

# Insights for the PDF4EIC effort from HEP, lattice QCD, and radiology

**Pavel Nadolsky**

Southern Methodist University

# PDF4LHC

**Uncertainty quantification**

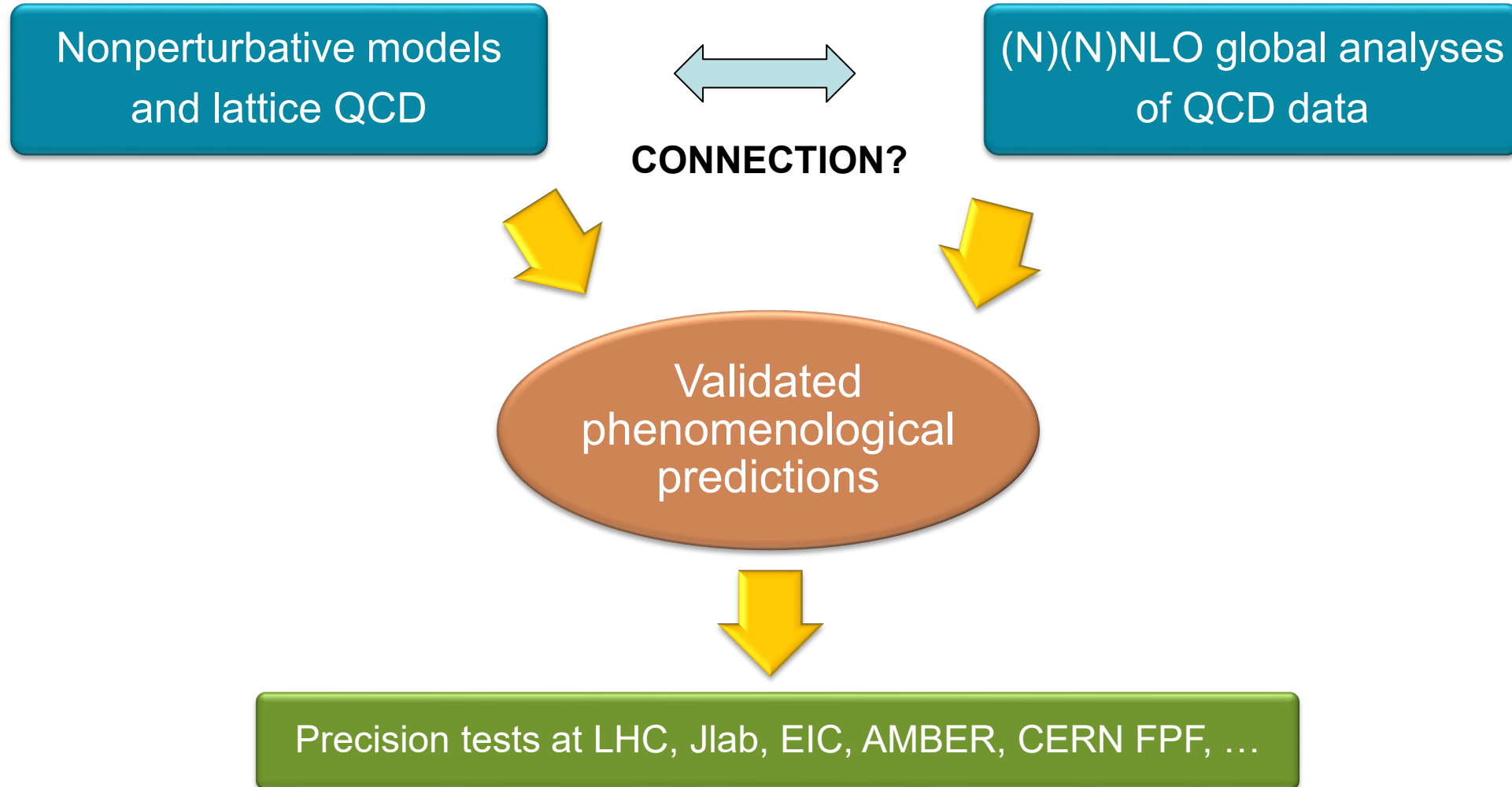
**Reproducibility & replicability**

**AI across STEM**

**New funding paradigm**

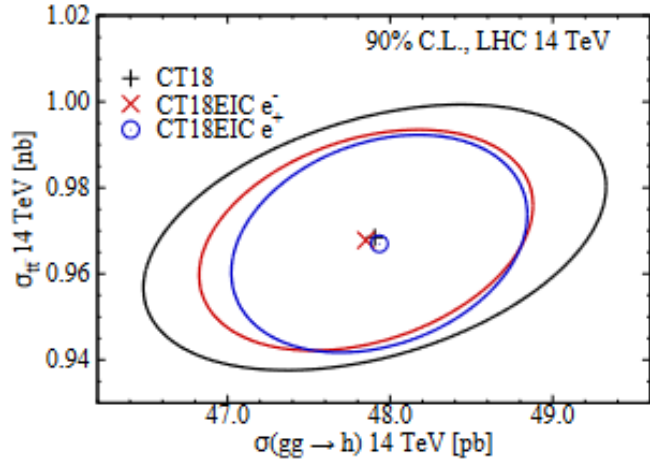


# New insights about 3-dimensional structure of hadrons

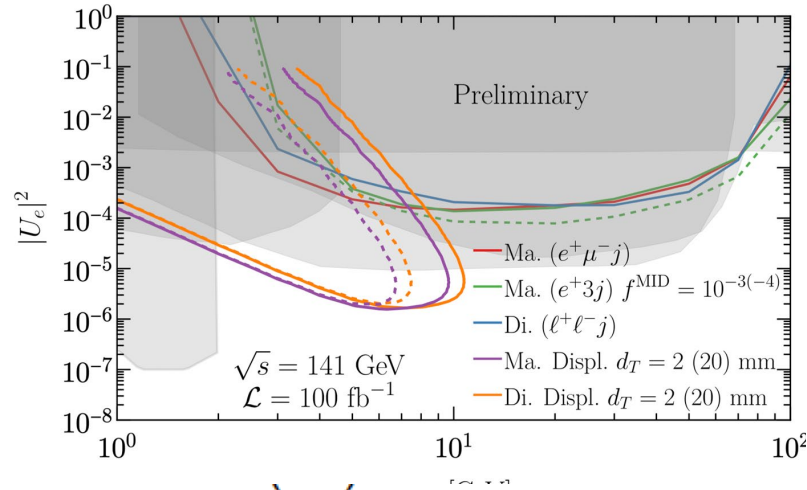


# Electron-Ion Collider: potentially a wealth of complex studies

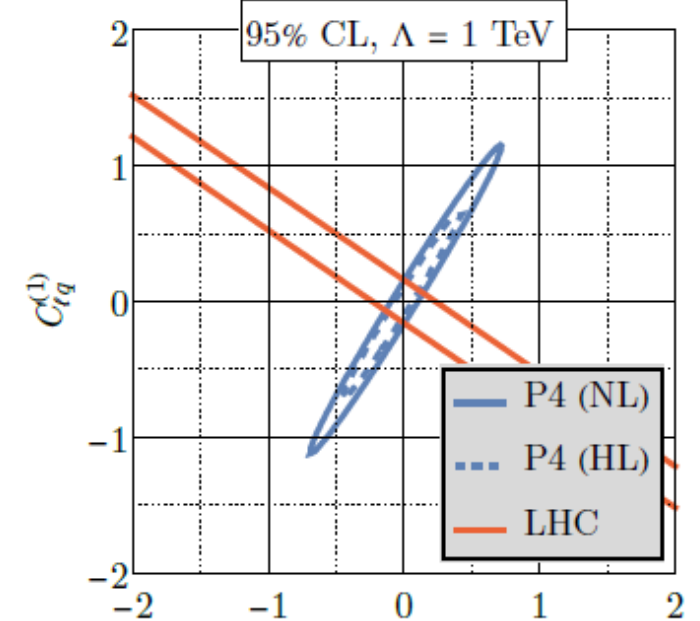
PDFs: arXiv:2103.05419



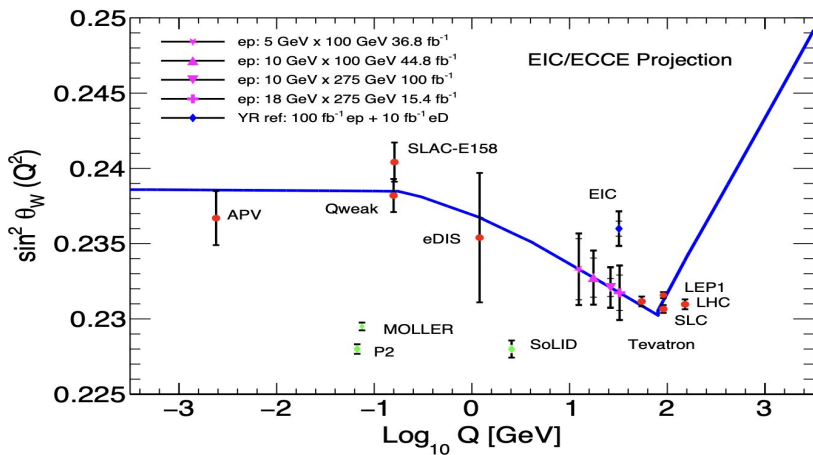
heavy neutral lepton searches arXiv: 2203.06705



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

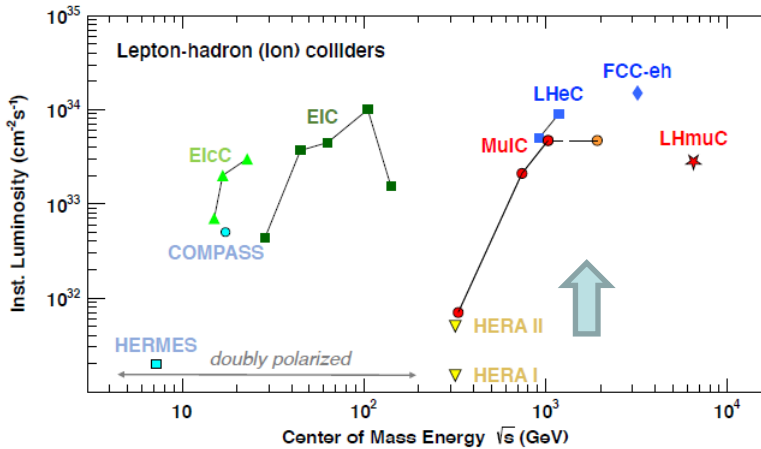
Lorentz/CPT violations  
 A. R. Vieira et al., 1911.04002

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

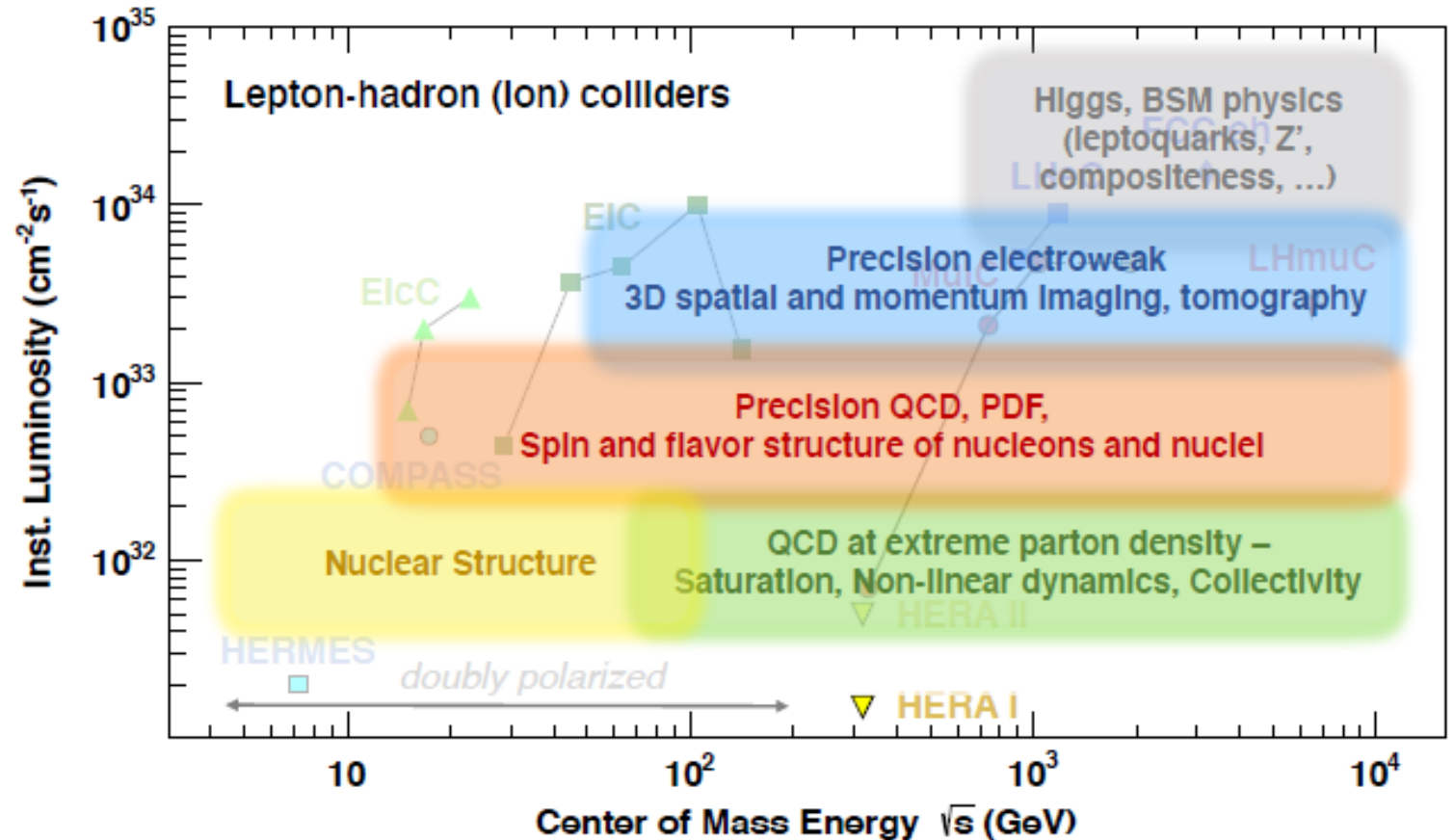
# The Muon-Ion Collider, Large Hadron Electron Collider, FCC-eh

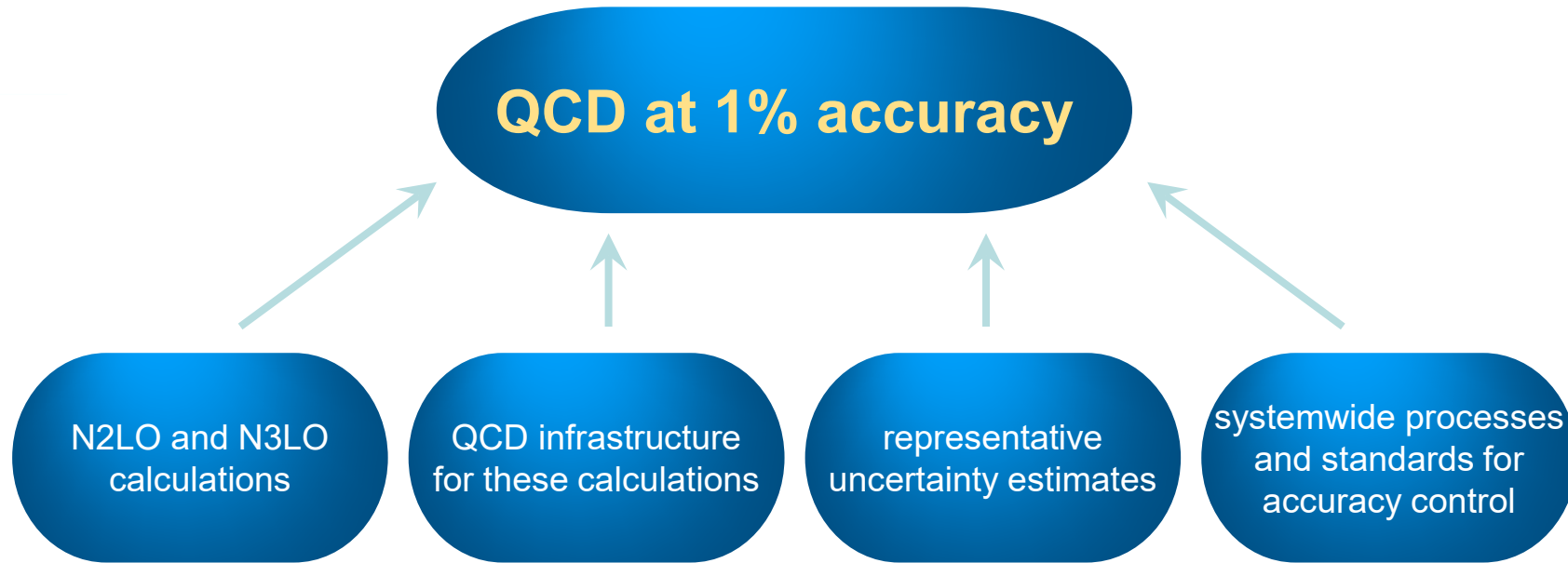
D. Acosta et al., “The Potential of a TeV-Scale Muon-Ion Collider,” [arXiv:2203.06258](https://arxiv.org/abs/2203.06258) [hep-ph]

LHeC, FCC-he Study Group, [arXiv:1206.2913](https://arxiv.org/abs/1206.2913), [2007.14491](https://arxiv.org/abs/2007.14491)



Exceptional machines for BSM discoveries, Higgs physics such as measurement of  $\kappa_{H \rightarrow c\bar{c}}$ , and SM tests at (sub)percent precision





Lots of promise in this area

Parton showers, fast NxLO interfaces, PDFs, ... must be comparably accurate

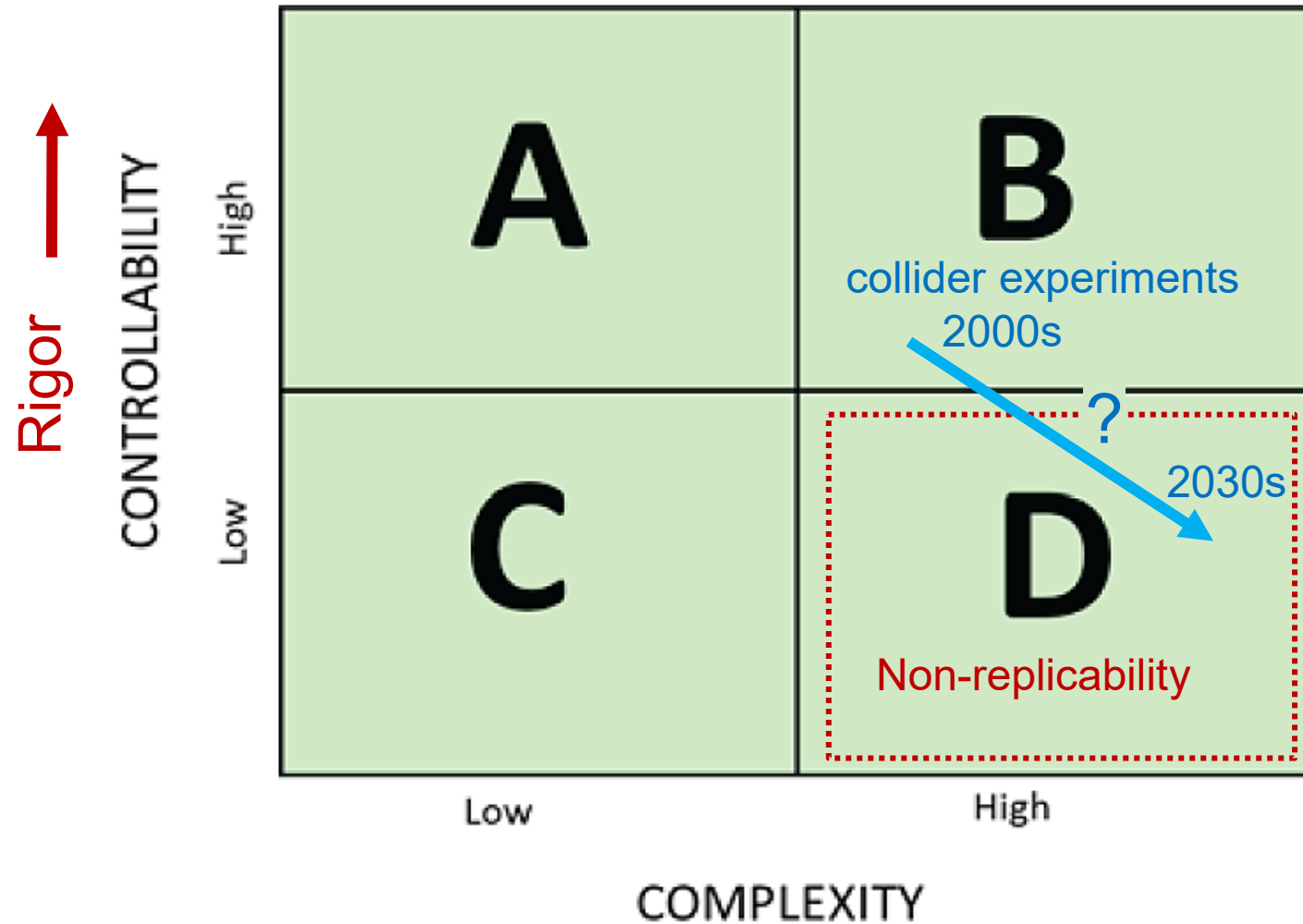
or The Importance of Being Earnest with Systematic Errors (experiment+theory; traditional or AI/ML)

This must be a part of the precision-focused community culture

Publishing statistical models: Getting the most out of particle physics experiments  
 Kyle Cranmer (New York U.), Sabine Kraml (LPSC, Grenoble), Harrison B. Prosper (Florida State U.), Philip Bechtle (Bonn U.), Florian U. Bernlochner (Bonn U.) [Show All\(33\)](#)  
 Sep 10, 2021  
 60 pages  
 Published in: *SciPost Phys.* 12 (2022) 1, 037, *SciPost Phys.* 12 (2022) 037  
 Published: Jan 25, 2022  
 e-Print: 2109.04981 [hep-ph]

2023 US DOE Funding Opportunity Announcement  
 DE-FOA-0000315  
**Advancing *Uncertainty Quantification* in Modeling, Simulation, and Analysis of *Complex Systems***

# A looming risk for particle physics



Based on Fig. 5.2 in  
"REPRODUCIBILITY AND  
REPLICABILITY IN SCIENCE"

# Reproducibility, Replicability, Rigor: definitions

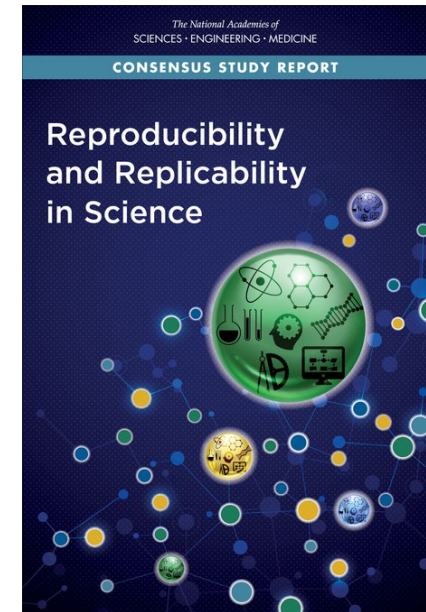
**Reproducibility** is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis.

**Replicability** is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

**Rigor** -- the strict application of the scientific method to ensure robust and unbiased experimental design -- makes replication of a study more likely

Definitions adopted from “*REPRODUCIBILITY AND REPLICABILITY IN SCIENCE*”, Conclusion 3.1  
*National Academy of Sciences, 2019, <https://doi.org/10.17226/25303>*

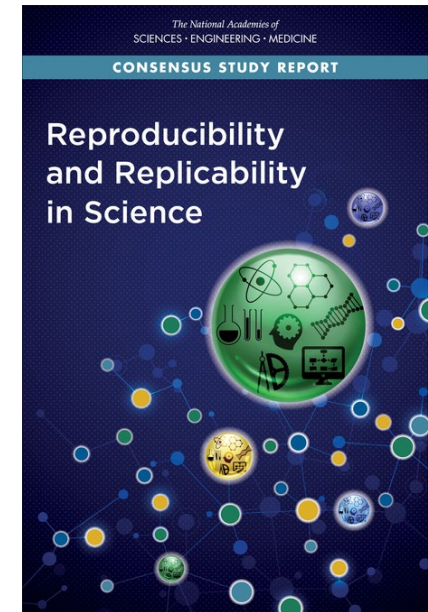
# RRR



# Universal factors affecting replicability

- complexity of the system under study;
- understanding of the number and relations among variables within the system under study;
- ability to control the variables;
- levels of noise within the system (or signal to noise ratios);
- mismatch of scale of the phenomena and the scale at which it can be measured;
- stability across time and space of the underlying principles;
- fidelity of the available measures to the underlying system under study (e.g., direct or indirect measurements);
- prior probability (pre-experimental plausibility) of the scientific hypothesis.

From “*REPRODUCIBILITY AND REPLICABILITY IN SCIENCE*”  
National Academy of Sciences, 2019, <https://doi.org/10.17226/25303>





# Strategies for improving replicability and reproducibility

Preselection of planned studies based on their likely replicability

Detailed documentation of methods and uncertainty quantification in the publications

Training of researchers in relevant statistical methods

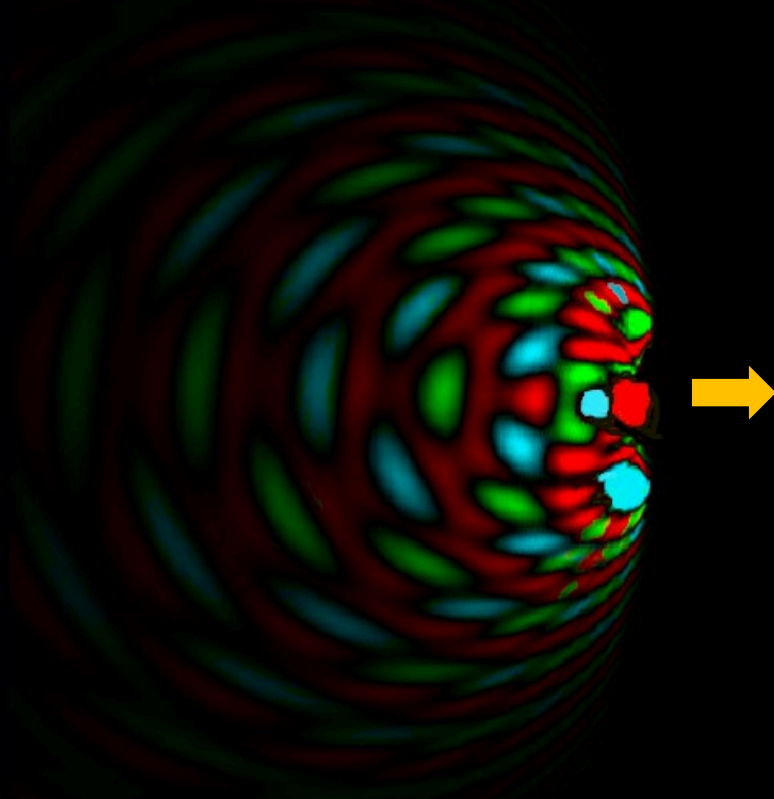
Journal policies that encourage replicability

Support from the funding agencies for the research infrastructure and collaborations focusing on replicability

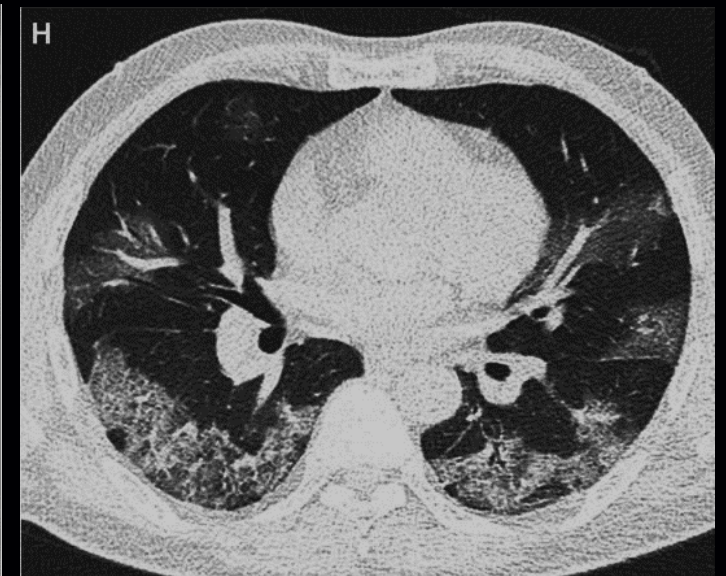
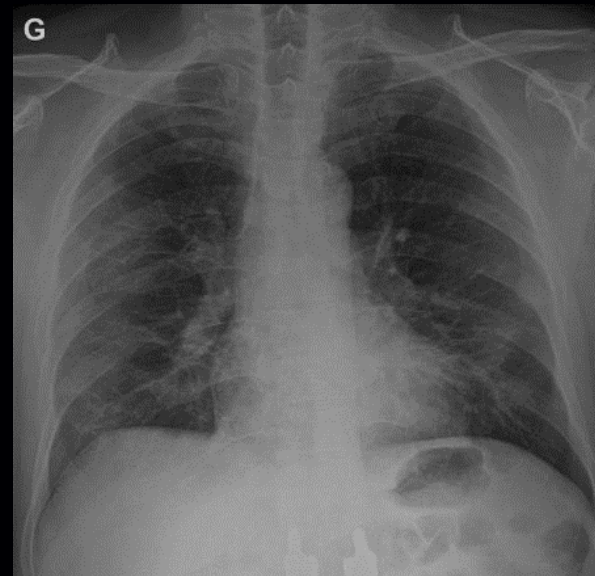
Support for open publication of the analysis codes and key data, using agreed-upon formats

“Skin-in-the-game” incentives for researchers to produce replicable results

# Particle physicists and radiology doctors infer from complex images



A proton at an *ep* collider moving with speed  $V \approx c$  to the right



A 3-dim tomographic image of a COVID-19 patient

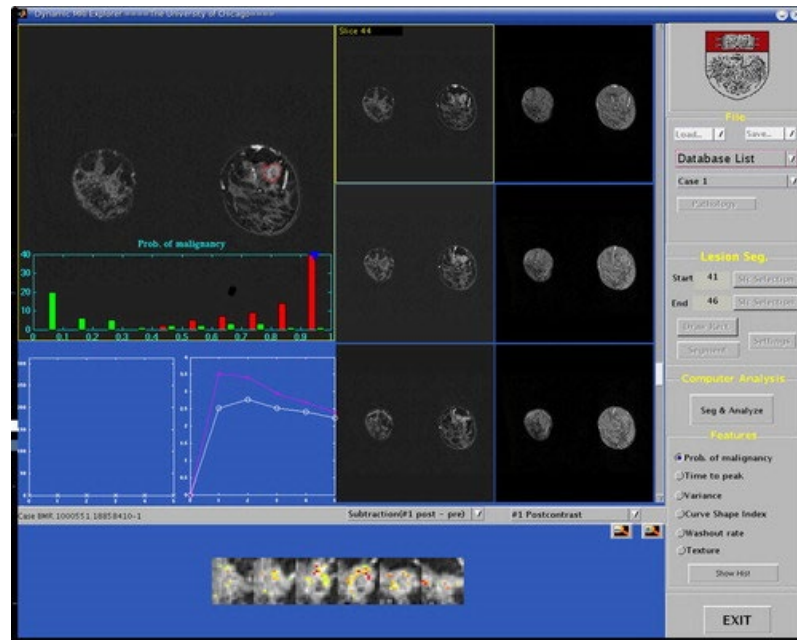
3-dim hadron femtography at the EIC

# Particle physicists and radiology doctors address analogous questions in statistics

> Radiology. 2011 Mar;258(3):696-704. doi: 10.1148/radiol.10100409. Epub 2011 Jan 6.

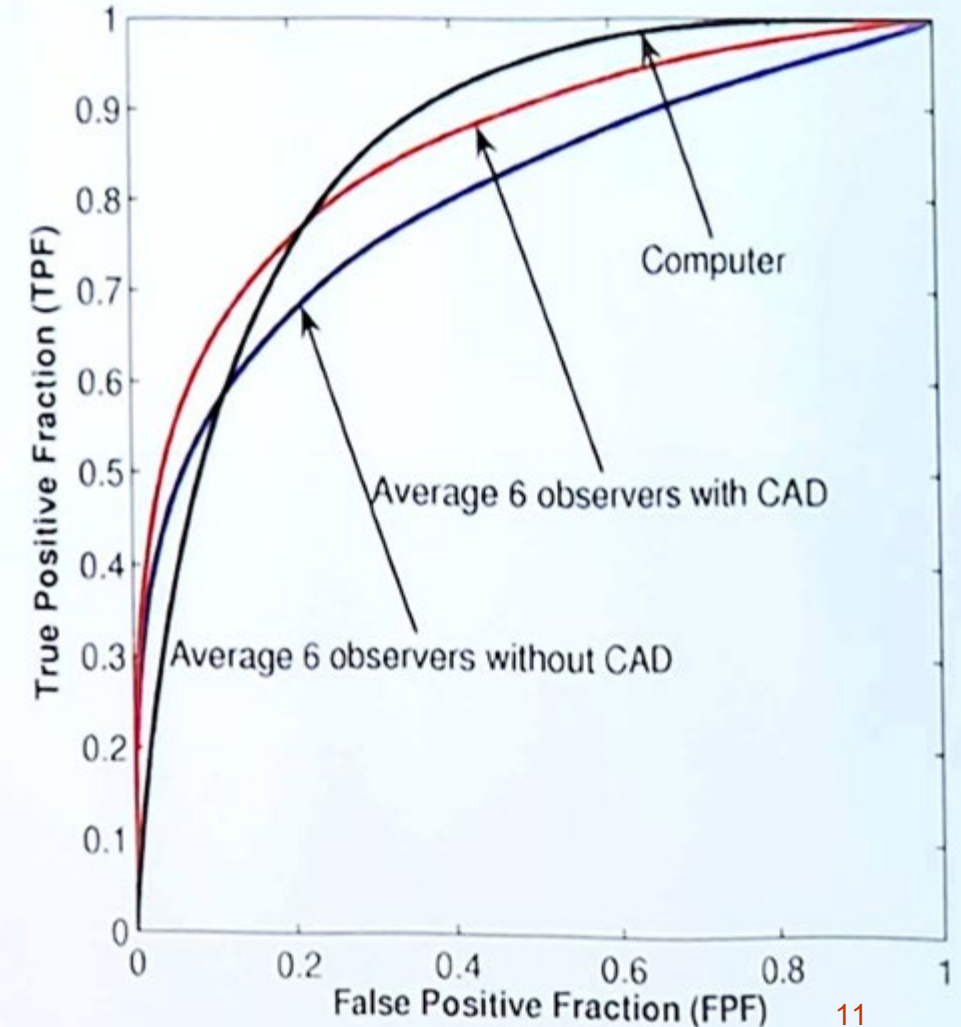
## Evaluation of clinical breast MR imaging performed with prototype computer-aided diagnosis breast MR imaging workstation: reader study

Akiko Shimauchi<sup>1</sup>, Maryellen L Giger, Neha Bhooshan, Li Lan, Lorenzo L Pesce, John K Lee, Hiroyuki Abe, Gillian M Newstead



**CADx:** Task of Distinguishing between Malignant & Benign Lesions on **Breast MRI**

### Performance of the End User



M. Giger (U Chicago), FNAL, July 20, 2023

# AI and replicability in radiology

“Within health care, the US Food and Drug Administration has already cleared 523 devices that use AI—75% of them for use in radiology.”

“... AI can help verify what we already know by addressing **science’s replicability crisis**. **Around 70% of scientists report having been unable to reproduce another scientist’s experiment—a disheartening figure**. As AI lowers the cost and effort of running experiments, it will in some cases be easier to replicate results or conclude that they can’t be replicated, contributing to a greater trust in science.”

[Eric Schmidt, This is how AI will transform the way science gets done, MIT Technology Review, 2023-07-05](#)

Statistical inference from hadron scattering data and medical images bear many similarities. The medical community working on AI is very large and well-funded.

During the COVID-19 pandemic, thousands of medical AI applications were developed to diagnose and cure the disease. Most have failed.

In response to this replicability crisis, the US medical community took numerous actions to implement systemwide infrastructure, standards, and procedures for organizing the data and quantifying uncertainties in AI-assisted analyses.

# What went wrong with AI/ML methods for COVID imaging?

MIT  
Technology  
Review

Featured Topics Newsletters Events Podcasts

Artificial intelligence / Machine learning

400+

## Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

nature machine intelligence

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature machine intelligence > analyses > article

Analysis | Open Access | Published: 15 March 2021

### Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts , Derek Driggs, Matthew Thorpe, Julian Gibbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence 3, 199–217 (2021) | Cite this article

## 1) Poor quality of COVID imaging data

- Mislabeled data
- Multiple unknown sources
- Duplicate data (training and testing)
- No traceability, limited quality control
- Lack of external validation

## 2) Lack of collaboration/communication between AI/ML experts and biomedical experts

- Need for valid ground truth
- Need for independent test set

## 3) Bias and diversity

- Data collected for a specific clinical task
- Specific populations, lack of diversity
- Single expert score, data sources correlated with 'truth', ...



# MIDRC

---

MEDICAL IMAGING AND DATA RESOURCE CENTER.

<https://www.midrc.org/>

Medical Imaging Community in response to the COVID-19 pandemic

M. Giger (U Chicago), FNAL, July 20, 2023



NIBIB COVID-19 Contract 75N92020D00021

- An open, curated, diverse image data commons
- A partnership between the AAPM, ACR and RSNA, supported by NIBIB, hosted at University of Chicago, and on the Gen3 data platform

## Two scientific components of MIDRC:

1. Open Discovery Data Commons
2. Machine Intelligence Computational Capabilities

3. The center uses a private subset of data to validate statistical rigor and replicability of the proposed (AI-assisted or not) algorithms

# MIDRC by the Numbers

**309,270**

Imaging Studies  
Ingested

**152,772**

Imaging Studies  
released to the  
Public

**156,498**

Imaging Studies  
undergoing quality &  
harmonization

**377**

Total Data  
Downloads  
this month

**60,801**

Cases

**13.27 TB**

Total size  
Published

**47**

Publications

**115+**

Presentations

**29**

Algorithms

**586**

Registered  
Users

**100+**

Investigators

**416**

Collaborating  
Institutions

To date, MIDRC has focused on medical imaging and data of COVID-19 patients, and the imaging studies made available to the public have mainly been chest imaging. Currently, however, in order to keep up with developments in the pandemic, imaging studies associated with post acute sequelae of COVID-19 (PASC, 'long COVID') are actively being collected as well as different imaging modalities and various additional organ systems (such as the heart or brain).

Future plans include expansion to become a wider comprehensive resource for all Institutes at NIH, with focused medical imaging data commons of chronic disease (e.g., diabetes, chronic liver disease, coronary artery disease, COPD and emphysema), and other infectious pandemics. Additionally, the sequestering of some of the MIDRC data in a separate data commons not accessible to the public for validation and testing will provide a valuable resource for data science challenges and a path to long-term sustainability through industry support for translation to - and approval of - clinical use which will impact public health worldwide. [Learn more...](#)

<https://www.midrc.org/>, accessed on 2023-09-17



New DIS and forward physics experiments that run **concurrently** with, or after the HL-LHC, open unique opportunities to understand nonperturbative QCD. They also create a strong synergistic effect in both SM and BSM studies

Progress on this program, especially in precision measurements, increasingly depends on cross-cutting research and replicability of complex measurements

**Precision QCD may soon walk into a replicability crisis. The experience from radiology and other fields suggests community-wide strategies for avoiding it.**



# The PDF4LHC working group

<https://www.hep.ucl.ac.uk/pdf4lhc/>

The PDF4LHC working group...

- ...was formed in the early 2010's to advise on development and applications of nucleon PDFs at the LHC
- ...includes a steering committee and participants from all major PDF fitting groups
- ...publishes a periodic recommendation on the PDF applications for a wide range of LHC users

See “**The PDF4LHC21 combination of global PDF fits for the LHC Run III**”, R. Ball et al., [2203.05506](#)

- ...provides combinations of PDF error sets to streamline estimates of PDF uncertainties for most LHC applications;

such PDF4LHC PDFs are constructed using the Hessian→MC conversion (*G. Watt, R. Thorne, [1205.4024](#); T.Hou et al., [1607.06066](#)*) and either the META (*J. Gao, P. Nadolsky, [1401.0013](#)*) or MC2Hessian (*S. Carazza et al., [1505.06736](#), [1602.00005](#)*) combination algorithms

- ...performs **benchmarking comparisons of fitting codes** and other validations aimed to improve **reproducibility and replicability** of global analyses
- ... can serve as a model for an EIC-centered effort

# Possible activities for the PDF4EIC community

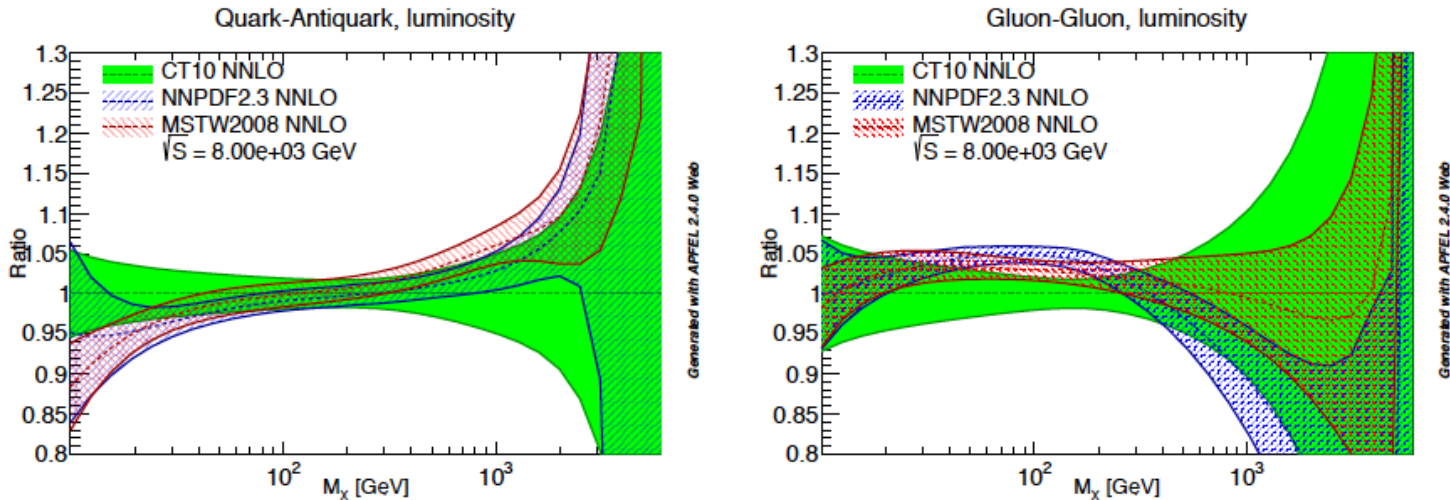
- common physics goals
  - ⇒ learn about the 3D hadron structure!
- shared resources
  - LHAPDF-like repository for interpolations of polarized/TMD/GPD PDFs? For  $\chi^2$  values for error PDFs? Other outputs of the fits?
  - Coordinated software development for global fits?
- agreed-upon practices
  - presentation of data and theory predictions? RIVET for the EIC?
  - common definitions of PDF uncertainties?
  - a common standard for PDF validation tests?
- benchmarking studies
  - explore experimental constraints on various types of PDFs and from various available and future processes at (N)(N)LO using the  $L_2$  sensitivity and other techniques



# Replicability of PDF uncertainties from 2012 to 2023

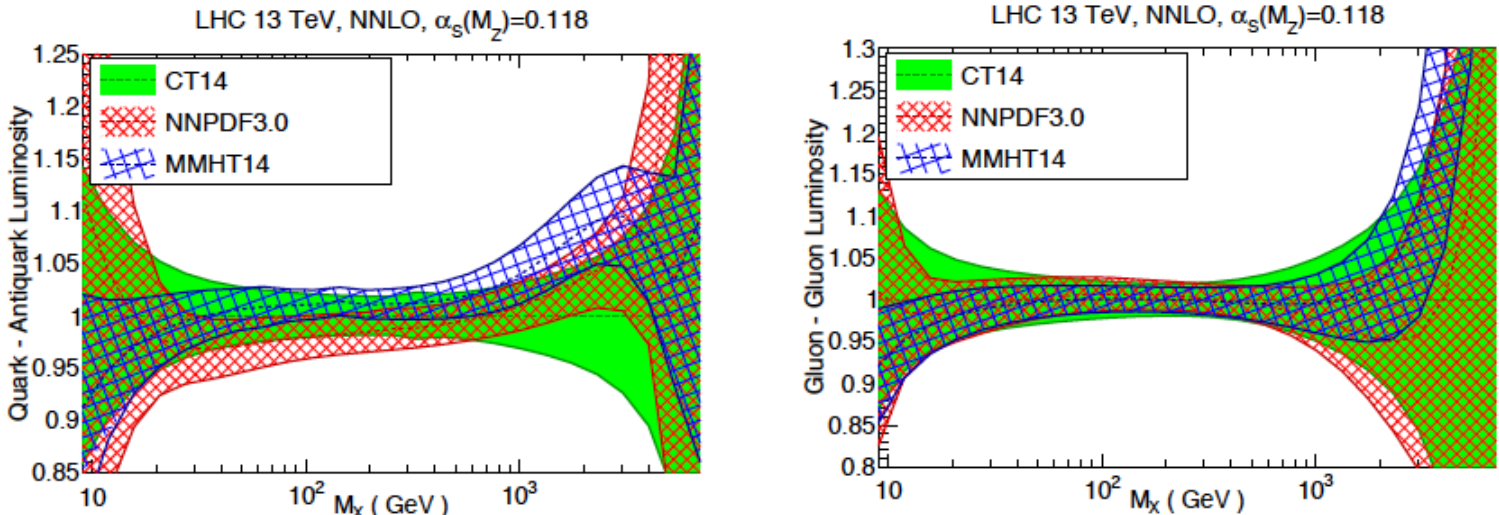
# 2012→2015: Agreement between NNLO PDFs greatly improved

2015



Note in particular the changes in the  $gg$  luminosity, especially important in the Higgs mass region

Figure 1: Comparison of the  $q\bar{q}$  (left) and  $gg$  (right) PDF luminosities at the LHC 8 TeV for CT10, MSTW2008 and NNPDF2.3. Results are shown normalized to the central value of CT10.



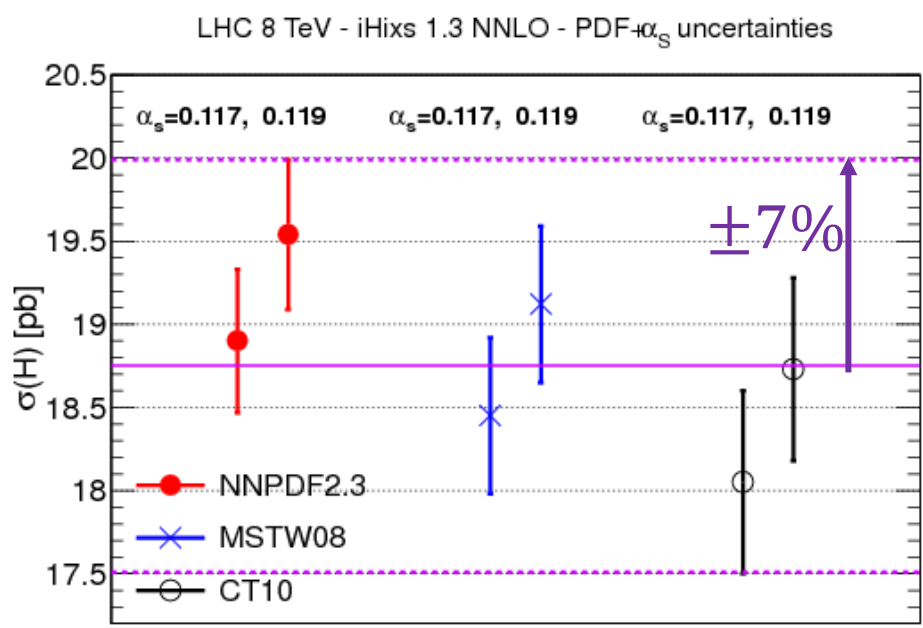
**LHC data has been added for all 3 new PDFs, but most of change is due to changes in formalisms**

# PDF4LHC15 benchmarking of codes reduced the PDF error on Higgs cross sections

2015

2012:  $\delta_{PDF} \approx 7\%$

2015:  $\delta_{PDF} \approx 2 - 3\%$



R. Ball et al., arXiv:1211.5142

Disagreement in central values

## $\sigma(gg \rightarrow H^0)$ at NNLO

	CT14	MMHT2014	NNPDF3.0
8 TeV	18.66 pb	18.65 pb	18.77 pb
	-2.2%	-1.9%	-1.8%
	+2.0%	+1.4%	+1.8%
13 TeV	42.68 pb	42.70 pb	42.97 pb
	-2.4%	-1.8%	-1.9%
	+2.0%	+1.3%	+1.9%

J.Huston, PDF4LHC, April 2015

Good agreement of central values

**N3LO scale dependence on  $\sigma_H$  is <3%**

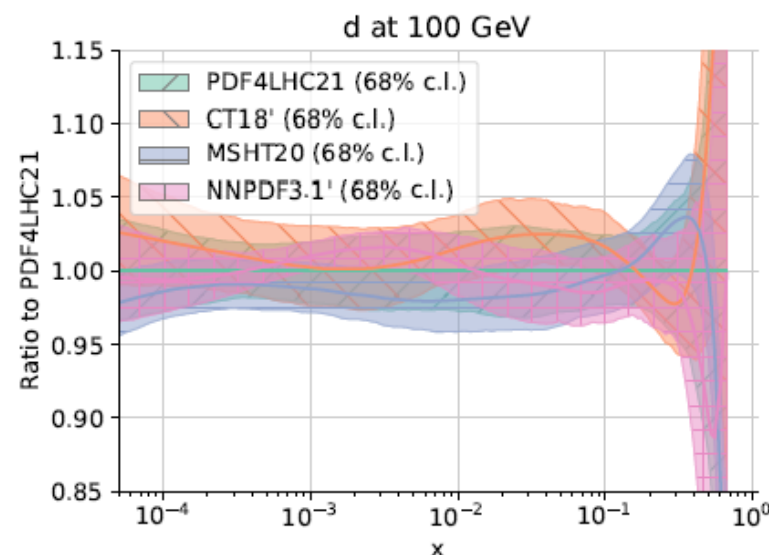
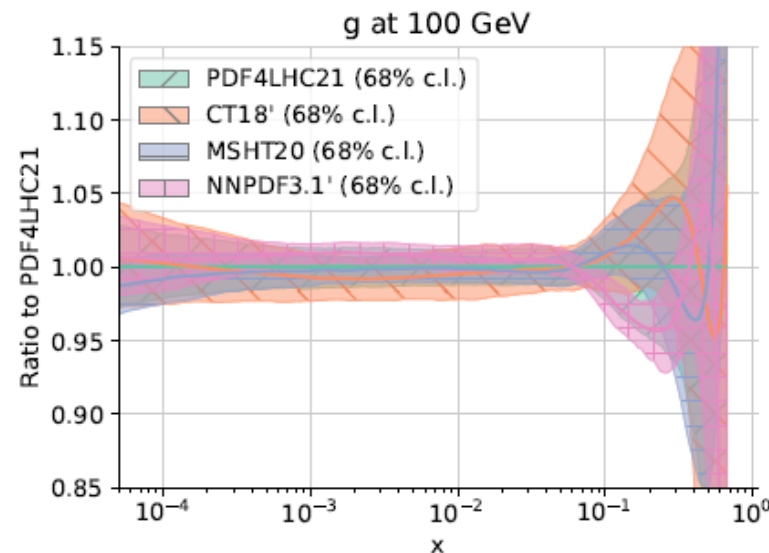
**Similar agreement for  $t\bar{t}$  cross sections**

# PDF4LHC21 recommendation and combined PDFs

2022

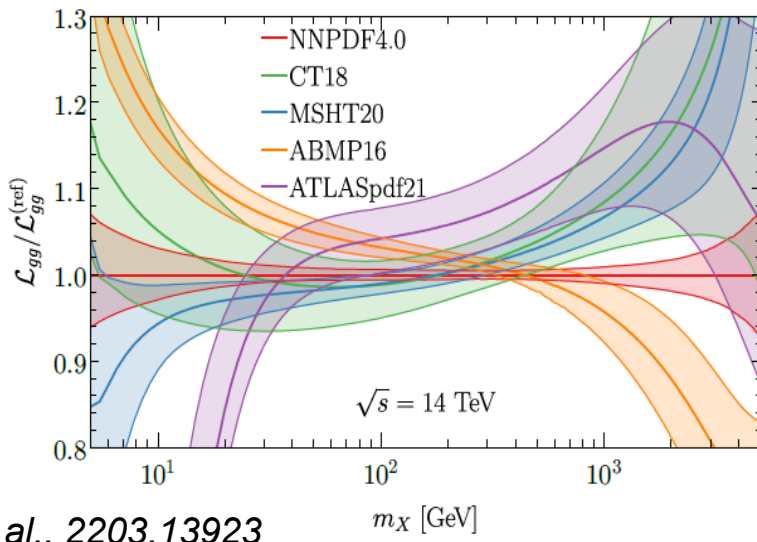
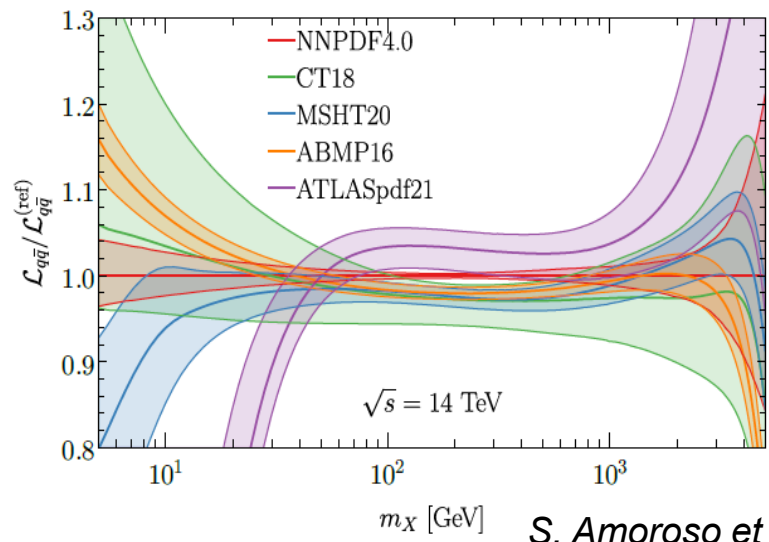
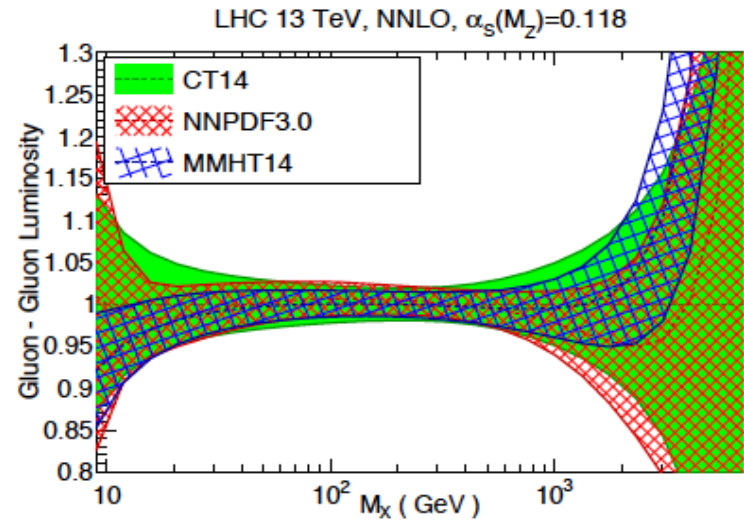
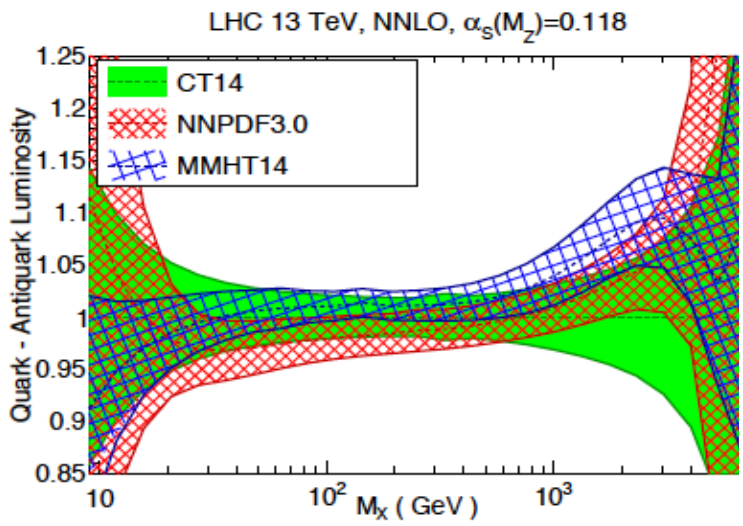
- A comprehensive recommendation for usage of PDFs at the LHC
- Replaces the PDF4LHC15 recommendation
- A detailed benchmarking comparison of global fits by three main groups
- Combined PDF4LHC21 NNLO PDFs based on CT18', MSHT'20, and NNPDF3.1.1 ensembles for BSM searches, measurements of moderate precision, theory predictions
- Provided as 40-member Hessian PDFs and 100-member Monte-Carlo PDFs of comparable accuracy

arXiv:2203.05506



# 2015→2023: The agreement of NNLO proton PDFs got worse, not better

2023



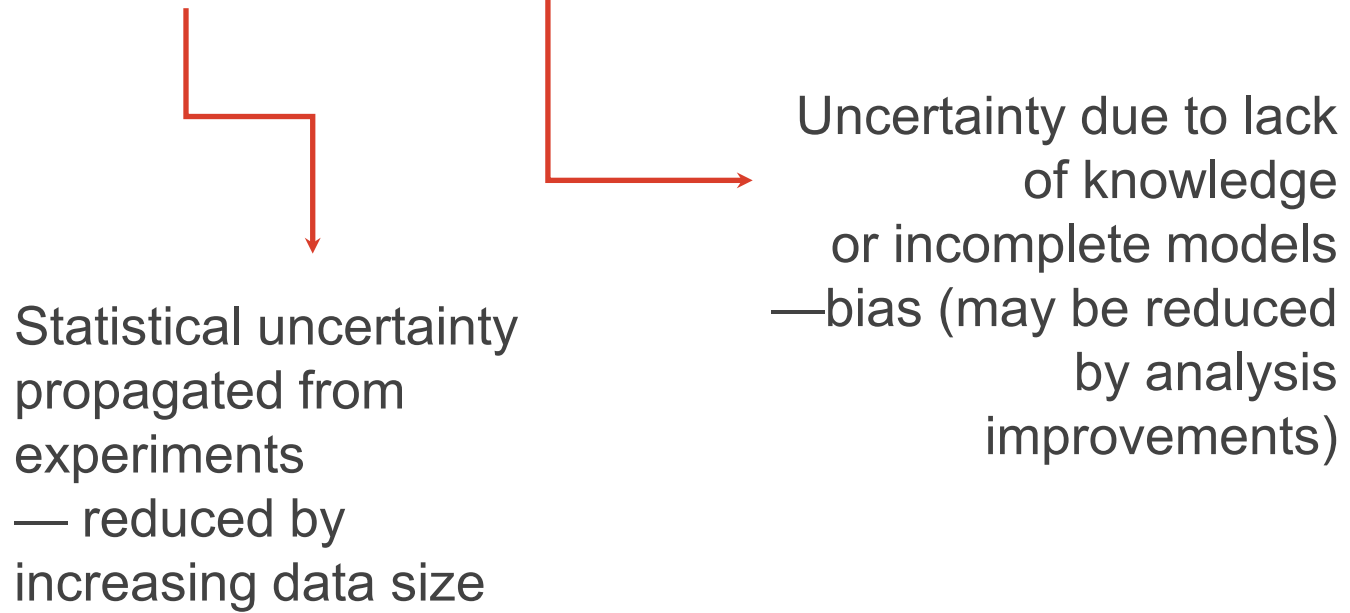
S. Amoroso et al., 2203.13923

The fitting groups and PDF4LHC21 study identified some possible reasons:

1. insufficient agreement between the fitted experiments (**systematic uncertainties**)
2. differences in the fitting methodologies (**tolerance**)
3. **more fundamental reasons**



## aleatory vs. epistemic uncertainties



# Epistemic PDF uncertainty...

...reflects **methodological choices** such as PDF functional forms or NN architecture and hyperparameters.

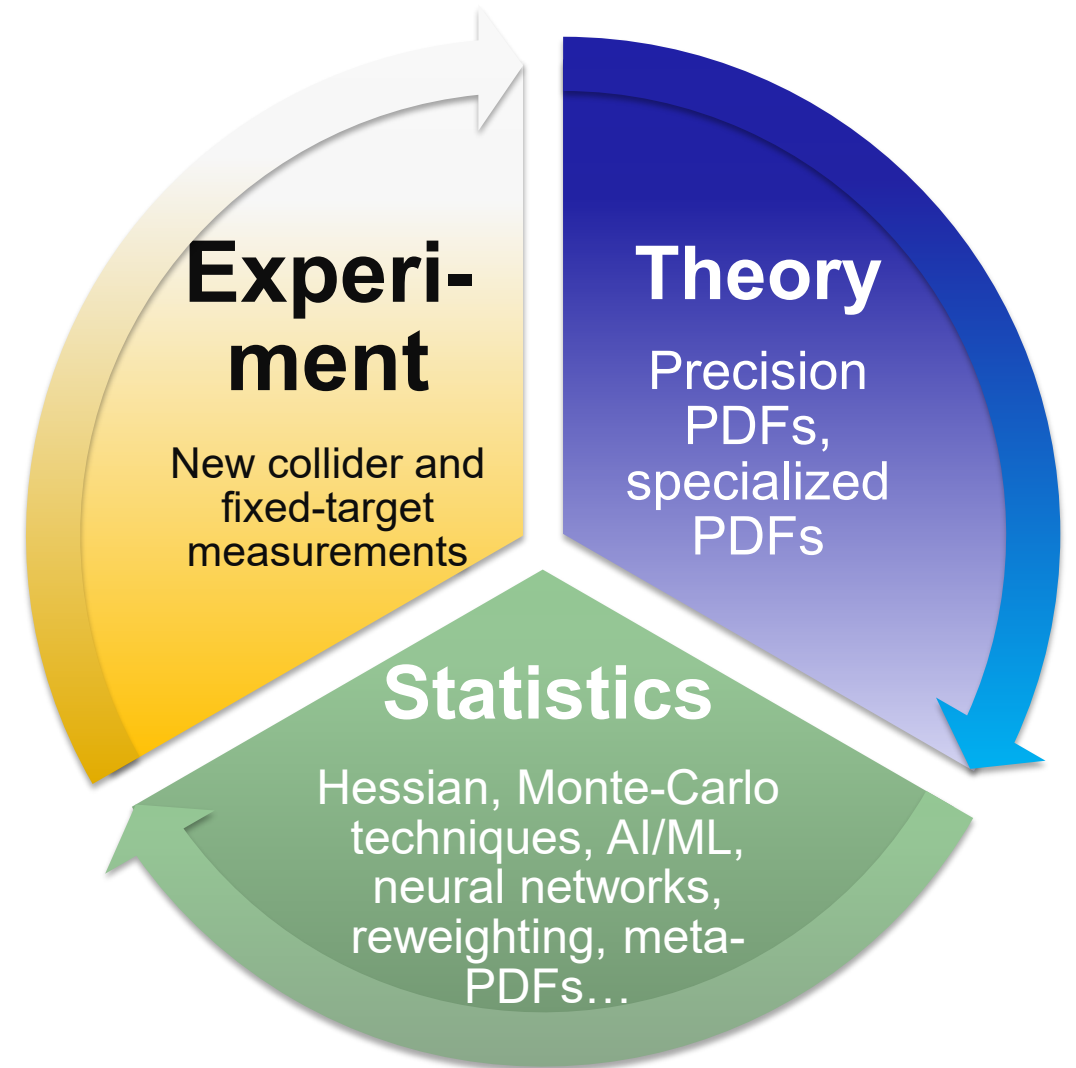
... can dominate the full uncertainty when experimental and theoretical uncertainties are small.

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

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

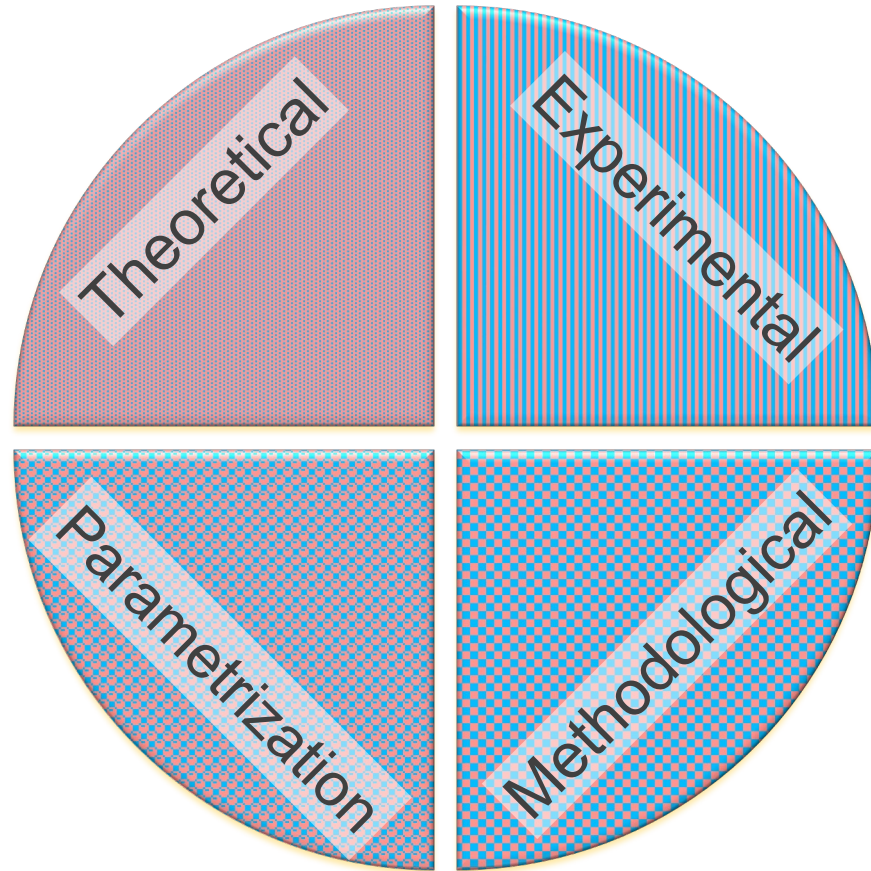
⇒ sampling over choices of experiments, PDF/NN functional space, models of correlated uncertainties...

⇒ in addition to sampling over data fluctuations





Components of a global QCD fit

## Components of PDF uncertainty



In each category, one must maximize

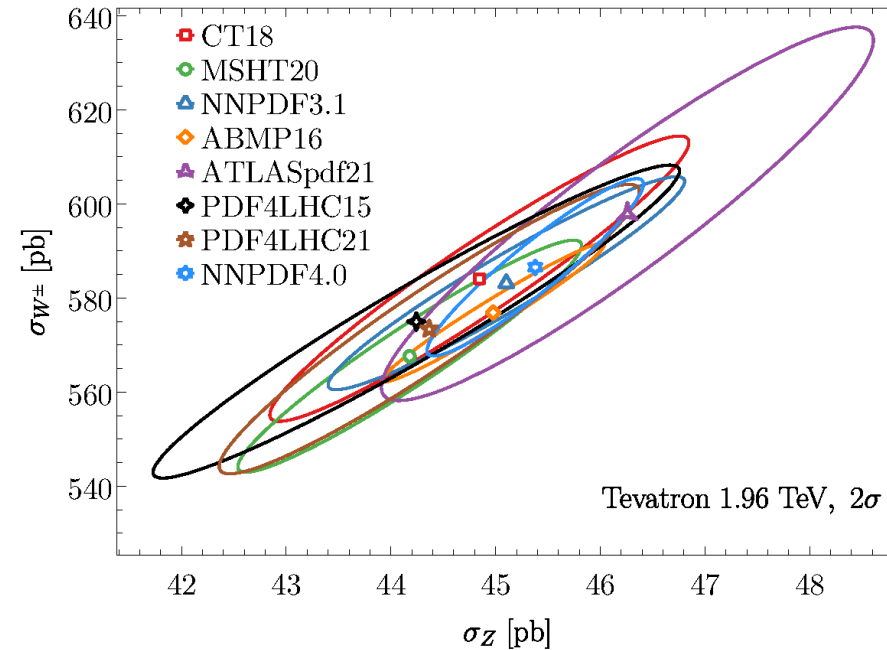
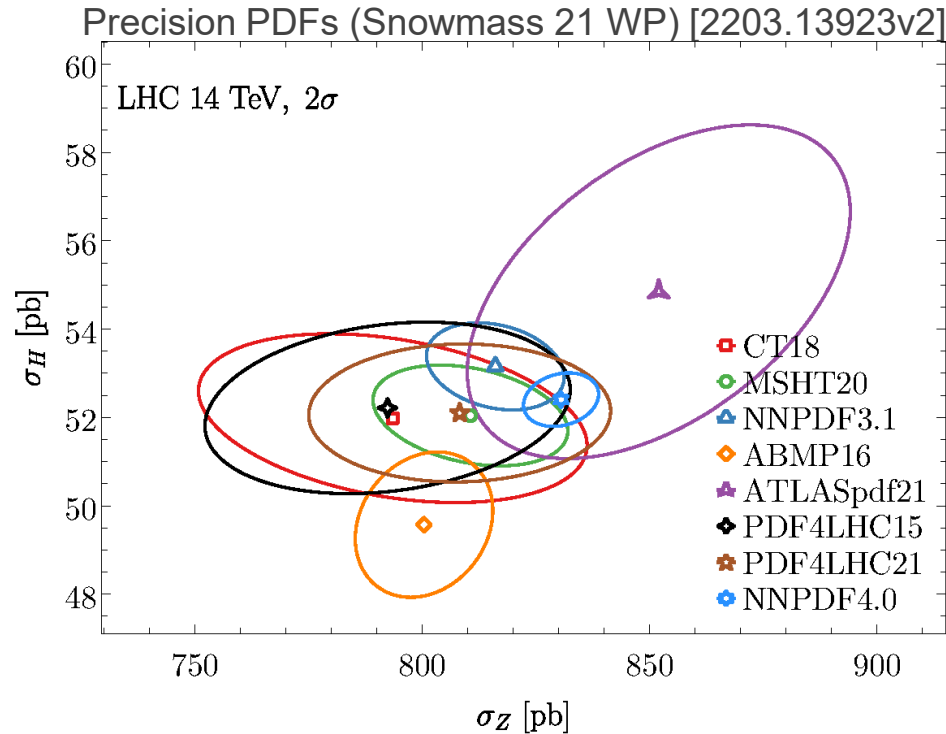
 **PDF fitting accuracy**  
(accuracy of experimental, theoretical and other inputs)

 **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]

# The tolerance puzzle

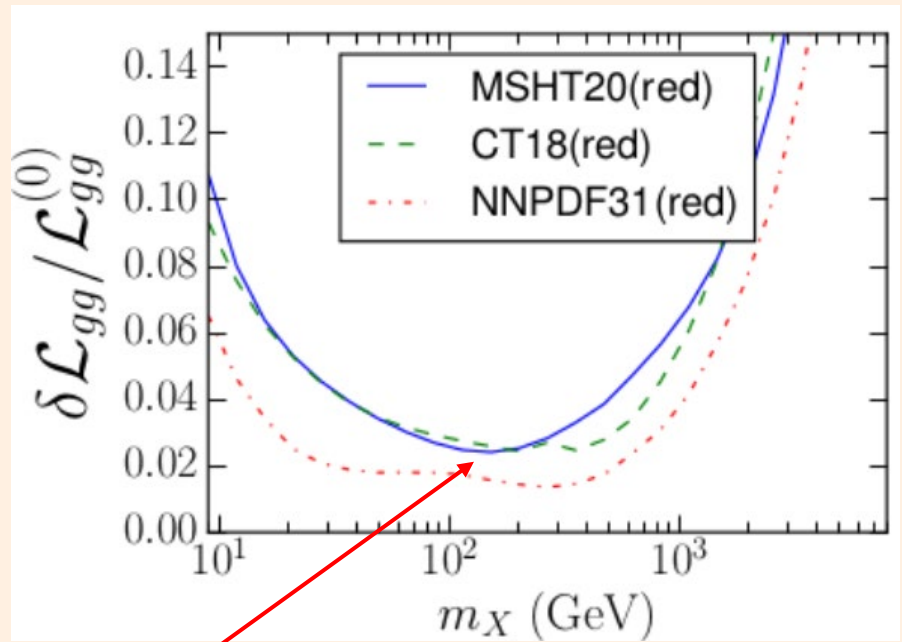
Why do groups fitting similar data sets obtain different PDF uncertainties?



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

# Tolerances explained by epistemic uncertainties

Relative PDF uncertainties on the  $gg$  luminosity at 14 TeV in three PDF4LHC21 fits to the **identical** reduced global data set



× 1.5 – 2 difference

While the fitted data sets are identical or similar in several such analyses, the differences in uncertainties can be explained by methodological choices adopted by the PDF fitting groups.

NNPDF3.1' and especially 4.0 (based on the NN's+ MC technique) tend to give smaller nominal uncertainties in data-constrained regions than CT18 or MSHT20

**Epistemic uncertainties explain many such differences**

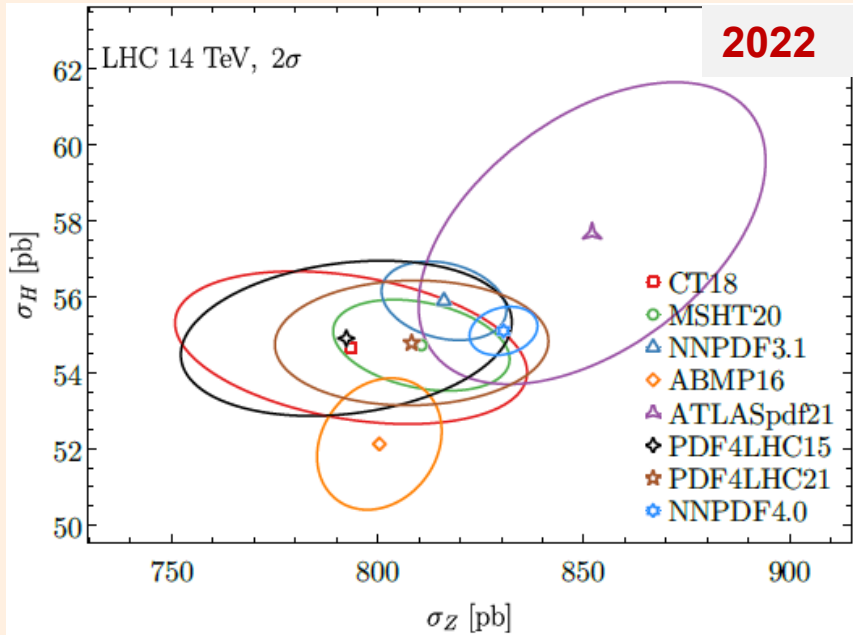
Details in [arXiv:2203.05506](https://arxiv.org/abs/2203.05506), [arXiv:2205.10444](https://arxiv.org/abs/2205.10444)

More in Aurore's talk tomorrow

# A \$10,000,000 question for the precision PDF analysis

Snowmass'2021 whitepaper  
 "Proton structure at the precision frontier"

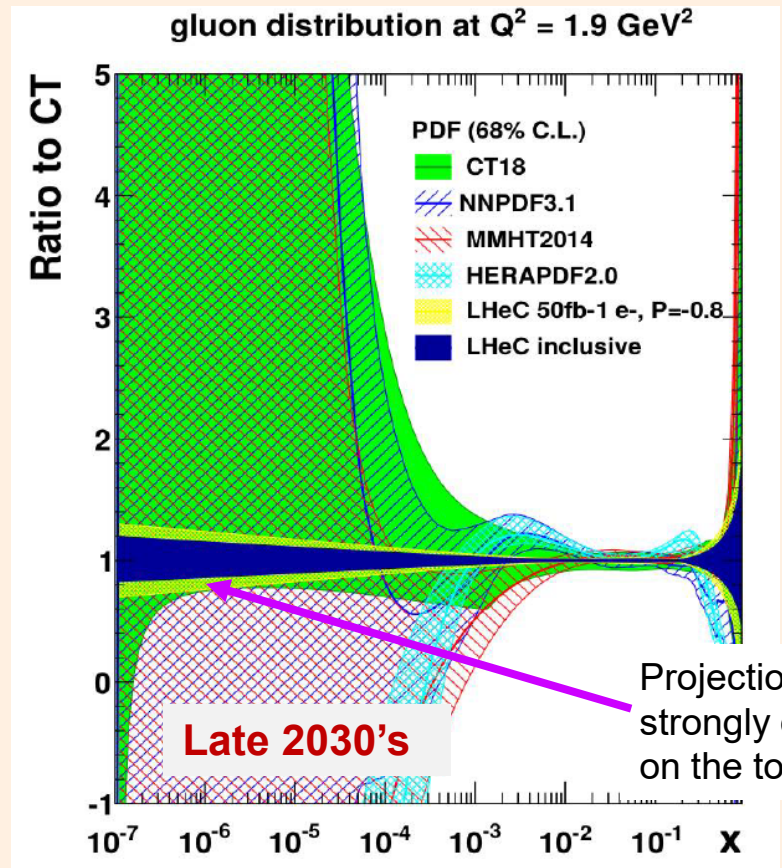
## How do we get from here...



NNLO  $Z^0$  and  $gg \rightarrow H^0$  cross sections at the LHC, and 95% CL PDF uncertainties predicted with recent PDF sets.

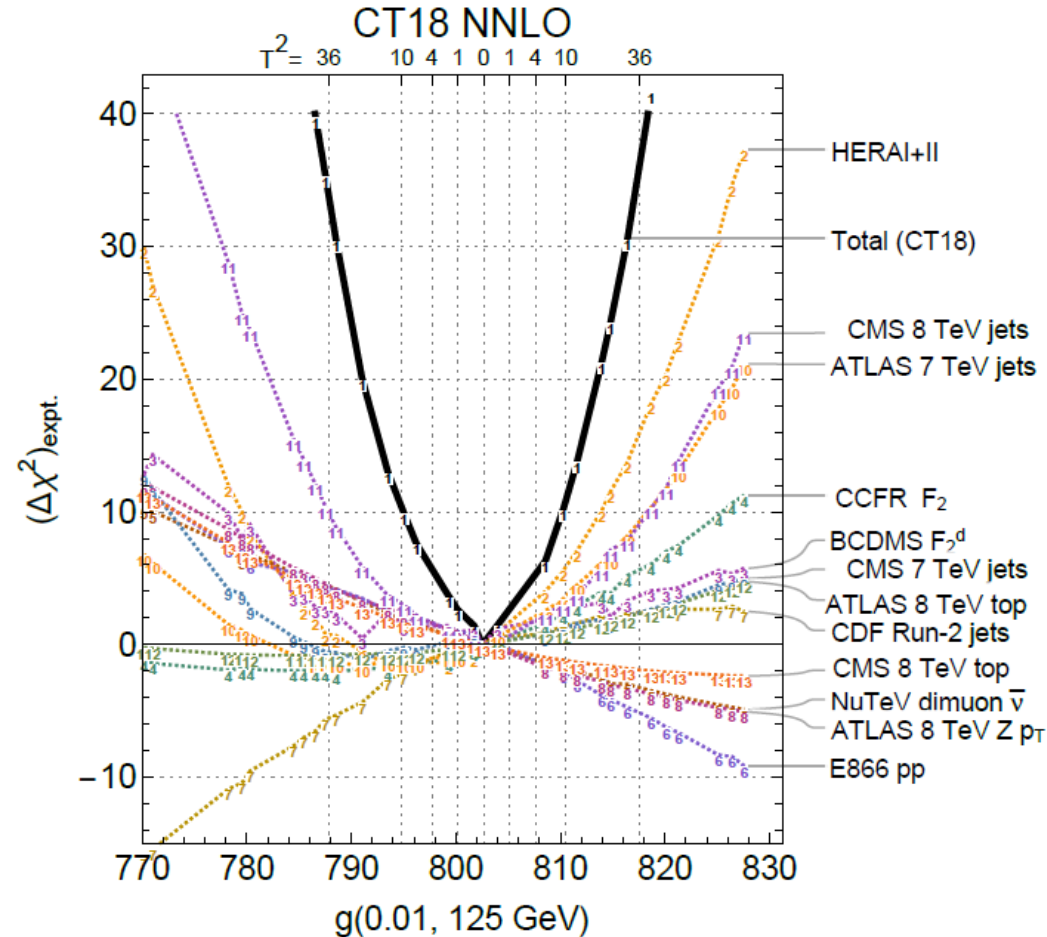
While the fitted data sets are similar in several of these analyses, the observed differences reflect to substantial degree the different methodological choices adopted by the PDF fitting groups.

## ...to here?

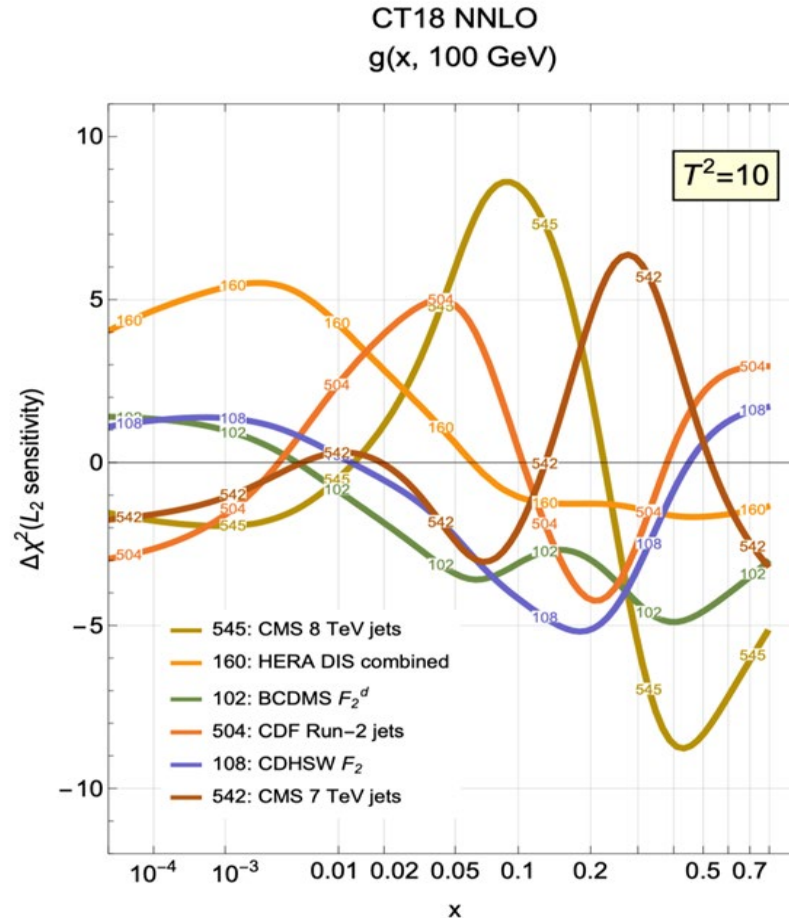


# Strong tensions among experimental measurements reflect non-replicability

Several techniques were developed to inspect and reduce tensions in the global fits

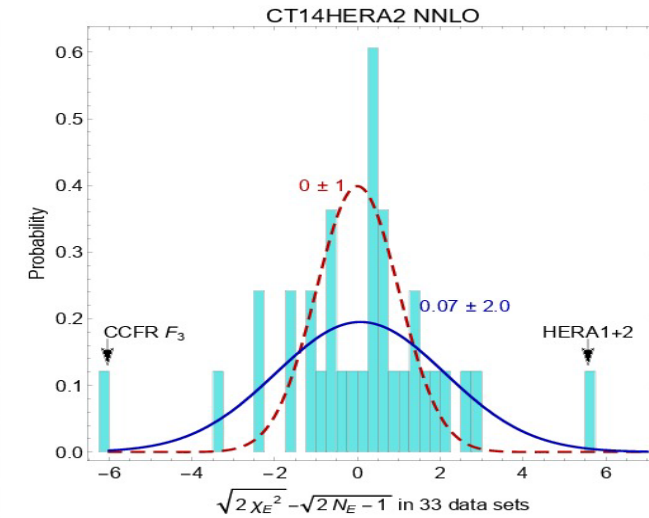


**Lagrange multiplier scans**  
 Stump et al., [hep-ph/0101051](https://arxiv.org/abs/hep-ph/0101051)



**L<sub>2</sub> sensitivities for ATLAS, CJ, CT, MSHT PDFs**

T. J. Hobbs et al., [1904.00222](https://arxiv.org/abs/1904.00222); A. Accardi et al., [2102.01107](https://arxiv.org/abs/2102.01107); X. Jing et al., [2306.03918](https://arxiv.org/abs/2306.03918)



**Effective Gaussian variables**  
 H.-L. Lai et al., [1007.2241](https://arxiv.org/abs/1007.2241)

# Ongoing studies of systematic uncertainties are essential and still insufficient

- from the experiment side

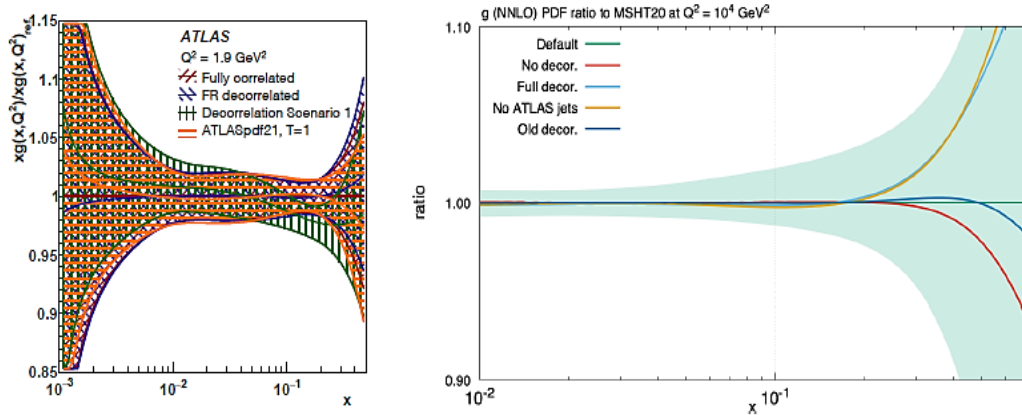


FIG. 9. Difference in the gluon PDF shown in ratio to the ATLASpdf21 (default) gluon (left). This default uses Decorrelation Scenario 2 and this is compared to the use of Full Correlation, Full decorrelation of the flavour response systematic and Decorrelation Scenario 1. The effect of no decorrelation, the default correlation of [9], the decorrelation in [362], and full decorrelation for the MSHT20 gluon (right).

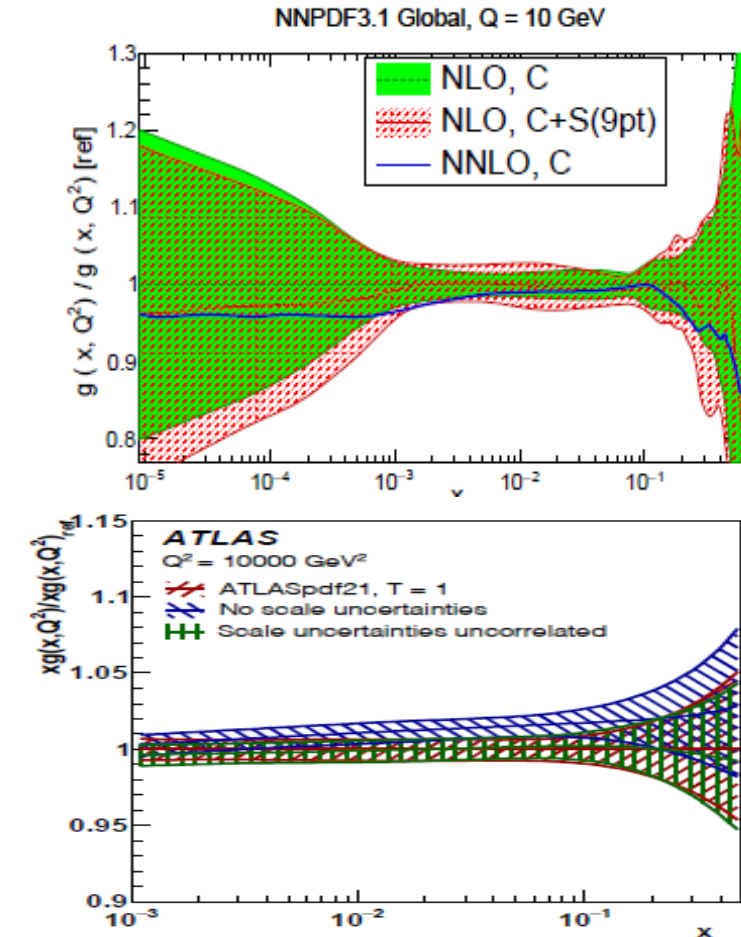
*S. Amoroso et al., 2203.13923, Sec. 5.A*

Strong dependence on the definition of corr. syst. errors raises 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]

- from the theory side



Examples: studies of theory uncertainties in the PDFs by NNPDF3.1 and ATLAS21



# Two common forms of $\chi^2$ in PDF fits

## 1. In terms of nuisance parameters $\lambda_{\alpha,exp}$

$$\chi^2 = \sum_{i=1}^{N_{pt}} \frac{[D_i + \sum_{\alpha} \beta_{i,\alpha}^{exp} \lambda_{\alpha,exp} - T_i]^2}{s_i^2} + \sum_{\alpha} \lambda_{\alpha,exp}^2$$

## 2. In terms of the covariance matrix

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

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

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

↓ algebraic minimization of  $\chi^2$  with respect to  $\lambda_{\alpha,exp}$

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

Experiments publish  $\sigma_{i,\alpha}$  (up to hundreds per data set). To reconstruct  $\beta_{i,\alpha}$ , we need to decide on the normalizations  $X_i$ . Possible choices:

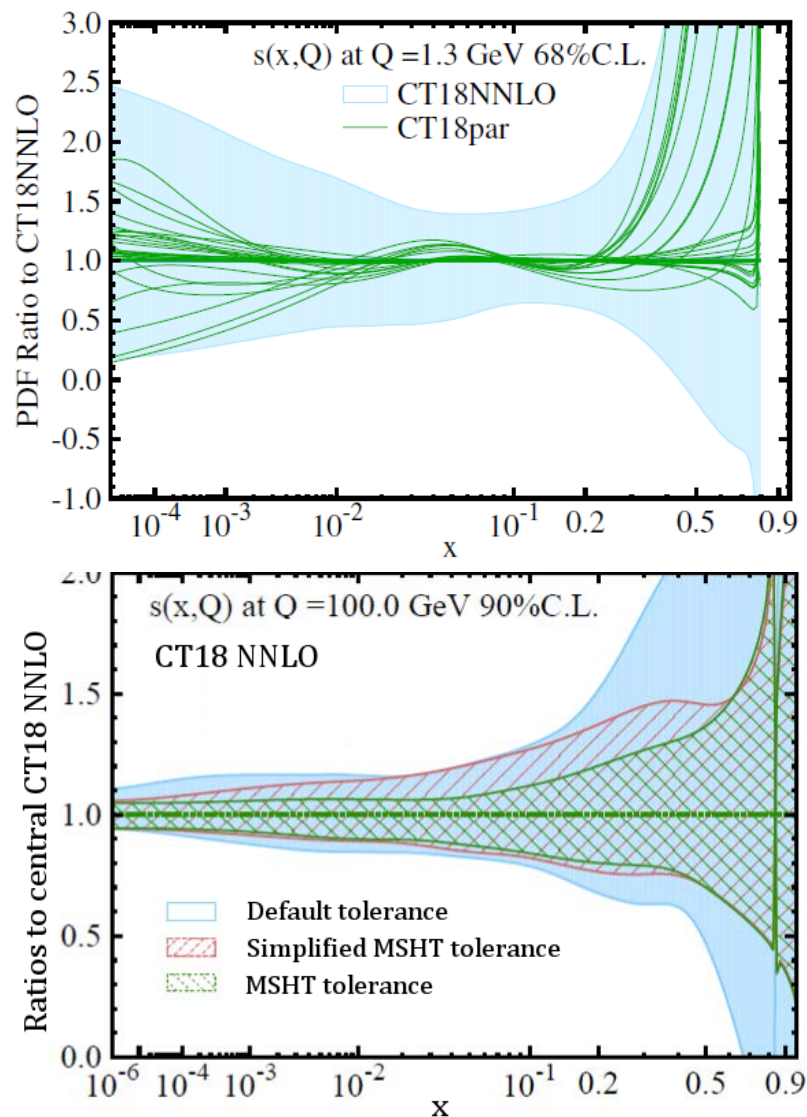
- $X_i = D_i$  : “**experimental scheme**”; can result in a bias
- $X_i = \text{fixed or varied } T_i$  : “ **$t_0, T, \text{ extended } T$  schemes**”; can result in (different) biases

# Goodness-of-fit functions in PDF analyses

Analysis	$\chi^2$ prescription to fit PDFs	$\chi^2$ prescription to compare PDFs	Comments
HERAPDF	HERA	HERA	
CT	Extended $T$ + addl. prior	Extended $T$ , Experimental	
MSHT'20	$T$	$T$	
NNPDF4.0	$t_0$ + addl. prior with fluctuated cross-sampled data	Experimental or $t_0$ with unfluctuated full data	$t_0$ prescription has pre- and post-NNPDF3.0 realizations
...			
Hopscotch'2022	N/A	Experimental or $t_0$ [2022] with unfluctuated data	

Different prescriptions reflect modeling of additive and multiplicative systematic errors in covariance matrices. **Neither prescription is complete because of the bias-variance dilemma. The  $\chi^2$  definition can strongly affect the PDF uncertainty.**

## Sampling of PDF parametrizations in global fits



**Upper figure:** A large part of the CT18 PDF uncertainty accounts for the sampling over 250-350 parametrization forms, possible choices of fitted experiments and fitting parameters, definitions of  $\chi^2$

**Lower figure:** this approach sometimes enlarges the uncertainties compared to the other groups, reflecting the chosen goodness-of-fit (tolerance) criterion more than the strength of experimental constraints

However, more restrictive tolerance criteria elevate the risk of sampling biases.

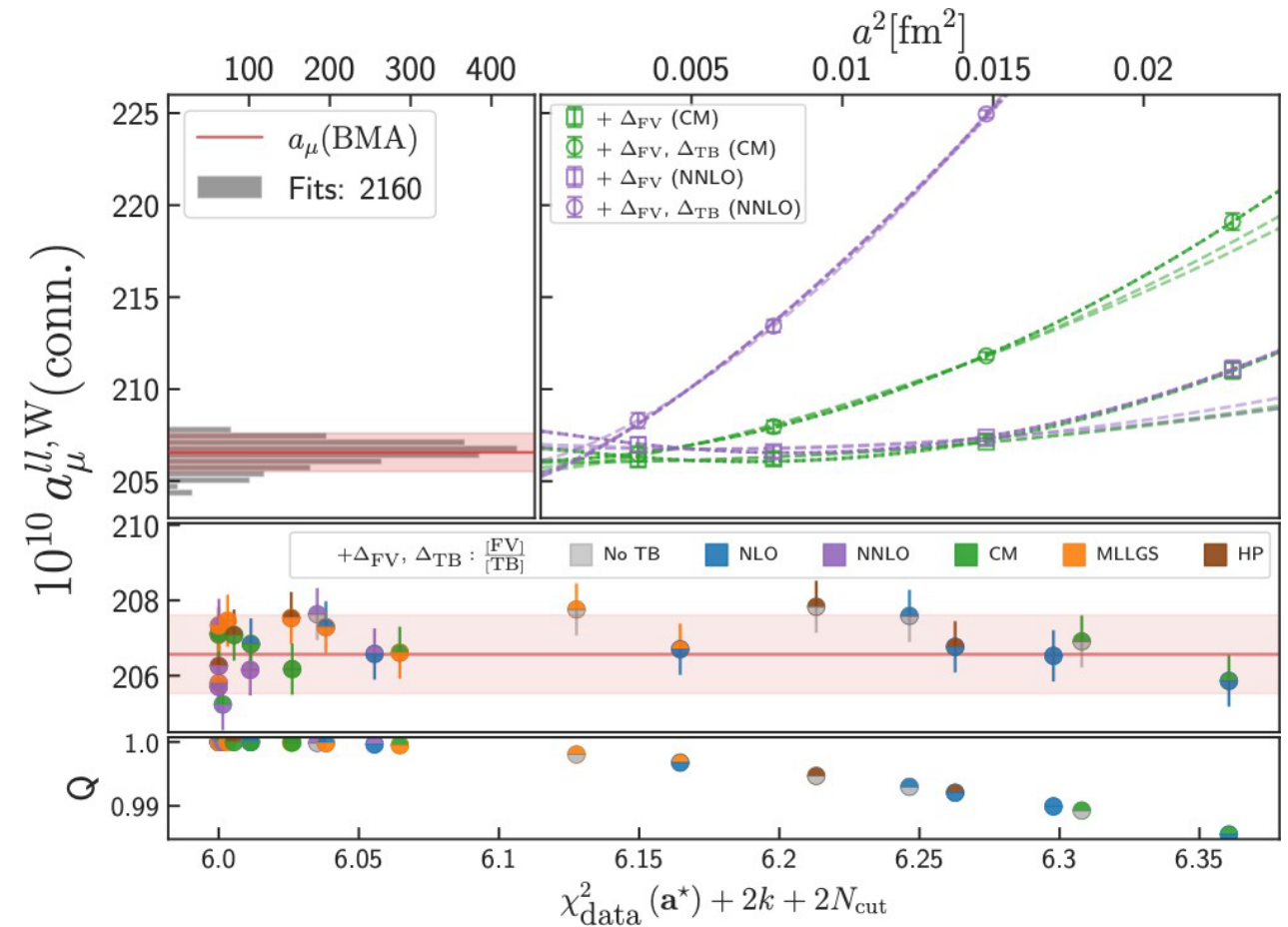
**Easier to examine these issues for specific QCD observables than in abstract**

# Systematic uncertainties in lattice QCD

# Lattice QCD deals with highly challenging syst. uncertainties

Strong interest in these issues, vigorous community

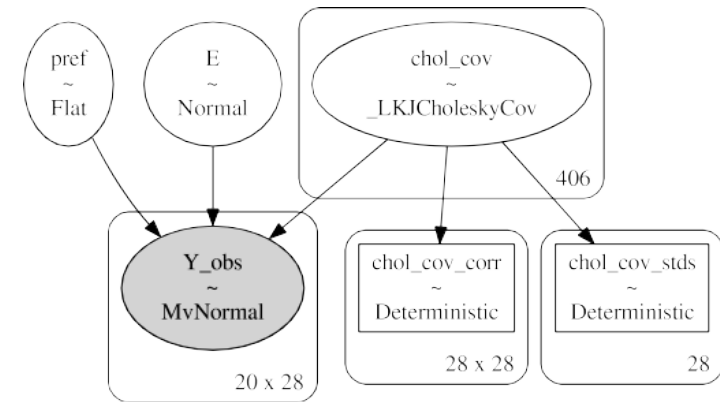
Opportunities for productive collaborations between PDF and lattice experts



- **Example 1:** (g-2) HVP intermediate window (see talk by [S. Lahert, Tue @2:10 PM](#))
- **2160 fit variations** - discretization, finite volume, mass corrections...model average gives a final combined estimate + error bar.

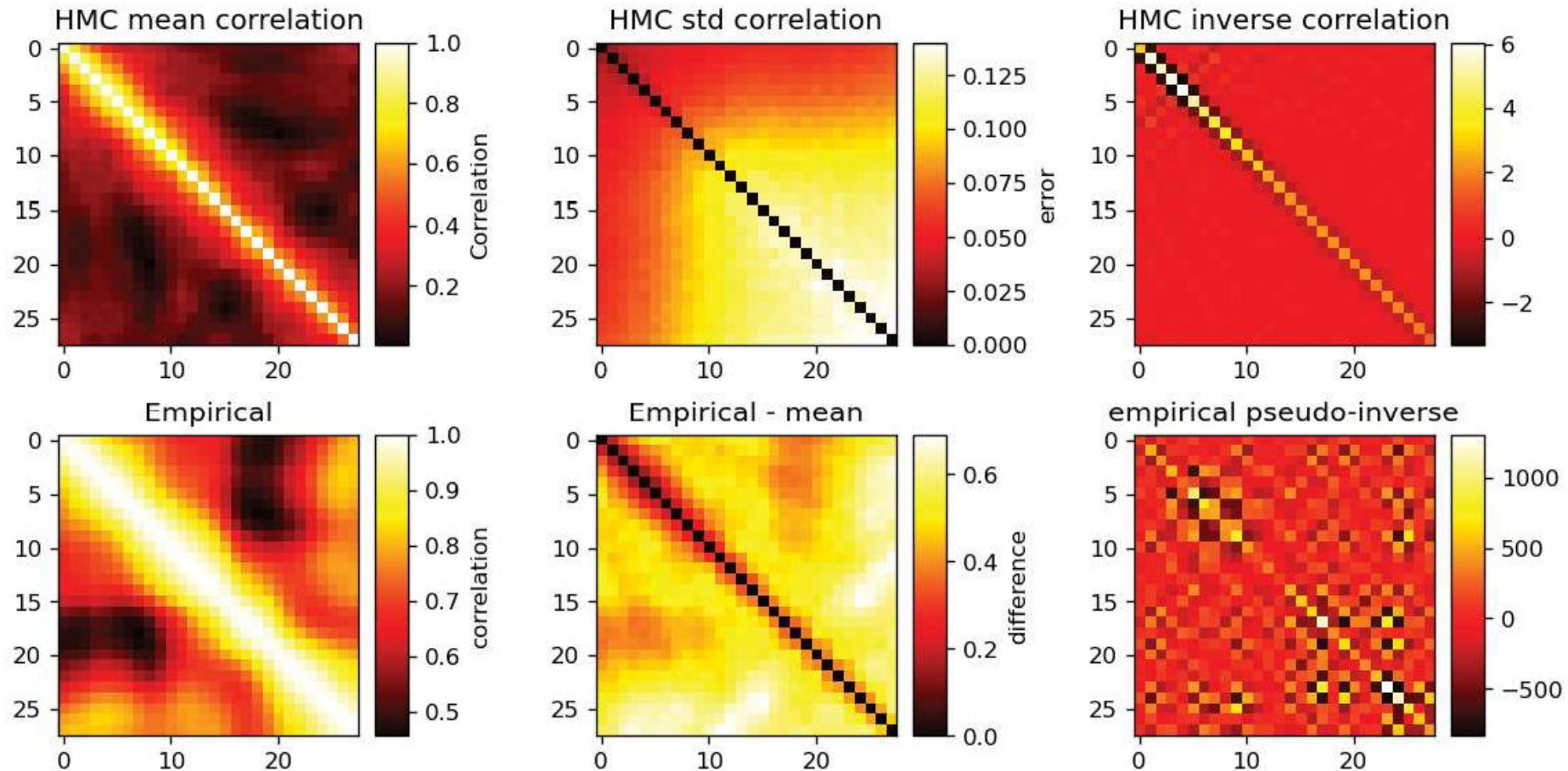
# Inferring a covariance (1)

- > Generalised Least Square is notoriously delicate in LQCD
- > The *sample* covariance is not the *true* covariance
- > We spend a lot of effort on evaluating the uncertainty on the mean but usually neglect the **uncertainty on the covariance!**
- > In practice often leads to get **badly conditioned or non-positive** matrices
- > Some regularisations are well-motivated but it does not propagate uncertainty, nor communicate with the model



# Infering a covariance (2)

Regularizations and truncations of a cov matrix may disagree among themselves



# Will AI/ML help?

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

**What about uncertainty estimation [exploring all good fits]?**

A common resampling procedure used by experimentalists and theorists:

1. Train a neural network model  $T_i$  with  $N_{\text{par}}$  (hyper)parameters on a randomly fluctuated replica of discrete data  $D_i$ . Repeat  $N_{\text{rep}}$  times. In a typical application:  $N_{\text{par}} > 10^2$ ,  $N_{\text{rep}} < 10^4$ .
2. Out of  $N_{\text{rep}}$  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  $N_{\text{rep}}$  replicas does one need?**



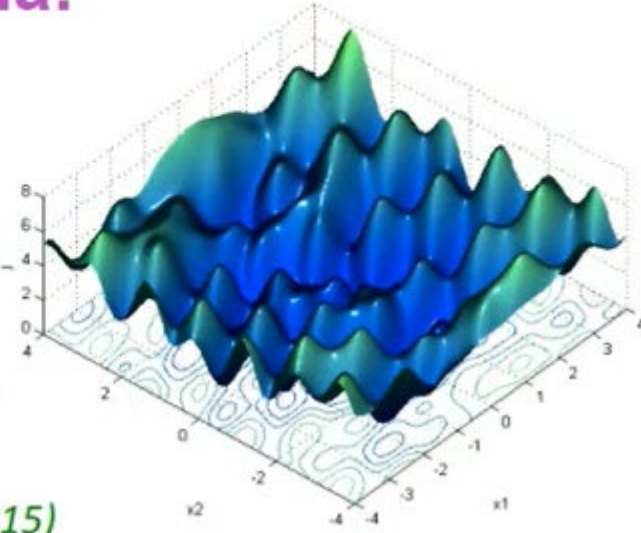
Multidimensional mini-landscape,  
or  
does the minimization of  $\chi^2$  work?

# 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)

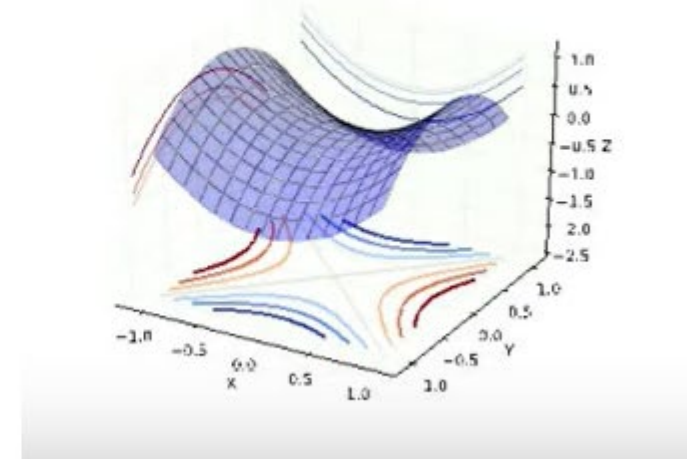
*(Dauphin et al NIPS'2014, Choromanska et al AISTATS'2015)*



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



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

...as well as with representative exploration of uncertainties

Article

# Unrepresentative big surveys significantly overestimated US vaccine uptake

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

Received: 18 June 2021

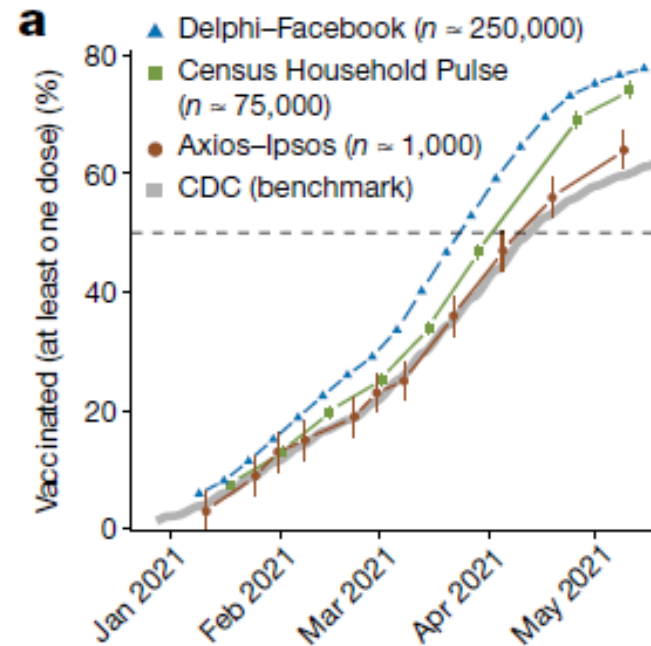
Accepted: 29 October 2021

Published online: 8 December 2021

Check for updates

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

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 Paradox<sup>1</sup>. 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<sup>2,3</sup> (about 250,000 responses per week) and Census Household Pulse<sup>4</sup> (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 panel<sup>5</sup> with about 1,000 responses per week following survey research best practices<sup>6</sup> provided reliable estimates and uncertainty quantification. We decompose observed error using a recent analytic framework<sup>7</sup> 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

# 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



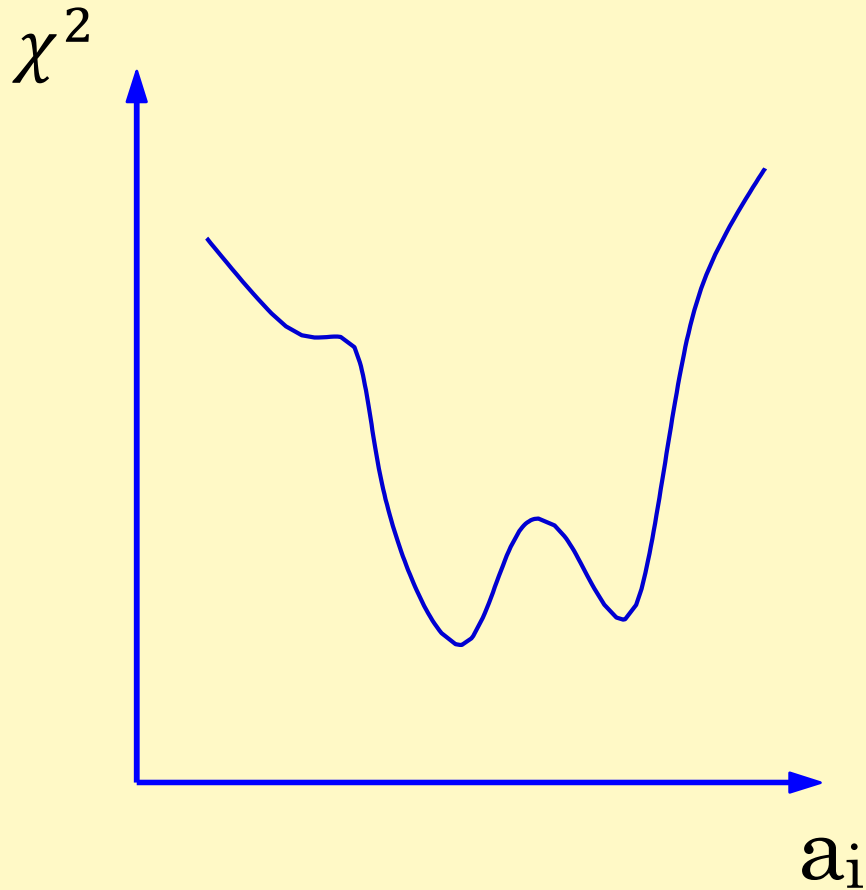
We knew about the PDF mini-landscape (now viewed as a saddle-point manifold) for 20+ years!

Profound implications for uncertainty quantification

Justification of the PDF tolerance due to

- incomplete agreement of experiments
- epistemic uncertainty

# Multi-dimensional PDF error analysis

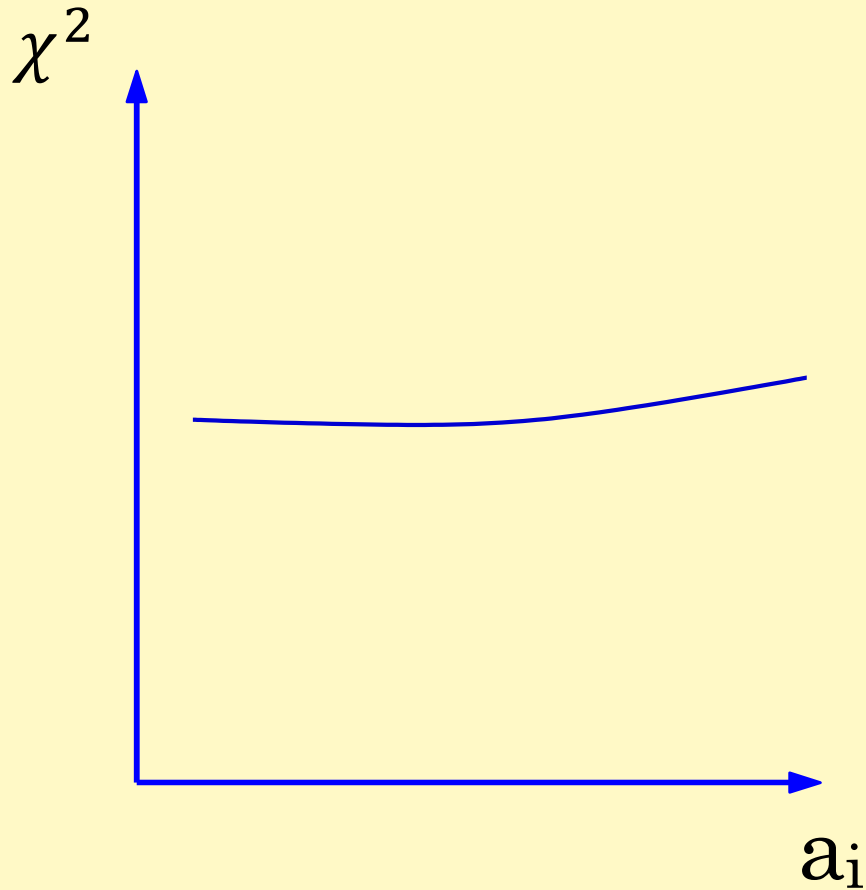


## Pitfalls to avoid

- “Landscape”
  - disagreements between the experiments

<https://online.kitp.ucsb.edu//online/lhc08/nadolsky/>

# Multi-dimensional PDF error analysis



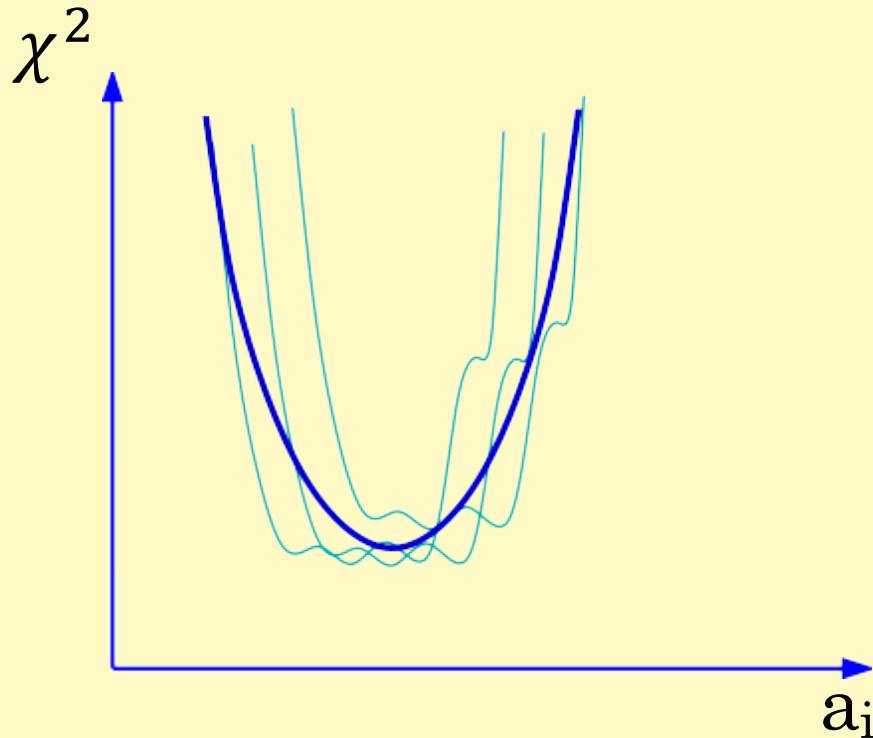
## Pitfalls to avoid

- Flat directions
  - unconstrained combinations of PDF parameters

<https://online.kitp.ucsb.edu//online/lhc08/nadolsky/>



# Multi-dimensional PDF error analysis

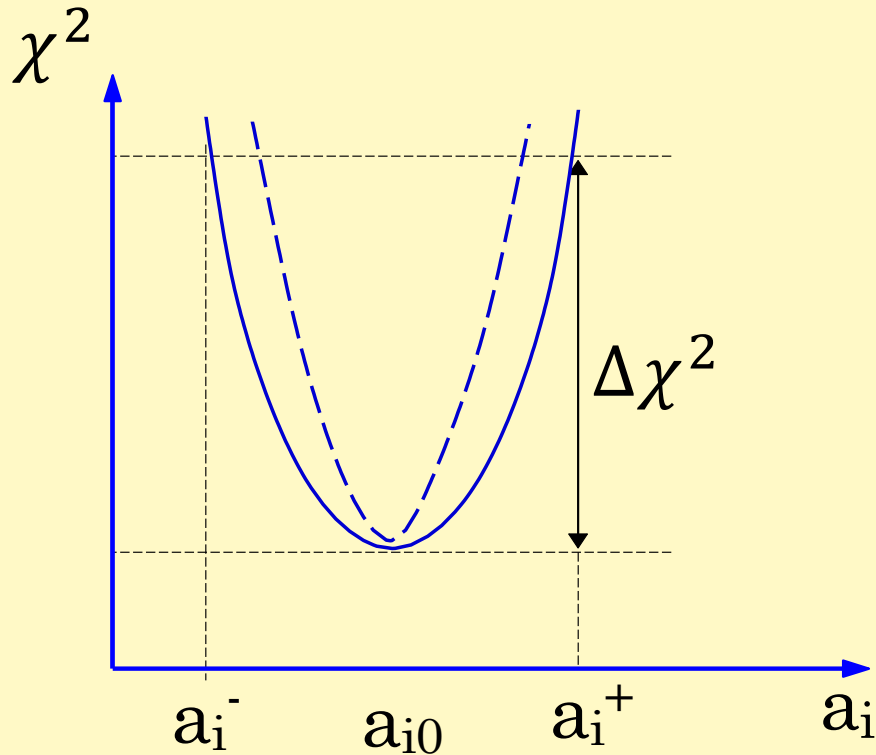


## The actual $\chi^2$ function shows

- a well pronounced global minimum near  $\chi_0^2$
- weak tensions between data sets in the vicinity of  $\chi_0^2$  (mini-landscape)
- some dependence on assumptions about flat directions

The likelihood is approximately described by a quadratic  $\chi^2$  with a revised tolerance condition  $\Delta\chi^2 \leq T^2$

# Multi-dimensional PDF error analysis



## The actual $\chi^2$ function shows

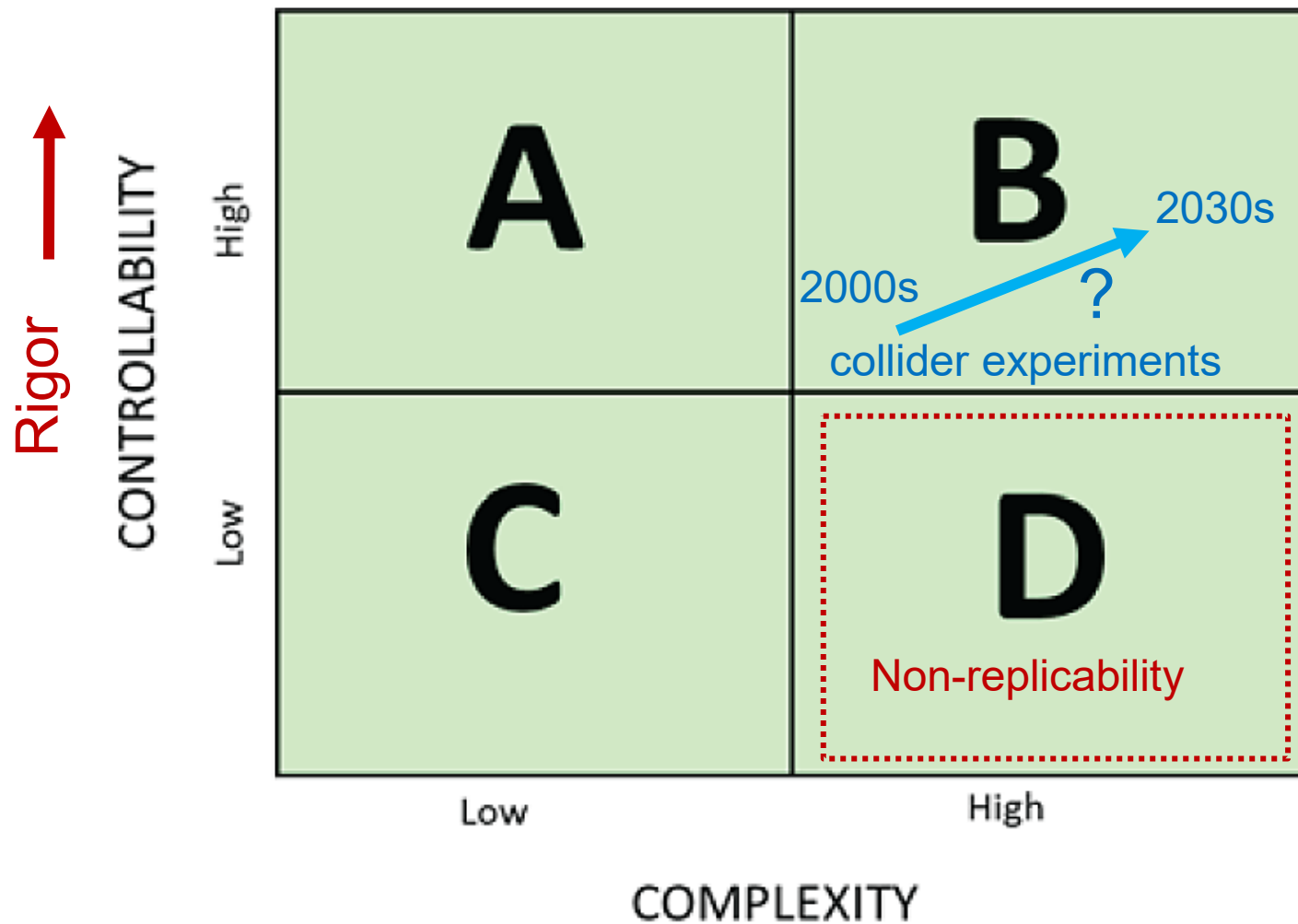
- a well pronounced global minimum near  $\chi_0^2$
- weak tensions between data sets in the vicinity of  $\chi_0^2$  (mini-landscape)
- some dependence on assumptions about flat directions

The likelihood is approximately described by a quadratic  $\chi^2$  with a revised tolerance condition  $\Delta\chi^2 \leq T^2$

# Seeing through the forest



# Our hope



Based on Fig. 5.2 in  
"REPRODUCIBILITY AND  
REPLICABILITY IN SCIENCE"

# Adopting the replicability mindset for the EIC

Complex STEM fields encounter RRR challenges. The EIC is not an exception.

These issues are very important when aleatory and perturbative uncertainties are small.

Early and broad adoption of the replicability mindset brings many advantages and is often cost-saving for research programs

This mindset encourages innovation within a framework that assures scientific rigor and standard practices.

It motivates researchers to have “the skin in the game” of replicable uncertainty quantification.

Much is known about the factors promoting RRR. Collider physics can learn from mistakes and successes in other fields such as AI in medicine.

Uncertainty quantification is often streamlined, and RRR improved, by reducing dimensionality of the problem.

# Possible PDF4EIC activities

- common physics goals
  - ⇒ learn about the 3D hadron structure!
- shared resources
  - LHAPDF-like repository for interpolations of polarized/TMD/GPD PDFs? For  $\chi^2$  values for error PDFs? Other outputs of the fits?
  - Coordinated software development for global fits?
- agreed-upon practices
  - presentation of data and theory predictions? RIVET for the EIC?
  - common definitions of PDF uncertainties?
  - a common standard for PDF validation tests?
- benchmarking studies
  - explore experimental constraints on various types of PDFs and from various available and future processes at (N)(N)LO using the  $L_2$  sensitivity and other techniques



# PDF wish list for systematic uncertainties

## A proposal

Fundamental issues in propagating systematic uncertainties. Some possible remedies:

1. More complete representations for experimental likelihoods that do not need reverse engineering
2. Agreed-upon nomenclature for leading syst. sources
3. Is reducing dimensionality of published correlation matrices advisable? Is there a standard for it? E.g., fewer nuisance parameters; collect less relevant/certain nuisance parameters into one uncorrelated error; etc.
4. Mathematical consistency of covariance/correlation matrices (see Z. Kassabov et al.)
5. How do different implementations of syst. errors affect pulls on PDFs?  $L_2$  sensitivities to nuisance parameters
6. ...

# Backup



# Recommendations for improving replicability of studies

All researchers should include a clear, specific, and complete description of how a reported result was reached, ... including

- a clear description of all methods, instruments, materials, procedures, measurements, and other variables involved in the study;
- a clear description of the analysis of data and decisions for inclusion/exclusion of some data;
- for results that depend on statistical inference, a description of the analytic decisions and when these decisions were made and whether the study is exploratory or confirmatory;
- a discussion of the expected constraints on generality, such as which methodological features the authors think could be varied without affecting the result and which must remain constant;
- reporting of precision or statistical power; and
- a discussion of the uncertainty of the measurements, results, and inferences.

Researchers who use statistical inference analyses should be trained to use them properly.

Funding agencies and organizations should consider investing in R & D of open-source, usable tools and infrastructure that support reproducibility for a broad range of studies across different domains in a seamless fashion.

Journals should consider ways to ensure computational reproducibility for studies to the extent it is ethically and legally possible.

From “*REPRODUCIBILITY AND REPLICABILITY IN SCIENCE*”, <https://doi.org/10.17226/25303>