

Parton distributions: the tolerance puzzle and big-data paradox

Pavel Nadolsky

Southern Methodist University

With A. Courtoy,

J. Huston, K. Xie, M. Yan, C.-P. Yuan

Manuscript in preparation

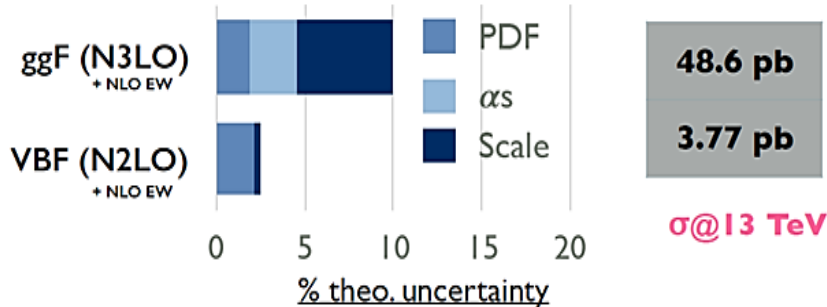
PDF uncertainties:
balancing **precision** and
robustness

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

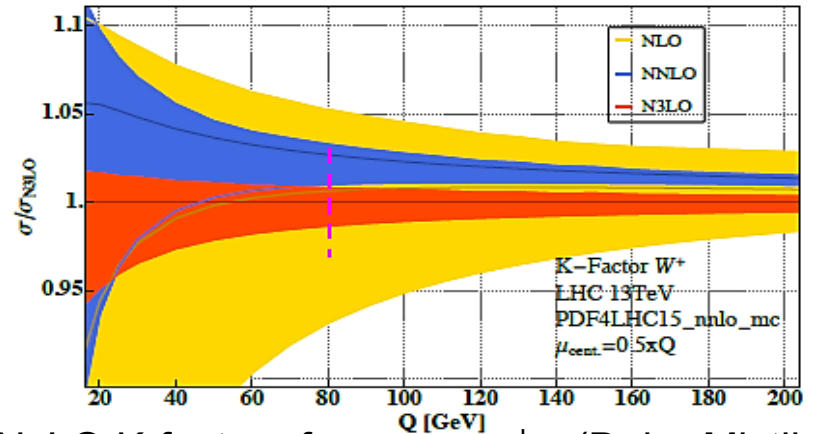


Reducing PDF and α_s uncertainties in EW/BSM physics at hadron colliders

Some key uncertainties in the (HL-)LHC Higgs physics are due to PDFs



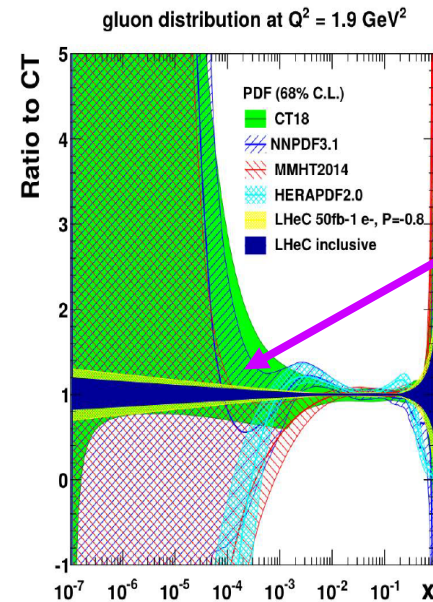
New N2LO/N3LO calculations precisely predict relevant PQCD cross sections



NxLO K-factors for $pp \rightarrow W^+ X$ (Duhr, Mistlberger)

High-luminosity measurements at HL-LHC and planned DIS experiments (EIC, LHeC, Muon-Ion Collider,...) + the progress in PQCD hold the potential to dramatically increase the precision of PDFs.

This advancement critically depends on understanding various sources of uncertainties in PDFs

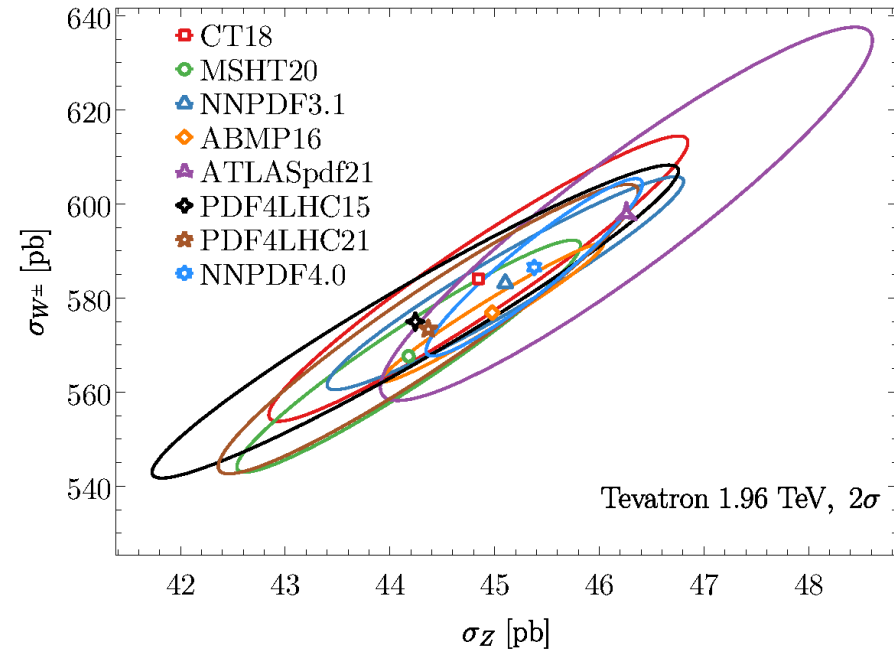
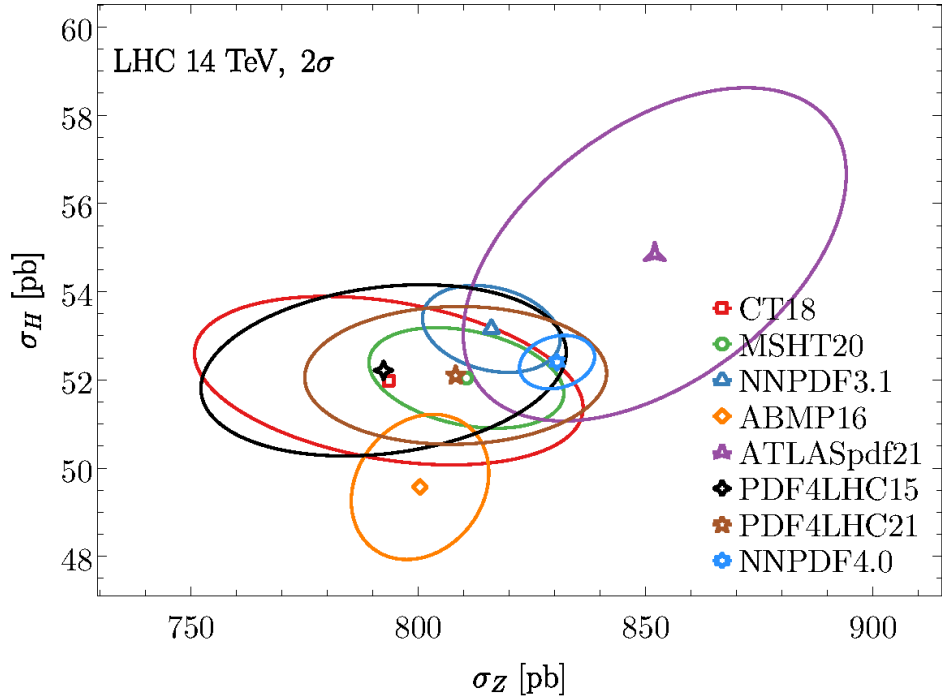


an optimistic post-LHeC uncertainty, requiring all advancements in the PDF fitting methodology

Recent advancements in determination of unpolarized PDFs

CT18, MSHT20, NNPDF4.0, ATLASpdf21 as well as PDF4LHC21

Precision PDFs (Snowmass 21 WP) [2203.13923v2]

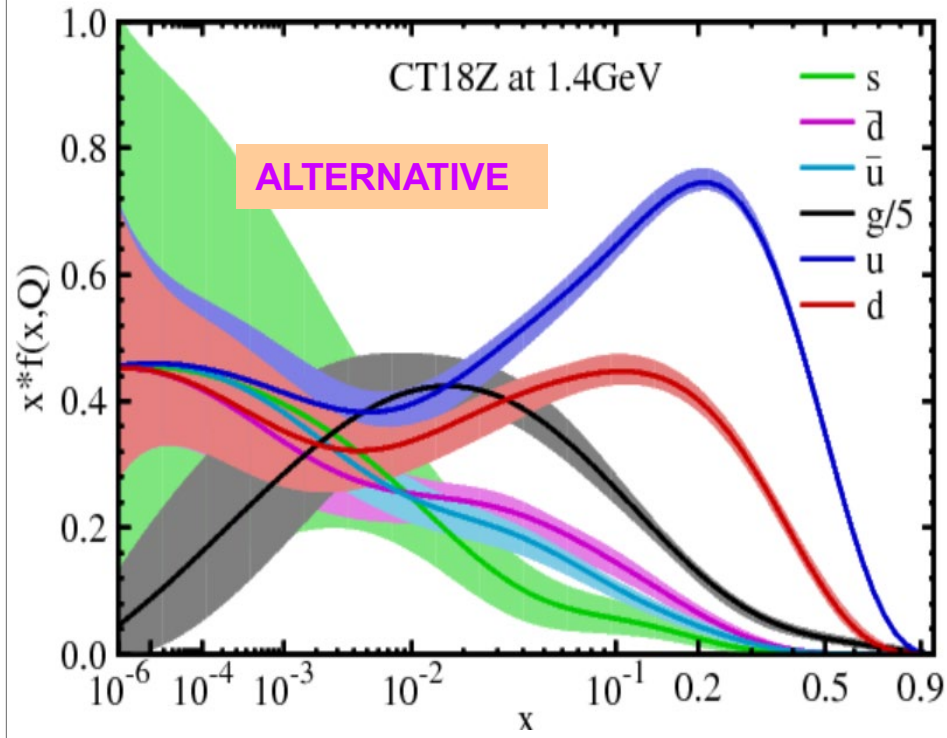
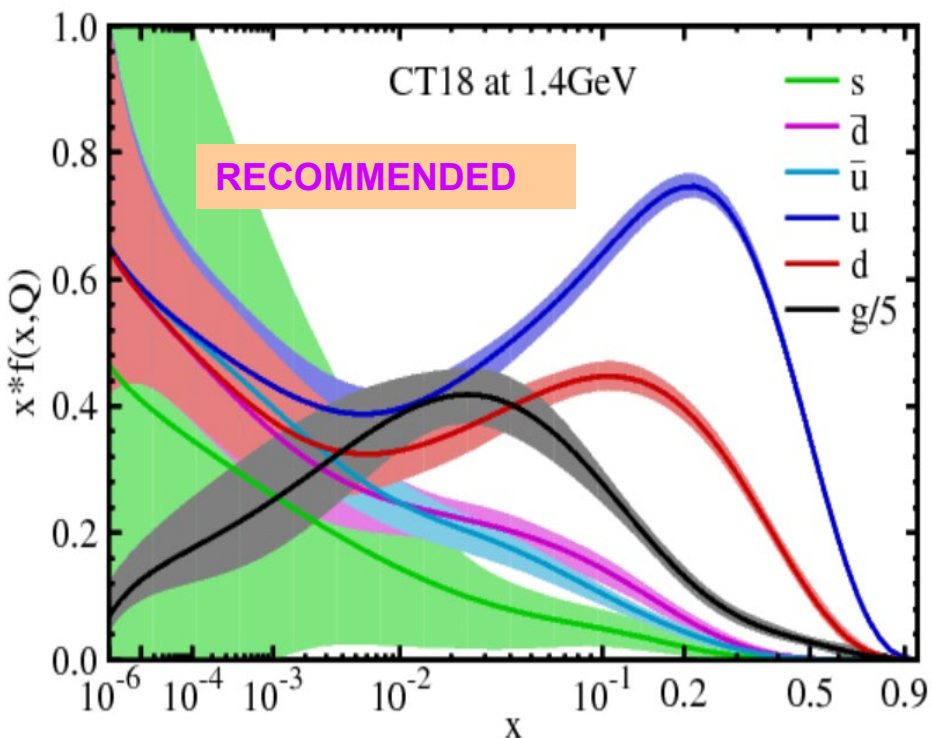


- Z^0 and $gg \rightarrow H^0$ cross sections at the LHC
- Z^0 and W^\pm cross sections at the CDF II
- 95% CL PDF uncertainties predicted with recent PDF sets.

CT18 parton distributions

PRD 103 (2021) 014013

Four PDF ensembles: CT18 (default), A, X, and Z



* **CT18 (N)NLO PDF set is recommended for the majority of LHC applications**

- CT18Z has enhanced gluon and strange PDFs at $x \sim 10^{-4}$, and reduced light-quark PDFs at $x < 10^{-2}$. The CT18Z maximizes the differences from CT18 PDFs, while preserving about the same goodness-of-fit as for CT18.
- CT18A and CT18X include some features of CT18Z, lie between CT18 and CT18Z.

CT18+CT18Z = robust NNLO PDF uncertainties

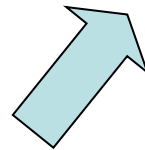
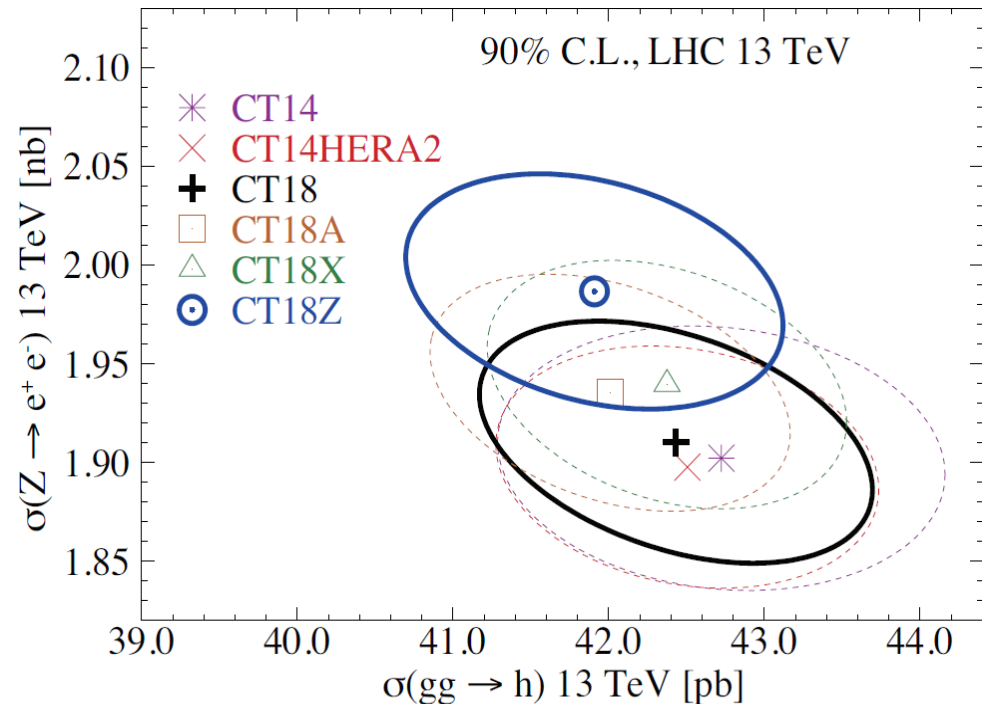
PRD 103 (2021) 014013

When **robust estimates** of PDF errors are needed, CTEQ-TEA recommends to combine predictions based on CT18 and CT18Z PDFs. This quantifies displacements of central PDFs that fall out of the nominal 90% CL CT18 uncertainty bands

CT18Z combines modifications made in CT18A and CT18X ensembles:

- **CT18A:** include ATLAS 7 TeV W/Z data, enhanced s quark PDF
- **CT18X:** mimic small- x resummation, different g PDF

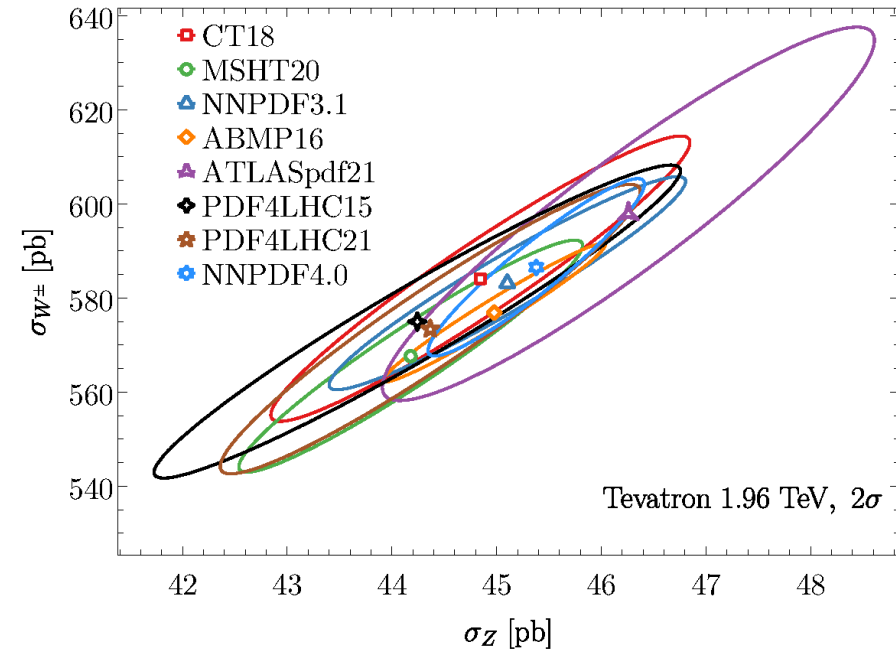
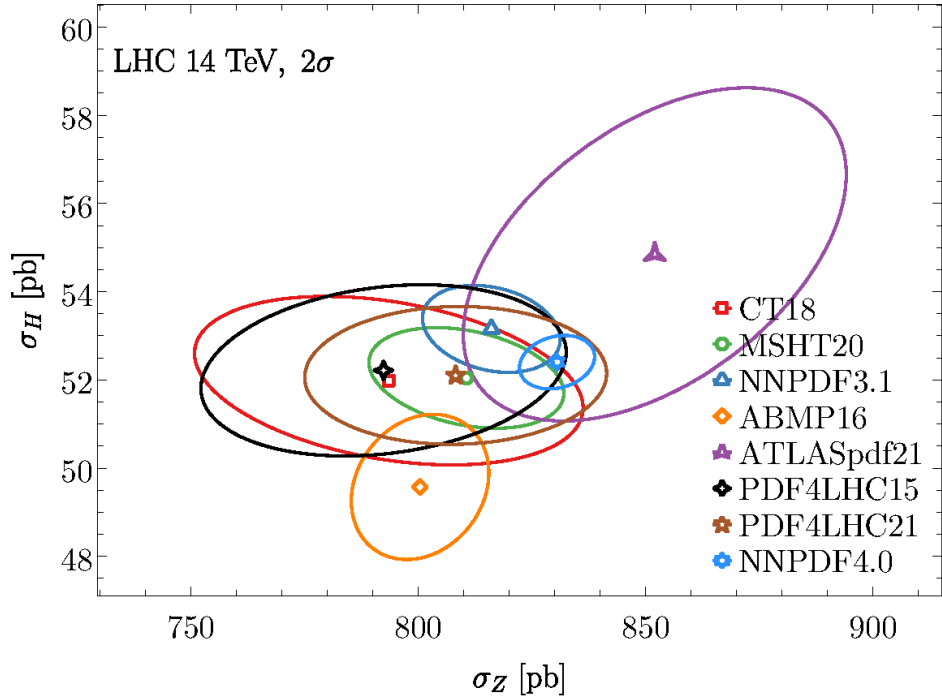
Taken together, CT18 and CT18Z largely cover the spread of central predictions obtained with different assumptions and selections of experiments



The tolerance puzzle

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

Precision PDFs (Snowmass 21 WP) [2203.13923v2]



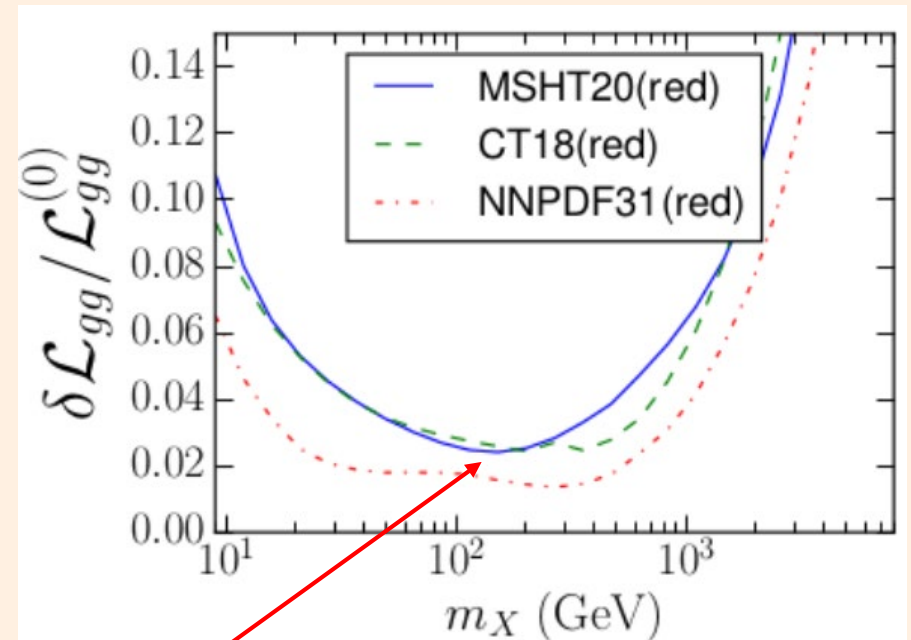
The answer has direct implications for high-stake experiments such as W boson mass measurement

The tolerance puzzle

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.

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

Insights from statistics of sampling and MC integration in many dimensions

Bad news: The tolerance puzzle is *intractable* in too 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, Woźniakowski, 1997]

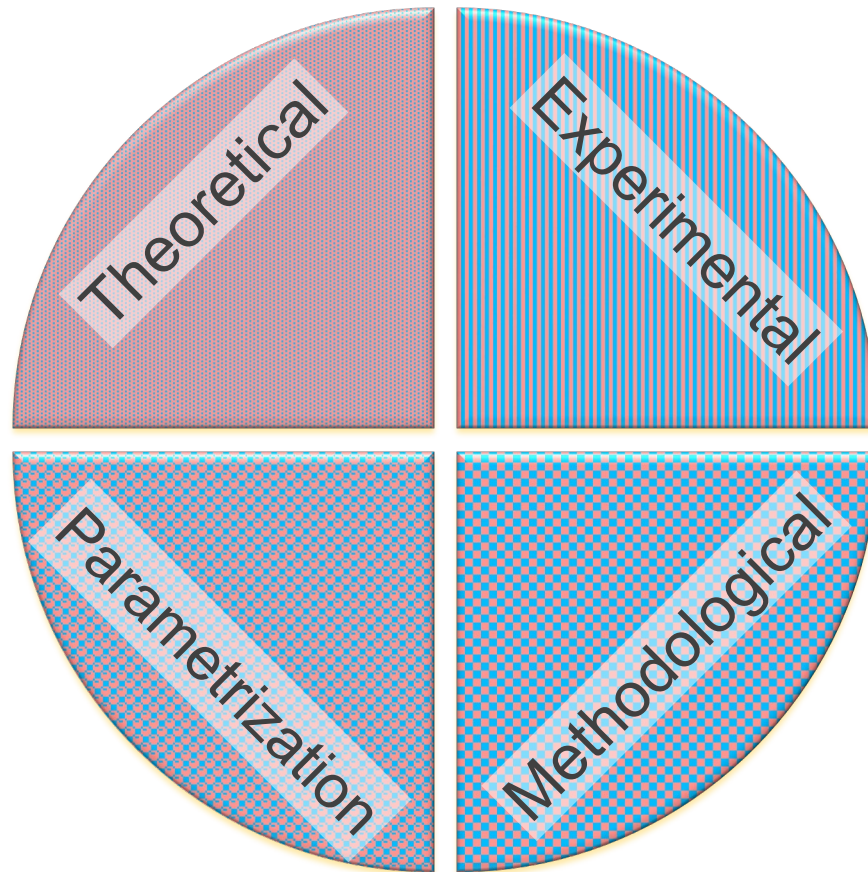
[Hickernell, MCQMC 2016, 1702.01487]

Good news: the expectation values can be estimated with fewer replicas if only few effective large dimensions contribute the bulk of the uncertainty


⇒ **hopscotch scans** to robustly sample the PDF uncertainty for specific QCD observables


see details in the manuscript

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](https://arxiv.org/abs/1905.06957)

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

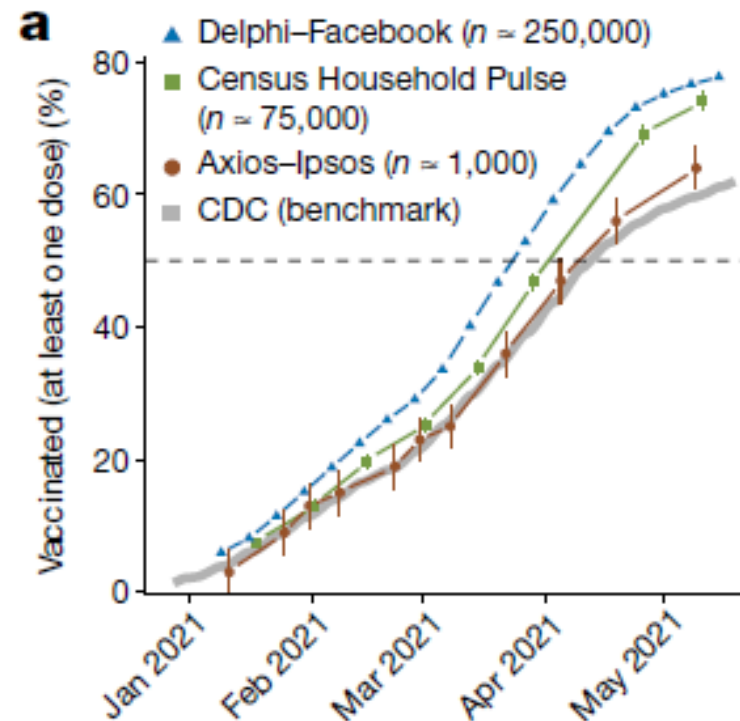
Published online: 8 December 2021

Check for updates

Valerie C. Bradley^{1,2}, Shiro Kuriwaki^{1,2}, Michael Isakov², Dino Sejdinovic¹, Xiao-Li Meng⁴ & Seth Flaxman^{2,3*}

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 Paradox¹. 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 Pulse⁴ (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 minuscule margins of error on the incorrect estimates. By contrast, an Axios–Ipsos online panel⁵ with about 1,000 responses per week following survey research best practices⁶ provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework⁷ 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 $n \rightarrow \infty$, under a set of hypotheses, it is usually expected that the sample deviation on an observable μ decreases as

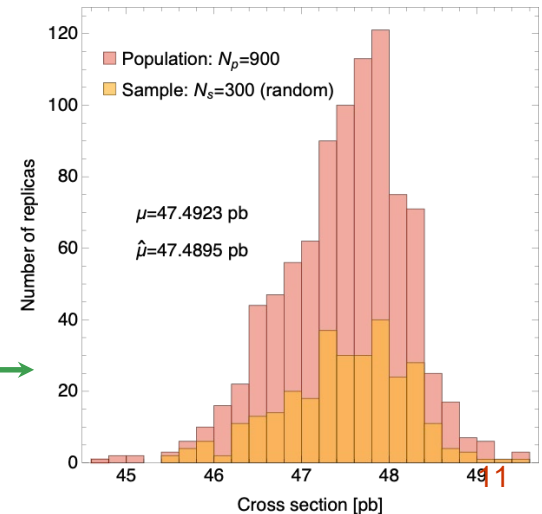
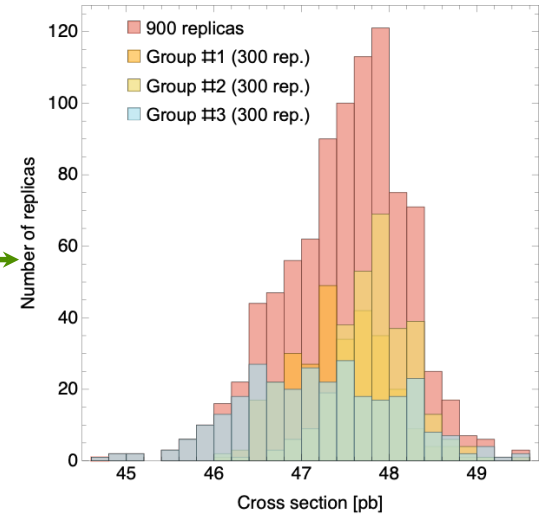
$$\mu - \hat{\mu} \propto \sigma_{std} / \sqrt{n}$$

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.

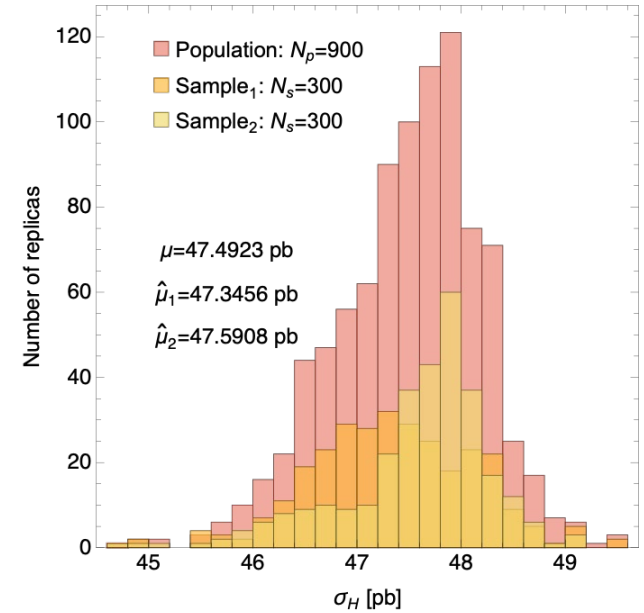


Trio identity

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 σ_{std}/\sqrt{n} !



Quality of the sample is as important as quantity.



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

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

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

$$\mu - \hat{\mu} = \text{Corr}[\text{observable}, \text{sampling algorithm}] \times \sqrt{\frac{N}{n} - 1} \times \sigma_{std}(\text{observable})$$

depends on the sampling algorithm

decreases as σ_{std}/\sqrt{n} for random sampling

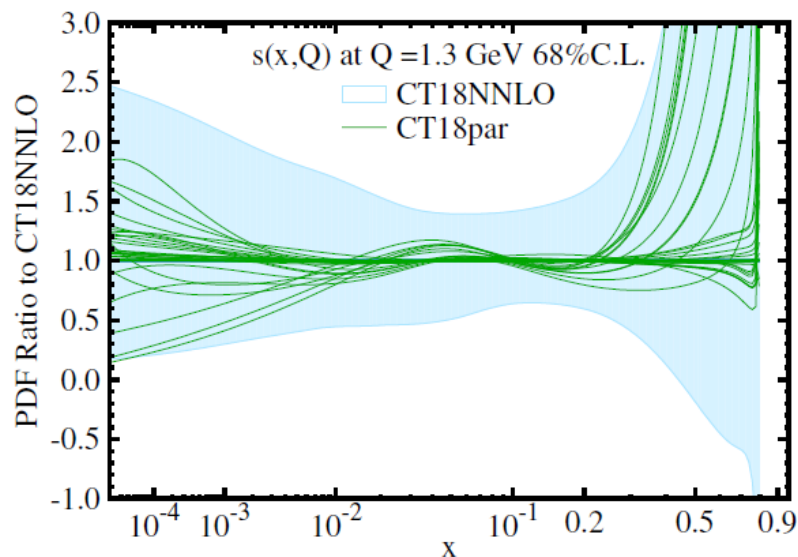
[X.-L. Meng, The Annals of Applied Statistics, Vol. 12 (2018), p. 685]
[Hickernell, MCQMC 2016, 1702.01487]

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

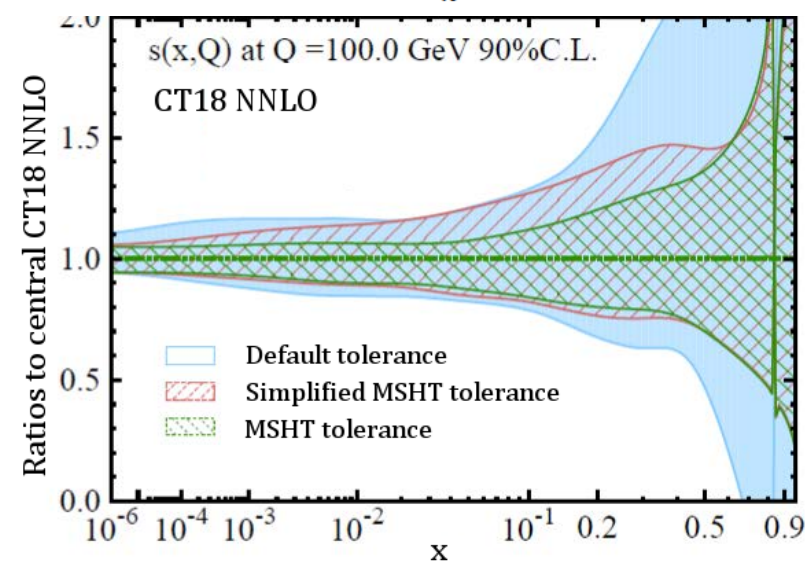
1. The sample deviation can be large if $\text{Corr}[\dots]$ does not decrease as $o(1/\sqrt{N})$
2. Standard error estimates can be misleadingly small.
3. **Control for sampling biases is critical** to avoid the situation described as the **Big Data Paradox** [Meng]:

The bigger the data, the surer we fool ourselves.

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:

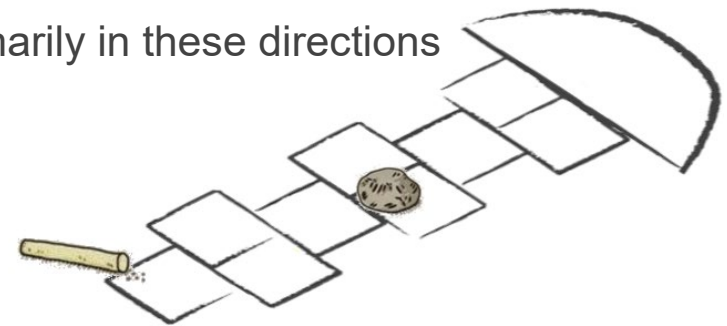
estimation of the PDF sampling uncertainty on a QCD cross section σ_{QCD}

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



ELEMENTOS PARA LLEGAR AL CIELO

Based on the ideas of [Hickernell, MCQMC 2016, 1702.01487]
[Sloan, Woźniakowski, 1997]

A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

Step 1

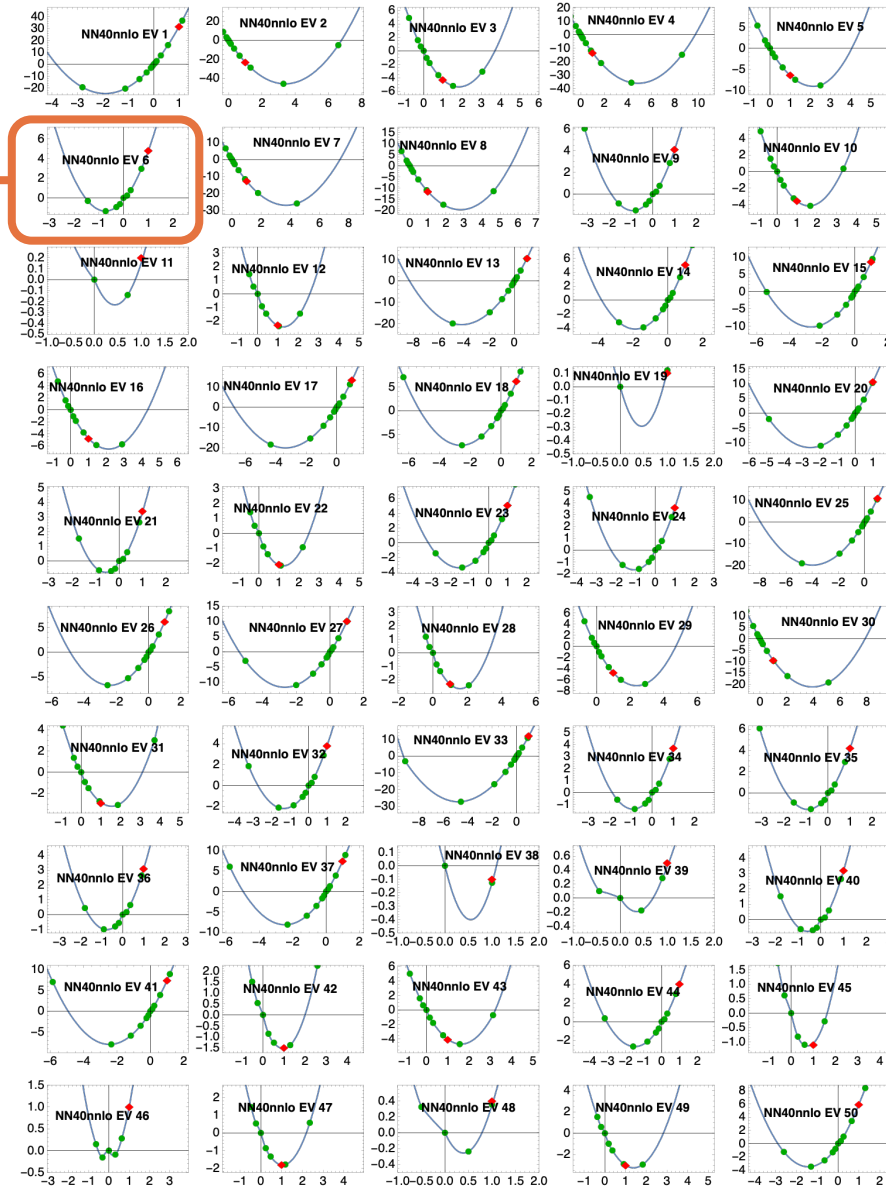
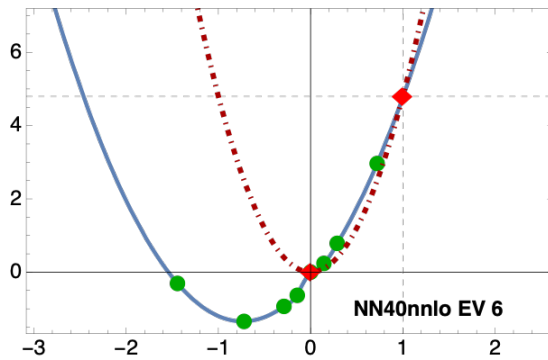
The NNPDF4.0 Hessian set ($n = 50$) defines a coordinate system on a manifold corresponding to the largest variations of the PDF uncertainty — **red dots and curve**.

[NNPDF, 2109.02653]

Step 2

Using the public NNPDF code, scan “ χ^2_{tot} along the 50 EV directions to identify a hypercube corresponding to $\Delta\chi^2 \leq T^2$ where $T^2 > 0$ is a user-selected value).

Lagrange multiplier scan confirms the approximate Gaussian profiles, but suggests that there exist solutions with lower χ^2 — **green dots and blue curve**. [Shown for “exp” χ^2 here.]



A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

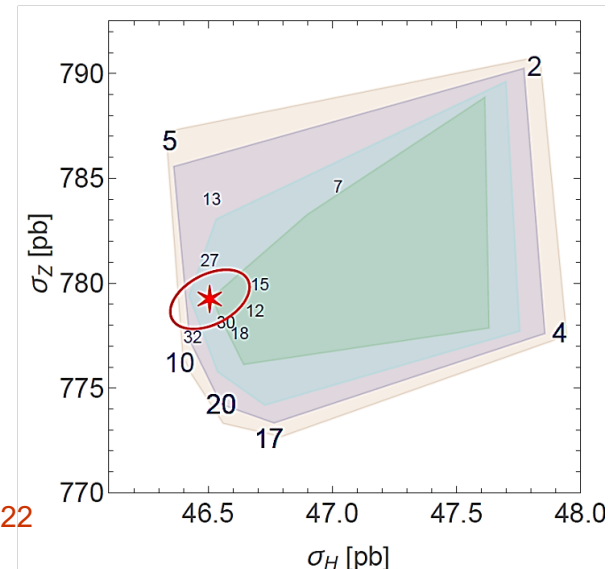
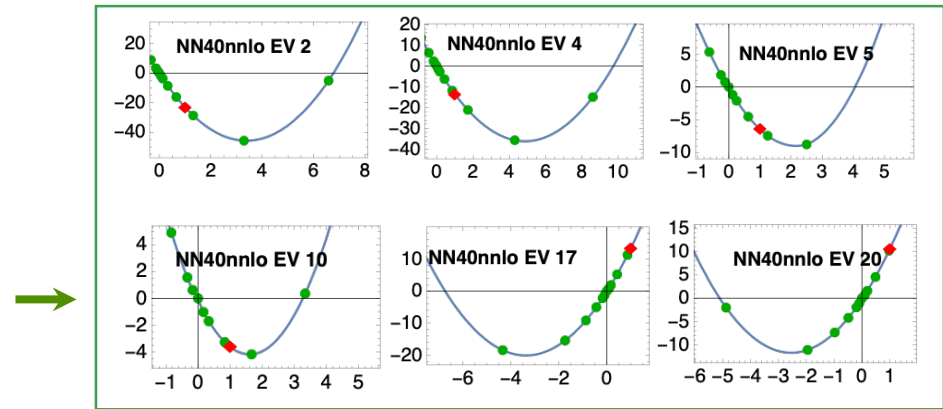
Step 3

Guidance from specific cross sections: we identify 4-7 EV directions that give the largest displacements for a given $\Delta\chi^2$ per pair.

E.g., σ_Z vs. σ_H is represented by a convex hull hexagon with the corners corresponding to “large” EV directions 2, 4, 5, 10, 17, 20.

Other directions generally give smaller displacements.

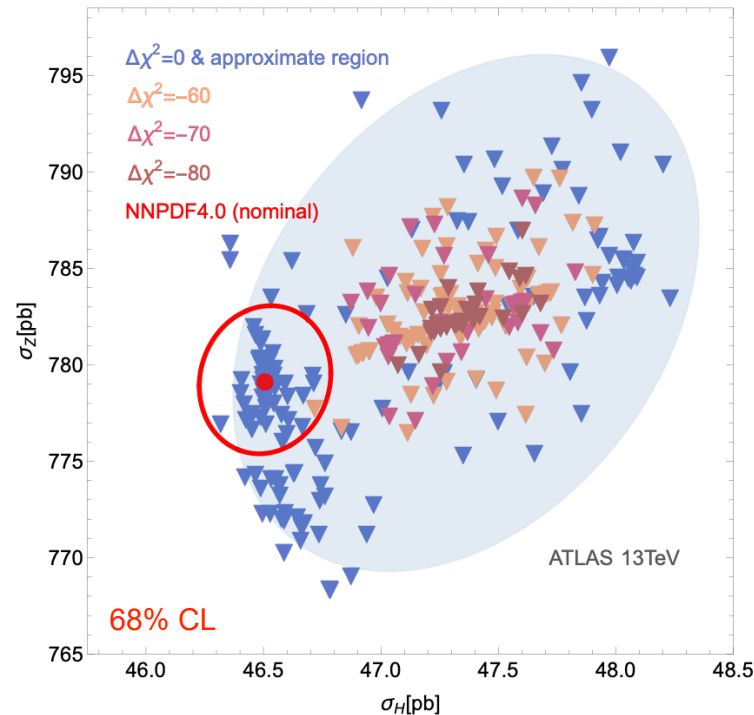
The contours are for $\Delta\chi^2 = +10, 0, -10, -20$ w.r.t. NNPDF4.0 replica 0 (red).



A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

Step 4

For each pair of cross sections, we sample its “large” central slice of the hypercube uniformly with 300 replicas. Sort the $n_{pairs} \times 300$ resulting replicas according to their $\Delta\chi^2$ w.r.t. to NN40 replica 0.



Each of the $\Delta\chi^2 = 0 \pm 3$ replicas is an acceptable PDF set from the NNPDF4.0 fit.

The blue ellipse (constructed using a convex hull method) is an approximate region containing all found replicas with $\Delta\chi^2 = 0 \pm 3$.

[Anwar, Hamilton, Nadolsky, 1901.05511]

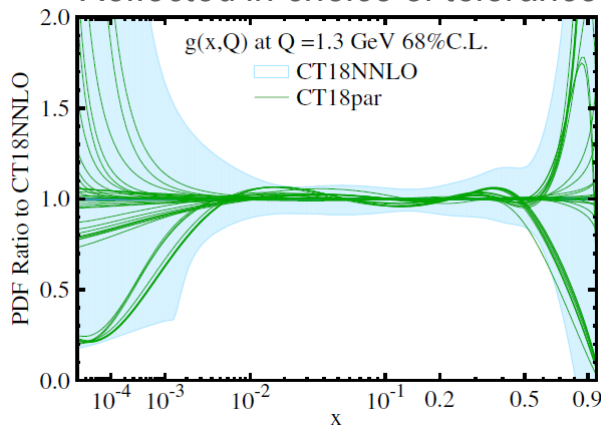
The blue area is larger than the nominal NNPDF4.0 uncertainty (red ellipse).

The hopscotch scans: NNPDF4.0 vs CT18 uncertainties

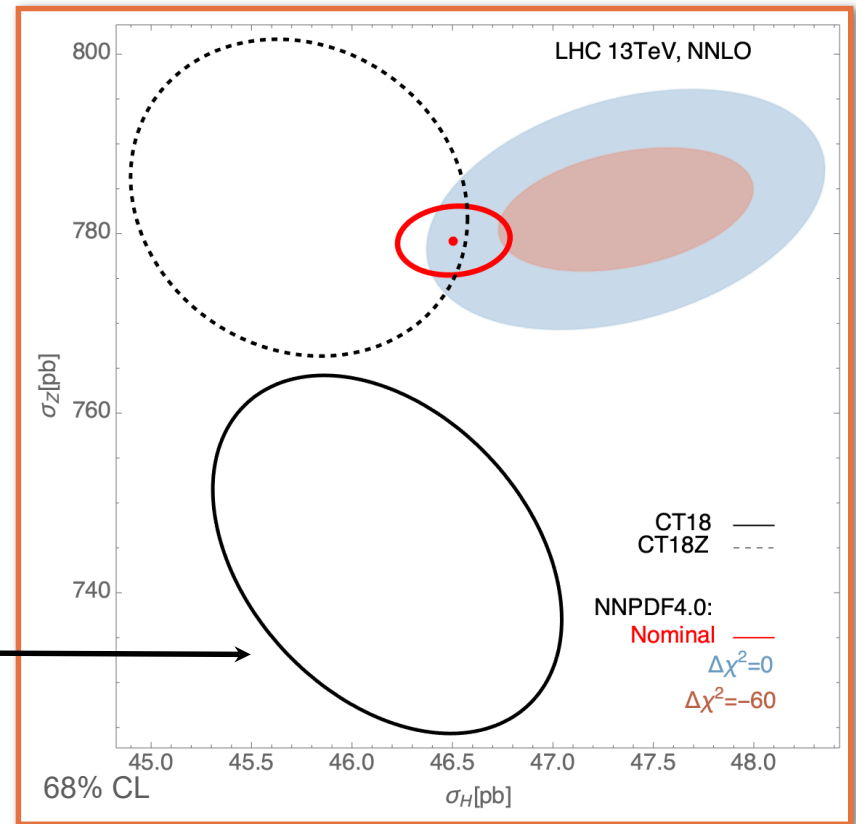
CT18 PDF uncertainty:

Accounts for the sampling over 250-350 parametrization forms and possible choices of fitted experiments and fitting parameters.

Reflected in choice of tolerance.



[Hou et al, Phys.Rev.D 103 (2021)]

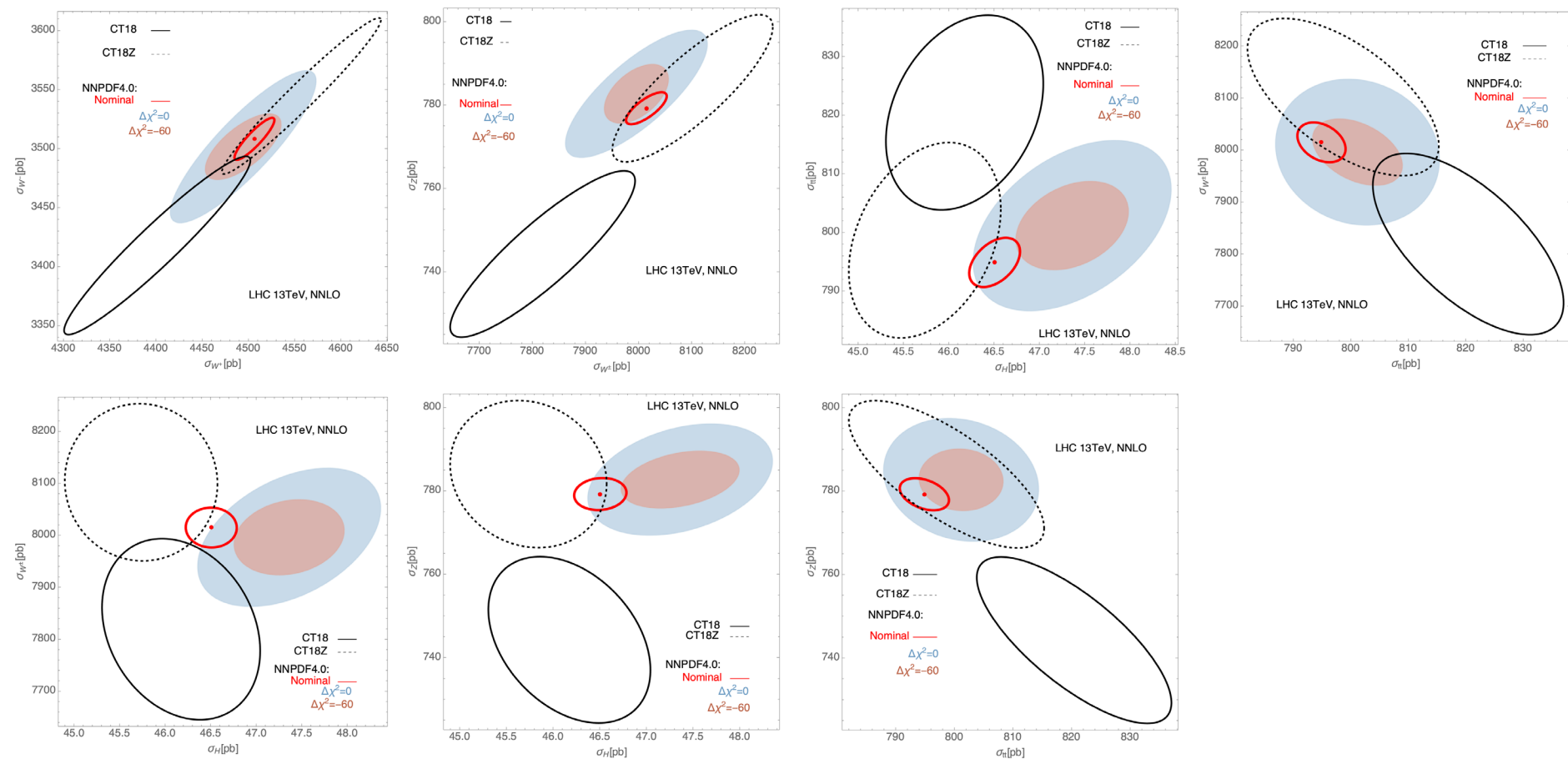


Blue and brown filled ellipses:

- areas of possible solutions corresponding to an equal ($\Delta\chi^2 = 0$) or lower ($\Delta\chi^2 = -60$) chi square w.r.t. the nominal solution
- found through the hopscotch scan — a dimensionality reduction method.
- size of blue areas comparable to 68% CL CT18 ellipses

The hopscotch scans: NNPDF4.0 vs CT18 uncertainties

PRELIMINARY



Ellipses at 68% CL

2022-05-13

P. Nadolsky, LoopFest 2022

20

A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

The chosen form of χ^2 affects the PDF uncertainty!

The previous slides discuss the “experimental” prescription for correlated syst. errors adopted in the χ^2/N_{pt} tables of the NNPDF4.0 publication

The NNPDF4.0 replicas are trained by optimizing χ^2 in the t_0 prescription.

Experimental prescription:

$$\chi_{tot}^2/N_{pt} = 1.160.$$

t_0 prescription:

$$\chi_{tot}^2/N_{pt} = 1.233.$$

$$\chi_{tot}^2(\text{exp}) - \chi^2(t_0) = -340 \text{ for } 4618 \text{ data points}$$

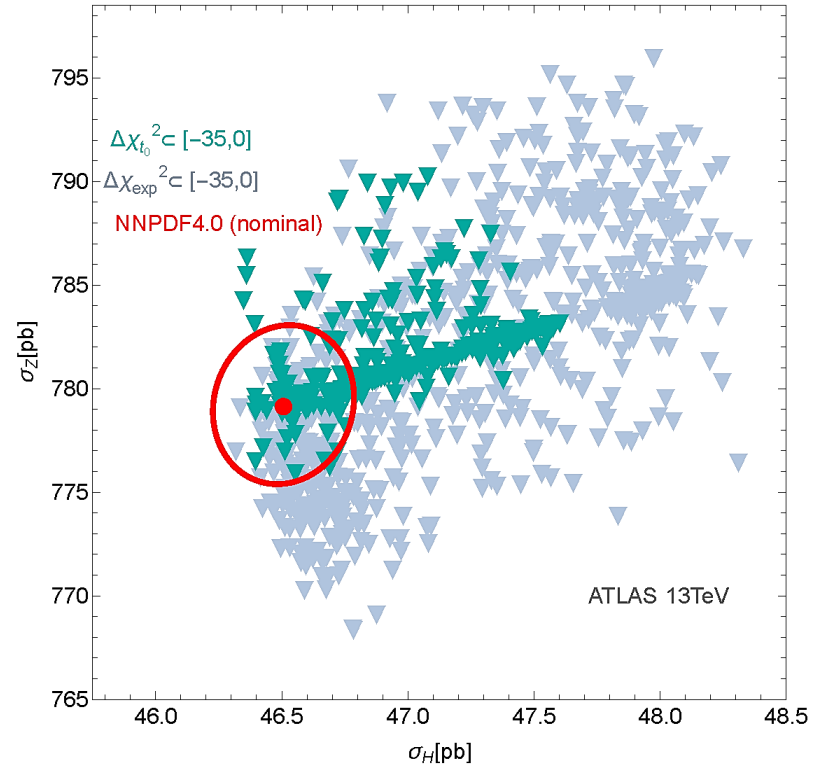
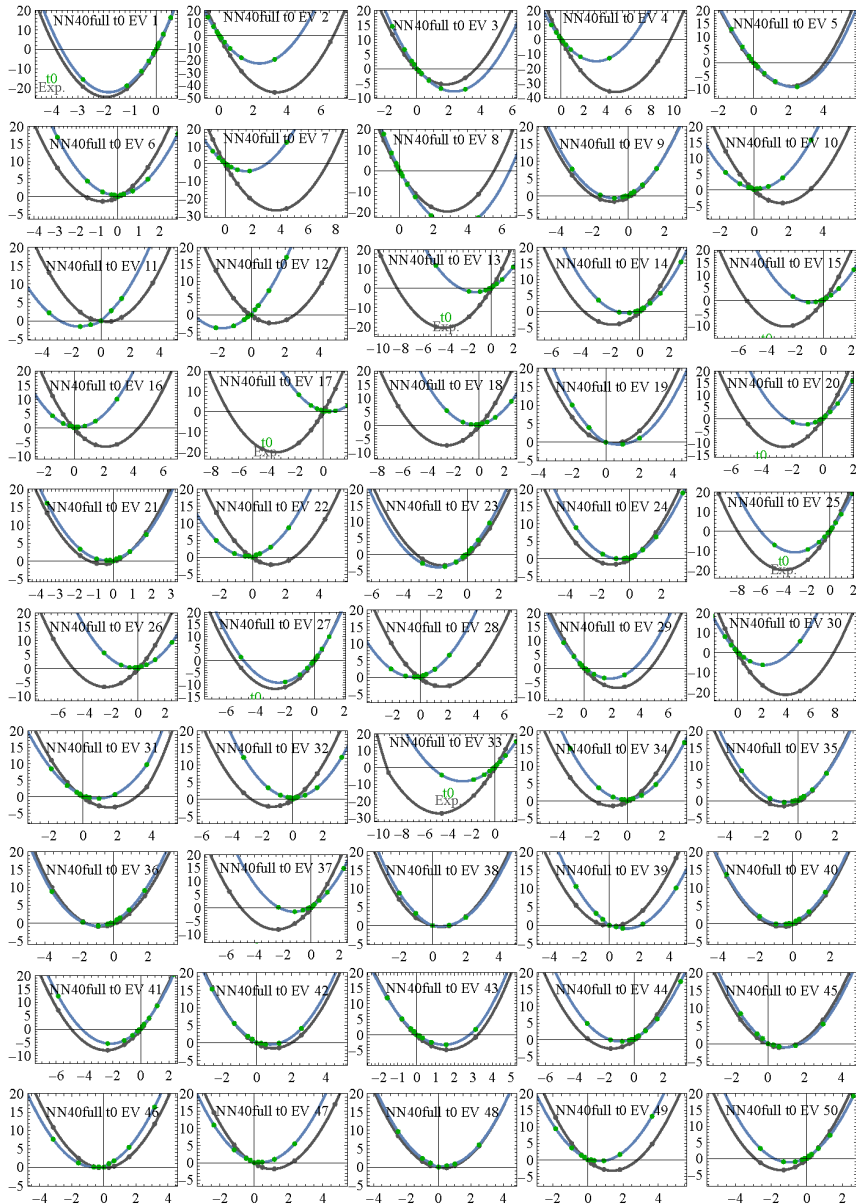
Larger differences from the nominal NN40 confidence regions

Smaller differences

Other prescriptions for χ^2 are in use by the other groups

Hopscotch scans, t_0 vs. experimental χ^2

in progress



Main observations hold.
 Generally smaller shifts from the nominal NN4.0 predictions than with the “exp” prescription

Conclusions

What is the faithful PDF uncertainty on QCD cross sections?

PDF uncertainties in high-stake measurements (Higgs cross sections, W mass...) should be examined for *robustness of sampling*.

Sampling biases may arise in PDF fits operating with large samples of data, multiparametric functional forms, elaborate models of correlated errors.

The trio identity may take over the law of large numbers.

An undetected sampling bias may result in a wrong prediction with a low nominal uncertainty.

Sample deviations may limit accuracy of the PDF uncertainties and explain some differences between the PDF sets.

Experience with big surveys and Monte-Carlo integration shows how to quantify such deviations for QCD parameters or cross sections.

⇒ possible framework for systematic study of parametrization within CT.

Hopscotch scans illustrated for the NNPDF4.0 —thanks to the publicly available code.

Applicable to other analyses using similar methodology and a large enough parameter space — e.g. for polarized PDFs.

Backup

Robustness with respect to systematic errors

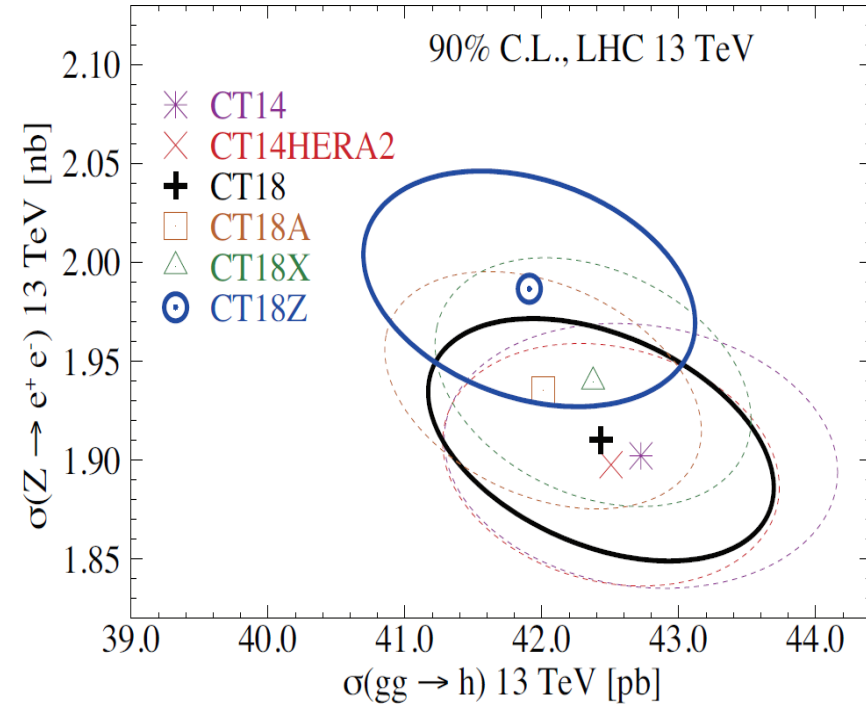
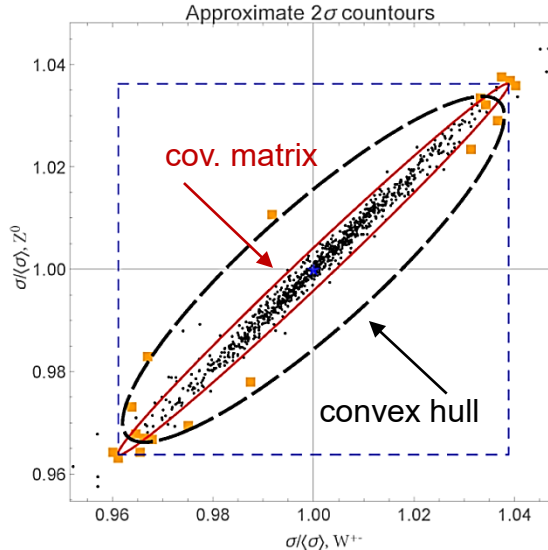
Strong dependence on the definition of corr.
syst. errors may raise a general concern:

Overreliance on Gaussian distributions and covariance matrices for poorly understood effects may produce very wrong uncertainty estimates

[N. Taleb, *Black Swan & Antifragile*]

For instance, the cov. matrix may overestimate the correlation among discrete data points, resulting in a too aggressive error estimate

[Anwar, Hamilton, P.N., arXiv:1905.05111]



The CT18 uncertainties aim to be **robust** by largely covering the spread of central predictions obtained with different assumptions and correlation models

[See also Kassabov et al., [2109.02653](https://arxiv.org/abs/2109.02653), Sec. 4.2.1]