Parton distributions, big-data paradox, and intrinsic charm

Pavel Nadolsky
Southern Methodist University

With A. Courtoy, M. Guzzi, T. Hobbs,
J. Huston, K. Xie, M. Yan, C.-P. Yuan

and members of the
CTEQ-TEA (Tung Et. Al.) working group

PDF uncertainties: balancing **precision** and **robustness**

The critical role of controlling for **sampling biases** in QCD analyses

Contents, part 1

1. The HL-LHC and Tevatron physics programs require accurate parton distribution functions (PDFs) in the proton
   - CTEQ-TEA (**CT18**), **PDF4LHC21,** and other recent NNLO **PDFs**

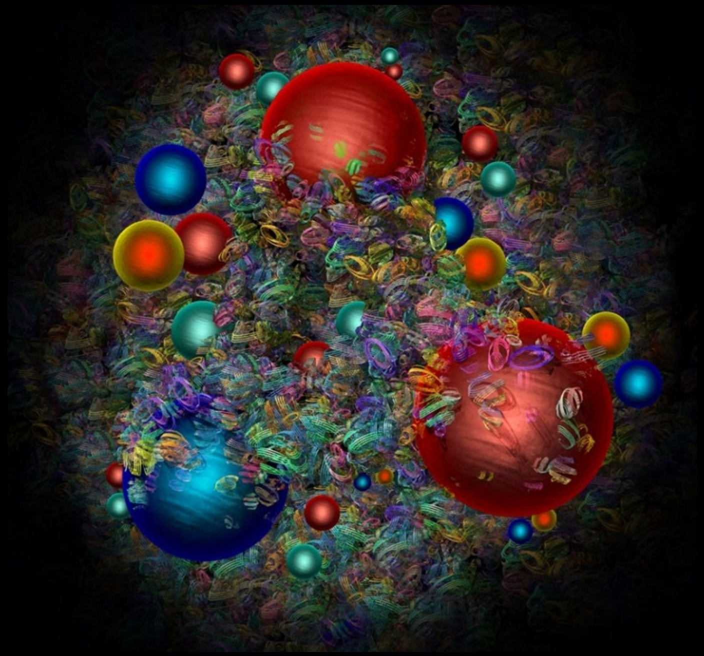
2. The **tolerance puzzle**: how well do we know the PDFs?
   - **The big data paradox**
     - ➢ quality of data and representative sampling of PDF solutions may matter more than (N)NNLO accuracy of individual solutions

   - Part 2, at LPC Physics Forum, Thursday, 1pm: **hopscotch scans**, the role of experimental systematic uncertainties

3. **Nonperturbative** (**intrinsic**) **charm** production
   - What is it? What is the experimental evidence?

A proton at rest

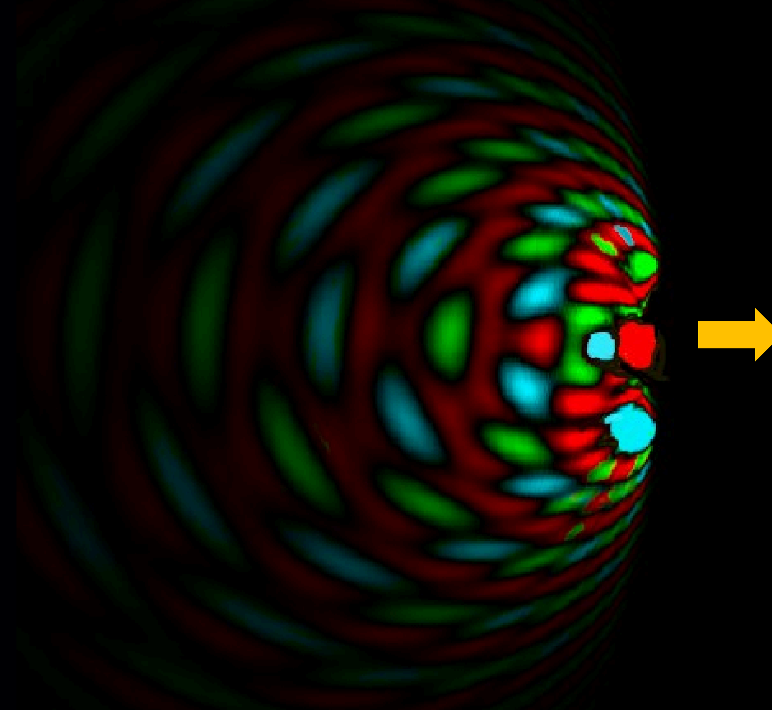
$V \approx 0$



- Nonperturbative and lattice QCD models of proton structure

A proton at a collider

moving with speed $V \approx c$ to the right

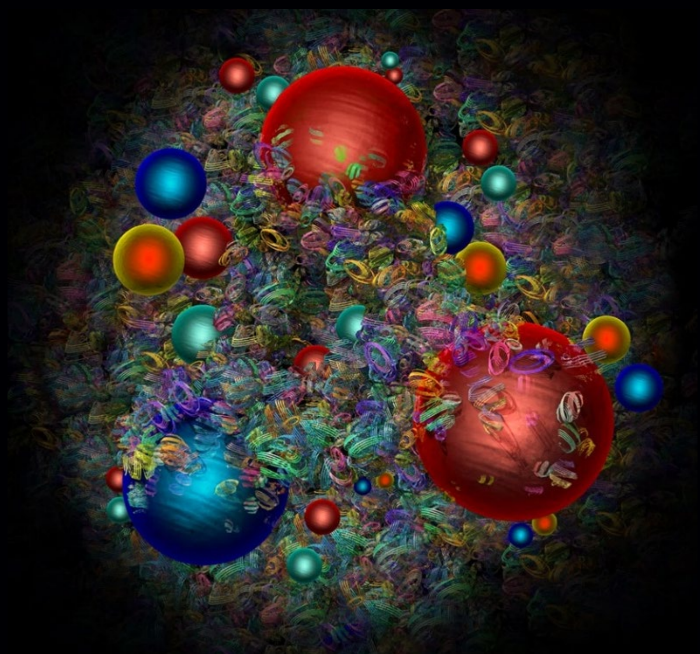


QCD factorization:

- short-distance perturbative expansions on the light front
- **universal long-distance functions**

A proton at rest

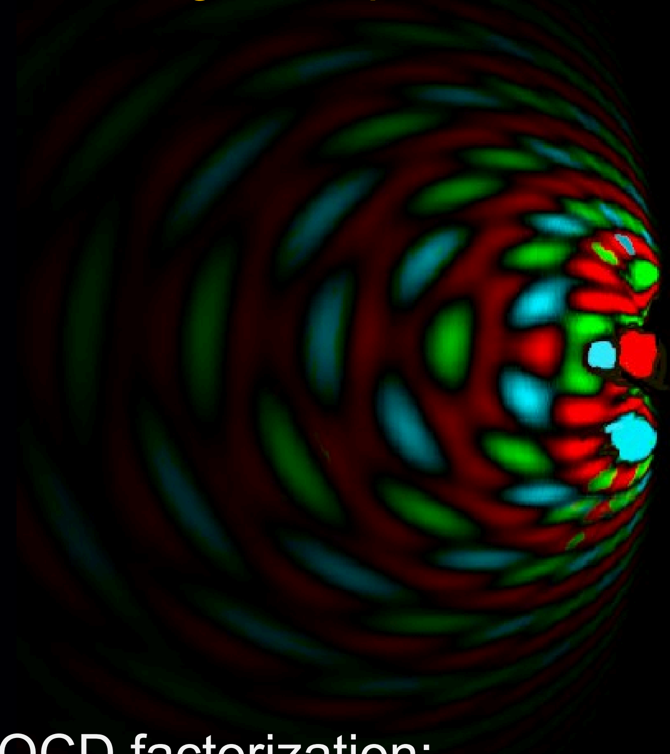
$V \approx 0$



- Nonperturbative and lattice QCD models of proton structure

A proton at a collider

moving with speed $V \approx c$ to the right

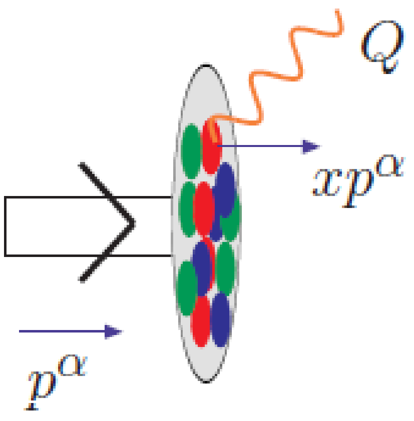


QCD factorization:

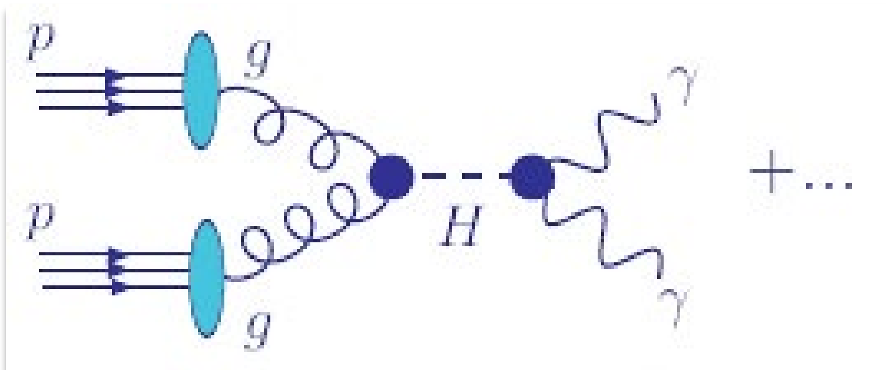
- short-distance perturbative expansions on the light front
- **universal long-distance functions**

$$f_{a/h}(x, Q)$$

Unpolarized collinear parton distributions $f_{a/h}(x, Q)$ are associated with probabilities for finding a parton a with the "+" momentum xp^+ in a hadron h with the "+" momentum p^+ for $p^+ \to \infty$, at a resolution scale $Q > 1$ GeV



Parton distributions describe long-distance dynamics in high-energy collisions



$$\sigma_{pp\to H\to\gamma\gamma X}(Q) = \sum_{a,b=g,q,\bar{q}} \int_0^1 d\xi_a \int_0^1 d\xi_b \hat{\sigma}_{ab\to H\to\gamma\gamma}\left(\frac{x_a}{\xi_a}, \frac{x_b}{\xi_b}, \frac{Q}{\mu_R}, \frac{Q}{\mu_F}; \alpha_s(\mu_R)\right)$$

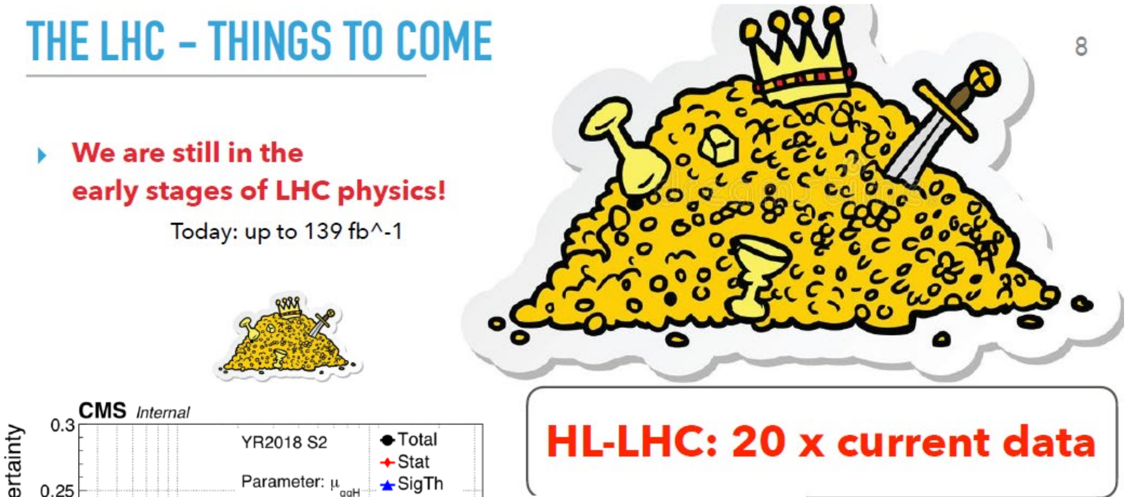
$$\times f_a(\xi_a, \mu_F) f_b(\xi_b, \mu_F) + O\left(\frac{\Lambda_{QCD}^2}{Q^2}\right)$$

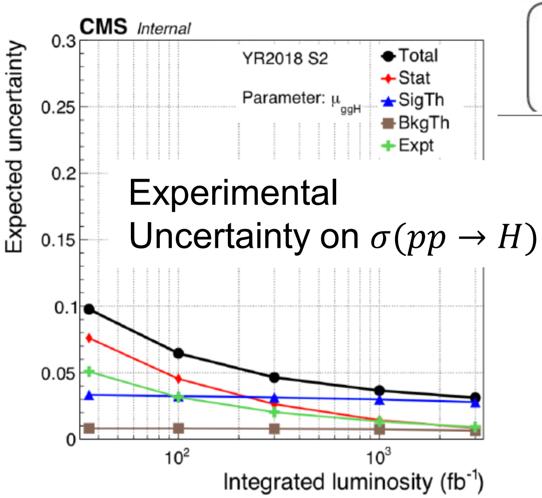
$\hat{\sigma}$ is the hard cross section; computed order-by-order in $\alpha_s(\mu_R)$

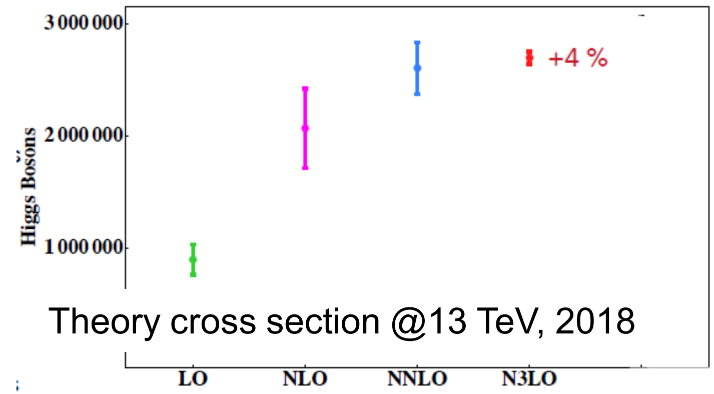
$f_a(x, \mu_F)$ is the distribution for parton a with momentum fraction x, at scale μ_F

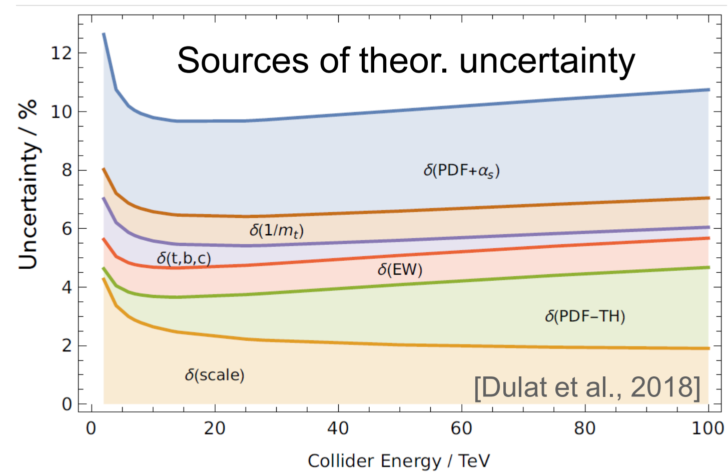
Higgs physics relies on QCD

B. Mistlberger, CTEQ SS22









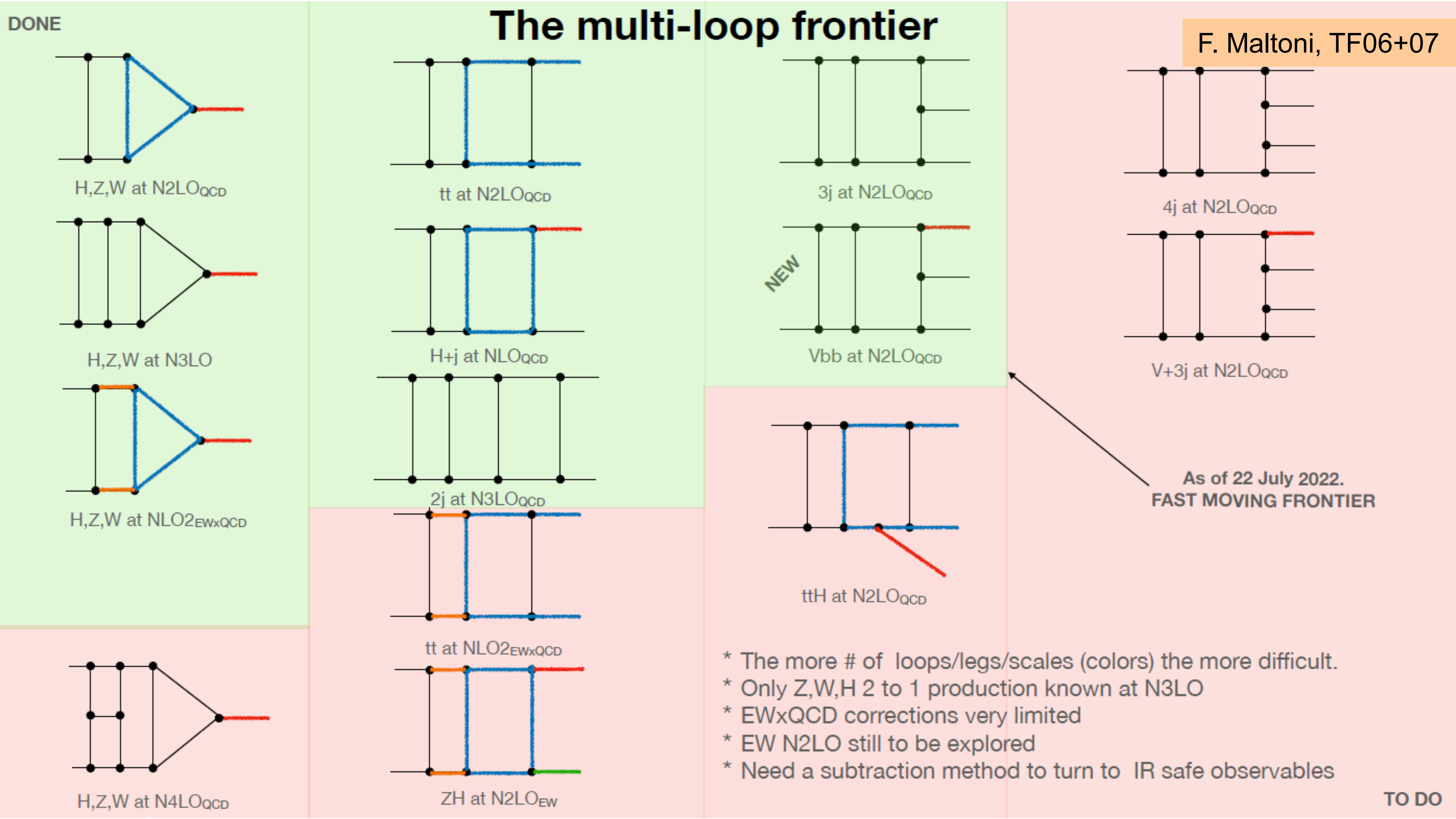
2022 Les Houches wish list for PQCD calculations for hadron colliders

TABLE IV. Summary of the LesHouches precision wishlist for hadron colliders [545]. HTL stands for calculations in heavy top limit, VBF* stands for structure function approximation.

process	known	desired
$pp \to H$	$\mathrm{N^3LO_{HTL}}$, $\mathrm{NNLO^{(t)}_{QCD}}$, $\mathrm{N^{(1,1)}LO^{(HTL)}_{QCD\otimes EW}}$	$\mathrm{N^4LO_{HTL}}$ (incl.), $\mathrm{NNLO^{(b,c)}_{QCD}}$
$pp \to H + j$	$\mathrm{NNLO_{HTL}}$, $\mathrm{NLO_{QCD}}$, $\mathrm{N^{(1,1)}LO_{QCD\otimes EW}}$	$\mathrm{NNLO_{HTL} \otimes NLO_{QCD} + NLO_{EW}}$
$pp \to H + 2j$	$\mathrm{NLO_{HTL} \otimes LO_{QCD}}$ $\mathrm{N^3LO^{(VBF^*)}_{QCD}}$ (incl.), $\mathrm{NNLO^{(VBF^*)}_{QCD}}$, $\mathrm{NLO^{(VBF)}_{EW}}$	$\mathrm{NNLO_{HTL} \otimes NLO_{QCD} + NLO_{EW}}$, $\mathrm{NNLO^{(VBF)}_{QCD}}$
$pp \to H + 3j$	$\mathrm{NLO_{HTL}}$, $\mathrm{NLO^{(VBF)}_{QCD}}$	$\mathrm{NLO_{QCD} + NLO_{EW}}$
$pp \to VH$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$, $\mathrm{NLO^{(t,b)}_{gg \to HZ}}$	
$pp \to VH + j$	$\mathrm{NNLO_{QCD}}$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$
$pp \to HH$	$\mathrm{N^3LO_{HTL} \otimes NLO_{QCD}}$	$\mathrm{NLO_{EW}}$
$pp \to HHH$	$\mathrm{NNLO_{HTL}}$	
$pp \to H + t\bar{t}$	$\mathrm{NLO_{QCD} + NLO_{EW}}$, $\mathrm{NNLO_{QCD}}$ (off-diag.)	$\mathrm{NNLO_{QCD}}$
$pp \to H + t/\bar{t}$	$\mathrm{NLO_{QCD}}$	$\mathrm{NNLO_{QCD}}$, $\mathrm{NLO_{QCD} + NLO_{EW}}$
$pp \to V$	$\mathrm{N^3LO_{QCD}}$, $\mathrm{N^{(1,1)}LO_{QCD\otimes EW}}$, $\mathrm{NLO_{EW}}$	$\mathrm{N^3LO_{QCD} + N^{(1,1)}LO_{QCD\otimes EW}}$, $\mathrm{N^2LO_{EW}}$
$pp \to VV'$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$, $+\mathrm{NLO_{QCD}}$ (gg)	$\mathrm{NLO_{QCD}}$ (gg,massive loops)
$pp \to V + j$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$	hadronic decays
$pp \to V + 2j$	$\mathrm{NLO_{QCD} + NLO_{EW}}$, $\mathrm{NLO_{EW}}$	$\mathrm{NNLO_{QCD}}$
$pp \to V + b\bar{b}$	$\mathrm{NLO_{QCD}}$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$
$pp \to VV' + 1j$	$\mathrm{NLO_{QCD} + NLO_{EW}}$	$\mathrm{NNLO_{QCD}}$
$pp \to VV' + 2j$	$\mathrm{NLO_{QCD}}$ (QCD), $\mathrm{NLO_{QCD} + NLO_{EW}}$ (EW)	Full $\mathrm{NLO_{QCD} + NLO_{EW}}$
$pp \to W^+W^+ + 2j$	Full $\mathrm{NLO_{QCD} + NLO_{EW}}$	
$pp \to W^+W^- + 2j$	$\mathrm{NLO_{QCD} + NLO_{EW}}$ (EW component)	
$pp \to W^+Z + 2j$	$\mathrm{NLO_{QCD} + NLO_{EW}}$ (EW component)	
$pp \to ZZ + 2j$	Full $\mathrm{NLO_{QCD} + NLO_{EW}}$	
$pp \to VV'V''$	$\mathrm{NLO_{QCD}}$, $\mathrm{NLO_{EW}}$ (w/o decays)	$\mathrm{NLO_{QCD} + NLO_{EW}}$
$pp \to W^\pm W^+ W^-$	$\mathrm{NLO_{QCD} + NLO_{EW}}$	
$pp \to \gamma\gamma$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$	$\mathrm{N^3LO_{QCD}}$
$pp \to \gamma + j$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$	$\mathrm{N^3LO_{QCD}}$
$pp \to \gamma\gamma + j$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$, $+\mathrm{NLO_{QCD}}$ (gg channel)	
$pp \to \gamma\gamma\gamma$	$\mathrm{NNLO_{QCD}}$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$
$pp \to 2\,\mathrm{jets}$	$\mathrm{NNLO_{QCD}}$, $\mathrm{NLO_{QCD} + NLO_{EW}}$	$\mathrm{N^3LO_{QCD} + NLO_{EW}}$
$pp \to 3\,\mathrm{jets}$	$\mathrm{NNLO_{QCD} + NLO_{EW}}$	
$pp \to t\bar{t}$	$\mathrm{NNLO_{QCD}}$ (w/ decays)+ $\mathrm{NLO_{EW}}$ (w/o decays) $\mathrm{NLO_{QCD} + NLO_{EW}}$ (w/ decays, off-shell) $\mathrm{NNLO_{QCD}}$	$\mathrm{N^3LO_{QCD}}$
$pp \to t\bar{t} + j$	$\mathrm{NLO_{QCD}}$ (w/ decays, off-shell) $\mathrm{NLO_{EW}}$ (w/o decays)	$\mathrm{NNLO_{QCD} + NLO_{EW}}$ (w/ decays)
$pp \to t\bar{t} + 2j$	$\mathrm{NLO_{QCD}}$ (w/o decays)	$\mathrm{NLO_{QCD} + NLO_{EW}}$ (w/ decays)
$pp \to t\bar{t} + Z$	$\mathrm{NLO_{QCD} + NLO_{EW}}$ (w/o decays) $\mathrm{NLO_{QCD}}$ (w/ decays, off-shell)	$\mathrm{NNLO_{QCD} + NLO_{EW}}$ (w/ decays)
$pp \to t\bar{t} + W$	$\mathrm{NLO_{QCD} + NLO_{EW}}$ (w/ decays, off-shell)	$\mathrm{NNLO_{QCD} + NLO_{EW}}$ (w/ decays)
$pp \to t/\bar{t}$	$\mathrm{NNLO_{QCD}}$*(w/ decays) $\mathrm{NLO_{EW}}$ (w/o decays)	$\mathrm{NNLO_{QCD} + NLO_{EW}}$ (w/ decays)
$pp \to tZj$	$\mathrm{NLO_{QCD} + NLO_{EW}}$ (w/ decays)	$\mathrm{NNLO_{QCD} + NLO_{EW}}$ (w/o decays)

A. Huss, J. Huston, S. Jones, and M. Pellen, "Report on the standard model precision wishlist,". arXiv:2207.02122

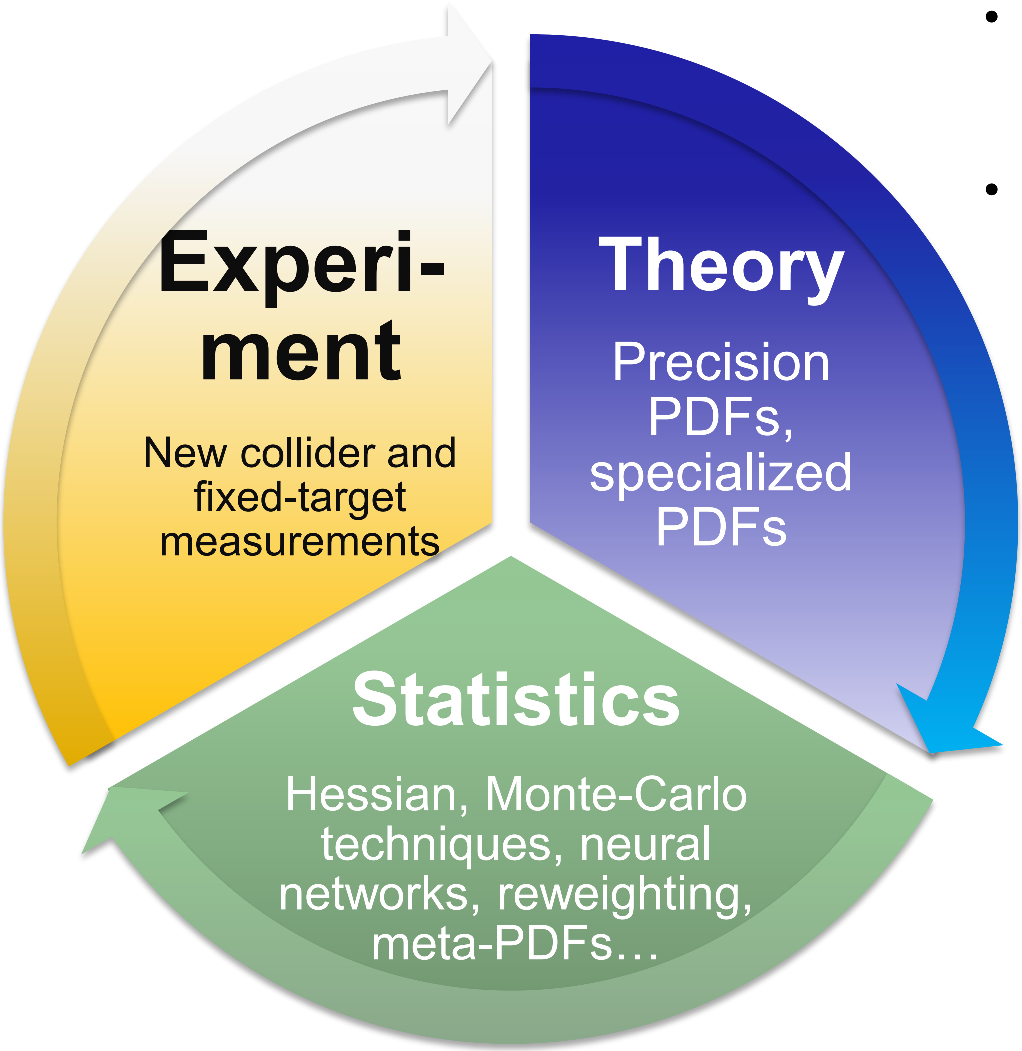
LHC experiments need accurate QCD predictions



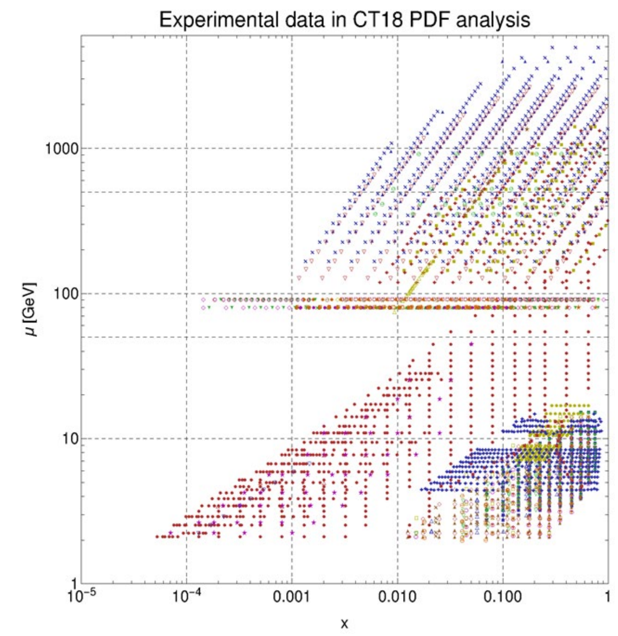
Dramatic advances in **perturbative** computations of NLO/NNLO/N3LO hard cross sections $\hat{\sigma}$.

To make use of them, accuracy of PDFs $f_{a/p}(x,Q)$ must keep up

Global fits of PDFs



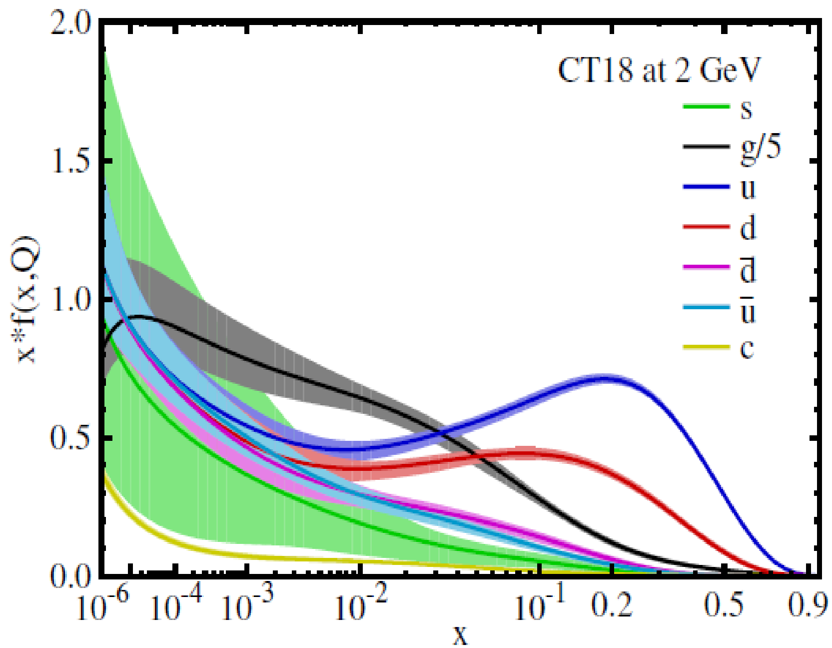
- Fits of PDFs is a rich subject at the intersection of QCD experiment, theory, and statistics
- They compare QCD computations up to NNLO with a variety of experiments probing various PDF combinations

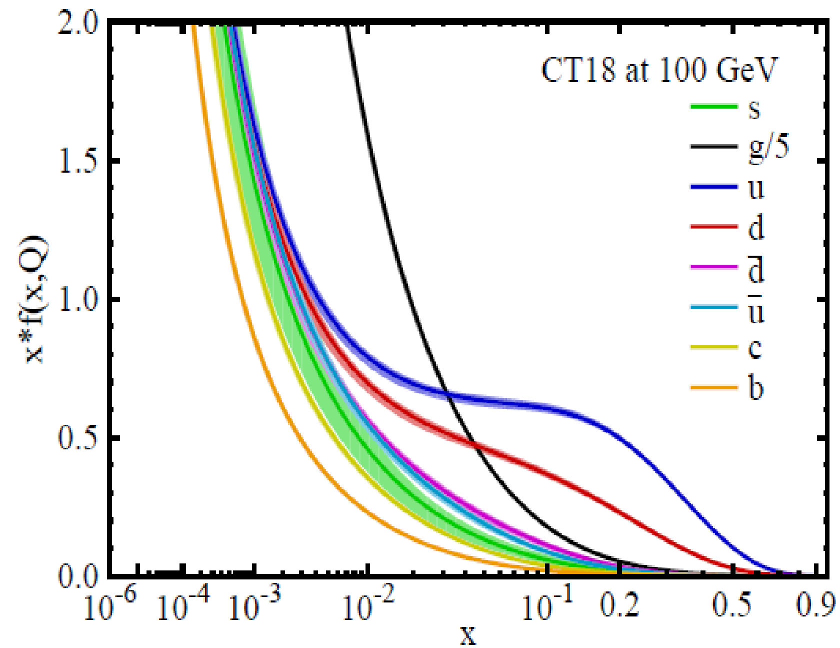


Examples from CTEQ-TEA studies and the Snowmass'2021 whitepaper "Proton structure at the precision frontier" arXiv:2203.13923

CT18 parton distributions

Recent PDFs from the CTEQ-TEA group arXiv:1912.10053 [hep-ph]

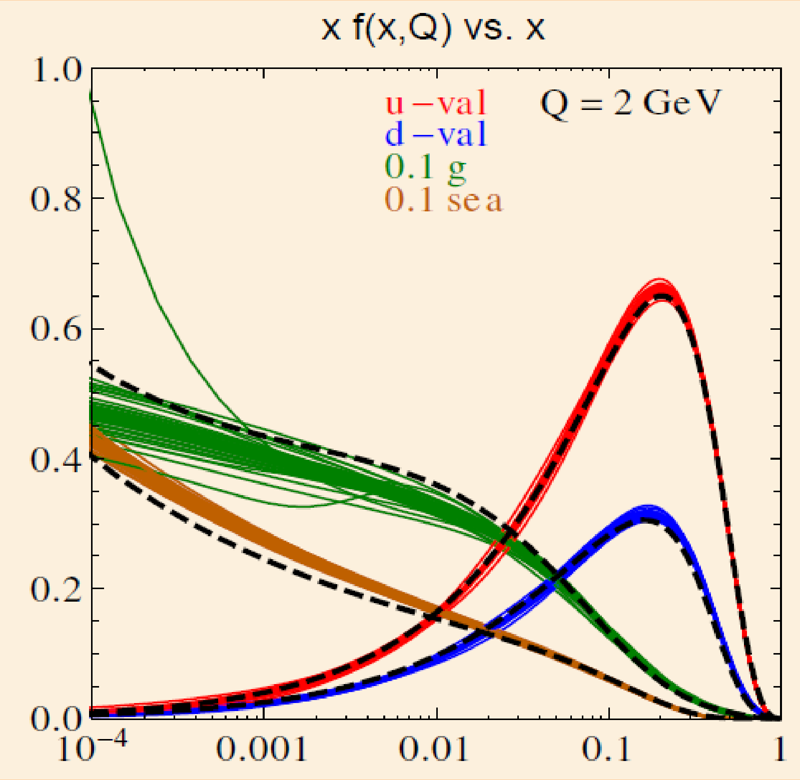




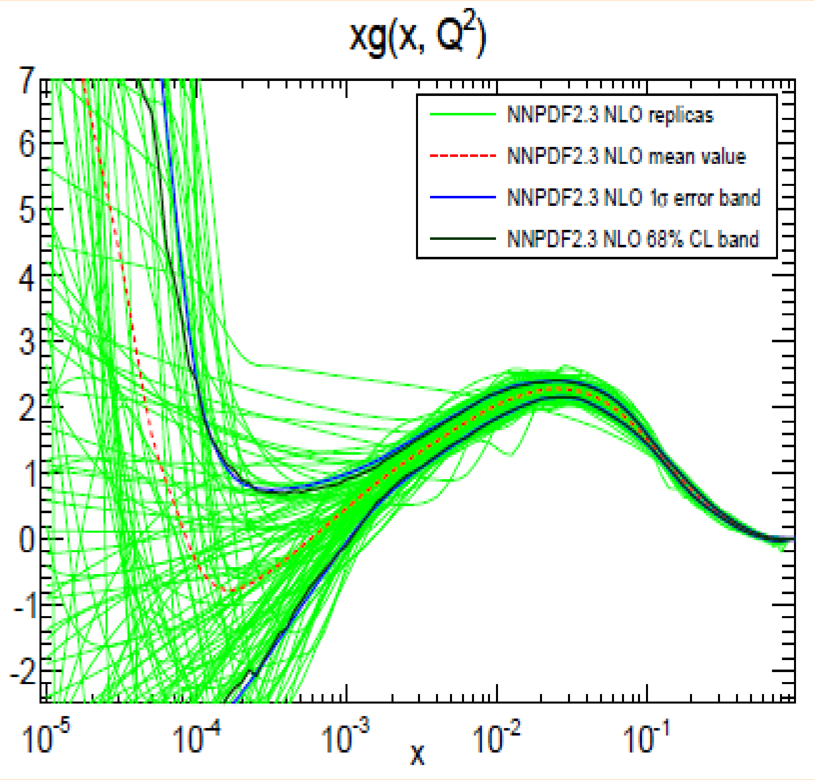
- Precise experimental data sets from ep collider HERA, LHC, Tevatron, fixed-target experiments
- Next-to-next-to-leading order (NNLO) accuracy in QCD coupling α_s
- Flexible parametric forms
- Central PDFs and bands of estimated uncertainty
- Four PDF ensembles: **CT18** (recommended), **CT18Z** (alternative), **A**, **X**

Two types of modern error PDFs

Analytic parametrizations +
Hessian PDF eigenvector sets
(ABM, CTEQ, HERA, MMHT,…)



Neural network parameterizations
+ Monte Carlo PDF replicas
(NNPDF)



Two powerful, complementary representations.
Hessian PDFs can be converted into MC ones, and vice versa.

Comparisons of the latest PDF sets…

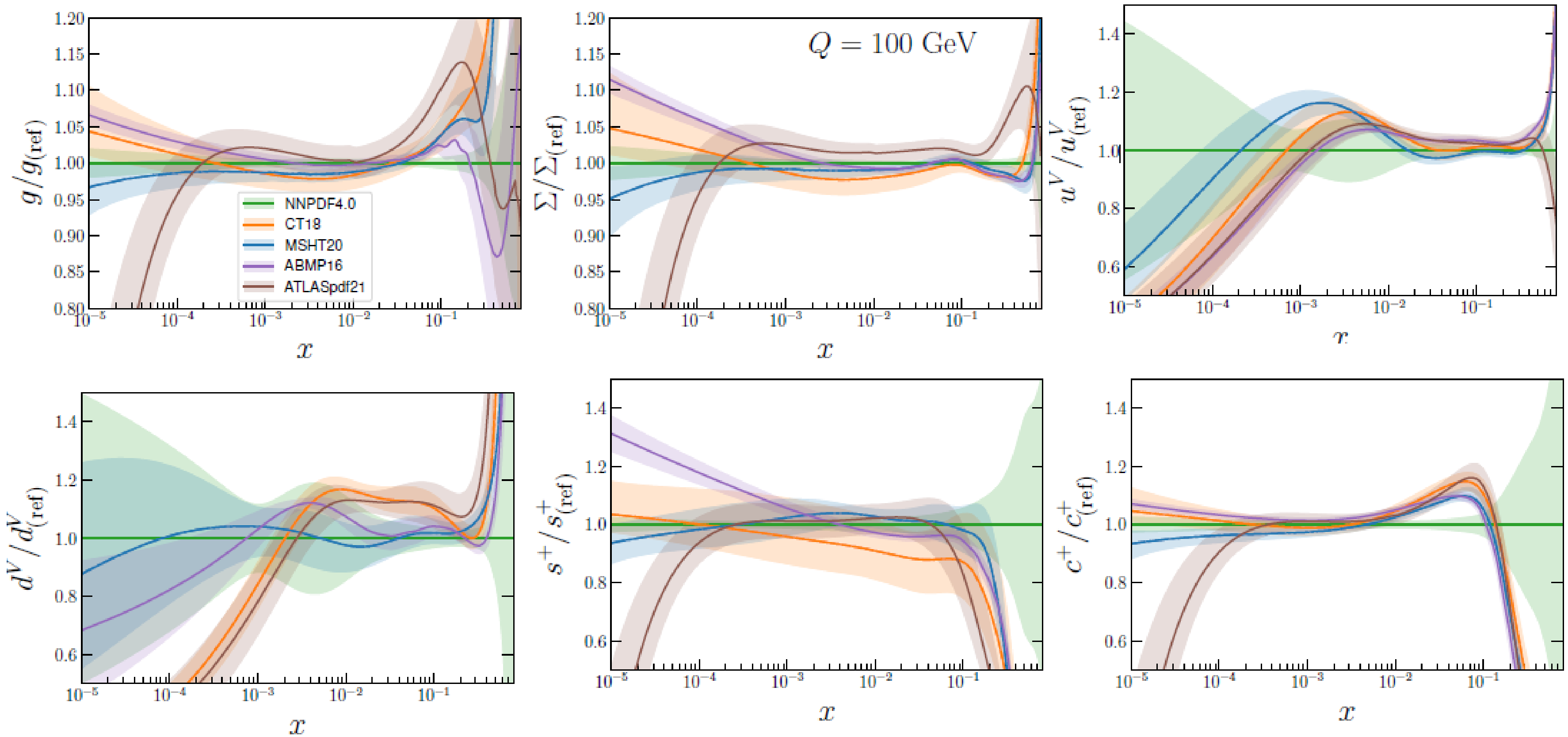


FIG. 2. Comparison of the PDFs at $Q = 100$ GeV. The PDFs shown are the N2LO sets of NNPDF4.0, CT18, MSHT20, ABMP16 with $\alpha_s(M_Z) = 0.118$, and ATLASpdf21. The ratio to the NNPDF4.0 central value and the relative 1σ uncertainty are shown for the gluon g, singlet Σ, total strangeness $s^+ = s + \bar{s}$, total charm $c^+ = c + \bar{c}$, up valence u^V and down valence d^V PDFs.

… PDF uncertainties…

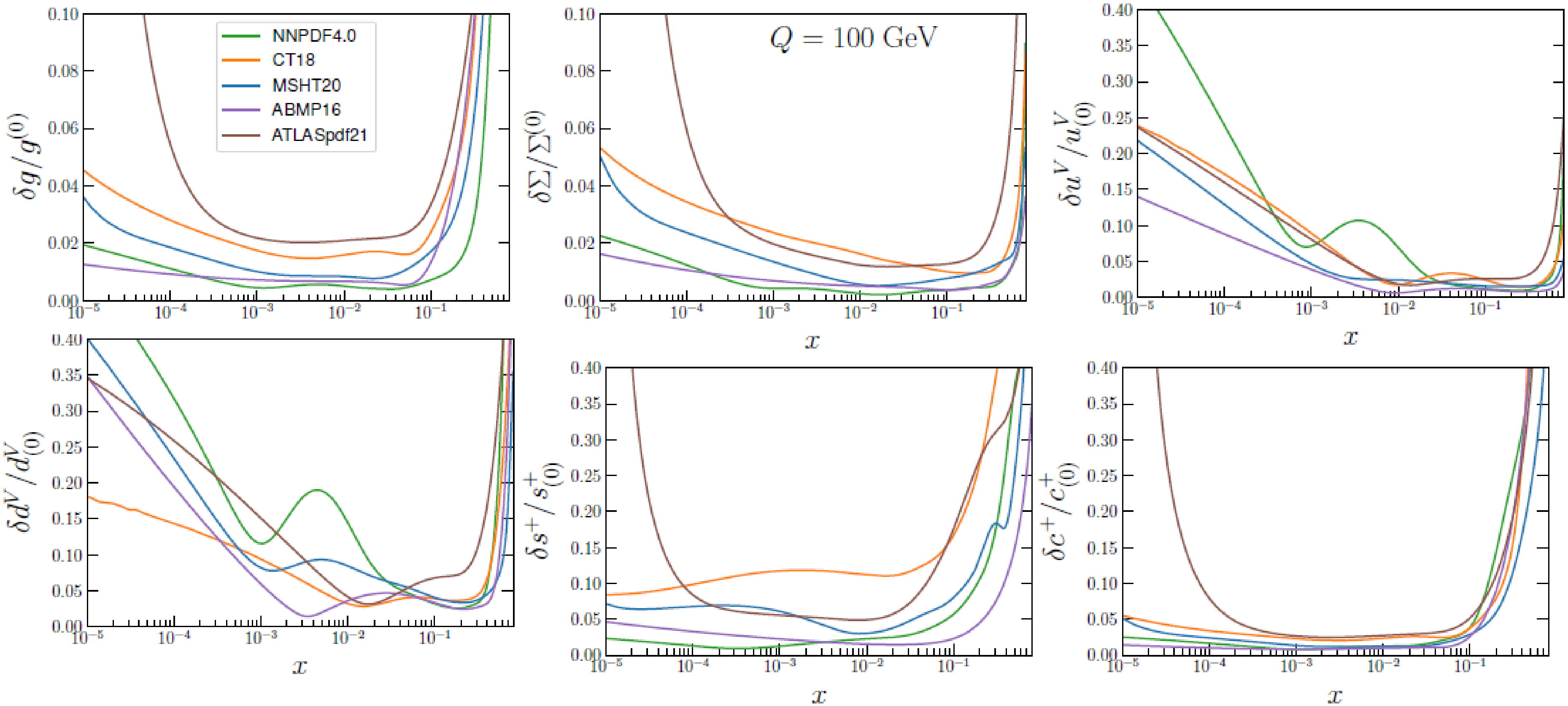


FIG. 3. Comparison of the symmetrized PDF uncertainties at $Q = 100$ GeV for the gluon g, singlet Σ, total strangeness $s^+ = s + \bar{s}$, total charm $c^+ = c + \bar{c}$, up valence u^V and down valence d^V PDFs. The PDF sets shown are the N2LO sets of NNPDF4.0, CT18, MSHT20, ABMP16 with $\alpha_s(M_Z) = 0.118$ and ATLASpdf21.

… parton luminosities…

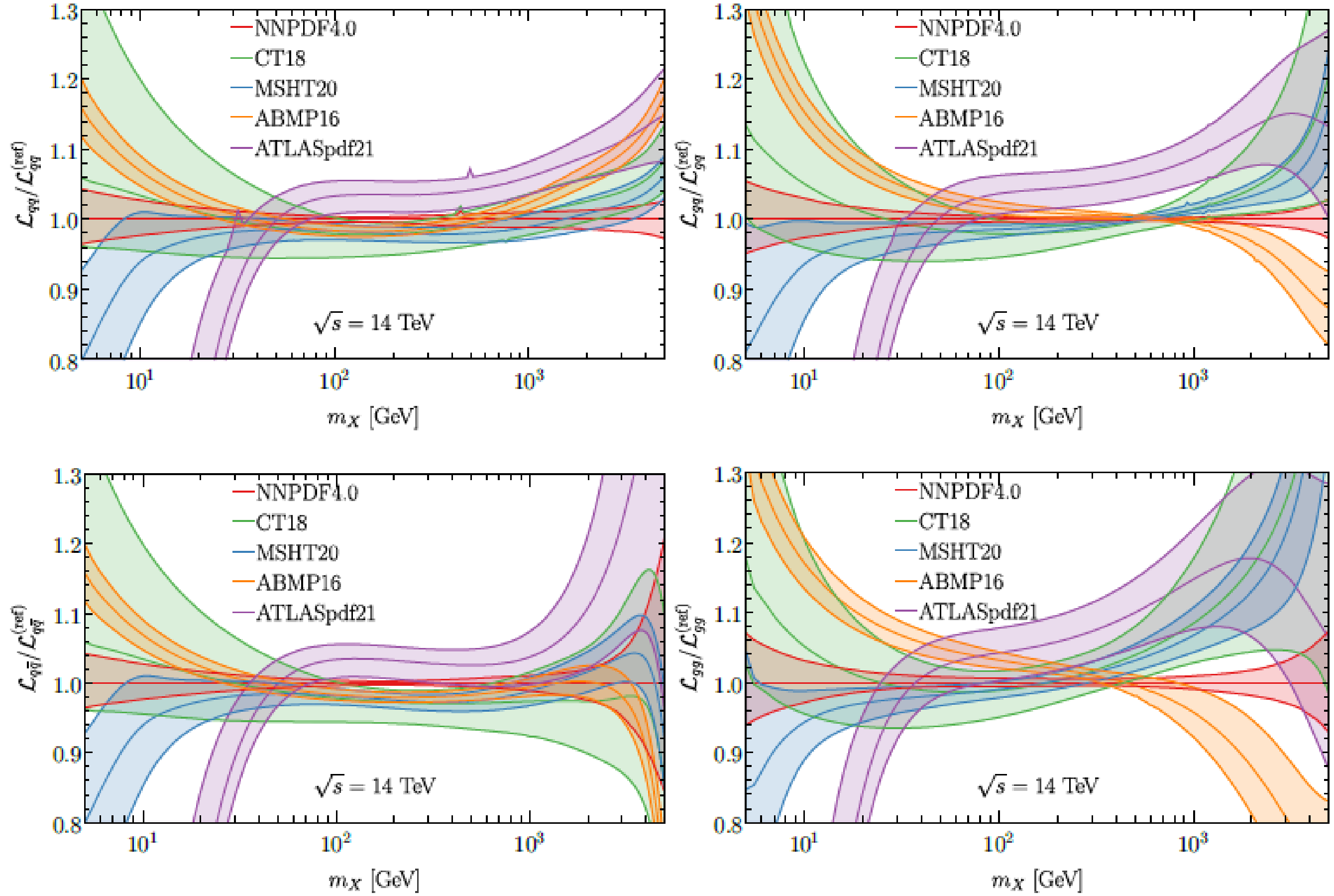
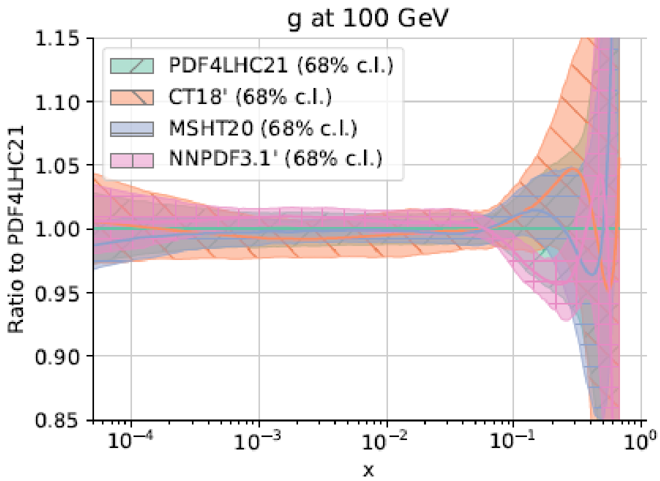


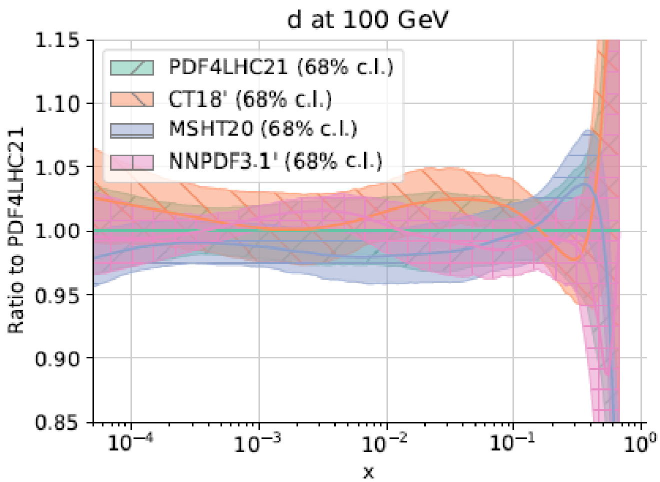
FIG. 4. Comparison, as a function of the invariant mass m_X, of the parton luminosities at $\sqrt{s} = 14$ TeV, computed using N2LO NNPDF4.0, CT18, MSHT20, ABMP16 with $\alpha_s(M_Z) = 0.118$, and ATLASpdf21. The ratio to the NNPDF4.0 central value and the relative 1σ uncertainty are shown for each parton combination.

PDF4LHC21 recommendation and combined PDFs

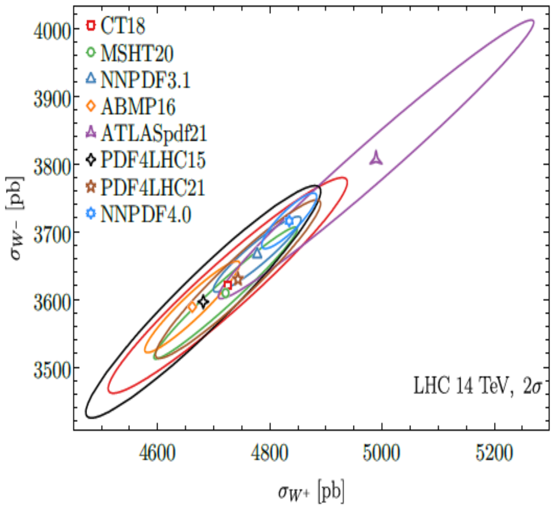
arXiv:2203.05506

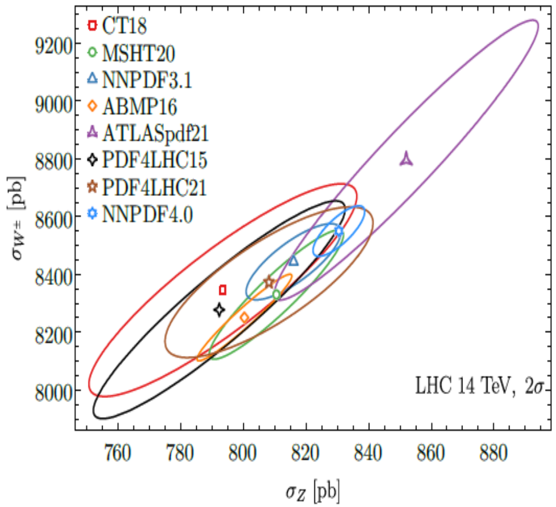
- A comprehensive recommendation for usage of PDFs at the LHC
- Replaces the PDF4LHC15 recommendation
- A detailed benchmarking comparison of global fits by three main groups
- Combined PDF4LHC21 NNLO PDFs based on CT18', MSHT20, and NNPDF3.1' ensembles. [The primes indicate minor changes in CT18 and NNPDF3.1 fits to maximize compatibility for the combination.]
- Suitable for BSM searches, measurements of moderate precision, theory predictions
- Provided as 40-member Hessian PDFs and 100-member Monte-Carlo PDFs of comparable accuracy

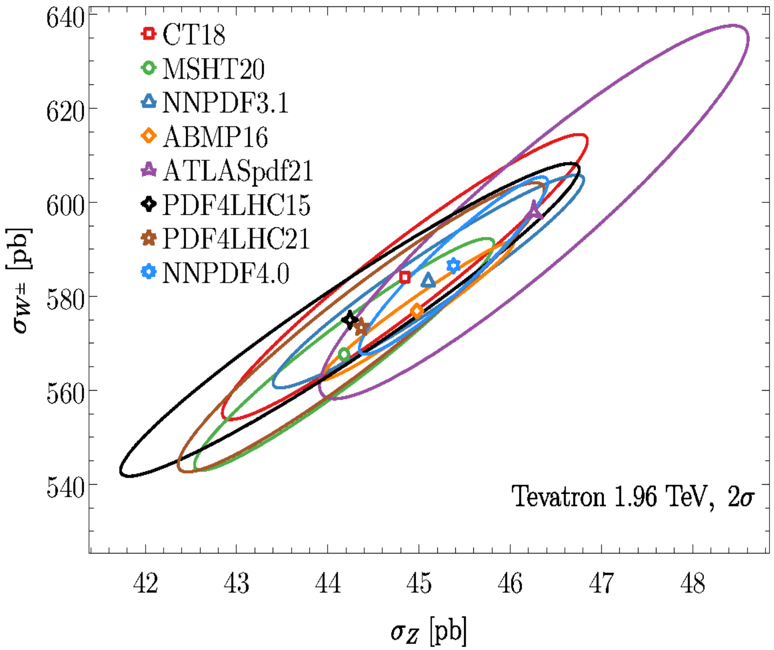


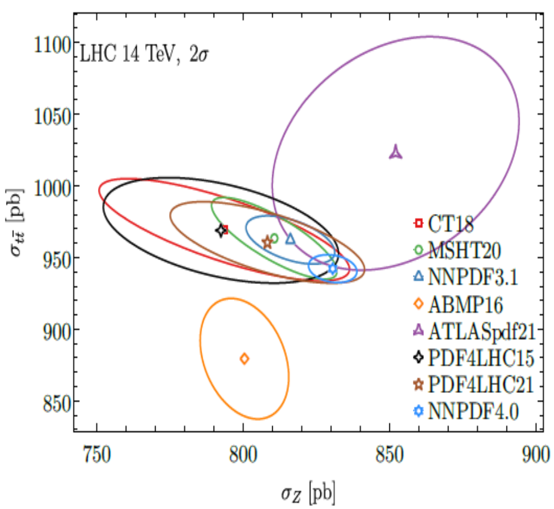


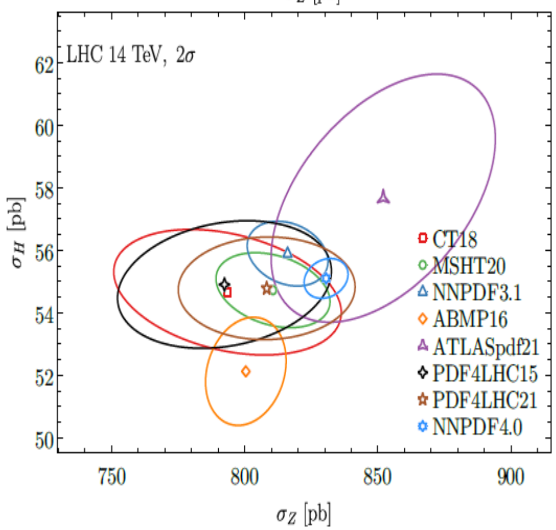
… predictions for LHC and Tevatron benchmark cross sections











PDF-related topics in Snowmass'13 [arXiv:1310.5189] and '21 studies

Topic	Status, 2013	Status, 2022
Achieved accuracy of PDFs	N2LO for evolution, DIS and vector boson produciton	N2LO for all key processes; N3LO for some processes
PDFs with NLO EW contributions	MSTW'04 QED, NNPDF2.3 QED	LuXQED and other photon PDFs from several groups; PDFs with leptons and massive bosons
PDFs with resummations	Small x (in progress)	Small-x and threshold resummations implemented in several PDF sets
Available LHC processes to determine nucleon PDFs	W/Z, single-incl. jet, high-p_T Z, $t\bar{t}$, $W+c$ production at 7 and 8 TeV	+ $t\bar{t}$, single-top, dijet, $\gamma/W/Z$ +jet, low-Q Drell Yan pairs, … at 7, 8, 13 TeV
Near-future experiments to probe PDFs	LHC Run-2 DIS: LHeC	LHC Run-3 DIS: EIC, LHeC, …
Benchmarking of PDFs for the LHC	PDF4LHC'2015 recommendation in preparation	PDF4LHC'21 recommendation issued
Precision analysis of specialized PDFs		Nuclear, meson, transverse-momentum dependent PDFs

NEW TASKS in the HL-LHC ERA:

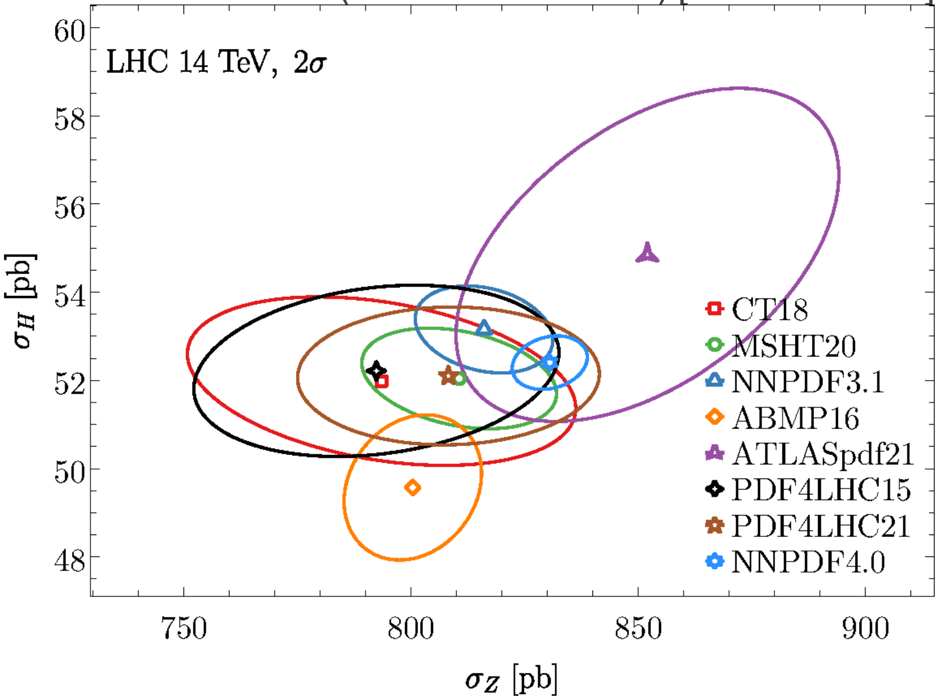
Obtain complete N2LO and N3LO predictions for PDF-sensitive processes	Improve models for correlated systematic errors	Find ways to constrain large-x PDFs without relying on nuclear targets
Develop and benchmark fast N2LO interfaces	Estimate N2LO theory uncertainties	New methods to combine PDF ensembles, estimate PDF uncertainties, deliver PDFs for applications

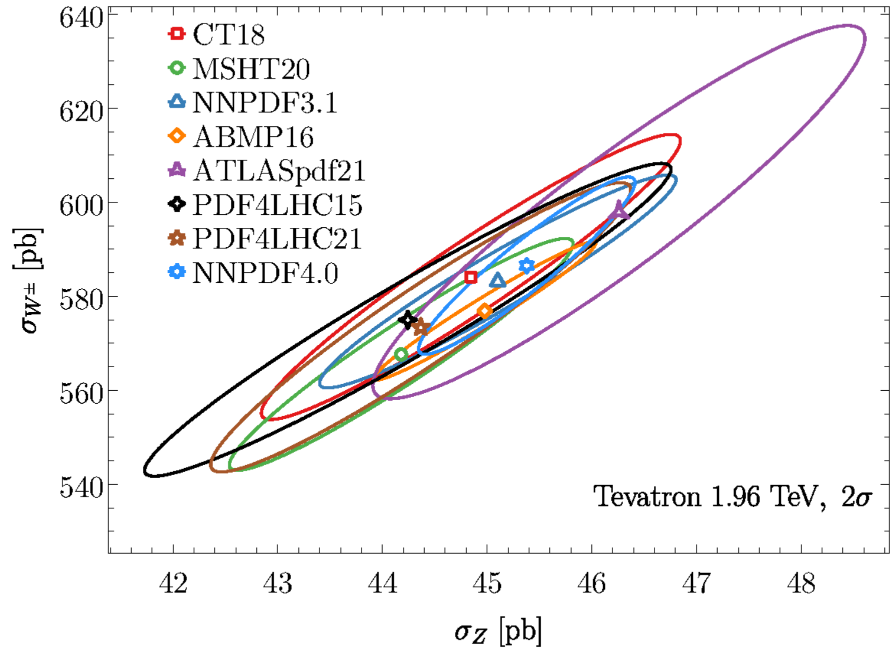
The tolerance puzzle

Courtoy, Huston, Nadolsky, Xie, Yan, Yuan, arXiv: [2205.10444](2205.10444)

Why do groups fitting similar data sets obtain different PDF uncertainties?

Precision PDFs (Snowmass 21 WP) [2203.13923v2]



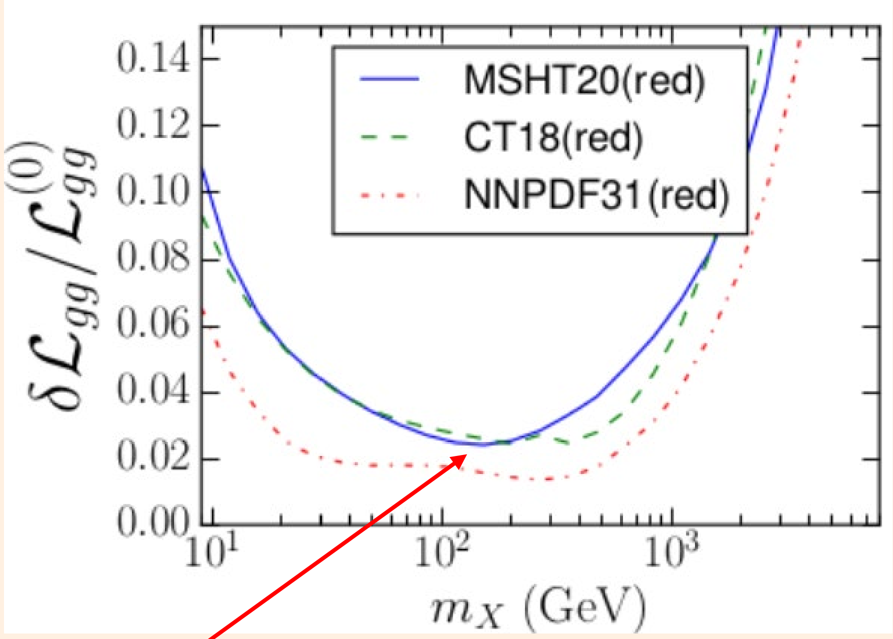


The answer has direct implications for high-stake experiments such as W boson mass measurement, tests of nonperturbative QCD models and lattice QCD, high-mass BSM searches, etc.

The tolerance puzzle

Relative PDF uncertainties on the gg luminosity at 14 TeV in three PDF4LHC21 fits to the **identical** reduced global data set

arXiv:2203.05506



× 1.5 – 2 difference

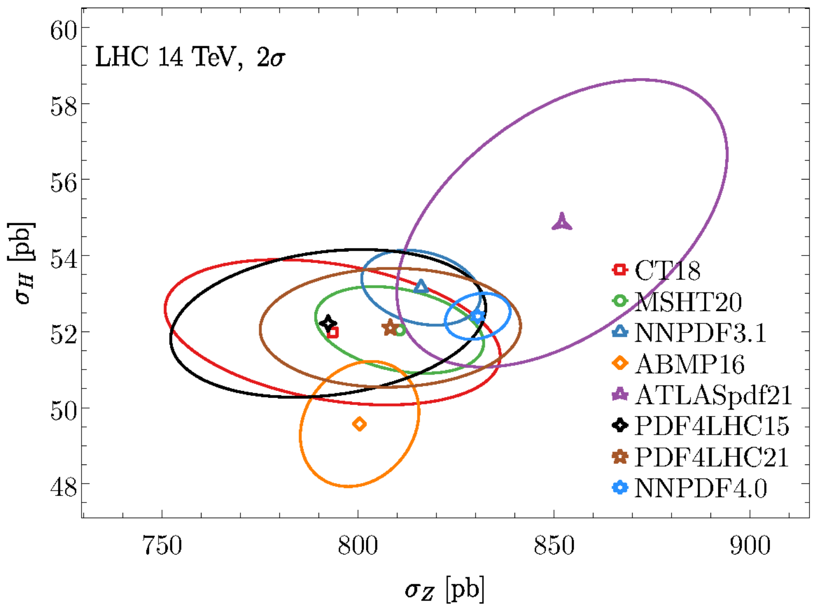
While the fitted data sets are identical or similar in several such analyses, the resulting PDF sets may differ because of methodological choices adopted by the PDF fitting groups.

NNPDF3.1' and especially 4.0 (based on the NN's+ MC technique) tend to give smaller uncertainties in data-constrained regions

Our findings I

- Large differences in uncertainty estimates can be due to **density of sampling** of multivariate probability distributions. This is a common issue reflecting geometry in many dimensions. It may lead to the "big data paradox" affecting also large population surveys (e.g, during the 2016 US presidential election) and quasi-MC integration.

[Xiao-Li Meng, https://tinyurl.com/XLMeng2019 and refs. below].



Sampling of PDF uncertainties for chosen cross sections is similar to population surveys.

Our findings II

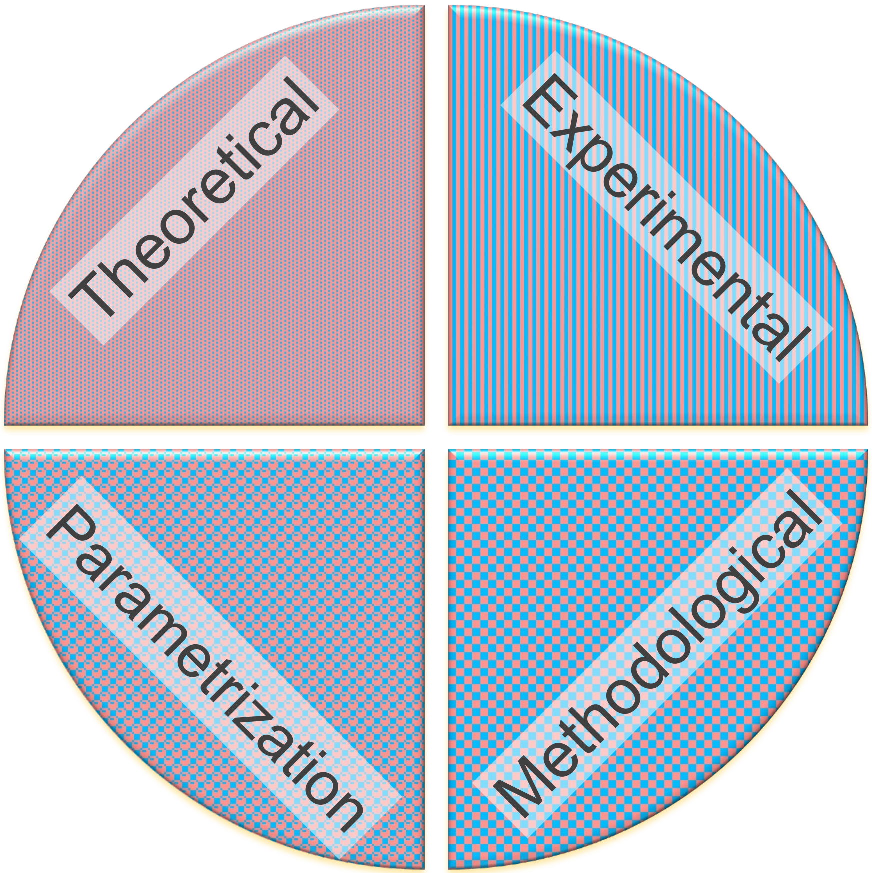
- **Bad news:** The tolerance puzzle is *intractable* in very complex fits
  - In a fit with N_{par} free parameters, the minimal number of PDF replicas to estimate the expectation values for $\forall\ \chi^2$ function grows as $N_{min} \geq 2^{N_{par}}$
  - Example: $N_{min} > 10^{30}$ for $N_{par} = 100$

[Sloan, Wo´zniakowski, 1997]

[Hickernell, MCQMC 2016, 1702.01487]

Good news: expectation values **for typical QCD observables** can be estimated with fewer replicas using a targeted sampling technique [a "**hopscotch scan**"]

Components of PDF uncertainty



In each category, one must maximize

PDF fitting accuracy (accuracy of experimental, theoretical and other inputs)

PDF sampling accuracy (adequacy of sampling of space of possible solutions)

NEW

Fitting/sampling classification is borrowed from the statistics of large-scale surveys [Xiao-Li Meng, *The Annals of Applied Statistics*, Vol. 12 (2018), p. 685]

Kovarik et al., arXiv: 1905.06957

Nature **v. 600** (2021) 695

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

https://doi.org/10.1038/s41586-021-04198-4

Received: 18 June 2021

Accepted: 29 October 2021

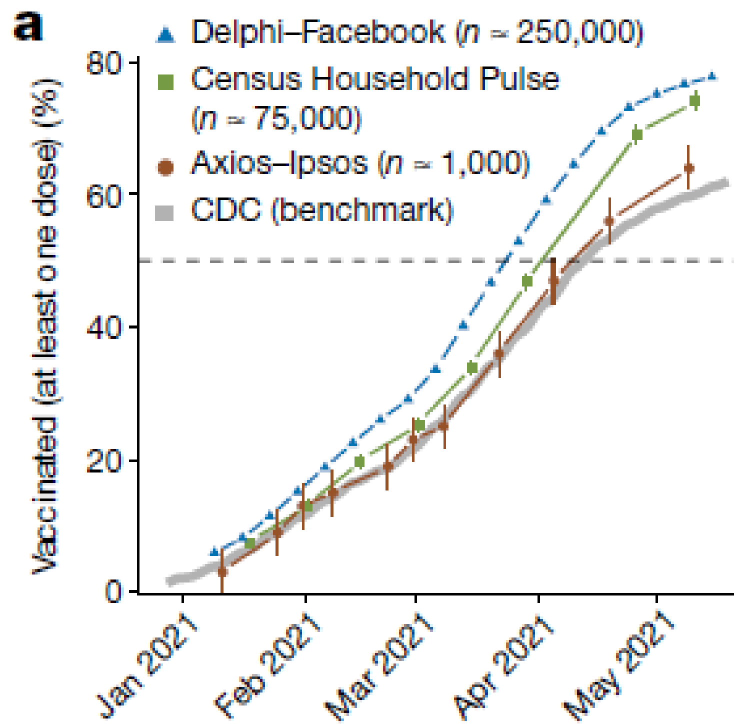
Published online: 8 December 2021

Check for updates

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5,6]

Surveys are a crucial tool for understanding public opinion and behaviour, and their accuracy depends on maintaining statistical representativeness of their target populations by minimizing biases from all sources. Increasing data size shrinks confidence intervals but magnifies the effect of survey bias: an instance of the Big Data Paradox[1]. Here we demonstrate this paradox in estimates of first-dose COVID-19 vaccine uptake in US adults from 9 January to 19 May 2021 from two large surveys: Delphi–Facebook[2,3] (about 250,000 responses per week) and Census Household Pulse[4] (about 75,000 every two weeks). In May 2021, Delphi–Facebook overestimated uptake by 17 percentage points (14–20 percentage points with 5% benchmark imprecision) and Census Household Pulse by 14 (11–17 percentage points with 5% benchmark imprecision), compared to a retroactively updated benchmark the Centers for Disease Control and Prevention published on 26 May 2021. Moreover, their large sample sizes led to miniscule margins of error on the incorrect estimates. By contrast, an Axios–Ipsos online panel[5] with about 1,000 responses per week following survey research best practices[6] provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework[1] to explain the inaccuracy in the three surveys. We then analyse the implications for vaccine hesitancy and willingness. We show how a survey of 250,000 respondents can produce an estimate of the population mean that is no more accurate than an estimate from a simple random sample of size 10. Our central message is that data quality matters more than data quantity, and that compensating the former with the latter is a mathematically provable losing proposition.

The Big Data Paradox in vaccine uptake



Surveys of the COVID-19 vaccination rate with very large samples of responses and small statistical uncertainties (Delphi-Facebook) greatly overestimated the actual vaccination rate published by the Center for Disease Control (CDC) after some time delay.

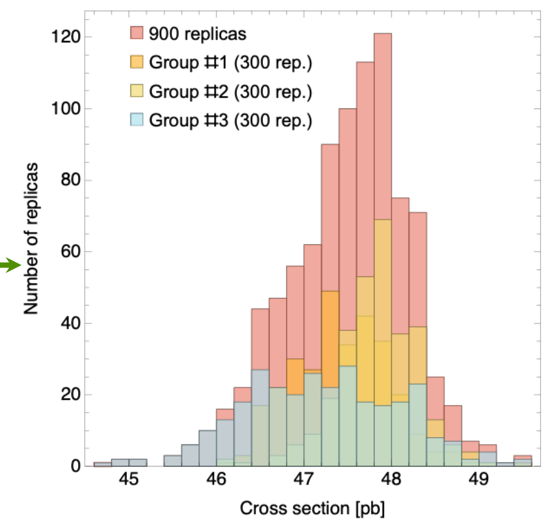
The discrepancy has been traced to the **sampling bias**. In contrast to the statistical error, the sampling bias can **grow** with the size of the sample.

Law of large numbers

With an increasing size of sample $n \to \infty$, under a set of hypotheses, it is usually expected that the sample *deviation* on an observable μ decreases as

$$\boxed{\mu - \hat{\mu} \propto \sigma_{std}/\sqrt{n}}$$

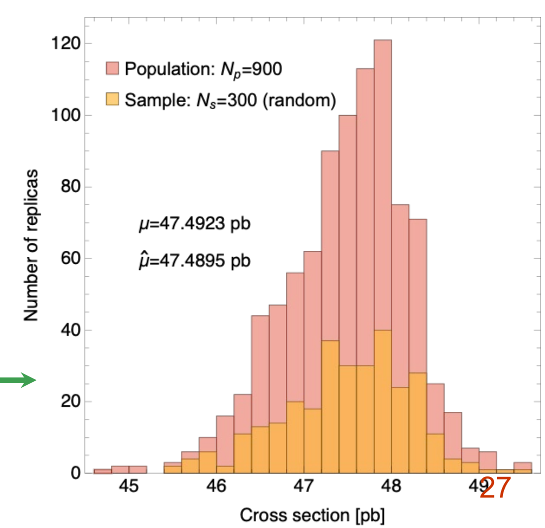
with σ_{std} the standard variation, μ and $\hat{\mu}$ the true and sample expectation values. *This is the law of large numbers.*



A toy sampling exercise

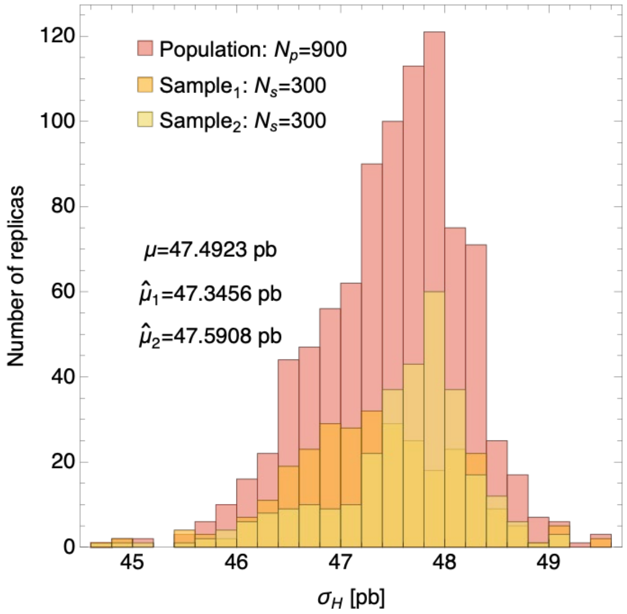
We take 300×3 groups of Higgs cross sections evaluated by 3 different groups (CT18', MSHT20, NNPDF3.1').

We **randomly** select 300 out of the 900 cross sections. The law of large numbers is fulfilled in this case: there is no bias.



Trio identity

If we **bias** the selection by taking 200 items from one group and 100 from another, the deviation $\mu - \hat{\mu}$ is no longer proportional to $\sigma_{std}/\sqrt{n}$!



Quality of the sample is as important as quantity.

The **trio identity** identifies three main contributions to the sample deviation:

$$\mu - \hat{\mu} = (\textit{confounding correlation}) \times (\textit{measure discrepancy}) \times (\textit{inherent problem difficulty})$$

This identity originates from the statistics of large-scale surveys
[Xiao-Li Meng, The Annals of Applied Statistics, Vol. 12 (2018), p. 685]

Trio identity, continued

A sample of n items from a population of size N can be described by an array R_j of sampling indicators =0 or 1, which shows that

$$\mu - \hat{\mu} = \underbrace{\mathrm{Corr}[observable, sampling\ algorithm]}_{\text{depends on the sampling algorithm}} \times \underbrace{\sqrt{\frac{N}{n} - 1} \times \sigma_{std}(observable)}_{\text{decreases as } \sigma_{std}/\sqrt{n} \text{ for random sampling}}$$

[X.-L. Meng, The Annals of Applied Statistics, Vol. 12 (2018), p. 685]
[Hickernell, MCQMC 2016, 1702.01487]

Consequences for large N (or large N_{par}):

1. The sample deviation can be large if Corr[…] does not decrease as $o(1/\sqrt{N})$
2. Standard error estimates can be misleadingly small.
3. **Control for sampling biases is critical** to avoid the situation described as the **Big Data Paradox** [Meng]:

The bigger the data, the surer we fool ourselves.

Multivariate parametric forms

A typical PDF set may depend on tens to several hundreds of free parameters

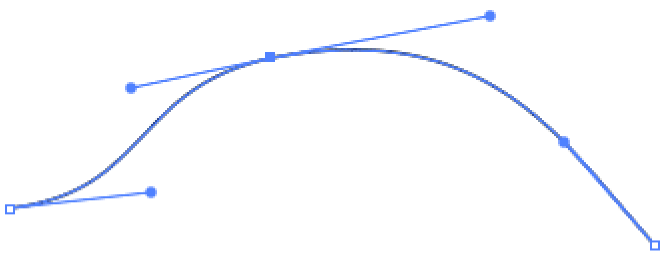
PDF functional forms must be flexible to accommodate a variety of behaviors
CT18 parametrizations at initial scale Q_0 are given by

$$\mathrm{f_a}\,(\mathrm{x}, \mathrm{Q_0}) = \mathrm{A}x^{a_1}(1-\mathrm{x})^{\mathrm{a_2}}\mathrm{B_a^{(n)}}(\mathrm{x}; \mathrm{a_3}, \mathrm{a_4}, \ldots)$$

$$\mathrm{B_a^{(n)}}(x) = \sum_{k=0}^{n} a_{k+2}\binom{n}{k} x^k (1-x)^{n-k}$$

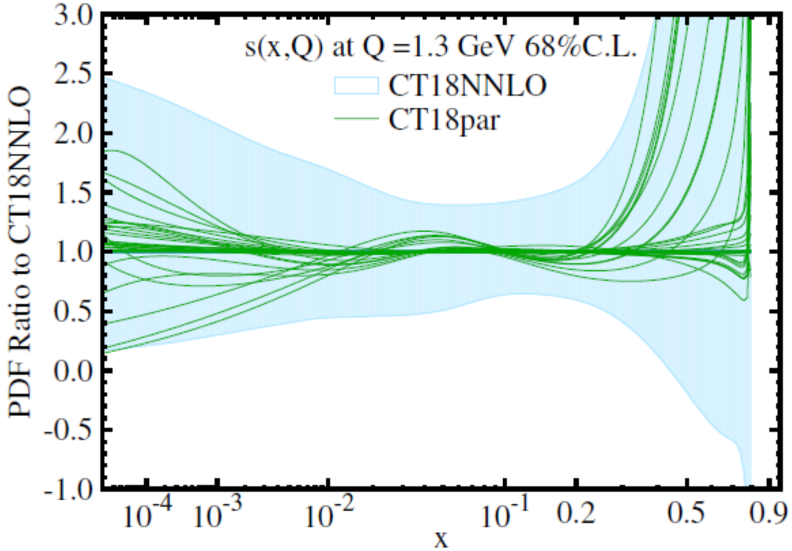
are **Bézier curves** – flexible polynomials familiar from vector graphics programs

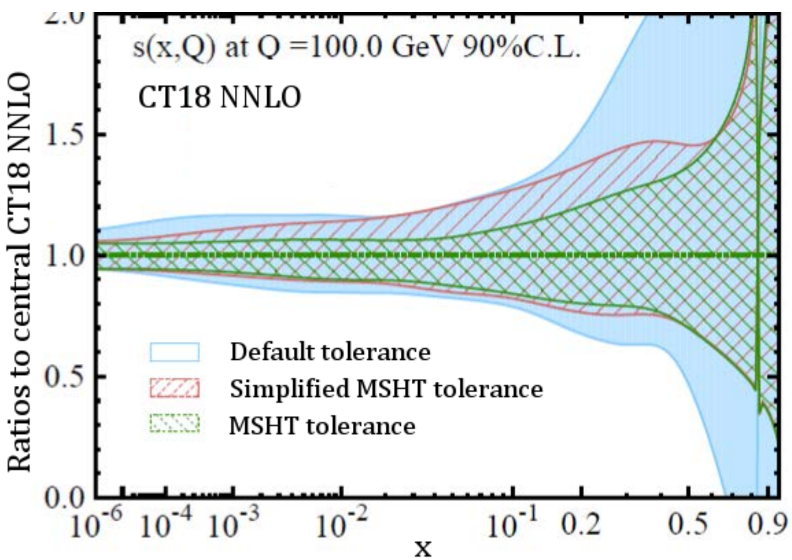
Bézier curves can mimic a variety of behaviors of PDFs and their uncertainties. A powerful alternative to neural networks!



[A. Courtoy, P. N., arXiv: 2011.10078]

Sampling of PDF parametrizations in global fits





Upper figure: A large part of the CT18 PDF uncertainty accounts for the sampling over 250-350 parametrization forms, possible choices of fitted experiments and fitting parameters, definitions of χ^2

Lower figure: this approach sometimes enlarges the uncertainties compared to the other groups, reflecting the chosen goodness-of-fit (tolerance) criterion more than the strength of experimental constraints

However, more restrictive tolerance criteria elevate the risk of sampling biases.

Easier to examine these issues **for specific QCD observables** than in abstract

Hopscotch scans:
estimation of the PDF sampling uncertainty on a QCD cross section σ_{QCD}

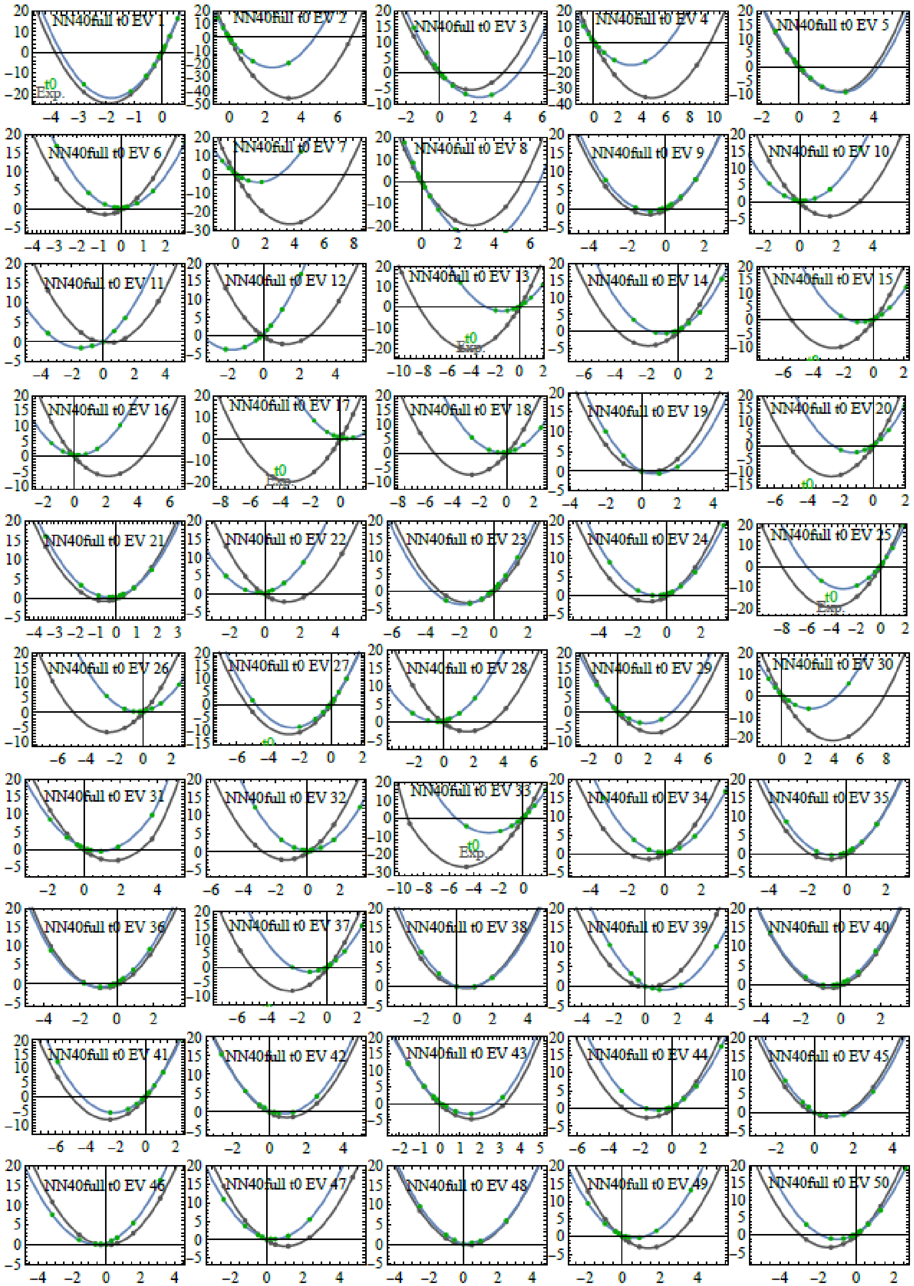
The release of a public code for NNPDF4.0's new methodology provides a perfect playground to explore the role of sampling.
[NNPDF, EPJC 81]

To sample the PDF dependence: sample primarily the coordinates with large variations of σ_{QCD}. We employ:

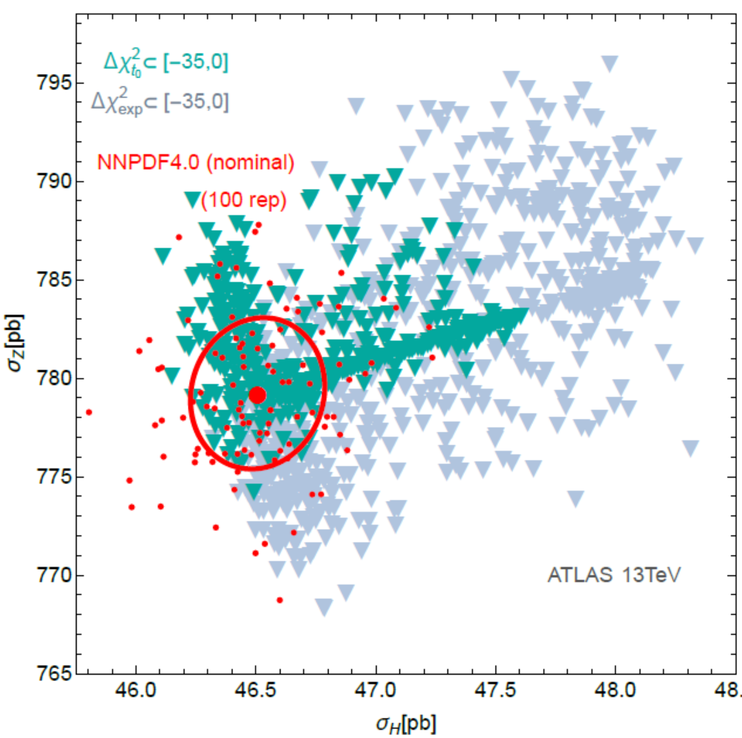
1. Basis coordinates in space of MC replicas. Naturally provided by the NNPDF4.0 Hessian set.
2. Knowledge of 4-8 "large dimensions" in PDF space controlling variation of σ
3. A moderate number of MC PDF replicas varying primarily in these directions

Based on the ideas of [Hickernell, MCQMC 2016, 1702.01487] [Sloan, Wo´zniakowski, 1997]

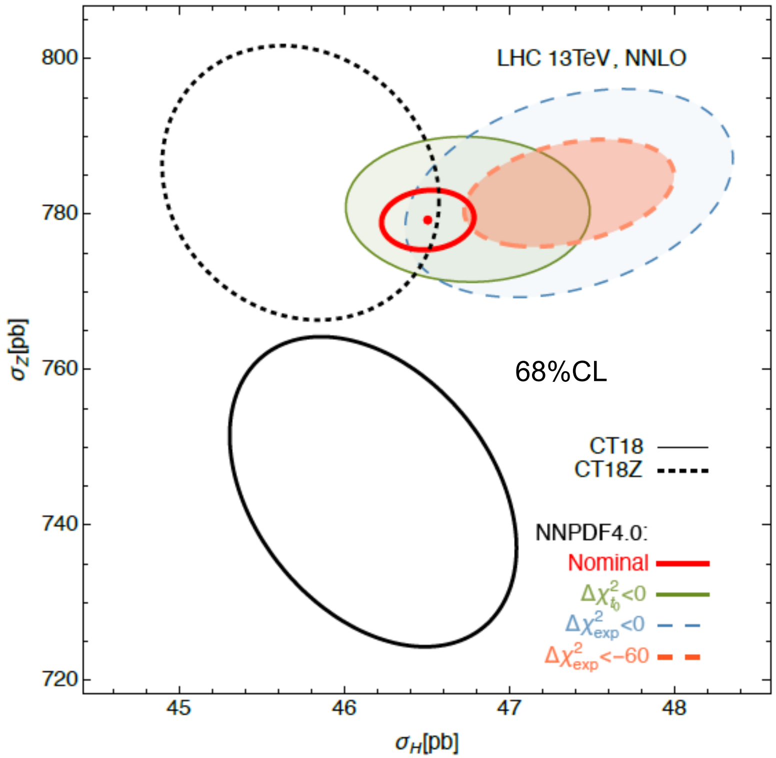
How the hopscotch solutions are found

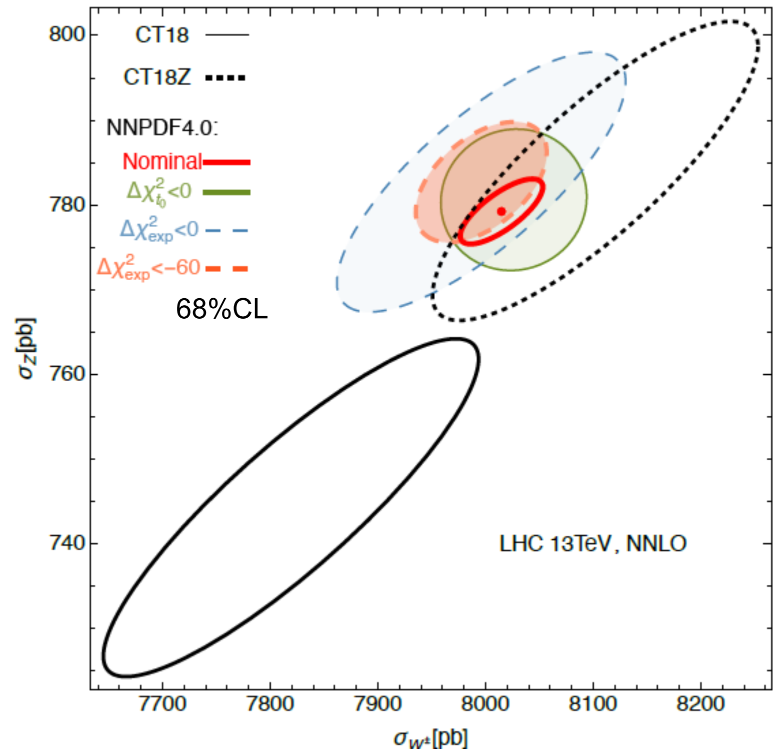


1. Examine the quasi-Gaussian χ^2 dependence along 50 Hessian EV directions
2. Perform high-density MC sampling of a span of a few EV directions that drive the specific PDF uncertainty



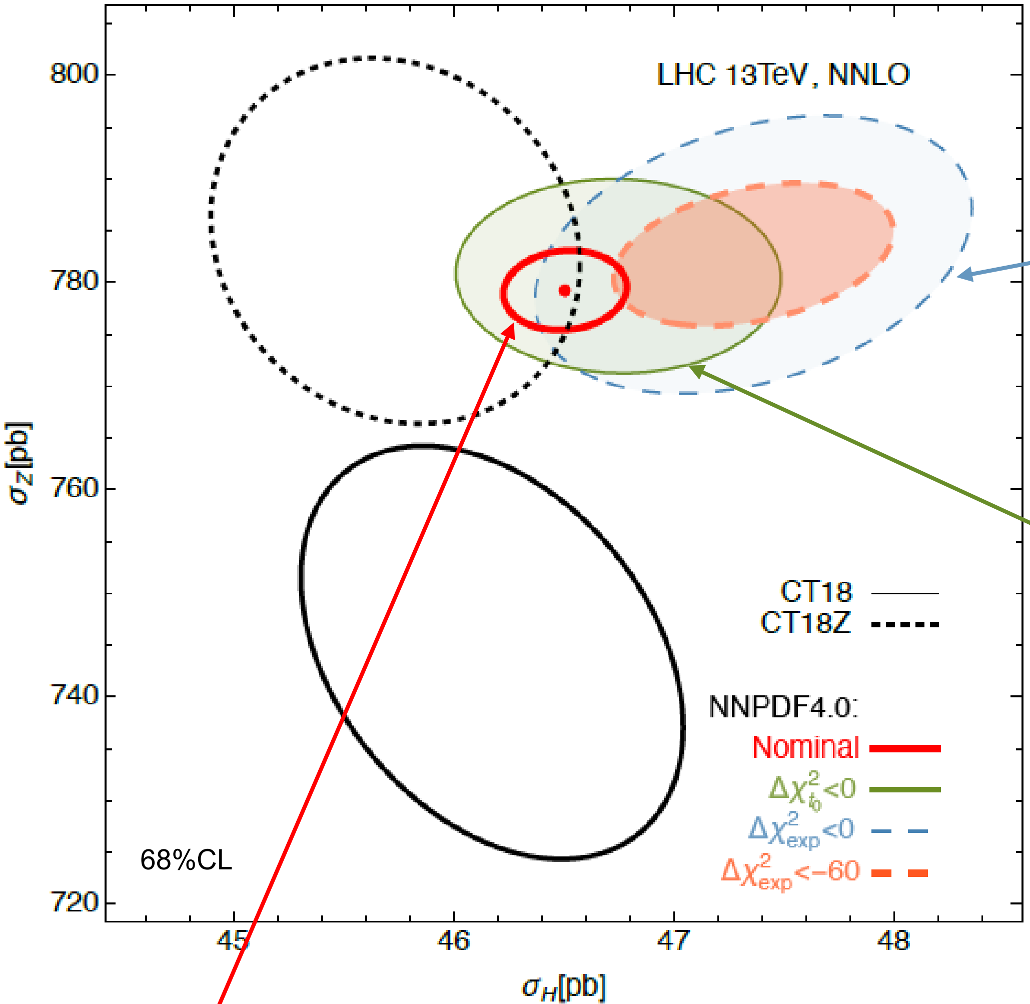
Monte-Carlo sampling of PDF parametrizations





Using the public NNPDF4.0 fitting code, we find well-behaving PDF solutions to the NN4.0 fit that have better χ^2 with respect to central data values (by as much as 35-80 units depending on the χ^2 definition) than the published replica 0. These replicas follow a regular pattern. They lie outside of the nominal (red) NN4.0 uncertainties.

Monte-Carlo sampling of PDF parametrizations



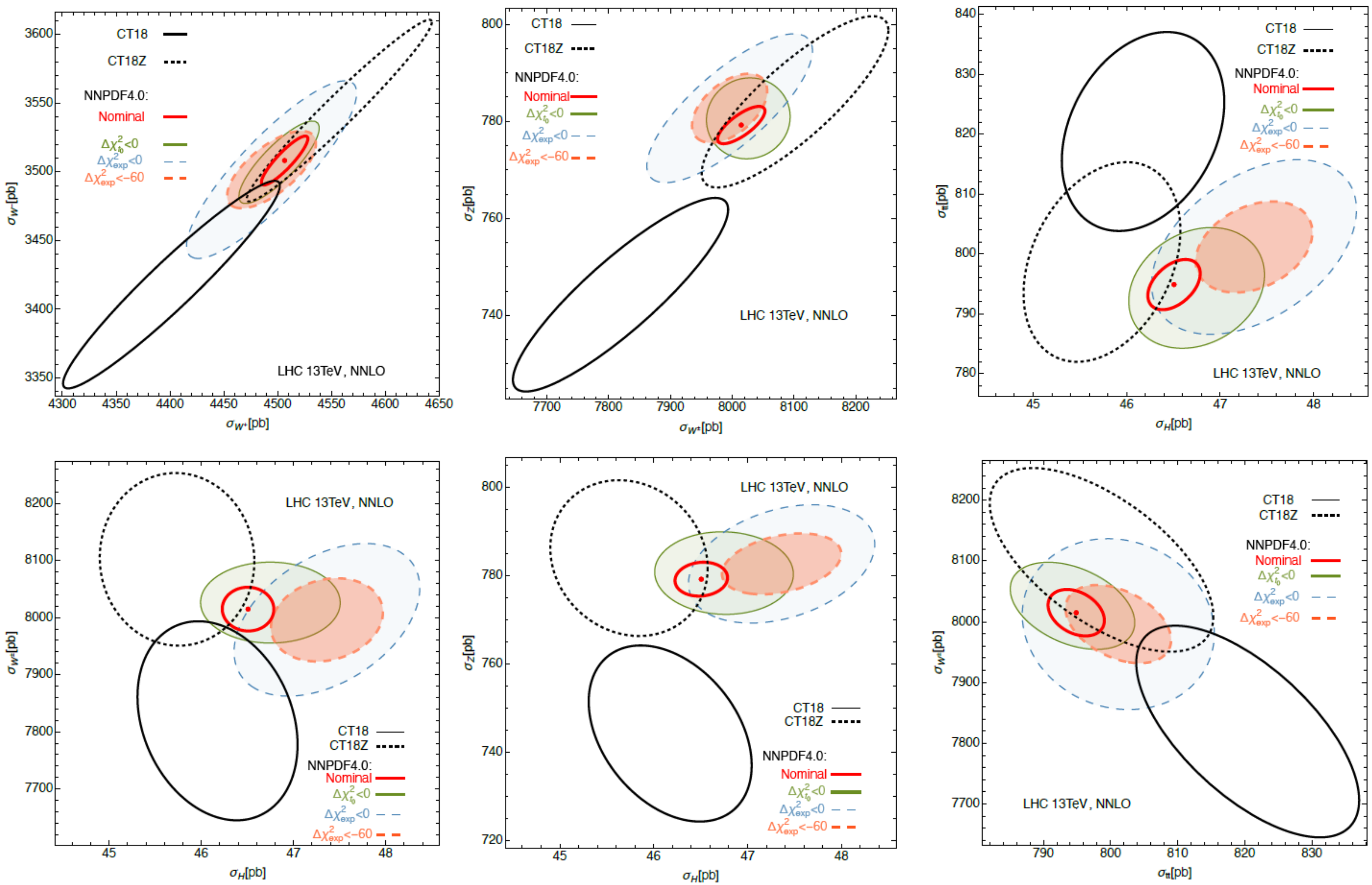
Regions containing (very) good solutions according to the experimental form of χ^2 (is used in χ^2 summary tables of the NN4.0 article, was a default in the NN4.0 public code)

Region containing good solutions according
to the most recent t_0 form of χ^2
(used to train NN4.0 replicas)

Nominal NN4.0 Hessian or MC 68%cl

These regions are approximate, at least as large as shown

The hopscotch scans: NNPDF4.0 vs CT18 uncertainties

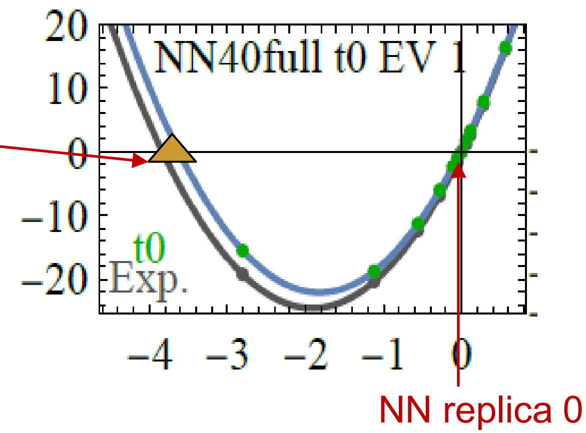


Ellipses at 68% CL

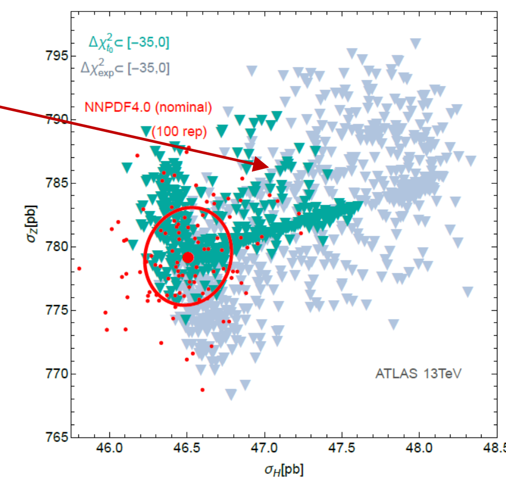
Hopscotch NN4.0 replicas

LHAPDF6 grids available at https://ct.hepforge.org/PDFs/2022hopscotch/

1. Alternative (second) EV sets with $\Delta\chi^2 = 0$, for 50 EV directions



2. A total 2329 PDF sets from hopscotch scans on $\sigma_Z, \sigma_{W^+}, \sigma_{W^-}, \sigma_H, \sigma_{t\bar{t}}$ total inclusive cross sections at the LHC 13 TeV

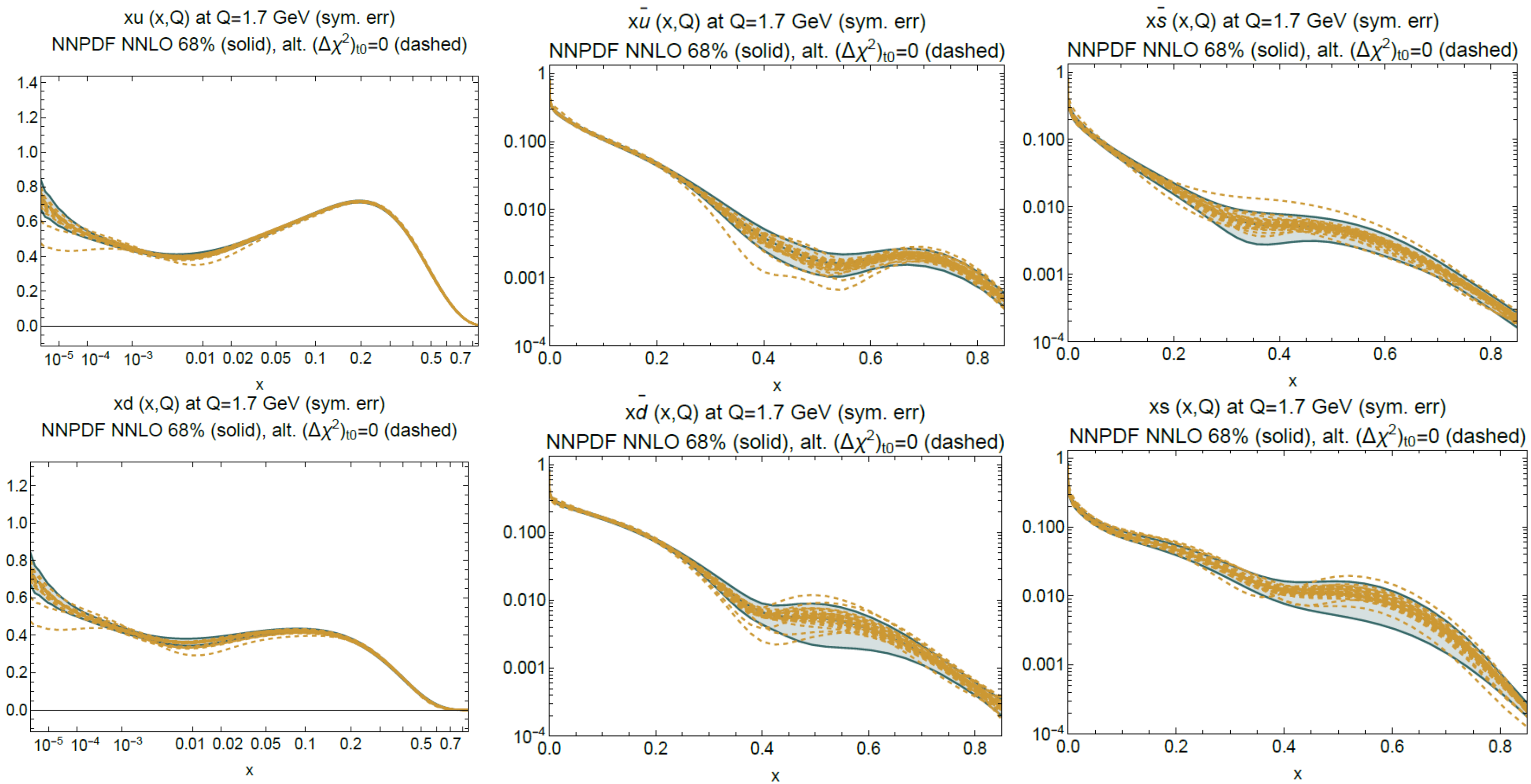


For $\chi^2_{t_0}$ and χ^2_{exp} definitions in the NNPDF4.0 code

Codes to generate LHAPDF grids for hopscotch replicas available by request.

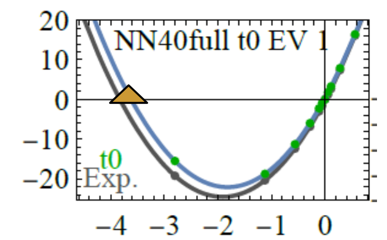
Hopscotch NN4.0 replicas

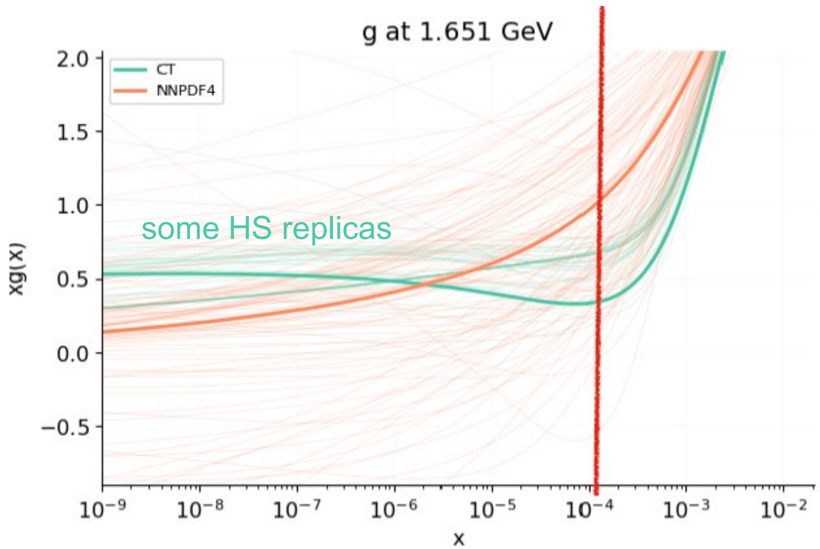
Error bands available at https://ct.hepforge.org/PDFs/2022hopscotch/



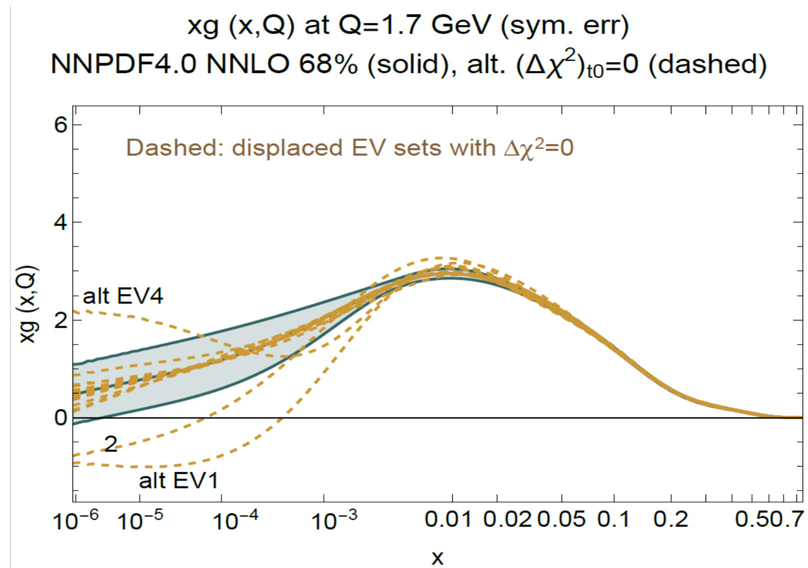
Nominal NN4.0 1σ bands and alternative $\Delta\chi^2_{t_0} = 0$ EV sets

Why doesn't NNPDF4.0 find HS solutions?





NNPDF authors find that some HS replicas fail the initial-stage overfitting test
(M. Ubiali, HP2 2022 workshop, Durham, 2022-09-22)



HS solutions have much lower χ^2 than NN MC replicas. HS PDFs are outside the 50-dim neighborhood of NN replica 0. We do not see evidence of "overfitting" according to CT18 criteria.

From arXiv: 2205.10444 v.3 , Sec. 3D

If the hopscotch solutions are acceptable, a natural question to raise is why they are not covered by the nominal NNPDF set. … As a possible hint, any hopscotch solution can be represented by a neural network in accord with the universal approximation theorems. The challenge of representative sampling in a high-dimensional space must therefore be also present in the NN approach. The nominal NNPDF replicas only resample the fitted data points while using a fixed methodology, with specific choices made on the NN architecture, the cost function, stopping and smoothness conditions. Finding a hopscotch solution in an NN approach may require variations in the training methodology, … which may thus constitute an unstated part of the uncertainty, together with the uncertainty due to the prescription for experimental systematic errors. The closure test … checks for the agreement of the PDFs with the pseudodata within the uncertainties. Yet it does not establish the full size of uncertainties in all directions, and neither it rules out potential subtle biases with the real data…

Hopscotch replicas enlarge the error bands

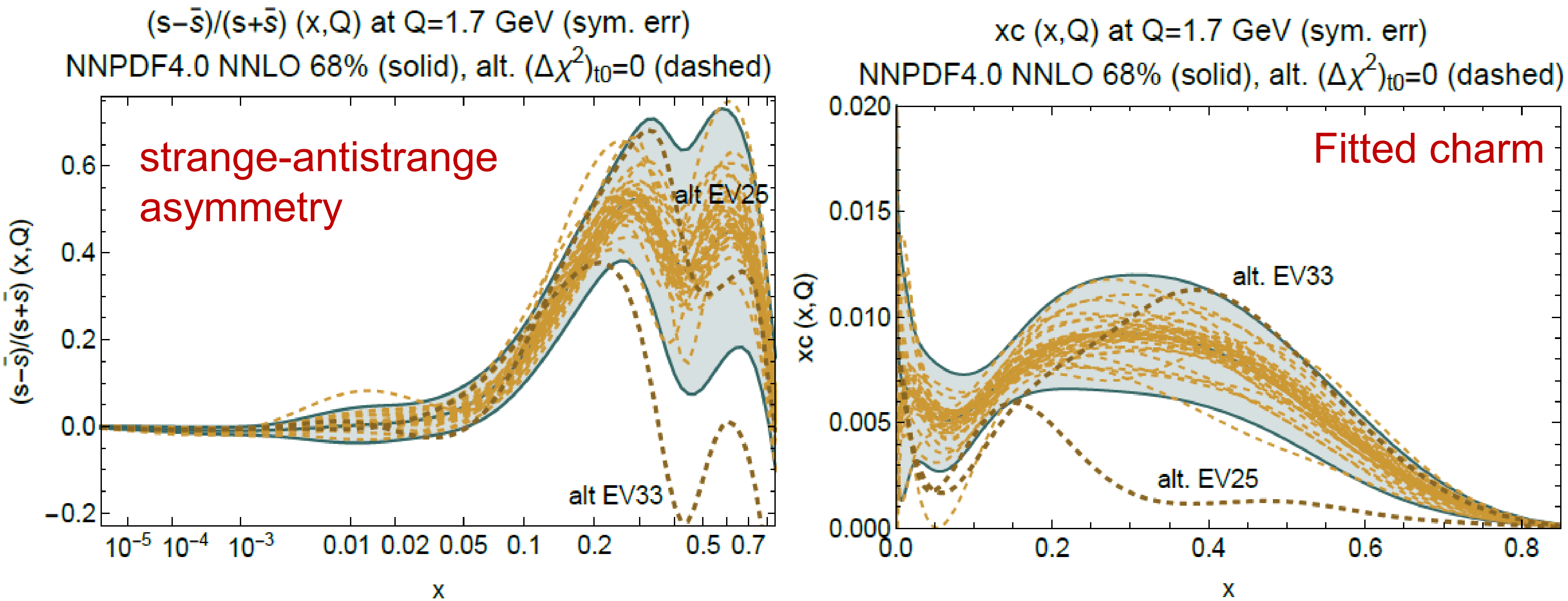
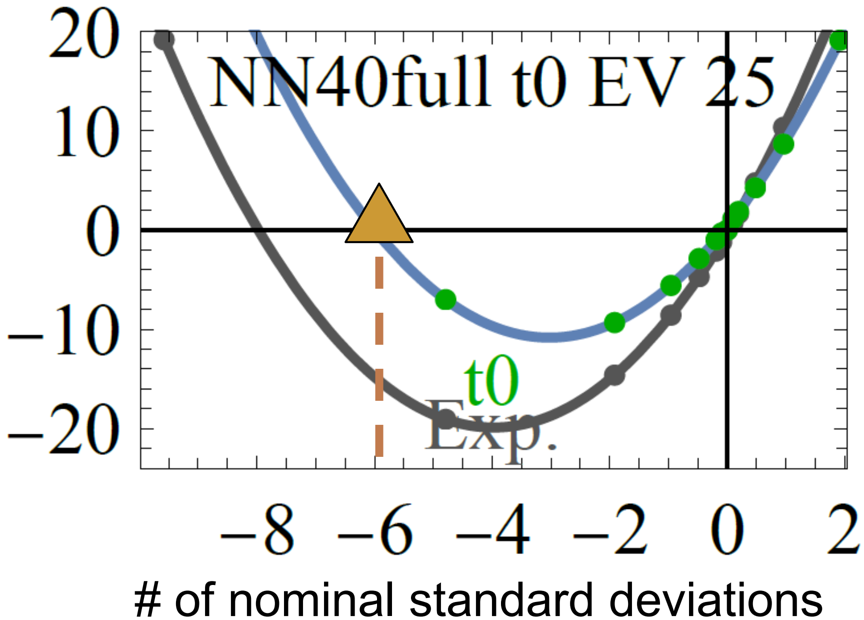
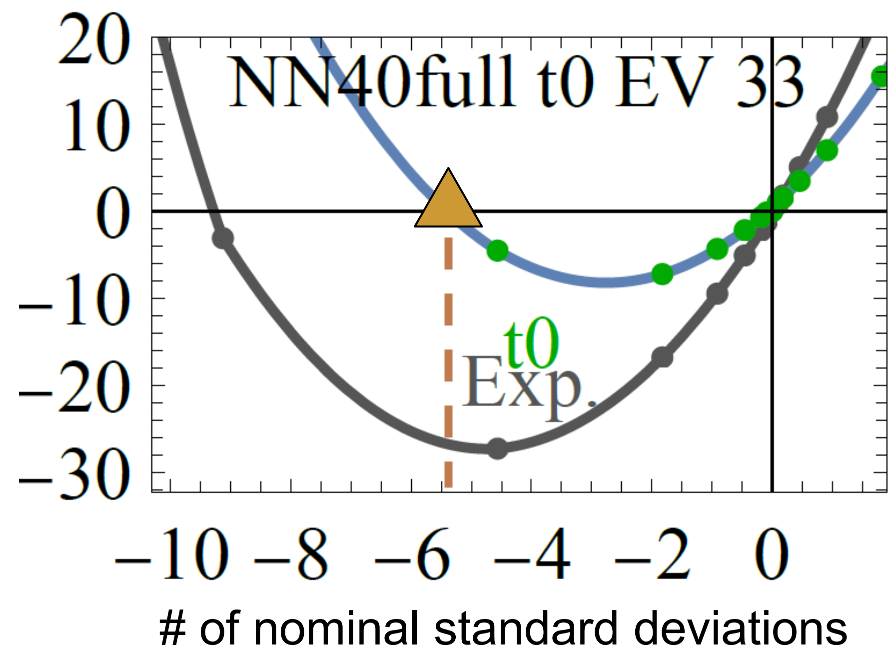


FIG. 9. Solid bands indicate the nominal 68% NNPDF4.0 uncertainties for strangeness asymmetry (left) and charm PDF (right) at $Q = 1.7$ GeV. The alternative EV sets with $\Delta\chi^2_{t_0} = 0$ are plotted as dashed lines.

At $x > 0.2$, $Q \approx Q_0 = 1.51$ GeV, the HS replicas reduce significance of $(s - \bar{s})/(s + \bar{s})$ $\approx 50\%$ (left) and $c(x, Q) \neq 0$ (right). This washes out the 3σ evidence for the "intrinsic charm" stated in R. Ball et al., Nature 608 no. 7923, (2022) 483.

Scans of the log-likelihood in EV directions 25 and 33





Fitted charm, intrinsic charm…

Are twist-2 NNLO contributions sufficient for describing the most precise experiments?

References:

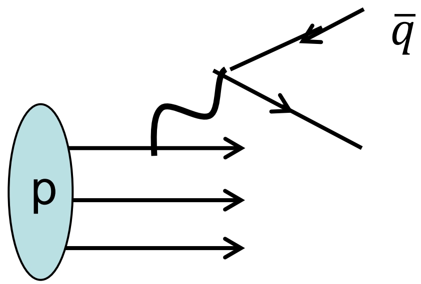
1. T.-J. Hou et al., JHEP 02 (2018) 059; 57 pages, 19 figures: QCD factorization with the NP charm and CT14 IC NNLO pheno analysis
2. M. Guzzi, T. J. Hobbs, K. Xie, et al., arXiv:2210.XXXXX; 10 pages: **new** CT18 IC analysis with the LHC Run-1 and 2 data

Models of the IC:

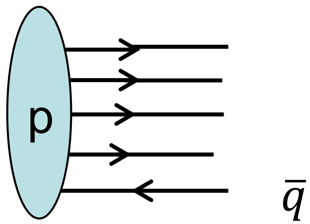
1. BHPS: Brodsky, Hoyer, Peterson, Sakai, PLB 93 (1980) 451
2. BHPS3: Bluemlein, PLB 753 (2016) 619
3. Meson-Baryon Cloud models (MBM): Hobbs, Londergan, Melnitchouk, PRD 89 (2014) 074008

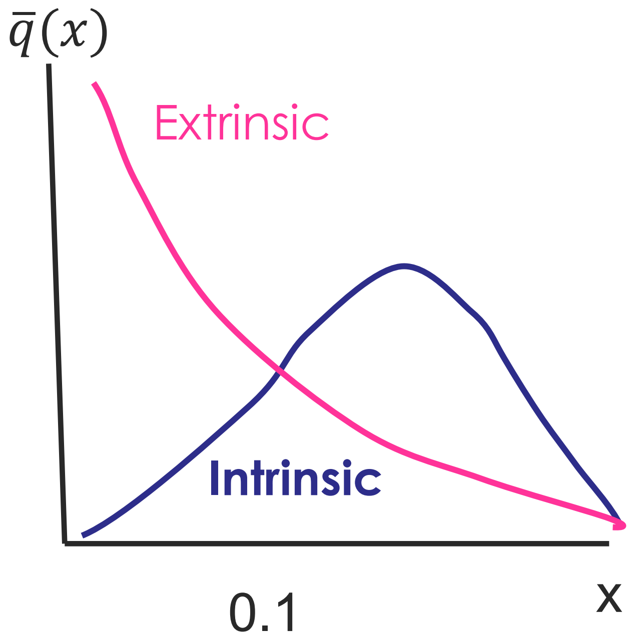
Extrinsic and intrinsic sea PDFs in nonperturbative models

"Extrinsic" sea
[maps on leading-power sea production from light flavors]



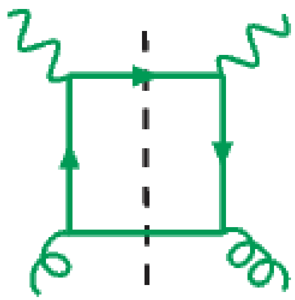
"Intrinsic" sea (excited Fock nonpert. states; beyond the leading-power production)



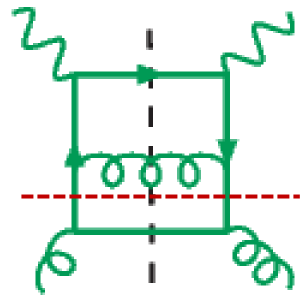


A twist-4 contribution in HERA DIS charm production ($\subset$ "intrinsic charm")

Twist-2
$\gamma^* g \to c\bar{c}$



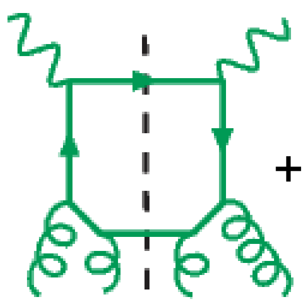
Order $\alpha_s(Q)$



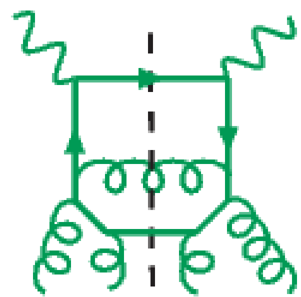
$\alpha_s^2(Q) \cdot \ln(Q^2/m_c^2)$

A ladder; must be resummed in $c(x,Q)$ in the $N_f = 4$ scheme at $Q^2 \gg m_c^2$; e.g., in the ACOT scheme

Twist-4
$\gamma^*(gg) \to c\bar{c}$



+…



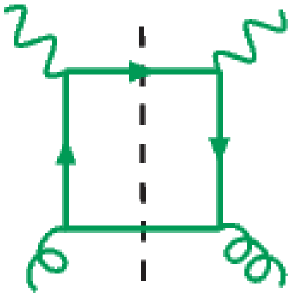
$\Lambda \lesssim 1$ GeV

$\alpha_s^2(Q) \cdot (\Lambda^2/Q^2)$
or $\alpha_s^2(Q) \cdot (\Lambda^2/m_c^2)$

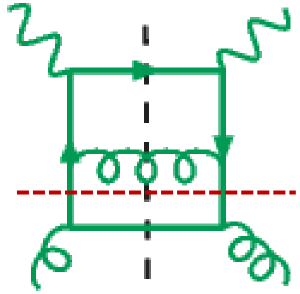
$\alpha_s^3(Q) \cdot (\Lambda^2/m_c^2)\ln(Q^2/m_c^2)$

A twist-4 contribution in HERA DIS charm production ($\subset$ "intrinsic charm")

Twist-2
$\gamma^* g \to c\bar{c}$



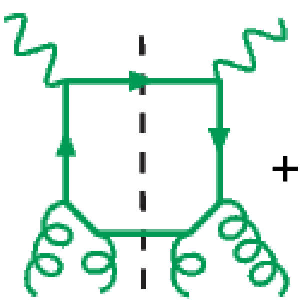
Order $\alpha_s(Q)$



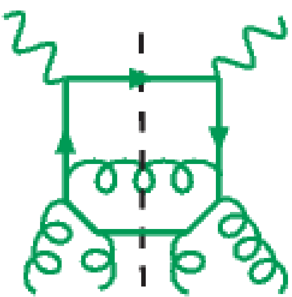
$\alpha_s^2(Q)\cdot\ln(Q^2/m_c^2)$

A ladder; must be resummed in $c(x,Q)$ in the $N_f = 4$ scheme at $Q^2 \gg m_c^2$; e.g., in the ACOT scheme

Twist-4
$\gamma^*(gg) \to c\bar{c}$



+…



$\Lambda \lesssim 1$ GeV

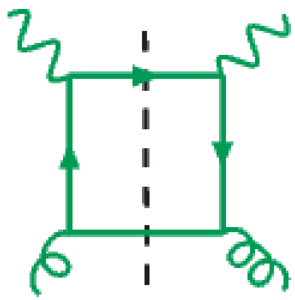
$\alpha_s^2(Q)\cdot(\Lambda^2/Q^2)$
or $\boxed{\alpha_s^2(Q)\cdot(\Lambda^2/m_c^2)}$

$\alpha_s^3(Q)\cdot(\Lambda^2/m_c^2)\ln(Q^2/m_c^2)$

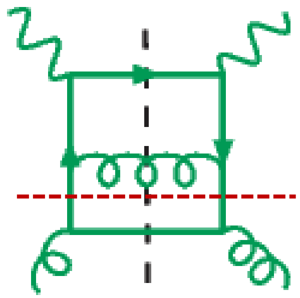
Can be of order 10% of the twist-2 α_s^2 term

A twist-4 contribution in HERA DIS charm production ($\subset$ "intrinsic charm")

Twist-2
$\gamma^* g \to c\bar{c}$



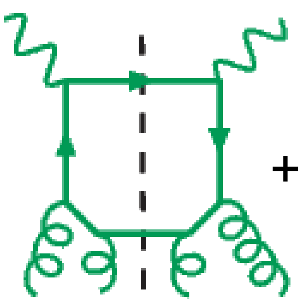
Order $\alpha_s(Q)$



$\alpha_s^2(Q) \cdot \ln(Q^2/m_c^2)$

A ladder; must be resummed in $c(x,Q)$ in the $N_f = 4$ scheme at $Q^2 \gg m_c^2$; e.g., in the ACOT scheme

Twist-4
$\gamma^*(gg) \to c\bar{c}$

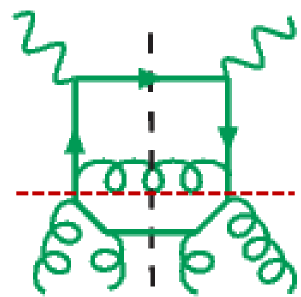


+…

$\alpha_s^2(Q) \cdot (\Lambda^2/Q^2)$
or $\boxed{\alpha_s^2(Q) \cdot (\Lambda^2/m_c^2)}$

$\Lambda \lesssim 1$ GeV

Can be of order ~10% of the twist-2 α_s^2 term



$\alpha_s^3(Q) \cdot (\Lambda^2/m_c^2) \ln(Q^2/m_c^2)$

The ladder subgraphs can be resummed as a part of $c(x,Q)$ in the $N_f = 4$ scheme at $Q^2 \gg m_c^2 > \Lambda^2$;

contributes to the boundary condition for $c(x,Q_0)$ at $Q_0 \approx m_c$;

obeys twist-2 DGLAP equations.

CT18 IC study: answers to important questions

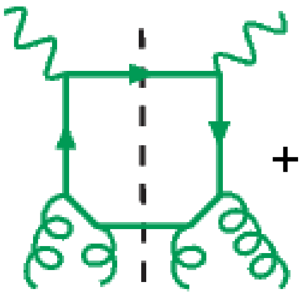
What are phenomenological constraints on the “intrinsic charm” from the global QCD data?

⇒ The CT18 charm PDFs allow a “nonperturbative” component carrying a total momentum fraction of $< 1\%$ at $Q \approx m_c$.

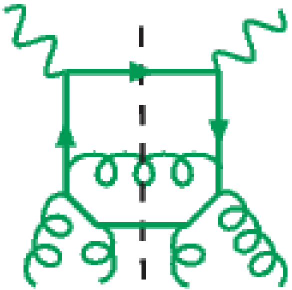
Can we estimate its impact on the LHC predictions?

Yes, based on the simplest approximation of the “nonperturbative” charm contribution.

Twist-4
$\gamma^*(gg) \to c\bar{c}$



+…



$\alpha_s^2(Q)\cdot(\Lambda^2/Q^2)$
or $\boxed{\alpha_s^2(Q)\cdot(\Lambda^2/m_c^2)}$

$\alpha_s^3(Q)\cdot(\Lambda^2/m_c^2)\ln(Q^2/m_c^2)$

Note:
“intrinsic charm” ≠ “fitted charm”

PDF fits may include a ``fitted charm'' PDF

``Fitted charm'' = ``higher-twist charm''
+ other (possibly not universal)
higher O(α_s) / higher power terms

QCD factorization theorem for DIS structure function $F(x,Q)$ [Collins, 1998]:

All α_s orders:

$$F(x,Q) = \sum_{a=0}^{N_f} \int_x^1 \frac{d\xi}{\xi} \mathcal{C}_a \left(\frac{x}{\xi}, \frac{Q}{\mu}, \frac{m_c}{\mu}; \alpha(\mu) \right) f_{a/p}(\xi,\mu) + \mathcal{O}(\Lambda^2/m_c^2, \Lambda^2/Q^2).$$

The PDF fits implement this formula up to (N)NLO ($N_{ord} = 1$ or 2):

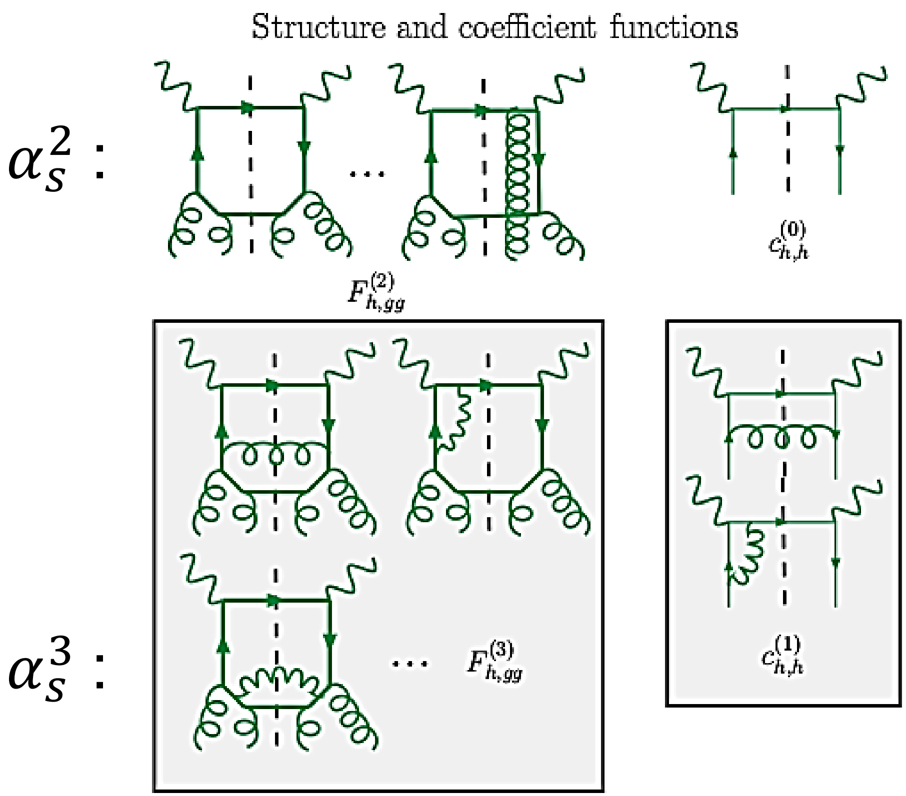
PDF fits:

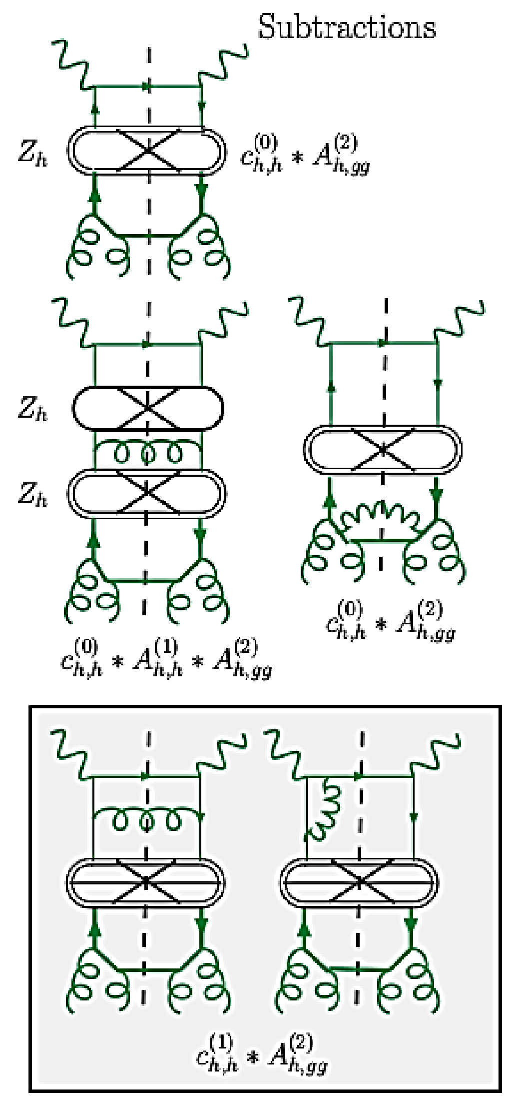
$$F(x,Q) = \sum_{a=0}^{N_f} \int_x^1 \frac{d\xi}{\xi} \mathcal{C}_a^{(N_{ord})} \left(\frac{x}{\xi}, \frac{Q}{\mu}, \frac{m_c}{\mu}; \alpha(\mu) \right) f_{a/p}^{(N_{ord})}(\xi,\mu).$$

The perturbative charm PDF component cancels at $Q \approx m_c$ up to a higher order

The 'fitted charm component' may approximate for missing terms of orders α_s^p with $p > N_{ord}$, or Λ^2/m_c^2, or Λ^2/Q^2

ACOT-like factorization for twist-4 charm contributions (an example)



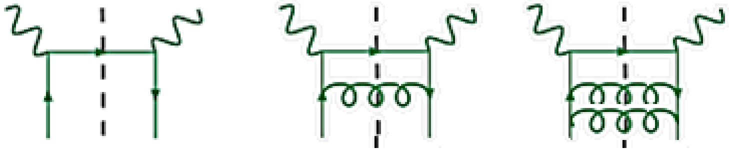


Many novel terms arise, involving twist-4 nonperturbative functions!

Intrinsic charm contributions, practical implementation

Keep only $c_{h,h} \otimes f_h$:

Discard $C^{(k)}_{h,gg} \otimes f_{gg}$, etc.



In the absence of full computation, we (and other groups) make the simplest approximation:

$$F_{IC}(x, Q_0) = [c_{h,h} \otimes f^{IC}_{c/p}](x, Q_0)$$

$\boldsymbol{c_{h,h}}$ **is the twist-2 charm DIS coefficient function** introduced to factorize the twist-4 ladder terms; defined according to the S-ACOT-χ scheme

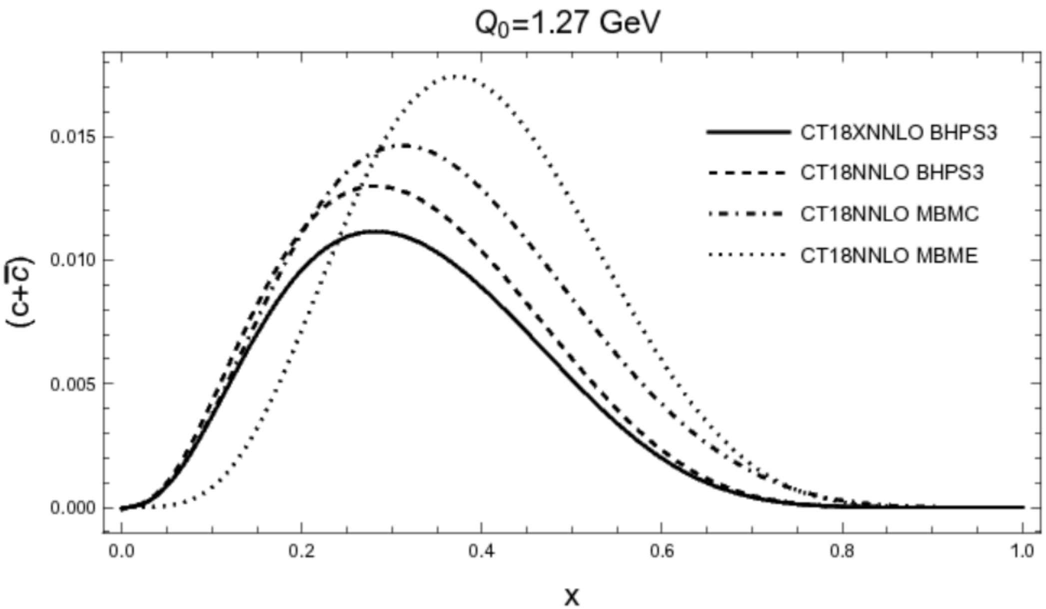
IC is compatible with any version of the ACOT scheme (cf. the paper)

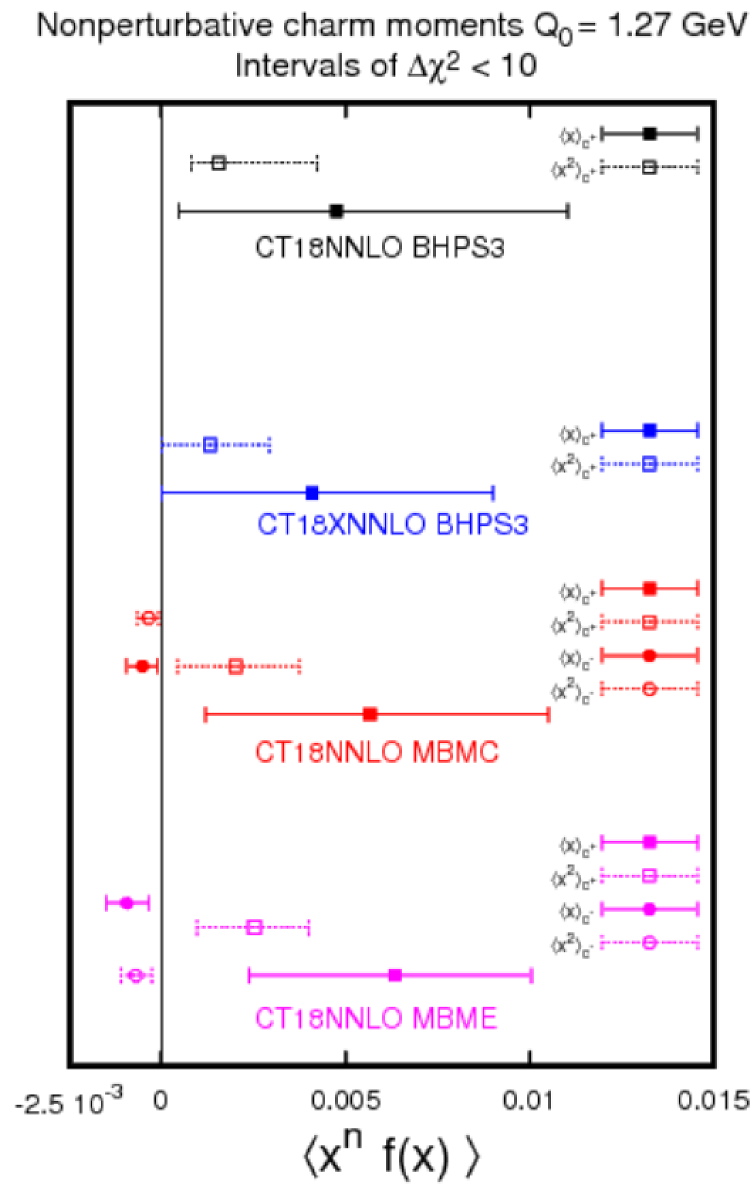
$\boldsymbol{f^{IC}_{c/p}(\xi, Q_0)}$ **is a nonperturbative charm parametrization**:

CT14 IC: $f^{IC}_{c/p}(\xi, Q_0)$ is a "**valence-like**" or a "**sea-like**" function, combined with the perturbative charm $f^{pert}_{c/p}$ from $g \to c\bar{c}$ splittings

CT18 IC NNLO analysis

IN PROGRESS





IC parametrizations with $\langle x\,(c+\bar{c})\rangle = 0 - 1\%$ at $Q \lesssim m_c$ allowed with high confidence

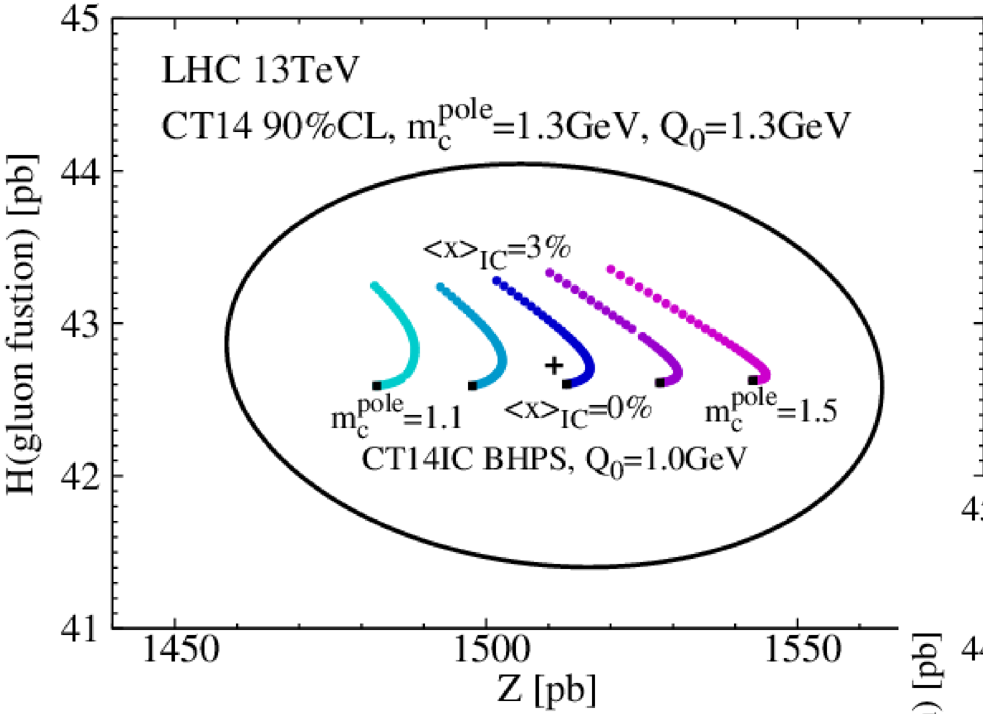
Preference for $\langle x \rangle_{IC} \neq 0$ is reduced compared to the CT14 IC study

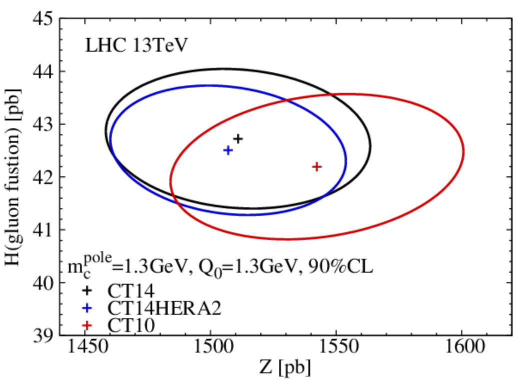
Impact of IC on physical observables

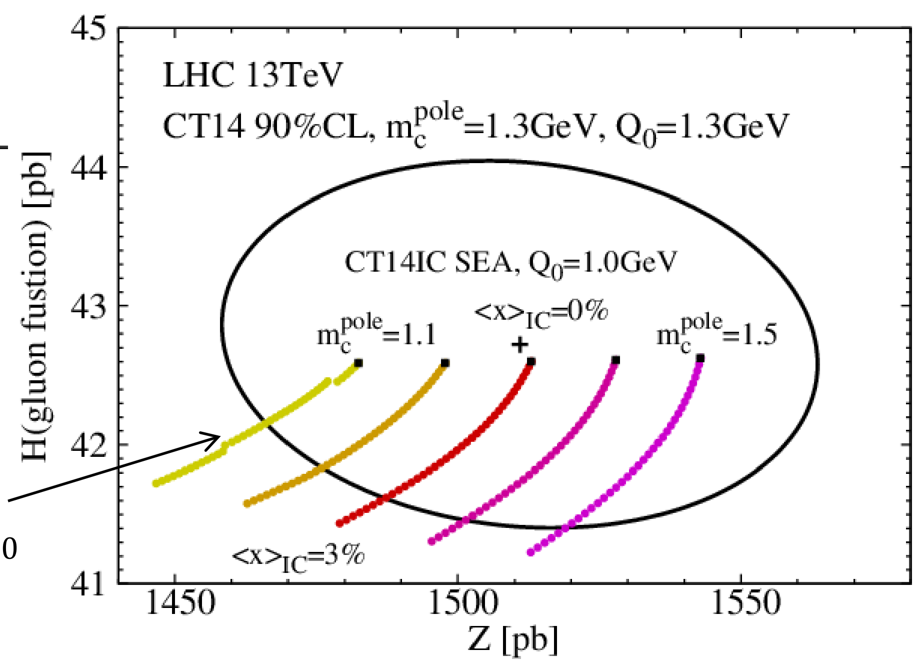
- Mild effects at the LHC
- Smoking gun signatures in SIDIS at the EIC

[Our estimates assume that the IC PDF component does not depend on the hard process.]

LHC: NNLO Total inclusive electroweak boson production cross sections $\sigma_{tot}(pp \to VX)$







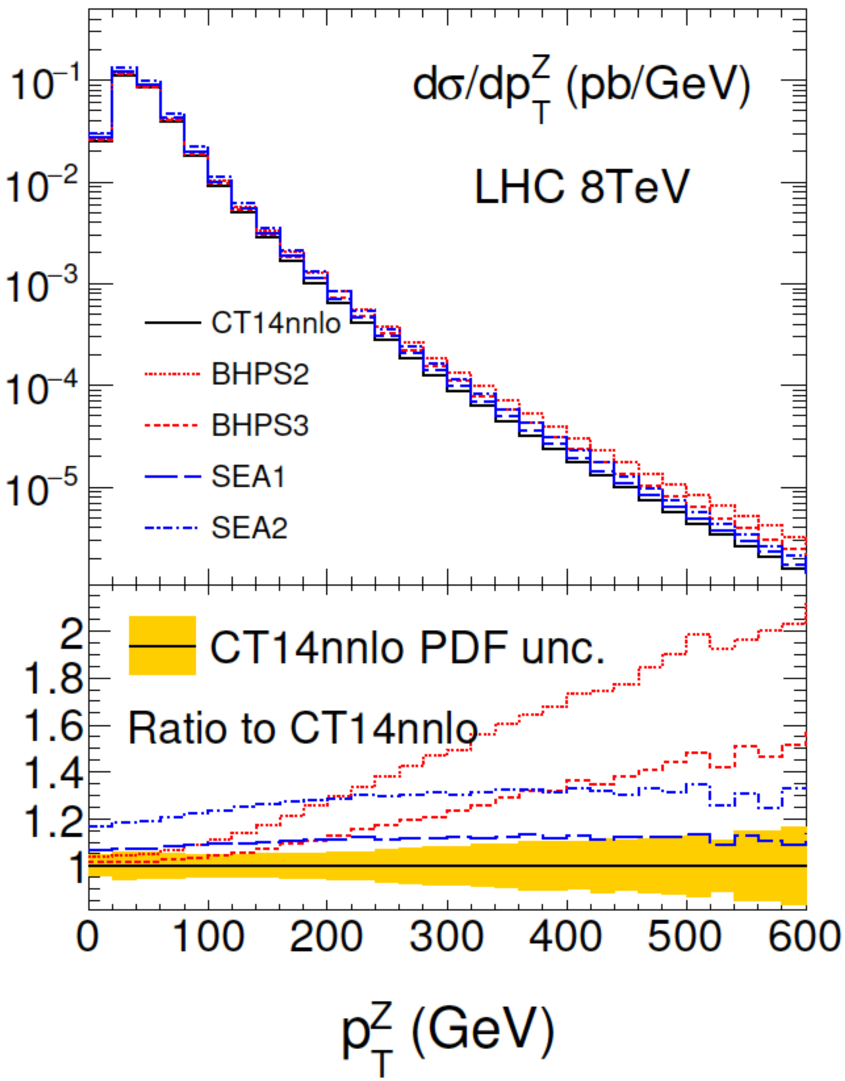
Disfavored,
$\Delta\chi^2_{global} > 100$

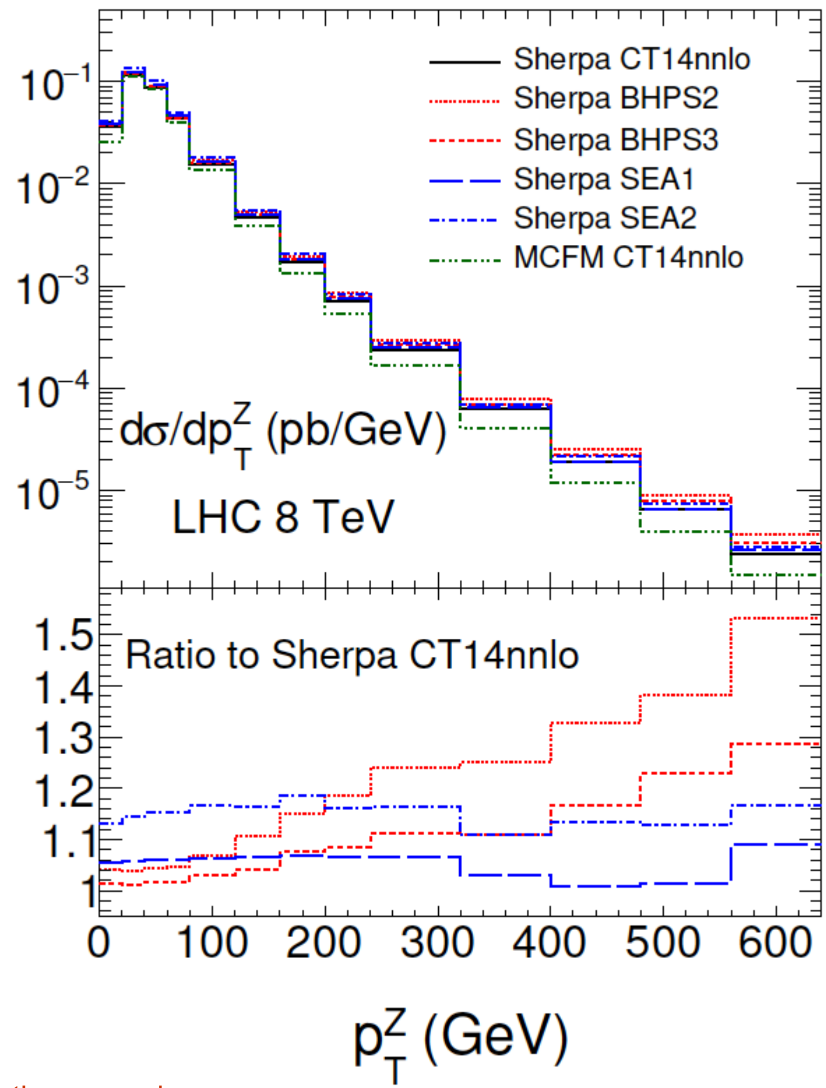
[Hou et al., arXiv:1707.00657]

LHC searches for intrinsic charm

[Hou et al., arXiv:1707.00657]

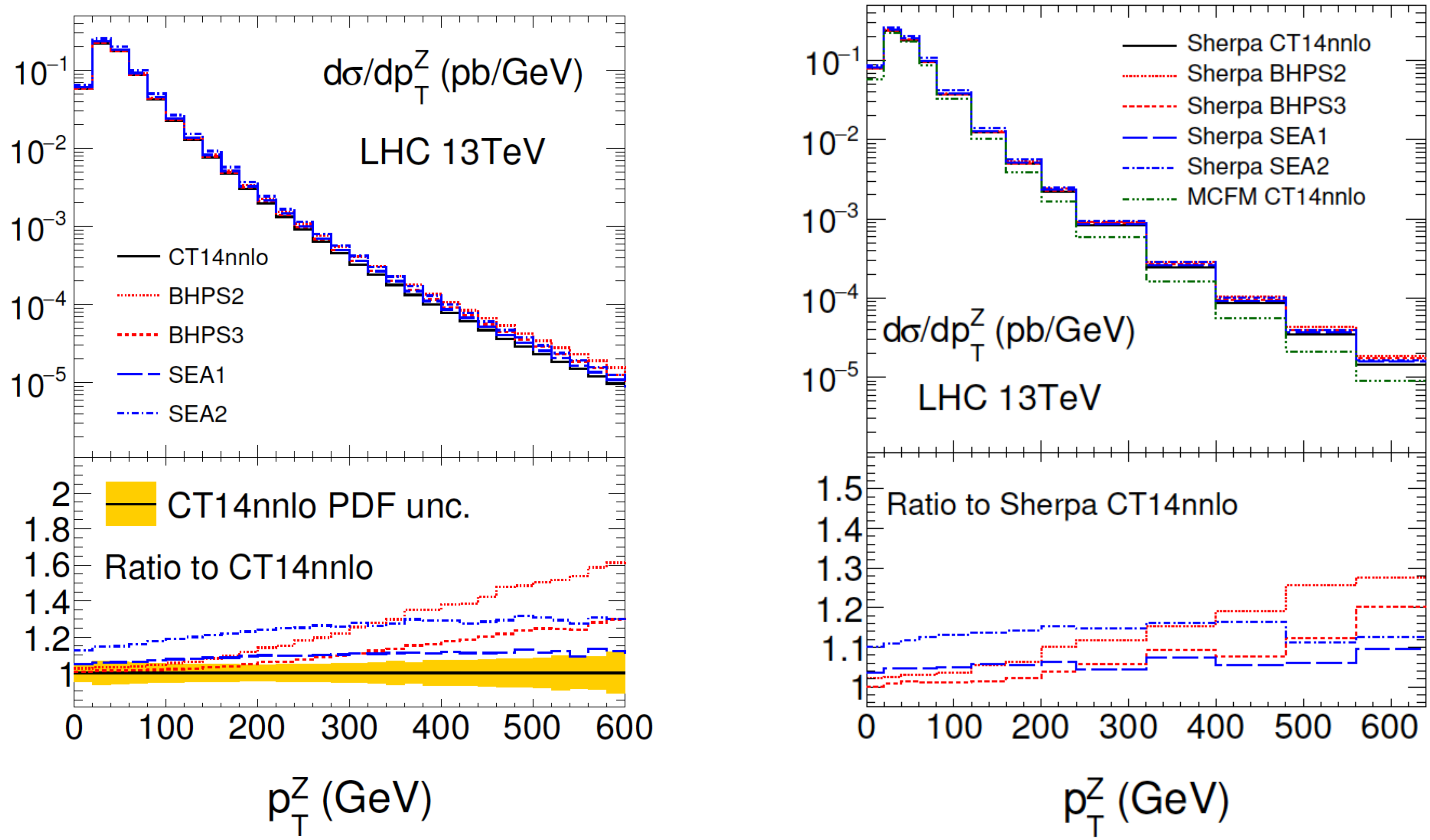
Z+c NLO computation with various models, without (left) and with parton shower (right)





Z+c NLO LHC 13 TeV

[Hou et al., arXiv:1707.00657]



The parton shower has the most significant effect in dampening the hard $p_T(Z)$ tail especially for BHPS fits. Sherpa predictions include HO tree-level MEs compared to MCFM and thus show enhancements in the harder $p_T(Z)$ region compared to MCFM. Similarly increasing or decreasing the number of multileg MEs in the merging changes the absolute level of p_T.

EIC, charm production

Orders-of-magnitude more events for some IC models

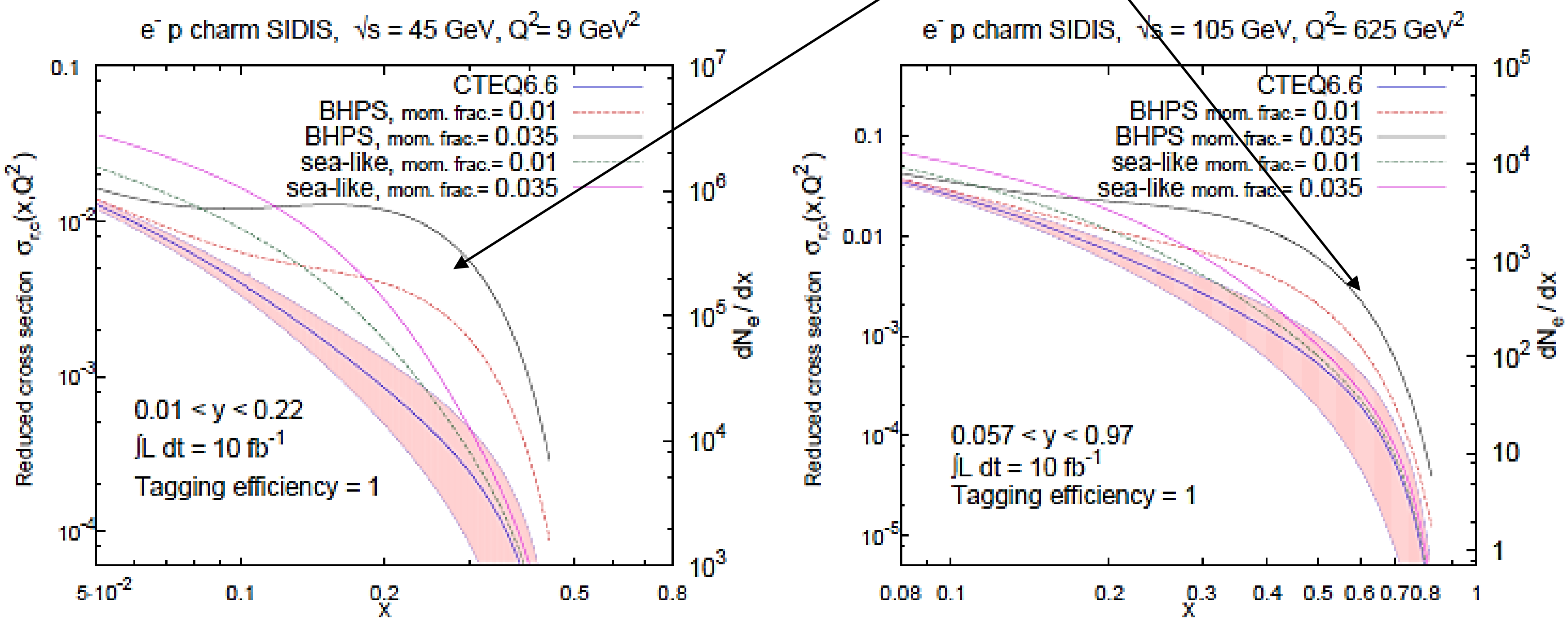


Figure 1.20. Charm contribution to the reduced NC e^-p DIS cross section at $\sqrt{s} = 45$ and 105 GeV. For each IC model, curves for charm momentum fractions of 1% and 3.5% are shown. For comparison we display the number of events dN_e/dx for $10\,\mathrm{fb}^{-1}$, assuming perfect charm tagging efficiency.

[Guzzi, Nadolsky, Olness, in arXiv:1108.1713; T. Hobbs, arXiv:1707.06711; Arratia et al., arXiv:2006.12520]

What is the faithful PDF uncertainty on QCD cross sections?

Our studies of CT, NNPDF, also MSHT fits show that the stated (as in CT18) or unstated (as in NN4.0) *uncertainty due to methodology* (parametrization/NN architecture, smoothness, data tensions, model for syst. errors, ...) is comparable to the impact of most recent data sets

PDF uncertainties in high-stake measurements (Higgs cross sections, W mass…) thus should be examined for *robustness of sampling over acceptable methodologies* and demonstrate *absence of biases* in this sampling.

Big data paradox: "the bigger the data, the surer we may fool ourselves".
Data analysis and (quasi-) MC integration with many (> 20) parameters are often at a risk of hard-to-detect, but dangerous sampling biases that take over the law of large numbers.

An undetected sampling bias may result in a wrong prediction with a low nominal uncertainty.
[X.-L. Meng, "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election," The Annals of Applied Statistics 12, (2018) 685.]

This experience also suggests how to verify PDF uncertainties on QCD parameters or cross sections using **hopscotch scans**. *[arXiv: 2205.10444, Sec. 2.]*

Hopscotch scans were illustrated for the NNPDF4.0 —thanks to the publicly available code.
Impact on the uncertainties at small and large x, PDF ratios, fitted charm, …
Insights applicable to other analyses using a large parameter space — CT/MSHT tolerance, polarized PDFs, etc.

Backup