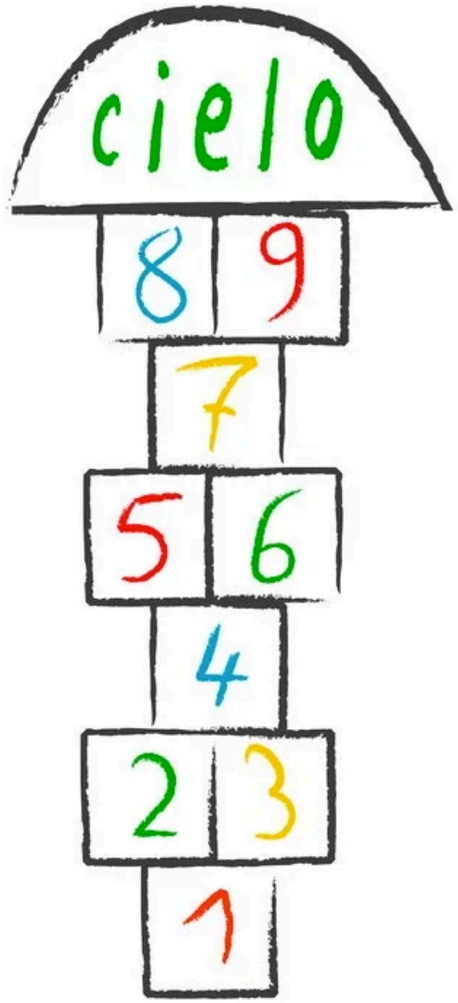
CT on uncertainties: Parton distributions need representative sampling

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



PDF4LHC meeting — November 2022

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

Our messages in [2205.10444]

Goodness-of-fit criteria come to the forefront as the community increasingly looks toward precision analyses using the PDFs.

Sampling uncertainty. Outside of HEP/NP, there is significant interest in statistical problems that are similar to the PDF tolerance problem. These studies introduce a fundamental distinction between the fitting uncertainty and sampling uncertainty, often overlooked in the PDF fits.

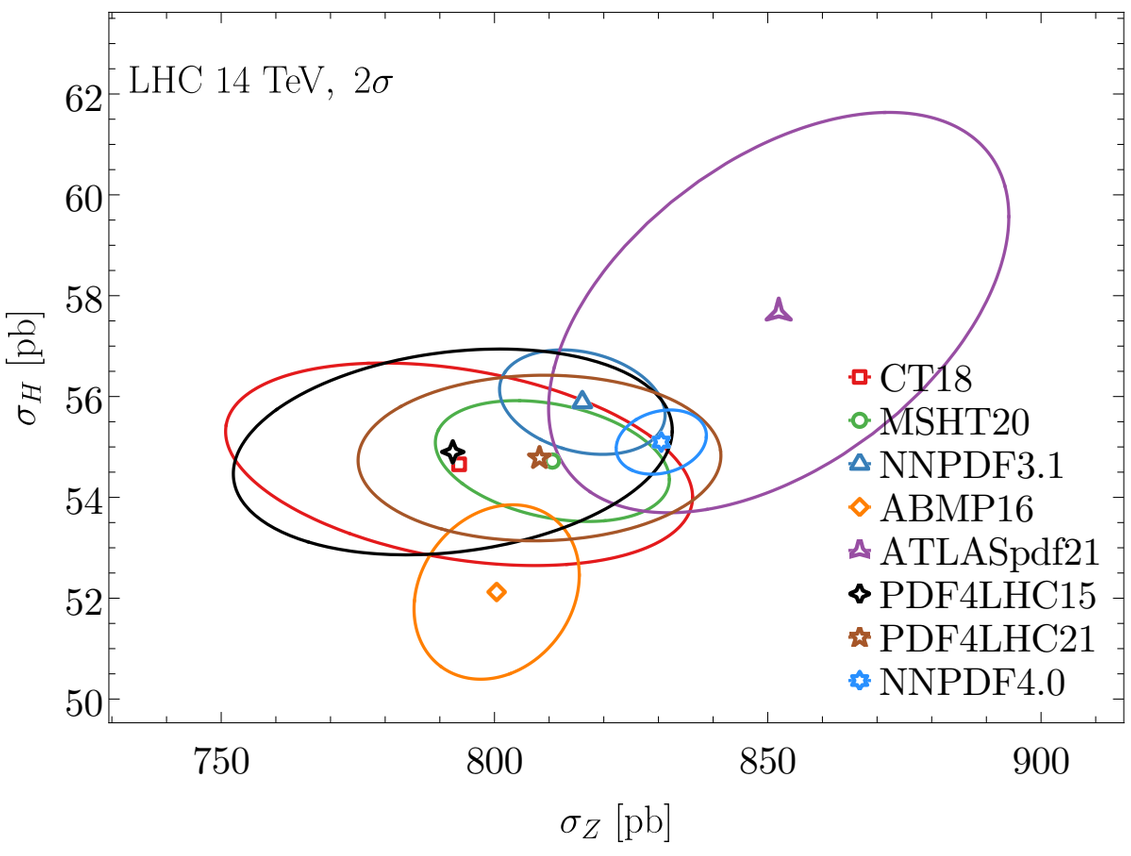
Effective parameter dimensions. Sampling uncertainty, while intractable in general, can be more reliably estimated for specific predictions in targeted low-dimensional searches like the "*hopscotch scan*" that we designed.

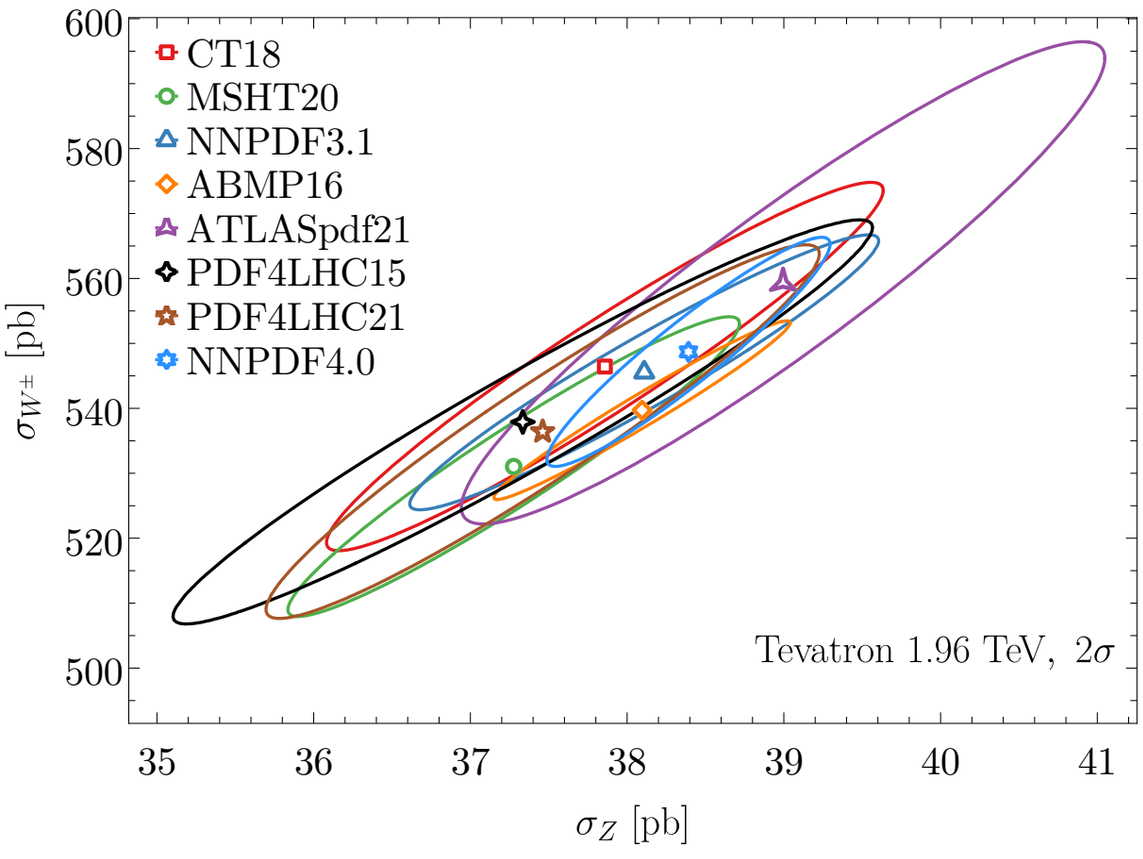
PDF tests outside of the fit. It is probably simpler to ascertain the quality of given PDFs than the process by which these PDFs were found.

How good is a PDF fit?

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

Precision PDFs (Snowmass 21 WP)
[2203.13923]





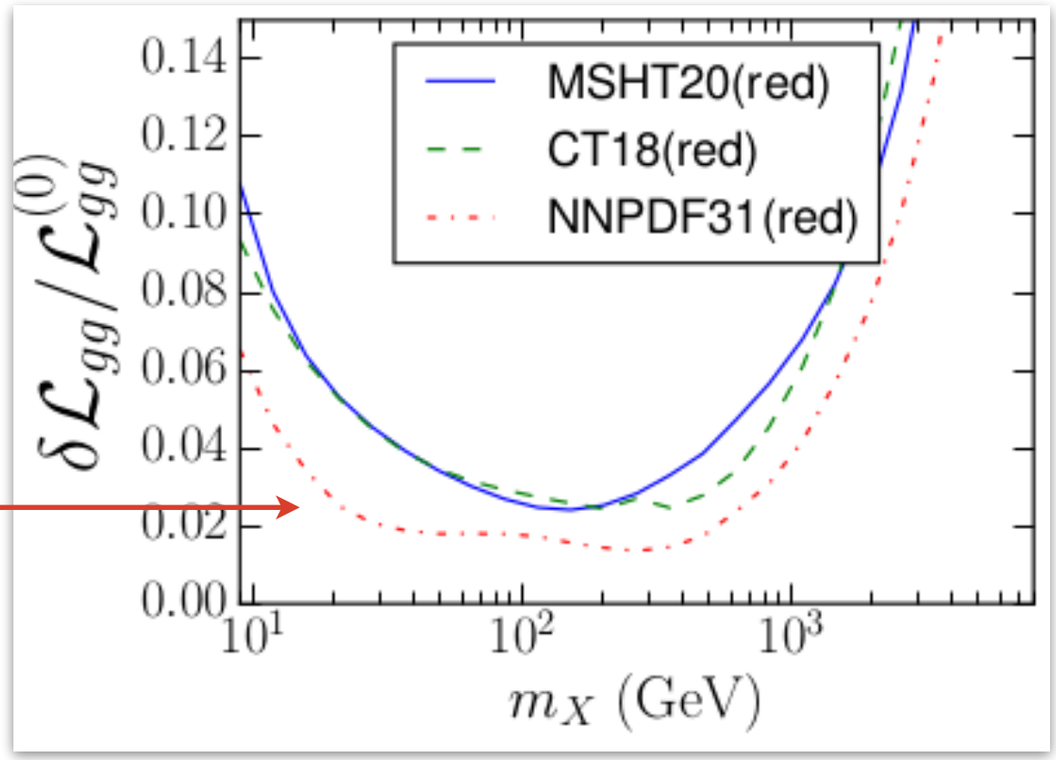
Uncertainty prescription (tolerance) in different groups

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.

NNPDF31(red) smaller by a factor of ~2



Monte-Carlo global analyses seem to lead to smaller uncertainties *wrt* Hessian fits.
Trend sustained by NNPDF4.0 set.

PDF4LHC21 [J.Phys.G 49]

The tolerance puzzle and the big-data paradox

Outside of HEP, there is significant interest in statistical problems that are similar to the PDF tolerance problem. These studies introduce a fundamental distinction between the **fitting uncertainty** and **sampling uncertainty**, often overlooked in the PDF fits.

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

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

https://doi.org/10.1038/s41586-021-04198-4
Received: 19 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

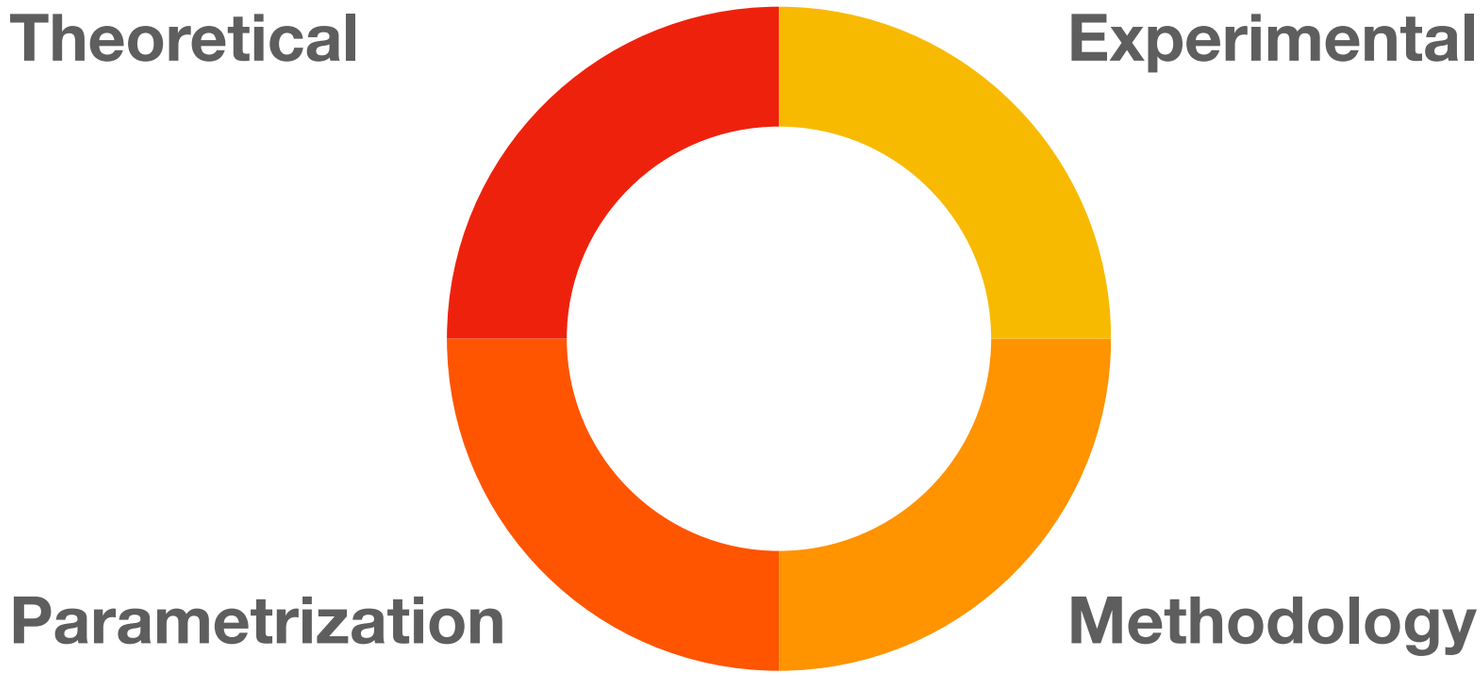
Arnald Puy[1,2,3]*, Pierfrancesco Beneventano[4], Simon A. Levin[2], Samuele Lo Piano[5], Tommaso Portaluri[6], Andrea Saltelli[3,7]

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

A new avenue to understand PDF tolerance



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

Goodness-of-fit applies to an individual best fit.

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

Sampling accuracy applies either to the tolerance or the number of error sets in a PDF ensemble.

This talk.

Sampling accuracy vs. fitting accuracy

Differences of tolerance prescriptions are in part due to sampling conventions.
There is evidence that PDF fitting groups doing the same analysis arrive at different conclusions because of their tolerance criteria.

PDF sampling takes place over experimental data sets, parametrization forms, hyperparameters, settings of fits, model approximations.

Biases in sampling are particularly risky in large-scale analyses.

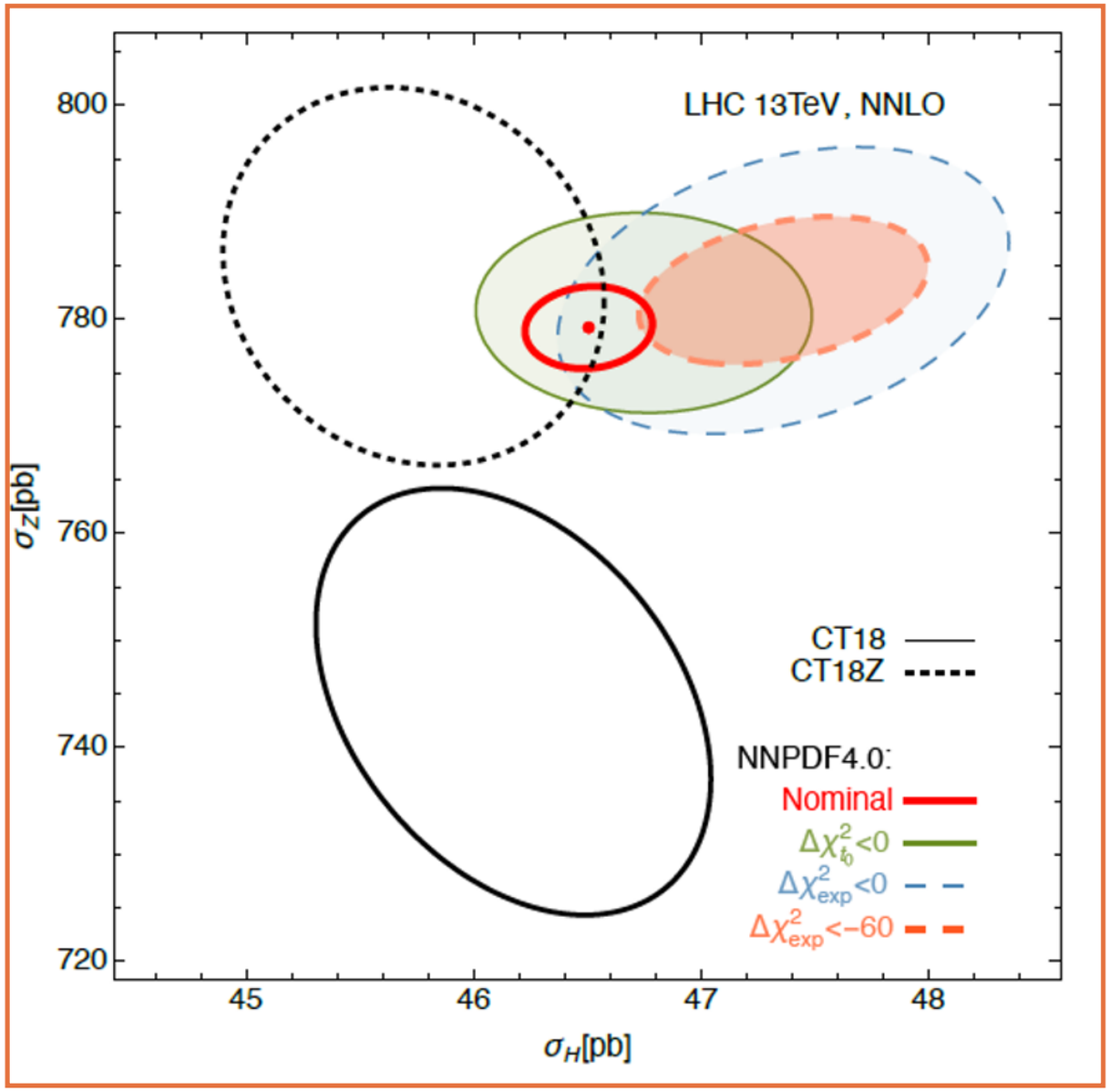
PDFs need representative sampling

[2205.10444]

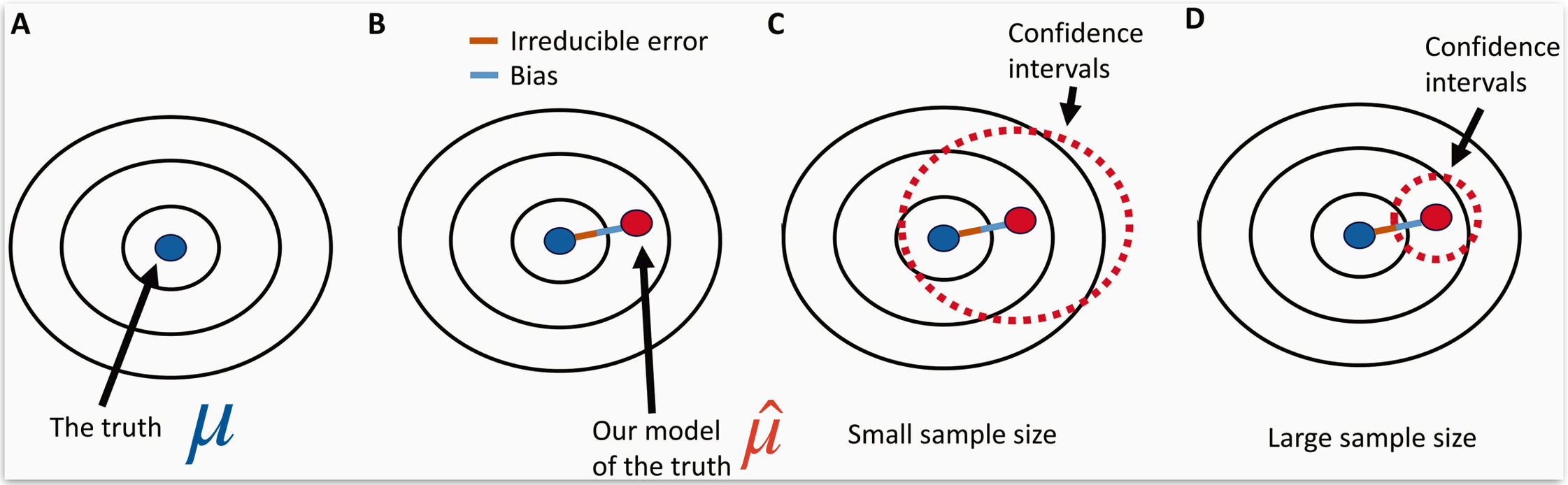
In this talk, sampling uncertainty will be estimated for specific predictions in targeted low-dimensional searches, called the **hopscotch scan,** that we designed.

Filled 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 to be described next.



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*, Irreducible error, Bias.

Illustration from:
Pavlos Msaouel (2022)
Cancer Investigation, 40:7, 567-576

Trio identity

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

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

depends on the sampling algorithm

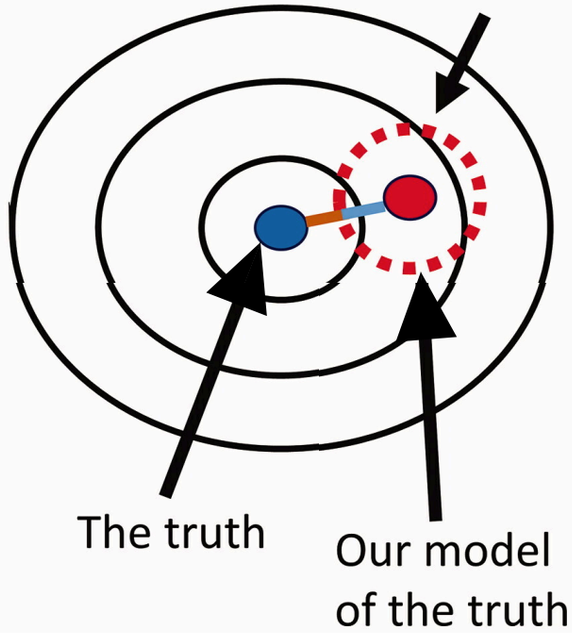
can tend to $\left(\sqrt{n}\right)^{-1}$ for random sampling

Irreducible error
Bias

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

Confidence intervals

Large sample size

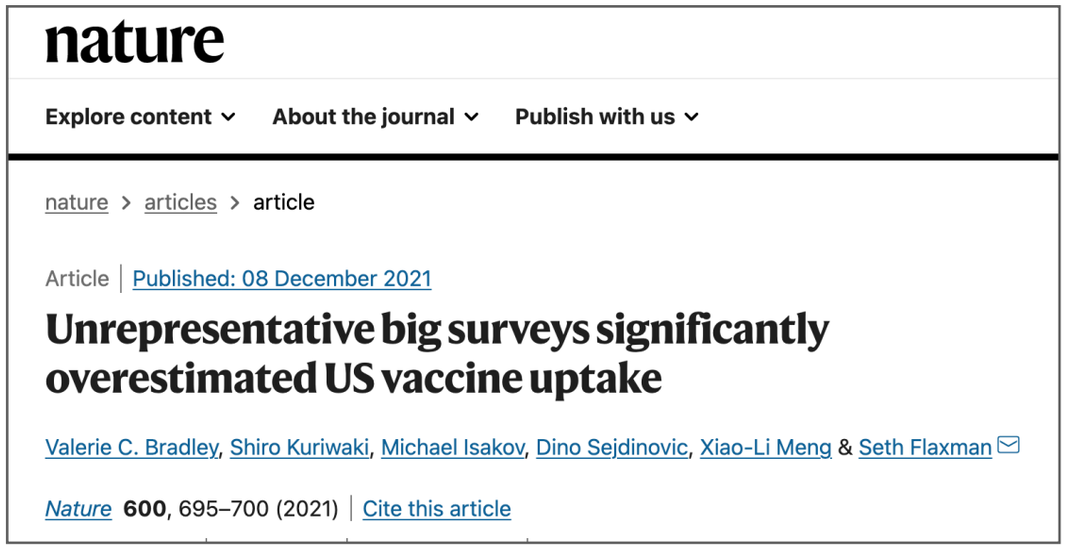


For a sample of n items from the population of size N, we can consider an array built by the random spanning of the binary responses of the $N - n$ (0) and n (1) items, so that

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

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 ⌄

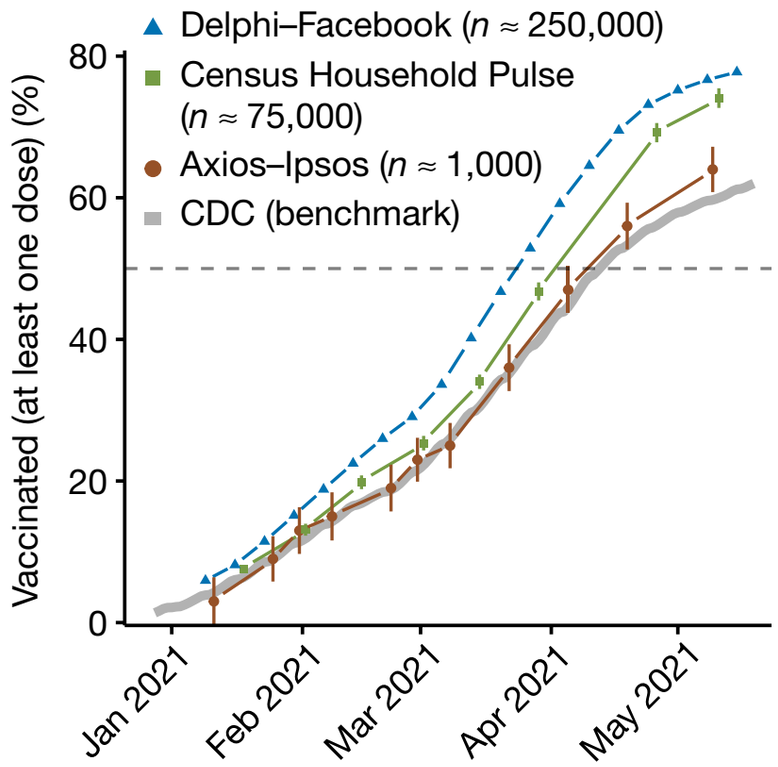
nature › articles › article

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 in PDF global analyses—I

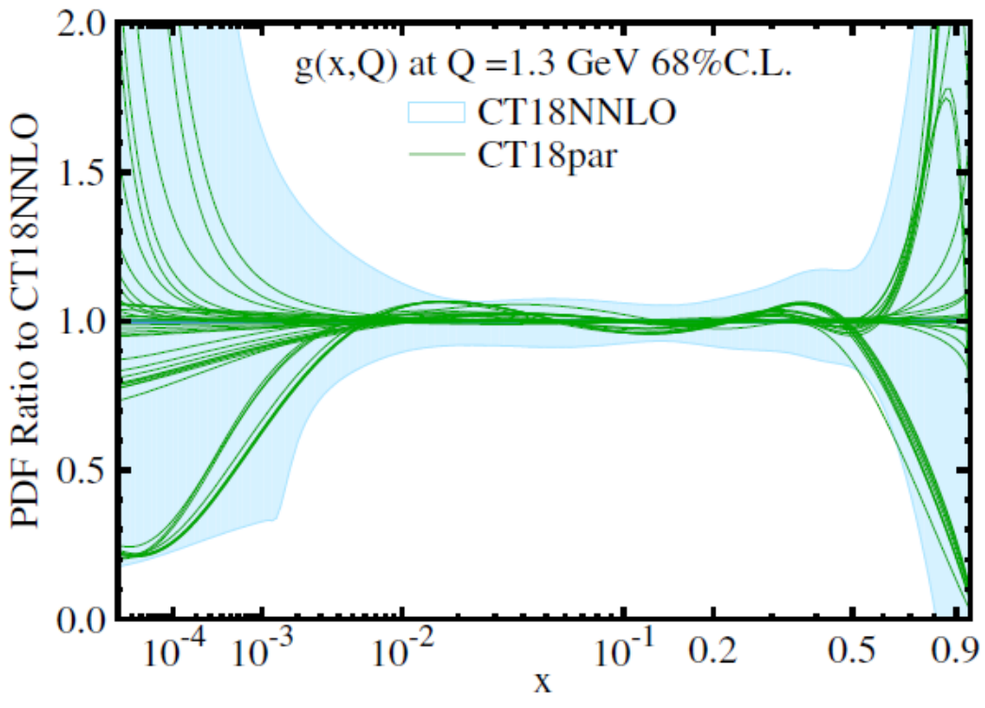
How do we know the “data+sampling defect=confounding correlation” of our analysis?

Hessian-based analysis:
objective function includes penalties, establishing the **tolerance criteria**.

Size of uncertainties reflect a series of confounding sources —selection of fitted experiments, treatment of correlated systematic errors, functional forms of PDFs, …

Verification that proper spanning of parameter space is compatible with total uncertainties (*a posteriori*). >300 functional forms are tested in CT18.

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



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

Sampling bias in PDF global analyses—II

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

Monte Carlo-based analysis:

optimization implies selection of hyperparameters

The usage of Neural Networks had as primary goal eliminating the biases associated with the choice of a specific functional form.

However, there are still many choices associated with the optimization:

- Number and with of the layers
- Activation functions and initialization
- Optimization algorithm (and associated parameters)
- Training length, stopping patience, etc.
- Strength of lagrange multipliers (positivity, integrability)

Collectively called “hyperparameters”, usually selected manually.

Do we understand sampling for QCD global analyses?

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

In general, an intractable problem.

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

Do we understand sampling for QCD global analyses?

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

In general, an intractable problem.

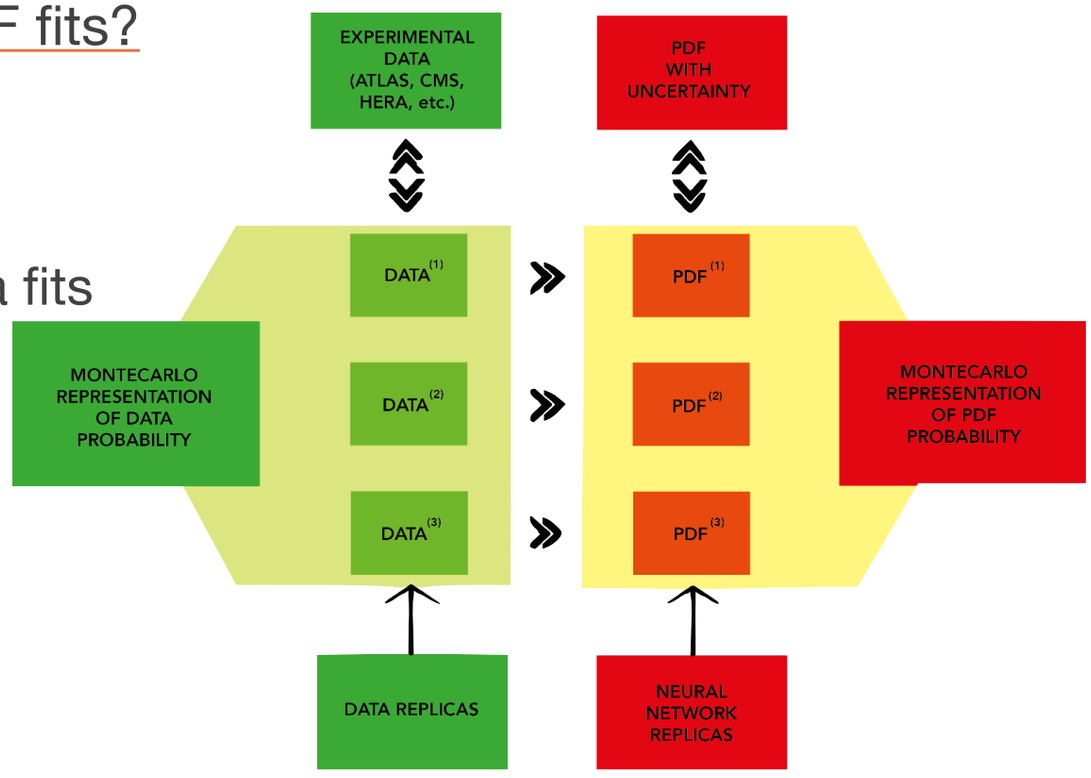
[Hickernell, MCQMC 2016, 1702.01487]
[Sloan, Wo´zniakowski, 1997]

1. Justification for tolerance criteria for Hessian-based PDF fits
2. How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling, as defined by NNPDF

- =bootstrap/resampling of random fluctuations in data
- expectations are then unweighted averages over replica fits

Such sampling does not include sampling over hyperparameters and priors.



Do we understand sampling for QCD global analyses?

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

In general, an intractable problem.

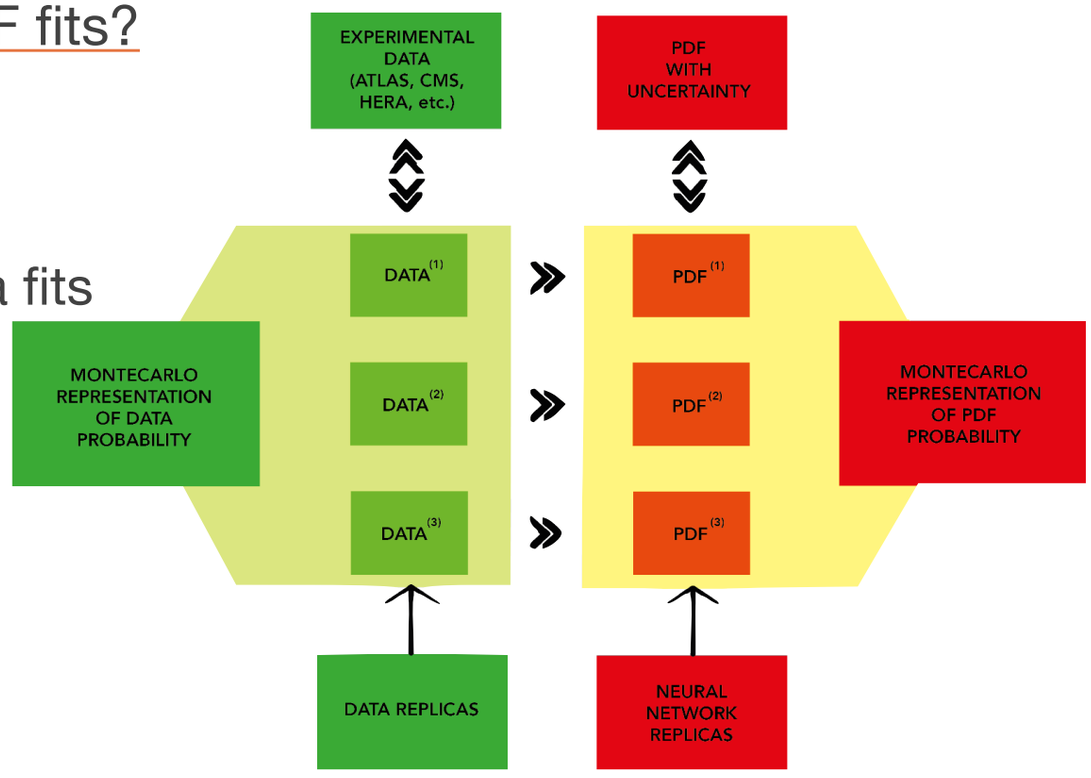
[Hickernell, MCQMC 2016, 1702.01487]
[Sloan, Wo´zniakowski, 1997]

1. Justification for tolerance criteria for Hessian-based PDF fits
2. How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling, as defined by NNPDF

- =bootstrap/resampling of random fluctuations in data
- expectations are then unweighted averages over replica fits

Such sampling does not include sampling over hyperparameters and priors.



Do we understand sampling for QCD global analyses?

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

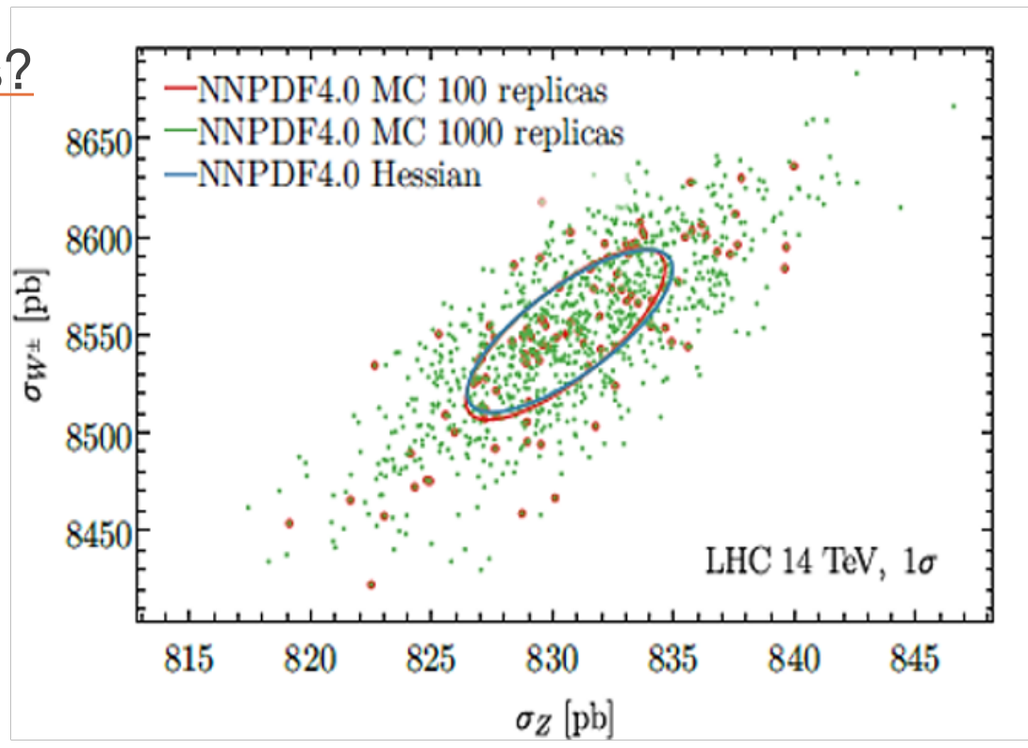
In general, an intractable problem.

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

1. Justification for tolerance criteria for Hessian-based PDF fits
2. How is sampling achieved in Monte Carlo-based PDF fits?

Importance sampling, as defined by NNPDF

- =bootstrap/resampling of random fluctuations in data
- expectations are then unweighted averages over replica fits



Such sampling does not include sampling over hyperparameters and priors.

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

Effective dimensions

Algorithm for observable-oriented verification of representative uncertainty

“Parton distributions need a representative sampling”

[2205.10444]

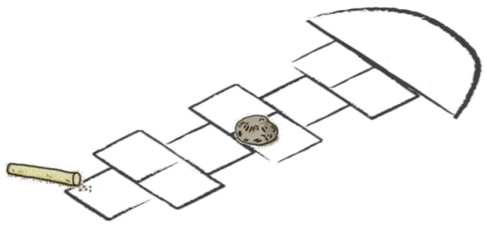
We determine dimensions of the problem from specific QCD observables: only few **effective large dimensions** contribute the bulk of the uncertainty.

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

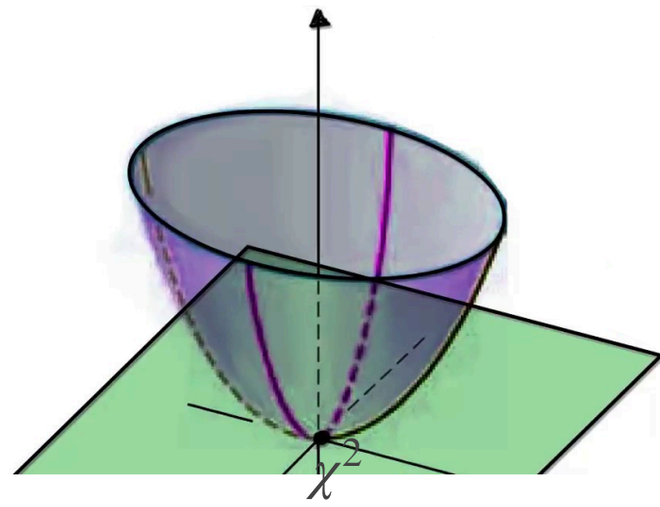
Using NNPDF4.0 public code, we then employ: n = the number of replicas/EV directions/…

1. Basis coordinates in the PDF space — Hessian representation of MC replicas
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?

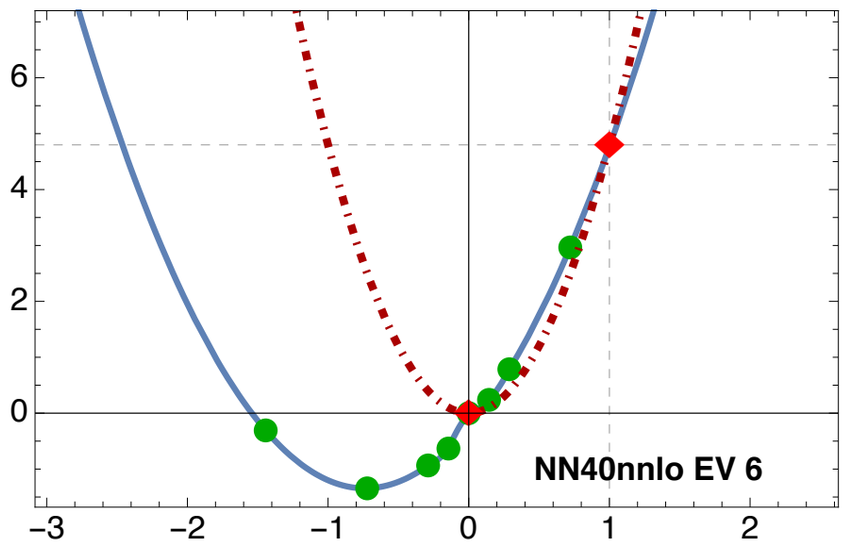


In the Hessian representation, the chi square behaves like a paraboloid of n_{param} dimensions, thus defining a global minimum.



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 EV direction for the NNPDF4.0 Hessian set, in blue. There are second EV sets with $\Delta\chi^2 = 0$, for all 50 EV directions.

Its shape indicates a larger paraboloid than the red curve provided by the NNPDF4.0 Hessian set.

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.

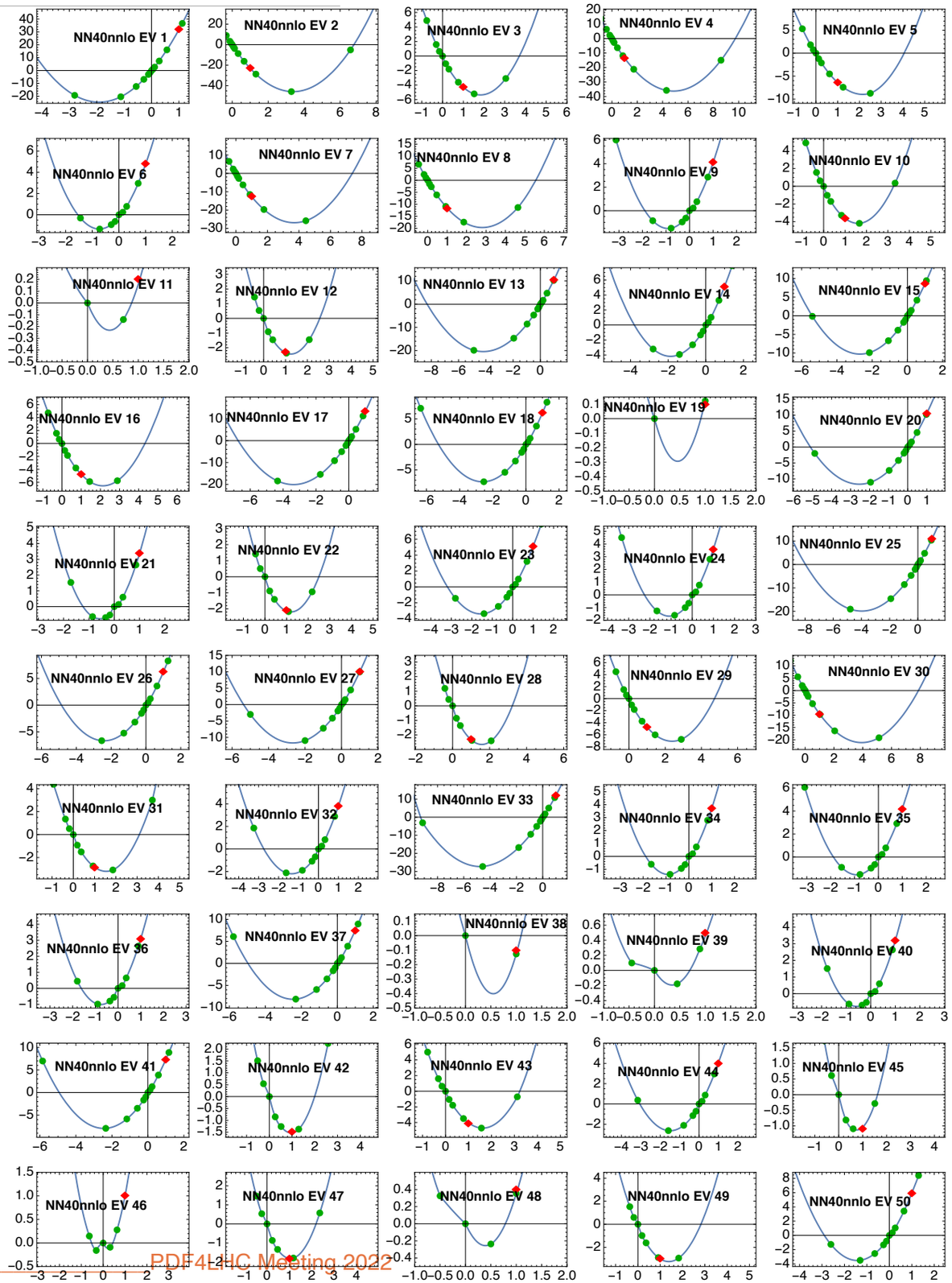
[NNPDF, 2109.02653]

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.

No fitting involved.



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.

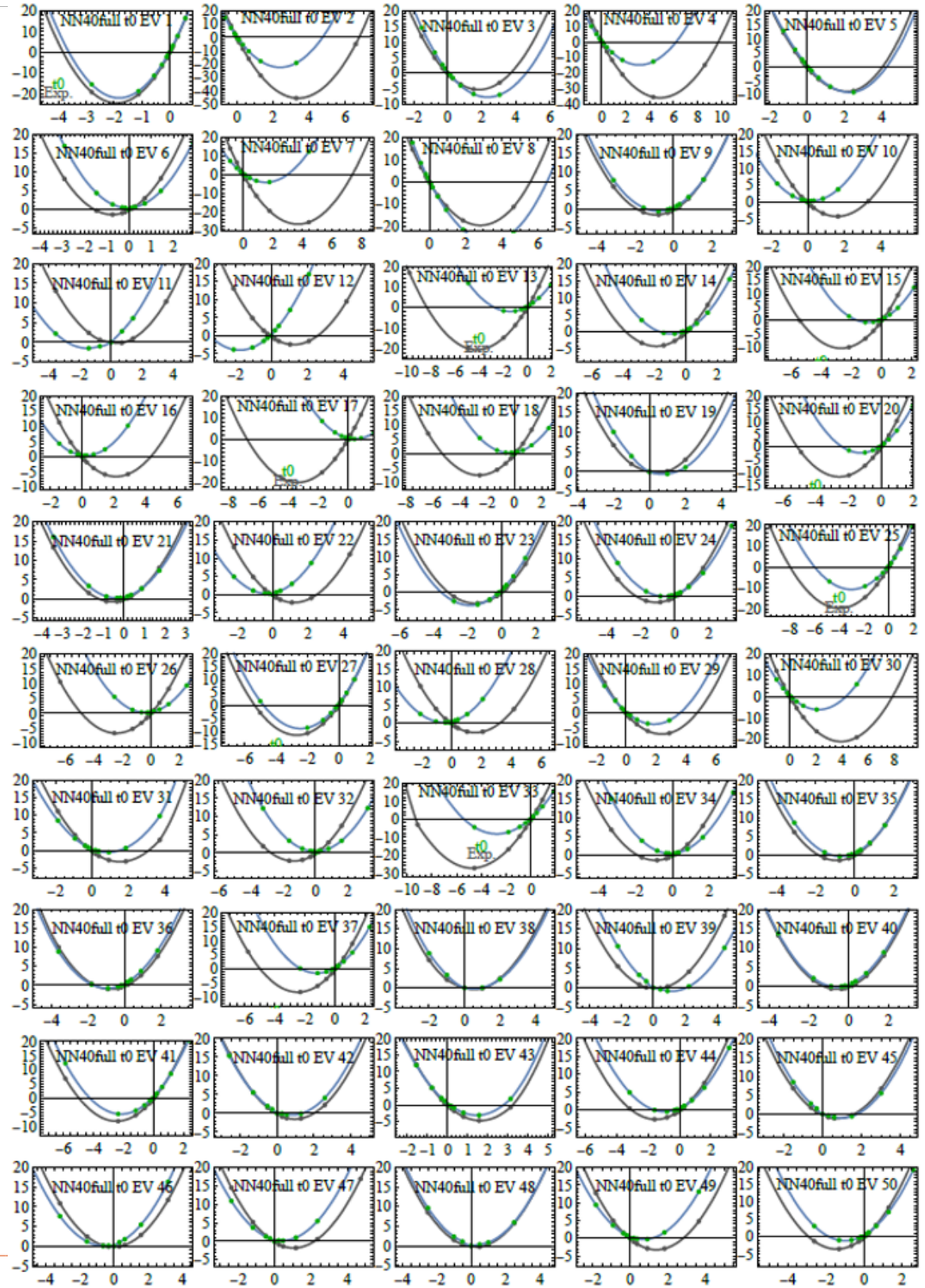
[NNPDF, 2109.02653]

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.

No fitting involved.

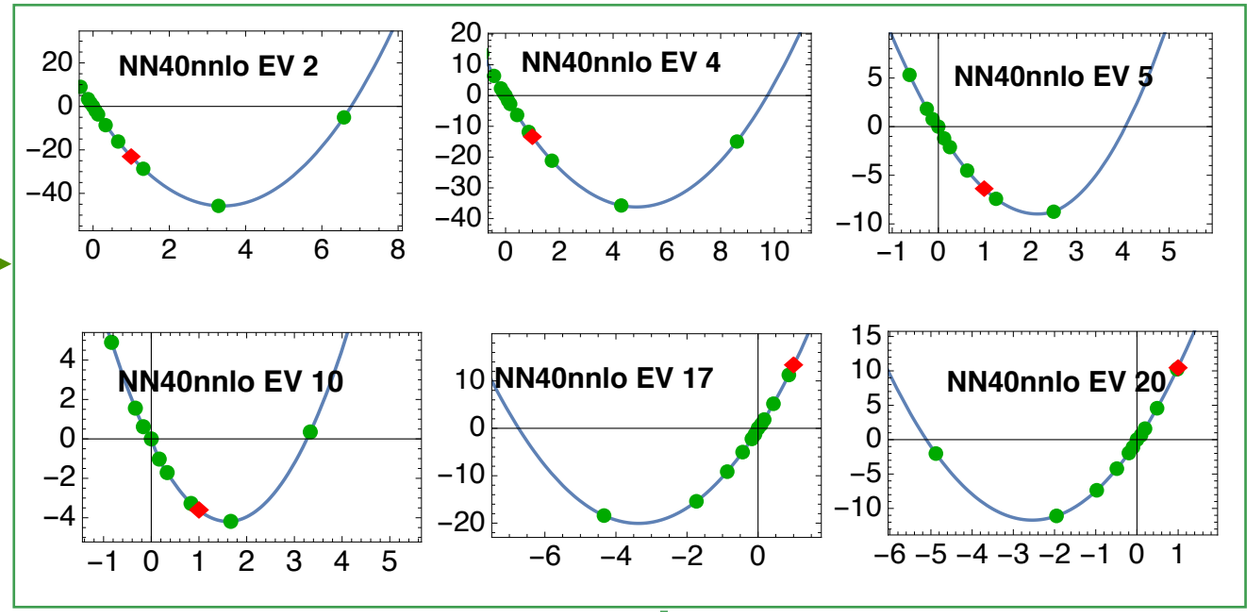


A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

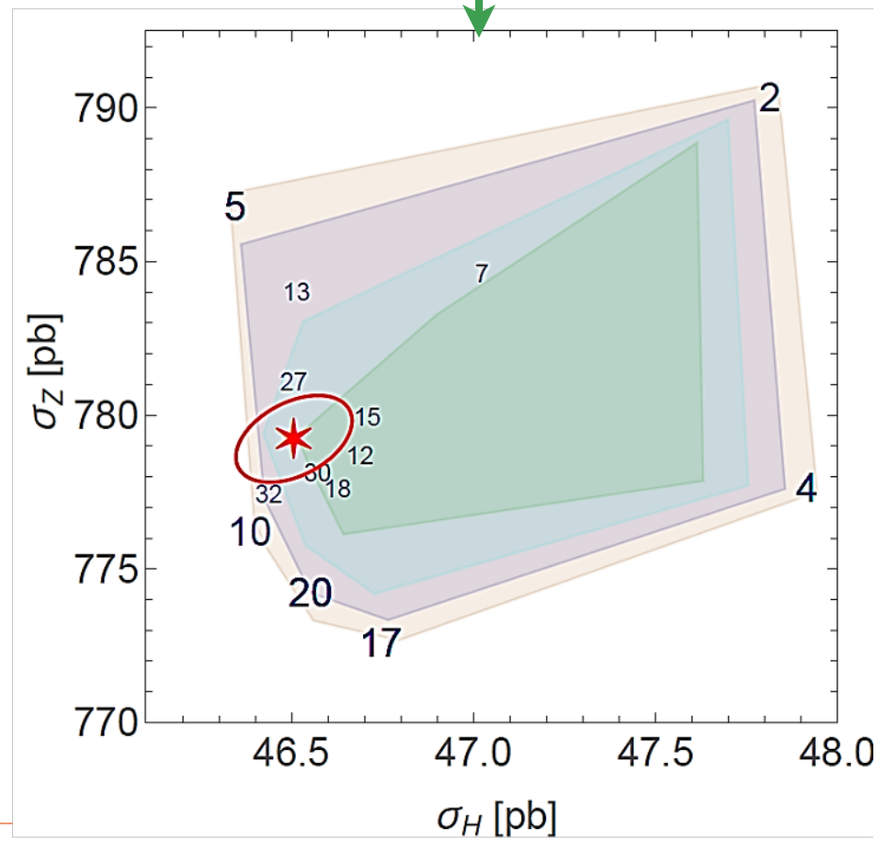
Step 3

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

Large EV directions are shared among various pairs of cross sections.



Construct the convex hulls 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

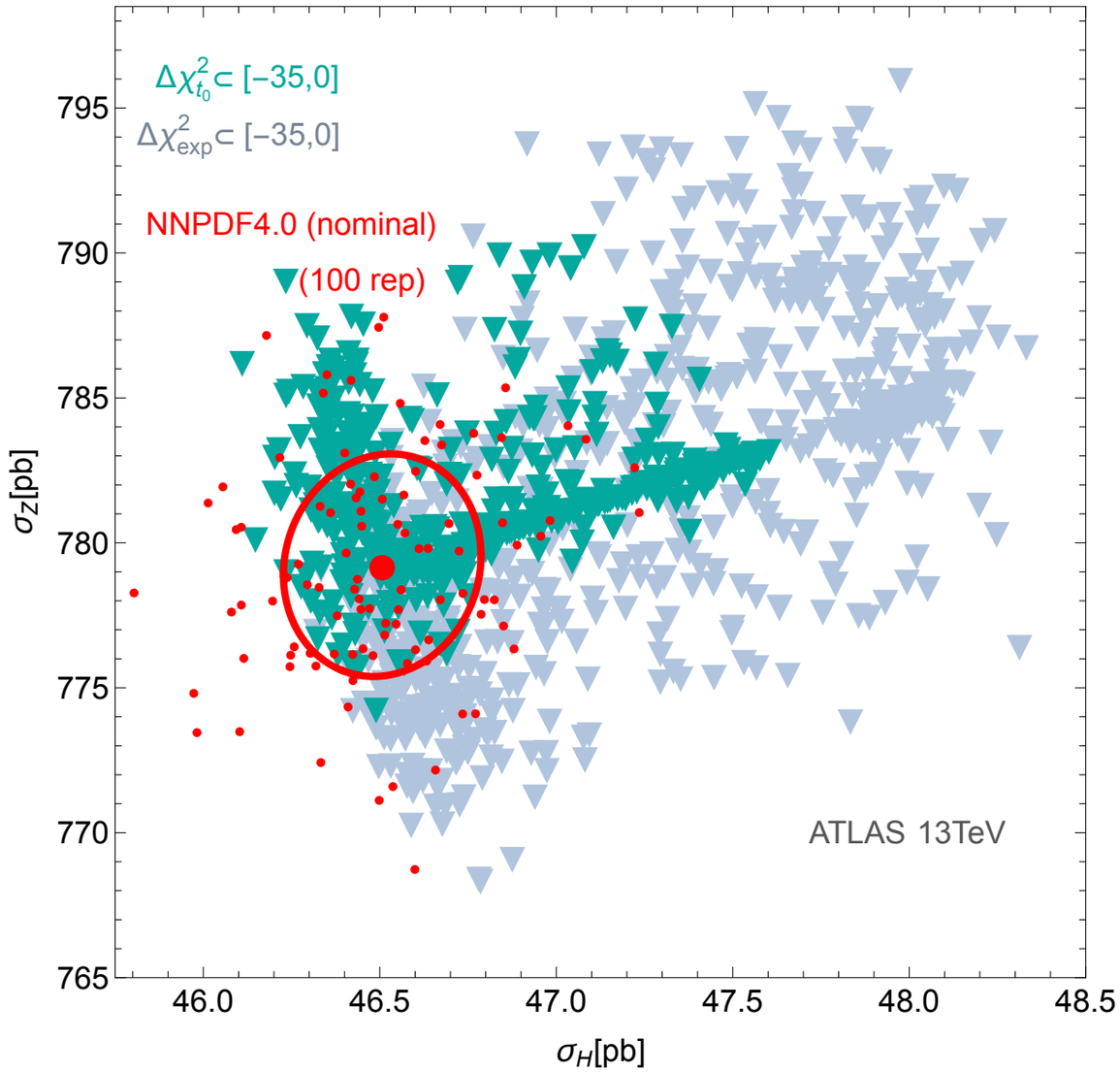
For each pair of cross sections, we generate 300 replicas by sampling uniformly along the “large” EV directions.

Sort the $n_{pairs} \times 300$ resulting replicas according to their $\Delta\chi^2$ *w.r.t.* to NN40 replica 0.

Hopscotch replicas are linear combinations of NNPDF4.0 Hessian EV.

Each of the solutions is an acceptable PDF set from the NNPDF4.0 fit.

High-density MC sampling of a span of a few EV directions that drive the specific PDF uncertainty.

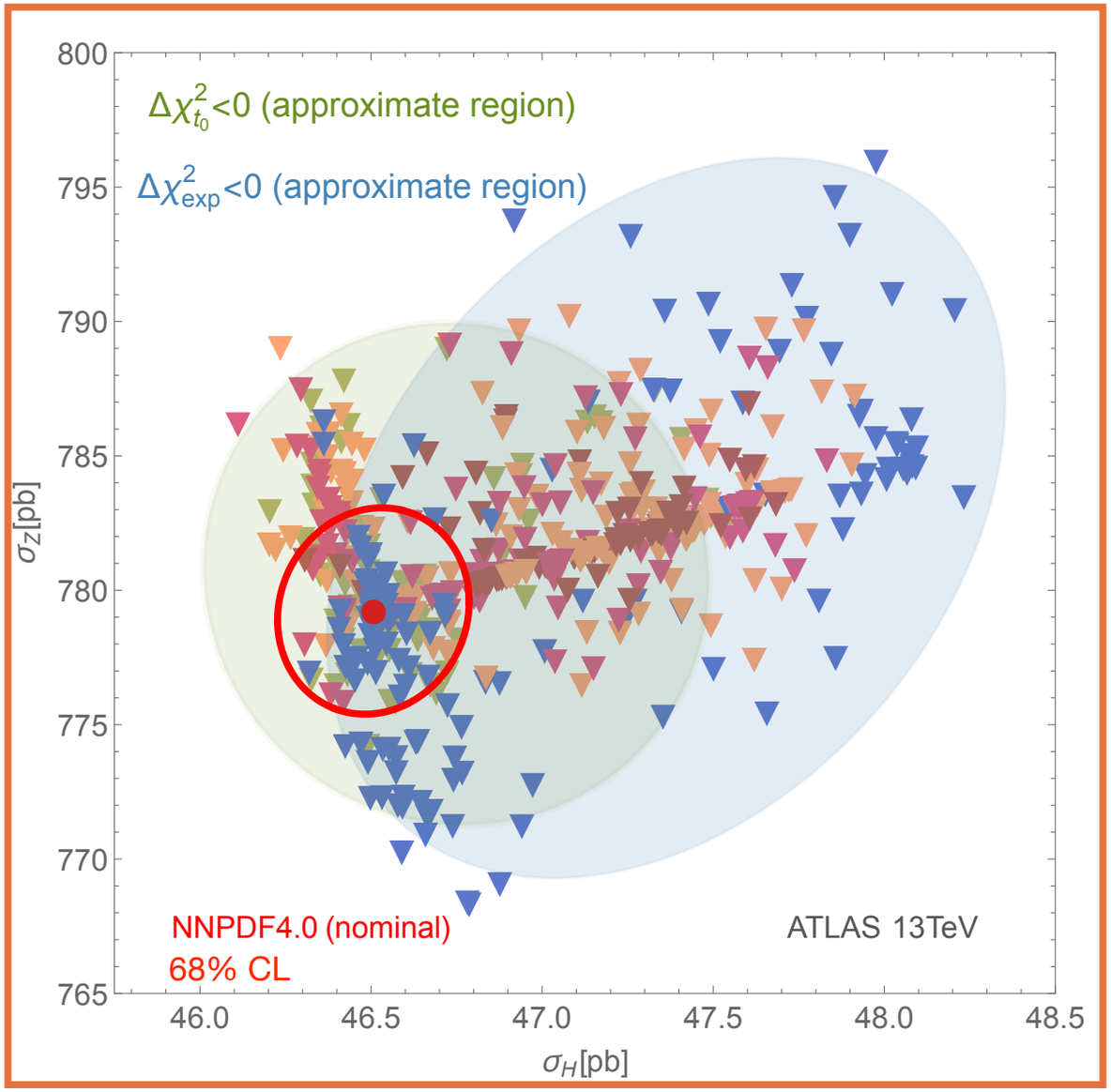


A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

The green and blue ellipses (constructed using a convex hull method) are approximate region containing all found replicas with $\Delta\chi^2 < 0$. They have no statistical meaning.

[Anwar, Hamilton, Nadolsky, 1901.05511]

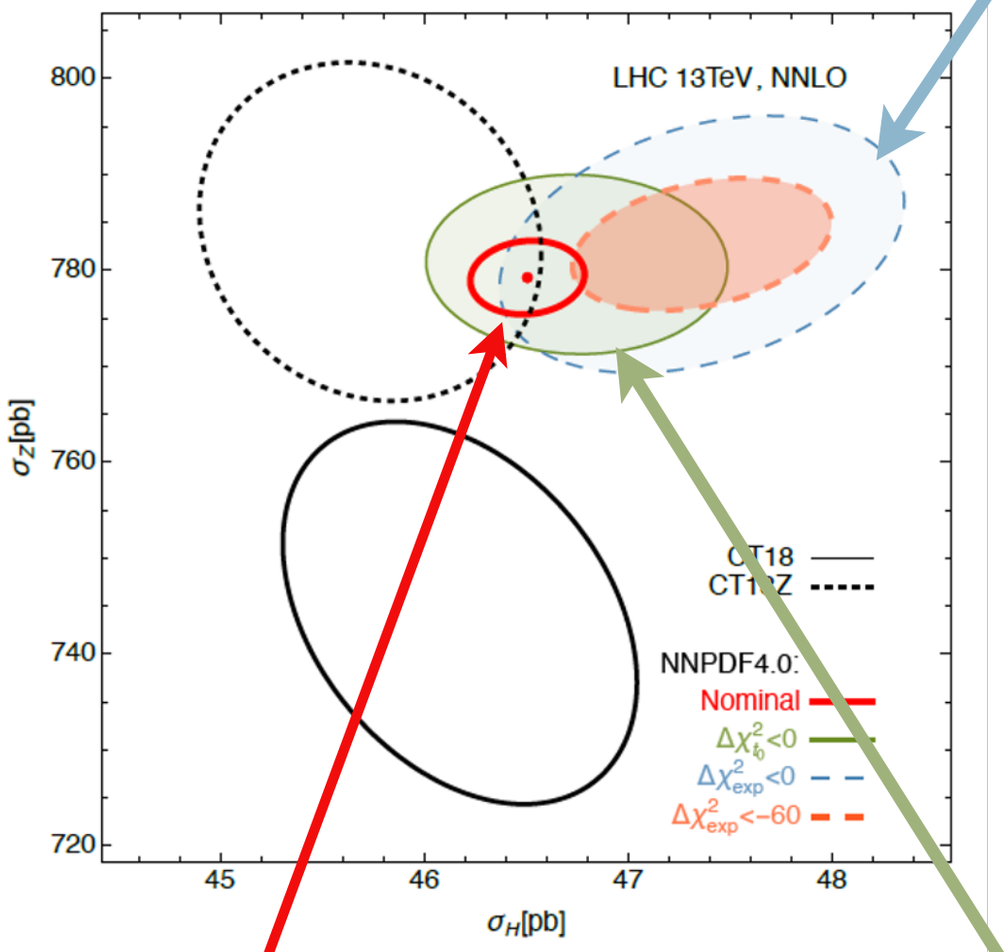
The green and blue areas are larger than the nominal NNPDF4.0 uncertainty (red ellipse).

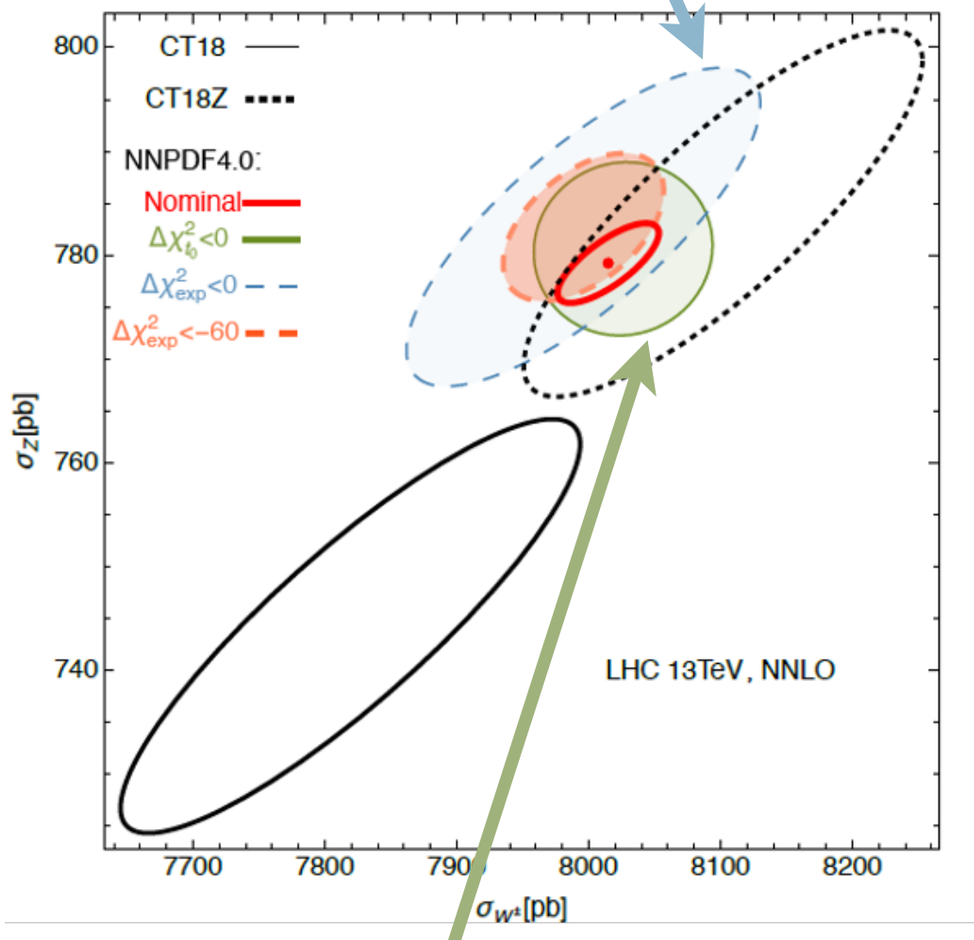


Monte-Carlo sampling sensitivity for PDFs

Regions containing (very) good solutions according to the experimental form of χ^2

(is used in χ^2 summary tables of the NN4.0 article, was a default in the NN4.0 public code)





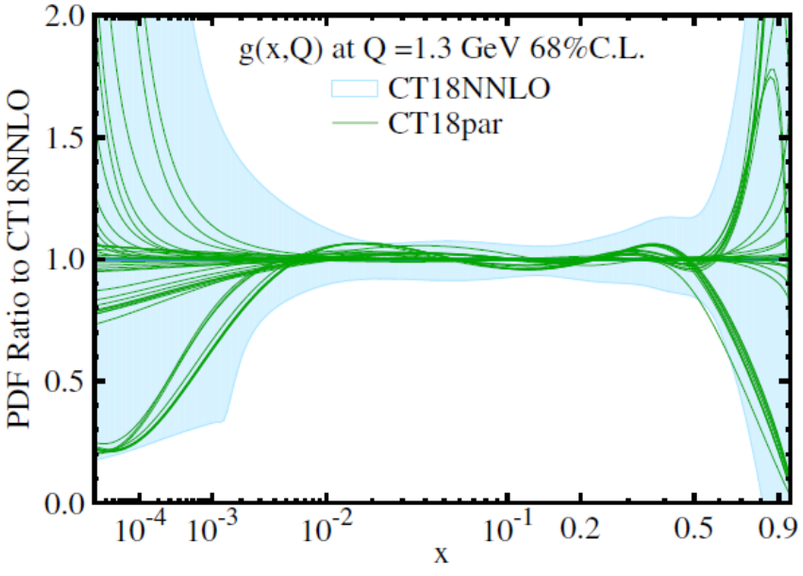
Nominal NN4.0 Hessian or MC 68%cl

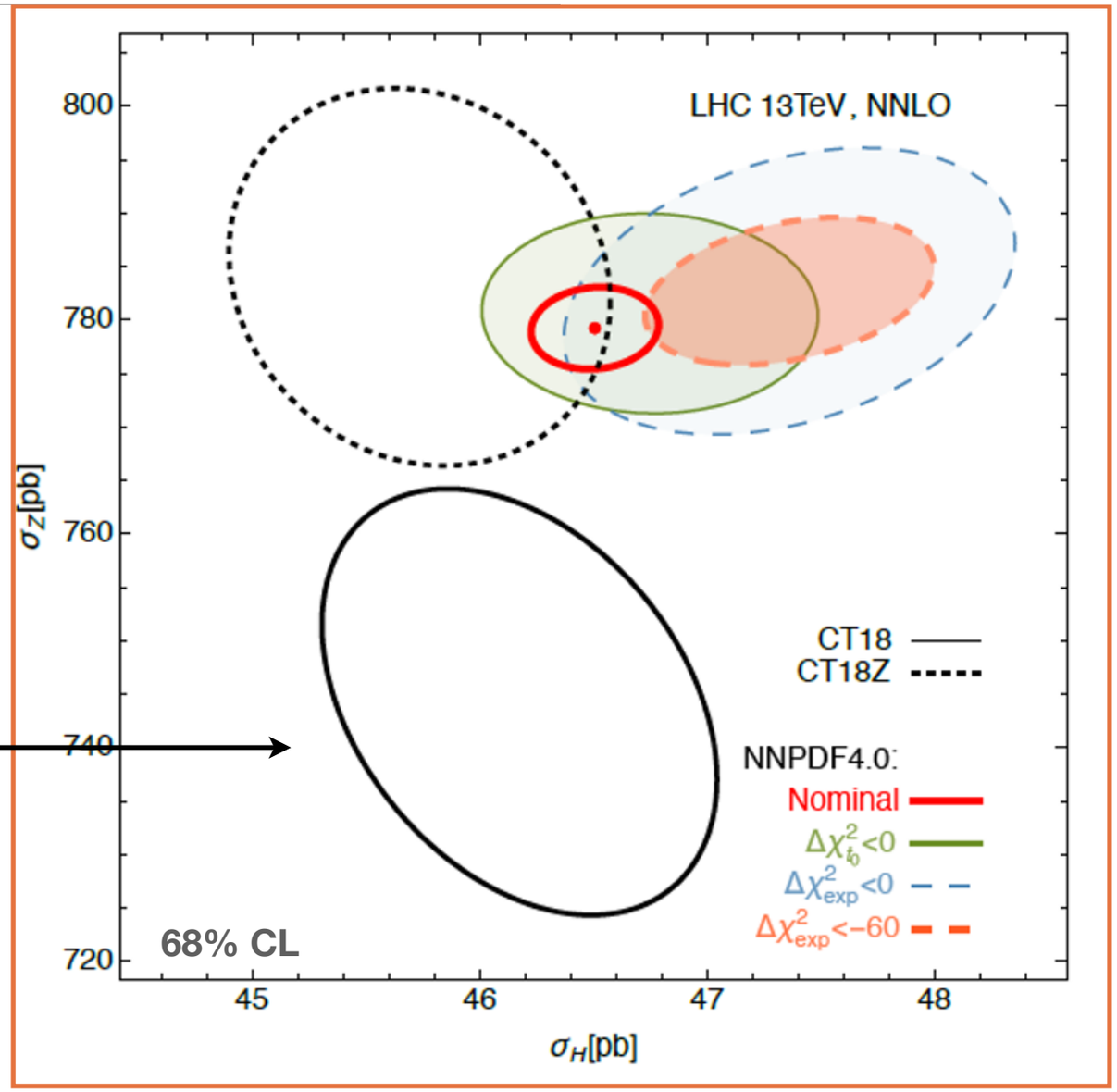
Region containing good solutions according to the most recent t_0 form of χ^2

(used to train NN4.0 replicas)

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.

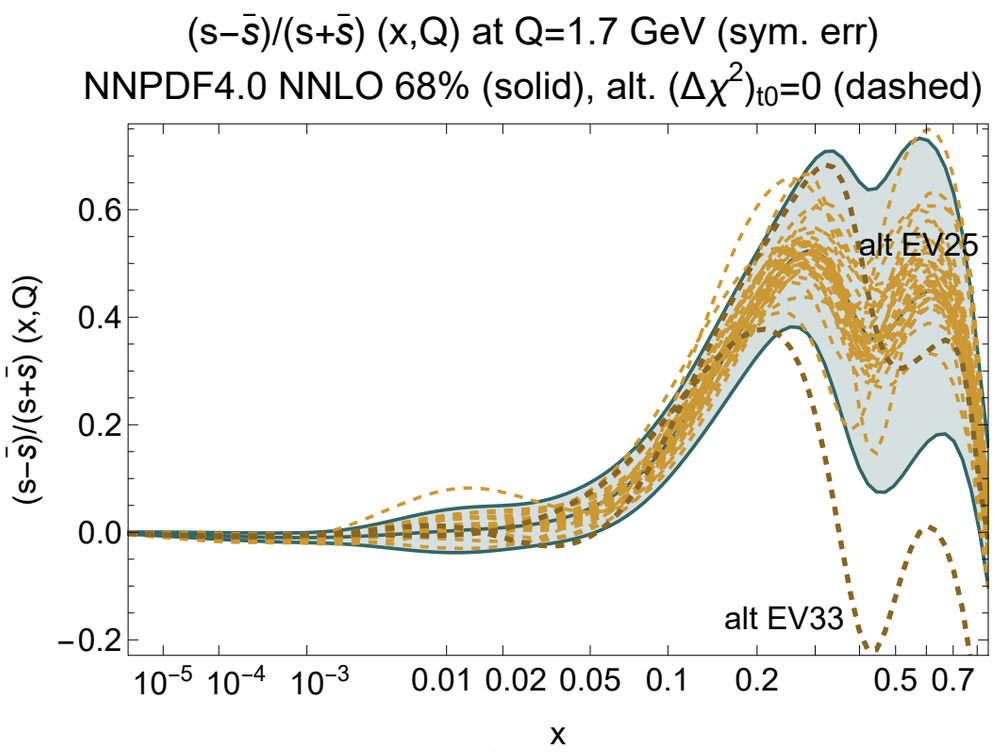


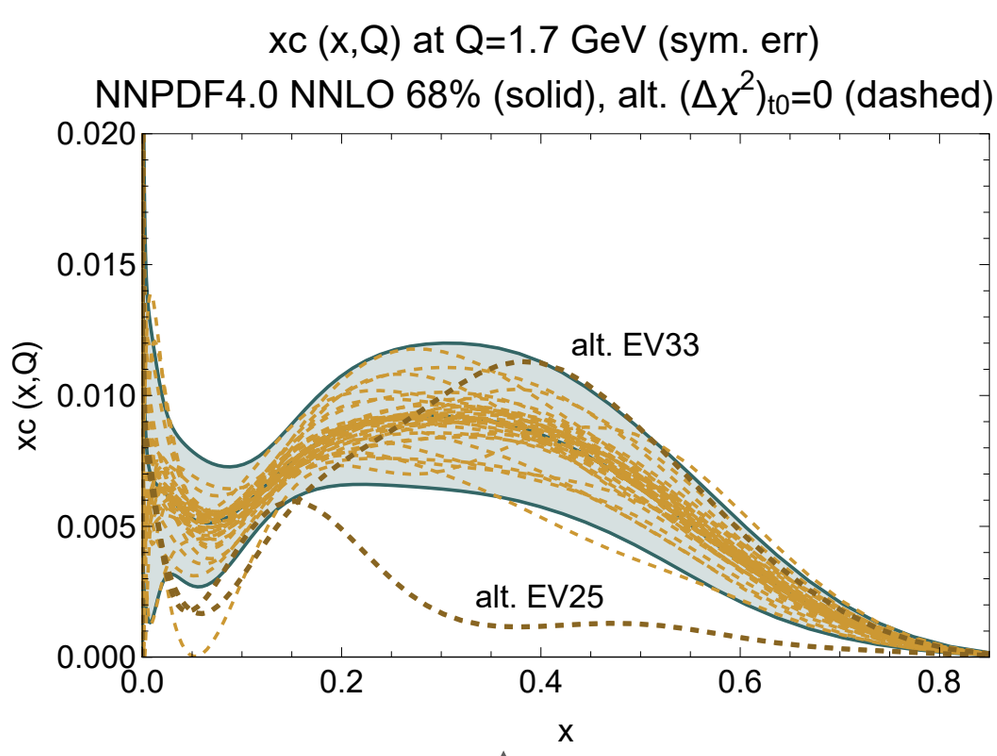


Filled color ellipses:

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

PDF tests outside the fit

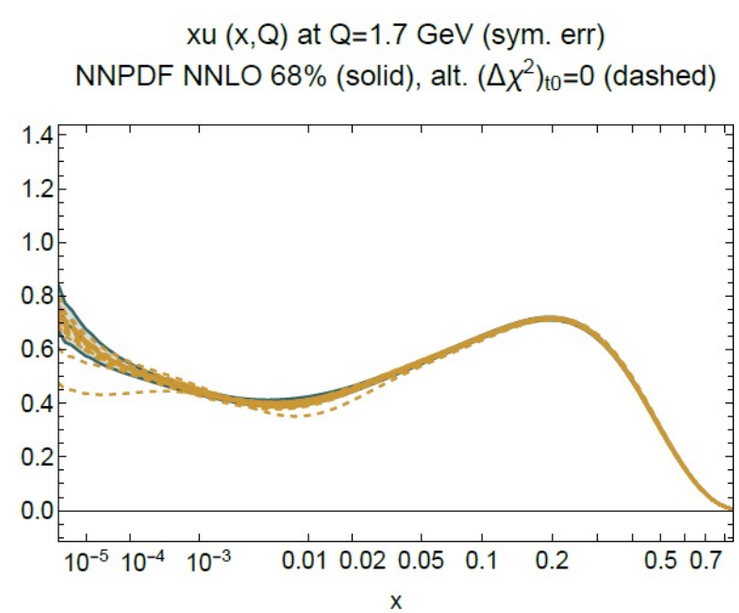


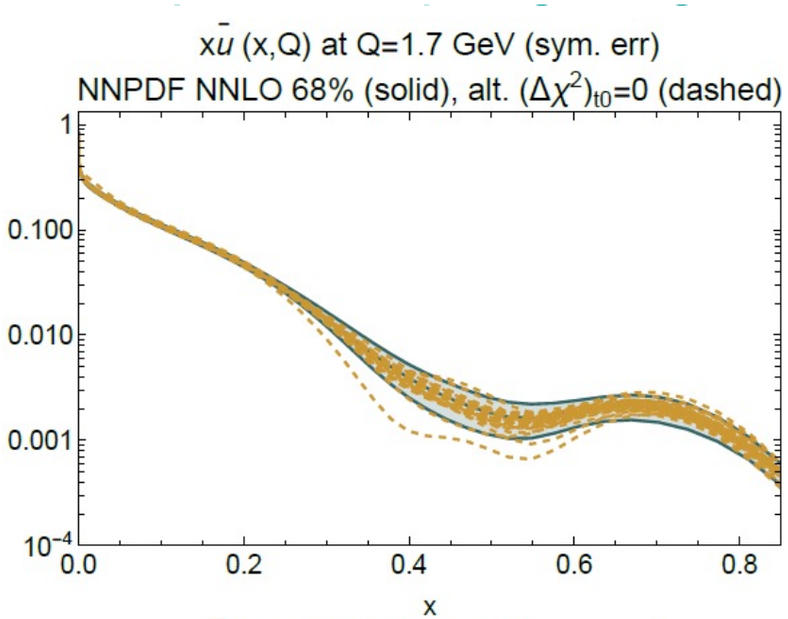


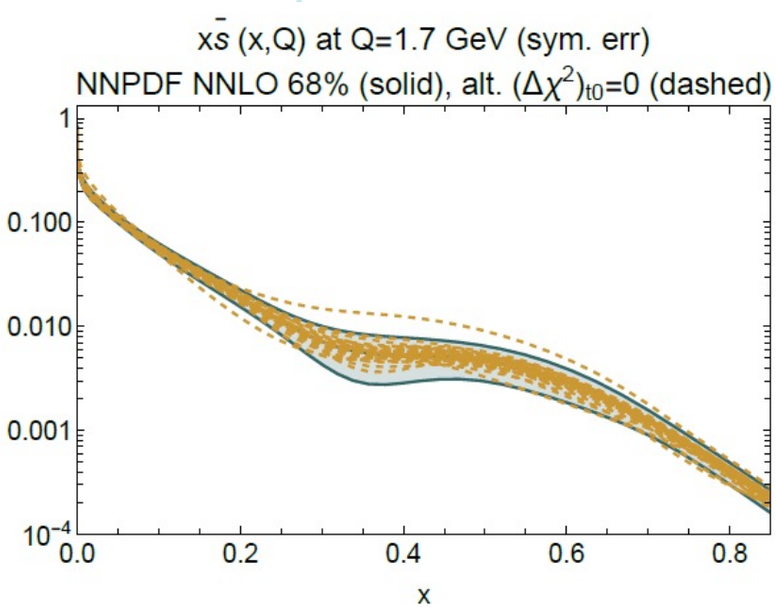
Hopscotch uncertainties wash out reported evidence for large positive strangeness asymmetry and non-zero intrinsic charm.

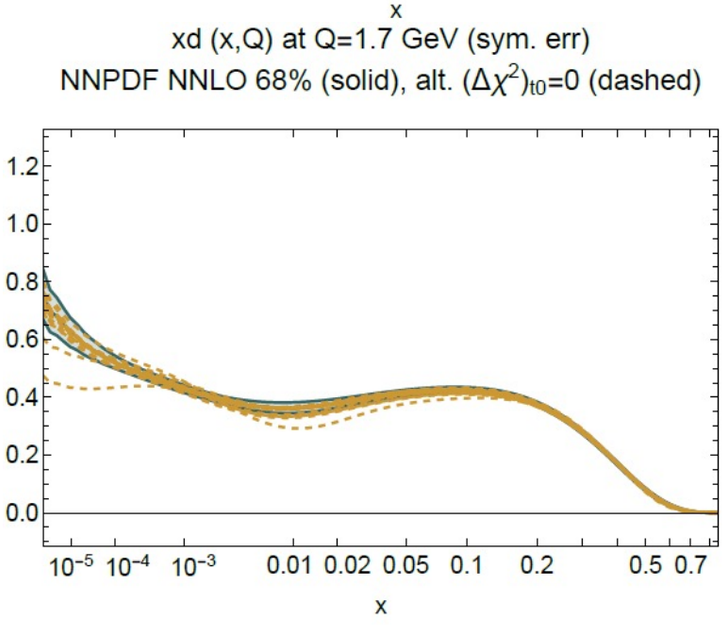
Hopscotch replicas

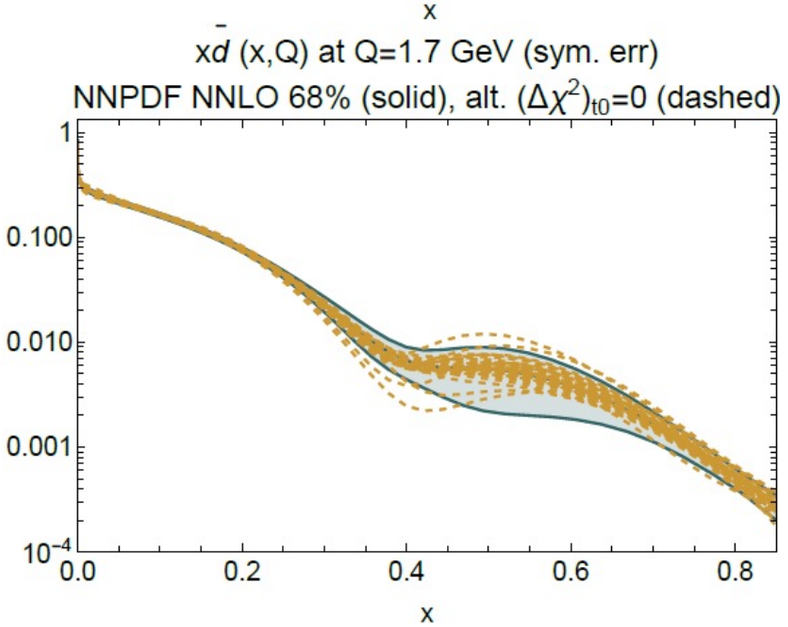
Nominal NN4.0 1σ bands and alternative $\Delta\chi^2_{t_0} = 0$ EV sets

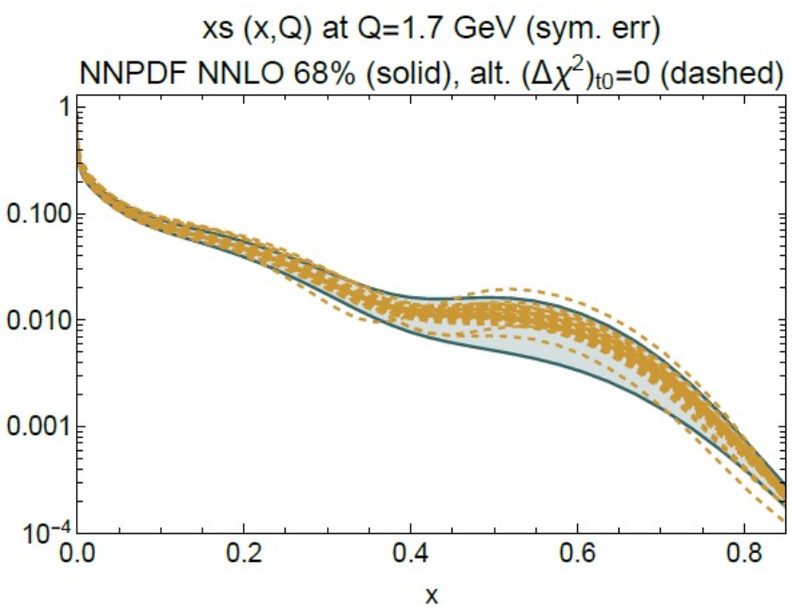






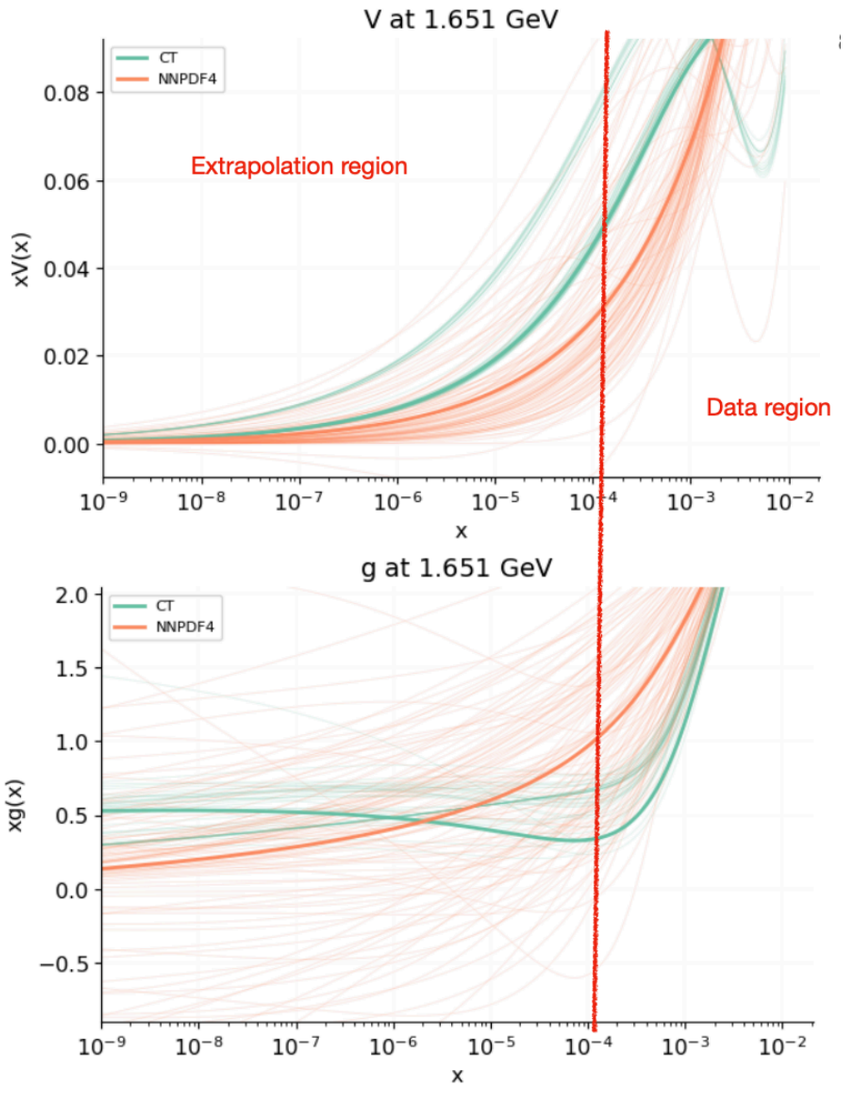






Error bands available at https://ct.hepforge.org/PDFs/2022hopscotch/

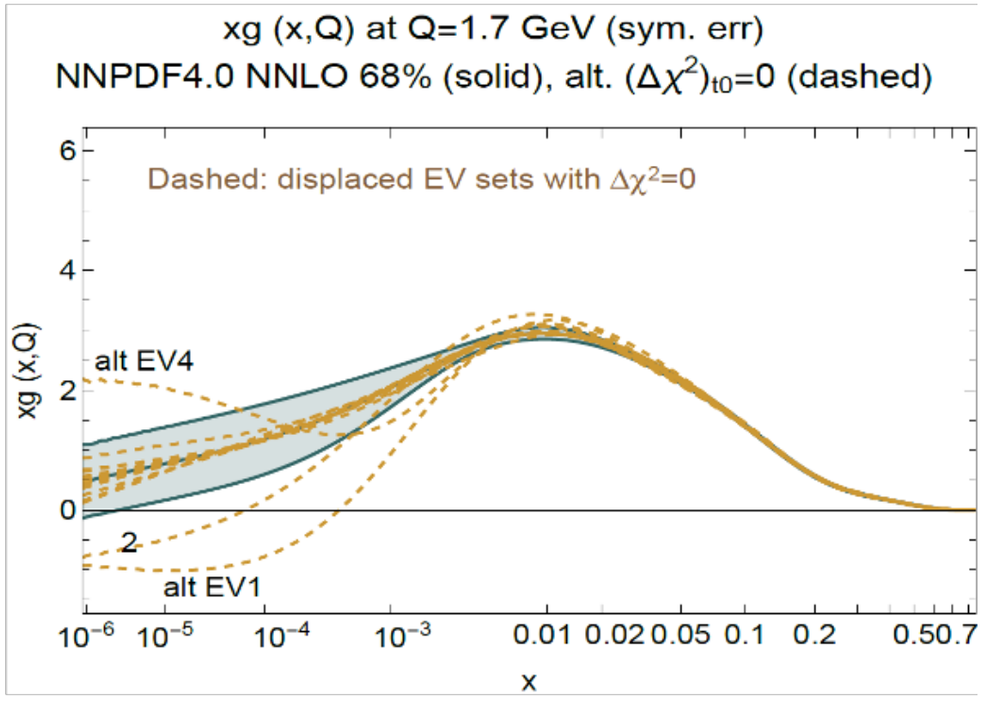
Hopscotch replicas and goodness-of-fit criteria



NNPDF prior disfavors some of the most displaced hopscotch replicas.

[M. Ubiali, HP2 2022 workshop, Durham, 2022/09]

The hopscotch replicas pass CT18 criteria for good individual solutions.



The ensemble of the hopscotch replicas are solutions that account for sampling bias. The hopscotch ensemble enlarges the NNPDF4.0 uncertainty for LHC benchmark cross section.

Conclusions

A new avenue to the tolerance puzzle is proposed through the study of the sampling uncertainties — a complementing source to the fitting uncertainty.

Highlights on the sampling uncertainties:

1. Tolerance criteria related to sampling choices. A PDF fit with few parameters and $\Delta\chi^2 = 1$ tolerance probably underestimates the parametric uncertainty.

2. Concept of effective large dimensions. Difficult to sample the full parameter space with many parameters without biases. A **hopscotch scan** intelligently reduces dimensionality of the relevant PDF parameter space for an observable under consideration.

3. Validating the final PDFs may be easier than understanding the respective fitting algorithm. Hopscotch algorithm is a **test outside the fit** to verify the PDF uncertainty for a specific QCD cross section or observable.

Conclusions

A new avenue to the tolerance puzzle is proposed through the study of the sampling uncertainties — a complementing source to the fitting uncertainty.

Highlights on the sampling uncertainties

Hopscotch scans, illustrated for the NNPDF4.0, **can be performed using public codes** (*LHAPDF + mcgen + xFitter/NNPDF fitting codes*).

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, TMDs, etc.

Of particular interest for future experiments — EIC, …

Back-up slides

Role of objective function

Hessian-based global analyses:

Figure-of-merit (function of the parametrization) and tolerance criteria will define the size of uncertainties.

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]

Chi-square definition

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

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

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

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

Please help to fill in

Goodness-of-fit functions

PDF Analysis	χ^2 prescription to fit PDFs	χ^2 prescription to compare PDFs	Comments
HERAPDF	HERA	HERA	
CT	Extended T +prior	Extended T	
MSHT'20	T [?]	T [?]	
NNPDF4.0	t_0 + prior with fluctuated data	Experimental or t_0 with unfluctuated data	t_0 prescription comes in pre- and post-NNPDF3.0 versions
…			
Hopscotch'2022	N/A	Experimental or t_0 [2022] with unfluctuated data	

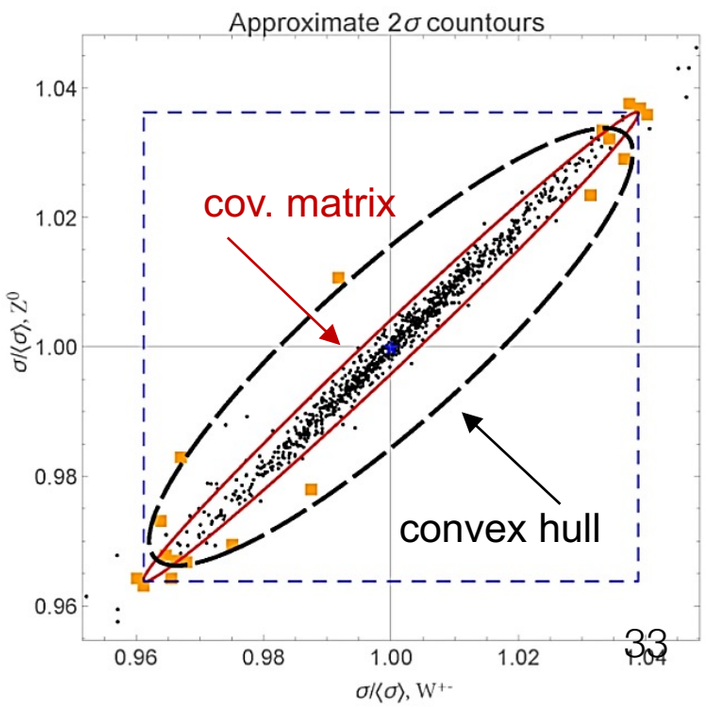
1. Systematic errors are approximate and can introduce biases in all analyses.
2. D'Agostini [1994] demonstrated a bias in an α_s fit to LEP data with one multiplicative normalization. In PDF fits, the pulls among several multiplicative factors may cancel, resulting in no substantial bias on the PDFs.

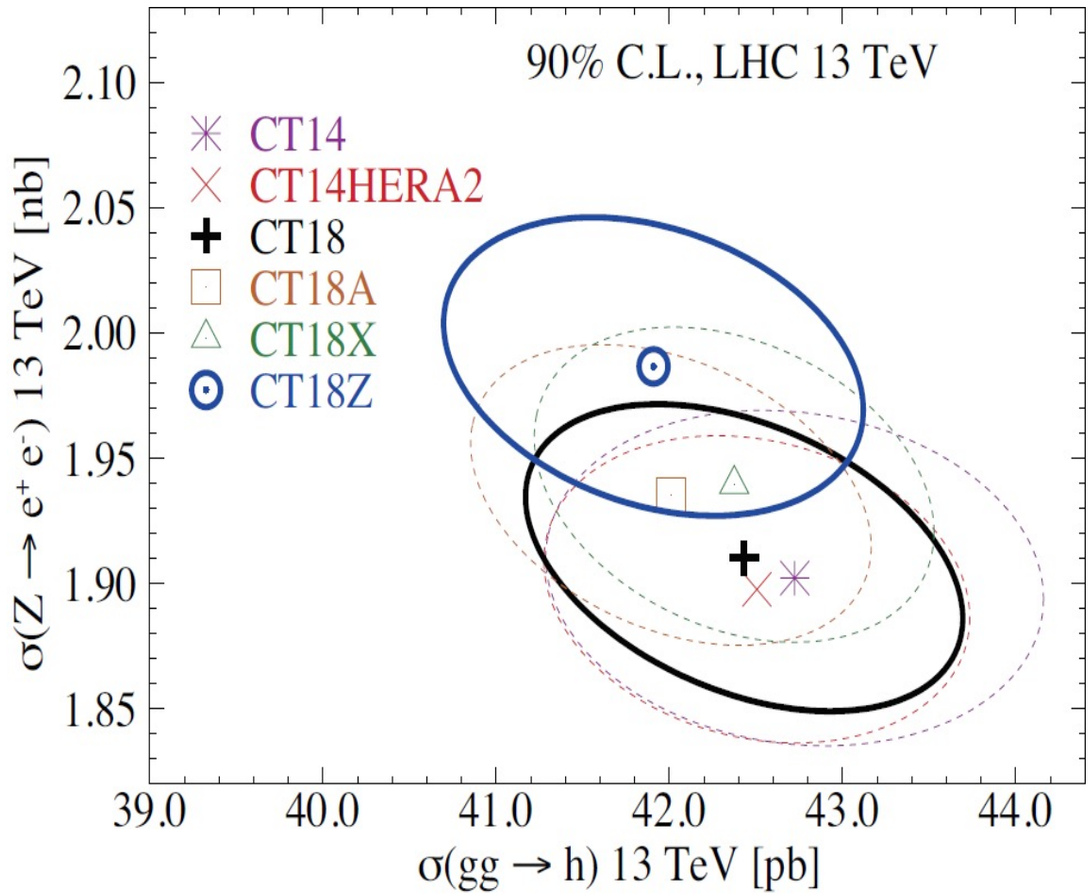
Toward **robust** PDF uncertainties

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 prescription in different groups

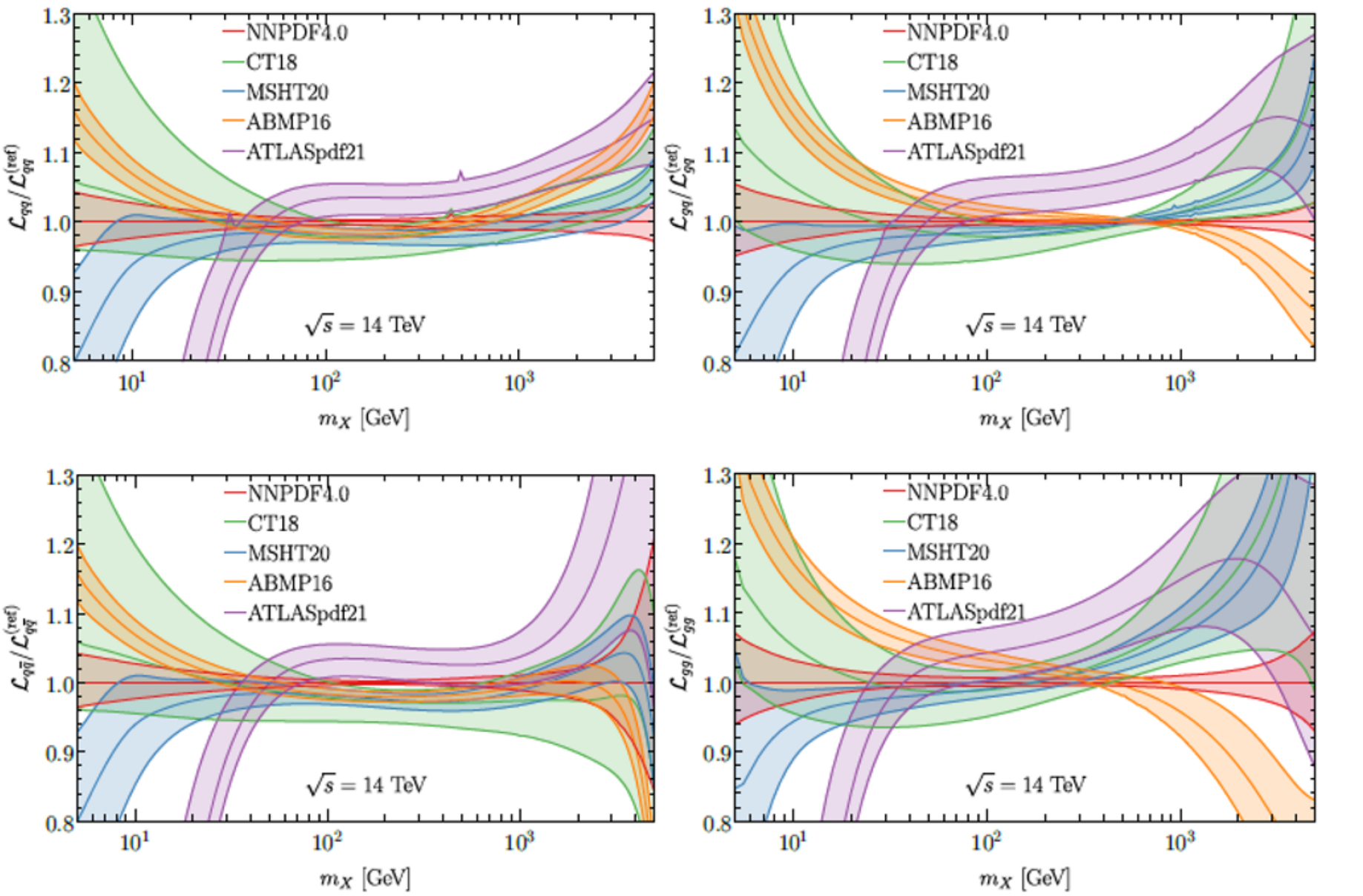


FIG. 4. Comparison, as a function of the invariant mass m_X, of the parton luminosities at $\sqrt{s} = 14$ TeV, computed using N2LO NNPDF4.0, CT18, MSHT20, ABMP16 with $\alpha_s(M_Z) = 0.118$, and ATLASpdf21. The ratio to the NNPDF4.0 central value and the relative 1σ uncertainty are shown for each parton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

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.

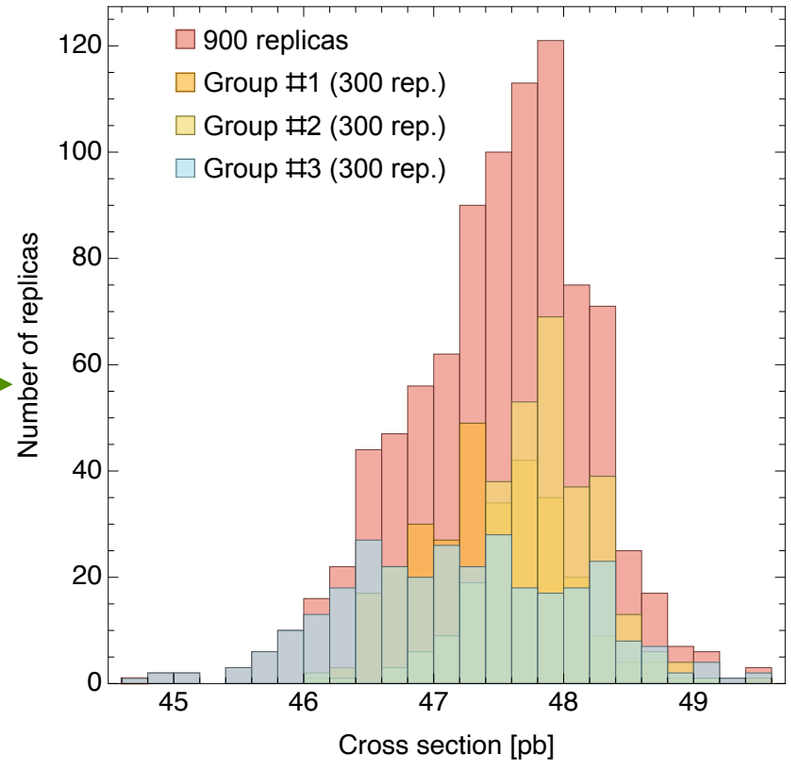
The kinematics cuts for the correlated uncertainties are fixed as the same of the NNPDF4.0 global analysis.

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.

Law of large numbers — Higgs XS

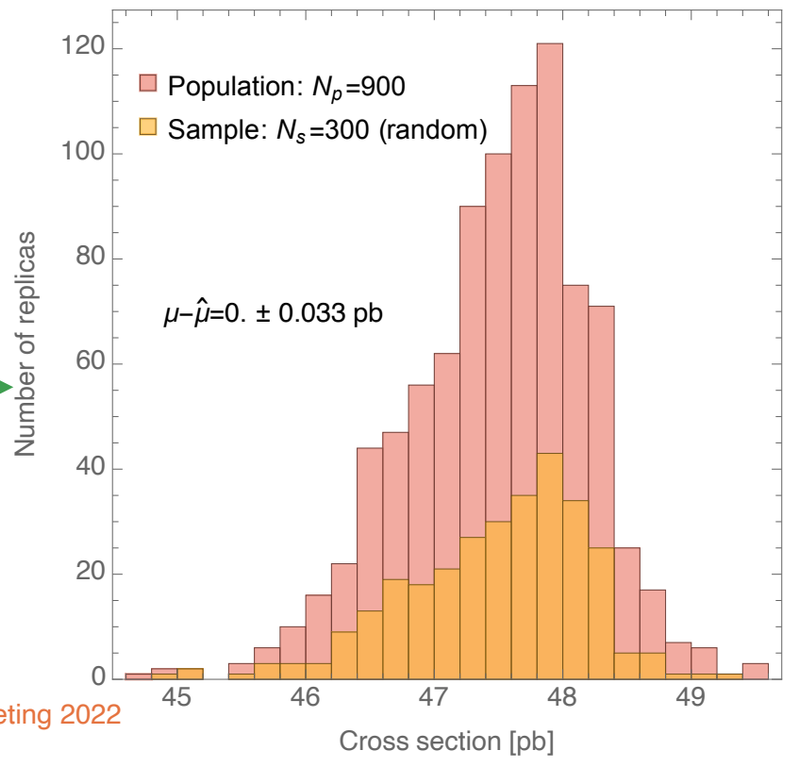
A toy sampling excercise

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



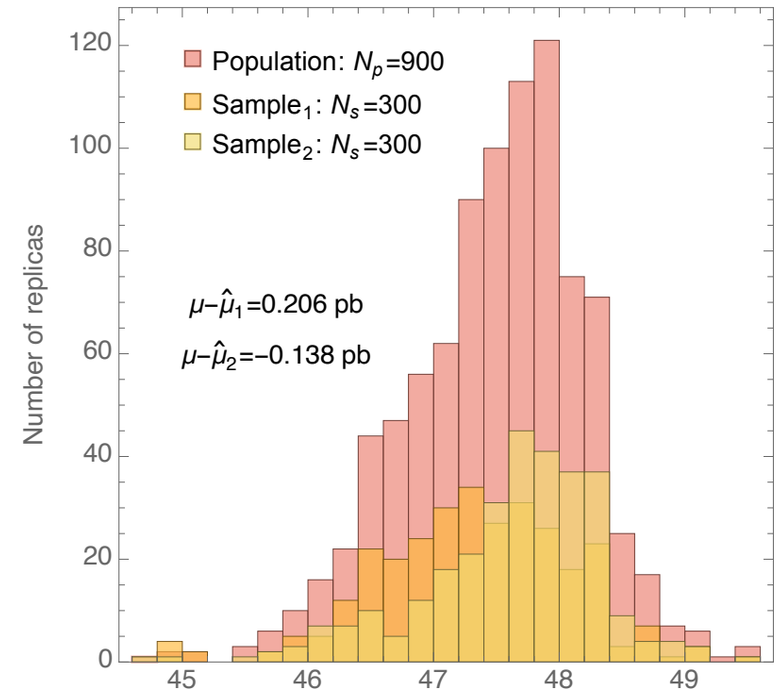
We **randomly** select 300 out of the 900 cross sections. The law of large number is fulfilled in this case: there is no bias in the original sampling of the 3 sets of Higgs cross sections.

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



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 $\sigma/\sqrt{n}$!

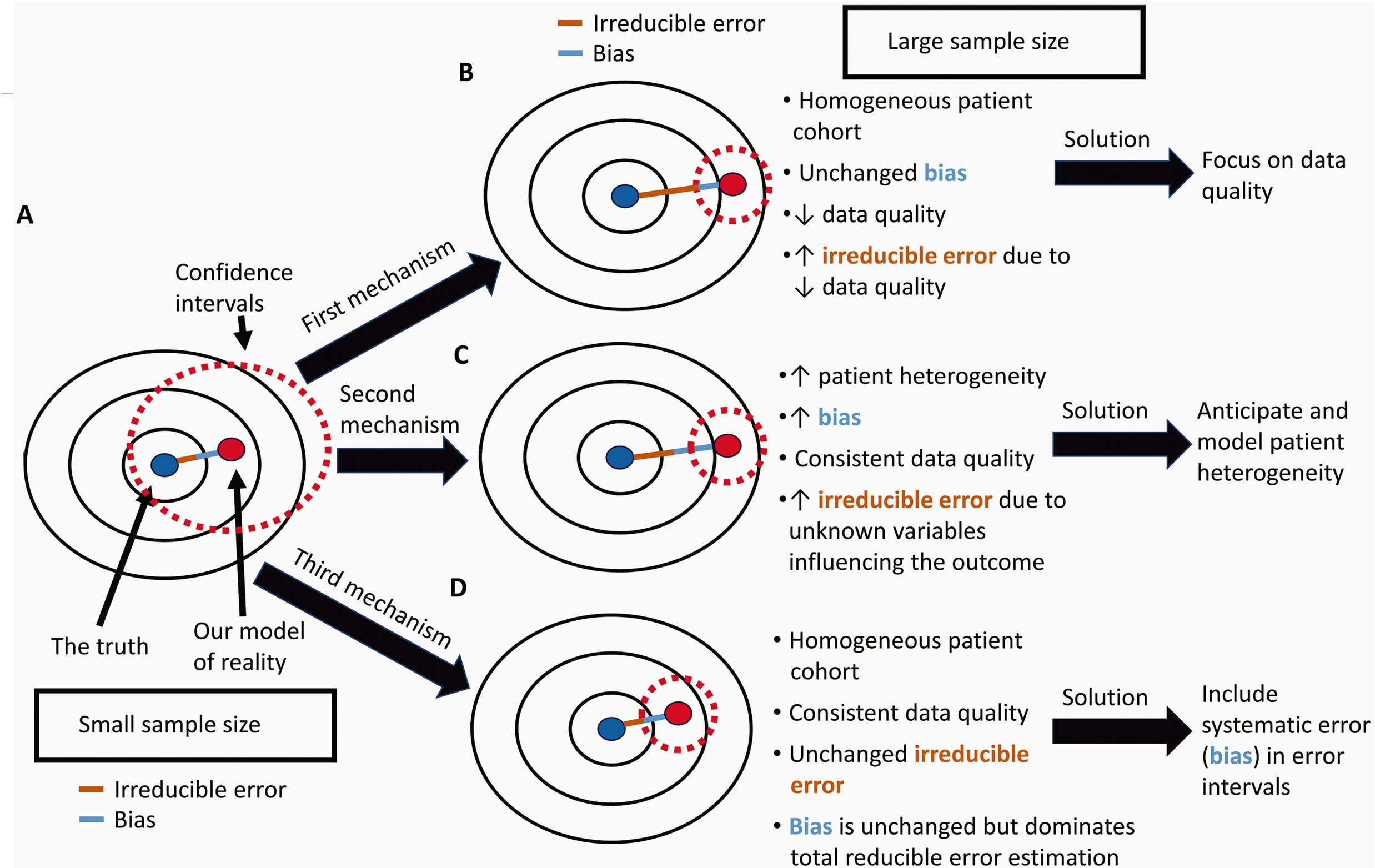


The law of large numbers disregards the *quality of the sampling* — distribution of n for a population size N/measure of the parameter space.

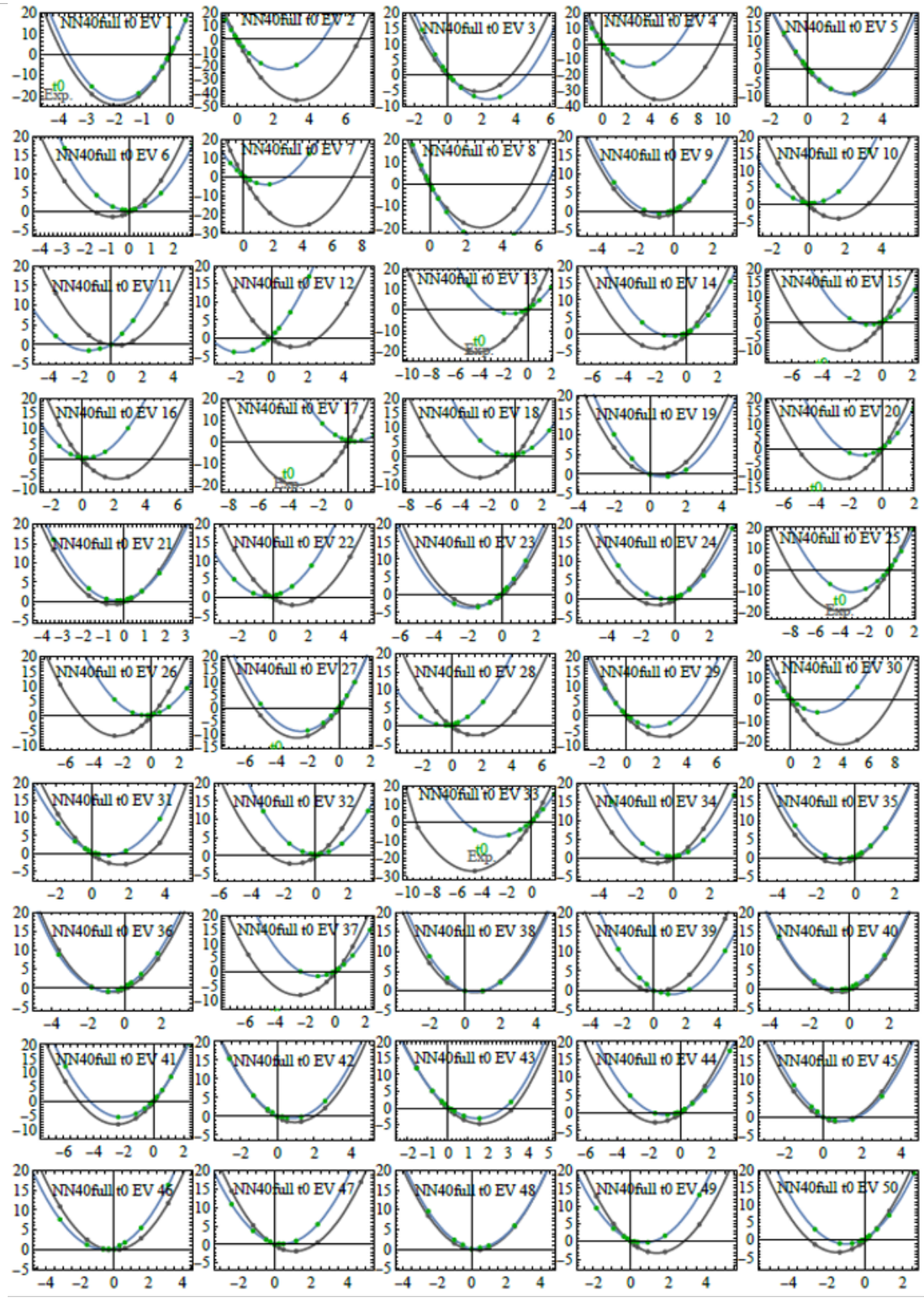
The **trio identity** remedies to that problem be accounting for the sampling bias:

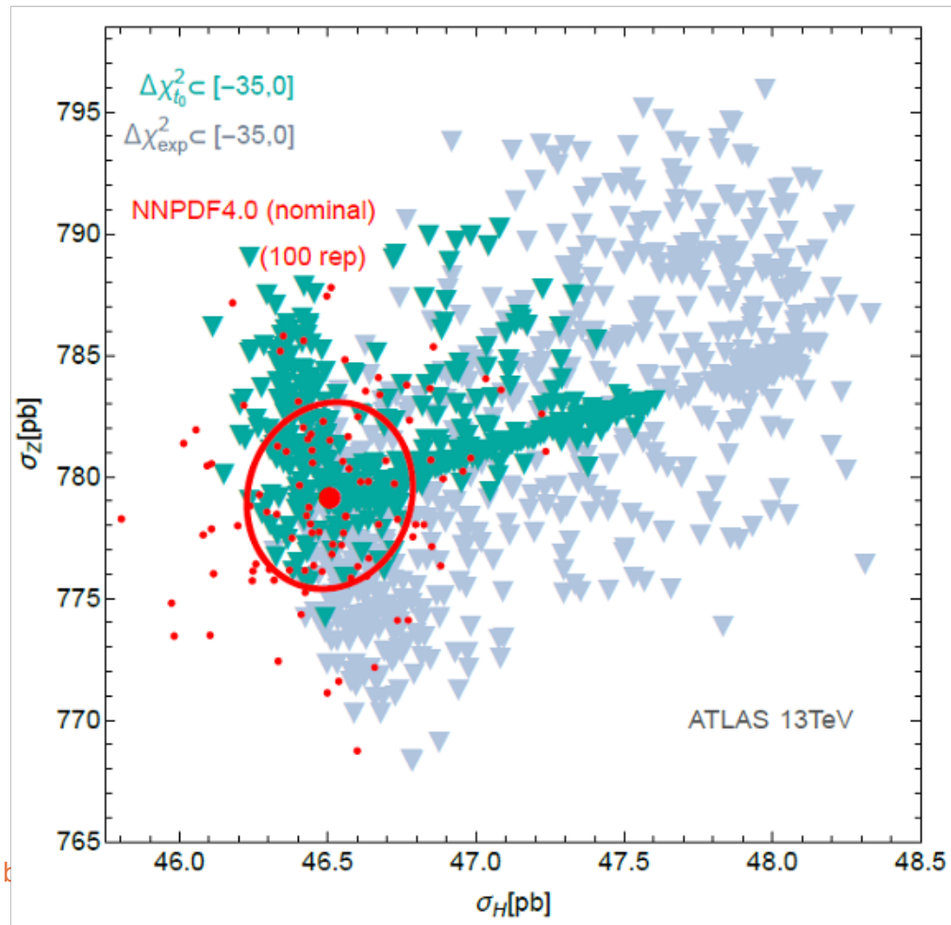
$$\mu - \hat{\mu} = (\text{data+sampling defect}) \times (\text{sampling discrepancy}) \times (\text{inherent problem difficulty})$$

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


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
Small sample size
Irreducible error
Bias
• 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

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

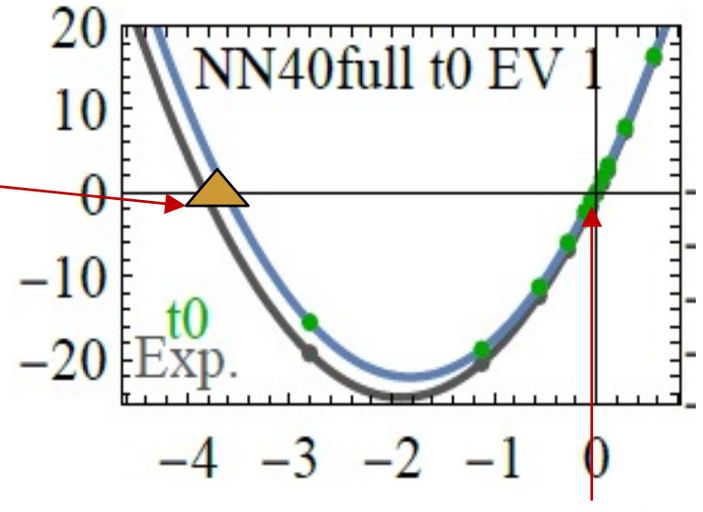




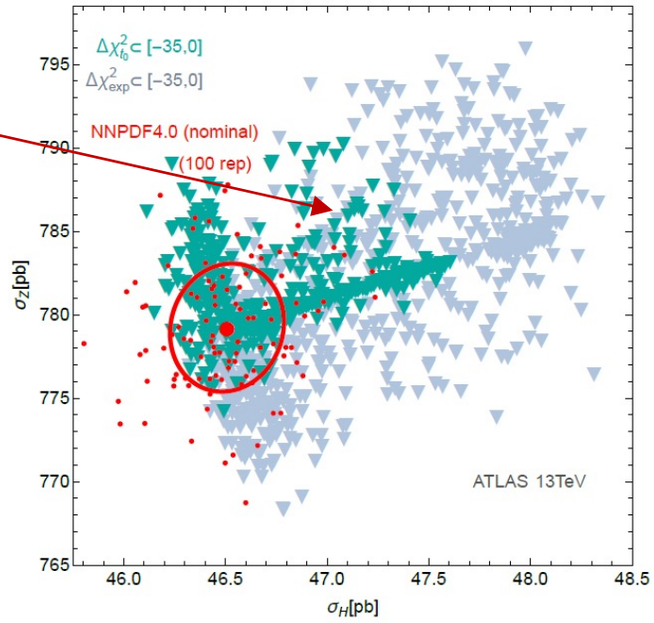
Hopscotch NN4.0 replicas

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

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



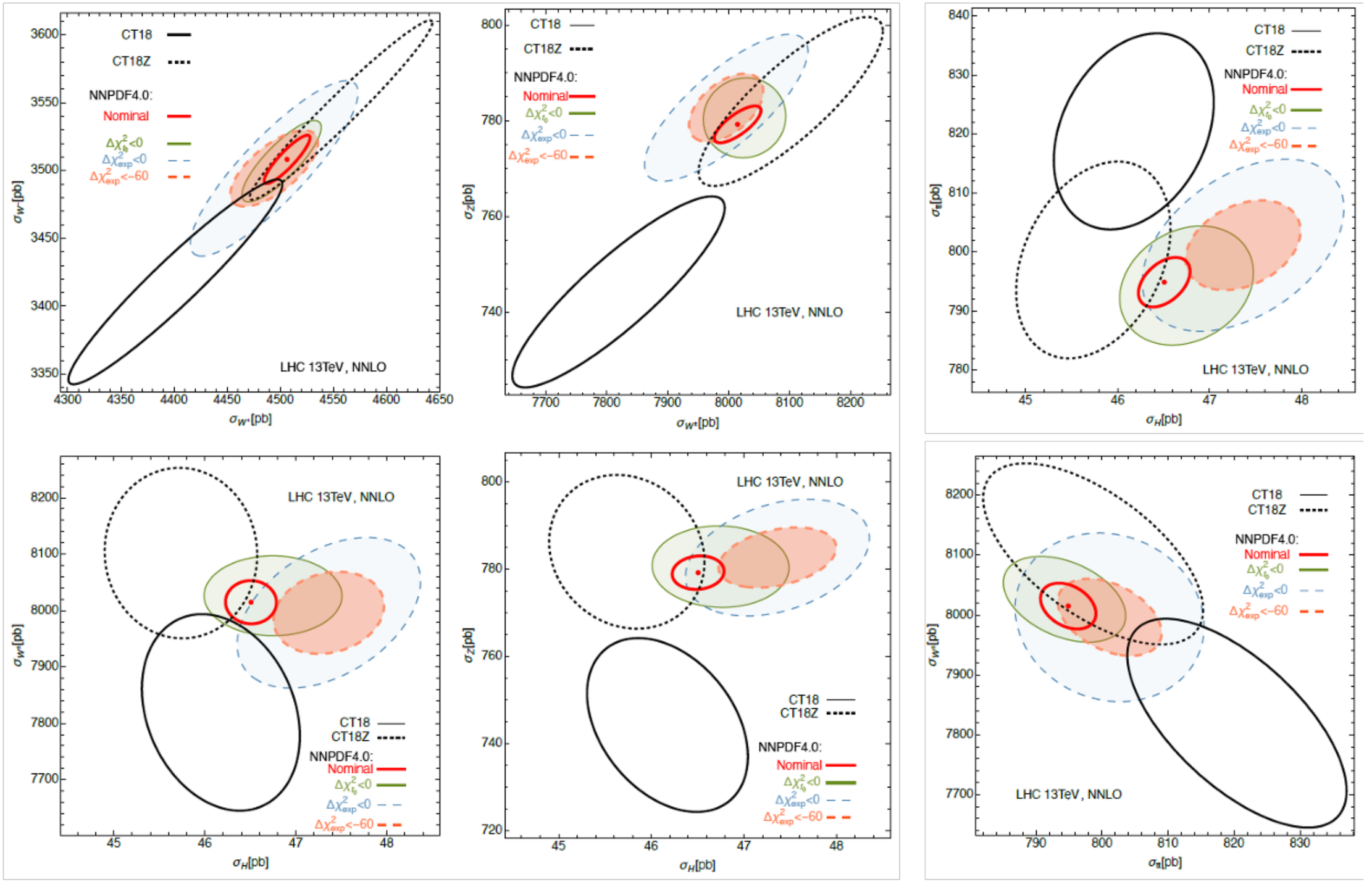
2. A total 2329 PDF sets from hopscotch scans on $\sigma_Z, \sigma_{W^+}, \sigma_{W^-}, \sigma_H, \sigma_{t\bar{t}}$ total inclusive cross sections at the LHC 13 TeV



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

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

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



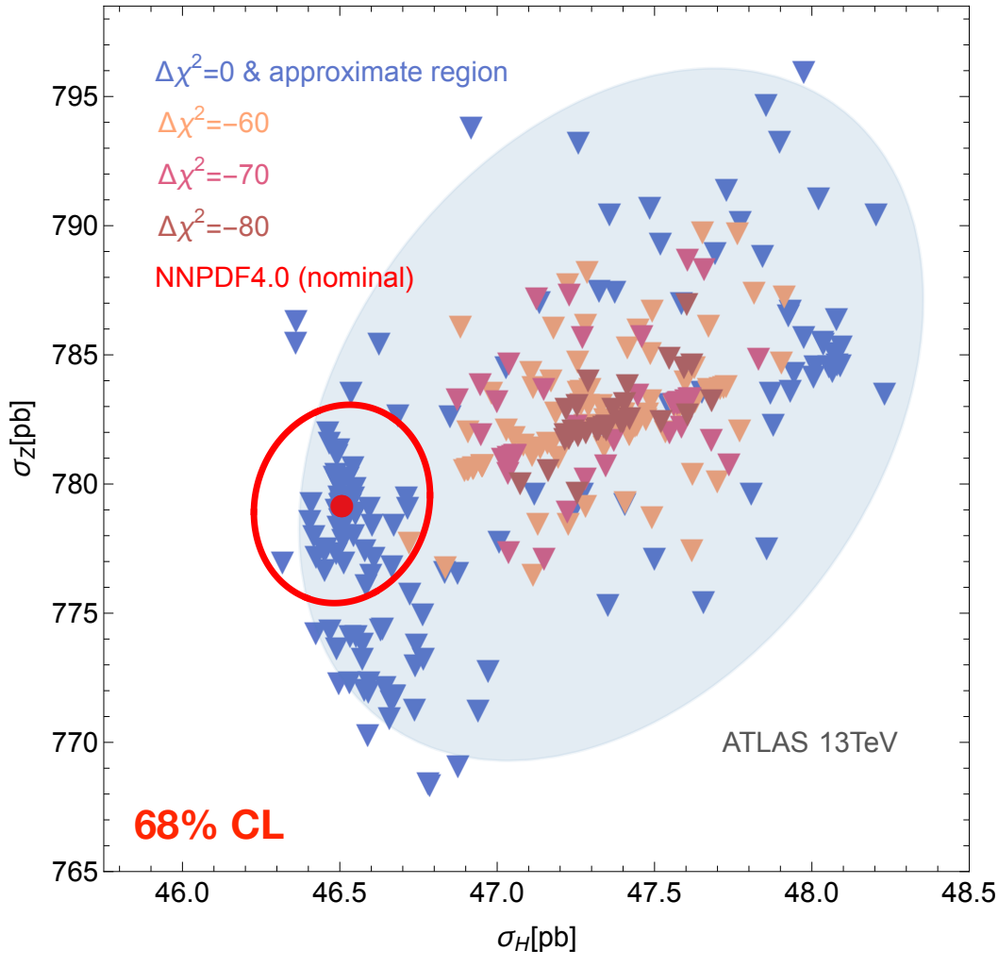
Ellipses at 68% CL

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).

Figures of merit in the NNPDF4.0 analysis I

1. χ^2 with respect to the central experimental values

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

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

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.

NNPDF4.0 use:

a. $X_i = D_i$: "**exp**erimental scheme"; can result in a bias

b. $X_i = \text{fixed } T_i$: "$\boldsymbol{t_0}$ scheme"; can result in a (different) bias

Figures of merit in the NNPDF4.0 analysis II

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

NNPDF4.0 use:

a. $X_i = D_i$: **exp**erimental scheme; can result in a bias

b. $X_i = \text{fixed } T_i$: $\boldsymbol{t_0}$ scheme; can result in a (different) bias

The conventions are neither complete nor unique. Ambiguity affects all groups.
See Appendix in 1211.5142.

2. **NNPDF4.0 trains MC replicas with χ^2 for fluctuated D_i, t_0 scheme, and replica selection (prior) conditions:**

$$\text{Cost}=\chi^2_{t_0}(T_i, D_i^{fluctuated}) + \chi^2_{prior}$$

3. NNPDF4.0 quotes the final unfluctuated χ^2 in the "exp" scheme.

Experimental scheme:
$\chi^2_{tot}/N_{pt} = 1.160\,.$

t_0 scheme:
$\chi^2_{tot}/N_{pt} = 1.233\,.$

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

PRIOR PROBABILITY IN PDF FITS

✓ PDF fitting example of inverse problem: aim to find a posterior probability of ***f*** given the data ***D***.

✓ Parametrization of PDFs: finite-dimensional problem.

$$f(x) \approx \tilde{f}(x, \theta) \in \mathcal{F}$$

✓ The posterior probability for the parametrization depends on both the figure of merit that maximises the data likelihood given the parameters and on prior probability ***H.***

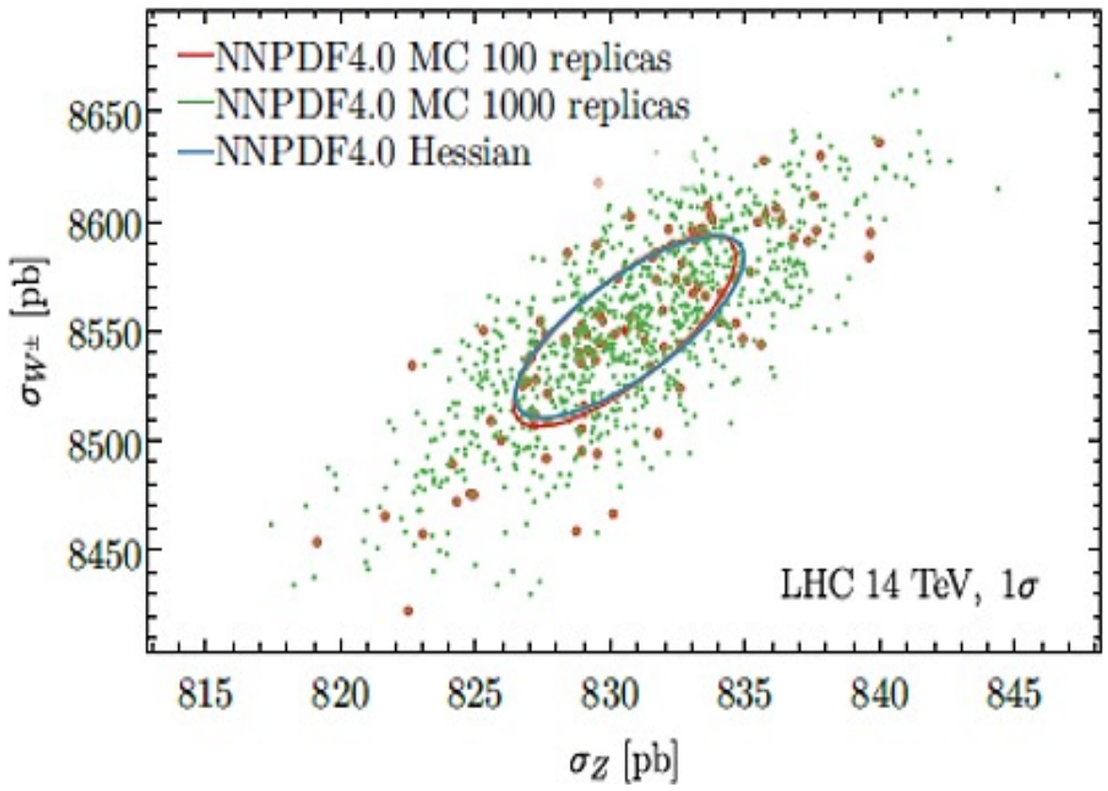
(M. Ubiali, HP2 2022 workshop, Durham, 2022-09-22)

Computing uncertainty ΔX

1. By unweighted averaging of predictions for 100 (or 1000) MC replicas:

$$\langle X \rangle = \frac{1}{N_{rep}} \sum_{i=1}^{N_{rep}} X_i \,; \quad \Delta X^2 = \langle (X - \langle X \rangle)^2 \rangle$$

(NNPDF calls it "**importance sampling**". The MC replicas are distributed according to the fluctuated data [Ball:2011gg] using the same training algorithm).



Replica 0 is the mean of 1000 MC replicas; has better unfluctuated χ^2 than MC replicas.

2. Using $N_{eig} = 50$ Hessian PDFs.

$$\Delta X^2 = \sum_{i=1}^{N_{eig}} (X_i - X_0)^2 \,.$$

NNPDF4.0 MC and Hessian uncertainties are in a good agreement.

From arXiv: 2205.10444 , Sec. 3D

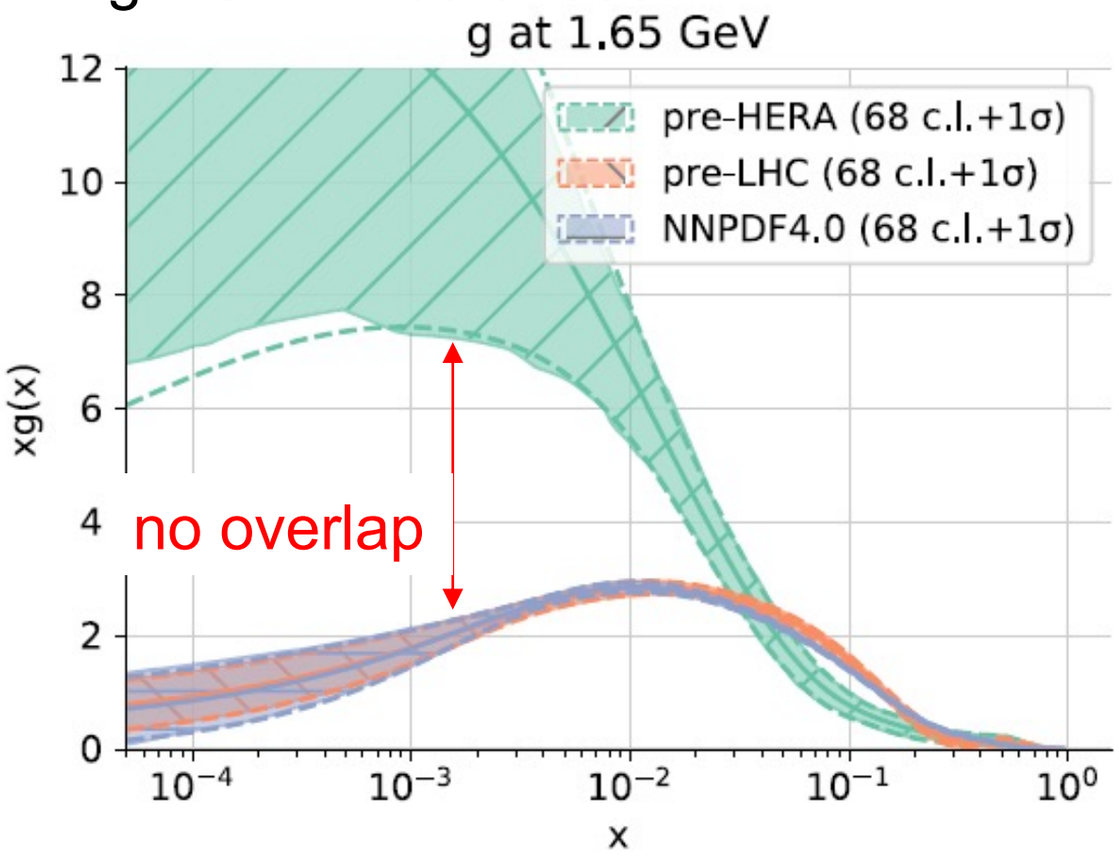
If the hopscotch solutions are acceptable, a natural question to raise is why they are not covered by the nominal NNPDF set. … As a possible hint, any hopscotch solution can be represented by a neural network in accord with the universal approximation theorems. The challenge of representative sampling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

What, exactly, did HERA do for us?

Evidence for non-trivial small-x dynamics depends on the uncertainty definitions

Example: a future test in NNPDF4.0

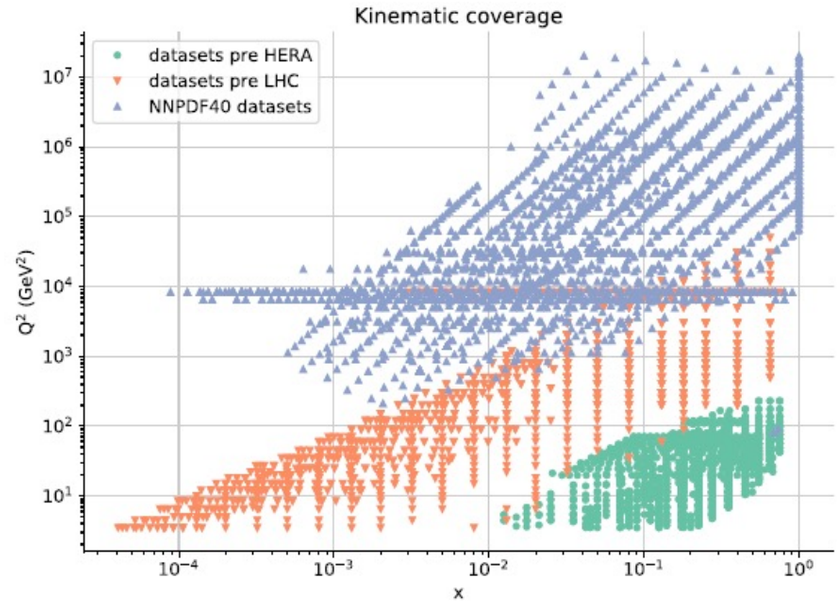
Fig. 29 in 2109.02653



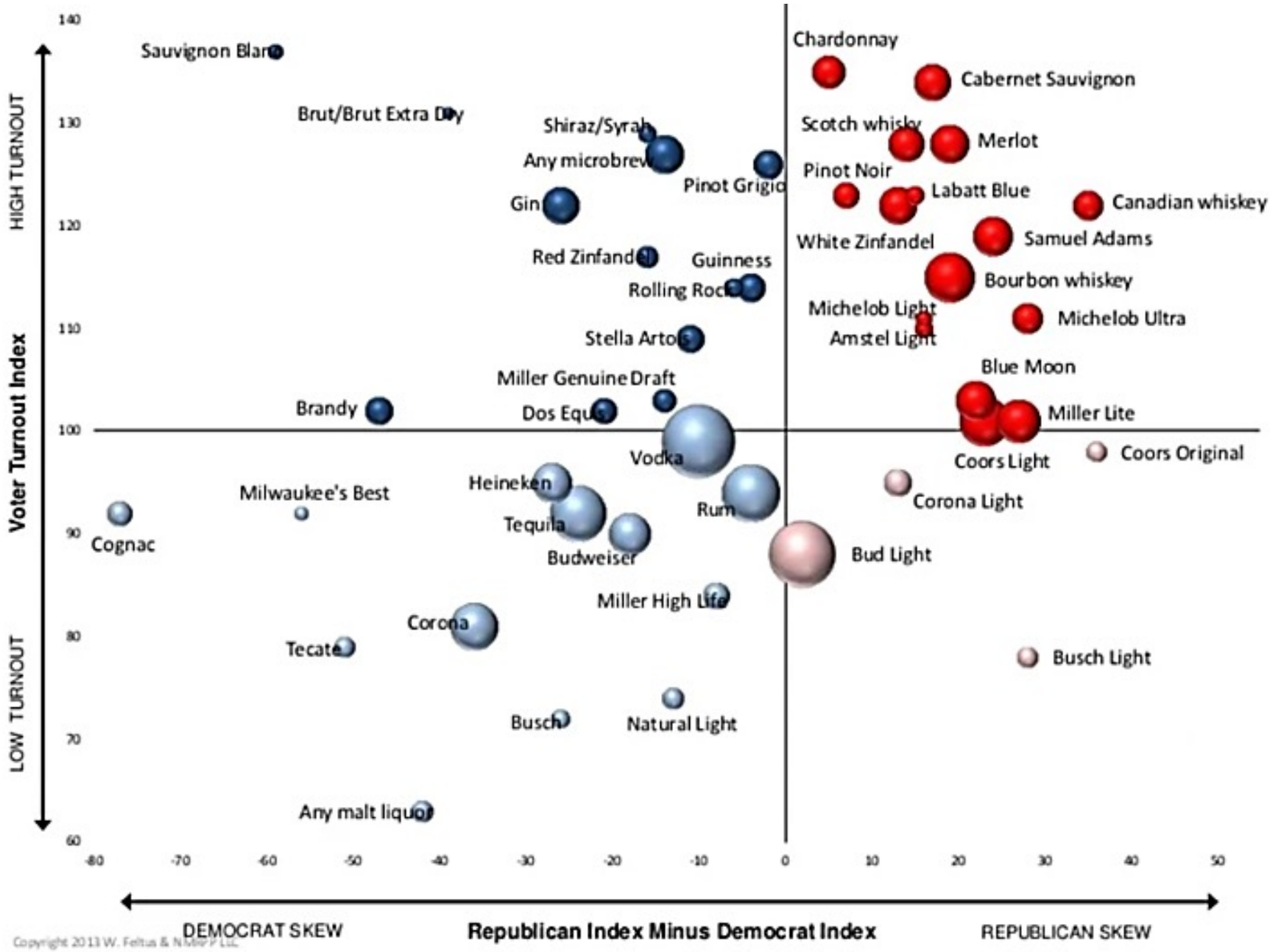
A fit only to the pre-HERA DIS & DY data prefers fast growth of the gluon at $x \to 0$, possibly reflecting a tension of BCDMS and NMC data. The growth is **reduced** by including the HERA data.

Historically, HERA was credited for establishing the fast small-x growth of the gluon (hard pomeron), not reducing the growth.

Which view is right?



Art and science of polling



Identification of the few discriminating variables is an art component of predictive polls

Here alcohol preferences and several other variables are proposed to weight election polls.

More plots [at https://tinyurl.com/yc67kkf9](https://tinyurl.com/yc67kkf9)

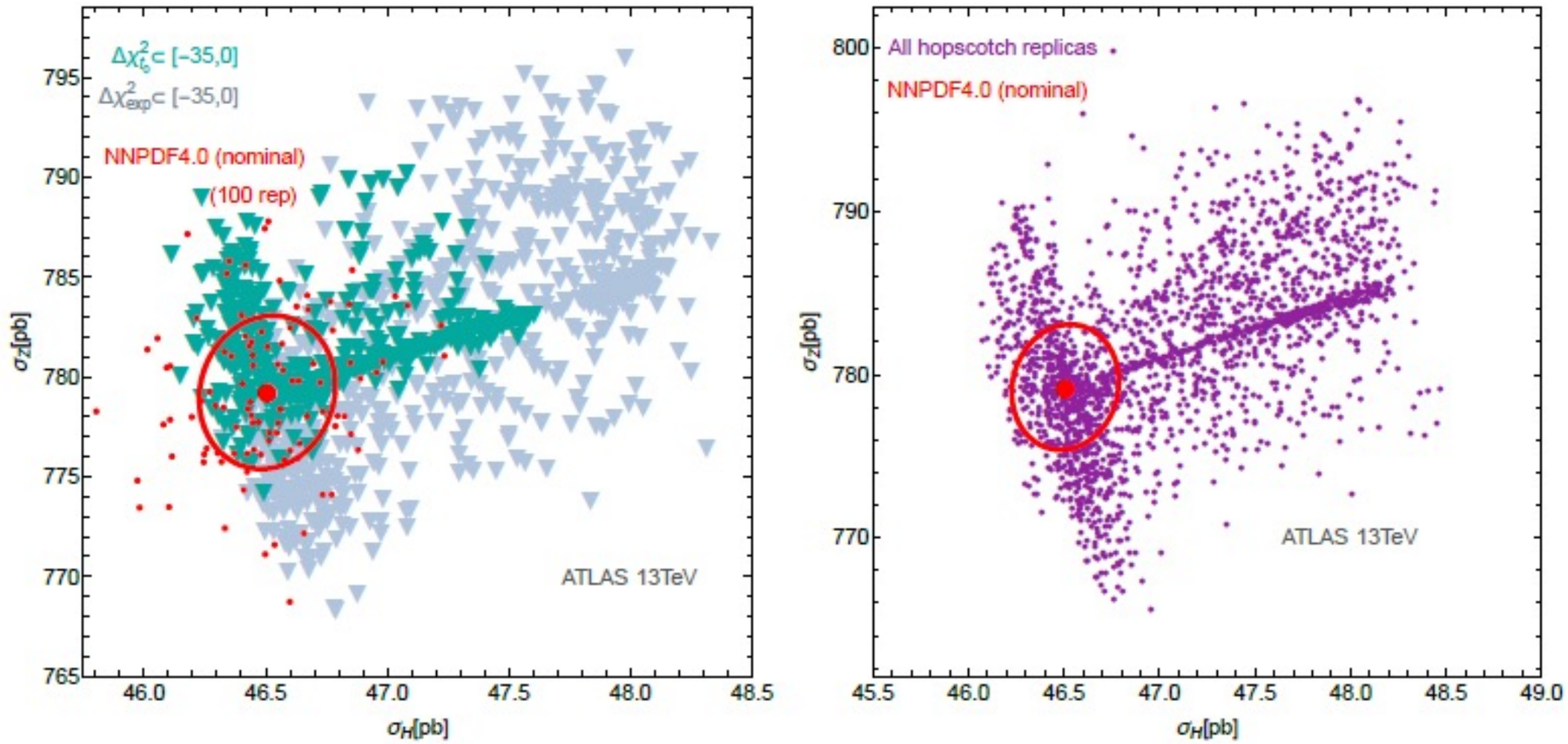
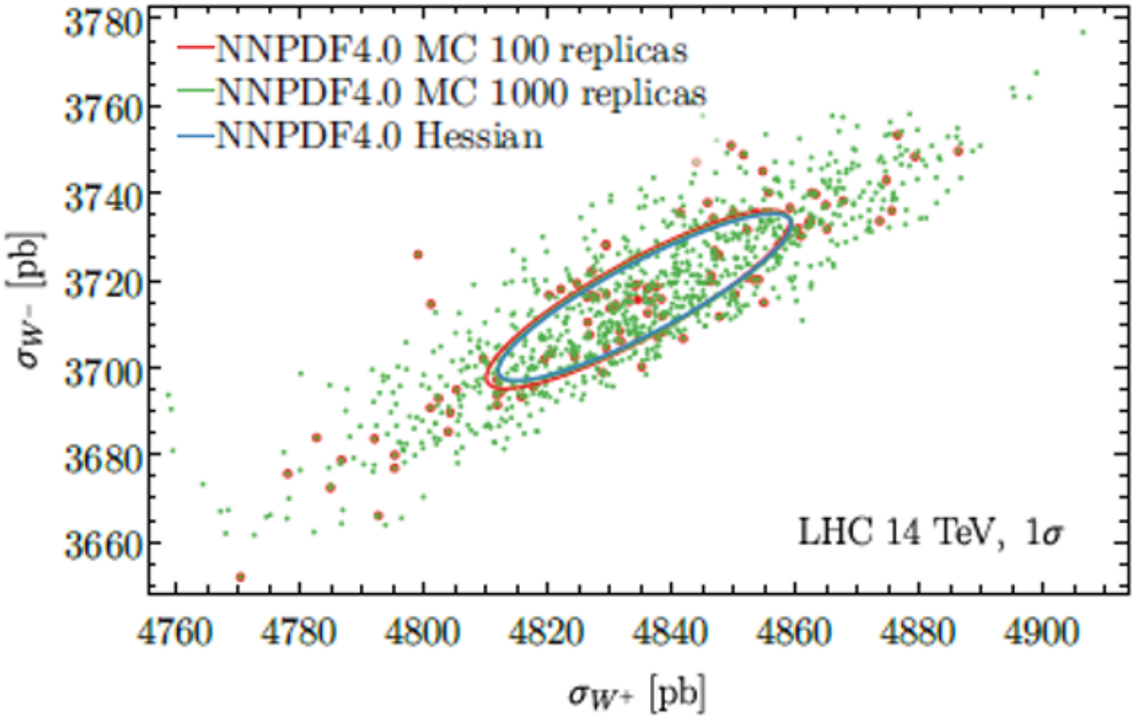
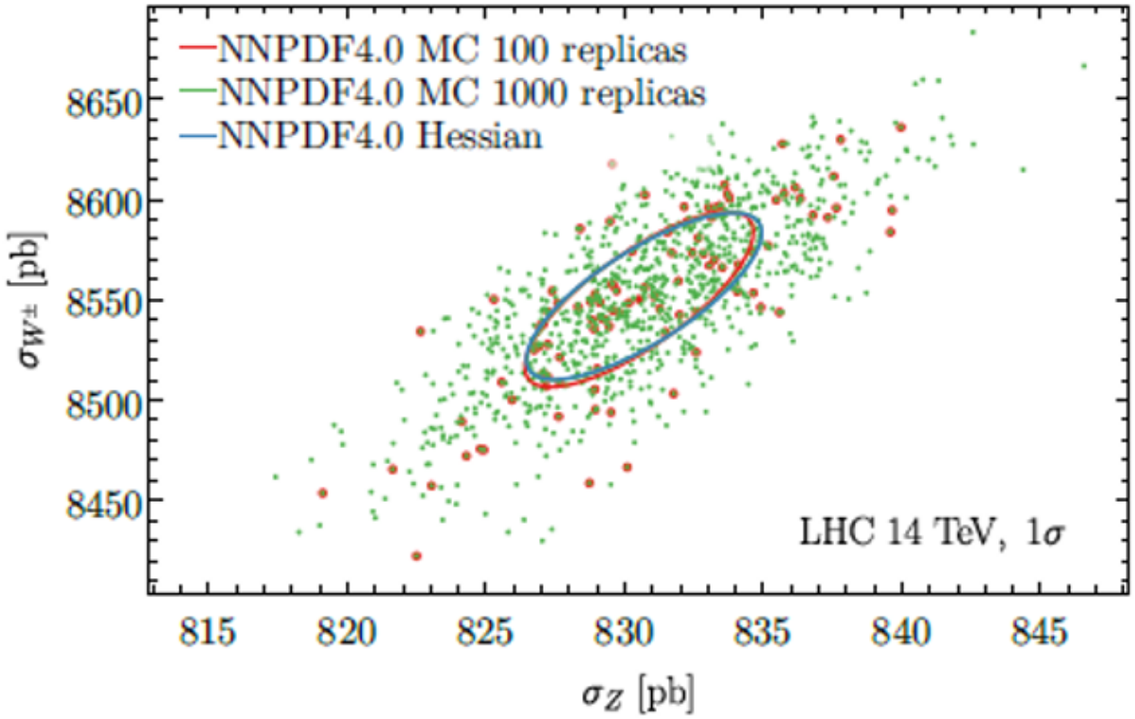
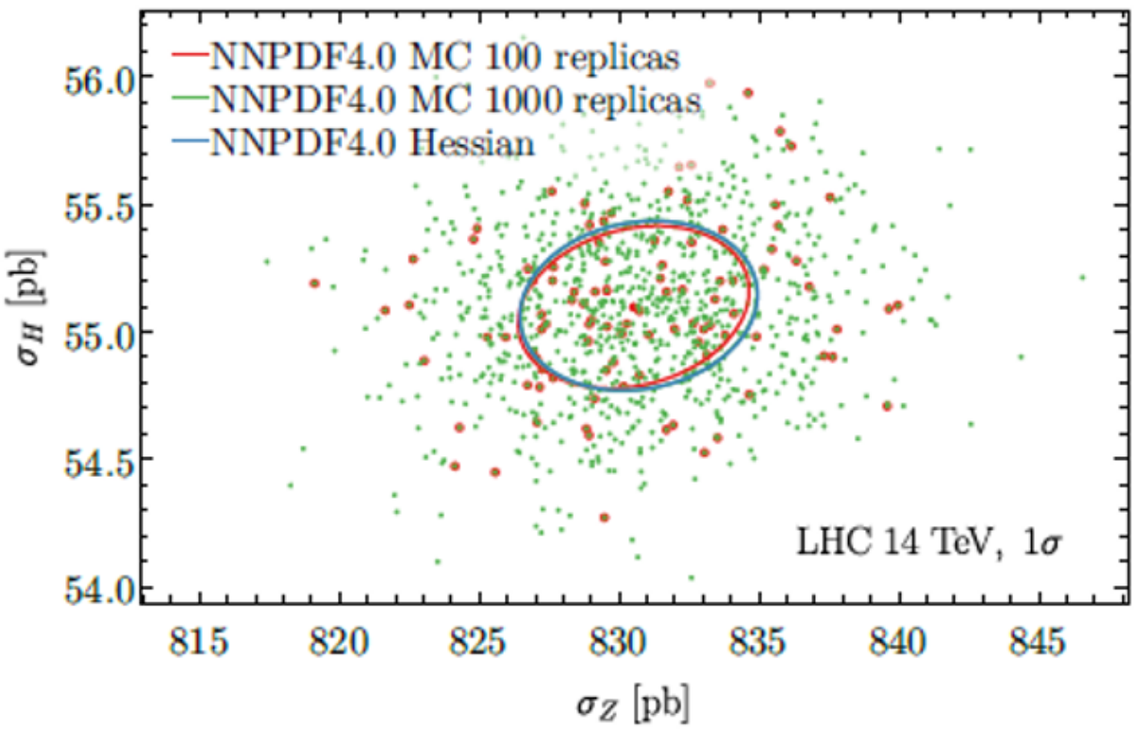


FIG. 7. Left: Hopscotch scan results for the Higgs vs. Z cross section for ATLAS at 13 TeV. Here we show clouds of alternative replicas that have $-35 \leq \Delta\chi^2 \leq 0$ with respect to the NNPDF4.0 central replica, where χ^2 is computed according to the t_0 (cyan) and experimental (grey) definitions. The red points indicate predictions with the 100-replica NNPDF4.0 ensemble. Right: The distribution of 2329 hopscotch replicas for any χ^2.







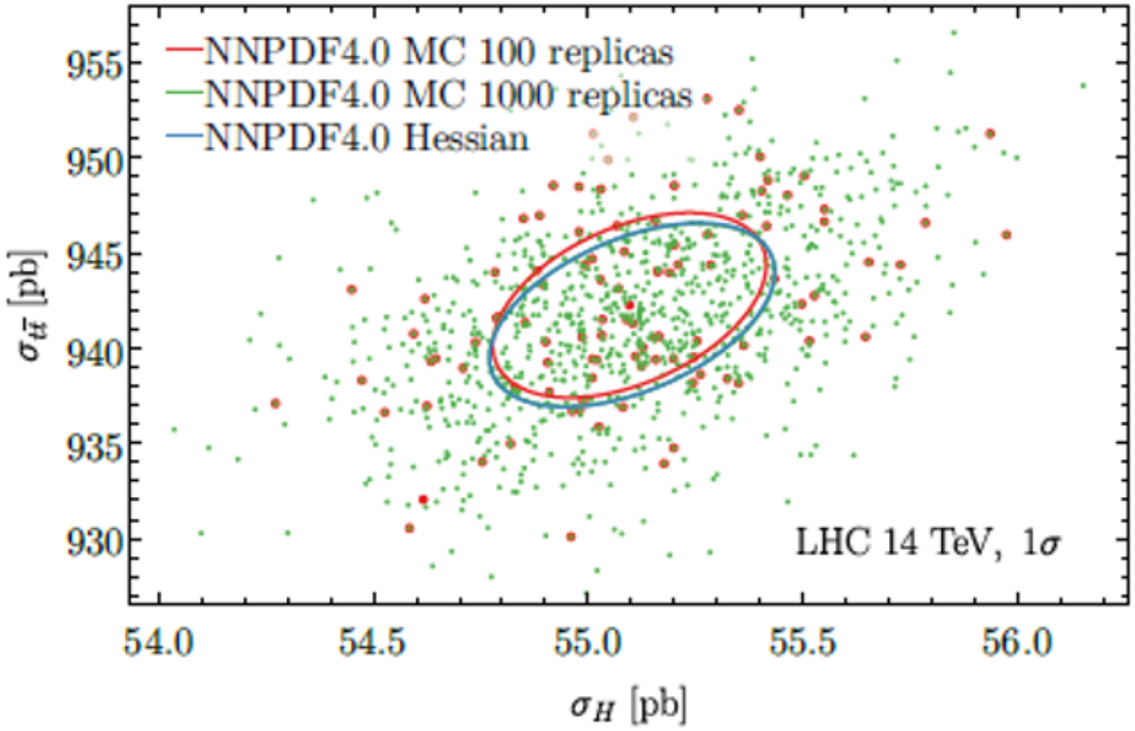


FIG. 8. LHC total cross sections at 14 TeV predicted using the NNPDF4.0 NNLO 1000-replica, 100-replica, and Hessian PDF ensembles. The ellipses indicate 1σ probability regions computed with each ensemble.