Theoretical and sampling uncertainties in global PDF fits

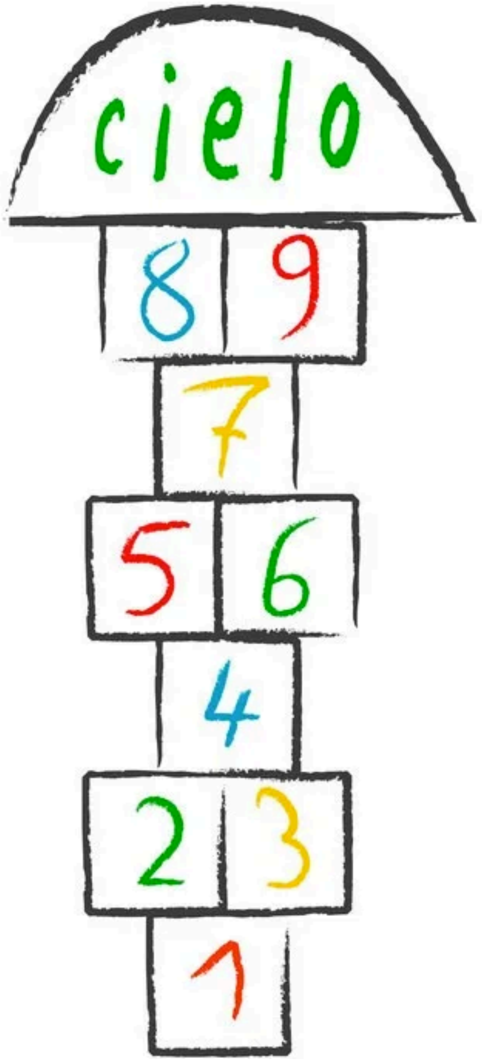
Aurore Courtoy *for the CT collaboration*

Instituto de Física
National Autonomous University of Mexico (UNAM)

CTEQ-TEA members

China: S. Dulat, J. Gao, T.-J. Hou, I. Sitiwaldi, M. Yan, and collaborators
Mexico: A. Courtoy
USA: T.J. Hobbs, M. Guzzi, J. Huston, P. Nadolsky, C. Schmidt, D. Stump, K. Xie, C.-P. Yuan

REF 2022 — online

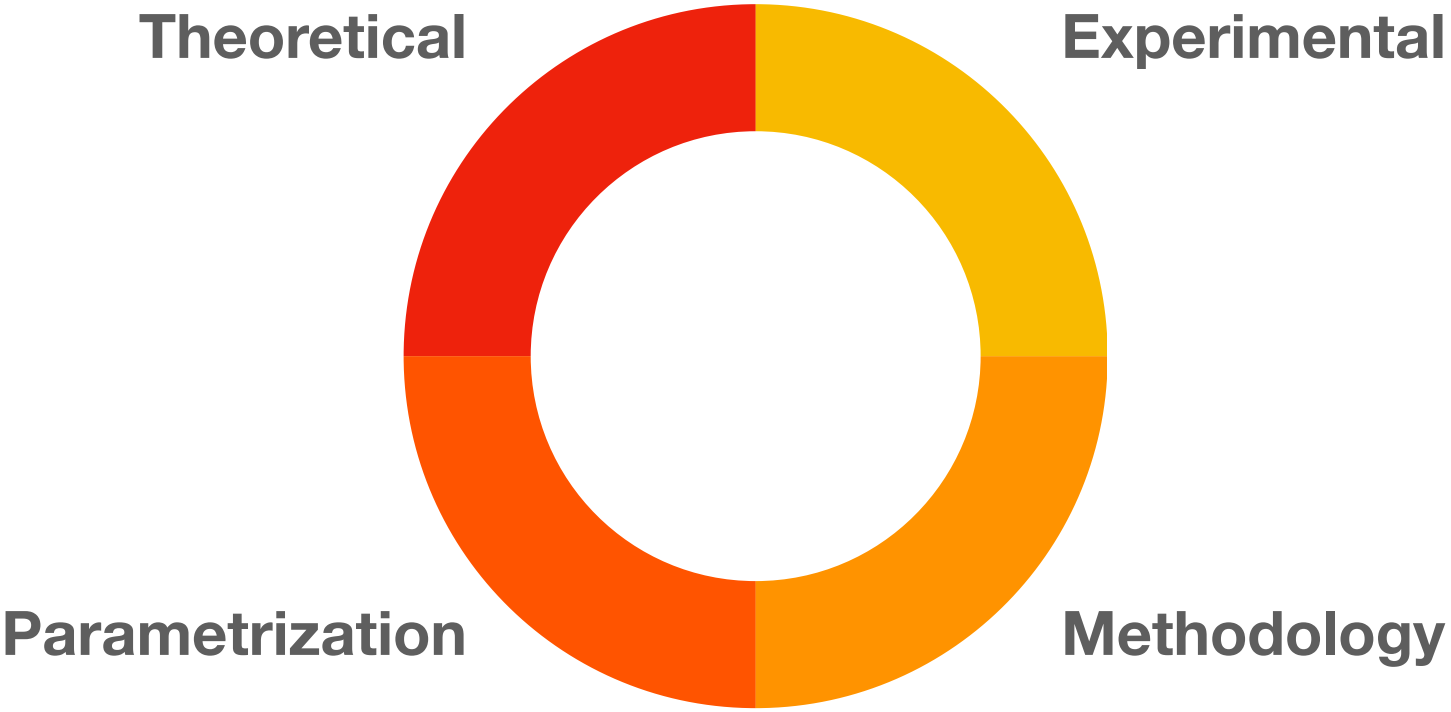


Instituto de Física UNAM

CONACYT
Consejo Nacional de Ciencia y Tecnología

dgapa
Dirección General de Asuntos del Personal Académico

Contributions to PDF uncertainties



Theoretical

Experimental

Parametrization

Methodology

In all four categories of uncertainties, we can further distinguish

PDF fitting accuracy and *PDF sampling accuracy.*

Accuracy in inputs —commonly integrated in global analyses.

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

A new avenue to understand PDF tolerance.

In this talk, we will discuss both — particular emphasis on sampling, though.

CT18 analysis in a nutshell

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

- Identify and include LHC data set available by mid-2018 with highest sensitivity to PDFs, using fast **Hessian techniques**.
- **Benchmark** predictions for newly implemented processes
- Examine ~**350 PDF parametrization forms** — more on this in a few slides
- Examine **QCD scale dependence** in key processes
- Validate results using a **strong set of goodness-of-fit tests**
- Examine agreement between experiments using diverse **statistical techniques**

CT18 analysis in a nutshell

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

- Identify and include LHC data set available by mid-2018 with highest sensitivity to PDFs, using fast **Hessian techniques**.
- **Benchmark** predictions for newly implemented processes
- Examine ~**350 PDF parametrization forms** — more on this in a few slides
- Examine **QCD scale dependence** in key processes
- Validate results using a **strong set of goodness-of-fit tests**
- Examine agreement between experiments using diverse **statistical techniques**

Four sets proposed:

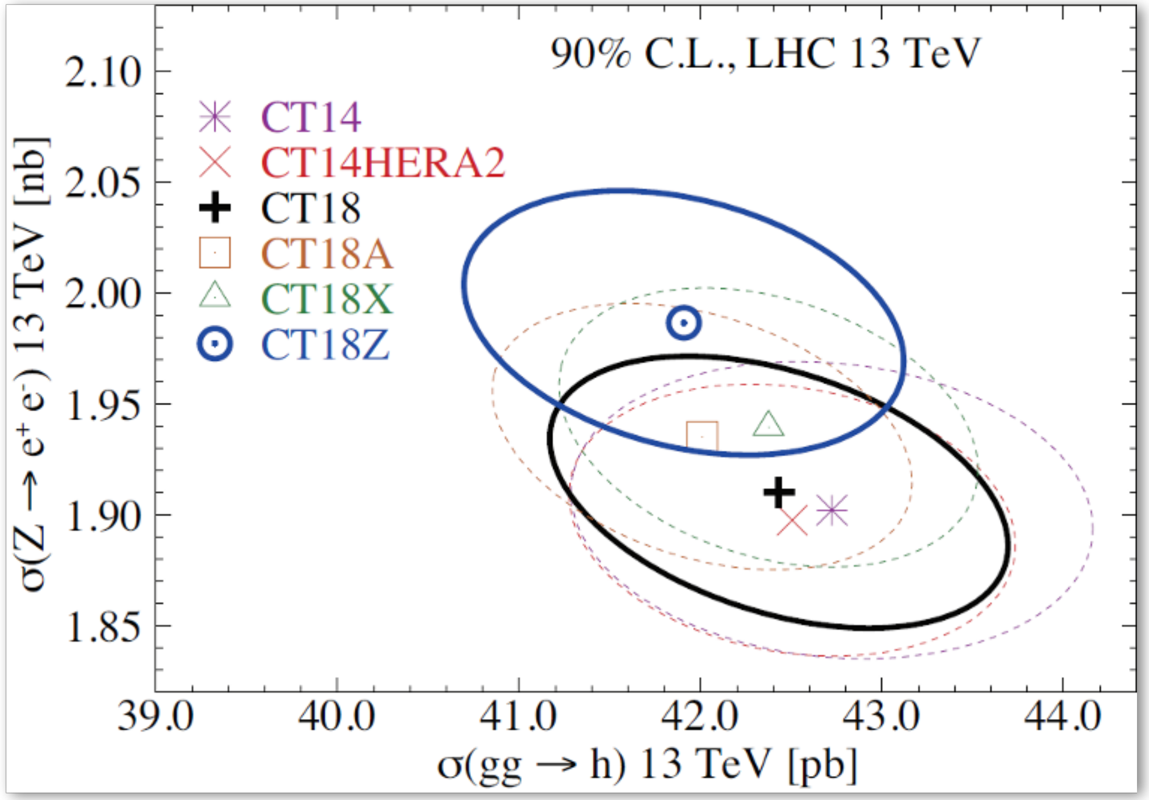
CT18 (nominal)

CT18A (include ATLAS 7TeV),

CT18X (DIS scale variation $\mu^2_{F,DIS} = 0.8^2\left(Q^2 + \frac{0.3GeV^2}{x^{0.3}}\right)$),

CT18Z (ATLAS 7TeV+scale variation)

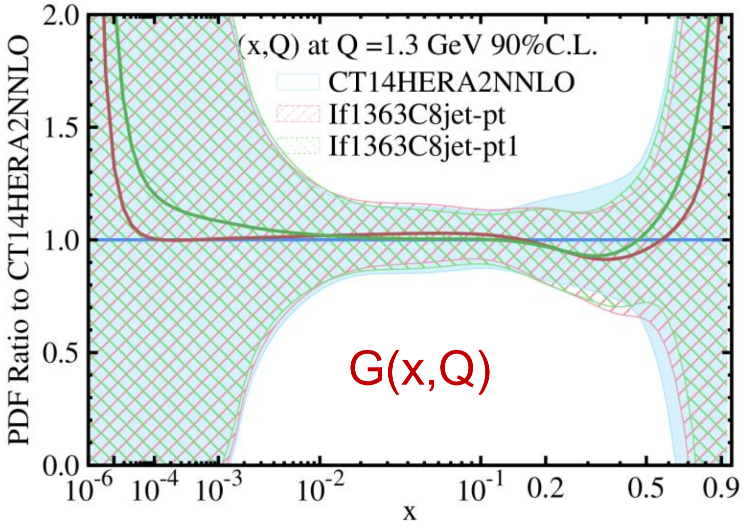
CT18 and *CT18Z* span the most different hypotheses, and the combination of the two represents the most complete uncertainty.



Theoretical uncertainties in CT18

Theory predictions and choice of scale

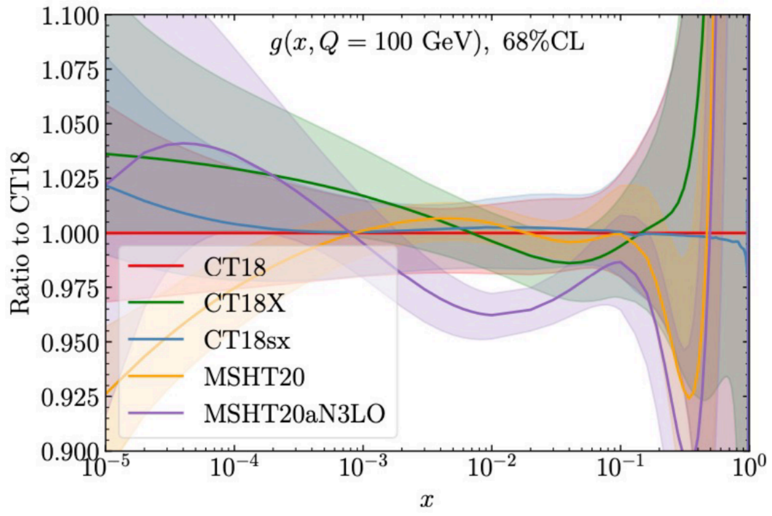
Choice of scale for inclusive jet data leads to a different gluon PDF yet contained in the CT uncertainty.
Resilience in global fit reflected through the tolerance.



Scale dependence and small-x resummation — K. Xie (in progress)

NNPDF and xFitter adopts BFKL to resum small-x logs. CT adopt a saturation DIS scale and obtain similar quality of description of data.

Small-x resummation enhances gluon PDF, similarly to N3LO (MSHT, see T. Cridge’s talk)

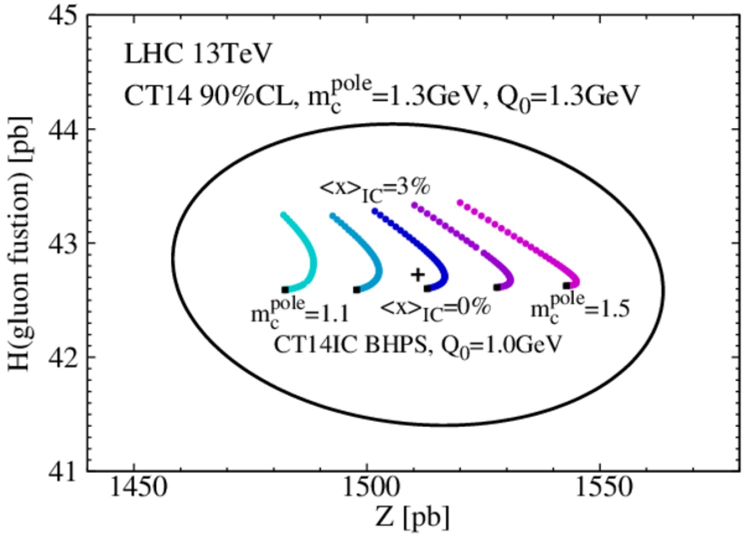


Dependence on m_c —CT14 Intrinsic Charm

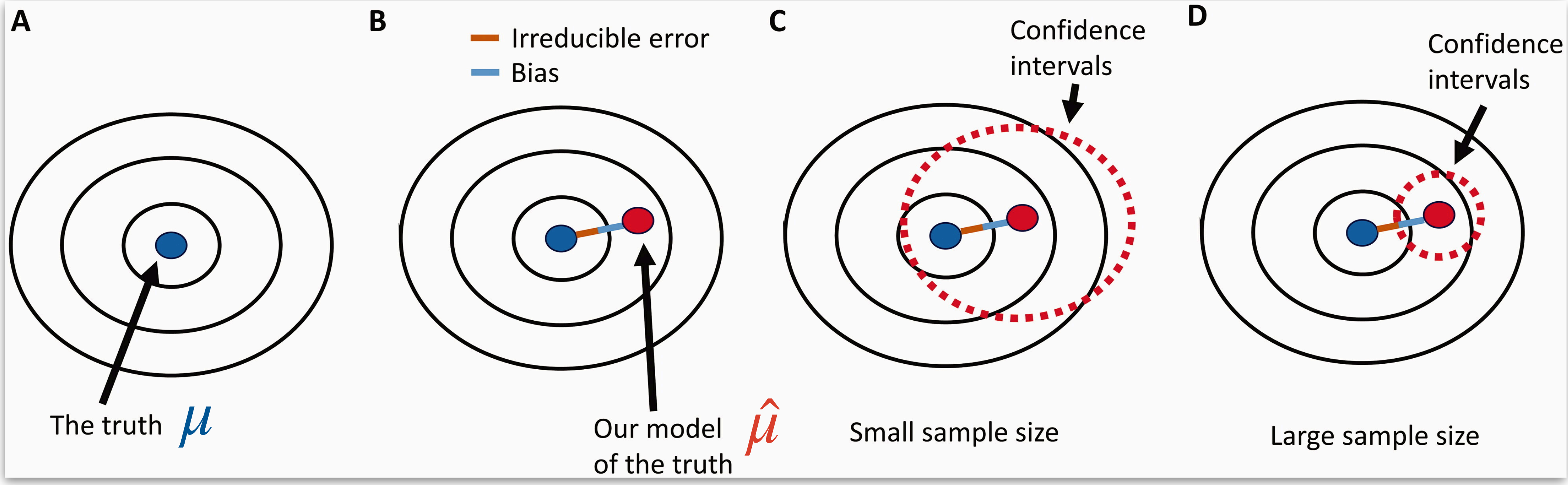
Study of dependence on the charm pole mass:

CT14 Intrinsic Charm analysis [Hou et al., arXiv:1707.00657]

CT18 Fitted Charm analysis (very soon)



From small to big data sets — sampling uncertainties



With an increasing size of sample $n \to \infty$, under a set of hypotheses, it is usually expected that the *deviation* on an observable decreases like $\left(\sqrt{n}\right)^{-1}$. *That's the law of large numbers.*

What uncertainties keep us from including *the truth,* μ?

The law of large numbers obviates the *quality of the sampling,* ▬ Irreducible error ▬ Bias .

Illustration from:
Pavlos Msaouel (2022)
The Big Data Paradox in Clinical Practice
Cancer Investigation, 40:7, 567-576

Trio identity

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

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

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

depends on the sampling algorithm

🟧 Irreducible error
🟦 Bias

≡ statistical model, quality of data,…

can tend to $\sigma/\sqrt{n}$ 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

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

Hickernell
MCQMC 2016
1702.01487

Sampling bias

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.

Sampling bias

The sample deviation can be large if the sampling is not sufficiently random.

Standard error estimates can be misleadingly small.

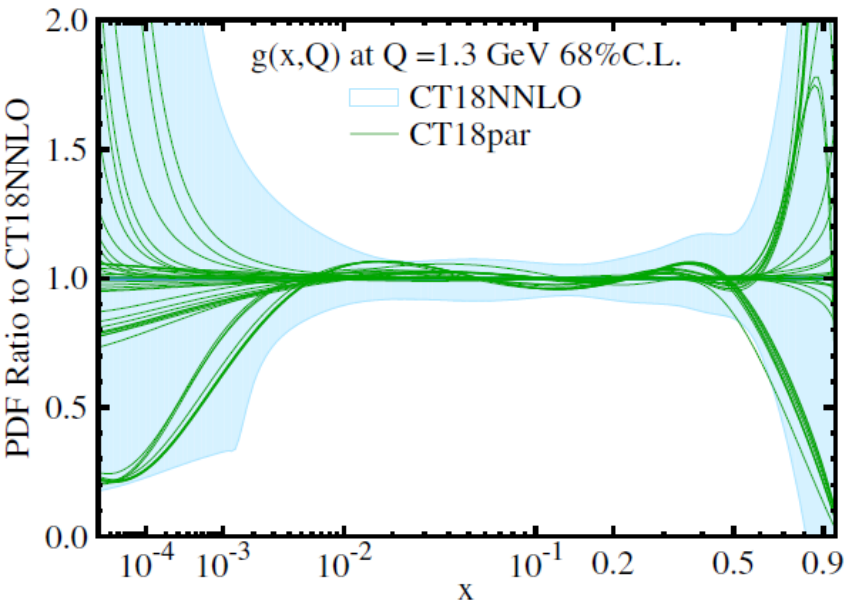
➪ critical role of controlling for **sampling biases** in determination of PDFs.

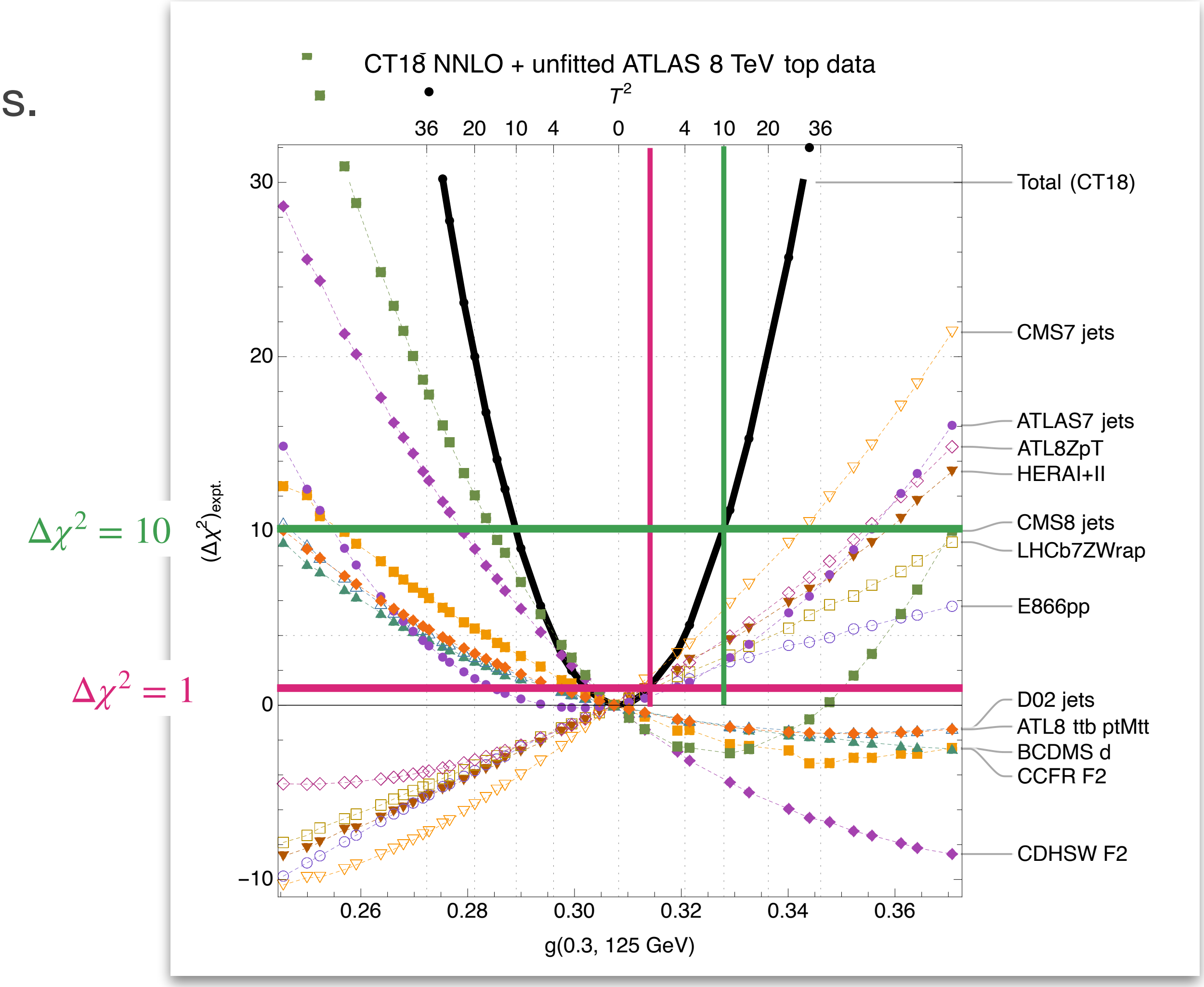
How do we know the “data+sampling defect=confounding correlation” of our analysis?

CT: tier-1 and tier-2 penalties related to tolerance criteria.

Size of uncertainties reflect a series of confounding sources.

Verification that proper spanning of parameter space is compatible with total uncertainties (*a posteriori*).



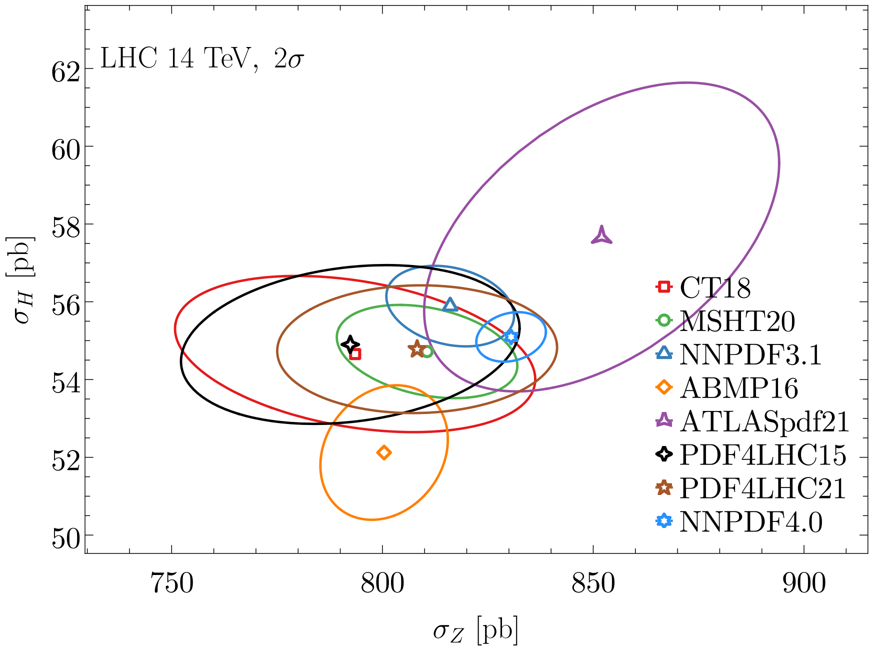


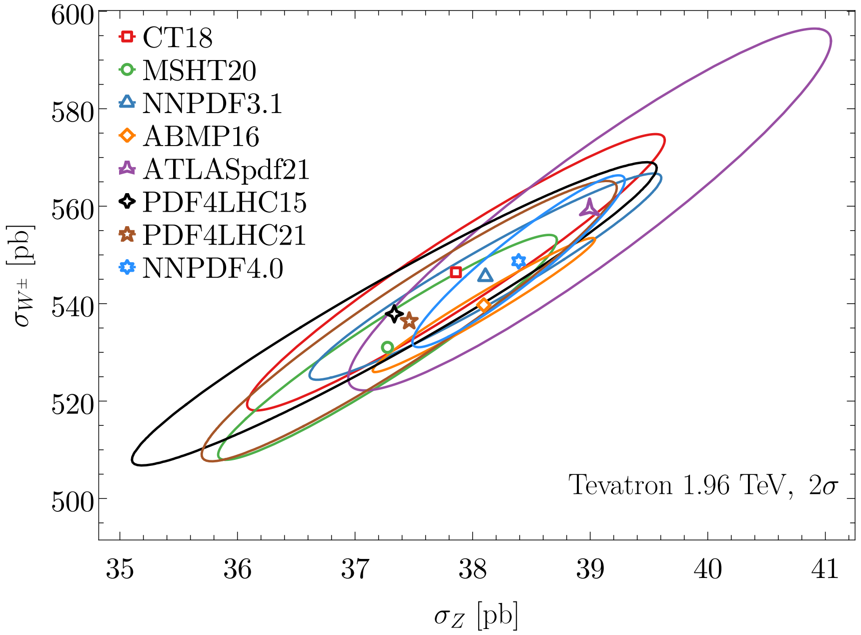
Hou et al, Phys.Rev.D 103 (2021)

Hopscotch scans

Parton distributions need a representative sampling
2205.10444

Algorithm for observable-oriented verification of representative uncertainty





To sample the PDF dependence for Monte Carlo-based global analyses:
sample primarily the coordinates with large variations of physical cross section σ.

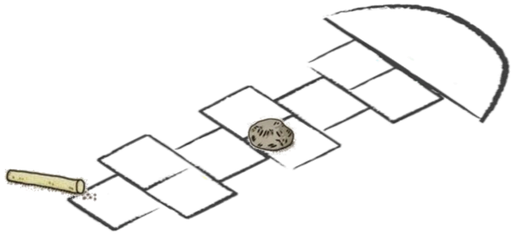
Using NNPDF4.0 public code, we then employ:

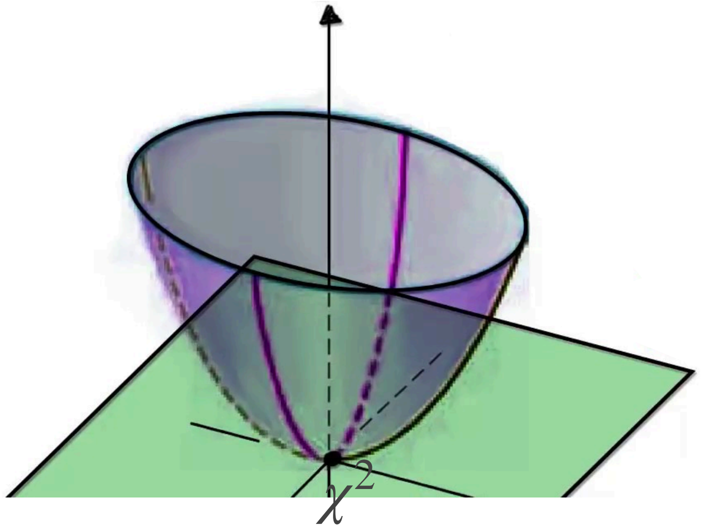
$n =$ the number of replicas/EV directions/…

1. Basis coordinates in the PDF space — Hessian representation
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,I.H.,Wo´zniakowski, 1997]

How to play hopscotch?

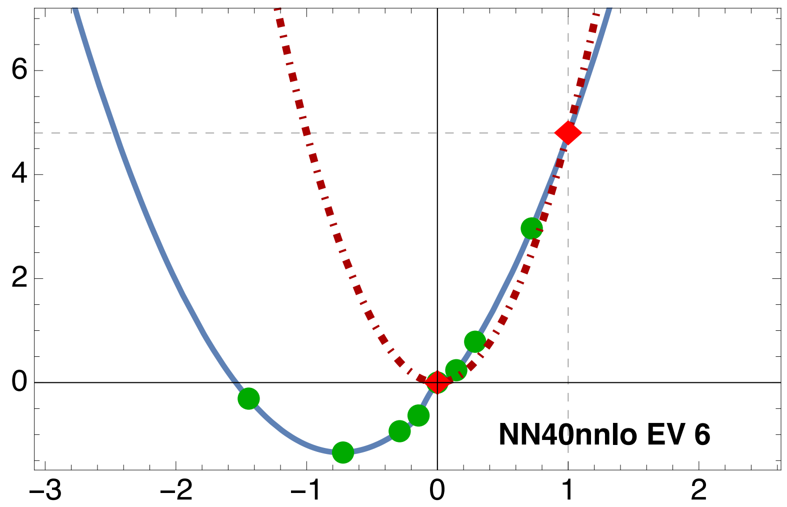




In the Hessian representation, the chi square behaves like a paraboloid of n_{param} dimensions, thus defining a global minimum.

Hessian and Monte Carlo representations of given PDF sets are shown to be compatible — convertions exist in both ways.

Hence, a chi-square paraboloid can also be defined for Monte Carlo-based analyses.

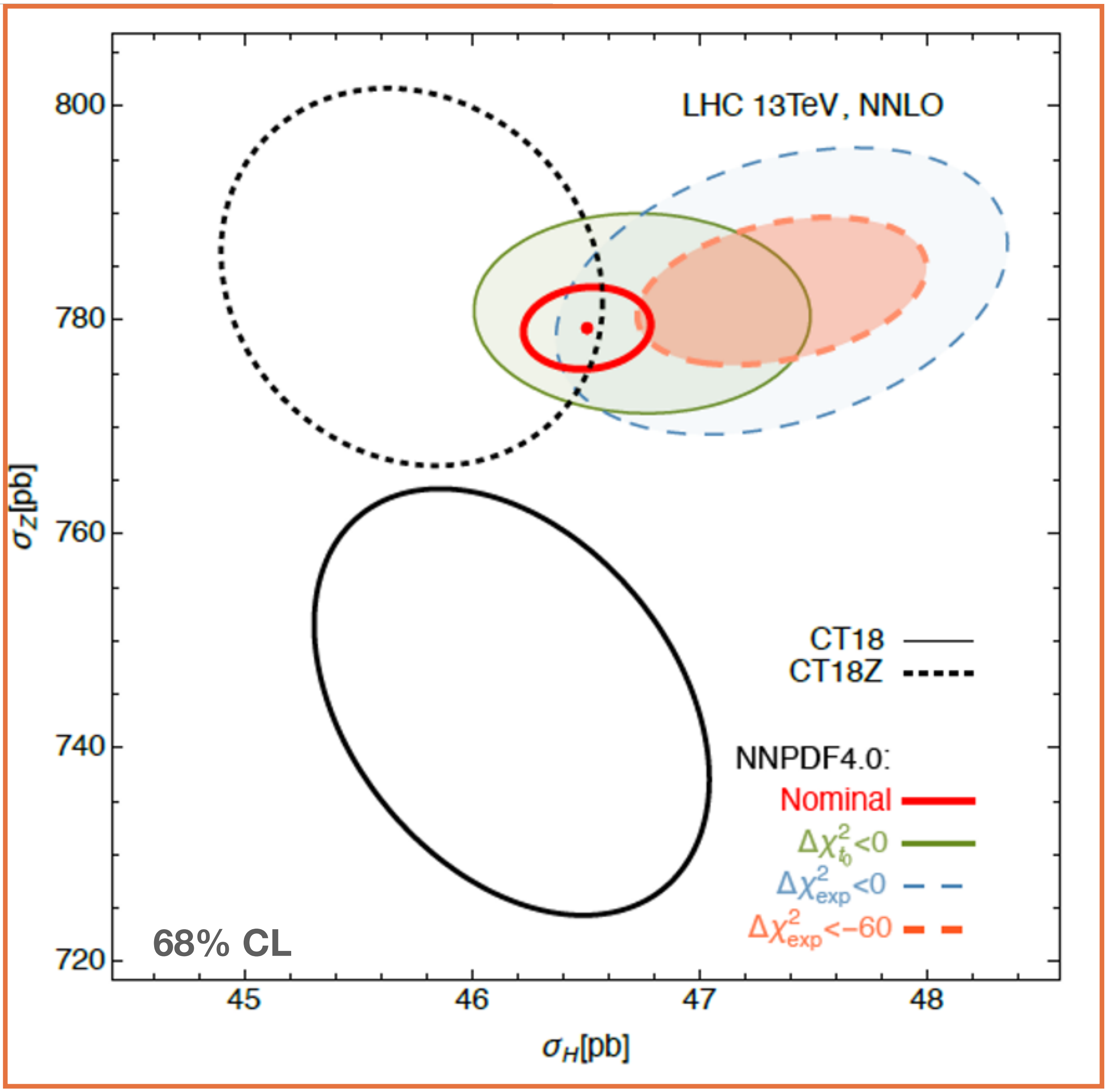


For example, here's a reconstructed eigenvector (EV) direction for the NNPDF4.0 set, in blue.

Its shape indicates a larger paraboloid than the red curve:

- we can throw the marker in (linear combinations of) the directions whose variation affect given cross sections the most
- we generate new replicas — the hopscotch replicas
- we draw the approximate regions defined by the latter for the cross sections of interest

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

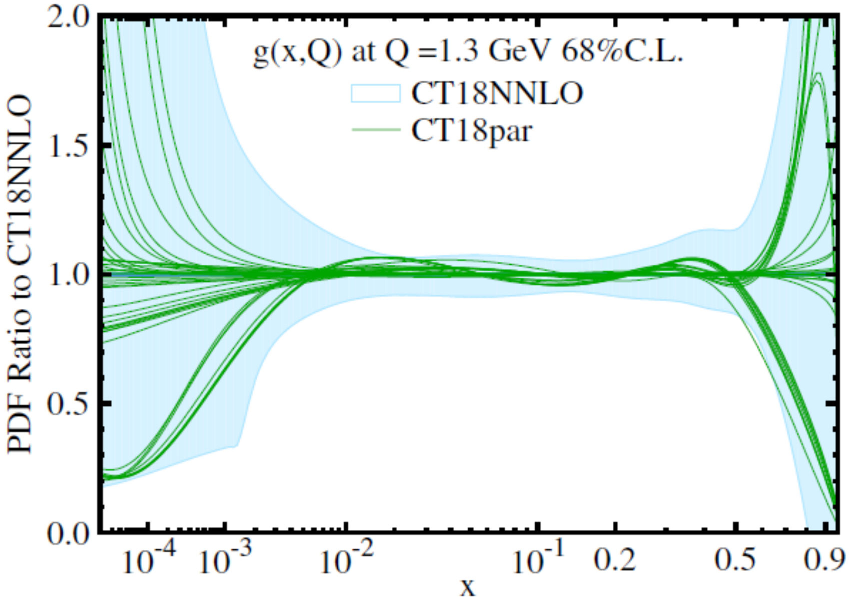


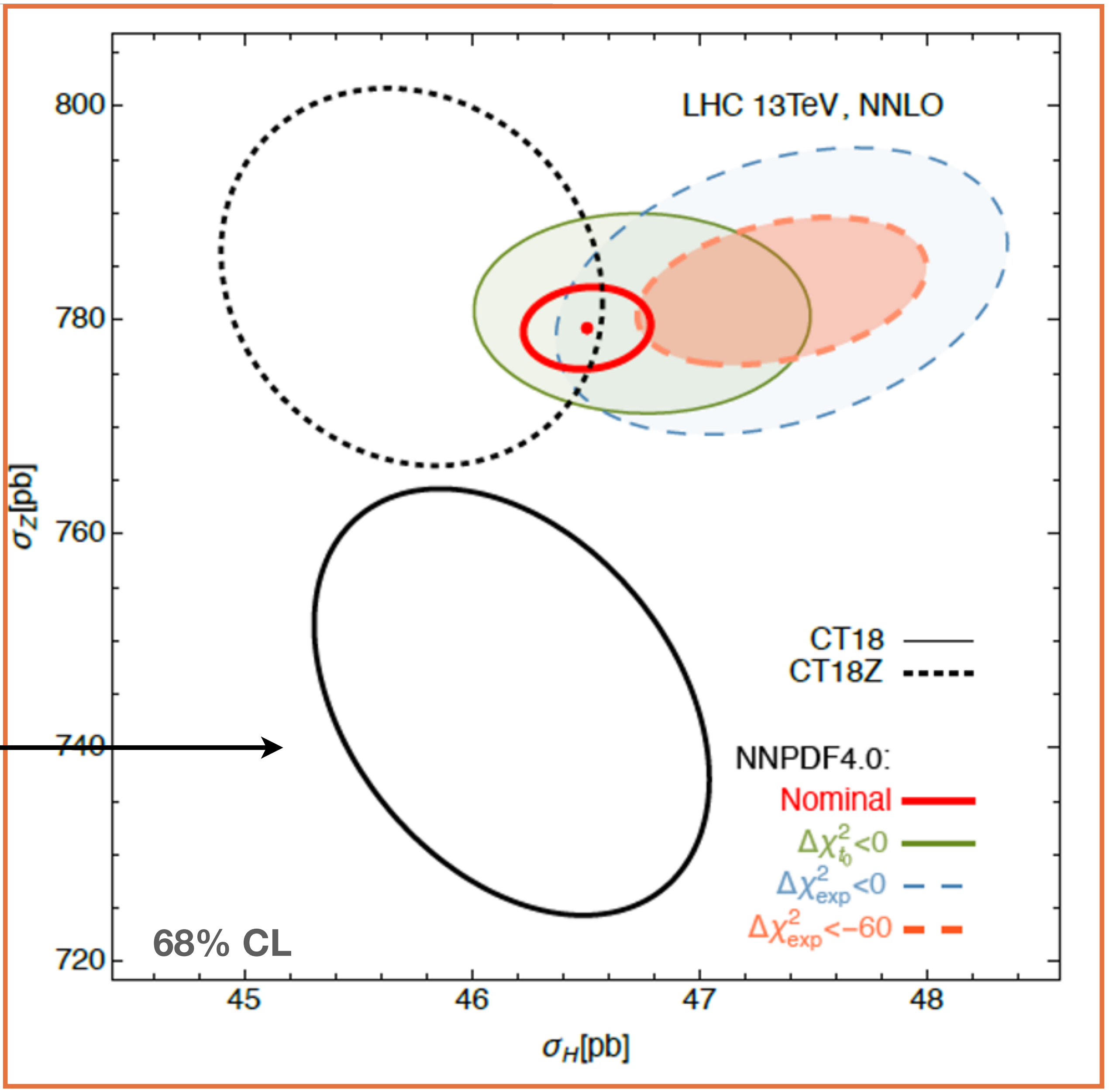
Color ellipses:

- areas of possible solutions corresponding to lower ($\Delta\chi^2 < 0$) w.r.t. the nominal solution
- found through the hopscotch scan — a dimensionality reduction method.

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

Monte Carlo uncertainties from sampling bias found through the hopscotch scans play a similar role as sampling of parameter space in Hessian uncertainties.





Color ellipses:

- areas of possible solutions corresponding to lower ($\Delta\chi^2 < 0$) w.r.t. the nominal solution
- found through the hopscotch scan — a dimensionality reduction method.

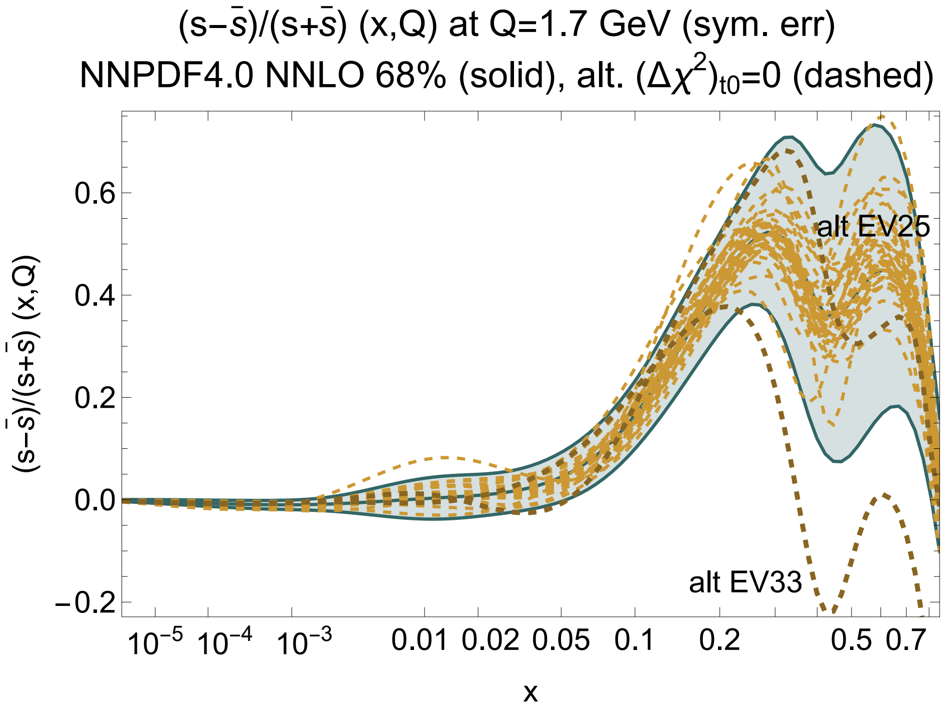
Monte Carlo and Hessian representation — role of constraints

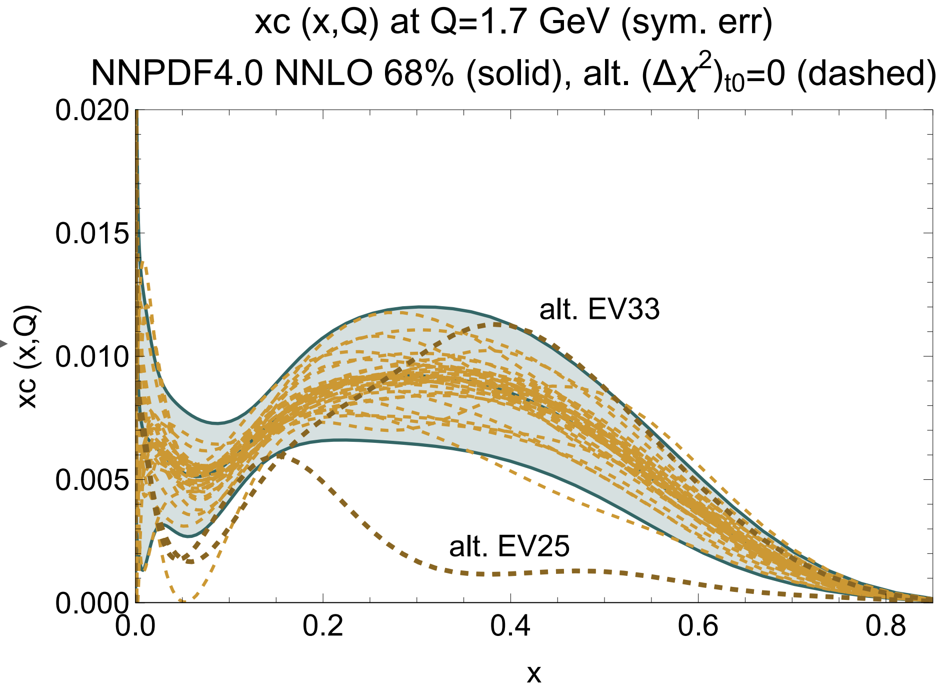
Role of constraints in global analyses: can act as *priors* to the final distributions.

Choice for positivity, integrability, large/small-x behavior, … will affect PDF sets in the interpolation region.

Hopscotch replicas pass all CT criteria:
need for a benchmark on constraints?

Hopscotch uncertainties wash out evidence for large positive strangeness asymmetry and non-zero intrinsic charm.





The understanding of theoretical constraints in MC vs. Hessian is very relevant to polarized PDFs, TMDs, etc.

Conclusions

The CT18 analysis includes various sources of theoretical uncertainties, displayed through various sets of PDFs. Further ongoing studies focus on understanding the interplay between theoretical, parametrization and methodological uncertainties.

Highlights on the sampling uncertainties:

1. A PDF fit with few parameters and $\Delta\chi^2 = 1$ tolerance probably underestimates the parametric uncertainty.

2. Difficult to sample the full parameter space with many parameters without biases. Analytic minimization like in CT18 and MSHT20 finds the global minimum and EV directions by construction. Validating the final PDFs may be easier than understanding the respective fitting algorithm.

3. A hopscotch scan intelligently reduces dimensionality of the relevant PDF parameter space. Can be performed using public codes (*LHAPDF + mcgen + xFitter/NNPDF fitting codes*) to verify the PDF uncertainty for a specific QCD cross section or observable.

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

Impact on the uncertainties at small and large x, PDF ratios, correlations, strangeness asymmetry, fitted charm, …

Insights applicable to other analyses using a large parameter space — CT/MSHT tolerance, polarized PDFs, etc.

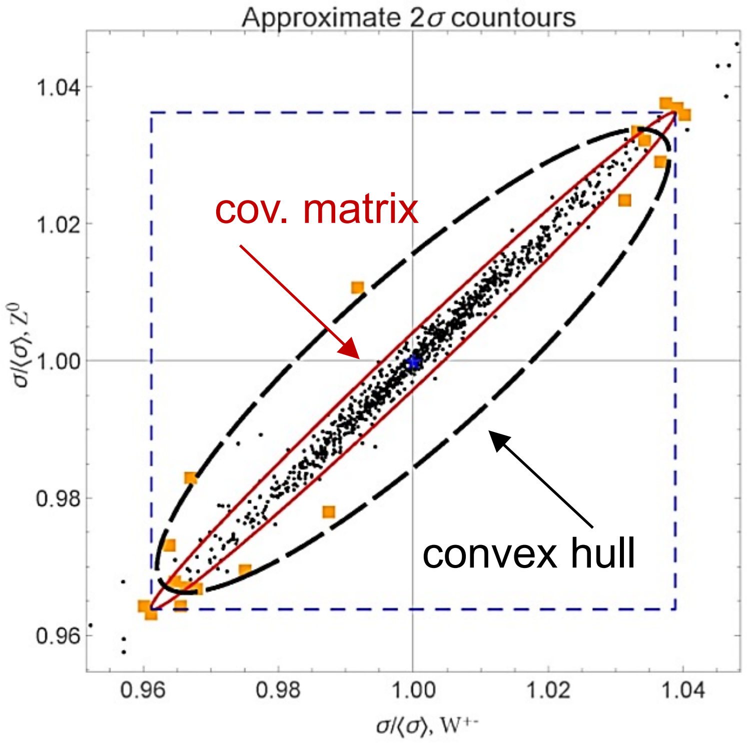
Back-up slides

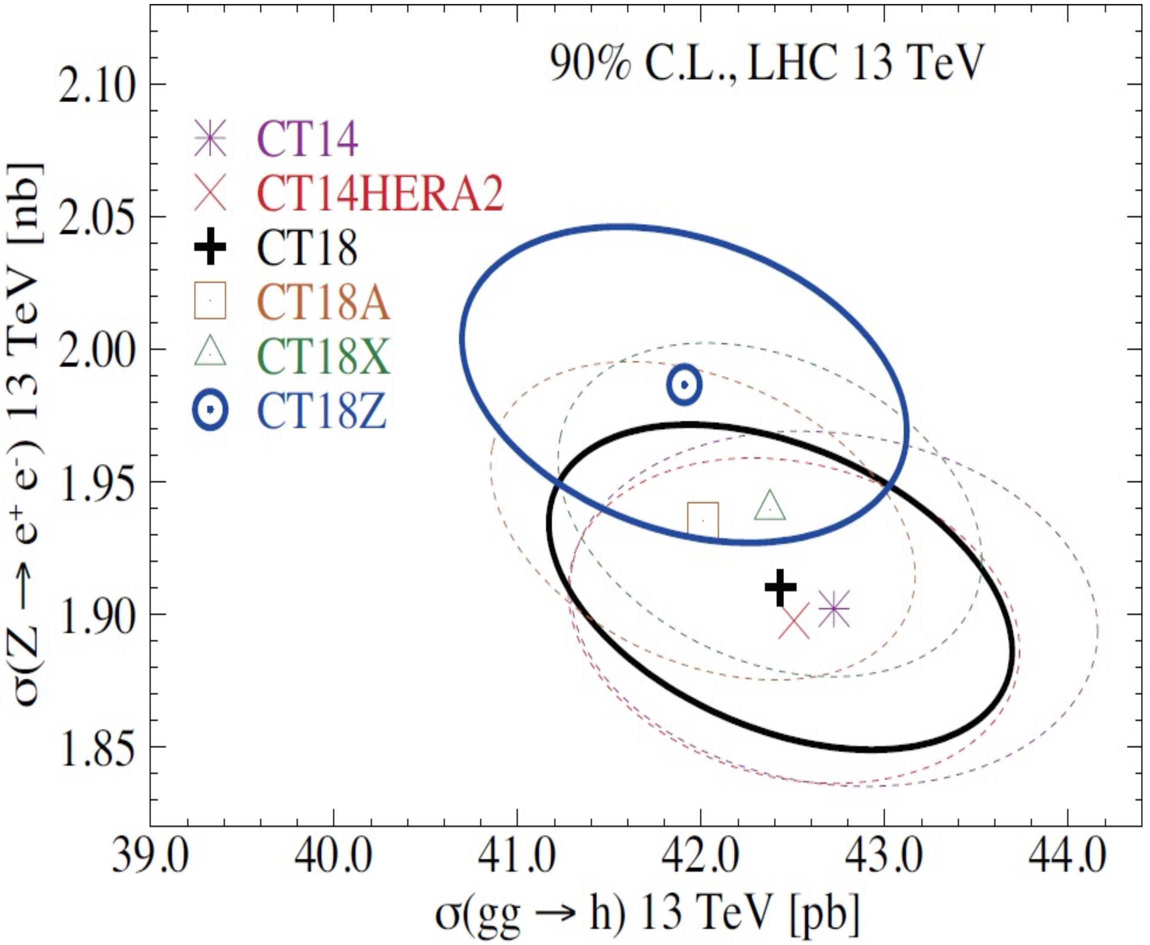
Toward **robust** PDF uncertainties

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

Setting for NNPDF4.0 code

The evaluation of χ^2 for NNPDF4.0 nnlo replicas is done by the public NNPDF code [NNPDF, EPJC 81], with its default setting.

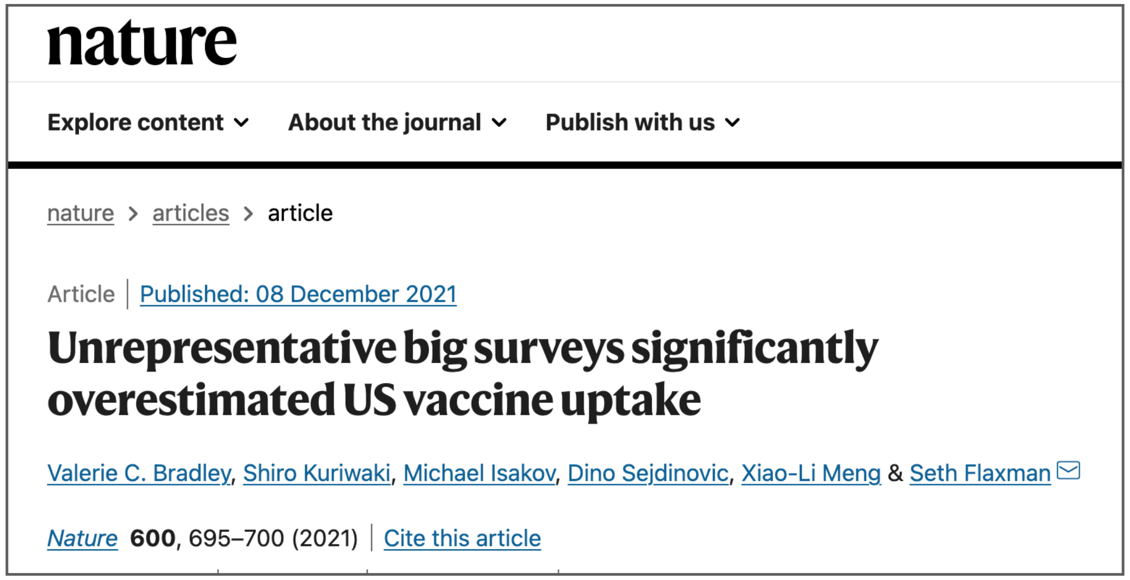
χ^2 is computed by the `perreplica_chi2_table` function of `validphys` program of the public NNPDF code.

The kinematics cuts for the correlated uncertainties are fixed as the same of the NNPDF4.0 global analysis.

The minimum value of Q^2 and W^2 for DIS measurements are hence chosen to be 3.49 GeV and 12.5 GeV respectively.

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.



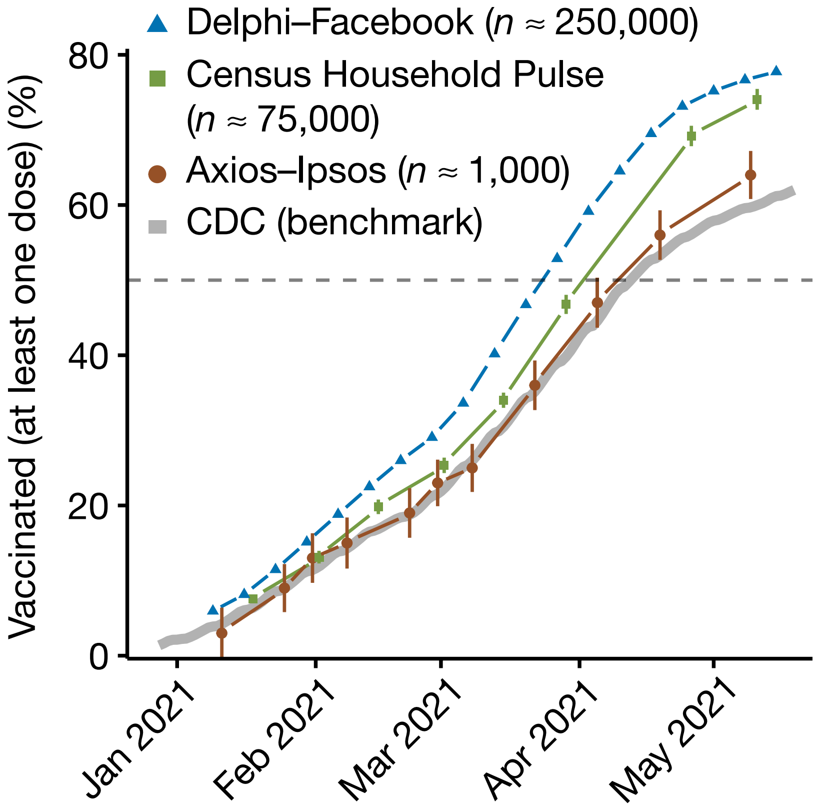
nature > articles > article

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

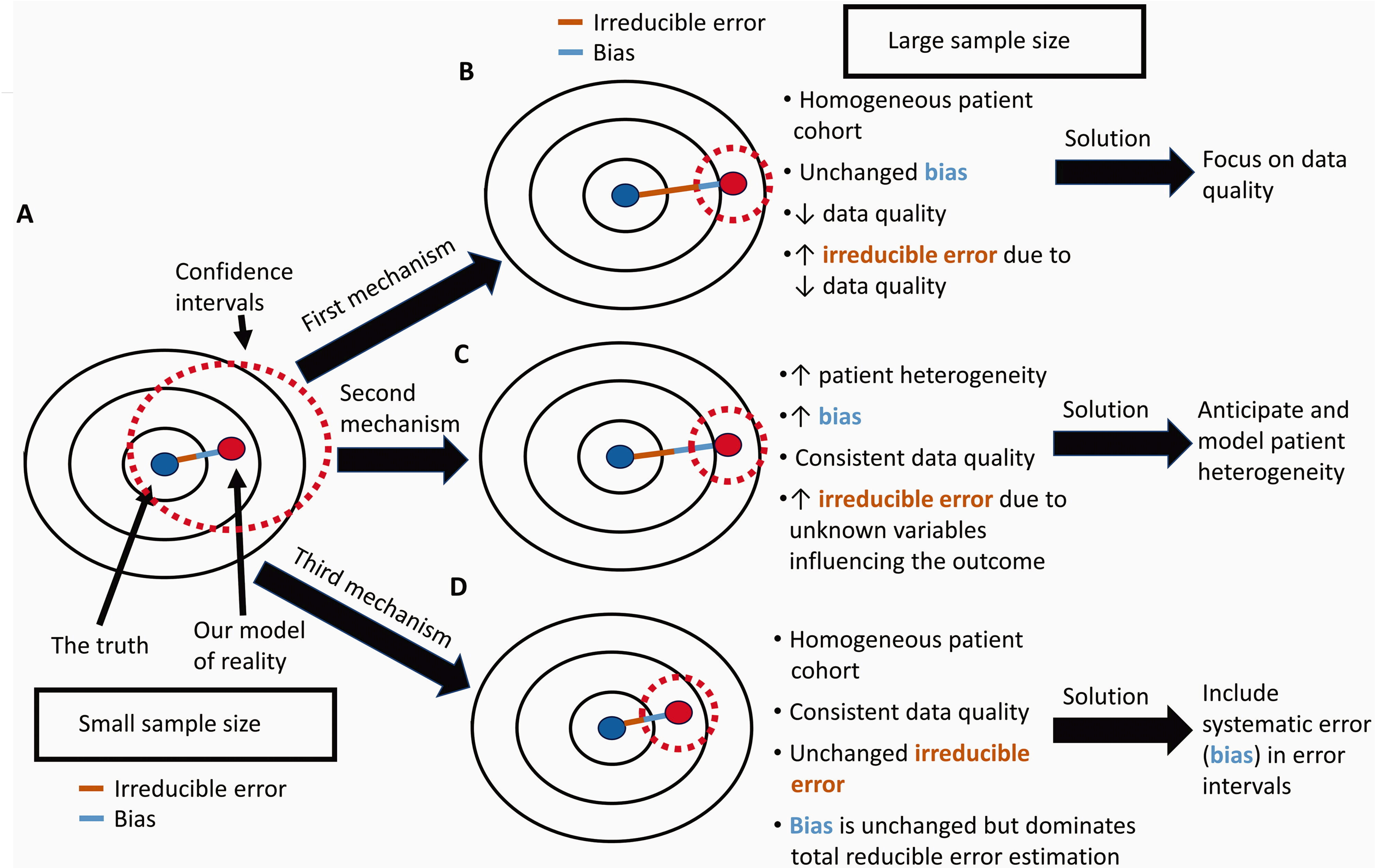
Nature **600**, 695–700 (2021) | Cite this article



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.



Large sample size

Small sample size

- Irreducible error
- Bias

B

- Homogeneous patient cohort
- Unchanged bias
- ↓ data quality
- ↑ **irreducible error** due to ↓ data quality

Solution → Focus on data quality

C

- ↑ patient heterogeneity
- ↑ **bias**
- Consistent data quality
- ↑ **irreducible error** due to unknown variables influencing the outcome

Solution → Anticipate and model patient heterogeneity

D

- Homogeneous patient cohort
- Consistent data quality
- Unchanged **irreducible error**
- **Bias** is unchanged but dominates total reducible error estimation

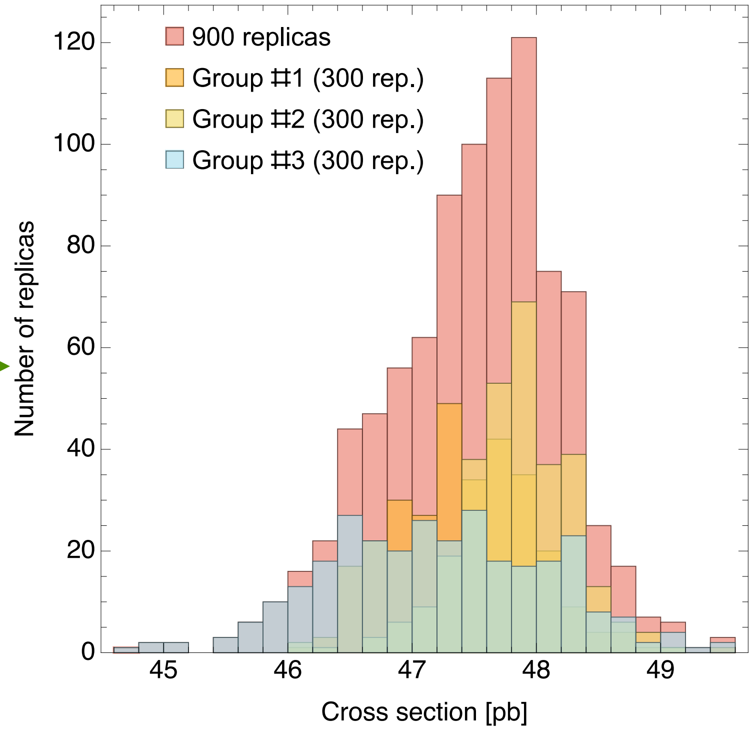
Solution → Include systematic error (**bias**) in error intervals

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

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

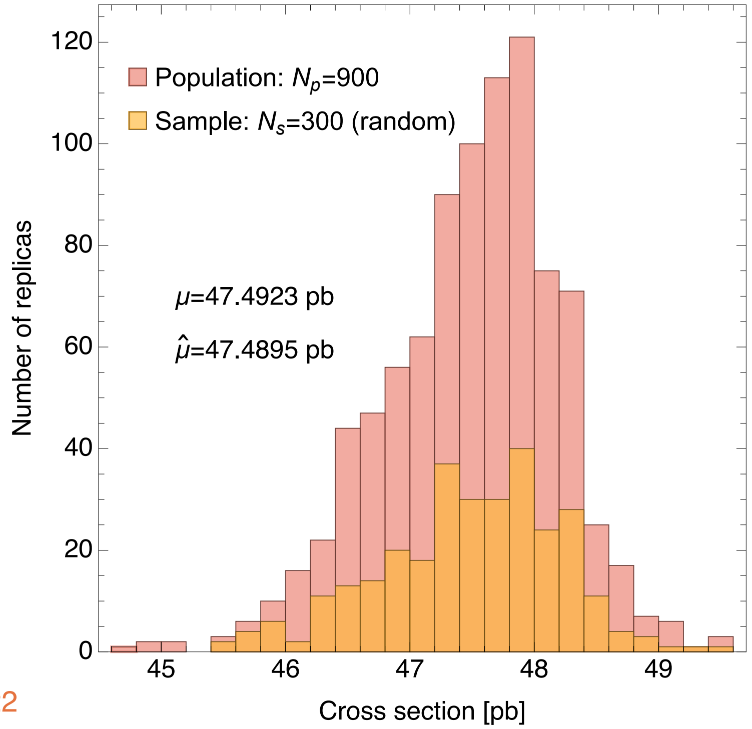
with σ the standard deviation, μ the true and $\hat{\mu}$ the determined values. *That's the law of large numbers.*



A toy sampling excercise

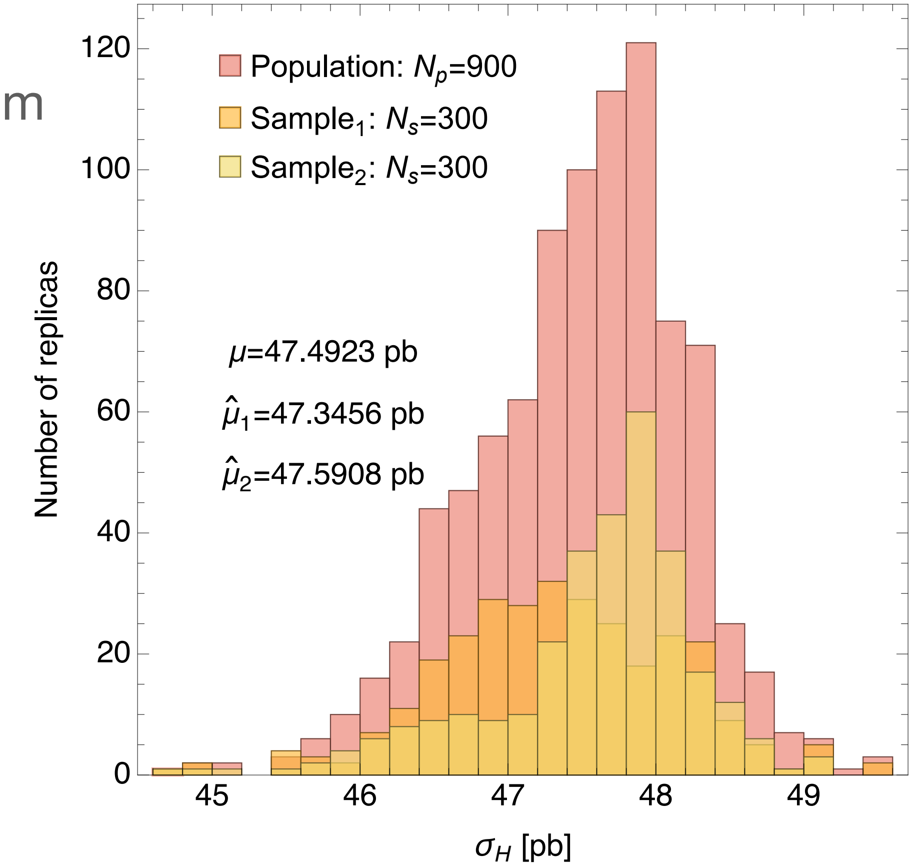
We take 300×3 groups of Higgs cross sections evaluated by 3 different groups.

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*.

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

$$\boxed{\mu - \hat{\mu} = \text{(data+sampling defect)} \times \text{(measure discrepancy)} \times \text{(inherent problem difficulty)}}$$

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

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

Step 1

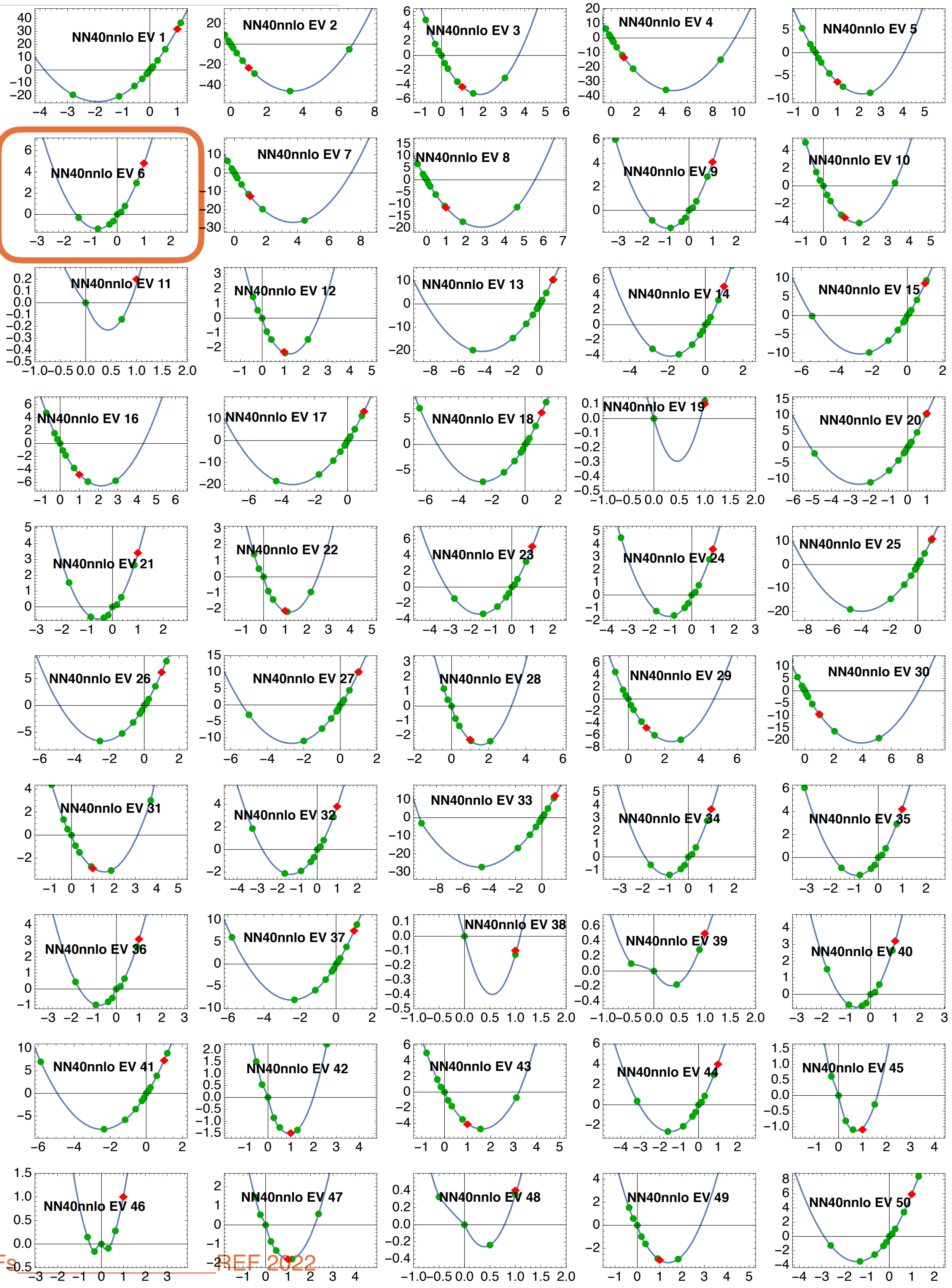
The NNPDF4.0 Hessian set ($n = 50$) defines a coordinate system on a manifold corresponding to the largest variations of the PDF uncertainty —red dots and curve.

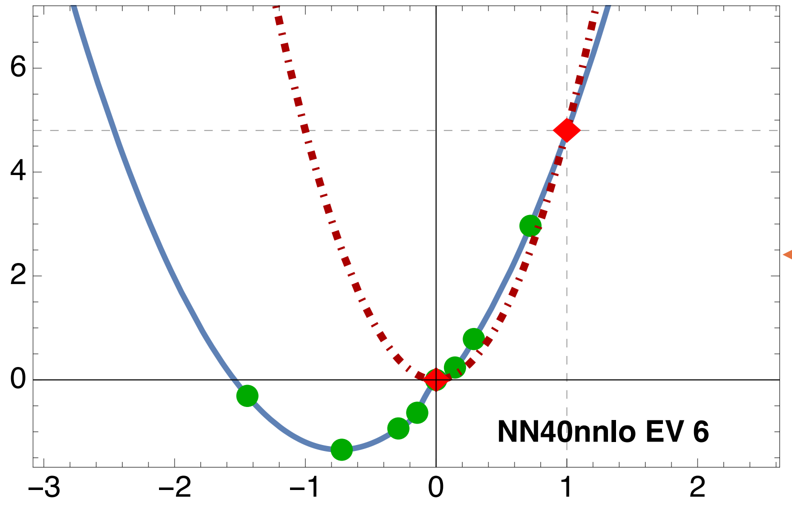
[NNPDF, 2109.02653]

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.



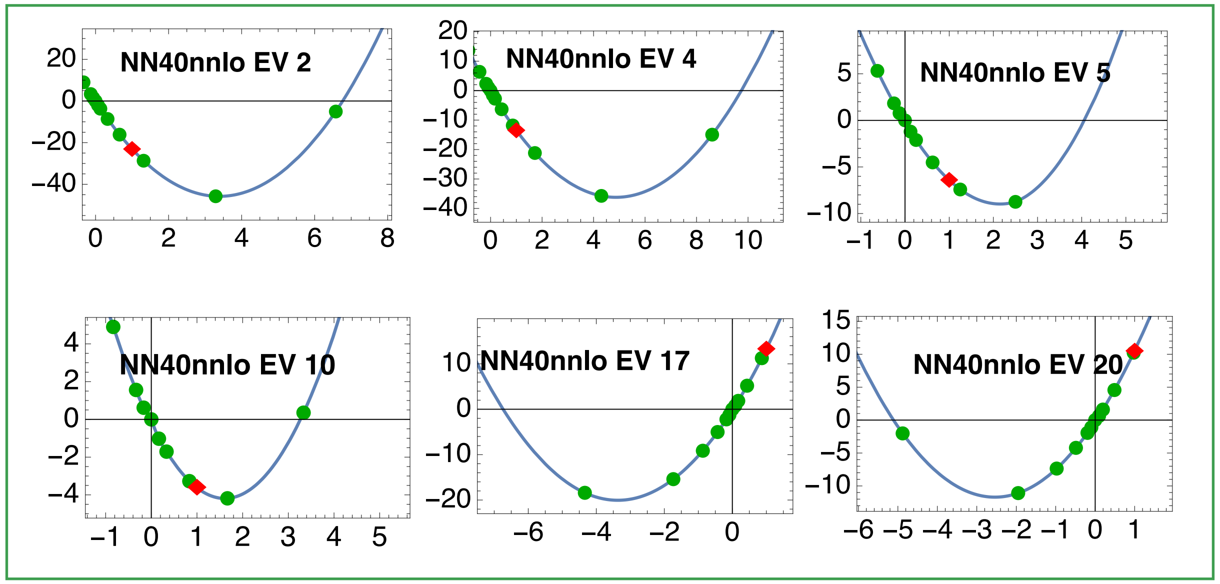


A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

Step 3

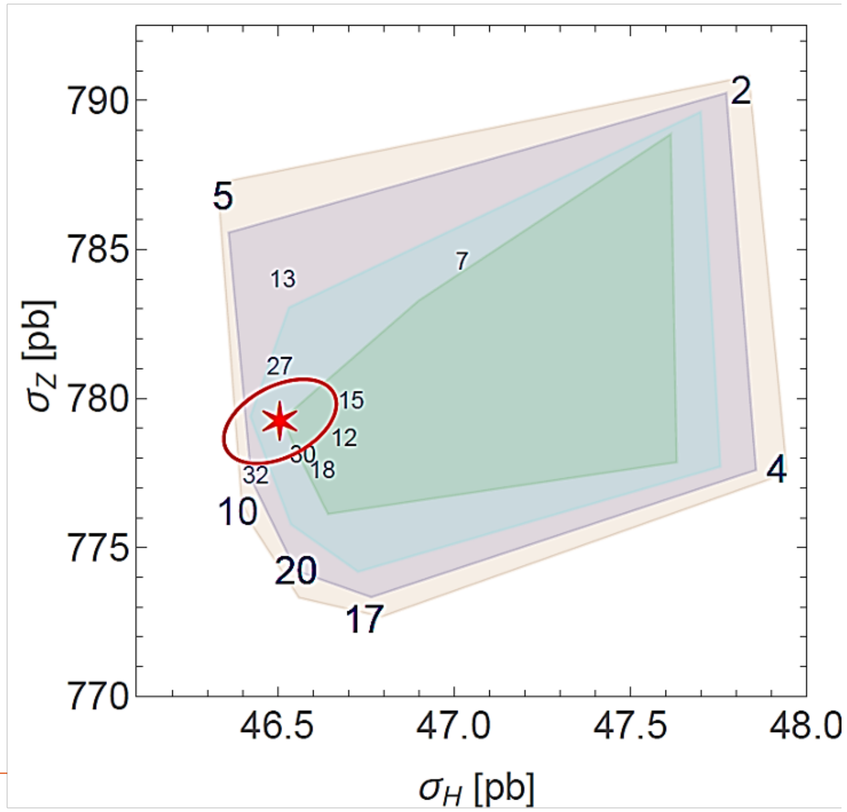
Guidance from specific cross sections:
we identify 4-7 EV directions that give the largest displacements for a given $\Delta\chi^2$ per pair.

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.

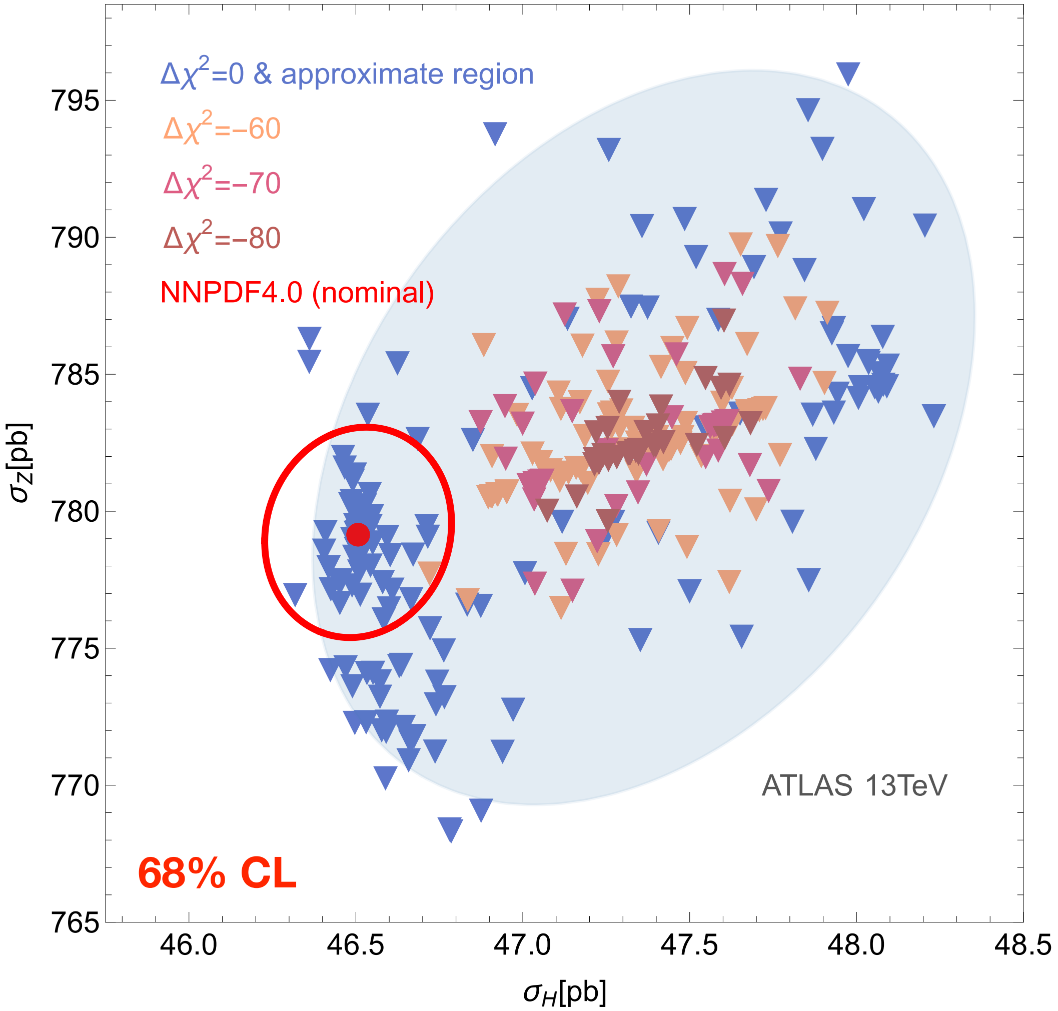
The contours are for $\Delta\chi^2 = +10, 0, -10, -20$ *w.r.t.* NNPDF4.0 replica 0 (red).



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, here for $\Delta\chi^2_{exp}$.



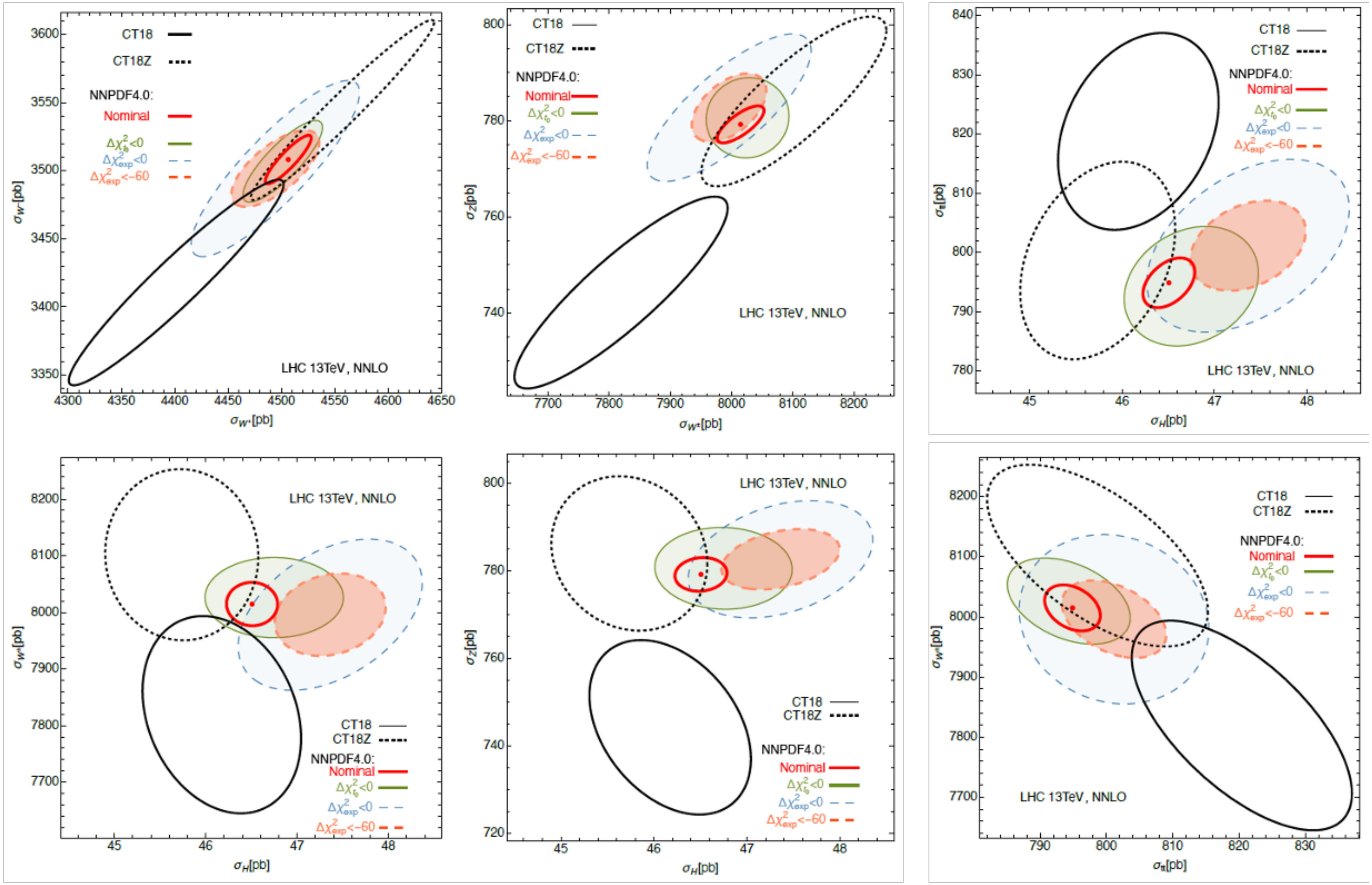
Each of the $\Delta\chi^2 = 0 \pm 3$ replicas is an acceptable PDF set from the NNPDF4.0 fit.

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).

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



Ellipses at 68% CL