PARTON DISTRIBUTIONS

OR

THE ART OF SOLVING ILL-POSED PROBLEMS

Stefano Forte
Università di Milano & INFN

1st COFI Workshop

San Juan, PR, July 17, 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006

SUMMARY

PDFs NOW

- FROM PDF4LHC15 TO NEW GENERATION SETS
- THE IMPACT OF LHC DATA
- RESOLVING ISUES WITH DATA: STRANGENESS & CHARM
- DATA VS. METHODOLOGY

FITTING THE METHODOLOGY

- DATASET OPTIMIZATION?
- MACHINE LEARNING PDFs
- HYPEROPTIMIZATION
- WHAT IS PROPER LEARNING?

THEORY UNCERTAINTIES

- PDF THEORY ERRORS?
- THE THEORY COVARIANCE MATRIX
- SCALE VARIATION AND ITS VALIDATION
- PDFs WITH THEORY ERRORS AND THEIR IMPACT

THE STATE OF THE ART

CONTEMPORARY PDF TIMELINE (ONLY PUBLISHED GLOBAL)

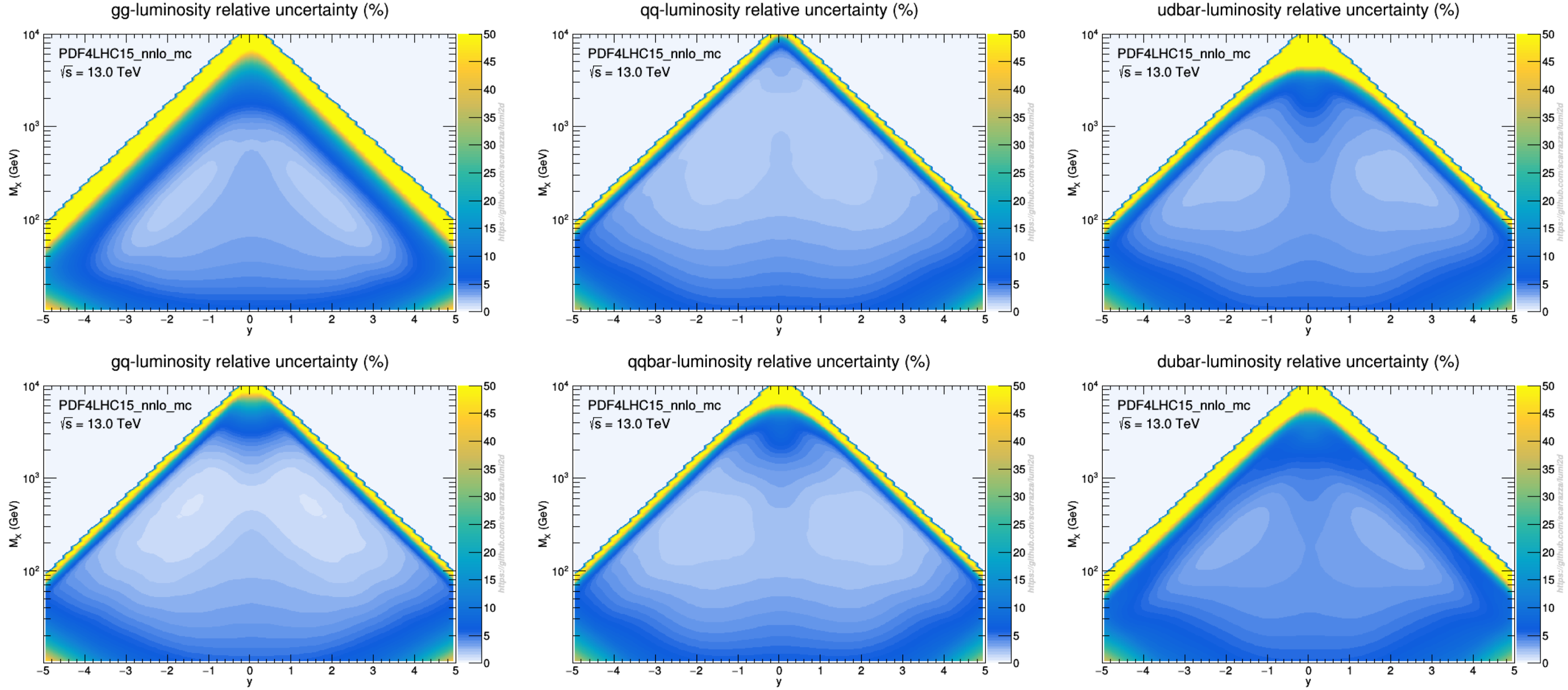
	2008		2009		2010		2011	2012		2013		2014		2015	2017	
SET / MONTH	CTEQ6.6 (02)	NNPDF1.0 (08)	MSTW (01)	ABKM09 (08)	NNPDF2.0 (02)	CT10 (NLO) (07)	NNPDF2.1 (NNLO) (07)	ABM11 (02)	NNPDF2.3 (07)	CT10 (NNLO) (02)	ABM12 (10)	NNPDF3.0 (10)	MMHT (12)	CT14 (06)	ABMP16 (01)	NNPDF3.1 (06)
F. T. DIS	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
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)

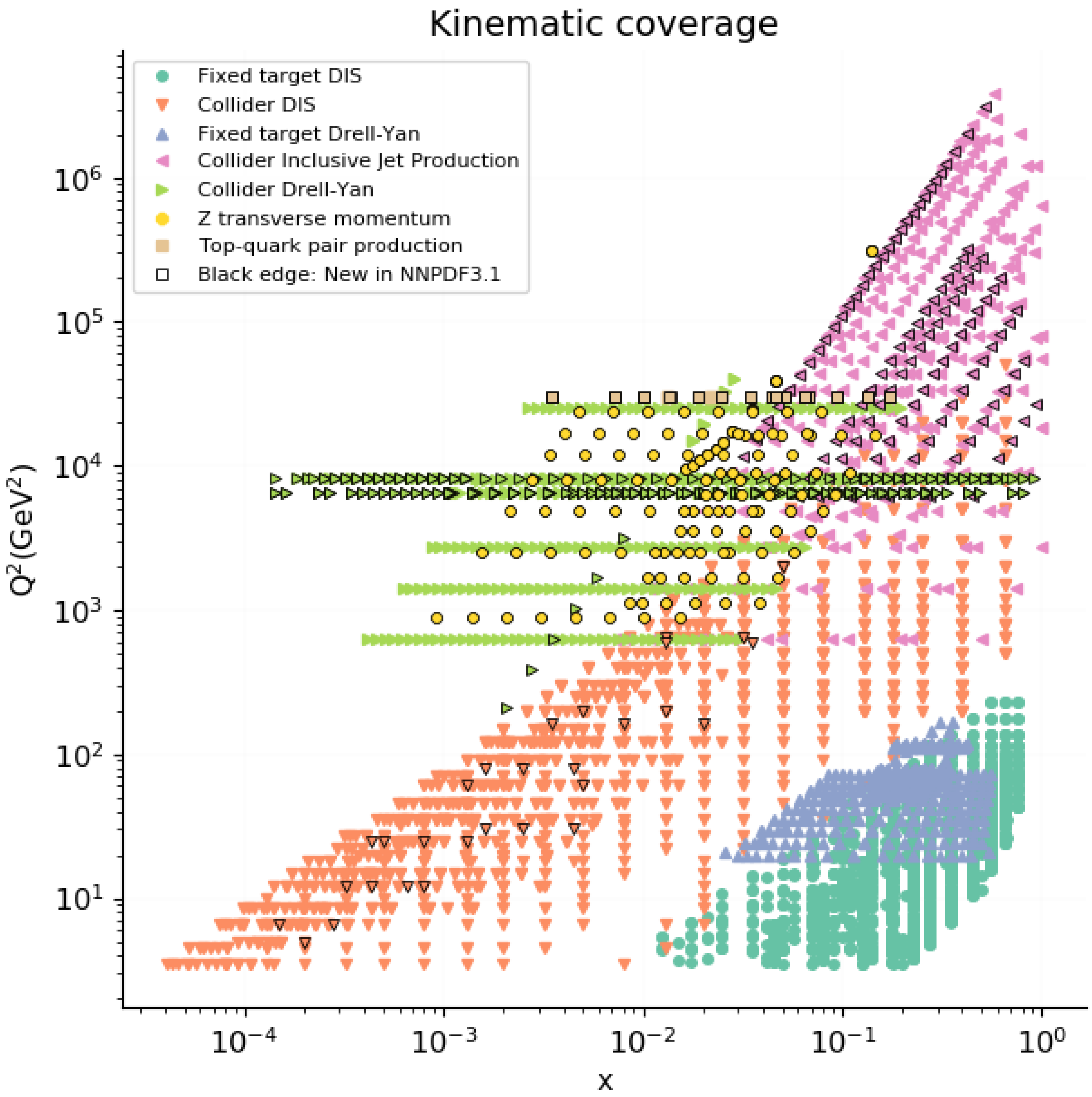
PDF4LHC15: PDF UNCERTAINTIES (NNLO)

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
- NO QUALITATIVE DIFFERENCE BETWEEN NLO AND NNLO

DATASET WIDENING
NNPDF3.0 vs NNPDF3.1



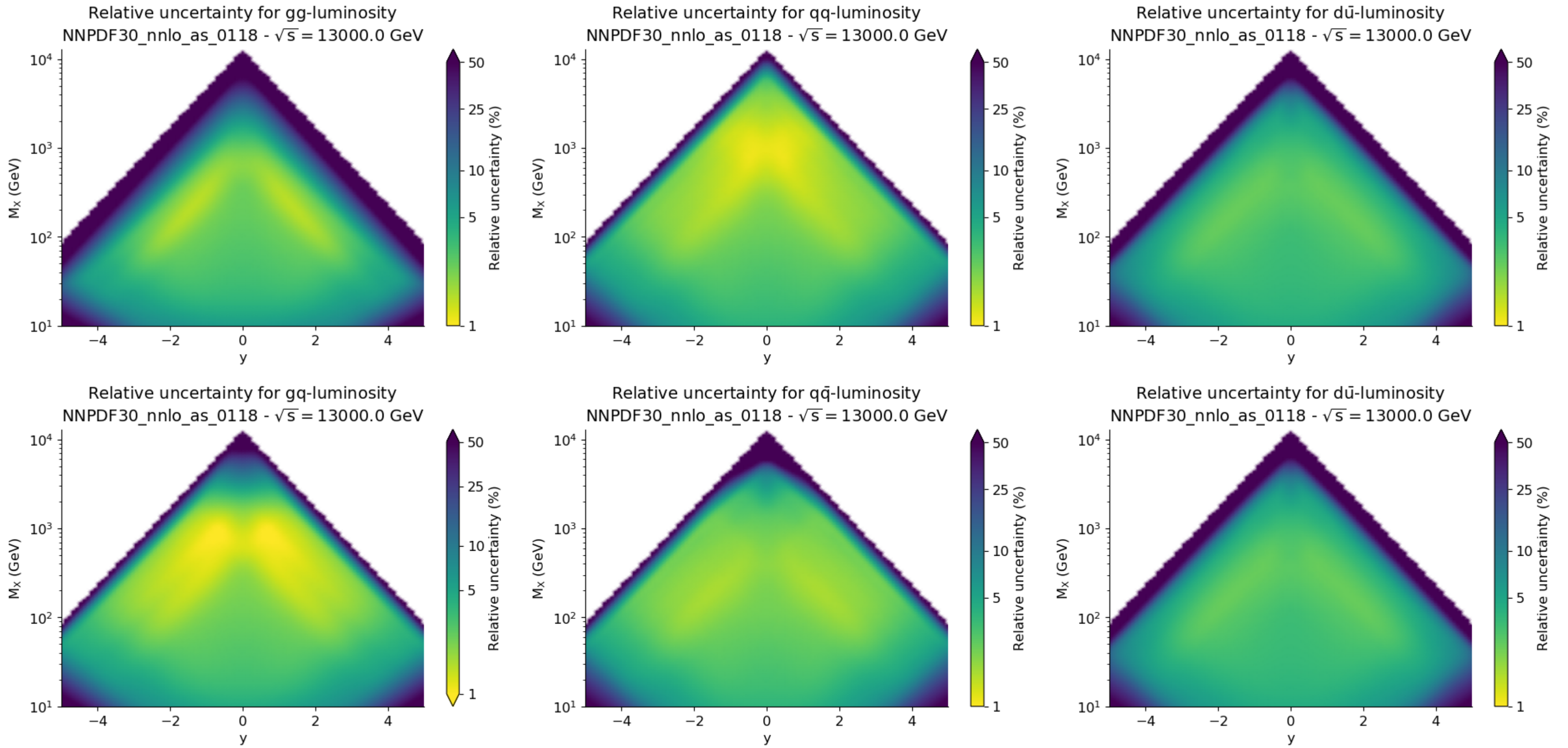
NEW DATA: (black edge)

- 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)

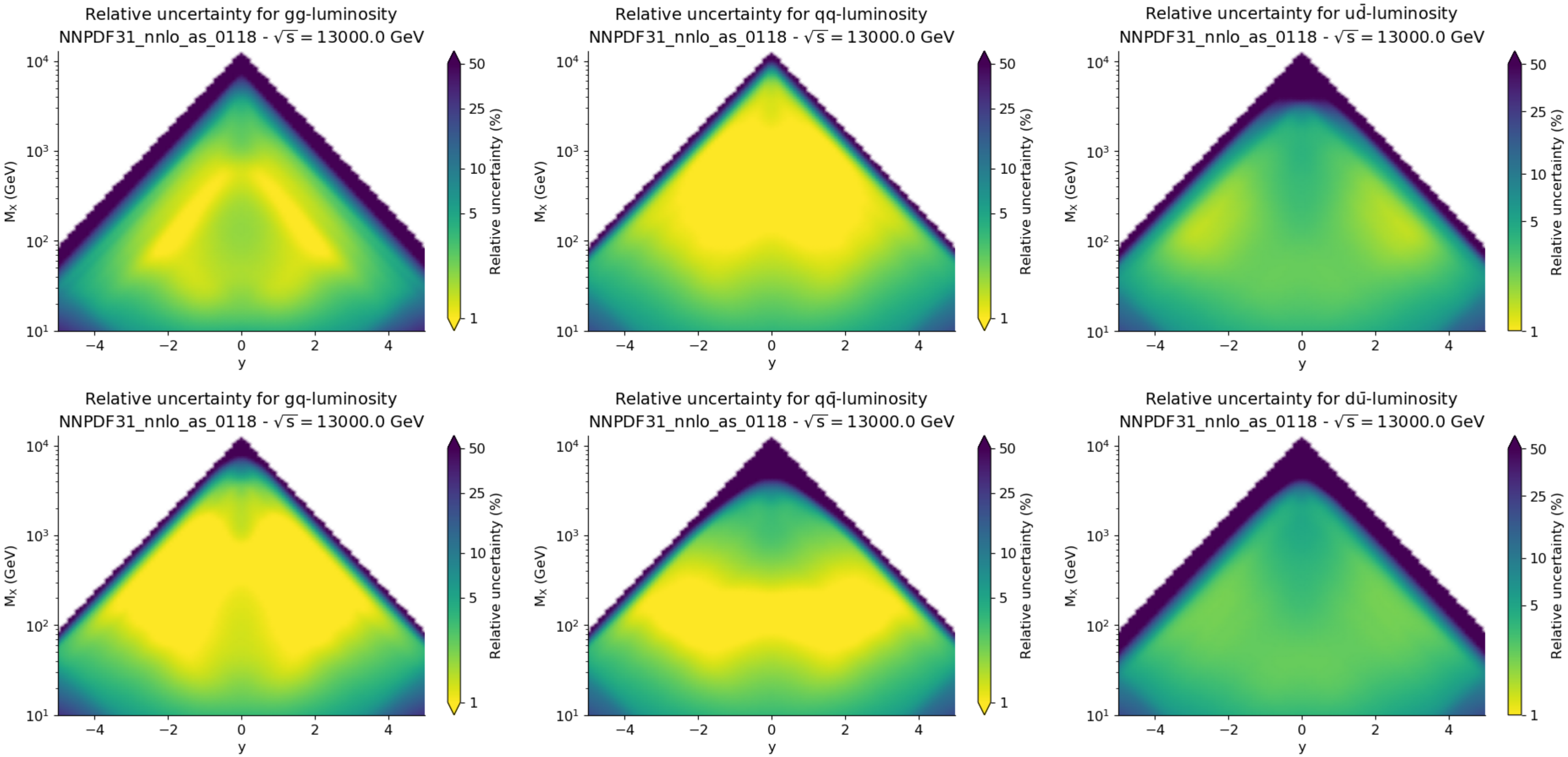
GLUON | SINGLET | FLAVORS



- 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

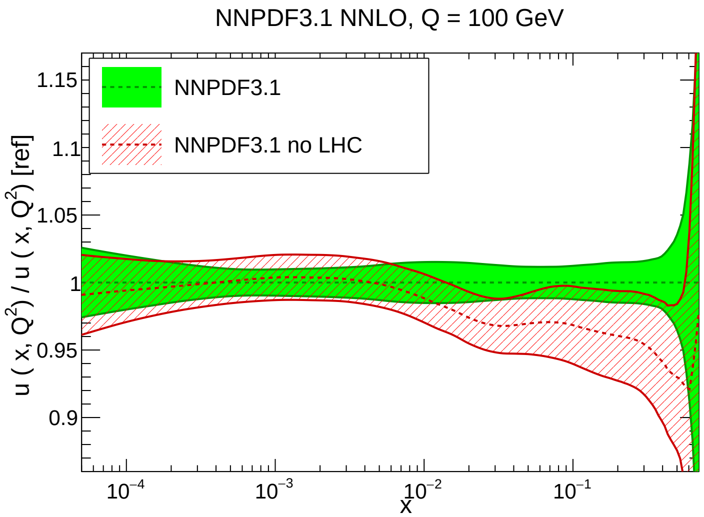


- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- TYPICAL UNCERTAINTIES IN DATA REGION $\sim 1 - 3\%$
- SWEET SPOT: VALENCE Q - G; 1% OR BELOW
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS
- NEW LHC DATA ⇒ SIZABLE REDUCTION IN UNCERTAINTIES

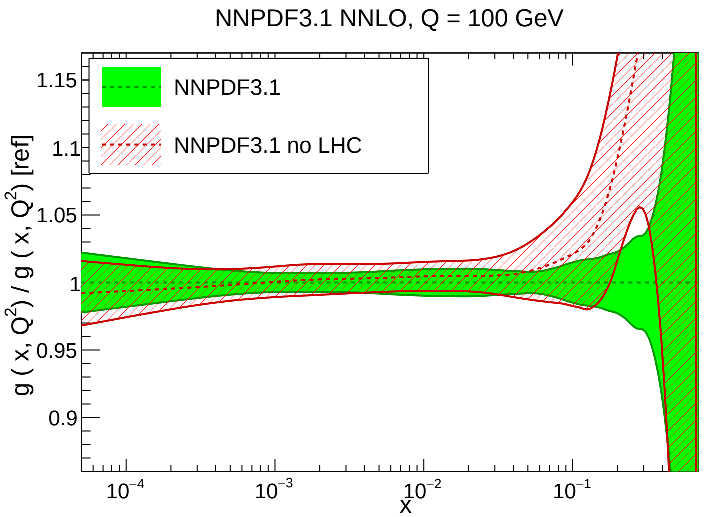
THE IMPACT OF LHC DATA

NEXT-GENERATION PDFS LARGELY DETERMINED BY LHC DATA: A FIRST!

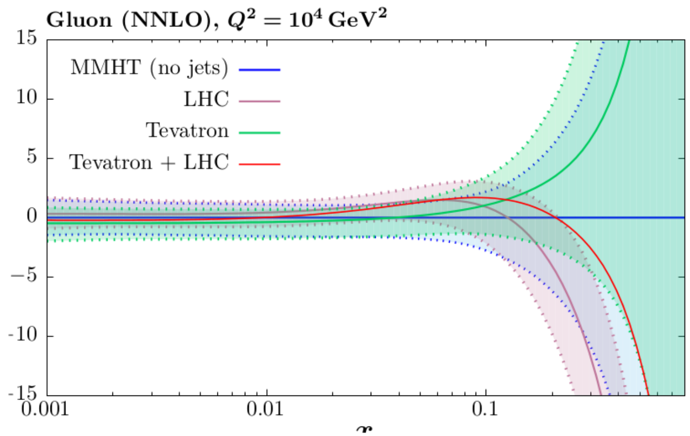
NNPDF3.1 up



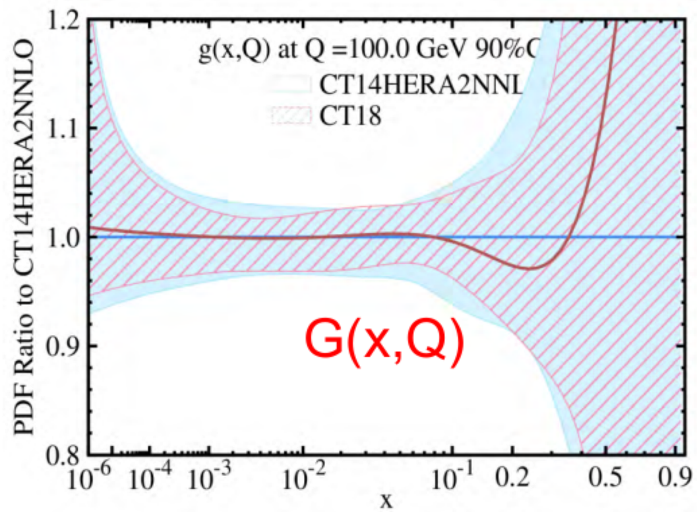
NNPDF3.1 glue



'MMHT' 19 glue (prelim., unpublished)



CT18 glue (preliminary, unpublished)



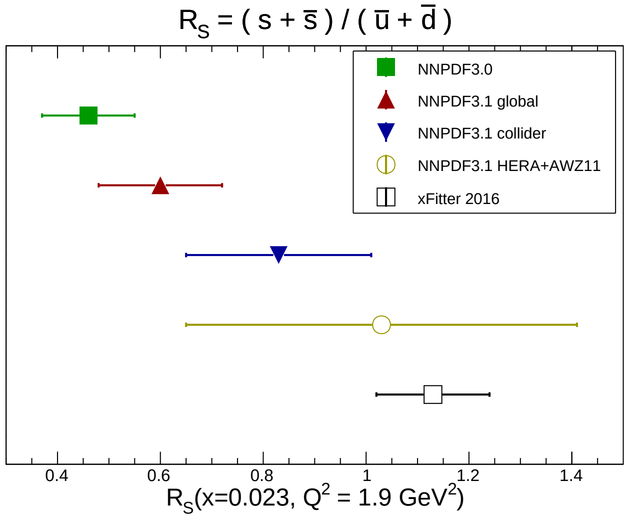
- SIGNIFICANT UNCERTAINTY REDUCTION
- MANY PDFS CHANGE BY MORE THAN ONE SIGMA
- BOTH FLAVOR SEPARATION & GLUON SIGNIFICANTLY AFFECTED

DATA VS. THEORY/METHODOLOGY
THE STRANGE PDF: DIS VS. W PRODUCTION

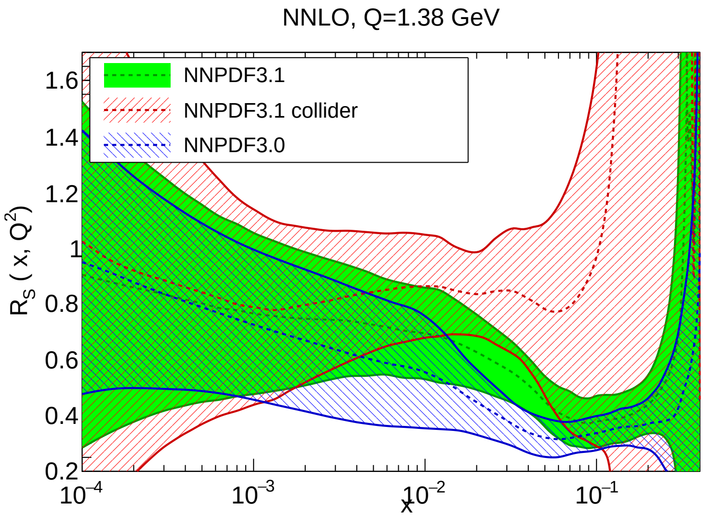
- STRANGE PDF CONTROLLED BY NEUTRINO DIS CHARM PRODUCTION + W PRODUCTION
- DIS DATA FAVOR "SUPPRESSED STRANGE" $\Rightarrow$ SMALL $R_s \equiv \frac{s+\bar{s}}{\bar{u}+\bar{d}}$
- ATLAS FAVORS ENHANCED STRANGENESS
- ATLAS IMPACT EXAGGERATED IN XFITTER ANALYSIS
- EVERYTHING CONSISTENT WITHIN UNCERTAINTIES IN GLOBAL FIT

THE STRANGENESS SUPPRESSION

XFITTER VS HERA+ATLAS VS. DIS ONLY VS ATLAS ONLY VS ALL



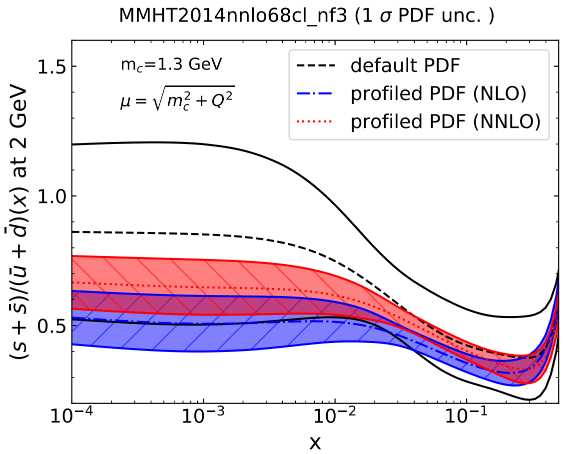
DIS ONLY VS ATLAS ONLY VS ALL



DATA VS. THEORY/METHODOLOGY
THE STRANGE PDF: DIS VS. W PRODUCTION

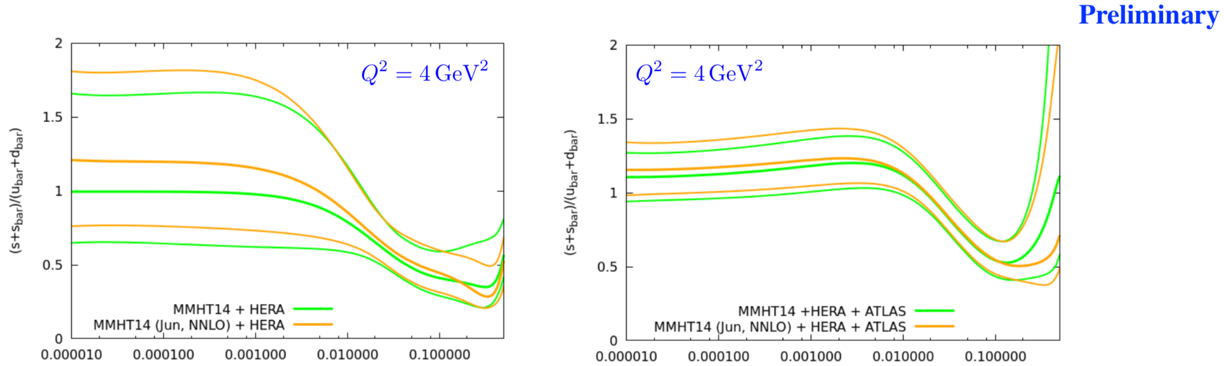
- MASSIVE CORRECTIONS TO CHARGED CURRENT DIS HITERTO INCLUDED TO NLO MASSLESS TO NNLO
- Gao, 2018 ⇒ NNLO COMPUTED
- STRANGENESS ENHANCED BY NNLO CORRECTIONS

HERAPDF +NLO CC DIS VS NNLO CC DIS



(Gao, 2108)

MMHT WITH NLO VS NNLO CC DIS



(Harland-Lang, Thorne, prelim.)

LESSONS:

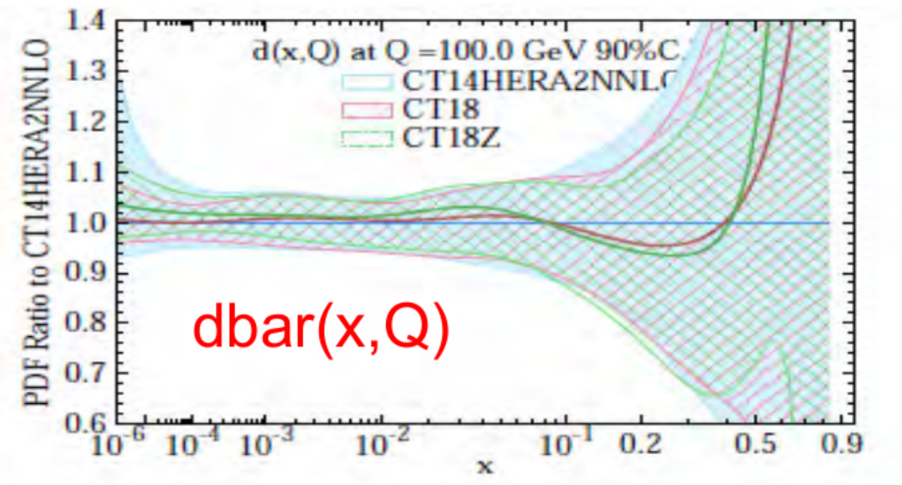
- IN A GLOBAL FIT DIFFERENT DATA ALWAYS PULL IN DIFFERENT DIRECTIONS!
- TENSIONS CAN BE RESOLVED BY BETTER THEORY

DATA VS. THEORY/METHODOLOGY
THE CHARM MASS AND TREATMENT
CT18 → CT18Z

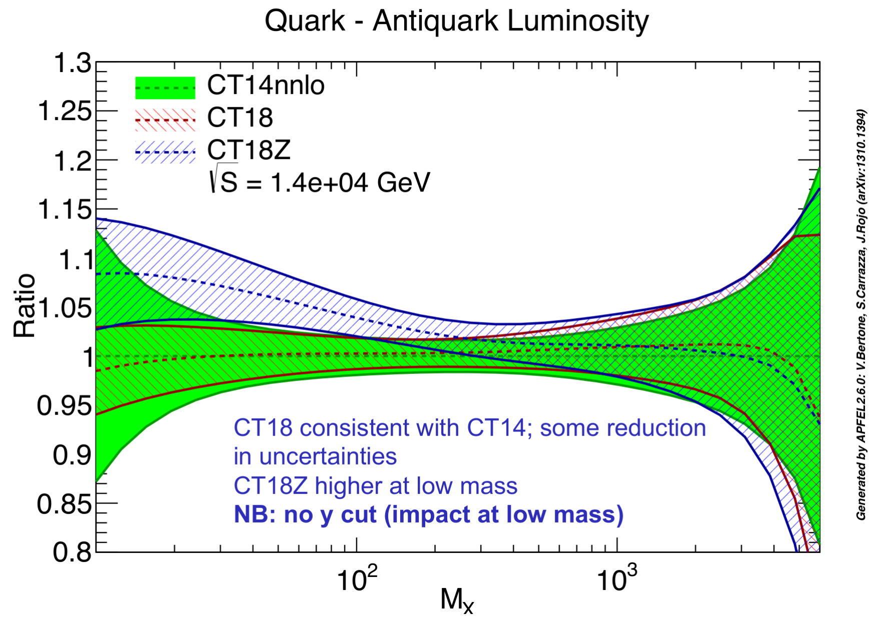
- ATLAS W AND Z 7TEV RAPIDITY INCLUDED
- CHARM MASS INCREASED
- x-DEPENDENT FACTORIZATION SCALE

CT18 VS. CT18Z (preliminary, unpublished)

DBAR PDF



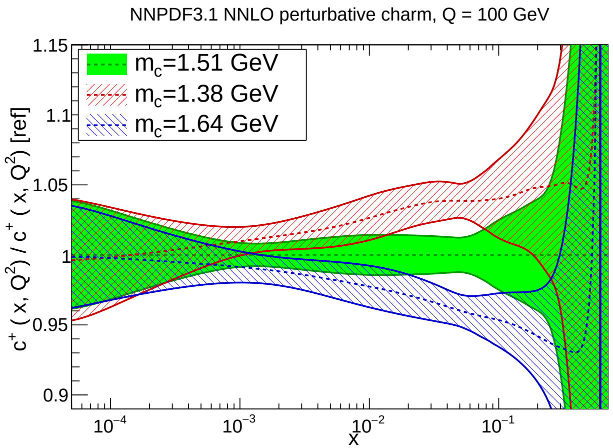
QQBAR LUMI



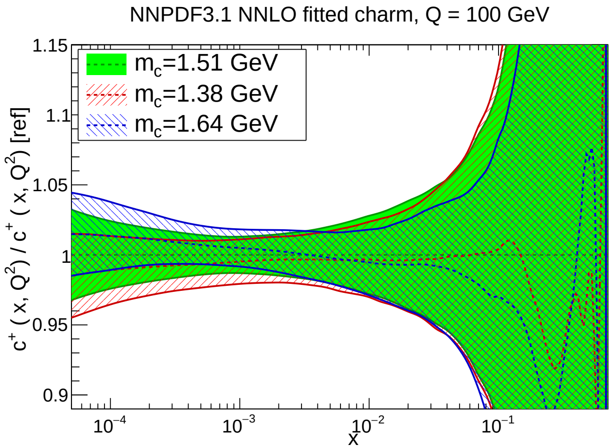
DATA VS. THEORY/METHODOLOGY
THE CHARM MASS AND TREATMENT
CHARM FROM DATA

- CHARM SHOULD NOT DEPEND STRONGLY ON CHARM MASS

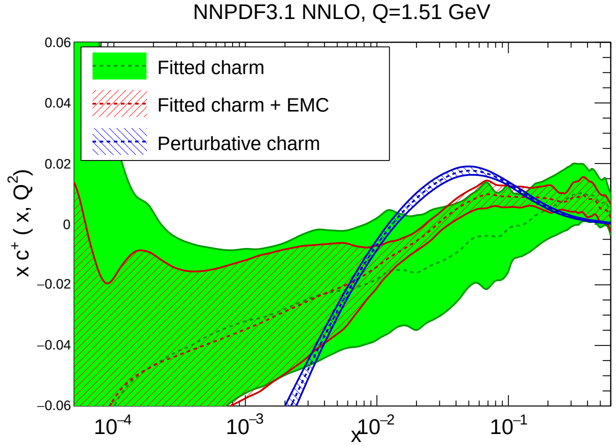
PERTURBATIVE CHARM VS m_c



FITTED CHARM VS m_c



FITTED VS. PERTURBATIVE CHARM



- ITS SHAPE SHOULD NOT BE DETERMINED BY FIRST-ORDER MATCHING (NO HIGHER NONTRIVIAL ORDERS KNOWN)

FITTED VS. PERTURBATIVE:
SUPPRESSED AT MEDIUM-SMALL x,
ENHANCED AT VERY SMALL, VERY LARGE x

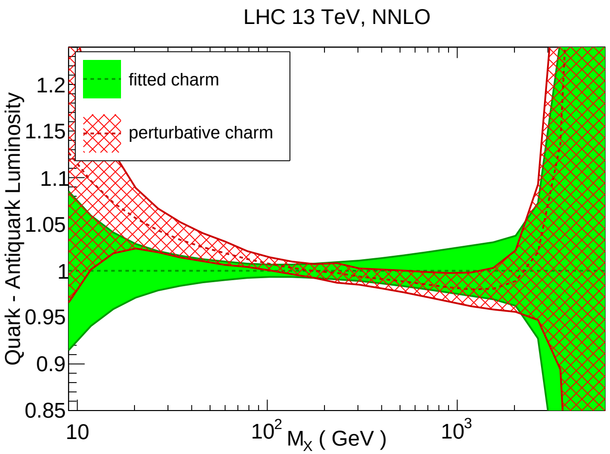
DATA VS. THEORY/METHODOLOGY

THE CHARM MASS AND TREATMENT

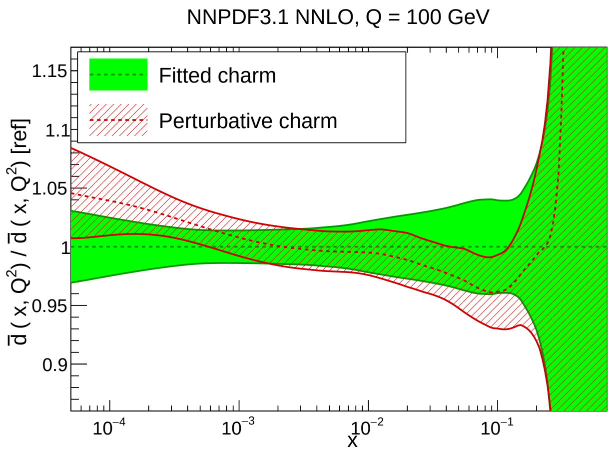
CHARM FROM DATA IMPACT ON LIGHT QUARK PDFS

FITTED VS. PERTURBATIVE CHARM

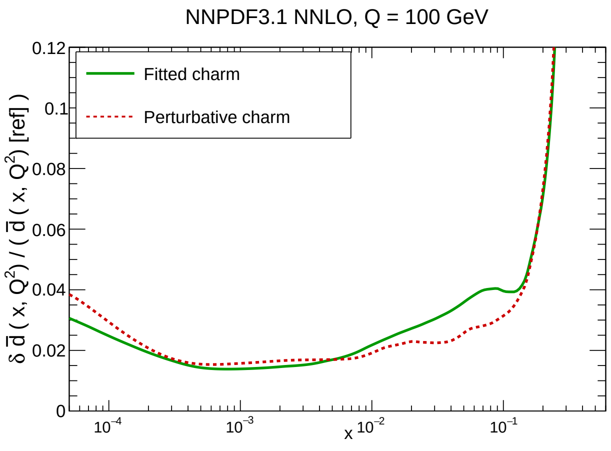
QQBAR LUMI



ANTIDOWN PDF



ANTIDOWN PDF UNCERTAINTY



- QUARK LUMI AFFECTED BECAUSE OF CHARM SUPPRESSION AT MEDIUM-x
- FLAVOR DECOMPOSITION ALTERED
- UNCERTAINTIES ON LIGHT QUARKS NOT SIGNIFICANTLY INCREASED

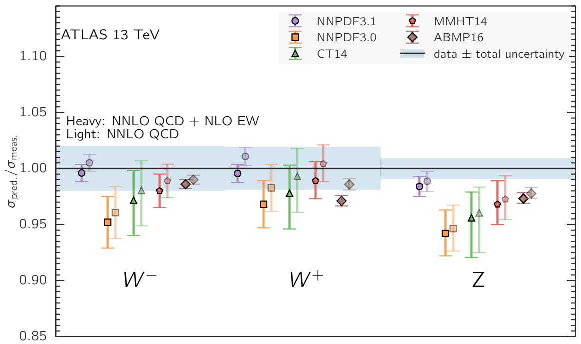
DATA VS. THEORY/METHODOLOGY

THE CHARM MASS AND TREATMENT

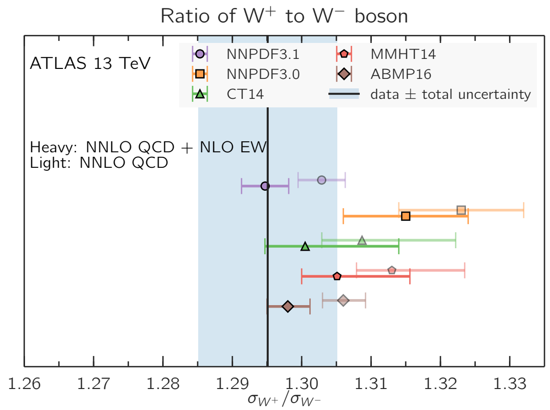
CHARM FROM DATA

IMPACT ON PHENOMENOLOGY

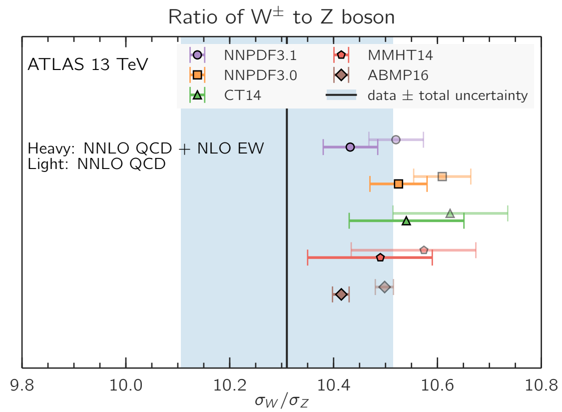
DRELL-YAN XSECTS



W^+/W^- XSECT RATIO



W/Z XSECT RATIO



- W, Z CROSS-SECTIONS AT 13 TEV IN PERFECT AGREEMENT WITH DATA THANKS TO FITTED CHARM!

LESSONS:

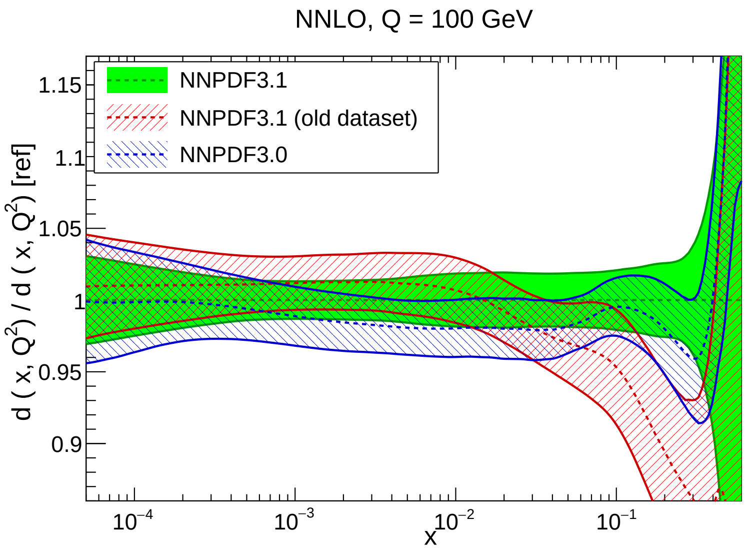
- TENSIONS CAN REVEAL METHODOLOGICAL ISSUES
- MORE LIKELY AS DATASET INCREASES, EXPERIMENTAL UNCERTAINTIES DECREASE
- RESOLVED BY MORE COMPLEX METHODOLOGY

DATA VS. METHODOLOGY

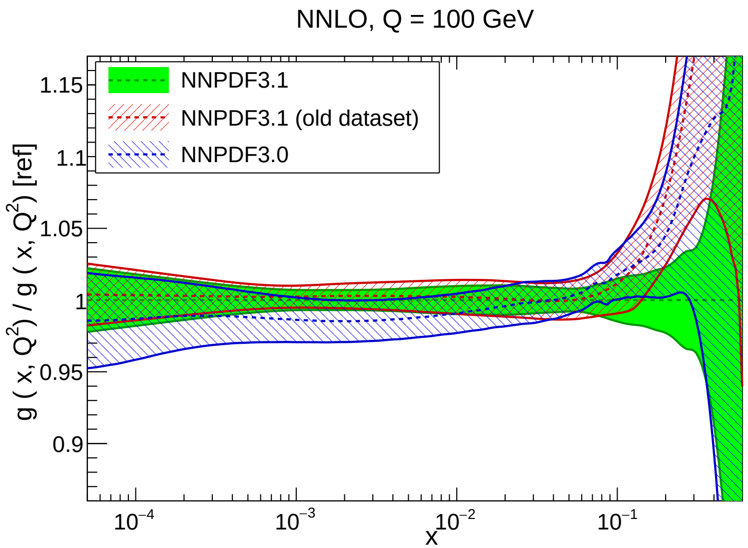
- NEW DATA ⇒ MAJOR METHODOLOGICAL CHOICES ⇒ SIGNIFICANT IMPACT
- NNPDF3.1 VS NNPDF3.0: DATA AND METHODOLOGY HAVE SIMILAR IMPACT

NNPDF3.0 VS. NNPDF3.1 VS. NNPDF3.1 W/ NNPDF3.0 DATASET

DOWN



GLUON

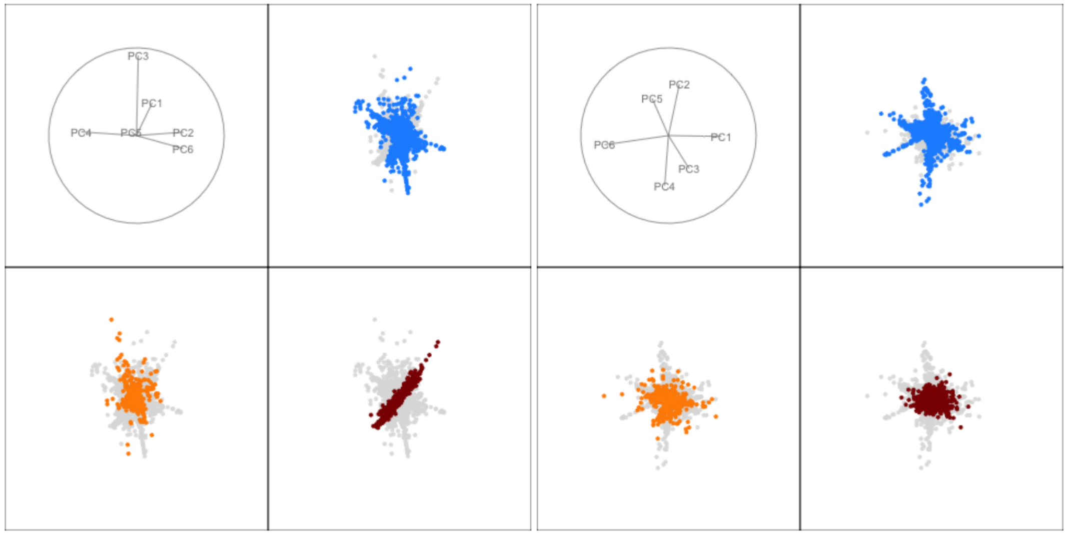


OPTIMIZING THE METHODOLOGY

DATA ⇒ THEORY ISSUES
VISUALIZATION TOOLS
TRACING THE IMPACT OF DATA

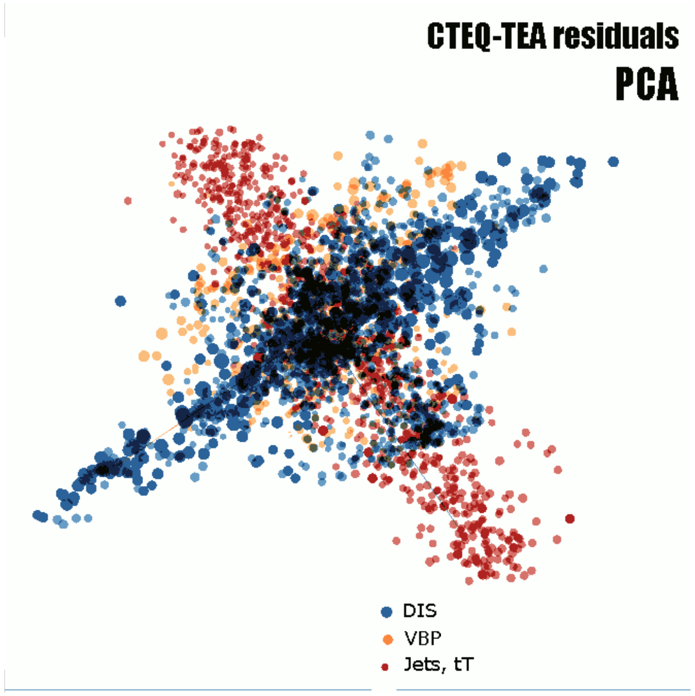
- DETERMINE THE SHIFT IN DATAPOINTS INDUCED BY CHANGES IN PDF
- CT14: 28 HESSIAN PARAMETER VARIATIONS
- VISUALIZE SHIFT VECTORS FOR 4000 DATAPOINTS
  - BY PCA
  - BY TOUR

TOUR OF CT14



(Cook, Laa, Valencia, 2018)

PCA PROJECTION OF CT14:
PDFSENSE

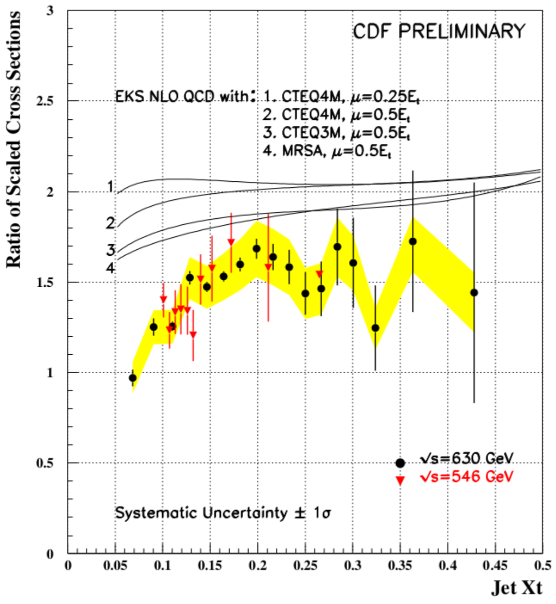


(Wang et al., 2018)

LESSONS FROM THE PAST: TWENTY YEARS BACK....

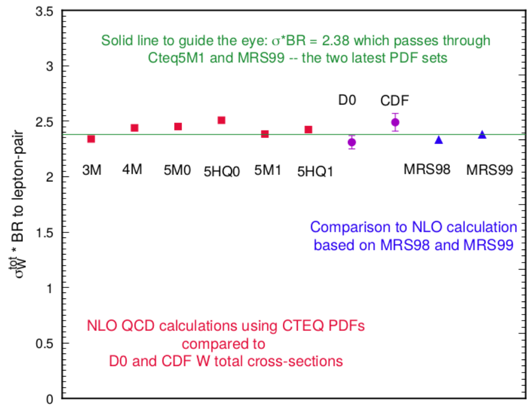
- NO PDF UNCERTAINTIES
- UNCERTAINTY ⇔ PDF SET VARIATION
- METOHDOLOGY ↔ AGREEMENT WITH DATA

CDF JETS



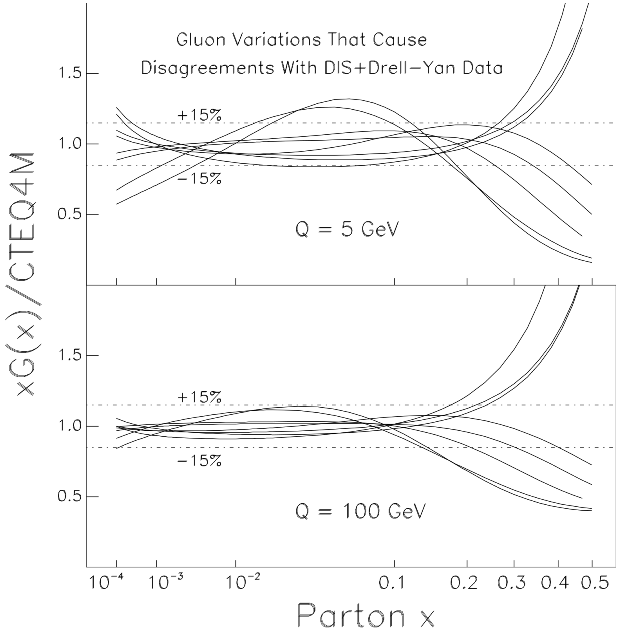
(Mangano, 1999)

W PRODUCTION



(Huston, Tung et al, 2000)

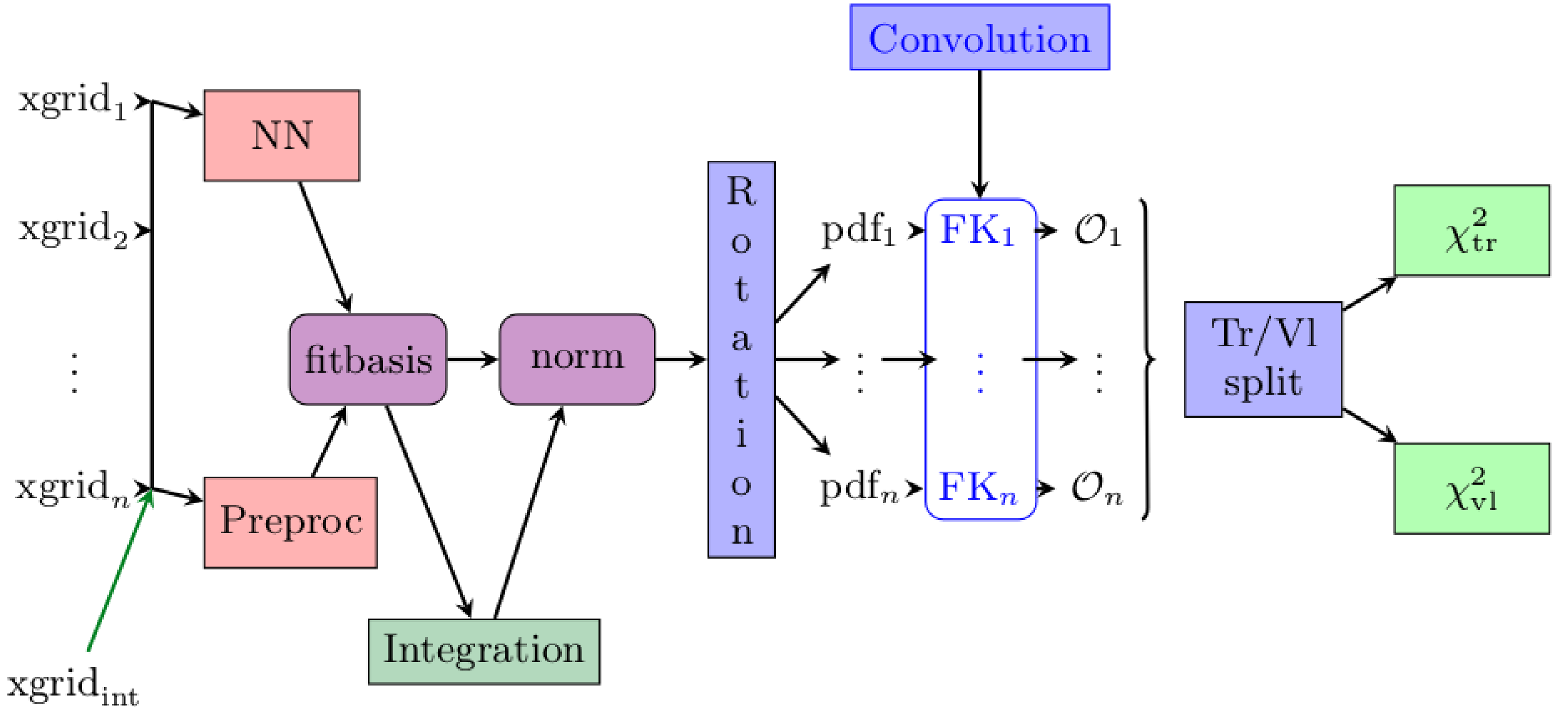
GLUON UNCERTAINTY



(Huston et al., 1998)

IS IT POSSIBLE TO DO BETTER?

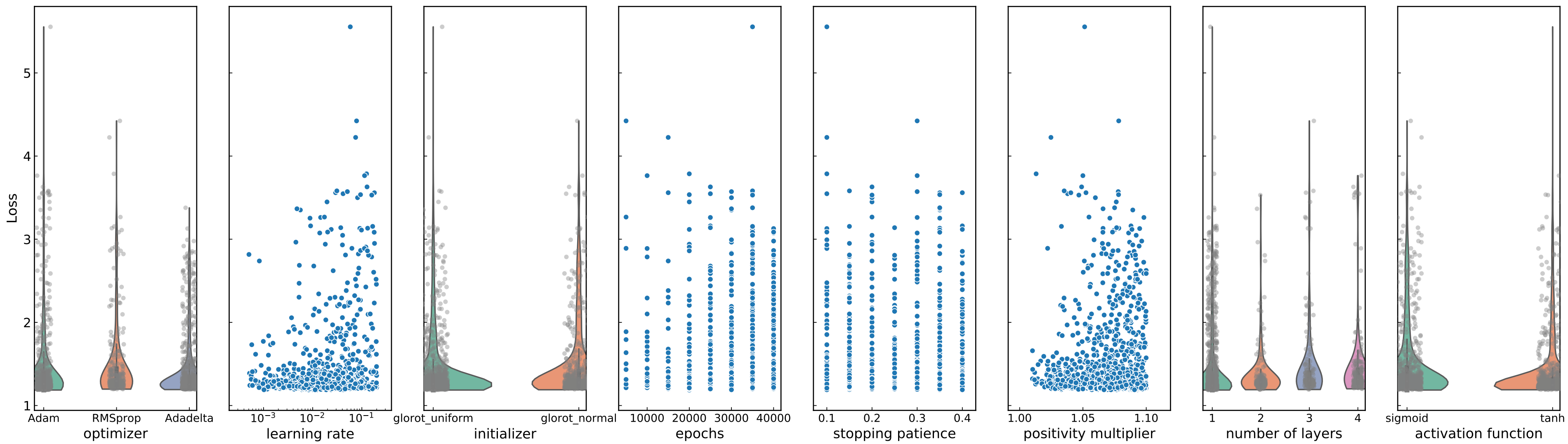
FITTING THE METHODOLOGY
THE N3FIT PROJECT



(Carrazza, Cruz-Martinez, 2019)

- PYTHON-BASED KERAS + TENSORFLOW FRAMEWORK
- EACH BLOCK INDEPENDENT LAYER
- CAN VARY ALL ASPECT OF METHODOLOGY
- GRADIENT DESCENT DETERMINISTIC MINIMIZATION
- ONE SINGLE NEURAL NET FOR ALL PDFs

FITTING THE METHODOLOGY
HYPEROPTIMIZATION SCANS

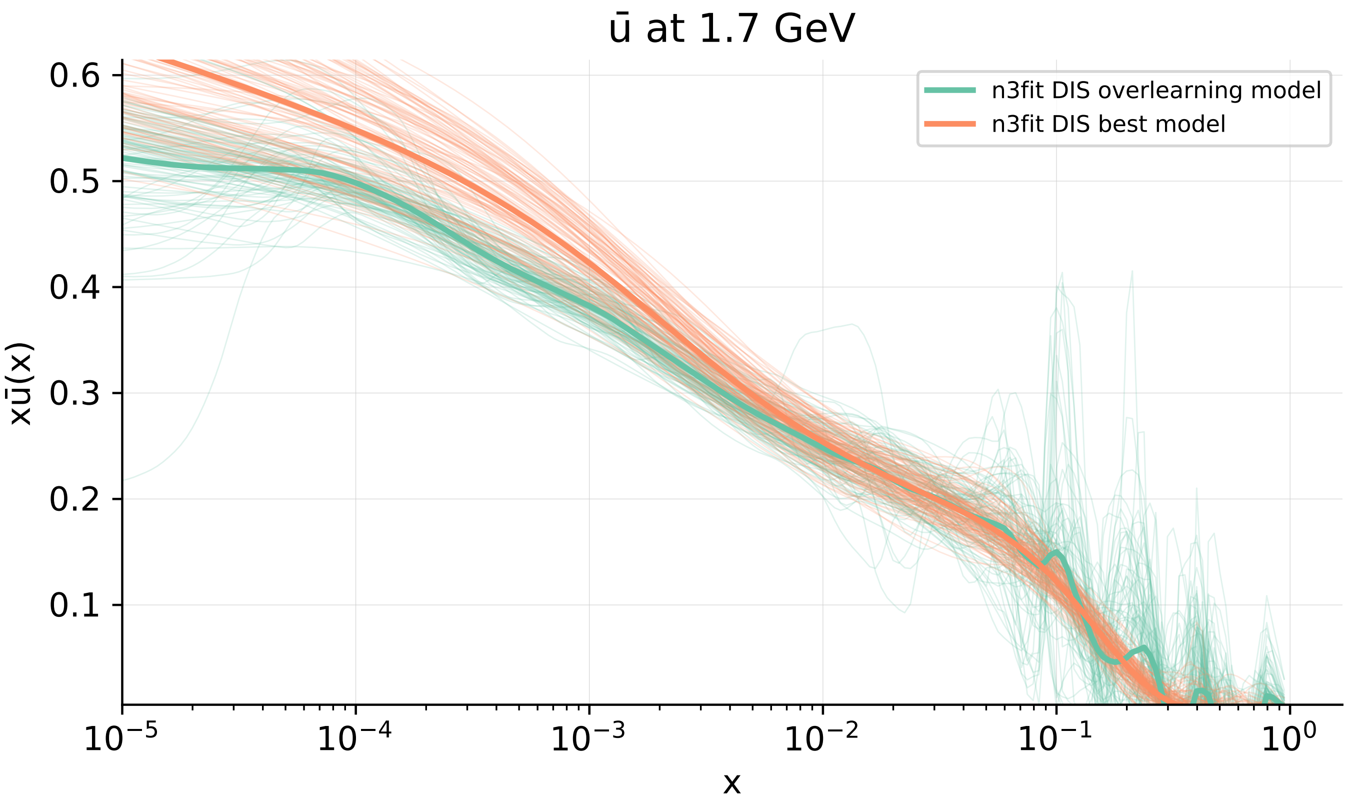


HYPEROPT PARAMETERS

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 (*)

- SCAN PARAMETER SPACE
- OPTIMIZE FIGURE OF MERIT
- BAYESIAN UPDATING

FITTING THE METHODOLOGY
OVERLEARNING VS. PROPER LEARNING



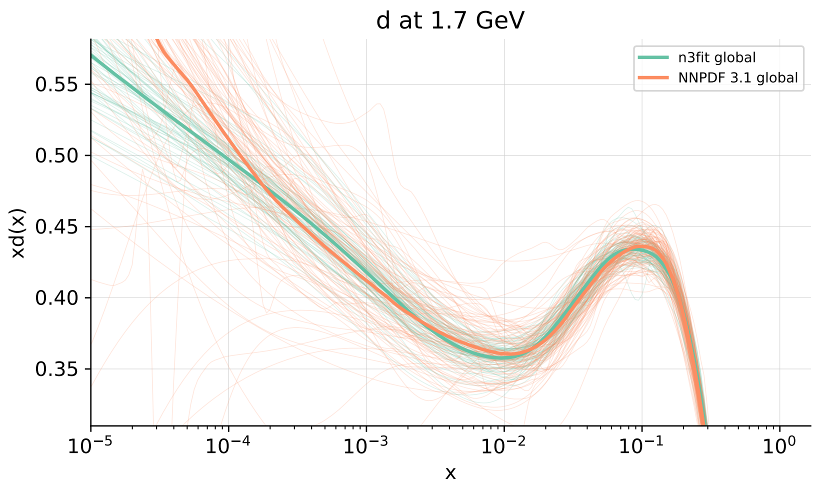
- CROSS-VALIDATION $\Rightarrow$ RANDOMLY DIVIDE DATA INTO TRAINING AND VALIDATION SETS
- MINIMIZE ON TRAINING; OPTIMIZE ON VALIDATION
- NO OVERLEARNING $\Rightarrow$ NOISE NOT LEARNT
- OVERFITTING $\Rightarrow$ $\chi^2_{\rm train} << \chi^2_{\rm valid}$!! & WIGGLY PDFS
- CORRELATIONS BETWEEN DATA IN A SET

PROPER LEARNING VS. OVERLEARNING
THE TEST SET METHOD

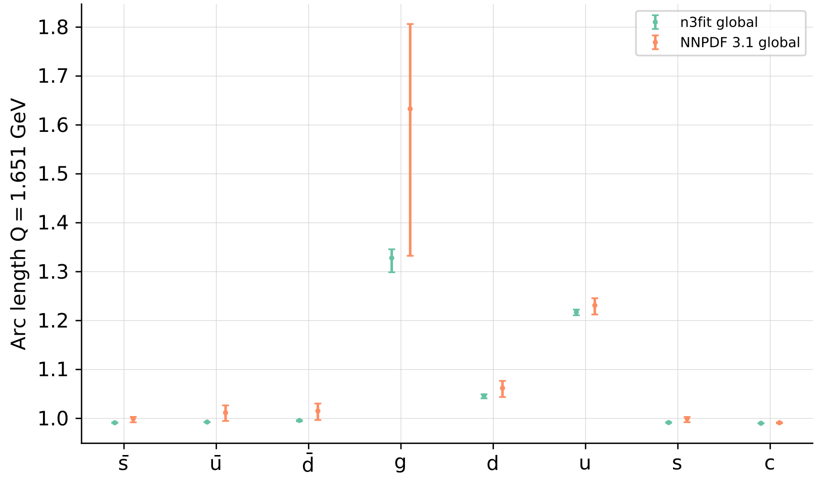
- NEED A COMPLETELY UNCORRELATED "TEST" SET
- OPTIMIZE ON WEIGHTED AVERAGE OF VALIDATION AND TEST
  $\Rightarrow$ NO OVERLEARNING

OPTIMIZED PDFs
n3fit VS NNPDF3.1

d QUARK



ARCLENGTHS



- NO OVERFITTING
- COMPARED TO NNPDF3.1
  - UNCERTAINTIES SOMEWHAT REDUCED
  - GREATER STABILITY $\Rightarrow$ FEWER REPLICAS FOR EQUAL ACCURACY

FITTING THE METHODOLOGY

WHAT IS "PROPER LEARNING"?

- FORECASTNG AN UNKNOWN TRUTH $\Rightarrow$ WHAT IS "OPTIMAL"?
- ARE STATISTICAL PROPERTIES OF THE ENSEMBLE RELEVANT?
- SHOULD THEORY PREJUDICE PLAY A ROLE?
- CLOSURE TESTING?
- REINFORCEMENT LEARNING? $\Rightarrow$ STAY TUNED!

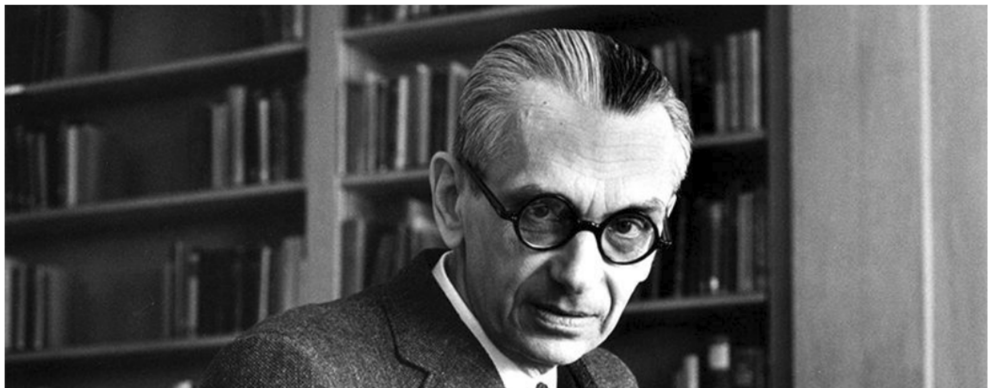
MENU nature International journal of science

NEWS · 08 JANUARY 2019

Machine learning leads mathematicians to unsolvable problem

Simple artificial-intelligence problem puts researchers up against a logical parc by famed mathematician Kurt Gödel.

Davide Castelvecchi



THEORY UNCERTAINTIES

MISSING HIGHER ORDER UNCERTAINTY ON FACTORIZED OBSERVABLES

$$\sigma = \hat{\sigma} \otimes f \otimes f$$

schematically

$$\sigma(M_w^2) = \hat{\sigma}(M_w^2)\left[\Gamma(M_w^2, Q^2)F_2(Q^2)\right]^2;\ \Gamma(M_w^2, Q^2) = \exp \int_{Q^2}^{M_W^2} \frac{d\alpha}{\beta(\alpha)}\gamma(\alpha)$$

- HADRONIC XSECT= PARTONIC XSEC TIMES PDFS (CONVOLUTION)
- PDFS ARE A PROXY FOR ANOTHER PROCESS (DIS)
- MUST EVOLVE BETWEEN TWO PROCESSES

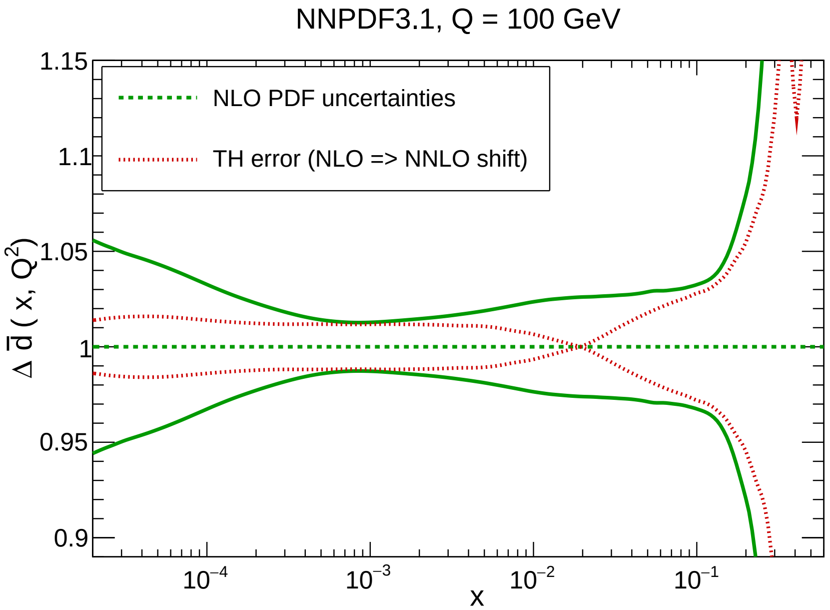
SOURCES OF MHOU UNCERTAINTY

- MHOU IN THE "DRELL-YAN" XSECT ⇒ STANDARD SCALE VARN.
- MHOU IN THE STRUCTURE FUNCTIONS ⇒ TH. UNCERTAINTY ON PDFS (1)
- MHOU IN THE EVOLUTION ⇒ TH. UNCERTAINTY ON PDFS (2)

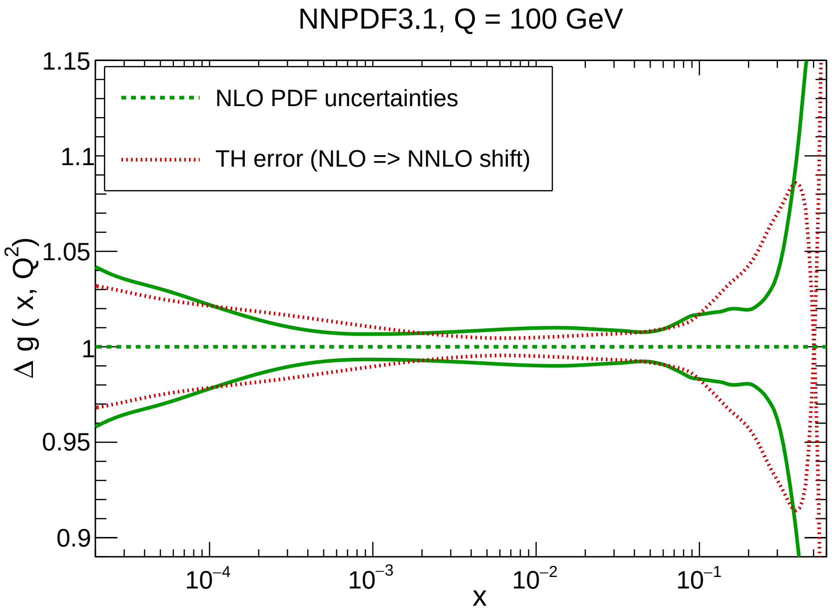
THE MISSING HIGHER ORDER UNCERTAINTY ON PDFs
HOW BIG IS IT?

NLO-NNLO SHIFT VS. NLO PDF UNCERTAINTY (NNPDF3.1)

ANTIDOWN



GLUON



- TODAY: NLO PDF & MHOU UNCERTAINTIES COMPARABLE
- NEAR FUTURE: SHOULD WE WORRY ABOUT NNLO MHOU?

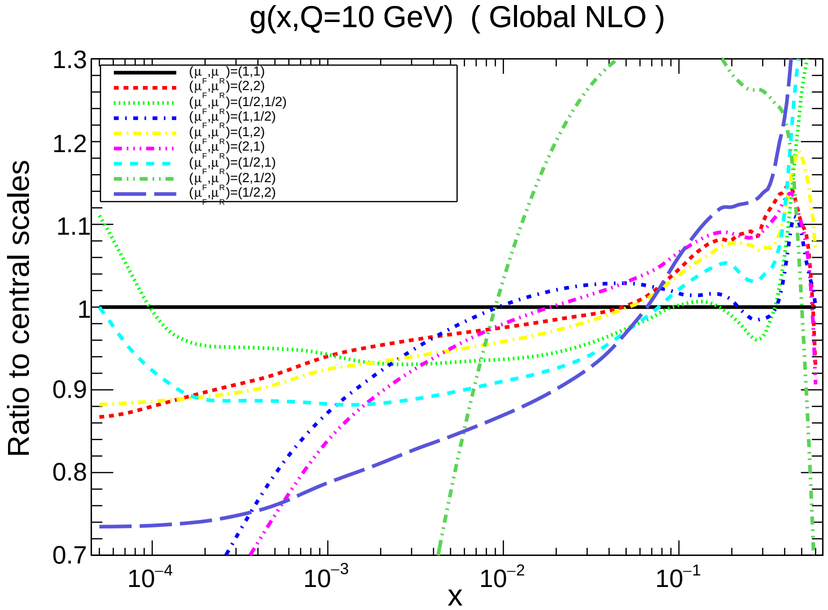
THE MISSING HIGHER ORDER UNCERTAINTY ON PDFs
CAN WE ESTIMATE IT?
SCALE VARIATION IN PDF FITTING

NAIVE IDEA FOR PDF MHOU ESTIMATE

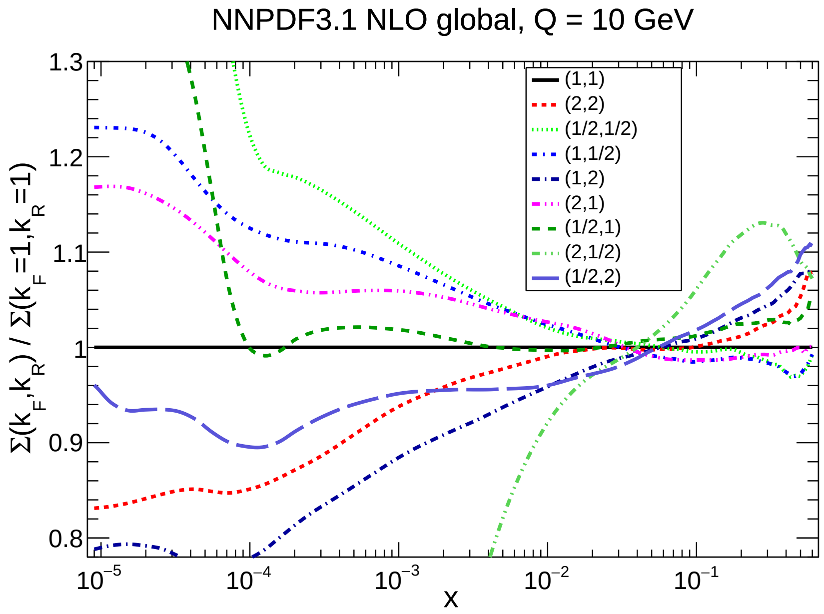
- PERFORM FIT WITH VARIOUS SCALE CHOICES
- TAKE ENVELOPE OF RESULTS
- 7-POINT ⇒ OK!; 9-POINT ⇒ UNSTABLE!
- RESULTS DEPEND STRONGLY ON THE CHOICE OF ENVELOPE

FITS WITH SCALE VARIATION

GLUON



QUARK SINGLET



THE THEORY COVARIANCE MATRIX

(NNPDF, 2019)

- ASSOCIATE MHOU TO NUISANCE PARAMETER $\Rightarrow$ THEORY COVARIANCE MATRIX S_{ij}
- $S_{ij} = \frac{1}{N}\sum_k \left(T_i^{(k)} - T_i^{(0)}\right)\left(T_j^{(k)} - T_j^{(0)}\right)$
  $\left(T_i^{(k)} - T_i^{(0)}\right)$: k-TH SHIFT OF i-TH DATAPOINT ABOUT CENTRAL PREDICTION $T_i^{(0)}$.
- SHIFT: GUESS FOR POSSIBLE MHO TERMS $\Rightarrow$ SCALE VARIATION

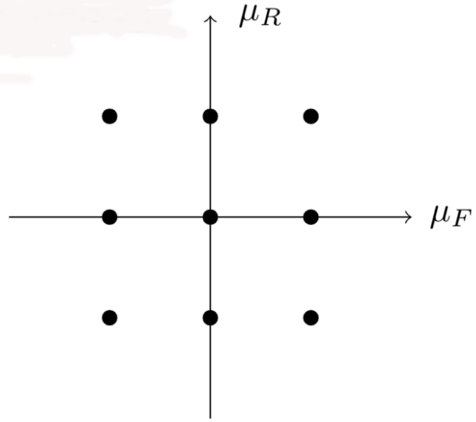
SCALE VARIATION

EXPERIMENTS AND PROCESSES

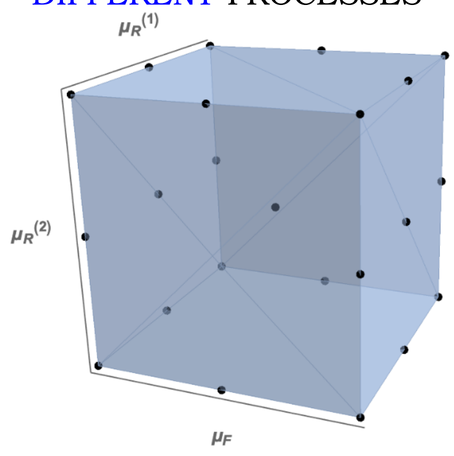
Process Type	Datasets
DIS NC	NMC, SLAC, BCDMS, HERA NC
DIS CC	NuTeV, CHORUS, HERA CC
DY	CDF, D0, ATLAS, CMS, LHCb (y, p_T, M_{ll})
JET	ATLAS, CMS inclusive jets
TOP	ATLAS, CMS total+differential cross-sections

NINE-POINT SCALE VAR.

SAME PROCESS



DIFFERENT PROCESSES



- CLASSIFY DATA INTO PROCESSES
- PICK A SET OF SCALE VARIATIONS
- DECIDE HOW TO CORRELATE SCALE VARIATION BETWEEN DIFFERENT PROCESSES
- RENORMALIZATION $\Rightarrow$ MATRIX ELEMENT; FACTORIZATION $\Rightarrow$ EVOLUTION

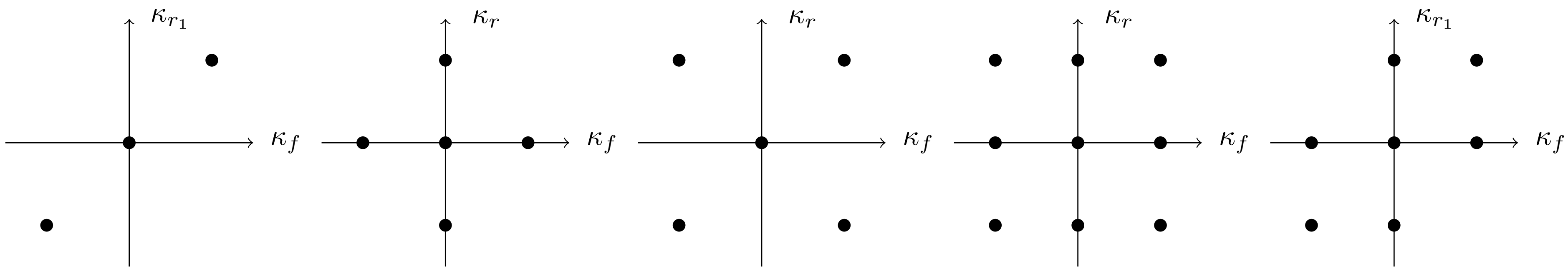
THE THEORY COVARIANCE MATRIX:
FACTORIZATION VS RENORMALIZATION SCALE

Scale	MHOU	'Traditional' name	'Modern' name[PDG]
μ_r	in hard xsec	—	renormalization scale
μ_f	in PDF evolution	renormalization scale	factorization scale
$\widetilde{\mu}$	in physical xsec	factorization scale	scale of the process

- Ren: $\mu_r \Rightarrow$ MHOU IN HARD CROSS SECTION
- Fact: $\mu_f \Rightarrow$ MHOU IN ANOMALOUS DIMENSION

PRESCRIPTIONS

- **3 POINT**: $\widetilde{\mu} = \mu_r = \mu_f$ UNCORRELATED BETWEEN PROCESSES
- **5 POINT**, **$\overline{5}$ POINT**, **9 POINT**: μ_r μ_f VARIED INDEPENDENTLY, μ_r UNCORRELATED, μ_f CORRELATED
- **7 POINT**: $\widetilde{\mu}$ ADDED TO 5 POINT



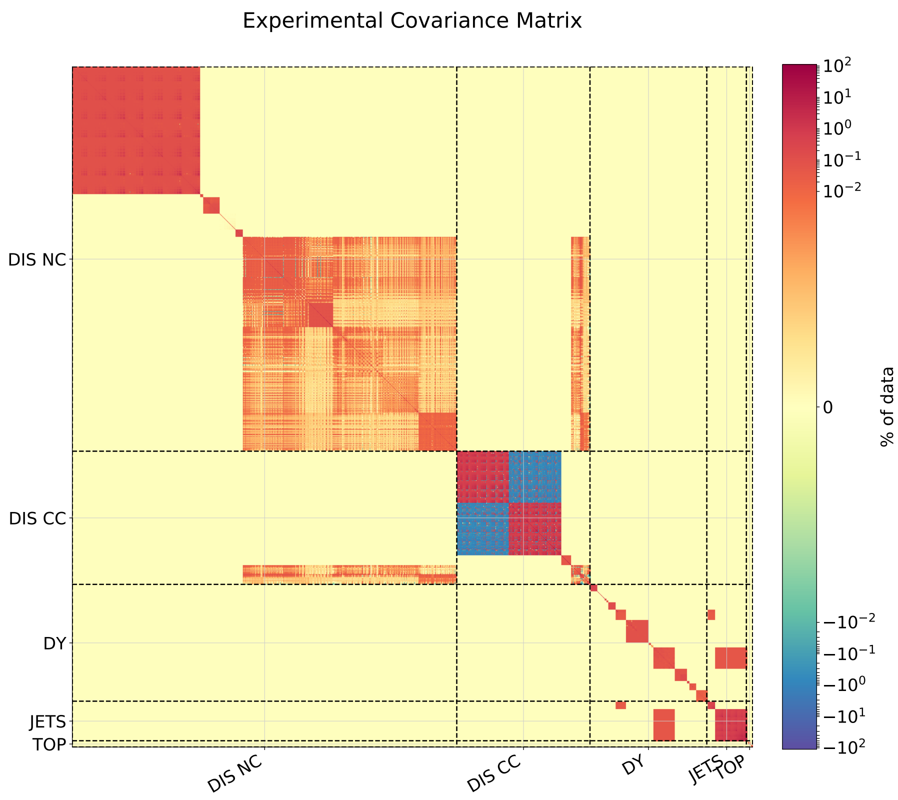
9 point, off-diagonal: $S^{(9\mathrm{pt})}_{i_1 j_2} = \frac{1}{24}\{2(\Delta^{+0}_{i_1} + \Delta^{++}_{i_1} + \Delta^{+-}_{i_1})(\Delta^{+0}_{j_2} + \Delta^{++}_{j_2} + \Delta^{+-}_{j_2}) + 2(\Delta^{-0}_{i_1} + \Delta^{-+}_{i_1} + \Delta^{--}_{i_1})(\Delta^{-0}_{j_2} + \Delta^{-+}_{j_2} + \Delta^{--}_{j_2})\} + 3(\Delta^{0+}_{i_1} + \Delta^{0-}_{i_1})(\Delta^{0+}_{j_2} + \Delta^{0-}_{j_2})\}.$

THE THEORY COVARIANCE MATRIX: CORRELATIONS

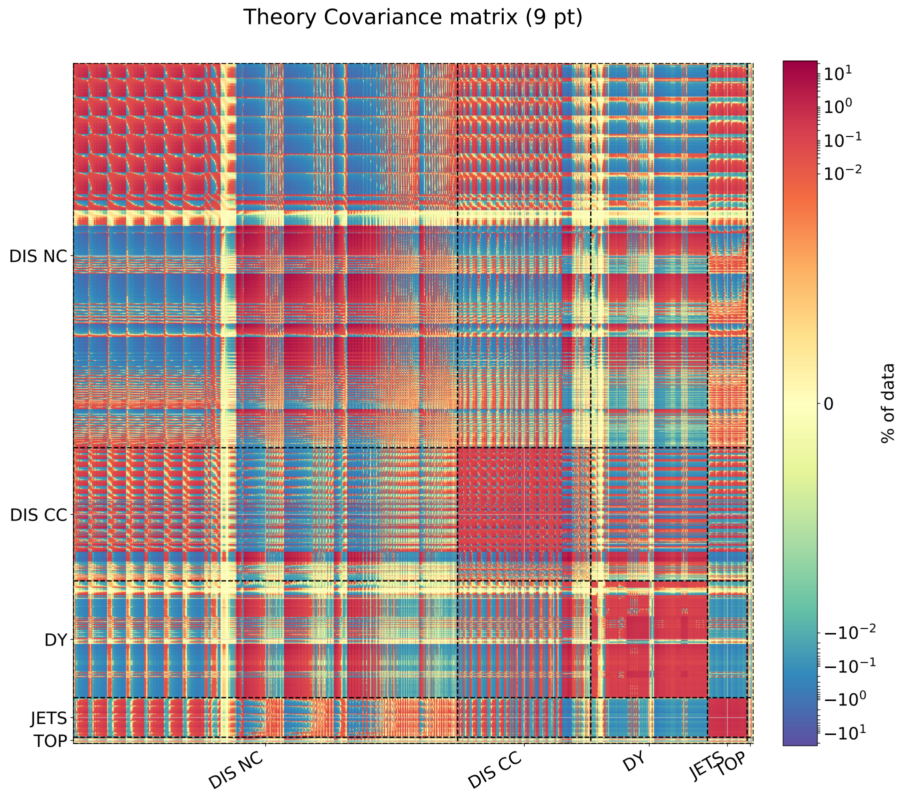
- INDEPENDENT NUISANCE PARAMETERS $\Rightarrow$ TH. AND EXP. ERRORS COMBINE IN QUADRATURE
  $\chi^2 = \sum_{i,j=1}^{N_{\rm dat}} \left(D_i - T_i^{(0)}\right) [S+C]^{-1}_{ij} \left(D_i - T_i^{(0)}\right)$
- REN. SCALE $\Rightarrow$ CORRELATIONS INDUCED BETWEEN EXPERIMENTALLY UNRELATED MEASUREMENTS OF SAME PROCESS
- FACT. SCALE $\Rightarrow$ CORRELATIONS INDUCED BETWEEN DIFFERENT PROCESSES

THE COVARIANCE MATRIX

EXPERIMENT



THEORY (9 PT)

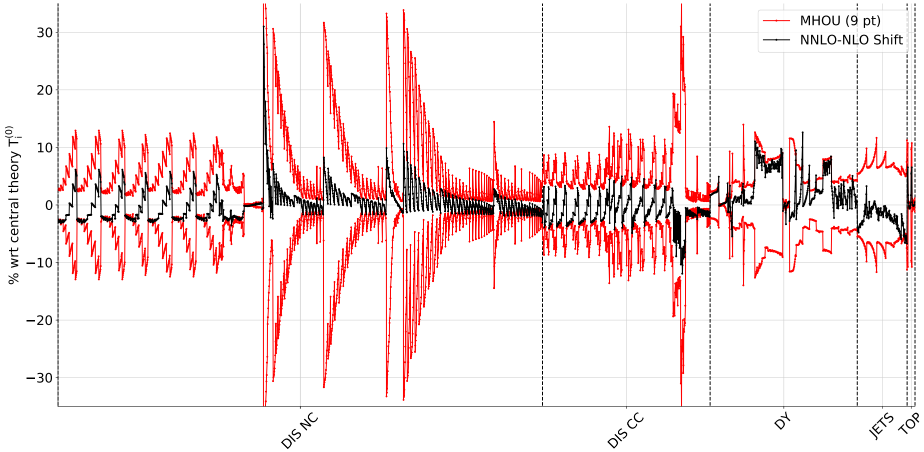


THE THEORY COVARIANCE MATRIX: VALIDATION

- COMPARE NLO THEORY COVMAT TO OBSERVED NLO-NNLO SHIFTS
- DETERMINE EIGENVECTORS e_i OF COVMAT $\Rightarrow$ 28 EVECS FOR 9PT, FIVE PROCESSES
- DETERMINE VECTOR OF SHIFTS δ
- DETERMINE PROJECTION OF δ IN SUBSPACE SPANNED BY e_i: IS IT CONTAINED IN IT?
- DETERMINE SIZE δ_i OF PROJECTIONS OF δ ALONG e_i: ARE THEY OF COMPARABLE SIZE?

9PT PRESCRIPTION

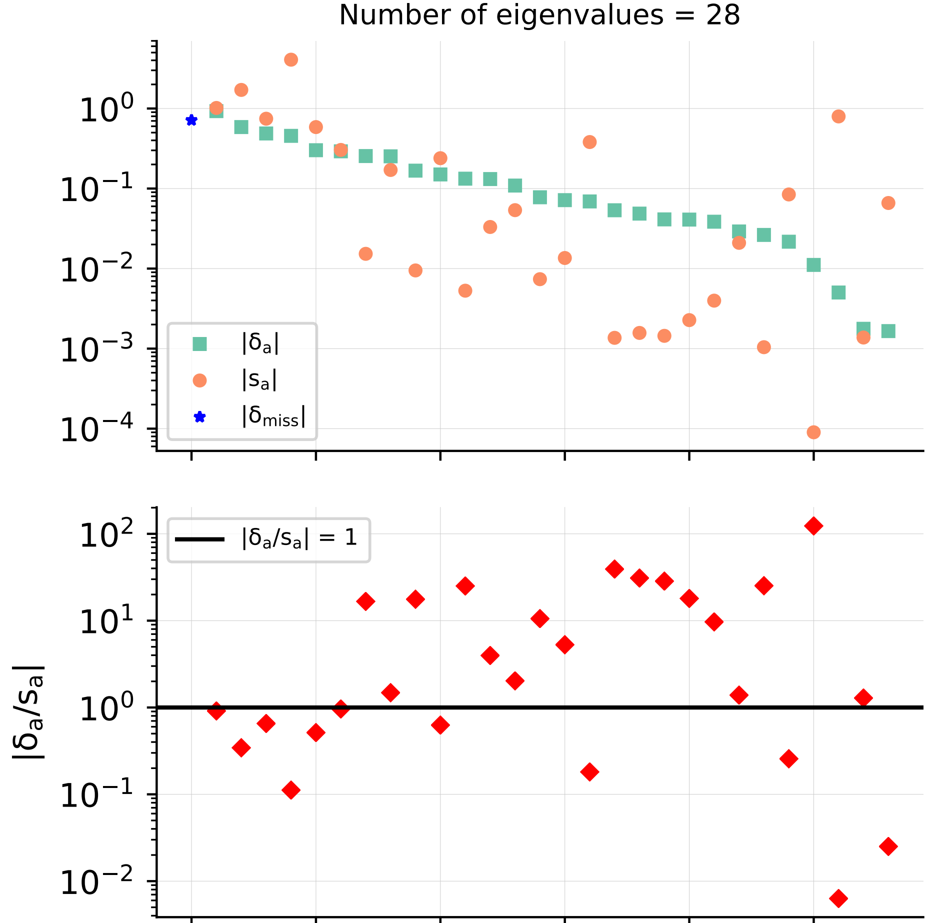
DIAGONAL UNCERTAINTIES VS. SHIFTS



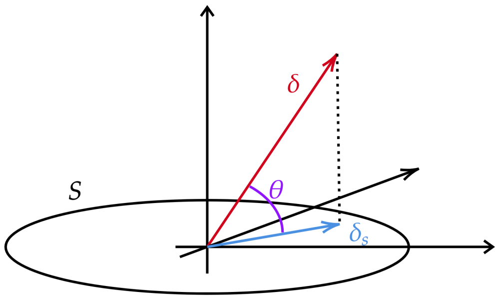
ANGLE BETWEEN SHIFT AND PROJECTED SHIFT

- DIS NC: 32°
- DIS CC: 16°
- DY: 21°
- JETS: 15°
- TOP: 3°
- GLOBAL: 27°

EIGENVALUES VS. SHIFT PROJECTIONS



SHIFT VECTOR VS PROJECTION



- ALL PRESCRIPTIONS BUT 3-PT PERFORM WELL
- ANGLE SCALES WITH NUMBER OF DATAPOINTS $\Rightarrow$ MORE POINTS, WORSE AGREEMENT
- ANGLE DOMINATED BY WORSE PROCESS

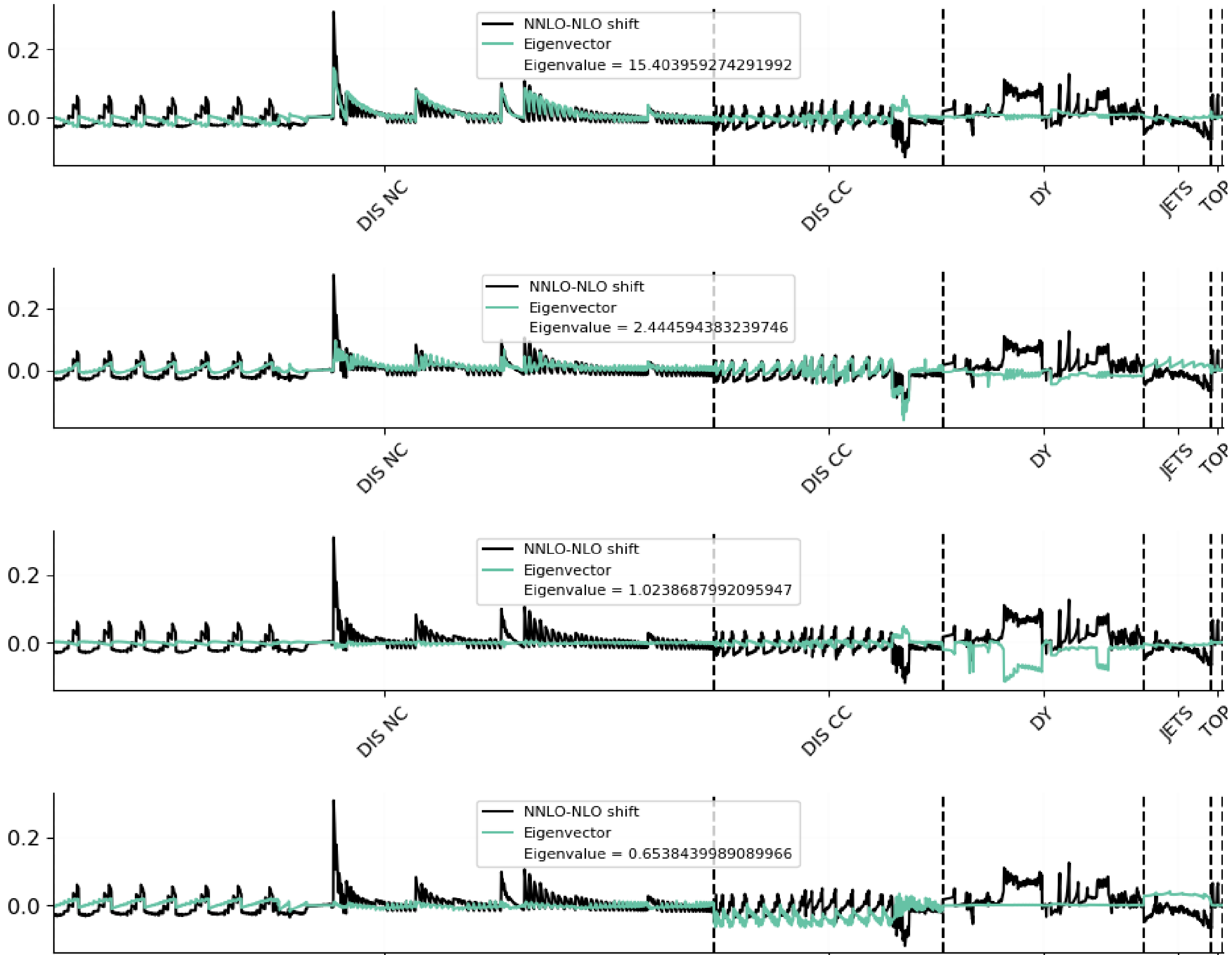
PRESCRIPTION	N_{sub}	θ
3-PT	6	52°
5-PT	8	33°
$\overline{5}$-PT	12	31°
7-PT	14	29°
9-PT	28	26°

PRESCRIPTION	θ				
	DIS NC	DIS CC	DY	JET	TOP
3-PT	54°	36°	39°	24°	12°
5-PT	39°	21°	25°	17°	11°
$\overline{5}$-PT	38°	17°	23°	22°	10°
7-PT	35°	17°	22°	16°	3°
9-PT	32°	16°	22°	14°	3°

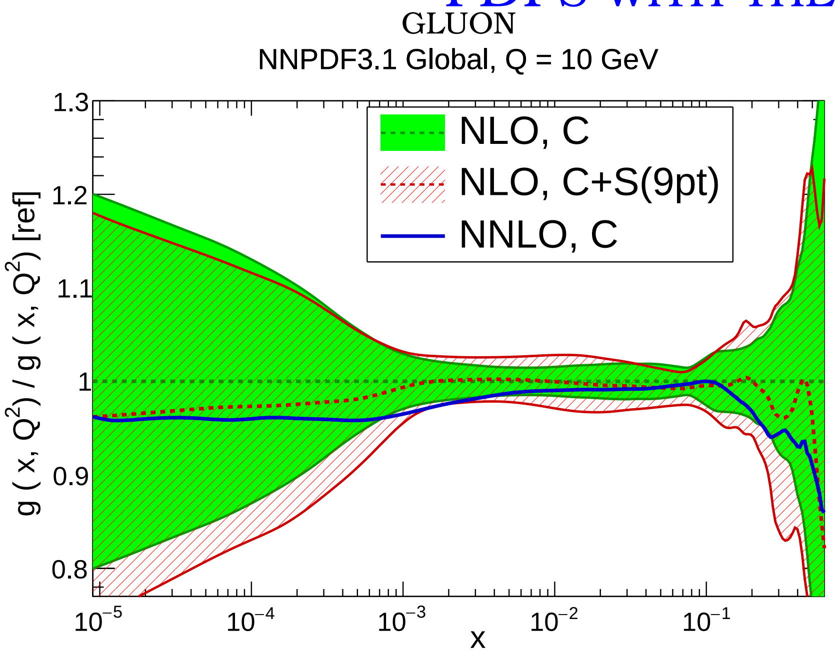
THE THEORY COVARIANCE MATRIX: VISUALIZING PROCESS IMPACT

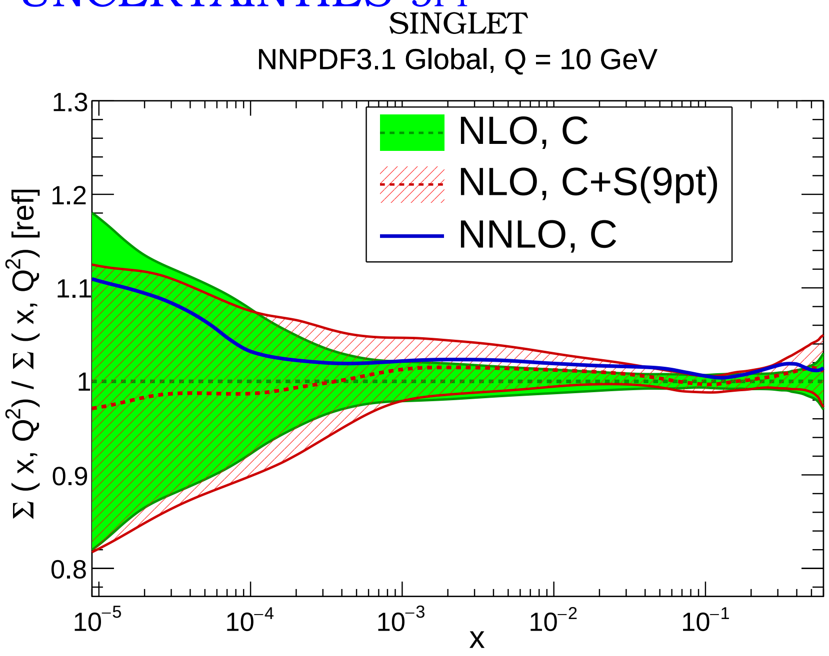
- PROJECT THE SHIFT VECTOR δ ON EACH EIGENVECTOR
- LOOK AT THE INDIVIDUAL ~ 3000 COMPONENTS
- GROUP POINTS BY PROCESS
- RELATION BETWEEN SCALE VARIATION EIGENVECTORS & PROCESSES

projection of the shift vector along the four dominant eigenvectors



PDFs WITH THEORY UNCERTAINTIES 9PT



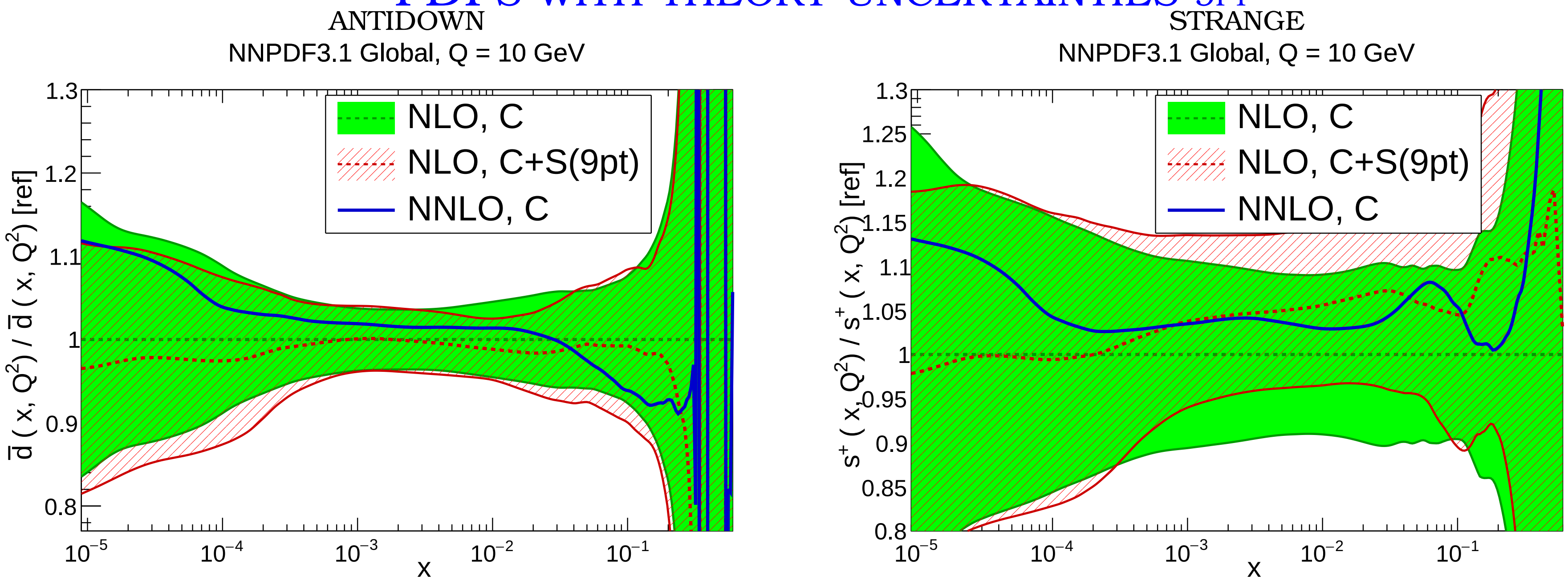


	C	$C+S^{(3\text{pt})}$	$C+S^{(9\text{pt})}$
χ^2	1.139	1.139	1.109
ϕ	0.314	0.310	0.315

- FIT QUALITY χ^2 IMPROVES
- RELATIVE ERROR ϕ ON PREDICTION DOES NOT CHANGE
- DATA REGION: PDF UNCERTAINTY ALMOST UNCHANGED
- EXTRAPOLATION REGION: PDF UNCERTAINTY SIGNIFICANTLY INCREASES
- CENTRAL VALUE MOVES TOWARDS KNOWN NNLO

EQUALLY PRECISE BUT MORE ACCURATE RESULT!

PDFs with theory uncertainties 9pt



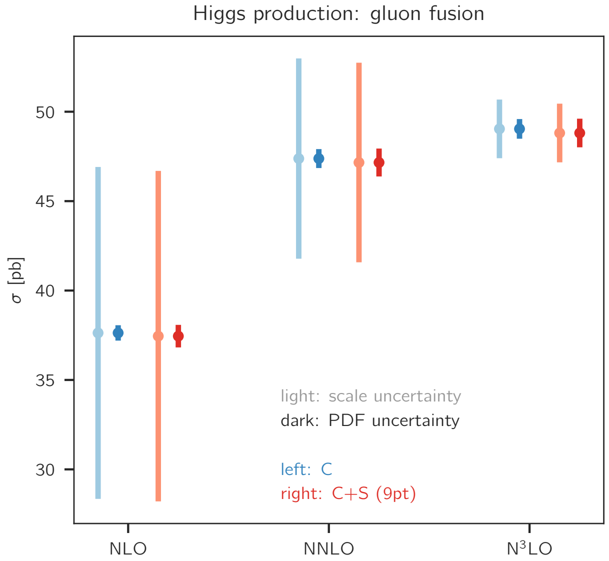
	C	$C+S^{(3\mathrm{pt})}$	$C+S^{(9\mathrm{pt})}$
χ^2	1.139	1.139	1.109
ϕ	0.314	0.310	0.315

- FIT QUALITY χ^2 IMPROVES
- RELATIVE ERROR ϕ ON PREDICTION DOES NOT CHANGE
- DATA REGION: PDF UNCERTAINTY ALMOST UNCHANGED
- EXTRAPOLATION REGION: PDF UNCERTAINTY SIGNIFICANTLY INCREASES
- CENTRAL VALUE MOVES TOWARDS KNOWN NNLO

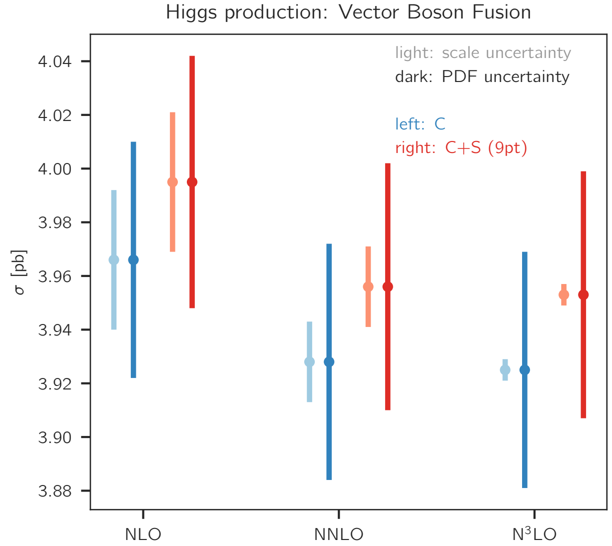
EQUALLY PRECISE BUT MORE ACCURATE RESULT!

PHENOMENOLOGY (LHC 13)

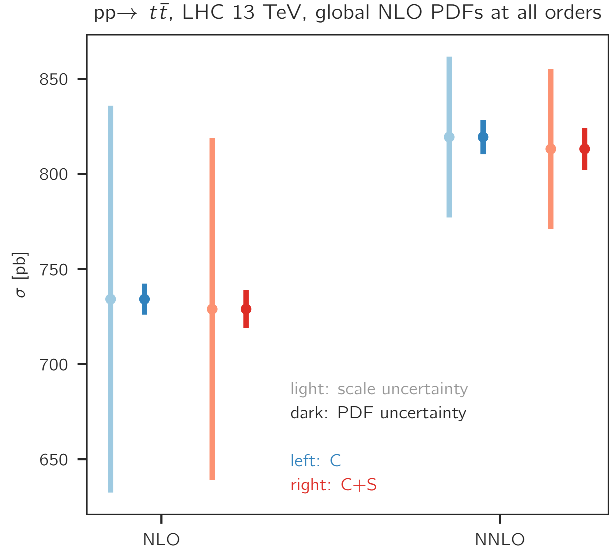
Higgs gg



Higgs VBF



Top pairs



USAGE: just compute PDF error as usual & combine with MHOU on hard matrix element computed with your preferred recipe

- GLUON FUSION, TOP
  - NO EFFECT ON CENTRAL VALUE
  - VISIBLE INCREASE OF UNCERTAINTIE
- VBF
  - MODERATE EFFECT ON UNCERTAINTIES
  - VISIBLE SHIFT OF CENTRAL VALUES

OUTLOOK

SUMMARY

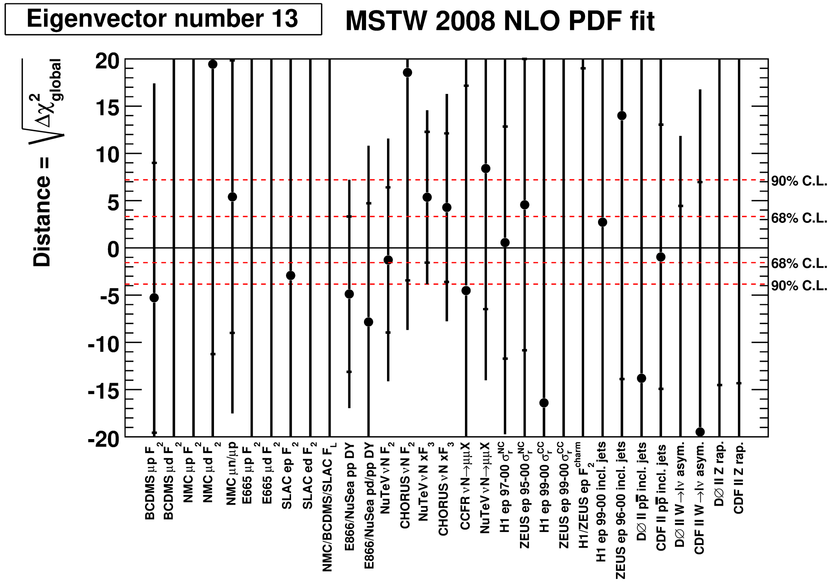
- BETTER DATA $\Leftrightarrow$ BETTER THEORY AND BETTER METHODOLOGY
- CAN WE REDUCE ARBITRARINESS IN PDF METHODOLOGY?
- CAN WE ASSESS ACCURATELY OUR THEORETICAL IGNORANCE?

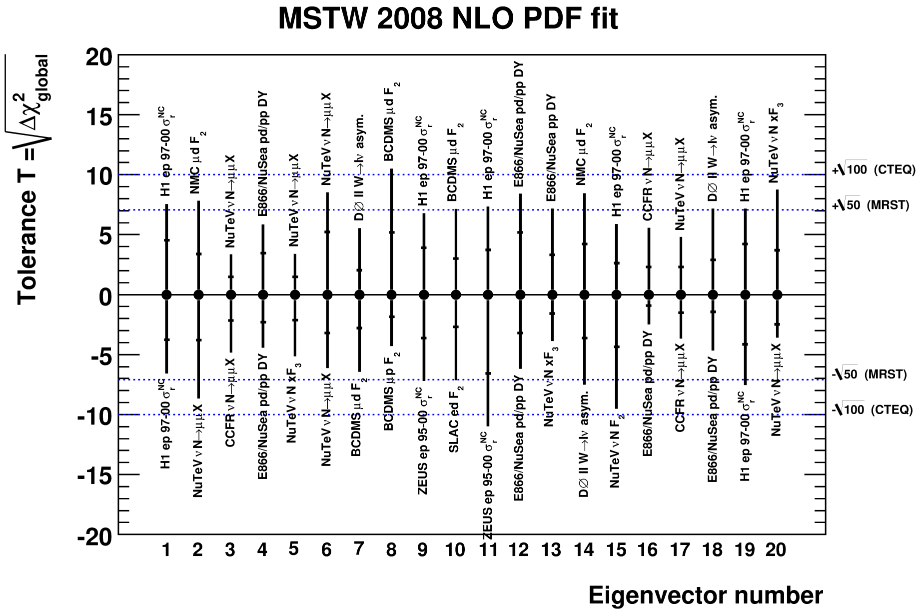
EXTRAS

PDF UNCERTAINTIES: TOLERANCE (MMHT-CT)

GLOBAL MSTW TOLERANCE

MSTW TOLERANCE PLOT FOR 13TH EIGENVEC.





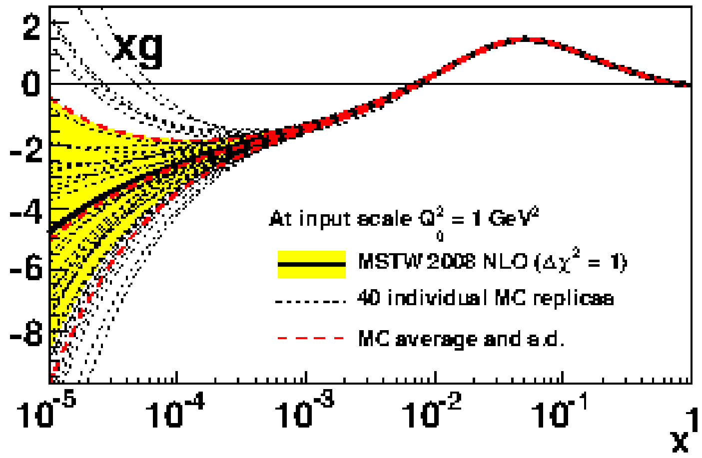
- (MSTW/MMHT) FOR EACH EIGENVECTOR IN PARAMETER SPACE DETERMINE CONFIDENCE LIMIT FOR THE DISTRIBUTION OF BEST-FITS OF EACH EXPERIMENT
- RESCALE $\Delta\chi^2 = T$ INTERVAL SUCH THAT CORRECT CONFIDENCE INTERVALS ARE REPRODUCED
- SIMILAR PROCEDURE ADOPTED BY CTEQ

WHAT ABOUT NNPDF?

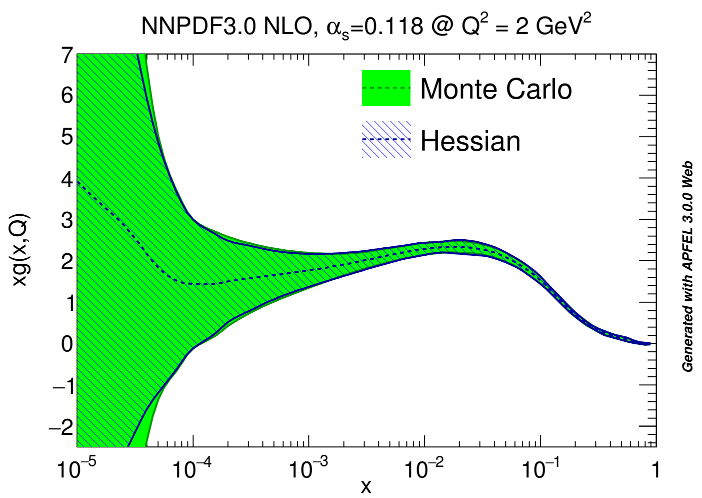
MC ⇔ HESSIAN

TWO DIFFERENT REPRESENTATIONS OF PDF UNCERTAINTIES

- 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)

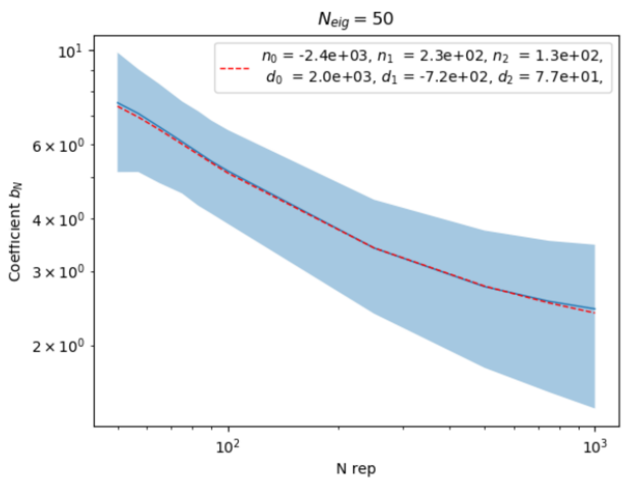
- TO CONVERT MONTE CARLO INTO HESSIAN, SAMPLE THE REPLICAS $f_i(x)$ AT A DISCRETE SET OF POINTS & CONSTRUCT THE ENSUING COVARIANCE MATRIX
- EIGENVECTORS OF THE COVARIANCE MATRIX AS A BASIS IN THE VECTOR SPACE SPANNED BY THE REPLICAS BY SINGULAR-VALUE DECOMPOSITION
- NUMBER OF DOMINANT EIGENVECTORS SIMILAR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

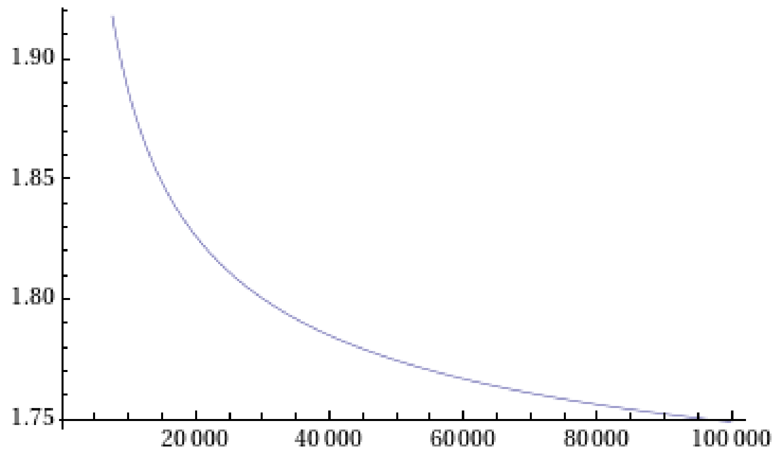
WHAT IS THE NNPDF "TOLERANCE"?

- PERFORM HESSIAN CONVERSION OF NNLO NNPDF3.1 PDFs
  50 OR 100 EIGENVECTORS
- DETERMINE χ^2 ALONG EACH EIGENVECTOR DIRECTION
- FIT A QUARTIC POLYNOMIAL
- STUDY DEPENDENCE ON
  NONGAUSSIANITY, NUMBER OF REPLICAS, NUMBER OF EIGENVECTORS,...

FINITE-SIZE EFFECTS

$\Delta\chi^2 = T^2$ VS NUMBER OF REPLICAS



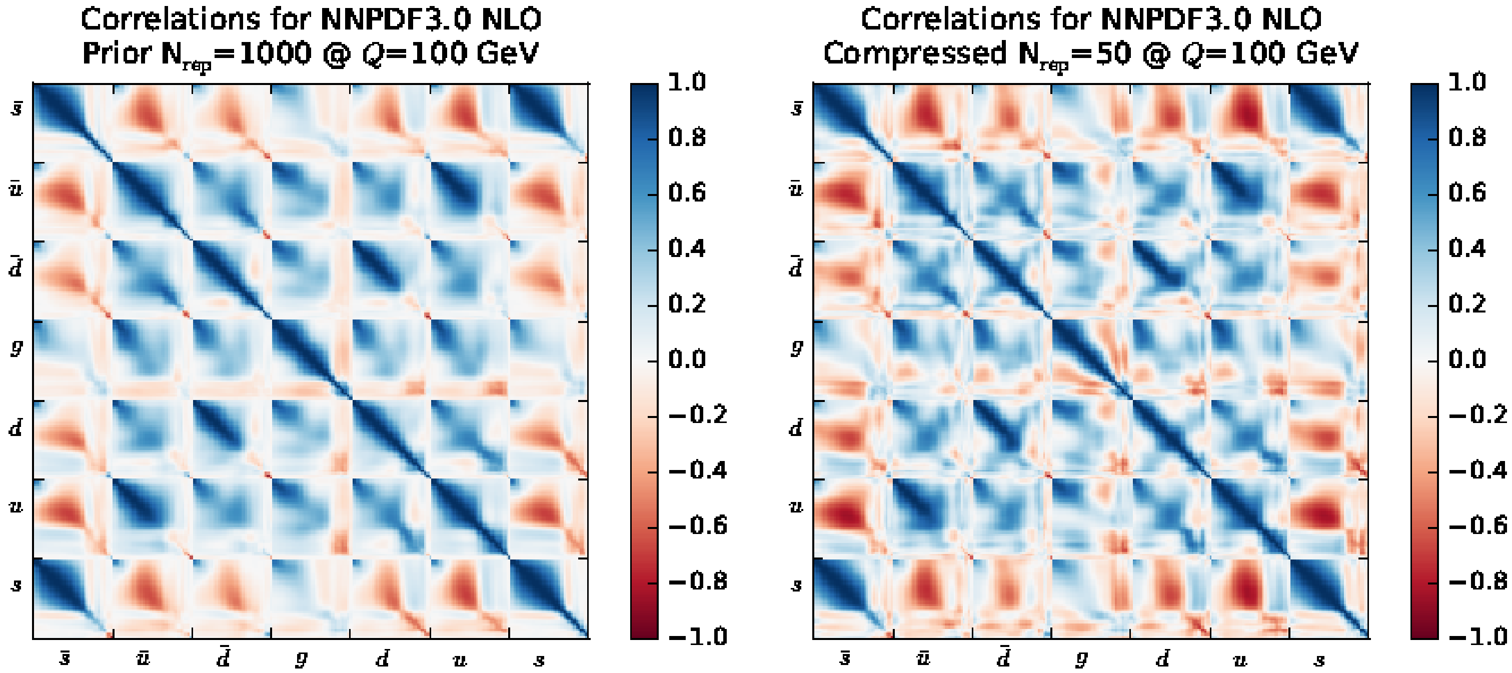


(Talon, MS thesis, 2019)

- NO SIGNIFICANT NONGAUSSIANITIY, DEVIATION FROM PARABOLIC,...
- SIGNIFICANT DEPENDENCE ON NUMBER OF REPLICAS
- ASYMPTOTIC TOLERANCE $T = 1.3 \pm 0.3$; $\Delta\chi^2 = 1.7 \pm 0.7$
- FOR $N_{\rm rep} = 100$, $T = 2.3$, EVEN FOR $N_{\rm rep} = 1000$, $T = 1.6$

DO WE HAVE TO FIT 10000 REPLICAS? DO WE HAVE TO USE 10000 REPLICAS?

SOLVING THE PROBLEM....
MONTECARLO COMPRESSION



(Carrazza, Latorre, Kassabov, Rojo, 2015)

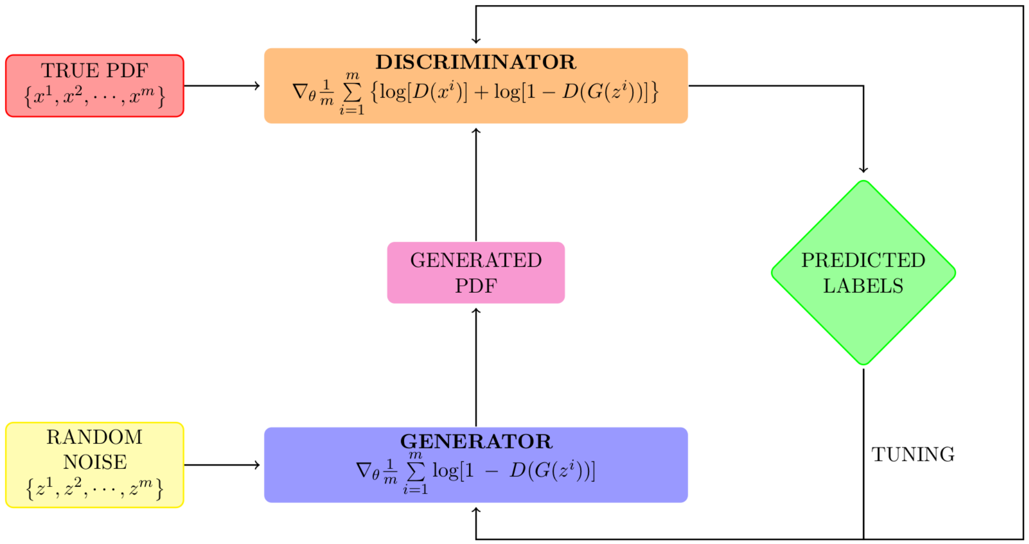
- START WITH LARGE REPLICA SAMPLE
- SELECT (BY GENETIC ALGORITHM) SUBSET OF REPLICAS $\Rightarrow$ STATISTICAL FEATURES OPTIMIZED TO PRIOR
- FOR ALL PDFs ON A GRID OF POINTS
  MINIMIZE DIFFERENCE OF FIRST FOUR MOMENTS, CORRELATIONS; OUTPUT OF KOLMOGOROV-SMIRNOV TEST (NUMBER OF REPLICAS BETWEEN MEAN AND σ, 2σ, INFINITY)
- 50 COMPRESSED REPLICA REPRODUCE 1000 REPLICA SET TO PRECENT ACCURACY

SOLVING THE PROBLEM....

GAN REPLICA GENERATION

- CAN WE REDUCE THE NUMBER OF COMPRESSED REPLICAS WITHOUT LOSS OF INFORMATION? SOLUTION FOR USER
- CAN WE INCREASE THE NUMBER OF REPLICAS WITHOUT REFITTING? SOLUTION FOR PDF FITTER

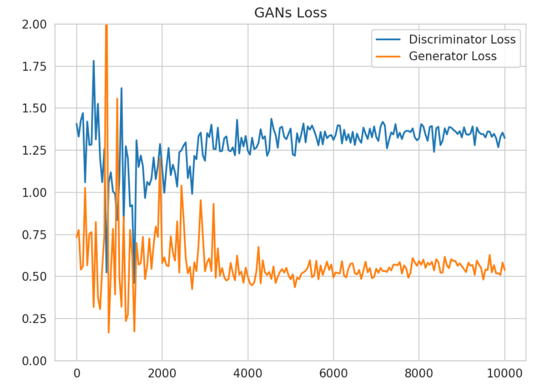
GENERATIVE ADVERSARIAL NETWORKS



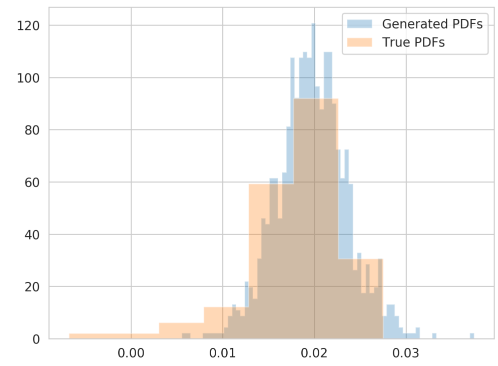
- TRAIN A NETWORK TO SIMULATE THE TRUE DISTRIBUTION (GENERATOR)
- TRAIN A NETWORK TO DISCRIMINATE TRUTH FROM SIMULATION (DISCRIMINATOR)
- TRAIN THE GENERATOR TO TRICK THE DISCRIMINATOR

SOLVING THE PROBLEM....
GAN REPLICA GENERATION

GAN TRAINING



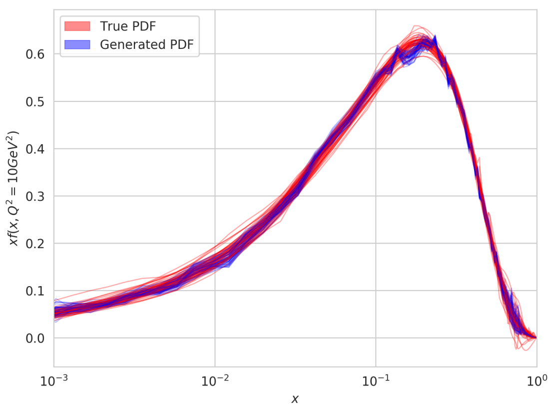
UP VALENCE AT FIXED x



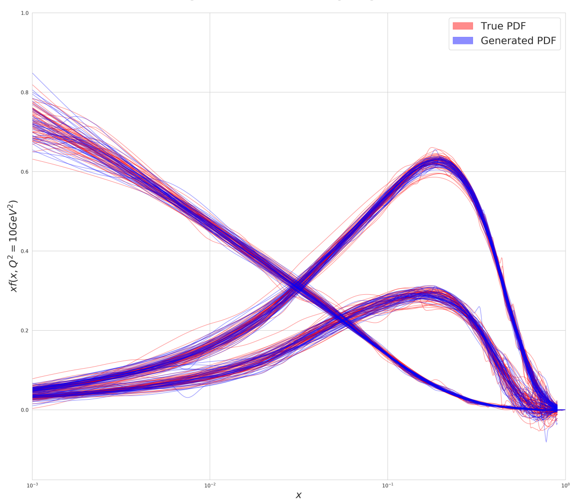
(Carrazza, Rabemananjara, preliminary)

- 1D GAN: REPRODUCE THE INFORMATION IN THE UNDERLYING REPLICA SET, BUT NO GAIN (WIGGLY REPLICAS)
  $\Rightarrow$ REDUCE THE NUMBER OF COMPRESSED REPLICA WITH FIXED NUMBER OF FITTED REPLICAS W/O INFORMATION LOSS
- 2D GAN: COMBINE CORRELATED INFORMATION FROM UNDERLYING REPLICA SET INFERRING THE TRUE UNDERLYING DISTTRIBUTION
  $\Rightarrow$ REDUCE THE NUMBER OF INPUT REPLICAS W/O INFORMATION LOSS

ONE-DIMENSIONAL



TWO-DIMENSIONAL

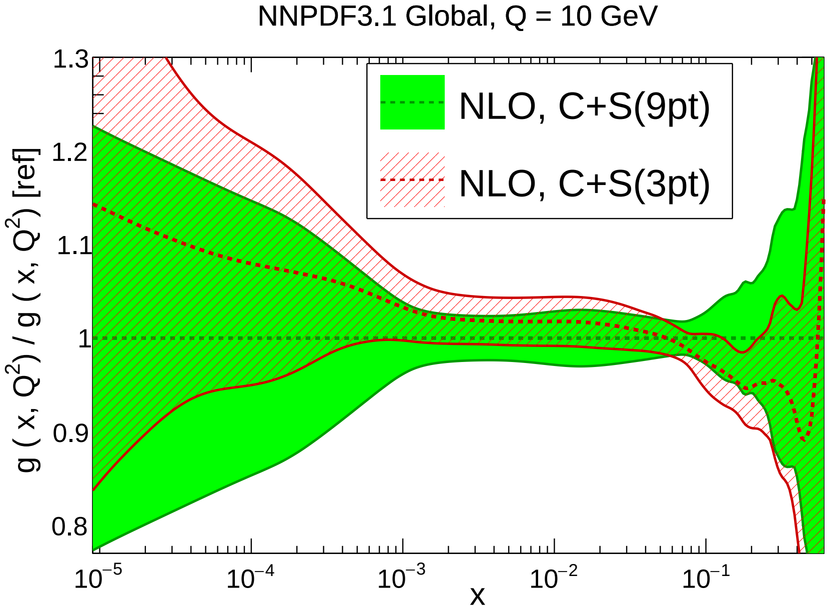


PDFs with theory uncertainties

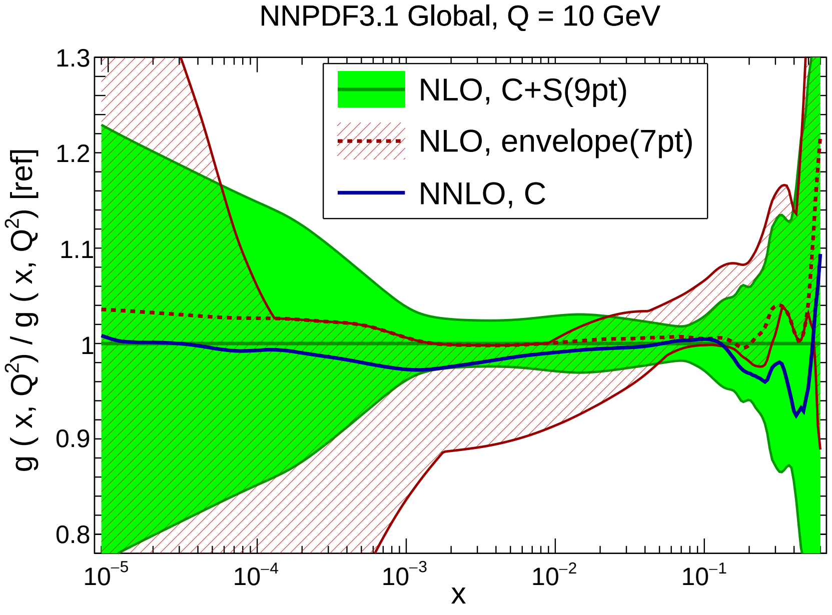
- results mildly dep. on prescription ⇒ 3pt closer to result w/o theory uncertainty
- "unstable" scale variations ⇒ no impact on fit
- 7pt envelope rather more conservative
  envelope does not include exp. uncertainty

the gluon PDF

3pt vs 9pt prescription

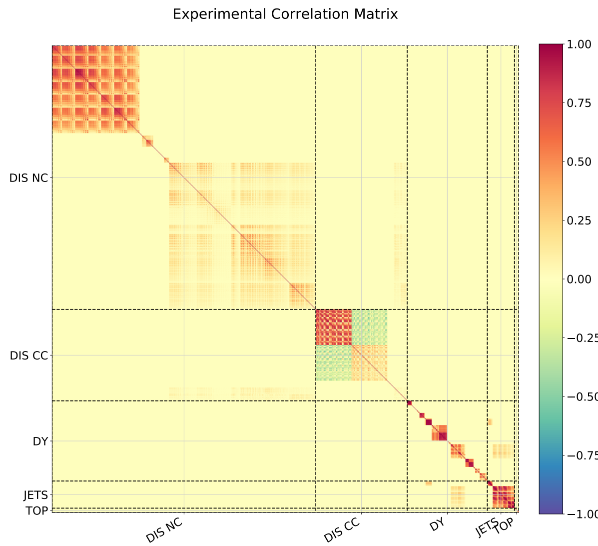


9pt prescription vs envelope

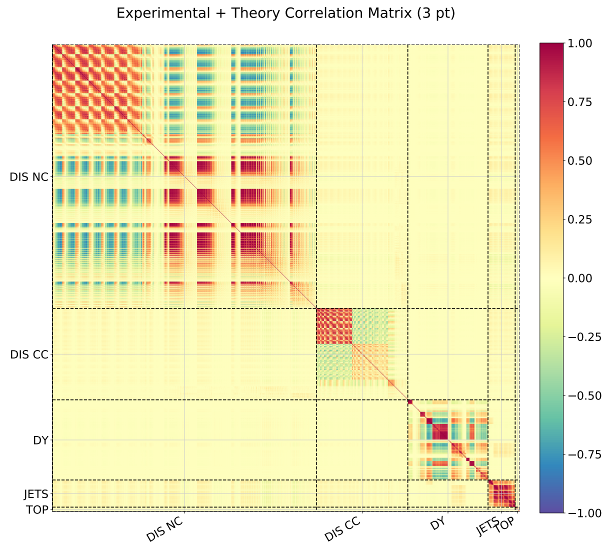


CORRELATION MATRICES

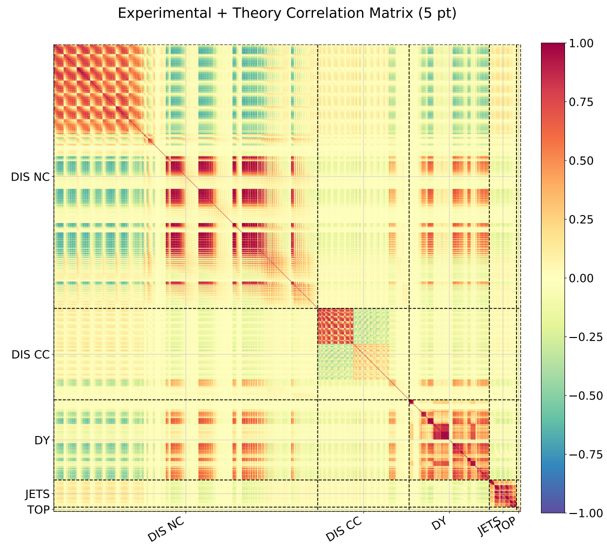
EXPERIMENT



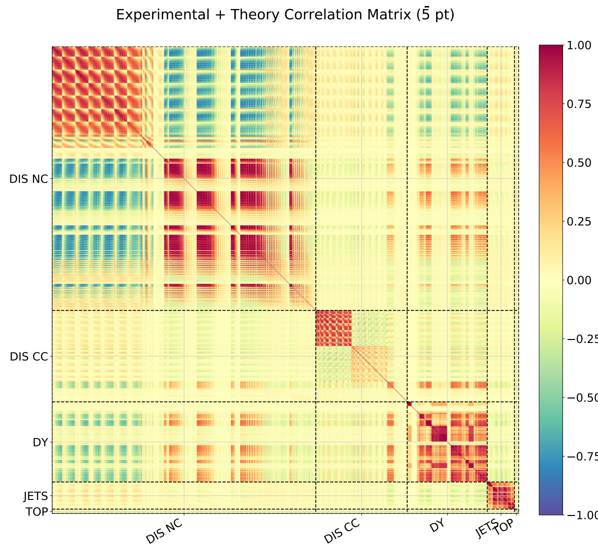
EXP+THEORY (3 PT)



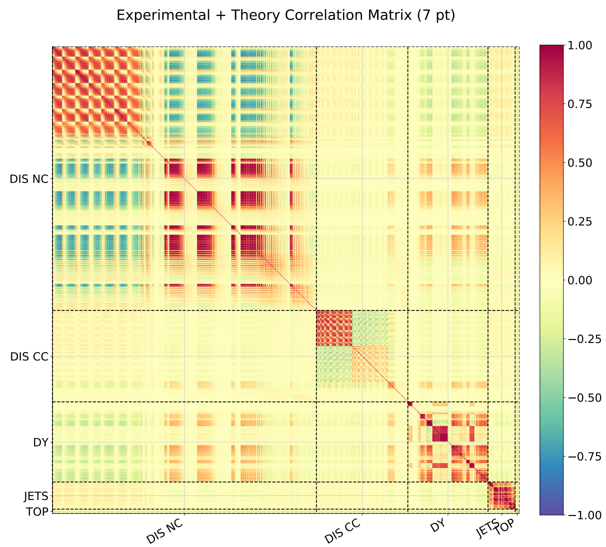
EXP+THEORY (5 PT)



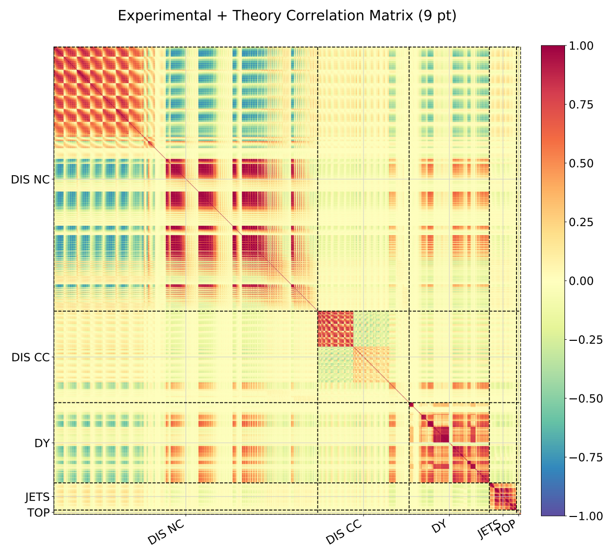
EXP+THEORY ($\bar{5}$ PT)



EXP+THEORY (7 PT)

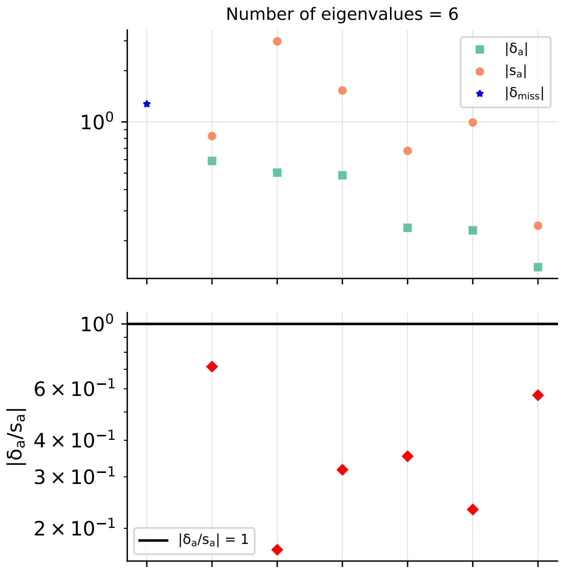


EXP+THEORY (9 PT)

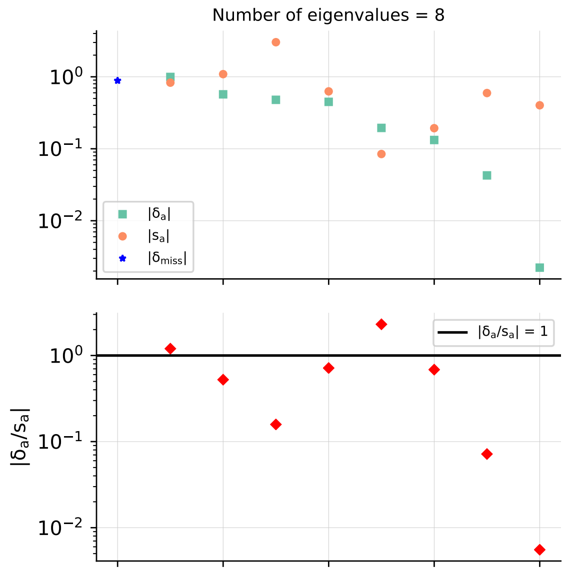


COVARIANCE EIGENVALUES vs. SHIFT PROJECTIONS

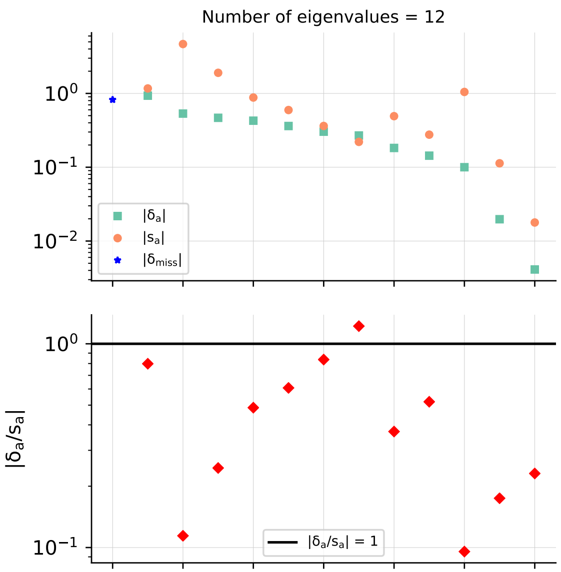
3 PT



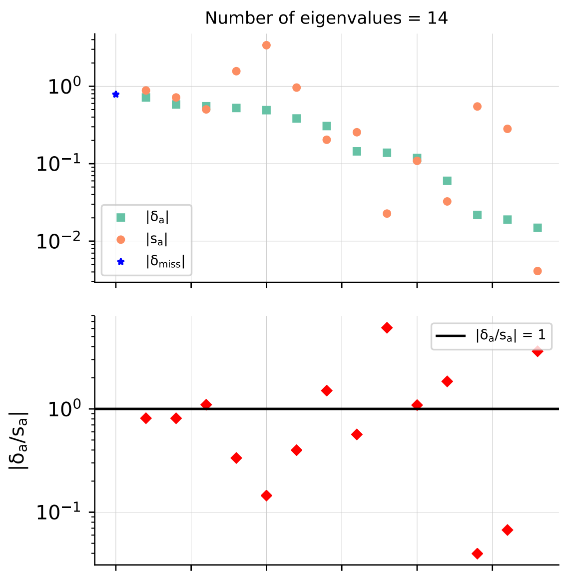
5 PT



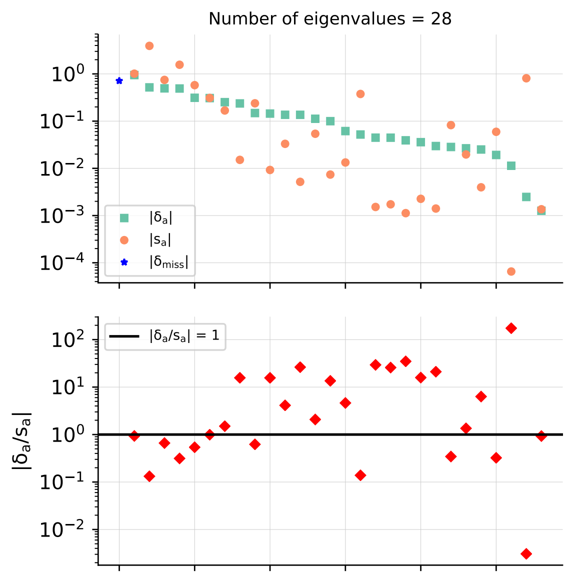
$\overline{5}$ PT



7 PT



9 PT



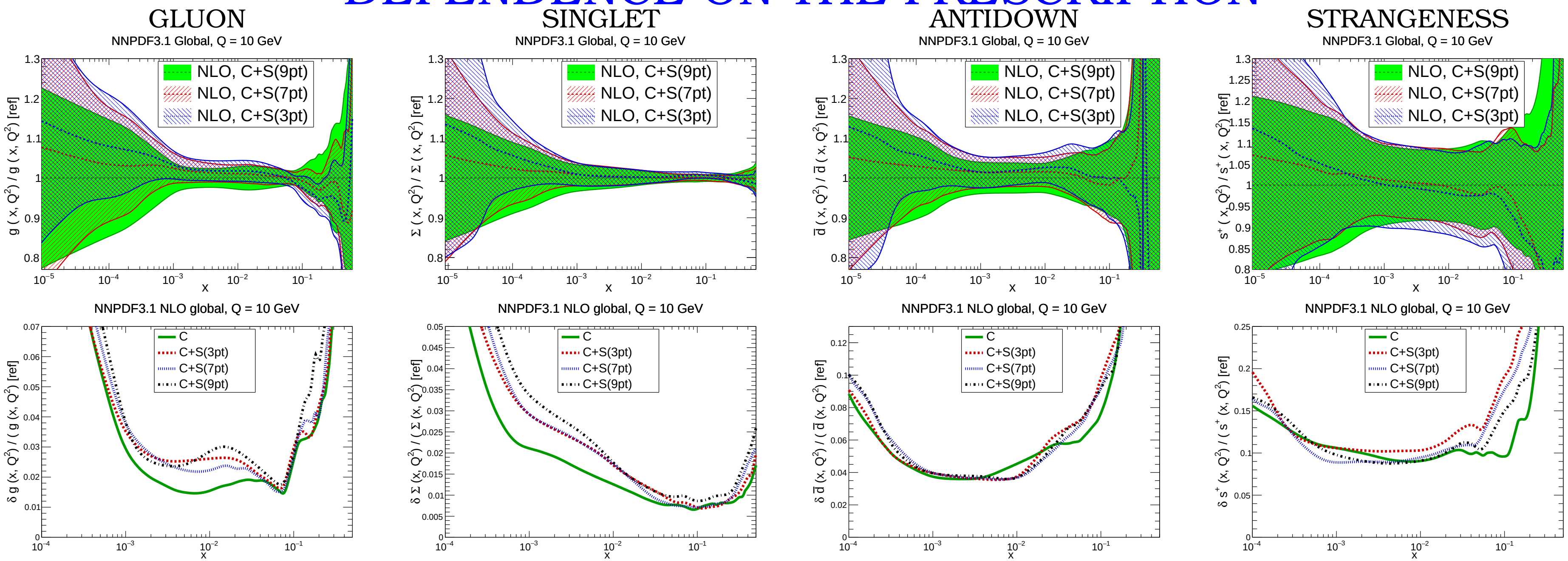
STATISTICAL INDICATORS

PROCESS	n_{dat}	χ^2/n_{dat} IN THE NNPDF3.1 GLOBAL FITS NLO C	$C+S^{(9\text{pt})}$	$C+S^{(7\text{pt})}$	$C+S^{(3\text{pt})}$	$C+S^{(9\text{pt})}_{\text{fit}}$	$C+S^{(9\text{pt})}_{\text{sampl}}$	NNLO C
DIS NC	1593	1.088	1.079	1.086	1.095	1.081	1.227	1.084
DIS CC	552	1.012	0.928	0.933	0.960	0.929	1.036	1.079
DY	484	1.486	1.447	1.485	1.483	1.461	1.434	1.231
JETS	164	0.907	0.839	0.858	0.901	0.848	0.911	0.950
TOP	26	1.260	1.012	1.016	1.077	1.001	1.264	1.068
TOTAL	2819	1.139	1.109	1.129	1.139	1.113	1.217	1.105

PROCESS	ϕ IN THE NNPDF3.1 GLOBAL FITS NLO C	$C+S^{(9\text{pt})}$	$C+S^{(7\text{pt})}$	$C+S^{(3\text{pt})}$	$C+S^{(9\text{pt})}_{\text{fit}}$	$C+S^{(9\text{pt})}_{\text{sampl}}$	NNLO C
DIS NC	0.266	0.268	0.262	0.261	0.261	1.137	0.305
DIS CC	0.389	0.376	0.367	0.391	0.369	0.502	0.471
DY	0.361	0.343	0.340	0.358	0.349	0.603	0.380
JETS	0.295	0.312	0.279	0.291	0.298	0.461	0.392
TOP	0.375	0.352	0.318	0.331	0.319	0.612	0.363
TOTAL	0.314	0.315	0.304	0.313	0.309	0.932	0.362

- MILD PRESCRIPTION DEPENDENCE
- COVMAT ONLY IN FITTING $\Rightarrow$ SAME CENTRAL VALUE, REDUCED UNCERTAINTY
  TH COVMAT RESOLVES TENSION

DEPENDENCE ON THE PRESCRIPTION



PDF THEORY ERROR AS A FIT UNCERTAINTY

- PDFs ARE DETERMINED BY MAXIMIZING THE LIKELIHOOD

$$P = N \exp - \left(\frac{d - t}{2\sigma^2_{exp}} \right)$$

d, t ARE REALLY VECTORS AND $1/\sigma^2$ THE INVERSE COVARIANCE MATRIX

- CAN VIEW THIS AS THE PROBABILITY OF THE THEORY t BEING CORRECT GIVEN DATA d, WHICH BY BAYES IS

$$P(t|d) \propto P(d|t)P(t)$$

- IF THEORY WAS KNOWN EXACTLY, THEN $P(t) = \delta(t - t^{\text{exact}})$
- IN ACTUAL FACT ONLY SOME PERTURBATIVE RESULT t_p IS EXACTLY KNOWN SO $t^{\text{exact}} = t_p + \Delta_p$, WHERE Δ_p INCLUDES MHO
- ASSUMING Δ TO BE GAUSSIANLY DISTRIBUTED, WITH UNCERTAINTY σ_{th} AND INTEGRATING OUT

$$P = N \exp \left[\frac{d - t_p}{2\left(\sigma^2_{exp} + \sigma^2_{th}\right)} \right]$$

- THEORETICAL UNCERTAINTY ADDED IN QUADRATURE, PROPAGATES INTO PDF UNCERTAINTY UPON MINIMIZATION
- SCALE VARIATION FOR EACH DATA POINT $\Rightarrow$ EIGENVECTOR OF COVARIANCE MATRIX (NUISANCE PARM.)