Toward precise and robust unpolarized PDFs

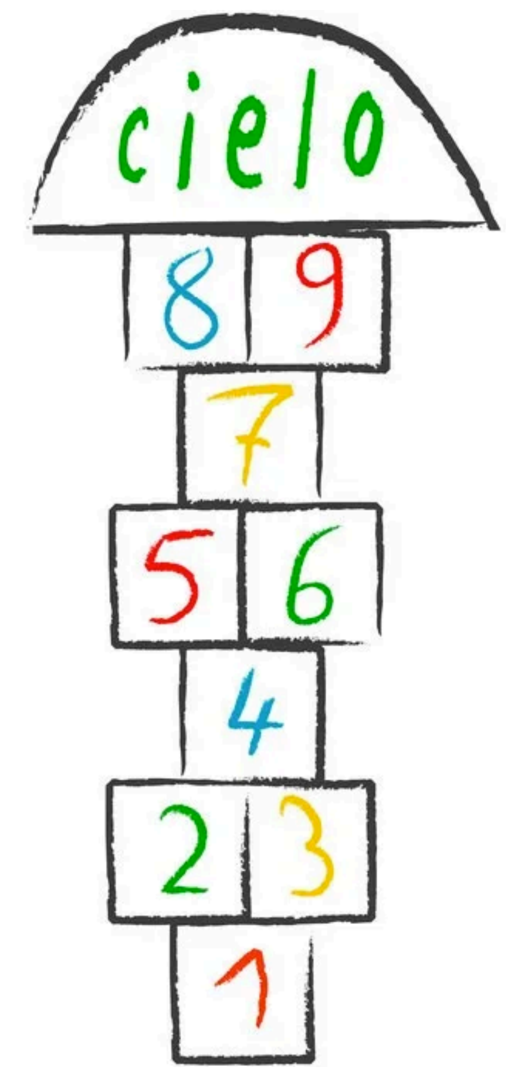
[Parton distributions need representative sampling, 2205.10444]

Aurore Courtoy

Instituto de Física
National Autonomous University of Mexico (UNAM)

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

JLab Theory Seminar — June 6, 2022



Instituto de Física UNAM

CONACYT
Consejo Nacional de Ciencia y Tecnología

dgapa
Dirección General de Asuntos del Personal Académico

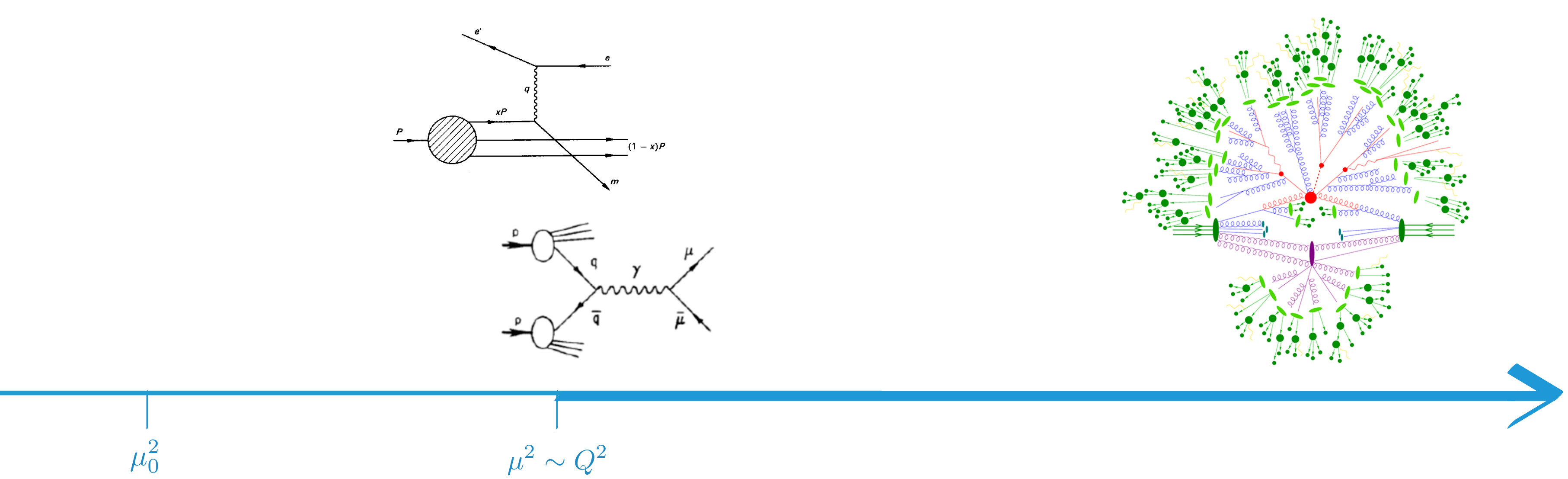
Phenomenology of proton structure

Parton Distribution Functions (PDFs) describe the distribution of quarks inside hadrons

Through PDFs, we can learn about nonperturbative dynamics at low energy.

PDFs contribute to high-energy processes, they are determined phenomenologically through global analyses.

Those global analyses rely on the availabilty of data in the Deep Inelastic regime, theoretical framework (e.g. pQCD at N^{x}LO) and a statistical methodology.



Parton Distribution Functions — towards phenomenology

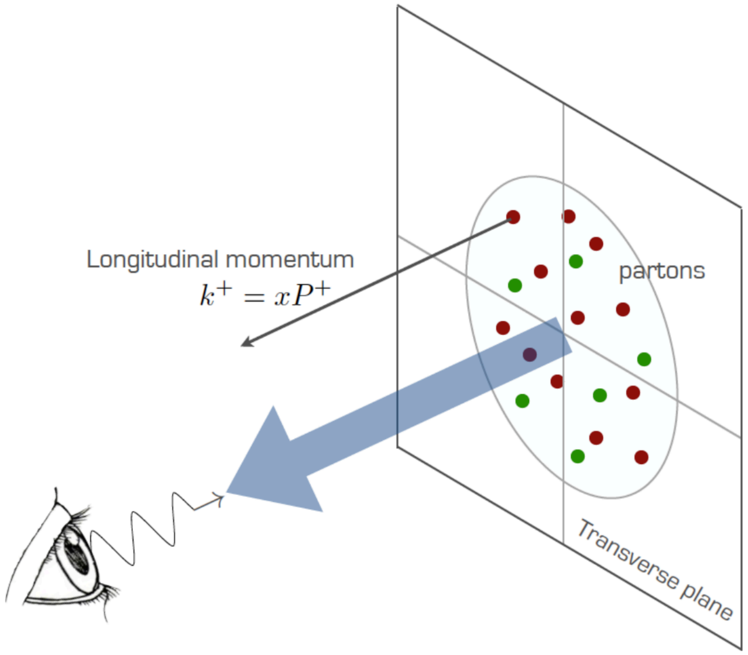
PDFs are nonperturbative objects.
They represent the distribution of quarks and gluons in a given configuration of the parent hadron.

(x, Q^2) interdependent variables

- [ ] x dependence difficult to predict in theory
- [x] Q^2 dependence known precisely in pQCD

Global analyses aim to extract the x-dependence of PDFs from data with minimal guidance from first principles.

- positivity constraints
- support in $x \in [0,1]$
- end-point: $f(x=1)=0$
- sum rules: $< x >_n = \int_0^1 dx\, x^{n-1} f(x)$
- …



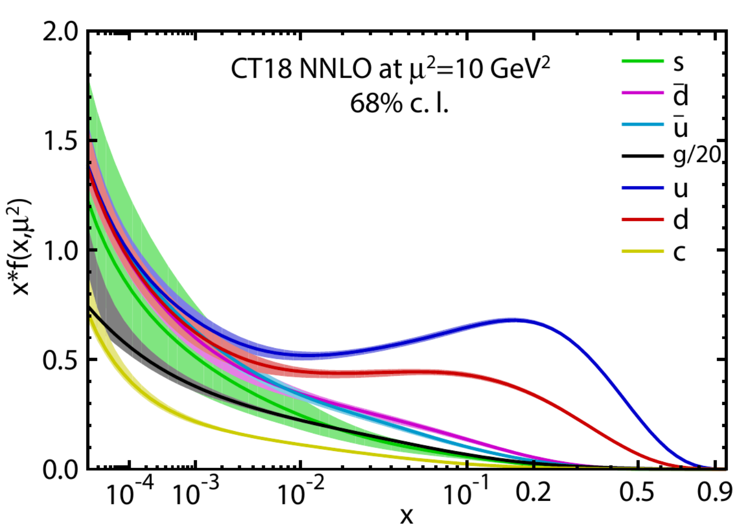
State-of-the-art analyses

Phenomenolgical analyses of PDFs combine the most updated/available versions of all the “ingredients.”

$$\phi_{ij}(k;p,s) = 2\pi \sum_X \int \frac{d^3\mathbf{P}_X}{2E_X} \delta^4(p-k-P_X) \langle p,s|\bar{\psi}_j(0)|X\rangle\langle X|\psi_i(0)|p,s\rangle$$

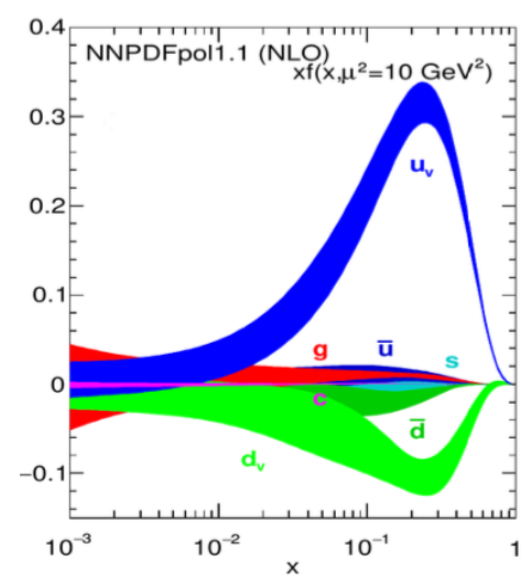
$$\boxed{\phi(x,s) = \frac{1}{2}\left[\mathbf{f_1}(x)\not{n}_+ + s_L g_1(x)\gamma^5\not{n}_+ + h_1 i\sigma_{\mu\nu}\gamma^5 n_+^\mu s_T^\nu\right]}$$

Unpolarized PDF



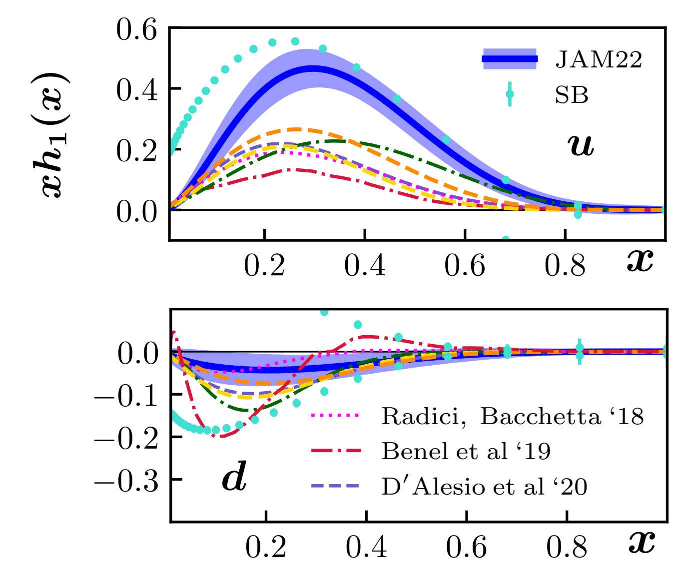
also NNPDF, MSHT,…

Helicity PDF

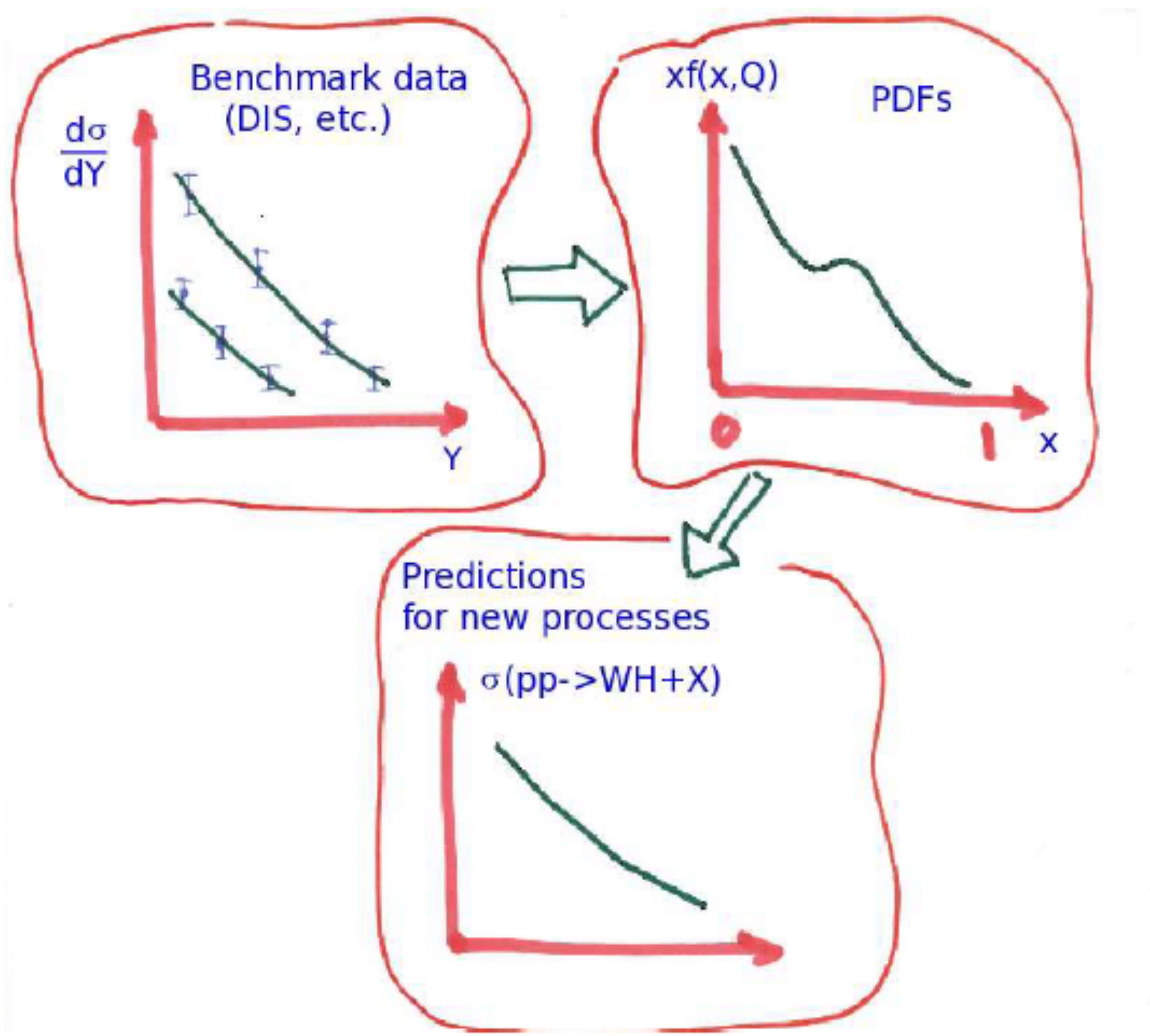


also JAM,…

Transversity PDF

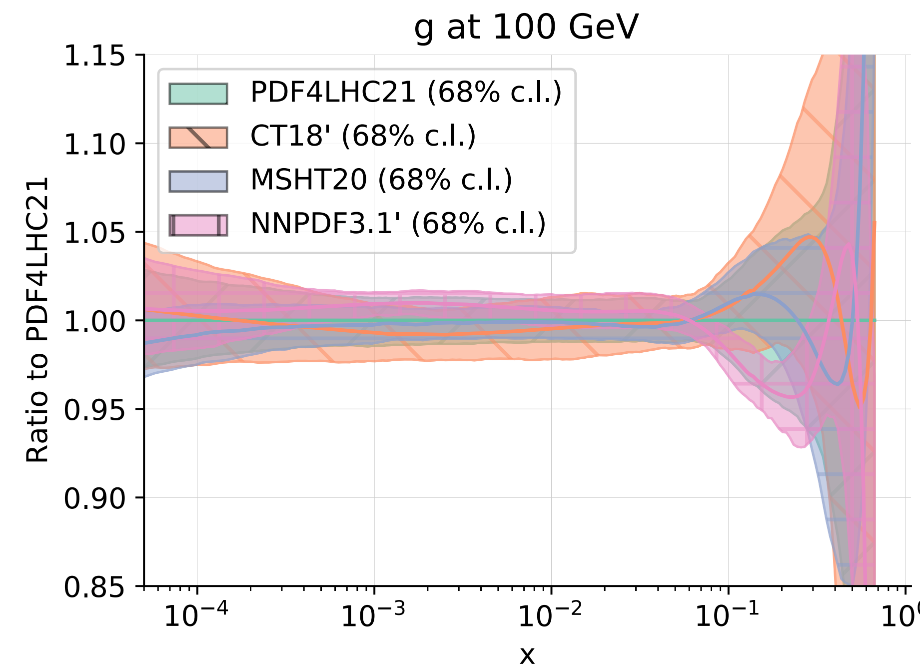


From non-perturbative QCD to EW/BSM/… physics

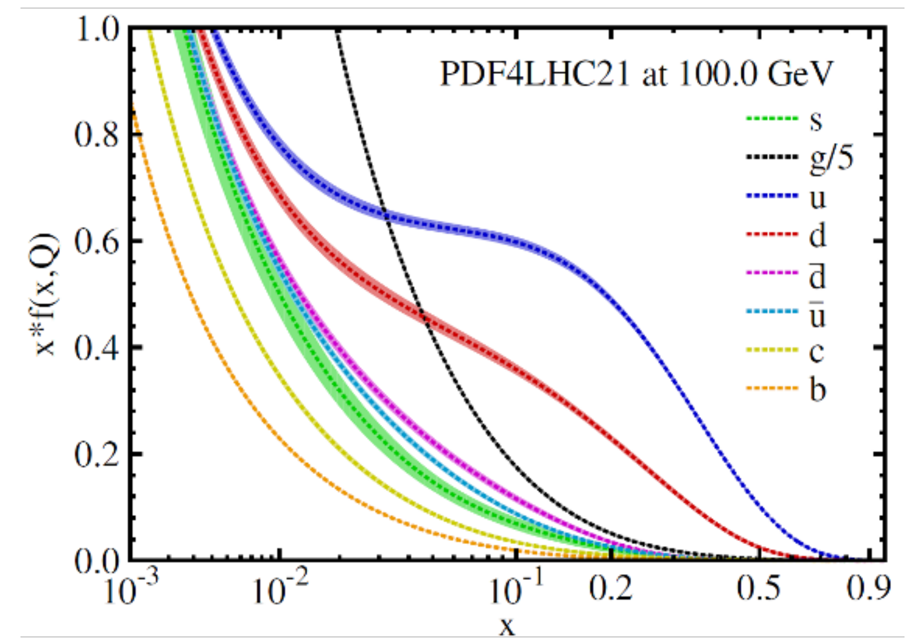


Benchmarking studies

Unpolarized PDFs from CT18, MSHT20 and NNPDF3.1.1. have been also thoroughly compared, benchmarked and combined during the PDF4LHC21 study.



PDF sets can, in turn, be used, say, at the LHC, Tevatron, EIC (future), … to predict cross sections,…

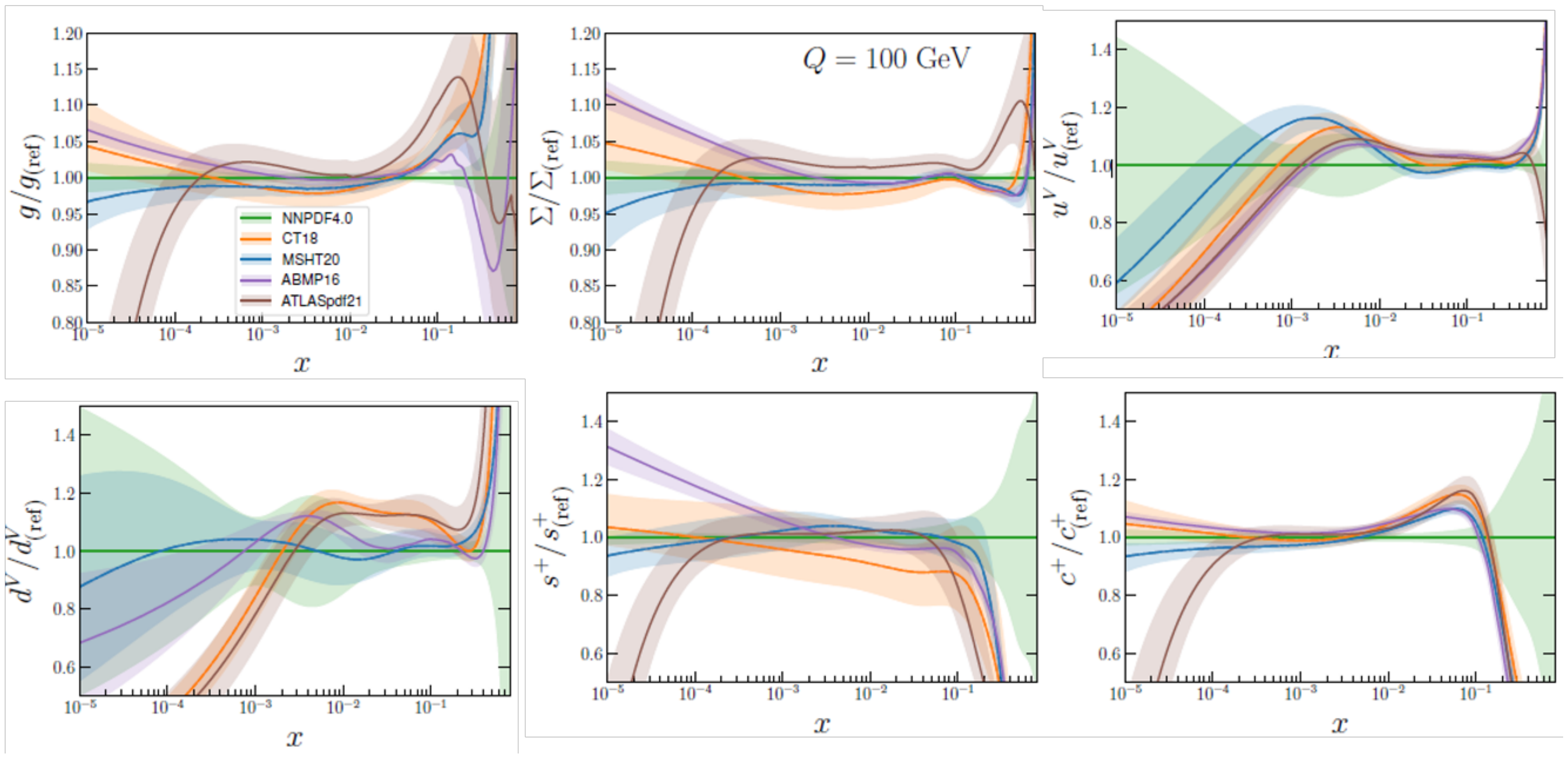


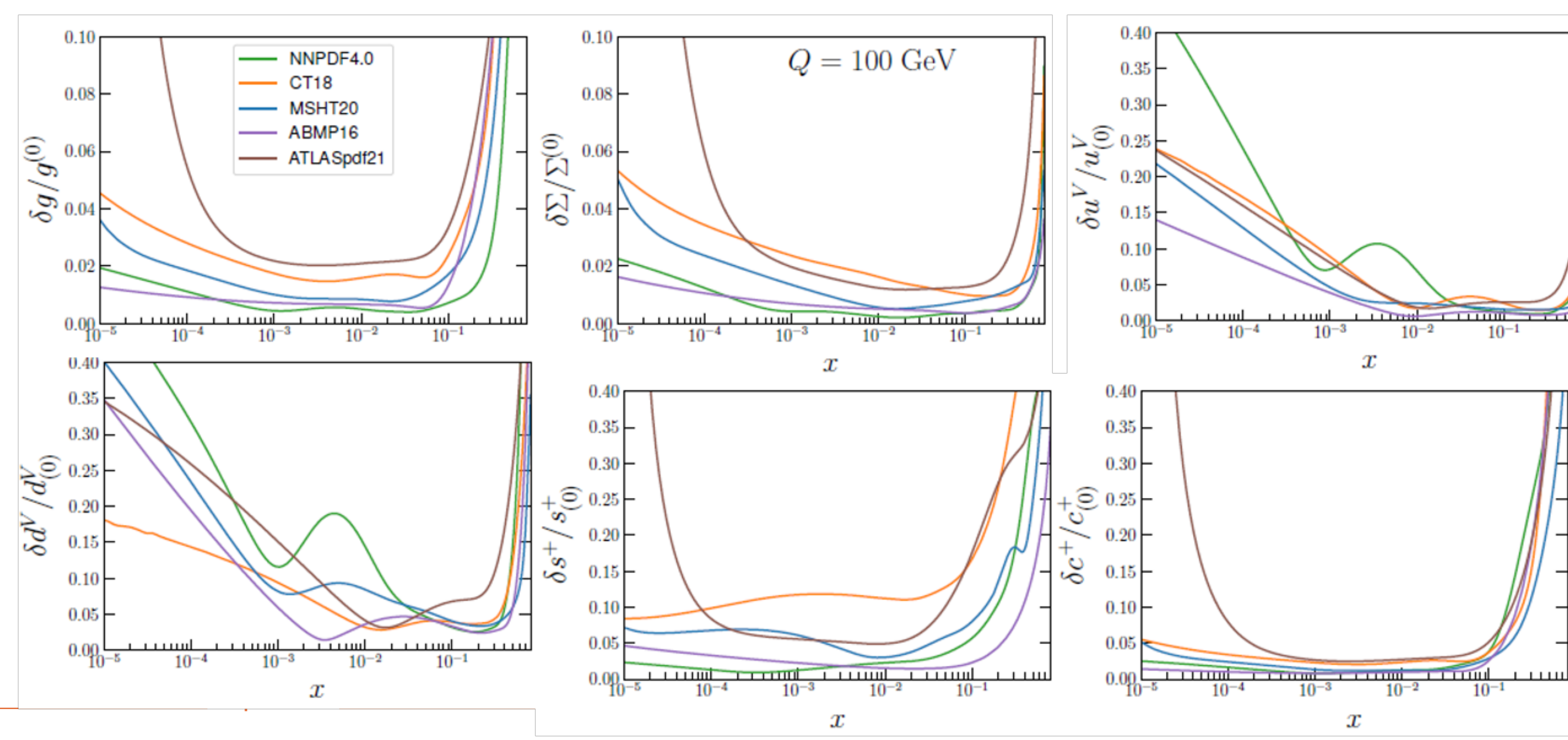
Recommendations of usage and combined set in [PDF4LHC Working Group, 2203.05506]

Recent advancements in the determination of unpolarized PDFs:

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

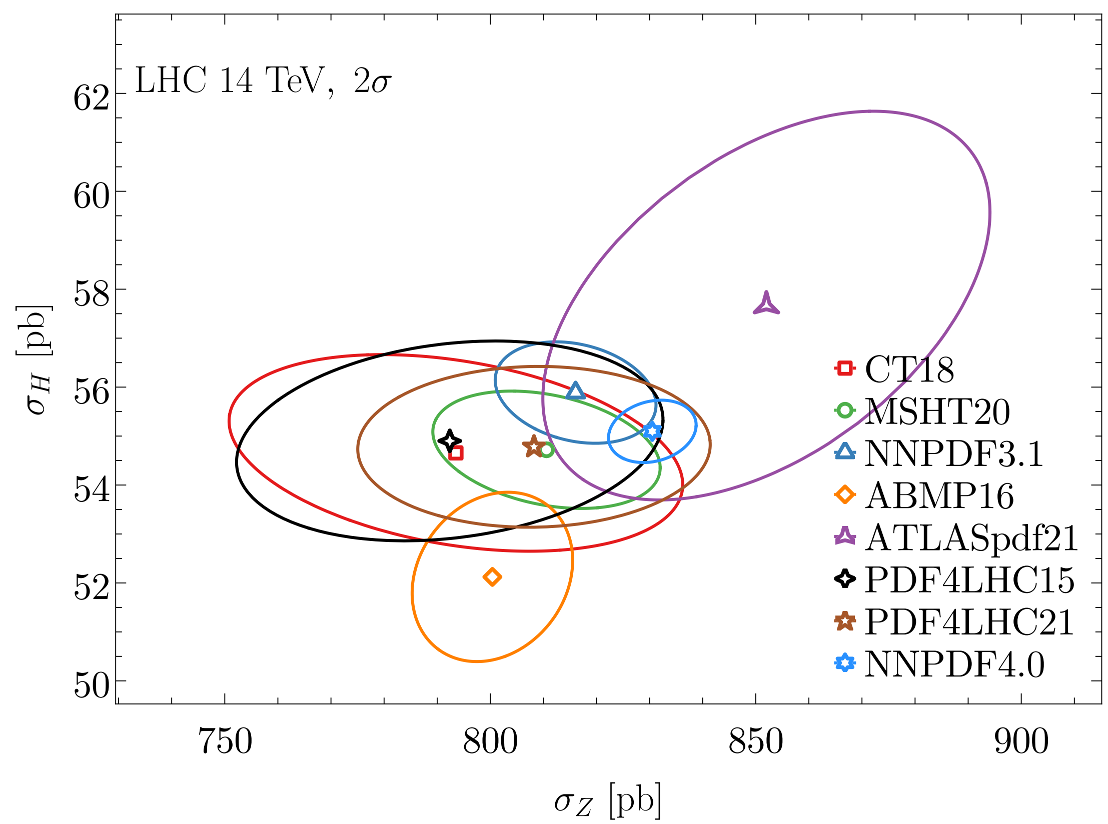
[Snowmass WP, 2203.13923]

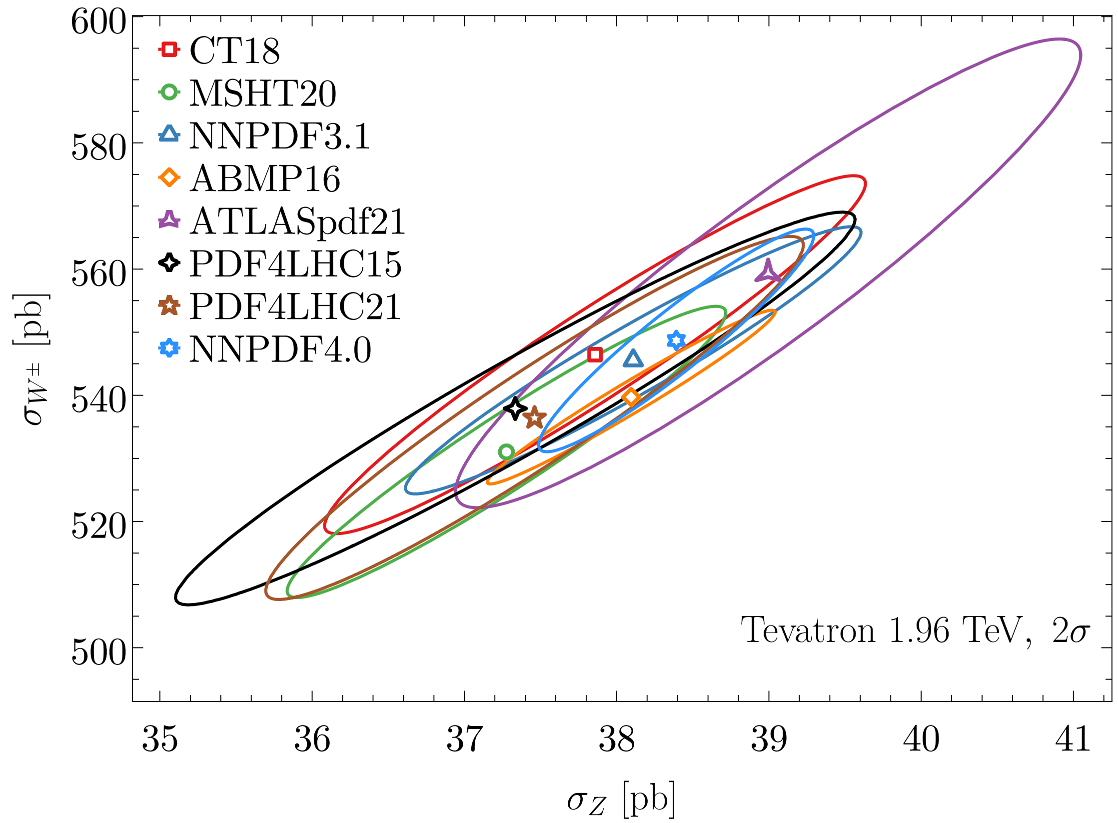




Uncertainties from global analyses of proton structure

Precision PDFs (Snowmass 21 WP) [2203.13923]





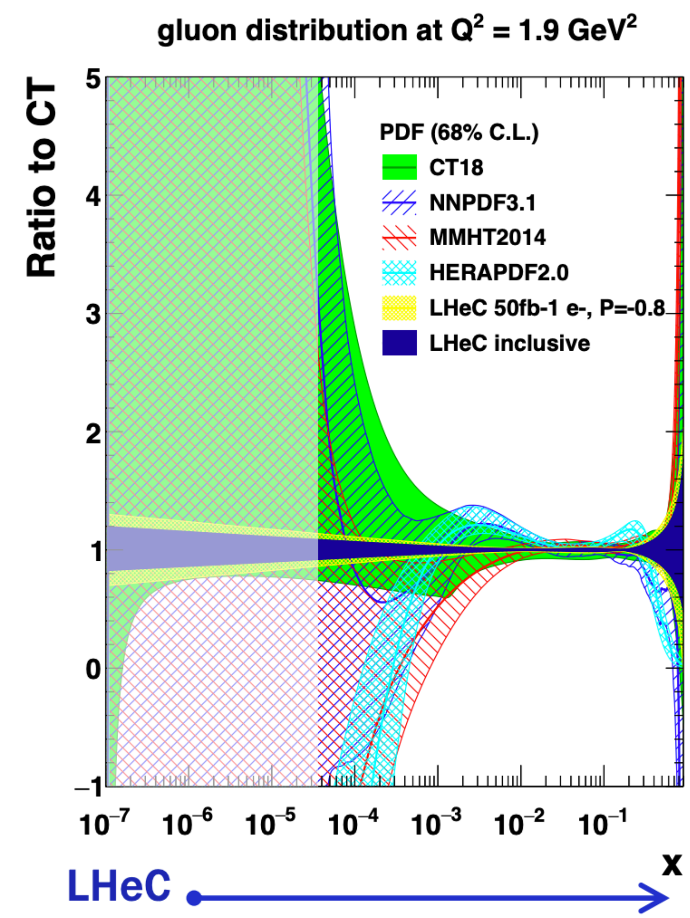
What is a faithful uncertainty coming from PDFs on those cross sections?

Reducing PDFs and α_s uncertainties for EW and BSM physics

Theoretical progress elevates precision on pQCD predictions.

Measurements of several SM parameters depend on PDF uncertainties.

Future experiments will potentially increase the precision of PDFs:
LHeC, EIC, HL-LHC,…



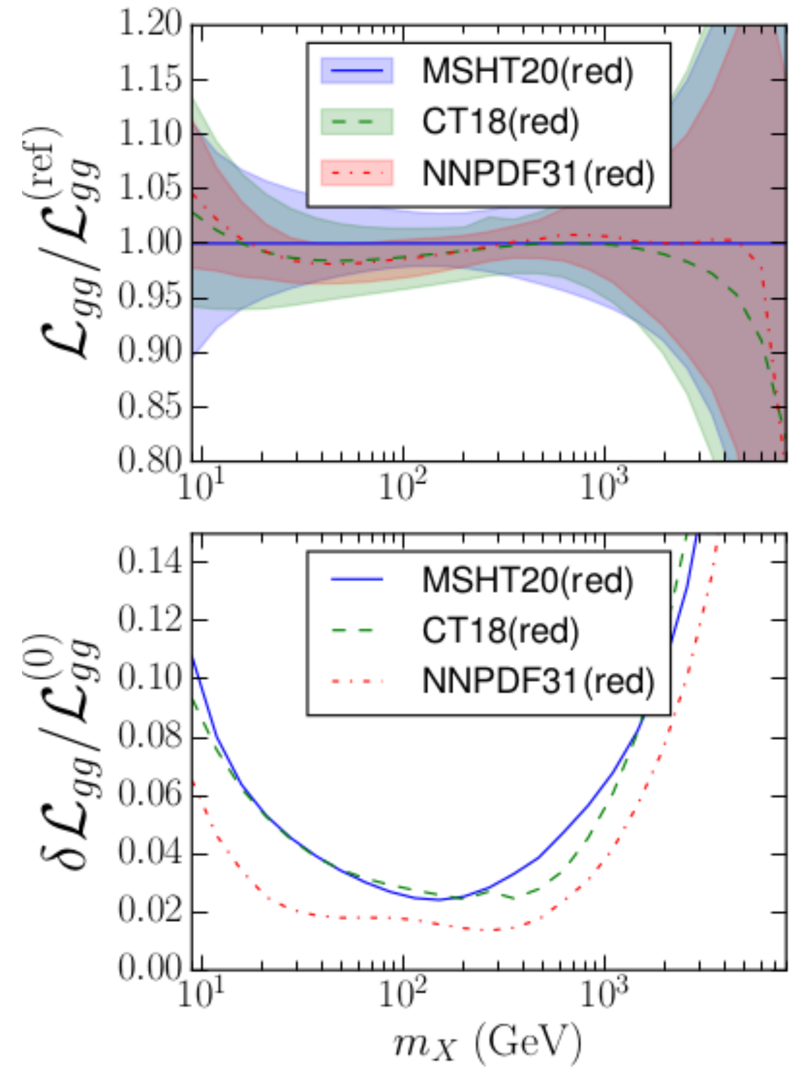
Plot from C. Gwenlan ICHEP 2020

Future global analyses will require **thorough understanding** of **various sources of uncertainties** in the PDF determination.

Do we understand the present uncertainty from PDF sets?

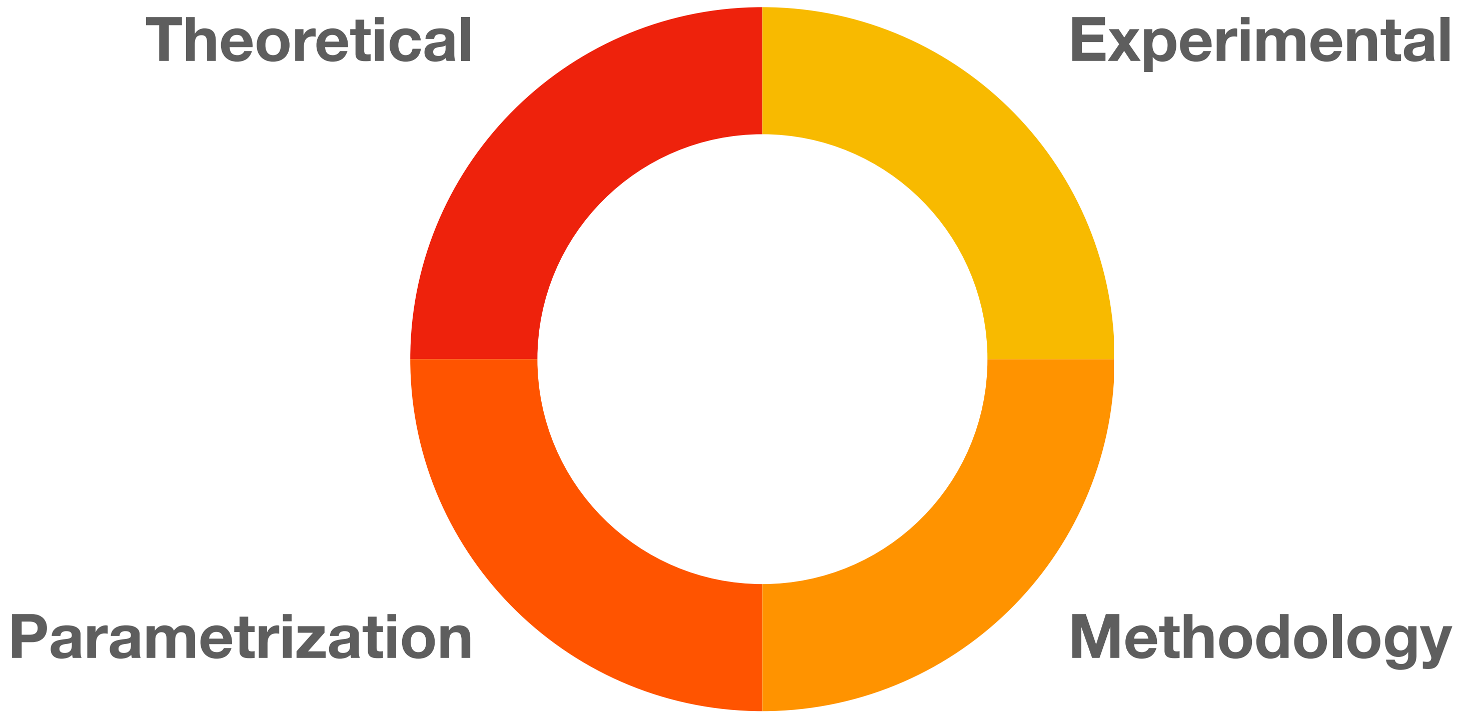
PDF4LHC21 benchmarking exercise:
comparison of uncertainties for same sets of data and QCD settings.

The uncertainties for CT18, MSHT20 and NNPDF3.1 reduced sets are still different. Key role played by methodology.



[PDF4LHC Working Group, 2203.05506]

Sampling biases contribute to PDF uncertainties



In all four categories of uncertainties, we can further distinguish

PDF fitting accuracy and *PDF sampling accuracy*.

Accuracy in theo., exp.,… inputs
—commonly integrated in global analyses.

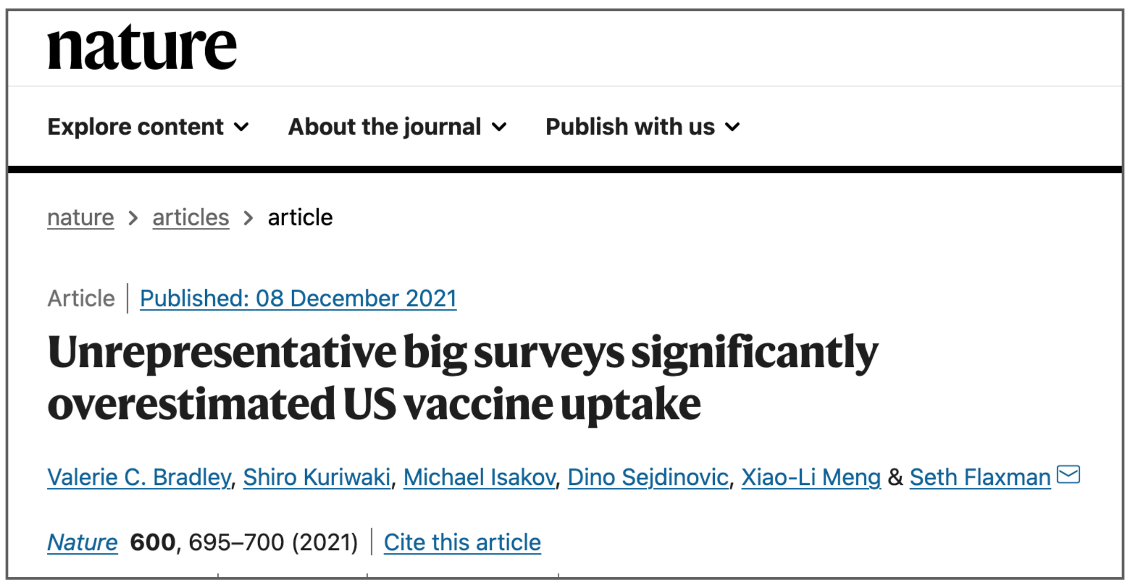
[Kovarik et al, Rev.Mod.Phys. 92 (2020)]

Adequacy of sampling the space of all
solutions — traditionally ignored.

Control for **sampling biases** in determination of PDFs plays a critical role.

Origin of sampling biases — experience with large population surveys

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.



nature

Explore content About the journal Publish with us

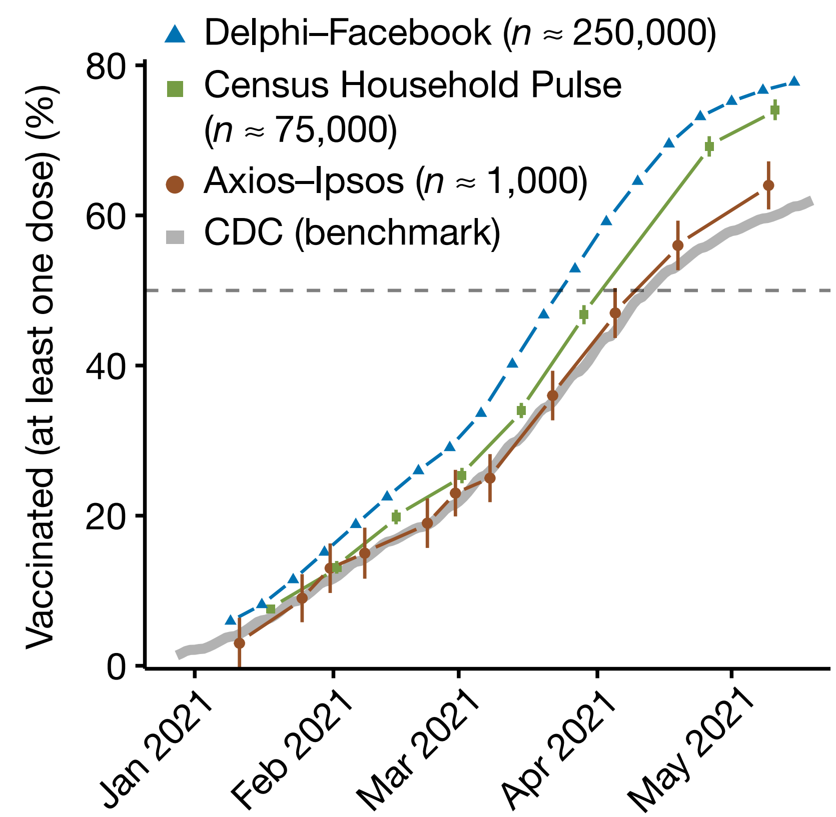
nature > articles > article

Article | Published: 08 December 2021

Unrepresentative big surveys significantly overestimated US vaccine uptake

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng & Seth Flaxman

Nature **600**, 695–700 (2021) | Cite this article



Based on
[Xiao-Li Meng, The Annals of Applied Statistics, Vol. 12 (2018), p. 685]

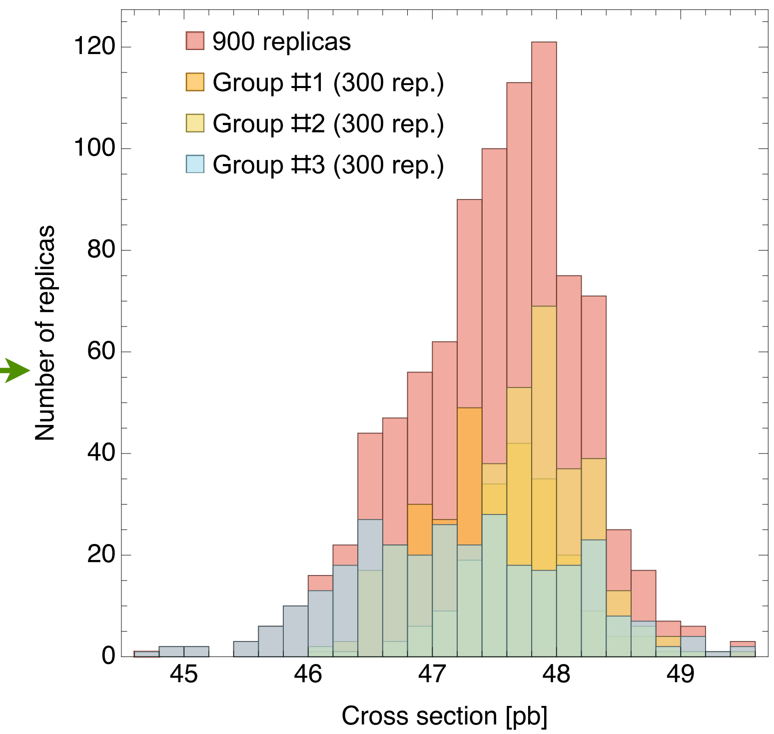
The deviation has been traced to the **sampling bias.**
In contrast to the statistical error, the sampling bias can involve growth with the size of the sample.

Law of large numbers

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

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

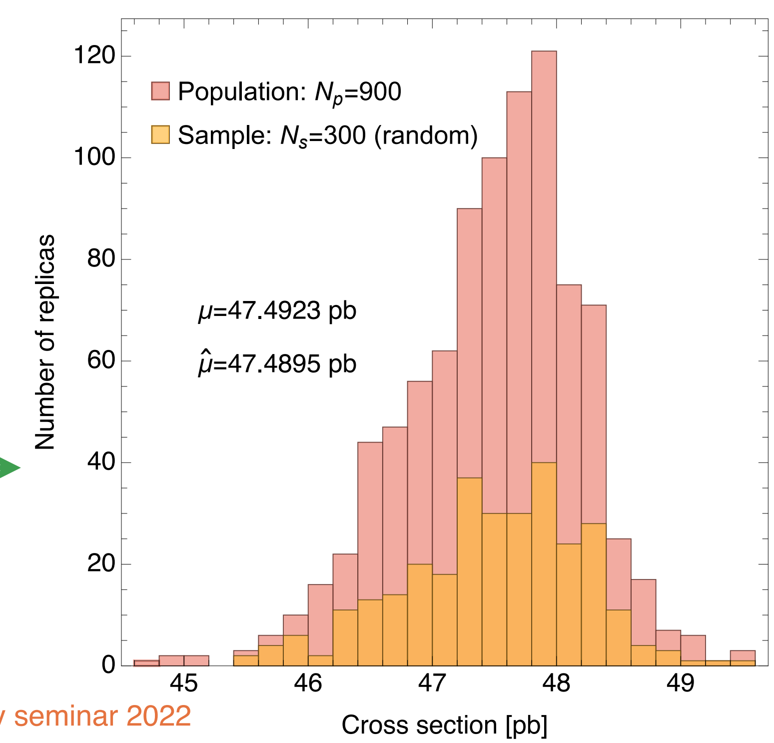
with σ the standard deviation, μ the true and $\hat{\mu}$ the determined values. *That's the law of large numbers.*



A toy sampling excercise

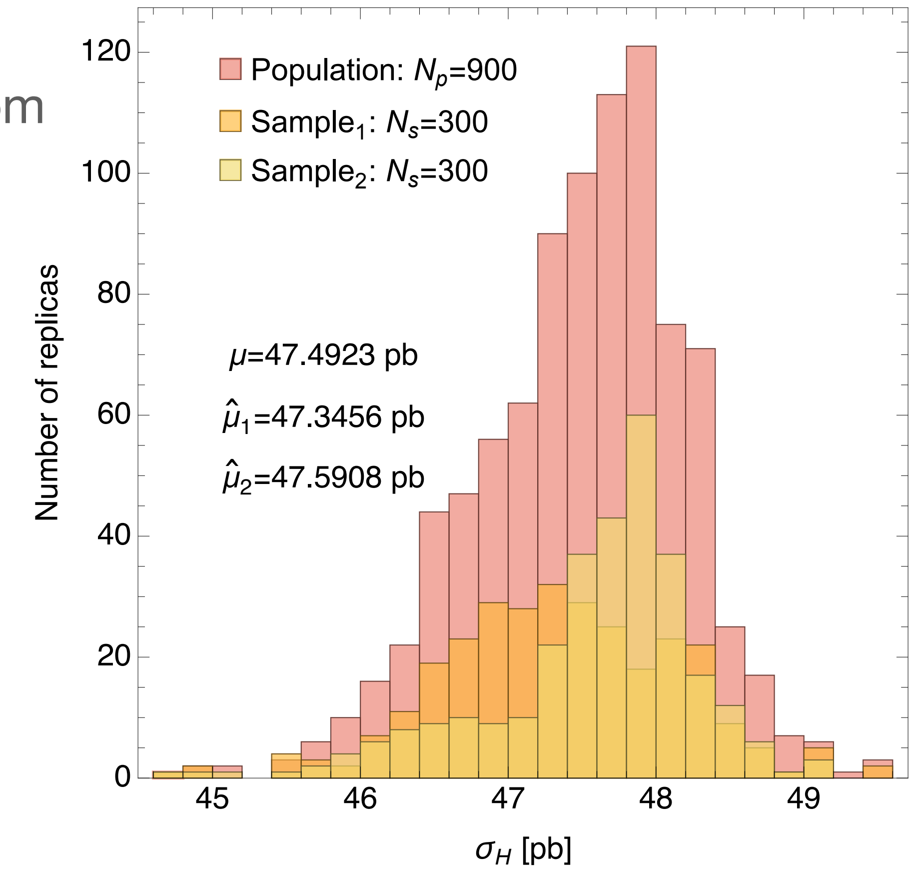
We take 300×3 groups of Higgs cross sections evaluated by 3 different groups.

We **randomly** select 300 out of the 900 cross sections.
The law of large number 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 $\sigma/\sqrt{n}$!

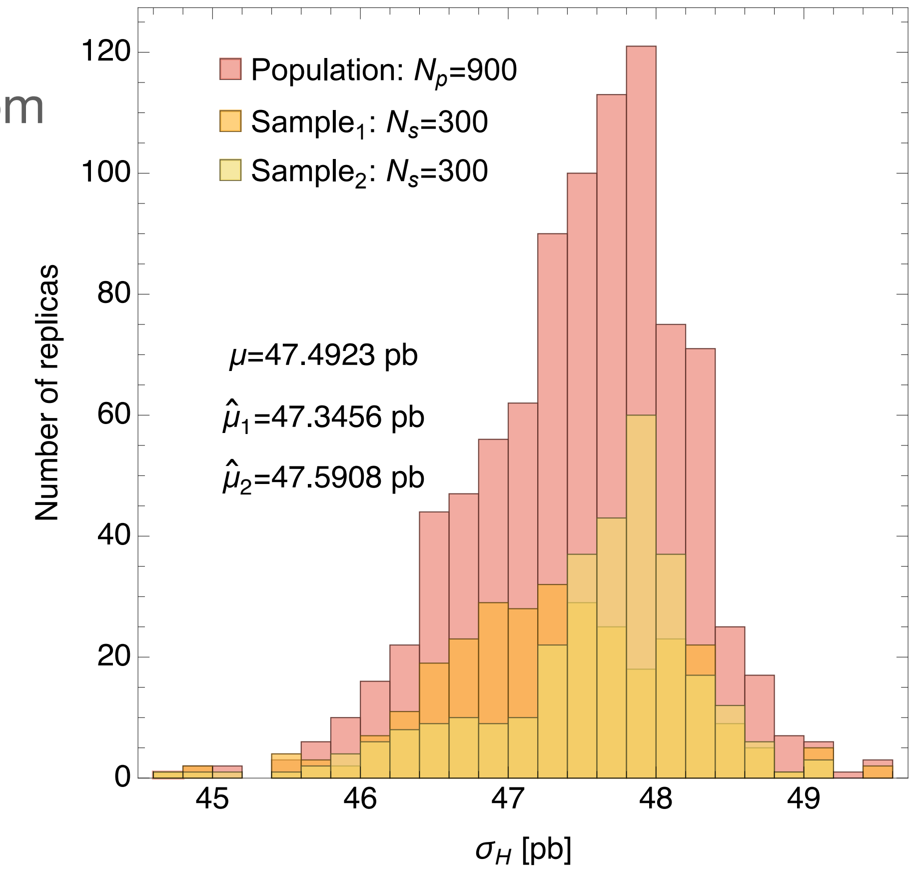


The law of large numbers obviates the *quality of the sampling*.

Role of the distribution of n for a population size N/measure of the parameter space in estimating the deviation $\mu - \hat{\mu}$?

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 $\sigma/\sqrt{n}$!



The law of large numbers obviates the *quality of the sampling*.

Role of the distribution of n for a population size N/measure of the parameter space in estimating the deviation $\mu - \hat{\mu}$?

The **trio identity** remedies to that problem be accounting for sampling bias:

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

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

$$\boxed{\mu - \hat{\mu} = \underbrace{(\text{confounding correlation})}_{\text{depends on the sampling algorithm}} \times \underbrace{(\text{measure discrepancy}) \times (\text{inherent problem difficulty})}_{\text{can tend to } \sigma/\sqrt{n} \text{ for random sampling}}}$$

For a sample of n items from the population of size N, we can consider an array built by the random spanning of the binary responses of the $N - n$ (0) and n (1) items, so that

$$\boxed{\mu - \hat{\mu} = \mathrm{Corr}[\text{observable, sampling quality}] \times \sqrt{\frac{N}{n} - 1} \times \sigma(\text{observable})}$$

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

The sample deviation can be large if the sampling is not sufficiently random.
Standard error estimates can be misleadingly small.

➪ critical role of controlling for **sampling biases** in determination of PDFs.

Uncertainty on QCD observables — the hopscotch

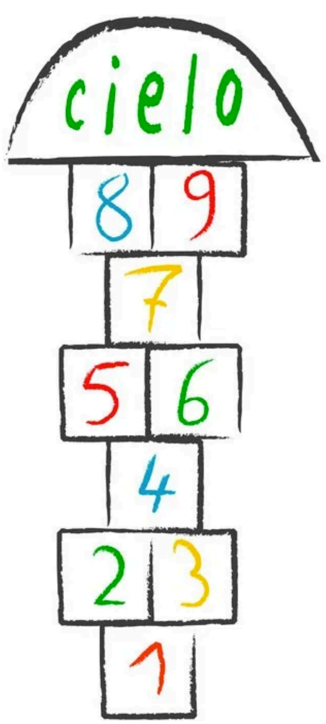
Sampling of multidimensional spaces ($d \gg 20$) can be exponentially inefficient and requires $n > 2^d$ replicas to obtain a convergent expectation value.

[Hickernell, MCQMC 2016, 1702.01487]
[Sloan,I.H.,Wo´zniakowski, 1997]

Specific QCD observables: only few effective large dimensions contribute the bulk of the uncertainty.
E.g. compressing MC PDFs into a Hessian set: we construct a basis to identify such large dimensions.

Hopscotch scans:

estimation of a representative uncertainty on a cross section σ.



The release of a public code for NNPDF4.0's new methodology provide a perfect playground to explore the role of sampling.

[NNPDF, EPJC 81]

Hopscotch scans

Estimation of a representative uncertainty on a cross section σ

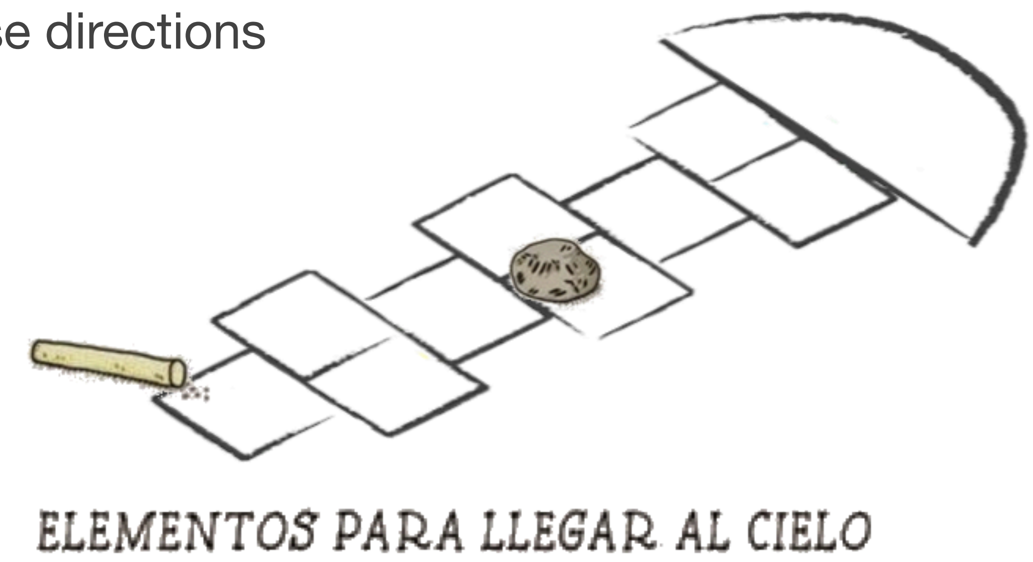
To sample the PDF dependence: sample primarily the coordinates with large variations of σ.

We employ:

1. Basis coordinates in the PDF space

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



Details in preprint [2205.10444]

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

Step 1

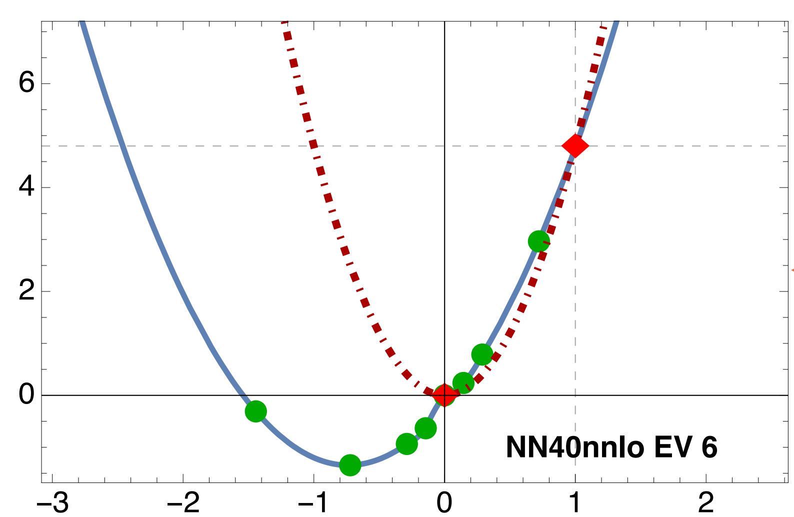
The NNPDF4.0 Hessian set ($n = 50$) defines a coordinate system on a manifold corresponding to the largest variations of the PDF uncertainty —red dots and curve.

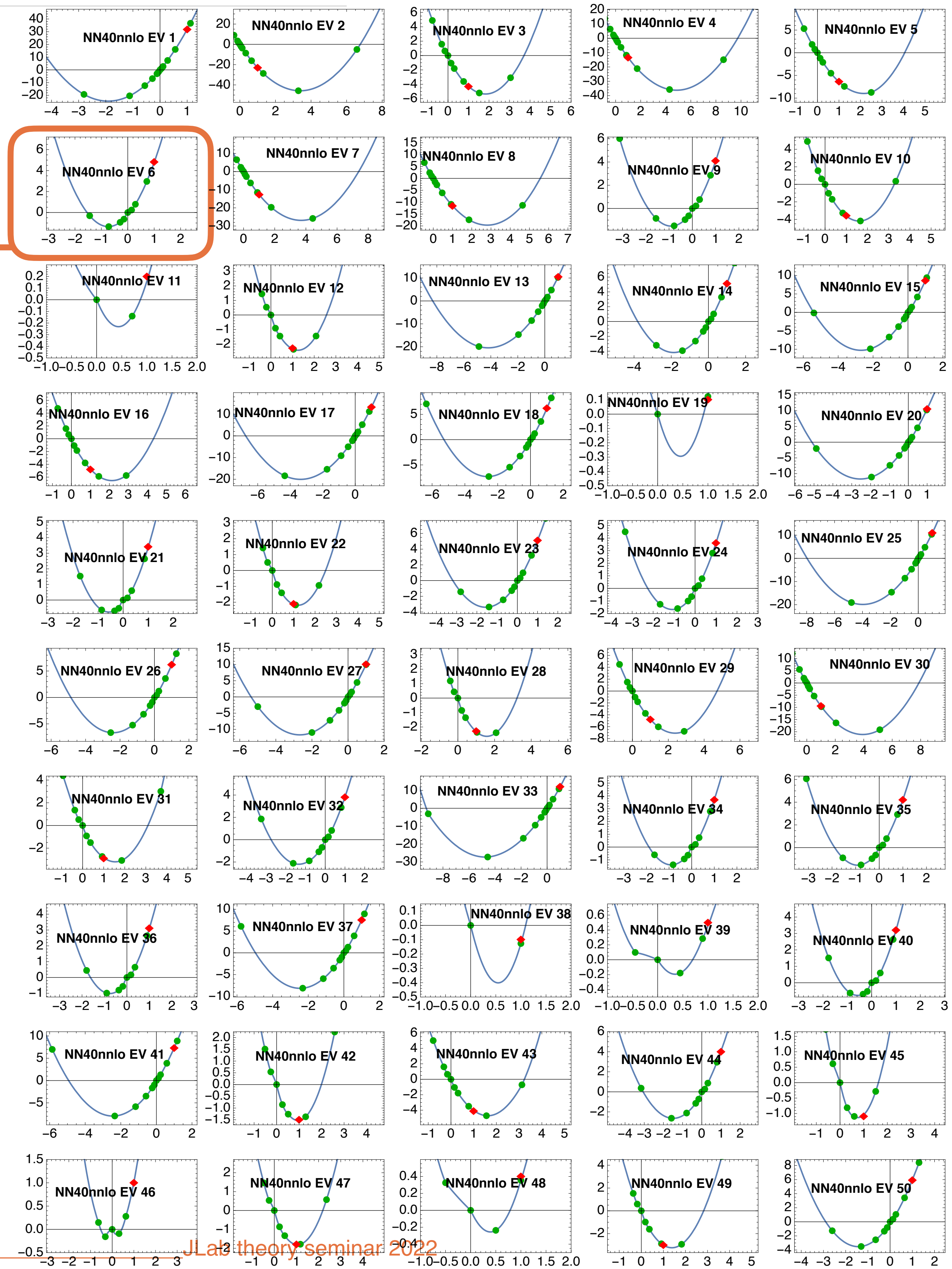
[NNPDF, 2109.02653]

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 suggest that there exist solutions with lower χ^2 — green dots and blue curve.



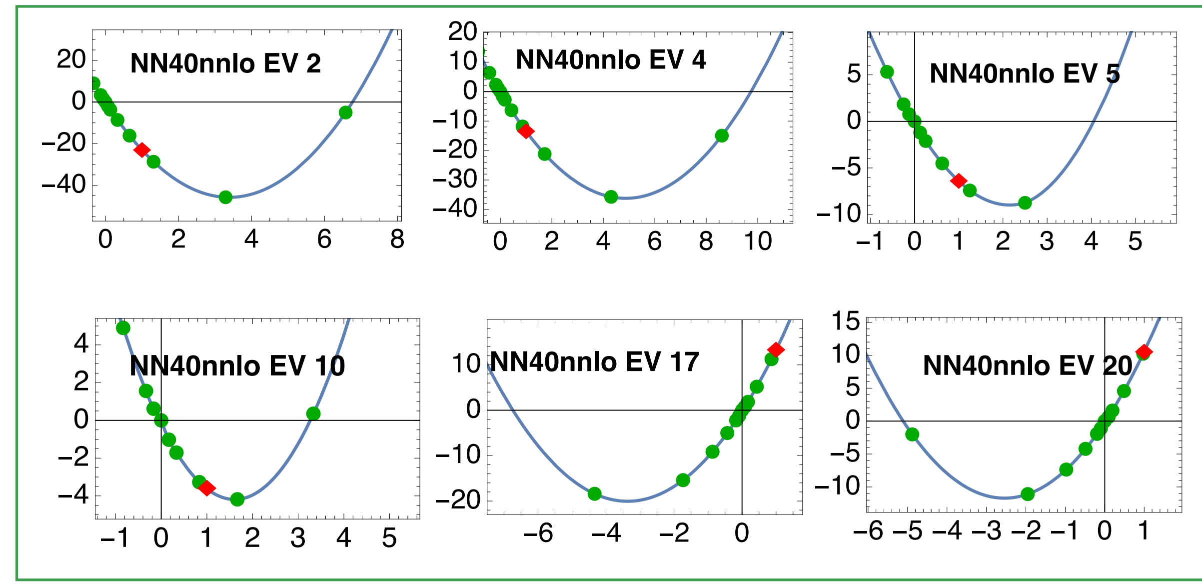


A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

Step 3

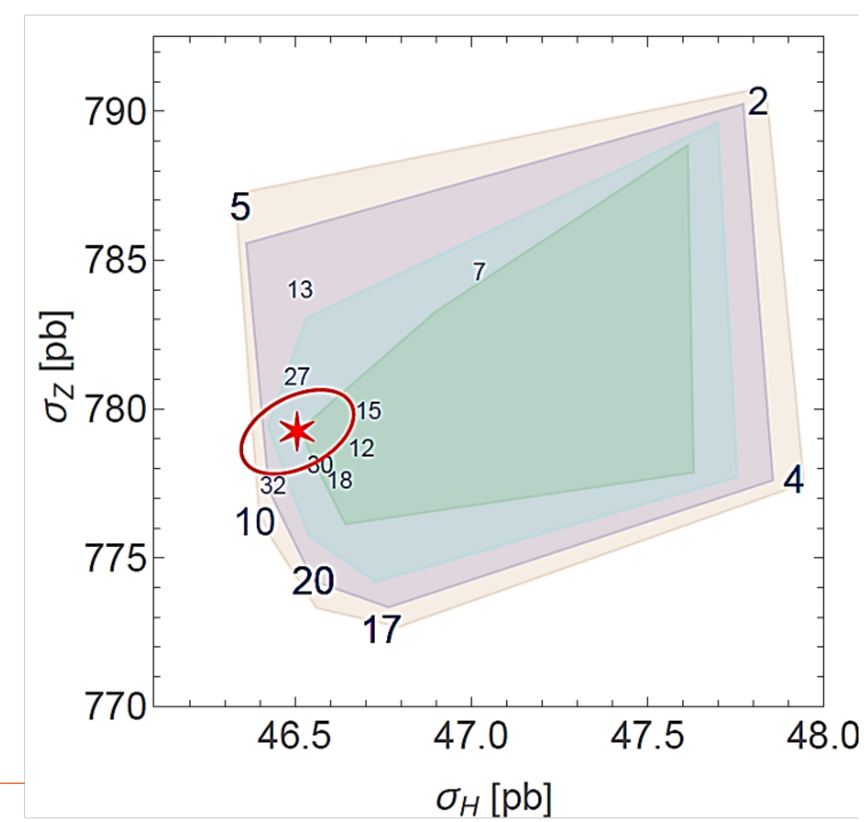
Guidance from specific cross sections:
we identify 4-7 EV directions that give the largest displacements for a given $\Delta\chi^2$ per pair.

E.g., σ_Z vs. σ_H is represented by the 6 corners of a projected "octahedron," corresponding to large EV directions: 2, 4, 5, 10, 17, 20.



Other directions generally give smaller displacements.
Large EV directions are shared among various pairs of cross sections.

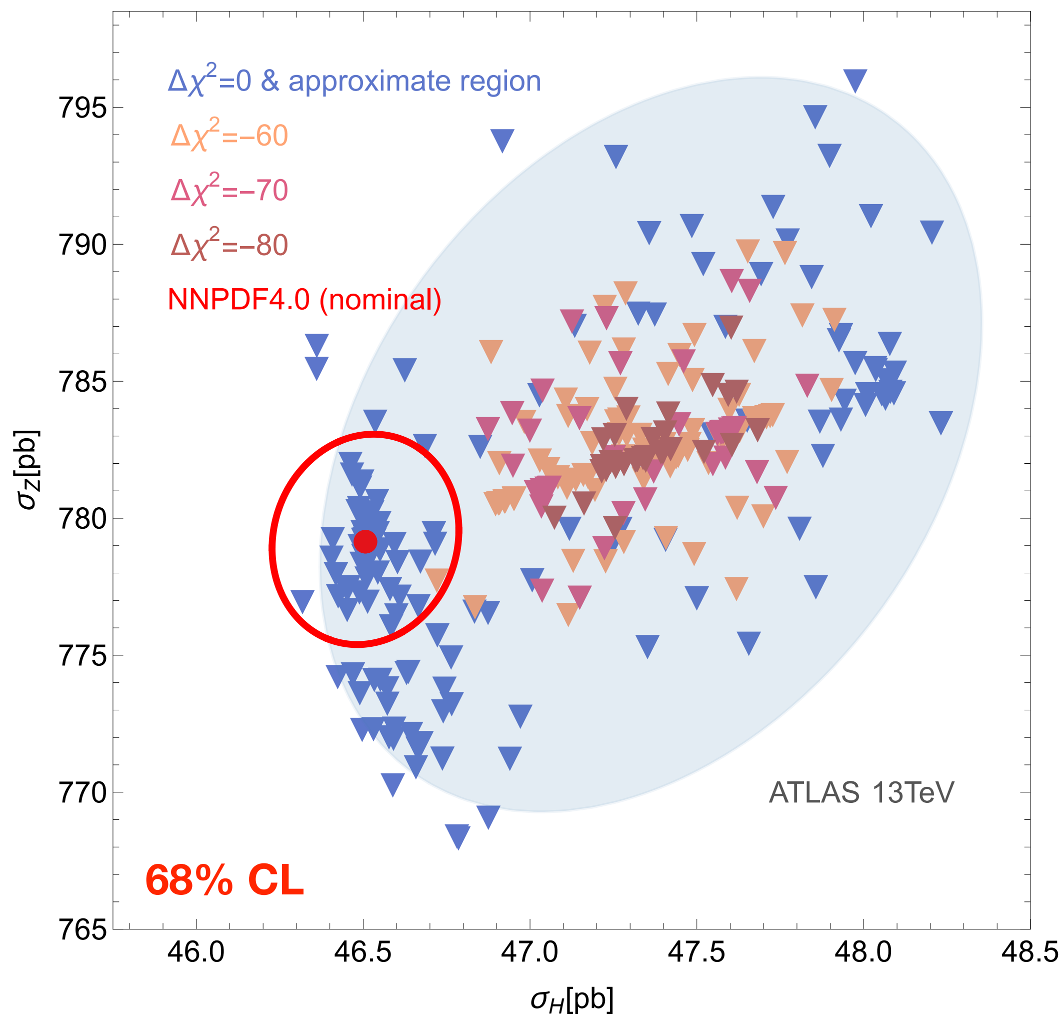
The contours are for $\Delta\chi^2 = +10, 0, -10, -20$ *w.r.t.* NNPDF4.0 replica 0 (red).



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

Step 4

For each pair of cross sections, we generate 300 replicas by sampling uniformly along the large EV directions. Sort the $n_{pairs} \times 300$ resulting replicas according to their $\Delta\chi^2$ *w.r.t.* to NN40 replica 0.



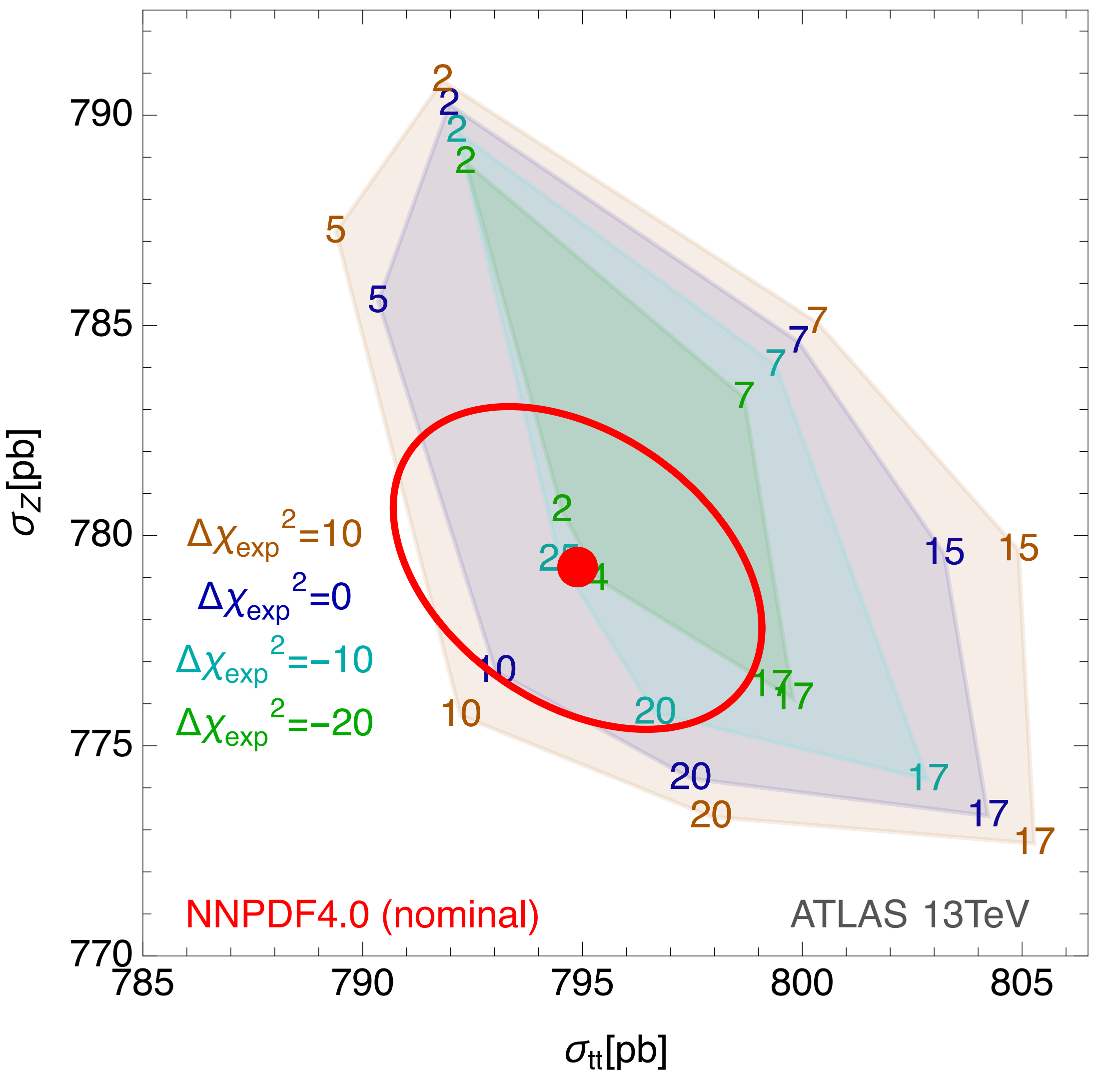
Each of the $\Delta\chi^2 = 0 \pm 3$ replicas is an acceptable PDF set from the NNPDF4.0 fit.

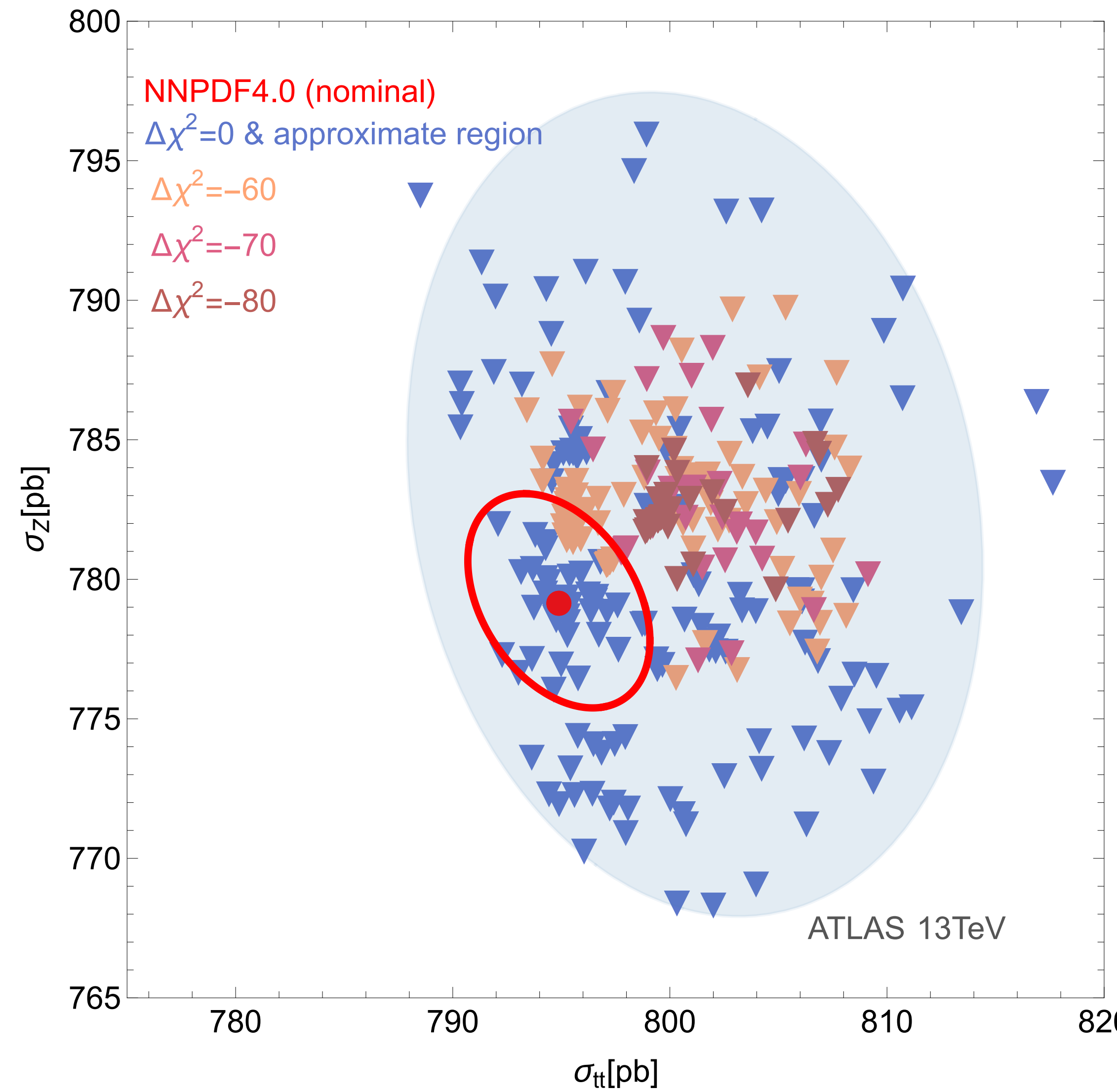
The blue ellipse (constructed using a convex hull method) is an approximate region containing all found replicas with $\Delta\chi^2 = 0 \pm 3$.

[Anwar, Hamilton, Nadolsky, 1901.05511]

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

A hopscotch scan: details

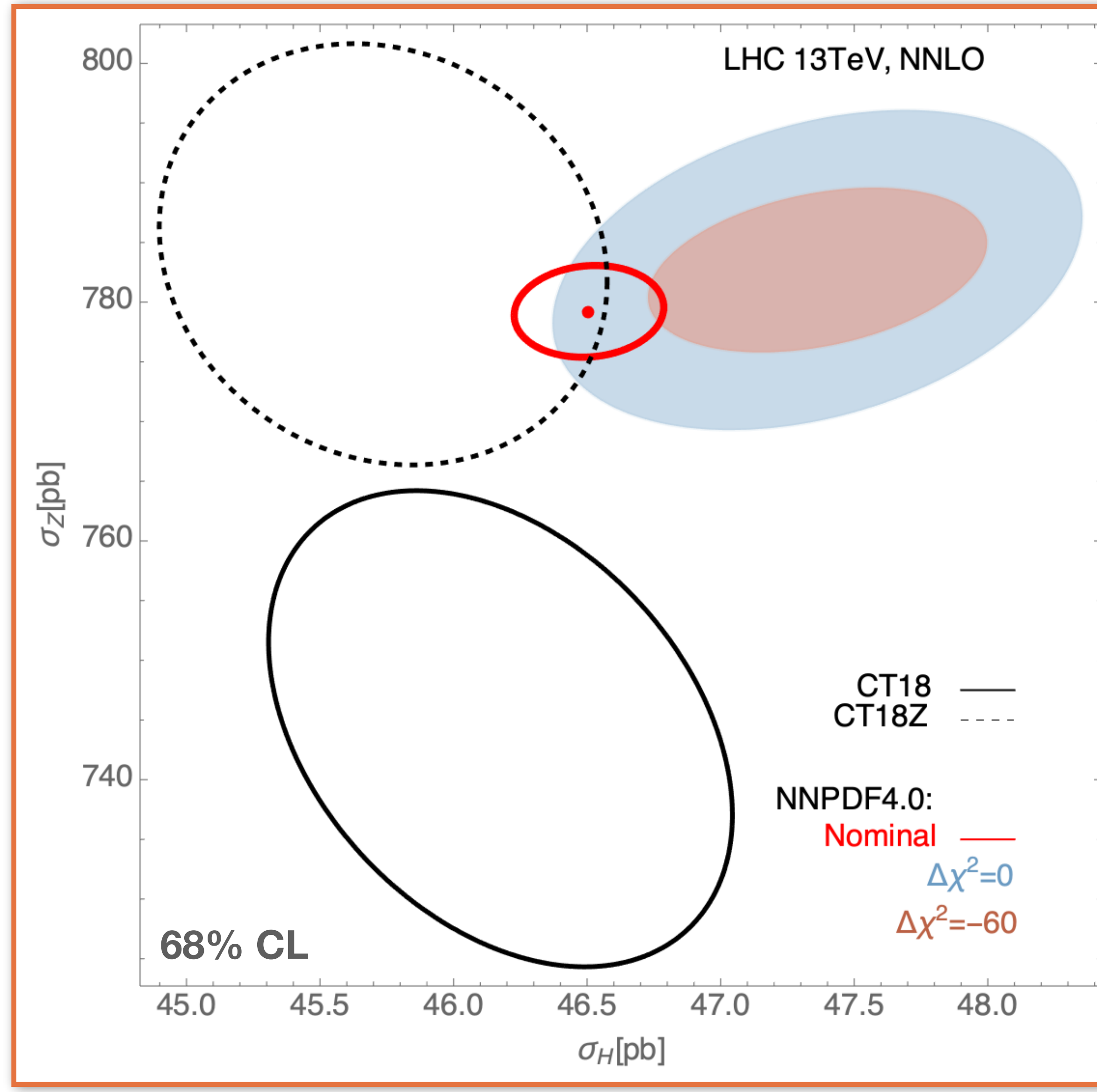




Explored n_{pairs} pairs of cross sections are:

- $\{\sigma_{t\bar{t}}, \sigma_Z\}$ EV directions 5, 2, 7, 15, 17 and 17, 20, 6, 10, 5
- $\{\sigma_Z, \sigma_{W^\pm}\}$ EV directions 2, 7, 23, 20, 17, 5
- $\{\sigma_{W^+}, \sigma_{W^-}\}$ EV directions 2, 13, 1,17, 14
- $\{\sigma_{t\bar{t}}, \sigma_H\}$ EV directions 8, 15, 17, 4, 2, 5

Monte-Carlo sampling for PDF parametrizations: cross sections for LHC



Blue and brown filled ellipses:

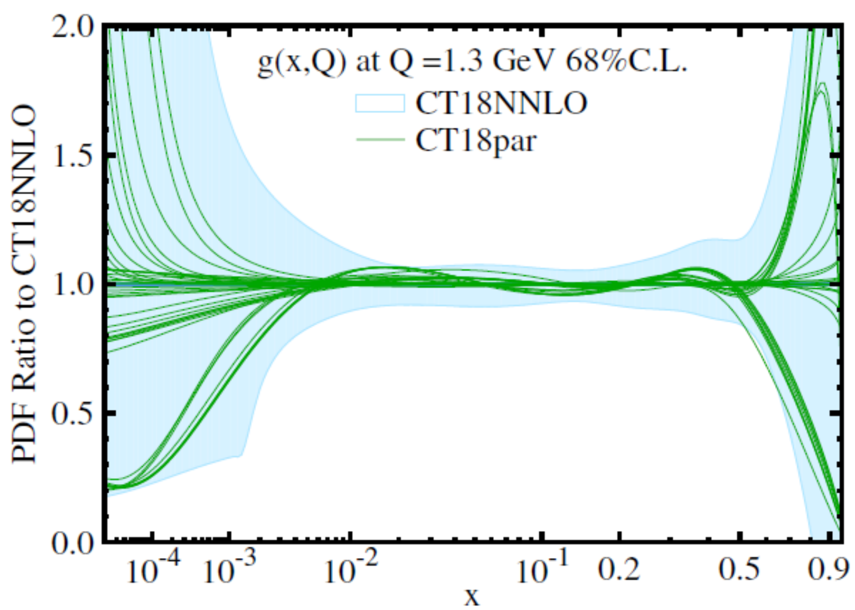
- areas of possible solutions corresponding to an equal $(\Delta\chi^2 = 0)$ or lower $(\Delta\chi^2 = -60)$ chi square w.r.t. the nominal solution
- found through the hopscotch scan — a dimensionality reduction method.
- size of blue areas comparable to 68% CL CT18 ellipses

Monte-Carlo sampling for PDF parametrizations: cross sections for LHC

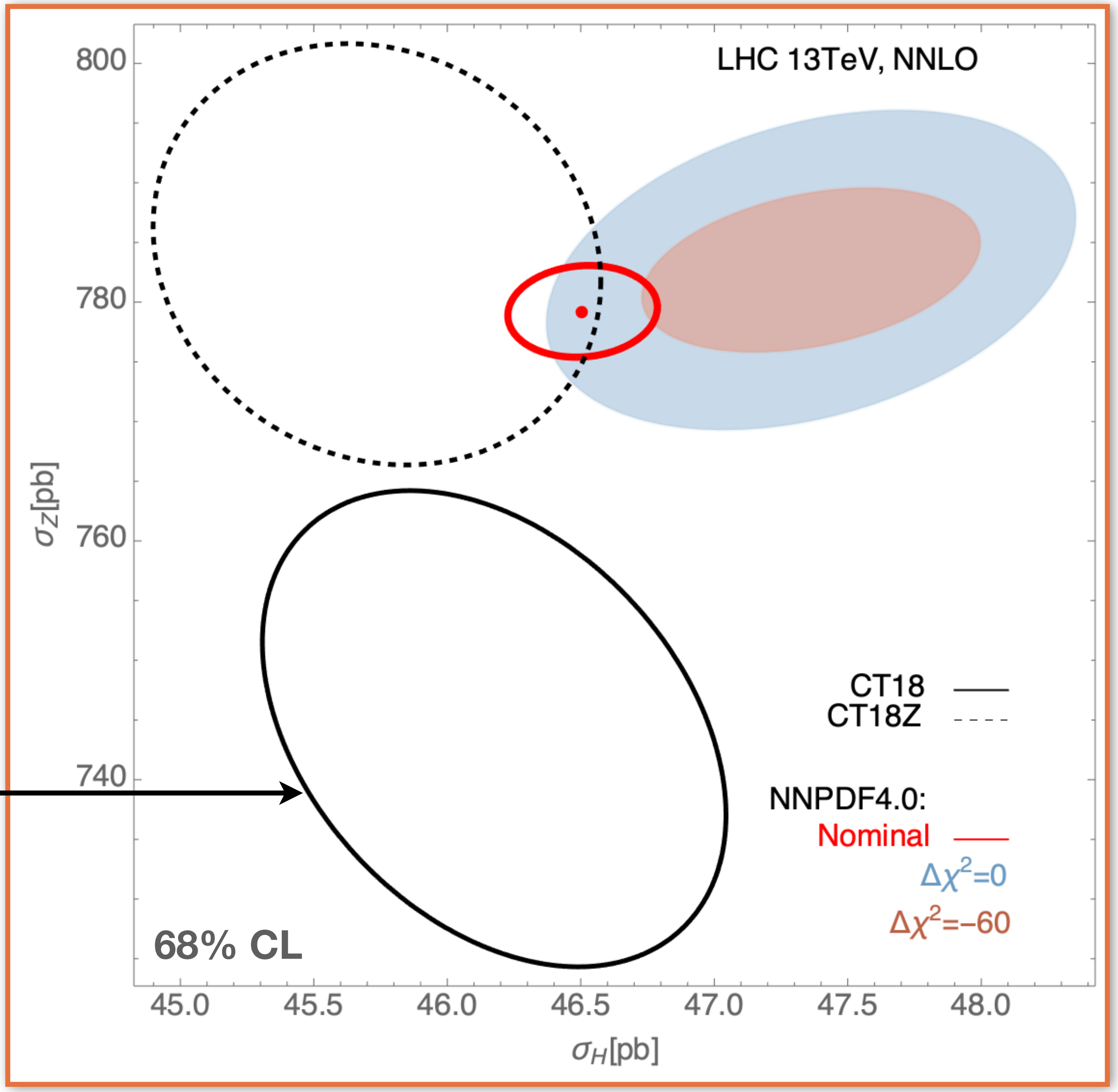
CT18 PDF uncertainty:

*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:

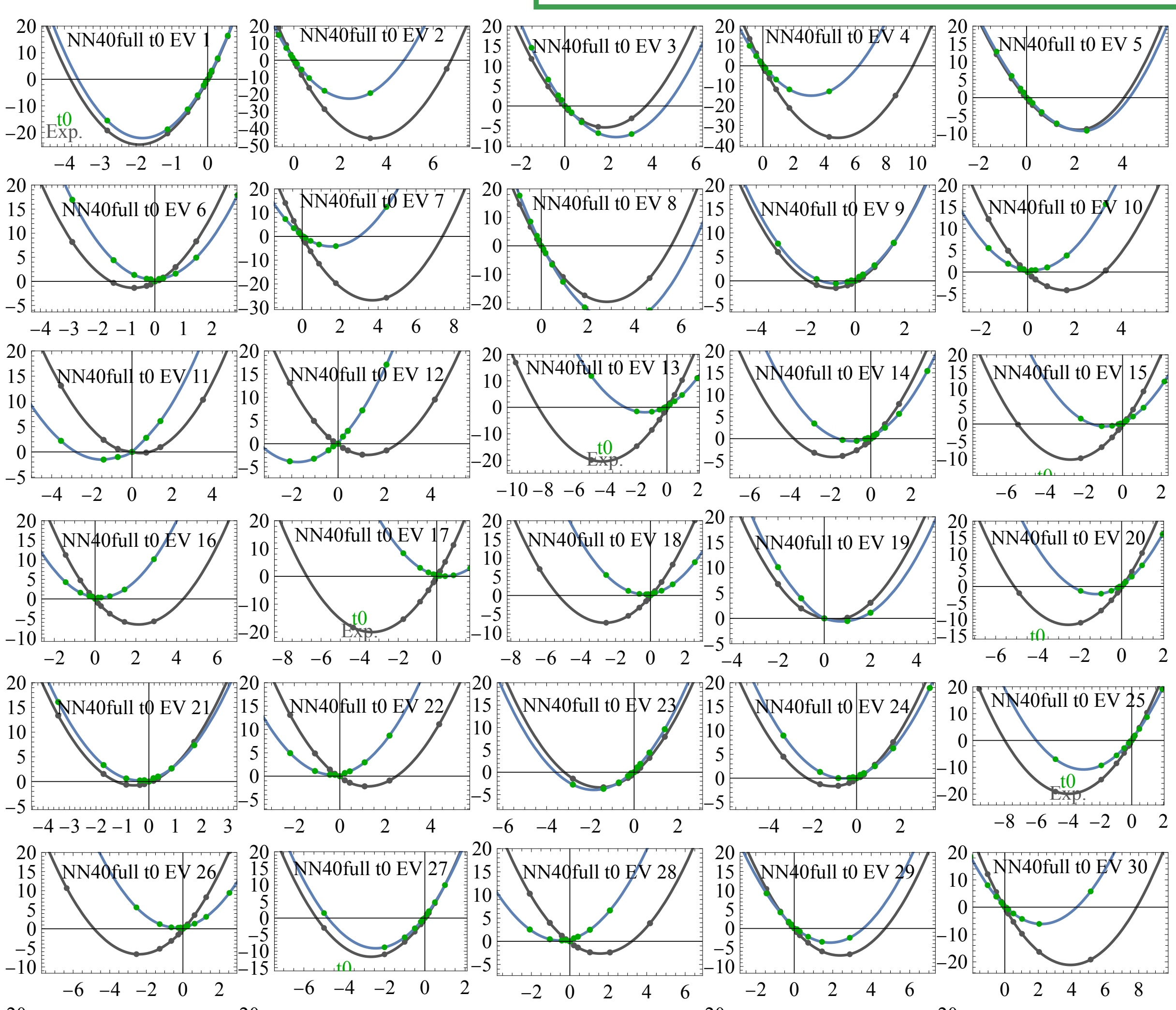
- areas of possible solutions corresponding to an equal ($\Delta\chi^2 = 0$) or lower ($\Delta\chi^2 = -60$) chi square w.r.t. the nominal solution
- found through the hopscotch scan — a dimensionality reduction method.
- size of blue areas comparable to 68% CL CT18 ellipses

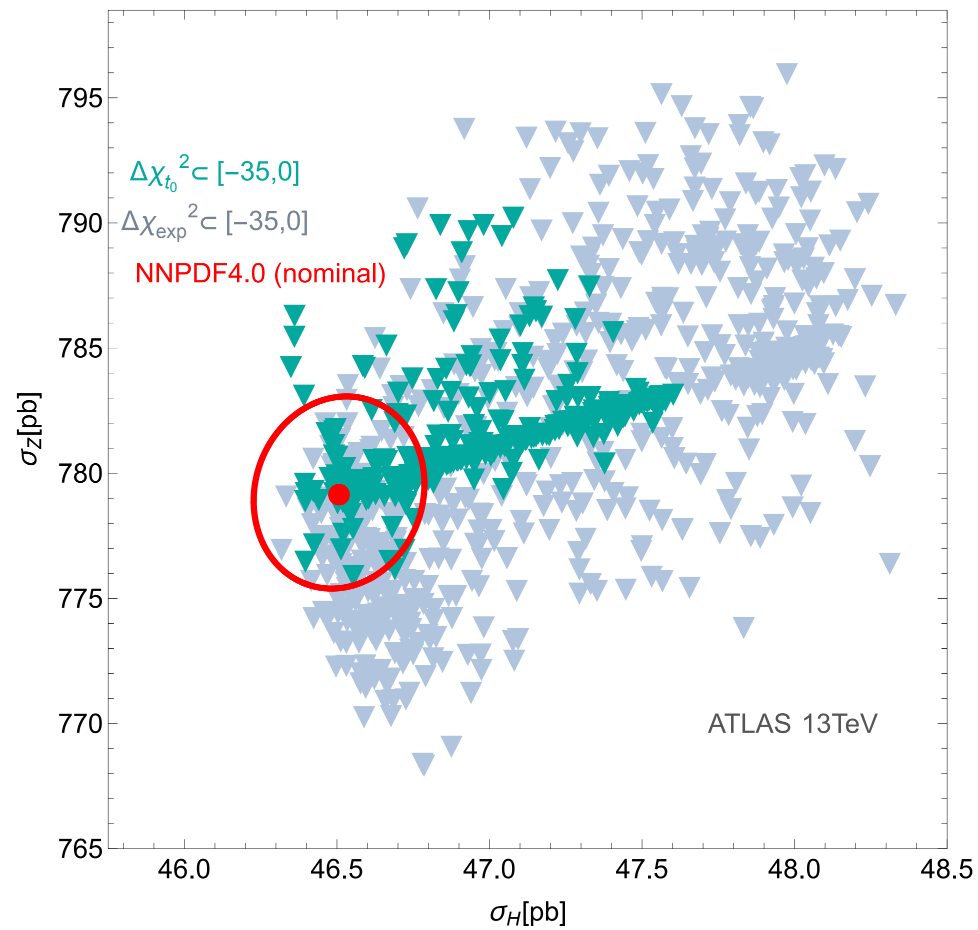
Monte-Carlo sampling for PDF parametrizations: choice of χ^2 form

Different definitions for the χ^2 form will affect the PDF uncertainty

- Experimental prescription for correlated systematic errors— used in our work.
- NNPDF4.0 uses the t_0-prescription definition for their tabulated χ^2
- Other groups use different prescriptions.

Some of the 50 EV sets for t_0 and exp. prescriptions

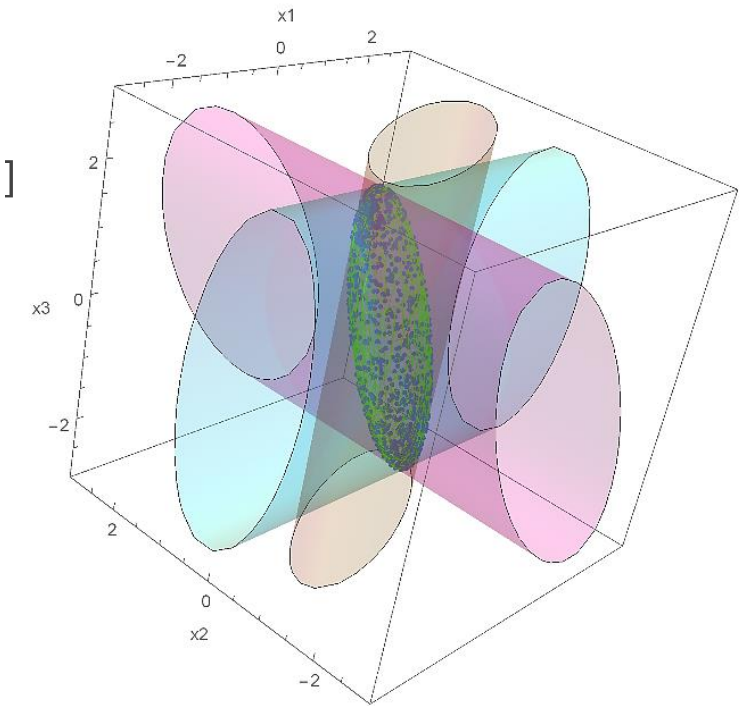




Main observations hold.
Generally smaller shifts from the nominal NN4.0 predictions than with the "exp" prescription.

Convex-hull ellipse reconstruction

[Anwar, Hamilton, Nadolsky, 1901.05511]

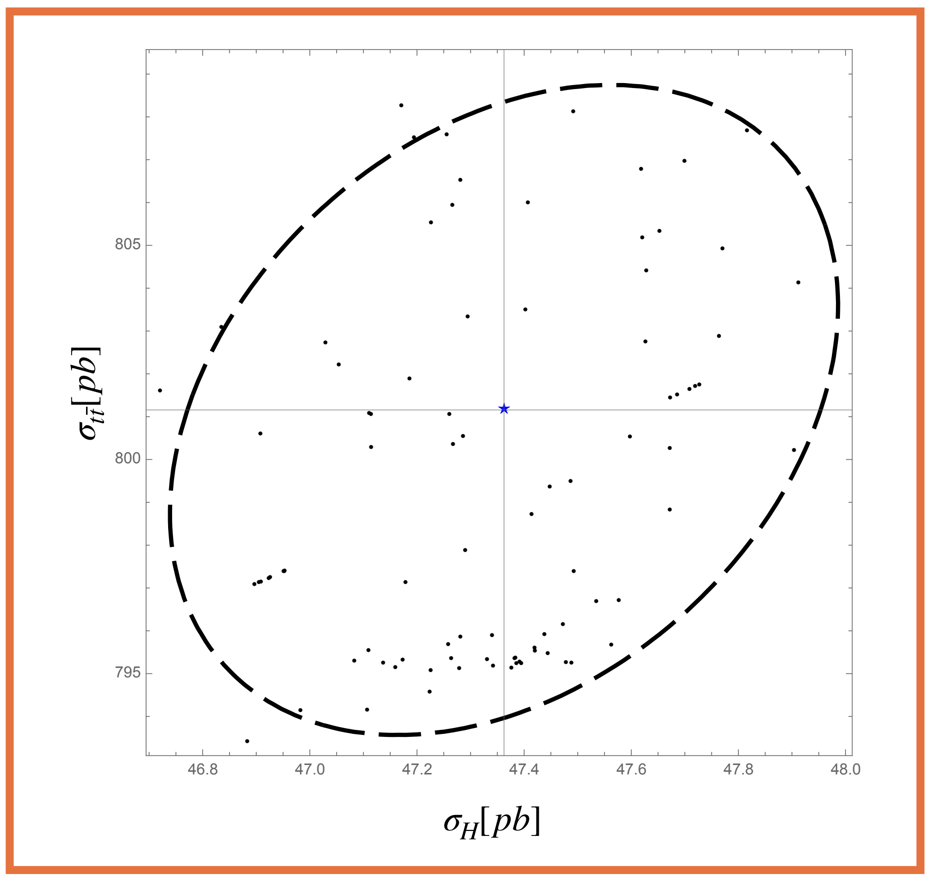


Projection of ellipsoids on 2D planes: pairs of XS values.

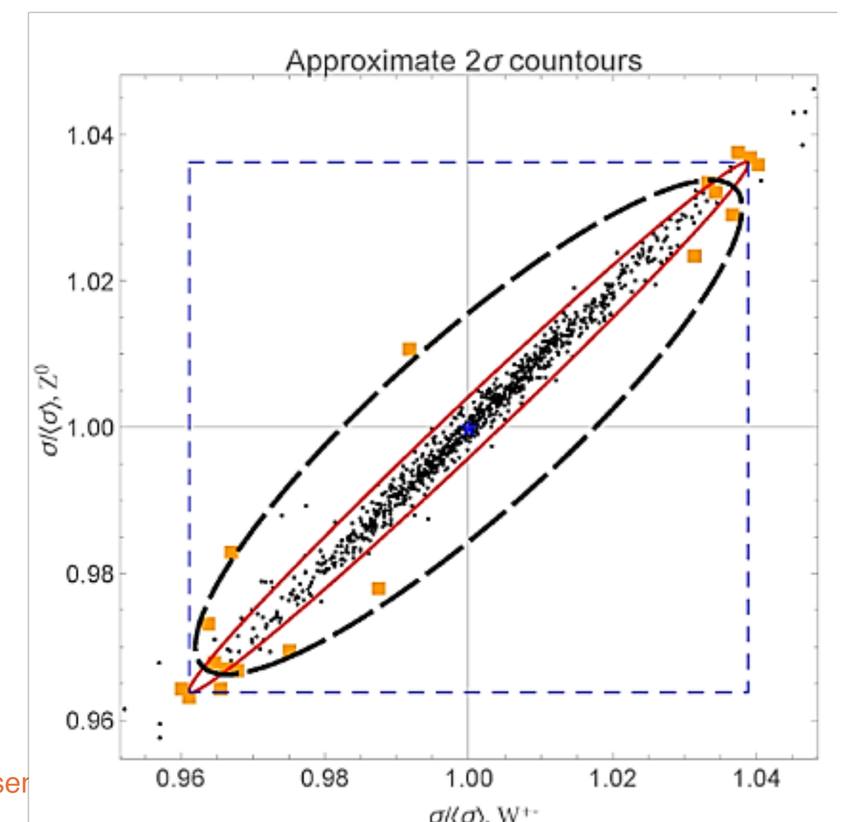
We reconstruct the ellipsoid from the scatter plots using a convex hull method. It provides an approximate region containing all found replicas with a given $\Delta\chi^2$.

Illustrated here for $\Delta\chi^2 = -60 \pm 3$. Differs from the covariance matrix calculation.

Ellipses centered in $\min(\chi^2_{Hopscotch})$.



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



Conclusions

What is a faithful uncertainty coming from PDFs on those 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** or **multiparametric functional forms.** *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.

Conclusions

What is a faithful uncertainty coming from PDFs on those 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** or **multiparametric functional forms**. *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 deviation may limit reduction of the PDF uncertainties and may explain some differences between the uncertainties of 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.

Conclusions

What is a faithful uncertainty coming from PDFs on those 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** or **multiparametric functional forms.** *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 deviation may limit reduction of the PDF uncertainties and may explain some differences between the uncertainties of 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.

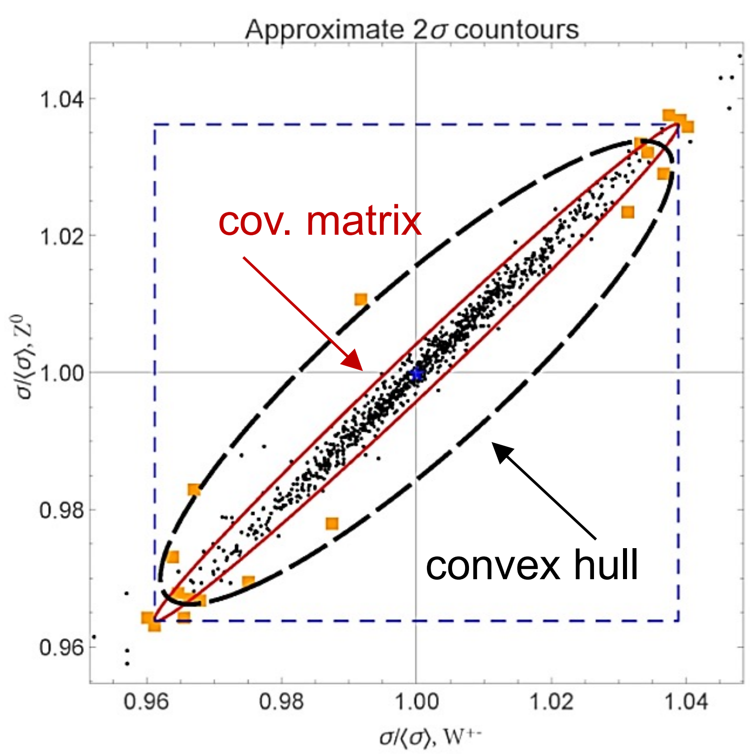
Back-up slides

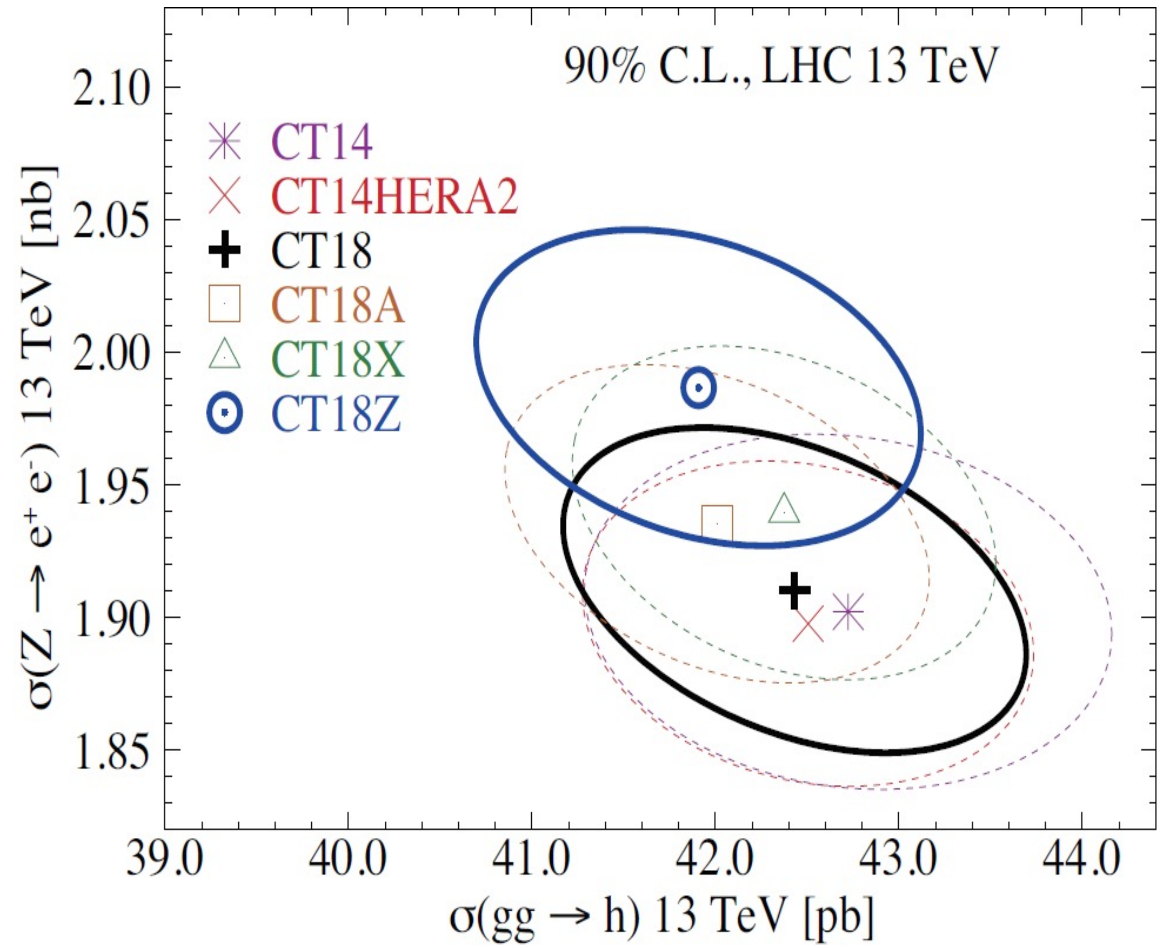
Toward **robust** PDF uncertainties

Strong dependence on the definition of corr. syst. errors would 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]

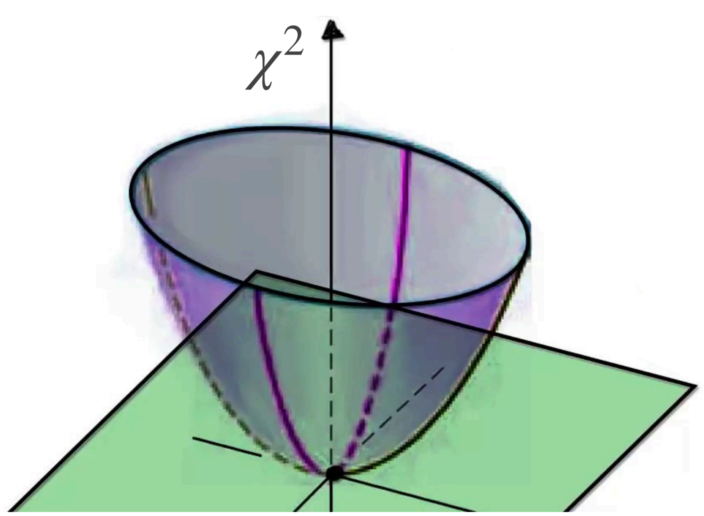




The CT18/CT18Z uncertainties aim to be **robust**: they largely cover the spread of central predictions obtained with different selections of experiments and assumptions about systematic uncertainties

Uncertainty on QCD observables — the hopscotch

Hessian methods are based on the paraboloid behavior of the χ^2 function — PDF eigenvector set naturally renders the coordinates giving the largest contribution to a determined value $\hat{\mu}$, with the Principal Component Analysis or a related method.



Uncertainty on QCD observables — the hopscotch

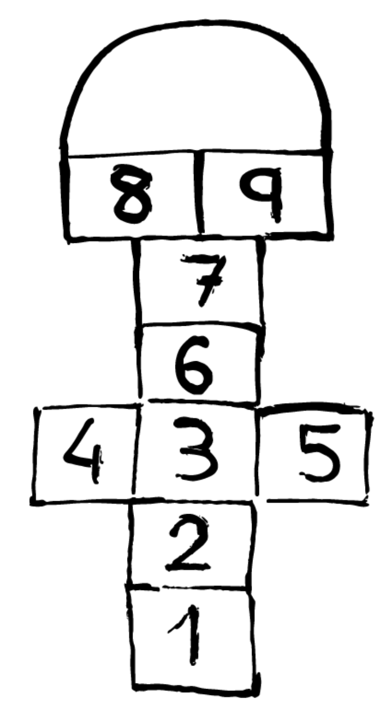
Sampling of multidimensional spaces ($d \gg 20$) can be exponentially inefficient and require $n > 2^d$ replicas to obtain a convergent expectation value.

[Hickernell, MCQMC 2016, 1702.01487]
[Sloan,I.H.,Wo´zniakowski, 1997]

Specific QCD observables: only few effective large dimensions contribute the bulk of the uncertainty.
E.g. compressing MC PDFs into a Hessian set: we construct a basis to identify such large dimensions.

Hopscotch scans:

estimation of a representative uncertainty on a cross section.



The release of a public code for NNPDF4.0's new methodology provide a perfect playground to explore the role of sampling.

[NNPDF, EPJC 81]

The 100 replica set (red points) of NNPDF4.0 gives rise to the red ellipse.
→ those points do not correspond to a same $\Delta\chi^2$ value, not comparable to the triangles