Deep-Inelastic Scattering and Collinear Physics

Second European School on the Physics of the EIC and Related Topics

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

23-24 June 2025

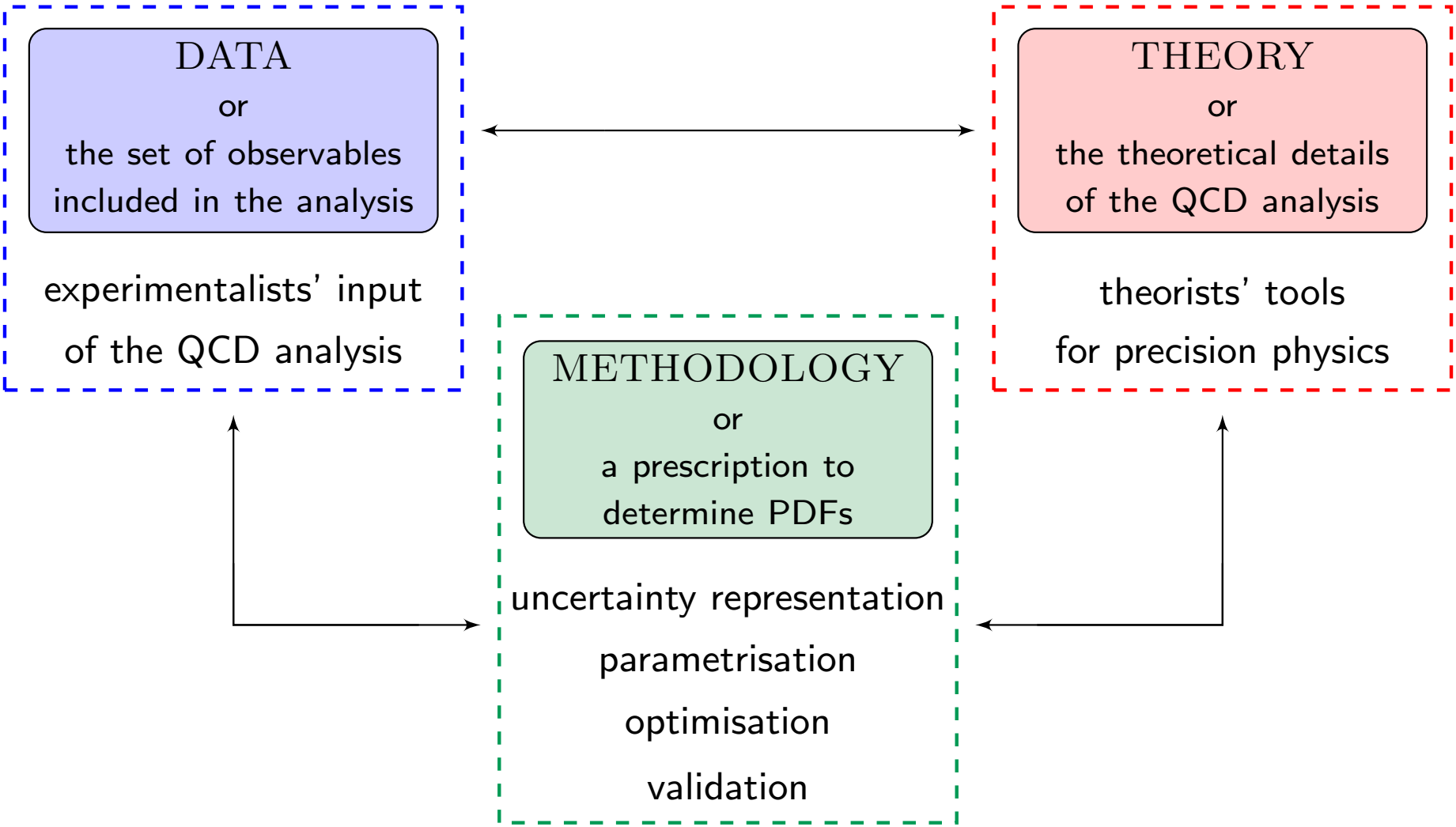
UNIVERSITÀ DI TORINO

Summary of Lecture 1

1. Deep Inelastic Scattering has been, is, and will be a crucial laboratory of QCD
   $\longrightarrow$ hadronic structure encoded in unpolarised and polarised PDFs
   $\longrightarrow$ PDFs are related to physical observales via factorisation and evolution
   $\longrightarrow$ (critically different) qualitative PDF features are driven by this theoretical framework
2. PDFs are a limiting factor for precision and discovery
   $\longrightarrow$ unpolarised PDFs: SM and BSM physics at the LHC
   $\longrightarrow$ polarised PDFs: contribution of partons' spin to the proton spin
3. PDFs are determined from experimental data by means of parametric regression
   $\longrightarrow$ need to define data, theory, and methodology
4. Different physical observables constrain different PDF combinations
   $\longrightarrow$ fixed-target NC DIS: u and d
   $\longrightarrow$ fixed-target CC DIS: s and $\bar{s}$
   $\longrightarrow$ HERA NC and CC DIS: u, $\bar{u}$, d, $\bar{d}$, g (scaling violations and tagged DIS)
   $\longrightarrow$ fixed-target DY: u and d at large x
   $\longrightarrow$ collider DY: u, $\bar{u}$, d, $\bar{d}$, s
   $\longrightarrow$ collider DY+c: s (W) and c (Z)
   $\longrightarrow$ Zp_T, $t\bar{t}$, jets: g
   $\longrightarrow$ only a small fraction of the above is available for polarised PDFs

Lecture 2: Data, theoretical, and methodological accuracy in PDF determination

The ingredients of PDF determination



Each of these ingredients is a source of uncertainty in the PDF determination

Each of these ingredients require to make choices which lead to different PDF sets

Overview of current PDF determinations

	NNPDF4.0	MSHT20	CT18	HERAPDF2.0	CJ22	ABMP16
Fixed-target DIS	☑	☑	☑	☒	☑	☑
JLAB	☒	☒	☒	☒	☑	☒
HERA I+II	☑	☑	☑	☑	☑	☑
HERA jets	☑	☒	☒	☑	☒	☒
Fixed target DY	☑	☑	☑	☒	☑	☑
Tevatron W, Z	☑	☑	☑	☒	☑	☑
LHC vector boson	☑	☑	☑	☒	☑	☑
LHC $W + c$ $Z + c$	☑	☒	☒	☒	☒	☒
Tevatron jets	☑	☑	☑	☒	☑	☒
LHC jets	☑	☑	☑	☒	☒	☒
LHC top	☑	☑	☒	☒	☒	☑
LHC single t	☑	☒	☒	☒	☒	☒
LHC prompt γ	☑	☒	☒	☒	☒	☒
statistical treatment	Monte Carlo	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2 = 1$	Hessian $\Delta\chi^2 = 1.645$	Hessian $\Delta\chi^2 = 1$
parametrisation	Neural Network	Chebyschev pol.	Bernstein pol.	polynomial	polynomial	polynomial
HQ scheme	FONLL	TR′	ACOT-χ	TR′	ACOT-χ	FFN
accuracy	aN^{3}LO	aN^{3}LO	NNLO	NNLO	NLO	NNLO
latest update	EPJ C82 (2022) 428	EPJ C81 (2021) 341	PRD 103 (2021) 014013	EPJ C82 (2022) 243	PRD 107 (2023) 113005	PRD 96 (2017) 014011

All PDF sets are available as (x, Q^2) interpolation grids through the LHAPDF library

Parton Distribution Functions

Lecture 2: Data, Theoretical and Methodological Accuracy in PDF Determination

Outline

2.1 Which data constrain which PDFs?
- DIS observables
- non-DIS observables

2.2 Can we improve the fit quality by improving the theory?
- heavy quarks
- missing higher order uncertainties

2.3 Why is the methodology important?
- parametrisation
- optimisation
- uncertainty representation
- validation of uncertainties
- PDF benchmarks

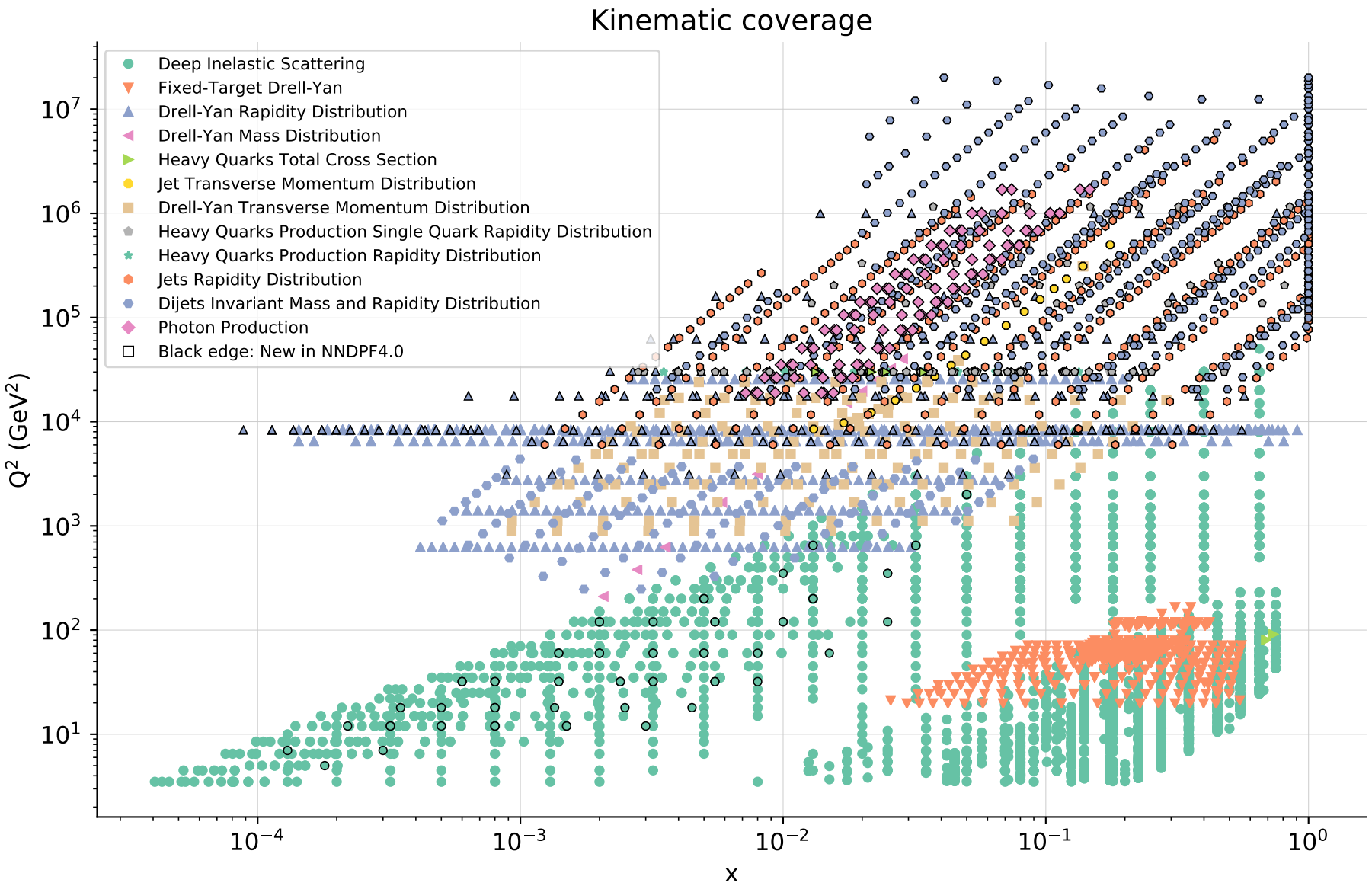
I will focus on a limited selection of recent results

I will not talk about some very interesting topics (*e.g.* aN^{3}LO PDFs, interplay between fitting PDFs and New Physics, non parametric regression models, . . .)

See also lectures by M. Ubiali and the ML hands-on session by P. Zurita

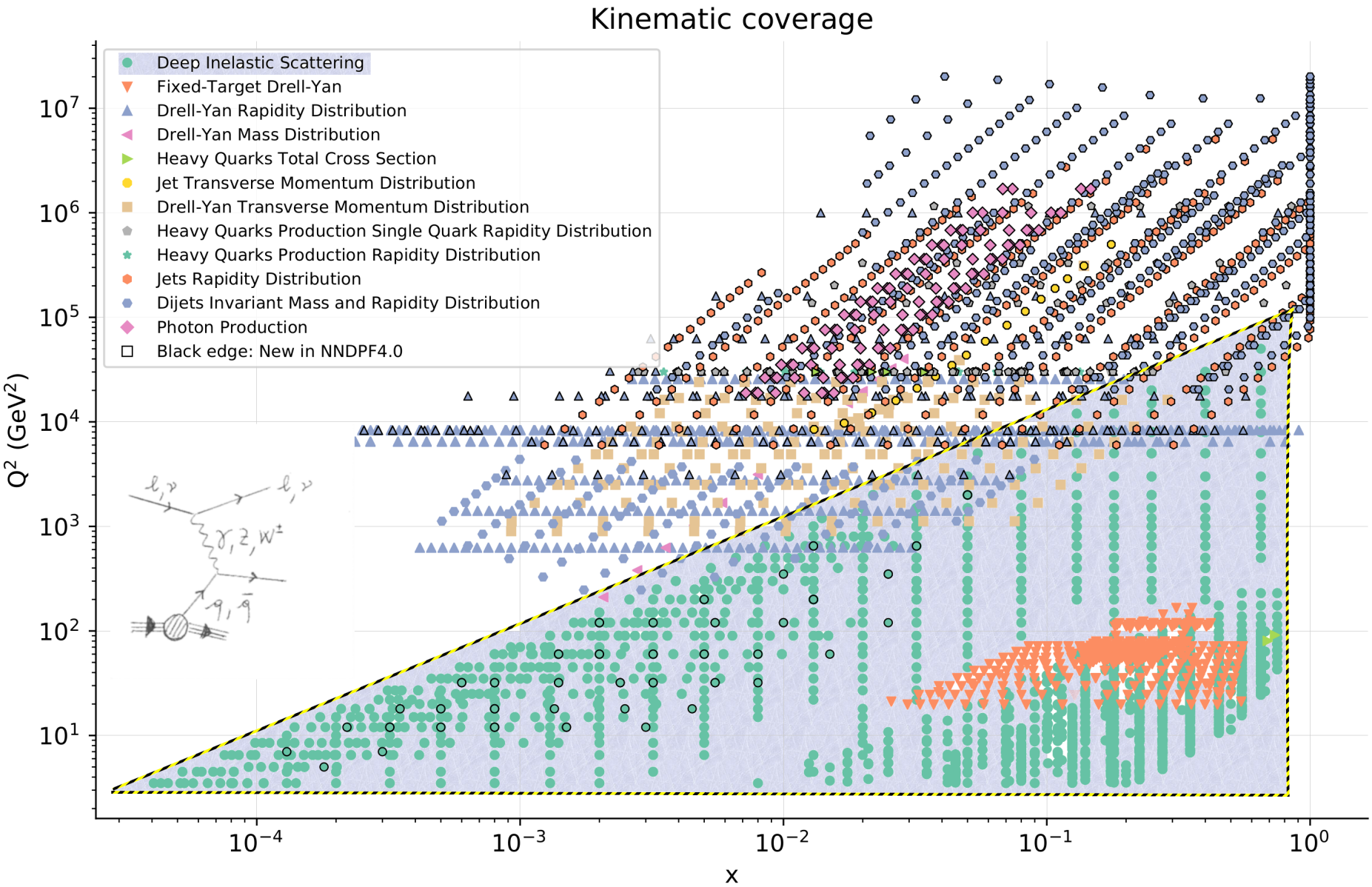
1.4 Data

Overview of experimental data



$$N_{\rm dat} = 4618$$

Deep Inelastic Scattering



Deep Inelastic Scattering

Re-write the cross section in terms of structure functions (F_2, F_3, F_L)

$$\frac{d^2\sigma^i}{dxdy} \propto Y_+ F_2^i \mp Y_- x F_3^i - y^2 F_L^i \quad Y_\pm = 1 \pm (1-y)^2 \quad F_L^i = F_2^i - 2x F_1^i \quad i = \mathrm{NC}(\gamma, \mathrm{Z}, \gamma\mathrm{Z}), \mathrm{CC}(\mathrm{W}^\pm)$$

NC DIS ($\ell p \to \ell X$) at LO (NMC, SLAC, BCDMS, HERA)

$$\left[F_2^\gamma, F_2^{\gamma Z}, F_2^Z\right] = x \sum_q \left[e_q^2, 2e_q g_V^q, (g_V^q)^2 + (g_A^q)^2\right] (q + \bar{q})$$

$$\left[F_3^\gamma, F_3^{\gamma Z}, F_3^Z\right] = \sum_q \left[0, 2e_q g_A^q, 2g_V g_A^q\right] (q - \bar{q})$$

$$\left[F_L^\gamma, F_L^{\gamma Z}, F_L^Z\right] = [0, 0, 0]$$

CC DIS ($\ell^- p \to \nu X$ or $\bar{\nu} p \to \ell^+ X$) at LO (CHORUS, NuTeV, HERA)

$$\left[F_2^{W^-}, F_2^{W^-}\right] = 2x \left[(u + \bar{d} + \bar{s} + c \dots), (d + \bar{u} + \bar{c} + s \dots)\right]$$

$$\left[F_3^{W^-}, F_3^{W^+}\right] = 2 \left[(u - \bar{d} - \bar{s} + c \dots), (d - \bar{u} - \bar{c} + s \dots)\right]$$

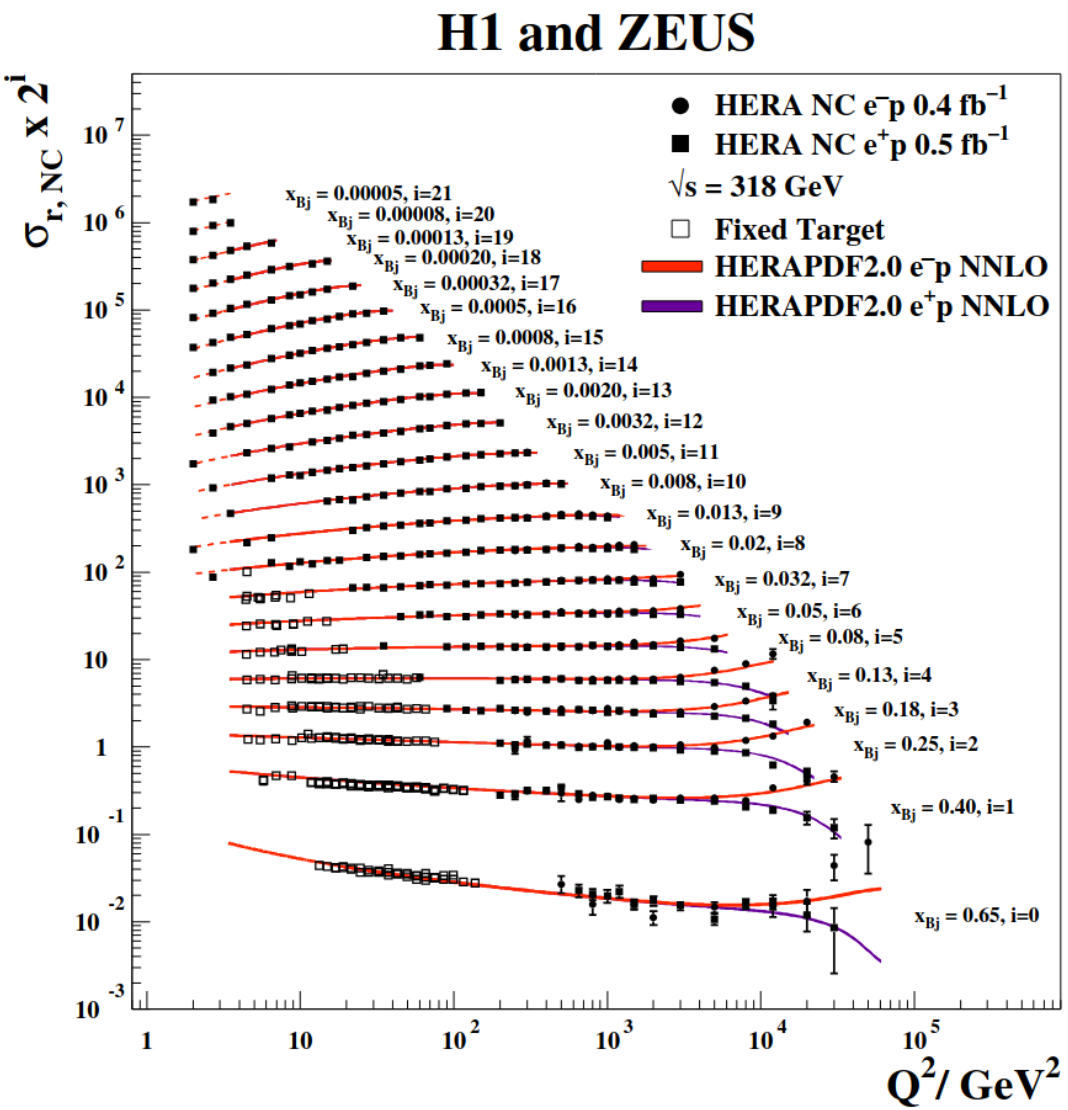
$$\left[F_L^{W^+}, F_L^{W^-}\right] = [0, 0]$$

isospin $(p \to n)$: $u^p = d^n \qquad d^p = u^n \qquad \bar{u}^p = \bar{d}^n \qquad \bar{d}^p = \bar{u}^n$

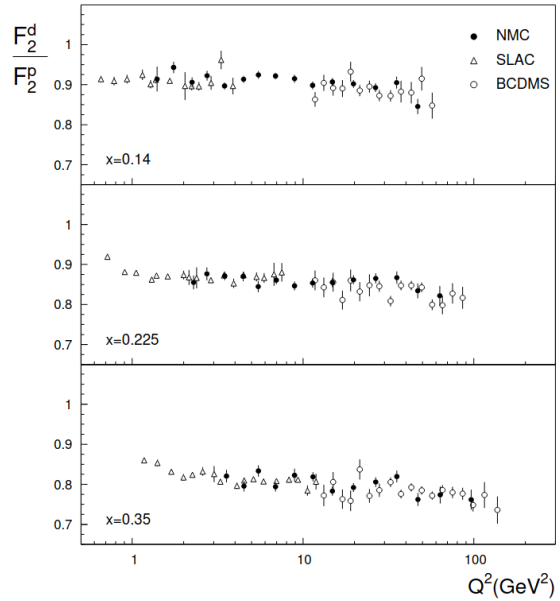
deuteron target approximated as the average of one proton and one neutron

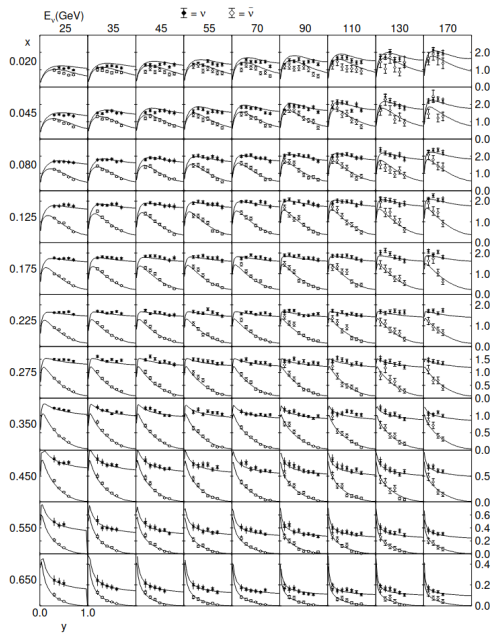
$$Q^2 \geq Q^2_{\min} \sim 1 \text{ GeV}^2 \qquad W^2 = Q^2(1-x)/x \geq W^2_{\min} \sim 10 \text{ GeV}^2$$

Deep Inelastic Scattering



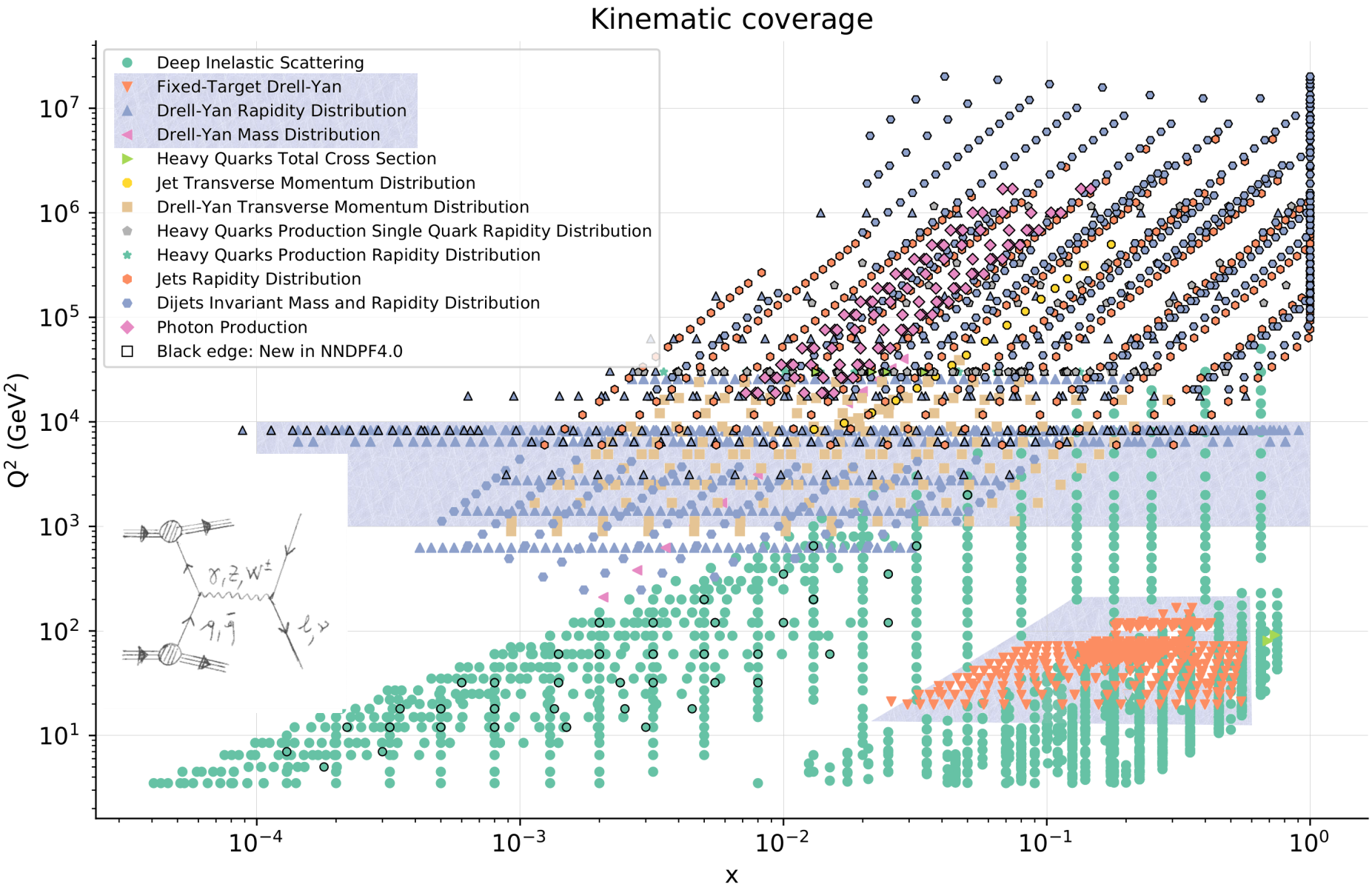
[HERA, EPJ C75 (2015) 580]





[NMC, NPB 487 (1997) 3; CHORUS, PLB 632 (2006) 65]

Drell-Yan


Kinematic coverage
Deep Inelastic Scattering
Fixed-Target Drell-Yan
Drell-Yan Rapidity Distribution
Drell-Yan Mass Distribution
Heavy Quarks Total Cross Section
Jet Transverse Momentum Distribution
Drell-Yan Transverse Momentum Distribution
Heavy Quarks Production Single Quark Rapidity Distribution
Heavy Quarks Production Rapidity Distribution
Jets Rapidity Distribution
Dijets Invariant Mass and Rapidity Distribution
Photon Production
Black edge: New in NNDPF4.0
Q^2 (GeV^2)
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
x
$\gamma, Z, W^\pm$
$q, \bar{q}$
l, ν

Drell-Yan

Work out the cross section differential in the rapidity of the lepton pair

$$\frac{d\sigma^i}{dy} \propto A^i \mathcal{L}^i \qquad i = \mathrm{NC}(\gamma, \mathrm{Z}, \gamma\mathrm{Z}), \mathrm{CC}(\mathrm{W}^\pm)$$

NC DY (γ, Z) at LO

$$\left[A^\gamma \mathcal{L}^\gamma, A^Z \mathcal{L}^Z\right] = \left[\sum_q e_q^2 q\bar{q}, \sum_{q,q'} |V_{qq'}^{\mathrm{CKM}}| q\bar{q}'\right]$$

CC DY $(W^\pm)$ at LO

$$\left[A^{W^\pm} \mathcal{L}^{W^\pm}\right] = \left[\sum_q \left((g_V^q)^2 + (g_A^q)^2\right) q\bar{q}\right]$$

isospin $(p \to n)$: $u^p = d^n \qquad d^p = u^n \qquad \bar{u}^p = \bar{d}^n \qquad \bar{d}^p = \bar{u}^n$

deuteron target approximated as the average of one proton and one neutron

different experiments measure different cross section combinations

$$\frac{\sigma_{pn}^Z}{\sigma_{pp}^Z} \approx \frac{4/9 u\bar{d} + 1/9 d\bar{u}}{4/9 u\bar{u} + 1/9 d\bar{d}} \to \frac{\bar{d}}{\bar{u}}$$

DY p/d asymmetry (NuSea, SeaQuest)

$$\frac{\sigma_{p\bar{p}}^{W^+}}{\sigma_{p\bar{p}}^{W^-}} \approx \frac{ud + \bar{d}\bar{u}}{du + \bar{u}\bar{d}} \to \frac{ud}{du}$$

$W^\pm$ asymmetry (CDF, D0)

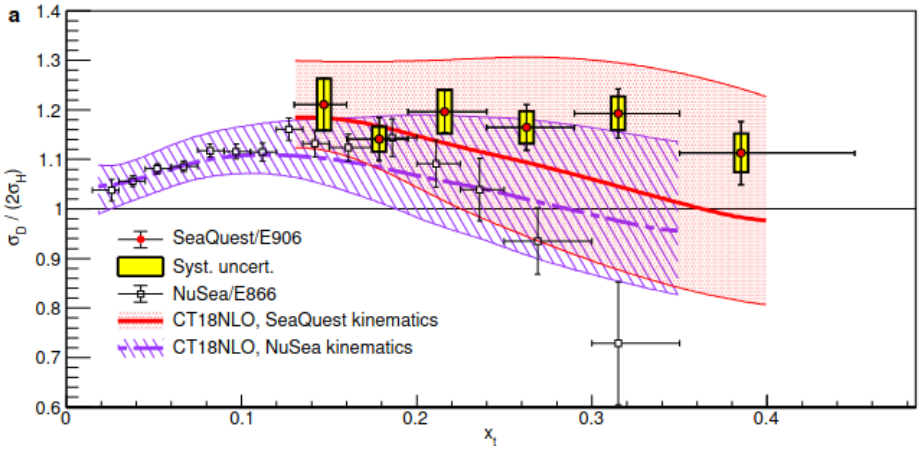
$$\sigma_{pp}^{W^+} \approx u\bar{d} + c\bar{s} \quad \sigma_{pp}^Z \approx u\bar{u} + d\bar{d} + s\bar{s} \to s, \bar{s}$$

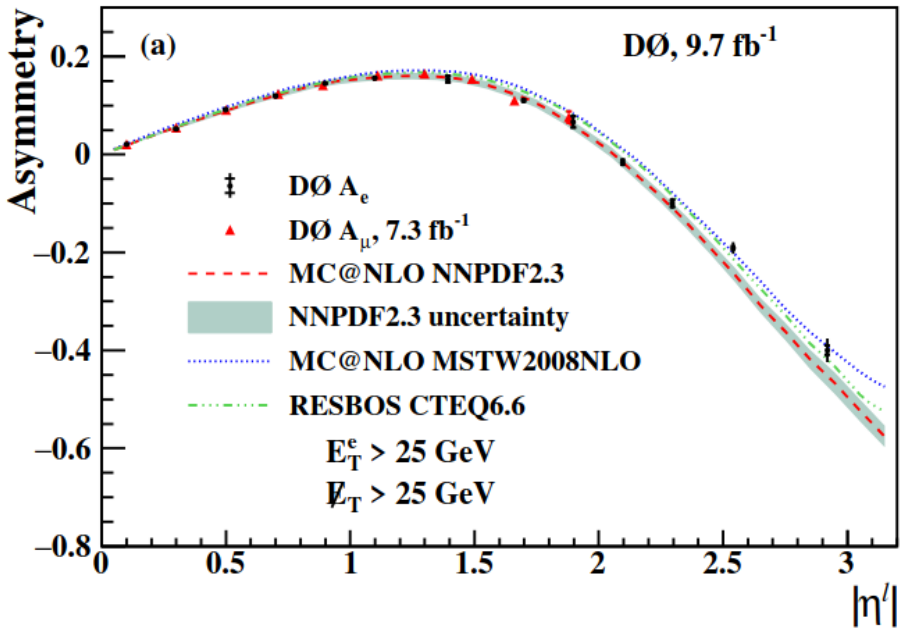
$W^\pm$ and Z production (ATLAS, CMS, LHCb)

$$\frac{\sigma_{pp}^{W^+}}{\sigma_{pp}^{W^-}} \approx \frac{u\bar{d} + \bar{d}u}{d\bar{u} + \bar{u}d} \to \bar{u} - \bar{d}$$

$W^\pm$ muon asymmetry (ATLAS, CMS)

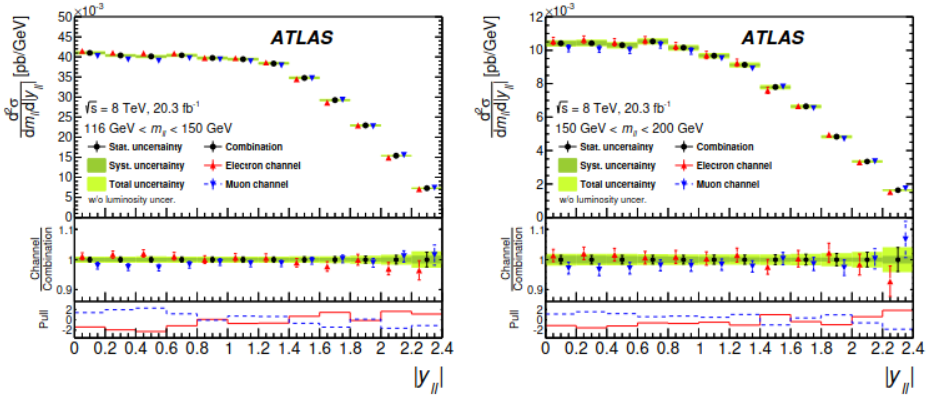
Drell-Yan

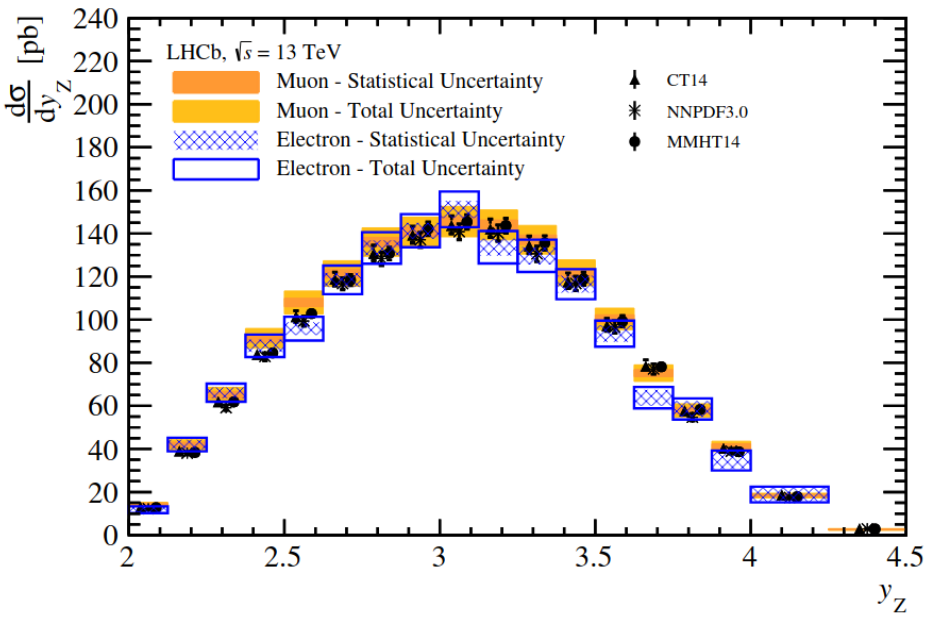




[SeaQuest, Nature 590 (2021) 7847]

[D0, PRD 91 (2015) 032007]

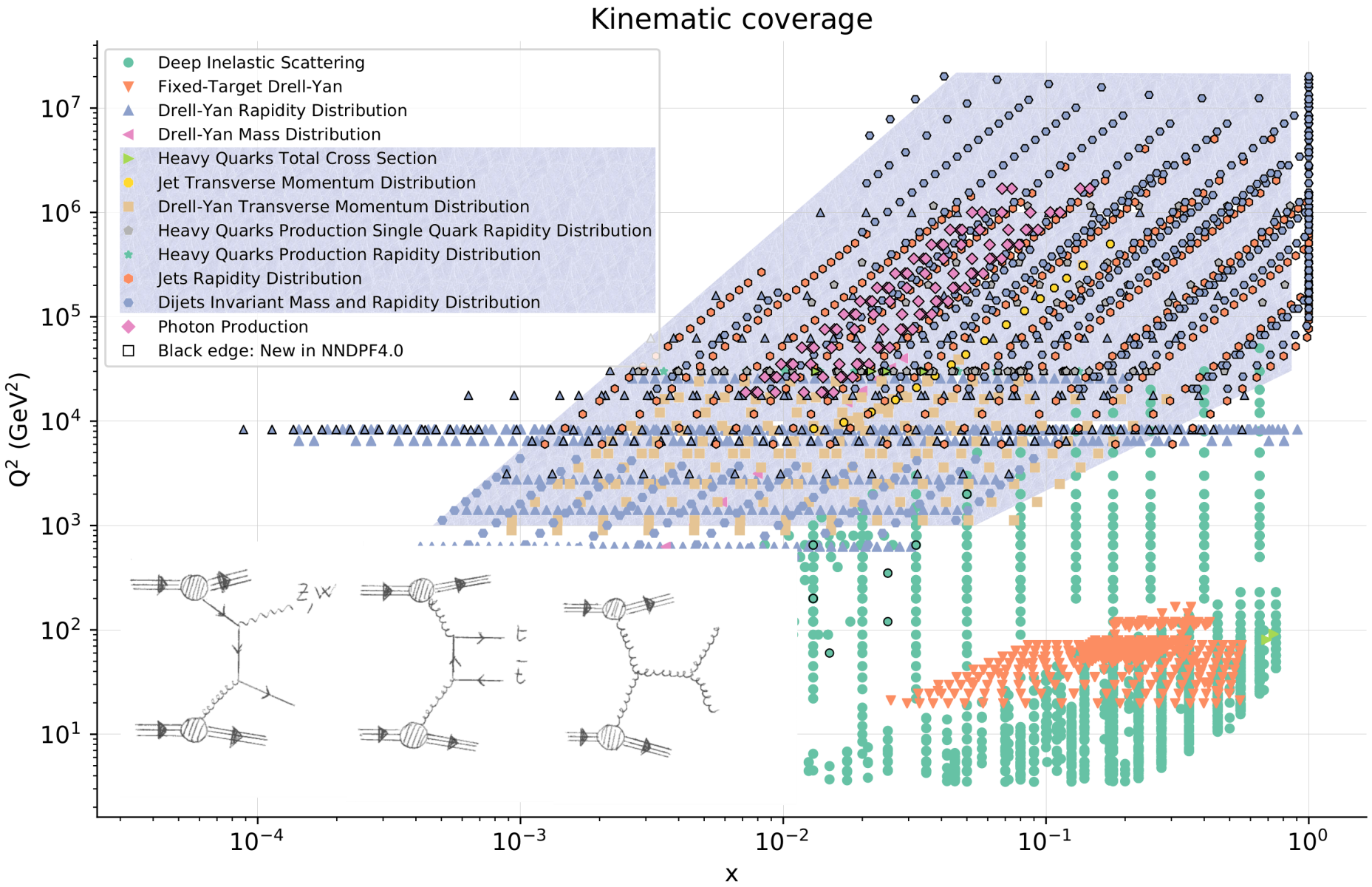




[ATLAS, JHEP 08 (2016) 009]

[LHCb, JHEP 09 (2016) 136]

Zp_T, $t\bar{t}$, Single-Inclusive Jet, and Dijet Production



Zp_T, $t\bar{t}$, Single-Inclusive Jet, and Dijet Production

Various differential distributions, all proportional to the gluon PDF at LO

Zp_T production: $\frac{d\sigma^2}{dp_T^Z dm_{\ell\ell}}$, $\frac{d\sigma^2}{dp_T^Z dy_Z}$ (ATLAS) $\frac{d\sigma}{dp_T^Z}$ (CMS)

need one final-state parton, then initial-state quark and gluon are on the same footing

wide p_T range, constraints on a wide x (typically intermediate) and Q^2 range

$t\bar{t}$ production: $\frac{d\sigma}{dp_T^t}$, $\frac{d\sigma}{dy^t}$, $\frac{d\sigma}{dy^{t\bar{t}}}$, $\frac{d\sigma}{dm^{t\bar{t}}}$ (ATLAS and CMS)

process initiated by two gluons in the initial state

differential cross sections reconstructed at parton level (additional systematics)

normalise by $\sigma_{\rm tot}^{t\bar{t}}$ (systematics largely cancel, but loose control on PDF shape)

particle-level cross sections are theoretically more complicated

wide rapidity range, constraints on the large-x region

single-inclusive jet and dijet production: $\frac{d\sigma^2}{dydp_T}$, $\frac{d\sigma^2}{dy_{1,2}dm_{1,2}}$ (HERA, ATLAS and CMS)

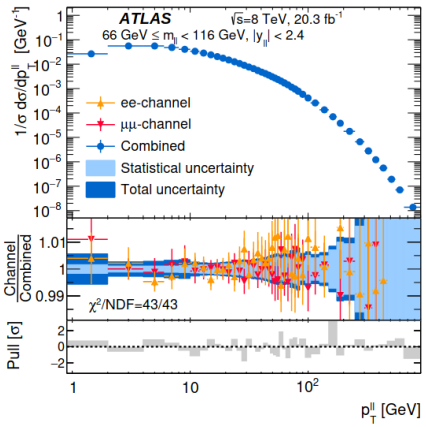
process initiated by two gluons in the initial state

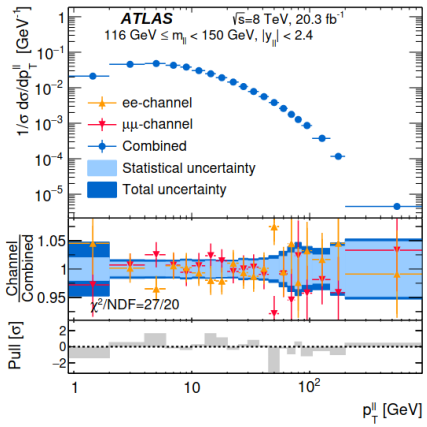
be careful with systematic uncertainties (mostly driven by jet energy reconstruction)

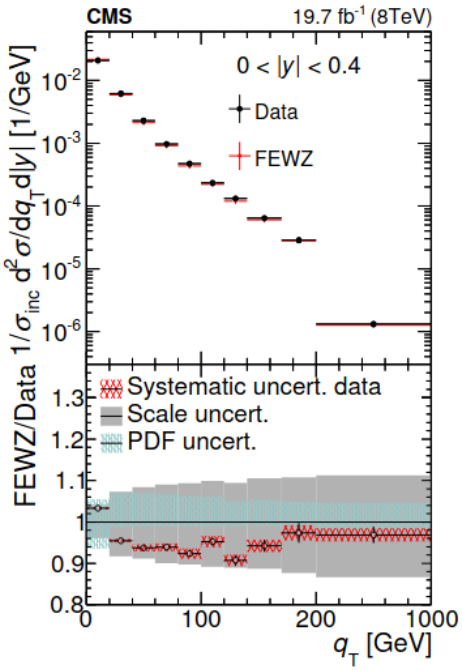
can also be measured in DIS

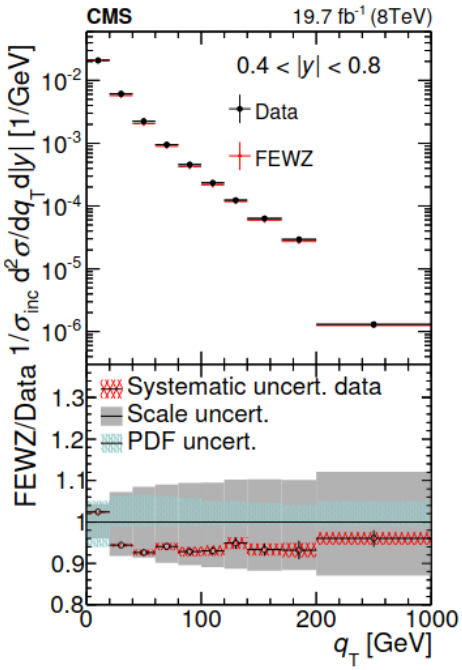
wide rapidity range, constraints on the large-x region

Zp_T, $t\bar{t}$, Single-Inclusive Jet, and Dijet Production



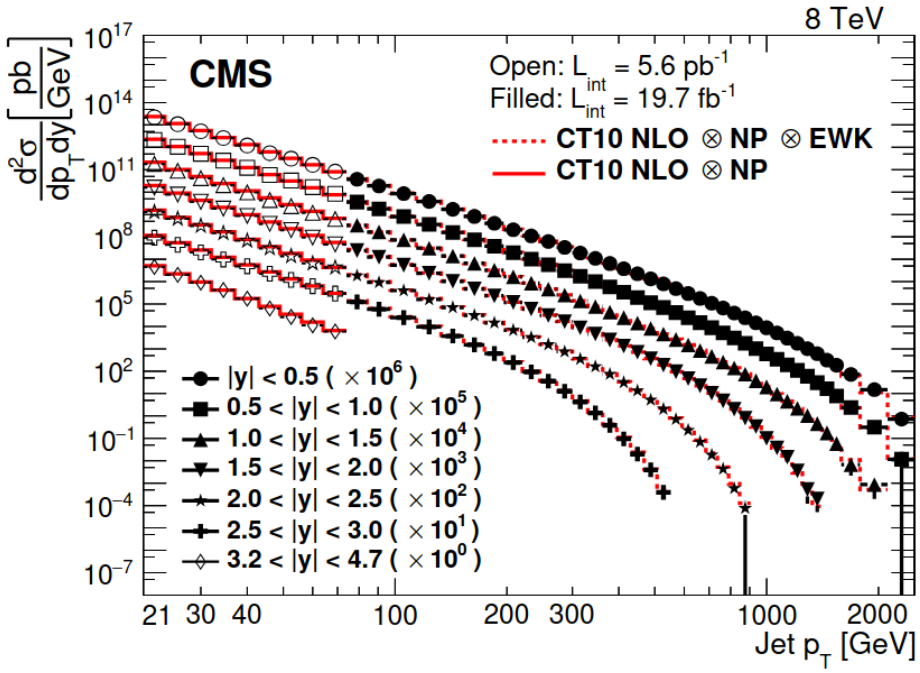


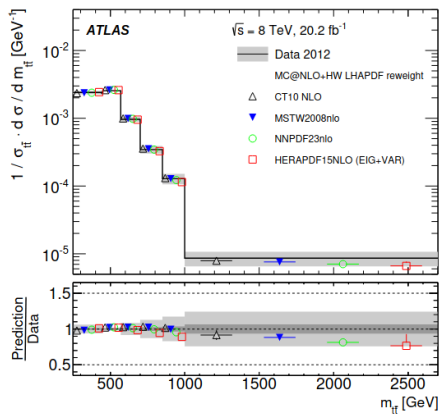


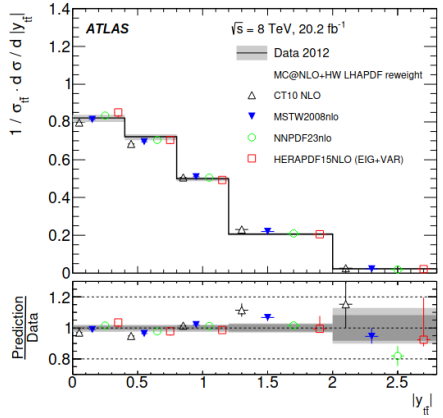


[ATLAS, EPJ C76 (2016) 291]

[CMS PhysLB 749 (2015) 187]







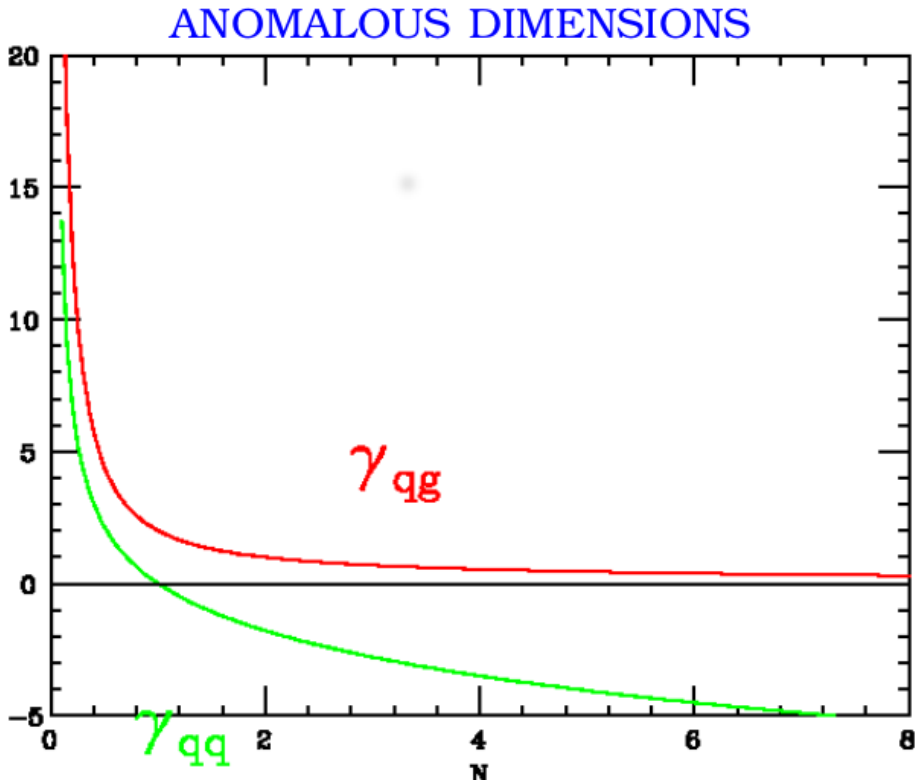
[CMS, JHEP 03 (2017) 156]

[ATLAS, EPJ C76 (2016) 538]

Scaling Violations and Heavy Flavour Production

Scale dependence in DIS
of singlet structure function

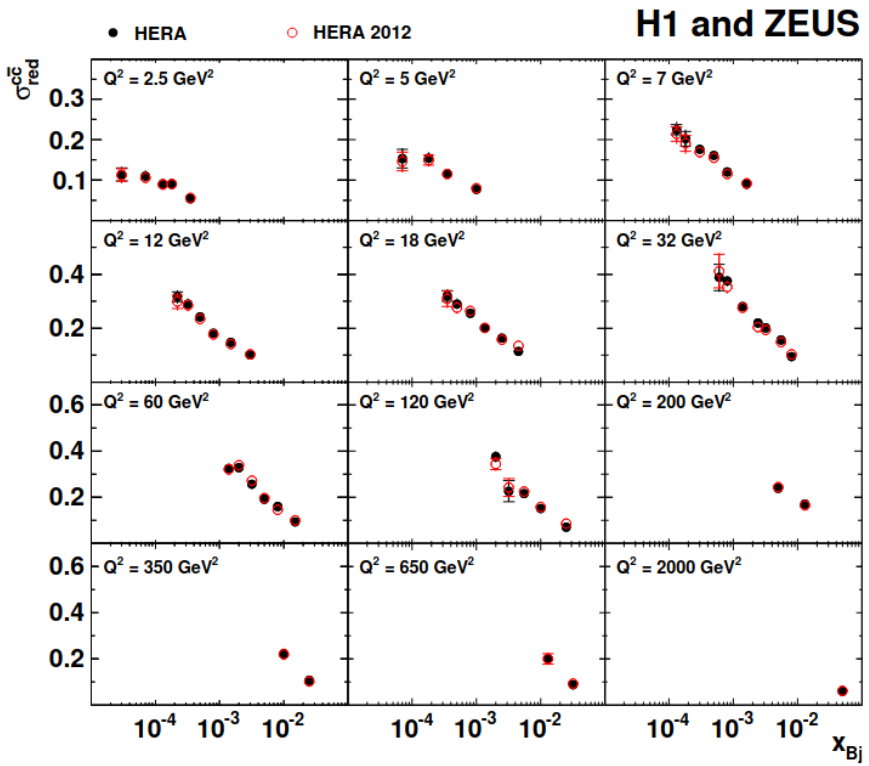
$$\frac{d}{d\ln(Q^2)} F_2^{\Sigma} \approx \frac{\alpha_s}{2\pi} \left[\gamma_{qq} f_{\Sigma} + 2n_f \gamma_{gq} f_g\right]$$



the gluon PDF can be determined at small x
from DIS scaling violations (from HERA)

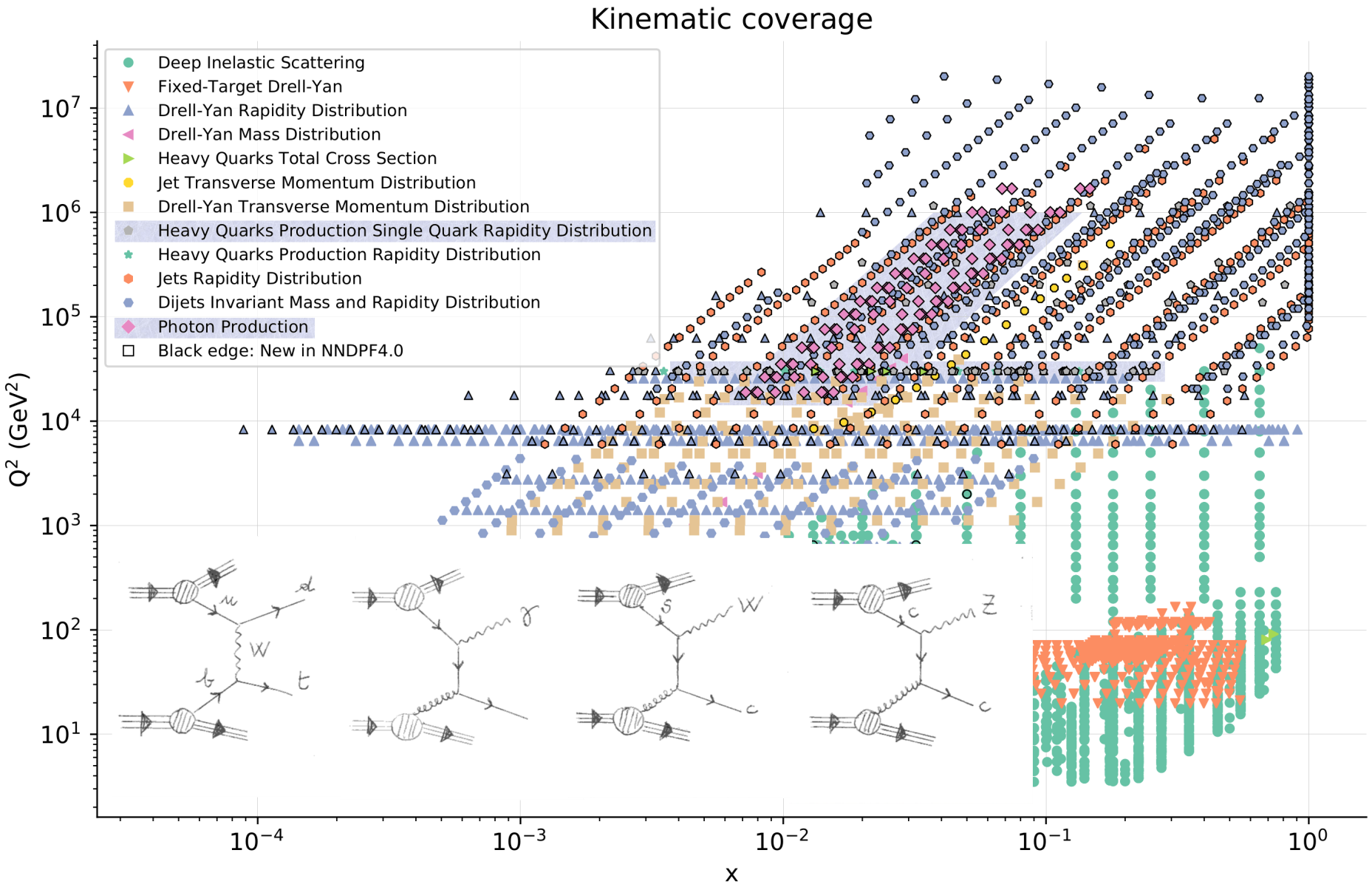
Heavy quark production in DIS
initiated by gluons

$$\sigma_{\rm red}^{c,b} = F_2^{c,b} - \frac{y^2}{1+(1-y)^2} F_L^{c,b}$$



the gluon PDF is determined by tagging
a c or a b quark in the final state (HERA)

Single t, Direct γ, W, Z+c-jets



Single t, Direct γ, W, Z+c-jets

Other processes, currently limited by experimental uncertainties

Single t production (t-channel): $\frac{d\sigma}{dp_T^t}$, $\frac{d\sigma}{dy^t}$ $\frac{d\sigma}{dp_T^{\bar{t}}}$, $\frac{d\sigma}{dy^{\bar{t}}}$ (ATLAS, CMS)

partonic cross sections similar to CC DIS; t reconstructed from Wb decay

potential sensitivity to $\bar{u}$ and $\bar{d}$, also through ratios of t and $\bar{t}$ production

potential currently limited by large experimental uncertainties

Prompt γ production: $\frac{d\sigma^2}{dE_T^\gamma y^\gamma}$ (ATLAS and CMS)

gluon-quark-initiated Compton scattering

potential sensitivity to the gluon PDF

potential currently limited by large experimental uncertainites

W, Z + charm-tagged jets: $\frac{d\sigma}{dy}$ (ATLAS and CMS)

$W + c$: sensitivity to strange PDF (and $s - \bar{s}$ asymmetry)

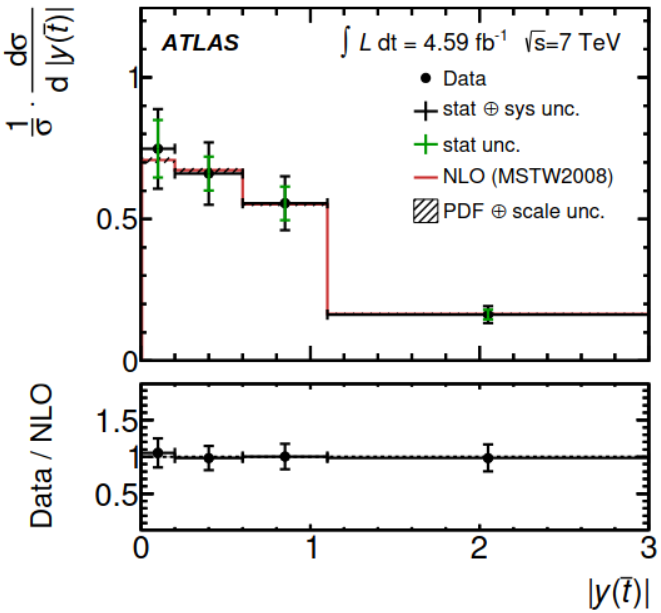
$W + Z$: sensitivity to charm PDF (including intrinsic charm and $c - \bar{c}$ asymmetry)

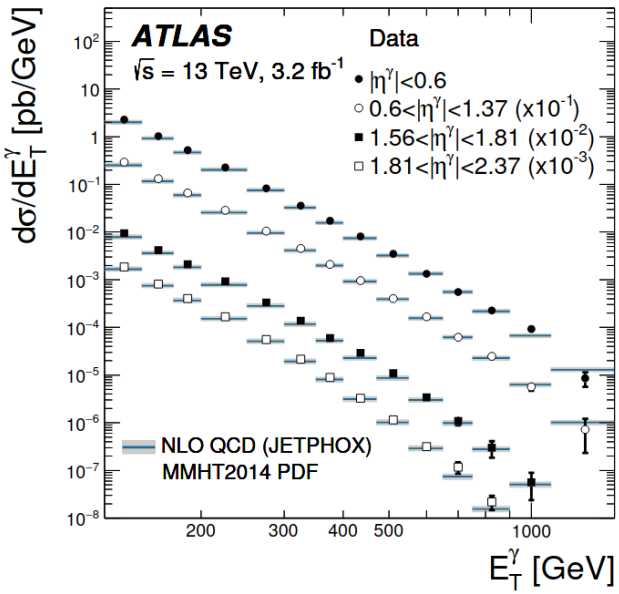
be careful with systematic uncertainties (due to jet tagging algorithm)

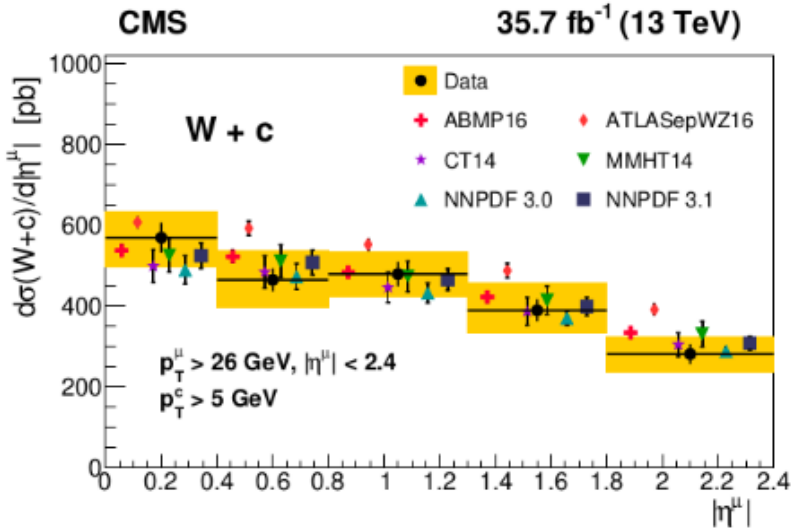
More exclusive processes: double gauge boson production, multijet production, ...

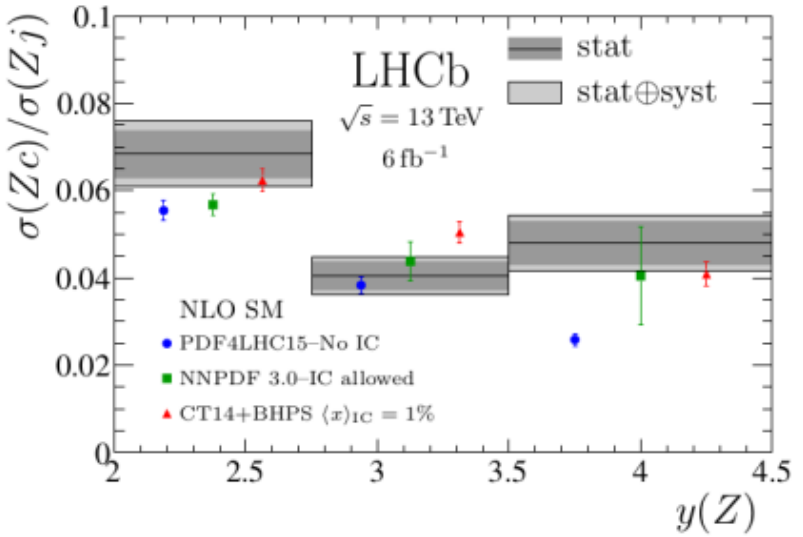
generally less precise and potentially contaminated by BSM physics

Single t, Prompt γ, W, Z+c-jets









[ATLAS, PRD 90 (2014) 112006; PLB 770 (2017) 473]

[CMS, EPJ C79 (2019) 269 ; LHCb, PRL 128 (2022) 082001]

All the data together

	Hadronic Process	Partonic Process	PDFs probed	x coverage
Lepton-nucleon	$\ell^{\pm}\{p,n\} \to \ell^{\pm} + X$	$\gamma^* q \to q$	$q, \bar{q}, g$	$x \gtrsim 0.01$
	$\ell^{\pm} n/p \to \ell^{\pm} + X$	$\gamma^* d/u \to d/u$	d/u	$x \gtrsim 0.01$
	$\nu(\bar{\nu})N \to \mu^-(\mu^+) + X$	$W^* q \to q'$	$q, \bar{q}$	$0.01 \lesssim x \lesssim 0.5$
	$\nu N \to \mu^- \mu^+ + X$	$W^* s \to c$	s	$0.01 \lesssim x \lesssim 0.2$
	$\bar{\nu} N \to \mu^+ \mu^- + X$	$W^* \bar{s} \to \bar{c}$	$\bar{s}$	$0.01 \lesssim x \lesssim 0.2$
	$e^{\pm} p \to e^{\pm} + X$	$\gamma^* q \to q$	$g, q, \bar{q}$	$0.0001 \lesssim x \lesssim 0.1$
	$e^+ p \to \bar{\nu} + X$	$W^+\{d,s\} \to \{u,c\}$	d, s	$x \gtrsim 0.01$
	$e^{\pm} p \to e^{\pm} c\bar{c} + X$	$\gamma^* c \to c, \gamma^* g \to c\bar{c}$	c, g	$0.0001 \lesssim x \lesssim 0.1$
	$e^{\pm} p \to jet(s) + X$	$\gamma^* g \to q\bar{q}$	g	$0.01 \lesssim x \lesssim 0.1$
Proton-(anti)proton	$pp \to \mu^+ \mu^- + X$	$u\bar{u}, d\bar{d} \to \gamma^*$	$\bar{q}$	$0.015 \lesssim x \lesssim 0.35$
	$pn/pp \to \mu^+ \mu^- + X$	$(u\bar{d})/(u\bar{u}) \to \gamma^*$	$\bar{d}/\bar{u}$	$0.015 \lesssim x \lesssim 0.35$
	$p\bar{p}(pp) \to jet(s) + X$	$gg, qg, qq \to jets$	g, q	$0.005 \lesssim x \lesssim 0.5$
	$p\bar{p} \to (W^{\pm} \to \ell^{\pm}\nu) + X$	$ud \to W^+, \bar{u}\bar{d} \to W^-$	$u, d, \bar{u}, \bar{d}$	$x \gtrsim 0.05$
	$pp \to (W^{\pm} \to \ell^{\pm}\nu) + X$	$u\bar{d} \to W^+, d\bar{u} \to W^-$	$u, d, \bar{u}, \bar{d}, (g)$	$x \gtrsim 0.001$
	$p\bar{p}(pp) \to (Z \to \ell^+\ell^-) + X$	$uu, dd(u\bar{u}, d\bar{d}) \to Z$	$u, d(g)$	$x \gtrsim 0.001$
	$pp \to (W + c) + X$	$gs \to W^- c, g\bar{s} \to W^+ \bar{c}$	$s, \bar{s}$	$x \sim 0.01$
	$pp \to (Z + c) + X$	$gc \to Zc, g\bar{c} \to Z\bar{c}$	$c, \bar{c}$	$x \sim 0.01$
	$pp \to t\bar{t} + X$	$gg \to t\bar{t}$	g	$x \sim 0.1$
	$pp \to t, \bar{t} + X$	$gq \to t, \bar{t}q$	u, d	$x \sim 0.01$
	$pp \to \gamma + X$	$gq \to \gamma q$	g	$x \sim 0.01$

Overview of current unpolarised PDF determinations

	NNPDF4.0	MSHT20	CT18	HERAPDF2.0	CJ22	ABMP16
Fixed-target DIS	☑	☑	☑	☒	☑	☑
JLAB	☒	☒	☒	☒	☑	☒
HERA I+II	☑	☑	☑	☑	☑	☑
HERA jets	☑	☒	☒	☑	☒	☒
Fixed target DY	☑	☑	☑	☒	☑	☑
Tevatron W, Z	☑	☑	☑	☒	☑	☑
LHC vector boson	☑	☑	☑	☒	☑	☑
LHC $W + c$ $Z + c$	☑	☒	☒	☒	☒	☒
Tevatron jets	☑	☑	☑	☒	☑	☒
LHC jets	☑	☑	☑	☒	☒	☒
LHC top	☑	☑	☒	☒	☒	☑
LHC single t	☑	☒	☒	☒	☒	☒
LHC prompt γ	☑	☒	☒	☒	☒	☒
statistical treatment	Monte Carlo	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2 = 1$	Hessian $\Delta\chi^2 = 1.645$	Hessian $\Delta\chi^2 = 1$
parametrisation	Neural Network	Chebyschev pol.	Bernstein pol.	polynomial	polynomial	polynomial
HQ scheme	FONLL	TR′	ACOT-χ	TR′	ACOT-χ	FFN
accuracy	aN^{3}LO	aN^{3}LO	NNLO	NNLO	NLO	NNLO
latest update	EPJ C82 (2022) 428	EPJ C81 (2021) 341	PRD 103 (2021) 014013	EPJ C82 (2022) 243	PRD 107 (2023) 113005	PRD 96 (2017) 014011

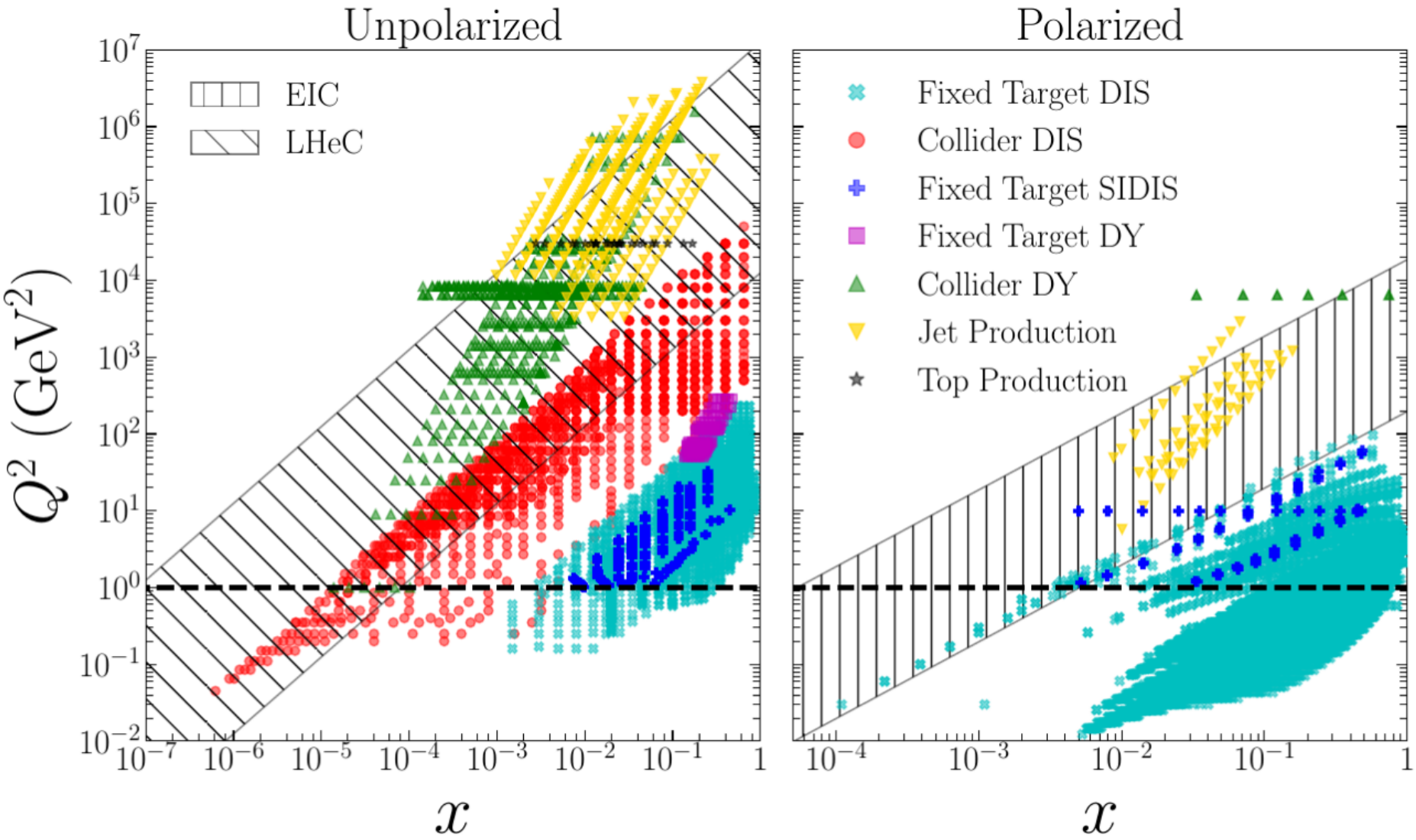
All PDF sets are available as (x, Q^2) interpolation grids through the LHAPDF library

And polarised PDFs?

	Hadronic Process	Partonic Process	PDFs probed	x coverage
Lepton-nucleon	$\ell^\pm\{p,n\} \to \ell^\pm + X$	$\gamma^* q \to q$	$\Delta q, \Delta\bar{q}, \Delta g$	$x \gtrsim 0.003$
	$\ell^\pm\{p,n\} \to \ell^\pm + h + X$	$\gamma^* q \to q$	$\Delta q, \Delta\bar{q}, \Delta g$	$0.005 \lesssim x \lesssim 0.5$
Proton-proton	$pp \to jet(s) + X$	$gg, qg, qq \to jets$	$\Delta g, \Delta q$	$0.05 \lesssim x \lesssim 0.2$
	$pp \to (W^\pm \to \ell^\pm\nu) + X$	$u\bar{d} \to W^+,\ d\bar{u} \to W^-$	$\Delta u, \Delta d, \Delta\bar{u}, \Delta\bar{d}, (\delta g)$	$0.05 \lesssim x \lesssim 0.4$
	$pp \to h + X$	$gg, qg, qq \to h$	$\Delta g, \Delta q$	$0.05 \lesssim x \lesssim 0.4$

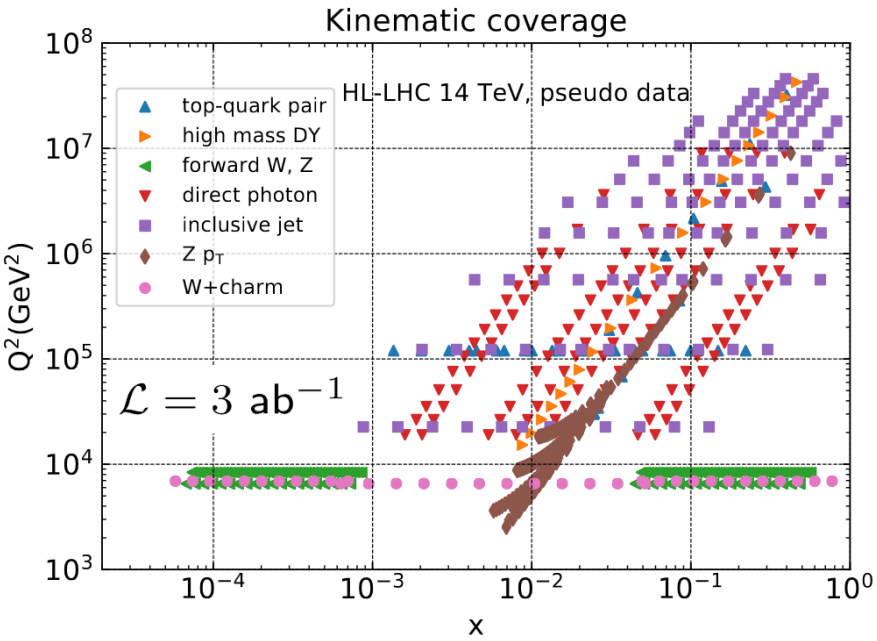
	NNPDFpol2.0	MAPPDFpol1.0	BDSSV25	JAM25
Fixed-target DIS	☑	☑	☑	☑
Fixed-target SIDIS	☒	☑	☑	☑
RHIC vector boson	☑	☒	☑	☑
RHIC jets	☑	☒	☑	☑
RHIC hadrons	☒	☒	☑	☒
statistical treatment	Monte Carlo	Monte Carlo	Monte Carlo	Monte Carlo
parametrisation	Neural Network	polynomial	polynomial	polynomial
HQ scheme	FONLL	ZM	ZM	ZM
accuracy	NNLO	NNLO	NNLO	NLO
latest update	arXiv:2503.11814	PLB 865 (2025) 139497	PRL 133 (2024) 15	arXiv:2506.13616

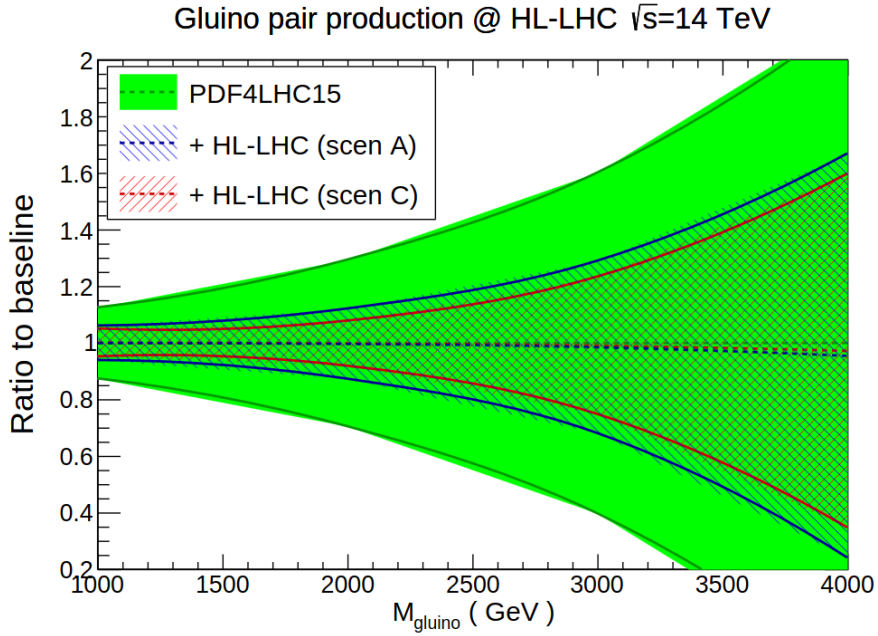
Unpolarised and polarised PDFs: data

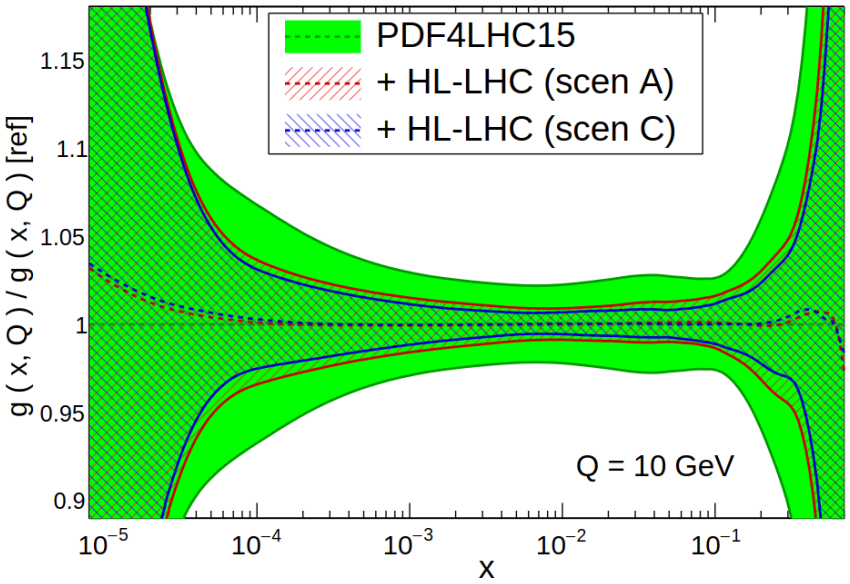


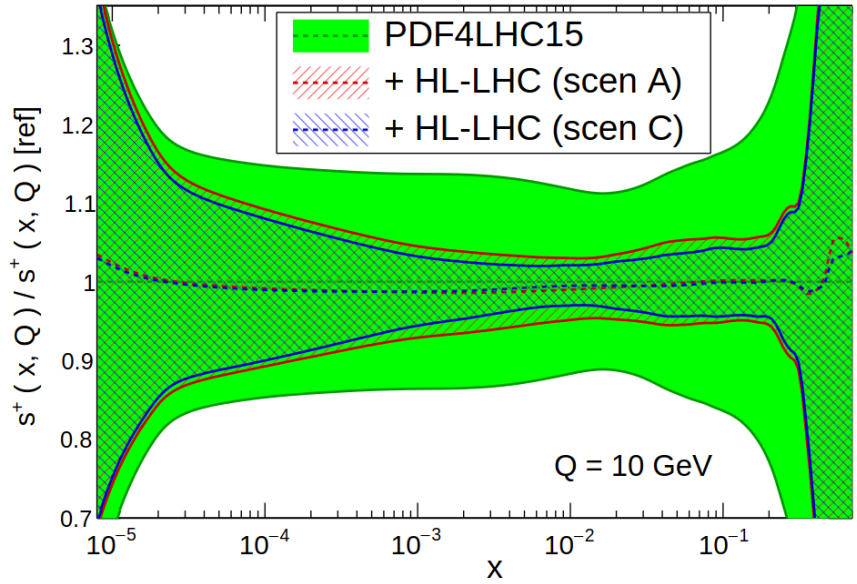
Thousands of data for unpolarised PDFs; hundreds of data for unpolarised PDFs

Impact of future data: HL-LHC



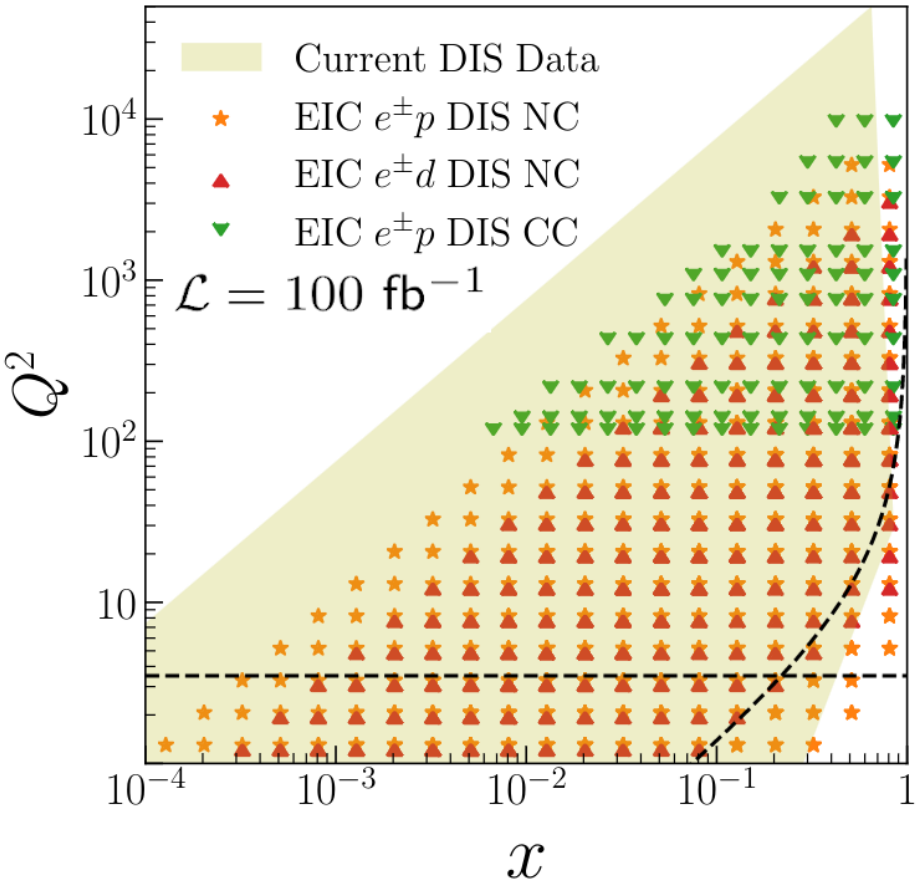






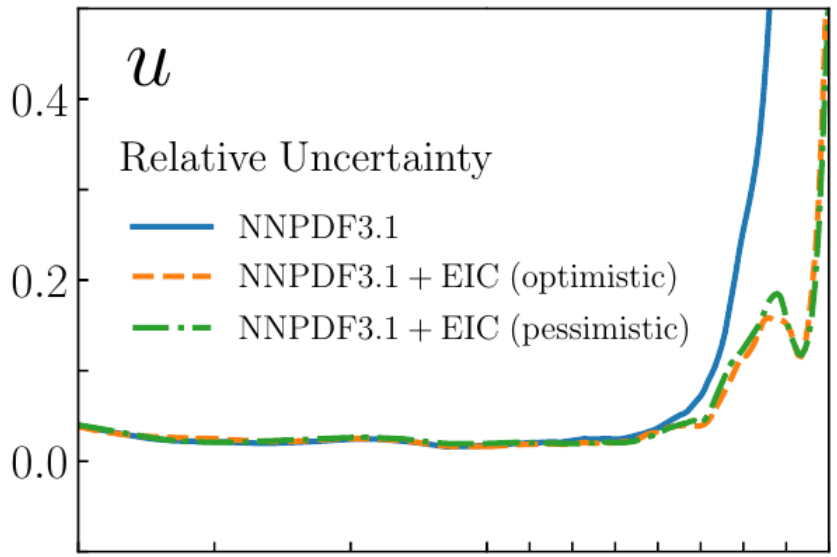
[EPJ C78 (2018) 962]

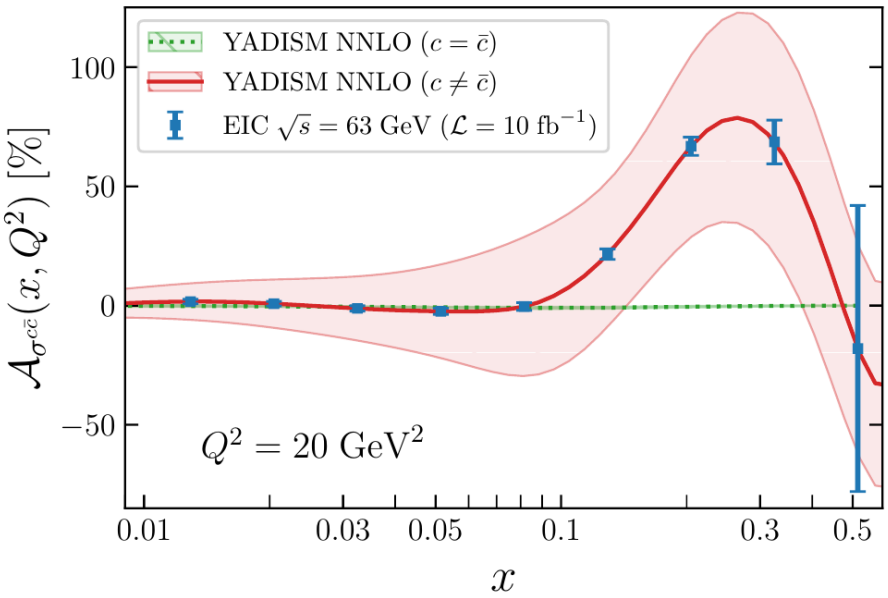
Impact of future data: unpolarised EIC



$E_\ell \times E_p$ [GeV]: 18× 275; 10× 100; 5× 100

$$\mathcal{A}_{\sigma^{c\bar{c}}} = \frac{\sigma^{c}_{\rm red} - \sigma^{\bar{c}}_{\rm red}}{\sigma^{c\bar{c}}_{\rm red}}$$

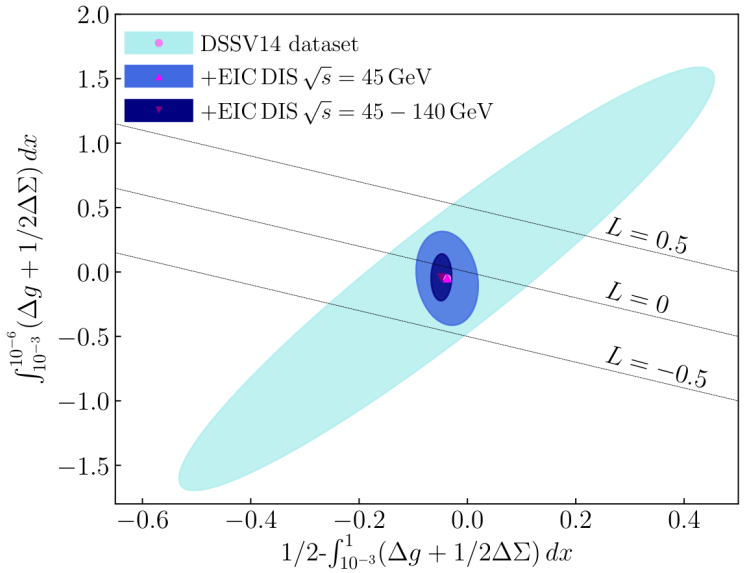




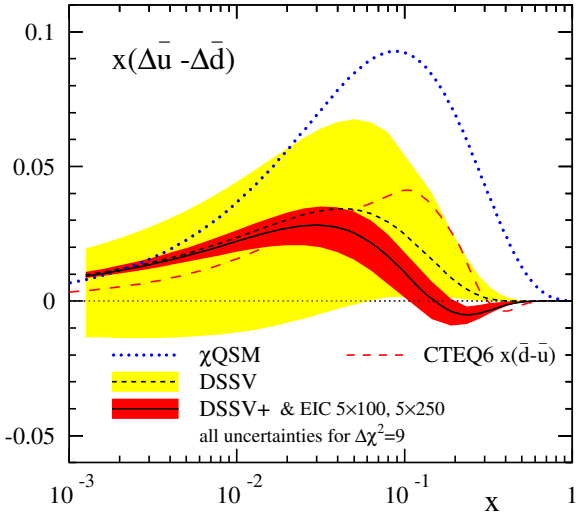
[PRD 103 (2021) 096005; PRD 109 (2024) L091501]

Impact of future data: polarised EIC

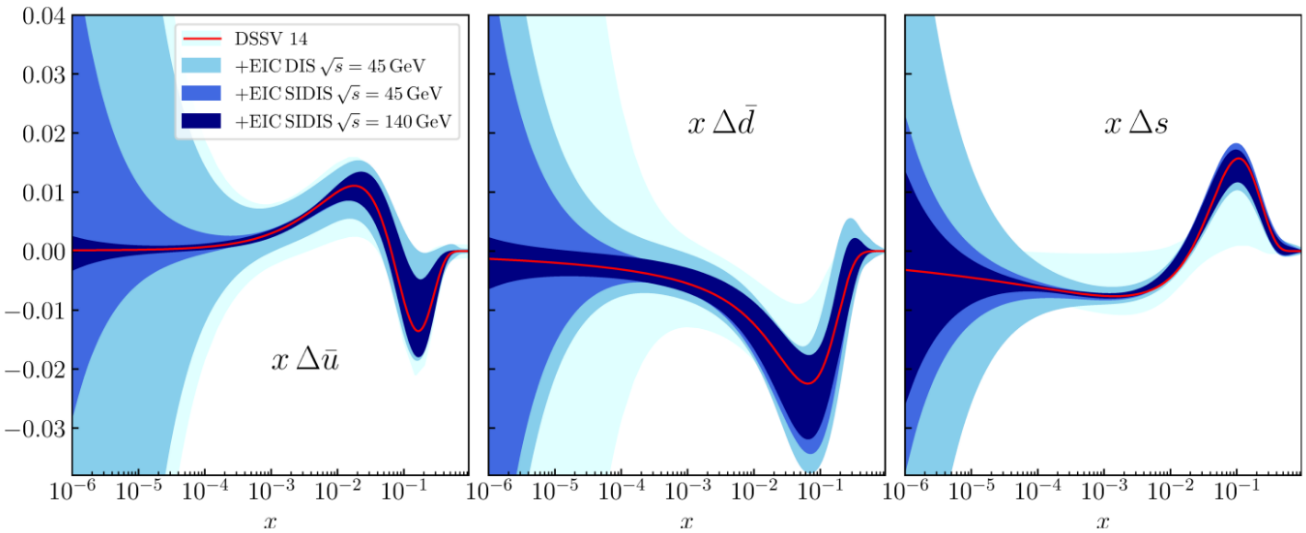
NC DIS: quark and gluon spin



CC DIS: light antiquark asymmetry

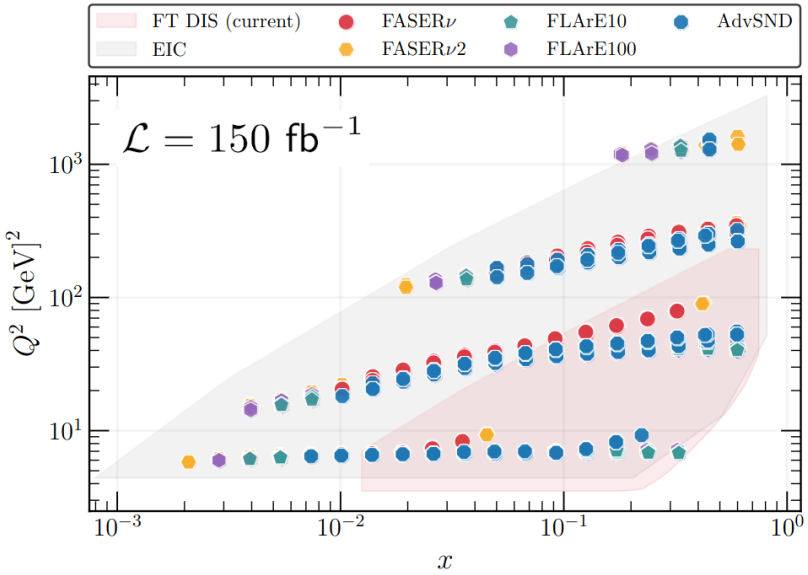


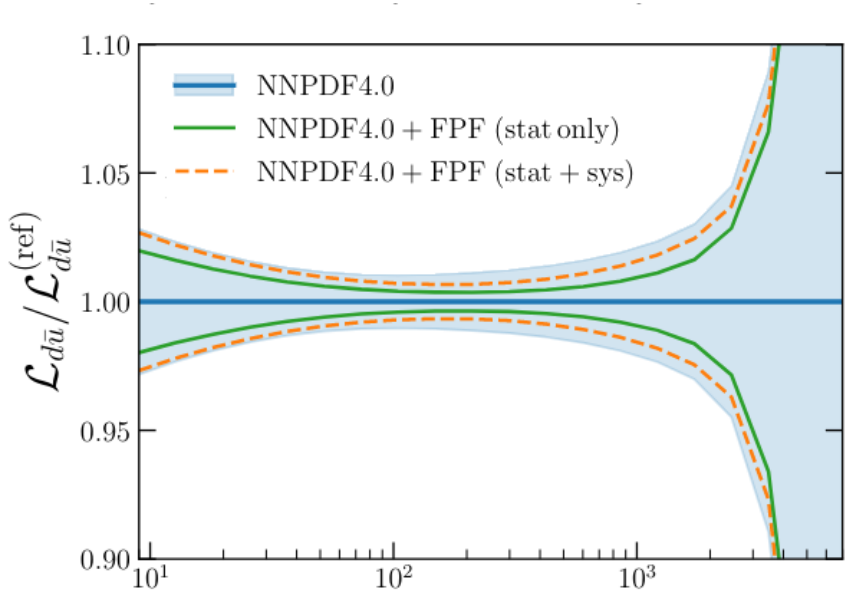
NC SIDIS with π and K: antiquark PDFs

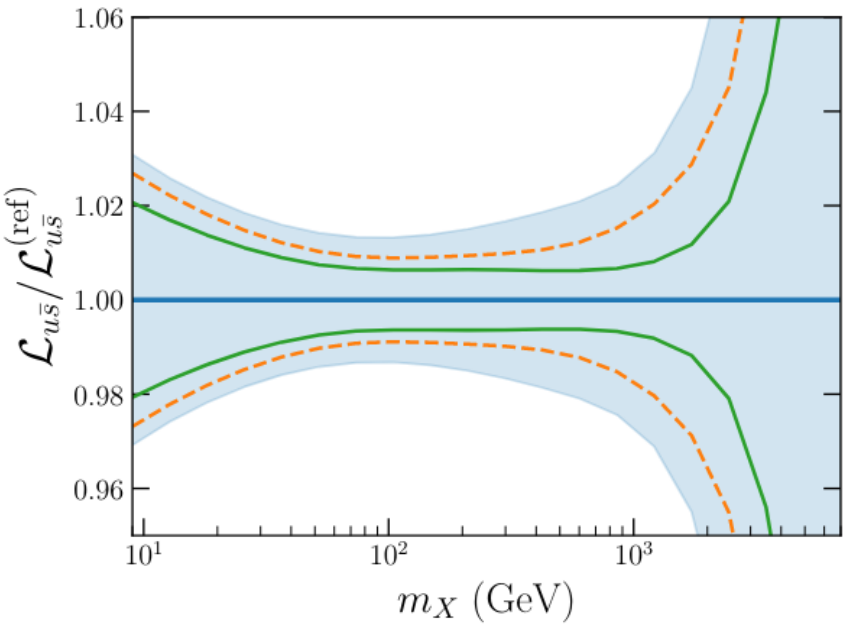


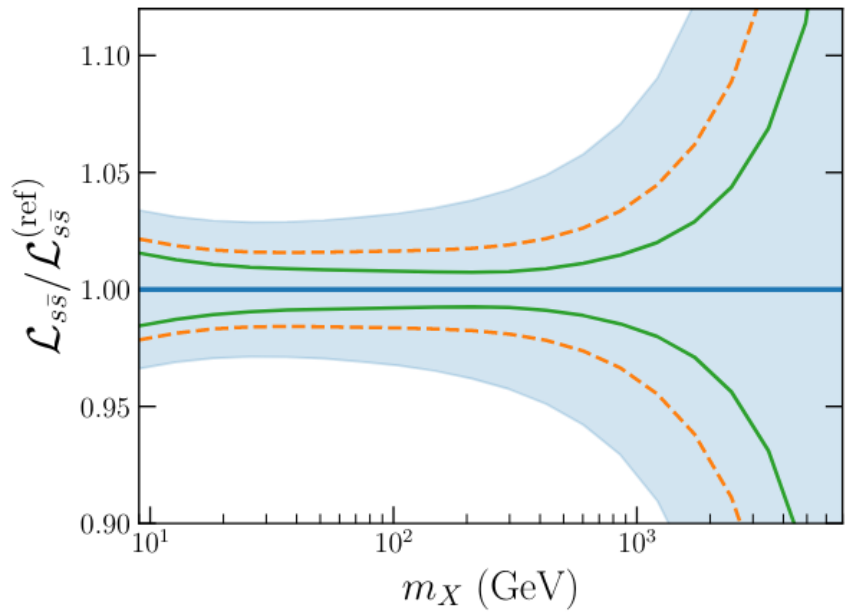
[PRD 88 (2013) 114025; PRD 102 (2020)]

Impact of future data: FPF





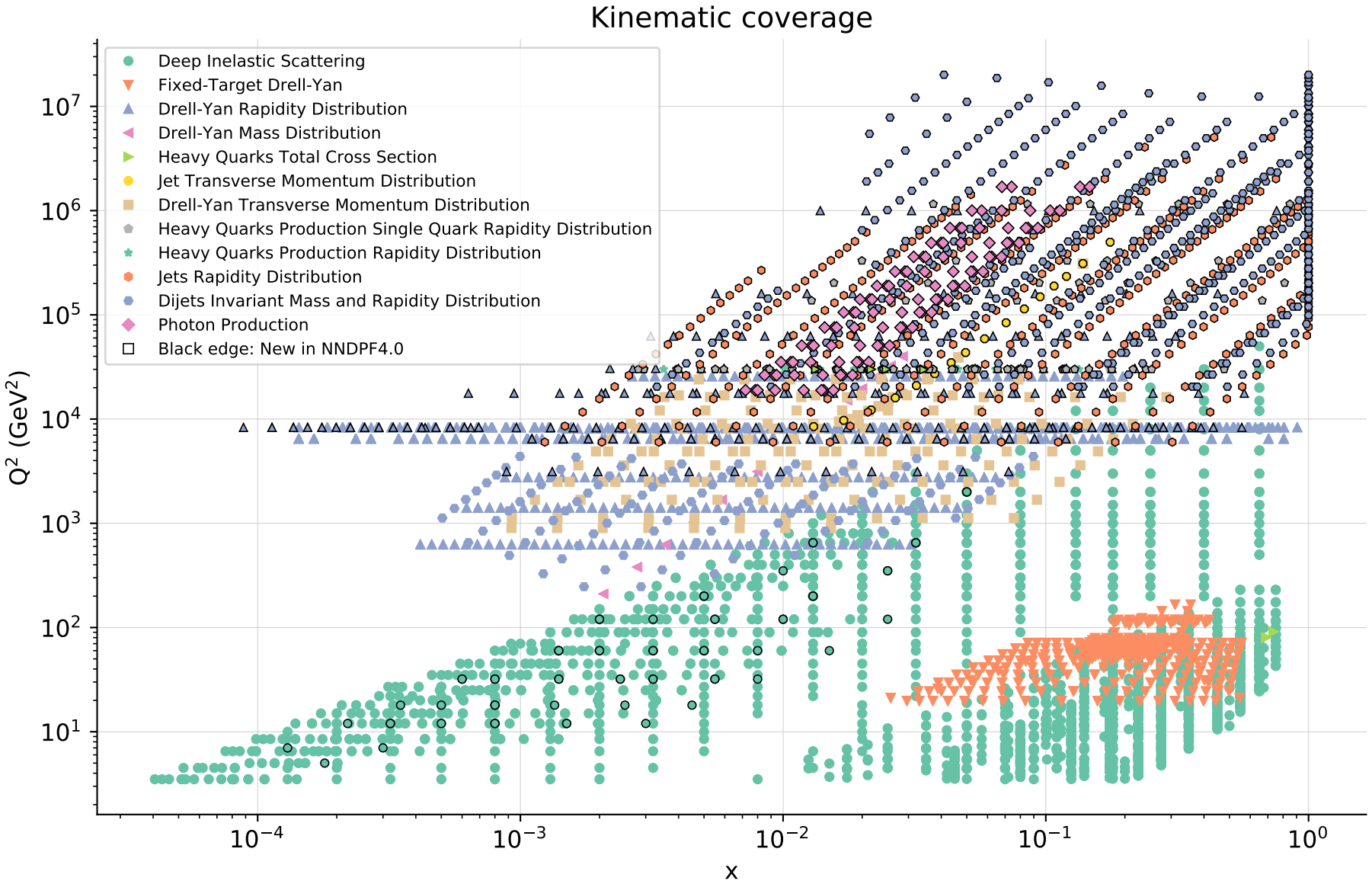




[EPJ C84 (2024) 369]

2.1 Theory

Can we improve the fit quality by improving the theory?



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

2.1.1 Heavy Quarks

Heavy Quarks in DIS

Two possible factorisation schemes for DIS structure functions

$\overline{\text{MS}}$ scheme

Heavy quarks are treated as massless (zero-mass scheme)

corrections proportional to $\ln(Q^2/m_h^2)$ are resummed to all orders by DGLAP

corrections that are $\mathcal{O}(m_h^2/Q^2)$ are neglected

This scheme is appropriate when $Q^2 \gg m_h^2$

Decoupling scheme

Heavy quarks are treated as massive (massive scheme)

corrections proportional to $\ln(Q^2/m_h^2)$ are treated at fixed order

corrections that are $\mathcal{O}(m_h^2/Q^2)$ are included

This scheme is appropriate when $Q^2 \sim m_h^2$

The third way: match the two schemes

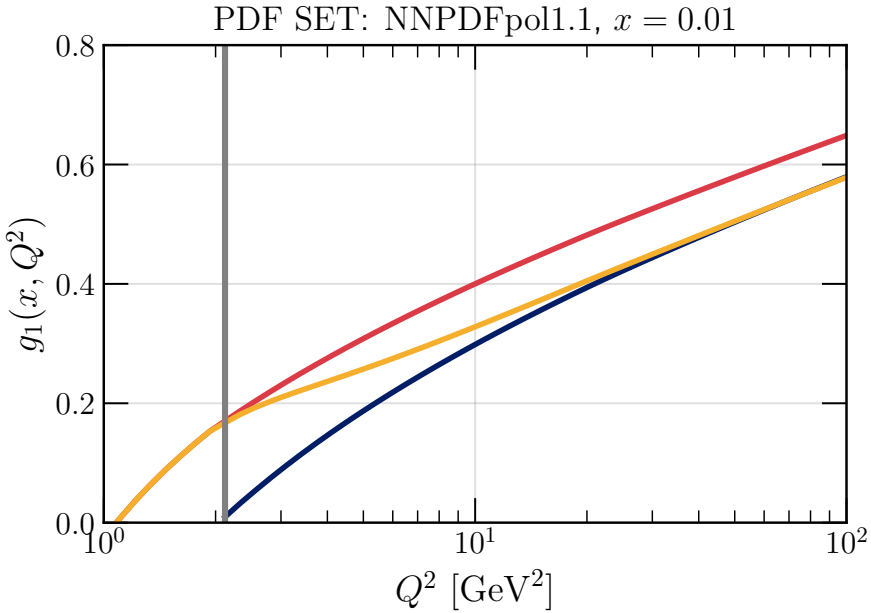
General-mass variable-flavour number schemes (ACOT, S-ACOT, TR, FONLL, ...)

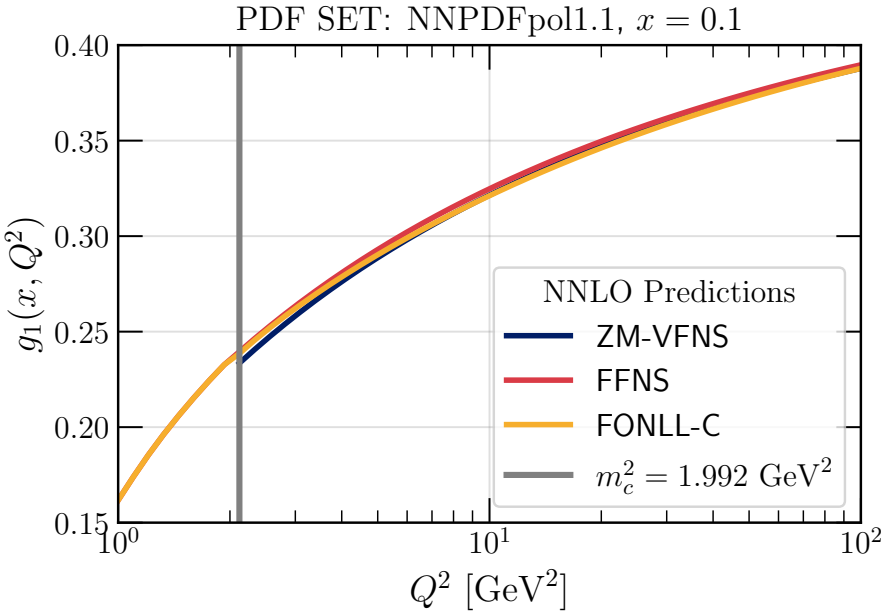
use $\overline{\text{MS}}$ for $Q^2 \gg m_h^2$ with full mass dependence retained

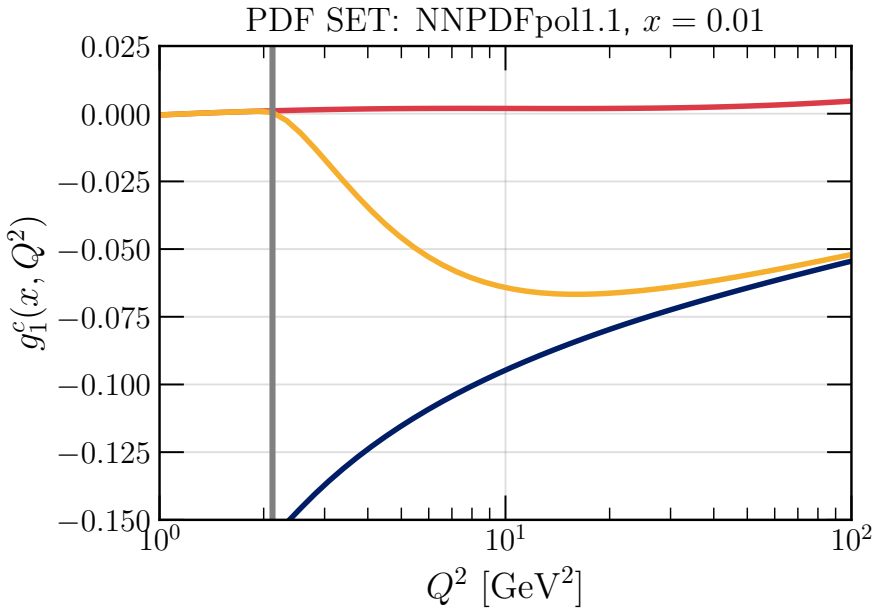
keep all flavour sin running DGLAP

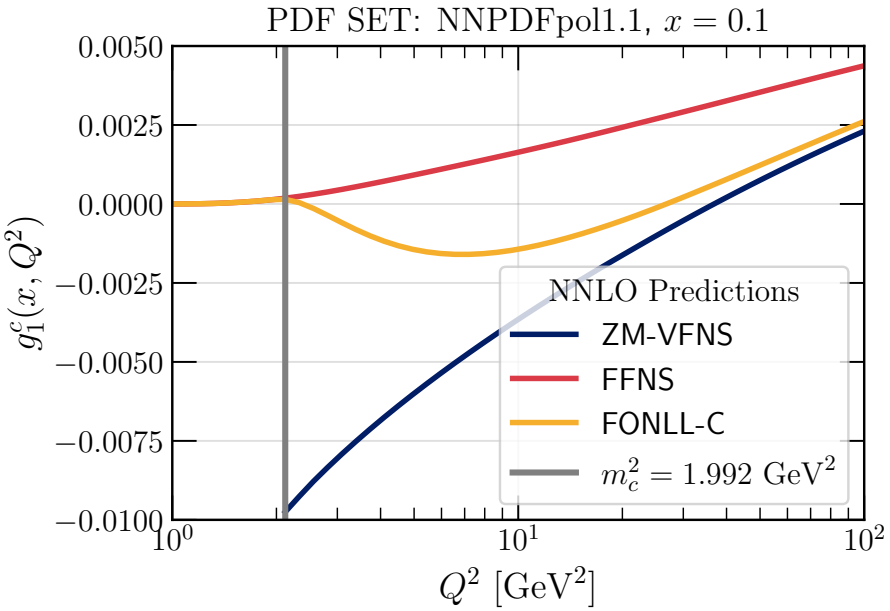
subtract double counting terms

Example: the g_1 and g_1^c structure functions at the EIC









2.1.2 Missing Higher Order Uncertainties

Perturbative Accuracy in PDF Determination

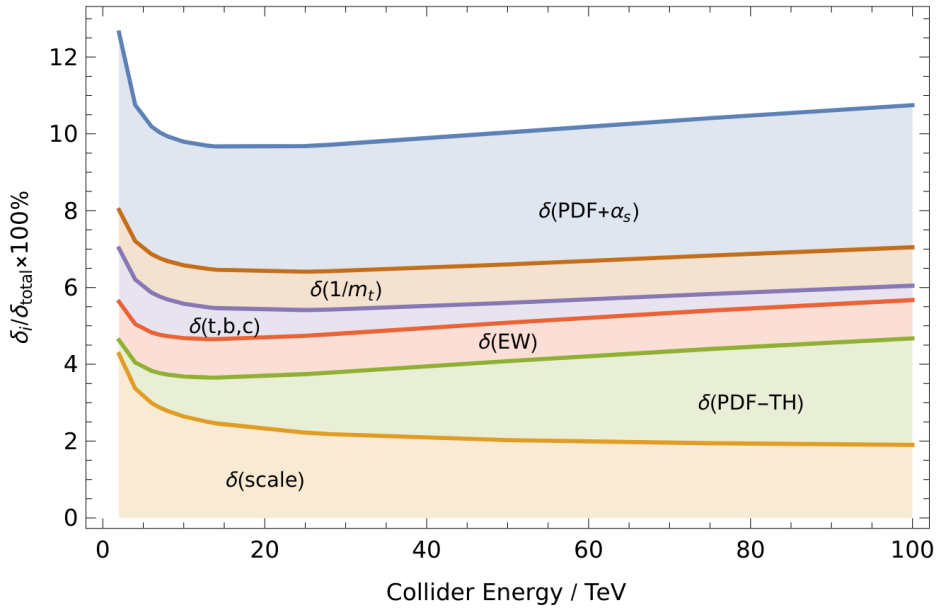
NNLO is the precision frontier for PDF determination

N3LO is the precision frontier for partonic cross sections

Mismatch between perturbative order of partonic cross sections and accuracy of PDFs may become a significant source of uncertainty

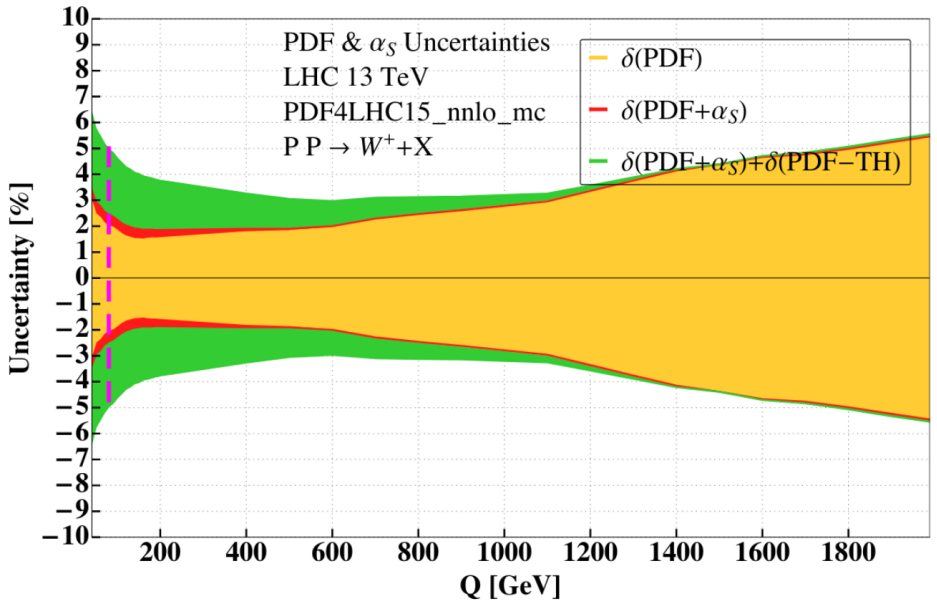
$$\hat{\sigma} = \alpha_s^p \hat{\sigma}_0 + \alpha_s^{p+1} \hat{\sigma}_1 + \alpha_s^{p+2} \hat{\sigma}_2 + \mathcal{O}(\alpha_s^{p+3}) \qquad \delta(\text{PDF} - \text{TH}) = \frac{1}{2} \left| \frac{\sigma^{(2)}_{\text{NNLO-PDFs}} - \sigma^{(2)}_{\text{NLO-PDFs}}}{\sigma^{(2)}_{\text{NNLO-PDFs}}} \right|$$

Higgs production in gluon-gluon fusion



[CERN Yellow Rep.Monogr. 7 (2019) 221]

W^+ boson production in CC Drell-Yan



[JHEP 11 (2020) 143]

MHOUs and Scale Variations

As an example, let us consider the NS DIS structure function

$$F_2^{\rm NS}(N,Q^2) = xC_{\rm NS}(N,\alpha_s(Q^2))\exp\left[\int_{Q_0^2}^{Q^2}\frac{d\lambda^2}{\lambda}\gamma_{\rm NS}\left(N,\alpha_s(\mu^2)\right)\right]f_{\rm NS}(Q_0^2)$$

Sources of MHOUs

$$\gamma_{\rm NS}^{\rm N^kLO}(N,\alpha_s) = \alpha_s\gamma_{\rm NS}^{(0)} + \alpha_s^2\gamma_{\rm NS}^{(1)} + \ldots \alpha_s^{k+1}\gamma_{\rm NS}^{(k)}$$

$$C_{\rm NS}^{\rm N^kLO}(N,\alpha_s) = 1 + \alpha_s C_{\rm NS}^{(1)} + \ldots \alpha_s^{k}C_{\rm NS}^{(k)}$$

Scale variations

Idea: $\alpha_s(\kappa^2\mu^2) = \alpha_s(\mu^2)[1+\mathcal{O}(\alpha_s)]$

at N$^{\rm k}$LO differences due to higher orders are related to the QCD β function up to β_k

$\bar{C}_{\rm NS}(\alpha_s(\kappa_r^2\mu^2,\kappa_r^2)) = C_{\rm NS}(\alpha_s(\mu^2))[1+\mathcal{O}(\alpha_s)]$ fixes $\bar{C}^{(k)}$ in terms of $C^{(k)}$

$\bar{\gamma}_{\rm NS}(\alpha_s(\kappa_f^2\mu^2,\kappa_f^2)) = \gamma_{\rm NS}(\alpha_s(\mu^2))[1+\mathcal{O}(\alpha_s)]$ fixes $\bar{\gamma}^{(k)}$ in terms of $\gamma^{(k)}$

$$\Delta C_{\rm NS} = \bar{C}_{\rm NS}(\alpha_s(\kappa_r^2\mu^2,\kappa_r^2)) - C_{\rm NS}(\alpha_s(\mu^2))$$

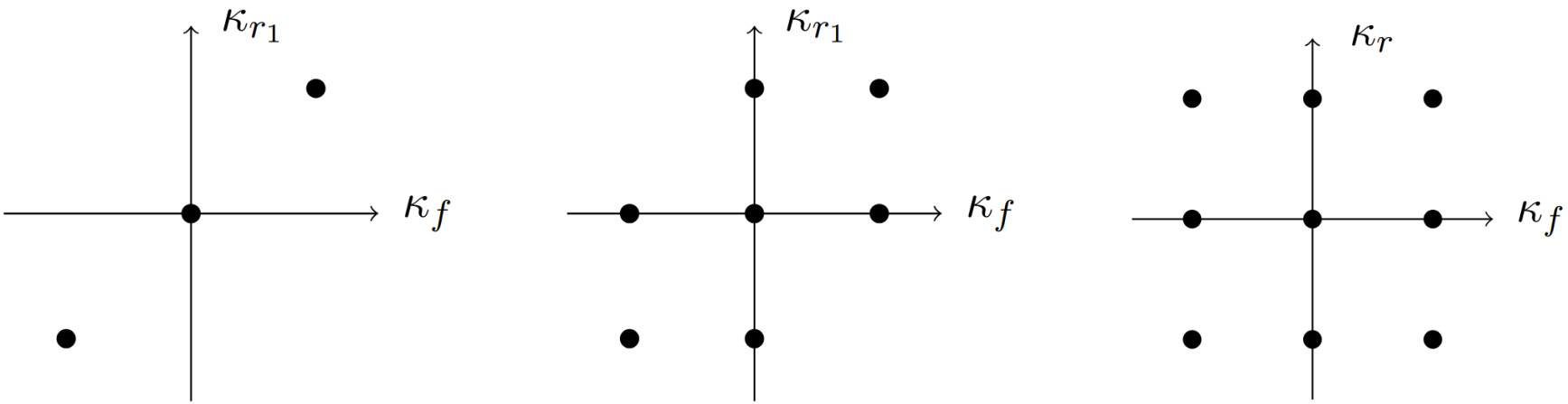
renormalisation scale (at which UV divergences are subtracted) $\mu_r = \kappa_r\mu$

$$\Delta\gamma_{\rm NS} = \bar{\gamma}_{\rm NS}(\alpha_s(\kappa_f^2\mu^2,\kappa_f^2)) - \gamma_{\rm NS}(\alpha_s(\mu^2))$$

factorisation scale (at which collinear divergences are factorised) $\mu_f = \kappa_f\mu$

Propagate ΔC and $\Delta\gamma$ into Δf

Scale Variations: Prescriptions



Vary μ_r and μ_f about μ_0

Pick a set of possible variations

3-points: $\mu_r = \mu_f$, $\kappa_{r,f} = 2, 1/2$

7-points: μ_r, μ_f varied independently, $\kappa_{r,f} = 2, 1/2$, remove $\mu_r/\mu_f = 4$

9-points: μ_r, μ_f varied independently, $\kappa_{r,f} = 2, 1/2$

To estimate MHOUs, take the envelope,
i.e. the difference between the largest and smallest predictions

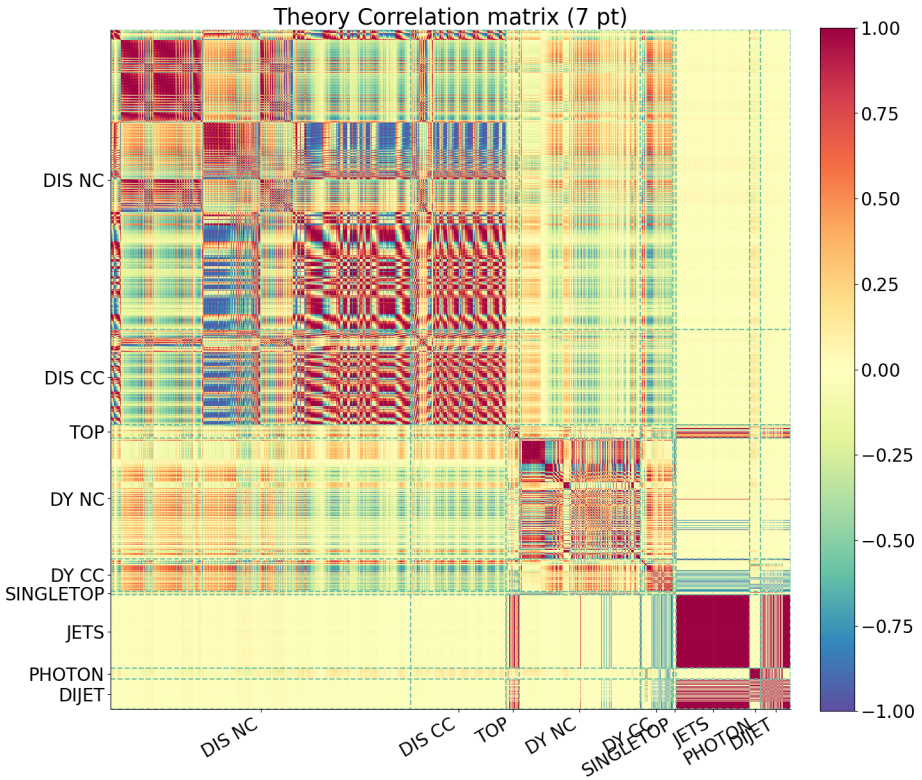
A Theory Covariance Matrix

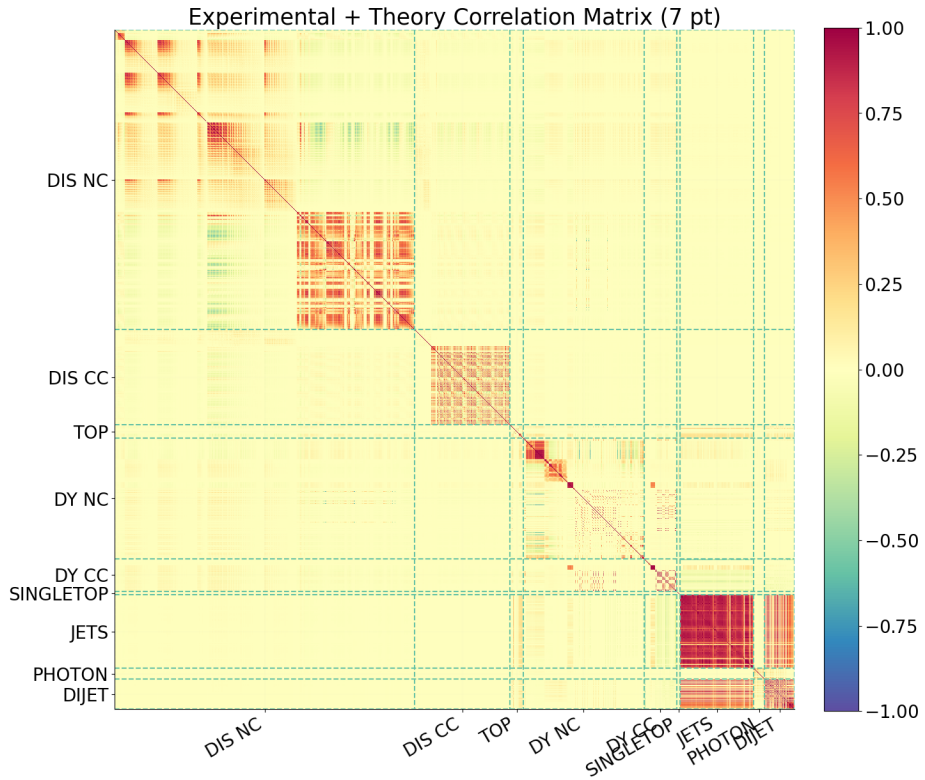
Assuming that theory uncertainties are (a) Gaussian and (b) independent from experimental uncertainties, modify the figure of merit to account for theory errors

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

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

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





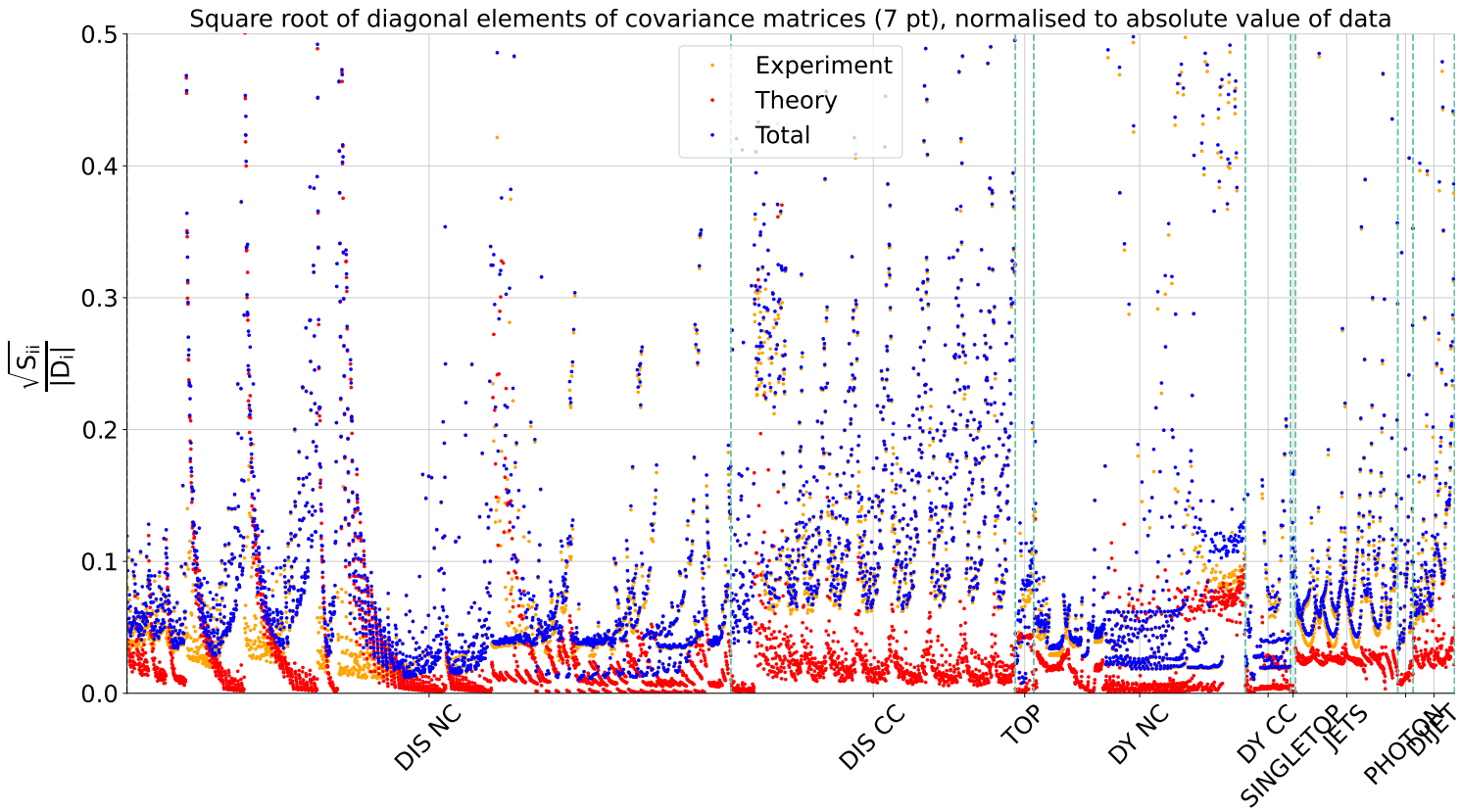
A Theory Covariance Matrix

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

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



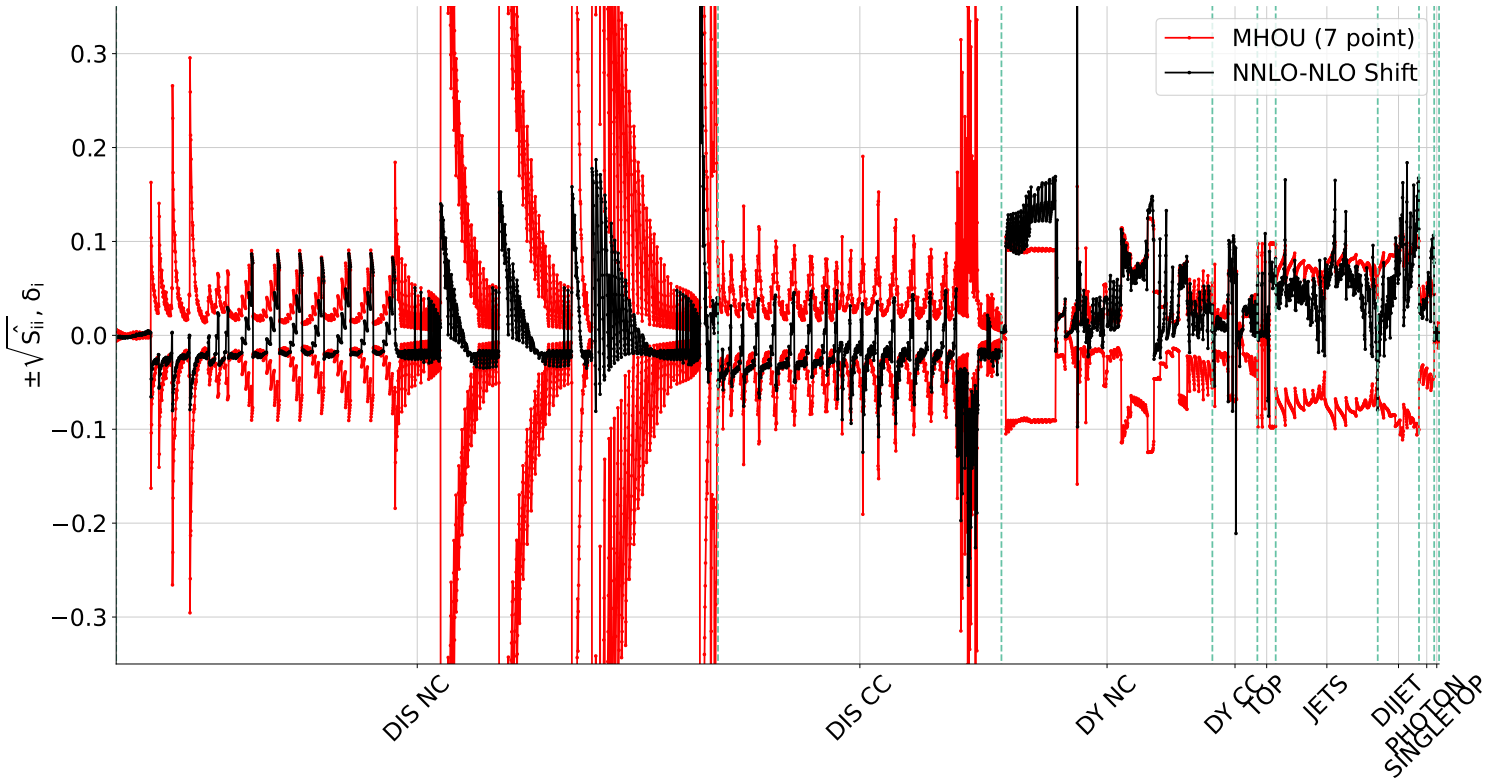
A Theory Covariance Matrix

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

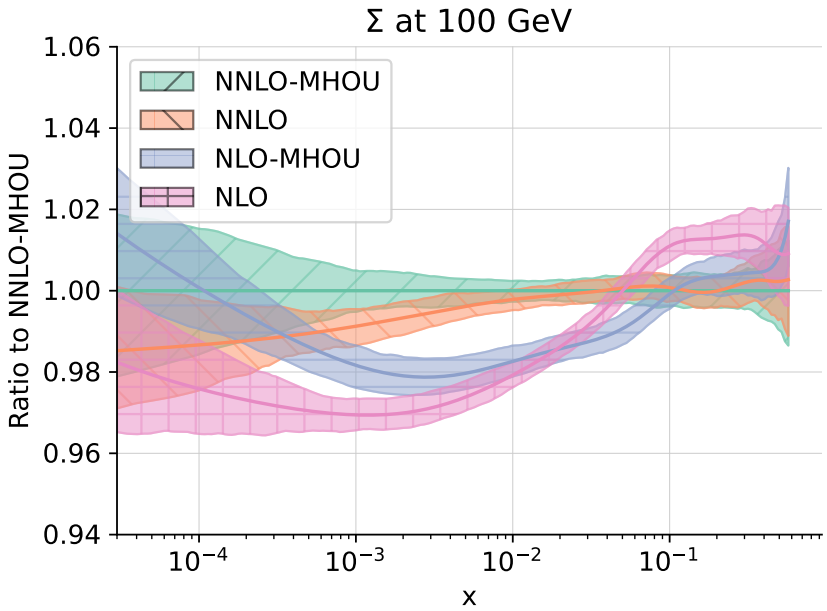
$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} (D_i - T_i)(\mathrm{cov}_{\rm exp} + \mathrm{cov}_{\rm th})^{-1}_{ij}(D_j - T_j);\ (\mathrm{cov}_{\rm th})_{ij} = \frac{1}{N}\sum_k^N \Delta_i^{(k)}\Delta_j^{(k)};\ \Delta_i^{(k)} \equiv T_i^{(k)} - T_i$$

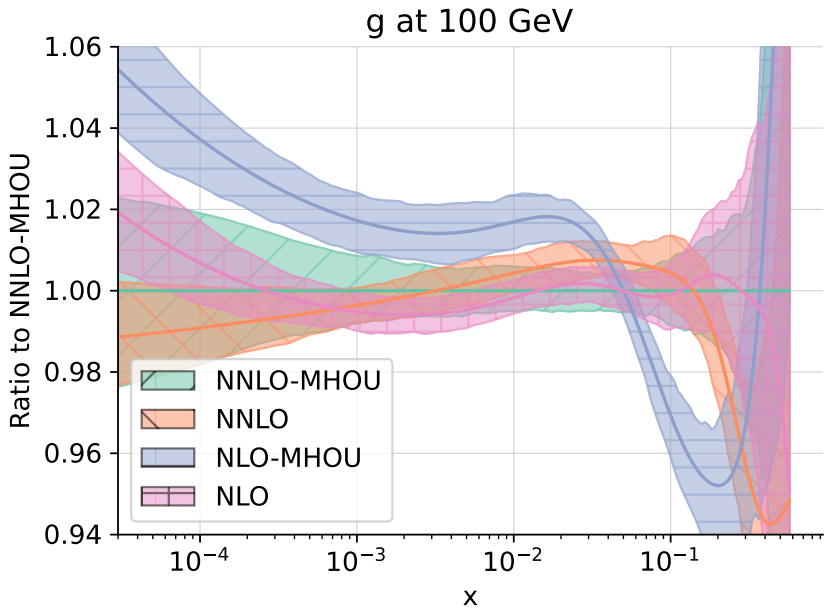
Problem reduced to estimate the th. cov. matrix, *e.g.* in terms of nuisance parameters

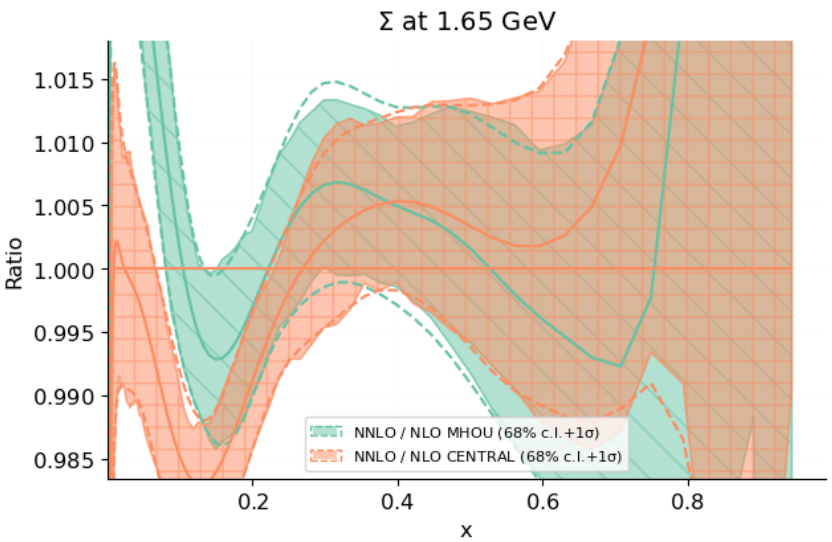
$$\Delta_i^{(k)} = T_i(\mu_R, \mu_F) - T_i(\mu_{R,0}, \mu_{F,0});\ \text{vary scales in } \frac{1}{2} \leq \frac{\mu_F}{\mu_{F,0}}, \frac{\mu_R}{\mu_{R,0}} \leq 2$$

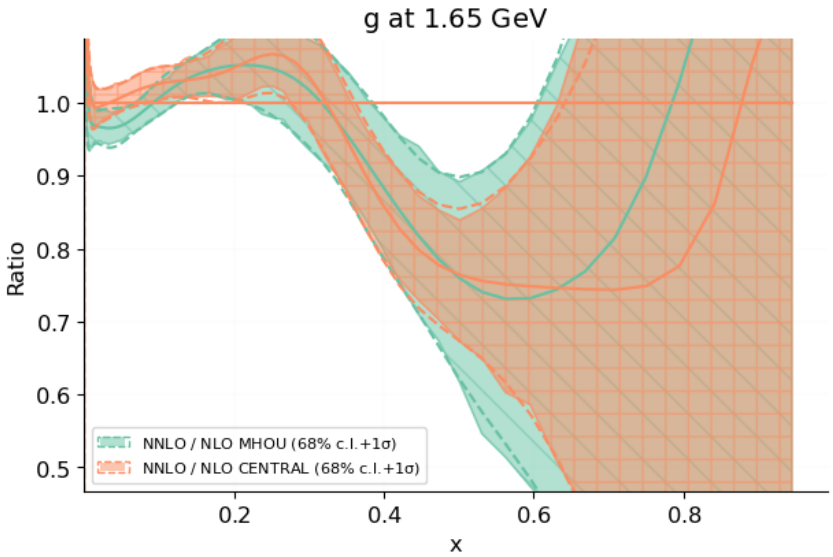


Impact on Parton Distributions





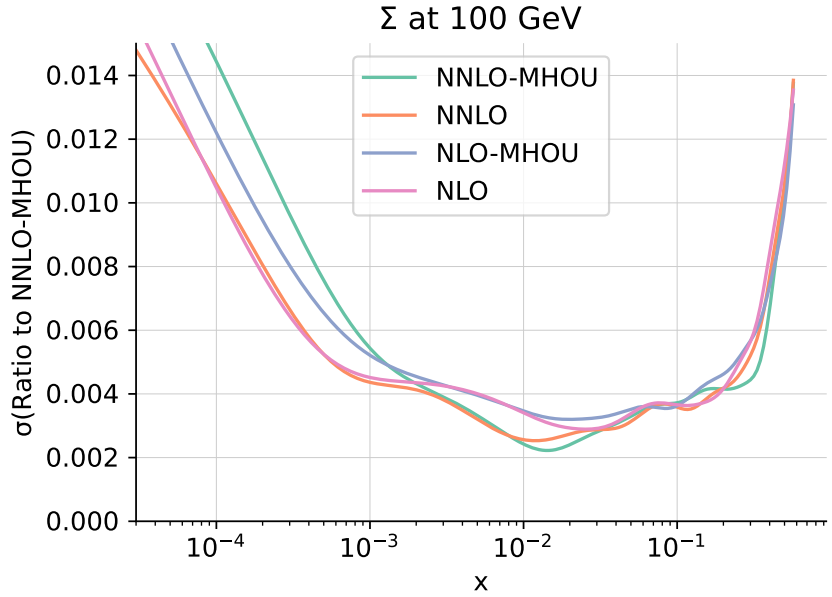


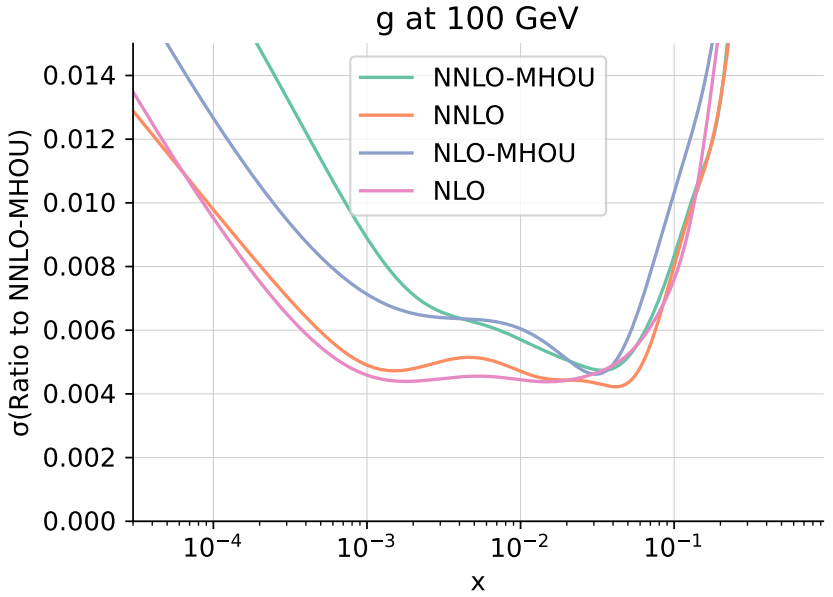


Faster perturbative convergence when MHOU are incorporated into PDFs

[EPJ C79 (2019) 838; ibid. 931; EPJ C84 (2024) 517]

Impact on Uncertainties and Fit Quality





Dataset	$N_{\rm dat}$	NLO		NNLO	
		no MHOU	MHOU	no MHOU	MHOU
DIS NC	2100	1.30	1.22	1.23	1.20
DIS CC	989	0.92	0.87	0.90	0.90
DY NC	736	2.01	1.71	1.20	1.15
DY CC	157	1.48	1.42	1.48	1.37
Top pairs	64	2.08	1.24	1.21	1.43
Single-inclusive jets	356	0.84	0.82	0.96	0.81
Dijets	144	1.52	1.84	2.04	1.71
Prompt photons	53	0.59	0.49	0.75	0.67
Single top	17	0.36	0.35	0.36	0.38
Total	4616	1.34	1.23	1.17	1.13

Overall (rather small) variation of uncertainties. Tensions relieved: improvement in χ^2

[EPJ C79 (2019) 838; ibid. 931; EPJ C84 (2024) 517]

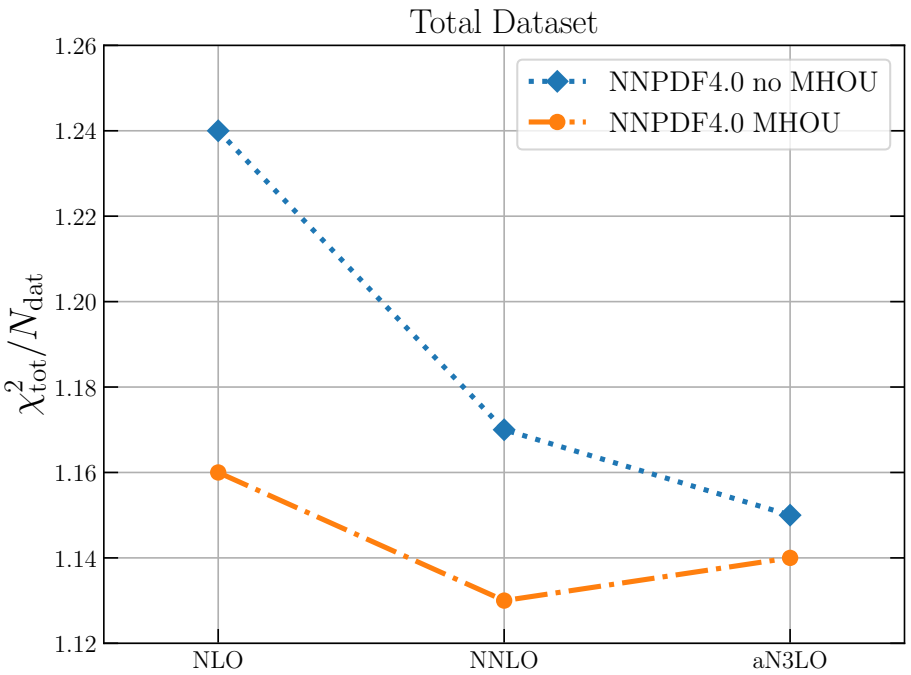
What Happens at aN^{3}LO?

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

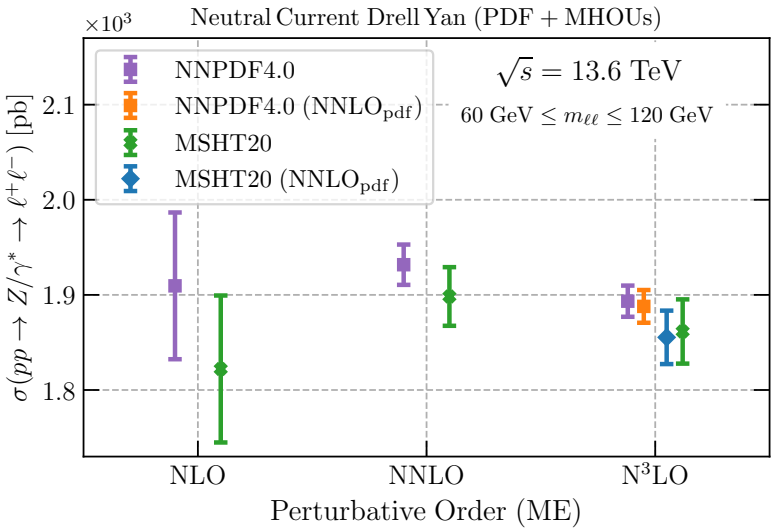
Fit quality improves with perturbative order

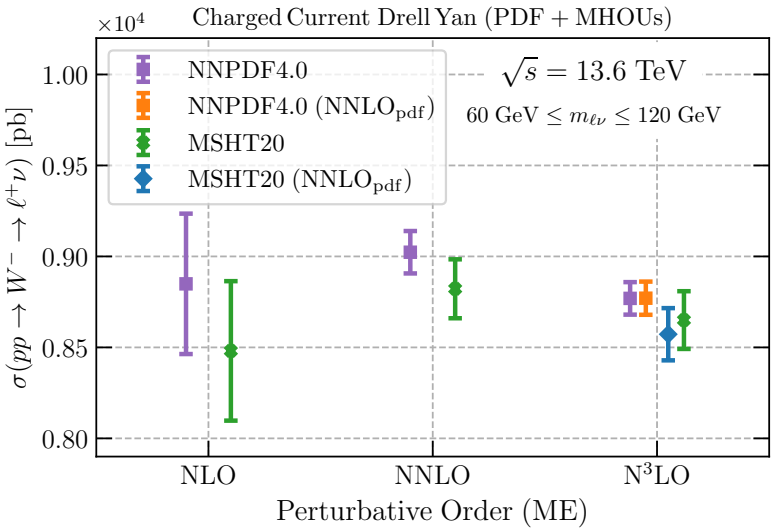
Fit quality almost independent from perturbative order when MHOU are included

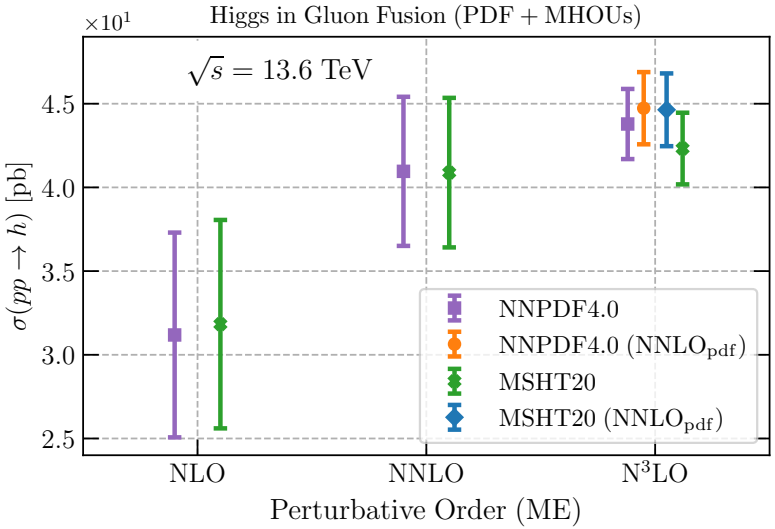
Data whose theoretical description is affected by large scale uncertainties are deweighted in favour of more perturbatively stable data

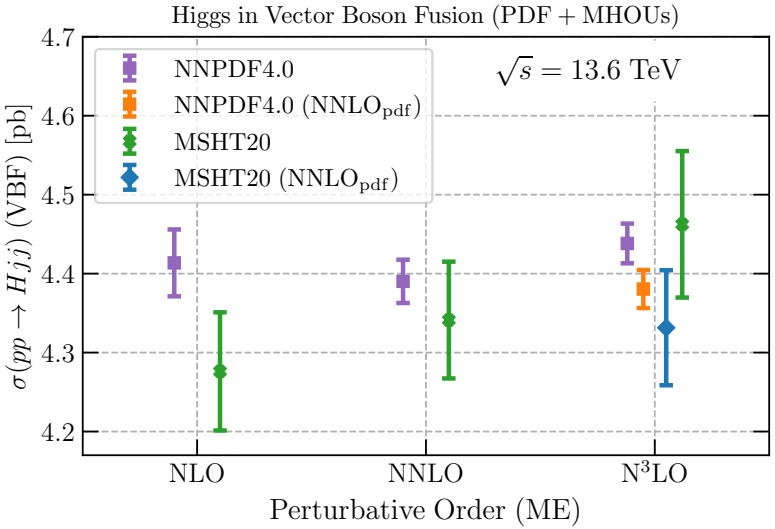


Impact on Inclusive Cross Sections







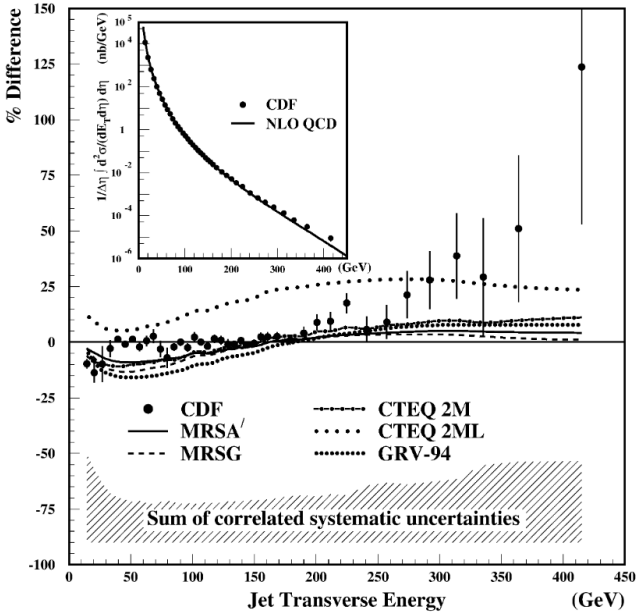


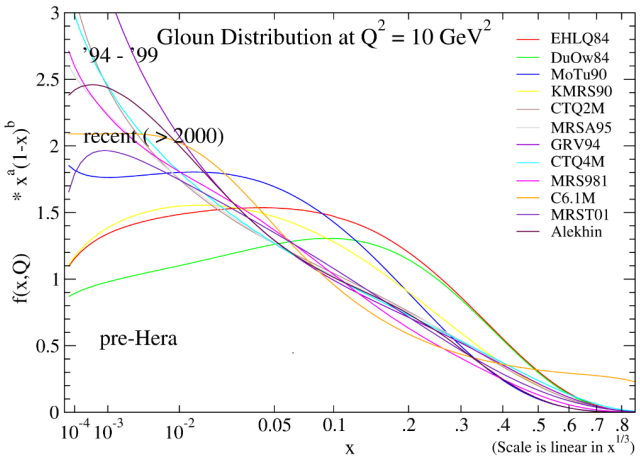
Effect of using aN^{3}LO PDFs instead of NNLO PDFs in N^{3}LO predictions is small

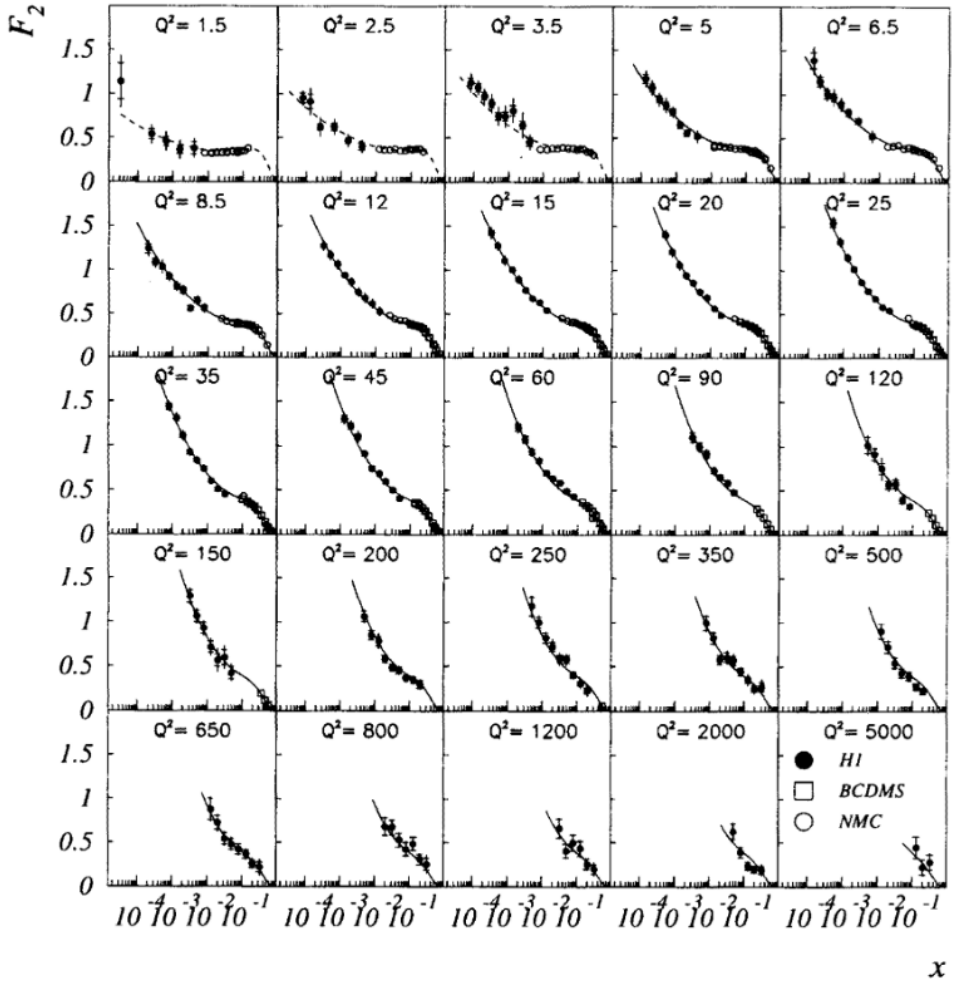
Good consistency between NNPDF4.0 [EPJ C84 (2024) 659] and MSHT20 [EPJ C83 (2023) 185]

2.2 Methodology

Why is the methodology important?



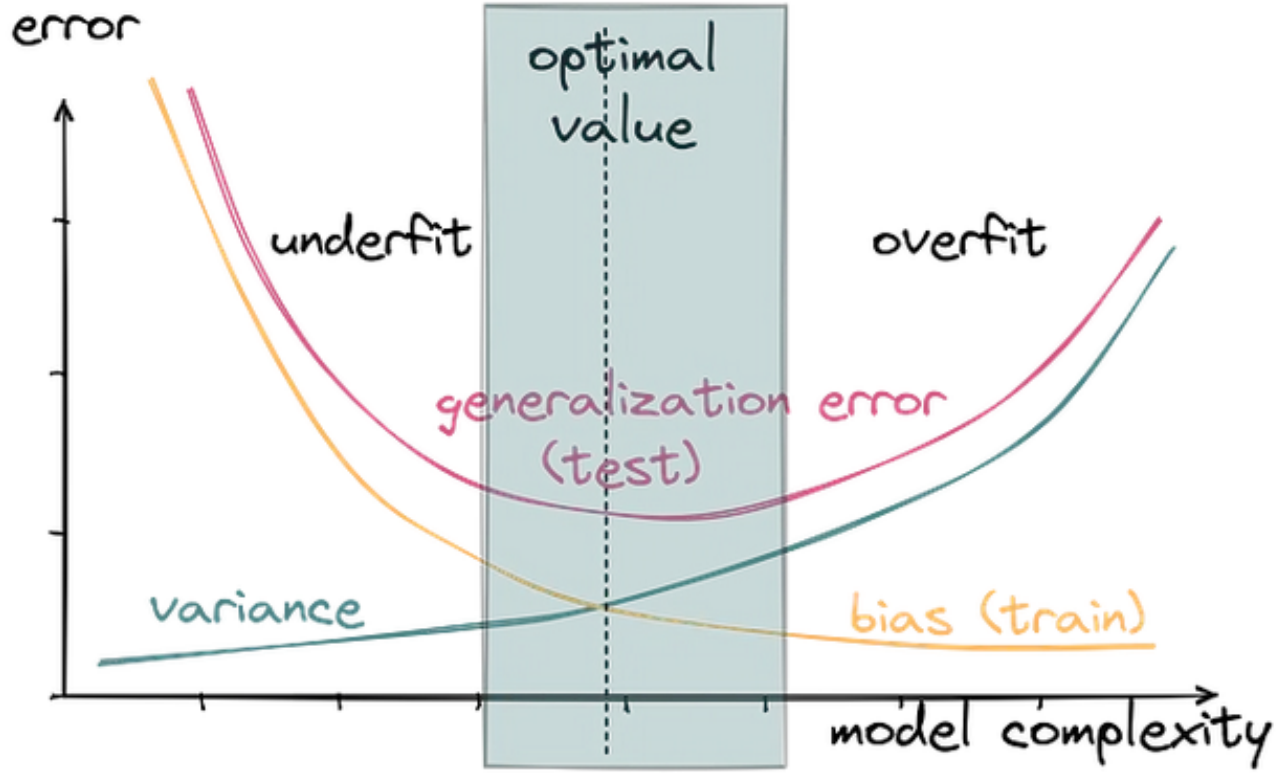


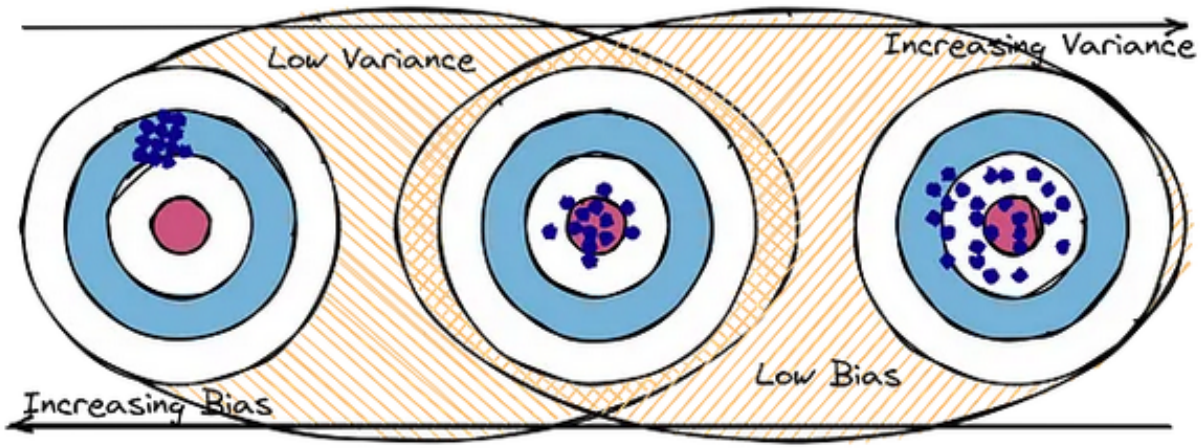


circa 1995: small-x rise of HERA F_2^p and CDF jet discrepancy

The methodology is crucial if we aim at percent-level accurate PDFs

Accuracy vs precision or bias vs variance





What are the ingredients of a fitting methodology?

parametrisation

polynomials/neural network(s)

is there a bias due to the parametrisation?

optimisation

(adaptive) gradient descent

is the parameter space explored efficiently?

uncertainty representation

Hessian/bootstrap of experimental uncertainties

what is the statistical meaning of uncertainties?

validation

closure tests (what happens if I know in advance the underlying law that I am fitting?)

are interpolation and extrapolation uncertainties statistically faithful?

benchmark

PDF4LHC working group

are PDFs obtained independently by various groups equivalent?

2.2.1 Parametrisation

Parametrisation: general features

Problem projected onto the finite-dimensional space of parameters

Choose a parametrisation at an initial scale Q_0^2
for each independent parton i (or a combination of them)

$$\boxed{x f_i(z, Q_0^2) = A_i \, x^{a_i} \, (1-x)^{b_i} \, \mathscr{F}_i(x, \{c_{f_i}\})}$$

small x

$$x f_i(x, Q_0^2) \xrightarrow{x \to 0} x^{a_i}$$

$$\xrightarrow[\text{interpolation in between}]{\mathscr{F}_i(x, \{c_{f_i}\}) \xrightarrow[x \to 1]{x \to 0} \text{finite}}$$

large x

$$x f_i(x, Q_0^2) \xrightarrow{x \to 1} (1-x)^{b_i}$$

The problem is reduced to the determination of the finite set of parameters $\{c_{f_i}\}$

The interpolating function $\mathscr{F}_i(x, \{c_{f_i}\})$ should be sufficiently

GENERAL (the range of PDF behaviours in the space of functions should not be limited)

SMOOTH (PDFs are implicitly assumed ot be smooth functions)

FLEXIBLE (it should be able to adapt to a variety of data and processes)

to describe the data with minimal bias

Parametrisation: two alternative choices

1. Polynomial (Bernstein, Chebyschev) parametrisation, *e.g.*

$$\mathscr{F}_i = 1 + \sum_{i=1}^{n} a_i T_i^{\rm Ch}\left(y(x)\right) \qquad y = 1 - 2\sqrt{x}$$

in terms of a (relatively) small set of parameters ($\mathcal{O}(30)$ per PDF set)

$$\{\mathbf{a}\} = \{a_i, b_i, \gamma_i, \delta_i\}$$

$\Rightarrow$ **smooth behavior (a desirable feature for a PDF)**
$\Rightarrow$ **potential source of bias if the parametrisation is too rigid**

2. Redundant parametrisation, *e.g.*

a neural network

in terms of a huge set of parameters ($\mathcal{O}(200)$ per PDF set)

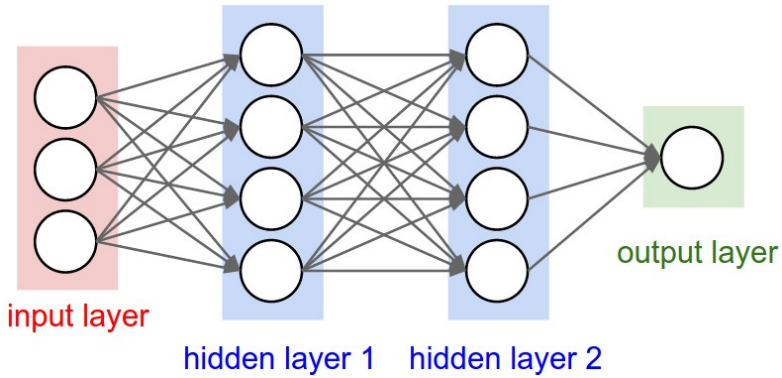
$$\{\mathbf{a}\} = \{\omega_{ij}^{(L-1),f_i}, \theta_i^{(L),f_i}\}$$

$\Rightarrow$ **potentially non-smooth**
$\Rightarrow$ **bias due to the parametrisation reduced as much as possible**

Parametrisation: what a neural network exactly is?

A convenient functional form providing a flexible parametrization
used as a generator of random functions in the PDF space

EXAMPLE: MULTY-LAYER FEED-FORWARD PERCEPTRON



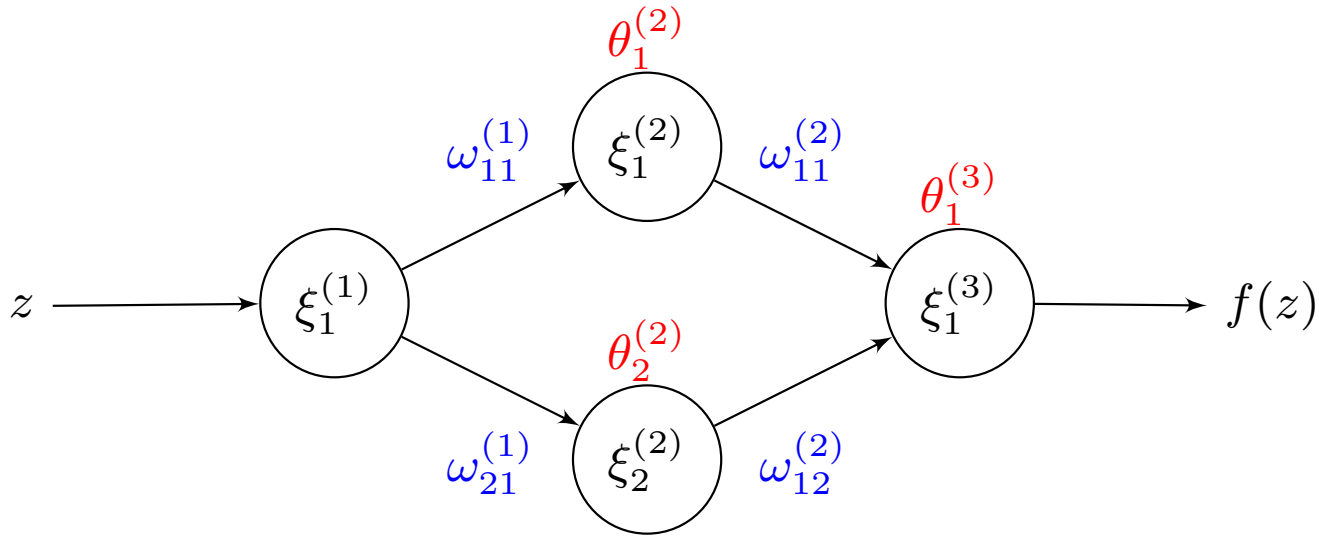
$$\xi_i^{(l)} = g\left(\sum_j^{n_l-1} \omega_{ij}^{(l-1)} \xi_j^{(l-1)} - \theta_i^{(l)}\right)$$

$$g(y) = \frac{1}{1+e^{-y}}$$

- made of neurons grouped into layers (define the architecture)
- each neuron receives input from neurons in the preceding layer (feed-forward NN)
- activation $\xi_i^{(l)}$ determined by a set of parameters (weights and thresholds)
- activation determined according to a non-linear function (except the last layer)

Parametrisation: what a neural network exactly is?

EXAMPLE: THE SIMPLEST 1-2-1 MULTI-LAYER FEED-FORWARD PERCEPTRON

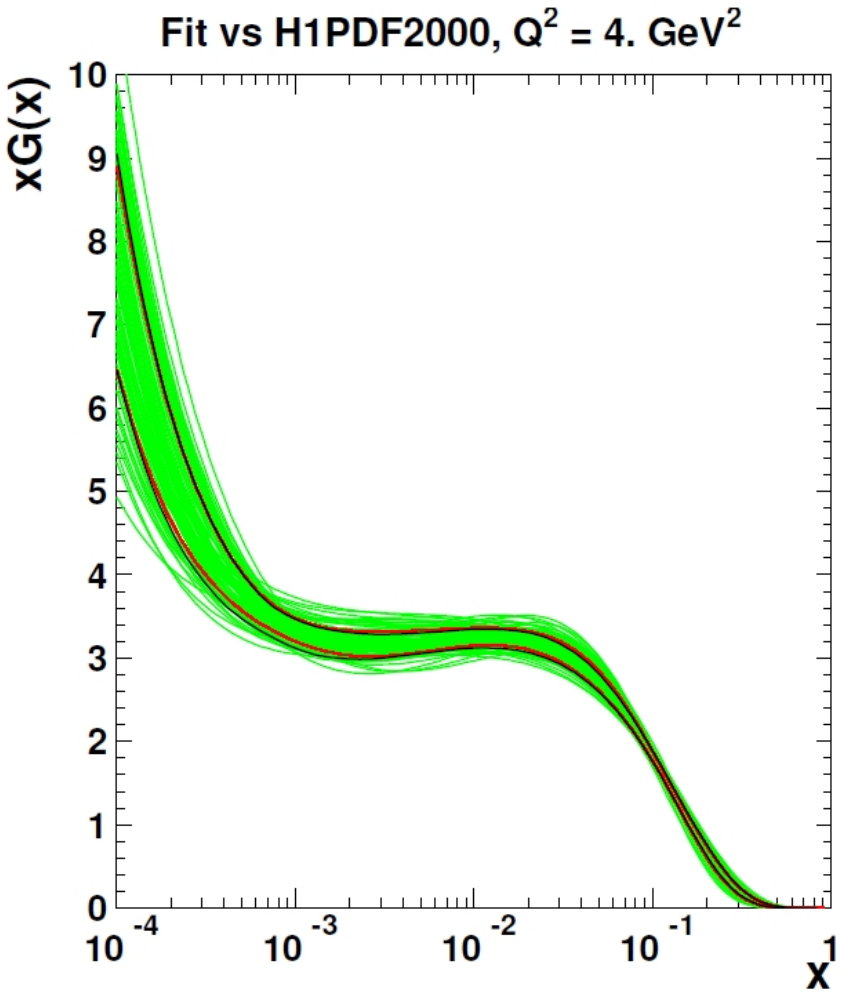


$$f(z) \equiv \xi_1^{(3)} = \left\{ 1 + \exp\left[\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1 + e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1 + e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}} \right] \right\}^{-1}$$

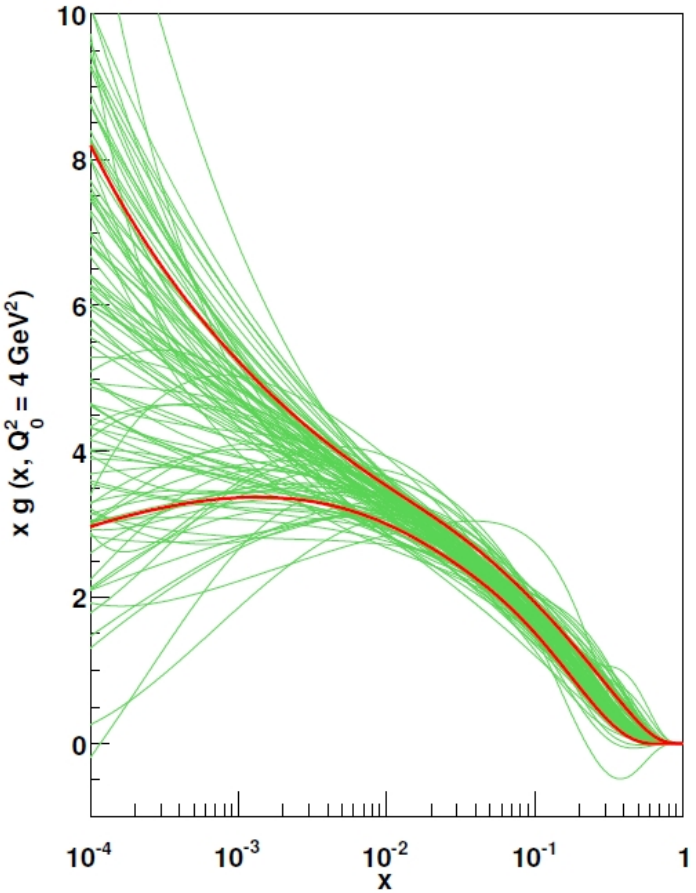
Recall: $\xi_i^{(l)} = g\left(\sum_j^{n_l - 1} \omega_{ij}^{(l-1)} \xi_j^{(l-1)} - \theta_i^{(l)} \right)$; $g(z) = \frac{1}{1 + e^{-z}}$

Parametrisation: standard *vs* redundant

HERA-LHC 2009 PDF benchmark



simple parametrization



redundant parametrization (NN)

2.2.2 Optimisation

Fit quality

1. Define the fit quality (the χ^2 function)

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

with the experimental covariance matrix

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

s_i are N_{dat} uncorrelated uncertainties (statistic + uncorrealted systematic ucnertainties)
$\sigma_{i,\alpha}^{(c)}$ 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

2. Find the best-fit configuration of parameters $\{\mathbf{a_0}\}$ which minimise the χ^2
3. Treat conveniently
   - uncorrelated/correlated uncertainties
     need not to overestimate uncertainties and to let the χ^2 be faithful
   - additive/multiplicative uncertainties
     need to avoid the D'Agostini bias

Parameter optimisation: general framework

Optimisation usually performed by means of simple gradient descent:
compute and minimise the gradient of the fit quality with respect to the fit parameters

$$\frac{\partial \chi^2}{\partial a_i}\,, \qquad \text{for } i = 1, \ldots, N_{\rm par}$$

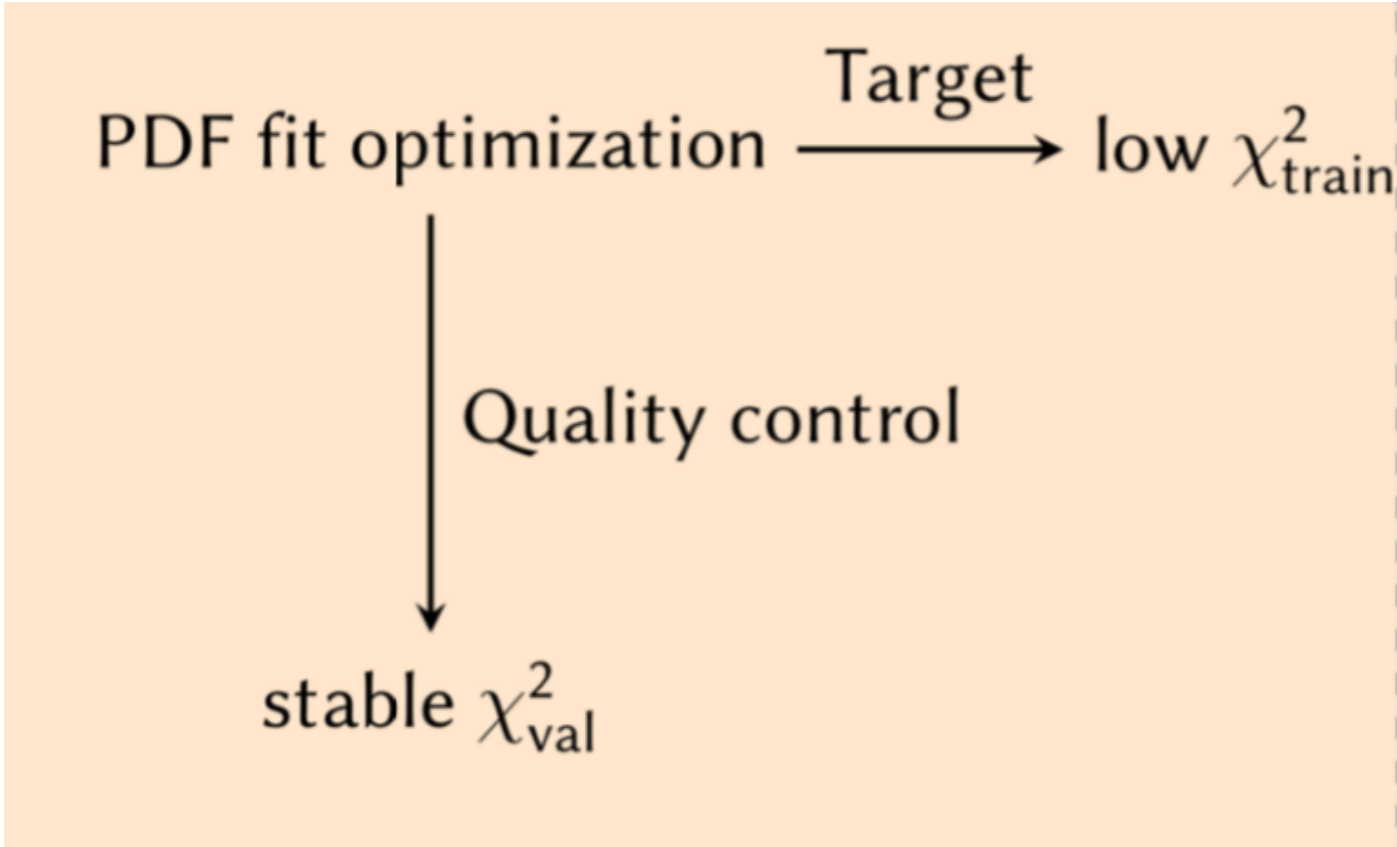
Optimisation should minimise the noise in the χ^2 driven by noisy experimental data

Additional complications in case of a redundant parametrisation (huge parameter space)

1. need to explore the parameter space as uniformly as possible
   (in order to avoid stopping the fit in a local minimum)
2. need for a computationally efficient minimisation
   (non-trivial relationship between FFs and observables via convolution)
3. need to define a criterion for minimisation stopping
   (avoid learning statistical fluctuations of the data)

Alternative algorithms:
genetic algorithms, adaptive algorithms, ...

Optimisation: training and validation



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

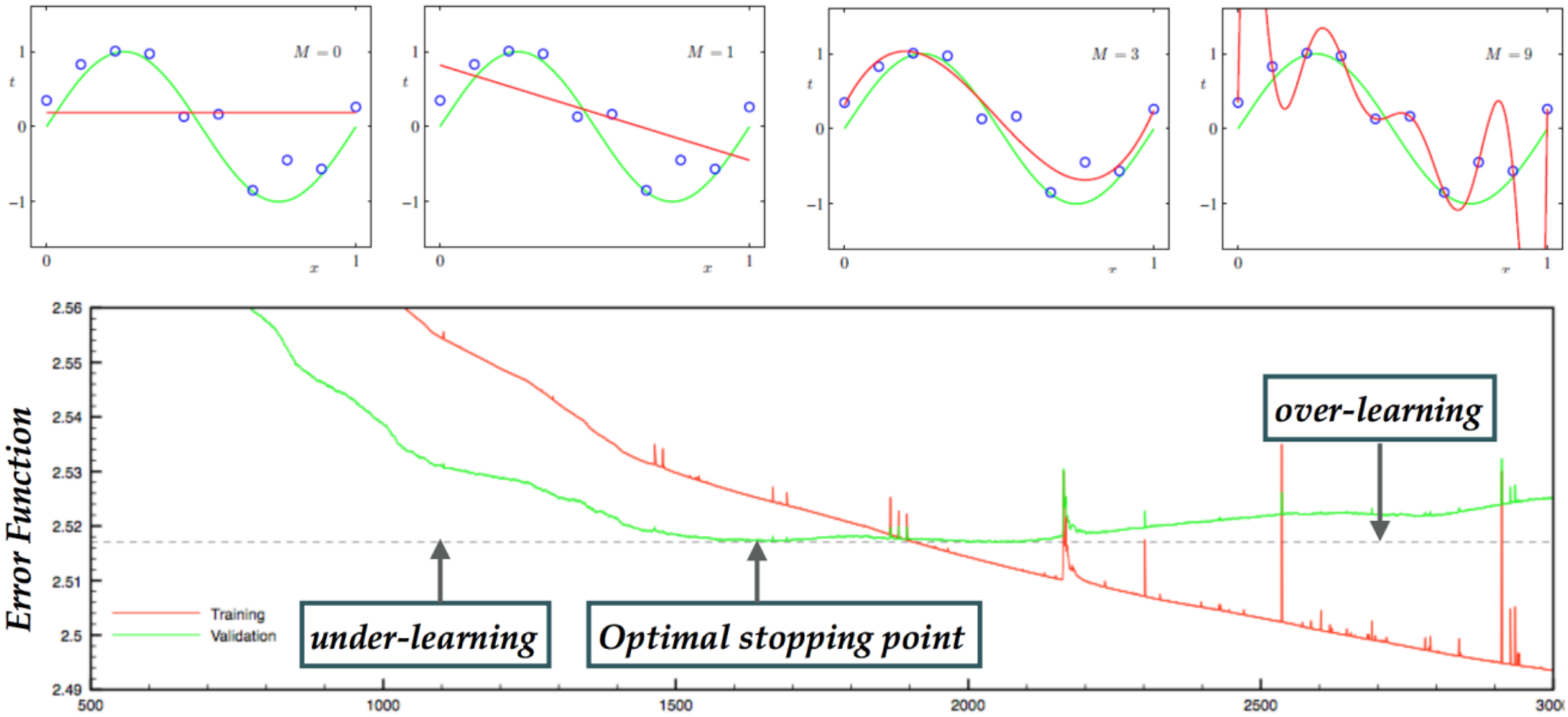
$$\left(\mathrm{cov}\right)_{ij} = \delta_{ij} s_i^2 + \left(\sum_{\alpha}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} + \sum_{\alpha}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \right) D_i D_j$$

Optimisation: stopping criterion

Divide the data into two subsets (training & validation)

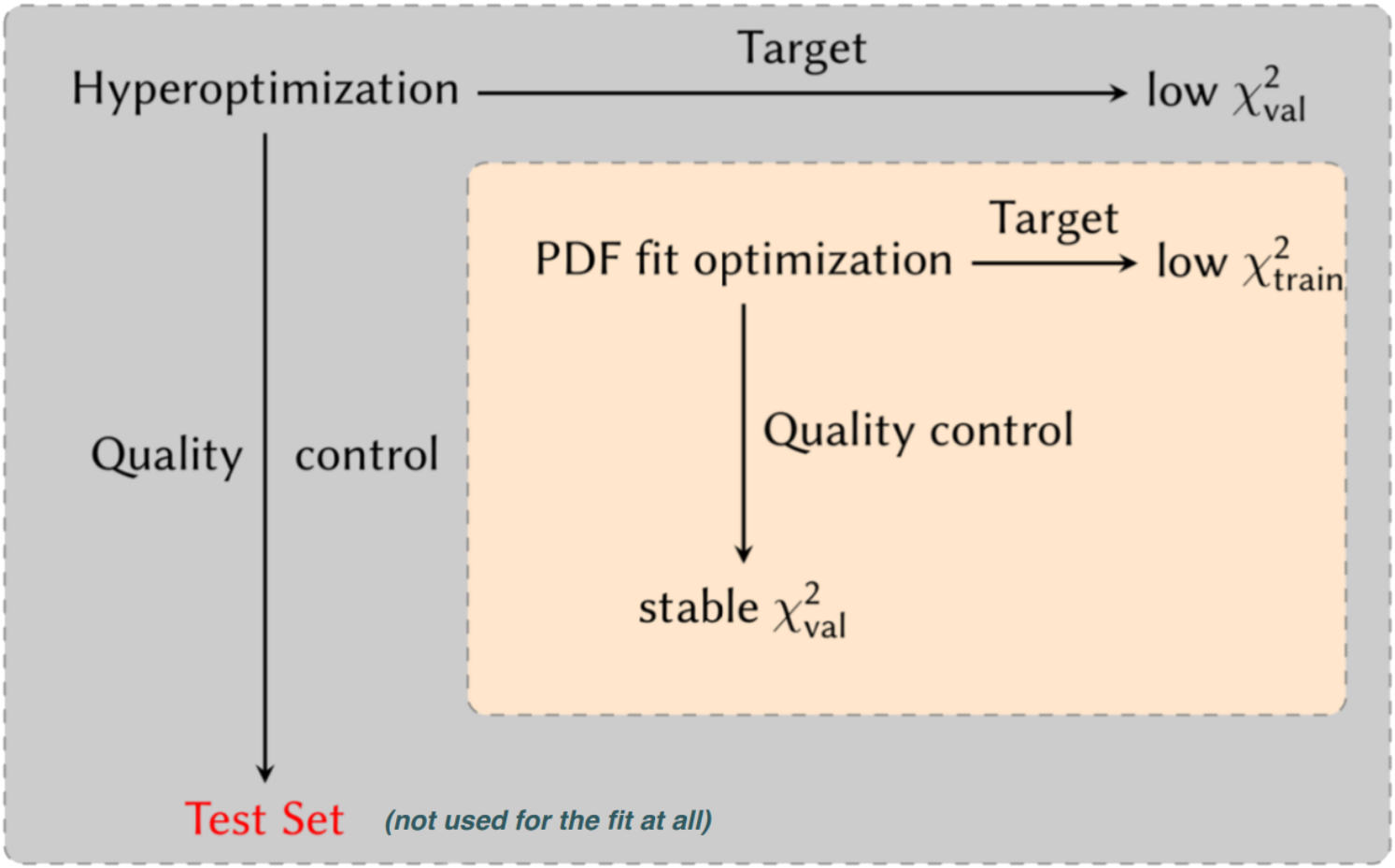
Train the NN on the training subset and compute χ^2 for each subset

Stop when the training loss reaches the absolute minimum



The best fit does not coincide with the absolute minimum of the χ^2

Hypertoprimisation: fitting the methodology



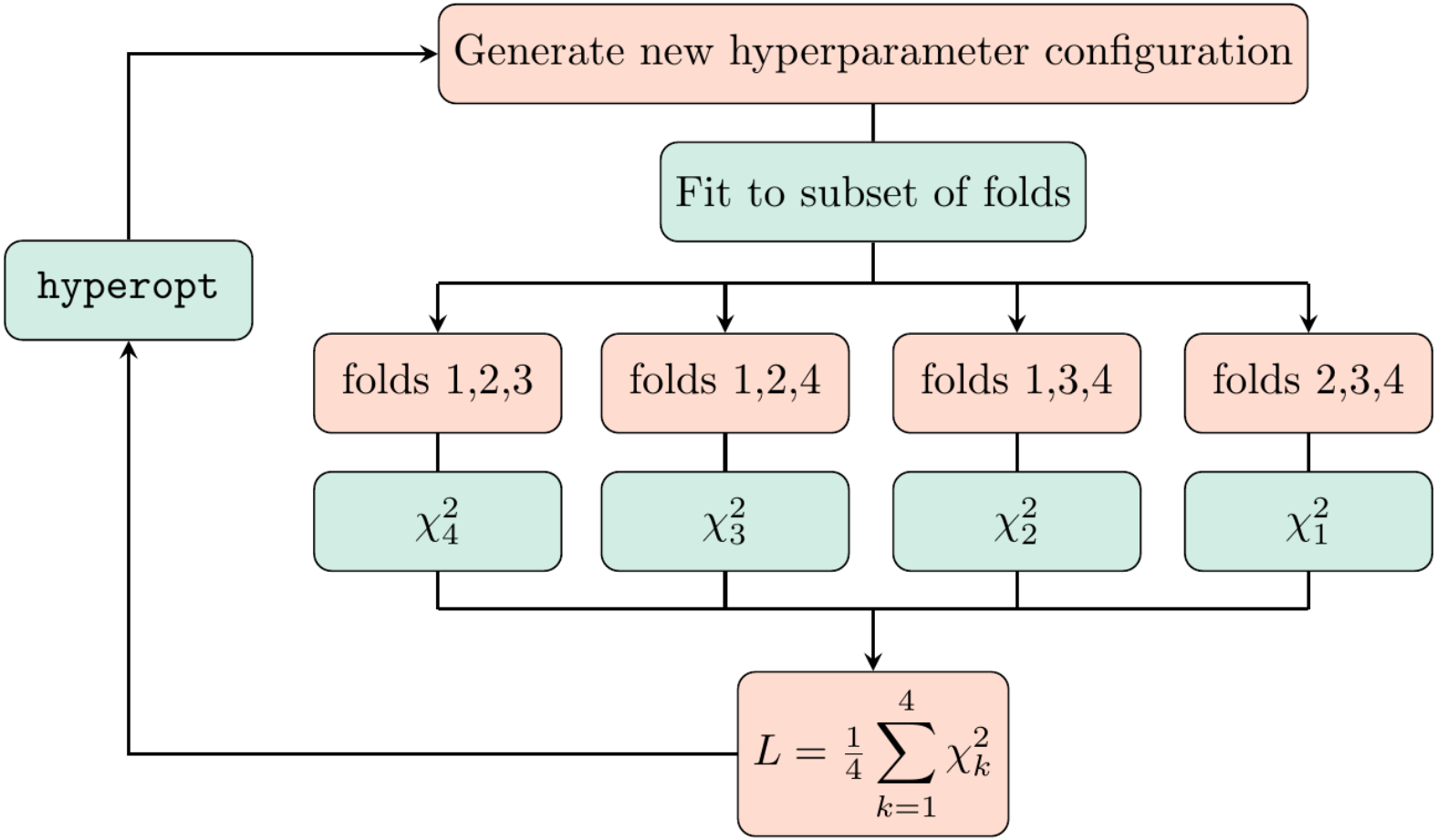
Compare to a Test Set (new set of data previously not used at all)

Who picks the Test Set? Automatic generalisation based on K foldings

Divide the data into n representative sets, fit $n-1$ sets and use the n-th set as test set

Hyperoptimise on mean and standard deviation of $\chi^2_{\text{test},i}$, $i = 1 \dots n$

Hyperoptimisation: K-folding



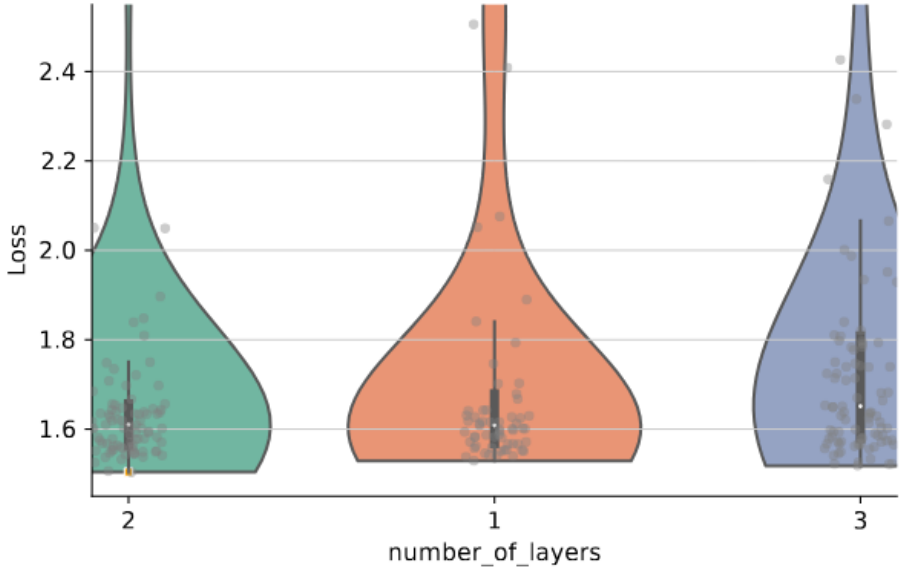
Compare to a Test Set (new set of data previously not used at all)

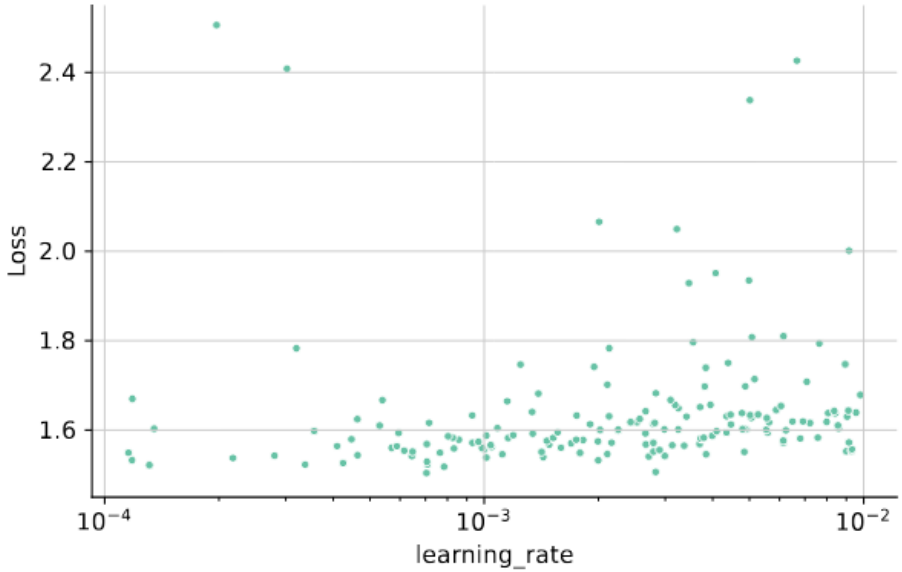
Who picks the Test Set? Automatic generalisation based on K foldings

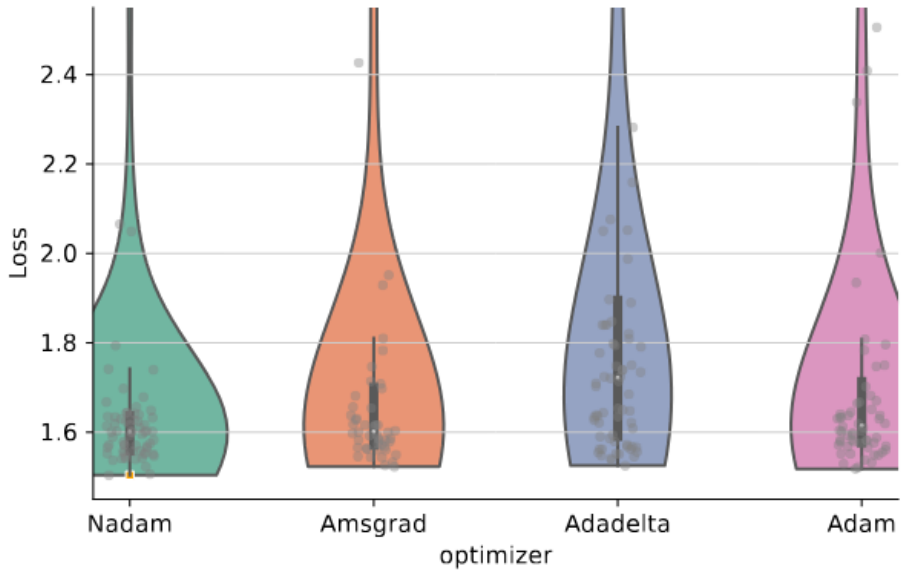
Divide the data into n representative sets, fit $n-1$ sets and use the n-th set as test set

Hyperoptimise on mean and standard deviation of $\chi^2_{\text{test},i}$, $i = 1 \ldots n$

Hyperparameters







2.2.3 Uncertainty representation

The Hessian method: general strategy

1. Expand the χ^2 about its global minimum at first (nontrivial) order

$$\chi^2\{\mathbf{a}\} \approx \chi^2\{\mathbf{a_0}\} + \delta a^i H_{ij} \delta a^j, \qquad H_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2\{\mathbf{a}\}}{\partial a_i \partial a_j} \right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}}$$

2. Assume linear error propagation for any observable $\mathcal{O}$ depending on $\{\mathbf{a}\}$

$$\langle \mathcal{O}\{\mathbf{a}\} \rangle \approx \mathcal{O}\{\mathbf{a_0}\} + a_i \left. \frac{\partial \mathcal{O}\{\mathbf{a}\}}{\partial a_i} \right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}} \qquad \sigma_{\mathcal{O}\{\mathbf{a}\}} \approx \sigma_{ij} \frac{\partial \mathcal{O}\{\mathbf{a}\}}{\partial a_i} \left. \frac{\partial \mathcal{O}\{\mathbf{a}\}}{\partial a_j} \right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}}$$

3. Determine σ_{ij} from H_{ij} from maximum likelihood (under Gaussian hypothesis)

$$\sigma_{ij}^{-1} = \left. \frac{\partial^2 \chi^2\{\mathbf{a}\}}{\partial a_i \partial a_j} \right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}} = H_{ij}$$

4. A C.L. about the best fit is obtained as the volume (in parameter space) about $\chi^2\{\mathbf{a_0}\}$ that corresponds to a fixed increase of the χ^2; for Gaussian uncertainties:

$$68\% \text{ C.L.} \iff \Delta\chi^2 = \chi^2\{\mathbf{a}\} - \chi^2\{\mathbf{a_0}\} = 1$$

The Hessian method: some remarks

1. Compact representation and computation of observables and their uncertainties

$$\langle \mathcal{O}[f(x,Q^2)] \rangle = \mathcal{O}[f_0(x,Q^2)]$$

$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \frac{1}{2} \left[\sum_{i=1}^{N_{\rm par}} \left(\mathcal{O}[f_i(x,Q^2)] - \mathcal{O}[f_0(x,Q^2)] \right)^2 \right]^{1/2}$$

2. Parameters can always be adjusted so that all eigenvalues of H_{ij} are equal to one (diagonalise H_{ij} and rescale the eigenvectors by their eigenvalues)

$$\delta a_i H_{ij} \delta a_j = \sum_{i=1}^{N_{\rm par}} \left[a_i'(a_i) \right]^2 \Longleftrightarrow \sigma_{\mathcal{O}\{\mathbf{a}'\}} = \left| \nabla' \mathcal{O}\{\mathbf{a}'\} \right|$$

   The total contribution to the uncertainty due to two different sources (possibly correlated) is obtained by simply adding them in quadrature

3. Any rotation in the space of parameters preserves the gradient (one can diagonalise a chosen observable without spoiling the result)

4. Unmanageable Hessian matrix if the numer of parameters is huge

The Monte Carlo method: general strategy

1. Generate (art) replicas of (exp) data according to the distribution

$$\mathcal{O}_i^{(art)(k)} = \mathcal{O}_i^{(exp)} + r_i^{(k)} \sigma_{\mathcal{O}_i} \,, \qquad i = 1, \ldots N_{\rm dat} \,, \qquad k = 1, \ldots, N_{\rm rep}$$

   where $r_i^{(k)}$ are (Gaussianly distributed) random numbers for each k-th replica ($r_i^{(k)}$ can be generated with any distribution, not neccesarily Gaussian)

2. Perform a fit for each replica $k = 1, \ldots, N_{\rm rep}$

3. Compact computation of observables and their uncertainties (PDF replicas are equally probable members of a statistical ensemble)

$$\langle \mathcal{O}[f(x,Q^2)] \rangle = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} \mathcal{O}[f^{(k)}(x,Q^2)]$$

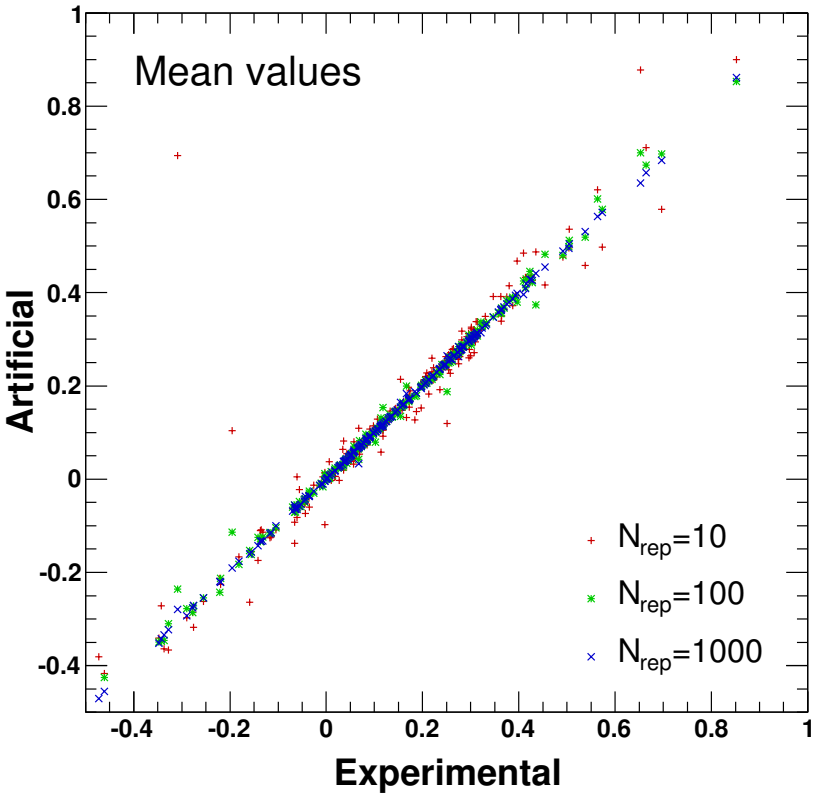
$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \left[\frac{1}{N_{\rm rep}-1} \sum_{k=1}^{N_{\rm rep}} \left(\mathcal{O}[f^{(k)}(x,Q^2)] - \langle \mathcal{O}[f(x,Q^2)] \rangle \right)^2 \right]^{1/2}$$

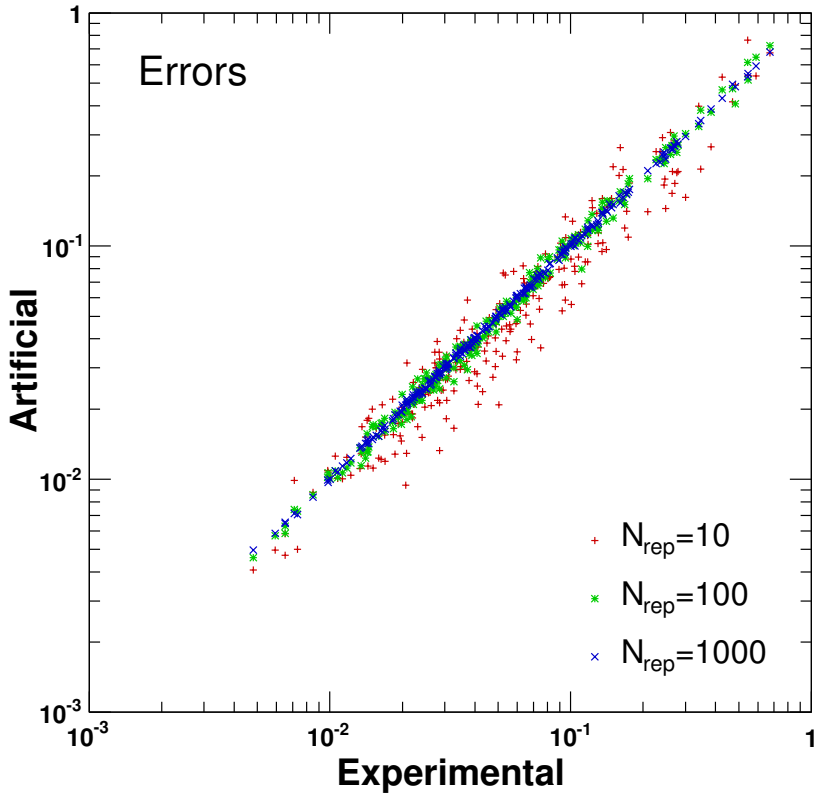
$\Rightarrow$ **no need to rely on linear approximation**

$\Rightarrow$ **computational expensive: need to perform $N_{\rm rep}$ fits instead of one**

The Monte Carlo method: determining the sample size

Require that the average over the replicas reproduces the central value of the original experimental data to a desired accuracy (the standard deviation reproduces the error and so on)

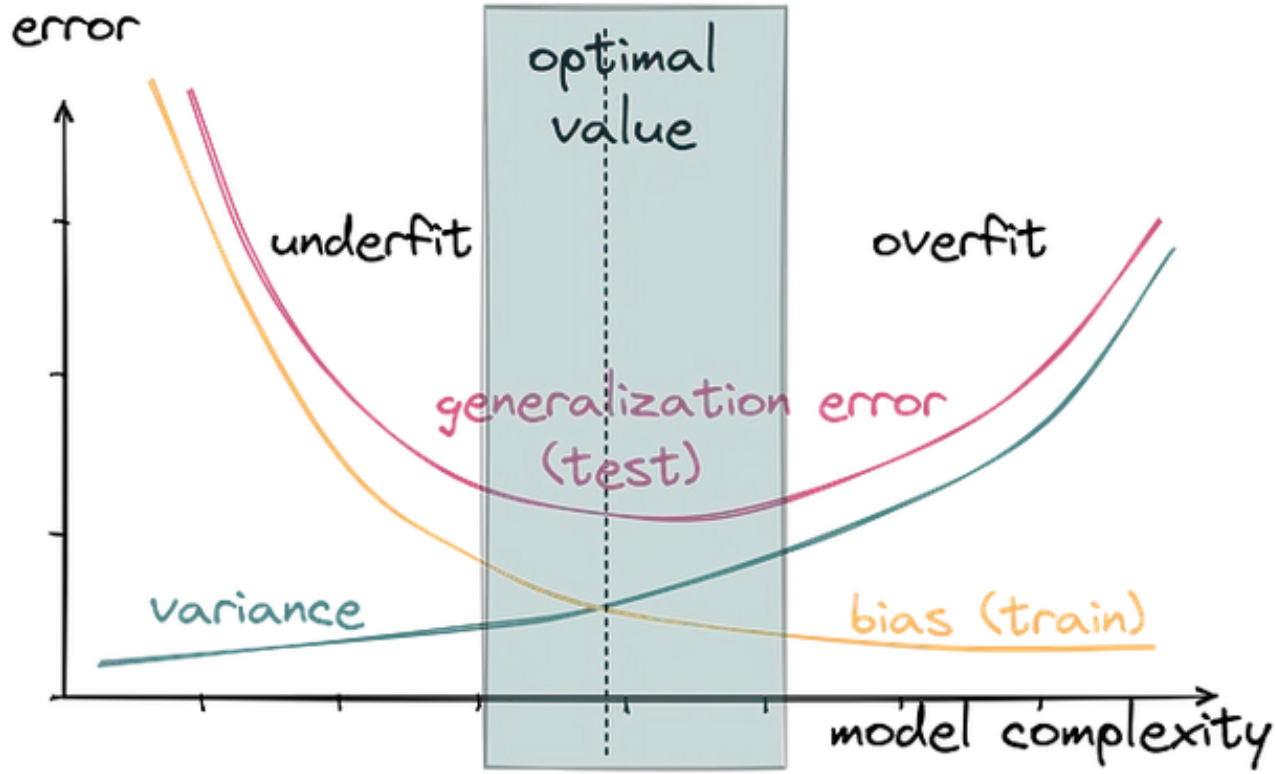


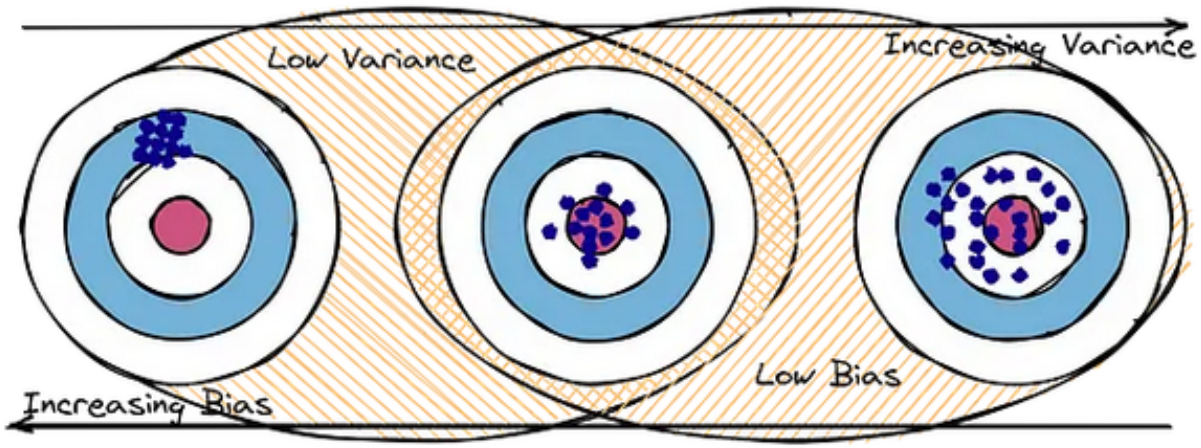


Accuracy of few % requires ~ 100 replicas

2.2.4 Validation

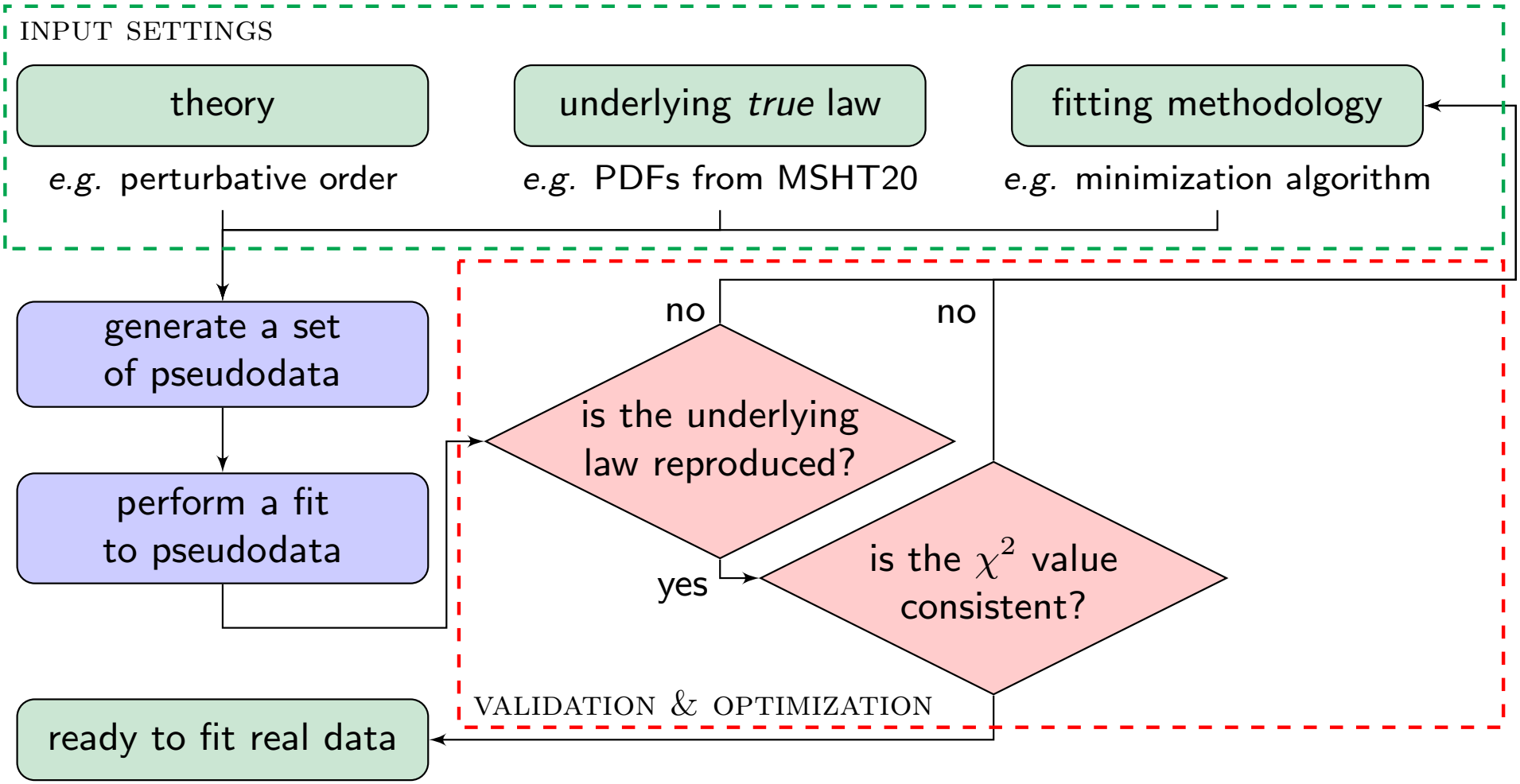
Accuracy vs precision or bias vs variance





Closure tests: general idea [JHEP 1504 (2015) 040]

Validation and optimisation of the fitting strategy with known underlying physical law

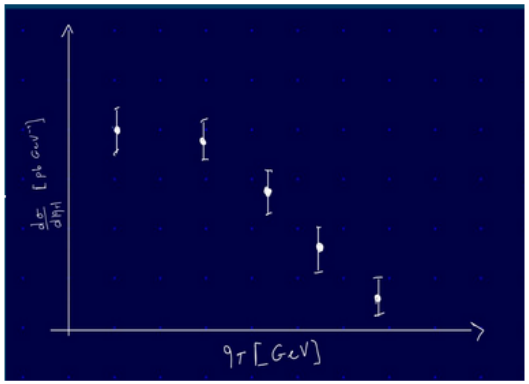


Full control of procedural uncertainties

Closure Tests: Levels

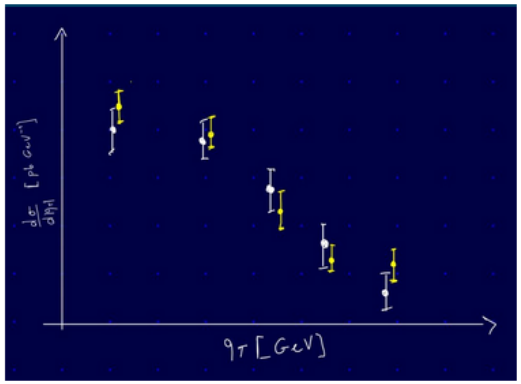
Level 0

no fluctuations



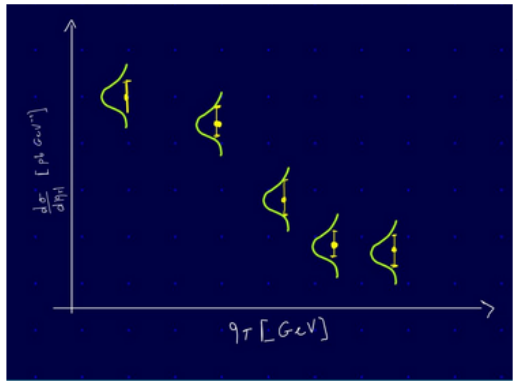
Level 1

Gaussian fluctuation

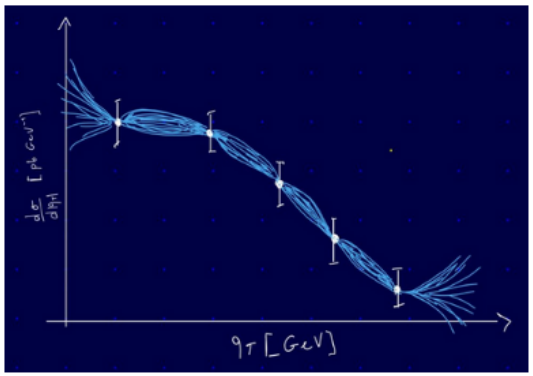


Level 2

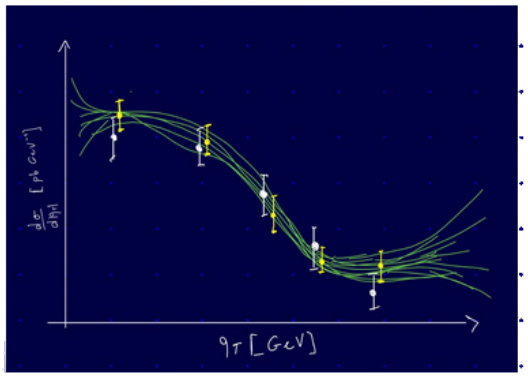
Monte Carlo replicas



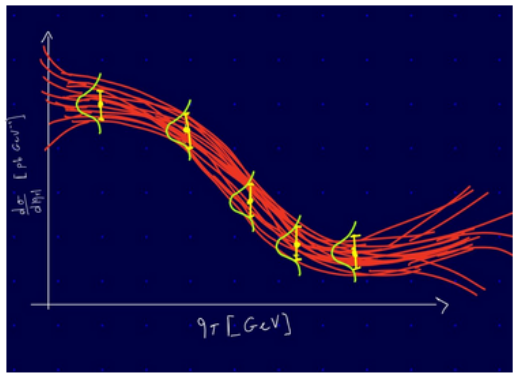
interpolation uncertainty



fuctional uncertainty



data uncertainty



Closure tests at work

Data region: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

Define bias and variance

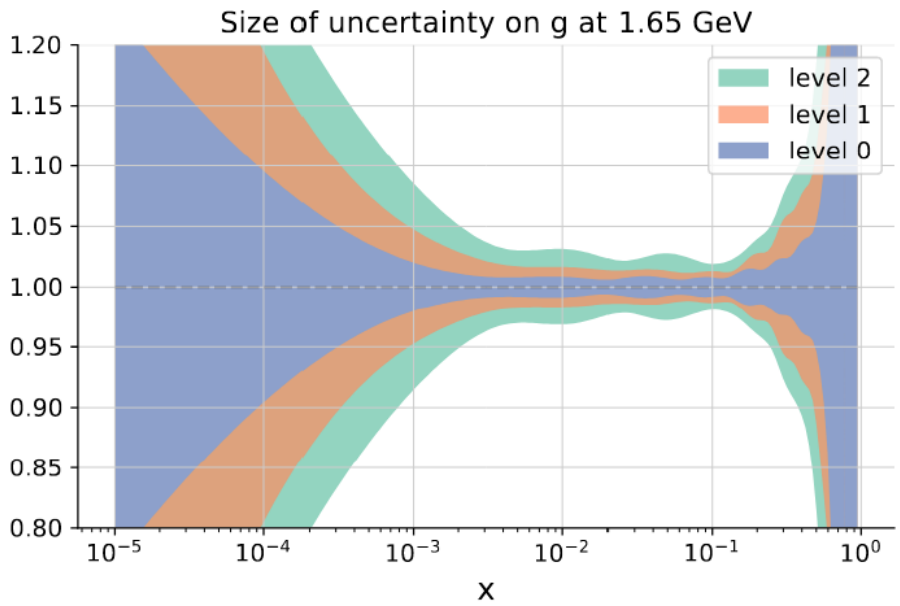
bias difference of central prediction and truth

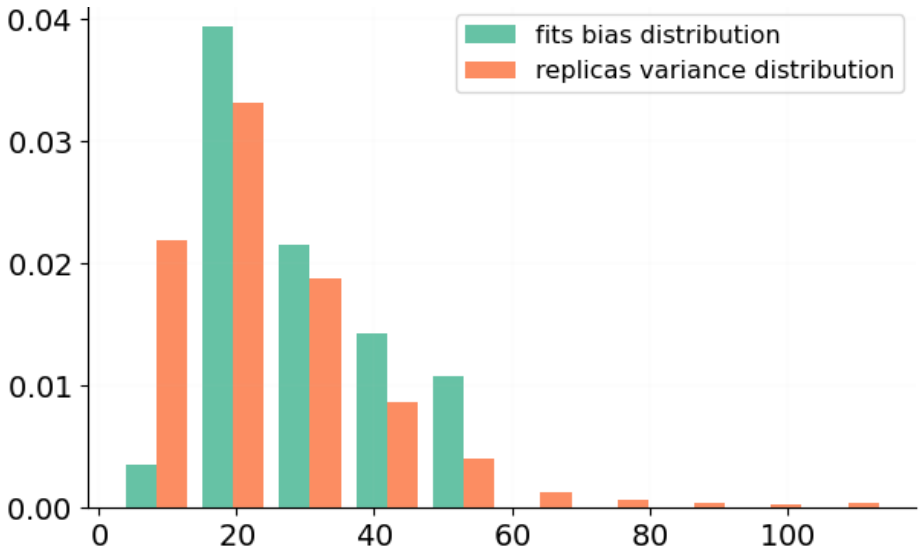
variance uncertainty of replica predictions

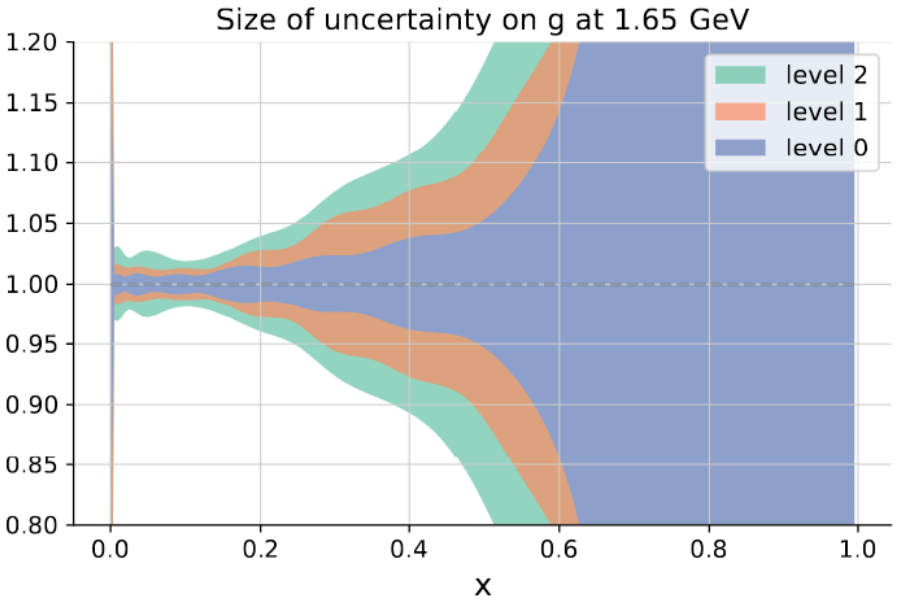
If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each

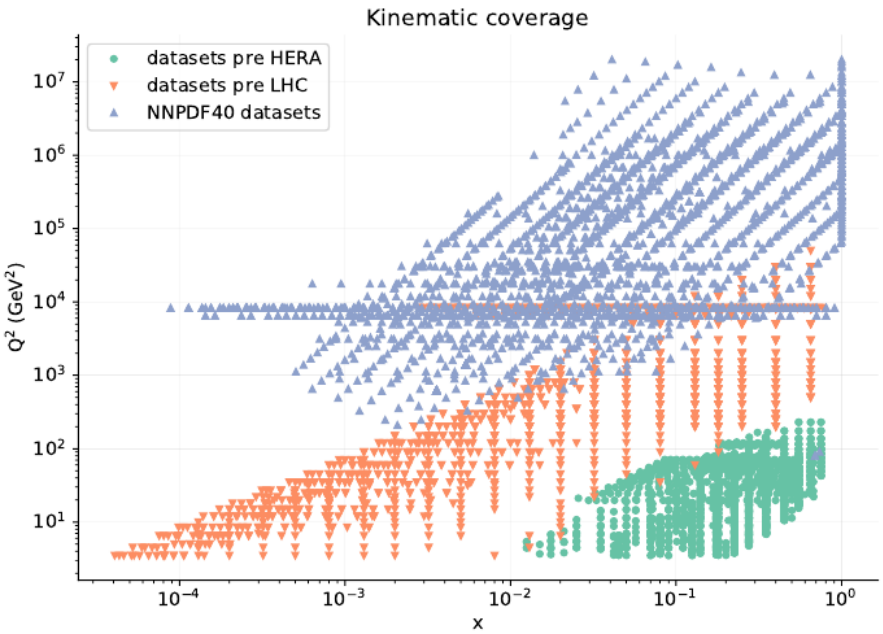


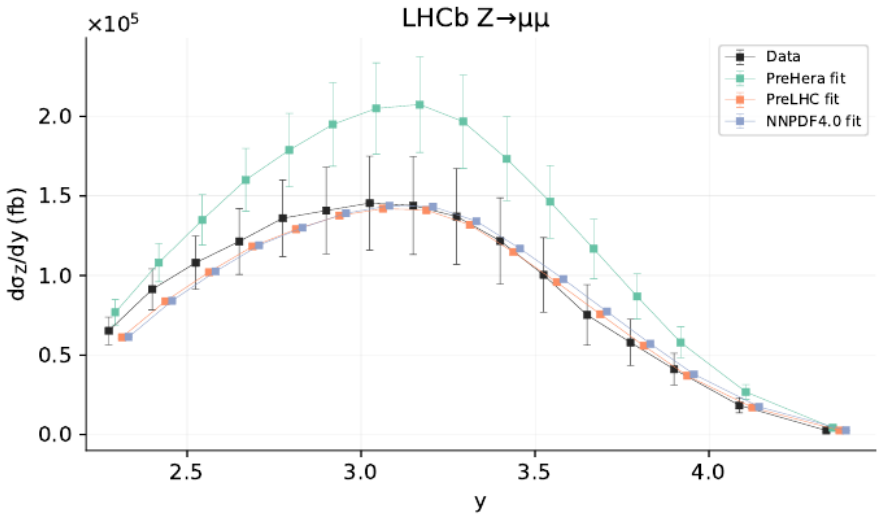




[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

Future tests



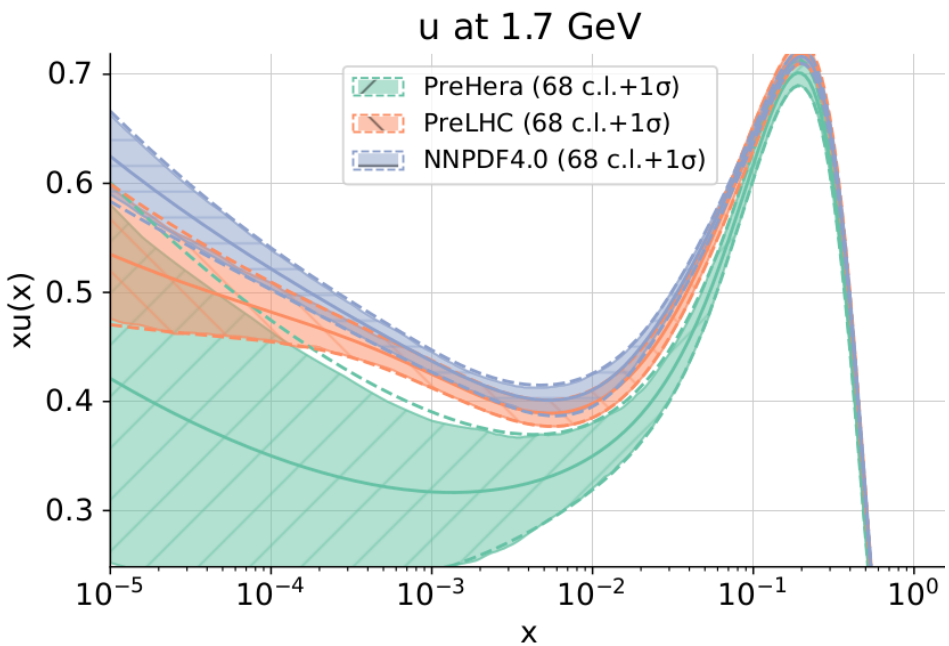


Extrapolation regions: future test

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

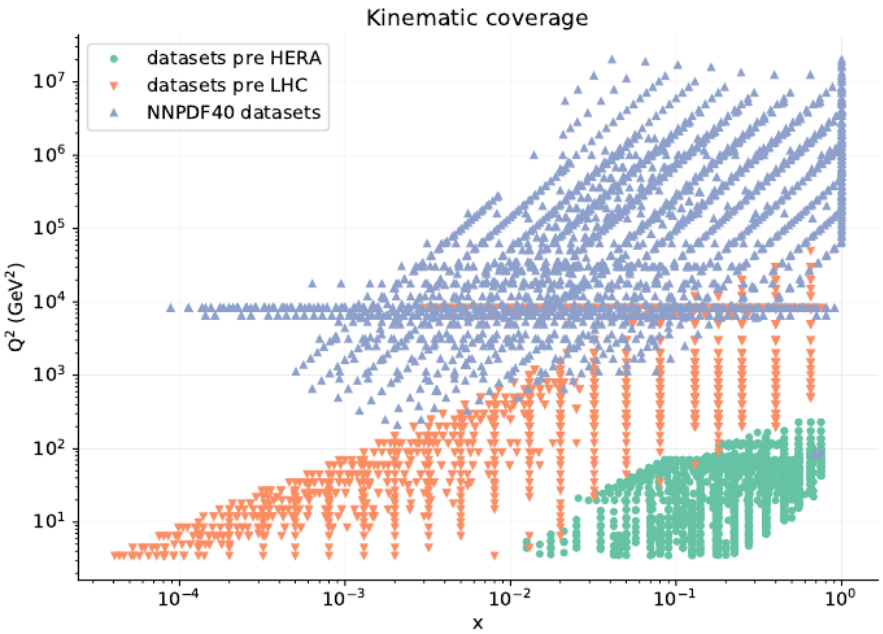
Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA	1.09	1.01	0.90
pre-LHC	1.21	1.20	23.1
NNPDF4.0	1.29	3.30	23.1

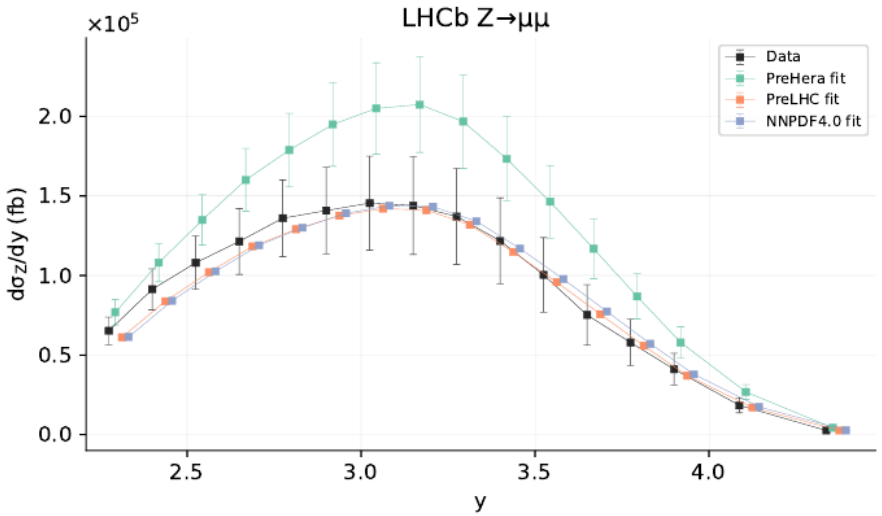
Only exp. cov. matrix



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

Future tests



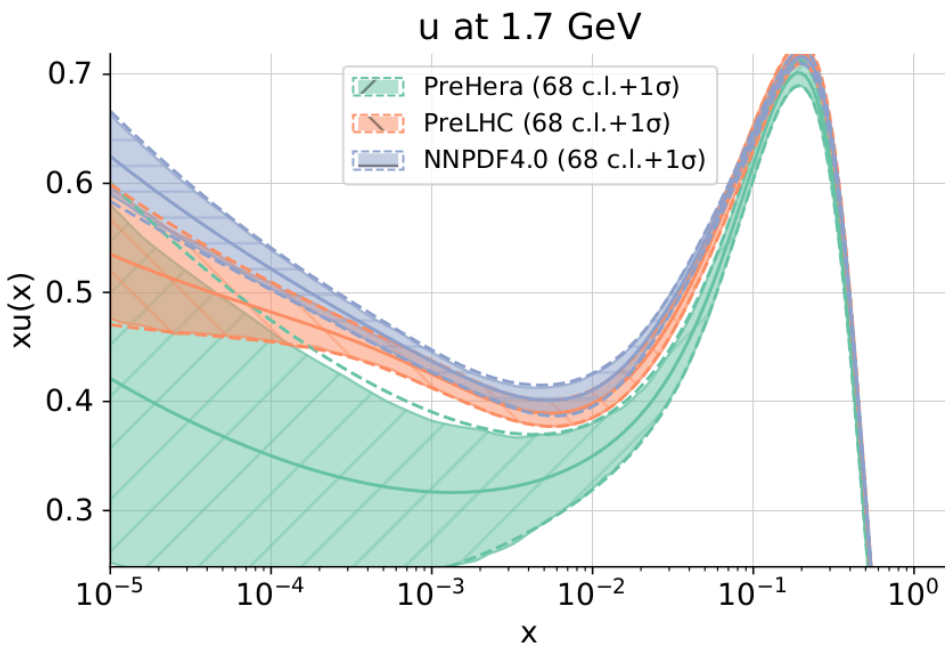


Extrapolation regions: future test

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



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

2.3 Summary of Lecture 2

Overview of current PDF determinations

	NNPDF4.0	MSHT20	CT18	HERAPDF2.0	CJ22	ABMP16
Fixed-target DIS	☑	☑	☑	☒	☑	☑
JLAB	☒	☒	☒	☒	☑	☒
HERA I+II	☑	☑	☑	☑	☑	☑
HERA jets	☑	☒	☒	☑	☒	☒
Fixed target DY	☑	☑	☑	☒	☑	☑
Tevatron W, Z	☑	☑	☑	☒	☑	☑
LHC vector boson	☑	☑	☑	☒	☑	☑
LHC $W + c$ $Z + c$	☑	☒	☒	☒	☒	☒
Tevatron jets	☑	☑	☑	☒	☑	☒
LHC jets	☑	☑	☑	☒	☒	☒
LHC top	☑	☑	☒	☒	☒	☑
LHC single t	☑	☒	☒	☒	☒	☒
LHC prompt γ	☑	☒	☒	☒	☒	☒
statistical treatment	Monte Carlo	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2 = 1$	Hessian $\Delta\chi^2 = 1.645$	Hessian $\Delta\chi^2 = 1$
parametrisation	Neural Network	Chebyschev pol.	Bernstein pol.	polynomial	polynomial	polynomial
HQ scheme	FONLL	TR′	ACOT-χ	TR′	ACOT-χ	FFN
accuracy	aN^{3}LO	aN^{3}LO	NNLO	NNLO	NLO	NNLO
latest update	EPJ C82 (2022) 428	EPJ C81 (2021) 341	PRD 103 (2021) 014013	EPJ C82 (2022) 243	PRD 107 (2023) 113005	PRD 96 (2017) 014011

All PDF sets are available as (x, Q^2) interpolation grids through the LHAPDF library

Summary of Lecture 2

1. PDF accuracy can be improved by improving the theory
   - $\longrightarrow$ proper treatment of heavy quarks is mandatory to describe DIS data
   - $\longrightarrow$ evidence for intrinsic charm in the proton
   - $\longrightarrow$ MHOUs can be estimated by scale variations
   - $\longrightarrow$ inclusion of MHOUs stabilises fit quality
   - $\longrightarrow$ electroweak corrections modify DGLAP equations
   - $\longrightarrow$ the photon PDF is determined very precisely
   - $\longrightarrow$ inclusion of photon PDFs impacts the gluon PDF
2. Devising the methodology is essential to minimise bias and variance
   - $\longrightarrow$ bias is a measure of accuracy, variance is a measure of precision
   - $\longrightarrow$ choices of parametrisation (polynomial vs neural network)
   - $\longrightarrow$ choices of uncertainty representation (Hessian vs Monte Carlo)
   - $\longrightarrow$ are all sources of bias and variance
   - $\longrightarrow$ closure tests are a way to validate PDF uncertainites in the *seen* region
   - $\longrightarrow$ future tests validate the generalisation power in the *unseen* region

Thank you