Uncertainty quantification for PDFs

Aurore Courtoy

Instituto de Física
Universidad Nacional Autónoma de México (UNAM)

CONACYT
Consejo Nacional de Ciencia y Tecnología

dgapa
Dirección General de Asuntos del Personal Académico

xFitter workshop

CERN/online

05/03/23

Instituto de Física UNAM

Towards epistemic parton distributions

Mainly based in the following publication

“Parton distributions need representative sampling”
[Phys.Rev.D 107] ***arXiv version more complete***

CTEQ-TEA collaboration

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

and forthcoming studies.

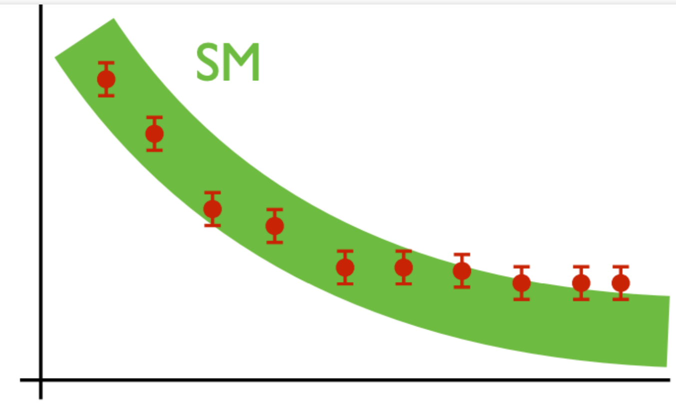
Application of concept of epistemic PDF uncertainties — next talk by L. Kotz

Challenges in global analyses

Keynote talks at DIS'23 (3 weeks ago)

Daniel de Florian: need for precision

- Most likely look for "new interactions"
- Small deviations from SM : PRECISION
- EFT description / BSM model



- **Precision** is the name of the game for the next decades (Higgs sector)

Marteen Boonekamp: need for accuracy

- Experiments WELCOME the ongoing inclusion of theoretical uncertainties in PDF fits.
- Still, very difficult to understand the significance of differences between results obtained using different PDF sets
  - Very interesting discussion in WG1
  - better uncertainty decomposition required

3

- M_W is such an active field, all of a sudden!
- Uncertainty propagation for this measurement currently *almost* broken by the PDFs – we should improve, and the discussions this week were extremely helpful

23

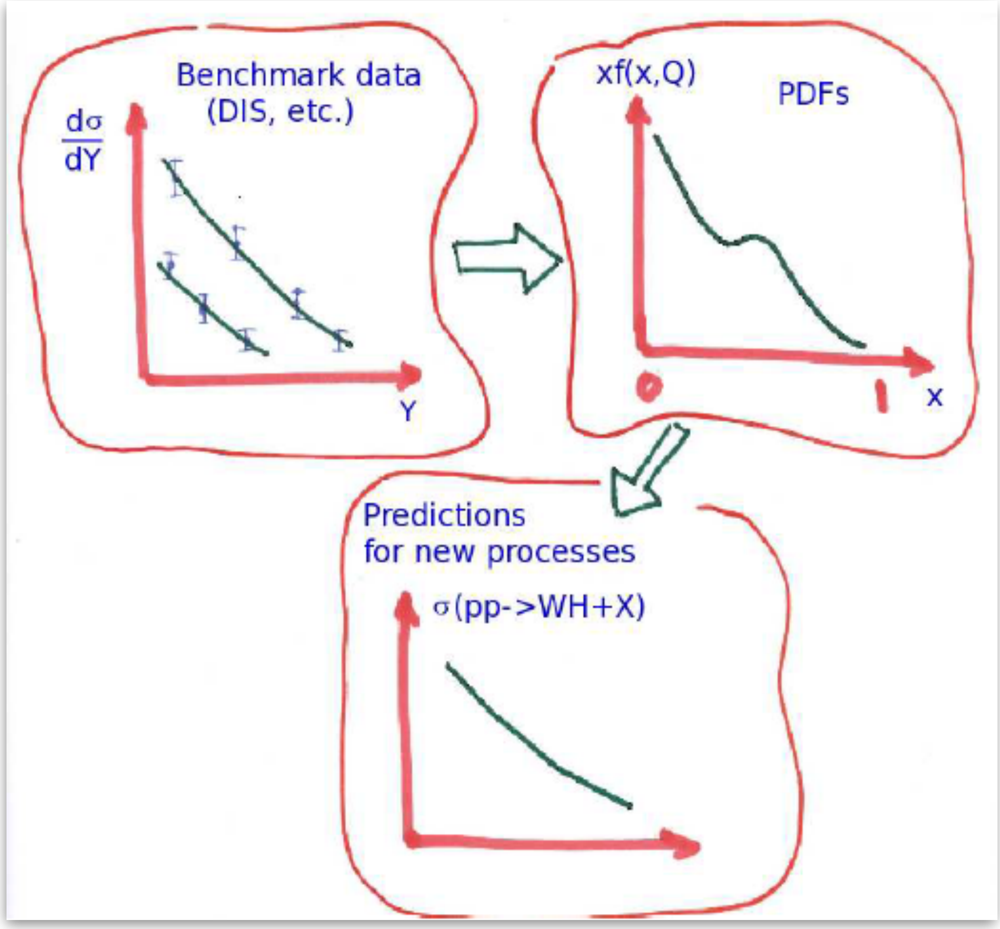
The shape of parton distributions

Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Shape in x extracted from data that are sensitive to specific PDF flavors, etc.

I. hints of behavior of partons at low scales

II. predictions for other (new) processes



Classes of *first principle* constraints for x-dependence

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

Model evaluation of x-dependence (in parallel to data learning)

- use QFT description of $f(x)$ together with model description of hadron wave function (non trivial to define)
- ensure symmetries are fulfilled

The shape of parton distributions

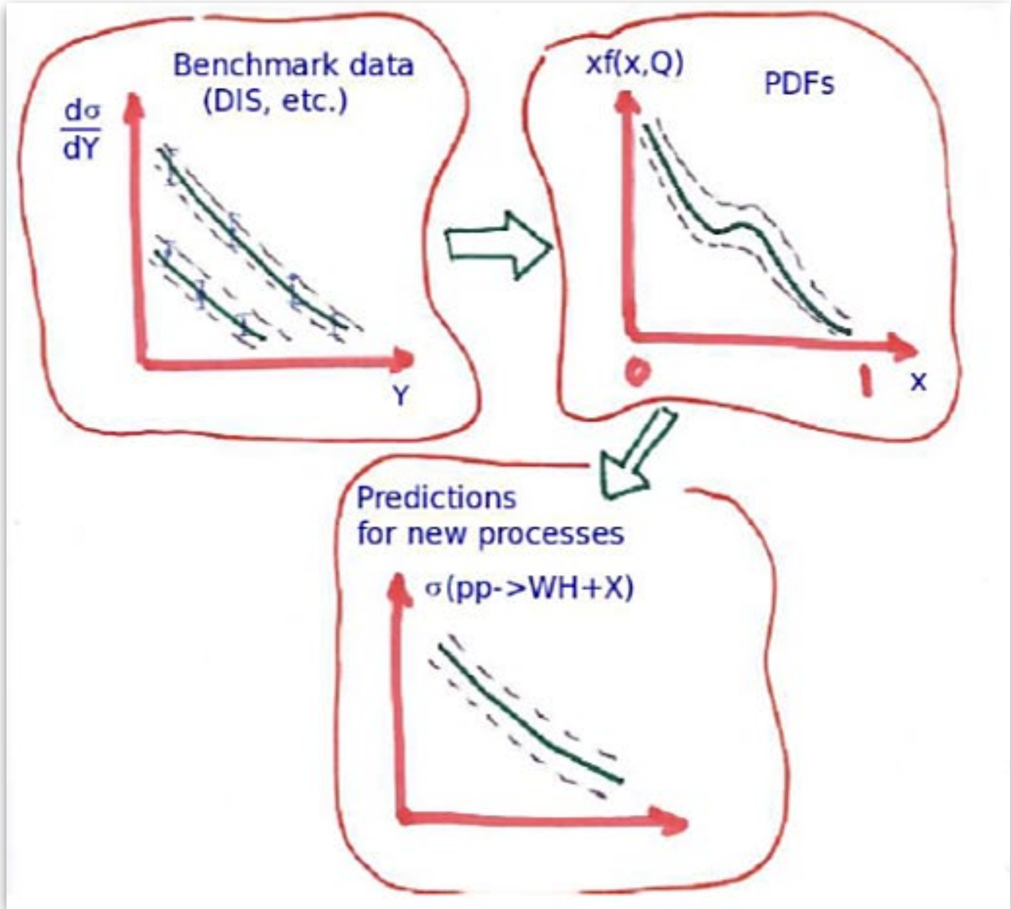
Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Uncertainty propagates from data and methodology to the PDF determination

I. assessment of uncertainty magnitude is key

II. advanced statistical problem

III. evolving topic in the era of AI/ML



The shape of parton distributions

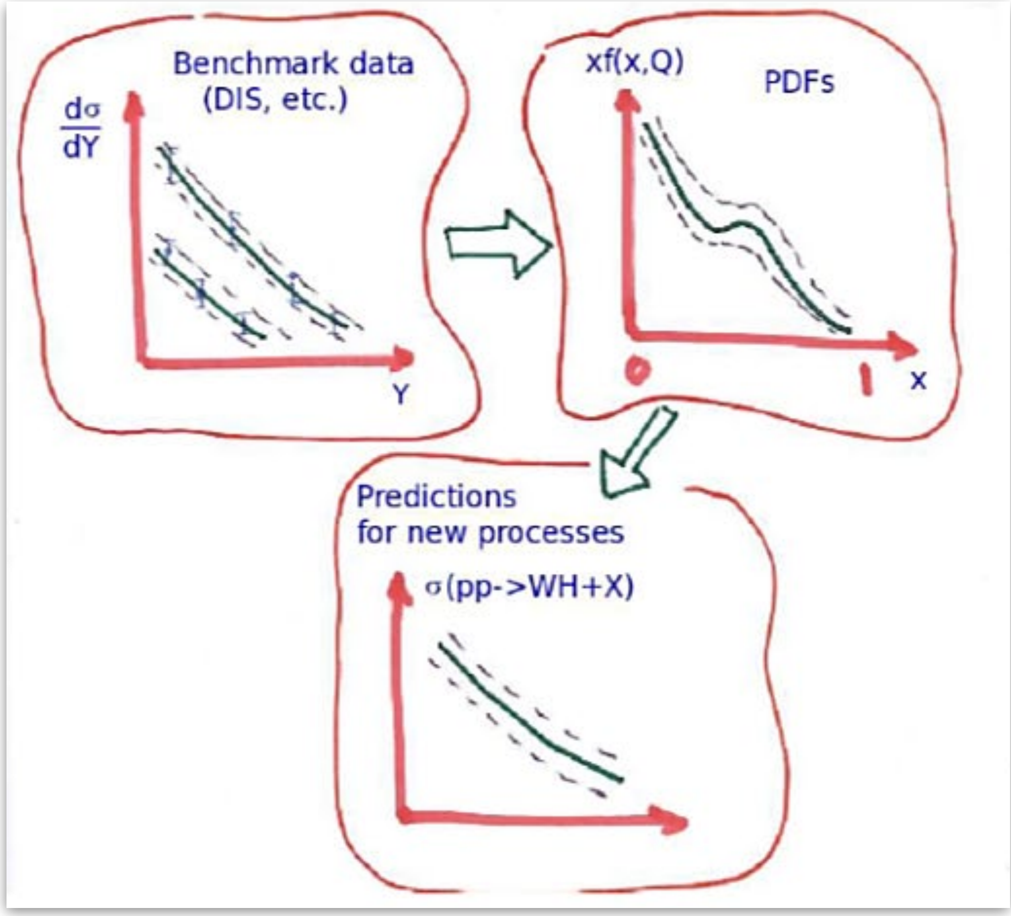
Low-energy QCD dynamics, encapsulated in PDFs, are learned from experimental data.

Uncertainty propagates from data and methodology to the PDF determination

I. assessment of uncertainty magnitude is key

II. advanced statistical problem

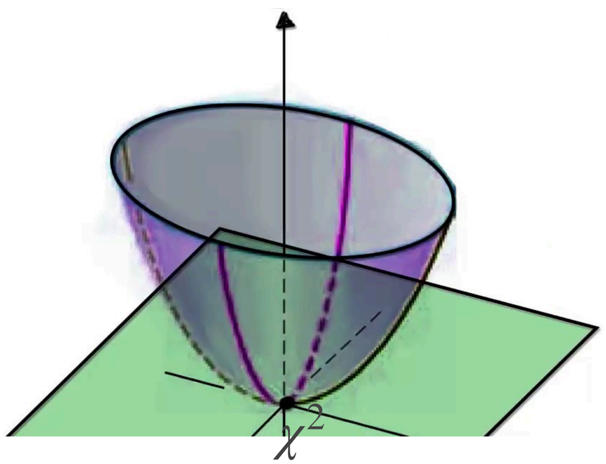
III. evolving topic in the era of AI/ML



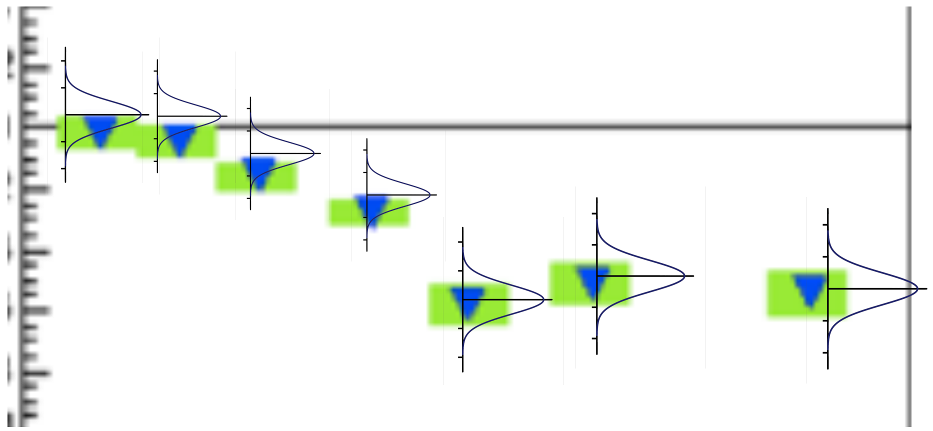
Epistemic vs. aleatory uncertainties

Uncertainty due to lack of knowledge —bias (may be reduced)

Statistical uncertainty propagated from experiments — irreducible

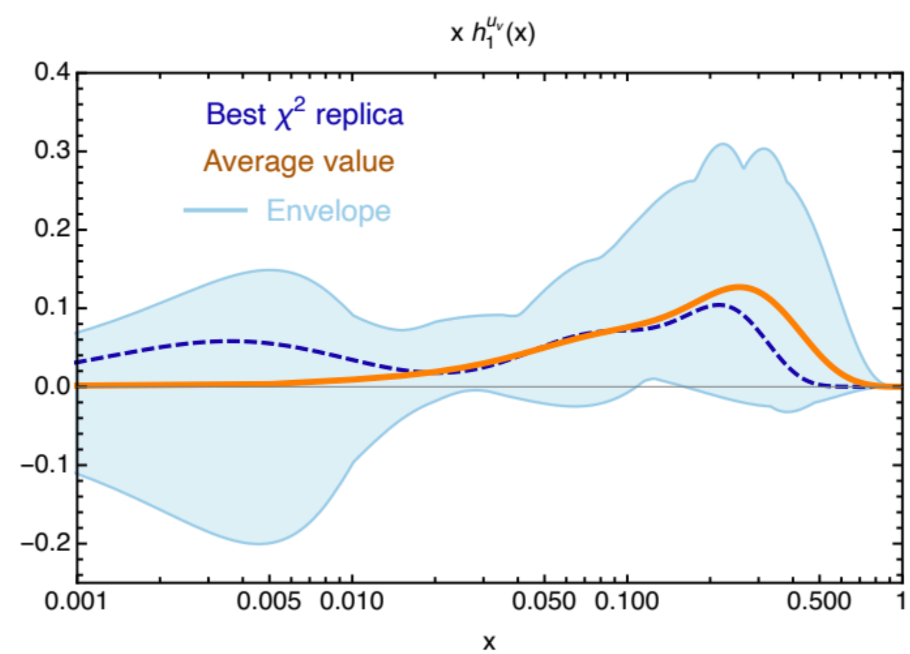


Hessian methodology finds the global minimum and explores the parameter space.



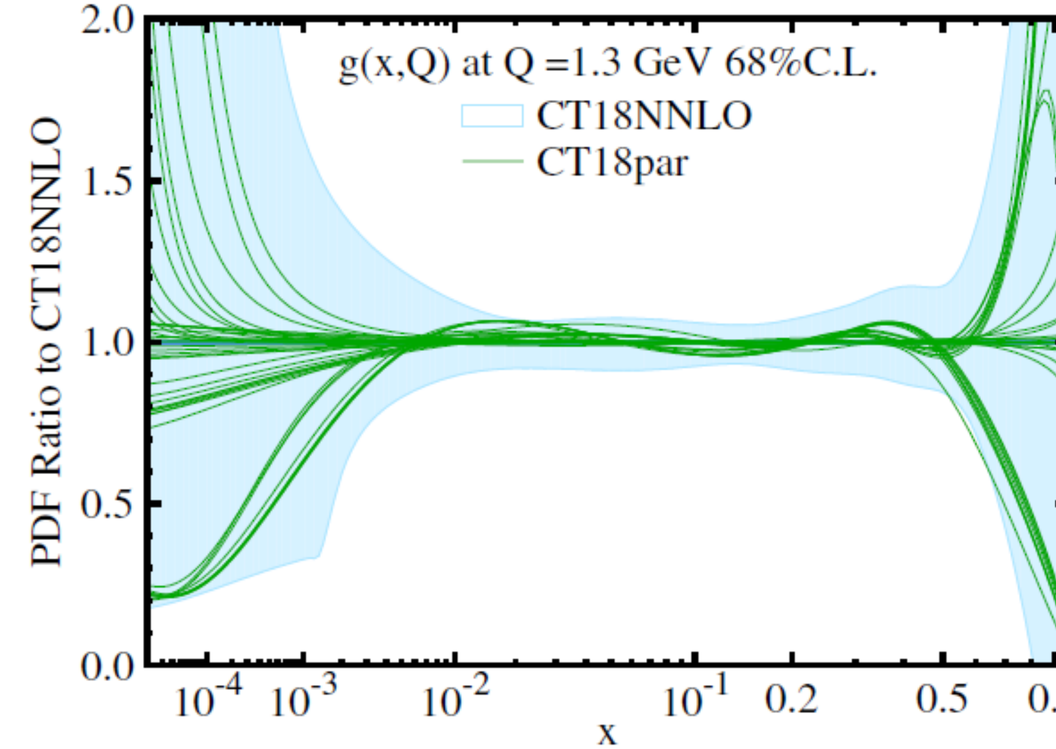
Monte-Carlo methodology (neural network, AI/ML) replicates fluctuated data, then optimizes each replica (up to training).

Illustration on *bootstrap* probability distribution with average value vs. the best replica for $N_{par} \sim 7$.



In multivariate analyses, sampling occurs at various levels — parameter space, bootstrap but also priors, … In large-dimensional problems, sampling is complex.

Epistemic uncertainties in global analyses

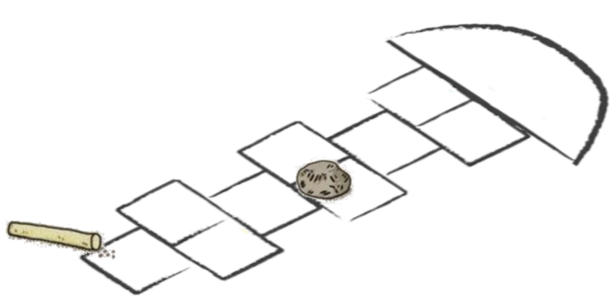


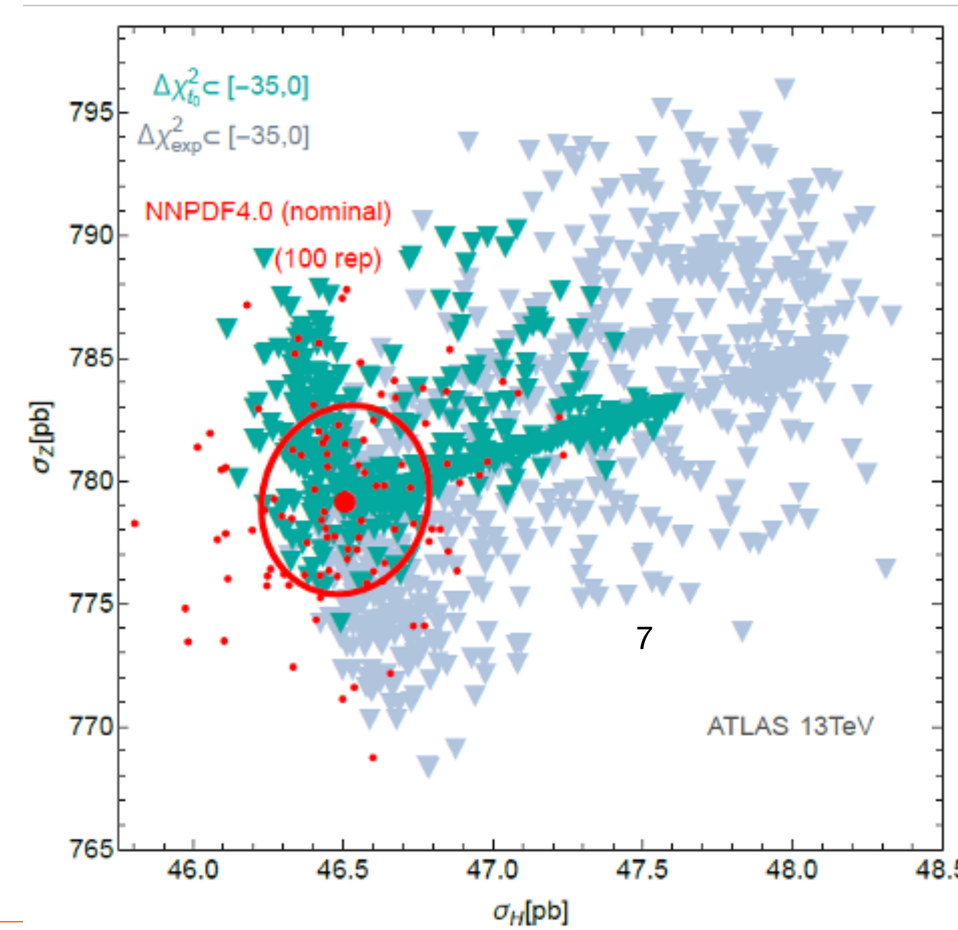
<u>CT18 PDF uncertainty:</u>

Hessian-based methodology
Inclusive of sampling bias/lack of knowledge

<u>Monte Carlo-based PDF uncertainty:</u>

Higher-dimensional space
The “hopscotch” algorithm quantifies bias due to lack of knowledge





Hypothesis testing and parton distributions

Hypothesis testing of theoretical predictions relies on

1. available data in x range, as well as value of Q,
2. sensitivity of data to the hypothesis,
3. quality of the data,
4. uncertainties found in the fits.

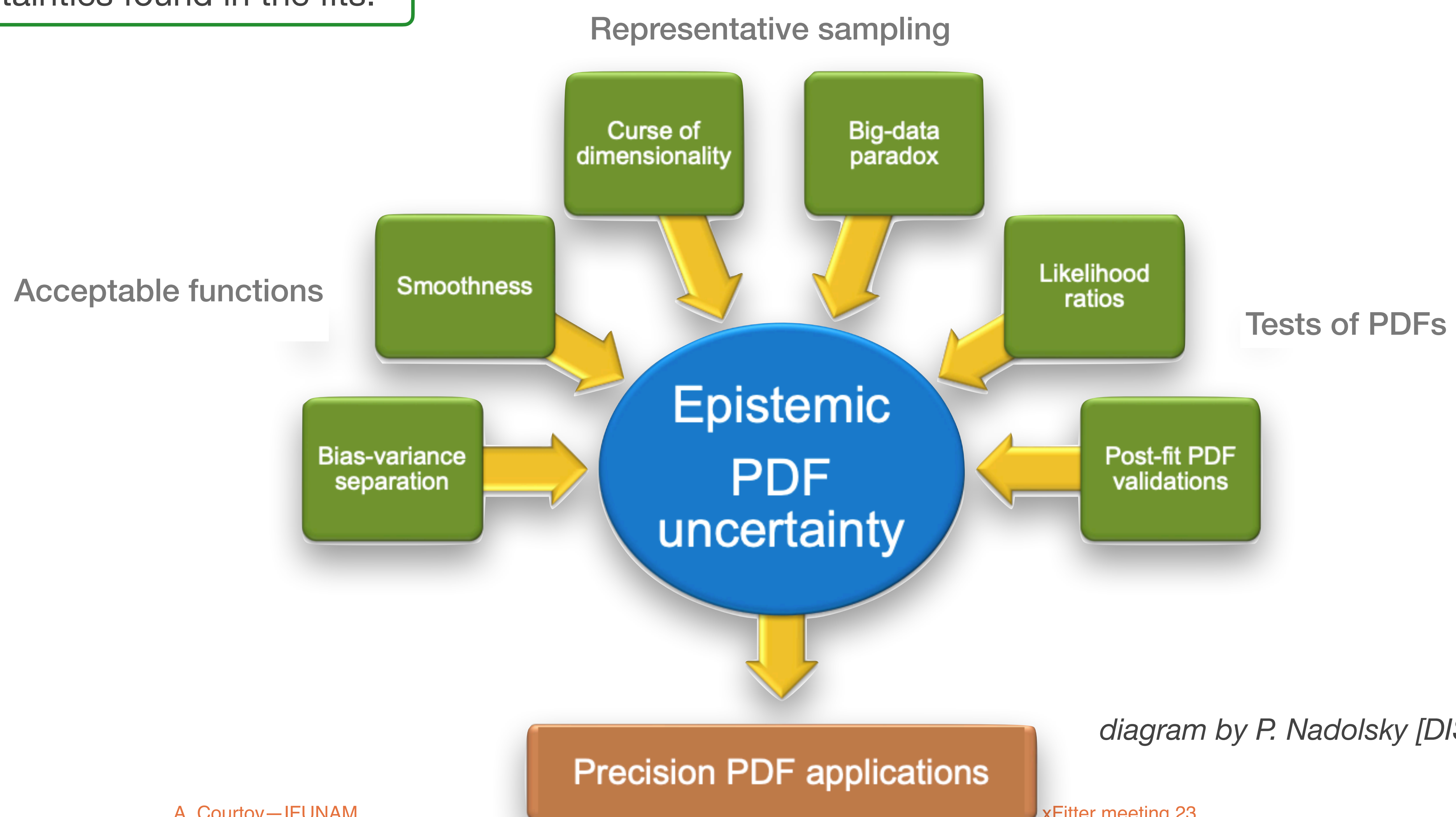


diagram by P. Nadolsky [DIS2023]

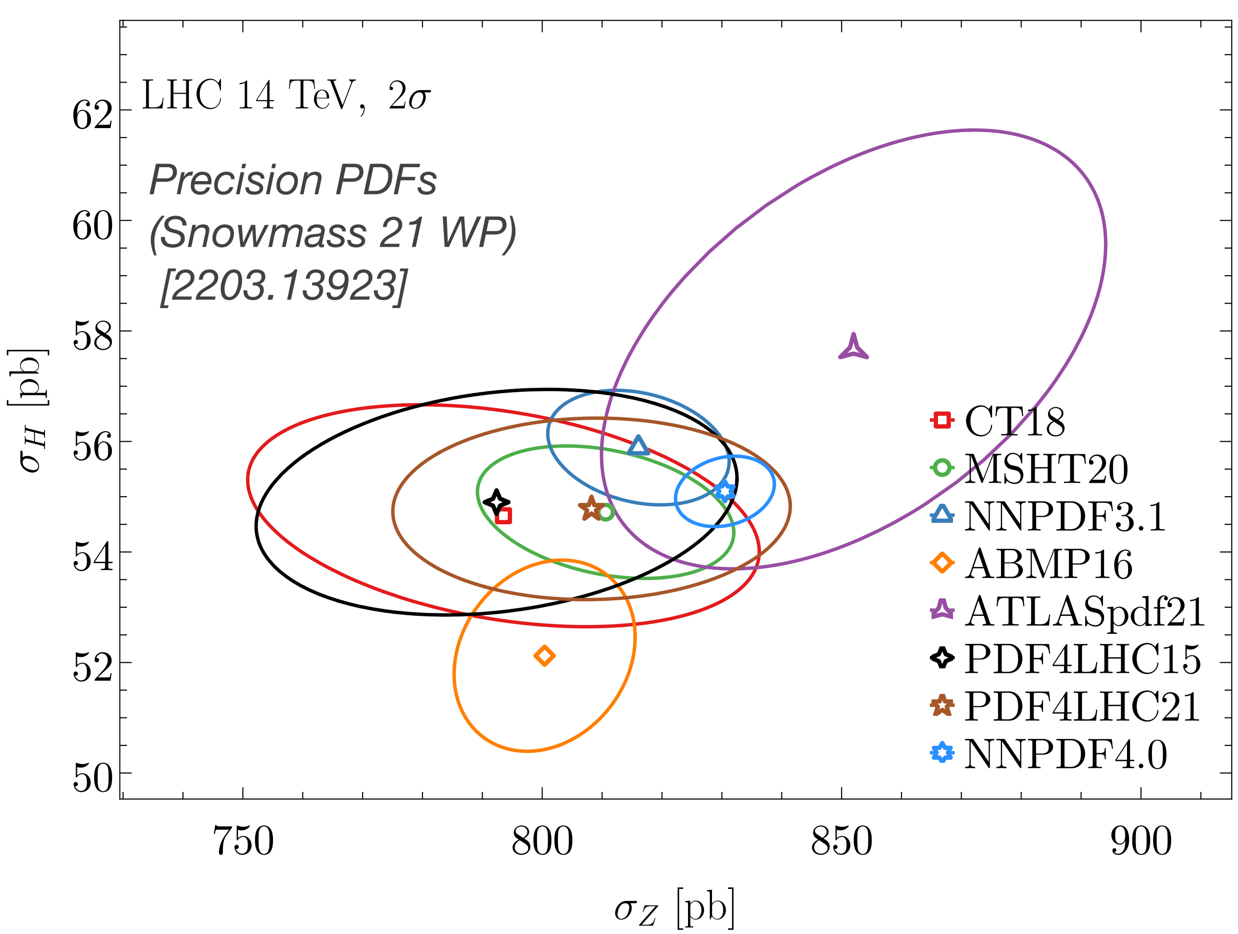
Criteria for PDF uncertainties

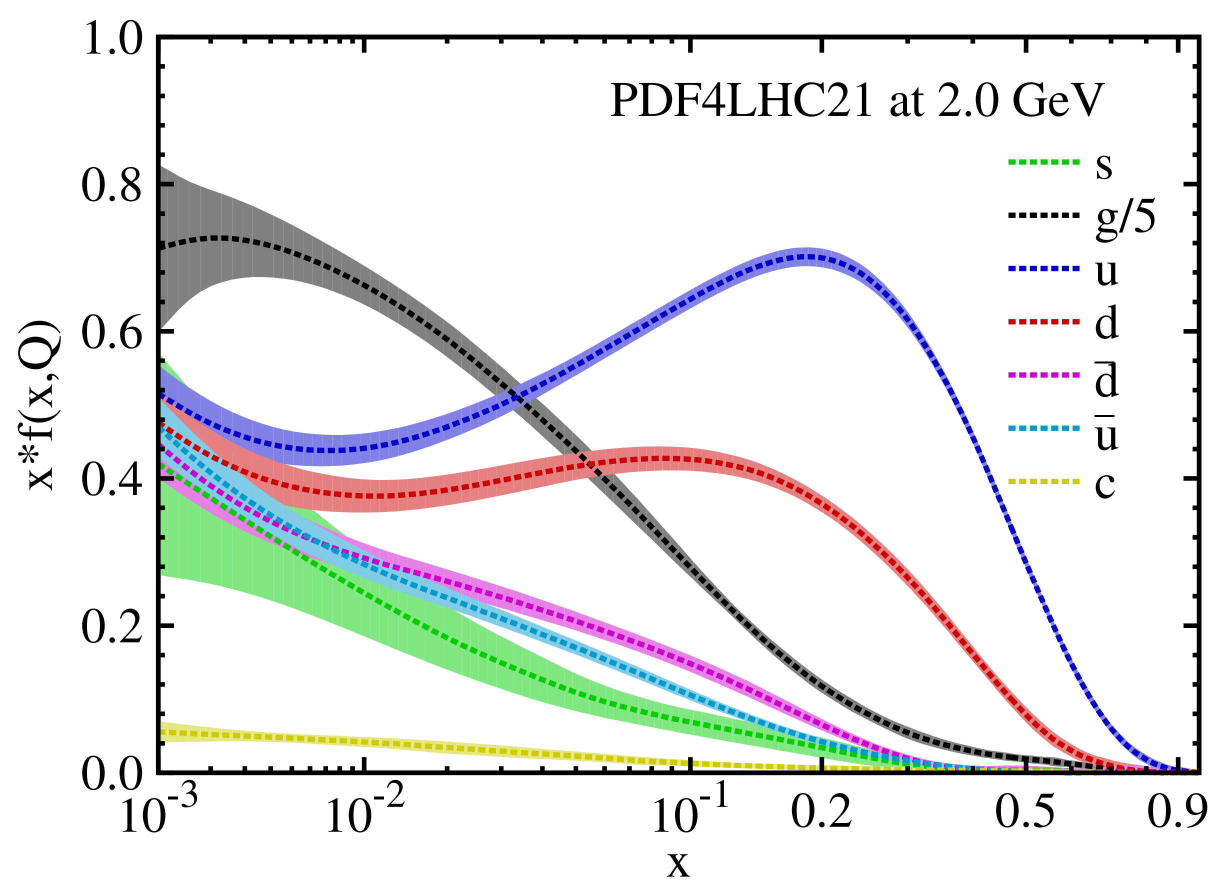
Recent advancements in the determination of unpolarized PDFs:
CT18, MSHT20, NNPDF4.0, ATLASpdf21 as well as PDF4LHC21.

PDF4LHC21:

benchmarking and combination of the leader PDF sets, CT, MSHT & NNPDF, for the run III of the LHC.

[Ball, et al, J.Phys.G 49 (2022)]





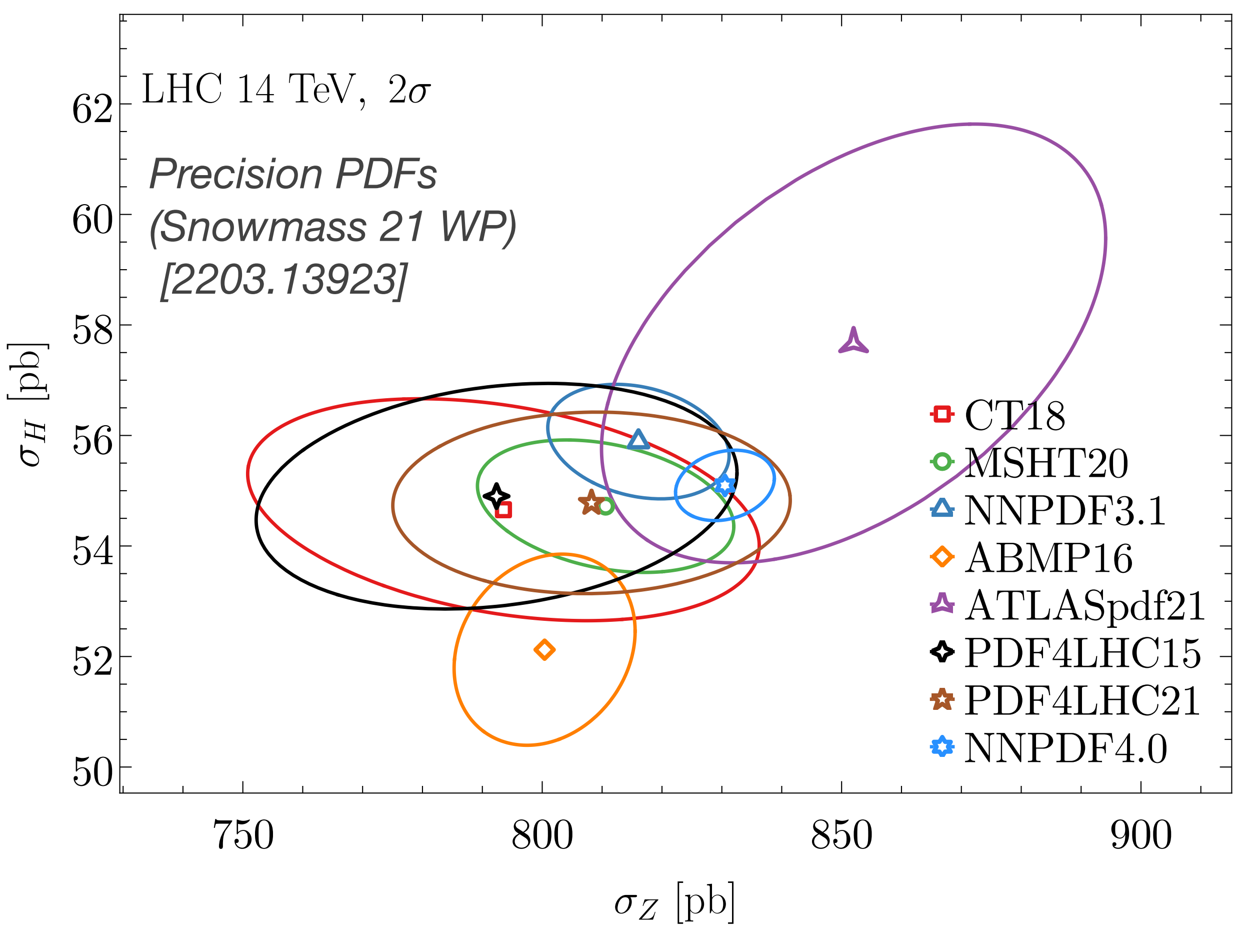
Criteria for PDF uncertainties

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

PDF4LHC21:

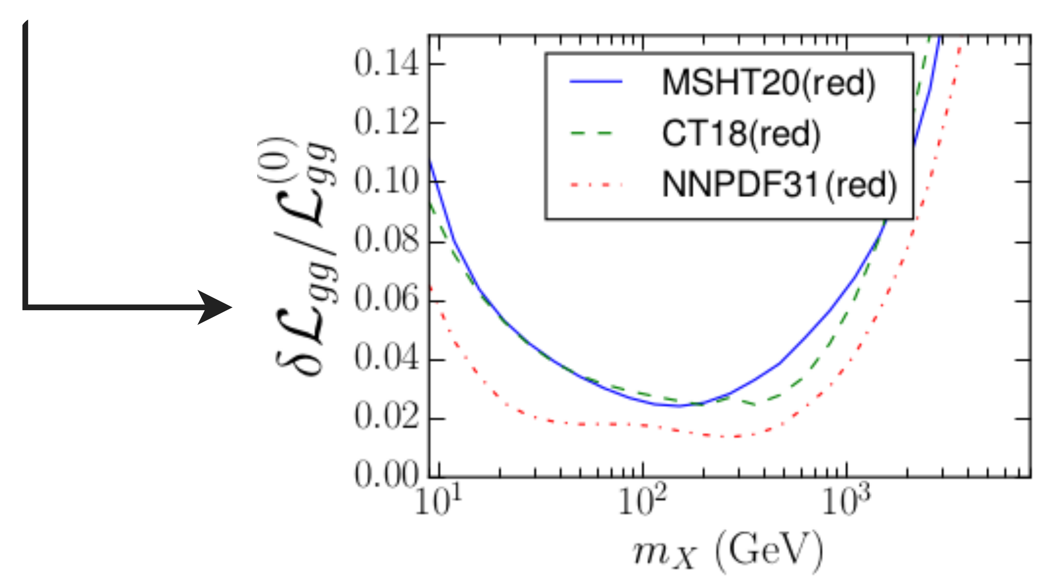
benchmarking and combination of the leader PDF sets, CT, MSHT & NNPDF, for the run III of the LHC.

[Ball, et al, J.Phys.G 49 (2022)]



What is the origin of the differences in size of correlation ellipses among various fits?

PDF4LHC21 excercise highlights the role of methodology. Monte Carlo-based analysis (NNPDF) gives smaller uncertainties.



Key role played by methodology

Outside of HEP/NP, there is significant interest in statistical problems that are similar to the assessment of uncertainties for PDF.
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: 18 June 2021

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5,12]

SCIENCE ADVANCES | RESEARCH ARTICLE

MATHEMATICS

Models with higher effective dimensions tend to produce more uncertain estimates

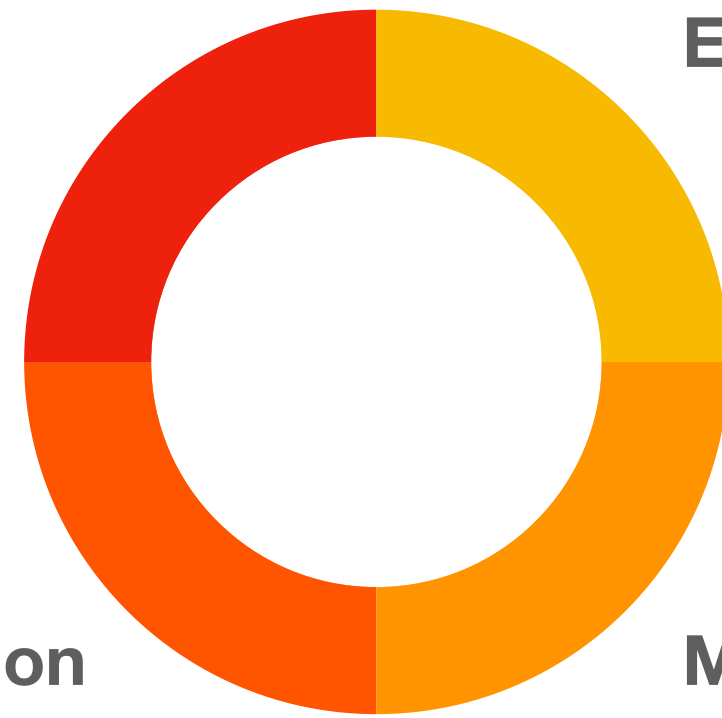
Arnald Puy[1,2,3]*, Pierfrancesco Beneventano[4], Simon A. Levin[2], Samuele Lo Piano[5], Tommaso Portaluri[6], Andrea Saltelli[3,7]

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

On uncertainty quantification



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

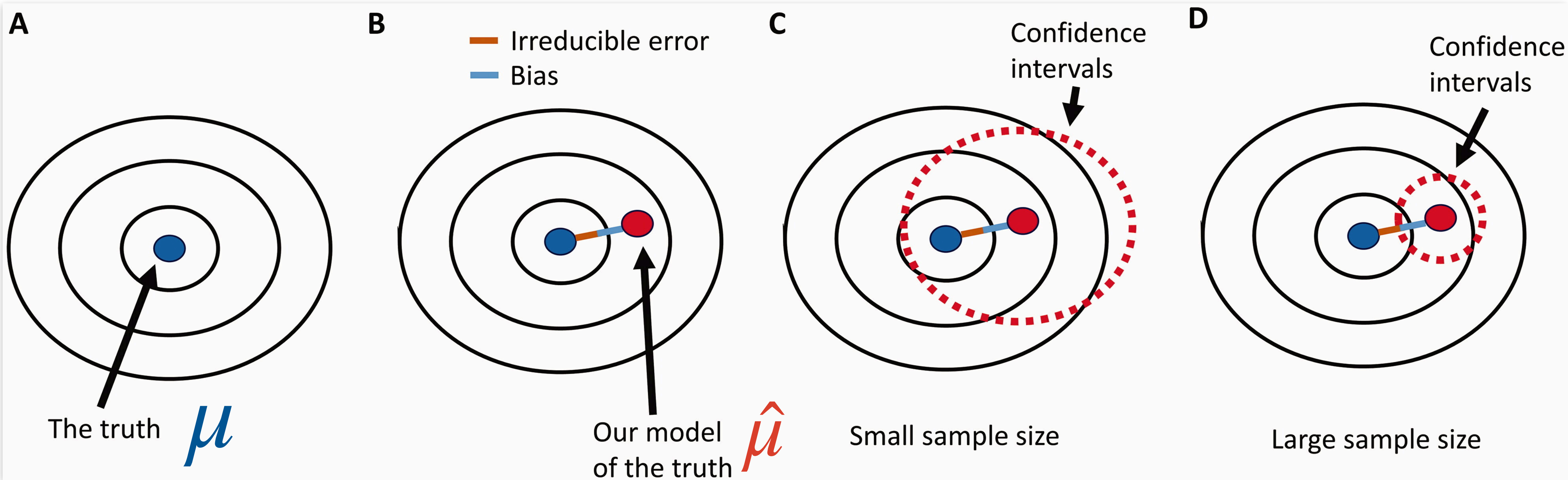
Goodness-of-fit applies to an individual best fit.

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

→ This talk.

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

Sampling bias and big-data paradox

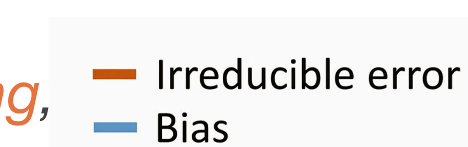


Pavlos Msaouel (2022)
Cancer Investigation, 40:7, 567-576

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*, 

.

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

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

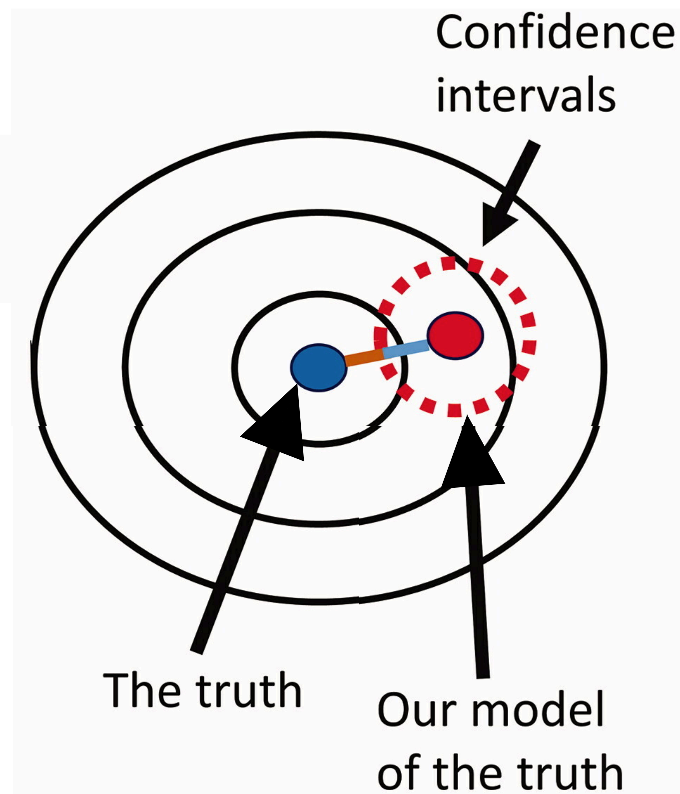
depends on the sampling algorithm

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

Irreducible error
Bias

≡ statistical model, quality of data,…

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

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

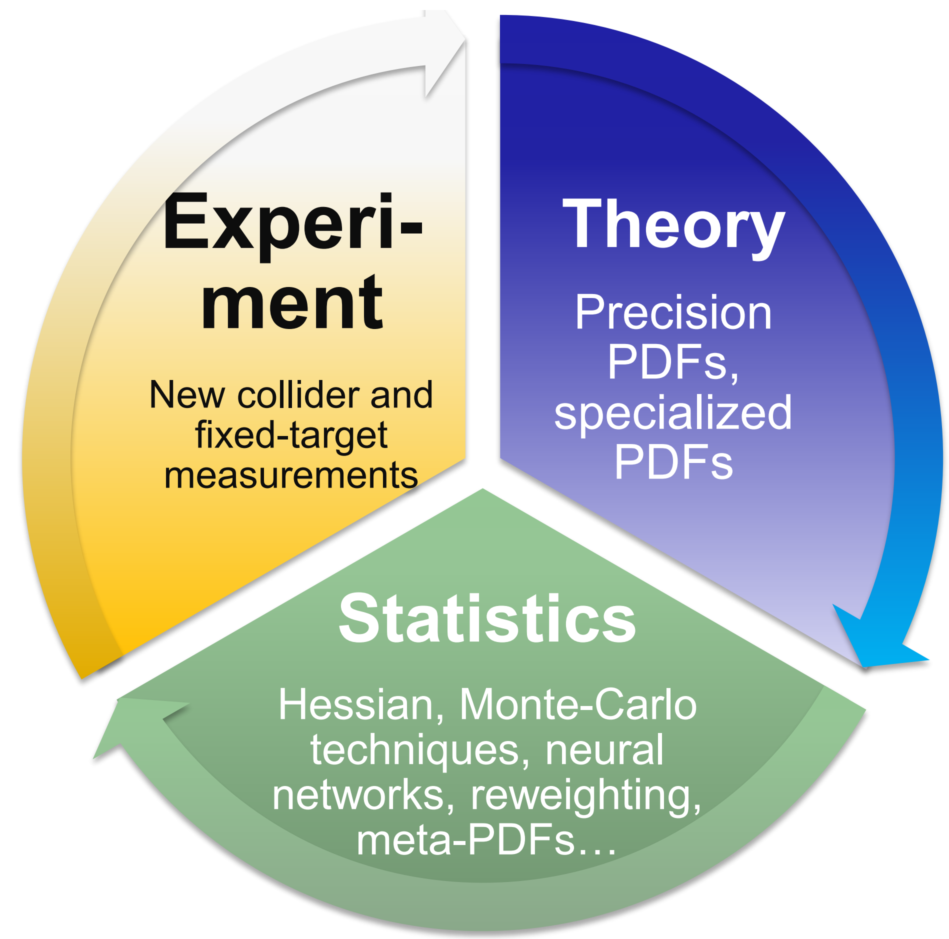
Sampling bias in PDF global analyses—I

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

Methodological choices are reflected through the epistemic uncertainty, including biases from sampling.

Priors, including choice of functional form or Bayesian *priors*, contribute to restrict the sampling quality.

Representative sampling accounts for the confounding correlation, and can ultimately be used to optimize its contribution, e.g. through the study of largest effective dimensions.



⇨ dimensionality reduction (effective dimensions) vs. phase space reduction (priors)

Sampling bias in PDF global analyses—I

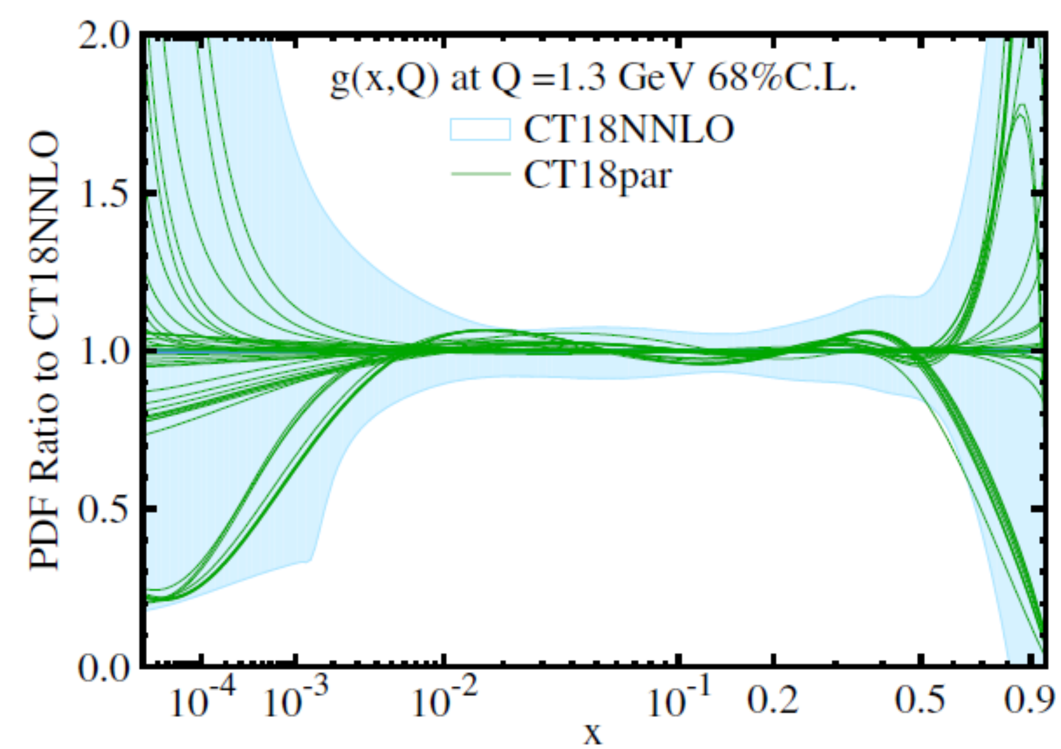
How do we know the "data+sampling defect=confounding correlation" of our analysis?

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

On which basis are PDFs accepted or rejected?

Likelihood ratios:

two replicas can be ordered according to their relative likelihood or relative prior.

$$\frac{P(T_2|D)}{P(T_1|D)} = \frac{P(D|T_2)}{P(D|T_1)} \times \frac{P(T_2)}{P(T_1)}$$

$\equiv r_{\text{posterior}}$ $\equiv r_{\text{likelihood}}$ $\equiv r_{\text{prior}}$

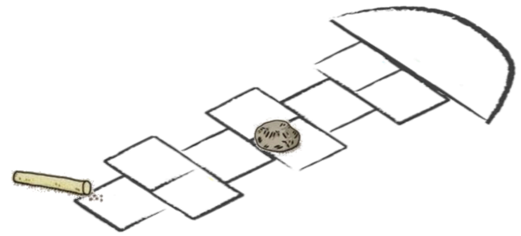
aleatory epistemic + aleatory probabilities

Prior: replica can be discarded based on $P(T_2) < P(T_1)$ even for $r_{likelihood} \sim 1$

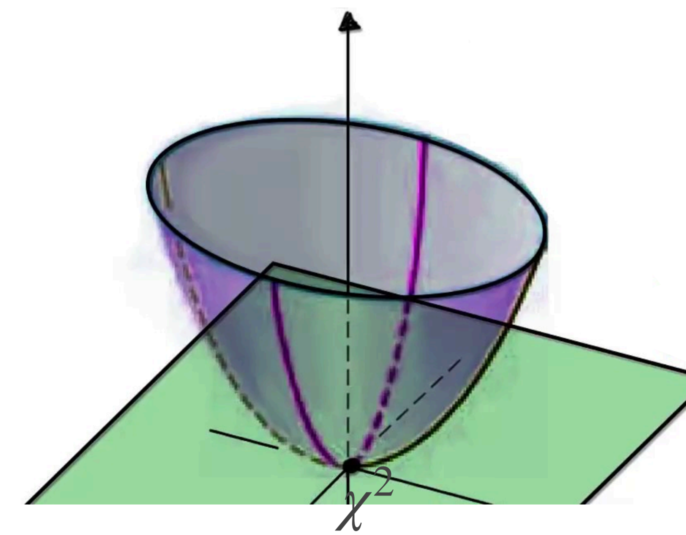
Likelihood: replica can be accepted based on $r_{likelihood} = \frac{P(D\,|\,T_2)}{P(D\,|\,T_1)} \sim 1$ when $P(T_2) \sim P(T_1)$

talk by J. Huston

How to play hopscotch?

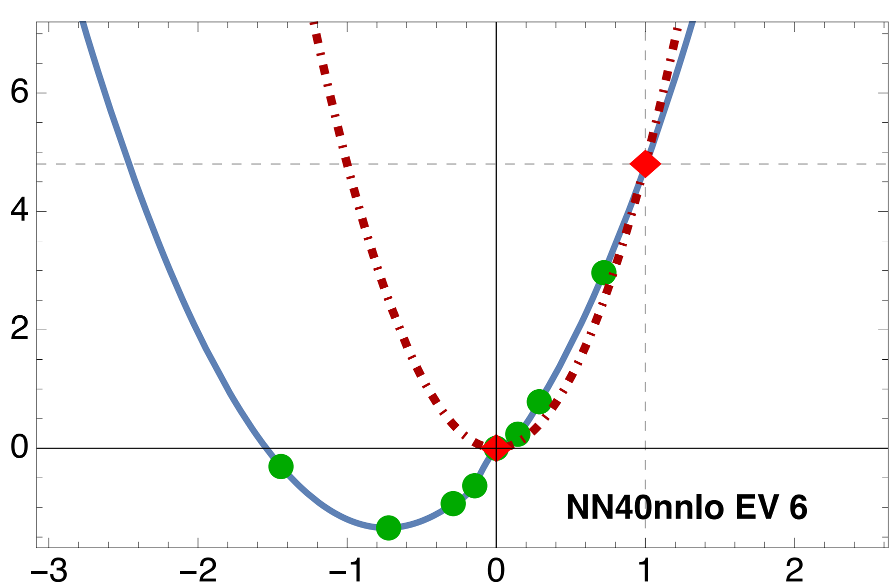


In the Hessian representation, the chi square behaves like a paraboloid of n_{param} dimensions, thus defining a global minimum.



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.

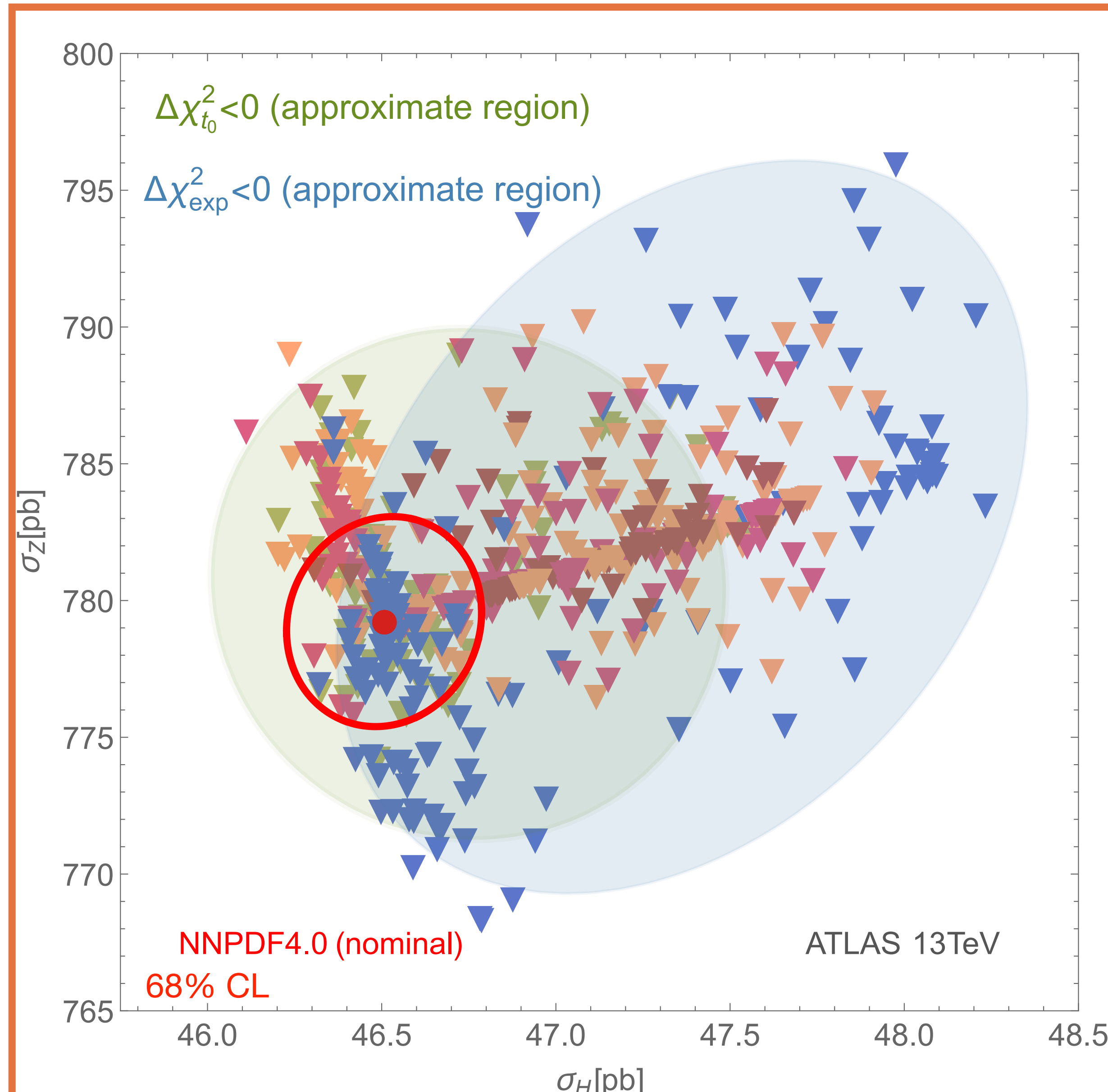
A hopscotch scan of LHC cross sections for NNPDF4.0 PDFs

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 the statistical meaning related to a likelihood-ratio test.

[Anwar, Hamilton, Nadolsky, 1901.05511]

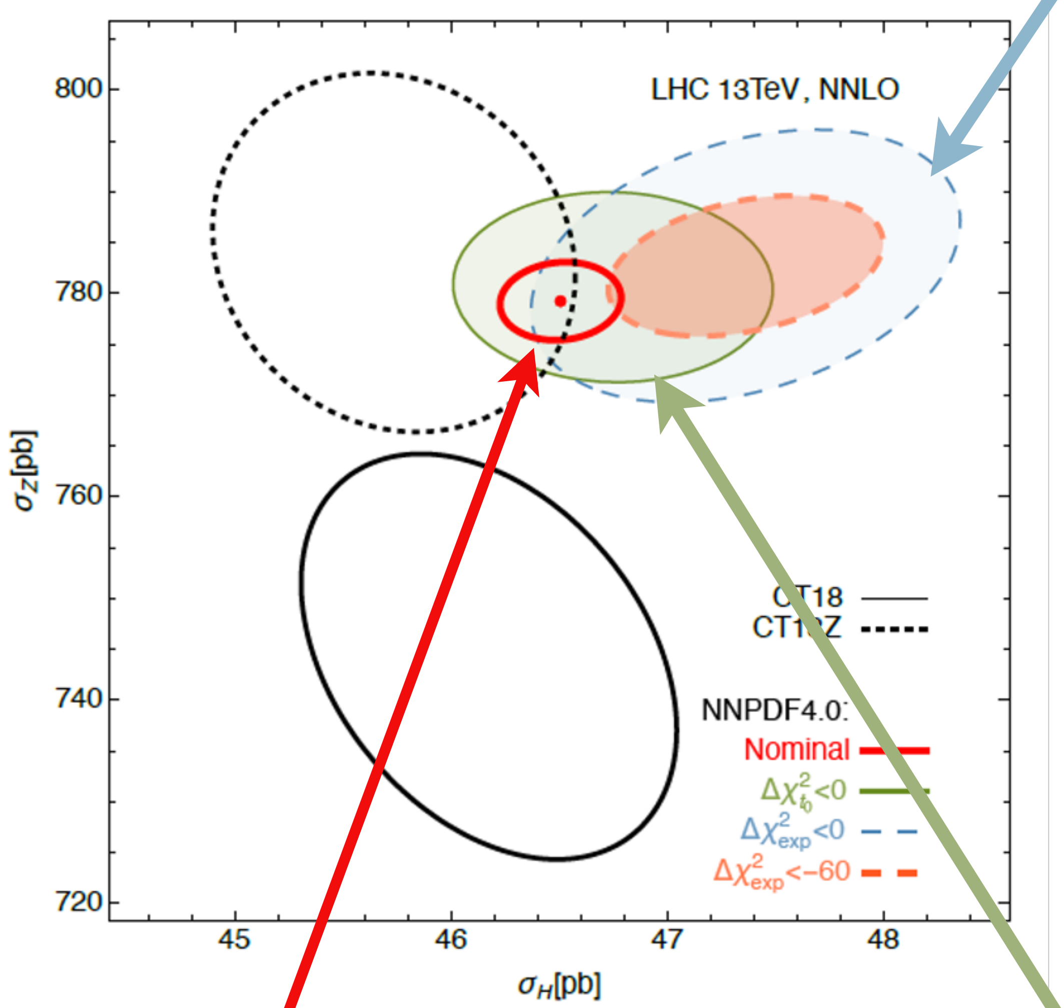
The green and blue areas are larger than the nominal NNPDF4.0 uncertainty (red ellipse).

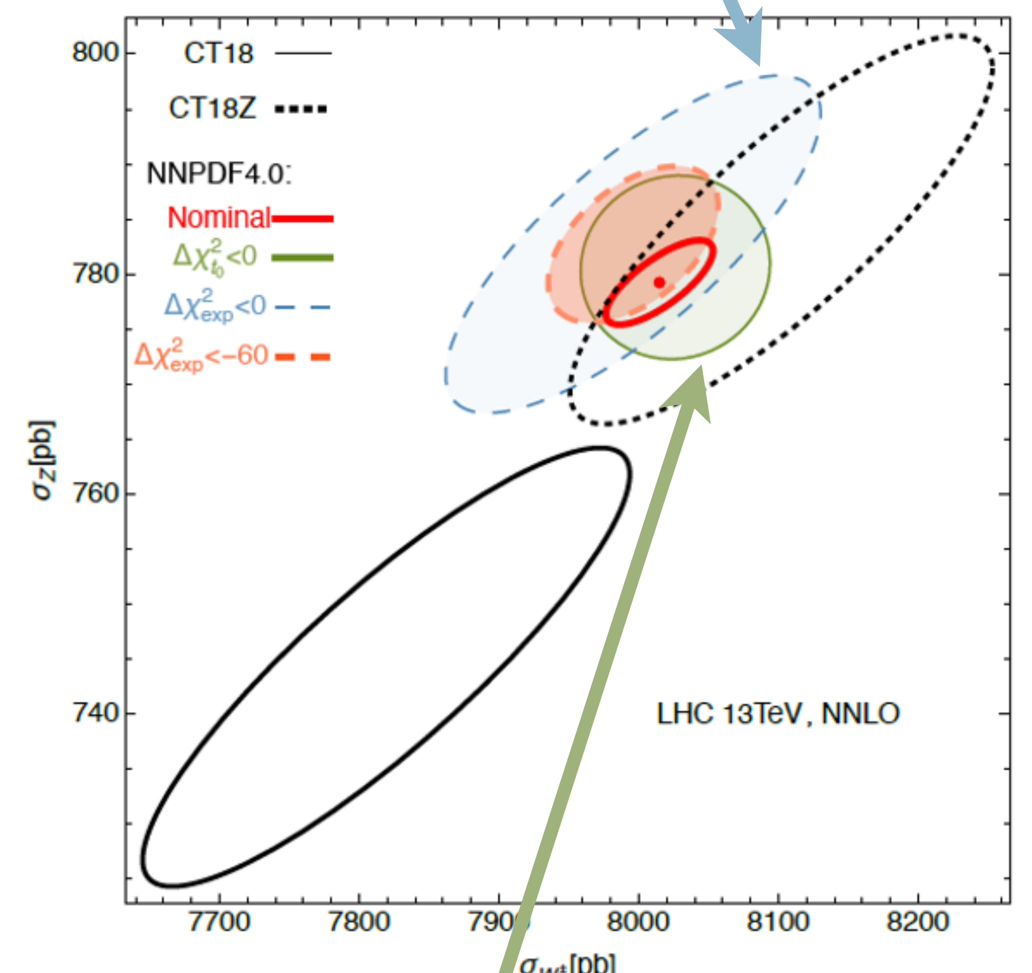


Monte-Carlo sampling sensitivity for PDFs

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

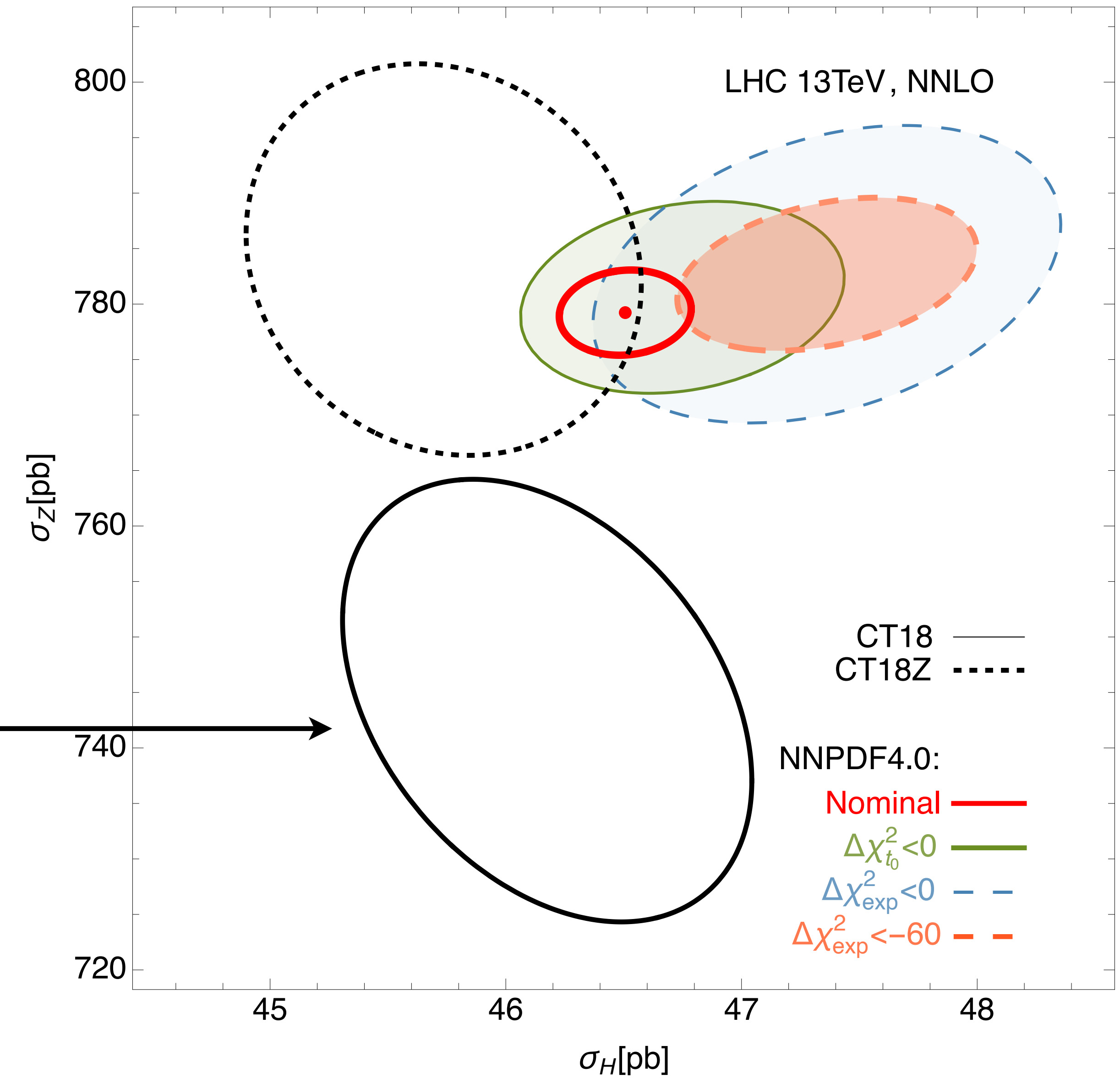
Region containing good solutions according to the most recent t_0 form of χ^2

(used to train NN4.0 replicas)

Representative sampling — the hopscotch algorithm

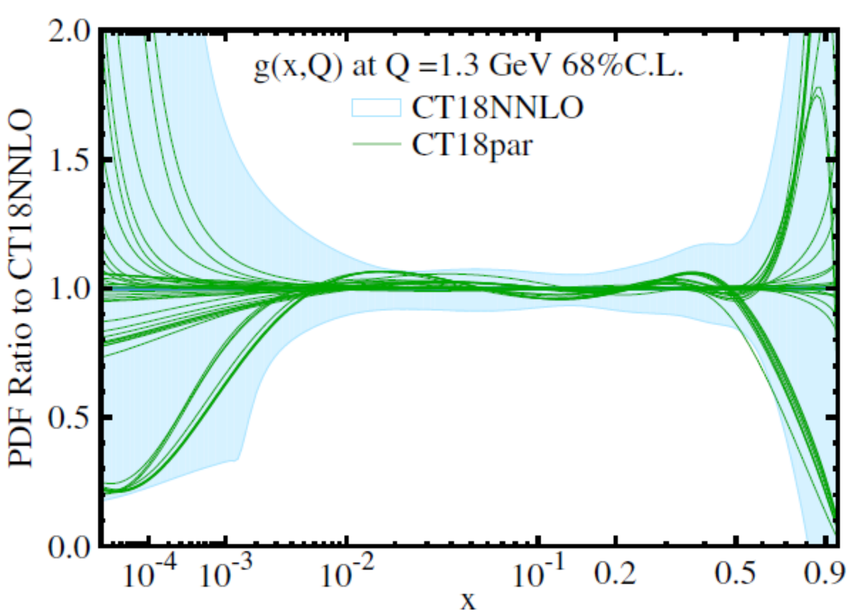
*[**AC**, Huston, Nadolsky, Xie, Yan & Yuan, 2205.10444, PRD107]*

We have devised an algorithm that focuses on the effective dimensions relevant for observables, to challenge Monte Carlo-based analyses. The resulting uncertainty is larger than the nominal one, shown here for (σ_H, σ_Z).



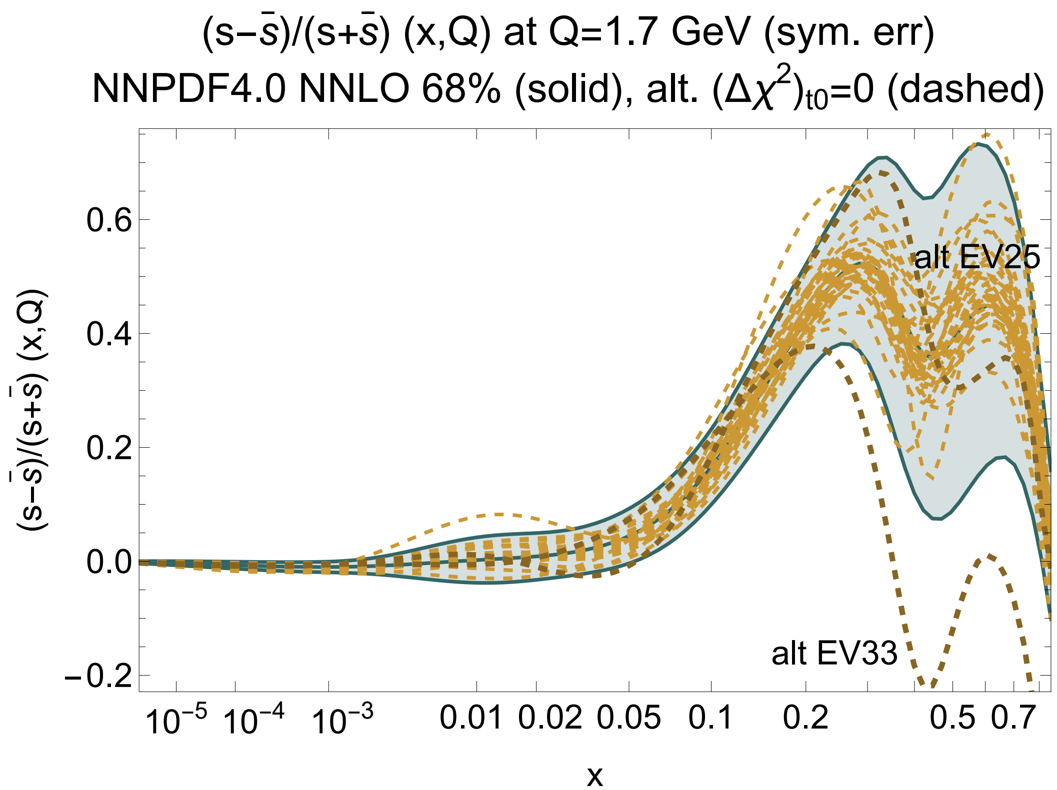
Parton distributions need a representative sampling

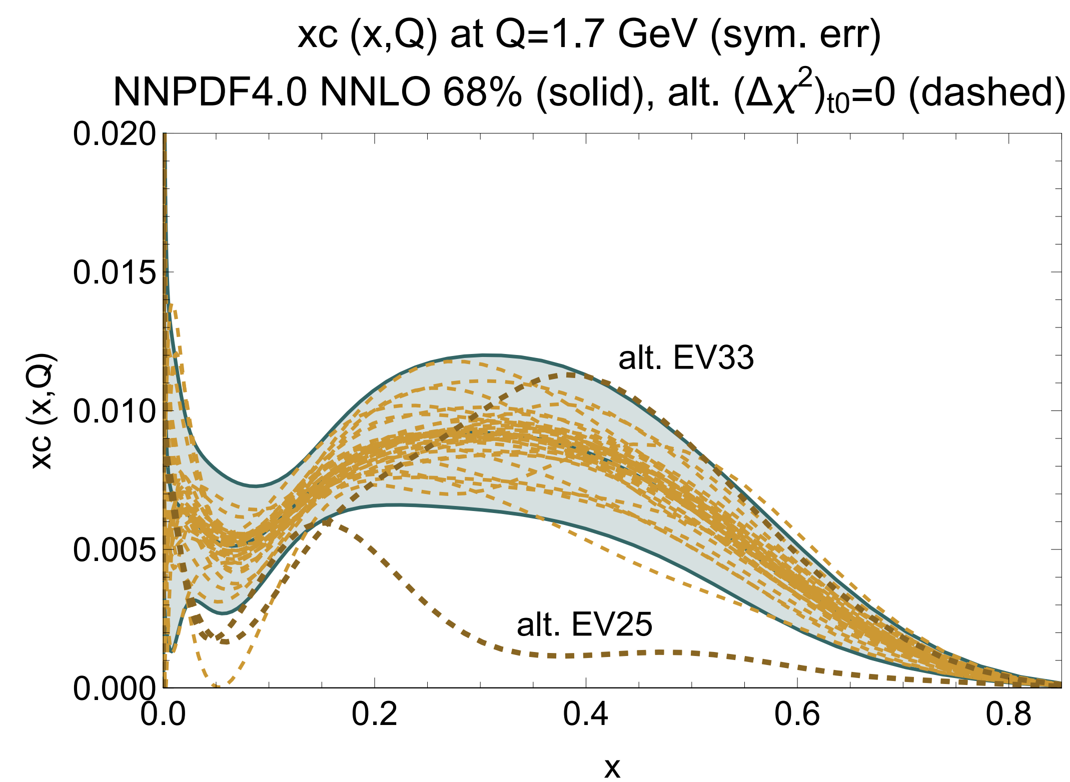
*Monte Carlo uncertainties from **sampling bias** found through our dimensional reduction method play a similar role as sampling of parameter space in Hessian uncertainties.*



PDF tests outside the fit

➡ <u>Hypothesis testing for local true or false statements</u> — *is there XXX* anywhere in the x region?*





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

* *fitted charm, strangeness asymmetry,…*

Hypothesis testing for the pion PDF

⇨ Hypothesis testing for functional behavior constraints — ***do PDFs fall off like $(1-x)^\beta$?***

Polynomial mimicry prevents functional behaviors from being validated as *if and only if* conditions.

Mathematical equivalence of polynomials of different orders can be illustrated with Bézier curves.
QCD corrections, at low and large Q^2, also inhibit the $(1-x)^\beta$ power to be tested.

*[**AC** & P. Nadolsky, Phys.Rev.D 103 (2021)]*

Hypothesis testing for the pion PDF

➪ Hypothesis testing for functional behavior constraints — *do PDFs fall off like* $(1-x)^\beta$*?*

Polynomial mimicry prevents functional behaviors from being validated as *if and only if* conditions.

Mathematical equivalence of polynomials of different orders can be illustrated with Bézier curves.
QCD corrections, at low and large Q^2, also inhibit the $(1-x)^\beta$ power to be tested.

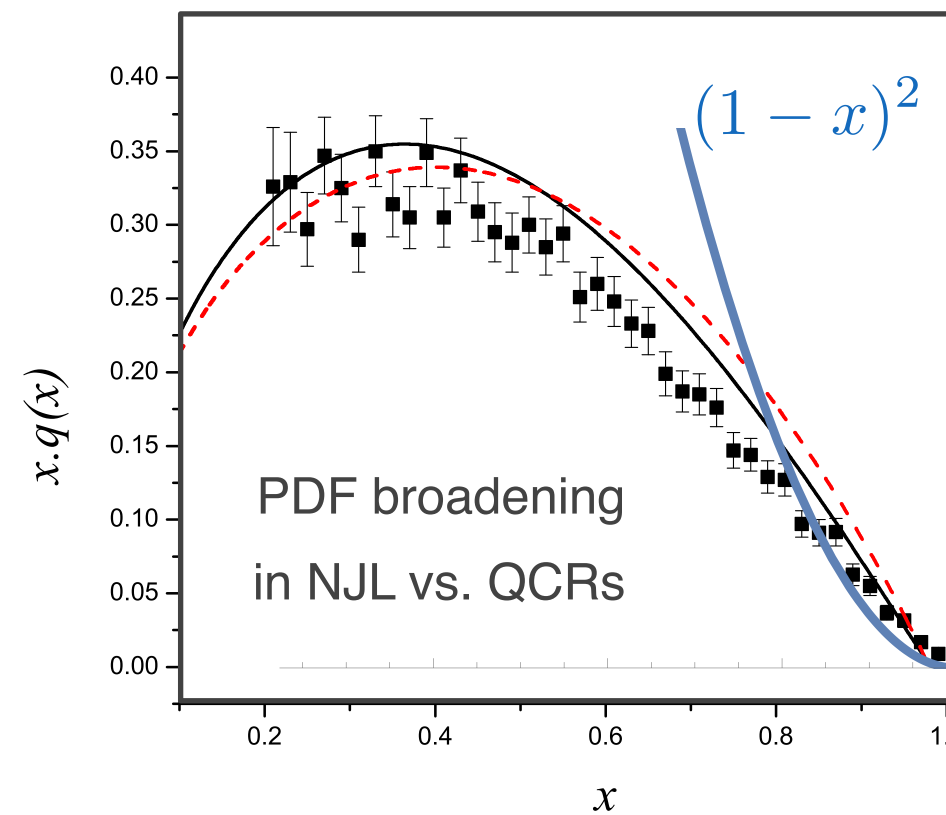
*[**AC** & P. Nadolsky, Phys.Rev.D 103 (2021)]*

Emergence of Hadronic Mass: broadens the PDF at Q_0^2

Quark counting rules: $(1\text{-}x)^2$ tail at mid-Q^2 values

➪ concurring effects that will not be distinguishable at a scale $Q^2 > Q_0^2$.

Bézier curve to improve sampling of parameter space for the pion PDF —next talk by L. Kotz



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.

Moving toward epistemic PDF uncertainty!