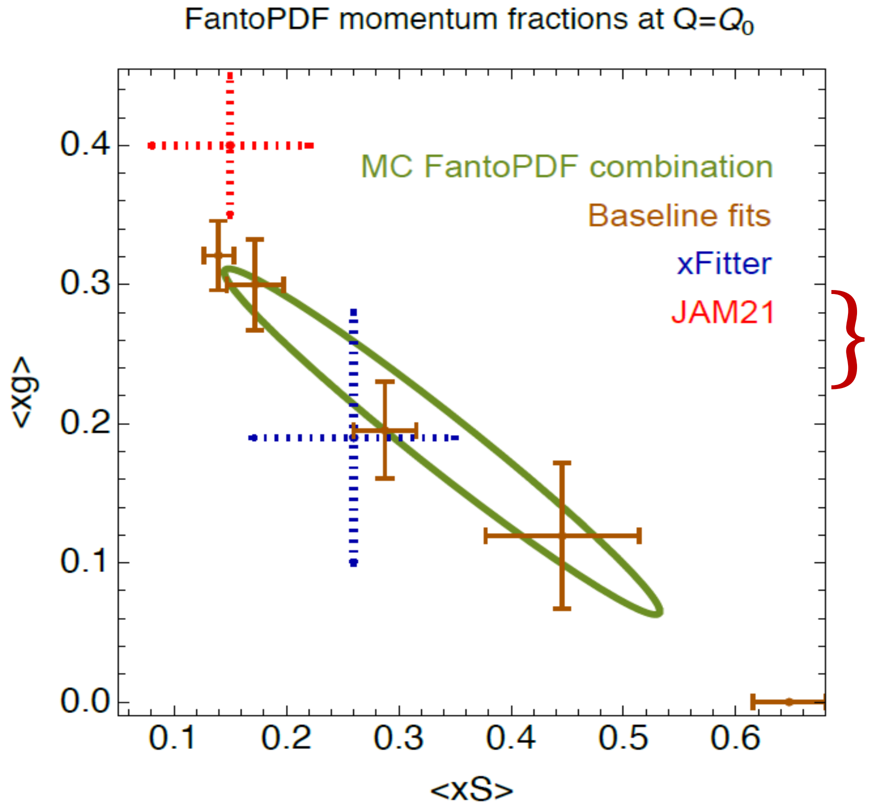
Epistemic uncertainties of parton distributions

Pavel Nadolsky
Southern Methodist University

With A. Courtoy, T. J. Hobbs, M. Guzzi, J. Huston, L. Kotz, F. Olness, M. Ponce Chavez, K. Xie, M. Yan, C.-P. Yuan,

and CTEQ-TEA (Tung Et Al.)
Global QCD analysis group



SMU CTEQ CONAHCYT CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS I·AN Inter-American Network of Networks QCD

3rd PDFLattice workshop by the the global-fitting and lattice-QCD communities

November 18-20, 2024 at Jefferson Lab

- Co-organized by Jefferson Lab and CTEQ collaboration
- *This workshop will be dedicated to uncertainty quantification of nonperturbative correlation functions in phenomenology and lattice QCD, and the best ways to integrate lattice inputs as the calculations mature. The workshop will focus on collinear PDFs without excluding uncertainty quantification studies of other hadronic functions, such as GPDs and TMDs, especially if there are lessons that can be learned and applied to PDFs.*

PDFs are the simplest nonperturbative functions for hadron structure

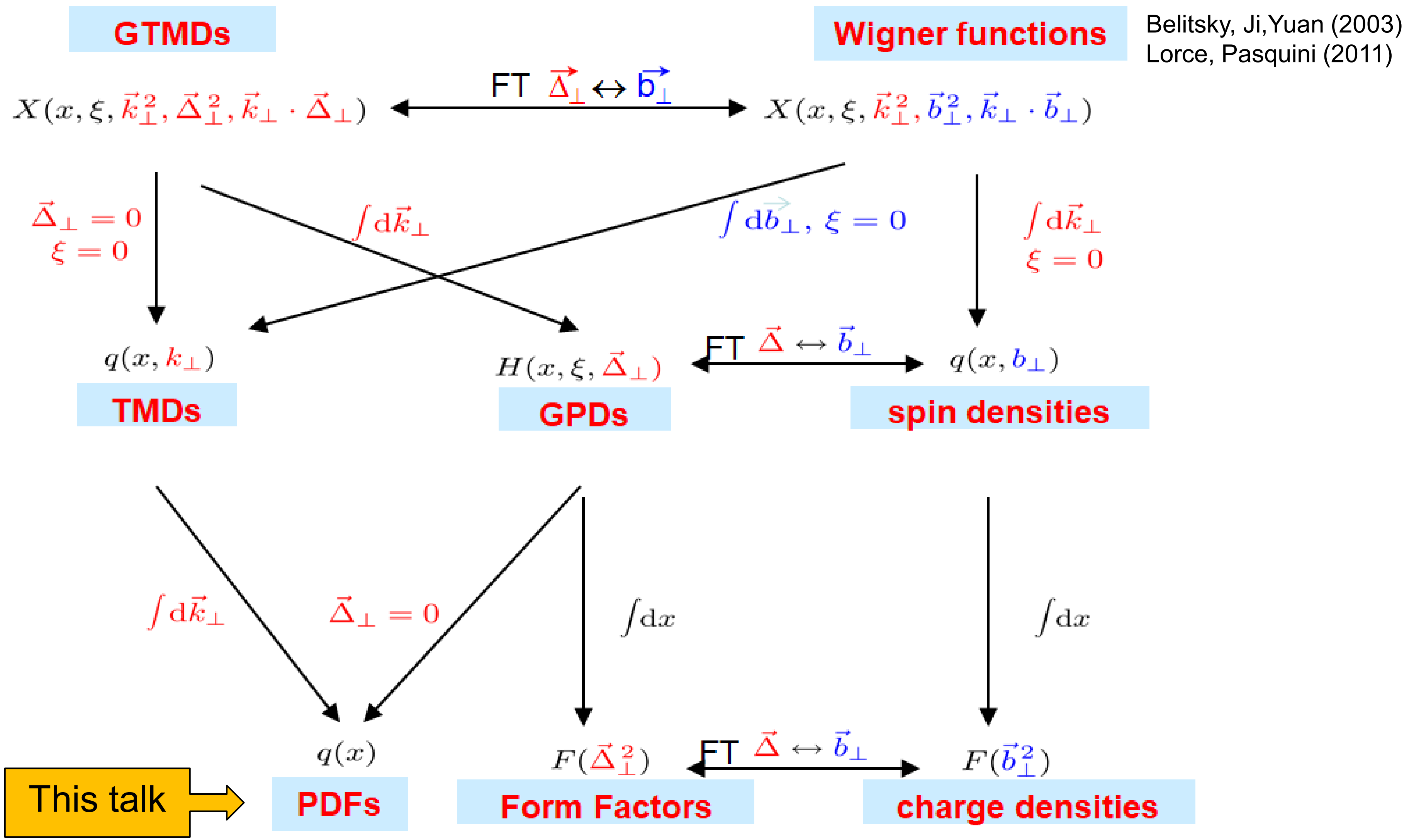
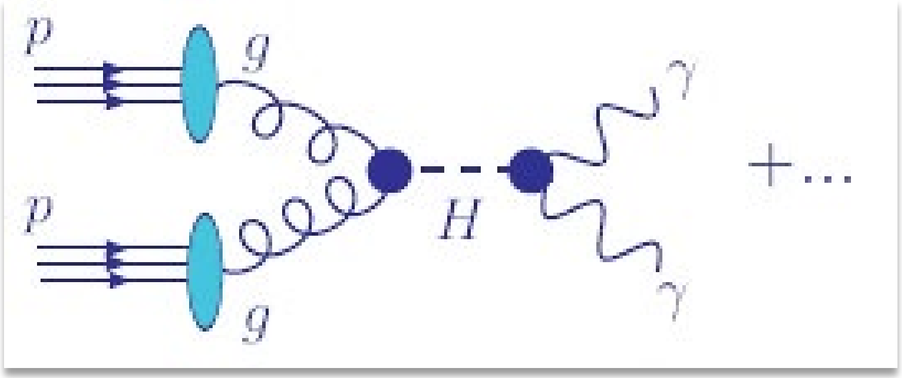


Figure: B. Pasquini, C. Lorcé

Parton distributions describe long-distance dynamics in high-energy collisions

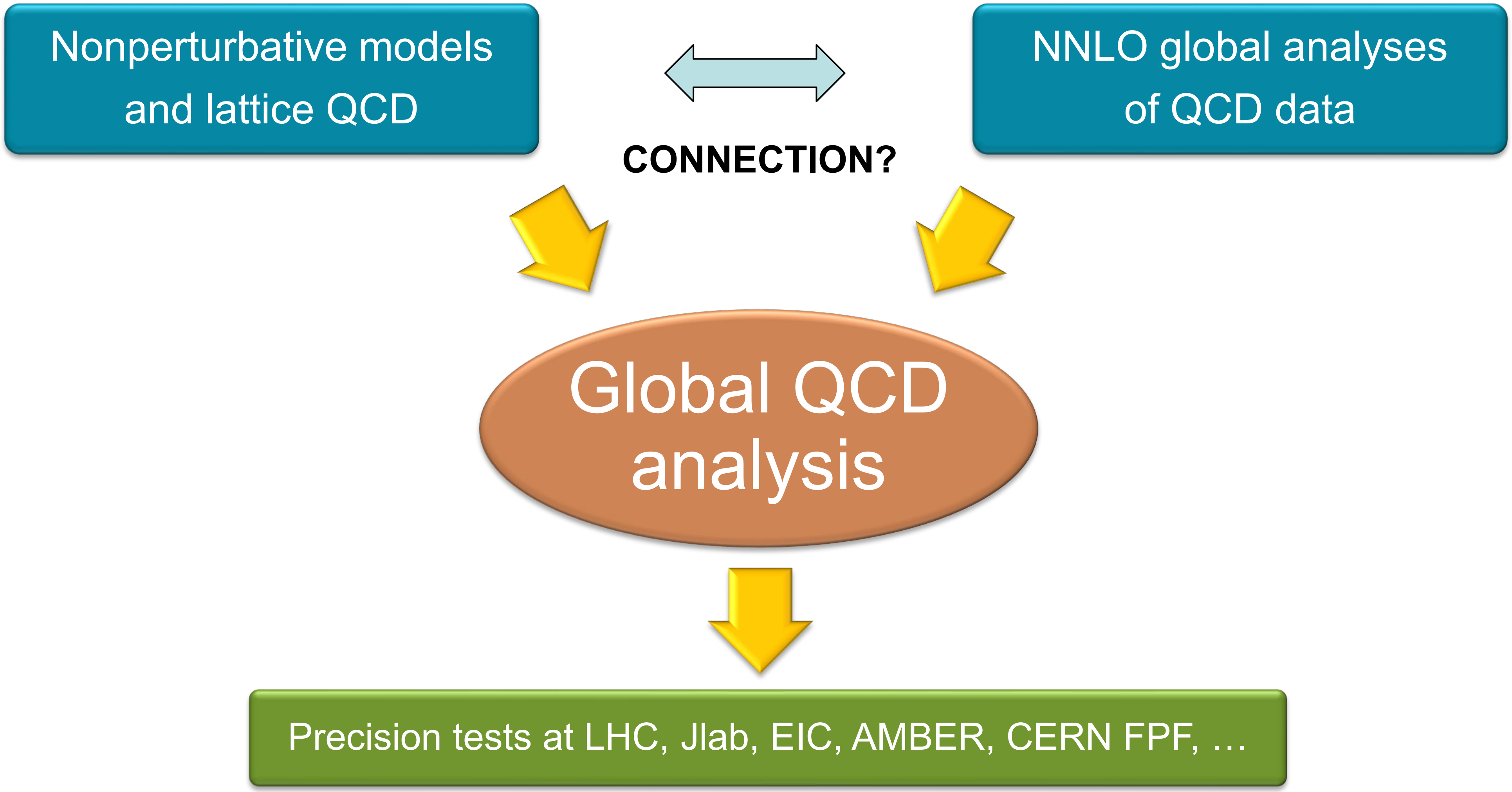


$$\sigma_{pp\to H\to\gamma\gamma X}(Q) = \sum_{a,b=g,q,\bar{q}} \int_0^1 d\xi_a \int_0^1 d\xi_b \hat{\sigma}_{ab\to H\to\gamma\gamma}\left(\frac{x_a}{\xi_a}, \frac{x_b}{\xi_b}, \frac{Q}{\mu_R}, \frac{Q}{\mu_F}; \alpha_s(\mu_R)\right)$$
$$\times f_a(\xi_a, \mu_F) f_b(\xi_b, \mu_F) + O\left(\frac{\Lambda_{QCD}^2}{Q^2}\right)$$

$\hat{\sigma}$ is the hard cross section; computed order-by-order in $\alpha_s(\mu_R)$

$\boldsymbol{f_a(x, \mu_F)}$ **is the distribution for parton $\boldsymbol{a}$ with momentum fraction $\boldsymbol{x}$, at scale $\boldsymbol{\mu_F}$**

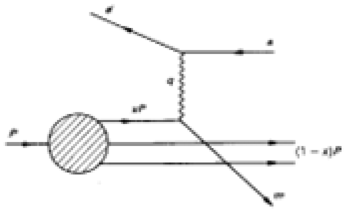
New insights about unpolarized parton distribution functions

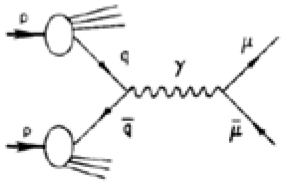


PDFs in nonperturbative QCD

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

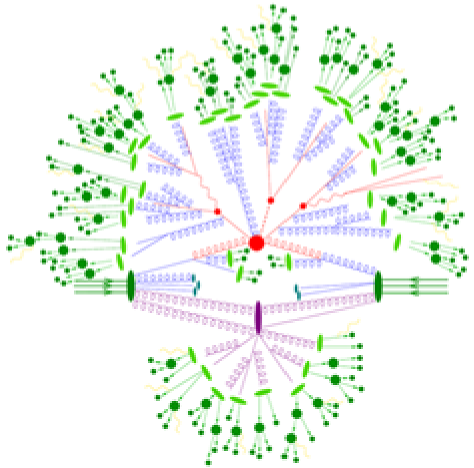
⇨ we can learn about nonperturbative dynamics by comparing predictions to data for the simplest scattering processes (DIS and DY)





Phenomenological PDFs

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

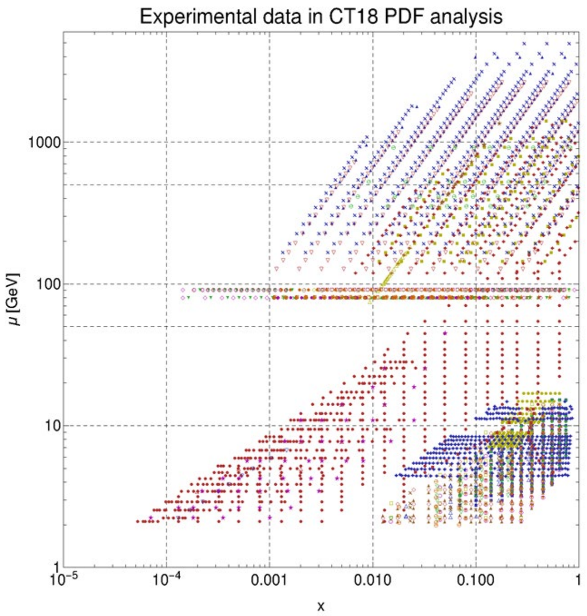


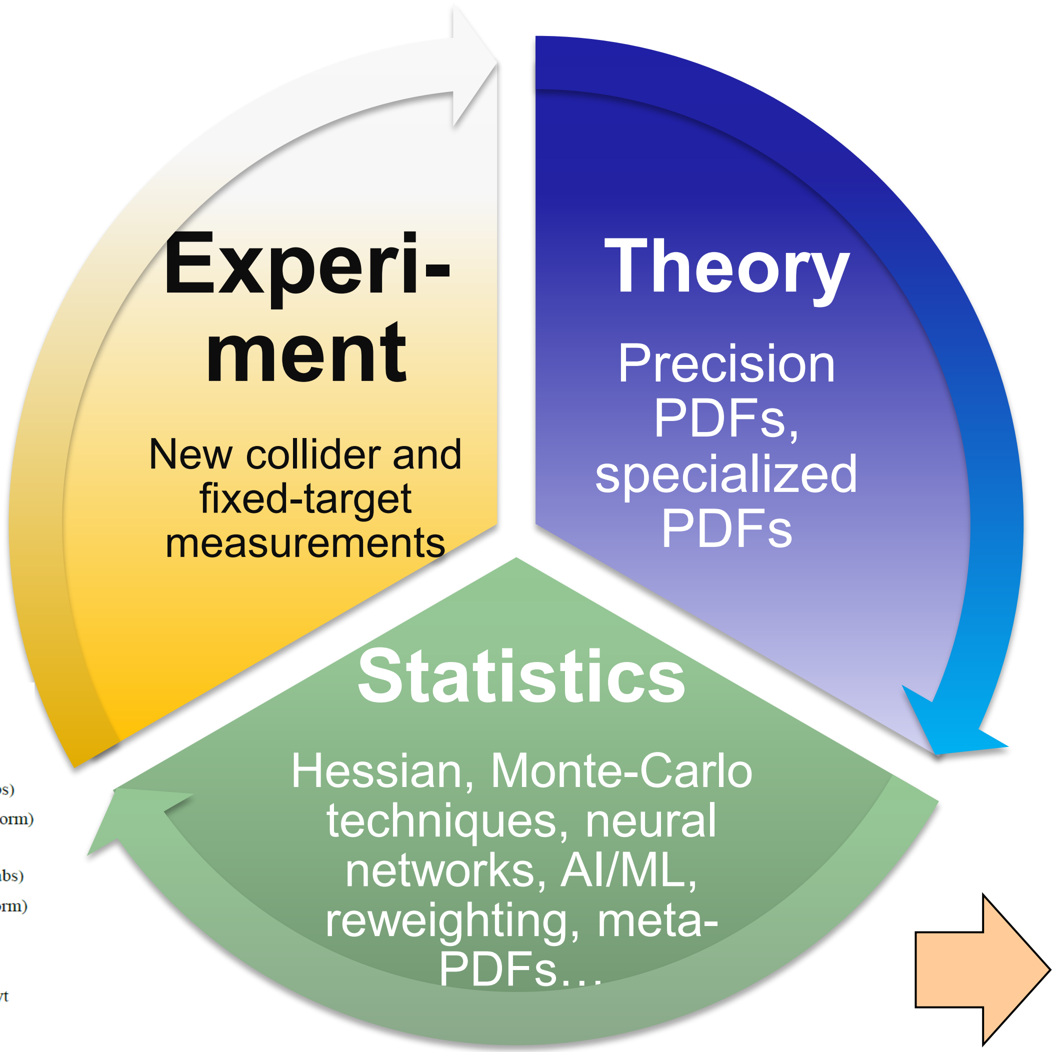
⇨ pheno PDFs are determined from analyzing many processes with complex scattering dynamics

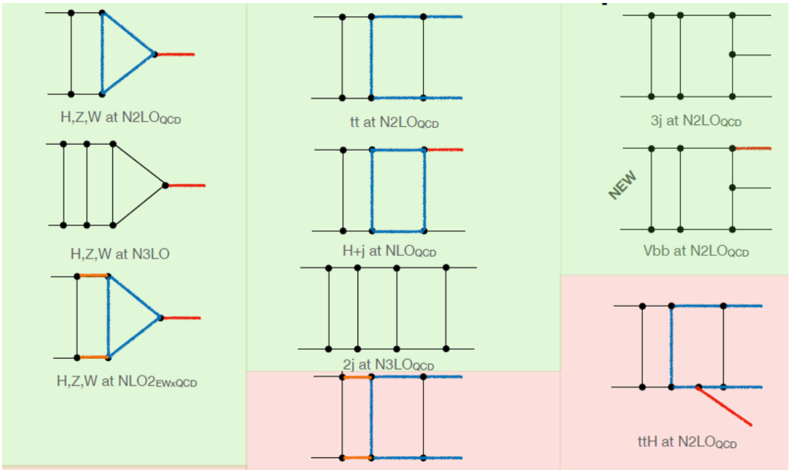
How to relate the x dependence of the perturbative and nonperturbative pictures?

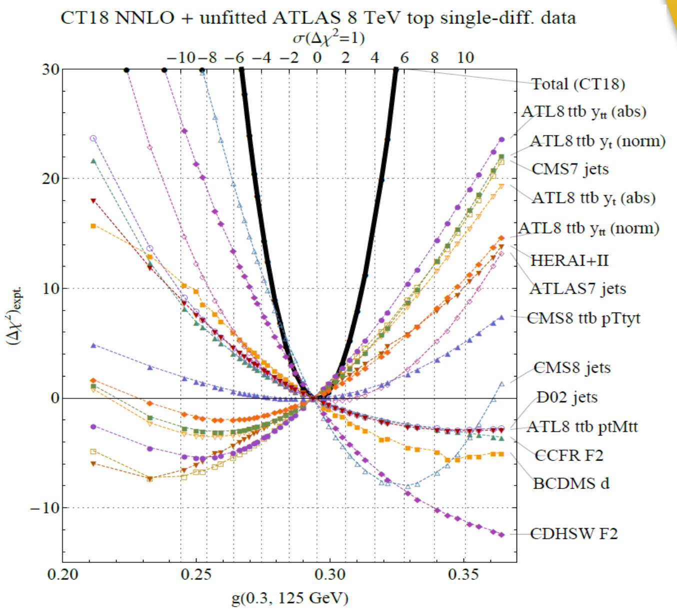
Does the evidence from primordial dynamics survive PQCD radiation?

Global fits of proton scattering data at (N)NNLO accuracy

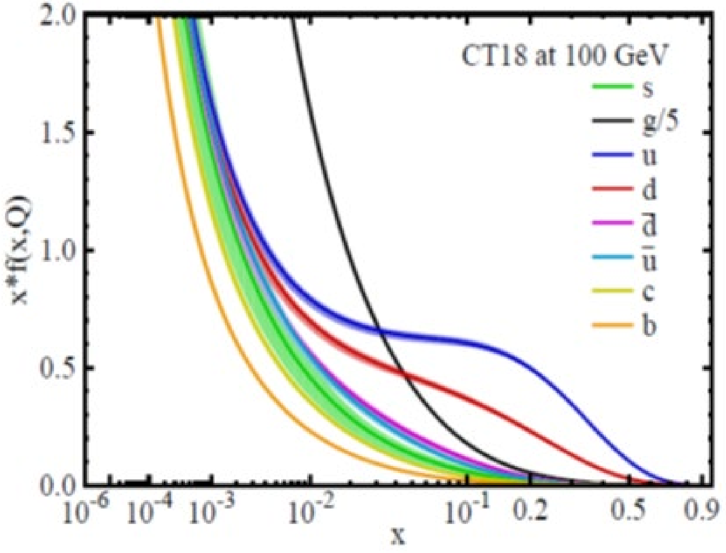








Parton distribution functions with uncertainties



Global fits of proton scattering data at (N)NNLO accuracy

A profound **inverse problem** with many parameters and a wide range of implications

Multiloop QCD and EW computations

Exploration of most complex experimental data sets

Accurate and fast high-performance computing

A testing bed for multidimensional uncertainty quantification, ML/AI, …

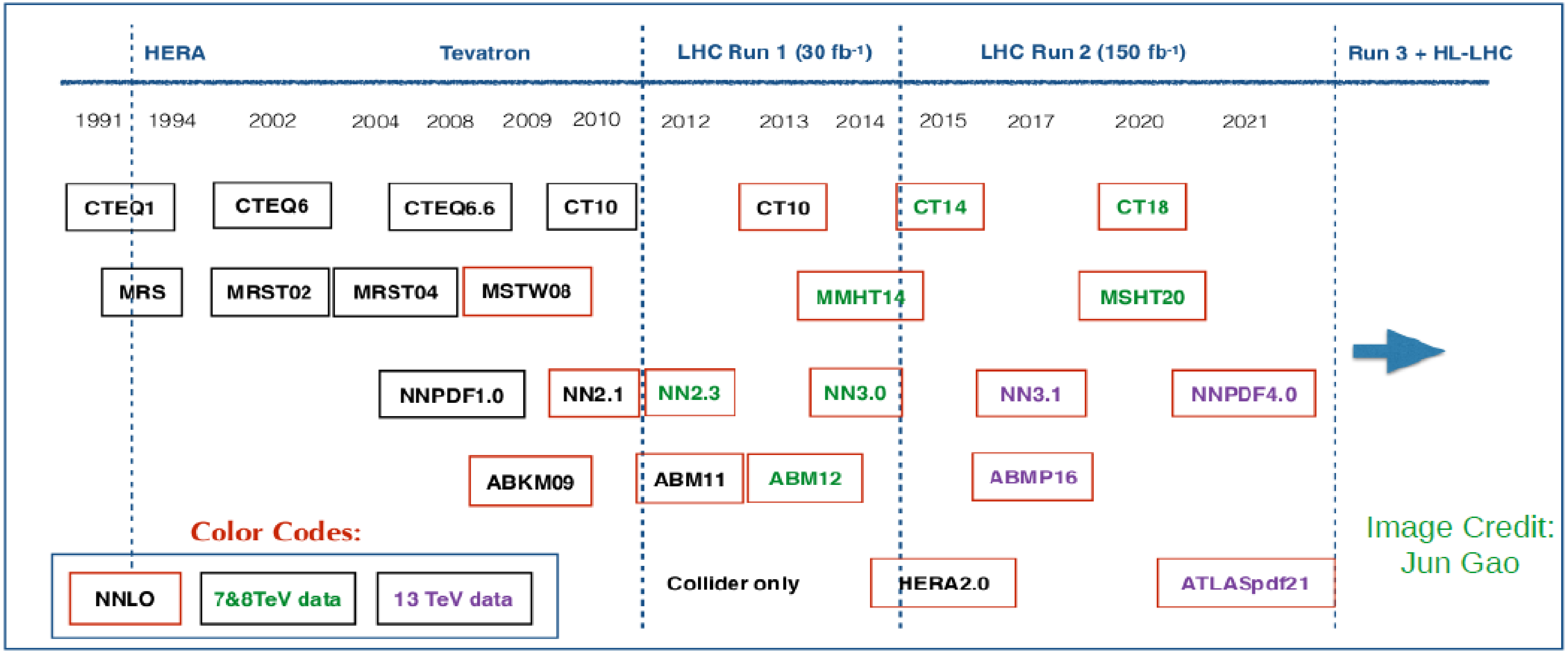
Snowmass'21 whitepaper: Proton structrure at the precision frontier

S. Amoroso et al., Acta Physica Polonica B 53 (2022) 12, A1

A summary of recent trends in the global analysis of proton PDFs

1. Status of modern NNLO PDFs and their applications
2. Future experiments to constrain PDFs
3. Theory of PDF analysis at N2LO and N3LO
4. New methodological advancements
   - Experimental systematic uncertainties in PDF fits
   - Theoretical uncertainties in PDF fits
   - Machine learning/AI connections
5. Delivery of PDFs; PDF ensemble correlations in critical applications
6. PDFs and QCD coupling strength on the lattice
7. Nuclear, meson, transverse-momentum dependent PDFs
8. Public PDF fitting codes
9. Fast (N)NLO interfaces
10. PDF4LHC21 recommendation and PDF4LHC21 PDFs for the LHC analyses

Phenomenological PDF analyses for a nucleon



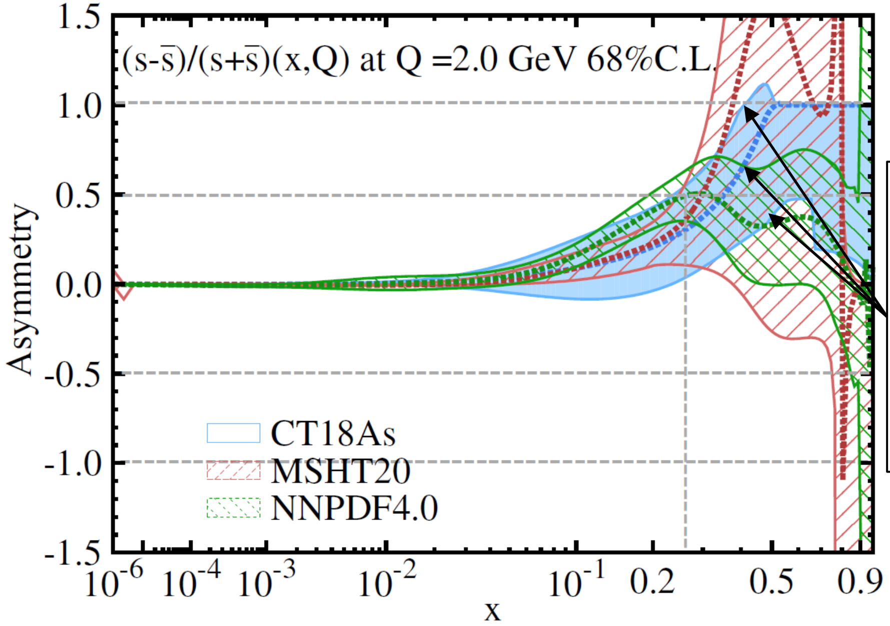
Pursued by several groups – ABM, ATLAS, **CTEQ-TEA (CT)**, CTEQ-Jlab, MSHT, NNPDF, JAM, …

Precision state-of-the art: NNLO QCD + NLO EW; partial N3LO results (NNPDF and MSHT groups)

Data from fixed-target experiments and colliders (HERA, Tevatron, LHC, …) and **increasingly lattice QCD**

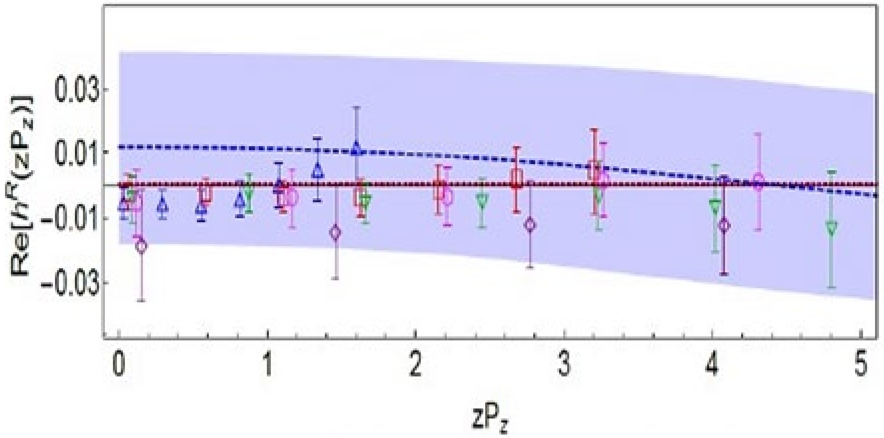
CT18As NNLO: Strangeness asymmetry with a lattice QCD constraint

T.-J. Hou et al., arXiv: 2204.07944



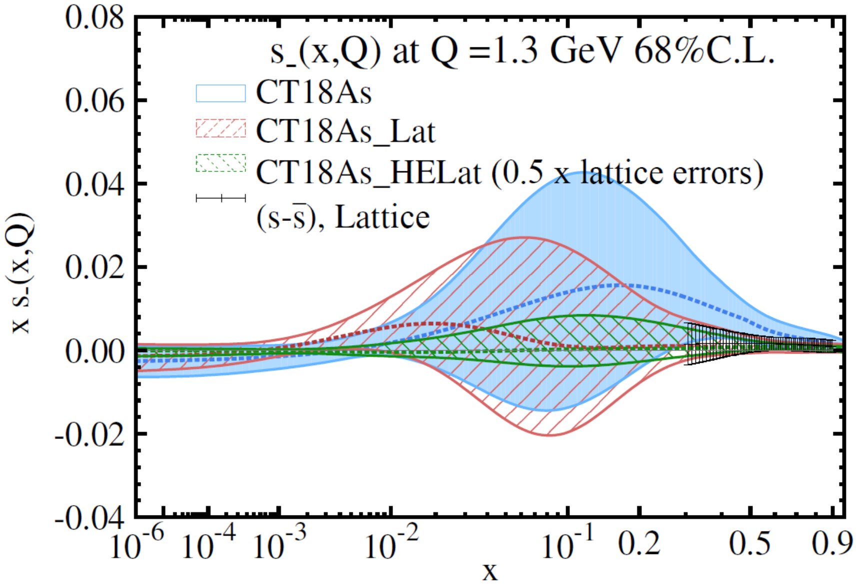
differences reflect the pulls of LHC and other experiments

Asymmetry nominally reaches $\approx 50\%$ at $x \approx 0.25$ in three global fits. **Is there a dynamical mechanism to produce it at such x?**

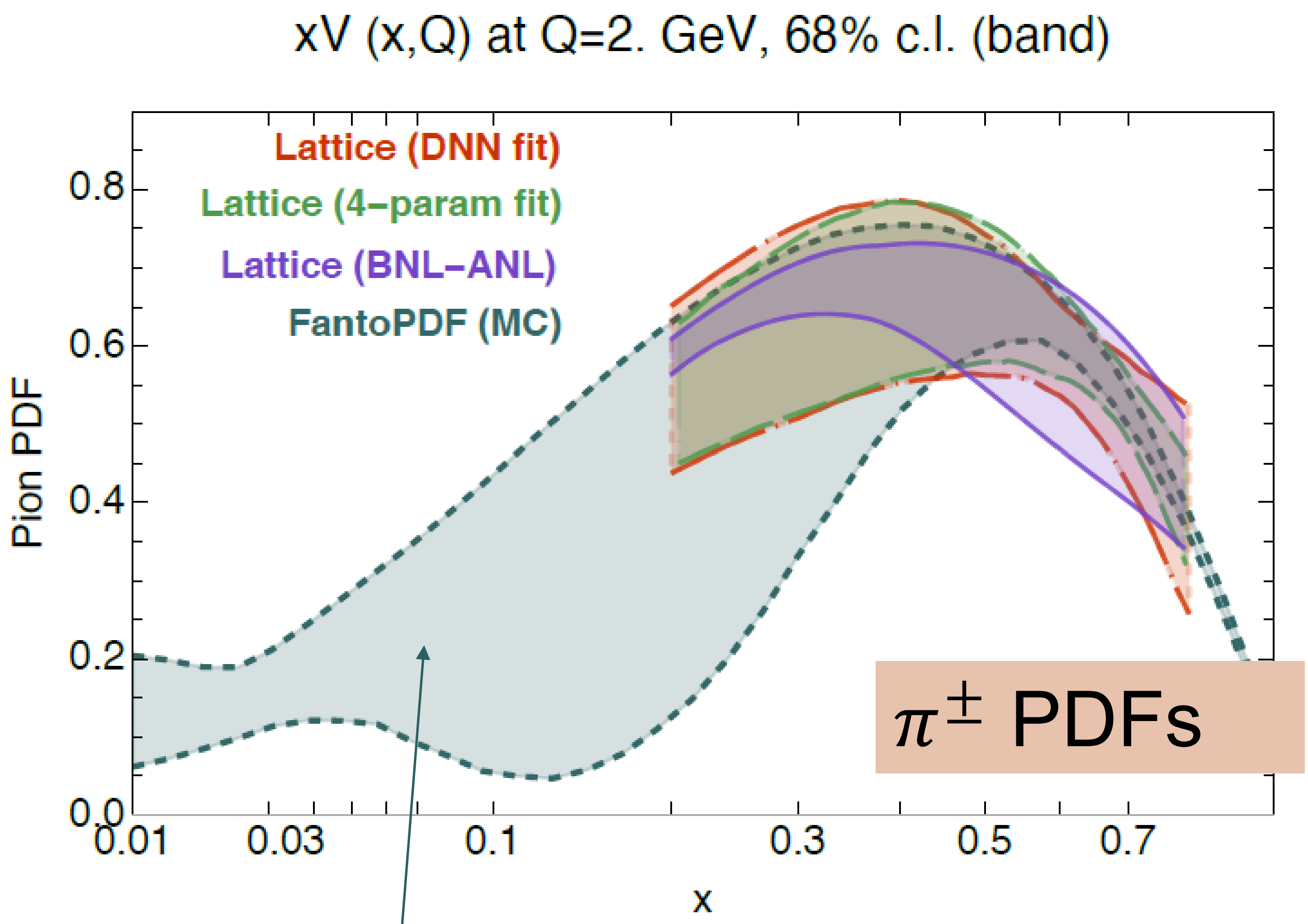


Include lattice data on s_- obtained by the MSULat/quasi-PDF method (2005.12015, Zhang, Lin, Yoon)

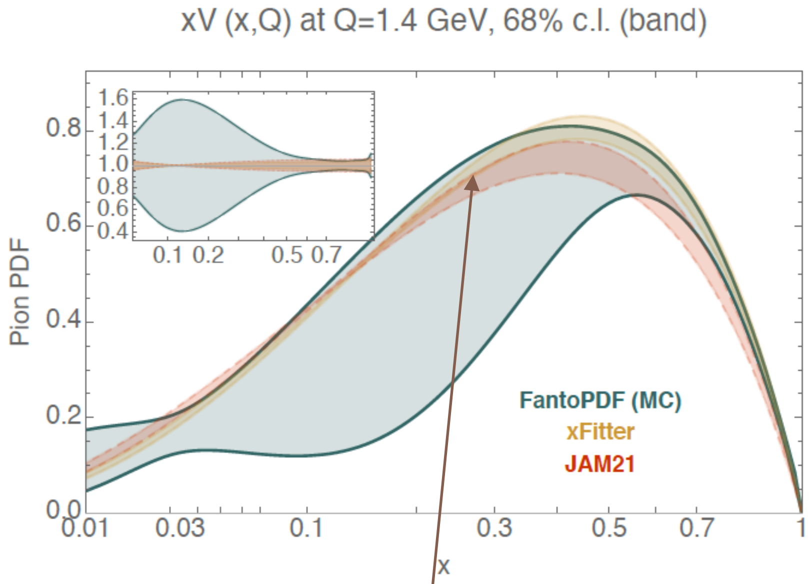
The lattice QCD prediction disfavors a large $s_-(x,Q)$ at $x > 0.3$ ⇒ reduction in $s_-(x,Q)/s_+(x,Q)$ in CT18As_Lat fit



Lattice QCD already predicts some features of PDFs from first principles

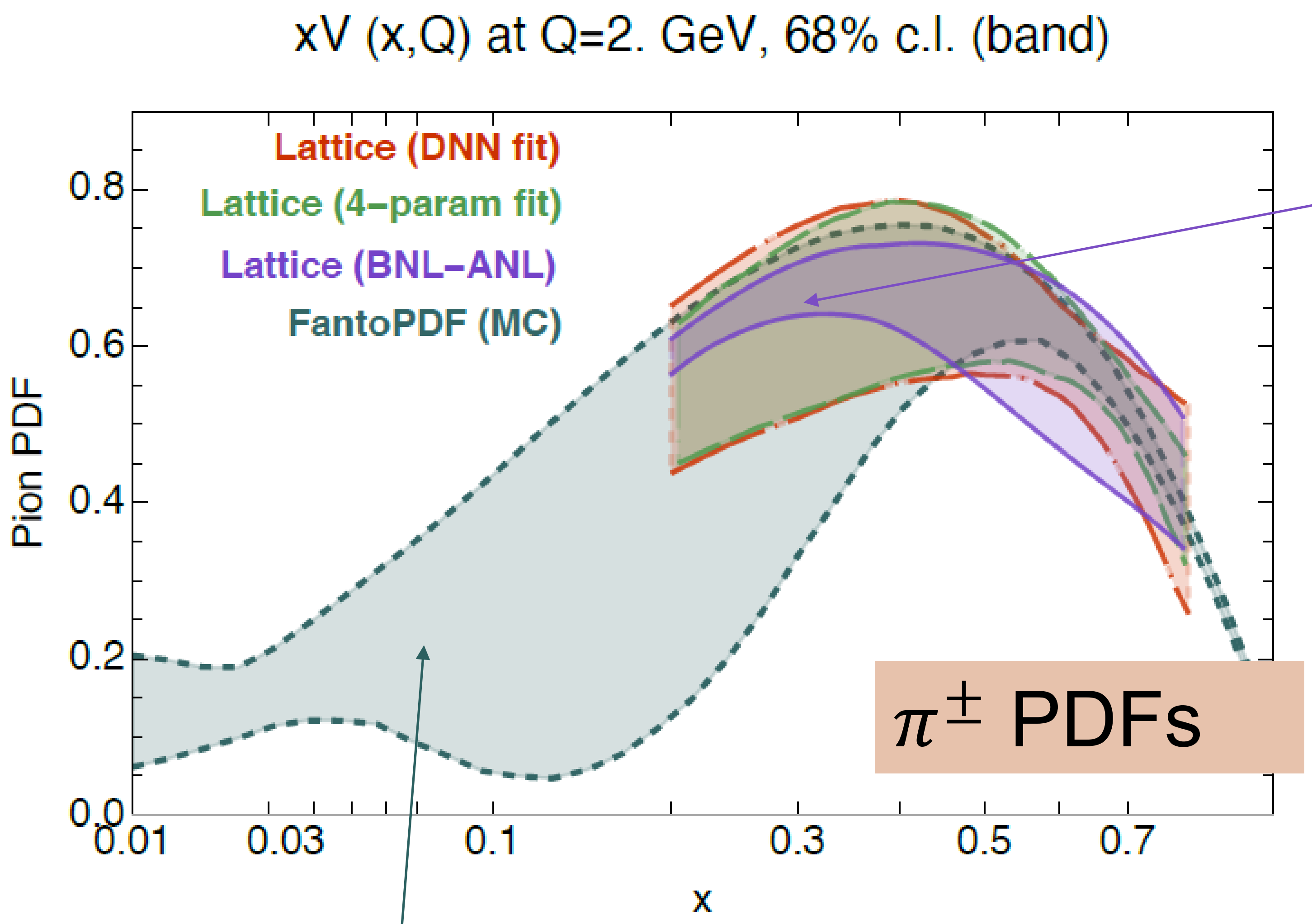


Phenomenological analysis, including the parametrization dependence
L. Kotz, A. Courtoy, M. Chavez, P. N., F. Olness, arXiv:2311.08447

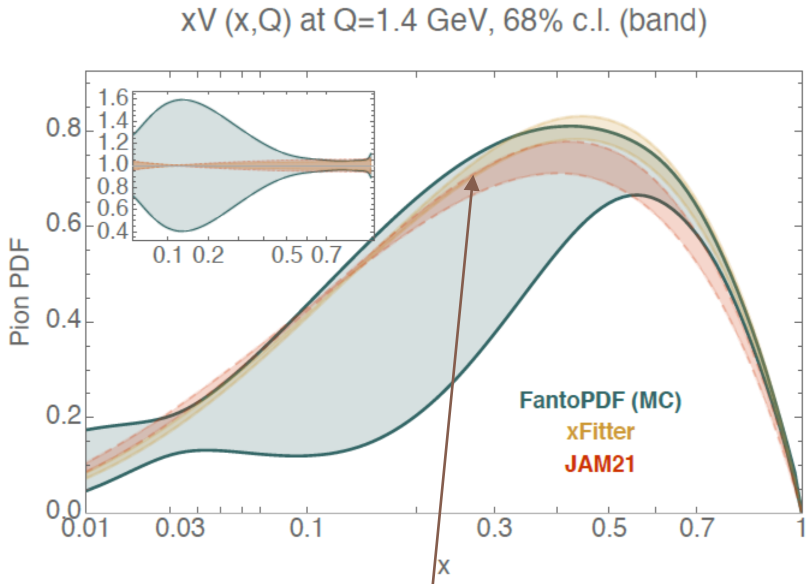


without parametrization dependence

Lattice QCD already predicts some features of PDFs from first principles



are the lattice uncertainties fully estimated?



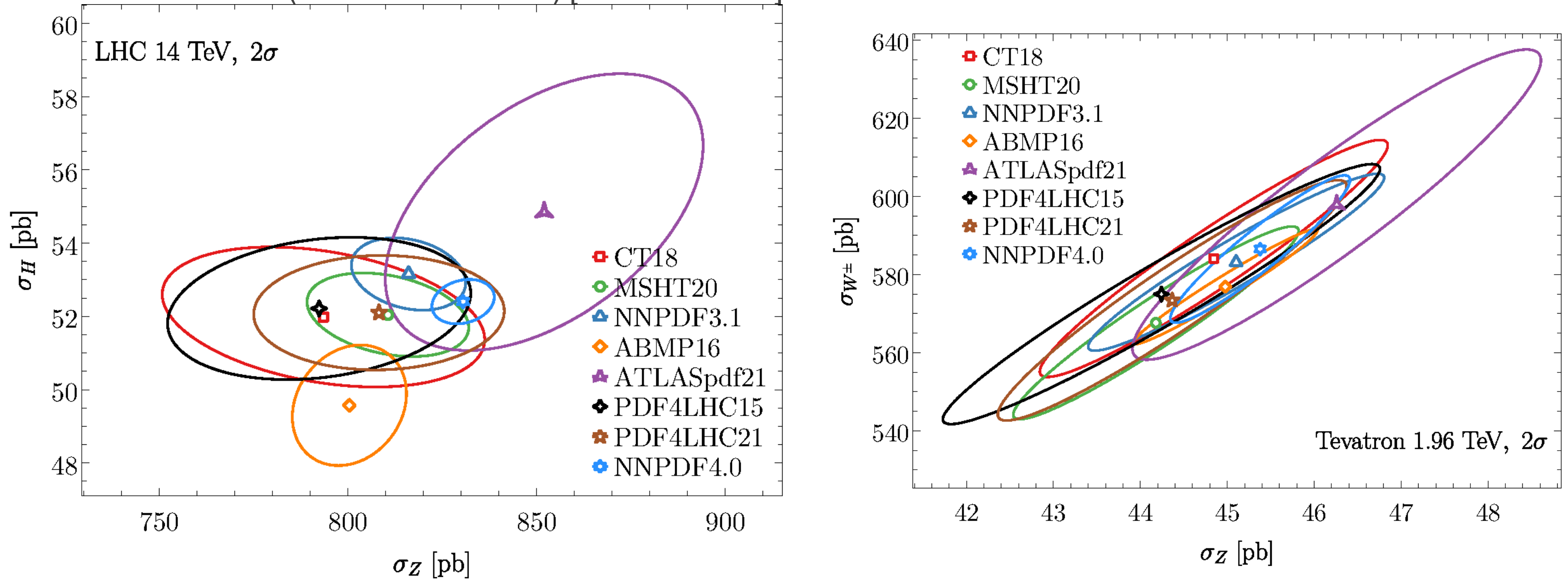
Phenomenological analysis, including the parametrization dependence
L. Kotz, A. Courtoy, M. Chavez, P. N., F. Olness, arXiv:2311.08447

without parametrization dependence

The tolerance puzzle

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

Precision PDFs (Snowmass 21 WP) [2203.13923v2]



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.

The role of epistemic uncertainties in understanding the PDF tolerance

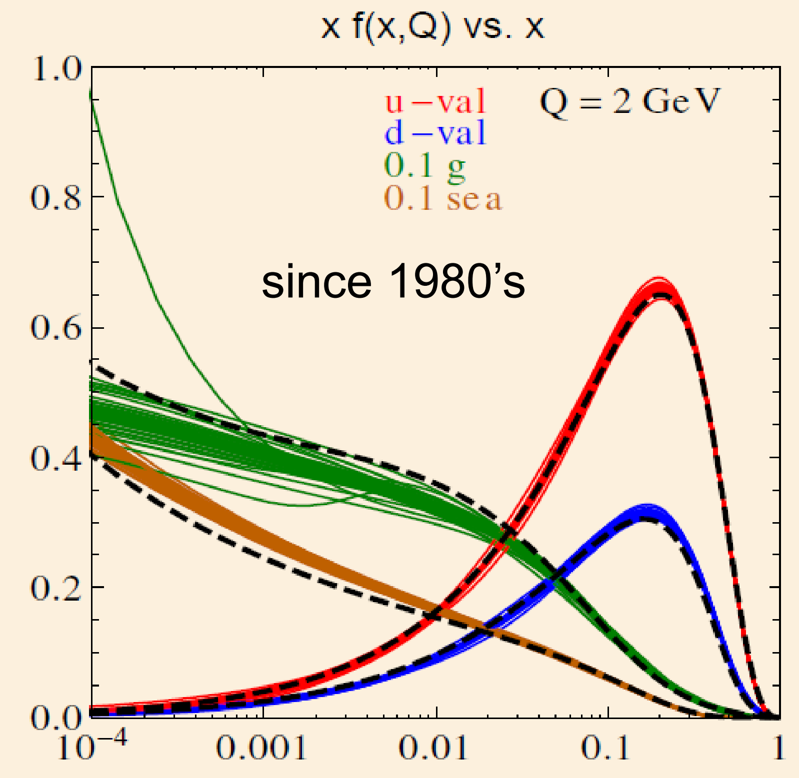
epistemic vs. **aleatory** uncertainties

Statistical uncertainty propagated from experiments — reduced by increasing data size

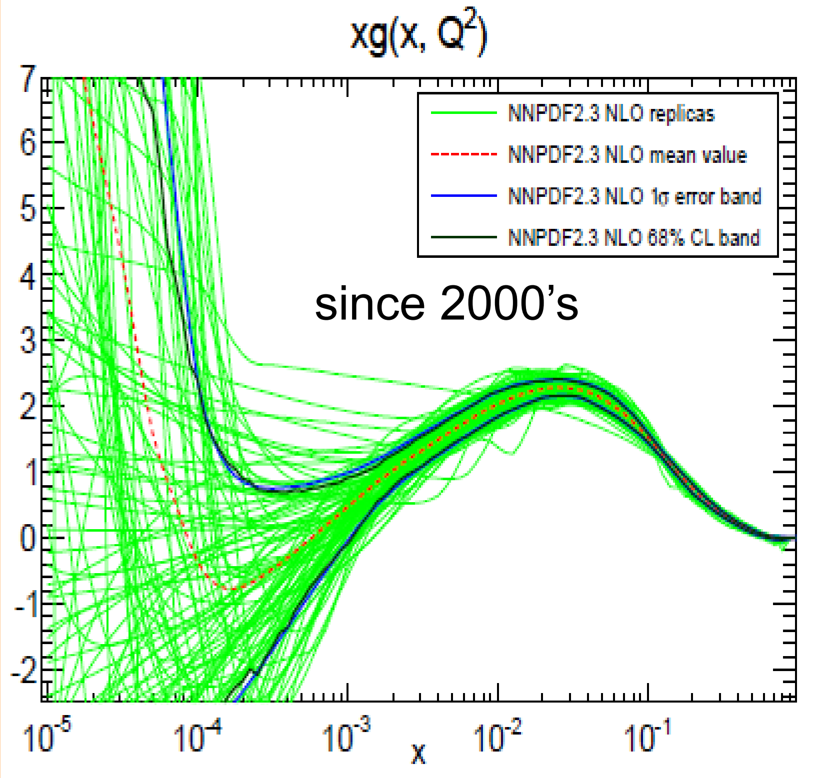
Uncertainty due to lack of knowledge —bias (may be reduced by analysis improvements)

Two types of modern error PDFs

Analytic parametrizations +
Hessian PDF eigenvector sets
(ABM, CTEQ, HERA, MMHT,…)



Neural network parameterizations
+ Monte Carlo PDF replicas
(NNPDF)



a textbook application of ML in particle physics

Two powerful, complementary representations.
Hessian PDFs can be converted into MC ones, and vice versa.

Epistemic PDF uncertainty…

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

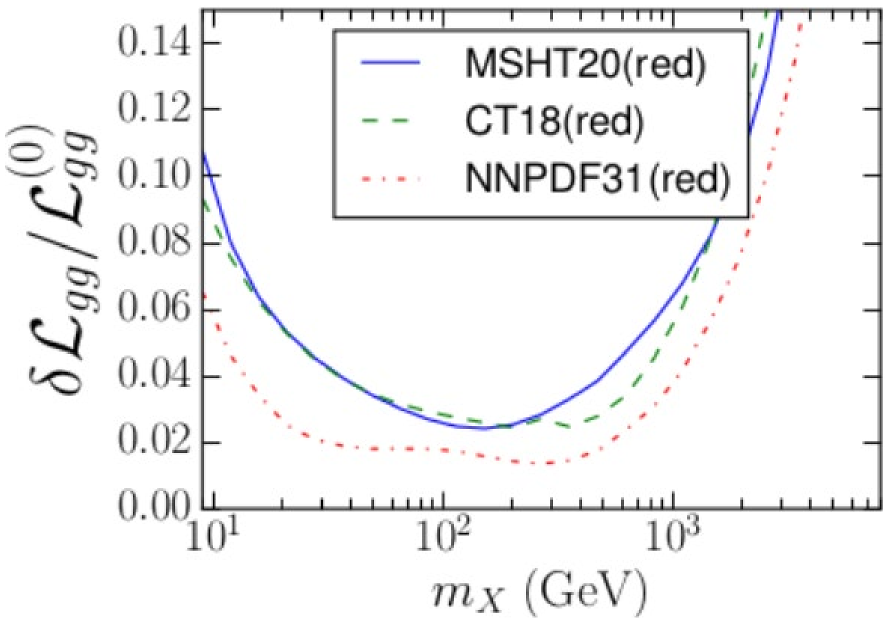
… can dominate the full uncertainty when experimental and theoretical uncertainties are small.

…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**, **arXiv:2205.10444**

A likelihood-ratio test of models T_1 and T_2

From Bayes theorem, it follows that

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

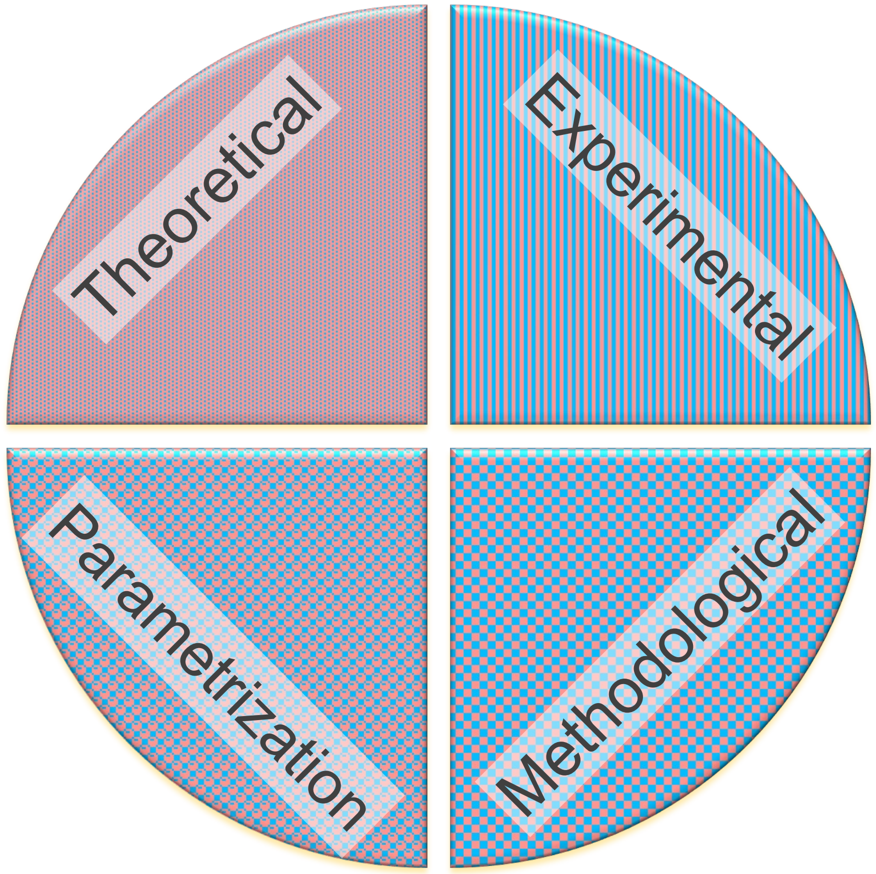
$\equiv r_{\mathrm{posterior}}$ $\equiv r_{\mathrm{likelihood}}$ $\equiv r_{\mathrm{prior}}$

aleatory   epistemic + aleatory

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

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

Components of PDF uncertainty



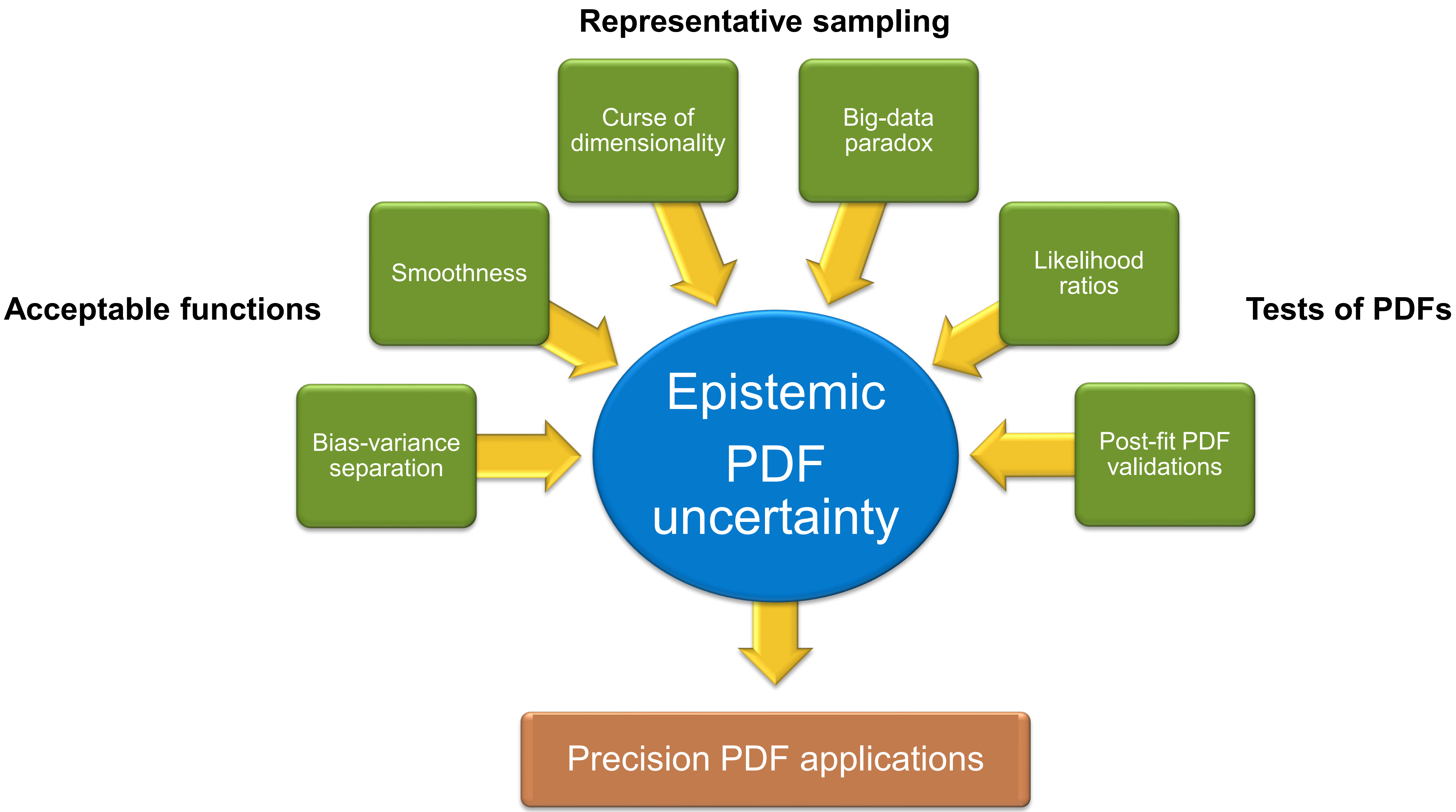
4 types of uncertainties from Kovarik et al., arXiv: 1905.06957

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]


Representative sampling
Curse of dimensionality
Big-data paradox
Acceptable functions
Smoothness
Likelihood ratios
Tests of PDFs
Bias-variance separation
Epistemic PDF uncertainty
Post-fit PDF validations
Precision PDF applications

Estimating the epistemic uncertainty is hard because statistics with many parameters is different!

In typical applications, especially AI/ML ones:

1. **As a rule, there is no single global minimum of** χ^2 (or another cost function)
   - "Best fits" are dominated by saddle points with the same low χ^2
2. **The law of large numbers may not work**
   - uncertainty may not decrease as $1/\sqrt{N_{rep}}$, leading to the **big-data paradox** [Xiao-Li Meng, 2018]:

The bigger the data, the surer we fool ourselves.

3. **Replication of complex measurements is daunting**

Some insights from our recent work

1. **Log-likelihood.** The commonly used χ^2 forms,

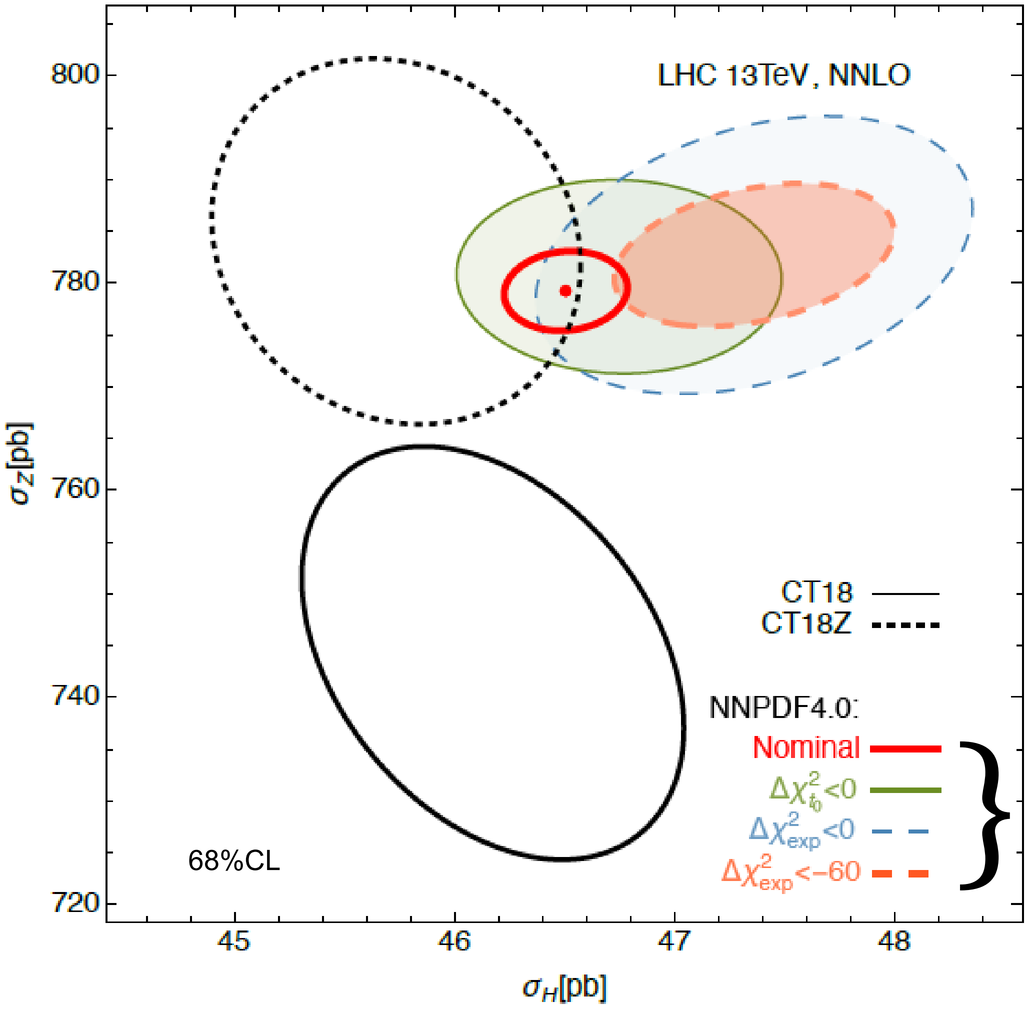
$$\chi^2 = \sum_{i=1}^{N_{pt}} \frac{\left[D_i + \sum_\alpha \beta_{i,\alpha}^{\exp} \lambda_{\alpha,\exp} - T_i\right]^2}{s_i^2} + \sum_\alpha \lambda_{\alpha,\exp}^2 \quad \text{and} \quad \chi^2 = \sum_{i,j}^{N_{pt}} (T_i - D_i)(\mathrm{cov}^{-1})_{ij}(T_j - D_j),$$

are deficient when there are many systematic parameters. **[in preparation]**

2. **ML/AI-based methods.** Any neural network is associated with a Bayesian prior. The prior depends on the architecture, hyperparameters, and training procedure. It is a source of an epistemic uncertainty. **[A. Courtoy et al.,arXiv:2205.10444].**

Example: different χ^2 treatments produce discrepant uncertainty estimates

Details in
A. Courtoy et al.,
arXiv:2205.10444



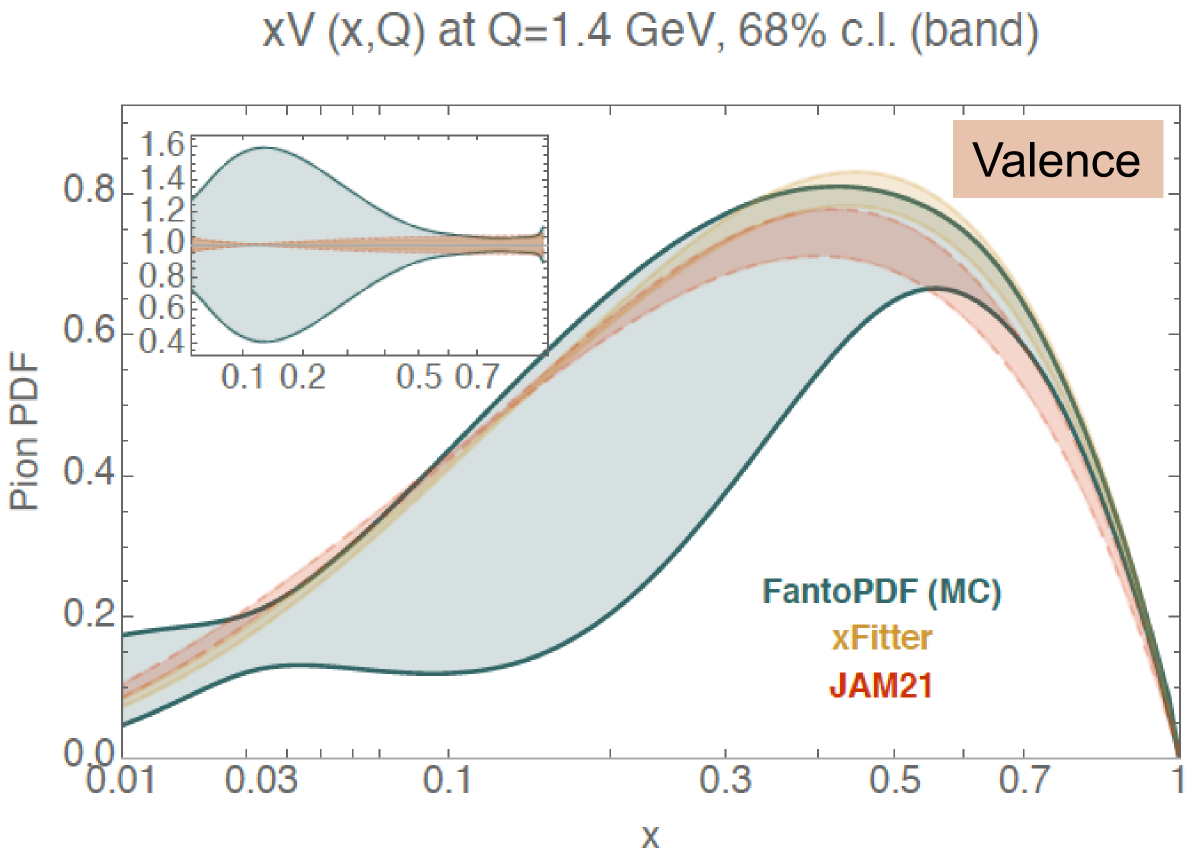
obtained with the same NNPDF4.0 fitting code using a "**hopscotch scan**" of the PDF param. space

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

Fantômas: the parametrization uncertainty on the valence pion PDF

$\pi^{\pm}$ NEW

L. Kotz, A. Courtoy, M. Chavez, P. Nadolsky, F. Olness, and others, arXiv:2309.00152, arXiv:2311.08447



We obtained an NLO PDF error ensemble for charged pions from experimental data in **xFitter** using a C++ module **Fantômas** to parameterize PDFs using **Bézier curves**

These polynomial curves are universal approximators.

The Fantomas PDF error band is based on ~ 100 alternative parametrization forms with the same or better χ^2 as in the 2021 xFitter study *[Novikov et al., arXiv:2002.02902]*

The PDF error bands are enlarged compared to xFitter'20 and JAM'21 due to estimating the parametrization uncertainty using the Fantômas & METAPDF *[arXiv:1401.00013]* techniques

I·AN QCD Network of Networks Inter-American

Bézier curve

Bézier curves are convenient for interpolating discrete data

The interpolation through Bézier curves is unique if the polynomial degree= (# points-1), there's a closed-form solution to the problem,

$$\mathcal{B}^{(n)}(x) = \sum_{l=0}^{n} c_l \; B_{n,l}(x)$$

with the Bernstein pol.

$$B_{n,l}(x) \equiv \binom{l}{n} x^l (1-x)^{n-l}.$$

The Bézier curve can be expressed as a product of matrices:

- $\underline{T}$ is the vector of x^l
- $\underline{\underline{M}}$ is the matrix of binomial coefficients
- $\underline{C}$ is the vector of Bézier coefficient, c_l, to be determined

$$\mathcal{B} = \underline{T} \cdot \underline{\underline{M}} \cdot \underline{C}$$

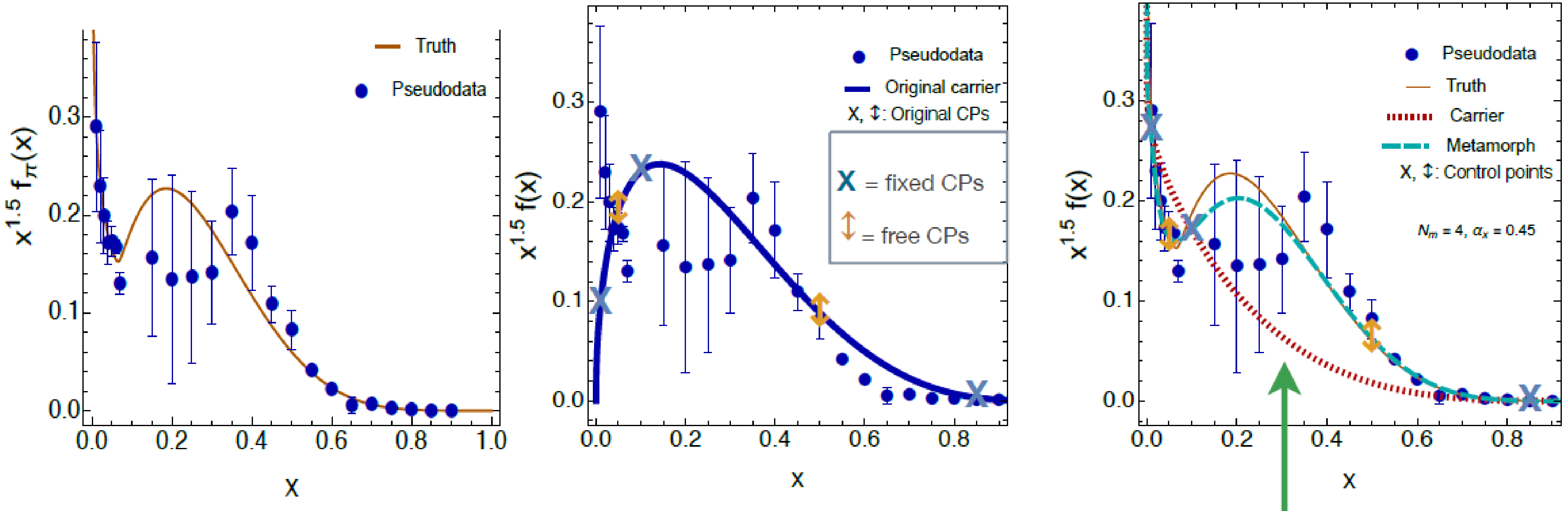
We can evaluate the Bézier curve at chosen **control points**, to get a vector of $\mathscr{B} \to \underline{P}$

$\underline{\underline{T}}$ is now a matrix of x^l expressed at the control points.

$$\underline{P} = \underline{\underline{T}} \cdot \underline{\underline{M}} \cdot \underline{C}$$

Slide by A. Courtoy

Bézier-curve methodology for global analyses — toy model



`metamorph` fit:

$$x\,q(x,Q_0^2) = A'_q\, x^{B_q}(1-x)^{C_q} \times \left(1+\mathcal{B}^{(N_m)}(x^{\alpha_x},Q_0^2;\underline{v})\right)$$

with N_m= # CPs-1 for a square-matrices system.

Shift of the control points $(\delta D_q, \ldots)$ replace free parameters

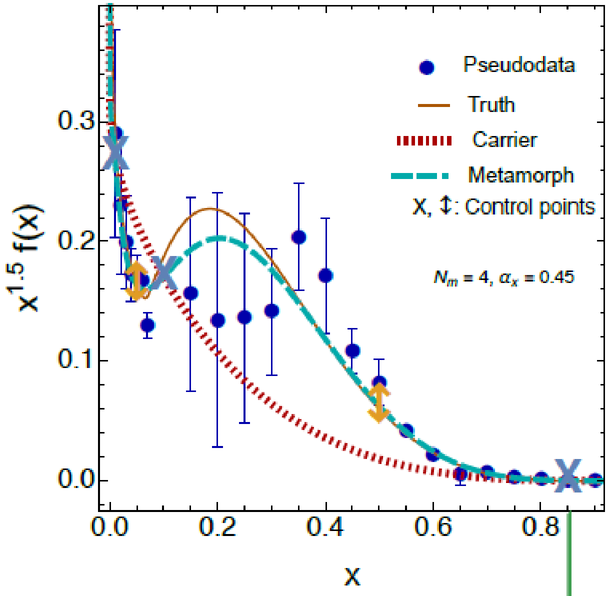
N_m= degree of polynomial can vary

δB_q & δC_q allow the carrier to vary

α_x can vary

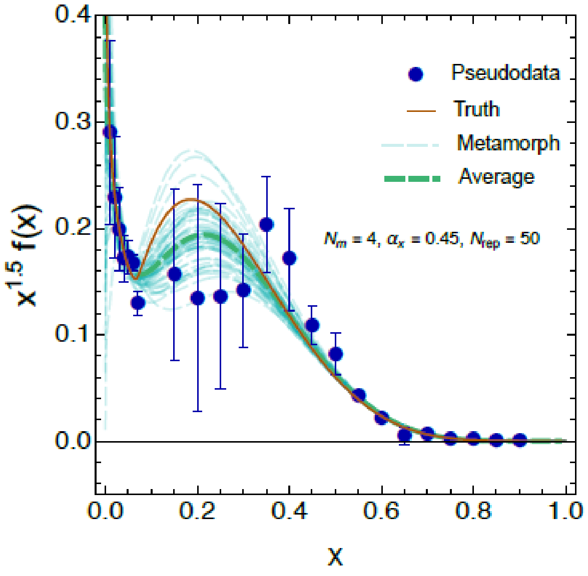
Slide by A. Courtoy

Bézier-curve methodology for global analyses — toy model



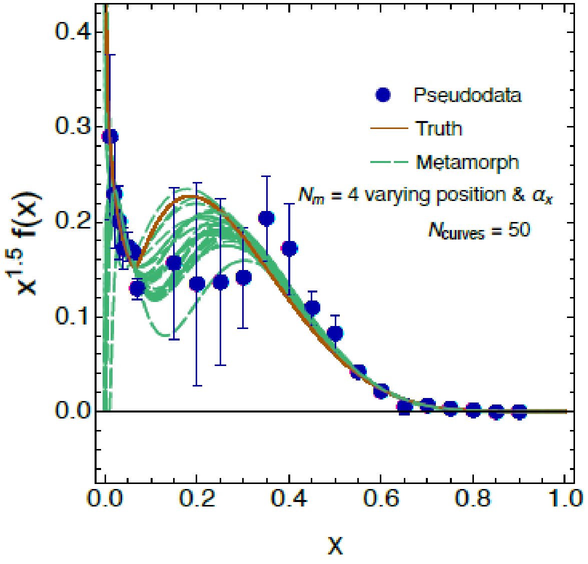
if bootstrapped

sampling on the distribution of data uncertainties



if sampled over `metamorph` settings

sampling over parametrizations



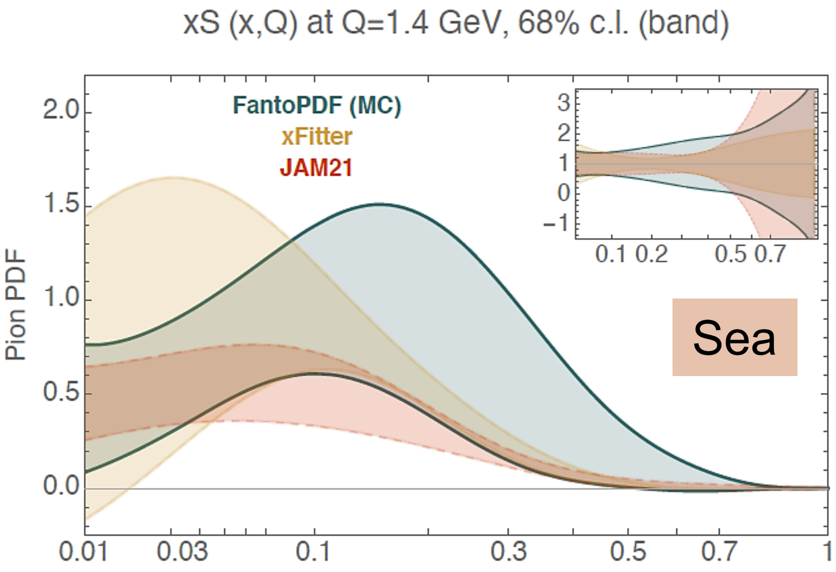
Both samplings can be done in the same analysis, they are not mutually exclusive.

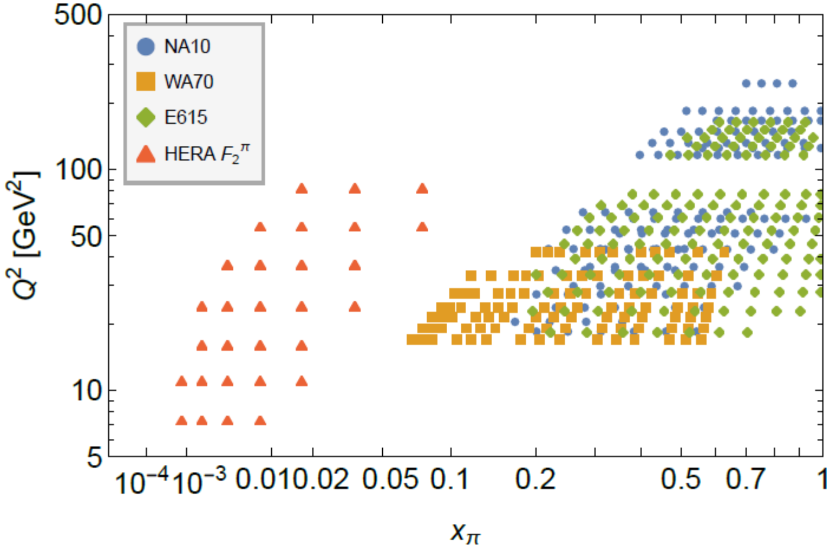
Slide by A. Courtoy

Fantômas pion PDFs: other results

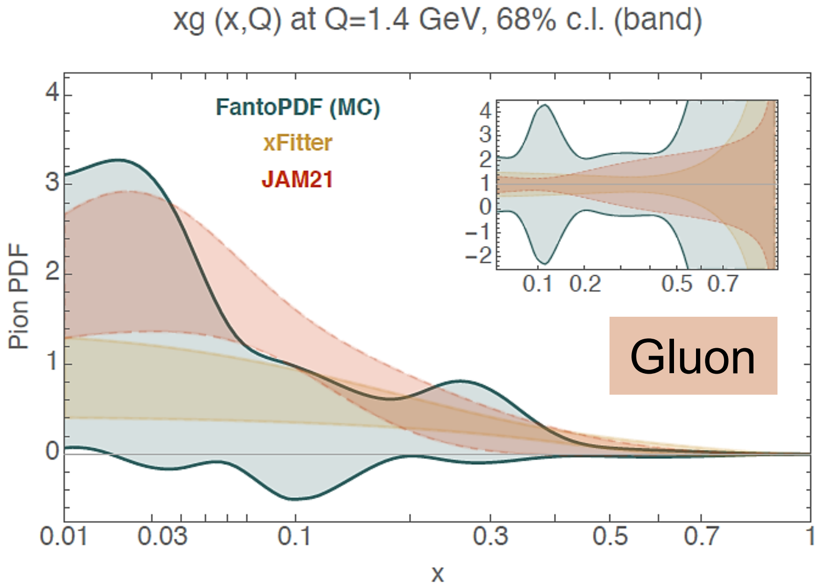
L. Kotz, A. Courtoy, M. Chavez, P. Nadolsky, F. Olness, and others, arXiv:2309.00152, arXiv:2311.08447

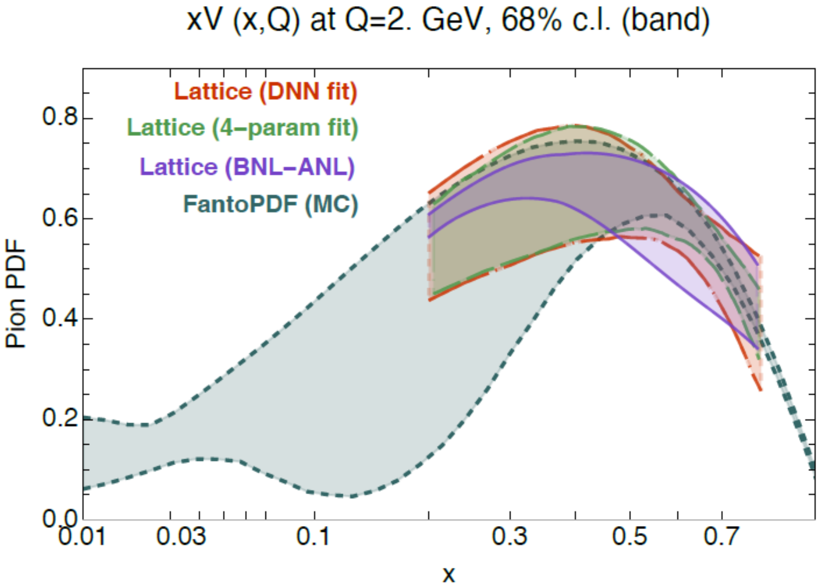
$\pi^{\pm}$ NEW



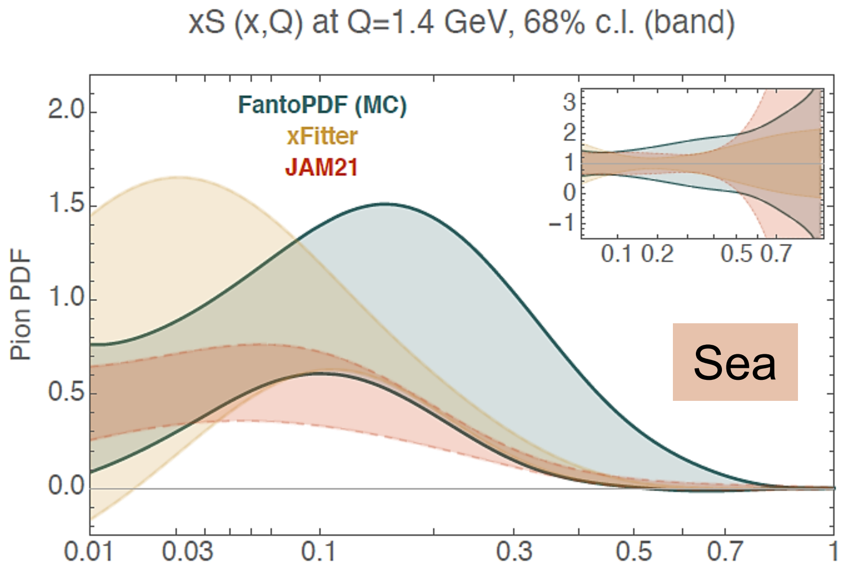


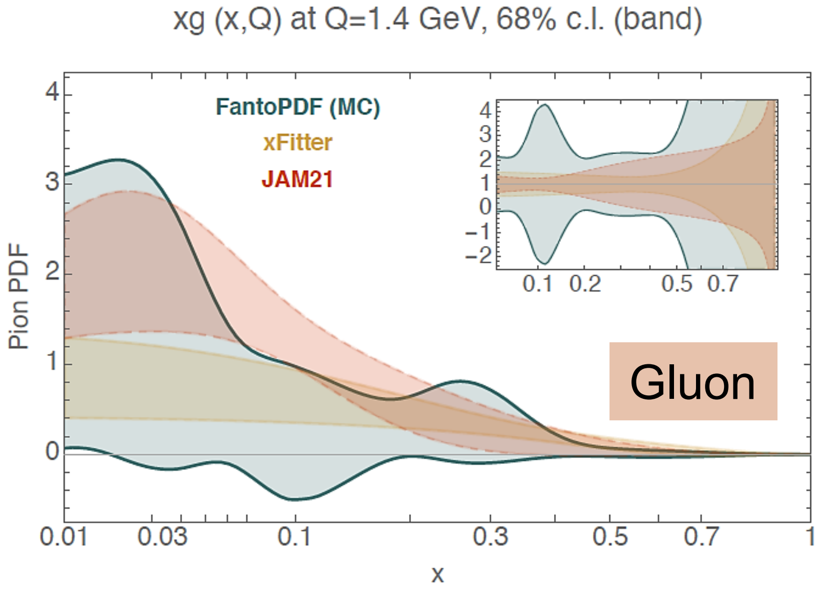
H1 leading-neutron data is included





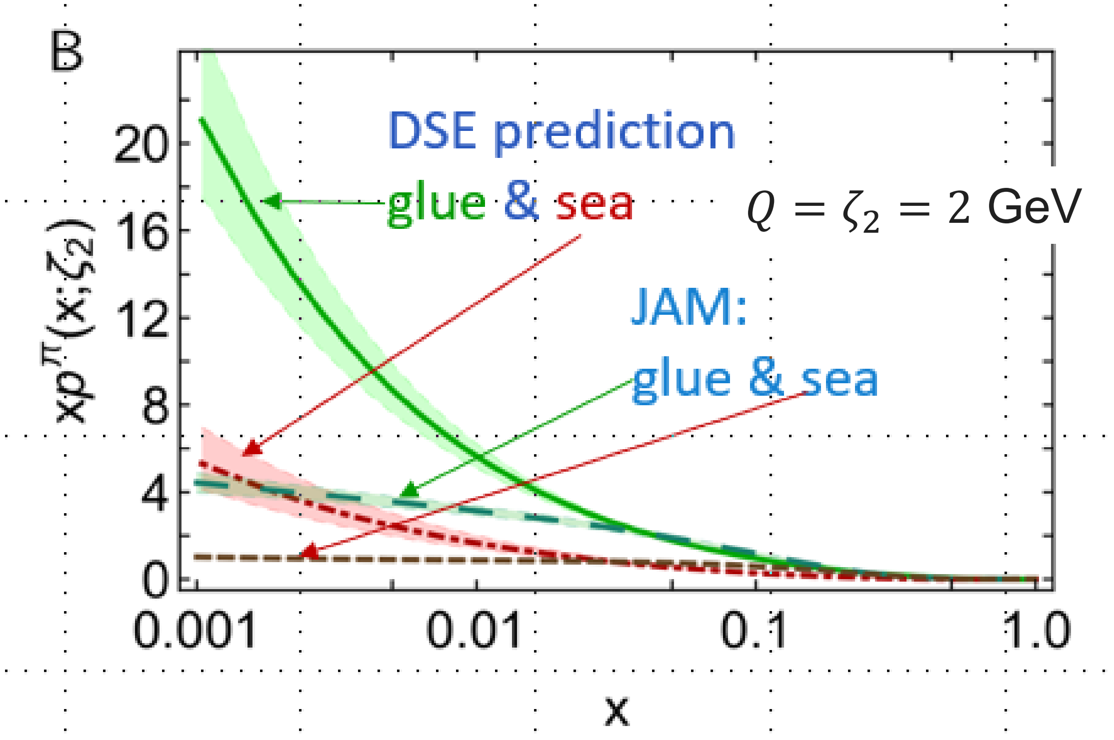
NLO pion PDFs
(Fantômas, JAM, xFitter)





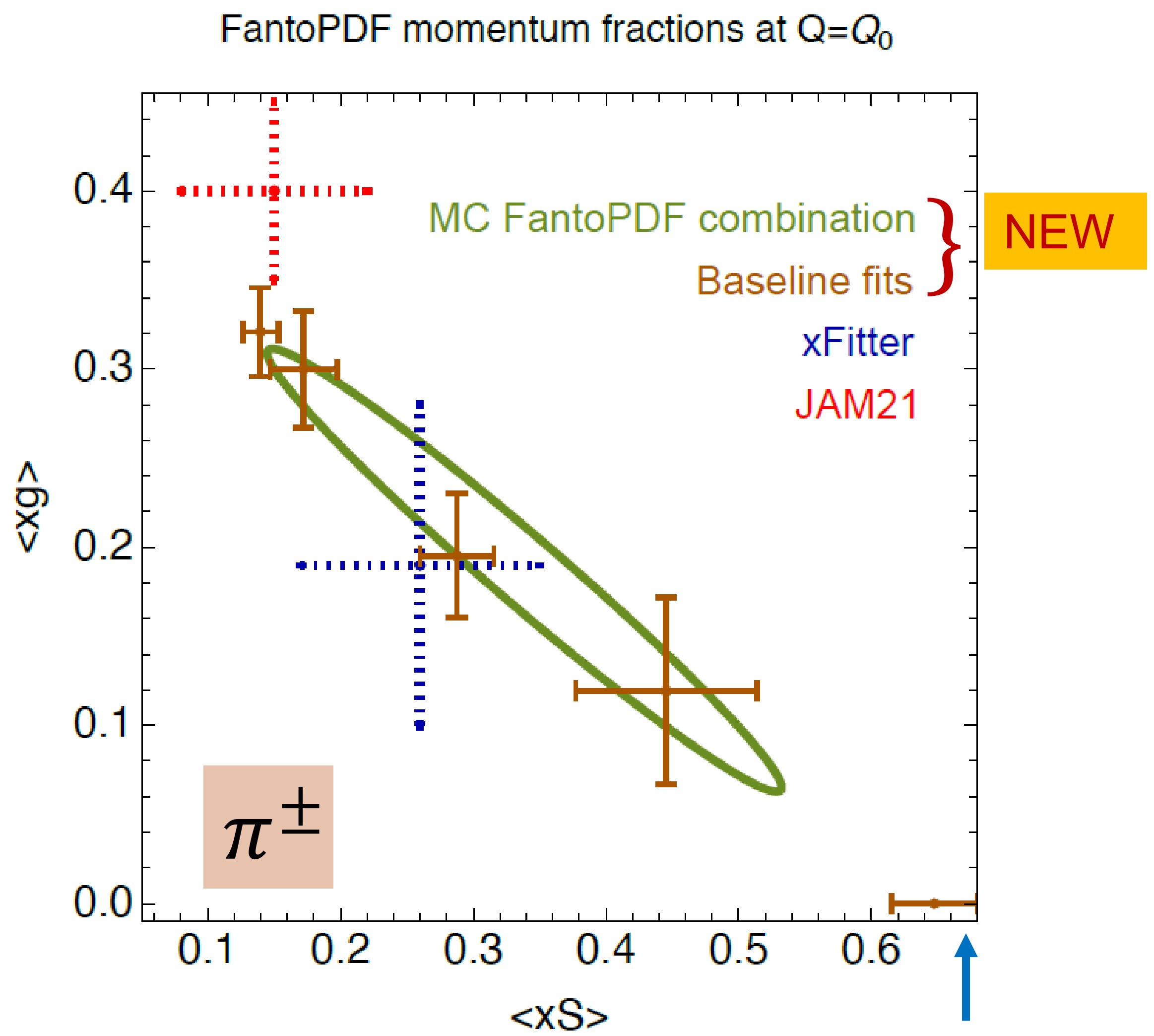
LO Dyson-Schwinger predictions

(Ding:2019lwe,Sufian:2019bol, …)

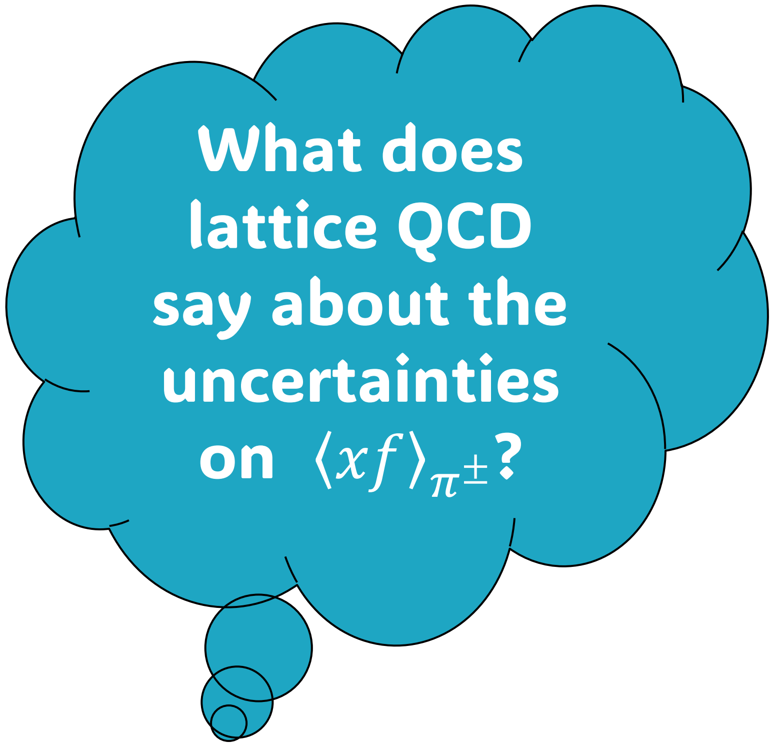


Leading-order DGLAP evolution necessitates the faster rise of glue and sea at $x \to 0$ than at NLO. Naturally explains different growth rates at $x \to 0$ between LO DSE and NLO pheno PDFs.

Fantômas pion PDFs: sea and gluon momentum fractions



a zero-momentum gluon is experimentally allowed



Moments from the lattice

Name	Q[GeV]	$\langle x(u+\bar{u})_{\pi^+}\rangle$	$\langle xg\rangle$
FantoPDF	2	0.331(25)	0.24(10)
HadStruct [19]	2	0.2541(33)	–
[Gao et al., PRD102]	3.2	0.216(19)(8)	–
ETM [46]	2	0.261(3)(6)	–
ETM [91]	2	$0.601(28)\vert_{u+d}$	0.52(11)
[Meyer et al., PRD77]	2	–	0.37(8)(12)
[Shanahan et al., PRD99]	2	–	0.61(9)
[MSU, 2310.12034]	2	–	0.364(38)(36)
ZeRo Coll. [95]	2	0.245(15)	–
[Martinelli et al., PLB196]	7	0.02	–

Lattice can access either quarks or gluons
— only the recent ETM coll. results have both.

All lattice numbers correspond to different lattice configurations.

Gluon momentum fraction varies greatly, and sometime lead to more than 100% total momentum carried!

Some calculations find that the amount of momentum carried by the gluons in the pion and the proton must be about the same.

Name	Q [GeV]	$\langle xV\rangle$	$\langle xS\rangle$	$\langle xg\rangle$
FantoPDF (DY+γ+LN)	$\sqrt{1.9}$	0.49(8)	0.34(19)	0.18(12)
xFitter [9] (DY+γ)	$\sqrt{1.9}$	0.55(6)	0.26(15)	0.19(16)
xFitter w/o scale variation	$\sqrt{1.9}$	0.55(2)	0.26(9)	0.19(9)
JAM'18 [8] (DY)	1.27	0.60(1)	0.30(5)	0.10(5)
JAM'18 [8] (DY+LN)	1.27	0.54(1)	0.16(2)	0.30(2)
JAM'21 [11] (DY+LN)	1.27	0.53(2)	0.14(4)	0.34(6)
CT18 NLO (proton)	$\sqrt{1.9}$	0.443(6)	0.160(10)	0.396(10)

Slide by A. Courtoy

$$C_V^{eff}(Q) \equiv \frac{\partial\left(xV(x,Q)\right)}{\partial(1-x)}$$

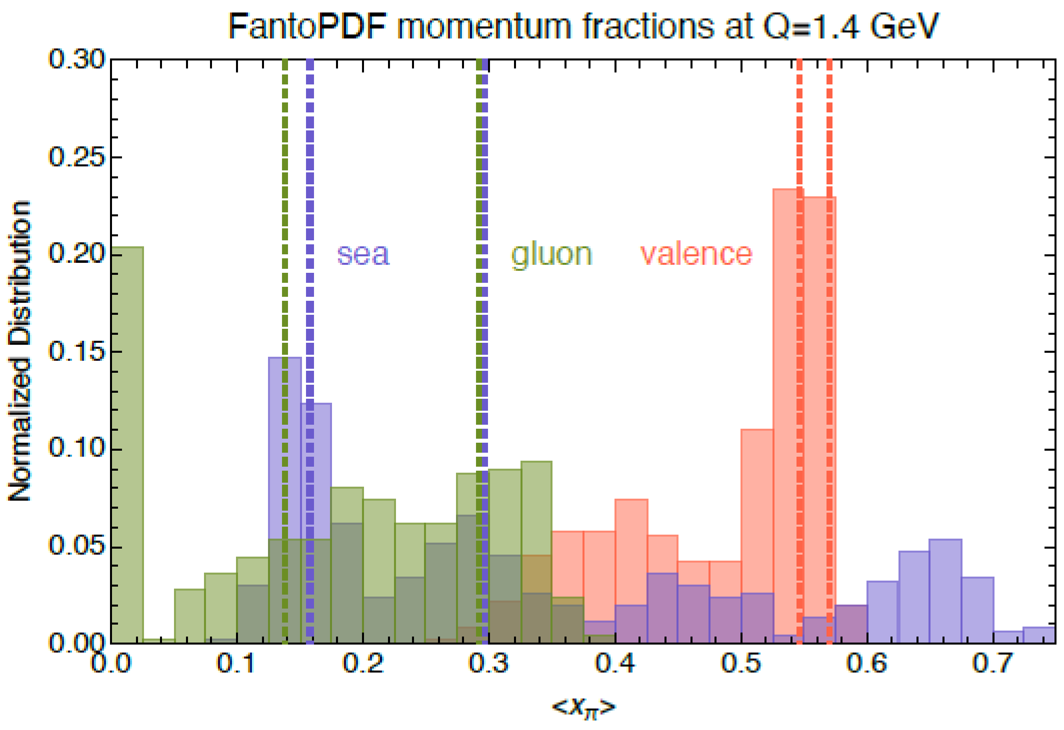


FIG. 11. The histograms represent momentum fractions for the valence (red), gluon (green) and sea (blue) PDFs from 500 MC **FantoPDF** distributions generated from five candidate fits. The histograms are not symmetric as a consequence of parametrization dependence. Vertical boundaries represent the extrema of momentum fractions for pre-**Fantômas** fits with DY+γ data only (Fig. 5). These results are at the initial scale Q_0.

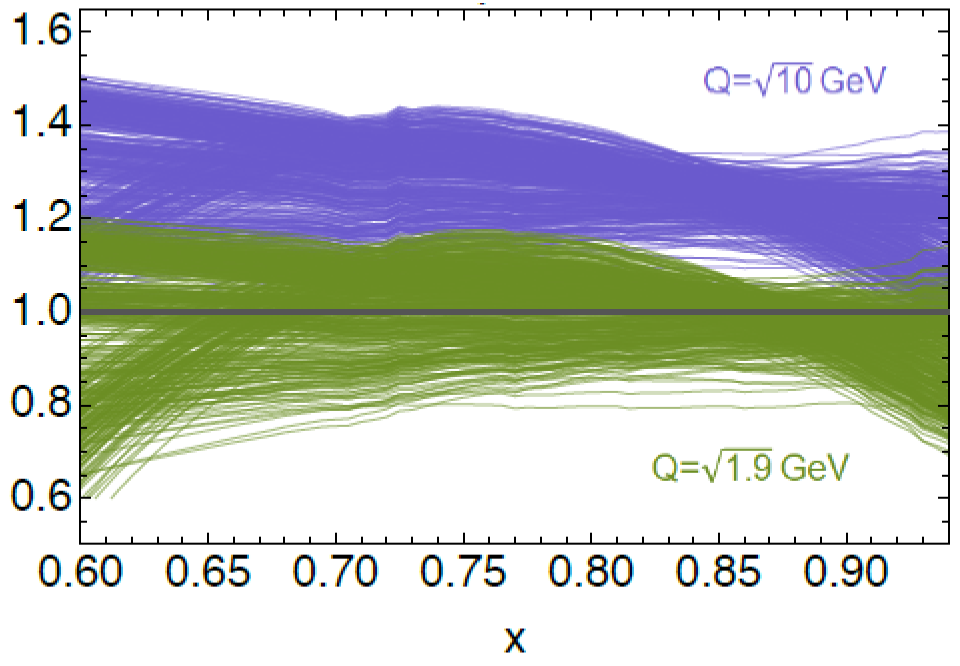
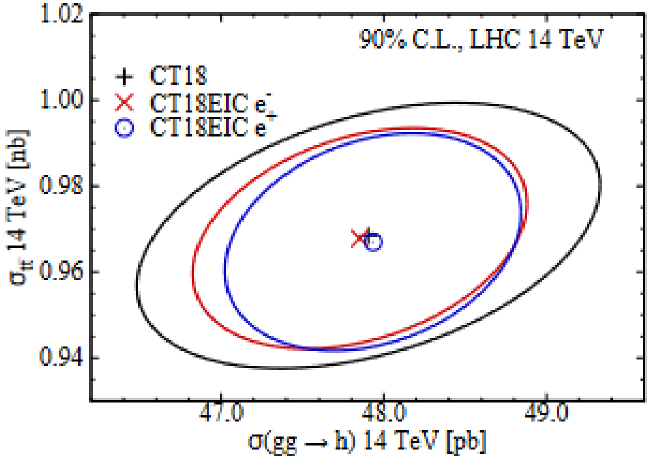


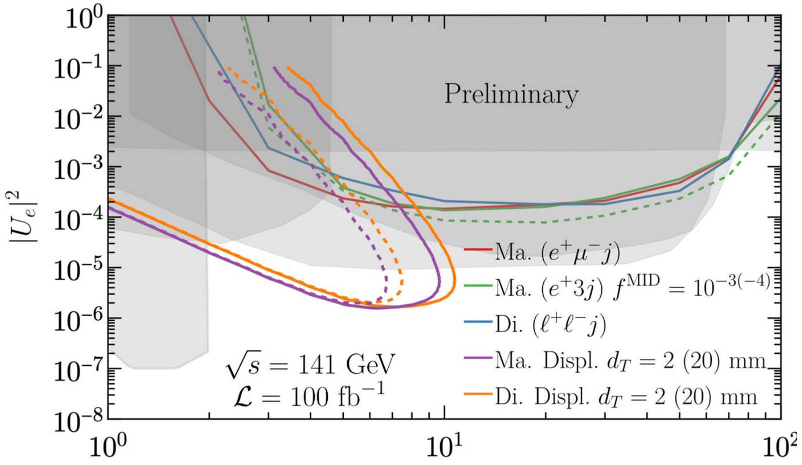
FIG. 12. The effective $(1-x)$ exponent of the valence PDF in the **FantoPDF** ensemble – the definition is given in [48]. In green, the effective exponent at $Q_0 = \sqrt{1.9}$ GeV and, in blue, at $\sqrt{10}$ GeV. The plot is cut at $x = 0.94$ for grid-extrapolation reasons. We have verified analytically that the highlighted Bézier curves of Fig. 7 converge to $C_V^{\rm eff} = 1$ at most for $x \to 1$ at Q_0.

Electron-Ion Collider: potentially a wealth of complex studies

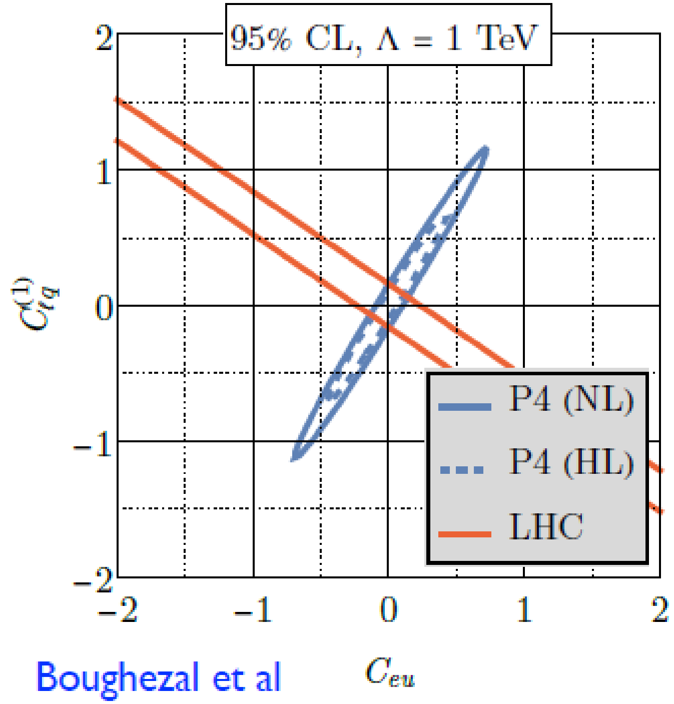
PDFs: arXiv:2103.05419



heavy neutral lepton searches arXiv: 2203.06705

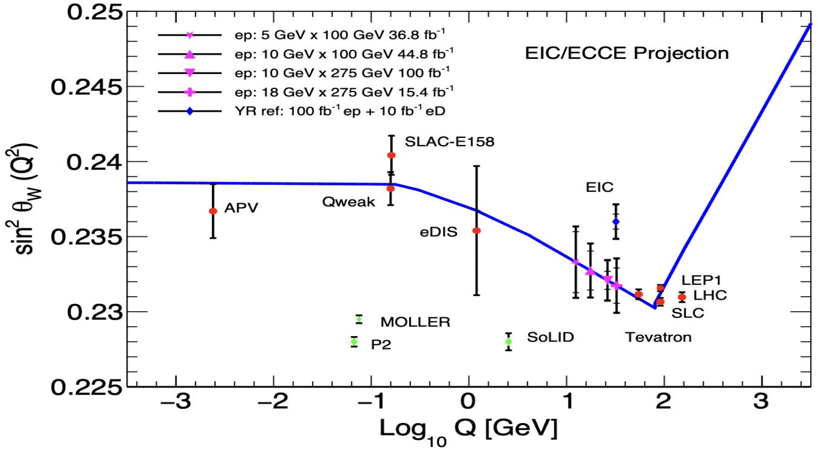


SMEFT Wilson coefficients



Boughezal et al
2004.00748, 2204.07557

weak mixing angle arXiv: 2203.13199



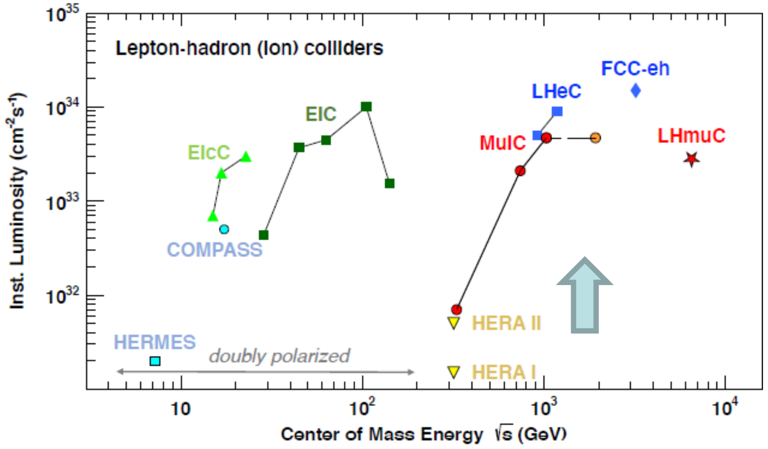
	EIC	**LHC**
$\lvert c_u^{XX} - c_u^{YY}\rvert$	0.37	15
$\lvert c_u^{XY}\rvert$	0.13	2.7
$\lvert c_u^{XZ}\rvert$	0.11	7.3
$\lvert c_u^{YZ}\rvert$	0.12	7.1
$\lvert a_{\rm Su}^{(5)TXX} - a_{\rm Su}^{(5)TYY}\rvert$	2.3	0.015
$\lvert a_{\rm Su}^{(5)TXY}\rvert$	0.34	0.0027
$\lvert a_{\rm Su}^{(5)TXZ}\rvert$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

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

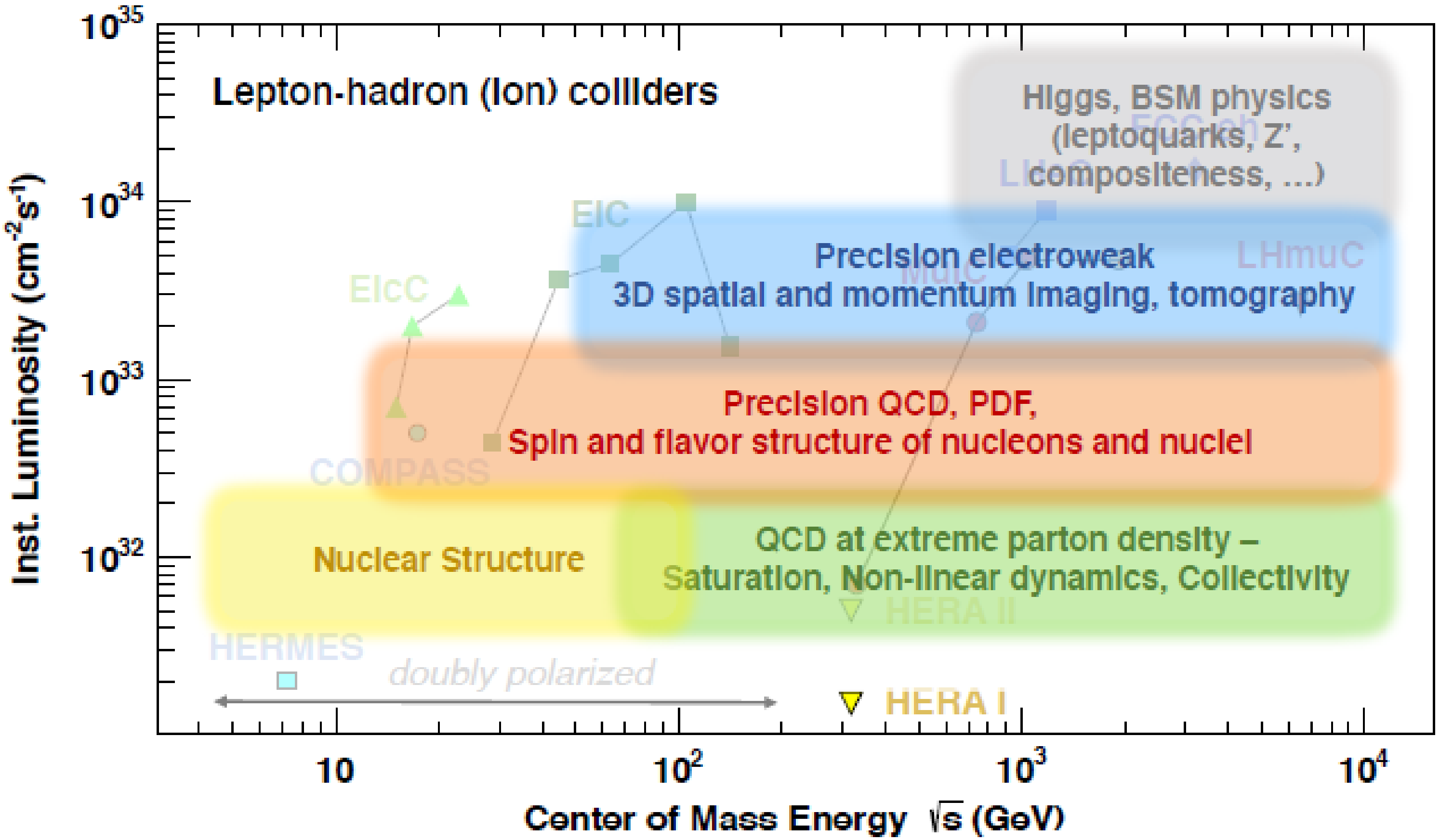
Abdul-Khalek et al., Snowmass 2021 whitepaper "EIC for HEP", 2203.13199

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 [hep-ph]
LHeC, FCC-he Study Group, arXiv:1206.2913, 2007.14491



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



Epistemic PDF uncertainty: recap

Epistemic uncertainty (due to parametrization, methodology, parametrization/NN architecture, smoothness, data tensions, model for syst. errors, ...) is increasingly important in upcoming studies 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,…