PARTON DISTRIBUTIONS
FOR THE XXI CENTURY

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

Istituto Nazionale di Fisica Nucleare

UNIVERSITÄT FREIBURG

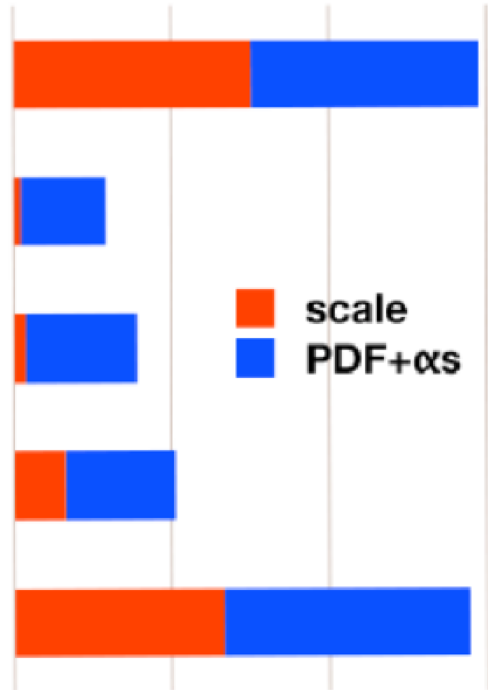
JANUARY 10, 2018

PROLOGUE

PAST (NOT SO LONG AGO)

HIGGS PRODUCTION

		σ (8 TeV)	uncertainty
NNLL QCD +NLO EW	gg→H	19.5 pb	14.7%
	VBF	1.56 pb	2.9%
NNLO QCD +NLO EW	WH	0.70 pb	3.9%
	ZH	0.39 pb	5.1%
NLO QCD	ttH	0.13 pb	14.4%



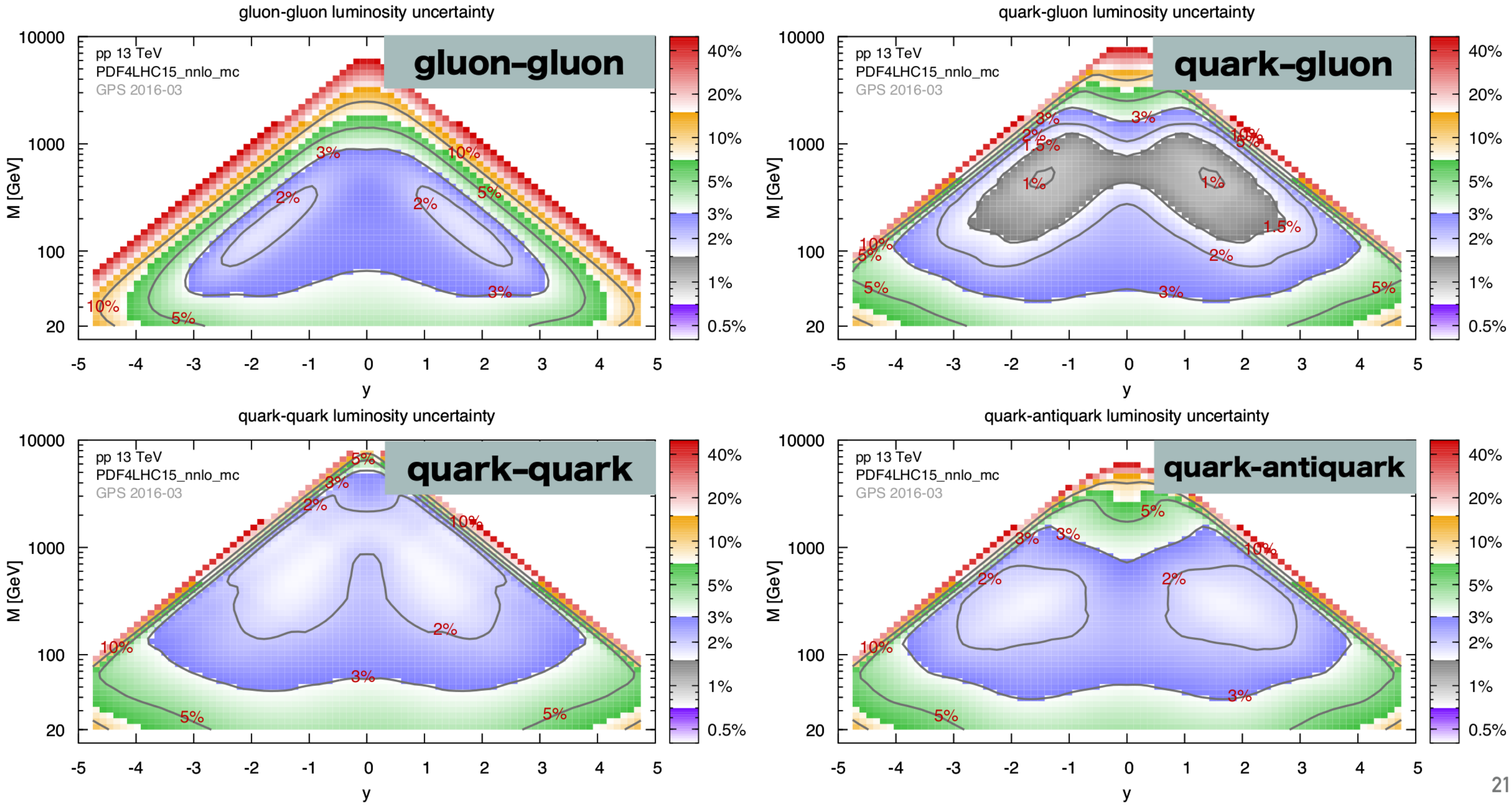
(J. Campbell, 2012)

PDF UNCERTAINTY EITHER DOMINANT, OR VERY LARGE, OR BOTH

TYPICAL PDF UNCERTAINTY $\sim 5-10\%$

PRESENT: THE PDF4LHC SET

LUMINOSITY UNCERTAINTIES VS RAPIDTY & MASS



G.P. Salam, 2016

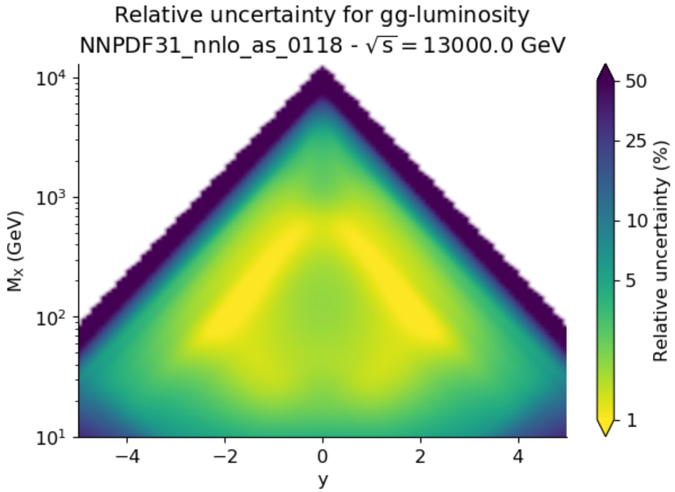
TYPICAL PDF UNCERTAINTY DOWN TO $\sim 2 - 5\%$

TOWARDS 1% PDF UNCERTAINTIES?

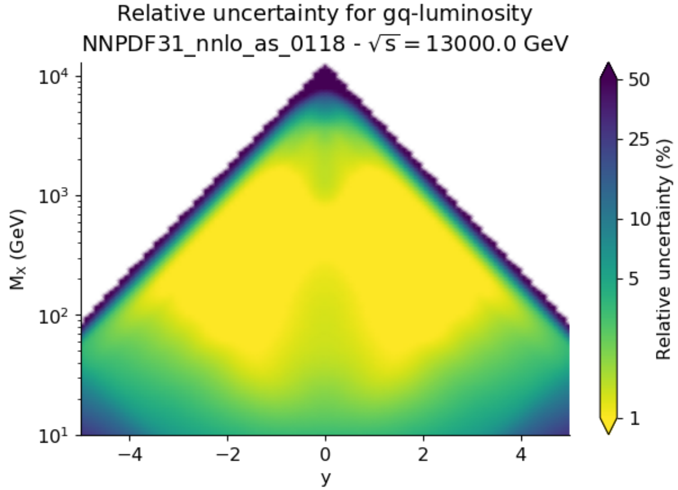
FUTURE: NNPDF3.1

LUMINOSITY UNCERTAINTIES VS RAPIDTY & MASS

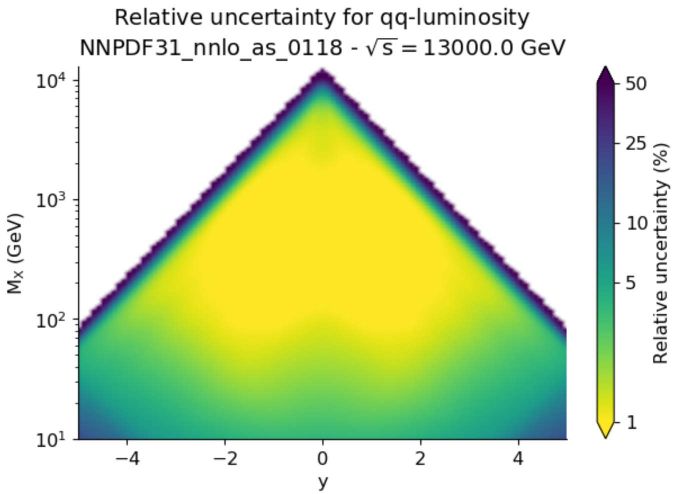
GLUON-GLUON



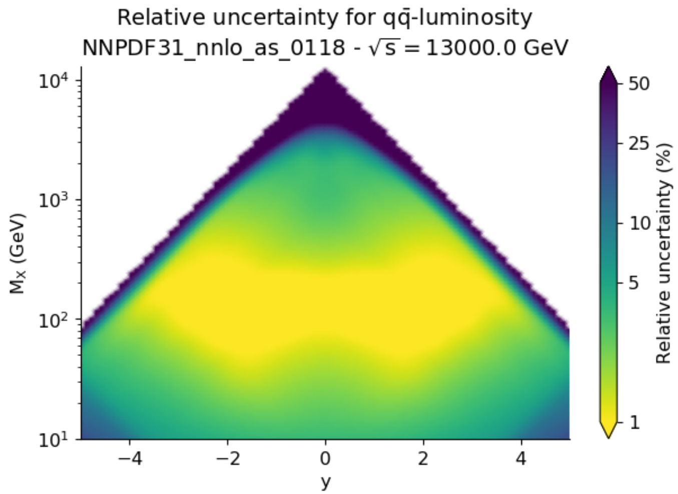
QUARK-GLUON



QUARK-QUARK



QUARK-ANTIQUARK



TYPICAL PDF UNCERTAINTY IN DATA REGION OF ORDER 1% !!
CAN WE BELIEVE IN 1% PDF UNCERTAINTIES? WHAT ARE THE CONSEQUENCES?

SUMMARY

THE IMPACT OF DATA

- WIDENING OF THE DATASET AND THE IMPACT OF LHC
- PDF UNCERTAINTIES
- FLAVOR SEPARATION & THE GLUON

METHODOLOGICAL ISSUES

- MONTE CARLO VS. HESSIAN
- PARAMETRIZATION ISSUES
- MINIMIZATION EFFICIENCY AND STATISTICAL TESTS
- CONTROLLING THE COVARIANCE MATRIX

THEORY ISSUES

- THE NNLO FRONTIER
- SMALL AND LARGE x RESUMMATION
- THE PHOTON PDF
- THE TREATMENT OF HEAVY QUARKS

THE IMPACT OF LHC DATA

CONTEMPORARY PDF TIMELINE (ONLY PUBLISHED GLOBAL)

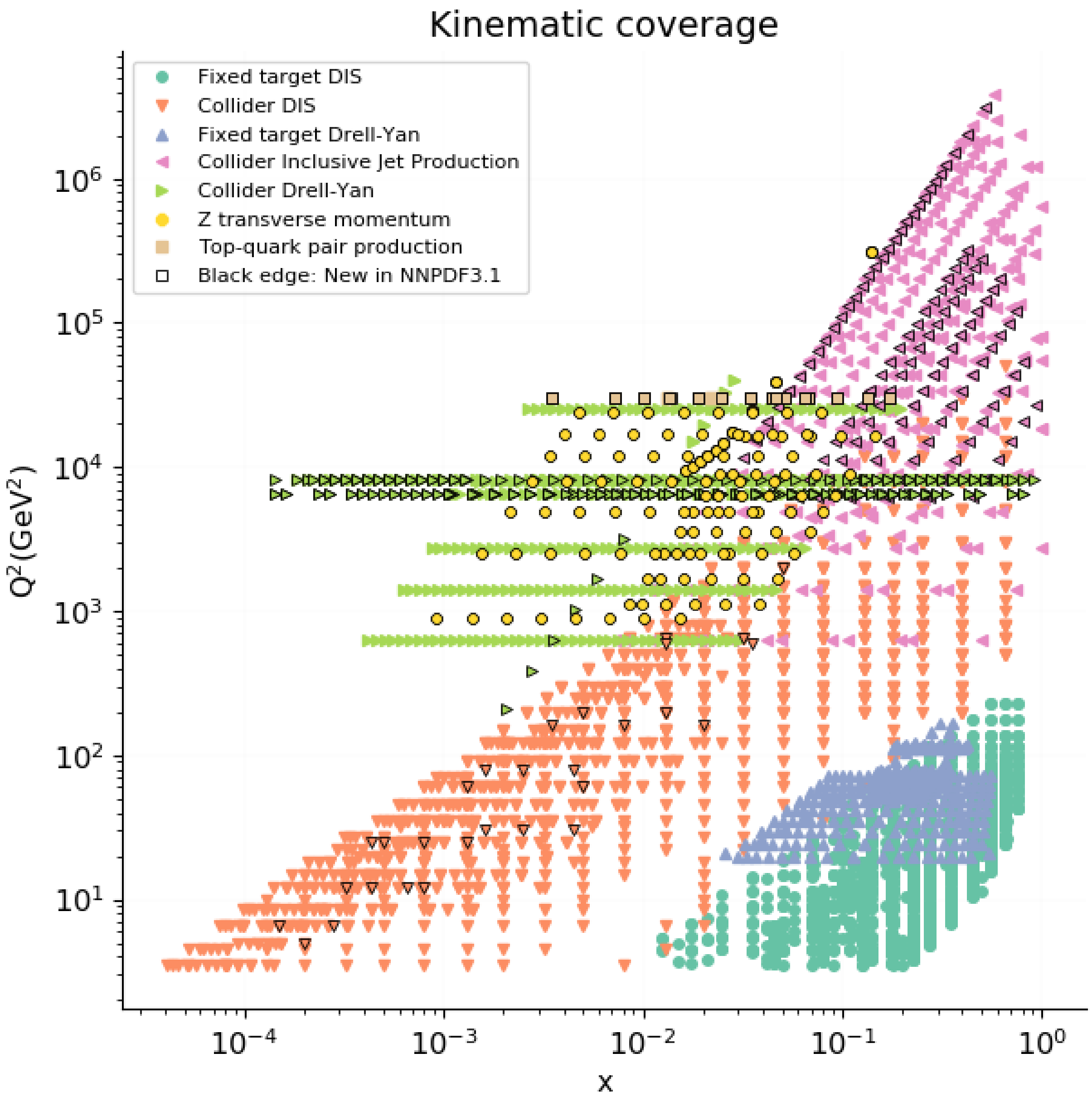
	2008		2009		2010		2011	2012		2013		2014		2015	2017	
SET	CTEQ6.6	NNPDF1.0	MSTW	ABKM09	NNPDF2.0	CT10 (NLO)	NNPDF2.1 (NNLO)	ABM11	NNPDF2.3	CT10 (NNLO)	ABM12	NNPDF3.0	MMHT	CT14	ABMP16	NNPDF3.1
MONTH	(02)	(08)	(01)	(08)	(02)	(07)	(07)	(02)	(07)	(02)	(10)	(10)	(12)	(06)	(01)	(06)
F. T. DIS	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ZEUS+H1-HI	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
COMB. HI	✗	✗	✗	✗	✔	✗	✔	✗	✔	✗	✔	✔	✗	✗	✔	✔
ZEUS+H1-HII	✗	✗	✗	✗	✗	✗	some	✗	✗	some	✗	✔	✗	✗	✔	✔
HERA JETS	✗	✗	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗
F. T. DY	✔	✗	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
TEV W+Z	✔	✗	✔	✗	✔	✔	✔	✗	✔	✔	✗	✔	✔	✔	✗	✔
LHC W+Z	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	some	✔	✔	✔	some	✔
TEV JETS	✔	✗	✔	✗	✔	✔	✗	✔	✔	✔	✗	✔	✔	✔	✗	✔
LHC JETS	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✔	✔	✗	✔
TOP TOTAL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✔	✔
SINGLE TOP TOTAL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗
TOP DIFFERENTIAL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔
W p_T	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗
W+C	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗
Z p_T	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔

THEORY PROGRESS:

- MSTW, ABKM: all NNLO; NNPDF NNLO since 07/11 (2.1), CT since 02/13 (CT10);
  NNPDF THRESHOLD RESUMMATION (3.0RESUM, 07/15), SMALL x RESUMMATION (3.1SX, 10/17)
- MSTW, CT, NNPDF all GM-VFN; NNPDF since 01/11 (2.1);
  ABM FFN+ZM-VFN since 01/17 (ABMP16)
- NNPDF FITTED CHARM since 05/16 (NNPDF3IC)
- PHOTON PDF: (mrst2004qed), NNPDF2.3QED (08/13), NNPDF3.0QED (06/16), NNPDF3.1LUXQED (12/17)

DATASET WIDENING

NNPDF3.0 vs NNPDF3.1

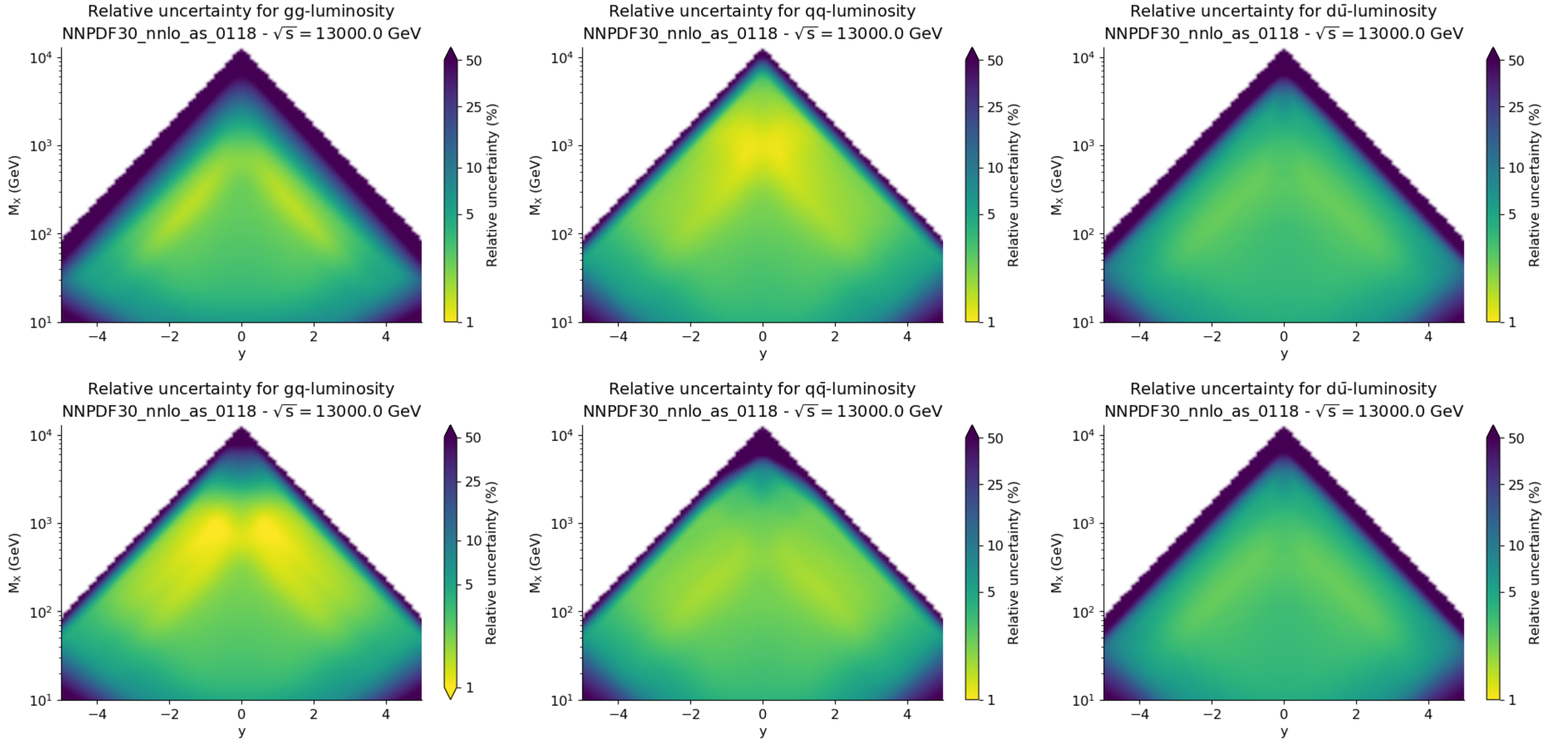


NEW DATA: (BLACK EDGE)

- HERA COMBINED F_2^b
- D0 W LEPTON ASYMMETRY
- ATLAS W, Z 2011, HIGH & LOW MASS DY 2011; CMS $W^{\pm}$ RAPIDITY 8TEV LHCB W, Z 7TEV & 8TEV
- ATLAS 7TEV JETS 2011, CMS 2.76TEV JETS
- ATLAS & CMS TOP DIFFERENTIAL RAPIDITY
- ATLAS Z p_T DIFFERENTIAL RAPIDITY & INVARIANT MASS 8TEV, CMS Z p_T DIFFERENTIAL RAPIDITY 8TEV

THE IMPACT OF LHC DATA
PDF UNCERTAINTIES IN DETAIL: NNPDF3.0 (NNLO)

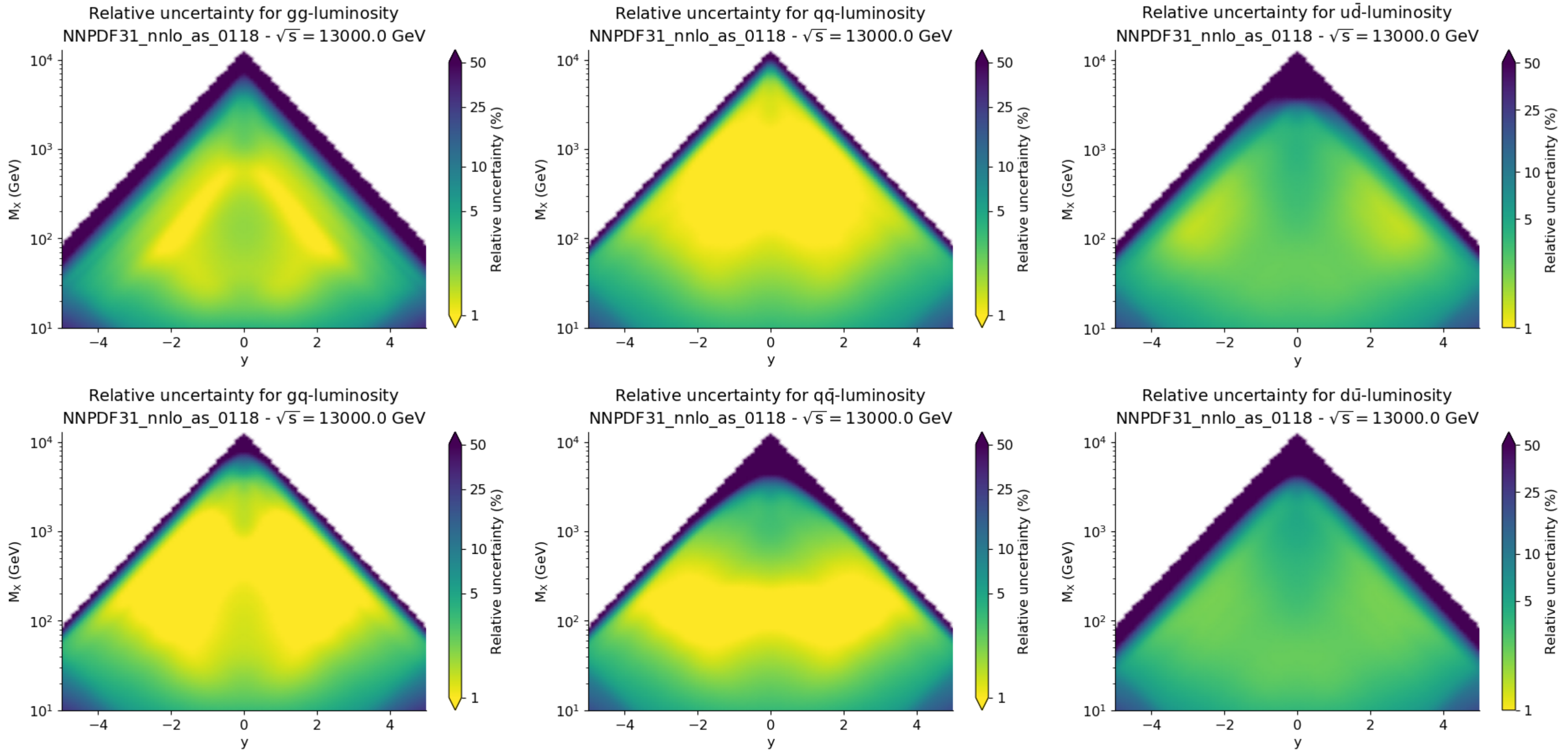
GLUON SINGLET FLAVORS



- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- TYPICAL UNCERTAINTIES IN DATA REGION $\sim 3-5\%$
- SWEET SPOT: VALENCE Q - G; DOWN TO 1%
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS

THE IMPACT OF LHC DATA
PDF UNCERTAINTIES IN DETAIL: NNPDF3.1 (NNLO)

GLUON | SINGLET | FLAVORS



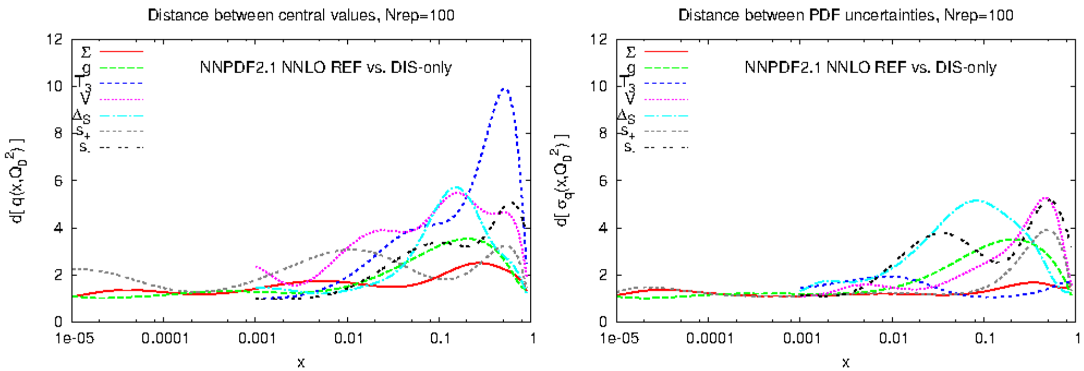
- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- TYPICAL UNCERTAINTIES IN DATA REGION $\sim 1-3\%$
- SWEET SPOT: VALENCE Q - G; 1% OR BELOW
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS
- NEW LHC DATA $\Rightarrow$ SIZABLE REDUCTION IN UNCERTAINTIES

THE IMPACT OF LHC DATA

BEFORE LHC: PDFS MOSTLY DETERMINED BY DIS DATA

NNPDF2.1 VS NNPDF2.1 DIS ONLY

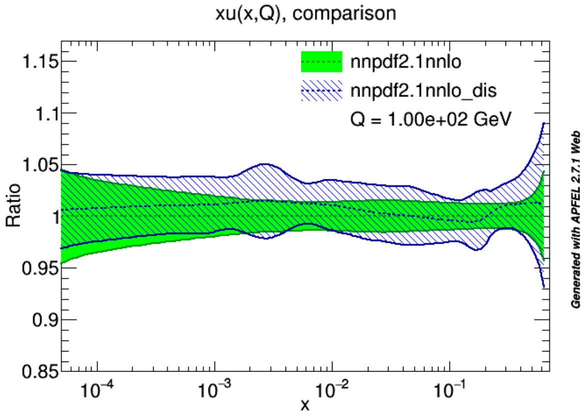
DISTANCES (difference in units of st. dev.)



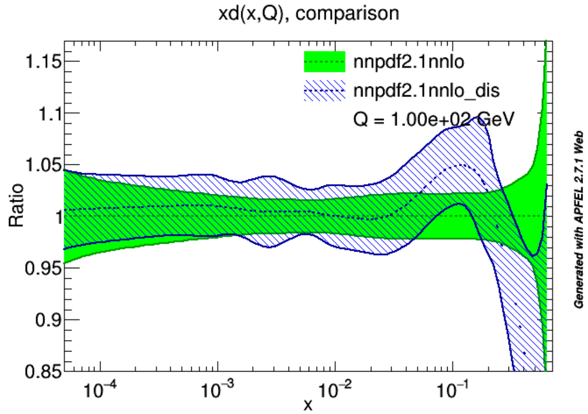
$d = 10 \Leftrightarrow$ one sigma difference

PDF COMPARISON

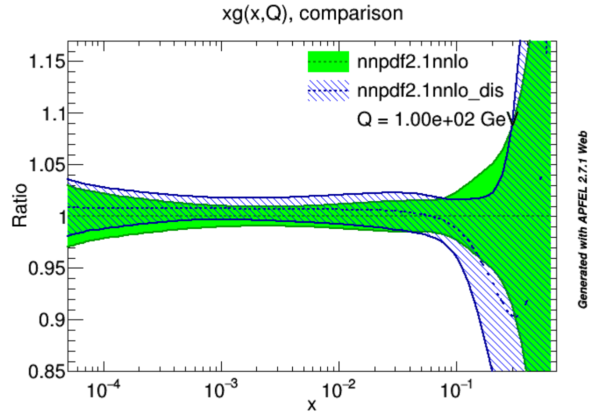
UP



DOWN



GLUON



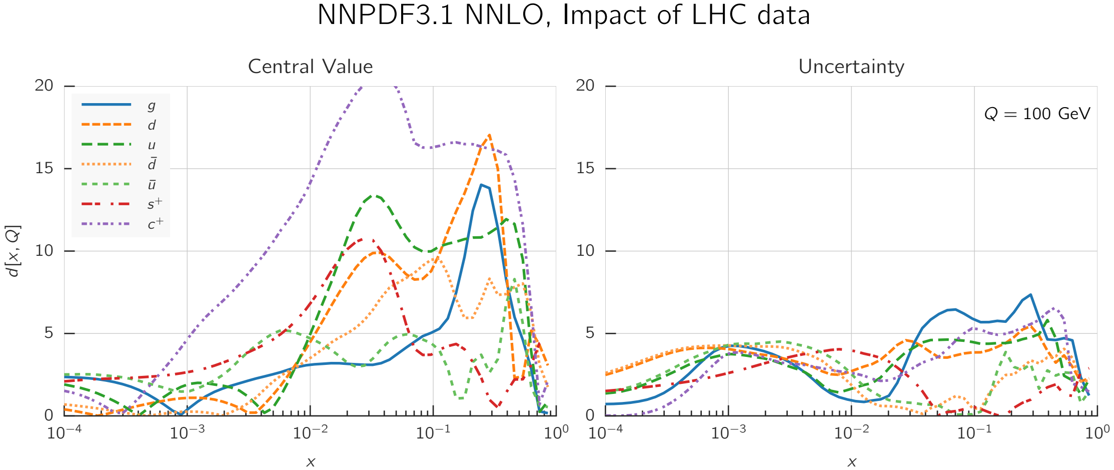
- ALL DIFFERENCES BELOW ONE SIGMA
- ONLY UP-DOWN SEPARATION SIGNIFICANTLY AFFECTED

THE IMPACT OF LHC DATA

NOW: PDFs LARGELY DETERMINED BY LHC DATA

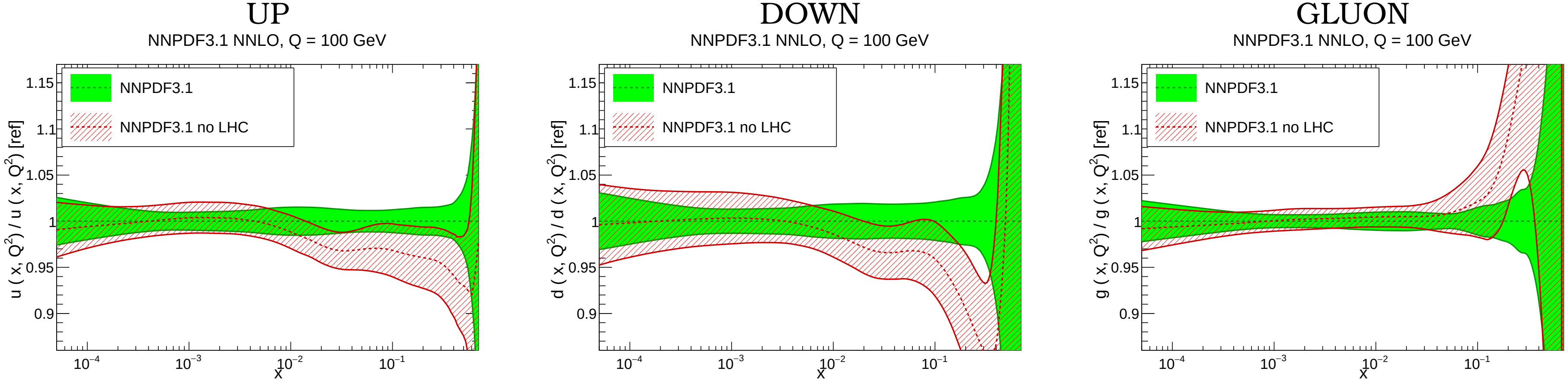
NNPDF3.1 VS NNPDF3.1 NO LHC

DISTANCES (difference in units of st. dev.)



$d = 10 \Leftrightarrow$ one sigma difference

PDF COMPARISON

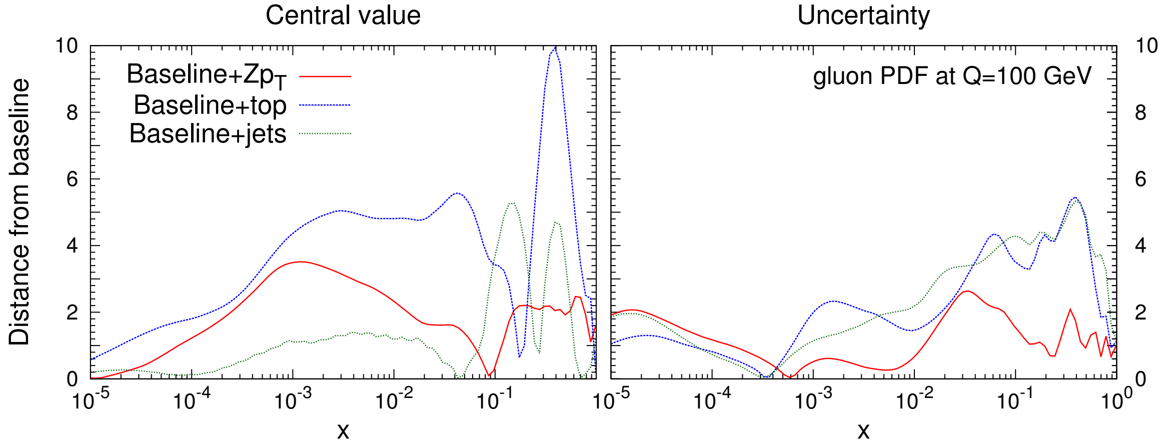


- MANY PDFs CHANGE BY MORE THAN ONE SIGMA
- BOTH FLAVOR SEPARATION & GLUON SIGNIFICANTLY AFFECTED

THE IMPACT OF LHC DATA
THE GLUON

- BEFORE LHC ⇒ DIS SCALING VIOLATIONS, TEV JETS AT LARGE X
- AFTER LHC ⇒ JETS; $Z\ p_t$, TOP

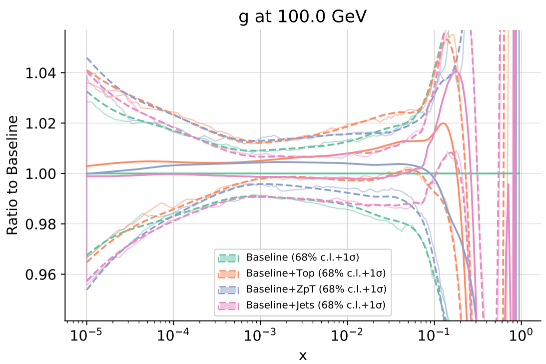
DISTANCES (difference in units of st. dev.)



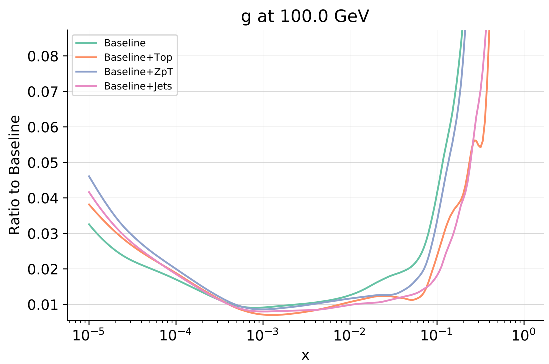
(Nocera, Ubiali, 2017)

PDF COMPARISON: GLUON

CENTRAL VALUE



UNCERTAINTY

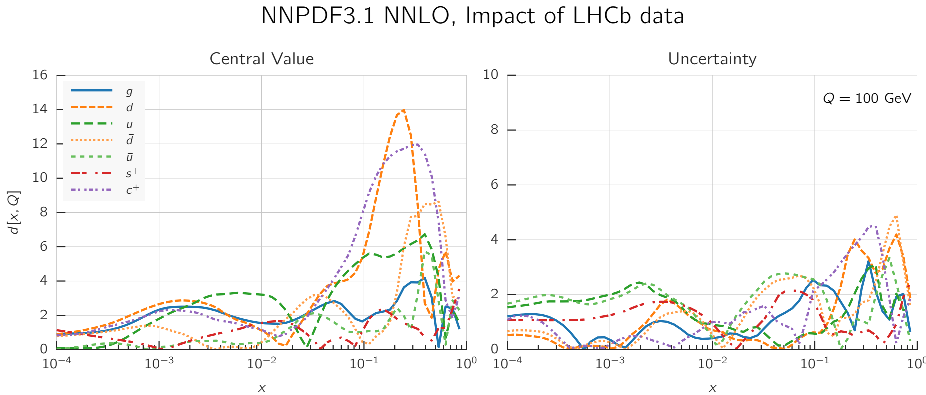


- TOP HAS LARGEST IMPACT, FOLLOWED BY JETS
- ALL LHC DATA PULL CENTRAL VALUE IN SAME DIRECTION!

THE IMPACT OF LHC DATA
FLAVOR SEPARATION

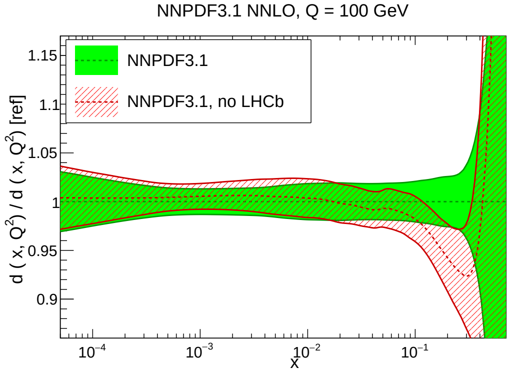
- BEFORE LHC ⇒ CC DIS, TEV FIXED-TARGET DY, W ASYM.
- AFTER LHC ⇒ WIDE RANGE OF W, Z PRODUCTION DATA

IMPACT OF LHCB
DISTANCES (difference in units of st. dev.)

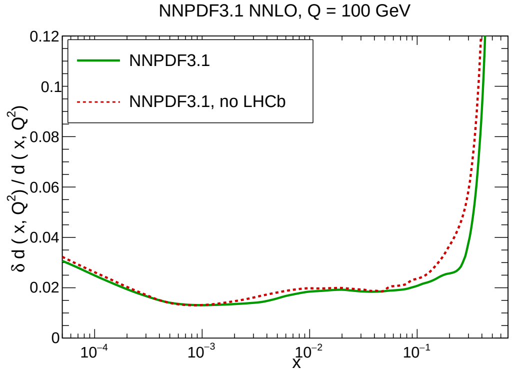


PDF COMPARISON: DOWN

CENTRAL VALUE



UNCERTAINTY



- SIZABLE SHIFT OF CENTRAL VALUE BY ALMOST ONE SIGMA
- LARGE x UNCERTAINTY DOWN BY LARGE FACTOR!

NEW DATA: SUMMARY

- LHC DATA NOW HAVE THE DOMINANT IMPACT ON PDFs
- METHODOLOGY AND THEORY MUST ACCORDINGLY ADAPT

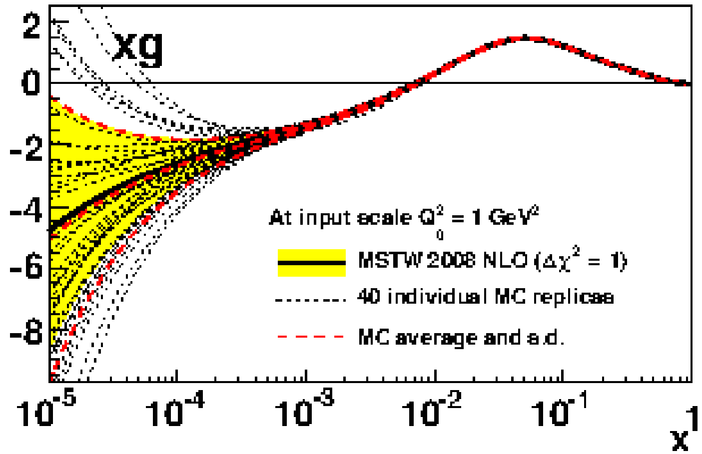
THE LIMITS OF METHODOLOGY

PDF parametrization & delivery

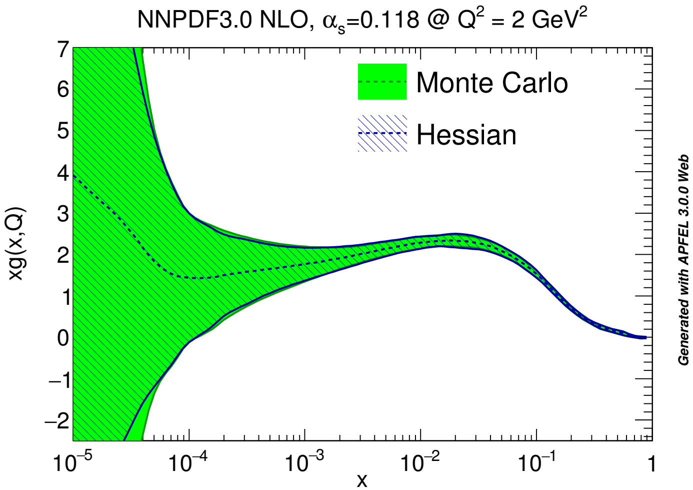
- traditionally, two delivery methods for PDFs
- Hessian a central PDF set, & error sets corresponding to eigenvectors of the covariance matrix in parameter space
  ADVANTAGE: efficient representation of uncertaity
  DISADVANTAGES: assumes gaussianity
- Montecarlo a set of PDF replicas which represents the probability in PDF space (so the mean unbiasedly estimates the central value &c)
  ADVANTAGE: faithful representation of probability
  DISADVANTAGES: may need large number of replicas
- traditionally, delivery ⇔ parametrization/minimization
  Hessian used with relatively simple functional forms (small numbers of parameters) ⇔ Hessian minimization

PROGRESS I
MC $\Leftrightarrow$ HESSIAN

- TO CONVERT HESSIAN INTO MONTECARLO GENERATE MULTIGAUSSIAN REPLICAS IN PARAMETER SPACE
- ACCURATE WHEN NUMBER OF REPLICAS SIMILAR TO THAT WHICH REPRODUCES DATA



(Thorne, Watt, 2012)

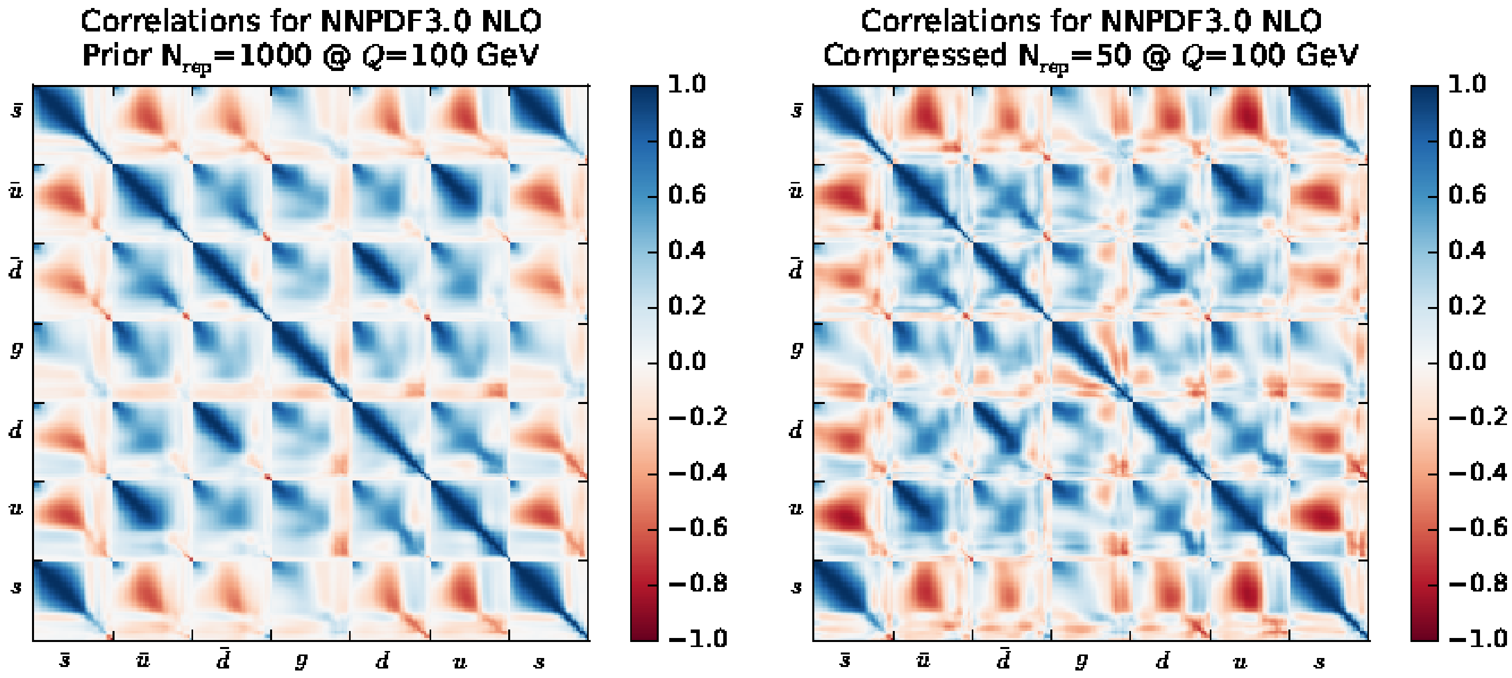


(Carrazza, SF, Kassabov, Rojo, 2015)

- TO CONVERT MONTE CARLO INTO HESSIAN, SAMPLE THE REPLICAS $f_i(x)$ AT A DISCRETE SET OF POINTS & CONSTRUCT THE ENSUING COVARIANCE MATRIX
- EIGENVECTORS OF THE COVARIANCE MATRIX AS A BASIS IN THE VECTOR SPACE SPANNED BY THE REPLICAS BY SINGULAR-VALUE DECOMPOSITION
- NUMBER OF DOMINANT EIGENVECTORS SIMILAR TO NUMBER OF REPLICAS $\Rightarrow$ ACCURATE REPRESENTATION

PROGRESS II

MONTECARLO COMPRESSION



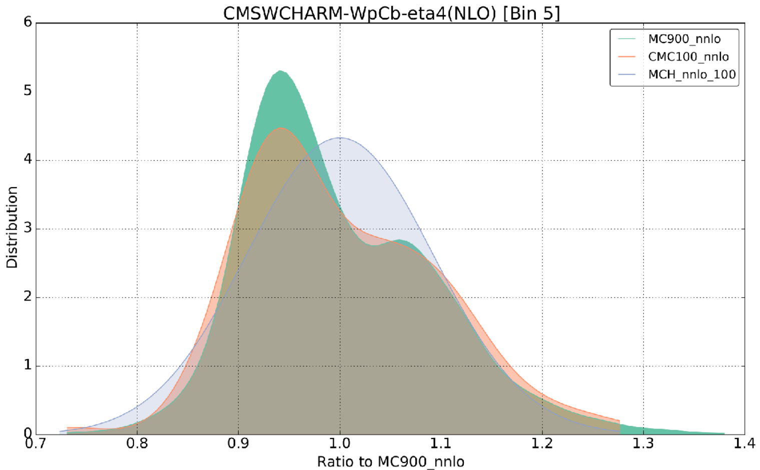
(Carrazza, Latorre, Kassabov, Rojo, 2015)

- CONSTRUCT A VERY LARGE REPLICA SAMPLE
- SELECT (BY GENETIC ALGORITHM) A SUBSET OF REPLICAS WHOSE STATISTICAL FEATURES ARE AS CLOSE AS POSSIBLE TO THOSE OF THE PRIOR
- $\Rightarrow$ FOR ALL PDFs ON A GRID OF POINTS// MINIMIZE DIFFERENCE OF: FIRST FOUR MOMENTS, CORRELATIONS; OUTPUT OF KOLMOGOROV-SMIRNOV TEST (NUMBER OF REPLICAS BETWEEN MEAN AND σ, 2σ, INFINITY)
- 50 COMPRESSED REPLICA REPRODUCE 1000 REPLICA SET TO PRECENT ACCURACY

NONGAUSSIAN BEHAVIOUR

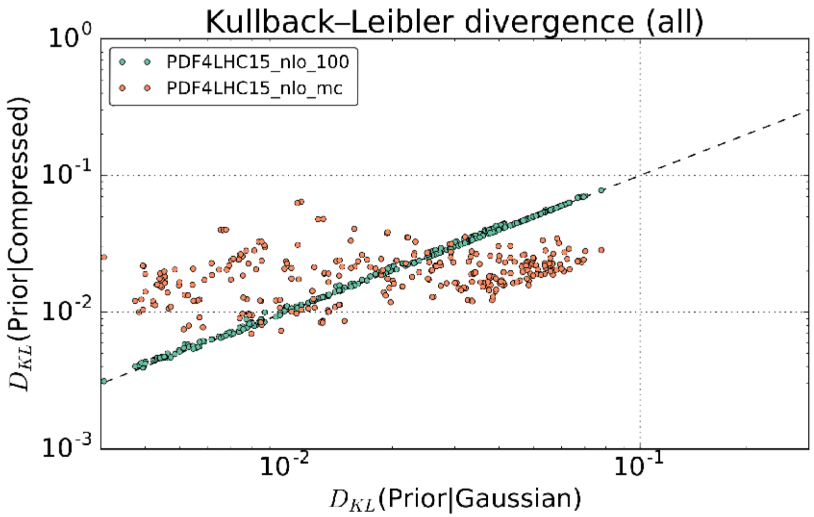
MONTE CARLO COMPARED TO HESSIAN

CMS $W + c$ production



- DEVIATION FROM GAUSSIANITY E.G. AT LARGE x DUE TO LARGE UNCERTAINTY + POSITIVITY BOUNDS
  $\Rightarrow$ RELEVANT FOR SEARCHES
- CANNOT BE REPRODUCED IN HESSIAN FRAMEWORK
- WELL REPRODUCED BY COMPRESSED MC

- DEFINE KULLBACK-LEIBLER DIVERGENCE
  $D_{\rm KL} = \int_{-\infty}^{\infty} P(x) \frac{\ln P(x)}{\ln Q(x)}\, dx$
  BETWEEN A PRIOR P AND ITS REPRESENTATION Q
- $D_{\rm KL}$ BETWEEN PRIOR AND HESSIAN DEPENDS ON DEGREE OF GAUSSIANITY
- $D_{\rm KL}$ BETWEEN PRIOR AND COMPRESSED MC DOES NOT



CAN (A) GAUGE WHEN MC IS MORE ADVANTAGEOUS THAN HESSIAN;
(B) ASSESS THE ACCURACY OF COMPRESSION

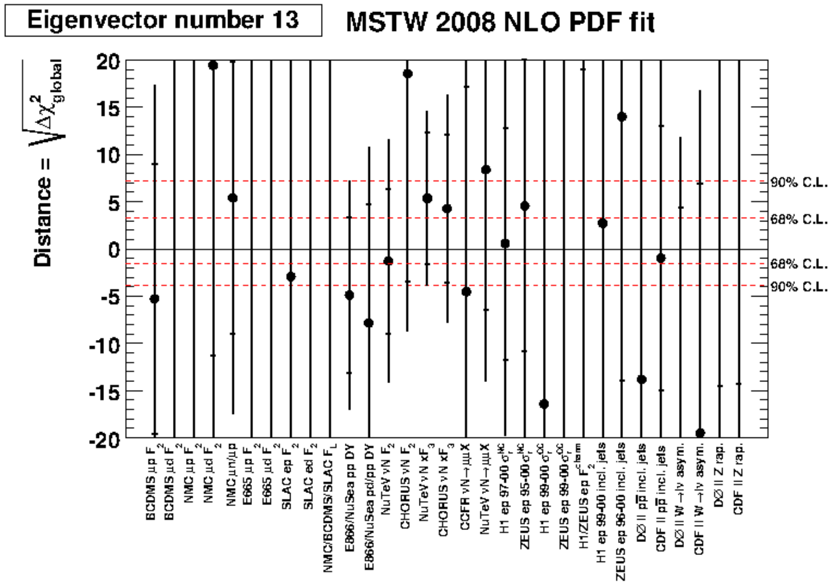
PDF PARAMETRIZATION ISSUES

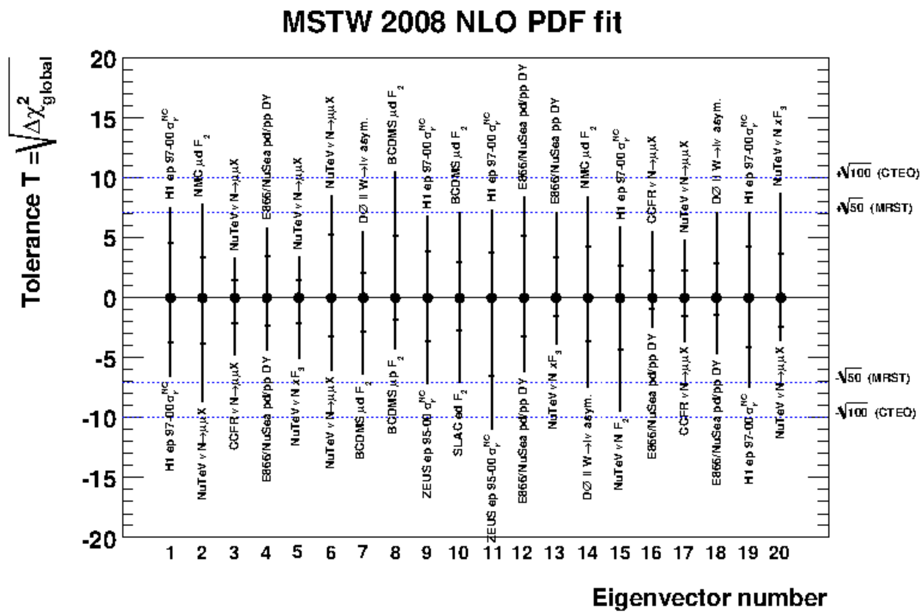
- Q: WHY ARE PDF UNCERTAINTIES ON GLOBAL FITS OF SIMLAR SIZE?
  - SIMILAR DATASETS
  - BUT DIFFERENT PROCEDURES
- A: UNCERTAINTY TUNING

TOLERANCE (MMHT-CT)

GLOBAL MSTW TOLERANCE

MSTW TOLERANCE PLOT FOR 13TH EIGENVEC.



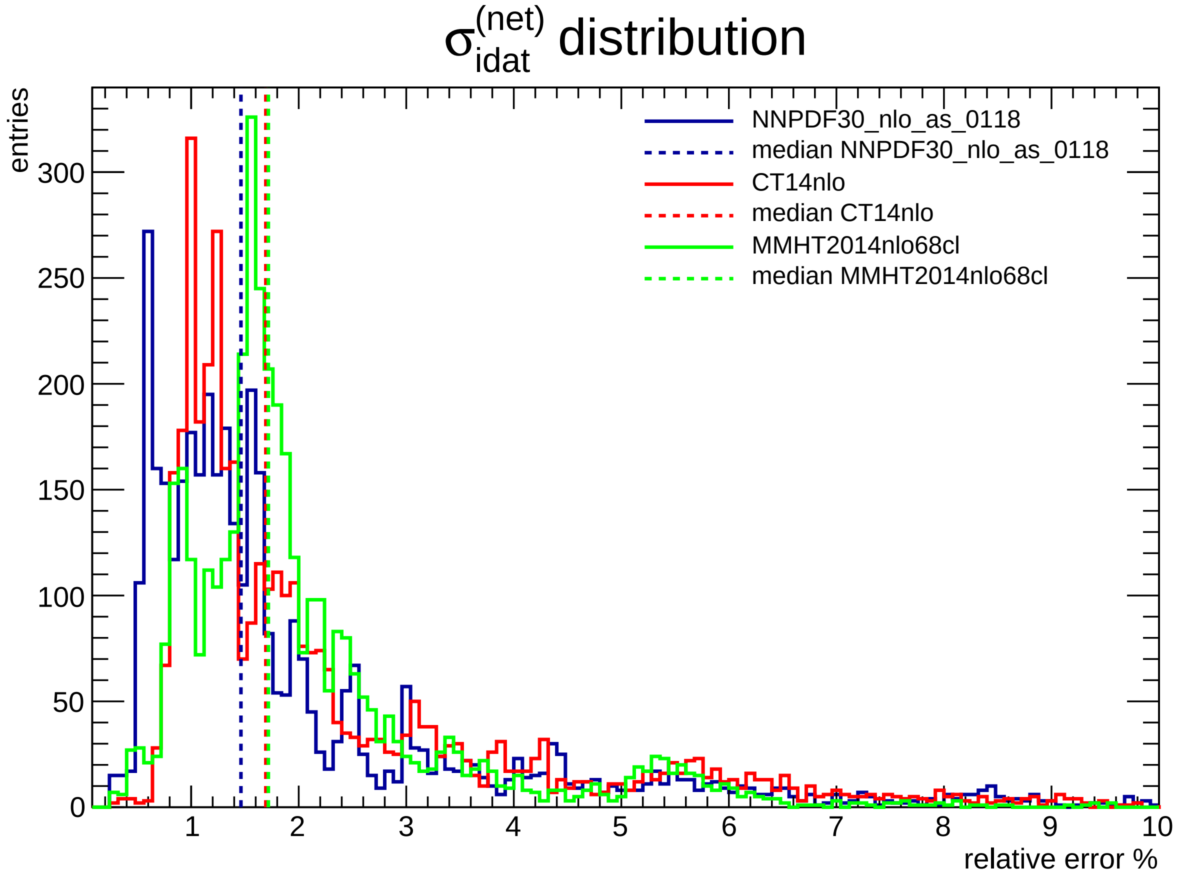


- (MSTW/MMHT) FOR EACH EIGENVECTOR IN PARAMETER SPACE DETERMINE CONFIDENCE LIMIT FOR THE DISTRIBUTION OF BEST-FITS OF EACH EXPERIMENT
- RESCALE $\Delta\chi^2 = T$ INTERVAL SUCH THAT CORRECT CONFIDENCE INTERVALS ARE REPRODUCED
- WHY DO WE NEED TOLERANCE?
- DO WE UNDERSTAND PDF UNCERTAINTIES?

PDF UNCERTAINTIES: HOW MUCH DO THEY VARY?

- COMPUTE PERCENTAGE PDF UNCERTAINTY ON ALL DATA INCLUDED IN GLOBAL FIT
- COMPARE GLOBAL FITS

PERCENTAGE PDF UNCERTAINTY ON PREDICTIONS



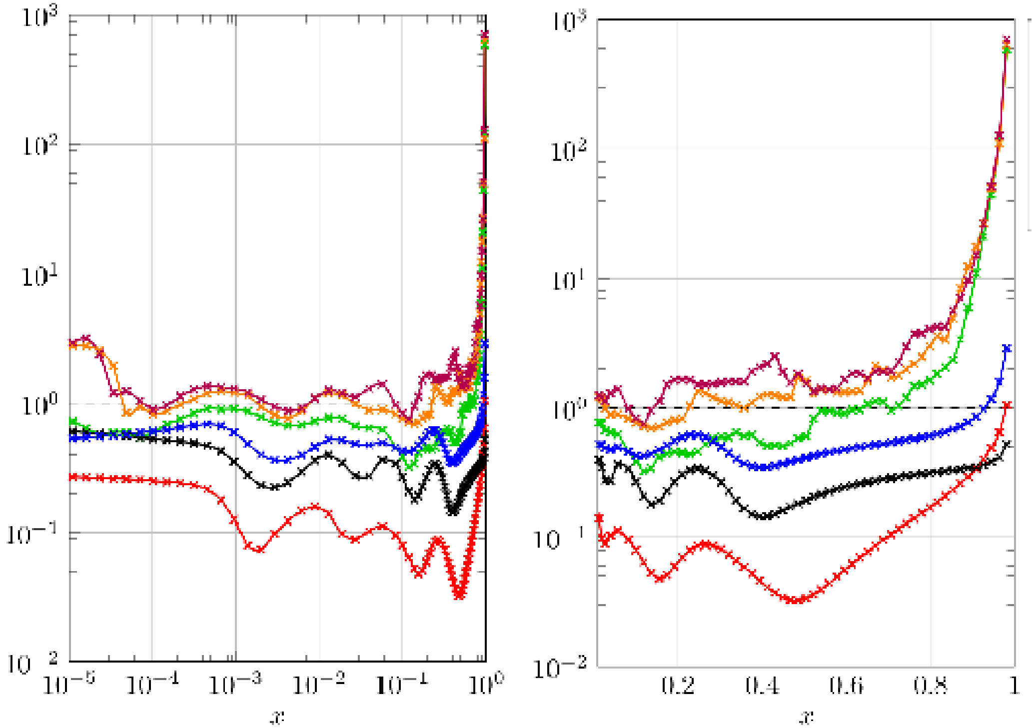
- MEDIAN SIMILAR
- DISTRIBUTION VERY DIFFERENT!
- NNPDF: SMALLER MODE, BUT FAT TAIL ⇔ GREATER FLEXIBILITY

CLOSURE TESTING
BASIC IDEA

- ASSUME PDFS KNOWN: GENERATE FAKE EXPERIMENTAL DATA
- CAN DECIDE DATA UNCERTAINTY (ZERO, OR AS IN REAL DATA, OR . . .)
- FIT PDFS TO FAKE DATA:
  - LEVEL 0: ZERO UNCERTAINTY
    - CHECK WHETHER MINIMZATION EFFICIENT
    - CHECK FOR INTERPOLATION UNCERTAINTY
  - LEVEL 1: DATA UNCERTAINTY, BUT NO REPLICAS
    - CHECK FOR UNIQUENESS OF BEST FIT ⇒ “FUNCTIONAL” UNCERTAINTY (Pumplin, 2010)
  - LEVEL 2: AS IN STANDARD PROCEDURE
    - CHECK WHETHER TRUE VALUE GAUSSIANLY DISTRIBUTED ABOUT FIT
    - CHECK WHETHER UNCERTAINTIES FAITHFUL

CLOSURE-TESTING: THE PARAMETRIZATION DEPENDENCE

GLUON PDF UNCERTAINTY NORMALIZED TO MSTW08



(C. Mascaretti, 2016)

- CLOSURE TEST PERFORMED WITH DATA GENERATED BASED ON MST08 FUNCTIONAL FORM
- REFITTED EITHER WITH NNPDF OR MSTW-CT FUNCTIONAL FORM
- LEVEL 0: VANISHING DATA UNCERTAINTY
  - MSTW-CT: FIT HAS ZERO UNCERTAINTY
  - NNPDF: ABOUT HALF OF TOTAL UNCERTAINTY
- LEVEL 1: NOMINAL DATA UNCERTAINTY, BUT REPLICAS FITTED W/O PSEUDODATA
  - MSTW-CT: FIT HAS SMALL UNCERTAINTY
  - NNPDF: ABOUT 2/3 OF FINAL UNCERTAINTY
- LEVEL 2
  - NNPDF UNCERTAINTY LARGER THAN MSTW-CT
  - NNPDF UNCERTAINTY SIMILAR TO MSTW WITH TOLERANCE

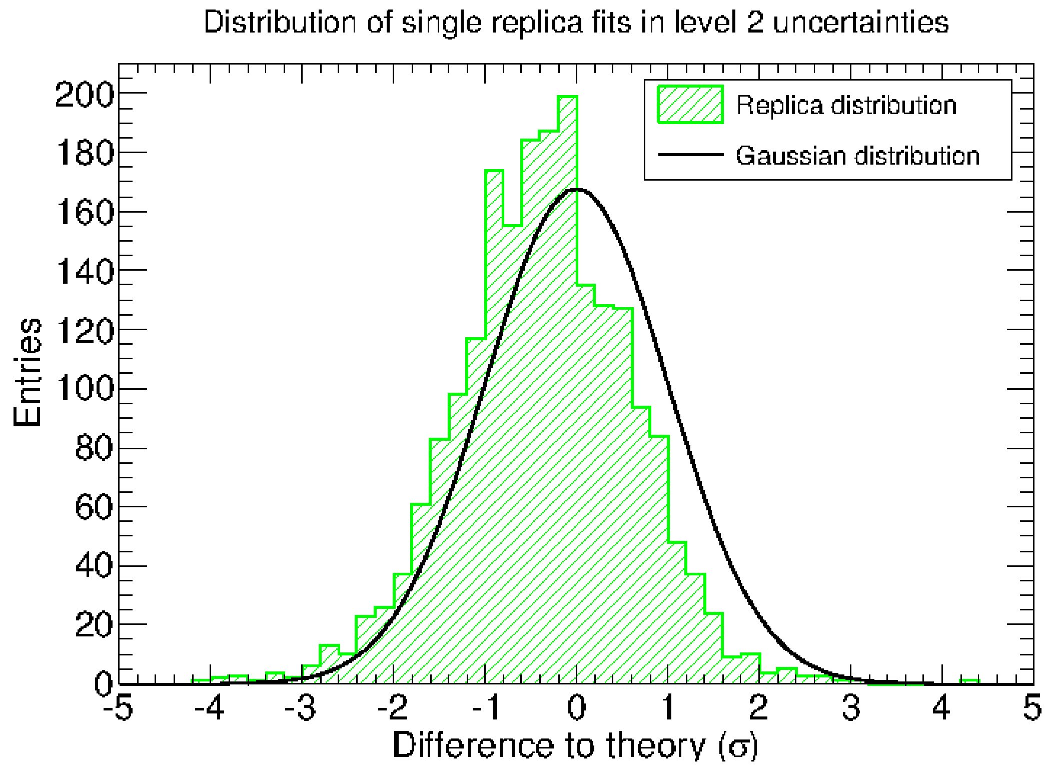
"STANDARD" PARAMETRIZATION MISSES INTERPOLATION & FUNCTIONAL UNCERTAINTY?

THE $\Delta\chi^2$ PROBLEM

- TOLERANCE MIGHT COMPENSATE FOR MISSING FUNCTIONAL UNCERTAINTY
- BUT WHAT IS $\Delta\chi^2$ FOR AN NNPDF FIT?
- CAN ANSWER USING HESSIAN CONVERSION! $\Delta\chi^2 = 16 \pm 15$
  - NON-PARABOLIC BEHAVIOUR NEAR MINIMUM ON SCALE OF UNCERTAINTIES?
  - INEFFICIENCY OF THE MINIMIZATION PROCEDURE?

CLOSURE-TESTING THE PDF UNCERTAINTIES
RESULTS

UNCERTAINTIES: DISTRIBUTION OF DEVIATIONS BETWEEN FITTED AND "TRUE" PDFS, SAMPLED AT 20 POINTS BETWEEN 10^{-5} AND 1

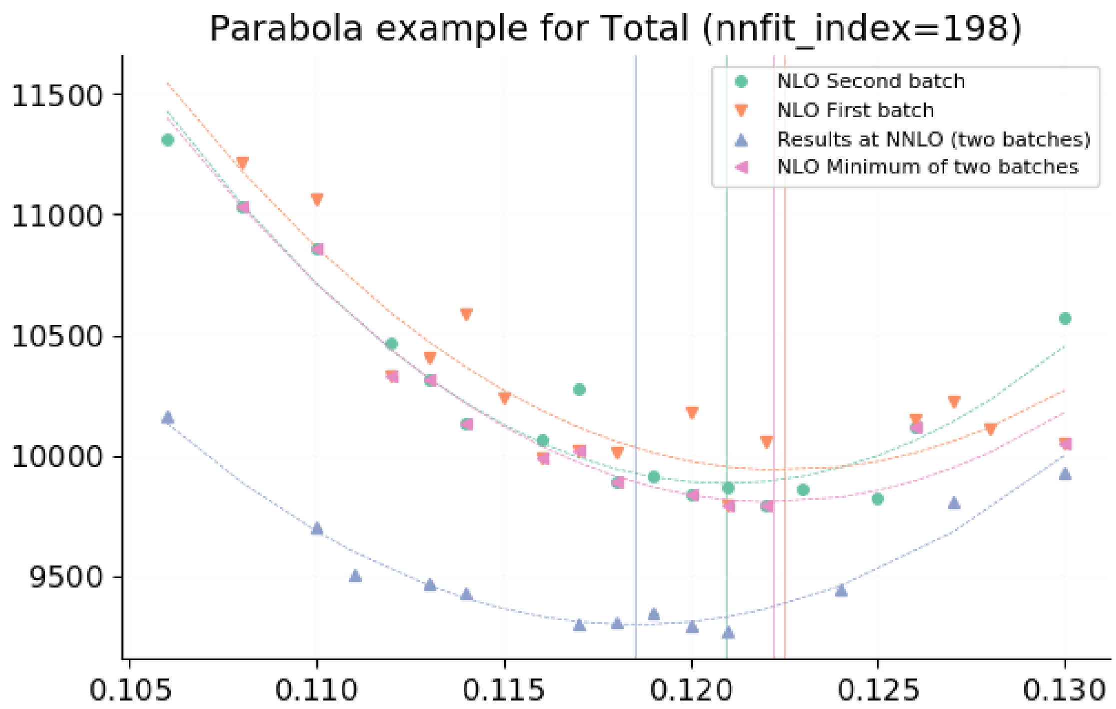


FIND 0.699% FOR ONE-SIGMA, 0.948% FOR TWO-SIGMA C.L.

- PDF UNCERTAINTIES ARE FAITHFUL
- BUT ARE THEY THE SMALLEST FROM GIVEN DATA?

MORE EFFICIENT MINIMIZATION?

- LOOK AT α_s DEPENDENCE (CORRELATED REPLICAS)
- SIGNIFICANT FLUCTUATIONS ABOUT PARABOLIC SHAPE
  NOT DUE TO FINITE-SIZE MONTE CARLO SAMPLE



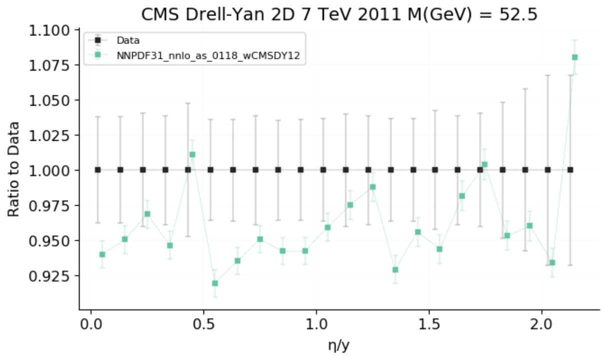
BATCH MINIMIZATION

- MINIMIZE EACH REPLICA MORE THEN ONCE & KEEP BEST RESULTS
- SIGNIFICANT STABILIZATION

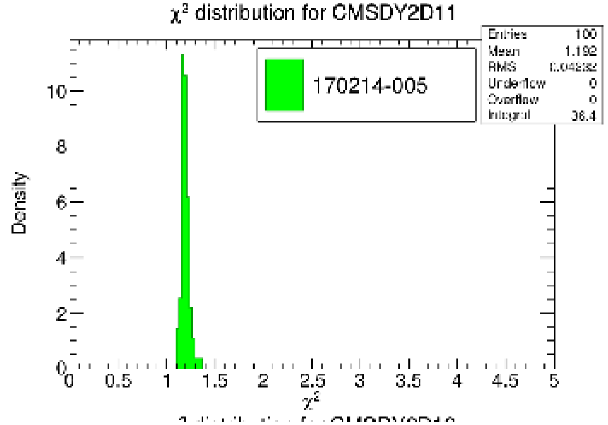
CORRELATIONS & THE COVARIANCE MATRIX

THE CMS DOUBLE-DIFFERENTIAL DRELL-YAN 2011

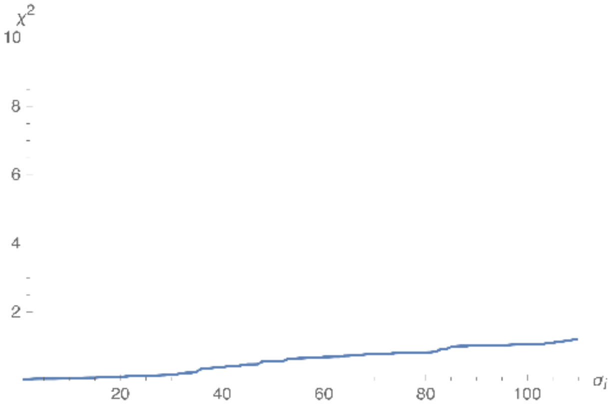
DATA/THEORY VS. DATA BIN



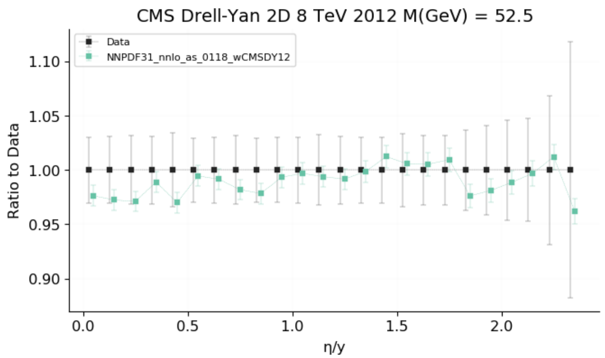
χ^2/dof HIST. OVER REPLICAS

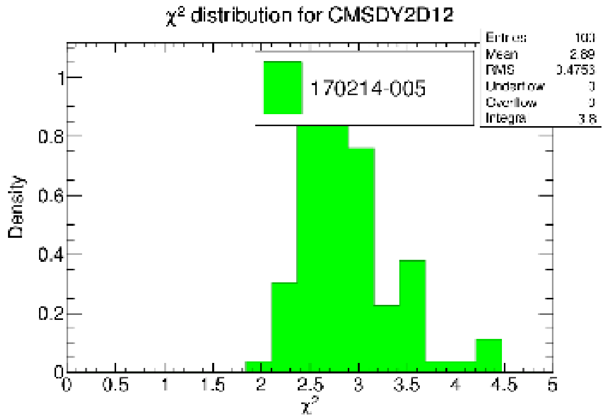


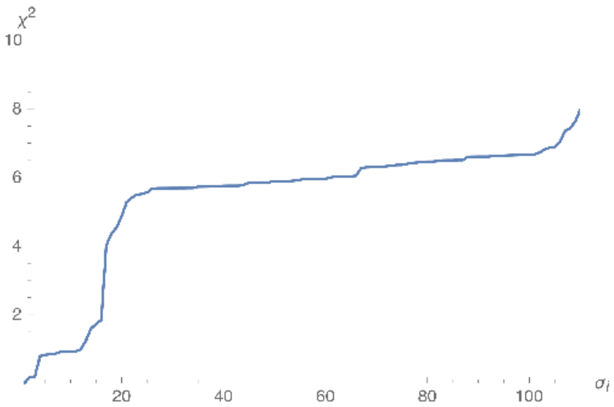
χ^2 AS COVMAT EIGVECS ADDED



2012





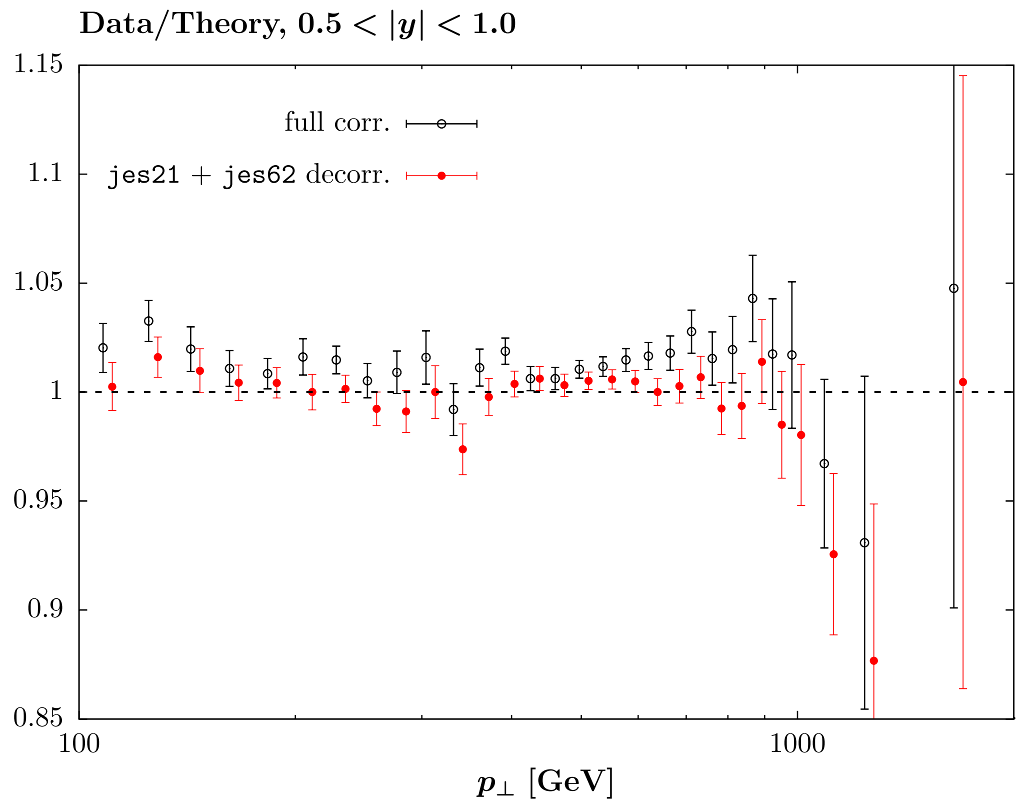


- FROM 2011 TO 2012, UNCORRELATED UNCERTAINTIES DOWN TO SUB-PERMILLE
- 2011: $\chi^2/dof \sim 1$; 2012: IMPOSSIBLE TO FIT BETTER THAN $\chi^2/dof \sim 3$
- PATHOLOGICAL BEHAVIOUR OF COVARIANCE MATRIX $\Rightarrow$ WHAT IS THE UNCERTAINTY ON IT?

CORRELATIONS & THE COVARIANCE MATRIX

THE ATLAS 7TeV JETS

- EACH RAPIDITY BIN CAN BE FITTED WITH $\chi^2/dof \sim 1$
- EACH LEADS TO INDISTIGUISHABLE BEST-FIT PDFs
- IF ALL BINS FITTED SIMULTANEOUSLY, $\chi^2/dof \sim 3$

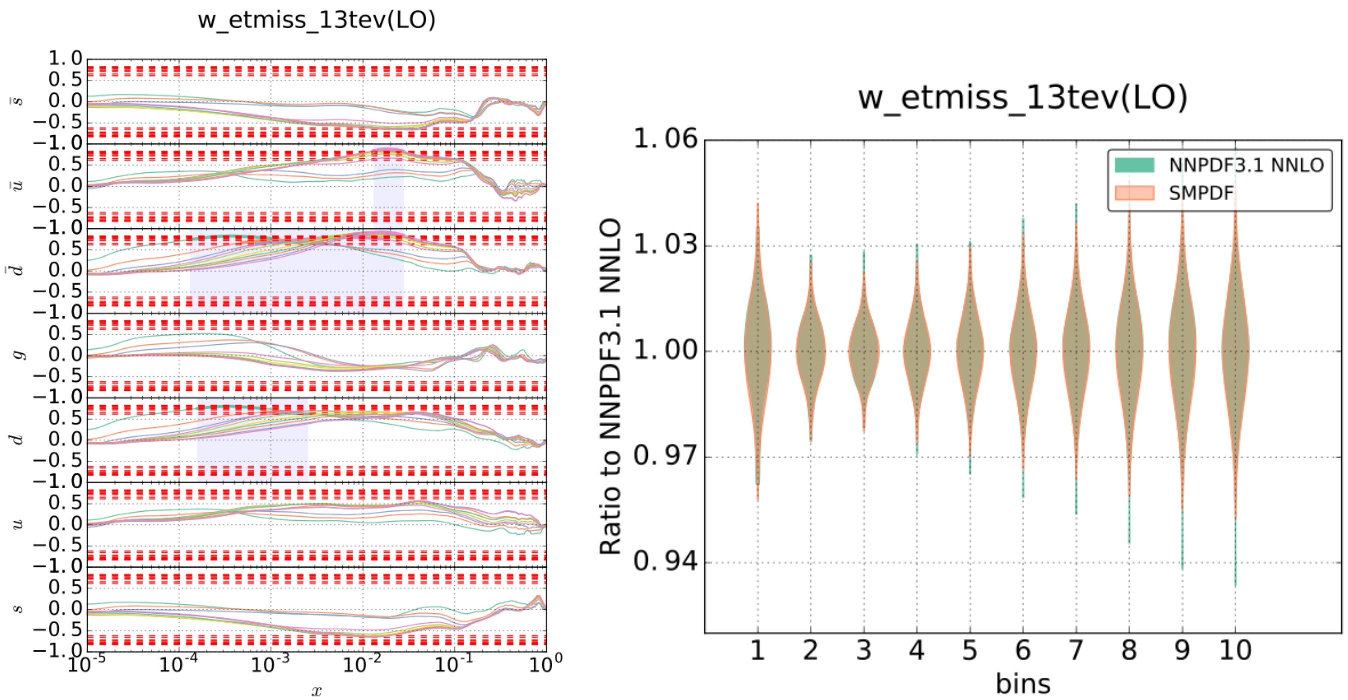


(Harland-Lang, Martin, Thorne, 1016)

- MISESTIMATED CORRELATIONS?
- CAN SINGLE OUT WHICH CORRELATION OUGHT TO BE REMOVED

SMPDF
A POWERFUL TOOL

- OLD ASPIRATION: PDFS OPTIMIZED TO PROCESSES (Pumplin 2009)
- SELECT SUBSET OF THE COVARIANCE MATRIX CORRELATED TO A GIVEN SET OF PROCESSES
- PERFORM SVD ON THE REDUCED COVARIANCE MATRIX, SELECT DOMINANT EIGENVECTOR, PROJECT OUT ORTHOGONAL SUBSPACE
- ITERATE UNTIL DESIRED ACCURACY REACHED
- CAN ADD PROCESSES TO GIVEN SET; CAN COMBINE DIFFERENT OPTIMIZED SETS
- WEB INTERFACE AVAILABLE



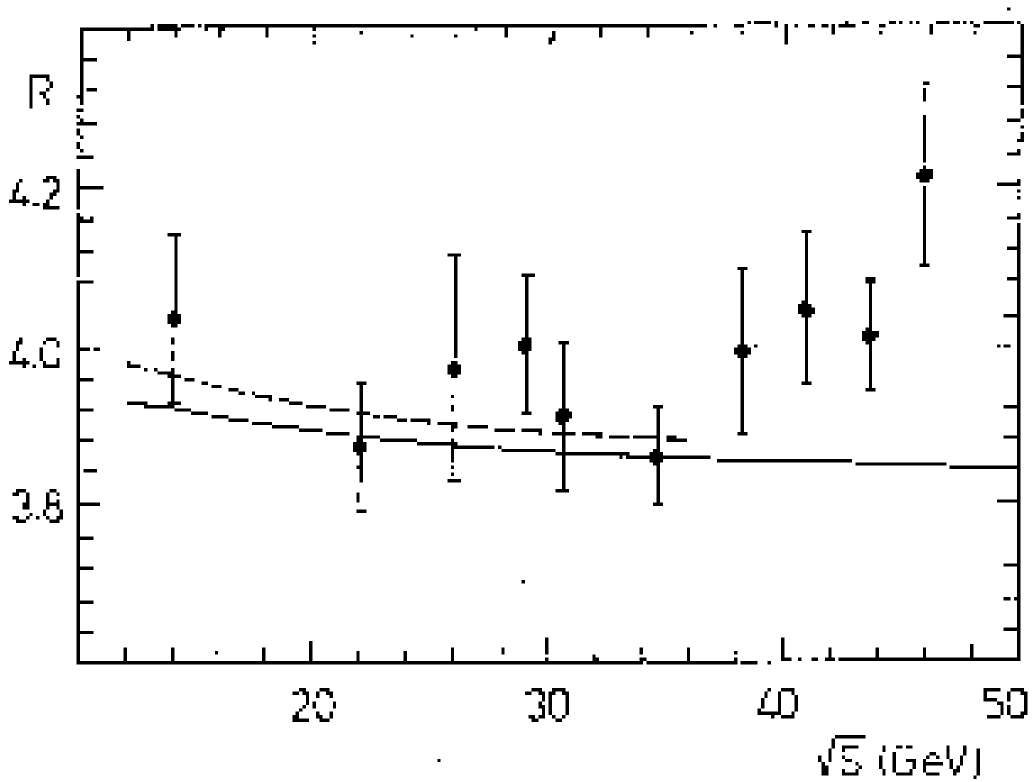
(Carrazza, SF, Kassabov, Rojo, 2016)

- EG ggH, $Hb\bar{b}$, W $E_T^{\rm miss}$ $\Rightarrow$ 11 EIGENVECTORS
- STUDY CORRELATIONS OF PDFS TO DATA AND AMONG THEMSELVES!

AN OLD PROBLEM

THE D'AGOSTINI BIAS

$$R = \frac{e^+e^- \to \text{hadrons}}{e^+e^- \to \mu^+\mu^-}$$



(CELLO collab., 1987)

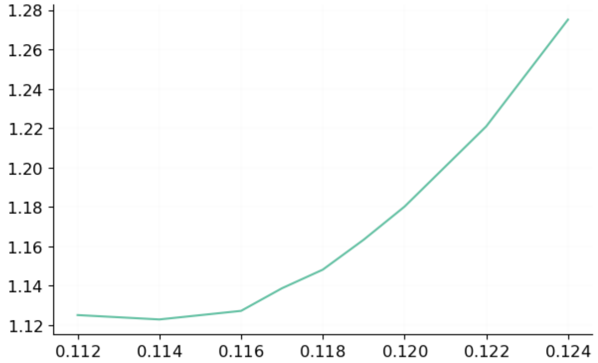
- MULTIPLICATIVE UNCERTAINTIES IN COVARIANCE MATRIX
$\Rightarrow$ FIT BIASED DOWNWARDS IF DATA INCONSISTENT (d'Agostini, 1994)
EQUIVALENT TO RESCALING DATA BUT NOT UNCERTAINTIES

- MUST USE ITERATIVE PROCEDURE
COVARIANCE MATRIX COMPUTED FROM PREVIOUS FIT (NNPDF, 2010)

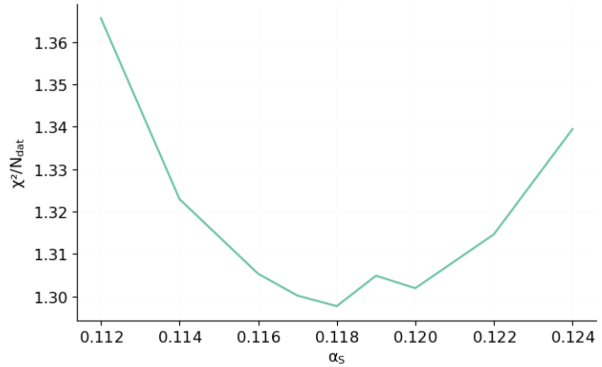
THE D'AGOSTINI BIAS

A SUBTLE EXAMPLE: α_s IN A PDF FIT

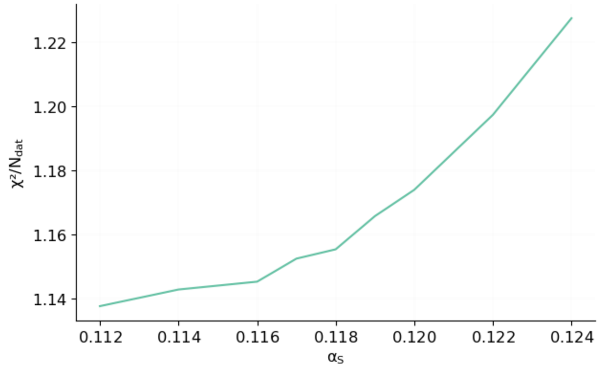
GLOBAL



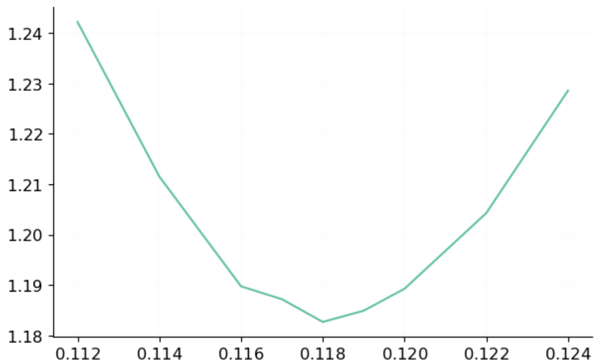
ADDITIVE (NMC, FT)

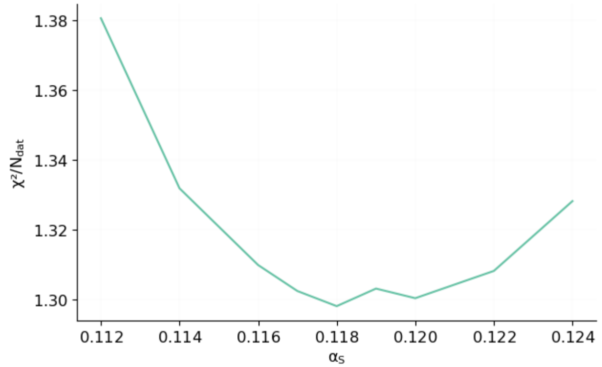


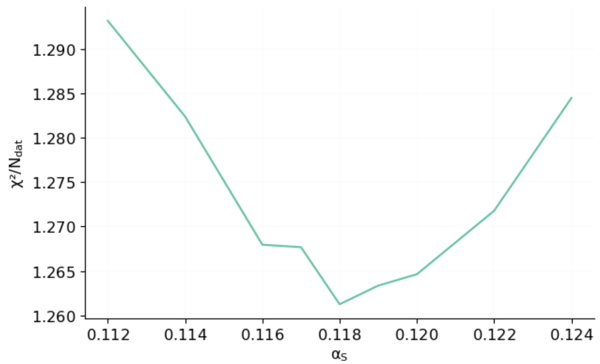
MULTIPLICATIVE (HERA, COLL)



CONSISTENT (ITERATIVE)







- χ^2 COMPUTED FROM COVARIANCE MATRIX $\Rightarrow$ BIASED LOW FIT FAVORED
- LESS EVOLUTION $\Leftrightarrow$ LOW α_s
- ONLY WHEN MULTIPLICATIVE UNCERTAINTIES DOMINATE
  COLLIDER ONLY, NOT FIXED TARGET

METHODOLOGY: SUMMARY

- STATISTICAL ANALYSIS TOOLS NECESSARY TO COPE WITH DATA ACCURACY
- PDF UNCERTAINTIES ARE FAITHFUL, BUT NOT OPTIMAL