^{-1}=4$ $\Delta\sigma_r$	$\lvert\Delta\rho\rvert$	$\delta^{-1}=5$ $\Delta\sigma_r$	$\lvert\Delta\rho\rvert$	$\delta^{-1}=7$ $\Delta\sigma_r$	$\lvert\Delta\rho\rvert$
ATLAS W, Z 7 TeV	9.0	94.4	0.50	21.9	0.19	8.63	0.09	4.15	0.05	2.12	0.02	0.50	0.01

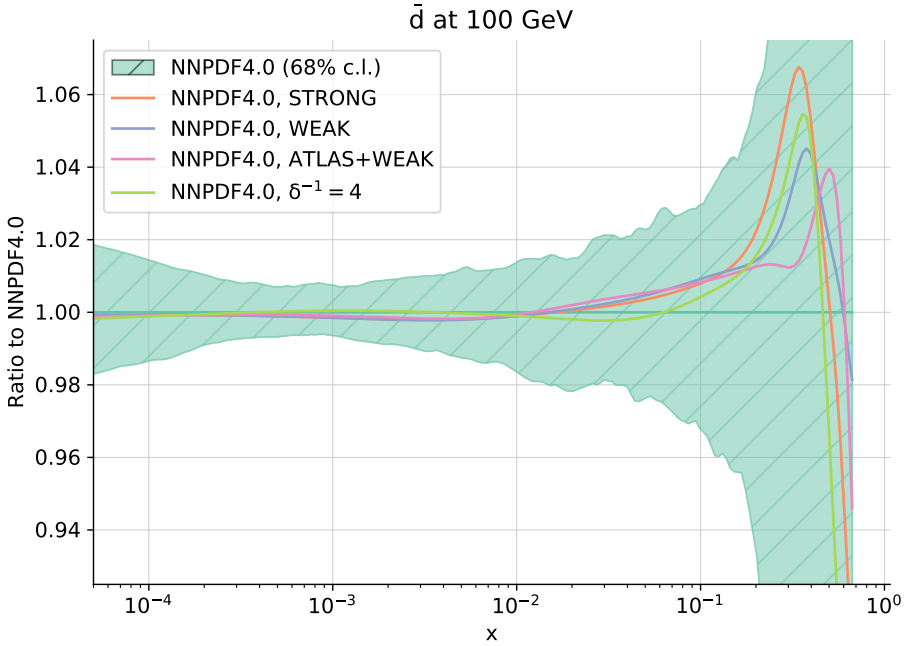
Regularising the NNPDF4.0 data set [EPJ C82 (2022) 956]

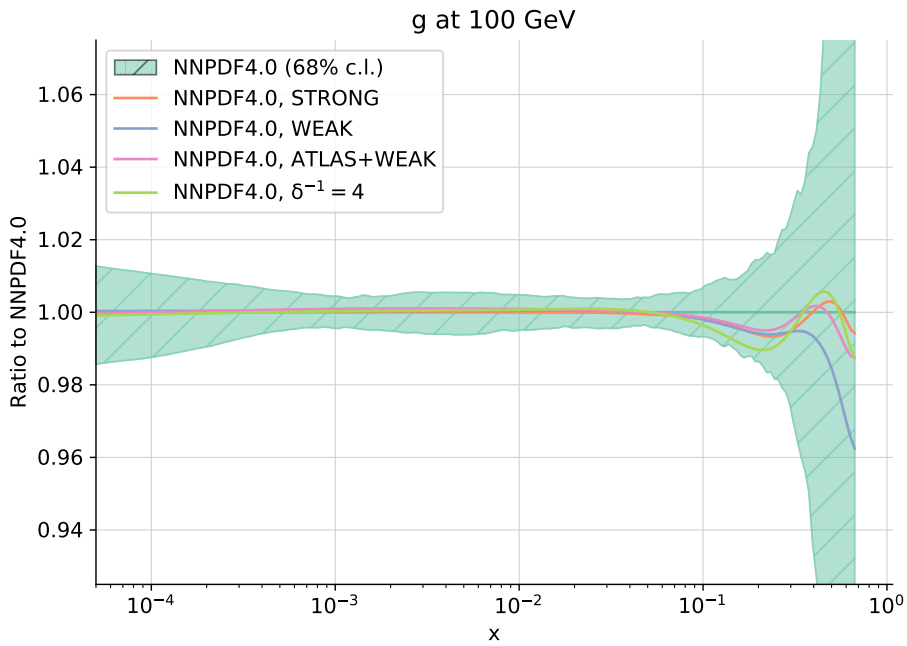




		$\chi^2/N_{\rm dat}$						
Data set	$N_{\rm dat}$	NNPDF4.0	$\delta^{-1}=1$	$\delta^{-1}=2$	$\delta^{-1}=3$	$\delta^{-1}=4$	$\delta^{-1}=5$	$\delta^{-1}=7$
Deep-inelastic scattering	3089	1.12	0.64	1.02	1.09	1.11	1.12	1.12
Fixed-target Drell-Yan	195	0.98	0.48	0.90	0.96	0.97	0.97	0.99
Tevatron Drell-Yan	65	1.11	0.48	0.71	0.85	0.93	1.02	1.10
ATLAS total	679	1.24	0.50	0.84	0.97	1.04	1.10	1.19
W, Z 7 TeV CC	46	1.92	0.31	0.74	0.94	1.21	1.47	1.76
CMS total	474	1.31	0.39	0.83	1.08	1.21	1.26	1.28
LHCb total	116	1.55	0.73	1.41	1.53	1.56	1.55	1.55
Total	4618	1.16	0.58	0.97	1.07	1.11	1.13	1.15

Regularising the NNPDF4.0 data set [EPJ C82 (2022) 956]





		$\chi^2/N_{\rm dat}$					
Data set	$N_{\rm dat}$	NNPDF4.0	STRONG	WEAK	ATLAS	ATLAS+WEAK	$\delta^{-1} = 4$
Deep-inelastic scattering	3089	1.12	1.12	1.12	1.12	1.12	1.11
Fixed-target Drell-Yan	195	0.98	1.00	0.99	0.99	0.99	0.97
Tevatron Drell-Yan	65	1.11	1.10	1.09	1.09	1.10	0.93
ATLAS total	679	1.24	1.24	1.24	1.23	1.24	1.04
CMS total	474	1.31	1.31	1.31	1.31	1.30	1.21
LHCb total	116	1.55	1.56	1.54	1.55	1.55	1.56
Total	4618	1.16	1.16	1.16	1.16	1.16	1.11

Inconsistent closure tests

Generate pseudodata with statistical and systematic uncertainties

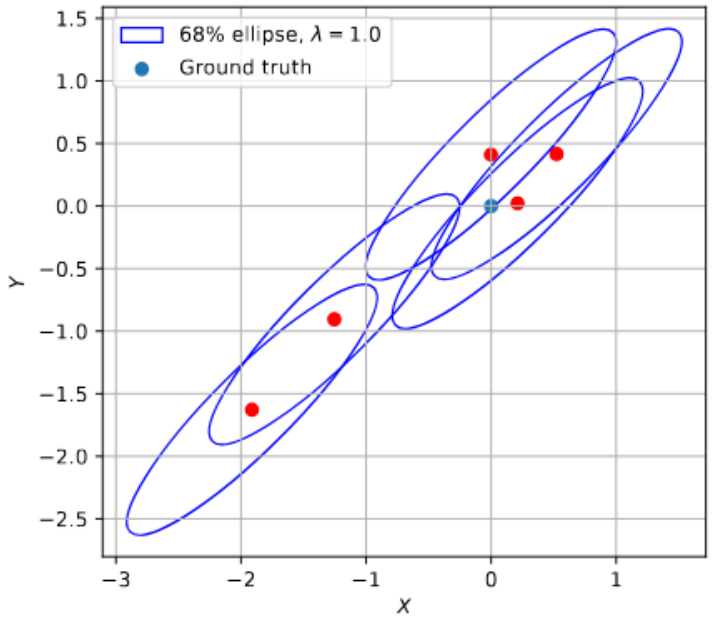
$$C = C^{\text{stat}} + C^{\text{syst}} \qquad C_{ij}^{\text{syst}} = \sum_k \Delta_i^k \Delta_j^k \qquad \Delta_i^k \text{ is the k-th sys. unc. for the i-th point}$$

Assume systematic uncertainties are underestimated and perform a fit with

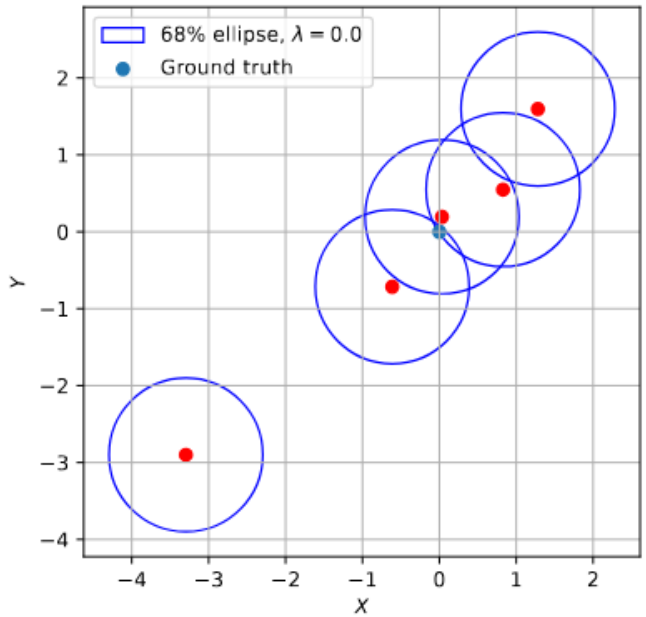
$$\Delta_i^k \to \lambda \Delta_i^k \qquad \lambda = 1 \text{ consistency} \qquad \lambda = 0 \text{ extreme inconsistency}$$

PREDICTED UNCERTAINTY ON GENERATED DATA

CONSISTENT



EXTREME INCONSISTENCY



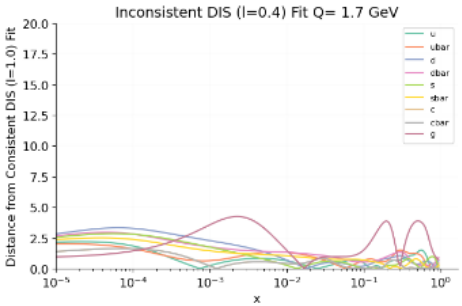
[arXiv:2503.17447]

Inconsistent closure tests

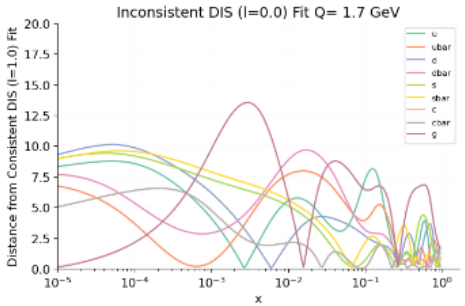
DISTANCES: $\frac{f^{\text{consistent}} - f^{\text{inconsistent}}}{\sigma^{\text{PDF}}/\sqrt{N_{\text{rep}}}}$; $d \sim 1 \Rightarrow$ STATISTICAL EQUIVALENCE

DIS: BULK INCONSISTENCY

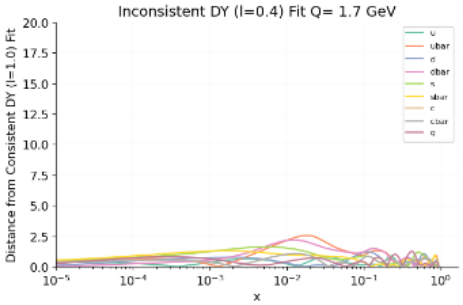
$\lambda = 0.4$: MODEL CORRECTS

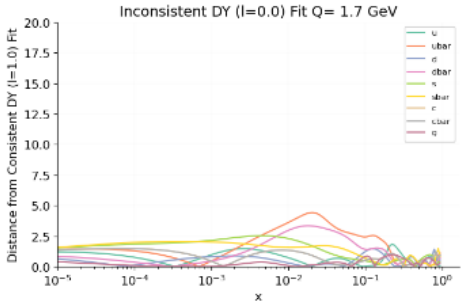


$\lambda = 0$: MODEL FAILS

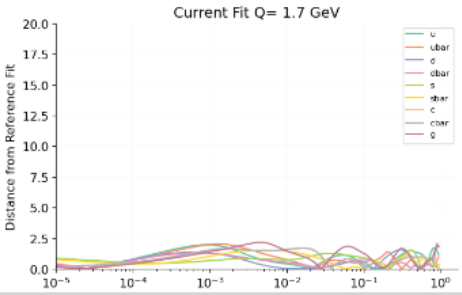


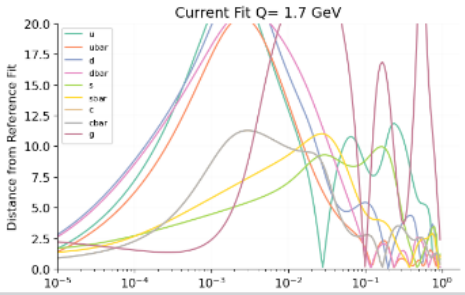
DY: SINGLE DATASET INCONSISTENCY





JETS: HIGH-IMPACT INCONSISTENCY





The ML model corrects for inconsistency except in extreme cases [arXiv:2503.17447]

2 Multiplicative uncertainties

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 normalisation 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

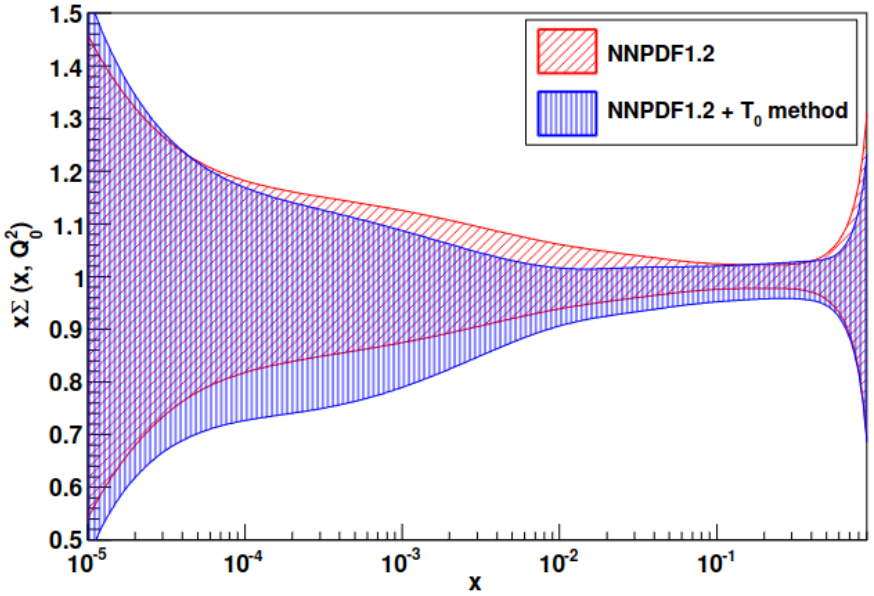
$$(\mathrm{cov})_{ij} \to (\mathrm{cov}_{t_0})_{ij} \Longleftrightarrow (s_i^2 + \sigma^2 d_i^2)\delta_{ij} \to s_i^2\delta_{ij} + \sigma^2 t_0^2$$

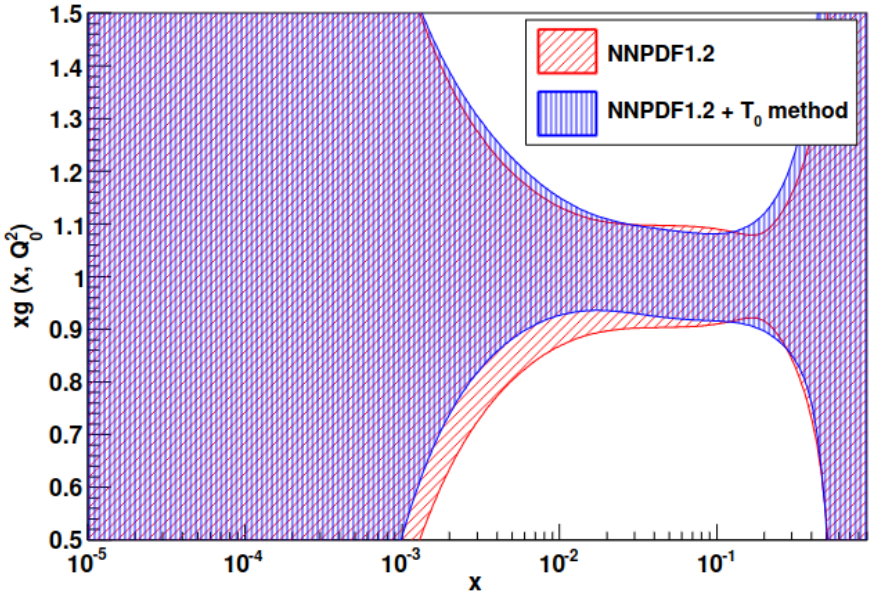
$$\left(\mathrm{cov}_{t_0}^{-1}\right)_{ij} = \frac{\delta_{ij}}{s_i^2} - \frac{\sigma^2 t_0^2}{s_i^2 s_j^2}\frac{\Sigma^2}{\Sigma^2 + \sigma^2 t_0^2} \Longleftrightarrow t_0 = w \qquad v_t = \Sigma^2 + \sigma^2 w^2$$

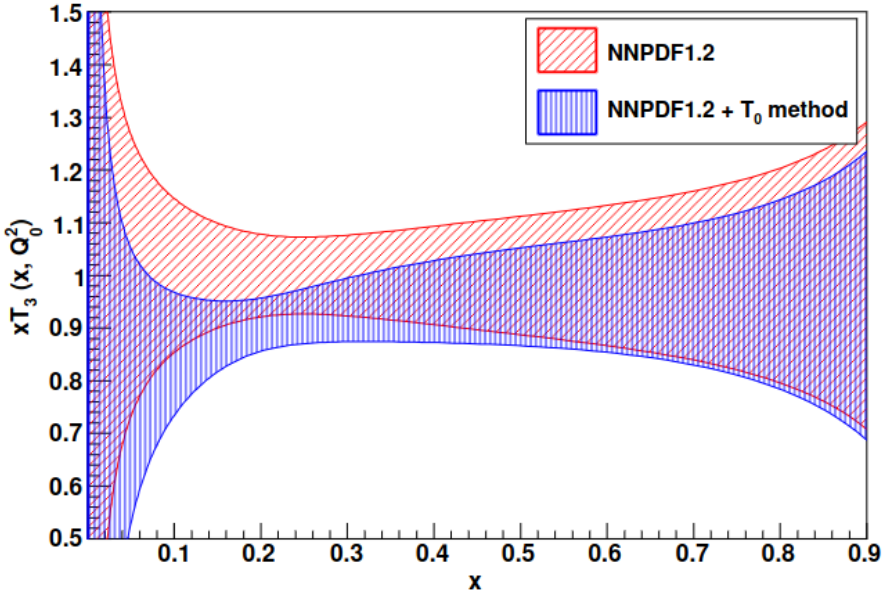
$\longrightarrow$ recover the unbiased estimators for t_0 and v_t, provided that w is tuned to t_0
$\longrightarrow$ w can be tuned to t_0 via an iterative procedure

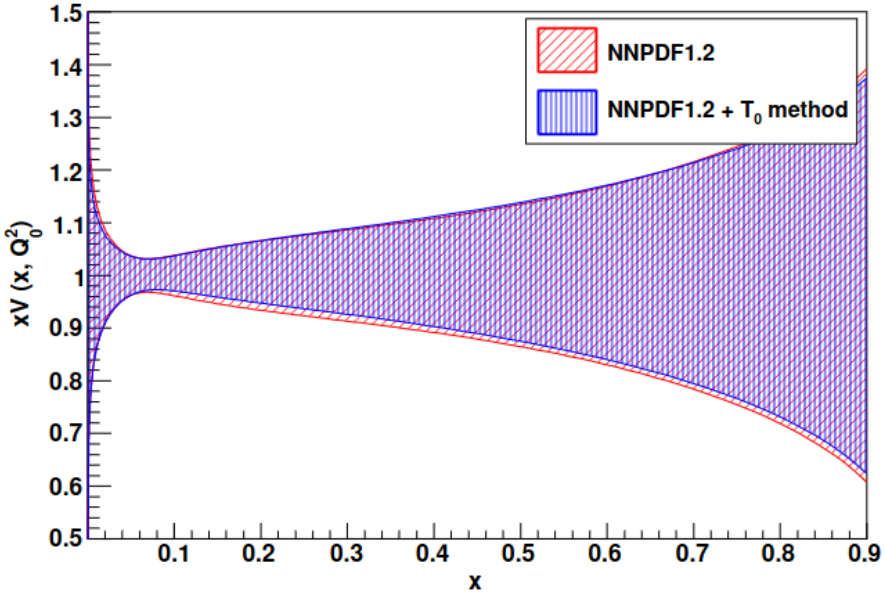
The d'Agostini bias and its solution can be generalised to more than one experiment with different normalisation errors (per experiment/per data point)

Normalisation uncertainties: impact on PDFs [JHEP 1005 (2010) 075]

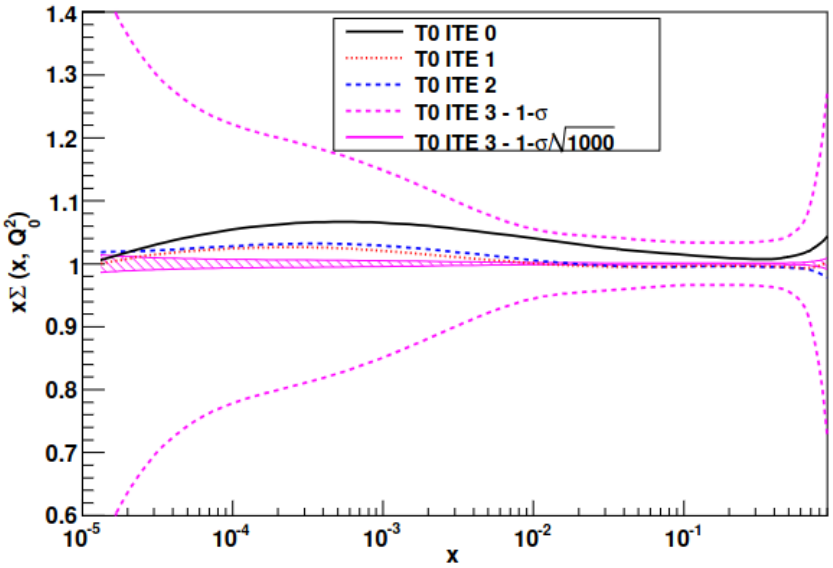


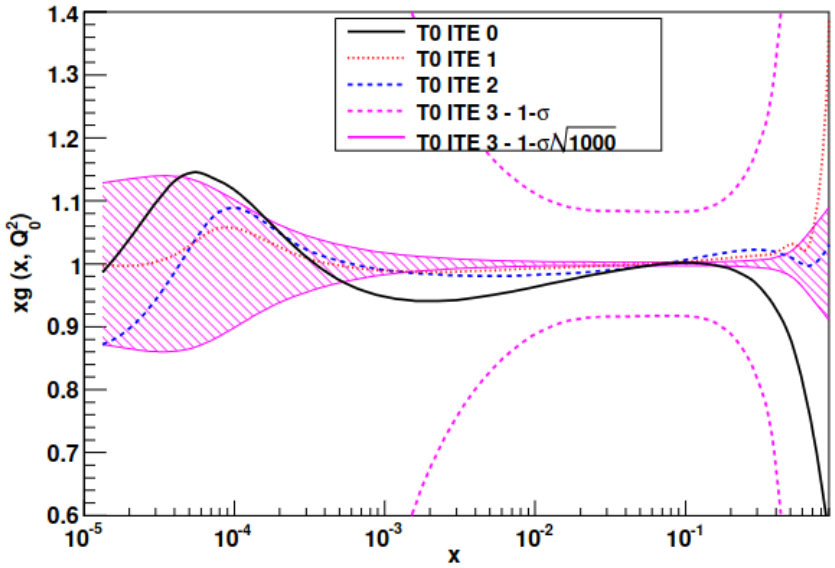


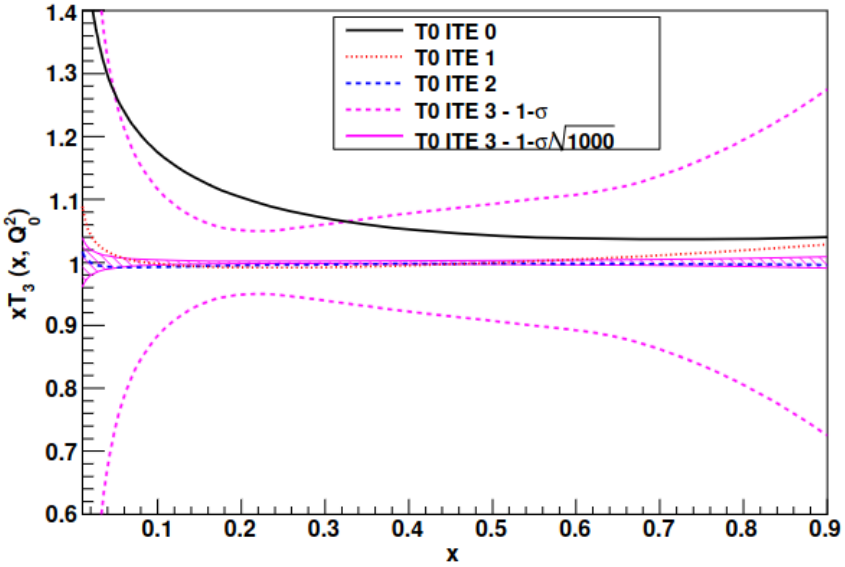


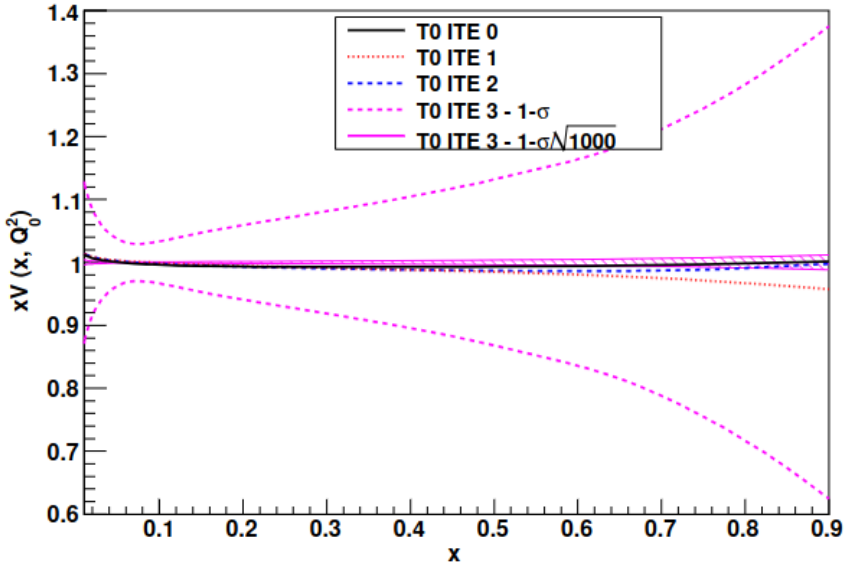


Does the t_0 procedure converge? [JHEP 1005 (2010) 075]









The t_0 prescription and SM parameters

Assume true underlying PDFs and $\alpha_s(M_Z) = \alpha_s^0 = 0.118$

Generate data distributed according to the experimental covariance matrix

Determine PDFs and α_s simultaneously from a fit to the data

Repeat the exercise N_r times

$$\langle \alpha_s \rangle = \frac{\sum_{j=1}^{N_r} \frac{\alpha_s^{(j)}}{\left(\sigma_\alpha^{(j)}\right)^2}}{\sum_{j=1}^{N_r} \frac{1}{\left(\sigma_\alpha^{(j)}\right)^2}} \qquad \langle \sigma_\alpha \rangle = \frac{1}{\sqrt{\sum_{j=1}^{N_r} \frac{1}{\left(\sigma_\alpha^{(j)}\right)^2}}} \qquad R_{bv} = \sqrt{\frac{1}{N_r} \sum_j^{N_r} \left(\frac{\alpha_s^j - \alpha_s^0}{\alpha_s^j} \right)^2}$$

weighted mean | weighted uncertainty | bias/variance ratio

New problem: theory predictions depend on α_s. Should we vary T_0 with α_s or not?

Vary T_0 with α_s: 25 runs of the universe

$\langle \alpha_s \rangle = 0.119450 \quad \langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000077 \quad R_{bv} = 3.80 \pm 0.16$

Keep T_0 fixed: 25 runs of the universe

$\langle \alpha_s \rangle = 0.118152 \quad \langle \sigma_\alpha \rangle / \sqrt{N_r} = 0.000070 \quad R_{bv} = 0.97 \pm 0.11$

[EPJ C85 (2025) 1001]

3 Summary

Conclusions

Uncertainty treatment is a central part of modern global fits.

Correlated systematics can affect fit quality and sometimes the PDFs themselves.

Regularisation helps when correlations are imperfectly known.

The t_0 method avoids bias from multiplicative uncertainties.

Joint fits with SM parameters require special care.

Reliable uncertainty treatment is essential for precision phenomenology.

Conclusions

Uncertainty treatment is a central part of modern global fits.

Correlated systematics can affect fit quality and sometimes the PDFs themselves.

Regularisation helps when correlations are imperfectly known.

The t_0 method avoids bias from multiplicative uncertainties.

Joint fits with SM parameters require special care.

Reliable uncertainty treatment is essential for precision phenomenology.

Thank you