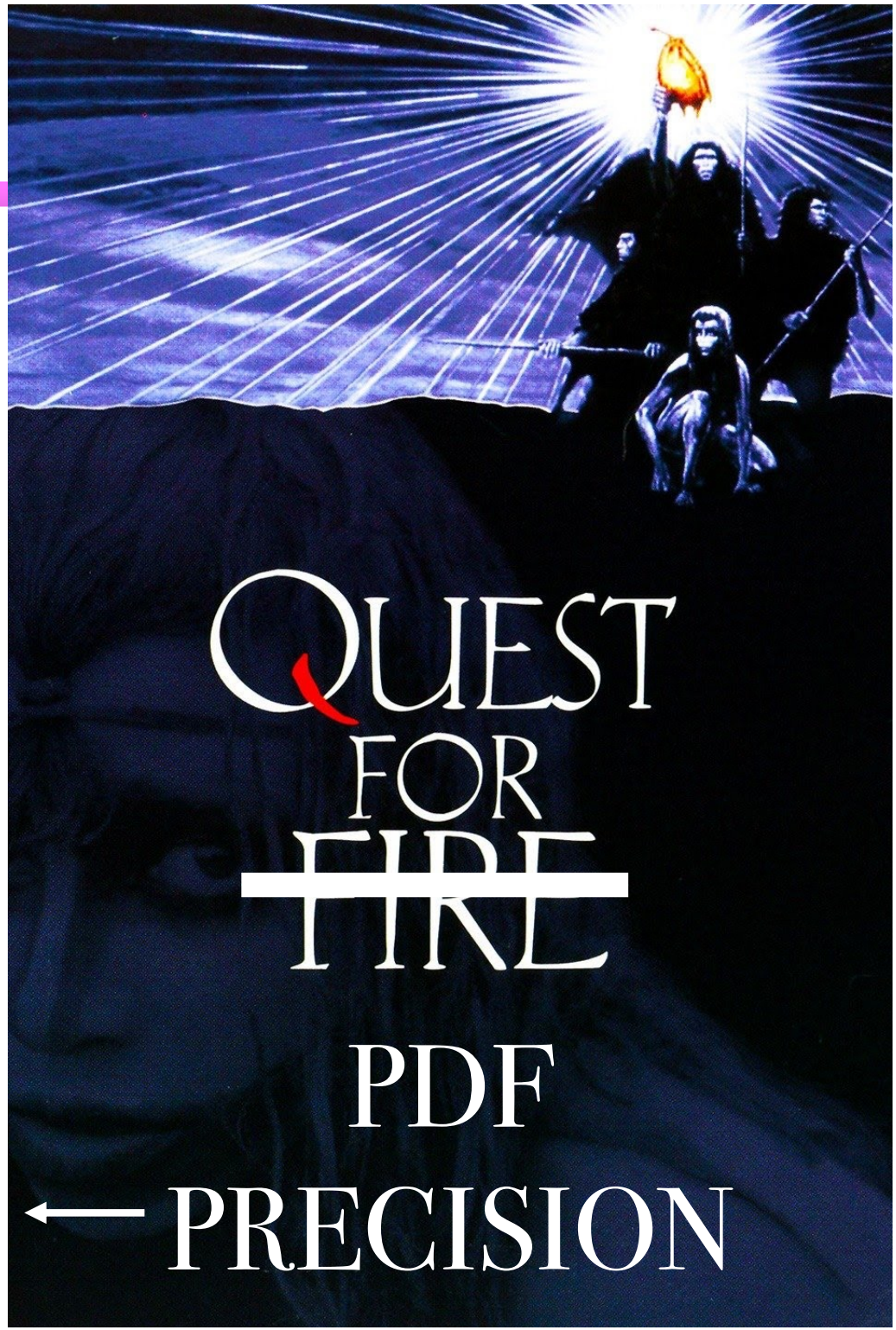
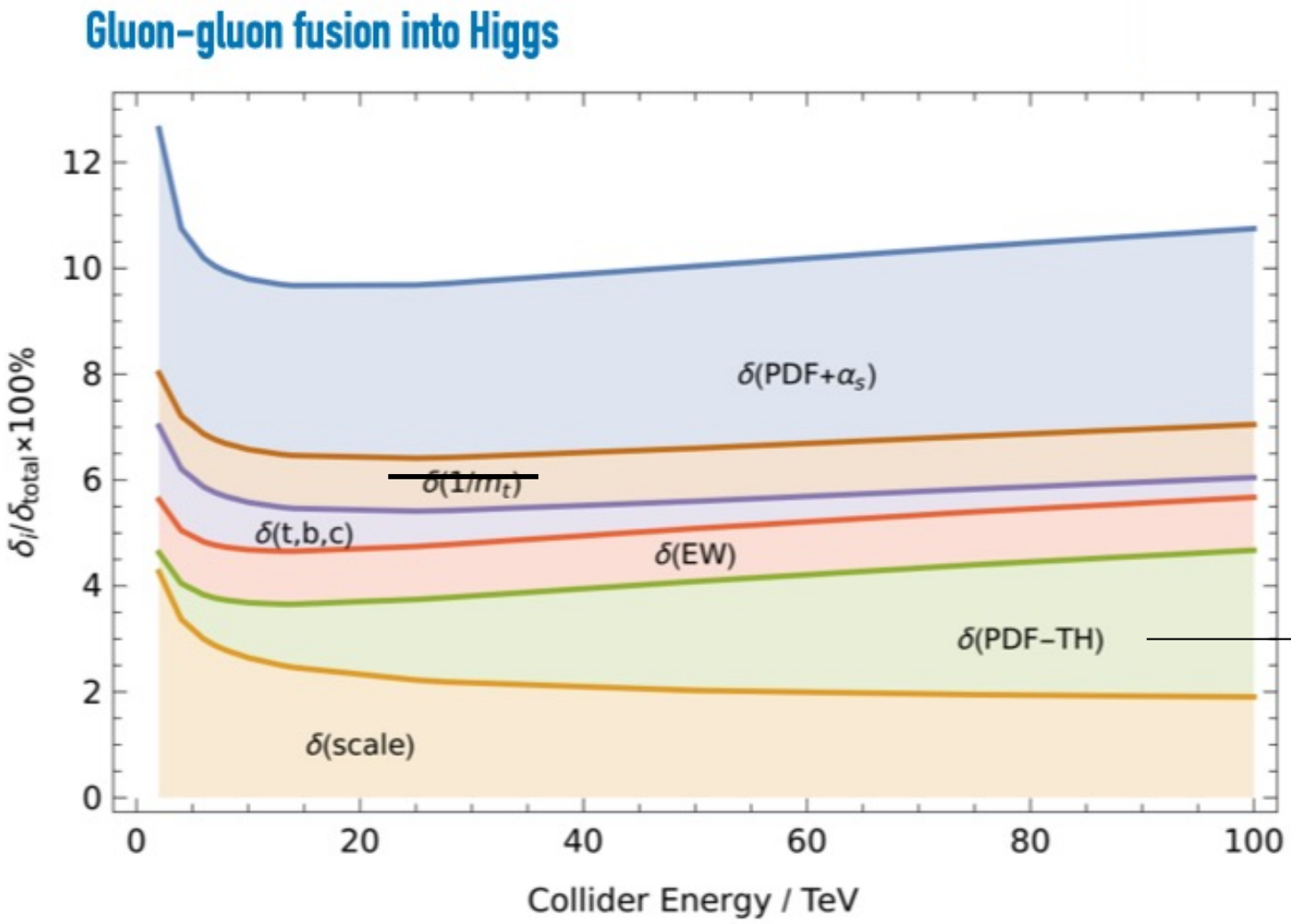
J. Huston
Michigan State University

HP2 Workshop
Newcastle, UK
Sept. 2022

and accuracy



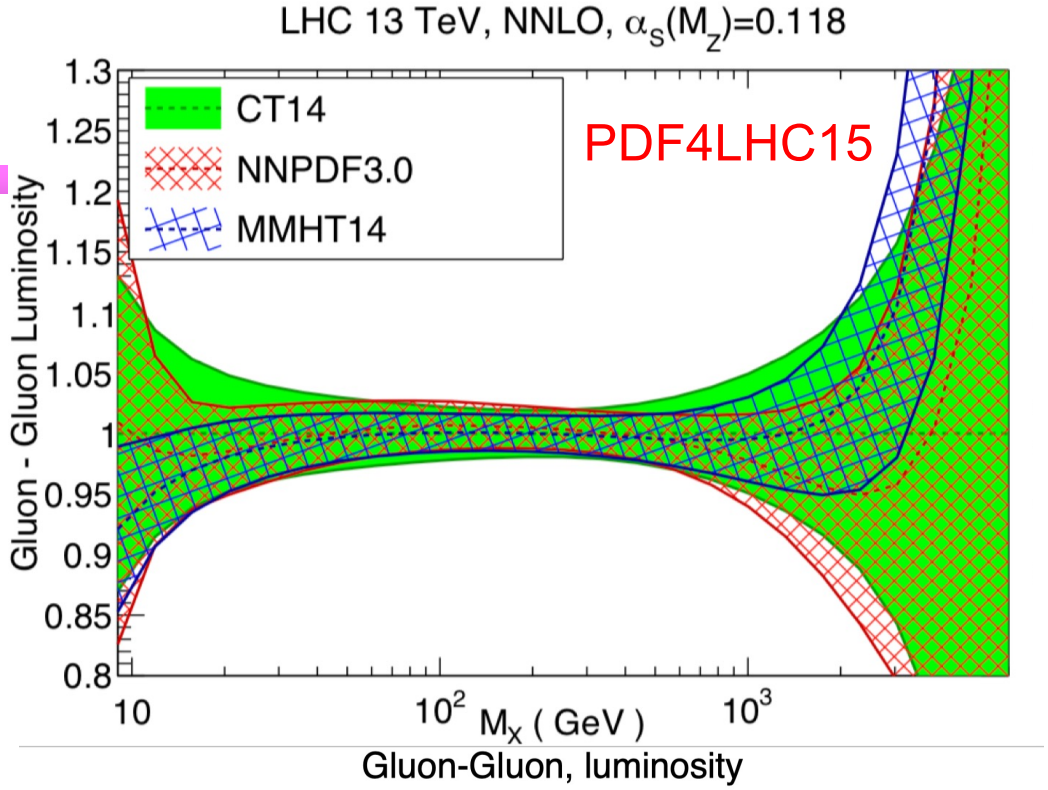
Better precision for PDFs easy to motivate

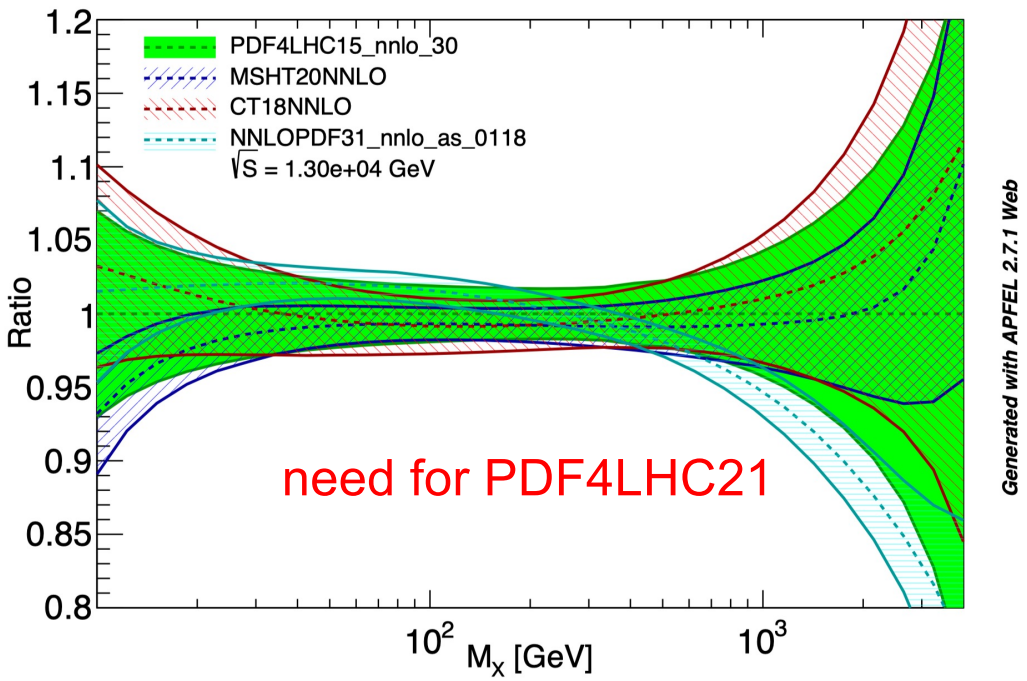


see Robert's talk

Global PDF fits

- Much of the information regarding PDFs comes from the three global PDF fit groups: CT, MSHT, NNPDF
- CT and MSHT both use a Hessian-based approach (for the determination of the central PDF and the uncertainties), while NNPDF uses a Monte Carlo replica approach (although the Monte Carlo replica basis can be converted into a Hessian basis, and indeed this is often the format that allows the easiest understanding of the uncertainties)

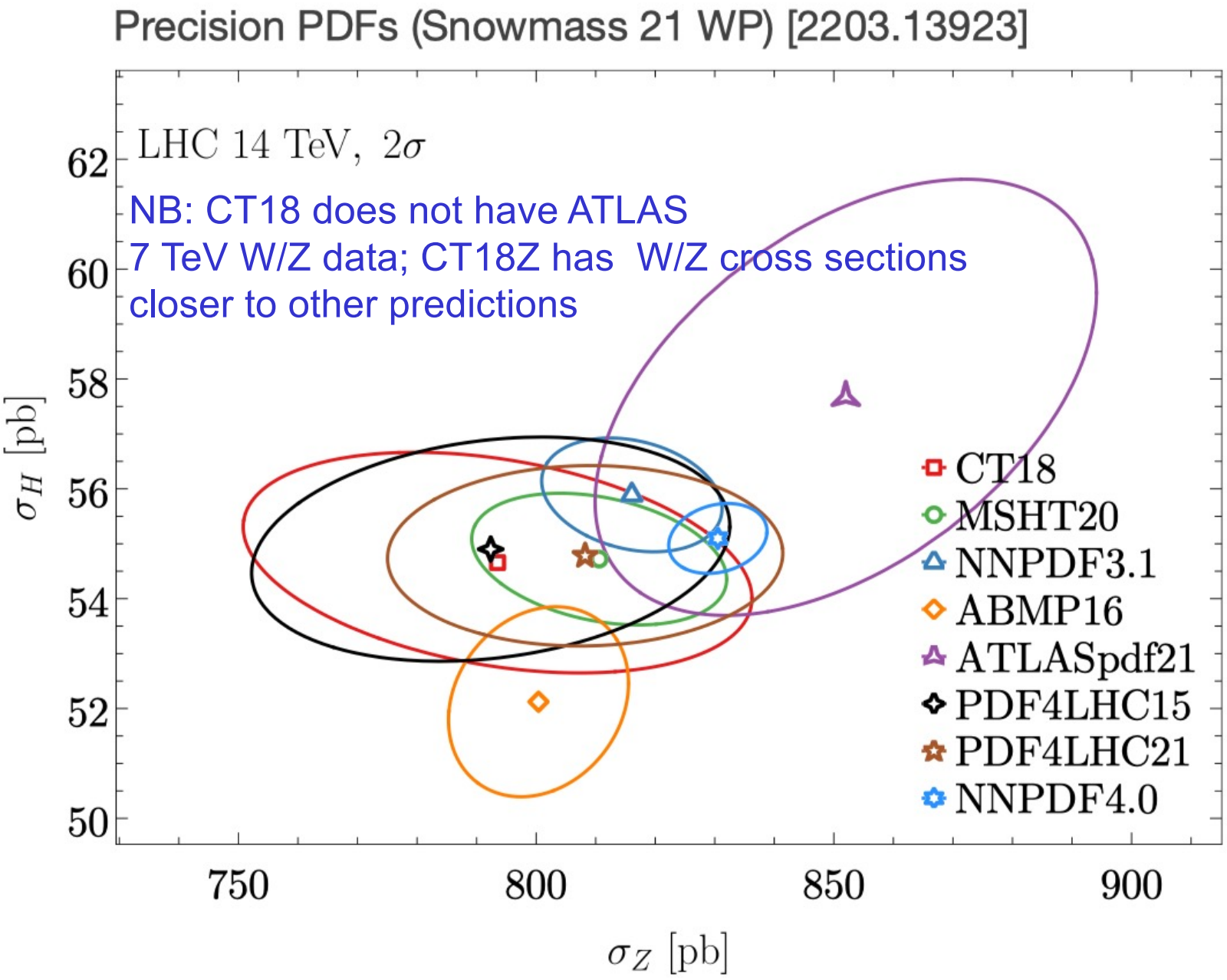




Aside: uncertainties

- PDF uncertainties depend first of all on the experimental uncertainties of the data
- Data from two measurements, or even from within the same measurement, can both be very precise, but the result of adding both to the PDF fit can be an increase in the PDF uncertainty (or more likely) a smaller decrease in uncertainty than expected) if the data are in tension with each other
- The resultant PDF uncertainty relies on the definition of a tolerance, i.e. what is a significant increase from the global minimum χ^2, i.e. PDF uncertainty can be adjusted by changing the tolerance
- $\Delta\chi^2=1$ is not applicable for ~4000 data points from different experiments
- NB: all groups see tensions; the relevant χ^2 values show that the fits do not correspond to zero tension (see tables in PDF4LHC21 doc)
- NB: CT (Tier 2) and MSHT (dynamic tolerance) have introduced criteria to restrict the pull of data sets that disagree with global fit
  - MSHT criterion is sometimes stricter

Useful to look at pairs of cross sections



NNPDF3.1 uncertainty is smaller than either CT18 or MSHT20

NNPDF4.0 uncertainty is smaller still

NNPDF4.0 Higgs σ is closer to CT18, MSHT20

What is the best estimate of the uncertainty? and the central value?

Reduced data set fits: PDF4LHC21

- Diverse enough to provide information for all PDFs
- Sparse enough that uncertainties should be very similar for all 3 PDFs
- Origins of differences of PDFs
  - due to variations of experimental input, treatment of systematic errors, different theory settings, fitting methodologies?
  - so for benchmarking, use common theory settings (i.e. perturbative charm, m_{charm}=1.4 GeV, s=sbar at input scale, $\alpha_s(m_Z)$=0.118, positive-definite PDFs, no deuteron or nuclear corrections…)
  - add several data sets to NNPDF3.1->3.1' (closer to 4.0)

Dataset	Reference	Dataset	Reference
BCDMS proton, deuteron DIS	[155, 156]	LHCb 8 TeV $Z \to ee$	[62]
NMC deuteron to proton ratio DIS	[157]	ATLAS 7 TeV high precision W, Z (2016)	[63]
NuTeV νN dimuon	[158]	D0 Z rapidity	[159]
HERA I+II inclusive DIS	[60]	CMS 7 TeV electron asymmetry	[160]
E866 Drell-Yan ratio pd/pp DIS	[161]	ATLAS 7 TeV W, Z rapidity (2011)	[149]
LHCb 7, 8 TeV W, Z rapidity	[61, 65]	CMS 8 TeV inclusive jet	[69]

Table 3.1. The measurements included in the initial round of PDF fits to a reduced dataset, together with the corresponding publication reference. This dataset is chosen as the largest subset of data fit by CT18, MHST20, and NNPDF3.1 in an (almost) identical manner.

Reduced fits

- Central values agree reasonably well
- …as do uncertainties at higher x
- There are some differences, for example at low x for the gluon distribution; this is a region nominally not well constrained by data

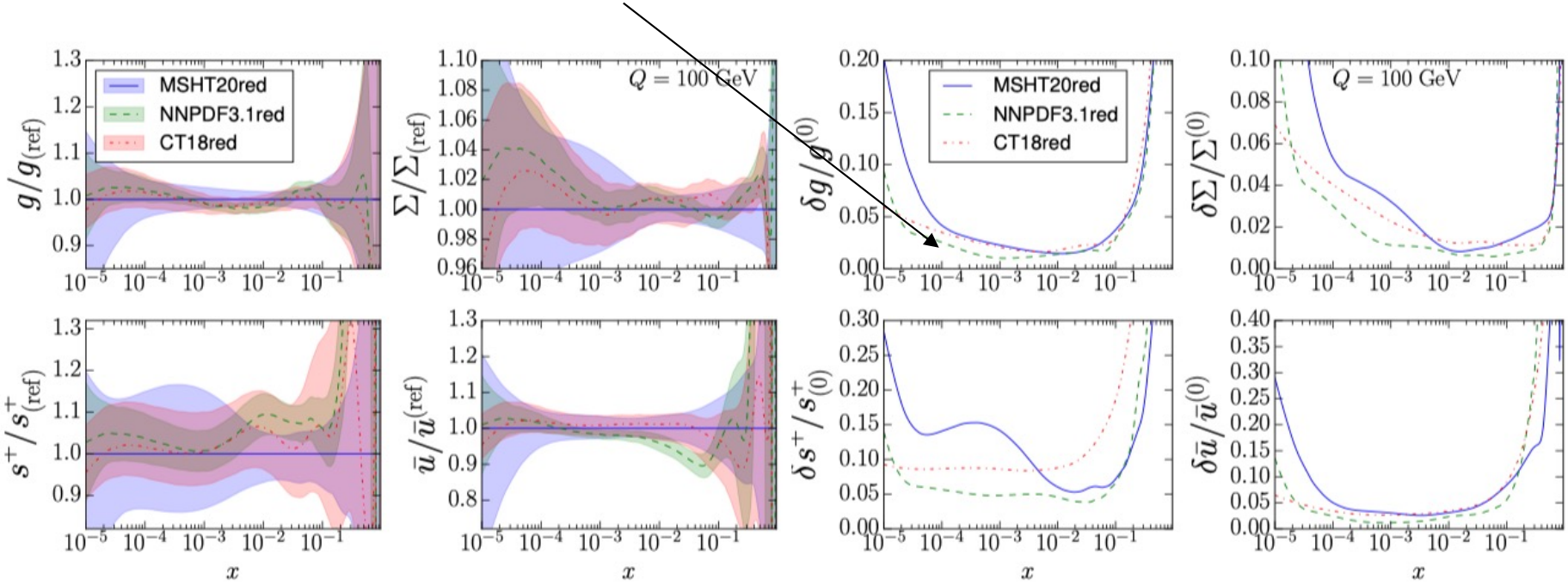
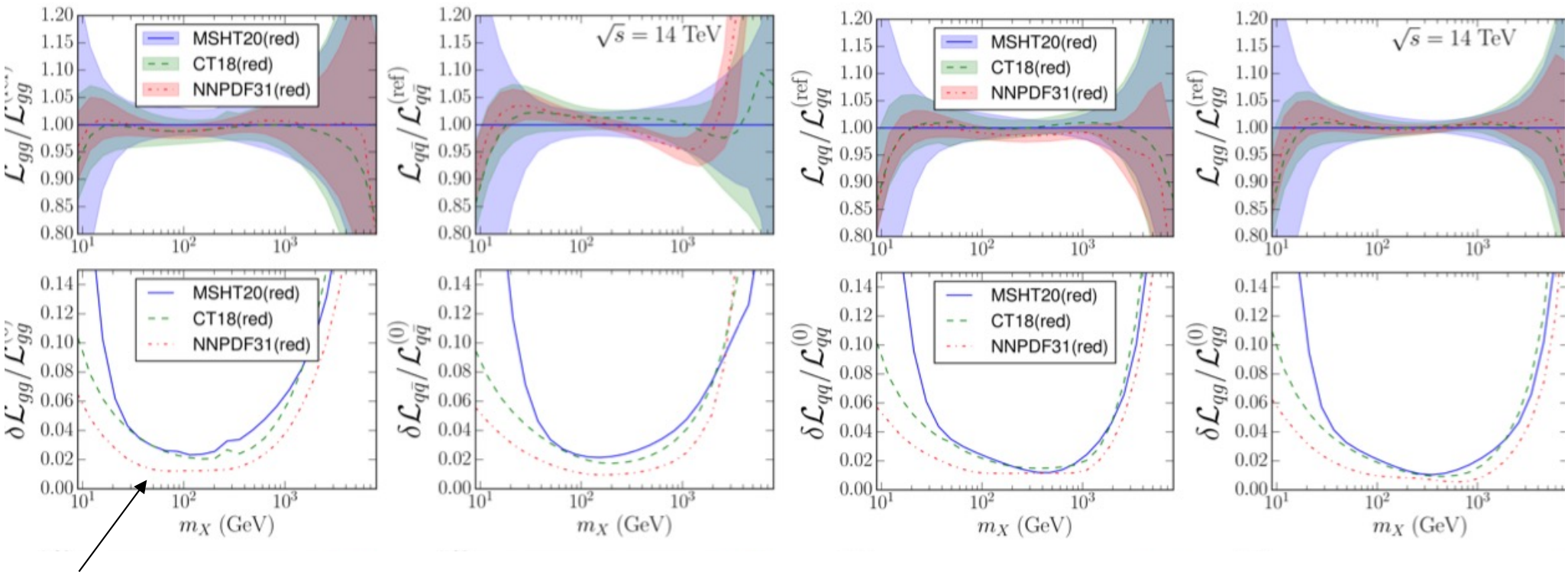


Figure 3.4. Comparison between the reduced PDF fits from the three groups, in the same format as in Fig. 3.1. For the three groups, PDF errors correspond to 1σ intervals. In the left panels, PDFs are displayed normalised to the central value of the MSHT20 reduced PDF set.

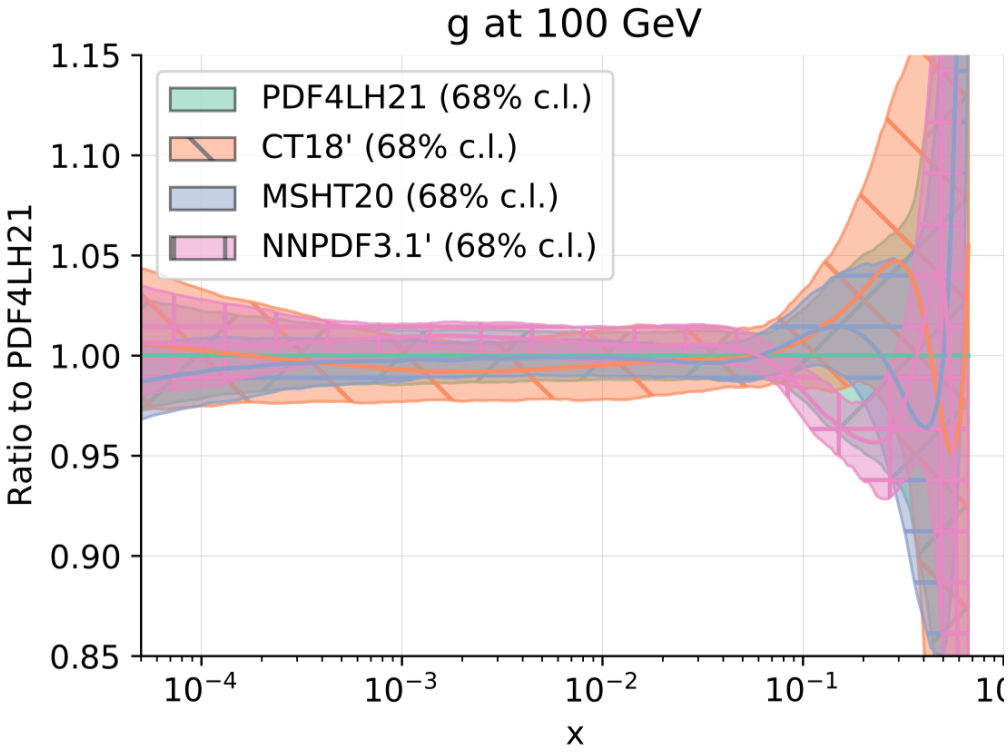
PDF luminosities for reduced fits

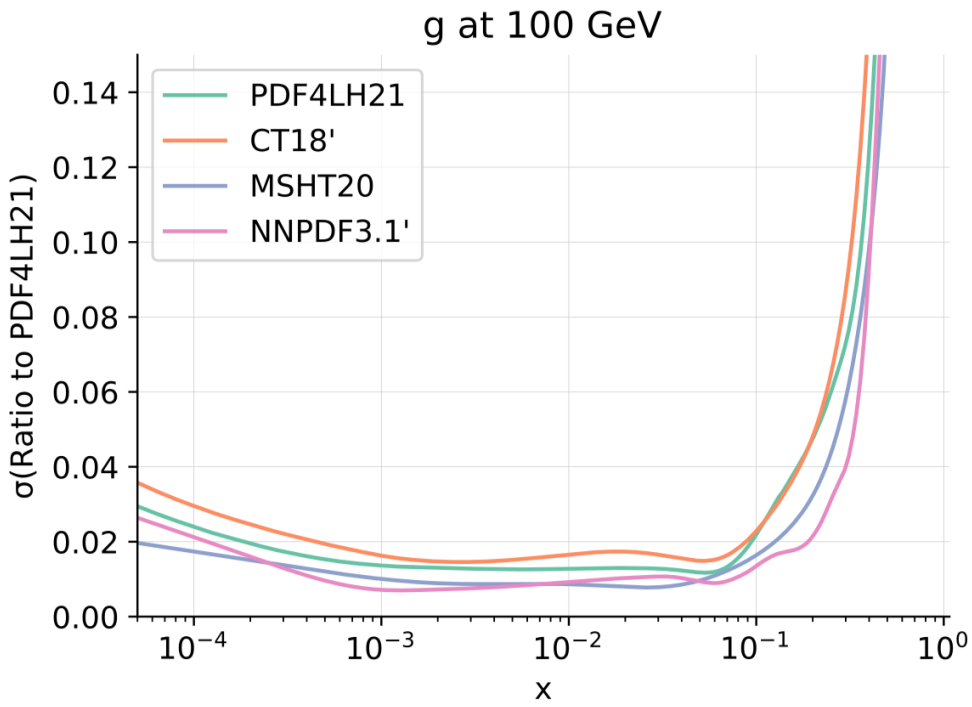


NNPDF3.1 has significantly reduced gg uncertainty using the same set of data; this implie their effective tolerance (for the same data information) is smaller than for CT or MSHT; the effect is even larger with NNPDF4.0. Due just to smaller gluon uncertainty? Maybe correlations are also different?

Figure 3.5. Comparison of the partonic luminosities between the CT18, MSHT20, and NNPDF3.1 reduced fits at $\sqrt{s} = 14$ TeV as a function of the invariant mass of the produced final state m_X. From left to right we show the gluon-gluon, quark-antiquark, quark-quark and quark-gluon luminosities, normalised to the central value of the MSHT20 prediction, together with the associated 1σ relative PDF uncertainties. The upper panels display the luminosities evaluated without any restriction on the final-state rapidity y_X, while the bottom panels instead account for a rapidity cut of $|y_X| < 2.5$ which restricts the produced final state to lie within the ATLAS/CMS central acceptance region.

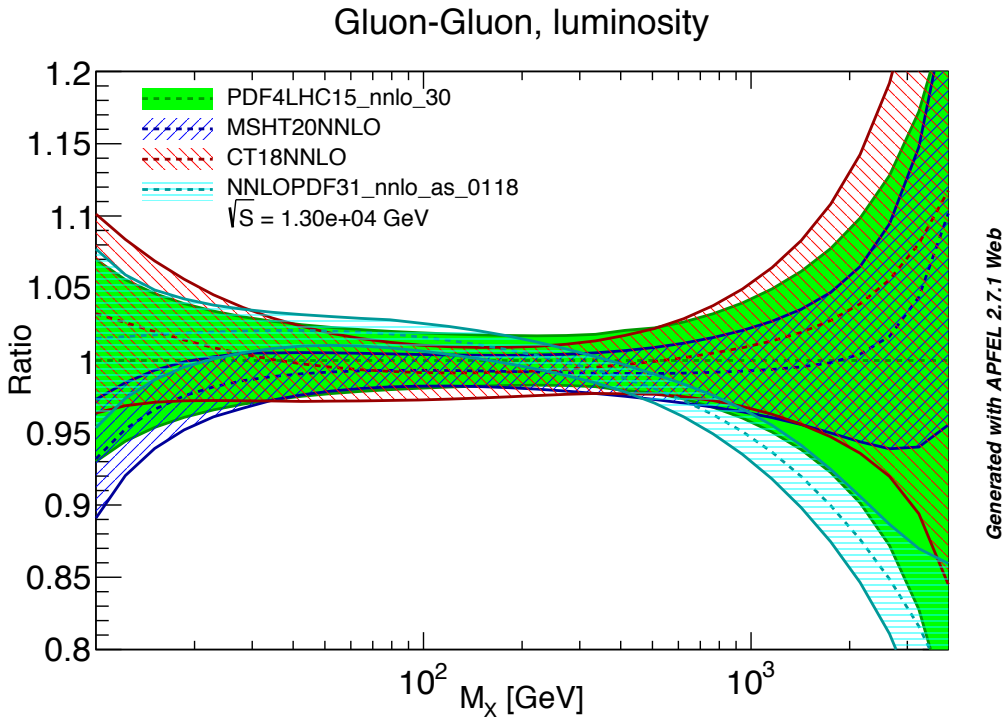
Gluon for PDF4LHC21

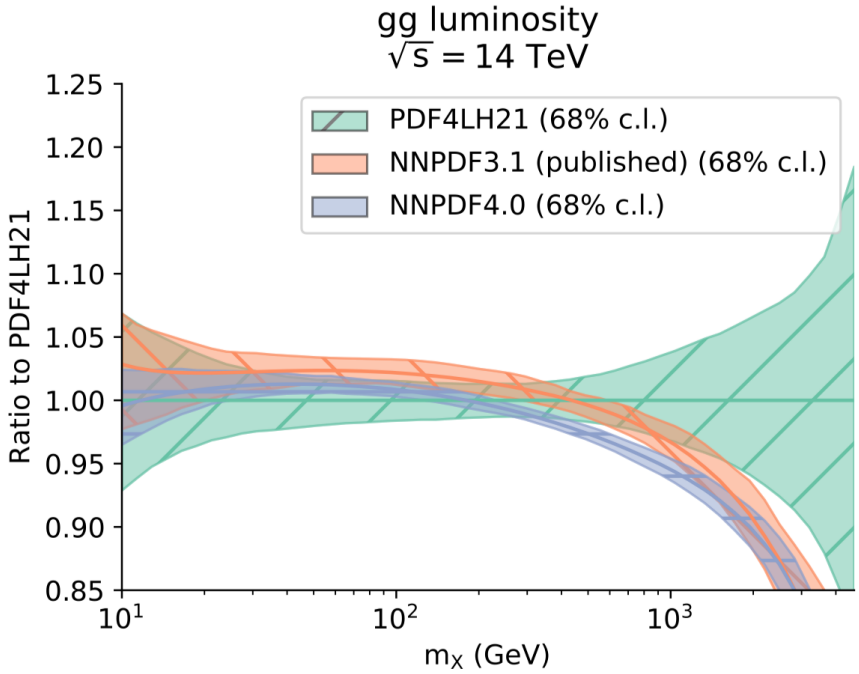




The prime signifies modifications from the original PDF needed for combination; in the CT18' case, use mc=1.4 GeV instead of 1.3 GeV): in the NNPDF3.1' case, several major new datasets added (which came after the publication of NNPDF3.1) -> "halfway to NNPDF4.0"

PDF luminosities for full fits

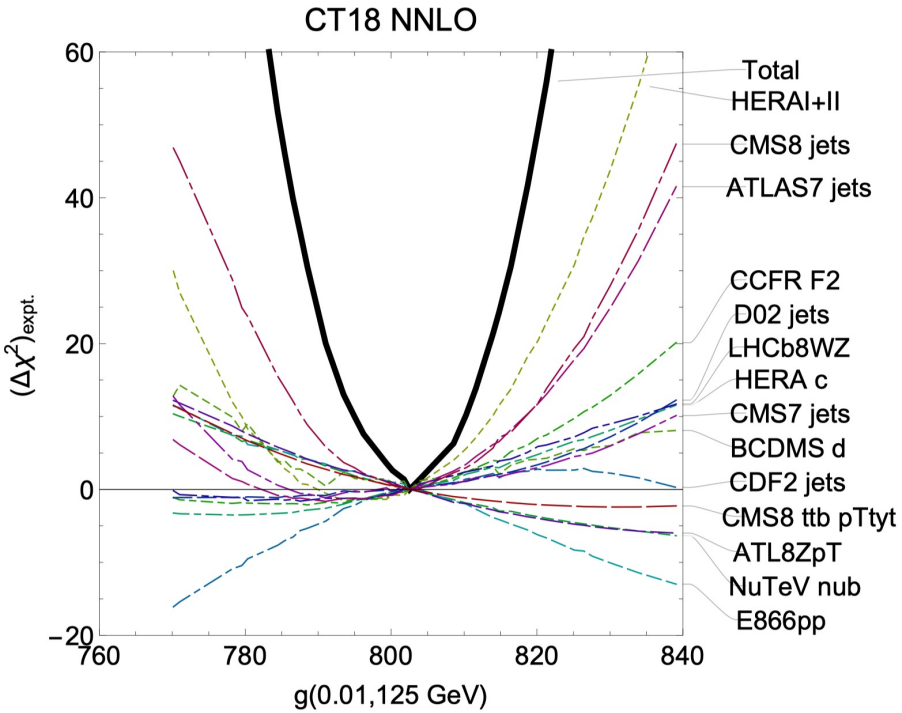




NNPDF4.0 has a larger data set than 3.1, but the crucial data sets are already in 3.1' used for the PDF4LHC21 combination. The NNPDF formalism overemphasizes small data sets (using them both for the training and validation samples), so additional (small) data sets may create artificial PDF uncertainty reductions.

Some useful tools for understanding how PDF fits work

Lagrange Multiplier scans

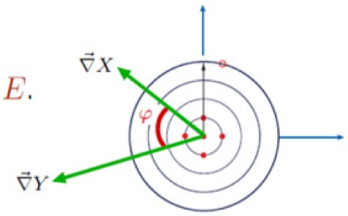


Varied preferences for the gluon distribution from the different data sets, but the net result is reasonably parabolic

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.



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

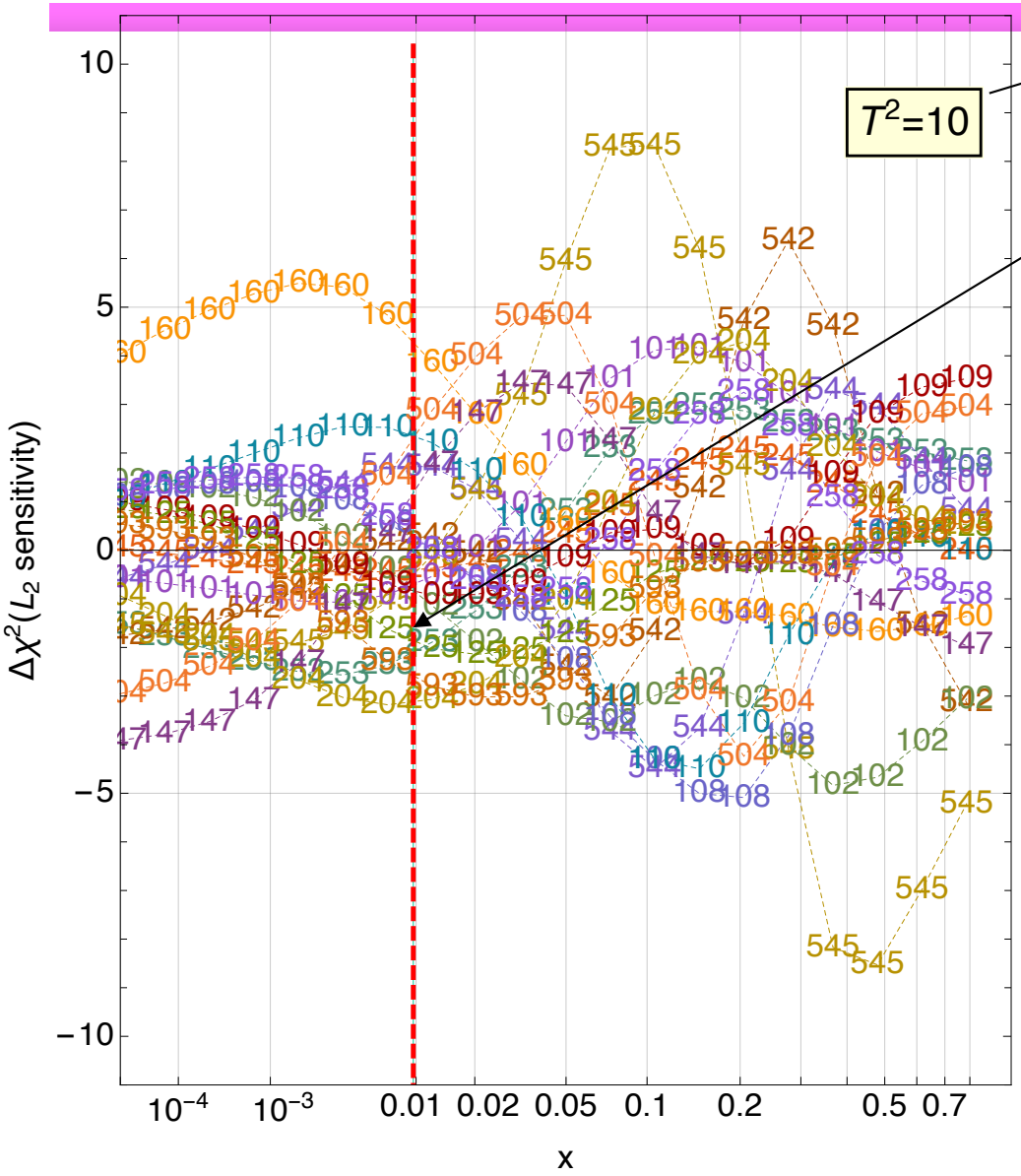
A fast version of the Lagrange Multiplier scan of χ^2_E along the direction of $f_a(x_i, Q_i)$!

Data sets that have a large statistical significance, and a large sensitivity to the variable under investigation, will tend to have a large L2 sensitivity. The sum of all L2 sensitivities in a kinematic region should be close to zero (sum of all of the pulls).

CT18 NNLO
g(x, 100 GeV)

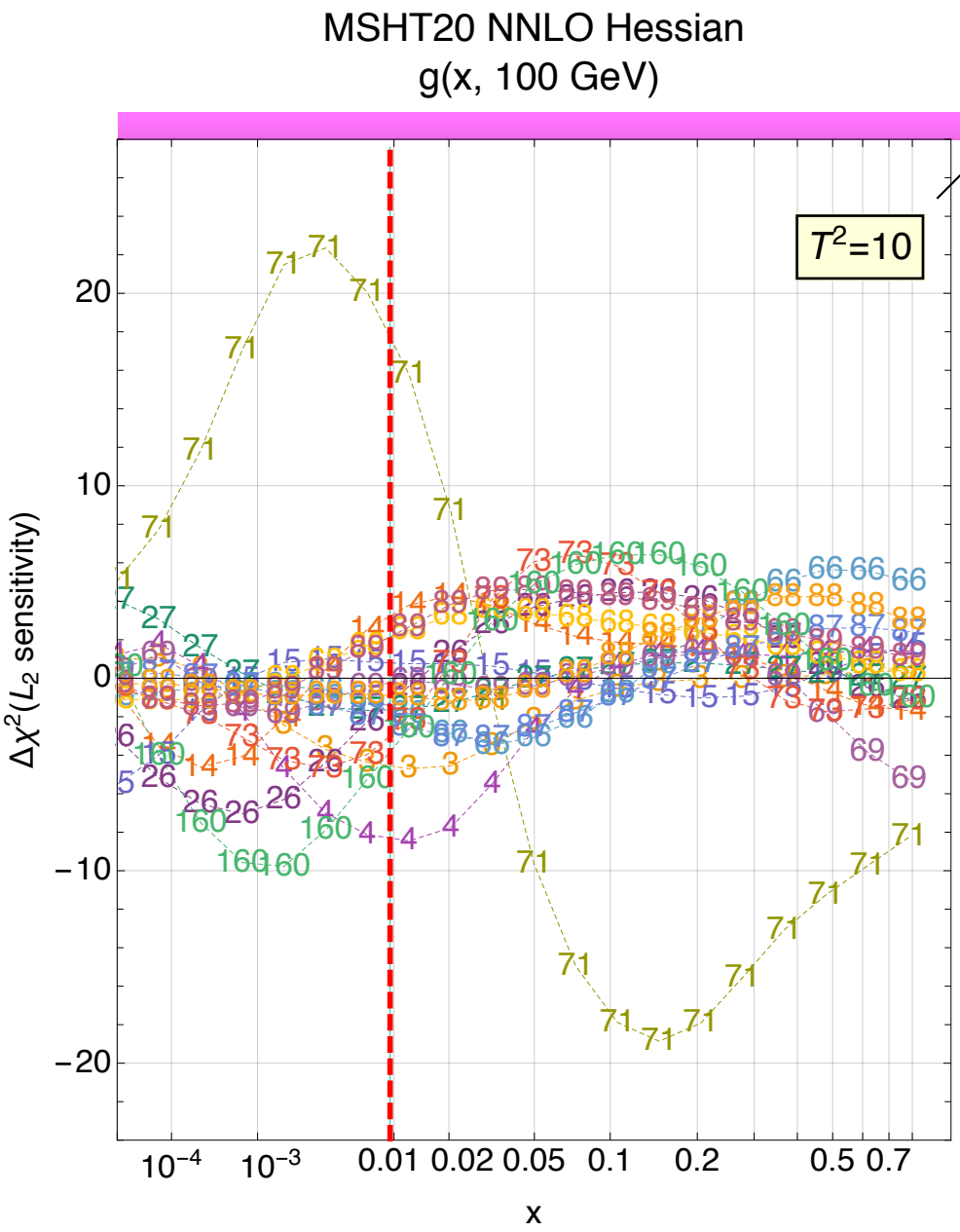
Take the gluon at x~0.01

evaluated at fixed tolerance



245	LHCb7ZWrap
253	ATL8ZpT
542	CMS7jtR7y6
544	ATL7jtR6u
545	CMS 8 TeV jet [incl.]
160	HERA DIS combined
101	BCDMS F2p
102	BCDMS F2d
108	cdhswf2
109	cdhswf3
110	ccfrf2.mi
125	NuTvNbChXN
147	Hn1X0c
204	e866ppxf
504	cdf2jtCor2
593	NuTeV Combined
258	LHCb W/Z 7 and 8 TeV

HERA DIS wants to pull the gluon up, a number of other experiments want to pull it down



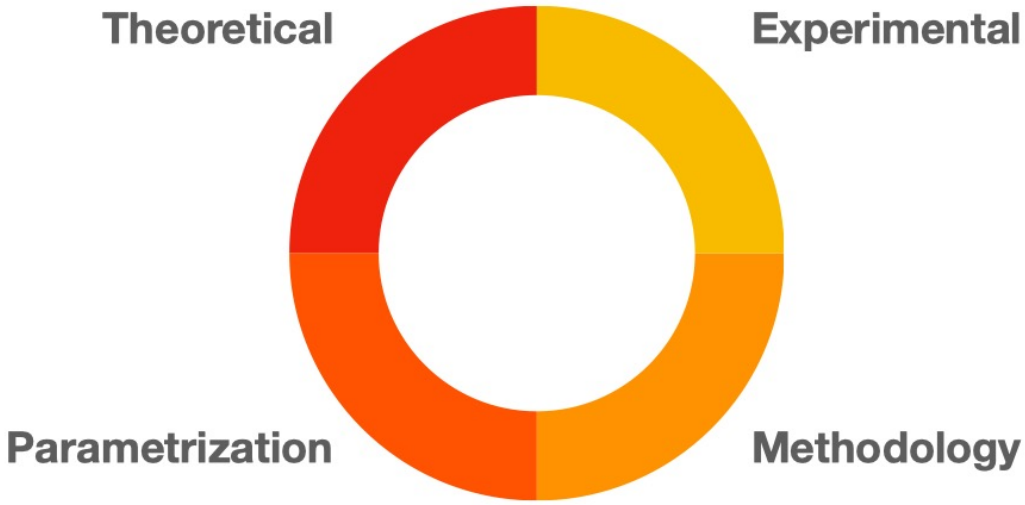
evaluated at fixed tolerance

3	NMC p	69	CMS 7TeV jets
4	NMC d	71	ATLAS 8TeV Z pt
14	F2charm	73	CMS 8TeV jets
15	NMC/BCDMS/SLAC/HERA FL	87	CMS 2.76TeV jets
26	HERA e+ p NC 920 GeV	88	CMS 8TeV single differential ttbar
27	HERA e− p NC 575 GeV	89	ATLAS 8TeV double diff. Z
66	ATLAS 7TeV jets	160	HERA DIS Combined
68	ATLAS 7TeV high precision WZ		

ATLAS 8 TeV Z p_T data tries to pull gluon down at x~0.01; sum of some of the other experiments is trying to pull it up; why is Z p_T so important to the gluon? worth investigating further

Hopscotch:arXiV:2205.10444

- Contributions to PDF uncertainties include
  - experimental errors of the data
  - parametrization uncertainties (CT18 uncertainty incorporates effect of trying out hundreds of parametrization forms)
  - theoretical uncertainties/limitations
  - methodology, including sampling accuracy for Monte Carlo fitting
    - the sampling accuracy has typically been ignored
    - ->hopscotch scans

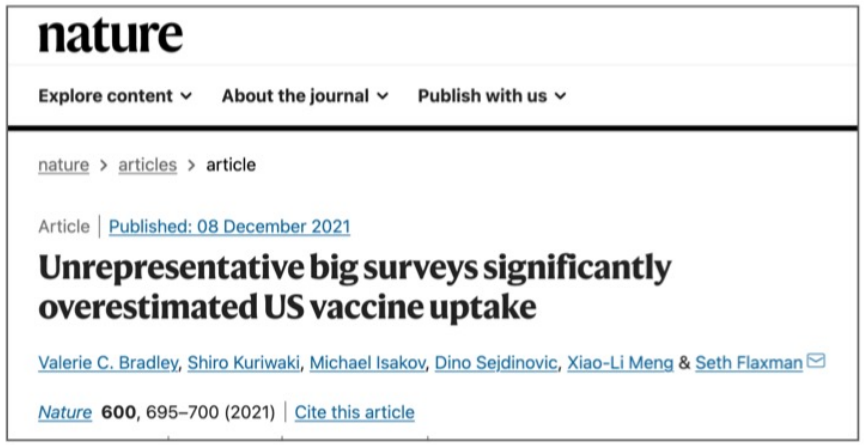


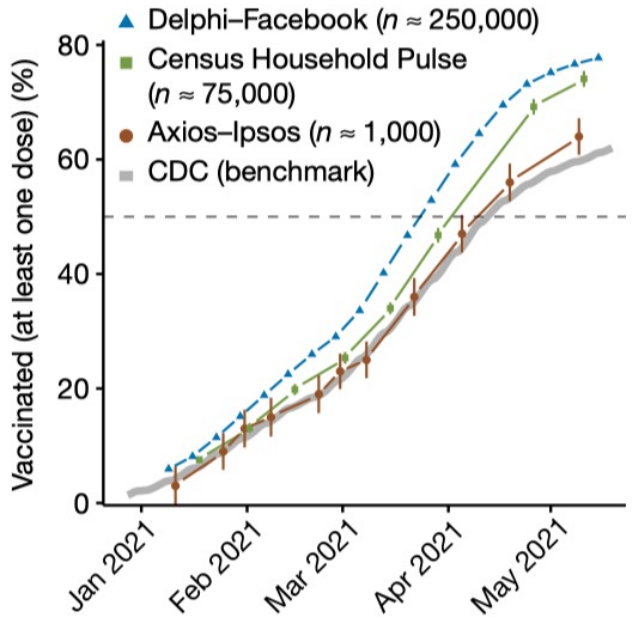
Control of sampling biases in the determination of PDFs can play a critical role

Aurore Courtoy: DIS 2022

Origin of sampling biases — experience with large population surveys

Surveys of the COVID-19 vaccination rate with very large samples of responses and small statistical uncertainties *(Delphi-Facebook)* greatly overestimated the actual vaccination rate published by the Center for Disease Control *(CDC)* after some time delay.

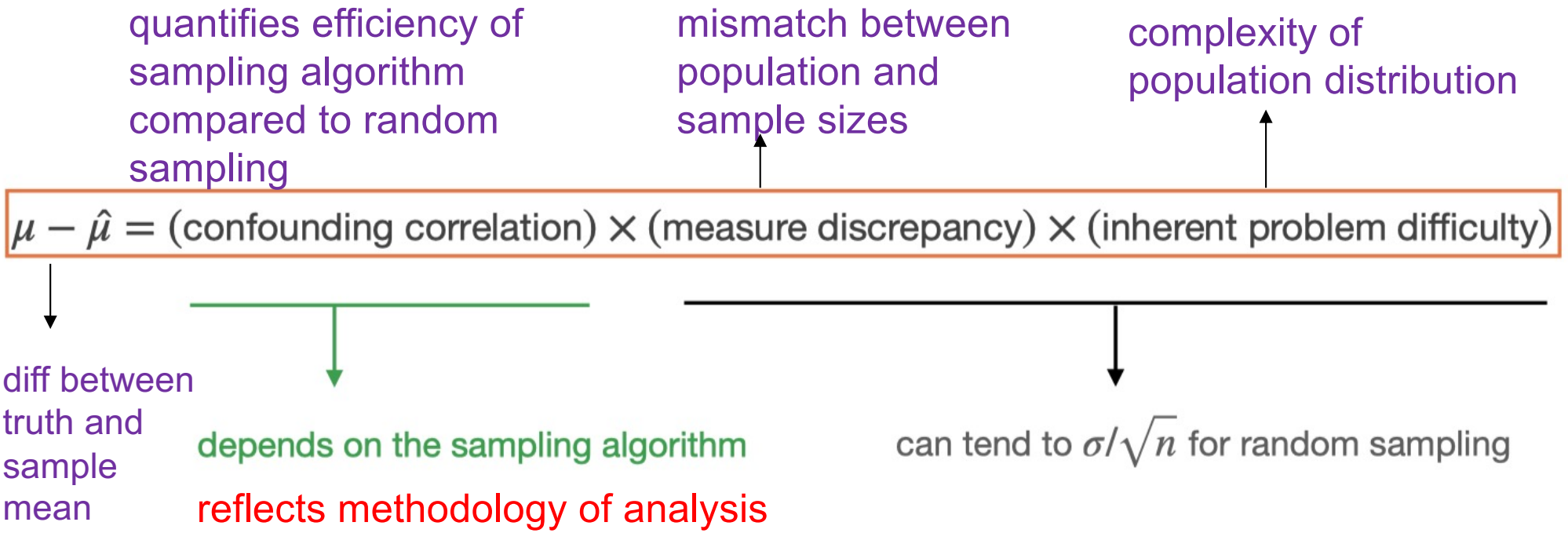




Based on
[Xiao-Li Meng, The Annals of Applied Statistics, Vol. 12 (2018), p. 685]

The deviation has been traced to the **sampling bias.**
In contrast to the statistical error, the sampling bias can involve growth with the size of the sample.

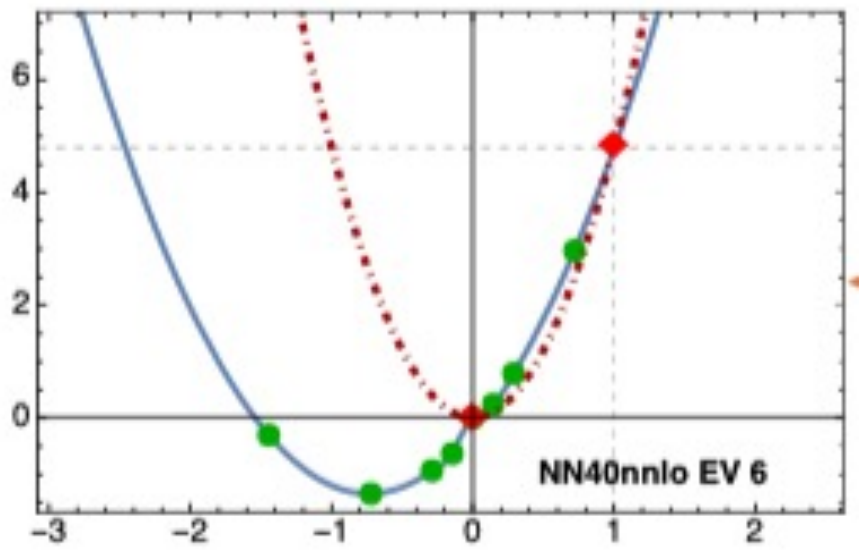
Trio identity (for sample expectation deviation)



- Sample deviation can be large if sampling is not sufficiently random
  - standard error estimates may be too small
- Methodological correlations play a central role in precise PDF analyses, together with data-driven and theory-driven correlations

Eigenvectors

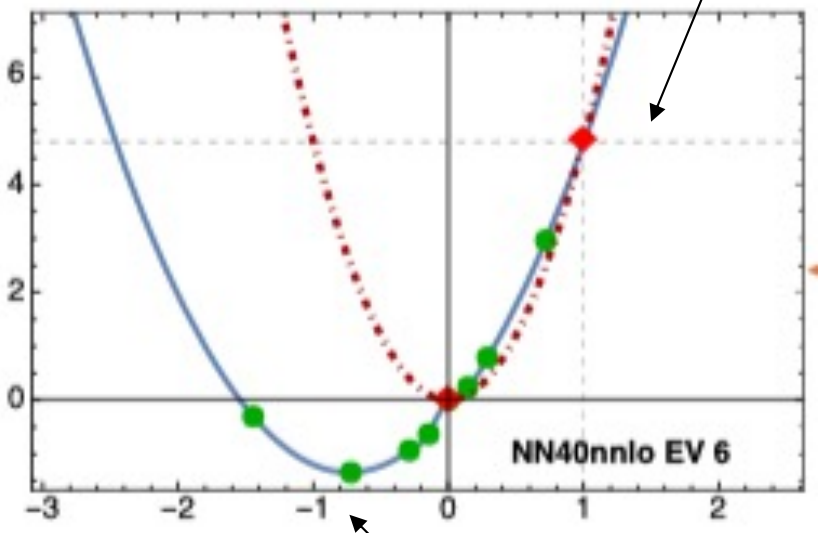
- Sampling of multi-dimensional spaces (d>>20) can be exponentially inefficient and require n > 2^d replicas for reasonable convergence
- A study of this multi-dimensional space for NNPDF is possible due to the public release of the NNPDF4.0 code
- Use published NNPDF4.0 Hessian basis (n=50), converted from MC replicas
  - total χ^2 of each eigenvector set varies, as large as +35 and as low as -25 (wrt replica 0); the majority no larger than 5-10 units in magnitude; only 1 error set per EV
- Can determine χ^2 at green points, where, for some eigenvectors, lower χ^2 solutions evident (displaced from 0)



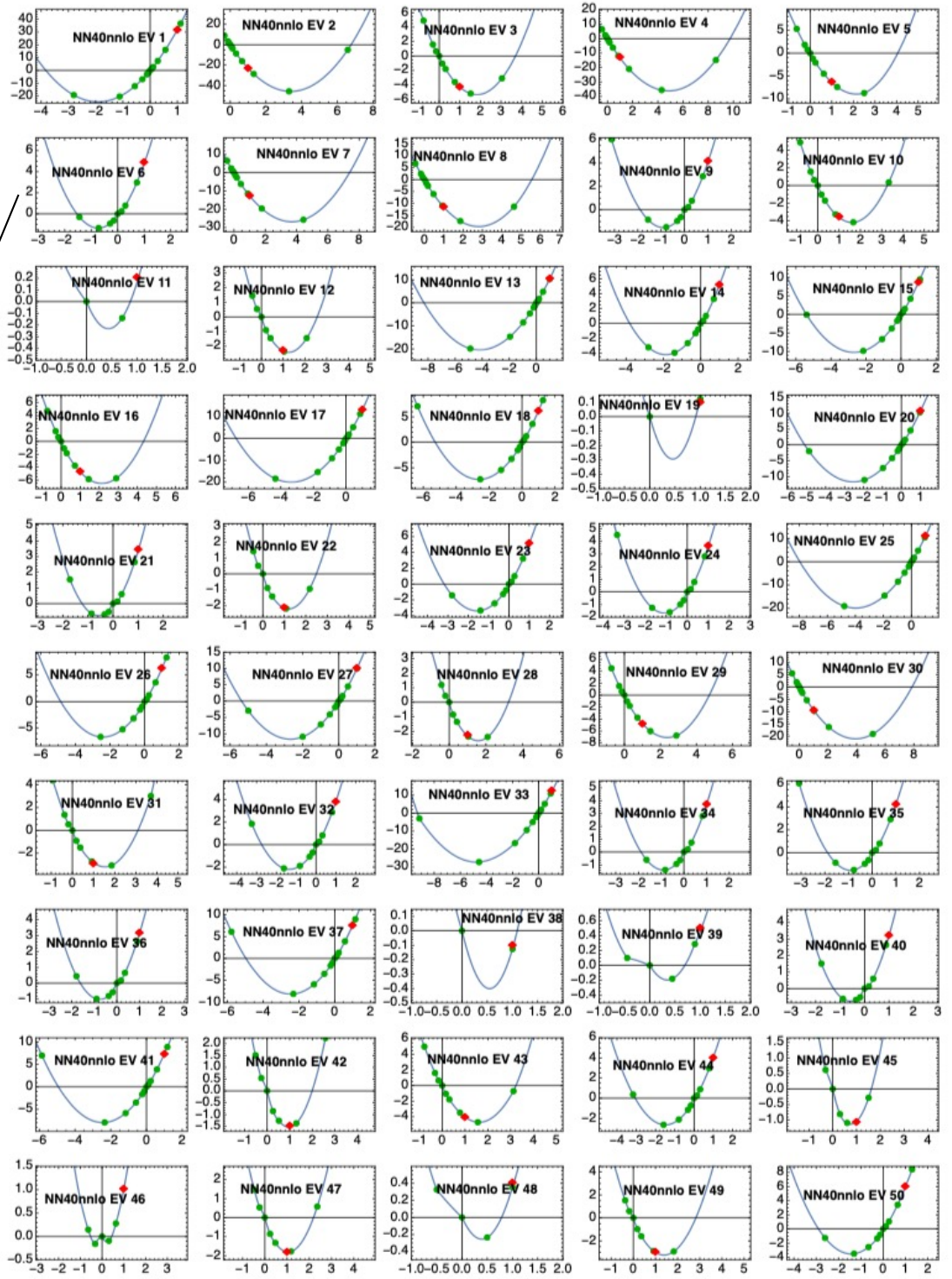
red points correspond to replica 0 and EV6

evaluate χ^2 at 16 points per eigenvector; quadratic behavior observed, i.e. Gaussian uncertainties

- …for a number of eigenvectors, it appears that there are solutions with lower values of χ^2, in some cases substantially lower
- doesn't happen by definition for Hessian-based global fits

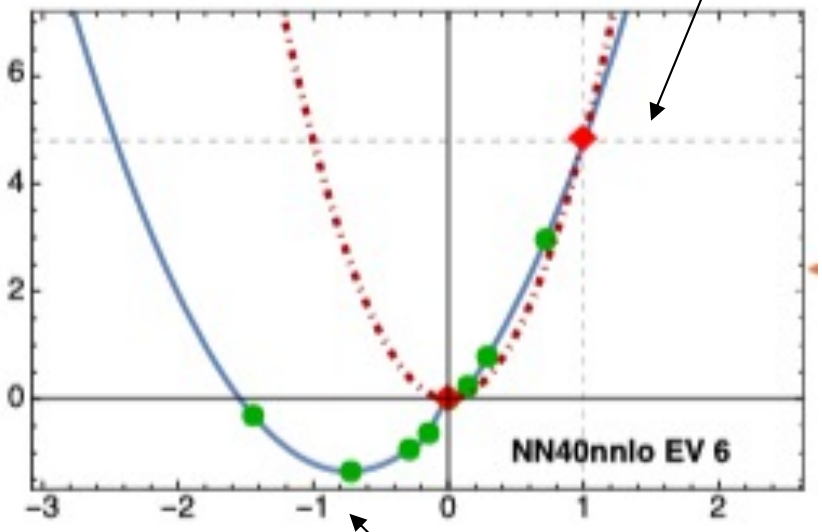


lower χ^2 solution

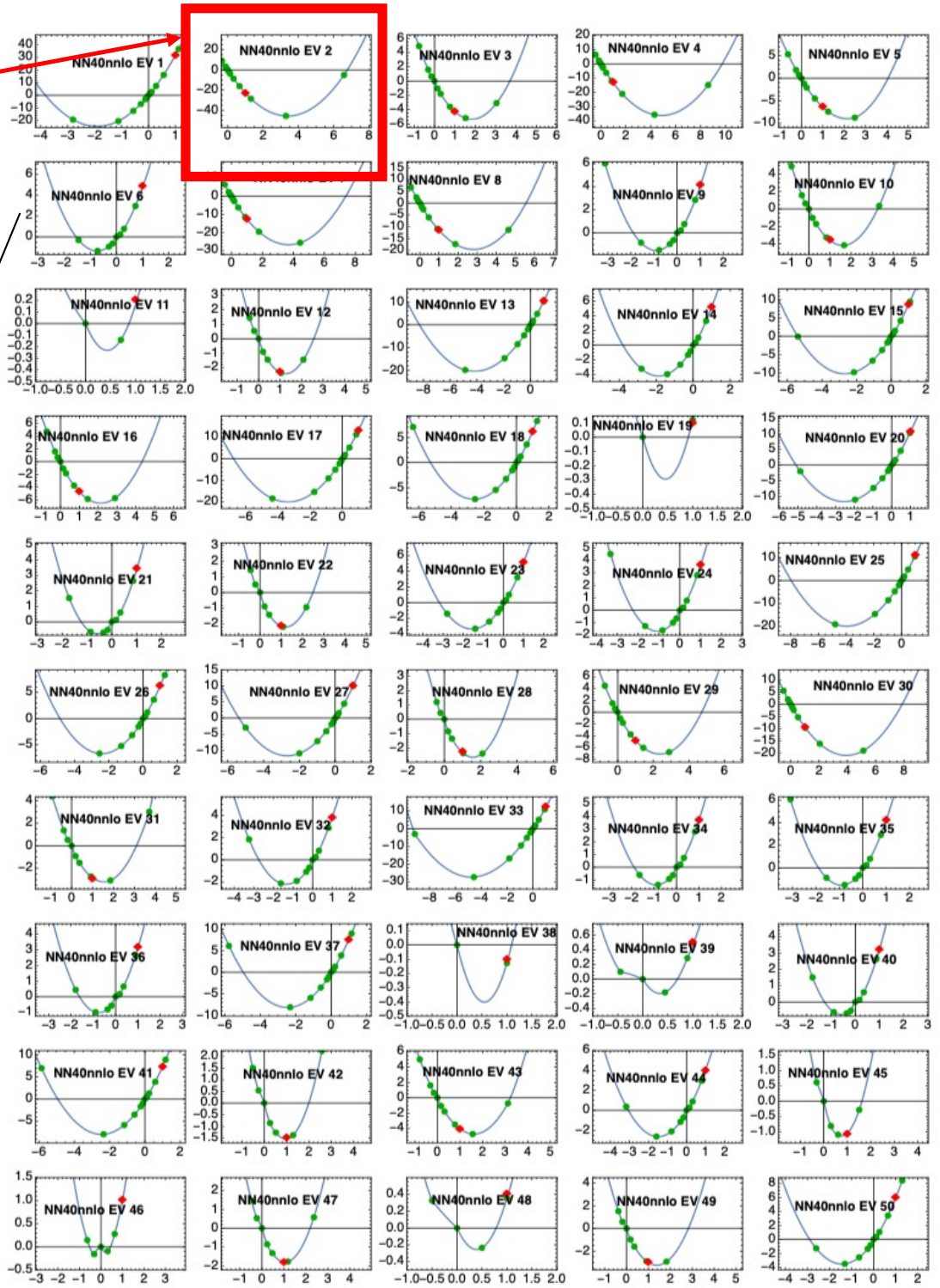


for some eigenvectors, the change in the χ^2 minimum is large

- …for a number of eigenvectors, it appears that there are solutions with lower values of χ^2, in some cases substantially lower
- doesn't happen by definition for Hessian-based global fits

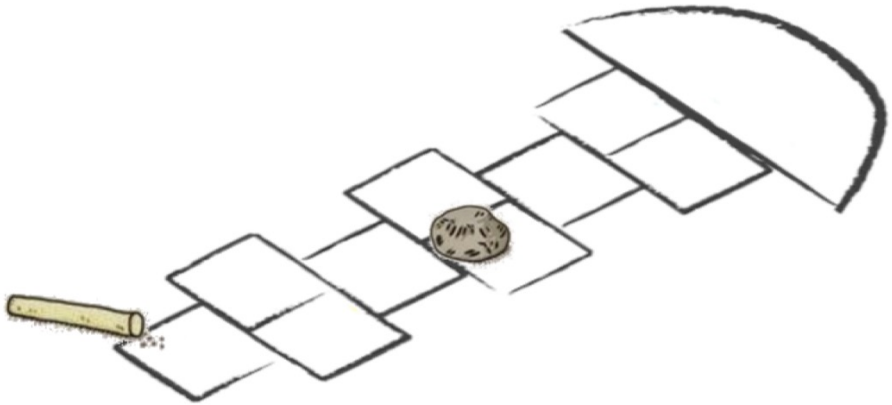


lower χ^2 solution



Hopscotch scans

- Scan along 50 EV directions to identify a hypercube corresponding to $\Delta\chi^2 < T^2$ (T is the tolerance, user-chosen)
- Confirm Gaussian profiles in each eigenvector direction with LM scans

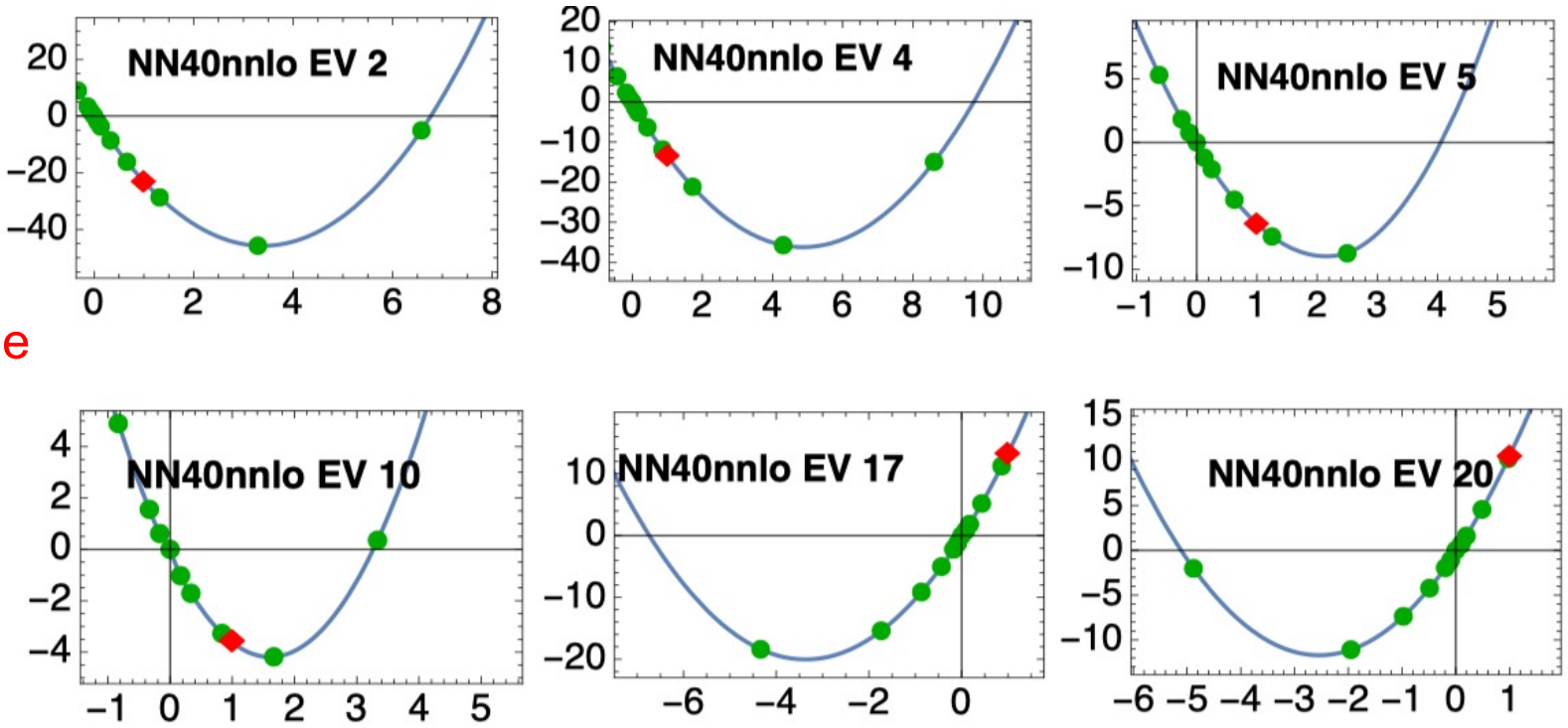


- Concentrate on 4-8 large dimensions in the PDF eigenvector space controlling the large variations of the cross sections under investigation
- Generate replicas varying primarily in these directions; this is not a search for the true global minimum

finding the displaced global minimum in the whole 50-dimensional space is computationally expensive; replica generation is a stochastic exploration; the minimum lies within error ellipses (see later)

Scans of LHC cross sections

- Identify limited number (4-7) of eigenvector directions that give the largest displacements for a given $\Delta\chi^2$ per pair of cross sections
- …for example for σ_Z vs σ_H
- note that both the values of the widths and the minima vary



call these the pole sets

- Generate 300 replicas along the above large eigenvector directions; sort the replicas into $\Delta\chi^2$ wrt NNPDF4.0 replica 0

Construct polygons from pole set eigenvectors

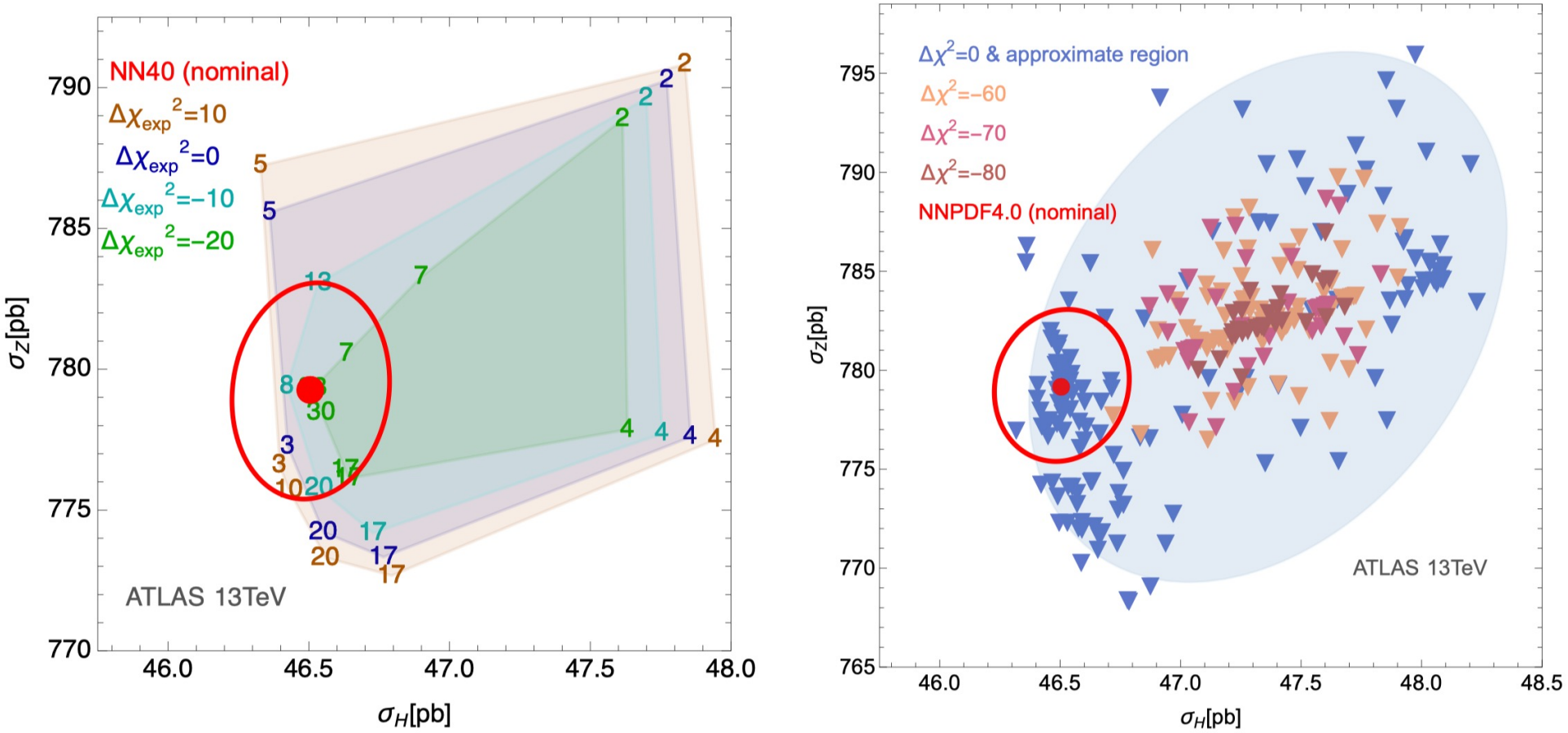
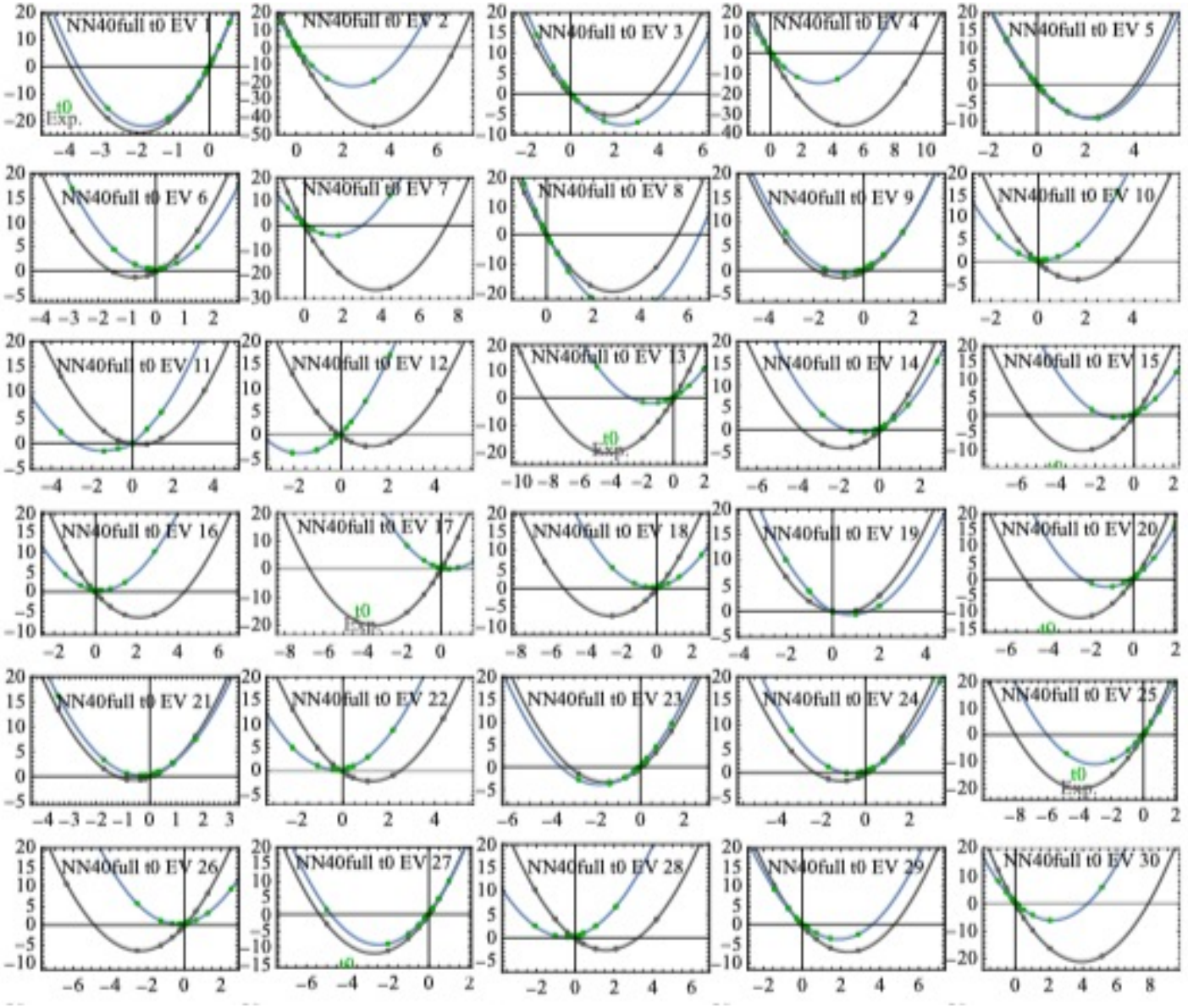


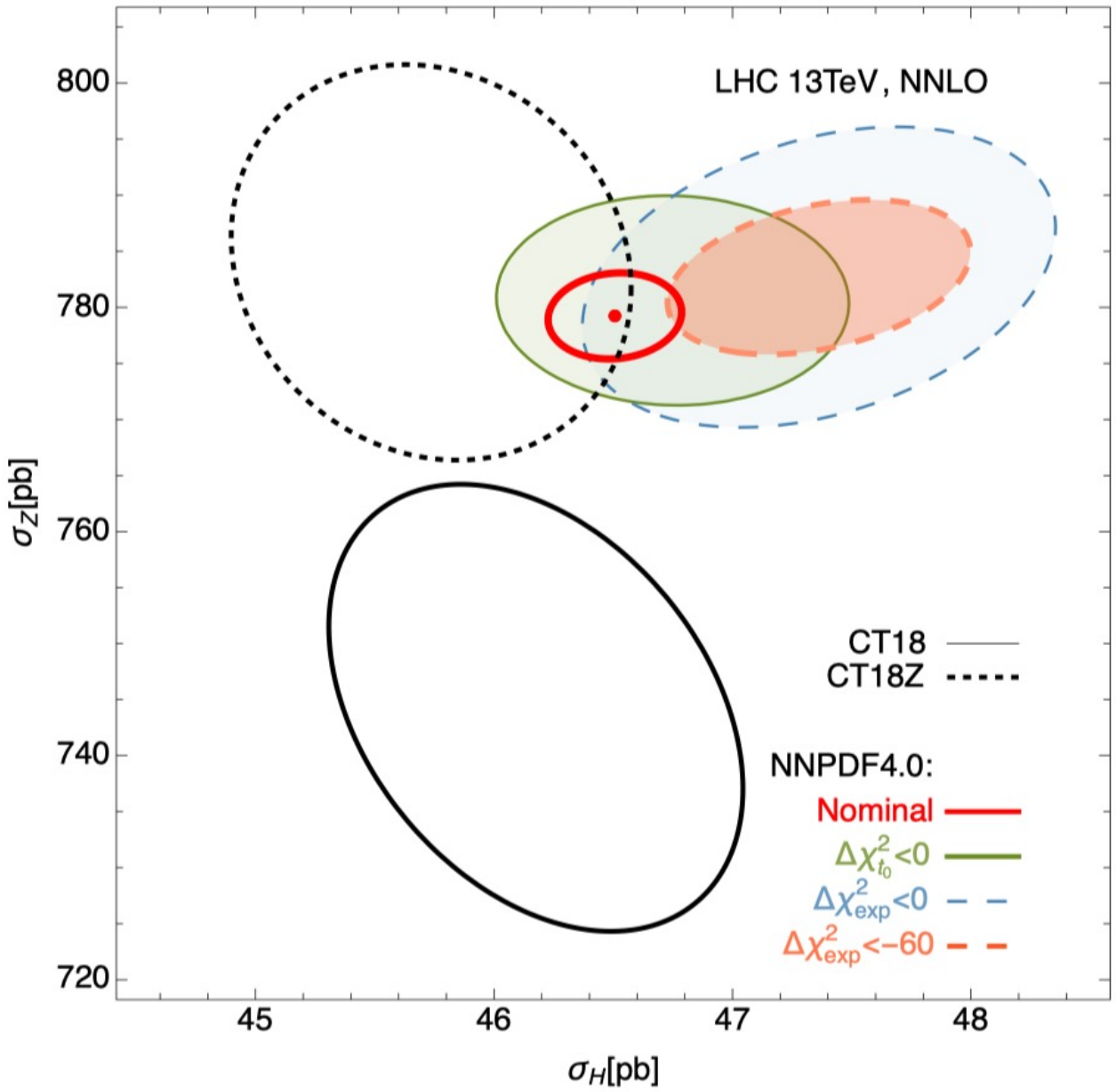
FIG. 5. Intermediate hopscotch scan results for Z vs. $t\bar{t}$ cross sections (upper row) and Z vs. Higgs boson cross sections (lower row) for ATLAS at 13 TeV. See the Appendix for details of the computation. The left panels shows polygons formed by the pole sets with $\Delta\chi^2 = +10$, 0, -10, and -20. In the right panels, the blue triangles correspond to $\Delta\chi^2 = 0$, with replicas with lower $\Delta\chi^2$ shown in increasing hue. Blue ellipses are approximate regions fitted to the $\Delta\chi^2 = 0$ boundary points. Red ellipses correspond to the 68% probability regions from the published NNPDF4.0 Hessian set.

NB: experimental error definitions

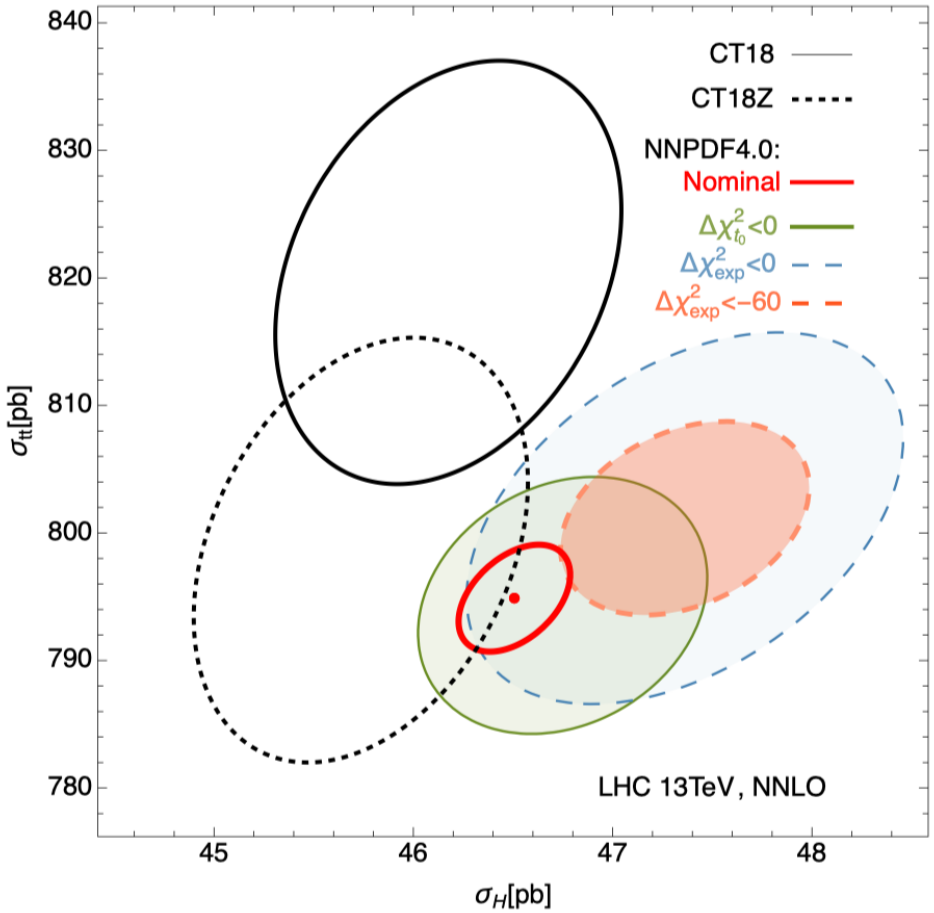
- Different definitions for χ^2 form can affect the PDF uncertainty, i.e. t_o vs experimental
- NNPDF4.0 uses t_o; shifts somewhat smaller than with experimental prescription (used by NNPDF for tabulated χ^2 values), but conclusions still stand

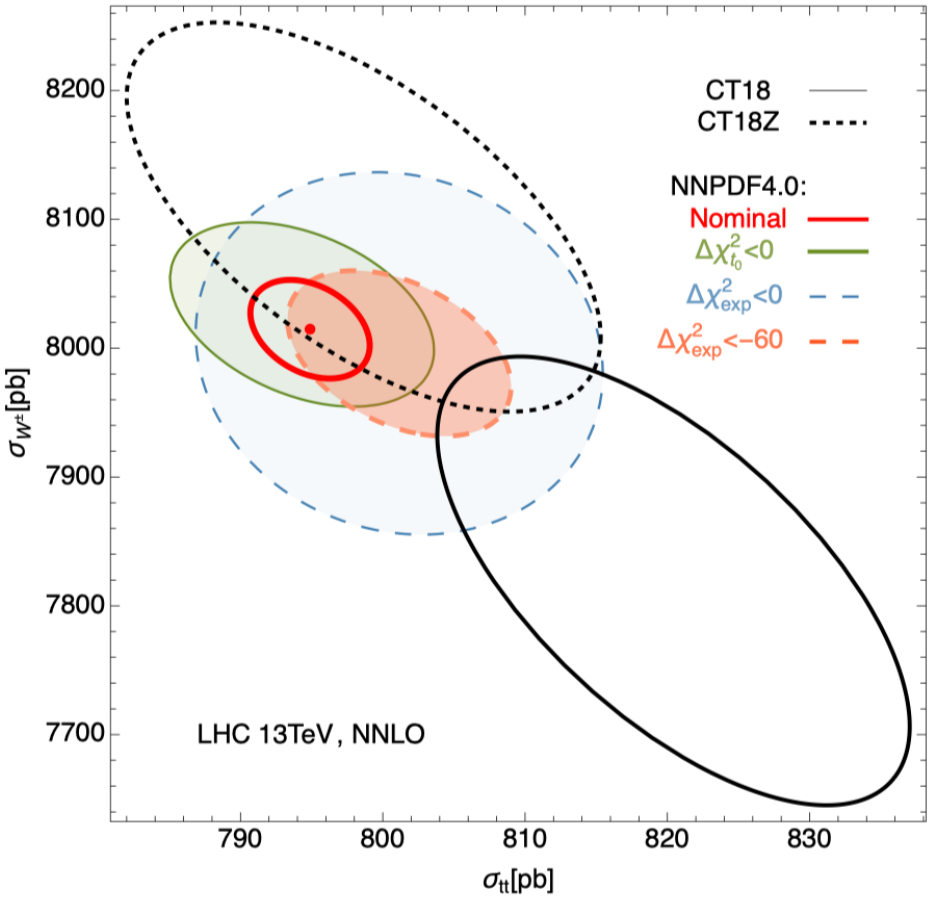


Approximate error ellipses for σ_Z vs σ_H



Two other cross section pairs

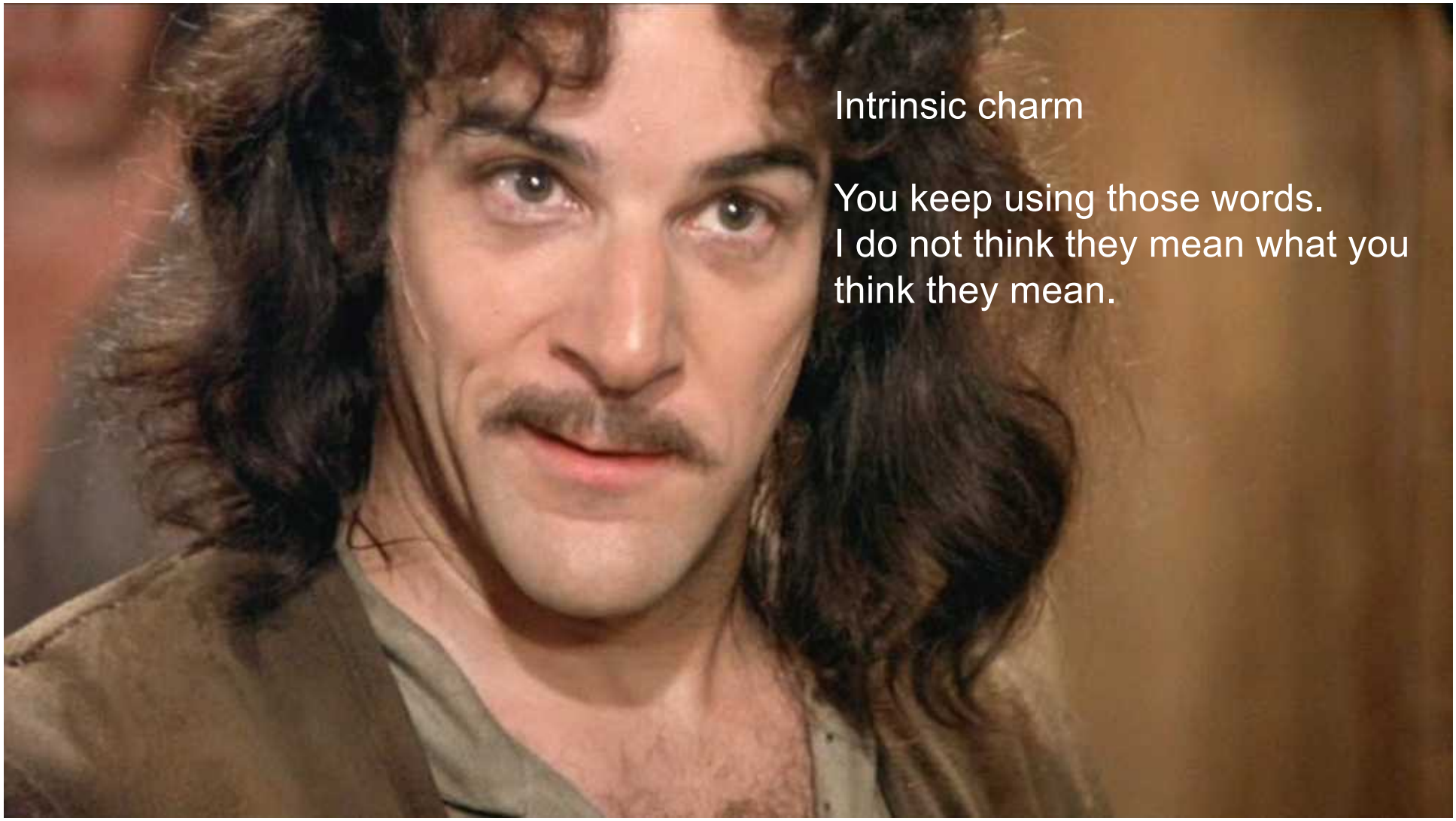




Summary

- Determination of central PDF values and of uncertainties has come a long way
- There is still work to do to provide a more rigorous understanding, especially of the different techniques used to determine the central values and the errors on PDFs
- Paradoxically, increasing the data sample and the parametric space may increase the sample expectation deviation
- In this study, we have shown that the NNPDF4.0 fitting code allows alternate solutions that appear to satisfy NNPDF requirements with similar or lower χ^2
- Trio identity equation may help to design a procedure that reduces any bias in the PDF determination
- Grids used in this study will be made available on LHAPDF, and additional plots at https://ct.hepforge.org/PDFs/2022hopscotch/

Afterward


Intrinsic charm
You keep using those words.
I do not think they mean what you think they mean.

Extras

CT18

Experimental data set E		$N_{\rm pt}$	$\chi^2/N_{\rm pt}$	S
LHCb 7 TeV 1.0 fb^{-1} W/Z forward rapidity	[61]	33	1.63 (1.21)	2.3 (0.9)
LHCb 8 TeV 2.0 fb^{-1} $Z \to e^-e^+$ forward rapidity	[62]	17	1.04 (1.06)	0.2 (0.3)
ATLAS 7 TeV 4.6 fb^{-1}, W/Z combined‡	[63]	34	8.45 (2.61)	16 (5.1)
CMS 8 TeV 18.8 fb^{-1} muon charge asymmetry A_{ch}	[64]	11	1.04 (1.10)	0.2 (0.3)
LHCb 8 TeV 2.0 fb^{-1} W/Z cross sec.	[65]	34	2.17 (1.75)	4.0 (2.7)
ATLAS 8 TeV 20.3 fb^{-1}, Z p_T cross sec.	[66]	27	1.12 (1.05)	0.5 (0.2)
CMS 7 TeV 5 fb^{-1}, single incl. jets, $R = 0.7$	[67]	158	1.23 (1.19)	2.0 (1.7)
ATLAS 7 TeV 4.5 fb^{-1}, single incl. jets, $R = 0.6$	[68]	140	1.45 (1.45)	3.4 (3.4)
CMS 8 TeV 19.7 fb^{-1}, single incl. jets, $R = 0.7$, (extended)	[69]	185	1.14 (1.12)	1.3 (1.2)
CMS 8 TeV 19.7 fb^{-1}, $t\bar{t}$ norm. double-diff. top p_T and y	[70]	16	1.18 (1.19)	0.6 (0.6)
ATLAS 8 TeV 20.3 fb^{-1}, $t\bar{t}$ p_T^t and $m_{t\bar{t}}$ abs. spectrum	[71]	15	0.63 (0.71)	-1.1 (-0.8)

Table 2.1. Numbers of points, χ^2/N_{pt}, and the effective Gaussian variables for the newly added LHC measurements in the CT18 and CT18Z NNLO fits. The ATLAS 7 TeV W/Z data (4.6 fb^{-1}), labelled by ‡, are included in the CT18A and CT18Z global fits, but not in CT18 and CT18X.

Spartyness, a variable that describes the goodness of fit, taking into account the number data points; expect S to be in the range of -1 to 1.

If S>>1, that means the data is poorly fit; if S<<1, that means the fit is too good, and possibly the errors are overestimated

$$S_E = \sqrt{2\chi^2_E} - \sqrt{2N_{\rm pt,E} - 1}$$

MSHT20

Experimental data set	N_{pt}	$\chi^2/N_{\rm pt}$	S
D0 W asymmetry [106]	14	0.86	-0.3
$\sigma_{t\bar{t}}$ Tevatron +CMS+ATLAS 7, 8 TeV [107]- [108]	17	0.85	-0.4
LHCb 7+8 TeV $W + Z$ [61, 62]	67	1.48	2.6
LHCb 8 TeV e [65]	17	1.54	1.5
CMS 8 TeV W [64]	22	0.58	-1.5
ATLAS 7 TeV jets $R = 0.6$ [68]	140	1.59	4.4
CMS 7 TeV $W + c$ [102]	10	0.86	-0.2
ATLAS 7 TeV W, Z [63]	61	1.91	4.3
CMS 7 TeV jets $R = 0.7$ [67]	158	1.11	1.0
ATLAS 8 TeV Zp_T [66]	104	1.81	5.0
CMS 8 TeV jets [69]	174	1.50	4.2
ATLAS 8 TeV $t\bar{t} \to l + $j single-diff [71]	25	1.02	0.1
ATLAS 8 TeV $t\bar{t} \to l^+l^-$ single-diff [109]	5	0.68	-0.4
ATLAS 8 TeV high-mass Drell-Yan [110]	48	1.18	0.9
ATLAS 8 TeV $W^{+,-}$ + jet [111]	32	0.60	-1.7
CMS 8 TeV $(d\sigma_{t\bar{t}}/dp_{T,t}dy_t)/\sigma_{t\bar{t}}$ [70]	15	1.50	1.3
ATLAS 8 TeV W^+, W^- [100]	22	2.61	4.2
CMS 2.76 TeV jets [112]	81	1.27	1.7
CMS 8 TeV $t\bar{t}$ y_t distribution [113]	9	1.47	1.0
ATLAS 8 TeV double differential Z [99]	59	1.45	2.3

Table 2.2. Numbers of points, fit qualities $\chi^2/N_{\rm pt}$ and S values for new collider data added to the NNLO MSHT20 fit.

Note the trouble fitting the ATLAS W/Z data

Definitions for CT18'/NNPDF3.1'

- CT18->CT18': m_c=1.4 GeV,m_b=4.75 GeV
- NNPDF3.1->NNPDF3.1': same as above plus some additions to the data set (in some ways NNPDF3.1' is a transition from 3.1-> 4.0)
- No MSHT20' since the above are the heavy quark mass values they normally use

	NNPDF3.1 [15]			NNPDF3.1′		
Experimental data set	$N_{\rm pt}$	$\chi^2/N_{\rm pt}$	S	$N_{\rm pt}$	$\chi^2/N_{\rm pt}$	S
D0 W electron asymmetry [121]	8	2.70	+2.70	11	3.07	+3.64
D0 W muon asymmetry [122]	9	1.56	+1.18	9	1.58	+1.21
ATLAS low-mass DY 7 TeV [123]	6	0.90	−0.03	6	0.89	−0.05
ATLAS W,Z 7 TeV [63]	34	2.14	+3.88	61	1.99	+4.58
ATLAS Z p_T 8 TeV (p_T, $m_{\ell\ell}$) [66]	44	0.93	−0.28	44	0.94	−0.23
ATLAS Z p_T 8 TeV (p_T, y_Z) [66]	48	0.94	−0.25	48	0.95	−0.20
ATLAS single-inclusive jets 7 TeV ($R = 0.6$) [68]	31	1.07	+0.33	140	1.25	+2.00
ATLAS $\sigma_{t\bar{t}}^{\rm tot}$ 7, 8, 13 TeV [124,125]	3	0.86	+0.04	3	0.95	+0.15
ATLAS $t\bar{t}$ ℓ+jets 8 TeV ($1/\sigma\ d\sigma/dy_t$) [71]	9	1.45	+0.99	4	3.56	+2.69
CMS W rapidity 8 TeV [64]	22	1.01	+0.11	22	1.03	+0.17
CMS Z p_T 8 TeV [126]	28	1.32	+1.18	28	1.34	+1.25
CMS single-inclusive jets 2.76 TeV [112]	81	1.03	+0.23	—	—	—
CMS single-inclusive jets 8 TeV [69]	—	—	—	185	1.30	+2.72
CMS $\sigma_{t\bar{t}}^{\rm tot}$ 7, 8, 13 TeV [127,128]	3	0.20	−1.14	3	0.18	−1.20
CMS $t\bar{t}$ ℓ+jets 8 TeV ($1/\sigma\ d\sigma/dy_{t\bar{t}}$) [113]	9	0.94	−0.01	9	1.67	+1.36
CMS $t\bar{t}$ 2D 2ℓ 8 TeV ($1/\sigma\ {\rm d}\sigma/{\rm d}y_t dm_{t\bar{t}}$) [70]	—	—	—	16	0.81	−0.48
LHCb $W,Z \to \mu$ 7 TeV [61]	29	1.76	+1.55	29	1.96	+3.11
LHCb $W,Z \to \mu$ 8 TeV [65]	30	1.37	+1.39	30	1.36	+1.35

Note the trouble fitting the ATLAS W/Z data

important addition

Table 2.3. The numbers of points, $\chi^2/N_{\rm pt}$ and S values for new collider data in the NNPDF3.1 fit [15] and in the NNPDF3.1′ fit variant adopted in the present combination. The Tevatron and LHC data sets already included in NNPDF3.0 are kept in NNPDF3.1, but not necessarily in NNPDF3.1′. These are not indicated in the table. Note that, despite the number of LHC data points is larger in NNPDF3.1′ than that in NNPDF3.1, the total number of data points in the two analyses is similar, mainly because the Tevatron single-inclusive jet measurements (not indicated in the table) are no longer included in NNPDF3.1′. See text for details.

Combination

- Generate 300 MC replicas of each of the 3 PDFs and combine

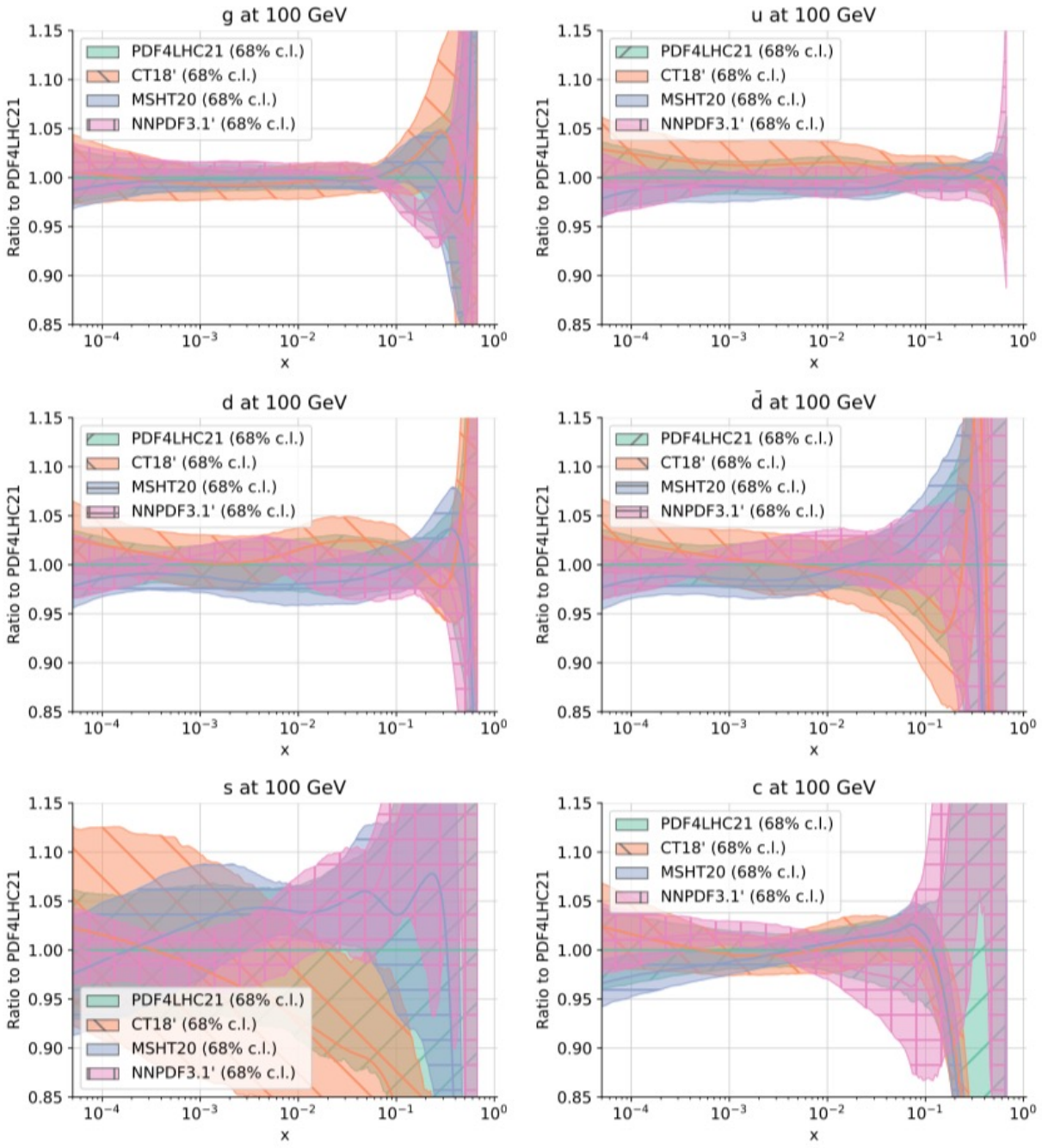


Figure 4.1. Comparison of the PDF4LHC21 combination (composed by $N_{\rm rep} = 900$ replicas) with the three constituent sets at $Q = 100$ GeV, normalised to the central value of the former and with their respective 68%CL uncertainty bands. In the case of the Hessian sets (CT18′ and MSHT20) we display their Monte Carlo representation composed by $N_{\rm rep} = 300$ replicas generated according to Eq. (4.3). The NNPDF3.1′ band is also constituted by $N_{\rm rep} = 300$ (native) replicas.

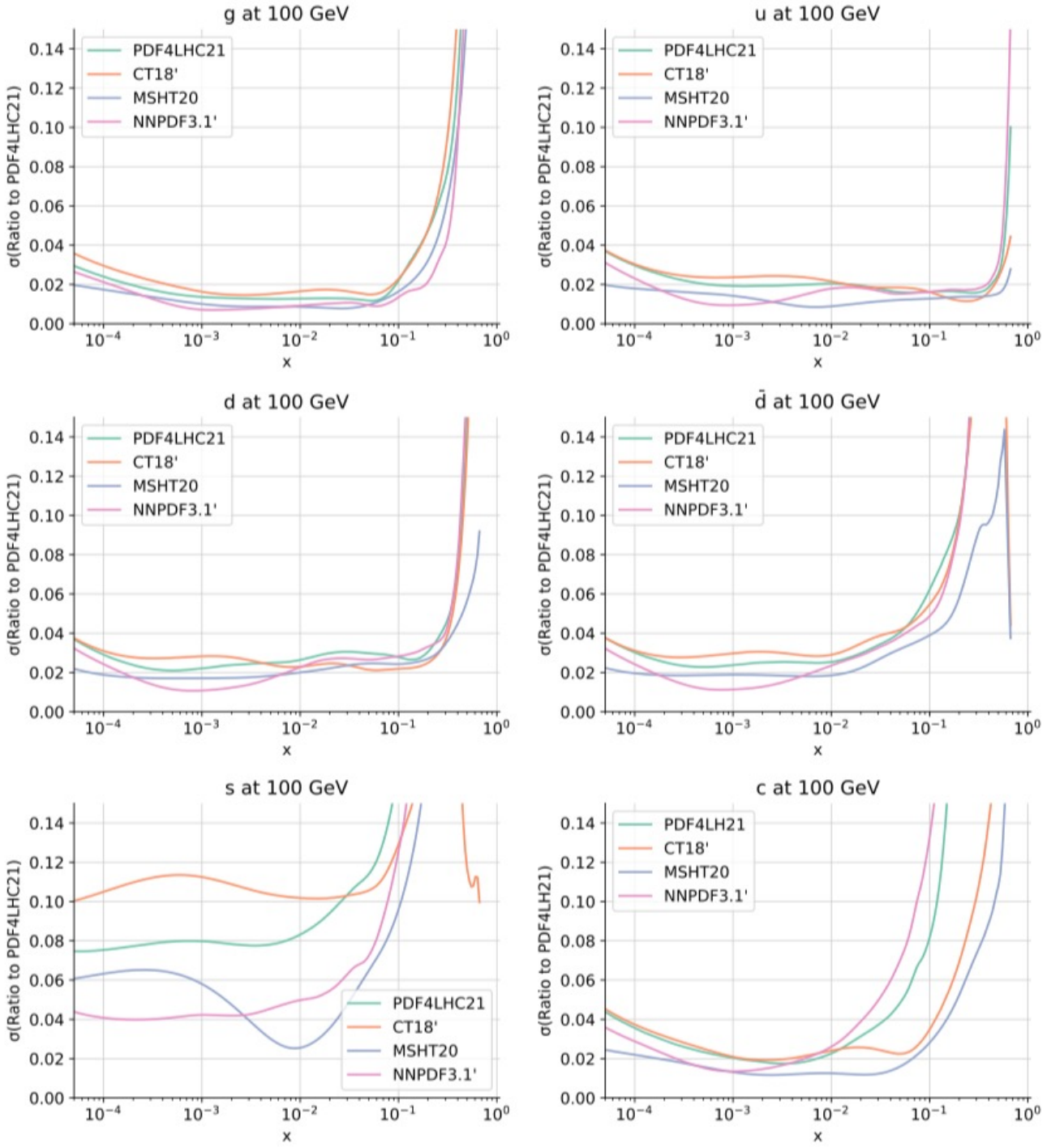


Figure 4.2. Same as Fig. 4.1 now showing the relative PDF 68% CL uncertainties (normalised to the PDF4LHC21 central value) of the four PDF sets.

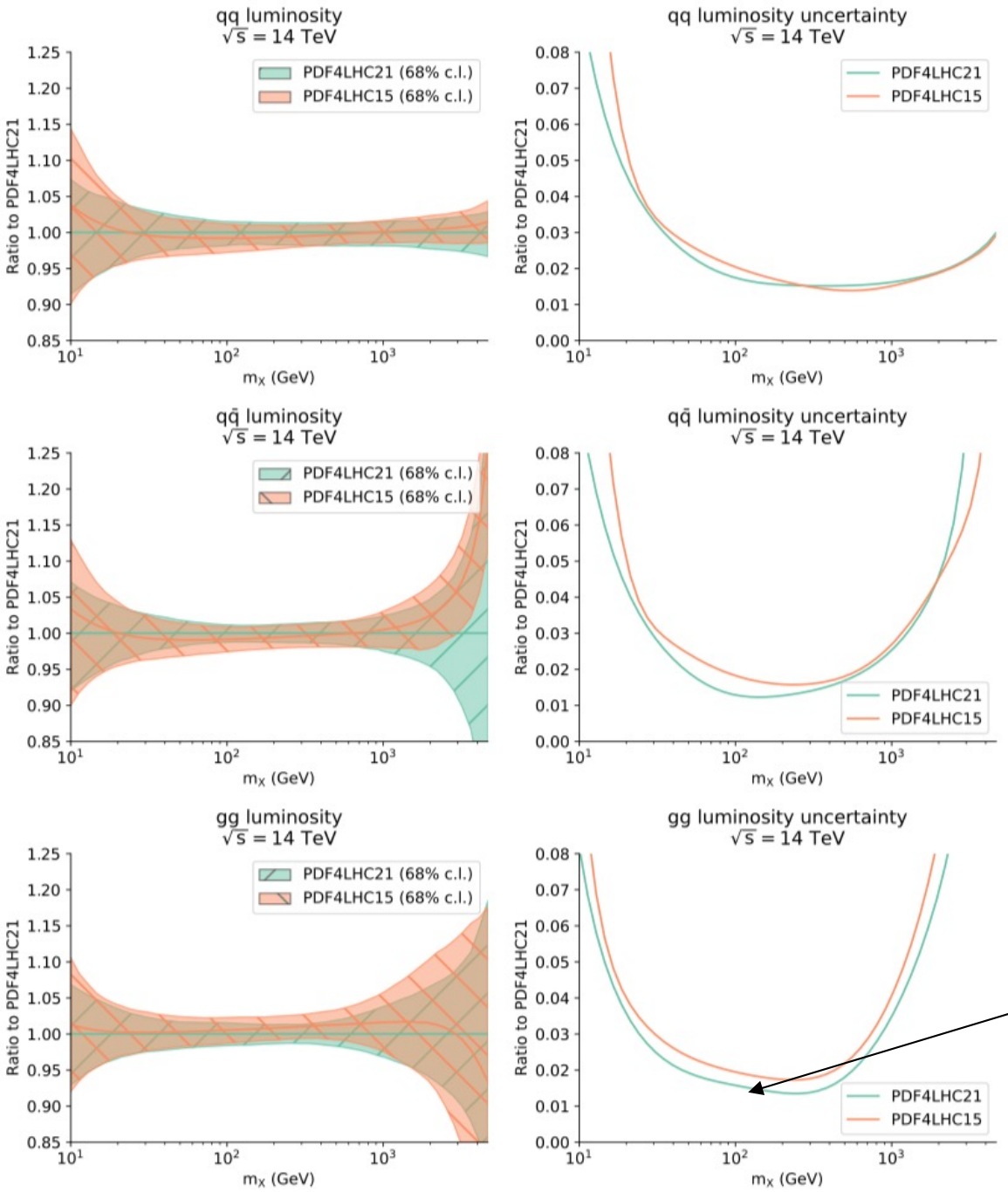


Figure 4.11. Comparison of the partonic luminosities at $\sqrt{s} = 14$ TeV between PDF4LHC15 and PDF4LHC21. In both cases, the original sets with $N_{\rm rep} = 900$ have been used. Results are shown for the quark-quark, quark-antiquark, and gluon-gluon luminosities as a function of the final state invariant mass m_X, and normalised to the central value of the PDF4LHC21 prediction. The right panels display the corresponding 68% CL relative uncertainties.

reduction in uncertainty for gg fusion

- It can be useful to look at 2-D ellipses comparing cross sections

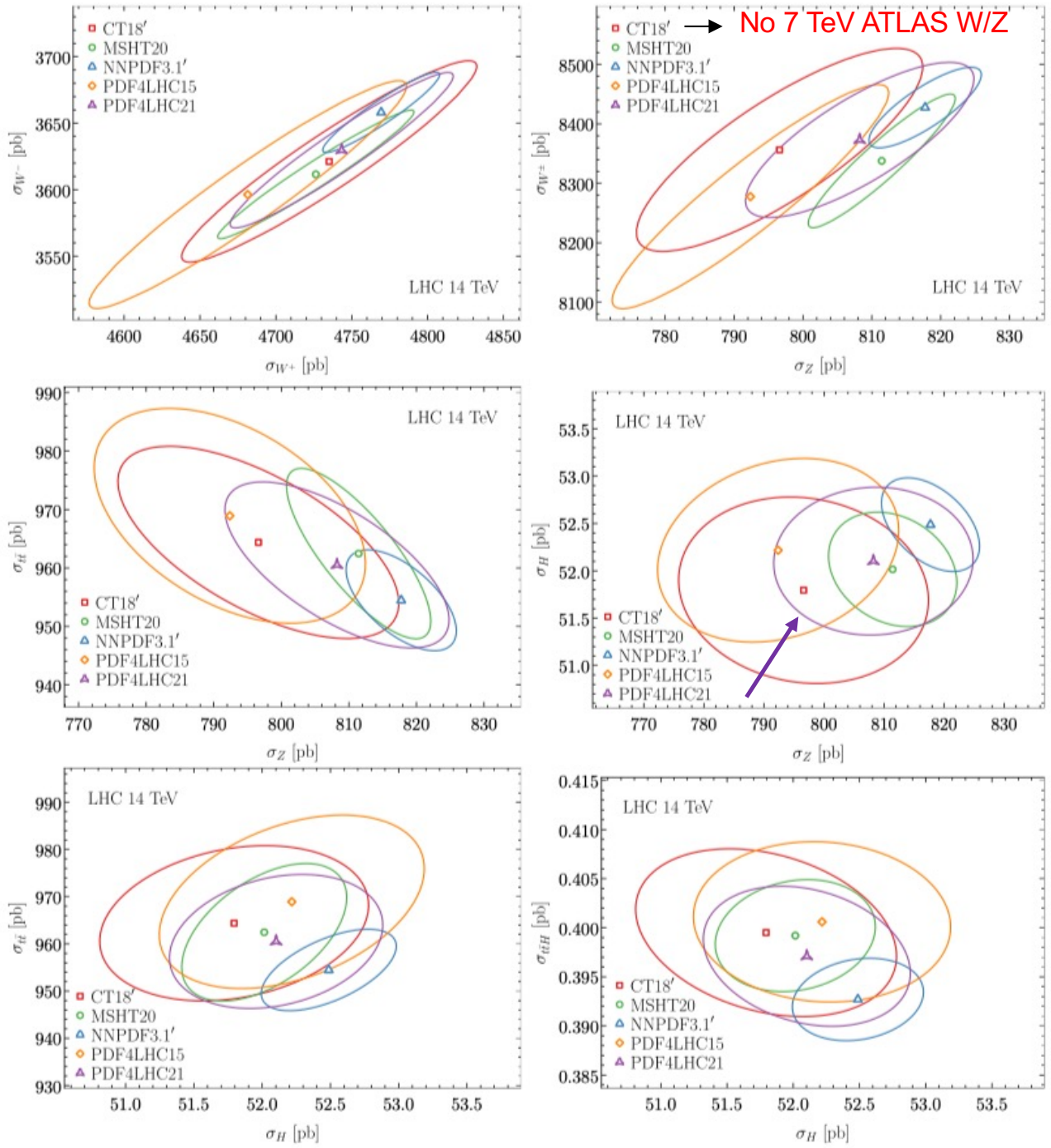


Figure 5.2. The 1σ ellipses for pairs of inclusive cross sections among $W^{\pm}, Z, t\bar{t}, H, t\bar{t}H$ production at the LHC 14 TeV. The $W^{\pm}/Z$ cross sections are defined in the ATLAS 13 TeV fiducial volume [170], while others correspond to the full phase space. See text for details of the theory calculations.