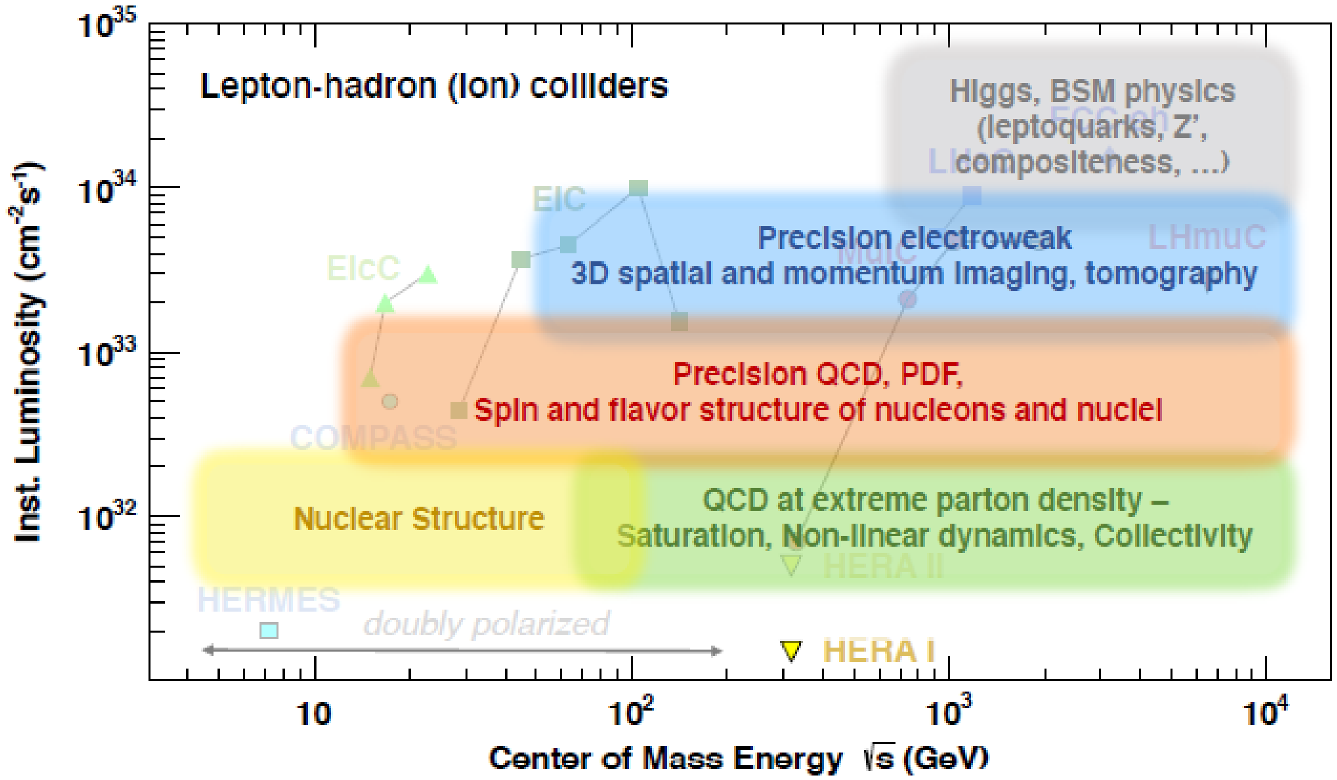
Parton Distribution Functions in the EIC era

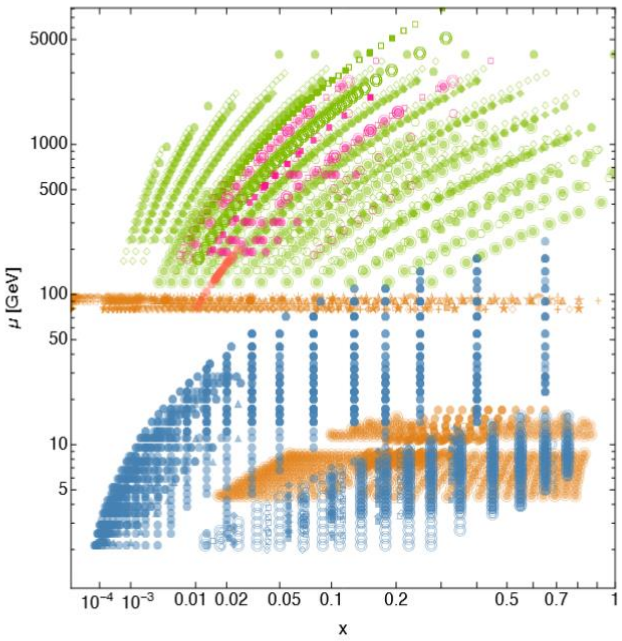
Lecture 2

Pavel Nadolsky
Michigan State University
CTEQ-TEA (Tung Et Al.) group



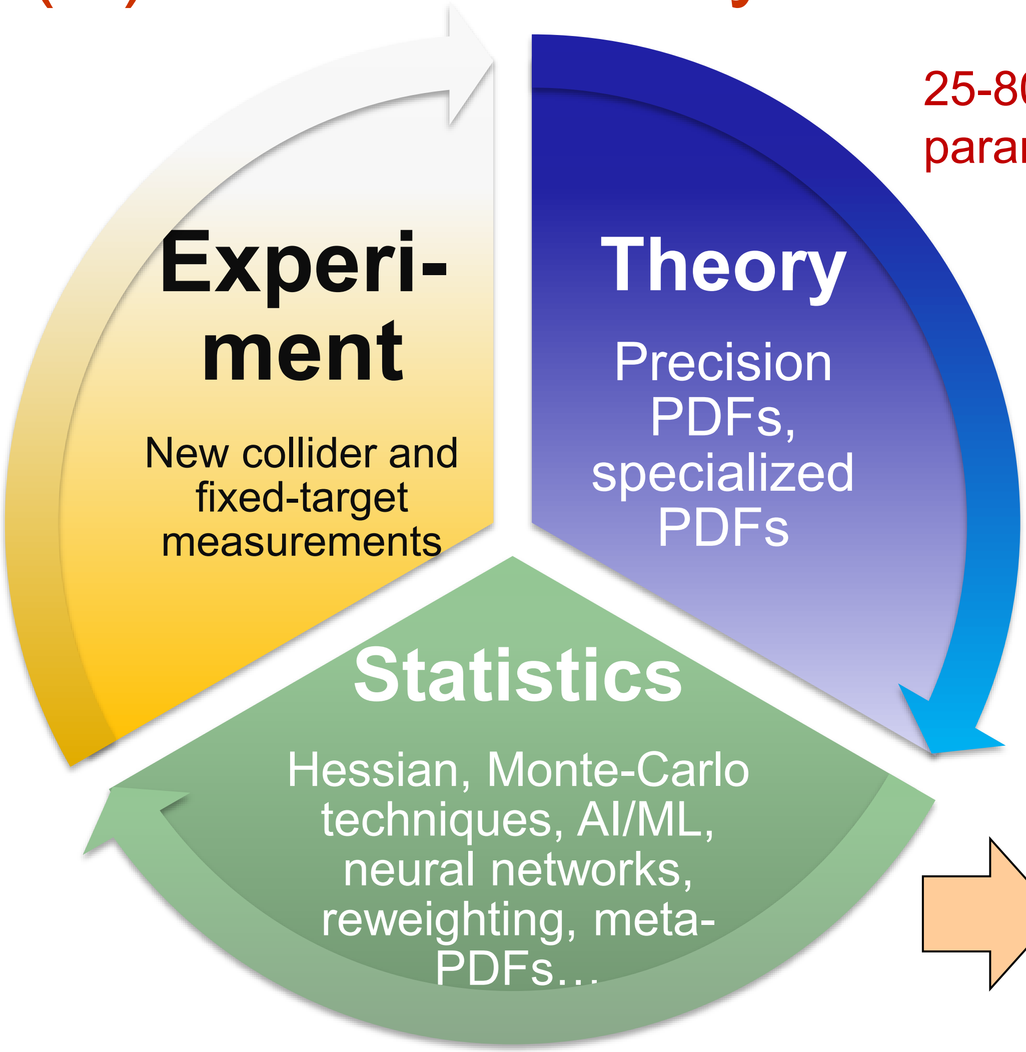
Unpolarized PDFs will be in high demand by many EIC measurements

Global fits of PDFs at (N)NNLO accuracy



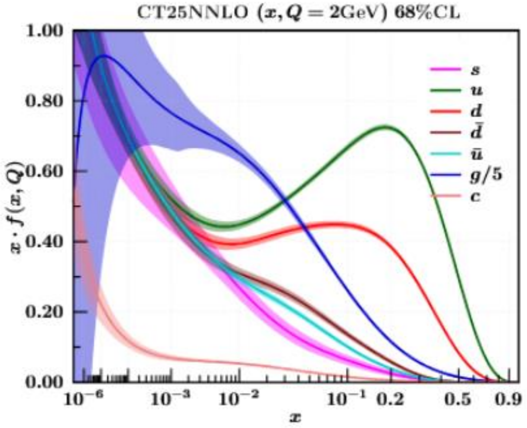
Constraints from 50+ experiments

Many systematic factors (up to several thousand)



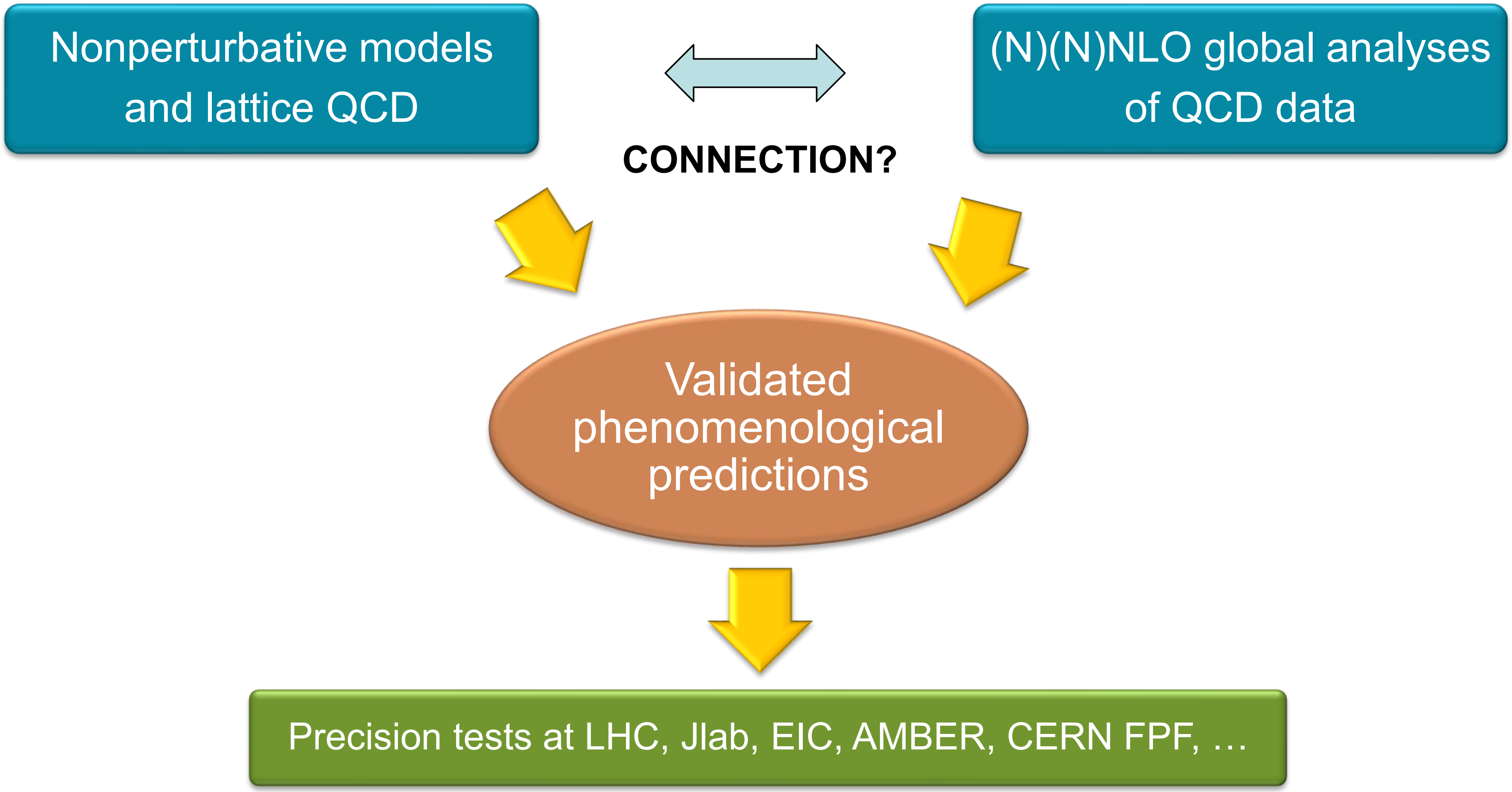
25-800 free PDF parameters

PDFs with uncertainties



Targeted accuracy < 2%

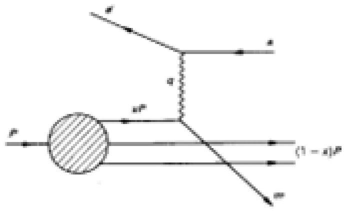
New insights about 3-dimensional structure of hadrons

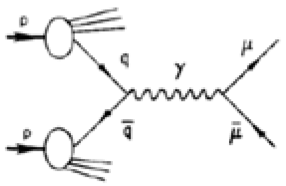


PDFs in nonperturbative QCD

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

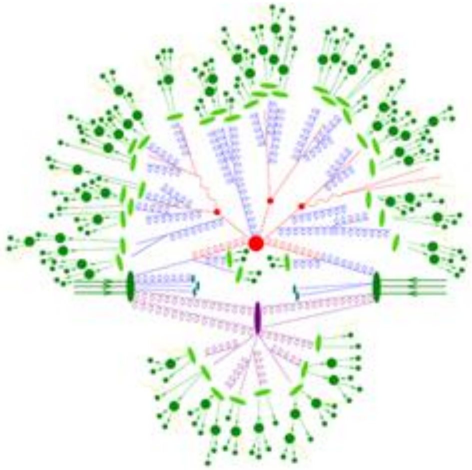
⇨ we can learn about nonperturbative dynamics by comparing predictions to data for the simplest scattering processes (DIS and DY)





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?

Frequently asked questions from users of PDFs

1. What is the meaning (definition) of my PDF?
2. How large are higher-order QCD contributions?
3. How should the PDF uncertainties be estimated?
4. How much am I biased by my parametrization of PDFs?
5. Are there hidden PDF uncertainties?
6. Are there **reliable** shortcuts to get the complete answer? [Error PDF reweighting, vector techniques, …]

These questions are best understood for unpolarized collinear PDFs

They must be addressed when trying to measure or compute the nuclear, polarized, meson, TMD, GPD, quasi-, pseudo- PDFs

References

1. Karol Kovařík, Pavel Nadolsky, Dave Soper, **Hadron structure in high-energy collisions,** *Rev.Mod.Phys.* 92 (2020) 4, 045003
   – An introduction to theoretical and statistical aspects of modern PDF fits

2. J. Ethier, E. Nocera, Parton Distributions in Nucleons and Nuclei, *Ann.Rev.Nucl.Part.Sci.* 70 (2020) 43

   Review of PDF fits from the NNPDF perspectives

3. S. Amoroso et al., **Proton structure at the precision frontier**, Acta Physica Polonica B 53 (2022) 12, A1

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

4. J. Campell, J. Huston, F. Krauss, **The Black Book of Quantum Chromodynamics,** free Adobe PDF available online

NNLO and partial N3LO PDFs

All PDF fits solve a single inverse problem

PDF determination in global QCD analyses

Inverse problem

Given a set of data D, determine $p(f|D)$ in the space of functions $f : [0,1] \to \mathbb{R}$

Solution: parametric regression

Approximate $p(f|D)$ with its projection in the space of parameters $p(\boldsymbol{\theta}|D)$

$$x f_i(x, Q_0^2) = A_{f_i}\, x^{a_{f_i}}\, (1-x)^{b_{f_i}}\, \mathscr{F}(x, \{c_{f_i}\})$$

Determine $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$ as MAP $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D)$

$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} [T_i[\boldsymbol{\theta}] - D_i](\mathrm{cov}^{-1})_{ij}[T_j[\boldsymbol{\theta}] - D_j]$$

Use a prescription to compute expectation values and uncertainties of observables

$$E[\mathcal{O}] = \int \mathcal{D}f \mathcal{P}(f|D)\mathcal{O}(f) \qquad V[\mathcal{O}] = \int \mathcal{D}f \mathcal{P}(f|D)[\mathcal{O}(f) - E[\mathcal{O}]]^2$$

Monte Carlo: $\mathcal{P}(f|D) \longrightarrow \{f_k\}$

$E[\mathcal{O}] \approx \frac{1}{N}\sum_k \mathcal{O}(f_k)$

$V[\mathcal{O}] \approx \frac{1}{N}\sum_k [\mathcal{O}(f_k) - E[\mathcal{O}]]^2$

Maximum likelihood: $\mathcal{P}(f|D) \longrightarrow f_0$

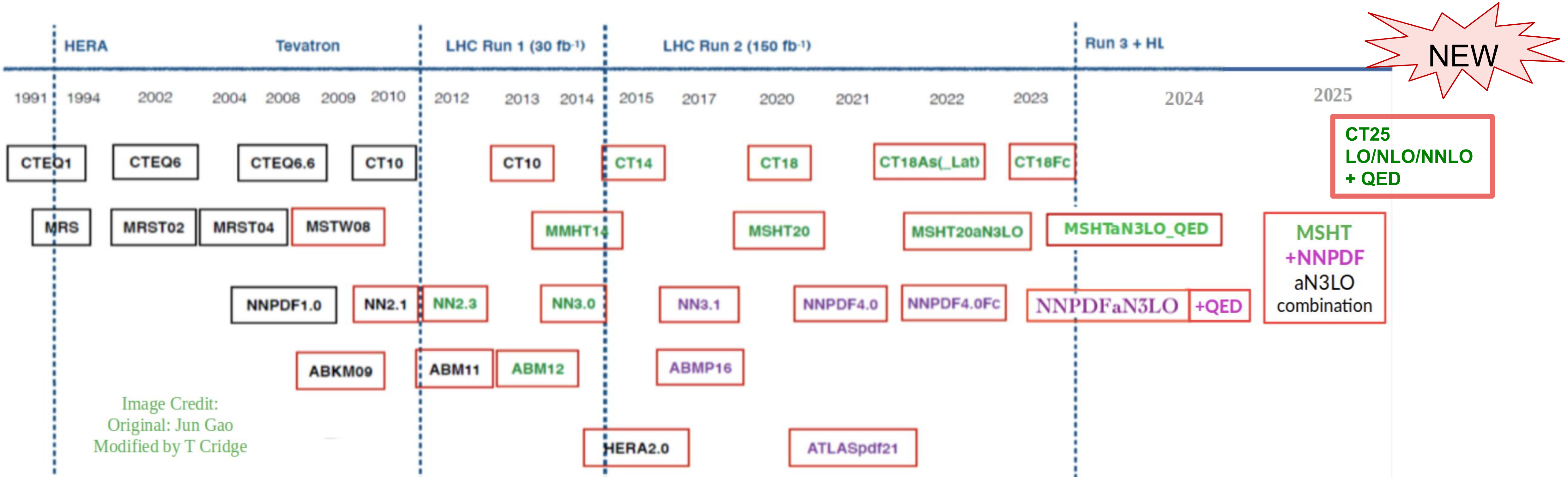
$E[\mathcal{O}] \approx \mathcal{O}(f_0)$

$V[\mathcal{O}] \approx$ Hessian, $\Delta\chi^2$envelope, ...

[For details, see e.g. Ann.Rev.Nucl.Part.Sci. **70** (2020) 43; Rev.Mod.Phys. **92** (2020) 045003]

E. Nocera

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, …)

Classes of PDFs

General-purpose

For (N)(N)NLO calculations with $N_f \leq 5$ active quark flavors

From several groups:
ABMP, ATLAS21,
CTEQ-Jlab (CJ'2025)
HERA2.0
CT25 (D-TOL)
MSHT'20
NNPDF4.0

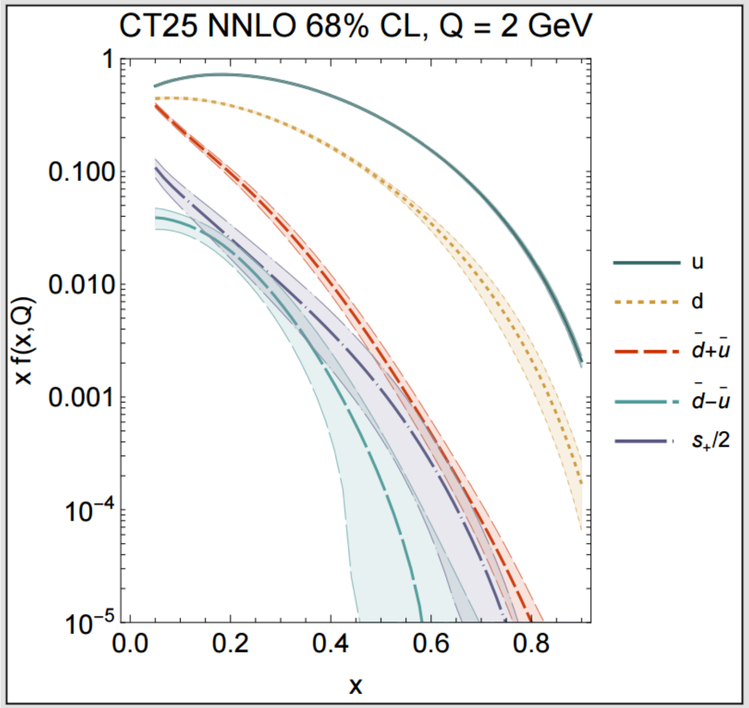
Specialized

For instance, for CTEQ-TEA:
CT25 $N_f = 3$, 4, 6
CT25 NLO and LO
CT25pd (no heavy-nucleus),
CT25m (PDF4LHC-like)
CT25FlatP (proton-only, flat prior)
…
CT18 Fitted Charm **[2211.01387]**
CT18 QCD+QED **[2305.10497]**

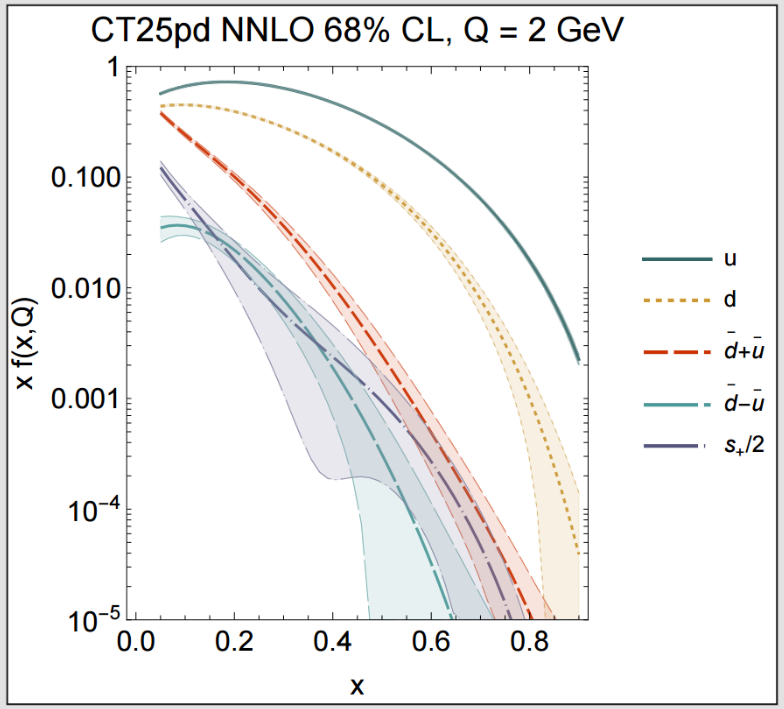
Combined **[2203.05506]**

PDF4LHC'21=CT18+MMHT'20+NNPDF3.1

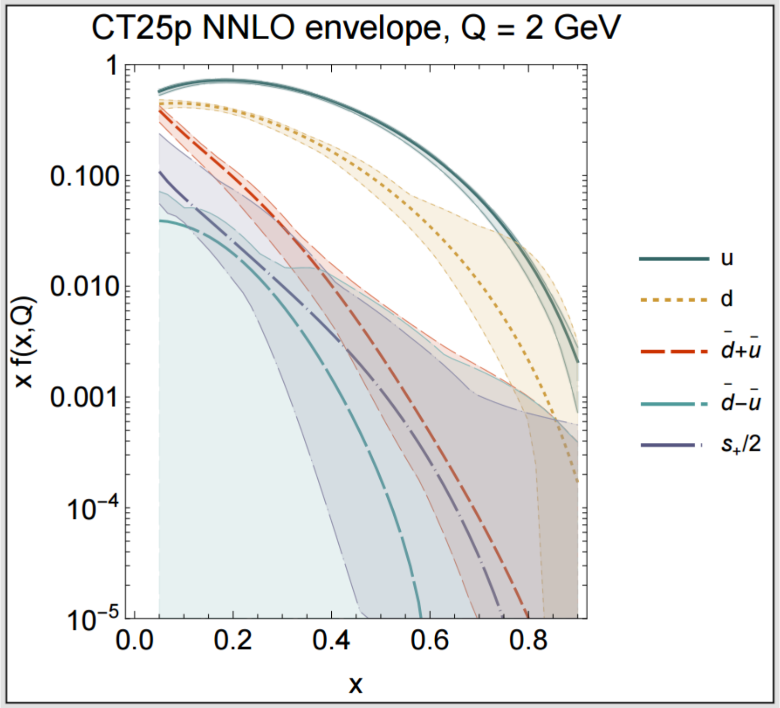
Example: quark sea PDFs in the EIC kinematic region



Full CT25 dataset



Proton-deuteron dataset; dimuon SIDIS on heavy nuclei excluded, weaker constraint on strangeness PDFs



Proton-only constraints, extended range of PDF parametrizations; no sea flavor separation at $x > 0.1$

SIDIS at the EIC will help sea flavor separation in the shown x region

Accurate (NNLO) PDFs

CT18 NNLO/CT18Z NNLO
MSHT20, NNPDF3.1/4.0 NNLO
ABMP, ATLAS, … NNLO

→ multiple PDF solutions consistent with observations

Potentially more accurate (aN3LO) PDFs

MSHT20 aN3LO
NNPDF4.0 aN3LO

→ Different from NNLO? More consistent? Unique?

Mixed NNLO-N3LO calculations

"NNLO+"

E.g., a possible CT18 NNLO+ prescription (out of several)

1. Use CT18Z NNLO or CT18 NNLO error sets
2. Central prediction: take the average of predictions with $\hat{\sigma}_{NNLO}$ and $\hat{\sigma}_{N3LO}$
3. Scale uncertainty: compute using $\hat{\sigma}_{N3LO}$

Necessary components of an N3LO PDF analysis

Component		Availability
Splitting functions		Partial N3LO
Hard cross sections	• DIS, light flavors	Full N3LO
	• NC DIS, heavy flavors	Full N3LO (Blümlein et al.), not yet in fitting codes
	• Vector boson production	Full N3LO for some processes, fixed N3LO/NLO K-factor tables
	• CC DIS, jet, $t\bar{t}$ production	N2LO
	• $pp \to W + c$, $pp \to Z + b$, $\mathrm{pp} \to b$	NLO (massive); NNLO (ZM)

Looking forward to including all components **exactly and fully** to reduce the QCD scale uncertainty and certify the N3LO accuracy in the near future.

CTEQ-TEA and other groups include some N3LO contributions in their fitting codes.
MSHT and NNPDF published NNLO+ (aN3LO) fits. Certified N3LO accuracy is still in the future.

Requirements for PDF parameterizations

Requirements for PDF parameterizations

A. A valid set of $f_{a/p}(x,Q)$ must satisfy QCD sum rules

Valence sum rule

$$\int_0^1 [u(x,Q) - \bar{u}(x,Q)]\, dx = 2 \qquad \int_0^1 [d(x,Q) - \bar{d}(x,Q)]\, dx = 1$$

$$\int_0^1 [s(x,Q) - \bar{s}(x,Q)]\, dx = 0$$

A proton has net quantum numbers of 2 u quarks + 1 d quark

Momentum sum rule

$$[\text{proton}] \equiv \sum_{a=g,q,\bar{q}} \int_0^1 x\, f_{a/p}(x,Q)\, dx = 1$$

momenta of all partons must add up to the proton's momentum

Through this rule, normalization of $g(x,Q)$ is tied to the first moments of quark PDFs

Requirements for PDF parameterizations

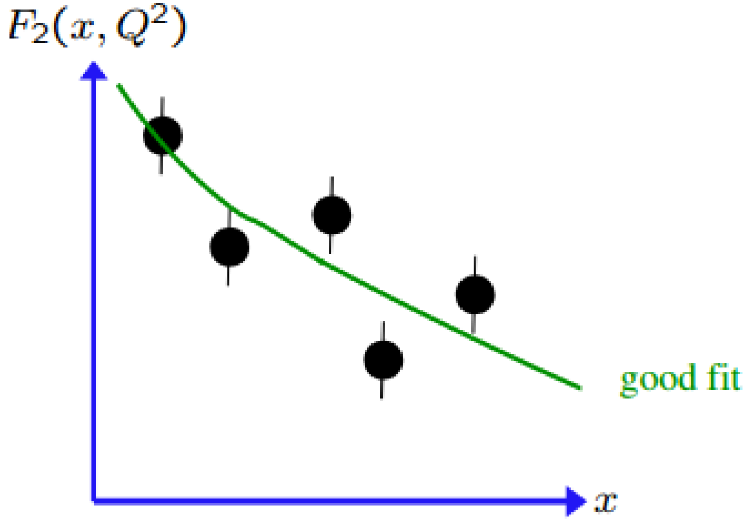
B. A valid PDF set must **not** produce unphysical predictions for observable quantities

Example

- Any concievable hadronic cross section σ must be non-negative: $\sigma \geq 0$
  - this is typically realized by requiring $f_{a/p}(x,Q) > 0$
- Any cross section asymmetry A must lie in the range $-1 \leq A \leq 1$
  - this constrains the range of allowed PDF pararametrizations
- Independent parametrizations for $s - \bar{s}$,γ, c may be difficult to constrain or separate from other factors

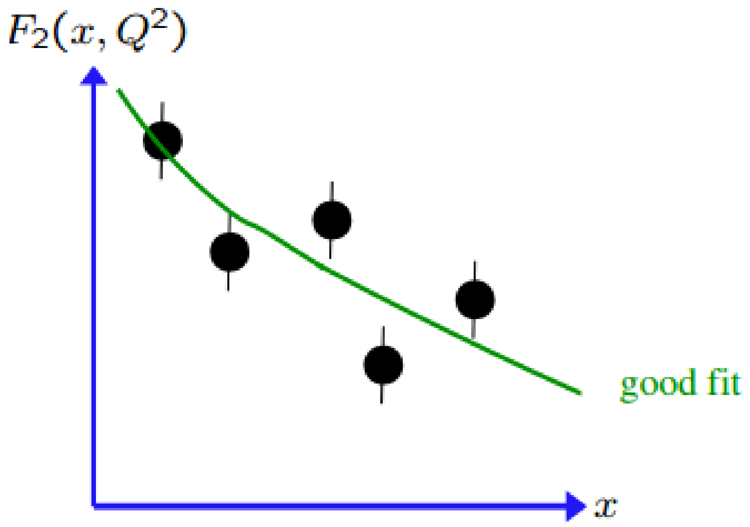
Requirements for PDF parameterizations

PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without violating QCD constraints (sum rules, positivity, ...) or reproducing random fluctuations



Requirements for PDF parameterizations

PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without violating QCD constraints (sum rules, positivity, ...) or reproducing random fluctuations



Traditional solution

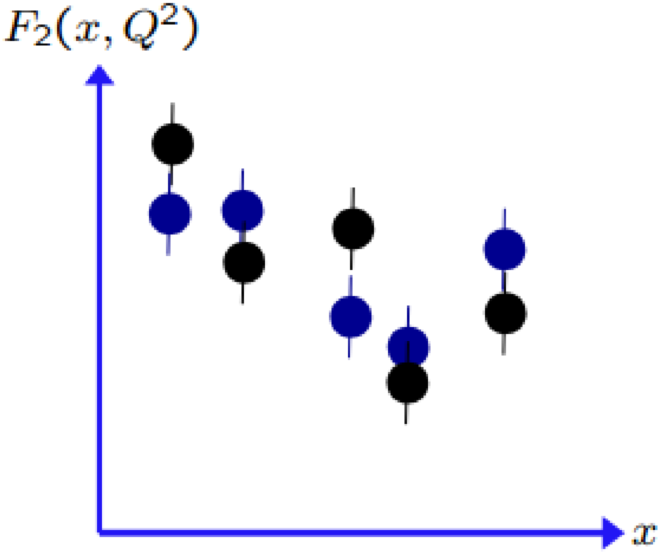
"Theoretically motivated" functions with a few parameters

$$f_{i/p}(x,Q_0) = a_0 x^{a_1}(1-x)^{a_2} \times F(x;\, a_3, a_4, \ldots)$$

- $x \to 0$: $f \propto x^{a_1}$ – Regge-like behavior
- $x \to 1$: $f \propto (1-x)^{a_2}$ – quark counting rules

Requirements for PDF parameterizations

PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without violating QCD constraints (sum rules, positivity, ...) or reproducing random fluctuations



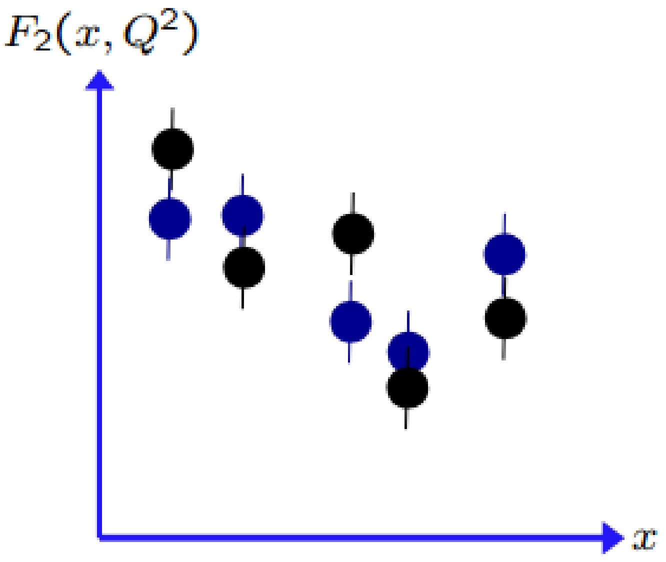
Resampling and bootstrapping

Neural Network PDF collaboration

- Generate N replicas of the experimental data, randomly scattered around the original data in accordance with the probability suggested by the experimental errors
- Divide the replicas into a fitting sample and control sample

Requirements for PDF parameterizations

PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without violating QCD constraints (sum rules, positivity, ...) or reproducing random fluctuations



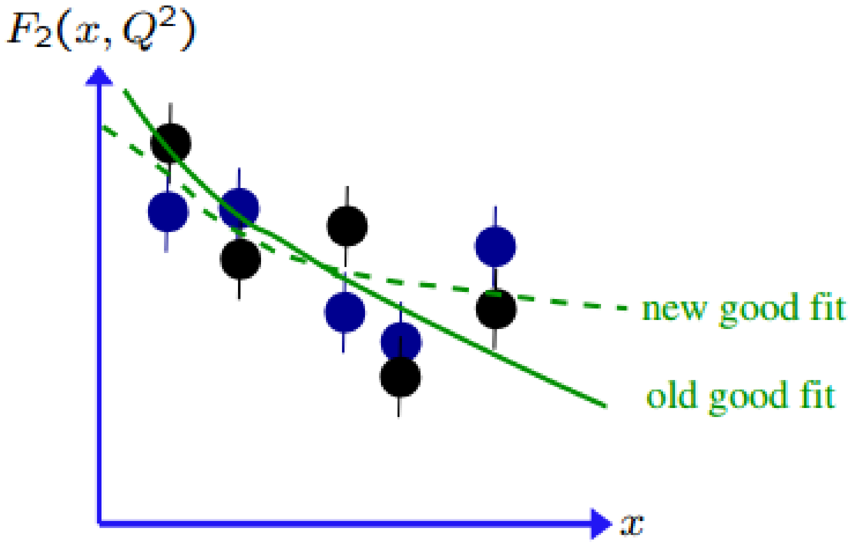
Resampling and bootstrapping

Neural Network PDF collaboration

- Parametrize $f_{a/p}(x,Q)$ by ultra-flexible functions — neural networks
- A statistical theorem states that any function can be approximated by a neural network with a sufficient number of nodes (in practice, of order 10)

Requirements for PDF parameterizations

PDF parametrizations for $f_{a/p}(x,Q)$ must be "flexible just enough" to reach agreement with the data, without violating QCD constraints (sum rules, positivity, ...) or reproducing random fluctuations

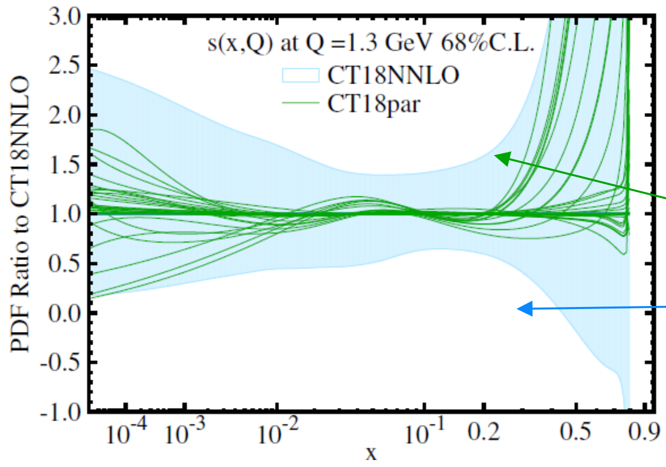


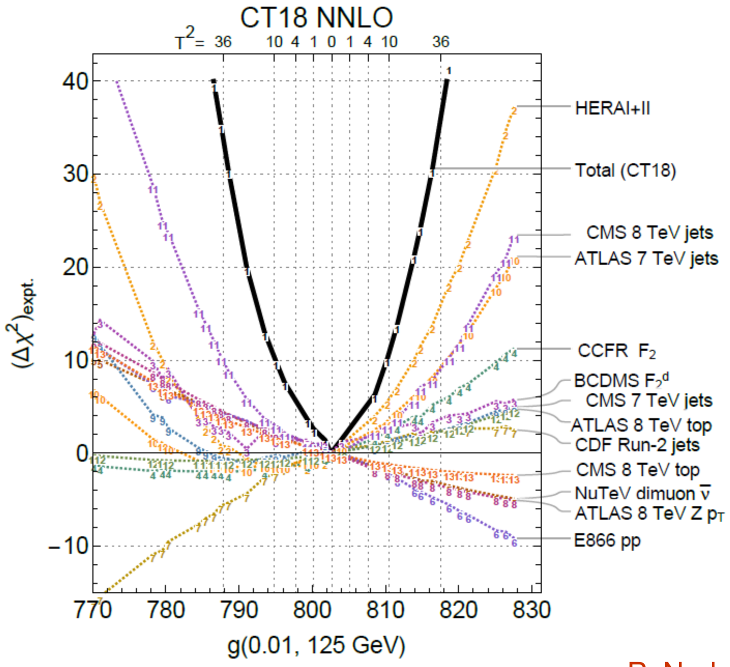
Resampling and bootstrapping

Neural Network PDF collaboration

- Fit the neural nets to the fitting sample, while demanding good agreement with the control sample
- Smoothness of $f_{a/p}(x,Q)$ is preserved, despite its nominal flexibility

Hessian PDFs, uncertainties (CT18 PDFs)





"Bayesian exploration with Gaussian emulation"

preliminary PDFs for alternative parametrizations

final uncertainty with one parametrization

Preliminary fits explore experimental, theoretical, parametrization, methodological uncertainties

The final PDFs are released as one quasi-Gaussian (**Hessian**) error set (50-60) that approximates the total uncertainty due to the above factors.

These error sets are constructed with a fixed choice of polynomial parametrization forms. The totality of error sources (not only experimental) is emulated by introducing **tolerance** T: the final 1σ uncertainty corresponds to $\Delta\chi^2 = T^2 \sim 10 - 30$, rather than $T^2 = 1$.

An alternative: Neural-network PDFs

Use **bootstrap** to estimate **aleatory** data fluctuations for a fixed training methodology (called "importance sampling" by NNPDF)

Parametrize PDFs using CNNs with optimized hyperparameters and restricted by prior conditions (positivity of cross sections, etc.)

[The whole fit is based on one "tuned" NN architecture]

All fitting codes, especially the NNPDF one, employ grid techniques for fast integration

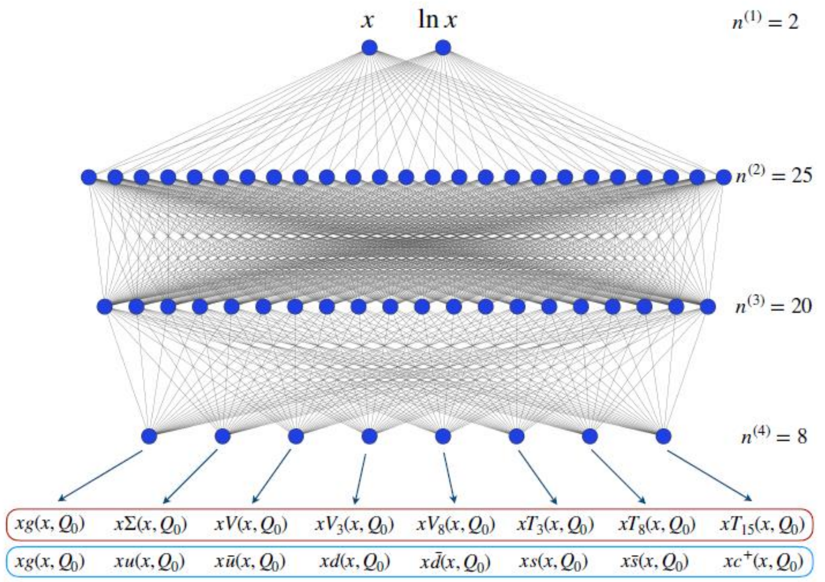


Figure 3.9. The neural network architecture adopted for NNPDF4.0. A single network is used, whose eight output values are the PDFs in the evolution (red) or the flavor basis (blue box). The architecture displayed corresponds to the optimal choice in the evolution basis; the optimal architecture in the flavor basis is different as indicated by Table 3.3).

NNPDF4.0 PDF ensemble,
R. Ball et al., arXiv:2109.02653

PDF uncertainties and the tolerance puzzle

Comparisons of the latest PDF sets

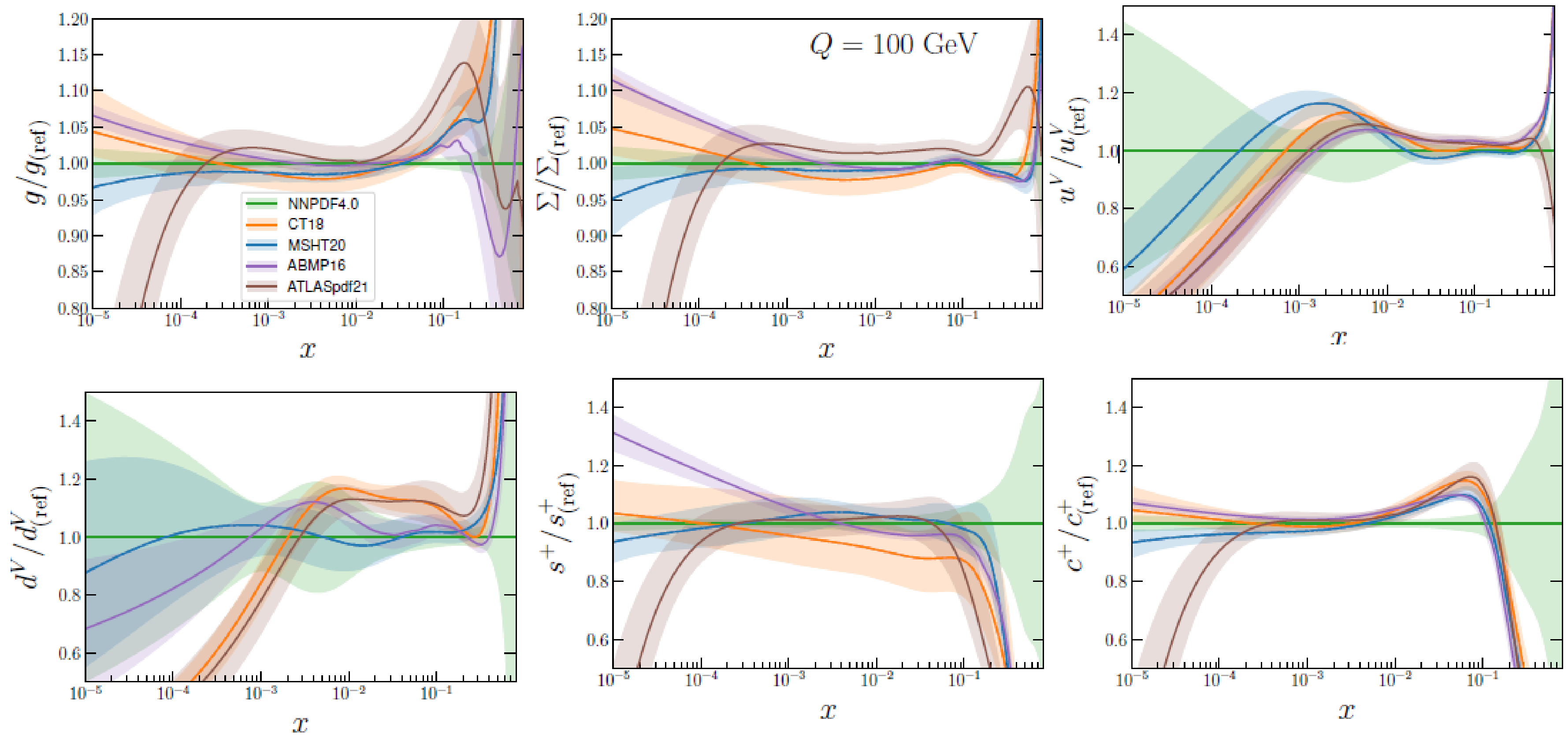
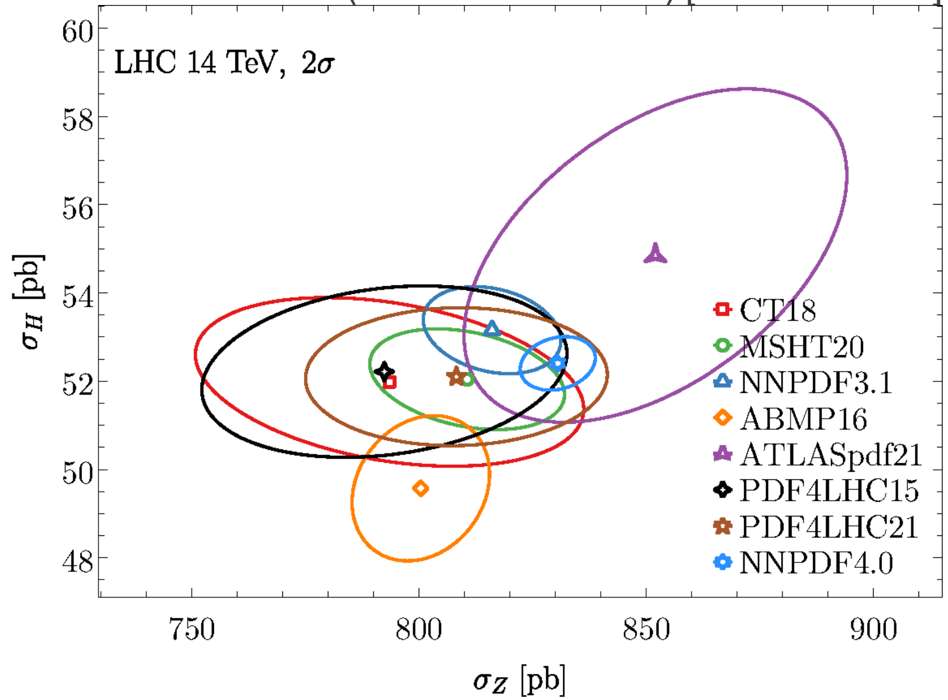


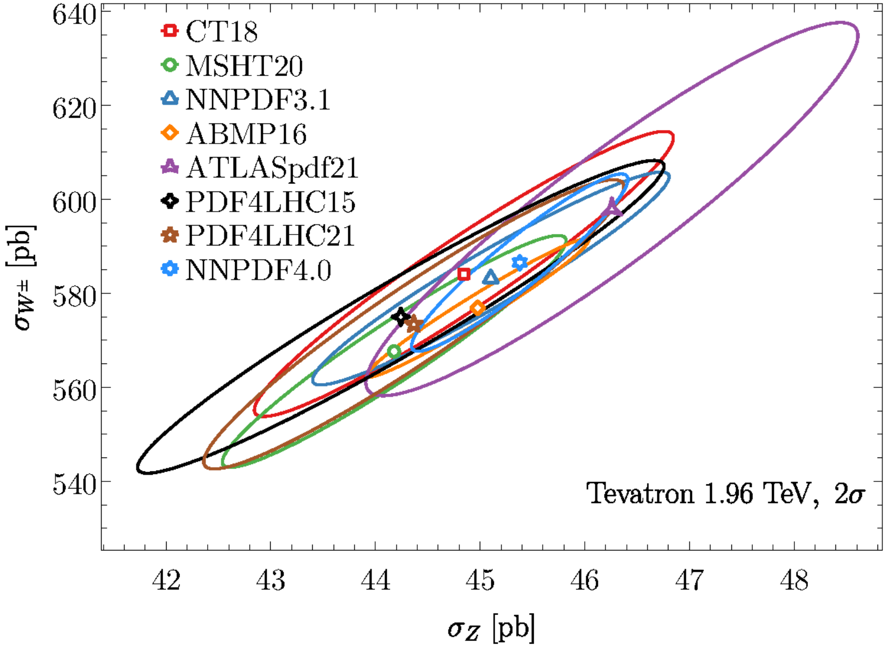
FIG. 2. Comparison of the PDFs at $Q = 100$ GeV. The PDFs shown are the N2LO sets of NNPDF4.0, CT18, MSHT20, ABMP16 with $\alpha_s(M_Z) = 0.118$, and ATLASpdf21. The ratio to the NNPDF4.0 central value and the relative 1σ uncertainty are shown for the gluon g, singlet Σ, total strangeness $s^+ = s +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

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.

PDF uncertainty: pheno classification

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.

(3 and 4:) associated with the **epistemic** uncertainty; explain several differences among the PDF fits

Kovarik et al., arXiv: 1905.06957

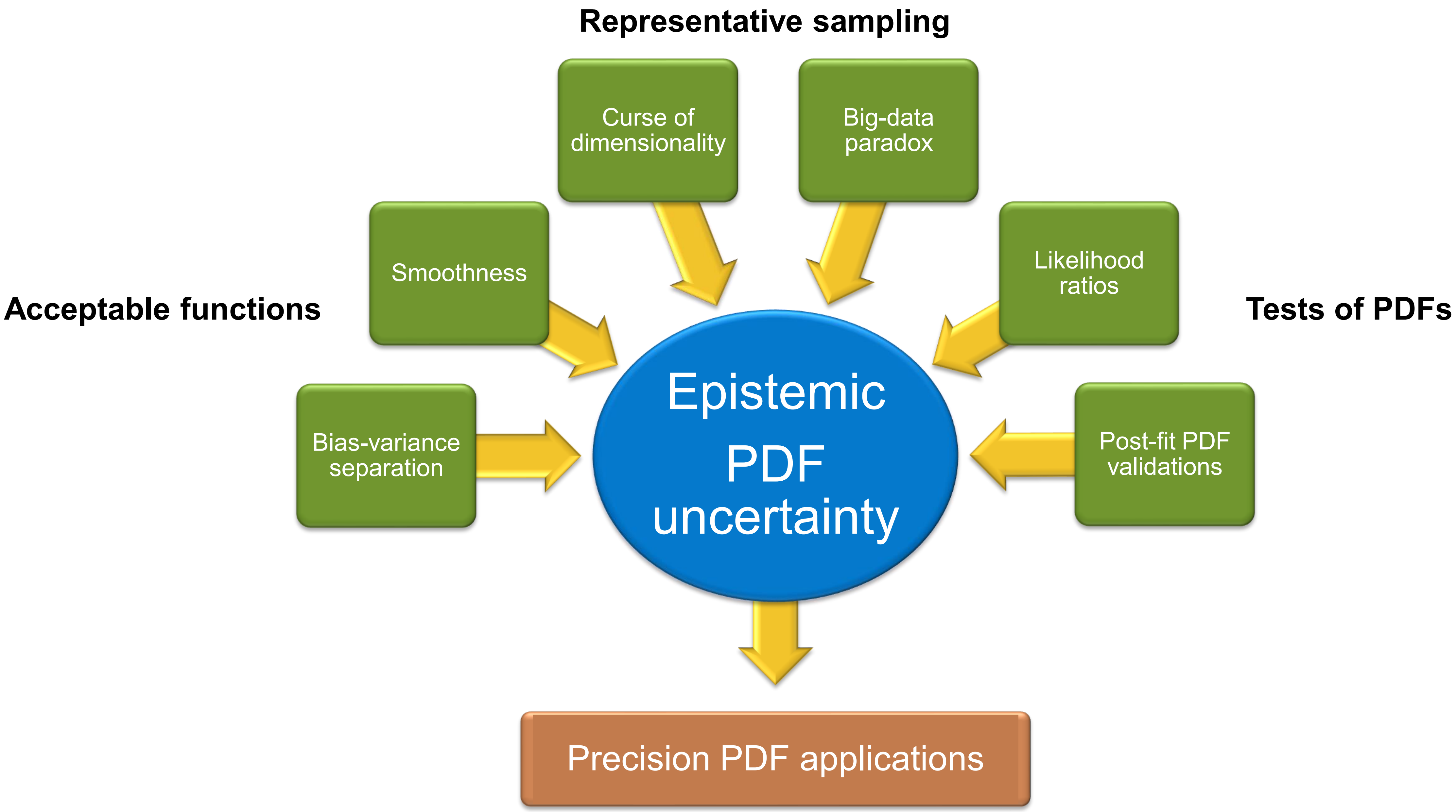
PDF uncertainty: information theory classification

Malinin, Gales, 2018
Courtoy, PN, et al, 2311.08447, 2205.10444
Hobbs, Kriesten, Gomprecht, 2023-24

Aleatory (dicey) uncertainty: statistical, propagated from experiments, reduced by increasing data size

Epistemic uncertainty: due to lack of knowledge, bias

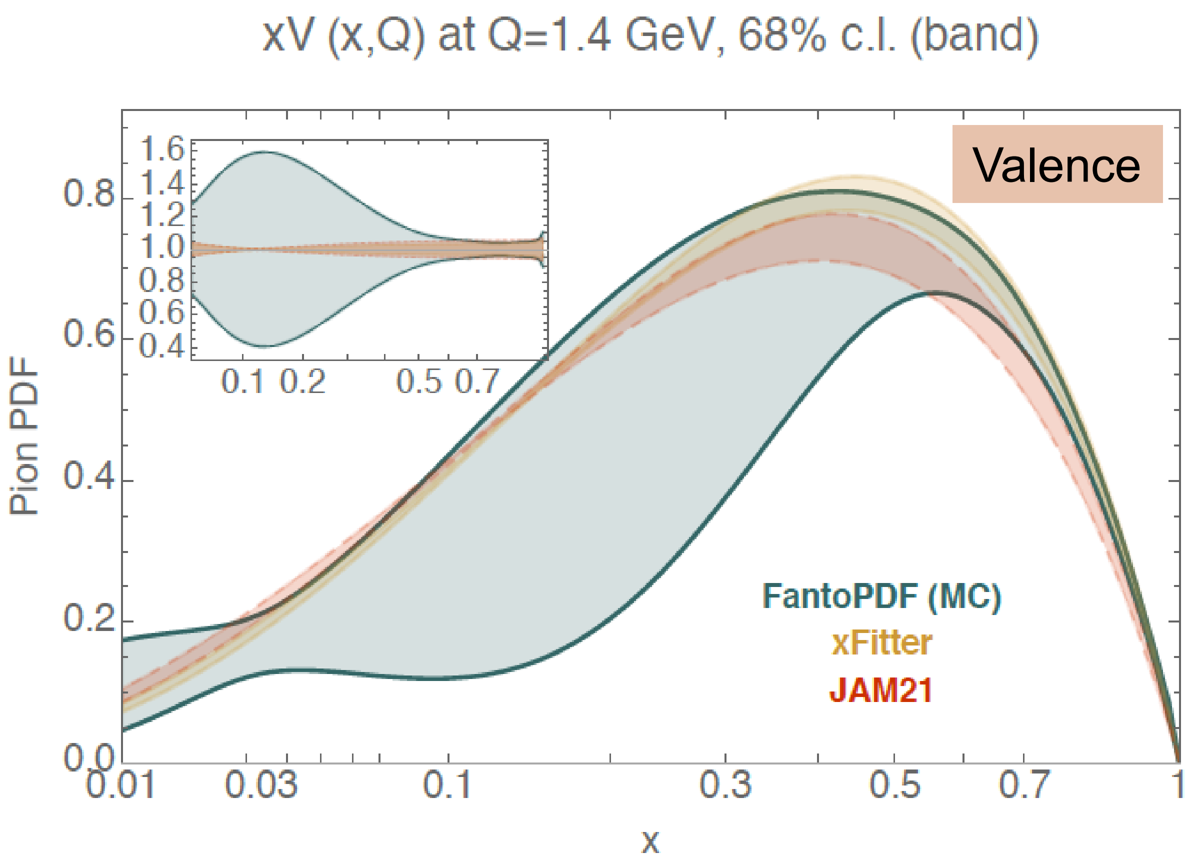
- model uncertainty: reduced by improving the model
- distributional uncertainty: reduced by representative sampling


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

Example of parametrization uncertainty: pion valence PDFs at NLO

L. Kotz, arXiv:2309.00152, arXiv:2311.08447

$\pi^{\pm}$



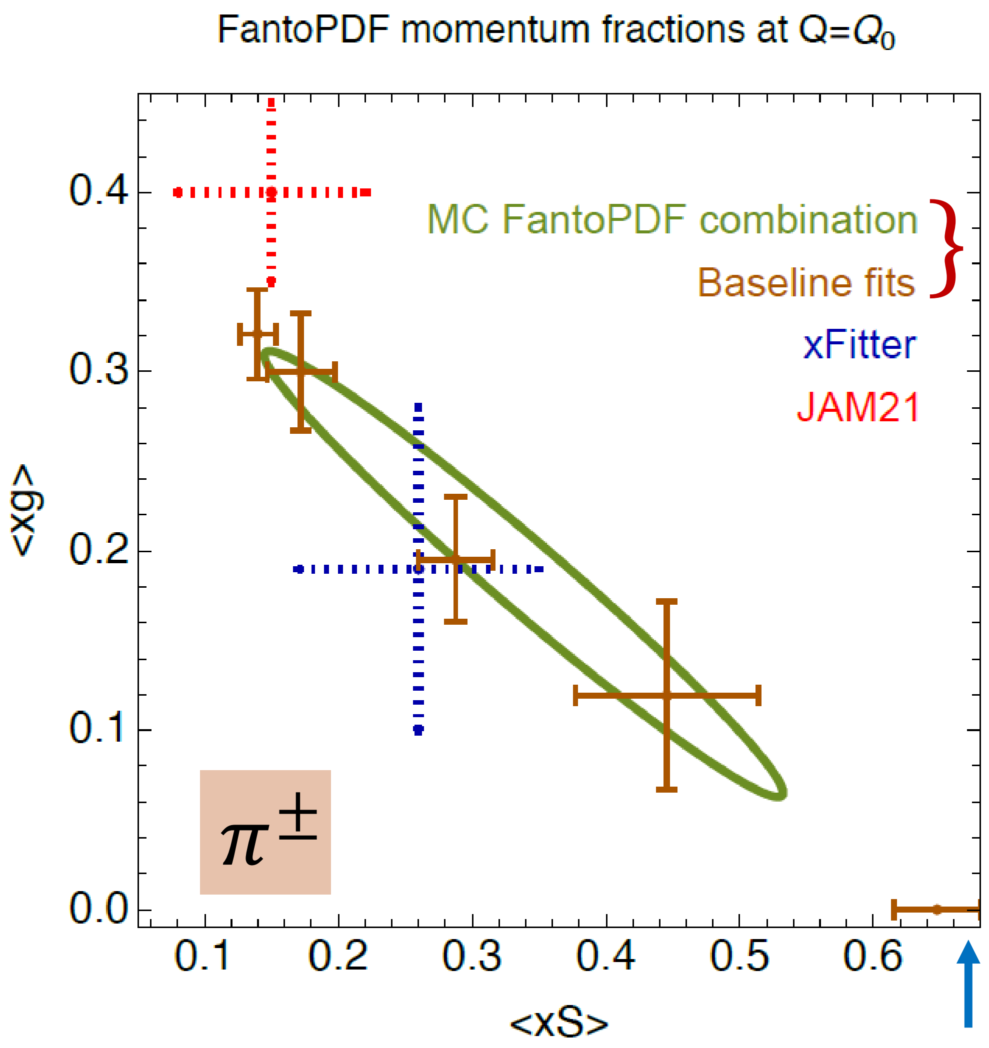
Narrow bands: xFitter'20 and JAM'21 fits with fixed PDF parametrization forms

Fanto1.0 PDF band: fit to the xFitter data based on ~ 100 alternative parametrization forms with the same or better χ^2 as in the 2020 xFitter study *[Novikov et al., arXiv:2002.02902]*

Fanto1.0 PDF parametrizations are constructed from **Bézier curves (universal polynomial approximators)** implemented in a public C++ program **Fantômas** and combined with the **METAPDF** *[arXiv:1401.00013]* technique

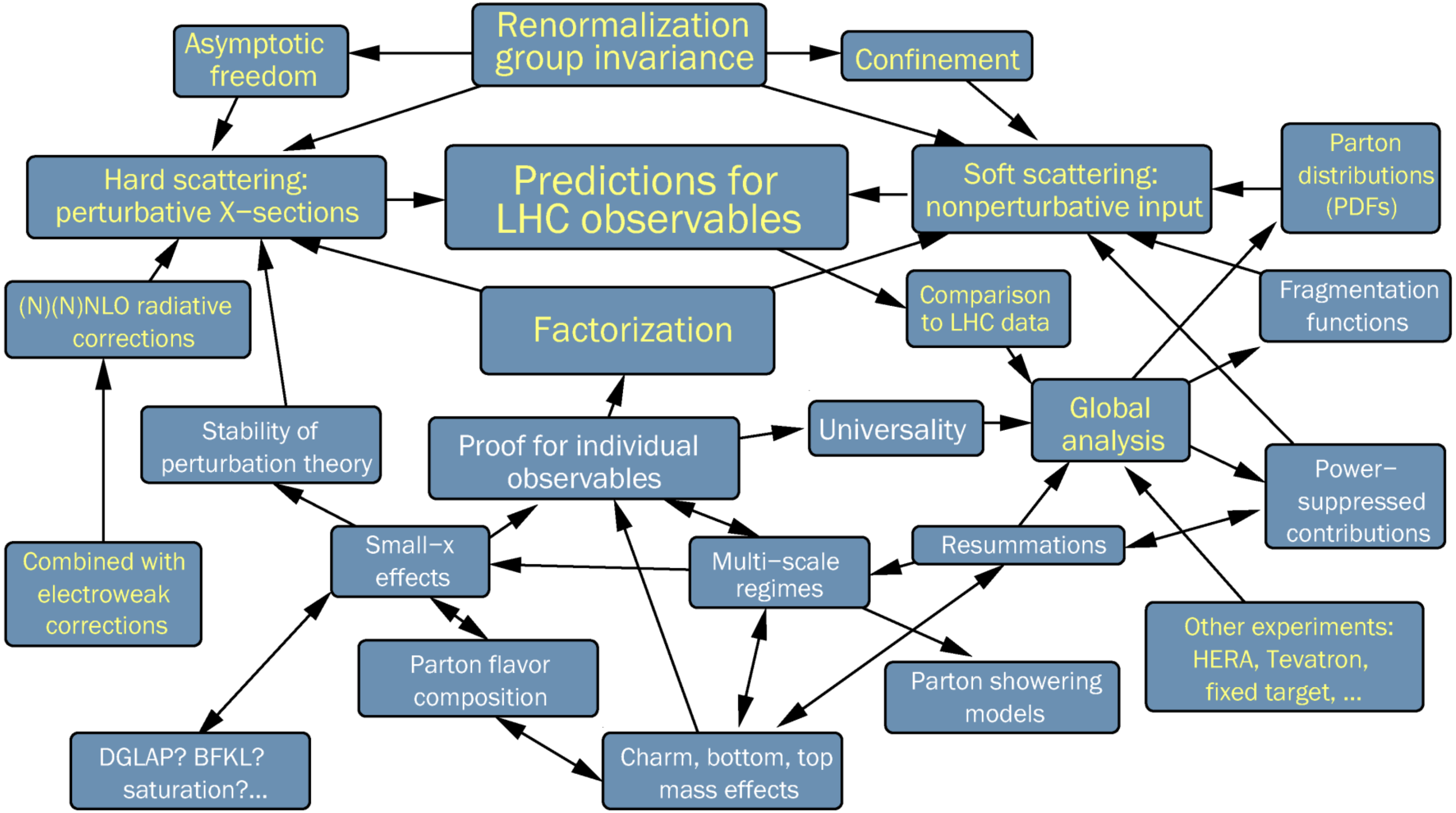
Parametrization uncertainty enlarges the error bands in regions with limited data

Fantômas pion PDFs: sea and gluon momentum fractions



a zero-momentum gluon is experimentally allowed

Parametrization uncertainty enlarges the error bands in regions with limited data



No analysis is an island *entire of itself*

- Accuracy is determined both by the individual calculation and its ambient connections
- **Aleatory** and **epistemic** uncertainties both play a role

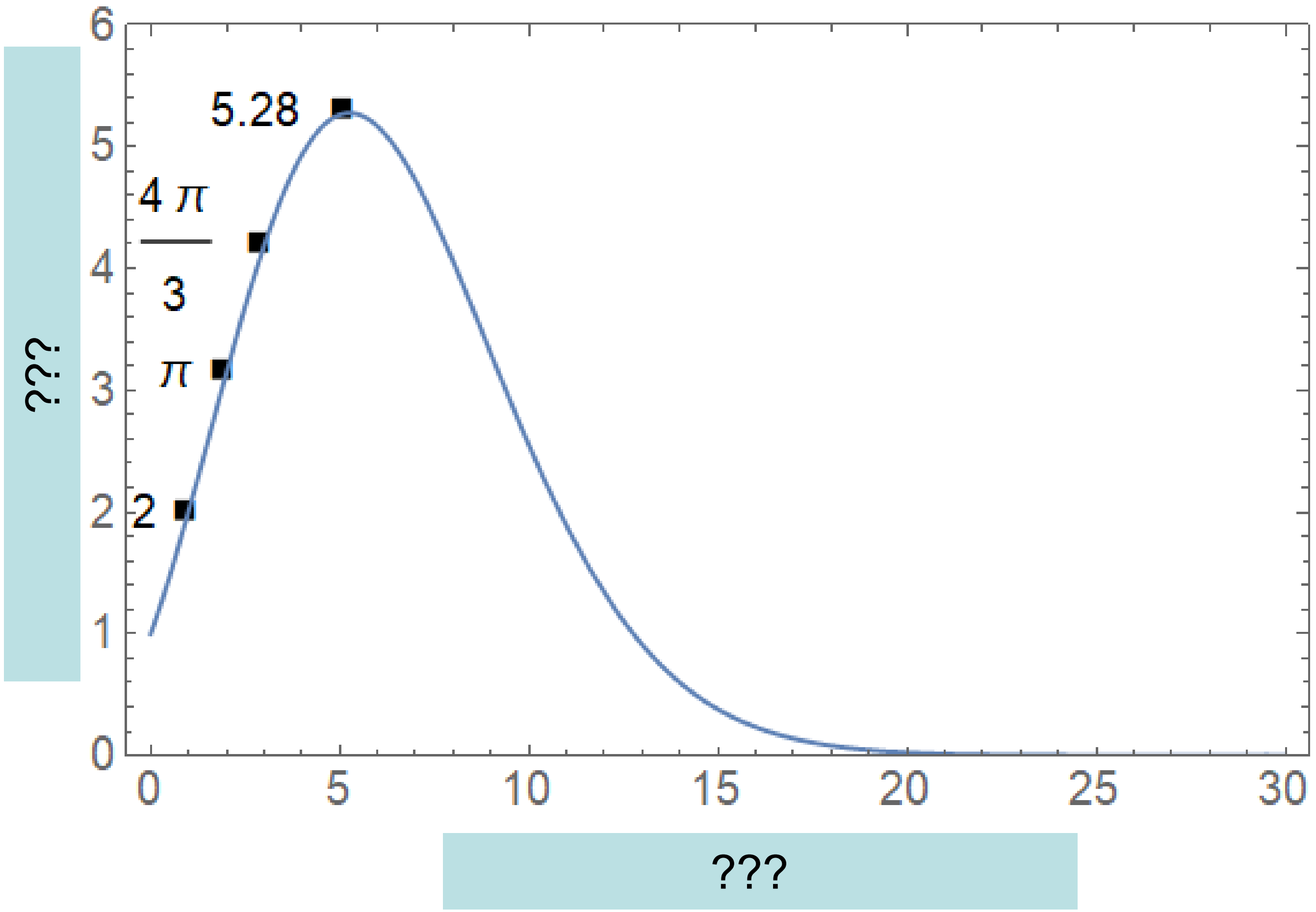
Warning! Statistics with >20 degrees of freedom disagrees with many common expectations

⇒ Estimating epistemic uncertainty is difficult

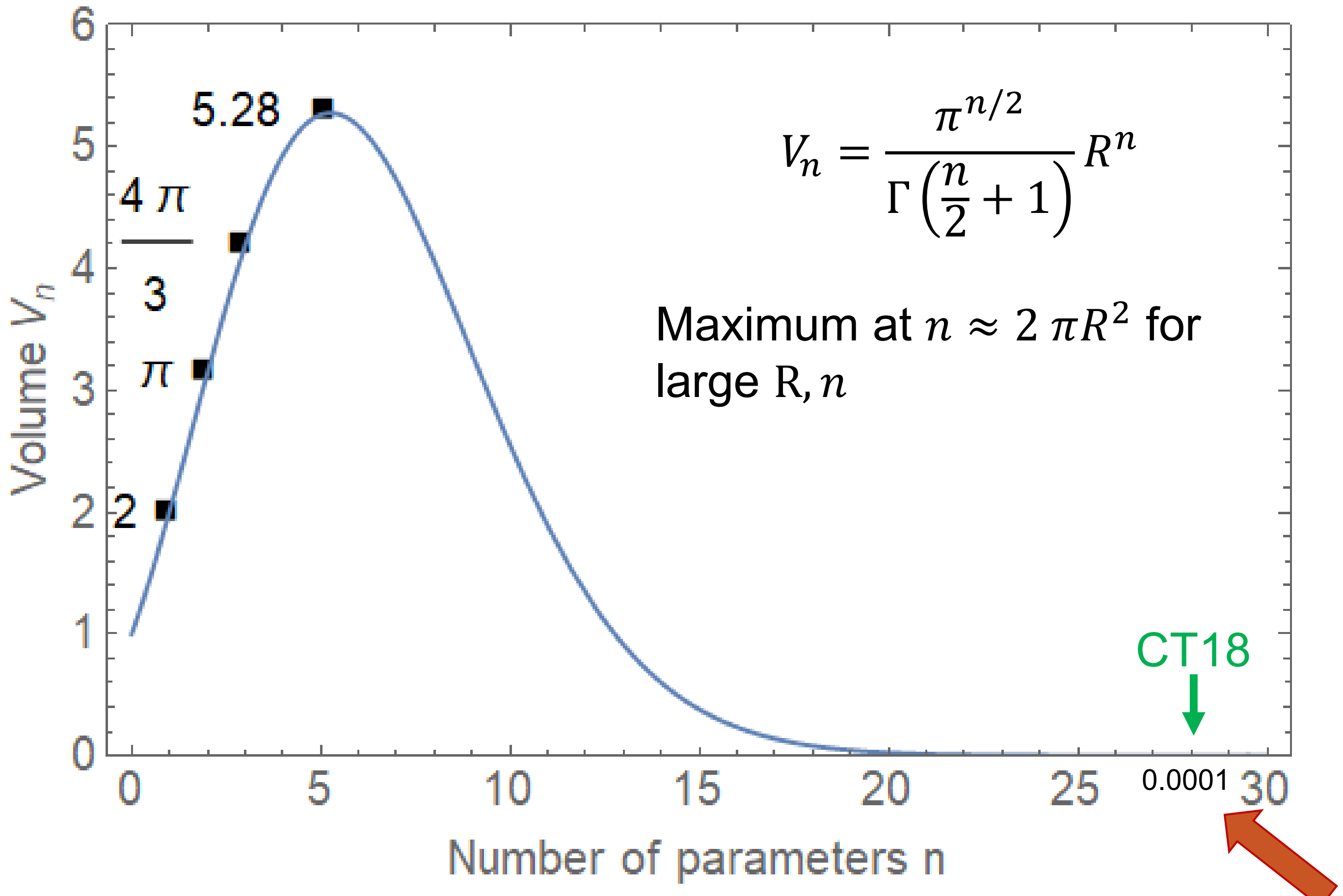
- Counterintuitive geometry of multidimensional functions
- Possible violations of the central limit theorem and other standard results
- Inefficient sampling of multidimensional probabilities

⇒ Implications for PDF fits, uncertainty quantification with AI/ML, replicability of results, …

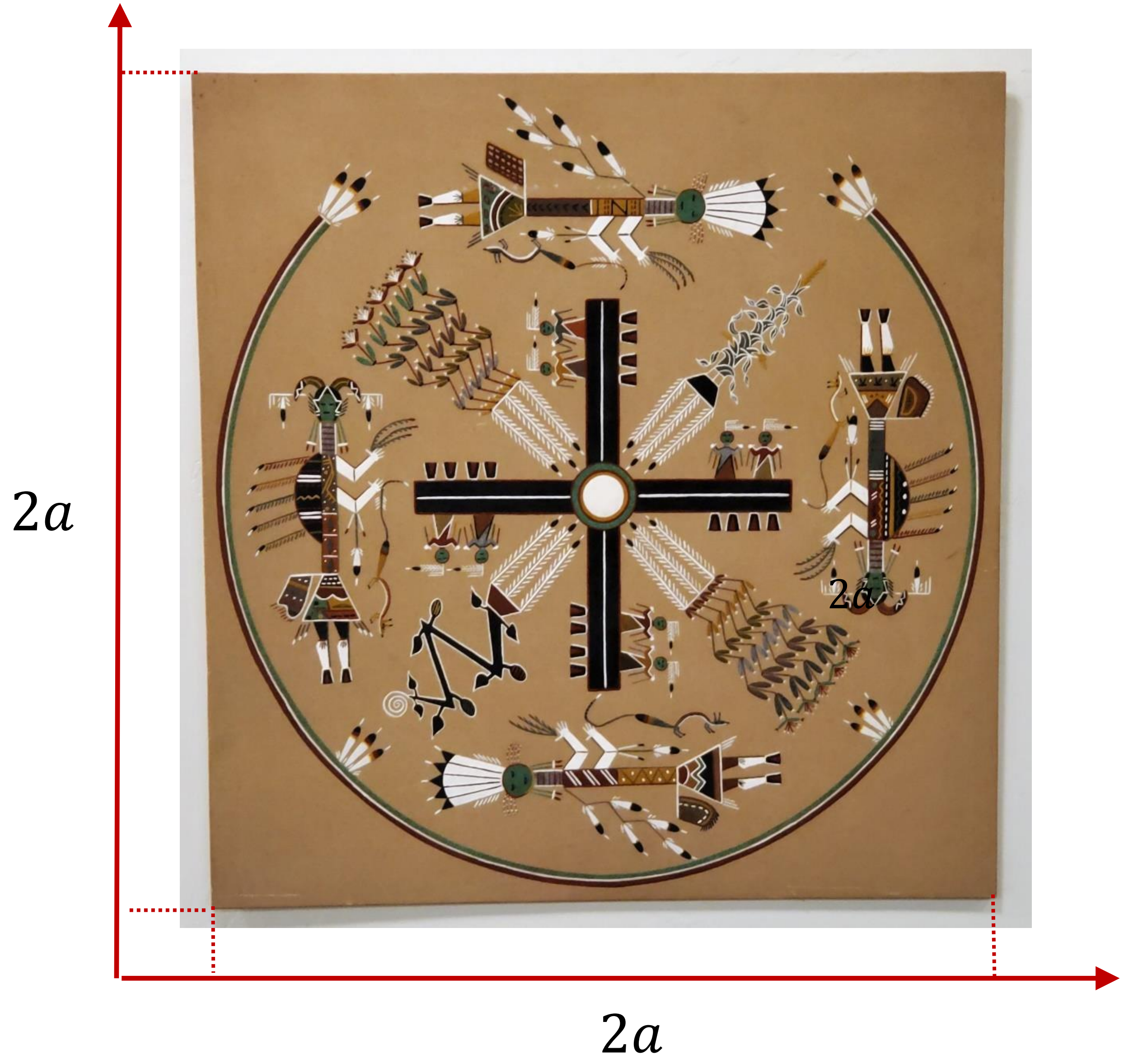
Think-pair share (1minute). Label the x axis and y axis



Think-pair share (1minute). Label the x axis and y axis



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

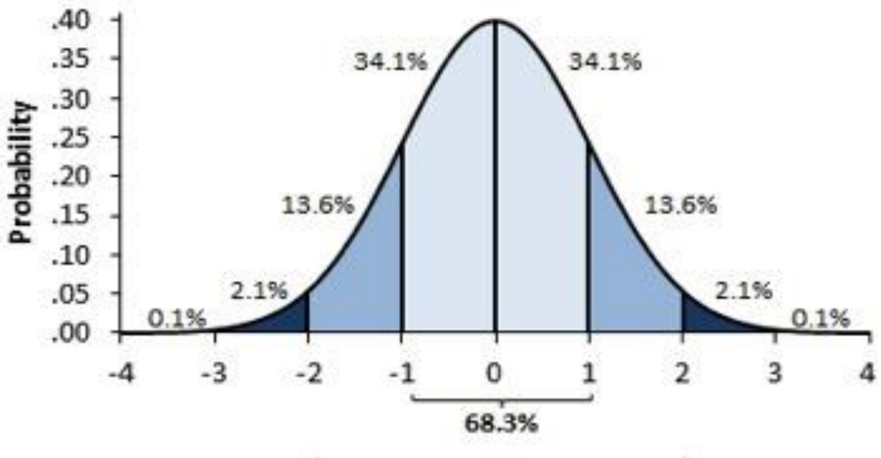
$$\frac{V_{sphere}}{V_{cube}} \approx \frac{0.0009}{2^{25}}$$

$$\approx \mathbf{3 \cdot 10^{-11}}$$

Image: sand painting, SMU-in-Taos

An n-dimensional **standard normal** distribution

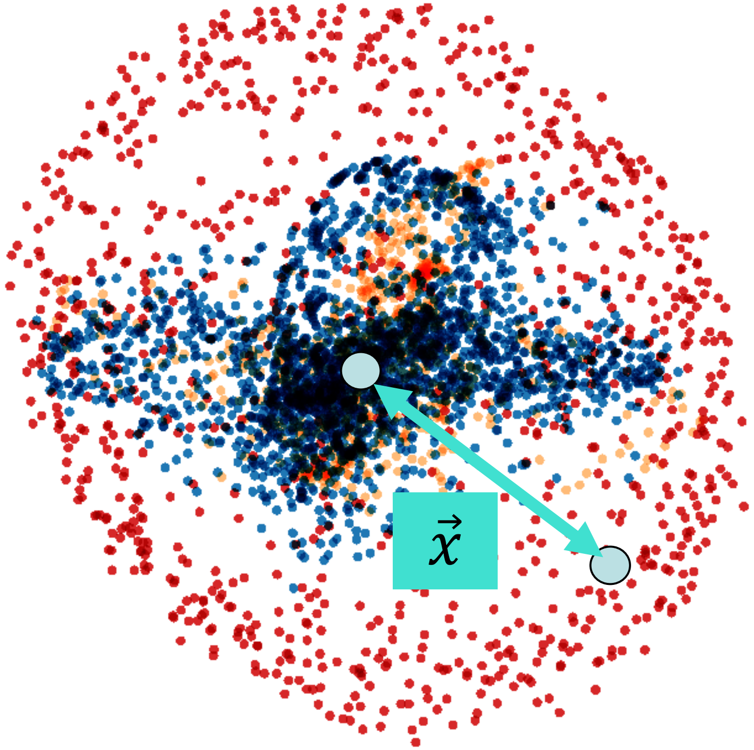
$$P(\vec{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\vec{x}^2}{2}\right)$$



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

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

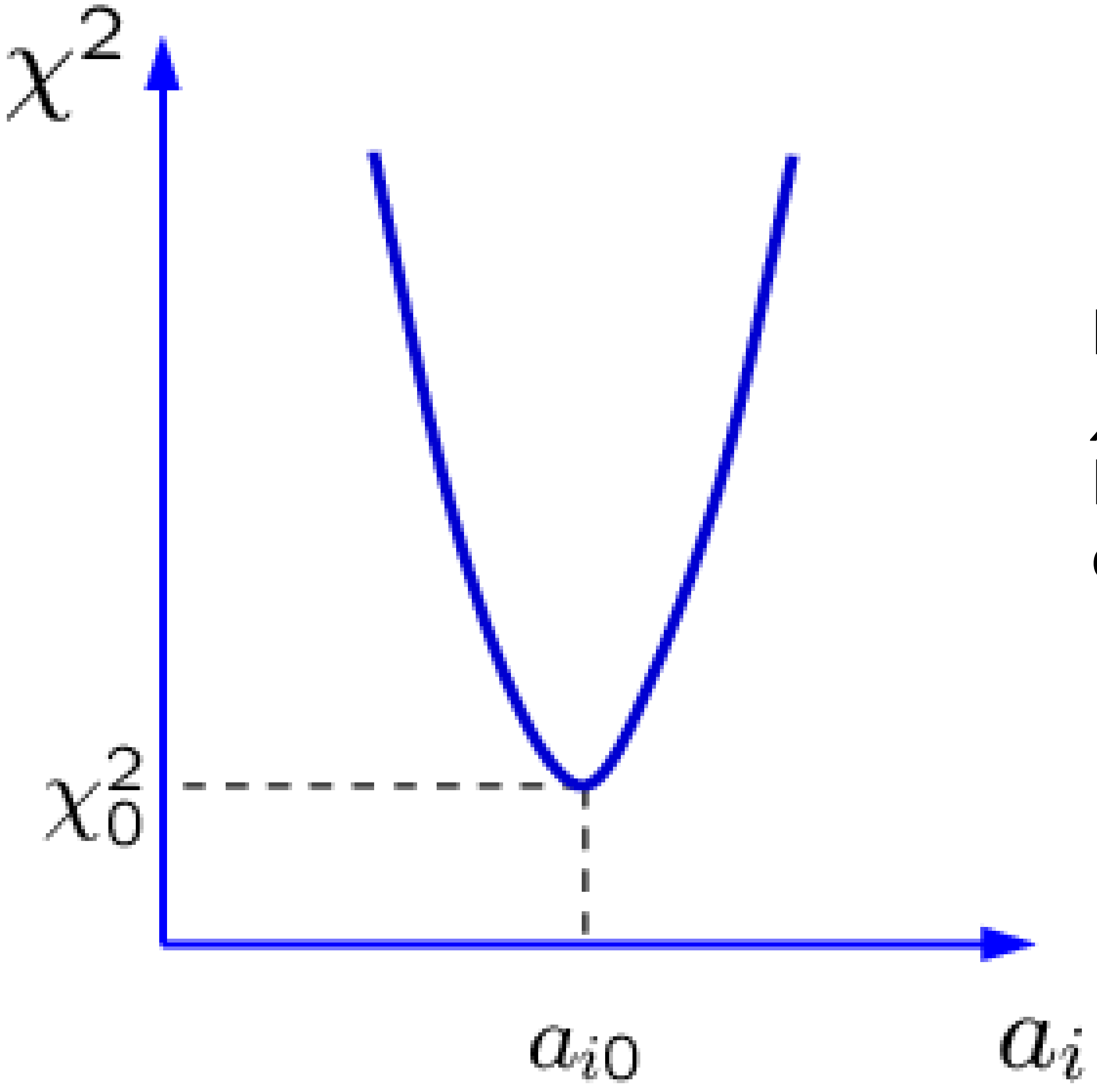
$\sqrt{n} \approx 5$ for $n = 25$

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

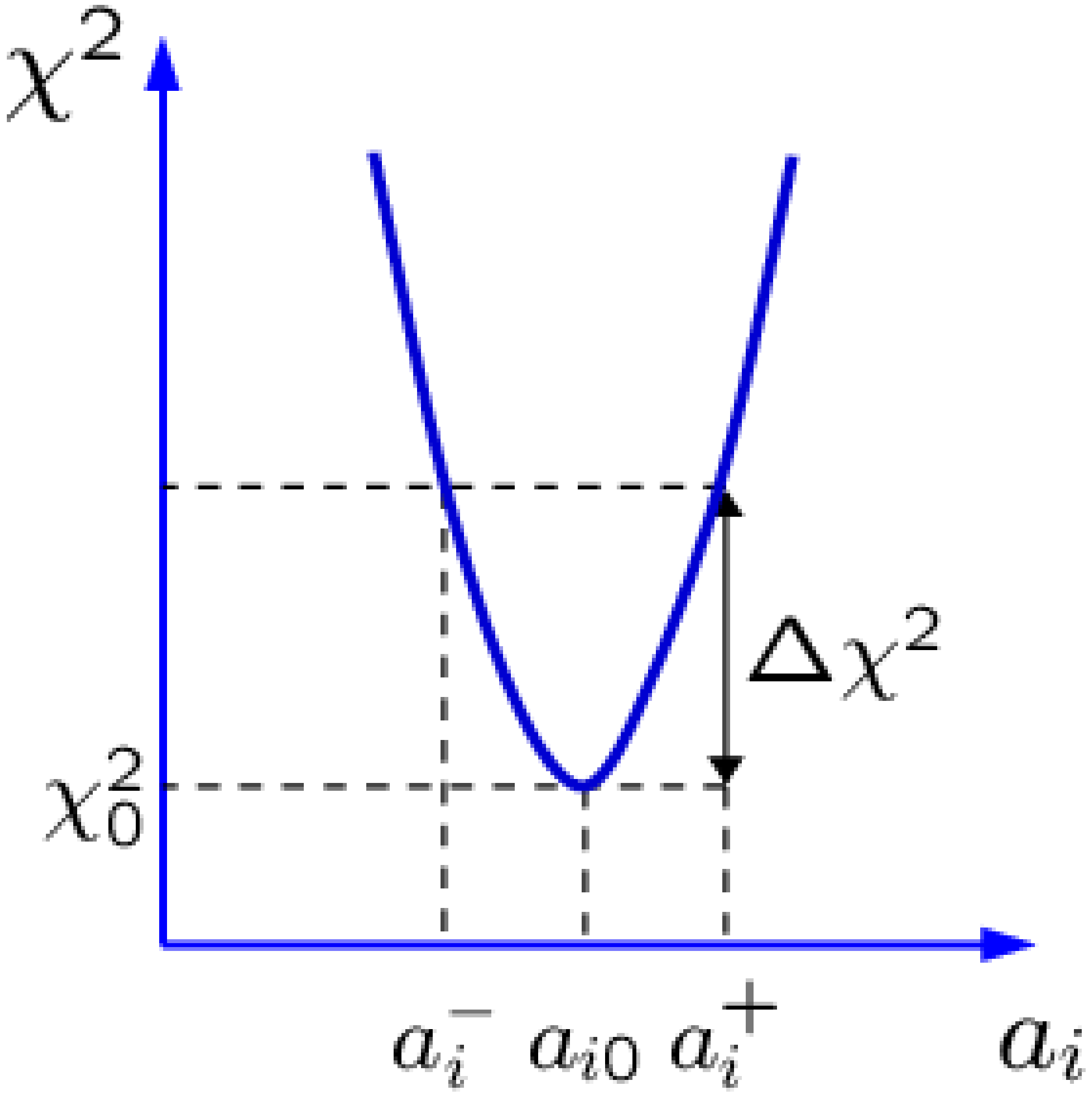
Quantification of PDF uncertainty in the Hessian formalism

Multi-dimensional error analysis



Minimization of a log-likelihood function $\chi^2$with respect to ~ 40 theoretical (mostly PDF) parameters $\{a_i\}$ and > 2000 experimental systematical parameters

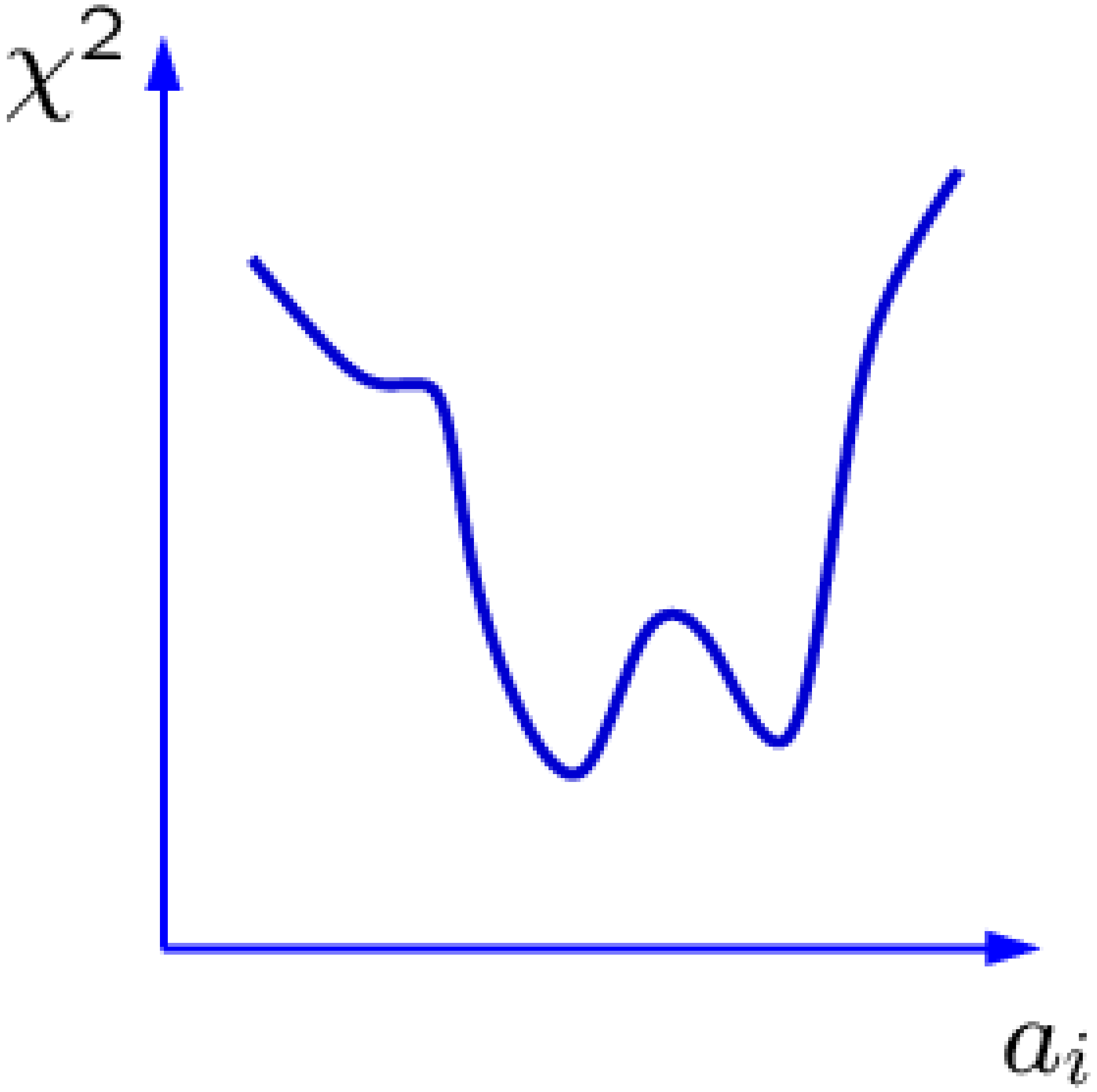
Multi-dimensional error analysis



- Establish a confidence region for $\{a_i\}$ for a given tolerated increase in χ^2
- In the ideal case of perfectly compatible Gaussian errors, 68% c.l. on a physical observable X corresponds to $\Delta\chi^2 = 1$ independently of the number N of PDF parameters

See, e.g., P. Bevington, K. Robinson, Data analysis and error reduction for the physical sciences

Multi-dimensional error analysis



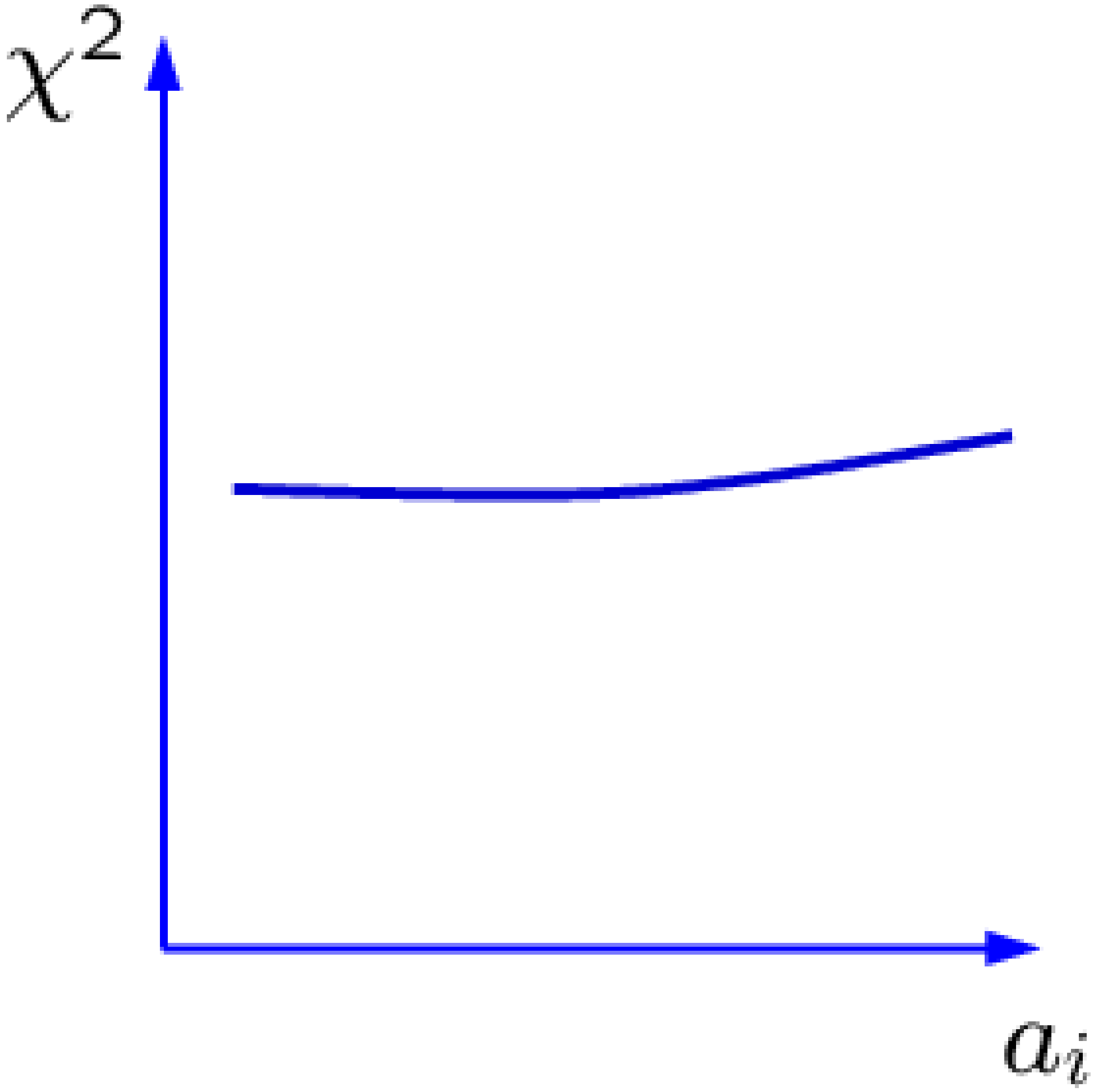
Pitfalls to avoid

``Landscape''

- disagreements between the experiments

In the worst situation, significant disagreements between M experimental data sets can produce up to $N \sim M!$ possible solutions for PDF's, with $N \sim 10^{500}$ reached for ``only'' about 200 data sets

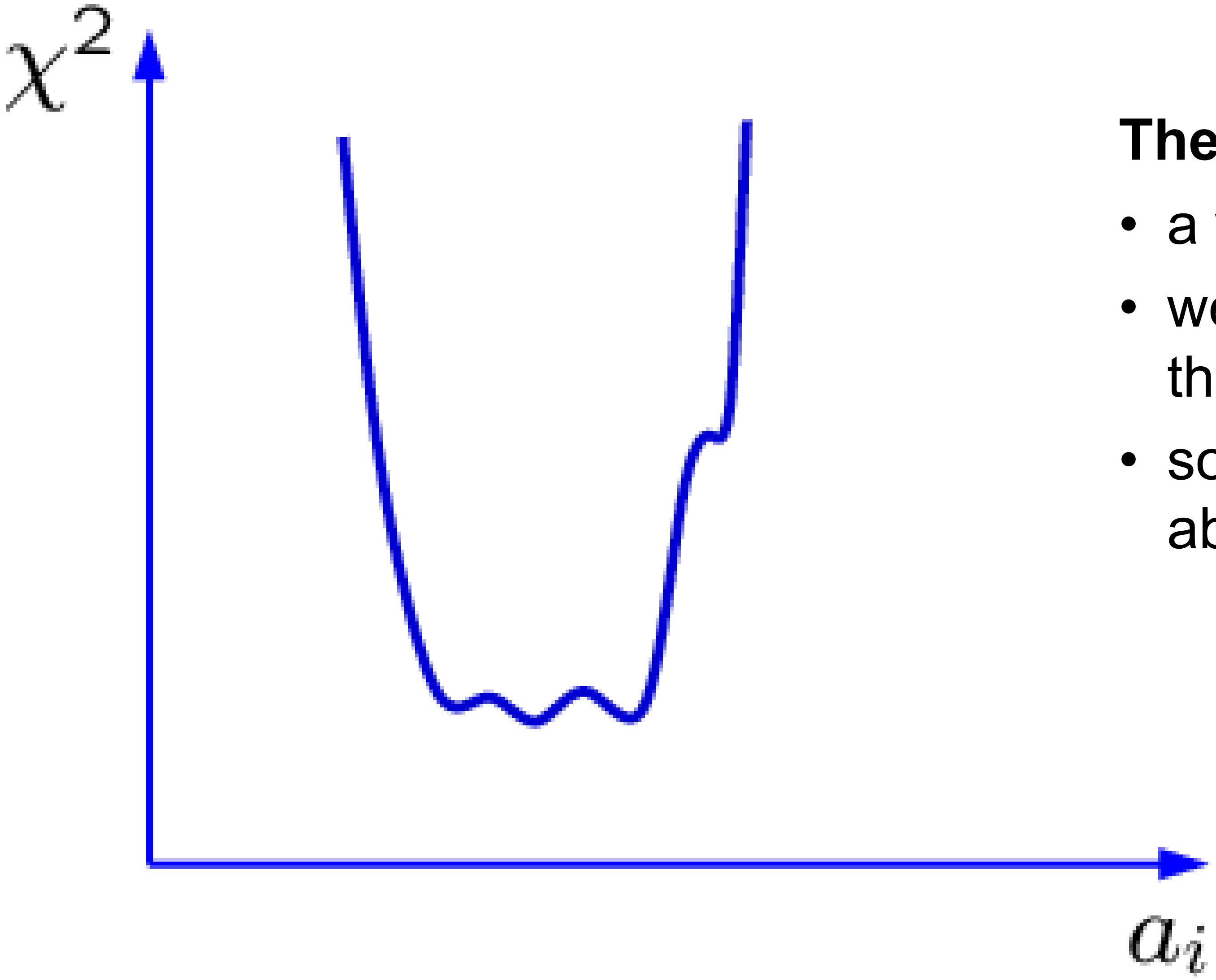
Multi-dimensional error analysis



Pitfalls to avoid

- Flat directions = unconstrained combinations of PDF parameters
- dependence on free theoretical parameters, especially in the PDF parametrization
- impossible to derive reliable PDF error sets

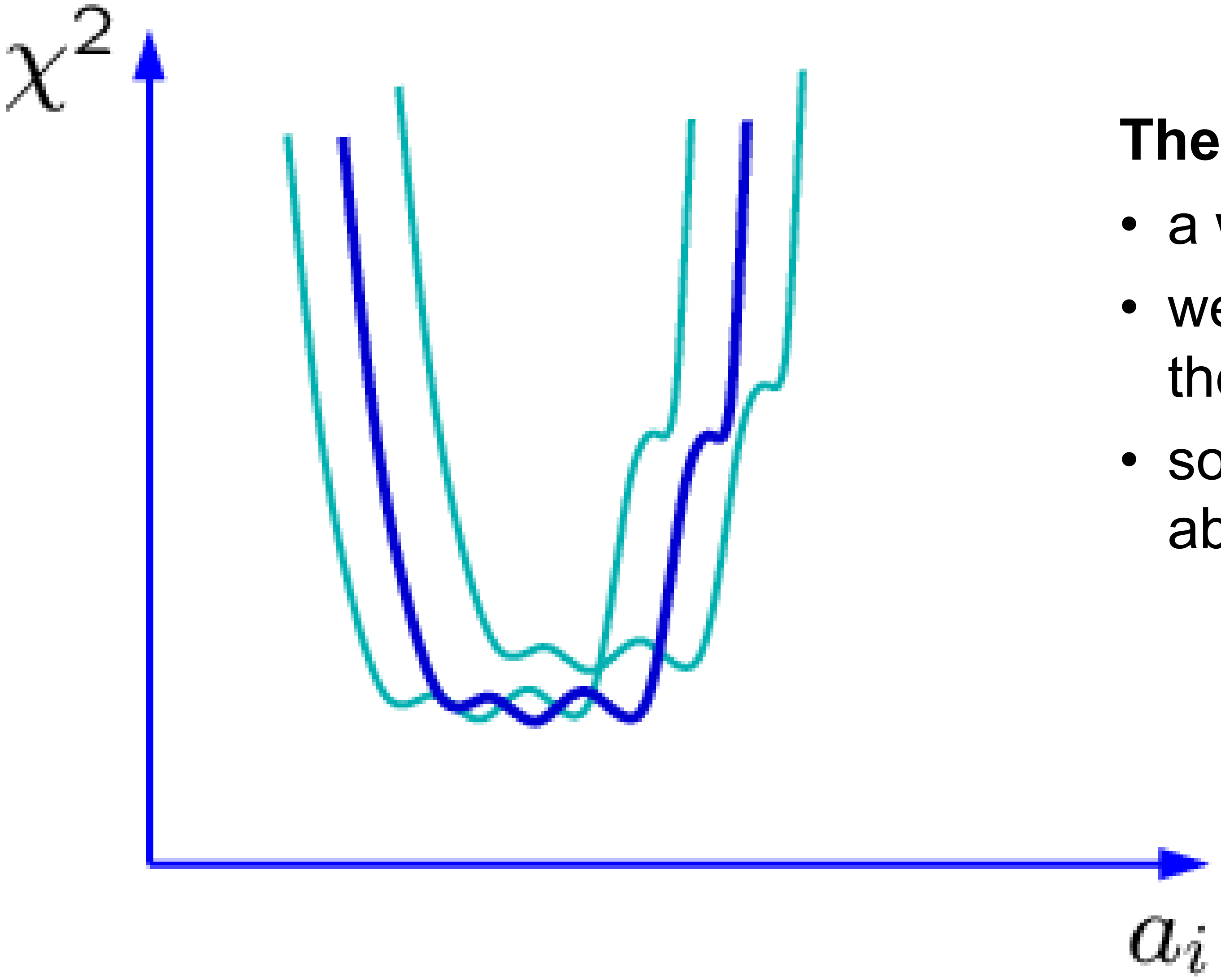
Multi-dimensional error analysis



The actual χ^2 function shows

- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

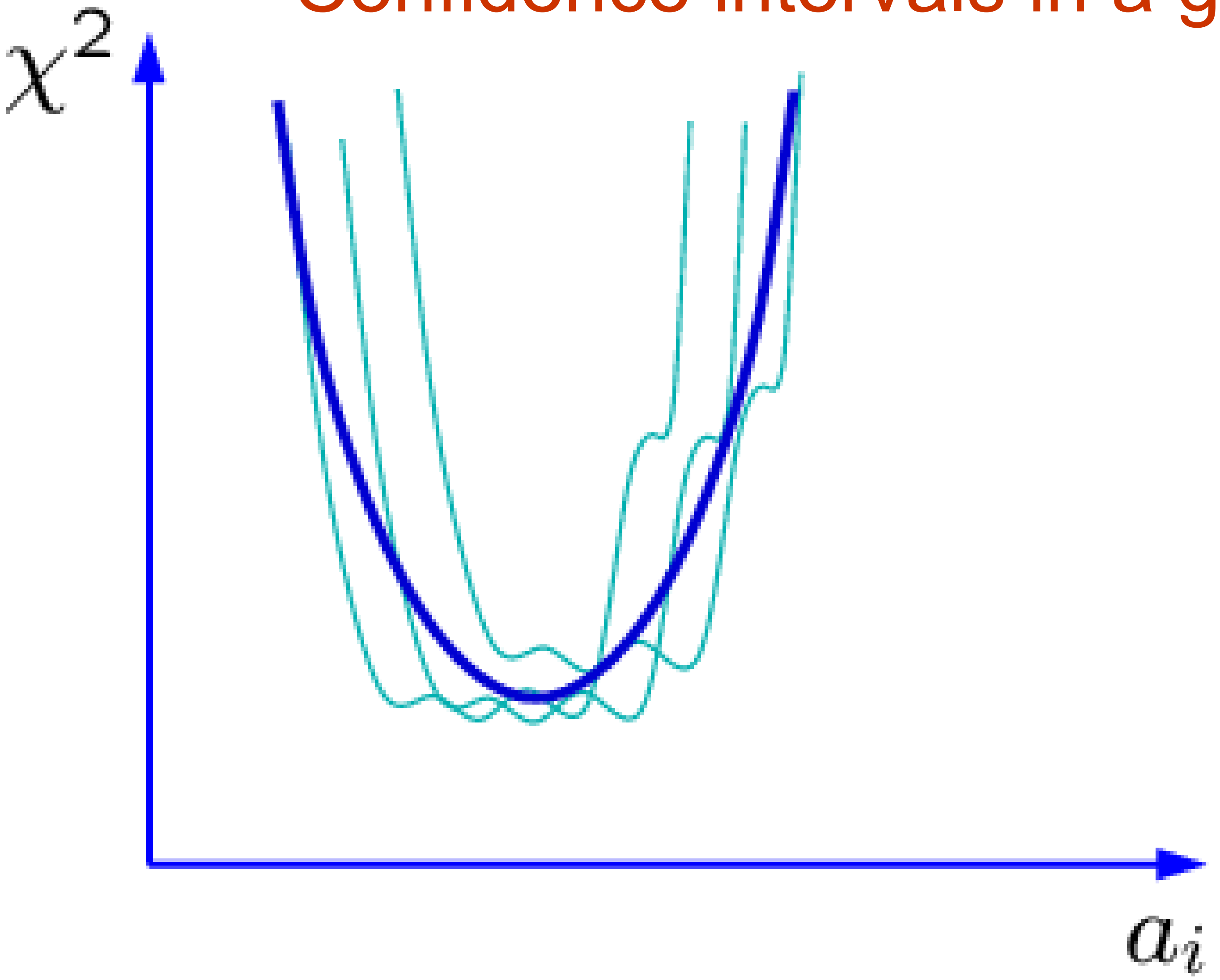
Multi-dimensional error analysis



The actual χ^2 function shows

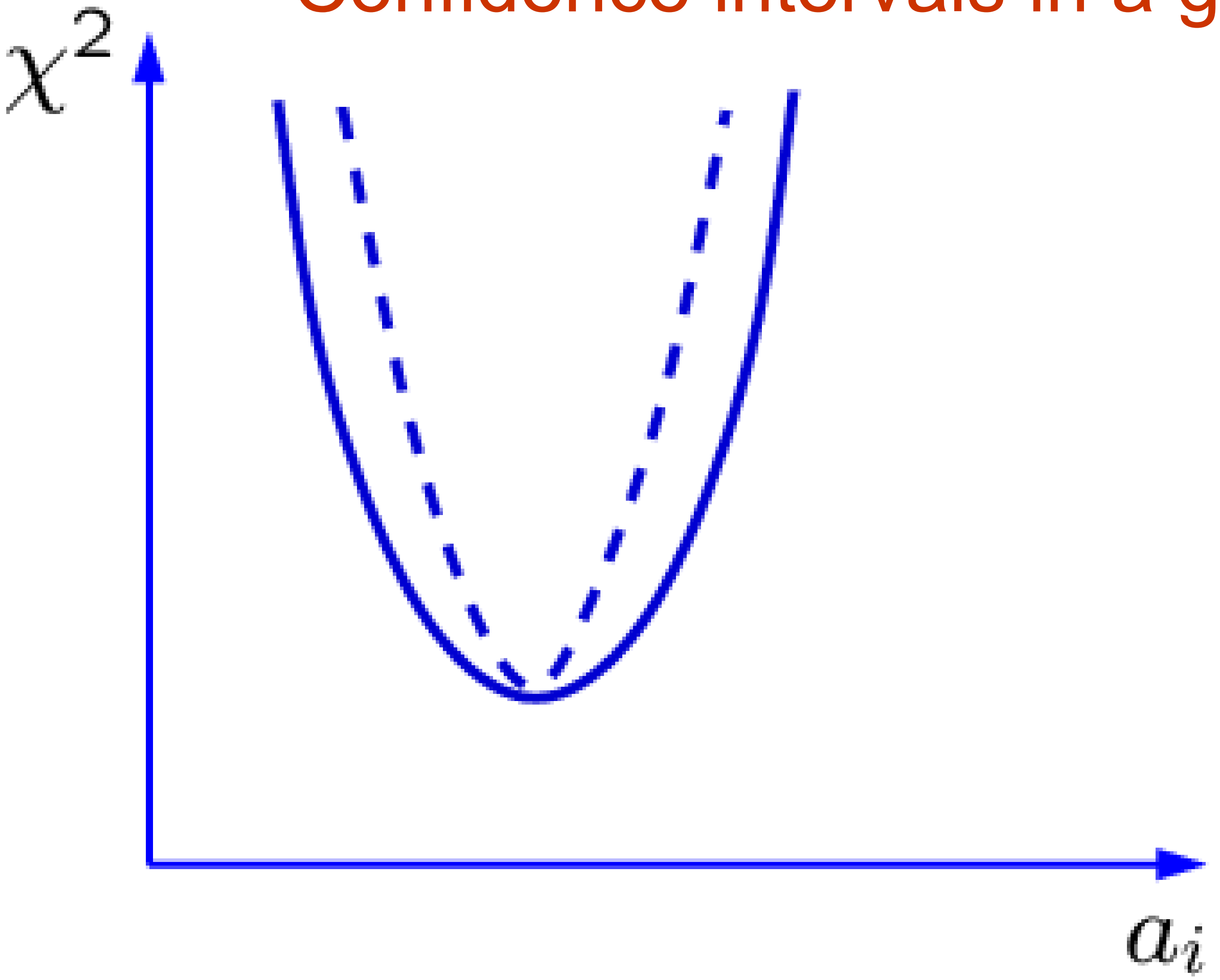
- a well pronounced global minimum χ_0^2
- weak tensions between data sets in the vicinity of χ_0^2 (mini-landscape)
- some dependence on assumptions about flat directions

Confidence intervals in a global PDF analysis



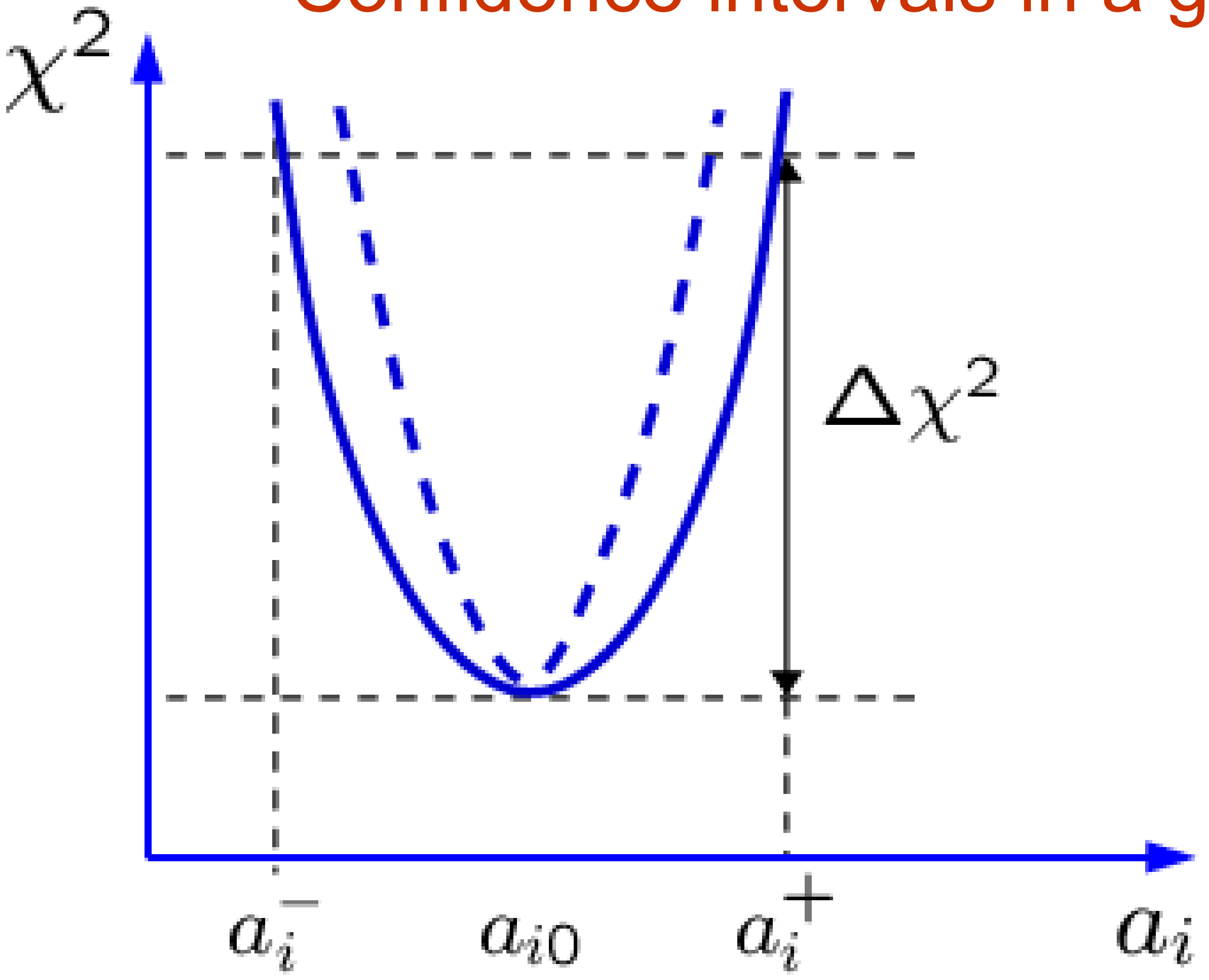
The likelihood is approximately described by a quadratic χ^2 with a revised tolerance condition $\Delta\chi^2 \le T^2$

Confidence intervals in a global PDF analysis



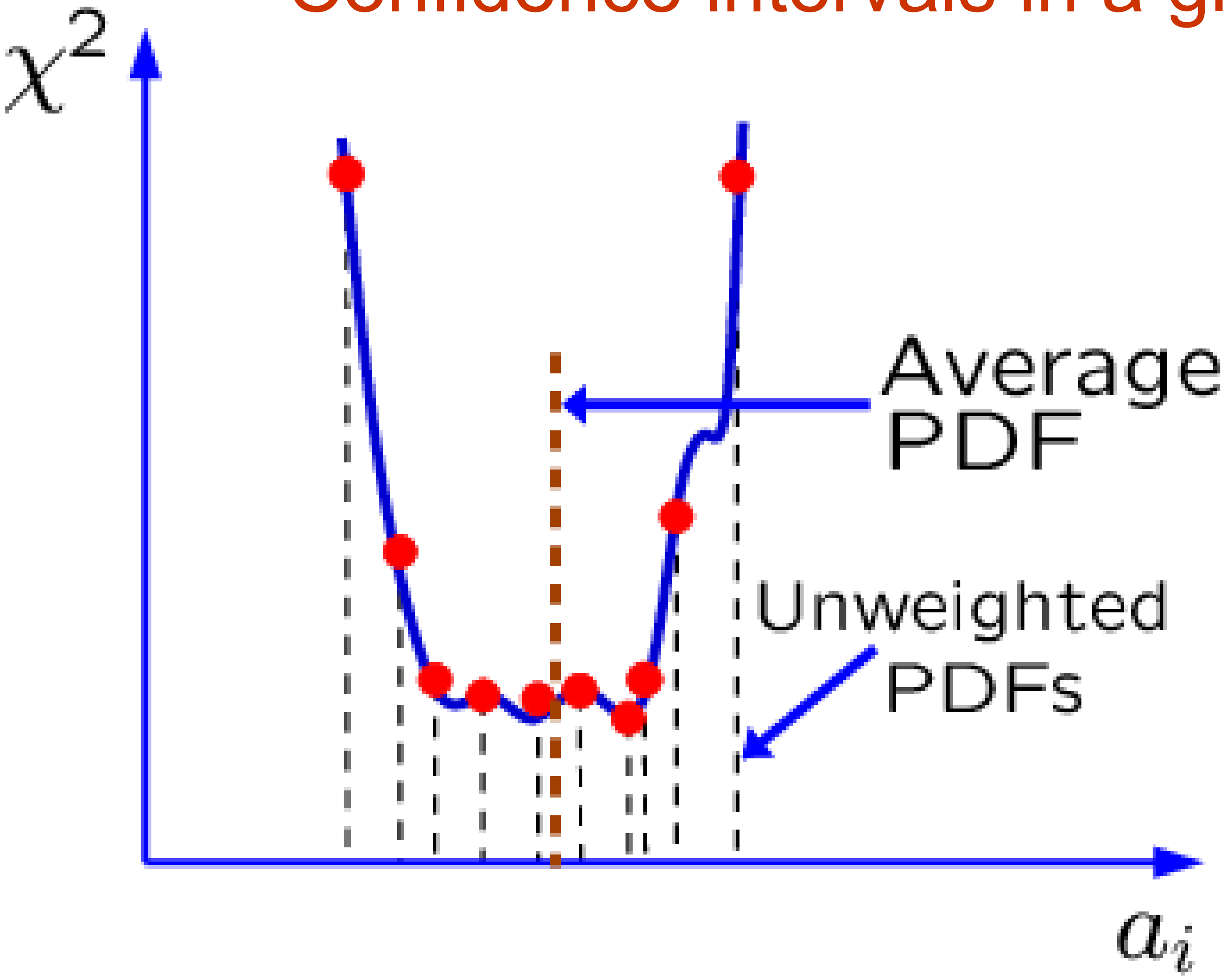
The likelihood is approximately described by a quadratic χ^2 with a revised tolerance condition $\Delta\chi^2 \leq T^2$

Confidence intervals in a global PDF analysis



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

Confidence intervals in a global PDF analysis

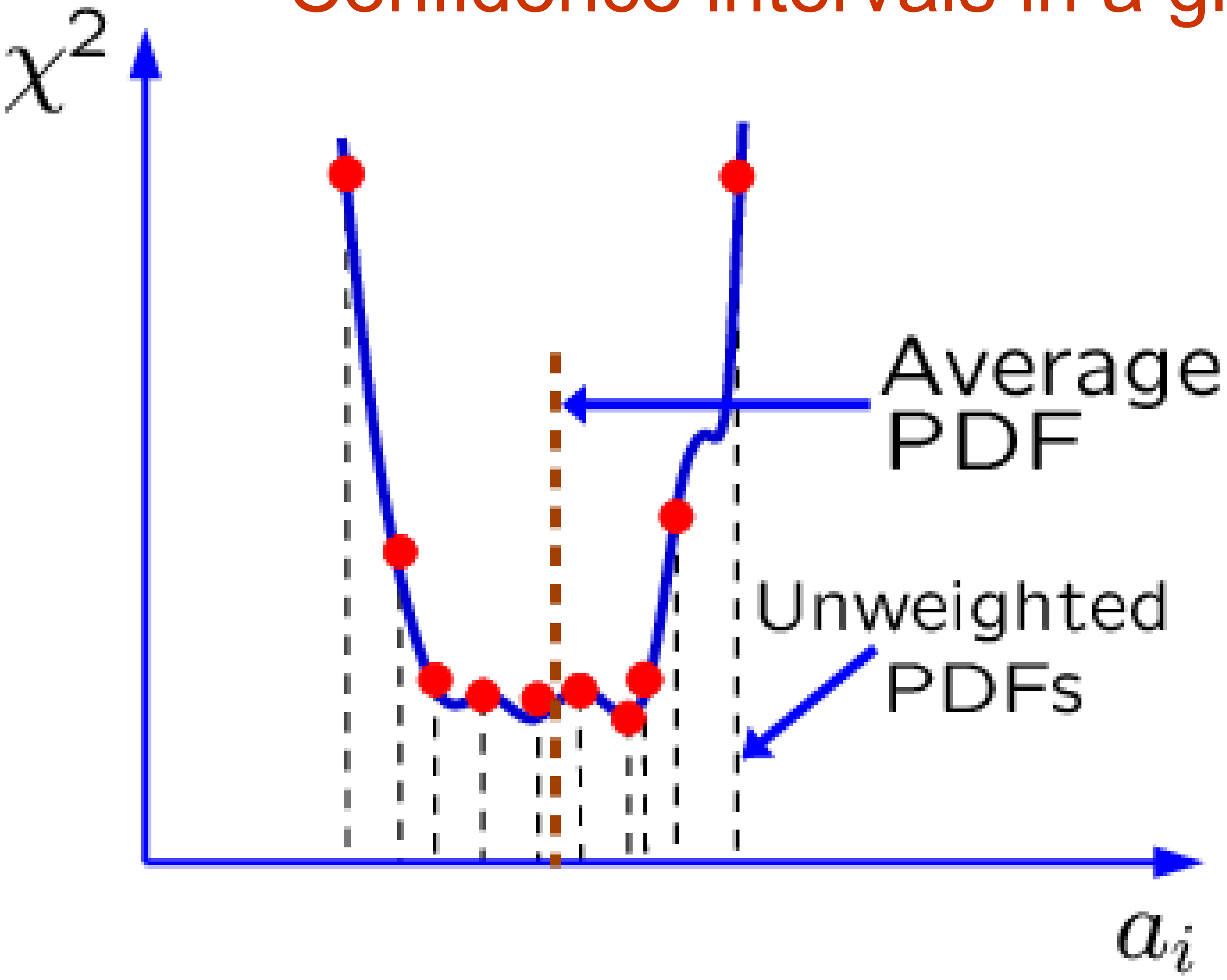


Monte-Carlo sampling of the PDF parameter space

🙂 A very general approach that

- realizes stochastic sampling of the probability distribution [*Alekhin; Giele, Keller, Kosower; NNPDF, JAM, …]*
- can parametrize PDF's by flexible neural networks [NNPDF, …]
- does not rely on smoothness of χ^2 or Gaussian approximations

Confidence intervals in a global PDF analysis



Monte-Carlo sampling of the PDF parameter space

☹ ■ Most individual replicas are **very bad** fits at $\sim \sqrt{N_{par}}$ std. deviations away

■ A good fit is obtained by averaging over the replicas

[arXiv:1607.06066, 2205.10444]

Another question [1 minute]

Given a QCD observable O, can you tell which parton flavors drive the PDF uncertainty on O? How?

Another question [1 minute]

Given a QCD observable O, can you tell which parton flavors drive the PDF uncertainty on O? How?

Let's use **Hessian correlations and sensitivities** to answer this question

A fast test of experimental constraints using L_2 sensitivity

T.J. Hobbs et al., 1904.00022

- Requires:
  - Hessian or Monte Carlo error PDFs
  - χ^2 values for fitted or envisioned experiments for each error PDF

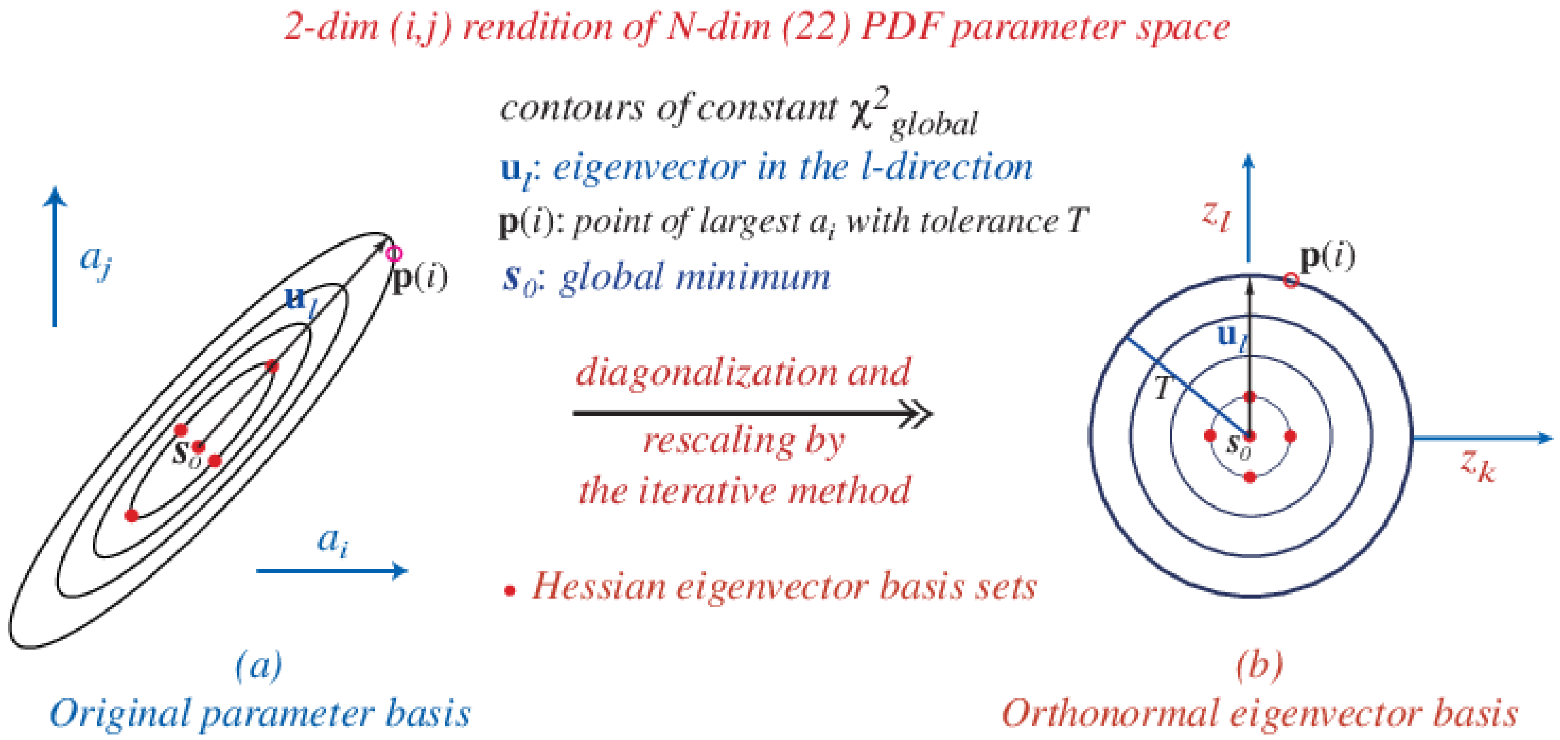
- Quantifies:
  - strengths of constraints from individual experiments on given PDFs
  - agreement among the experiments (universality of the best-fit PDFs)
  - sensitivity of processes not included in the global fit

L_2 sensitivities facilitate comparisons among Hessian fits: CJ and CT **[2102.01107]**; ATLAS, CT, and MSHT **[2306.0391]**

If data point residuals for each error PDF set are also provided, a related $\boldsymbol{L_1}$ **sensitivity [B. T. Wang et al., 1803.02777]** can be computed to visualize kinematic distributions of experimental constraints in the x-Q plane

Tolerance hypersphere in the PDF space

Hessian method: Pumplin et al., 2001



A hyperellipse $\Delta\chi^2 \leq T^2$ in space of N physical PDF parameters $\{a_i\}$ is mapped onto a filled hypersphere of radius T in space of N orthonormal PDF parameters $\{z_i\}$

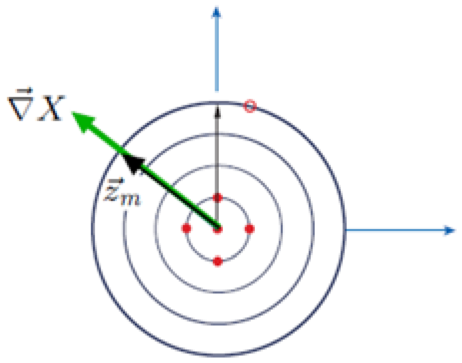
Tolerance hypersphere in the PDF space

2-dim (i,j) rendition of N-dim (26) PDF parameter space

A symmetric PDF error for a physical observable X is given by

$$\Delta X = \vec{\nabla} X \cdot \vec{z}_m = \left|\vec{\nabla} X\right|$$

$$= \frac{1}{2}\sqrt{\sum_{i=1}^{N}\left(X_i^{(+)} - X_i^{(-)}\right)^2}$$



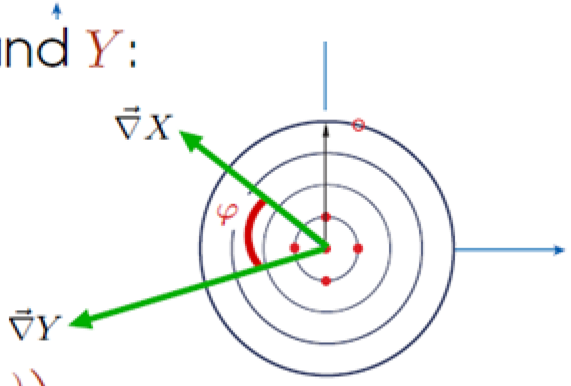
(b)
Orthonormal eigenvector basis

Correlation cosine for observables X and Y:

hep-ph/0110378

$$\cos\varphi = \frac{\vec{\nabla} X \cdot \vec{\nabla} Y}{\Delta X \Delta Y} =$$

$$\frac{1}{4\Delta X\,\Delta Y}\sum_{i=1}^{N}\left(X_i^{(+)} - X_i^{(-)}\right)\left(Y_i^{(+)} - Y_i^{(-)}\right)$$

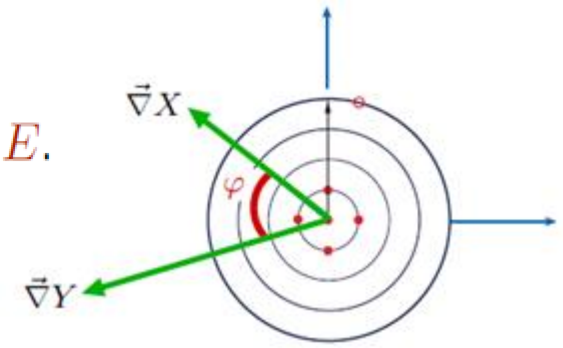


(b)
Orthonormal eigenvector basis

L_2 sensitivity, definition

$S_{f,L_2}(E)$ for experiment E is the estimated $\Delta\chi^2_E$ for this experiment when a PDF $f_a(x_i,Q_i)$ increases by the +68% c.l. Hessian PDF uncertainty

Take $X = f_a(x_i, Q_i)$ or $\sigma(f)$; $Y = \chi^2_E$ for experiment E.

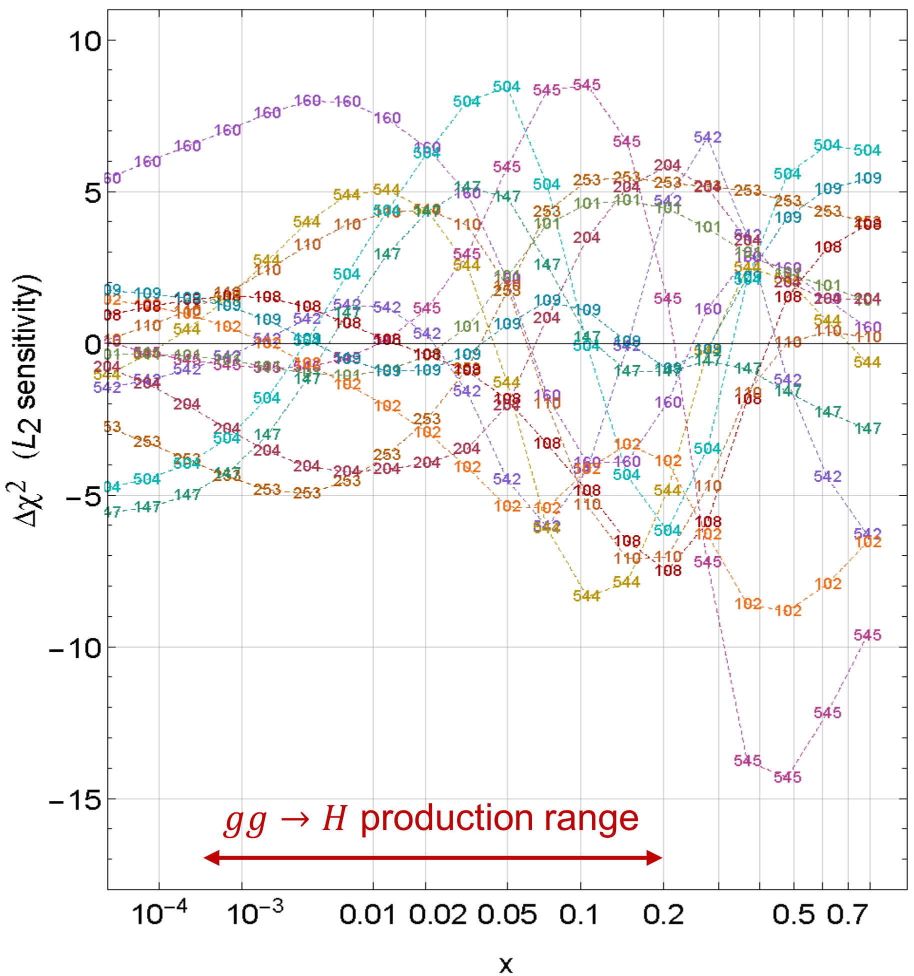


$$S_{f,L_2} \equiv \Delta Y(\vec{z}_{m,X}) = \vec{\nabla}Y \cdot \vec{z}_{m,X} = \vec{\nabla}Y \cdot \frac{\vec{\nabla}X}{|\vec{\nabla}X|} = \Delta Y \cos\varphi.$$

A fast version of the Lagrange Multiplier scan of χ^2_E along the direction of $f_a(x_i,Q_i)$!

Estimated χ^2 pulls from experiments

(L_2 sensitivity, T. J. Hobbs et al., arXiv:1904.00222)



CT18 NNLO, gluon at Q=100 GeV

15 core-minutes

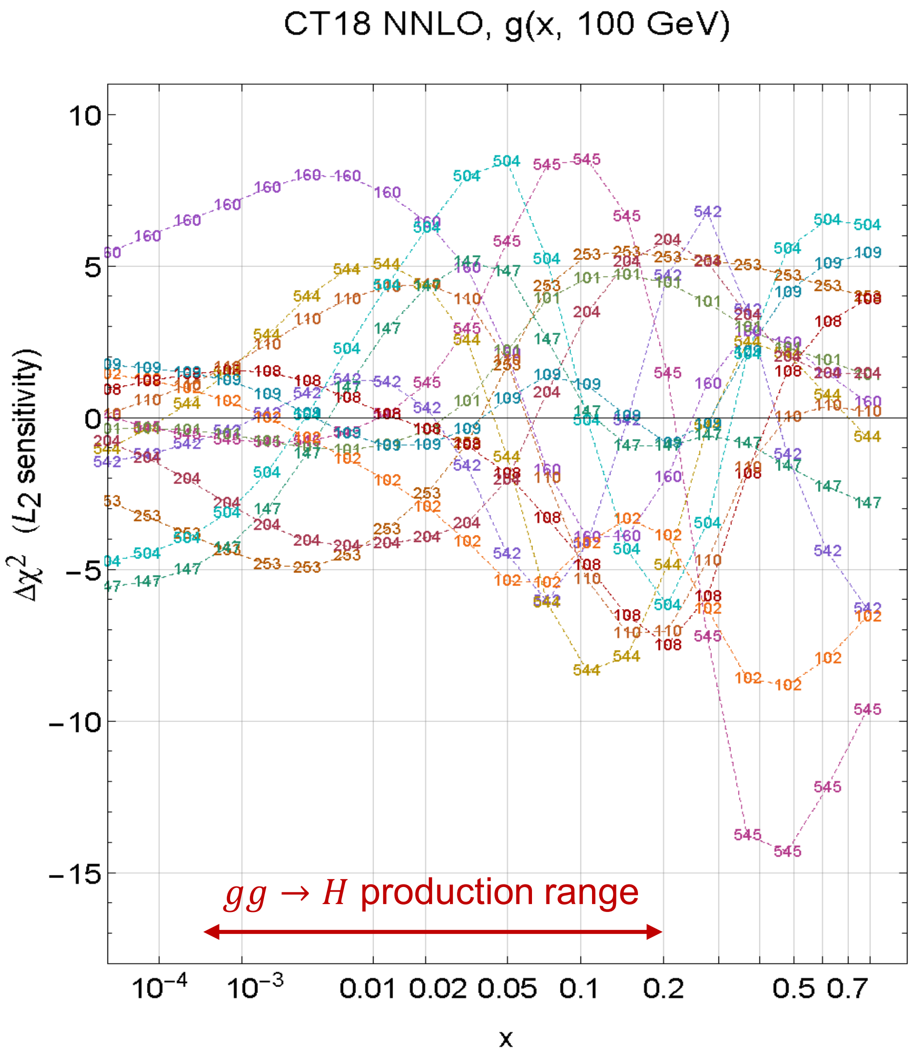
Most sensitive experiments

253 ATL8ZpTbT	109 cdhswf3
542 CMS7jtR7y6T	110 ccfrf2.mi
544 ATL7jtR6uT	147 Hn1X0c
545 CMS8jtR7T	204 e866ppxf
160 HERAIpII	504 cdf2jtCor2
101 BcdF2pCor	
102 BcdF2dCor	
108 cdhswf2	

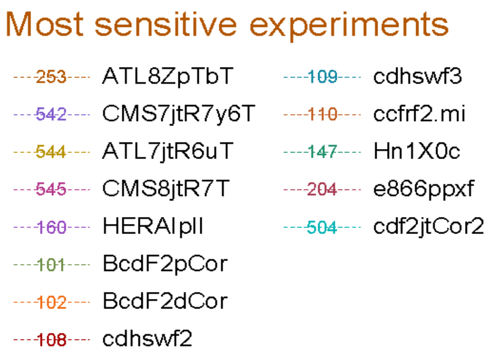
Experiments with large $\Delta\chi^2 > 0$ [$\Delta\chi^2 < 0$] pull $g(x,Q)$ in the negative [positive] direction at the shown x

Estimated χ^2 pulls from experiments

(L_2 sensitivity, T. J. Hobbs et al., arXiv:1904.00222)



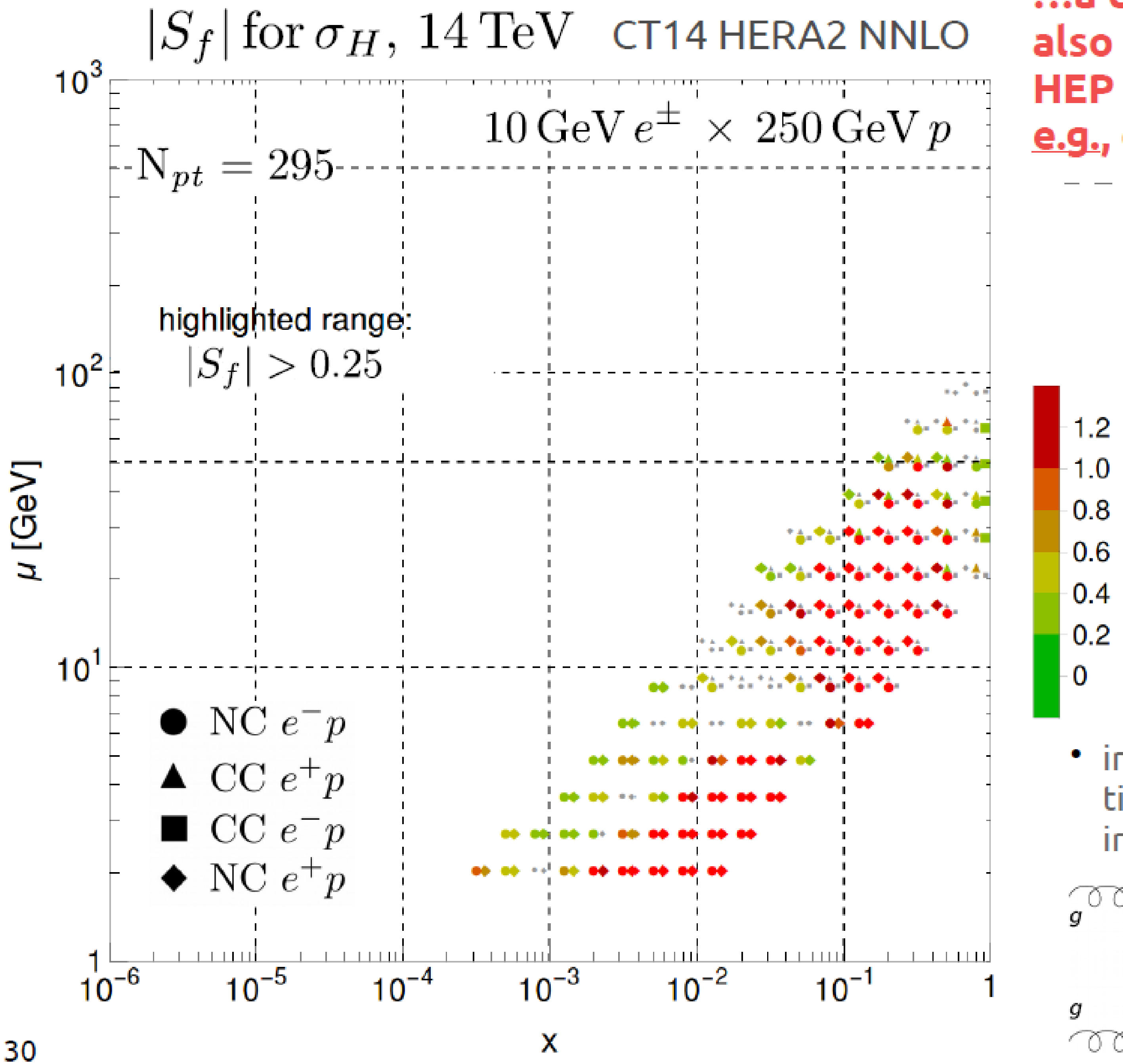
CT18 NNLO, gluon at Q=100 GeV



Note opposite pulls (tensions) in some x ranges between HERA I+II DIS (ID=160); CDF (504), ATLAS 7 (544), CMS 7 (542), CMS 8 jet (545) production; E866pp DY (204); ATLAS 8 Z pT (253) production; BCDMS and CDHSW DIS

What about *future experiments*
…like the EIC or LHeC?
especially, in the context of other measurements at HL-LHC

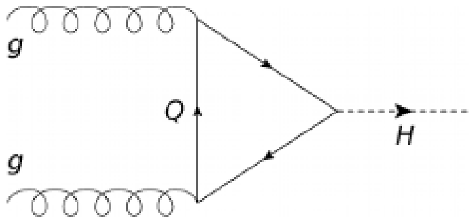
- **EIC and LHeC** PDFSense projections by Hobbs and Wang
- Compared to **HL-LHC** projections by Abdul Khalek, Bailey, Gao, Harland-Lang, Rojo **[arXiv:1810.03039]**



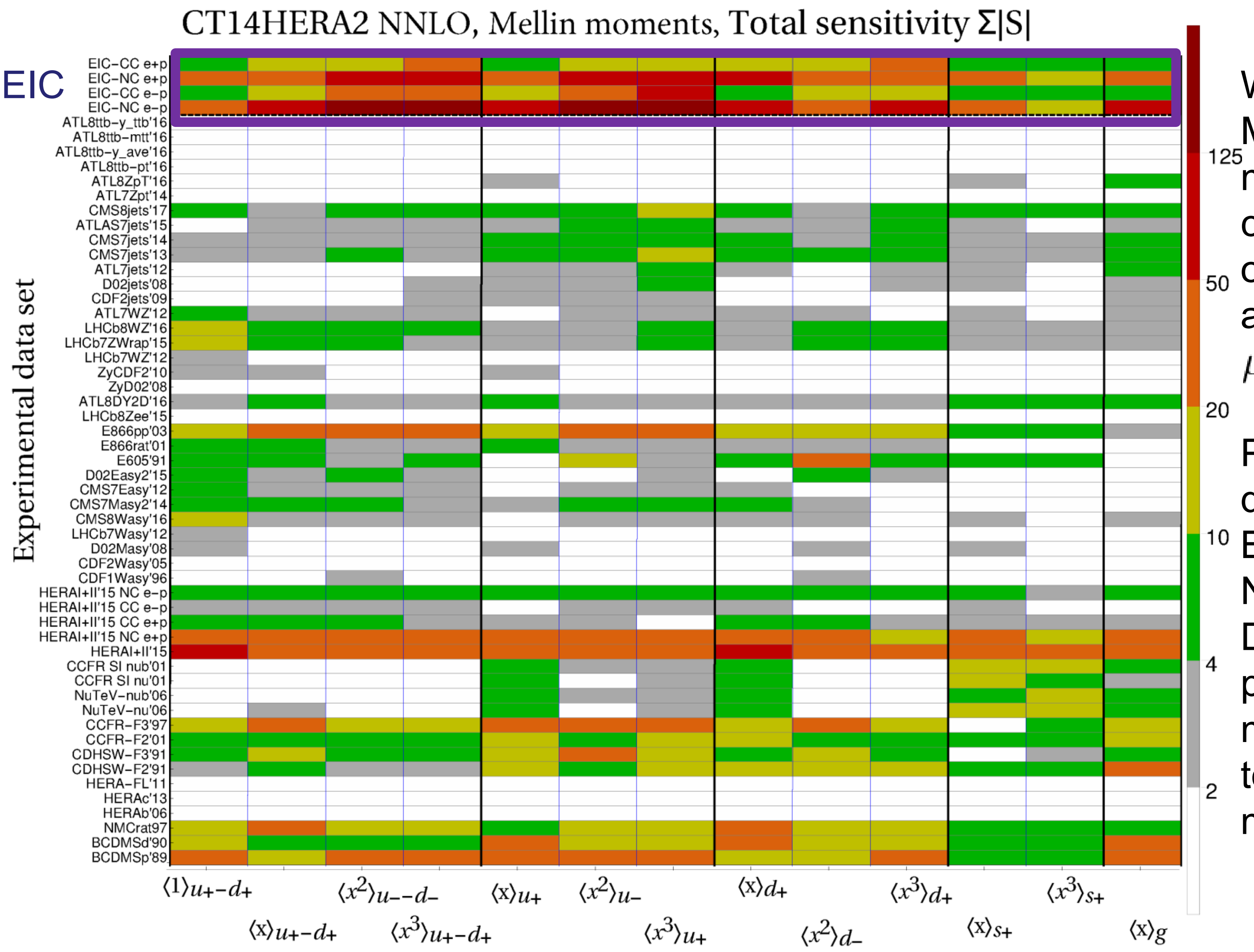
…a US-based EIC will also have important HEP consequences, e.g., on Higgs physics

- the impact of an EIC upon the theoretical predictions for inclusive Higgs production arises from a very broad region of the kinematical space it can access

- impact rather closely tied to that of the integrated gluon PDF:



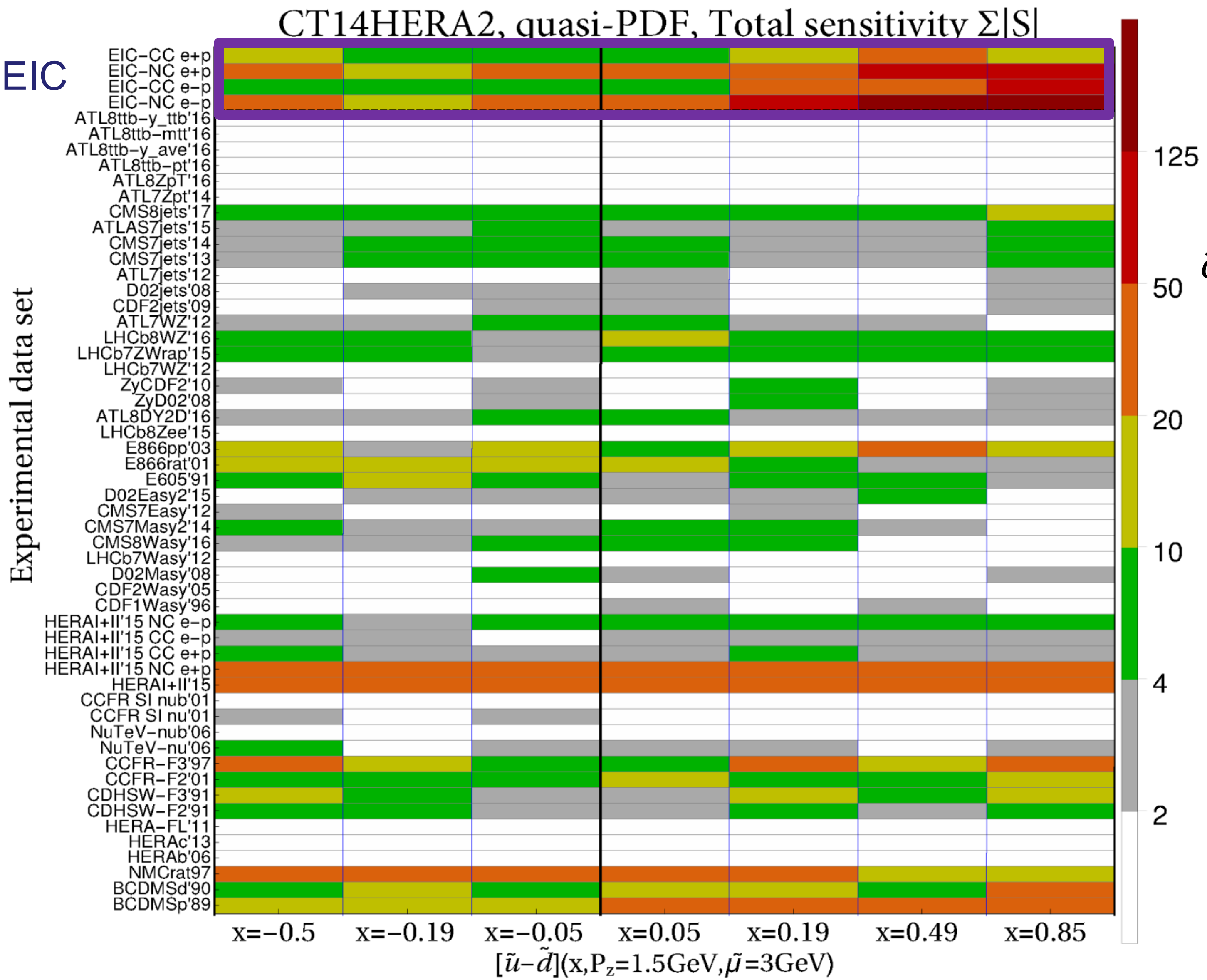
Total sensitivity to Mellin moments



We show Mellin moments computable on the lattice at the scale $\mu_F = 2\,\mathrm{GeV}$

Pre-2019 data: HERA, BCDMS, NMC, E866 DY pair production are most sensitive to the moments

Total sensitivity to lattice quasi-PDFs



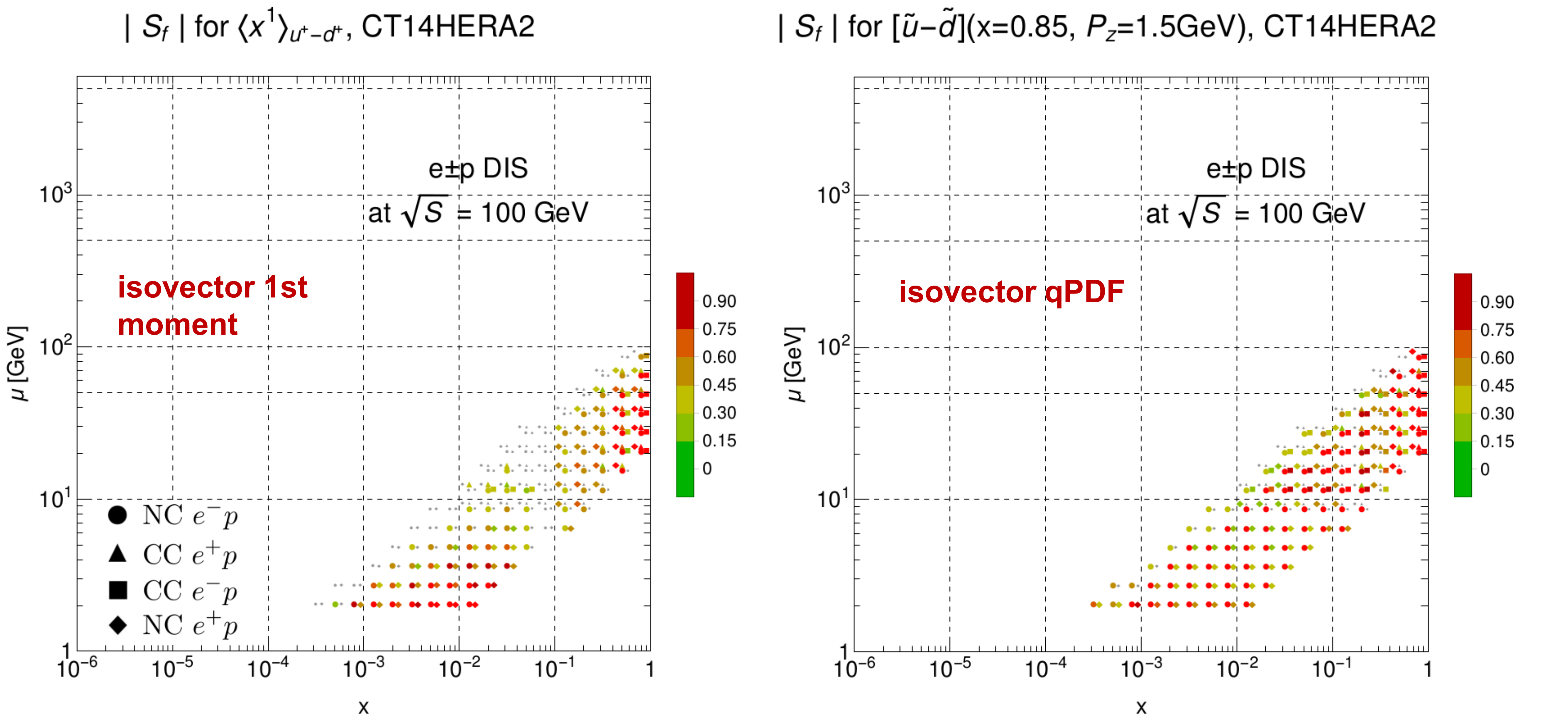
$$\tilde{q}(x) \equiv \begin{cases} q(x), \\ x > 0; \\ -\bar{q}(|x|), \\ x < 0 \end{cases}$$

$Q = 3$ GeV,
$P_z = 1.5$ GeV

An EIC would drive lattice phenomenology

- A high-luminosity lepton-hadron collider will impose very tight constraints on many lattice observables; below, the isovector first moment and qPDF

- Many of the experiments most sensitive to PDF Mellin moments and qPDFs involve nuclear targets ⟶ **eA data from EIC would sharpen knowledge of nuclear corrections**



Key points: future experiments

- The HL-LHC, EIC, and LHeC have **complementary potentials,** make a powerful physics case in combination. Highlights:
  - **HL-LHC:** reach to high Q, a variety of processes sensitive to PDFs; high sensitivity to the gluon, u and d antiquarks
  - **LHeC:** reach to $x < 10^{-6}$; high sensitivity to small-x gluon, as well as d and s quarks (esp. at $x \to 1$) in the clean $e^{\pm}p$ scattering environment; independent of BSM physics
  - **EIC:** supercedes the bulk of fixed-target DIS experiments; offers unique reach in the region of $x > 0.1$ necessary for tests of lattice QCD and nonperturbative QCD models; nucleon and nuclear beams; flavor separation via SIDIS; polarized beams

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

A rich domain of SM phenomenology!

Impact on a wide range of HEP and NP studies

Multiloop QCD and EW computations

Exploration of most complex experimental data sets

Accurate and fast high-performance computing

Frontier statistical inference in many dimensions

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

} opportunities for conceptual breakthroughs

Recap: foundations of multivariate fits

1. Fitting as learning
   a. Meaning of uncertainties: Bayesian, frequentist, Hessian, Monte-Carlo,…
   b. Goodness-of-fit criteria: χ^2 is not the only measure!
   c. Wilks' theorem: the likelihood ratio as the fundamental quantity for hypothesis/parameter testing
   d. Aleatoric, model, distributional uncertainties in an ML-based approach
   e. Averaging over model uncertainty
   f. Fitting the likelihood and priors (a Gaussian model mixture)
2. Dependence on the number of parameters N_{par}
   a. Parsimony: Occam's razor, information criteria, naturalness…
   b. Curse of dimensionality
   c. Big-data paradox in sampling with many N_{par}
3. Fundamental limitations
   a. Dominance of saddle points in non-convex optimization with many N_{par}
   b. Bias-variance ambiguity
   c. Impact on systematic uncertainties
   d. "No free lunch" theorems

…MUCH TO LEARN AND EXPLORE

BACKUP

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

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

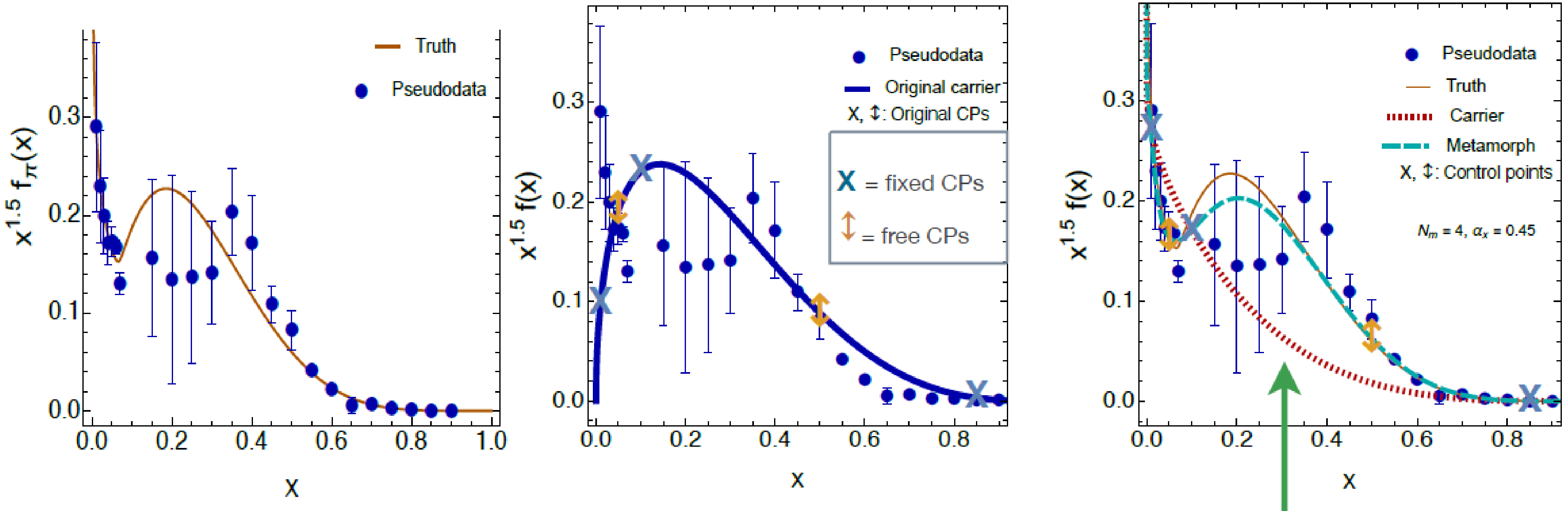
We can evaluate the Bézier curve at chosen **control points**, to get a vector of $\mathscr{B} \to \underline{P}$

$\underline{\underline{T}}$ 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

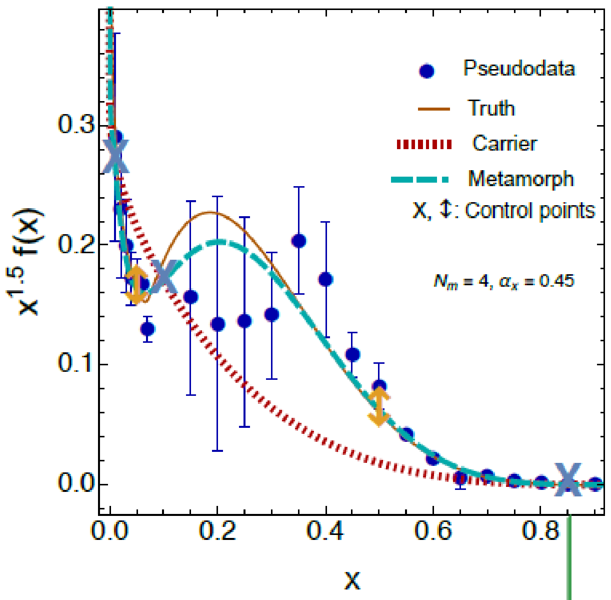
N_m= degree of polynomial can vary

δB_q & δC_q allow the carrier to vary

α_x can vary

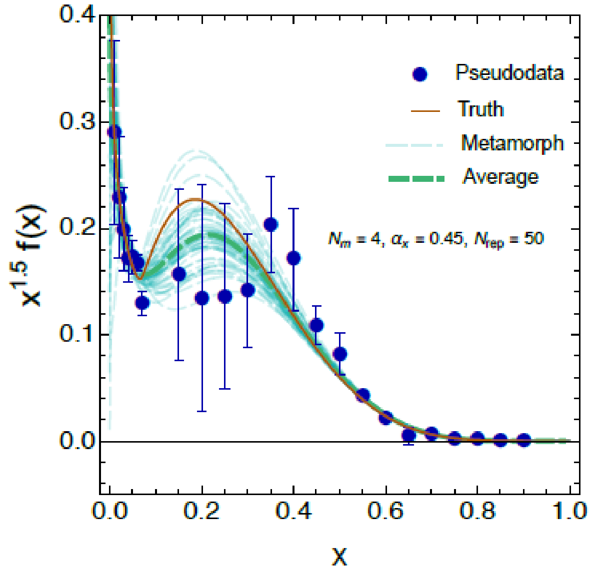
Slide by A. Courtoy

Bézier-curve methodology for global analyses — toy model



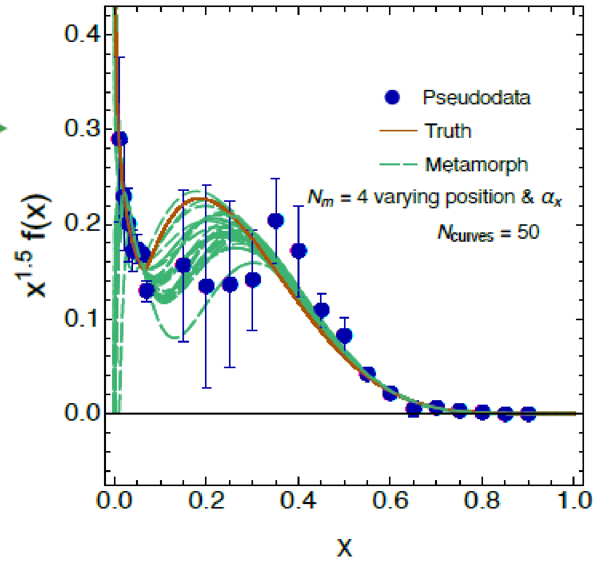
if bootstrapped

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

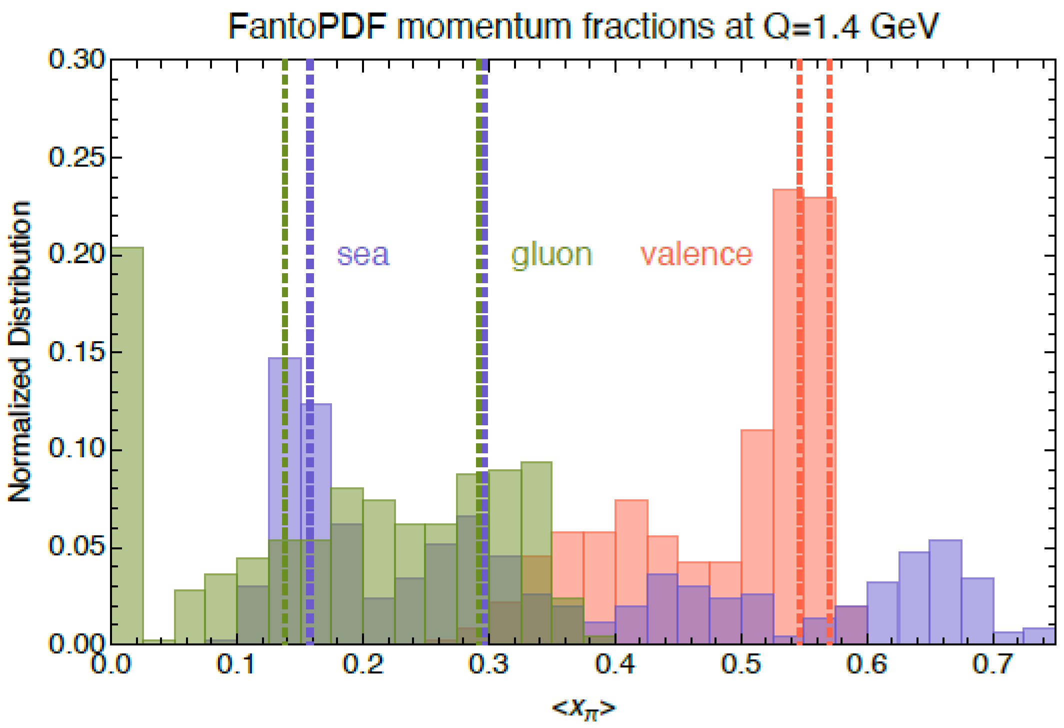


FIG. 11. The histograms represent momentum fractions for the valence (red), gluon (green) and sea (blue) PDFs from 500 MC FantoPDF distributions generated from five candidate fits. The histograms are not symmetric as a consequence of parametrization dependence. Vertical boundaries represent the extrema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

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

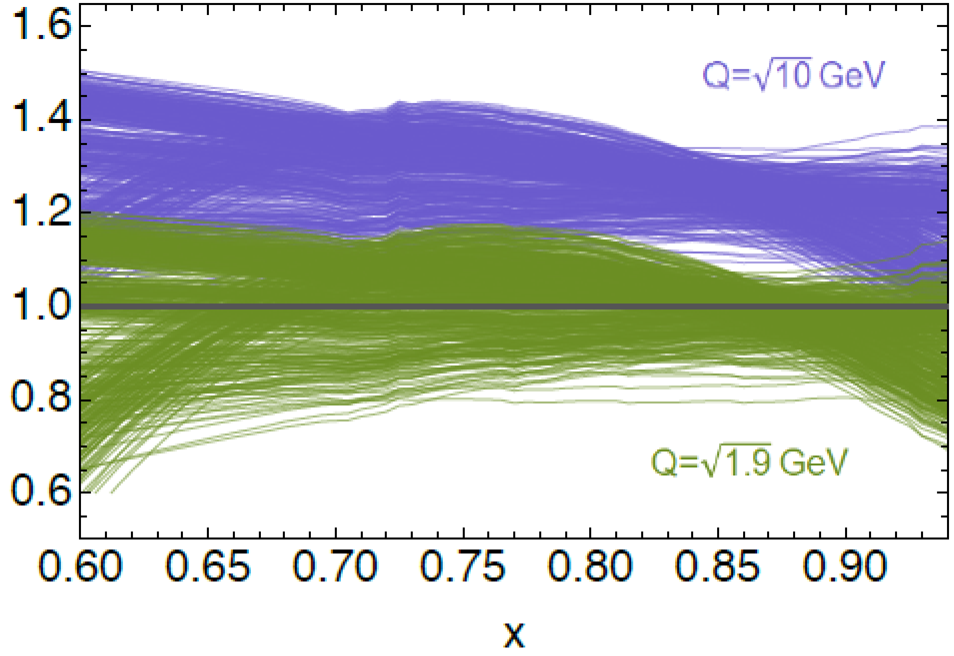
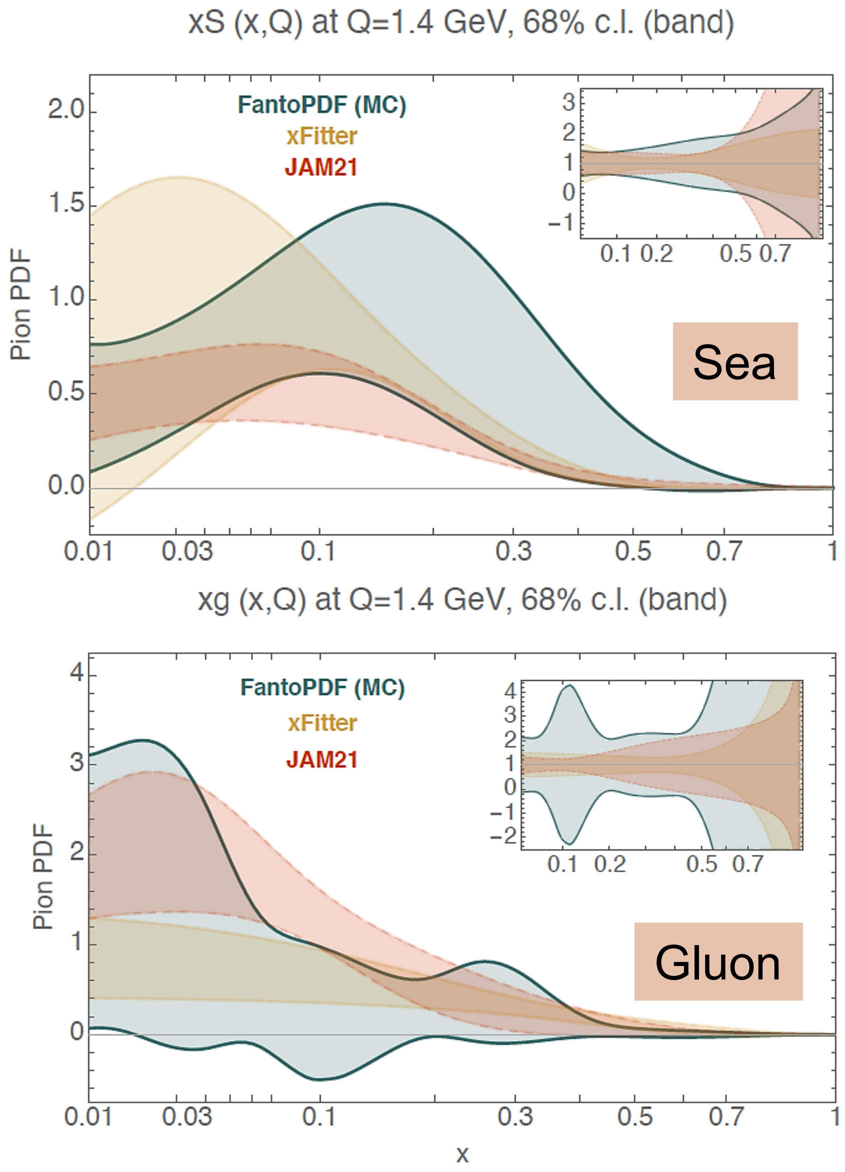


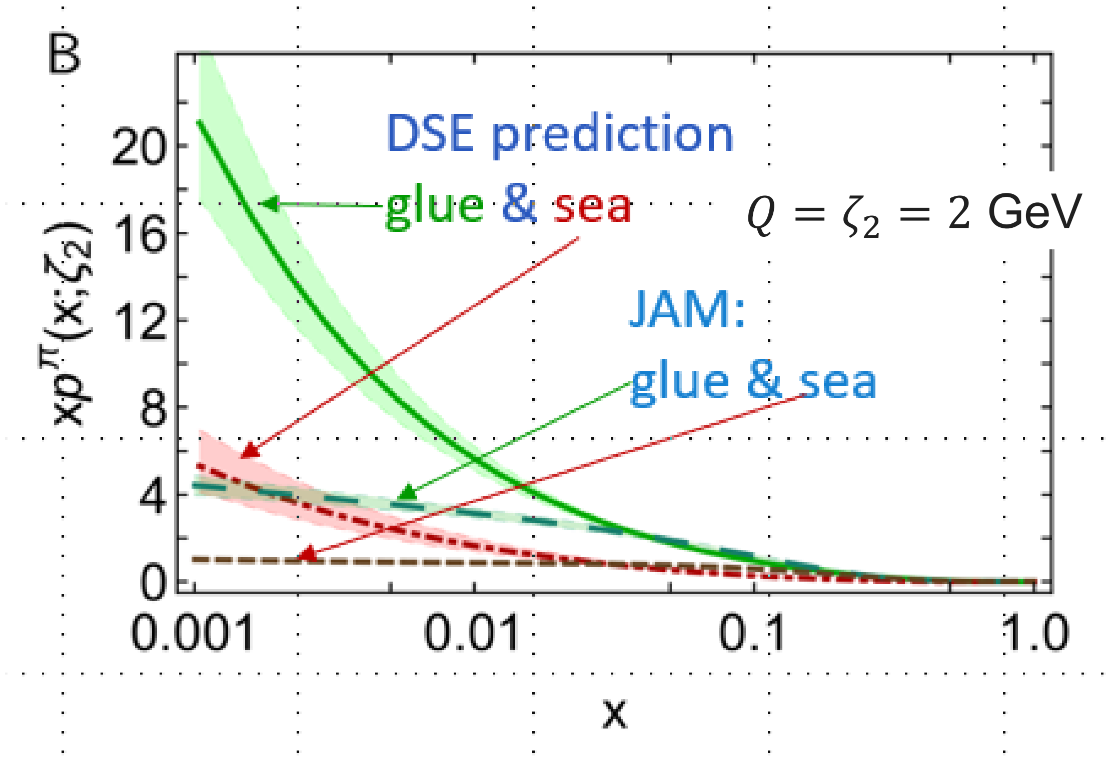
FIG. 12. The effective $(1-x)$ exponent of the valence PDF in the FantoPDF ensemble – the definition is given in [48]. In green, the effective exponent at $Q_0 = \sqrt{1.9}$ GeV and, in blue, at $\sqrt{10}$ GeV. The plot is cut at $x = 0.94$ for grid-extrapolation reasons. We have verified analytically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

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.

What about *future experiments*
…like the EIC or LHeC?
especially, in the context of other measurements at HL-LHC

- **EIC and LHeC** PDFSense projections by Hobbs and Wang
- Compared to **HL-LHC** projections by Abdul Khalek, Bailey, Gao, Harland-Lang, Rojo **[arXiv:1810.03039]**

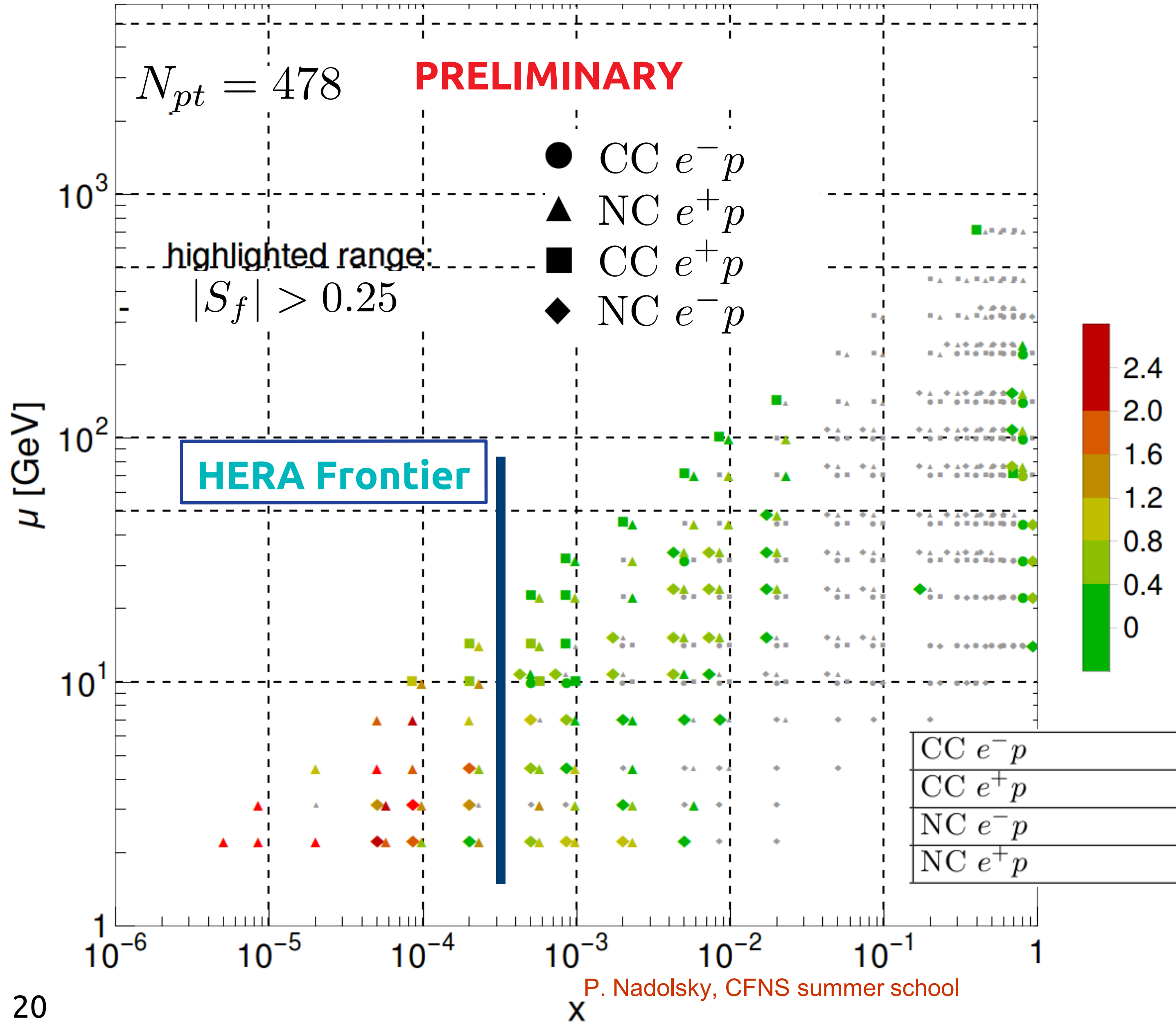
a high-energy Electron-Ion Collider, Large Hadron-electron Collider

- an ep (eA) collider to achieve **high luminosities** > 1000 times that of HERA
  - access a wide range of x, including $x \sim 10^{-6}$
  - explore the dynamics of gluon saturation; greatly improve PDF precision; perform SM tests; and many other physics goals
- can perform a sensitivity analysis of Monte Carlo generated reduced NC/CC cross sections **[Klein & Radescu, LHeC-Note-2013-002 PHY]**

  60 GeV $\mathrm{e}^{\pm}$ on 1 or 7 TeV p

- pseudodata generated by randomly fluctuating about the PDF4LHC15 NNLO prediction according to putative LHeC uncorrelated errors – based on **100 fb-1** of data

$|S_f|$ for $g(x,\mu)$, PDF4LHC15 NNLO



	N_{pt}	
CC e^-p	93	circle
CC e^+p	109	square
NC e^-p	128	diamond
NC e^+p	148	triangle

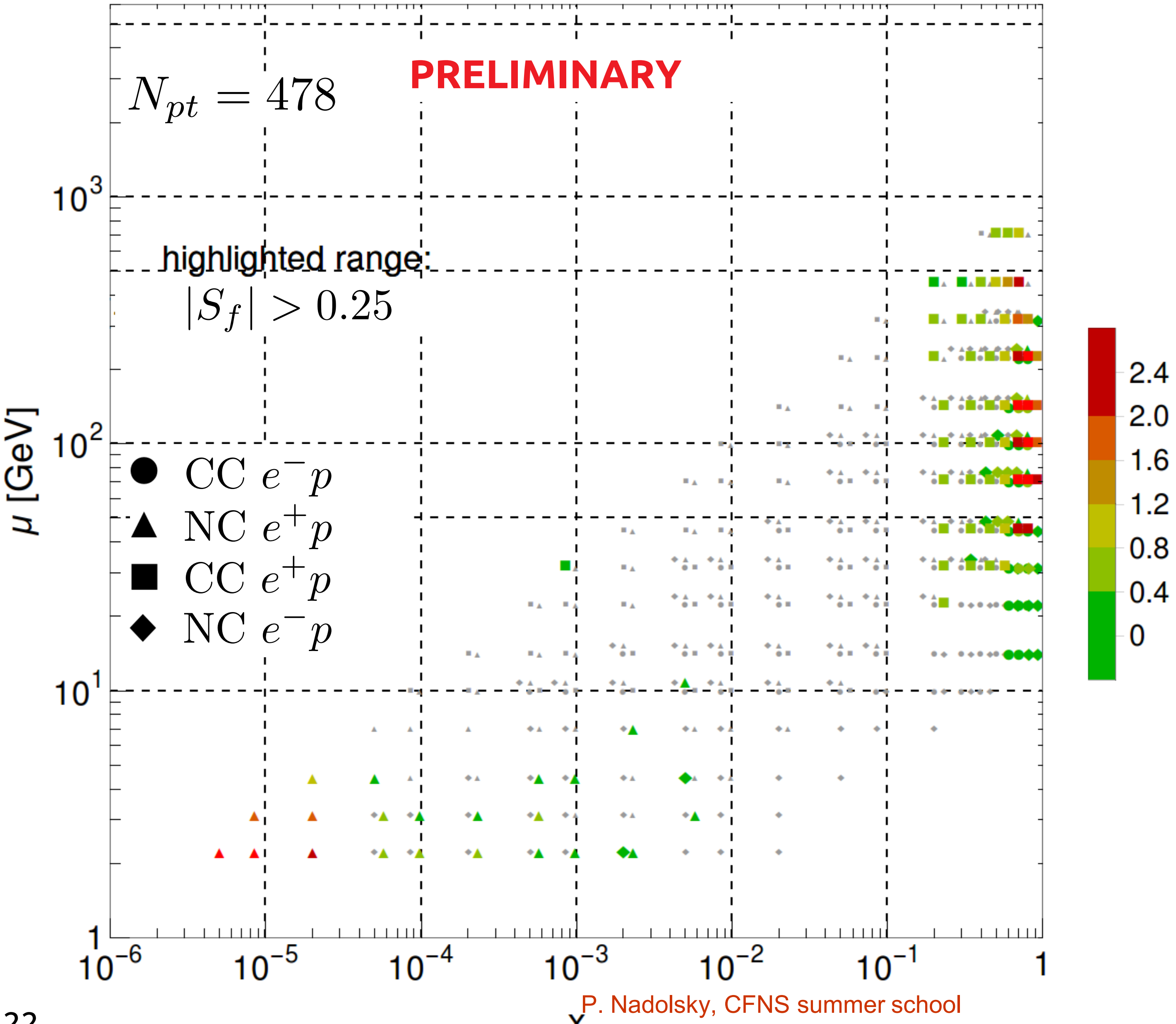
LHeC

- **very strong sensitivities along the frontier of the HERA Run II data, $x \lesssim 10^{-4}$**
- **In addition, important constraints for $x \lesssim 10^{-2}$**

...and high $x, \ \mu$

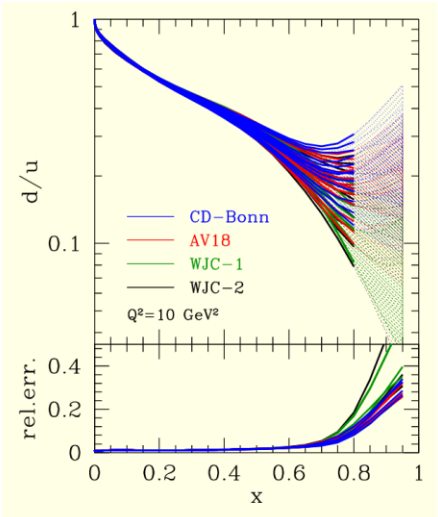
20

| S_f | for d/u(x,mu), PDF4LHC15 NNLO



LHeC

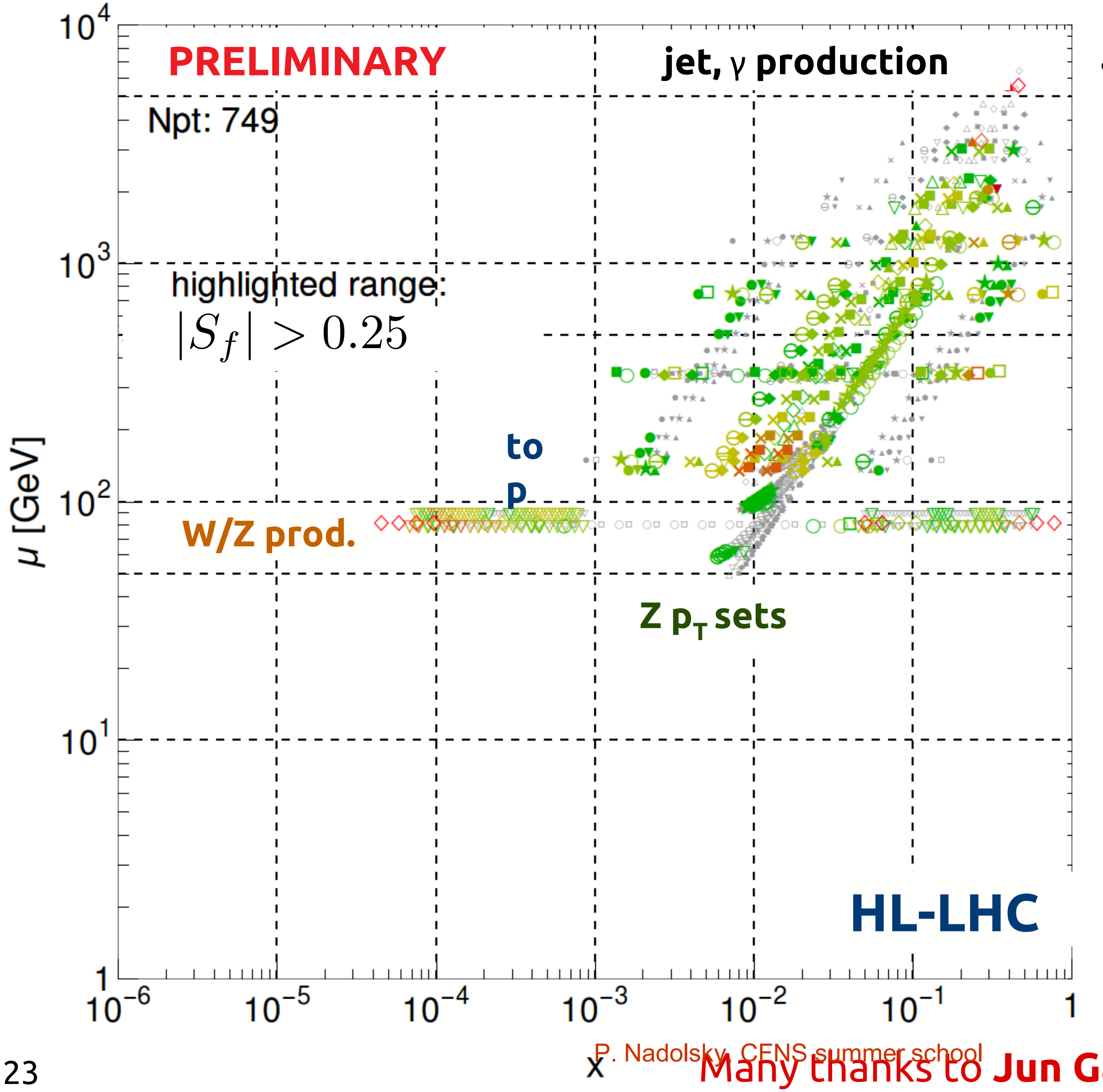
- LHeC's high luminosity may give it a reach to high enough x to help resolve the stubborn d/u question
  - …without a nuclear target…



J.F. Owens, PAVI11

$|S_f|$ for $g(x,\mu)$, PDF4LHC15 NNLO

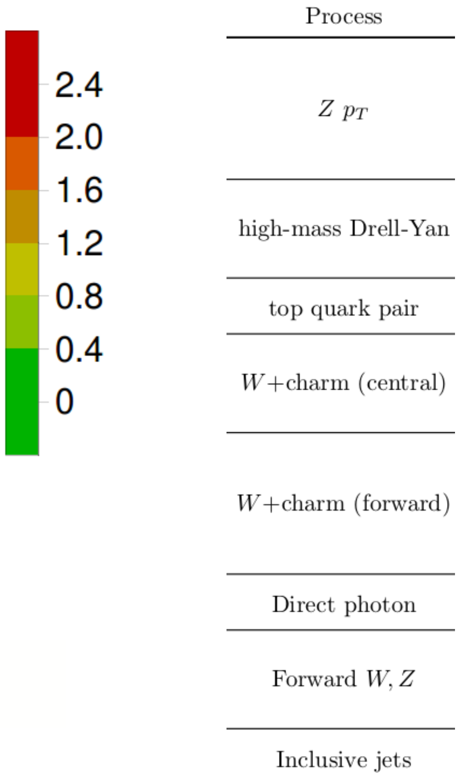
HL-LHC



- **very large integrated luminosities:**

$\mathcal{L} = 3\ ab^{-1}$ (CMS/ATLAS)

$\mathcal{L} = 0.3\ ab^{-1}$ (LHCb)



Many thanks to **Jun Gao** for the pseudodata!

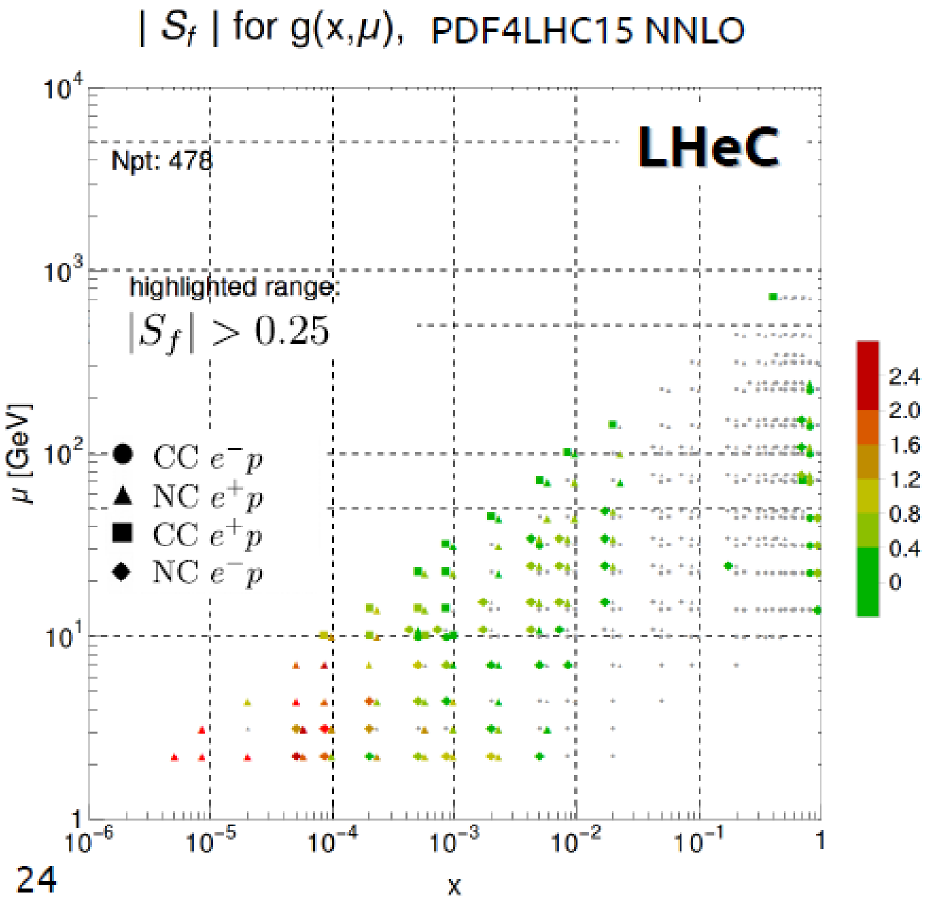
now directly compare the LHeC vs. HL-LHC flavor sensitivities*

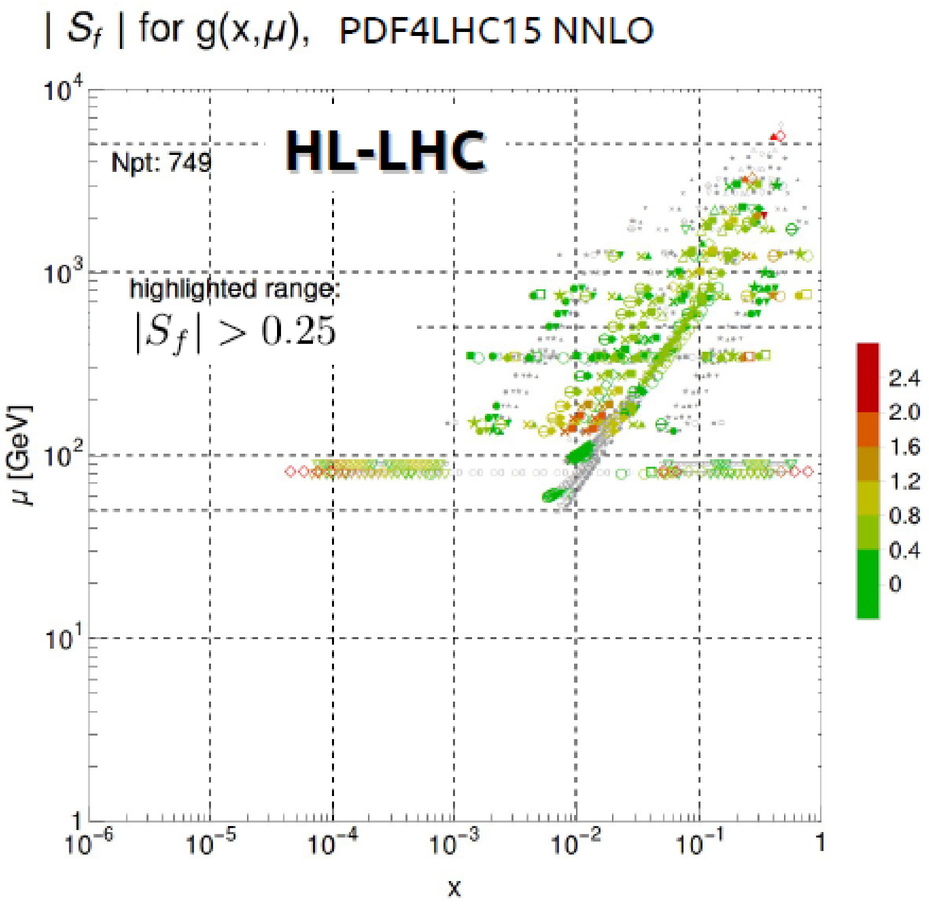
$g(x, \mu)$

*noting the much larger integrated luminosity of the HL-LHC pseudo-data

→ we compare both kinematic regions of especially strong sensitivity, and the aggregated impact of each experiment:

$$|S_g^{\mathrm{LHeC}}| = 151.4 \quad < \quad |S_g^{\mathrm{HL-LHC}}| = 244.6$$



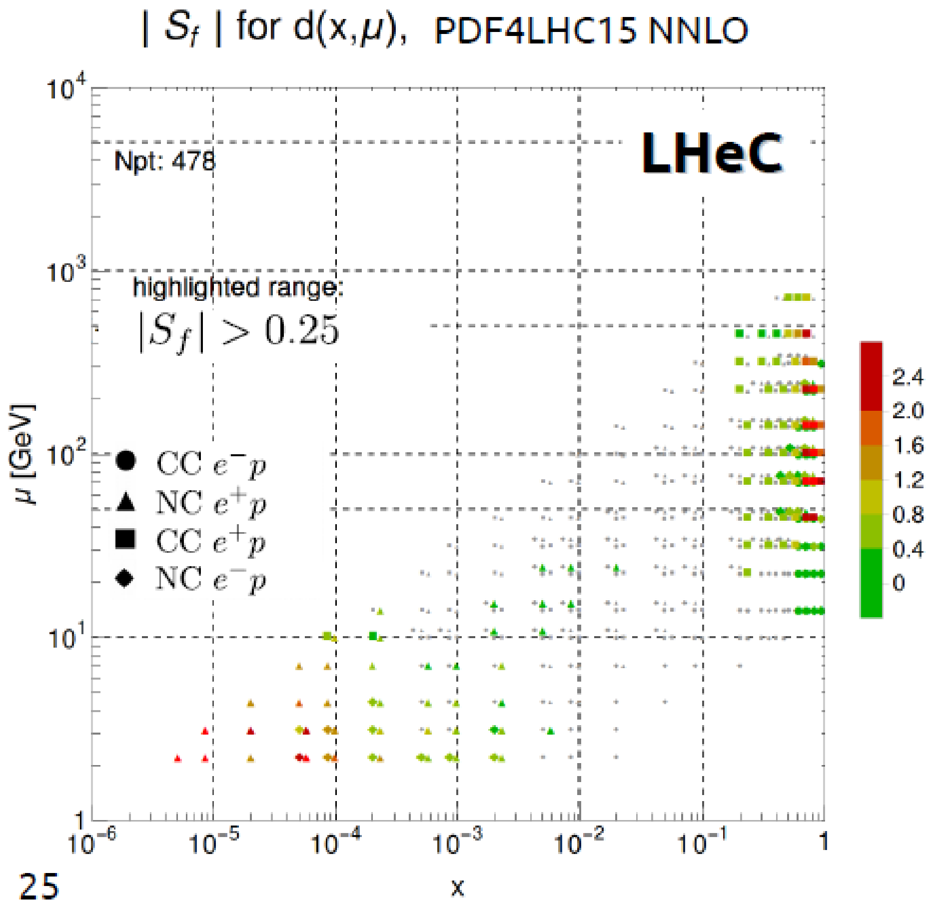


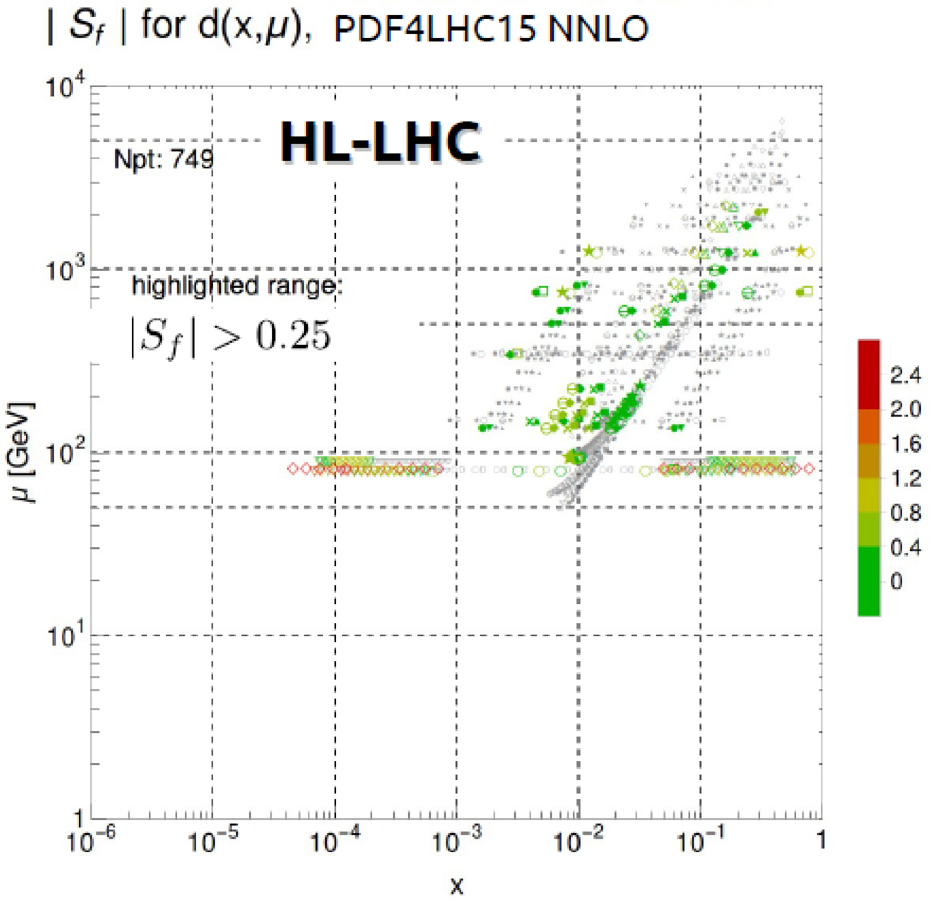
even with a small (pseudo)data set, LHeC enjoys strong sensitivity to down-type distributions!

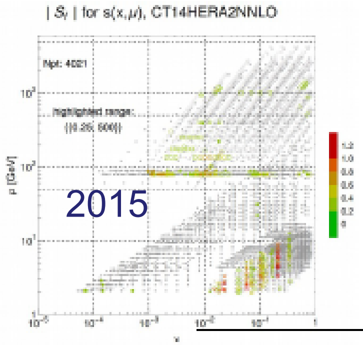
d(x,μ)

→ especially in a fashion complementary to HL-LHC, at very high/low x

$$|S_d^{\mathrm{LHeC}}| = 214.3 \quad > \quad |S_d^{\mathrm{HL-LHC}}| = 170.8$$





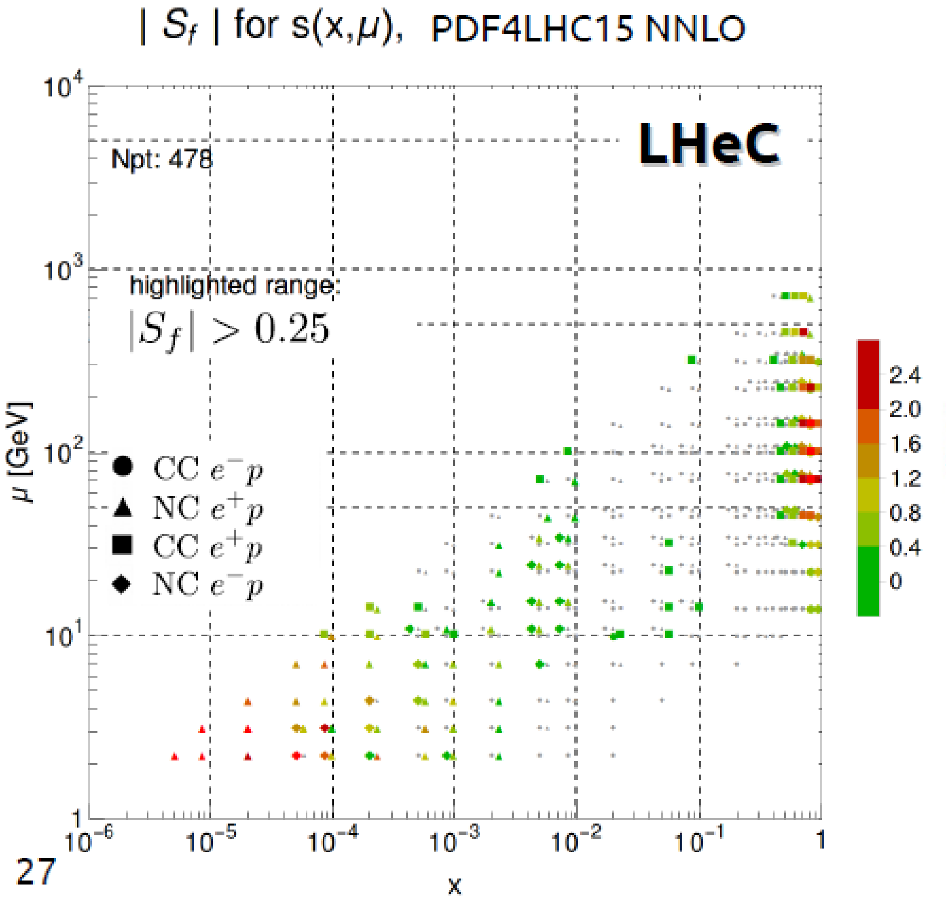


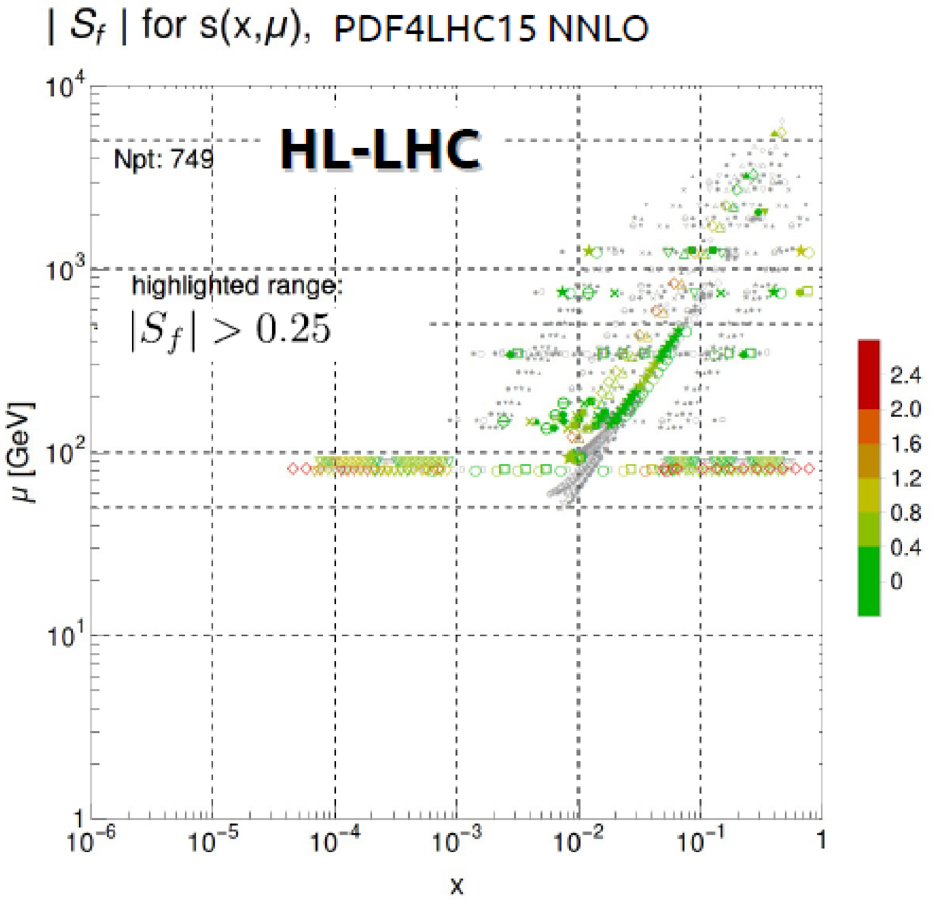
→ LHeC especially portends significantly heightened knowledge of nucleon strangeness

- CTEQ-TEA constraints come primarily through older fixed-target data and Tevatron data (and LHC Run I)

s(x, μ)

$$|S_s^{\mathrm{LHeC}}| = 214.1 \quad \gtrsim \quad |S_s^{\mathrm{HL-LHC}}| = 184.8$$



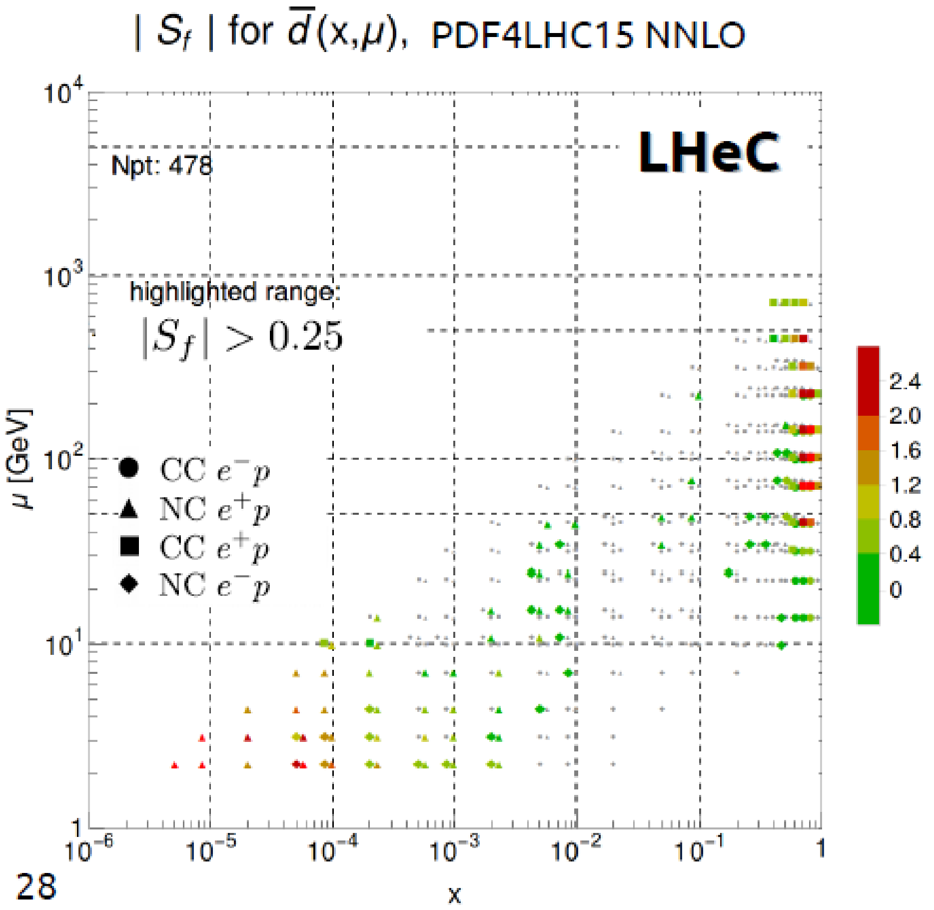


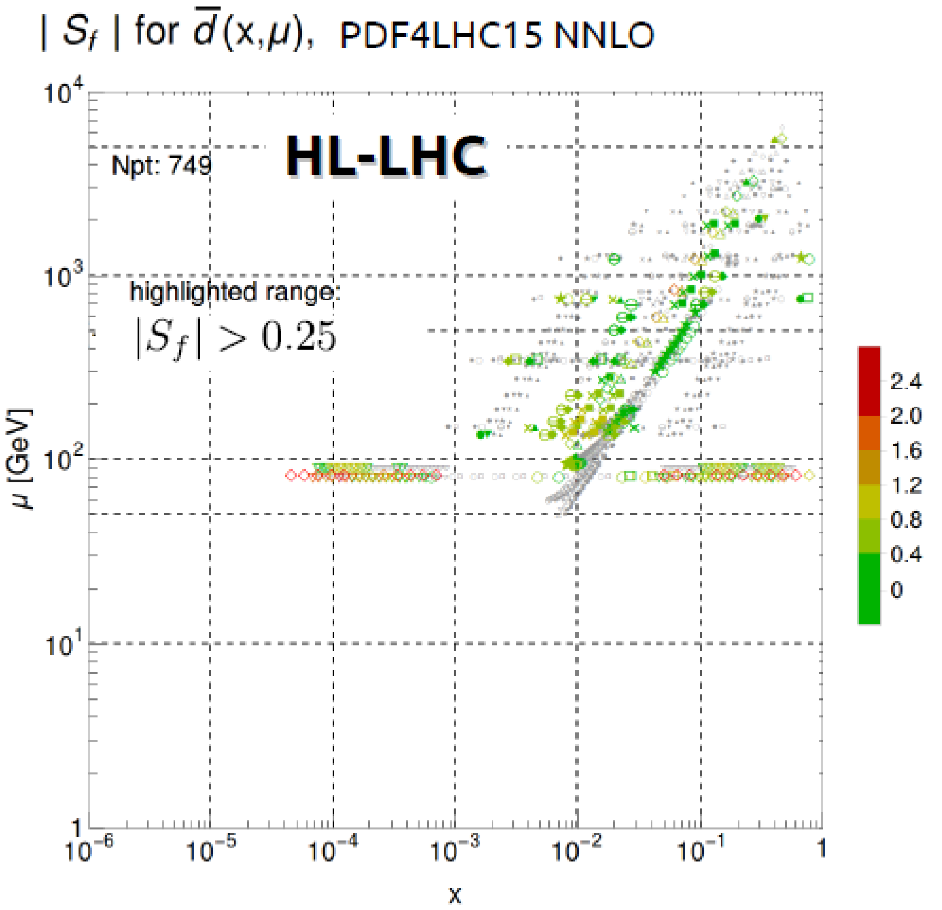
in the SU(2) quark sea, the LHeC 100 fb^{-1} set imposes constraints of magnitude comparable to HL-LHC

$\bar{d}(x,\mu)$

…these again predominate at the extrema of x

$$|S_{\bar{d}}^{\mathrm{LHeC}}| = 192.7 \quad \sim \quad |S_{\bar{d}}^{\mathrm{HL-LHC}}| = 199.4$$





- **lattice QCD** calculations continue to improve and will be increasingly useful as inputs into QCD global analyses

PDF-Lattice whitepaper – Lin et al., PPNP100, 107 (2018); arXiv:1711.07916.

- the PDF-Lattice relationship is *synergistic* :

→ PDF phenomenologists deliver benchmarks to challenge the Lattice

↕

→ Lattice delivers theoretical priors for QCD global fits

PDFSense analysis – Hobbs, Wang, Nadolsky and Olness, arXiv:1904.00022.

---

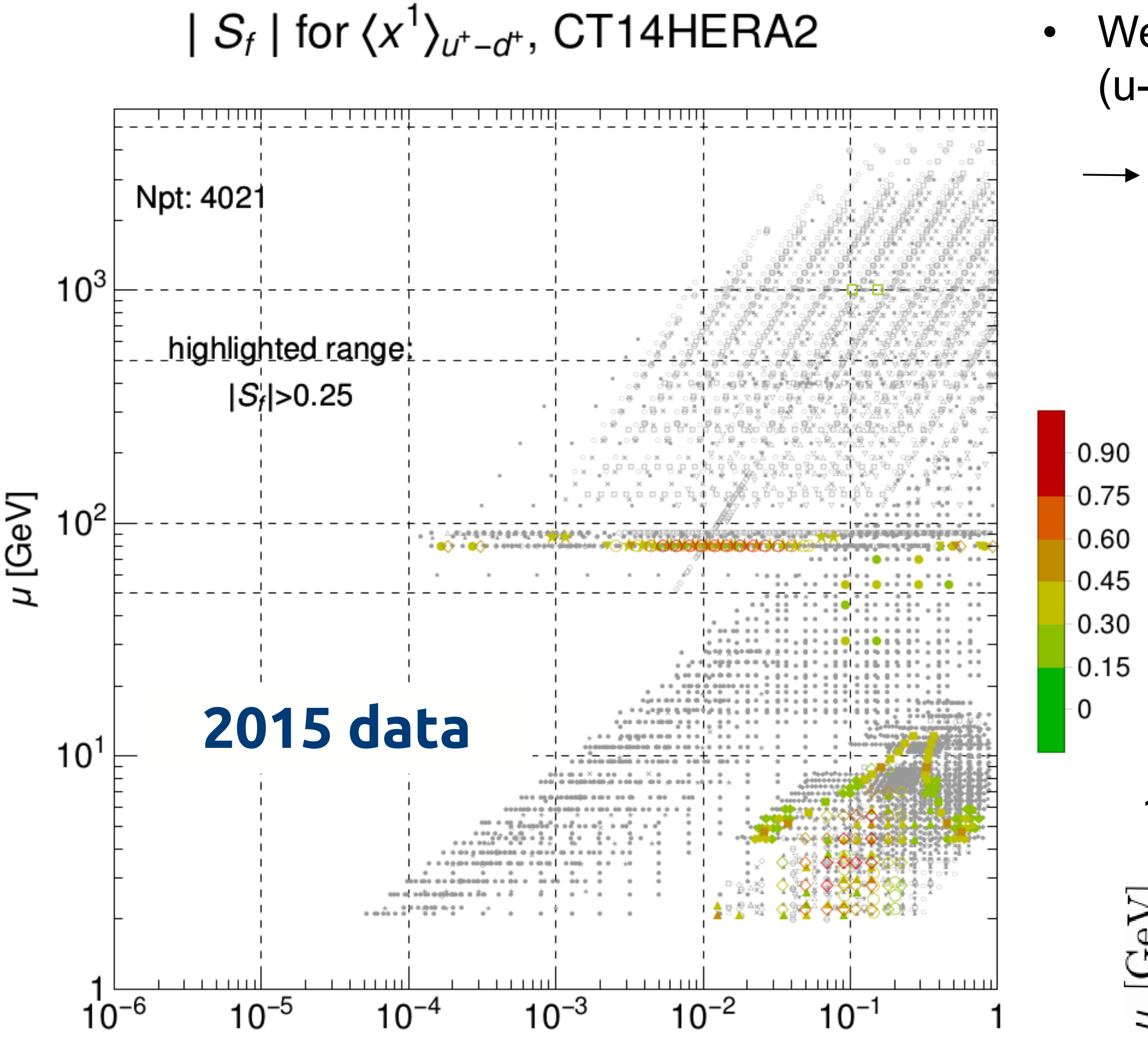
1. Mellin moments from lattice

$$\langle x^n \rangle_q = \int_0^1 dx\; x^n \left[q(x) + (-1)^{n+1}\, \bar{q}(x) \right] \qquad \to \langle x^{1,3,\cdots} \rangle_{q^+},\; \langle x^{2,4,\cdots} \rangle_{q^-}$$

2. Quasi-PDFs (qPDFs) from lattice

$$\widetilde{q}(x, P_z, \widetilde{\mu}) = \int_{-\infty}^{\infty} \frac{dz}{4\pi} e^{ixP_z z} \langle P | \overline{\psi}(z) \gamma^z U(z,0) \psi(0) | P \rangle$$

Sensitivity maps: isovector moments

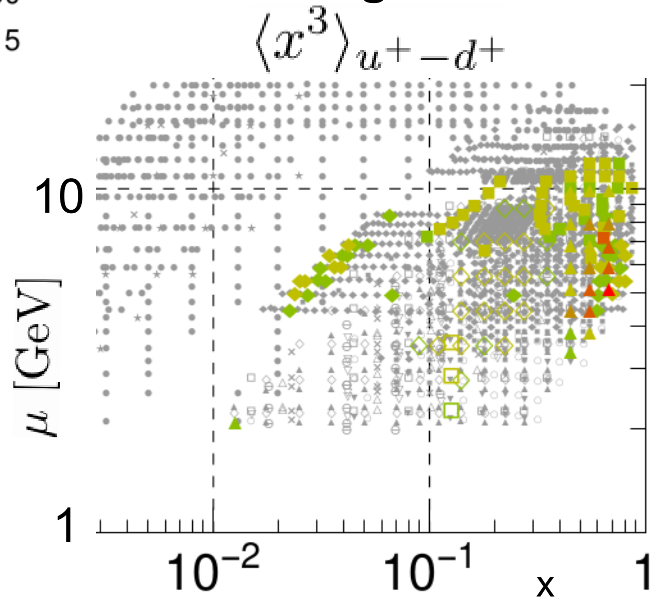


Hobbs et al., arXiv:1904.00222

- We focus on **isovector** (u-d) PDF combinations

→ on the lattice, these are more readily computed, since flavor non-singlet combinations do not receive disconnected insertions

- **Higher-order moments increasingly constrained by higher x**, **fixed-target data**



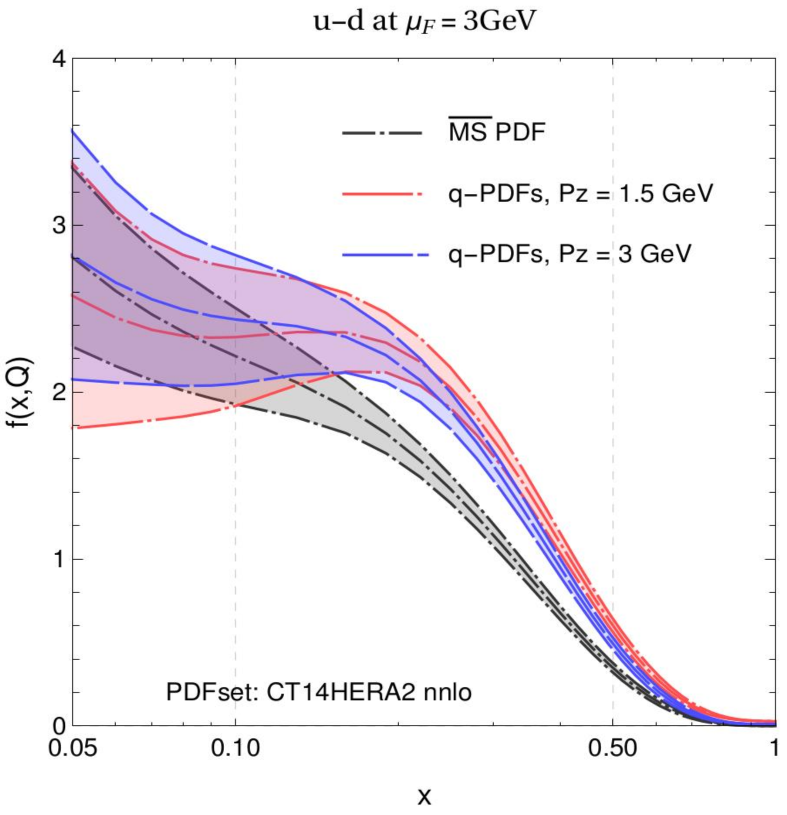
- Also, can **predict** the quasi-PDFs using phenomenological CT PDFs

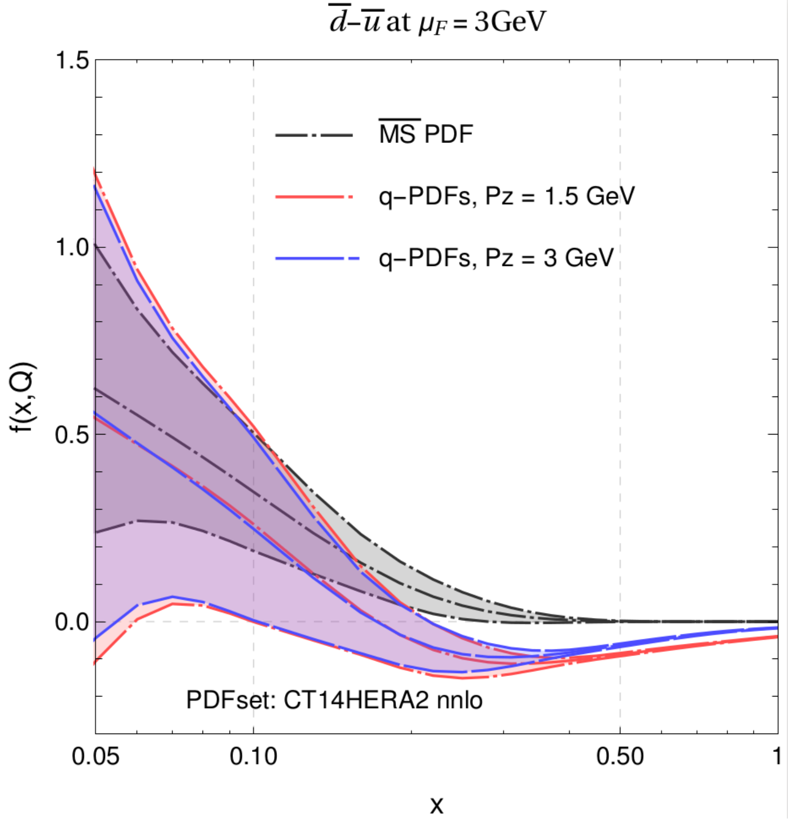
$$\widetilde{q}(x, P_z, \widetilde{\mu}) = \int dy\, Z\left(\frac{x}{y}, \frac{\Lambda}{P_z}, \frac{\mu}{P_z}\right) q(y, \mu) + \mathcal{O}\left(\frac{\Lambda^2}{P_z^2}, \frac{M^2}{P_z^2}\right)$$

Output qPDF;
To be compared against
lattice QCD

Input $\overline{MS}$ PDF (CT14, etc.)

The dependence of qPDFs on the nucleon boost P_z mostly saturates for P_z>1.5 GeV





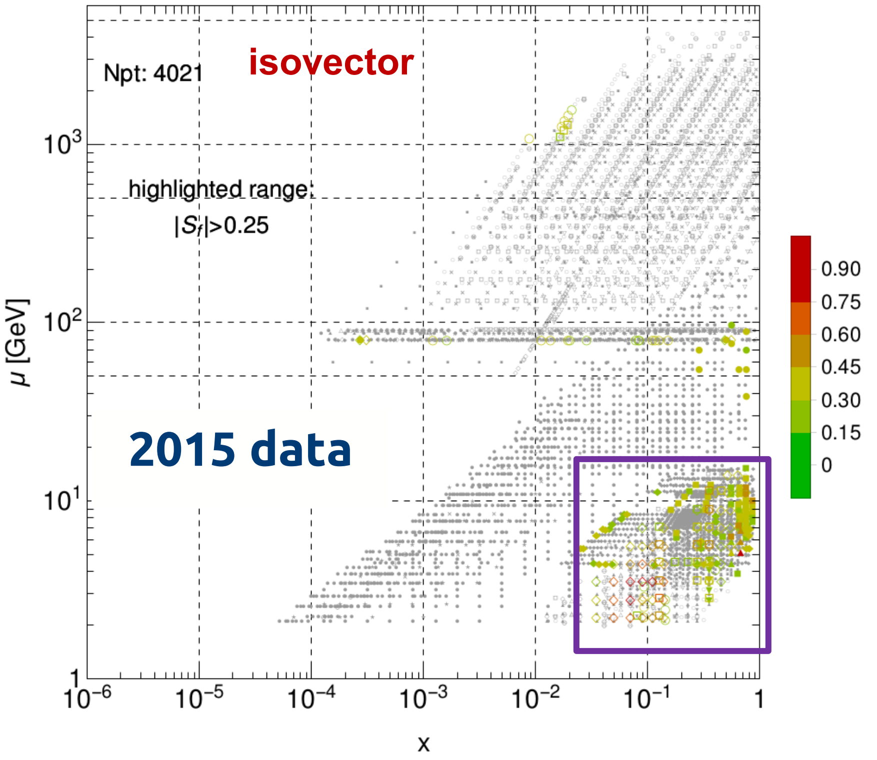
Sensitivity to the "derived" isovector q-PDF

$$\widetilde{q}(x, P_z, \widetilde{\mu}) = \int dy\, Z\left(\frac{x}{y}, \frac{\Lambda}{P_z}, \frac{\mu}{P_z}\right) q(y, \mu) + \mathcal{O}\left(\frac{\Lambda^2}{P_z^2}, \frac{M^2}{P_z^2}\right)$$

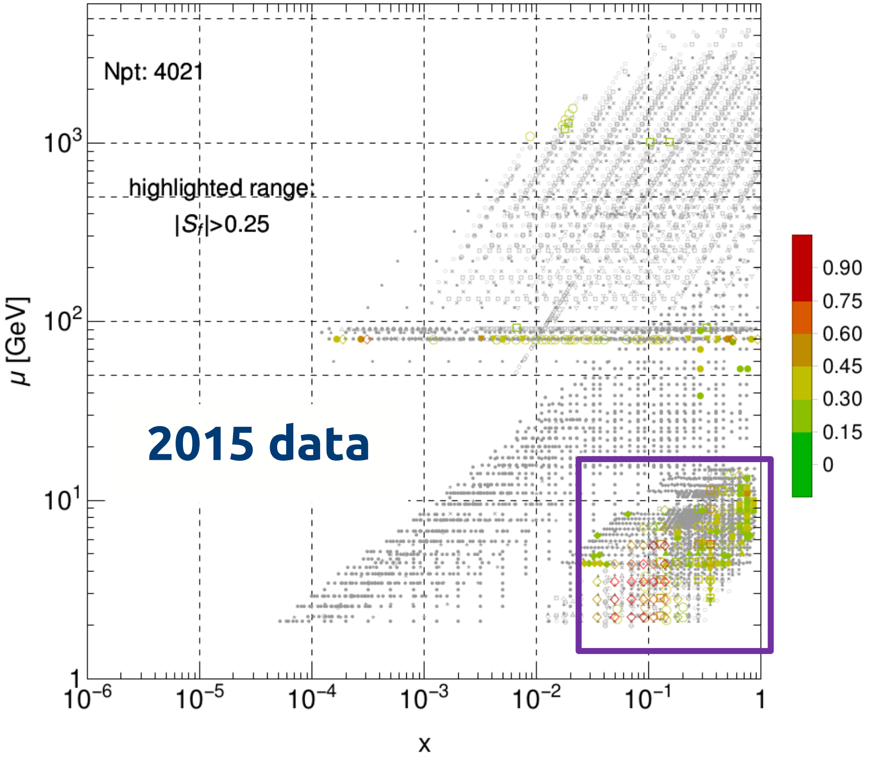
Output qPDF;
To be compared against
lattice QCD

Input $\overline{MS}$ PDF (CT14, etc.)

| S_f | for $[\tilde{u}-\tilde{d}](x=0.85,\ P_z=1.5\text{GeV})$, CT14HERA2



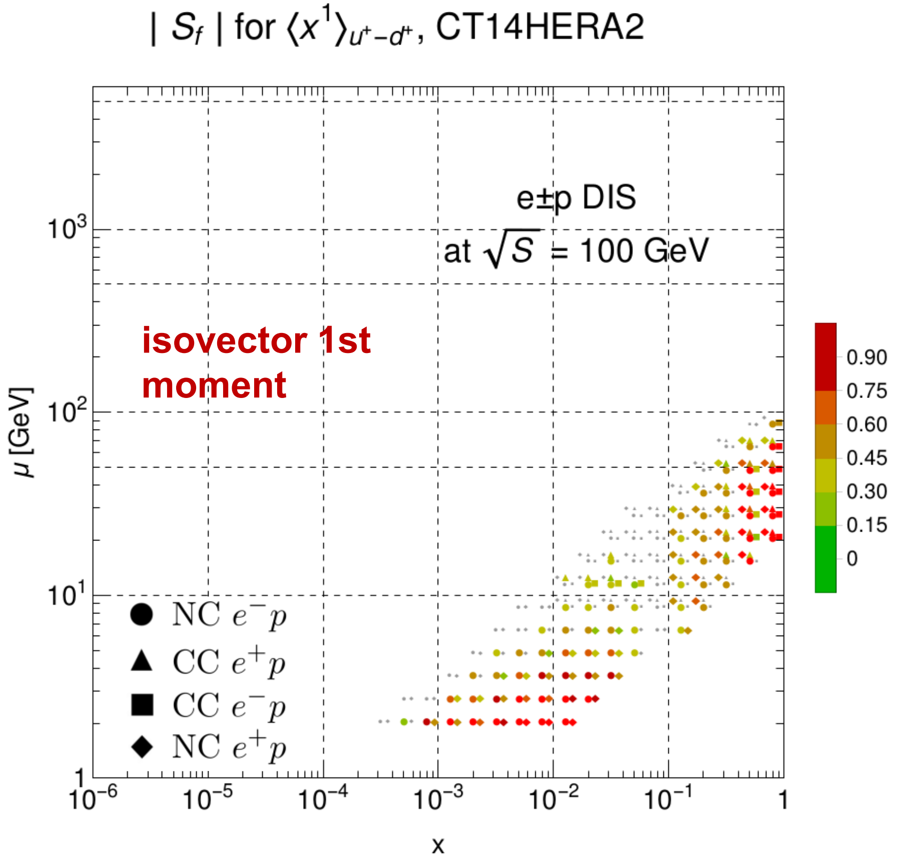
| S_f | for $[\tilde{u}-\tilde{d}](x=0.85,\ P_z=3\text{GeV})$, CT14HERA2

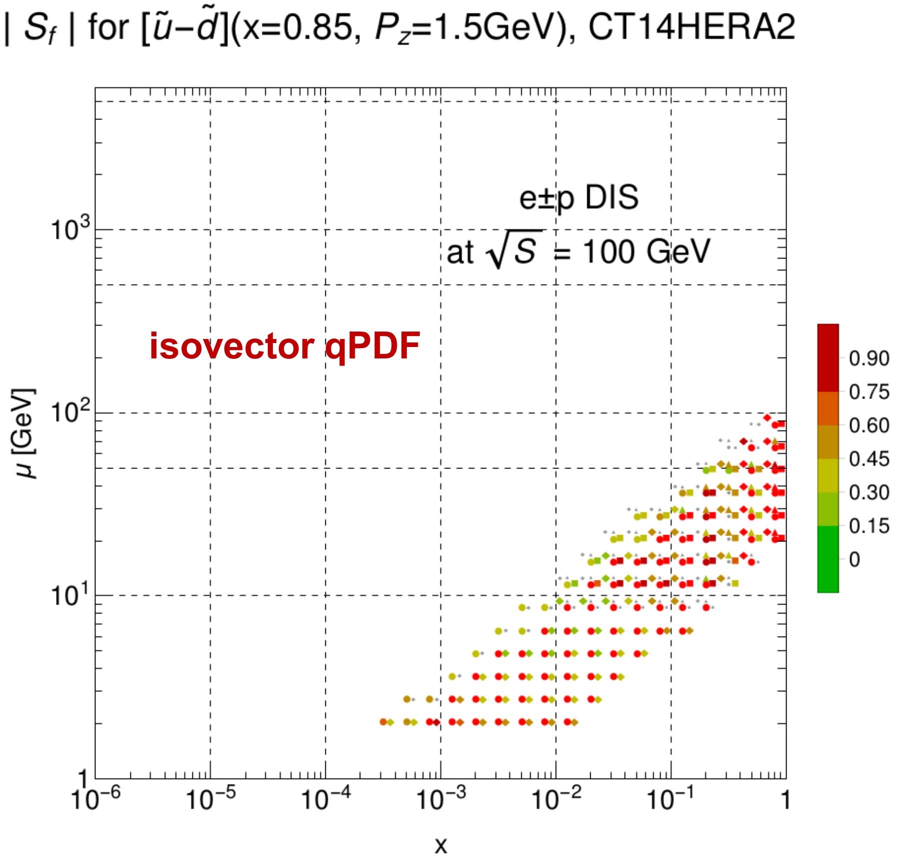


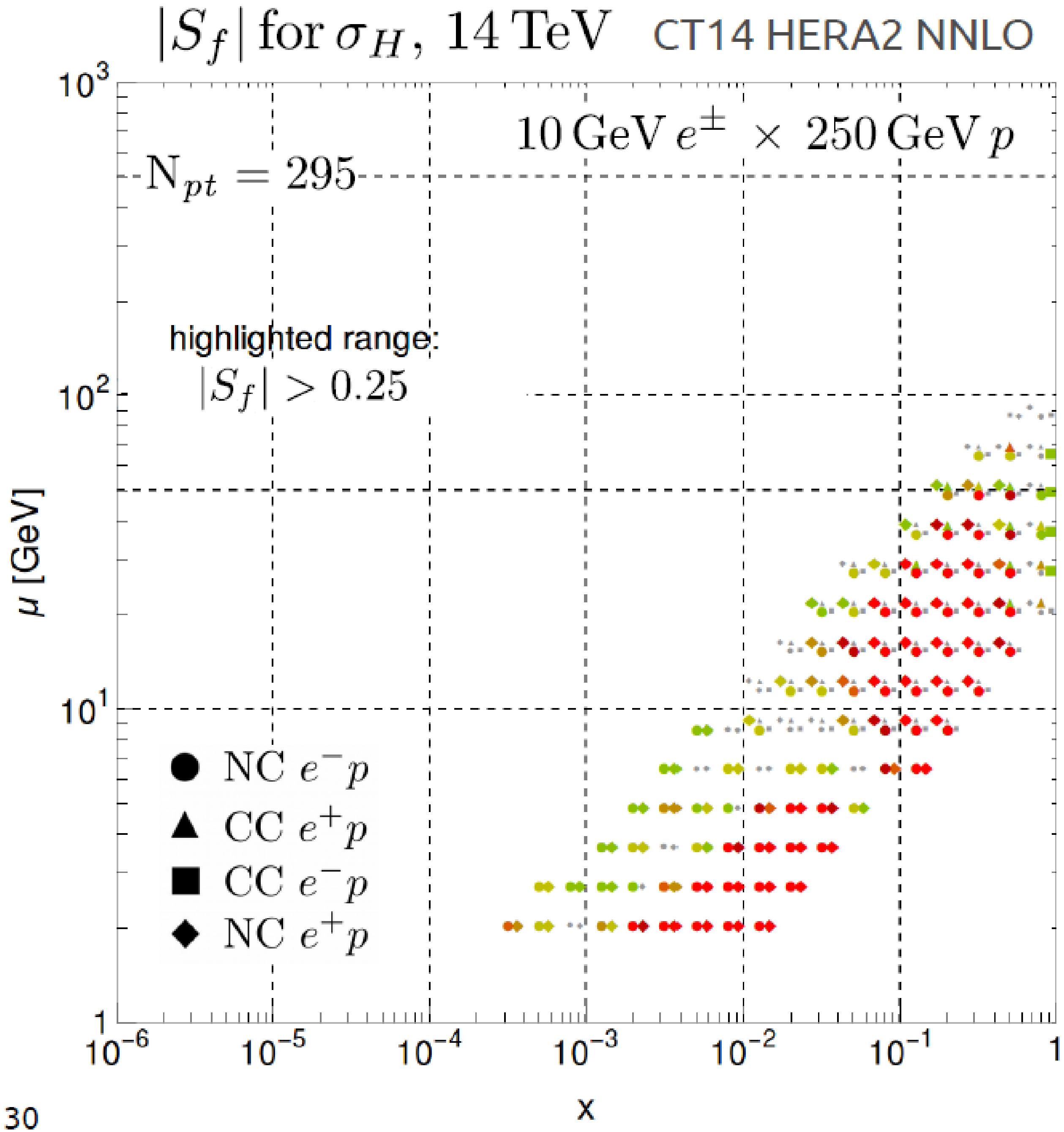
An EIC would drive lattice phenomenology

- A high-luminosity lepton-hadron collider will impose very tight constraints on many lattice observables; below, the isovector first moment and qPDF

- Many of the experiments most sensitive to PDF Mellin moments and qPDFs involve nuclear targets ⟶ **eA data from EIC would sharpen knowledge of nuclear corrections**



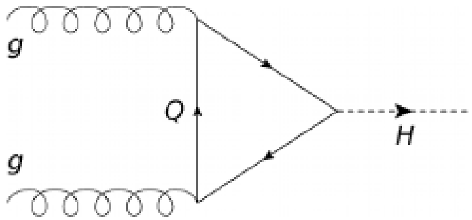




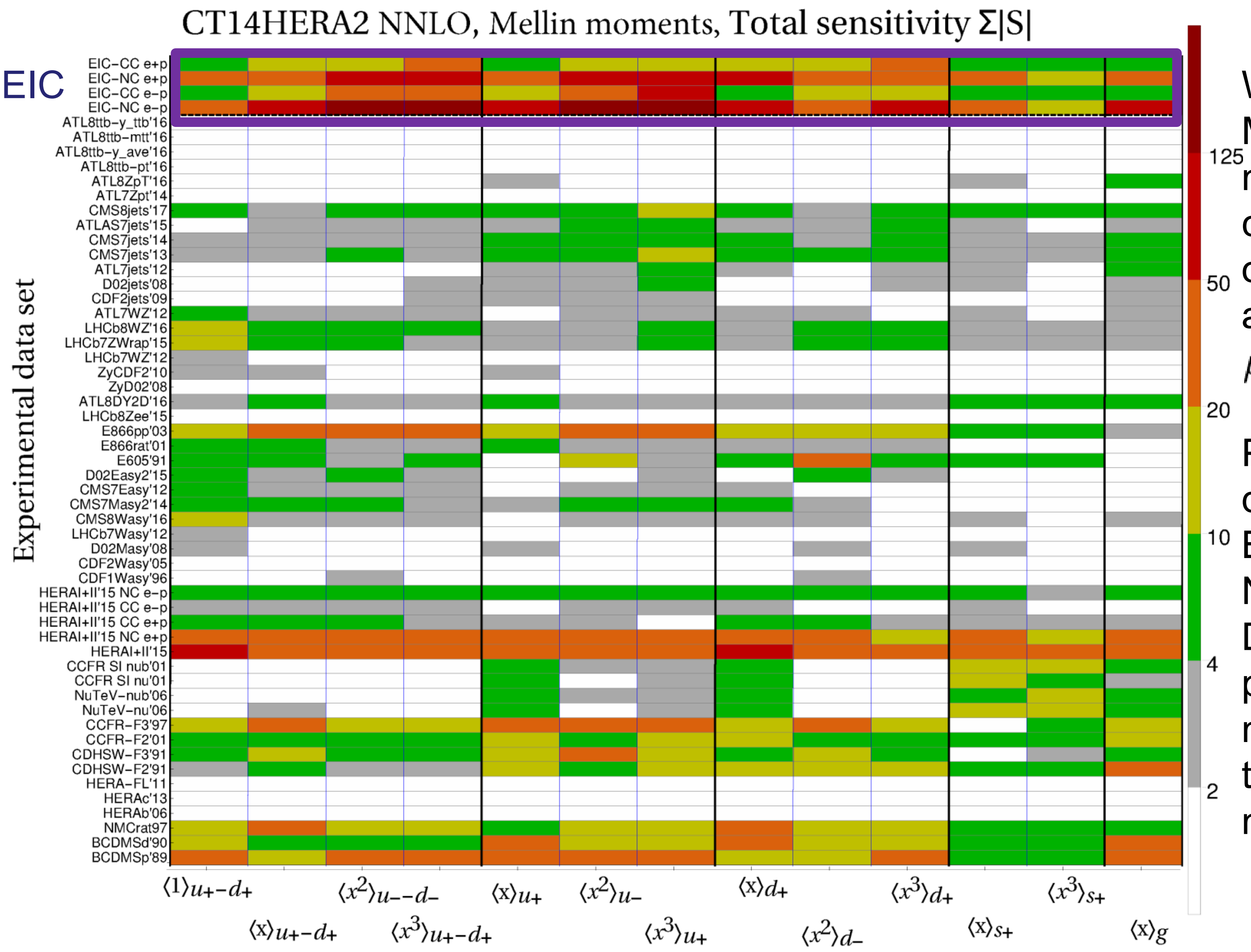
...a US-based EIC will also have important HEP consequences, e.g., on Higgs physics

- the impact of an EIC upon the theoretical predictions for inclusive Higgs production arises from a very broad region of the kinematical space it can access

- impact rather closely tied to that of the integrated gluon PDF:



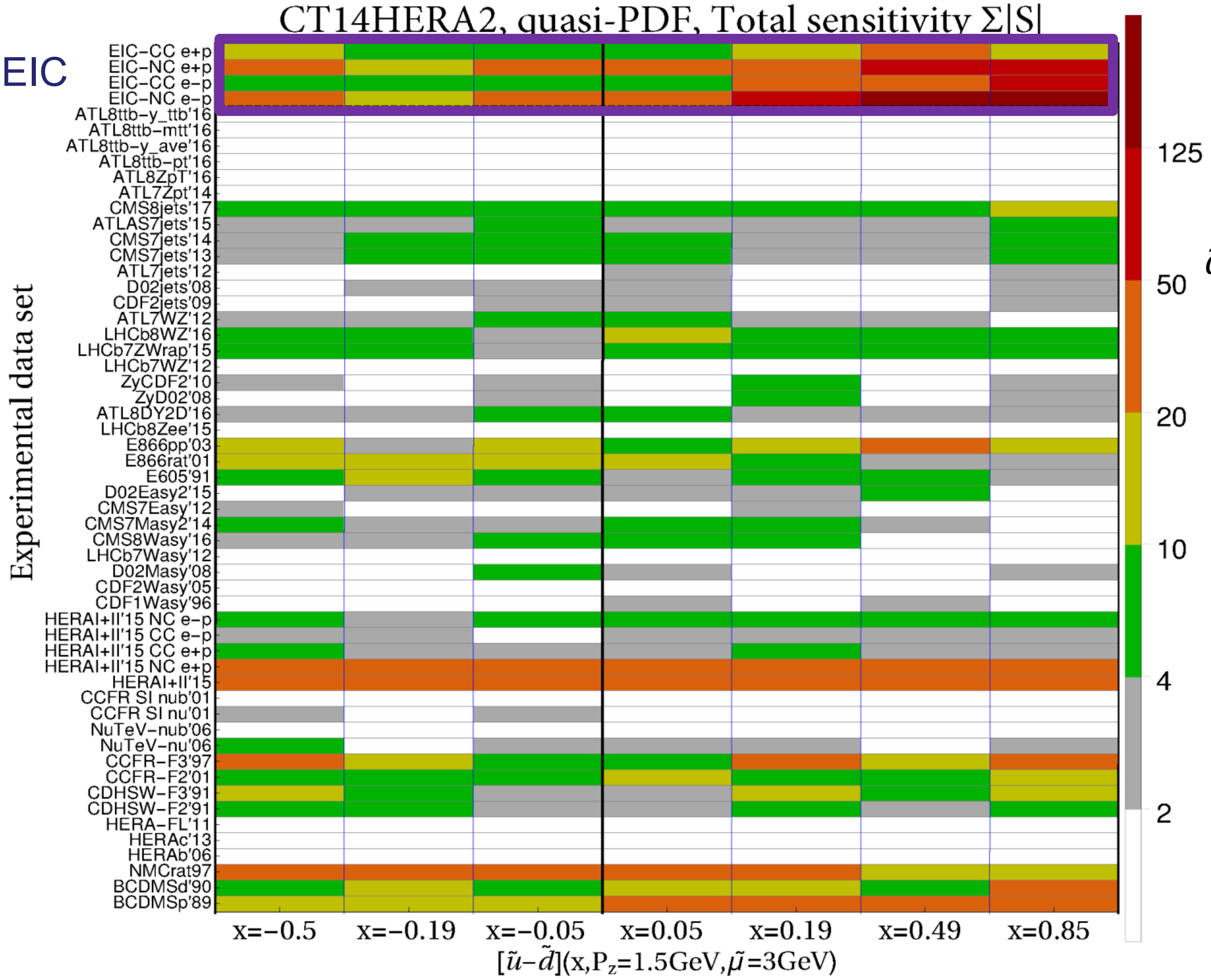
Total sensitivity to Mellin moments



We show Mellin moments computable on the lattice at the scale $\mu_F = 2\,\mathrm{GeV}$

Pre-2019 data: HERA, BCDMS, NMC, E866 DY pair production are most sensitive to the moments

Total sensitivity to lattice quasi-PDFs



$$\tilde{q}(x) \equiv \begin{cases} q(x), \\ x > 0; \\ -\bar{q}(|x|), \\ x < 0 \end{cases}$$

$Q = 3$ GeV,
$P_z = 1.5$ GeV

Key points: future experiments

- The HL-LHC, EIC, and LHeC have **complementary potentials,** make a powerful physics case in combination. Highlights:
  - **HL-LHC:** reach to high Q, a variety of processes sensitive to PDFs; high sensitivity to the gluon, u and d antiquarks
  - **LHeC:** reach to $x < 10^{-6}$; high sensitivity to small-x gluon, as well as d and s quarks (esp. at $x \to 1$) in the clean $e^{\pm}p$ scattering environment; independent of BSM physics
  - **EIC:** supercedes the bulk of fixed-target DIS experiments; offers unique reach in the region of $x > 0.1$ necessary for tests of lattice QCD and nonperturbative QCD models; nucleon and nuclear beams; flavor separation via SIDIS; polarized beams