NNPDF

PAST AND FUTURE

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

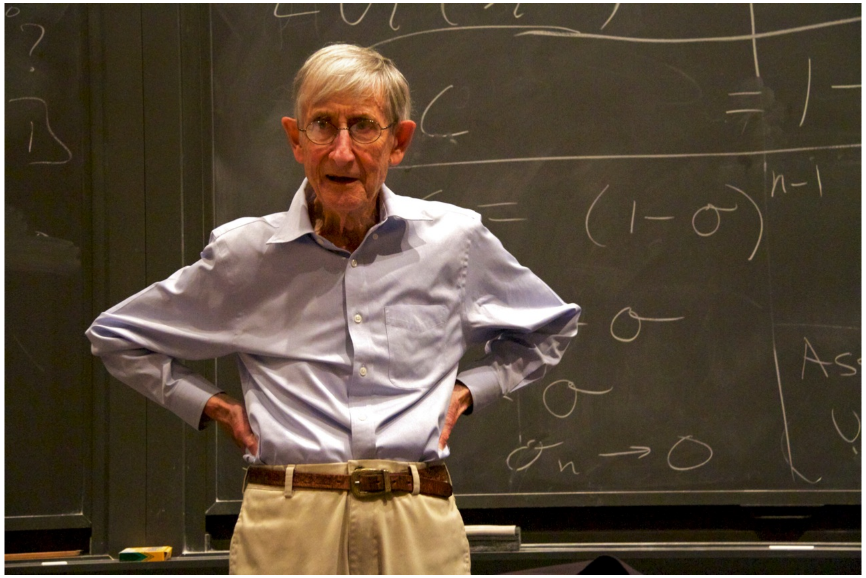
Istituto Nazionale di Fisica Nucleare

NNPDF+N^{3}PDF MEETING

GARGNANO, SEPTEMBER 17, 2018

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

ENERGY FRONTIER vs. ACCURACY FRONTIER



"There are historical reasons not to expect too much from the LHC. (...)

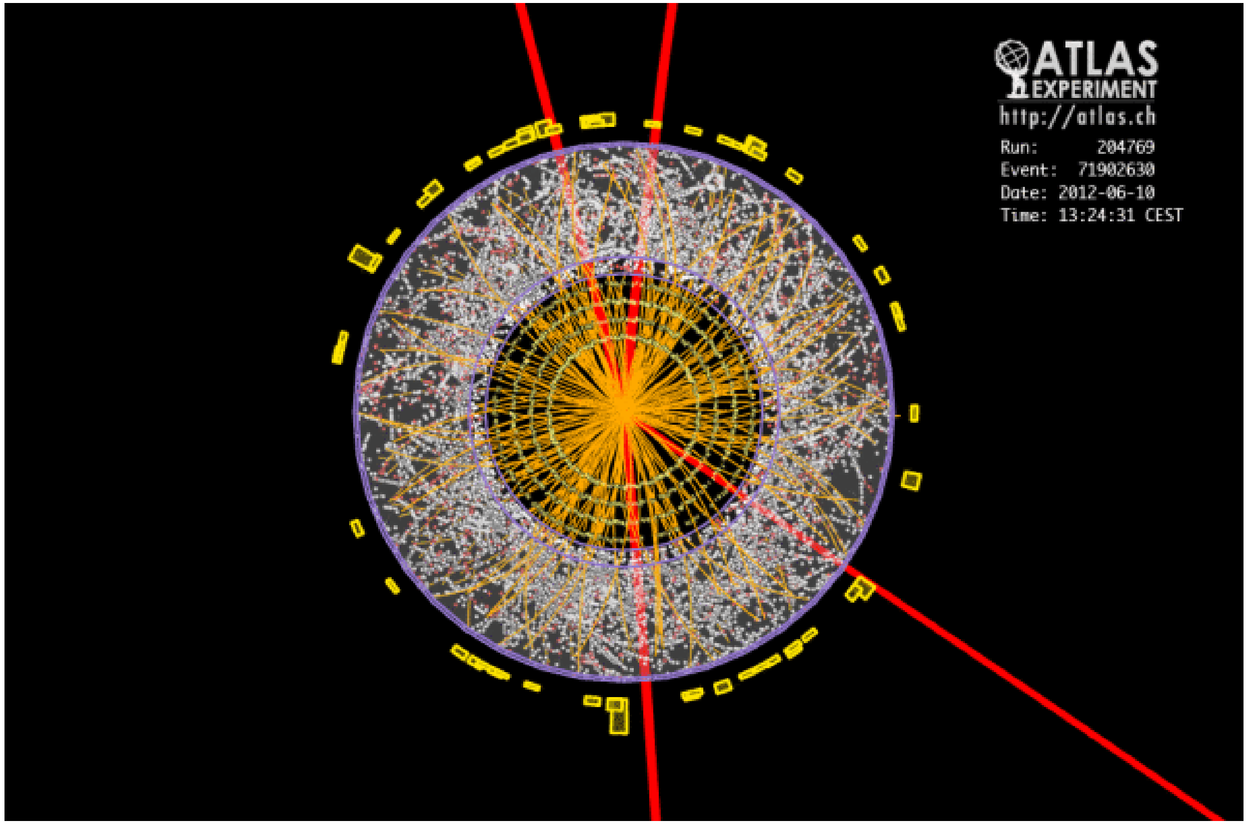
There have been sixteen important discoveries" (in HEP)
"between 1945 and 2008:
four discoveries on the energy frontier,
four on the rarity frontier,
eight on the accuracy frontier"

Freeman Dyson, 2008

PHYSICS AT THE LHC

"There are two reasons to be skeptical about the importance of the LHC: one technical and one historical".

HIGGS BOSON PRODUCED IN pp COLLISION, DECAY INTO TWO $\mu^+\mu^-$ PAIRS



"The technical weakness of the LHC arises from the nature of the collisions that it studies. These are collisions of protons with protons, and they have the unfortunate habit of being messy"

Freeman Dyson, 2008

PREHISTORY: DISCOVERY AT A HADRON COLLIDER

THE DISCOVERY OF THE W

THEORETICAL PREDICTION

42 *G. Altarelli et al. / Vector boson production*

TABLE 2

Values (in nb) of the total cross sections for $W^{\pm}$ and Z^0 production

$\sqrt{S}$ (GeV)	$W^+ + W^-$ GHR	$W^+ + W^-$ DO1	$W^+ + W^-$ DO2	Z^0 GHR	Z^0 DO1	Z^0 DO2	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ GHR	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ DO1	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ DO2
540	4.2	4.3	4.1	1.3	1.3	1.2	3.1	3.4	3.5
700	6.2	6.3	6.1	2.0	1.9	1.8	3.1	3.3	3.4
1000	9.5	9.5	9.6	3.1	3.0	2.9	3.1	3.2	3.3
1300	12.5	12.5	12.9	4.0	3.9	3.9	3.1	3.2	3.3
1600	15.5	15.6	16.5	5.0	4.8	5.0	3.1	3.2	3.3

ALTARELLI, ELLIS, GRECO, MARTINELLI, 1984

EXPERIMENTAL DISCOVERY

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-EP/85-108
11 July 1985

W PRODUCTION PROPERTIES AT THE CERN SPS COLLIDER

UA1 Collaboration, CERN, Geneva, Switzerland

Aachen[1] – Amsterdam (NIKHEF)[2] – Annecy (LAPP)[3] – Birmingham[4] – CERN[5] – Harvard[6] – Helsinki[7] – Kiel[8] – London (Imperial College[9] and Queen Mary College[10]) – Padua[11] – Paris (Coll. de France)[12] – Riverside[13] – Rome[14] – Rutherford Appleton Lab.[15] – Saclay (CEN)[16] – Victoria[17] – Vienna[18] – Wisconsin[19] Collaboration

The corresponding experimental result for the 1984 data at $\sqrt{s}$ = 630 GeV is

$$(\sigma \cdot B)_W = 0.63 \pm 0.05\ (\pm 0.09)\ \text{nb}.$$

This is in agreement with the theoretical expectation [14] of $0.47^{+0.14}_{-0.08}$ nb. We note that the 15%

- AGREEMENT AND UNCERTAINTIES AT 20% CONSIDERED TO BE SATISFACTORY
- RESULTS FROM DIFFERENT PDF SETS DIFFER BY AT LEAST 5%
- NO WAY TO ESTIMATE PDF UNCERTAINTIES

PREHISTORY: DISCOVERY AT A HADRON COLLIDER

THE DISCOVERY OF THE W

THEORETICAL PREDICTION

42 *G. Altarelli et al. / Vector boson production*

TABLE 2
Values (in nb) of the total cross sections for $W^{\pm}$ and Z^0 production

$\sqrt{S}$ (GeV)	$W^+ + W^-$ GHR	$W^+ + W^-$ DO1	$W^+ + W^-$ DO2	Z^0 GHR	Z^0 DO1	Z^0 DO2	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ GHR	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ DO1	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ DO2
540	4.2	4.3	4.1	1.3	1.3	1.2	3.1	3.4	3.5
700	6.2	6.3	6.1	2.0	1.9	1.8	3.1	3.3	3.4
1000	9.5	9.5	9.6	3.1	3.0	2.9	3.1	3.2	3.3
1300	12.5	12.5	12.9	4.0	3.9	3.9	3.1	3.2	3.3
1600	15.5	15.6	16.5	5.0	4.8	5.0	3.1	3.2	3.3

ALTARELLI, ELLIS, GRECO, MARTINELLI, 1984

PDFS IN 1984

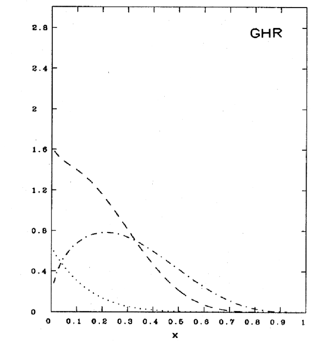


FIG. 25. Parton distributions of Glück, Hoffmann, and Reya (1982), at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and q_v (dotted line).

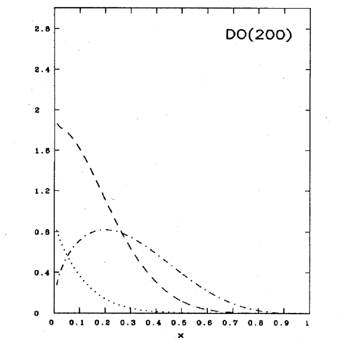


FIG. 27. "Soft-gluon" ($\Lambda=200$ MeV) parton distributions of Duke and Owens (1984) at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and $q_v(x)$ (dotted line).

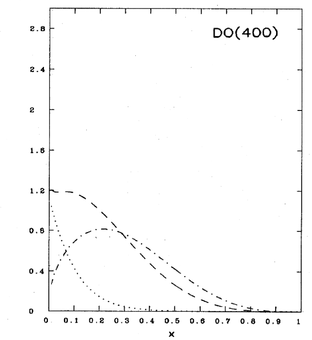


FIG. 26. "Hard-gluon" ($\Lambda=400$ MeV) parton distributions of Duke and Owens (1984) at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and $q_v(x)$ (dotted line).

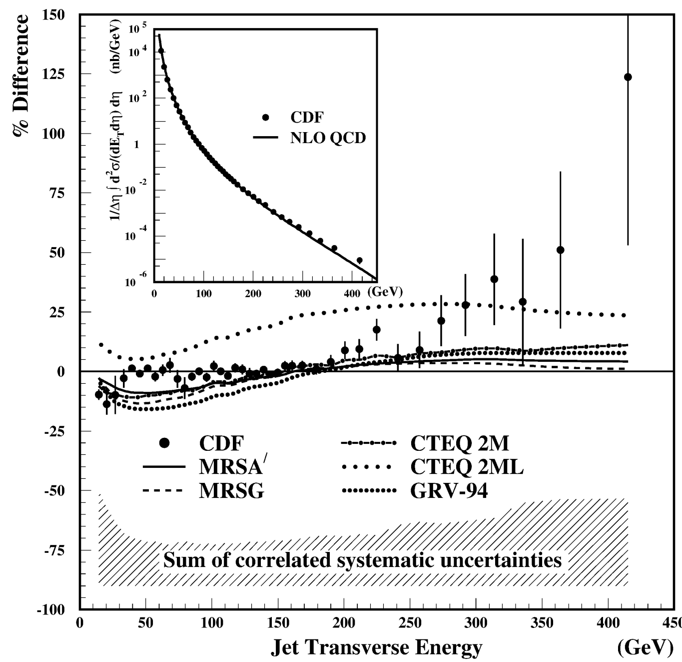
Rev. Mod. Phys., Vol. 56, No. 4, October 1984

GHR VS DUKE-OWENS

- AGREEMENT AND UNCERTAINTIES AT 20% CONSIDERED TO BE SATISFACTORY
- RESULTS FROM DIFFERENT PDF SETS DIFFER BY AT LEAST 5%
- NO WAY TO ESTIMATE PDF UNCERTAINTIES

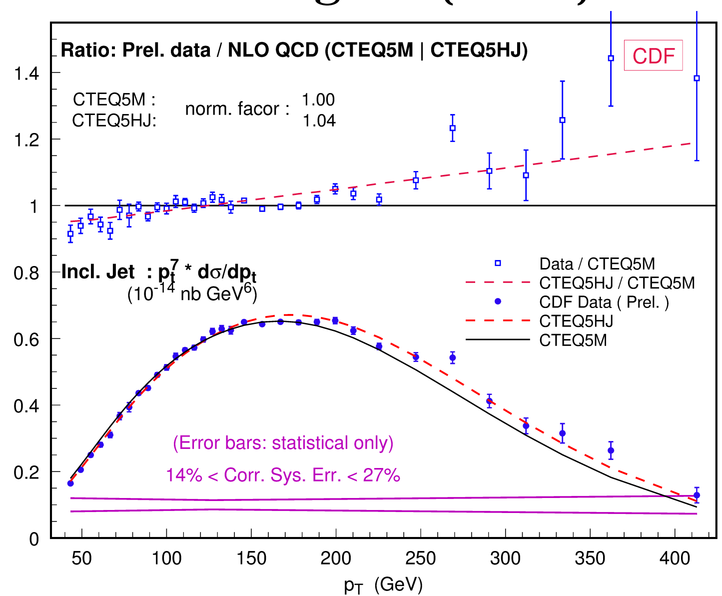
ANCIENT HISTORY: THE CDF LARGE E_T JETS

- DISCREPANCY BETWEEN QCD CALCULATION AND CDF JET DATA (1995)
- EVIDENCE FOR QUARK COMPOSITENESS?
- BUT NO INFO ON PARTON UNCERTAINTY $\Rightarrow$ RESULT STRONGLY DEPENDS ON GLUON AT $x \gtrsim 0.1$

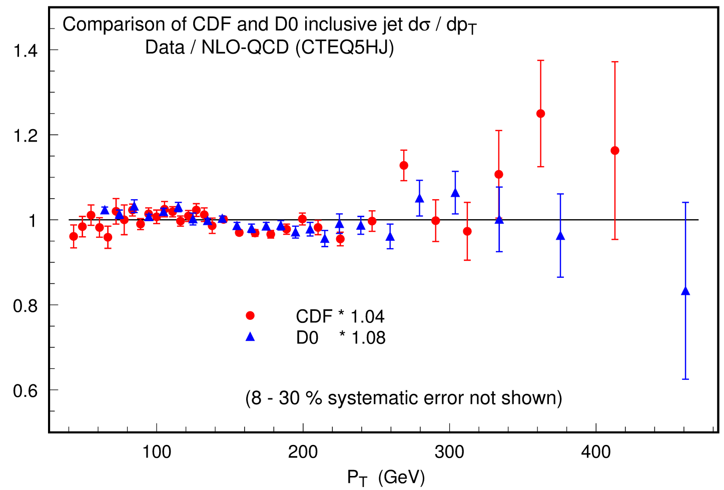


DISCREPANCY REMOVED IF JET DATA INCLUDED IN THE FIT

NEW CTEQ FIT (1996)

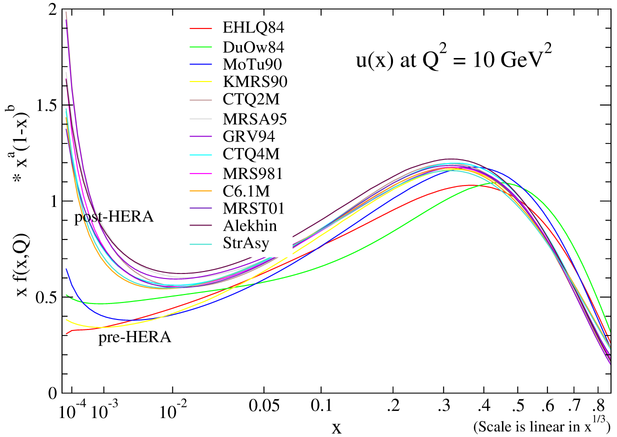


FINAL CTEQ FIT (1998)

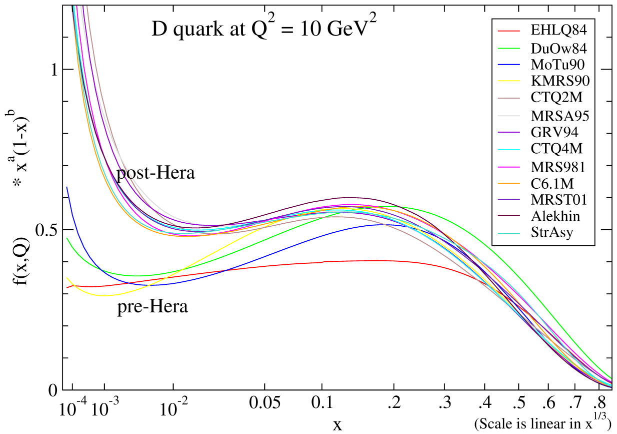


THE DARK AGES:
PDFs IN THE EARLY 2000

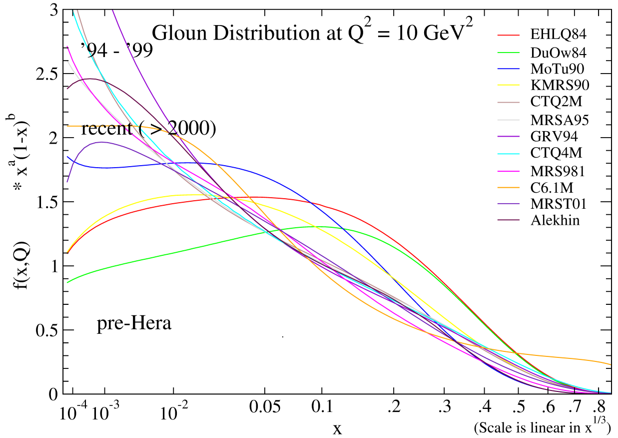
UP



DOWN



GLUON



W.K.Tung, DIS 2004

THE DARK AGES:
SEEKING A RENAISSANCE

D. Kosower, 1999

- FOR A SINGLE QUANTITY, WE QUOTE 1 SIGMA ERRORS: VALUE$\pm$ ERROR
- FOR A PAIR OF NUMBERS, WE QUOTE A 1 SIGMA ELLIPSE
- FOR A FUNCTION, WE NEED AN "ERROR BAR" IN A SPACE OF FUNCTIONS

MUST DETERMINE THE PROBABILITY DENSITY (MEASURE) $\mathcal{P}[f_i(x)]$
IN THE SPACE OF PARTON DISTRIBUTION FUNCTIONS $f_i(x)$ (i=quark, antiquark, gluon)

EXPECTATION VALUE OF $\sigma\,[f_i(x)] \Rightarrow$ FUNCTIONAL INTEGRAL

$$\Big\langle \sigma\,[f_i(x)]\,\Big\rangle = \int \mathcal{D}f_i\, \sigma\,[f_i(x)]\; \mathcal{P}[f_i],$$

MUST DETERMINE AN INFINITE–DIMENSIONAL OBJECT
FROM A FINITE SET OF DATA POINTS

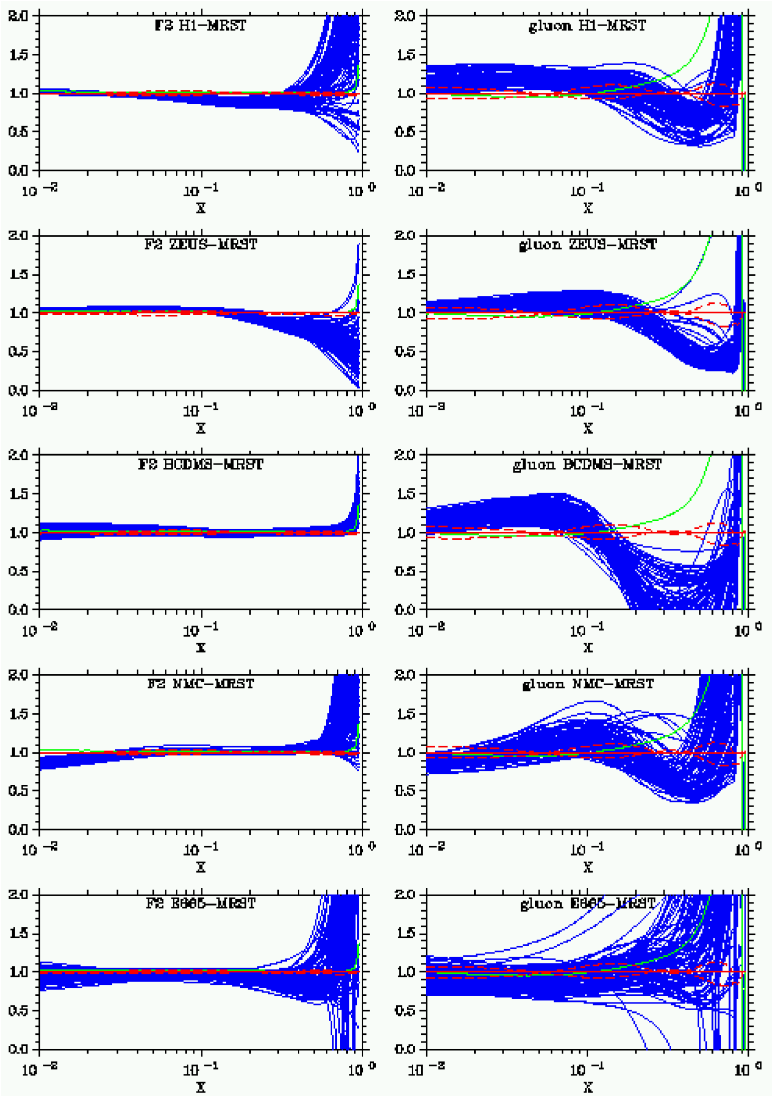
THE BAYESIAN MONTE CARLO APPROACH

(GIELE, KOSOWER, KELLER 2001)

- generate a Monte-Carlo sample of fcts. with "reasonable" prior distn. (e.g. an available parton set) → representation of probability functional $\mathcal{P}[f_i]$
- calculate observables with functional integral
- update probability using Bayesian inference on MC sample: better agreement with data → more functions in sample
- iterate until convergence achieved

PROBLEM IS MADE FINITE-DIMENSIONAL BY THE CHOICE OF PRIOR, BUT RESULT DO NOT DEPEND ON THE CHOICE IF SUFFICIENTLY GENERAL
HARD TO HANDLE "FLAT DIRECTIONS"
(Monte Carlo replicas which lead to same agreement with data);
COMPUTATIONALLY VERY INTENSIVE;
DIFFICULT TO ACHIEVE INDEP. FROM PRIOR

RESULT: FERMI PARTONS



$F_2^{\rm singlet}$ AND GLUON RATIOS FERMI/MRST

ONLY SUBSET OF DATA FITTED (H1, E665, BCDMS DIS DATA)

GOOD AGREEMENT WITH TEVATRON W XSECT

TROUBLE WITH VALUE OF α_s

MODERN TIMES:
THE HERA-LHC WORKSHOP

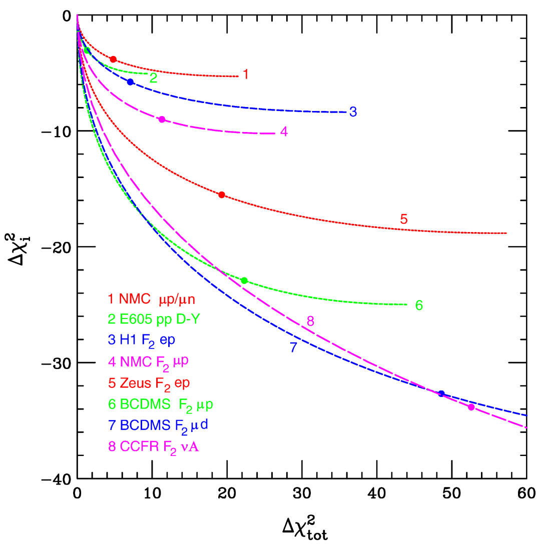
. . . this is when Dyson made his comments!

THE TOLERANCE PROBLEM

2002: FIRST PDFS WITH UNCERTAINTIES

- DETERMINE EIGENVECTORS OF χ^2 PARABOLOID
- DETERMINE 90% C.L. FOR EACH EXPT. ALONG EACH EIGENVECTOR
- DETERMINE MOST RESTRICTIVE INTERVAL ABOUT GLOBAL MINIMUM (TOLERANCE)
- $$\Delta\chi^2 = 100$$

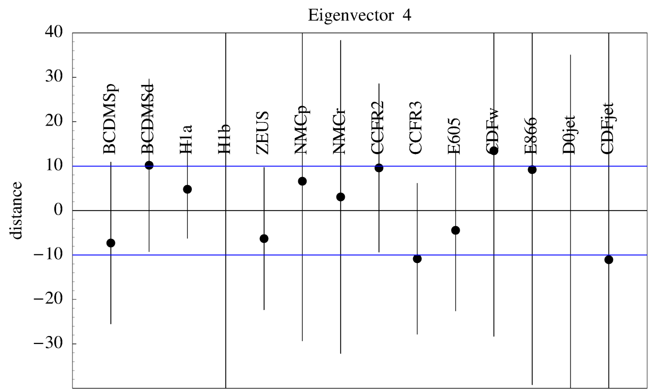
MINIMUM χ_i^2 VS GLOBAL χ^2



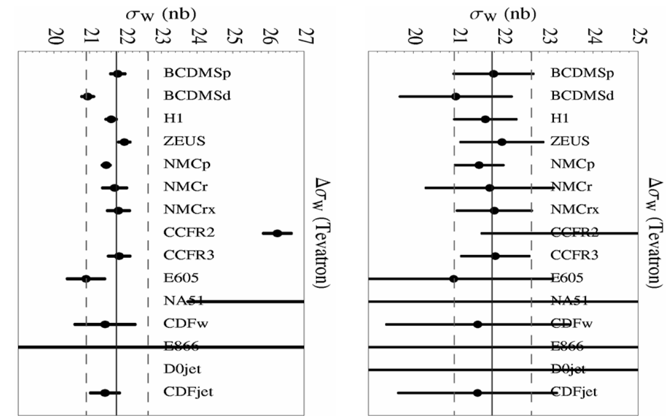
Collins, Pumplin 2001

CCFR, BCDMS INCOMPATIBLE

TOLERANCE PLOT
FOR 4TH EIGENVEC.



σ_W: ONE σVS. TOLERANCE

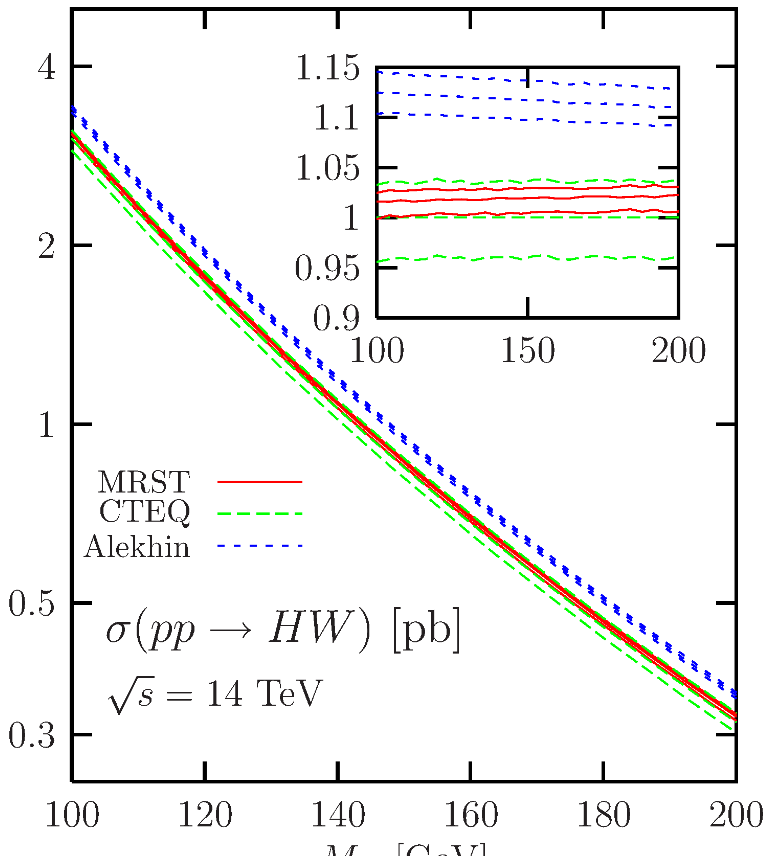


(CTEQ6, 2002-2007)

THE INCOMPATIBILITY PROBLEM

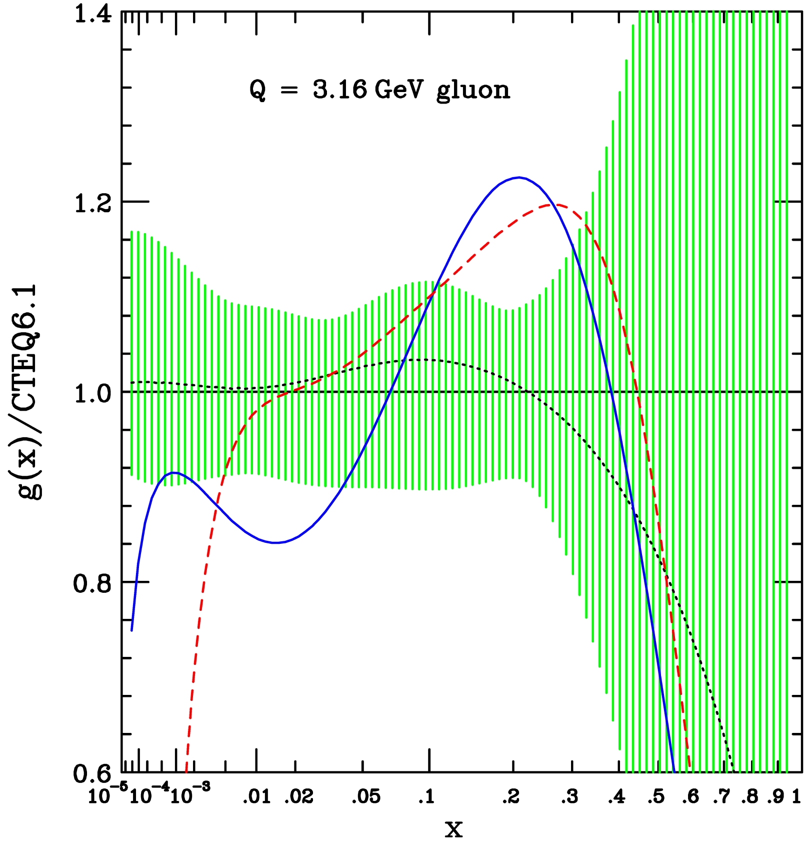
PARTON SETS DO NOT AGREE WITHIN RESPECTIVE ERRORS...

PHYSICAL OBSERVABLE: HIGGS PRODUCTION AT LHC



DJOUADI AND FERRAG, 2004

PARTON DISTRIBUTIONS: MRST/CTEQ GLUON



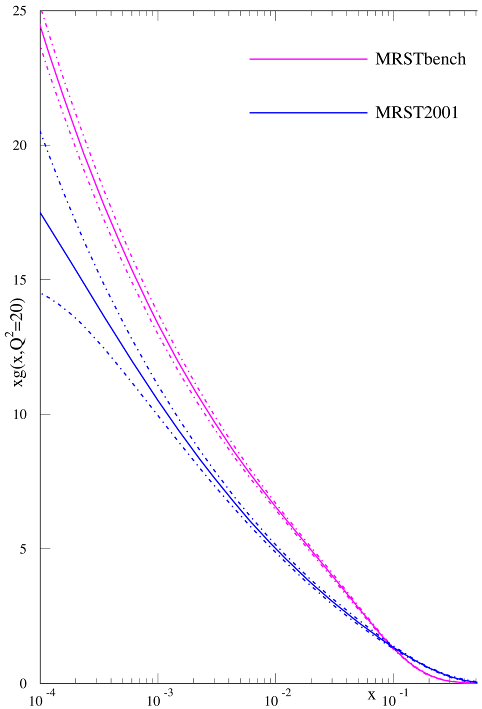
- ALEKHIN VS. MRST/CTEQ $\to$ PREDICTIONS FOR ASSOCIATE HIGGS W PRODUCTION @ LHC DO NOT AGREE WITHIN RESPECTIVE ERRORS
- MRST VS. CTEQ GLUONS DO NOT AGREE WITHIN RESPECTIVE ERRORS

ARE MORE DATA ENOUGH TO RESOLVE THE DISCREPANCIES?

THE HERA-LHC BENCHMARK PROBLEM

- RESTRICTED AND VERY CONSISTENT DATASET USED
- RESULTS COMPARED TO THEN-BEST RESULT FROM FULL DATASET

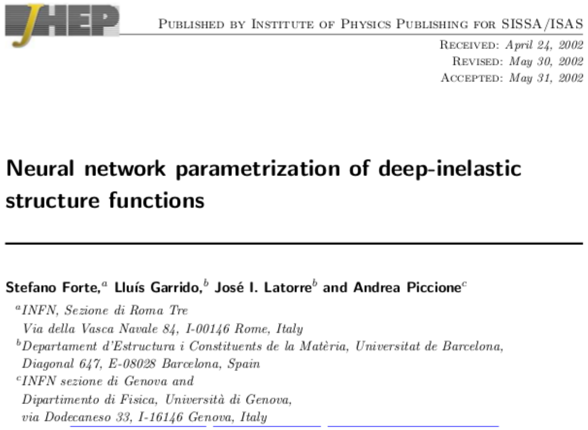
BENCHMARK VS DEFAULT GLUON



"...the partons extracted using a very limited data set are completely incompatible, even allowing for the uncertainties, with those obtained from a global fit with an identical treatment of errors...The comparison illustrates the problems in determining the true uncertainty on parton distributions." (R.Thorne, HERALHC, 2005)

ENLIGHTENMENT AND MODERN TIMES

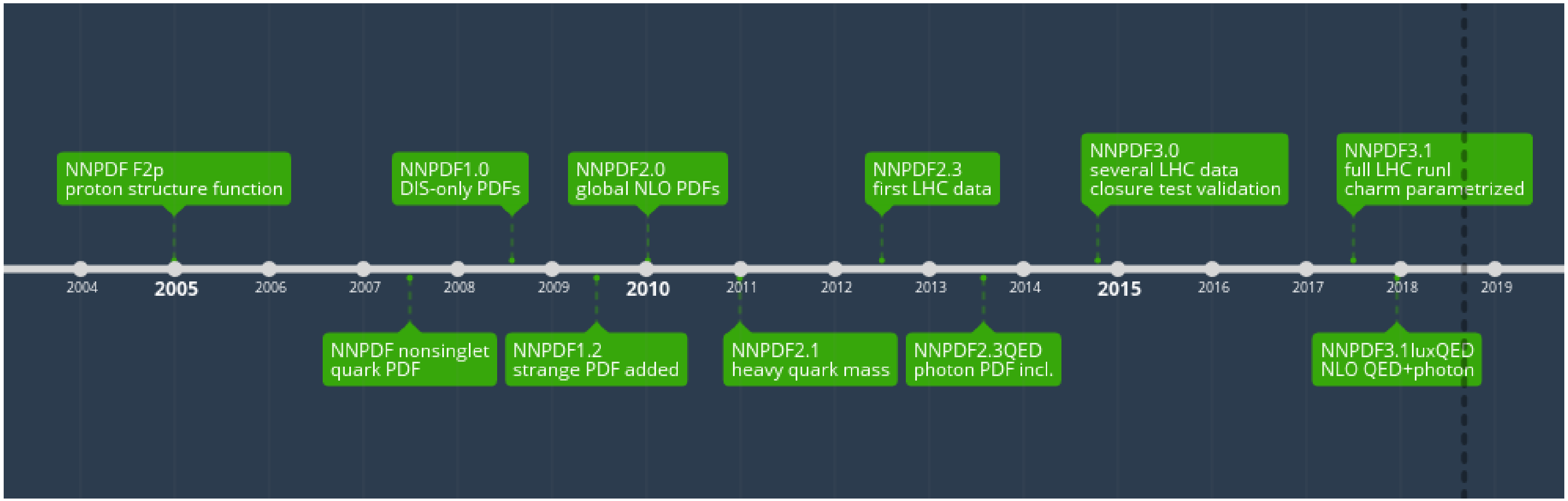
FROM THE PROOF OF CONCEPT…


Published by Institute of Physics Publishing for SISSA/ISAS
Received: April 24, 2002
Revised: May 30, 2002
Accepted: May 31, 2002

Neural network parametrization of deep-inelastic structure functions

Stefano Forte,[a] Lluís Garrido,[b] José I. Latorre[b] and Andrea Piccione[c]
[a] INFN, Sezione di Roma Tre
Via della Vasca Navale 84, I-00146 Rome, Italy
[b] Departament d'Estructura i Constituents de la Matèria, Universitat de Barcelona,
Diagonal 647, E-08028 Barcelona, Spain
[c] INFN sezione di Genova and
Dipartimento di Fisica, Università di Genova,
via Dodecaneso 33, I-16146 Genova, Italy

…TO THE NNPDF TIMELINE



THE NEURAL MONTE CARLO

THE NNPDF COLLABORATION

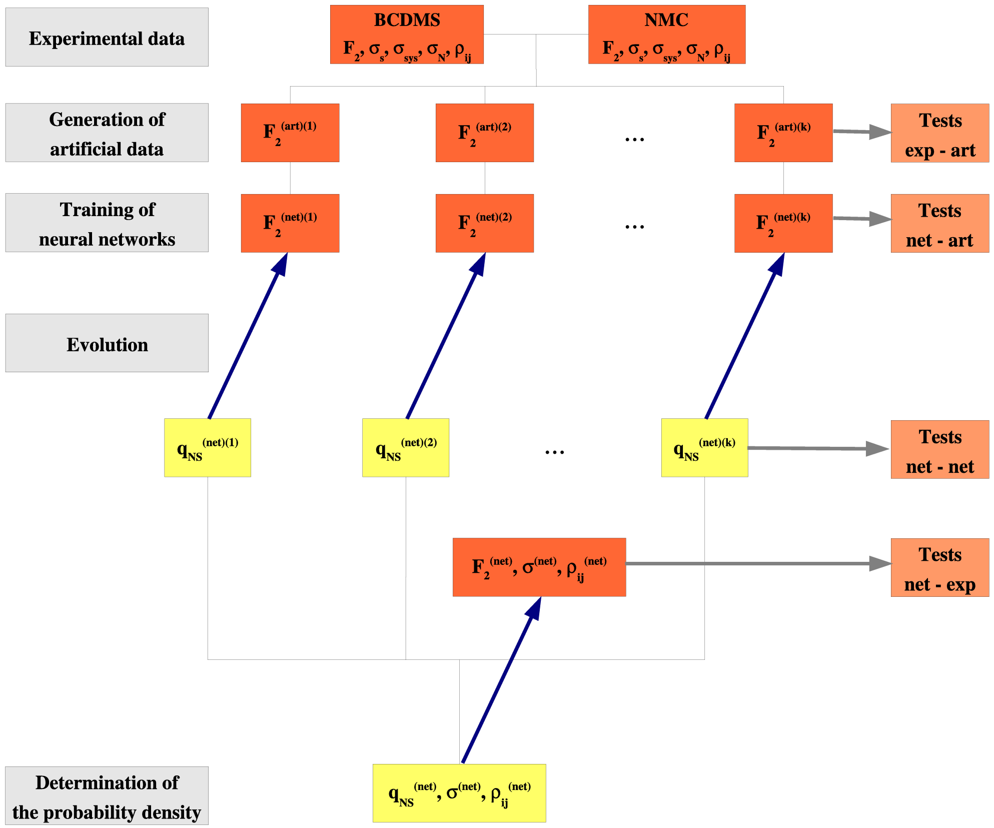
(2004: Del Debbio, SF, Latorre, Piccione, Rojo; 2007: +Ball, Guffanti, Ubiali)

BASIC IDEA: USE NEURAL NETWORKS AS UNIVERSAL UNBIASED INTERPOLANTS

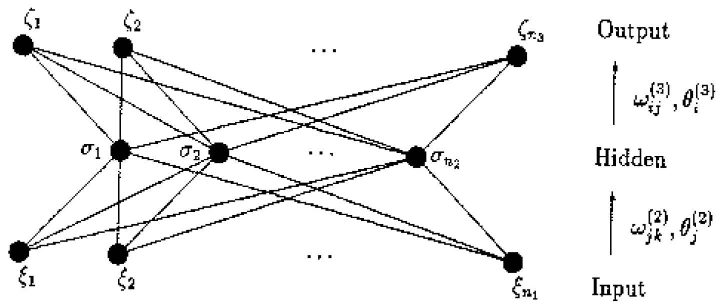
- GENERATE A SET OF MONTE CARLO REPLICAS $\sigma^{(k)}(p_i)$ OF THE ORIGINAL DATASET $\sigma^{(\mathrm{data})}(p_i)$
  $\Rightarrow$ REPRESENTATION OF $\mathcal{P}[\sigma(p_i)]$ AT DISCRETE SET OF POINTS p_i
- TRAIN A NEURAL NET FOR EACH PDF ON EACH REPLICA, THUS OBTAINING A NEURAL REPRESENTATION OF THE PDFS $f_i^{(net),(k)}$
- THE SET OF NEURAL NETS IS A REPRESENTATION OF THE PROBABILITY DENSITY:

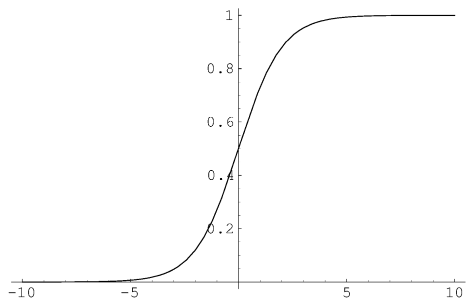
$$\left\langle \sigma\left[f_i\right] \right\rangle = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \sigma\left[f_i^{(net)(k)}\right]$$

(plot from the 2002 paper)



NEURAL NETWORKS
STRUCTURE





MULTILAYER FEED-FORWARD NETWORKS

- Each neuron receives input from neurons in preceding layer and feeds output to neurons in subsequent layer
- Activation determined by weights and thresholds
  $\xi_i = g\left(\sum_j \omega_{ij}\xi_j - \theta_i\right)$
- Sigmoid activation function
  $g(x) = \frac{1}{1+e^{-\beta x}}$

- WEIGHTS & THRESHOLDS CAN BE ADJUSTED SO THAT SIGMOIDS ARE IN CROSSOVER NONLINEAR REGION
- THANKS TO NONLINEAR BEHAVIOUR, ANY FUNCTION CAN BE EXPANDED OVER BASIS OF $g(x)$, $g(g(x))$, $g(g(g(x)))$ ...
- CAN CHOOSE REDUNDANT ARCHITECTURE (NO. OF LAYERS & NODES) TO MAKE SURE NO SMOOTHING BIAS IS INTRODUCED

ISSUES AND PROGRESS:

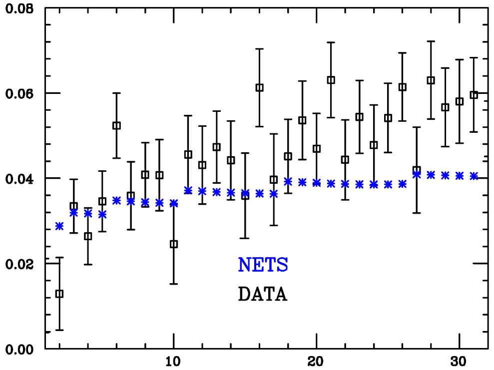
- CONSISTENCY AND ROBUSTNESS
  - NN ARCHITECTURE AND PREPROCESSING
  - OVER- VS. UNDERLEARNING AND STOPPING
  - DATA WEIGHTING AND CONSISTENCY
- RELIABILITY
  - DEPENDENCE OF UNCERTAINTY ON DATASET
  - FUNCTIONAL BIAS OR LACK THEREOF
  - INCONSISTENT DATA HANDLING
- THEORY ISSUES
  - α_s
  - HEAVY QUARKS

OVER/UNDERLEARNING 2002

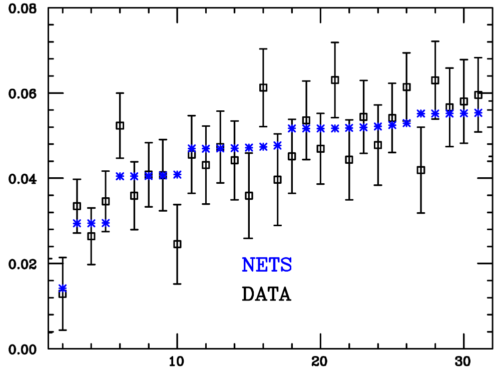
TRAINING BY BACK-PROPAGATION

- START WITH RANDOM NETWORK & COMPUTE OUTPUT FOR GIVEN INPUT (F_2 FOR GIVEN (x, Q^2))
- COMPARE COMPUTED OUTPUT TO DESIRED OUTPUT BY MEANS OF ENERGY FUNCTION ($e.g.$ χ^2)
- VARY WEIGHTS AND THRESHOLDS ALONG DIRECTION OF STEEPEST DESCENT OF ENERGY FUNCTION $\Rightarrow$ CAN BE DONE BY BACK-PROPAGATION
- ITERATE

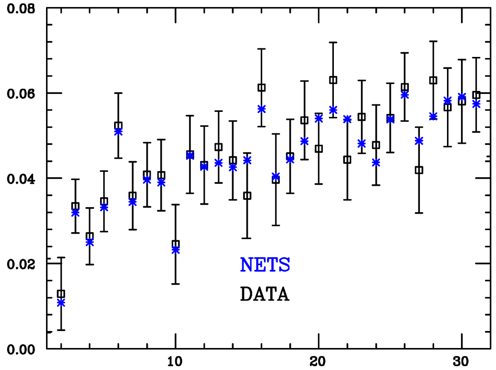
UNDERLEARNING



"PROPER" LEARNING



OVERLEARNING



WHEN SHOULD TRAINING STOP?
WHICH IS THE APPROPRIATE ENERGY FUNCTION?

OPTIMAL TRAINING

WITH LONG ENOUGH TRAINING & BIG ENOUGH NETWORK,
PREDICTION GOES THROUGH ALL POINTS

any error function proportional to (data-nets) will do: vanishes at minimum.

Q: DO WE REALLY WANT THIS?
NAIVE A: SURE! Then when averaging over MC sample, at (x, Q^2) of datapoints averaging over nets is *identical* to averaging over data
OBJECTION: WHAT IF WE HAVE TWO MEASUREMENTS AT THE SAME (x, Q^2)?

PERFORM WEIGHTED AVERAGE $\frac{F_2^{(1)}/\sigma_1+F_2^{(2)}/\sigma_2}{1/\sigma_1+1/\sigma_2}$ BEFORE DATA GENERATION.

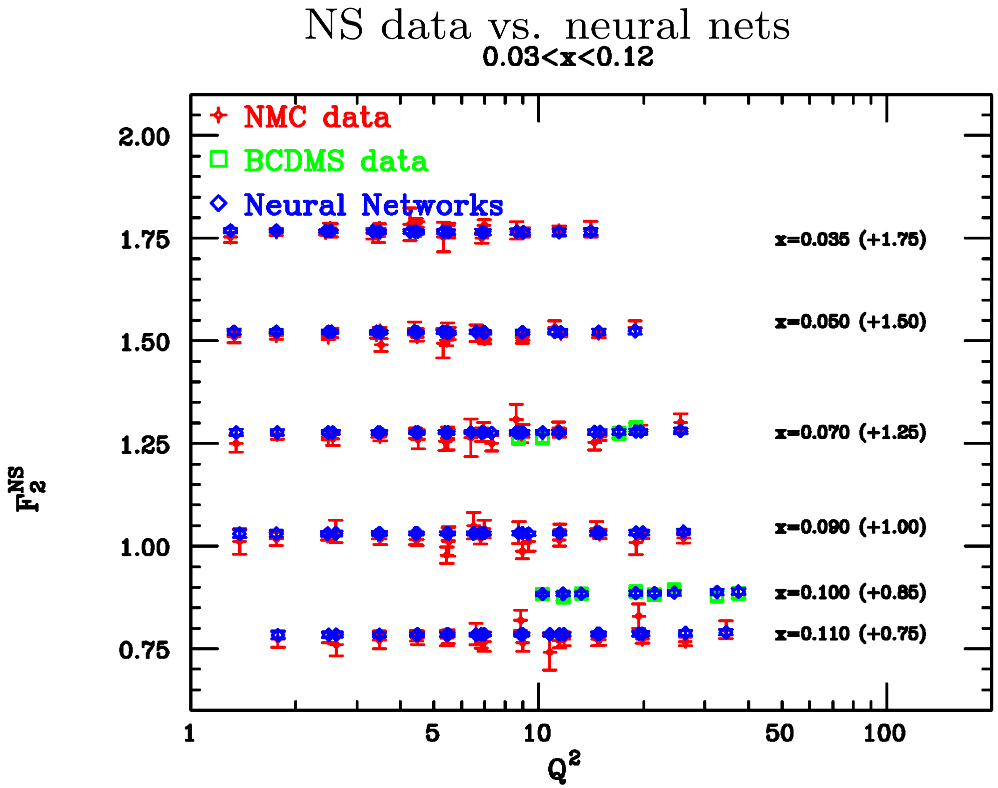
BUT WHAT IF WE HAVE TWO MEASUREMENTS AT (x_i, Q_i^2) WHICH ARE VERY CLOSE?
F_2 IS NOT A FRACTAL!
CLEVER A: •ERROR FUNCTION → USUAL LOG-LIKELIHOOD

$$E^{(k)}[\omega,\theta] = \sum_{i=1}^{N_{dat}} \frac{\left(F_i^{(art)(k)} - F_i^{(net)(k)}\right)^2}{\sigma_{i,s}^{(exp)\,2}}$$

•ESTABLISH FIXED TRAINING LENGTH SUCH THAT $\frac{E^{(k)}[\omega,\theta]}{N_{dat}} \approx 1$

FAITHFUL UNCERTAINTY VS. BIAS 2002

COMBINING DATA



IN NONSINGLET CASE,
AVERAGE VARIANCE OF NETS << STAT. ERROR OF DATA (FACTOR 3–4)
IS IT DUE TO SMOOTHING BIAS?
OR IS IT DUE TO COMBINING DATA?
recall error on weighted average
$\sigma = \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2} < \sigma_i$
CAN CONSTRUCT A STATISTICAL INDICATOR TO TELL!

Average error $\langle E \rangle = \frac{1}{N_{rep}} \sum_{n=1}^{N_{rep}} \sum_{i=1}^{N_{dat}} \frac{\left(F_i^{(art)(n)} - F_i^{(net)(n)}\right)^2}{\sigma_{i,s}^{(exp)\,2}}$ ($n \to$ replica; $i \to$ datapoint)

"Central" error $\langle \tilde{E} \rangle = \frac{1}{N_{rep}} \sum_{n=1}^{N_{rep}} \sum_{i=1}^{N_{dat}} \frac{\left(F_i^{(exp)} - F_i^{(net)(n)}\right)^2}{\sigma_{i,s}^{(exp)\,2}}$

Bias indicator $\mathcal{R} \equiv \langle \tilde{E} \rangle / \langle E \rangle$: if $\sigma_{net} << \sigma_{exp}$ then
$\mathcal{R} \approx 1 \Rightarrow$ BIAS; $\mathcal{R} \approx 1/2 \Rightarrow$ ERROR REDUCTION HERE $\mathcal{R} = 0.58$ (0.53 NMC only)

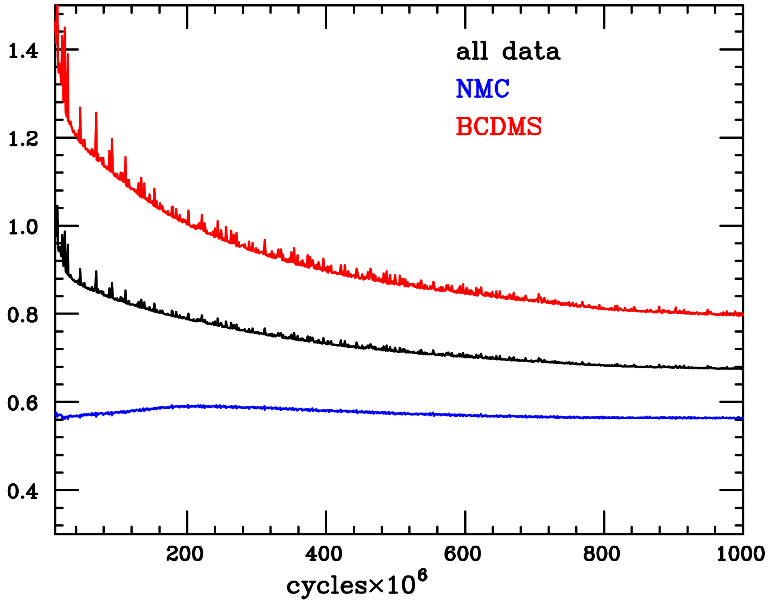
WEIGHTED TRAINING 2002

STUDY DEPENDENCE OF ERROR FCTN $E^{(0)} = \frac{1}{N_{dat}} \sum_{i=1}^{N_{dat}} \frac{\left(F_i^{(exp)} - F_i^{(net)(0)}\right)^2}{\sigma_{i,s}^{(exp)\,2}}$ ON TRAINING LENGTH FOR NET TRAINED ON CENTRAL VALUES

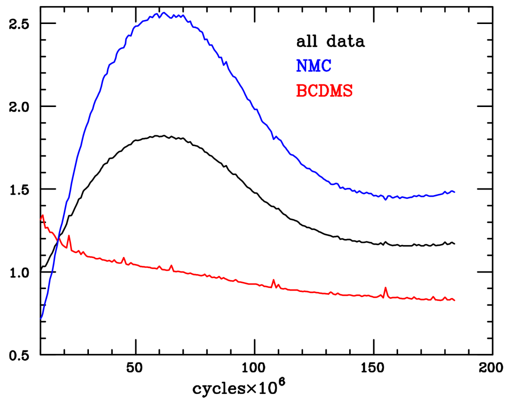
INHOMOGENEOUS ERRORS

NS: AFTER $\sim 10^7$ TRAINING CYCLES, $E^{(0)} \approx 1$ BUT WIDE SPREAD BETWEEN DATASETS
$\Rightarrow$ NMC OVERLEARNT & BCDMS UNDERLEARNT

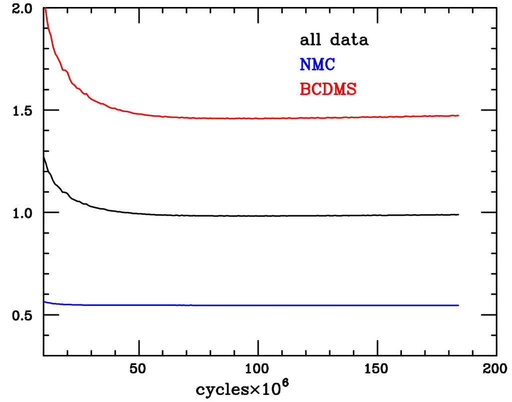
training on all data



training on BCDMS



training on NMC



- EACH DATASET PREDICTS THE OTHER
  $\Rightarrow$ FULL COMPATIBILITY
- BCDMS HARDER TO LEARD THAN NMC (SMALLER ERRORS)

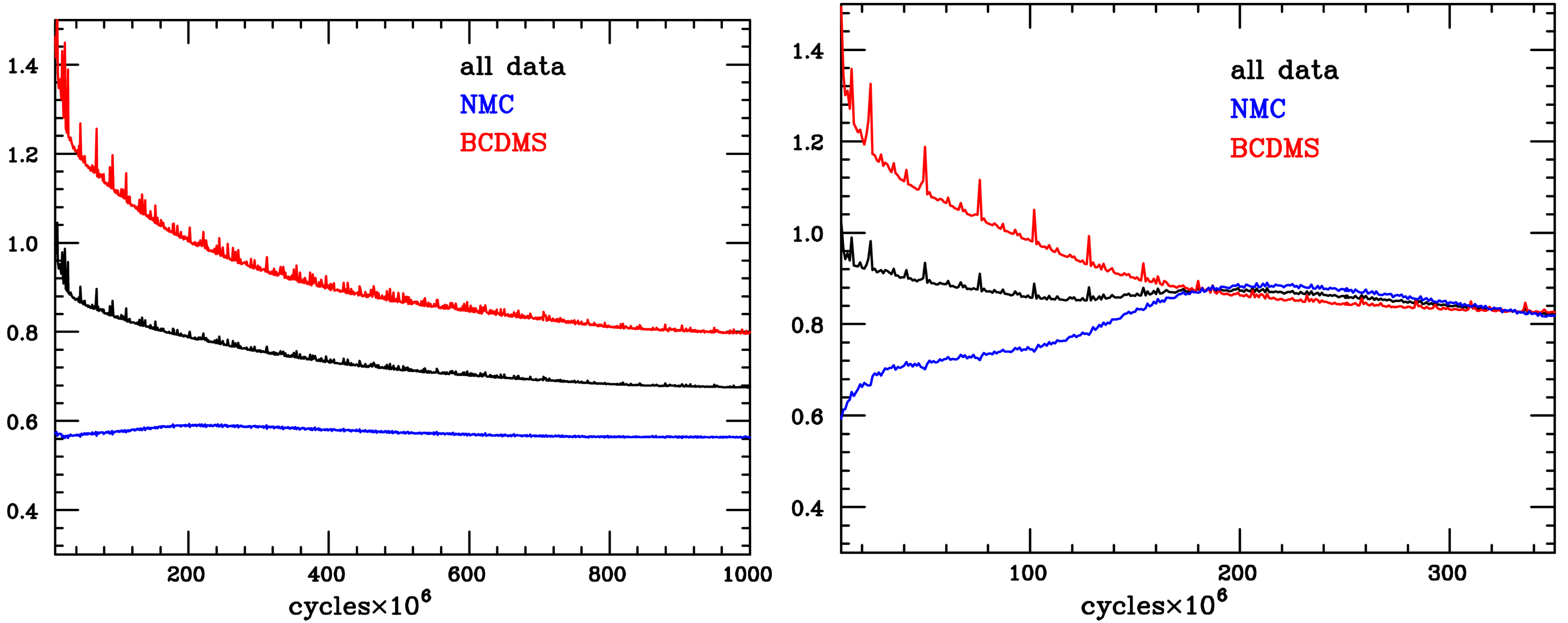
INHOMOGENEOUS ERRORS cont'd

NETS ARE GETTING TRAPPED IN LOCAL MIN. OF THE DATA WHICH ARE LEARNT FASTER

global min. can only be reached at overlearning point

SOLUTION: WEIGHTED TRAINING

uniform training 90% BCDMS 10 % NMC



- convergence of two experiments reached fast by weighted training
- at convergence, $E^{(0)} \approx 1$
- after convergence, $E^{(0)}$ for two experiment slowly improve at same rate, oscillating about each other $\Rightarrow$ global minimum found

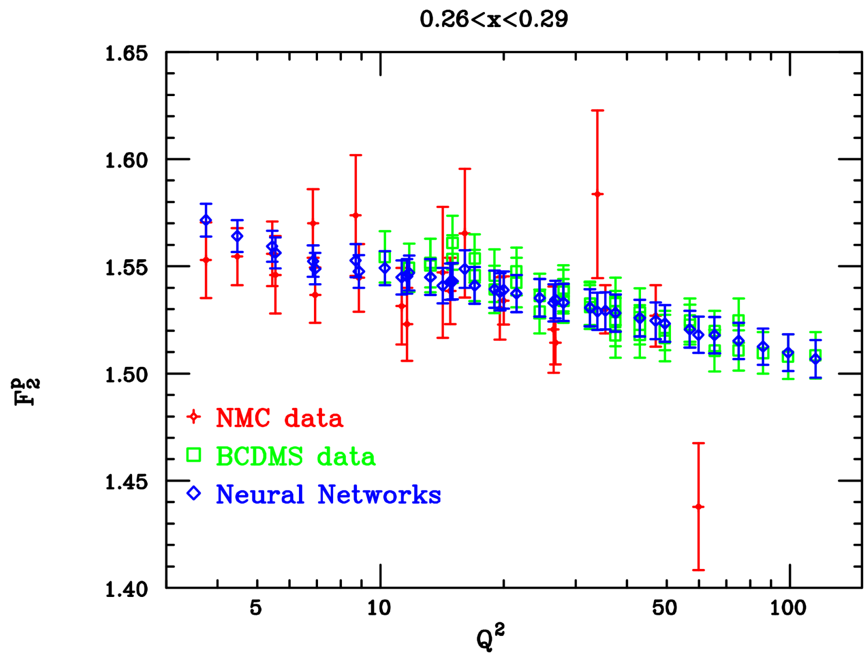
INCOMPATIBLE DATA 2002

INCOMPATIBLE DATA

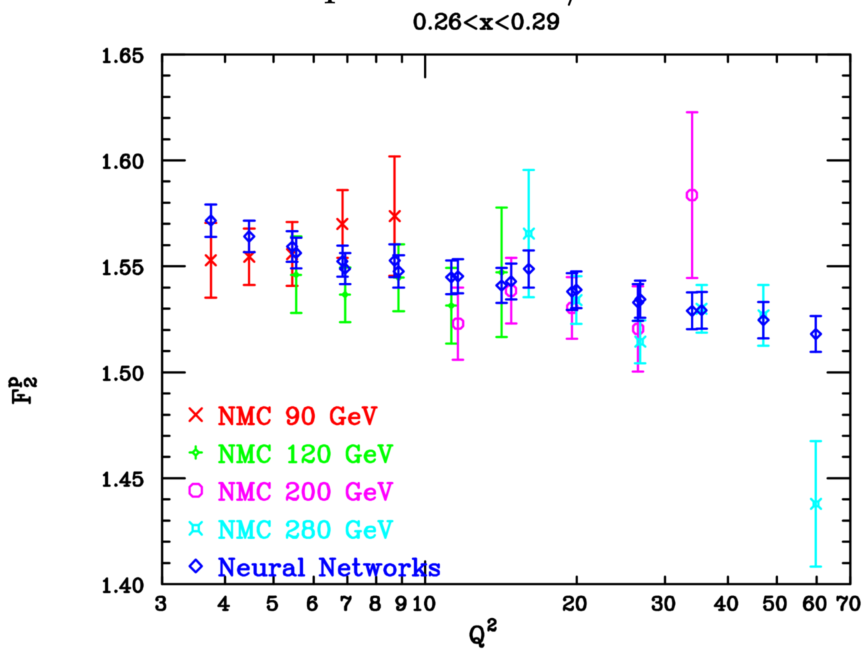
- FOR PROTON FITS, CONVERGENCE ACHIEVED, BUT $E^{(0)} \gtrsim 1.4$ EVEN W. VERY LONG TRAINING
- for NMC data $E^{(0)} \gtrsim 1.6$ (training with all data)
- for NMC data $E^{(0)} \gtrsim 2.2$ (training with NMC only)
- ALL OTHER STATISTICAL INDICATORS OK

SOME NMC DATA ARE INCOMPATIBLE WITH OTHER DATA

Blow-up of proton data/nets



NMC proton data/nets

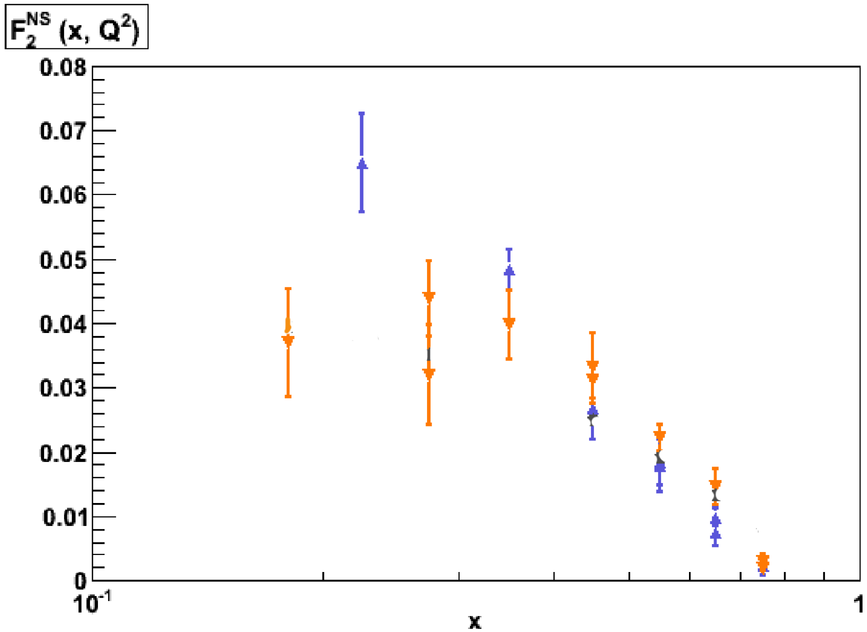


NEURAL NET DISCARDS INCONSISTENT DATA & PROVIDES GOOD FIT TO THE REST

STOPPING 2007

MINIMIZE BY GENETIC ALGORITHM:
AT EACH GENERATION, THE χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

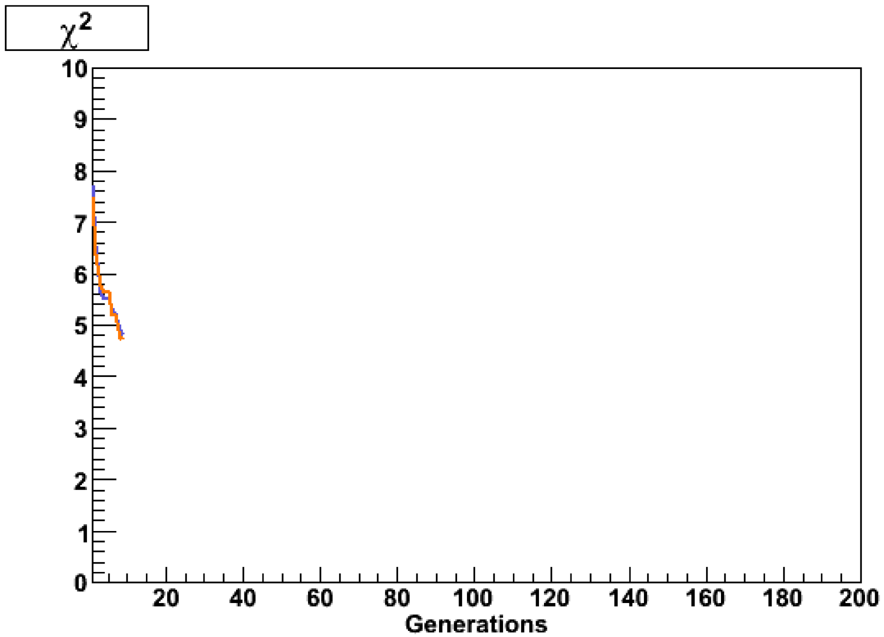


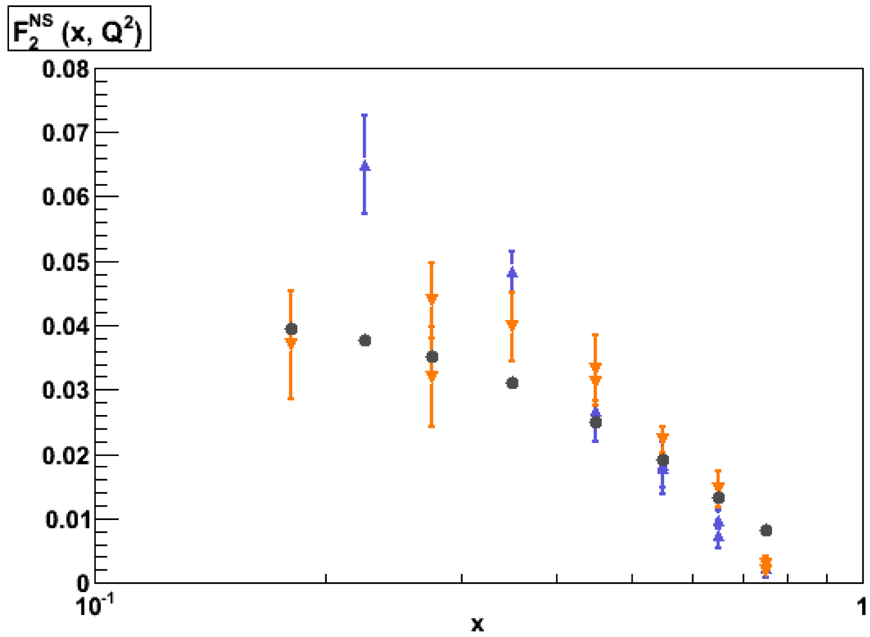
STOPPING 2007

MINIMIZE BY GENETIC ALGORITHM:
AT EACH GENERATION, THE χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

GO!



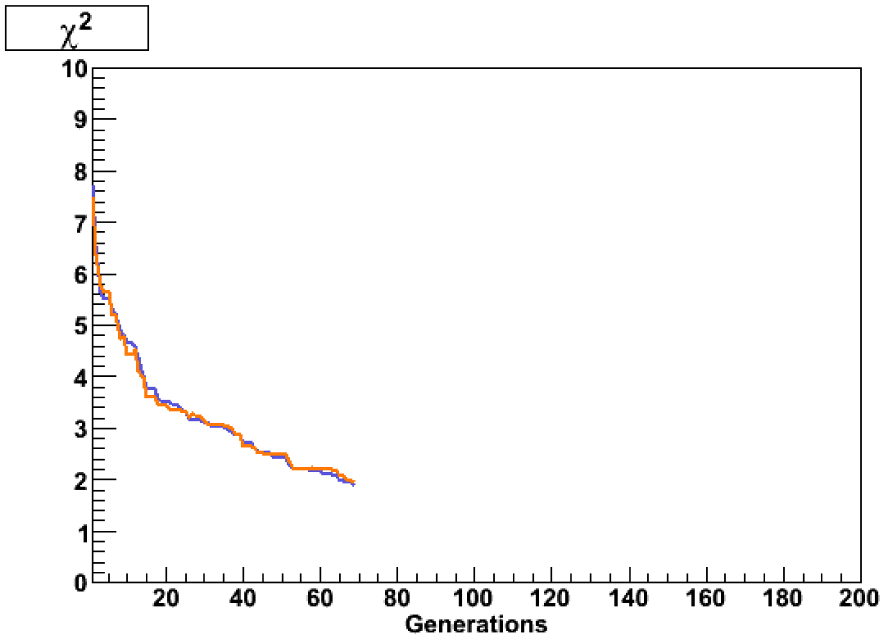


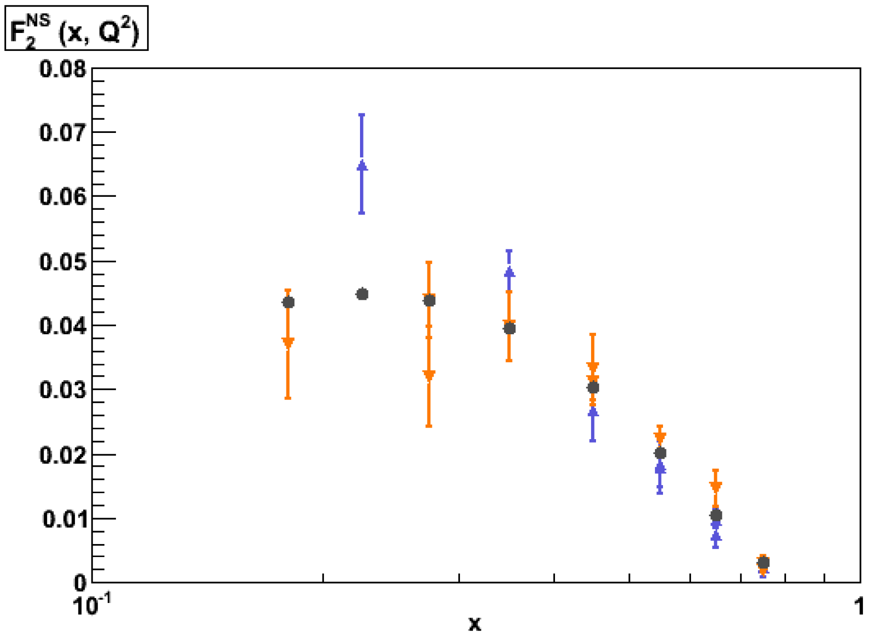
STOPPING 2007

MINIMIZE BY GENETIC ALGORITHM:
AT EACH GENERATION, THE χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

STOP!



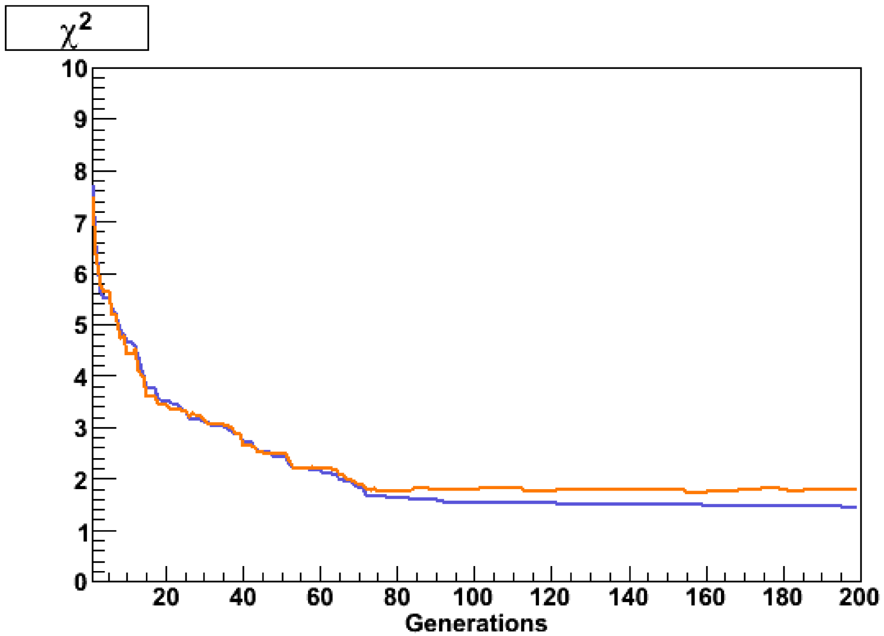


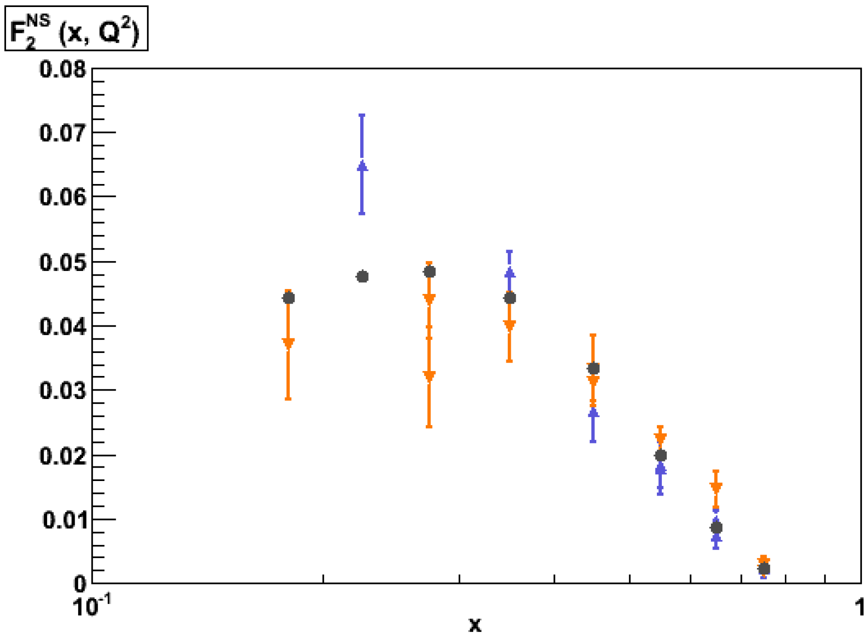
STOPPING 2007

MINIMIZE BY GENETIC ALGORITHM:
AT EACH GENERATION, THE χ^2 EITHER UNCHANGED OR DECREASING

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

TOO LATE!

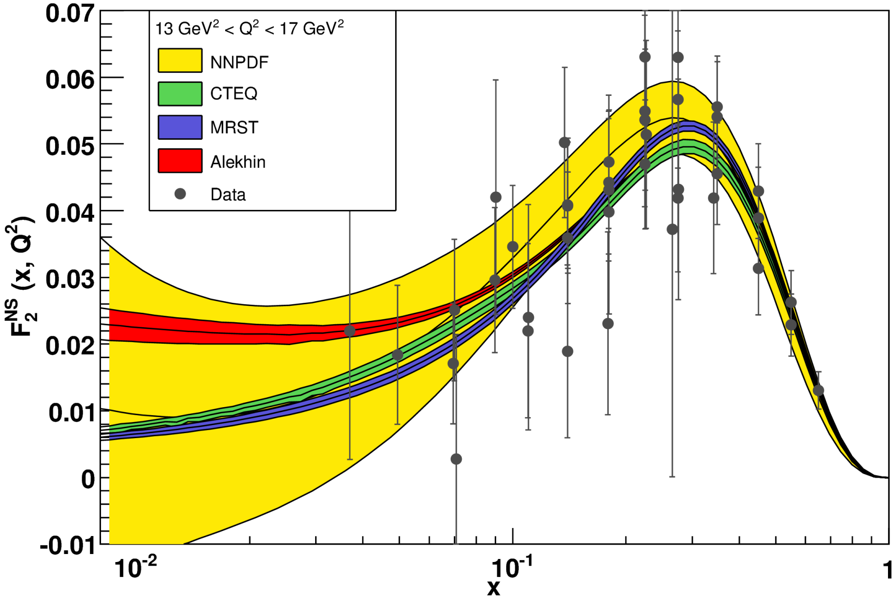




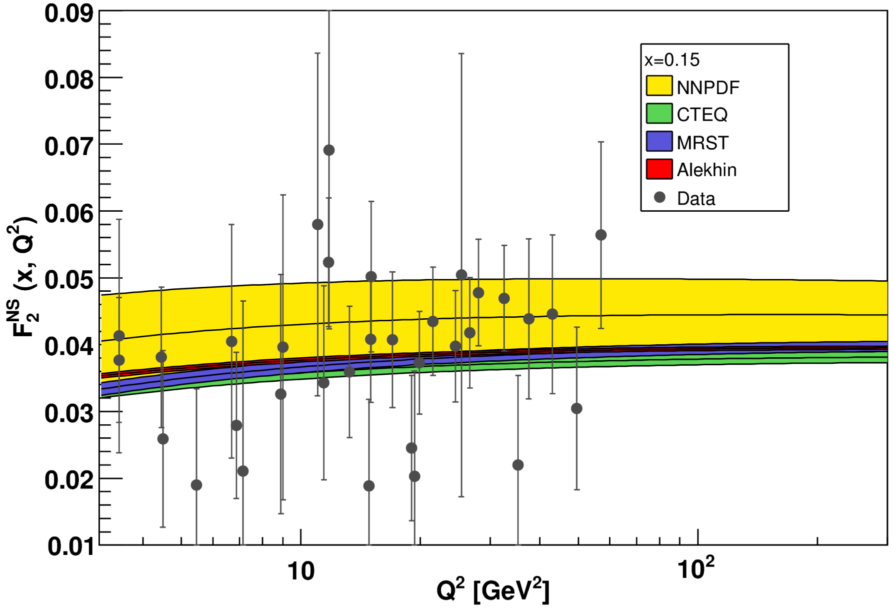
THE FIRST PDF: NONSINGLET (NNPDF 2007)

NLO RESULTS: THE STRUCTURE FUNCTION $F_2^{\mathrm{NS}}(x, Q^2)$

VS x AT $Q^2 = 15\ \mathrm{GeV}^2$



VS Q^2 AT $x = 0.15$

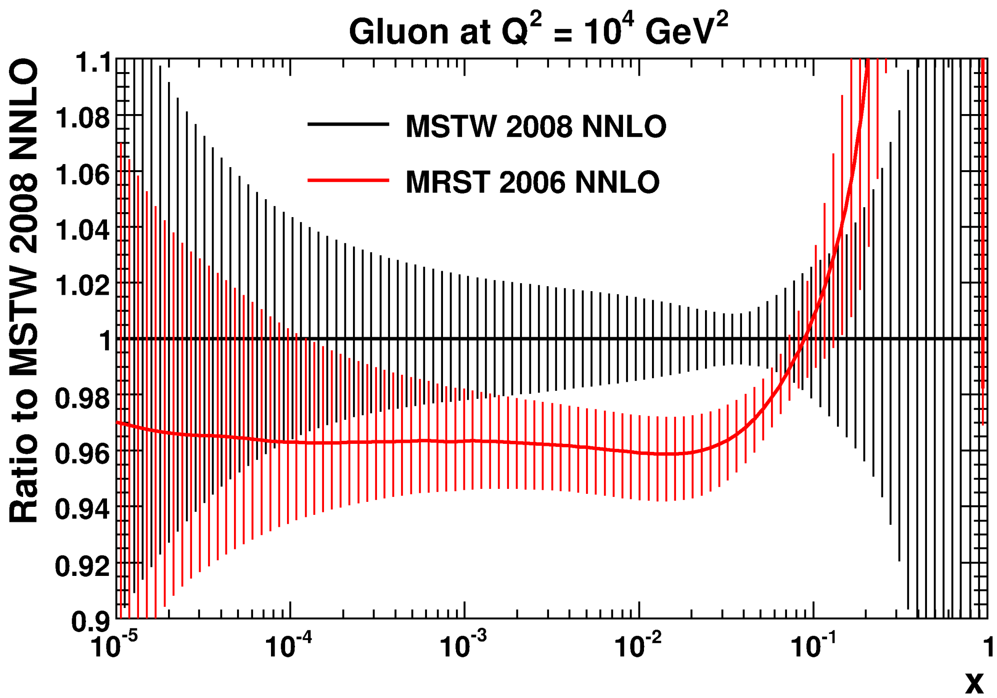


- COMPATIBLE WITH EXISTING FITS WITHIN ERROR
  (even when they disagee with each other)
- UNCERTAINTY MUCH LARGER IN EXTRAPOLATION BUT ALSO IN DATA REGION
  (note no other global fit data constrain q_{NS})
- CENTRAL FIT DISAGREES WITH EXISTING FITS IN VALENCE REGION
  $0.1 \le x \le 0.3$

THE DYSON PROBLEM: 2008

UNCERTAINTIES IN MSTW/CTEQ FITS OFTEN GO UP WHEN DATA ARE ADDED, BECAUSE OF THE NEED TO ADD PARAMETERS

Smaller high-x gluon (and slightly smaller α_S) results in larger small-x gluon – now shown at NNLO.

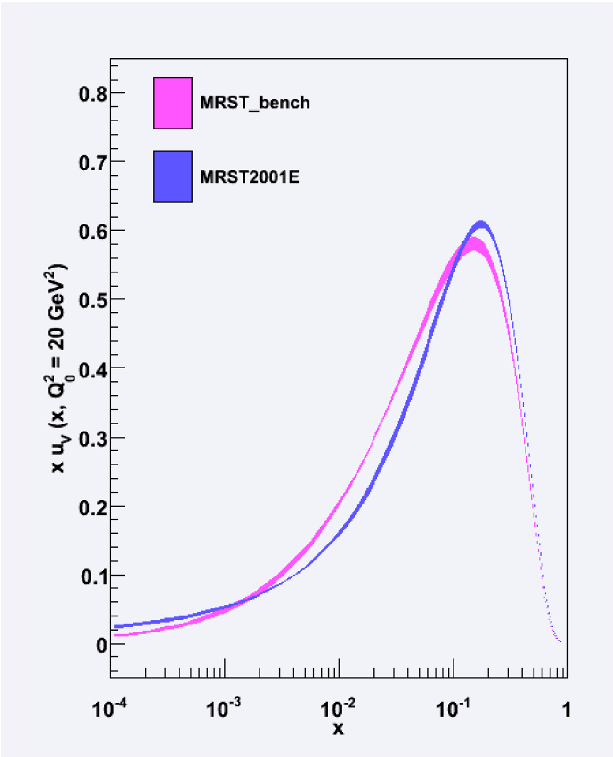


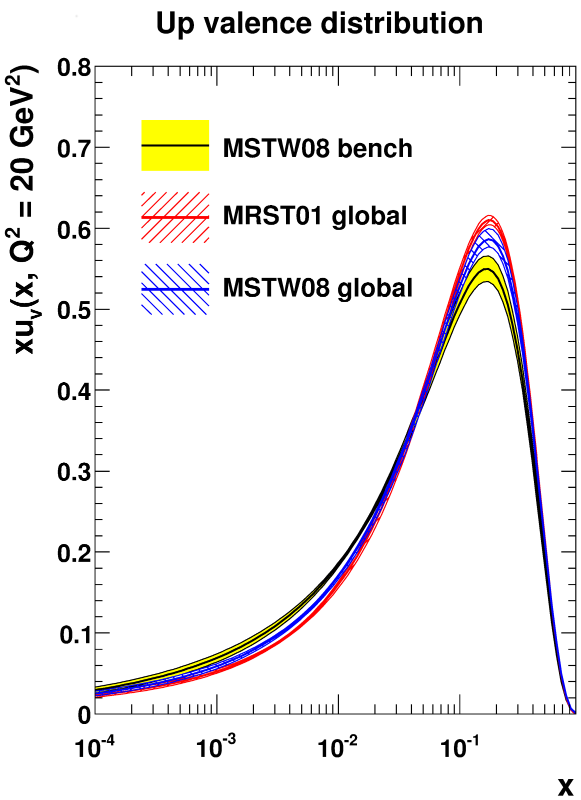
Larger small-x uncertainty due to extrat free parameter.

R. Thorne, HERALHC2008

THE 2ND HERALHC BENCHMARK (2008)

- PERFORM A MRST (MRSTBENCH) FIT TO A CONSISTENT SUBSET OF DATA, USE $\Delta\chi^2 = 1$
  $\Rightarrow$ RESULTS NOT CONSISTENT, UNCERTAINTY DOES NOT GROW AS DATASET DECREASES
- ...BUT MRST WAS DONE WITH TOLERANCE 50: REPEAT WITH DYNAMICAL TOLERANCE (MSTW08BENCH)
- IMPROVEMENT, BUT PROBLEM NOT SOLVED
  $\Rightarrow$ MUST TUNE PARAMETRIZATION AND STATISTICAL TREATMENT TO DATASET





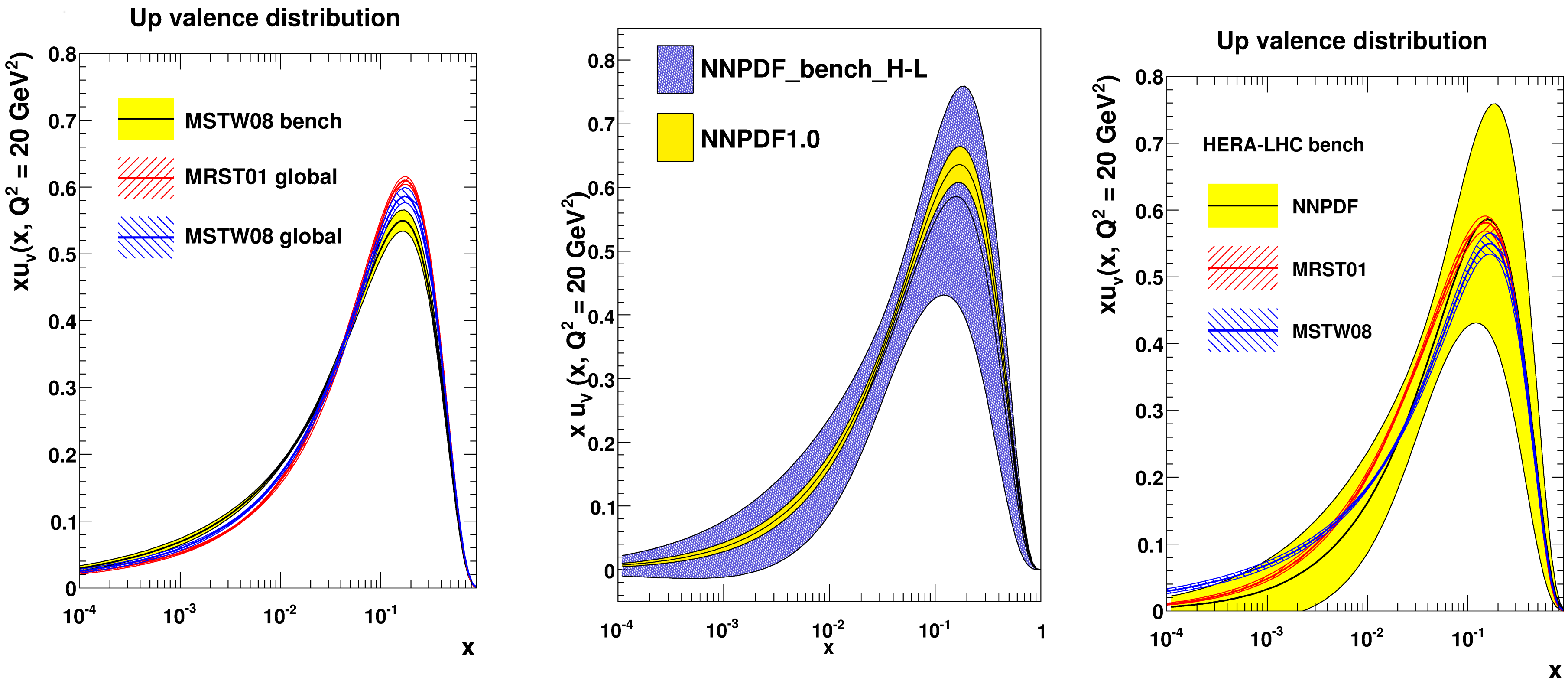
WHAT DETERMINES PDF UNCERTAINTIES?

THE NNPDF1.0 REVOLUTION (2008)

MRST/MSTW: BENCH VS REF

NNPDF: BENCH VS REF

NNPDF BENCH VS MRST/MSTW BENCH



- SINGLE PARAMETRIZATION AND STAT. TREATMENT CAN ACCOMMODATE DIFFERENT DATASETS
- IMPACT OF DATA CAN BE STUDIED INDEPENDENT OF THEORETICAL FRAMEWORK

“FUNCTIONAL” PDF UNCERTAINTIES
THE 2008 PDF4LHC NNPDF STUDY

Thanks to J. Pumplin

- FIT TO REPLICAS VS. FIT TO DATA PARTITIONS $\Leftrightarrow$
  $\Leftrightarrow$FLUCTUATION OF DATA (TRUE) VS. FLUCTUATION OF REPLICAS (NOMINAL)
- FIT TO PARTITIONS VS. FIT TO A SINGLE PARTITION $\Leftrightarrow$
  $\Leftrightarrow$ UNCERTAINTY DUE TO DATA VS. UNCERTAINTY DUE TO OTHER SOURCES
- OPTIMAL FIT VS. OVERLEARNING FIT $\Leftrightarrow$
  $\Leftrightarrow$ UNDERLYING LAW VS. STATISTICAL NOISE

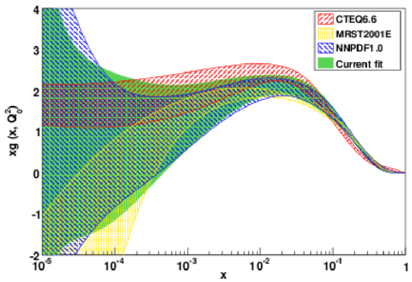
WHERE IS THE UNCERTAINTY COMING FROM?

FIT TO REPLICAS VS RANDOM SUBSET OF CENTRAL VAL.S

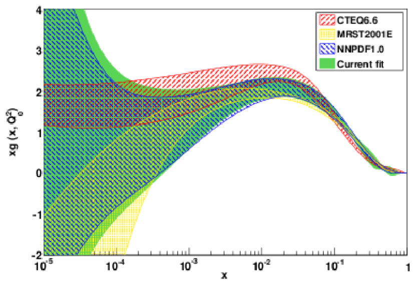
	REPLICAS	CENTRAL V.
χ^2	1.32	1.32
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035

GLUE

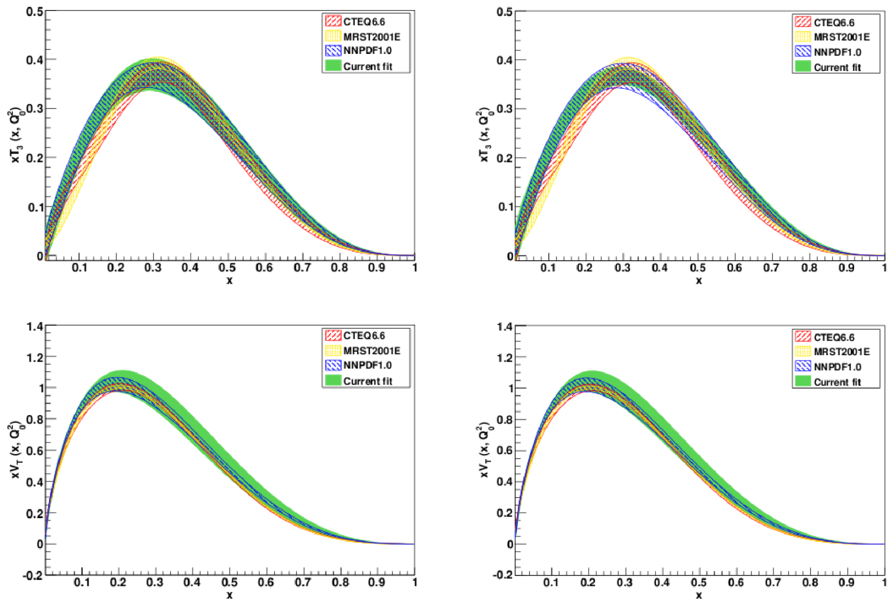
replicas



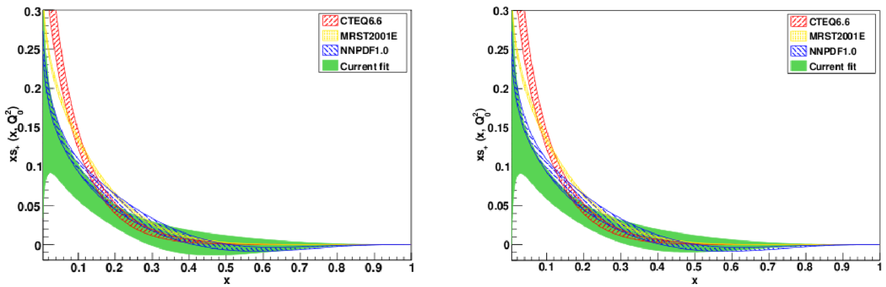
c. vals.



LIGHT QUARKS



STRANGE



- QUALITY OF FIT &PDFs UNCHANGED
- REDUCTION OF $\langle\chi^2\rangle_{\rm rep}$ BY FACTOR $\sim 2 \Rightarrow$ FLUCTUATIONS ABOUT TRUE VALUE HALVED
- UNCERTAINTY ON DATA ONLY REDUCED BY 1.1 $\Rightarrow$ EXPT. UNCERTAINTIES UNDERESTIMATED OR UNDERLYING INCOMPRESSIBLE UNCERTAINTY

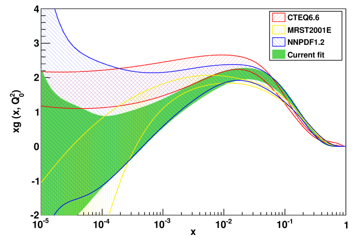
WHERE IS THE UNCERTAINTY COMING FROM?

CENTRAL VALUES: VARYING PARTITION VS FIXED PARTITION

	REPLICAS	CENTRAL VALUE	FIXED PARTITION
χ^2	1.32	1.32	$\sim$1.3
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	$\sim 1.6 \pm 0.2$
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035	$\sim$0.03

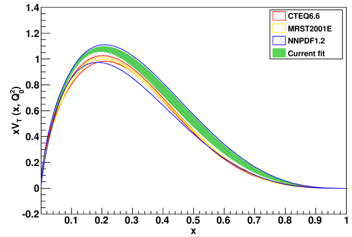
fixed partition results obtained averaging over 5 different choices of partition (100 replicas each); more partitions needed for accurate results

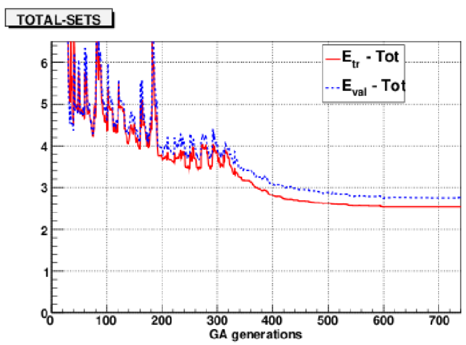
GLUE

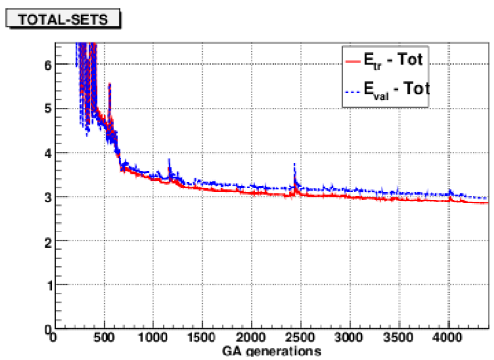


- QUALITY OF FIT UNCHANGED
- $\langle\chi^2\rangle_{\rm rep}$ UNCHANGED $\Rightarrow$ CENTRAL FIT UNCHANGED
- UNCERTAINTY ON PREDICTION (I.E. ON PDFS) REDUCED

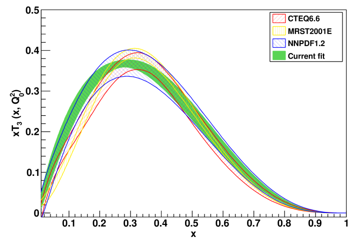
VALENCE





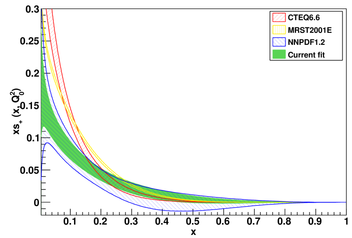


TRIPLET



FUNCTIONAL UNCERTAINTY

STRANGE



- MORE THAN HALF OF UNCERTAINTY DUE TO "FUNCTIONAL FORM": $\langle\sigma^{\rm dat}\rangle = \sim 0.3$ SMALLER FOR HERA DATA
- REMAINING UNCERTAINTY ROUGHLY SCALES WITH DATA UNCERTAINTY: $\langle\sigma^{\rm dat}\rangle = \sim 0.005$ CENT.; $\langle\sigma^{\rm dat}\rangle = \sim 0.009$ REP.

ARE WE CONSTRAINED BY THE FUNCTIONAL FORM?

REMOVE STOPPING: OVERLEARNING FIT

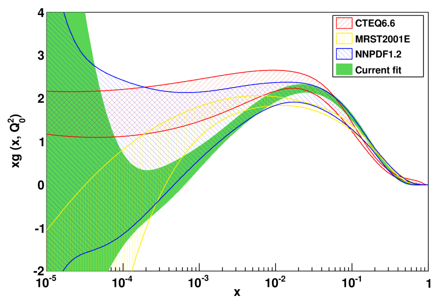
PERFORM A FIT WITH A FIXED, VERY LARGE NUMBER OF GA GENERATIONS:
25000 gens. (AVERAGE 1000 gens. FOR STANDARD FIT)

	STANDARD STOPPING			FIXED LONG	
	REPLICAS	CENTRAL VALUE	FIXED PARTITION	REPLICAS	CENTRAL VALUE
χ^2	1.32	1.32	$\sim$1.3	1.18	1.19
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	$\sim 1.6 \pm 0.2$	2.43 ± 0.13	1.29 ± 0.06
$\langle\chi^2_{\rm tr}\rangle_{\rm rep}$	2.76	1.59	$\sim$1.6	2.40	1.27
$\langle\chi^2_{\rm val}\rangle_{\rm rep}$	2.80	1.61	$\sim$1.6	2.47	1.30
$\langle\sigma^{\rm dat}\rangle$	0.039	0.035	$\sim$0.03	0.032	0.019

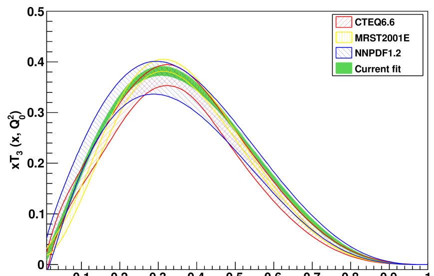
χ^2 OF THE GLOBAL FIT DECREASES A LOT!
IS IT REALLY OVERLEARNING?

- PERCENTAGE DIFFERENCE BETWEEN VALIDATION AND TRAINING $\langle\chi^2\rangle_{\rm rep}$ MORE THAN DOUBLED (FROM 1.5% TO 3%) (NOTE 1650 DATA POINTS EACH)
- SOME PDFs HAVE FUNNY SHAPES
- REDUCTION OF $\langle\sigma^{\rm dat}\rangle$ BY FACTOR $1.7 > \sqrt{2}$ WHEN GOING FROM REPLICAS TO CENTRAL VALUES
- AMOUNT OF OVERLEARNING SMALL, $\Leftrightarrow$ $\langle\chi^2\rangle_{\rm rep}$ DOUBLES WHEN GOING FROM CENTRAL VALS. TO REPLICAS, SHOULD REMAIN UNCHANGED FOR EXTREME OVERLEARNING

GLUON



TRIPLET



YES!

WHERE IS THE UNCERTAINTY COMING FROM?

WHEN THE BEST FIT IS NOT AT THE MINIMUM

	STANDARD STOPPING			FIXED LONG	
	REPLICAS	CENTRAL VALUE	FIXED PARTITION	REPLICAS	CENTRAL VALUE
χ^2	1.32	1.32	1.35	1.18	1.19
$\langle\chi^2\rangle_{\rm rep}$	2.79 ± 0.24	1.65 ± 0.20	1.60 ± 0.19	2.43 ± 0.13	1.29 ± 0.06
$\langle\sigma^{\rm dat}\rangle$	0.39	0.35	0.28	0.32	0.19

- FIT QUALITY:
  - "FUNCTIONAL" UNCERTAINTY SUPPRESSED IN OVERLEARNING FITS:
    $\Rightarrow \langle\sigma^{\rm dat}\rangle \approx 0.2 \Rightarrow$ "DATA" UNCERTAINTY
  - FLUCTUATION OF $\langle\chi^2\rangle_{\rm rep}$ FOR OVERLEARNING FIT STATISTICAL:
    $$\sigma = \sqrt{\frac{2}{N_{\rm dat}}} \approx 0.05$$
  - FLUCTUATION OF $\langle\chi^2\rangle_{\rm rep}$ IN STANDARD FIT MUCH LARGER:
    CONTROLLED BY DISTANCE FROM THE MINIMUM
    IF $\Delta\chi^2 = 1$ DUE TO UNDERLYING PARM AT $\chi^2_{\rm min}$, THEN ONE SIGMA VARIATION AROUND
    $\chi^2_0 > \chi^2_{\rm min}$ EQUALS $\sqrt{\chi^2_0 - \chi^2_{\rm min}}$
- DATA INCONSISTENCY: FOR STANDARD FIT, VALUE OF $\chi^2 = 1.3 > 1$
  $\Rightarrow$ ERRORS UNDERESTIMATED BY 30%

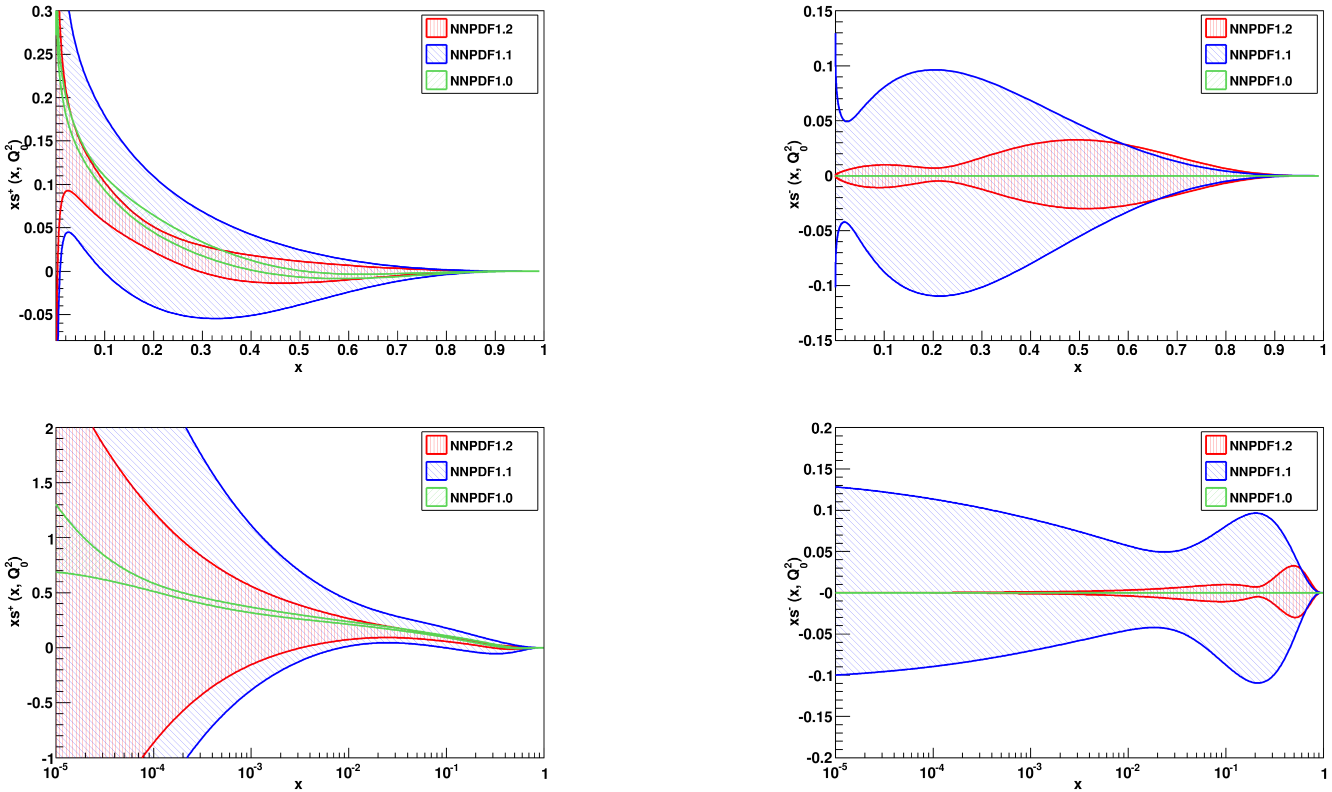
FROM NNPDF1.0 TO NNPDF1.2: STRANGENESS

- STRANGENESS ALMOST UNCONSTRAINED BY INCLUSIVE DIS DATA
  NNPDF1.1: s, $\bar{s}$ (actually $s^{\pm}$) indep. parametrized, no dimuon data

- IN PARTON FITS UP TO 2009 $\rightarrow$ STRANGENESS FIXED BY ASSUMPTION
  NNPDF1.0: $s(x, Q_0^2) = \bar{s}(x, Q_0^2)$, $s + \bar{s} = \frac{1}{2}(\bar{u} + \bar{d})$

- IN CURRENT PARTON FITS $\rightarrow$ STRANGENESS FIXED BY DIS DIMUON PRODUCTION $\nu + s \rightarrow c$
  NNPDF1.2: s, $\bar{s}$ (actually $s^{\pm}$) indep. parametrized, dimuon data

STRANGE PDFS



PDFs AND α_s: NNPDF1.2 (2009)

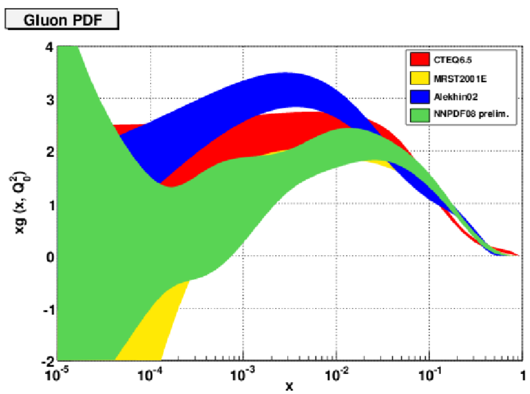
NO α_s DEPENDENCE

- DETERMINE DISTANCE d IN UNITS OF STANDARD DEVIATION OF THE MEAN $s = \frac{\sigma}{\sqrt{N_{\rm rep}}}$ FOR EACH PDF$\Rightarrow$ TWO DIFFERENT SUBSETS OF REPLICAS OF SAME FIT $\langle d \rangle = 1$
- DISTANCE BETWEEN PDFS WITH $\Delta\alpha_s = \pm 0.002$ COMPATIBLE WITH STATISTICAL FLUCTUATIONS for all PDFs, central values and uncertainties, data and extrapolation regions
- RECOMMENDED: TO ESTIMATE UNCERTAINTY, VARY α_s WITH PDFs FIXED AT STANDARD NNPDF SET

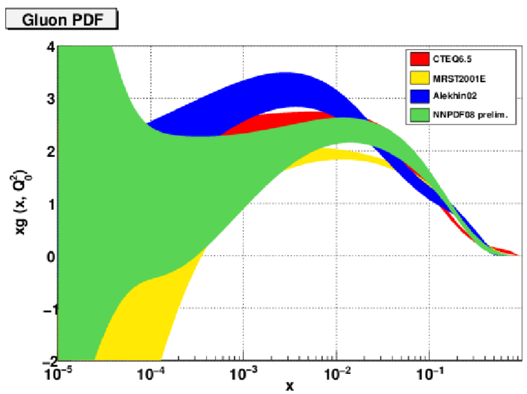
$\alpha_s(M_Z^2)$	0.117		0.121	
χ^2	1.35		1.33	
	Data	Extra	Data	Extra
$\Sigma(x, Q_0^2)$				
$\langle d[q] \rangle$	1.72	1.05	0.73	0.81
$\langle d[\sigma] \rangle$	1.05	1.03	1.22	0.95
$g(x, Q_0^2)$				
$\langle d[q] \rangle$	4.68	2.29	4.12	0.71
$\langle d[\sigma] \rangle$	1.00	0.91	0.88	0.83
$T_3(x, Q_0^2)$				
$\langle d[q] \rangle$	0.71	0.71	1.55	0.96
$\langle d[\sigma] \rangle$	0.93	0.75	1.11	0.78
$V(x, Q_0^2)$				
$\langle d[q] \rangle$	0.92	0.74	1.89	1.72
$\langle d[\sigma] \rangle$	0.94	0.71	0.67	0.65
$\Delta_S(x, Q_0^2)$				
$\langle d[q] \rangle$	0.74	0.58	0.86	1.36
$\langle d[\sigma] \rangle$	0.67	0.83	0.78	0.76

NNPDF GLUON

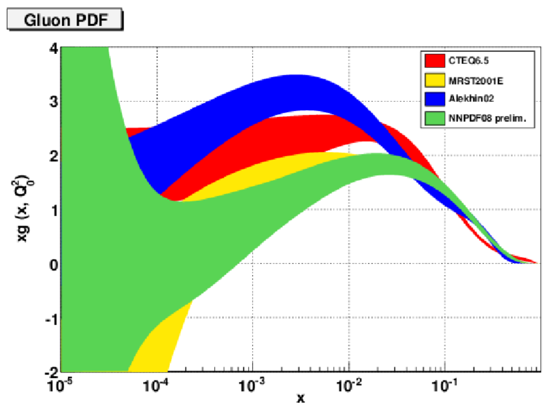
$\alpha_s = 0.119$



$\alpha_s = 0.117$



$\alpha_s = 0.121$



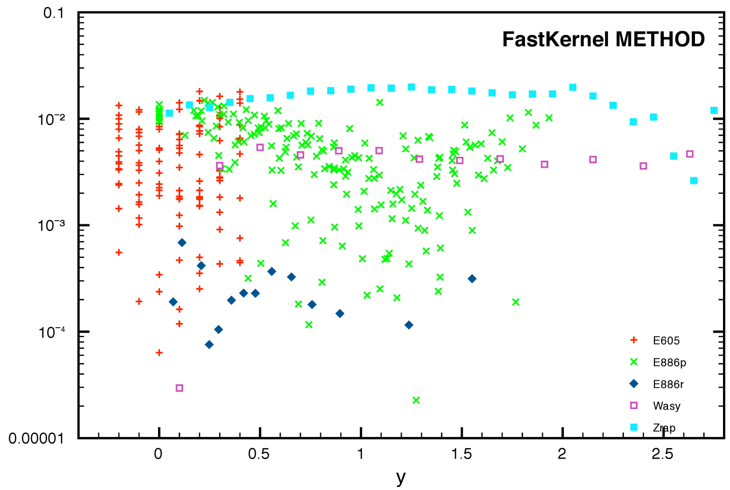
THE CONTEMPORARY ERA: 2010 AND BEYOND
NNPDF2.0: THE PROBLEM OF NLO FITTING

- OTHER EXISTING GLOBAL FITS ARE NOT FULLY NLO
  MSTW, CTEQ: DRELL-YAN TREATED AT LO+ K–FACTORS
- OTHER EXISTING NLO FITS ARE NOT GLOBAL
  HERAPDF ONLY DIS, ALEKHIN (ABKM) DIS+SOME FIXED-TARGET DRELL-YAN
- BOTTLENECK: FAST COMPUTATION OF DOUBLE CONVOLUTIONS FOR HADRONIC PROCESSES
  MELLIN SPACE APPROACH INCOVENIENT FOR JETS, & FOR GENERAL PARTON PARAMETRIZATIONS

NNPDF2.0: THE FIRST GLOBAL NLO FIT

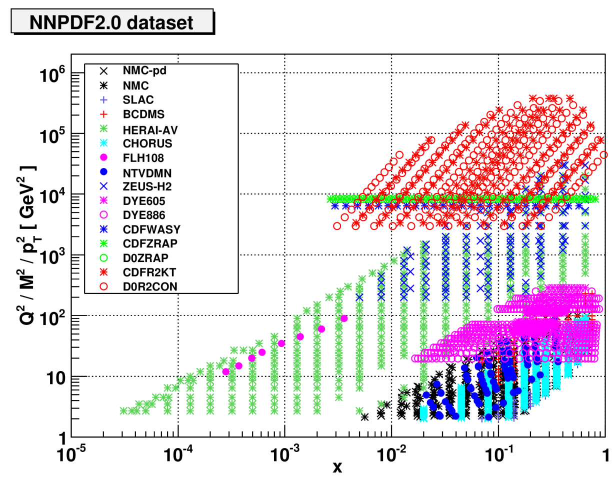
- GRID-BASED METHODS: EXPANSION OF PDFS ON BASES OF POLYNOMIALS, PRECOMPUTE CONVOLUTION WITH BASIS FUNCTIONS (Pascaud, Zomer, 2001)
- FASTNLO: FAST INTERFACE FOR JET CROSS SECTIONS (Kluge, Rabbertz, Wobisch 2006)
- NNPDF2.0 USES FASTKERNEL: GRID METHOD INTERFACED TO N-SPACE COMPUTATION OF GLAP GREEN FUNCTIONS, INTERFACED TO FASTNLO FOR JETS AND TO SUITABLE FAST-DY (NNPDF, 2010)
- MORE RECENTLY APPLGRID: OPTIMIZED GRID, POTENTIALLY UNIVERSAL INTERFACE, IMPLEMENTED FOR JETS, W AND Z PRODUCTION (Carli et al., 2010)

FASTKERNEL PERC. ACCURACY

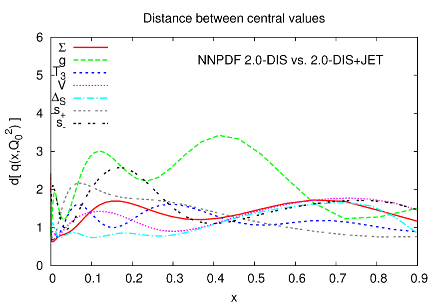


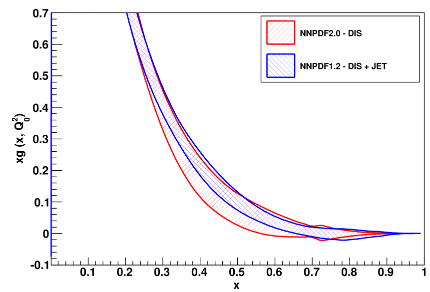
THE POWER OF GLOBAL FITTING: CONSISTENCY
DIS DATA VS. JET DATA

	DIS	DIS+JET	NNPDF2.0
χ^2_{tot}	1.20	1.18	1.21
NMC-pd	0.85	0.86	0.99
NMC	1.69	1.66	1.69
SLAC	1.37	1.31	1.34
BCDMS	1.26	1.27	1.27
HERAI	1.13	1.13	1.14
CHORUS	1.13	1.11	1.18
FLH108	1.51	1.49	1.49
NTVDMN	0.71	0.75	0.67
ZEUS-H2	1.50	1.49	1.51
CDFR2KT	**0.91**	**0.79**	0.80
D0R2CON	**1.00**	**0.93**	0.93
DYE605	7.32	10.35	0.88
DYE866	2.24	2.59	1.28
CDFWASY	13.06	14.13	1.85
CDFZRAP	3.12	3.31	2.02
D0ZRAP	0.65	0.68	0.47



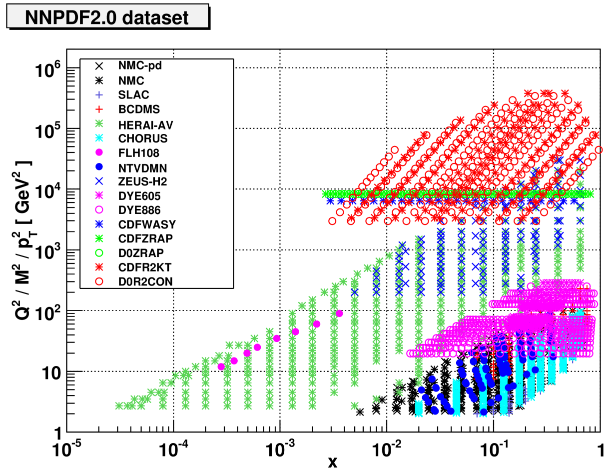
- HIGH E_T JET DATA WELL REPRODUCED EVEN WHEN NOT FITTED $\Rightarrow$ LARGE x GLUON WELL DETERMINED BY SCALING VIOLATIONS!
- SIGNIFICANT IMPROVEMENT IN LARGE x GLUON ACCURACY
- OTHER PDFs UNCHANGED



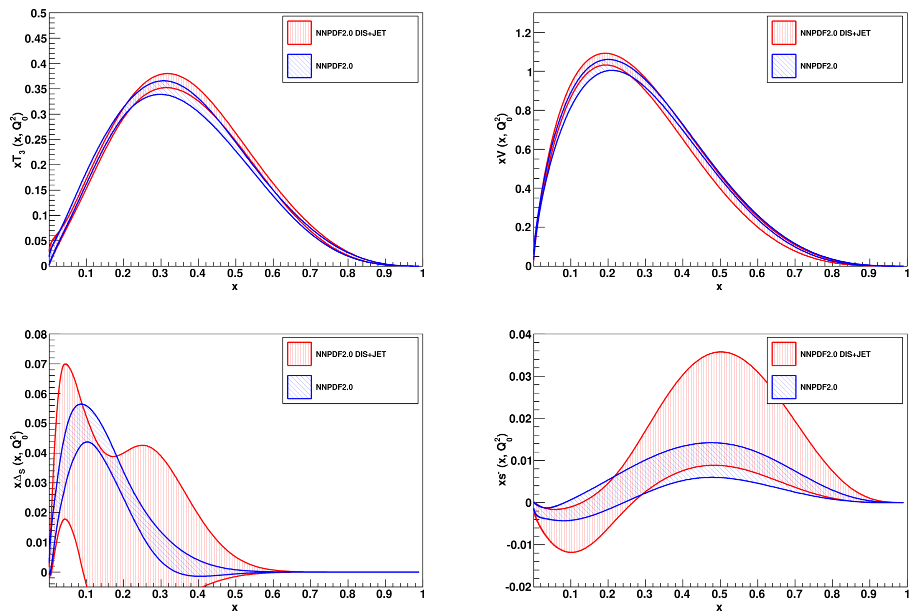


THE POWER OF GLOBAL FITTING: ACCURACY
DIS+JETS VS. DRELL-YAN (AND W, Z) DATA

	DIS	DIS+JET	NNPDF2.0
$\chi^2_{\rm tot}$	1.20	1.18	1.21
NMC-pd	0.85	0.86	0.99
NMC	1.69	1.66	1.69
SLAC	1.37	1.31	1.34
BCDMS	1.26	1.27	1.27
HERAI	1.13	1.13	1.14
CHORUS	1.13	1.11	1.18
FLH108	1.51	1.49	1.49
NTVDMN	0.71	0.75	0.67
ZEUS-H2	1.50	1.49	1.51
CDFR2KT	0.91	0.79	0.80
D0R2CON	1.00	0.93	0.93
DYE605	7.32	**10.35**	**0.88**
DYE866	2.24	**2.59**	**1.28**
CDFWASY	13.06	**14.13**	**1.85**
CDFZRAP	3.12	**3.31**	**2.02**
D0ZRAP	0.65	**0.68**	**0.47**



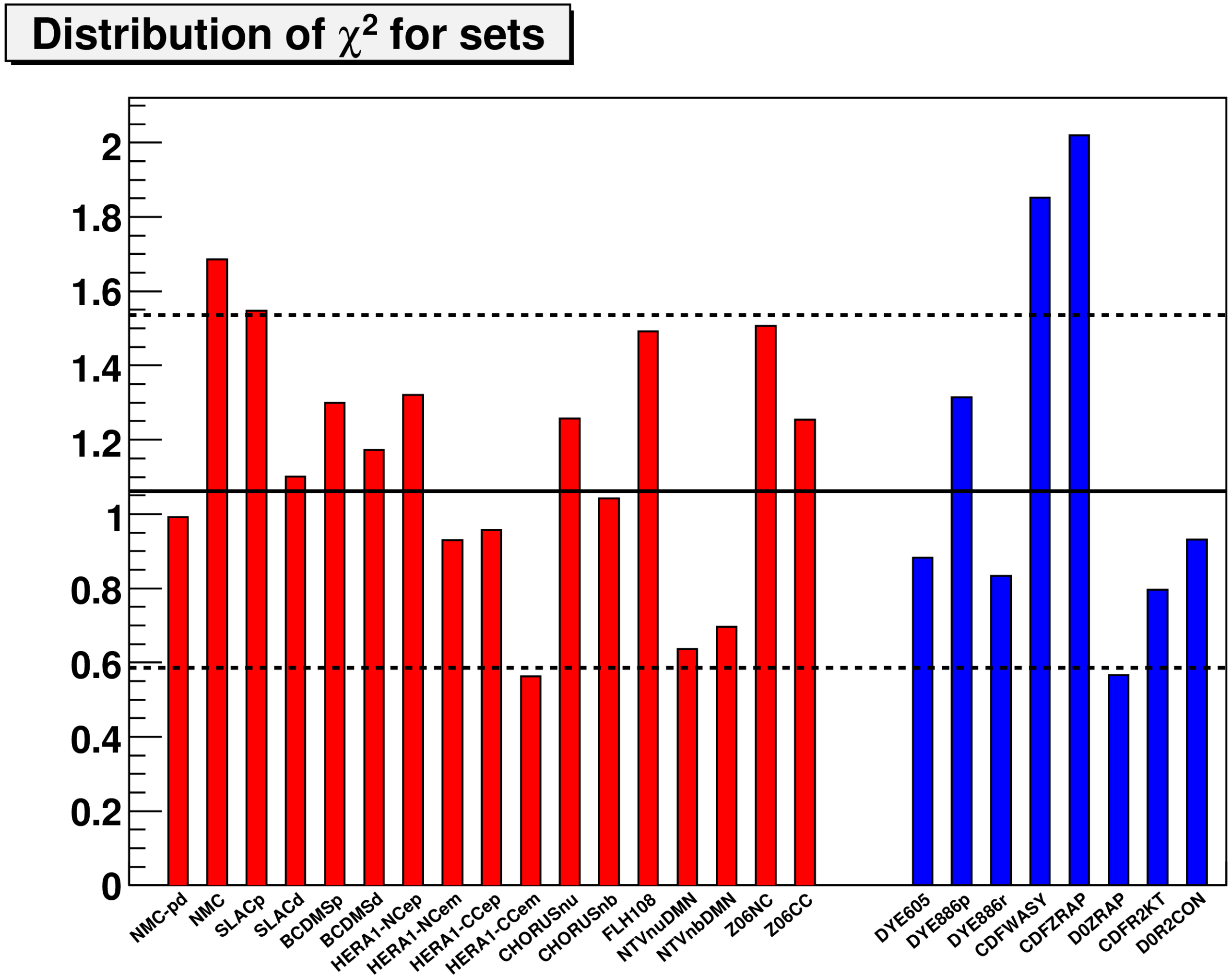
- VERY SUBSTANTIAL IMPROVEMENT IN FIT QUALITY WHEN DATA INCLUDED $\Rightarrow$ SOME PDF COMBINATIONS POORLY DETERMINED WITHOUT THESE DATA
- HUGE IMPROVEMENT IN SEA ASYM $\bar{u}-\bar{d}$ & STRANGENESS $s-\bar{s}$
- SIGNIFICANT IMPROVEMENT IN TOTAL VALENCE $(\sum_i (q_i - \bar{q}_i))$ & ISOTRIPLET $(u+\bar{u}-(d+\bar{d}))$



GLOBAL FITS & DATA COMPATIBILITY

FIT QUALITY: DIS DATA AND HADRONIC DATA

NNPDF2.0



- NO OBVIOUS MUTUAL TENSION BETWEEN DIS AND HADRONIC DATA
- CLEAR SIGN OF INTERNAL DATA INCONSISTENCIES (NMC DIS DATA, CDF Z AND W RAPIDITY DISTRIBUTIONS)

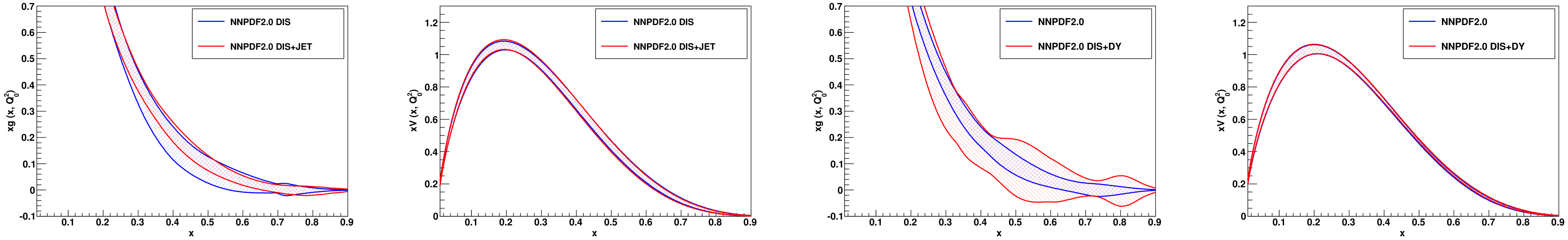
GLOBAL FITS & DATA COMPATIBILITY

DIS VS. HADRONIC DATA

A SENSITIVE TEST: IS THE IMPACT OF A DATASET INDEP. OF THE DATA IT IS ADDED TO?

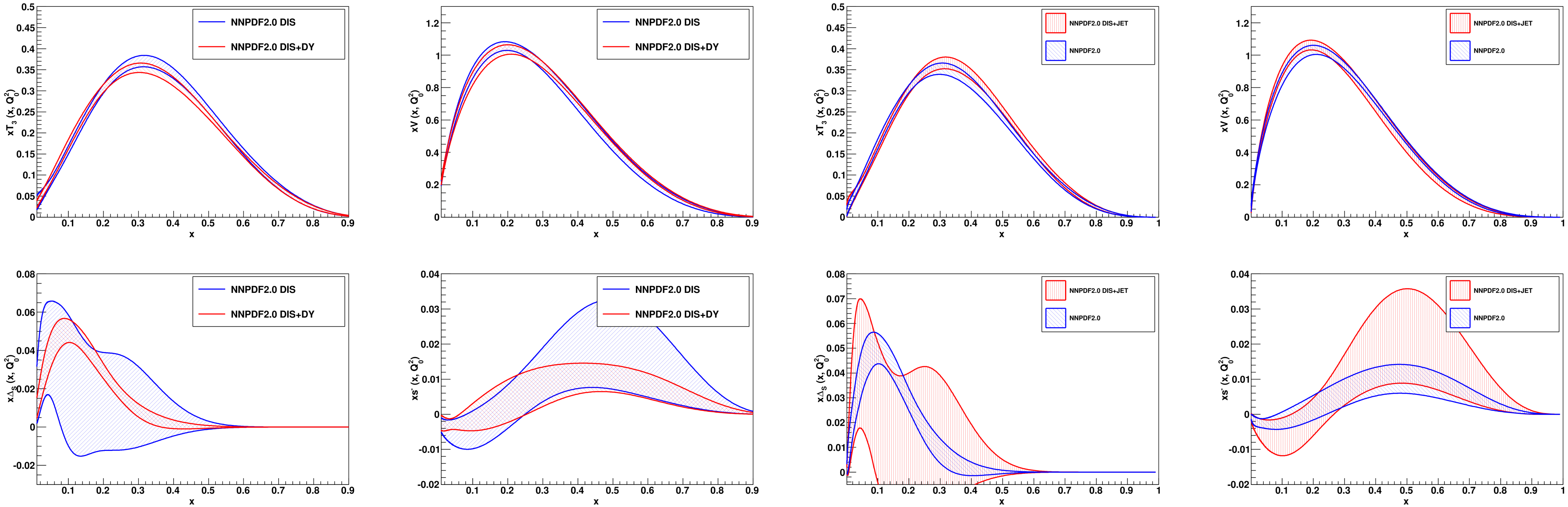
ADDING JET DATA. . .

. . . TO DIS DATA . . . TO DIS+DY DATA



ADDING DRELL-YAN DATA. . .

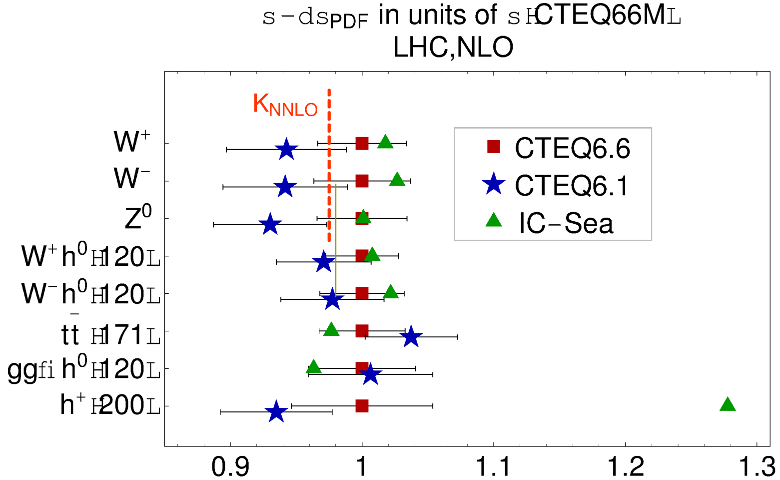
. . . TO DIS DATA . . . TO DIS+JET DATA



FITS COMMUTE $\Rightarrow$ GOOD COMPATIBILITY!

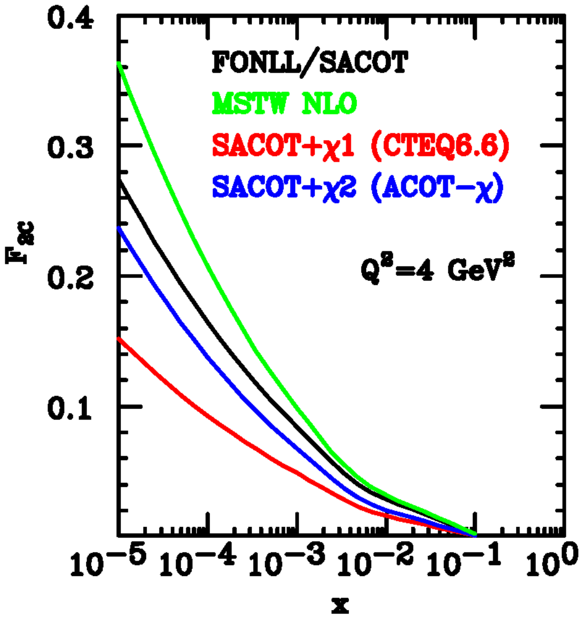
HEAVY QUARK MASS EFFECTS
FONLL & NNPDF2.1

- MANY FITS (CTEQ<6, NNPDF, ALEKHIN<09) TREAT CHARM AS MASSLESS ABOVE THRESHOLD ⇒ "ZMVFN" SCHEME
- COMBINED MATCHED SCHEMES AVAILABLE SINCE LONG (ACOT94, FONLL98) INCLUDING CHARM MASS ALONG WITH LL RESUMMATION; ALTERNATIVE TR/TR' PROCEDURE IMPLEMENTED SINCE '98 IN MRST
- WHEN CTEQ IMPLEMENTED ACOT IN 2008, SURPRISING CHANGE CTEQ61→CTEQ6.5 IN σ_W, & AGREEMENT WITH MRST SPOILED (LATER RESTORED)



(Nadolsky et al., 2008)

RECENT PROGRESS:
THE LES HOUCHES 2009 BENCHMARKS

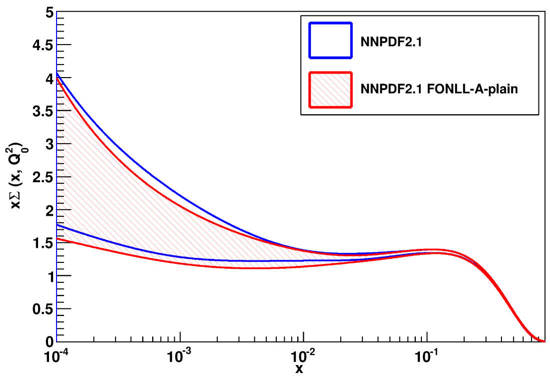


(Rojo et al., 2010)

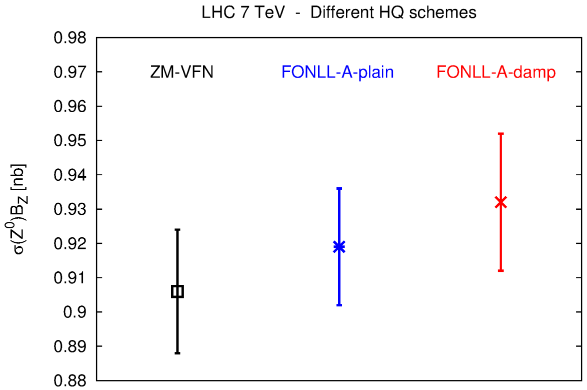
- FONLL PRESCRIPTION RECENTLY ALSO IMPLEMENTED FOR DIS, AVAILABLE TO $O(\alpha_s^2)$ (LIKE TR/TR', ACOT ONLY TO $O(\alpha_s)$)
- $O(\alpha_s)$ FONLL, ACOT COINCIDE EXACTLY, TR' DIFFERS BY SUBLEADING $O(\alpha_s^2(m_c))$ Q^2–INDEP. TERM
- VARIOUS PRESCRIPTIONS FOR HANDLING SUBLEADING TERMS ("χ–SCALING"): DIFFERENCES SIZABLE

HQ AMBIGUITIES ON PDFS (2010)

CORRELATION m_c/PDFS



DEP. OF CHARM ON m_c



- FURTHER UNCERTAINTY ON CHARM → LIGHT PDFS DUE TO SUBLEADING TERMS AT THRESHOLD
- DAMPED SUBLEADING TERMS ⇒ LESS CHARM ⇒ MORE LIGHT QUARKS (LARGER σ_Z)

TOTAL UNCERTAINTY

- PDF+m_c+THRESHOLD SUBL TERMS
- $\Delta m_c = 0.1$ GEV (PDG POLE MASS ∼ 0.15 GEV, HOWEVER UNCERTAINTY CAN BE REDUCED BY CLEVER USE OF $\overline{MS}$ MASS (ALEKHIN, MOCH, 2010))+
- PDF UNCERTAINTY INCREASED BY ∼ 30% DUE TO HQ AT LHC7 (∼ 40% AT LHC14)

LHC 7 TEV	$W^+B_{l\nu}$ [NB]	$W^-B_{l\nu}$ [NB]	$Z^0B_{l\bar{l}}$ [NB]
$m_c = 1.414$ GEV	5.99 ± 0.14	4.09 ± 0.09	0.932 ± 0.02
$m_c = 1.5$ GEV	6.06 ± 0.17	4.14 ± 0.12	0.943 ± 0.024
$m_c = 1.6$ GEV	6.11 ± 0.14	4.17 ± 0.10	0.951 ± 0.020
$m_c = 1.7$ GEV	6.14 ± 0.14	4.19 ± 0.09	0.956 ± 0.019
$\delta_{\rm PDF}$	0.14	0.09	0.019
$\delta_{\rm PDF+m_c}$	0.15	0.10	0.021
$\delta_{\rm PDF+m_c+GM}$	0.19	0.12	0.025
$\rho\,[\sigma, m_c]$	0.44	0.41	0.48

LESSON FOR LHC

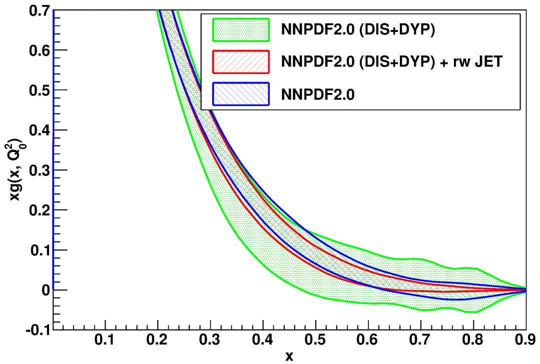
- HQ MASS & THRESHOLD BEHAVIOUR AFFECT HIGH ENERGY OBSERVABLES THROUGH THE SIZE OF THE CHARM AND BOTTOM PDFS
- COMBINED HERA DATA WILL LEAD TO MUCH IMPROVED KNOWLEDGE OF BOTH
- HOWEVER, EVENTUALLY, IT WOULD BE BETTER TO DETERMINE THE c AND b SIZE AT LHC WITHOUT HAVING TO RELY ON LOW-ENERGY DATA

NNPDF2.1:REWEIGHTING (2011)

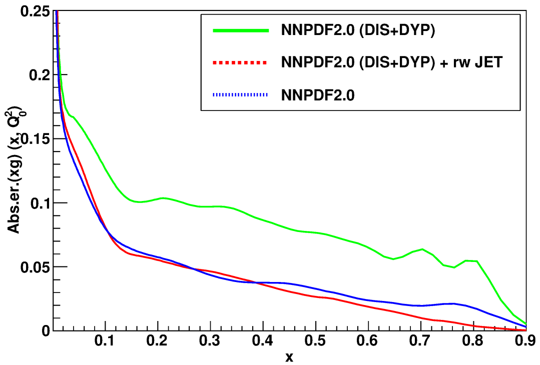
INCLUSION OF JET DATA: REWEIGHTING VS. REFITTING

NNPDF2.0DIS+DY VS. NNPDF2.0FULL

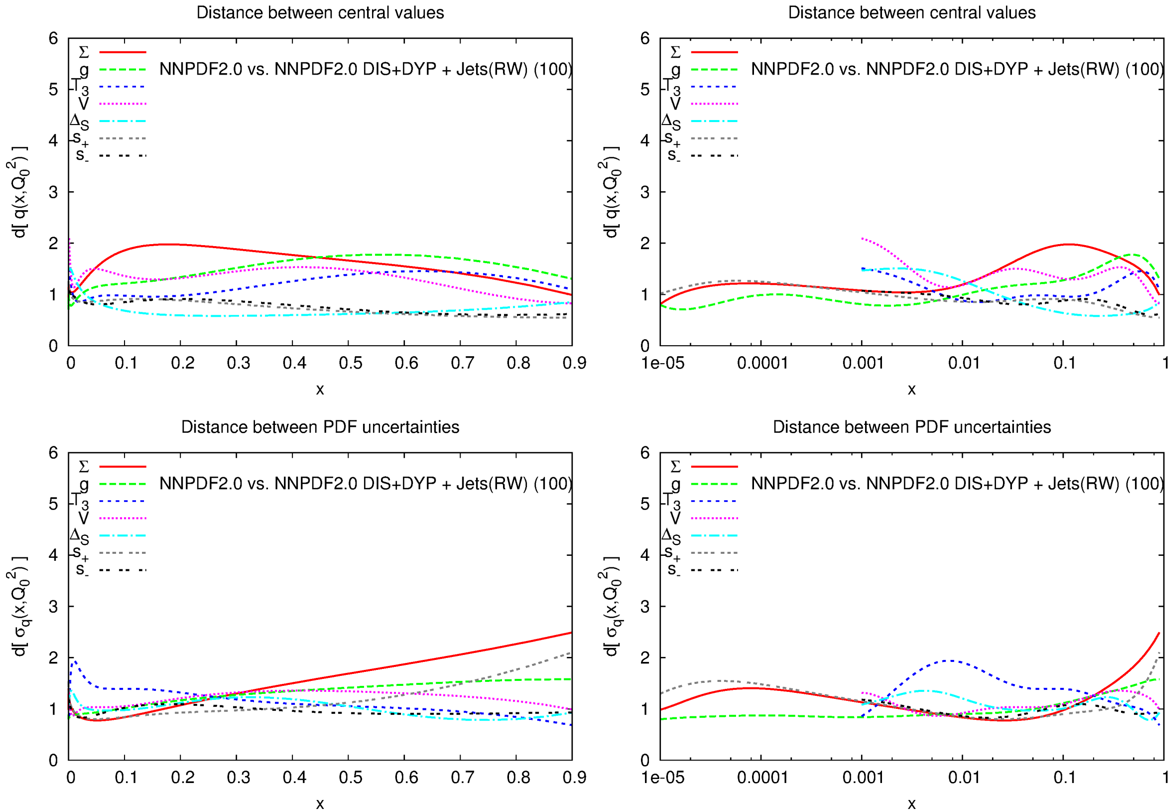
GLUON



GLUON UNCERTAINTY



DISTANCES



$d \sim 1 \Rightarrow$ STATISTICAL EQUIVALENCE

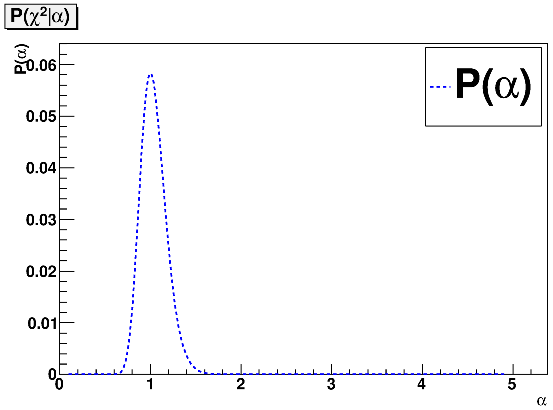
$(d = n \Leftrightarrow n\ \sigma$ DISCREPANCY)

STATISTICAL INTERPRETATION VALIDATED

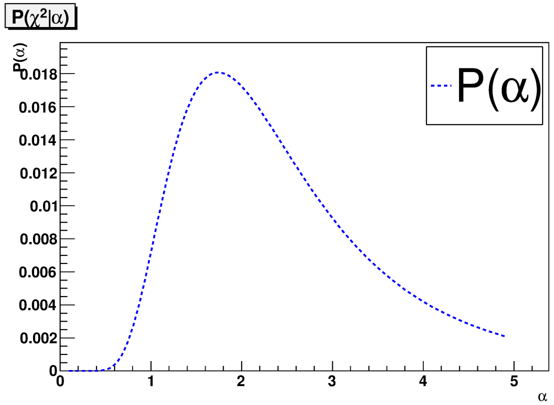
REWEIGHTING & DATA CONSISTENCY

- INCONSISTENT DATA $\Leftrightarrow$ UNDERESTIMATED UNCERTAINTIES
- RESCALE ALL UNCERTAINTIES IN A GIVEN EXPERIMENT BY SOME FACTOR α: $\chi^2_\alpha = \chi^2/\alpha$ (TOLERANCE)
- DETERMINE PROBABILITY DISTRIBUTION OF α VALUES BY BAYES' THEOREM $\Rightarrow$ REWEIGHTING: $\mathcal{P}(\alpha) = \frac{\mathcal{N}}{\alpha}\sum_{k=1}^{N} w_k w_k(\alpha)$.

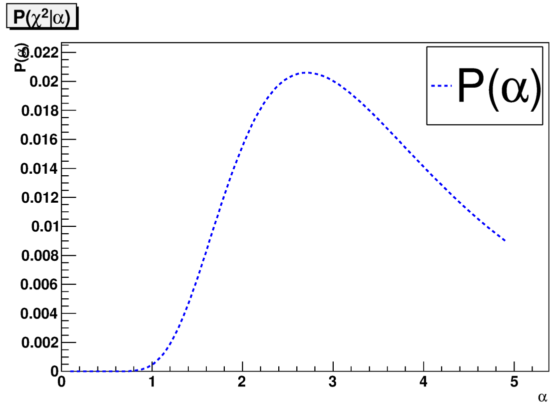
JETS



D0 e INCLUSIVE



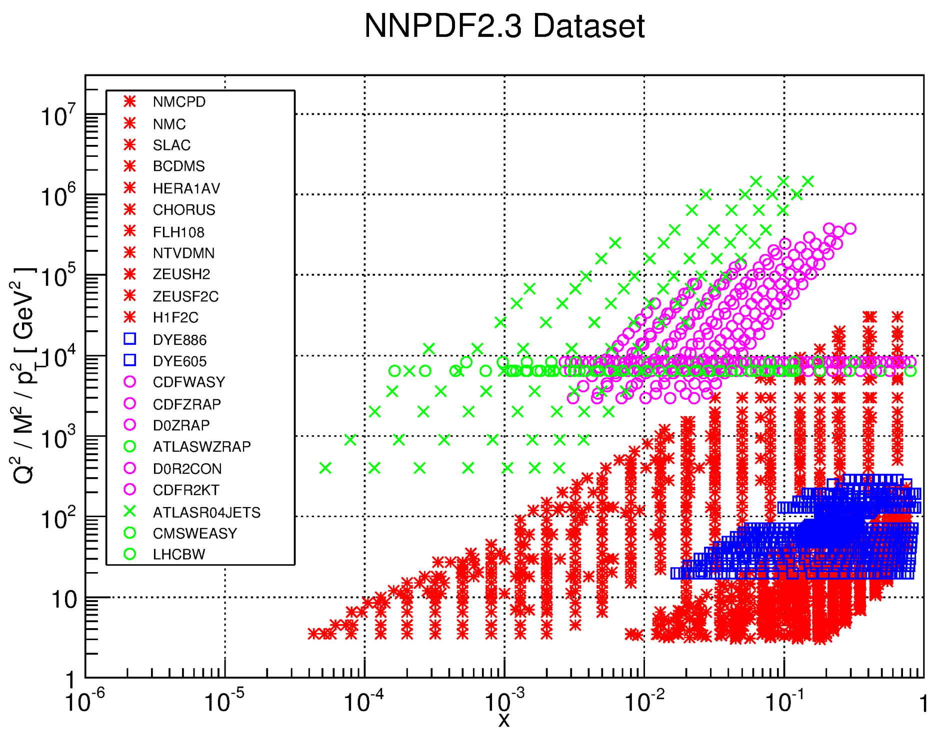
D0 e HIGH E_T



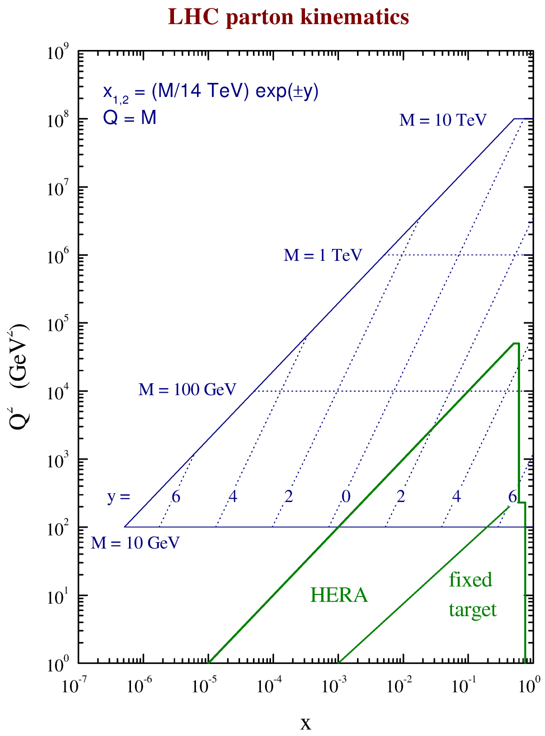
- JETS: $\Rightarrow$ CONSISTENT DATA
- $W^\pm$ CHARGE ASYMMETRIES, D0 INCLUSIVE e DATA $\Rightarrow$ UNCERTAINTIES UNDERESTIMATED BY $\sim 30\%$ (PROB. PEAKS AT $\alpha \sim 1.6$)
- $W^\pm$ CHARGE ASYMMETRIES, D0 e DATA WITH $E_T > 35$ GEV $\Rightarrow$ INCONSISTENT DATA

NNPDF2.3: LHC DATA!

$$\sigma_X(s, M_X^2) = \sum_{a,b} \int_{x_{\min}}^1 dx_1\, dx_2\, f_{a/h_1}(x_1) f_{b/h_2}(x_2) \hat{\sigma}_{q_a q_b \to X}\left(x_1 x_2 s, M_X^2\right)$$



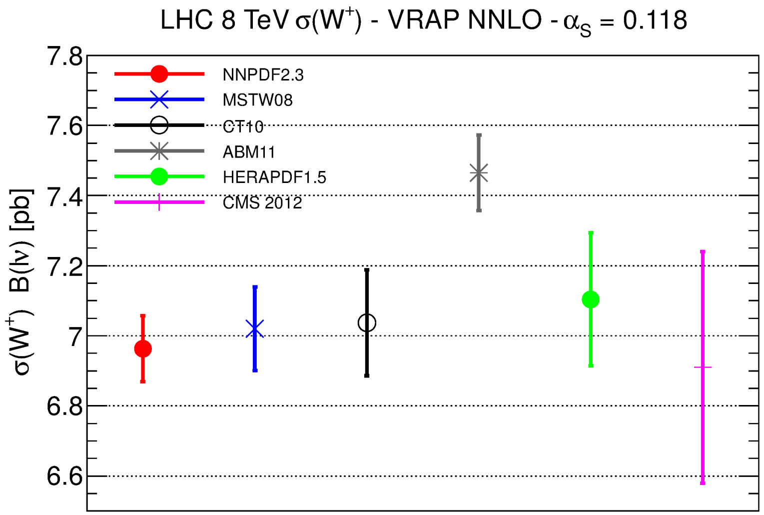
LHC KINEMATICS



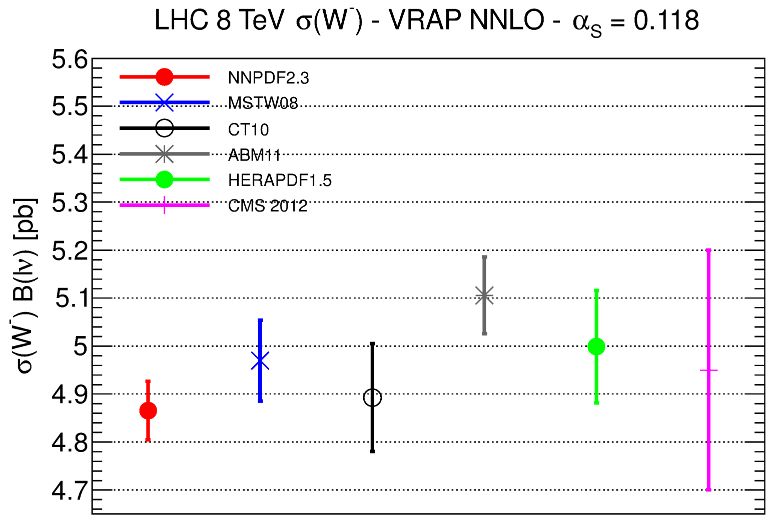
	MSTW08	CT10	NNPDF2.3	HERAPDF1.5	ABM11	JR09
HERA DIS	✔	✔	✔	✔	✔	✔
FIXED-TARGET DIS	✔	✔	✔	✗	✔	✔
FIXED-TARGET DY	✔	✔	✔	✗	✔	✔
TEVATRON W+Z+JETS	✔	✔	✔	✗	✗	✗
LHC W+Z+JETS	✗	✗	✔	✗	✗	✗

LHC EW standard candles at Higgs discovery (2012)

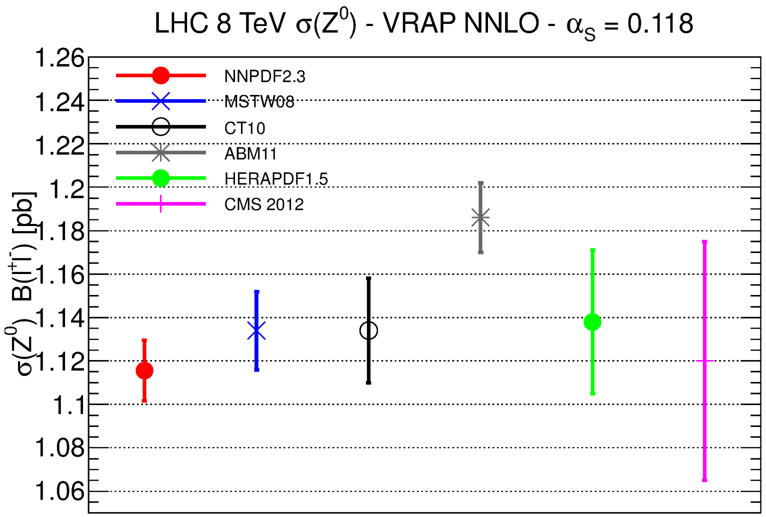
W^+



W^-



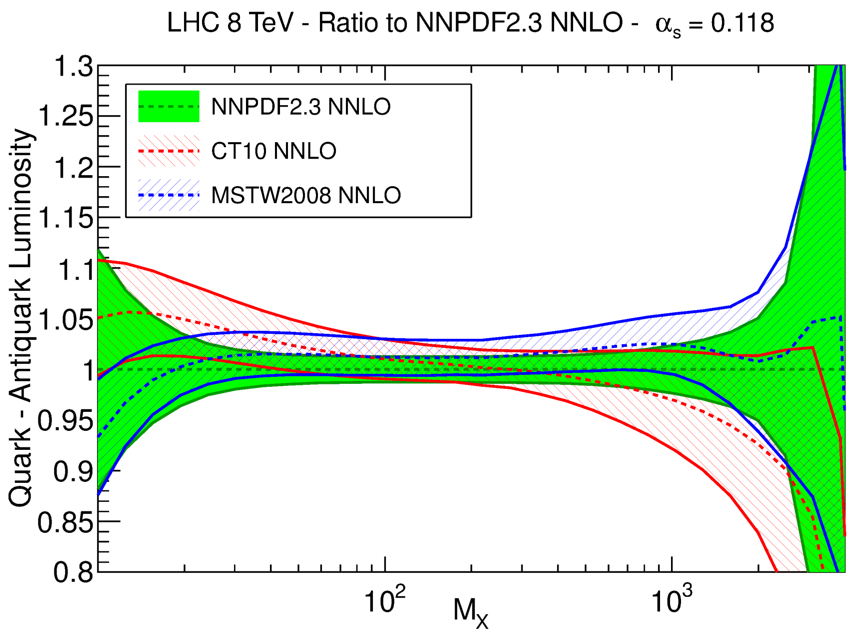
Z



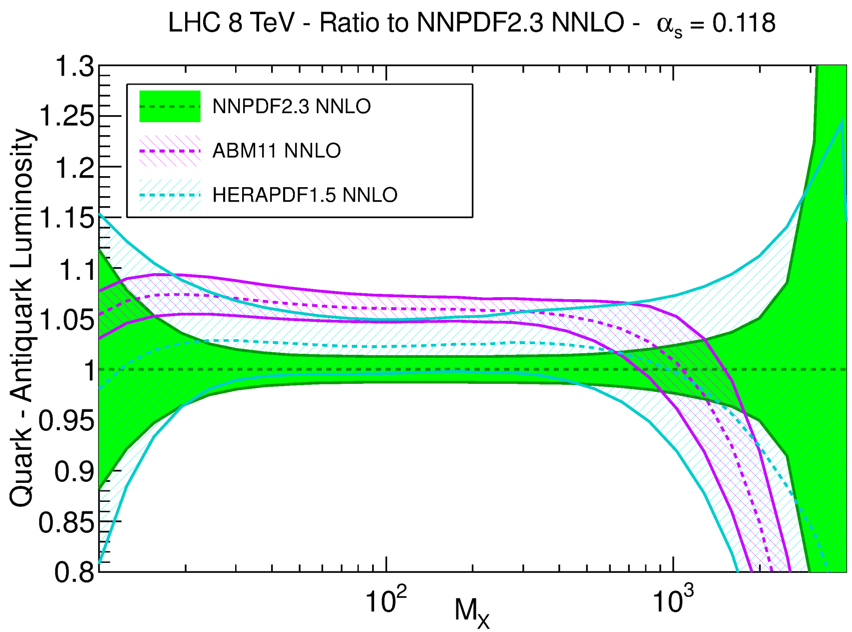
- Global fits in good mutual agreement
- DIS-only fit safe (HERAPDF) safe, but large uncertainty
- Weak dependence on α_s
- LHC data soon to provide competitive constraints

PARTON LUMINOSITIES: QUARK SECTOR $(q\bar{q})$

GLOBAL PDF SETS
(ratio to NNPDF2.3)



OTHER PDF SETS
(ratio to NNPDF2.3)

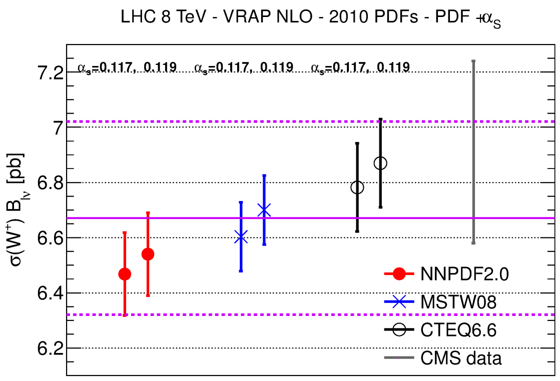


- CROSS-SECTIONS REFLECT UNDERLYING LUMINOSITIES
  FEWER DATA $\rightarrow$ LARGER UNCERTAINTIES (OR SYSTEMATIC BIAS)
- GLOBAL SETS: GOOD AGREEMENT IN THE REGION OF THE EW SCALE
- UNCERTAINTIES BLOW UP FOR LARGE-MASS FINAL STATES

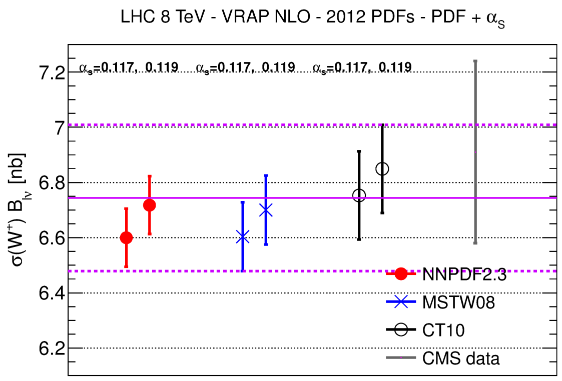
CONTINUOUS PROGRESS

GLOBAL PDF SETS: THE W^+ CROSS-SECTION

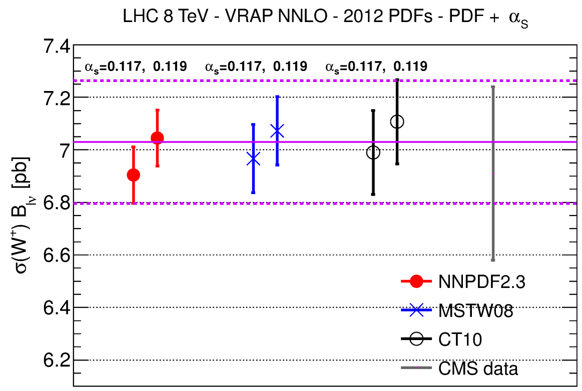
2010 NLO PDFs



2012 NLO PDFs



2012 NNLO PDFs



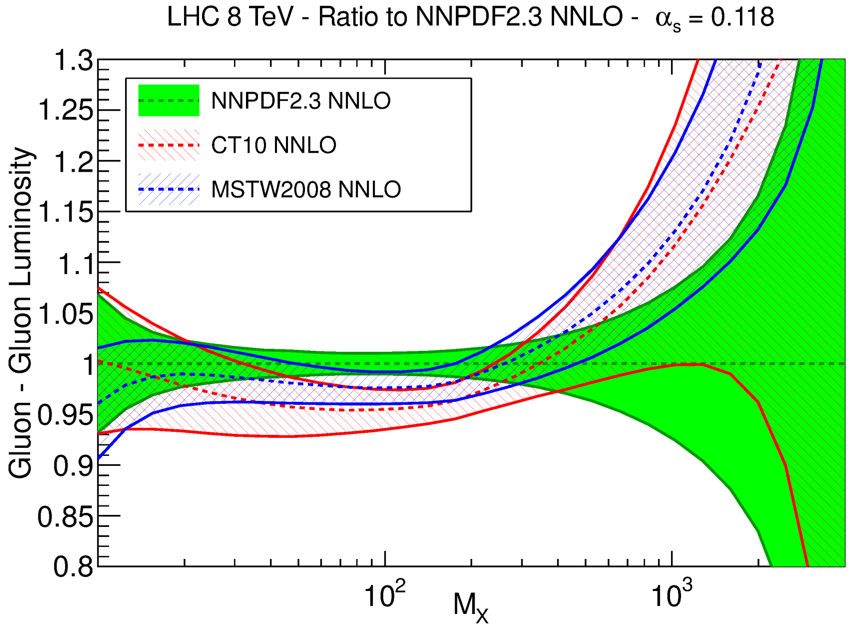
- Each datapoint includes PDF+α_s uncertainty; $\Delta\alpha_s = 0.001$
- $\alpha_s = 0.117$ and $\alpha_s = 0.119$ predictions given for each set (note all PDFs depend on α_s)
- horizontal (purple) line show envelope of predictions

IMPROVEMENTS

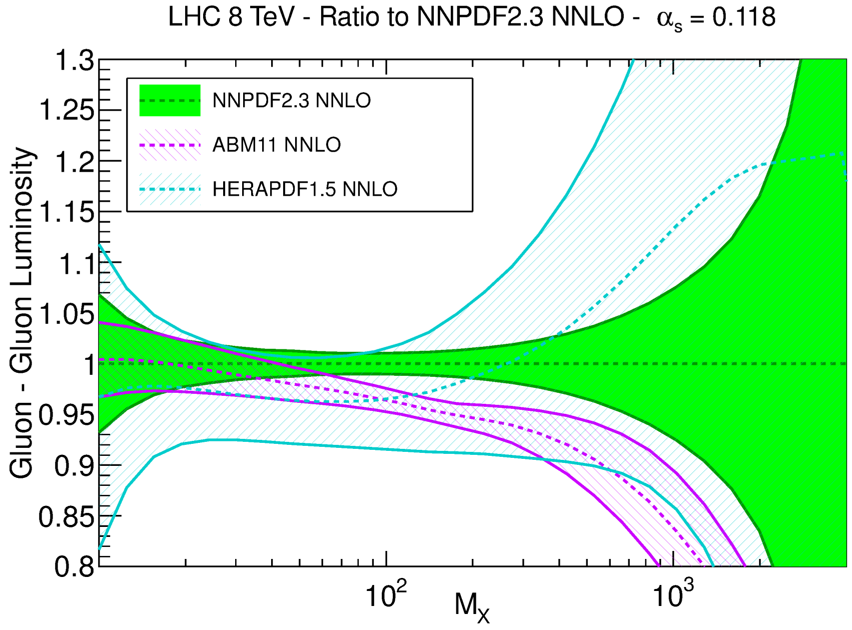
- MORE GENERAL PARAMETRIZATION (CTEQ, MSTW)
- NNLO FITS AVAILABLE (NNPDF, CTEQ)
- FULL TREATMENT OF CHARM MASS (NNPDF)
- CONTINUOUS BENCHMARKING

PARTON LUMINOSITIES: GLUON SECTOR

GLOBAL PDF SETS
(ratio to NNPDF2.3)



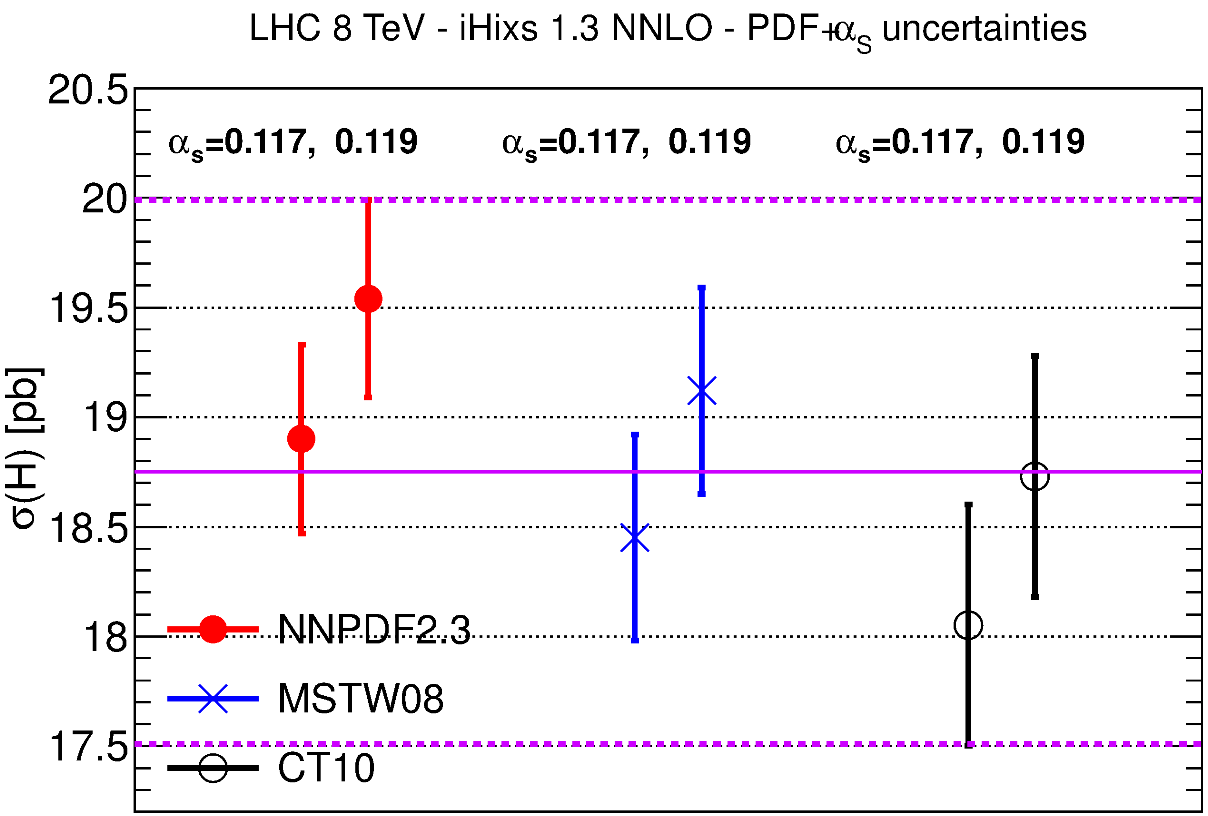
OTHER PDF SETS
(ratio to NNPDF2.3)



- FEWER DATA $\rightarrow$ LARGER UNCERTAINTIES (OR SYSTEMATIC BIAS)
- GLOBAL SETS: NOT SO GOOD AGREEMENT IN THE REGION OF THE EW SCALE
- UNCERTAINTIES BLOW UP FOR LARGE-MASS FINAL STATES

THE FIRST PDF4LHC PRESCRIPTION

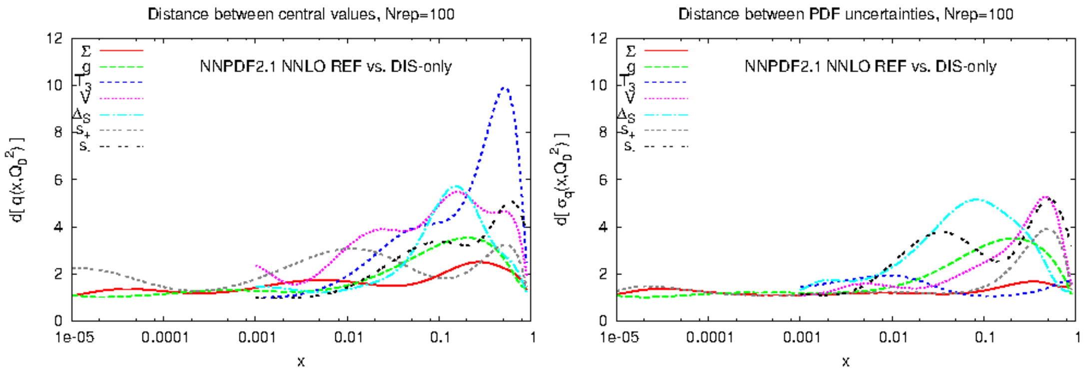
HIGGS IN GLUON FUSION



- HOW CAN ONE HANDLE DISCREPANCIES WHICH ARE NOT UNDERSTOOD?
- CONSERVATIVE ANSWER: TAKE THE ENVELOPE OF RESULTS

THE IMPACT OF LHC DATA: NNPDF2.X
PDFS MOSTLY DETERMINED BY DIS DATA
NNPDF2.1 VS NNPDF2.1 DIS ONLY

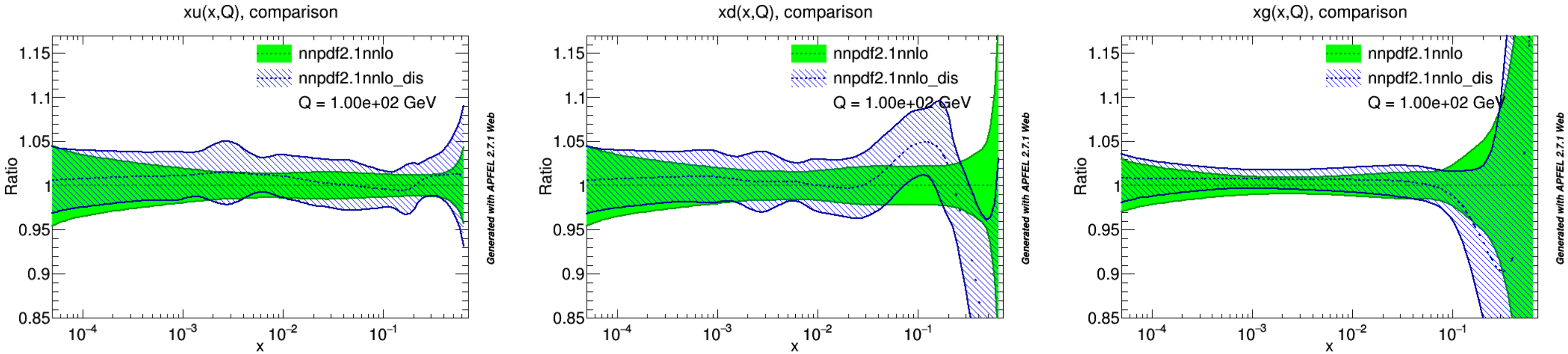
DISTANCES (difference in units of st. dev.)



$d = 10 \Leftrightarrow$ one sigma difference

PDF COMPARISON

UP DOWN GLUON

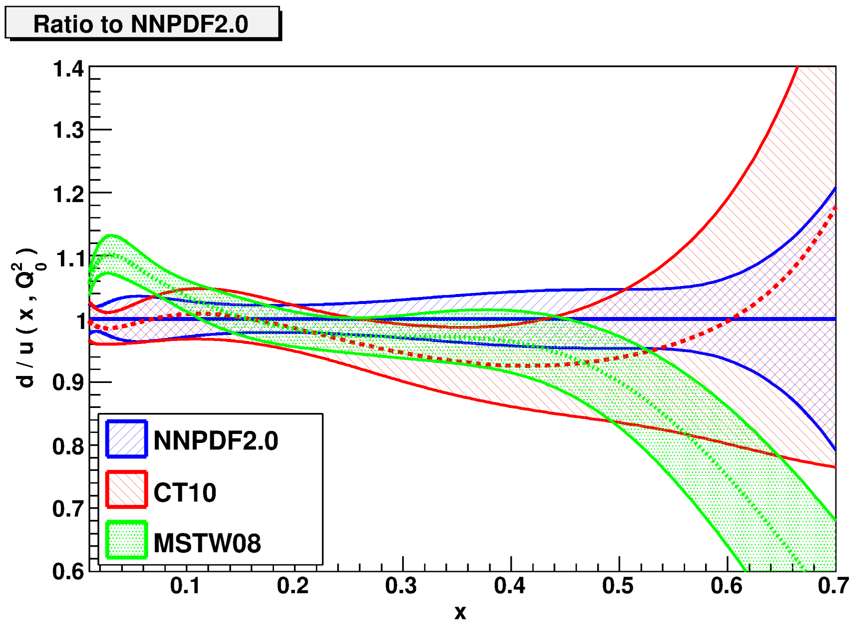


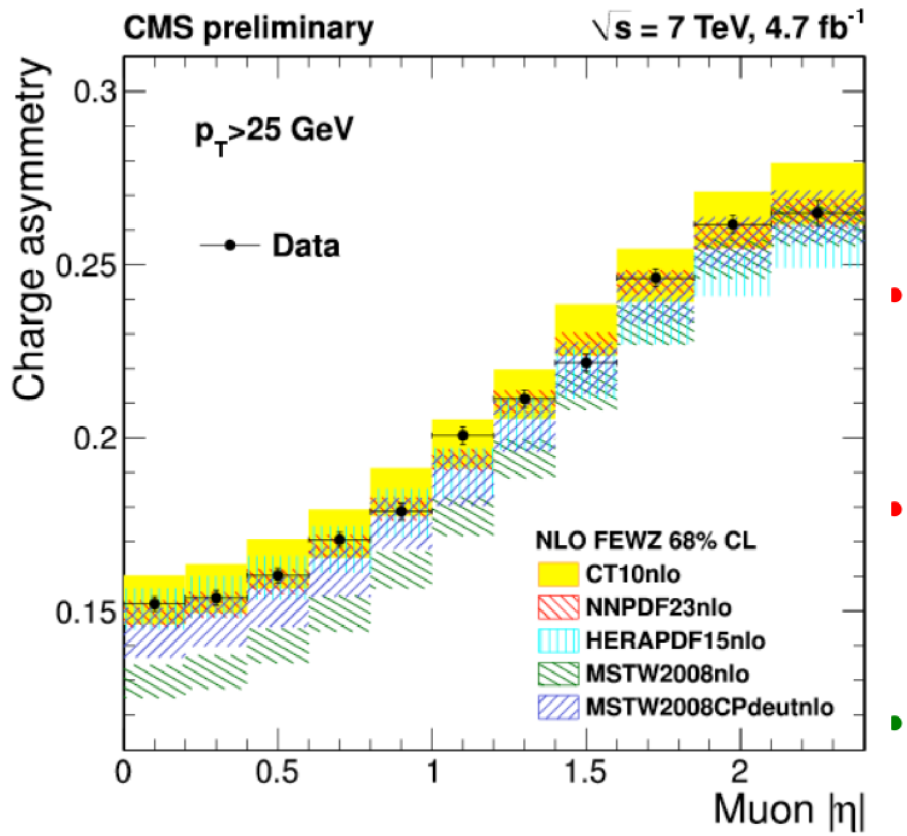
- ALL DIFFERENCES BELOW ONE SIGMA
- ONLY UP-DOWN SEPARATION SIGNIFICANTLY AFFECTED

THE IMPACT OF LHC DATA
RESOLVING DISCREPANCIES
AN EXAMPLE: THE d/u RATIO

THE CMS W ASYMMETRY

THE d/u RATIO





- LONG-STANDING DISCREPANCY IN THE d/u RATIO BETWEEN MSTW AND OTHER GLOBAL FITS
- RESLOVED BY CMS W ASYMMETRY DATA
- EXPLAINED BY INSUFFICIENTLY FLEXIBLE PDF PARAMETRIZATION $\rightarrow$ NEW MSTW08DEUT SET

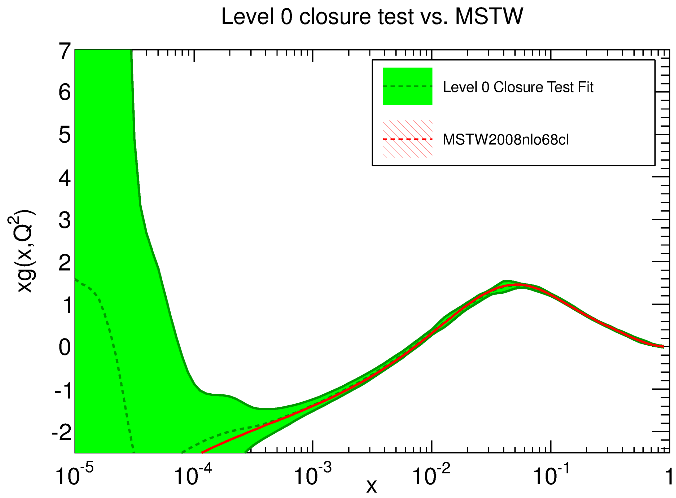
NNPDF3.0: CLOSURE TESTS (2014)

LEVEL 0

- ASSUME VANISHING EXPERIMENTAL UNCERTAINTY
- MUST BE ABLE TO GET $\chi^2 = 0$
- UNCERTAINTY AT DATA POINTS TENDS TO ZERO (NOT NECESSARILY ON PDF!)

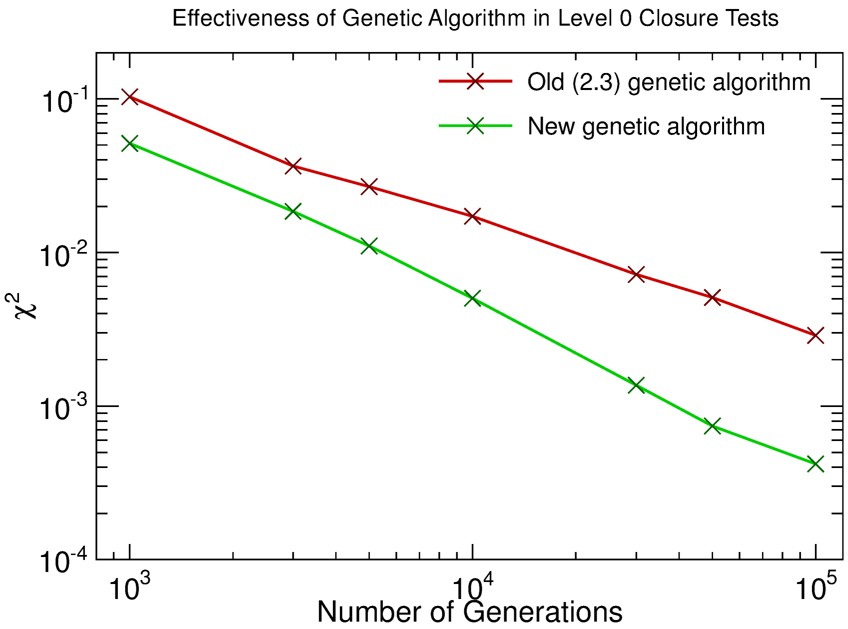
  DEFINE $\phi \equiv \sqrt{\langle \chi^2_{rep} \rangle - \chi^2}$, EQUALS FIT UNCERTAINTY/DATA UNCERTAINTY; CHECK $\phi \to 0$
- BEST FIT ON TOP OF "TRUTH" IN DATA REGION

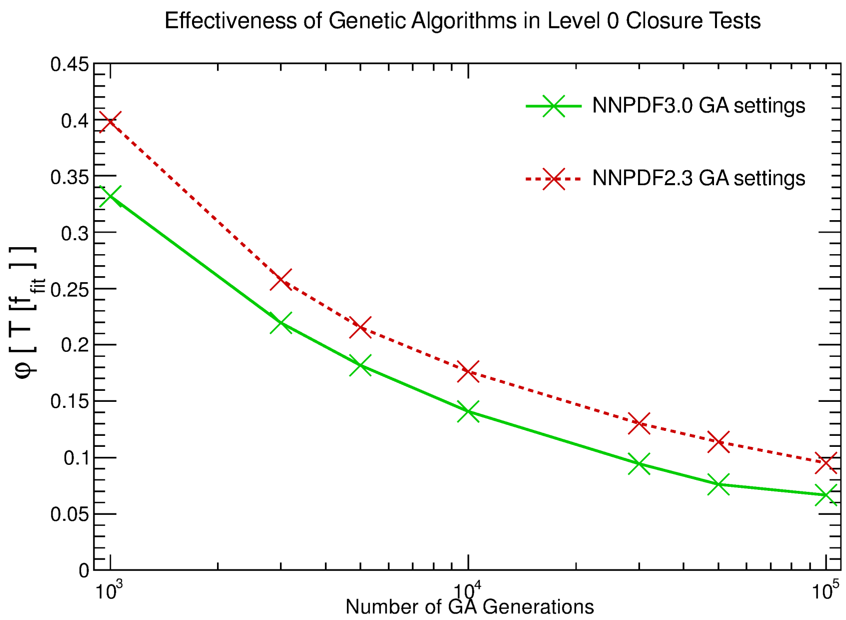
THE GLUON



FRACTIONAL UNCERTAINTY VS TRAINING LENGTH

χ^2 VS TRAINING LENGTH

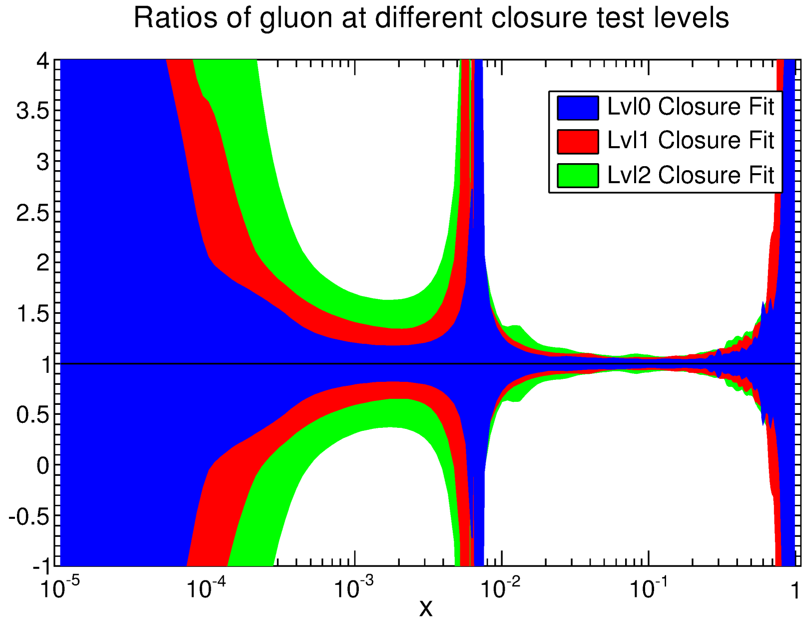


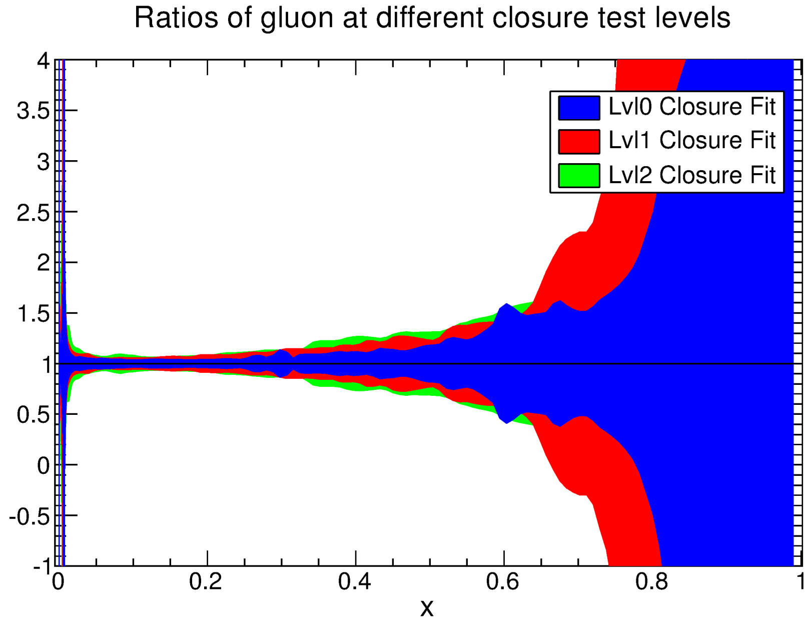


LEVEL-0, LEVEL-1 AND LEVEL-2

- LEVEL 0: FAKE DATA GENERATED WITH NO UNCERTAINTY
  $\to$ INTERPOLATION AND EXTRAPOLATION UNCERTAINTY
- LEVEL 1-2: FAKE DATA GENERATED WITH SAME UNCERTAINTY AS REAL DATA (INCLUDING CORRELATIONS)
- LEVEL 1: NO PSEUDODATA REPLICAS:
  $\Rightarrow$ REPLICAS FITTED TO SAME DATA OVER AND OVER AGAIN
  $\to$ FUNCTIONAL UNCERTAINTY DUE TO INFINITY OF EQUIVALENT MINIMA
- LEVEL 2: STANDARD NNPDF METHODOLOGY
  $\Rightarrow$ REPLICAS FITTED TO PSEUDODATA REPLICAS
  $\to$ DATA UNCERTAINTY
- THREE SOURCES OF UNCERTAINTY COMPARABLE IN DATA REGION

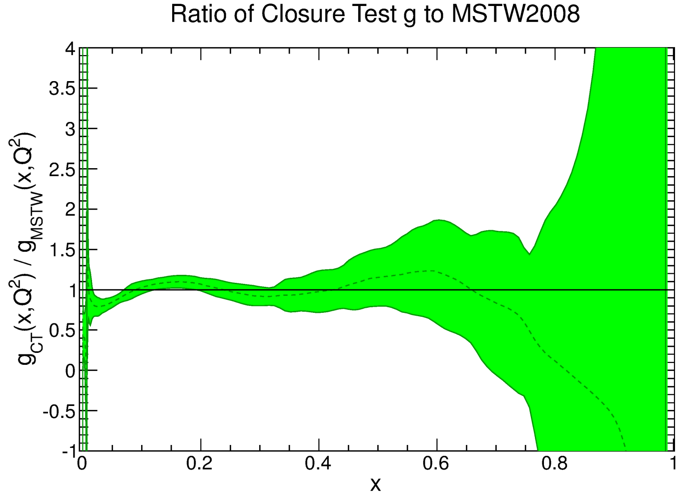
THE GLUON: LEVEL 0, LEVEL 1 AND LEVEL 2





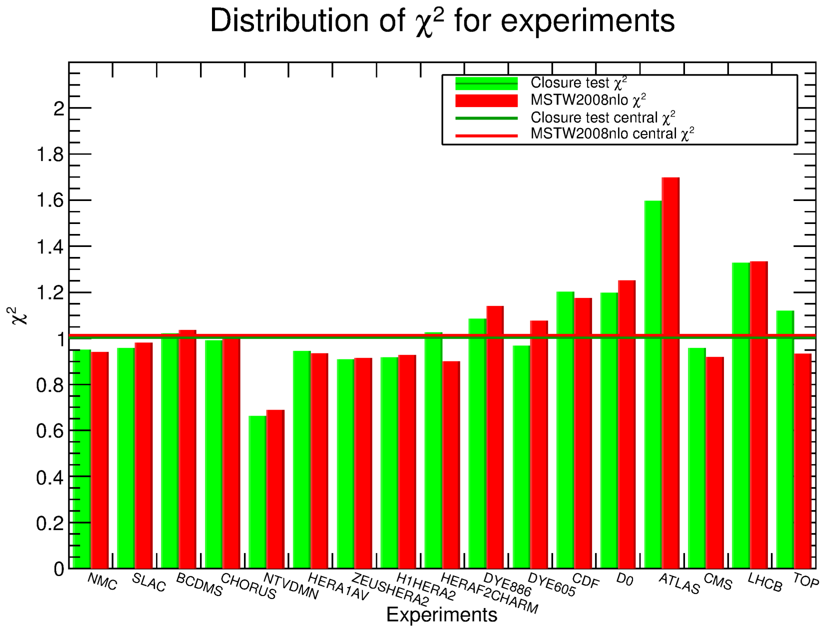
LEVEL-2: CENTRAL VALUES AND UNCERTAINTIES

THE GLUON: FITTED/"TRUE"

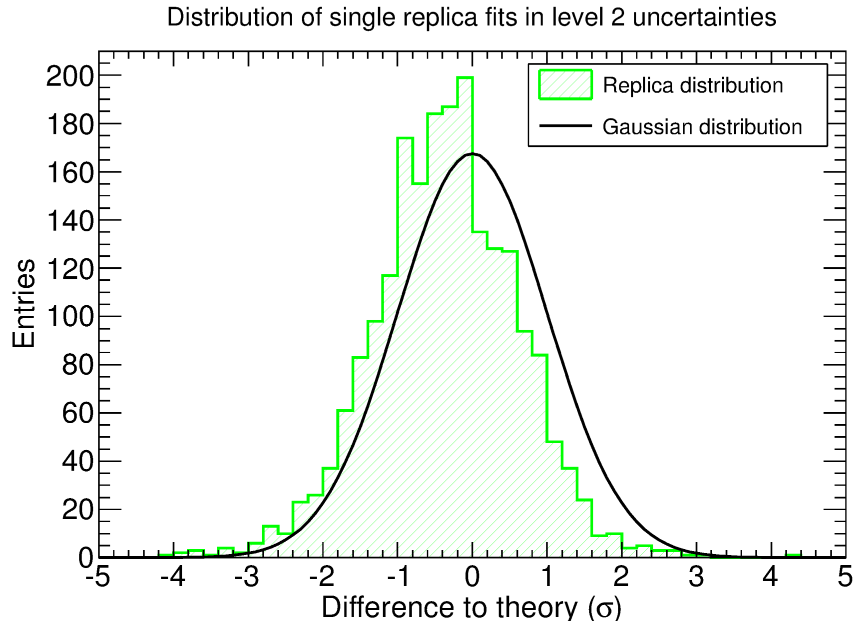


- CENTRAL VALUES:
  COMPARE FITTED VS. "TRUE" χ^2
  BOTH FOR INDIVIDUAL EXPERIMENTS
  & TOTAL DATASET
  FOR TOTAL $\Delta\chi^2 = 0.001 \pm 0.003$
- UNCERTAINTIES: DISTRIBUTION OF DEVIATIONS BETWEEN FITTED AND "TRUE" PDFs
  SAMPLED AT 20 POINTS BETWEEN 10^{-5} AND 1
  FIND 0.699% FOR ONE-SIGMA,
  0.948% FOR TWO-SIGMA C.L.

LEVEL-2 FITTED χ^2 VS "TRUE"



NORM. DISTRIBUTION OF DEVIATIONS



LEVEL-2 STABILITY TESTS

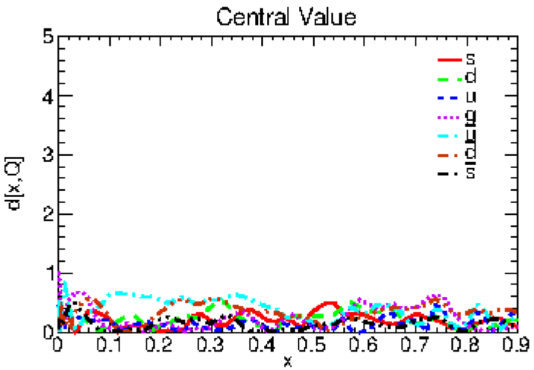
- CHANGE UNDERLYING PDF SET (CT10, NNPDF2.3)
- INCREASE MAXIMUM GA TRAINING LENGTH TO 80K
  TESTS EFFICIENCY OF CROSS-VALIDATION
- INCREASE NN ARCHITECTURE TO 2-20-15-1
  NUMBER OF FREE PARAMETRES INCREASE BY MORE THAT $10\times$
- CHANGE PDF PARAMETRIZATION BASIS
  OLD: ISOTRIPLET, $\bar{u}-\bar{d}$, $s+\bar{s}$, $s-\bar{s}$;
  NEW: ISOTRIPLET, SU(3)-OCTET, BOTH TOTAL $(q+\bar{q})$ & VALENCE $(q-\bar{q})$

STATISTICAL EQUIVALENCE!

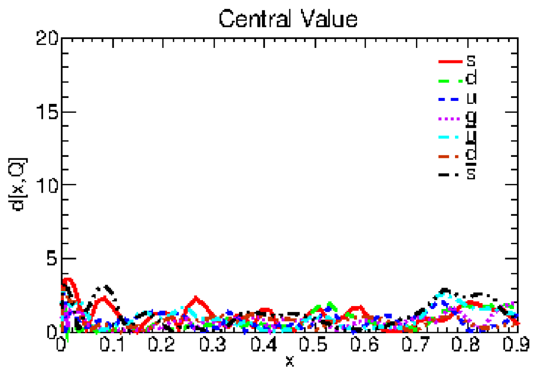
DISTANCES BETWEEN REF. AND NEW FIT:

difference in unites of standard deviation of the mean

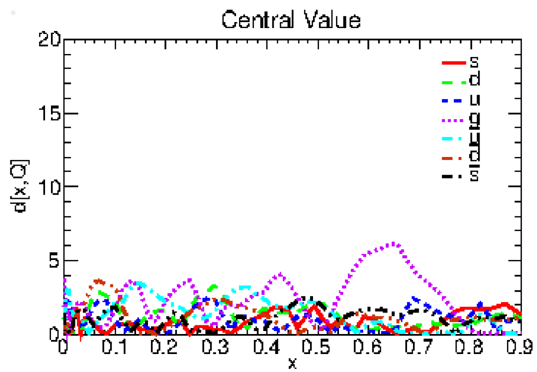
30K GENS VS 80K GENS



2.3 BASIS VS 3.0 BASIS



300 VS 37 PARMS

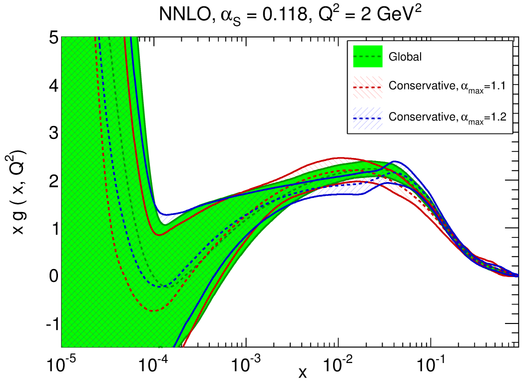


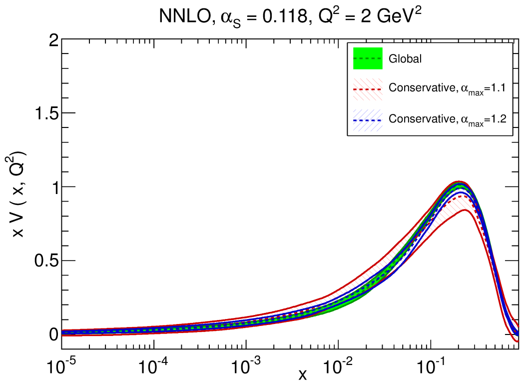
NNPDF3.0: DATA CONSISTENCY

- RESCALE ALL UNCERTAINTIES $\sigma \to \alpha\sigma$: $\chi^2 \to \chi^2/\alpha^2$ FOR A GIVEN EXPERIMENT
- DETERMINE PROBABILITY DISTRIBUTION $P(\alpha)$ (USING BAYES)
- DISCARD ALL EXPERIMENT FOR WHICH $P(\alpha)$ PEAKS WELL ABOVE ONE
  TWO OUT OF MEDIAN, MODE, MEAN, GREATER THAN $\alpha_{threshold}$
- $\chi^2 = 1.29$ FOR NNLO GOBAL, BECOMES $\chi^2 = 1.16$ FOR $\alpha_{threshold} = 1.3$, $\chi^2 = 1.10$ FOR $\alpha_{threshold} = 1.2$, $\chi^2 = 1.01$ FOR $\alpha_{threshold} = 1.1$, BUT CONSIDERABLE DETERIORATION OF UNCERTAINTIES

CONSERV. VS. DEFAULT

GLUON AND VALENCE





α PEAK FOR EXPERIMENTS DISCARDED IN CONS. FIT

WHEN INCLUDED OR EXCLUDED FROM FIT

	NNLO global fit			NNLO cons. fit $\alpha_{max} = 1.1$		
Experiment	mean	mode	median	mean	mode	median
NMC $\sigma^{NC,p}$	1.27	1.26	1.27	1.50	1.45	1.48
SLAC	1.13	1.09	1.12	1.61	1.37	1.48
BCDMS	1.20	1.19	1.20	2.02	1.86	1.92
CHORUS	1.10	1.09	1.09	2.55	1.69	2.32
ZEUS HERA-II	1.25	1.24	1.25	1.38	1.33	1.36
H1 HERA-II	1.35	1.34	1.34	1.51	1.47	1.49
HERA σ^c_{NC}	1.14	1.11	1.13	1.13	1.09	1.12
E886 p	1.15	1.14	1.15	2.18	1.62	2.03
CDF Z rapidity	1.39	1.32	1.36	1.56	1.40	1.50
CDF Run-II k_t jets	1.15	1.12	1.14	1.25	1.18	1.22
ATLAS W, Z 2010	1.17	1.12	1.15	1.38	1.25	1.32
ATLAS high-mass DY	1.00	1.34	1.63	1.63	1.19	1.45
CMS W muon asy	1.60	1.40	1.53	2.90	2.48	2.81
CMS W+c total	1.50	1.09	1.33	1.85	1.37	1.67
CMS W+c ratio	2.00	1.39	1.69	2.12	1.58	1.94
CMS 2D DY 2011	1.28	1.27	1.28	1.29	1.28	1.29
LHCb	1.20	1.12	1.17	1.58	1.22	1.48

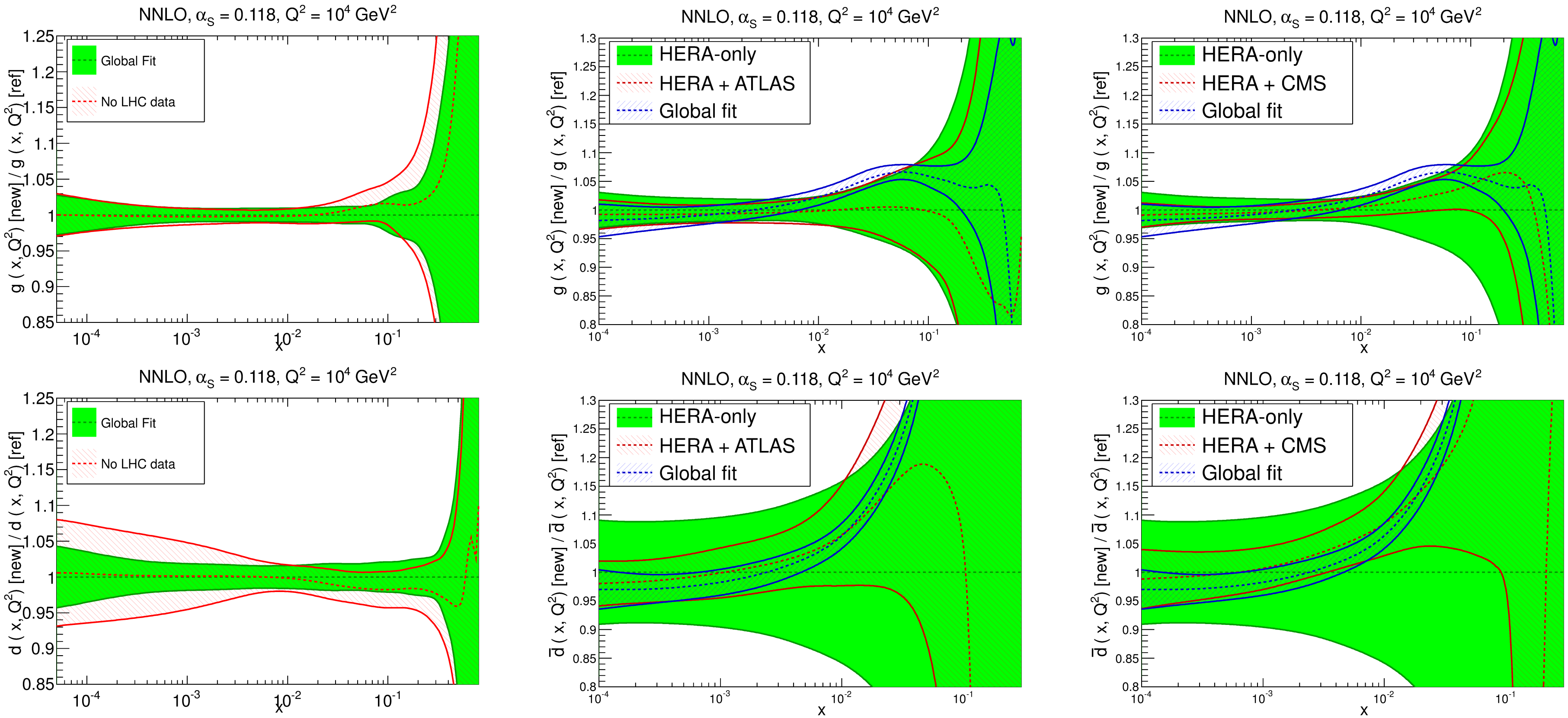
NNPDF3.0: THE IMPACT OF LHC (AND HERA) DATA

- OVERALL MEASURE OF IMPACT:
  $\phi \Rightarrow$ FIT UNCERTAINTY/DATA UNCERTAINTY
- HERA-II IMPACT SIZABLE
- IMPACT OF LHC DATA MODERATE BUT VISIBLE
- IMPACT OF CMS OR ATLAS COMPARABLE TO (MODERATE) IMPACT OF NON-LHC, NON-HERA DATA

FRACTIONAL UNCERTAINTY

Dataset	φ_{χ^2} NLO	φ_{χ^2} NNLO
Global	0.291	0.302
HERA-I	0.453	0.439
HERA all	0.375	0.343
HERA+ATLAS	0.391	0.318
HERA+CMS	0.315	0.345
Conservative	0.422	0.478
no LHC	0.312	0.316

GLOBAL VS NO LHC: g& d GLOBAL VS HERA+ATLAS: g& $\bar{d}$ GLOBAL VS HERA+CMS: g& $\bar{d}$



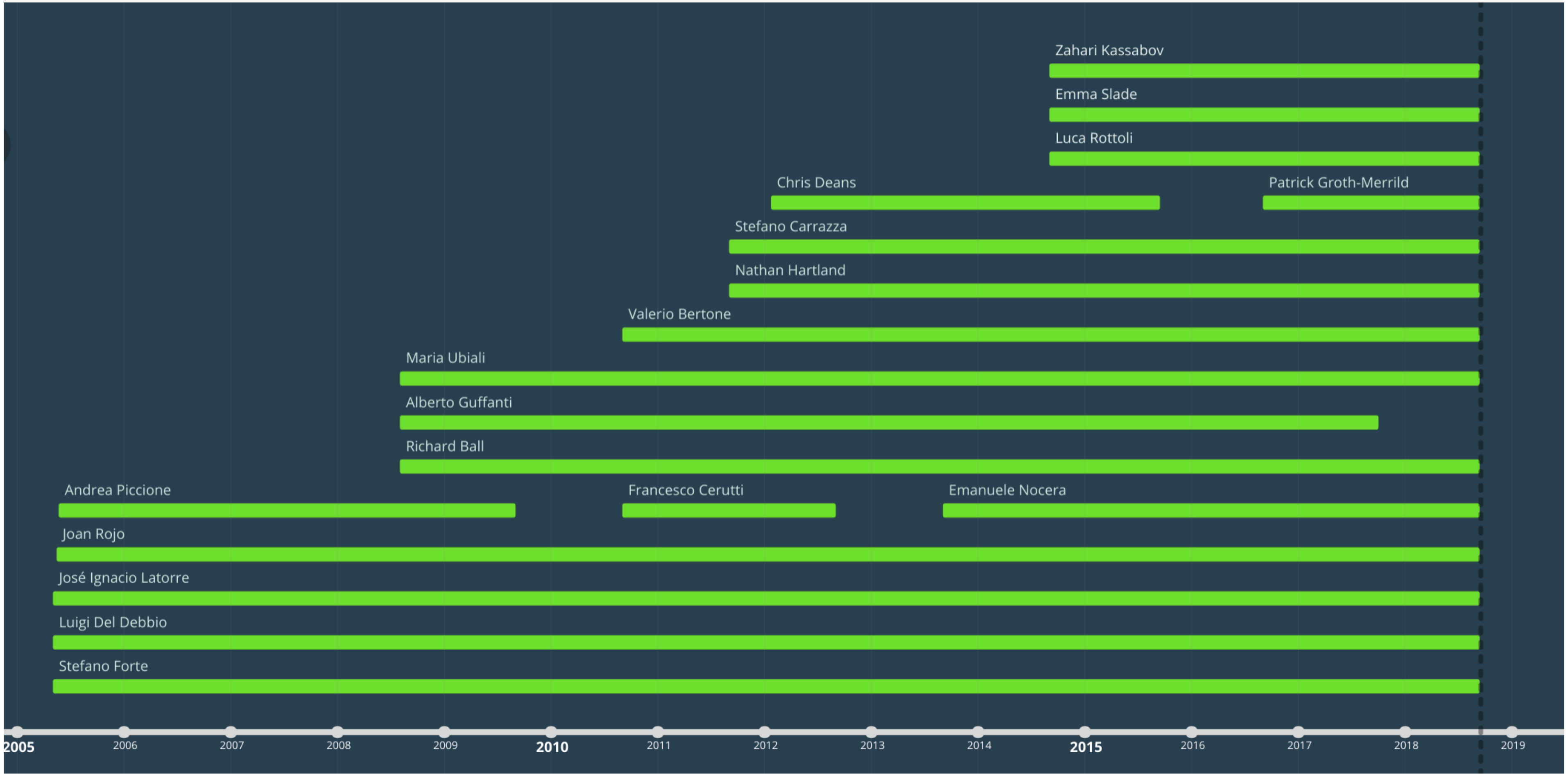
THE LAST FEW YEARS (2015-2018)

- THE CHARM PDF
- MC-H, COMPRESSOR & SM-PDFs
- PDF4LHC15
- RESUMMED PDFs (THRESHOLD & SMALL x)
- NNPDF 3.1
  - INDEPENDENTLY PARAMETRIZED CHARM
  - LHC DATA
- LUX-QED PDFs
- NNFF1.0, NNFF1.1
- α_s

TECHNICAL “DETAILS”: CHOICES/PROGRESS

- CODE STORAGE, TASK MANAGEMENT, REPOSITORIES
  - SVN $\Rightarrow$ GIT
  - VALIDPHYS
- PARAMETRIZATION & MINIMIZATION
  - PREPROCESSING: TUNED VS RANDOM RANGE
  - t_0 FOR MULTIPLICATIVE UNCERTAINTIES
  - POSITIVITY
  - GA: FIXED VS. NODAL MUTATION
  - TARGETED WEIGHTED TRAINING
- STOPPING
  - THRESHOLD VS LOOKBACK
- COMPUTATION OF PHYSICAL PROCESSES
  - APFEL $\rightarrow$ APFELCOMB
  - `FastKernel` $\Rightarrow$ FKTABLES

PEOPLE



THANK YOU!

& WELCOME TO:

Tommaso Giani, Rabah Abdul Kaleh, Rosalyn Pearson, Cameron Voisey, Michael Wilson

"Io stimo più il trovare un vero, benché di cosa leggiera, che il disputar lungamente delle massime questioni senza verità nissuna"

Galileo Galilei, letter to Tommaso Campanella