NNPDF

WHERE DO WE COME FROM? WHAT ARE WE? WHERE ARE WE GOING?

STEFANO FORTE

MILAN UNIVERSITY & INFN

FOR THE COLLABORATION: R. D. BALL, L. DEL DEBBIO,

S.F., A. GUFFANTI, J. I. LATORRE, J. ROJO, M. UBIALI

PDF4LHC & TH INSITITUTE

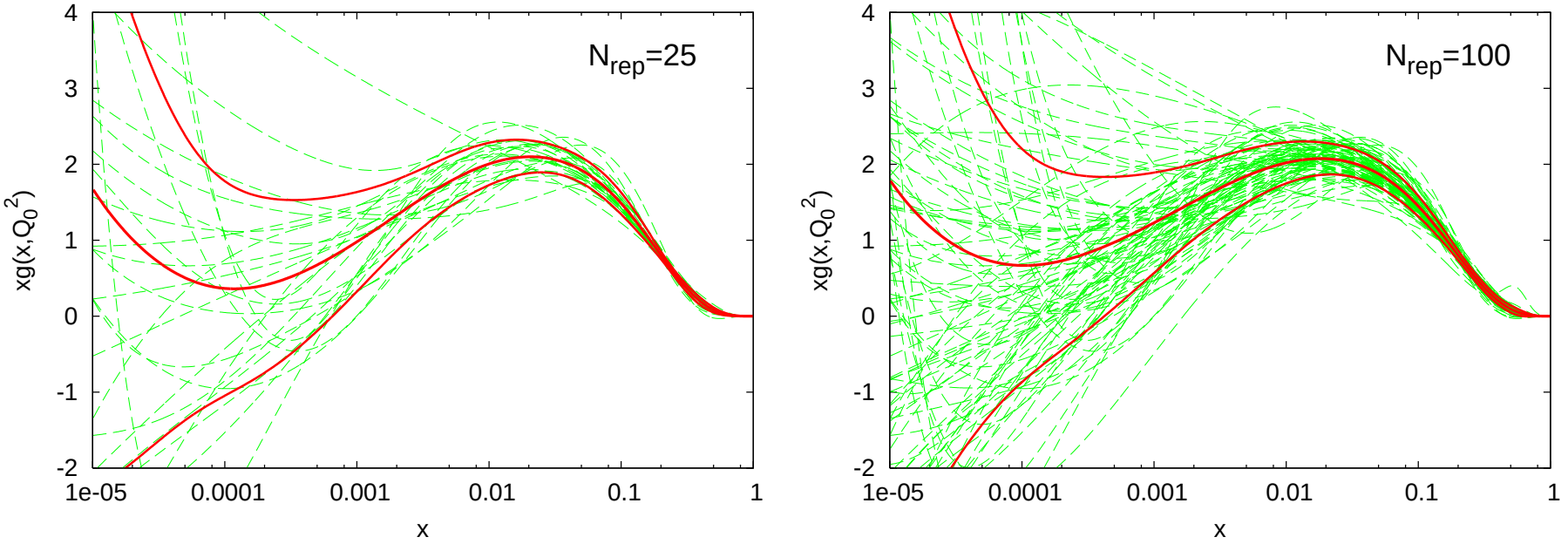
CERN, AUGUST 6, 2009

WHAT ARE WE?

WHAT DO WE DELIVER?

A MONTE CARLO SET OF REPLICAS OF PDFs

Example: the gluon distribution in the NNPDF1.0 set



- ENSEMBLE OF REPLICAS $\leftrightarrow$ PROBABILITY DISTRIBUTION OF PDFs
- EXPECTED CENTRAL VALUE $\leftrightarrow$ MEAN; UNCERTAINTY $\leftrightarrow$ STANDARD DEVIATION
- ANY FEATURES OF DISTRIBUTION CAN BE DETERMINED
  (C.L. INTERVALS, CORRELATIONS...)

WHERE DO WE COME FROM?

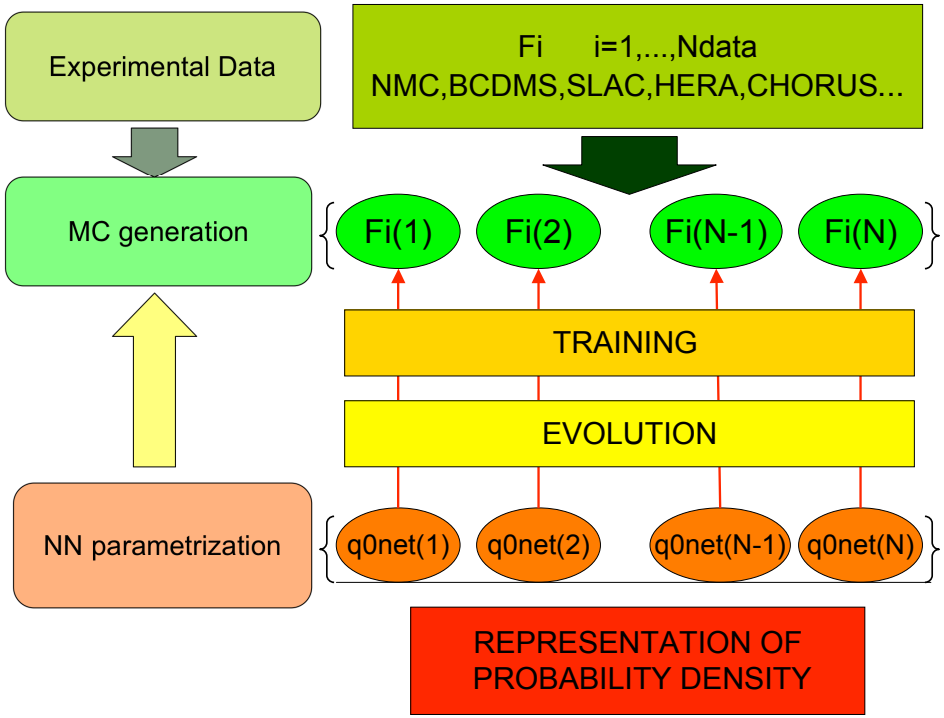
WHAT DO WE START FROM?

A MONTE CARLO SET OF REPLICAS OF DATA

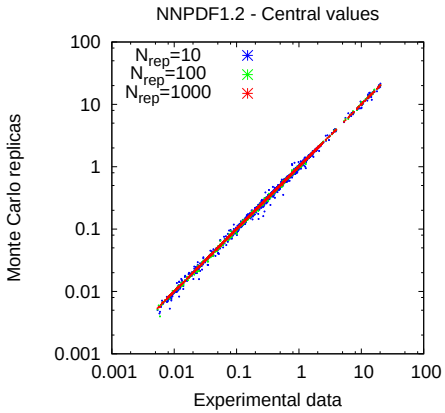
- SPACE OF FUNCTIONS HUGE

  5 BINS FOR 10 PTS$\times$ 5 FCTNS $\rightarrow 5^{50} \sim 10^{34}$ BINS

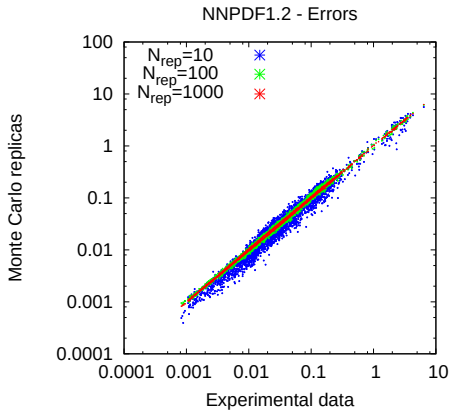
- THE DATA TELL US WHICH ONES ARE POPULATED (IMPORTANCE SAMPLING)



replica averages vs. central values



replica standard dev. vs. uncertainties



10 REPLICAS ENOUGH FOR CENTRAL VALS, 100 FOR UNCERTAINTIES, 1000 FOR CORRELNS

WHERE DO WE COME FROM?

HOW DO WE GET PDF REPLICAS FROM DATA REPLICAS?

NEURAL NETWORK PARM+ CROSS-VALIDATION METHOD

- EACH PDF$\leftrightarrow$ NEURAL NETWORK PARAMETRIZED BY 37 PARAMETERS
- NNPDF1.2: $37 \otimes 7 = 259$ PARMS
  (COMP. MSTW08 $\rightarrow$ 28 FREE PARAMETERS)
  "INFINITE" NUMBER OF PARAMETERS$\Rightarrow$ CAN REPRESENT ANY FUNCTION
- COMPLEX SHAPES (LARGE NO.OF PARAMETERS) REQUIRE LONGER FITTING
- FIT STOPS WHEN QUALITY OF FIT TO RANDOMLY SELECTED "VALIDATION" DATA (NOT FITTED) STOPS IMPROVING
- CAN OBTAIN A FIT WITH χ^2 LOWER THAN BEST FIT ("OVERLEARNING")

WHERE DO WE COME FROM?

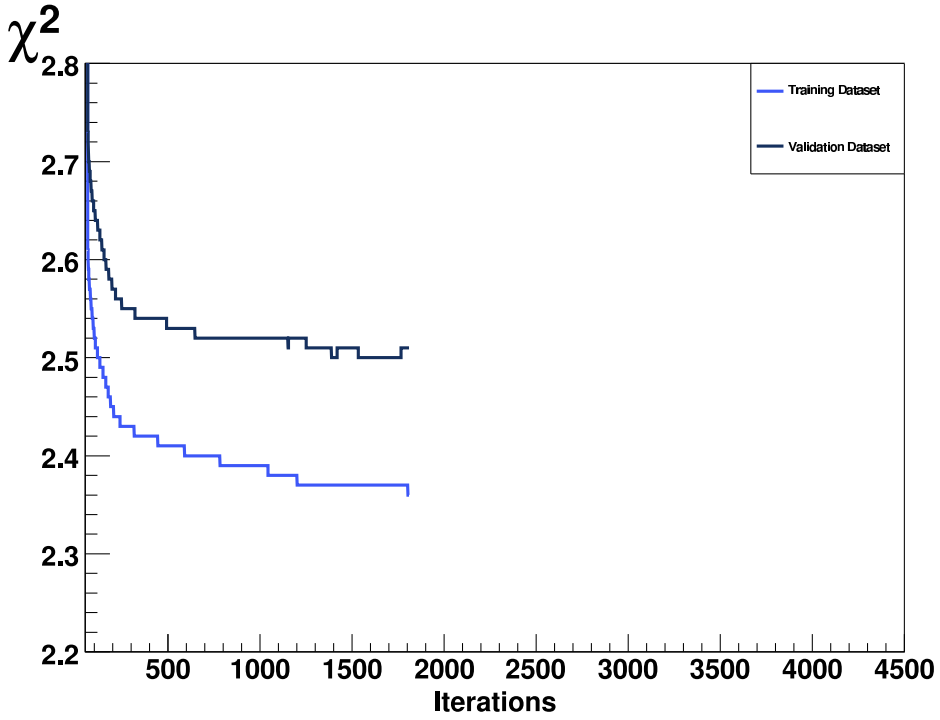
HOW DO WE GET PDF REPLICAS FROM DATA REPLICAS?

CROSS-VALIDATION

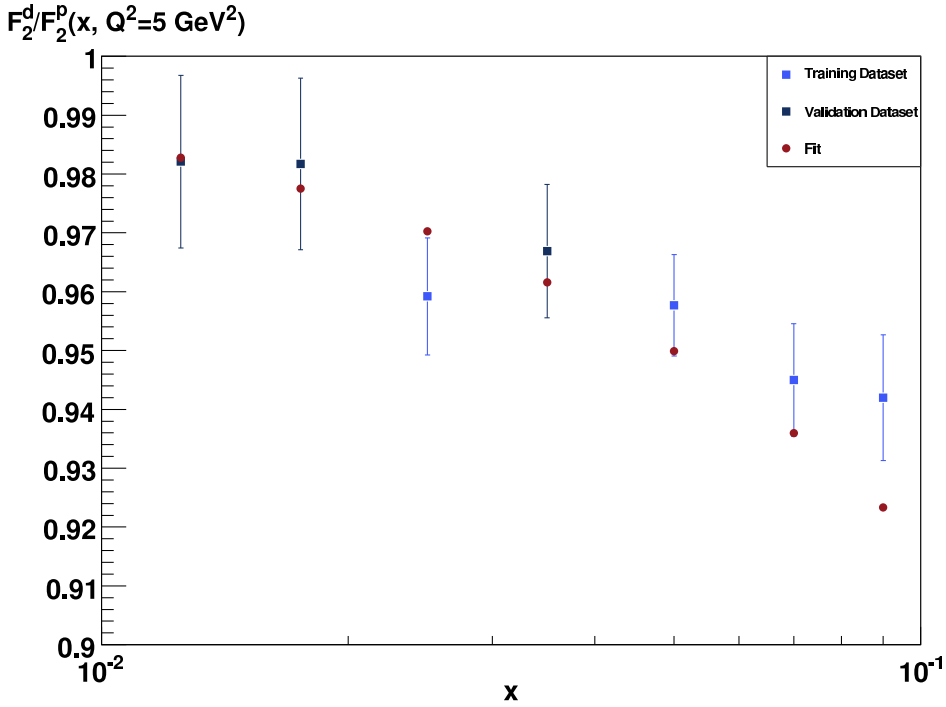
- OPTIMAL FIT OBTAINED WHEN QUALITY OF FIT TO VALIDATION (CONTROL) DATA STOPS IMPROVING
- POSSIBILITY OF OVERFITTING GUARANTESS THAT MINIMUM NOT DRIVEN BY PARAMETRIZATION

OPTIMAL FITTING

χ^2



FIT TO DATA



WHERE DO WE COME FROM?

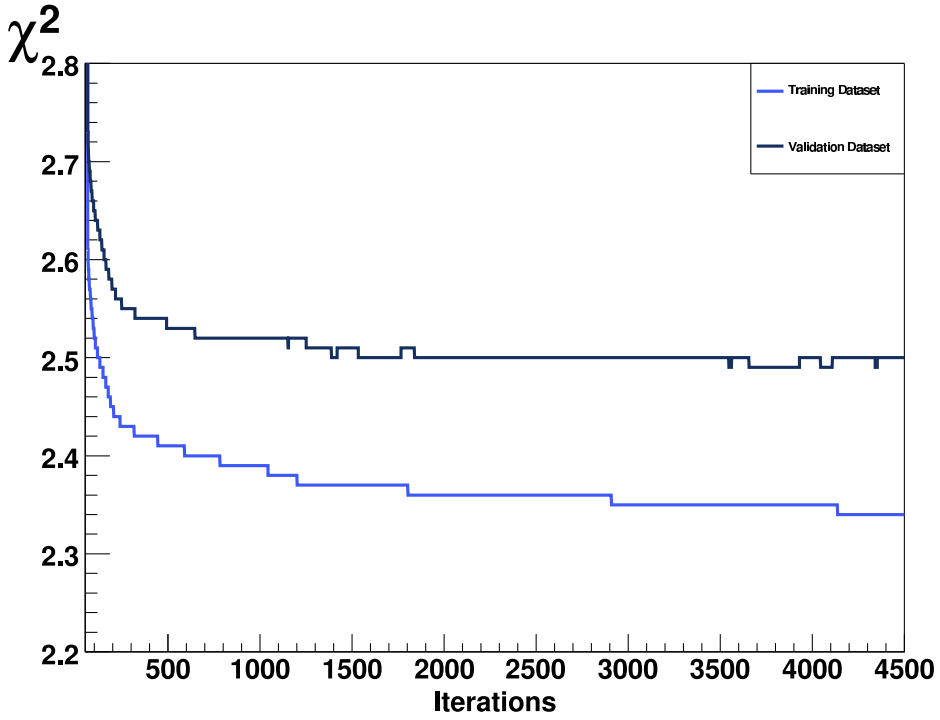
HOW DO WE GET PDF REPLICAS FROM DATA REPLICAS?

CROSS-VALIDATION

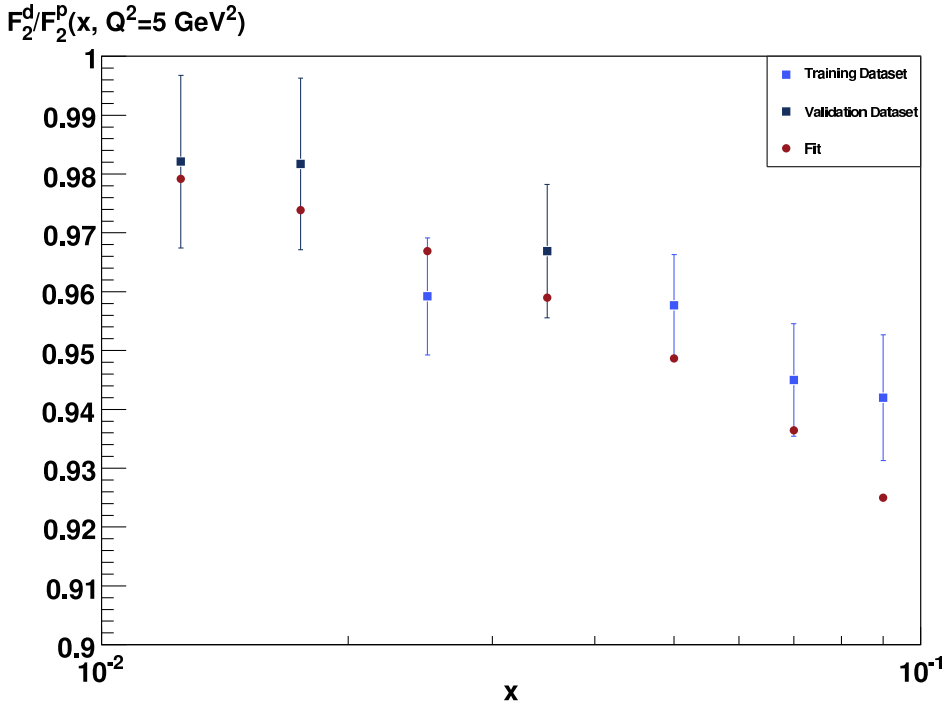
- OPTIMAL FIT OBTAINED WHEN QUALITY OF FIT TO VALIDATION (CONTROL) DATA STOPS IMPROVING
- POSSIBILITY OF OVERFITTING GUARANTESS THAT MINIMUM NOT DRIVEN BY PARAMETRIZATION

OVERFITTING

χ^2



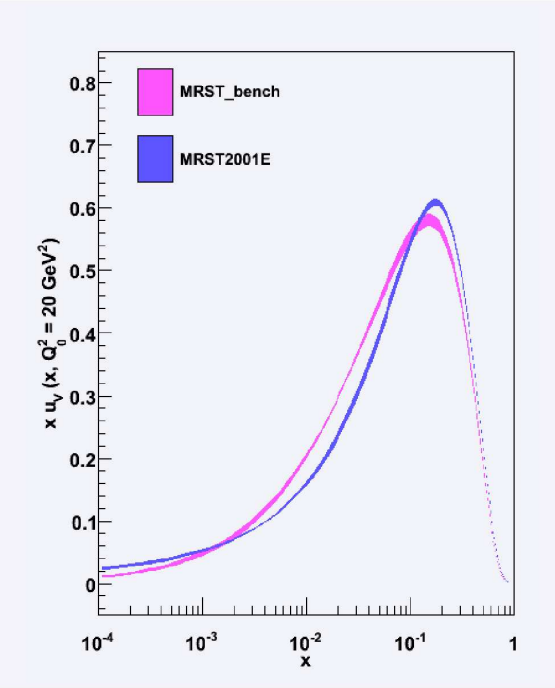
FIT TO DATA

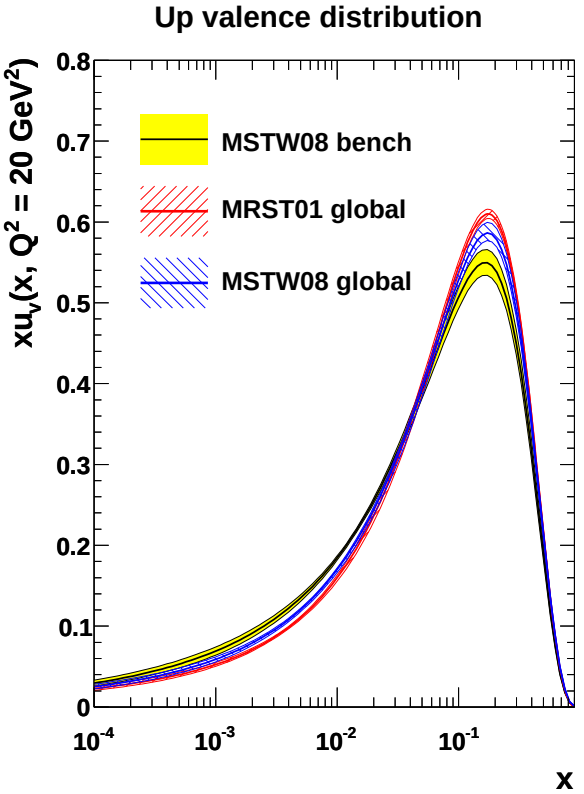


WHERE DO WE COME FROM?
WHY ARE WE DOING ALL THIS?
ISSUES IN THE STANDARD APPROACH

BENCHMARK FITS (HERALHC 2005-2008)

- PERFORM A FIT TO A CONSISTENT SUBSET OF DATA, USE $\Delta\chi^2 = 1$
- RESULTS NOT CONSISTENT, BUT IMPROVE W/ MORE GENERAL PARM.(MSTW08 VS MRST01)



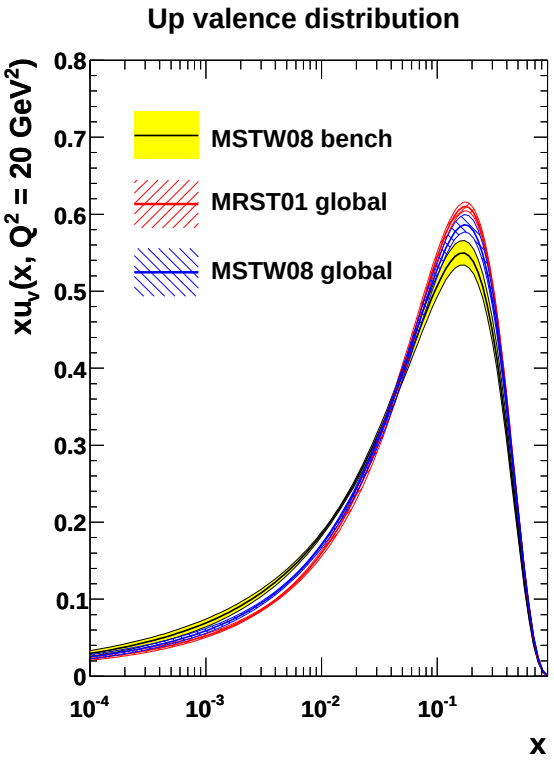


- UNCERTAINTY DOES NOT GO DOWN AS DATASET INCREASES
- MUST TUNE PARAMETRIZATION AND STATISTICAL TREATMENT (TOLERANCE) TO DATASET

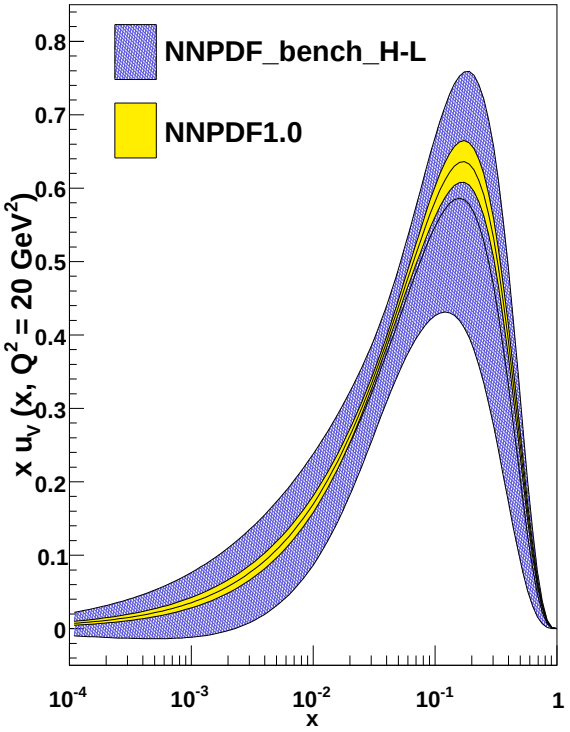
WHO ARE WE?

IS THIS ISSUE SOLVED?

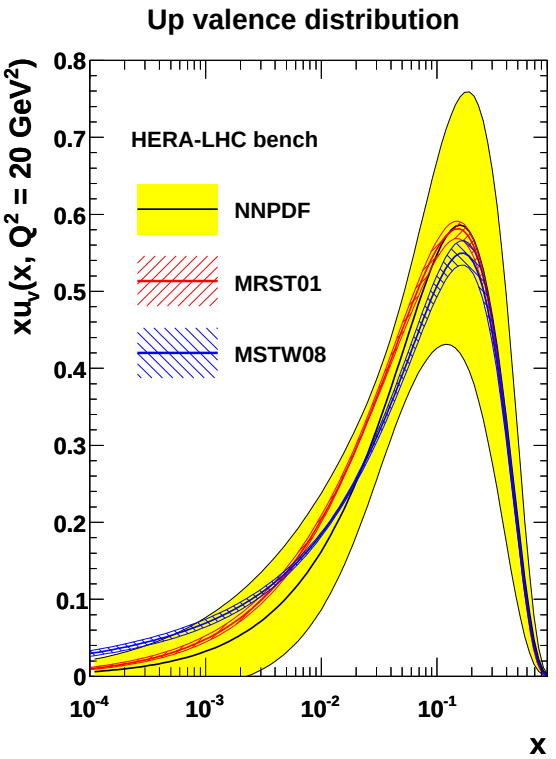
MRST/MSTW: BENCH VS REF



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

WHO ARE WE?

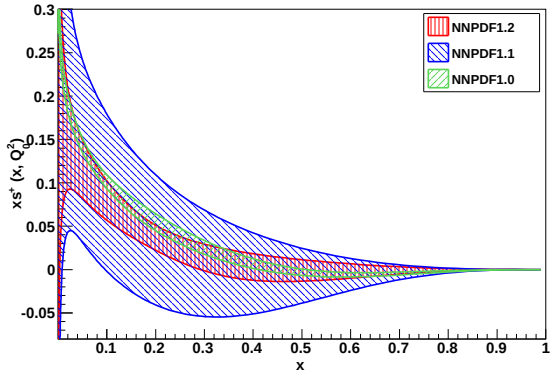
WHAT IS THIS GOOD FOR?

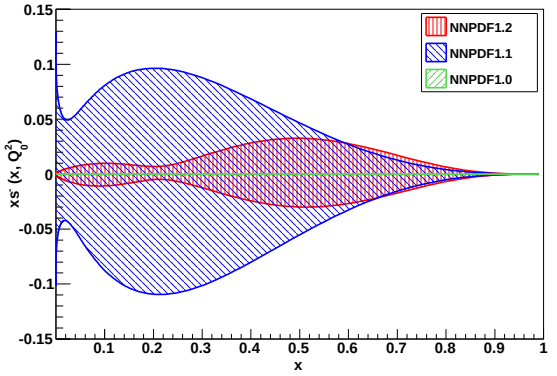
DETERMINATION OF WEAKLY CONSTRAINED QUANTITIES

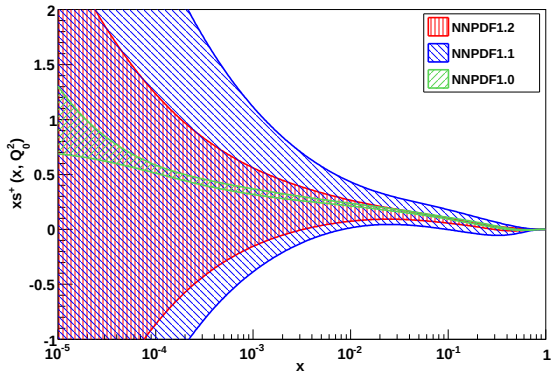
THE STRANGE PDF

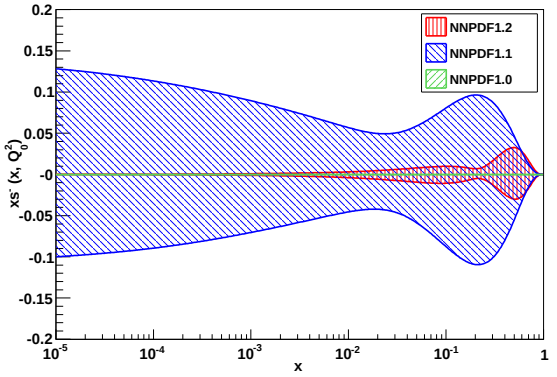
- NNPDF1.0: $s(x,Q_0^2) = \bar{s}(x,Q_0^2)$, $s+\bar{s} = \frac{1}{2}(\bar{u}+\bar{d})$, NO DIMUON DATA
- NNPDF1.1: s, $\bar{s}$ (ACTUALLY $s^{\pm}$) INDEP. PARAMETRIZED, NO DIMUON DATA
- NNPDF1.2: s, $\bar{s}$ (ACTUALLY $s^{\pm}$) INDEP. PARAMETRIZED, DIMUON DATA

STRANGE PDFS







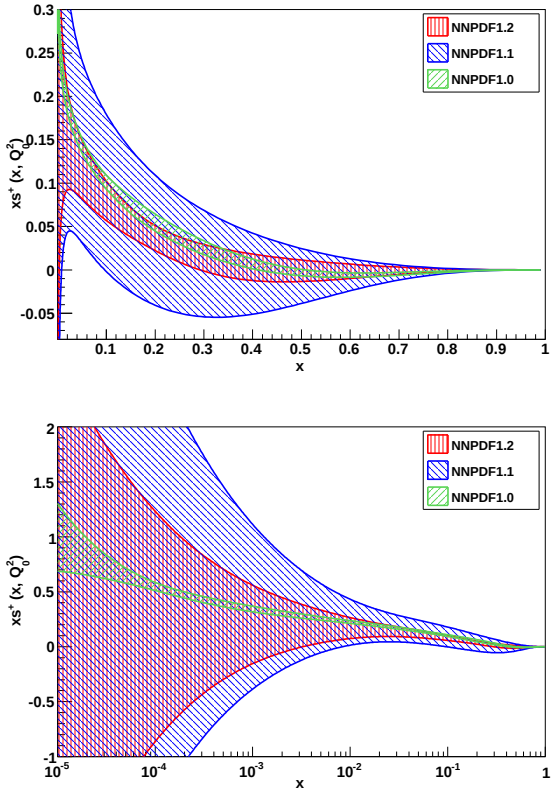


WHO ARE WE?

WHAT IS THIS GOOD FOR?

COMPATIBILITY CAN BE CHECKED QUANTITATIVELY

THE STRANGE PDF



NNPDF1.2 vs. NNPDF1.1		
	DATA	EXTRAPOLATION
$\Sigma(x, Q_0^2)$	$5\ 10^{-4} \leq x \leq 0.1$	$10^{-5} \leq x \leq 10^{-4}$
$\langle d[q] \rangle$	2.7	1.2
$\langle d[\sigma] \rangle$	3.1	1.8
$g(x, Q_0^2)$	$5\ 10^{-4} \leq x \leq 0.1$	$10^{-5} \leq x \leq 10^{-4}$
$\langle d[q] \rangle$	2.4	2.0
$\langle d[\sigma] \rangle$	1.3	1.4
$T_3(x, Q_0^2)$	$0.05 \leq x \leq 0.75$	$10^{-3} \leq x \leq 10^{-2}$
$\langle d[q] \rangle$	1.5	0.9
$\langle d[\sigma] \rangle$	1.1	1.2
$V(x, Q_0^2)$	$0.1 \leq x \leq 0.6$	$3\ 10^{-3} \leq x \leq 3\ 10^{-2}$
$\langle d[q] \rangle$	1.1	1.0
$\langle d[\sigma] \rangle$	1.3	1.4
$\Delta_S(x, Q_0^2)$	$0.1 \leq x \leq 0.6$	$3\ 10^{-3} \leq x \leq 3\ 10^{-2}$
$\langle d[q] \rangle$	0.8	0.8
$\langle d[\sigma] \rangle$	1.3	1.1
$s^+(x, Q_0^2)$	$5\ 10^{-4} \leq x \leq 0.1$	$10^{-5} \leq x \leq 10^{-4}$
$\langle d[q] \rangle$	2.0	1.6
$\langle d[\sigma] \rangle$	4.5	1.8
$s^-(x, Q_0^2)$	$0.1 \leq x \leq 0.6$	$3\ 10^{-3} \leq x \leq 3\ 10^{-2}$
$\langle d[q] \rangle$	1.1	1.3
$\langle d[\sigma] \rangle$	6.1	4.6

DISTANCE OF EXP. VALUES RESCALED BY $\sigma/\sqrt{N_{rep}}$

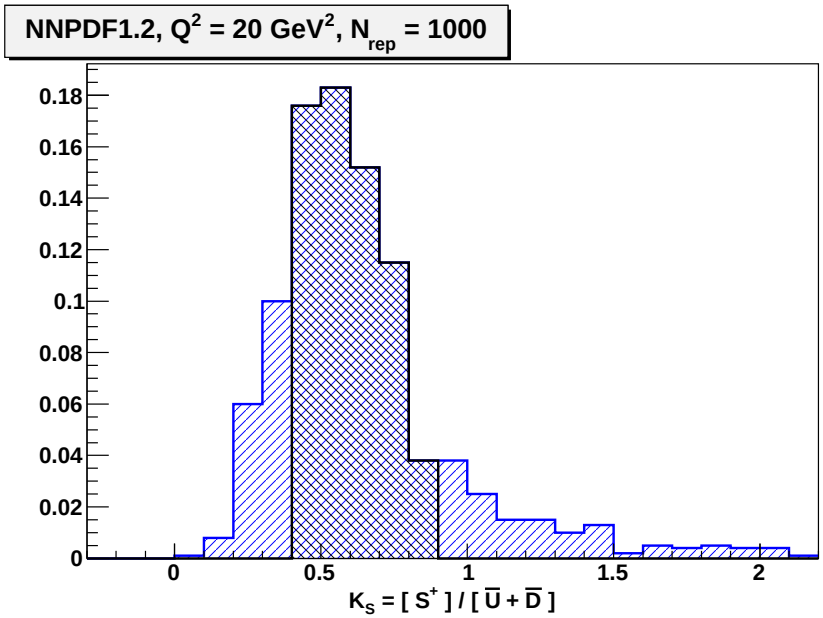
WHO ARE WE?

WHAT IS THIS GOOD FOR?

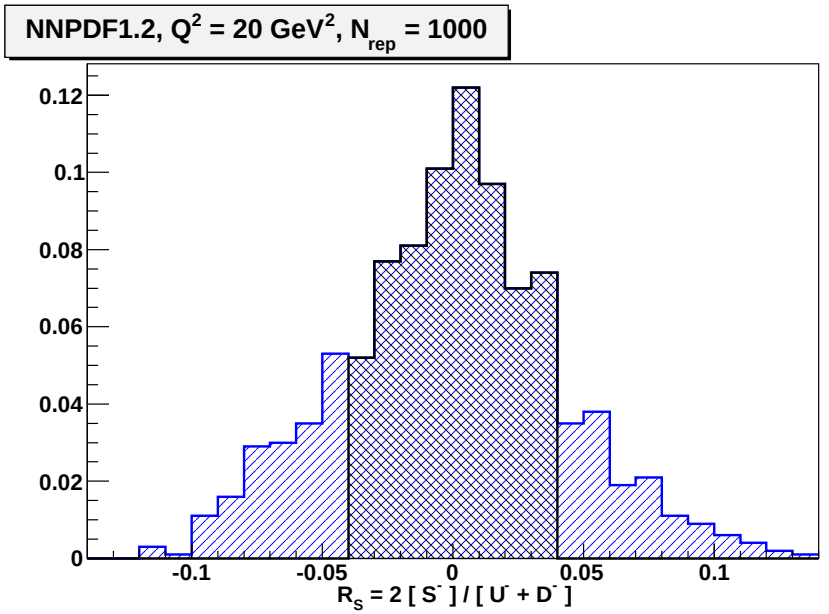
CAN DETERMINE RELIABLY LARGE & ASYMMETRIC UNCERTAINTIES

THE STRANGE MOMENTUM FRACTIONS

$s + \bar{s}$



$s - \bar{s}$



Analysis	$K_S\left(Q^2 = 20\ \mathrm{GeV}^2\right)$
NNPDF1.2	$0.71^{+0.19}_{-0.31}$
MSTW08	0.56 ± 0.03
CTEQ66	0.72 ± 0.05
AKP08	0.59 ± 0.08

Analysis	$\left[S^-\right](Q^2) \cdot 10^3$	Q^2 GeV^2	$\left[S^-\right](20\ GeV^2) \cdot 10^3$
NNPDF1.2	0.5 ± 8.6	20	0.5 ± 8.6
MSTW08	2.4 ± 2.0	1	1.4 ± 1.2
CTEQ6.5S	2.0 ± 1.8	1.69	1.2 ± 1.1
CTEQ6.1S	1.5 ± 1.5	1.69	1.2 ± 1.2
AKP08	1.0 ± 1.3	20	1.0 ± 1.3
NuTeV07	2.2 ± 1.3	16	1.3 ± 0.8
BPPZ03	1.8 ± 3.8	20	1.8 ± 3.8

AN IMPLICATION: THE "NuTeV ANOMALY" IS GONE

WHO ARE WE?

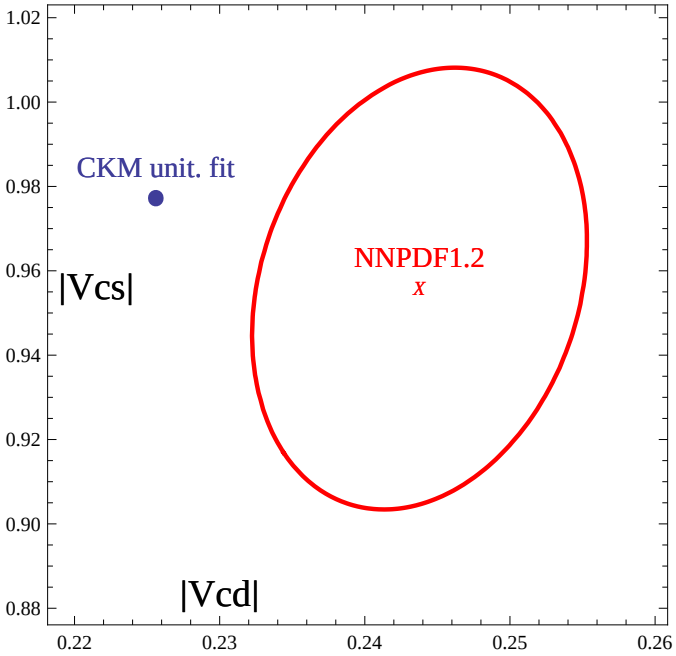
WHAT IS THIS GOOD FOR?

CAN DISENTANGLE PHYSICAL PARAMETERS FROM PDFs

DETERMINATION OF CKM PARMS FROM DIS

$$F_2^{\nu,c} = x\left[C_{2,q}\otimes\left(|V_{\rm cd}|^2(u+d)+2|V_{\rm cs}|^2 s\right)+C_{2,g}\otimes g\right]$$

$$F_2^{\bar\nu,c} = x\left[C_{2,q}\otimes\left(|V_{\rm cd}|^2(\bar u+\bar d)+2|V_{\rm cs}|^2 \bar s\right)+C_{2,g}\otimes g\right]$$



Analysis	Determination	$\lvert V_{cs}\rvert$
NNPDF1.2	Direct from global PDF analysis	$0.96 \pm 0.07^{\rm tot}$
CDHS	LO from $\nu N \to \mu^+\mu^- X$	≥ 0.59 (90% C.L.)
CCFR	NLO from $\nu N \to \mu^+\mu^- X$	≥ 0.74 (90% C.L.)
PDG08	Average from D decays	1.04 ± 0.06
Hocker	Average from $\nu N \to \mu^+\mu^- X$	1.04 ± 0.16
DELPHI	Direct from $W^+ \to c\bar s$ decays	$0.94^{+0.32}_{-0.26} \pm 0.13$
PDG08	CKM unitarity fit	0.97334 ± 0.00023

Analysis	Determination	$\lvert V_{cd}\rvert$
NNPDF1.2	Direct from global PDF analysis	$0.244 \pm 0.019^{\rm tot}$
CDHS	LO from $\nu N \to \mu^+\mu^- X$	0.24 ± 0.03
CCFR	NLO from $\nu N \to \mu^+\mu^- X$	$0.232^{+0.017}_{-0.019}$
PDG08	Average from $\nu N \to \mu^+\mu^- X$	0.230 ± 0.011
PDG08	Average from $D \to K/\pi l\nu$ decays	0.218 ± 0.023
PDG08	CKM unitarity fit	0.2256 ± 0.0010

$\Rightarrow$ MOST PRECISE AVAILABLE DIRECT DETERMINATION OF V_{cs}

WHO ARE WE?

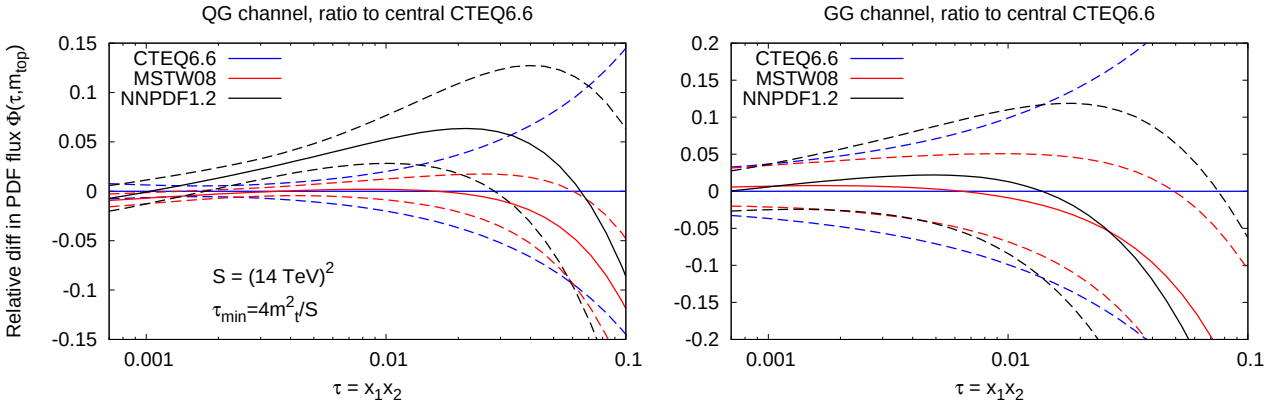
WHAT IS THIS GOOD FOR?

CAN TRUST RESULTS!

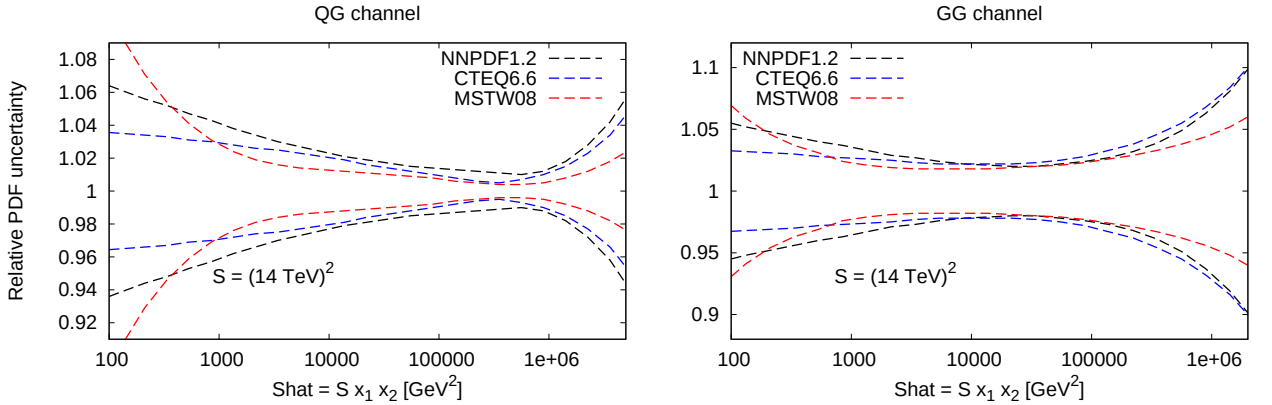
PARTON FLUXES AT THE LHC

$$\Phi_{ij}\left(\tau,\mu^2\right)=\int_0^1 dx_1\int_0^1 dx_2\, q_i\left(x_1,\mu^2\right) q_j\left(x_2,\mu^2\right)\delta\left(x_1x_2-\tau\right)$$

FLUX



RELATIVE UNCERTAINTY ON FLUX



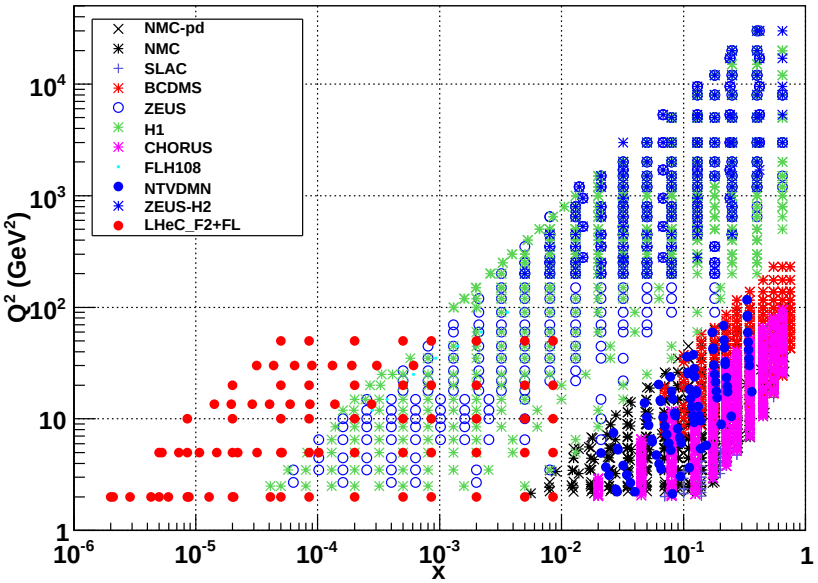
- UNCERTAINTIES IN FUTURE (GLOBAL) NNPDF FIT CAN ONLY DECREASE
- SENSITIVITY TO α_s OF NNPDF PARTONS NEGLIGIBLE

WHERE ARE WE GOING?

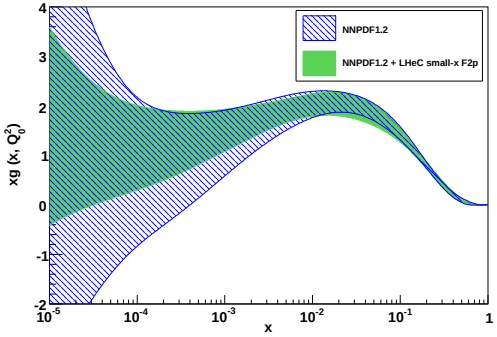
WHAT IS THIS GOOD FOR?

CAN STUDY UNCERTAINTIES AT FUTURE ACCELERATORS

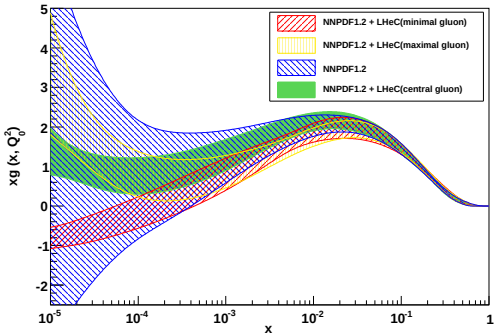
GLUON DETERMINATION AT THE LHeC



IMPACT OF LHeC F_2 DATA



LHeC $F_2 + F_L$ DATA



CAN DISENTANGLE DIFFERENT SCENARIOS FOR SMALL x GLUON BEHAVIOUR

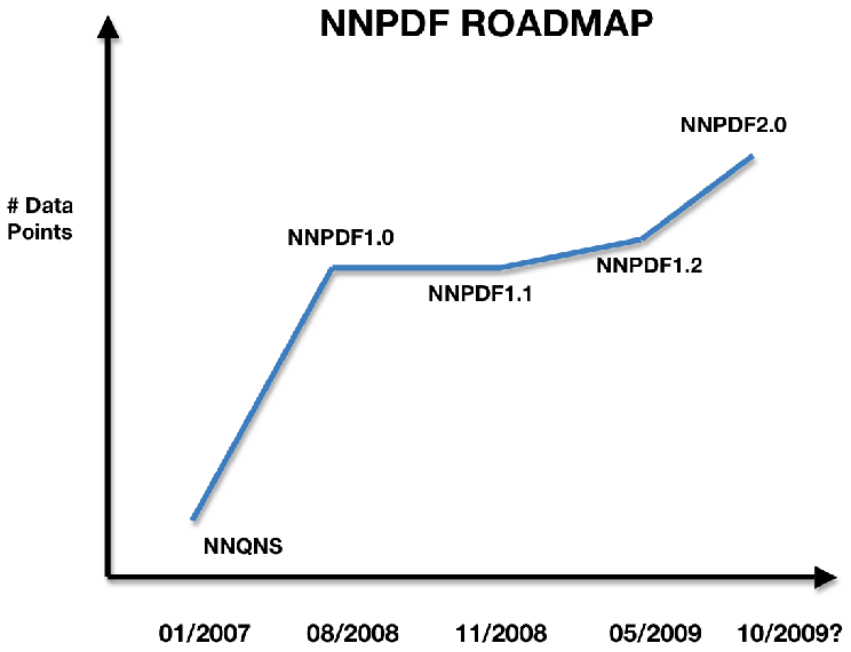
- THE MONTE CARLO SAMPLE CAN BE USED DIRECTLY FOR SUCH STUDIES WITHOUT REFITTING:
  JUST REWEIGHT SAMPLE ACCORDING TO PROJECTED DATA
- LHC STUDIES ALONG THESE LINES POSSIBLE/IN PREPARATION
  (SEE MCNULTY'S TALK)

WHERE ARE WE GOING?

WHAT NEXT?

ROADMAP

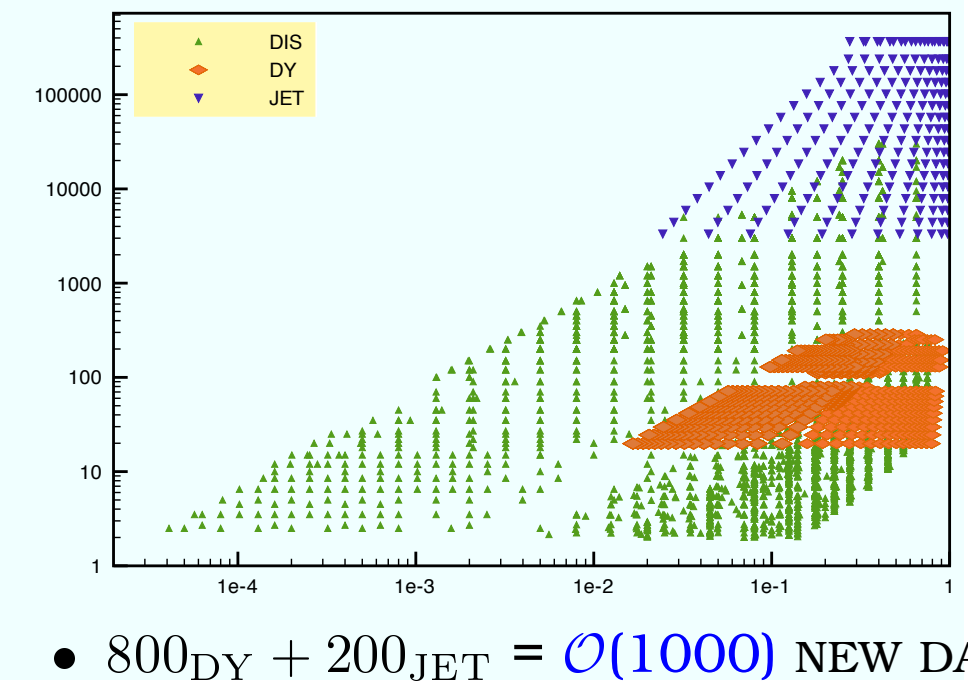
- 2007 $\rightarrow$ FIRST (NONSINGLET) FIT
  500 DIS DATA; ONE PDF
- 2008 $\rightarrow$ FIRST PARTON SET
  4000 DIS DATA; FIVE, THEN SEVEN PDFS
- 2009 $\rightarrow$ FIRST GLOBAL FIT
  4000 DATA: DIS, DY, JETS; SEVEN PDFS



WHERE ARE WE GOING?

WHAT NEXT?

NNPDF2.0



OBS	DATA SET
$d\sigma^{\rm DY}/dM^2dy$	E605
$d\sigma^{\rm DY}/dM^2dx_F$	E886
W ASYM.	D0/CDF
Z RAP. DISTR.	D0/CDF
INCL. $\sigma^{\rm jet}$	CDF(k_T)
INCL. $\sigma^{\rm jet}$	D0(CONE)

- $800_{\rm DY} + 200_{\rm JET}$ = $\mathcal{O}(1000)$ NEW DATA
- FASTDY ALGORITHM $\Rightarrow$ FIRST TRULY NLO FIT
- DEFECT: STILL ZM-VFN SCHEME FOR HEAVY QUARKS

SUMMARY

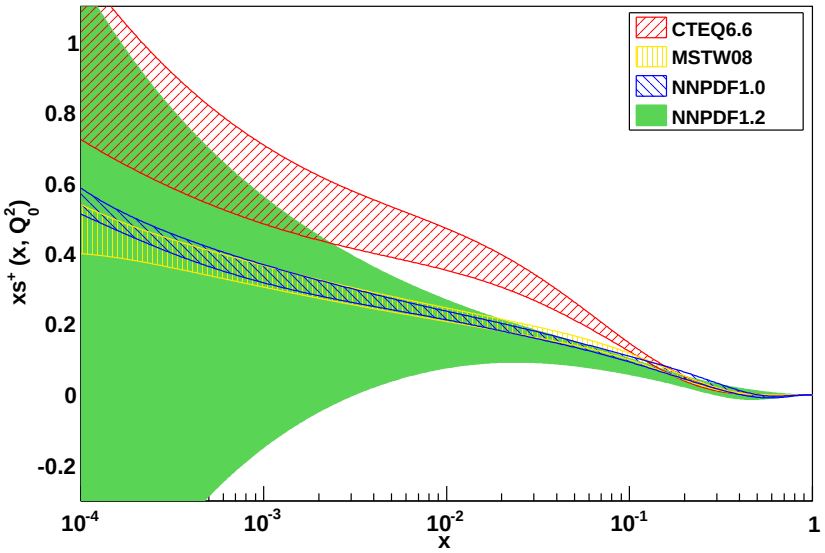
- THE NNPDF APPROACH TRIES TO AVOID AS MUCH AS POSSIBLE
  - PARAMETRIZATION BIAS
  - SUBJECTIVE CHOICES IN UNCERTAINTY ESTIMATE AND COMBINATION
  - ASSUMPTIONS ON THE SHAPE OF THE UNDERLYING PROBABILITY DISTRIBUTION (GAUSSIANITY, SYMMETRY, LINEAR ERROR PROPAGATION. . .)
- A SET OF PDFS BASED ON THIS ASSUMPTIONS IS AVAILABLE, IT HAS BEEN USED FOR PHYSICAL APPLICATIONS AND IT IS ALREADY USEFUL FOR COMPETITIVE UNCERTAINTY ESTIMATE
- A GLOBAL SET OF PDFS BASED ON THIS METHODOLOGY IS BEHIND THE CORNER
- IT WILL ALLOW STUDIES OF DATA COMPATIBILITY, UNCERTAINTY ASSESSMENT, AND IT WILL VALIDATE OTHER PDF STUDIES

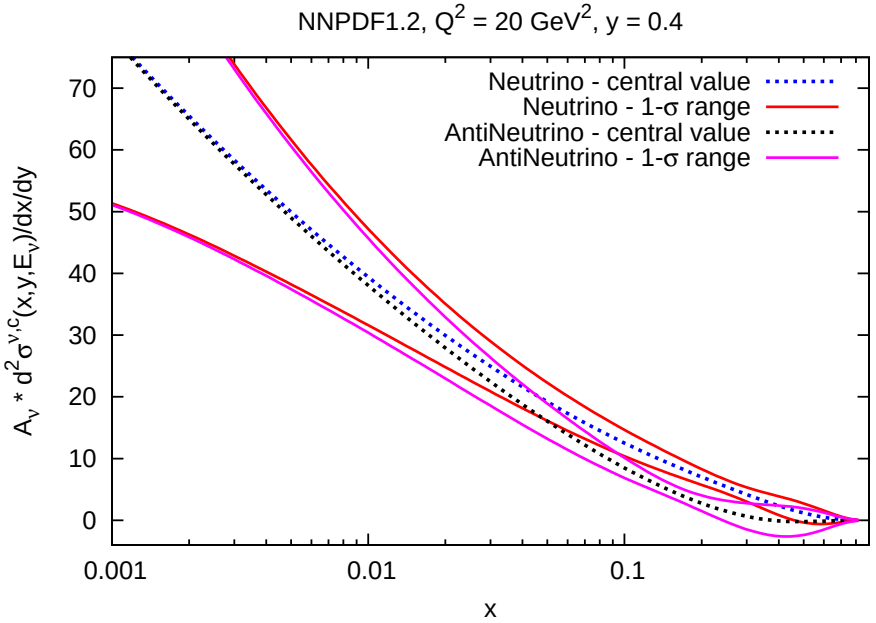
EXTRAS

POSITIVITY

- PDFs can go negative, provided physical cross sections remain positive definite
- (in principle) this should be true for ANY observable
- in practice, mostly relevant close to the data edge
- NNPDF1.x: positivity of F_L enforced

STRANGE PDF AND DIMUON XSECT

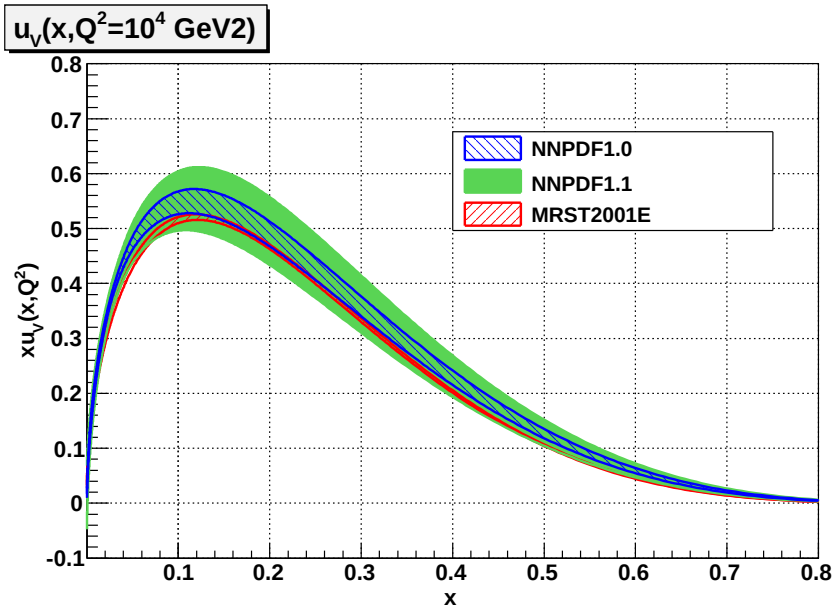


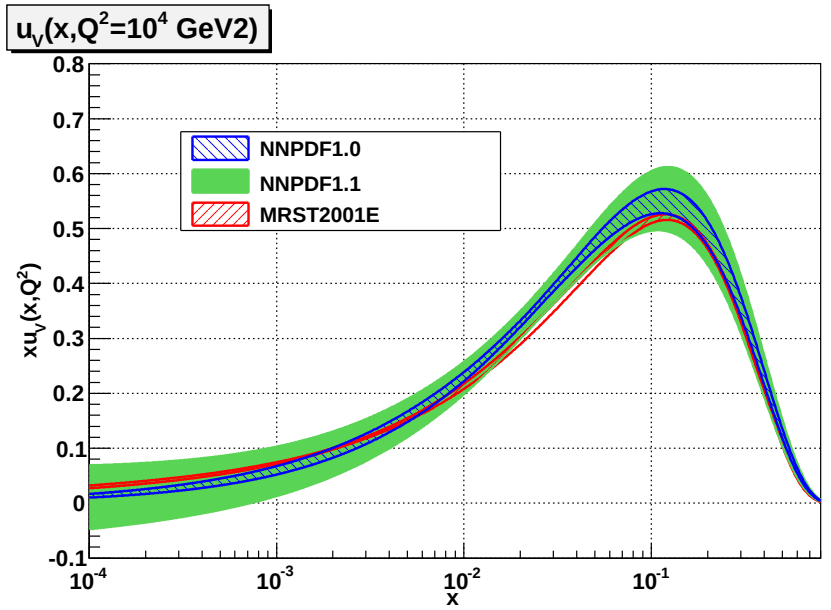


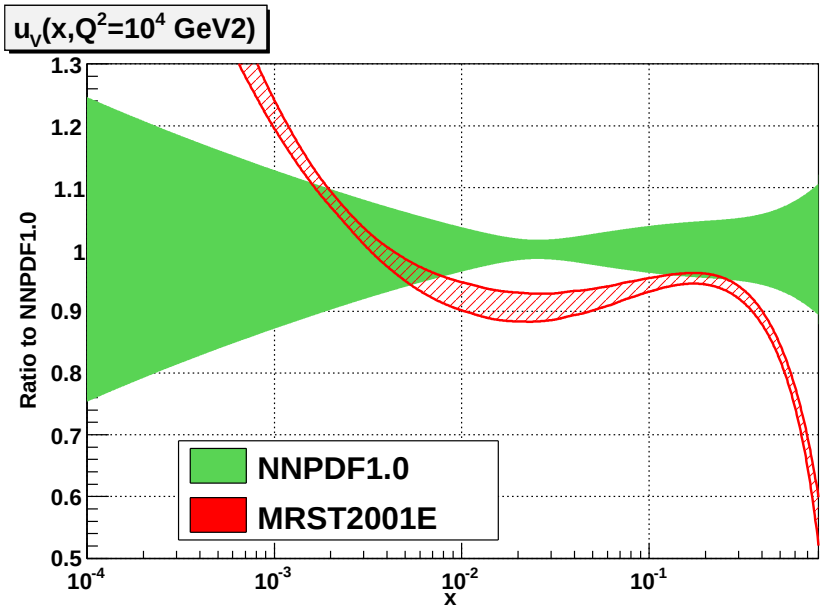
- STRANGENESS GOES NEGATIVE PRETTY SOON...
- BUT THE CROSS-SECTION REMAINS POSITIVE

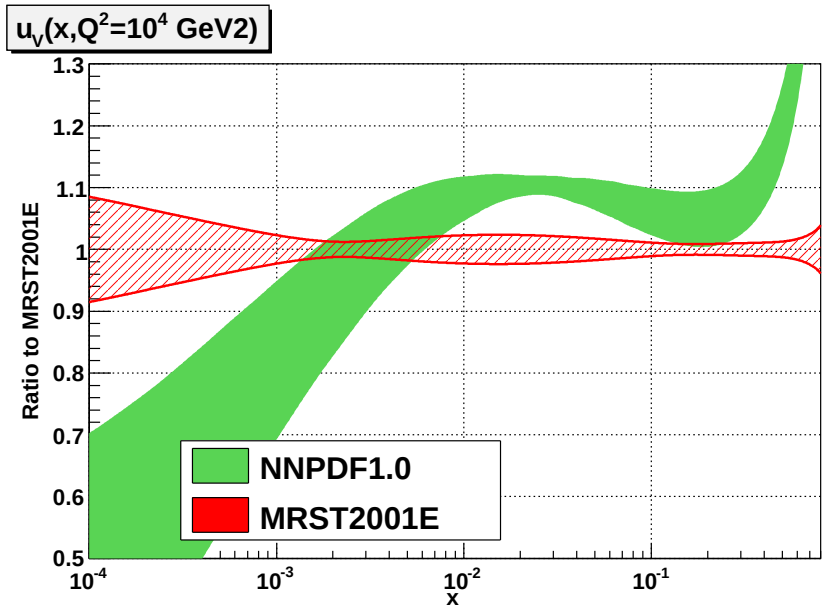
COMPARISON PLOTS

BEWARE OF RATIOS!





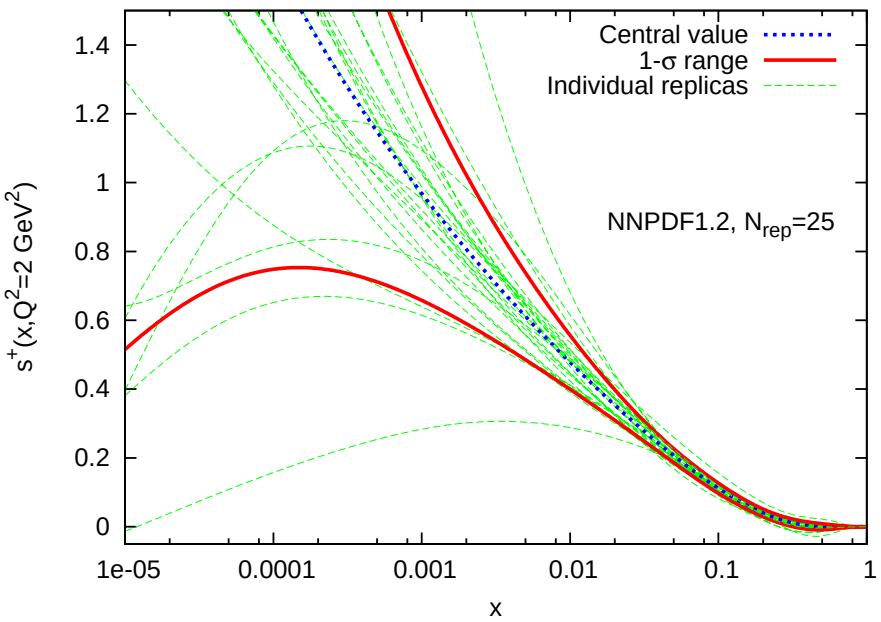


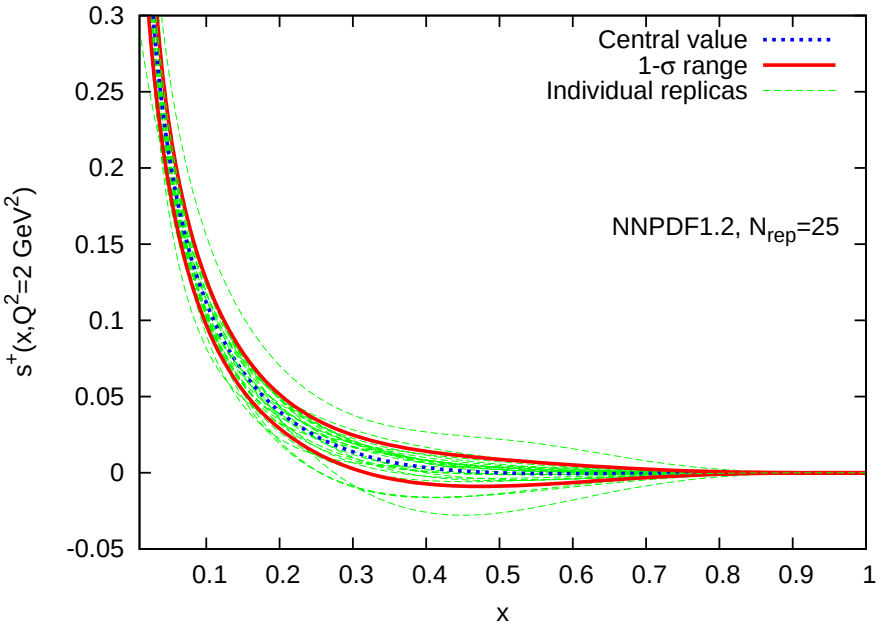


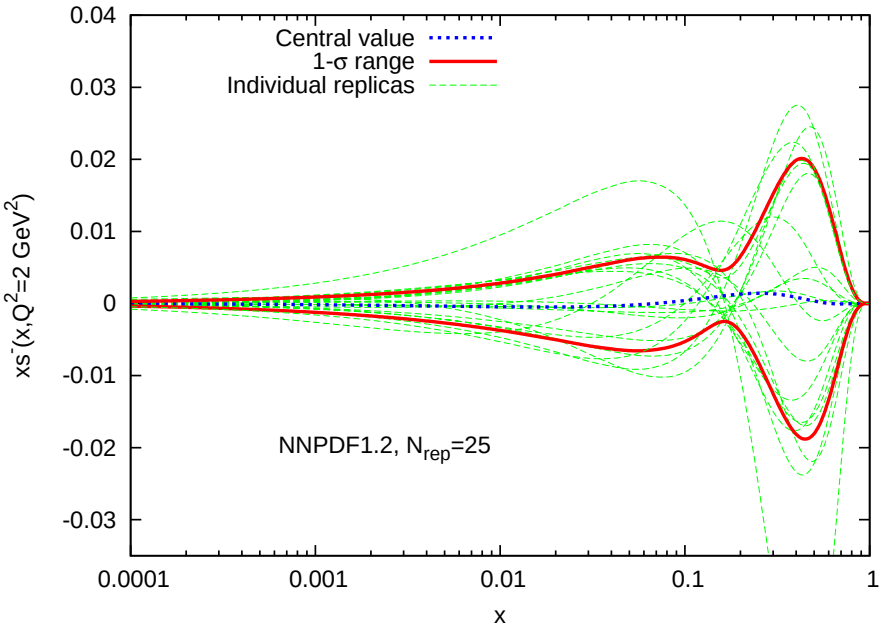
THE PDF IN THE NUMERATOR ALWAYS LOOKS SILLY!

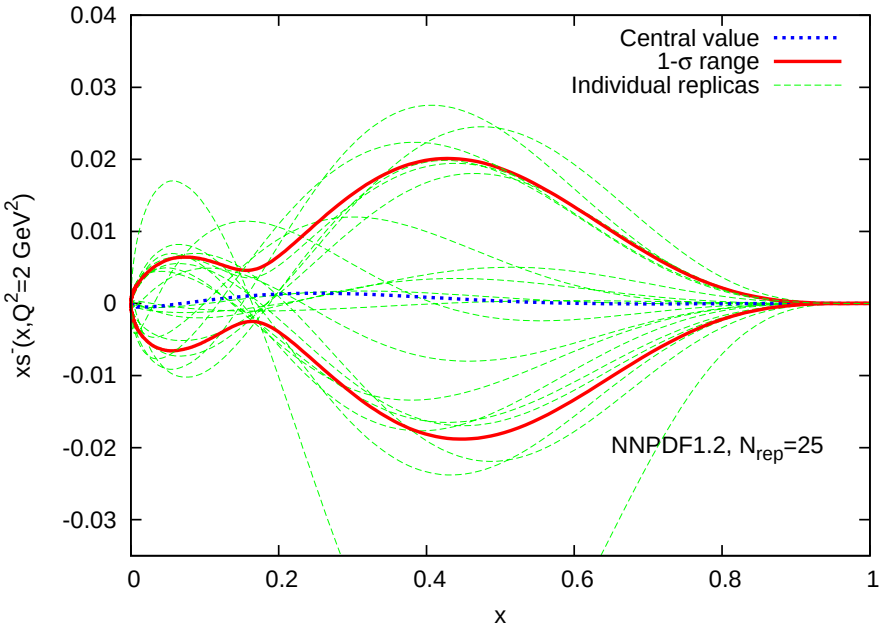
REPLICAS: STRANGENESS

FLEXIBLE PARAMETRZATION WITH LITTLE INFORMATION









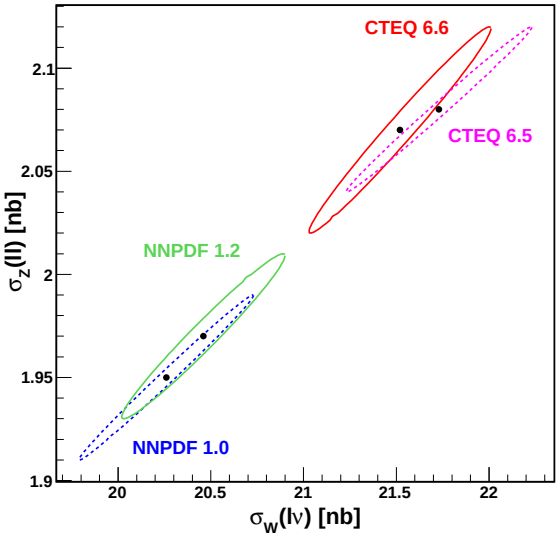
s^- MAY CHANGE SIGN SEVERAL TIMES!

IMPACT ON SOME LHC STANDARD CANDLES:

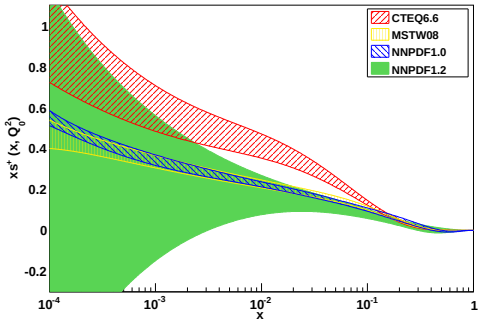
IT COULD BE WORSE

- TOTAL W/Z XSCT DOMINATED BY CENTRAL RAPIDITY REGION
  $\Rightarrow$ UNCERTAINTY ON STRANGENESS UNDER CONTROL
- AWAY FROM CENTRAL REGION UNCERTAINTY MUCH LARGER

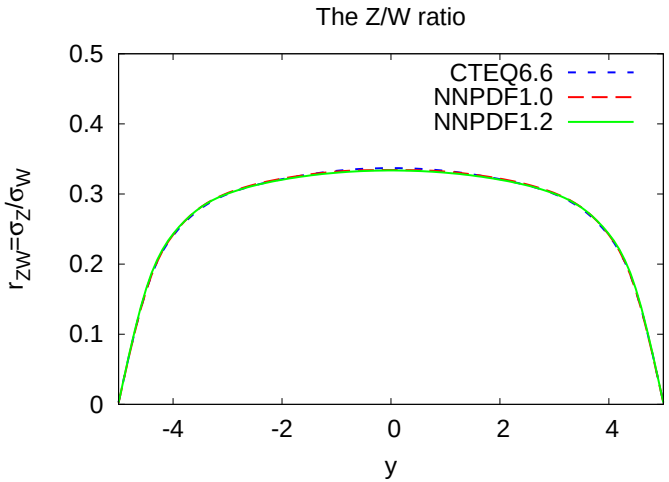
TOTAL XSECTNS



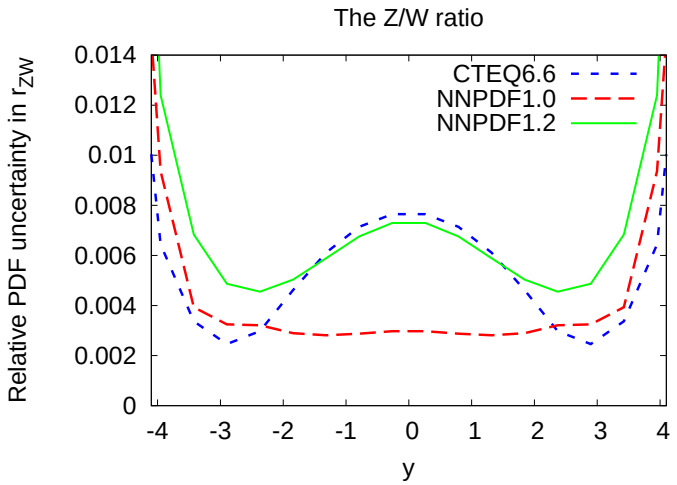
STRANGENESS (REMINDER)



RAPIDITY DISTRIBUTION



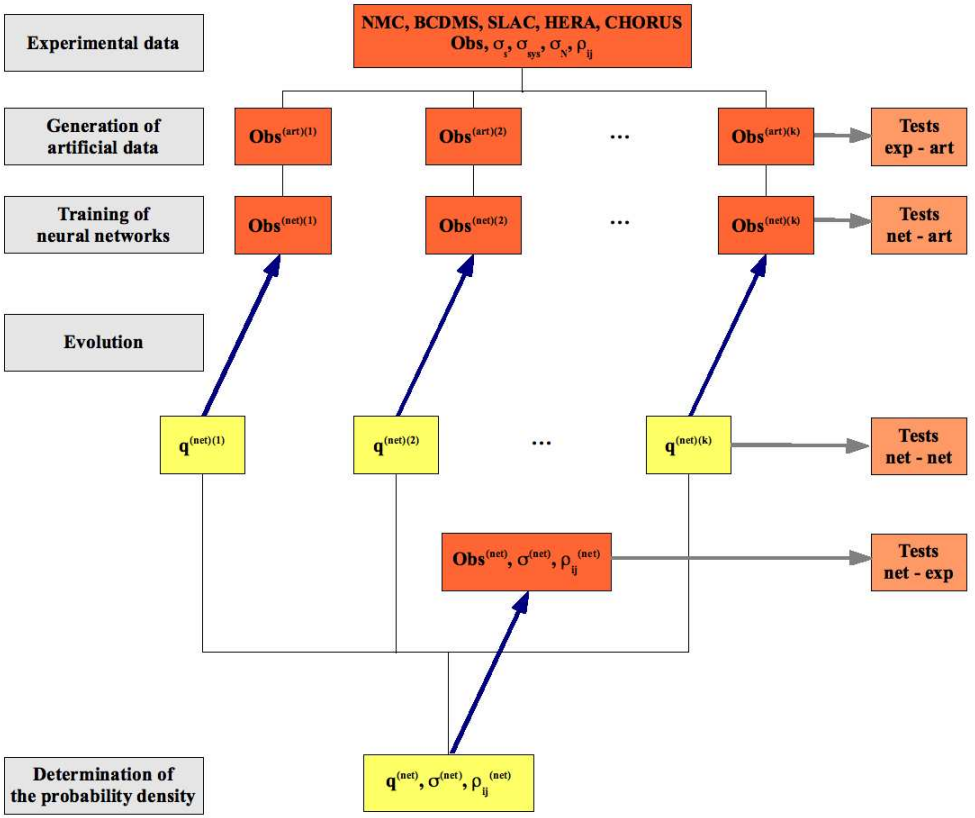
PERC. UNCERTAINTY



THE NEURAL MONTE CARLO

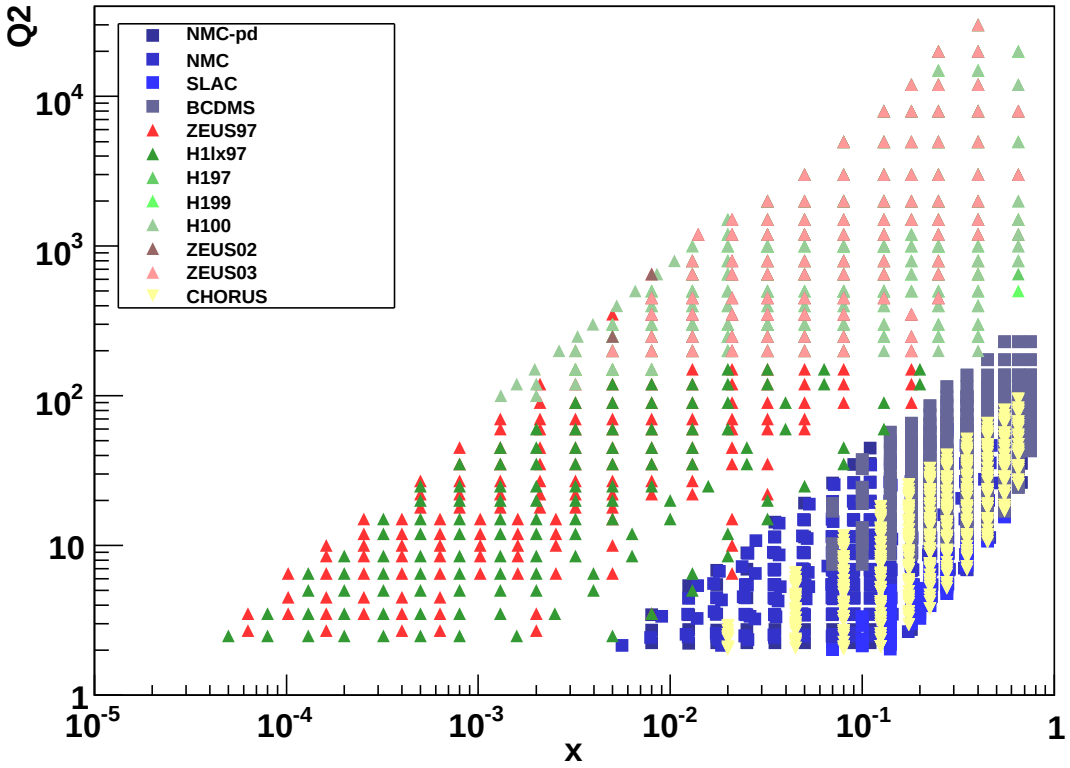
- GENERATE A SET OF MONTE CARLO REPLICAS $\sigma^{(k)}(p_i)$ OF THE ORIGINAL DATASET $\sigma^{(\mathrm{data})}(p_i)$
  $\Rightarrow$ REPRESENTATION OF $\mathcal{P}[\sigma(p_i)]$ AT DISCRETE SET OF POINTS p_i
- TRAIN A NEURAL NET FOR EACH PDF ON EACH REPLICA, THUS OBTAINING A NEURAL REPRESENTATION OF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

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



FEATURES OF THE FIT

THE DATASET



$Q^2 > 2$ GeV^2; $W^2 > 12.5$ GeV^2

NAME	DATA POINTS	TARGET
NMC_PD	153	F_2^d / F_2^p
NMC	245	F_2^p
SLAC	47 (47)	$F_2^{p(d)}$
BCDMS	333 (248)	$F_2^{p(d)}$
ZEUS97	240 (29)	$\tilde{\sigma}^{NC(CC),+}$
ZEUS02	92 (26)	$\tilde{\sigma}^{NC(CC),-}$
ZEUS03	90 (30)	$\tilde{\sigma}^{NC(CC),+}$
H1LX97	135	$\tilde{\sigma}^{NC,+}$
H197	130 (25)	$\tilde{\sigma}^{NC(CC),+}$
H199	139 (28)	$\tilde{\sigma}^{NC(CC),-}$
H100	147 (28)	$\tilde{\sigma}^{NC(CC),+}$
H108	8	F_L
CHORUS	471 (471)	$\tilde{\sigma}^{\nu(\bar{\nu})}$
TOTAL	3161	

FEATURES OF THE FIT

THEORY

- NLO EVOLUTION (N – SPACE, EXPANDED)
- ZM-VFN SCHEME FOR THRESHOLDS
- $\alpha_s(M_z) = 0.119$, PDFS GIVEN AT $Q_0^2 = 2$ GeV^2
- TARGET-MASS CORRECTIONS INCLUDED UP TO TWIST FOUR

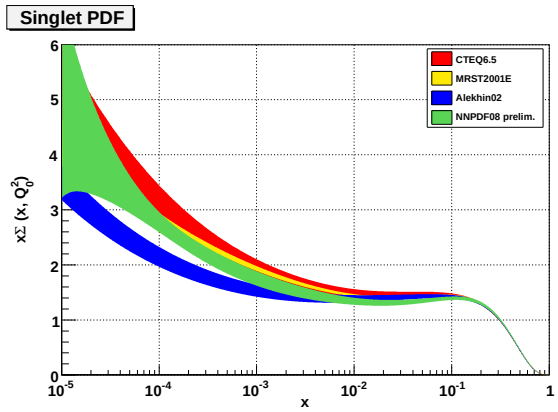
BASIS FUNCTIONS AND PARAMETRIZATION

- FIVE INDEPENDENT PDFS: SINGLET, GLUON, TOTAL VALENCE, TRIPLET, $\bar{d} - \bar{u}$.
- SYMMETRIC STRANGE SEA $s(x) = \bar{s}(x)$,
  PROPORTIONAL TO NON-STRANGE SEA, $\bar{s}(x) = \frac{C}{2}(\bar{u}(x) + \bar{d}(x))$, (C = 0.5)
- ALL PDFS PARAMETRIZED BY A 2-5-3-1 NEURAL NETWORK:
  $37 \times 5 = 185$ PARAMETERS
- MOMENTUM AND VALENCE SUM RULES ENFORCED STRICTLY
- POSITIVITY OF F_L ENFORCED for $x \geq 10^{-7}$, $Q^2 \geq 2$ GeV^2

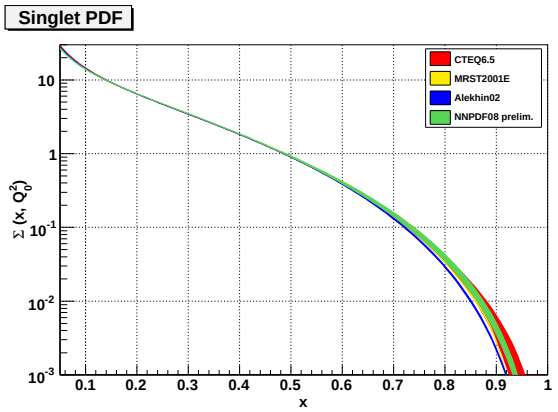
RESULTS

PARTON DISTRIBUTIONS

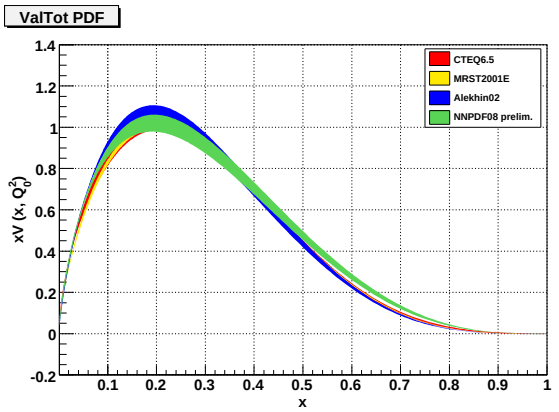
SINGLET



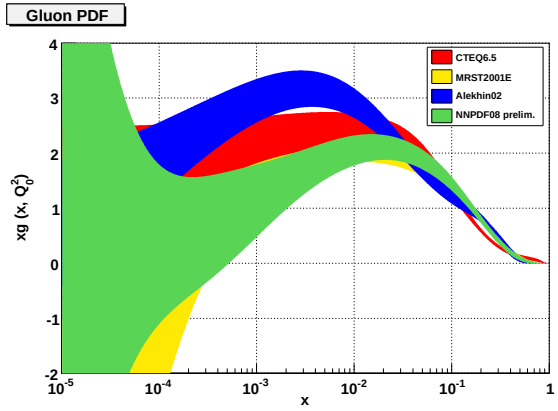
SINGLET



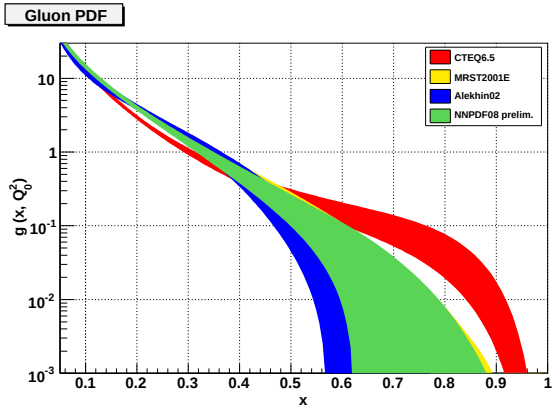
VALENCE



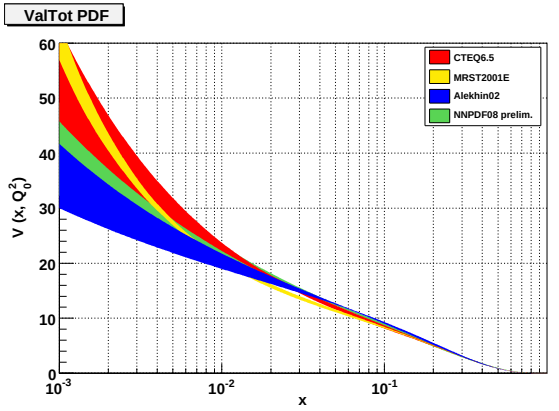
GLUON



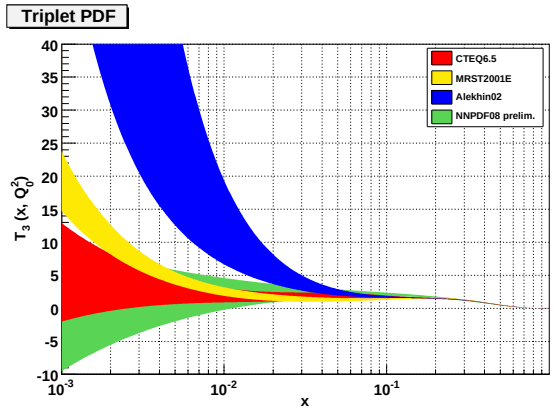
GLUON



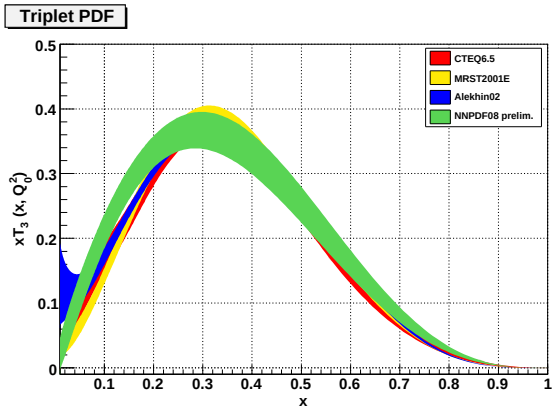
VALENCE



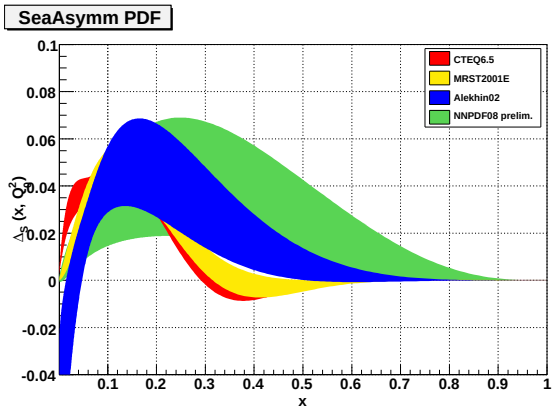
TRIPLET



TRIPLET



SEA ASYM.

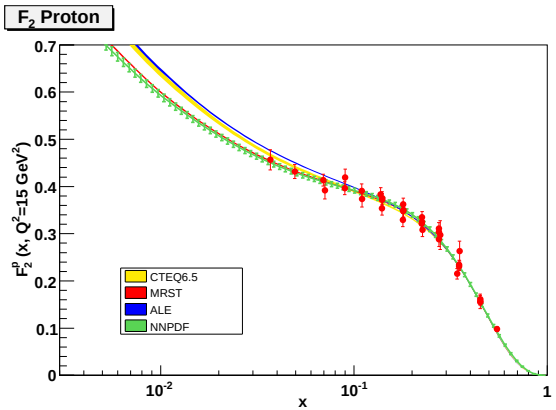


RESULTS

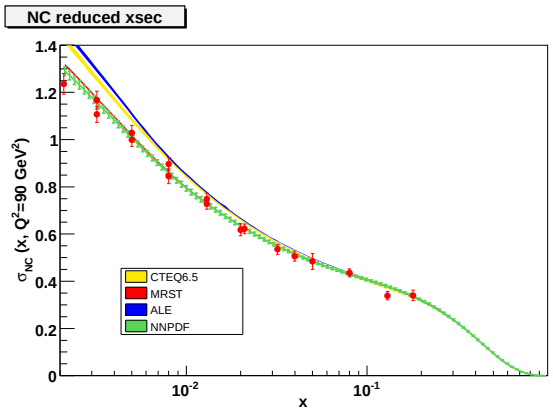
PHYSICAL OBSERVABLES

STRUCTURE FUNCTIONS

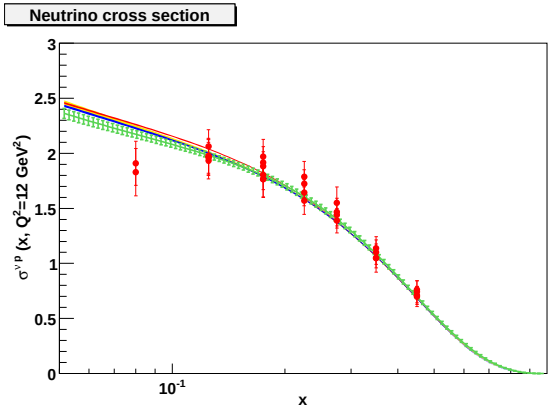
PROTON F_2



NC XSECT.

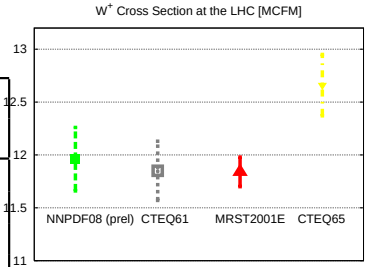


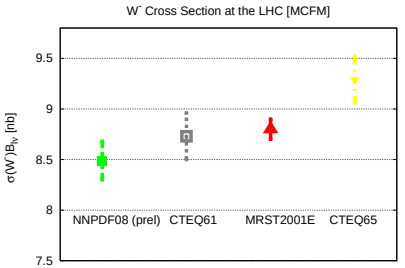
NEUTRINO XSECT.



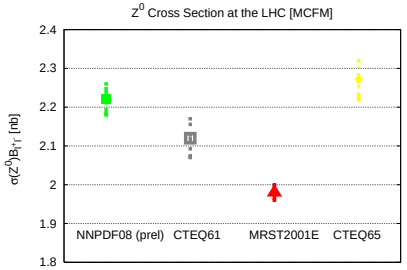
TOTAL CROSS-SECTIONS AT LHC, NLO FROM MCFM

-	$\sigma_{W+}\mathcal{B}_{l+\nu}$ [nb]	$\Delta\sigma/\sigma$ $W+$	$\sigma_{W-}\mathcal{B}_{l-\nu}$ [nb]	$\Delta\sigma/\sigma$ $W-$
NNPDF08	11.96 ± 0.30	2.5%	8.49 ± 0.19	2.3%
CTEQ6.5	12.66 ± 0.29	2.3%	9.29 ± 0.23	2.5%
CTEQ6.1	11.85 ± 0.28	2.4%	8.73 ± 0.23	2.6%
MRST01	11.84 ± 0.14	1.2%	8.80 ± 0.10	1.1%





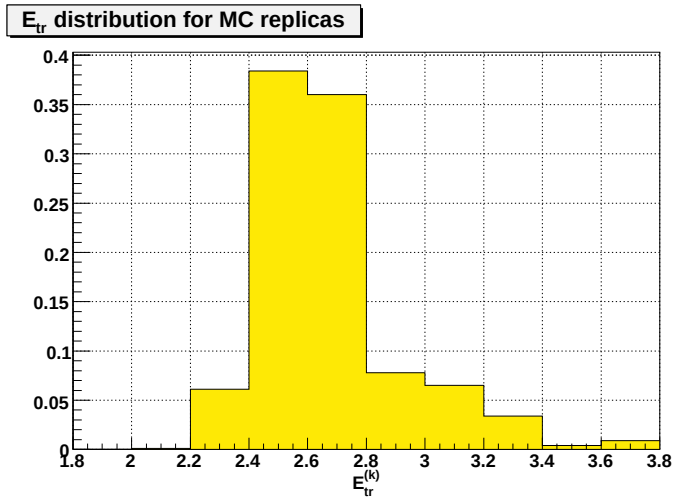
	$\sigma_Z\mathcal{B}_{l+l-}$ [nb]	$\Delta\sigma/\sigma$ Z	$\sigma_{t\bar{t}}$ [pb]	$\Delta\sigma/\sigma$ $t\bar{t}$	σ_H [pb]	$\Delta\sigma/\sigma$ H
NNPDF08	2.22 ± 0.04	2.0%	1014 ± 24	2.3%	35.79 ± 1.04	3.0%
CTEQ6.5	2.27 ± 0.05	2.2%	942 ± 19	2.0%	37.51 ± 0.80	2.2%
CTEQ6.1	2.12 ± 0.05	2.3%	970 ± 18	1.9%	38.50 ± 0.85	2.2%
MRST01	1.98 ± 0.02	1.0%	1013 ± 13	1.3%	37.52 ± 0.40	1.1%



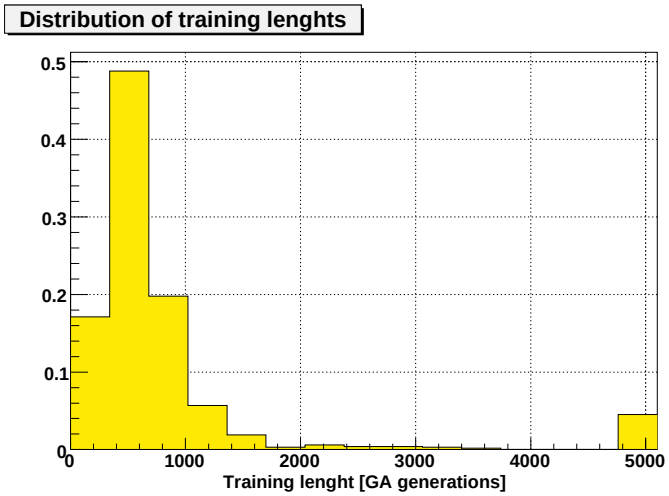
RESULTS

GENERAL STATISTICAL FEATURES

χ^2 AT STOPPING



TRAINING LENGTHS



$\chi^2_{\rm tot}$	1.34
$\langle E \rangle$	2.71
$\langle E_{\rm tr} \rangle$	2.68
$\langle E_{\rm val} \rangle$	2.72
$\langle {\rm TL} \rangle$	824.23
$N_{\rm rep}$	1000

- POISSONIAN DISTRIBUTION OF TRAINING LENGTHS
- BEST FIT $\chi^2 = 1.34$: MINOR DATA INCOMPATIBILITIES (?)

PARAMETRIZATION INDEPENDENCE:
METHODOLOGY

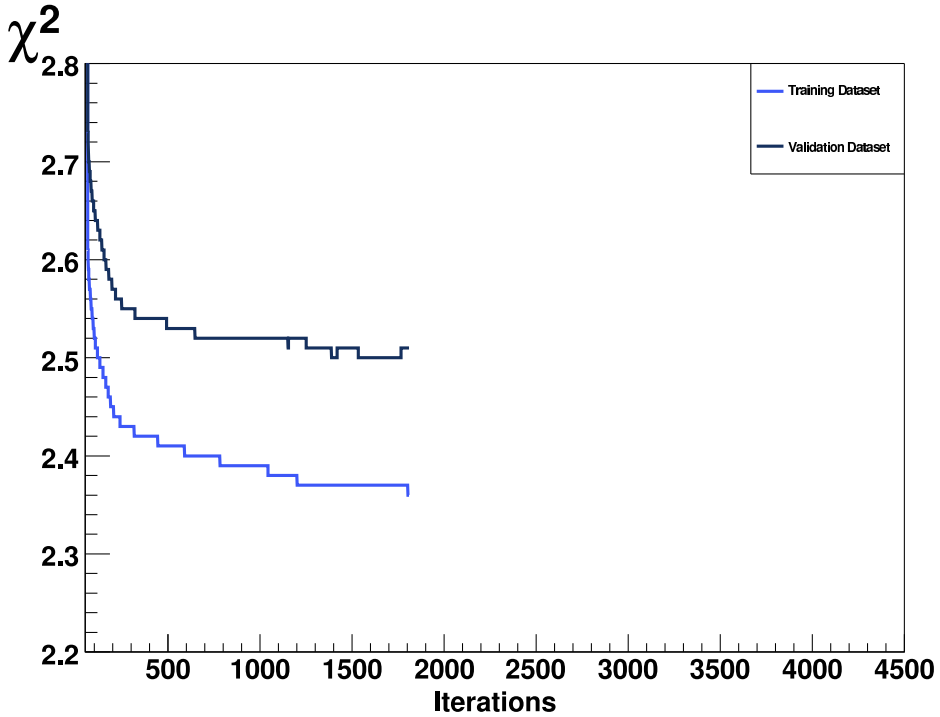
- EFFECTIVELY INFINITE NUMBER OF PARAMETERS
  $\Rightarrow$ CAN REPRESENT ANY FUNCTION
- COMPLEX SHAPES (LARGE NO.OF PARAMETERS) REQUIRE LONGER FITTING
- FIT STOPS WHEN QUALITY OF FIT TO RANDOMLY SELECTED "VALIDATION" DATA (NOT FITTED) STOPS IMPROVING
- CAN OBTAIN A FIT WITH χ^2 LOWER THAN BEST FIT ("OVERLEARNING")

PARAMETRIZATION INDEPENDENCE:
REDUNDANCY AND OVERLEARNING

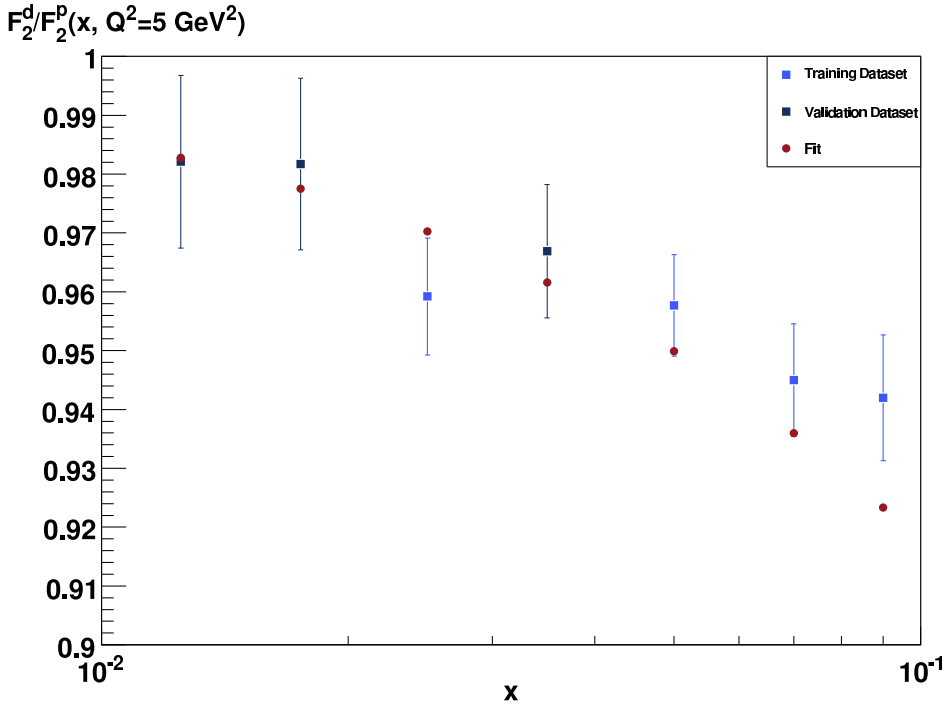
- OPTIMAL FIT OBTAINED WHEN QUALITY OF FIT TO VALIDATION (CONTROL) DATA STOPS IMPROVING
- POSSIBILITY OF OVERFITTING GUARANTESS THAT MINIMUM NOT DRIVEN BY PARAMETRIZATION

OPTIMAL FITTING

χ^2



FIT TO DATA

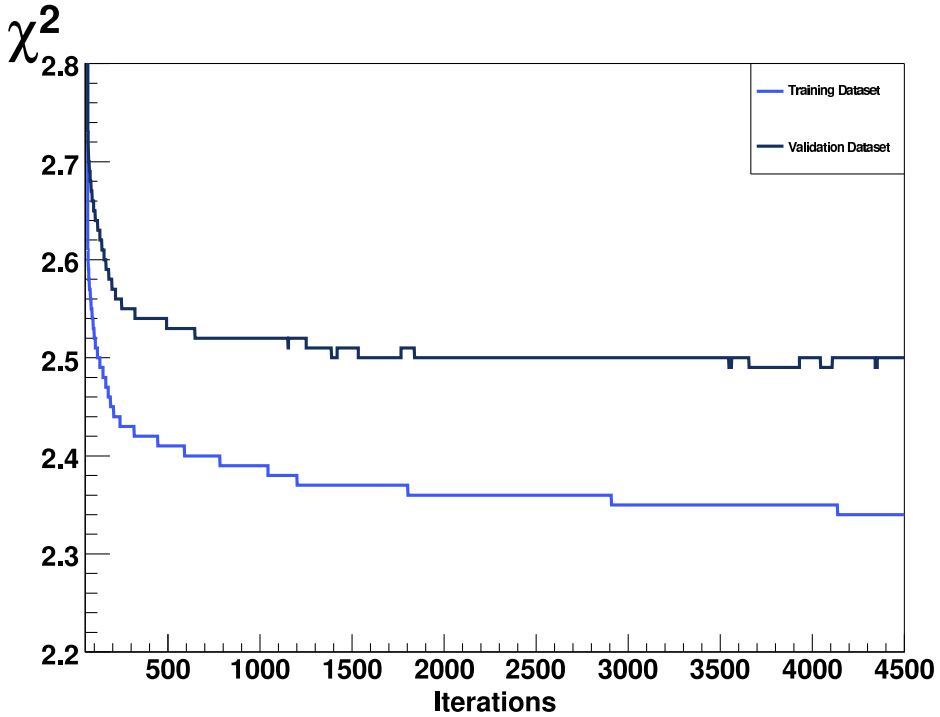


PARAMETRIZATION INDEPENDENCE:
REDUNDANCY AND OVERLEARNING

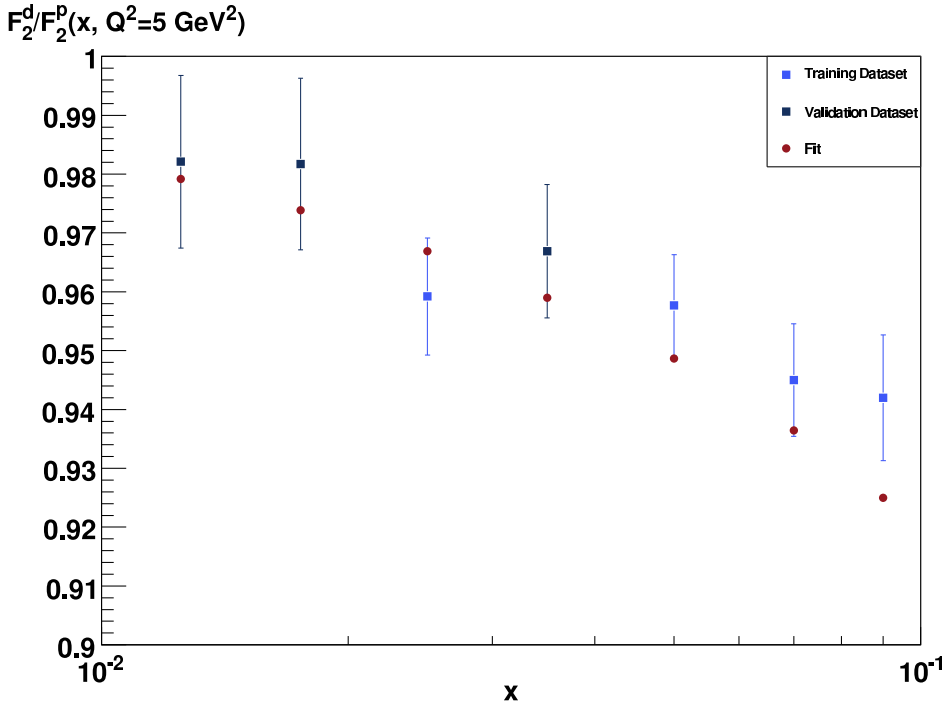
- OPTIMAL FIT OBTAINED WHEN QUALITY OF FIT TO VALIDATION (CONTROL) DATA STOPS IMPROVING
- POSSIBILITY OF OVERFITTING GUARANTESS THAT MINIMUM NOT DRIVEN BY PARAMETRIZATION

OVERFITTING

χ^2



FIT TO DATA

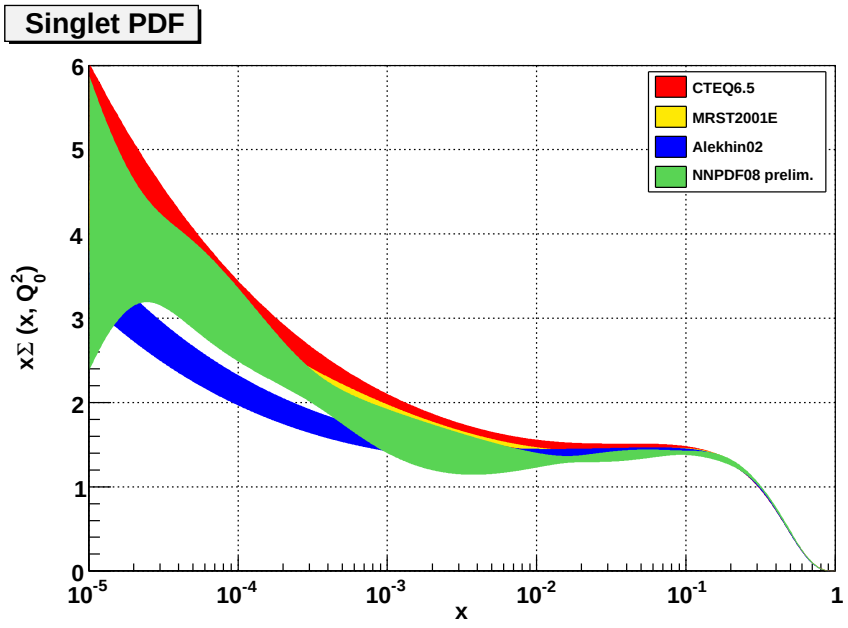


PARAMETRIZATION INDEPENDENCE:
LACK OF BIAS

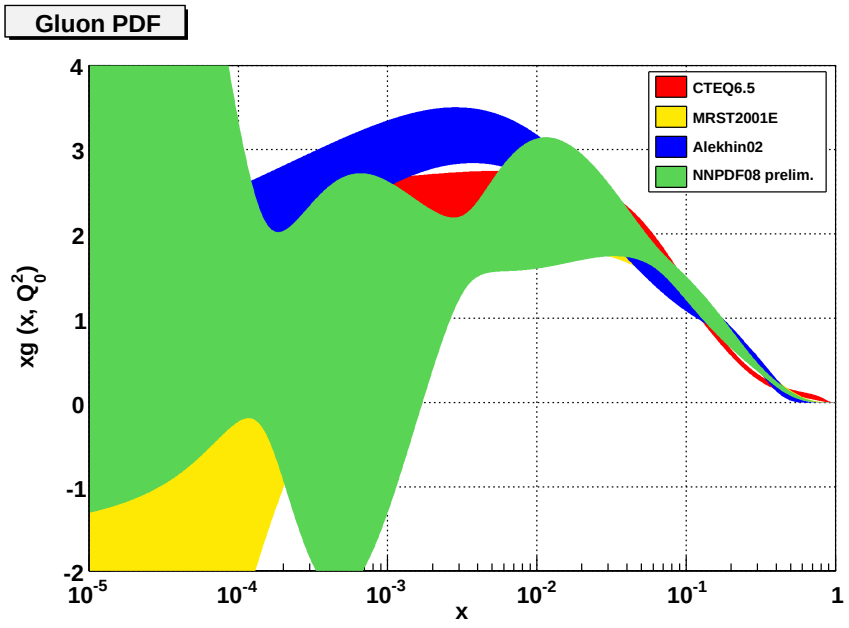
- IRREGULAR OR KNOTTY SHAPES ALLOWED IF DATA FLUCTUATE
- STATISTICS SHOW WHETHER THE EFFECT IS REAL

10 REPLICAS

SINGLET



GLUON

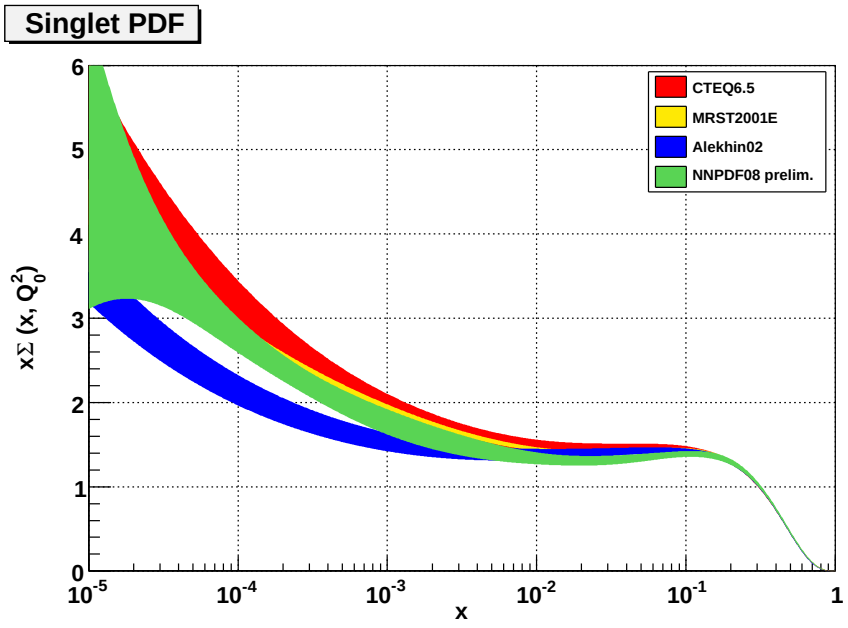


PARAMETRIZATION INDEPENDENCE:
LACK OF BIAS

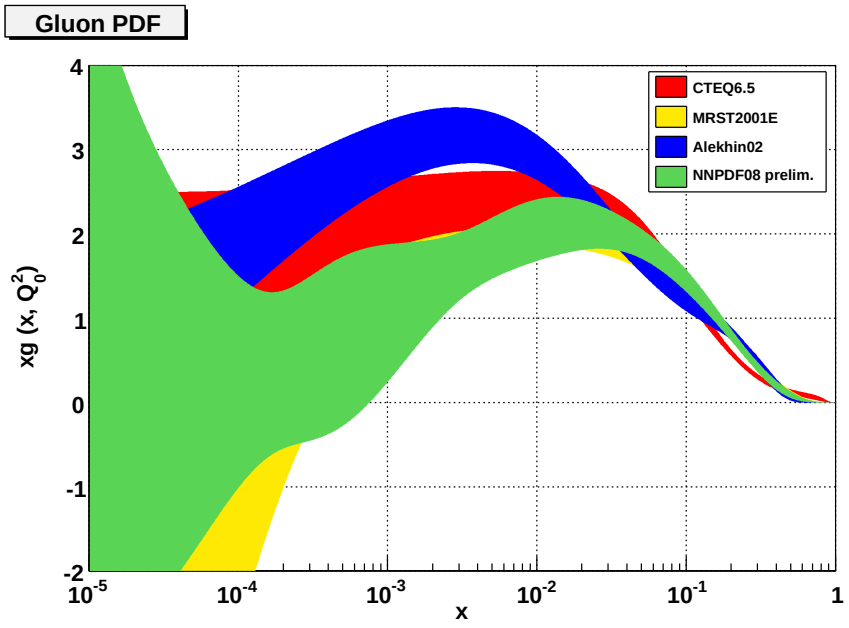
- IRREGULAR OR KNOTTY SHAPES ALLOWED IF DATA FLUCTUATE
- STATISTICS SHOW WHETHER THE EFFECT IS REAL

100 REPLICAS

SINGLET



GLUON

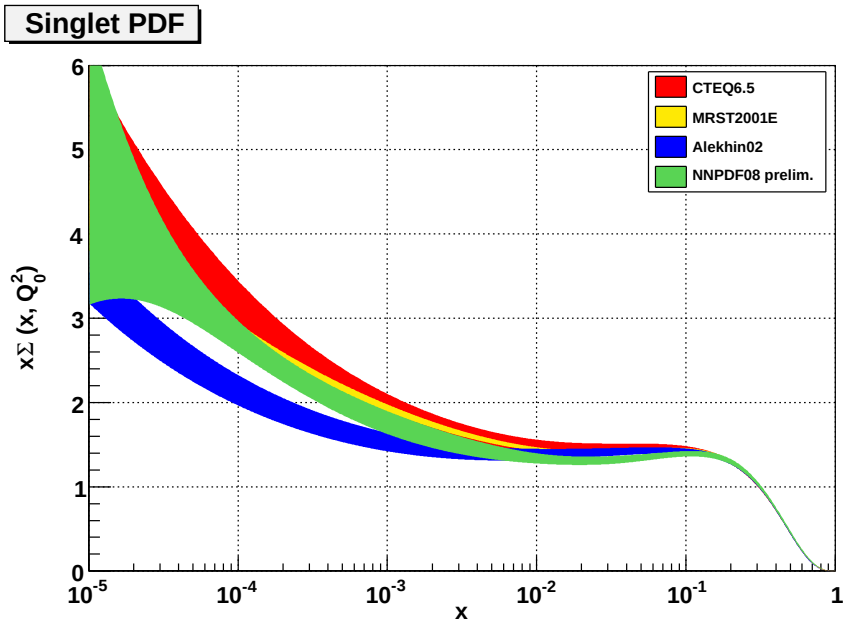


PARAMETRIZATION INDEPENDENCE:
LACK OF BIAS

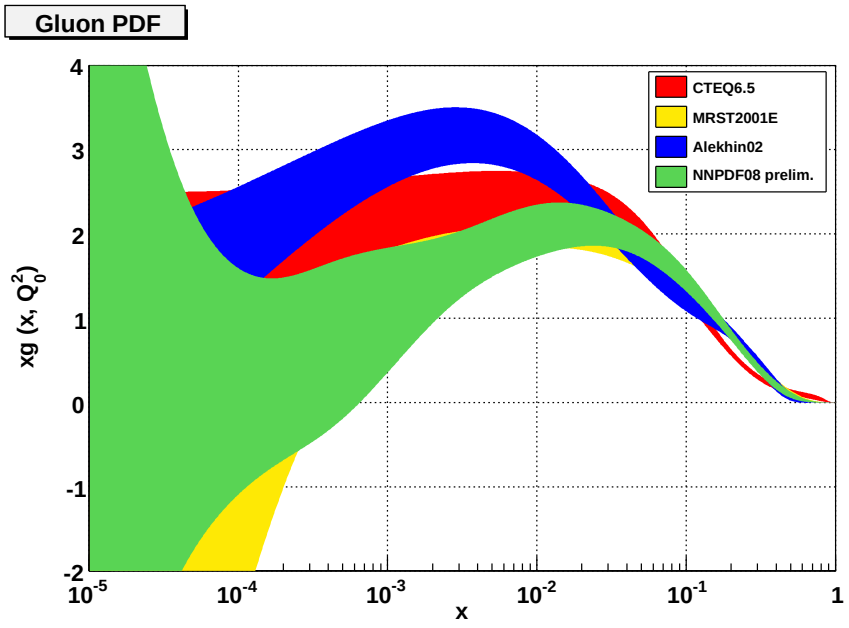
- IRREGULAR OR KNOTTY SHAPES ALLOWED IF DATA FLUCTUATE
- STATISTICS SHOW WHETHER THE EFFECT IS REAL

200 REPLICAS

SINGLET



GLUON

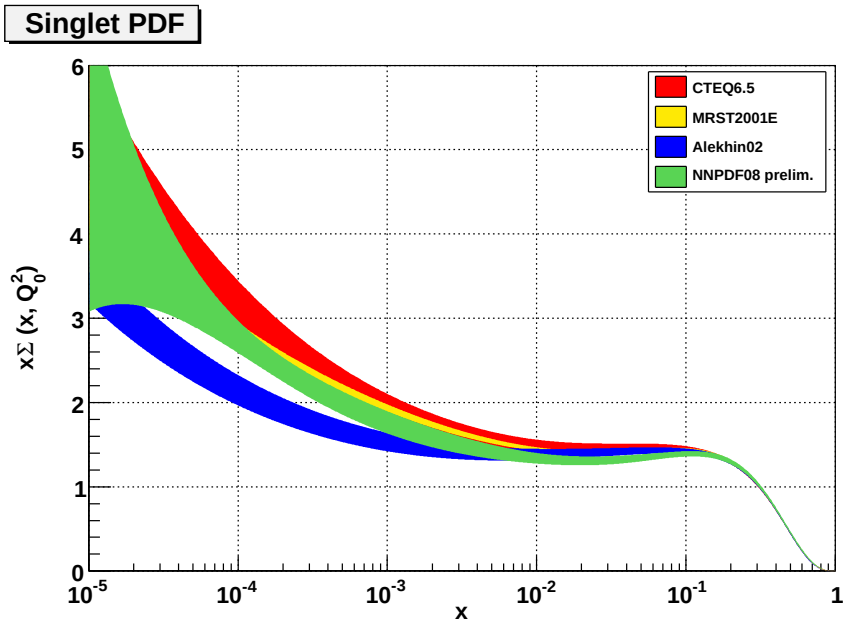


PARAMETRIZATION INDEPENDENCE:
LACK OF BIAS

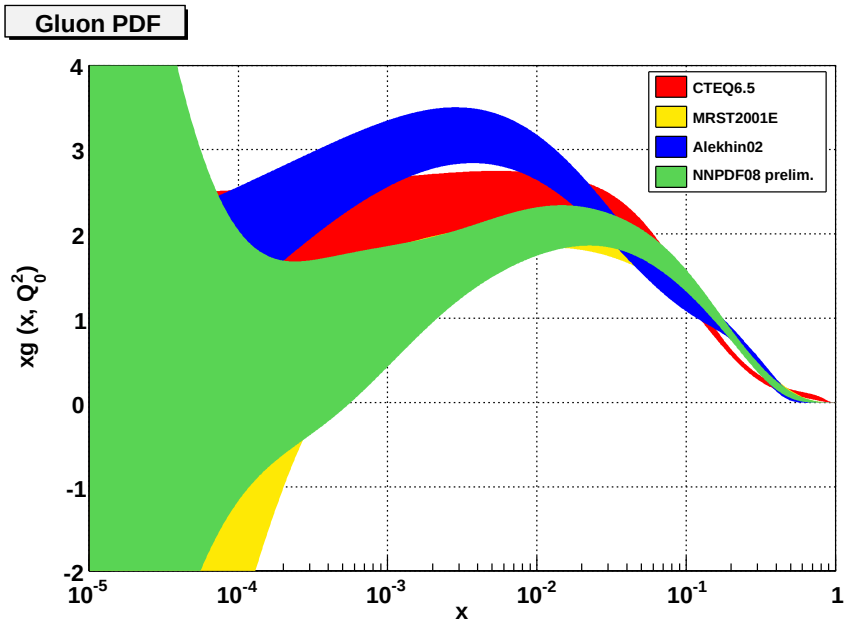
- IRREGULAR OR KNOTTY SHAPES ALLOWED IF DATA FLUCTUATE
- STATISTICS SHOW WHETHER THE EFFECT IS REAL

400 REPLICAS

SINGLET



GLUON

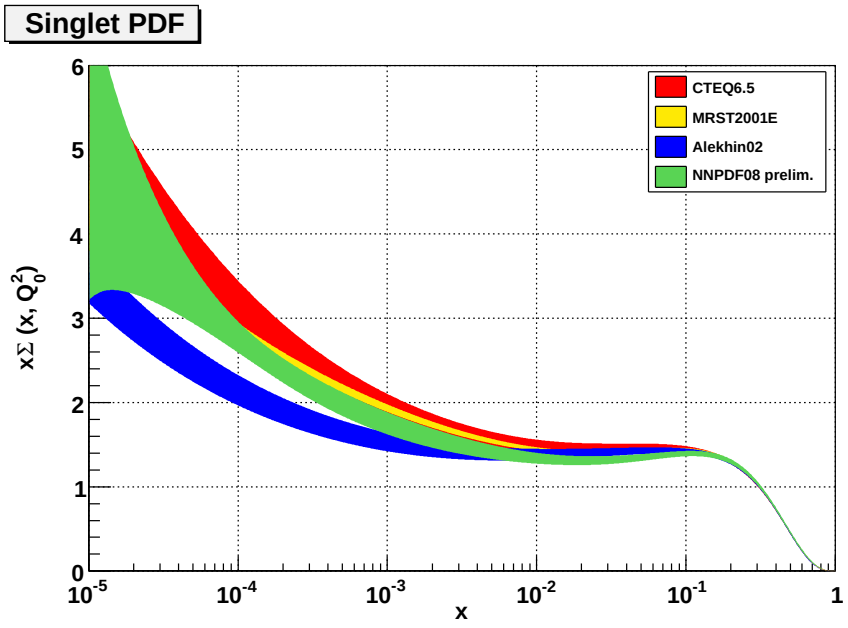


PARAMETRIZATION INDEPENDENCE:
LACK OF BIAS

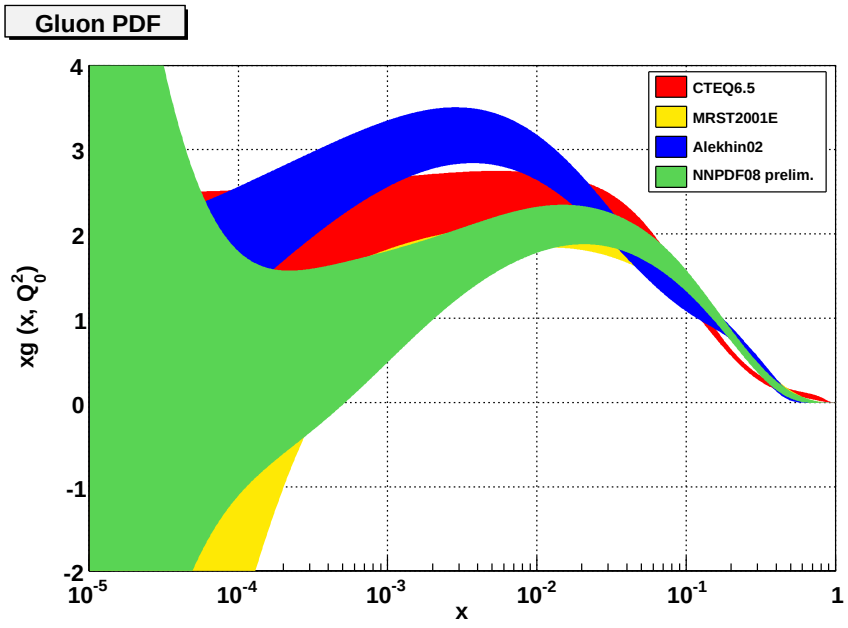
- IRREGULAR OR KNOTTY SHAPES ALLOWED IF DATA FLUCTUATE
- STATISTICS SHOW WHETHER THE EFFECT IS REAL

1000 REPLICAS

SINGLET



GLUON



PARAMETRIZATION INDEPENDENCE:
STATISTICAL STABILITY

COMPARE DISTANCE IN UNITS OF SIGMA OF RESULTS OBTAINED WITH DIFFERENT ASSUMPTIONS

- DISTANCE IN UNITS OF SIGMA

$$\langle d[q]\rangle = \sqrt{\left\langle \frac{\left(\langle q_i\rangle_{(1)} - \langle q_i\rangle_{(2)}\right)^2}{\sigma^2[q_i^{(1)}] + \sigma^2[q_i^{(2)}]} \right\rangle_{\rm dat}}$$

- NOTE σ $\Rightarrow$ ERROR ON AVERAGE
  = (ERROR ON q_i)$/\sqrt{N}$
  with 100 replicas,$d = 1$
  $\rightarrow$ fits differ by 1/10 of nominal error

- TEST PREDICTIONS
  FOR CENTRAL VALUES & ERRORS

DISTANCE BETWEEN STANDARD & FIT WITH SMALLER NEURAL NETS
2-4-3-1 VS 2-5-3-1 ARCHITECTURE
(31 vs. 37 parms per net)

	DATA	EXTRAPOLATION
SINGLET	$0.005 \le x \le 0.1$	$10^{-4} \le x \le 10^{-3}$
$\langle d[q]\rangle$	0.96	1.32
$\langle d[\sigma]\rangle$	1.23	1.32
GLUON	$0.005 \le x \le 0.1$	$10^{-4} \le x \le 10^{-3}$
$\langle d[q]\rangle$	1.40	1.13
$\langle d[\sigma]\rangle$	1.17	1.06
VALENCE	$0.1 \le x \le 0.6$	$0.03 \le x \le 0.3$
$\langle d[q]\rangle$	1.40	0.93
$\langle d[\sigma]\rangle$	1.09	0.96
TRIPLET	$0.05 \le x \le 0.75$	$0.01 \le x \le 0.1$
$\langle d[q]\rangle$	1.05	1.09
$\langle d[\sigma]\rangle$	1.68	2.5

PARAMETRIZATION INDEPENDENCE: THE "HERALHC BENCHMARK"

$Q^2 > 9$ GeV^2; $W^2 > 15$ GeV^2

REDUCED DATASET $\Rightarrow$ WIDER ERROR BAND

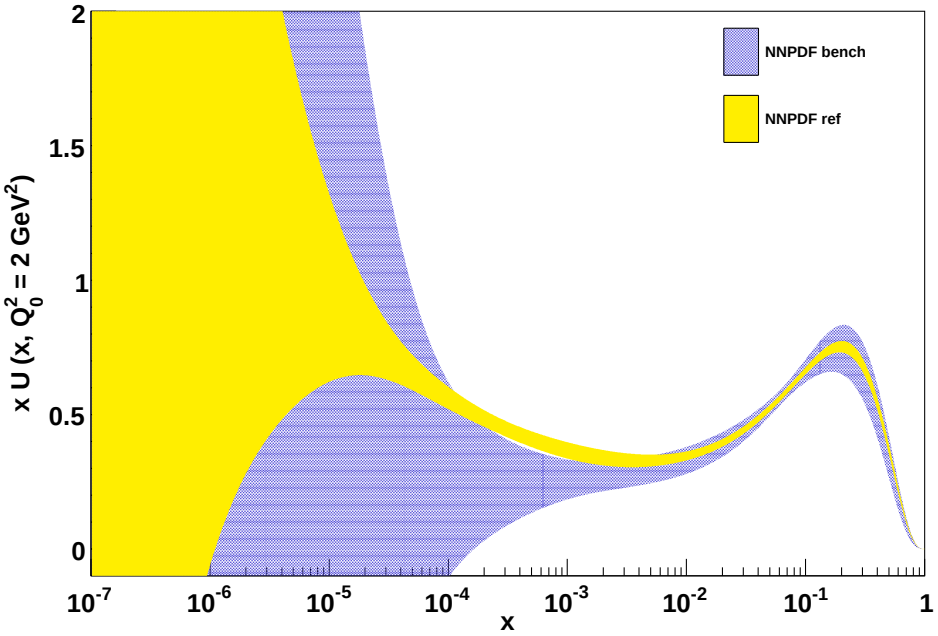
from 3161 to 773 datapoints

reduced info on small x sea (no low Q^2 data)

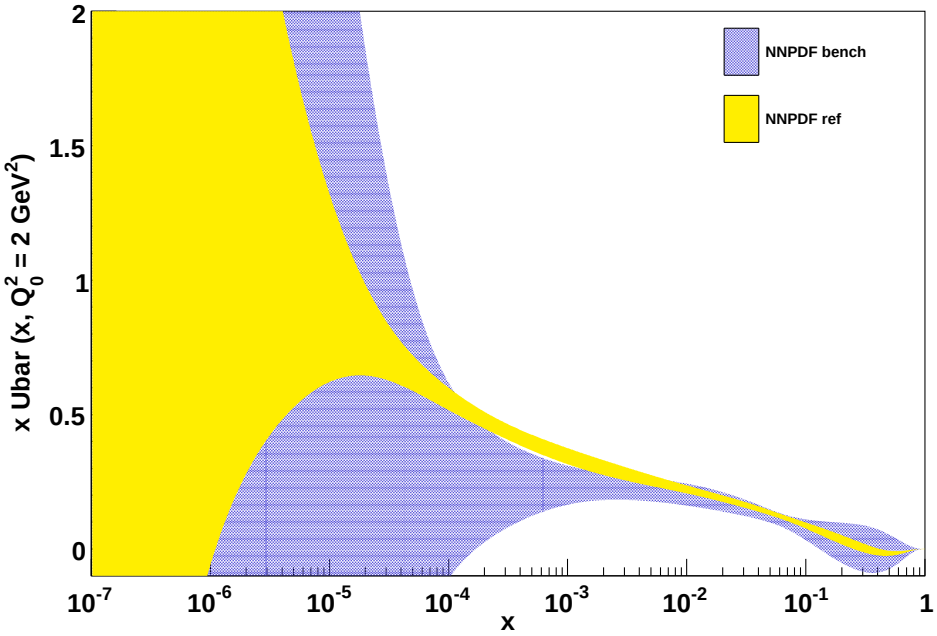
& large x valence (no neutrino data)

NAME	DATA POINTS	TARGET
NMC_PD	73	F_2^d / F_2^p
NMC	95	F_2^p
BCDMS	322	F_2^p
ZEUS97	206	F_2^p
H1LX97	77	F_2^p
TOTAL	773	

UP QUARK



UP ANTIQUARK



RESULTS COMPATIBLE TO WITHIN LESS THAN TWO SIGMA

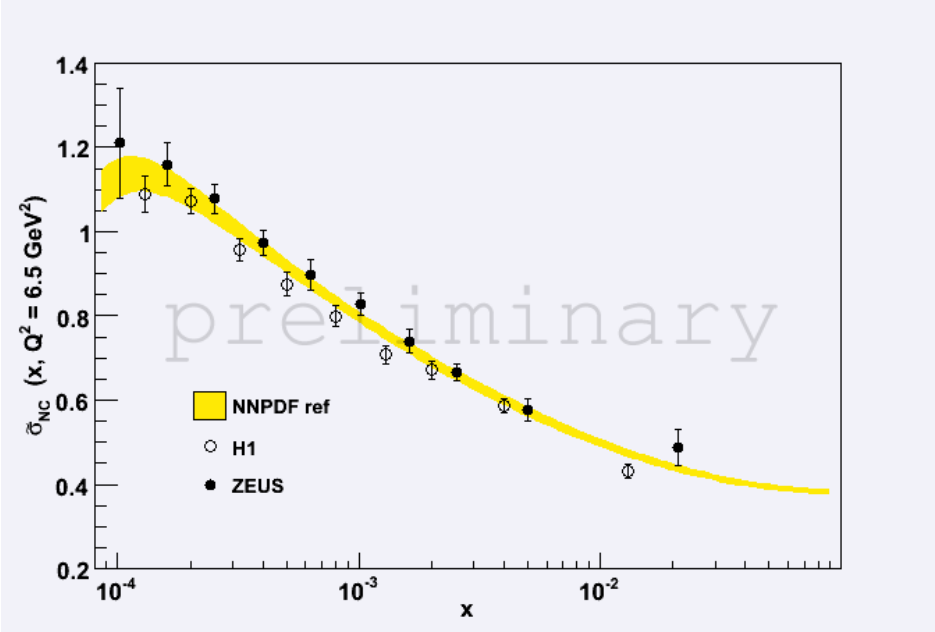
PARAMETRIZATION INDEPENDENCE:

THE "HERALHC BENCHMARK":INCOMPATIBLE DATA

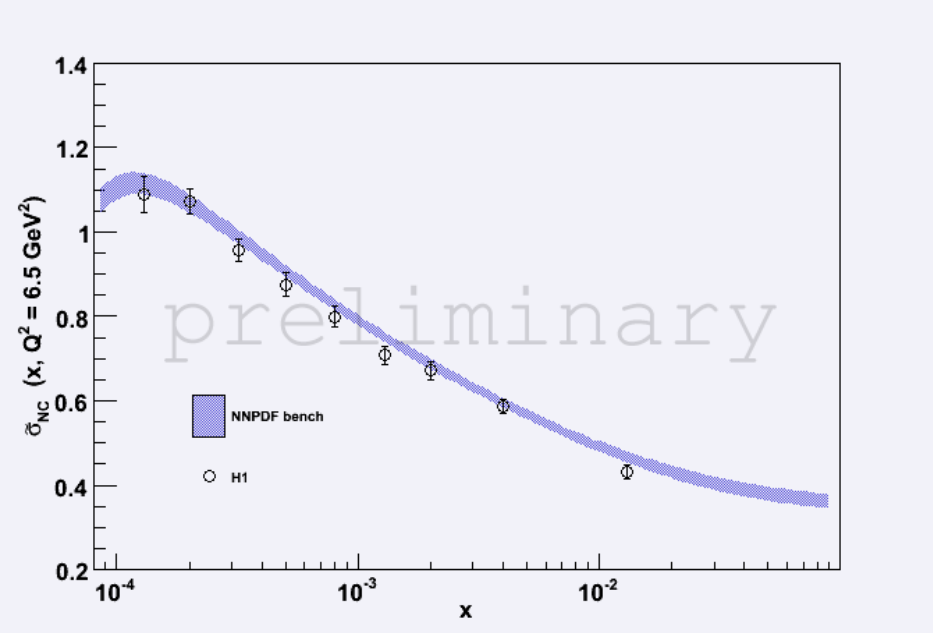
THE SMALL x CC REDUCED CROSS SECTION

results presented at the HERALHC workshop (NNPDF preliminary fit)

FULL DATASET



REDUCED "BENCHMARK" DATASET



NO ERROR REDUCTION WHEN DATA IN WIDER DATA SET ARE INCOMPATIBLE

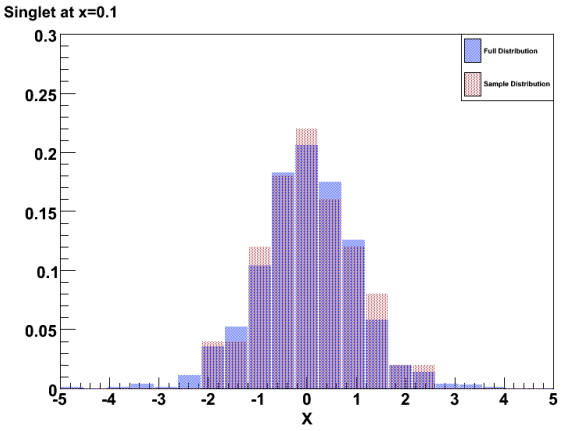
DELIVERY:
RESTRICTED SAMPLE OF REPLICAS

- WIDE SAMPLE OF PSEUDODATA ENDURES NO BIAS
- IMPRACTICAL TO AVERAGE OVER THOUSAND(S) OF REPLICAS
- SELECT SUBSET OF REPLICAS WITH APPROXIMATELY SAME STATISTICAL DISTRIBUTION AS FULL SET

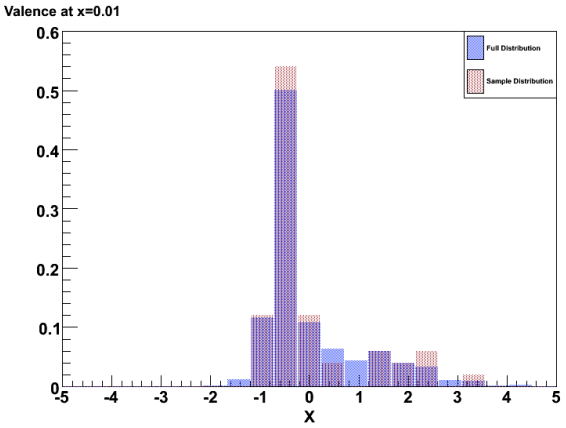
- construct histogram for # of replicas n sigma away from mean
- compare result for subset & full
- minimize relative entropy of two histograms
  $S =$
  $=$
  $$\sum_{\text{i bins}} \left(p_i^{(1)} - p_i^{(2)}\right) \ln \frac{p_i^{(1)}}{p_i^{(2)}}$$
- select with genetic algorithm subset which minimizes S

FULL (1000) VS. REDUCED (50) PROBABILITY HISTOGRAMS

SINGLET AT $x = 0.1$

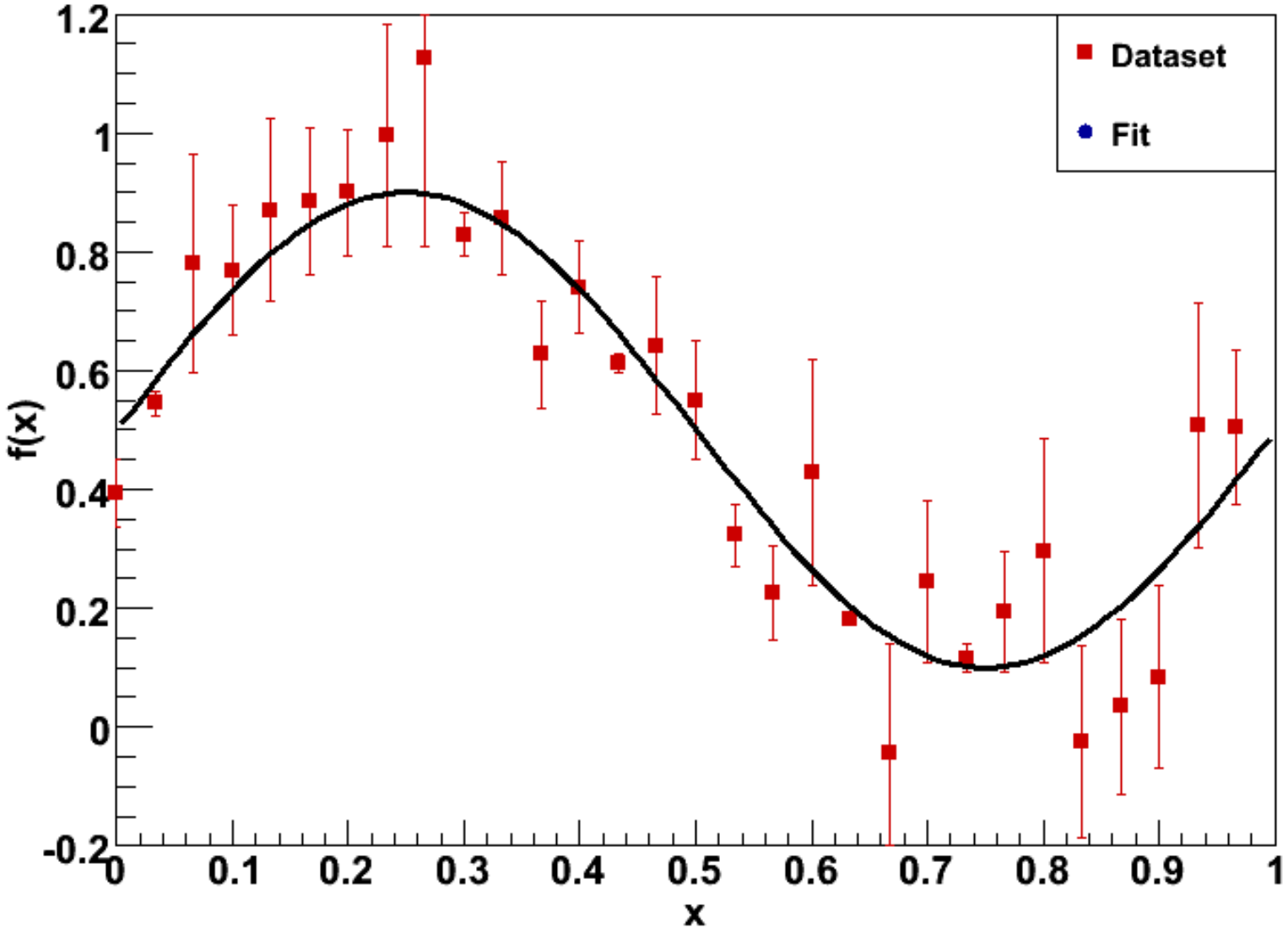


VALENCE AT $x = 0.01$



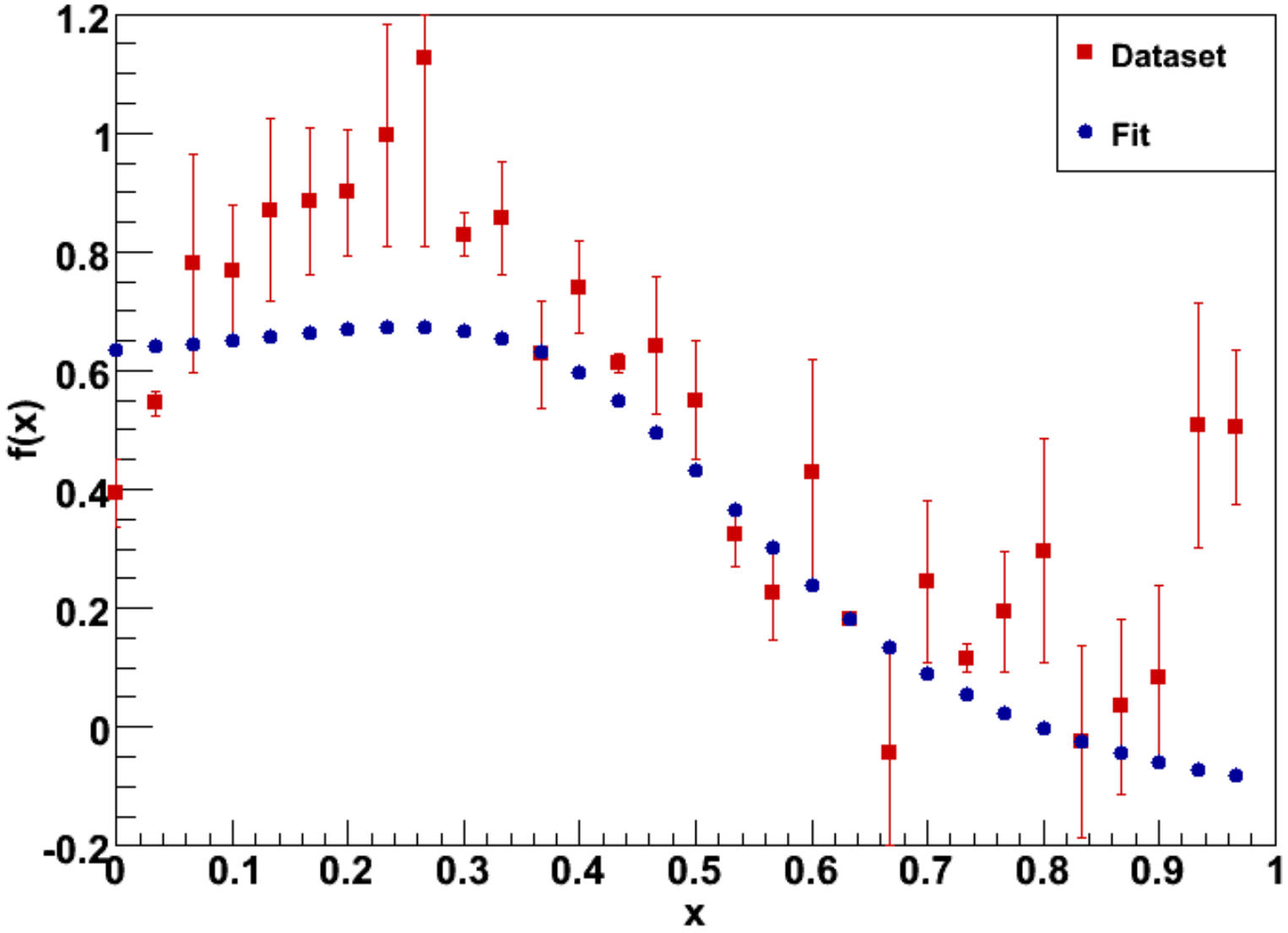
Optimal fitting vs. Overlearning: an example

the true function



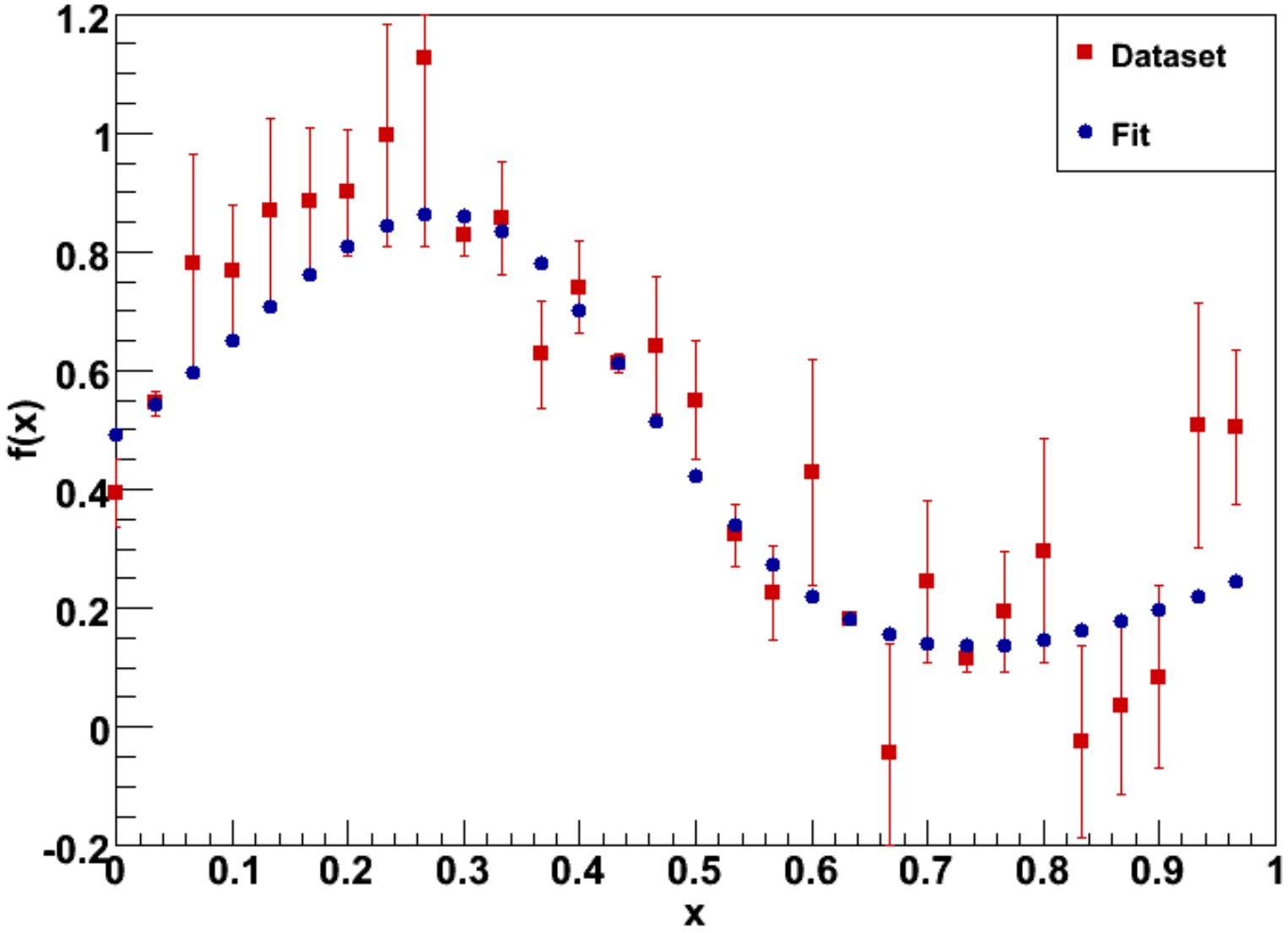
Optimal fitting vs. Overlearning: an example

UNDERLEARNING



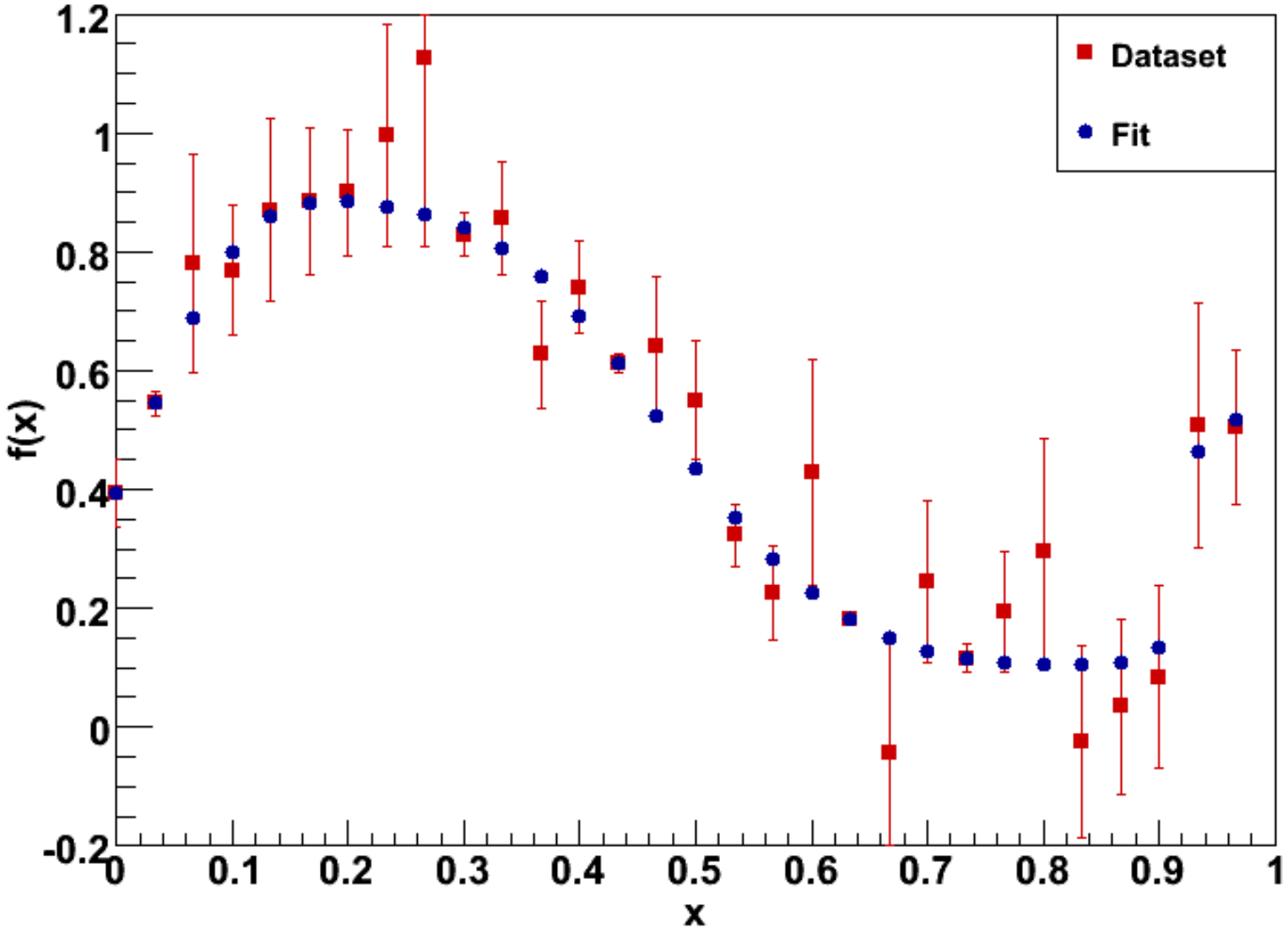
Optimal fitting vs. Overlearning: an example

optimal fit

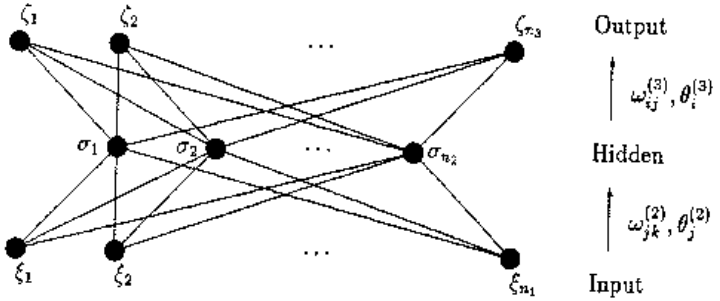


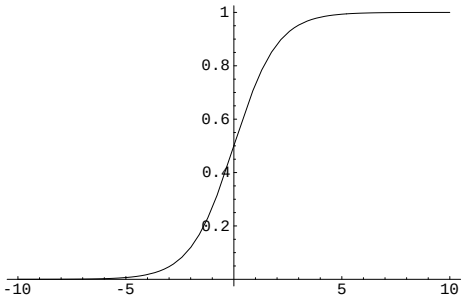
Optimal fitting vs. Overlearning: an example

overlearning



WHAT ARE NEURAL NETWORKS?





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

$$g(x) = \frac{1}{1+e^{-\beta x}}$$

JUST ANOTHER SET OF BASIS FUNCTIONS!

A 1-2-1 NN: $f(x) = \dfrac{1}{1+e^{\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1+e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1+e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}}}}$

ANY FUNCTION CAN BE REPRESENTED BY A SUFFICIENTLY BIG NEURAL NETWORK

LESS PARAMETERS $\rightarrow$ SMOOTHER FUNCTIONS

WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?

WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?
IN A NEURAL FIT, SMOOTHNESS DECREASES AS FIT QUALITY IMPROVES:

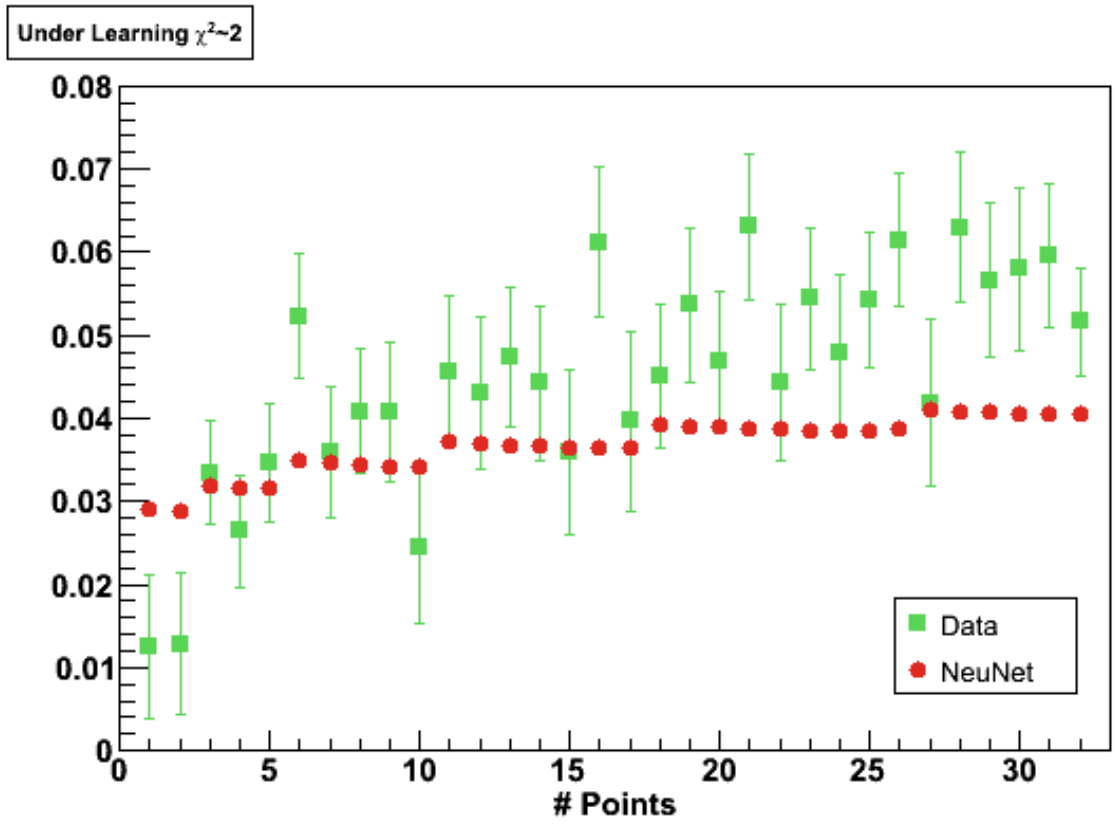
WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?
IN A NEURAL FIT, SMOOTHNESS DECREASES AS FIT QUALITY IMPROVES:

UNDERLEARNING



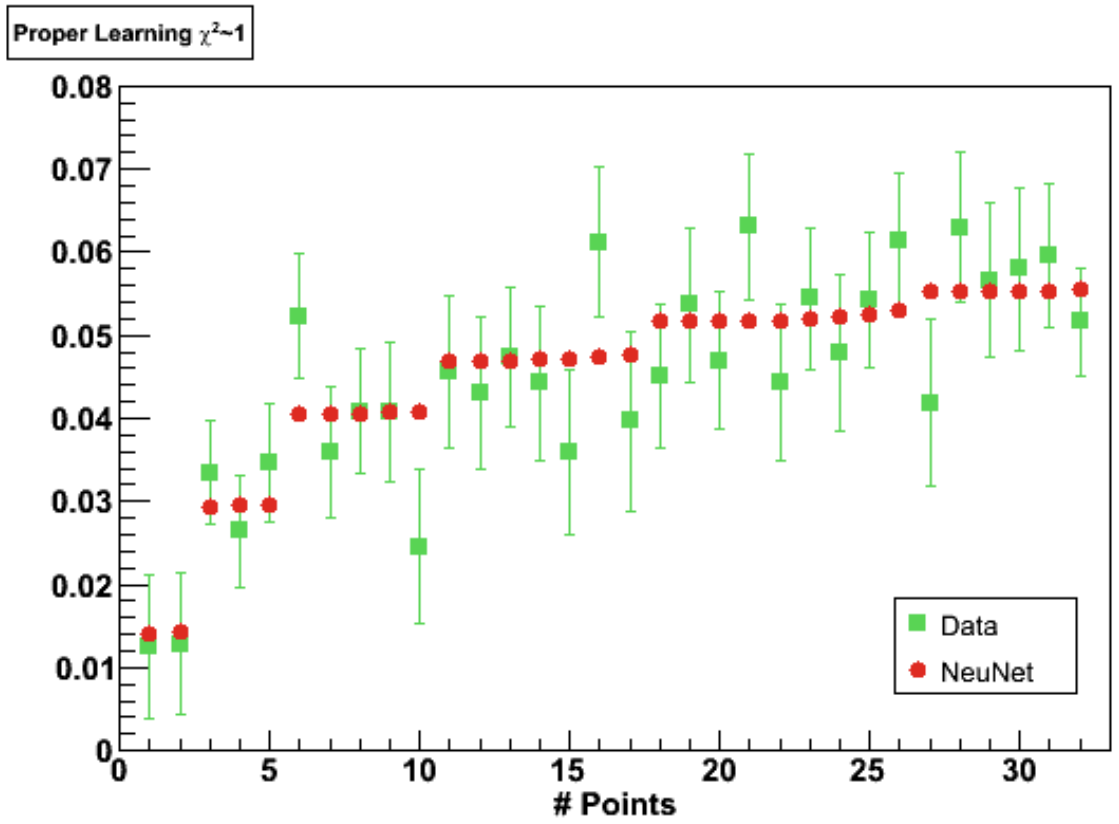
WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?
IN A NEURAL FIT, SMOOTHNESS DECREASES AS FIT QUALITY IMPROVES:

PROPER LEARNING



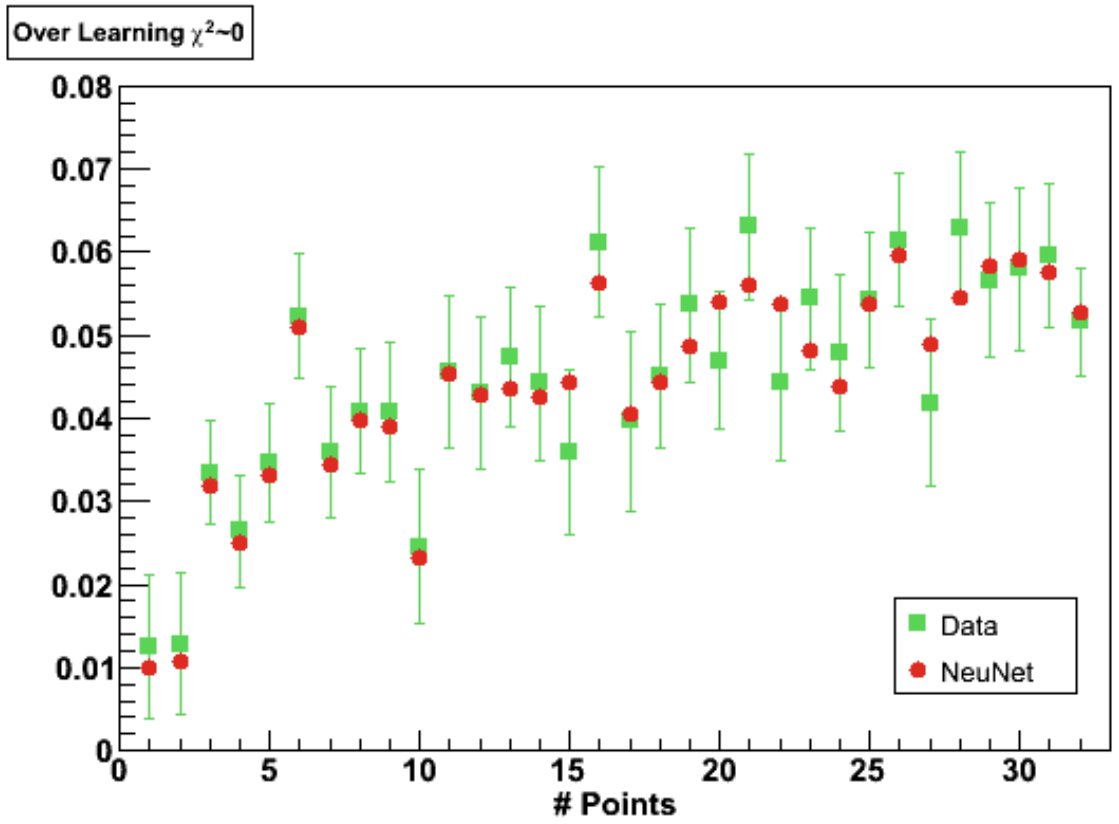
WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?
IN A NEURAL FIT, SMOOTHNESS DECREASES AS FIT QUALITY IMPROVES:

OVERLEARNING



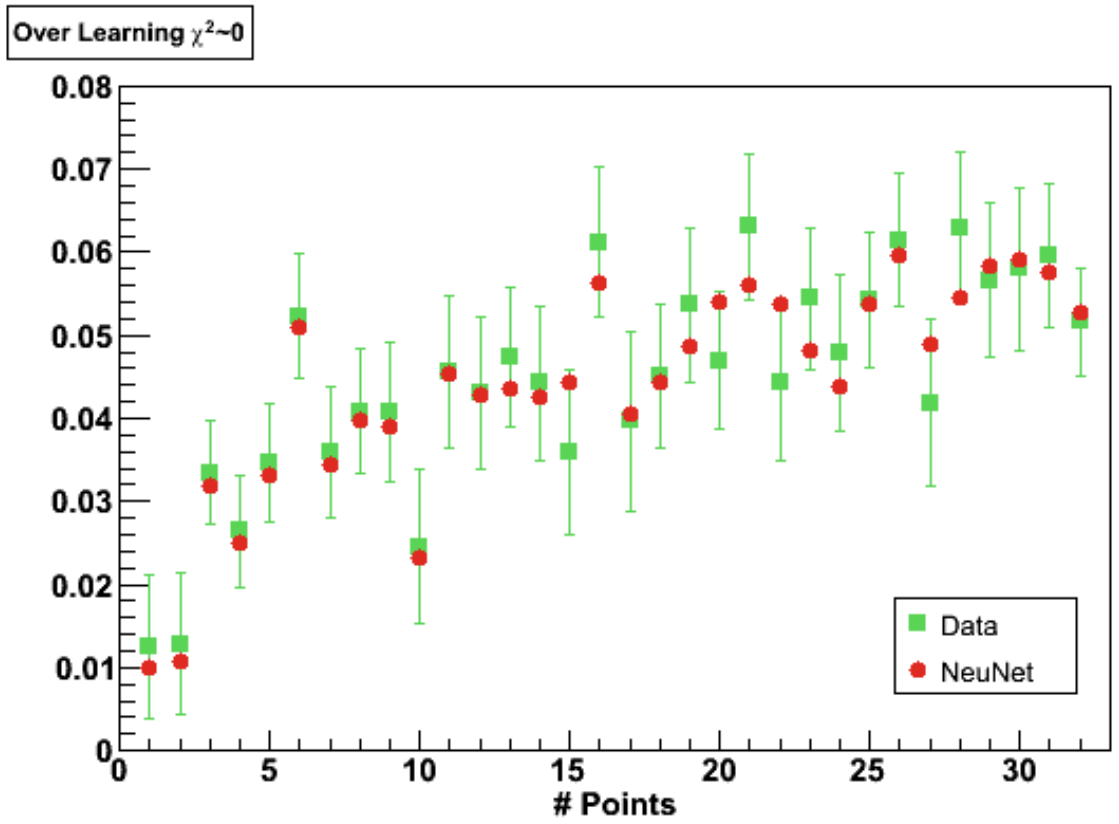
WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?
IN A NEURAL FIT, SMOOTHNESS DECREASES AS FIT QUALITY IMPROVES:

OVERLEARNING



A: STOP THE FIT BEFORE OVERLEARNING SETS IN!

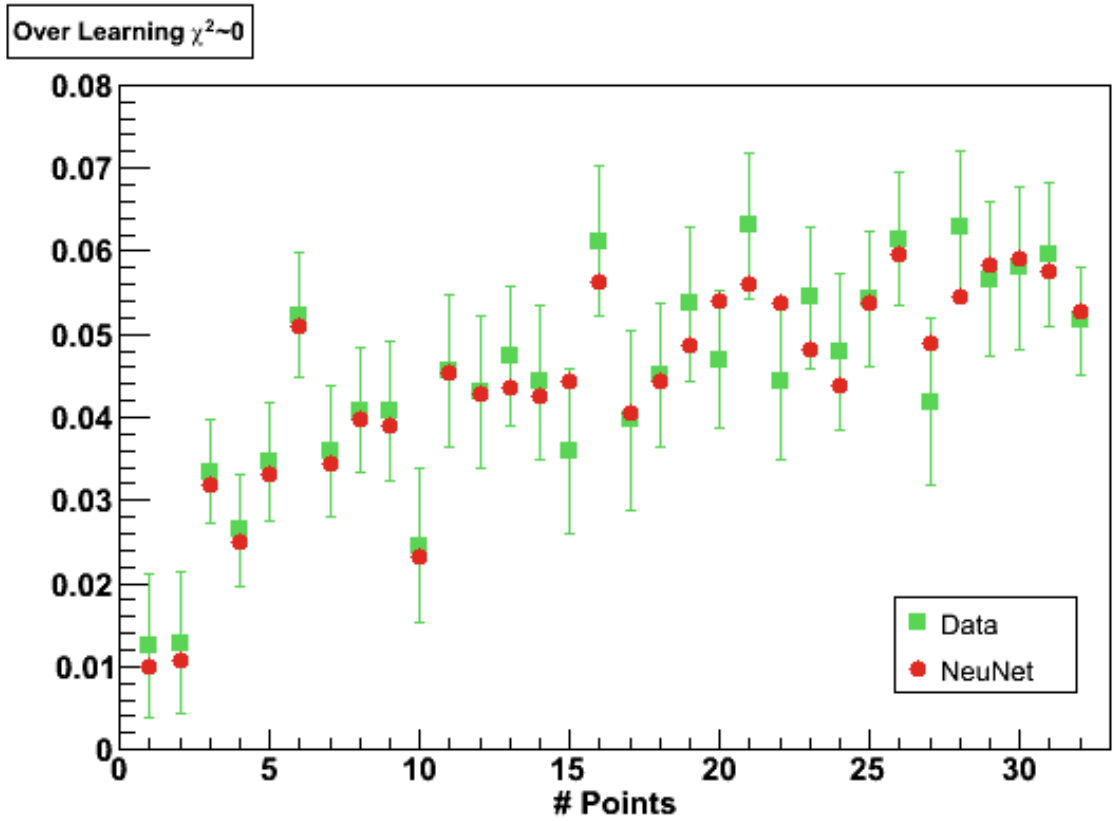
WHY NEURAL NETWORKS?

IN A STANDARD FIT, ONE LOOKS FOR MINIMUM χ^2 WITH GIVEN FINITE PARM.

- IF THE BASIS IS TOO LARGE, THE FIT NEVER CONVERGES
- IF THE BASIS IS TOO SMALL, THE FIT IS BIASED

Q: HOW CAN ONE BE SURE THAT THE COMPROMISE IS UNBIASED?
IN A NEURAL FIT, SMOOTHNESS DECREASES AS FIT QUALITY IMPROVES:

OVERLEARNING

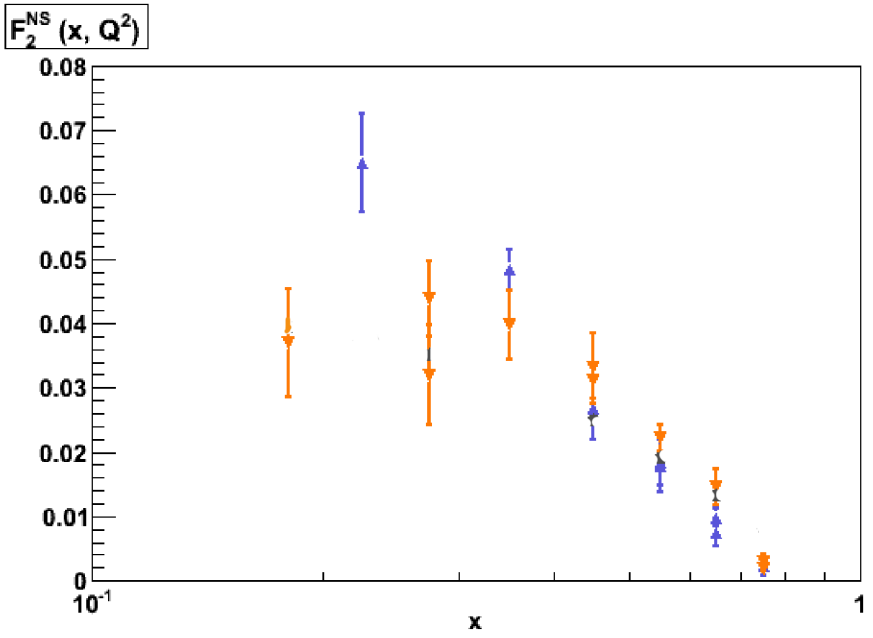


A: STOP THE FIT BEFORE OVERLEARNING SETS IN!
COULD BE DONE WITH STANDARD PARAMETRIZATIONS, BUT VERY INEFFICIENTLY

THE STOPPING CRITERION

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

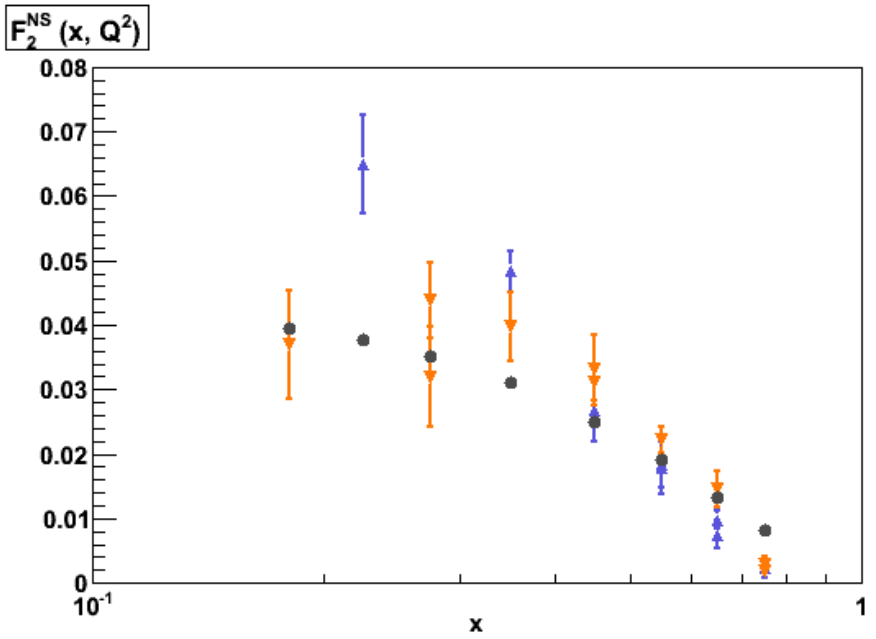


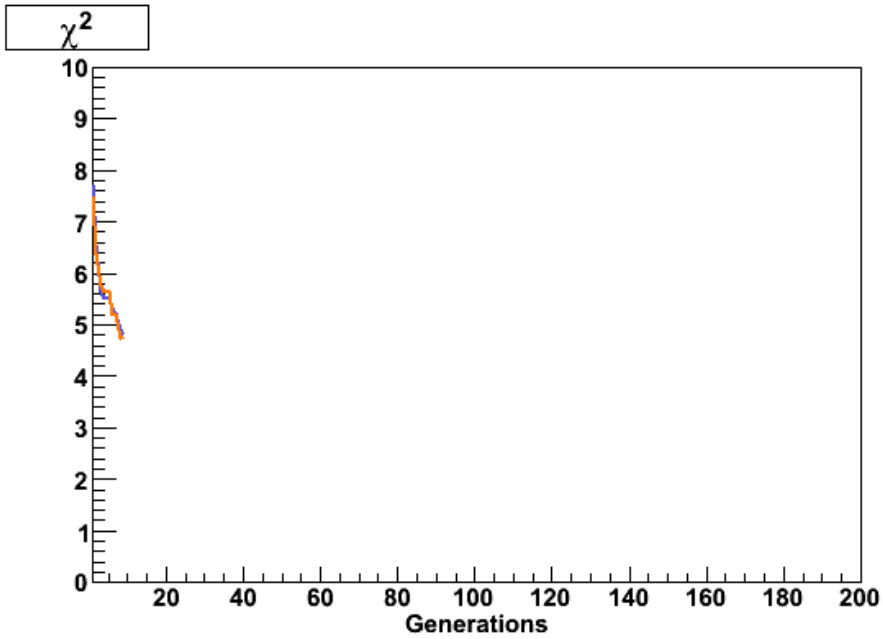
THE STOPPING CRITERION

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!



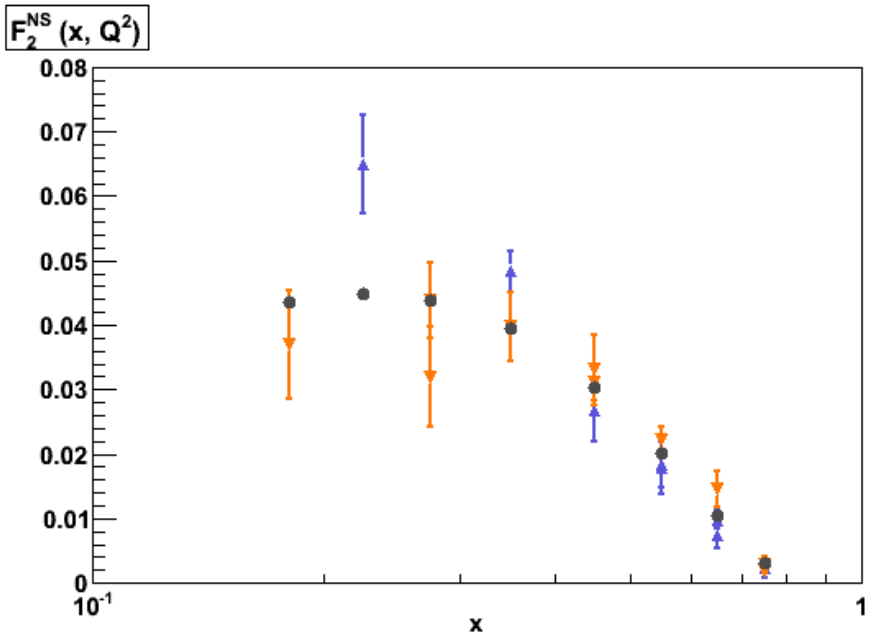


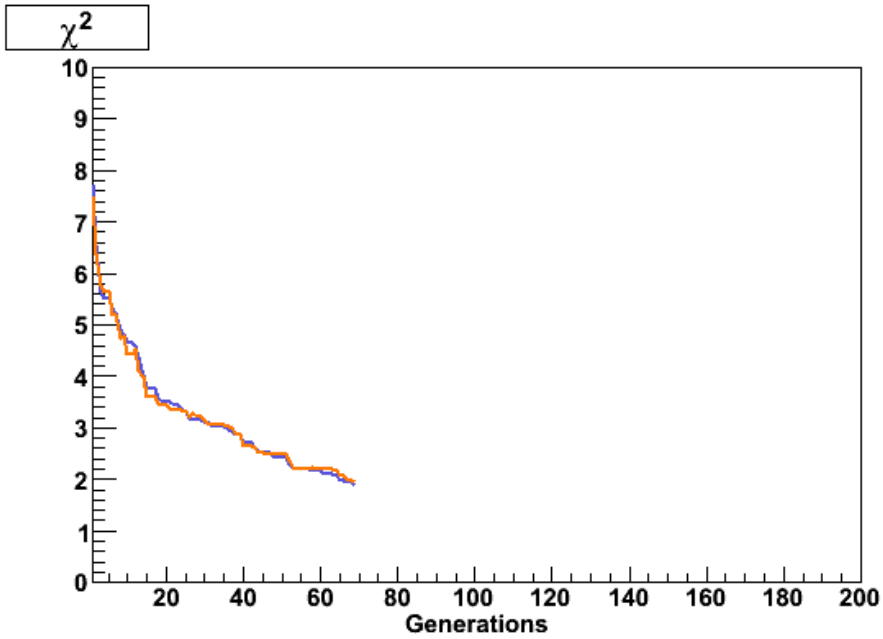
THE STOPPING CRITERION

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!



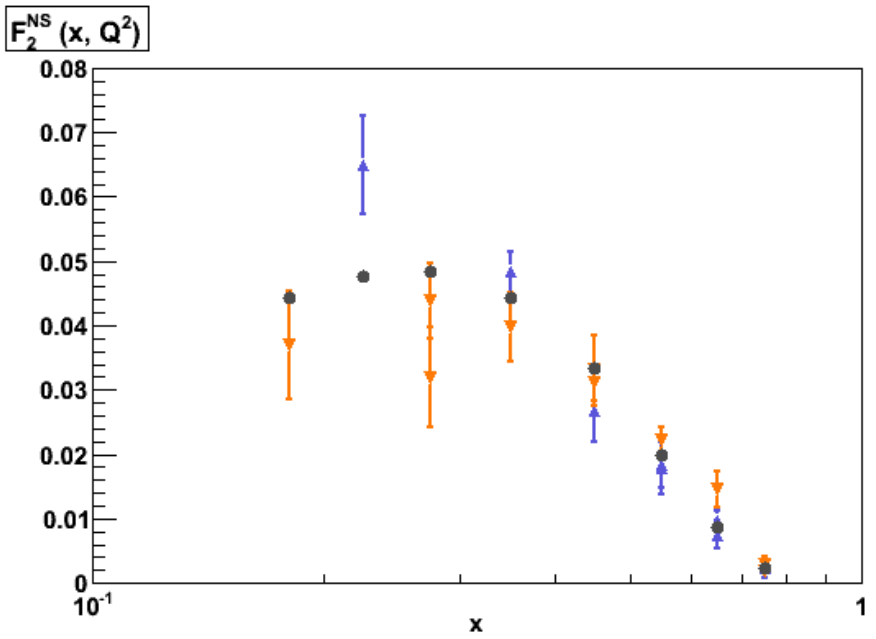


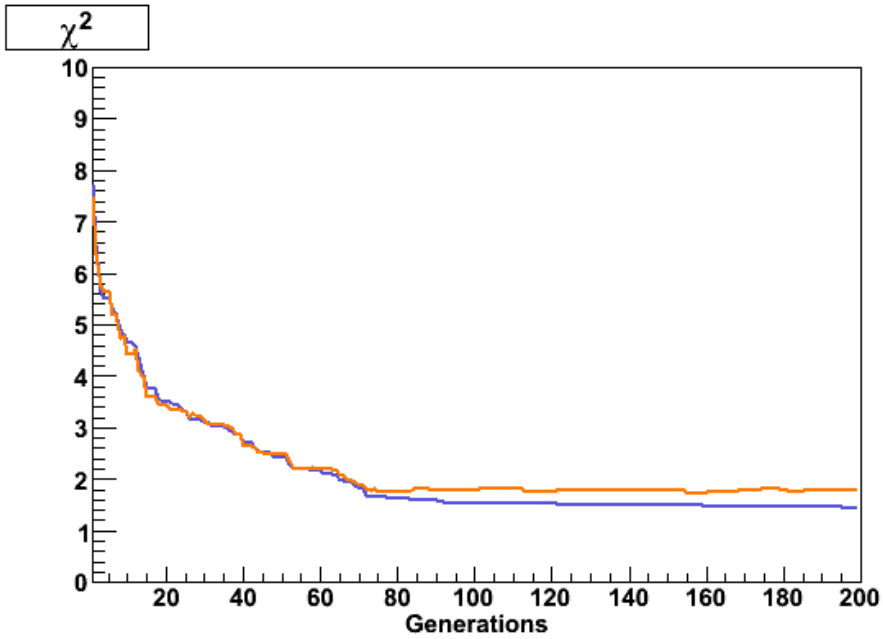
THE STOPPING CRITERION

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!





Monte Carlo data generation

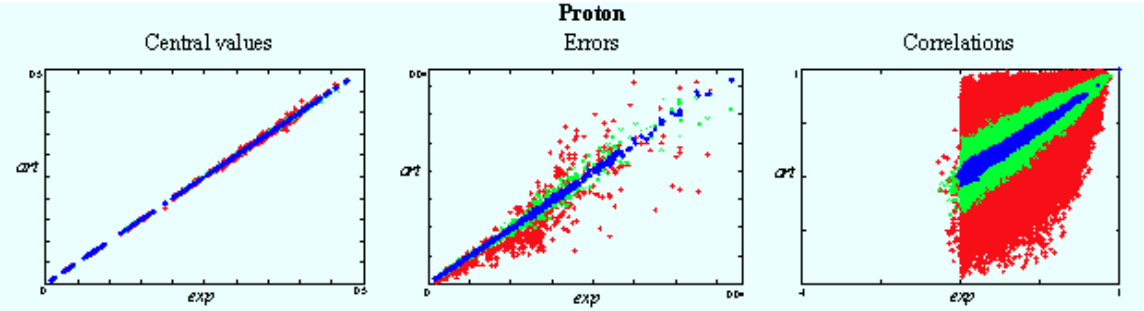
- BCDMS+ NMC proton & deuteron F_2 data (full correlated systematics available), taken at 4 beam energies
- on top of stat. errors, 4 systematics + 1 normalization (NMC) or 6 systematics + 1 absolute & 2 relative normalizations (BCDMS), with various forms of correlation (full, or for each target, or for each beam energy)

Generate data according to a multigaussian distribution

$$F_i^{(art)\,(k)} =$$

$$(1 + r_5^{(k)}\,\sigma_N)\sqrt{1 + r_{i,6}^{(k)}\,\sigma_{N_t}}\sqrt{1 + r_{i,7}^{(k)}\,\sigma_{N_b}}\left[F_i^{(exp)} + \frac{r_{i,1}^{(k)}\,f_b + r_{i,2}^{(k)}\,f_{i,s} + r_{i,3}^{(k)}\,f_{i,r}}{100}F_i^{(exp)} + r_{i,s}^{(k)}\,\sigma_s^i\right]$$

r univariate gaussian random nos., one $r_{i,s}$ for each data, but single $r_{i,j}$ for all correlated data



scatter plot art. vs. exp. for 10 (red) 100 (green) and 1000 (blue) replicas

need 1000 replicas to reproduce correlations to percent accuracy

QUESTION: (F. OLNESS)

HOW "STRANGE" IS THE STRANGE PDF?

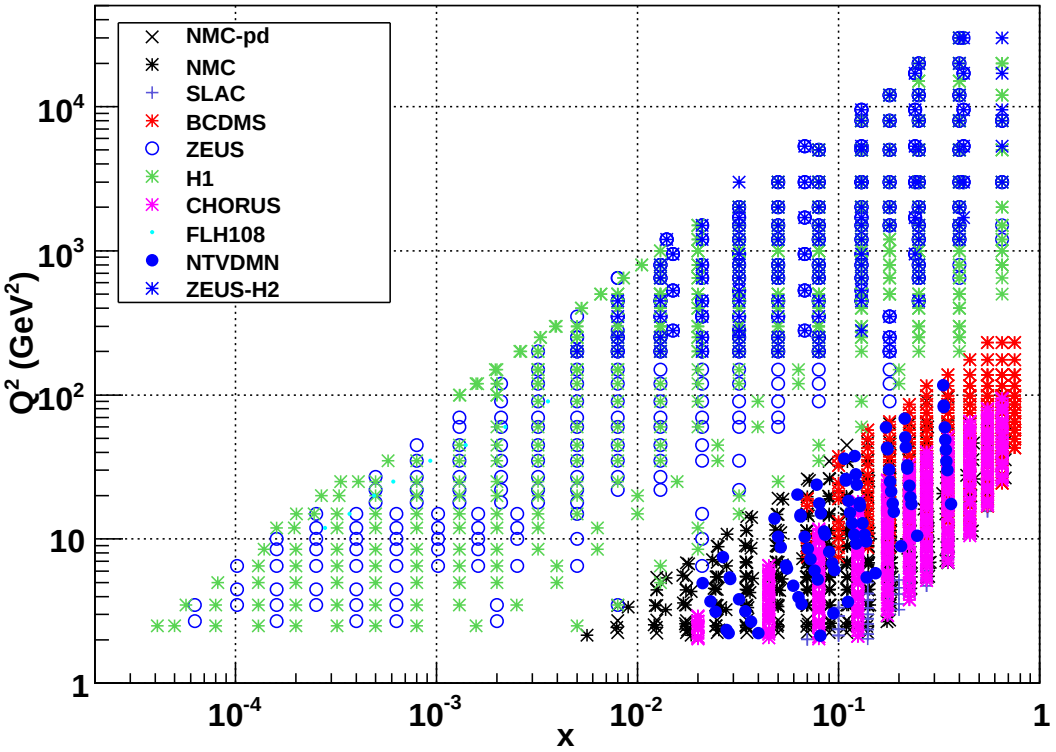
QUESTION: (F. OLNESS)

HOW "STRANGE" IS THE STRANGE PDF?

ANSWER: THE NNPDF1.2 SET

THE DATA SET

- SEVEN PDFS INDEP. PARAMETRIZED
- 74 EXTRA FREE PARMS ($s^{\pm}$) IN COMPARISON TO NNPDF1.0
- NUTEV DIMUON DATA USED TO CONSTRAIN STRANGENESS

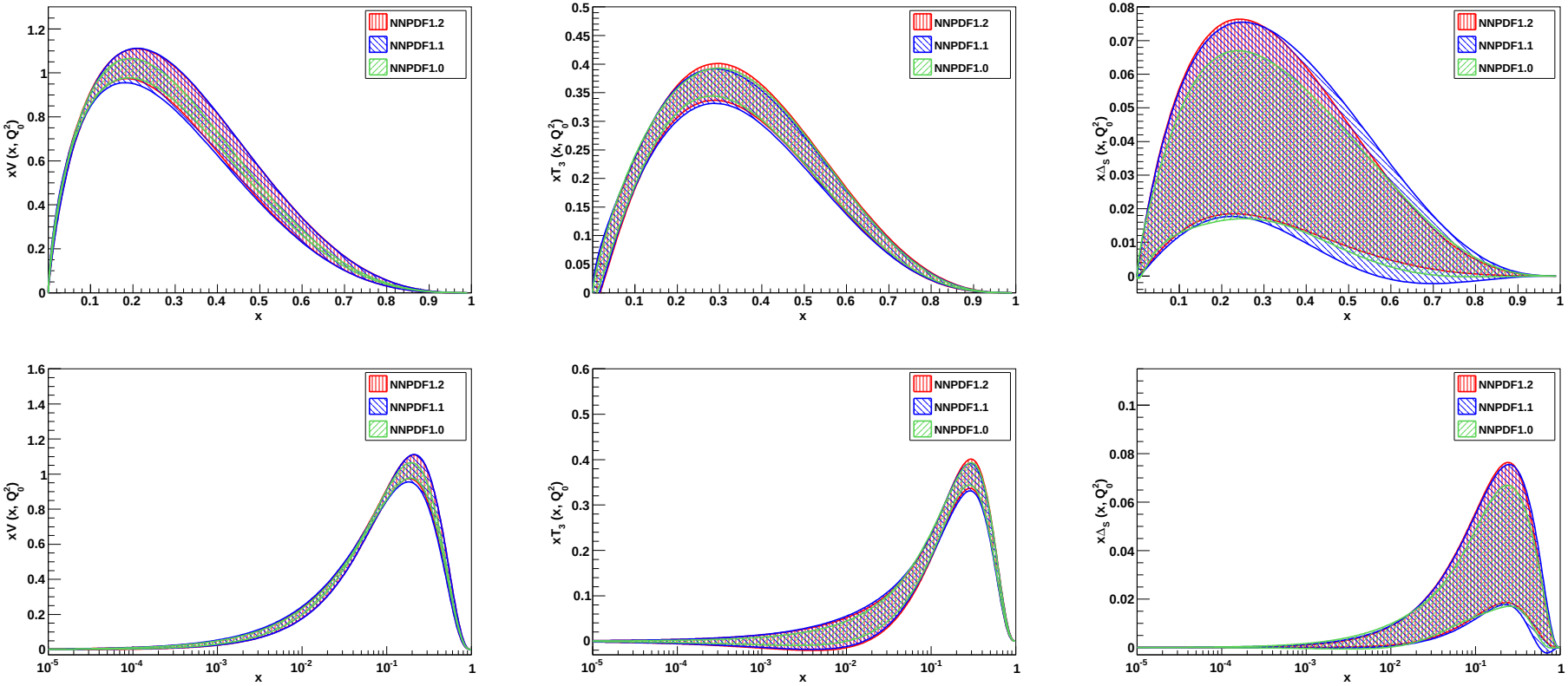


STABILITY

COMPARISON TO PREVIOUS NNPDF SETS

- NNPDF1.0: $s(x,Q_0^2) = \bar{s}(x,Q_0^2)$, $s+\bar{s} = \frac{1}{2}(\bar{u}+\bar{d})$
- NNPDF1.1: s, $\bar{s}$ (ACTUALLY $s^{\pm}$) INDEP. PARAMETRIZED, NO DIMUON DATA
- NNPDF1.2: s, $\bar{s}$ INDEP. PARAMETRIZED, DIMUON DATA

NONSTRANGE PDFS

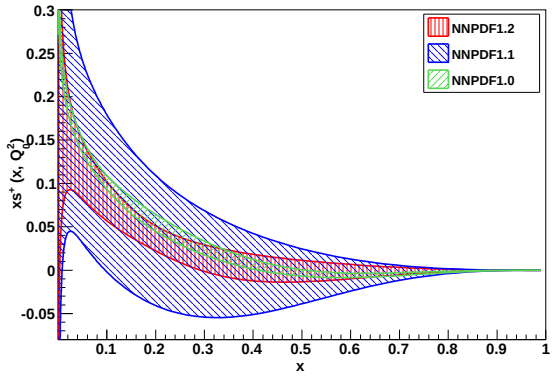


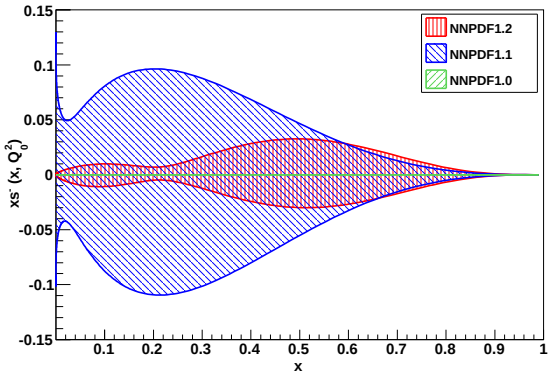
DETERMINING STRANGENESS

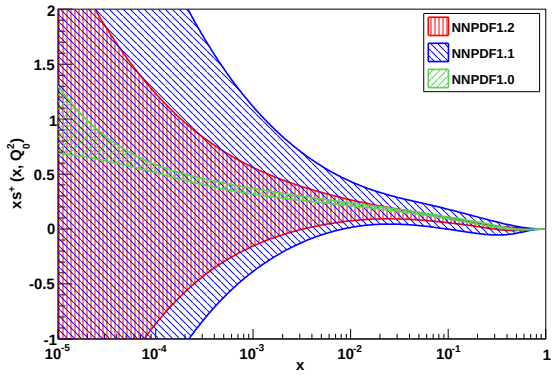
COMPARISON TO PREVIOUS NNPDF SETS

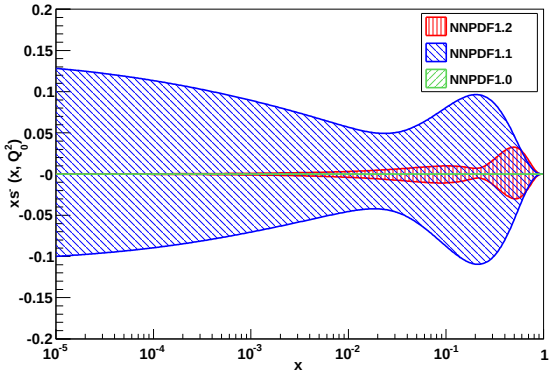
- NNPDF1.0: STRANGENESS UNCERTAINTY UNDERESTIMATED
- NNPDF1.1: STRANGENESS UNCERTAINTY HUGE
- NNPDF1.2: STRANGENESS UNCERTAINTY UNDER CONTROL

STRANGE PDFS







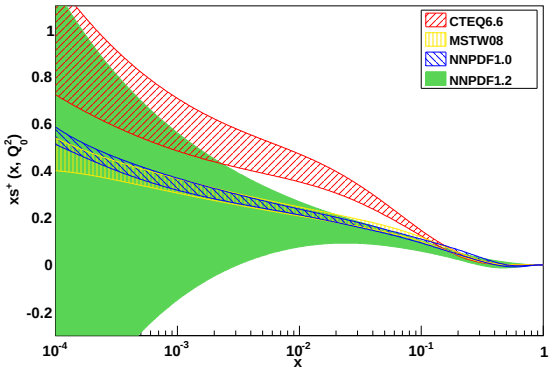


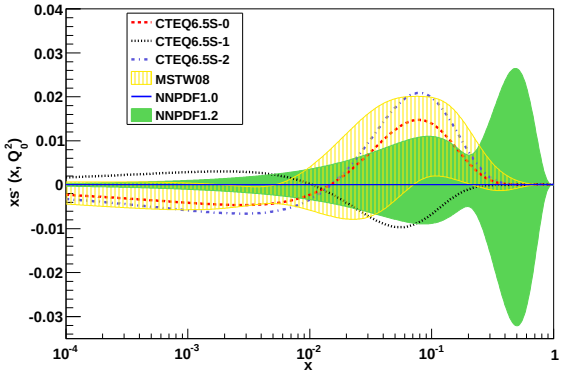
DETERMINING STRANGENESS

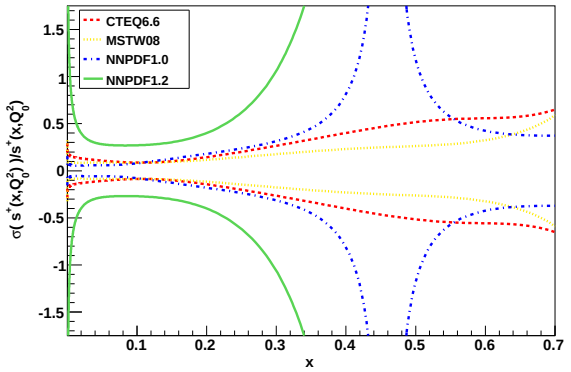
COMPARISON TO OTHER NNPDF SETS

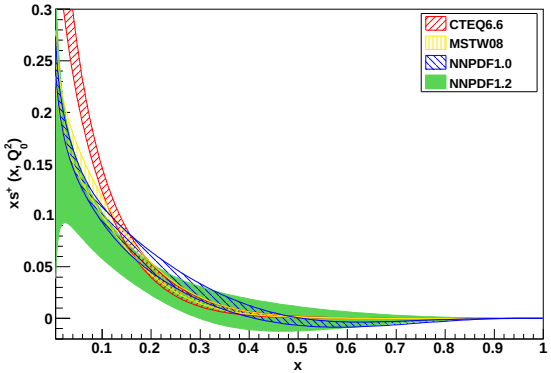
- CTEQ6.6: $s = \bar{s}$, s^+ PARM W. TWO FREE PARAMETERS
- MSTW08: s^+ & s^- PARM W. TWO FREE PARAMETERS EACH
- EVERYBODY ENFORCES STRANGENESS SUM RULE

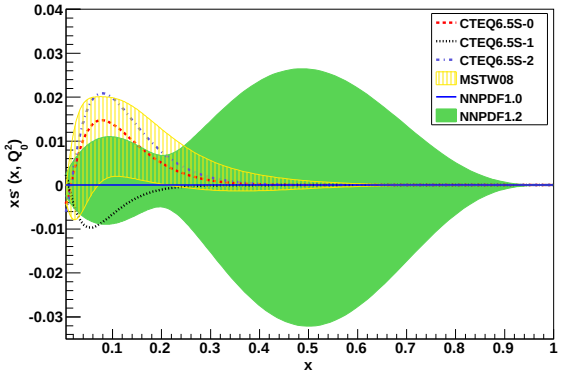
STRANGE PDFS

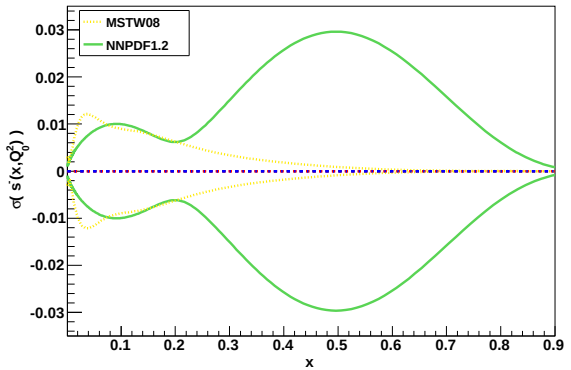












THE NUTEV ANOMALY

THE NUTEV ANOMALY

IS GONE

$$R_{\rm PW} \equiv \frac{\sigma(\nu\mathcal{N}\to\nu X)-\sigma(\bar\nu\mathcal{N}\to\bar\nu X)}{\sigma(\nu\mathcal{N}\to\ell X)-\sigma(\bar\nu\mathcal{N}\to\bar\ell X)} = \frac{1}{2}-\sin^2\theta_{\rm W}+\left(\frac{(U^-D^-)+(C^--S^-)}{\mathcal{Q}^-}\frac{1}{6}\left(3-7\sin^2\theta_{\rm W}\right)\right) ,$$

- STRANGENESS VALENCE MOMENTUM ASSUMED BY NUTEV TO VANISH
- UNCERTAINTY ON IT DOMINATES UNCERTAINTY ON THE PW RATIO