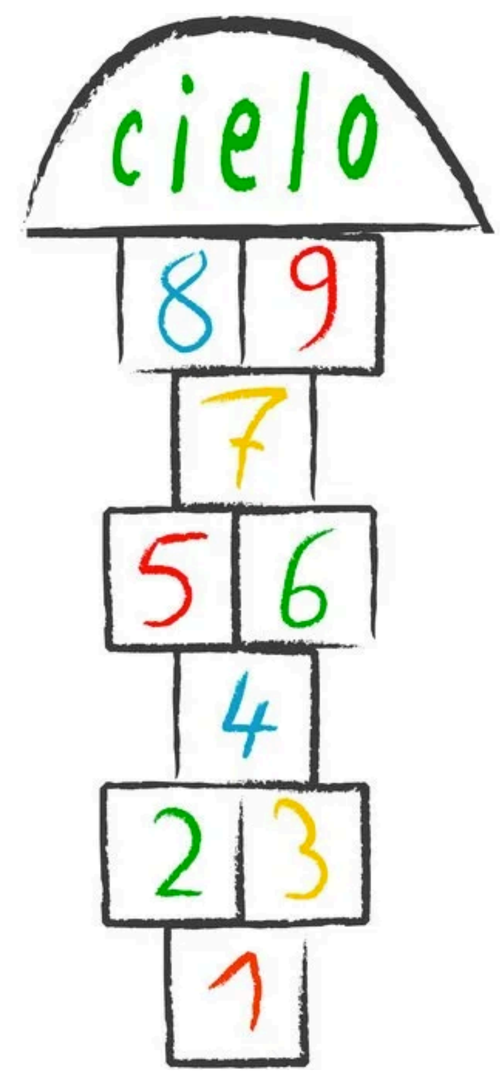
Parton distributions need representative sampling

Aurore Courtoy *for the CT collaboration*

Instituto de Física
National Autonomous University of Mexico (UNAM)

CTEQ-TEA members

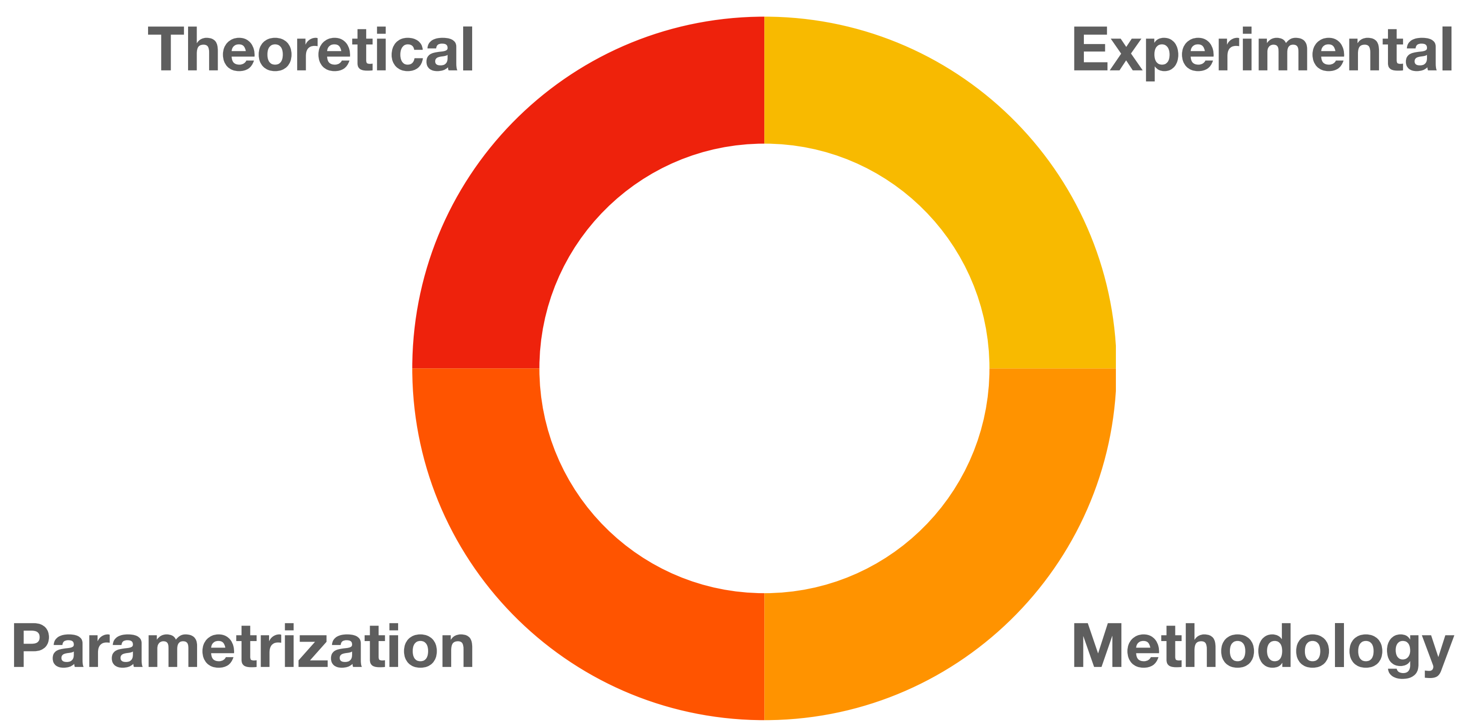
China: S. Dulat, J. Gao, T.-J. Hou, I. Sitiwaldi, M. Yan, and collaborators
Mexico: A. Courtoy
USA: T.J. Hobbs, M. Guzzi, J. Huston, P. Nadolsky, C. Schmidt, D. Stump, K. Xie, C.-P. Yuan



CTEQ meeting 2022

Instituto de Física UNAM
CONACYT
Consejo Nacional de Ciencia y Tecnología
dgapa
Dirección General de Asuntos del Personal Académico

Contributions to PDF uncertainties



In all four categories of uncertainties, we can further distinguish *PDF fitting accuracy* and *PDF sampling accuracy*.

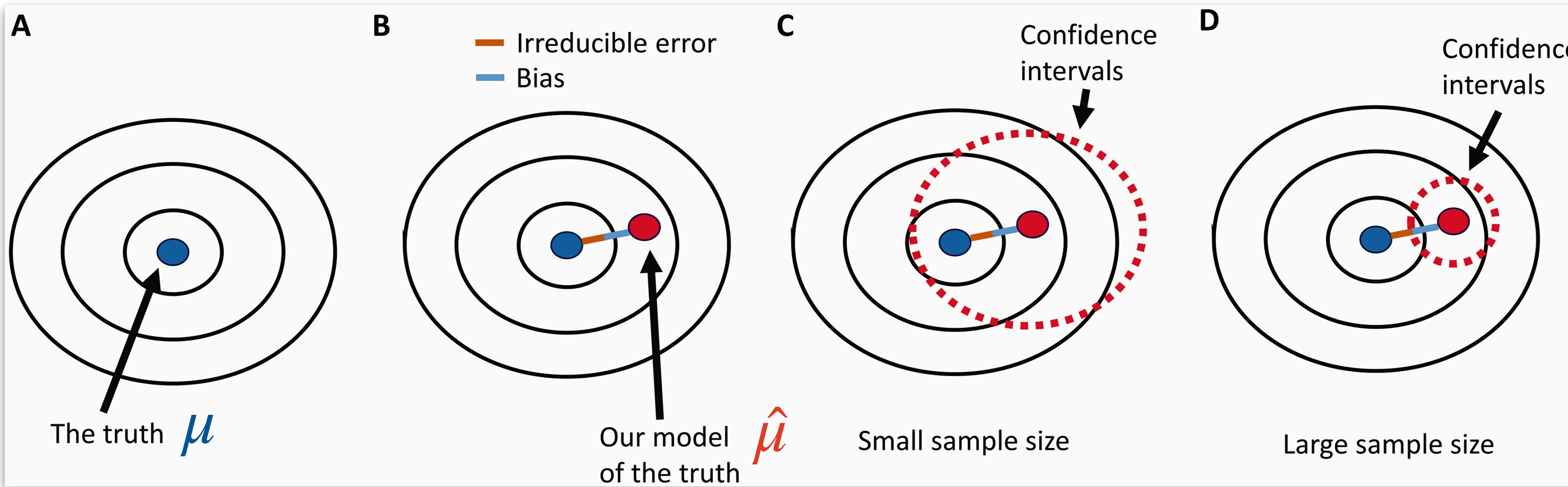
Accuracy in inputs —commonly integrated in global analyses.

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

A new avenue to understand PDF tolerance.

This talk.

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 disregards the *quality of the sampling*,

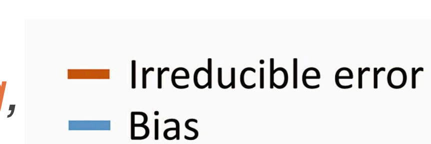


Illustration from:
Pavlos Msaouel (2022)
The Big Data Paradox in Clinical Practice
Cancer Investigation, 40:7, 567-576

Law of large numbers — Higgs XS

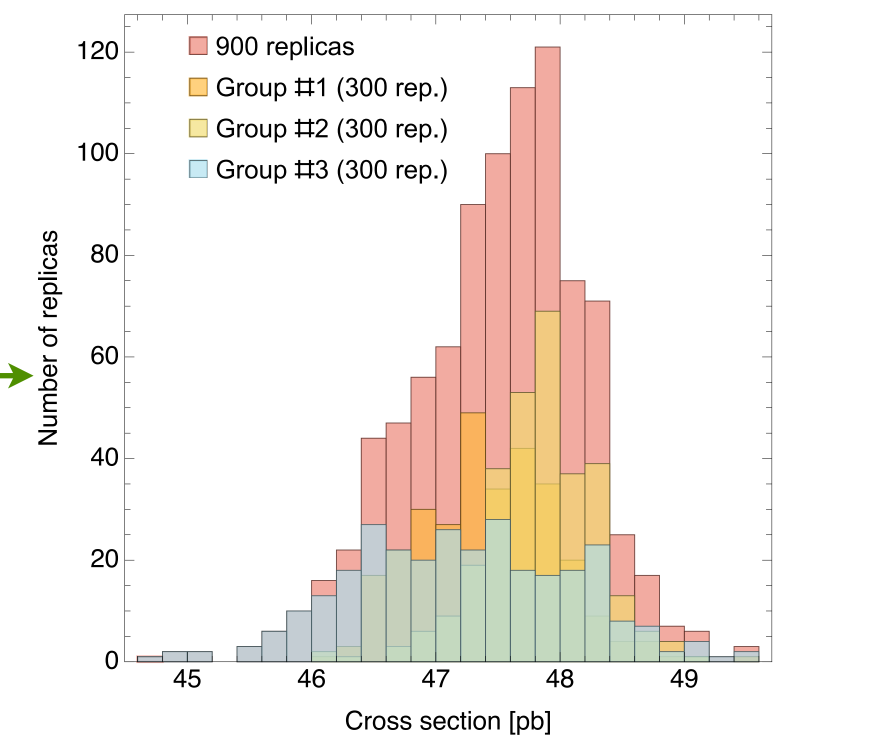
With an increasing size of sample $n \to \infty$, under a set of hypotheses, it is usually expected that the *deviation* on an observable

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

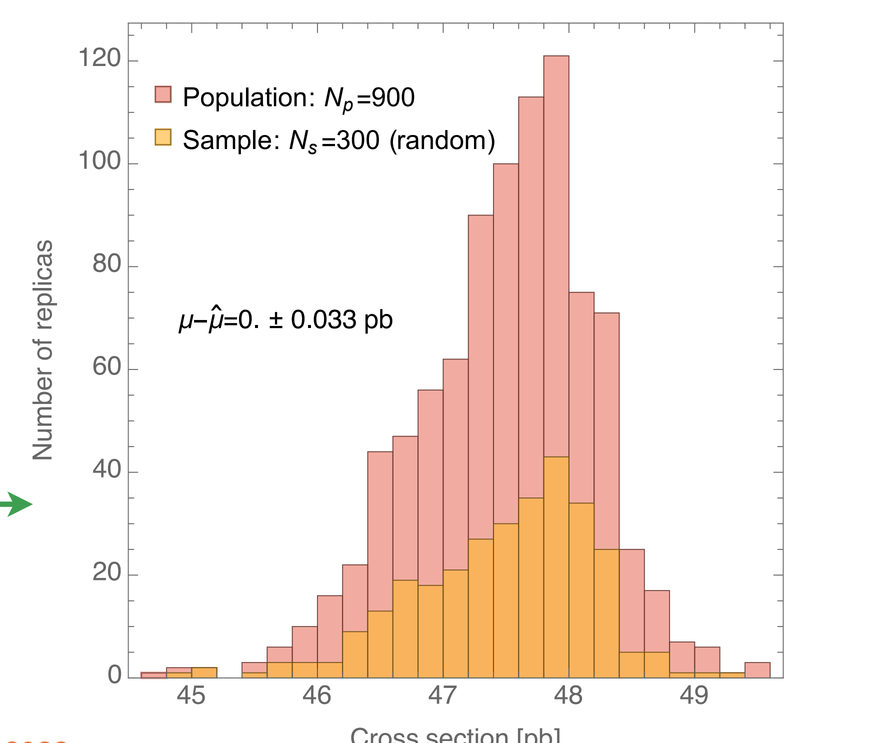
with σ the standard deviation, μ the true and $\hat{\mu}$ the determined values. *That's the law of large numbers.*

A toy sampling excercise

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

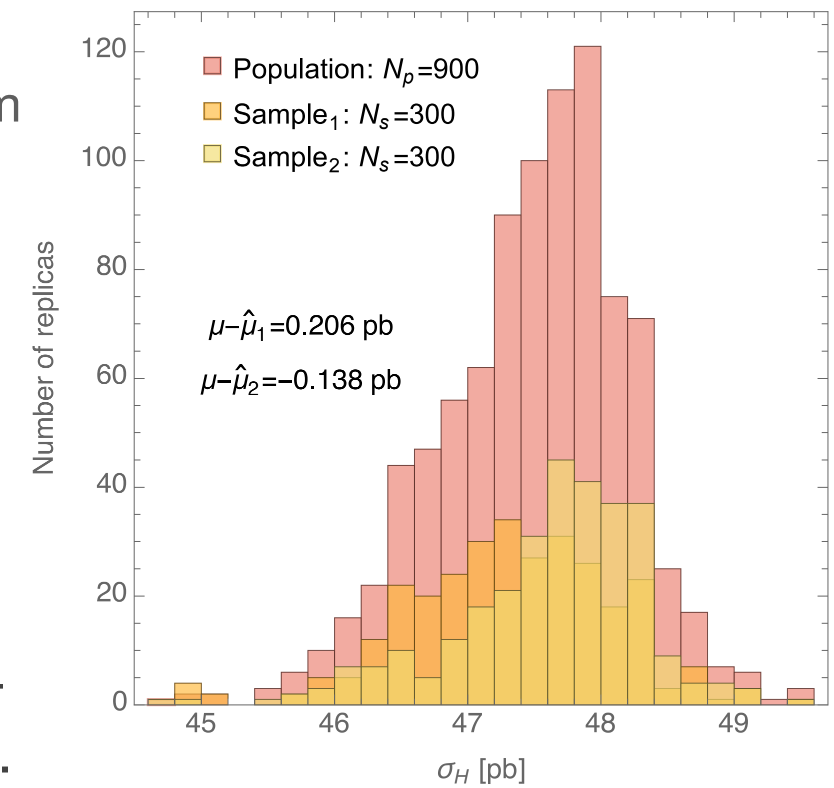


We **randomly** select 300 out of the 900 cross sections.
The law of large number is fulfilled in this case: there is no bias in the original sampling of the 3 sets of Higgs cross sections.



Trio identity— Higgs XS

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 disregards the *quality of the sampling* — distribution of n for a population size N/measure of the parameter space.

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

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

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

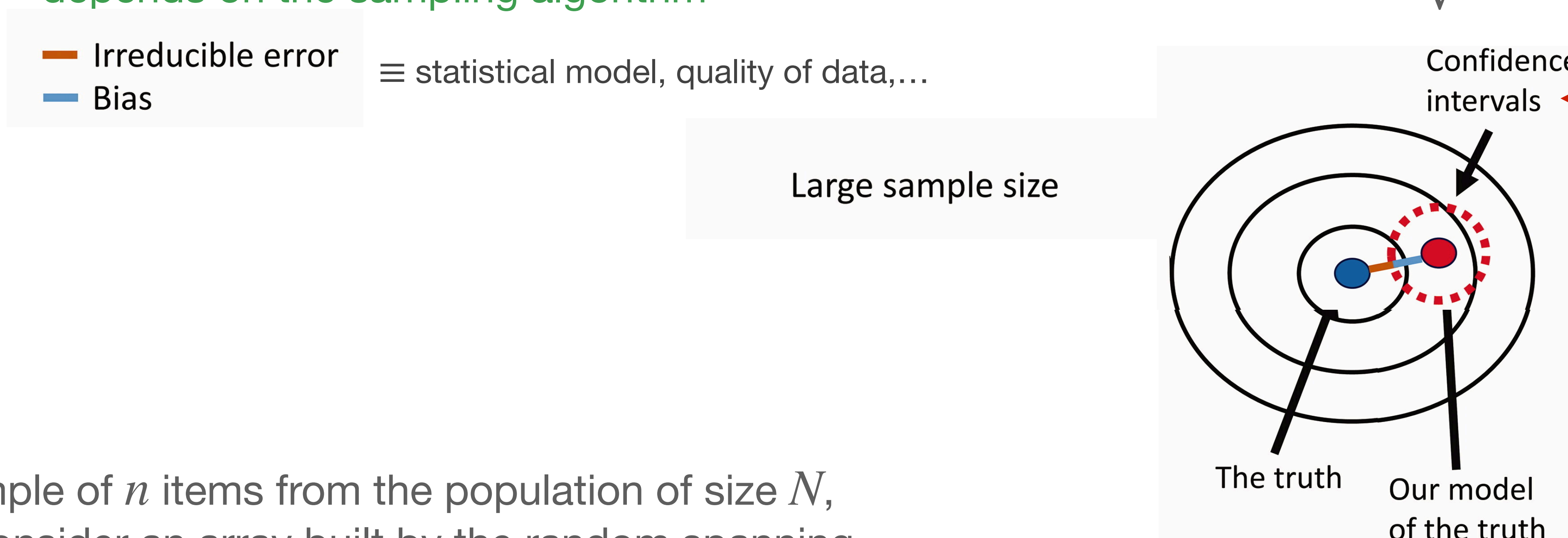
Xiao-Li Meng
The Annals of Applied Statistics
Vol. 12 (2018), p. 685

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

depends on the sampling algorithm

can tend to $\sigma/\sqrt{n}$ for random sampling

$\equiv$ statistical model, quality of data,…



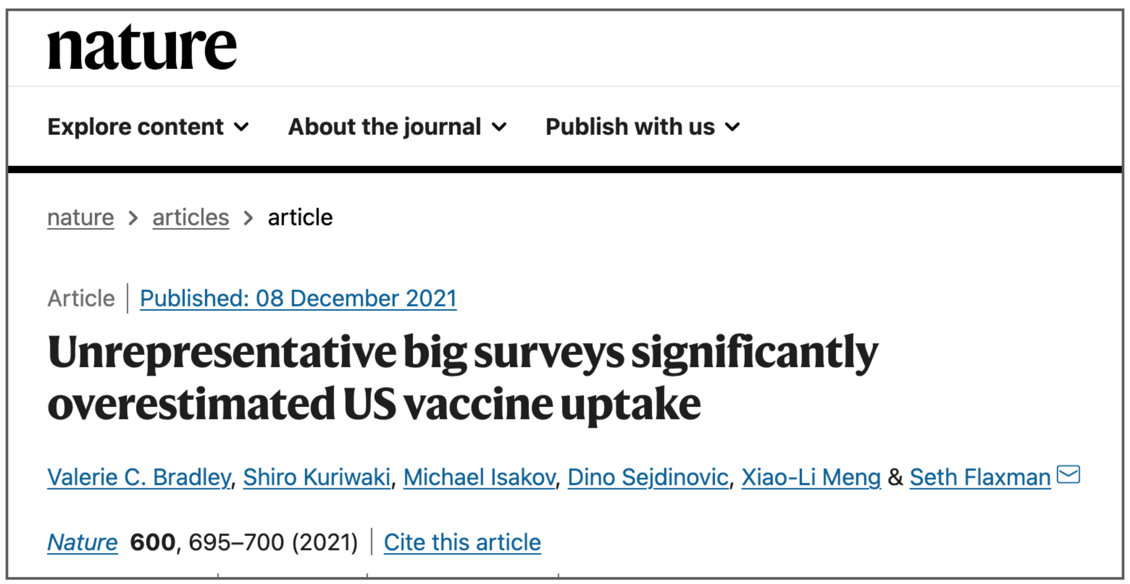
For a sample of n items from the population of size N, we can consider an array built by the random spanning of the binary responses of the $N-n$ (0) and n (1) items, so that

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

Hickernell
MCQMC 2016
1702.01487

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

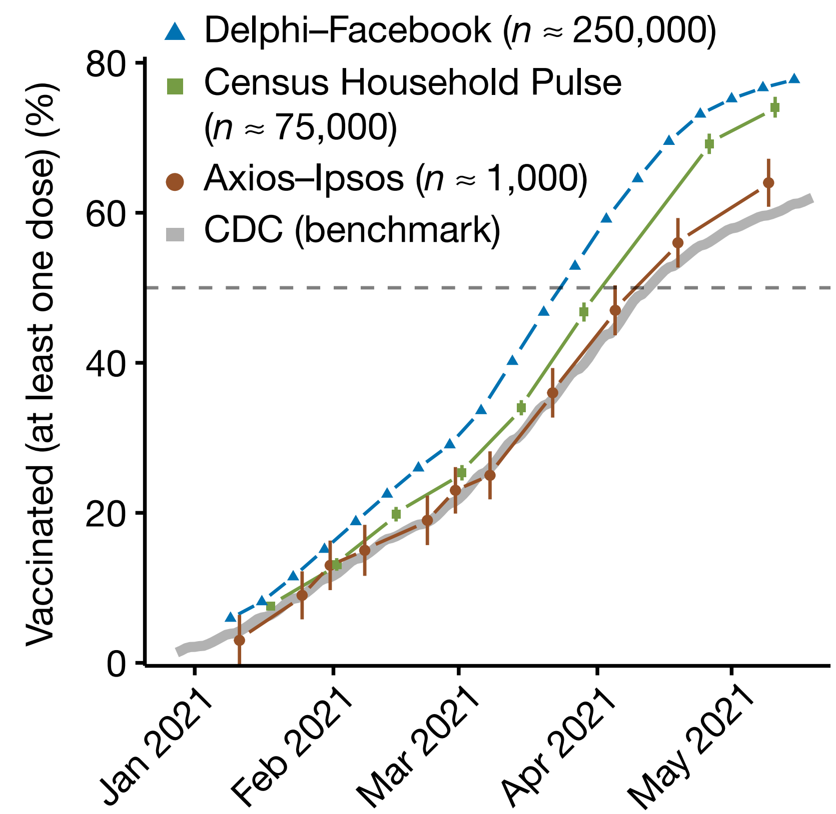
nature > articles > article

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.

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?

Sampling bias

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

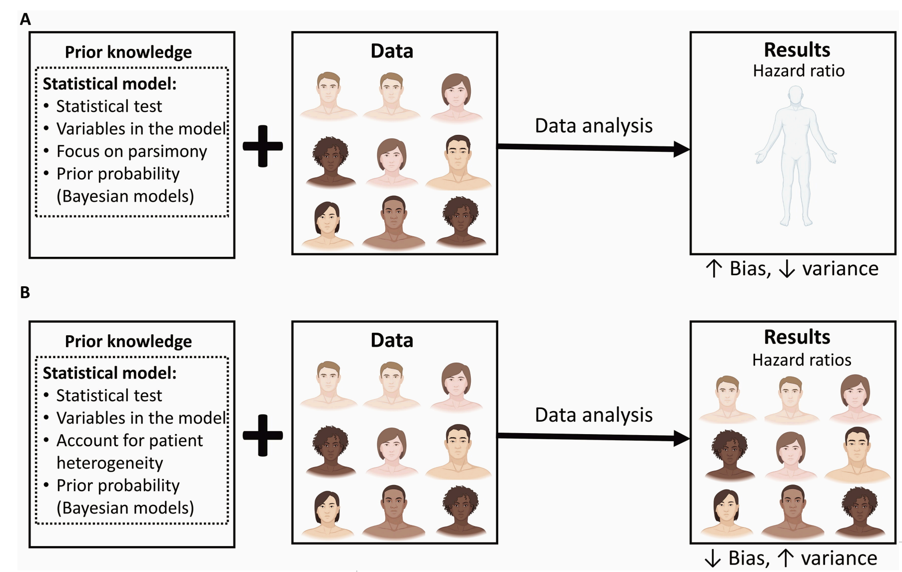
Standard error estimates can be misleadingly small.

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

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

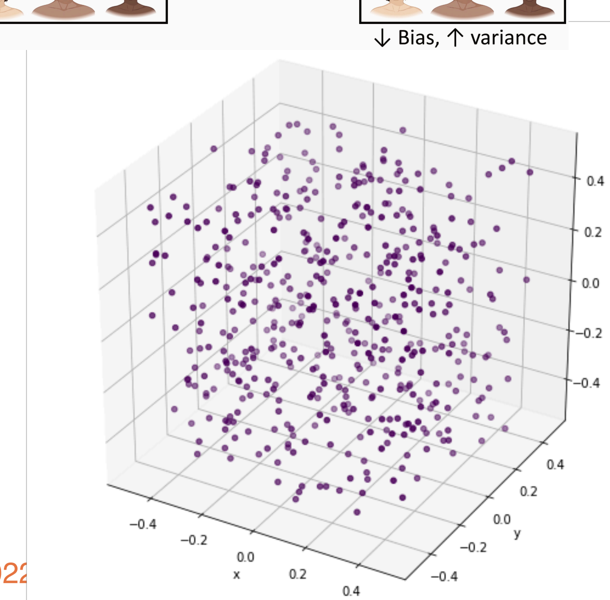
Tractable problems like the vaccination rate, presidential elections or clinical practice can benchmark their confounding correlation.

e.g. [Msaouel, Cancer Investigation, 40:7, 567-576]



In some cases, Monte Carlo integration problems can optimize their sampling by considering the effect of the confounding correlation.

e.g. [Hickernell, MCQMC 2016, 1702.01487]

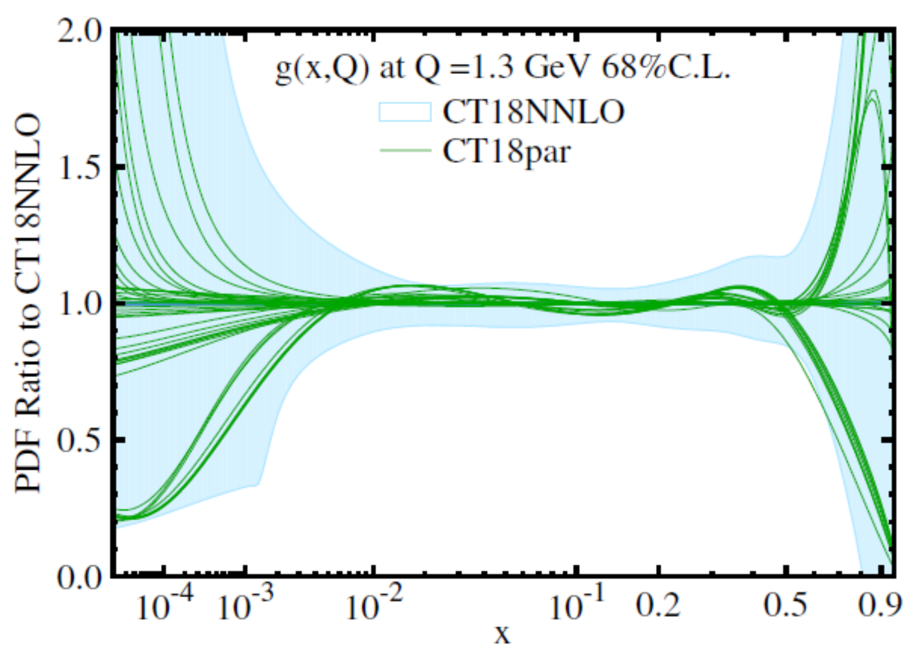


Sampling bias in PDF global analyses

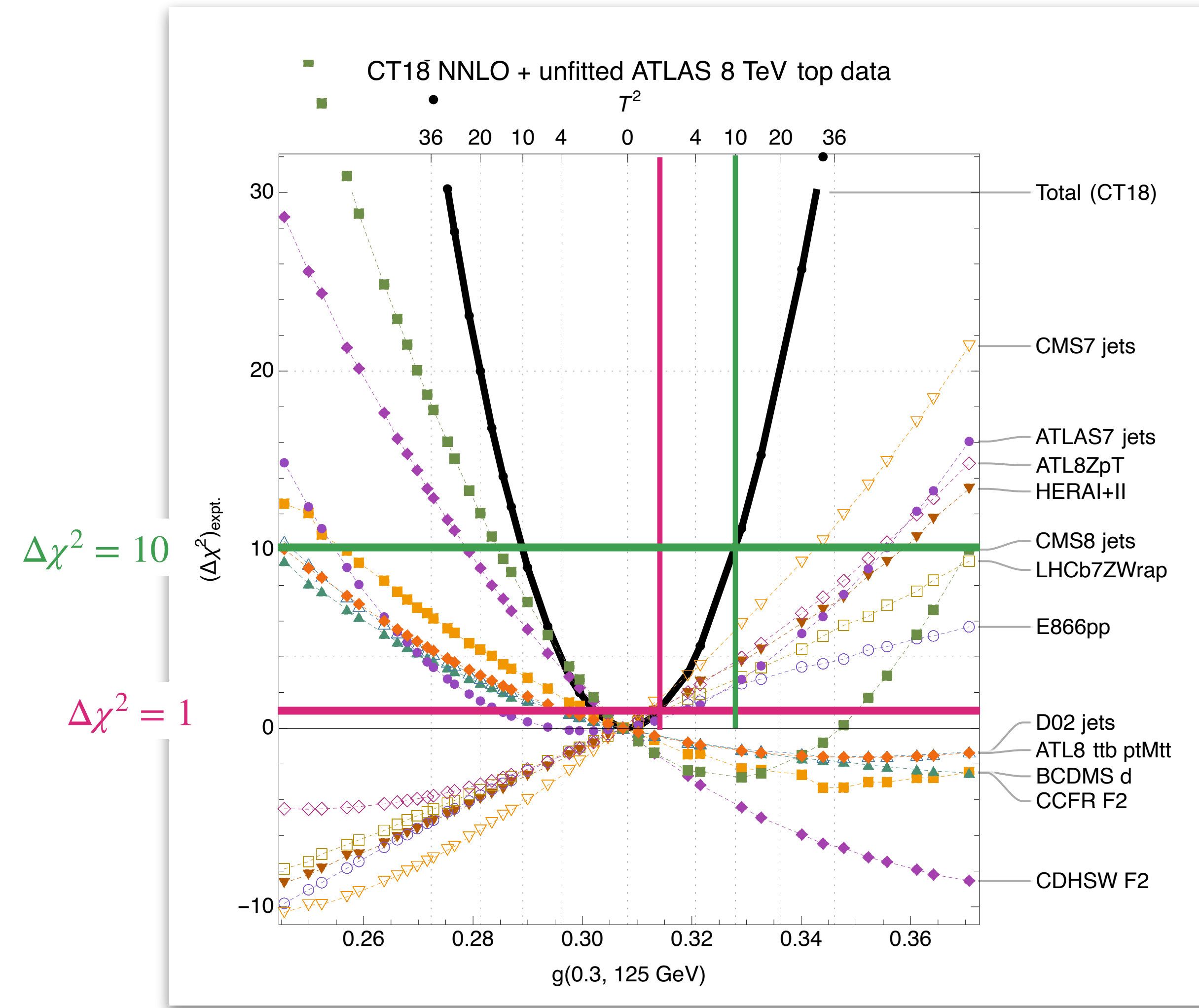
How do we know the "data+sampling defect=confounding correlation" of our analysis?

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*).



Dimensions of the problem given by the number of parameters=eigenvector (EV) directions.



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

Sampling matters for PDF global analyses

PDF analyses are affected by the bias/variance balance due to the high number of dimensions of the problem.

Sampling bias must be studied to faithfully reconstruct uncertainties.

That’s our take-away message.

Increasing interest in bias/variance dilemma in high-dimensional problems

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

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

https://doi.org/10.1038/s41586-021-04198-4
Received: 18 June 2021

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

SCIENCE ADVANCES | RESEARCH ARTICLE

MATHEMATICS

Models with higher effective dimensions tend to produce more uncertain estimates

Arnald Puy[1,2,3]*, Pierfrancesco Beneventano[4], Simon A. Levin[2], Samuele Lo Piano[5], Tommaso Portaluri[6], Andrea Saltelli[3,7]

The Big Data Paradox in Clinical Practice

Pavlos Msaouel

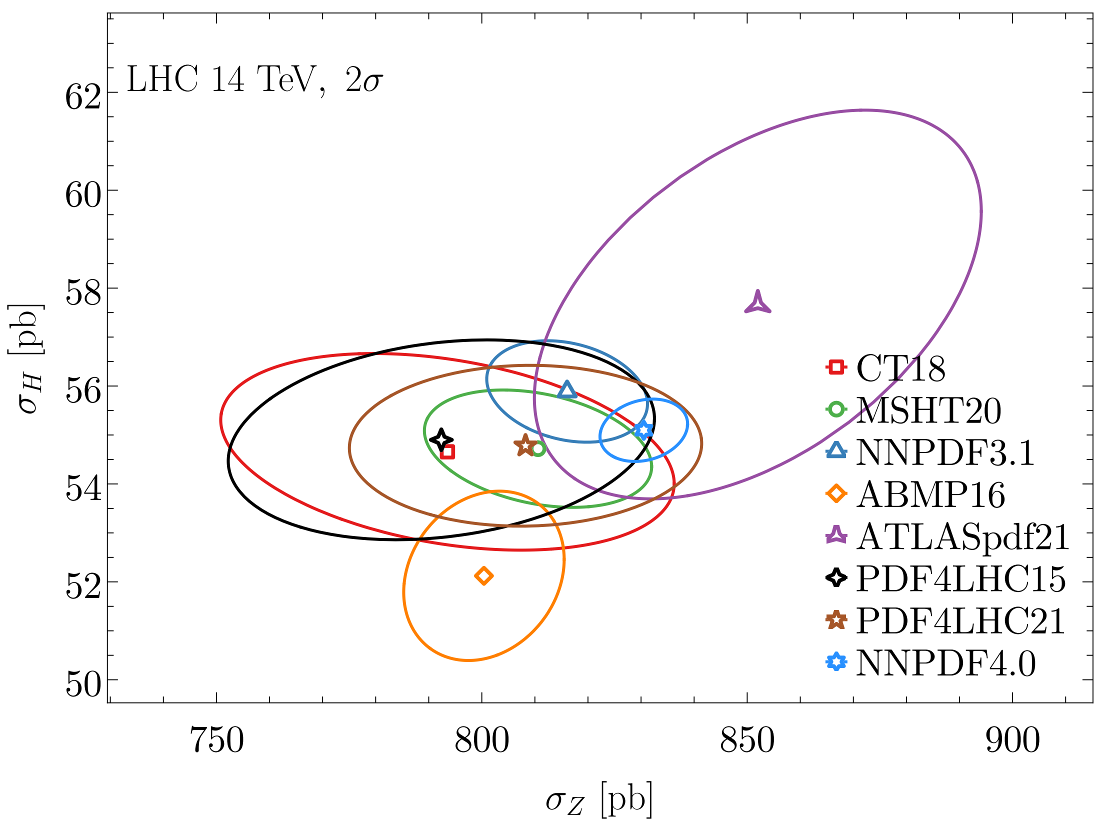
To cite this article: Pavlos Msaouel (2022) The Big Data Paradox in Clinical Practice, Cancer Investigation, 40:7, 567-576, DOI: 10.1080/07357907.2022.2084621

Uncertainties from global analyses of proton structure

Now focusing on the details of uncertainties for PDF analyses.

Recent advancements in the determination of unpolarized PDFs:
CT18, MSHT20, NNPDF4.0, ATLASpdf21 as well as PDF4LHC21.

Precision PDFs (Snowmass 21 WP) [2203.13923]



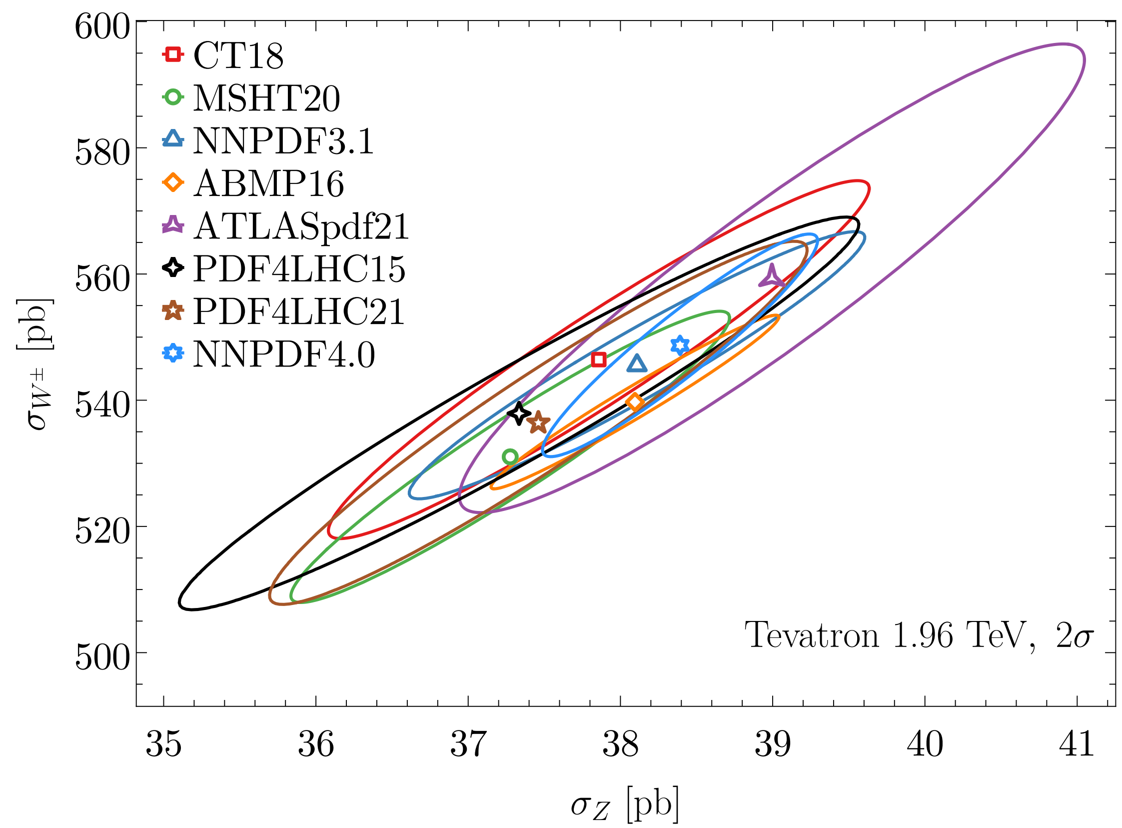


Figure-of-merit/objective function and priors

Chi-square definition

$$\chi^2 = \sum_{i,j}^{N_{pt}} (T_i - D_i)(\mathrm{cov}^{-1})_{ij}(T_j - D_j)$$

$$(\mathrm{cov})_{ij} \equiv s_i^2 \delta_{ij} + \sum_{\alpha=1}^{N_\lambda} \beta_{i,\alpha}\beta_{j,\alpha},$$

$$\beta_{i,\alpha} = \sigma_{i,\alpha} X_i,$$

D_i, T_i, s_i are the central data, theory, uncorrelated error
$\beta_{i,\alpha}$ is the correlation matrix for N_λ nuisance parameters.

Experiments publish $\sigma_{i,\alpha}$.
To reconstruct $\beta_{i,\alpha}$, we need to decide on the normalizations X_i.

Choices:

- $X_i = D_i$: "**experimental scheme**"; can result in a bias
- $X_i =$ "fixed" T_i : "t_0 **scheme**"; can result in a (different) bias

For Hessian-based global analyses:

Figure-of-merit and tolerance criteria will define the size of uncertainties.

For Monte Carlo-based global analyses:

"The posterior probability for the parametrization depends on both the figure-of-merit […] given the parameters and on the prior probability." NNPDF [M. Ubiali, HP2 2022 workshop, Durham, 2022/09]

Do we understand sampling for QCD global analyses?

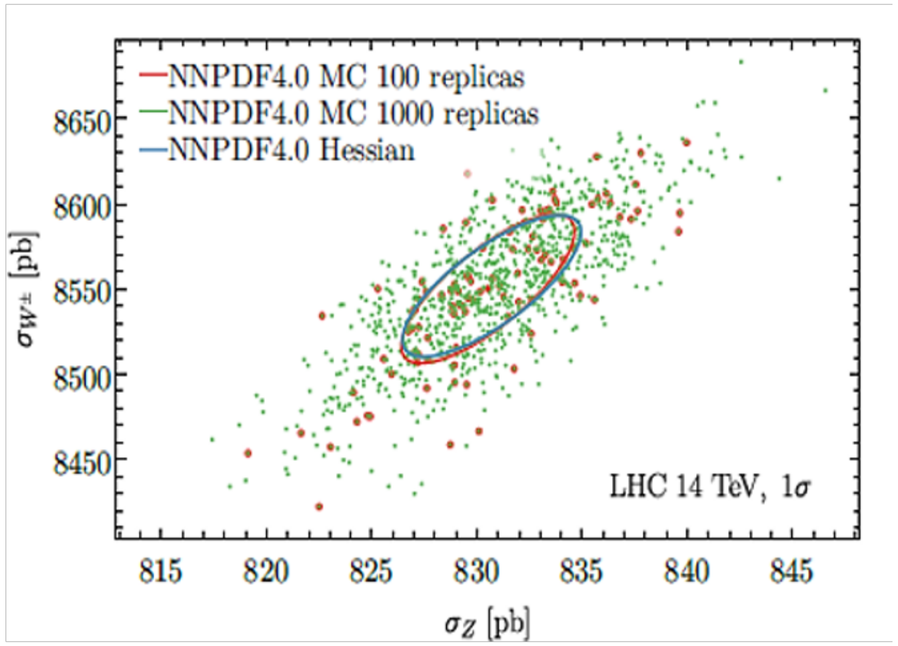
Sampling of multidimensional spaces ($d \gg 20$) can be exponentially inefficient and require $n > 2^d$ replicas to obtain a convergent expectation value.
Most probably an intractable problem.

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

How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling — sampling on the space of the data/bootstrap/resampling of data.

Uncertainties are then unweighted averages.



➪ Caveat: we found that Hessian and MC uncertainties are in good agreement.

Do we understand sampling for QCD global analyses?

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

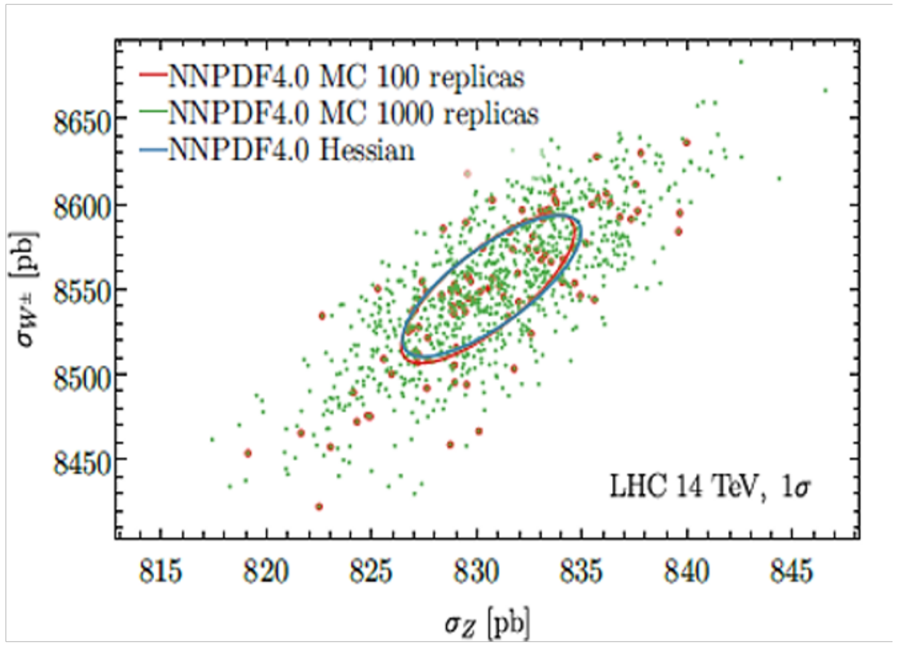
[Hickernell, MCQMC 2016, 1702.01487]
[Sloan,I.H.,Wo´zniakowski, 1997]

How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling — sampling on the space of the data/bootstrap/resampling of data.

Uncertainties are then unweighted averages.

➪ Caveat: we found that Hessian and MC uncertainties are in good agreement.



Algorithm for observable-oriented verification of representative uncertainty

Specific QCD observables: only few effective large dimensions contribute the bulk of the uncertainty.

Hopscotch scans

Algorithm for observable-oriented verification of representative uncertainty

"Parton distributions need a representative sampling"

[AC et al, 2205.10444]

We determine dimensions of the problem from specific QCD observables: only few **effective large dimensions** contribute the bulk of the uncertainty.

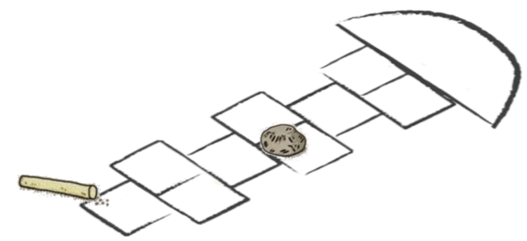
To sample the PDF dependence for Monte Carlo-based global analyses:
sample primarily the coordinates with large variations of physical cross section σ.

Using NNPDF4.0 public code, we then employ:

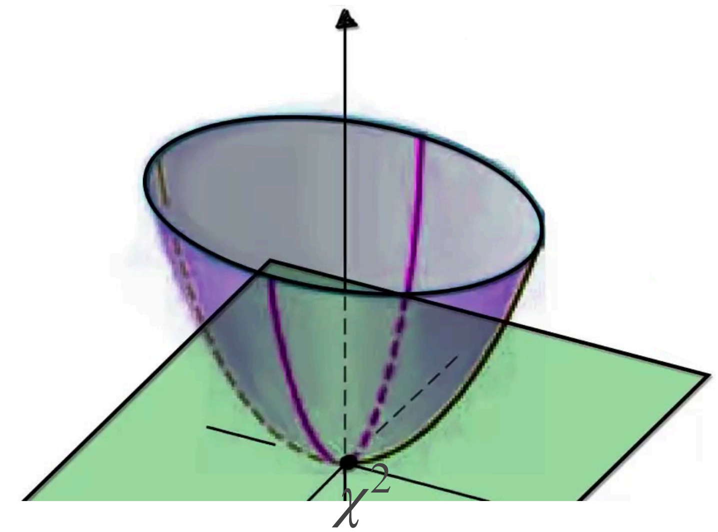
$n =$ the number of replicas/EV directions/…

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

How to play hopscotch?

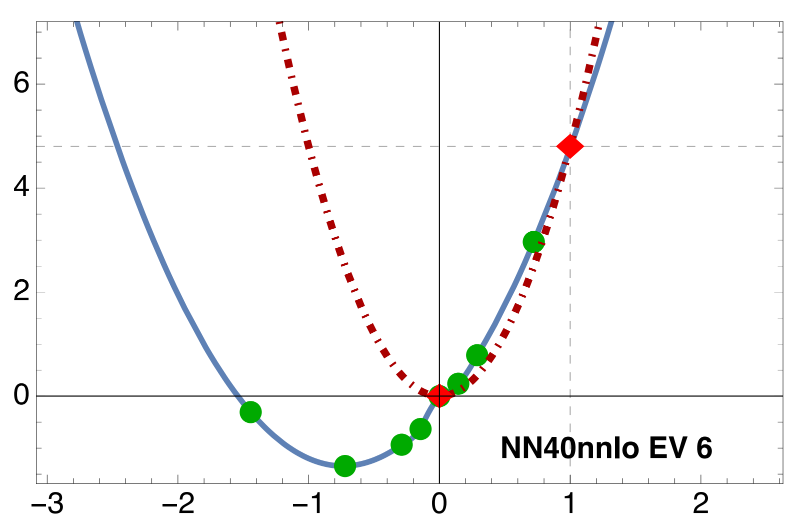


In the Hessian representation, the chi square behaves like a paraboloid of n_{param} dimensions, thus defining a global minimum.



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

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

Step 1

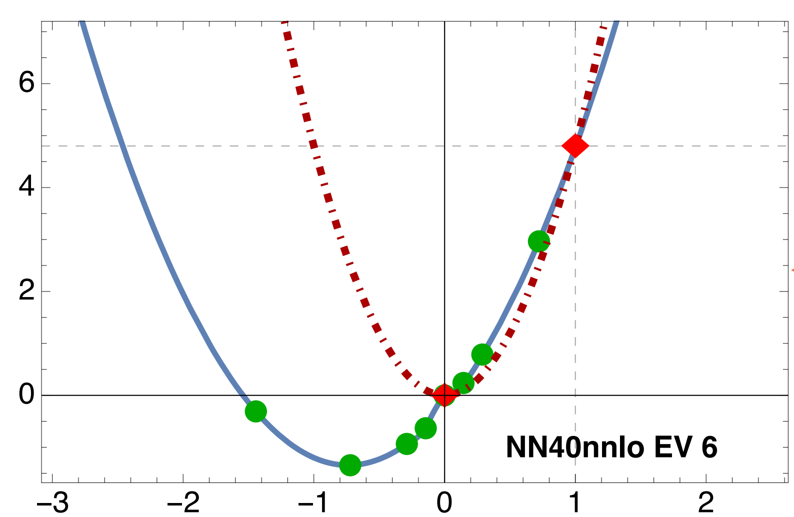
The NNPDF4.0 Hessian set ($n = 50$) defines a coordinate system on a manifold corresponding to the largest variations of the PDF uncertainty —red dots and curve.

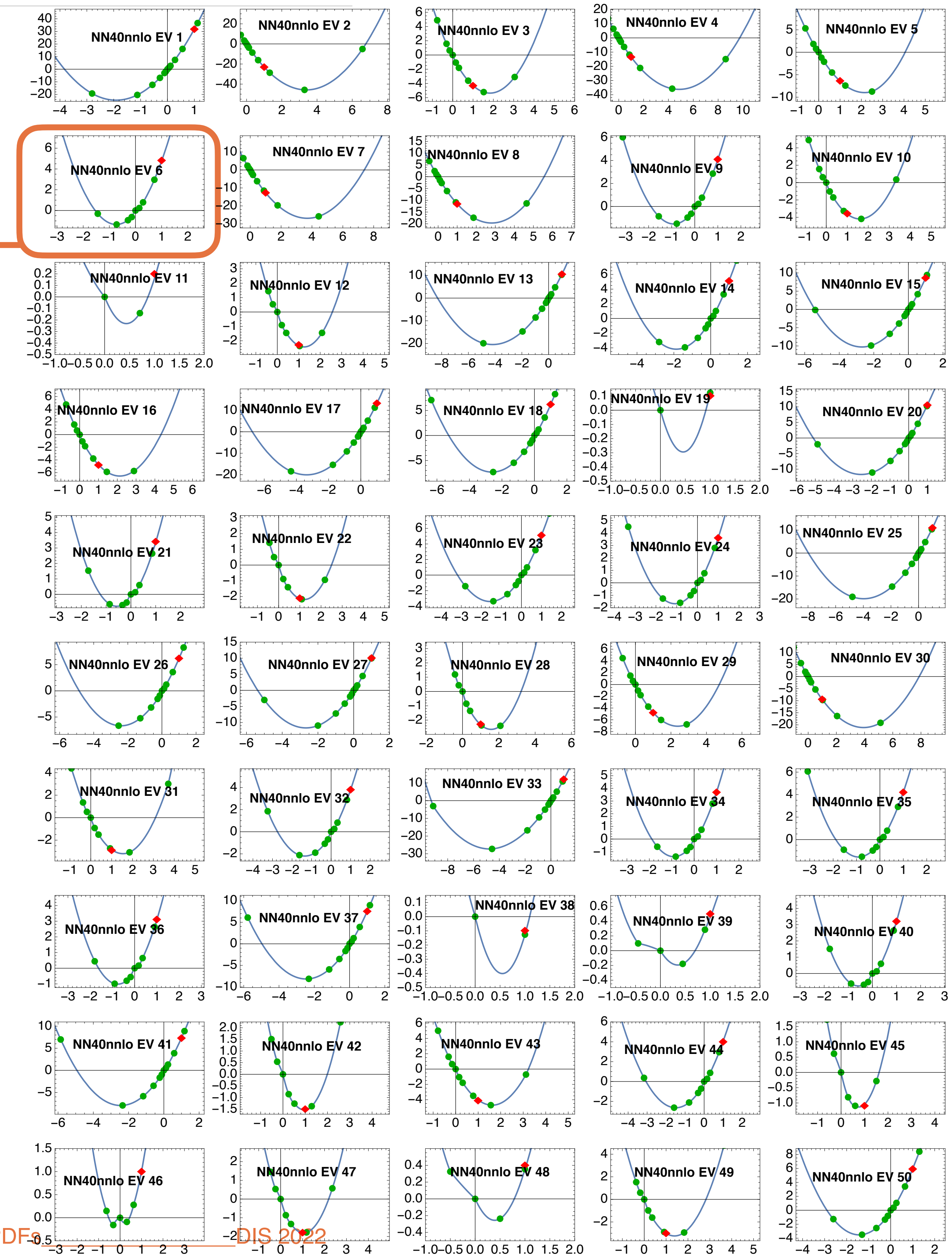
[NNPDF, 2109.02653]

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.



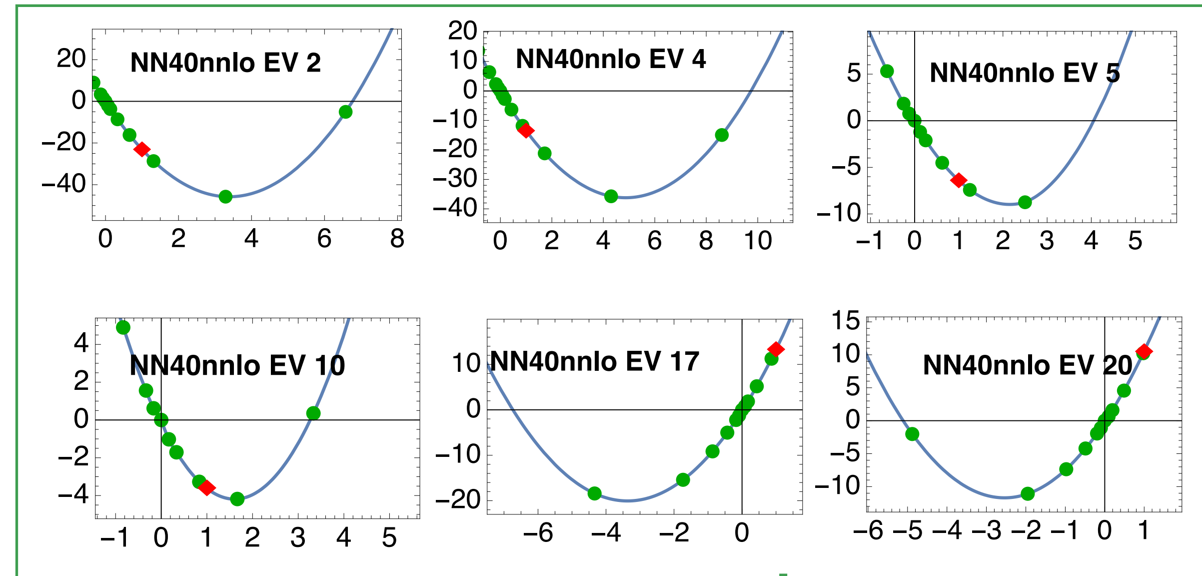


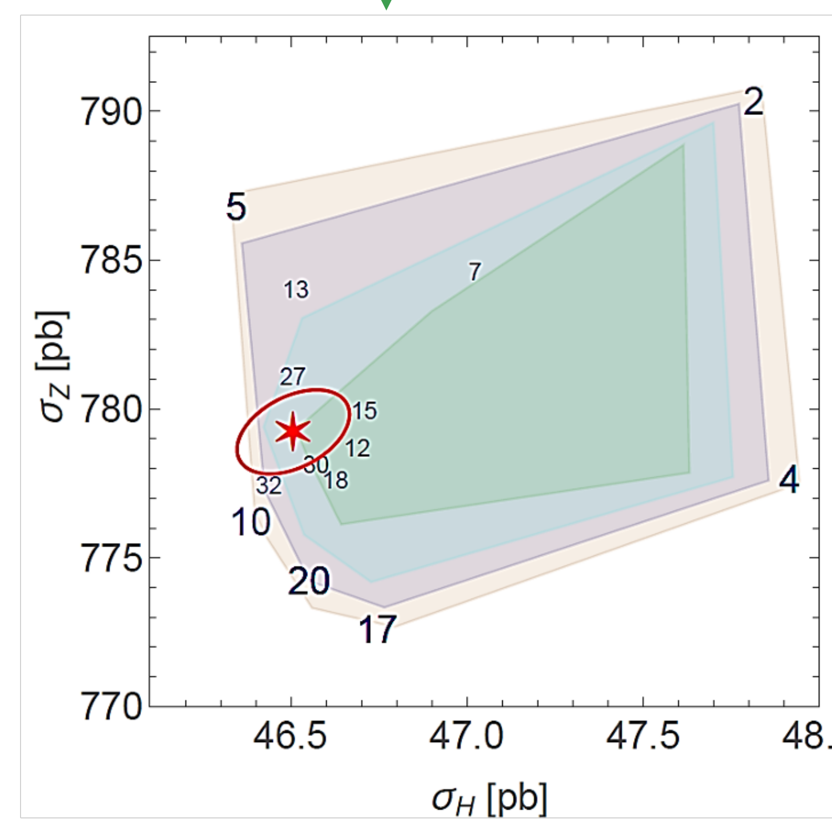
A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

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.

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



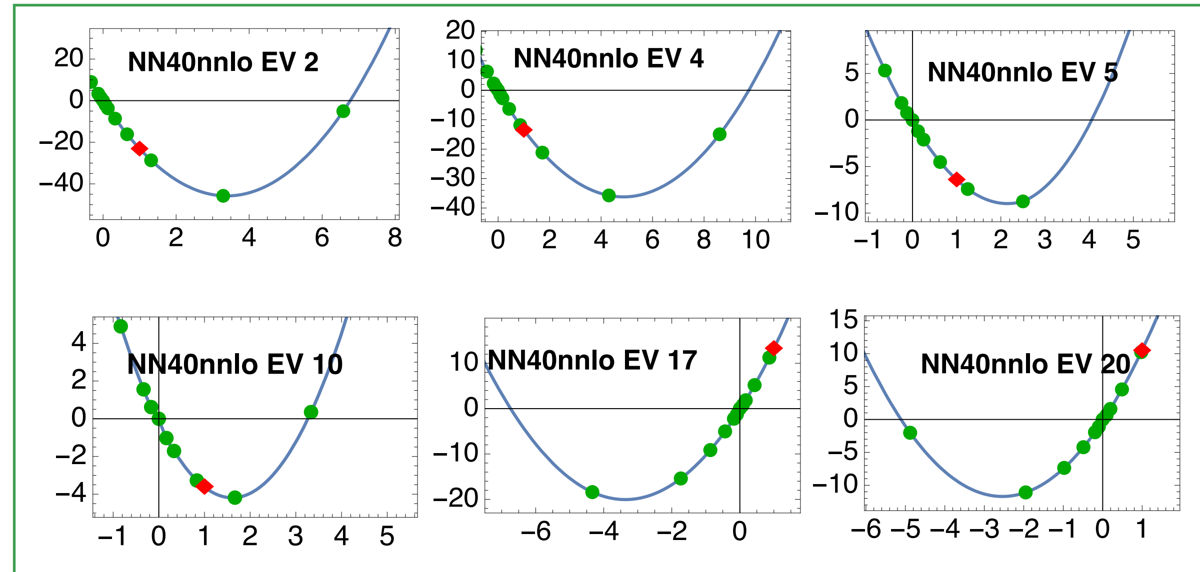


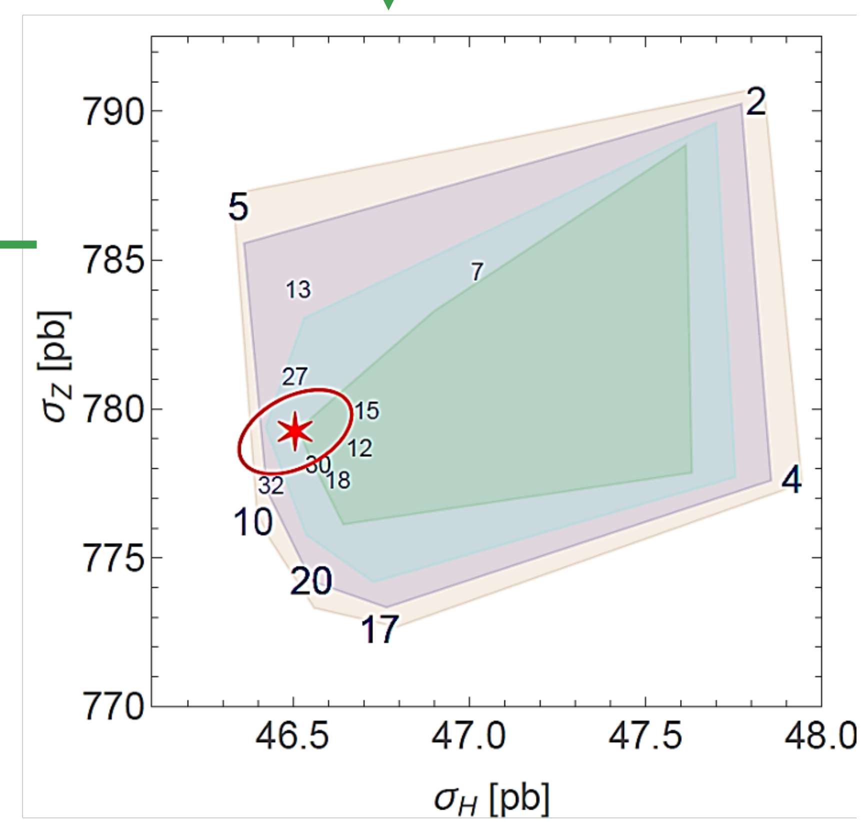
A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

Step 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.

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

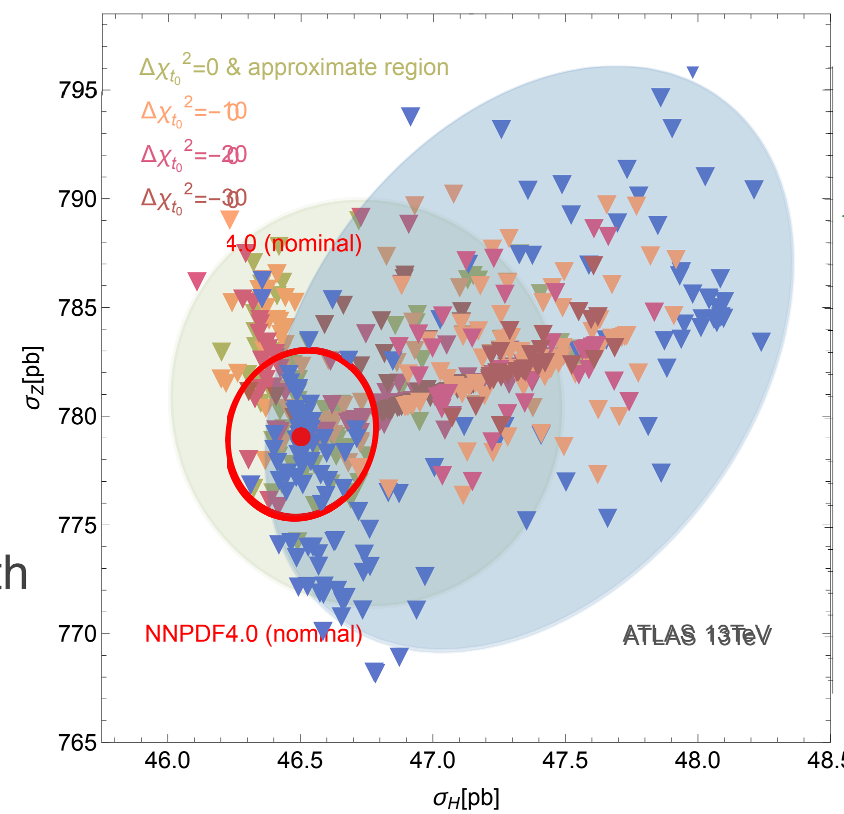




Step 4

For each pair of cross sections, we generate 300 replicas by sampling uniformly along the “large” EV directions.

The color ellipse is an approximate region containing all found replicas with $\Delta\chi^2_{exp/t_0} < 0$.



[Anwar, Hamilton, Nadolsky, 1901.05511]

Monte-Carlo sampling sensitivity for PDFs

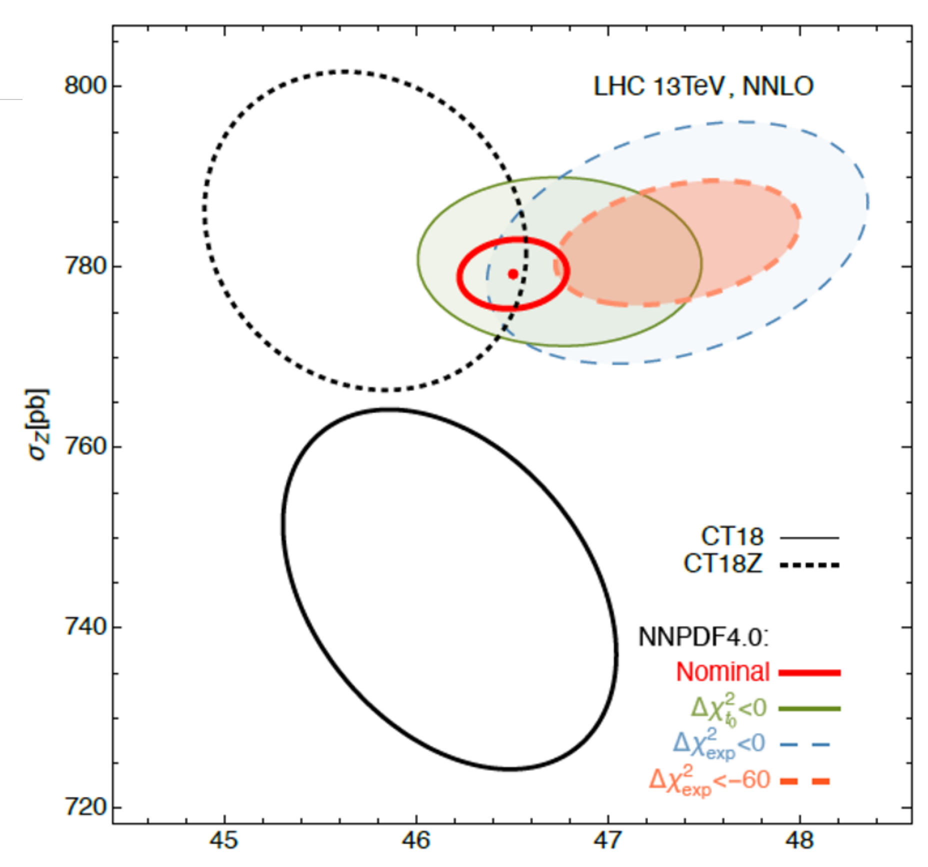
Cross sections for Higgs vs. Z and for $W^{\pm}$ vs. Z for LHC at 13TeV

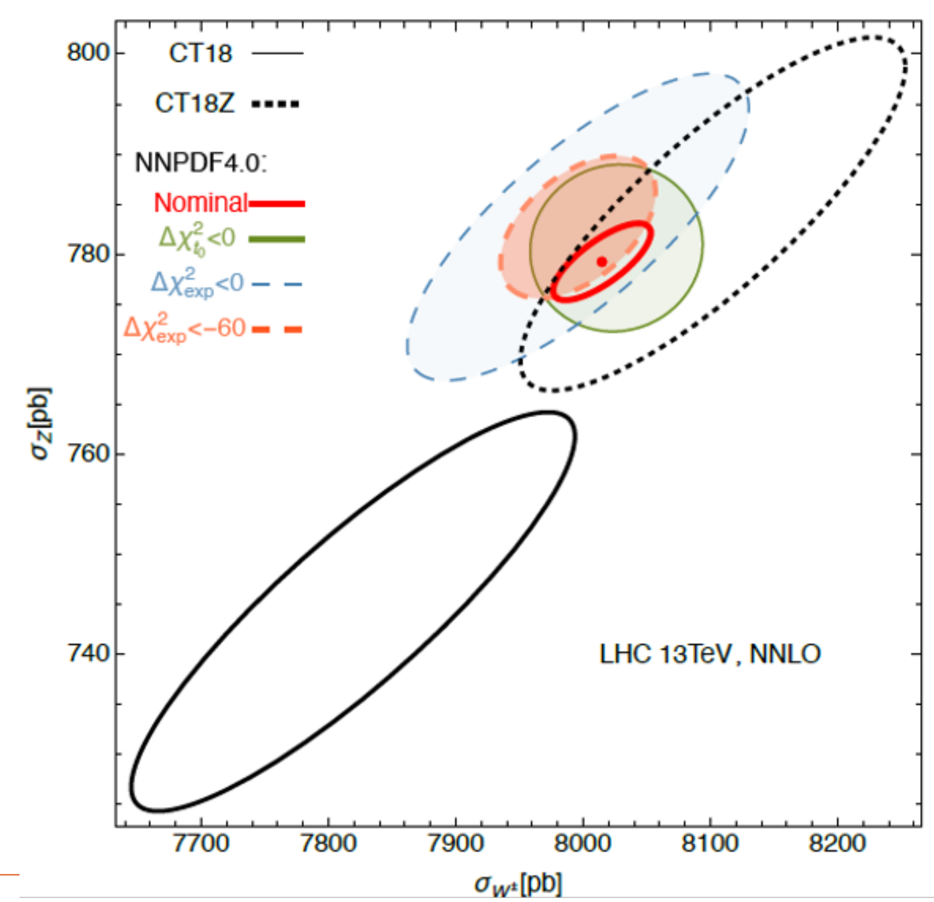
Legend:

- CT18NNLO & CT18Z (NNLO)
- Nominal NNPDF4.0
- Green ellipses for t_0 prescription for the objective function:
  - found through **the hopscotch scan — a dimensional reduction method.**
- Blue and brown filled ellipses for experimental χ^2 prescription:
  - areas of possible solutions corresponding to an equal or lower $(\Delta\chi^2 < 0)$, and even $(\Delta\chi^2 < -60)$ chi square w.r.t. the nominal solution

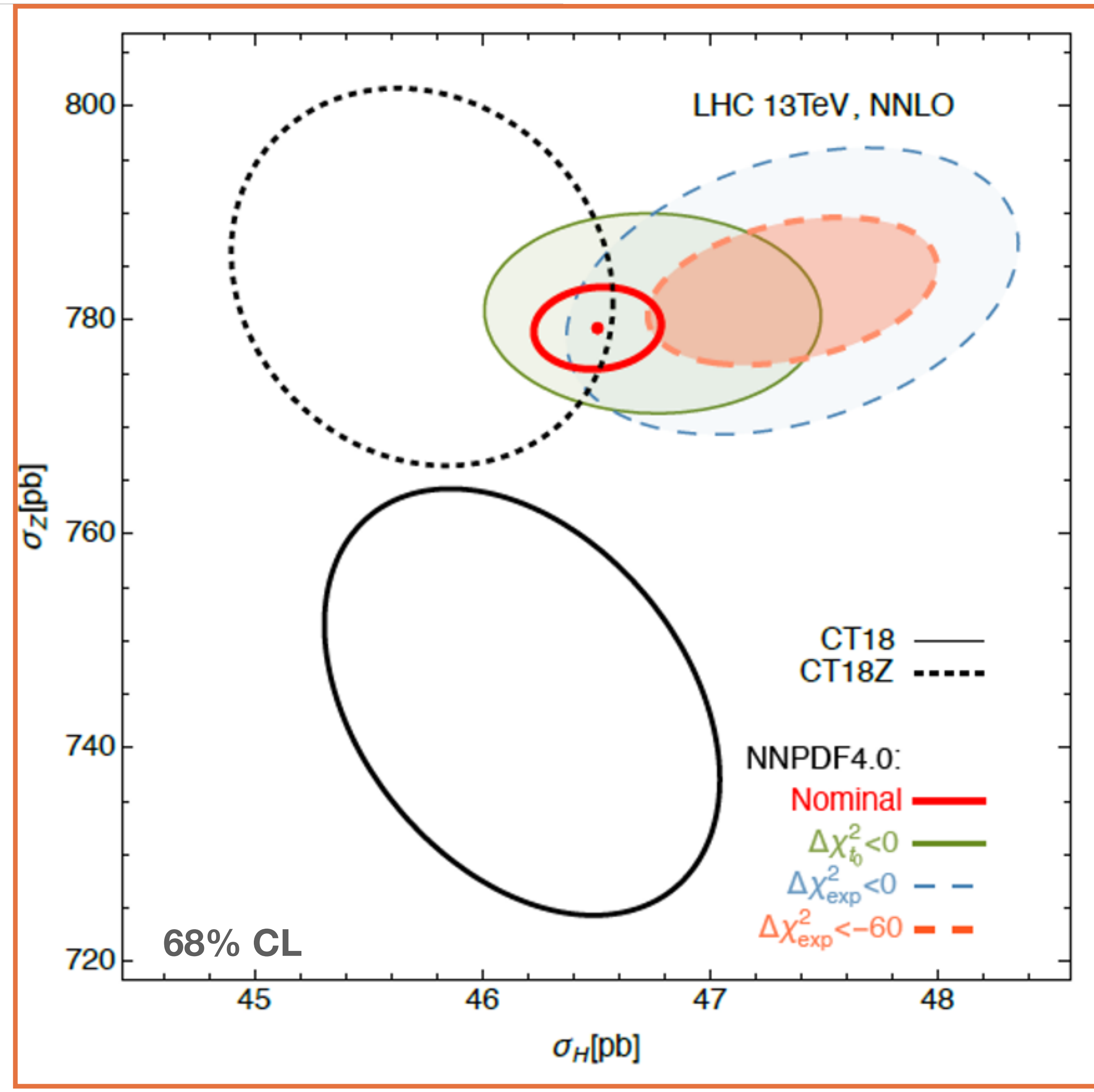
Hopscotch scans illustrated for the NNPDF4.0 —thanks to the publicly available code.

Applicable to other global analyses using similar methodology and a large enough parameter space.





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

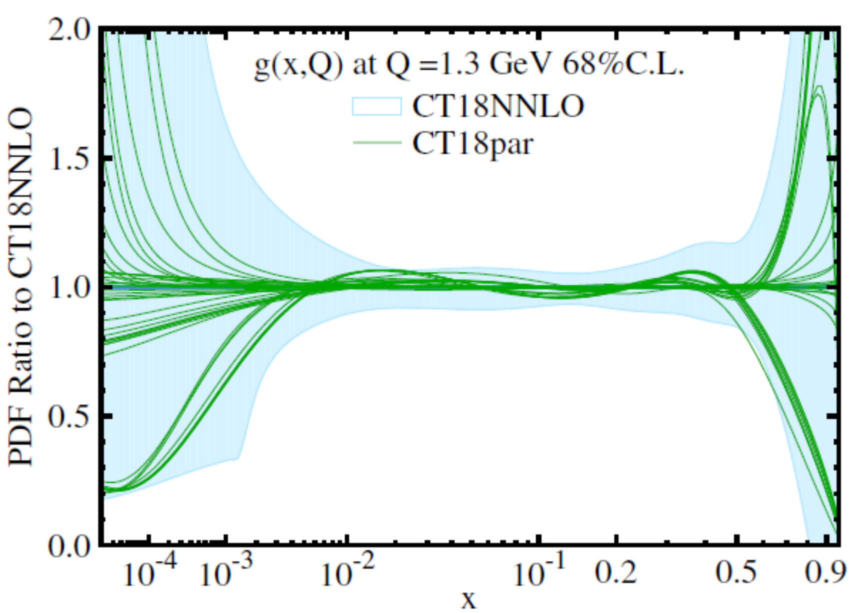


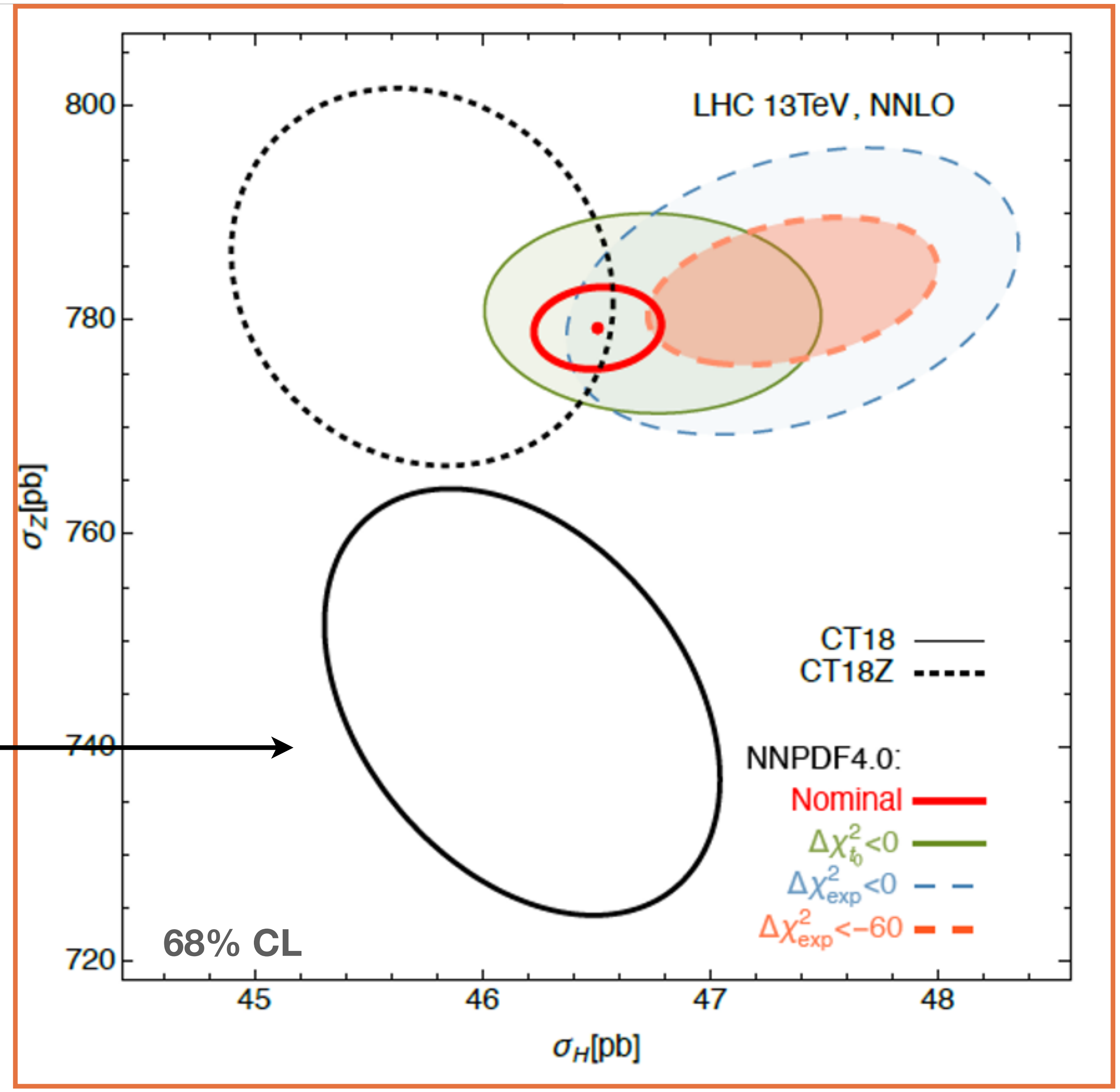
Color ellipses:

- 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:

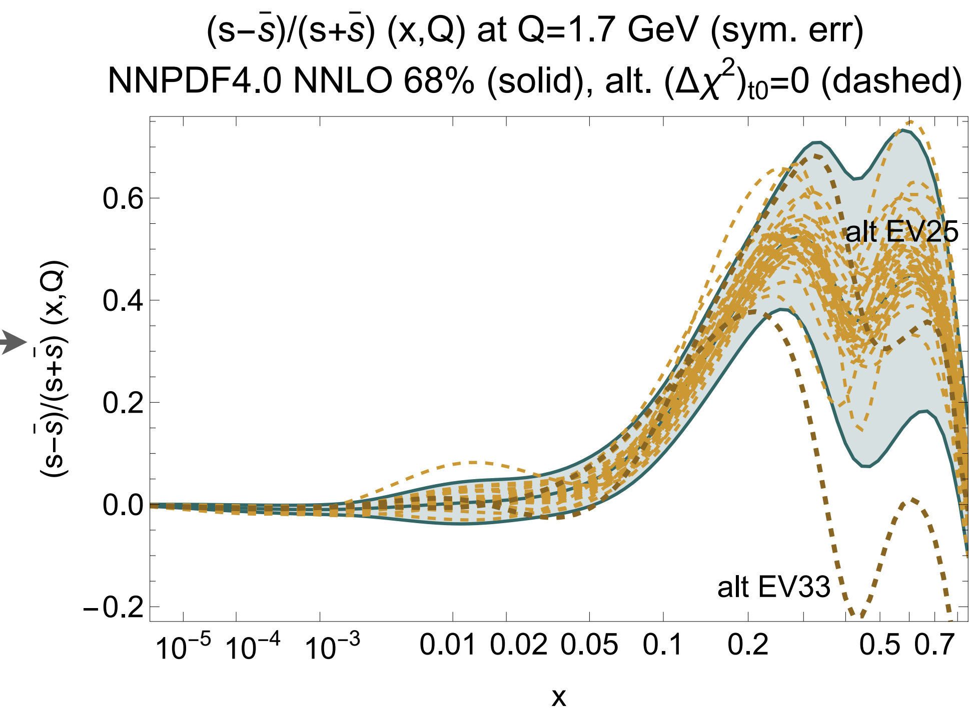
- areas of possible solutions corresponding to lower $(\Delta\chi^2 < 0)$ w.r.t. the nominal solution
- found through the hopscotch scan — a dimensionality reduction method.

Monte Carlo and Hessian representation — role of constraints

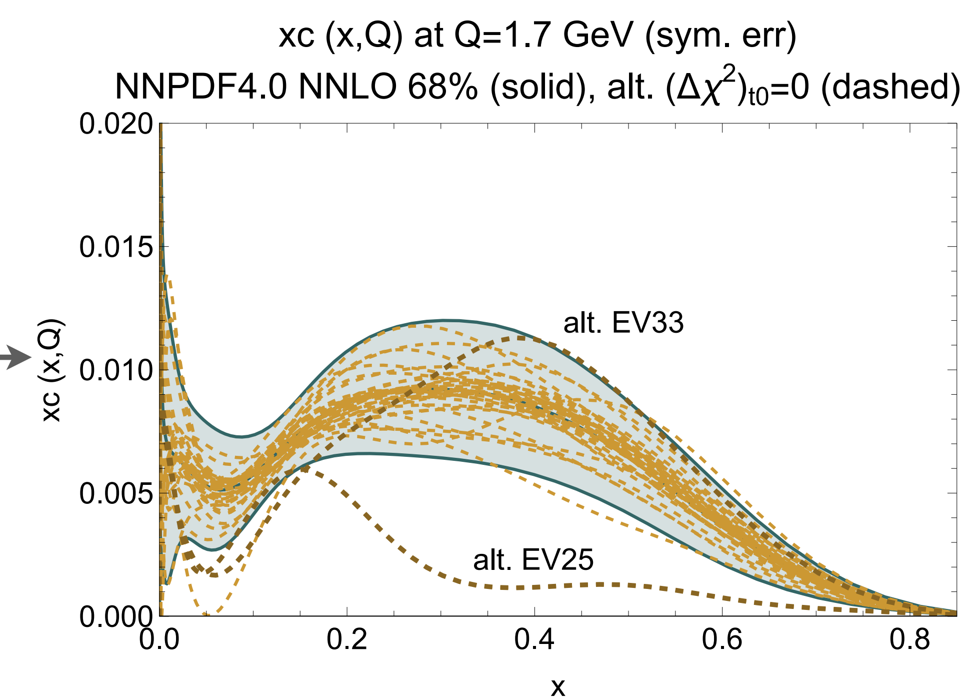
Role of constraints in global analyses: can act as *priors* to the final distributions.

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. 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.
4. Needs to be formally connected to known ML concerns — e.g. *no free lunch theorems* (more soon)

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

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:

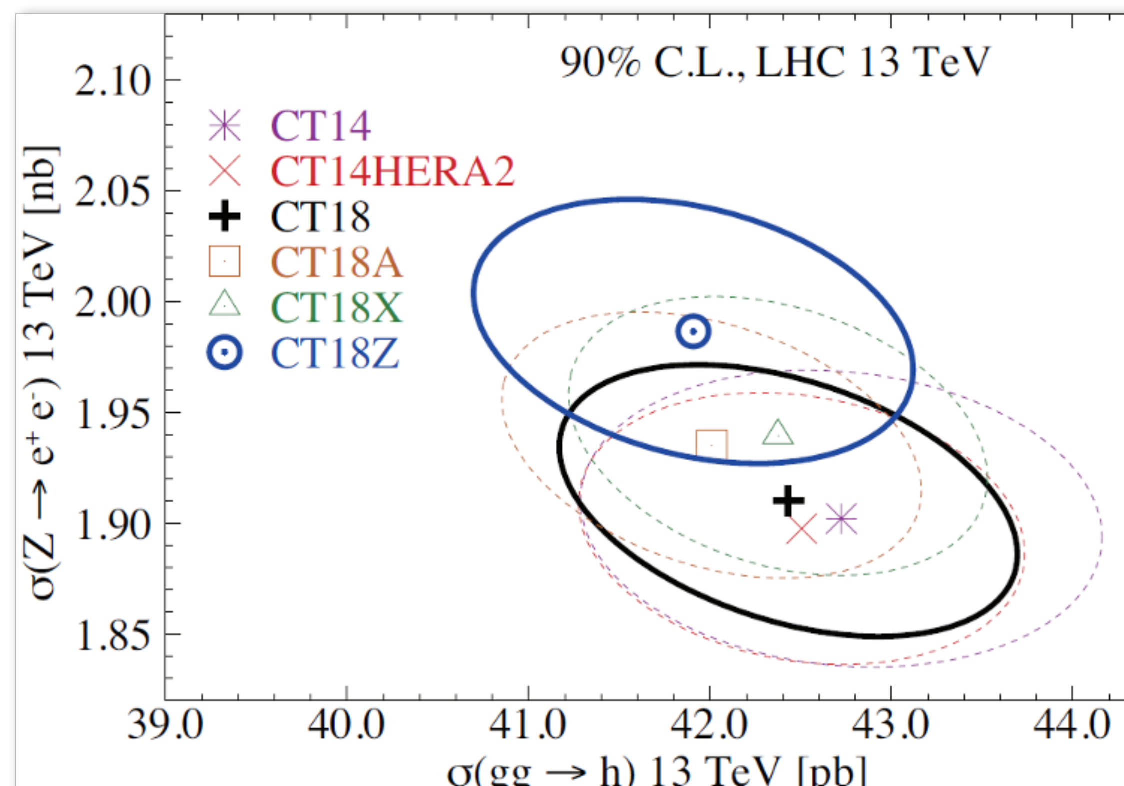
CT18 (nominal)

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)

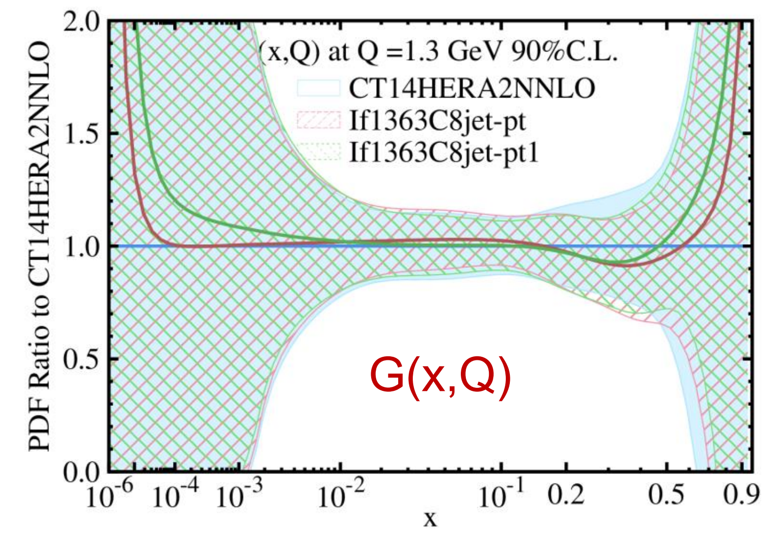
CT18 and *CT18Z* span the most different hypotheses, and the combination of the two represents the most complete uncertainty.



Theoretical uncertainties in CT18

Theory predictions and choice of scale

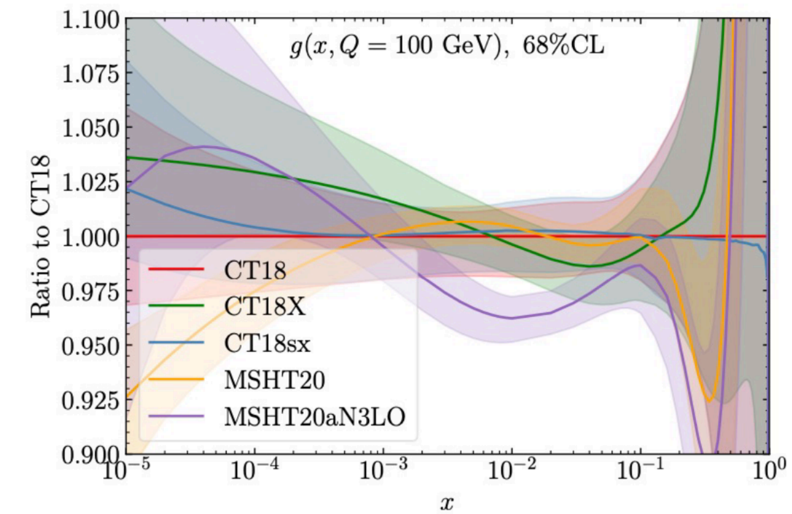
Choice of scale for inclusive jet data leads to a different gluon PDF yet contained in the CT uncertainty.
Resilience in global fit reflected through the tolerance.



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.

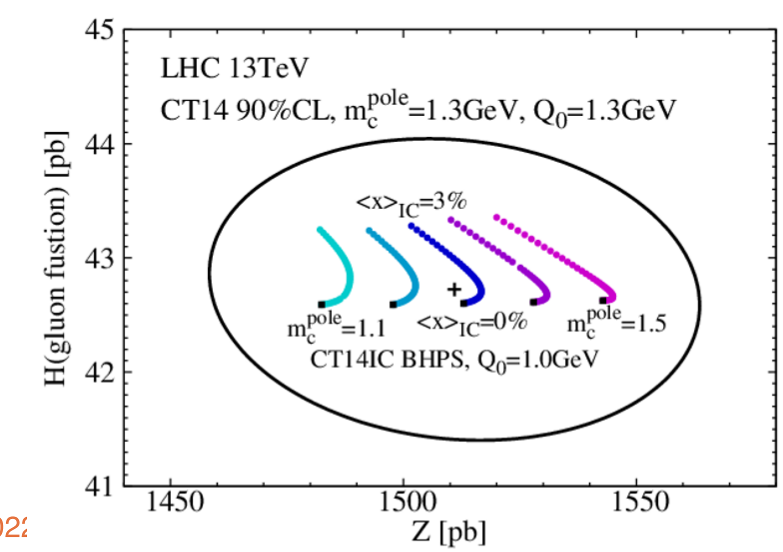
Small-x resummation enhances gluon PDF, similarly to N3LO (MSHT, see T. Cridge's talk)



Dependence on m_c —CT14 Intrinsic Charm

Study of dependence on the charm pole mass:

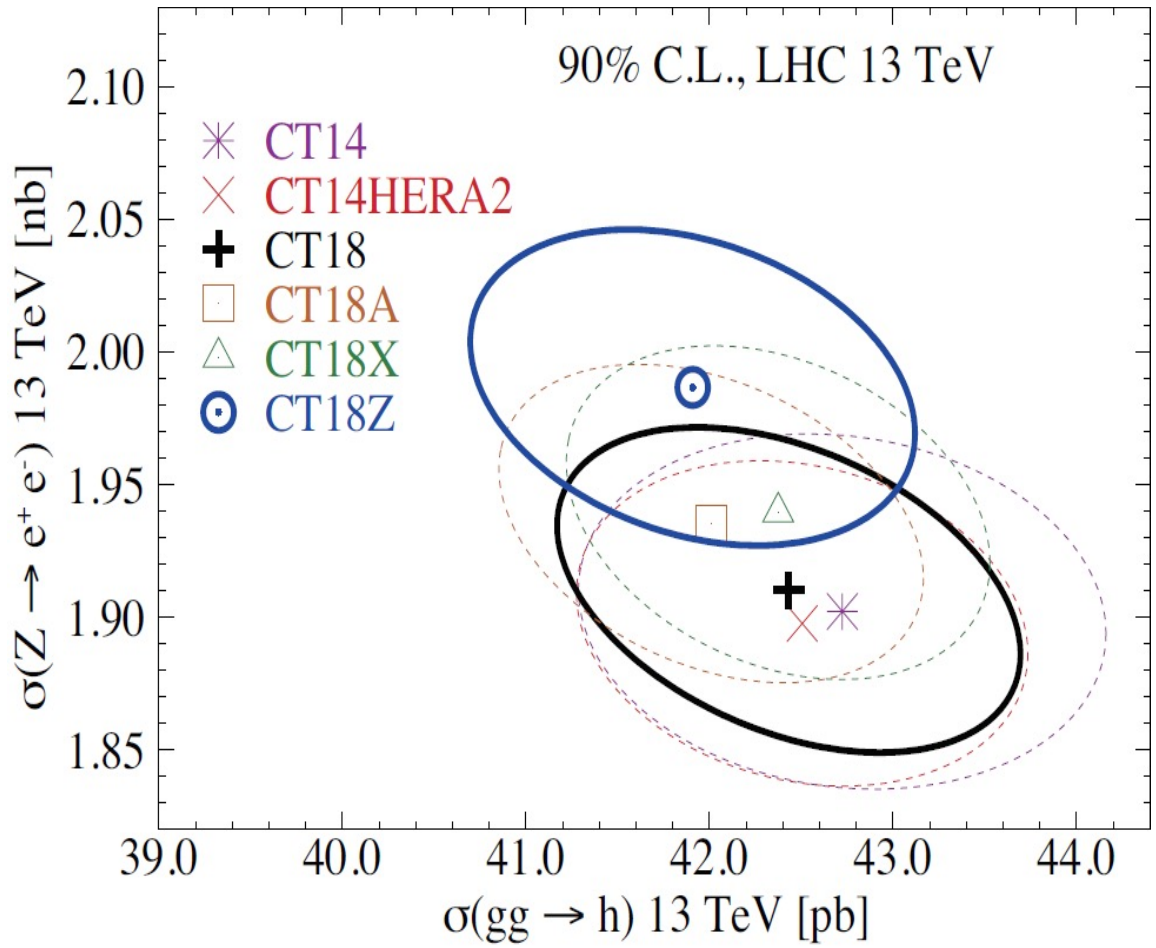
- CT14 Intrinsic Charm analysis [Hou et al., arXiv:1707.00657]
- CT18 Fitted Charm analysis (very soon)



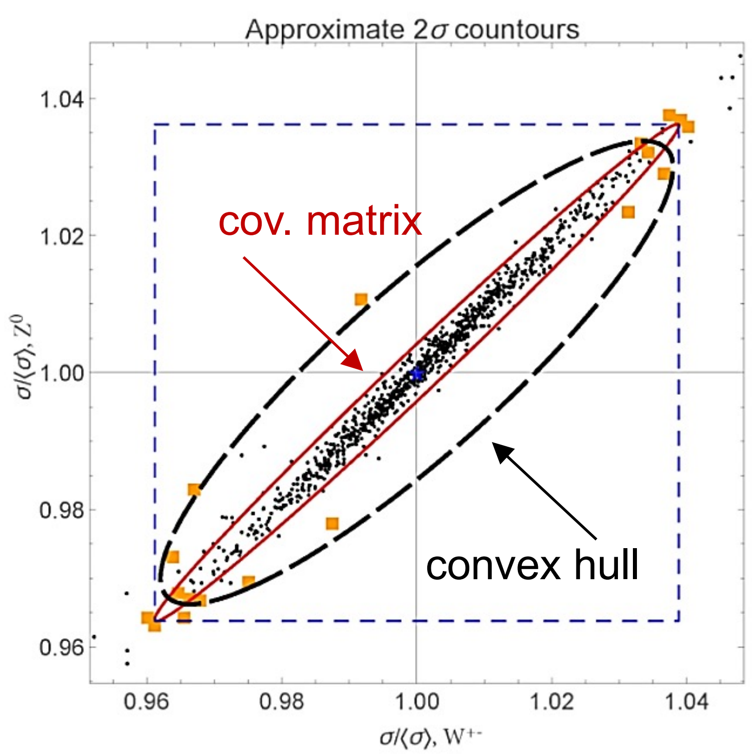
Toward **robust** PDF uncertainties

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

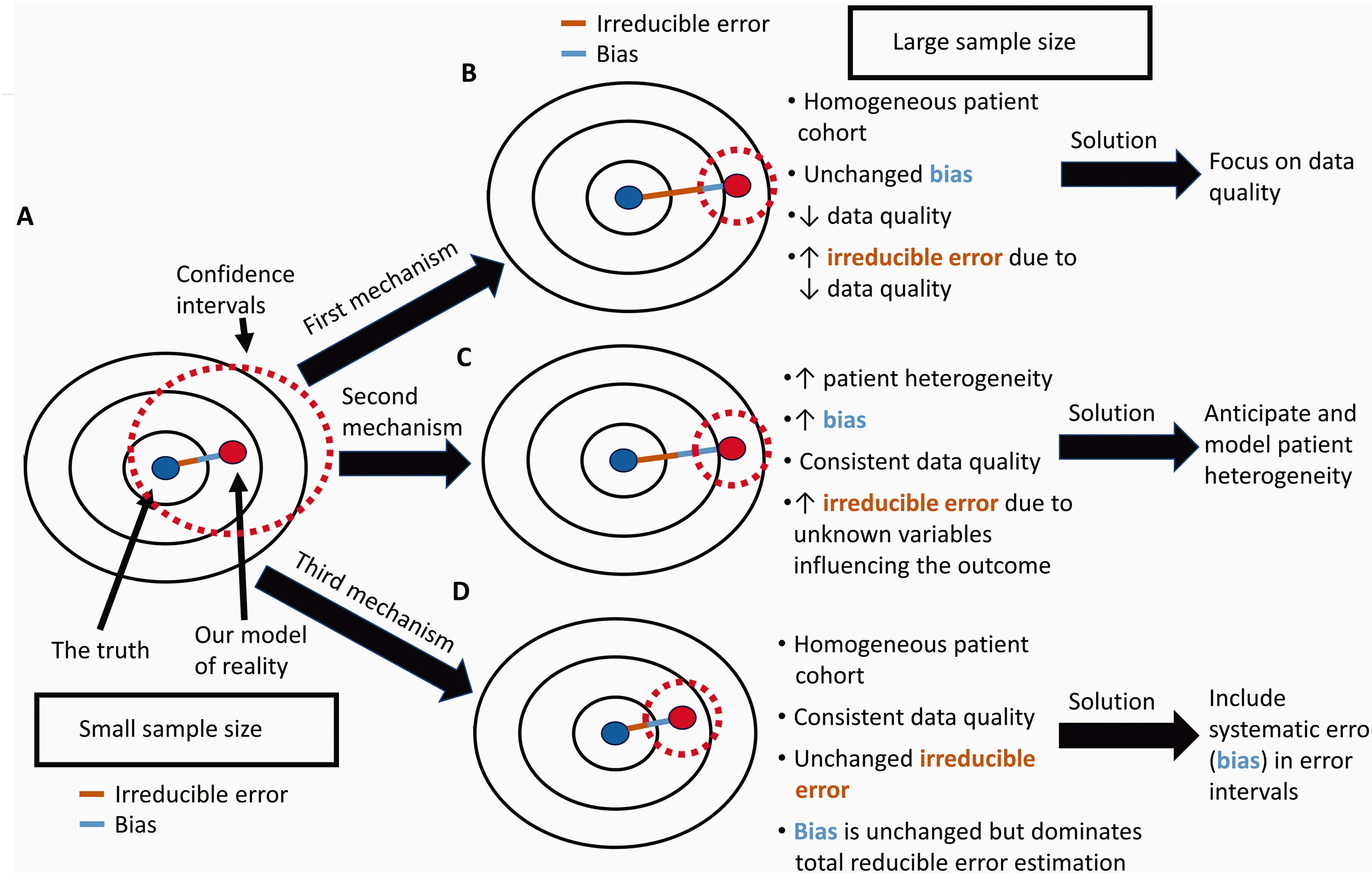
Setting for NNPDF4.0 code

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.

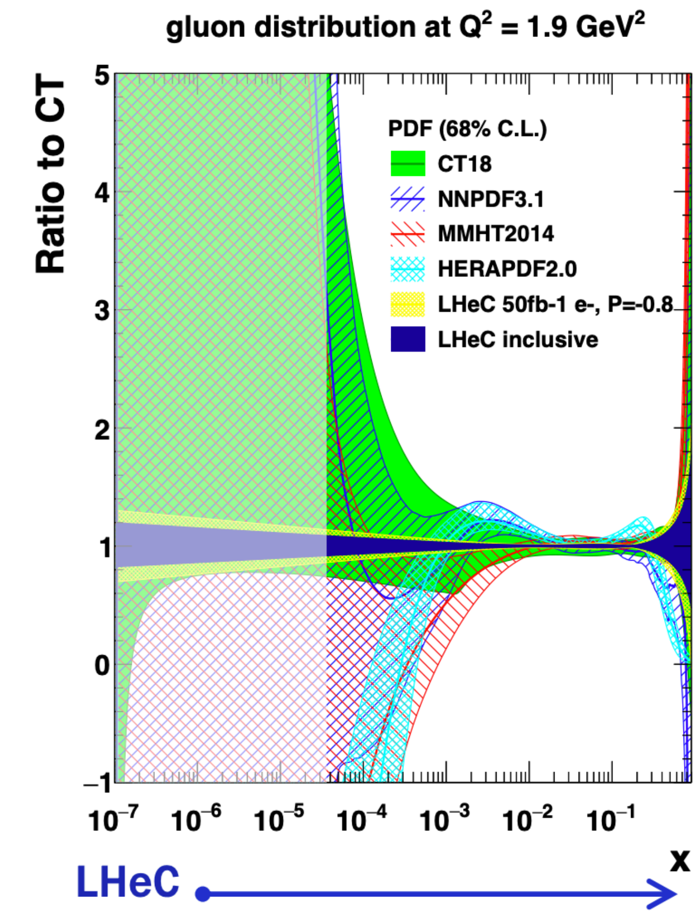

Irreducible error
Bias
Large sample size
B
Homogeneous patient cohort
Unchanged bias
↓ data quality
↑ irreducible error due to ↓ data quality
Solution
Focus on data quality
A
Confidence intervals
First mechanism
C
Second mechanism
↑ patient heterogeneity
↑ bias
Consistent data quality
↑ irreducible error due to unknown variables influencing the outcome
Solution
Anticipate and model patient heterogeneity
Third mechanism
D
The truth
Our model of reality
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
Small sample size
Irreducible error
Bias

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,…

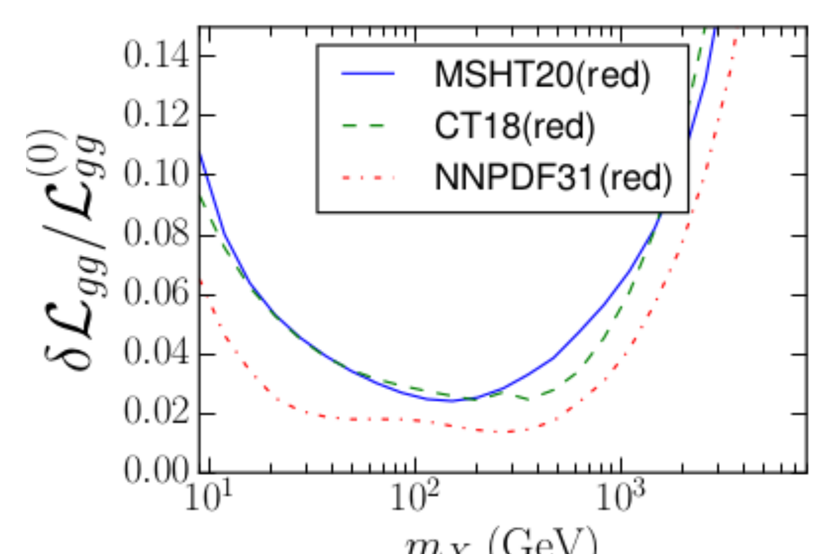


Plot from C. Gwenlan ICHEP 2020

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

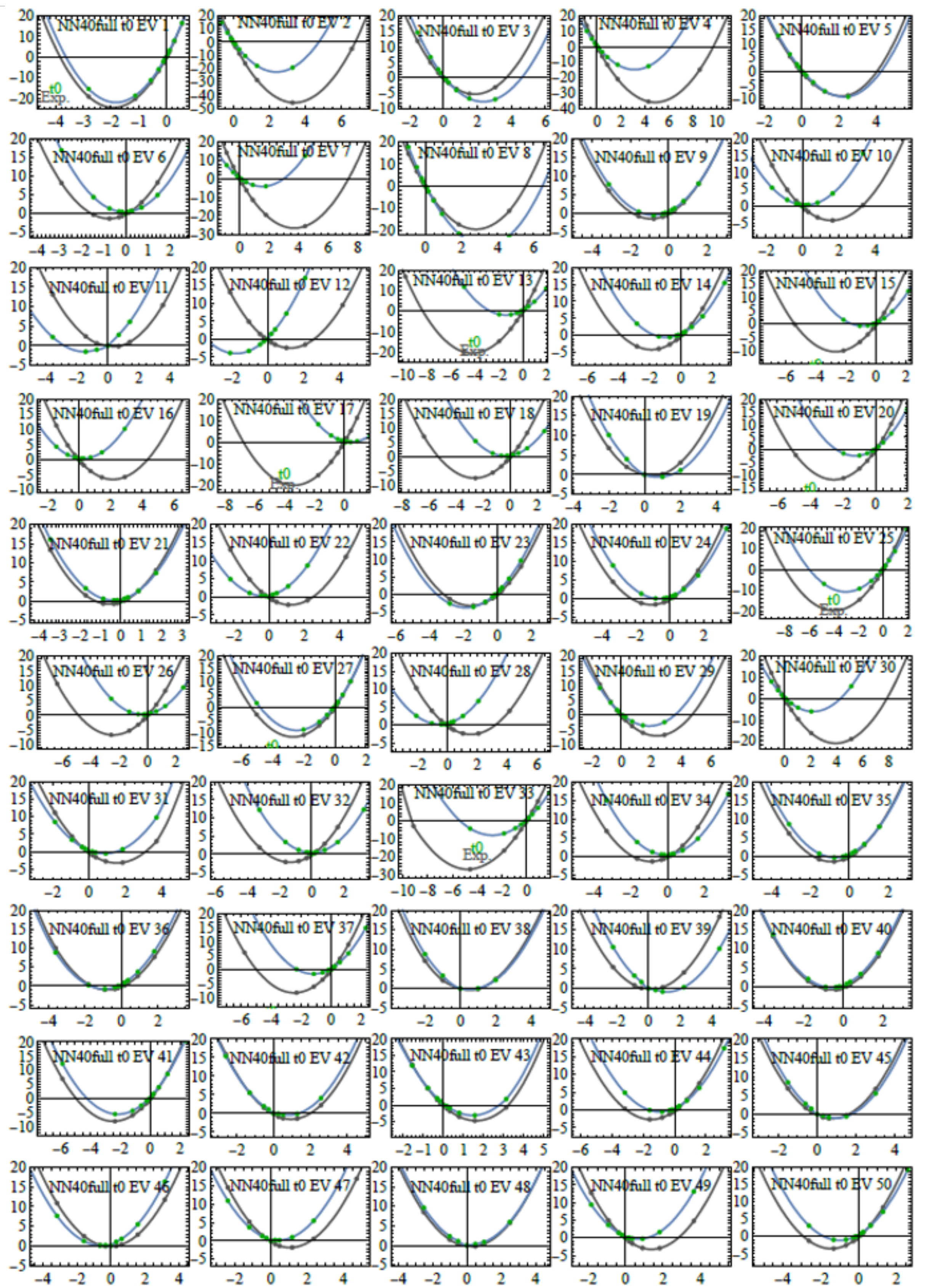
PDF4LHC21 benchmarking exercise:
comparison of uncertainties for same sets of data and QCD settings.

PDF4LHC21 [2203.05506]

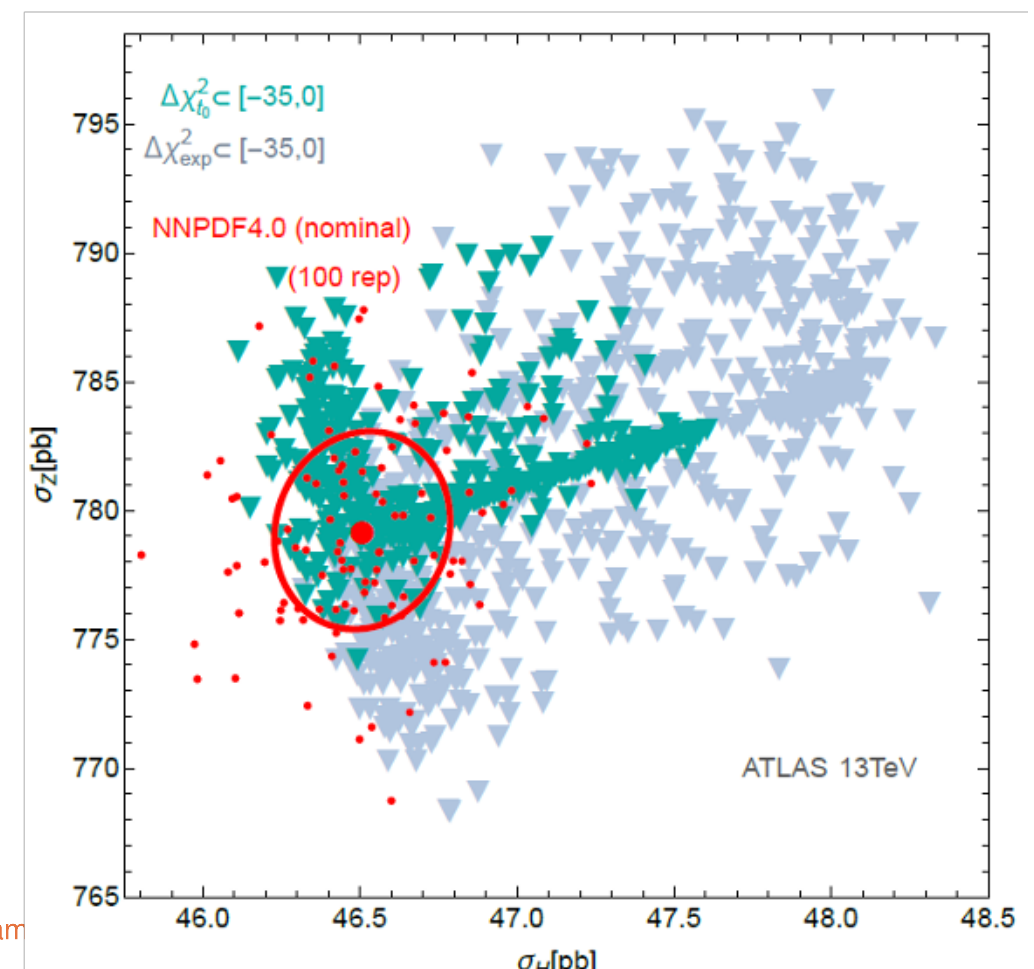


The uncertainties for CT18, MSHT20 and NNPDF3.1 reduced sets are still different. Key role played by methodology.

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



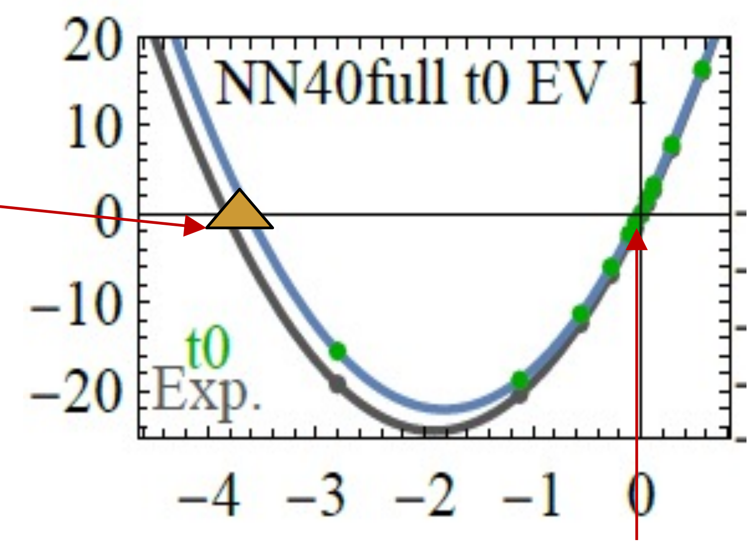
Sam



Hopscotch NN4.0 replicas

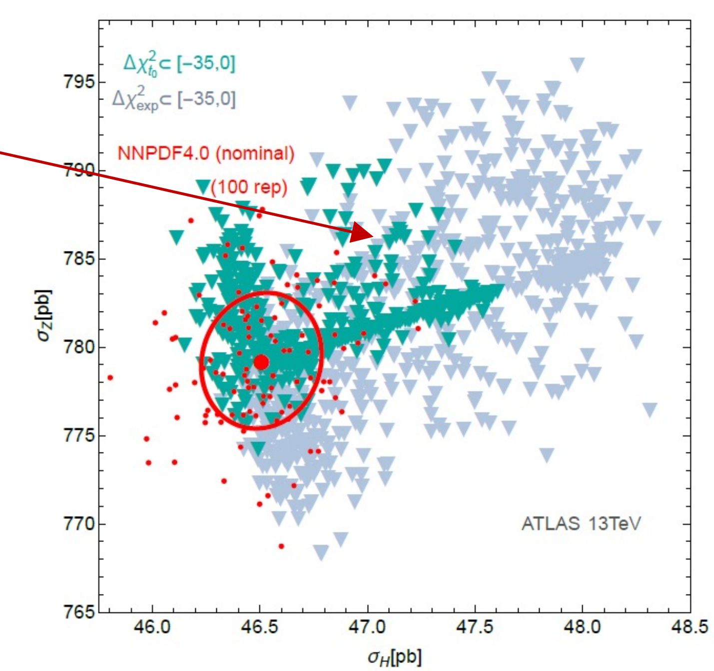
LHAPDF6 grids available at https://ct.hepforge.org/PDFs/2022hopscotch/

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



NN replica 0

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



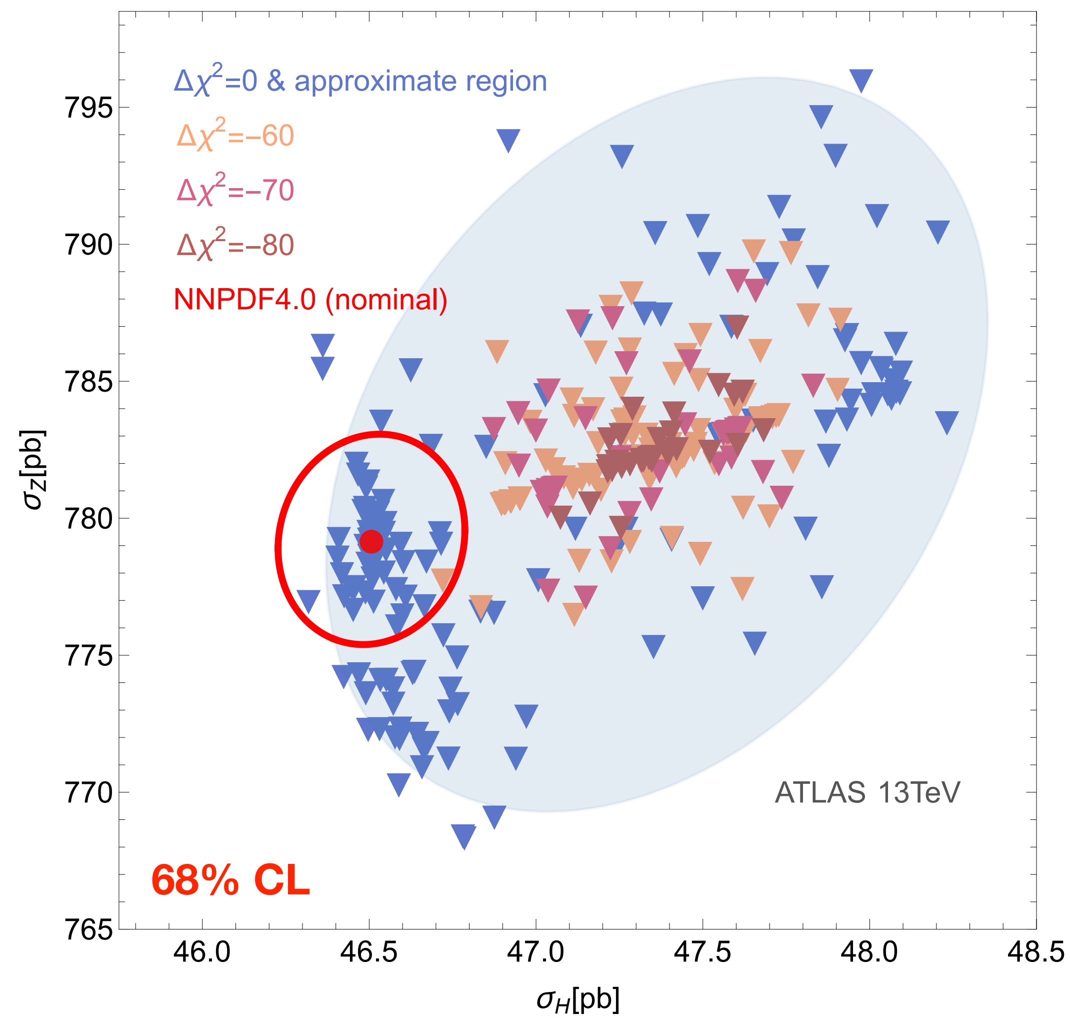
For $\chi^2_{t_0}$ and χ^2_{exp} definitions in the NNPDF4.0 code

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

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 $\Delta\chi^2_{exp}$.



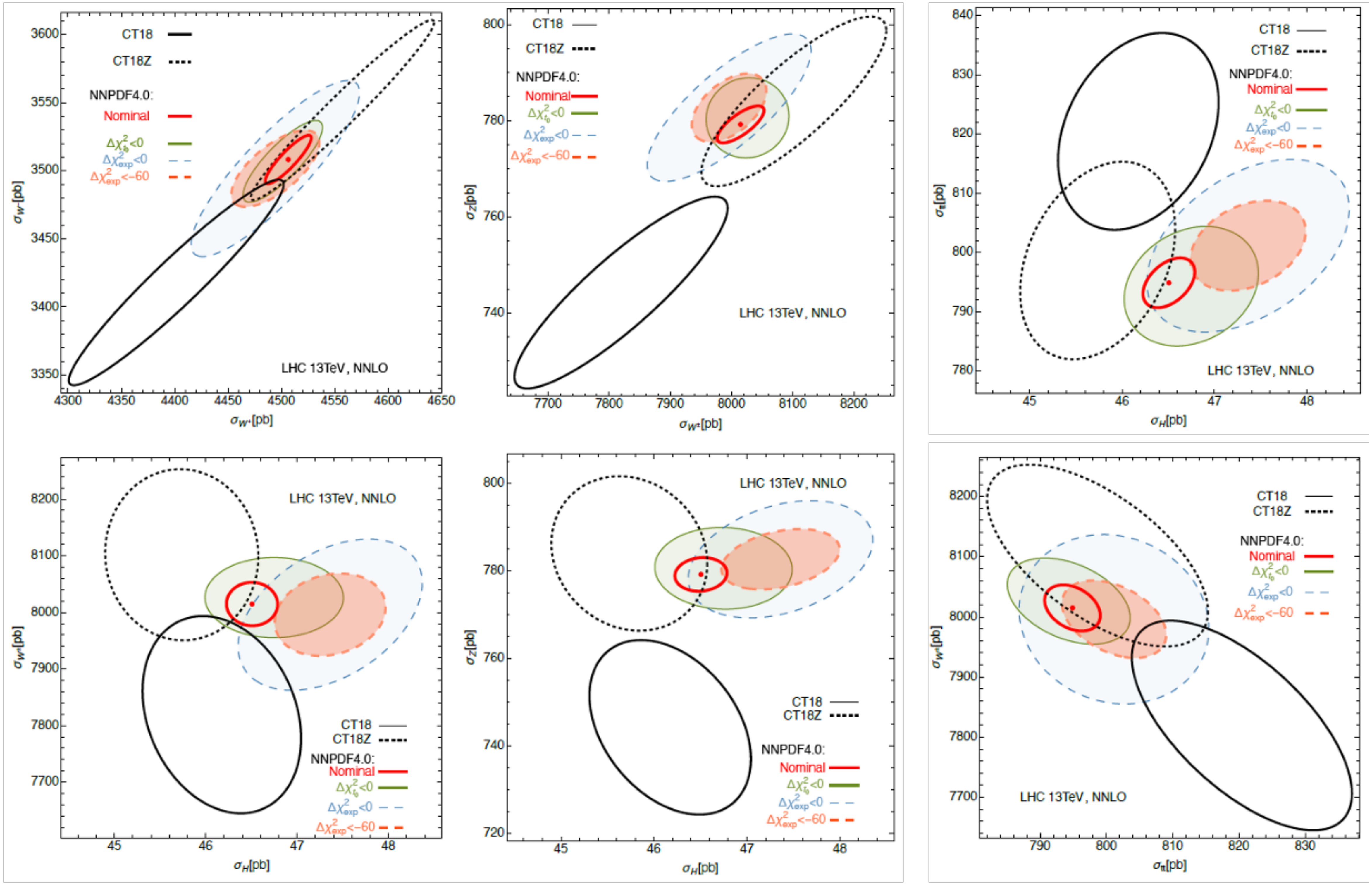
Each of the $\Delta\chi^2 = 0 \pm 3$ replicas is an acceptable PDF set from the NNPDF4.0 fit.

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