



European Research Council
Established by the European Commission

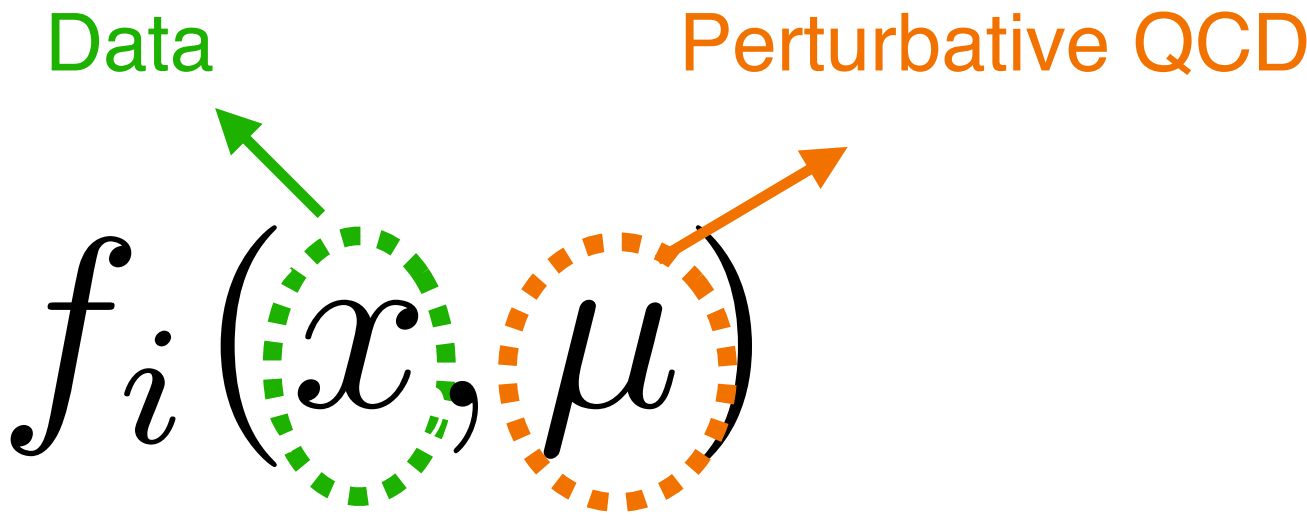


MARIA UBIALI
UNIVERSITY OF CAMBRIDGE

RECENT UPDATES ON PDF DETERMINATIONS

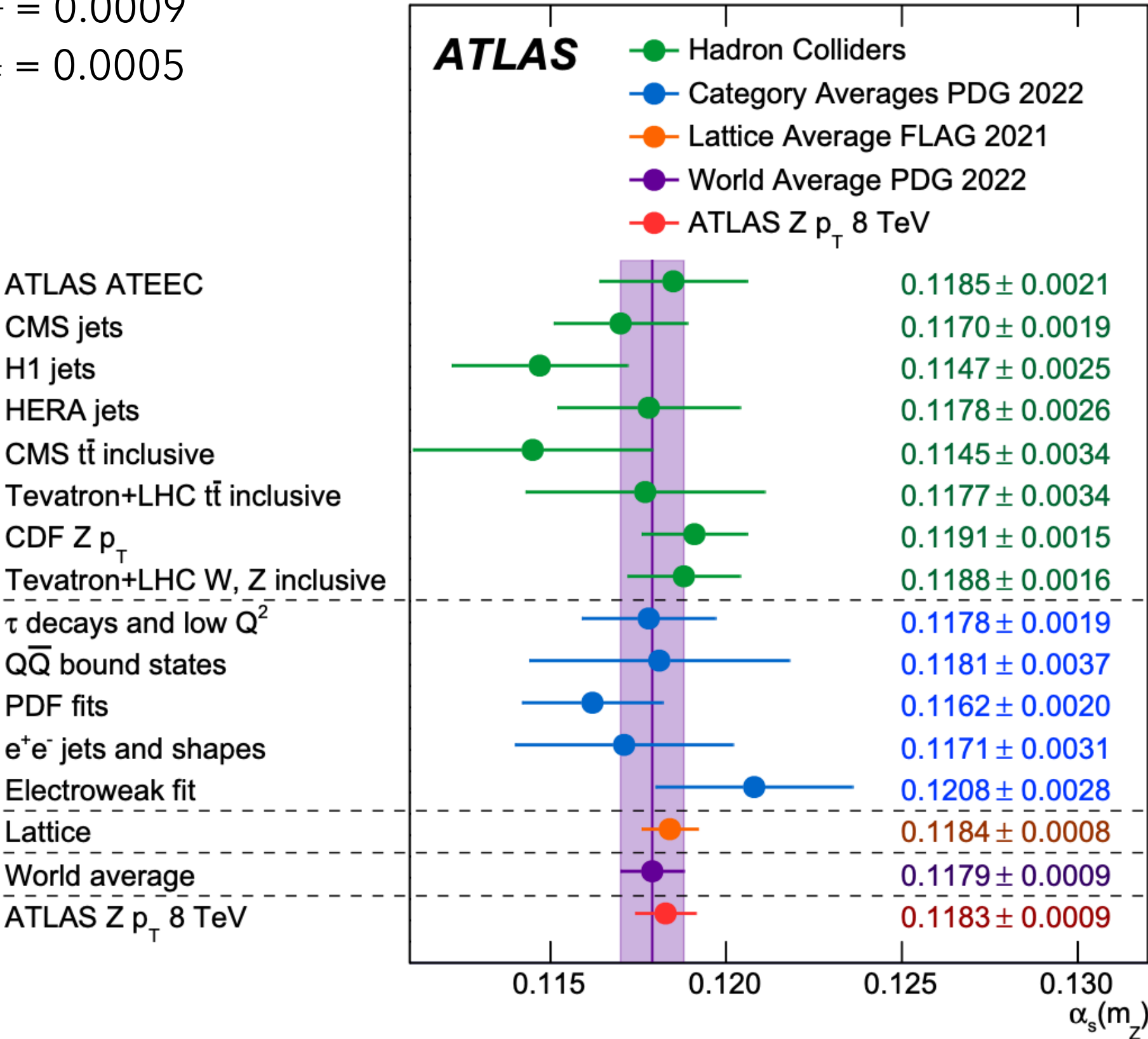
PDFs CRUCIAL INPUT OF LHC PHYSICS

$$d\sigma^{pp\rightarrow ab} = \sum_{i,j} f_i \otimes f_j \otimes d\hat{\sigma}^{ij\rightarrow ab} + \dots$$

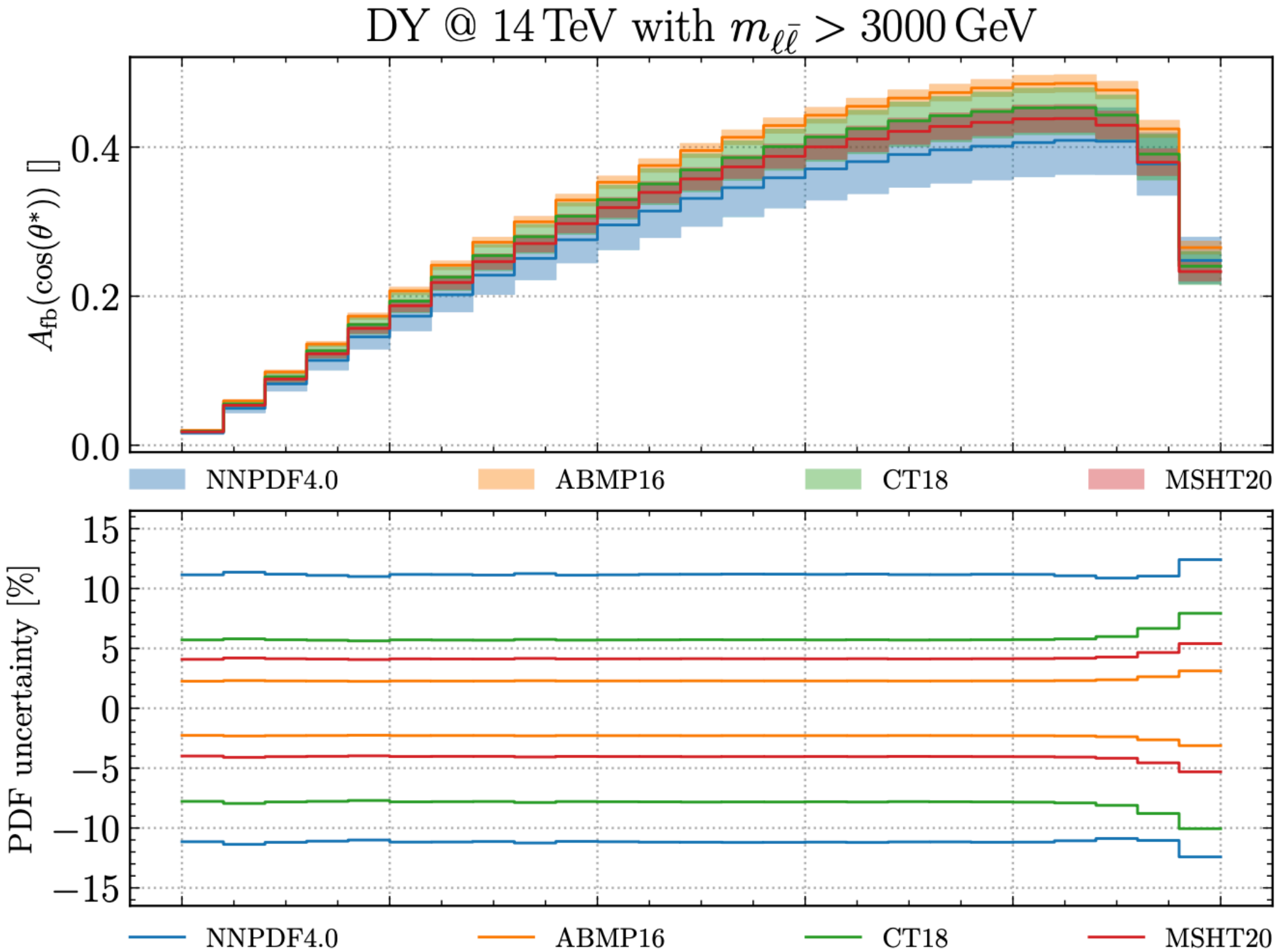


ATLAS α_s determination

$\Delta_{\text{TOT}} = 0.0009$
 $\Delta_{\text{PDF}} = 0.0005$



Heavy Z' search in DY Forward-Backward asymmetry



PDF DETERMINATION: PROGRESS AND STATUS

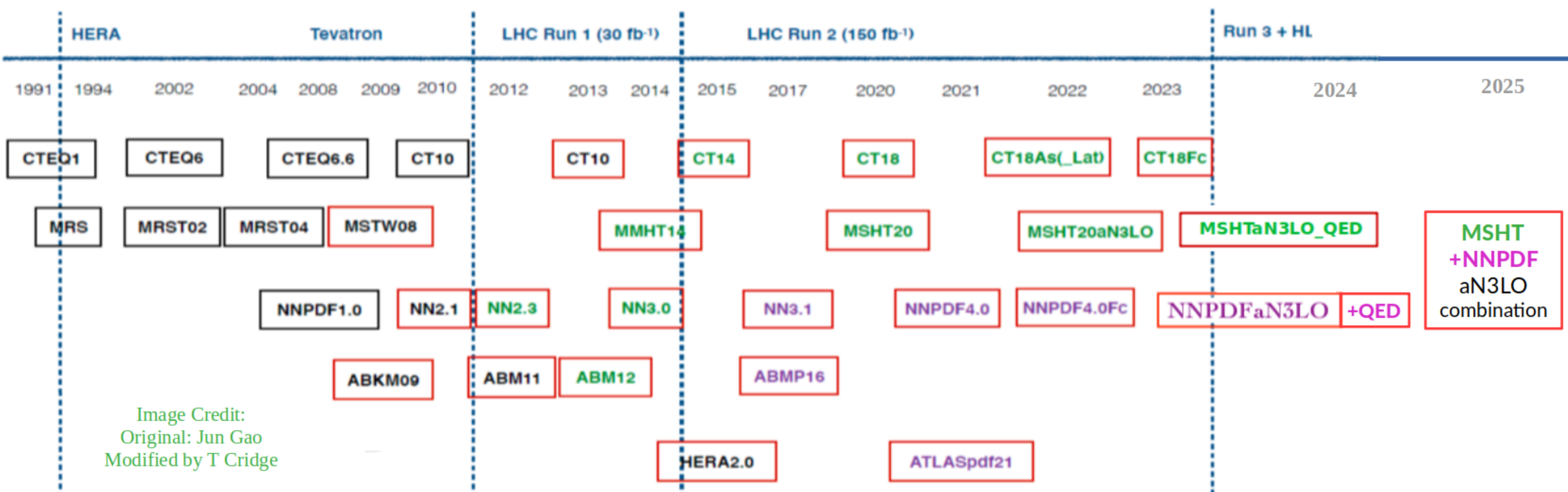
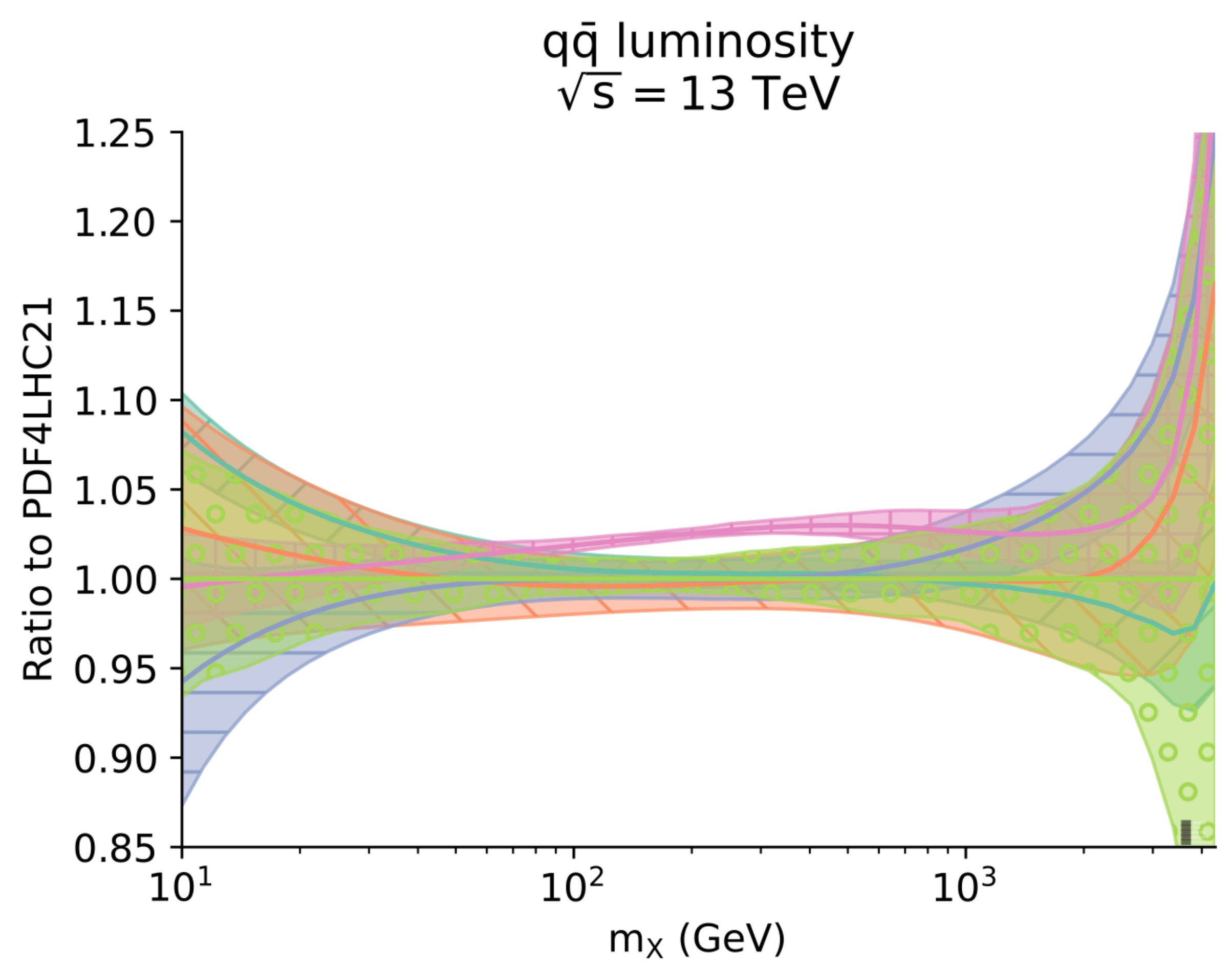
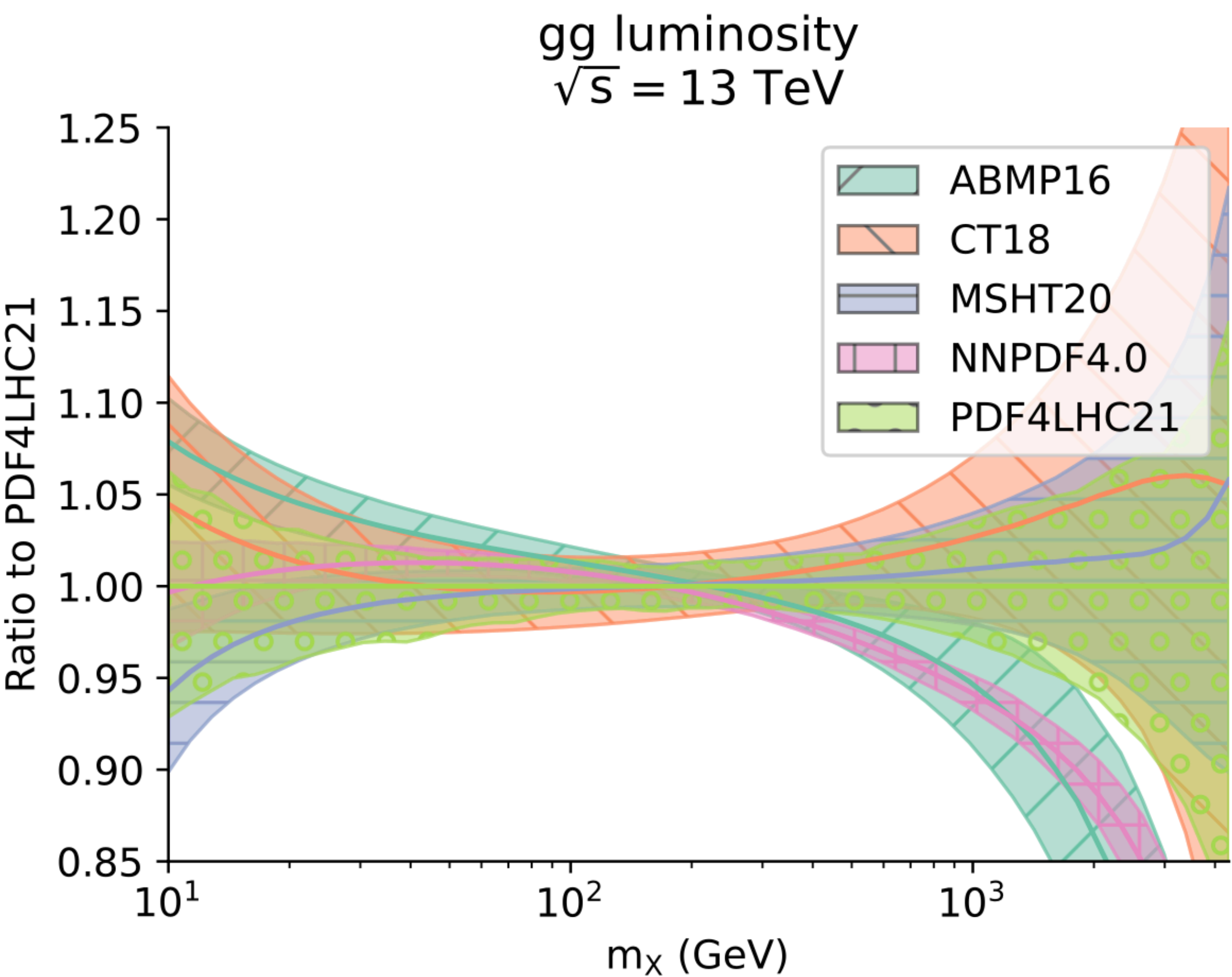


Image Credit:
Original: Jun Gao
Modified by T Cridge



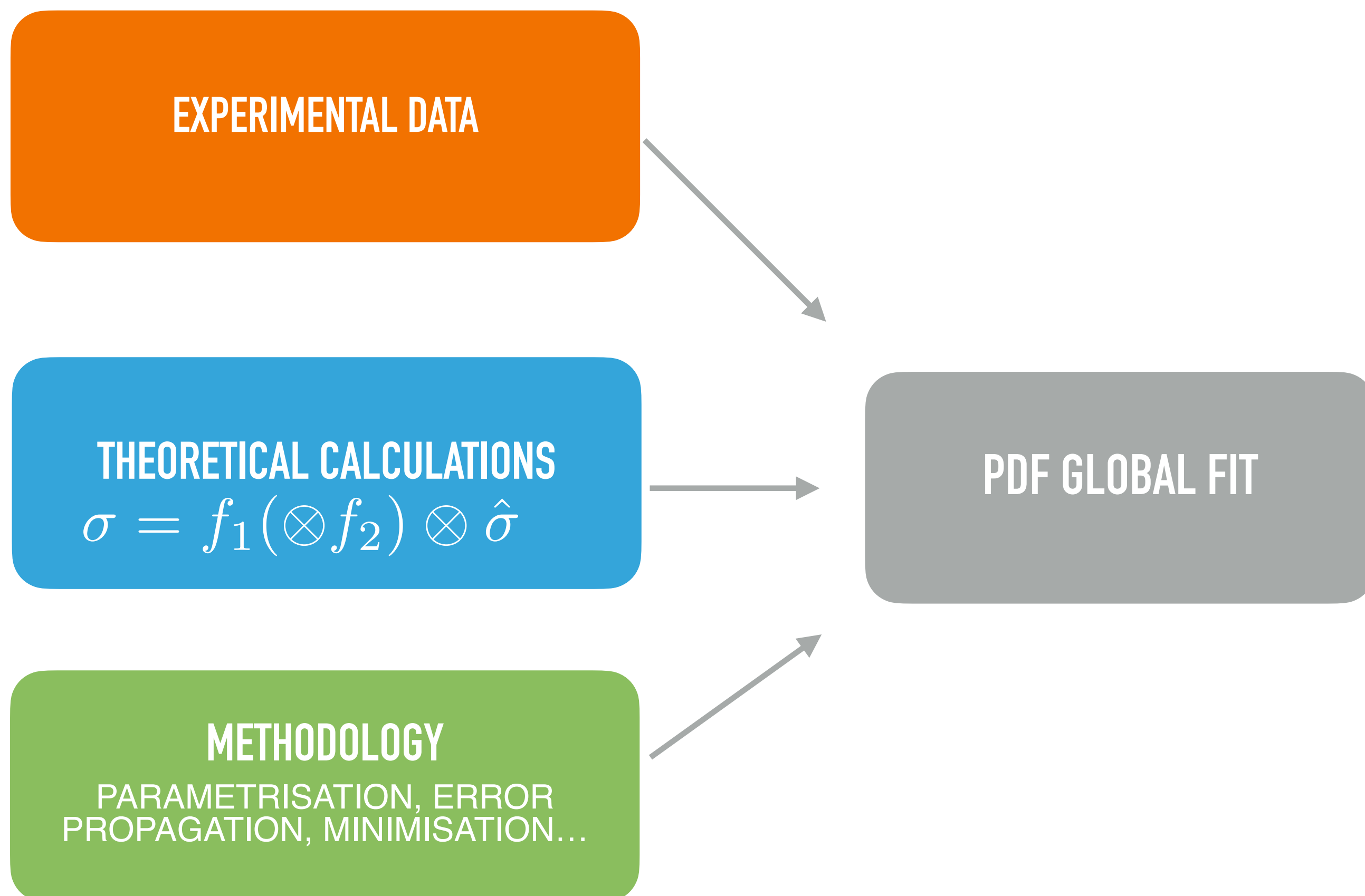


Parton luminosities:

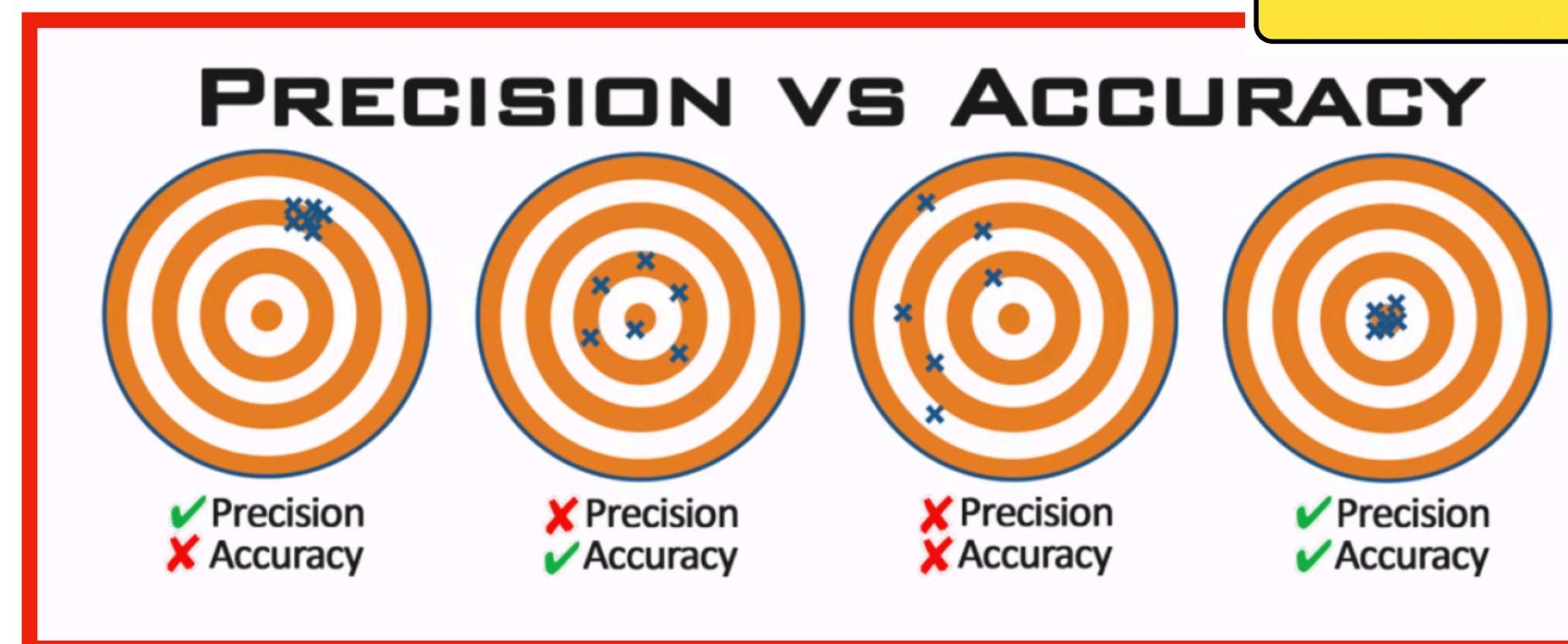
$$\Phi_{ab}(m_X, S) = \frac{1}{S} \int_{\tau}^1 \frac{dx}{x} f_a(x, m_X) f_b\left(\frac{\tau}{x}, m_X\right) \quad \tau = \frac{m_X^2}{S}$$

PDF DETERMINATION: PROGRESS AND STATUS

- Progress highlights
 - ➔ **Experimental data**: many more precise measurements constraining PDFs
 - ➔ **Theory**: heavy quark schemes, aN3LO PDFs, inclusion of MHOUs in PDF error bands, EW and QED corrections, initial photon...
 - ➔ **Methodology**: uncertainty propagation, parametrisation bias, effect of priors, closure tests,...



Precision and accuracy of PDF determinations has a limit that is determined by precision and accuracy of all ingredients entering a global fit



CURRENT SM PRECISION CHALLENGE

PDF set	$\alpha_s(m_Z)$	PDF uncertainty	g [GeV ²]	q [GeV ⁴]
MSHT20 [37]	0.11839	0.00040	0.44	−0.07
NNPDF4.0 [84]	0.11779	0.00024	0.50	−0.08
CT18A [29]	0.11982	0.00050	0.36	−0.03
HERAPDF2.0 [65]	0.11890	0.00027	0.40	−0.04

ATLAS collaboration, arXiv:2309.12986

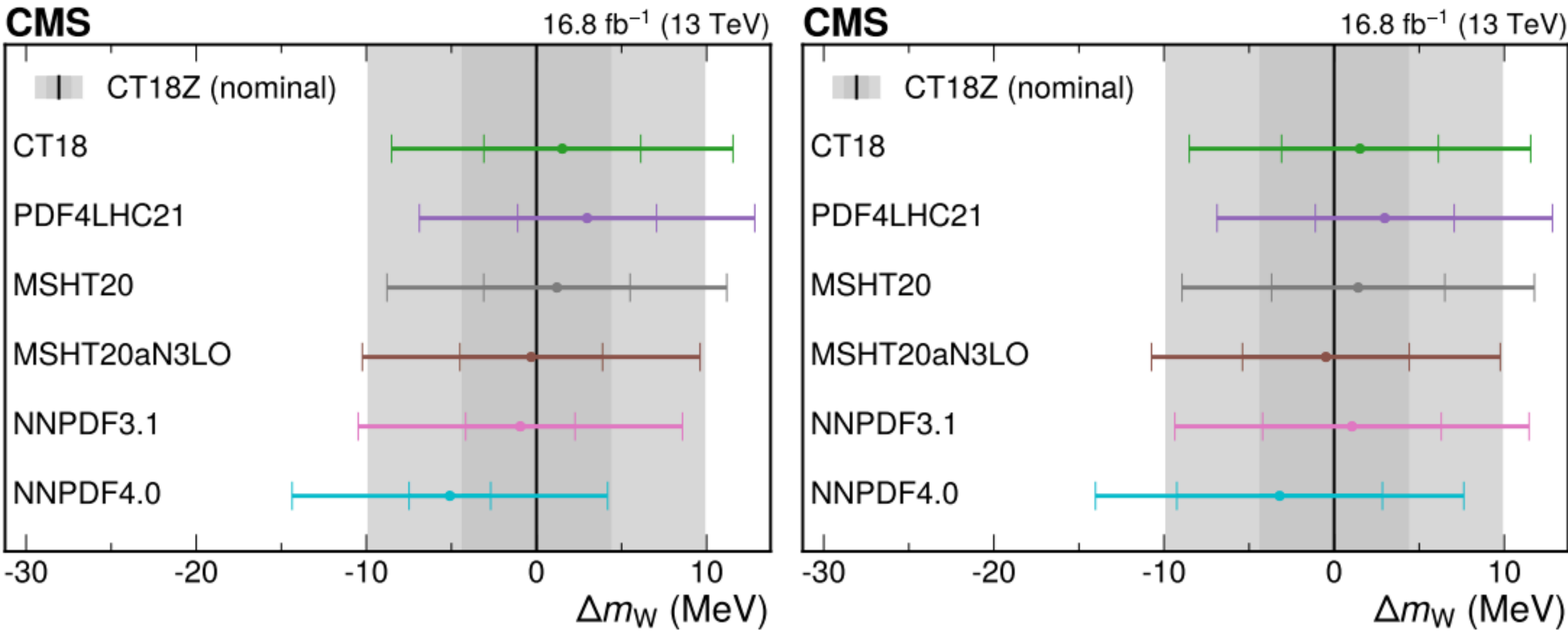
Significantly different results obtained when using different input PDFs.

- ATLAS determination of α_s from neutral-current Drell-Yan Z pT measurement most precise α_s determination ever performed from a single experiment
 $\Delta_{\alpha,\text{TOT}} = 9 \times 10^{-4}$ with $\Delta_{\alpha,\text{PDF}} = 5 \times 10^{-4}$ (MSHT20aN3LO)
But
 $\delta_{\alpha,\text{PDF}} = 20 \times 10^{-4}$ (shift CT18A versus NNPDF4.0)

CURRENT SM PRECISION CHALLENGE

PDF set	$\alpha_s(m_Z)$	PDF uncertainty	g [GeV ²]	q [GeV ⁴]
MSHT20 [37]	0.11839	0.00040	0.44	-0.07
NNPDF4.0 [84]	0.11779	0.00024	0.50	-0.08
CT18A [29]	0.11982	0.00050	0.36	-0.03
HERAPDF2.0 [65]	0.11890	0.00027	0.40	-0.04

ATLAS collaboration, arXiv:2309.12986



CMS collaboration, arXiv: 2412.13872

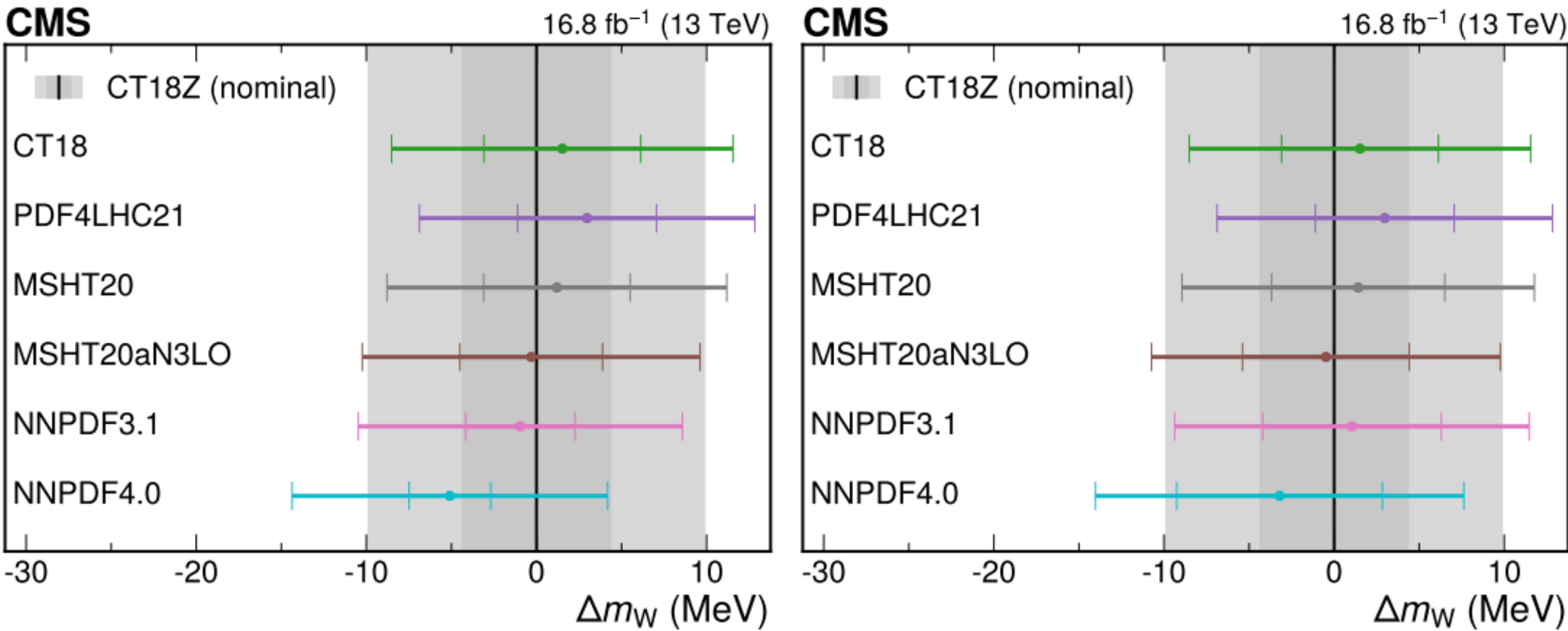
Significantly different results obtained when using different input PDFs.

- ATLAS determination of α_s from neutral-current Drell-Yan Z pT measurement most precise α_s determination ever performed from a single experiment
 $\Delta_{\alpha, \text{TOT}} = 9 \times 10^{-4}$ with $\Delta_{\alpha, \text{PDF}} = 5 \times 10^{-4}$ (MSHT20aN3LO)
But
 $\delta_{\alpha, \text{PDF}} = 20 \times 10^{-4}$ (shift CT18A versus NNPDF4.0)
- Similar situation in M_W determination [ATLAS-CONF-2023-004]
 $\Delta_{M_W, \text{PDF}} = 7.7$ (14.6) MeV in $p_{T, l}(m_\tau)$ channel (CT18)
But
 $\delta_{M_W, \text{PDFS}} = 17$ (21) MeV (shift MSHT20 versus NNPDF4.0)
- CMS M_W : inflate PDF uncertainties by a factor and profile with CMS data decreases shift between PDF set (but effectively giving much larger weight to a given measurement as compared to all experiments that determine PDFs in global fit)

CURRENT SM PRECISION CHALLENGE

PDF set	$\alpha_s(m_Z)$	PDF uncertainty	g [GeV ²]	q [GeV ⁴]
MSHT20 [37]	0.11839	0.00040	0.44	-0.07
NNPDF4.0 [84]	0.11779	0.00024	0.50	-0.08
CT18A [29]	0.11982	0.00050	0.36	-0.03
HERAPDF2.0 [65]	0.11890	0.00027	0.40	-0.04

ATLAS collaboration, arXiv:2309.12986



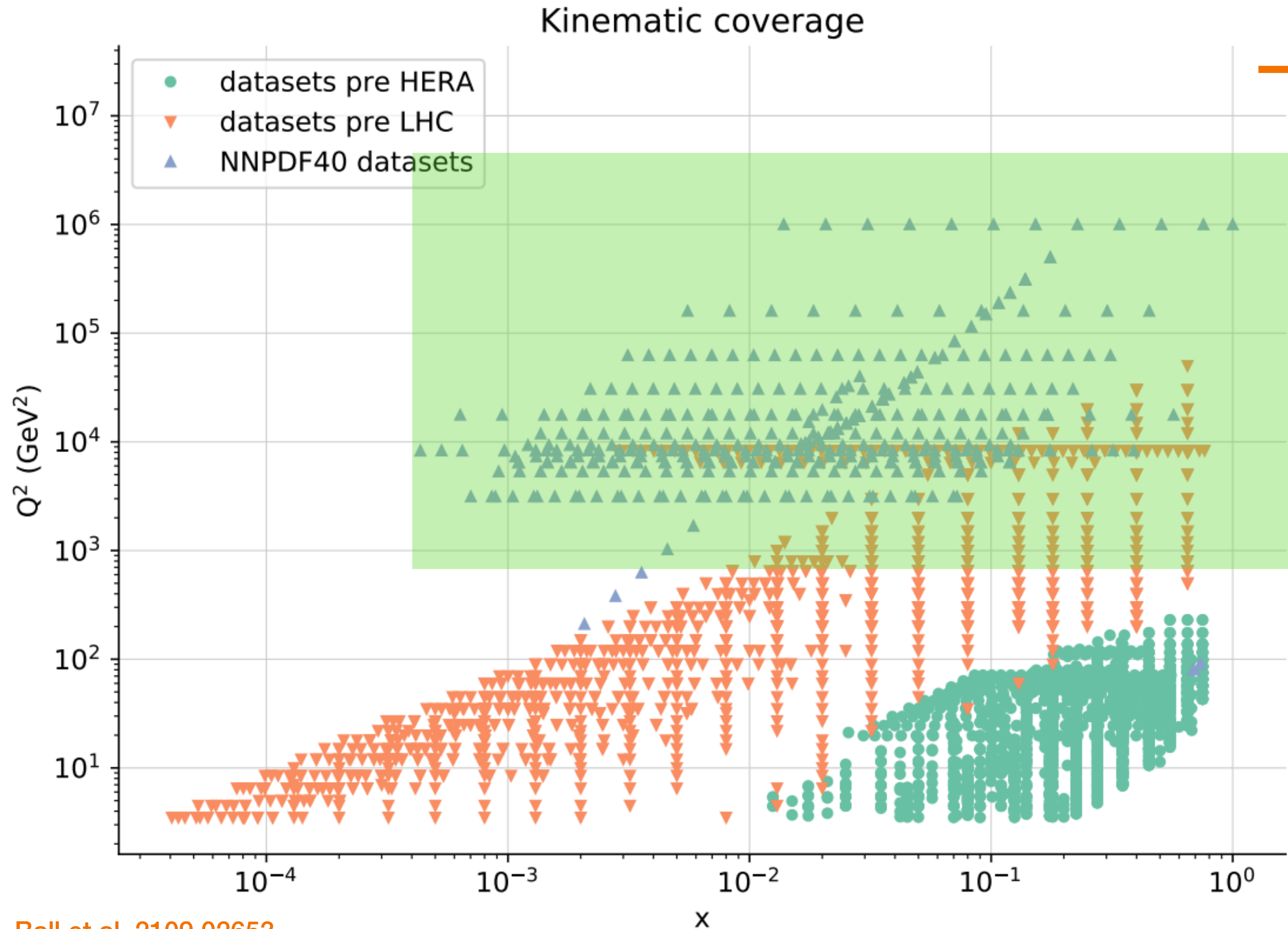
CMS collaboration, arXiv: 2412.13872

How to improve this situation?

Significantly different results obtained when using different input PDFs.

- ATLAS determination of α_s from neutral-current Drell-Yan Z pT measurement most precise α_s determination ever performed from a single experiment
 $\Delta_{\alpha, \text{TOT}} = 9 \times 10^{-4}$ with $\Delta_{\alpha, \text{PDF}} = 5 \times 10^{-4}$ (MSHT20aN3LO)
But
 $\delta_{\alpha, \text{PDF}} = 20 \times 10^{-4}$ (shift CT18A versus NNPDF4.0)
- Similar situation in M_W determination [ATLAS-CONF-2023-004]
 $\Delta_{M_W, \text{PDF}} = 7.7$ (14.6) MeV in $p_{T, l}(m_\tau)$ channel (CT18)
But
 $\delta_{M_W, \text{PDFS}} = 17$ (21) MeV (shift MSHT20 versus NNPDF4.0)
- CMS M_W : inflate PDF uncertainties by a factor and profile with CMS data decreases shift between PDF set (but effectively giving much larger weight to a given measurement as compared to all experiments that determine PDFs in global fit)

1) Experimental data: status and frontiers

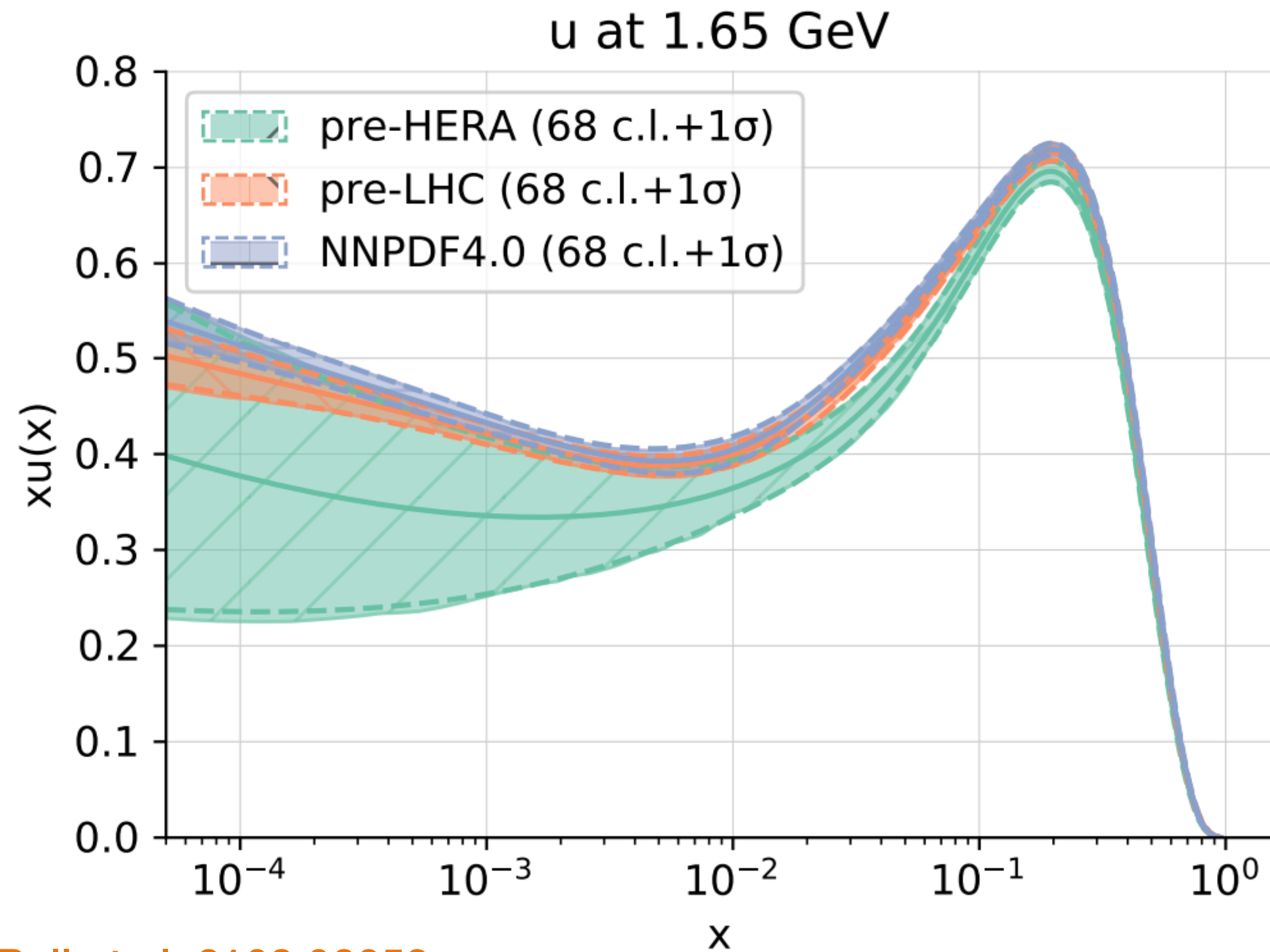


NNPDF4.0:
O(5000) datapoints,
about 30% from LHC

- Modern data show increasingly small statistical uncertainty and dominating correlated systematic uncertainties make global PDF fits challenging

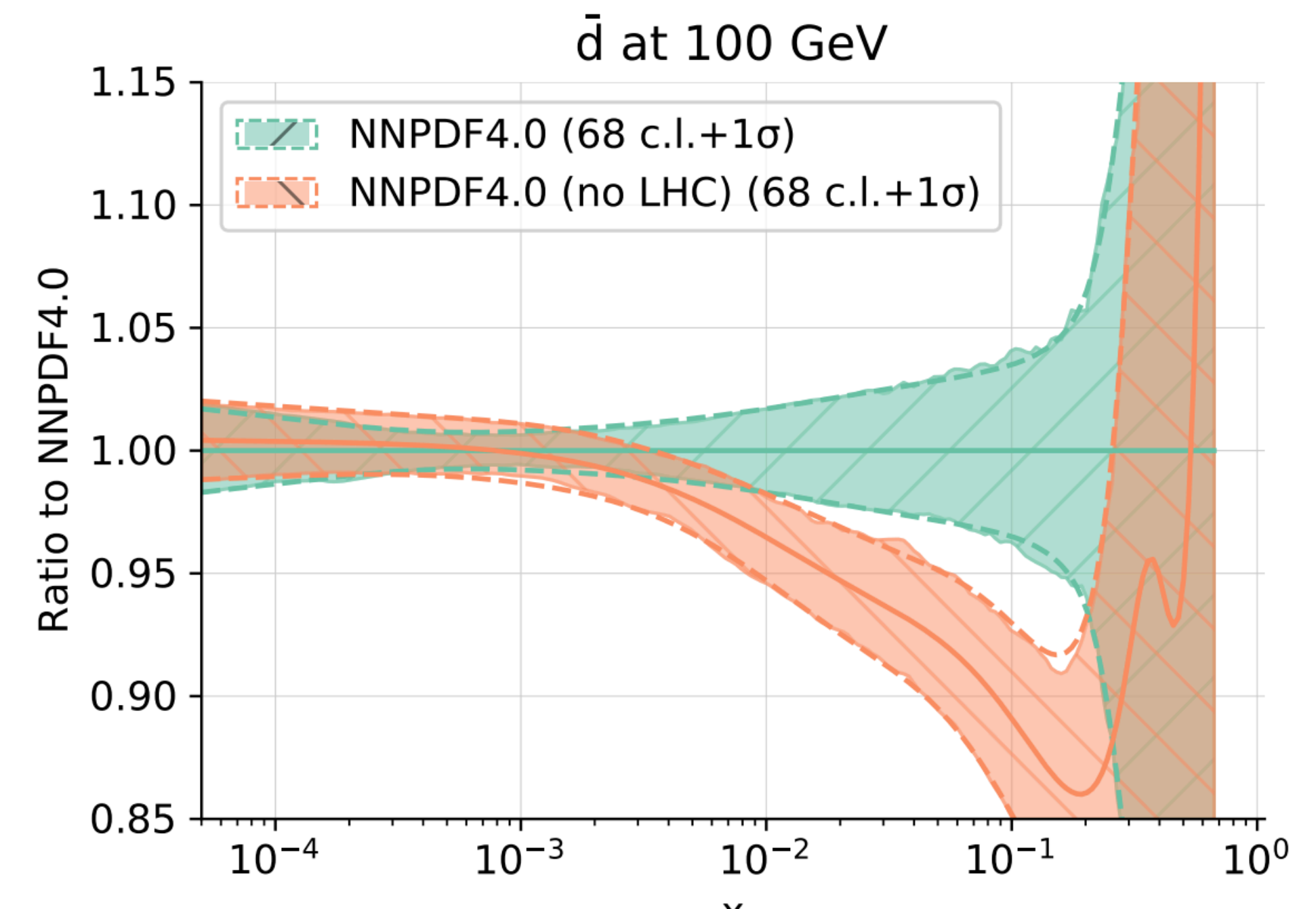
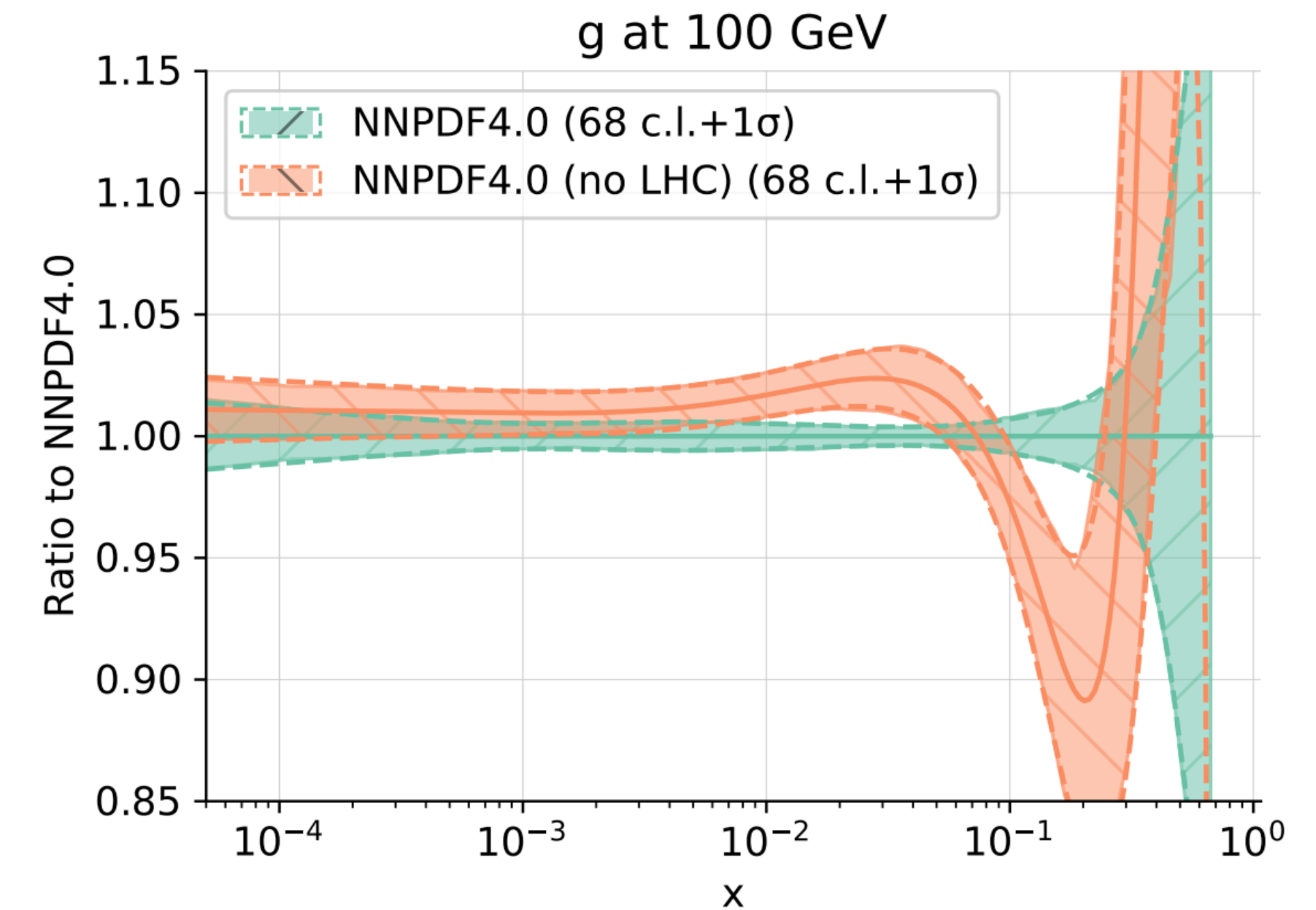
Kassabov, Nocera, Wilson 2207.00690
Bailey, Harland-Lang 1909.10541

EXPERIMENTAL DATA

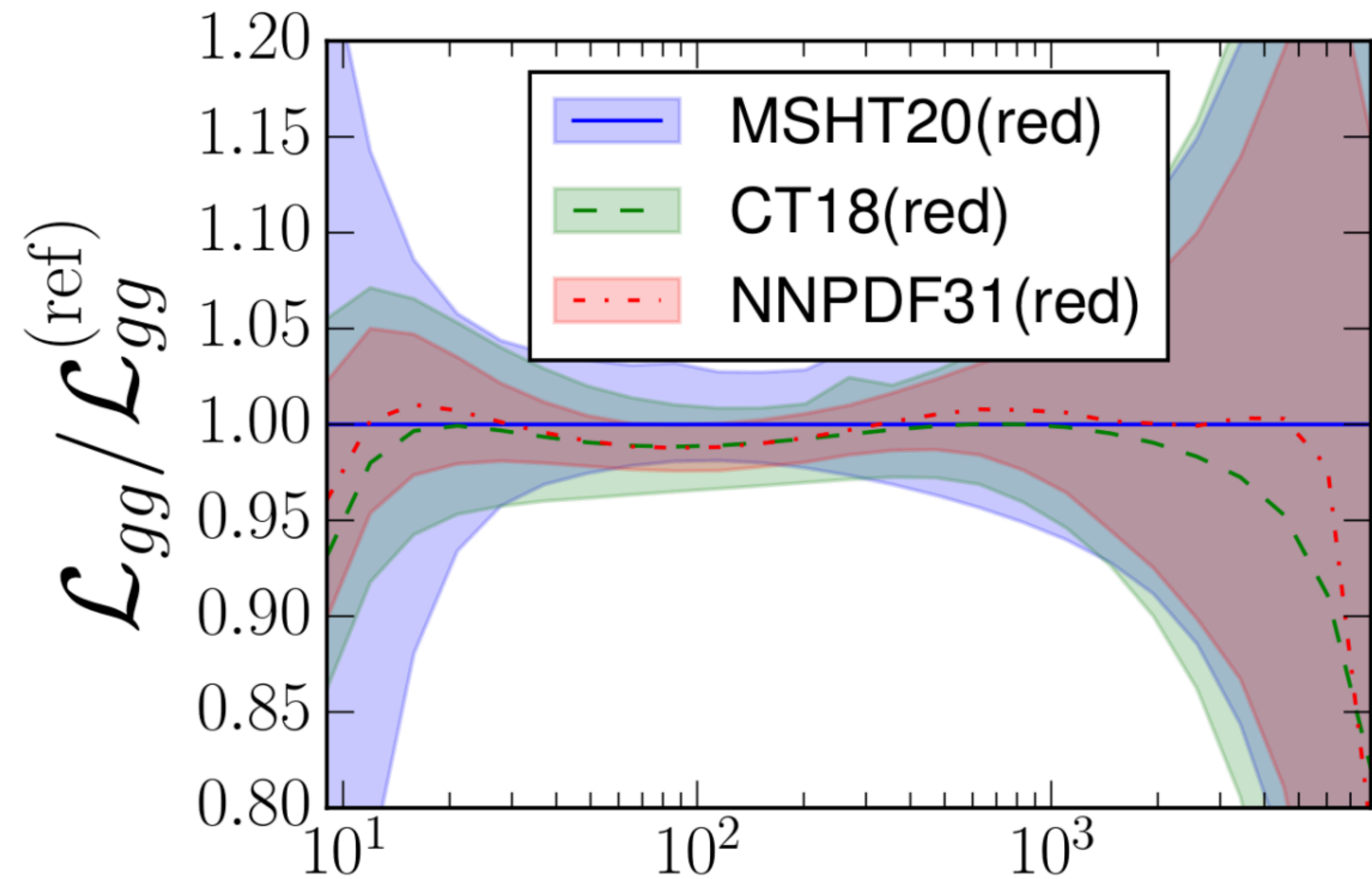
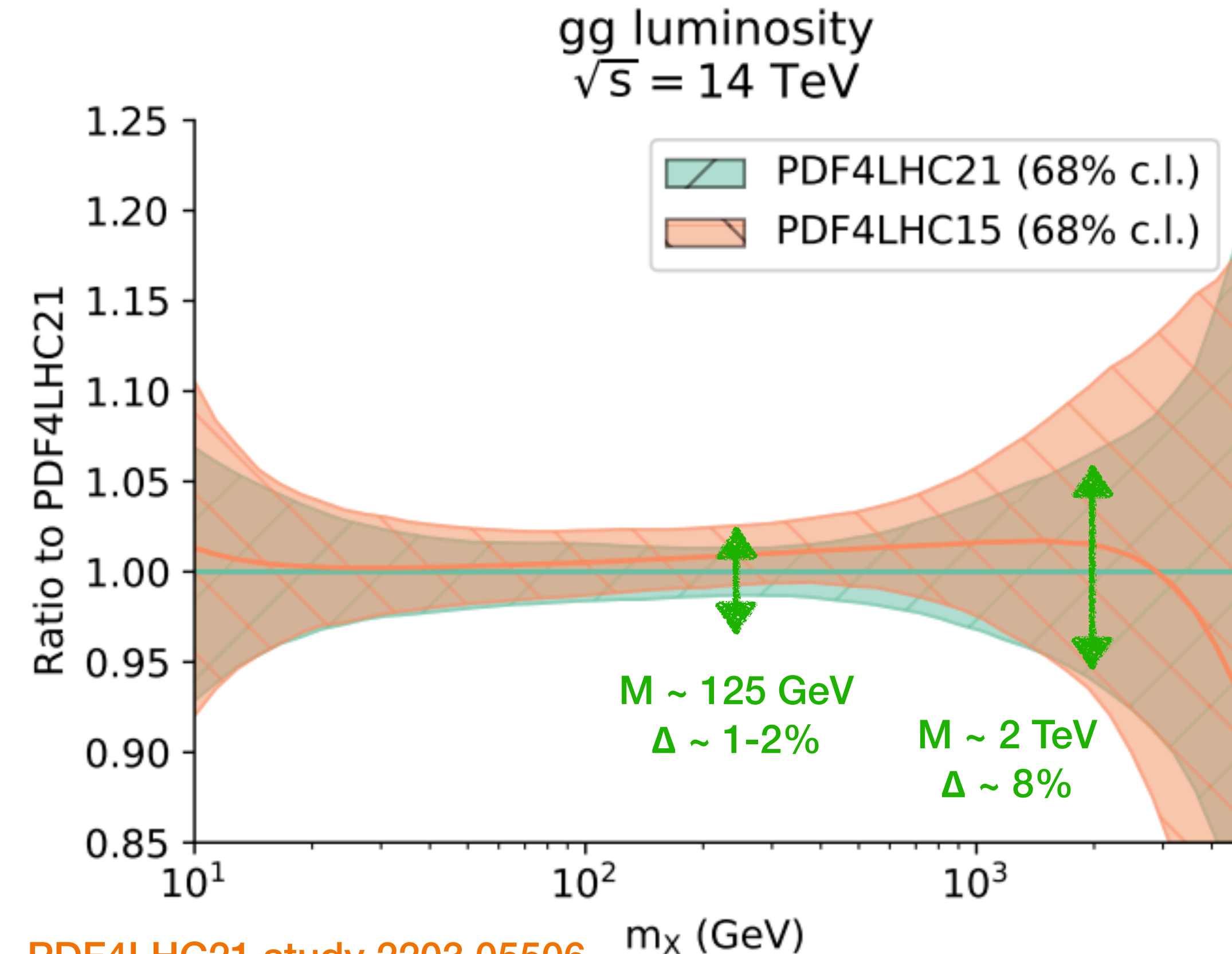


Ball et al, 2109.02653

- **HERA** data crucial to constrain quark valence (up and to less extent down valence) across intermediate to small x and gluon at small x
- **LHC** high energy data crucial to provide additional constraints to PDFs, in particular in medium- to large- x gluon and quarks.
- Some **tension** with older fixed-target Drell-Yan and DIS data visible in the large- x region (especially gluon and anti-quarks)



DATA-DRIVEN PROGRESS



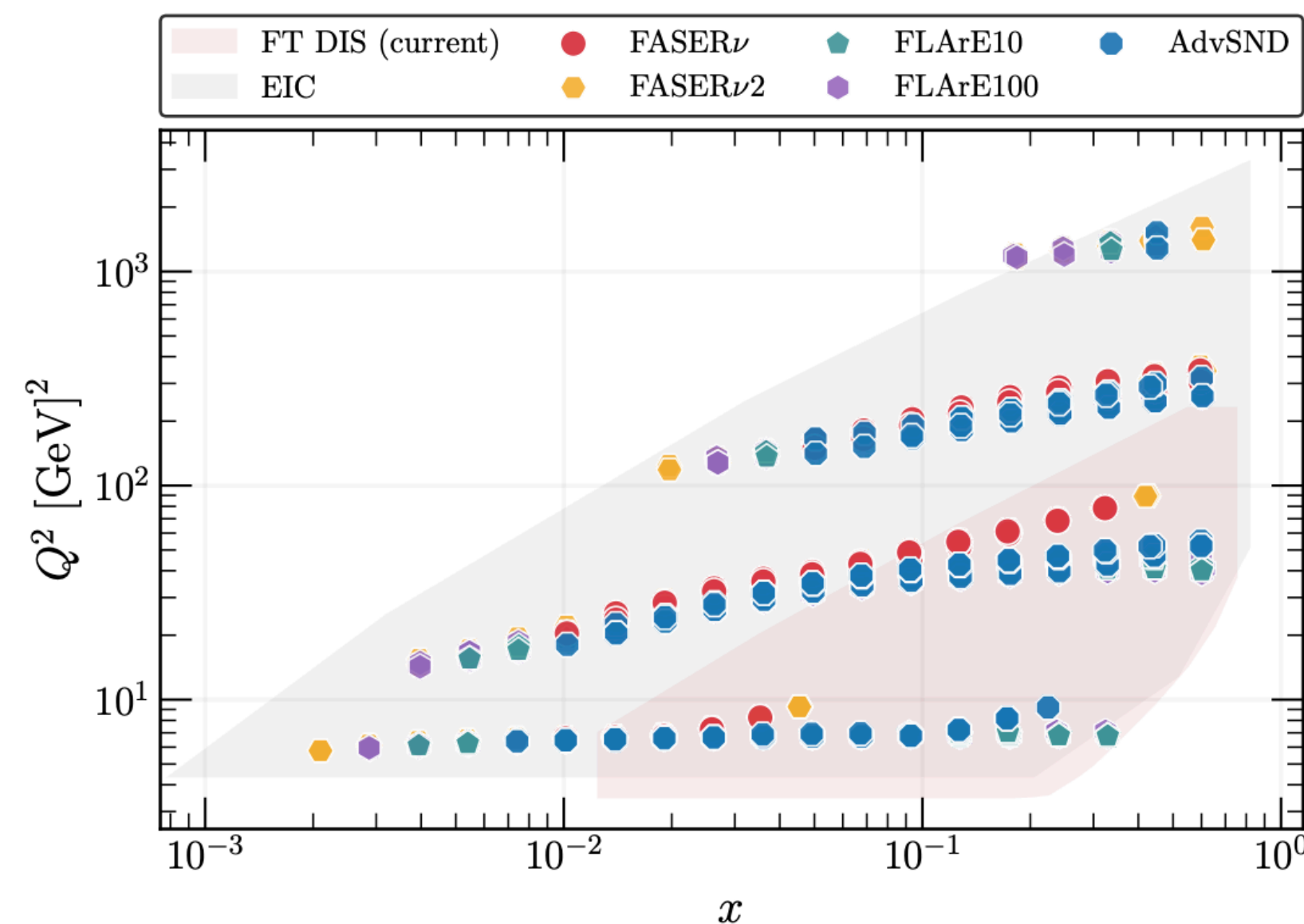
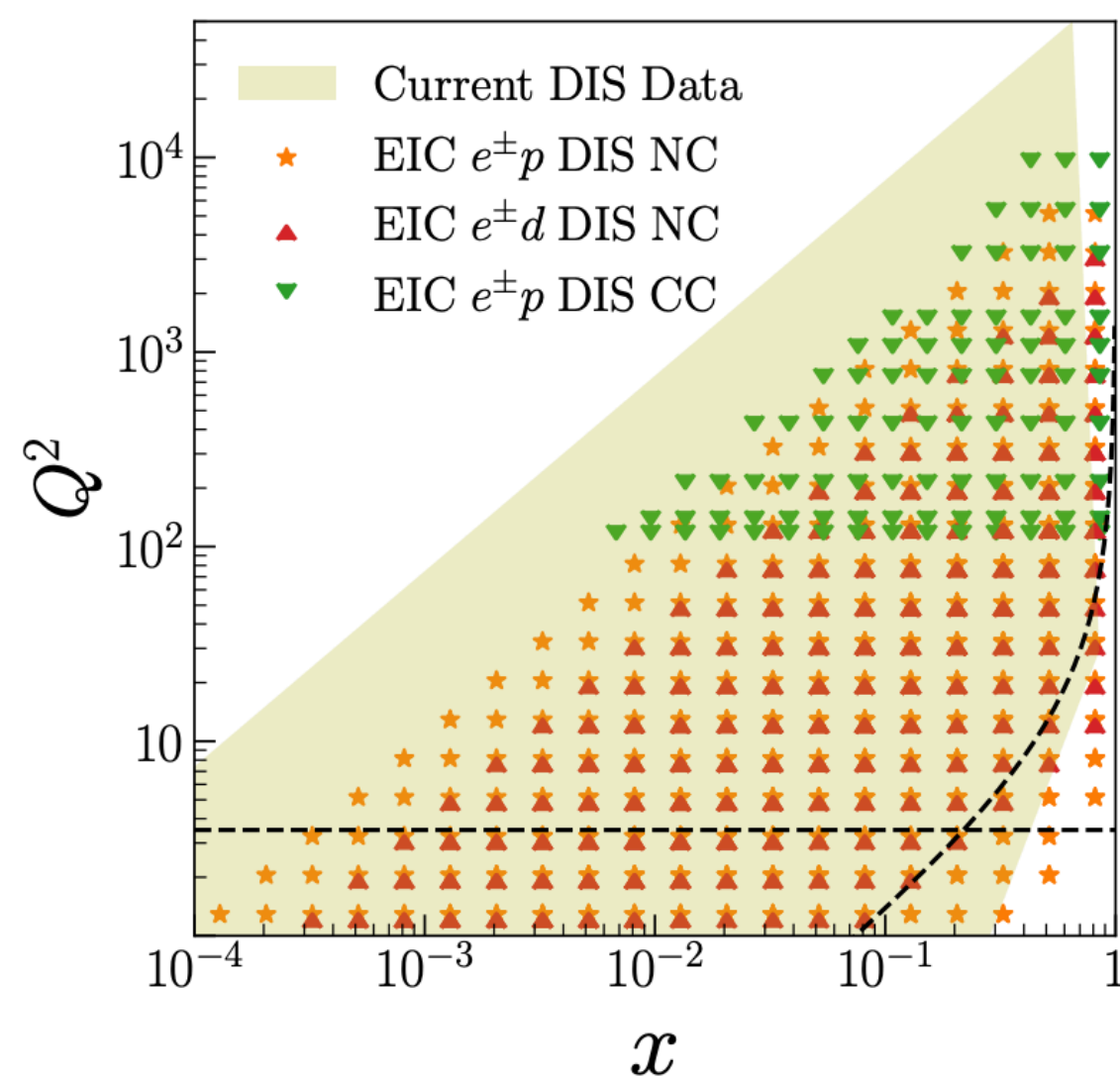
- In recent updates from global PDF fitting collaborations **effect of LHC data driving PDF uncertainties down**
→ PDF4LHC21 combination of NNPDF31', CT18' and MSHT20' smaller uncertainties than PDF4LHC15
- Benchmark among NNPDF3.1, MSHT20 and CT18: overall agreement, which improves once common dataset is used, **differences in uncertainties** associated with methodologies with $\Delta_{\text{CT18}} \gtrsim \Delta_{\text{MSHT20}} \gtrsim \Delta_{\text{NNPDF31}}$
- Increased precision of the data and their strong correlation demands methodological improvements (e. g. **more flexible parametrisation** that shifts PDFs outside nominal error band) and **>1 PDF fits** to accomodate some tension (CT18 vs CT18Z)

FRONTIER: LARGE-X PDFs AND HIGH-ENERGY DATA

- Large-x PDF affected by large uncertainties
- Reducing PDF uncertainties in large-x region is essential to characterise any possible new physics and to interpret deviations in the tails of experimental distributions
- Looking forward: **HL-LHC (and FCC-hh)**

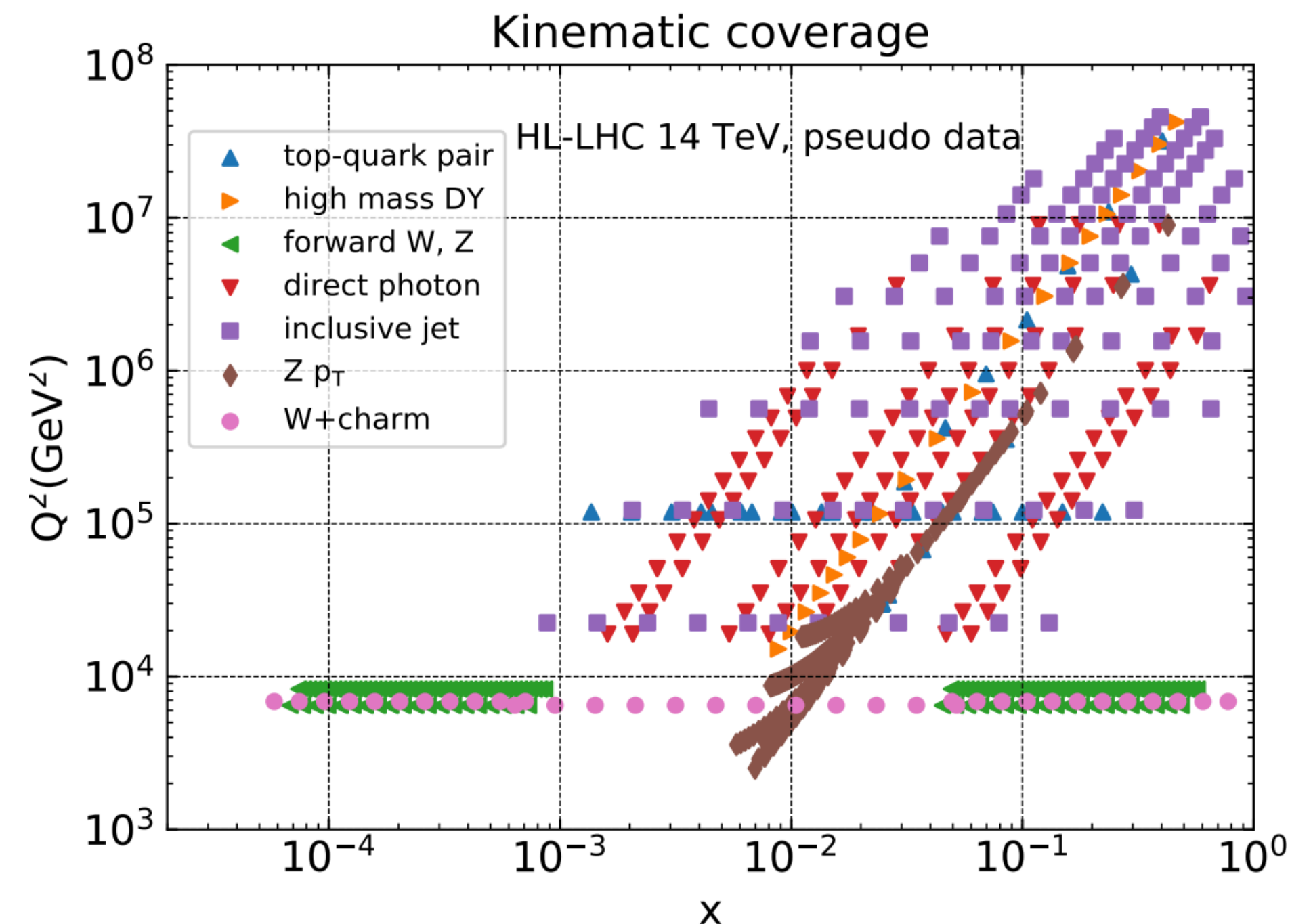
Large $x \leftrightarrow$ Large M and/or Large Y

$$x_{1,2} \approx \frac{M}{\sqrt{S}} e^{\pm Y}$$



Khalek et al, 2102.00018

J Cruz-Martinez et al 2309.09581



Khalek et al, 1810.03639

- Looking forward:
EIC, Forward Physics Facilities, Lattice...

Large $x \leftrightarrow$ Lower S

FRONTIER: PDFs AND BSM INTERPLAY

10/19

- Assume you generate synthetic data in PDF fit assuming a given BSM model but fit PDFs assuming SM
- Can PDFs absorb signs of new physics? In some model **YES**

$$\sigma_{obs} = \hat{\sigma}_{BSM} \otimes \mathcal{L}_{SM} \approx \hat{\sigma}_{SM} \otimes \mathcal{L}_{BSM-biased}$$

- Serious consequences if BSM-induced bias on PDFs undetected: missing indirect signs of BSM or see spurious deviations
- Important to tackle the issue:
 - ➔ Explore conservative fits
 - ➔ Simultaneous SMEFT&PDF fits
 - ➔ Complementary large-x constraints (EIC, FPF, lattice)
 - ➔ Broaden understanding of interplay to more realistic BSM models
 - ➔ Explore more sectors besides Drell-Yan and top (jets, di-jets, flavor..)

Cole et al 2602.20235

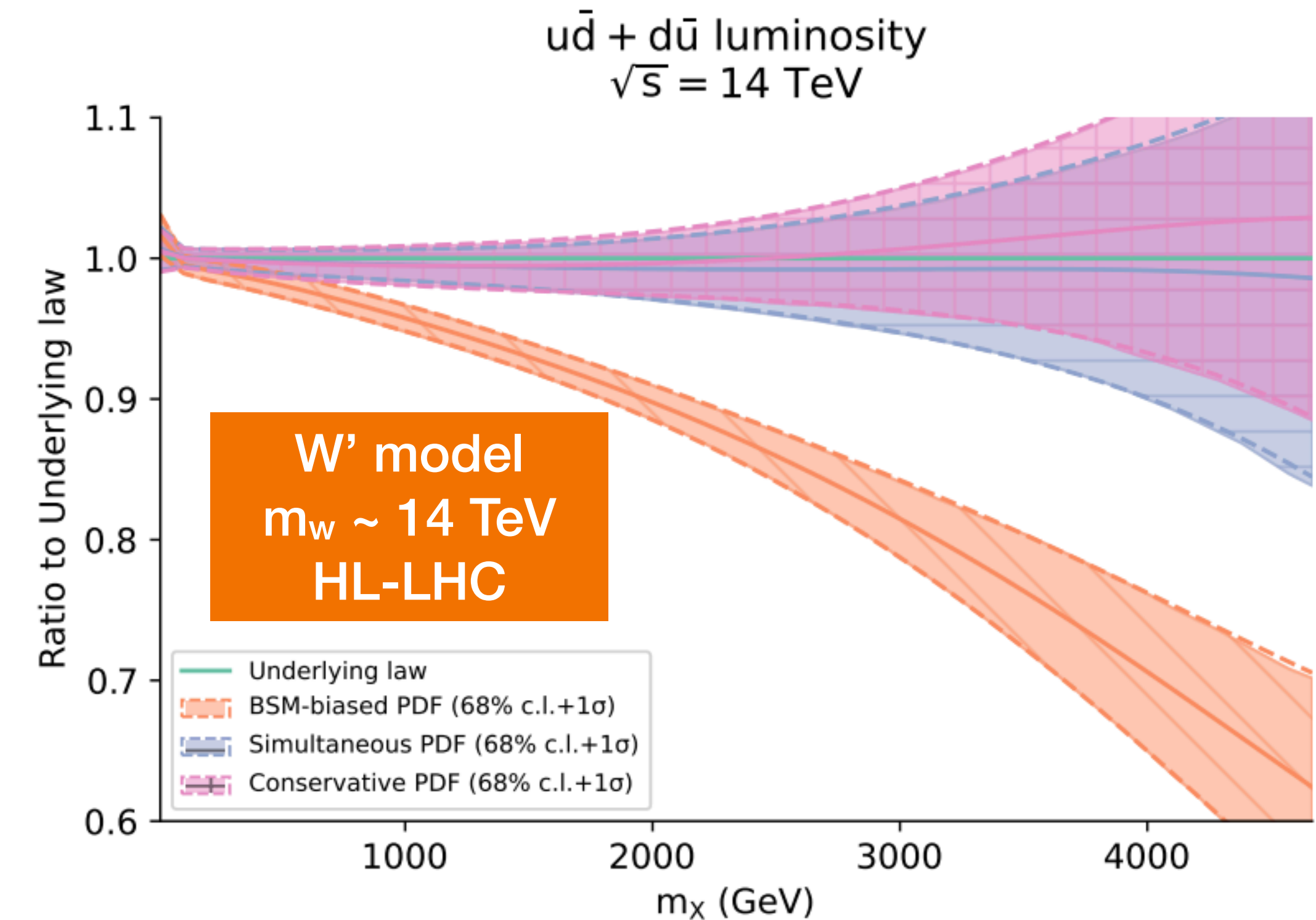
Hammou et al 2307.10370, 2410.00963

Costantini et al 2402.03308

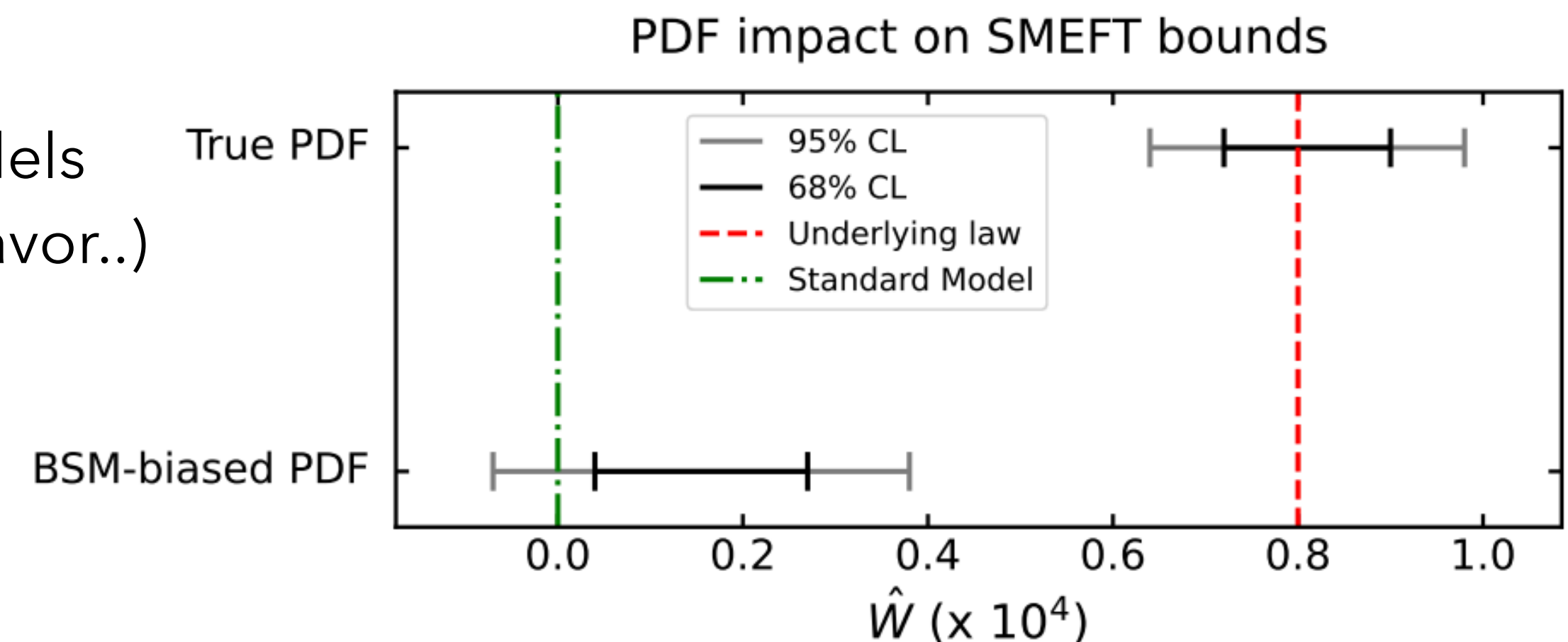
Kassabov et al 2303.06159

Greljo et al 2104.02723

Iranipour et al 2201.07240 ...

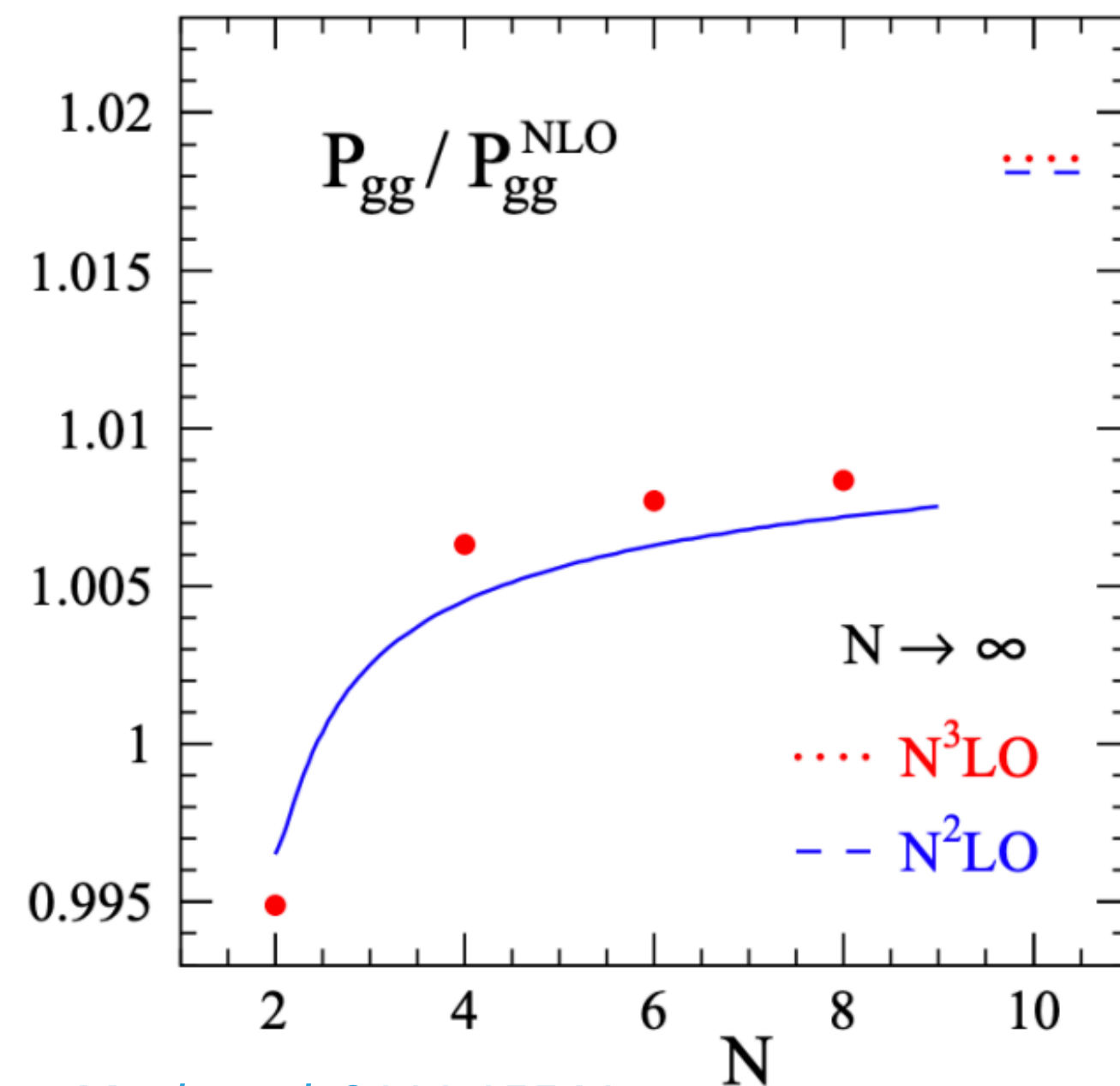


Cole et al 2602.20235



2) Theory: status and frontiers

PROGRESS: THEORY



Moch et al, 2111.15561

First approximate N3LO PDF sets by MSHT and NNPDF

[McGowan et al 2207.04739, Ball et al:2402.18635]

- Order-by-order convergence improved in both PDF sets and in phenomenological predictions
- Rather good agreement between collaborations' results considering different input included and strategies for the approximation

DGLAP
splitting
functions

$$P_{ij}(x, \alpha_s) = \alpha_s P_{ij}^{(0)}(x) + \alpha_s^2 P_{ij}^{(1)}(x) + \alpha_s^3 P_{ij}^{(2)}(x) + \alpha_s^4 P_{ij}^{(3)}(x) + \alpha_s^5 P_{ij}^{(4)}(x)$$

✓ LO (1973)

✓ NLO (1982)

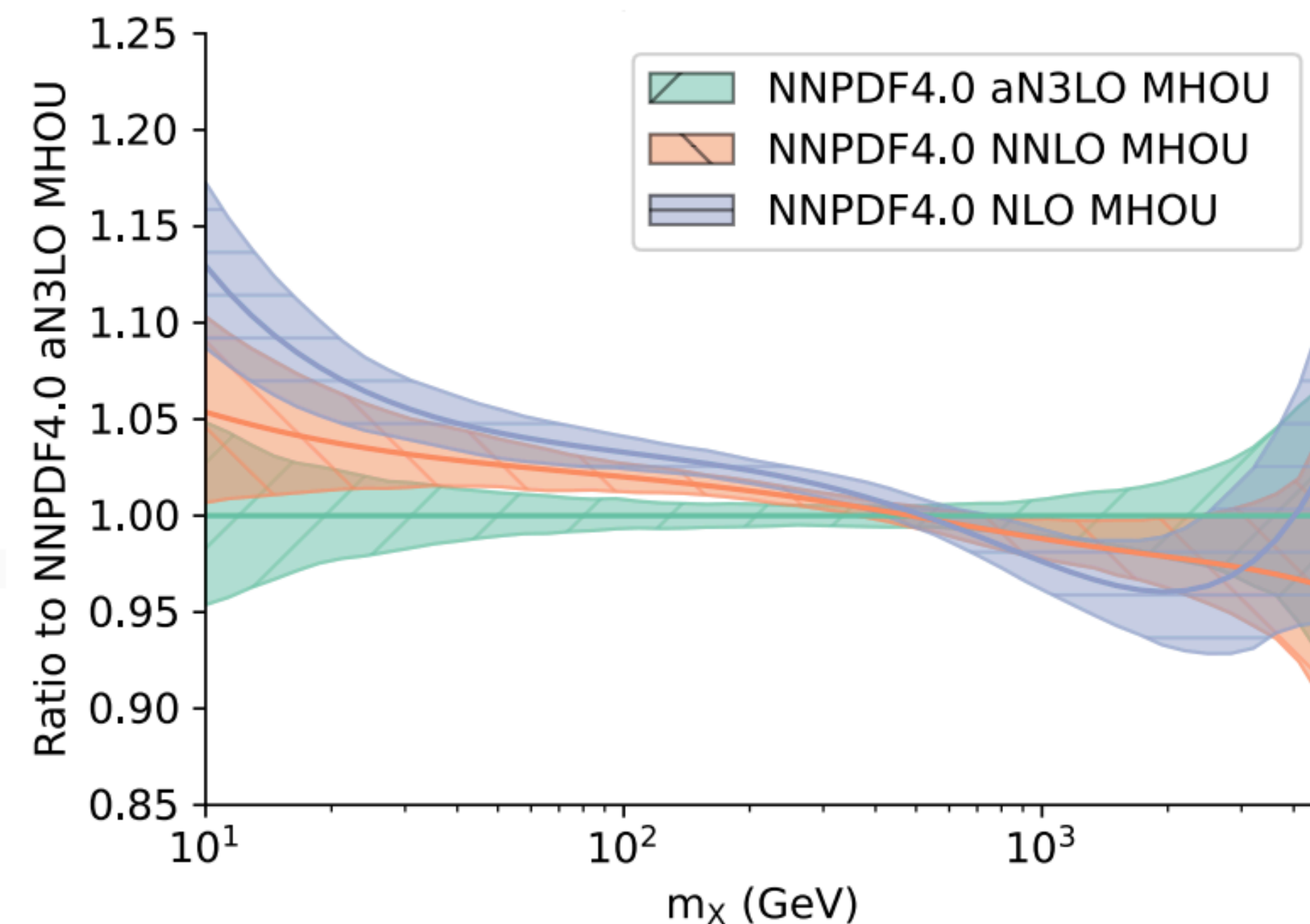
✓ NNLO (2004)

N³LO
partial results

N⁴LO
partial results

Huge progress towards full N3LO

- Splitting functions known in several limits and enough momenta to build reasonable approximation ([Blumlein, Moch, Gehrmann, von Manteuffel, Sotnikov, Yang, Davies, Vogt, Bonvini, Marzani,...](#) see benchmark and references in [arXiv:2406.16188](#))
- However partonic cross sections in PDF fits only at NNLO
- Inclusion of MHOU in PDF errors improves convergence and accounts for IHOU



PROGRESS: THEORY

Error estimate before aN3LO PDFs

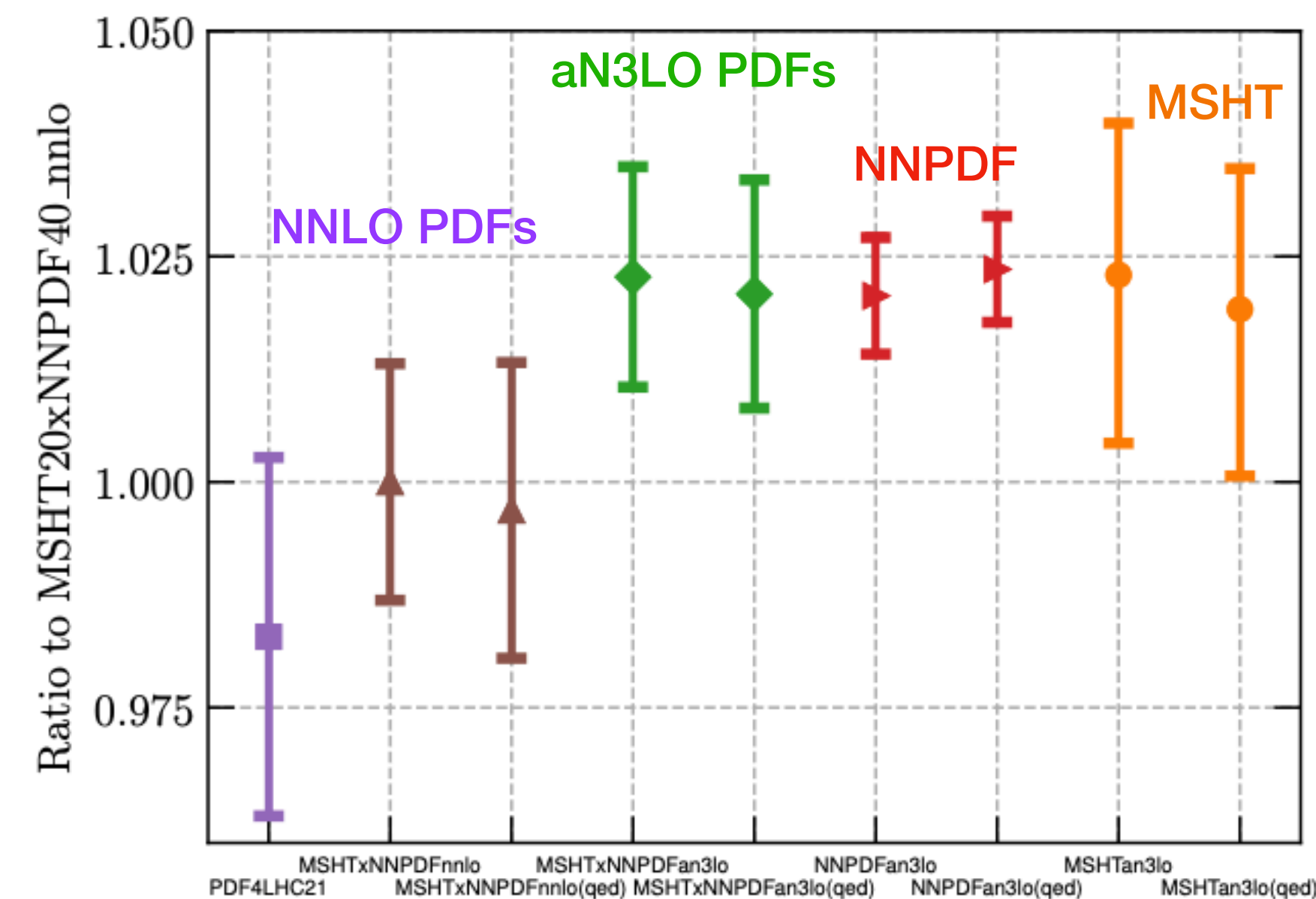
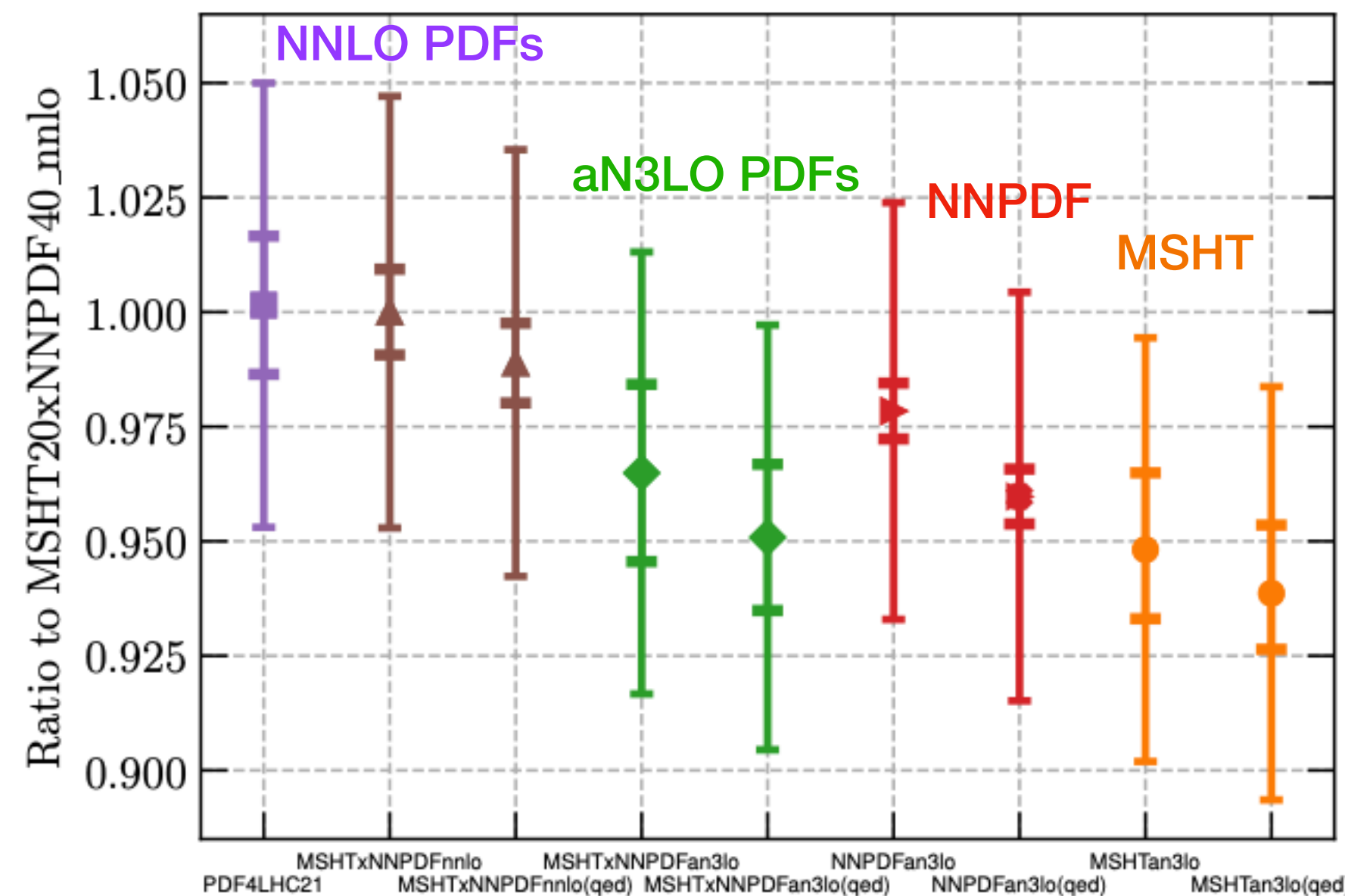
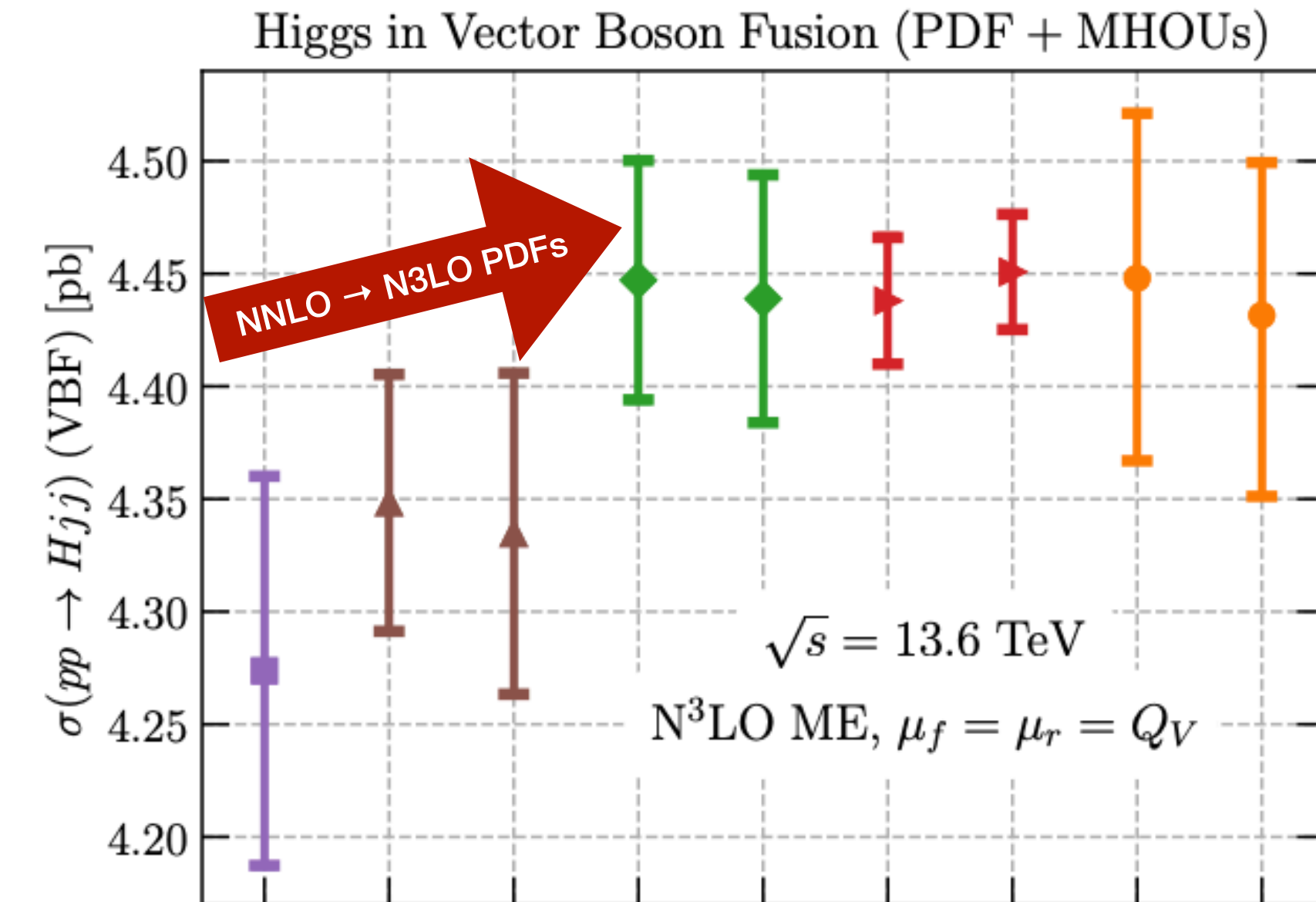
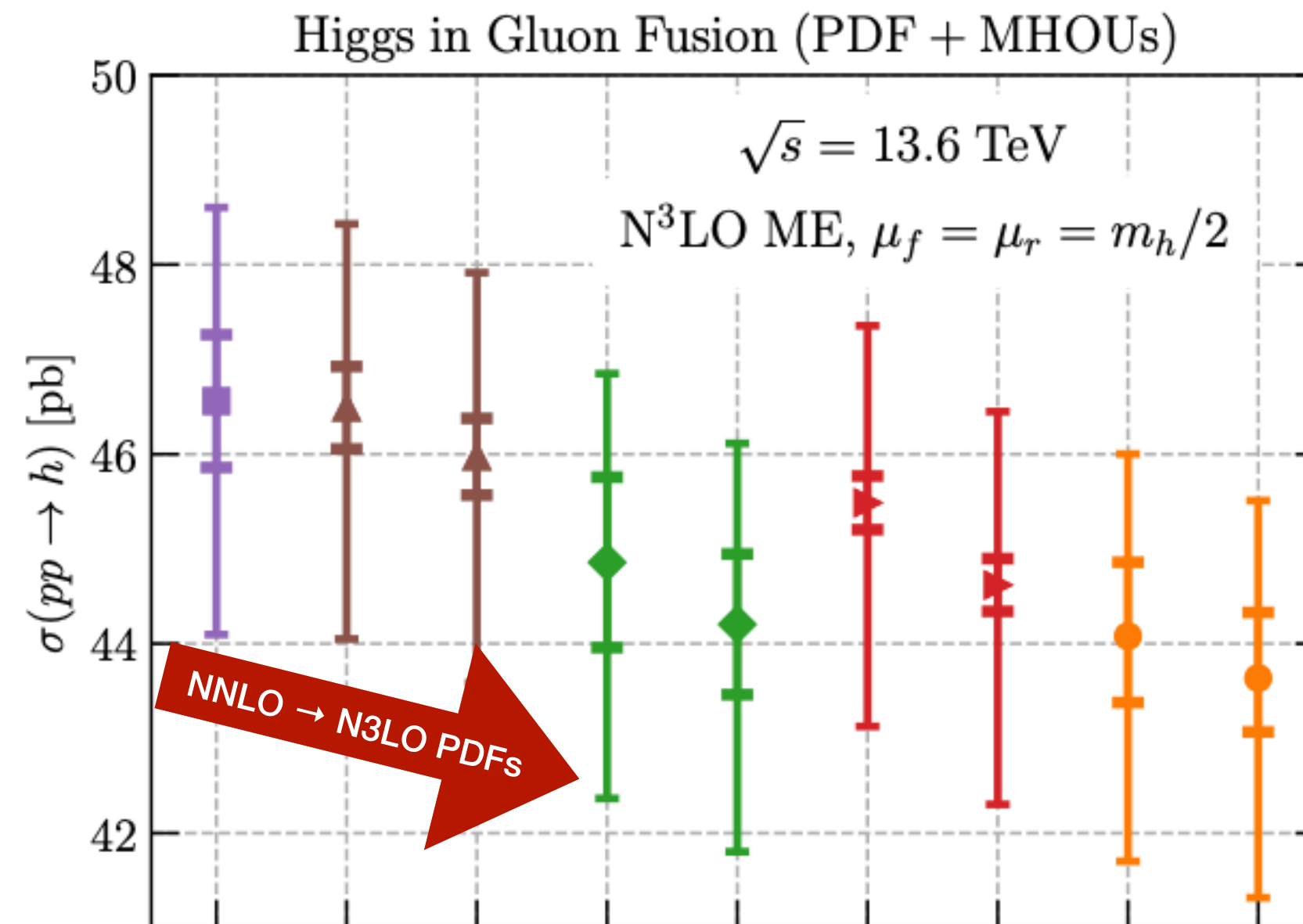
$$\Delta_{\text{NNLO}}^{\text{app}} \equiv \frac{1}{2} \left| \frac{\sigma_{\text{NNLO-PDF}}^{\text{NNLO}} - \sigma_{\text{NLO-PDF}}^{\text{NNLO}}}{\sigma_{\text{NNLO-PDF}}^{\text{NNLO}}} \right|$$

Actual error using aN3LO PDFs

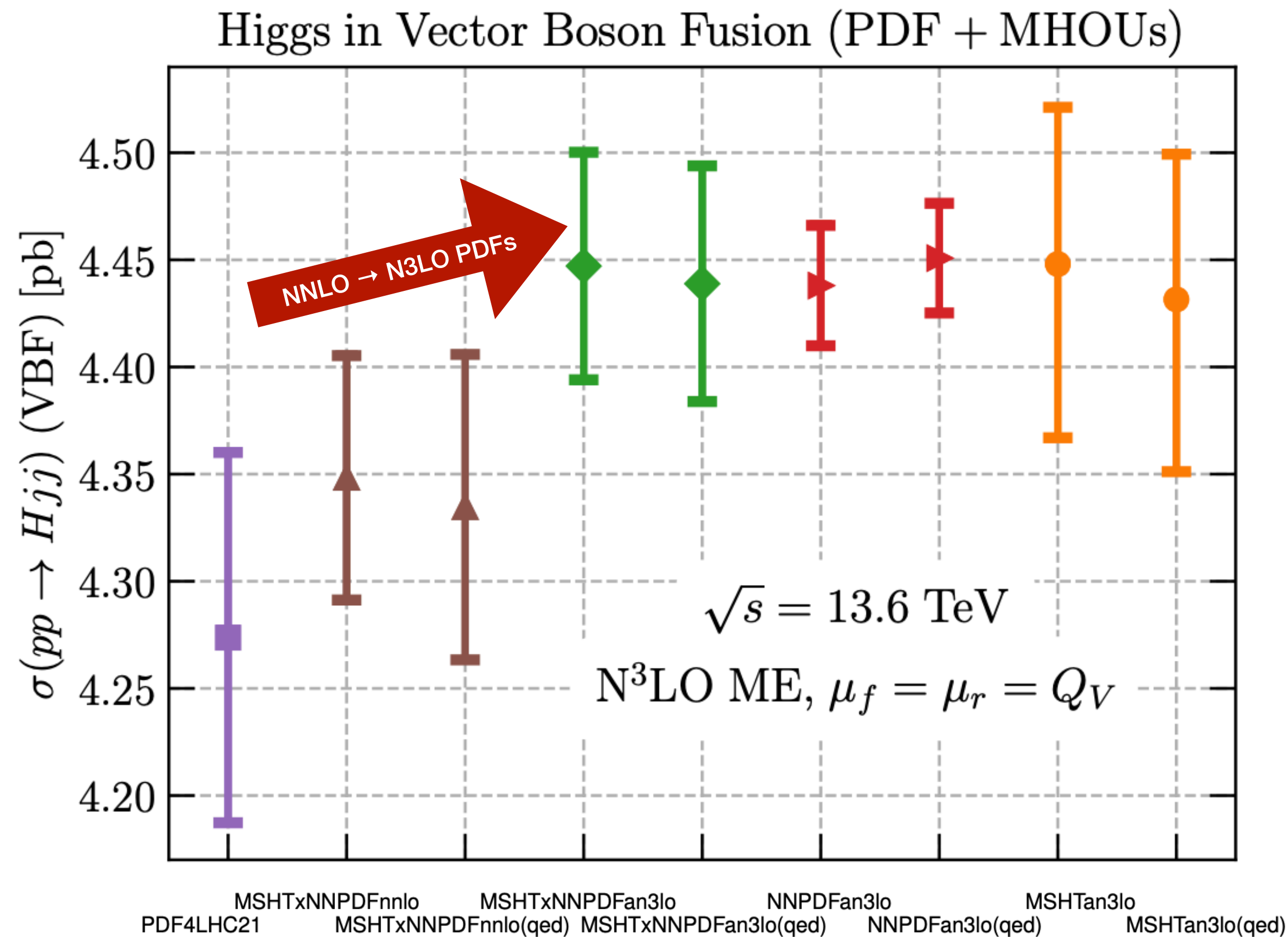
$$\Delta_{\text{NNLO}}^{\text{exact}} \equiv \left| \frac{\sigma_{\text{N}^3\text{LO-PDF}}^{\text{N}^3\text{LO}} - \sigma_{\text{NNLO-PDF}}^{\text{N}^3\text{LO}}}{\sigma_{\text{N}^3\text{LO-PDF}}^{\text{N}^3\text{LO}}} \right|$$

	ggF	VBF
$\Delta_{\text{N}^3\text{LO}}$ (approx)	1.6%	0.5%
$\Delta_{\text{N}^3\text{LO}}$ (actual)	3.3%	2.3%

Previous estimates of the effect of NNLO/N3LO mismatch were **optimistic**



FRONTIER: HIGHER ORDERS



e.g. Higgs in VBF @ N3LO
[MSHT+NNPDF arXiv:2411.05373](#)

- Shift between results obtained with NNLO PDFs and with aN3LO PDFs more significant than expected
- Next steps:
 - As more N3LO theory ingredients become available, **aN3LO PDFs updated** by improving convergence of results obtained with different methodologies
 - **Exact N3LO evolution**, exact NNLO grids for partonic cross section in PDF fits and N3LO k-factors
- Every step of the way becomes computationally more costly and complex for a smaller and smaller improvements. But **extremely needed!**
- What about other effects? Resummations, higher twists, power corrections...[\[Ball, Chiefa 2511.14387\]](#)
- Estimate of incomplete N3LO terms and MHO uncertainties still computed via scale variations: is there a better way that can be exported in PDF fits?

3) Methodology: status and frontiers

PROGRESS: TESTING THE METHODOLOGY

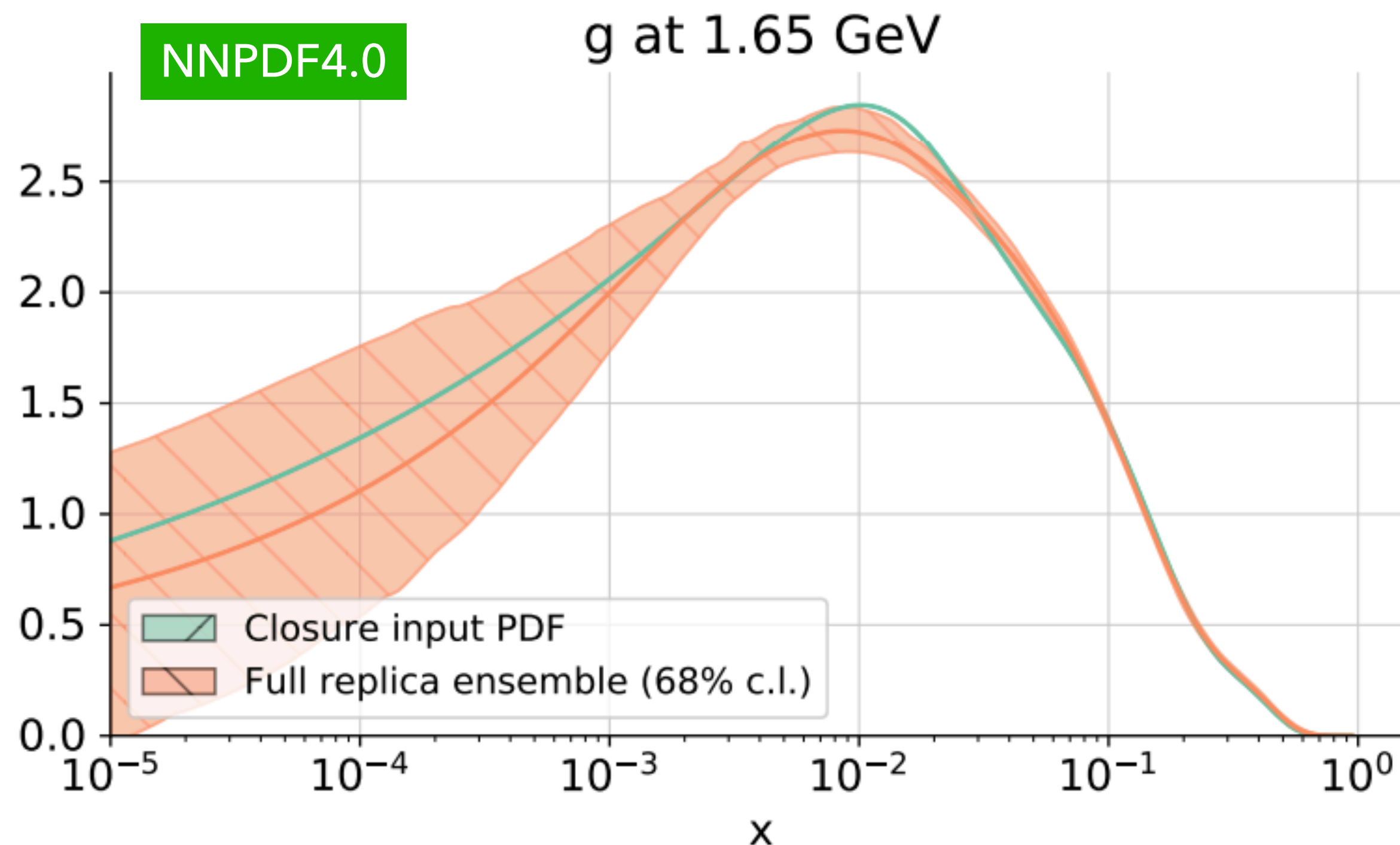
- Closure tests: key study to **statistically validate** various PDF fitting methodologies and **assess faithfulness of PDF uncertainties**
- Imagine we knew the law of Nature $G(w^0)$ and the true PDFs w^0 is fitting methodology able to reproduce it? Is the uncertainty faithful?
- Can also be used to check the effect of inconsistencies in the data
[Harland-Lang et al, arXiv:2407.07944, Costantini et al: arXiv:2503.17447]

$$y_0 = \mathcal{G}(w^0) + \eta_i$$

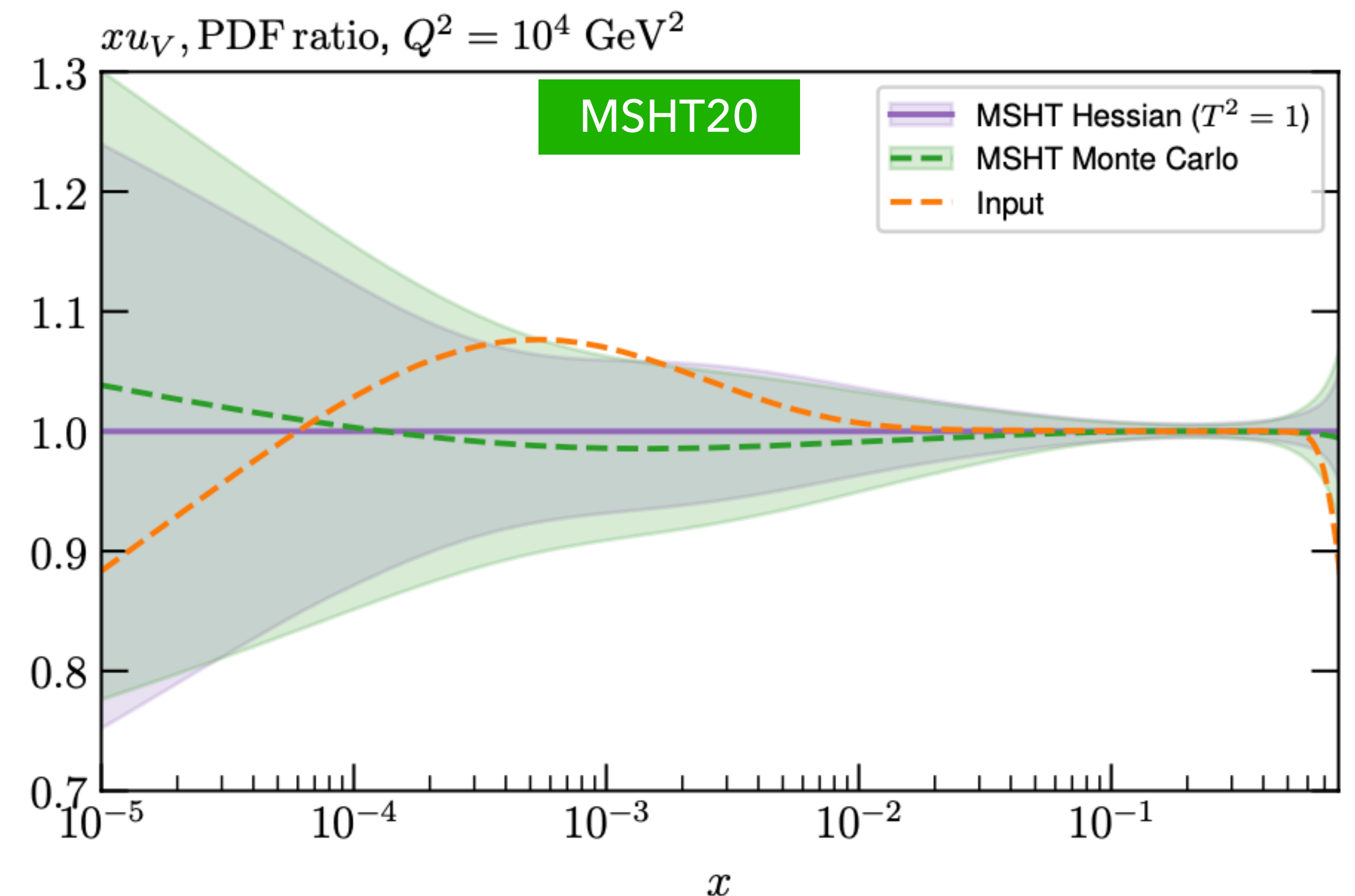
Law of Nature

$\hat{\sigma}_{\text{NNLO}} \otimes (f \otimes f)$ "true" PDF

Experimental noise



Del Debbio et al, [arXiv: 2111.05787]



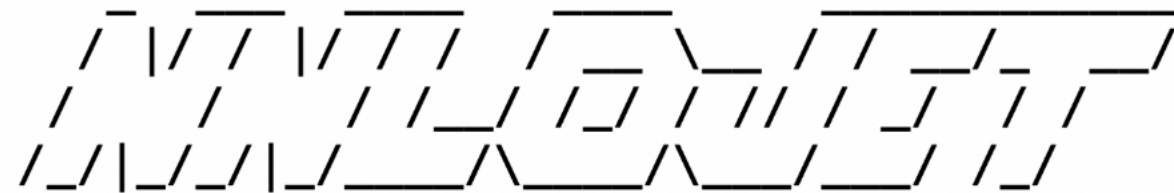
Harland-Lang et al, arXiv:2407.07944

PROGRESS: TESTING PDF PERFORMANCE ON UNSEEN DATA

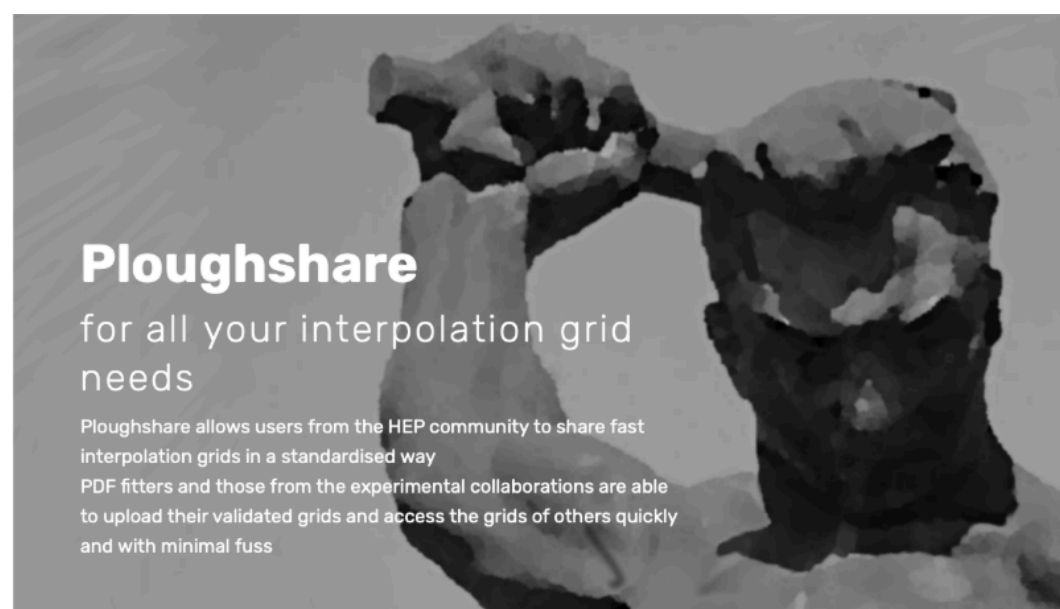
- To discriminate among PDF sets, test all global PDF sets against **new precise** data from LHC data & DIS HERA jets data. How well do various PDF sets describe data that are **not yet** included in the fit?
- Generalisation test** is crucial and possible using exact NNLO predictions (no k-factor approximation) thanks to PineAPPL, NNLOJET, MATRIX and Ploughshare

PineAPPL

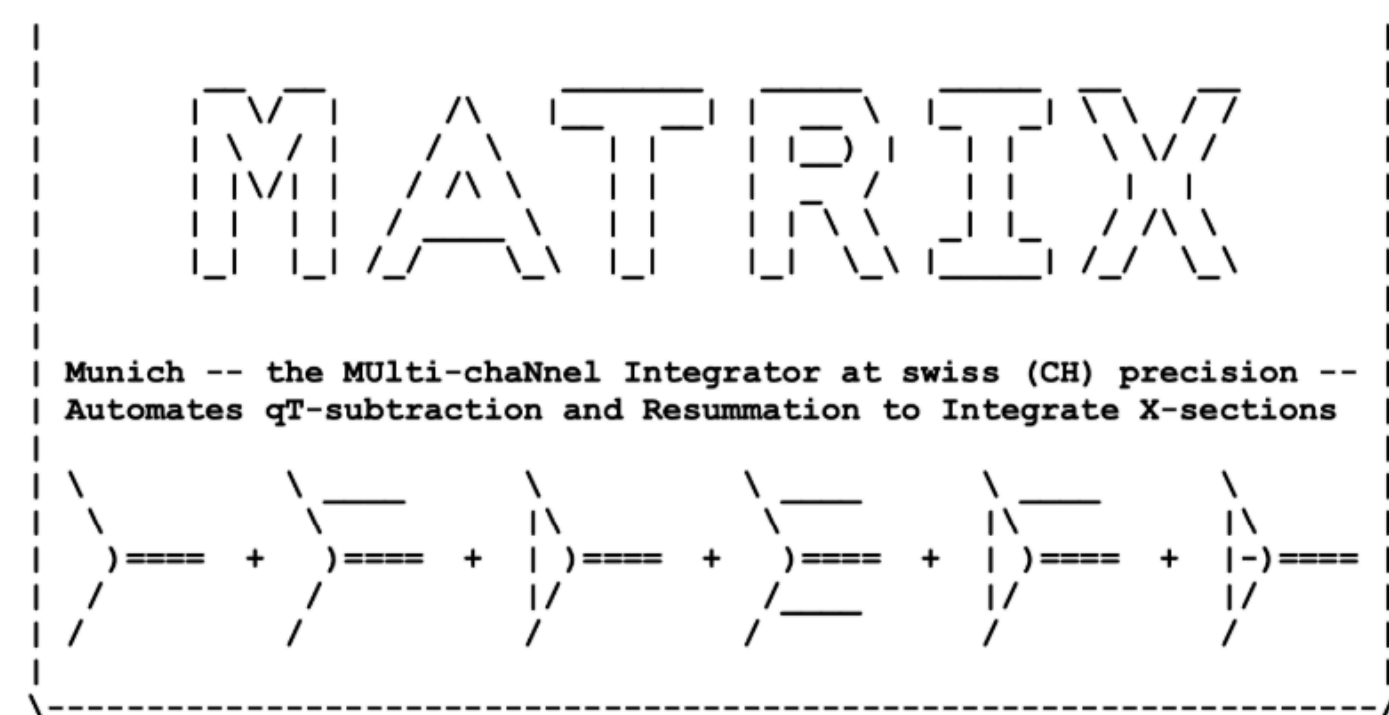
Carrazza et al: 2008.12789
Barontini et al: 2302.12124



Gehrman-De Ridder et al: 1507.02850,
1605.04295, 2301.11827, Britzger et al:
2207.13735, Cruz-Martinez et al: 2501.13167



Ploughshare project, APPLgrid, FastNLO
Kluge et al: hep-ph/0609285
Carli et al: 0911.2985
Czakon et al: 1704.08551



Grazzini et al: 1711.06631

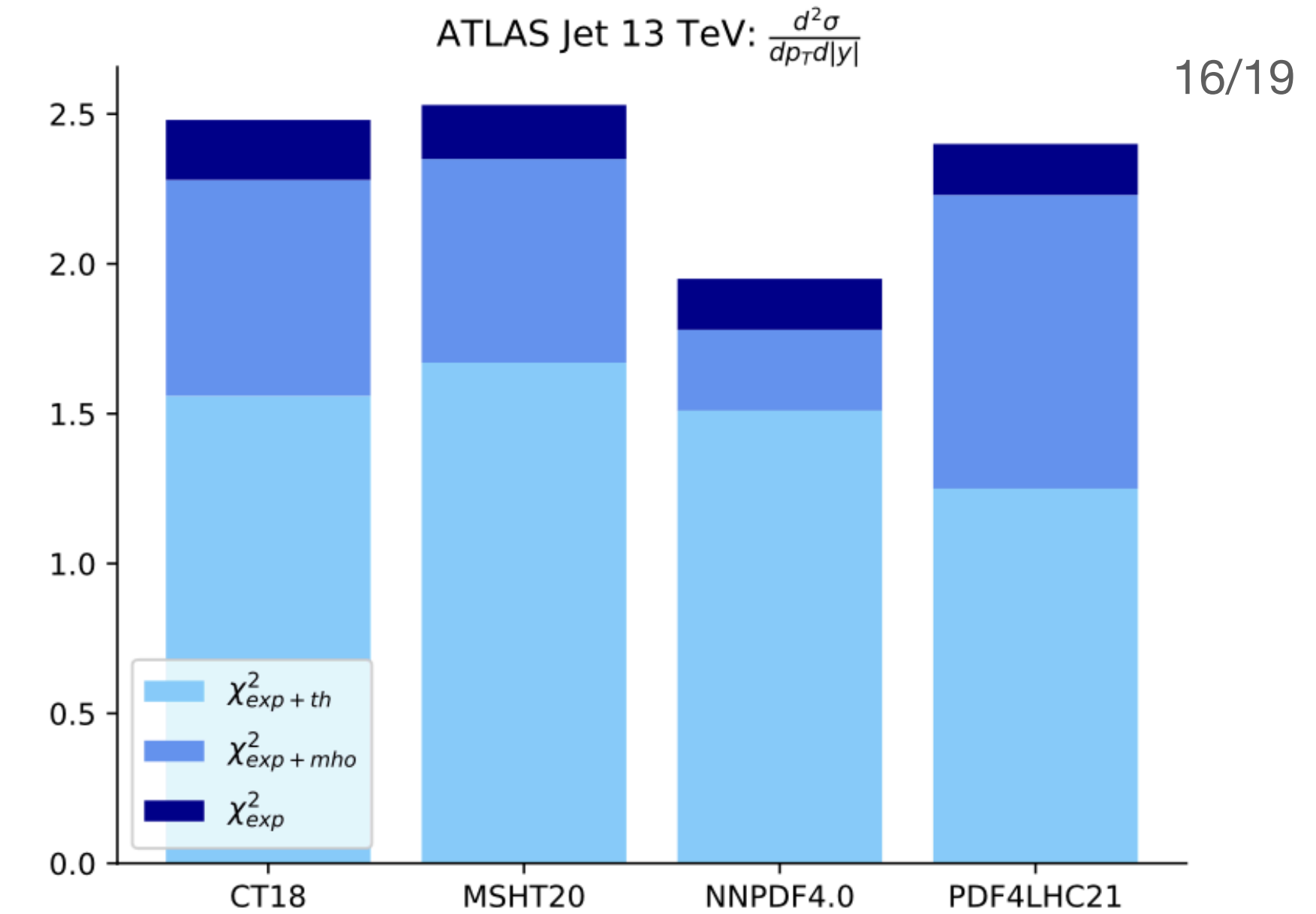
- Data-theory agreement with unseen data must be measured by considering **all sources of uncertainties**: experimental, missing higher order (MHOU), PDF uncertainties, α_s uncertainties by including all bin-by-bin correlations

$$\chi^2 = \sum_{i,j=1}^{N_{\text{dat}}} \left(T_i^{(0)} - D_i \right) (\text{cov}^{-1})_{ij} \left(T_j^{(0)} - D_j \right)$$

$$\chi_{\text{exp+mho+pdf+as}}^2 \leftrightarrow \text{COV} = \text{COV}_{\text{exp}} + \text{COV}_{\text{mho}} + \text{COV}_{\text{pdf}} + \text{COV}_{\text{as}}$$

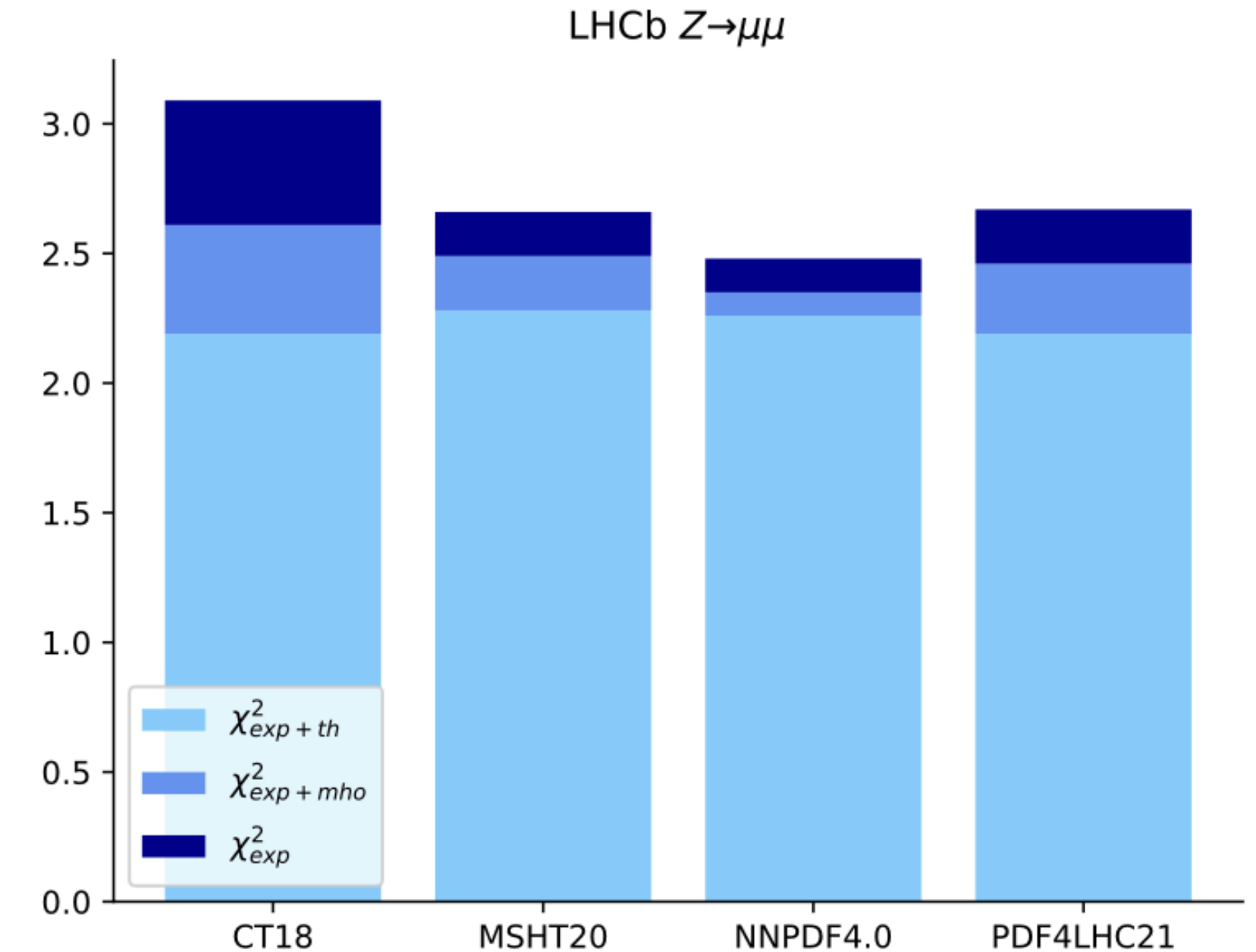
HOW WELL DO MODERN PDF SETS PERFORM?

- Recently analysed 25 experimental distributions LHC@13 TeV and HERA jets measurements not yet included in any PDF analysis: inclusive jets, dijets, top differential distributions, W and Z production, Z pT, DIS jets
[Chiefa et al, arXiv:2501.10359]
- Crucial to keep into account all sources of uncertainties and all bin-by-bin corrections.
- For some observables the inclusion of MHO uncertainties or PDF uncertainties changes the picture as compared to pure experimental χ^2
- NNPDF4.0, despite being the most precise (smallest PDF uncertainties in data region) yields good agreement with data
- LHC@13 TeV measurements do not strongly discriminate among PDF sets so far, keep testing!.



ATLAS incl. jet sec at 13 TeV, R=0.6, anti-kT

ATLAS collaboration [arXiv:1711.02692]



LHCb Z forward production at 13 TeV

LHCb collaboration [arXiv:2112.07458]

- Given some model constraints and given $O(5000)$ experimental data points $\mathbf{D}$ want to determine the set of functions $\{f\}$ (with $i = 1, \dots, n_{\text{partons}}$) and **estimate their uncertainty**
- Want to find a infinite-dimensional object from a finite number of information.
- Hence, PDF inference is an example of a **non-linear infinite-dimensional** inverse problem.
- Mapping $\mathbf{D}$ into f is mathematically ill defined, nobody knows the true $f = f^*$
- Best we can do it to find best f given the data $\mathbf{D}$.

$$P(f|\mathbf{D}) \propto P(\mathbf{D}|f) P(f)$$

✓ Likelihood

? Prior

$$P(\mathbf{D}|f) \sim \exp(-\chi^2/2)$$

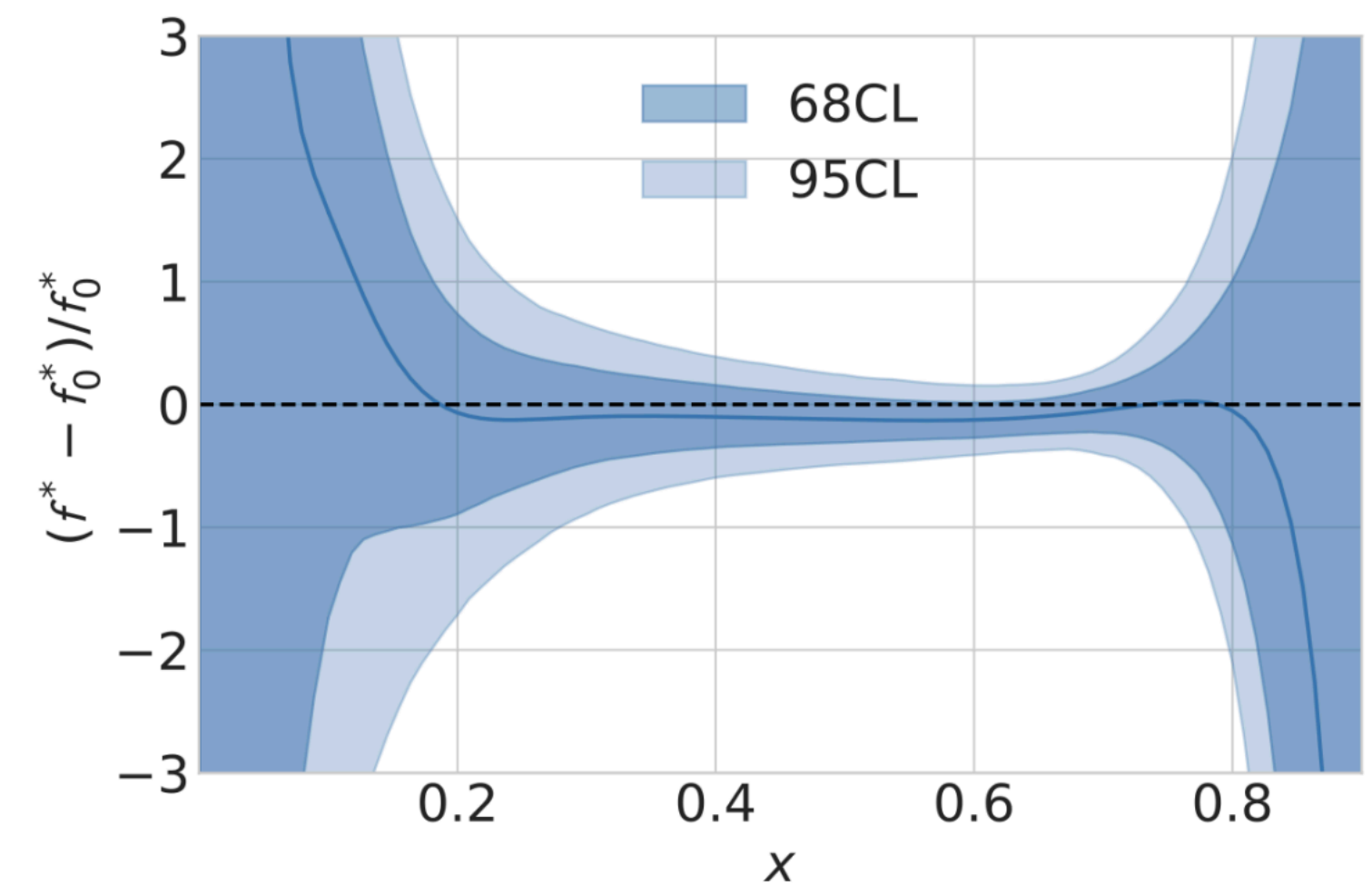
$$\chi^2 = \sum_{a,b=1}^{N_{\text{dat}}} (T_a(\{f\}, \vec{c}) - D_a) \Sigma_{ab}^{-1} (T_b(\{f\}, \vec{c}) - D_b)$$

Two ways of assuming a prior:

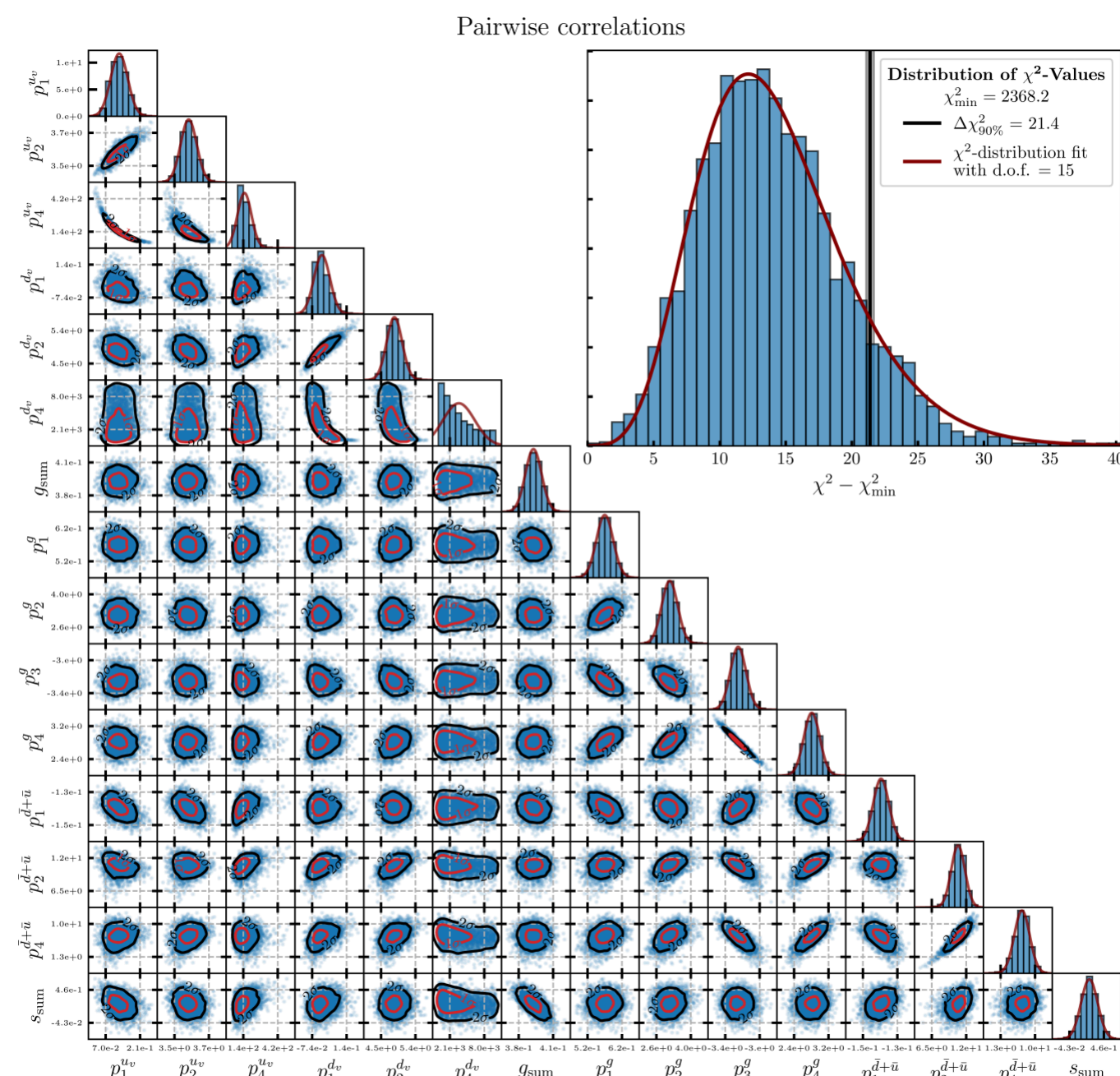
- **Explicit parametrization** (parametrical modelling)
Function is projected on a vector space of parameters θ
- **Non-parametrical inference** (NNs):
use data to also infer probable 'smoothness' of $f(x)$ and thus infer probability function $P(f)$ throughout the space of functions
- **No explicitly Bayesian PDF fits so far!**

FRONTIER: TOWARDS BAYESIAN PDF FITS

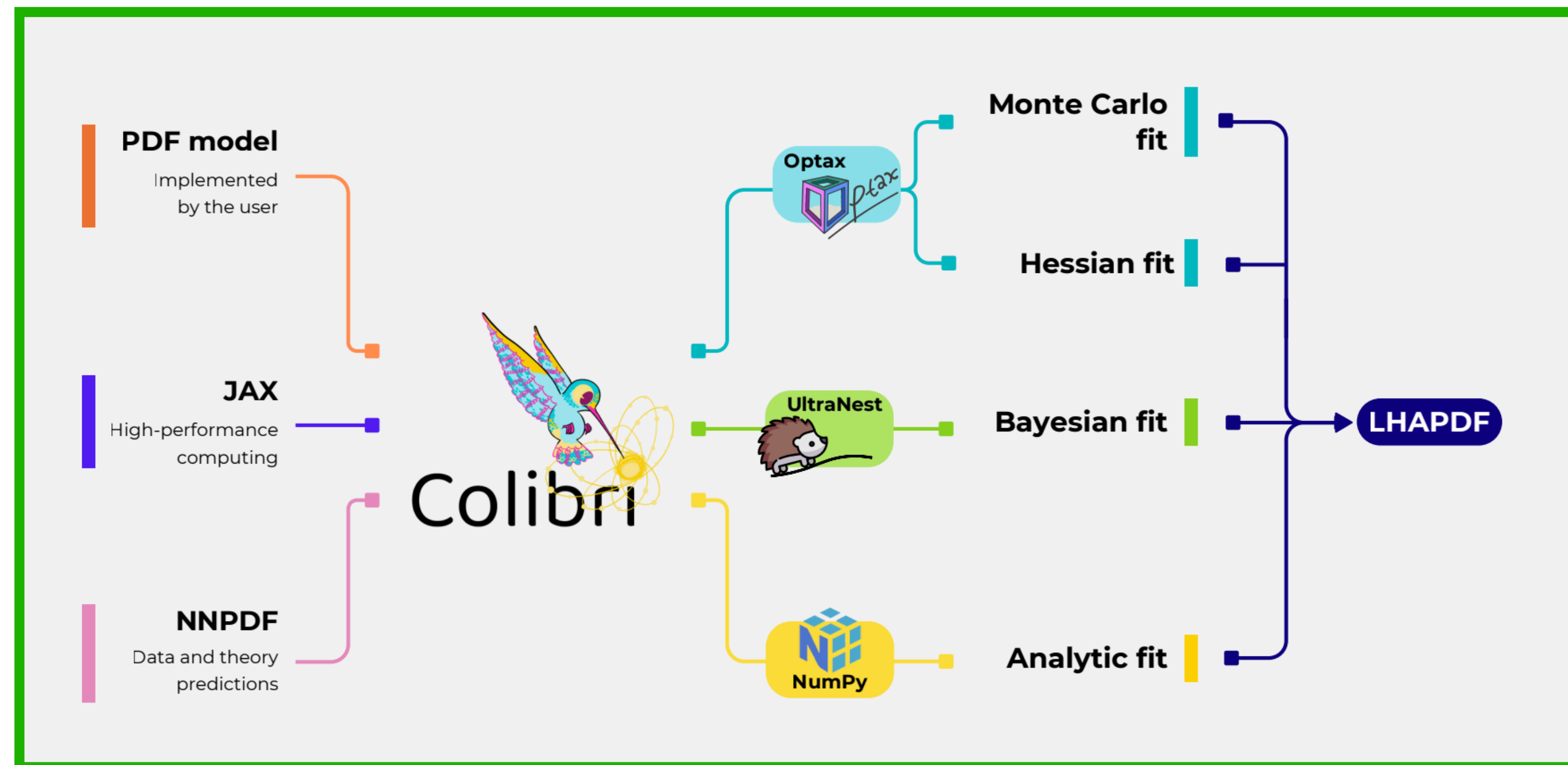
- First proof-of-concepts and public tools for explicitly Bayesian PDF fits
- Development would allow to
 - ➔ Determine model(s) with largest evidence (balance between χ^2 and complexity)
 - ➔ Have explicit control on prior choices.
 - ➔ Bayesian average models based on evidence (instead of PDF4LHC combination)
- Prior can be extended to SM / BSM parameters along with PDF parameters (simultaneous fits)



Candido et al: 2404.07573



Risse et al: 2510.16158



M. Costantini, L. Mantani, J. Moore, V. Schütze Sanchez, MU, arXiv:2510.03391

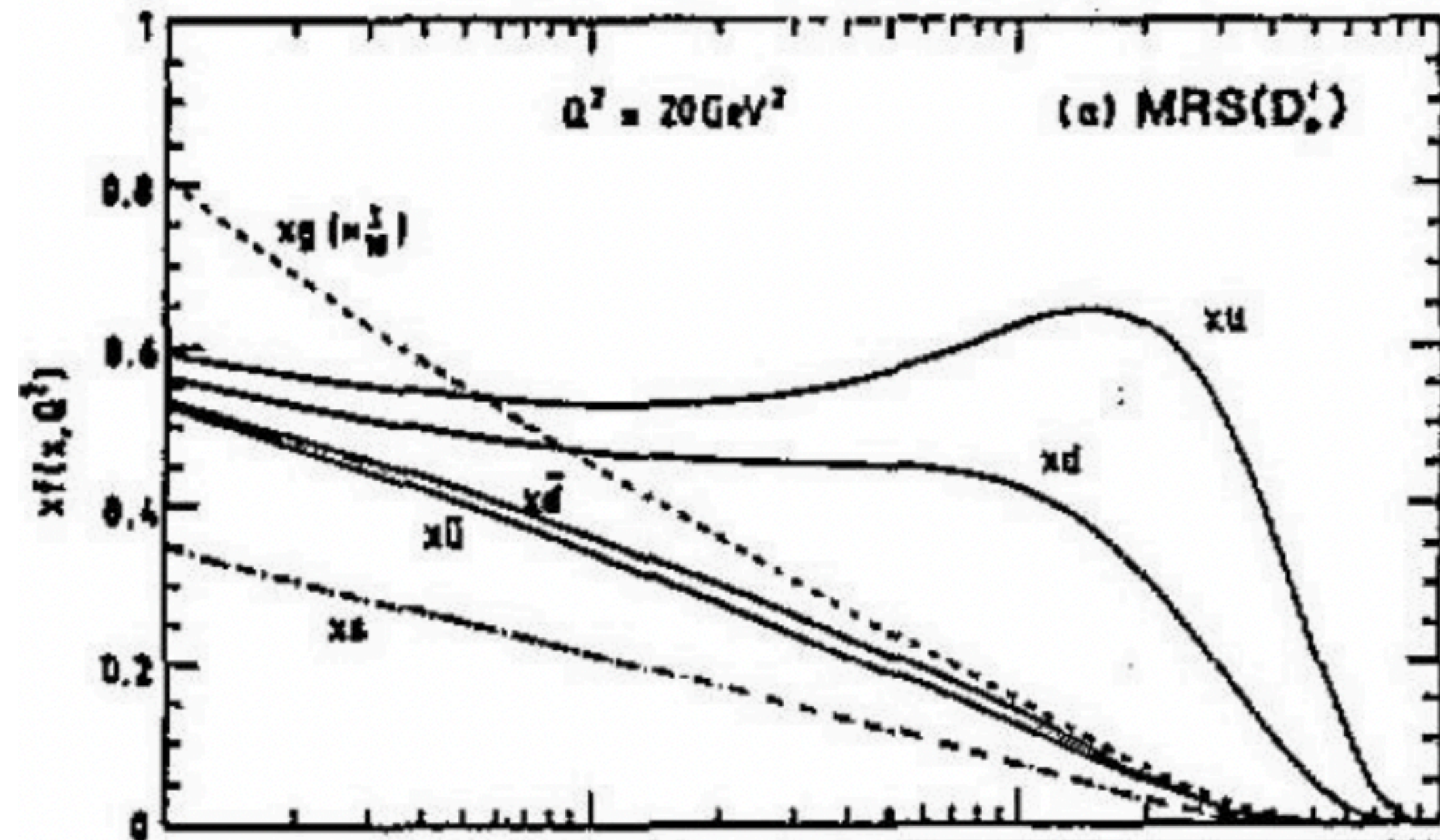
CONCLUSIONS AND OUTLOOK

- In an era of precision at LHC, need precise and accurate PDFs
- Lots of progress in theoretical framework:
aN3LO PDFs, inclusion of MHOUs in PDF error bands, QED corrections
- New exciting avenues such as interplay between heavy BSM and PDFs, hinting to complementarity of low-energy experiments (EIC, FFP) and lattice to constrain large- x PDFs
- Simultaneous determination of PDFs alongside SM/BSM parameters might become crucial for interpretation of HL-LHC
- Tools like closure tests and generalisation tests are crucial to test methodology robustness, effects of theory/exp inconsistencies
- Crucial reproducibility & open-source tools (and fitting codes)
- Bayesian PDF (+SM/BSM) fits new exciting frontier for statistical model average and good control on prior assumptions.



EXTRA MATERIAL

PARAMETRICAL APPROACH



Martins, Roberts, Stirling (1993)

20 free parameters

- Explicit parametrisation:

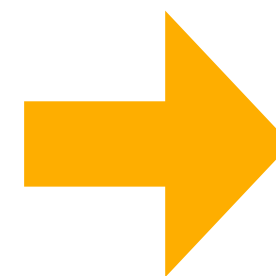
$$f(x, \mu_0) = a_0 x^{a_1} (1 - x)^{a_2} P(x, \{a_3, \dots, a_{N_{\text{par}}}\})$$

N_{par} : not too many, not too few

- Assume that $P[f]$ has uniform support in this particular space of functions
- Maximise $P(f|D)$ by maximising $P(D|f)$ - **maximum likelihood**
- Gives best fit PDFs by minimising χ^2 in $\{a_i\}$ space and uncertainties by setting C.L.s on χ^2 contour

$$\Delta\chi^2 = 1 \Leftrightarrow 68\% \text{ C.L.}$$

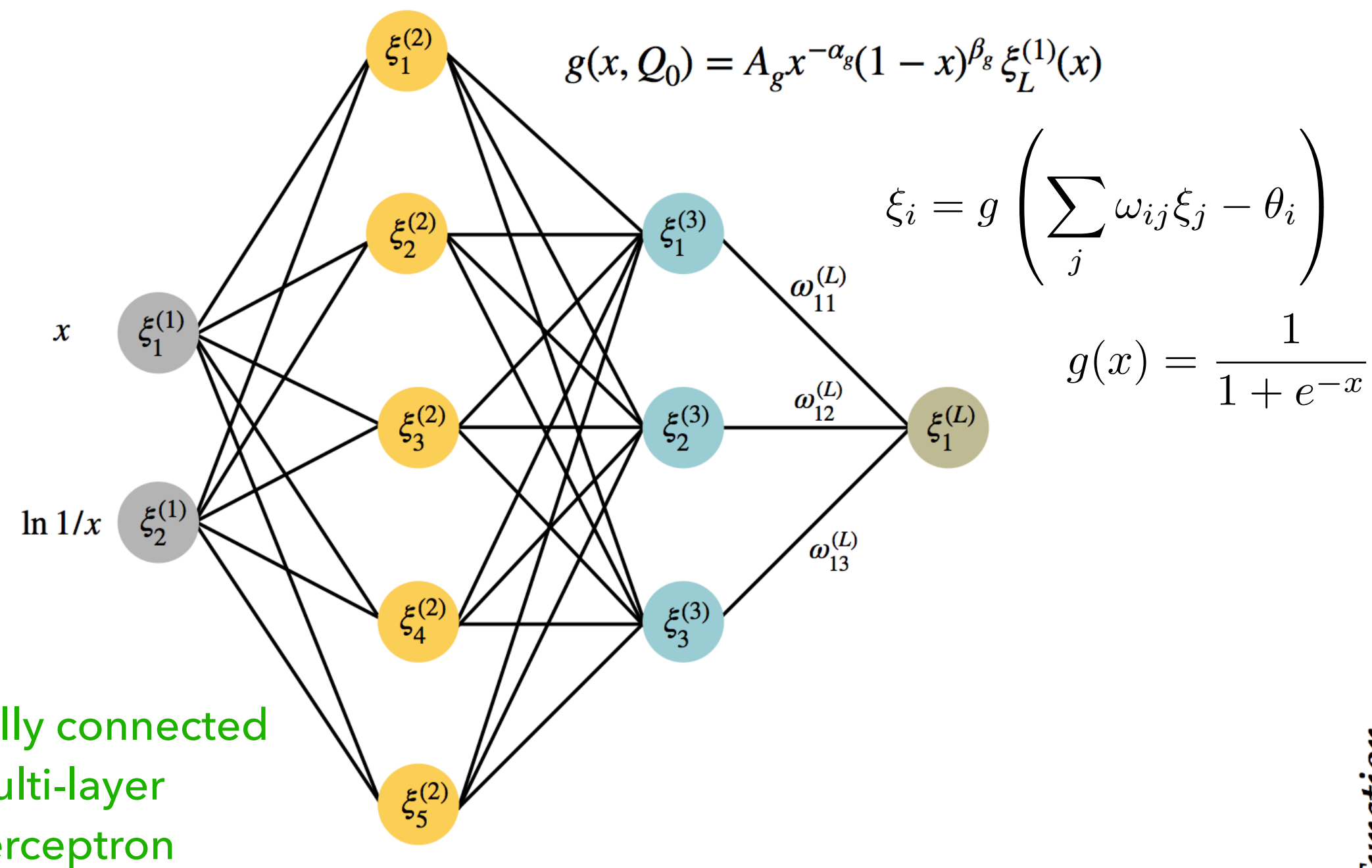
Problem: precision $\ll$ accuracy
Possible reasons: data inconsistencies or over-constraining modelling



Solution: inflate experimental uncertainties by a factor T such that $\Delta\chi^2 = T^2$ until **($T \sim 5-10$)**
precision $\sim$ accuracy

NEURAL NETWORK APPROACH

- Alternatively, choose parametrisation so large that in principle can fit any conceivable function f .
- One option is to use NNs [Forte, Latorre 2002]
- First PDF fit using NN to parametrise PDFs [Ball, Del Debbio, Forte, Guffanti, Piccione, Rojo, MU, 2008]

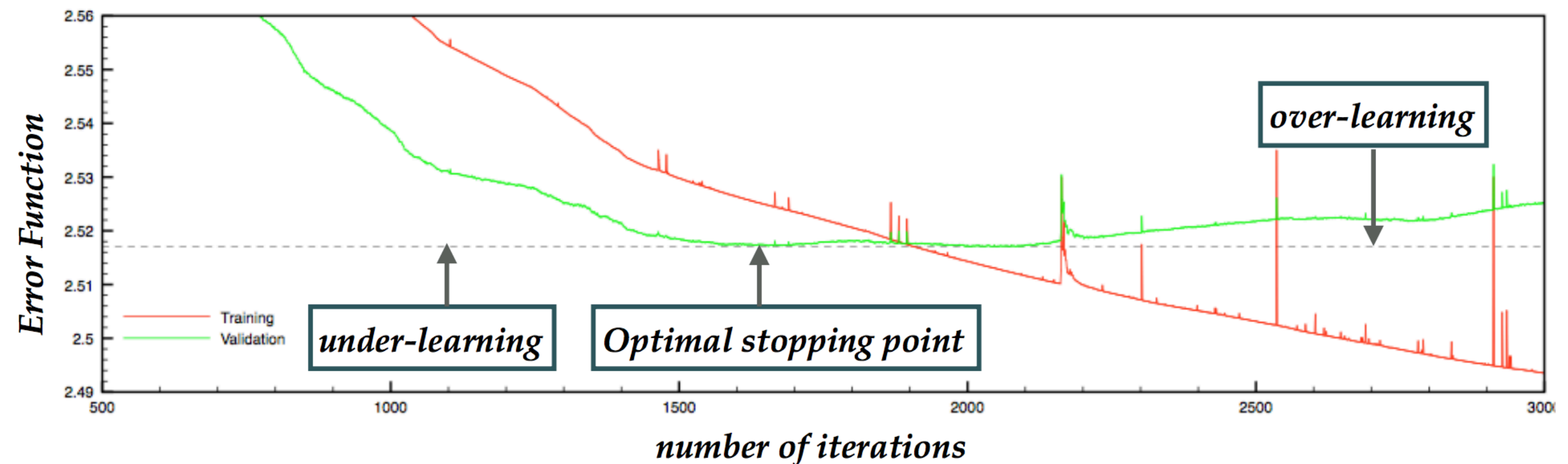


$$g(x, Q_0) = A_g x^{-\alpha_g} (1-x)^{\beta_g} \xi_L^{(1)}(x)$$

$$\xi_i = g \left(\sum_j \omega_{ij} \xi_j - \theta_i \right)$$

$$g(x) = \frac{1}{1 + e^{-x}}$$

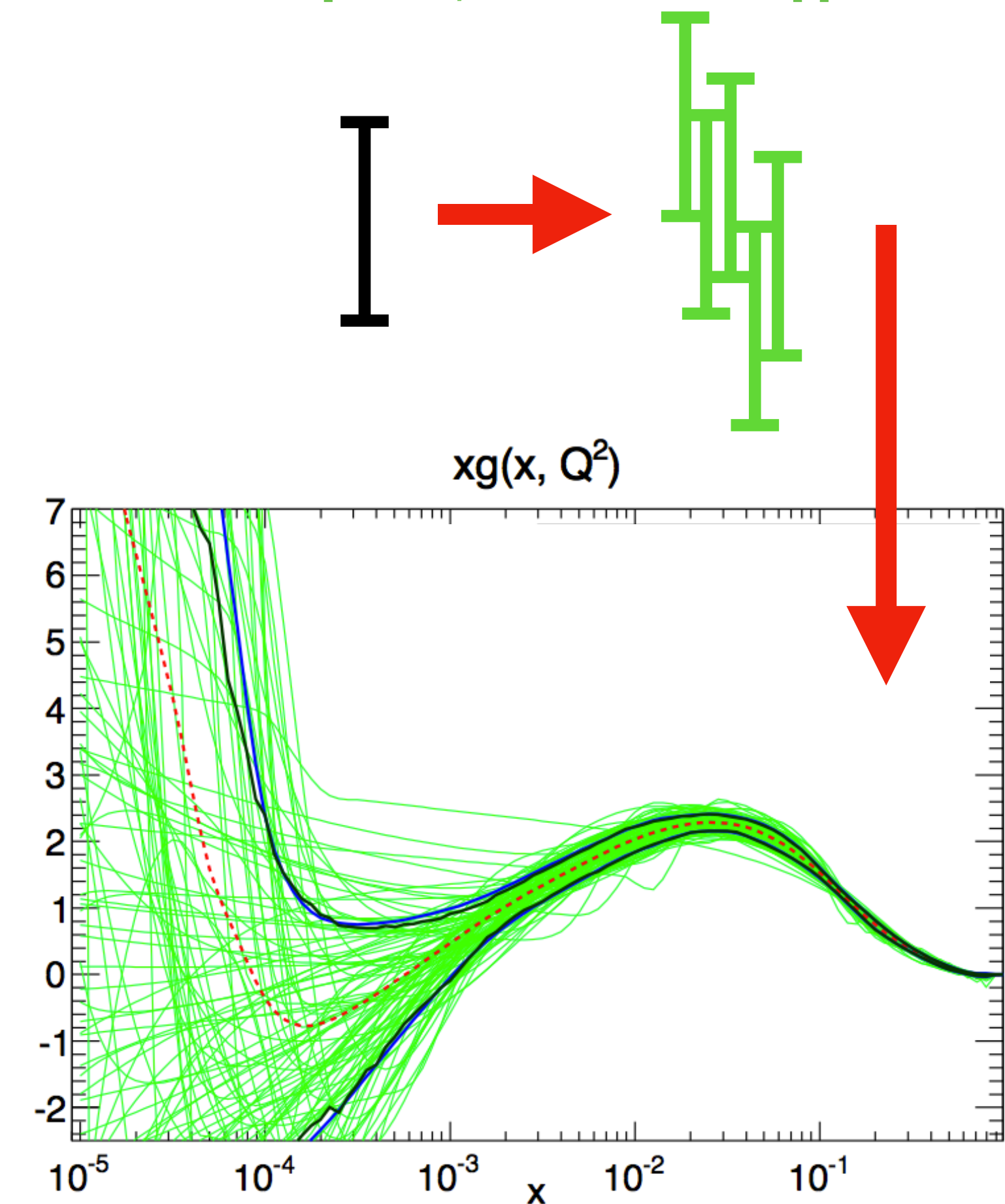
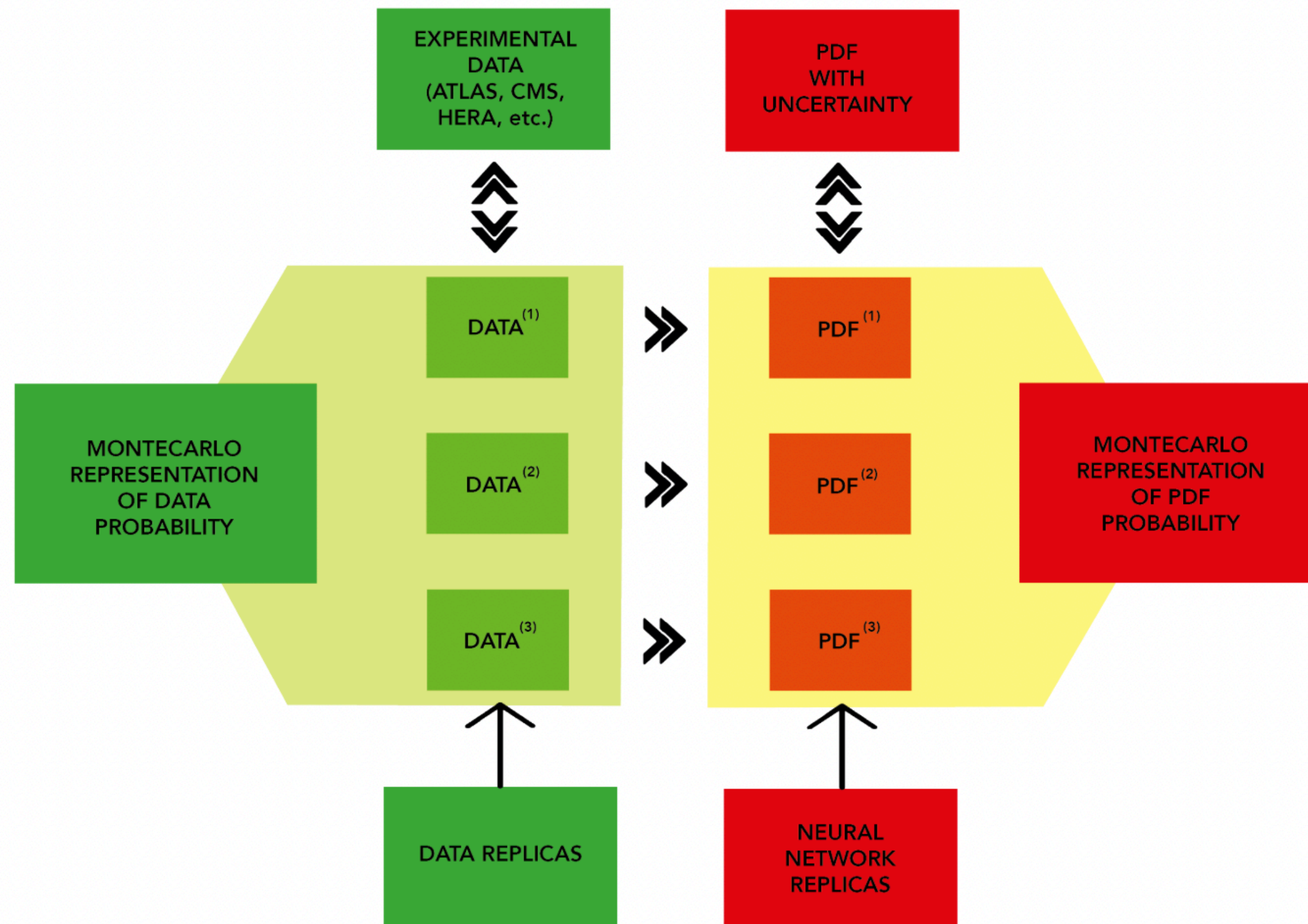
- Neural Networks: each independent PDFs associated to an unbiased and flexible parametrisation: $O(300)$ parameters of seven (2-5-3-1) fully connected multi-layer perceptron versus $O(30)$ in polynomial parametrisation
- Training and validation splitting to avoid over learning



$$\xi_1^{(3)}(\xi_1^{(1)}) = \frac{1}{1 + e^{\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1 + e^{\theta_1^{(2)} - \xi_1^{(1)} \omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1 + e^{\theta_2^{(2)} - \xi_1^{(1)} \omega_{21}^{(1)}}}}}}$$

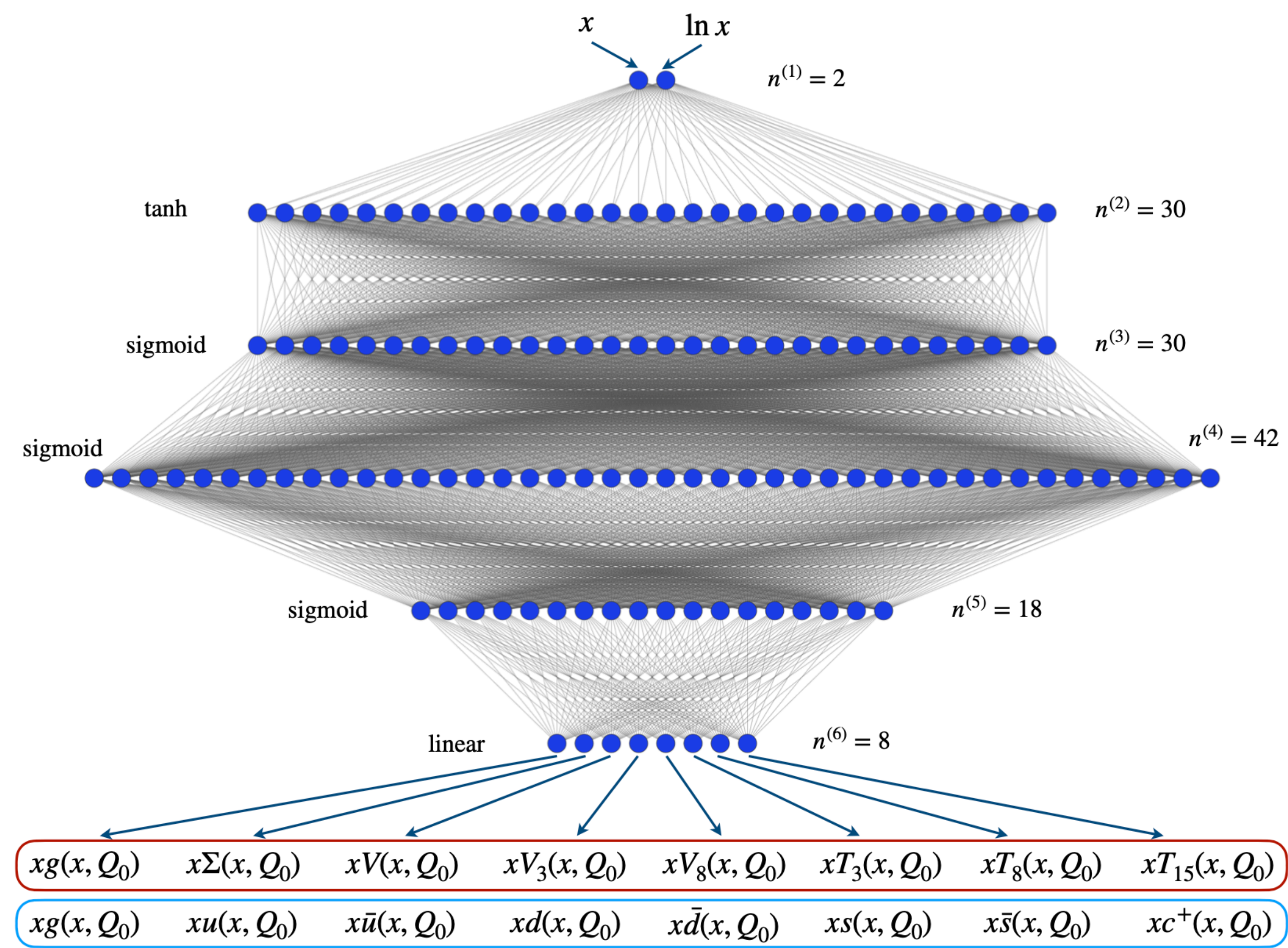
NEURAL NETWORK APPROACH

- Alternatively, choose parametrisation so large that in principle can fit any conceivable function f .
- One option is to use NNs [Forte, Latorre 2002]
- First PDF fit using NN to parametrise PDFs [Ball, Del Debbio, Forte, Guffanti, Piccione, Rojo, MU, 2008]
- Hessian approach not applicable. Use Monte Carlo - or bootstrap - error propagation by importance sampling [Giele, Kosover 1993] [Forte et al, 2006]

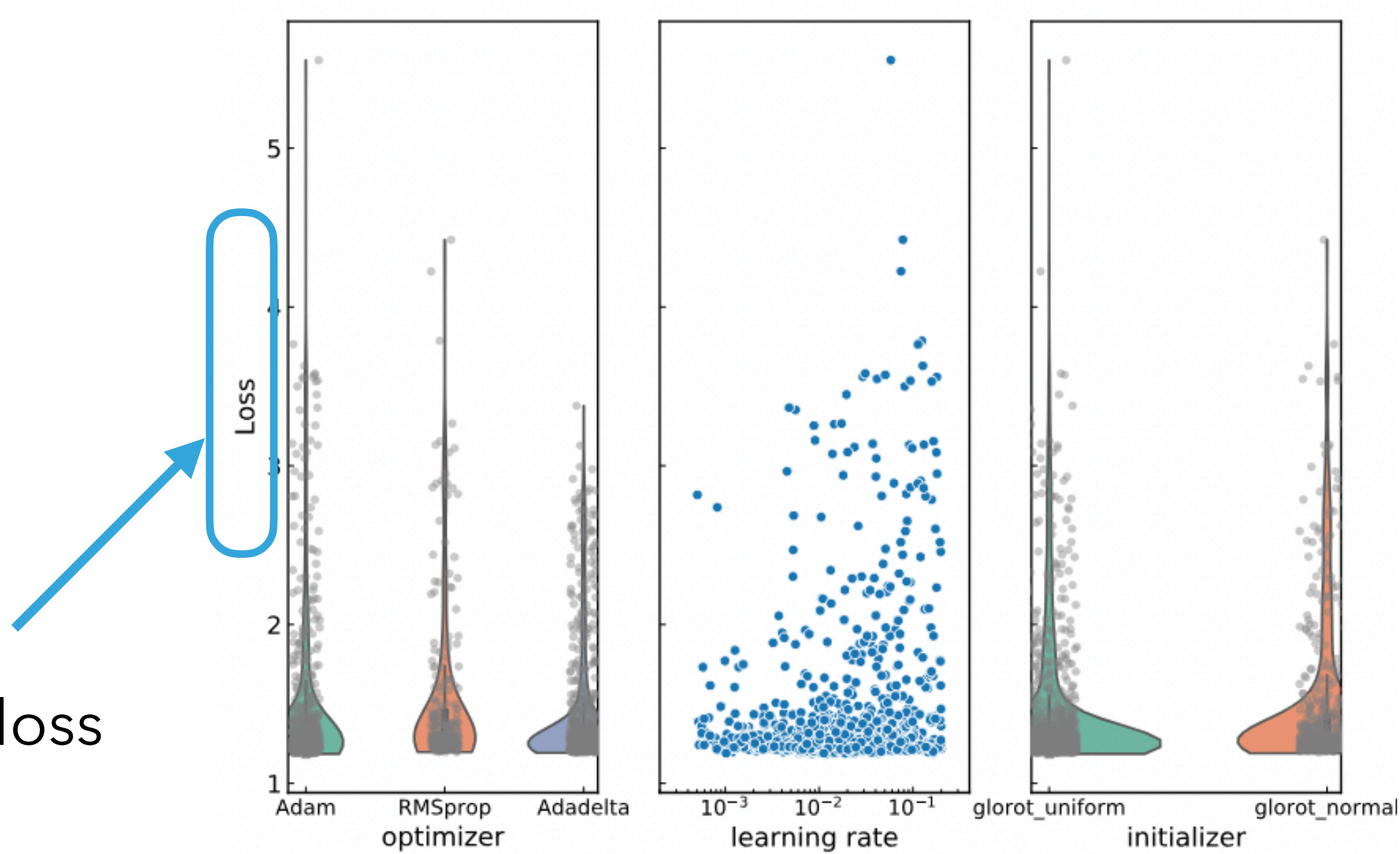


NNPDF4.0 METHODOLOGY

- NNPDF4.0: Single deep Neural Network (763 parameters) [Ball et al, 2021 arXiv:2109.02653]
- Hyper-parameter optimisation via K-fold procedure: scan parameter space and optimise K-fold loss.



K-fold loss



List of hyper-parameter

NEURAL NETWORK	FIT OPTIONS
NUMBER OF LAYERS (*)	OPTIMIZER (*)
SIZE OF EACH LAYER	INITIAL LEARNING RATE (*)
DROPOUT	MAXIMUM NUMBER OF EPOCHS (*)
ACTIVATION FUNCTIONS (*)	STOPPING PATIENCE (*)
INITIALIZATION FUNCTIONS (*)	POSITIVITY MULTIPLIER (*)

BAYESIAN METHODOLOGY

- We can now turn to the problem of inferring the PDF parameters $\mathbf{w}$ from the available data via numerical sampling

$$p(\mathbf{w}|\mathbf{D}) = \frac{p(\mathbf{D}|\mathbf{w})p(\mathbf{w})}{\mathcal{Z}}$$

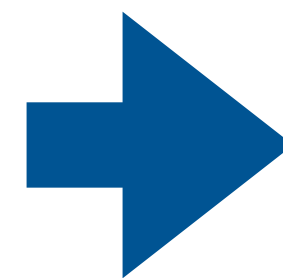
Likelihood $p(\mathbf{D}|\mathbf{w}) \propto \exp\left(-\frac{1}{2}[\chi_{\text{data}}^2(\mathbf{D}, \mathbf{w}) + \chi_{\text{pos}}^2(\mathbf{w}) + \chi_{\text{integ}}^2(\mathbf{w})]\right)$

Prior

Evidence or marginal likelihood

- How many parameters do we need?

$$Z = p(\mathbf{D} | \mathcal{M}_N) = \int d\mathbf{w} p(\mathbf{w} | \mathbf{D}, \mathcal{M}_N) p(\mathbf{w})$$



- Determine model(s) with largest evidence (balance between max likelihood and complexity)
- Bayesian average of models with best evidence
- Bayesian posterior update: start from large uniform prior and compute posterior analytically on linear data and use posterior as a prior to do numerical sampling of final posterior

Model with N weights = N vectors in the POD basis

$$\ln Z = -\frac{1}{2}\chi^2 + \frac{N}{2}\ln(2\pi) + \ln\left(\frac{\sqrt{|(X^T\Sigma^{-1}X)^{-1}|}}{\prod_i(b_i - a_i)}\right)$$

A LINEAR PDF PARAMETRISATION



Desirable qualities for candidate PDF space

- Should respect known theoretical constraints such as **small- and large-x scaling** and **sum rules**

$$f(x) \sim A x^{1-\alpha} (1-x)^{\beta}$$

- Should be flexible enough to explore the space of candidate PDFs amongst $C^1[0,1]^{n_f}$
- Should have a fair degree of complexity to efficiently fit model parameters

A NEURAL NETWORK ENSEMBLE TO SPAN PDF SPACE

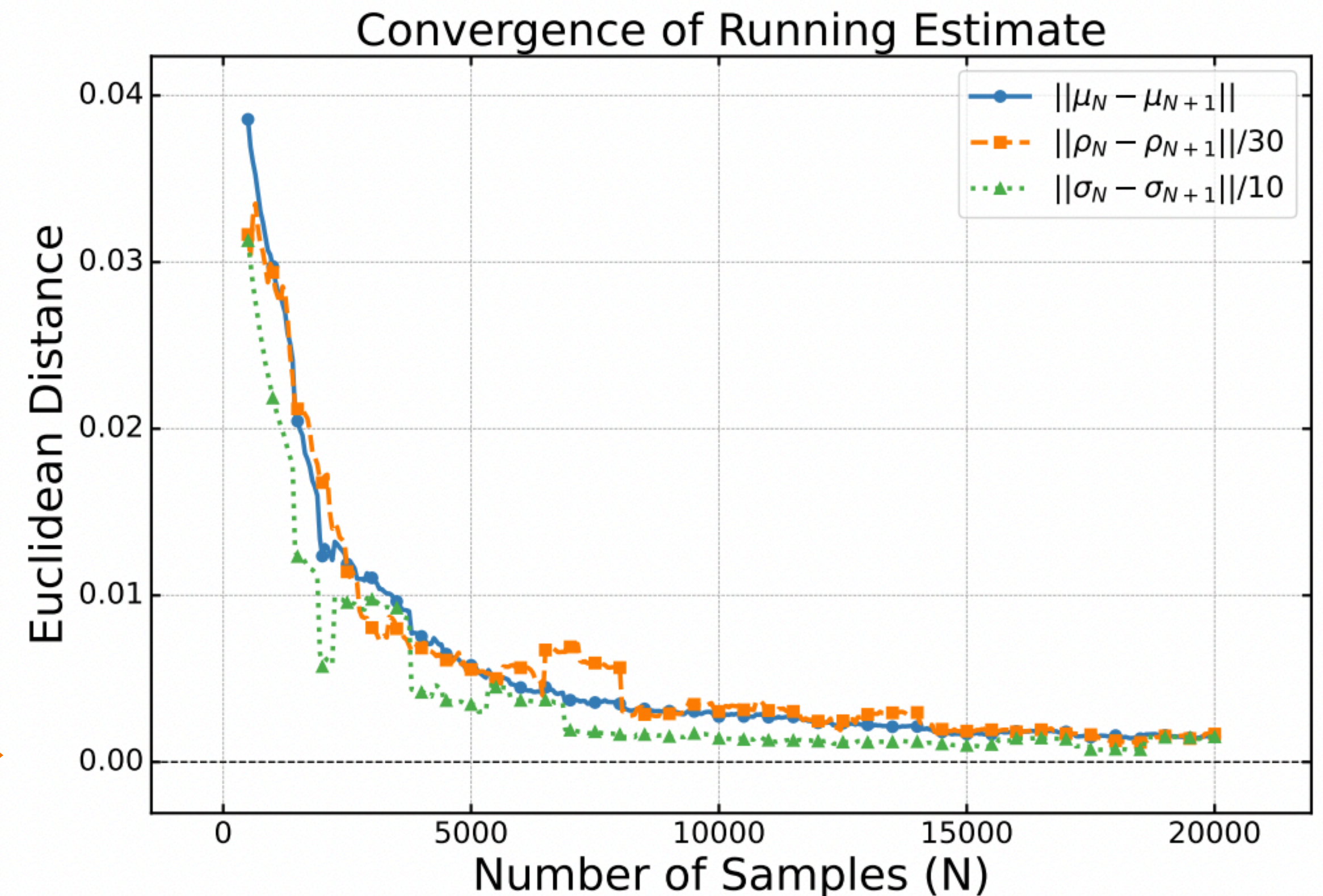
$$xg_m^j(x, Q_0; \theta) = A_m^j x^{1-\alpha_m^j} (1-x)^{\beta_m^j} \text{NN}_m(x; \theta),$$

$$m = 1, \dots, M$$

Pre-processing
exponents
randomly picked
in a suitable range
(uniformly)

Output of a randomly initialised NN
Here NNPDF4.0 architecture
(fully connected feedforward NN
with input layer of 2 neurons, two
hidden layers with 25 and 20
neurons & n_f outputs).
NN weights randomly initialised
with Glorot normal distribution

**M = 20K reproduces first
momenta of NN ensembles to
sufficient accuracy**



POD FOR A LINEAR PDF PARAMETRISATION

- Proper Orthogonal Decomposition (POD): decompose NN ensemble in its principal modes, from $M=20K$ samples to $N = O(10 - 100)$ basis functions $\rightarrow$ dimensional reduction of $\mathcal{H}$ to a linear functional space, where orthogonal basis functions $\{\varphi_k\}$ with $k = 1, \dots, N$ are organised in order of importance

$$f_{\mathbf{w}}(x) = \mathbf{w}^T \boldsymbol{\varphi}(x) = \varphi_0(x) + \sum_{k=1}^N w_k \varphi_k(x);$$

Basis functions
 $\boldsymbol{\varphi} = (\varphi_0(x), \varphi_1(x), \dots, \varphi_N(x))$

Weights = parameters $\mathbf{w} = (1, w_1, \dots, w_N)$

- POD provides most efficient way of capturing key features of an infinite-dimensional ensemble of functions using only a finite, often surprisingly small, number of 'modes'.
- Theoretical constraints (sum rules, integrability..) can be imposed directly on the weights $\mathbf{w}$

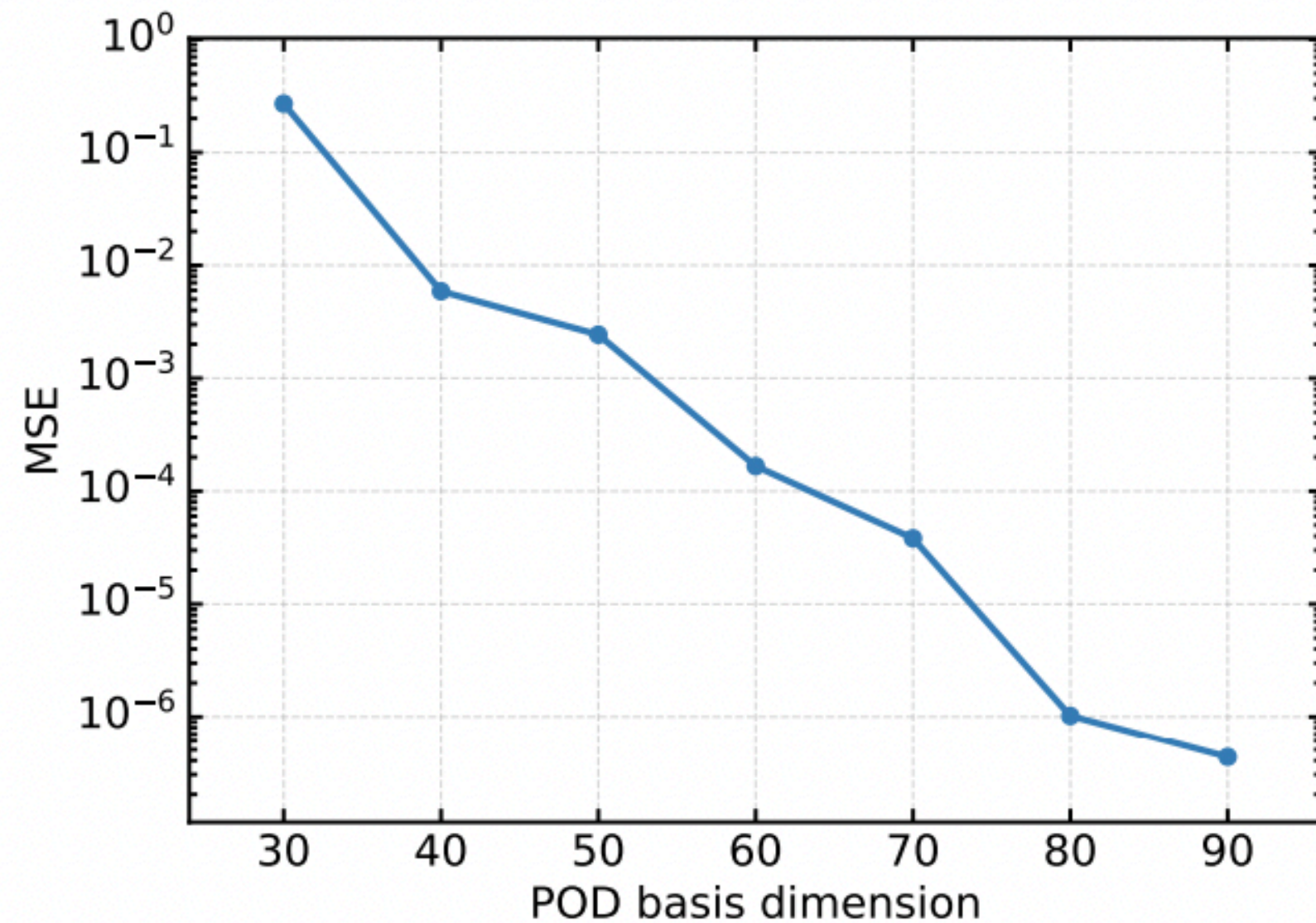
ASSESSING COMPLETENESS AND GENERALIZATION

- Completeness: any PDF produced using a random initialisation of the NN can be well-approximated by the linear mode in POD basis
- Generalization: any PDF parametrisation (from all available in LHAPDF) can be well-approximated by the linear model in POD basis

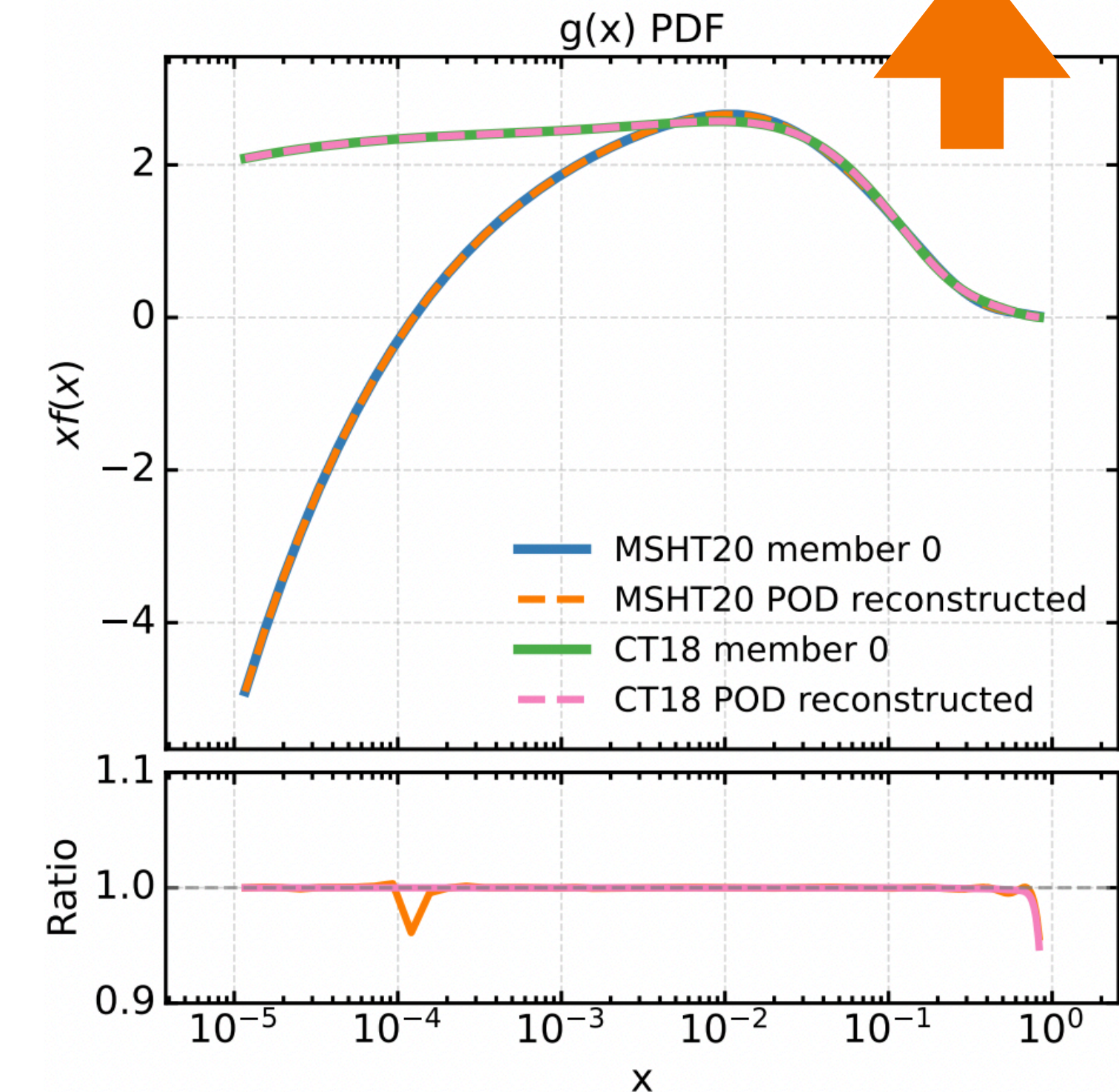
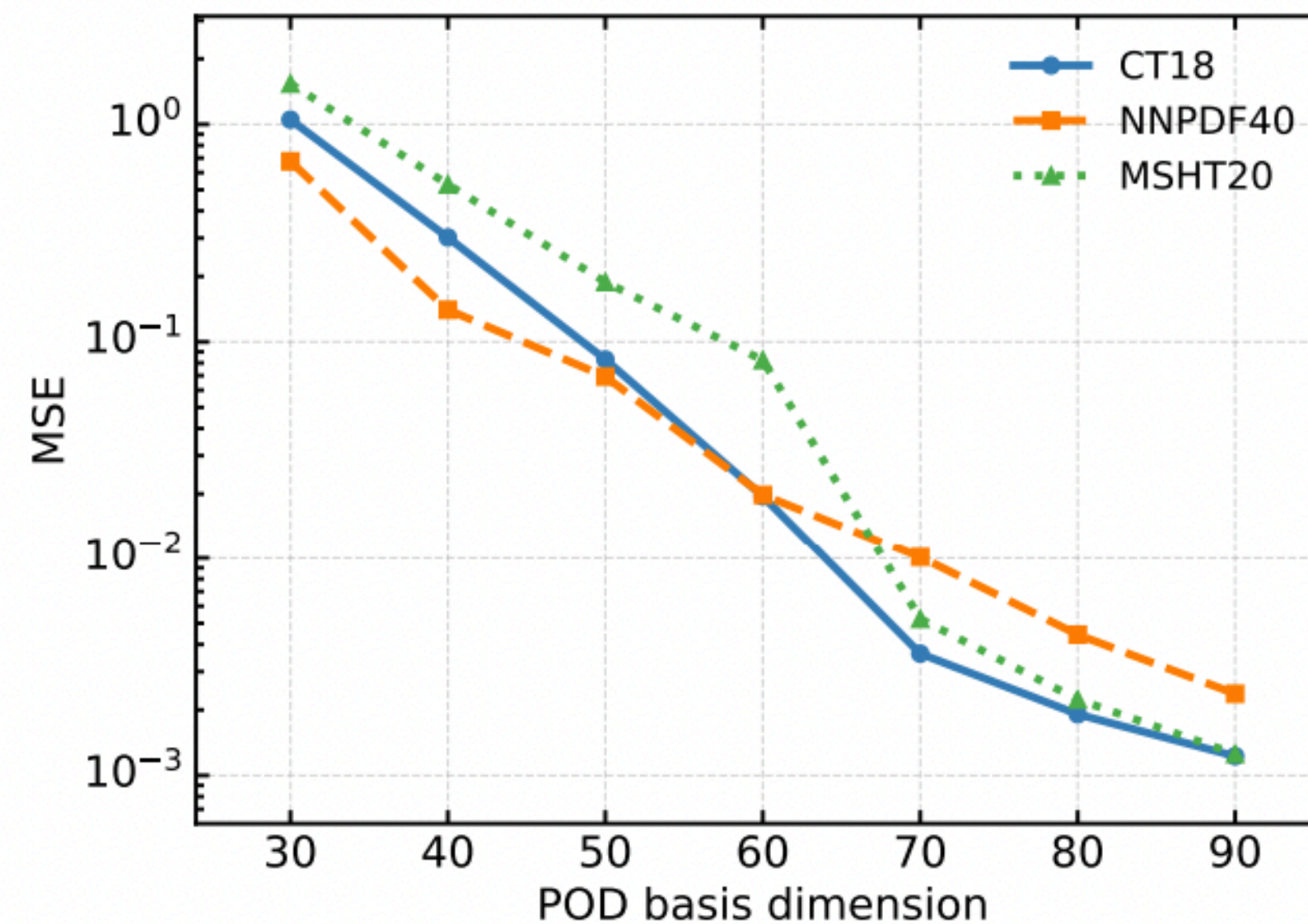
$$\text{MSE} = \frac{1}{N_{\text{targ}}} \sum_{i=1}^{N_{\text{targ}}} \left\| \mathbf{f}_T^{(i)} - \mathbf{f}_{\mathbf{w}^{(i)}} \right\|_2^2 = \frac{1}{N_{\text{targ}}} \sum_{i=1}^{N_{\text{targ}}} \left\| \mathbf{f}_T^{(i)} - \left(\varphi_0 + \sum_{n=1}^N w_n^{(i)} (\varphi_n - \varphi_0) \right) \right\|_2^2$$

80 weights enough to reproduce other parametrisation

COMPLETENESS



GENERALIZATION



APPLICATION: A CLOSURE TEST OF DIS FIT

- Underlying law: model obtained with N=40 weights

$$\mathbf{f}_{\text{in}} = \varphi_0 + \sum_{k=1}^N w_k^{\text{in}} (\varphi_k - \varphi_0)$$

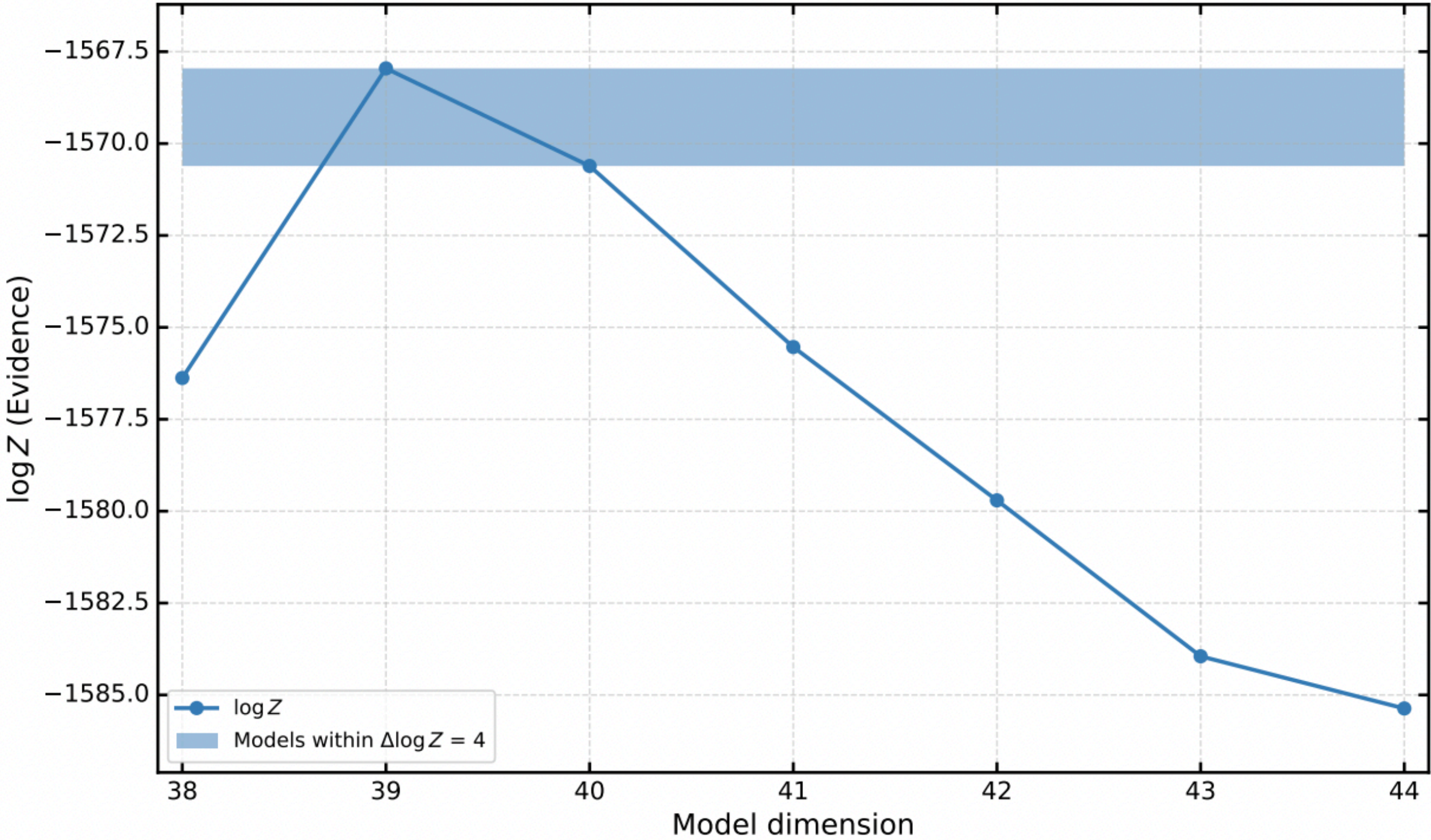


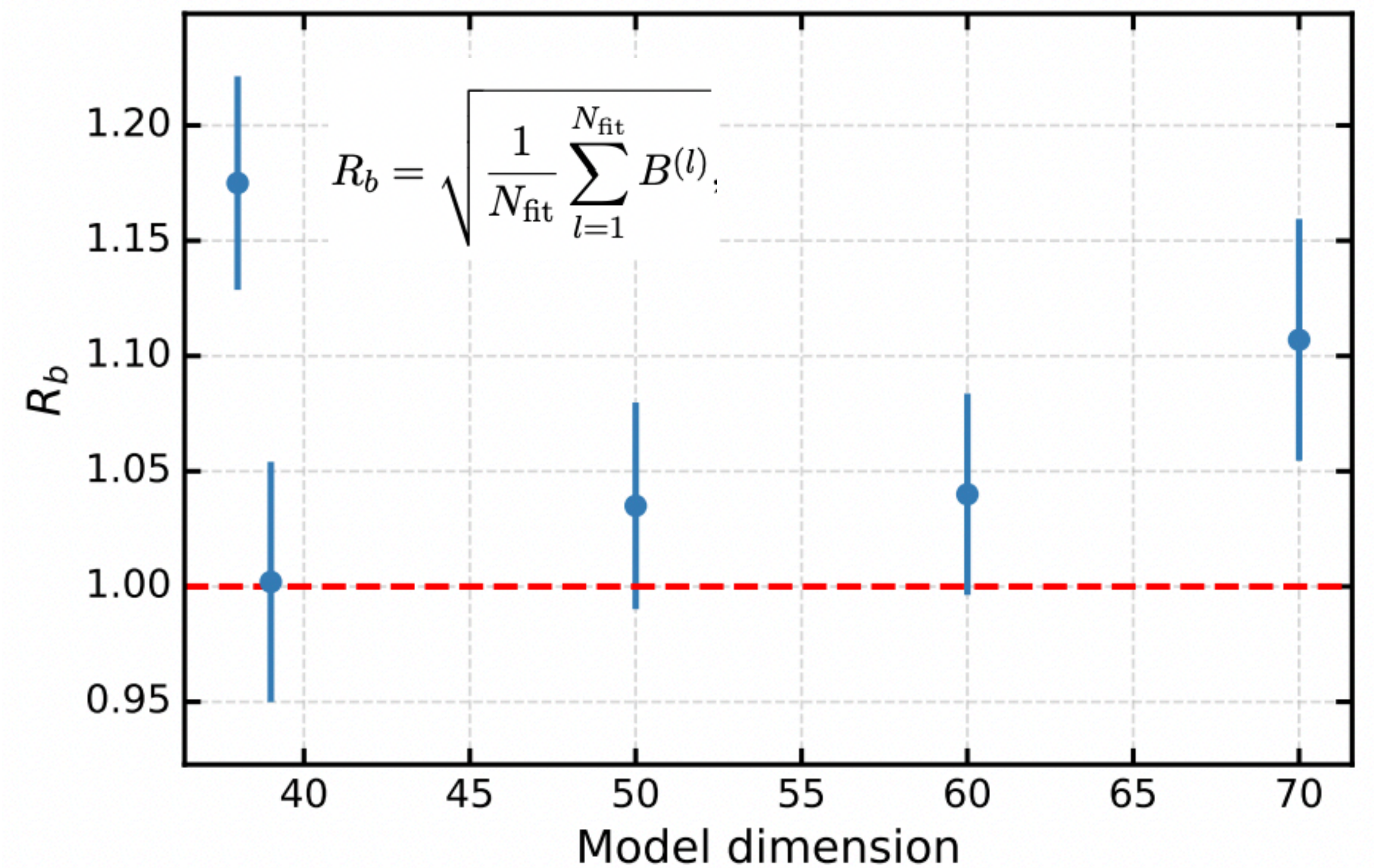
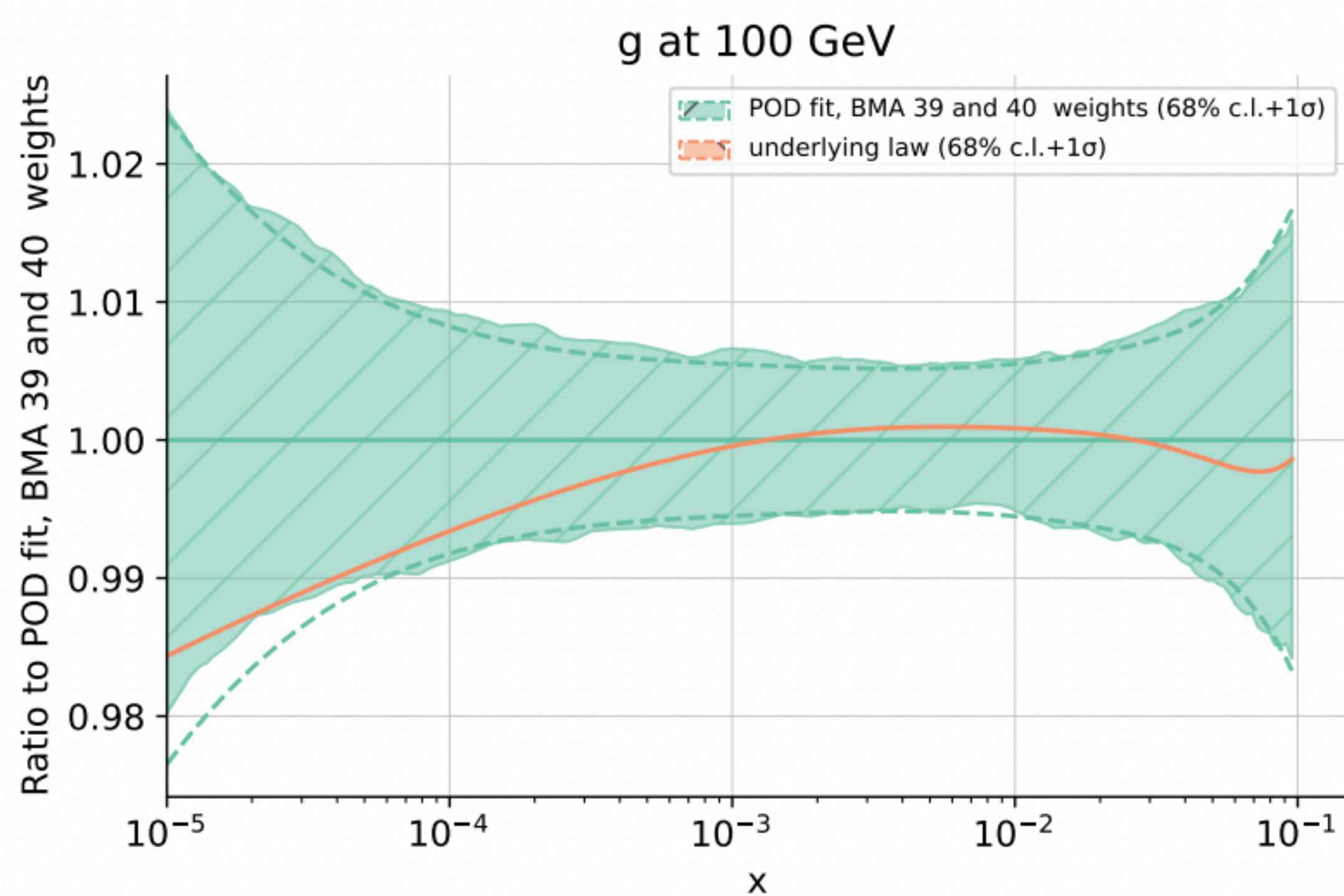
Table 3: Goodness-of-fit and model posterior probabilities for the two most likely models.

Model Complexity	χ^2/N_{data}	$p(\mathcal{M}_k D)$
39	1.033	0.934
40	1.032	0.066

Dataset	N_{data}	Experiment	Model
<i>Numerical fit</i>			
NMC F_2^d/F_2^p [55]	121	Deuteron	Ratio
SLAC F_2^p [56]	33	Deuteron	Linear
SLAC F_2^d [56]	34	Deuteron	Linear
BCDMS F_2^p [57]	333	Deuteron	Linear
BCDMS F_2^d [57]	248	Deuteron	Linear
<i>Analytical fit</i>			
(*) NMC $\sigma^{\text{NC},p}$ [58]	204	NMC	Linear
CHORUS σ_{CC}^ν [59]	416	Nuclear	Linear
CHORUS $\sigma_{CC}^{\bar{\nu}}$ [59]	416	Nuclear	Linear
NuTeV dimuon σ_{CC}^ν [60]	39	Nuclear	Linear
(*) NuTeV dimuon $\sigma_{CC}^{\bar{\nu}}$ [60]	37	Nuclear	Linear
HERA I+II $\sigma_{CC}^{e^+p}$ [61]	39	HERA	Linear
HERA I+II charm σ_{NC} [62]	37	HERA	Linear
HERA I+II $\sigma_{NC}^{e^-p}$ (320 GeV)	159	HERA	Linear
(*) HERA I+II $\sigma_{NC}^{e^+p}$ (460 GeV)	204	HERA	Linear
HERA I+II $\sigma_{NC}^{e^+p}$ (575 GeV)	254	HERA	Linear
HERA I+II $\sigma_{NC}^{e^+p}$ (820 GeV)	70	HERA	Linear
HERA I+II $\sigma_{NC}^{e^+p}$ (920 GeV)	377	HERA	Linear
(*) HERA I+II $\sigma_{CC}^{e^-p}$ [61]	42	HERA	Linear
HERA I+II bottom σ_{NC} [62]	26	HERA	Linear
Total	3089		

APPLICATION: A CLOSURE TEST OF DIS FIT

M. Costantini, L. Mantani, J. Moore, MU, arXiv:2507.16913



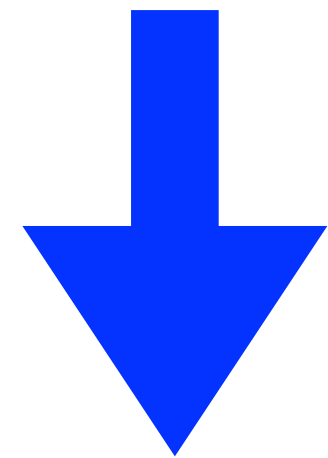
$$B^{(l)} = \frac{1}{N_{\text{data}}} \sum_{i,j=1}^{N_{\text{data}}} \left(\mathbb{E}_{\epsilon} T_i[f_k^{(l)}] - D_i^{L_0} \right) (C_{\text{PDF}})_{ij}^{-1} \left(\mathbb{E}_{\epsilon} T_j[f_k^{(l)}] - D_j^{L_0} \right)$$

- Model-averaged posterior successfully recovers underlying law within estimated uncertainties.
- Posterior from average of models with best evidence faithfully reproduces statistical distribution of the data
- Under- and over-parametrised models systematically underestimate uncertainties in data region.

EXTRACTING PARAMETERS FROM DATA

$$\chi^2 = \sum_{a,b=1}^{N_{\text{dat}}} (T_a(\{f\}, \vec{c}) - D_a) \Sigma_{ab}^{-1} (T_b(\{f\}, \vec{c}) - D_b)$$

$$T^{(\text{ep})}(\{f\}, \vec{c}) = \sum_{i=1}^{n_{\text{partons}}} \int_x^1 dy f_i(y; \vec{c}) \sigma_i \left(\frac{x}{y}; \vec{c} \right) \equiv \sum_{i=1}^{n_{\text{partons}}} [f_i \otimes \sigma_i](x; \vec{c})$$



PDFs



SM parameters ($\alpha_s(M_Z)$, M_W , θ_W ...), or BSM parameters (SMEFT WCs ...)



$$T(\{f\}, c) = T^{\text{SM}}(\{f\}, c = 0)(1 + k_{\text{lin}} \cdot c + k_{\text{quad}} \cdot c^2)$$

✓ In a PDF fit typically

$$T(\{f\}) = T^{\text{SM}}(\{f\}, c = 0)$$

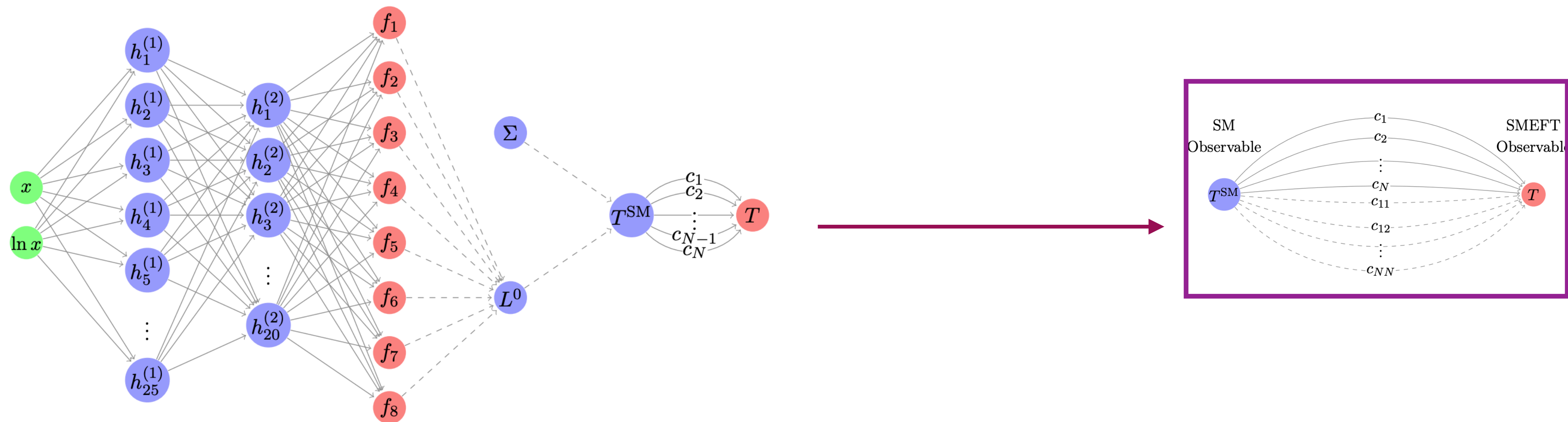
✓ In a fit of SMEFT Wilson Coefficients

$$T(c) = T^{\text{SM}}(\{f\} = \{f\}^*)(1 + k_{\text{lin}} \cdot c + k_{\text{quad}} \cdot c^2)$$

THE SIMUNET SOLUTION

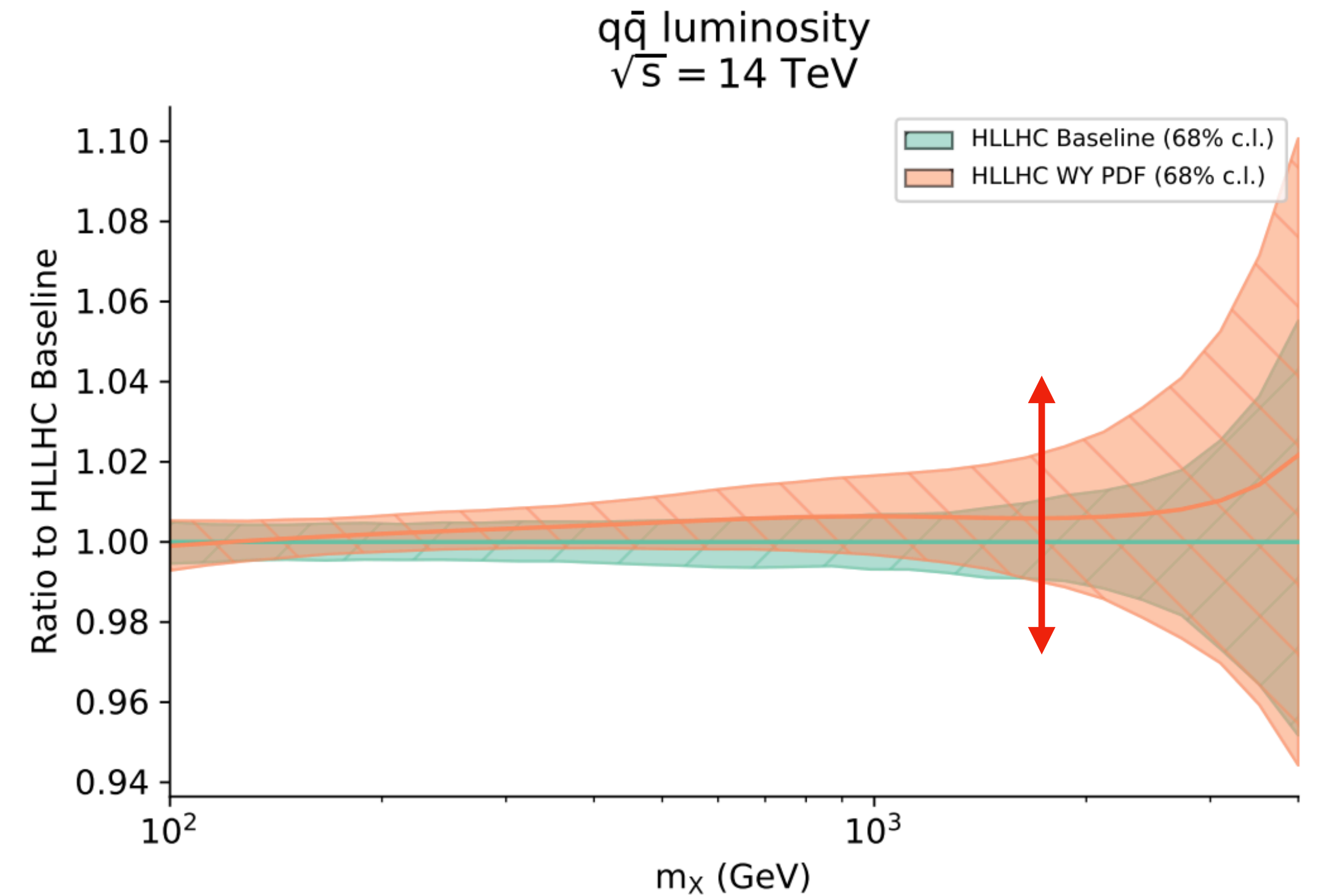
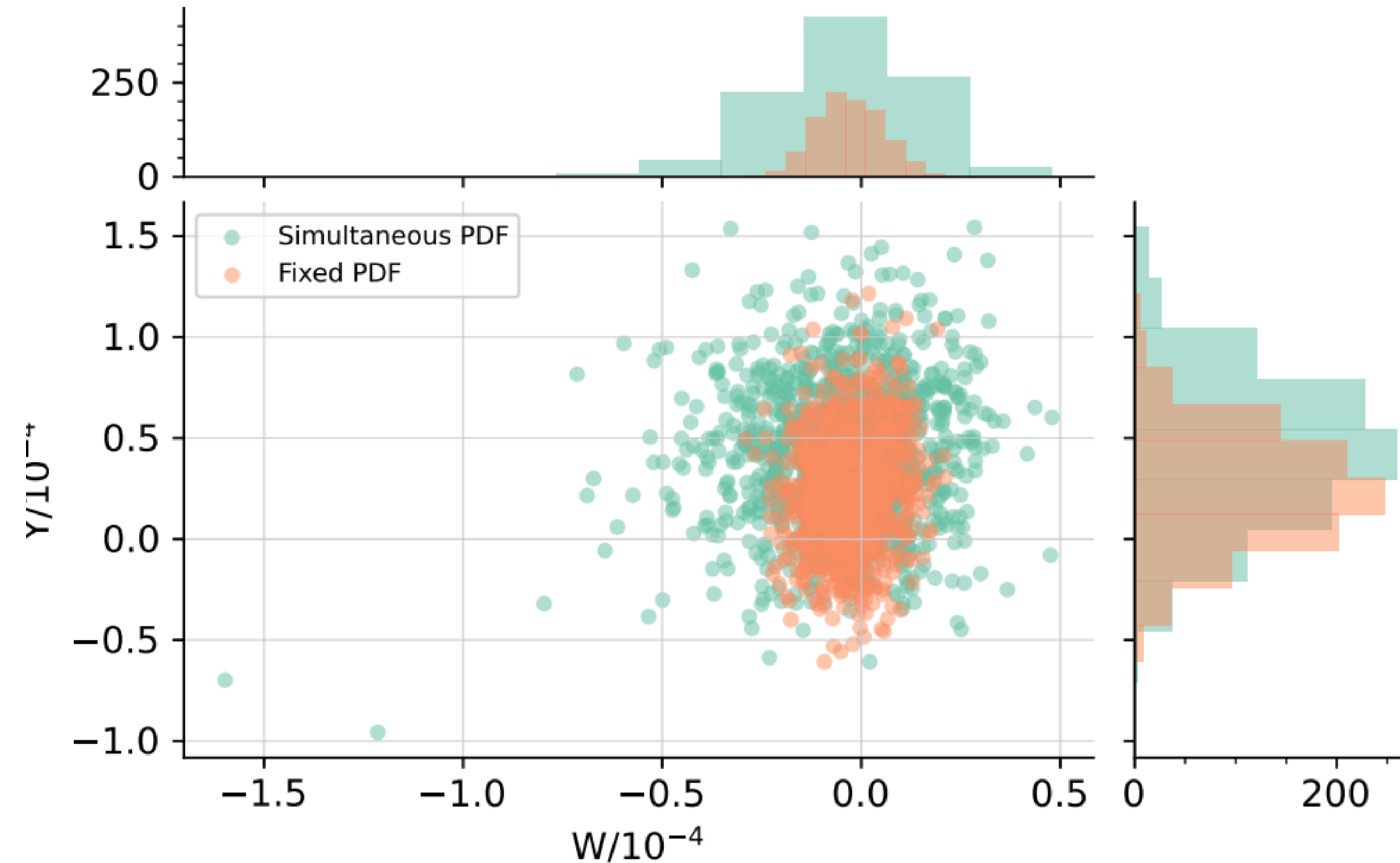
- ▶ The idea: take a PDF fit based on NNPDF4.0 methodology and make dependence of observables on physics parameters $\{c_i\}$ explicit before computing the loss function (e.g. adding SMEFT corrections, or expanding observables in terms of SM precision parameters)
- ▶ Perform minimisation of loss function over $\hat{\theta} = \theta \cup \{c_i\}$ by adding new layer to the deep neural network used in NNPDF4.0

Input layer Hidden layer 1 Hidden layer 2 PDF flavours Convolution step SM Observable SMEFT Observable



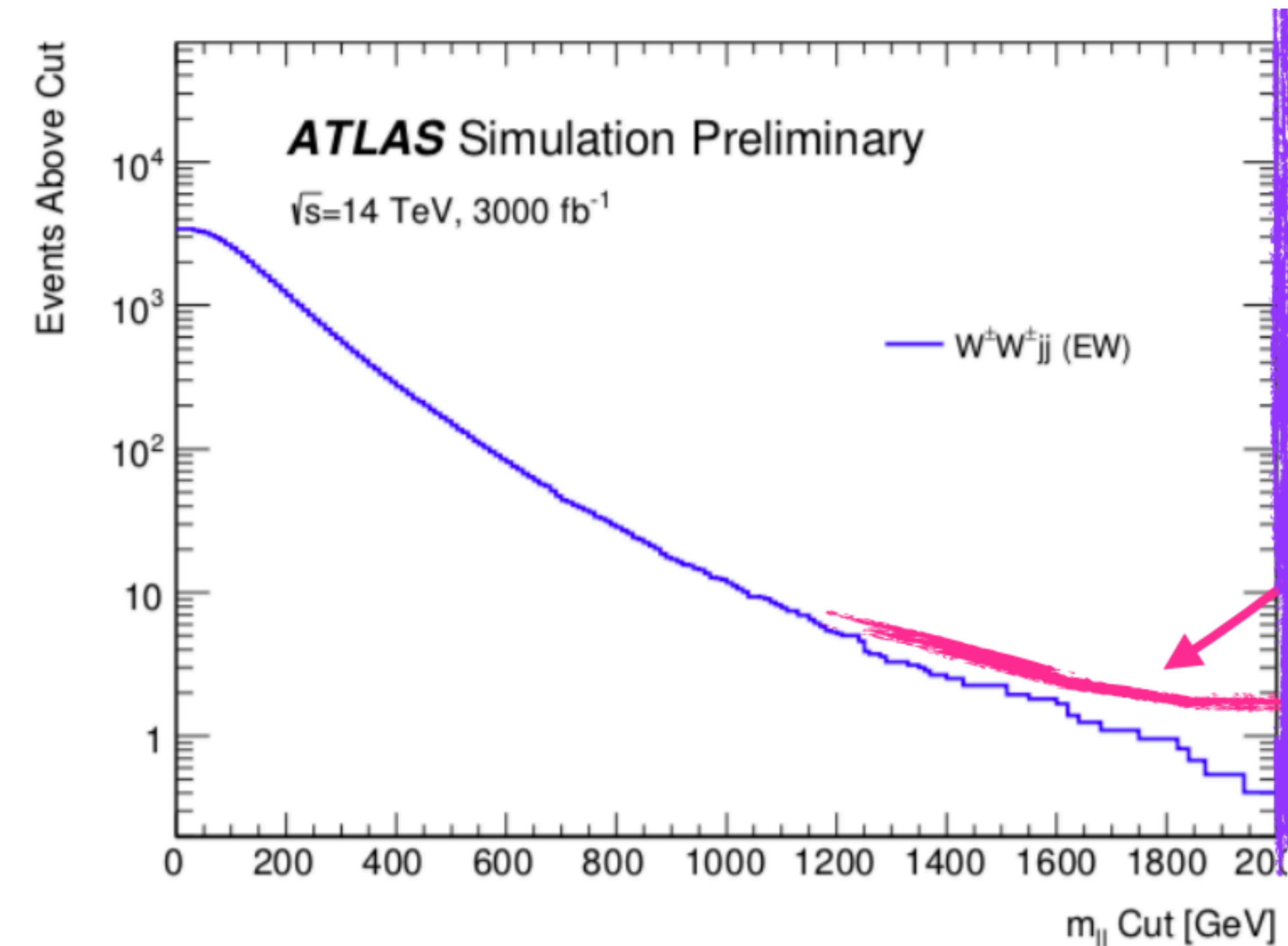
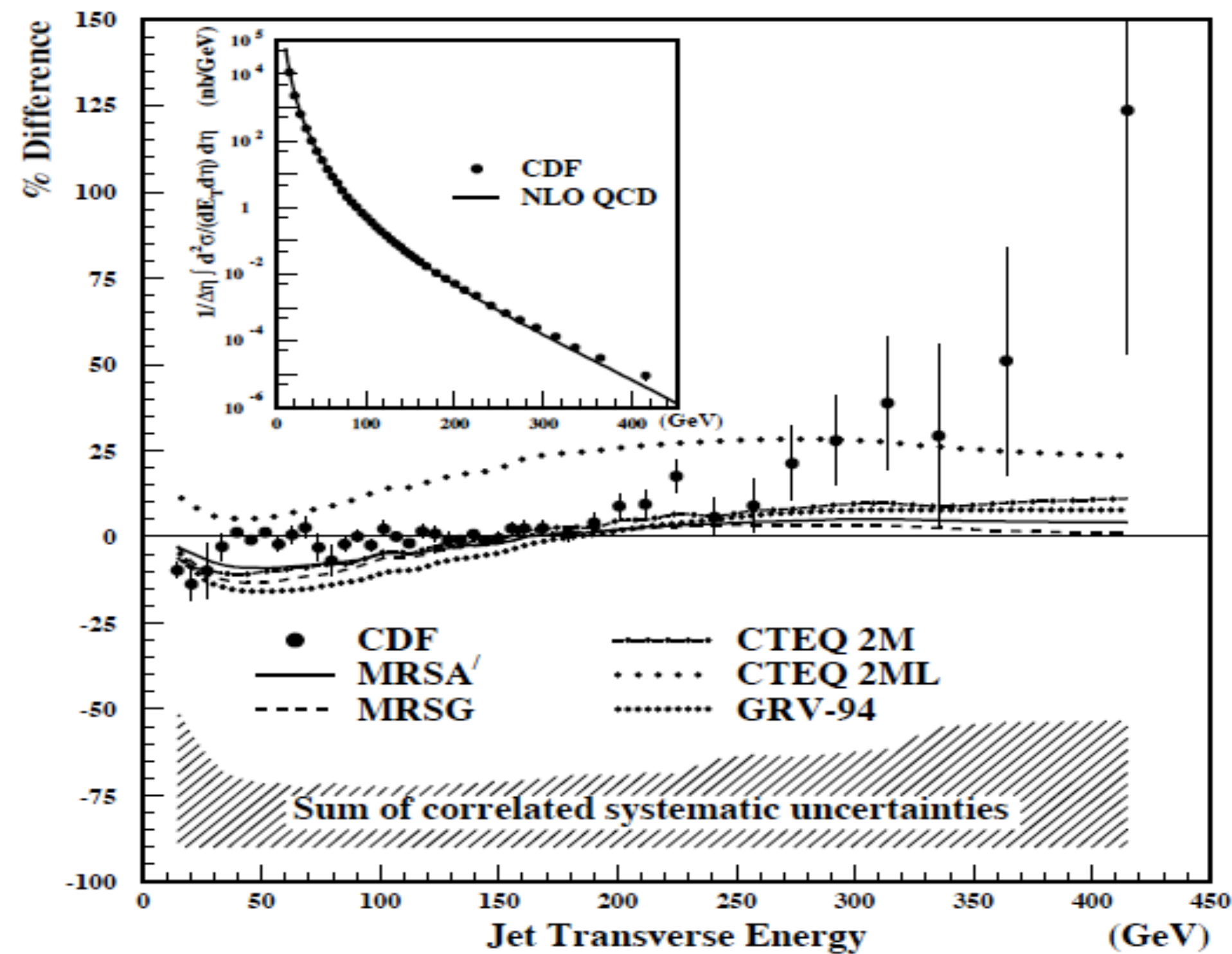
EXAMPLE: DRELL-YAN DATA @HL-LHC

S. Iranipour, MU - arXiv: 2201.07240



- ✓ Simultaneous analysis of PDFs and W&Y SMEFT coefficient using simuNET method shows that at HL-LHC the effect of interplay becomes important as WCs bounds broaden and PDF uncertainties change significantly once SMEFT effects allowed in theory predictions entering PDF fit
- ✓ In this study the dependence on c was **linear**

FRONTIER: LARGE-X PDFs AND HIGH-ENERGY DATA



CA Lee, HL/HE-LHC Jamboree, 1 March 2019

➡ Discrepancy between QCD calculation and CDF jets data (1995)

At that time no information on PDF uncertainties and theory predictions strongly depended on gluon shape at $x > 0.1$.
Once data included in the CTEQ fit, discrepancy disappeared.

➡ Deviations from SM predictions in high energy tails (>2025)

New physics or limited understanding of proton structure?