

Precision and replicability of parton distributions II

Pavel Nadolsky
Southern Methodist University
CTEQ-TEA global analysis group



Contents

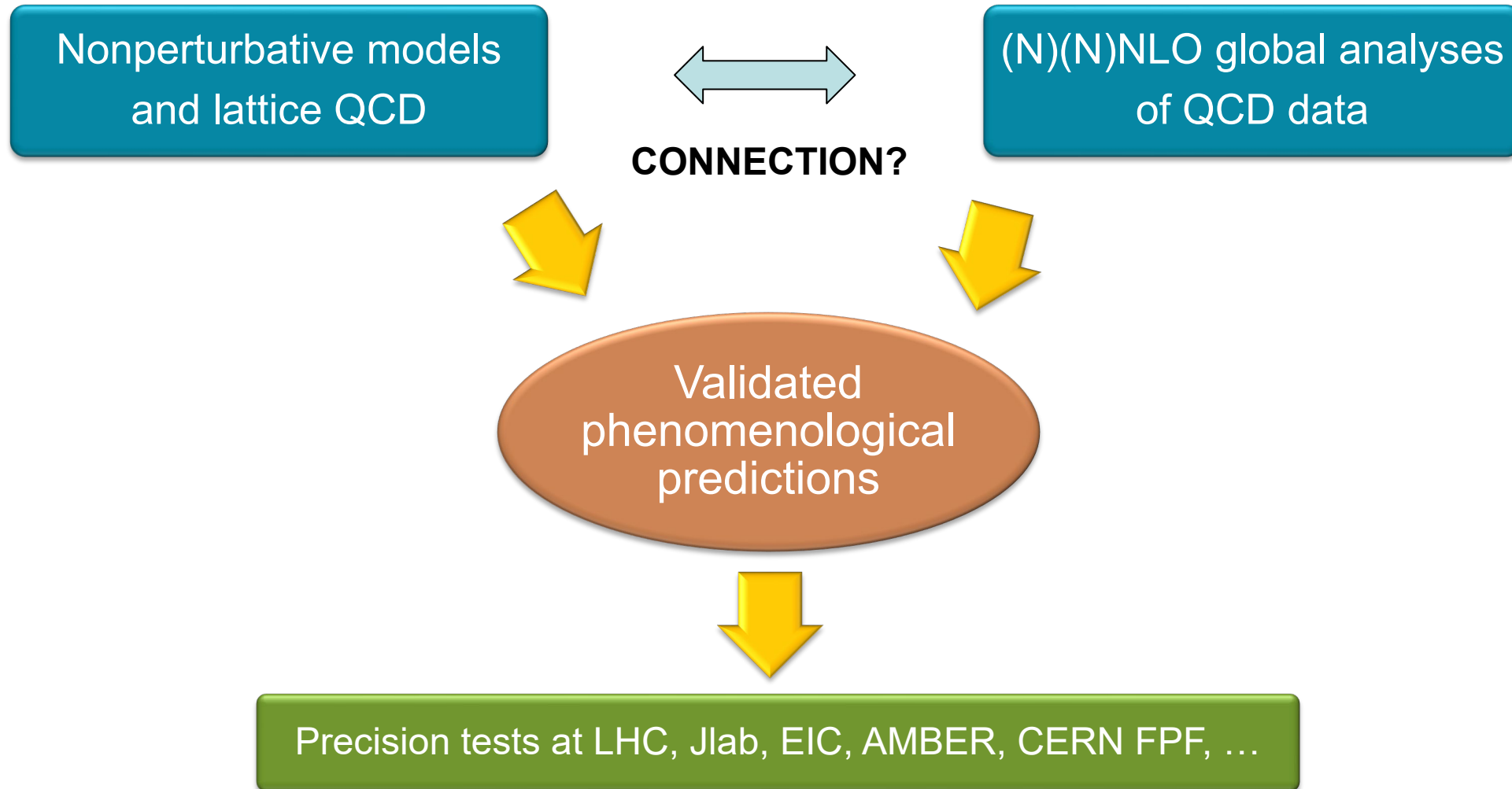
Week 1

1. Dynamic images of hadrons in the AI era
2. Collinear PDFs, their applications and determinations
3. **RRR**: rigor, reproducibility, replicability in the PDF analysis

Week 4

4. Uncertainty quantification on parton distributions
 - Tolerance puzzle
 - Less known aspects of multivariate statistics

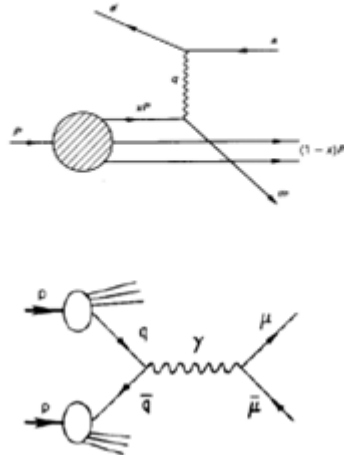
New insights about 3-dimensional structure of hadrons



PDFs in nonperturbative QCD

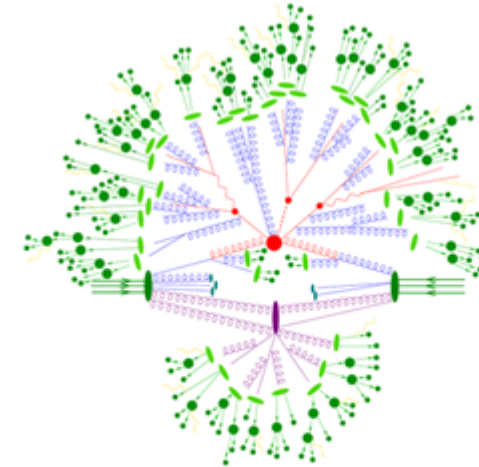
Relevant for processes
at $Q^2 \approx 1 \text{ GeV}^2$?

⇒ we can learn about nonperturbative dynamics by comparing predictions to data for the simplest scattering processes (DIS and DY)



Phenomenological PDFs

Determined from processes
at $Q^2 \gg 1 \text{ GeV}^2$



⇒ pheno PDFs are determined from analyzing many processes with complex scattering dynamics

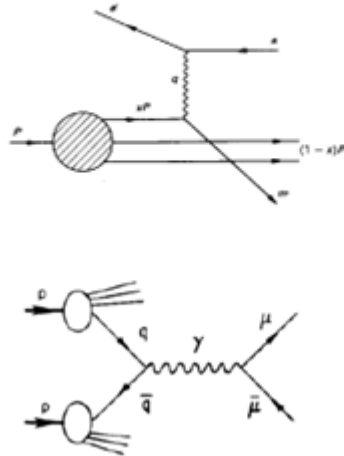
How to relate the x dependence of the perturbative and nonperturbative pictures?

Does the evidence from primordial dynamics survive PQCD radiation?

PDFs in nonperturbative QCD

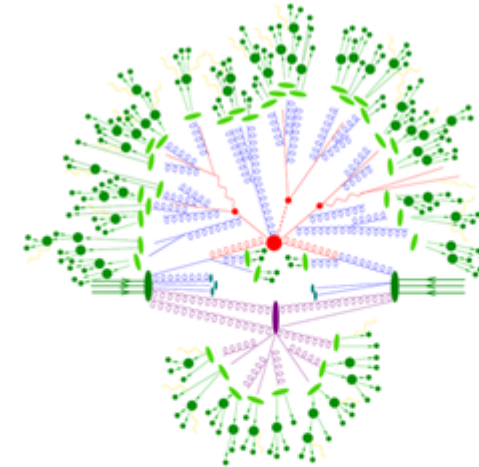
Relevant for processes
at $Q^2 \approx 1 \text{ GeV}^2$

⇒ we can learn about nonperturbative dynamics by comparing predictions to data for the simplest scattering processes (DIS and DY)



Phenomenological PDFs

Determined from processes
at $Q^2 \gg 1 \text{ GeV}^2$



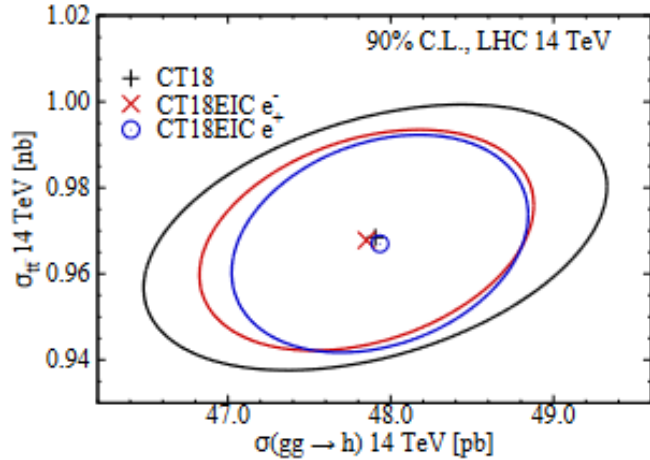
⇒ pheno PDFs are determined from analyzing many processes with complex scattering dynamics

How to relate the x dependence of the perturbative and nonperturbative pictures?

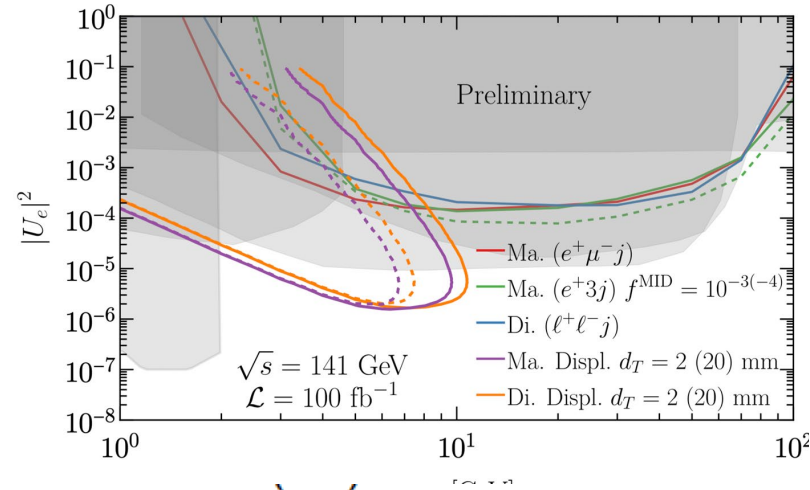
Does the evidence from primordial dynamics survive PQCD radiation?

Electron-Ion Collider: potentially a wealth of complex studies

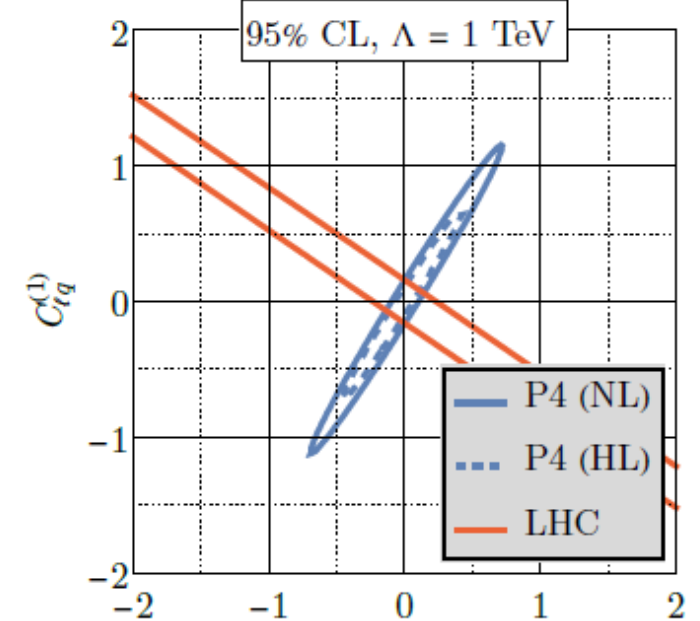
PDFs: arXiv:2103.05419



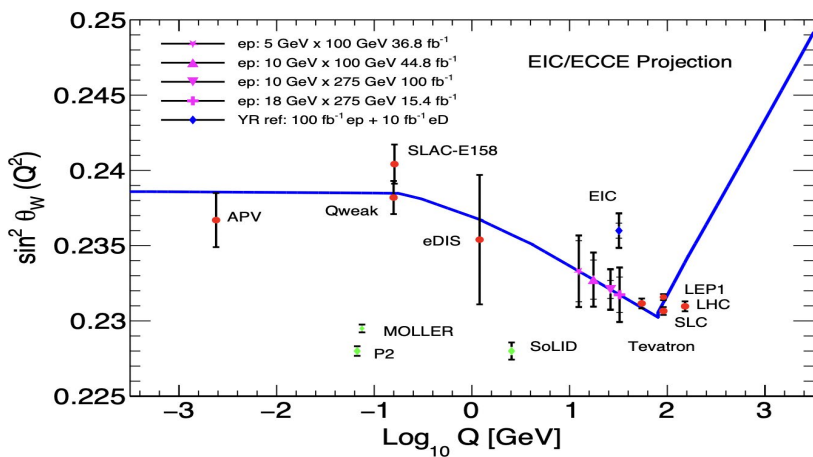
heavy neutral lepton searches arXiv: 2203.06705



SMEFT Wilson coefficients



weak mixing angle arXiv: 2203.13199



	EIC	LHC
$ c_u^{XX} - c_u^{YY} $	0.37	15
$ c_u^{XY} $	0.13	2.7
$ c_u^{XZ} $	0.11	7.3
$ c_u^{YZ} $	0.12	7.1
$ a_{Su}^{(5)TXX} - a_{Su}^{(5)TYY} $	2.3	0.015
$ a_{Su}^{(5)TXY} $	0.34	0.0027
$ a_{Su}^{(5)TXZ} $	0.13	0.0072
$ a_{Su}^{(5)TYZ} $	0.12	0.0070

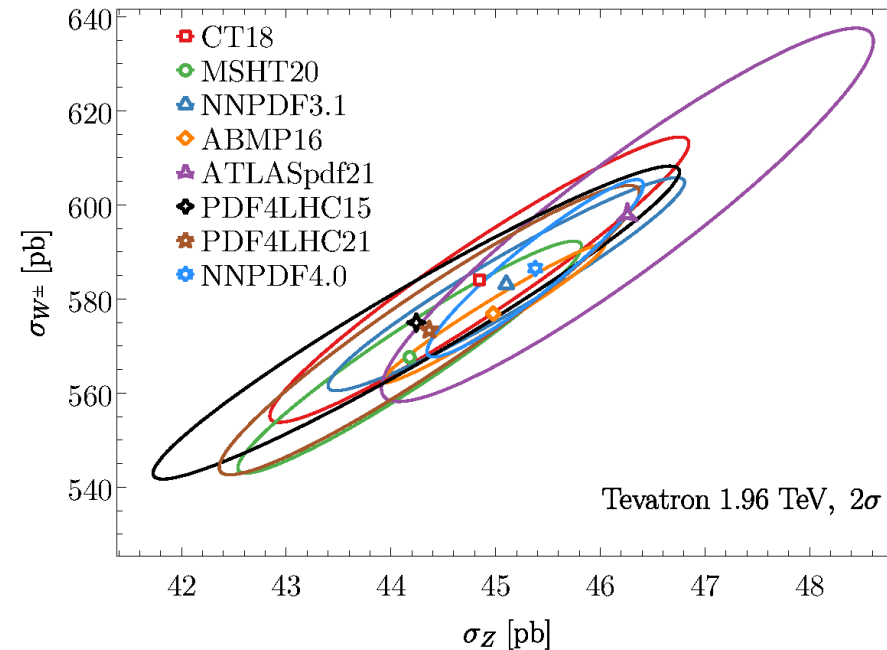
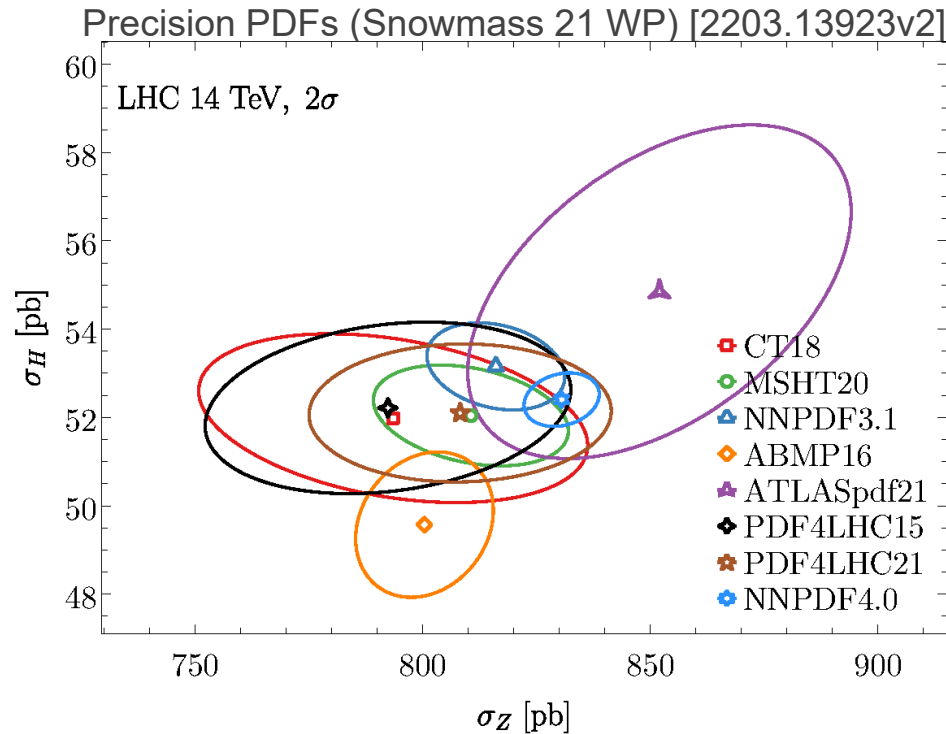
Boughezal et al
2004.00748, .2204.07557

Lorentz/CPT violations
A. R. Vieira et al., 1911.04002

Abdul-Khalek et al., Snowmass 2021 whitepaper
"EIC for HEP", [2203.13199](https://arxiv.org/abs/2203.13199)

The tolerance puzzle

Why do groups fitting similar data sets obtain different PDF uncertainties?



The answer has direct implications for high-stake experiments such as 3D femtography, W boson mass measurement, tests of nonperturbative QCD models and lattice QCD, high-mass BSM searches, etc.

Comparisons of the latest PDF sets

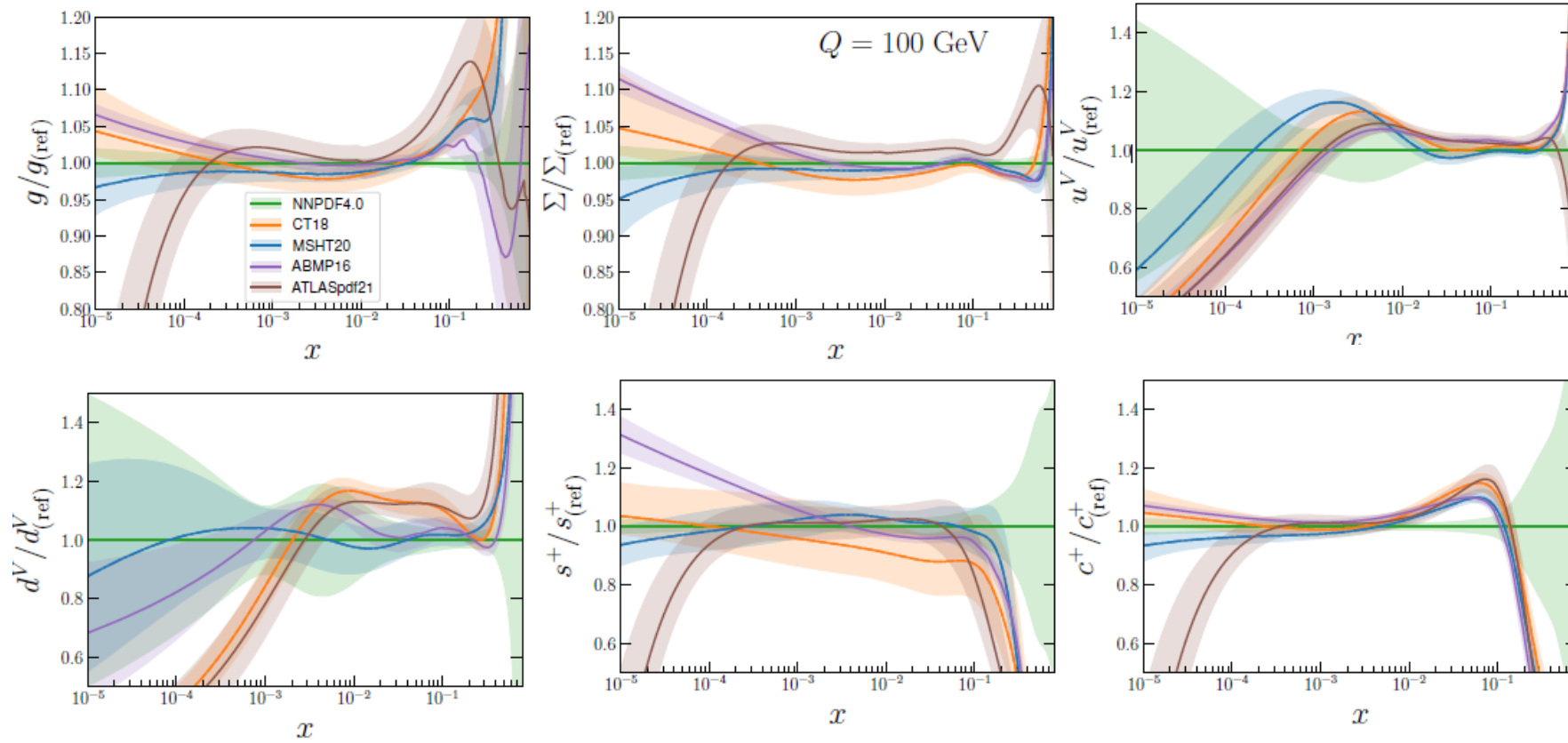


FIG. 2. Comparison of the PDFs at $Q = 100$ GeV. The PDFs shown are the N2LO sets of NNPDF4.0, CT18, MSHT20, ABMP16 with $\alpha_s(M_Z) = 0.118$, and ATLASpdf21. The ratio to the NNPDF4.0 central value and the relative 1σ uncertainty are shown for the gluon g , singlet Σ , total strangeness $s^+ = s + \bar{s}$, total charm $c^+ = c + \bar{c}$, up valence u^V and down valence d^V PDFs.

Statistics with many parameters is different!

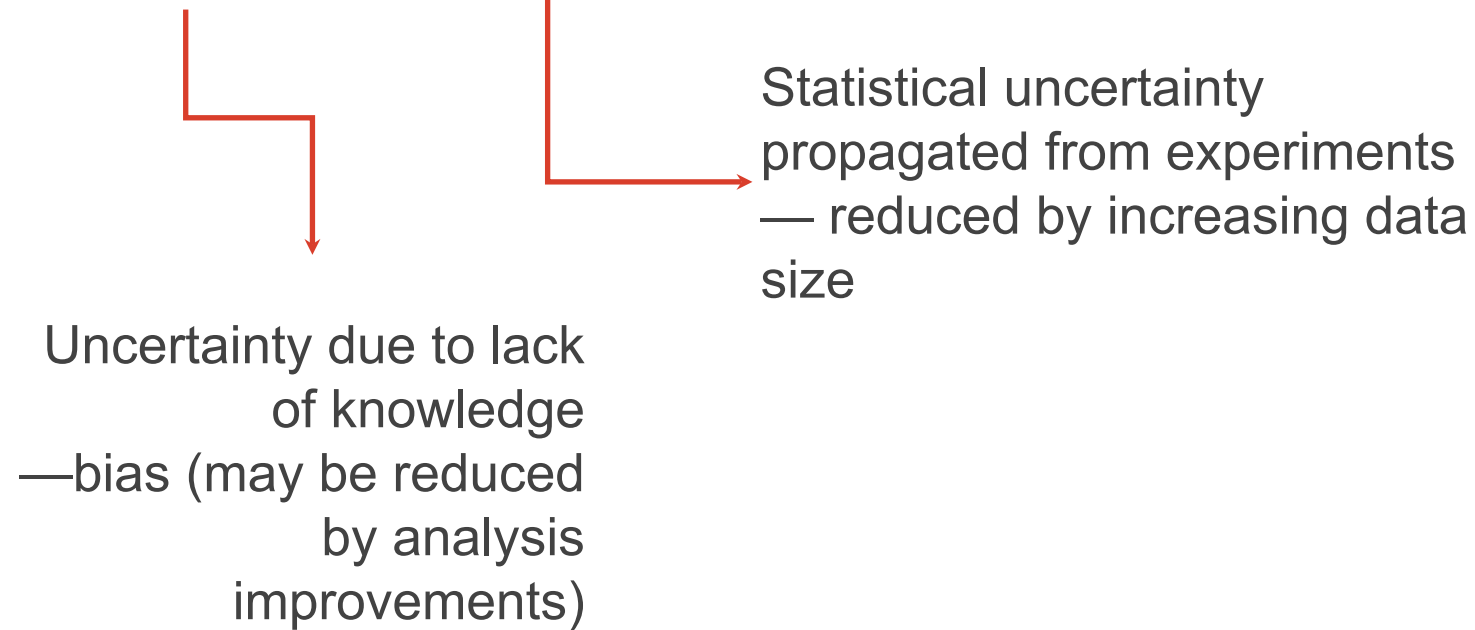
1. Epistemic uncertainties may dominate when other uncertainties are suppressed

More often than not, the realistic 1σ PDF uncertainty does not correspond to $\Delta\chi^2 = 1$.

2. Common estimations of systematic uncertainties are incomplete because...
 - a. **There is no single global minimum of χ^2** (or another cost function)
 - b. **The law of large numbers may not work**
 - uncertainty may not decrease as $1/\sqrt{N_{\text{rep}}}$, leading to the **big-data paradox** [Xiao-Li Meng, 2018]:

The bigger the data, the surer we fool ourselves.

epistemic vs. aleatory uncertainties



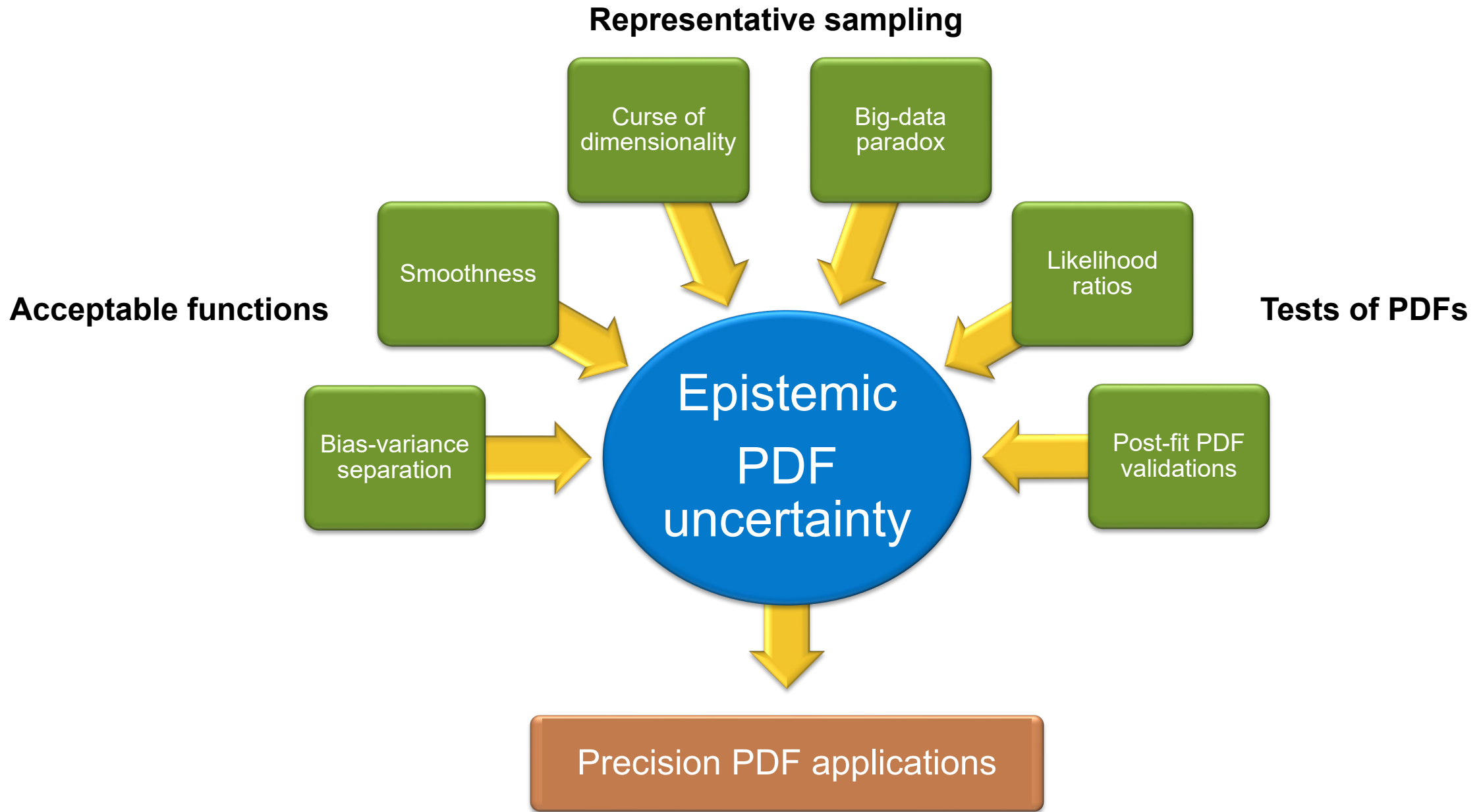
Sources of the uncertainty on PDFs

1. **Experimental uncertainties**, e.g., statistical, correlated and uncorrelated systematic uncertainties of each experimental data set;
2. **Theoretical uncertainties** due to the absent radiative contributions, approximations in parton showering simulations
3. **Parameterization uncertainties** associated with the choice of the PDF functional form or AI/ML replica training algorithm
 - contribute at least a half of the CT18 total PDF uncertainty
4. **Methodological uncertainties** associated with the selection of experimental data sets, fitting procedures, and goodness-of-fit criteria.



associated with the **epistemic** uncertainty

Kovarik et al., arXiv: [1905.06957](https://arxiv.org/abs/1905.06957)



Epistemic PDF uncertainty...

...reflects **methodological choices** such as PDF functional forms, NN architecture and hyperparameters, or model for systematic uncertainties

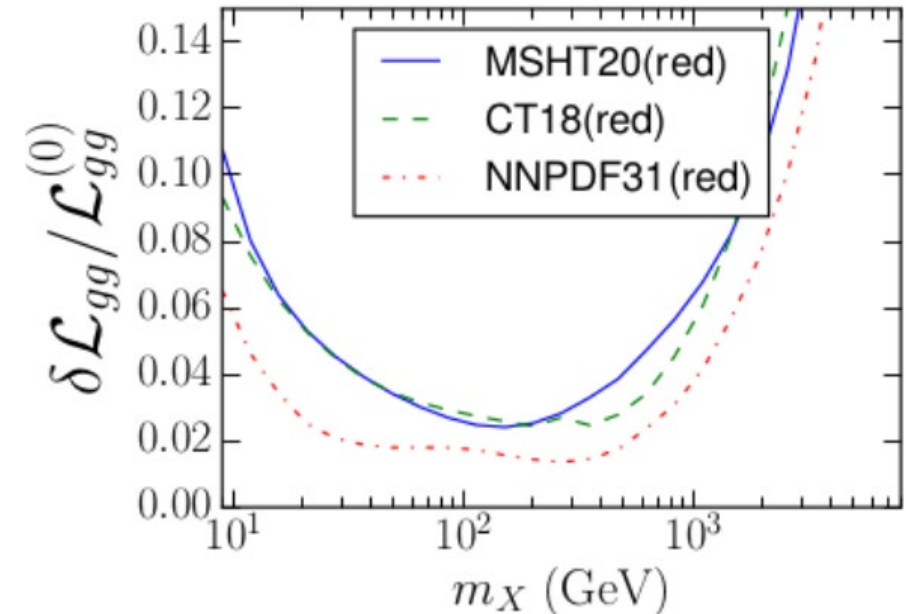
... can dominate the full uncertainty when experimental and theoretical uncertainties are small.

...is associated with the **prior probability**.

... can be estimated by **representative sampling** of the PDF solutions obtained with acceptable methodologies.

⇒ sampling over choices of experiments, PDF/NN functional space, models of correlated uncertainties...

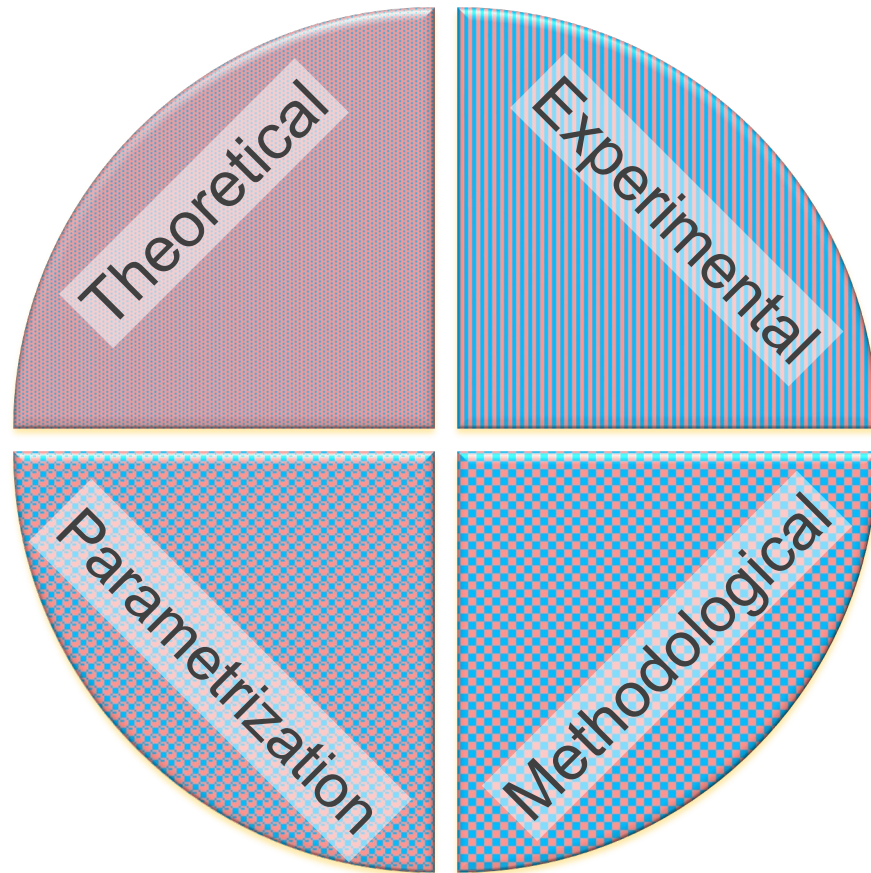
⇒ in addition to sampling over data fluctuations




Epistemic uncertainties explain many of the differences among the sizes of PDF uncertainties by CT, MSHT, and NNPDF global fits to the same or similar data


Details in [arXiv:2203.05506](https://arxiv.org/abs/2203.05506), [arXiv:2205.10444](https://arxiv.org/abs/2205.10444)

Components of PDF uncertainty



In each category, one must maximize

 **PDF fitting accuracy**
(accuracy of experimental, theoretical and other inputs)

 **PDF sampling accuracy**
(adequacy of sampling in space of possible solutions)

Fitting/sampling classification is borrowed from the statistics of large-scale surveys [Xiao-Li Meng, *The Annals of Applied Statistics*, Vol. 12 (2018), p. 685]

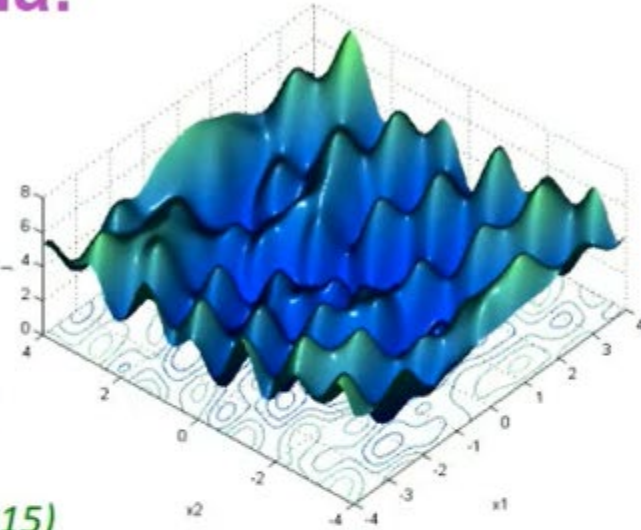
A multidimensional mini-landscape,
in which
the global χ^2 minimum is rare

Not so terrible local minima: convexity is not needed

Myth busted:

- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are relatively close to the bottom (global minimum error)

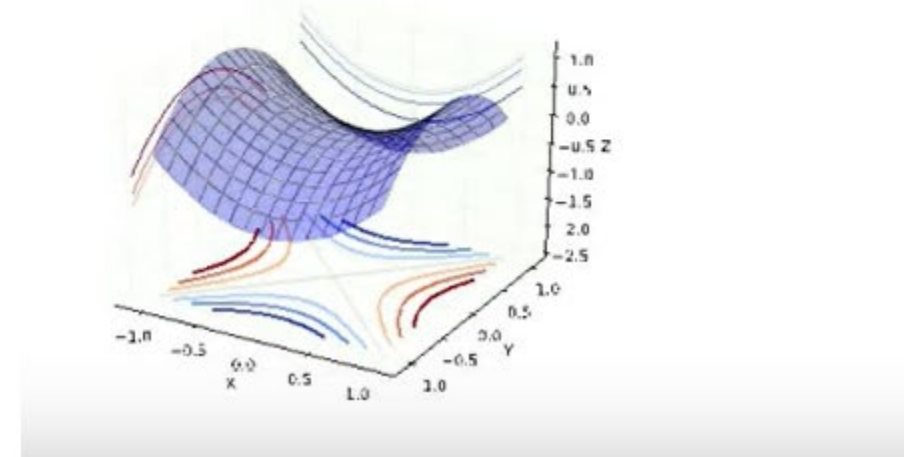
(Dauphin et al NIPS'2014, Choromanska et al AISTATS'2015)



Global minimum: all $\frac{\partial^2 \chi^2}{\partial a_i \partial a_j} > 0$ (improbable)

Saddle point: some $\frac{\partial^2 \chi^2}{\partial a_i \partial a_j} > 0$ (probable)

An average global minimum: in properly chosen coordinates, $\frac{\partial^2 \chi^2}{\partial z_i \partial z_j} > 0$ for dominant coordinate components



Y. Bengio, 2019 Turing lecture ([YouTube](#))

Many dimensions introduce major difficulties with identifying a global minimum...

An important question concerns the distribution of critical points (maxima, minima, and saddle points) of such functions. Results from random matrix theory applied to spherical spin glasses have shown that these functions have a combinatorially large number of saddle points. Loss surfaces for large neural nets have many local minima that are essentially equivalent from the point of view of the test error, and these minima tend to be highly degenerate, with many eigenvalues of the Hessian near zero.

We empirically verify several hypotheses regarding learning with large-size networks:

- For large-size networks, most local minima are equivalent and yield similar performance on a test set.
- The probability of finding a “bad” (high value) local minimum is non-zero for small-size networks and decreases quickly with network size.
- Struggling to find the global minimum on the training set (as opposed to one of the many good local ones) is not useful in practice and may lead to overfitting.

The Loss Surfaces of Multilayer Networks

A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. LeCun PMLR 38:192-204, 2015

Many dimensions introduce major difficulties with identifying a global minimum...

...as well as with representative exploration of uncertainties

The Big Data Paradox in vaccine uptake

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

<https://doi.org/10.1038/s41586-021-04198-4>

Received: 18 June 2021

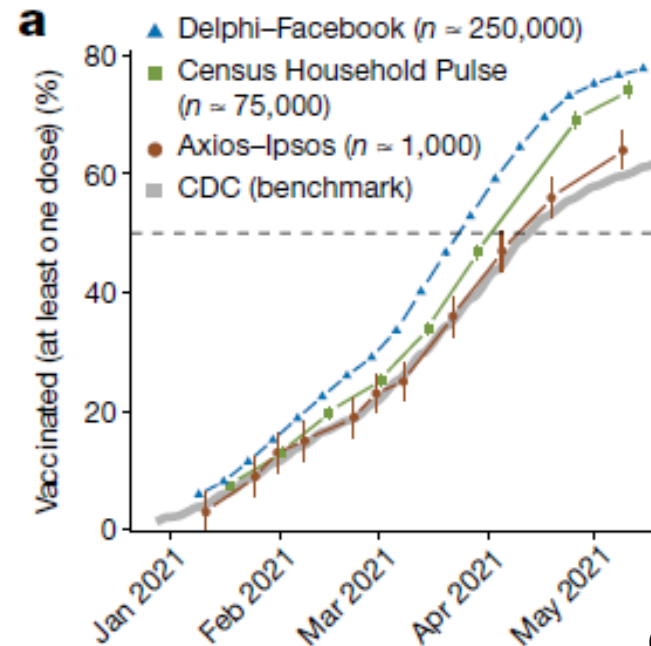
Accepted: 29 October 2021

Published online: 8 December 2021

Check for updates

Valerie C. Bradley^{1,2}, Shiro Kuriwaki^{2,3}, Michael Isakov³, Dino Sejdinovic³, Xiao-Li Meng⁴ & Seth Flaxman⁵✉

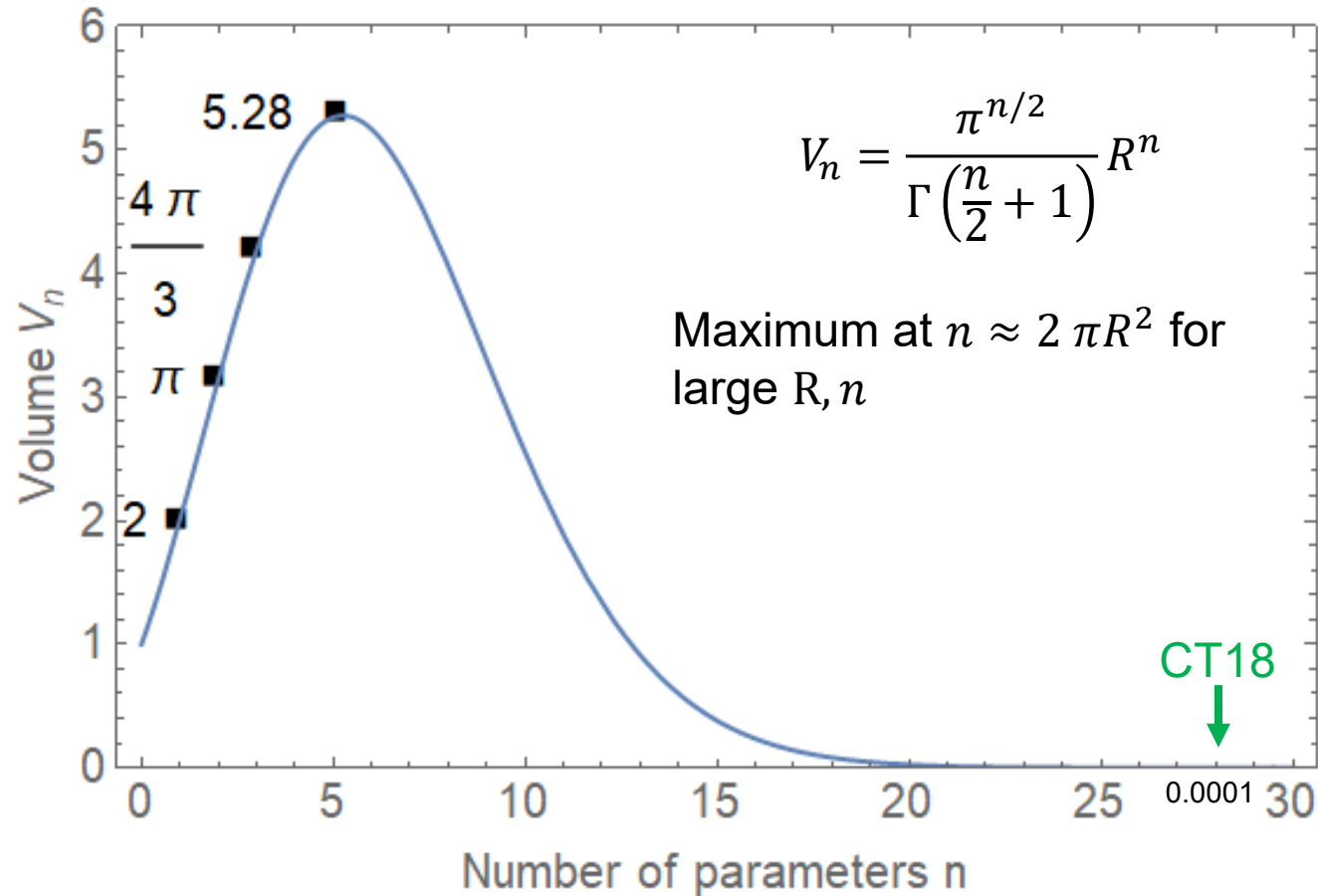
Surveys are a crucial tool for understanding public opinion and behaviour, and their accuracy depends on maintaining statistical representativeness of their target populations by minimizing biases from all sources. Increasing data size shrinks confidence intervals but magnifies the effect of survey bias: an instance of the Big Data Paradox¹. Here we demonstrate this paradox in estimates of first-dose COVID-19 vaccine uptake in US adults from 9 January to 19 May 2021 from two large surveys: Delphi–Facebook^{2,3} (about 250,000 responses per week) and Census Household Pulse⁴ (about 75,000 every two weeks). In May 2021, Delphi–Facebook overestimated uptake by 17 percentage points (14–20 percentage points with 5% benchmark imprecision) and Census Household Pulse by 14 (11–17 percentage points with 5% benchmark imprecision), compared to a retroactively updated benchmark the Centers for Disease Control and Prevention published on 26 May 2021. Moreover, their large sample sizes led to minuscule margins of error on the incorrect estimates. By contrast, an Axios–Ipsos online panel⁵ with about 1,000 responses per week following survey research best practices⁶ provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework⁷ to explain the inaccuracy in the three surveys. We then analyse the implications for vaccine hesitancy and willingness. We show how a survey of 250,000 respondents can produce an estimate of the population mean that is no more accurate than an estimate from a simple random sample of size 10. Our central message is that data quality matters more than data quantity, and that compensating the former with the latter is a mathematically provable losing proposition.



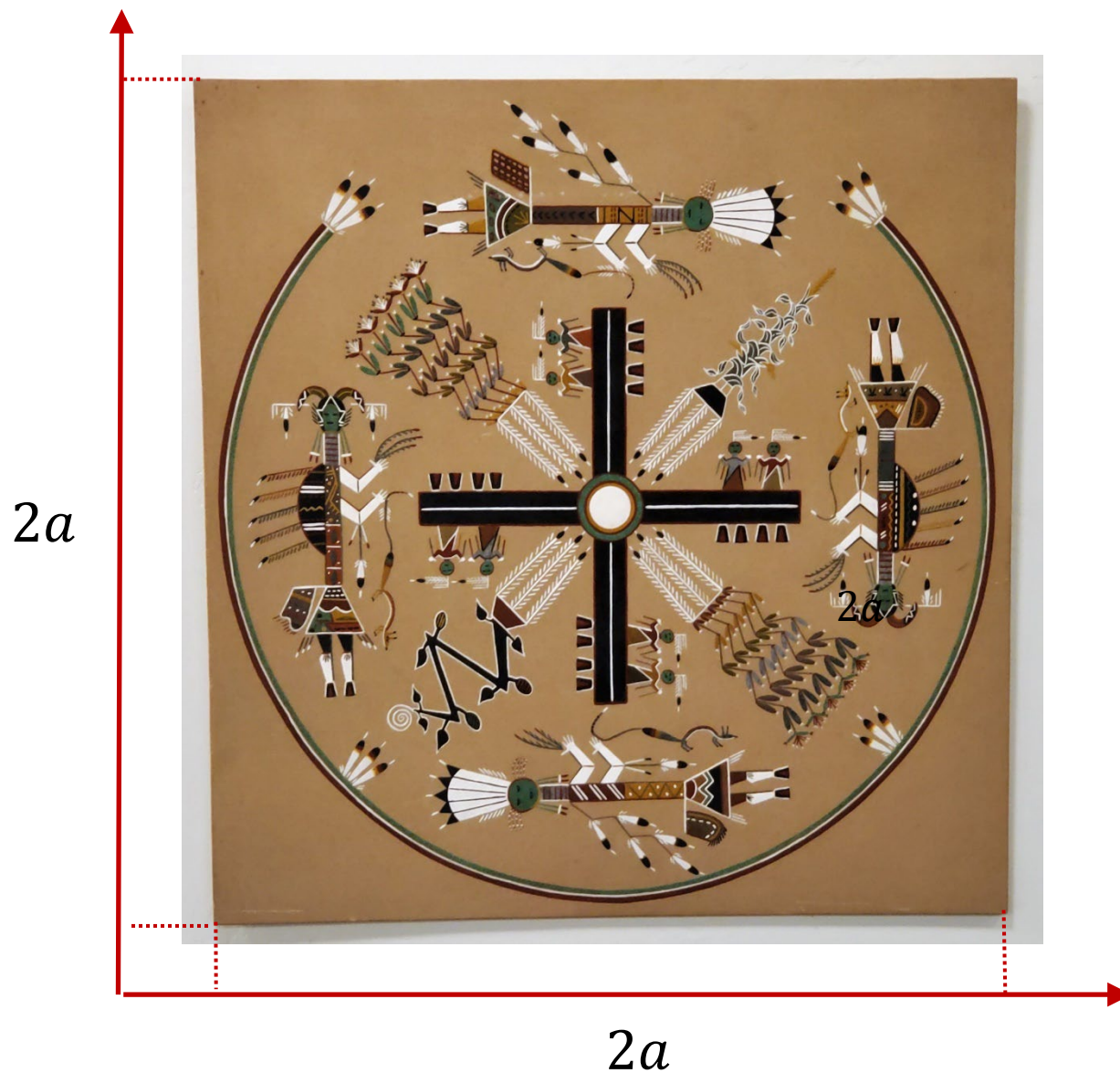
[Nature](#) v. 600 (2021) 695

Courtoy et al., PRD 107 (2023) 034008

Volume of a unit ball in n dimensions



**The Curse
of Dimensionality!**



Compare:

- the volume of a cube with side $2a$
- the volume of a sphere with radius a

- $n=2$

$$\frac{V_{sphere}}{V_{cube}} = \frac{\pi}{4} \approx \mathbf{0.8}$$

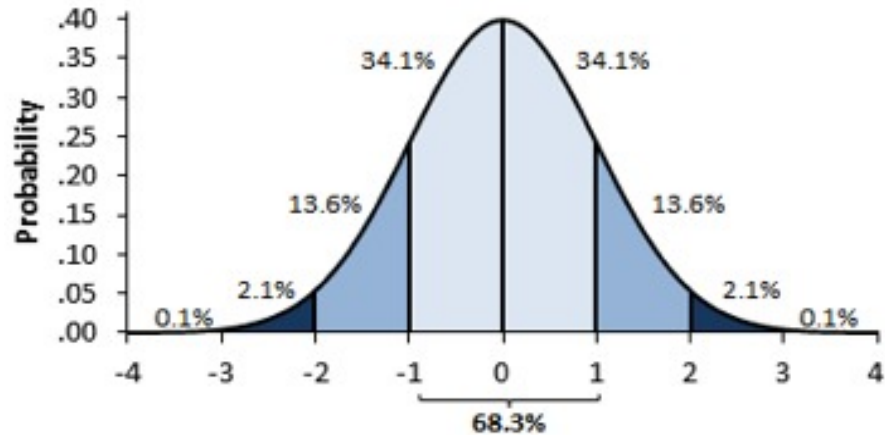
- $n=25$

$$\frac{V_{sphere}}{V_{cube}} \approx \frac{0.0009}{2^{25}} \approx \mathbf{3 \cdot 10^{-11}}$$

Image: sand painting, SMU-in-Taos

An n-dimensional standard normal distribution

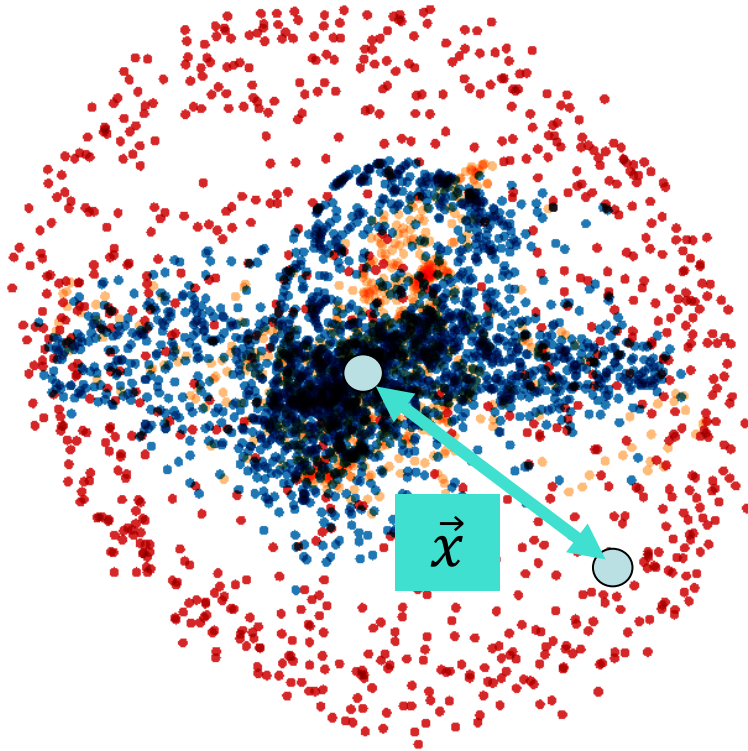
$$P(\vec{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\vec{x}^2}{2}\right)$$



Any 1-dim. projection contains 68% of the elements in the interval
 $-1 < x_i < 1$

An n-dimensional **standard normal** distribution

$$P(\vec{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\vec{x}^2}{2}\right)$$



The mean distance of an element from the center (“truth”) at $\vec{x} = 0$ is

$$\langle |\vec{x}| \rangle \approx \sqrt{n}$$

$$\sqrt{n} \approx 5 \text{ for } n = 25$$

In a large- n **normal** distribution, a single element is likely to be very **abnormal** (be $\sim \sqrt{n} \sigma$ away from the “truth”) in some direction(s)

Hou et al., [arXiv:1607.06066](https://arxiv.org/abs/1607.06066)

Law of large numbers

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

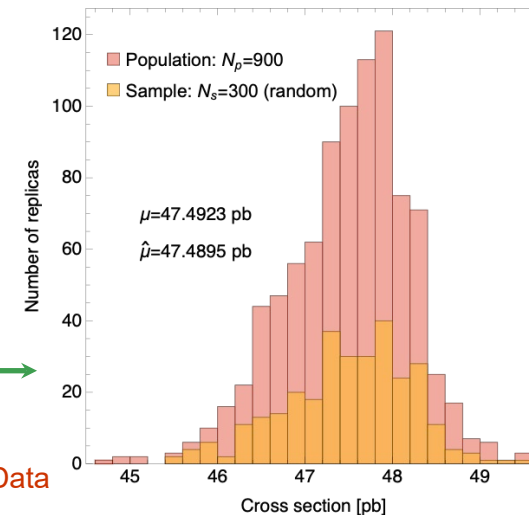
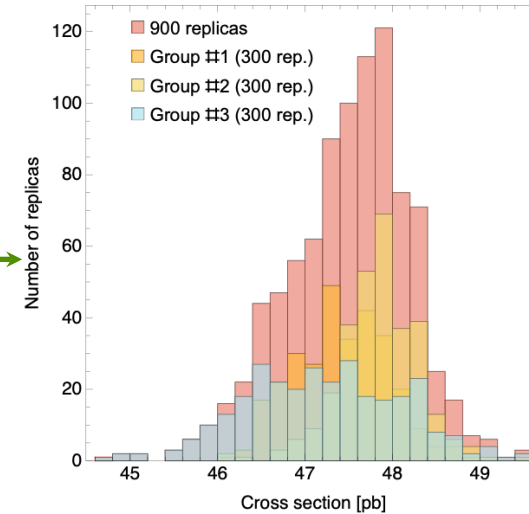
$$\mu - \hat{\mu} \propto \sigma_{std} / \sqrt{n}$$

with σ_{std} the standard variation, μ and $\hat{\mu}$ the true and sample expectation values. *This is the law of large numbers.*

A toy sampling exercise

We take 300×3 groups of **Higgs cross sections** evaluated by 3 different groups (CT18', MSHT20, NNPDF3.1').

We **randomly** select 300 out of the 900 cross sections. The law of large numbers is fulfilled in this case: there is no bias.

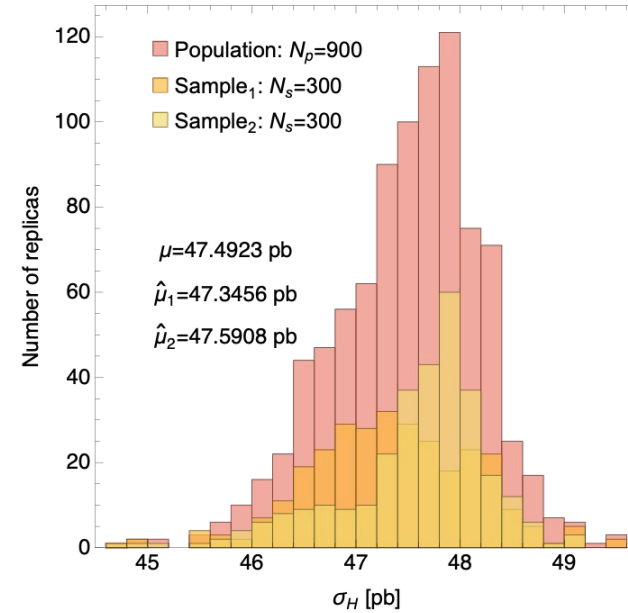


Trio identity

If we **bias** the selection by taking 200 items from one group and 100 from another, the deviation $\mu - \hat{\mu}$ is no longer proportional to σ_{std}/\sqrt{n} !



Quality of the sample is as important as quantity.



The **trio identity** identifies three main contributions to the sample deviation:

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

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

Trio identity, continued

A sample of n items from a population of size N can be described by an array R_j of sampling indicators =0 or 1, which shows that

$$\mu - \hat{\mu} = \underbrace{\text{Corr}[\text{observable}, \text{sampling algorithm}]}_{\text{depends on the sampling algorithm}} \times \underbrace{\sqrt{\frac{N}{n} - 1} \times \sigma_{std}(\text{observable})}_{\text{decreases as } \sigma_{std}/\sqrt{n} \text{ for random sampling}}$$

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

Consequences for large N (or large N_{par}):

1. The sample deviation can be large if $\text{Corr}[\dots]$ does not decrease as $o(1/\sqrt{N})$
2. Standard error estimates can be misleadingly small.
3. **Control for sampling biases is critical** to avoid the situation described as the **Big Data Paradox** [Meng]:

The bigger the data, the surer we fool ourselves.

Complexity and PDF tolerance

- **Bad news:** The tolerance puzzle is *intractable* in very complex fits
 - In a fit with N_{par} free parameters, the minimal number of PDF replicas to estimate the expectation values for $\forall \chi^2$ function grows as $N_{min} \geq 2^{N_{par}}$
 - Example: $N_{min} > 10^{30}$ for $N_{par} = 100$

[Sloan, Woźniakowski, 1997]

[Hickernell, MCQMC 2016, 1702.01487]

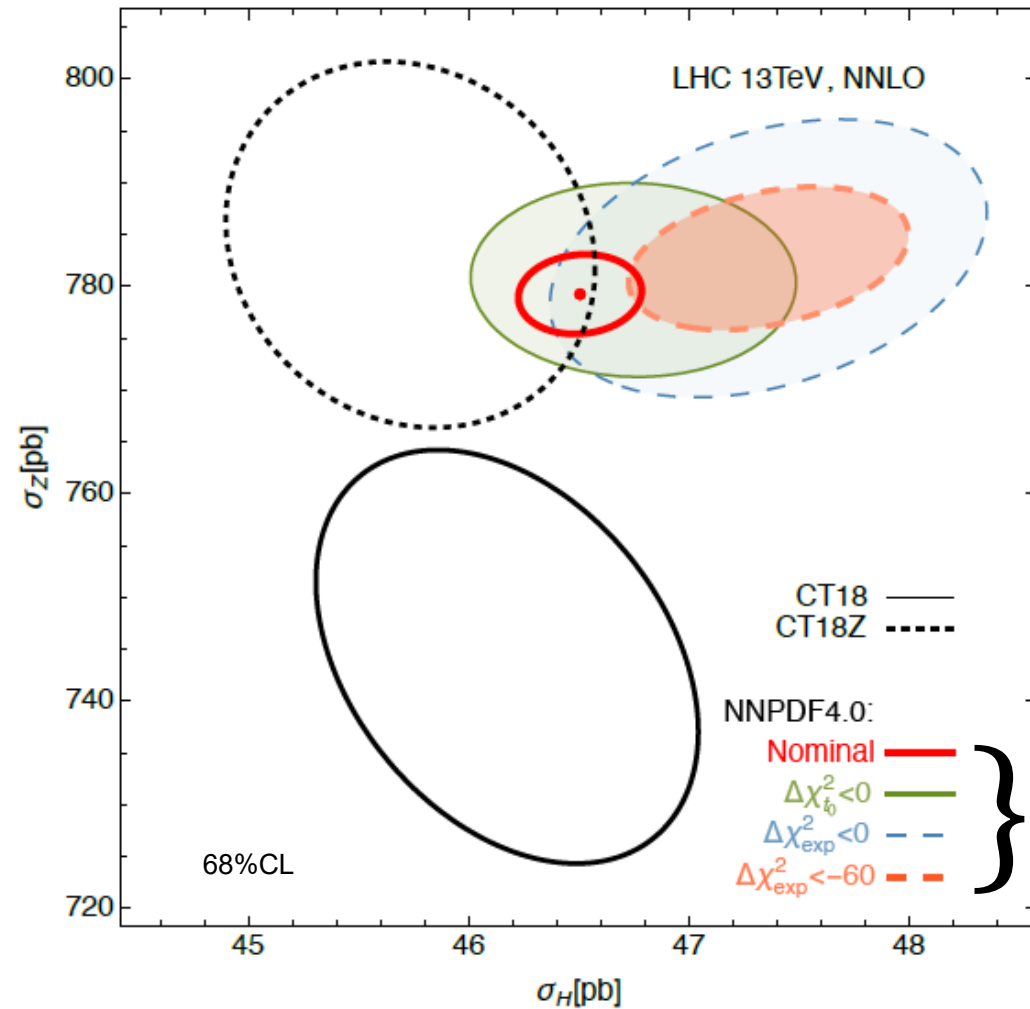
Good news: expectation values for **typical QCD observables** can be estimated with fewer replicas by reducing dimensionality of the problem or a targeted sampling technique.

Example: a “**hopscotch scan**”, see 2205.10444



Example: the impact of epistemic uncertainty on NNLO Higgs and Z cross sections

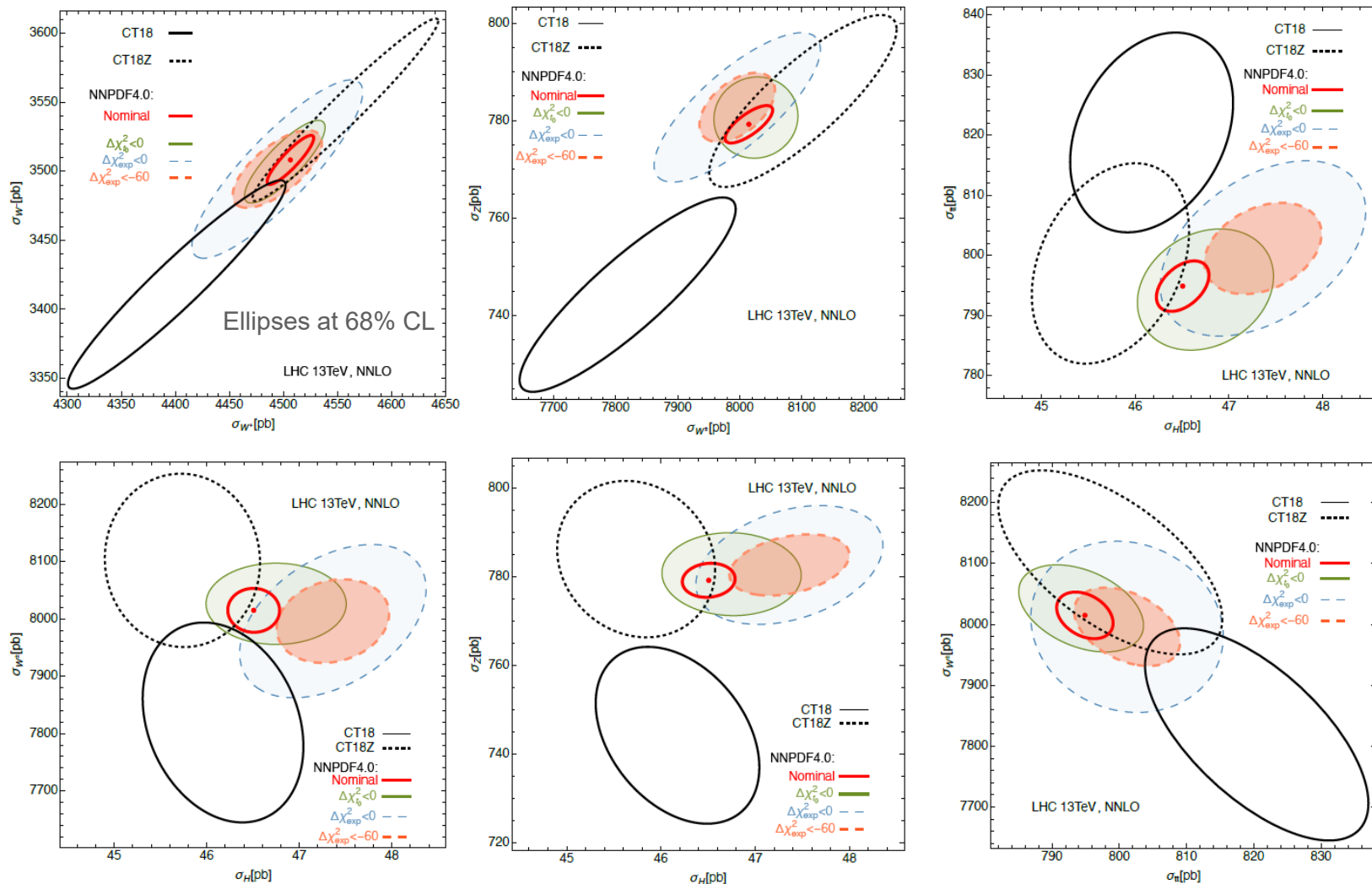
Details in
A. Courtoy et al.,
[arXiv:2205.10444](https://arxiv.org/abs/2205.10444)



obtained with the same NNPDF4.0 fitting code
using a “**hopscotch scan**” of the PDF param. space

all ellipses contain acceptable predictions
according to the likelihood-ratio test
Nominal NN4.0 uncertainty is too small!

Impact of epistemic uncertainties on other cross sections



The ellipses are projections of 68% c.l. ellipsoids in N_{par} -dim. spaces

$N_{par} = 28$ and 50 for CT18 and NNPDF4.0 Hessian PDFs

Weak and strong goodness-of-fit criteria

Kovarik, P. N., Soper, **arXiv:1905.06957**

Weak (common) goodness-of-fit (GOF) criterion

Based on the global χ^2

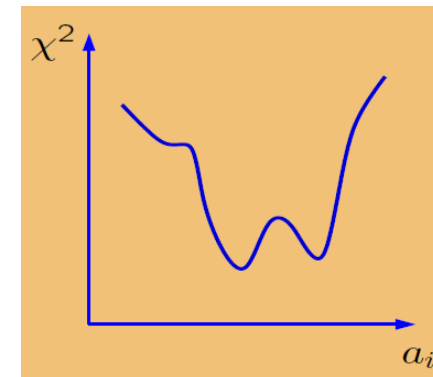
A fit of a PDF model to N_{exp} experiments with N_{pt} points ($N_{pt} \gg 1$) is good at the probability level p if $\chi_{global}^2 \equiv \sum_{n=1}^{N_{exp}} \chi_n^2$ satisfies

$$P(\chi^2 \geq \chi_{global}^2, N_{pt}) \geq p; \quad e.g.$$

$$|\chi_{global}^2 - N_{pt}| \lesssim \sqrt{2N_{pt}} \quad \text{for } p = 0.68$$

Even when the weak GOF criterion is satisfied, parts of data can be poorly fitted

Then, **tensions between experiments** may lead to **multiple solutions** or **local χ^2 minima** for some PDF combinations



An excellent fit requires more than a good global χ^2

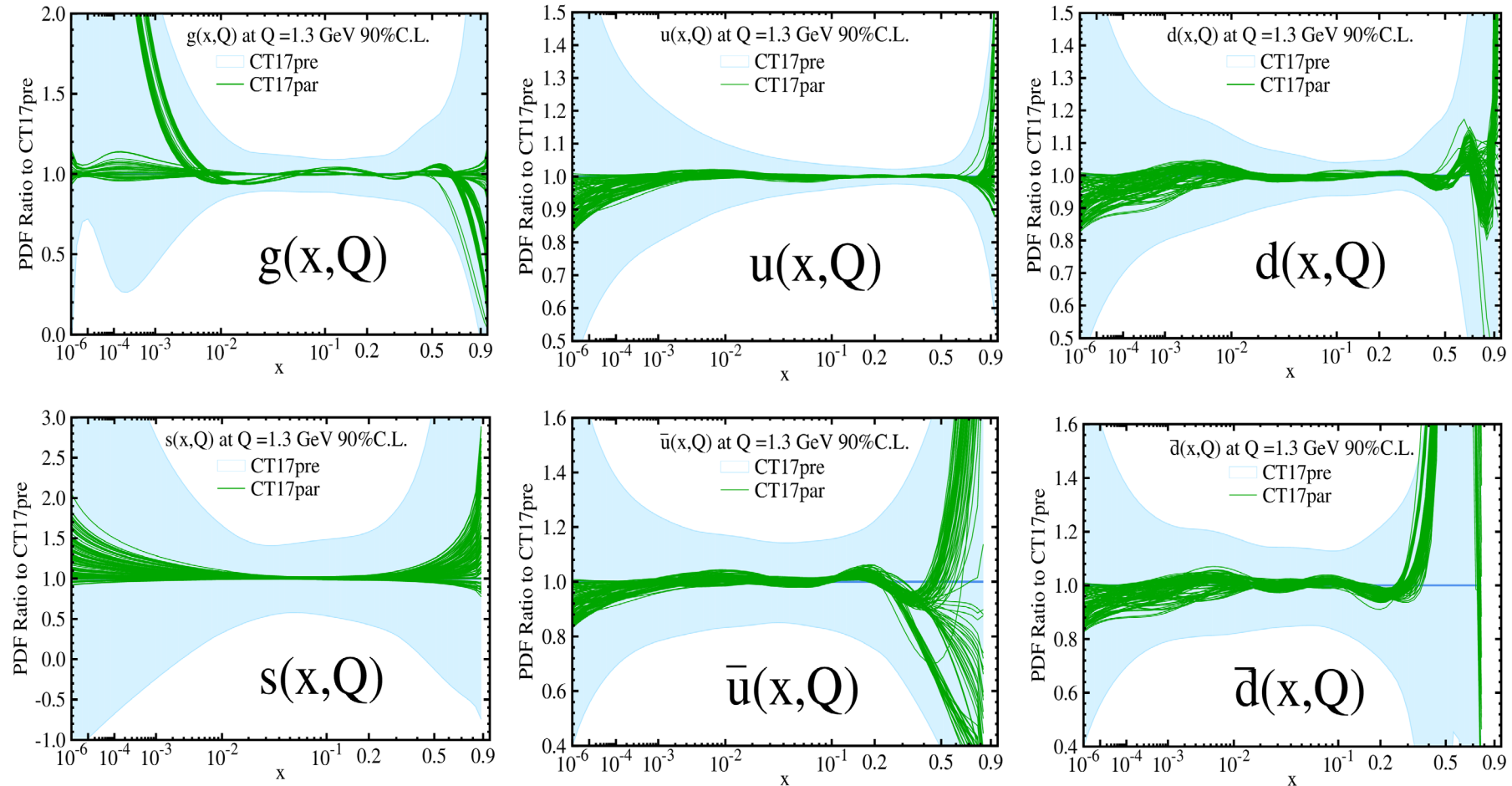
It passes a number of quality tests, called together the **strong set of goodness-of-fit criteria**

1. Each possible partition n of the global data set has a good χ^2
 - differences between theory and data for this partition are indistinguishable from random fluctuations
 - $P(\{\chi_n^2\}) \geq 0.68$ for the distribution of χ_n^2 over N_{part} partitions
2. Best-fit nuisance parameters obey the expected probability distribution
3. **Resampling test:** the data are neither underfitted nor overfitted
4. A closure test is passed, such as the one used in NNPDF 3.x
5. ...

Functional forms of PDFs and resampling test

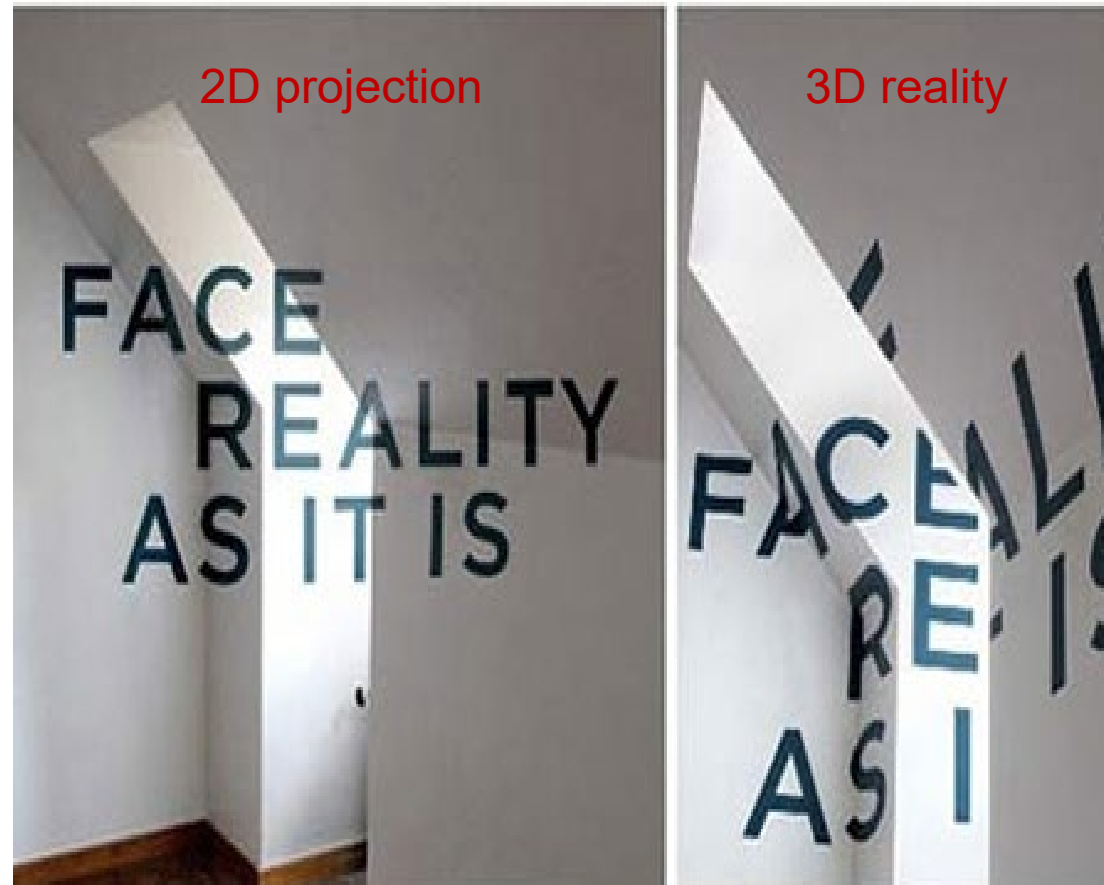
The uncertainty due to the PDF functional form contributes as much as 50% of the total PDF uncertainty in CT fits. The CT18 analysis estimates this uncertainty using 100 trial functional forms.

Explore various non-perturbative parametrization forms of PDFs



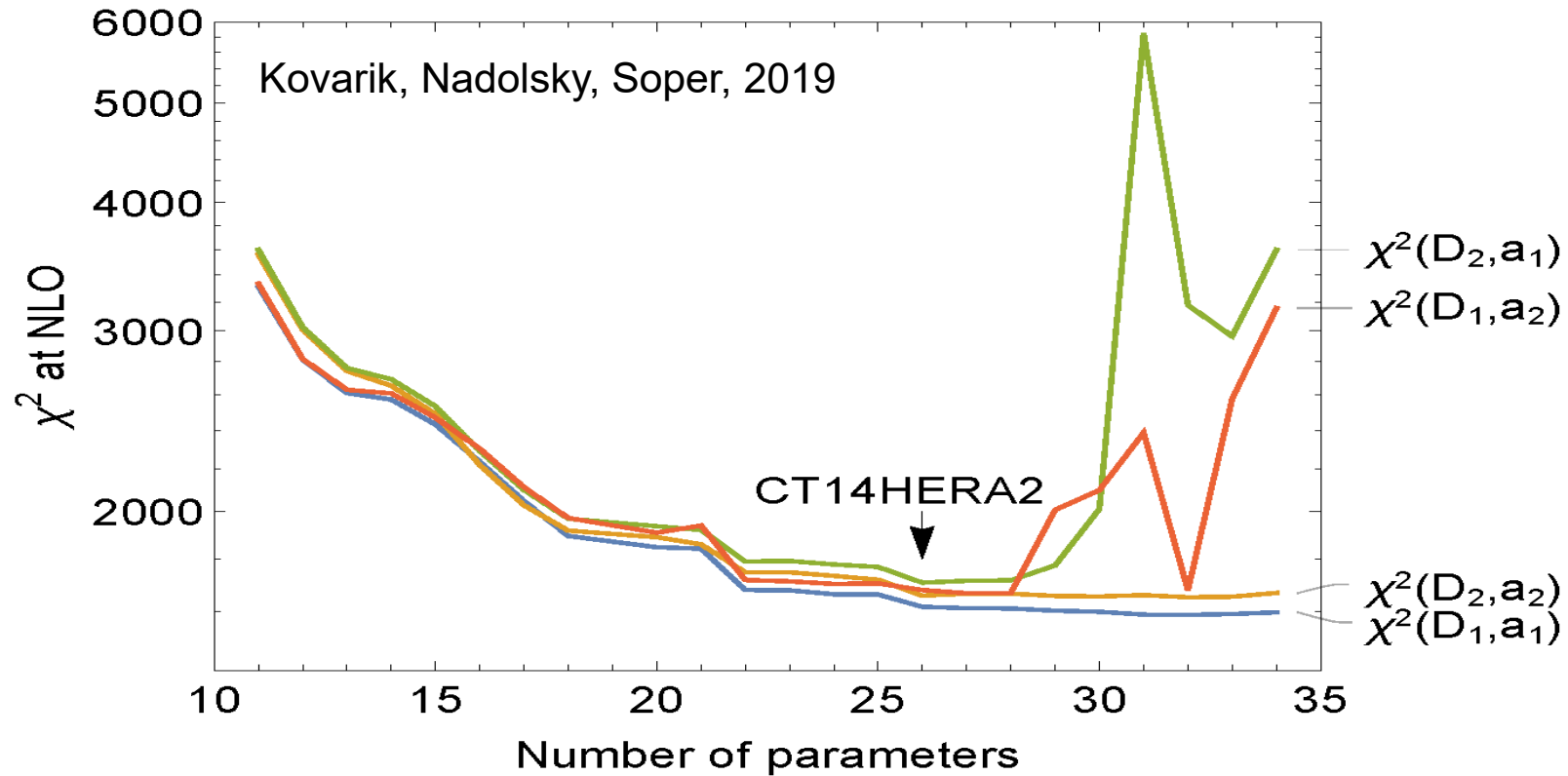
- CT17par – sample result of using various non-perturbative parametrization forms.
- No data constrain very large x or very small x regions.

If too few parameters



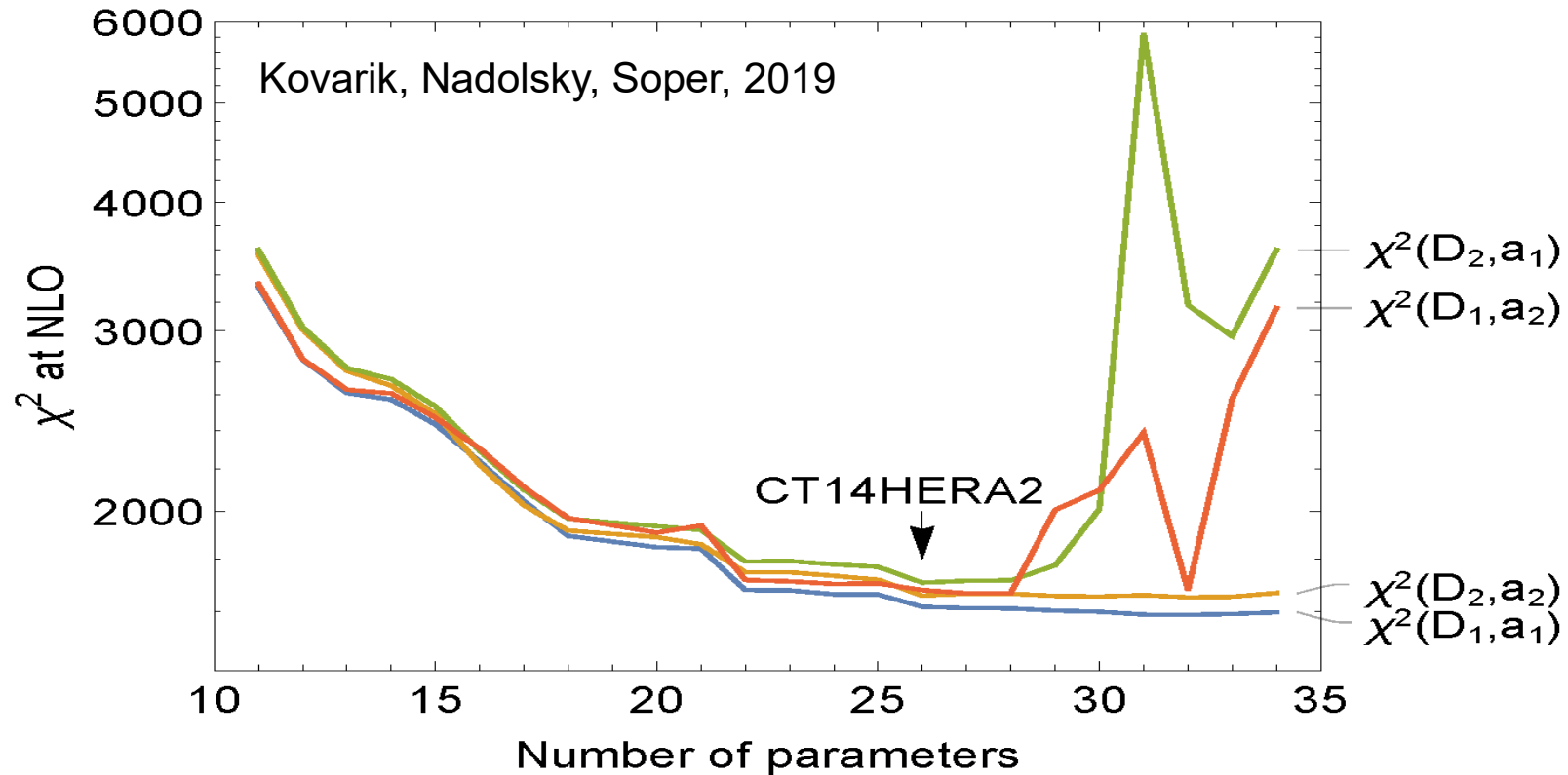
The solution can be consistent and false

If too many parameters



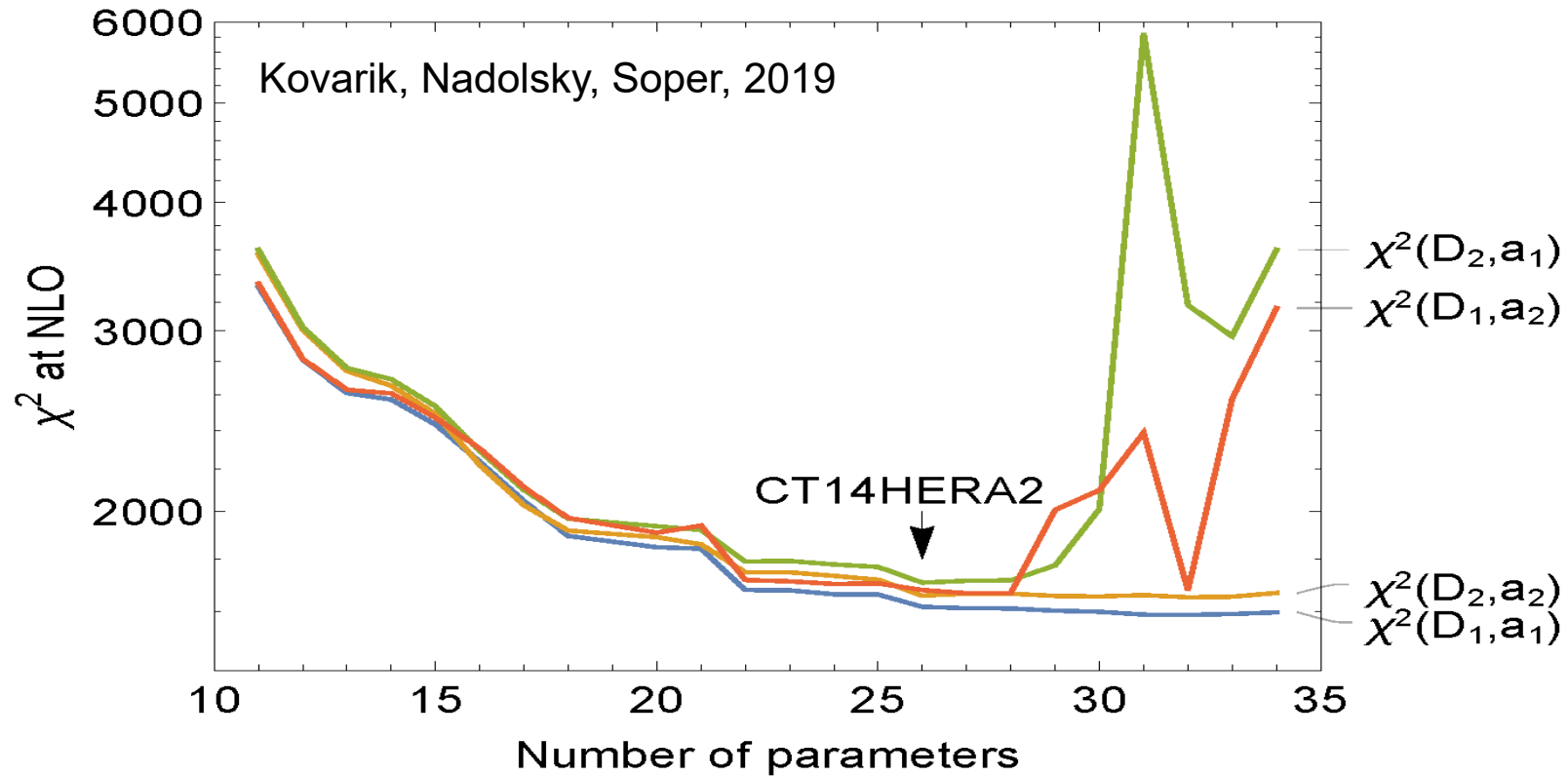
- Randomly split the CT14HERA data set into two halves, D_1 and D_2
- Find parameter vectors a_1 and a_2 from the best fits for D_1 and D_2 , respectively

If too many parameters



- **Fitted samples:** $\chi^2(D_1, a_1)$ and $\chi^2(D_2, a_2)$ uniformly decrease with the number of parameters; eventually the fits become unstable (“fitting noise”)
- **Control samples:** $\chi^2(D_2, a_1)$ and $\chi^2(D_1, a_2)$ fluctuate when the number of parameters is larger than about 30

If too many parameters

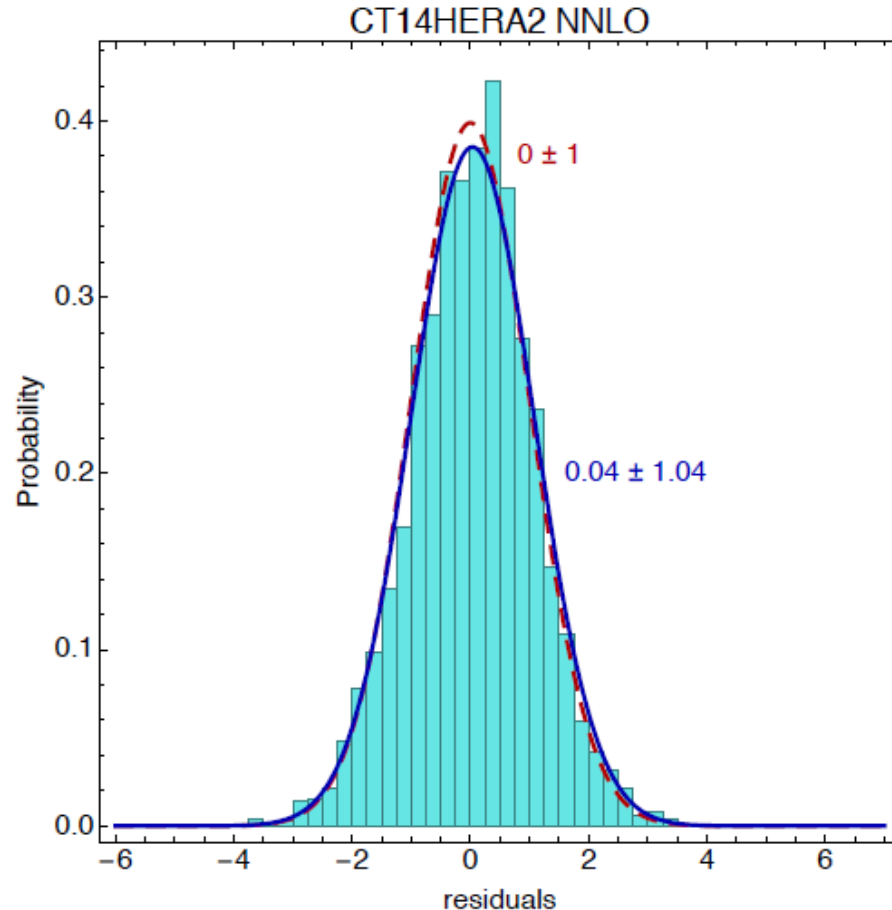


≈ 30 parameters (26 in CT14HERA2) is optimal for describing the CT14HERA2 data set. 15 parameters or less is optimal for nuclear PDFs

How well are the data described?

Note: It is convenient to define $S_n(\chi^2, N_{pt})$ that approximately obeys the standard normal distribution (mean=0, width=1) independently of N_{pt}

Example: data residuals r_n

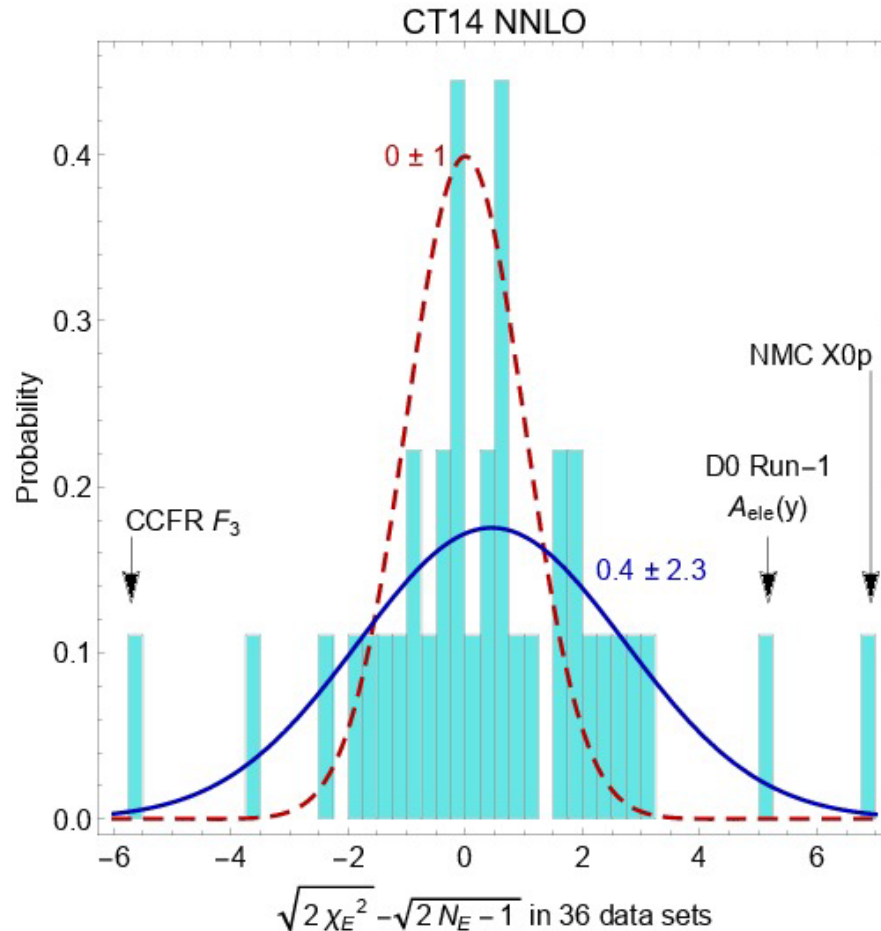


$$r_n \equiv \frac{T_n(\{a\}) - D_n^{shifted}(\{a\})}{\sigma_n^{uncorrelated}}$$

The distribution of residuals is consistent with the standard normal distribution

Full definition of r_n in the backup slides

Example: individual experiments



Define

$$S_n(\chi^2, N_{pt}) \equiv \sqrt{2\chi^2} - \sqrt{2N_{pt} - 1}$$

$S_n(\chi_n^2, N_{pt,n})$ are Gaussian distributed with mean 0 and variance 1 for $N_{pt,n} \geq 10$

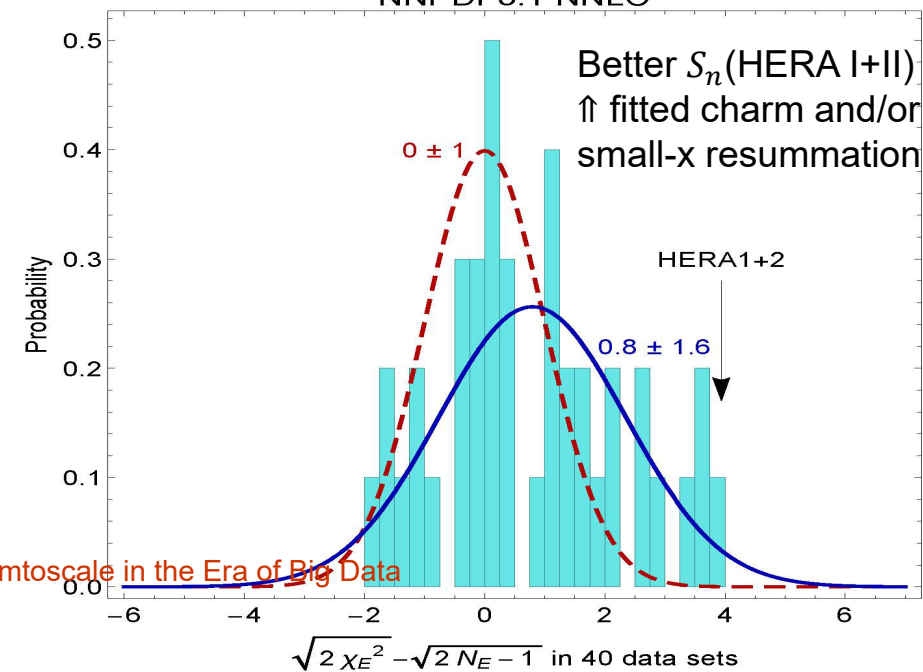
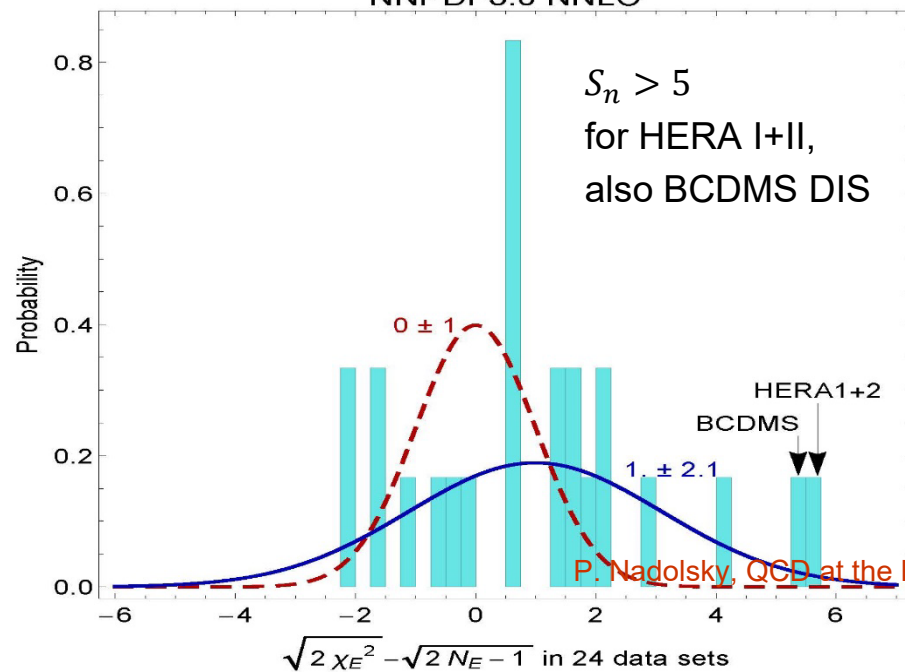
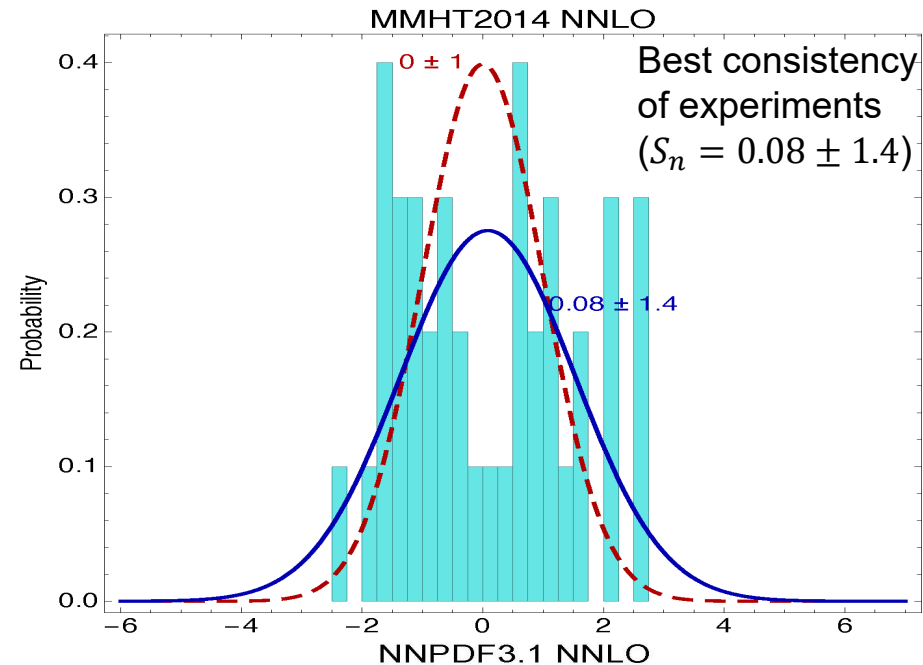
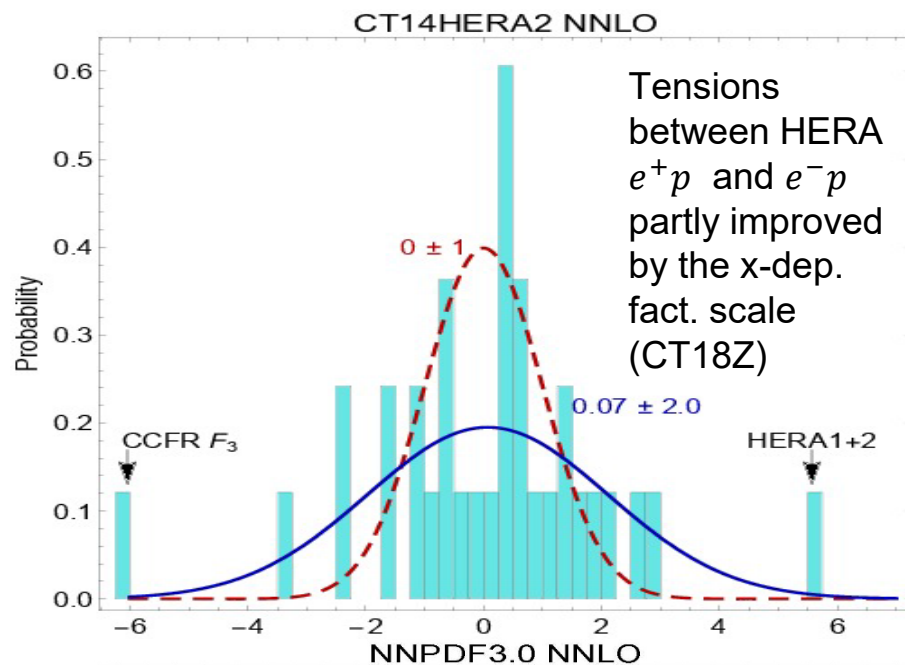
[R.A.Fisher, 1925]

Even more accurate (χ^2, N_{pt}) :

T.Lewis, 1988

An empirical S_n distribution can be compared to $N(0,1)$ visually or using a statistical (KS or related) test

Effective Gaussian variables



Epistemic PDF uncertainty: recap

Epistemic uncertainty (due to parametrization, methodology, parametrization/NN architecture, smoothness, data tensions, model for syst. errors, ...) is increasingly important in NNLO global fits as experimental and theoretical uncertainties decrease. We make progress in understanding it.

With $O(10 - 1000)$ free parameters, including nuisance parameters, the $\Delta\chi^2 = 1$ criterion for 1σ PDF uncertainties is almost certainly incomplete. Stop using it “as is”. There are strong mathematical reasons.

Nominal PDF uncertainties in high-stake measurements at the HL-LHC and EIC thus should be tested for *robustness of sampling over acceptable methodologies* and demonstrate *absence of biases* in this sampling.

Public tools for this are increasingly available: xFitter, NNPDF code, ePump, Fantômas, MP4LHC,...

Backup

AI/ML techniques are superb for finding an excellent fit to data.

Are these techniques adequate for uncertainty estimation [exploring all good fits]?

A common resampling procedure used by experimentalists and theorists:

1. Train a neural network model T_i with N_{par} (hyper)parameters on a randomly fluctuated replica of discrete data D_i . Repeat N_{rep} times. In a typical application: $N_{\text{par}} > 10^2$, $N_{\text{rep}} < 10^4$.
2. Out of N_{rep} replicas T_i with “good” description of data [i.e., with a high likelihood $P(D_i|T_i) \propto e^{-\chi^2(D_i,T_i)/2}$], discard “badly behaving” (overfitted, not smooth, ...) replicas
3. Estimate the uncertainties of T_i using the remaining “well-behaved” replicas

Is this procedure rigorous? How many N_{rep} replicas does one need?

A likelihood-ratio test of NN models T_1 and T_2

From Bayes theorem, it follows that

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

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

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

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

aleatory

epistemic + aleatory

Suppose replicas T_1 and T_2 have the same χ^2 [$r_{\text{likelihood}} = \exp\left(\frac{\chi_1^2 - \chi_2^2}{2}\right) = 1$], but T_2 is disfavored compared to T_1 [$r_{\text{posterior}} \ll 1$].

This only happens if $r_{\text{prior}} \ll 1$: T_2 is discarded based on its **prior** probability.