The CT18 QCD analysis with the LHC experimental data

Pavel Nadolsky

Southern Methodist University

CTEQ-TEA (Tung et al.) working group

China Northeastern University: T.-J. Hou

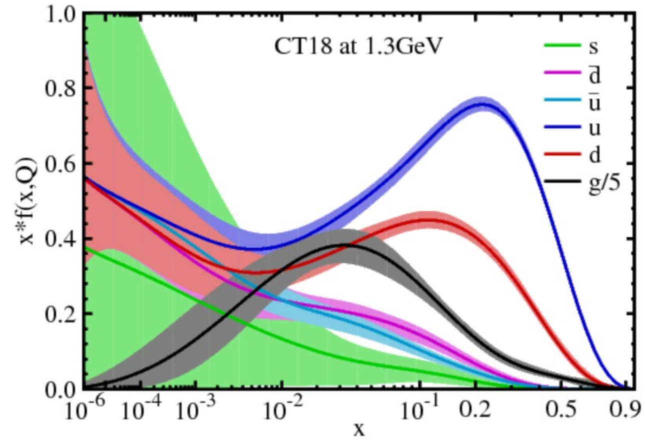
Kennesaw State University: M. Guzzi

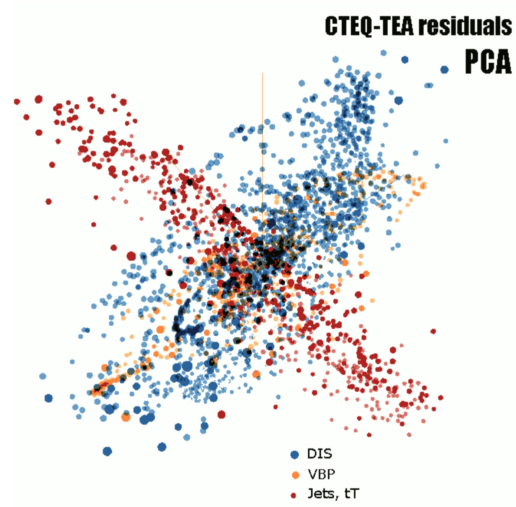
Michigan State U.: J. Huston, J. Pumplin, D. Stump, C. Schmidt, J. Winter, C.-P. Yuan

Shanghai Jiao Tong University: J. Gao

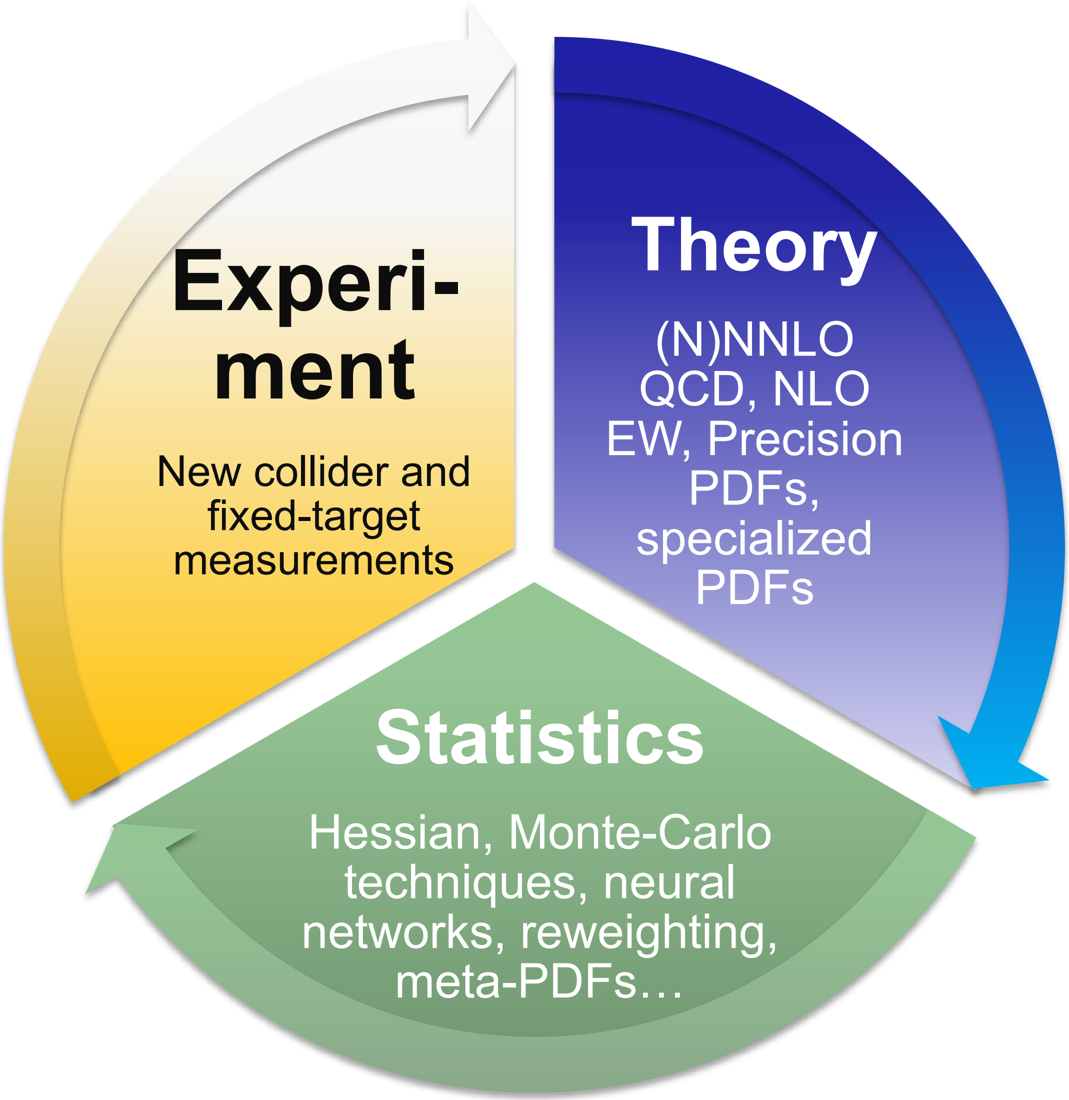
Southern Methodist University: T. Hobbs, P.N., B.T.Wang, K. Xie

Xinjiang University: S. Dulat, I. Sitiwaldi





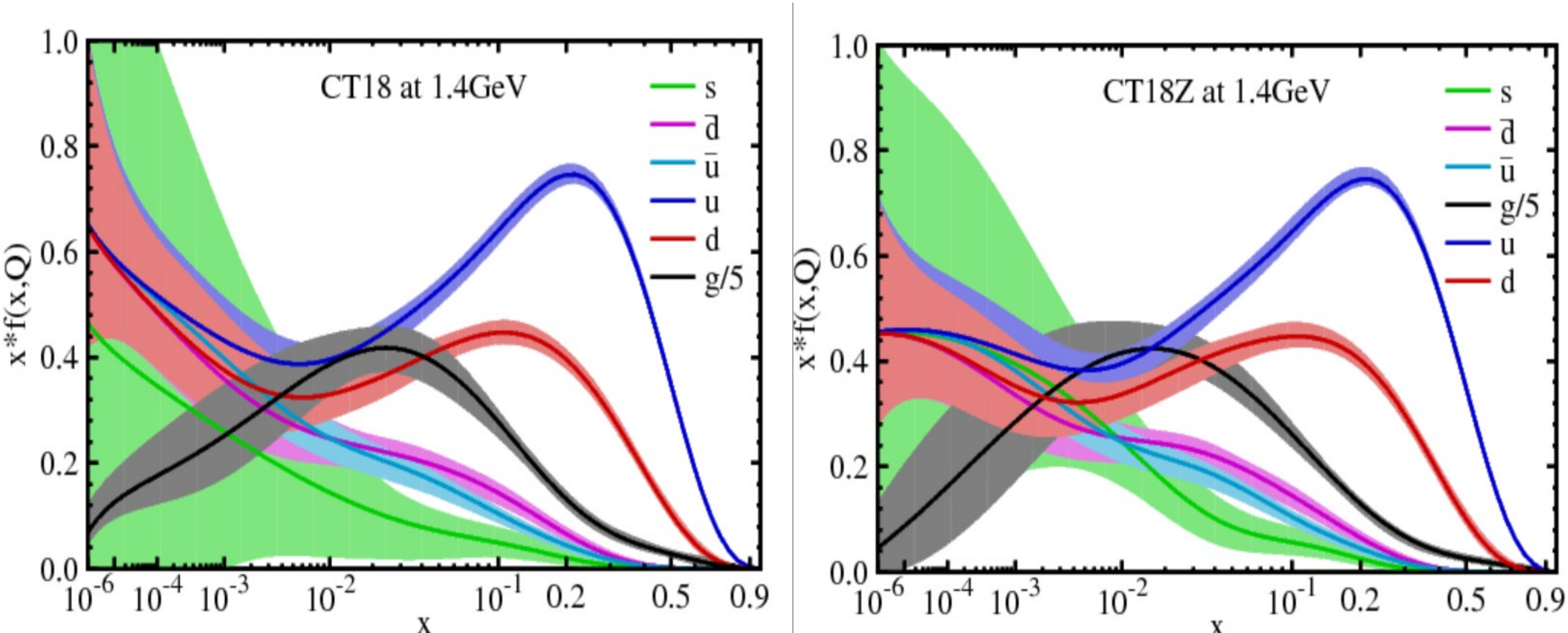
Frontiers of the PDF analysis



Significant advances on all frontiers will be necessary to meet the targets of the HL-LHC program

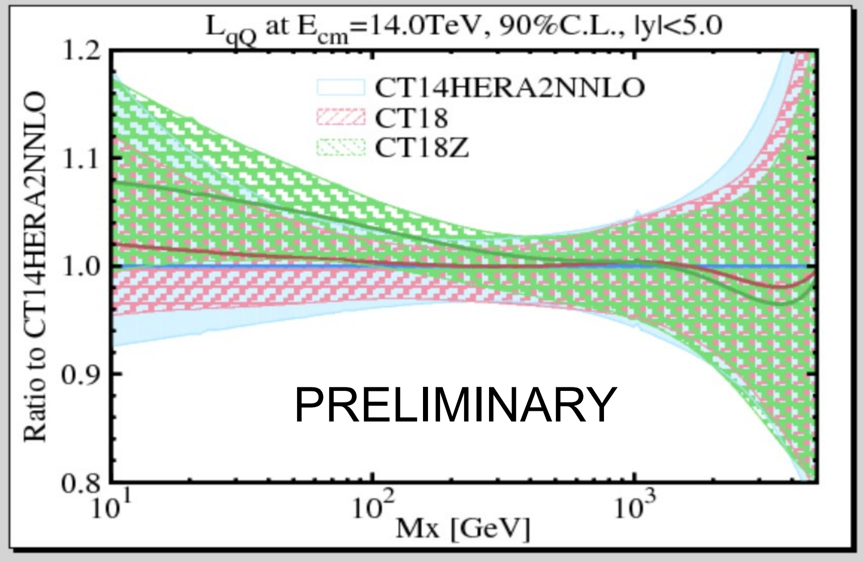
CT18 parton distributions

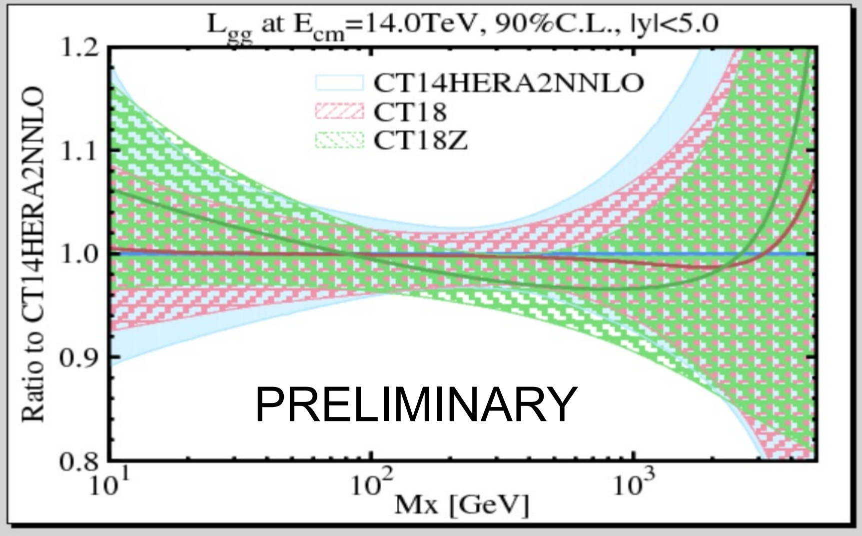
Four PDF ensembles: CT18 (default), A, X, and Z

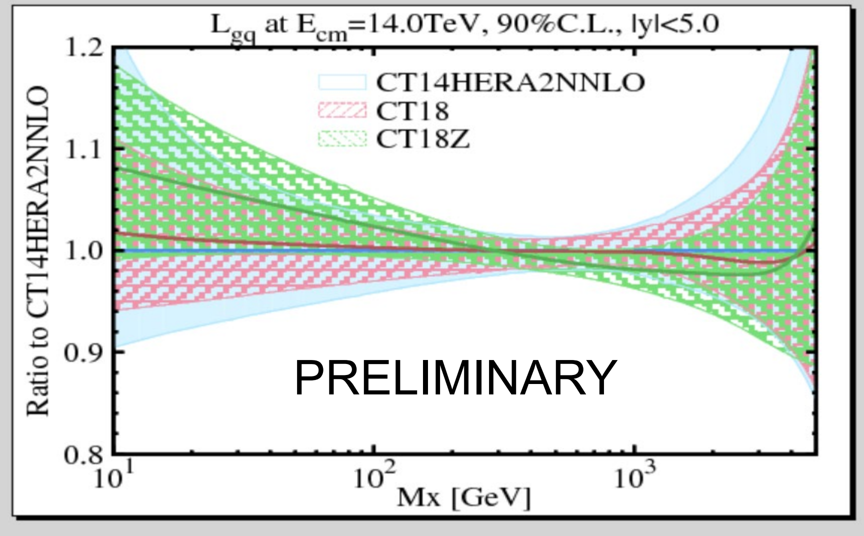


CT18Z has enhanced gluon and strange PDFs at $x \sim 10^{-4}$, and reduced light-quark PDFs at $x < 10^{-2}$. The CT18Z fit is performed so as to maximize the differences from CT18 PDFs, while preserving about the same goodness-of-fit as for CT18. CT18A and CT18X include some features of CT18Z

CT18/CT18Z parton luminosities



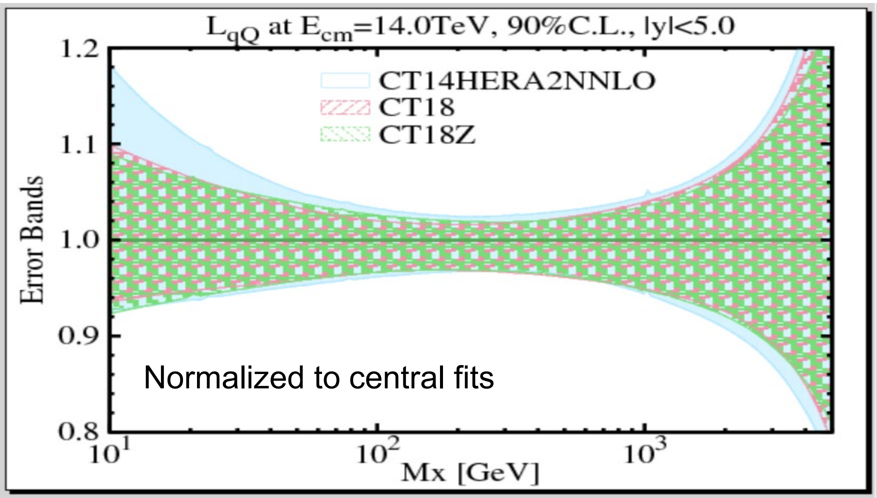


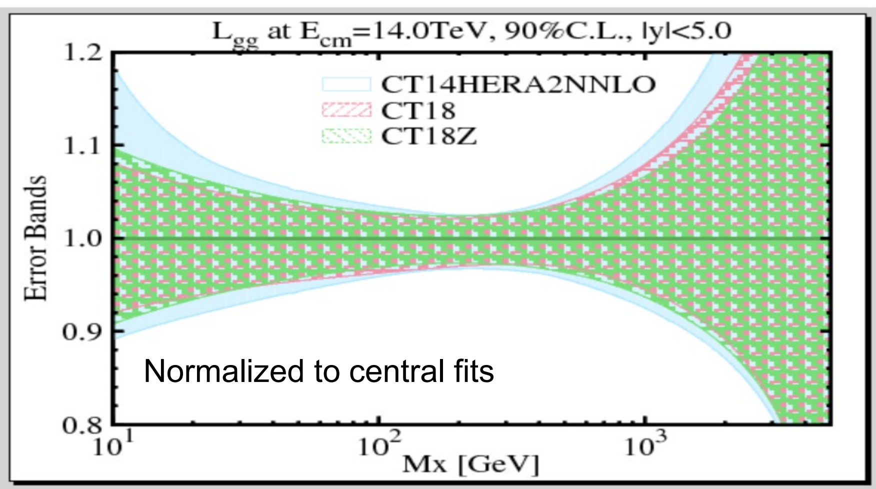


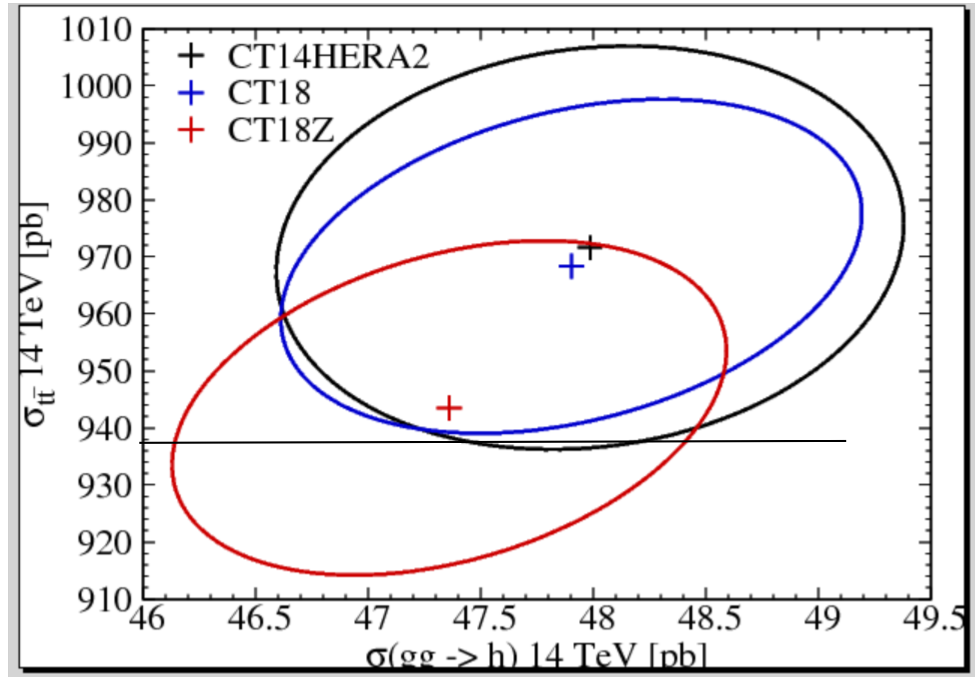
CT18 consistent with CT14

CT18Z has a somewhat different shape, especially at low invariant masses M_X

Mild reduction in nominal PDF error bands and cross section uncertainties





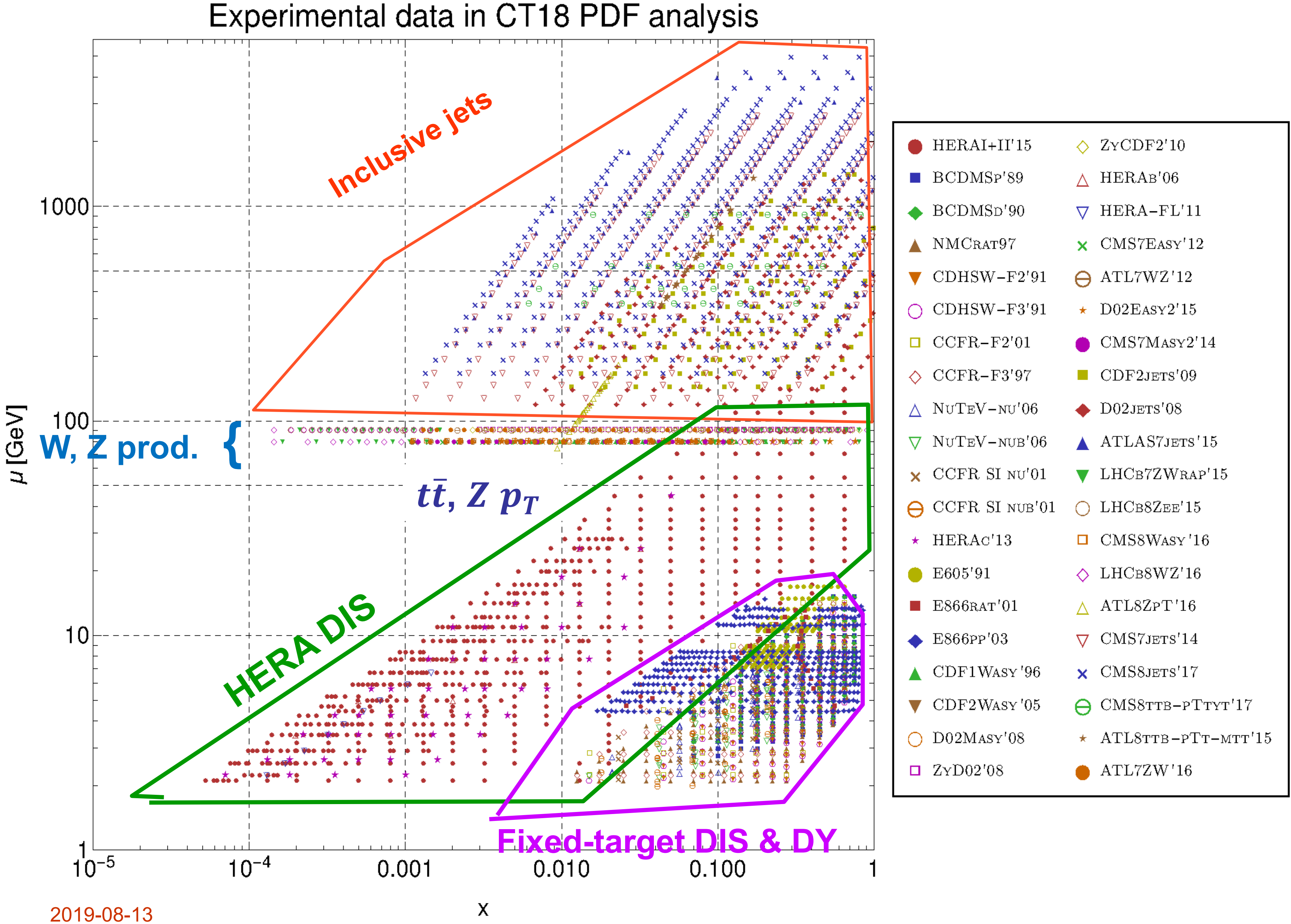


CT18Z $gg \to H$ and $t\bar{t}$ production cross sections lower by about 1 and 2.5% compared to CT14HERA2

PRELIMINARY

CT18 in a nutshell

- Start with CT14-HERA2 (HERAI+II combined data released after publication of CT14)
- Examine a wide range of PDF parameterizations
- Use as much relevant LHC data as possible using applgrid/fastNLO interfaces to data sets, with NNLO/NLO K-factors, or fastNNLO tables in the case of top pair production. **Benchmark the predictions!**
- Examine **QCD scale dependence** in key processes
- Implement **parallelization** of the global PDF fitting to allow for faster turn-around time
- Validate the results using a **strong set of goodness-of-fit tests** *(Kovarik, PN, Soper,* arXiv:1905.06957*)*
- Use diverse statistical techniques (**PDFSense, ePump, Gaussian variables, Lagrange Multiplier scans**) to examine agreement between experiments



Experimental data in CT18 PDF analysis
Inclusive jets
W, Z prod.
$t\bar{t}$, $Z\ p_T$
HERA DIS
Fixed-target DIS & DY
μ [GeV]
x
1000
100
10
1
10^{-5}
10^{-4}
0.001
0.010
0.100
1
HERAI+II'15
BCDMSp'89
BCDMSd'90
NMCrat97
CDHSW-F2'91
CDHSW-F3'91
CCFR-F2'01
CCFR-F3'97
NuTeV-nu'06
NuTeV-nub'06
CCFR SI nu'01
CCFR SI nub'01
HERAc'13
E605'91
E866rat'01
E866pp'03
CDF1Wasy'96
CDF2Wasy'05
D02Masy'08
ZyD02'08
ZyCDF2'10
HERAb'06
HERA-FL'11
CMS7Easy'12
ATL7WZ'12
D02Easy2'15
CMS7Masy2'14
CDF2jets'09
D02jets'08
ATLAS7jets'15
LHCb7ZWrap'15
LHCb8Zee'15
CMS8Wasy'16
LHCb8WZ'16
ATL8ZpT'16
CMS7jets'14
CMS8jets'17
CMS8ttb-pTtyt'17
ATL8ttb-pTt-mtt'15
ATL7ZW'16

2019-08-13

New LHC datasets for CT18

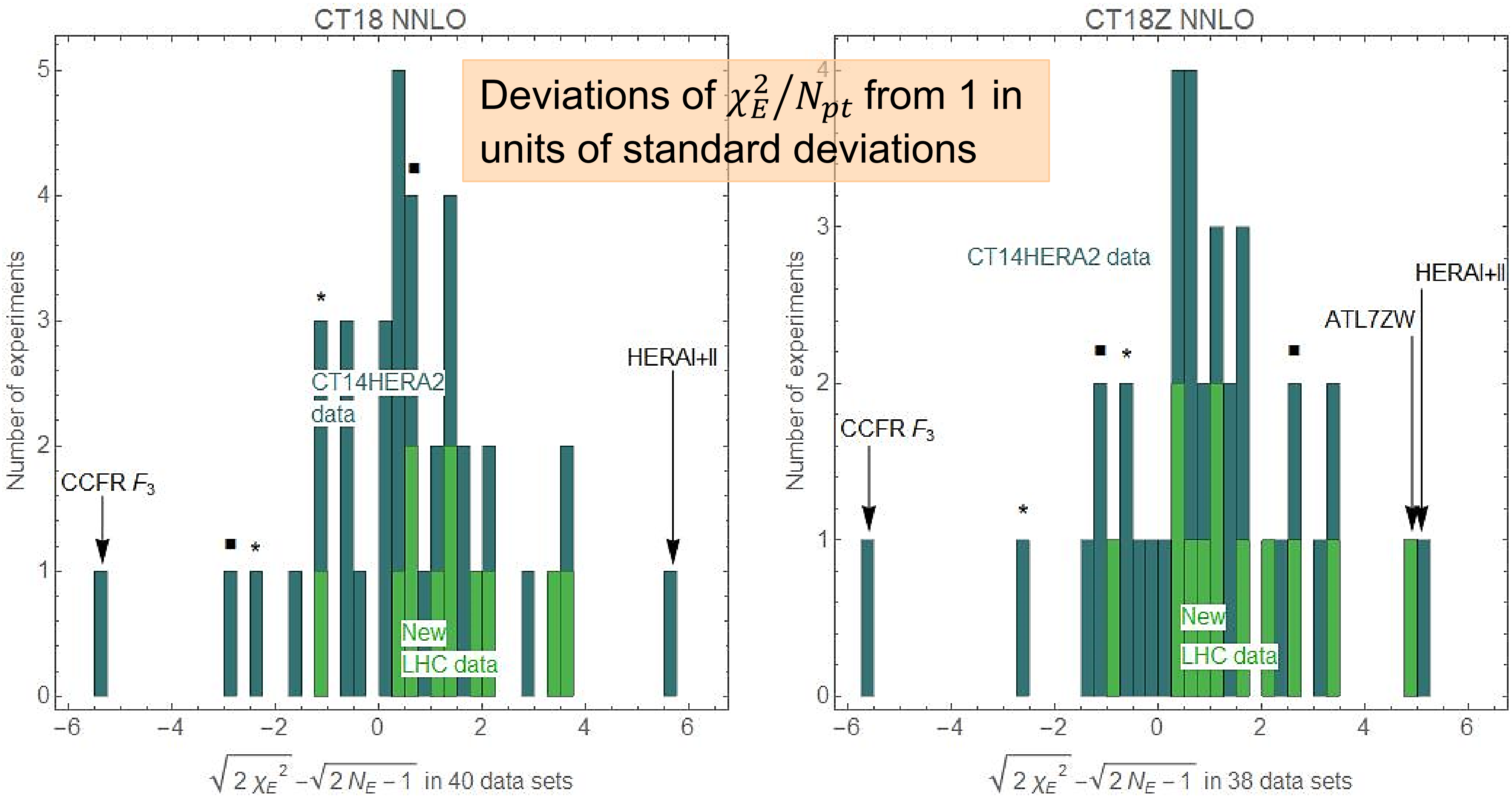
1. 245 1505.07024 LHCb Z (W) muon rapidity at 7 TeV(applgrid)
2. 246 1503.00963 LHCb 8 TeV Z rapidity (applgrid);
3. 249 1603.01803 CMS W lepton asymmetry at 8 TeV (applgrid)
4. 250 1511.08039 LHCb Z (W) muon rapidity at 8 TeV(applgrid)
5. 253 1512.02192 ATLAS 7 TeV Z pT (applgrid)
6. 542 1406.0324 CMS incl. jet at 7 TeV with R=0.7 (fastNLO)
7. 544 1410.8857 ATLAS incl. jet at 7 TeV with R=0.6 (applgrid)
8. 545 1609.05331 CMS incl. jet at 8 TeV with R=0.7 (fastNLO)
9. 565 1511.04716 ATLAS 8 TeV tT pT diff. distributions (fastNNLO)
10. 567 1511.04716 ATLAS 8 TeV tT mtT diff. distributions (fastNNLO)
11. 573 1703.01630 CMS 8 TeV tT (pT , yt) double diff. distributions (fastNNLO)

12. 248 1612.03016 ATLAS 7 TeV Z and W rapidity (applgrid)->CT18Z
  - also uses a special small-x factorization scale, charm mass m_c=1.4 GeV
  - serious changes in PDFs, so warrants a separate PDF

CT18 (CT18Z) NNLO

13 (14) new LHC experiments with 665 (711) data points

LHC experiments, especially ATLAS 7 TeV Z, W production (only in the CT18A and Z fits) tend to have elevated χ_n^2/N_{pt} in global fits



CT14 PDFs with HERA1+2 (=HERA2) combination

Phys.Rev. D95 (2017) 034003

Separate the four HERA2 DIS processes;
(Q_{cut} = 2 GeV)

	N_{pts}	$\chi^2_{red.} / N_{pts}$
NC e^+p	880	1.11
CC e^+p	39	1.10
NC e^-p	159	1.45
CC e^-p	42	1.52
totals		
[reduced χ^2] /N	1120	1.17
χ^2 / N	1120	1.25
R^2 / N	1120	0.08

e^+p data are fitted fine

e^-p data are fitted poorly

reduced χ^2 values

χ^2 = [reduced χ^2] + R^2

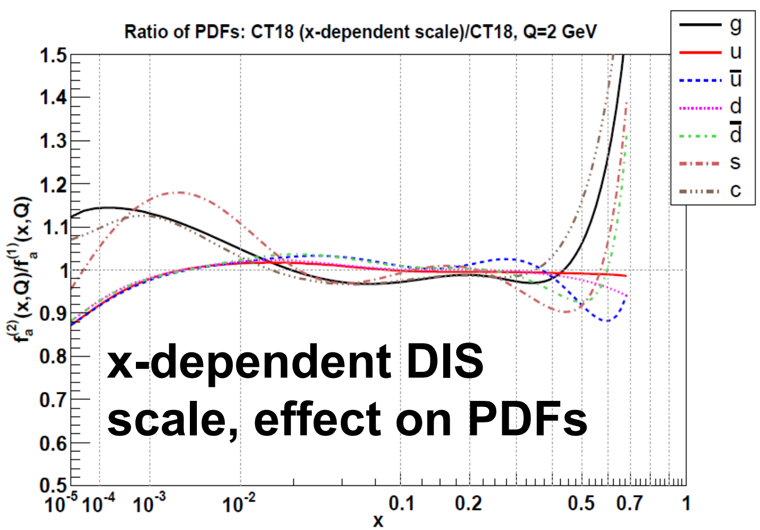
The quadratic penalty for 162 systematic errors = 87.5

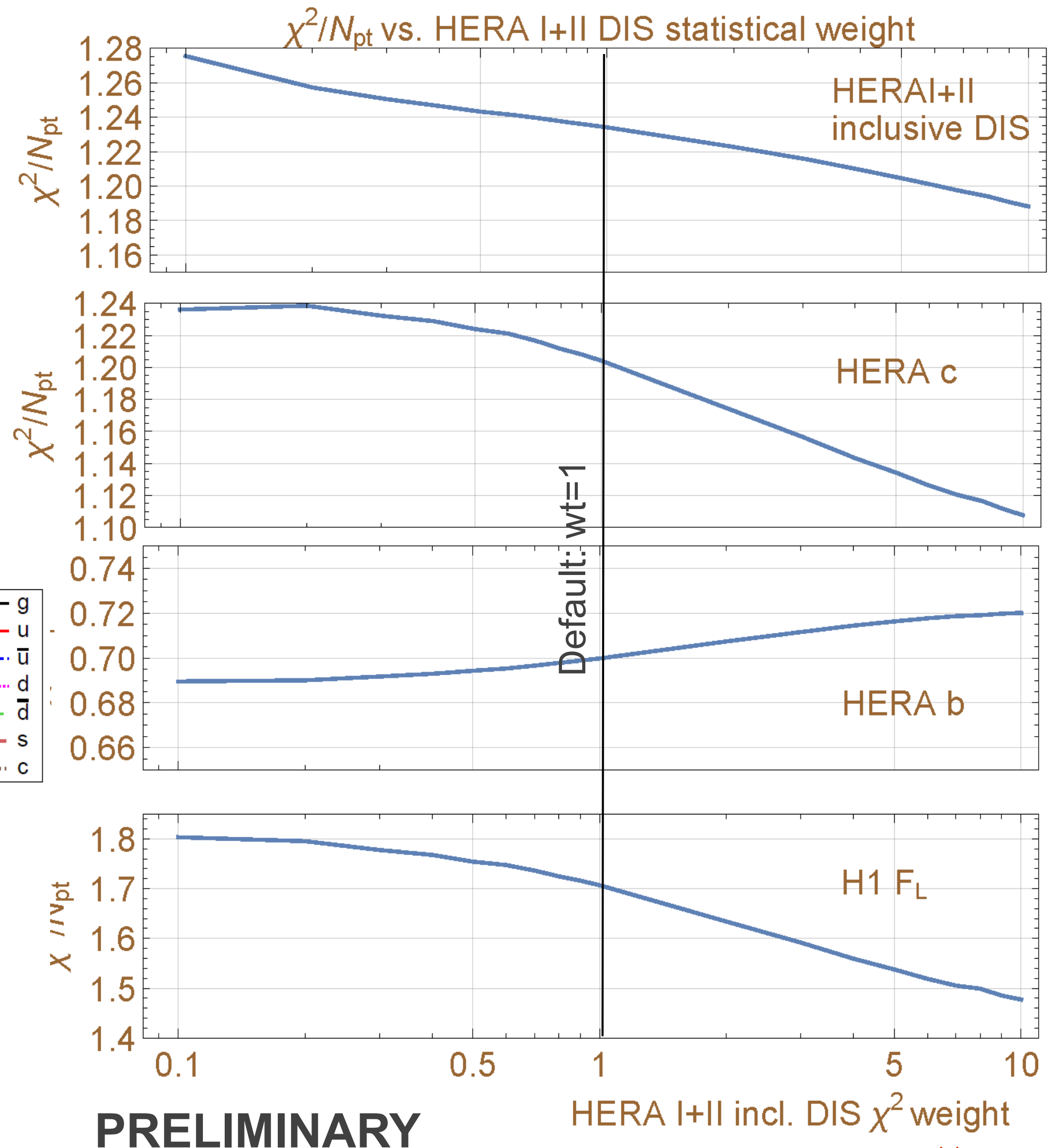
Fair (not perfect) agreement; can be mildly improved by the QCD scale choice

CT18X and Z: a special factorization scale in DIS

The CT18Z fits uses a $\mu_{DIS,X}$ scale that reproduces many features of NNLO-NLLx fits with $\ln(1/x)$ resummation by the NNPDF [arXiv:1710.05935] and xFitter [1802.0064] groups.

$$\mu_{DIS,X}^2 = 0.8^2 \left(Q^2 + \frac{0.3\ GeV^2}{x^{0.3}} \right)$$

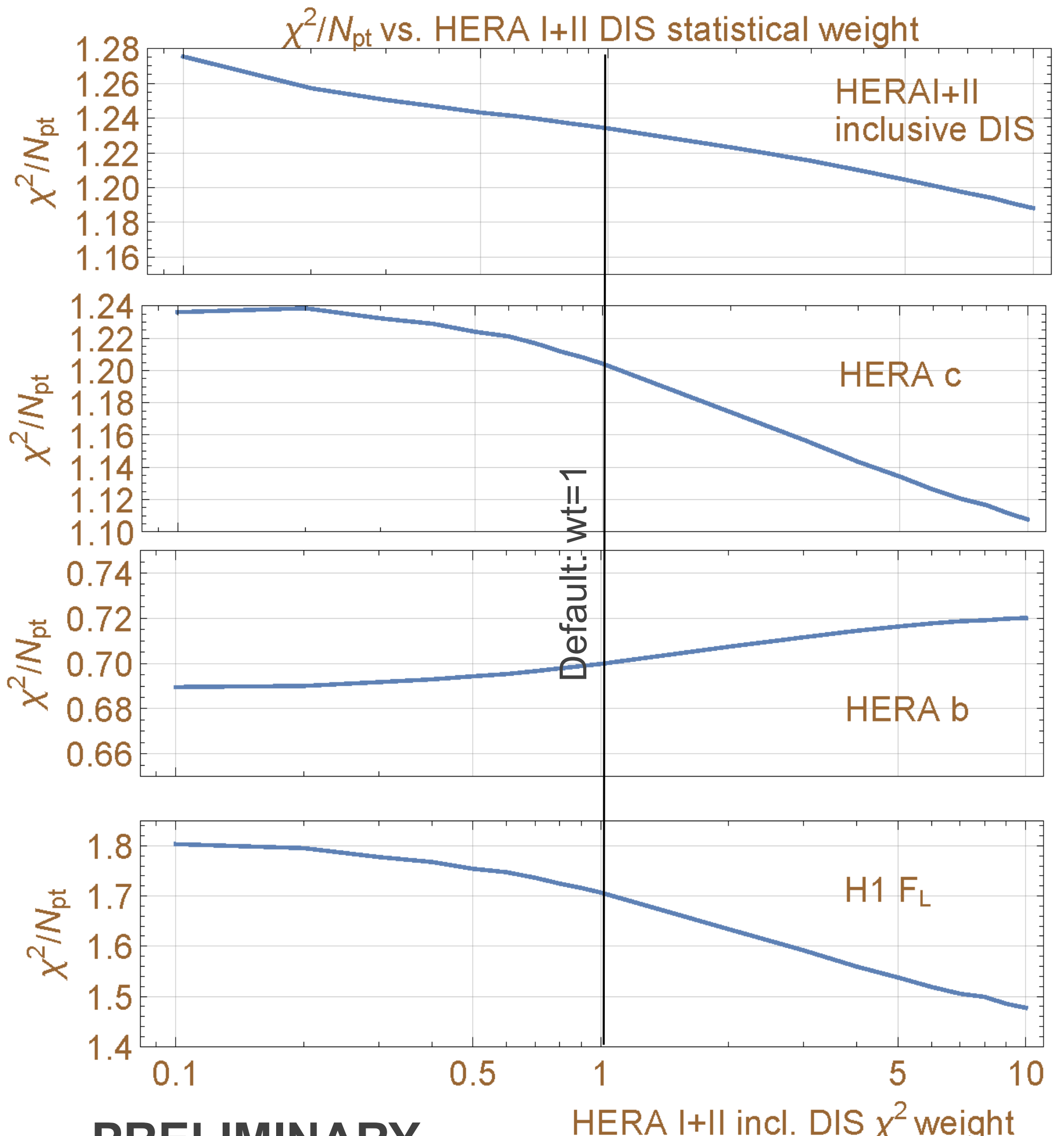




CT18X and Z: a special factorization scale in DIS

Right: when the χ^2 weight for the **inclusive** HERA I+II DIS is increased to $wt = 10$ to suppress pulls from the other experiments, χ^2_{CT18Z}/N_{pt} for HERA I+II DIS **and** HERA charm production decreases to about the same levels as in HERA-only NNLO+NLLx fits by other groups.

- **NNLO with an x-dependent scale is statistically indistinguishable from BFKL resummation in the CT18 x-Q region (Q > 2 GeV)**



PRELIMINARY

Theory input

Obs.	Expt.	fast table	NLO code	K-factors	R,F scales
Inclusive jet	ATL 7 CMS 7/8	APPLgrid fastNLO	NLOJet++	NNLOJet	p_T, p_T^1
p_T^Z	ATL 8	APPLgrid	MCFM	NNLOJet	$\sqrt{Q^2 + p_{T,Z}^2}$
W/Z rapidity W asymmetry	LHCb 7/8 ATL 7 CMS 8	APPLgrid	MCFM/aMCfast	FEWZ/MCFM	$M_{W,Z}$
DY (low,high mass)	ATL 7/8 CMS 8	APPLgrid	MCFM/aMCfast	FEWZ/MCFM	Q_{ll}
$t\bar{t}$	ATL 8 CMS 8	fastNNLO			$\frac{H_T}{4}, \frac{m_T}{2}$

when justified, a small Monte-Carlo error (typically 0.5%) added for NNLO/NLO K-factors

Theory calculations must be benchmarked before the PDF4LHC'20 combination!

One program/scale not sufficient for understanding theory uncertainties

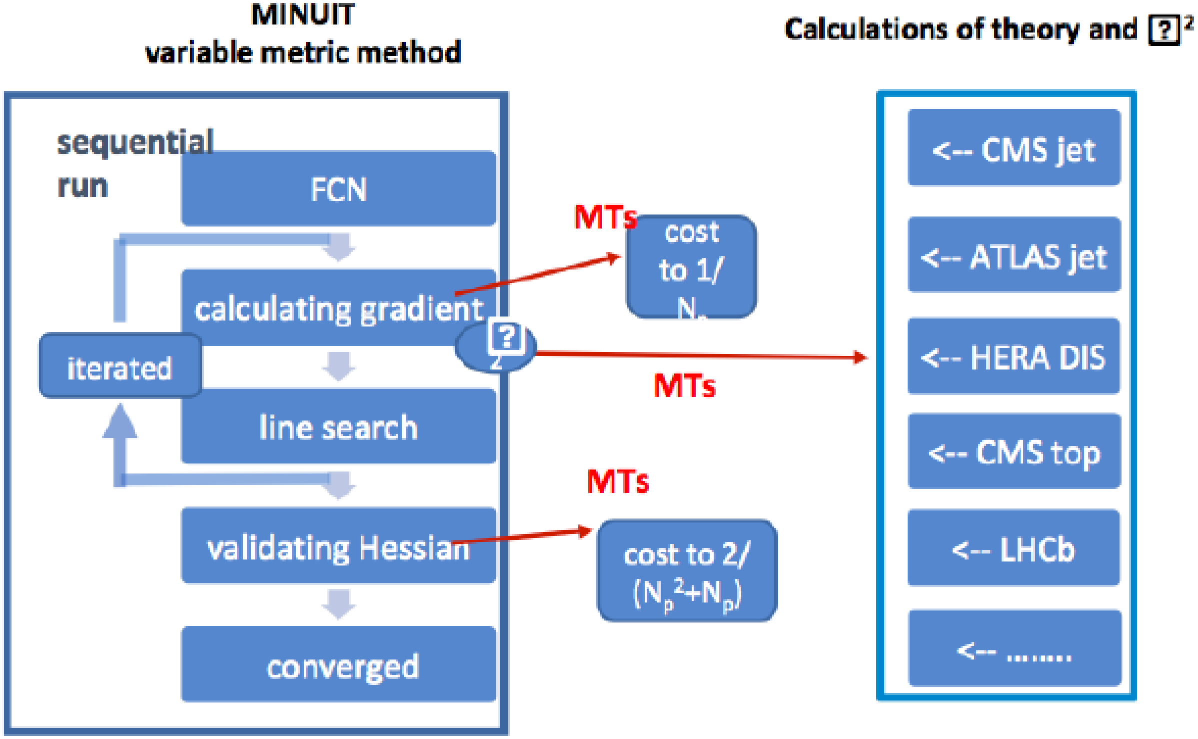
Theoretical calculations for vector boson production

ID	Obs.	Expt.	fast table	NLO code	K-factors	$\mu_{R,F}$
245 246 250 249	$y_{\mu\mu}, \eta_\mu$ y_{ee} $y_{\mu\mu}, \eta_\mu$ $A(\mu)$	LHCb7ZW LHCb8Z LHC8ZW CMS8W	APPLgrid	MCFM/aMCfast	MCFM/FEWZ	$M_{Z,W}$
253	p_T^{ll}	ATL8Z	APPLgrid	MCFM	NNLOJet	M_T^{ll}
201 203 204	$\sqrt{\tau}, y$ $\sigma_{pd}/\sigma_{pp}, x_F$ Q, x_F	E605 E866 E866	CTEQ		FEWZ	Q_{ll}
225 227 234 281	$A(e)$ $A(e)$ $A(\mu)$ $A(e)$	CDF1Z CDF2W D02W D02W	CTEQ		ResBos	Q_{ll} M_W
260 261	y_{ll} y_{ll}	D02 CDF2	CTEQ		VRAP	Q_{ll}
266 267 268	$A(\mu)$ $A(e)$ $y_{ll}, \eta_l, A(l)$	CMS7W CMS7W ATL7ZW(2012)	CTEQ		ResBos	M_W $M_{Z,W}$
248	y_{ll}, η_l	ATL7ZW(2016)	APPLgrid	MCFM/aMCfast	MCFM/FEWZ	$M_{Z,W}$

2019-08-13

Fitting code parallelization with multi-threads

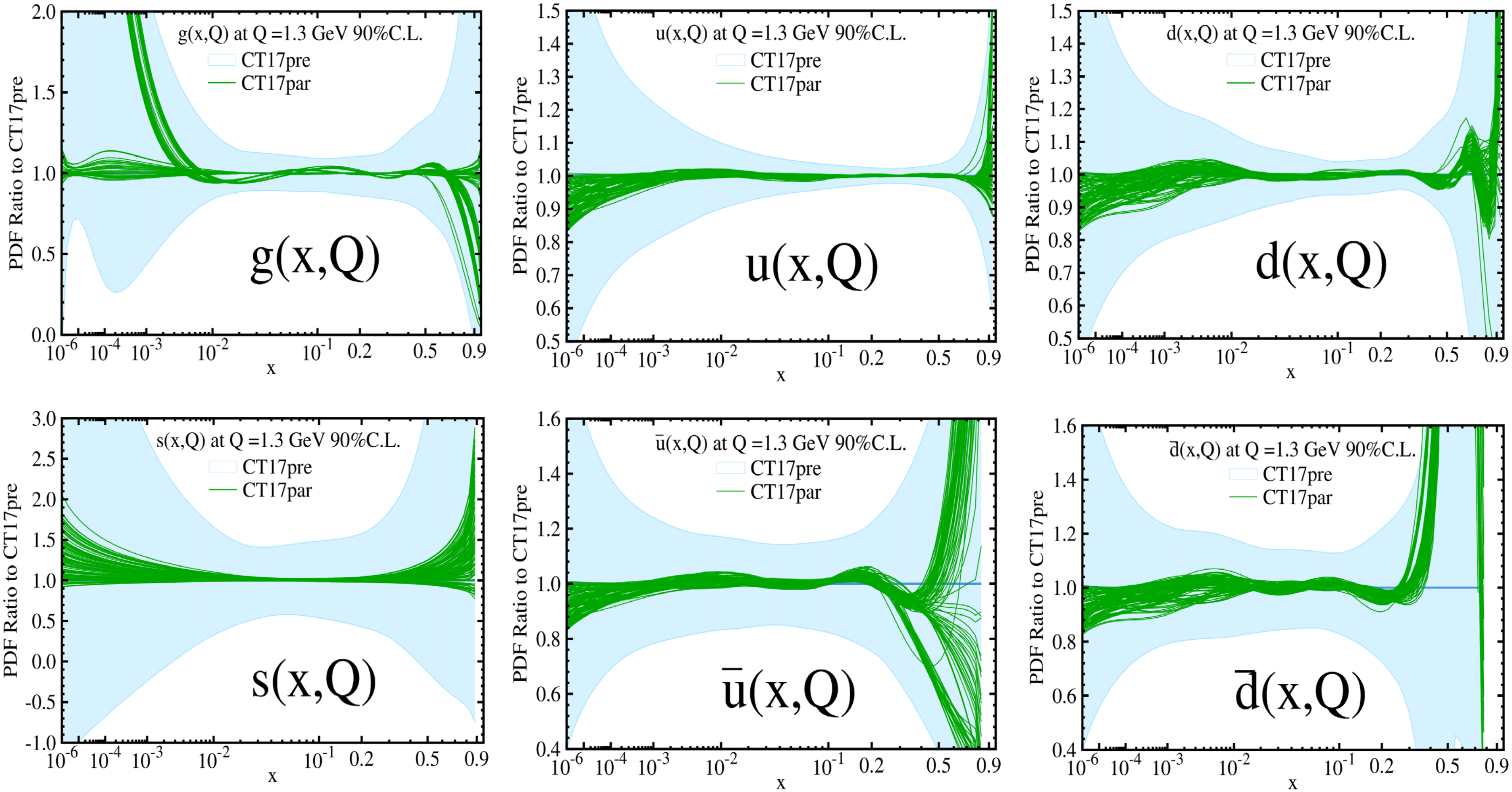
upgrade to a parallelized version of the fitting code, two-layer parallelization: 1. through rearrangement of the minimization algorithm; 2. via redistribution of the data sets



Layer 1: after all a factor of 4~5 improvement on speed is achieved!

Layer 2: further speed up by a factor of 2, depending on data sets included

Explore various non-perturbative parametrization forms of PDFs



- CT17par – sample result of using various non-perturbative parametrization forms.
- No data constrain very large x or very small x regions.

The questions we ask:

Which of 30+ eligible LHC experiments provide promising constraints on the CTEQ-TEA PDFs?

Do the LHC experiments agree among themselves and with other experiments?

The questions we ask:

Which of 30+ eligible LHC experiments provide promising constraints on the CTEQ-TEA PDFs?

Do the LHC experiments agree among themselves and with other experiments?

A **consistent** answer emerges from a powerful combination of four methods:

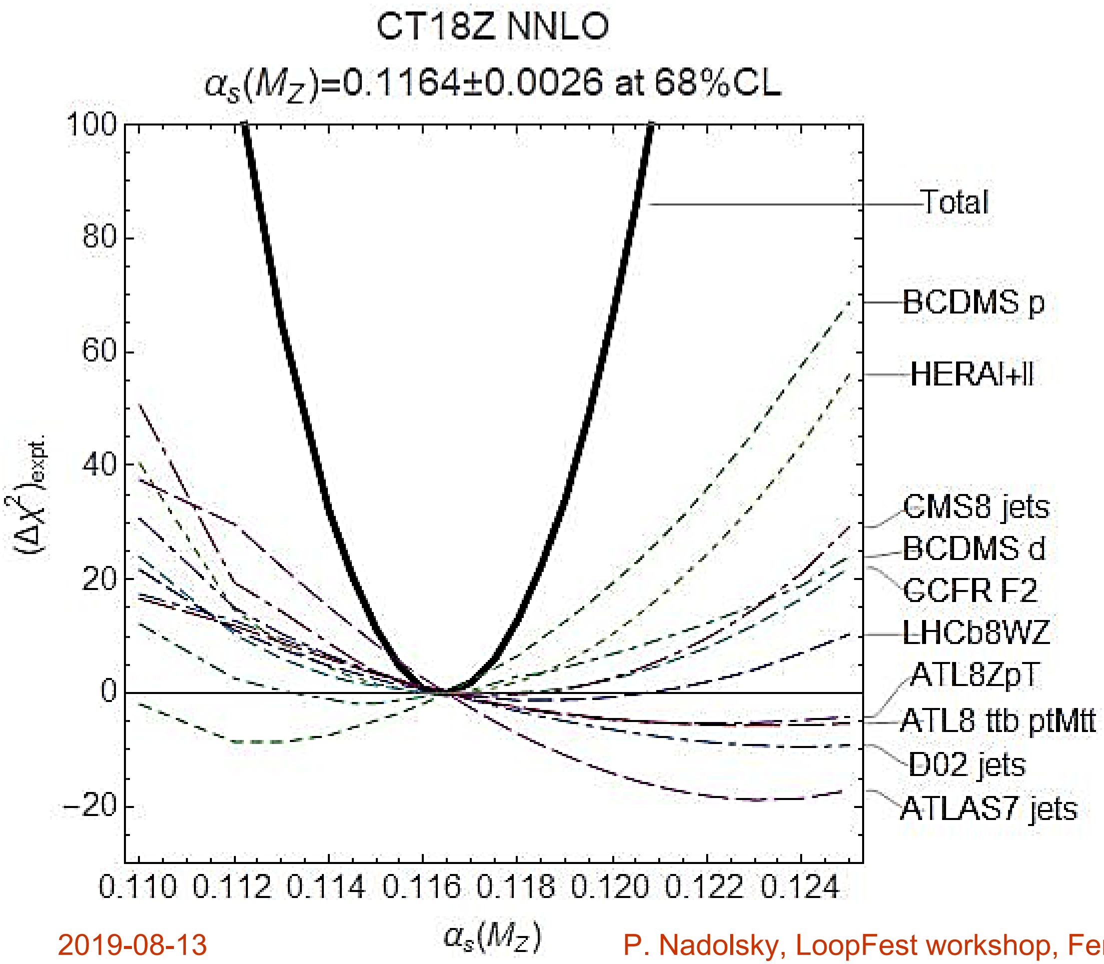
1. **Lagrange multiplier scans** } slow, most accurate
2. **PDFSense and L_2 sensitivity**
3. **ePump** [Schmidt, Pumplin, Yuan, PRD 98, 094005]
4. **Effective Gaussian variables**

} Fast approximations (items 2–4)

Lagrange Multiplier (LM) Scan: $\alpha_s(M_Z)$

The LM scan technique is introduced in **Stump et al., Phys.Rev. D65 (2001) 014012**

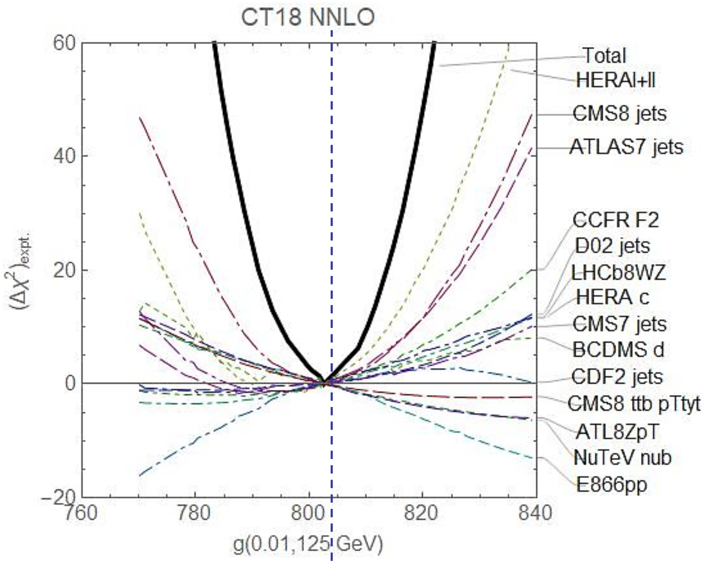
☺ **Detailed dependence of χ^2** ☹ **slow; refitting on a supercomputing cluster**

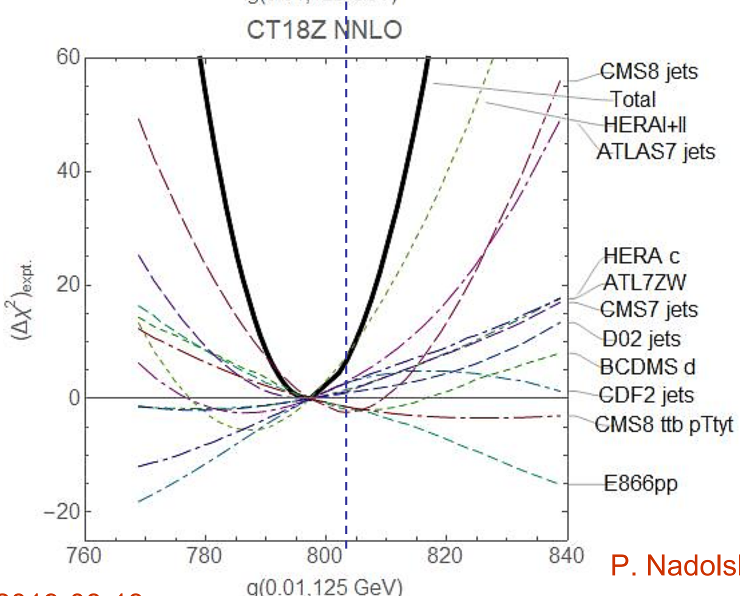


Examine changes in χ^2 for
- all experiments (“Total”)
- individual experiments

$\alpha_s(m_Z)$ from global fit closer to 0.117 than to 0.118, primarily due to pulls from HERA and BCDMS DIS experiments

Lagrange Multiplier scan: $g(0.01, 125\ GeV)$





Upper row: CT18

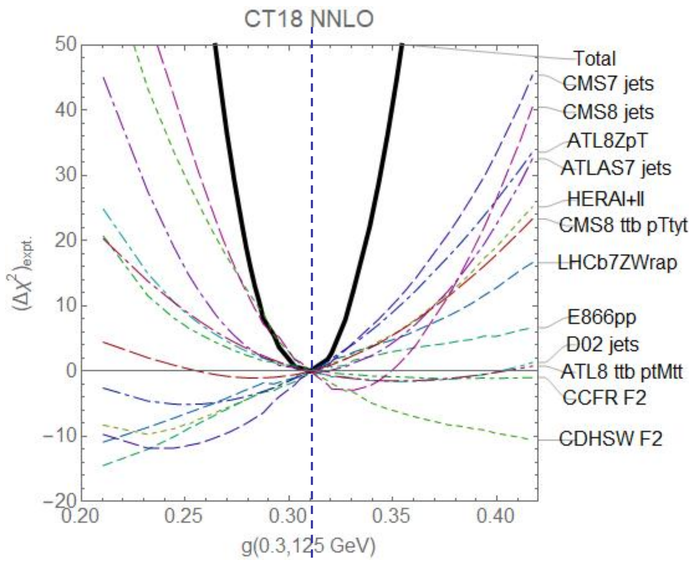
6500 core hours

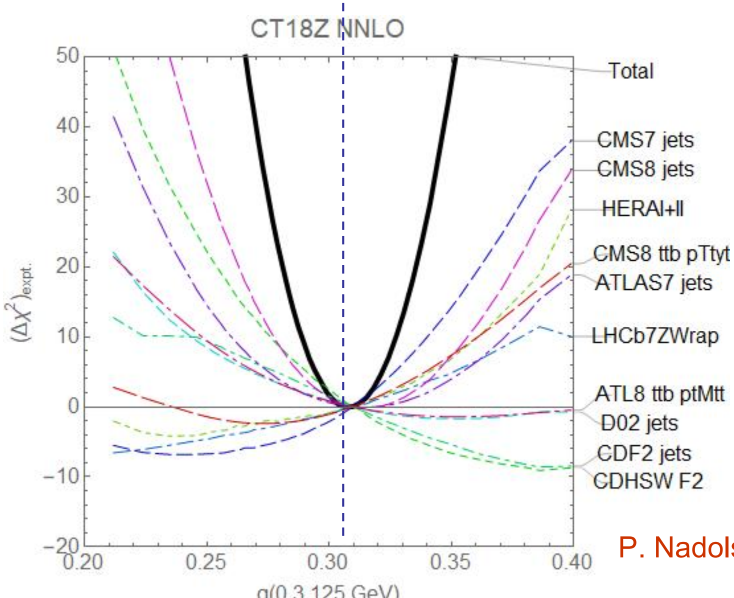
- HERAI+II data set provides the dominant constraint, followed by ATLAS, CDF2, CMS, D02 jet production, HERA charm,…
- $t\bar{t}$ double-diff. cross sections provide weaker constraints

Lower row: CT18Z

- CT18Z: a 1% lower NNLO gluon in the Higgs production region than for CT14/CT18

Lagrange Multiplier scan: $g(0.3, 125\ GeV)$





Upper/lower rows: CT18/CT18Z

Good overall agreement. But observe opposite pulls from ATLAS7/CMS7 jet production and CMS8 jet production

Similarly, ATLAS $t\bar{t}$ distributions $\mathrm{d}^2\sigma/(dp_{T,t}dm_{t\bar{t}})$ and CMS $t\bar{t}$ distributions $\mathrm{d}^2\sigma/(dp_{T,t}dy_{t,ave})$ at 8 TeV impose weak opposite pulls

Constraints from ATLAS 8 $Z\ p_T$ production data are moderate and still affected by NNLO scale uncertainty

PDFSense program

Approximate LM scans, **fast… and many other insights**

PDFSense program: fast surveys of QCD data using a vector data technique

Authors: Tim Hobbs, Bo Ting Wang, et al.: arXiv:1803.02777, 1904.00222, 1907.00988

Language: Mathematica

Inputs: vectors of fitted data residuals $r_i(\vec{a}) \equiv [T_i(\vec{a}\,) - D_i^{shifted}\,(\vec{a})]/\sigma_i$ and predicted cross sections evaluated for Hessian or MC error PDFs

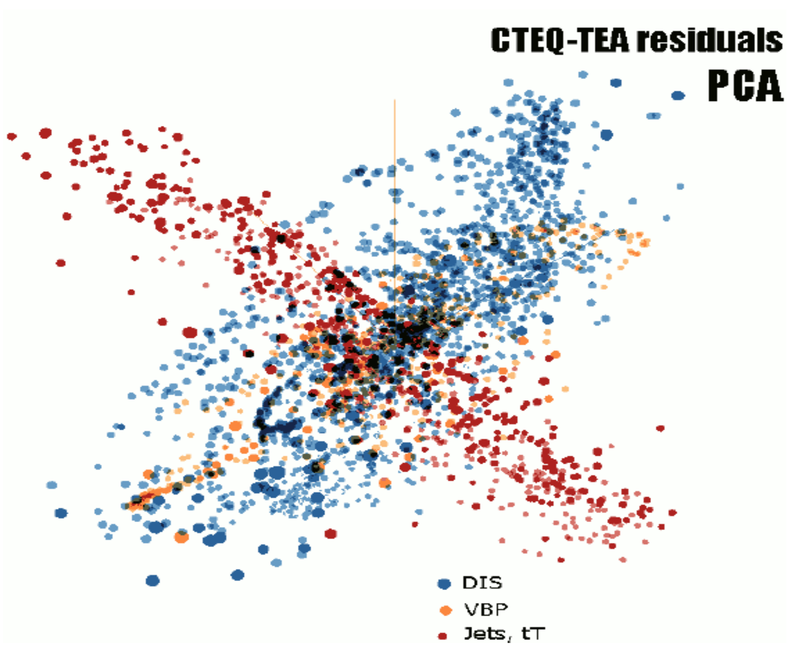
Outputs: Hessian **sensitivity** variables S_f and $S_{f,L2}$: easy-to-compute indicators of data point sensitivity to PDFs in the presence of experimental errors

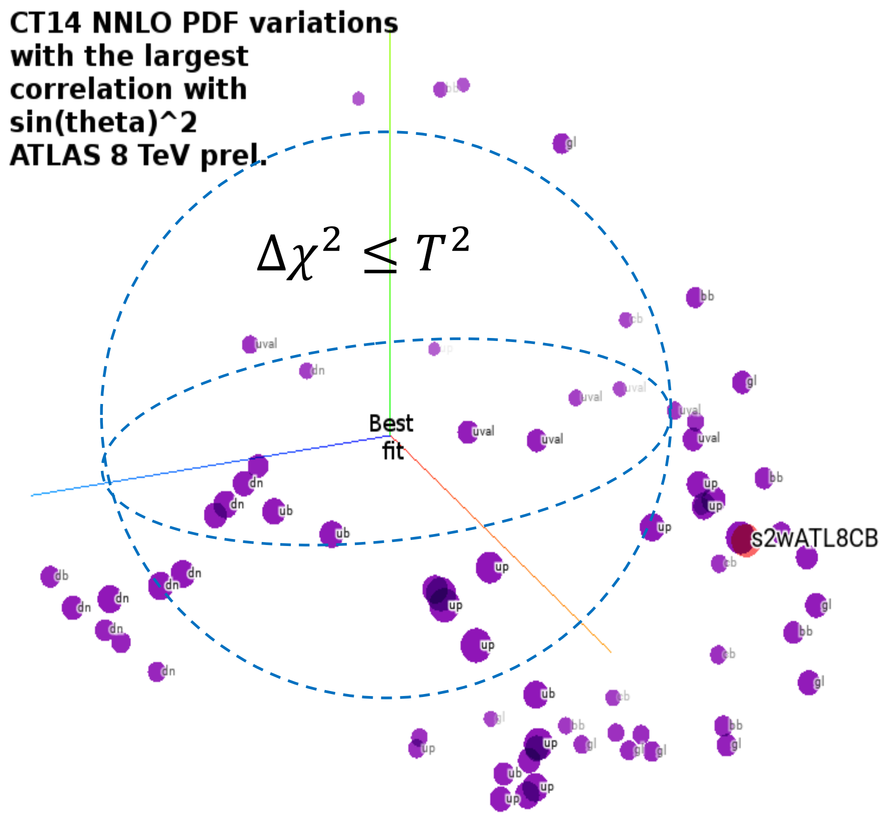
Available by request

Definitions in the backup

Vectors of data point residuals…

… carry detailed information about sensitivity of individual experimental data points to PDFs; can be studied using statistical packages (TensorFlow, Mathematica,…)

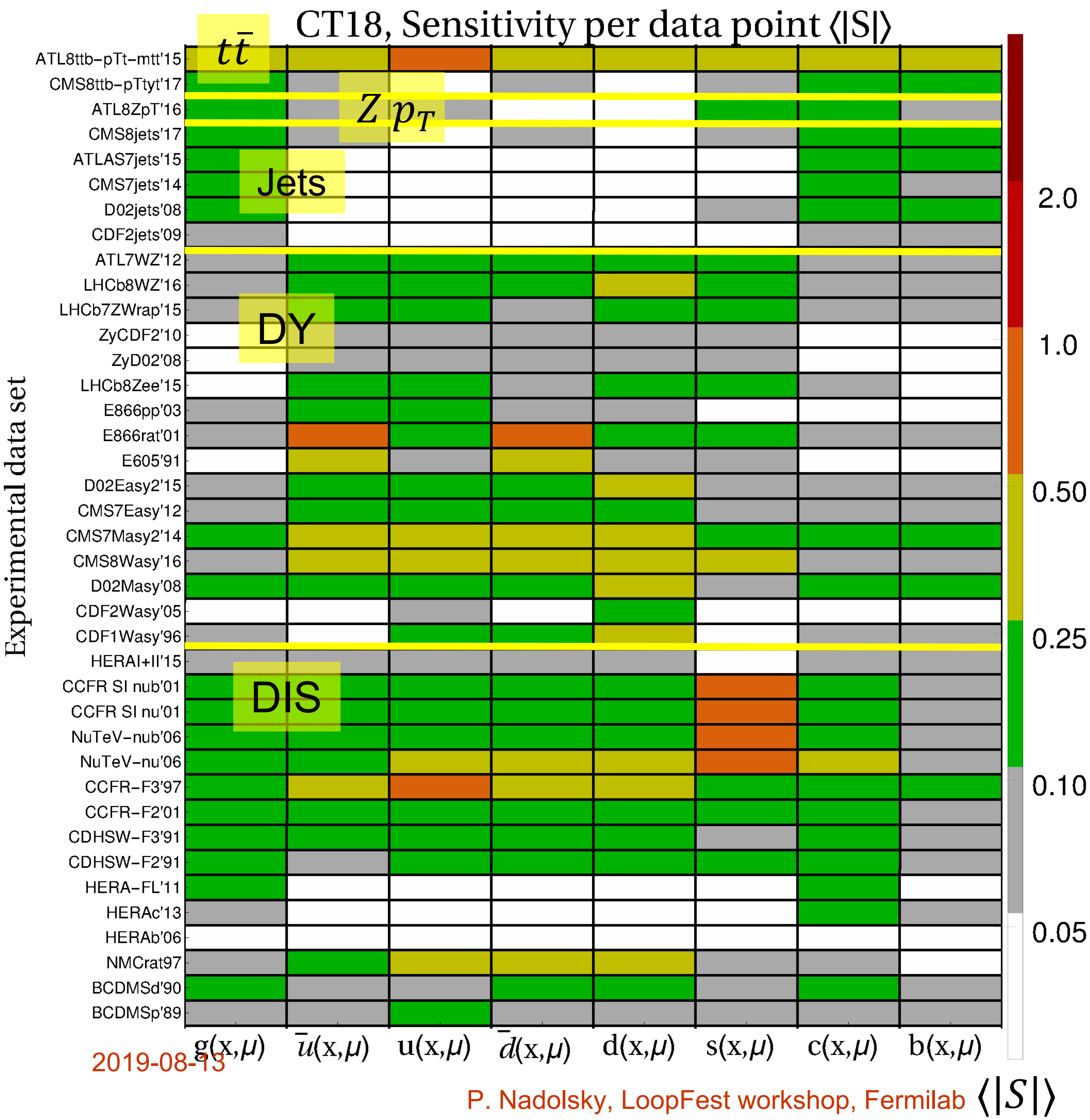




Principal Component Analysis (PCA) visualizes the 56-dim. manifold by reducing it to 10 dimensions (à la META PDFs)

Using Hessian PDFs

Sensitivity of hadronic experiments to PDFs



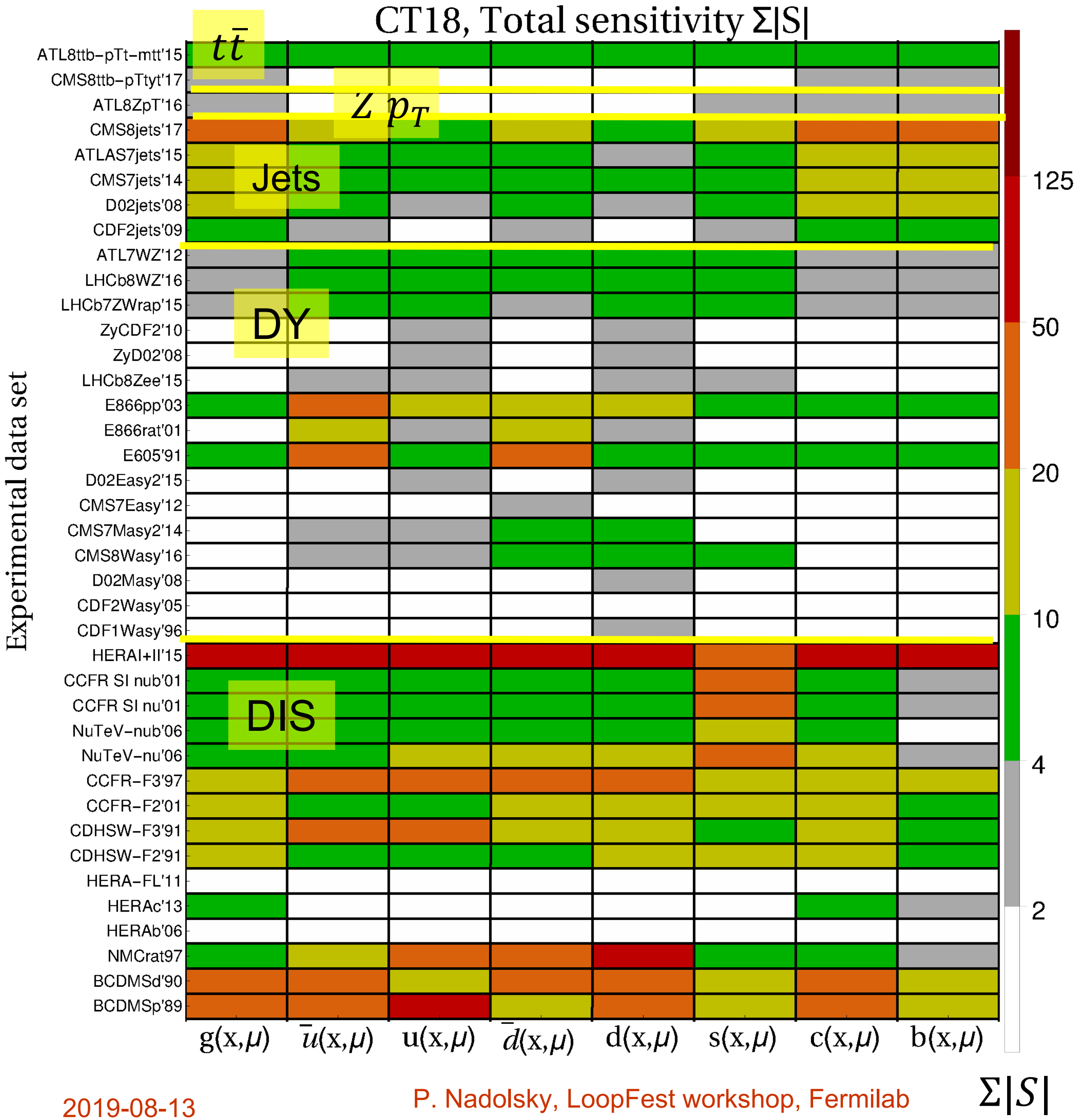
Computed using the **PDFSense** method **[arXiv:0803.02777]**

Average sensitivity to $f_a(x_i, \mu_i)$ **per data point**

- defined in the backup
- expt. and PDF errors included

Red bars = most sensitive experiments

Sensitivity of hadronic experiments to PDFs

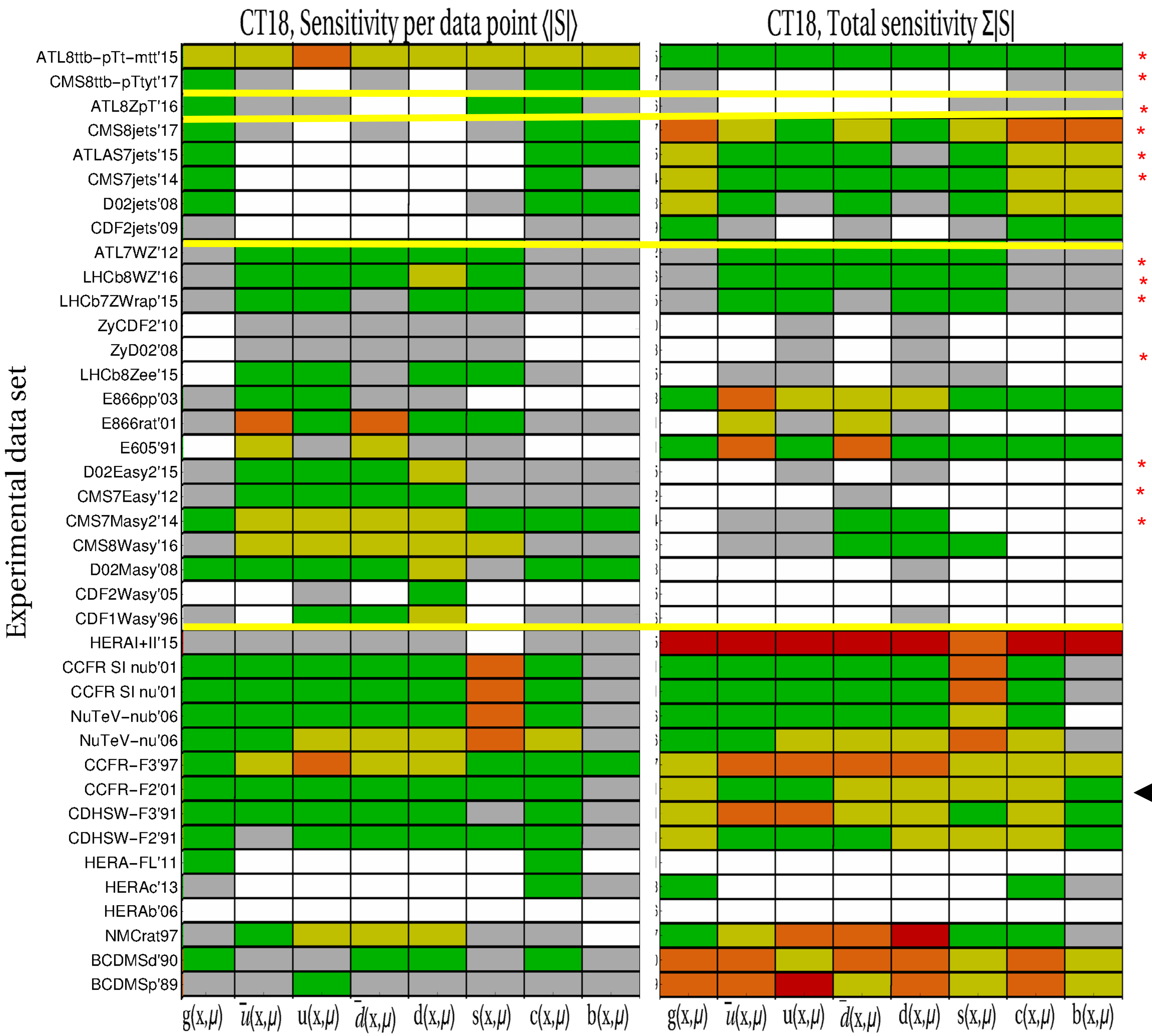


Total sensitivity to $f_a(x_i, \mu_i)$, **summed over data points**

$$\sum_{points} |S_{f,i}|$$

Computed using the **PDFSense** code **[arXiv:0803.02777]**

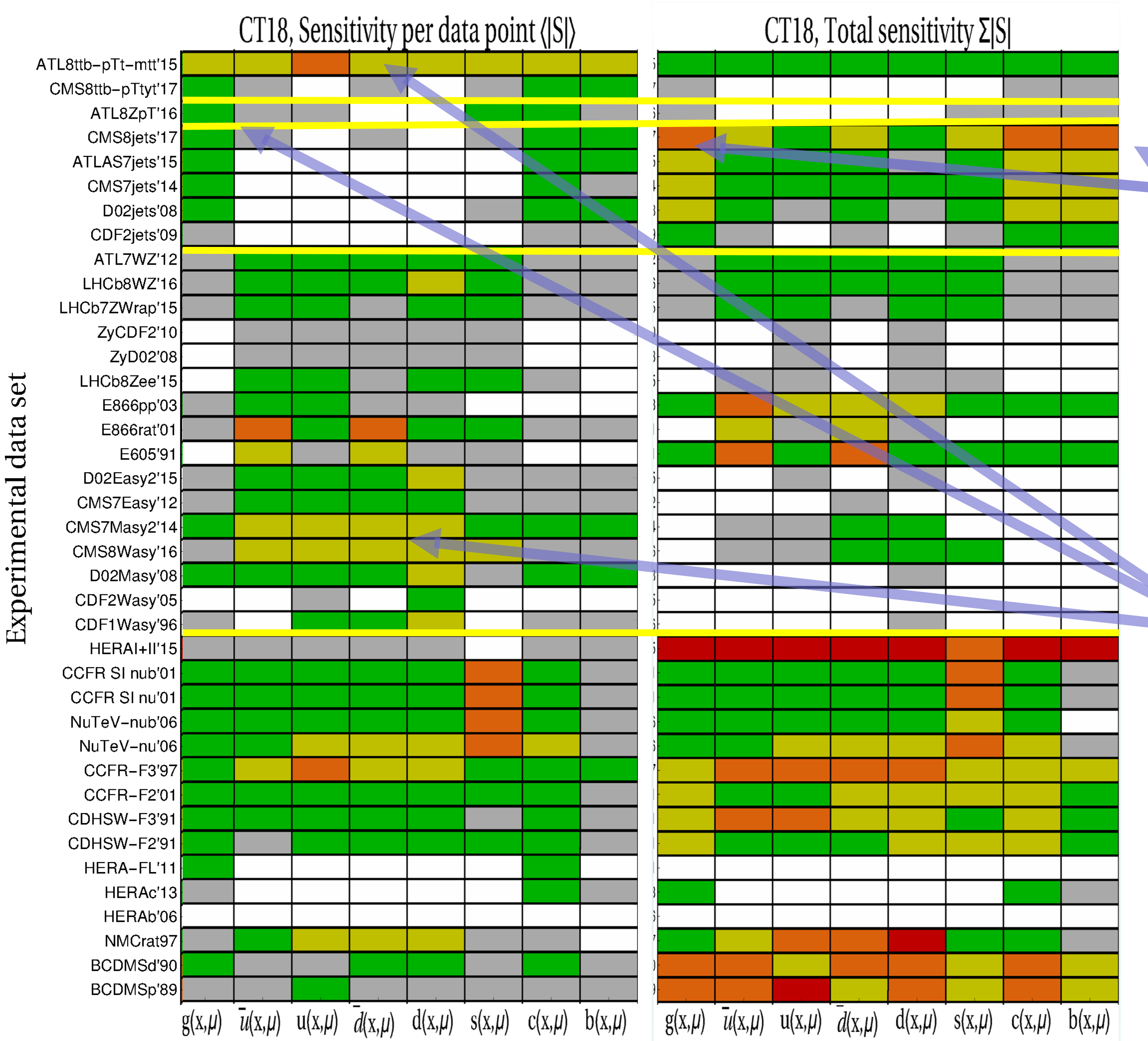
Sensitivity of hadronic experiments to PDFs



The LHC data sets (*) hold a great promise – if they agree

HERA I+II, BCDMS, NMC, DIS data sets dominate experimental constraints. Large numbers of data points matter!

Sensitivity of hadronic experiments to PDFs

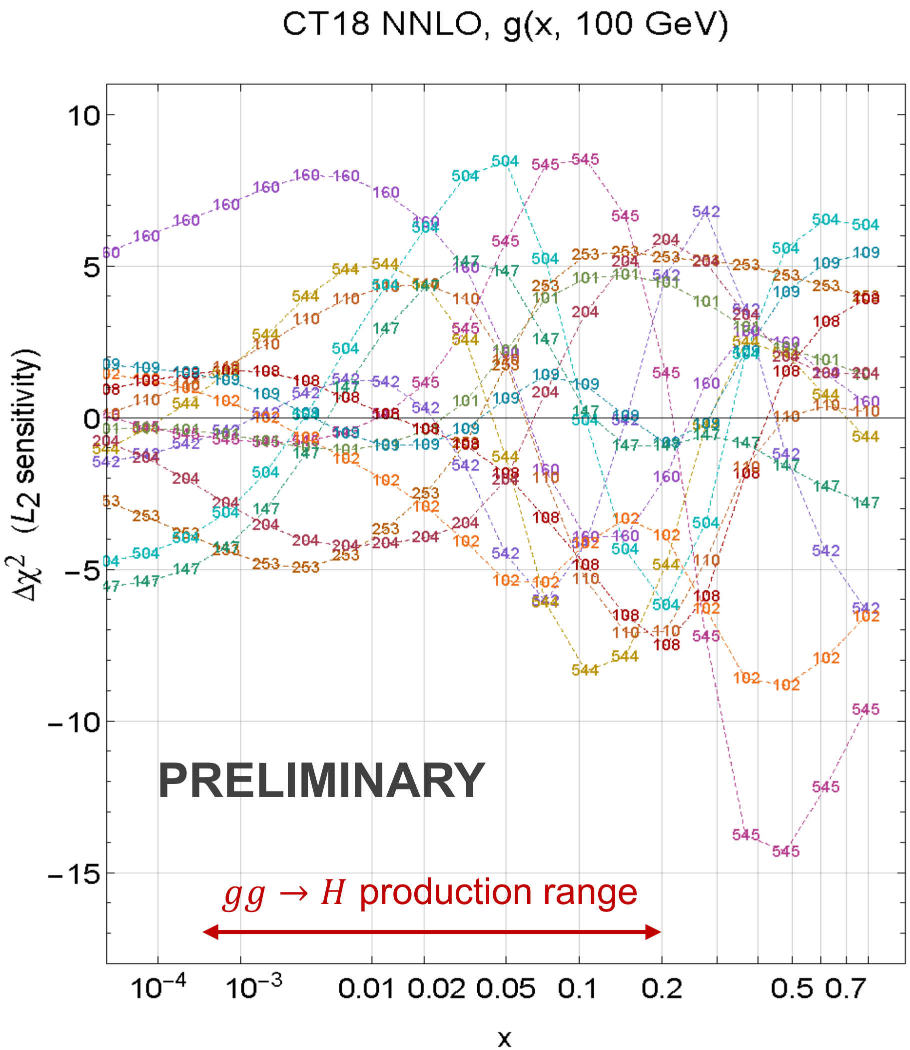


CMS 7 & 8 TeV single-inclusive jet production has highest total sensitivity ($N_{pt} > 100$), modest sensitivity per data point

$t\bar{t}$, CMS W asy, high-p_T Z production have high sensitivity per data point, smaller total sensitivity ($N_{pt} \sim 10 - 20$)

Estimated χ^2 pulls from experiments

(L_2 sensitivity, arXiv:1904.00222, v. 2)



CT18 NNLO, gluon at Q=100 GeV

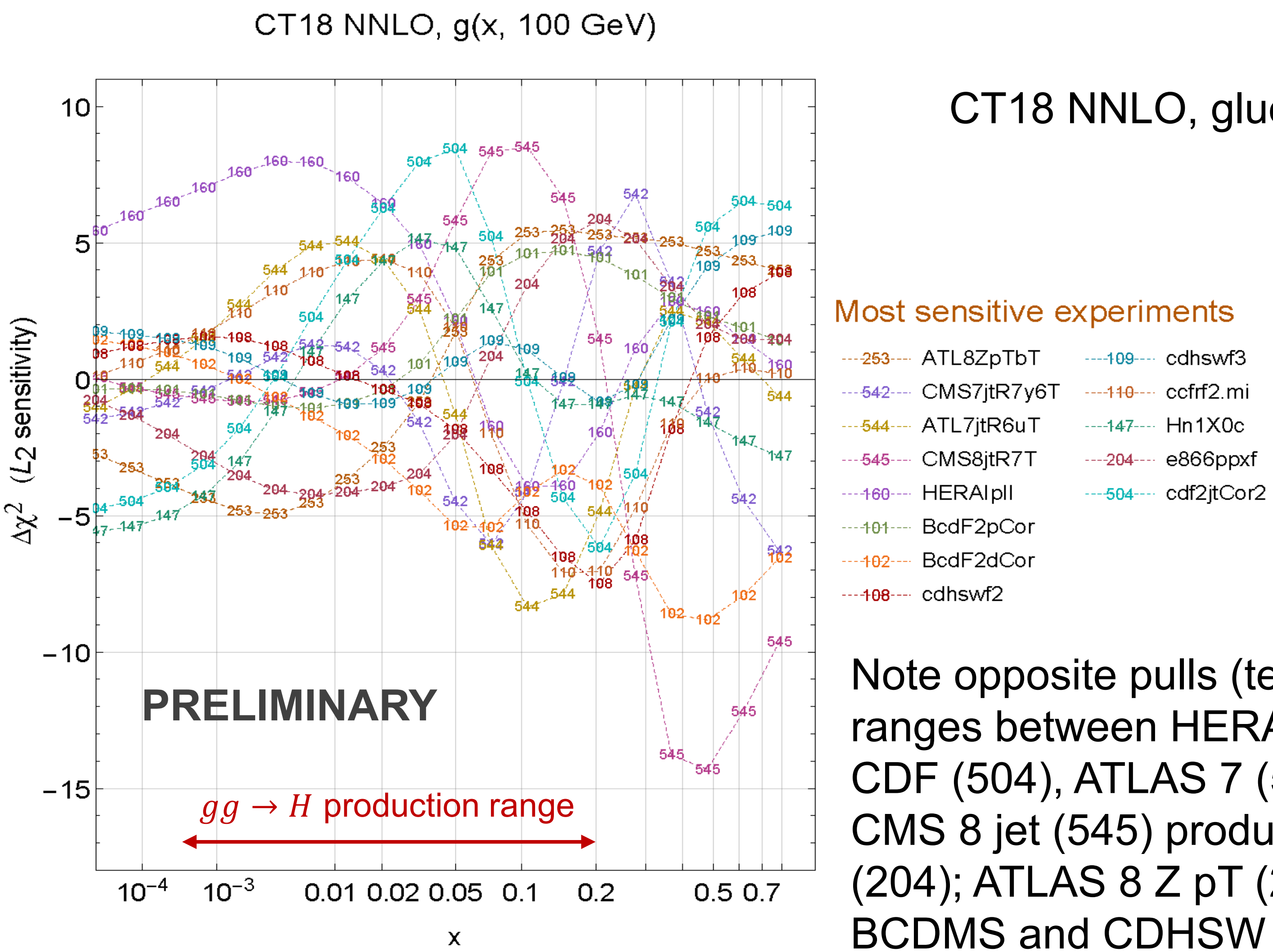
15 core-minutes

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 using CT18 Hessian PDFs

Estimated χ^2 pulls from experiments

(L_2 sensitivity, arXiv:1904.00222, v. 2)

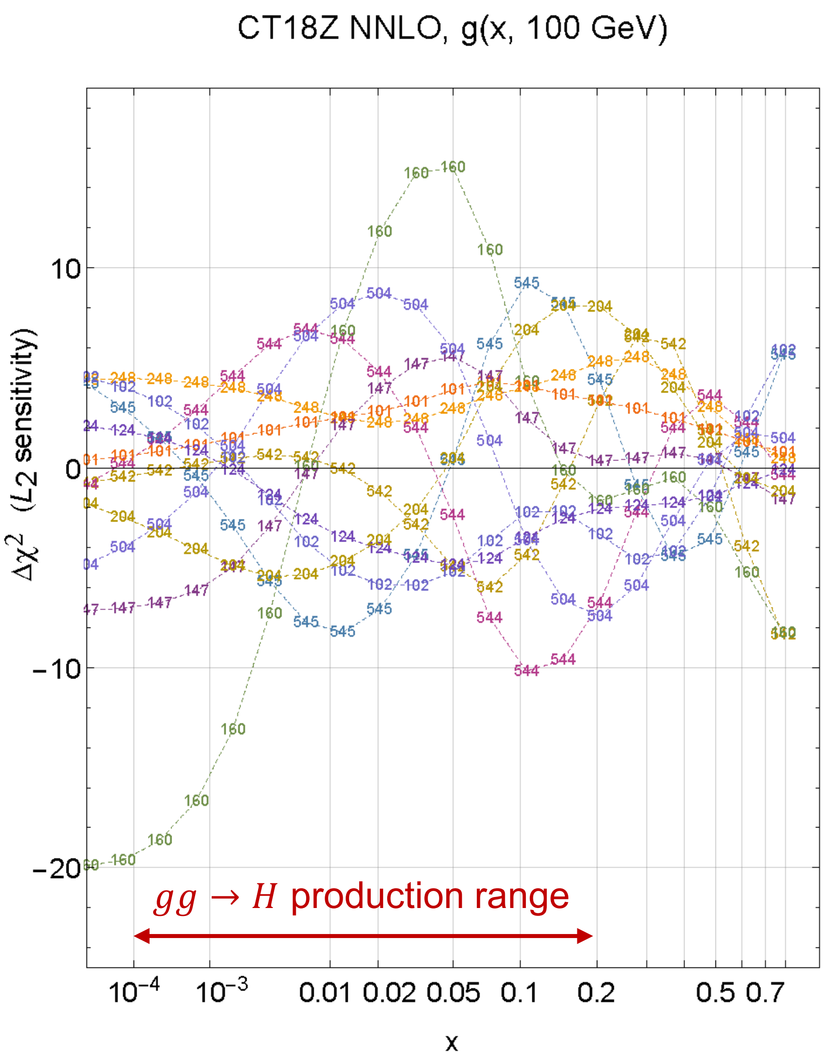


CT18 NNLO, gluon at Q=100 GeV

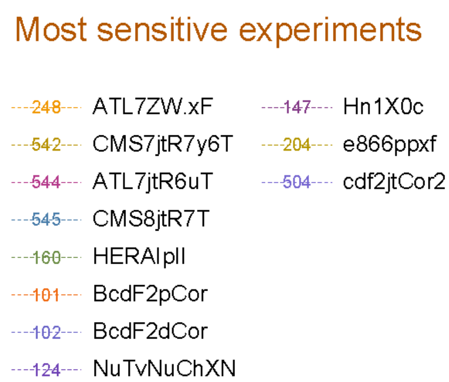
Note opposite pulls (tensions) in some x ranges between HERA I+II DIS (ID=160); CDF (504), ATLAS 7 (544), CMS 7 (542), CMS 8 jet (545) production; E866pp DY (204); ATLAS 8 Z pT (253) production; BCDMS and CDHSW DIS

Estimated χ^2 pulls from experiments

(L_2 sensitivity, arXiv:1904.00222, v. 2)



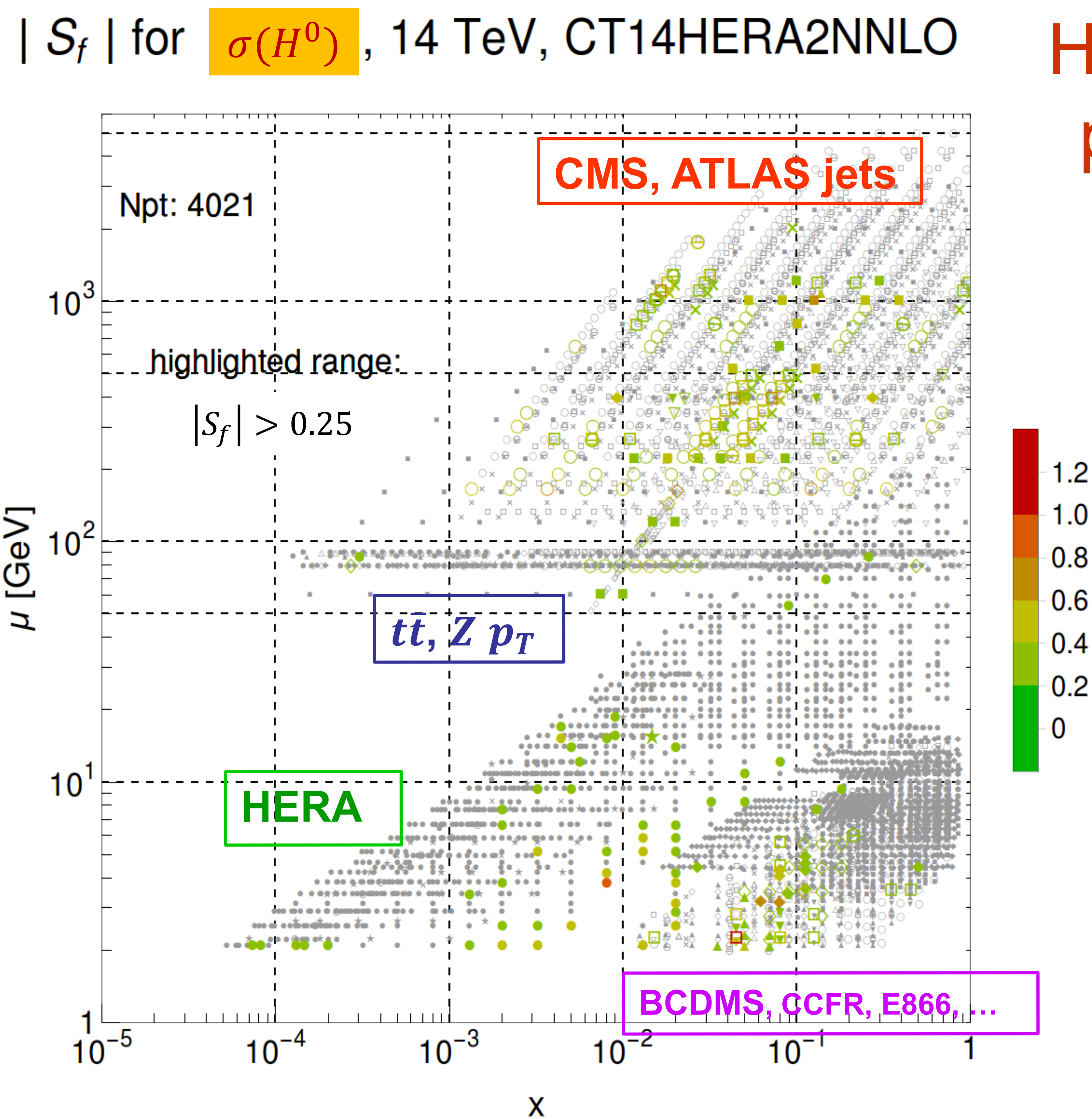
Same for CT18Z NNLO



Constraints from HERA I+II DIS (ID=160) follow a different trend from CT18 NNLO because of the x −dependent QCD scale

PRELIMINARY

Higgs boson production



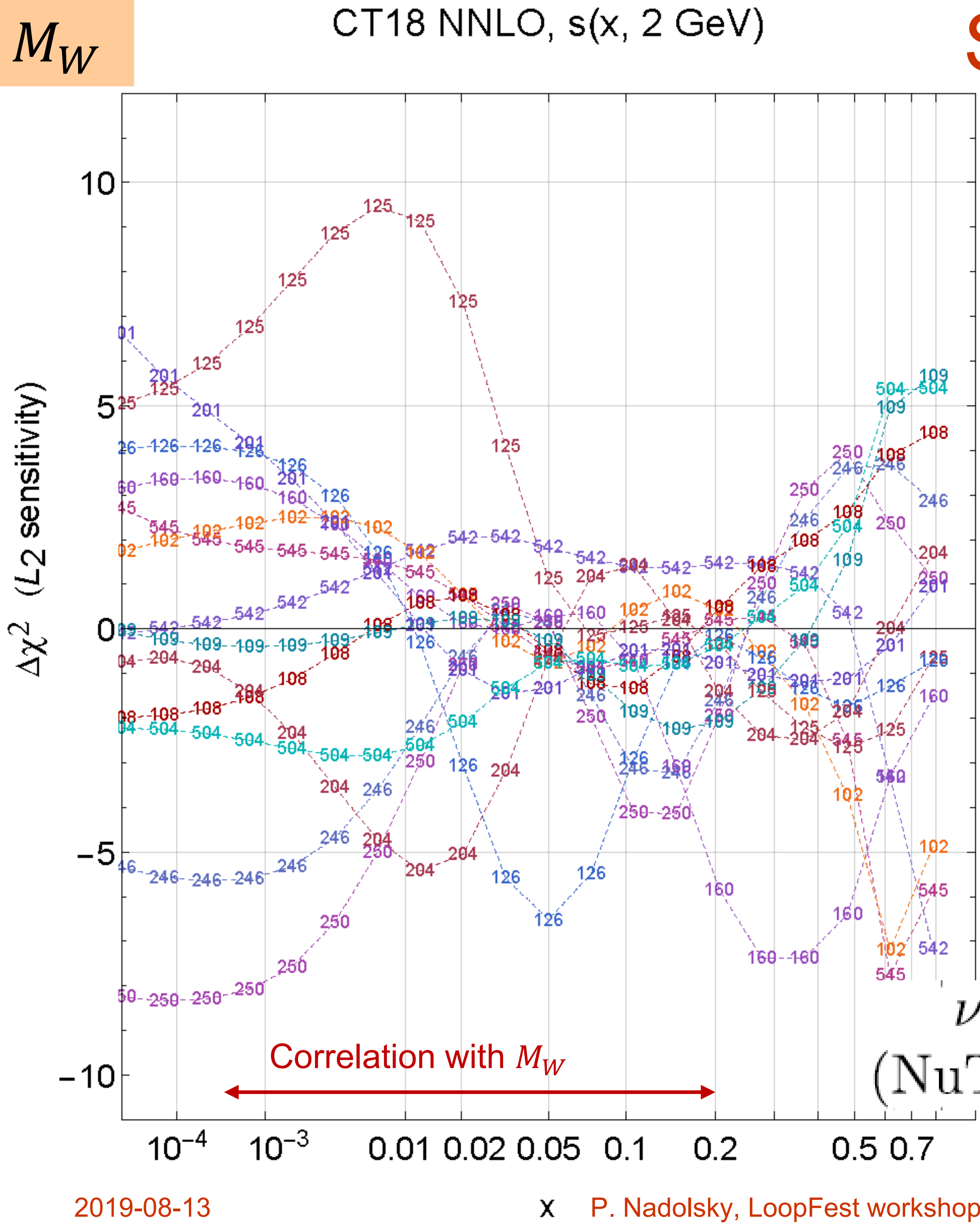
HERA DIS still has the dominant sensitivity!

CMS 8 TeV jets is the next expt. after HERA sensitive to $\sigma_H(14$ TeV); jet scale uncertainty dampens $|S_f|$ for jets

Good correlations C_f with some points in E866, BCDMS, CCFR, CMS WASY, $Z\ p_T$ and $t\bar{t}$ production; but not as many points with high $|S_f|$ in these processes

Strangeness, CT18

M_W

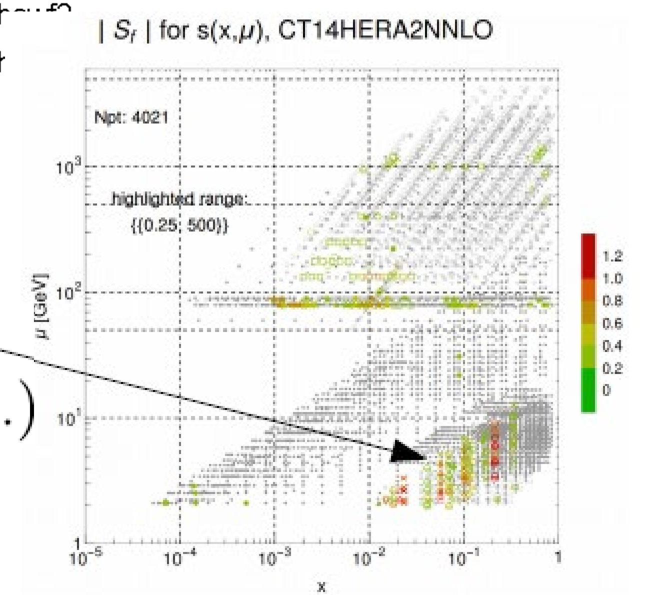


This analysis can be extended to strangeness, which affects M_W extraction through Z boson recoil calibration

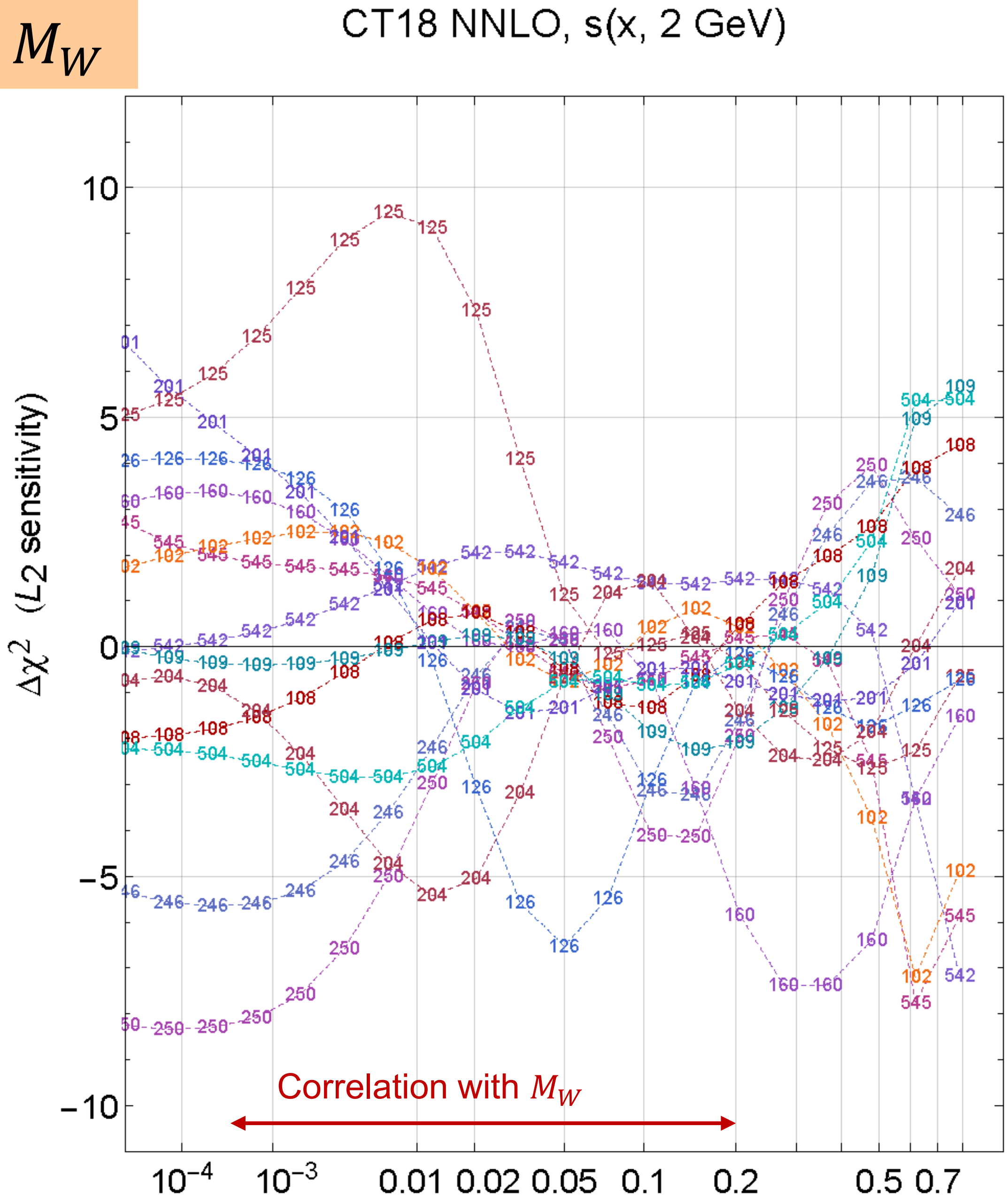
Most sensitive experiments

246	LHCb8Zeer	125	NuTvNbChXN
250	LHCb8WZ	126	CcfrNuChXN
542	CMS7jtR7y6T	201	e605
545	CMS8jtR7T	204	e866ppxf
160	HERAIpII	504	cdf2jtCor2
102	BcdF2dCor		
108	cdhswf2		
109	cdh		

νDIS (NuTeV, ...)

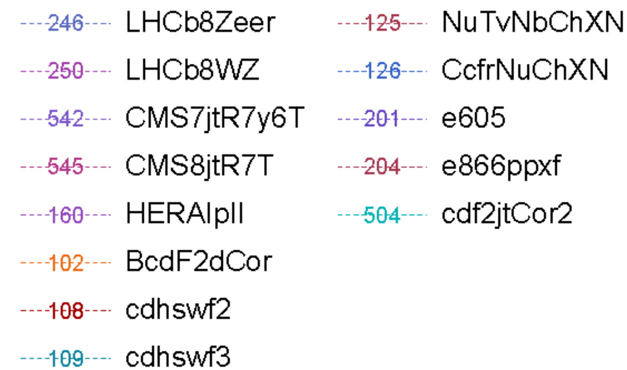


M_W



This analysis can be extended to strangeness, which affects M_W extraction through Z boson recoil calibration

Most sensitive experiments

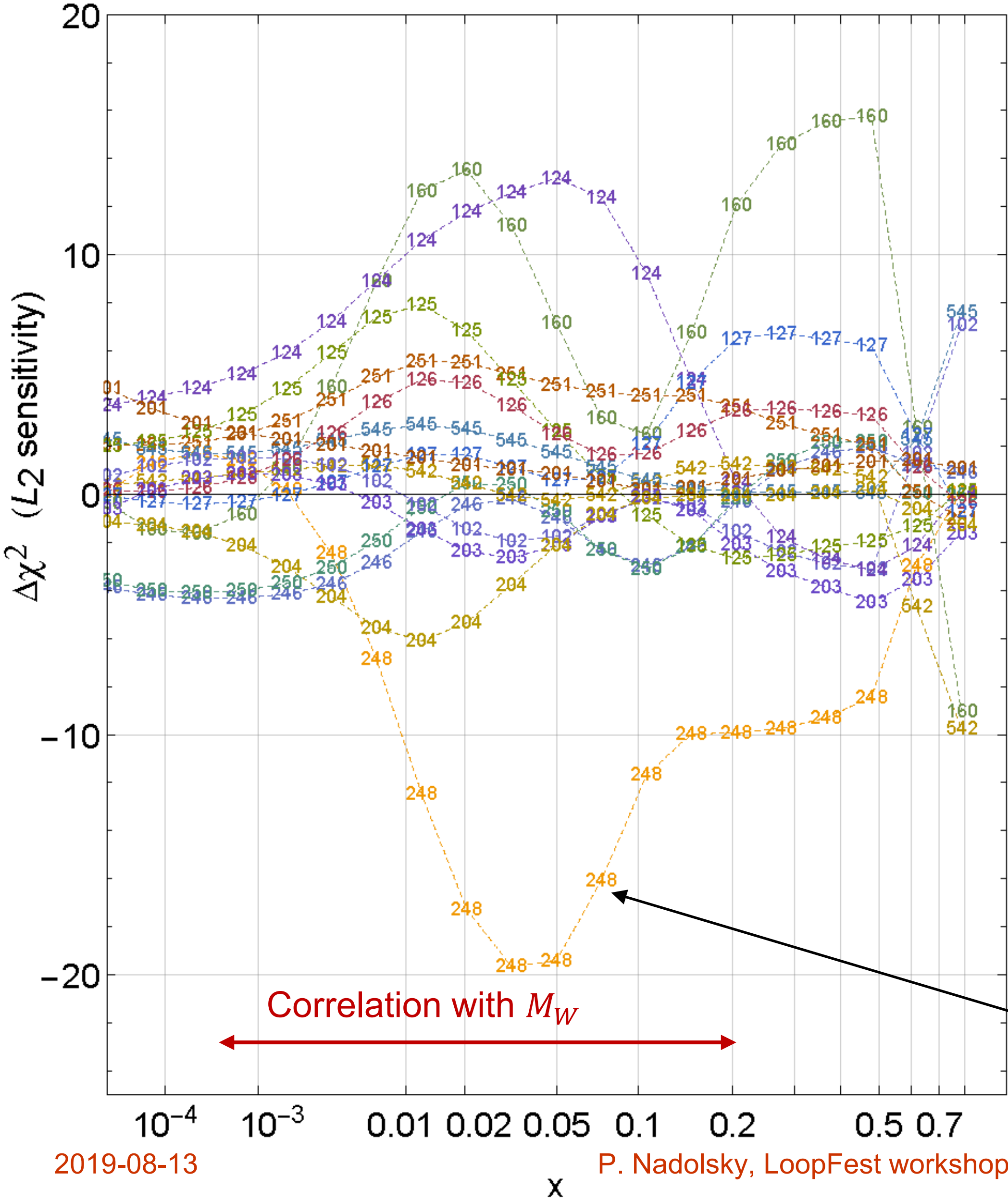


A tension trend between DIS (NuTeV, CCFR, HERA) and Drell-Yan (LHCb W/Z, E866 pp, …) experiments

Strangeness, CT18Z

M_W

CT18Z NNLO, s(x, 2 GeV)



Most sensitive experiments

- 246 LHCb8Zeer
- 248 ATL7ZW.xF
- 250 LHCb8WZ
- 251 ATL8DY
- 542 CMS7jtR7y6T
- 545 CMS8jtR7T
- 160 HERAIpII
- 102 BcdF2dCor
- 124 NuTvNuChXN
- 125 NuTvNbChXN
- 126 CcfrNuChXN
- 127 CcfrNbChXN
- 201 e605
- 203 e866f
- 204 e866ppxf

A tension trend between DIS (HERA I+II, CCFR, NuTeV) and Drell-Yan (ATLAS 7 Z/W, LHCb W/Z, E866 pp, …) experiments

pronounced pull of ATLAS7 Z/W data

Key points, the CT18(Z) global QCD analysis

- modest reduction in the PDF uncertainties compared to CT14
- DIS experiments dominate constraints on PDFs
- LHC Run-1 and 2 processes (jet, W/Z, high-p_T Z, $t\bar{t}$, $W+c$, …production) will provide promising constraints once they are brought into mutual agreement
- NNLO DIS cross sections with an x-dependent factorization scale behave like NNLO+NNLx resummed ones, are incorporated in CT18Z PDFs with the modified small-x gluon and strangeness
- **Future reduction of NNLO PDF uncertainties is not automatic. The goals of the HL-LHC program demand a broad coordinated effort to eliminate tensions between experimental measurements that were identified using several techniques (L_2 sensitivity, LM scans,…)**

Backup

CT18(Z), χ^2 values

ID#	Experimental data set		$N_{pt,n}$	χ^2_n	$\chi^2_n/N_{pt,n}$	S_n
160	HERAI+II 1 fb^{-1}, H1 and ZEUS NC and CC $e^{\pm}p$ reduced cross sec. comb.	[32]	1120	1405.1(1370.2)	1.25(1.22)	5.6(5.0)
101	BCDMS F_2^p	[33]	337	376.3(385.8)	1.12(1.14)	1.5(1.8)
102	BCDMS F_2^d	[34]	250	288.3(289.3)	1.15(1.16)	1.7(1.7)
104	NMC F_2^d/F_2^p	[35]	123	123.3(114.5)	1.00(0.93)	0.061(-0.51)
108	CDHSW† F_2^p	[36]	85	85.1	1.00	0.061
109	CDHSW† F_3^p	[36]	96	83.1	0.865	-0.93
110	CCFR F_2^p	[37]	69	78.3(74.6)	1.13(1.08)	0.81(0.52)
111	CCFR xF_3^p	[38]	86	33.7(29.5)	0.391(0.343)	-5.3(-5.9)
124	NuTeV $\nu\mu\mu$ SIDIS	[39]	38	19.3(29.7)	0.508(0.781)	-2.6(-0.96)
125	NuTeV $\bar{\nu}\mu\mu$ SIDIS	[39]	33	36.5(54.7)	1.11(1.66)	0.50(2.3)
126	CCFR $\nu\mu\mu$ SIDIS	[40]	40	29.2(33.0)	0.729(0.825)	-1.3(-0.76)
127	CCFR $\bar{\nu}\mu\mu$ SIDIS	[40]	38	20.1(20.8)	0.530(0.550)	-2.4(-2.3)
145	H1 σ_r^b	[41]	10	6.8(7.1)	0.682(0.710)	-0.65(-0.57)
147	Combined HERA charm production	[42]	47	58.6(54.7)	1.25(1.16)	1.2(0.82)
169	H1 F_L	[43]	9	17.1(14.5)	1.90(1.61)	1.7(1.2)
201	E605 Drell-Yan process	[44]	119	100.3(98.0)	0.843(0.824)	-1.2(-1.4)
203	E866 Drell-Yan process $\sigma_{pd}/(2\sigma_{pp})$	[45]	15	10.0(12.2)	0.670(0.813)	-0.90(-0.43)
204	E866 Drell-Yan process $Q^3 d^2\sigma_{pp}/(dQdx_F)$	[46]	184	240.2(239.3)	1.31(1.30)	2.7(2.7)
225	CDF Run-1 electron A_{ch}, $p_{T\ell} > 25$ GeV	[47]	11	9.1(9.2)	0.828(0.835)	-0.28(-0.27)
227	CDF Run-2 electron A_{ch}, $p_{T\ell} > 25$ GeV	[48]	11	13.6(13.3)	1.23(1.21)	0.65(0.61)
234	DØ Run-2 muon A_{ch}, $p_{T\ell} > 20$ GeV	[49]	9	9.3(9.2)	1.04(1.02)	0.23(0.19)
260	DØ Run-2 Z rapidity	[50]	28	17.0(19.0)	0.606(0.680)	-1.6(-1.3)
261	CDF Run-2 Z rapidity	[51]	29	49.6(62.6)	1.71(2.16)	2.3(3.4)
266	CMS 7 TeV 4.7 fb^{-1}, muon A_{ch}, $p_{T\ell} > 35$ GeV	[52]	11	8.6(13.5)	0.785(1.23)	-0.40(0.64)
267	CMS 7 TeV 840 pb^{-1}, electron A_{ch}, $p_{T\ell} > 35$ GeV	[53]	11	12.2(16.8)	1.11(1.53)	0.39(1.2)
268	ATLAS 7 TeV 35 pb^{-1} W/Z cross sec., A_{ch}	[54]	41	44.1	1.08	0.41
281	DØ Run-2 9.7 fb^{-1} electron A_{ch}, $p_{T\ell} > 25$ GeV	[55]	13	24.4(20.8)	1.88(1.60)	1.9(1.4)
504	CDF Run-2 inclusive jet production	[56]	72	109.9(107.6)	1.53(1.49)	2.8(2.6)
514	DØ Run-2 inclusive jet production	[57]	110	114.4(115.9)	1.04(1.05)	0.33(0.43)
245	LHCb 7 TeV 1.0 fb^{-1} W/Z forward rapidity cross sec.	[23]	33	50.8(41.4)	1.54(1.25)	2.0(1.0)
246	LHCb 8 TeV 2.0 fb^{-1} $Z \to e^- e^+$ forward rapidity cross. sec.	[24]	17	23.4(21.1)	1.37(1.24)	1.1(0.8)
248	ATLAS‡ 7 TeV 4.6 fb^{-1}, W/Z combined cross sec.	[16]	34	(80.2)	(2.36)	(4.2)
249	CMS 8 TeV 18.8 fb^{-1} W cross sec. and A_{ch}	[22]	11	10.8(11.0)	0.98(0.99)	0.09(0.1)
250	LHCb 8 TeV 2.0fb^{-1} W/Z cross sec.	[25]	34	70.2(58.3)	2.07(1.71)	3.5(2.5)
251	ATLAS 8 TeV 20.3 fb^{-1} single diff. high-mass cross sec.	[58]	12	18.7(?)	1.56(?)	1.3(?)
253	ATLAS 8 TeV 20.3 fb^{-1}, Z p_T cross sec.	[27]	27	31.4(29.4)	1.16(1.09)	0.7(0.4)
542	CMS 7 TeV 5 fb^{-1}, single incl. jet cross sec., $R = 0.7$ (extended in y)	[59]	158	208.7(204.24)	1.32(1.29)	2.6(2.4)
544	ATLAS 7 TeV 4.5 fb^{-1}, single incl. jet cross sec., $R = 0.6$	[60]	140	204.6(205.2)	1.46(1.47)	3.4(3.5)
545	CMS 8 TeV 19.7 fb^{-1}, single incl. jet cross sec., $R = 0.7$, (extended in y)	[61]	185	249.4(229.1)	1.35(1.24)	3.1(2.2)
573	CMS 8 TeV 19.7 fb^{-1}, $t\bar{t}$ norm. double-diff. top p_T & y cross sec.	[62]	16	30.4(26.3)	1.90(1.64)	2.1(1.6)
580	ATLAS 8 TeV 20.3^{-1}, $t\bar{t}$ p_T^t and $m_{t\bar{t}}$ abs. spectrum	[63]	15	15.4(20.5)	1.03(1.36)	0.2(1.0)

Data sets employed in the CT18(Z) analysis. The numbers in round brackets are for the CT18Z fit. $N_{pt,n}$, χ^2 are the number of points and value of χ^2 for the n-th experiment at the global minimum. S_n is the effective Gaussian parameter quantifying agreement with each experiment.

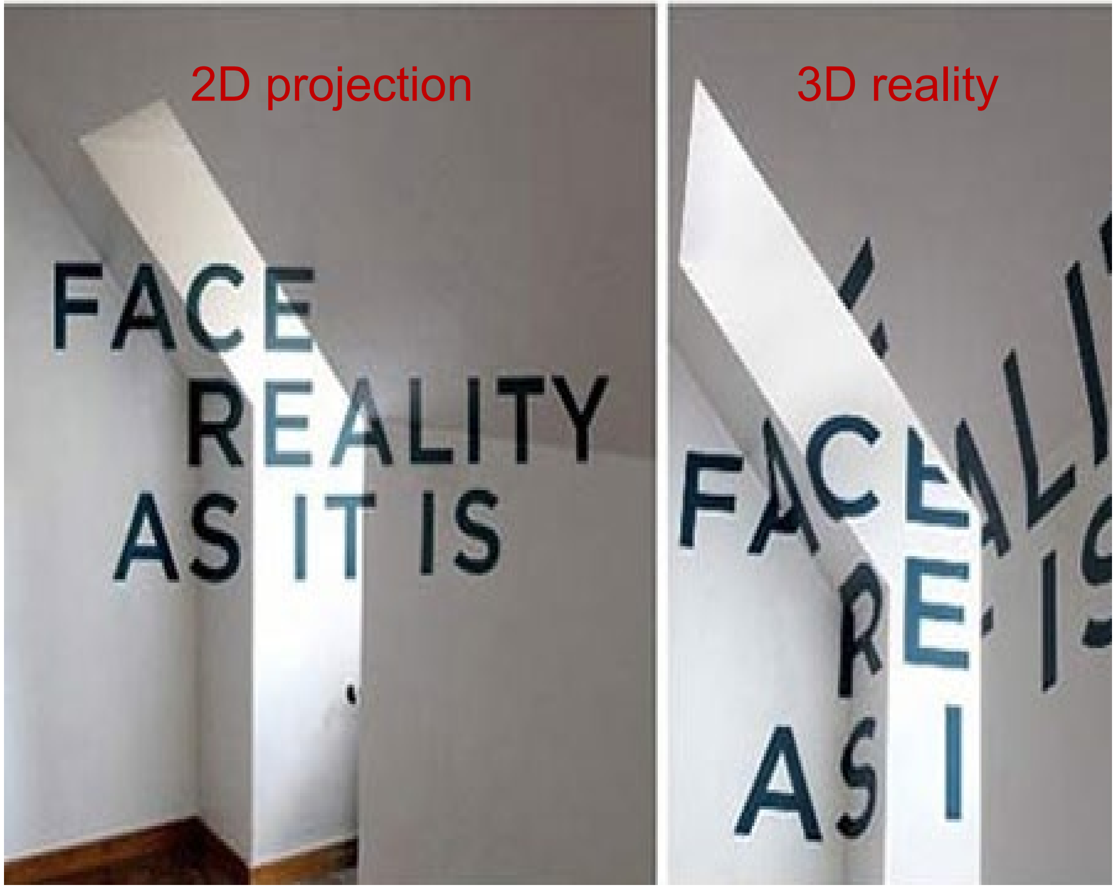
CT14: parametrization forms

- CT14 relaxes restrictions on several PDF combinations that were enforced in CT10. [These combinations were not constrained by the pre-LHC data.]
  - The assumptions $\frac{\bar{d}(x,Q_0)}{\bar{u}(x,Q_0)} \to 1$, $u_v(x,Q_0) \sim d_v(x,Q_0) \propto x^{A_{1v}}$ with $A_{1v} \approx -\frac{1}{2}$ at $x < 10^{-3}$ are relaxed once LHC W/Z data are included
  - CT14 parametrization for $s(x,Q)$ includes extra parameters
- Candidate CT14 fits have 30-35 free parameters
- In general, $\mathrm{f_a}\,(\mathrm{x}, \mathrm{Q_0}) = \mathrm{A}x^{a_1}(1-\mathrm{x})^{\mathrm{a_2}}\mathrm{P_a}(\mathrm{x})$
- CT10 assumed $P_a(x) = \exp(a_0 + a_3\sqrt{x} + a_4 x + a_5\, x^2)$
  - exponential form conveniently enforces positive definite behavior
  - but power law behaviors from a_1 and a_2 may not dominate
- In CT14, $\mathrm{P_a(x) = G_a(x)F_a(z)}$, where $\mathrm{G_a(x)}$ is a smooth factor
  - $\mathrm{z} = 1 - 1(1-\sqrt{\mathrm{x}})^{\mathrm{a_3}}$ preserves desired Regge-like behavior at low x and high x (with $\mathrm{a_3}$>0)
- Express $F_a(z)$ as a linear combination of Bernstein polynomials:

$$z^4, 4z^3(1-z), 6z^2(1-z)^2\,, 4z(1-z)^3, (1-z)^4$$

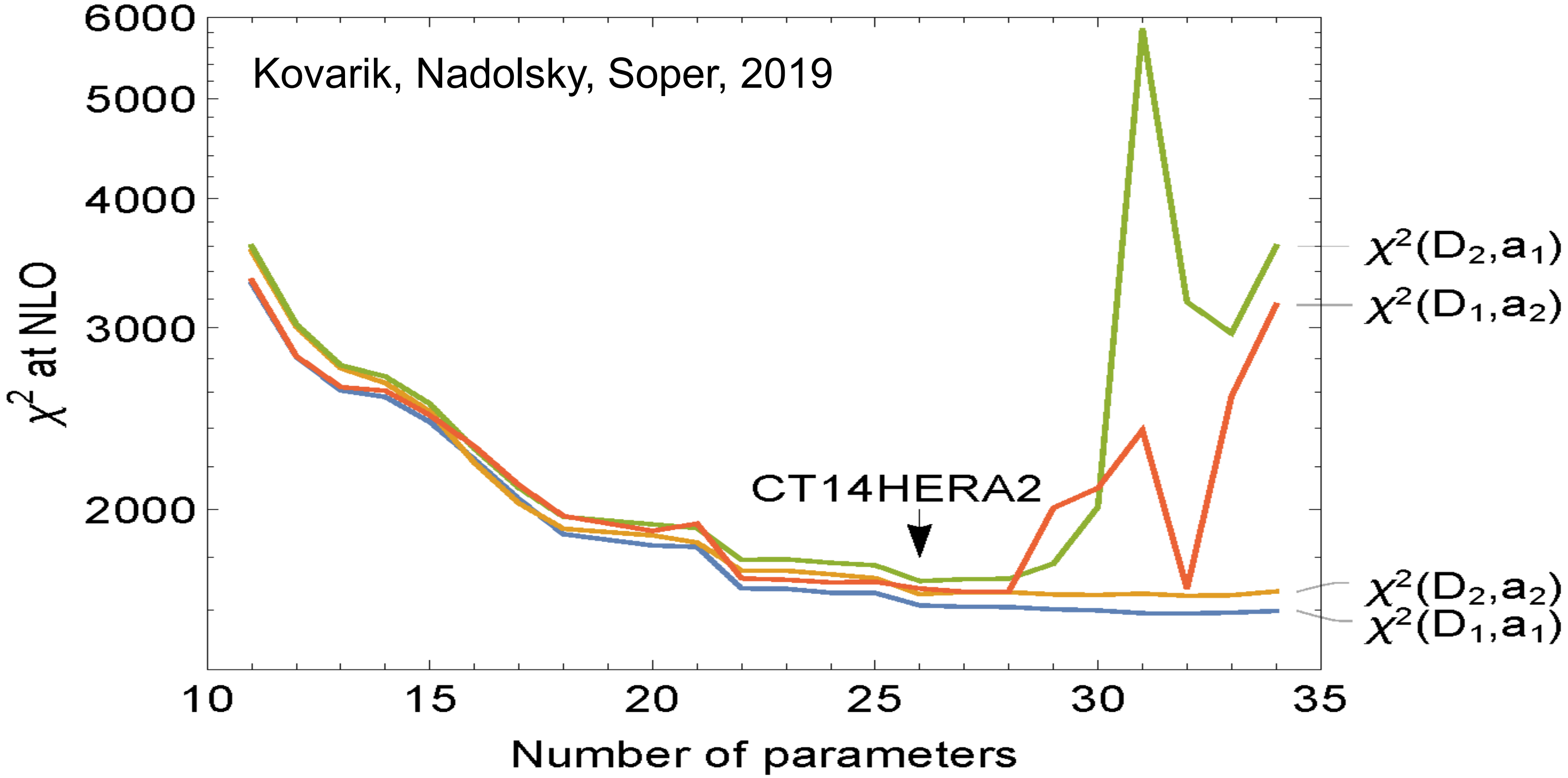
  - each basis polynomial has a single peak, with peaks at different values of z; reduces correlations among parameters

If too few parameters



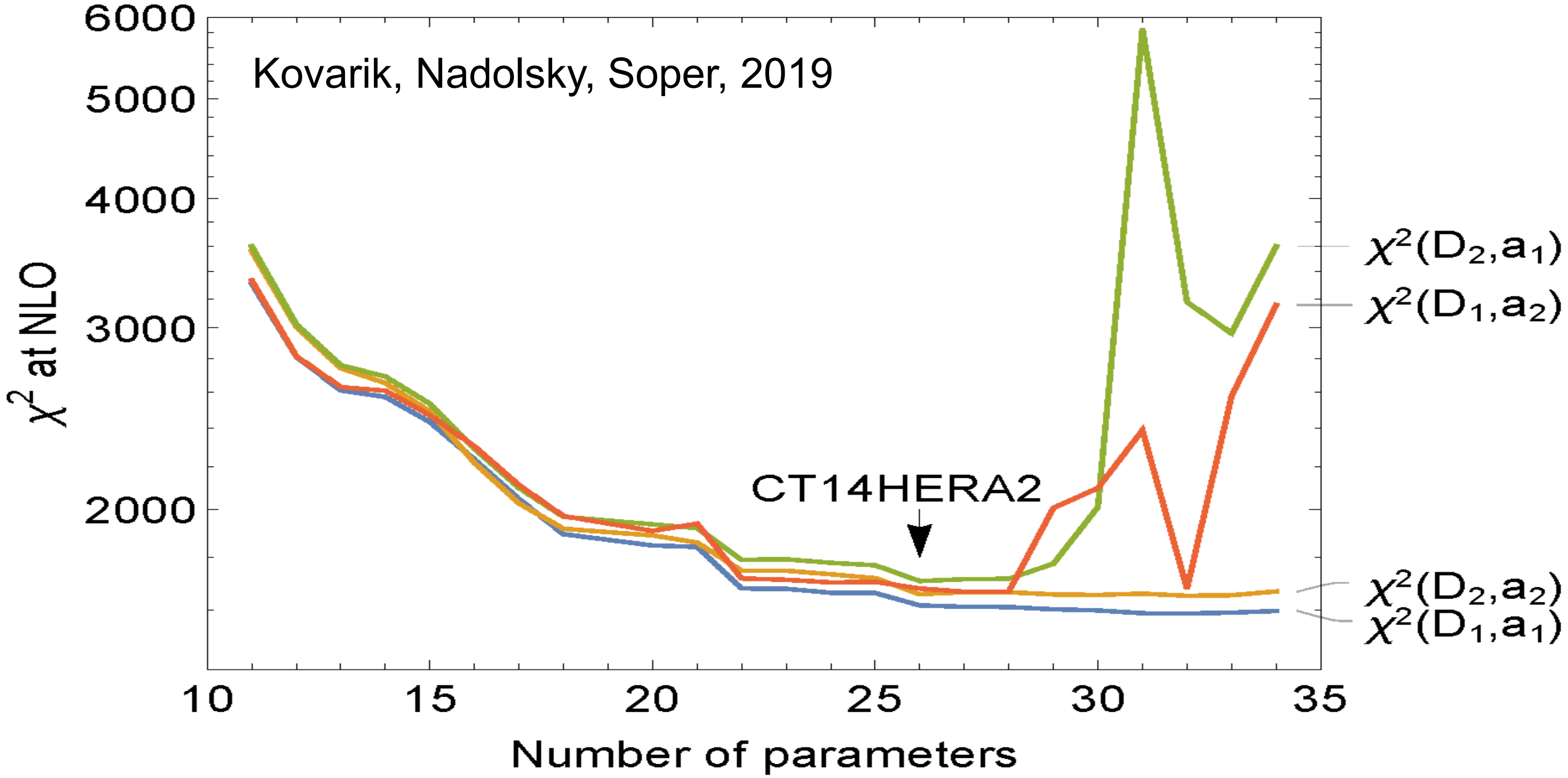
The solution can be consistent and false

If too many parameters



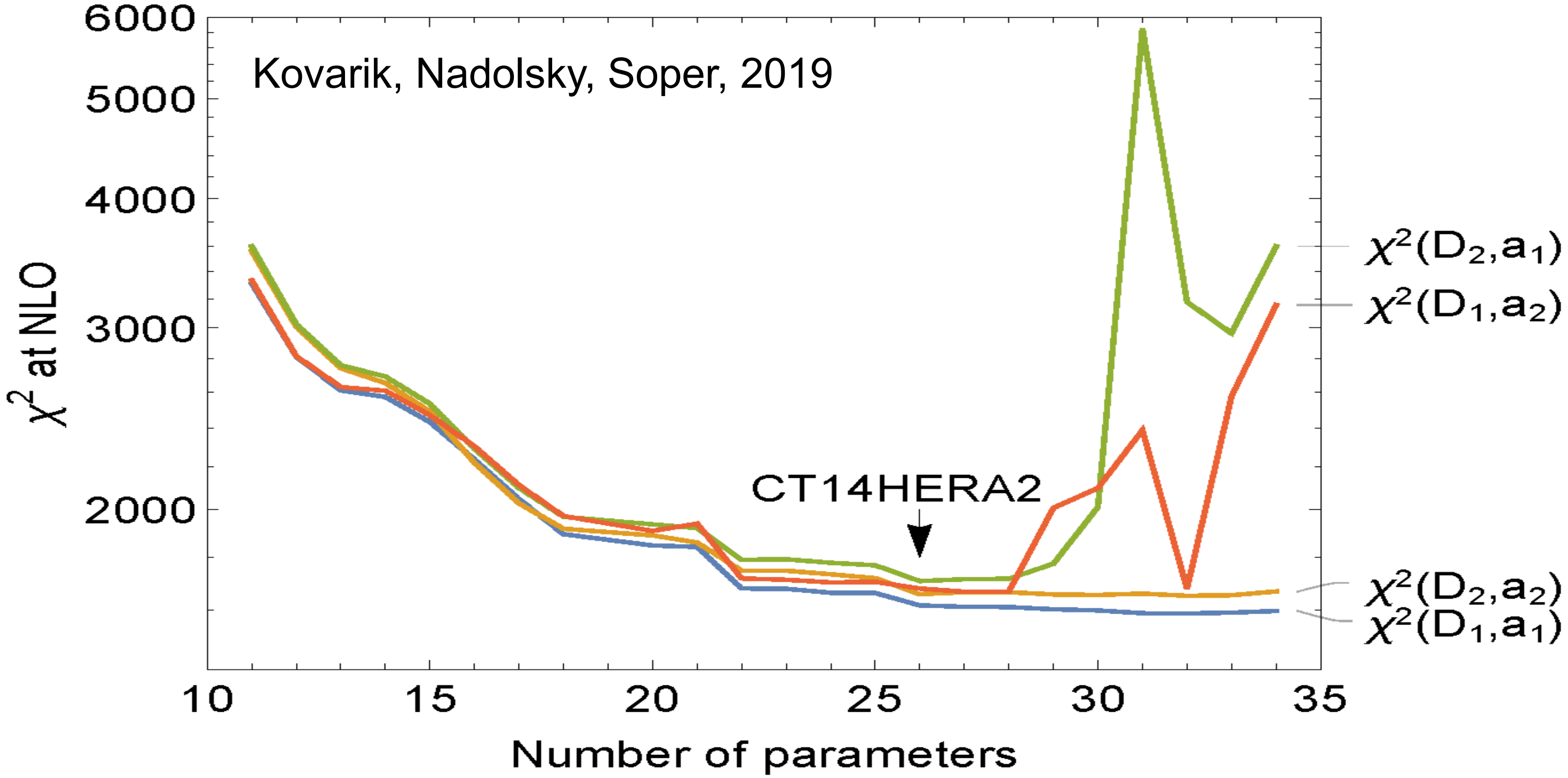
- Randomly split the CT14HERA data set into two halves, D_1 and D_2
- Find parameter vectors a_1 and a_2 from the best fits for D_1 and D_2, respectively

If too many parameters



- **Fitted samples:** $\chi^2(D_1, a_1)$ and $\chi^2(D_2, a_2)$ uniformly decrease with the number of parameters
- **Control samples:** $\chi^2(D_2, a_1)$ and $\chi^2(D_1, a_2)$ fluctuate when the number of parameters is larger than about 30

If too many parameters



$\lesssim 30$ parameters (26 in CT14HERA2) is optimal for describing the CT14HERA2 data set

A shifted residual r_i

$r_i(\vec{a}) = \frac{T_i(\vec{a}) - D_i^{sh}(\vec{a})}{s_i}$ are N_{pt} **shifted residuals** for point i, PDF parameters $\vec{a}$

$\bar{\lambda}_\alpha(\vec{a})$ are N_λ **optimized nuisance parameters** (dependent on $\vec{a}$)

The $\chi^2(\vec{a})$ for experiment E is

$$\chi^2(\vec{a}) = \sum_{i=1}^{N_{pt}} r_i^2\,(\vec{a}) + \sum_{\alpha=1}^{N_\lambda} \bar{\lambda}_\alpha^2\,(\vec{a}) \approx \sum_{i=1}^{N_{pt}} r_i^2\,(\vec{a})$$

$T_i(\vec{a})$ is the theory prediction for PDF parameters $\vec{a}$

D_i^{sh} is the data value **including the optimal systematic shift**

$$D_i^{sh}(\vec{a}) = D_i - \sum_{\alpha=1}^{N_\lambda} \beta_{i\alpha}\bar{\lambda}_\alpha(\vec{a})$$

s_i is the uncorrelated error

$r_i(\vec{a})$ and $\bar{\lambda}_\alpha(\vec{a})$ are tabulated or extracted from the cov. matrix

Finding shifted residuals r_i from the covariance matrix

The CTEQ-TEA fit returns tables of $r_i(\vec{a})$ and $\bar{\lambda}_\alpha(\vec{a})$ for every i and α

Alternatively, they can be found from the covariance matrix:

$$r_i(\vec{a}) = s_i \sum_{j=1}^{N_{pt}} (\mathrm{cov}^{-1})_{ij} \, (T_j(\vec{a}) - D_j) \,, \qquad \bar{\lambda}_\alpha(\vec{a}) = \sum_{i,j=1}^{N_{pt}} (\mathrm{cov}^{-1})_{ij} \frac{\beta_{i\alpha}}{s_i} \frac{(T_j(\vec{a}) - D_j)}{s_j}$$

Vectors of data residuals

For every data point i, construct a vector of residuals $r_i(\vec{a}_k^{\pm})$ for 2N Hessian eigenvectors. $\mathrm{k} = 1, \dots, N$, with $N = 28$ for CT14 NNLO:

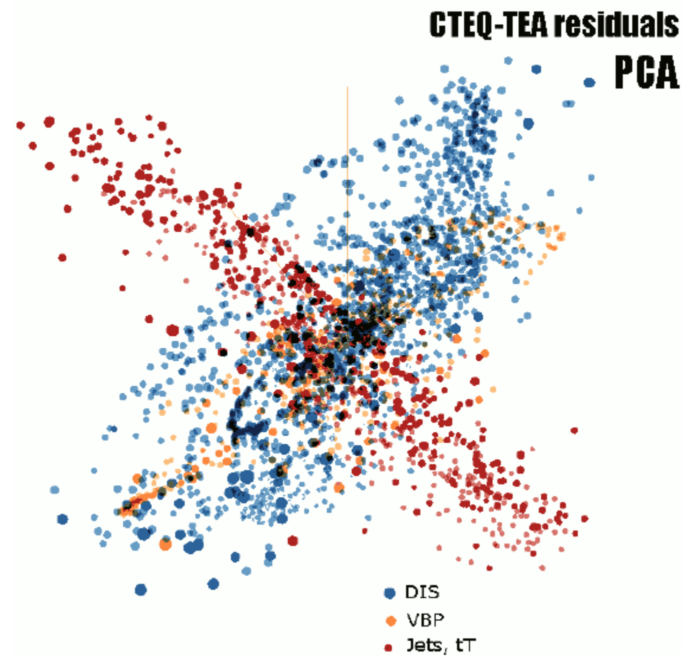
$$\vec{\delta}_i = \{\delta_{i,1}^{+}, \delta_{i,1}^{-}, \dots, \delta_{i,N}^{+}, \delta_{i,N}^{-}\} \qquad [\mathrm{N} = 28]$$
$$\delta_{i,k}^{\pm} \equiv \left(r_i(\vec{a}_k^{\pm}) - r_i(\vec{a}_0)\right) / \langle r_0 \rangle_E$$

-- a 56-dim vector normalized to $\langle r_0 \rangle_E$, the root-mean-squared residual for the experiment E for the central fit $\vec{a}_0$

$$\langle r_0 \rangle_E \equiv \sqrt{\frac{1}{N_{pt}} \sum_{i=1}^{N_{pt}} r_i^2(\vec{a}_0)} \approx \sqrt{\frac{\chi_E^2\,(\vec{a}_0)}{N_{pt}}}$$

$\langle r_0 \rangle_E \approx 1$ in a good fit to E

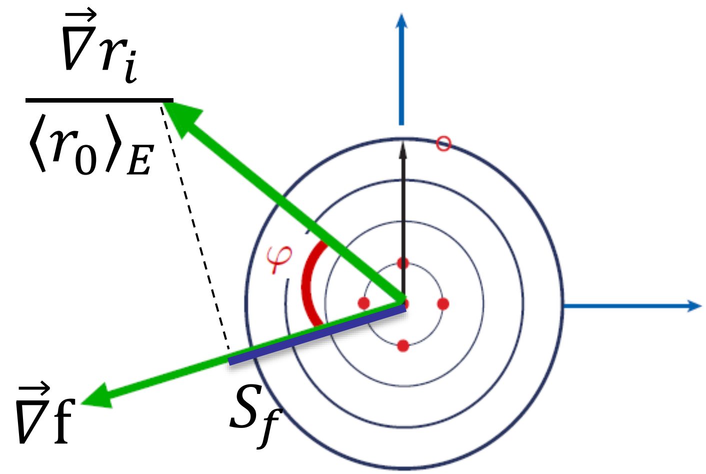
r_i is defined in the backup



The TensorFlow Embedding Projector (http://projector.tensorflow.org) represents CT14HERA2 $\vec{\delta}_i$ vectors by their 10 principal components indicated by scatter points. A sample 3-dim. projection of the 56-dim. manifold is shown above. A symmetric 28-dim. representation can be alternatively used.

Correlation C_f and sensitivity S_f

The relation of data point i on the PDF dependence of f can be estimated by:



- $C_f \equiv \mathrm{Corr}[\rho_i(\vec{a})), f(\vec{a})] = cos\varphi$

$\vec{\rho}_i \equiv \vec{\nabla} r_i / \langle r_0 \rangle_E$ -- gradient of r_i normalized to the r.m.s. average residual in expt E;

$$\left(\vec{\nabla} r_i\right)_k = \left(r_i(\vec{a}_k^+) - r_i(\vec{a}_k^-)\right)/2$$

C_f is **independent** of the experimental and PDF uncertainties. In the figures, take $|C_f| \gtrsim 0.7$ to indicate a large correlation.

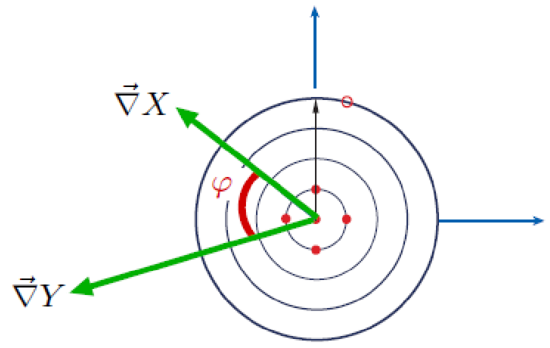
- $S_f \equiv |\vec{\rho}_i| cos\varphi = C_f \frac{\Delta r_i}{\langle r_0 \rangle_E}$ -- projection of $\vec{\rho}_i(\vec{a})$ on $\vec{\nabla} f$

S_f is proportional to $\cos\varphi$ and the ratio of the PDF uncertainty to the experimental uncertainty. We can sum $|S_f|$.
In the figures, take $|S_f| > 0.25$ to be significant.

L_2 sensitivity, definition

Tolerance hypersphere in the PDF space

2-dim (i,j) rendition of N-dim (22) PDF parameter space

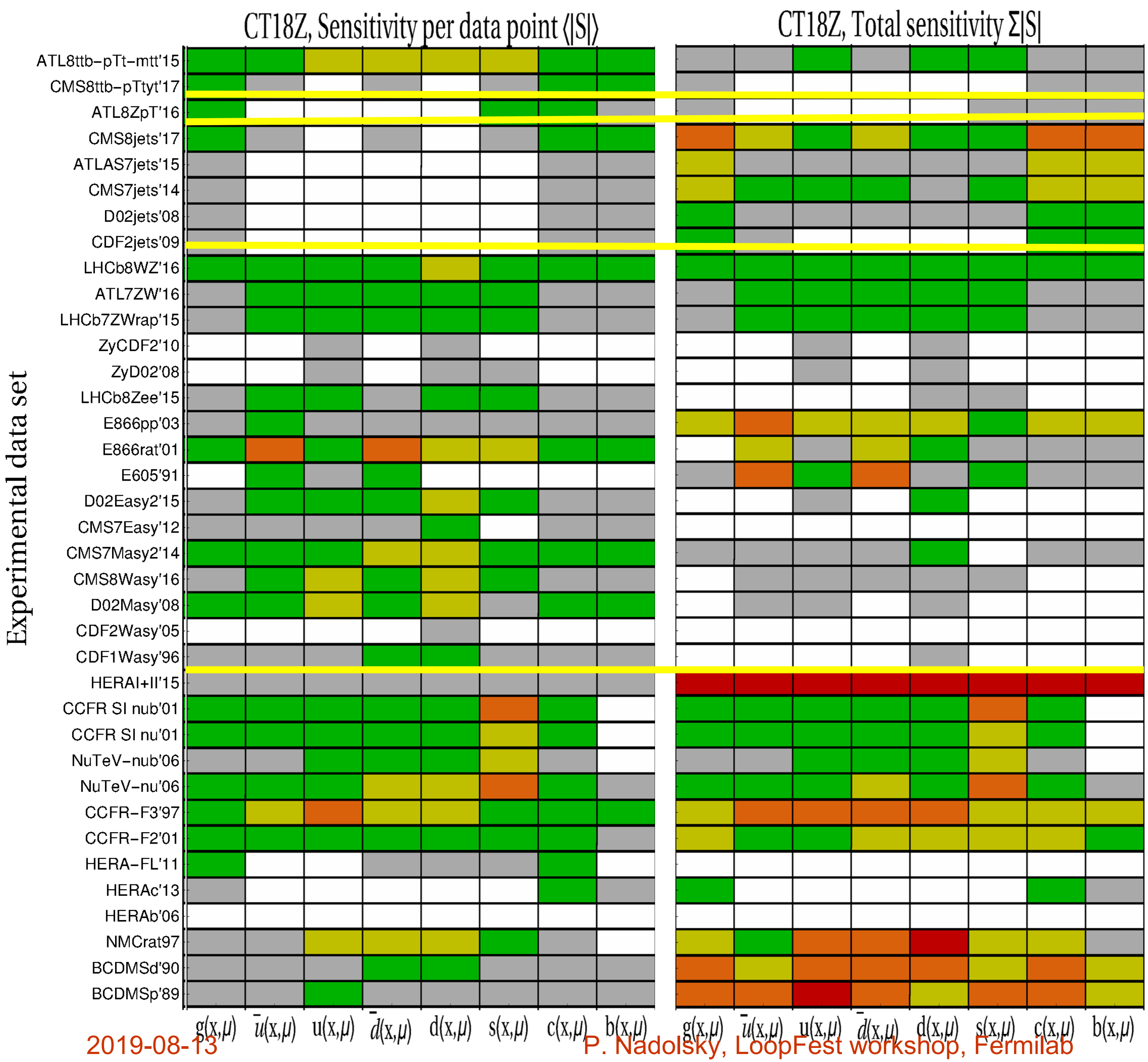


(b)
Orthonormal eigenvector basis

L_2 sensitivity. Take $X = f_a(x_i, Q_i)$ or $\sigma(f)$; $Y = \chi^2_E$ for experiment E. Find $\Delta Y(\vec{z}_{m,X})$ for the displacement $|\vec{z}_{m,X}| = 1$ along the direction $\vec{\nabla} X / \left|\vec{\nabla} X\right|$ (corresponding to $\Delta\chi^2_{tot} = T^2$ and $X(\vec{z}) = X(0) + \Delta X$):

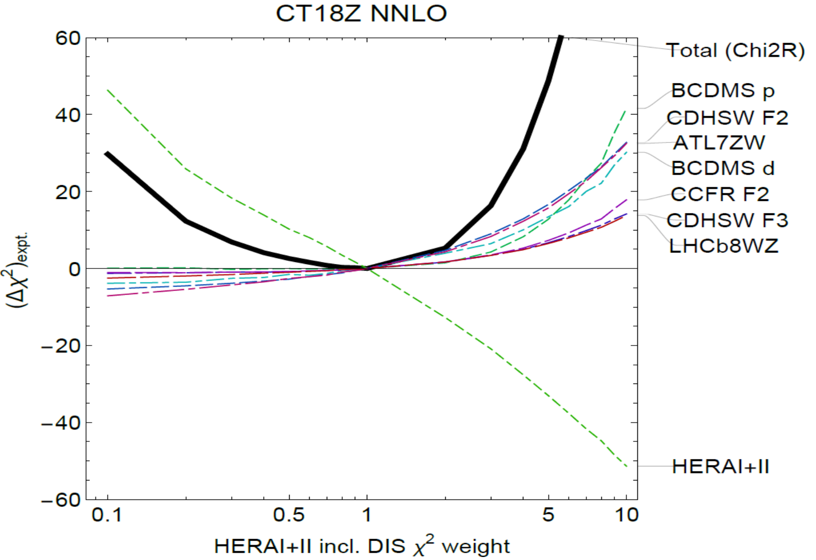
$$S_{f,L_2} \equiv \Delta Y(\vec{z}_{m,X}) = \vec{\nabla} Y \cdot \vec{z}_{m,X} = \vec{\nabla} Y \cdot \frac{\vec{\nabla} X}{|\vec{\nabla} X|} = \Delta Y \cos\varphi.$$

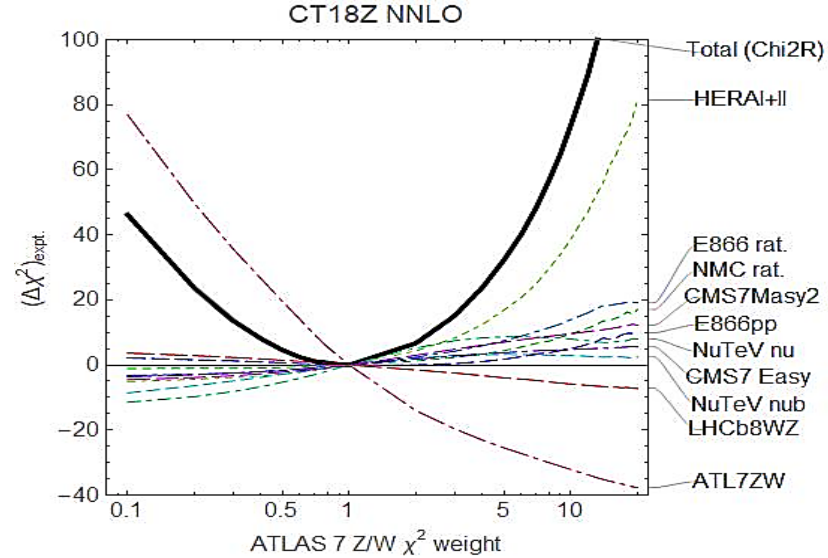
Sensitivity of hadronic experiments to PDFs

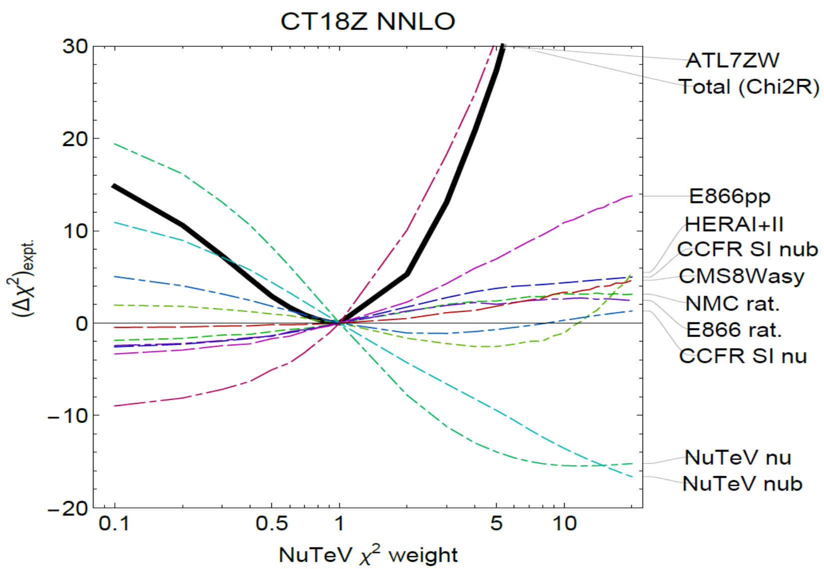


For the CT18Z NNLO data set

LM scans on χ^2 weights of HERA I+II, ATLAS 7 Z/W, and NuTeV data

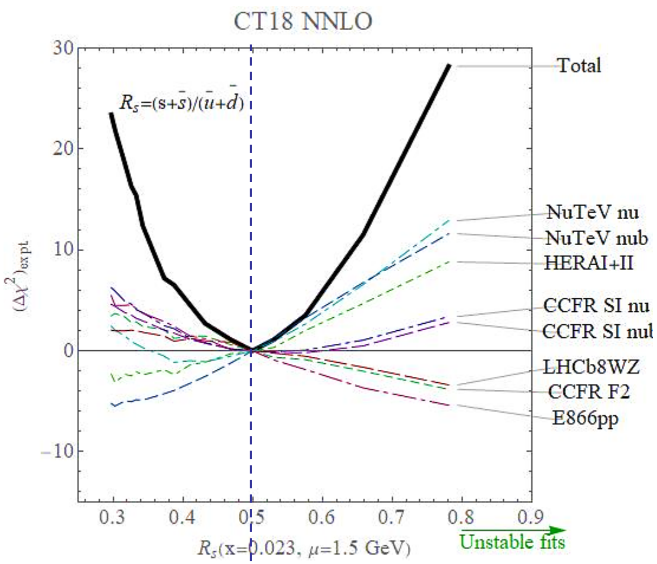


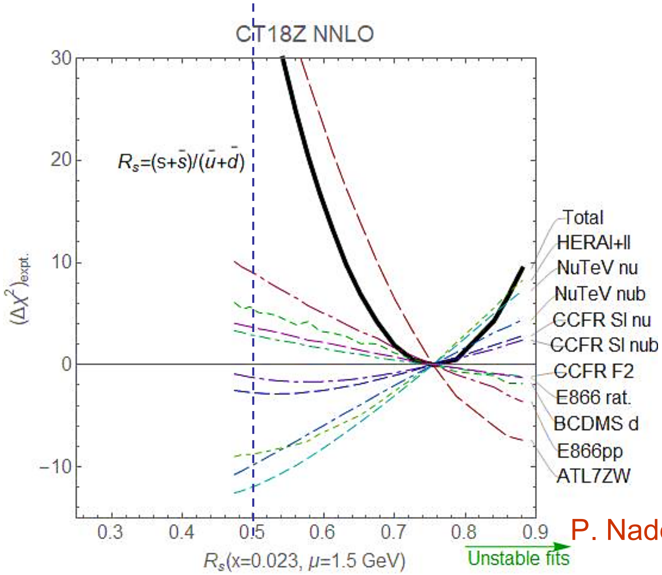




Fits with varied weights and LM scans reveal a disagreement between important DIS [primarily HERA, CCFR, NuTeV,…] and DY [primarily ATL7ZW, E866, LHCb8WZ,…] experiments. This is more pronounced for large$-x$ gluon as well as strangeness.

Lagrange Multiplier scan: $R_s(x = 0.023, \mu = 1.5\ GeV)$

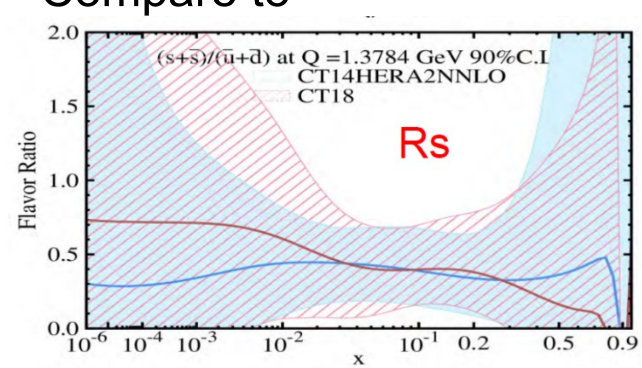




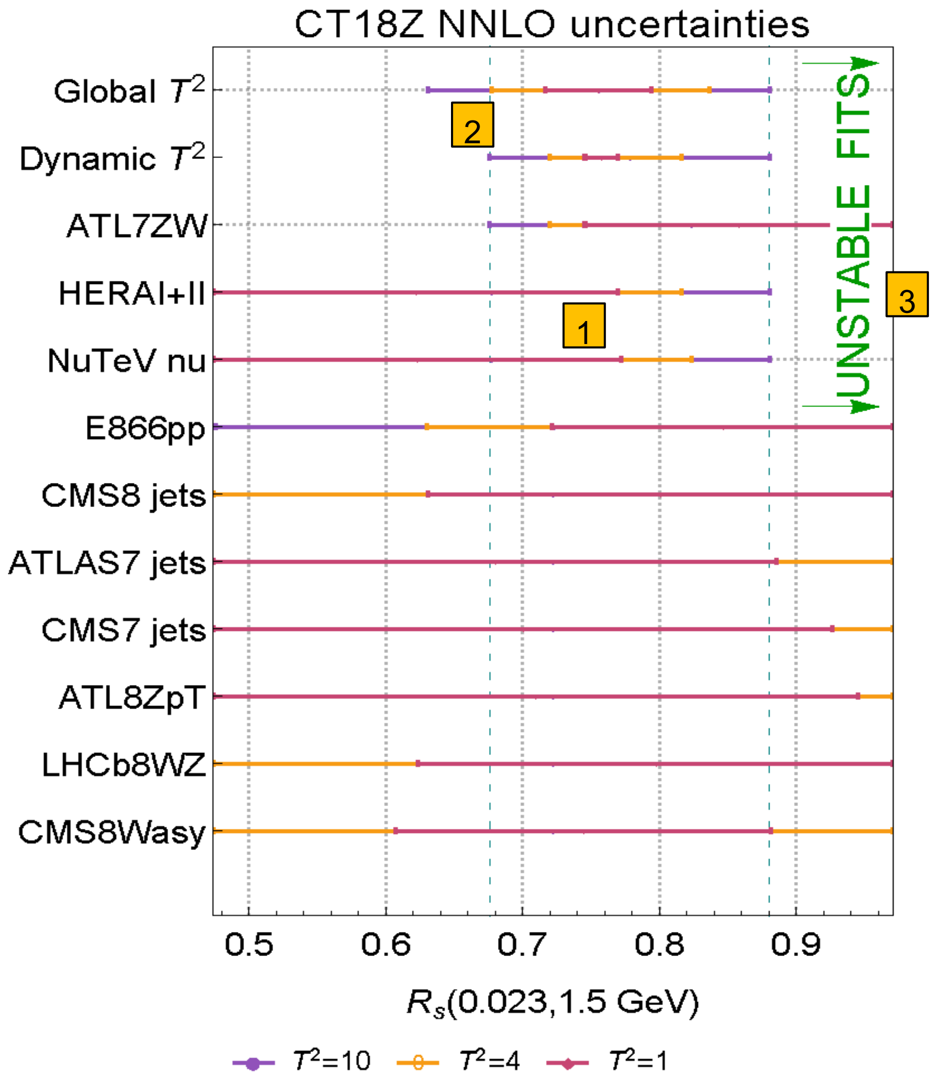
The CT18Z strangeness is increased primarily as a result of including the ATLAS 7 TeV W/Z production data (not in CT18), as well as because of using the DIS saturation scale and $m_c^{pole} = 1.4$ GeV

In either CT18 or CT18Z fit, observe instability in the fits for $R_s > 1$ at $x = 0.01 - 0.1$

Compare to



Effect on PDF uncertainties



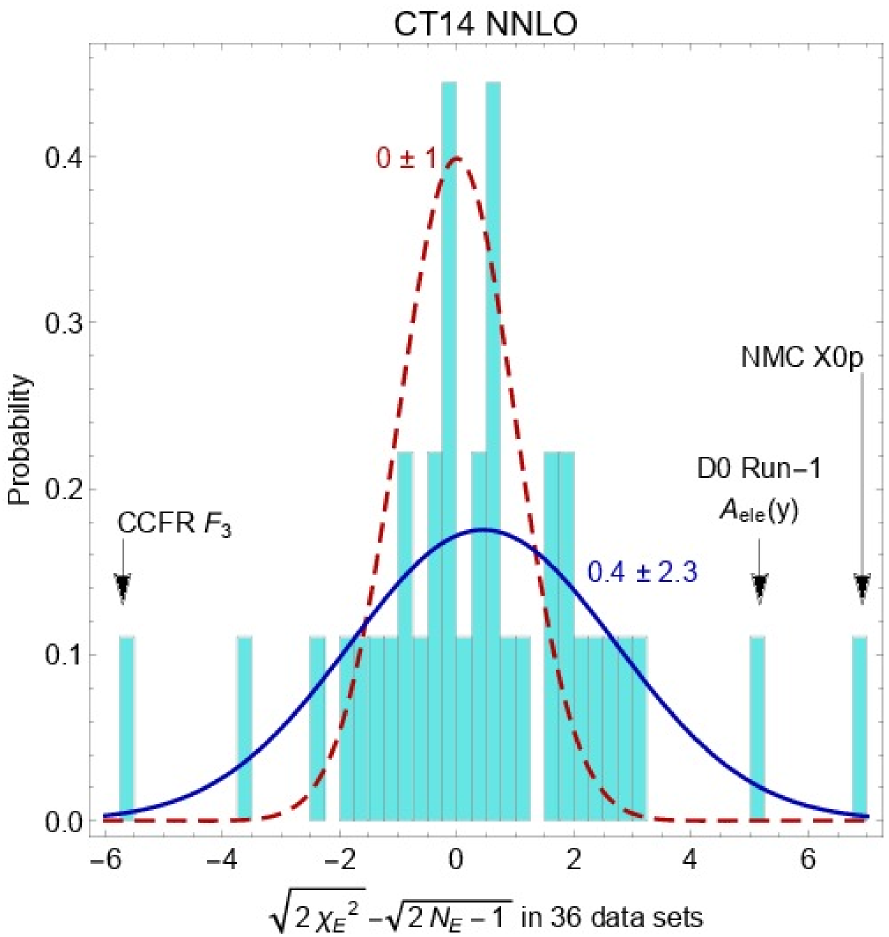
The LM scan reveals details not captured by other methods

1 **Nonlinearities:** the error bands for tolerance $T^2 = 1,4,10$ may not scale according to the Gaussian distribution

2 **Tensions:** in the affected direction(s), the global tolerance and **especially dynamic tolerance** may underestimate the true PDF error.

3 χ^2 **instability:** Neither the "global T^2" nor "dynamic T^2" reflect instability of fits at $R_s > 0.9$

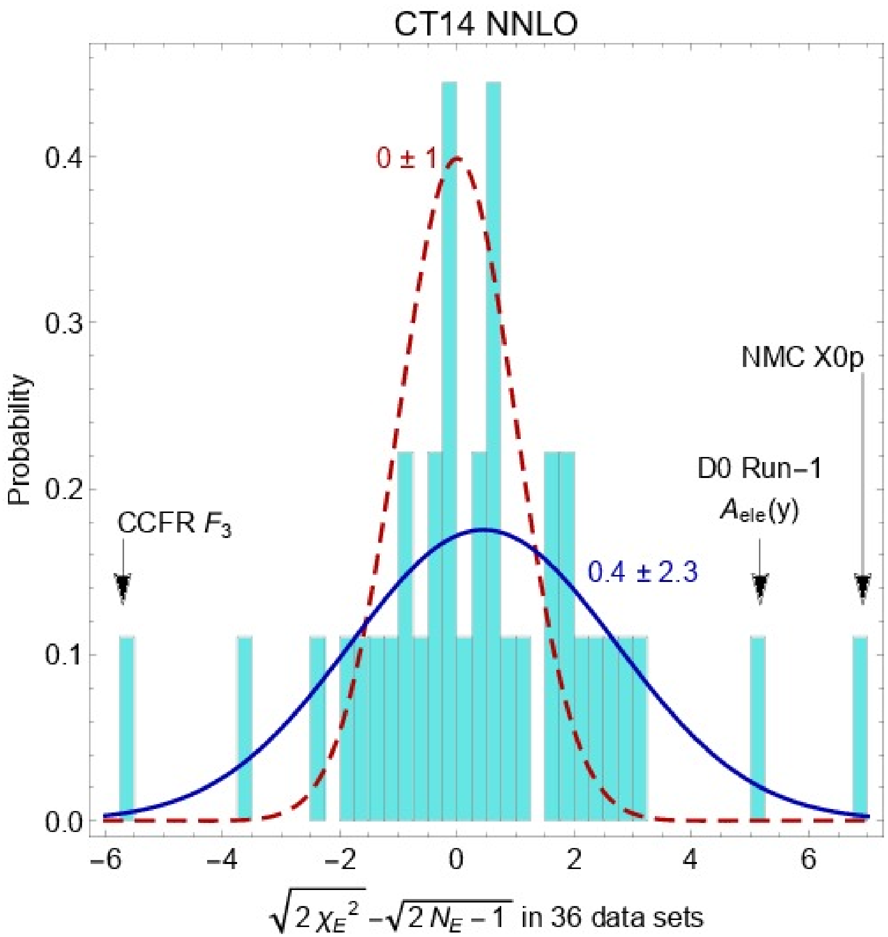
Effective Gaussian variables



Define $S_n(\chi^2, N_{pt})$ for experiment n so that, in a perfect fit, it would approximately obey the standard normal distribution $\mathcal{N}(0,1)$ (mean=0, half-width=1) independently of $N_{pt,n}$

[H.-L. Lai et al., arXiv:1007.2241; S.Dulat et al., arXiv:1309.0025; K. Kovarik, P.N., D. Soper, arXiv:1905.06957]

Effective Gaussian variables



$$S_n(\chi^2, N_{pt}) \equiv \sqrt{2\chi^2} - \sqrt{2N_{pt}-1}$$

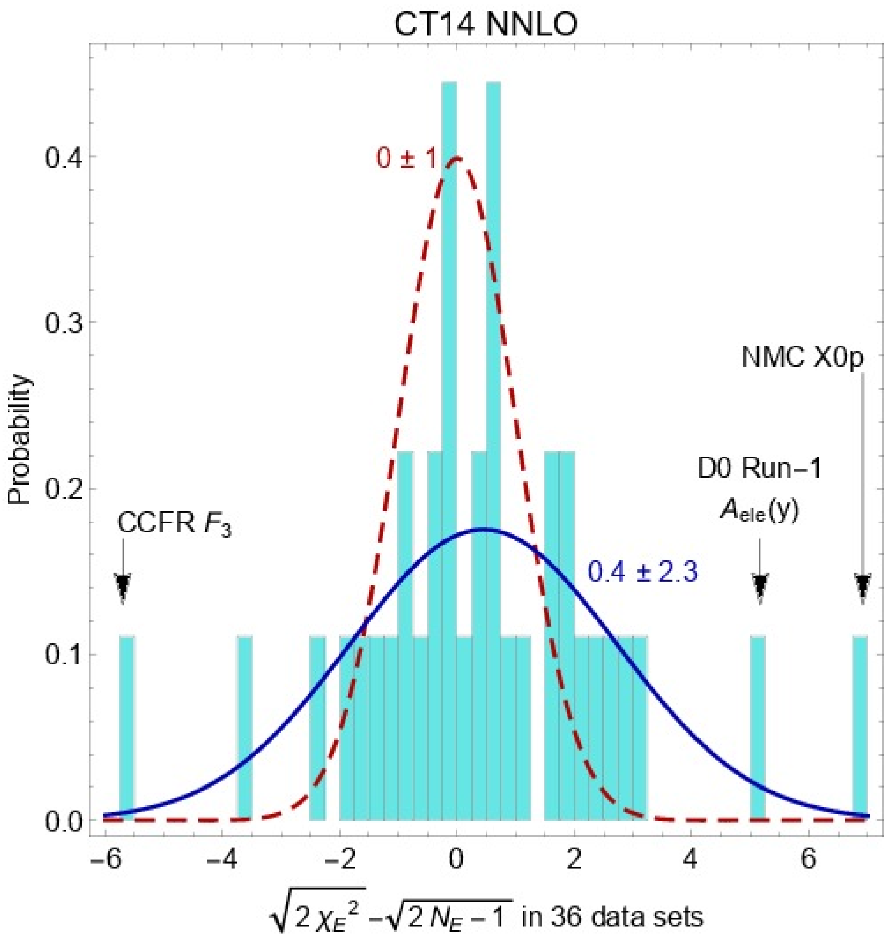
$S_n(\chi_n^2, N_{pt,n})$ are Gaussian distributed with mean 0 and variance 1 for $N_{pt,n} \geq 10$

[R.A.Fisher, 1925]

Even more accurate (χ^2, N_{pt}):
T.Lewis, 1988

An empirical S_n distribution can be compared to $\mathcal{N}(0,1)$ visually or using a statistical (Anderson-Darling, Kolmogorov-Smirnov, …) test

Effective Gaussian variables



Some S_n are too big or too small in a global fit

CT14 NNLO:

- $S_n > 4$ for NMC DIS ep cross section and D0 Run-1 electron charge asymmetry
- These data sets are eliminated in CT14HERA2/CT18 fits
- The rest of CT14 experiments are reasonably consistent; $S_n \sim \mathcal{N}(0.3, 1.6)$
- Qualitatively similar S_n distributions for MMHT, NNPDF3.X

Effective Gaussian variables